HAPHAZARD REALITY

IOT IS A METAPHOR

Shoumen Palit Austin Datta

Principles and Practice of Connectivity and Convergence
The education of a boy, may influence the fate of a man.
The education of a girl, may change the destiny of a nation.
HAPHAZARD REALITY – IOT IS A METAPHOR
PRINCIPLES AND PRACTICE OF CONNECTIVITY AND CONVERGENCE

A Collection of Amorphous Ideas

by

Shoumen Palit Austin Datta
HAPHAZARD REALITY – IOT IS A METAPHOR

Shoumen Palit Austin Datta

© Copyright

This collection is available from the MIT Library and is free from any copyright protection. These ideas may be used for any purpose including for-profit operations.
TABLE OF CONTENTS

CHAPTER 1 • CONNECTING ATOMS TO BITS ................................................................. 3
CHAPTER 2 • CYBERSECURITY .................................................................................. 47
CHAPTER 3 • EMERGENCE OF DIGITAL TWINS ..................................................... 61

FOLLOWING ARE ATTACHED WITH THIS DOCUMENT (SEPARATE PDFs)

CHAPTER 4 • ARTIFICIAL INTELLIGENCE
CHAPTER 5 • DIGITAL ENTERPRISE X.0
CHAPTER 6 • HEALTHCARE – DIGITAL IN 4D
CHAPTER 7 • HEALTHCARE – MEDICAL IoT
CHAPTER 8 • HEALTHCARE – Book Chapter
HEALTHCARE from the MIT LIBRARY • https://dspace.mit.edu/handle/1721.1/107893

APPENDIX I • IPv6 in identification (preBlockchain)
APPENDIX II • Big Data (before “Big Data” era)
APPENDIX III • Adaptive Value Networks

COMMENTARY [A] • Social Disequilibrium
COMMENTARY [B] • Commencement Address
COMMENTARY [C] • World Sensor Organization

MORE INFO from the MIT LIBRARY • https://dspace.mit.edu/handle/1721.1/111021

Please find • ADDITIONAL MATERIAL inside “CHAPTERS” (zipped folder)
Please find • APPROX 6000 PPT SLIDES/CHARTS in “RE-VIEW-IOT” (zipped folder)
A FEW PAGES ARE LEFT BLANK TO FACILITATE TWO SIDED PRINTING
CHAPTER 1 • CONNECTING ATOMS TO BITS

WHY IT MAY NOT BE ENOUGH
IOT IS A DESIGN METAPHOR for DIGITAL TRANSFORMATION X.0
INTRODUCTORY ESSAY

CONNECTING ATOMS TO BITS

Shoumen Palit Austin Datta

Senior Member, MIT Auto-ID Labs, Research Affiliate, Dept of Mechanical Engineering
Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139

Senior Scientist, Medical Interoperability Program, MDPnP Laboratory
Massachusetts General Hospital, Harvard Medical School
Partners MGH Research, 65 Landsdowne St, Cambridge, MA 02139

Managing Director, NSF Center for Robots and Sensors for Human Well-Being
RoSeHuB, School of Engineering Technology, Purdue Polytechnic, Purdue University,
156 Knoy Hall, W Lafayette, IN 47907

shoumen@mit.edu • sdatta8@mgh.harvard.edu
ABSTRACT

Threads of amorphous discussions, historical references and atypical examples, the author admits, is unconventional. In one prior engagement, the author was repeatedly excoriated for garrulity. Some people do not like to read more than a page, managers seek executive summaries, executives limit reading to bullet points and VCs prefer elevator pitches, under thirty seconds. The trend in this article may disappoint those individuals.

Synthesis of information helps us to illustrate the evolution of thought. Digital transformation, is a pragmatic topic of interest to industry and business, but it is vast. The scope of digital penetration remains unfathomable. Innovation is quintessential in this context. Old ideas must be viewed by the youth with a set of different eyes. Hence, the “think different” adage rings true. Pursuit of ideas, must reflect on a large number of issues, combined and connected with diverse domains. Leaders must resist the “hype begets hype” trend and ignore slick marketing attempts to perpetuate hype cycles.

If this approach discombobulates your comfort zone, or if you prefer convention, then, you are at the mental stage possessed by billions. That stage will change, sooner or later. In some regions, we are almost at the end of the lag, and ready to begin a climb. The metaphorical climb is not magical. It is due to the evolution of IoT and digital transformation. On this journey, rational digital economics will serve as our compass, as we navigate through the sign posts of progress, and attempt to map the road, ahead.

Fig 0: Hockey stick model illustrating a preparatory phase followed by rapid growth.
PREFACE

The ideas relevant to the internet of things (IoT) has a rich history. Perhaps, it was seeded by the short stories authored by Isaac Asimov. In 1953, Asimov’s “Sally” was conceived as an autonomous vehicle⁴, cruising our streets in 2057. Herbert Simon⁵ in 1986 and Mark Weiser⁶ in 1991 are notable visionaries who contributed to the evolution of our thoughts, related to the metaphor of internet of things (IoT). In 1999, at the MIT Auto ID Center, Kevin Ashton, a marketing expert, loaned to MIT from Procter and Gamble (P&G), proposed the term internet of things⁷ or IoT. In the past few years, it has been suggested⁸ that the exact term (IoT) was first used by Peter Lewis⁹ in 1985, which was even before the seminal talk by Herbert Simon in 1986. Parallel observations and discoveries are nothing new, as acknowledged in a recent publication¹⁰ by Sanjay Sarma.

In this discussion, I have offered “my version” of a sense of the future, partially with respect to IoT. My purpose is not to be correct or right about these ideas, but to propose them, try to disseminate these ideas, allow experts to dig deeper and provide students with a few clues, about the principles and practice of connectivity and convergence. It is a sentiment borrowed¹¹ from Pierre-Gilles de Gennes (1932-2007).

This article is aimed at informing [a] experts, who may talk, generally, to other experts in their field, most of the time, [b] faculty, who are digging deeper and deeper, into fewer and fewer micro-domains, accumulating vast repertoire of knowledge and research publications, but may not have the time to explore the connected world, or the networked society, or the chasm precipitated by the socio-economic disequilibrium catalyzed by the quantum leaps of technology, [c] industry executives, forced to focus on financials, but eager to view the big picture, yet handicapped to provide leadership due to the alarming frequency with which the next quarter rapidly approaches, and P&L results which influences the ‘writing on the wall’ by analysts, peddling prosperity [d] heads of government agencies or aides to legislative bodies or think tank gurus or the heads of governments, in quest of ideas or vision for public goods, or tools for economic growth, [e] administrators of global organizations, institutions, non-governmental organizations and multi-national platforms (standards bodies, policy roundtables, economic forums), who may influence the direction of future growth, development of science as a service to society, the pursuit of human values, and the elusive quest for égalité.

The most important purpose of this article is to inform students, about known unknowns, clues to unknown unknowns, and how to challenge conventional wisdom, about IoT concepts, in order to develop a broad spectrum of IoT umbrella initiatives. But, in conclusion, I point out that digital transformation and IoT is only a part of the picture.

My purpose is not to provide answers, even if the content may appear, in parts, to offer suggestions/solutions. Ideas in the article, often disjointed, may trigger the reader to delve into her own thoughts on the topic. The content is supposed to act as “hooks” or unfinished structures, the reader may choose to extend, or continue to build, or demolish, the idea. The failure of my ideas may be considered a success if it triggers new ideas.

My emphasis on debate, dissent and discussion of ideas, is due to my belief that IoT, a digital-by-design metaphor, represents connectivity and convergence, partially, in a plethora ways, known and unknown, at this time. Scientists, engineers, professors and professionals, cannot provide all the answers, cannot imagine the future where students may travel, and cannot address tomorrow’s initiatives, armed only with yesterday’s tools.
Hence, my zeal to distribute these and other ‘different’ thoughts, not to be revered, but to serve as whipping blocks, to be proven incorrect, to be questioned. Perhaps, in an attempt to prove them wrong, or in questioning the suggestions, the next generation of leaders, and students, may find the triggers to ignite their imagination. They may invent and innovate solutions, with tangible benefits, which could lift many boats.

Commodities makes the world go around. Energy, water, food, health, education, sanitation, communication and civil safety, are the pillars of life on earth, as we know.

GDP, CMOS, IPv6, SVM, GARCH, RBC, ANN, are parts of the “Alphabet” soup which the primary care nurse recommends, in addition to an Apple a day, to keep the doctor away. This soup, and its many ingredients, may help to deliver the commodities.

IoT is a metaphor for solutions and applications, in the ‘service’ of these and other commodities, to the global population, for socio-economic progress of humanity. Multi-domain application of the IoT metaphor has the potential to catalyze social re-engineering. By lowering transaction costs, IoT lowers the barrier to entry. Hence, IoT may penetrate markets of billions, rather than a few nations, with the affluent millions. The tsunami of changes on the road ahead will not be fostered by didactic pedagogy in closed loop disciplines or any one nation or a few institutions or corporate oligopolies.

Mobility has, and will continue to redefine, security, energy, education, health, finance and emergency response systems. Conglomerates and products may be reformed by personalization, experiences and services. Vertical integration may regress in the face of horizontal platform aggregation, for consumers and industries. Consumption of things and objects may be substituted by data of things. Consumers may only pay for integrated information at the point of need or point of care or point of transaction. Subscriptions to services and premiums (insurance) may be transformed by pay-per-use micro-payments. The penchant for next quarter earnings may no longer be a key performance indicator.

Decisions and processes may be optimized by humans in the loop, using tools from AI, ML and DL, but humans may not be replaced by AI avatars, much to the chagrin of Hollywood. AI and automation will accelerate repetitive manufacturing, may aid humans in dangerous jobs (cleaning nuclear leaks, spills), improve precision and accuracy in certain domains, offer assistance to people immersed in monotonous tasks, provide occasional relief to workers for bio-breaks, assist the transportation industry and serve laughter, through comic relief, by winning against humans in GO or may jeopardize George Alexander "Alex" Trebek’s job as the host for Jeopardy. In the world at large, except the space station, automation may not serve us tea via robots, deliver palliative care to geriatric patients or care for pre-term infants in the neonatal intensive care unit.

The metaphor of IoT, and ideas in this article, may seem haphazard, yet real. This is borrowed from Haphazard Reality by Hendrik Brugt Gerhard Casimir (1909-2000), the famed Dutch physicist, known for the eponymous Casimir effect and the Casimir-operator in quantum mechanics. Part of the ideas are also borrowed, from the past, present and the future. It is presented here after processing through a different lens, often synthesized in the context of different thinking and hopefully, may be reviewed by different audiences. Part of the problem is my inclination to write in complete sentences and not in bullet points or executive summaries with a predicted “time stamp” of how long it will take an “average” reader to read the material (as often found “stamped” on articles published by the august Harvard Business Review and McKinsey Quarterly).
If *reading* was equivalent to comprehension, then we may not have witnessed the impact of invention of printing, for example, on poetry. When a poem can be read, and re-read, the emphasis shifts from the rhyme and the rhythm, to the allusion and the allegory. The latter is akin to a painting, it adds color to our *imagination*.

Part of the strength of this approach (you may choose to disagree) are its frequent, and haphazard, digressions, to connect the dots, from a myriad of sources, disciplines, and history. This is not a conventional mechanism to *teach* you any subject, but it may help you to *think, learn, connect* and, ask *questions*, hopefully, the *correct questions*. Asking the correct questions, remains the most difficult task, through the ages.

---

INTRODUCTION

The thorn in any transformation is the perspective of the leadership. Pre-set mental boundaries, in some humans, once imprinted, engages in a herculean attempt to gain comfort, and then, controls the metaphorical turf or territory. It clings desperately and tries very hard, to remain oblivious of change, it prefers to stay cloistered, remain inside the box and try to disregard advances, as fads, soon to evaporate from reality.

The ‘hockey stick’ illustration (Fig 0) and generic sigmoidal curves, exhibit an initial period of diminished activity, or lag, represented by a slow growth (flat) trajectory. This is contributed by, among other things, the resistance to change. The latter weighs down attempts to adopt change, even when necessary. It creates barriers to embrace the next new economy. Digital economics, at this time, perhaps, is the next “new” economy.

Figure 2 reflects a ‘hockey stick’ embedded in the sigmoidal trajectory, emerging from the work of Norman Poire, an economist and economic historian. Poire’s model uses observations attributed to Austrian economist Joseph Schumpeter, circa 1940, who noted a 50+ year cycle of ebb and flow, accompanying “new economies” leading to global economic growth and developments.

In the illustration below (Fig 2), I have included my projections (circa 2000), to highlight the progression from the physical world of atoms, to the transformation, in terms of the digital economy, where information about atoms may be represented by bits (Digital Twins) or bits, per se, may be information or data from decisions, processes and things, referring to the internet of things or IoT and industrial IoT. A few projections did not materialize but exceptions may prove the rule, albeit, partially. Other models by Poire (see Figure 24) suggests that the 2020’s may witness transformative changes.

Fig 2: Advances follow a pattern, punctuated by certain intervals, in models by Poire.
THE PROPHASE FOR DIGITAL TRANSFORMATION

Before eukaryotic cells can replicate (cell division in plants, animals, including humans), the cell undergoes a prolonged phase of internal changes. The cell takes time to prepare, prior to the commencement of active phases, preceding the actual division of the cell. This phase is referred to as the prophase in the cell’s cycle of replication or division.

As an analogy, the prophase for digital transformation may be divided into two parts, prophase 1 and prophase 2. Prophase 1 starts with hand stencils, which appears to be about 35,000 years old. The culmination of prophase 1 may be the Cuneiform script, circa 3500 BC, in Sumeria (Mesopotamia). It was probably the first time that thoughts or ideas, were transferred from the mind, via the stencils and then the scripts or pictographs, to tablets. Prophase 2 started in the Far East, with wood block printing circa 200 AD, but sprang to life with the discovery of the printing press by Johannes Gutenberg, circa 1440.

From the cuneiform script to the printing press spans about 5,000 years. This is the time it took to create tools for democratization of information. How many people read the Epic of Gilgamesh, which was written in cuneiform and carved on the Deluge tablet\(^\text{13}\) circa 2000 BC? How many read “If” by Joseph Rudyard Kipling? One had to be a Queen to have the resources, to read the Epic of Gilgamesh. The transaction cost, was too high. In the post-printing press society, “If” was accessible to people during Kipling’s lifetime (born on 30 December 1865 in Bombay and died on 18 January 1936, in London). In the 21st Century, ‘If’ may be downloaded for free. If the transaction cost is approaching zero, then please thank the digital economy, for lowering the barriers to market penetration.

Transaction cost\(^\text{14}\), pioneered as an economic staple by Ronald Coase (1937), is not the only lubricant for digital transformation. The human transformation to capture information and replicate information (stone tablets, papyrus, paper, books, bits) have had profound impact. The King as well as the citizen, the CEO as well as the clerk, the President as well as the janitor, can, now, access the same information. Ideas by Marshall McLuhan\(^\text{15}\) promotes that anyone, anywhere, may consume the same information. We are now able to democratize data due to advances in digitization and digital transformation.

Fig 3: The President and janitor may access the same information in a digital democracy.
The broad spectrum of digital transformation, which includes a future where Digital Twins\textsuperscript{16} may exist, may transform Marshall McLuhan’s suggestion to reality, for diverse verticals, ranging from manufacturing to robotics, energy systems to healthcare.

The preparation for digital transformation, the prophase, is nearing its concluding segment in OECD nations, but still struggling, in pockets, within ASEAN and USAN nations. The discrepancy between the economic zones is mostly due to structure of investments in infrastructure. Meaningful convergence of technologies is catalytic for digital transformation to grow sufficient roots, adequately fertilize markets in order to boost economies of scale, and improve profitable outcomes for economic development.

**TOOLS FOR DIGITAL TRANSFORMATION**

It bears reminding that the plural of anecdote is not evidence. This observation concerns Microsoft and the context is the infrastructure for digital transformation. In other words, the tools we need, must evolve and be available to the masses, in order to support digital transformation. Figure 4 shows the market cap of top 10 companies from 2001-2011. During this period\textsuperscript{17}, US companies faced a return to the economic phase of 17\textsuperscript{th} century England when NASDAQ lost 78\% of its value (5046.86 to 1114.11) between March 11, 2000 and October 9, 2002. The “dot com” bubble was attributed, by some observers, to laissez faire economics. In contrast, the collapse of the Lehman Brothers in the US, ignited a global recession. The collapse was manufactured by deliberately structuring financial instruments, to peddle greed, by a few financial institutions. The S&P 500 declined 57\% from its high in October 2007 of 1576 to its low in March 2009 of 676. Only Microsoft (and ExxonMobil) remained in the top 10, by market cap, during this decade (2001-2011) of “dot com” and “sub-prime” mediated financial turbulence.

<table>
<thead>
<tr>
<th>2001</th>
<th>Name</th>
<th>HQ</th>
<th>Industry</th>
<th>Market Cap USD million</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GE</td>
<td>US</td>
<td>Various</td>
<td>477,406</td>
</tr>
<tr>
<td>2</td>
<td>Cisco</td>
<td>US</td>
<td>Network</td>
<td>304,699</td>
</tr>
<tr>
<td>3</td>
<td>Exxon Mobil</td>
<td>US</td>
<td>Oil &amp; Gas</td>
<td>286,367</td>
</tr>
<tr>
<td>4</td>
<td>Pfizer</td>
<td>US</td>
<td>Pharma</td>
<td>263,996</td>
</tr>
<tr>
<td>5</td>
<td>Microsoft</td>
<td>US</td>
<td>Software</td>
<td>258,438</td>
</tr>
<tr>
<td>6</td>
<td>Wal-Mart</td>
<td>US</td>
<td>Retail</td>
<td>250,935</td>
</tr>
<tr>
<td>7</td>
<td>Citigroup</td>
<td>US</td>
<td>Banking</td>
<td>250,148</td>
</tr>
<tr>
<td>8</td>
<td>Vodafone</td>
<td>UK</td>
<td>Telco</td>
<td>237,175</td>
</tr>
<tr>
<td>9</td>
<td>Intel</td>
<td>US</td>
<td>Computer</td>
<td>227,048</td>
</tr>
<tr>
<td>10</td>
<td>Royal Dutch Shell</td>
<td>NL/UK</td>
<td>Oil &amp; Gas</td>
<td>205,340</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2008</th>
<th>Name</th>
<th>HQ</th>
<th>Industry</th>
<th>Market Cap USD million</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exxon Mobil</td>
<td>US</td>
<td>Oil and gas</td>
<td>403,366</td>
</tr>
<tr>
<td>2</td>
<td>PetroChina</td>
<td>China</td>
<td>Oil and gas</td>
<td>325,320</td>
</tr>
<tr>
<td>3</td>
<td>Apple Inc</td>
<td>US</td>
<td>Various</td>
<td>253,674</td>
</tr>
<tr>
<td>4</td>
<td>Microsoft</td>
<td>US</td>
<td>Software</td>
<td>243,687</td>
</tr>
<tr>
<td>5</td>
<td>Wal-Mart</td>
<td>US</td>
<td>Retail</td>
<td>295,905</td>
</tr>
<tr>
<td>6</td>
<td>B&amp;G</td>
<td>US</td>
<td>Retail</td>
<td>211,480</td>
</tr>
<tr>
<td>7</td>
<td>Industrial and Commercial Bank of China</td>
<td>China</td>
<td>Banking</td>
<td>208,397</td>
</tr>
<tr>
<td>8</td>
<td>Erichges Nethals</td>
<td>US</td>
<td>Insurance</td>
<td>202,901</td>
</tr>
<tr>
<td>9</td>
<td>China Mobile</td>
<td>China</td>
<td>Telecom</td>
<td>198,358</td>
</tr>
<tr>
<td>10</td>
<td>J &amp; J</td>
<td>US</td>
<td>Health care</td>
<td>183,602</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2011</th>
<th>Name</th>
<th>HQ</th>
<th>Industry</th>
<th>Market Cap USD million</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exxon Mobil</td>
<td>US</td>
<td>Oil and gas</td>
<td>417,566.7</td>
</tr>
<tr>
<td>2</td>
<td>PetroChina</td>
<td>China</td>
<td>Oil and gas</td>
<td>334,139.2</td>
</tr>
<tr>
<td>3</td>
<td>Apple Inc</td>
<td>US</td>
<td>Various</td>
<td>311,072.1</td>
</tr>
<tr>
<td>4</td>
<td>Industrial and Commercial Bank of China</td>
<td>China</td>
<td>Banking</td>
<td>251,078.1</td>
</tr>
<tr>
<td>5</td>
<td>Petrobras</td>
<td>China</td>
<td>Oil and gas</td>
<td>247,417.6</td>
</tr>
<tr>
<td>6</td>
<td>BP</td>
<td>UK</td>
<td>Oil and gas</td>
<td>247,079.5</td>
</tr>
<tr>
<td>7</td>
<td>China Construction Bank</td>
<td>China</td>
<td>Banking</td>
<td>231,606.6</td>
</tr>
<tr>
<td>8</td>
<td>Royal Dutch Shell</td>
<td>NL/UK</td>
<td>Oil &amp; Gas</td>
<td>224,128.7</td>
</tr>
<tr>
<td>9</td>
<td>Cheesecake Corporation</td>
<td>US</td>
<td>Oil and gas</td>
<td>215,790.6</td>
</tr>
<tr>
<td>10</td>
<td>Microsoft</td>
<td>US</td>
<td>Software</td>
<td>213,338.4</td>
</tr>
</tbody>
</table>

Fig 4: Microsoft survived the economic turbulence (Top 10 list by market capitalization)
In another analysis (Fig 5), which covers about a decade (from 2005 to 2015), economic volatility created a sinusoidal path, but failed to eject Microsoft (and ExxonMobil) from the top 10 list\(^\text{18}\) of globally dominant companies.

![Fig 5: Global tech dominance by US companies (illustration from Goldman Sachs)](image)

Can we make any sense about digital transformation from these illustrations? It may be a bit heretic to arrive at any conclusion, but perhaps there is a suggestion lurking between the lines. The only other US company that kept pace with Microsoft in these lists, is ExxonMobil. It is a commodity company. Will it be too much of a stretch to suggest that Microsoft, too, may be viewed as “digital” commodity company?

Technology companies are producing and servicing the digital commodities necessary to pursue digital transformation. Tech companies are purveyor of tools, to accelerate the adoption and diffusion of digital transformation. Tech companies servicing digital commodities may continue to amplify their profitability, seed global economic growth and lead the digital economy. If digital economy were to grow deeper roots, all forms of digital transformation must be pervasive and form a part of our daily lives. The latter is one definition of a commodity. If WiFi isn’t a commodity, then what is it?

Digital tools and their applications are commodities, for citizens in a connected world, the networked physical world of things or objects and the networked society. From that perspective, the portfolio representing the primary source of revenue, for five tech behemoths (Fig 6), suggests why Microsoft, as a purveyor of digital commodities, may weather storms, endure turbulence and survive economic volatility. Connectivity and commodities (products) are essential, hence, continued optimism for Apple and Amazon.
Products may be simplified into two categories – push and pull. Products in the pull category are “things” (products, services) which are consumed (B2B, B2C, C2C). Push is distributed, almost irrespective of demand. Microsoft (Fig 6) produces and serves the digital commodities sector, the “pull” category because consumers consume. Advertisement (ads), which are pushed, may have restricted viability. In other words, there is only so much that an advertisement can expect to achieve. Hence, the dependence on ads, sooner or later, may approach its growth limit. Disproportionate reliance on ads as a revenue stream, offers room for doubt about the viability of such a business model. Yet, in the short term, both tech giants (Fig 6) seem infallible, in their approach. Perhaps there is still enough advertising budget yet to move from print and other audio-visual media to the digital domain, which can tailor customer-specific ads.

The question of how much advertisements can achieve, in terms of sales, is worth exploring. Perhaps a model can be created (may exist) where a product or service (being advertised) may be assigned a penetration score based on [1] cost of product or service [2] purchasing power parity vs the “need” for the product/service [3] demographic spread of income vs disposable income or discretionary spending by households/individuals and [4] other factors, for example, brand recognition, social trajectory (status), philanthropy.

Using an “area under the ROC curve” tool, the model may suggest the boundaries of buying power. No matter how much you advertise, the outcome may be restricted within a buying power “area” under the ROC (Receiver Operating Characteristic) curve.

ROC analysis is part of "Signal Detection Theory" developed during World War II for analyses of radar images. Its current usage spans a vast spectrum of applications, including medical applications\textsuperscript{20} pioneered\textsuperscript{21} by Charles Metz\textsuperscript{22}.

---

\textsuperscript{19} Fig 6: Revenue streams of US technology behemoths and the “ads” percentage in each.

\textsuperscript{20}▪ Design Metaphor – Principles and Practice of Connecting Atoms to Bits • Shoumen Datta
The importance of data in decision making is not a new facet for the 21st century. Bargaining was the key *modus operandi* to negotiate dynamic pricing, in the bazaars of Mohenjo-Daro and Harappa, during the “modern part” (3500 BCE-1800 BCE) of the Indus Valley Civilization which dates back to 7000 BCE. Bargaining was the lifeblood of commerce for traders on the Silk Roads and Silk Routes (500 BCE to 500 CE).

From 2nd century BC to end of 14th century AD, a trade route originated from Chang'an (now Xian) in the east and ended at the Mediterranean in the west, beyond Venice. Hence, linking China with the Roman Empire. The Great Bazaar in Istanbul is rooted in the Silk Route. Kapalıçarşı or Büyük Çarşı, began as a small strong room, the Bedestan, built in 1461. It grew to cover 100 acres, and now has 18 gates and about 4000 shops, drawing about 91 million annual visitors. All kinds of bargaining behavior evolved; complex negotiations done simply by hand gestures, carried out beneath the cover of a shawl to prevent prying eyes from knowing what deals were being struck.

Bargaining, hence, was the first form of data arbitrage or information arbitrage. The importance of data is neither unique to digital transformation nor did it dawn upon us in the recent past. Data and decisions are inextricably linked and has always been, even before the evolution of human speech. Imagine humans using gestures to signal or warn that an animal is approaching.

It is ironic that pre-Neanderthalic gesture communication is being re-invented by BMW to enhance the ultimate driving experience. Gesture appears to be the new toy in the connected world but has the potential to serve humans with speech impediments.

Fig 7: Silk Roads and Silk Routes – trading valid for the period 500 BCE to 500 CE. In 1877 the term "Seidenstraße" (Die Seidenstrassen, literally "Silk Road") was coined by the German geographer, cartographer and explorer Ferdinand von Richthofen.

---

14 • Design Metaphor – Principles and Practice of Connecting Atoms to Bits • Shoumen Datta
Perhaps even more incredible, is the use by the United States, the ethos of the messengers on the Persian Royal Road. During the Achaemenid Empire (500-330 BCE), the Persian Royal Road ran from Susa (modern day Iran), to the Mediterranean Sea in Asia Minor (modern day Turkey). It served as one of the main arteries of the Silk Road, the network of trade routes, formally established during the Han Dynasty of China, which linked the regions of the ancient world, for commerce. The Persians maintained the Royal Road carefully and, in time, expanded it through smaller side roads. These paths eventually crossed down into India, across Mesopotamia, and over into Egypt.

The Persian Royal Road featured postal stations along the route, with fresh horses, for envoys to quickly deliver messages throughout the empire. Herodotus, writing of the speed and efficiency of the Persian messengers, stated:

“There is nothing in the world that travels faster than these Persian couriers. Neither snow, nor rain, nor heat, nor darkness of night prevents these couriers from completing their designated stages with utmost speed.”

These lines, from his Histories, 8.98, would, centuries later, form the creed of the United States of America’s postal service (USPS).

Data, and its time-sensitive arbitrage, as well as exchange and communication, are central to human evolution. The amount of data, its visual representation and extracting its contextual significance, has evolved over time. Digital transformation is yet another chapter in this evolution, as we sharpen our repertoire of analytical tools and continue to add new techniques for analysis of data, to make sense of its meaning and visualize data.

Fig 8: Visualization – (Left) Positions of Sun, Moon, Planets, from 950 AD, Europe. (Right) Positions of Earth and Venus (?) over 8 years (science). A fractal pattern?
In the evolution of internet of things, the principal emphasis is on data. In other words, we can think of IoT, to serve as a design metaphor for DoT or “data of things” implied in the principle of connecting bits to bits. That is, connecting not only data about things (bits about atoms) but also data about data or information store (bits about bits).

Because these data stores and things must be connected to be a part of a system or ecosystem, the network of the internet serves as the connectivity medium and gives rise to the recent concept of NoT or network of things advocated by NIST. NoT reinvents a concept which shares common grounds with Metcalfe’s Law.

Metcalfe’s law states the value of a telecommunications network is proportional to the square of the number of connected users of the system \(n^2\). Attributed to Robert Metcalfe in regard to Ethernet, Metcalfe’s law was originally presented, circa 1980, not in terms of users, but rather of "compatible communicating devices" (for example, fax machines, telephones, etc.). Later, with the globalization of the internet and the rapid diffusion of the concept of IoT, the idea was carried over to include, perhaps erroneously, users and networks. Its original intent was to describe Ethernet connectivity.

The significance of the “network effect” in these concepts relates to the value of information from one node versus a network of nodes. A multitude of nodes can act as a swarm. Data from such swarms may lend itself to analysis using the principles of swarm intelligence, for example, ant-based algorithms, used in network optimization routines and sensor networks in addition to emerging applications in digital twins (see Chapter 3 in _Haphazard Reality – IoT is a Metaphor_).

Fig 9: Two phones can make one connection, 5 can make 10 connections, and 12 can make 66 connections \([12(12-1)/2]\). Metcalfe’s Law is related to the fact that the number of unique connections in a network of a number of nodes \(n\) can be expressed as the triangular number \(n(n - 1)/2\), which is proportional to \(n^2\) asymptotically (ref 39). In other words, the apparent value of the network increases with the number of connections. Hence, Metcalfe’s Law, adapted to serve special interests, highlights the “value” of social networks. The latter has created grand illusions, catalyzes various forms of delusions and increasingly, we witness the transmutation of tabloid fodder, from speculation, to truth.
HYPE OF ARTIFICIAL INTELLIGENCE IN DIGITAL TRANSFORMATION

The present day unscrupulous marketing, about the power of, and benefits from, artificial intelligence, is reminiscent of the campaign of lies orchestrated by the US sugar industry. These “lies” stretch over the past 50 years and the most egregious act by the sugar industry was to bribe Harvard researchers to publish false results, to shield the ill effects of excessive use of sugar, exposed in a JAMA paper on September 12, 2016.

The exponential volume of deliberate false marketing and claims about AI by certain corporate behemoths, is increasingly amplified through ill-informed “me-too” social networkers, craving self-promotion.

This disingenuous AI barrage has all but drowned out the reasoned voices of credible scientists and genuine thought leaders, who are reluctant to be pro-active and abide by a code of decorum, prefers understatement, and are generally not narcissistic.

Artificial Intelligence (AI) and its lack of intelligence, is neither news nor new (see Chapter 4 in Haphazard Reality – IoT is a Metaphor). The term AI is a misnomer but has caught the public imagination. John McCarthy coined the term and used it in the August 31, 1955, proposal for the Dartmouth Conference to be held during the Summer of 1956. The usefulness of AI is a fact but not in the manner advertised, at present.

Making sense of data is central to profitability in the digital economy. Traders who were bargaining in the ancient bazaars had a limited set of data (for example, price of goods). The data set was sufficient for humans, with good memory, to develop astute bargaining skills, through data arbitrage, and excel in dynamic price optimization.

Even super-humans, with photographic memory, may not be able to navigate the pricing data from one single vendor, Amazon, which sells over 480 million products in the US. Amazon's product selection expanded by 235 million since Q1 of 2016. It has added, approximately half a million new products (485,000 SKUs), per day.

Dealing with such immense data volume is a major handicap for humans and the primary benefit from computation and data analytic tools, including AI type applications. Imagine the delivery planning optimization problem for goods delivery from Amazon.

Fig 10: Algorithms are key for transportation routes and sequence for delivery (ref 51)
Fig 11: Artificial Intelligence and Machine Learning are umbrella terms for many different tool sets and subsets. Knowledge of the problem space, context of the data, structure and quality of data, feature engineering and deep conceptual understanding of convergence, holds the key, in selecting the appropriate portfolio of techniques and algorithms, which, if combined, in proper sequence, may deliver an optimal outcome.
In the real world application space, selection of AI/ML tools must be context-dependent. In practice, the individuals who have industry knowledge, for example, in retail and fast moving consumer goods, may not be skilled as supply chain optimization analysts. The latter is classically the domain of specialists in operations research (OR). The retail industry may employ OR experts, to grasp the nuances of the retail business, in the context of the markets (understanding demand, demographics, distribution, etc).

In the realm of digital supply chain, the retail analyst is expected to use or apply AI/ML tools to optimize outcomes (for example, Fig 12). The confluence of knowledge, understanding of data, and selection of algorithms, must converge, for harmonization of the processes. Analysts must frame the correct questions, expect actionable information, and use it to improve decision systems, before the value of the data/information perishes.

Fig 12: Supply Chain Production, Planning and Optimization, using AI/ML tools. The coordination of these tasks within the enterprise, in the context of digital transformation, may be, still, largely in a “power-point” phase. Connecting data acquisition tools (RFID, sensors, market data) with the data curation domain, followed by the data analytics tools, in context of feature selection, with respect to the business/market logic, is the essence of the design at the heart of the internet of things concept. Connecting data about things, processes and decisions, is a part of the broad spectrum of the atoms to bits paradigm.

Transforming enterprises to digital enterprises is a herculean undertaking. It may take decades before actual transformation takes hold, survives the “pilot” phase and is accepted by business units, as a part of their operations. Human resource changes are necessary to transform ideas into reality, and, the seat of greatest uncertainty. Perspective of the leadership, and the ability of the executive management to lead these changes, are often in doubt. Even when the leaders seem to understand and champion the change, the pace of change is, at best, sluggish, or contained within a specific domain or business unit. Diffusion of change throughout the organization is difficult. The latter created buzz about “change management” and “open innovation” which are often lame and bit vague.
To add insult to injury, big consulting firms take advantage of these “gaps” and rake in billable hours. Publicists and marketing moguls, in pursuit of sound bites and media attention, create and disseminate, fantasies, which are empty, vapid and callous (Fig 13).

Fig 13: Mindless drivel created as AI and sourced by a “show-biz” organization as a window dressing, to peddle AI tools and enhance the “optics” of erudite imagery.

Fig 14: Model application using AI tools (SAP screen shot taken on June 1, 2000). Use of RFID, sensor data, and software agents, to reduce out of stock for inventory management.
Figure 14 illustrates the pragmatic benefits of digital transformation, one may expect, from reasonable use of tools (given their limitations), sourced from the portfolio of techniques, broadly referred to as AI/ML (artificial intelligence / machine learning). AI lacks intelligence, naturally. That is not the problem. The lack of the ability of humans, to understand why AI lacks intelligence, is the fuel which perpetuates the conundrum. But, enigma sells news, pseudo-sensational news accelerates publicity and the combination, spells profit. Hence, the penchant for industry to prolong the myth of intelligence in AI.

Figure 15 was created with great fanfare but failed to be functionalized due to business reasons. Open Source Interconnect (OSI) model shown in Figure 16 (bottom) is the structure of the infrastructure which makes it possible for us to access bits. But access is not the same as connectivity, in the context of process. Connecting business partners and domains (Figure 15) with business processes (Figure 16, IoT layers,) may be far more difficult. Access and analysis of bits, from intra-enterprise data silos, is impotent.

Failure to implement the AI application (outlined in Figure 14) or execute the P&G vision (Figure 15) is due to the immense procedural difficulties and intrinsic human resistance, to link process, and data flow, in context of the business relationships, which must converge, to generate the outcome (illustrated in Fig 16, top; seven layers of IoT). Connecting bits (Fig 16, bottom) to atoms (Fig 16, top) is a simple design metaphor, in principle, but difficult to execute in practice. We are still in the "power-point" phase.
Fig 16: Structure of connectivity (atoms to bits) - OSI\textsuperscript{55} vs IoT (top). Dynamics of cloud, fog and mist computing, and QoS\textsuperscript{56}, will be influenced\textsuperscript{57} by bandwidth, latency and jitter.

22 • Design Metaphor – Principles and Practice of Connecting Atoms to Bits • Shoumen Datta
Systems integration of AI tools at various data analytical nodes (core vs edge, cloud vs fog computing, real-time mist computing) within an ecosystem, is far from reality. Add-on or external data analytical engines are, logically, the first point of entry for new tools, including AI/ML. The analytical outcome is fed back into the system or the decision process. Processing high volume unstructured data is immersed in trial and error.

Application of AI/ML tools are a part of this “learning” process. The learning process itself is bereft with problems. The algorithms are plagued with misconceptions about how the human brain works, in particular, the neural networks (see Chapter 4 in *Haphazard Reality – IoT is a Metaphor*), which generates artificial neural networks.

An example is the backpropagation algorithm which is a staple in the AI world. In 1986, Geoffrey Hinton co-authored the seminal paper, which, three decades later, is central to the marketing explosion of AI. In YYZ, on Sep 13, 2017, Hinton said his method should be dispensed with and he is "deeply suspicious" of the back-propagation method, which underlies the hype marketed as AI, including the capacity to sort through photos and talk to Siri. "My view is throw it all away and start again," Hinton said.

![Backpropagation algorithm](Figure 11)

**Before Weight Adjustment**

**Parameters**

For \( w_2 = 5 \land x = 2 \land w_1 = 3.5 \)

Where \( MAE_1 = w_1x - y \land MAE_2 = w_2x - y \)

For \( w_2x = 10 \land w_1x = 7 \land y = 4 \)

\[
\begin{align*}
  f(x) &= \frac{MAE_1}{2} \\
  g(x) &= \frac{MAE_2}{2}
\end{align*}
\]

**Backpropagation of Error**

\[
g'(x) = g'(f(3)) \cdot 10
\]

\[
g'(x) = \frac{1}{2} \cdot 3.2 \cdot 10 = \frac{32}{2} = 16
\]

**Derivative of error**

Rubens Zimbres

Fig 17: Backpropagation algorithm (see Figure 11), in principle, is an error correction routine. It is a form of “supervised” learning to train artificial neural networks (ANN). After the training, the ANN should do what the trainer expects (the output), for example, recognize a hot dog, from a photo of a hot dog vendor, on the streets of New York City. If the output of ANN fails to recognize the “hot dog” (if the output is *baseball bat*) then the training algorithm, goes back and adjusts the “weights” associated with each step, in order to drive the neural net to output “hot dog” as the outcome of supervised learning.
Error correction is a staple and invaluable mathematical tool, in science and engineering. It is of critical importance in many fields. Without error correction codes, information theory\textsuperscript{64} may not function. Applications in finance\textsuperscript{65} and supply chain\textsuperscript{66} pre-dates the buzz of “big data” by at least a quarter century. Hence, error correction routine in backpropagation, in principle, is extremely valuable, and conceptually, is essential.

The discontent about backpropagation stems from the forced fitting of weights to deliver \textit{anticipated learning}. Adjusting weights is a blend of stochastic and heuristic methods. It is not reflective of biomimetic processes in higher organisms and unlikely to represent neural networks in humans. Thus, the call to \textit{“start again”} is both judicious and justified. The application of error correction, in the context of training neural networks, may be revisited or reviewed by “new” eyes, to deliver new tools, to complement or even replace backpropagation method. However, true AI/ML tools are \textit{essential} for our future.

Fig 18: Machine Learning (ML) may be one way to detect anomalies\textsuperscript{67} in large data sets. The importance of AI/ML in data analytics will depend on the question and its context.
Moving forward, beyond backpropagation, Geoffrey E Hinton of Google Brain has now, thankfully, re-invented a grand idea that was shining in relative obscurity, for more than half a century. Hinton’s "capsule" papers are “the emperor’s new clothes” made from Marvin Minsky’s "cube-on-cube" (Hinton’s “capsule”). Minsky’s idea dates back several decades but it was published in a book format in 1985. The illustration from "Society of Mind" is reproduced in Figure 19. In combination with Agent based hierarchical solutions (Agent, Agents, Agency, Agencies), the principles and practice of using ant-based algorithms in multi-agent systems have been around since 1960’s.

Fig 19: The network concept is obvious in this illustration by Marvin Minsky (ref 69). The abstraction is loosely based on synaptic connections made by neurons (in brains) and may have also inspired Charles Babbage, circa 1850, to lay the foundation of the design (but not build) the difference engine (ref 77). Each unit of data or information or parameter or variable or value (binary) may be captured in cubes in Marvin Minsky’s diagram. It may be extrapolated to the novel “capsule” concept, suddenly re-invented by Geoffrey E Hinton, of Google Brain, and presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017, Long Beach, California, USA). The power of the “cube-on-cube” concept is evident due to several “re-inventions” of this principle and certain forms of commercial products which may have borrowed its salient features (for example, the “containers” software). Cube-on-cube reappears in several chapters in my book Haphazard Reality – IoT is a Metaphor to illustrate various ideas, including the concept of hash tables and blockchain as well as the key essence of connectivity.
ELUSIVE QUEST FOR BIO-INSPIRATION IN DIGITAL TRANSFORMATION

From the dawn of computer design the management of architectural complexity (lower levels of logical circuits to higher levels, nested above one another) has drawn on the biology of neurons, in general, and neuronal scaling, in particular, as a bio-inspired model. The history of artificial intelligence research, and difference engines, proposed by Charles Babbage, appear to take cues from neural structures found in worms, animals, mammals and, of course, humans. The human brain is often compared to a computer. It is a gross error, and distortion, of astronomical proportions. The brain is not a computer.

In terms of an analogy, the roughest approximation, if one may dare to simplify, is the notion that the brain may partially perform functions resembling a very advanced systems integration platform. The brain is not a computer but computation may be a part of the platform analogy. The elusive quest for bio-inspired systems is justified because “nature” created structures and functions which successfully survived evolutionary tests.

Lessons from bio-mimicry suggests that we should invest heavily on sensors and sensing. Why? Because the epidermis (for example, human skin) is a whole body touch sensor (pressure). Other sensors are located, all around the skull (optical-eyes, auditory-ears, olfactory-nose, gustatory-tongue). Bio communication medium is electrochemical (analogy - telecommunication) and operates in a hub-spoke system. Brain is the hub and the spoke (motor neurons) are muscles. Signal processing happens in the brain, hence, the analogy, brain as a systems integration platform, with a “cloudy” component. Other signal processing (reflex arc) occurs through the spinal cord (analogy to fog or mist computing), for example, the patellar reflex (also, analogous to in-network processing).

Digital transformation is bursting at its seams with sensors, but making sense of data, to yield actionable information, is the quagmire of “haphazard reality” in IoT.

Fig 20: The myth of the human “large” brain. *Homo sapiens* are not special. Large brains appear several times in the mammalian radiation (Mya = Million years ago).
CRITICAL FOR DIGITAL TRANSFORMATION: CYBERSECURITY

Inviting digital annihilation is inevitable, without cybersecurity. If data and information exchange is desired beyond the closed “stand-alone” paradigm, then security cannot be “layered” as an after-thought.

Implementing security by design, through engineering principles in the backbone, is not only prudent, but may boost profit and brand recognition. The example from the auto industry, illustrated below, is of historic interest, with a message for practitioners, to think first about, and integrate cybersecurity, with digital transformation.

Fig 21: E J Claghorn patented a safety belt in 1885 but medical studies and work from Griswold and Volvo’s Bohlin were necessary for widespread adoption. Swedish automobile Saab GT750 (Gran Turismo 750 produced between 1958 and 1960) introduced at the New York International Auto Show, in 1958, featured safety belts, as a standard.
The dramatic cybersecurity story of 2016 came to a quiet conclusion on December 8, 2017, in an Anchorage (Alaska) courtroom, as three young American computer savants pleaded guilty to masterminding the unprecedented Mirai botnet, powered by unsecured internet-of-things devices like security cameras, wireless routers, that unleashed DDoS attack using an internet device on October 21, 2016. The Mirai "packet" strategy (PCAP) targeted IPv4 addresses and related DNS registries (Dyn).

Could IPv6 offer better protection for cybersecurity by engineering design? The economics of the resistance, to move from IPv4 to IPv6, is due to behemoths in the routing business, who have a lot invested in IPv4 and are simply guarding their profits.

DIGITAL TRANSFORMATION X.0
Do you have your compass? We are at the dawn of a new evolution. The road ahead is still under construction, plagued with ambiguity, potholes and gross missteps (Fig 22). Constructive convergence, and synthesis of ideas, may not be so easy.

Fig 22: Wearables are a new direction for the consumer internet of things (IoT). It is hyped up to measure vital signs and a tool for remote monitoring. The potential exists but the current practice is deeply flawed. In part, because software technicians are meddling in medicine. The heart monitor on the wrist band displays pulse rate of 40 BPM (beats per minute) but the medically verifiable heart rate recorded by Dr Wayne Whitwam, MD, is 147 BPM. This type of an error may spell the difference between life and death.
In *Darwin’s Middle Road* Stephen Jay Gould wrote, "if genius has any common denominator, I would propose breadth of interest and the ability to construct fruitful analogies between fields." This article hints at a lot of different domains and offers clues. I have alluded to diverse domains, which must converge, in order to transform the vision and weave it into the fabric of digital transformation.

The principle of connecting atoms to bits may be pervasive and it may gain ubiquity, in view of increasing computational power and decreasing cost of computation, as well as storage. However, the ideas relevant to internet of things (IoT), industrial IoT (IIoT), data of things (DoT) and network of things (NoT), are only a small part of the connected world. The forthcoming tsunami of changes will be far greater than digital transformation. We need changes to improve quality of life for billions of people.

For readers still willing to explore digital transformation, I may suggest my book “*Haphazard Reality - IoT is a Metaphor*” which is available from the MIT Library.

**TEMPORARY CONCLUSION: SIGN POSTS ON THE DIGITAL SILK ROAD**

Transformation is not limited to the digital economy. The scope of the economy itself is subject to change, based on other purveyors of civilization, for example, energy, food, water, health and diffusion of education. The cycle of economic evolution may span several generations and cover centuries, before these changes appear striking (Figure 23).

Prediction about the course of these changes are not easy but they are in abundant supply. Careful economic studies are few and far between. It is generally agreed that in the modern world (early 21st Century), we are equally, or more concerned, with access to sustainable and renewable energy, respectful of the ecological and environmental issues.

Hence, breakthrough in energy science and engineering, as well as, healthcare and sanitation, can change the future of nations, and as a consequence, the global economy.

![Fig 23: Changes in the share of GDP (gross domestic product) for the past 2,000 years.](image)
Science, engineering, and technological feasibility, will determine whether good ideas from energy research, may find their way to reality of execution, and adoption. To change the economic indicators of the world, the availability of such energy must reach the billions. Paths to adoption are going to be inextricably linked to investment. It may be co-dependent on other forms of infrastructure, for example, communications and roads.

Fig 24: Norman Poire’s (ref 10) sinusoidal “march of reason” based on work by several economists including Joseph Alois Schumpeter\(^\text{94}\), an Austrian-born political economist, at Harvard University (8 Feb 1883 – 8 Jan 1950) and the original Kondratieff\(^\text{95}\) Waves.

Based on the sinusoidal pattern (Figure 24), it appears that the next decade (circa 2026), may witness the emergence of new global direction, followed by transformative technological progress, poised to emerge at the dawn of the 22\(^{nd}\) Century.

One may be tempted to predict that the technological breakthrough (circa 2105) may involve energy. The sourcing of energy will revolutionize the global supply chain. Perhaps the development of hydrogen fuel cells, fashioned as an average sized mobile phone (form factor), or even better, a credit card sized fuel cell, in every pocket, may be the then Presidential\(^\text{96}\) slogan. The future may increasingly witness FCV (fuel cell vehicles) and a decline in EV (electric vehicles). The form factor of PEM fuel cells may be the size of a credit card or slim iPhone. Will one size fit all? Perhaps not. But, imagine you can insert the credit card (cc) configuration, proton-exchange membrane fuel cell, also known as polymer electrolyte membrane (PEM) fuel cell (PEMFC), in your laptop or smartphone or tablet, to power the devices. Now re-imagine the energy supply chain.

The same ccPEMFC may be inserted in the dashboard of your automobile, as its energy source. The “range anxiety” justifiably, will be of concern. How far the car may travel for “one dose” of ccPEMFC? But, before that, re-imagine the smart city concept.
Assume a vehicle can travel 10 miles per unit ccPEMFC. You are a taxi in New York where the average distance per trip is 2.6 miles. Have you lost or gained, in the context of the *massive shift* from lugging around 700 kg in battery weight (for EVs) or 15-30 gallons of gasoline (ICE) *versus* dash-board ccPEMFC weighing 10-100 grams?

Swap your energy-exhausted ccPEMFC with a fresh ccPEMFC and you have just entered the era of “swappable atoms” where the form factor of portable mobile energy enables you to **swap atoms**. Carry a 6-pack of ccPEMFC or swap ccPEMFC in a grocery store or convenience store, eg 711. The reality of the ‘swap’ may be in practice in India.

![Yellow taxis provide an average of 485,000 trips/day. The average trip distance is 2.6 miles.](image)

India’s electric vehicle revolution will begin with auto-rickshaws running on swappable batteries

![Fig 24: Complementarity between “atoms to bits” & the paradigm of “swappable atoms”](image)
Fuel cell "cards" (ccPEMFC) may fade away in favor of energy in the form of a microUSB flash drive. *Re-imagine* the mobile, grid-free, swappable ccPEMFC but in a new reincarnation: a slim flash-drive inserted into a microUSB-like port.

The energy source may be *solid* hydrogen. Contrary to the standard textbooks of chemistry, J. D. Bernal suggested\(^9\) that gaseous substances, under very high pressure, may be transformed into metallic or valence lattices. In 1935, Wigner and Huntington (see reference 99) theoretically calculated the energy of a body-centered lattice of hydrogen as a function of the lattice constant (which corresponds to a density many times higher than that of the ordinary, molecular lattice of solid hydrogen).

I did not read or understand Eugene Wigner’s paper until Glenn Seaborg\(^10\) (element Seaborgium\(^11\) atomic number 106) explained the prediction. The notion of solid hydrogen, in a thimble-sized container, was the basis of a suggestion, I proposed, in a publication\(^12\) in 2005-2006 (published by TEKES, Government of Finland).

At that time (2005), I did not mention solid hydrogen and did not state that one unit of the slim flash-drive-like microUSB device may be sufficient to operate a Boeing 787 or Airbus 380 or the largest container ship, the Hong Kong-registered CSCL Globe, measuring more than 400m (1,312ft) and carrying about 20,000 standard containers.

![Fig 25: (Left) Energy for a vehicle using a slim credit card sized form factor, proposed in 2005. (Right) Schematic representation of layered lattice of graphite. Metallic hydrogen may assume this type of lattice structure, according to Wigner & Huntington (see ref 99).](image-url)
In 2017, the impossible was made possible\textsuperscript{103} by Ranga Dias, a graduate student at Harvard University, who claimed the formation of metastable (solid) Hydrogen\textsuperscript{104} at a pressure of 495 GigaPascals (Fig 26). As with all seminal discoveries, congratulations\textsuperscript{105} and greetings poured in from all corners of the globe (Fig 27 quoted from reference 105).

![Fig 26: The formation of metallic Hydrogen by Harvard University student Ranga Dias.](image)

"From our point of view it's not convincing,"
Mikhail Eremets (MaxPlanck Institute)

"The word garbage cannot really describe it,"
Eugene Gregoryariz (CSEC)

![Fig 27: In Praise of Imperfection – congratulations poured in from other great scientists.](image)
Packed with metastable hydrogen (solid hydrogen), anything that needs energy ("electricity") may wish to incorporate a standard microUSB port in the design of the device. Insert pH (power Hydrogen) to a phone, laptop, automobile, airplane, building, lawn mower and submarine. Today we may plug in a EV in our homes to charge the vehicle. The fusion-less, grid-less, powerless, struggling economies may change for the better when they can plug-in their homes to the future FCV, to power up their living.

The form factor of portable, mobile, grid-free energy may undergo a radical evolution, which may reach its zenith around 2105-2120 (Figure 24). The metaphor of connecting atoms to bits (IoT) may embrace the complementarity which will evolve from swappable atoms. The paradigm shift in energy cost, access and distribution, will impact all supply chains. It may catalyze an unstoppable wave of freedom, development and economic growth. Power to the people will cease to be merely a political platitude.

A portfolio of inexpensive energy will swing open the final frontier for business and commerce, by further accelerating the creation of markets on the continent of Africa.

The benefits from the energy reshuffle will meet with resistance in the advanced Western nations, due to the dead weight of old technology, linked to the political lobby of the wealthy few. Asia and Africa may leapfrog into the future with solar, hydrogen and fusion, as a part of their renewable, carbon-neutral and sustainable, energy portfolio.

The strength of science behind the chemistry of metastable hydrogen makes it less speculative to suggest that the future of hydrogen-based energy, may be packaged in a micro-USB mini-device, with metallic hydrogen\textsuperscript{106} as the fuel. For immobile objects and traditional functions (buildings), the commercialization\textsuperscript{107} of fusion energy\textsuperscript{108} may be an equally important economic leap for the 22\textsuperscript{nd} Century (see Figure 2). Solar, hydrogen and fusion, may co-exist as pragmatic commercial solutions, globally, in the year 2121.

But what about the expected change in global direction (predicted in Fig 24), in the next decade? It is tempting to speculate that the global forces for good, may begin to tackle gender\textsuperscript{109} inequality. Boosting women’s opportunities, in the labor market, could add US$12 trillion to the annual global GDP, over the next decade. Gender parity will indeed be a new direction, a transformation which will change the fabric of civilization.

Fig 28: Ubiquitous computing meets ubiquitous energy in the 22\textsuperscript{nd} century.
EPILOGUE

Connecting atoms to bits is an important step for global development and decision systems. Digital transformation is one of many key areas that must march in lockstep, with other facets. The Digital Silk Road promises anything but to be smooth as silk. The forecast about energy, and supply chain, will add turbulence to anticipated changes. The journey will be mired in false promises, snake-oil sales and rabid sensationalism.

No one may escape the travails of time, especially those inflated with hubris. The cyclical rise and fall of General Electric is one example. On August 28, 2000, GE was at its prime with a stock price of $34.55 (Fig 29). More recently, on 20 December 2016, shares of GE were $32.25 per share. By 29 December 2017, GE's share price was nearly halved to $17.45 per share. In a recent analysis, from six CEO regimes at GE (Wilson, Cordiner, Borch, Jones, Welch and Immelt), the megalomania of Jack Welch may have disproportionately sowed the seeds of GE’s decline but Jeff Immelt is not without blame. Transforming GE to an energy company, or to a software company, are good slogans, for on-stage glitz and glamour, for “Minds and Machines” showmanship.

Fig 29: GE share prices from 1962-2017 (top) vs changes in GE leadership (bottom).
The economic disequilibrium\textsuperscript{118} due to technological chasm may be tearing apart\textsuperscript{119} the fabric of society, according to one\textsuperscript{120} report. The undue emphasis on digital transformation in this and other articles, is, therefore, an alarming example of bias. GE is an example of what happens when leaders lead by hubris or fall for hype\textsuperscript{121} and act as marketing agents rather than visionaries. Quick and dirty may win, but, for how long?

Based on my very limited knowledge and multiple other inadequacies, it may be an act of stupidity, on my part, to suggest what domains are equally (or more) important.

Feeding the world tops the list, not only because of its short term need (daily hunger) but its long term impact on the evolution of the brain\textsuperscript{122}. In parallel, reasonable elimination of sugar\textsuperscript{123} in our diet, and in processed foods, may be crucial for nutrition.

The bitter politics of sugar\textsuperscript{124} and its collusion with big pharma, may be sowing the seeds of an epidemic, called diabetes. Hence, to empower people, self-monitoring of blood glucose\textsuperscript{125} is quintessential, in order to prevent excess sugar in their daily diet.

Fig 30: Over the course of a 15-year study\textsuperscript{126} on added sugar and heart disease, participants who took in 25% or more of their daily calories as sugar were more than twice as likely to die from heart disease as those whose diets included less than 10% added sugar.
Inexpensive clean energy may be a key economic performance indicator, which will influence industry, commerce, workforce, financial arbitrage and globalization. The path to inexpensive energy may still run through political minefields (see Figure 1).

It is of general opinion that the trio of sanitation, water and health, as well as healthcare, are critical necessities for billions, mostly ignored by the affluent world. Emergency management, telecommunication and roads (infrastructure) are the bread and butter of the connected world. 5G is icing on the cake, not an immediate staple.

Advancing any domain for human progress is rooted in science and engineering. The tools for public goods, for example, nanotech and graphene for solar photovoltaic cells, as well as, desalination, and purification, of water, are fruits of academic research.

It is ironic that GE’s strength is in energy and healthcare. It is still taking the company on a downward trajectory. One reason may be its extreme affinity for the Western idea of business profitability. GE aims to serve its shareholders and maximize profit. If a business, such as GE, were to serve the community and, then, rake in the micro-payments from billions of users, it may generate trillions in revenue, over time.

The Western business school ethos of more and more profit is the yardstick of capitalism. In part, it fuels the lifestyle and luxuries we take for granted in the West. We are not without blame for the greed that may decimate families due to medical bills. Yet, we are too timid and cautious to raise our voices or act with purpose. The cost of prescription medicine in the US (Table 1), by law, cannot be subjected to price control.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glybera</td>
<td>$1,000,000</td>
</tr>
<tr>
<td>Luxturna</td>
<td>$850,000</td>
</tr>
<tr>
<td>Spinraza</td>
<td>$750,000</td>
</tr>
<tr>
<td>Soliris/Ionis</td>
<td>$750,000</td>
</tr>
<tr>
<td>Kymriah</td>
<td>$475,000</td>
</tr>
<tr>
<td>Elaprase</td>
<td>$375,000 pa</td>
</tr>
<tr>
<td>Naglazyme</td>
<td>$365,000 pa</td>
</tr>
<tr>
<td>Folotyn</td>
<td>$360,000 pa</td>
</tr>
<tr>
<td>Cinryze</td>
<td>$350,000 pa</td>
</tr>
<tr>
<td>ACTH</td>
<td>$300,000</td>
</tr>
<tr>
<td>Myozyme</td>
<td>$300,000</td>
</tr>
<tr>
<td>Arcalyst</td>
<td>$250,000 pa</td>
</tr>
<tr>
<td>Cerezyme</td>
<td>$200,000</td>
</tr>
<tr>
<td>Fabrazyme</td>
<td>$200,000 pa</td>
</tr>
<tr>
<td>Aldurazyme</td>
<td>$200,000</td>
</tr>
</tbody>
</table>

Table 1: The adulation for cure for blindness at a cost of $425,000 per eye per dose far outweighs the muted outcry about the outrageous acts by US and EU pharmaceuticals.

A pillar of Western business is optimizing “shareholder value” which may be archaic to some but still peddled by the Jack Welch school. It is clear from demand of “executive education” courses from elite US universities that the emergent economies are keen to emulate the Western practices, based more on pillars and less on building bridges.

The Chinese and Indian business schools are preaching personal wealth creation and how to catalyze ROI. But, what worked for the affluent Western society (US and EU) may not be reproducible in the East, with billions of hungry people who can neither afford healthcare nor education. The Eastern penchant for Western business practices is a form of myopia that neither Joshua Silver nor his mission nor his vision can cure.

Globally, there is a lack of understanding. Tools and technologies, by themselves, are impotent for sustained improvements, hence, are incapable of generating long term profitability and wealth of nations. When tools and technologies are a part of a system, in the systems approach to a solution, then, the outcome may favor long term profitability.
Use of *ad hoc* tools are often mired by initial failures. The use of RFID, as an identification tool for tracking and tracing goods in the supply chain, is still a grand idea. In 2000, retailers expected *that* productivity, instantly, simply by slapping on a RFID tag. It was catalyzed by hype and the stampede for a new source of *billable hours* from a variety of consultants and consulting firms, oblivious of the *systems* approach.

The lessons of *productivity* are steeped in *complementarity*. You cannot have one without the other. Complementarity is about *systems* where multiple tools and many technologies, both new and old, may *converge* to produce an impact, greater than the sum of the parts. Elegant illustrated by the history of electrification, similar lessons apply to most tools (RFID, sensors, AI, ML, blockchain, 3D printing, digital foundry).

Taken together, what does it mean for global transformation and its subset, the digital transformation? Unfortunately we shall witness catastrophic clashes, especially with respect to the ongoing parade of expectations from AI and its apparently supernatural powers. Even more troubling is the observation that once-admired companies, mammoth consulting firms and so-called “captains” of industries are marketing untruths.

Having said all of this, we must hasten to return to reality, to face the brutal facts in Figure 31, about the global gaps which continues the morbidity. It is not enough to blame the policies and politics of unethical humans in the loop, pre-occupied with treachery, lechery and debauchery, rather than helping to catalyze science to serve society, deliver human values and dignity.

Each one of us has a role to play, no matter how small. Our existence is based on compassion and knowledge. Compassion without knowledge is ineffective. Knowledge without compassion is inhuman. In our pursuit of ideas, we must strive to think about the correct questions. Wrong questions will generate wrong answers. Humanity needs compassionate dreamers, an ethical sense of higher purpose, and humility.

---

**Fig 31: Remediable injustice?** How lack of adequate sanitation may be robbing billions of children in the world, from a healthy childhood (ref 92). Can science help society?
Le vrai point d'honneur [d'un scientifique] n'est pas d'être toujours dans le vrai. Il est d'oser, de proposer des idées neuves, et ensuite de les vérifier. La Science des Rêves, Science et Vie Junior, 214, (18 May 2007), page 13


https://dspace.mit.edu/handle/1721.1/107989

http://www.investopedia.com/features/crashes/


http://www.visualcapitalist.com/chart-5-tech-giants-make-billions/


https://en.wikipedia.org/wiki/Harappa

https://en.wikipedia.org/wiki/Mehrgarh

https://www.ancient.eu/Silk_Road/

https://www.goodreads.com/book/show/251270.Stories_From_The_Silk_Road

https://en.wikipedia.org/wiki/Grand_Bazaar,_Istanbul

http://driving.ca/bmw/7-series/auto-news/news/how-it-works-bmw-gesture-control

http://www.myo.com/

http://www.silkroutes.net/orient/mapssilkroutestrade.htm

www.ancient.eu/Silk_Road/

https://en.wikipedia.org/wiki/United_States_Postal_Service_creed

http://data-art.net/resources/history_of_vis.php
104 http://science.sciencemag.org/content/early/2017/01/25/science.aal1579/tab-figures-data
105 https://www2.ph.ed.ac.uk/~gja/thermo/HydrogenGI.pdf
107 www.sciencealert.com/the-uk-has-just-switch-on-its-tokamak-nuclear-fusion-reactor
111 https://en.wikipedia.org/wiki/One_Belt_One_Road_Initiative
113 https://seekingalpha.com/article/4125284-ge-fall
117 http://www.macrotrends.net/stocks/charts/GE/prices/genl-electric-stock-price-history
118 https://journals.fe.up.pt/index.php/IJMAI/article/view/190/133
122 http://www.pnas.org/content/109/Supplement_1/10661.full.pdf
124 http://blogs.plos.org/publichealth/2013/03/28/the-not-so-sweet-politics-of-sugar-consumption/
126 https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/1819573
130 www.vanguardngr.com/2018/01/treatment-blindness-sell-850000-us/
131 https://www.health.harvard.edu/blog/the-11-most-expensive-medications-201202094228
133 http://www.qcatz.oxy.ac.uk/node/625
134 http://www.globaleyesightnow.org/self-adjusting-glasses/
Shoumen - I finally had some time to complete a thorough reading of your recent paper, "Adapting Decisions, Optimizing Facts and Predicting Figures." It was certainly thought-provoking. As you know, we have been thinking about some of these ideas, but it is nice to see them woven together more completely than I have before. I think the trick for industry will be to fearlessly use these ideas instead of rejecting them out of ignorance, cynicism, or short-sightedness.

JAW

17 June 2004
Jeffrey A. Wilke

Jeffrey A. Wilke
CEO
AMAZON

Jeff Wilke (JAW) has served as CEO Worldwide Consumer since April 2016. From February 2012 to April 2016, Jeff served as Senior Vice President, Consumer Business, from January 2007 until February 2012, he served as Senior Vice President, North American Retail, and from January 2002 until December 2006, he was Senior Vice President, Worldwide Operations. Jeff joined Amazon.com as Vice President and General Manager, Operations in September 1999. He left AlliedSignal (Honeywell) where he was Vice-President and General Manager, Pharmaceutical Fine Chemicals. Jeff spent the preceding six years in a variety of operations and general management assignments in the chemical, polymer, and electronics industries. Jeff did his graduate work (MBA and MS in Chemical Engineering) at MIT's Leaders for Global Operations (formerly Leaders for Manufacturing) program where he focused on Total Quality and Process Improvement techniques. He began his working career writing software and leading software development at Andersen Consulting (now Accenture). Jeff also holds a BSE degree in Chemical Engineering (summa cum laude) from Princeton University, NJ.
Re: Requesting harshest comments

Rodney Brooks [rbrooks@rethinkrobotics.com]

To: Shoumen Peiit Austin Datta

Saturday, August 20, 2016 10:42 PM

Hi Shoumen,

Not drivel at all. I generally agree with your sentiments.

You asked for comments. It is a little hard to read in places. I think I succeed because I know what you are trying to say -- I am pre-tuned to understanding your criticisms of AI. Others who feel that deep learning is indistinguishable from human (or super human) intelligence, whatever that can possibly mean, may have more trouble appreciating your points. I recommend working with a hard nosed editor who does not already know what you are trying to say so that the language and flow gets tuned up to be more accessible.

--Rod

Rodney Brooks | Chairman and CTO
ReThink Robotics, Inc


Rodney Brooks is the Panasonic Professor of Robotics (emeritus) at MIT.

He is a robotics entrepreneur and Founder, Chairman and CTO of Rethink Robotics (formerly Heartland Robotics). He is also a Founder, former Board Member (1990 - 2011) and former CTO (1990 - 2008) of iRobot Corp (Nasdaq: IRBT). Dr. Brooks is the former Director (1997 - 2007) of the MIT Artificial Intelligence Laboratory and then the MIT Computer Science & Artificial Intelligence Laboratory (CSAIL). He received degrees in pure mathematics from the Flinders University of South Australia and a Ph.D. in Computer Science from Stanford University in 1981. He held research positions at Carnegie Mellon University and MIT, and a faculty position at Stanford before joining the faculty of MIT in 1984. He has published many papers in computer vision, artificial intelligence, robotics, and artificial life.

Dr. Brooks served for many years as a member of the International Scientific Advisory Group (ISAG) of National Information and Communication Technology Australia (NICTA), and on the Global Innovation and Technology Advisory Council of John Deere & Co. He is currently an Economist at Xconomy and a regular contributor to the Edge. Since June 2014 he has been a member of the Visiting Committee on Advanced Technology VCAT at the National Institute of Standards and Technology NIST. Since June 2015 he has been an external member of GE's Robotics Advisory Council. Since January 2016 he has been Deputy Chairman of the Advisory Board of Toyota Research Institute.

Dr. Brooks is a Member of the National Academy of Engineering (NAE), a Founding Fellow of the Association for the Advancement of Artificial Intelligence (AAAI), a Fellow of the American Academy of Arts & Sciences (AAAS), a Fellow of the American Association for the Advancement of Science (the other AAAS), a Fellow of the Association for Computing Machinery (ACM), a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), a Corresponding Member of the Australian Academy of Science (AAS) and a Foreign Fellow of the Australian Academy of Technological Sciences and Engineering (ATSE).

He won the Computers and Thought Award at the 1991 IJCAI (International Joint Conference on Artificial Intelligence). In 2008 he won the IEEE Inaba Technical Award for Innovation Leading to Production. In 2014 he won the Robotics Industry Association's Engelberger Robotics Award for Leadership. And he won the 2015 IEEE Robotics and Automation Award. He was awarded honorary Doctor of Science degrees from Flinders University in 2016, and Worcester Polytechnic Institute in 2017.
CHAPTER 2 ▪ CYBERSECURITY

Personal Security Agents: Modular Models represent People, Process, Atoms & Bits

Abstract
The DDoS attack using an internet device on October 21, 2016, was predicted by experts and generated a flurry of suggestions from pundits. What if the attacks were not limited to social media but disrupted implanted devices in patients with cardiac arrhythmia? The potential of personal security agents (PSA) is proposed as a modular tool to model people, process, bits and atoms (objects) with layers to address trust, privacy, security and wrappers, within these layers, to include cognitive firewalls to mitigate other risks.

Fig 0 – Is cybersecurity essential for medical devices1 operated by wireless signals?

1 IoT is a Metaphor - Principles & Practice of Connectivity & Convergence ▪ Shoumen Datta
Distributed Trust Management

The distributed denial of service (DDoS) attack originating from IoT type devices infected by the Mirai botnet\[2\] was predicted\[3\] by experts. The news about tracking individuals by name and location\[4\] to profit from advertisements, was also perpetrated by experts. Objects and people appear to be equally vulnerable to cyber-intrusion.

The concept\[5\] of personal security agent (PSA) may not be new. PSA may not be restricted to individuals. Devices may have PSA linked to a registry, for example GHR\[6\] maintained by CNRI\[7\]. Other, less robust, examples include mobile device management tool and rosy predictions\[8\] about MDM, which are attracting a new\[9\] class of entrants.

In light of revelations\[10\], which continues to accumulate, it is worth considering security-as-a-service. Manufacturers of devices may provide a layer of security, but, other independent layer(s) may be procured, installed and managed by the user, based on source(s), trusted by the user. It may be analogous to accessing a secure box in a bank vault. It requires two keys to open, an example of a modular approach in an analog world.

Let us extend this over-simplification to a hotel room. The key card or mobile code, to open the door, may not allow the help (house-keeping) to enter the room, if the occupant dead-bolts or uses the security chain, from inside the room.

Re-imagine the scenario where neither the dead-bolt nor the security chain is affixed to the door surface, facing the inside of the room. On entering the hotel room, you find a bag of screws, a screw driver and four relevant holes on the door frame to insert and fix the dead-bolt or chain. You can use the widgets provided by the hotel or use your own widgets, as long as it is a standard that fits relevant “holes” in the hotel door frame.

In the digital world, there are no widgets. The digital equivalent (screws and holes) may be a dynamic API. The digital equivalent for dead-bolt is a digital certificate, with mobile duo authentication, provided by a third party trusted vendor (of your choice). Once you enter the room, the mobile app on your smartphone, which you used to enter the room code, prompts you to select a personal encrypted key, from a menu of providers. You choose a key, authenticate and transmit it, to the hotel door lock. Neither room service nor hotel manager can use any over-riding “master” code (a version of a public key) to unlock your hotel room door, unless you authenticate, and transmit your private key, to the electronic door lock. Dual keys will unlock the hotel door.

If you are in the shower, and the breakfast trolley is outside your door, you can use your smartphone to transmit your key to the lock, without getting out of the shower.

However, if you suffered a myocardial infarction and passed away in the shower, you need something more than breakfast. The room service assistant delivering the breakfast trolley gets concerned and calls the hotel security. Hotel manager calls 911. Upon arrival, emergency services uses a “blue button” emergency code, which over-rides private keys, to enable access to the hotel room, when there is a medical emergency.

For a more dramatic example, a jewel thief is hiding inside a room, after a heist. Tim McGee from NCIS can hack the digital key to unlock the door. But, if Zeevah is not quick enough to burst inside the room, the DID system (digital intruder detection system) is activated to sound an alarm. The jewel thief escapes through the window.

In another scenario, Mr Bond is inside a hotel villa. Dr No is trying to enter the villa. Dr No simultaneously hacks the digital key and silences the DID system. DID detects intruder activity but cannot trigger the sound software to elicit an audible signal.
Dr No hacked the DID alarm by using the 3rd party sound wave software (think 3rd party tire pressure sensor in a highly secure autonomous vehicle) necessary to create the audible response when triggered by the DID system. Dr No could not hack DID, but just turning off the audible alarm, accomplished the task, functionally, rendering the DID impotent by inactivating the sound.

However, despite the technical wizardry of Dr No, he is stumped to find, not one, but a triple layer cascade, engineered by the MI6 boffin Q for Mr Bond’s protection.

The MI6 version of the standard key offers a triple layer key. For the intruder, it will take three times the effort, to hack three keys, in order to complete unlocking. It is a waterfall cascade, requiring three jumps. For cybersecurity, it is a time trap. Hacking the first key gets you in, but you must hack the second key, and a third key, to complete the set. The time trap offers the bonafide user to take precaution. The tri-layered DID system triggers back-up signals for intruder detection, anticipating inactivation of the sound.

Dr No is a sitting duck in his DeLorean parked outside the entrance of the hotel villa occupied by Mr Bond and Miss Eva Gaelle Green. Alerted by the redundancy of the DID system, Vesper Lynd springs into action, from the adjoining villa. She perforates Dr No’s left vertebral artery with her Walther PPK, even before Dr No can utter, yes or no.

Modus Operandi

The drama aside, the suggestion involves the purchase of security-as-a-service (SECaaS) from one or more trusted vendor(s). This is, at this time, a hypothetical idea.

This is a business opportunity in cybersecurity. Distributed modularity for ad hoc cybersecurity via mobile delivery, installation, activation, intruder detection and third party response connectivity (for example, to police, emergency, FEMA or other humans).

A similar approach, but without mobile delivery, may be installed via a flash drive with microUSB, which may be inserted in the hotel door lock, to deliver the user preferred private key to the electronic door lock. In future, all door locks may have microUSB ports as well as compliance with WiFi, Bluetooth and other wireless protocol standards, to download, install and activate device drivers for security keys, DID systems.

Security as a service (SECaaS) on a microUSB flash drive may port security not only to hotel doors but any device, including millions using FPGAs. SECaaS drives may be bought from gas station (petrol pump) vending machine, at Walmart or Tesco or 711. It can be personalized by re-formatting the drive using a biometric tool (finger print or voice). The drive based system may be better protected from hacking, or time spoofing, compared to mobile delivery of the security app via wireless telecommunications.

Variations may create other forms, formats and ASICS. The essence is a SECaaS system which is modular, and distributed with such granularity, that the sheer volume may pose a hurdle for intruders, to identify which one to track and hack. The delivery of multi-component apps through different media (physical flash drive, mobile keypad) introduces complexity, a source of resistance. The number of providers decreases the potential for agency or industry sponsored espionage (by installing software ports with universal access or “back” doors to siphon and transmit data, unbeknownst to the user).

Albeit imperfect, mobile or USB SECaaS, may confer some degree of security to any system or device, which can be controlled by the user. Absolute reliance on external security systems may be plagued with trust issues, related to suppliers the provider may use, in order to deliver the service. User controlled SECaaS enhances the user experience.
SECaaS providers may include NGOs in the developing economies, government-academic-industry consortia, consumer watchdogs, global organizations or standards body (eg GS1, EAN, ITU[11], IEEE). These organizations may offer an independent service as a pay-per-use micro-payment. The assumption is that an august “external” agency, removed from direct corporate influence, is less likely to pilfer or compromise security. Hence, we may trust their security-as-a-service (SECaaS) offer.

Vendors of devices (shop floor machines, healthcare robots, heart monitors[12], airplane turbines, valves for oil pipelines, refrigerators, automobile parts, prosthetics, phones, hotels) will offer APIs to digitally receive, install and activate the security service (SECaaS). In future, devices which generate, transmit or acquire data, may not be sold without APIs. Interface standards will enable users to install one or more layers (think containerization of the sand-box concept) of security protocols, from one or more SECaaS providers, from different global regions, to create a secure cascade, as a system.

Mobility makes it imperative that security-as-a-service is user controlled and calls for new software tools. SECaaS may draw on vendors specializing in a smorgasbord of dynamic security engines, for example, random number generation tools, prime number cryptography, biometrics (finger print, iris scan) and composites (RFID EPC + SSN).

Partners in the SECaaS ecosystem requires secure storage (cloud, fog, mist) and connectivity with other domains (intruder detection, non-obvious relationship analysis, fraud monitoring). Data protection[13] rules and new policies[14] makes it essential for users to store and access their SECaaS data in their preferred nations. Redundancies introduced by the user may make it difficult to penetrate all layers of security, which did not originate from device vendors, or may be difficult to hack from distributed (cloud) storage, wrapped in cognitive firewalls, in separate locations or different geographies.

A Complex Path

Device-linkage is only one version of a personal security agent (PSA) for objects. But how an individual wishes to interact with the cyber-world, what data one wants to share, or which information one chooses to keep private, is inextricably linked to one’s identity. Protecting this identity, and keeping it secure, is another function for the PSA. Hence, it helps to differentiate between object-PSA (OPSA) and individual-PSA (IPSA).

Fig 1 – World Bank proposes to endow each individual with a digital identity. It may be a vehicle to embed individual personal security agents, for example, social security system.
Starting with medical records and continuing with the World Bank Group’s Identification for Development (ID4D) initiative, also advocated by GSMA, we need tools to deliver identity, security and privacy, for any citizen, with a digital footprint linked to their digital identity. The medium of delivery for IPSA may be linked to OPSA, because humans need a medium to interact with cyberspace. That medium is provided by the convergence of objects, hence, IoT, and telecommunications, via various applications.

There is little debate to refute the need for national policy and global tools. The latter must be redundant and distributed with very high fault tolerance. These tools must be capable of ad hoc dynamic composition when the status quo is challenged due to threats arising from breach of cybersecurity. Because post hoc security is useless in the IoT-by-design paradigm, the emphasis is on the dynamic composable architecture, and its ability to “discover” atoms and bits, in real-time, related to people, process and/or events.

IoT related cybersecurity is in need of IPv6, a key protocol for connectivity. It is a fundamental routing layer in packet communications. The rapid diffusion of connectivity catalyzed by IoT is increasingly and inextricably linked with our lives, perhaps, our digital twin representations, which may include information about information. In this scenario, information arbitrage by “avatars” may help decision support systems.

Individual Personal Security Agents (IPSA) may protect privacy, regulate data sharing and communicate (including emergencies) between individuals and their digital world(s), which may include digital twins, especially for industrial systems.

IPSA must be globally unique and coupled with individual id systems eg social security number (USA) or Aadhaar (UIDAI, India). We propose a digital twin approach (see Chapter 3), handled by creating a mobile software intelligent agent, for each citizen of the world (think of IPSA as an avatar from ‘Second Life’ games, 2003).

To tamper-proof the digital footprint and protect the digital records of this agent (IPSA), we may seek tools from “blockchain” (see Appendix I) to document, authenticate and grant permissions (‘handshake’), for interoperability and multi-tasking across diverse applications (eg healthcare, banking, fintech, e-vote and remote 3D printing-on-demand). Using public key cryptography and personal agents to protect private keys, PSA may be equipped with tools to de-identify data, in course of case-specific dynamic composition of responses. Hence, the potential to use IPSA in e-vote, or sharing de-identified private data, for example, healthcare data collection, for census or public policy surveys.

The agency of agents (APSA), acting on your behalf (IPSA), or on behalf of machine (OPSA) components or devices, associated with you, must be trained, updated and maintained, to be in tune with your personal likes/dislikes/preferences for allowing or not allowing your data (location, medical data) to be shared (or not), when external agents or bots query your digital ecosystem. What if a mapping service queries your phone for your location data? Who will protect you and offer privacy, if you so desire?

Your IPSA may be pre-programmed by you to respond, according to your preferences, which you may change using another remote device (OPSA), or modify associated dependencies, using digital assistants (time of day, office or home, travelling for business or in clandestine meetings, medical status or trigger emergency blue button over-ride). Perhaps similar approaches are necessary for Agents (OPSA) overseeing sensors, machine parts, medical devices, turbine blades, PLC controllers, smart grids, automobile brake pads, water filters and trillions of inanimate objects or things or processes using IoT, as a digital by design metaphor, in developing applications.
APSA, IPSA, OPSA and other PSAs, need standards. The road taken by trusted organizations may drive standards based operations, for security-as-a-service, to evolve. However, standards or policies for every possible situation cannot be conceived \textit{a priori}. Systems must be installed to trigger dynamic composition of \textit{ad hoc} micro-directives\cite{25}.

Open data sharing may be as essential as selective de-identification schemes, when anonymity and privacy must be balanced, in the interest of the individual. For example, Sam collapses on the steps of Vittoriano\cite{26} due to exhaustion and rushed to Ospedale Fatebenefratelli, nearby. But, they are unable to access her Epic-locked EHR from Massachusetts General Hospital. Sam is injected with a steroid and drifts to a coma. When staff in Rome speaks with Sam’s physician in Cambridge, they learn that Sam is diabetic. Injecting the steroid was a nail on her coffin, but, her death was preventable.

In certain scenarios, IPSA and OPSA, if adequately tuned, may be a life saver. In other cases, the responses handled by the agents may be denuded of certain data or values to protect personal ID (convergence of public key encryption\cite{27} with editable blockchain principles and IPv6 for mobile e-vote). Machines and devices (OPSA) may need data, information and metrics, to remain vigilant for intruders, as well as industrial espionage.

**Bio-inspired Modularity, Dynamic Connectivity, Composability and Convergence**

The community of PSAs must converge, connect, communicate and continuously monitor, to curate diverse digital threads, to analyze threats, estimate risk and synthesize the threshold necessary to elicit an alert, alarm or instantiate action, to preserve security.

The example of community as a \textit{function} based on components as a \textit{form} is a robust time-tested bio-inspired theme of modularity. One example is anchored in the principles surrounding the evolution of nano-biomes with respect to life forms in the oceans\cite{28}. Rather than combining all life functions into a single organism, the nano-biome works as a network of specialists, each with a special form (module) that can only exist as a community. The forms must converge in a complex spatio-temporal interplay, to give rise to systemic interactions, which, in turn, will manifest the desired function.

In the modular approach to cybersecurity, the form may be equivalent to agents, each specifically created to execute a particular task or role. When the agents aggregate, as an agency, to support a system, the overall outcome from convergence of these forms generates the function, in this case, an agency tasked to protect the security of the system.

How do we determine that the individual agents and their related tasks are secure?

This is where one begins to appreciate the value of modularity, convergence and the formation of agencies (groups of agents) which enhances the function (cybersecurity).

For example, you receive a SMS from your wife to warm up the lunch casserole for 5 minutes in the microwave. You proceed to perform this task and enjoy your lunch. What if you received a message from your wife to heat up the lunch for 500 minutes in the microwave? You (the human) wouldn’t comply with the command. Would you? Your sense of \textit{what is reasonable} prevents the execution of the message. This action represents the concept\cite{29} of a “\textit{cognitive}” firewall, designed to raise an alarm based on what is reasonable, as determined by our sense of cognition (Joshua E Siegel, PhD thesis, MIT; Sanjay Sarma, personal communication). If a security system is semantically challenged, it will not be able to distinguish between the \textit{values} in the field, 5 vs 500, because both are numbers, hence, allowable, unless there is a pre-set range, coded in the instruction.
Consider a command to a temperature sensor in a critical environment (cooling tower in a nuclear facility or combustion chamber in a jet engine or turbine). Usually, an external command may trigger the sensor to sense the temperature and report back to the data center every 5 minutes. An intruder-designed action or malware mimics the command but changes the time to 5 seconds. This action may appear benign but the battery life of sensor may be depleted in a few hours and sensor will cease to function.

If the temperature of the cooling tower or combustion chamber exceed the limits, it may result in a meltdown or some other form of catastrophe, including loss of lives, injuries and contamination, due to failure of cybersecurity. The sensor and the micro-system functioned exactly in the manner it was designed - sense temperature on demand. But, the micro-system was not designed to reason that there may be a breach of security because the task demanded – sense and respond every five seconds – did not include within its scope the validation process, whether or not a particular command “makes sense” for the whole system, taking into consideration the battery life of the hardware.

A cognitive firewall and its “supervisor” function, if installed, as separate yet linked modules, could evaluate the system’s “health” and test incoming commands, to ensure that a particular command “makes sense” for the connected system. Supervisor may proactively monitor the system’s pattern and identify anomalous behavior, due to system model breakdown, physical system failure, or other self-inflicted mechanisms.

Labelling this function as “cognitive” may invite justifiable criticism because the science of cognition is far more complex. Use of “cognition” is at best like using a new language without a dictionary or knowledge of the grammar. It may be analogous to learning a language like a child - through imitation and trend identification in complex examples of linguistic usage.

Similarly, claims of “intelligence” by unscrupulous corporations are vapid, trite and shoddy, if one considers the rigor of neural underpinnings of intelligence, compared with the pedestrian, apocalyptic, momentum attributed to artificial intelligence by those who are uninformed or ill-informed or inspired solely by profit.

Connectivity of these forms and agents, via IPv6, is one mechanism by which each entity level unit may communicate, when the systems are distributed, or where long range interactions are essential. The concept of cube-on-cube was proposed by Marvin Minsky, MIT (page 311 in this PDF). The convergence of Minsky’s cube-on-cube, with connectivity between the cubes, using IPv6 as the medium, is touted by companies, claiming “containerization” as a new and novel software innovation.

Connectivity using IPv6 draws on an earlier (2006) paper which is included in Appendix I. In that paper, I suggested using the internet protocol version six (IPv6) format to uniquely identify, not only things (objects in IoT), but processes, relationships (syntax, semantics) and interfaces (sensors). The design key in the earlier paper relied on using the 128-bit IPv6 global standard (which offers 3.4x10^38 unique “names or addresses”) to uniquely identify every instance of any transaction, any state change and follow their trail, even when distributed (routed). If re-imagined, the 2006 paper may appear to contain elements related to the concept of blockchain triggered by the bitcoin principle, which is related to the idea of digital ledgers. Implementing digital ledgers may monetize PSA but there aren’t any low hanging fruits, in this scheme.
Fig 2 – Marvin Minsky’s bio-inspired cube-on-cube concept [30] is extrapolated from, and may be a representation of, the topological connectivity between neurons or neuronal circuits, representing the form but not the function of biological neural networks.

Paths to Monetization

New business growth from security-as-a-service (cybersecurity software) may serve about 10 billion consumers by 2050 and trillions of B2B operations, much sooner. One must create/sell/lease/rent/train the Agents (IPSA, OPSA) for security-as-a-service. Monetization of software for security-as-a-service (SECaaS) calls for innovation, using the principle of “digital ledger” for innovative tracking of each instance of action, generally an event, and logging its associated changes (for example, change in state function) or relationships, for example, a log that reflects the pre- and post- states as well as the provider (vendor) associated with those states. The latter is essential for billing.

The management of micro-payments from a pay-per-use model, needs service requests, and service delivery documentation, as well as quality of service (QoS) compliance. The trusted vendor must be dissociated from the service delivery, once the service is activated, in order to reduce the risk, from point of entry, for the threat. The sales of the software license by the trusted vendor may not be the only transaction (in the old model a fixed payment was offered for a fixed term). If the trusted vendor wishes to charge by usage, then the ability of the trusted vendor to monitor the use of the service is pivotal. Dissociating the trusted vendor from keeping tabs on your security service is essential to improve security for the user. The latter introduces loss of opportunity for the trusted vendor, in this pay-per-use model, because the vendor is in the dark about how many times the user is accessing the security-as-a-service (SECaaS) application.
Consider the camera on the front door of your house. You are in Princeton, NJ and your smartphone app shows FedEx Fiona walking up your driveway to deliver a package. Fiona rings the bell. You open the door to your home in Cambridge, MA while visiting Tom (203 Lewis Thomas Lab) using remote key pad on your Iris app connected to the Schlage digital lock on your door. Fiona goes inside the house. She exits from the house and door locks behind her. You see on your app Fiona walking down the drive. What did FedEx Fiona do inside your home? She left the package on the table. What else did she do? Did she re-apply her lipstick? Did she use the toilet?

You wouldn’t know what happened inside your home unless you have a camera inside. Trusted vendor for security-as-a-service needs the equivalent of a “camera outside” to know when you ping SECaaS. It chooses to remain oblivious of your use (what you accomplish). The latter constitutes data from “camera inside” the house but trusted vendor is dissociated from that function. Trusted vendor does not need “inside” data because the trusted vendor charges micro-payments based on pay-per-use each time you ping the SECaaS app. It does not matter what you do but what matters is the duration of the use (unit rate or cost) and the time of the day (traffic volume, bandwidth, latency may be factors in the quality of service or priority queue determination). Thus, the timestamp is important for monetization. It is equivalent to house entry/exit data captured by the external camera. Privacy inside the house (data) remains unshared.

Distributed digital ledger, in practice, if combined with IPv6 transaction identity, can guarantee authenticity to bill for time-based micro-payment for service delivery (in any industry, eg, automobile, retail, healthcare). Each unit of this distributed digital ledger is in the form of an agent module. The nature of the service can “drag and drop” the “lego” “block” for the selected service units (modules) necessary to complete the function. The same modular principle, which applies to each distributed task sub-unit, is applicable in the distributed digital ledger paradigm.

Concurrent execution, co-location and semantic interoperability between standards and/or platforms, are key elements of this vision, if we wish to transform the suggestions to implementations. But, who will build these blocks? Part of the answer may germinate the next massive business opportunity, if distributed security is modularized to the lowest common denominator. To demolish the barrier to entry, a “pay-a-penny-per-use” SECaaS may enable any individual to “buy” security using a mobile app, from any virtual store, in a manner similar to buying a lock and a key, from a convenience store, to secure the suitcase or luggage, before handing it over to the airline, when travelling.

Fig 3 – The lego business of cybersecurity, convergence of system of ecosystems

55 • IoT is a Metaphor - Principles & Practice of Connectivity & Convergence • Shoumen Datta
The sign posts on the road ahead suggest confluence of telecommunications, mobile apps, data, analytics, AI, IoT, blockchain, edge services, with individual identity and identification. The system must map humans and machines, tools and widgets, components and systems, with digital precision, in the correct sequence, to track and trace, each instance. One aim is to uniquely identify that event or process or instance which provided the opportunity for the intruder to enter or introduce a threat, which may be propagated or percolate in the system, and eventually, cause direct or indirect harm.

This granularity is one reason why one may find the 128-bit format of IPv6 such an interesting scheme and use the IPv6 unique addresses, as a globally unique id (GUID). These concepts, digitally threaded with IPv6, may be a mechanism to improve cybersecurity, by engineering security, by design. There is ample room for convergence between IPv6 and blockchain, with selective use of public/private key encryption, for digital object architecture[^39] and IPSA.

How can we (can we?) use the 40 bits, or an extension of the security domain in the current IPv6 design, to serve as a cybersecurity base in the engineering design? Digital crypto-tokens, concealed in an alphanumeric stretch, may be connected with software security agents to authenticate (handshake) transactions, data transmissions or user activated action (the nature of which may be immensely diverse and vast in number).

The hypothetical concept of a set of cascading locks, is suggested. The header of the lock may be part of the 40-bit design of IPv6. Data related to the lock and its functional activation (I/O, open/close) may reside in a separate agent. It may mimic how RFID[^40] EPC[^41] contains a reference[^42] to the location where the actual data[^43] (or data agent) is stored, as a part of the broader context of the EPC information system[^44].

The locks may only open (allow, activate) with a digital key, or digital token, which must be generated in real-time (if triggered), using reference data (authentication?) secured by an agent in another location (potential for network verification at the edge). The “open lock” status in tier-x, could trigger the process to open the lock in tier-y (next lock in the cascade), using information (dependencies) from tier-x. This hypothetical cascade of locks, and the sequential effect (outcome), may offer the ability to trap an intruder, in time. The system may sacrifice a few locks but eventually the aberration due to the intrusion or anomaly (if detected through alerts or alarms) can turn off the cascade (remaining locks, sequences in queue) to prevent the remaining steps (remaining locks stay locked). This is the type of function one may also expect from the supervisory layer of the cognitive firewall, if a threat is detected or potential intruder is identified.

The hypothetical idea of cascading functions (with lagged dependencies) is my attempt, to theoretically propose a tool, or mechanism, to prevent or contain an attack (intruder detection, repulsion, protection). I do not possess the depth of knowledge to determine if cascading functions are at all possible, or can be implemented, in practice. Because the suggestion does not violate any natural laws of physics, it may not be impossible. On the other hand, this suggestion may trigger a better idea, in the reader.

If feasible, and deployed, this system of security may be useful for autonomous transportation (prevent vehicle from being hacked) or machines (prevent turning off the engine when a plane is in flight) or healthcare (prevent over-dose of morphine in post-operative surgical care) or energy grid brown-outs (time spoofing synchrophasors by creating anomalies in time-sensitive networking or inducing protocols to malfunction).
Concluding Commentary

Being Digital\cite{45} may not benefit from any hasty conclusions about people, bits and atoms. Controversies due to cybersecurity will continue to accumulate. Questions may soon outweigh answers, and good answers may be very few, and far between.

Is cybersecurity essential, at its maximum strength, for every instance, in any or all digital transactions? If it is, then, it may pose an insurmountable economic barrier.

In a common sense approach, cybersecurity needs to be enhanced for high value targets and the level of security may be increased or decreased, by choice. For example, the US President has more security personnel than the Governor of a State. The systems at the Pentagon are better protected than online systems for Ticketmaster. On the other hand, Ms Julia Roberts may be accompanied by a few security personnel when travelling to Boston but her entourage may include a few dozen security members if she arrives in Bangalore or Bahrain.

Transforming these simple scenarios to technical solutions, may not be difficult, at least in principle. For the President we could use HACMS\cite{46} and task Galois\cite{47} with its execution, under the watchful eyes of Kathleen Fisher\cite{48} and John Launchbury\cite{49}. For the Governor, we may use enterprise version of ESET\cite{50} while Ms Roberts may summon more ‘bodies’ if necessary. The latter may be analogous to the one layer of personal key for the hotel door versus the triple layer cascade of keys, configured by Q for Mr Bond.

The financial burden of cybersecurity and assurances for safety, may drain our resources, to the point where the actual task and function of the system may suffer, to provide for, and guarantee, security. Then there must be provision for litigation costs, if the security malfunctions or events are not covered by indemnity clauses. We may create an almost-impenetrable digital fortress, with nothing in it, of value, to be guarded.

In FY 2015, US Department of Defense spending was $598 billion, compared to US Department of Education annual budget of $68 billion and the US Department of Health and Human Services budget of $77 billion. For FY 2018, the US Senate approved $700 billion National Defense Authorization Act on 18th September 2017. The writing is on the wall.

Cybersecurity by engineering design, within the data transmission layers, may be one approach. The idea is to detect the intrusion and/or to catch the intruding agent, very early. The digital nature of the intrusion, and the digital access points the intruder must use, makes it reasonable to assume that the nefarious agent uses the same infrastructure, and same routing protocols, to reach its desired target(s). Thus, engineering security by design, using some facet of IPv6, could be helpful.

Distributing cybersecurity, in the modular approach, allows room for immense granularity. A cacophony of providers, vying for a piece of the business, must peddle its brand reputation and use different delivery mechanisms. The user will add extra layers of security, perhaps at the application layers. The strength of this approach is rooted in its diversity and may be boosted by its complexity, which must be sufficiently compromised, in order to create a breach.

Monetization of distributed cybersecurity using a pay-per-use micro-payment model may be lucrative. If 10% of the global population, uses one security function per day, and if you charge 1 penny (1 US cent) per use, then you earn $2.5 billion per year. With trillions of end points in IoT networks, billions of people in the world using millions of devices, is it possible that projected earnings of $2.5 billion per year is exaggerated?
PSA, IPSA, OPSA and distributed cybersecurity tools are not yet in our vernacular but may not be too far behind. Agent based cybersecurity is an old concept which must be re-viewed with new eyes. Dynamic composability is quintessential for cybersecurity due to volatility of environments. The ability to maintain a modest presence (for example, security guards at the entrance) but the capability to call in the cavalry, if an attack is imminent (for example, swarms of mobile agents), is the type of push-pull optimization that cybersecurity demands, to face future challenges. We need basic protection for a nominal SECaaS fee and pay-per-use on-demand security-on-steroids option, when faced with a threat, attack or intruder.

These concepts may not have a mainstream following, yet. These are not “low hanging fruits” to entice the corporate world. Harvesting low hanging fruits require low level skills. The pursuit of cybersecurity calls for brilliant, creative, and unusual, minds, and investments which can tune the engine, rather than merely polish the chrome.

Fig 4 - Cybersecurity may be analogous to the propellers of a ship. You may not see it, but the vessel cannot move forward, without it. Security is the not-so-secret sauce for diffusion, adoption, and profitable outcomes, we expect, from digital transformation.
27 https://dash.harvard.edu/bitstream/handle/1/12362600/Deterministic%20Public-Key.pdf?sequence=1
28 http://pubs.acs.org/doi/pdf/10.1021/acsnano.5b07826
29 https://dspace.mit.edu/handle/1721.1/104456
31 http://www.acad.bg/ebook/ml/Society%20of%20Mind.pdf
32 https://docs.docker.com/engine/userguide/networking/default_network/ipv6/
34 https://bitsonblocks.net/2015/09/09/a-gentle-introduction-to-blockchain-technology/
35 https://www.weforum.org/agenda/2016/06/blockchain-explained-simply/
36 http://www.mail-archive.com/cryptography@metzdowd.com/msg09959.html
37 https://bitcoin.org/bitcoin.pdf
44 https://doi.org/10.1016/j.compind.2017.10.004
45 http://web.stanford.edu/class/sts175/NewFiles/Negroponte%20Being%20Digital.pdf
47 https://galois.com/
48 https://www.cs.tufts.edu/~kfisher/Kathleen_Fisher/Home.html
49 https://galois.com/team/john-launchbury/
50 https://www.eset.com/us/
CHAPTER 3 • EMERGENCE OF DIGITAL TWINS

Abstract

Multiple forms of digital transformation are imminent. Digital Twins represent one concept, where we may use tools and technologies to “map” data (bits) from objects (atoms). It is gaining momentum because the “map” can act as a “compass” to reveal the status of atoms (things, devices, components, machines, people), process visibility and real-time transparency. Adoption of digital proxies, or digital duplicates, may face some hurdles due to lack of semantic interoperability between architectures, standards and ontologies. The technologies necessary for automated discovery are in short supply. Progress depends on the convergence of information technology, operational technology and protocol-agnostic telecommunications. Making sense of the data, ability to curate data, and perform data analytics, at the edge, or mist, rather than in the fog or cloud, is key to value. Delivering algorithm engines to the edge, are crucial for edge analytics, if latency is detrimental. The confluence of these, and other factors, may chart the future path for Digital Twins. The number of unknown unknowns, and the known unknowns, in this process, makes it imperative to create global infrastructure and organize groups, to pursue the development of fundamental building blocks. We need distinctly new ideas and unconventional research, to generate creative and innovative solutions.

1 Introduction

Being Digital (Nicholas Negroponte, 1996)1 and When Things Start To Think (Neil Gershenfeld, 2000)2 introduced the public to the emergence and the rise of smart machines. About a decade later, Jeff Immelt of GE started to market these ideas in the minds and machines campaign, claiming a future where self-organizing systems, subsystems, multi-component subunits and modular embedded code, shall define the next generation of adaptive intelligent machines. When can we expect that “next” generation to emerge?

In 1513, the discovery of the isthmus at Panama by Vasco Núñez de Balboa triggered the idea of creating a trans-oceanic canal. Francisco Lopez de Gomara suggested (in his book, in 1552) Panama, Nicaragua, Darien and Tehuantepec as choices for a canal. Not for another 300 years, not until the 19th century, would the canal building actually commence under the leadership of Ferdinand Marie Vicomte de Lesseps, a French diplomat. Ferdinand de Lesseps (19 Nov 1805 to 7 Dec 1894) could not complete the Panama Canal and did not live to see the successful completion of the Panama Canal, in 1914, by the US Army Corps of Engineers3.

Creating intelligent adaptive machines faces a similar uphill battle. Our optimism is not unfounded but it may be burdened by the dead weight of old technology. Paving the path for new theories, new concepts and new forms of connectivity, in engineering design, of future systems, may lead to intelligent (?) machines. Perhaps, a greater challenge may be introducing cognition in systems, due to our wobbly and poor understanding of what constitutes intelligence4 (AI).

Companies, in quest of low hanging fruits are afraid to delve deeper. They are retrofitting existing machines by slapping on sensors to designate them as connected. Workflow on steroids is peddled as
intelligence in analytics. Others are collecting and feeding big (noisy) data sets to existing software systems and claiming “cognitive software” in use. The tapestry of buzz words and patch-work of programs are introducing glaring gaps, generating errors, callous disregard for physical safety, and inept approach to cybersecurity in general.

One reason for the confusion, perhaps, is our general inability to ask correct questions. These are some of the questions from the field. What machines, devices and systems may be built with the tools and technologies at hand? How should we build and use them? Do we really want to just connect everything to collect big volume of data? Is it really all about data? How can we teach machines to achieve specific goals? Are these the correct questions to ask? Are these the questions even worth answering?

These and other related questions may find some answers hidden in bio-inspired design principles. Progress in bio-MEMS, bio-NEMS and molecular machines are elements which may converge with AI in an over-arching strategic plan. Integrating that plan to inform engineering design is the Holy Grail. The command, control, and coordination of bio-inspired engineering design requires hardware-software synchronization, by design (not as an after-thought).

Time-synchronized hardware-software integration is one hallmark of cyber-physical systems (CPS) which is the foundation of embedded systems. The concept of digital twins emerged from NASA, perhaps because time guarantee (concurrence) in embedded systems is critical for aero/astronautics. To advance the diffusion of the digital twin concept, we must adopt practices borrowed from CPS.

The current advocacy to advance the principles and practice of digital twins, from manufacturing to healthcare, calls for connectivity, by design. Systems should be able to discover, inherit, evaluate and share intelligence across sub-components, and coordinate to turn on/off modular code, embedded in sub-systems. We may monitor, analyze, control units (PLC) and sub-units, in real-time (sensors, actuation) and visualize operations not only at the system level, but, perhaps, the entire ecosystem. The latter may accelerate the diffusion of digital transformation, using a systems engineering approach.

2 The Rationale for Digital Twins

2.1 Signal vs Noise – IoT vs Digital Transformation

The term IoT may have been coined by Peter Lewis (1985) or at the MIT Auto ID Center (1999), but the past, present and future concepts of IoT have been brewing for almost a century. Milestones include Isaac Asimov’s “Sally” the fictional autonomous car, Herbert Simon’s seminal paper (“talk to the computer”), Hiroshi Ishii’s idea of “Tangible Bits” (People, Bits and Atoms), Mark Weiser’s paper “Activating Everyday Objects” as well as the 1991 article (Scientific American) and the vision of the “networked physical world” by Sanjay Sarma, David Brock and Kevin Ashton (2001). The IoT roadmap promises to be even more dynamic in the future and scholarly discussions, including one by Alain Louchez outlines the layers of influence.

IoT is a digital-by-design metaphor and a paradigm for ubiquitous connectivity. The value proposition rests on proper use of plethora of tools and technologies that must converge to make sense of the data. The hypothetical transparency is of little use without
the data of things, if we wish to profit from IoT applications. On the other hand, digital transformation is a cacophony of ideas open to innovation from wireless systems as well as broadband communication and forthcoming 5G, for time critical operations. The latter, if combined with 8K visualization, may catalyze robotic surgery. Masses may benefit from standard surgical procedures, such as laparoscopic cholecystectomy, appendicitis and phacoemulsification (removal of cataracts).

CNC machines, ERP, Web 2.0, fixed-task robots are examples of waves of digital transformation in business. The 2012 proposal from Sanjay Sarma of MIT Auto ID Labs to pursue a Cloud of Things initiative resonated globally and the concept was promoted by others (Finland, France and South Korea, to name a few). The next wave appears to be the transition from manufacturing products (items to be sold), to the creation of a service ecosystem, around the product, to sell service as a pay-per-use model. Digital transformation includes establishing a digital leash to monitor, promote, connect, track and trace, in order to monetize every point of contact in the relationship (digital CRM), not only once (sales of product), but over the life time of the customer. Hence, product lifecycle management evolves to digital PLM with quality of service (not product delivery) as the KPI, and monetization tool. Quality of service (QoS) emerges as the readiness metric to gauge customer satisfaction. If QoS metrics are maintained by provider(s) or manufacturer(s), then the client or customer is expected to pay for the QoS level associated with the product-service, per contractual agreement, as long as the customer is consuming the benefit, that is, the lifecycle of the user, not only the product.

In instances where the product is not an object (eg, telco) the business models are inextricably linked to “outcomes” the customer expects. Monetization of digital transformation from an outcome-based model is complex, due to the ecosystem of players and alliances. It is not easy to optimize and arrive at the point of convergence, to deliver the outcome as a seamless function, which involves end-to-end value chains, operating as a pre-agreed platform.

The task associated with monitoring each instance of engagement for micro-revenue collection, and its disbursement, is complex. We need to track each instance, and maintain a record of connectivity, in an irrefutable evidence log (eg blockchain). The latter may act as a digital ledger to validate fractional micro-payments, due from each point of contact (PoC). The digital id of the service, delivered at the PoC, identifies the member of the supply chain providing the unit of service, at that specific instance. The latter may be a part of the sum of services, in the portfolio, that defines the service, and each may be weighted by a specific QoS, the customer expects. The customer pays for the final outcome (value in the value chain). The sum of the parts must be delivered before the value perishes. The duration of that value may be widely divergent (compare retail vegetables in a grocery store versus life of the data which predicts risk of diabetes).

Synthesis of the parts to act as a seamless function is the challenge. Who will build the parts of the platform, which will be sufficiently open, and interoperable, to connect with the innumerable end points, at the edge? Who will build the blocks for the digital modules? Who will build the blocks for the blockchains? The business is building the blocks is not for the faint-hearted.
2.2 Digital Twins

Scenario

Schlumberger is monitoring a drill-head in operation, on a drilling platform, in the Outer Hebrides to determine the MTBF (mean time between failure) metric, and trigger replacement (maintenance), to prevent work stoppage on the rig. The camera at the tip of the drill-head, and drill-head (drill-case) sensors (e.g., vibration, temperature, gyroscope, accelerometer) transmits (wired, wireless) video, audio and other data, which must be analyzed as close to real-time as possible, with respect to object identification, precision geolocation and process linkage. AI/ML analytics updates MTBF metrics. Depending on MTBF range (80%, 90%) as decided by business logic (when to replace), the drill-head spare parts supply chain (service, fulfillment) must be connected to auto-trigger the “head” when the MTBF range is reached. Purchase orders [supplier(s)] are followed by transport and logistics for delivery, and workforce scheduling, to execute the replacement prior to breakage (payment, contracts, invoices, and accounts payable, must be connected). Data about drill-head, and lag time for each process/operation is captured by the operations management team, at a remote location, for future studies or collective evaluations. Can we visualize this entire end-to-end process as a Digital Twin, operating in real time?

In our current modus operandi this operation involves a plethora of operational silos (OT, drilling operation, mechanical engineering, systems, supply chain, finance, human resources), software (connectivity between locations, cloud store, cybersecurity) and hardware (not only the spare parts and drill-head but also the computational hardware, e.g., servers at different locations which are essential for the IT infrastructure).

The concept of DIGITAL TWIN posits that the flow of data, process and decision (this hypothetical scenario) is captured in a software avatar that mimics the operation or offers, at the least, a digital proxy.

The 3D “twin” or its “lite” version, the digital proxy, may be visualized by an analyst or manager, on a location-agnostic mobile device (iPhone, iSkin). Drilling down on a schematic illustration with the icon “drill-head” may link to the live video-feed from the drill-head camera, which opens up on a new GUI (tab or window). Data fields (features, attributes, characteristics) related to the drill in operation (pressure, torque, depth, temperature, rotations) are visible by clicking on the icon for the drill. A plot showing the data, approaching the MTBF metric, may be instantiated using a command (“plot data”). It shows the live sensor data feeding the dynamic plot, displaying the characteristics of the drill-head, and the rate at which it is approaching the MTBF. The range may be set by humans (or the system), using prescriptive and/or predictive values, based on “learned” values from machine learning tools operating in a “diagnostic” mode.

Will it allow for “what if” analysis?

Can the analyst viewing the Digital Twin, change the MTBF range and explore how downstream processes may change (see principles of http://senseable.mit.edu)?
The digital proxy for supply chain should spring into action, displaying delivery lag times from different suppliers, and cost of normal vs expedited delivery. The material composition of the alloy, used in manufacturing the drill-head, should be visible. The analyst may use an *ad hoc* selection process and identify a new vendor. Can the system trigger process workflow to alert the people (roles) along the way to clear the requisition, and generate purchase order, for the new supplier? Can it auto-verify the new supplier to check credentials, inventory, cost, transportation scheduling, quality of service reports and customer reviews from prior contracts?

We are viewing Digital Twin or its digital proxy on a device. We watch the drill-head in action. A window displays real-time data/analytics approaching MTBF. Using a different app, we identify a supplier to custom-design and 3D print-on-demand a drill-head with precision fit. The supplier ([www.quickparts.3dsystems.com/social-solutions/](http://www.quickparts.3dsystems.com/social-solutions/)) downloads the video feed (from the cloud) of the drill-head operating in Outer Hebrides. The manager monitoring the end-to-end chain [a] selects the team of engineers who will replace the 3D printed drill using a HR menu which lists skill sets, proficiencies and years of expertise by category [b] pre-sets the command on the digital twin to actuate the replacement supply chain process, when MTBF reaches 72% because fulfillment takes 21 days, and by then (21 days later) the MTBF is predicted to reach 85% (code red: replace).

Each sub-unit provider must collaborate and synchronize (systems, standards, semantic interoperability) their role in the operation, and the *representation* of their function, in the digital proxy, or in the digital twin model, in near real-time. The plethora of system providers, suppliers, 3rd party software, analytics, cloud storage, fog providers and hardware component manufacturers - *also* - expects to be paid for the outcome.

The design of content, and connectivity between such system of systems, calls for new principles of model-based systems engineering, to integrate global standards, to anchor architectures responsible to drive the digital-by-design paradigm. New models must inculcate the IoT digital-by-design metaphor with respect to connectivity by design, interoperability between standards by design, and cybersecurity by design. The silos of OT, IT and telecommunications must converge, to create this new digital-by-design foundation for digital proxies and/or digital twins (digital proxy plus 3D models).

In this paradigm shift, objects and things may not be baptized after birth to follow the digital persuasion, but will be born free of analog baggage. It may not need a path to digital transformation because they will be *born digital*.

**Configuring Digital Twins: Creating the Blocks in the Blockchain**

Lessons from cyber-physical systems (CPS) with respect to operational time synchronization may be key for certain forms of architecture for digital twins. Without open repositories, the process of creating (building) digital twins and the adoption of digital twins may be restricted to an industrial oligopoly. The vast majority of users cannot deploy an army of engineers to create custom digital twins. Rapid diffusion of digital twins calls for open source entity level models of sub-components (units). Think of each SKU listed in a BOM (bill of materials). Next, imagine each sub-system with unit parts, to serve as the “block” or base level unit, which needs to be created (built). The “old world” notion may have stopped at the physical manifestation - the actual unit made of tangible material. In a digital twin world, we need a digital version.
In the era of digital twins, we will call on the source, that is, the CAD/CAM model owner of that unit, to create, and contribute to a common repository (?), the software representation of the unit, replete with the physics of the material, and the engineering characteristics of its operational function. For example, the physics of the part will inherit natural laws, which governs all entities. For example, if a spare part were to fall off a table (on this planet), it will fall down at a rate defined by the acceleration due to gravity of 9.8 m/s². The latter is an inherited attribute from the laws of physics (characteristics forming the base in a ‘layer cake’ model).

To the informed mind, it is clear, we have encountered and entered the domain of semantics and ontologies. Digital twins of the granular units (parts), to be useful, must be connected by their relationships to the relevant data feeds from sensors/gateways. These entity relationship models and parts (connectivity) must be accessible to managers or analysts who can drag and drop the parts from the repositories on a “sense” table (device GUI). The ontology of entity level relationship models for digital twins may use bio-inspired principles. Elements from disease models, for example, a bio-surveillance model is shown below in Figure 1.

To the untrained eye the cartoons may not suggest the cryptic complexity of the foundation. These foundations are “layer cakes” (eg, TCP/IP) which must be able to communicate with other “layers” (for example, semantic web) built on other principles, concepts or ontological frameworks. It is imperative we minimize the number of such architectures, to ease interoperability between architectures (requires interoperability between standards, access to open data dictionaries, and shared ontologies).

The abstraction of the building blocks necessary for digital twins, may be similar, in principle, to the building blocks necessary to implement the use of blockchains, as a trusted digital ledger of connected instances. Who can we trust to build the blocks? This question is critical to practitioners of digital twins and blockchains. These concepts share with IoT the digital-by-design metaphor. The “block” in IoT may be the integrated platform, synthesized from subunits or blocks, containing data, of things.

The rate limiting step, which defines the functionality, is inextricably linked with, and driven by, connectivity. In order to deliver value, connectivity must span a broad spectrum of dynamic ecosystems. Implementation of such connectivity must be protocol-agnostic, location agnostic and time sensitive (maximize transmission, minimize steps), with respect to the “sense and response” latency, between the edge and the core. Since IoT is expected to connect trillions of things, scalability will be a key enabler.
Have we encountered such “block” and “connectivity” concepts elsewhere? The common answer is Agent systems. Marvin Minsky’s brain connections related abstraction “cube-on-cube” illustrates this concept where each cube is a software Agent. It is relevant to this topic because each cube may be viewed as a “block” in the blockchain or a baseline ‘unit’ in the digital twin paradigm (digital proxy of physical entity). The origin of the concept from software Agents, emphasizes the link to semantics, ontology, and related roots, which could go as deep as neural networks, cognition and even epistemology.

Fig. 2 - Illustration from page 315 (Appendix: Brain Connections) from Society of Mind by Marvin Minsky. MIT, 1985

The illustration (cube-on-cube) simplifies Minsky’s abstraction and the principle of “blocks” to represent objects, data, process and decisions (outcomes). The blocks, when connected, can synthesize a variety of entities or networks joined by common digital threads. Alignment of appropriate blocks can create platforms for implementation of IoT. Parts and sub-units can be configured, to create a digital twin of any machine (drill-head). Instances and units of transactions, represented as blocks, may constitute a digital ledger of events, similar to application of financial transactions using blockchain.

For a scenario at hand, please consider the act of driving your automobile (if you can still use a gas guzzler with ICE, that is, an internal combustion engine) to the proximity of a gas dispenser stand, in a gas station (petrol pump).

Your car recognizes “arrival” at the gas station, correlates with low fuel reserve and unlocks the gas inlet. The petrol pump recognizes that your car is within the necessary proximity to the dispenser and recalls your choice for unleaded product. The nozzle from dispenser discovers the gas inlet, and commences fill-up, when your inlet allows, and confirms that the nozzle delivers petrol, not diesel. Once refueling completes, you see a green icon on your dashboard. You receive a SMS, indicating completion of fueling. The latter auto-triggered a financial transaction, to match the cost of fuel. Your bank confirms payment over iSkin or a smartphone app. It also informs your wife (authorized routing) that you have topped off the vehicle with gasoline (petrol).
The convergence of IoT, digital twins and blockchain is evident. The ecosystem of enterprises, when dissociated by modular structures and associated by function, in an operational sequence, presents a series of steps, which can be sub-divided into “blocks” which are not only things/objects but software Agents, units, processes, authentication, authorization, decisions, outliers, feedback, security, metrics and dependencies.46

Who will build these blocks?

Fig. 3 - Blocks – a seamless operation represents an array of functions converging from diverse partner companies47 (Left https://i.io.ua/img_su/small/0010/63/00106374_n1.jpg and cartoon on the right is taken from 48)

As it is with IoT, there will not be any one industry, or company, which may claim to be the front-runner. The modular building blocks, for the many domains spanning and overlapping IoT, digital twins and blockchains (not limited to financial transactions) are quintessential to the global economy.

The idea of a distributed team, or teams entrusted to architect these blocks, may seem reasonable. The fractured state of the world and the intrinsic impact of natural language (in)competencies on creating semantic dictionaries and ontological frameworks introduces severe socio-technical incongruencies.

Hence, credible academic leadership of industry-government consortia, in partnership with global organizations or standardization bodies, may be a prudent option.

If a few global alliances create the blocks, and agree to establish the tools for interoperability, then we may anticipate a future global repository49 for these digital blocks to accelerate global digital transformation. The ubiquitous need for principles and practice of connectivity50 is salient to this discussion.

The value expected from connectivity, assumes operation of multiple ecosystems, which must converge, to deliver the value. What are the layers and components, necessary for this engineering ecosystem?
Table 1. Layers, and components for this engineering ecosystem

<table>
<thead>
<tr>
<th>Layer</th>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Infrastructure</td>
<td>Scaffolds which includes energy, internet engineering, telco networks</td>
</tr>
<tr>
<td>T</td>
<td>Telecommunications</td>
<td>Backbone of connectivity which enables location agnostic operations</td>
</tr>
<tr>
<td>P</td>
<td>Protocol</td>
<td>Transaction triggered response operating agnostic of protocol-specificity</td>
</tr>
<tr>
<td>D</td>
<td>Discovery</td>
<td>Blocks/entities must find each other in order to communicate (think RDF)</td>
</tr>
<tr>
<td>C</td>
<td>Connectivity</td>
<td>Glue that enables digital transformation unless restricted by boundaries</td>
</tr>
<tr>
<td>S</td>
<td>Sense</td>
<td>Data acquired from points of interaction to understand status / attributes</td>
</tr>
<tr>
<td>R</td>
<td>Response</td>
<td>Analytics driven action/actuation based on integrating diverse knowledge</td>
</tr>
<tr>
<td>O</td>
<td>Operate</td>
<td>Outcome as pre-determined or change direction if influenced by factors</td>
</tr>
<tr>
<td>A</td>
<td>Adapt</td>
<td>Ability to remain dynamic and agile by recalibrating operations (eg SCM)</td>
</tr>
<tr>
<td>K</td>
<td>Knowledge</td>
<td>Learnings from operation (store/delete), dissemination, update analytics</td>
</tr>
</tbody>
</table>

Local and global providers, who supply products and services, germane to each layer (and several sub-layers within each layer), may not practice standard operating procedures (SOP). When volatility is the norm, it is wishful to expect SOP or groups in distant corners of the world to conform.

The task of interoperability, and the ability to automate interoperability by “discovering” what is necessary to start communication or cross-check resources, becomes pivotal. It is a tool not yet available. Do we need this tool to discover and replenish the gaps, for functional interoperability to commence?

It is expected that automating interoperability, may lead to auto-generation of APIs when interfaces “discover” that they cannot “talk” between models, data holders, tables, devices. It may trigger an automated mechanism to understand what needs to be understood between the systems, and then obtain the “glue” (for example, creates remote function call to source a “patch” from a repository) to facilitate interoperability. APIs are enablers of interoperability but creating API is not synonymous with instituting interoperability between systems.

True interoperability involves the arduous task of semantic interoperability between systems, to facilitate data-driven multi-system performance optimization.

If someone speaks to me in Hebrew, I must know that I am listening to Hebrew before using Google translate to communicate in Hebrew. If I had a tool to auto-detect languages, then it could help trigger (CNN/RNN51) Hebrew translation on my iPhone. An “avatar” or Siri or Cortana or Alexa, can then, guide my exchange.

**Discovery** tools for auto-detection of attributes, and characteristics between entities, Agents, and models, are a core part of the digital-by-design metaphor, for example, IoT. In the machine tools arena, for example, if a square widget doesn’t fit a round hole, the problem must be reported in a manner that can call for a solution. In this case, the nature of discovery, may trigger the need for software as a robot. Perhaps, endowing the software with some form of mobility, to trigger the adequate response, through connectivity between diverse systems which may not be interoperable.

Research in connectivity may help develop a new generation of digital semantic sensors to sense (what we may not know) what needs sensing. The concept of “sensing the need to sense” must include mobility to be useful in a multitude of environments. It must also move beyond semantic detection and include certain levels of cognition. What we need is a cognitive compass.
The process of discovery, in addition to other elements, is a mix of semantic and cognitive modules. The inability to discover objects or identifying the wrong object or perturbing time-critical discovery (implants, medical devices) are potential pathways to compromising safety, security, privacy, authorization and cybersecurity. Intruders may be sophisticated to avoid semantic detection (5, 50 and 500, are, semantically, numbers, albeit, in different ranges). The importance of cognition and need to introduce cognitive firewalls\(^{53}\) may be gain a wider following. Cognitive “supervisors” may be necessary to keep an eye on the semantic processes.

Thus, claims for ubiquitous computing, first, must find tools for discovery. The illusion of trillions of “things” connected via IoT are delusional, and hyperbole, marketed by uninformed publicists. Unless objects can safely \textit{discover} each other, they may not connect (assuming that connectivity is protocol agnostic). Implementing tools and technologies\(^{54}\) central to discovery, and adequate diffusion of cybersecurity as a service (SECaaS), may be required, \textit{a priori}, to accelerate digital transformation.

Digital twins may depend on safe “discovery” using semantic properties and cognitive rules for auto-configuration (think \textit{ad hoc} auto-configuration of mesh networks) to create the desired assembly (machine floor, medical devices attached to patient, turbines, water purification). Digital twins will inherit attributes of the physical components and physics of the system. Ontology based semantic interoperability by design depends on entity level relationships.

Distributed digital twins created by different sources, approved by cognitive supervisors, and connected by secure systems, may communicate, and form \textit{swarms}, to help us make better decisions (one agent vs an agency or one ant vs a swarm of ants) employing the popular concepts of swarm intelligence.\(^{55}\)
Fraternal Twins: The First Born – Digital or Physical?

A century ago (1916), a theory about freemartins (the female of the heterosexual twins of cattle) generated interest about rare monozygotic twins in cattle. This led to the discovery of Müllerian Inhibiting Substance (MIS). Research by Patricia Donahoe is beginning to unravel the role and therapeutic potential of MIS. It appears that the human genetic program is inherently female (which came first - male or female, the chicken or the egg). If the foetus was left to differentiate without MIS, foetal development of müllerian ducts will produce a female child. The “default” option of human genetics is programmed to produce females. Males exist due to inhibition of the development of müllerian ducts by MIS and differentiation of the Wolffian ducts by foetal testosterone. Females have evolved by design but males appear to be a modification or a by-product of evolution.

This very distant digression is intended to make the point that a fundamental plan, a base, exists in nature. The female plan is copied (duplicated) to produce the male, albeit, with modifications, catalyzed by a starter dose of MIS.

The concept of digital twins assumes we are creating a digital duplicate of the physical entity. But, the born digital metaphor may be, in reality, digital conception. In the industrial arena “things” may not be created without an engineering plan, technical specs and CAD/CAM models. The physical entity must be conceived as a digital model, before birth of the product. In the world of machines and parts, being conceived as digital is the design, to use a pre-natal metaphor. The physical entity is the post-natal stage.

A plethora of exceptions may prove the rule. One such case are add-ons, for example, humans (and animals), attached to medical devices in homes or hospitals. We can create digital twins of this combined state, to monitor their physiology, but that “twin” or digital proxy may not be referred to as born digital. In this instance, for medical purposes, we can think of digital twins as a combinatorial outcome, for monitoring and acquisition of data. By closing the loop and enabling feedback from medical digital twins, we can save lives, thousands, or even millions, of lives.

Think about a helicopter which may not be created as a physical entity unless we have a CAD/CAM version and create a simulation (using differential equations) to test the operation (rotation of the blades for lift-off). In recent models from manufacturers (Boeing 787), the pilot is subservient to the simulated model in the auto-pilot. An image conjured by the latter generated the apocryphal statement that in an airplane of the future there may be only two creatures in the cockpit. A pilot and a dog. The role of the dog is to stop the pilot from touching the controls. The role of the pilot is to feed the dog.

If we reverse the logic of the digital twins we have discussed thus far, one might propose the digital blueprint as the primordial layer and the physical entity to be the fraternal twin (perhaps with limited mobility, if thinking about machines).

The digital blueprint and the simulated models which exists today, may be rapidly engineered with data feeds from the physical operation, to approach the “live” concept of digital twins. In proposing this modus operandi, we move closer to the domain of cyber-physical systems. Time dependencies need time guaranteed software which can understand the semantics of time and resist cyberthreats or time spoofing, by using cognitive firewalls, if time criticality is pivotal for its operation. For example, from (t=0) the decision to apply the brakes to the actual act of braking (t=n) to stop an automobile.
The old world of engineering V models (Fig 5), spanning requirements, design, test, to user documentation. The upper tiers of the V represent the modelling and design phases where engineers build features based on product requirements, which flow to subsequent phases of the development, and manufacturing lifecycles. Individual features of a product are represented in the software as pieces of modular source code that can be "turned" on or off, allowing assets to be generated based on a particular set of features (variant configuration). This accumulation of source code-based assets provides the foundation for creating physical products, but may be leveraged simultaneously to create a corresponding digital twin.

Thus, the physical product is born digital. The digital twin can be found using unique RFID identifiers, or other forms of component serial numbers, once the device is manufactured, and may operate through intelligent PLM platforms, throughout the product’s life. The latter may ensure that the digital twin or the digital proxy, is auto-updated. Standards and certifications are key to adoption of this approach.

**Confluence of Swarms Through Data Fusion**

The digression about the conceptual see-saw, about whose twin is it anyway, is a thread of reasoning, not a barrier. It may make it easier to create the open repositories needed in the process of digital transformation. The road ahead for digital twins, digital proxies and digital transformation is fraught with problems and also brimming with potential.

Driving fusions (please see illustrations below) through collaborative ecosystems may be one path to profit. There may not be a “winner takes it all” version in a cognitive digital twin economy.
Fig. 6 – Digital Transformation: Energy Equilibrium – Elusive Quest for the Digital Mitochondria? To maintain homeostasis of energy production, distribution and load balancing. Adapting to multiple sources and types of energy obtained from diverse producers (domestic, commercial) with variable end points (homes, roadside charger, factories, mobile delivery). Left image from 70 and right image from 71.

Fig. 7 – Digital Transformation: The Transportation Alloy – Alliance of Autonomy, IoT, Telecommunications and 3D Printing72 (cartoon 73)

To profit from fusion in the digital-by-design era, collaborative efforts74 may be one way forward. Examples of convergence (above) may be coupled with their operational digital twins or digital proxies. Information arbitrage from a wide cross-section of similar operations (from many devices) may provide a glimpse of patterns, which were previously unobtainable, due to the focus on one or few operations.

Hence, standards are key, followed by interoperability between standards and other facilitators (converters, adapters, translators, multi-homing) to increase connectivity and reduce incompatibility. The transaction cost may increase and reduce margin of profit, incur losses and downgrade the brand, if architectural and/or structural discrepancies, continue to pose barriers to function and/or outcome, expected by the customer.

A typical laptop deals with 250 compatibility standards. About 20% of the standards are from individual companies, while 44% are from consortia and 36% from accredited standards development organizations. The complexity in laptops will be dwarfed by the variability expected for Digital Twins. The use of ICE (Internalizing Complementary Externalities) like principles may evolve to create working solutions.75
Complementarity, compatibility and interoperability, assures us that we can expect visibility not only of one operation (which is what one industry or one team may monitor) but a group of hundreds or thousands of such operations. This massive data set may help us to understand patterns, predict faults, detect anomalies and use true “big” data, data curation and higher level metadata, to feed other functions, such as data driven policy, security threats and intruder detection using cognitive pattern analytics.

Consider the cartoon (below, left) of a physical event and assume that we have a digital twin of that operation that an analyst or manager can remotely view to “see” or monitor the physical operation in progress. What if the operation on the left is repeated?

![Fig. 7 – Physical System (cartoon adapted from Lego)](image)

If the physical event (left) is not an isolated scenario, then digital duplication (right) may generate a form of digital transformation, representative of a digital swarm or flock.  

This digital vision of aggregated events, may generate big data and metadata, from precision patterns, which may be extracted or extrapolated with respect to process, performance and profitability. Any one instance may not offer sufficient incisive insight, but applying the principles of swarm intelligence to hundreds of instances, may provide wealth of information (not only data) which could enhance decision support systems and improve the monetization potential.

![Fig. 8 – Robots (cartoon from 78)](image)
Swarms of robots (networks, connected via combination of cloud, fog, mist) are likely to generate massive amounts of data. This data, if acquired, analyzed, and used in feedback control, may optimize process, reduce waste and compress production time. Use of “intelligent” decision systems platforms, various levels of automation, and predictive analytics, may transform the concept of ‘zero’ downtime to manufacturing reality.

Data from swarms may improve detection of anomalies, predictive analytics (if equipment needs a part or replacement) and errors or red herrings in the swarm may indicate security threat, breach or may elicit unusual activity alerts.

The blockchain-like digital ledgers, in the backbone of the digital twins, may be useful in identifying the point of anomaly and associated objects or humans in the loop. Combined with advances in hack-proof code and cognitive firewalls, this approach may add a new dimension to systems cybersecurity.

The swarm and flock approach, if applied to the developing notion of smart cities, may offer relatively precise information, from digital twin operations of scale-free networks in urban digital transformation. Monitoring digital proxies of water valves, in operation, to control or regulate water waste, water security and water pollution. From an engineering systems point of view, the digital abstraction is applicable to city-level applications. Cities are inter-dependent cascade of systems and networks such as energy networks, traffic networks, sewer networks, communication networks, road networks and emergency response networks. The latter can make a difference between life and death.

The vision of network convergence may be crippled and remain impotent without architectures which are resilient, fault tolerant, and uses standards which are dynamic.

But, interoperability between standards are rather difficult when competition fuels mistrust, spurs acrimony and short-term profits are the life-blood of the industry. Digital transformations calls for confluence of ideas beyond the horizon and new roads to reach the luminous summit. Investment in scientific vision is often viewed with reservation, excessive caution, undue scepticism and even disdain. The latter is most unfortunate for the progress of civilization.

Digital twins, IoT, blockchains, AI and swarm intelligence may re-define our imagination and future vision of globalization. That sense of the future requires businesses to re-think about ROI and profits, re-configure micro-payments and micro-revenue models, but not at the expense of investment in R&D. The latter is quintessential for innovation, and a tool to catalyze the principles of digital economics, to accelerate globalization.

In part, this idea originates from Marshall McLuhan that anyone, anywhere, may consume the same information. Digital Twins are another example of how this idea may be the reality for humans and machines, at multiple levels.

For example, a manufacturing plant in China, may be operating a component or machine sub-system, (atoms), which is represented by the Digital Twin or the data model of the digital proxy (bits). The global supply chain analyst, in India, monitors the status of the part via PLM. SCM intelligent decision support can trigger a replacement part from a supplier in Tampere, Finland or the component may be 3D printed by DDM Systems in Atlanta, GA and shipped to the factory in Dalian, China.
Taking this idea one step further, if the transport cost of bits approach zero, the transport to Dalian may be replaced, by sending the bits from the 3D architect in USA, to the 3D printer in Dalian, China, to 3D print the part. Digital Twins, through the process of digital transformation, shall lead us, perhaps, to the true digital economy, which once was an idea, implicit in the trade model, proposed by Paul Krugman\textsuperscript{83}, 40 years ago.

3 Concluding Commentary

At the dawn of the 21\textsuperscript{st} Century, the internet was viewed by different groups to serve different functions. In some cases, it served as a storage platform\textsuperscript{84}, others perceived it as a copying machine\textsuperscript{85}, and economists\textsuperscript{86} explored its ability to reduce transaction cost (for example communication, replication, transportation, tracking, verification, search) and democratization of information, as a catalyst for global economic growth.

Digital transformation, in the 21\textsuperscript{st} Century, is the ability to represent atoms, in terms of bits. It stems from the seminal work by Claude Shannon\textsuperscript{87} which grew roots during the 20\textsuperscript{th} Century (Shannon information theory).

Digital transformation is made possible by the internet. Digital Twins and its “lite” version, digital proxy, are a part of the fabric of digital transformation, likely to affect most enterprises, worldwide, willing to duplicate the physical model for remote monitoring, viewing and controlling, based on a digital format (smartphones, tablets).

It is almost justified to think about the internet as a “giant copying machine” which can “copy” physical objects (atoms), to generate corresponding Digital Twins, a representation of information about the atoms, in terms of bits. Hence, we are not dealing with entirely new concepts, just new modes of expression.

Digital Twins are akin to the “emperor’s new clothes” which are made of pre-existing conceptual yarns. We have added new vernacular and embedded the fabric with potentially new widgets (for example, the use of blockchain, as a verification tool).

In this article, we call for an open source approach, to create the “blocks” or modules, necessary to democratize the \textit{ad hoc} and \textit{en masse} configuration of Digital Twins, by \textbf{non-experts}. The latter may no longer limit the use and application of Digital Twins in the hands of experts, alone.

Digital Twins may evolve as a tool, unconstrained by domains, beyond the boundaries of high performing economic regions, and contribute to economic growth, through open source platforms for digitization. Economic\textsuperscript{88} growth from such “digital dark matter” and other intangible benefits to the global economy, remains unmeasured. The impact from Digital Twins, digital proxies and digital duplicates, and cumulative benefits, may be only limited, by our imagination.

Industry must embrace change, imagine paths to reduce transaction cost, and shoulder the need to balance uncertainty, which may accompany the dynamics of digital economics. Leaders must proactively support the call for creating structures, necessary to pursue collaborative initiatives\textsuperscript{89} through investment in massive workforce development, skills training, digital learning\textsuperscript{90}, education, research, institutional advancement and the pursuit of dignity.
Notes

3 www.simonandschuster.com/books/The-Path-Between-the-Seas/David-McCullough/9780743262132
8 www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
9 www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf
11 http://leeseshia.org
12 https://cse.nd.edu/seminars/cse-seminar-series-rahul-mangharam
15 https://thingful.net/
18 http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=34057
20 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.130&rep=rep1&type=pdf
23 http://autoid.mit.edu/people-2

77 • IoT is a Metaphor - Principles & Practice of Connectivity & Convergence • Shoumen Datta


www.seattletimes.com/opinion/editorials/another-nobel-prize-for-uw-scientist-illustrates-value-of-pure-science


http://pioneer.netserv.chula.ac.th/~kkornkar/inter%20trade%20course/a%20model%20of%20Innovation.pdf


http://faculty.csuci.edu/minder.chen/MBA550/reading/hal-varian-information-rules-chapter-1.pdf

http://www.nber.org/papers/w23684

http://www.itsoc.org/resources/Shannon-Centenary/shannon-work-legacy-paper

https://pdfs.semanticscholar.org/32ff/6098dc5d26551107935b78ef4a34f0b9aece.pdf

http://autoid.mit.edu/iot_research_initiative

http://odl.mit.edu/

---

Dear Shoumen,

Thank you so much! This is the BEST report I ever read on the IoT, Industrial Internet, whatever it’s called. I like the evidence-based analysis, the notion of “impotence” of II without data and data analytics, the description of II around the dimensions of Technology, Strategy and Organisation (with an emphasis on culture change), the detailed analysis and predictions about application fields, etc. So well done!

Gerald Santucci
Head of Unit “Knowledge Sharing” at European Commission

---

80 • IoT is a Metaphor - Principles & Practice of Connectivity & Convergence • Shoumen Datta
THIS PAGE IS LEFT BLANK
The elusive quest for intelligence in artificial intelligence prompts us to consider that instituting human-level intelligence in systems may be (still) in the realm of utopia. In about a quarter century, we have witnessed the winter of AI (1990), being transformed and transported to the zenith of tabloid fodder, about AI (2015). The discussion, at hand, is about the elements that constitute the canonical idea of intelligence. The delivery of intelligence as a pay-per-use-service, popping out of an app or from a shrink-wrapped software or AI-as-a-Service, is in contrast to the bio-inspired view of intelligence, formed from a tapestry of events, often cross-pollinated by instances, each with its own microcosm of experiences, and learnings, which may not be discrete, all-or-none functions but continuous, over space and time. The enterprise world may not require, aspire or desire such an engaged solution, to improve its services for enabling digital transformation, through the deployment of digital twins, for example. One might ask whether the "work-flow on steroids" version of decision support may substitute for intelligence? Are we harking back to the era of rule based expert systems? The image conjured by the publicity machines offers solutions with human-level AI and preposterous claims, about capturing the "brain in a box" by 2020. Even emulating arthropods (leave alone cephalopods) may be difficult, in terms of rational AI. Perhaps we can try to focus on worms (*Caenorhabditis elegans*) which may offer what businesses may need, to *pareto-quench* its thirst, for so-called intelligent applications based on elements of AI.
Is Intelligence an Illusion in Artificial Intelligence?

It’s better to keep your mouth closed and be thought a fool than to open your mouth and remove all doubt • Twain

I am an AI optimist. My article on AI (Agents: Where Artificial Intelligence Meets Natural Stupidity) is here http://dspace.mit.edu/handle/1721.1/41914 (2002) • Dr Shoumen Palit Austin Datta, Research Affiliate, MIT Auto-ID Labs, Department of Mechanical Engineering MIT • Please explore website http://autoid.mit.edu
The Elusive Quest for Intelligence in Artificial Intelligence

Is intelligence an illusion in artificial intelligence? This essay is an academic blog. The opinions are due to the author.

Dr Shoumen Palit Austin Datta, MIT Auto-ID Labs, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and MD PnP Lab, Massachusetts General Hospital, Harvard Medical School

The promise, and pessimism, about AI (artificial intelligence) charts a sinusoidal path. The questions about “intelligence” in AI persists. Many more questions are meeting with fewer answers. Intelligence in AI may be an advanced form of multi-level dynamics with/without a higher order of structured complex networks or (even better) temporal networks. The semantics depends on, and is colored by, one’s view of systems or interpretation of what is intelligence, what constitutes proof of intelligence, and the nature of complexity.

Figure 1: The Blind Men and the Elephant • John Godfrey Saxe (1816-1887)

In systems thinking a common analogy is that of six blind men from Indostan (India) touching various parts of an elephant and claiming that “elephant is a tree” (man who touched the leg), “it is like a rope” said the blind man who touched the tail. ”Like a snake“ (man who touched the trunk). ”It is like a big hand fan“ (man who touched the ear). “Like a huge wall” said the man who touched the belly of the elephant. “Like a spear” said the blind man who touched the tusk. They were misguided in describing their part and wrong about the whole picture. The definition of intelligence may be colored by the professional bias of the interpreter. Intelligence is not a point. It is a fabric or an array or collective continuum of network systems which may not be boxed with human skills, yet (the latter is by far the most ludicrous and incredible claim, at the present time, under our known circumstances).
Why this mad pursuit?

University of Cambridge, which had done best at teaching mathematics, is the one from amongst whose graduates, have come more of the English poets. While Oxford, which has specialized in the humanities, has tended to turn out writers who have attained, on the whole, a high level of mediocrity. By the time one has discussed literature with a witty and learned professor, one knows what has been achieved, and how good it is. You become respectful, and begin to wonder, who am I to do better? (Dialogues of Alfred North Whitehead (1861-1947) as recorded by Lucien Price. Little, Brown ▪ Boston, 1954)

The definitive treatise by Stuart J Russell and Peter Norvig (Artificial Intelligence: A Modern Approach, 3rd edition, 2015) should suffice to discourage all amateurs (myself included) to desist from testifying about the trials and tribulations with respect to AI.

Who am I to do better?

In this neo-Norvig-ean era of human-level AI claims, we find LISP programming language, the 2nd oldest language, since FORTRAN. It is still regarded as a powerful AI tool. Figure 1 in the first LISP paper (John McCarthy, MIT) illustrates “Representation of S-Expressions by List Structure” which takes us back to the history of perceptrons elaborated in the seminal paper by Warren McCulloch and Walter Pitts (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity in Bulletin of Mathematical Biophysics 5 115-133 (marks the dawn of AI).

The perceptron algorithm created by Frank Rosenblatt, was championed by the US Navy, in fueling the “intelligence” controversy, to the extent, that The New York Times reported (in 1958) the perceptron to be "the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence." A decade later, Marvin Minsky and Seymour Papert observed (in Perceptrons, MIT, 1969) that such statements may be guilty of the wildest exaggeration. However, Minsky is no stranger to extending the euphoria, and feeding the frenzy with predictions about AI’s rosy potential. Extending the scope of intelligence in AI has always been in vogue and continues unabated.

It is well nigh impossible to over emphasize that perceptrons, decision trees, recursion functions, etc, are good topological representations and manifestations of the anatomy, which are visible, in the context of neural connectivity, and the neural networks, which we observe, for example, in worms, eg Caenorhabditis elegans (Robert Horvitz, MIT; figure 3).

Abstractions of these networks, eg, cube-on-cube (The Society of Mind by Marvin Minsky, MIT, 1985) and “learning” mechanisms (The Organization of Behavior by Donald O Hebb, 1949) may be building blocks of intelligence. These processes populate fields with values from the user environment which can be selectively used (per contra hard coded defined sets). NEST Learning Thermostat uses input values, to tune your preferred temperatures.
Elements of the equation/rule-based (brittle, static) structures caused the bust of expert systems. It ended the lure of *The AI Business* (Winston and Prendergast, MIT, 1984) before the rise of artificial neural networks (ANN, popularity circa 1990). Topology and (dubious) synaptic weights, in stochastic models, are suggestive of flexible infrastructure. The intent is to improve profitability from data. But, is it really an application of intelligence in AI?

Rather than partial differential equations, exploding due to increase in state functions (due to the large number of parameters), use of Agents allow each variable to be represented as a single-function entity. Collective output from an Agency of Agents improves predictive or prescriptive precision, compared to operations research applications (see Figure 2). The behavior of Agents and Agencies using “AI” concepts originated from the principles of stigmergy (Pierre-Paul Grasse, 1959) which continues to evolve.

**Think Data before Thinking of AI**

The recent surge in the hype associated with “big data” has navigated profitability from analytics to the front and center. Intelligence is marketed as a commodity in this scenario.

![Figure 2: Pursuit of the Luminous Summit ($\text{Intelligence as a Service}$)](image)
In order to market intelligence as a service, the AI paradigm is being refurbished as a commodity (brain in a box), and touted to business and industry as an essential tool to reach the luminous summit [$]. Winning at games using ANN is advocated as intelligence. Smart and intelligence are emerging as speculative tabloid fodder. Witnessing the rapid transmutation of tabloid fodder from speculation to business truth is deeply troubling.

Claims about “original” thinking in containerization of data and processes are good ideas, but are concepts proposed almost half a century ago by Marvin Minsky (page 315 in the original book or search page 311 in this PDF version of the book). Connecting entities (containers) using IPv6 resonates with ideas suggested about a decade ago. However, it is reassuring that the concepts are not lost, but are being developed to advance the march of digital transformation (see Digital Twins https://dspace.mit.edu/handle/1721.1/104429).

Data, especially the volume of curated data, and the fundamental tools of knowledge representation (KR), are required for deploying AI and machine learning (ML) techniques. Data, AI, ML, are topics of intensive investigation, grand debates and energetic discourse, for at least half a century. However, the accelerating pace of rancor in this field, stems from marketing pundits and less informed celebrities, of a certain type.

Philosophical notions, linguistics and epistemology, in convergence with logic, ontology and computable models, fuels KR. Comprehending these principles, in combination with conventional mathematics, statistics, computer science and neurology, begins to form the foundation of AI analytics. To profit from AI applications, depth of domain knowledge must be added, to this foundation.

Those who can grasp this confluence, are converts, proselytizing about abstractions. The call for abstractions, which can hide the complexity of these principles, are essential for mass consumption of AI/ML applications. Intuitive, friendly, user interfaces (think Lego Mindstorms) which can be configured by the masses, lacking in specialized skills, are the “killer” apps, which will transform business, commerce, government, academia and any instance, any event, any operation, where decision may be necessary, before the next step.

Excruciatingly detailed level of minutiae are germane to KR frameworks. It remains quite inaccessible to the global population, except a few thousand computer science graduates. Minor changes in the sequence of symbols, in predicate logic (predicate calculus), can make a crucial difference in the meaning, as illustrated in the following example from John Sowa:

\[
(\forall x)(\exists y)((\text{woman}(x) \land \text{dept}(x,MechE)) \supset (\text{man}(y) \land \text{hometown}(y,Boston) \land \text{married}(x,y))) [I] \\
(\exists y)(\forall x)((\text{woman}(x) \land \text{dept}(x,MechE)) \supset (\text{man}(y) \land \text{hometown}(y,Boston) \land \text{married}(x,y))) [II] \\
(\forall x)(\exists y)((\text{man}(x) \land \text{dept}(x,MechE)) \supset (\text{woman}(y) \land \text{hometown}(y,Boston) \land \text{married}(x,y))) [III] \\
(\exists y)(\forall x)((\text{man}(x) \land \text{dept}(x,MechE)) \supset (\text{woman}(y) \land \text{hometown}(y,Boston) \land \text{married}(x,y))) [IV]
\]
Is Intelligence an Illusion in Artificial Intelligence?

It’s better to keep your mouth closed and be thought a fool than to open your mouth and remove all doubt ● Twain

[I] in English and French, respectively, may read as follows:
Every woman in department of Mech Engineering married a man who came from Boston
Chaque femme du département de génie mécanique a épousé un homme venu de Boston

[II] in English and French, respectively, may read as follows:
A man who came from Boston married every woman in department of Mech Engineering
Un homme venu de Boston a épousé toute femme dans le département de génie mécanique

[III] in English and French, respectively, may read as follows:
Every man in department of Mech Engineering married a woman who came from Boston
Tout homme dans le département de génie mécanique a épousé une femme venant de Boston

[IV] in English and French, respectively, may read as follows:
A woman who came from Boston married every man in department of Mech Engineering
Une femme venant de Boston a marié tous les hommes dans le département de génie mécanique

The difference between the meaning of [I] and [IV] leaves room for social disequilibrium.
The tiny change which could ignite such an immense social cataclysm may be attributed to
an almost imperceptible alteration of the order of the symbols between [I] and [IV]. The
latter may be obvious, only to a handful of people, in the entire world. Hence, the clamor for
abstraction.

Tools for the masses must sweep these intricacies under “big buttons” which one can drag
and drop, to convey their intent, through rudimentary natural language contexts. Anyone
can create programs, without the faintest idea of what is a program, or any knowledge of
programming, whatsoever (Lego Mindstorms). When outcome, function, and output, can be
optimized by the masses, the floodgates to profit from rational AI apps, may cease to close.
It will be a part of the fabric of our daily lives (quote from Herbert Simon, CMU).

Before we reach that halcyon era, we must deal with data. Data must feed algorithms, some
of which, may be based on principles of AI and ML. However, most gains arise from great
features, not great ML algorithms. Often, a heuristic model can solve 80% of the problem
even without ML. Feature engineering is a critical skill in data science. Incorrect features
will lead to irrelevant or imperfect training for algorithms. Before using a model live or
exporting a model, the performance of the model must be checked using data. Anomalies in
performance is an indicator of problems either with data, features or the model or all of the
above. Continuous checking of the AUC (area under ROC (receiver operating characteristic)
curve) is a prudent approach to validate performance of models and their suitability.

Pg 6 ● I am an AI optimist. My article on AI (Agents: Where Artificial Intelligence Meets Natural Stupidity) is
here http://dspace.mit.edu/handle/1721.1/41914 (2002) ● Dr Shoumen Palit Austin Datta, Research Affiliate,
MIT Auto-ID Labs, Department of Mechanical Engineering MIT ● Please explore website http://autoid.mit.edu
Is Intelligence an Illusion in Artificial Intelligence?

It's better to keep your mouth closed and be thought a fool than to open your mouth and remove all doubt ● Twain

Human-level AI?

It is hard to improve the statement from Rodney Brooks, hence, here is a verbatim quote: “We have found a way to build fixed topology networks of our finite state machines which can perform learning, as an isolated subsystem, at levels comparable to these examples. At the moment, of course, we are in the very position we lambasted most AI workers for, earlier in this paper. We have an isolated module of a system working and the inputs and outputs have been left dangling.” (Intelligence without Representation, 1991)

Learning triggers profound, sustained, often long term changes in our neural networks at many levels that we cannot even begin to understand, or grasp, its cognitive repercussions. Almost all assumptions made by McCulloch & Pitts (1943) may be violated (Appendix 1).

The “all or none” phenomena assumed by McCulloch & Pitts (1943) is relevant from a mechanical perspective if one assumes (perhaps incorrectly) that input data is supposed to transduce a signal, and the resultant action potential (neuronal activation), may be one form of a proof of learning. Neurologists may strenuously and vociferously take exception.

AI experts may wish adopt this view, about “learning” in the AI context. The neurological state of learning, cognition, and behavior is usually a continuous function, modulated by evolutionary weights, which are not subject, in the least, to the limitations of discrete-state machines. Application of machine learning models are often inconsistent and incorrect.

Discrete systems have a finite (countable) number of states, which may be described in precise mathematical models. The computer is a finite state machine which may be viewed as a discrete system. The brain is not a computer. The neural infrastructure and networks are not finite state machines. Imposing any such model (real-world continuous systems) or ill-advised abstraction or gross extrapolation (by those not so well informed) may only perpetrate great lengths of fantasy, about intelligence, and learning, related to AI systems.

“Of the vast stream of sense data that pour into our nervous systems, we are aware of few and we name still fewer. For it is the fact that even percepts are wordless. Only by necessity do we put a vocabulary to what we touch, see, taste, and smell, and to such sounds as we hear, that are not themselves words. We look at a landscape, a rich carving and majestic architecture of a cathedral, listen to the development of harmonies in a symphony, or admire special skill in games and find ourselves woefully lacking in ability to describe our percepts. Words, as we very rightly say, fail us, either to describe the plain facts of these experiences or to impart to others, our feelings.” (G Jefferson CBE, FRS, MS, FRCS, Professor of Neurosurgery in The Mind of Mechanical Man in British Medical Journal, 25th June 1949). The author was aware of “Dr Wiener of Boston, his entertaining book Cybernetics (1948).”
Alan Turing was cognizant of the over-reach in claiming “intelligence” in AI and outlined potential objections including Godel’s theorem (mathematical objection) and “Argument from Consciousness” which he reproduced from Professor Geoffrey Jefferson as a quote (from his Lister Oration, 1949) "Not until a machine can write a sonnet, or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain. That is, not only write but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants.” A. M. Turing (1950) Computing Machinery and Intelligence. Mind 49 433-460 (PDF)

Page 452 (see Appendix 2) removes any doubt that Turing had grave doubts regarding claims of intelligence in the context of computers. Turing’s suggested starting point is “the child machine” (Appendix 2). Then he proposes to add the roles or processes of “evolution” “hereditary material” “mutation” “education” and “natural selection” in order to mature “the child machine” to “imitate an adult human mind” as a path forward to intelligence. To understand even vaguely what happens after “the initial state of mind, say at birth” the reader is urged to review Patterns in the Mind by Ray Jackendoff (1966) and then take into consideration the field of linguistics and natural language development (1970, PhD thesis of Terry Winograd, MIT http://hci.stanford.edu/winograd/shrdlu/AITR-235.pdf).

For all this to happen, we must process information encoded via developmental and environmental signals. Hence, the suggestion, research and convergence on the concept of molecular logic gates. The complexity of the process may help deter one from concluding that we are dealing with intelligence, with respect to computers, machinery or AI systems.

However, the human spirit and the fabric of scientific research cannot step away from problems, even if all available reason suggests that something is impossible, at the time. It is with this fervor the 1956 Dartmouth Summer Research Project on Artificial Intelligence (June 17 - August 16) was proposed in 1955 by a visionary group of eminent and erudite academic scholars (www.aaai.org/ojs/index.php/aimagazine/article/view/1904/1802).

The proposal (see Appendix 3) admits it is a “conjecture” but continued “that every aspect of learning, or any other feature of intelligence, can, in principle, be so precisely described that a machine can be made to simulate it.”

Great strides (Appendix 4) have been made, yet the 1956 Summer Research “conjecture” looms overhead. Progress of AI is evident from the 1145 page book by Russell and Norvig (AI - A Modern Approach, 3rd ed). We are also learning how decisions can be made without a brain in cognitive organisms (unicellular mould Physarum polycephalum).
**Neurobiology 101**

Topology and weights are the foundational underpinnings of artificial neural nets (ANN) which are the mainstay of AI systems. How reliable are these extrapolations? Are we still talking about AI? Let me reiterate what Rodney Brooks has stated but in a different vein.

Structural design of network topology aims to mimic the commonly observed organization of neurons. Topology based on neural organization (small world networks) may be fraught with errors, as evidenced by studies on wiring configuration and neuroanatomical analysis which reveals differences in circuit architecture and connectivity, if viewed at mesoscopic vs microscopic scale. On a mesoscopic scale, seemingly random networks exhibit consistent properties. It may be difficult, if not impossible, to extract useful/meaningful abstractions from these counter-intuitive non-linear yet dynamic structure-function complementarities.

In ANN, weights are assigned to signify connectivity strengths (the links between the perceptrons). These are arbitrary, at best, because synaptic weights between neurons and clusters are subjected to conditions not well understood. Even if one acquired neuro-physiological data related to frequency variations of action potentials (~200 Hertz), in an attempt to understand synaptic weights between neurons, the results may be inconclusive. The complexity may be compounded by the fact nerve transmissions are modified by ions, electrical threshold potential, chemical neuro-transmitters and their location in the brain.

In synaptic design, one assumes the all-or-none process (Appendix 1) and the weights are modeled based on extrapolation from “inferential changes” which are in the order of milliseconds to seconds (hence, subject to observation, data collection and extrapolation). But, the nature of the connectivity and resultant weight is also influenced by epigenetic factors (time scale – seconds to days), ontogenic factors (days to years) and phylogenetic factors, which are the result of generations or are derived from the evolutionary time scale, as noted in Appendix 2. Hence, the nature of the weight deduced from “inferential” changes (primarily sense and response mechanisms) are only the tip of the iceberg. We are almost completely in the dark about the nature of the influence from these and many other factors.

Synaptic weights ascribed in stochastic models are based on heuristics. This approach may not be sufficiently informed to account for the complex distribution of spike rates, observed synaptic weights and intrinsic excitability for neurons, in different areas of the brain, for example, auditory or visual cortex, hippocampus, cerebellum, striatum, midbrain nuclei. These processes underlie neuroplasticity and the potential for continuous configuration and reconfiguration of the brain through dynamic Hebbian Learning (Donald O. Hebb).
By **definition**, in Hebbian learning (one of the oldest learning algorithms), a synapse between two neurons is strengthened when the neurons on either side of the synapse (input and output) have highly correlated outputs. Following the analogy to an artificial system (ANN), the tap weight is increased with high correlation between two sequential neurons. Capturing this dynamic change in weights and including this change appropriately in the function of the algorithm, requires feeding the algorithm. In other words, getting the correct data with the correct features (discrete features and/or crosses) to the algorithm.

Figure 3: Distribution of neural spikes and intrinsic excitability from various parts of the brain indicate a range of responses. A representation of artificial neural network (center). Scheler, Gabriele (2017) Logarithmic distributions prove that intrinsic learning is Hebbian. *F1000Research* **6** 1222 (https://f1000research.com/articles/6-1222/v1)

Heuristic weights used in ANN (Figure 3, center) may not resemble reality, with respect to observed pattern of tap weights, in biological neural nets. Evidence of this discrepancy is based on observations of variability (Figure 3), central to biological processes, learning and neuroplasticity. This information may be useful when training artificial neural nets (ANN).
Untrained memory network is a structural representation (Figure 4, left panel) of a minuscule segment of the 100 trillion synapses in the (human) brain. When humans or animals learn certain behaviors, the synaptic relationship changes. A class of the latter type may be referred to as Hebbian learning (a synapse between two neurons is strengthened when the neurons on either side of the synapse, input and output, have highly correlated outputs). But, learning is not an unimodal action. Animals and humans are multi-sensory, multi-modal and multi-dimensional learners. Hence, an event cannot be mapped to change in one synapse. A simple learning, may impact a neural network, no matter how small.

Figure 4: Untrained memory network (left panel). Trained memory network (right panel)

Re-visiting the classic example of Ivan Pavlov's dog (Spot) the “food” signal is a complex multi-sensory input. For simplicity we dissect it as follows: [1] auditory signal (Spot, Food!), [2] olfactory signal (smell of food) and [3] optical signal (sight of food). Assuming that Spot did not understand Russian or English, using the Hebbian learning principle, calling out "Spot, Food!" is just a sound, albeit emphatic, but still a sound, but it follows a specific form, a pattern, which repeats (reinforcement learning). In course of these experiments, Ivan Pavlov’s dogs may have “learned” to differentiate between arbitrary background chatter and the call "Spot, Food!" by their owner. Because of the natural language ambiguity, the blue input (Figure 4, right panel) manifests as a turquoise output. The choice of the closely related colors may indicate reduced synaptic weight, and perhaps insinuate a different strength of learning in this memory circuit. However, the auditory signal (Spot, Food!), in combination, with other signals, sends the correct message (sounds like food) to the neural controller. The combinatorial and synergistic effects, are weighted differently, and are important determinants in memory circuits.
The blue input signal synapses with the green input signal, as if to validate the incoming olfactory signal (green cluster), since the sense of smell is a signature stimuli, for dogs. The output from the olfactory signal (input, green) is definitive (output, green) and strengthens the output message (food). The sight of food (input, red) adds to the combined output but it may be weak (output, pink, related color to red, input) because humans are unclear if dogs are capable of **visually** differentiating between chicken tikka masala v beef stroganoff.

I digress to make the point about “humans” understanding or not understanding. Logic, ontology and programming of computable models, is colored by the syntax and semantics of humans, their natural languages and their habitats. The same problem approached by an English, Chinese and Indian person, may produce three, related, but different outputs. This underscores the critical importance of semantics and interoperability between standards and interoperability between platforms. Hence, the role of interfaces is key to abstraction.

Returning to neural nets and weights, the point of this explanation is to convey that neural networks, naturally, lead to memory circuits. In this case, the dog’s memory circuit gets trained, over time (hence, these are linked to time series) to sight, sound and smell. The training creates patterns, corresponding to the input signals feeding the network. The dynamic nature of this training is represented by colors of the neural cell clusters, with colors representing strength of connections.

The complexity of this very simple event illustrates how learning, synaptic weights and modifiers, influences memory. In this case, you can execute associative control logic by having an integrator/arbitrator neural cluster (sight and sound). The degree to which we can abstract these principles and create ANNs closely resembling natural neural nets remains an open question.

But, it should not discourage humans from trying to use bio-inspired models in developing and using ANNs to support “intelligent” decision support systems. No matter how intricate and multifactorial natural neural nets may be, humans are persevering in their attempt to create ANNs and use it as effectively, as possible, even if “intelligence” remains a mirage.

For example, you can plug in the memory network for the stimulus represented by green neural cell cluster (Fig 4) and track the memory of an event, executing the associative logic.

---

Similarly, we can integrate associative control clusters and create a complex of associative control logic. In Figure 5 (right), two associative control circuits are integrated. Each circuit feeds a gray memory circuit above it. The two memory circuits tie in to a control circuit (black node) that feeds into a memory circuit at the apex (gray node). This topology can execute a complex set of behavioral controls in response to a complex set of multi-source stimuli. It has an internal memory of the actions taken in response to the situation. The latter is reminiscent of human action, that is, the ability to remember actions in a given set of circumstances. The ability of humans to show restraint and the ability to “forget” are important elements which may be driven by higher level cognition. Pragmatic applications of AI are keen to capture part of this “human problem solving and processing ability” due to neural networks. It is certainly worthy of mimicry.

![Figure 5: Simple associative control (L) and association complex with internal memory (R)](image)

Taken together, perhaps we shall invest more thought in the design of ANNs by re-visiting assumptions which are generally incorrect because we are still significantly uninformed. Having said that, one must hasten to add, that, no matter how approximately correct the synthetic weights, may be, it may not be impossible to conceive building better ANNs with partially unsure topologies and unqualified numbers (Figure 3) using the Pareto Principle. Perhaps we can achieve 80% of the goal, using the approximations, at 20% of the cost.

Clever integration of tools and techniques (for example, Luenberger Observer, back propagation algorithms) may allow us to tune and re-tune AI systems in a dynamic data-driven manner and improve the learning, as discussed above, to generate actionable information. Over-fitting the model may cause harm, for example, in collision avoidance. Understanding the principles that govern these processes (neurobiology) may improve the thinking of analysts to use better design principles while constructing dynamic models.
Scholars continue to discuss new ways of using robots to make robots, create self-healing intelligent machines and adaptive machines to optimize up-time. Thinkers are conjuring up ways to harness the developmental foundations of neurons – neurogenesis. Emulation of neural development using computational AI systems can incorporate characteristics of natural neural systems into engineering design. Scientists are claiming that rather than designing neural networks, emulation of neurogenesis shall enable us to generate neural networks to serve dynamic and even more complex systems of the future. This emerging field of programmable artificial neurogenesis appears to call for a meta-design paradigm which may begin with components (objects?). It aims to build higher order intelligent systems which will adapt (to demands, environment, resources) without re-programming component level entities. When components are updated, the changes will be propagated, via appropriate “learning” functions, up/down hierarchies.

The great desire to emulate the grand vision latent in intelligence, cognition and the brain, works almost as an aphrodisiac. The immense powers of biology, and the ability to distil and capture even an iota of that potential, from bio-inspired systems, through convergence with computation, will continue to be a Holy Grail. Here is one example. We have about 3 billion base pairs (A-T, G-C) in the human genome (3x10⁹) which codes for about 10,000 – 20,000 genes resulting in a human body with 100 trillion cells (1x10¹⁴). At least a third of the approximately 20,000 different genes that make up the human genome are active (expressed) in the brain. We have about 8.5x10¹⁰ neural cells (there are an equivalent number of glial cells). Each neural cell connects on an average with 1,000 other neural cells to create about 1x10¹⁴ neural connections (100 trillion synapses). This is the natural neural network which makes us human, brews intelligence and cognition. In terms of compression ratio, the ratio approaches 10¹¹ (7,000 genes creating 1x10¹⁴ connections). From the non-biological world, the most effective compression algorithm CMIX doesn’t even come close. The illegal 42.zip bomb which unfolds to 4.5 petabytes (pb) from a 42 kilobytes (kb) single symbol zip, approaches a compression ratio of 10¹¹ in an artificial circumstance devoid of intelligence. The human compression of 10¹¹ offers sustainable, real, life-long intelligence.

**Conclusion – These ‘intelligent’ machines may never be intelligent in a human sense (p339)**

A quantum leap, still cryptic within the unknown unknowns, may unleash what is intelligence within AI, in the future. We must continue to explore far and wide, emulate insects and think about the Octopus. “We can only see a short distance ahead, but we can see plenty there that needs to be done.” Convergence of tools (statistics, math) with data curation (noise vs signal) is replete with promise and profitability even if it lacks (human-level) intelligence. We must continue to explore ways to profit from AI (Appendix 6).
A Logical Calculus of Ideas Immanent in Nervous Activity

adequate. But for nets undergoing both alterations, we can substitute equivalent fictitious nets composed of neurons whose connections and thresholds are unaltered. But one point must be made clear: neither of us conceives the formal equivalence to be a factual explanation. Per contra—we regard facilitation and extinction as dependent upon continuous changes in threshold related to electrical and chemical variables, such as after-potentials and ionic concentrations; and learning as an enduring change which can survive sleep, anaesthesia, convulsions and coma. The importance of the formal equivalence lies in this: that the alterations actually underlying facilitation, extinction and learning in no way affect the conclusions which follow from the formal treatment of the activity of nervous nets, and the relations of the corresponding propositions remain those of the logic of propositions.

The nervous system contains many circular paths, whose activity so regenerates the excitation of any participant neuron that reference to time past becomes indefinite, although it still implies that afferent activity has realized one of a certain class of configurations over time. Precise specification of these implications by means of recursive functions, and determination of those that can be embodied in the activity of nervous nets, completes the theory.

THE THEORY: NETS WITHOUT CIRCLES

We shall make the following physical assumptions for our calculus.

1. The activity of the neuron is an “all-or-none” process.
2. A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous activity and position on the neuron.
3. The only significant delay within the nervous system is synaptic delay.
4. The activity of any inhibitory synapse absolutely prevents excitation of the neuron at that time.
5. The structure of the net does not change with time.
APPENDIX – 2

Page 452

A. M. Turing (1950) Computing Machinery and Intelligence. Mind 49 433-460

In the process of trying to imitate an adult human mind we are bound to think a good deal about the process which has brought it to the state that it is in. We may notice three components.

(a) The initial state of the mind, say at birth,

(b) The education to which it has been subjected,

(c) Other experience, not to be described as education, to which it has been subjected.

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child’s? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer’s. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

We have thus divided our problem into two parts. The child programme and the education process. These two remain very closely connected. We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns. One can then try another and see if it is better or worse. There is an obvious connection between this process and evolution, by the identifications

Structure of the child machine = hereditary material

Changes of the child machine = mutation,

Natural selection = judgment of the experimenter

Pg 16 • I am an AI optimist. My article on AI (Agents: Where Artificial Intelligence Meets Natural Stupidity) is here http://dspace.mit.edu/handle/1721.1/41914 (2002) • Dr Shoumen Palit Austin Datta, Research Affiliate, MIT Auto-ID Labs, Department of Mechanical Engineering MIT • Please explore website http://autoid.mit.edu
Is Intelligence an Illusion in Artificial Intelligence?

It’s better to keep your mouth closed and be thought a fool than to open your mouth and remove all doubt. — Twain

APPENDIX – 3

1955 Dartmouth Summer Research Proposal


A Proposal for the
DA RTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE
June 17 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1) Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2) How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning.
Is Intelligence an Illusion in Artificial Intelligence?

It’s better to keep your mouth closed and be thought a fool than to open your mouth and remove all doubt ● Twain

APPENDIX – 4

1943 Warren S McCulloch and Walter H Pitts A Logical Calculus of the Ideas Immanent in Nervous Activity in Bulletin of Mathematical Biophysics 5 115-133
http://www.cse.chalmers.se/~coquand/AUTOMATA/mcp.pdf

1949 Geoffrey Jefferson The Mind of Mechanical Man in British Medical Journal June 25

1950 Alan M. Turing Computing Machinery and Intelligence in Mind 49 433-460

1955 Proposal for Dartmouth Summer Research Project on Artificial Intelligence

1960 Marvin Minsky Steps Toward Artificial Intelligence in Proceedings of the IRE
http://web.media.mit.edu/~minsky/papers/steps.html
http://www.cs.utexas.edu/~jsinapov/teaching/cs378/readings/W2/Minsky60steps.pdf

1966 Ray Jackendoff Patterns in the Mind

1970 Natural Language – PhD Thesis of Terry Winograd, MIT

1974 Marvin Minsky A Framework for Representing Knowledge in MIT AI Lab Memo 306

1984 Patrick H Winston Artificial Intelligence: A Perspective
https://mitpress.mit.edu/sites/default/files/titles/content/9780262570770_sch_0001.pdf

1984 Winston & Prendergast The AI Business https://mitpress.mit.edu/books/ai-business


2015 Stuart J Russell and Peter Norvig Artificial Intelligence - A Modern Approach (3rd ed)

Pg 18 ● I am an AI optimist. My article on AI (Agents: Where Artificial Intelligence Meets Natural Stupidity) is here http://dspace.mit.edu/handle/1721.1/41914 (2002) ● Dr Shoumen Palit Austin Datta, Research Affiliate, MIT Auto-ID Labs, Department of Mechanical Engineering MIT ● Please explore website http://autoid.mit.edu
APPENDIX – 5 – What business needs for sustainable long term profitability

As an AI optimist, one cannot leave “AI” hanging without some form of a path forward to deploy the treasure of principles. The discussion about lack of “intelligence” in AI should not indicate the author’s lack of support for the ideas that are represented through AI. Informed individuals are likely to agree that human level intelligence is still utopian. A cursory reading of Chapter 1 of Russell and Norvig’s seminal work (2015) will enlighten even a pre-teen student that the flight of an air plane is not artificial flight which mimics birds and artificial intelligence or AI is not similar to human intelligence. The introduction of the term “AI” by John McCarthy during the summer of 1956 was a genius stroke of public relations and marketing coup completely unbeknownst to John McCarthy or others in 1956.

Definitions are often useless and narrow but if one must re-define “AI” then consider this: The science and art of designing engineering tools, systems and platforms to perform tasks that generally requires human input, as a gradient of responses (segmented by simplicity or complexity) based on rational, contextual, and optimal attributes of human reasoning.

Those seeking to distill the “value proposition” of AI may consider this verbose statement: The value of AI is in performance of tasks, using one or more tools or systems, repeatedly and accurately, often, at a very high frequency, often, with immense sets of structured or even unstructured data, in various possible combinations of the features, of the data, using a dynamic selection of algorithms and solvers, from the system or platforms or several other sources (cloud, fog), in near real-time, or with negligible latency which the outcome can tolerate. The delivery of the value of AI is inextricably linked to the strategic value of data and/in connected networks. The value of AI analytics may be reaped at the edge or core or in sub-systems, if the outcome or output can effectively aid decision support systems, within bounded latency, to improve performance and/or actuate/execute events, without further human input, in near real time, or at the correct time, to optimize the scenario or improve profitability from the decision or outcome based on AI data analytics.

I am an AI optimist. My article on AI (Agents: Where Artificial Intelligence Meets Natural Stupidity) is here http://dspace.mit.edu/handle/1721.1/41914 (2002) ▪ Dr Shoumen Palit Austin Datta, Research Affiliate, MIT Auto-ID Labs, Department of Mechanical Engineering, MIT ▪ Please explore website http://autoid.mit.edu
Is Intelligence an Illusion in Artificial Intelligence?

It's better to keep your mouth closed and be thought a fool than to open your mouth and remove all doubt ● Twain

Organizations and corporations are perhaps most interested in the AI economy. Hence, the application and deployment of AI tool kits. To achieve these goals, at least in part, diffusion of the principles and practice of AI tools must precede the P&L quest for profitability. The most critical catalyst for breaking new ground is information arbitrage. AI is no exception.

The industry needs skilled communicators with content knowledge, those who can simplify but refrain from being simplistic, explain without dilution or downsizing the material and reach the executive core as well as practitioners without making it a journey of “billable hours” consulting. The communicators must also understand strategy and management.

Strategic communication of the science and engineering, related to the principles and practice of AI, requires a synthesis of logic, ontology and computation. Taken together, the trinity, is the foundation of knowledge representation (KR). The abstraction of KR may be industry agnostic for certain classes and industry specific in many instances. Widespread understanding of these elements and the participation of the industry specific knowledge experts in their development, is the Holy Grail, which is essential for systems to benefit and profit from AI. The strategic value of this modus operandi must be in balance rather than serving AI as a one shoe fits all or a panacea for all the woes of decision support systems.

To be driven by data, AI tools cannot exclude the principles of data handling, feature engineering, data curation and how to use dynamic high volume data to extract actionable information. These are not apps which teenagers can build in a few hours. A most sincere concerted effort is a pre-requisite. This is a field where certain tools and standards exist. Inclusion, interoperability between standards and integration may be essential, in certain domains. The ability to cross-pollinate existing and available tools from different sectors with new, novel or unique capabilities, requires a dynamic platform, agile in its ability to adopt and adapt. These attributes are missing in certain vendors who promote their “all-in-one” package or install cryptic boundaries (example: closed data dictionaries or semantic incompatibilities between knowledge representation modules). In a volatile economy, any restriction may become a detriment to global businesses and an anathema for profitability.

These roadblocks are nothing new and are usually organizational. The general disregard for experts and academic-industry collaborations, fuels the “illusion of expertise” which often retards innovation. Evolution of digital transformation can easily fall prey to market forces which manufactures these veneers or silos. One role of the visionary leader or the entrepreneurial corporation is to recognize the mirage created by illusions of completeness and prevent the delusion from asphyxiating the transformation of vision into reality.
Businesses need intuition and imagination to inculcate vision. The leadership may invest in building the capacity to execute in a manner that catalyzes entrepreneurial innovation. The path to AI driven data analytics is calling out for leaders. Also, AI calls for confluence, even more than other broad areas. Context of AI applications are pivotal in order to reduce the garbage-in/garbage-out syndrome. Cloud, cybersecurity, computation and communication must converge for the data to be available, secure, analyzed and used before the value of the information perishes. These and other factors are critical for success of AI applications.

For businesses to stay in business it must look beyond business, for networks of innovation which will depend on the supply chain of global talent. The excitement about new tools and technologies may die a premature death in a manner similar to the hypothetical demise of the industrial revolution, if the world ran out of coal. Data is not the new coal, or new oil or new goal. Talent and education was and still remains the coal to ignite our imagination. The wealth of nations must be invested to cultivate talent through education. The US ought to be deeply concerned by the diminishing number of STEM graduates from its institutions.

According to the World Economic Forum, US produced 568,000 STEM graduates in 2016. If 1% of these students (5,680) are in the general area of AI, and if 1% of that number (568) are proficient in communicating about AI, and if 1% of the strategic experts may choose to work for your business, then the human capital you may have to promote AI education, strategy and usability in your business, amounts to less than 6 employees, if you are lucky! The supply chain of talent is the ultimate rate limiting step. Education needs an avenue to amplify its reach, increase the pool of talent, globally. The new path is digital, it is digital learning. For corporations with global ambitions, sponsorship of education, digital learning and academic-industry partnerships, are sign posts on the road, to long term profitability.

Pg 21 I am an AI optimist. My article on AI (Agents: Where Artificial Intelligence Meets Natural Stupidity) is here http://dspace.mit.edu/handle/1721.1/41914 (2002) • Dr Shoumen Palit Austin Datta, Research Affiliate, MIT Auto-ID Labs, Department of Mechanical Engineering MIT • Please explore website http://autoid.mit.edu
Prescription for Profit

Dr. Shoumen Palit Austin Datta
Senior Member, MIT Auto-ID Labs
Research Affiliate, Department of Mechanical Engineering
Massachusetts Institute of Technology, Room 35-203, Cambridge, Massachusetts 02139, USA
Senior Scientist, Medical Interoperability Program, MDPnP Laboratory
Massachusetts General Hospital, Harvard Medical School
Partners MGH Research Building, 65 Landsdowne Street, Cambridge, MA 02139, USA
shoumen@mit.edu, sdatta8@mgh.harvard.edu

Letter from Academia

AI is being marketed as a panacea solution or a complex “black box” by the PR spin doctors. How to profit from AI in business applications is still unclear. The lack of understanding of knowledge representation, data structures and feature engineering, are a few of the core underlying problems, devoid of easy solutions. This short guide is a note on strategy with respect to the use of AI tool kits. What is necessary for rational use and integration of AI tools with business are humans.

Keywords. AI, Digital Twins, Rational Agents, Logic, Ontology, Knowledge Representation, Data, Analytics, Computational Rationality, Neural Networks, Cognition, Machine Learning, ERP, Syntax, Semantics, Synapse, Speech, Vision, Neurobiology, Innovation, Management.

1 Pragmatic use of Artificial Intelligence

Pragmatic use of Artificial Intelligence (AI), which can catalyze corporations to profit from applications of AI, is the ultimate goal for business and industry. Academia could help industry achieve this goal, albeit in part, the rational part.

The probability of bursting the public relations bubble, and the hype about the promises made on behalf of AI, are increasing. It may be reduced if industry and management better understood why the world must opt to lower its expectations of “intelligence” as an outcome from AI tools.

The suggestions in this letter does not distract from the rational possibilities of using the principles of AI in data analytics, decision support, and even, in automation.

During 1955-56, the term “AI” rather than “computational rationality” was used to describe a “new” and “emerging” field. The difference between the terms is a matter of states because “intelligence” is continuous (core attribute of many biological processes) while “computational rationality” is a discrete process. The latter is explained by the boundaries of limited rationality which systems generate, based on computable models.

The cognitive glue, necessary to bond discrete events to form the continuum, may be a cherished objective, but remains an illusion for science, and delusion for engineers, at this time. This brings to mind the pithy words of John Searle "brains cause minds" as
if to say that a mere collection of cells (neurons, glia) will lead to thought, action and consciousness. It is true, the brain is a collection of these cells, but does that suffice to serve as a platform, to extrapolate the brain to the scope of the human mind?

To illustrate the issue on page 7 (see https://dspace.mit.edu/handle/1721.1/108000), I refer to (https://link.springer.com/content/pdf/10.1007%2Fs11633-017-1093-8.pdf) a recent paper. It starts with states (each binary bit has two states, 0 and 1) and memory storage capacity. A human brain has \(1 \times 10^{14}\) neural cells (100 trillion synapses) with approximately \(2 \times 10^{15}\) states, equivalent to a storage capacity of 500 terabytes (assume 4.6 bits of information stored by each synapse). Hence, \(2 \times 10^{15}\) is the number of synaptic connections in a human brain. Is this, then, the capacity of the human mind?

In our mind, cognition allows us to read, write, create, and understand language, simple and complex. Our vision can distinguish topology of objects, colors, size, depth, shades. Our five senses, in combination, can respond to an array of input, to produce a vast (unknown) number, and variety, of output. If artificial neural nets claim to “capture” the brain and if we can scientifically describe this capture as “brain in a box” then, this network, if etched on a “neuromorphic chip” is the sum total of “intelligence” that we may rely on, for all our activities. By this rationale, “intelligence” is governed by the maximum number of states of our synapses. That number is about \(2 \times 10^{15}\) and that is, by this account, the total number of instances or the magnitude of combinations of our thoughts. Is this a true statement?

As Danko Nikolic points out, an English speaker’s vocabulary has about 15,000 words which consists of 5% adverbs, 20% adjectives, 20% verbs and 55% nouns (750, 3000, 3000, 8250 words in each of the four categories, respectively). From those numbers, we can calculate the number of all combinations, of sentences, of different lengths. For four word sentences, consisting of a noun, followed by a verb and ending with a noun plus an adjective, we obtain \(8250 \times 3000 \times 8250 \times 3000\) or about \(2 \times 10^{18}\) combinations. We have not even pondered about the semantic boundaries of the syntax in the four-word sentences. This number \(2 \times 10^{18}\) is already bigger than the limit that is posed by the total number of synapses in the brain \(2 \times 10^{15}\). By this reasoning, there isn’t enough memory in our brains to generate a different response even for sentences with 4 words!

At this stage, we have only considered “speech” and limited our expression to 4-word sentences. Limited by the storage capacity of the total number of synapses in our brain \(2 \times 10^{15}\) we will not be able to see, hear, taste or touch, among other things. Do we still wish to continue, and support marketing campaigns, suggesting that deep neural nets are equivalent to biological intelligence, which powers AI? Hence, is there intelligence in AI?

The fact that humans possess at least five senses, and do much more than what \(2 \times 10^{15}\) synapses may allow, is due to the fact that this number is an anatomical representation of the number of discrete connections. This is the structure of the organizational aspect of the anatomy and topology of the human brain. Structure is not the same as function in the same manner that anatomy (human skeleton) is not equivalent to physiology and physiological function (human organism), even though the skeleton (structure) is quintessential for physiology (function).

It is function that generates the amorphous quality of intelligence and makes humans intelligent. The numbers in the structure are discrete. The numbers matter, of course.
With 302 neurons, potential structural relationships in *Caenorhabditis elegans* (worms) may not qualify to provide intelligent functions or even pattern recognition.

The function of intelligence is best perceived as a *continuous* fabric, inextricably linked with data, rules, patterns, experiences, knowledge and learnings to inform or support decisions.

The almost unlimited number of connected continuity, the underpinning of intelligent human action, is a result of $2 \times 10^{15}$ synapses which are being formed, and re-formed, connected and disconnected, re-connected and re-configured, in an asynchronous, dynamic manner, in response to signals, perceived, received, in processing or being transmitted. Signals may originate from diverse sources (internal, external, autocrine, endocrine) or may be presented to sensory interfaces in a multitude of shapes or forms.

Continuity is not an attribute of a computable model. The term AI was less appropriate than “computational rationality” in 1956 and it is even less appropriate, today. But, we may use the term AI, for the sake of posterity, its magnetic image and public imprint.

The term intelligence is supposed to present a mental image relating evolution of words, objects, ideas, in terms of meaning and context. It is not a discrete, structural, one to one syntax, which can be translated. It is an *interpretation*, based on semantics, and by extension, logic, and ontology. The fact that intelligence may not be amenable to simple syntactic translation was demonstrated by the almost abject failure of the Russian to English translation during 1960s (prior to use of convolutional or recursive neural nets).

The “artificial” architecture of intelligence may have literary roots. Perhaps, a reference to *Leviathan* by Thomas Hobbes (1651) or similar, from that school of thought. Hobbes argued for “artificial animal” based on observation that the heart is a spring, nerves are strings and joints are wheels. Attempts to mimic birds and develop “artificial flight” did not lead to aviation. The right approach by the Wright Brothers was to view flight as a function of aerodynamics, which gave birth to the airline industry. Reality of flying, for human use, was not a reproduction of the fantasy of viewing birds in flight.

2 From Taylorisms to Terabytes

The movement from Taylorisms to terabytes needs AI, and its tools. Hence, AI, despite its limitations and a handicapped terminology, presents opportunities for companies to automate business processes. But, fantasy driven scenarios, about winning at GO or poker, may not suffice for integrating AI or ML applications, in the real world. While ERP implementations enhanced competitiveness, several companies also uncovered nightmares. The promised opportunity from ERP never came to fruition, for some. Do we have a sense of *déjà vu* with AI? The rain on the AI parade falls mostly on input data and the output/outcome. Unless reliably automated, the outcome requires people to do something with the information. Is it actionable? AI analytics cannot help if input data is noisy or corrupt. How do you know the data or the outcome is of poor quality?

AI and ML can augment performance. In case of AI (more than ERP) those changes create highly skilled tasks which require education, prudence and domain expertise, *from humans*. Businesses are forever in an elusive quest for “low hanging fruit” without gaining the wisdom from repeated failures. The pursuit of “low hanging fruits” require only low level skills. That *modus operandi* may not help, at all, to profit from AI.
Generating value from AI by recruiting more data scientists is an amorphous escape clause. Several domains converge under the umbrella of data science, which makes it impossible to ascribe the term data scientist to any one individual. Data science is a team sport. Bringing the talent together, and synthesizing the unpacked problems, are tasks that few companies can execute because companies do not have, or rarely employ, strategic cube-on-cube thinkers.

Companies do not even know, that they do not know, that they lack trans-disciplinary cross-pollinators. Companies and HR are unable to comprehend that they need people with broad spectrum of knowledge “cubes” and a matrix of experiences, unlike those that can fit in a box. “Thinking different” is not a principle that HR departments can practice. Hence, the clamor for data scientists but lack of jobs describing the need for out-of-the-box thinkers, followed by an absence of zeal, to pursue the road not taken.

Thinkers are pivotal to assist teams to dissect problems into components, to identify the confluence of domains, and underpinnings of potential solutions. Creative thinkers are key to assist the leaders to move the fulcrum and mentor the rank and file to frame the correct questions. Hiring and allowing cube-on-cube thinkers to form agile, case-dependent teams, staffed with vertical experts, across silos (network of business units), may be the first step to profitability, from advanced applications, which are fueled by convergence, such as, AI, analytics, robotics and nanotechnology.

Data science must start with data. Data must be acquired, processed and curated to serve the business needs. Hence, the critical demand for domain experts, and field knowledge providers, who must help identify the obvious, common, and uncommon “features” that businesses are seeking. Then, add non-obvious relationship analyses, and garnish with unconventional wisdom. To harvest the latter, perhaps crowd sourcing may be useful.

Organized data, using the principles of knowledge representation and application of logic and ontology, is a starting point, to construct computable models/structures of the domains of interest (agnostic of industry, vertical or horizontal). In the computational phase, we can use algorithms and tools from AI including ML, DL, ANN, CNN, RNN.

The trinity of out-of-the-box thinkers, who can connect the cubes, with field knowledge providers, and computational experts, is the “secret sauce” which must be continuously stirred, shaken, configured and re-configured, to blend the correct team, case by case, to profit from AI, and use the ability of AI, in problem solving. This approach and grasp of the extended fabric, is lacking in businesses and absent within corporate leadership.

The marketing hype, which is furiously polishing the chrome, on the AI engine, may help to explode the bubble and trigger a second AI winter. Global warming will be essential to thaw the AI ice age. But, before we boil the ocean, let us try to warm up to what may be necessary, the prerequisites, what is missing, how deep is the abyss and how education may bridge the chasm. Let us imagine, we have managed to fast forward to the spring of AI. Assume, AI in the tool kit is generating probabilistic output.

As pointed out by Jeanne Ross, an AI application indicates that a lead has a 95% chance of converting into a sale, while another has a 60% chance. Should we assume the salesperson knows what to do with that information?

ML applications may help lawyers identify appropriate legal precedents, help vendor management teams ensure compliance with contracts, assist financial institutions to gauge risk. These systems use ML to perform mundane tasks. Systems can learn to
develop spreadsheets and search databases for relevant information. But, in order to generate competitive advantage from ML (AI), we may need skilled humans to process the outcome. Hence, companies must redesign accountabilities, motivate employees to deploy ML tools, when they believe it may enhance outcomes. The educated workforce of the future must possess higher order skills, capable of consuming intelligence, and trigger actions to benefit, and hopefully, profit, from the deployment of AI tool kits.

Hence, these AI tools must be capable of use by general employees. The tools may be drag and drop interfaces representing abstractions. The employee may not need to deal with the computational complexities, programming principles and Boolean operators.

To use these abstracted tools and intuitive interfaces, the educated workforce, in future, must possess skills which catalyzes the consumption of intelligence, the outcome. The educated consumer is the best customer for the future of AI and to profit from AI tools.

This raises a critical issue concerning K-12 education and how learning must be adapted to deal with the imminent socio-economic disequilibrium. Education must address the changing face of the supply chain of talent as well as the ingredients which are necessary to future-proof workforce preparation and make the workforce future-ready.

The future is not about apocalyptic reduction of employment. It is about a refresh of skills, which must be updated and upgraded, for the humans in the loop, to play relevant roles. We need the AI tool kit to help reduce uncertainty, and better manage, volatility.

To achieve that goal, executives need to appreciate the principles of data analytics and AI. Leaders must support education, inculcate insight and remain eager to learn, before they leap to manage. Leaders must institute internal education and external learning liaisons, where thinkers are viewed as assets and not as cost centers. Leaders must stress on understanding how AI works rather than blindly purchasing “black box” solutions.

The digital world will still need to serve analog communities. AI may lead to profit, if allowed to offer reliable computational assistance to the workforce, customers, and the global ecosystem of consumers, seeking credible, rational, near real-time, and perhaps predictive, decision support.

All things considered, the path to profitability rests with the imagination and the vision of the executive management and their counterparts in academia, and government.

Corporate leaders must evolve in their leadership roles. Leaders must assume the risk of leadership. Leaders must engage to provide broader guidance, bring parties to the table (competitors) and advocate for interoperability of architectures, to enable digital connectivity. Without security, digital transformation could be annihilated. Without connectivity, without data from different systems and ecosystems, without knowledge of what is beyond the boundary, the ability of AI, analytics and tools such as blockchain applications, will be curtailed. The outcome will be less valuable, less actionable, less profitable.

Hence, leaders must champion digital transformation by leading, and inspiring global teams, and navigating businesses to lift many boats, not just their personal yachts.

Acknowledgments. 1) This “letter” also appears as APPENDIX 6 in the essay “03.AI” available from the MIT Library – see folder “CHAPTERS” https://dspace.mit.edu/handle/1721.1/106496. 2) The Fatal Flaw of AI Implementation by Jeanne Ross, MIT SMR (14 July 2017) served as a source of a few examples and the reference to enterprise resource planning.
THIS PAGE IS LEFT BLANK
Digital Enterprise X.0

A blog about building blocks

Shoumen Palit Austin Datta
Senior Member, MIT Auto-ID Labs and Research Affiliate, Department of Mechanical Engineering
Massachusetts Institute of Technology, Room 35-203, Cambridge, Massachusetts 02139, USA
shoumen@mit.edu • http://bit.ly/IOT-MIT

Senior Scientist, Medical Interoperability Program, MDPnP Labs
Massachusetts General Hospital, Harvard Medical School
Partners MGH Research Building, 65 Landsdowne Street, Cambridge, MA 02139, USA
sdatta8@mgh.harvard.edu • http://bit.ly/MIT-IOT
Do not borrow

Artificial Intelligence Project--RLR and MIT Computation Center
Memo 17--Programs With Common Sense

by

John McCarthy

SUMMARY

Interesting work is being done in programming computers to
solve problems which require a high degree of intelligence in
humans. Moreover, certain elementary verbal reasoning processes
so simple that they can be carried out by any non-feebie-minded
human have yet to be simulated by machine programs.

This paper will discuss programs to manipulate in a suita-
bble formal language (most likely a part of the predicate cal-
culus) common instrumental statements. The basic program will
draw immediate conclusions from a list of premises. These con-
cclusions will be either declarative or imperative sentences.
Then an imperative sentence is deduced the program takes a cor-
responding action. These actions may include printing sen-
tences, moving sentences on lists, and reinitiating the basic
deduction process on these lists.

Facilities will be provided for communication with humans
in the system via manual intervention and display devices con-
ected to the computer.

Fig 0: The fact that a program can find a solution in principle does not mean that the program contains
any of the mechanisms needed to find it in practice. (1956 ● MIT AI Lab Memo # 17 by John McCarthy)
PREFACE

Experts and non-experts have already commented that this blog is shoddy and second grade in style and content. It lacks several seminal issues indicative of the author’s lack of knowledge, depth and business experience. I concur. Therefore, please proceed with caution. The intent (please see appendix) was to offer suggestions about paths that may be different or diverges from conventional wisdom or what is in vogue. The disjointed digressions are an attempt to connect the “dots” and that process may be chaotic. This is not an instruction manual, recipe or a roadmap. It is a compass. It points in so many directions because of what may be possible, the yet unseen tangled networks and the looming digital quagmire.

---

 Aren’t we are all apprentices in a craft where no one ever becomes a master? from Ernest Hemingway

ABSTRACT

The approaching wave of digitalization may dwarf the immense economic consequences of what could have happened if we were to flood the Sahara Desert. In this essay, we may not even scratch the surface of convergence which is the quintessential foundation of the imminent transformational tsunami. Flooding the Sahara [1] is a metaphor. We present vast opportunities for business solutions in terms of development and sales of services under the banner of Digital Enterprise X.0 (DEX) and its potential for global diffusion. Modular services may offer delayed differentiation using “lego bricks” or building blocks aggregated to form platforms, which can rapidly configure trillions of outcomes for billions of customer choices, using variant configuration and dynamic composition. The combinatorial efficiency may be optimized by a “menu” to select choices triggered by customer preferences. Sense-aware bots may “drag and drop” or “plug and play” Agent-based entities at mobile edge devices (or interact in a mist or fog). It may execute functions (data analytics - predictive, prescriptive, diagnostic) at the edge or in the cloud, linked to objects, processes, decisions, patterns or things connected to the internet [2]. Hence, a convergence, of enterprise systems, with IoT, a conceptual digital design metaphor. Synthesis of each element, in sequence, offers a micro-revenue harvesting potential for each creator, developer, provider of each function in each micro-service. The trans-disciplinary convergence and systems integration, necessary to arrive at the outcome, is rewarded when the end customer pays, after the value of the service is realized or the outcome is delivered. Revenue collection “at the end of the tunnel” may require a channel master, brand integrity, security and interoperability between platforms (local and global, private and public, open and proprietary). Monetization demands tools and technologies to document every event in an irrefutable digital ledger. The latter is essential for billing purposes, if the partners in the ecosystem may wish to claim their fraction of nano-earnings from each micro-payment.
BACKGROUND

Digital transformation is still embryonic and evolving. The process is punctuated by islands of hype and hyperbole[^3] which makes it difficult to synthesize the fabric, in order to capture the convergence and continuum, characteristic of most enterprises. The alternative approach is to discuss discrete events, such as, tools, technologies and applications, that is, business as usual in terms of products or services. Compartmentalization of business solutions further fragments the collective use of data and diminishes the potential for profitability which may ensue from data analytics, if the interdependencies between silos, were analyzed. Static data storage may not be beneficial in this age of data driven decisions.

We are drowning in data yet suffering from poverty of information and knowledge. The classical chasm between information technology (IT) and operational technology (OT) is a reminder that the enterprise was not designed to interface or analyze external/internal data to boost efficiency, reduce waste and take advantage of risk pooling. Systems integration to promote collective “intelligence” by dissolving the archaic IT/OT boundary is not a core competency for most enterprise. Thus, supporting businesses are still creating discrete products and pushing independent domain specific business services touting their use and efficacy in a manner which resembles a hammer looking for a nail. Not every product or service is a solution and certainly one solution may not fit all. In some cases, the solution may not be a solution at all[^4] but a public relations gimmick to impress the uninformed through game shows or board games and marketing chutzpah. The elusive quest for wisdom about “disruptive”[^5] digital transformation from dubious management gurus[^6] with a widely known predilection for self-promotion[^7] is tantamount to dissemination of dead ideas[^8] from live people[^9].

WHAT IS THE EXPECTED OUTCOME?

Pulling the digital thread through the eye of the needle relies on the assumption that every needle has an eye. Once threaded, we must be able to display a measurable outcome, that is, the needle must move, hopefully, in a manner which may be sustainable and supports an ethos of ethical profitability.

Neither the hand-waving about the prospects for digital transformation nor the glib PR of estimated economic growth due to digitalization[^10] may be realized, even partially, unless we focus on building the foundation, at the systems level. We have to make sure that there are “new eyes” in every system, if the process of digital threading is expected to succeed. Convergence of digital threads are quintessential to the systems integration process, which in turn, is key to linking the systems of systems, at the heart of the enterprise, both analog and digital. Solutions are likely to incorporate elements, we have seen, or what already may exist. Innovation germinates from imagination, and thinking, what no one else has thought[^11], by visualizing creative juxtaposition of blocks or bricks, to create new designs or patterns, and to implement them for systemic improvements. A catalyst for entrepreneurial economic growth is not so much to obtain novel facts or delve into research, but to uncover unique ways of thinking about existing facts, data, tools and technologies, percolating them in “various speeds” in a matrix of unusual or imaginative combinations. Frequent departures from conventional wisdom, creative destruction of existing rules and application of the principle of non-obvious relationship[^12] are vitamins for progress.
Imagine standing in front of a machine on the floor of a New York Times printing facility (Figure 1). You take out your smartphone and click on the app marked DEX. What you expect is the software avatar of the hardware, i.e., machine, in front of you (aisle 8A in Figure 1). You see a blinking icon in a network diagram displaying other connected machines (8A, 8B, 7A in Figure 1). You tap the blinking icon to focus on the digital duplicate of 8A, that is, the Digital Twin [13] of the paper-cutting and sorting machine (8A in Figure 1). You tap on the parts icon to drill down to the details, for example, performance (blade speed), wear and tear (MTBF of blade), lubrication of the hydraulic cutting arm (oil pressure), temperature inside the press which holds down the paper stack (to optimize performance of the blade) and conveyor belt surface aberrations (data from mounted camera under the belt analyzed using optical recognition).

This scenario is part of an expected outcome relevant to, for example, efficiency. We need metrics to measure the value of key performance indicators (KPI), for example, down-time, based on the data (outlined in above paragraph as blade speed, MTBF of blade, oil pressure, temperature, belt health) and data analytics, to understand the meaning of the data, in terms of information, related to efficiency.

The data thus converted into information can update and inform the status of the machine (8A in Fig 1). Individual machine data, when combined, can further inform the KPIs for the entire machine floor or the entire operational system. Instead of one machine, we can acquire data and analytics from a swarm of machines and rather than understanding the status of one machine, we optimize the entire operation.

The transaction cost [14] to obtain each piece of data can be aggregated and the monetary value of the information is deduced. To perform comparisons, we model and run “what if” simulations to compare the operation with and without data, with respect to the sequence of events in the cascade or the final outcome (may be affected) of this operation, due to an unplanned failure of a machine (preventable).
If two machines shut-down due to over-heating without an alert (absence of temperature sensor data) what will be the impact on the operation? Will decreasing the output affect distribution or logistics or the supply chain? Can it impact quality of service for delivery and customer satisfaction? Will it damage brand image? Will the ripple effect influence the ecosystem of partners? Can it reduce EVA (economic value added) parameters? Each event in this multi-pronged process may have a value or can be assigned a weight in addition to transaction cost (positive or negative). Comparing the simulated transaction cost between [a] acquisition of data/information to inform and alert vs [b] operation running without these tools, may provide the rationale and justification to invest in digital transformation or not.

HOW TO PREPARE TO ACHIEVE THE EXPECTED OUTCOME?

This discussion must begin by reiterating the obvious - no one solution or architecture or strategy may suffice to address the needs of Digital Enterprise X.0 (DEX). Complete disregard for archaic or monolithic approaches (for example, IBM DB2, SAP R3) may be a bit unwise. Salivating for new tools (for example, Google Spanner, IBM Watson) may be replete with copious shortcomings.

The disturbing efforts by some behemoths claiming “brain in a box” by 2020 can only amplify hype through yellow journalism. AI apocalypse is a marketing attempt devoid of scientific foundation. For example, engine wants to reduce NOx emissions without increasing carbon dioxide. The outcome depends on a plethora of parameters which must be optimized, such as, controlling multiple points of injection of the fuel, how to determine efficiency by balancing fuel composition, temperature, humidity data from hygrometer, etc. It is difficult for humans to monitor, adapt and regulate NOx emissions to the lowest level because the optimization requires too many experts to learn, observe, agree and predict the thresholds and the range, before each equation can be solved and the results applied to the process. A portfolio of algorithms and solvers using the principles of machine learning (ML) may be used. Is this AI or artificial intelligence at work? This is ML at work using contextual higher order reasoning (CHOR) in real-time. Publicists call it AI. This is the crux of the discontent, over-selling of what is intelligence.

A portfolio of solvers is not equivalent to intelligence. What is intelligence if we are discussing actions at the level of a worm or an ant? Solvers rely on data but forced fitting data to models is fitting noise. Processed data may be denuded of its informational content just as processed food can lose its nutritional value. Just because the data was smoothed does not mean that the data is smooth. Hence, curation of data through a data “forge” and then harvesting the data based on feature engineering is a prelude to composing betters models. But, models per se are fraught with problems. Most models are incapacitated in the face of dynamic change in data or fails to be useful if the rules or patterns in the data may change, indicating the need for feature engineering re-design.

Automatic code generation based on models continues to be an useful tool for equation based models (EBM) but may lack the dynamic attributes increasingly important due to the volatility of digital transformation. Modularity, dynamic composition and mathematical models to drive the best outcome in face of uncertainty are the general principles of rational Agents and multi-Agent systems. The latter is essential for enterprise architecture as a key component of AI and will enable us to profit from use of AI.
Hence, we propose introducing Agent based models (ABM) with EBM. Meta-modeling principles from model driven engineering [23] could be useful in digital transformation if the meta approach may be better tuned to accommodate dynamic changes expected in the future. Knowledge representation and symbolics, for example, in feature engineering for data analytics, will be avoided in this blog. Rather we shall focus on interoperability and connectivity between platforms and applications. Convergence of various cyberphysical systems (CPS) or CPS SoS (system of systems) with time sensitive networking will present even more complex challenges [24] where interoperability between platforms may solicit more attention. However, tuning the engine vs polishing the chrome, may continue to haunt us. It is difficult to find the balance, coalesce teams, adapt and/or adopt with respect to state, context and dynamic composability needs, at present and in the future. Examples of progress may be found in GE Predix, Siemens MindSphere, Hitachi Lumada and IBM Bluemix, in addition to other less known IoT platforms.

BUILDING BLOCKS OF THE DIGITAL ENTERPRISE X.0 (DEX) – A BRICK BY BRICK ACCOUNT OF DEX

The preparation necessary to substantiate the introductory level of Digital Enterprise X.0 (DEX) outlined under “expected outcome” in the preceding section, offers the grand potential for new lines of business.

The evidence from GE’s Predix suggests that GE may use this platform for its own business units and sell the platform as a service. This is the new business. The scope and reach may be far greater than what appears to be on offer from any vendor, at this time.

This solutions space is analogous to fractals [25] or Minsky’s cube-on-cube paradigm [26] rather than point or line. It will evolve and change as the semantic context of DEX adapts to diverse domains, from manufacturing to medicine to mobility, serving different needs in Memphis, Maputo and Mumbai.

The vision, investment and determination necessary to create this new line of business may be uncommon. Few companies possess a portfolio of operations where DEX systems may be developed, evaluated and implemented. The customer and supplier base of these companies may be the first users. Hence, expect a network boost from creating the business of DEX.

Not all components of DEX may be created de novo. DEX needs to tap into applications (start-ups) and potential advances (academia). Commodity tools (sensors) and services (cloud, storage) are likely to be procured. Foresight is crucial to converge and orchestrate near-term needs, profitability and long-term sustainability of the vision.

If the expression “low hanging fruit” creeps in early discussions, that is an ominous sign and the plight of DEX may be compromised. It is abundantly clear that incremental gains must be a part of this journey but such gains may not be “low hanging” from the onset of this process.

Slapping a sensor on a machine to harvest vibration signals from a pump is not indicative of digital transformation. Similar mistakes [27] were made at the dawn of the 20th century with electrification [28]. It was repeated [29] two decades ago when retailers thought affixing a RFID tag offered transparency and the digital supply chain was a fait accompli.
Convergence of the need for data with acquisition tools to harvest data (for example, from RFID, sensor) is step one in a systems approach. Subjecting the data to appropriate analytical tools may help to make “sense” of the data in step two. Extracting actionable information or knowledge makes the data valuable (sense and response) to the supply chain network or partners. The architecture and systems engineering principles must converge with the business process and the context of the data, with respect to the operation, where the information may be actionable.

Latent in the “preparation” is the need for organizational vision which must sufficiently percolate within senior leadership. They must be imbued with the spirit to grasp beyond their reach and embrace innovation uncertainty with a pledge to proceed beyond proof of concept to large scale implementation. The rationale for the latter is based on the principle in Young’s double slit experiment conducted by Akira Tonomura [30] in HCRL, Kokubunji [31]. There must be a consensus to move mountains and remain cognizant of the fact that there are mountains beyond mountains [32]. The action plan may commence by acknowledging convergence is as important as connectivity between digital threads.

How to connect the entities to drive the “outcome” is a key question. This is a convergence issue which manifests as systems integration, has many moveable parts which may rapidly change. Components may become relics within the time span of implementation due to short half-lives. Functions often may not scale and non-linearity is the norm. Response or lag time in proof of concept (PoC) may not extrapolate in large scale implementations due to non-deterministic real-world variabilities which one may not capture in deterministic models used for PoC. These known unknowns behaves poorly at the systems level (Figure 3). Early attempts to create systems to generate tangible outcomes [33] resulted in the “Store of the Future” in 2002 [34]. The PoC may be best described as a successful failure.

---

Fig 2: Digital Twin of a pump showing virtual (VR) sensor. The context of a pump within the operation (water purification, air recirculation, oil flow) is key to the value of data vs information. Including sensors to identify pump metrics is perhaps an incomplete approach, incongruent with the convergent vision of operational efficiency in the context of the customer (pump is one asset in the operation of the plant).
The complexity of the components, which are systems within systems, and integration with data and analytics, remains challenging. It may have inspired Amazon GO, the no-checkout convenience store of 2017. GO is wobbly, incomplete and still a PoC, on steroids. But, Amazon’s “Store of the Future” is still not yet good to go [35].

Fig 3: Dr David Brock attaches RFID tags to objects (1999), at the MIT Auto ID Center, which created part of the vision (The Networked Physical World) and further catalyzed digital transformation and DEX [36].

WHICH STRUCTURAL ELEMENTS AND COMPONENTS ARE CRUCIAL FOR CONNECTIVITY?

Over the past half century we have made progress in connecting various software modules with a host of different interfaces. In a simplistic form, we would like to connect processes and data in a manner that it can use the outcome (decisions) to inform and improve the process (classic feedback loop). Process and data must stay decoupled but encapsulating process and data in software “wrappers” may generate myriad of different forms. Perhaps the most useful are functional modular representations or “lego” blocks, which can be used for complex combinatorial creations, to develop layers of successively higher order functions, subject to rapid assembly, dis-assembly and dynamic re-assembly, if necessary.

The Agent Network Architecture proposed by Patti Maes [37] was based on a 1989 proposal [38]. Some of the ideas [39] are still germinating at the hands of Google, Apple, Microsoft and others [40]. The potential revolution cryptic within the vision of Agent Network Architecture (ANA) was recognized by experts [41] and gained academic momentum [42]. However, its use in industry [43] may still be sluggish [44], at best.

Agents based architecture provides the tools critical for interface components. These tools are key to facilitate connectivity and interoperability between the blocks or modules or units of Agents. Thirty years later, Agents are still struggling to cross the bridge to the business world. There weren’t any “low hanging fruits” then and there may not be any “low hanging fruits” now. The myopia prevalent in the business world leads to the stampede for “low hanging fruits” but without the understanding that “low hanging fruits” require low level skills. The rational Agent based approach is the underpinning of AI.
Development, deployment and diffusion of ANA and other Agent architectures are pregnant with solutions [45]. Skills necessary to design, evaluate and implement these architectures are neither trivial nor automated. The “winter” of AI [46], the resurgence of AI hype [47] and the lack of CS graduates [48] are hampering the arduous tasks which must be undertaken if the so-called fourth industrial revolution, characterized by ubiquitous connectivity, is expected to yield actual profit. It appears that these discussions are increasingly part of corporate marketing budgets rather than R&D or education investments. All things digital, IoT and AI, in particular, is prime tabloid fodder and electrifies cocktail banter at the gatherings of glitterati in a highly secure winter chalet, feeding the frenzy of paparazzi [49]. AI is rather useful if we may think less of human level intelligence and more of higher order reasoning.

![Fig 4: BS (L) & PhD (R) degrees in CS (computer science) from US universities [50]. Too few in numbers to supply the workforce necessary to catalyze convergence (IoT, AI, ML, Agents, Robotics, Data, AR, VR).](image)

Connectivity by powerpoint (Fig 5, left) is witnessing exponential growth [51]. But dissecting these “boxes” and reducing each to its basal functional granularity may reveal multiple layers in each “box” and each with its own multi-dimensional matrix (but not always) representing complex interactions, dependencies and potential for interactions between building blocks and function modules (Fig 5, right).

In Agent based architecture, each function is encapsulated as an Agent with interfaces which can connect to other Agents encapsulating other functions (process, decision, location, data feed, sensors, analytical engine, edge parameter, cloud store). Each unit or block or box may comprise of multiple Agents in accordance with the “layers” of functions that each represents. Each Agent or a collection of Agents (referred to as an Agency) within a “box” may serve as a module capable of executing one function. It can be used independently (1 Agent) or as a collection of function-specific Agents (Agency).

Thus, Agents and Agencies (representing these “boxes” at the level of desired granularity) can be mixed and matched without the need to re-tool or re-structure or re-program the entire function or operation. In equation based models (EBM), the interrelationships and dependencies are refractory to dynamic adaptability because they are hard-coded. Hence, addition or deletion of functions (boxes) alters the equation and renders it useless or incompatible unless re-programmed or re-structured or re-designed.
In contrast, in Agent based models (ABM), dynamic composability is normal since each Agent or Agency can interact with any other Agent or Agency as long as the Agent interface enables the communication. The Agent interface can control the connectivity, because it can be configured by design to regulate the nature of data sharing process, by differentiating between sharing data vs information. The Agent acts as a “wrapper” within which the decoupled modules of process definition (feature) and data co-exists in separate sub-modules. They can function as 1 entity for data/information sharing or differentially parsed based on business or security logic. If a bot pings your location, your PSA (personal security agent) may trigger selective permeability (Figure 8). The Agent shares information “not in the office” but protects your actual GPS location data because of your selective privacy settings in the Agent interface layer [52].

![Diagram of Agent based models](image)

Fig 5: Connectivity by design require architectural modularity and interoperability between the modules. Elements are available from object-oriented analysis and design principles [53] and related advances [54]

Is this *modus operandi* applicable to all enterprise architectures in order to align businesses with DEX? Unlikely. As previously stated, digital transformation must be a value-driven process guided by metrics including transaction cost analysis. But, most businesses lack granular transaction cost information. The economy, as a whole, still considers Agent based models [55] as an academic [56] storm in a tea cup [57].

In preparing to augment connectivity, businesses must undertake an analysis of the structural elements in their existing software architecture, especially the interfaces (API), which influences system behavior and collaboration between elements (within the existing architecture framework).

Enterprise architecture [58] was not created for digital transformation or DEX. It is critical to understand which structural elements can be or should be changed to Agent based architecture (ABM) to augment connectivity between elements. Which elements are likely to provide the alignment necessary to adapt between static, dynamic and real-time events in the age of 5G, GIS, AI, ML, M2M, IoT and industrial IoT?
This is a powerful analysis and should be documented as a template for business process engineering or re-design (BPR). The template may be a key learning and form a part of the service provided through the new line of DEX business, as a part of the business solutions arm of the corporation (how GE created the new business of GE Digital as a part of the GE conglomerate). DEX will offer new services to move the rest of the business world (SME) on the road to digital transformation through DEX implementations.

Interfaces to and from data sources and sinks, as well as analytics at the core and the edge, are increasingly commonplace in domains as diverse as finance and fashion. Interoperability between elements and interfaces (both internal and external) shall determine system behavior and interaction between ecosystems or platforms. Assigning a priority and value to interfaces, with respect to stability and security, may reduce or balance the risk of innovation. Risk is implicit and may be omnipresent in transformation from a product-based ERP state of mind to a service driven economy, which is “always-on” and mobile, digital and connected, fragmented yet linked to a plethora of partners, regulators, influencers, customers and constantly bombarded by digital environments and ecosystems (Figure 6).

<table>
<thead>
<tr>
<th>IT &amp; Infrastructure</th>
<th>Google</th>
<th>Amazon</th>
<th>Facebook</th>
<th>Apple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial intelligence</td>
<td>Google Assistant</td>
<td>Amazon Alexa</td>
<td>Facebook</td>
<td>Siri</td>
</tr>
<tr>
<td>Hardware Devices</td>
<td>Google Nexus</td>
<td>Amazon Fire TV</td>
<td>Facebook Home</td>
<td>Apple Watch</td>
</tr>
<tr>
<td>Communication &amp; Messaging</td>
<td>Google Hangouts</td>
<td>Amazon Kindle</td>
<td>Facebook Messenger</td>
<td>Siri</td>
</tr>
<tr>
<td>Digital Media &amp; Entertainment</td>
<td>Google Play Store</td>
<td>Amazon Music</td>
<td>Facebook Gaming</td>
<td>iTunes</td>
</tr>
<tr>
<td>Connected Car &amp; e-Mobility</td>
<td>Google Play Pass</td>
<td>Amazon Echo</td>
<td>Facebook Integration</td>
<td>CorPay</td>
</tr>
<tr>
<td>E-Commerce &amp; Retail</td>
<td>Google Shopping</td>
<td>Amazon Payments</td>
<td>Facebook Buy-Button</td>
<td>Apple Pay</td>
</tr>
<tr>
<td>FinTech &amp; Retail</td>
<td>Google Pay</td>
<td>Amazon Associates</td>
<td>Facebook Business</td>
<td>Apple Maps</td>
</tr>
<tr>
<td>Navigation &amp; Location services</td>
<td>Google Maps</td>
<td>Amazon Maps API</td>
<td>Facebook Business</td>
<td>Apple Search Ads</td>
</tr>
<tr>
<td>Advertising</td>
<td>Google AdExchange</td>
<td>Amazon Advertising</td>
<td>Facebook Audience Network</td>
<td>Apple Search Ads</td>
</tr>
</tbody>
</table>

Fig 6: The global digital environment is faced with and reacts to these captivating influencers. Can B2B and B2C interactions adapt to and adopt the de facto standards of communication and connectivity?

Preparing to transform the architecture to enable DEX is just one part of progress. It may be useful to use a “bill of materials” (BOM) approach in evaluating key expected outcomes relevant to the operation (for example, manufacturing, water purification, retail, commodities) or business services (logistics, supply chain, finance, business solutions). What do you wish to change or improve or make more efficient or reduce waste or increase productivity? Which nodes, if improved, may lead to decreased uncertainty, better margins, less waste, higher quality of service and will aid greater brand recognition?
Once these targets are identified, that is, each outcome you expect to influence profitability, the task begins to create process maps of dependencies and network of layers which may have an impact on the outcome. For each dependency and/or sub-layer and/or sub-node, it is vital to map its data/analytics flows/engines and the metrics associated with the input/output of data/decisions to measure and determine the performance of that dependency/sub-layer/sub-node with respect to a part or the entire operation, that is, the network of layers/dependencies/nodes, each in its appropriate semantic context.

This approach may benefit from principles of business dynamics using structured analysis in terms of systems engineering thinking and convergence with system dynamics [59]. Software tools (SyntheSim, iThink, Powersim) may help in creating models [60] and may make the tasks less onerous. Classical SD (system dynamics) may still be relevant in the age of DEX. But, the practitioner must be creative and innovative in applying SD to model development for DEX purposes. Blind use of SD as a recipe may be incongruous in view of the EBM infrastructure of classical SD versus the need to move to a digital by design strategy. The latter is volatile by nature and data drivers are often subject to network effects.

In addition, models must be able to simulate adaptability to dynamic situations which oscillate between closed-loop/open-loop with variable feedback control loops, if any, and may be subjected to networks of inter-dependent/inter-related, asynchronous, non-linear and decentralized dynamics. Perhaps we need to determine [1] which nodes, sub-nodes, instances, can perform as modular structural elements in Agents based models (ABM), [2] the interface with legacy systems and [3] dynamic co-existence between ERP/EBM systems versus ABM.

Fig 7: Use of SD (system dynamics) modeling [58] to understand dependencies and networks which may influence outcomes. SD is in need of creativity to adapt to DEX, the digital enterprise, by including, for example, data flow, information streams, analytical engines, dynamic composition, risk and security.
Connectivity between structural elements in the architecture of business solutions will continue to pose challenges due to multiplicity of legacy systems built on different principles. Continuation of mainframe servers may serve as an example and reminder of the dead weight of old technology that still co-exists. How has the processing of payroll changed due to the digital onslaught? Externally, the change is the medium of delivery - paper checks vs electronic transfers. The processes for payroll including payment cycles and deductions may not have changed significantly since the first GE data processor in 1959 [61].

Connectivity between structural elements, old and new, must be able to respond to and interface with the network. Connectivity between layers, domains and external systems becomes more crucial in the age of DEX. The drum roll about connectivity has reached a fever pitch in the context of IoT and IIoT. The tools and connectivity between tools and interfaces between software and hardware must continue to evolve and are subjected to rapid replacement/substitution/upgrade based on customer demand. If the customer wants to know the price of gasoline in 3 different zip codes, the app or service provider must connect the data streams, design the visualization appropriately and work within the telecom standards, protocols and latency boundaries to deliver the user experience capable of capturing a critical mass of repeat customers. The important principle here is interoperability between systems.

![Diagram of architectural components](image)

Fig 8: Architectural components must evolve and serve syntactic, semantic and pragmatic demands [62]. Selective permeability (cartoon, right) may be necessary for data security and privacy between layers.

The nature and reach of connectivity, and its interactions with the ecosystem, is so crucial for the final customer-facing outcome that the new line of business solutions and DEX leaders cannot be constrained by lack of interoperability, between domains, including third party API and systems on-demand. Evolving open platforms or platforms that are partially open (for example, GE’s Predix) are indicative of a future trend. Interoperability between platforms and systems are driven by interfaces (API). Agent based interfaces are interoperable by design and may be selectively permeable for security. EBM products using specific software languages or closed data dictionaries or specifically designed to restrict semantic interoperability (for example, Epic electronic health record systems) are road blocks on the path to global DEX. The digital tsunami may annihilate closed approaches, especially in the emerging markets.
Connectivity between the (hardware) physical machine in aisle 8A (Figure 1) and its digital (software) representation is a part of the “liquid” interface of DEX as outlined in “expected outcome” segment of this article. If every piece of hardware - that is, every machine manufactured by a different equipment factory - were to use proprietary interfaces for its data port (for example, data from sensors) then interfaces for data analysis may be incompatible. Software and hardware are likely to be poles apart in aligning their product interfaces for data communications unless they belong to the same club [63].

An Agent model of the machine (hardware) may be interoperable with a number of Agent-based software data and analysis tools. The Agent model of the machine is not one Agent but an Agency made up of many Agents representing each discrete component in the machine (parts, sensors, belts, ball bearings, gears). The software representation (this Agent model) of the hardware is the Digital Twin concept [13]. Digital Twins are not restricted to machines but can represent humans. For example, a human may be uniquely identified on the smartphone screen, of a RN or MD in a hospital who can view and monitor real time data about the patient’s heart rate, oxygen saturation, end tidal carbon dioxide and blood glucose level from sensor data or diagnostic devices [64].

Without the type of connectivity and interoperability described above, the “expected outcome” with reference to Figure 1 may be difficult to achieve. The modularity of the Agent infrastructure - bricks and blocks - can be combined to create components in manufacturing plants, bridges, turbines, airplanes, automobiles, EKG machines, pulse oximeters, mammography, sensors, oil pipelines, water pumps, satellites, violins and agricultural machinery. Access to digital object repositories [65] may enable a “drag and drop” tool to create any machine or object or tool as a duplicate or Digital Twin [13]. Activating the software to connect the data from the object or machine with the Digital Twin or avatar calls for a plethora of other interfaces and telco components which must “discover” [66] the hardware (machine), “discover” the sensor data stream from that specific machine (in a shop floor of thousands of machines) and connect it to the specific software model on the smartphone or tablet of the user. The pivotal role of “discovery” and the tools for automatic identification and secure communication of data must converge to fuel the digital convergence, in order to accomplish the “expected outcome” for DEX.

The hype about trillions of objects connected to the internet (in IoT and industrial IoT) is oblivious of the fundamentals [1] that a connection to the internet is not equivalent to a function and [2] that IoT is not a solution or a tool or technology but rather a design metaphor. Functional value of connected objects depends on discovery [67] of the appropriate objects in context of the operation and relation to the data streams. Taken together, it may create an ad hoc dynamic composition of objects which offer value when constituted as a function and the ensuing network of things [68] provides an outcome. An analogy is the ad hoc creation of a mesh network [69] which also highlights the role of standards and interfaces. Connectivity of objects offers limited value unless the connectivity is in the context of an operation. For example, the efficiency of pumps is important to the manufacturer but the customer who buys the pump to execute an operation prefers information on the whole operation in terms of all the key performance indicators, not just the motors or the pumps [70]. The outcome economy is less interested in tools and more interested in the final outcome. DEX is also interested in performance and outcome.
Components of connectivity have been discussed *ad nauseam* in millions of technical and business publications but a few other issues deserves to be mentioned.

Telecommunication standards are immensely critical in bridging domains. It must be flexible to switch between protocols as IP applications (software) move from closed to open to hostile environments (PLC in boiler room of a plant) and interacts with multitude of objects, data and processes using diverse modes of communication. It is the central theme of application dependent networks which posits that applications will trigger how the network will respond because the application is the focus or the master (it is the monetization node) while the telecom network is a medium (slave) for delivery. The network must respond to the needs of the application, by switching to appropriate fixed or wireless or hybrid technologies, for example, multi-protocol label switching (MPLS) network.

It may be obvious that the application running on the Digital Twin of the machine in aisle 8A (Figure 1) will be deemed inefficient if the sensor data from the machine failed to populate the Digital Twin sensor database with the pressure or temperature data. One reason for such inefficiency may be the nature of the telecommunication signal on a shop floor densely packed with machines spewing exabytes of data. If the communication link suffers from interference (WiFi or Bluetooth or VPN or controller area network, LPWA) or transmits null data signals (occurs with RF data in a dense data environment or if reflected by metal in the infrastructure) then the Digital Twin application fails to provide the desired outcome.

Connectivity in the age of DEX, therefore, requires new thinking in design, construction and utilization of telecommunications networks to prioritize network applications and service demands. DEX expects significant application-efficiency gains for networks to ensure robust connectivity adhering to the QoS with respect to latency, bandwidth and data rates.

To respond to these targets, over the past few years application triggered systems (for example, ACI from Cisco, ADN from Huawei) are trying to meet future application demands, especially for 5G, for example, by decoupling the services plane. Applications do not need to know the underlying network complexity, which must be navigated, in order to deliver the service – the application.

Not everything is about high data volume, broadband and zero latency. There are a growing group of functions in the IoT/IIoT domain where proprietary low-power wide-area (LPWA) networks are providing services for very low bandwidth, very low data, very low power and very low cost devices.

In preparing for the digital transformation tsunami, it goes without saying that reliability of connectivity will be paramount in a DEX environment because we are increasingly reliant on data driven decision support systems. The medium of telecommunications is the principal conduit for acquisition of data. If this medium shows “cracks” and the data falls through these “cracks” then the analytical engines will view data streams or data flows with missing data sets. The analysis by the analytical engines will be based on data presented to the algorithms. The outcome or prediction will be flawed to the extent of the significance and context of the missing data (machine shop vs pediatric cardiology ICU in a hospital).
If the flawed data is for a MTBF (mean time between failures) value then the outcome may be an unanticipated machine downtime due to fractured blade or rotor or spindle. The flaw in data acquisition may be due to network traffic congestion, inconsistencies or interference from other signals or time [71] spoofing (systems security breach). If the missing data pockets were for heart rate from a pediatric patient in the ICU, with cardiac arrhythmia, then the outcome may be fatal and a funeral, bit too soon.

Fig 9: Niche IoT applications use low-power wide-area (LPWA) networks using proprietary products [72]

The complexity and variety of network evolution follows the new uses of networks for different types of applications which may be created in remote corners of the world and/or crowd sourced. Traditional voice communication relied on hierarchical network architecture and follows a Poisson distribution model because phone calls are likely to remain at a constant frequency and duration, over time. But internet (IP) communications follow power-law distribution model [73] where most users connect via central nodes. Machine-to-machine (M2M) communication, which may be in progress in the scenario depicted in Figure 1, may be comprised of many different cases, each with a different need.

For example, autonomous vehicles and robotic surgery both demand near-zero latency, extreme bandwidth and ultra-high reliability. The environment for a vehicle rapidly changes as it moves on a crowded street lined with tall buildings or inside a tunnel or a bridge or climbs the Twin Peaks. The environment for robotic surgery may be an urban hospital from where the surgeon is “performing” a laparoscopic cholecystectomy (gall bladder surgery) on a patient in Petit Bois Island, Mississippi.
These scenarios emphasize that connectivity demands are very diverse. To meet these demands we must be able to slice, dice and splice dynamic programmable network-agnostic networks on top of fiber, copper, wireless, LTE and future 5G, using open source platforms and protocols, for example, software defined networking (SDN) and network function virtualization (NFV). Network flexibility is vital. It is enhanced by dividing the network and using open source software, as far as appropriate, to manage the decoupled data plane (traffic forwarding), control plane (devices) and services plane (applications). If a connected vehicle needs software download or must upload sensor data, the customer expects the outcome to be download/upload. In the outcome economy, the customer will not care if it is a wired or wireless connection, SDN or NFV. Implementation must offer complete connectivity for the customer.

For the telecommunication industry the concept of monetizing the “pipe” is increasingly irrelevant because the revenue is realized only when billions of end customers access or use the services. It is hard to differentiate “pipes” which must adhere to standards and protocols that are globally accepted. The value-added portfolio for telco may be in real-time data analytics, which is not their core competency. Hence, it is important for progressive telecoms and OEMs to create partnerships across a diverse cross-section of other component providers in order to multiply their sources of partial revenue (micro-payment, nano pay-per-use). The service will only earn revenue when delivered to the end user.

Sharing fractional micro-payments may require new tools for micro-event billing segmented by micro-sequences which may include micro-services and require seamless convergence to deliver the end service (outcome). The modularity of micro-services and use of software-oriented architecture may be better suited to Agent based architectures. SOA may enable harmonious co-existence of Agent-based models (ABM) and equation-based models (EBM) to optimize dynamic composition when serving the on-demand service request (application triggered). Services in each layer may have a different provider. Each provider in the ecosystem will claim a fraction of the end-revenue which is earned only when the service is delivered at a satisfactory quality of service level. A potential scheme for this approach was suggested and if analyzed, it resembles, at least, in principle, a digital ledger or blockchain.

Fig 10: Who will build the blocks? New lines of businesses may emerge in the domain of DEX blocks which are not limited to enterprise architecture could include cybersecurity and Digital Twins. Both instances are influenced by AI and the rational Agent approach may the common building block.
Fig 11: Digital Twins are within our reach? Cartoon from article by Aaron Parrott and Lane Warshaw, Deloitte [82]
WHY INTEROPERABILITY BETWEEN PLATFORMS IS SYNONYMOUS WITH PERFORMANCE AND PROFIT?

A challenge for open platforms providers is that open is synonymous with free in the digital economy. A quandary for open platform providers is the lack of monetization systems integration, that is, seamless convergence between micro-billing tools, digital ledger of micro-service delivery confirmation and time guarantees for time sensitive networks or real-time events. The ecosystem is strained when seamless functionality is compromised due to lack of interoperability between services or platform (assuming that connectivity is not punctured and the data flow is agnostic of the protocols and the telecom networks).

The sketch (Figure 12) outlines the diverse domains that must interact through the mobile platform to provide the service you demand (the expected outcome). The plethora of service providers include hardware and software vendors, telecom, security systems, asset owners (airline, car, garage, hotel, restaurant, wearables) and other public/private agencies (map, traffic, news, weather). Each provider wants a share of the revenue. The customer only pays when the end-to-end connectivity offers value.

![Figure 12: Your mobile device](image)

The consumer centric IoT-esque illustration in Figure 12 contains the essential structural elements also common in B2B or intra-enterprise scenarios. Convergence of Agents-based micro-services, ABM, EBM, data, connectivity, analytical engines and decision support “sense & response” systems augmented with security and higher order reasoning tools (ML, DL, AI) are at the core of Digital Enterprise X.0 [83].

Modular units encapsulating core functions, processes or data holders, which can be sourced from a common services (SaaS) repository (across multiple verticals or business units) may offer benefits in terms of risk pooling, reducing waste and re-use or recycling of software components. Once identified, the units may be connected through interoperable, configurable and adaptable interfaces. The resultant architecture can be recombined with other units to increase the combinatorial potential which may be essential to capture and map the process cascade, in terms of functional outcome. These units must be capable of “discovery” and amenable to be “discovered” when interacting with and between platforms.

How the modular units/blocks will be combined, to create the higher order structures, which maps to the process or the platform, is a rate limiting factor. The compilation (drag and drop) may be slow if humans are necessary to create these units per process maps or events/data flow. Automating this process and creating a “supervisory” Agent to guide the process, when triggered, is an option. The ability to use natural language descriptions as a trigger will facilitate the process and empower non-technical users to create appropriate outcomes (Figure 13) by selecting and stitching the components using verbal commands and logic operators (using set operators to specify epistemic constraints).

For example, a non-technical TSA employee encounters an individual and wants to know if that person was in pre-recorded photo bank. To set up the search parameters, the program requires defining what algorithm based facial recognition application may be used. To select the algorithm based app, one must understand the parameters that makes the algorithm useful. The TSA employee is unable to perform this task and sends the request to the technical desk to search for a match with the person the TSA employee encountered. The delay in arriving at the outcome was sufficiently long. In the interim, the person of interest had ample time to board the aircraft and depart.

Fig 13: Connectivity between cubes[^82] or modular units (blocks, bricks, Agents) can contain functions with multiple variables and create higher order Multi-Agent networks which can converge on platforms.

[^82]: Reference link to figure 13
[^83]: Reference link to Digital Enterprise X.0
One solution is programming abstraction [84]. If the program understands (semantics) natural language descriptions, then, it concludes from the natural language query that the TSA agent wants to find a match of the person within the pre-recorded material. The program solves for the required algorithm parameters, processes the request and provides an output, which may be instrumental for airport security. Software automation by implementing programming abstractions may be useful to configure dynamic architectures on demand. Interoperability between programming interfaces will catalyze seamless connectivity between cubes [85]. The hype about AI bots in automation [86] is just that - hype. Programming abstraction is a tool defined by the rigors of computer science and software engineering.

Platforms may be viewed as higher level structures to aggregate, converge and synthesize digital building blocks. Modular units shuffle and re-shuffle within dynamic transformation processes where decoupling and recoupling may design, re-design and optimize the structural elements that can deliver the expected outcome. This layer is closer to the customer’s expectation of the digital experience at the edge. The interaction at the edge favors, and often demands, pay-per-use models rather than stoic contracts. Connecting to the platform is the bridge the customer chooses. If the customer selects to cross the bridge then the possibility of monetization increases. The transaction cost to access, use and re-visit the platform (if satisfied) will drive and define service diffusion [87], create markets, predict market penetration and may be indicative of the potential for sustainable ethical profitability [88].

Digital transformation is hard at work to convert products to services. For example, Rolls Royce may no longer sell jet engines but “uptime” services which offers engine performance as the outcome [89]. The pay per performance metric eliminates the multi-million dollar price tag for the engines. By lowering the barrier to entry, Rolls Royce engines can serve low cost airlines. Reducing barriers to entry in emerging markets requires different strategic perspectives. US management gurus often insist on chanting the same mantras from an analog world, burdened with the dead weight of old technology, and affordable to less than 10% of the world’s population. Digital tools may help us leapfrog the old world inhibitions.

Fig 14: Strategy and management ideas used in US and EU unlikely to help the emerging economies.
The platform principle enables service aggregation from any person or behemoth or small or medium enterprises, if the services and tools are designed to interface with one or more open templates. For example, assume the pump manufacturer Grundfos is still living in an analog world of products but anticipates charting a digital transformation path to embrace DEX. Pumps are one component in a plant where the outcome may be purified drinking water. Creating a digital twin of the plant with its share of pumps and various other equipment is one way to create a new line of business. Analytics may predict which equipment may need service or replacement. Predictive analytics coupled with supply chain improves efficiency and reduces downtime. The outcome is a steady supply of clean drinking water. In the analog model, the pumps and data from pumps may be the sole focus of the pump manufacturer. In the digital paradigm, the contribution of Grundfos to create the digital twin of the plant and monitor its pump for optimum efficiency may lift many boats. It will earn the customer’s trust. The customer can rely on “Grundfos Digital” to monitor the entire plant operation using Digital Twins. Isn’t that a new source of revenue? Will Grundfos still continue the old world modus operandi of charging for pumps and maintenance? Can Grundfos think different and envision the global challenge of the digital tsunami? Can it transform its revenue model to reflect micro-earnings based on service/performance related to the outcome? Why not make the product “free” but charge a penny per gallon of clean water leaving the plant? Is this a new idea? Not at all. Please ask the telecommunications and the telephone industry.

Thinking different is applicable to existing platforms which must change if the platform wishes to serve the masses. Nano-satellites [90] are using different tools and technologies [91] which are clashing with the multi-million dollar satellite industry by lowering unit costs (to thousands of dollars) for satellite service. Innovation in education, using credibility as a platform to build technical services, is the hallmark of edX, developed by MIT [92] and Harvard. Massive open online courses (MOOC) are evolving [93] and often are disastrous [94]. But, it discovers students determined to benefit from available academic wealth [95]. By lowering the barrier to entry, MIT is accelerating the diffusion of excellence, usually at no cost to the student. For example, the microMasters program in Supply Chain Management [96] is free unless you want a credential. It costs $1,350 for the entire program. Successful candidates who may earn the credential can apply to MIT for a Masters in Engineering (M. Eng) degree in Supply Chain Management (SCM). Credit from the online program makes it possible to complete M. Eng in SCM in one term at MIT at half the cost [97]. The full-time residential program lasts for 10 months or two full terms at MIT.

The edX platform running the MITx microMasters program [98] is capable of harvesting the crème de la crème students who may further bolster the credibility, reputation and prestige of the program and the institution. In the recently completed SCM microMasters, around 200,000 students registered and 2,000 sought a credential ($2.7 million in tuition fees). In the absence of the SCM online program, the capacity for accepting students into the residential program at MIT is about 40. Student tuition is $71,000 for the full 10-month residential program (M. Eng in SCM). If 40 paying students attended MIT then they would pay $2.84 million to MIT. By reducing the barrier to entry from $71,000 (full time) to $1,350 (credential), the program earned $2.7 million. Was it a financial loss? Perhaps. But, the edX microMaster’s strategy identified the best of the best brains. Later, during Fall 2017, about 40 students from the 2,000 cohort will join MIT and complete M. Eng in SCM at half the cost of residential tuition (plus living expenses).
Fig 15: Pay-per-use micro-revenue model is digital by design. It touches every facet of an user everyday in a manner that is distributed over time without making time to access create a barrier to entry. For example, if an educational course is distributed in unit packages and if the charging tool operates are per unit (each unit may be 5 minutes in duration) then even full-time employees can complete a course in many steps by accessing and learning a few units, between breaks or when they have some spare time. Instead of an upfront payment of $100 for an online course of 1000 units, the user pays $0.10 per unit. Using a pay-as-you-go self-paced system converts almost all individuals to potential students if they have access to mobile device, telecom infrastructure and an incentive leading to economic growth.

Perhaps counter-intuitive but profitability of platform providers may be directly proportional to the open source strength of the platform. In other words, if platform vendor(s) offer “almost free” access to the platform then the potential for profit increases dramatically and sustained over the long term. The logic behind this apparently illogical statement is firmly rooted in the principle of micro-payments. If tier two and tier three providers can create touch points to the platform without high licensing costs then the ecosystem of function providers and app creators, from all over the world, will exponentially flock to associate with the “channel masters” platform. Consider Amazon’s network of other vendors who use Amazon’s brand to market their products (especially the category of used/new book sellers). The global diffusion of the platform and its connectivity with providers from all around the world also unleashes the creation of new markets all around the world, naturally. The more end users and customers use the services and touch the platform, the greater is the market penetration, hence, more trips to the bank.

If each pay per transaction is affordable and within reach of 80% of the population, both in terms of cost and services they actually need and use, (health, insurance, security, education but not flamboyance, ie, wardrobe matching, luxury cruise, LVMH), then every point of contact is a potential point of sale, hence, a point of revenue, even if it is only a penny per transaction. This is a profitable idea [99]. Micro-earners include PayPal, PayTM, Amazon Pay, Apple Pay, Google Pay, Samsung Pay, Ali Pay, Citi Pay and others. It goes without saying that DEX platforms and access must be essentially global and truly mobile [100].
Fig 16: Pay-per-use micro-revenue model is mobile by design. Emerging economies are the markets [97].

Fig 17: Rationale for “Miss Moneypenny” paradigm: Pay-per-use micro-revenue model at a penny per transaction may be an ethical profitability model. It may work due to volume of transactions [101] which may continue to increase [102] and already proven to be quite lucrative for the financial sector [103].
HEALTHCARE PLATFORMS

The corner of the world referred to as South-east Asia, stretches from Pakistan to Philippines. It is home to about 80% of the world’s population and enjoys a certain market envy. It may be the testing ground for the Miss Moneypenny paradigm (Figure 17). If businesses can use platforms as service delivery tools or engagement points and if pricing per access is limited to one or two pennies per transaction, it is likely to become feasible. The South-east Asian market of 5.5 billion is poor in several pockets and many still earn around $1.90 per day [104]. Two pennies per day is more than 2% of their daily earnings. This is a context that is front and center of businesses in India, China and Africa. This is why it is necessary to abandon the strategies developed and disseminated in US and EU (Figure 14) as paths to profitability.

Providing health and healthcare access using pay-penny-per-use platforms will test the art of the possible. Healthcare does not discriminate between the wealthy, poor or the impoverished [105]. It is the access to healthcare by the poor and from impoverished nations that is limited. 10 million children die annually due to preventable causes [106] and 3 million neonatal deaths are due to pre-term births [107].

Fig 18: Good downward trend for child mortality [108] but overall access to healthcare is still infinitesimal.

Healthcare is in need of solutions where technology can help to keep platforms open yet find routes to ethical profitability for the vendors via value added services. The deep complexity of security [109] and privacy infrastructure in healthcare creates quagmire. It discourages technology innovators. But, health workers are unlikely to disrupt this impasse. Innovation is expected from the technology leaders, but in convergence with people and processes that must be compliant with the ethics and ethos of healthcare.

Platform as an aggregator is the lowest common denominator in this discussion. The vast landscape of healthcare makes any discussion abysmally incomplete. Preventive medicine may serve as a simplistic use case where low cost sensors or gateways may be used in combination with mass produced low cost non-invasive devices to monitor certain vital signs, if they have access to the devices and mobile phones, perhaps with hot swappable bio-nano-sensor arrays on “sticks” connected via micro-USB interfaces [110].
Indicators such as heart rate \(^{[111]}\), blood glucose \(^{[112]}\), hemoglobin \(^{[113]}\) and urine analysis \(^{[114]}\) may be useful. Data uploaded to a common platform, when analyzed, may create a molecular metabolomics profile, which may be instrumental in molecular diagnostics, if necessary \(^{[115]}\). Pattern changes in these data sets may be indicative of physiological malfunction. Depending on the sensitivity of the sensor data \(^{[116]}\) and analytical precision, early detection, diagnosis and intervention may reduce morbidity and mortality. It may reduce acuity and hence the cost of healthcare from emergencies and hospitalization.

Fig 19: Exhaled breath condensate (EBC) is collected. Nitrite content is measured electrochemically using graphene oxide sensor. Increased levels of nitrogen oxides (NO) are associated with inflammatory disease states such as asthma, chronic obstructive pulmonary disease (COPD) and cystic fibrosis (CF). The increased level of exhaled NO in asthma may be due to an increased expression of inducible NO synthase (iNOS) in bronchial epithelium. Given the relative stability of nitrite in EBC, it is a promising biomarker of chronic respiratory inflammation. The nano-sensitivity of probe-free/label-free sensor for the detection of nitrite in EBC enables early detection and reduces cost of hospitalization \(^{[117]}\).
The transition from classical sensors \[^{118}\] to nanowire and graphene \[^{114}\] sensors \[^{119}\] may increase sensitivity for sensing changes \[^{120}\] in disease \[^{121}\], microbiomes \[^{122}\], and water \[^{123}\]. The latter is a medium of transmission for many pathogens. Some of these tools and technologies are nearly two decades old. What is lacking in healthcare is an informed vision and the convergence of technical tools on platforms. The leadership for systems integration is suffering from paralysis by analysis. The situation may be similar to use of RFID in retail stores which lacked ignition \[^{124}\] circa 2002 vs the 2017 “back to the future” attempt to re-invent the past through Amazon GO, the new no-checkout grocery store \[^{125}\].

Fig 20: More people (about 5+ billion) live inside this circle than anywhere else in the world. Five cities (small red circle) represents a market of 100 million (Dhaka, Calcutta, Mumbai, New Delhi, Karachi).

If convergence of healthcare platforms were a reality, the long-term profitability seems to be immense. Consider New Delhi, with an estimated population of 40 million by 2030. Assume 10% of the population subscribes to health platforms. The platform vendor earns $0.01 per day as a micro-payment from the customer to access/upload data/analytics through the platform. That translates to an annual revenue of US$14.6 million for one point of contact on the platform from one city. If 1% of the 5.5 billion people in SE Asia paid $0.10 per event per day then the platform revenue soars to US$2 billion per year. This is not a “low hanging fruit” and seeks an astute visionary with a penchant to make a difference to create the convergence for healthcare, for the underserved. Hence, the “Amazon” of healthcare is still out there.
Convergence in healthcare requires thinking of healthcare as a fabric rather than as a visit to the doctor. Healthcare vendors start the thinking with billing codes. Hence, it is hard to realize that a 30 minute hospital visit, after a patient survives a myocardial infarction (heart attack), leaves the patient at risk for 99.999% of the time. Where is the system to monitor the patient? We have tools but not the systems because the businesses in the business of systems integration know precious little about health. The healthcare industry still cannot find the “billing code” for remote monitoring of a cardiac patient in order to prevent future emergencies or reduce re-hospitalization. Emergency visits[^126], hospital stay, diagnostic procedures[^127] and unnecessary routines[^128] are profit centers for US healthcare industry.

Healthcare platform innovation may not germinate from Accenture (systems integrator) or CVS Health or McKesson. Electronic health/medical record (EHR/EMR database) providers are not promoting global platforms but poster perfect for epic greed[^129], skilled in information blocking[^130] by developing closed systems and preventing semantic interoperability, deliberately, by design. For healthcare information technology (HIT), the inability to access EHR/EMR information is often fatal[^131] but continues unabated.

Perhaps it is hard to change systems from inside. Tesla and Google changed the auto industry rather than GM or Toyota. YouTube and Facebook re-invented media, not ABC, NBC or BBC. Space exploration entered a new era with SpaceX and Virgin Galactic, not through Boeing or Airbus. Retail giant Wal-Mart is struggling to catch-up with innovator Amazon. Hence, the elusive quest continues, for a leader to rise and implement micro-payment based systems innovation for healthcare as a service, for the rest of us. Is it going to address all healthcare woes? No, but, it may be helpful to billions. The near complete lack of such tools and lack of EHR/EMR in most of SE Asia makes it a fertile ground for “open” platforms. The immense scale of implementation may be unsettling to providers who may think in terms of US and EU.

The ecosystem of device manufacturers and their convergence with platform providers through mobile infrastructure to acquire, deliver and distribute data, in a manner that is safe and secure, is a tall order. The open source movement is generally focused on access to data. Lack of data or inability to combine data to extract information is one reason why medical errors are a leading cause of death in the US[^132].

Enabling access to data from devices is an anathema for manufacturers. It results in interoperability issues[^133] and asphyxiates open source approaches[^134] vital for healthcare. Lack of medical device interoperability issues are more pronounced in the affluent nations (US and EU). The inability of emerging economies to afford US-based EHR/EMR systems (Epic, Cerner) is a blessing and a catalyst for open source integrated clinical environments[^135] to improve healthcare and safety.

Justifiably, monetization of the open source model deserves attention. Data acquisition from devices, the data model and interfaces, are some of the essential “open” elements. Safety, security and privacy must be demonstrated in the open model. Authorized access to raw data must be kept open and almost free for patients, hospitals, healthcare workers and anybody else who may be legally permitted to use.
A simplistic analogy is a person calling a bank or checking the online statement to find out what is the balance in the checking account. This data (balance in the account) cannot be “closed” to the account holder. If the account holder asks how to optimize her rate of interest or investment, ie, seeking advice (analysis), then the bank can charge a consulting fee for an appointment to discuss individual needs. For access to data in terms of bank balance, the bank can charge 10 cents each time a customer calls or a nano-payment of 1 cent each time the customer wishes to query the balance through the online portal.

In the open source healthcare platform model, the vendor may not restrict the data but may impose a micro-payment for accessing the platform to access the data. Customers may understand that there is a cost to develop, implement, maintain and secure the infrastructure which enables these transactions.

It may be a quantum leap forward if the healthcare industry is able to obtain raw data from devices and use an open platform model as a tool for convergence. If it is your data, then, you should be able to access your health data, anytime, anywhere. But, perhaps, there will be pay-per-access nano-payment.

If the person observes that the blood glucose level is 139mg/dl and wishes to know what it means, then it triggers a different application, which may require the services of an analytical engine to process the data and respond to the query. This analytical engine may be provided by the platform vendor or may be sourced from a third party, similar to Apple/Google app stores. Access to this app is not free and the query may cost the customer 10 cents each time. If 1% of the world’s population (~7.5 billion people) asked one question each month then it could generate $90 million in annual revenue from 1 question.

The monetization model distributes fractional micro-payments to the app creator, platform host, telecom provider and other micro-service vendors. If the system required thirty different vendors to converge for the systems integration to answer that question then each vendor may receive 1/30th of the $90 million revenue or $3 million per annum (if revenue was shared equally among the partners).

Dynamic pricing of queries may be based on the complexity of the question. Dynamic composition of micro-services, analytical tools, algorithm engines or portfolio of solvers must be sourced, aggregated, activated and executed in order to respond. It may not be difficult to understand why the structural elements of this architecture must be modular, blocks to be “discovered” if necessary, interoperable, agile, compatible, perhaps capable of natural language processing and preferably Agent based tools.

Platform analytics will evolve as a key tool for monetization for open platforms and medical internet of things (MIoT). Security [136], privacy and data de-identification may play pivotal roles. The question of “intelligence” in analytics and how the platform will address context of input data remains a challenge. Reasoning (erroneously referred to as AI) must occur in a specific context for specific goal(s). Medical expert systems accepting list of symptoms as input and generating output (most likely diagnosis) ignores the basic tenets of medicine: hypothesis-directed gathering of information, complex task of interpreting sensory data (obtain uncertain indicators of symptoms) and the goal of curing the patient. The latter may involve treating less likely but potentially dangerous condition rather than likely but harmless ones.
Artificial intelligence (AI) is a potent tool but may be an unfortunate yet very attractive choice of words to capture the public imagination. Reference to intelligence in AI is due to its use of the phenotype of neural networks, with input and output (I/O), as a core [82] feature of artificial neural networks (ANN). The lack of erudition in AI was mimicked when an AI program was requested to name colors (Figure 21). However, the world has embraced AI and etched it in its psyche. AI appears to possess that Midas touch. All things AI sells faster than hot scones with clotted cream. For a few, witnessing the transmutation of tabloid fodder from speculation to truth may be deeply troubling because there is no “intelligence” in AI [137] with respect to human intelligence. It was never there, may not be there now, or in the near future.

Yet, the tools of AI are critical in the connected global economy to catalyze growth, reduce inefficiencies, make sense of data. Collectively AI presents immense potential to increase profitability for practitioners.

Fig 21: I gave the neural network a list of about 7700 paint colors along with their RGB values. (RGB = red, green, blue). Could the neural network learn to invent new paint colors and give them attractive names? The invention of paint by numbers may not be bad [138] but the names are less desirable.
Yet it is an egregious error of judgement when venerable corporations manipulate the media [139] in the name of AI. IBM spun a story how IBM’s AI flagship Watson could improve cancer treatment [140]. It was unsuccessful yet PricewaterhouseCoopers walked away with $23 million for the project, which, despite contractual stipulations, was not implemented at the MD Anderson Cancer Center, University of Texas.

If we can look beyond the name and the fake publicity, the tools (Figure 22) collectively referred to as AI, are indeed extremely useful. Higher order reasoning is indeed necessary for digital transformation and is an important monetization tool for analytics in DEX, the digital enterprise, made up of rational Agents.

Fig 22: The combination of tools which are collectively referred to as AI by the media and businesses.

Fig 23: So called augmented intelligence [AI] in this diagram is rather pedestrian quality workflow [141]. Converting “work packages” (box) to multi-Agent systems may enable us to profit from principles of AI.
The scholastic effort in this field over the past half century or more has resulted in breakthroughs that makes it possible for systems and computers to process speech and recognize photographs, among other advances. The field has progressed from rule based expert systems [142] to pattern recognition systems which uses artificial neural networks [143] and/or statistical methods [144]. Unfortunately, the exaggerated views verging on narcissistic stupidity [145] appears to be more fashionable (Figure 25).

Fig 24: Brief History of AI – The evolution of rule based systems and pattern recognition [146] by ANN

Fig 25: Sensationalism is a hard habit to resist for marketing, even if it is a form of misrepresentation.
Perhaps we can use this AI awareness that has gripped the global imagination. It seems that almost everything in the business world strives to be laced with AI or shrewdly insert AI in the folds (Figure 26). We have moved from hot dog eating contests as spectator sports to watching numbers creep up on AI acquisition charts (Figure 27). Perhaps there is a silver lining. May be school students dreaming about six figure starting salaries will be inclined to take a deeper look at Khan Academy [147] or edX or MOOCs [148] to understand that mathematics, computer science, linguistics, epistemology, neurology, engineering and basic science are indeed salient to understanding the foundation of AI and learn how to apply AI.

Fig 26: Why is AI missing in this ad? Not really. It is carefully concealed to stir the viewer’s imagination.

Fig 27: Number of AI start-ups acquired [149]. Like watching balls jump up the tube in a lottery drawing.
Contextual higher order reasoning (CHOR) sounds insipid compared to the magnetism of AI and unlikely to become a household name. Artificial neural network (ANN) tools, including convolutional neural networks (CNN) and recursive neural networks (RNN), in combination with predictive analytical tools (ANN and statistical techniques) drives machine learning (ML) analytics. Speech, NLP and vision are components necessary for many industries and image analysis is integral to medical diagnostics.

These tools are expected to be part of analytical engines or apps [150] associated with the open data platform for healthcare (for example, openICE [131], the integrated clinical environment). These tools may or may not be created by the platform lead but sourced from contributors including crowdsourcing. Verification and validation [151] of these solvers and apps may be performed for the channel master (platform lead) before they are uploaded. Who will choose the correct solvers or apps? If the consumer does not know how to choose, the importance of programming abstraction (page 22) increases. NLP can guide these choices by extracting information from user’s speech, if the platform is quite sophisticated.

Monetization of apps [152] and sharing micro-dividends from micro-payments, harvested by app stores or app creators for crowdsourced apps, will present financial challenges that must be addressed. Apps with “security badges” “pre-validation” “five star reviews” are likely to be “preferred” and will rise to claim profitability, if the monetization model of the ecosystem is approved and online. At a systems level, nano-payments for micro-services is a horizontal need for Digital Enterprise X.0, or any version of digital transformation, in an economy where the exponential clamor for blockchain may be justified.

Platforms for remote monitoring, telemedicine and in-hospital care, are in need of open source tools to aggregate data, structure and function. Without decision support and predictive analytics, healthcare platforms may be impotent to serve as the conduit for contextual connectivity and convergence.

Fig 28: Contextual Higher Order Reasoning (CHOR) as another name for Artificial Intelligence (AI) may never see the light of the day. Decades ago, knowledge representation was introduced (it is pivotal for enterprise architecture) and the term “intelligent reasoning” [153]. Is CHOR akin to intelligent reasoning?
LOOKING FORWARD

The illustration below is "Rue Future" (Future Street) from Eugene Alfred Henard's (1849-1923) article, "The Cities of the Future" (1911). Henard (architect of Paris and from 1880 a life-long employee and advocate of public works in Paris) looks into the future and sees movement towards underground (or enclosed) vehicular traffic, "smart" buildings, pneumatic tubing for vacuum cleaners ("almost sure to come into general use"), an improvement in the system for water delivery and removal, replacing coal with natural gas. His ideas, if implemented in Paris, would cost $420 million, then, $15 billion, today [154].

Fig 29: Size, scale and vision of “Rue Future” is comparable to the Digital Enterprise X.0 (DEX) proposal.

DEX is comparable to the foresight in Rue Future. Hence, it is difficult to conclude any discussion of DEX and its building blocks because we haven’t even started the journey, in earnest. Digital transformation must be viewed in the context of demand, demographics, markets, economic trends in the context of the socio-cultural milieu. Digital enterprises must embrace digital twins, optimize security and mobility. The challenges of DEX may not be solved by yesterday’s thinking or today’s solutions. Opportunities for the diffusion of DEX may arise from unexpected quarters, unchartered waters and may even unleash untapped potential.
Fig 30: Flooding the Sahara Desert could have changed Africa. Technology should serve humanity.

Fig 31: Testament of Youth [155]
Fig 32: Money left on the table? Gender parity may increase GDP up to $12 trillion in 2025 [156]. The education of women and gender parity at the workplace is an economic no-brainer to increase profit.

<table>
<thead>
<tr>
<th>Region</th>
<th>Incremental 2025 GDP over business-as-usual scenario, %</th>
<th>Incremental GDP, $ trillion</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>16%</td>
<td>0.7</td>
</tr>
<tr>
<td>Latin America</td>
<td>14%</td>
<td>1.1</td>
</tr>
<tr>
<td>China</td>
<td>12%</td>
<td>2.5</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>12%</td>
<td>0.3</td>
</tr>
<tr>
<td>North America and Oceania</td>
<td>11%</td>
<td>3.1</td>
</tr>
<tr>
<td>World</td>
<td>11%</td>
<td>11.8</td>
</tr>
<tr>
<td>Middle East and North Africa</td>
<td>11%</td>
<td>0.6</td>
</tr>
<tr>
<td>South Asia (excl. India)</td>
<td>11%</td>
<td>0.1</td>
</tr>
<tr>
<td>Western Europe</td>
<td>9%</td>
<td>2.1</td>
</tr>
<tr>
<td>Eastern Europe and Central Asia</td>
<td>9%</td>
<td>0.4</td>
</tr>
<tr>
<td>East and Southeast Asia (excl. China)</td>
<td>8%</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Fig 33: In 1920, Goddard outlined how a rocket might reach the moon. NY Times wrote he “lacked knowledge ladled out daily in high schools” and argued thrust was not possible in a vacuum [157].
Fig 34: Apple employees Jeff Robbin and Bill Kincaid developed iTunes in the 1990’s but left Apple to market it as SoundJam (1998). Apple bought SoundJam and re-introduced it as iTunes in 2000. In 2001, the debut of iPod in the US catapulted the concept of micro-payment based crowd-sourced apps on a platform. Combining the concept of iPod and the almost parallel release of i-mode (2001) platform by NTT DoCoMo in Japan, it wasn’t hard for the public to predict that the iPod may also serve as a mobile device and phone, as suggested in a 2005 article reproduced above [158]. The control of objects using a mobile device was a concept introduced around 1999 in the form of IoT by the MIT Auto-ID Center. The concept of IoT with that of a mobile device linked to telecom was one rationale for the suggestion (cartoon above) that in future we will control automobiles with a phone. The figure shows the potential of “plugging in” or inserting a “stick” in the automobile. This concept was not elaborated in the 2005 article because it was too far out as science fiction, at that time. The concept of inserting a “stick” to control the automobile was based on ideas surrounding metastable hydrogen which proposed the potential of hydrogen to behave as a metal [159] and hence “packaged” as a source of energy for hydrogen powered objects, for example, the hydrogen powered automobile model showed in the cartoon (above). The disclosure of this idea at this time is appropriate because the presence of metastable hydrogen hypothesized by Eugene Wigner in 1935 has been confirmed to exist [160]. It is a first step by graduate student Ranga Dias at Harvard University [161] on the road to driving an automobile or flying an airplane using a highly compact USB thumb drive type “stick” as shown in the cartoon from 2005 using the iPod form factor. The future stick may be no larger than a thimble or perhaps similar to a credit card sized pack of energy. This form factor may change the automobile industry which will no longer need solar fuel cells or batteries which weigh thousands of kilograms and reduces range. Taxicabs and auto-rickshaws can use credit card size energy packs and swap the energy cards in retail stores or convenience stores when depleted [162]. Mobility of atoms and its connection with bits (data, information) is a catalyst for the untethered but networked physical world - a digital by design metaphor suggested by Sanjay Sarma [163] and others, as a conceptual basis for diffusion of IoT.
In our excitement to engage with the explosion of emerging advances, tools, ideas, IoT, and applications, we have devoted less attention to the foundation of semantic interoperability \cite{164} necessary to make sense of the data. Meaning of the data drives its value and extracting actionable information is the Holy Grail. At the heart of this process is the convergence of data, which is essential for relevant analysis, but that cannot happen unless we optimize connectivity between data holders (streams, lakes, fog, cloud), which must understand (cognitive tools?) the context of the data, in order to select, channel or curate the data, before subjecting it to associations, algorithms and analytics. Cross-pollination and information arbitrage depends on semantic interoperability between modules, building blocks, systems, ecosystems and end-user facing applications (where the rubber meets the road in terms of revenue, profit and loss).

We will be handicapped if we fail to “look forward” in addressing the ontological basis of semantic interoperability between entities. Agent based architectures are one approach. A global discussion between standards bodies, ontologists, linguists, business user groups, developers and enterprise architects are essential to improve systems interoperability, if and where Agents may fear to tread \cite{165}.

Fig 35: In 1987 John A. Zachman, an IBM researcher, proposed the Zachman Framework, a concept of what is involved in information system architecture \cite{166}. Organizations may not have one architecture, but a range of diagrams/documents representing different aspects or views and/at different stages \cite{167}.

Fig 36: Models sub-layers to ontologies. Enterprise architecture as an ontology of symbolic systems \cite{168}.
CONCLUSION

It is possible the transformational tsunami will wipe out the need or significance of DEX. The discussion about forward looking thoughts may be wishful thinking. Quest for open source \[169\] may be stunted. The academic discussion what constitutes intelligence may be obliterated by the market forces driving AI. The financial muscle behind market driven AI (Figure 37) can crush dissent and defund dissenters.

However, it is reassuring that AI leaders in some of the corporations are exceptionally brilliant engineers (for example, Peter Norvig, Andrew Ng) who are aware of the deep shortcomings of intelligence in AI and they pay attention to neuro-scientists (Tomaso Poggio). Yet, the marketing machines keep churning out chest-thumping tabloid fodder, as if it were the truth. In the middle-ages “artificial flight” attempted to mimic birds. The Wright brothers turned to wind tunnels and aerodynamics, which removed “artificial flight” from our vernacular and gave birth to the aviation industry. AI will generate deep reasoning (DR) and profitable analytics but not intelligence. But, truth always lags behind, limping on the arm of Time (Baltasar Gracián y Morales). Is the truth in AI for a few? Is this discussion about AI a storm in tea cup? Will it suffice if the market acknowledges this is not intelligence or AI but DR or deep reasoning?

Why AI is devoid of intelligence may be unintelligible to individuals without knowledge of neurology or the desire to understand what does it mean to sense or taste or hear. The various aspects of neurology essential to grasp the concept of intelligence, even in its rudimentary form, may include anatomy and physiology of the nervous system and a comparative perspective of neural systems (octopus vs worms). As discussed elsewhere \[47\], the physical representation of neurons in our brain is not representative of the functional outcome. This is the single most crucial factor in this debate. Just because computational systems can create neuro-morphic structures, does not mean it can re-create the function or pass the Turing Test. Cryptic and deep in the functional context are states of intelligence, intuition, perception and characteristics that makes us, humans, intelligent. Harnessing the amorphous attributes are beyond engineering systems because the biology still remains undiscovered and our elusive quest continues.

Fig 37: R&D budgets of AI-intensive firms and select countries (OECD, 2015). Bubbles brewing trouble?
Fig 38: Innovation abhors conventional wisdom and driven by vision reflecting confluence of ideas

Fig 39: Geography of biotech innovation (Ernst & Young, 2016) proportional to research and education?
Quantum leaps are integral to progress of science (Galileo Galilei) and society (Leo Varadkar). The revelation “what is intelligence” may be breaking news, tomorrow on BBC, from Corina Logan. Hence, paradoxes may be converted into paradigms. What is heresy today may be in harmony, tomorrow. We must view the past (Fig 38) with the present (Fig 39) and accept change (Fig 40) in our future (Fig 41). But, it is encouraging to observe the diffusion of change, even if they are only in illustrations (Fig 42). Things that we may not wish to change are our norms of respect, égalité, and a sense of dignity [170].

Fig 40: French Revolution 2.0 (La deuxième partie débute environ deux cents ans plus tard).

Fig 41: Le parti naissant du président français Emmanuel Macron La République en Marche et son modède d’allié centenaire sécurisent 350 des 577 sièges à l’Assemblée nationale. Macron réfute Voltaire? En 1741, Voltaire [171] écrivit "sans la voix de la la Maure, & le canard de Vaucanson vous n’auriez rien qui fit ressouvenir de la gloire de la France." Macron voudrait différer. N’est-ce pas?
Fig 42: The manufacturing enterprise illustrated by Deepesh Nanda suggests convergence of digital threads in DEX.

REFERENCES

1 https://ia902205.us.archive.org/23/items/floodingsaharaa01mackgoog/floodingsaharaa01mackgoog.pdf
3 http://callingbullshit.org
5 http://www.newyorker.com/magazine/2014/06/23/the-disruption-machine
6 http://www.economist.com/node/11848432
7 https://hbr.org/1993/07/building-a-learning-organization
8 https://1.cdn.edl.io/ewMbRRD0v8rfYEBfrRN4JnsbvgPQ2g28N1oGvi9HYX2s0mC.pdf
12 http://cass-mt.pnnl.gov/docs/ia3-2013/brockman.pdf
13 https://dspace.mit.edu/handle/1721.1/107989
The Transformational Tsunami by Dr Shoumen Datta

http://cbcl.mit.edu/cbcl/publications/memos-reports.html

https://pdfs.semanticscholar.org/7de2/0692c774798016e06d2eb0ade959c61df350.pdf


http://ubiquity.acm.org/article.cfm?id=2822533


www.mdpnp.org and www.openice.info

http://www.cnri.reston.va.us/k-w.html

http://ubiquity.acm.org/article.cfm?id=2822529


www.open-mesh.org


http://www.ieee802.org/1/pages/tsn.html


https://opencore.com/ressources/what-are-microservices


https://dspace.mit.edu/handle/1721.1/41902

https://dspace.mit.edu/handle/1721.1/107988


http://www.acad.bg/ebook/ml/Society%20of%20Mind.pdf


https://books.google.com/books/about/Good_Profit.html?id=Kp_GBgAAQBAJ&source=kp_cover

https://hbr.org/2016/06/manufacturing-companies-need-to-sell-outcomes-not-products

https://www.nasa.gov/mission_pages/station/research/experiments/1326.html

http://www.accion-systems.com/team/

http://odl.mit.edu/

http://media.proquest.com/media/pq/classic/doc/3389125941/fmt/ai/rep/NPDF?_s=wUTlu9wjfAltqKUqmHEHxqi6w5E%3D


APPENDIX

The Great Wave at Kanagawa, 1829 by Katsushika Hokusai (modified)

Digital Tsunami

The objective of this blog is to suggest tools and structures that businesses may use or develop in order to benefit from the tsunami of digital transformation, create new lines of business and venture into new global markets. The task is immense and as a consequence any cohesive synthesis is vastly imperfect, as indicated in the “preface” on page 2. Despite its critical shortcomings, the appendix (this section) will try to provide a context for individuals who may be less acquainted with the current wave of digitalization.

Imagine a well-built grand highway.
Now, re-imagine a well-built grand highway but in the middle of nowhere!

Consider that within a few miles from this highway there are people living in the area. You are here to sell Tesla Model 3. By now you have grasped the conundrum. Haven’t you? Your potential customers near the highway are strongly suggesting that if you wish to sell cars, please help them to build roads.

But that is a different task, the job of building roads. Other potential customers are taking you (on their horses) to show some of dirt roads and even stretches of ill-maintained paved roads which are peppered with pot holes. Can you help them repair pot holes or can you sell them faster horses, instead of Tesla?

In the analogy to digital transformation, imagine the highway to represent the digital platform of convergence. The car is your application or transaction or enterprise, on that platform, which enables you to connect to the ecosystem, using the platform (that is, you travel from A to B to C and may engage in commerce, if the highway was connected to the people and businesses via roads, lanes and bridges).

The job of building roads and repairing pot holes, dampens the thrill of driving the Tesla down the highway. Digital transformation projects are failing or pilots are struggling to break away from small scale implementations because we are interested in the exciting aspects of the digital revolution, IoT and AR/VR. We are less interested to help bridge the gap from legacy, devices and monolithic systems.
This blog suggests that decomposition of the enterprise into modular blocks and enabling connectivity between the building blocks to synthesize the customer facing outcome may be a multi-step and tedious process, but one that is mandatory, if we expect to profit from the vast potential of digital enterprise.

To profit, the customer facing outcome must be the reverse of what is mentioned in the last paragraph. Customers are seeking solutions to their problems not just “patches” or information about the problem when unpacked and reduced to tractable sub-problems. Dissection is quintessential but the customer will reward (pay) the value provider who can synthesize a “systemic” solution not just install a tool or technology as a contribution to the “whole” solution. Hence, “tool” providers (Intel, Cisco, Microsoft, Oracle, SAP, IBM) are all consulting companies. Tools are not produced in the context of the enterprise but externally with guidance from knowledge experts. The demand from system integrators (Accenture, Deloitte, KPMG, Wipro, Infosys, TCS) is to “stitch” the “patches” from “tool” providers and expect a harmonious and functional “quilt” to serve as a solution. Exception to this model are conglomerates like GE, Siemens and Hitachi. They are evolving from manufacturing and finance to systems companies recognizing the underpinning of data, information and knowledge which is at the core of profitability in a connected world. The creation of GE Digital and its software evolution through Predix, is one example. It reflects a thinking about systemic solutions approach to problems rather than depend on vendors or system integrators. Both are essential in the ecosystem but business solutions from businesses may offer a different flavor because they view the whole “elephant” rather than look at it in parts.

The Blind Men and the Elephant ● John Godfrey Saxe (1816-1887)

In systems thinking a common analogy is that of six blind men from Indostan (India) touching various parts of an elephant and claiming that “elephant is a tree” (man who touched the leg), “it is like a rope” said the blind man who touched the tail. “Like a snake” (man who touched the trunk). “It is like a big hand fan” (man who touched the ear). “Like a huge wall” said the man who touched the belly of the elephant. “Like a spear” said the blind man who touched the tusk. They were misguided in describing the part they touched and as a consequence their interpretation is incorrect about the whole animal.
The illusion of expertise peddled by vendors and integrators are true in the context of their “part” of the solution. The latter was not unsuitable for an analog world where the company had no idea of what happened to the product once it left the manufacturing plant. In the service paradigm, connectivity is the bread and butter. Customer experience is king. The ubiquity of digital transformation has the potential to constant communicate and harvest outcomes from each user layer or machine or device. The inability of the system to use that data or extract the information is directly proportional to its profitability or lack of it.

Hence, the “illusion of expertise” which was once suitable and sufficient for discrete analog enterprises, is now a chronic recurring challenge which handcuffs systemic innovation. The vast complexity of business evolution in its always connected digital transformation _modus operandi_ is an entirely new playing field. There is plenty of room for conglomerates to profit from unleashing _intrapreneurs_ and supporting unconventional entrepreneurial innovation to splice with conventional wisdom. Leaders must recognize the illusions (and delusions) in traditional business as usual and help to remove those roadblocks to spur real transformation, that is, the whole “elephant” is your customer.

The whole “elephant” is the customer. An “analog” approach to “parts” may be insufficient to increase profitability when continuous connectivity of service, and “intelligent” decision support systems in near real-time, are desired outcomes and key performance indicators (KPI) expected from digital enterprises. The metaphor of building blocks in this blog now must help build the metaphorical elephant which is a connected system, a biological enterprise. One cannot use the mouse as a model to build an elephant.
Digital in 4D

COMBINE

Telecommunications (COM) + Bio (B) Info (I) Nano (N) Ecology-Environment (E)

Shoumen Palit Austin Datta
This blog is dedicated to the memory of Mrs Sheila O’Sullivan. If diagnosis and treatment of her breast cancer, had access to these tools, there would be so much reason to celebrate her life, rather than her death anniversary, on 25\textsuperscript{th} September.
The ideas and suggestions in this blog are not uncommon. This is neither an unique proposal nor a report about a new widget or novel technology. The strength, and purpose of this blog, is to seek convergence. Suggested scenarios, highlights the synthesis and confluence of tools, without any new inventions. The suggestions offer room for innovation, social business, economic progress, plethora of entrepreneurial opportunities, ways to improve quality of life and lift many boats.

Suggestions in this blog may not provide any final solution, it is a compass to create solutions, far beyond the scope of this discussion. The primary focus is on detection and orders of magnitude we may gain, in terms of sensitivity of detection, from micro-arrays vs nano-arrays. Unless we can detect, we can neither measure nor diagnose. Unless we detect long before the event precipitates, the actions and analytics necessary to trigger predictions for prevention, may be moot and sterile.

PROBLEM SPACE - SOLUTION PORTAL

US DHS CBP officer at Boston Logan Airport Global Entry kiosk hands me a tube, and instructs me to spit inside the tube, after I take it over to the attendant, in the holding area. I am returning back to the US, after giving a talk at the Internet Forum in Goma, DRC. During the event, I lodged at the Lake Kivu Serena Hotel in Gisenyi, Rwanda. To re-enter the US, I must be tested for Zika virus, first detected in 1947 in the Zika forest, Entebbe, Uganda, near Lake Victoria. The forest is more than 500km east of the locations I visited. Goma and Gisenyi, are both located on the shores of Lac Kivu, but in different countries, Democratic Republic of Congo and Rwanda, respectively.

My sputum is diluted, and a miniscule drop is applied (using a micropipette), on a flash drive with a micro-USB. The attendant hands me the flash drive and suggests I insert it in my phone, after activating the DHS Global Entry app. The app display reveals a distinct red button. The dialog box prompts me to insert Foreign Agent Detection (FAD) flash-sensor drive. I insert the flash drive. In about 30 seconds, the red button turns blue. It reads: “SUCCESS! Foreign agent was not detected. Welcome to the United States of America.” I return the drive. On inquiry, I learn the assay[^1] used gold nanostars conjugated with antibodies to detect Zika non-structural protein 1 (NS1). The sensitivity to the biomarker was improved using surface enhanced Raman spectroscopy (SERS). The assay did not detect NS1 in my sputum. Hence, I was free to walk out of the secure area. A DHS agent re-checks my app for the blue button. I proceeded to baggage claim to pick up my bag.

ABSTRACT

This blog analyzes tools and technologies, to collect data, and the convergence, to transform this scenario (above) to reality. This is one of many potential[^2] applications in healthcare. However, ‘detection’ is useful in agriculture, security, environment, manufacturing, etc.
INTRODUCTION

The public health scenario, in practice, may prevent epidemics and pandemics. It may help identify the point of origin, pattern of dissemination and spread of infection, due to a pathogen.

The sensor, in the scenario, can sense any material which can be engineered\(^3\) and detected, within the limits of what can be sensed, if the limits of sensitivity, are useful for the purpose. Detection of air-borne pathogens\(^4\) or toxins\(^5\), arsenic\(^6\) in water, pesticides\(^7\) in soil\(^3\), bacteria in food\(^9\), methane\(^10\) from animals, gas\(^7\) in the environment, nerve agents\(^12\) in conflict zones\(^13\), chemical\(^14\) pollution and glucose\(^9\) in blood\(^6\), are a few, of the many use cases.

Transmission of the outcome (data resulting from action of the sensor) must generate a wireless signal\(^14\), which can be captured\(^7\) in an app\(^7\), is essential for connectivity and communication. It allows us to thread the “digital needle” in order to move the needle, and take advantage of the wave of digital transformation, and benefit from the imminent tsunami of digital economics\(^18\). Various efforts target only part of this end-to-end scenario (detect, connect, transmit, decide).

![Diagram of public health application](https://dspace.mit.edu/handle/1721.1/111021)

**Figure 1:** Simple illustration of public health application. The basic research, in principle, and the tools to connect the data, to an application, on a mobile device, are available. What are the key barriers to implementation? Growth in air travel and shipborne trade, are reasons for appearance of communicable disease pandemics, disease vector invasions, vector-borne disease movement\(^19\) and infected people. A risk mitigation strategy, without emphasis on detection, is blasphemous.
Figure 2: Monitoring bio-markers may help prevention, and treatment, in progress. Prediction of potential risks can reduce morbidity and mortality due to common diseases (stroke, congestive heart failure, COPD). What often happens during cancer treatment (upper panel) may be reduced by monitoring biomarkers with high sensitivity nano-arrays, to alleviate multi-factorial resistance.
APPLICATIONS – POPULATION STUDIES

If Twitter (Figure 3) can be used as a crowdsourcing app to map hate\(^ {20}\), imagine the immense benefit to public health (and epidemiology) if we can crowdsource sensor data, about air-borne pathogens or arsenic in water (for example, Bangladesh). Using the appropriate sensor, attached to a mobile device (microUSB flash drives with swappable slots for analyte-specific sensors), one can generate global maps of analyte distribution, and track and trace, any molecule, anywhere, anytime, if the sensor and signal transmission is compatible with smartphone microUSB platform.

Figure 3: Potential of crowd-sourced sensor data, in convergence with GIS, to map harmful agents.

Figure 4: Converge: crowd-sourced data and swarm\(^ {21}\) intelligence\(^ {22}\) for data analytics?
LOCAL DECISIONS USING GLOBAL PLATFORMS: SENSE-TRANSMIT-COMMUNICATE-RESPOND

Project Ara from NK Labs[23] (Cambridge, MA), failed to ignite the spark, at the hands of Google[24] but its potential, if viewed with ‘new’ eyes, may be innovative for sensor applications.

The swappable modular architecture may be the preferred “housing” infrastructure for the sensor and serve as a platform. The form factor of this housing platform is a standard (USB) connector based flash drive-like unit which can transmit, compute and communicate with smart devices, for example, phones, tablets, automobiles, cameras, lawn mowers, refrigerators, washing machines. For mobility applications, the device must be mobile, but for stationary monitoring, any widget, with a connector, may be adequate, as long as it is functional, within the end-to-end ecosystem.

Computational requirements within the housing platform may need tinyOS (operating system), tinyDB (database), tinyAPI (application programming interface). Seminal work by David Culler[25] and Howard Shrobe[26], among others, are available for innovators, to introduce these functions, within this form factor (housing platform). Security, identity, authorization, verification and data privacy are cybersecurity[27] issues which depend on data de-identification.

Identification of the housing platform, and creating an unique identification of the test subject, in combination with the tool (drive), may use RFID EPC code plus the IPv6 address of the device or combine with biometric id. This may be relevant for travel and immigration.

Figure 5: Any analyte, or pathogen, can be sensed, anywhere. If the sensor, and smartphone, can capture the data, it may be transmitted, anywhere. One assumes the transmission infrastructure exists, is protocol-agnostic and connects or uploads between wired or wireless, nodes or gateways. Examples of molecular diagnostics, illustrated, are benefits from medical internet of things (IoT). It may help preventive medicine, and diffusion of healthcare tools, to remote places in the world.
Figure 6: Arsenic contamination of water\cite{28} leads to severe morbidity, for many, in Bangladesh.

Figure 7: Pioneering work by Nokia. Connecting sensor data, for detection of Arsenic in water, with communications, may be rejuvenated in the era of IoT. Digital transformation, and detection of Arsenic, can help to improve the quality of life for inhabitants of Bangladesh. Millions of people may use a mobile phone, with attached sensor, to test any water, they encounter. The data can be uploaded to the cloud. Dynamic data analytics may provide precision maps of the areas to avoid. It may trigger measures for water safety\cite{29}, calibrate purification\cite{30} and preventive healthcare.
Unless we can make sense of the data from sensors, in near real-time, the value of detection may be zero. Hence, the importance of telecommunications in this end-to-end ecosystem. Nokia’s entry into “digital water” was not triggered by altruism but due to the business case, based on pay-per-use micro-payment business model (discussed in “Digital Enterprise” which is a part of “Digital Transformation” here [https://dspace.mit.edu/handle/1721.1/111021](https://dspace.mit.edu/handle/1721.1/111021)). Telco standards and interoperability between protocols, may help or hinder adoption of solutions, in different areas.

Figure 8: Sense and response systems (in discussion) must transmit data to centers or to the edge.

Figure 9: Competing standards and protocols in the short range (1) are an increasing problem. This is the telco segment where “detection” tools are likely to be used. Hence, compatibility and interoperability between protocols and standards, is a pre-requisite for function. The longer range (2) is used for other purposes, eg, smart grid, where there is less competition and more regulation.
Open standards and use of global standards, for example, Open Systems Interconnection (OSI) model, are important elements in the solutions approach. OSI (Figure 10) characterizes and standardizes the communication functions of a telecommunication or computing system. Its goal is the interoperability of diverse communication systems with standard protocols (Figure 11). The 7-layered OSI design is due to Charles William Bachman III (December 11, 1924 – July 13, 2017). Aspects of the OSI design are similar to ARPANET[3] and related networks, in other countries.

Figure 10: OSI model. ICT WAN (information communication technologies wide area network) standards are bifurcating to high-power/high-bandwidth and low-power/low-bandwidth segments. Most ICT WAN protocols aid in sensing (not in actuating). Operational Technology (OT) is not directly relevant to this discussion but it is an integral part of the 7-layer OSI model.

Figure 11: Standards relevant to OSI model. Open standards reference model from David Culler[22]
REVIEWING THE COMPONENTS OF THE SYSTEM – EXPLORING THE ROOM FOR INNOVATION

The scenario presented at the beginning of this blog starts with

[1] fluid (sputum) applied to
[2] sensor (FAD flash-sensor drive)
[3] which communicates via an app (data).

The sample containing the analyte or pathogen or bio-marker is the natural product, to be tested. Innovation is not applicable to this phase. Selecting alternate fluids may not be an innovation, for example, if an invasive test, involving blood were to be used, instead of the non-invasive sputum.

Signal detection, data transmission, storage, transport and visualization through an app, may not be as direct as procuring the components, in the bill of materials (BOM), from an “off-the-shelf” catalogue. However, these steps are used in millions of processes. Information is available for configurable components (for example, tinyOS and tinyDB). Communication systems must use standards and protocols. Apps, APIs, GUIs are generic software products. Taken together, it must converge to generate the result – the data – which must be analyzed, “intelligently” if possible[32].

Data analytics is of paramount importance. It is fertile grounds for innovation, to fuel decisions, following detection, and diagnosis. Unlocking actionable information from data is the Holy Grail. Data, analytics, AI, and platforms, are brimming with hype. Examples of how the cookie[33] can crumble[34] are cautionary tales. Even if you have a hammer, does not mean everything is a nail.

![Diagram](https://dspace.mit.edu/handle/1721.1/111021)

Figure 12: Making sense of data, with respect semantics and context[35] of data, is crucial to obtaining useful information. Data must deliver value to justify the investment in the system. The application of AI (artificial intelligence) algorithms, and its various manifestations, are a part of the portfolio of approaches to data analytics. Choice of algorithms, the type of solvers and models which may be used in analytical engines, may evolve with the application, or the application may trigger functions based on context, an old idea[36], reincarnated through SDN[37]. The nature of the data, transmission speed, bandwidth and the latency the data can tolerate, (before the data or information perishes in value), may dictate whether the analysis, and response, may occur via the cloud, the fog or mist or at the point of contact[38], that is, at the edge (as in the DHS scenario).
ROOM FOR INNOVATION?

We are left with the actual sensor. This is a critical point of convergence and demands fusion of transdisciplinary domains. Generating viable analyte-responsive sensors require orchestration of medical parameters, biology of molecules, chemistry of molecules and material, material science, biomedical device engineering, and manufacturing processes necessary to produce sensors.

There is plenty of room for innovation but one must start with asking correct questions. What are we planning to detect? This is where the depth of science, and knowledge about the subject area, is critical. The sensor will be worthless, unless identification of the target is unambiguous, and is relevant, from the perspective of biology and medicine (if applicable), in terms of function.

The next question is the knowledge about the complementarity of the detection process. For example, if we wish to detect an antigen or an epitope of an antigen, then its specific antibody must be available and presented, as the immobile component, on the surface of the sensor.

How many immobile components will be available for interaction on the surface of the sensor? This is where the “game changes” as we move from micro-arrays to nano-arrays, which may owe its origin to the principle of nanowires, but modified, over the past quarter century. The move to nano will increase the surface area but will that result in a concomitant increase in the sensitivity of detection? Detection ability of novel nanowire sensors are in the picogram per milliter (pg/ml) range and at a 100x reduction in cost compared to other manufacturing processes for silicon nanowire sensors (Marcie Black, Advanced Silicon Group, personal communication).

Sensitivity of detection, is one reason why nano-wire sensors (nanoarrays) could change the face of public health and preventive medicine. Coupled with digital transformation, and IoT, as a digital by design metaphor, the scope of digital health, and medical IoT are, truly immense.

Figure 13: The revolution in detection due to nano-arrays using silicon nanowires. Compare 5-10 molecules on other forms of nanowire devices versus this configuration (above) which has the potential to be far more sensitive due to >1 million nanowires per device (six orders of magnitude increase, per sensor surface). Theoretically, each nanowire can serve as an anchor for an analyte.
The biochemical challenge is retaining the functional viability after attaching the analyte to the nanowire. An assay to measure the amount of attached analytes cannot be used to extrapolate the gain in sensitivity due to nanowires. To estimate the gain in sensitivity, we must reproducibly assay the percentage of the attached molecules which are functionally active to detect the target. What is the tolerance for variability in assay conditions with respect to functional sensitivity? Is the nano-sensor suitable for multiple uses or single use? Will the complex formation (necessary for detection?) auto-dissociate to rejuvenate the nano-sensor for another cycle? In the absence of auto-dissociation, can the nano-sensor return (treated to return?) to basal or ground zero state? What is the nature of false positives? Is error correction and/or data curation/cleansing required to obtain the result? What is the detection signal when the analyte detects the target? In what ways this signal can be modulated? How is the signal converted to data? Attempting to answer these and other related questions, provides ample room for innovation in biology/biochemistry.

Can we create standard operating processes (SOP) for nano-arrays, using a modular approach? In this vision, the components, except the target, are pre-fab “kits” from vendors (for example, New England BioLabs). Democratization of the nano-array tools, beyond specialized labs to any clinic or industry, may benefit other domains, for example, exploration of microbiomes[41].

Figure 14: Network of interactions between gut microbia in obese children[35]. With low cost, routine nano-metabolomic assays, out-patient or primary care clinics, may detect molecules of interest, while they are still below the threshold (picogram or femtogram per milliliter) of disease or dysfunction. Nano-arrays for prevention, in home test kits, can use non-invasive body fluids (sputum, urine, feces) to detect target molecules, in a manner similar to pregnancy tests or blood glucose strips, available from pharmacies (CVS, Walgreens, in USA). Nano-kits may be purchased from brick-and-mortar stores or ordered online (WalMart, Petsmart) or delivered-by-drone to your door-step or window-sill (Amazon, Alibaba), in the near future.
CONFLUENCE OF IDEAS

The vast room for innovation, in optimizing sensitivity of detection of analytes, will determine the ultimate power of, and benefit from, nano-arrays. Availability of home nano-array kits in gas stations and convenience stores, can reach far corners of the world, where availability of medical professionals may not need a miracle, but is as rare as a solar eclipse. What was not feasible with microarrays may be possible with nano-arrays in the detection of agents for flu, cholera, plague, TB, malaria, leprosy, meningitis, diabetes, myocardial infarction and so many other diseases, which affect the rich and the poor. If detected, before the threshold, millions of lives can be saved.

Digital transformation catalyzed data collection, enables quantitative measurements, establishing metrics, changing the rigidity of “the model mind set” and driving key performance indicators. In the data driven economy, KPIs must be guided by the context of the problem or the environment, rather than feeding models from antiquity, limited by the scope of their equation based structure. In the digital era of near real-time data, the dynamic composability of agent based models and its modular components (parameters, features), must be continuously constructed, de-constructed and reconstructed, to keep pace with the volatility of data. The data must not be made to fit the model. The model must fit the observed data, to be useful, at least per the Pareto principle. Data from detection is no longer the “Epic of Gilgamesh” available to the Queen. Digital transformation can make this data accessible to anyone, anywhere and we may access that same information (please see page 5 in “Digital Transformation” here https://dspace.mit.edu/handle/1721.1/111021).

Data repeatedly re-appears to be at the heart of decisions. Suggestions in this blog are “obvious” as well as data dependent. Mobile phones analyzing sensor data can detect tremors related to Parkinson’s disease[42] and display in apps. In the near future, the smartphone, which you plug in the car, to charge, while you are driving, can also inform you if the tire-pressure is low or the car needs wheel balancing – without the automobile’s sensor. Sensors in phones (accelerometer, gyroscopic sensor) can “sense” the operation and use machine learning (ML) to predict status. Sensors can upload real-time data about emissions and record reduction of pollution by EVs.

![Figure 15: We do not need nano-array sensors to detect NOx (g/km). EU data from diesel cars[43]](https://dspace.mit.edu/handle/1721.1/111021)
The ability to connect the data to health monitoring or healthcare systems, unleashes multiple benefits, for the patient, individuals, families, local community, national governments and global organizations (PIH, WHO, MSF). Data and data analytics are essential elements in the closed loop abstraction, that is, cycle of data and analytics feeding the decision, and data from implementing the decision, feeding the next round, of the dynamic decision process. Data and contextual analytics, at the right time, may reduce [a] the 43 million cases\[44\] per annum, worldwide, due to medical errors, [b] 99,000 deaths per year due to healthcare related infections and [c] 770,000 injuries and deaths due to problems with medication. The outcome, hospital readmission, costs\[45\] an estimated US$17 billion, each year. This is the immense value, from connected data.

How do we collect data, how do we communicate it, how we store it, how we transmit it, how to query it, and what to do with it, in order to extract value for the decision process, with or without humans, are questions that must be addressed, case by case, because it is not formulaic. We have come a long way from the stethoscope of 1816. In about 200 hundred years, we can use a mobile smartphone, which can connect to an USB flash drive, with a sensor. It is a sign of the times.

The type of sensors suggested in this blog, are able to track and trace analytes, related to health and environment, among other things, using nano-wires for ultra-high sensitivity nano-arrays. Anybody who doubts “if” it will happen, may wish to pay attention to recent developments in chemistry. Two groups from the same department at the University of Manchester, Chilton & Mills\[46\] and Richard Layfield\[47\], have reported a dysprosium molecule $^{66}$Dy with switchable magnetic properties, single-molecule magnet (SMM), with the ability to store a single bit of data. The stunning chemistry of the dysprosium molecule runs parallel to the thrilling potential for advancing information technology. Molecular data storage using SMMs could handle about 30 terabits of data per square centimeter, or more than 25,000 GB of data ($>$3 terabytes), on a device the size of a flash drive. That is about 10-fold more than the current (256GB) data storage capacity of Apple's latest iPhone 8, which relies on magnetic nanoparticles.

Cross-pollination of multi-disciplinary approaches, and innovation, may guide us to create new vistas, to analyze analytes, in an attempt to inform the processes, which may improve the health of animals, humans, plants, and the collective health of the interconnected environment.

**Figure 16:** In this blog we have focused on the bottom of the pyramid. “Detection, Diagnosis and Decision” needs data. Nanosensors and mobile devices are tools we may use for collection of data.
Figure 17: US corporate R&D budgets. Tech companies (red bars) in the Top 20 list, by R&D spending, invested almost US$100 billion in research and occupy 5 out of the Top 5 positions[48]

Figure 18: Feature engineering is key to AI. Insufficient, noisy, unstructured data, with respect to features, renders these tools impotent. From[49] Ian Goodfellow in Deep Learning. MIT Press. 2016
CONCLUDING COMMENTARY

Science and engineering are at the heart of this blog. The identification of biomarkers and analytes, as molecules of interest, is fundamental. Scientists who can transform these ideas into reality, may not care much about the end-to-end qualitative scenario, outlined in this blog. In other words, majority of scientists will not consider reading a “blog” by a non-expert. Engineers who can fabricate the material, to create the nano-wires, will follow in the footsteps of scientists. The engineers in telecommunications, architecture and systems connectivity, will disregard the blog because it deals with “biology” or medicine. The computer scientists and the data scientists (we used to call them statisticians), will ignore the blog because there is “nothing in it about actual data” to discuss. Most business folks and investors will roll their eyes, because they are interested in ROI and business models, not in analytes or nano-wires or tools to save millions of lives. Foundations and “angels” will ignore this blog due to its lack of focus. Publishers of august journals will point out that there aren’t any new advances, in this blog. Other publication outlets will ask who is the intended audience, what is the target market? There are so many, so justified, and so good reasons, why this blog was a waste of time and considered junk mail, in in-boxes.

Ian Goodfellow, answering a question from the audience, remarked that the reason why those who tend to delve into deep learning are not sufficiently informed in neurology, and those who are in neurology, are not engaged with AI or deep pursuits in computer science, due to the fact that mathematics is not a priority in biology and biology is not a priority for artificial intelligence. He went on to add, computer scientists abhor learning declarative facts and prefers developing processes from first principles. But, neurology or biology, emphasizes the learning of facts. The chasm between the groups, and the fact that we may be at the level of arthropods (insects, bees), in terms of understanding neural networks, makes the process of AI and representational learning almost entirely, and precariously, dependent on trial and error. The science of AI is still an art.

Figure 19: The panel on the left attempts to fit data and reveals, once again, that common sense is uncommon. Extending that idea to neural networks, the right panel, shows how artificial neural networks (ANNs) are highly vulnerable due to their linear nature. In this example[59], applying a very small perturbation (+ 0.007) to the dataset (panda image recognition with 58% confidence), can dramatically change the output and emphasize an incorrect answer with very high confidence (image suggests gibbon with 99% confidence). This is due to the linearity of ANN, which is the principal tool for ML, DL and AI. Linearity is an artifact of ANN and not at an intrinsic property of neural networks. From the perspective of human neurology, humans are well equipped to “think back” and go “back” in thought and thinking, often in the presence of significant perturbation. Back propagation is difficult yet back propagation algorithms are used in training deep models.
In the ANN model, data perturbation by an input of 0.007 completely changed the output, and sharply increased the confidence of the result. Imagine what could happen to medical diagnosis! If using an AI tool for diagnosis, minor perturbations in cerebral metabolic rate (CMR) or GABA (gamma amino butyric acid, a neurotransmitter) or tissue plasminogen activator (tPA), may lead to a diagnosis of idiopathic epilepsy[51], when a normal person or animal, was suffering only from anxiety or stress due to duress. AI is useful but how can it be trusted without humans in the loop?

Figure 20: Home delivery-by-drone of nano-array kits. Arriving soon on a window-sill, near you? Measuring Arsenic in Bangladesh or Mercury[52] in Colombia or CEA in cancer, can save millions.

One expected outcome (illustrated in Figure 20), is not impossible, because we do not have to break any rules of physics, to make it happen. The barriers are due to the rules of finance and economics. An investor commented that it will take a major catastrophe in a wealthy nation to mobilize the detection tools, drive the convergence, and for industries, to seek the end-to-end solution. The fact that eleven million children die, annually, due to mainly preventable causes, is not sufficient reason to catalyze the convergence. About 70% of the deaths are attributable to six causes (diarrhea, malaria, neonatal infection, pneumonia, preterm delivery, and lack of oxygen at birth). The fact these deaths are concentrated in 10 countries, mostly, in Sub-Saharan Africa, is the fly in the ointment. If that “fly” from Bamako could fly and trigger a public health havoc in Boston, then, try to prevent angels from investing or foundations from opening the funding gates.
THIS PAGE IS LEFT BLANK
Healthcare

Shoumen Datta
Julian Goldman

This is the true joy in life, the being used for a purpose you consider a mighty one, the being a force of nature rather than a feverish, selfish clod of ailments and grievances complaining that the world will not devote itself to making you happy.

George Bernard Shaw
Digital Transformation of the Healthcare Value Chain: Emergence of Medical Internet of Things (MIoT) and an Integrated Clinical Environment (ICE) as a Platform

Table of Contents

Cover Page 1
Table of Contents Page 2
Abstract Page 3
Background Pages 4-7
In Praise Of Page 8
Introduction Pages 9-11
Problem Space Page 12
Solution Space Page 13
Innovation Space Pages 14-15
Digital Transformation of Healthcare Page 16
Medical IoT (Internet of Things) Pages 17-24
Beyond Errors Pages 25-27
The Pragmatic Horizon Pages 28-30
Near Future Pages 30-35
Global Public Health Pages 36-39
Nano Tools Pages 40-41
Temporary Conclusion Pages 42-43
References Pages 44-57
Digital Transformation of the Healthcare Value Chain: Emergence of Medical Internet of Things (MIoT) and an Integrated Clinical Environment (ICE) as a Platform

Shoumen Palit Austin Datta, PhD
Senior Member, MIT Auto-ID Labs and Research Affiliate, Dept of Mechanical Engineering
Massachusetts Institute of Technology
Senior Scientist, Medical Interoperability Program, MDPnP Labs
Massachusetts General Hospital, Harvard Medical School
shoumen@mit.edu or sdatta8@mgh.harvard.edu (additional info http://bit.ly/IOT-MIT)

Julian M Goldman, MD
Medical Director, Biomedical Engineering, Department of Anesthesiology
Director, Medical Interoperability Program, MDPnP Labs
Partners HealthCare System, Massachusetts General Hospital, Harvard Medical School
jmgoldman@mgh.harvard.edu (additional info http://www.juliangoldman.info/)

ABSTRACT

The complexity of the healthcare ecosystem and the trans-disciplinary convergence which is essential for its function, makes it difficult to address healthcare as one domain. Divided in various stages for ease of management, healthcare is quintessentially a continuum which commences at birth and may cease at death. In this paper, we have attempted to take a broad view but it still isn’t broad enough. We have focused on issues, such as, data and interoperability, yet it scratches only the surface of these sub-domains. For the specific problem at hand, we propose an “http” approach to adopt open standards to enable medical equipment and devices to share and synthesize data using ICE (integrated clinical environment), a medical device interoperability platform. ICE may help reduce deaths due to preventable medical errors. Thinking about the future of data in personal healthcare, we propose creating an ecosystem which contributes data from multiple health related domains. Data curation and analysis of the information may boost our health related knowledge. Improving infrastructure, increasing connectivity and decision support apps for communication, may help, among other things, to uncover facts, synthesize data and observations which may influence the future of global health. In our plight for lofty ideas, we must not forget that 11 million children die each year and about 70% of the deaths are attributable to six causes (diarrhea, malaria, neonatal infection, pneumonia, preterm delivery, and lack of oxygen at birth). The majority are preventable. Health care at homes and huts may profit from systems integration and mobility due to internet of things (IoT), especially, medical IoT.
BACKGROUND (US-CENTRIC)

The legendary editor of the august New England Journal of Medicine, Arnold S Relman (1923-2014), offered an incisive perspective when he delivered the 171st Annual Discourse at the Massachusetts Medical Society on May 21, 1980 [1].

“According to an article in the Wall Street Journal of December 27, 1979, the net earnings of health-care corporations with public stock shares rose by 30 to 35 per cent in 1979 and are expected to increase another 20 to 25 per cent in 1980. A vice-president of Merrill Lynch appeared a few months ago (prior to this article) on "Wall Street Week," the public television program, to describe the attractions of health-care stocks. According to this authority, health care is now the basis of a huge private industry, which is growing rapidly, has a bright future, and is relatively invulnerable to recession. He predicted that the health business would soon capture a large share of the health-care market and said that the only major risk to investors was the threat of greater government control through the enactment of comprehensive national health insurance or through other forms of federal regulation” [1].

The sluggish progress of the US national health insurance initiatives [2] and the pugilistic powers in action [3] are indicative of the pugnacious influence of private health-care industries on national health policy. A broad national health-insurance program, with the inevitable federal regulation of costs, is an anathema to the medical-industrial complex, just as a national or global disarmament policy may be to the military-industrial complex. President Eisenhower was concerned about the unhealthy aspirations of the military-industrial complex and emphasized “we must guard against the acquisition of unwarranted influence” in his farewell address [4]. A similar admonition is applicable to the healthcare industry. Today’s medical-industrial complex is replete with aggressive enterprises vying for global dominance of epic proportions and maximizing profit taking at every step.

The transmutation of healthcare, once guided by the ethos of the Hippocratic Oath [5], to a commodity guided by the free market economy to improve efficiency and quality, is deeply flawed. Application of business values [6] and operational optimizations are uninformed efforts by well-intentioned bean counters. Patients are not “consumers” or “clients” and doctors are not “partners” or “service providers” in the classical context of Adam Smith [7]. However, business relationships and “best practices” may be applicable when a hospital or a clinic is purchasing bathroom tissue or contracting for janitorial services or outsourcing valet parking. The latter is an operational function but cannot be referred to as healthcare.

Hence, tools of operations management, for example, supply chain management, are useful only to a limited extent to things, equipment, objects or facilities, which are very likely to
be less uncertain in their use or application because the goods or inventory, usually, are uninfluenced or unrelated to the practice of healthcare. These observations, however, have had little impact on the vast cost of consulting fees extracted from the healthcare industry by major firms [8] who continue to pontificate about business best practices to healthcare.

Patients who are sick, or worried that they may be sick, generally, are neither capable of understanding their physiological status nor inclined to shop around for bargains. The value of life often far outweighs the consideration of cost. Patients and their families seek and demand the best care they can get irrespective of the price. Hence, the classic laws of supply and demand are ill suited because healthcare “consumers” may not subscribe to the usual incentives to be prudent, discriminating and frugal in their decision. This “decision” is not about profit and loss, even if glib management consultants may view it as a purchase.

Health care is neither synonymous nor interchangeable with “medical marketplace” because the canonical rules of competitive economic equilibrium [9] are unlikely to be applicable. The tireless pursuit of the for-profit business process consultants to inflect the so-called supply chain principles from the “marketplace” on to healthcare is fraught with problems. It exhibits a flagrant disregard for the foundational distinction that must choose value over cost in the practice of medicine or the effort necessary to save even a single life. Economies of scale or risk pooling are not always applicable or even desired in health care.

The root of the disequilibrium in healthcare is the heavy, often total, dependence of the patient (irrationally referred to as the consumer) on the medical practitioner (nurse or physician, viewed in the business framework as a service provider).

The business consultants, administrators and software packages peddling the operational principles extracted from retail stores or grocery chains or manufacturing plants may not be fully cognizant about the seminal work by Kenneth Arrow referred to as information inequality [10]. The latter catalyzed more in depth analysis of markets with asymmetric information pioneered by George A. Akerlof, A. Michael Spence and Joseph E. Stiglitz [11].

It is interesting to observe that few business students, middle managers and corporate executives are able to connect the fact that the Forrester Effect [12] which later morphed as the Bullwhip Effect [13] is partly due to information asymmetry [14]. The resurgence of RFID [15] to ignite the digital supply chain [16] has had limited impact on the reduction of information asymmetry. The reasons may include dead weight of old technology, lack of an engineering systems approach [17], inability to evolve out of the organizational “silo” frameworks and punctuated connectivity. Taken together, productivity gains, if used as a (key) performance indicator (KPI), remains a paradox [18] while limits on our data driven prediction abilities [19] makes predictive analytics more of an art and less of a paradigm.
The incisive foresight available from the application of the principles of information asymmetry to healthcare may explain why the usual assumptions about competitive free markets do not apply. There aren’t any known mechanism to the bridge the chasm of medical knowledge between the patient and physician in order to generate “equilibrium” through the establishment of “symmetric” information. The patient does not choose the plethora of medical tests or the regimen of procedures or plan for medication. Physicians decide the course of action. Hence, it is the physician who will influence 70% or 80% of all the expenditures associated with healthcare.

The potential for financial abuse, therefore, is obvious, when private for-profit companies enter the market. Private healthcare companies can conspire to influence the decisions of the physicians to maximize profit [20]. The physician is a profit center – actively engaged in the business of the medical marketplace which is a service industry. This scenario may fit the corner service station at the intersection of Happy and Healthy, serving gas on demand.

In Goldfarb vs Virginia State Bar, the US Supreme Court [21] handed down a landmark decision that found that the business activities of professionals were properly subject to antitrust law. Today, we are dealing with increasingly caustic consequences of that decision, as astutely pointed out by Late Arnold S Relman [22] in his 101st Shattuck Lecture at the Annual Meeting of the Massachusetts Medical Society in Boston on May 18, 1991 [23].

It would be a heresy to temporarily conclude this sketchy US-centric background without mentioning the influence of physicians and the organizations they champion (for example, the almighty AMA or America Medical Association). The forbidding political landscape in US healthcare reform was shaped, in part, by the physicians and their refusal to include government health insurance in the 1935 Social Security Act. In 1945, the AMA lobbied against President Harry Truman’s proposed universal health insurance program and delivered the fatal blow. In 1965, much to the chagrin of the AMA, it met with partial defeat when President Lyndon Baines Johnson created Medicare, a federal health insurance program for the elderly, and Medicaid, a combined federal-state program for the poor. President Bill Clinton’s failure was, in part, due to the opposition party who may have bullied the US Chamber of Commerce into withdrawing its support, as published by a senior member of the healthcare reform team, Paul Starr of Princeton University [24].

The timbre of compromises, tempered success and the catastrophic victory of President Obama’s Affordable Care Act on 23 March 2010 [25] capped a struggle for US healthcare reform that commenced even before 1935. The legal name of “Obamacare” is The Patient Protection and Affordable Care Act. One cornerstone of PPACA is patient protection. This paper and our mission is focused on patient safety. Without patient safety, it may be well-nigh impossible to adequately address what is necessary to ensure patient protection.
Physicians can no longer act collectively on matters affecting the economics of practice, whether their intent is to protect the public or simply to defend the interests of the profession. Advertising and marketing by individual physicians, groups of physicians, or medical facilities, which used to be regarded as unethical and were proscribed by organized medicine, are now protected - indeed, encouraged - by the Federal Trade Commission.

Advertisements now commonly extol the services of individual physicians or of hospital and ambulatory facilities staffed by physicians. Most of them go far beyond simply informing the public about the availability of medical services. Using the slick marketing techniques more appropriate for consumer goods, they lure, coax, and sometimes even frighten the public into using the services advertised.

I recently saw a particularly egregious example of this kind of advertising in the Los Angeles Times. A freestanding imaging center in southern California was urging the public to come for magnetic resonance imaging (MRI) studies in its new "open air" imager, without even suggesting the need for previous examination or referral by a physician. The advertisement listed a wide variety of common ailments about which the MRI scan might provide useful information — a stratagem calculated to attract large numbers of worried patients whose insurance coverage would pay the substantial fee for a test that was probably not indicated.

Many respectable institutions and reputable practitioners advertise in order to bring their services to the public's attention. But in medical advertising there is a fine line between informing and promoting; as competition grows, this line blurs. Increasingly, physicians and hospitals are using marketing and public relations techniques that can only be described as crassly commercial in appearance and intent.

Before it was placed under the protection of antitrust law, such advertising would have been discouraged by the American Medical Association (AMA) and viewed with disfavor by the vast majority of physicians. Now it is ubiquitous, on television and radio, on billboards, and in the popular print media. Of course, not all medical advertising is as sleazy.
IN PRAISE OF OTHER IMPERFECTIONS

Four common approaches are (a) private-sector (USA, Switzerland), (b) national health service (UK), (c) provincial government health insurance (Canada) and (d) social insurance (France, Netherlands, Germany). Founded in 1883, by Otto von Bismarck, Germany’s Statutory Health Insurance (SHI) program insures about 99.9% of the country’s population and is administered by private not-for-profit organizations, authorized by law to wield power on behalf of payers and providers. The federal government intervenes in the interest of public goods if the broader interests of society are neglected [27]. The SHI Health Care Structures Reform Law, passed in 2012, created a third care sector called integrated ambulatory specialist care [28]. The latter (IASC) and other efforts (in 2004) are an attempt to mitigate the negative impact on integration due to previous segregation of ambulatory and hospital care delivery. The segregation dates back to a 1931 decree by Chancellor Heinrich Brüning, which granted physicians in private practice a monopoly on ambulatory care and essentially prohibited hospitals from providing outpatient care (the emergency decree was a direct result of a physician strike).

Economist Robert J Evans posits that the ethos of equality in the provincial healthcare system of Canada reflects a principle similar to equality before law [29]. The Canadian system is an imperfect hybrid of public funding (government taxes), private providers and universal coverage which is comprehensive in scope, affordable, single-payer, provincially administered yet works as a “national” healthcare system. It reflects political compromises, in face of powerful opposition [30].

The lofty idealism in the Canadian system is but a poisoned chalice to the US medical-industrial complex. Big pharma, manufacturers and insurance companies poured more than half a billion US dollars to lobby US politicians [31] and even more was invested to commission studies with “alternative facts” to advocate that a US national health insurance scheme modeled on the Canadian system would not work. An excess of a quarter billion US dollars went into the coffers of political parties [32] to convince Americans that they will get inferior health care coverage and fewer choices with a Canadian system [33].

The US political momentum toward the lowest possible common denominator is a tool to preserve corporate interests in healthcare, for example, the lucrative dialysis business [34]. Measures and medication to reduce the need, cost and deaths due to dialysis [35] are often subjected to the vagaries of obfuscation or appears to be mired in chronic controversies [36] orchestrated with Machiavellian shrewdness. Public illiteracy of science and medicine goads forward profit optimization routines. Those who know and can shed light on these unscrupulous practices are failing to sustain their opprobrium and silence fuels repetition.
**INTRODUCTION**

The ills of the US medical-industrial complex are not unique to the US. Many of the medical equipment manufacturers [37] are global. The product-driven manufacturing business model prevalent in the equipment and device sector is nuclear. The product is supposed to spell profit for companies and any attempt to share parts, feature or information about the equipment or device is viewed as detrimental for profitability. This is a general observation and not a specific behavior exclusive to the US medical-industrial complex.

This is also where medical manufacturing must distinguish itself from other industry verticals. The equipment and devices used in medical and healthcare system must be optimized for performance which, generally, may focus on one physiological function and can be used on one individual or one patient, at a time. A heart rate monitor cannot multi-task and monitor Jane and Joe, simultaneously. Hence, the principles of business (best practices, risk pooling, supply chain, economies of scale and similar operational lessons) from most other domains are relatively sterile, since the end point of the activity for the medical manufacturer is the well-being of an animal or one human being, at a time.

The human being is not the patellar disc, myocardial infarction or glomerular nephritis. A human being may have one or more of those symptoms, syndromes or afflictions but the entire person or the “whole” patient must be the focus of health care (albeit, there may exist different stages of acuity, which may or may not be applicable in this scenario).

Safety is of paramount importance. Safety is about the status of the patient, as a whole.

Surgical wizardry may save the life of a person involved in an automobile collision. Yet the post-operative patient may succumb to patient-controlled over-dose of analgesia in an ICU.

These deaths due to errors are often due to lack of integration of patient safety processes at the level of the medical devices, several of which may be attached to monitor each patient.

Each device or equipment is usually dedicated to monitor one function and can be used for one patient at a time. The physiological functions, essential for monitoring, make up a long list and each function may require a specific device manufactured by a specific corporation. Several different corporations are supplying these devices and equipment. The functional integration of data from devices which can reflect the physiological “status” of the “whole” patient at the point of care (nurse, physician) is not a competency within the domain of device manufacturers. OEM’s are not experts in physiological status integration. Hence, perhaps, it is unfair to expect device and equipment manufacturers to bear the moral burden of integration, necessary for physiological status.
The physiological “sum” of data, from different devices, taken together, can make a significant difference between saving a life or death.

Errors leading to death of patients include errors arising out of lack of patient safety processes. A key element of patient safety is related to unintegrated device-level data. Devices are designed by manufacturers to present a pre-set, time-dependent, single function data point but usually lack integration with other single function monitors. This data segregation introduces (often, fatal) medical errors by preventing physiological status updates about the patient, as a whole, in a hospital or nursing facility or in private homes.

In the US, deaths due to medical errors (Figure 1) are the third leading cause of death \[^{38}\]. That is equivalent to one or two passenger-filled jumbo jets (B747) crashing every day.

---

Fig 1: An estimated 220,000 – 440,000 patients die due to medical errors in the USA \[^{38}\]
Deaths due to medical errors appear to be less documented in OECD nations. The relative virtues of the private Swiss system may be due to its demographics (population 8 million) and economics, 4th highest GDP in the world \[39\]. UK National Health Service data about adverse incidents is shown in Figure 2 \[40\]. Grouped as accidents under the leading causes of deaths in different geographies \[41\], deaths due to medical errors are unclear \[42\].

Fig 2: UK NHS Reporting – Trusts and Number of Incidents (bars) in the National \[40\] Reporting and Learning System. NRLS is a database where adverse incidents are reported.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Israel</th>
<th>EU-15 countries</th>
<th>USA</th>
<th>Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Malignant neoplasms</td>
<td>Malignant neoplasms</td>
<td>Heart disease</td>
<td>Malignant neoplasms</td>
</tr>
<tr>
<td>2</td>
<td>Heart disease</td>
<td>Heart disease</td>
<td>Malignant neoplasms</td>
<td>Heart disease</td>
</tr>
<tr>
<td>3</td>
<td>Cerebrovascular disease</td>
<td>Cerebrovascular disease</td>
<td>CLRD</td>
<td>Dementia</td>
</tr>
<tr>
<td>4</td>
<td>Diabetes</td>
<td>CLRD</td>
<td>Dementia</td>
<td>Cerebrovascular disease</td>
</tr>
<tr>
<td>5</td>
<td>Septicemia</td>
<td>Dementia</td>
<td>Cerebrovascular disease</td>
<td>CLRD</td>
</tr>
<tr>
<td>6</td>
<td>Kidney disease</td>
<td>Accidents</td>
<td>Accidents</td>
<td>Accidents</td>
</tr>
<tr>
<td>7</td>
<td>CLRD</td>
<td>Pneumonia &amp; Influenza</td>
<td>Alzheimer’s disease</td>
<td>Diabetes</td>
</tr>
<tr>
<td>8</td>
<td>Dementia</td>
<td>Diabetes</td>
<td>Diabetes</td>
<td>Alzheimer’s disease</td>
</tr>
<tr>
<td>9</td>
<td>Pneumonia &amp; Influenza</td>
<td>Alzheimer’s disease</td>
<td>Pneumonia &amp; Influenza</td>
<td>Pneumonia &amp; Influenza</td>
</tr>
<tr>
<td>10</td>
<td>Accidents</td>
<td>Kidney disease</td>
<td>Kidney disease</td>
<td>Kidney disease</td>
</tr>
</tbody>
</table>

Table 1: Leading Causes of Death as published by the Ministry of Health, Israel \[41\]
PROBLEM SPACE – Preventable Fatalities due to lack of an Integrated Clinical Environment

A patient had a laparoscopic cholecystectomy [gall bladder removal] performed under general anesthesia. At the surgeon’s request, a plain film x-ray was taken during a cholangiogram [bile duct x-ray]. The anesthesiologist stopped the ventilator for the film. The x-ray technician was unable to remove the film because of its position beneath the table. The anesthesiologist attempted to help but found it difficult because the gears on the table were jammed. Finally, the x-ray was removed and the procedure recommenced. At some point, the surgeon glanced at the EKG and noticed severe bradycardia. The ventilator was not restarted by the anesthesiologist. The ventilator is typically paused for 20–60 seconds to prevent motion-induced blurring of the image. The patient expired [43].

Cardiac (heart) surgery typically requires the use of cardiopulmonary bypass (CPB). During CPB, the CPB machine takes over both the pumping function of the heart and the ventilation function of the lung. Therefore, during CPB, the anesthesia machine ventilator is usually not needed, and is turned off to prevent unnecessary ventilation-induced lung movement that can interfere with surgery. During this period, physiological respiratory and circulatory monitors can be turned off or their alarm signals inactivated to prevent nuisance false alarm signals. At the conclusion of the CPB period, the heart resumes pumping blood, and the CPB machine pump is stopped. Lung ventilation must be resumed prior to discontinuation of CPB to prevent circulation of non-oxygenated blood which can cause organ damage. The anesthesia/surgical team has to remember to resume ventilation and manually re-start the anesthesia ventilator. Patient injuries and deaths occur when the team forgets or delays resumption of ventilation. This is a longstanding problem that continues to occur. Immediately following CPB, the heart and other major organs can be especially susceptible to injury from poorly oxygenated blood [43].

A 49-year-old woman underwent an uneventful total abdominal hysterectomy (removal of uterus and cervix through an abdominal incision) and bilateral salpingo-oophorectomy (removal of both sets of fallopian tube and ovary). Post-operatively, the patient complained of severe pain and received intravenous morphine sulfate in small increments. She began receiving a continuous infusion of morphine via a patient controlled analgesia (PCA) pump. A few hours after leaving the PACU [post anesthesia care unit] and arriving on the floor of the general ward, she was found pale with shallow breathing, a faint pulse and pinpoint pupils. The nursing staff called a “code” and the patient was resuscitated and transferred to the intensive care unit on a respirator [ventilator]. Based on family wishes, life support was withdrawn. The patient expired. A case review affirmed death due to PCA overdose [43].
SOLUTION SPACE – Improved Patient Safety due to an Integrated Clinical Environment

An elderly patient with end-stage renal failure was administered a standard intravenous (IV) insulin infusion protocol to manage her blood glucose, but no glucose was provided (oral or intravenous). Her blood glucose dropped to 33mg/dl and then rebounded to over 200mg/dl after glucose was administered \[^{43}\]. The fluctuation is harmful for the patient.

The extreme swings of hypoglycemia and hyperglycemia should be avoided. A patient receiving IV insulin infusion must be continuously monitored for blood glucose level. An integrated platform using a workflow may automatically adjust the IV syringe pump rate to deliver the insulin according to the real-time blood glucose levels provided by the glucose monitor (mobile data dashboard also available in real-time at point of care). Integration of active device data with other patient specific data (perhaps from electronic health/medical records, EHR/EMR) may optimize the physiologically desired patient-centric maintenance of blood glucose values. The patient-specific target range integrates in the decision process the patient’s record for weight, target glucose range, typical insulin dosage range (and correction factors, if any), glucose responsiveness to meals (insulin-to-carbohydrate ratio, glucose tolerance test data) and continuous assessment of vital signs (temperature, pulse oximetry, end tidal carbon dioxide, electrocardiograph, blood pressure, respiratory rate).

Hence, an integrated system will host a physiologic closed-loop control (PCLC) system algorithm to monitor and analyze the data from multiple devices as well as relevant health records in order to use an evidence-based (perhaps intelligent) decision tool to regulate the delivery of IV insulin to maintain the blood glucose values within the clinically desired range for this specific patient. To maintain homeostatic glucose levels within the target range, the system can also change/modify the glucose infusion rate using the integrated system-hosted algorithm. The IV insulin and glucose infusion changes can be automated and the system can execute changes to maintain physiological homeostatic balance without human intervention but will issue alarm/alert concomitant with the change (to mobile phone or point of contact nurse/physician) in case the medical situation warrants a medical professional to override the automated decision of the PCLC system algorithm.

The integrated system and algorithm can also trigger a software defined acuity-based contextual alarm and alert app, if patient safety was deemed at risk either for engineering reasons (device malfunction, null data, interface error) or medical causes (patient is at high risk for myocardial infarction). Hence, the maintenance of glucose levels and/or resultant physiological issue may reach beyond the design/scope of the PCLC bounded algorithm-driven decision process. The embedded logic may no longer effectively and safely manage the patient’s (homeostatic) outcome by adjusting IV insulin dose or glucose infusion \[^{43}\].
Making sense of sensors, sensor data and reflected radio frequency waves could lead to life-saving solutions. US reports indicate that nearly half a million patients may be dying each year due to preventable medical errors. One source of error is our lack of knowledge about the physiological status of the patient. Healthcare systems are sluggish in their progress and zeal to synthesize the data in order to provide the necessary information. Taken together, we are swimming in sensors and drowning in data yet slow to integrate the systems on a common platform to alleviate the burden of death due to errors. The common notion that such integration may be prohibitive in terms of cost is not supported by studies which claim that improved patient outcomes can be delivered within acceptable cost [44].

The importance of being frugal in healthcare carries a substantially heavy moral burden but ignoring fiscal reality and affordability may not be prudent [45]. In our haste to be penny-wise, we are often pound-foolish and court disasters. Suggestions in this paper are tainted by our lack of data and scenarios from other parts of the world but it may not be entirely callous to generalize [46] that elderly individuals are often rushed to emergencies due to falls. Accidents and emergencies are hemorrhaging [116] the healthcare system of its resources. Falls lead to hip replacements and brain injuries which are cost intensive [47].

It seems reasonable, therefore, that prevention of falls should be quite high on the list of priority for individuals over 65 years of age who are covered by Medicare and Medicaid in the US. Is it? In the US alone, more than 25,000 deaths [Figure 3] were reported in 2014 due to unintentional falls. The cost of treatment for falls in the US exceeded $31 billion in 2015. Millions are treated in emergency rooms and thousands hospitalized, each year [48].

![Unintentional Fall Death Rates, Adults 65+](Image)

Fig 3: Mortality and morbidity due to unintentional falls may be preventable [48].
Prevention of falls is one example where integration \cite{49} of medicine with innovative technologies \cite{50} offers solutions. The latter is driven by principles of convergence \cite{51} through collaborations between engineering \cite{52} and medicine \cite{53}. Confluence of ideas in advancing healthcare must include the entire spectrum including prevention, primary care, accidents and emergencies, palliative centers, hospices and dying with dignity \cite{54}.

Monitoring individuals in homes, hospitals and nursing facilities using reflected (RF) radio waves \cite{55} eliminates the need for sensors. Eliminating the need to wear a tag or sensor or transponder or any object on the body is helpful to elderly people who may be forgetful or suffering from neural afflictions. Analysis of reflected radio waves may predict (Figure 4) if a person is likely to fall. Reducing the time from the instance of the fall to medical attention is key to reducing long term brain damage, hip replacements and fatalities. Today, we have the tools to prevent mortality and morbidity due to unintentional falls, albeit, in part. As soon as an event (fall) occurs, the system can connect with ambulatory services and alert care-givers \cite{56}. Integration of the patient’s existing medical data with the patient 	extit{en route} to the hospital, may reduce delays in the emergency room and improve quality of care.

The principles and practice of connectivity, when taken together with an integrated clinical environment (ICE), may enable the system to combine device data in hospitals as well as data from home, wearables, edge devices, nutrition indicators and wellness monitors. A tsunami of change is imminent in preventive care, healthcare decision support systems and the medical-industrial complex. Science and engineering \cite{57} are catalyzing tools and technologies to advance medical \cite{58} cyber physical systems \cite{59} and fueling an explosion of the medical internet of things \cite{60} which can save lives and deliver some form of healthcare to billions \cite{61}. Decision support apps in real-time may help to prevent escalation to high acuity states by enabling a temporary “bridge” to deliver limited medical function a PoC.

![Fig 4: Predict the possibility of fall by analyzing the data from reflected radio waves \cite{55}](image)
DIGITAL TRANSFORMATION OF HEALTHCARE – Connected Value Chain of Medical IoT

The grand challenge implicit under the umbrella term medical internet of things is the ability to have an open mind and rise above the restricted view of medical IoT as a new thing or technology. IoT is a digital by design metaphor. By extension, medical IoT is a design metaphor, too. Why IoT is a design metaphor may be better understood if traced back to its roots in a seminal paper [62] and recent publications, in 1991 [63] and 2000 [15]. The design principle is not limited to healthcare or the medical domain. It is applicable to a vast number of domains [64] including manufacturing [65], transport [66] and energy [67].

The digital by design paradigm is made possible due to rapid diffusion of ubiquitous connectivity. The principle of connectivity, when transformed into practice, for example, Integrated Clinical Environment (ICE), is applicable to medicine and healthcare. As defined by ASTM F2761-09 standard [43], ICE is an architectural framework (work in progress) for integrating the clinical environment. It is a step toward enabling interoperability which is key to implementation, not only for medical IoT but anywhere we can apply IoT by design.

The global appeal of IoT as a design metaphor is based on the potential of IoT to spur economic growth [68]. Since hype is the main driver for public relations, the bombastic megalomaniac projections [69] are not predictions but perhaps clues to half-truths.

What appears to be true is the ability of the IoT as a design to reduce transaction cost [70]. The robust principles of transaction cost economics [71] when viewed in the context of IoT appears to support the definitive trajectory of IoT as a conceptual tool for global digital transformation expected to reshuffle prevalent trends and create new markets [72].

Observing the mere mundane fact of comparing prices of goods using your smart phone and then opting to buy the product you choose from the comfort of your home is a form of e-commerce. It is retail IoT as well, because connectivity may allow you to explore your purchase and your trigger to purchase may be due to the embedded IoT functions.

Use of medical IoT in the prevention of falls may reduce a portion of the $31 billion direct costs for treating unintentional falls in the US. Research driven entrepreneurial innovation created economic growth [73] and even a modest (10% - 20%) reduction in direct costs will save billions for US tax-payers. Taken together, medical IoT will reduce transaction cost.

Transaction cost in the value chain of ecosystems drives the cost of creating end to end connectivity. Digital solutions are rarely vertical and engage diverse supply chain partners. Digital transformation will morph classical models and supply network strategies. Services which offer data, value, information and knowledge synthesis, are preferred over products.

shoumen@mit.edu or sdatta8@mgh.harvard.edu and jmgoldman@mgh.harvard.edu
The lack of integration of data in the context of medical devices is a major contributor to lapses in patient safety in the US [74] and worldwide. Lack of granular data from and interoperability between medical devices (attached to a patient in a hospital or from remote monitoring at home) are factors which compromise patient safety. In the absence of appropriate data analysis and synthesis, the ability to produce actionable information relevant to the collective status of the patient must depend on medical professionals to combine the data at the point of care. In this context, integration and interoperability may improve efficiency, safety and security. It is anticipated that it will lower the rates of preventable medical errors and signal a significant improvement in quality of patient care.

The grand challenge of the medical internet of things (MIoT) is to sufficiently enable the deployment of patient-centric and context-aware networked medical systems in all care environments, ranging from hospital floors to operating rooms, intensive care units to home care units. Heterogeneous devices in each care environment would effectively share data (efficiently, safely and securely) to minimize preventable errors often introduced by humans.

Mobility of medical devices between different care environments or from patient to patient presents challenges. While in transit, they must remain secure yet “open” for “discovery” by other devices in much the same way that consumer IoT accessories, for example, Amazon Echo Dot, pairs with your smartphone or the laptop when it is in the cone of connectivity. Once discovered, the devices must self-organize and form the ad hoc mesh network of devices, specific to the patient. They need to interoperate with, verify and execute safe, authorized and compliant operational profiles.

This generic modus operandi is far from the norm. To transform this simple procedure to reality, we need common or standard architectures which will optimize and prioritize the necessary balance between utility, reliability and safety with those of accuracy, security and privacy. Unless widely adopted, reference architectures, blueprints or standards are sterile, impotent and quite useless [75].

The Integrated Clinical Environment (ICE) framework, as defined by the ASTM F2761-09 standard [43] is a step toward enabling the interoperable MIoT vision. ICE is an architecture for a medical data and device interoperability platform [76]. The debate rages on about open [77] vs partially closed platforms [78]. Platform is a new word for the conceptually similar “common bus” in computers [79]. ICE is a platform where devices may connect via standard interfaces or API’s in order to contribute, select or analyze data. Actionable information relevant to the patient can be updated, shared, published or routed. OpenICE is the reference implementation of this platform vision. The medical community may access OpenIce (www.openice.info/ and www.mdpnp.org) and adapt and/or adopt for their use.
Layering ICE with a variety of engines for distributed data acquisition, data curation and data analytics (dynamic composition of engines based on the nature of data), crowd-sourcing open apps to innovate using available (de-identified) data, introducing novel algorithms for non-obvious relationship analysis (correlations) and semantic data dictionaries, are sign posts for future development. We are keen to find unknown and unchartered paths beyond the horizon. Integrating cybersecurity, security, privacy, device discovery, authentication, standards, interoperability and data de-identification [61] are a few of the more immediate tasks. Security for interconnected and dynamically composable medical systems are a priority not only because they are critical to patient safety but also because they are mandated by HIPAA, the Health Insurance Portability and Accountability Act of 1996 [80].

In most environments, physical objects – the medical devices – are likely to be mixed and matched in an ad hoc fashion to serve the patient (dynamic system composition). It may be a conceptual stretch but compare it to an assembly of Lego blocks where the patient is the “base” to which the blocks are connected, either directly (device) or indirectly (printer or screen) or remotely (wireless heart rate monitor). In the rush to create a working solution for a patient in an emergency room, medical staff may “grab” what is available at first sight depending on the need (infusion pump, sphygmomanometer, plethysmograph, ventilator).

Did they verify the security mechanisms and authorizations which are configured on these devices? How will they support automatic verification that the system components being used are as intended in the correct clinical context? How will they determine that the components are authentic and authorized for use in that environment? Where can they log in to confirm that the devices are still approved by the hospital's biomedical engineering staff and certified to meet regulatory safety and effectiveness requirements?

In the long term future, as soon as a staff touches a medical equipment or a device, an automated association will be instantiated on her/his mobile device and/or in a control device which will serve as a digital footprint connecting a specific human with a specific object (for example, a baggage handler wearing a RFID and sensor embedded hand glove picks up a piece of luggage tagged with a “smart” label). The association can extend to other humans (in this case, nurses, physicians, patients) or objects (equipment, devices). In this manner, the staff rushing around to gather the devices and connect them to a patient is free to do her/his job while the “digital twin” evolves [81] to capture the patient, team of medical professionals and the devices connected to the patient. Each team member will receive a copy of the “digital twin” on their smart phones. The questions in the preceding paragraph will be digitally queried and unsatisfactory outcomes could trigger alarms or alerts (for example, app triggers message and the app is activated for all individuals in the decision support value chain to help change device or charge equipment or re-boot EKG machine).
The “digital twin” will capture the network of activities in progress in real-time, monitor for errors, serve as an active virtual assistant (avatar) to the team and store a continuous log of events. Currently, none of the above is available but digital twins \cite{79} are slowly emerging.

Today, as far as medical device communication is concerned, few of the existing or proposed standards, for dynamically composed and interoperable medical devices and information systems, include sufficiently comprehensive or flexible security mechanisms to meet current and future safety needs. There are gaps between required security properties and those that can be fulfilled even by combinations of currently standardized protocols \cite{44}. Safety considerations in these standardization efforts are effectively incomplete due to a lack of appropriate security analysis.

Regulators are also noting the importance of incorporating security for safety and privacy in the medical domain. The FDA is calling for medical device manufacturers to address cyber-security issues for the entire lifecycle of the device. It begins with the pre-market stage \cite{82} from initial design through deployment and end-of-life or post-market stage \cite{83}. FDA may use these draft guidelines as a basis for clearing medical device submissions \cite{84}.

Imminent fears of regulatory hurdles, associated with security, further discourages device manufacturers from delving deeper into installing security which may be subject to approval by the FDA \cite{85} for US markets. Elsewhere, the considerations may be different.

The device-centric view of security as a company-specific endeavor may be the primary barrier in this traditional hospital-focused effort which excludes the distributed network-centric view which is rapidly evolving \cite{86}. Cybersecurity, security and privacy related to any device and its data is a patient-centric issue and no longer a company or device-centric issue. Each device and its data is critical yet one part of a larger, distributed, temporal aggregate of a network of devices, data, decisions, events and records, linked over time, to a person and/or patient. It is a time series of events. The demand for total security and pervasive cybersecurity to ensure that privacy is not punctuated remains to be analyzed.

Solutions may evolve from alliances, organizations or groups working as a team bounded by goals which must rise above the mantra of amplifying shareholder values and take into account social responsibility to promote stakeholder values. The question without a good answer is the amount of security and privacy which is enough or is sufficient or is needed. Pervasive security and cybersecurity may be an illusion and costs may be truly prohibitive. Funding for pervasive cybersecurity may kill the patient by draining away the resources necessary for medical purposes. Perhaps passengers and drivers in body armors may have a lower risk of injury or fatality in case of an automobile accident but as a society we have decided to limit our sense of safety and precautions to safety belts and air bags in vehicles.
An analogy of punctuated security and privacy is the use of a tire pressure sensor that you bought online to monitor your car’s tire pressure using an app on your iPhone. A hacker went into the control system of your BMW using the external (not factory installed) tire pressure sensor and disabled your brakes. The shield of cybersecurity was punctuated. Similar events using other IoT devices were predicted in June 2015 [87] and massive attacks were observed on 21 October 2016 [88]. Healthcare is not immune from such attacks [89].

Healthcare is no longer bound by the traditional medical institution and it is no longer in complete control of the flow of health information. Hence, security considerations must be regarded and regulated as public goods. The new and fractured state of health, wellness and emergencies are distributed problems. It cannot be solved by hospitals, alone.

How much of which data must be protected or should be kept private will change in a dynamic manner depending on personal preferences and health coverage in different nations. Travel and tourism makes it imperative that humans receive treatment when outside their countries.

Taken together, issues of interoperability, dynamic composability, variant configuration and levels of operational flexibility are all important. These factors call for different thinking about principles, infrastructures and tools which can be globally deployable yet subject to local optimizations [90], for example, the cognitive firewall (J Siegel and S Sarma).

However, no single solution can act as a panacea for all scenarios and applications. We need to balance new approaches with the old [91] in an effort to converge healthy idealism with doses of pragmatism. Fear of short term vulnerabilities may, justifiably, evoke resistance to reduce the dead weight of old technologies even though pockets of conviction exist about the value of distributed agent-based approaches [85] as a robust long term prescription for digital transformation.

The ICE framework, as defined by the ASTM F2761-09 standard [43] provides an approach for integrating heterogeneous medical devices and coordinating their activities to automate clinical workflows. Medical devices that conform to the ICE standard, either natively or using after-market adapters, may interoperate with other ICE-compliant devices regardless of their manufacturing origin.

A similar paradigm exists in the computing and communications domain, leading to an explosion of devices supporting WiFi, LiFi, USB, Bluetooth, LoPAN, LTE, 5G and other telco standards. It is useful to think of ICE architecture, ICE standard and OpenICE reference implementation as foundational as the “http” or hypertext transfer protocol [92]. The global impact of the non-proprietary, open standards based http, needs little over-emphasis. The mission of ICE is similar to that of http for medical equipment, devices and any object used for healthcare. There is a time and place to exercise protectionism. This is not one of them.
ICE standard will enable dramatic improvements to patient safety because cross-vendor inter-device communications may significantly reduce preventable medical errors. Examples include patient transfers from the Operating Room (OR) to Intensive Care Units (ICU) and reducing false alarms or deaths due to Patient-Controlled Analgesia (PCA). In both of these examples, synthesis of data from diverse range of medical devices may enable the generation of real-time actionable information from the system to a point of care physician/nurse to prevent medical errors and reduce patient mortality and morbidity.

Fig 5: General architecture of ICE and an instantiation of it in a test setup at MD PnP Lab

Fig 6: Overview - Components of the ICE System (apps can connect distributed functions).
The *ICE Network Controller* is a high-assurance middleware that forwards data or commands to or from ICE applications and devices, ensures communication related quality-of-service and is agnostic as to the intended use of the clinical apps that it may support. It also manages the discovery and connection protocol for devices that wish to connect to the system.

In view of its central role in ICE communications, it is imperative to install high-performance and context-aware security support in the ICE network controller. The major functional security requirements for the ICE network controller are:

[a] authentication mechanisms for validating the identity of devices and apps, vouching for their provenance and ICE compliance,
[b] dynamic and flexible yet user friendly mechanisms for defining and enforcing access control policies for various ICE configurations in different environments,
[c] mechanism for secure device and app discovery,
[d] secure auditing mechanism and
[e] tools to guarantee the integrity, freshness (perishability) and confidentiality of data.

Functional requirements should minimize negative impact on non-functional requirements such as performance, availability, robustness and ease of use for clinicians and developers.

Safety considerations in standardization efforts are incomplete due to a lack of appropriate security analysis. To address this, we developed a prototype of ICE based on DDS or Data Distribution Service middleware \[93\] as the ICE Network Controller (please see Figure 5).

DDS is a communications API and an interoperability standard that provides a data-centric publish-subscribe model for integrating loosely coupled real-time distributed systems. DDS is data-centric. DDS separates state \[94\] management and data distribution from application logic and supports discoverable data models. This exposes data model to communication middleware, enabling DDS middleware to reason and optimize the performance of data movement within the system. To customize run-time behavior and achieve desired performance profile, DDS allows publishing and subscribing entities to express several quality-of-service (QoS) parameters. The offered versus requested QoS requirements of the participating entities are matched before any communication can proceed.

The *ICE Supervisor* provides separation/isolation-kernel-like data partitioning and time partitioning. It makes sure that the information cannot inadvertently leak between apps and apps cannot inadvertently interfere with one another. It provides real-time scheduling guarantees that the computation in one app cannot cause the performance of another to degrade or fail. It also provides a console that allows a clinician to launch apps, monitor their progress and provide user input during app execution. The ICE Network Controller and Supervisor may be incorporated together and deployed as a standalone ICE Manager.
ICE Applications are programs that accomplish a clinical objective by interacting with one or more devices attached to the network controller. As each app executes in the supervisor, it defines the intended use of the current ICE configuration. ICE medical devices never interact directly with each other (for safety and security reasons). All interactions are coordinated and controlled via the ICE apps. It is crucial that ICE apps exactly correspond to the specified task for which they were designed.

ICE Equipment Interfaces declares the functional capabilities of the device (for example, format of its data streams, commands to which it responds) along with non-functional properties of the data such as the rate at which data elements are streamed from the device. It is crucial that ICE Interfaces (and apps) are designed with considerations for usable security, for developers and clinical end-users.

ICE Data Logger is dedicated to logging communication and other important events within the ICE Network Controller and ICE Supervisor. Data logger should also record all security related events. The data logger can be used by medical professionals to re-create the events and actions.

It may also serve to protect against institutional vilification and indemnify the professionals in case of charges of death due to malpractice or insurance claims/payments. Physiological networks can interact and cross-react in many unpredictable ways which could lead to a cascade of deleterious events without human intervention (human is not a fault). In a very litigious society, as in the US, such incidents are often followed by “see you in court” tweets. Thus, the burden is on the medical domain to provide verifiable evidence of the sequence of events and identify each human associated with each stage of that process. Without the ICE data logger, this becomes a network reconstruction problem and network reconstruction remains an outstanding challenge [95] even in the absence of physiological complexity. ICE data logger can offer time-stamped series of the sequence of events which may exonerate the medical institution and professionals, if no human / preventable error was involved.

The concept of a cognitive supervisor and cognitive firewall has been proposed by Joshua Siegel and Sanjay Sarma (unpublished work; Joshua Siegel, PhD thesis, MIT 2016). The implementation of cognition with difficulties but the concept is perfectly logical. If the ICE Data Logger records multiple commands over a short time in a manner that might deplete a resource (battery life) or induce a DDoS (distributed denial of service) attack then the “cognitive firewall” springs into action by limiting the function or the performance or triggers an intruder detection or malfunction alert. The “cognitive supervisor” sits atop a pyramid of functions and monitors the functions or activates the cognitive firewall if and when necessary. The rationale and logic of the cognitive functions can be hard coded using known ranges and use of machine learning algorithms can extract lessons and store in a repository (memory) to use if an event appears to lie outside the hard coded boundaries.
Standards-based medical device interoperability can also provide real-time comprehensive population of a patient’s Electronic Health/Medical Record (EHR/EMR). It will permit the creation of integrated error-resistant medical systems and support advanced capabilities such as (a) automated System readiness assessment, (b) physiologic closed loop control of medication delivery, ventilation and fluid delivery, (c) decision support using machine learning and/or rudimentary artificial intelligence applications, (d) safety interlocks, (e) monitoring of device performance, (f) plug-and-play modularity to support “hot swapping” of replacement devices, (g) selection of “best of breed” components from competitive sources and (h) innovations to improve patient safety, treatment efficacy and workflow efficiency.

The vision of data synthesis and data interoperability depends on adoption of these (ICE) standards by healthcare delivery organizations and medical device manufacturers as well as endorsement by organizations (eg AMA) and regulatory bodies (eg FDA). Medical Device "Free Interoperability Requirements for the Enterprise” or MD FIRE [96] is a guide to coalesce the ecosystem necessary to create a critical mass which may initiate the steps to reduce preventable medical errors. The pursuit of ICE as a standard and interoperability between other ICE-like global standards are essential. ICE advocacy by organizations and adoption of ICE by choice or regulation will enhance patient protection and patient safety.
BEYOND ERRORS – The Opportunity and The Need

The problem of preventable medical errors is, at least in part, created by the penchant for profit. Proprietary semantic data dictionaries, locked data interfaces and upselling services are tools of the trade used by the medical-industrial complex. Adoption of open standards and open platforms [97] may partially ameliorate this dysfunction. But, the corporate need to monetize is also necessary to spur economic growth and incentivize innovation. The balance between idealism and pragmatism is crucial for our path to (ethical) profitability.

The far bigger opportunity in seeking better quality of EHR/EMR is not limited to potential for integration with data from devices and home monitors or inclusion of wearables. There are other sources of data in a fractured state or uncollected. Data from transcriptomics, genomics, proteomics, metabolomics, imaging, and pathology may reveal patterns or interplay between morphogens and morphogenesis or factors that influence immune responses [98] or eliminate pathogens [99] or analytes which predict heart failure [100].

This vast resource of data, if collected, combined and analyzed (if made available from an open repository) will aid systems biology to uncover clues cryptic in networks of proteins or genetic circuits or cellular signals. These signals connect, converge and collude in a myriad of ways to promote health or cause disease. Elucidating the elusive paths taken by these target molecules may be one part of the future progress of precision medicine [101]. The toilet in our bathroom may become our best ally [61] in preventive medicine.

Transforming these target molecules to prognostic tools may open the Pandora’s Box on population genomics and populations at risk. Global medicine must find ways to address health and healthcare issues for billions. Moving from expensive micro-arrays to very low cost nano-sensor [102] arrays or nano-wires [103] to detect single molecules may hold promise for diagnostics. However, the ability to detect is only a part of the equation. Reusability is a key driver to reduce cost. Analyte association is as important as dissociation to replenish and recycle the detection tool. The time, duration and circadian rhythm of the signal is pregnant with information. An “always-on” transmission is required to capture any signal, anytime. The form factor and stable vs unstable location (topical, epidermal, subdermal, subcutaneous, ingestible) will influence efficacy, accuracy and adoption. Invention [104] and innovation of new technologies [105] may be necessary to detect, transmit and capture signals in vivo.

Data leading to diagnostic tools and data logging of transmitted signals to identify patterns will need a variety of tools as well as new thinking and re-visiting old ideas. We will need
data curation, pattern analysis, non-obvious relations, chemistry, engineering, computer science, nanotechnology, telecommunications, software, system of systems and medical cyber physical systems. Confluence of transdisciplinary knowledge domains will be essential to synthesize the outcomes and coordinate the compass with the road map. This is yet another dimension of the anticipated digital transformation in medicine. Medical IoT is a part of this forthcoming tsunami from ubiquitous connectivity. The confluence of molecular connectivity with 3D/4D printing and distributed additive manufacturing-on-demand facilities may deliver medical tools to benefit billions. The new business of billions may flourish if rooted in the concept of micro-payments for services.

Current cost of hip replacement in India is about US$5,000 but at a per capita annual income of $1500, nominal ($5350 by purchasing power parity), it is still out of reach for many Indians (even though the cost is about one tenth of that in US). If this product was transformed into a service and distributed over the lifecycle of the implant (say 10 years) then the daily micro-payment amounts to $1.40 (or more depending on cost of service). For 5.6 billion people in the world surviving on less than $10 per day, the amount spent on healthcare will be minimal. Can the demographics, who need the hip replacement the most, afford to bear the cost of $1.40 per day for 10 years if they earn about $10 per day?

![Table 2: If 1% of the population of India and China could afford hip or knee replacement then the demand for hip & knee implants (including US & EU) may exceed 35 million [106]](image-url)
Advances in material science, chemistry of metal additive printing and lowering the capital expenses, related to the manufacturing infrastructure, may converge to lower the cost of the implant. Surgical costs and post-operative costs are, however, significant factors which may not be easily reduced. Spinal implants may need teams of orthopedic surgeons as well as neuro-surgeons. Patient-specific 3D printed implants will need access to imaging tools (MRI), design labs and technicians to produce the implant with a patient-specific fit.

Using an array of sensors to derive data from these implants (for example, for hip joints, femoroacetabular impingement) opens up yet another horizon. To go one step further, we can create a cavity within the implant and use the implant as a Trojan Horse to deliver a variety of molecular tools packed in that cavity. The challenge is to power the tools inside the implant for the lifecycle of the implant and communicate or extract digital signatures, data and signals from inside the implant. The external surface of the implant may serve as an internal “skin” embedded with nano-wire sensors. These nano-wires may serve as docking zones for bio-markers \[107\] or other analytes, which, when bound, elicit signals. Similar set of issues are applicable in identifying the signal over noise, replenishing the nano-sensor, capturing and transmitting the signal. In addition, the impact of angiogenesis and vascularization on the function of these nano-sensors remains to be determined.

Capturing signals released by our bodies may generate a wealth of data, for example, the temperature of our body (form of radiation/convection). Neural activity inside our bodies generate data we generally do not capture or analyze. This data may offer information about the state of neuromuscular or neurodegenerative diseases, such as ALS \[108\].

The field of terahertz radiation and imaging is yet another source of data which remains unexplored. Protein electrodynamics \[109\] suggests that proteins act as biological radios \[110\]. Proteins emit and absorb terahertz (THz) radiation \[111\]. Hence, protein signatures may be dynamic or may change, theoretically, depending on the stage of synthesis, modification or mutation. Hence, a potential for comparative analysis of protein signatures to identify modifications or mutations of protein structure. These changes may be implicated in the etiology of disease(s) or dysfunction. Detection of such signals may alter diagnosis and prognosis of disease states, many of which are not sufficiently expressed unless they reach a certain threshold or activation point. Problems pertaining to signal vs noise appears to corrupt data from terahertz imaging \[112\]. One wonders if innovative application of error correction techniques from information theory \[113\] or econometrics \[114\] or some combination of both (or modification of these principles) may help in curating data from terahertz imaging to identify authentic signals released by proteins.

Bringing together data from a variety of domains may revolutionize medicine and trigger thoughts that we don’t know how to think, yet. These suggestions may not be correct.
It is to “dare to propose” ideas \(^{[115]}\) and re-visit “dots” or renew thoughts which may be waiting to be connected or disconnected.

**THE PRAGMATIC HORIZON – The Influence of Digital Transformation on Retail Healthcare**

The tsunami of the principles and practice of connectivity is expected to usher in an unprecedented era of healthcare information technology that shall be woven into the daily fabric of our lives almost through our entire life-cycle, from conception to the grave.

Healthcare-associated infections may benefit from connectivity [Figure 8] and the connectivity as a medical IoT design principles, if used, may save lives. About 1 in 25 patients gets an infection each year while receiving medical care in the US \(^{[116]}\). Estimated 25,000 deaths per year from about 1 million infections each year from US healthcare at a projected cost of $30 billion per annum. In addition to logging prescribed antibiotics and their use \(^{[117]}\) it may be necessary to track and trace individual hand-washing habits and places, beds, patients, equipment they may touch. The tools for such visibility include RFID tracking of objects and their association with events. When added to the power of mobility and apps, this information can be available in real-time and potential nodes of infection predicted or identified before the spread commences (epidemic) or gains momentum.

Fig 8: Principles and practice of connectivity coupled with RFID and medical IoT (MIoT) may help in tracking, predicting and containing Healthcare-associated Infections \(^{[118]}\)
Digital by design is the fabric integrated above [Figure 8]. Businesses may use similar fabric in an era where IoT may be the dominant design metaphor. Digital entrepreneurship and intrapreneurship is exactly what the doctor ordered. Trans-disciplinary convergence of medicine and engineering may morph the brick and mortar health departments, clinics and pharmacies from its emergency or retail outlet concept. An integrated function for health and healthcare is necessary to decrease demand for high acuity events and the cost.

IoT is poised to recalibrate almost every facet of the business world by taking advantage of the progress of connectivity. Medical IoT will influence health and non-emergency health, as well as general wellness, by connecting precision metabolomics within primary care environments fostering the practice of preventive medicine and clinical attention/action.

Retail clinics and “pharmacies” will undergo transformation to create the 22nd Century service centers for medicine, perhaps something akin to “Jiffy Lube” (Boots, Walgreens, CVS) compared to a visit to Sears Auto Center (hospital). $0.99 Dexa scan, $1.99 PSA tests and $2.99 mammograms may be done in-store when shopping for milk, bread and eggs.

The transformation will be catalyzed by pioneers who will usher in, albeit in phases, convergence of a wide variety of precision medicine tools applicable on a massive scale and harvest metabolomics data from device-agnostic, protocol-agnostic, platform aggregators (for example, ICE) which will connect to streaming data inside and outside the body (humans, animals). Predictive analytics from person-specific data will be the digital path for clinical “sense and response” system and offer prescriptive analytics. This may serve many facets of preventive medicine, non-emergency medicine but may exclude sudden trauma and ambulatory scenarios.

Retail healthcare may serve as the future point of contact for the confluence of preventive medicine, precision medicine, primary care, tele-health and remote diagnostics. Retail health industries must reform their mission from selling drugs to acquiring data, analyzing and advocating in addition to building alliances to serve individuals who are not patients. The potential of digital by design health IoT will generate business growth and generate massive revenues through pay-per-use micro-revenue schemes. It may help those in the US who are less fortunate [119] and reduce the barrier to entry in L-26 countries [120] where health spending is less than US$50 per year per person for more than 2 billion people [121].

Imagination, invention and innovation must be coupled with wireless telecommunication based remote monitoring where changes in physiological status or alerts could trigger applications via intelligent agents using functional mesh (networks) for multi-directional multi-cast communication of data, information, analytics, intelligence and streams for real time decision support or at-home care or ambulatory access depending on the "sense and response" system of systems that provide one-on-one guidance at point of contact (POC).
The retail health industry must demonstrate this concept on a large scale, to build credibility. It must create the local and global ecosystem of competencies necessary to provide the end to end value chain. It must be driven by the principles of ethical profitability and in practice adopt a micro-payments model for pay-per-event services.

Cybersecurity, trust, authorization, validation, privacy, policy, regulatory compliance and authentication may require digital ledgers, for example, blockchain-like concepts, to track, trace and secure every instance and events related to every process and nested sub-process. Often data may have long shelf-lives (perishability of medical data).

The complexity calls for a global surge of and focus on, collective entrepreneurial as well as intra-preneurial recombinant innovation. It will create new lines of business and immense economic growth but not through traditional channels and existing business models or organizational status quo.

This calls for a new organizational platform approach where credible groups lead and coalesce tools from a diverse array of providers and champion a new form of delivery.

The leadership must embody the relentless pursuit of frontiers without the fear of failure to lift the future plight of humanity through distributed medical care beyond boundaries. One must continuously re-invent to re-align with new research, new inventions, new theories, new ideas, new science, new ways to help people and new customers, locally and globally. If one thinks that any one solution or company or provider or nation holds the key then one may be suffering from disequilibrium due to an incurable ailment commonly referred to (in the medical jargon) as solipsistic bliss.

NEAR FUTURE – Prologue to Nonlinearity, Known Unknowns and Unknown Unknowns

A systems approach to preventing errors by enabling medical device interoperability depends on data synthesis from the “network” of devices. It may prevent deaths by treating the whole patient rather than responding to separate signals. Insufficient strides and displaced focus are certainly not in short supply.

One proposal in this article is to gather device data and acquire data from other sources (genomics, microbiomes, terahertz radiation) in order to combine, synthesize and analyze. The outcome may reveal one or more networks, connectivity, patterns and relationships, both obvious and non-obvious [Figure 11].

Patterns are ubiquitous - from Fibonacci in nature to galaxies, even if it is amorphous to the naked eye in very large systems.
Network(s) thus revealed may be super-imposed (sub-imposed) on other networks in systems biology – network of gene circuits [131], network of cellular signals [132], neural network of networks [133]. Network layers are part of nature [left, Figure 8]. Cascade of layers and/or networks underlie many functions, such as the OSI model and TCP/IP which is at the core of the internet [center panel, Figure 8]. System of systems may connect and orchestrate public services for towns and communities [right panel, Figure 8] aspiring to join the global league of smart cities [134].

Fig 9: Layers, Networks and Cascades – the systems approach to form versus function

Fig 10: The principle of GIS (Google Map) helped to organize patient-centric data layers
Network analytics of such diverse data may need new tools for data curation, as a first step. The relationships within and between these networks may need new description. Some are likely to be linear, perhaps a few of the IFTTT type [136] but non-linearity may be the norm.

Linear associations may have failed to extract the link between gum disease and arthritis [137]. Common gingivitis is implicated in the etiopathogenesis of autoimmune rheumatoid arthritis (RA) due to molecular mimicry. Autoantibody production is triggered by epitope spreading. Porphyromonas gingivalis, found in the oral cavity is responsible for certain forms of periodontal diseases. The \( \alpha \)-enolase from \( P. \) gingivalis and humans share 82% homology at the 17-amino acid stretch of an immune-dominant region. Thus, antibodies against \( P. \) gingivalis \( \alpha \)-enolase can recognize the homologous human \( \alpha \)-enolase and promote the production of anti-human \( \alpha \)-enolase autoantibodies. Anti-citrullinated human \( \alpha \)-enolase antibodies were correlated with antibodies to \( P. \) gingivalis \( \alpha \)-enolase in patients with rheumatoid arthritis [138].

Relationships likely to evolve out of this data and analytics effort may influence the future and economics of nations with respect to cost of healthcare.
Treatment of gingivitis using tooth floss, good brushing technique, suitable toothpaste and common sense applications of oral hygiene coupled with low cost adjunctive antibiotic treatment [139] may result in favorable response at the primary care stage. Reducing the risk of autoimmune rheumatoid arthritis reduces healthcare transaction cost, reduces cost of prescription medicine, reduces loss of wages due to disability, reduces morbidity, reduces arthritic pain and improves the quality of life as well as individual productivity which contributes to the national GDP. Ignorance is not a strength given the astronomical difference in cost of prevention due to treatment at the primary care level versus in-patient or ambulatory care.

CDC estimates [140] that about half of US adults (64.7 million American adults) have mild, moderate or severe periodontitis. In adults 65 and older, prevalence rates increase to 70% or higher [141]. Rheumatoid arthritis (RA) affects 1% (70 million) of the world population.

What if we could nearly eliminate gingivitis related autoimmune rheumatoid arthritis?

The 17 amino acid immune-dominant sequence may be subjected to genome surgery [142] using CRISPR/Cas9 tools [143] in eukaryotic systems [144] to edit the coding region (exon) of the human α-enolase gene in a manner which may alter the triplet code for amino acids in the immune-dominant stretch but the amino acid substitution will be highly conservative with respect to the human α-enolase secondary and tertiary structure. It appears that the function of the human α-enolase protein may be refractory to minor modifications [145].

The cost and impact on global mortality and morbidity from autoimmune diseases is staggering. In a few cases, the etiopathogenesis is known and the target is defined. Imagine if gene editing techniques (Cre-lox, ZFNs, TALENs, CRISPR/Cas9) could help to reduce the number of people affected by autoimmune and other diseases. The tools are at hand.

![Fig 12: Selected autoimmune diseases and number of cases per 100,000 people](image)
Detection of polyps in the GI tract using disposable ingestible cameras coupled with non-surgical removal of polyps using biomarker-doped excretable micro- or nano-drones \(^{[147]}\) may make a difference between an apparently normal life and a life with cancer surgery. The infrastructure for human-mediated surgery or laparoscopic robo-surgery may not be affordable, yet, in many parts of the world. Ingestible cameras and nano-drones may be available as FMCG retail goods, drone-delivered by Amazon or ordered via Alibaba or found in Wal-Mart or in the corner convenience store (where you also swap your solid metallic hydrogen \(^{[148]}\) USB battery stick for your electric vehicle's source of energy \(^{[72]}\)).

When combined with 3D/4D printing, we may print drones-on-demand (locally) for more specific needs, for example, precision delivery of biopharmaceuticals (3D printed insulin from peptides instead of Humulin or recombinant insulin). These drones may be powered by stomach acid \(^{[149]}\), navigated by external wireless signals and carry sensors for analytes or metabolites or bio-markers to update decision systems with data from inside the body.

The breast cancer detecting bra \(^{[150]}\), data from the toilet bowl \(^{[151]}\) and non-invasive mapping of single nucleotide polymorphisms (coding or non-coding SNP) associated with disease markers \(^{[152]}\) may offer clues to improving healthcare and reducing the cost.

Fig 13: The Grand Convergence – Integration of Distributed Data and Information \(^{[153]}\)
Making sense of this grand convergence needs new types of data gathering tools\textsuperscript{154} and a combination of new data analysis tools with classical tools, such as, machine learning. Elements of rudimentary artificial intelligence\textsuperscript{155} in higher order reasoning apps may be useful (claims of “intelligence” in AI\textsuperscript{156} are generally due to commercial exaggeration). Vast numbers and types solvers or engines must be a part of this “analytics” portfolio. The choice of the solver may be triggered by the event or data stream or context awareness or semantic relationship. The architecture of such diversely distributed analytical complexity is uncommon in the current enterprise architecture world. These engines may operate at the edge\textsuperscript{157}, some in the mist, others in the fog and at the back-end (cloud). Efficacy will be determined by the limits of tolerable latency and time sensitive networking demands.

Crowd sourcing of solvers/engines may enable apps to embed the best in precision analytics. For example, a person in Timbuktu may develop an algorithm which can combine data from blood pressure with levels of uric acid, creatine and creatinine to predict with accuracy the onset of glomerular nephritis which can compromise renal function (kidney). The next gen architecture will have mechanisms to accept, verify and incorporate that app if approved by credible groups\textsuperscript{158} or institutions where industry may invest\textsuperscript{159} in order to partner with experts\textsuperscript{160} who may create or evaluate the tools.

The type and nature of data for diagnostics and therapy may improve in granularity and enable precision measurements as we begin to incorporate nanomedicine\textsuperscript{161} in our daily practice. The myth of the “silver bullet” may come to life if we could pin nano-particulates to target cells (eg cancer) and induce programmed cell death (apoptosis). Nano-markers may serve as biological cross-hairs, to create molecular identifying mechanisms or induce cellular suicide by design, using chemical or physical triggers, to specifically destroy rogue, dysfunctional, malignant or corrupted cells.

The many layers of complexity discussed above, if resolved, will save lives. In addition, we can save millions of lives (see table\textsuperscript{162} on the right) if we improve sanitation or create simple sensors to monitor (via an app) pollution in an early warning decision system to prevent major outbreaks of epidemics and pandemics which affect billions, worldwide, each year.
GLOBAL PUBLIC HEALTH – 70% of the 11 million child deaths each year are preventable

Diarrhea is the third leading cause of death in children aged 1-59 months (above). The etiology (Table 3, below) leaves little to the imagination [162]. Unlike cancer or idiopathic neuropathy, the molecular targets for vast number of preventable yet neglected diseases are known to a large extent. The mortality and morbidity due to these diseases are immense. The control of these diseases through prevention, early diagnosis, low cost pharmaceutical therapy and real-time responses, are imminently possible and feasible.

Table 3 – Etiology of Diarrhea reveals tools to detect and prevent millions of deaths annually

The solution to the problem requires more than characterizing the molecular etiology. Solutions in the real world calls for convergence of biomedical research, medicine as an engagement platform, engineering tools, telecommunication technologies and software based decision systems integrated with clinical/healthcare platforms at the point of care or point of origin (may be a house or a hut). The challenge is to bridge this chasm in order to materialize the convergence through cross-pollination and collaboration, local and global.
About 29,000 children under the age of 5 die every day, mainly from preventable causes [163].

According to UNICEF [163] more than 70% of almost 11 million child deaths every year are attributable to only six causes: diarrhea, malaria, neonatal infection, pneumonia, preterm delivery, and lack of oxygen at birth. These deaths occur in the developing world. South-central Asia has the highest number of neonatal deaths while sub-Saharan Africa has the highest rates. Two-thirds of deaths occur in 10 countries. The majority are preventable.

Six million of the eleven million children who die each year may be saved by low-tech, evidence-based, cost-effective measures such as vaccines, antibiotics, micronutrient supplementation, insecticide-treated bed nets, improved family care and breastfeeding.

About 80% of health care in developing countries occurs in the home and the majority of children die at home, without being seen by a health worker. Hence, mobility (the mobile phone) is key to connect home health care to the professional health system to reduce child mortality. Better integration among systems that deliver basic supplies and health services include treatment for diarrhea, especially the use of oral re-hydration salts which can be administered by a parent or family member at home, as soon as symptoms appear.

Fig 14: Oral rehydration calculator - smart phone app created by Dr Eric Jorge Nelson and Farhana Haque [164]. Oral rehydration therapy can prevent 1 million child deaths each year.
The importance of clean water \[^{165}\] for consumption (below) and management of excreted water cannot be overemphasized. Water systems to sense (nano-sensors for detection), monitor, (early) warn and actuate/execute control is key to eliminating the risk due to water-borne diseases, like Guinea worm and cholera, that undermine child survival and development, reduce productivity, raise healthcare costs and depress economic growth.

Fig 15: Early Detection of Arsenic in water may adopt IoT design principle to improve lives.
The diffusion of medical IoT (MIoT) by design when integrated with developments in molecular sensing using nano technology and microbiomes \cite{166} may provide us with tools. These tools (the sensor, the data, the app, the decision support, the execution response, the treatment, the patient) may form overlapping closed-loop segments \cite{167} which are parts of networks and other system of systems. Connectivity by design between these systems is the core concept of internet of things and one paradigm of this design is medical IoT. This is not a tool or one technology. It is a convergence of medicine, engineering, communication as well as myriad of services plus patients, families, medical staff, communities and cities.

Fig 16: Bulk of the mortality due a few defined causes - diarrhea, pneumonia, malaria \cite{168}

One should not infer that the recent euphoria about IoT serves as a panacea for healthcare. It is one approach which may increase the level of first response or primary care at the point where urgently needed. Other highly ignored \cite{169} approaches include supplements of vitamin A. If taken every four to six months, vitamin A may eliminate xerophthalmia as well as reduce child mortality from all causes by as much as 23%, measles deaths by 50% and deaths from diarrhea by 33% \cite{163}. The tremendous challenges for mapping the nodes in deeply entangled networks of interactions between diverse factors should not dissuade the application of modern tools and technologies in our pursuit to improve global public health outcomes.
The powerful reach of medical IoT in medicine, in general, and in preventive medicine, in particular, is based on the premise that signals that precede physiological perturbations are often detectable years before a threshold is reached when symptoms may be obvious.

How do we know what to detect? How to detect? How do we know what we have detected? Determining the target analyte (gas molecules, small molecules, macromolecules, peptides, nucleic acids, protein fragments, biomarkers) with sufficient specificity is key. That is the domain of omics, biochemistry and biophysics. Using the analyte to create a detection mechanism that can offer high signal to noise ratio is in the domain of material science, microfluidics and nanotechnology [170]. Capturing and transmitting the signal to a gateway is in the domain of radio waves (RF), telecom systems, IPv6 (sensor to internet direct) and other protocols. Replenishing the signaling mechanism of the nanowire sensor isn't trivial.

Do we know the content of the signal that we have captured? Do we know the meaning of that data? How is the data applicable to the specific problem and the general physiology? How should we act upon receiving this data? Can this data yield actionable information? Do we need to respond to that information? Is the response time sensitive? Should we expect events in real-time? Did we use time sensitive networking systems and protocols to ensure that the data did not perish? Should we store the data or the decision based on the data?

Most of the above questions may fall in the broad category of (biomedical) data analytics. There may not be a simple answer and each answer will depend on context and content. Hence, each question and answer, if any, may be patient specific. There may not be a "supply chain" modus operandi to pool the risk or aggregate the averages or choose a function where the error term may be “assumed” to be a homoskedastic distribution.

What we have not discussed here (but referred to in another section - ICE Data Logger) is the issue of data security during acquisition, distribution and sharing between devices and domains (eg hospital, home, patient, practitioner, parents, family member, care giver).

Notwithstanding the layers of complexity that accompanies the task of obtaining a bona fide signal from a sensor, one must ask, if we may have such a sensor for the purpose?

The advances that connect detection of a signal and subsequent visualization on mobile devices are a part of the design which we may continue to refer as medical internet of things (MIoT). Indeed a smartphone-based sensing strategy was developed which employs chemi-responsive nanomaterials integrated into the circuitry of commercial near-field communication (NFC) tags to achieve non-line-of-sight, portable, and inexpensive detection and discrimination of gas-phase chemicals (eg ammonia, hydrogen peroxide, cyclohexanone, and water) at part-per-thousand and part-per-million concentrations [171].
Fabrication of multiphase emulsions with controllably reconfigurable morphologies \[^{172}\] may offer optical changes which can be detected by a smart phone qualitatively and can also be processed to quantify the amount of analyte, for example, proteins or bacteria \[^{173}\].

Fig 17: The Future of Mobile Digital Diagnostics and Health – Detection of gas molecules (L) and quantification of bacterial content from Janus emulsions (R) using any smart phone.

Millions of children die from dehydration due to watery diarrhea as a result of cholera caused by some strains of the gram negative bacteria *Vibrio cholerae*, a potent water-borne pathogen, which causes epidemics and pandemics, often annually, in the developing world.

Imagine the global diffusion of mobility in the remote parts of the world where neither electricity nor automobiles may have reached. Millions of smart phone users armed with phone-adaptable micro USB swappable “Swager” sensors and the corresponding app can test water samples wherever they are located, as often as they may choose, from as many sources as possible. If the signal indicates presence of bacteria (eg *Vibrio cholerae*) or an analyte (eg Arsenic) then the crowd-sourced information is automatically transmitted via the telecommunication infrastructure (2G, 3G, LTE) to the cloud or fog or mist (repository data center) where data, information and decision systems converge to control the spread of pathogens or warn communities with respect to sewers, sanitation and supply of water.

The task ahead of us is to connect the dots and gather the experts to focus on the outcomes.
TEMPORARY CONCLUSION

Figure 18 epitomizes the Gandhi-esque aphorism that there may be enough for our need but perhaps never enough for our greed. CEOs at the 8 of the largest publicly traded US insurance companies made $171.8 million in total compensation in 2016 \(^{[174]}\) according to statements filed with the US Securities and Exchange Commission. The CEOs’ combined 2016 realized compensation would be enough to cover the average annual premium for about 59,150 people enrolled in the most popular plan on the HealthCare.gov federal marketplace in 2016 (before financial assistance).

<table>
<thead>
<tr>
<th>CEO</th>
<th>2016 compensation/realized compensation</th>
<th>Pct change from 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Neidorff</td>
<td>$21,966,983 Centene Corp. $32,161,754</td>
<td>5.85% -26.90%</td>
</tr>
<tr>
<td>Bruce Broussard</td>
<td>$19,722,400 Humana $17,019,300</td>
<td>90.74% 257.99%</td>
</tr>
<tr>
<td>Mark Bertolini</td>
<td>$18,662,306 Aetna $41,676,887</td>
<td>8.11% 49.22%</td>
</tr>
<tr>
<td>Stephen Hemsley</td>
<td>$17,765,612 UnitedHealth Group $33,368,652</td>
<td>22.37% 66.07%</td>
</tr>
<tr>
<td>Joseph Swedish</td>
<td>$16,455,697 Anthem $17,057,940</td>
<td>20.96% 8.54%</td>
</tr>
<tr>
<td>David Cordani</td>
<td>$15,279,857 Cigna Corp. $21,990,392</td>
<td>-11.72% -55.13%</td>
</tr>
<tr>
<td>Dr. J. Mario Molina</td>
<td>$10,048,868 Molina Healthcare $3,816,395</td>
<td>-2.11% -52.48%</td>
</tr>
<tr>
<td>Kenneth Burdick</td>
<td>$9,260,080 WellCare Health Plans $4,687,059</td>
<td>18.81% 42.30%</td>
</tr>
</tbody>
</table>

Fig 18: The combined realized compensation of 8 CEOs would be enough to cover the average annual premium for almost 60,000 people enrolled in the most popular plan on the HealthCare.gov federal marketplace in 2016.
Neither the US nor the nations aspiring to mimic US-style healthcare can afford to spend at current rates [Figure 19]. Where applicable, digital transformation, if enabled, may partially reduce transaction costs. Improving quality of care and increased access to healthcare for the masses may be facilitated by including digital health and medical internet of things in the portfolio of healthcare. But, these problems cannot be solved with yesterday’s tools. Emerging tools, systems and standards must be part of the solution.

Convergence of new thinking, new technologies with new sources of data requiring new platforms, may be essential. To create this new momentum, we need ecosystems which can be formed under leadership which is both credible and can offer guidance for collaboration between different domains. The momentum must be channeled to solve complex problems which are also large, multi-national and the tsunami of data may influence evidence-based policies which may help to lift many boats, not just a few yachts. The sense of purpose of this endeavor must be integrated with global public health and a grand sense of the future.

![Fig 19: US Personal Healthcare Spending (2013) by Age and Condition Category](image-url)

Fig 19: US Personal Healthcare Spending (2013) by Age and Condition Category [175]
REFERENCES


31 https://www.opensecrets.org/lobby/indus.php?id=H


34 http://catalyst.nejm.org/the-big-business-of-dialysis-care/


Pg. 46 ● THE ICE AGE COMETH ● shoumen@mit.edu or sdatta8@mgh.harvard.edu and jmgoldman@mgh.harvard.edu


https://www.cdc.gov/homeandrecreationalsafety/falls/fallcost.html

https://www.cdc.gov/injury/wisqars/

http://www.cimit.org/web/cimit/about

http://www.massgeneral.org/cimit/about/

https://cimit.org/about


http://www.community.harvard.edu/alumniimpact/

https://en.wikipedia.org/wiki/Kalighat_Home_for_the_Dying


http://precise.seas.upenn.edu/research/medical-cyber-physical-systems/

Houssam Abbas, Zhihao Jiang, Kuk Jin Jang, Marco Beccani, Jackson Liang, Sanjay Dixit, and Rahul Mangharam (2016) Computer Aided Clinical Trials for Implantable Cardiac Devices. http://repository.upenn.edu/mlab_papers


Pg. 47 ● THE ICE AGE COMETH ● shoumen@mit.edu or sdatta8@mgh.harvard.edu and jmgoldman@mgh.harvard.edu
60 Dimitrov, Dimiter V. (2016) *Medical Internet of Things and Big Data in Healthcare.* Healthcare Informatics Research 22 (3) 156-163
http://dx.doi.org/10.4258/hir.2016.22.3.156

61 Datta, S. (2016) Medical IoT (MIoT / 03_Healthcare.pdf) *Digital by Design*
https://dspace.mit.edu/handle/1721.1/104429

http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=34057

doi:10.1038/scientificamerican0991-94
www.nature.com/scientificamerican/journal/v265/n3/full/scientificamerican0991-94.html

64 www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf


70 Coase, Ronald H (1937) *The Nature of the Firm*
www.colorado.edu/ibs/es/alston/econ4504/readings/The%20Nature%20of%20the%20Firm%20by%20Coase.pdf

www.masonlec.org/site/rte_uploads/files/Williamson%20class%20background.pdf

https://dspace.mit.edu/handle/1721.1/104429


Premarket Submissions for Management of Cybersecurity in Medical Devices, Guidance for Industry & FDA Staff (October 2014)

Postmarket Management of Cybersecurity in Medical Devices, Draft Guidance for Industry and FDA Staff (January 2016)


https://www.theregister.co.uk/2016/10/21/dyn_dns_ddos_explained/


www.w3c.org


http://www.openehr.org/downloads/ehrcomponents


101 https://obamawhitehouse.archives.gov/precision-medicine


103 http://www.advancedsilicongroup.com/what-s-new.html and http://nebula.wsimg.com/d4e65619cd3841474c5d7e3a8d8dad69?AccessKeyId=47B549337A11E84045AE&disposition=0&alloworigin=1


www.pnas.org/content/105/19/6888.full.pdf
www.pnas.org/cgi/doi/10.1073/pnas.0706336105


http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf


116 https://www.cdc.gov/winnablebattles/healthcareassociatedinfections/


118 www.cdc.gov/winnablebattles


121 www.pih.org


128 http://jwilson.coe.uga.edu/emat6680/parveen/fib_nature.htm

129 http://thebeautifulstars.blogspot.com/2011_05_01_archive.html

130 https://www.cfa.harvard.edu/~mig/


140 https://www.perio.org/consumer/cdc-study.htm


146 https://rheumatoidarthritis.net/what-is-ra/ra-statistics/

147 http://robohub.org/bioinspired-robotics-1-swarm-collectives-with-radhika-nagpal/


150 https://gigaom.com/2015/02/10/heres-a-cancer-detecting-bra-thats-better-than-a-mammogram/


158 http://centerfordigitalhealthinnovation.org/

159 http://hitconsultant.net/2014/05/29/samsung-pumps-50m-into-digital-health-initiative/


161


163 https://www.unicef.org/mdg/childmortality.html

164 http://purl.stanford.edu/sv492bk0032


174 http://www.modernhealthcare.com/article/20170427/NEWS/170429877

Chapter 8

Future Healthcare: Bioinformatics, Nano-Sensors, and Emerging Innovations

Shoumen Palit Austin Datta

Contents

8.1 Introduction ..................................................................................................................248
8.2 Problem Space .............................................................................................................250
  8.2.1 Background .........................................................................................................250
  8.2.2 Focus ...................................................................................................................252
8.3 Solution Space .............................................................................................................253
  8.3.1 Existing Electronic Medical Records Systems .....................................................253
  8.3.2 Changing the Dynamics of Medical Data and Information Flow ..........................257
  8.3.3 Data Acquired through Remote Monitoring and Wireless Sensor Network ............265
  8.3.4 Innovation in Wireless Remote Monitoring and the Emergence of Nano-Butlers ..........273
8.4 Innovation Space: Molecular Semantics ...................................................................290
  8.4.1 Molecular Semantics Is about Structure Recognition ..................................290
8.1 Introduction

Proclaiming health as a human right [1] is a platitude when healthcare still remains inaccessible to millions, even in affluent nations, or billions elsewhere. Few dispute the fact that more than 30,000 children, less than 5 years old, die every day, many suffering from preventable diseases. According to the Institute of Medicine [2] at the National Academy of Sciences, healthcare is substantially underperforming on several dimensions: effectiveness, appropriateness, safety, cost, efficiency, prevention, and value. Increasing complexity in healthcare and regulatory steps is likely to accentuate current problems, unless reform efforts go beyond financing, to foster sustainable changes in the ethos, culture, practice, and delivery of healthcare. If the effectiveness of healthcare is to keep pace with the opportunity of prognostic, diagnostic, and treatment innovation, then the design must be based on systems thinking [3]. Information technology must be structured to assure access and application of evidence, at the right time, to facilitate continuous learning, and research insights, as a natural by-product of healthcare process. We need to reengineer the development of a research-driven learning healthcare organization [4] by integrating a systems engineering approach, to keep the individual in focus, while continuously improving concepts, quality, safety, knowledge, and value.

In particular, commitment to research may be emphasized by lessons from the past [5] to catalyze a future where creative cross-pollination of diverse concepts from unconventional [6] strategic thinkers is rewarded. It is equally essential to build multidisciplinary collaborative global research teams, imbued with the true spirit of discovery [7] in basic and applied domains. These teams must be enabled to drive an entrepreneurial [8] enterprise approach to create proof of concepts. Translation of unconventional concepts into reality may be guided by innovators with horizontal global vision rather than gatekeepers who prefer to stay “in the box” and avoid risks that leadership demands. The caveat in this process is the impatience of “practical” people for unconventional concepts, but the “nail on the coffin” is often driven by political polemists who also get impatient with concepts, no matter how justified or pragmatic, because it gets in their way to sell their plans [9]. The latter may block funding or policy that may better enable the conversion of unconventional vision into reality, only limited by our imagination [10].

Yet, unconventional thinkers and business leaders [11] are largely credited for globalization [12]. Striking transformations have occurred through decision
systems and process engineering, not only in established markets but also in creating new markets [13] despite omnipresent global uncertainty [14]. These changes have reshaped political economy [15], governments [16], the service industry, and manufacturing sector, including business process [17], software, hardware, banking, retail, airline safety, automobile industry, national security, and the business of the military industrial complex [18].

Lessons from the failures and the fruits of success, enjoyed by the business world, may not be irrelevant for exploration and/or adaptation by healthcare organizations, despite the irreconcilable differences that exist in the dynamics of mechanical versus biological and social systems. The current challenges in healthcare compel a fresh view of the organization, structure, and function of the delivery and monitoring processes in healthcare, not only for the industrial world that may afford the increasing cost (macro-payments) but for global healthcare services, in a manner accessible to, as well as feasible for, the billions (micro-payments). Financial sustainability of healthcare is a key issue in the design of innovative health services. The latter evokes the paradigm of services that may be deliverable for “micro-payments” rather than the current spending that claims a double digit share of the gross domestic product of rich nations.

One of many lessons from the business world is that to survive in business, businesses must exploit the power of “now” [19], perhaps best illustrated by the surge in using real-time data, almost for everything, through the use of radio frequency identification (RFID) tools [20]. To remain profitable in an RFID-driven world, industry must “adapt or die” [21]. Real-time consumer-driven supply chain [22] dynamics also determine the speed at which the industry must change to remain competitive [23]. Hence, business models are continuously focused on data, new technologies [24], cost reduction, and the quest for new growth without sacrificing quality of product and/or service. Systems that help to maintain “everyday low cost” at Wal-Mart and efficiencies at Dell that still allow “making boxes” a profitable pursuit deserve strategic exploration and integration of germane ideas to improve sustainable healthcare delivery. Healthcare systems were not designed with scientific principles in mind [25], and the ethos of “innovate or die” [26] is not salient to healthcare providers.

While software systems, like enterprise resource planning (ERP), generate benefits for business and industry, through some degree of integration of data and automation of planning, there are few healthcare electronic medical records (EMR) systems (EMRS) that have autonomous decision-making capability. Healthcare, even preventative or wellness, if available, is still dependent on expensive human resources for data acquisition, monitoring, analysis, reporting, and follow-up. Reengineering this cost structure and hospital-centric service model is necessary. Configuring a reliable business service model for health services may meet with entrenched resistance to change but may catalyze extended healthcare for billions and may even improve healthcare quality of service.
8.2 Problem Space

8.2.1 Background

The multitude of problems and issues in management of healthcare are beyond the scope of any paper or book. Here, I briefly touch upon only one issue: acquisition of medical data to improve healthcare. However, the fundamental nature of this issue forces us to address a series of integrated problems, some of which may be textbook cases in basic physiology [27] and biochemistry [28]. Therefore, it is well nigh impossible to do sufficient and equal justice to all the interdependent processes and areas.

The naive thrust of this issue is to reduce healthcare cost but with concomitant expansion of improved healthcare services for billions. Healthcare cost reduction is a hackneyed topic of discussion, but the thrust of this chapter is to suggest solutions where emerging ideas and innovation may help expand and improve healthcare service at a cost that may be soon sustainable even by the developing countries of the world. Hence, the solution space shall deal with the issue in focus: innovation in acquisition and analysis of medical data. It is obvious that one can remain oblivious of the fact that innovation in data collection calls for innovation in tools for data collection as well as analysis of data, to extract information and knowledge that can add value to healthcare services. Due to the latter, the problem (and solution) space of this chapter is bound to evolve in multiple directions, each indicating a further line of innovation.

One general problem in global healthcare is due to the mimicry of the Western model focused on acute-care hospital-centric view of what health “care” is supposed to deliver [29]. The aphorism that “better health is inherently less expensive than worse health” is equally applicable in the West and the East [30] yet seldom practiced. In fact, the profitability of the acute-care hospital-centric Western model appears to be the preferred line of health service delivery (Figure 8.1). It reasons that the word “care” must be omitted from healthcare [31] because the hospital-centric revenue model is at odds with the “care” that health services are expected to deliver for an individual in a patient-centric view of personalized healthcare.

I hasten to add that in general, the acute-care hospital-centric model still offers appreciable services when it concerns accidents and emergencies (A&E). It is vital to respond to the challenges of uncertainty in healthcare stemming from A&E. But, the criticism surfaces when the A&E modus operandi is extended to other areas of non-A&E healthcare. The systemic efficiencies necessary to respond to A&E situations must continue to be supported. Suggestions in this chapter or elsewhere about patient-centric personalized healthcare must not be viewed as a replacement but as a realignment of the existing system that is necessary for economic transformation to keep healthcare sustainable. A&E and non-A&E approaches are not mutually exclusive, and there is a need for elements of both systems to coexist for mutual enrichment.

Another problem that has evolved over the past few decades concerns an inability or incongruent response of the healthcare community to adopt advances in
information technology in order to develop an integrated systems approach to services. The human-driven decision model practiced by the medical community is adept in maintaining and/or sequestering data, information, and knowledge, about cases and patients in paper-driven silos. It prevents the improvement and development of a learning healthcare system based on insight and experience. Critics of information technology within the medical community will defend this status quo by pointing to issues of confidentiality of patient–doctor relationship, privacy of patient data, lack of standards for medical systems interoperability, and quagmire of ethical guidelines for medical knowledge diffusion. The critics are justified in their claims. But they also remain refractive to extract and use the principles that have increased profitability in the services industry and catapulted businesses such as GE, P&G, Nokia, and Wal-Mart to luminous heights of profitability. In addition, the chasm between medical education and engineering education creates a lack of awareness of the advanced tools and information technologies that exist and the potential for innovation in intelligent decision sciences, in order to provide data protection the critics demand and the patient and doctor deserve.

The business services approach to healthcare, admittedly, raises alarms if individuals or patients are to be viewed as cost-centers with the administration trying to function as a profit-center. This view assumes a literal translation of business services to health service, which is not what the proposal calls for. As an example of successful transformation of business service efficiencies, one may cite education, a domain with distinctly different dynamics from business. The Open University in

Figure 8.1 Reporting service lines as profitable (above 80% reporting) or unprofitable (below 80% reporting). (Adapted from Deloitte Consulting, The Future of Healthcare: An Outlook from the Perspective of Hospital CEOs, New York, 2005. With permission.)
United Kingdom [33] and University of Phoenix in the United States [34] have pioneered profitable business service type approaches, integrated with extensive use of information communication technologies (ICT), to deliver education of a reasonable standard for vast number of individuals, who were unable to access traditional higher education, for a variety of reasons.

Remote access to education had humble origins in “correspondence” courses but was transformed by the growth of the ICT sector and accelerated by the diffusion of the Internet. This is a form of “personalized” education that enables individuals to move ahead in careers of their choice and contribute to economic growth. This educational–economic transformation may offer a paradigm for the healthcare industry. Of course, the academic level of OU and other similar outfits may not train an individual to be the next “big bang” theorist or lead researcher to garner a Nobel Prize. But for those elite purposes, the academic system is well prepared with its select institutions. The elite academic institutions may be analogous to the “elite” equivalent in healthcare, which may be the acute care and A&E. However, the elite model in healthcare system may exclude or restrict, on grounds of profitability or cost, the accessibility of non-A&E type preventative or early diagnostic modes of healthcare (analogous to OU and similar outfits) for the masses.

8.2.2 Focus

To reiterate, the focus of this chapter relates to innovation in acquisition and analysis of medical data to improve healthcare by expanding coverage for the masses yet deliver greater value of healthcare services. Service orientation, systems architecture, and the use of software as infrastructure depend on data sources and analytics needed from healthcare monitoring, sensing, and responding to situations. The business services concept of data, information transparency between systems, as well as data exchange and/or interoperability issues may be more complex in this context due to regulatory and security constraints. It may be quite useful if the business service type approach can also introduce some degree of automated decision making, even based on rules and workflow, for non-exceptional healthcare case management.

The next level of decision making based on acquired data with reference to standards (e.g., pulse rate, blood pressure, and normal range of blood glucose) may require some basic algorithms based on simple artificial intelligence (AI) principles that can induce a learning healthcare approach when evaluating data about a specific individual or patient. For example, if the individual is otherwise “normal” even under a higher systolic or diastolic blood pressure (BP) reading, then the analytical algorithm can learn that the deviation from the medical standard reference model (120/80 mm Hg) is not readily a cause for alarm in this specific case since the individual is physiologically “normal” even under an elevated or lower than standard BP reference data. Hence, billions of learning instances are necessary for a global model.
This very important decision, to conclude an individual or patient is normal despite a slightly aberrant standard data, must be made, under most current circumstances, by a trained medical professional. That translates to cost. Aggregated over numbers of inpatient- and outpatient-related data, these cumulative costs soon begin to destabilize the financial infrastructure. Equally and perhaps more important is the time spent by the trained medical professional to review the data and arrive at the decision. Time spent for non-exception management siphons away valuable time from exception management and patients who indeed need attention. Therefore, it is not difficult to comprehend that small changes in the healthcare process pose minimal risk, quantifiable using business tools (see Ref. [201]), yet may improve quality of service and reduce cost.

Documenting the acquired data, for example, the blood pressure reading mentioned above, is the next level that deserves exploration due to the cascade of events that this data may trigger. With a few exceptions, even in the most advanced industrialized nations, paper-based documentation is the norm [35]. Unless paper-based documentation is the exception, rather than the norm, healthcare systems may continue to be crippled from displaying their true functional potential. Given human errors of data input, the transformation of preexisting and accumulating data as well as notes and decisions poses a major challenge. Without the available information on existing patients and cases, the ability of decision systems to arrive at non-exception-management-related decisions, on these existing patients and cases, may be seriously flawed and hence may be rendered unacceptable, to help deal with existing patient management.

Thus far, we have referred to patients, but what about individuals who are not patients yet? Wellness or preventative medicine and early risk identification are critical to reduce the probability that an individual shall become a patient or need acute care or emergency attention. The acute-care hospital-centric model is largely viewed as a failure to address this broader spectrum of personalized healthcare even though it may be well equipped to deal with A&E in nations big (e.g., the United States) and small (e.g., Ireland).

8.3 Solution Space

8.3.1 Existing Electronic Medical Records Systems

Before embarking on the discussion of emerging trends and potential for innovation to address the problem focus outlined above (Section 8.2.2), it may be pointed out that medical data captured in electronic format (EMRS) exist in practice in some form or the other [36]. It may offer architectural clues or serve as a basic template (starting point) for nations beginning to grapple with the problem of creating EMRS. However, expecting the current EMRS to serve as a “best practice” or benchmark may not be prudent. There is ample room for improvement of EMR,
which is essentially a generic data aggregation platform. The evolution of future versions or variations of generic EMR platform may not bear any resemblance to current systems that are generally command-driven, archival data stores, with little, if any, analytical capabilities, such as clinical decision support (see Ref. [63]).

Since 1907, the Mayo Clinic (the United States) claims to have kept unified medical records that exist in electronic format since 1993 [37] in a single database with 5 million records including patient files, x-rays, laboratory results, and electrocardiogram (ECG) records. Since 2004, it includes data mining and pattern-recognition tools to discover relationships among specific proteins, genetic makeup, and treatment responses (see Sections 8.3.2 and 8.5.1). Information can be shared between the different geographic locations of the Mayo Clinic, and physicians can conduct a virtual consultation on any patient because the EMRs are accessible from all three sites. This may represent the primordial role of data in personalized healthcare.

The VistA system in use by the Veterans Administration (the United States) hospitals covers 150 medical centers and 1400 sites across the country [38]. Eighty-five percent of 57 million outpatient visits and almost all inpatient notes are available online through VistA. In addition, 94% of outpatient prescriptions (equivalent to 200 million 30 day prescriptions) and almost all of inpatient prescriptions are entered in VistA (EMR) directly by the prescribing clinician. A study in 2004 compared VA versus non-VA patients in 12 communities and found that VA patients scored higher on quality of care, chronic disease care, and preventative healthcare.

The scope of benefits that can be derived from EMRS, as one component in the solution space, is only limited by our imagination, but, at present, several thorny challenges remain. Unlike the Mayo Clinic records and their visibility across the three different geographic locations, the infrastructure of VA or Partners HealthCare [39] is more extensive. Patients can move between locations and their treatment can change over time. This introduces major systemic issues, as follows. Partial solutions are suggested, if applicable.

Of immediate concern is unique identification of data (see Ref. [39]). How do we uniquely identify the patient, and patient records, that may be assigned different ID numbers in different locations as well as different records of the same patient that may be numbered according to the system in operation at a given location? The critical value of unique identification that unambiguously links the individual to his/her records, irrespective of the physical location of the hospital, does not need emphasis. The sheer number of individual records (laboratory tests, x-ray, CT scan, ultrasound, physician’s notes, medication, response) multiplies over time and may present a numbering scheme dilemma that requires a solution but without reinventing the wheel or introducing yet another “new” system.

Creating unique identification is not a competency of the healthcare industry. Hence, the health services may opt for an existing identification scheme that (1) can offer vast number of unique addresses [40] that can be organized in relationships or subclasses, (2) is truly portable, (3) Internet ready, (4) already in operation, and (5) globally pre-agreed in a manner that can aid adoption by the healthcare industry.
A proposal that can address this issue of unique identification of octillions of items using the globally agreed IPv6 format is presented in a separate paper (see Ref. [40]). Support for the potential use of IPv6-based identification may soon gain momentum [41], and the need for this approach in healthcare was highlighted in problems discussed elsewhere (see testimony that follows from Ref. [38]).

Personalized unique identification in healthcare may offer a robust id solution, but its value for the patient, who may move between various healthcare units (e.g., physiotherapy, stroke care, outpatient clinic) or hospitals, local and/or global, shall remain constrained without interoperability between systems of different healthcare providers, public and private, engaged with the patient. There is little value in deploying unique identification if an individual permanently resides in one location and receives healthcare from one medical professional, in one clinic or hospital, which is entirely self-contained in all its services and without any external interaction, guaranteeing total quality healthcare for the entire life cycle of the individual.

Interoperability segues to the issue of standards. The IPv6 standard governing the unique identification scheme mentioned above will make it possible to identify the unique number (“address”) in any healthcare system anywhere in the world, because the standard has made provisions for assigning that number or address to that record, or patient, in a manner that shall remain unique over the life of an individual. It is logical to anticipate that a “number re-claiming scheme” that may be deployed to claim back and reuse dormant numbers (e.g., a patient number may be claimed back and reused every 150 years if it is assumed that a human being is unlikely to live for more than 150 years).

Standards shall prominently feature in EMRS solutions space when and if the healthcare system begins to experiment with rule-based or intelligent decision sciences to help evolve clinical decision support (see Ref. [63]). Standards or “rules of operation” are immediately necessary for the granular and diverse quality of record-keeping that commences with patient history and physical examination. It is most often carried out by the local primary care physician or general practitioner (GP), in communities where primary care may be in operation and where EMRS may be available for record keeping. Of course, the same could commence for individuals who report to the A&E.

Standard “history” data in an EMRS that is also interoperable are a complex problem that must dig deep and wide for multidisciplinary convergence to generate a working solution. It is complicated by the multitude of descriptive syntax that may be used by the patient and the written form in which the medical professional scripts the information. The natural language (mother tongue), cognitive abilities, culture, education, values, and experience of both the patient and the physician or medical professional shall color the content and context of this history document. It is apparent that existing ICT and computer systems may find it difficult, if not impossible, to extract with reasonable precision the “meaning” of the history written in words and transform them to a standard format that is medically relevant for EMRS interoperability. Even more crucial is to “understand” the importance
of the phenotypic information relative to the context of the patient and his or her genotype, which, to be of value, must refer to a reference model of the community, geography, and environment where the patient lives or lived.

To the serendipitous reader, it may become obvious that the transformation from a paper-based localized healthcare system, with a limited number of professionals from a homogeneous background, to EMRS, may help to deliver global healthcare services, by any one, the myriad of diverse professionals. But it requires substantial research and innovation to develop standards [42]. The development of these standards requires integration with cognition and semantic theory. I will return to this issue when I discuss the proposal of molecular semantics (Section 8.4).

Before “boiling the biomedical ocean” in cognition and semantics, it may be useful if the healthcare system may globally agree to partially address the gulf between syntax and semantics in a “quick and dirty” approach, as adopted by business processes in global organizations [43] or the efforts spawned by the semantic web movement [44] to create biomedical ontologies [45]. These solutions, albeit temporary, if deployed and implemented in EMR type systems, are likely to offer some benefits, if the questions and content are less descriptive and are already expressed in a manner that is medically relevant, irrespective of the “background and culture” of the involved medical professional.

The development of “quick and dirty” rules and partial standards may help with (1) exchange of clinical data, (2) defining categories or circumstances when a physician in one healthcare organization can change or amend the problem list entry of a physician in another organization, (3) conditions under which clinical staff from one organization may discontinue medication prescribed by another clinician in the same organization or from another organization, and (4) types of data and information that must be secured by privacy policies and their enforcement, so that confidential data and patient information cannot be shared with, or released to, any external nonmedical organization without due authorization from the patient.

Finally, EMRS, even with its current healthcare service handicaps and restrictions, can still provide business value that may translate to cost savings. Because organization-specific EMRS like VistA operated by the VA cover a majority of its operations, the supply-demand profile of the operation can be deduced with a fair degree of precision. Using economies of scale, products, and services may be bought at a bargain. VistA offers an indication of volume of patients that are likely to be served and that volume information may be analyzed to forecast [46] inventory of supplies necessary to meet the projected demand. For perishable products (drugs, IV fluids, food), the design and management of supply chain [47] of vendors and partners can partner with the business operations unit to ensure adequate inventory of perishables to meet “peace” time and “war” time [48] type volatilities, analogous in the healthcare supply chain if one compares normal course of events versus epidemics, pandemics, natural disasters, or acts of terrorism. However, it is well nigh impossible to stress that the key performance indicators (KPI) for business
operations and inventory management or purchasing decisions must be based on different operating principles and are significantly distinct between business [49] and healthcare (see Section 8.5.1). Nevertheless, gaining business efficiencies in healthcare is not an automatic process simply because EMRS offers an aggregated view of potential consumption of products and services. It is for this reason that some countries with national health service, whether servicing a large [50] or small [51] population, may still suffer from an inability to take full advantage of the economies of scale to drive business efficiencies.

8.3.2 Changing the Dynamics of Medical Data and Information Flow

The thesis of this chapter outlined in Section 8.2.2 selects medical data as one conduit that may offer the potential to catalyze low-cost, high-quality healthcare services. An analysis of the nodes of origin of data in the context of their relationship to cost (to acquire data) and quality of service (decision based on data) may form the basis for suggesting how the current dynamics may benefit from a paradigm shift (Figure 8.2).

![Figure 8.2](image-url)
In some industrialized nations and in the developing world, the current healthcare practice may be loosely represented by the schematic outline shown in the top portion of Figure 8.2. In the current scenario, almost all points of interaction (denoted by rectangles) incur cost (denoted by large $ signs), and each action consumes time from professionals (denoted by “HR” and “Service HR”). The latter, as a result, is a drain on the available working hours of medical personnel and is an improper use of valuable time, which could have been, otherwise, put to better use by offering health “care” service focused on the elements connected to the treatment of the patient.

The paradigm shift outlined in the lower portion of Figure 8.2 is neither new nor an innovative breakthrough. It is a strategic process engineering that is driven by common sense. It suggests how basic integration of some tools and technologies may lead to savings in cost and improvement in quality of care by focusing the time of medical professionals on the patient, rather than on standard tasks and chores. In this scheme, the lack of the large $ signs under the points of interaction (rectangles) indicates potential for cost savings. Above the points of interaction (rectangles), the small $ signs imply that implementing and maintaining these changes are not free, but the cost is lower (than the large $ signs) and return on investment (ROI) is higher if the cost of installing the systems is amortized over their productive life cycle. Less time spent by medical personnel at various stages (lack of HR and service HR) leaves more time to focus on patient care.

Three elements that drive the paradigm shift (Figure 8.2) are (1) data monitoring, addressed in Sections 8.3.3 and 8.3.4, (2) electronic data capture or EMR, discussed in Section 8.3.1, and (3) diversion in data flow. In the remainder of this current section, I will discuss the latter since it is perhaps the key in the proposed paradigm shift.

Diversion of data flow leads to an information loop that channels acquired data to flow through a global reference model and directs the outcome of analytics to medical personnel if it appears to be an exception. By concentrating on the patients that need attention (exceptions), the system is able to improve its quality of service.

Critics are eager to pose the thorny question: how reliable is the machine-based analytical process? Due to the experimental nature of the scientific process, I find it reasonable to conclude that there may not be, ever, a unanimously acceptable, complete, and absolute scientific certainty, with which anyone can predict that a machine-learning process can be guaranteed to be foolproof. Errors in medical diagnosis by medical experts are rare. Thus, neither humans nor machines are entirely infallible. It follows, therefore, that the “thorny” question is the wrong question. Attempts to find the precise answer to the wrong question have, thus far, fractured the determination of, and seriously distracted, the healthcare and decision sciences experts’ (see Ref. [63]) effort, to focus on finding an approximately reasonable answer to the right question: how much and for what type of cases can we generally depend on the analytical tools in clinical decision support?

In the United States, healthcare spending was in excess of $2.1 trillion in 2007 and projected to double to $4.2 trillion by 2016. Thus, reducing 10% of health
Future Healthcare  ■  259

services workload through monitoring and analytics may save the United States nearly half a trillion dollars [52] in a few years. A small country like Ireland may save a couple billion euro [53] per year. It may also translate to a 10% improvement in the quality of service (QoS). Financial savings may be reinvested in community primary care centers, home help for independent living, and early risk identification, for example, for diabetes.

Therefore, the information loop proposed in the paradigm shift must be created and implemented, soon, even if the construction occurs in steps or modules. Inevitable “growing pains” are expected to accompany any change of direction. The “loop” is a generic expression that involves critical sub-elements of great depth that require precision knowledge. This chapter does not provide a prescription for building the sub-elements but strives to present examples that may convey the nature of these components. It goes without saying that the loop is not an “IT” job but demands collaboration between IT and medical experts.

Figure 8.2 alludes to a Global Reference Model (GRM) where the acquired data is fed via a database, such as an EMRS. The schema indicates that the data flow through the GRM into the intelligent analytics domain. What is unclear from the illustration and requires explanation is that the suggestion positions the GRM to represent an umbrella, or collection, of multiple modular databases, each embedded with some form of rule-based analytical engine that can evaluate the incoming data (streaming data) and use conventional workflow to query its database (specific to that module) to find relationships, homologies, or discrepancies based on its own stored data, specifically with reference to and in the context of that module of the GRM.

For example, a patient with elevated temperature suffering from severe bouts of coughing undergoes exploration in an outpatient clinic. A nurse records the blood pressure and temperature, draws blood for total blood count, and, based on the ethnicity of the patient, decides to take a sputum (saliva) sample in addition to administering a Manteaux test. It is an immunological test designed to detect tuberculosis (TB), an infectious disease [54], caused mainly by the microorganism *Mycobacterium tuberculosis*.

What types of modules in the GRM can process and analyze the data from this patient? Medical experts can define the nature of the GRM sub-modules, but, for the nonmedical reader, it may be of interest to note the following. Reference for normal blood pressure, correlated to age, is common, as is temperature. Total blood count is a common standard and is easily included in a GRM module. However, the GRM module that can deal with the results of the Manteaux test requires medical details as well as environmental details that relate to the epidemiology of TB, length of habitat of the patient in geographies where TB is prevalent, immunological profile of individuals with confirmed infection by *Mycobacterium tuberculosis*. The Manteaux test is an intracutaneous tuberculin skin test usually applied on the forearm and contains tuberculin purified protein derivative (PPD) to elicit the immune response that is visible on the epidermis within 2–3 days. For a patient who was born or lived in Southeast Asia or a resident of warm humid coastal areas
in the United States, a positive Manteaux test, but measuring less than 10 mm in transverse diameter of induration, as detected by gentle palpation at 48–72 h, is not indicative of TB but rather indicates tuberculin hypersensitivity, resulting from contact with nonpathogenic environmental mycobacteria or childhood vaccination by Bacillus Calmette-Guerin (BCG), which is an attenuated strain of *Mycobacterium bovis* [55].

The above example illustrates the web of relationships that the GRM and analytics must be able to extract and transmit, to the point-of-care medical professional, using a visualization interface such as personal digital assistant or Blackberry type mobile phone. However, irrespective of the apparent complexity of the above example to the nonmedical reader, most of the data and information mentioned above are already available in several databases and are classified under variety of topics, including the obvious heading of infectious diseases. There is no need to create, de novo, any basic medical data or information database. For example, from the scenario above, the sputum sample from the patient may hold several clues for early detection and diagnosis, based on advances in saliva-based biomarkers [56]. The data from sputum analysis may be transmitted to the GRM, and it can query the SKB or Salivaomics Knowledge Base [57] to extract the information for further analysis. Several such databases exist with specialized knowledge and information, which are accessible via the World Wide Web. Unfortunately, the traditional web works as a directed graph of pages with undifferentiated links between pages. This is not conducive to the type of relationships necessary for healthcare analytics. Emerging principles from social networking may be quite helpful for healthcare service analytics (see Section 8.5.1). Blogosphere (see Ref. [10]) has a much richer network structure in that there are more types of nodes that have more kinds of relations between the nodes. Deploying the principles of blogosphere in healthcare analysis may be quite promising.

Thus, the challenge is to find new ways or tools to identify and relate the selected sources that may serve as components of GRM through a virtual amalgam. GRM requires a mechanism to search and detect the information database and then query the database depending on the case or patient under investigation. For every patient, these strands of data-dependent, symptom-dependent, or test-dependent tasks must be created, in real time, on demand, perhaps, as a higher layer integrated abstraction in the form of an application module (poor choice of word but hopefully, it conveys the concept). For example, continuing the TB scenario, the data from the Manteaux test plus the symptom of cough and the eosinophil count from blood test may serve as three variables that may trigger an ad hoc application that asks, either as single queries or collectively, “What is the potential diagnosis if the Manteaux test reveals a 10 mm transverse diameter, chesty coughs are persistent, and eosinophil count is 8%?” To execute this process, the system may create an ad hoc application-specific interface (ASI), application-specific query (ASQ), and an application-specific relationship (ASR) that can act, either alone or as a bundled application, to probe relevant knowledge repositories or databases, to extract the information. This information
may be further refined by the intelligent analytics component in the information loop or transmitted to the point-of-care (POC) medical professional.

The scenario above may be loosely suggestive of a medical example of mash-up software [58] that is gaining popularity in business services utilizing SOA or service-oriented architecture [59]. SOA is touted to help business services remain agile and adaptable to meet the competitive challenges due to volatility of consumer preferences and uncertainty stemming from globalization of the supply chain. To capture sales, these business services, on demand, in real time, create a personalized web service and display a collage of objects, optimized for consumer choice. The collage is culled from different domains or databases that may or may not belong to the business but secure content on demand through licensing. The mash-up appears seamless through the wizardry of visualization tools. The consumer views it as a web page on a computer screen or mobile phone. The view may include, for instance, a company logo (a software object), price of a product (database table format, stored as an object) aimed at a market segment (extracts clients preferences and matches with stored classes of objects, e.g., sports), and ordering information (another standard shipping and handling object) with links to track and trace details provided by a third-party logistics provider (e.g., link to FedEx site).

The analogy in the above two paragraphs may fall short of the specifics or lack the precision that experts in respective domains (medicine and information technology) may demand, but it may convey the essence of case-specific exploration, on demand, integrated to knowledge discovery from databases, or other domains, necessary for delivering healthcare analytics. Each country may create their own GRM infrastructure to optimize how the GRM may be relevant to the nation and deliver value in medical analytics, at least in cases that are simple enough to be acted upon by machine intelligence, where reasonable confidence can be placed in the decision. Determining what is “simple” may vary, widely.

Triggered by patient data input, the search function of GRM may evoke the notion of Medical Google. One difference is that the search is not the end point of GRM but is for the Google search engine [60]. The granularity of the search process implicit in GRM also differs from that of Google in its quest for knowledge databases, followed by the extraction of relevant data and/or information and/or knowledge that must be first “discovered” and then resynthesized and presented to the intelligent analytical engine.

The “intelligent analysis” referred to in Figure 8.2 is a “place holder” for multiple analytical tools that are available and may be developed in future where the “learning” ability of the tools may be a key emphasis in addition to rule-based applications. The tools, for example, artificial neural networks (ANN), originate from the domain of artificial intelligence (AI), and new algorithms may continue to evolve. Since medicine is an intensely integrated science, the network of interrelationships between medical parameters and physiological function is key to understanding health. The plethora of reasons that may offer generic symptoms, for example, fever, makes it imperative that the point-of-care (POC) physician is sufficiently aware
of the spectrum of reasons why an individual may present the symptoms of fever. Presentation of a list of reasons may stem from the use of rule-based engines that may search and compile a list. However, the value of the “list” is limited unless the context of the patient and history is taken into consideration.

Rules, to combine and select the best possible match between the list and context, may be created, but the rigidity of rule-based selection may make it less reliable if compared to a set of algorithms based on AI principles that can “learn” and forget the subtle, or not so subtle, changes in the context of the patient that may include parameters, such as age, activity, profession, environment, habitat, and nutrition. For this reason, GRM and the intelligent analysis cluster of the information loop may use data cubes [61] and components that may be country specific, nation specific, or community specific but may also draw on synergies between demographically related co-localized nations (e.g., Scandinavia or Eurasian Steppes).

Let us consider an example where Jane, 11, contracts a fever on Monday morning after a hot weekend that she spent on the beach and swimming in the ocean. On Monday, Patrick, 71, complains of fever, too. He also has chronic obstructive lung disease (COLD). The attending physician may focus on determining whether Jane may have an ear infection while gearing to treat Patrick for chest infection. But can we approach this level of interaction and perhaps a treatment suggestion without incurring the cost and time of the physician? The use of acquired data (see Section 8.3.3) and decision sciences may offer a route to savings. To execute this interaction and offer a reliable, low-risk decision or treatment suggestion, intelligent analytics views the GRM-evaluated raw data plus list of possibilities for the fever and integrates the context with patient history. It also checks the GRM pharmacology module. The system transmits the exploratory analytical sequence log and diagnosis and recommends age-appropriate antibiotics, in each case.

Although the use of artificial neural networks (ANN) and artificial intelligence-based algorithms (e.g., ant-based algorithms) was mentioned only for business services applications (e.g., mash-up), the use of AI-based agent systems [62] can provide a robust and granular system. The experimental use of AI in clinical decision making [63] and productive implementation of AI in industry [64], business [65], business-to-business exchanges [66], and other applications [67] provides confidence that agent-based systems may become the norm in healthcare, too. Due to the interrelated nature of medicine, numerous parameters must be concomitantly evaluated for any decision, and each patient must be treated as an independent instance that is specific for that patient only. Data related to the patient always remain patient specific without sharing, aggregating, or clustering data, in any form, whatsoever. For each patient, the classes of data and volume of data points are likely to be quite high. All data points must be stored and relationships evaluated for diagnosis and prognosis. Therefore, the use of data cubes and the ability of agent systems to connect between all data points through the cube-on-cube organization of data cubes may make this approach particularly essential and beneficial for the billions of instances necessary for healthcare services.
Revisiting the earlier example where a Manteaux test was administered, from an agent perspective, the applications, in that discussion, may be dissociated into single agents, each with its specific task in relation data and exploration of the web of relationships with respect to the data or task assigned to the agent. In other words, the agent that holds the results of the Manteaux test (data = 10 mm) for the observed induration is charged with the task to find out, through the medium of the GRM, the implications of the observed data (10 mm). The development of intelligent agents systems [68] is within the grasp of current technology (see Ref. [182]) and can be implemented in healthcare systems. The “bundled” higher layer abstraction of individual applications (ASI+ASQ+ASR) mentioned earlier in this section is analogous to multi-agent systems [69] that work with agents in a hierarchical fashion with higher level agents tasked to integrate the information or data from lower-level agents. It is likely that multi-agent systems (MAS) may become the workhorse of the AI-based intelligent analytics in healthcare.

AI-based agents generally use programming languages from the open source domain. Hence, agents are highly mobile and suited to query a variety of databases for knowledge discovery using open source tools, for example, RDF or resource description framework [70] and OWL or ontology working language [71]. Agents are likely to form an integral part of the emerging semantic web [72]. But, agents need special interfaces if interacting with proprietary databases. Proprietary software vendors [73] deny RDF from accessing their data dictionaries through the use of proprietary programming language, for example, the use of proprietary ABAP programming by the software behemoth SAP AG. These problems may spur novel approaches. One data transformation tool called Morpheus [74] may facilitate extraction of data from various locations by transforming them into a common format, which is then sent to “holding tank” or a repository of transformations. Morpheus, used as a browser tool, may help to find a repository transformation that GRM (Figure 8.2) may be seeking. The transformation tool may drag and drop data or information in a format used by GRM.

In another vein, the open source movement is catalyzing the diffusion of software tools that may make it easier for agents to access proprietary formats through standard application programming interfaces (API) that still preserve the proprietary nature of the system but through a “translational interface” or flat file type format exchanges data or information. If the data or information sought by the agent is secured and in need of authorization for release, agent systems are capable of exchanging proofs to provide such authorization and release the data. The mobility of agents raises important questions about data security and its implications for healthcare. It is well documented that agent-based models are more robust to ensure data security by virtue of the AI algorithms used in its construction. At present, generally, most software architecture, for example, of the type used in EMRS, depends on equation-based models that are inherently far less secure.

Taken together, agent-based architecture may soon become pervasive in emerging healthcare systems. Agent-based software may form part of the infrastructure
of the healthcare system of systems (HSOS). HSOS not only consists of the components illustrated in Figure 8.2 but extends to include mobile agents that (1) monitor the functional status of medical devices, (2) schedule human resources, (3) aid in the planning of meal services to match nutritional needs or restrictions of patients, (4) guide business functions to benefit from economies of scale, and (5) oversee financial records to monitor the fiscal health of the healthcare service. Other forms of conventional non-AI software (Morpheus, mash-up, SOA, ERP, web services) may coexist within the agent-based software infrastructure for routine transactions.

Agent-based “learning” systems may augment the depth and precision of data mining and pattern recognition (see Sections 8.3.1 and 8.5.1). Rule-based data mining and pattern recognition may be out-of-date soon after “new” rules are updated. The latter may be particularly relevant to healthcare tools, data structures, and analysis of parameters in diagnosis or early risk identification. Systems must continuously learn, adapt, or improve to extract and use the subtle changes that may be indicative of future disease potential or can differentiate between closely related types of anomalies. Some of these parameters may differ or be altered sometimes, albeit slightly, between populations [75]. In the context of the patient, that can impact the outcome, considerably, to prevent false positives or wrong treatment. Agent systems can continuously “learn” about changes and hence offer greater confidence in the outcome of agent-based GRM-integrated intelligent analytics (AGRMIA) compared to rule-based tools. Algorithms for relation analysis are emerging from research on social networking [76] and reality mining [77] that may be adapted for use in AGRMIA (see Section 8.5.1).

Application of pattern recognition, in one study, achieved perfect discrimination (100% sensitivity, 100% specificity) of patients with ovarian cancer, including early-stage disease [78]. The study identified subset of proteomic biomarkers using mass spectroscopy of proteomic analysis of serum from ovarian cancer patients and cancer-free individuals. Statistical algorithms analyzed the mass spectral data and selected, using random field theory, all biomarkers that were significantly different in expression levels between affected and unaffected subjects. The best discriminating pattern was chosen among all significant biomarkers by using the best-subset discriminant analysis method (Linear Discriminant Analysis). Another study along the same lines developed an algorithm employing principal component analysis followed by linear discriminant analysis on data from mass spectrometry and achieved sensitivity, specificity, and positive predictive values above 97% on three ovarian cancer and one prostate cancer dataset [79]. Detection of ovarian cancer using sensitive molecular biomarkers is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer and for women with a genetic predisposition to breast cancer due to abnormalities in genes such as BRCA1 and BRCA2 [80].

Application of remote monitoring (see next section) of body fluids using protein microarray chips [81] that can transmit data, advanced mathematical tools for biomarker data analysis, AI-based intelligent pattern recognition, and agent-based
GRM information flow (AGRMIA), if taken together, may hold promise for global healthcare. Tools, such as mass spectroscopy (MS), provide clues about molecular identities of differentially expressed proteins and peptides in body fluids or in breath [82] that may be critical for early diagnosis. Agent-based systems, operating through the GRM (Figure 8.2), can extract these types of data as well as information catalogues of biomarkers from other fields [83] and apply the knowledge to patient data, to identify risk and improve diagnosis. Hence, MS analysis of protein or peptide biomarkers [84] in body fluids using micro-fabricated miniaturized MS device [85] operating as a low-cost wireless sensor may offer a general population-based assessment of proteomic pattern technology, as a screening tool for early risk identification for several diseases, to complement lab-on-a-chip type sensors for early detection of cardiovascular diseases (CVD) [86] and carcinomas [87]. A proposed systems approach, by which mass spectroscopic data (protein and peptide biomarkers) may be compared between systems, will be explored in Section 8.4 on molecular semantics.

8.3.3 Data Acquired through Remote Monitoring and Wireless Sensor Network

The paradigm shift in Figure 8.2 illustrates that data acquisition and transmission, even if partially assisted by the use of medical devices for remote monitoring tools and information communication technologies, may reduce cost and free up time for medical professionals. In principle, few can argue about the value of this approach. In Section 8.3.2, references were made to potential for remote monitoring and sensors to improve healthcare services. In this section, I focus on one remote monitoring device. The basic strategy, from a medical device perspective, may be similar for the majority of vital measurements (data) carried out by the primary care GP or at the hospital. Security of transmitted data and unauthorized access is preventable using agents. To guarantee even more stringent data security, recent research on PUF or Physical Unclonable Functions [88] may generate unique “fingerprints” that can distinguish identical chips or IC from the same manufacturing batch, that are used in bio-sensors and other medical devices.

Remote sensing technologies are well developed [89], yet their application to noninvasive, wearable bio-instrumentation capable of wireless transmission of reliable data has only emerged in the past few years. One innovative device, the Ring Sensor (Figure 8.3), has emerged from the convergence of robust self-organizing wireless radio frequency (RF) transmission and an improved photoplethysmographic (PPG) wearable sensor to monitor vital signs [90]. The Ring Sensor minimizes motion artifacts when measuring arterial blood volume waveforms and blood oxygen saturation, noninvasively and unobtrusively, from the wearer’s finger base. Figure 8.4 shows the results from Ring Sensor monitoring of heart rate (data transmitted through a wireless sensor network) and compares the results to conventional electrocardiogram and wired finger photoplethysmograph. The latter is susceptible to motion artifacts.
Figure 8.3  Noninvasive ring sensor for wireless monitoring of heart rate. (From Rhee, S. and Liu, S., An ultra-low power, self-organizing wireless network and its applications to non-invasive biomedical instrumentation, in IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communications, West Trenton, NJ, March 13, 2002. © 2002 IEEE.)

Figure 8.4  Heart rate monitoring by conventional, wired and wireless devices. (From Rhee, S. and Liu, S., An ultra-low power, self-organizing wireless network and its applications to non-invasive biomedical instrumentation, in IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communications, West Trenton, NJ, March 13, 2002. © 2002 IEEE.)
It does not require any stretch of imagination to slip on the Ring Sensor (or a refined version of a similar device shown in Figure 8.3) on a patient’s finger to monitor key vital signs, such as heart rate, continuously. Real-time streaming data under the “watchful eye” of a dedicated AI-based agent, embedded in the monitoring system or even in the Ring Sensor operating system (OS), monitor waveforms in real-time. It may be similar in principle to motes with TinyDB (database) and TinyOS [91]. If the Ring Sensor “senses” reasonable deviation of the PQRST wave in the context of the patient (rather than the PQRST standard in GRM or global reference model), then it immediately “responds” by sending an alert (code blue, code red) to the PDA or mobile phone of the medical professional on duty. Reference to context is important for patients with chronic CVD in order to prevent false alarms. Patients diagnosed with myocardial infarction or angina pectoris may display a PQRST waveform that may be different from the standard GRM version, but this altered PQRST waveform may be the “patient-specific normal” waveform. The data monitoring and analysis components of the system must be able to contextualize this difference.

This distinction in the analysis of monitored data highlights the need for caution to treat this suggestion as a medical device business bonanza. The challenge is to combine the sophistication of the waveform monitoring medical device (e.g., the Ring Sensor) with patient-specific context and history under the “supervision” of an agent (intelligent analytical tools). Agents may optimize the “sense, then respond” outcome, to be transmitted in a visual format comprehensible by a consultant cardiologist as well as a student nurse, to initiate the medical professional-driven response, decision, and treatment plan, if needed.

In practice, patients with cardiovascular problems are often required to wait. Intermittent monitoring of their vital signs is possible when the student nurse or trainee gets around to the patient. Physiological events that may happen in between the human resource-dependent monitoring may well determine the long-term morbidity (stroke victims) or mortality of the patient.

In the at-home scenario, Ring Sensor type applications offer greater value. Continuous monitoring is the key to preventative care, in this case, for people with chronic CVD or individuals at a high risk of CVD that may stem from other conditions, for example, increasing blood cholesterol or obstructive pulmonary diseases. Transmitted data from continuous real-time monitoring in the home, if subjected to real-time analysis (systems located in the community or primary care center or local hospital), are likely to (1) improve the quality of life for the patient; (2) reduce health service expenses by keeping the patient out of the hospital for longer periods; (3) optimize resource planning by creating a management plan and predicting when the patient may need to visit GP or outpatient clinic for non-acute follow-up or treatment; (4) reduce health service expenses and demand on service from A&E medical professionals by decreasing the probability for acute-care emergency services that may be required by the patient, more often, without remote monitoring provisions; (5) improve the overall quality of healthcare by extending
remote monitoring not only for patients but the at-risk group, as well, by using the Ring Sensor with an ultralow power wireless device, the “i-Bean,” which is an ad hoc, self-organizing network protocol that is as simple as plugging in a wireless WiFi router in home or office (Figure 8.5) or any location that can connect to the medical analytical system through the Internet; and (6) expand the reach of healthcare services (without mortgaging the treasury) by extending low-cost remote monitoring to an otherwise healthy demographic who may volunteer to keep an eye on their cardiovascular wellness profile.

The benefits from remote monitoring are undoubtedly robust, but it is also necessary to remain cautious because, whether remote or on-site, wireless or wired, local or global, healthcare produces data that may be difficult to interpret, and lack of proper interpretation may be fatal. The role of the medical professional and human-driven decisions, even for apparently routine instances, such as blood pressure check, may, sometimes, cast a doubt if data or symptoms are too generic. Table 8.1 (in Section 8.3.4) illustrates this point by highlighting that monitoring blood pressure and recording an elevated, lower, or normal reading, in some
Table 8.1  Single Gene Diseases That Elevate or Lower Blood Pressure

<table>
<thead>
<tr>
<th>Disease</th>
<th>Mutation</th>
<th>Molecular Mechanism</th>
<th>Effect on Blood Pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucocorticoid-remediable aldosteronism</td>
<td>Duplication of genes encoding aldosterone synthase and 11ß-hydroxylase, caused by an unequal crossover</td>
<td>Ectopic expression of a protein with aldosterone synthase activity regulated by corticotropin; increased plasma volume</td>
<td>Increased</td>
</tr>
<tr>
<td>Aldosterone synthase deficiency</td>
<td>Mutations in the gene encoding aldosterone synthase</td>
<td>Defective aldosterone synthase activity; decreased plasma volume</td>
<td>Decreased</td>
</tr>
<tr>
<td>21-Hydroxylase deficiency</td>
<td>Mutations in the gene encoding 21-hydroxylase</td>
<td>Absence of circulating aldosterone; decreased plasma volume</td>
<td>Decreased</td>
</tr>
<tr>
<td>Apparent mineralocorticoid excess</td>
<td>Mutation in the gene encoding 11ß-hydroxylase</td>
<td>Absence of circulating aldosterone; decreased plasma volume</td>
<td>Increased</td>
</tr>
<tr>
<td>Hypertension exacerbated by pregnancy</td>
<td>Mutation in the ligand-binding domain of the mineralocorticoid receptor</td>
<td>Activation of the mineralocorticoid receptor by steroids lacking 21-hydroxyl groups (probably due in part to the rise in progesterone levels during pregnancy)</td>
<td>Increased</td>
</tr>
<tr>
<td>Pseudo-hypoaldosteronism type I (autosomal dominant)</td>
<td>Loss-of-function mutations in mineralocorticoid receptor</td>
<td>Partial loss of function of the mineralocorticoid receptor, impairing salt reabsorption; improvement with age and a high-salt diet</td>
<td>Decreased</td>
</tr>
</tbody>
</table>

(continued)
Table 8.1 (continued)  Single Gene Diseases That Elevate or Lower Blood Pressure

<table>
<thead>
<tr>
<th>Disease</th>
<th>Mutation</th>
<th>Molecular Mechanism</th>
<th>Effect on Blood Pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liddle's syndrome</td>
<td>Mutations in the ENaC β or γ subunit</td>
<td>Deletion of the C-terminal domain of ENaC, resulting in increased ENaC activity</td>
<td>Increased</td>
</tr>
<tr>
<td>Pseudo-hypoaldosteronism type I (autosomal recessive)</td>
<td>Loss-of-function mutations in ENaC subunits</td>
<td>Impairment of ENaC subunits, which is not ameliorated by activation of the mineralocorticoid receptor by aldosterone; no improvement with age; massive salt supplementation required</td>
<td>Decreased</td>
</tr>
<tr>
<td>Gitelman's syndrome</td>
<td>Loss-of-function mutations in the sodium–chloride cotransporter of the distal convoluted tubule</td>
<td>Salt wasting from the distal convoluted tubule, leading to activation of the renin–angiotensin system; subsequent activation of the mineralocorticoid receptor increases ENaC activity, preserving salt homeostasis</td>
<td>Normal or decreased</td>
</tr>
<tr>
<td>Bartter's syndrome</td>
<td>Loss-of-function mutations in genes required for salt reabsorption in the thick ascending loop of Henle</td>
<td>Salt wasting in the thick ascending loop of Henle leads to activation of the renin–angiotensin system and the mineralocorticoid receptor, increased ENaC activity, and relative salt homeostasis</td>
<td>Normal or decreased</td>
</tr>
</tbody>
</table>


Note: ENaC denotes epithelial sodium channel.
cases, may not identify the reason or provide diagnosis based on that data, alone. Convergence of diagnostic tools is necessary to further any decision on such diagnosis. Current tools and tests may gradually undergo changes with the emergence of personalized healthcare, made possible by sequencing of the human genome [92] and development of genomics-based tools [93] (see Section 8.3.4).

The illustration in Figure 8.5 has yet another dimension that is particularly important in global healthcare, especially for emerging and developing economies where the number of and access to medical experts are limited. The goal is to support local data analysis either with information or connectivity that may enable access to experts, opinions, and resources. Various efforts to deliver information [94] through ICT [95] have been pioneered [96]. However, there is room for further advances with diffusion of broadband and other high-speed networks to the far corners of the world. It may transform the vision where a patient in a remote village clinic in Malawi may have access to an electrocardiograph (ECG) or low-cost Ring Sensor and can transmit that data through a standard network or innovative 3G system [97] that offers mobile phone service even from airplanes, during flight [98]. In the village in Malawi, a nurse practitioner may be the only medical professional in the clinic and may be unable to decipher the ECG and hence incapable of suggesting medication. It is here that the value of the transmitted ECG data becomes obvious. Through a consultancy network, on the other end of the world, a cardiologist [99] may review the data and diagnose cardiac arrhythmia due to repolarization abnormality (clinical effect) causing Long-QT syndrome [100].

This simple example and other types of analysis, which, in addition to connectivity, may also require computational power, for example, analysis of brain activity data from magnetoencephalography (MEG), may immensely benefit from grid computing [101]. Advances in microfabrication of atomic magnetometers could enable the development of precision magnetic resonance imaging (MRI) systems for self-monitoring, in any location [102].

The striking benefits that may emerge from the trinity of grid infrastructure, remote MRI monitoring, and intelligent or expert data analysis (AGRMIA) may be appreciated in view of the fact that mental health anomalies often display signs that may resist diagnosis due to lack of adequate expression of symptoms. Individuals may even fail to recognize that something is amiss because the rate of increment in the expression of some mental health conditions may be infinitesimal and over several years or decades. It may not be uncommon that individuals may even adapt to these changes as “age relevant” rather than differentiating them as potential symptoms of a disease and demanding medical exploration or treatment. Personal MRI devices coupled with microfabricated MS may create the remote ms-MRI personal monitoring device that could work as a noninvasive wearable wireless sensor that can be placed on the head to fit as a swimming cap. This development shall unleash a new horizon in “being digital” [103] in personalized medicine. In particular, remote monitoring using ms-MRI sensors may be instrumental for early detection of biochemical changes in the brain, either sporadic or due to aging.
The immeasurable value of ms-MRI remote sensors may be best illustrated by Alzheimer’s disease. It is a condition where the activity of the choline acetyl transferase (CAT) enzyme, responsible for the synthesis of acetylcholine, shows a 60%–90% decrease [104]. Acetylcholine is a key neurotransmitter and a marker for cholinergic neurons. The ability of ms-MRI to detect and profile (using MS) biological and biochemical molecules is the driving technology to determine the shape and concentration of acetylcholine molecules and hence by extrapolation determine the activity of CAT. The biochemical identification of molecules (and in future identify differences in the structure or shape of the molecules) is critical in Alzheimer’s disease and ms-MRI is one promising tool that is amenable to work as a wireless sensor for remote monitoring. The choice for ms-MRI over conventional MRI is based on the fact that conventional MRI only identifies physical structures. Recent developments in functional imaging using MRI have created the functional-MRI (fMRI) that can identify the rate of blood flow within a physical structure or area in the brain. Hence, fMRI may be useful to monitor learning disabilities where external stimuli may fail to activate certain regions of the brain, suggesting abnormalities. However, in Alzheimer’s disease, the biochemical loss of enzyme activity of choline acetyl transferase occurs in the cerebral cortex, hippocampus, and related areas, but cell counts of the neocortex and hippocampus of patients with Alzheimer’s disease did not reveal major reductions in numbers of cholinergic neurons when compared with age-matched controls. Thus, for the purpose of early detection, individuals developing Alzheimer’s disease may appear “normal” by conventional MRI analysis since the physical structure of the potentially affected areas of the brain remains unchanged as far as the numbers of neurons are concerned. The formation of plaques in the brain of patients affected by Alzheimer’s disease may be detected by MRI. Remote monitoring using ms-MRI wireless sensors may be also applicable to other conditions related to changes in neurotransmitter-related proteins and molecules in the brain, for example, Parkinson’s disease, Huntington’s disease, and some forms of dementia.

The adoption of these developments in personal remote monitoring of mental health coupled with the ability of individuals to obtain an expert opinion by transmitting the data as well as learn about the implication, of the changes recorded over time, may help determine a medical management plan that may improve the individual’s quality of life. The analysis of data may require computational resources that may be unavailable in many locations, even in affluent nations. The ability of the transmitted data via the local network to access “medical grid”-based expert services offers immense benefits. The access to these resources through the network and medical grid services, even from the developing nations [105], can reshape the fabric of global mental health.

In addition to MRI, grid computing has the potential to add remarkable value to other forms of remote biomedical imaging systems as well as bio-telematics [106] since current operations [107] are limited to groups [108] that engage in stand-alone point-to-point systems [109] without the benefit of a platform to aggregate medical
grid type services. Bringing together this platform may be similar in effort to creating the GRM and is expected to be a part of the GRM (Figure 8.2) that may run on a grid infrastructure. Several biomedical imaging databases are in existence, and taken together, they may form a biomedical imaging platform (BIP) tethered to a proposed “bMDs” services infrastructure [110] through an accessible open format where images (data) may be uploaded from anywhere in the world and viewed (analyzed) by expert(s) who will share observations and/or deliver their interpretation or diagnosis directly to the patient or the healthcare service provider. Tools for simulation and visualization are important and significant advances [111] can be resourced. Agents embedded in the architecture may monitor, device to device (D2D), machine to machine (M2M), and device to system (D2S) or vice versa (S2D), medical data to ensure data security and privacy issues.

Progress in the AI vision of autonomic computing may gradually transform BIP, either independently or in combination with an AGRMIA type infrastructure. Inclusion of embedded intelligence may provide opinions or recommendations or diagnosis or referrals (exception management) without active human intervention. The latter should be welcome news to the aging population in Europe and Japan who wish to remain independent and live in their own homes rather than in long-term healthcare communities that can drain national healthcare resources. Nations may be prudent to explore ms-MRI type high technology medical practices and find new ways to think about diseases [112] with long-term impact and challenge the medical governance [113] to reduce cost by investing in and accelerating the convergence of medical knowledge and engineering technologies. The convergence proposed in bMDs shall include advances such as ms-MRI remote monitoring platforms and may be a part of future healthcare tapestry.

Thus, grid computing is an enabling technology for healthcare service connectivity in much the same way that grid can facilitate business services [114]. In healthcare, as in business, grid computing may provide access to on-demand computational resources (that are in a different location) for real-time data processing and analysis, through grid-based tools, such as Globus [115].

8.3.4 Innovation in Wireless Remote Monitoring and the Emergence of Nano-Butlers

Nano-butlers is a facetious term but is expected to convey an “image” to suggest that nano (small)-tools and technologies act as “small butlers” serving the demands of healthcare. (Their “fees” are also small; hence, they work for “micro” payments.) The “tongue-in-cheek” image is expected to create an awareness that the detection of biological molecules including proteins or peptides at nano levels may be critical for the identification of biomarkers that may be associated with the risk of a disease. When the investment to develop nano-detection is recovered (return on investment) from the savings from reduced acute-care responses, if may be, the actual cost of nano (small)-detection will be small enough to enable global diffusion of the
tools, which can be sustained by micro (small) payments, mimicking the concept of micro-finance that has gained global acclaim [116] for alleviating poverty in some parts of the developing world. This approach, due to the inclusion of the term “nano,” may also draw some unfounded attention [117].

Early risk identification, prior to detectable symptoms, offers the potential to develop a management plan to contain the disease or even stop it from presenting any symptoms. This approach reduces the acute-care response that may be necessary if the condition was not detected and left unattended. A&E responses and acute-care are far more expensive and often increase morbidity and mortality if the response fails to be administered in near real-time. For example, from the time of onset of a cardiac attack, there is only a 30 min window for successful administration of tissue plasminogen activator (TPA) to dissolve clot(s) if a patient with CVD or stroke is the victim of thrombosis (blood clot). It is likely that more than 30 min may elapse between recognizing that a person is having a “heart attack” and the arrival of A&E services at the location, assuming not only that the paramedics will correctly diagnose the reason (clot) but also that they will have an inventory of TPA-type drugs in their mobile unit (ambulance). It is also assumed that the person is coordinated and coherent enough to call to A&E services if the individual is alone in the location.

Prevention through preventative healthcare may require, in the above scenario, the convergence of wireless sensors with remote monitoring technologies and data-driven analytics based on biomedical research knowledge bases. Advances in understanding the basic physiological, biochemical, and molecular relationships (Figure 8.6) that contribute to heart disease [118] offer hope for rigorous early detection mechanisms. There is ample evidence both from basic biological sciences [119] and clinical research [120] that early warning signs of CVD are amenable to identification from research on biomarkers. Several biomarkers for CVD are already in the market [121], but the commercial “kit” approach is far from the innovative potential of next generation diagnostics [122].

Innovation in detection is based on the generally applicable principle that physiological systems respond to thresholds. In other words, few, if any, reactions occur in the human body or fetus without a critical mass or concentration of molecules. If allowed to reach the “threshold” only, then the “rogue” molecules may trigger a cascade of events, which may, eventually, over time, present itself as a detectable symptom. Hence, molecular identification of the biomarkers and rogue molecules is vital for noninvasive detection. Tools for detection require convergence of the knowledge from identification of molecules from biomedical research with engineering-based detection technologies to determine the number and concentration of the molecules, beginning at the single molecule [123] level or at the pico (10–12) level, but most reliably at the nano level. A medical management plan or treatment must exist to prevent the concentration of the molecules from reaching the threshold, where it may commence the cascade of events leading to a disease or symptom or precipitating a heart attack. In case of fetal diagnosis, the management issues may be complex, but the early detection of sporadic or genetic diseases of the
unborn child may offer scope for medical intervention. Foolproof identification of specific biomarkers for multi-factorial diseases such as CVD is complex. Without reliable specificity, the next generation nano-diagnostic tools may offer less value. Therefore, the healthcare industry must remain vigilant to combine advances in one field with another through parallel investments both in medicine and engineering.

To illustrate, let me revisit the case of blood pressure (BP) measurement and what diagnostic information the BP data may provide if the BP is elevated, lower, or normal, compared to the standard reference. In short, the BP data, alone, provide little diagnostic value unless they are analyzed in conjunction with existing case history or other data. One reason for this is summarized in Table 8.1. If a definitive diagnosis of the BP data is sought, in some patients, it will be necessary to identify which particular gene is affected.

Sequencing of the human genome and advances in search tools and genomic technologies [124] makes it possible to extract the DNA sequence from the Human Genome Database [125] of the genes implicated in a disease (e.g., genes that may elevate or lower blood pressure). Based on the DNA sequence, anti-sense RNA may be used to determine gene expression profile. By extrapolation from the DNA
sequence, protein-based peptide fragments can be synthesized for use in noninvasive proteomics-based microarrays, using the lab-on-a-chip wireless sensor, to detect expression levels of one or more proteins and/or their mutant variations, in body fluids, that may be involved in the etiology of the condition under investigation. Based on the gene and protein expression profile, in some cases and some diseases, the promise of gene therapy may be realized by “silencing” harmful candidate genes using RNAi [126] and snRNA techniques that are already under intense commercial exploration and have reported some degree of success [127] for future therapeutic applications.

This scenario, starting with a possible gene, followed by expression profiling data remotely monitored by a wireless sensor and potential for selective silencing therapy of the disease, is a part of the evolution likely to chart the future of individualized medicine and personalized healthcare. The value and impact of this approach may not qualify as a “disruptive innovation” [128] but may reflect the systemic lessons from the age of introduction of the electric dynamo, at the turn of the twentieth century [129], the wisdom of which may have been ignored by other emerging technologies [130].

The pragmatic credibility of this vision garners support both from current practices, albeit in part, and information based on recent research that clearly points to the need for this vision. Current practices already use one or more of the steps outlined in this strategy, for example, the use of genomic [131] and proteomic biomarkers in the diagnosis of CVD [132] and some forms of cancer [133]. But even more important support for this vision draws on seminal medical research [134] that has identified one single human gene, for low-density lipoprotein (LDL) receptor-related protein 6 (LRP6), that is involved in the etiology of a specific type of CVD, referred to as coronary artery disease (CAD). It is a leading cause of death worldwide, and early detection is essential to save lives.

CAD is commonly caused by a constellation of risk factors called the Metabolic Syndrome, the symptoms of which include hyperlipidemia, hypertension, diabetes, and in addition, osteoporosis. Analysis of LRP6, the gene identified as responsible for CAD, has uncovered a single nucleotide substitution that changes the wild-type (normal) cytidine base to thymidine. This single nucleotide missense mutation, located in the protein-coding region (exon 9) of the human gene for LDL LRP6, causes a single amino acid substitution, which inserts the amino acid cysteine to replace the normal counterpart, arginine, at codon 611 (R611C). The importance of this finding and the fundamental significance of the LRP6 gene in human evolution are further highlighted by the extremely high degree of conservation of the protein sequence from humans to amphibians, such as frogs (Xenopus laevis). Among the species surveyed, the amino acid arginine (R) is conserved from frogs to humans (Table 8.2). Substitution of arginine with cysteine (R611C) creates havoc and results in CAD and accompanying diabetes, hyperlipidemia, hypertension, and osteoporosis.
### Table 8.2  Conserved Amino Acid Arginine (R) Is Substituted in LRP6 Gene and Causes Heart Disease

|                | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Human          | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Chimp          | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Monkey         | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Mouse          | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Rat            | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Dog            | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Cow            | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Opposum        | C   | L   | Y   | F   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | F   | E   | L   |
| Chicken        | C   | L   | Y   | R   | P   | Q   | G   | L   | R   | C   | A   | C   | P   | I   | G   | L   | E   | L   |
| Frog           | C   | L   | Y   | F   | P   | Q   | G   | P   | R   | C   | A   | C   | P   | I   | G   | L   | E   | L   |
Genomic technologies can help identify R611C mutation in at-risk populations even in the fetal state since this LRP6 is transmitted as an autosomal dominant trait. Lifelong management plan for inherited genetic diseases, for example, phenylketonuria, is common. Individuals homozygous for R611C, if diagnosed early, may follow a recommended lifestyle that may enable them to enjoy normal life expectancy. The evolutionary conservation of LRP6 gene and its protein product indicates a fundamental role of this protein in physiology. Indeed, LRP6 is involved in cellular signaling pathways, disruption of which leads to a plethora of problems. Proteins and other molecules, referred to as transcription factors [135], are often conserved and fundamental to gene expression [136] from bacteria to humans. Transcription factors can alter gene expression positively and negatively [137] or may serve as secondary or tertiary targets [138]. Early detection of nonfatal mutations in transcription factors is warranted because they may also cause profound physiological disturbances and present multiple symptoms, related in scope to mutation in LRP6.

While the case for early detection of genetic diseases needs little emphasis, the need for early detection of the bulk of sporadic cases such as type II diabetes mellitus and several other disease states deserves emphasis. Globalization has also created a need for remote monitoring. To make globalization work better for the world economy [139] it is imperative to contain infectious diseases by determining the risk and at-risk factors at the origin rather than at an immigration check-point of a country. The lightning spread of SARS, from its origin in Hong Kong to Toronto, in a few days, highlights the need for remote monitoring tools and the vulnerability of current healthcare system (in most nations) that is ill-equipped for global challenges and may actually aid an epidemic or pandemic.

Epidemics, however, are no longer limited to only infectious diseases. Type II diabetes may reach nearly epidemic proportions in many nations. It is alarming in some countries, where, despite a small population [140] a high number of sporadic cases of diabetes are documented. In addition, an equally high number of projected at-risk population are acknowledged, but the latter estimates exclude the segment of population projected to be defined as clinically obese. That can potentially increase the at-risk numbers for sporadic type II diabetes but remains unaccounted by the system. Individuals with other anomalies may also have diabetes, as pointed out in course of the discussion on LRP6 where heart disease and diabetes can occur simultaneously. Recent evidence suggests that in individuals without any genetic predisposition, there may be a direct effect of elevated cholesterol level on reducing insulin production by the β (beta) cells in the pancreas [141]. This increases the risk of diabetes for obese as well as for non-obese individuals who have elevated levels of cholesterol, without genetic predisposition.

Diagnosed diabetics often monitor their blood glucose levels using over the counter kits but its use in preventative healthcare remains dubious. For diabetics, the frequency of testing using kits is weekly or daily but expert interpretation and
advice may be far in between. Let us assume that a weekly outpatient visit to the clinic for blood glucose test and insulin therapy costs the healthcare service an average of $25 in direct cost and costs the economy another $25 in indirect costs, such as the cost of time spent by patient, cost of travel and decrease in productivity due to time taken off work by patient. If only 1% of the population require weekly attention (serious diabetics), then for a hypothetical nation with a population of 5 million, there will be 50,000 diabetics requiring this attention. For 50 visits a year at $50 a visit, the attention to 50,000 diabetics for simple monitoring of blood glucose and administration of insulin, will cost the nation $125 million per year. The focus on those who need it the most aggravates the unattended conditions in other diabetics and at-risk population, driving them to seek acute-care services or catapulting them to the $50 category. However, a far greater concern is the need for inpatient services if some of the diabetic patients need hospitalization. Can the national healthcare system respond adequately [142] if only 1% of the 50,000 diabetics (that is, 1% of the 1% documented diabetics in the population) need hospital beds? For example, in Ireland, overall, there are 2.9 beds available per 1000 people [143].

Preventative remote monitoring can alter this vicious cycle of crisis, reduce cost, and improve actual care.

Apart from individuals who are obese or juvenile diabetics or those with history of genetic predisposition to early onset diabetes, the definition of “early stage” in monitoring is rather vague for sporadic diabetics. There is little scientific rationale either to include or to exclude young adults and individuals with dynamic vivacity in the prime of their life and in their 40s or even younger. Hence, effective monitoring of blood glucose that is not shunned by the otherwise healthy population may require a lifestyle approach that offers a product and service that are easy to use, of low risk, of low maintenance, socially acceptable, of robust value, safe, and medically effective. Since the adoption of this device may be voluntary, it may be less attractive if the monitor is a visible wearable [144] or a “thing” that an individual must “remember” to do. Thus, nanotechnology-based [145] monitoring tools may function as “always-on” wireless nano-sensors which remain under the epidermis. It may make blood glucose monitoring of general population an attractive modus operandi for early risk identification and prevention of diabetes as well as associated morbidity, such as diabetic glaucoma, which can lead to partial or complete blindness in severe diabetics or juvenile diabetics, if left untreated.

Early detection of diabetes as a function of monitoring blood glucose concentration benefits from a plethora of glucose sensors developed over the past 25 years, but challenges still exist. A key advantage in the development of a miniaturized or nanoscale device that can quickly and reliably monitor glucose in vivo, is based on the fact that the level of blood glucose detection does not require nano-level detection. The benefit of a nano-device or nano-sensor is to make the
monitoring tool virtually unobtrusive to the user. The normal clinical range for blood glucose is in the millimolar (mM) range between 3.5 and 6.1 mM, but abnormal glucose levels may reach 20 mM. This concentration range can be easily monitored using electrochemical reactions. What is critical for diabetes is a tool that does not deter early-stage frequent vigilance about the changes in blood glucose levels, also measured in milligrams per deciliter of blood. Alterations may signal the ability or inability to maintain the standard equilibrium concentration of glucose (120 mg/dL or 3.5–6.1 mM). For example, if the blood glucose concentration in an individual takes longer to return to normal after meals, then it may signal germinating problems with glucose clearance and/or tolerance in the individual.

The innovation required to develop a wireless blood glucose nano-sensor that is capable of subcutaneous monitoring of blood glucose and transmission of the data to a wireless node may be simple and at hand. The core components are a nano-sensor [146] capable of detecting blood glucose concentration in vivo and a nano-radio [147] that can transmit the data. The combination is illustrated in Figure 8.7.

![Diagram of blood glucose monitoring and wireless data transmission](image)

**Figure 8.7** Blood glucose monitoring and wireless data transmission 24.7.365. (Adapted from Forzani, E.S. et al., *Nano Lett.*, 4, 1785, 2007; Jensen, K. et al., *Nano Lett.*, 7, 3508, 2007. With permission. © 2007 American Chemical Society.)
The open questions presented by this innovative potential may be divided into two broad categories: data acquisition and data transmission. The issues are as follows:

1. Glucose sensor data transmission by the nanotube radio illustrated in Figure 8.7 and other similar [148] devices is not proven because they are constructed as receivers, not data transmitters. But, in general, single-wall nanotubes (SWNT) are like single-mode fiber for electrons (Figure 8.7) and hence have data properties that were made to act as a receiver (Figure 8.7) for the nano-radio but may be altered to transmit the acquired data from the glucose sensor.

2. Functional co-fabrication or simply co-locating or “housing” glucose nano-sensor and data transmitter on a nonallergenic matrix or platform suitable for use as a subcutaneous implant.

3. Optimizing signal-to-noise ratio.

4. Interference minimized nano-communication link to wireless sensor node.

5. Physical locations for safe subcutaneous insertion per customer preference.

6. Procedure for safe extraction of sensor device with minimal discomfort.

7. Foolproof immobilization of implant.

8. Containment of degradation or breakage of components within the housing of the nano-device.

9. Customer’s ability to “forget” about sensor implant.

10. Customer control (agent-based web tool) over function of the device.

11. Customer control over data transmission or modulating the frequency of monitoring.

12. Low maintenance of sensor and transmitter.


14. Explore use of electrolyte gradient of the body or energy from movement to power data transmission.

The responses to the open questions are beyond the scope of this chapter, but the different mechanisms of detection by sensors are important to review, briefly, because wireless nano-communication must be an integral part of the sensor or sensor combination in order to reliably transmit the data outside the body.

The glucose sensor in Figure 8.7 detects glucose based on principles of electrochemistry. The assumption is that the sensor will perform in vivo as well as it has performed in body fluids tested in vitro. Specific detection of glucose is mediated by glucose oxidase, GO₅, an enzyme (hollow circles with pink borders, Figure 8.7) that is immobilized onto PANI-PAA, a conducting polymer (green area, Figure 8.7) made up of polyaniline (PANI) polymerized with PAA (poly(acrylic acid)). Upon exposure to glucose, GO₅, with the help of a natural coenzyme, flavin adenine dinucleotide (FAD), catalyzes the oxidation of glucose to gluconolactone and becomes reduced, [GO₅(FADH₂)], Step 1. The reduced GO₅ form, [GO₅(FADH₂)],
is regenerated via reoxidation by oxygen \((O_2)\) in solution to \(GO_s(FAD)\) and produces hydrogen peroxide \((H_2O_2)\), Step 2. Polyaniline, which exists in its reduced (red) state (PANIred), is oxidized (ox) by \(H_2O_2\) to PANIox and triggers an increase in polyaniline conductivity (Step 3) due to the sensitive dependence of polyaniline conductivity on its redox state. This change of conductivity is data, indicating glucose detection.

Step 1: \(Glucose + GO_s(FAD) \rightarrow gluconolactone + GO_s(FADH_2)\)
Step 2: \(GO_s(FADH_2) + O_2 \rightarrow GO_s(FAD) + H_2O_2\)
Step 3: \(H_2O_2 + PANIred \rightarrow H_2O + PANIox\) (increase in conductivity)

Other types of construction use carbon nanotubes (CNT) but may still use the electrochemical principles for detection of glucose on a CNT scaffold (instead of PANI-PAA, as shown in Figure 8.7). Nano-wires [149] represent one such type of construction. Nano-wire-based glucose biosensors [150] use CNT nano-electrode ensembles (NEE) for selective detection of glucose based on the high electro-catalytic effect and fast electron-transfer rate of CNT but employ the same electrochemical mechanism described above. \(GO_s\), glucose oxidase, is immobilized on CNT-NEE, instead of PANI-PAA, via carbodiimide chemistry by forming amide linkages between the amine \((NH_2)\) residues and carboxylic acid \((COOH)\) groups of the enzyme, \(GO_s\), covalently linked to the exposed tips of single CNT, illustrated as perpendicular black bars in Figure 8.8 (center). Numbers of CNT on a CNT-NEE are in the millions, with each nano-electrode being less than 100 nm in diameter, thereby, increasing sensitivity of the sensor (by analogy, the speed of a processor, e.g., Intel Pentium, is a function of the number of microprocessor circuits etched on the chip). The catalytic reduction of hydrogen peroxide liberated from the enzymatic reaction of glucose oxidase covalently immobilized on the CNT-NEE, in the presence of glucose and oxygen, leads to the selective detection of glucose. CNT are excellent electrochemical transducers, and each CNT serves as a nano-electrode that detects the change in current (conductivity) when glucose reacts with \(GO_s\)-linked CNT (the coupling drives the specificity of glucose detection) in a CNT-NEE sensor. The sensor effectively performs a selective electrochemical analysis of glucose in the presence of interference from common molecules, for example, acetaminophen \([AA]\), uric acid \([UA]\), and ascorbic acid \([AC]\), shown in Figure 8.8 (top panel). But most important, the sensor is sensitive to increments of glucose. Detecting fluctuations in concentration of blood glucose is the key to early detection in diabetes (Figure 8.8, lower panel and inset).

It is relevant to note that nano-wire sensors that detect changes in chemical potential, accompanying a target or analyte binding event, such as DNA or RNA hybridization, peptide interactions, or oxidation reduction in electrochemical reactions, can act as a field effect gate upon the nano-wire, thereby changing its conductance. This is similar, in principle, to how a field-effect transistor (FET) works [151].
Today, FETs are low cost and in extensive use in environmental and agricultural monitoring. Advances in technology have made the expensive electronic marvel of a “transistor radio” of the 1950s only suitable for infant’s toys mass manufactured in China and sold in discount chains. The transistor radio has evolved to dirt cheap FETs now used in animal farms to alert owners that the stench from the ammonia-rich waste from animal excreta in the holding tanks needs attention. The expensive transistor of yesterday is a low-cost (or no cost) component today that delivers significant value for environmental monitoring. It benefits the meat industry at such a negligible cost that it has only increased productivity of the meat industry and concomitant increase in global consumption. The evidence for the latter is gleaned from the beef and chicken consumption data from the United States and EU that exceeded 120 kg per person per year (330 g/day) compared to 16 kg per person per year (44 g/day) in China and India, combined [152].
Knowledge from research and applications of FET are available from archives that may date back to the initial discovery of transistors [153], nearly 70 years ago. The development of nano-wire, nano-sensors, and nano-communication may benefit from the experience and wisdom of the FET pioneers.

The principle of FET indeed may be crucial for nano-communication for wireless transmission of data from the in vivo sensor to a sensor communication link or node outside the body that can connect to the Internet. The characteristic of a CNT to act as fiber that can transport electrons is under scrutiny in order to use the material for high-speed data transmission in a variety of ways that includes the emerging field of plasmonics [154] that studies interactions between light and nanoscale particles and structures. Remote monitoring in vivo that may take advantage of light-emitting luciferase [155] enzyme-linked sensors, by immobilizing luciferase (on PANI-PAA matrix or) preferably on CNT-NEE type scaffolds, may find it useful to exploit the potential of nanoscale antennae that converts light into broadband electrical signals capable of carrying approximately one million times more data than existing systems [156].

The illustration in Figure 8.7 is simple to understand and conveys the image of the components necessary to drive the convergence of in vivo detection and transmission of data. But, caution is necessary to extrapolate the application of the components illustrated in Figure 8.7. Although it may be ideal for visualizing the concept, the components illustrated are not proven or guaranteed to be the combination of choice that could drive the development of an in vivo nano-device that is equally reliable as a detection tool and a nano-communication tool, for reasons that I shall explain in the next few paragraphs. But of course that argument holds for any innovation. We shall not know the outcome unless we attempt to create the device if there is even “just enough” reason that the innovation may bear fruit. Reasonably, one issue in the glucose nano-junction sensor is that it does not use nano-materials, that is, in this case, CNT, in its construction. Nano-wire glucose biosensors use CNT as the nano-electrode, as shown in the CNT-NEE sensor illustrated in Figure 8.8 (center). This may be an issue in terms of the ability to transmit the data out of the body.

If the nanotube radio receiver illustrated in Figure 8.7 can be modified or designed as an in vivo transmitter and if it can successfully detect the data emerging as the change in conductivity between the two states of PANI (PANI reduced versus PANI oxidized, as shown in Step 3) from the non-CNT glucose nano-junction sensor illustrated in Figure 8.7, then the nano-device may be produced, by the combination illustrated in Figure 8.7, for wireless in vivo glucose monitoring.

The preference for CNT-based sensors like CNT-NEE (Figure 8.8, center) is linked to the central need to transmit the data from the in vivo sensor and the potential for using individual nanotubes within CNT networks to carry information. Innovation in nano-communication using CNT communication network [157] is likely to become a core competency necessary for the future of nano-device
use in healthcare, in general. It is in this context that the principle of field effect transistors (FET) springs back into action. Current technology utilizes an entire CNT network as semiconducting material to construct a single FET. Several FETs are required to build traditional or legacy network equipment. The result is that there are many nanoscale networks embedded within each device (FET) that might be otherwise more effectively utilized for communication. In other words, the CNT network itself is the communication media, and individual CNT are the links. But, individual CNT and tube junctions (forming nodes) do not have the equivalent processing capability of a traditional network link and network node. To compensate for this, the system needs to leverage large numbers of CNT. The illustration in Figure 8.8 (center) shows individual CNT linked to GOx (black bars), but millions of CNT (black bars) make up an ensemble (CNT-NEE). The latter is precisely what is necessary for the single-wall CNT communication network (NanoCom) of the future. However, NanoCom may not function according to the traditional architecture of data communication layers [158]. Comparison of legacy communication and NanoCom highlights the changes necessary, as shown in Table 8.3.

At the bottom level (least sophisticated level), communication links may be between hosts and routers in a communication network, or they may be CNT overlapping at points that will be identified as nodes. A network functions by changing state. Data must either flow or be switched or routed through nodes. State may be implemented as a routing table on a router or an electromagnet field controlling the resistance within a specific area of a CNT network. Finally, a mechanism needs to be in place to control state (ascending level of sophistication, Table 8.3), be it a routing algorithm or FET gate voltages applied to a CNT network. The traditional networking protocol stack is inverted in this approach because, rather than the network layer being logically positioned above the physical and link layers, as in the

<table>
<thead>
<tr>
<th>Basic Network Components</th>
<th>Traditional Networking</th>
<th>Nanoscale Networking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocol</td>
<td>Processors</td>
<td>Gate control</td>
</tr>
<tr>
<td>State</td>
<td>Node memory</td>
<td>Semiconducting tube resistance</td>
</tr>
<tr>
<td>Network</td>
<td>Links</td>
<td>Nanotubes</td>
</tr>
</tbody>
</table>

standard OSI model [159], the CNT network and routing of information are an integral part of the physical layer.

Data transmission in a CNT network occurs via modulated current flow (changes in conductance) through the CNT network guided towards specific nano-destination addresses. The addresses identify spatially distinct areas of the CNT network that may be made up of nanosensor arrays. Since gate control is used to induce routes through the CNT network, nano-addresses are directly mapped to combinations of gates to be turned on that induce a path from a source to a destination. Is the “combination of gates” in any way analogous to the network, subnet, host type of partitions that specify a 32-bit IPv4 address of the type 151.193.204.72 or the 128-bit IPv6 format 21DA: 00D3: 0000: 2F3B: 02AA: 00FF: FE28: 9C5A (or equivalent 21DA: D3: 0: 2F3B: 2AA: FF: FE28: 9C5A with leading zero suppression)?

Does nano-addressing require an entirely new scheme for unique identification? Is unique identification necessary at the level of individual nano-addresses? If necessary, is a relativistic identification of information [160] necessary? Are “source” and “destination” comparable to client–server architecture?

These and several other open questions are likely to emerge. Figure 8.9 illustrates a conceptual network view of the CNT infrastructure. Superimposed on NanoCom are some of the medical benefits that make CNT-based sensors and data communication a powerful ally for healthcare improvements.

Diabetes is not the only disease that merits prevention and monitoring. For early detection and monitoring, several other diseases qualify prominently. Sensor

---

**Figure 8.9** NanoCom transmits data for multiple biomarkers from in vivo sensor nano-array. (From Bush, S.Y. and Li, Y., *Nano-Communications: A New Field? An Exploration into a Carbon Nanotube Communication Network*, Technical Information Series, GRC066, GE Global Research Center, Niskayuna, NY, 2006. With permission. © 2006 Interscience.)
Figure 8.10  Improving healthcare from the home to the hospital: “sense, then respond.”

nano-arrays may be one answer for a multifunctional detector, as conceptualized in Figure 8.10. Sensors constructed from nanotubes change their resistance based on the amount and specificity of the material or biomarker detected (sensed). Thus, the act of sensing may change the routing through the NanoCom network. In other words, in a staggered approach, the duty cycle and/or the sleep time [161] of the different CNT nano-sensors can be regulated to allow any one sensor to function in a specified interval and detect or sense its specific analyte (glucose, cholesterol, or acetylcholine, Figure 8.9). The change in the properties of the CNT, when the act of sensing is in progress, opens the appropriate “gate,” and when a gate is turned on, the nanotubes within the gate area become conducting. Then, the data acquired for that specific sensor are transmitted to an external node. For complex diseases, where multiple pieces of data and vital signs may be necessary to make an informed decision, either by humans or initially by an AI-based intelligent analytical system (AGRMIA, Figure 8.2), properly choosing the sequence of sensor, hence gates, to turn on, changes the current flow to the edges of NanoCom, the CNT network. The latter effectively creates a controlled NanoCom, which may act as or provide weights in a neural network for AGRMIA type system requiring a collection of different biomarker data, to determine the relative impact or relationships or values of variables that may be co-integrated [162], which, if taken
together, may better reflect the state of the patient and the status of the disease. The Metabolic Syndrome caused by the mutation in the LRP6 gene may be an example of a multi-factorial disease where multiple conditions are affected and may require simultaneous monitoring.

Although this chapter, thus far, may have synthesized a number of promising practical ideas, far better and innovative ideas and concepts about products and services, than the ones mentioned here, have perished. Rarely acknowledged is the observation that no matter how good an idea may be, it may only shine in obscurity unless there is a strategic plan that charts an appropriate use and adoption path in context of other complementary technologies and social awareness of its value. A common example is the introduction of the first hand-held “Newton,” the personal digital assistant (PDA), from Apple that few may recall. Apple only sold 140,000 Newton PDAs at its peak in 1993–1994 and soon discontinued production. A few years later, 3Com introduced the Palm Pilot PDA with features that were even primitive to Apple's Newton but with the option of Internet access. Today, it may be hard find anyone in the industrial world and professionals in the developing world who do not have a PDA, of some form or the other. The “social awareness” of the value of PDA accelerated with the penetration of the Internet. Newton, RIP, was slightly ahead of its time.

Calls to contain the unbridled cost of healthcare are demanding exploration of tools and technologies. The need for wireless remote monitoring and the use of nano-biosensors in healthcare are as vital as desalination projects, carbon sequestering, metabolic engineering, clean water, and clean air. Hence, innovative tools and technologies must also outline a strategic path for their integration and potential for adoption by healthcare systems without expecting a complete overhaul of the existing system, anytime soon.

The overall strategy for the use of wireless remote monitoring tools and the manner in which it may function in the landscape of healthcare cost reduction (Figure 8.2) is illustrated in Figure 8.10. The time saved by attending to the two individuals identified in Figure 8.10 instead of outpatient visits by seven individuals (Figure 8.10) saves cost of service and supplies and enables healthcare professionals to devote necessary time to those who need the attention, hence improving the quality (QoS) of healthcare.

The normal clinical concentration of glucose in blood is in the millimolar range, and that precludes the need for nano-level detection. However, early detection of other diseases may indeed be far more effective if detection is possible at the nano-level of proteins, peptides, molecules, or degraded macromolecules that may hold clues to problems in their embryonic stages. The task of medical research is to identify these markers, and the task of the technical experts is to find ways to identify these markers through remote sensing tools. Differential profiling of gene and protein expression between normal and disease states may be helpful to identify biomarkers that are only expressed in disease states. The amounts of such
disease-dependent biomarkers may benefit from pico or nano-level detection and offer “true” early detection.

Data and detection as components of the systems approach to future healthcare require medical science and engineering technology to create tools linked to systems that may be purchased, as ubiquitous generic plug-n-play commodities, from the local pharmacy or convenience store in a gas station or corner grocery store. Consider the revolutionary discovery of transistors that produced FETs. The field has been shaped by evolutionary market forces over the years, and FETs are now used as low-cost commodities in the design of environmental and agricultural monitoring. It is this trend that Figure 8.7 illustrates, in concept. The value they deliver through specificity and sensitivity is illustrated in Figure 8.8. The strategy for integration and adoption pathway is illustrated in Figure 8.10, which includes one of the key “behind the scene” drivers that may materialize as modular analytical engines of the AGRMIA type collection of resources, illustrated in Figure 8.2, connected on global grids. Taken together, remote monitoring by internal wireless nano-sensors or external sensors (as fashion rings, bracelets, wrist watches) will coevolve and diffuse as a lifestyle approach in healthcare, not because individuals are sick but because they prefer to stay healthy. The healthcare industry may benefit from business advice in order to take advantage of time compression. In other words, the time to market from idea (Figure 8.7) to adoption (Figure 8.10) may be shorter than the decades between the discovery of transistors and the use of FET as a dirt cheap commodity for agricultural and environmental monitoring.

This is the “writing on the wall” for what is in store for remote monitoring by wireless nano-sensors, the nano-butlers, that can deliver value at a reduced cost, for micro-payments, in local and global healthcare. Ignoring the suggestions in this and other reviews or the failure to accelerate the necessary convergence to create the suggested healthcare services as well as support research [163] may lead to chaos.

Healthcare imbalances will continue to proliferate in the United States, and severe consequences are also predicted for EU nations, particularly Ireland.Projected rise in age-related government spending as a share of GDP of Ireland, over the ext 40 years, is among the highest in the euro zone [164]. In the absence of reforms and adequate investments in Ireland, the advances in innovative convergence suggested here may be sluggish, at best. But, the cost of healthcare in Ireland, which is approaching €3000 per capita, may continue to increase (see Ref. [53]). Rising public debt in Ireland may force spending cuts with a concomitant decrease in the quality of life for those who need healthcare services. Cost containment in healthcare services may grossly reduce preventative measures, selective or elective procedures, palliative care, and all nonemergency services. Individualized medicine and personalized healthcare will be a matter of fiction, and A&E type acute-care services may be the healthcare skeleton. Enterprise, academia, and government can prevent this state of affairs.
8.4 Innovation Space: Molecular Semantics

The proposal of molecular semantics, whether right or wrong, does not impact the implementation of nano-butlers. Introducing the concept of molecular semantics as an independent paper may have been prudent, but the preliminary idea merits inclusion in this chapter to indicate the importance of structure in medical science. Current systems, such as EMRS and AGRMIA, may be, in general, incapable of dealing with structures unless dedicated programs are used.

8.4.1 Molecular Semantics Is about Structure Recognition

Classical semantics includes descriptive ontologies and extracting word relationships that form bulk of the thinking [165] prevalent presently to move from the syntactic to the semantic web. Molecular semantics may not be necessary for general usage but may aid special analytical applications in the future to uniquely identify molecular or chemical structures or units of structures or epitopes, in a manner that may have some similarity with the concept of digital semantics (see Ref. [40]). Unique identification of structures may enable diverse systems to compare structures in a catalogue or database with those that may be identified in some disease states, or query, if an identified structure has any known homologies or close similarities to one or more parts of chemical or biological molecules. The significance of partial structures and segments of structures or epitopes in healthcare diagnostics may be better appreciated from the discussion, later in this section, on autoimmune diseases and molecular mimicry [166].

To deliver value in healthcare, part of the solution calls for analytical tools to extract information from data. Healthcare presents three types of data that may need to work together, in some cases:

1. Numerical data in EMRS and AGRMIA platforms may be fed directly to the analytical engines.
2. Syntax from patient history and physician “notes” still on paper is likely to create syntax versus semantics nightmare if transcription is necessary to create EMRs. The syntactic web of today also plagues the business world. The semantic web movement and organic growth of ontological frameworks contributed by experts from various disciplines are continuing their valiant efforts to enable computer systems to “understand” and extract meaning from syntax. In this context, I have proposed a parallel approach to explore semantic and ontological frameworks using an IPv6 type unique identification scheme to enumerate data, information, and decision in a relativistic approach (see Ref. [40]).
3. Molecular patterns may be partially novel and especially relevant for healthcare diagnostics.
This section on molecular semantics is similar yet distinct from “digital semantics” proposed earlier (see Ref. [40]), but both share the concept of unique identification to enable global systems interoperability without ambiguity of identification or errors due to ambiguity in ontological frameworks that may be incomplete. Both proposals, however, are also likely to be incomplete.

The current proposal on molecular semantics addresses the third type of healthcare-relevant data structure, that is, molecular patterns or molecular structures. Organic chemistry and biomedical sciences place a great deal of emphasis on patterns and structures. Hence, it may be worthwhile to dare to forward this new idea of how to enable computer systems to recognize patterns and structures through the use of agreed units of molecular structure that form parts of macromolecules. This proposal freely borrows ideas from seminal works of great scholars but with rudimentary understanding of their depth and without any guilt. In addition to a few building blocks of linguistic theory [167], I have also stretched the outlines of semantic theory and cognition [168], perhaps to an unnatural and unreasonable extent.

Cognition as related to structure in natural language is mapped to the units of molecular structure in biological macromolecules. I have extrapolated the ideas and elements of the linguistic system to fit molecular semantics in a way that proposes the use of molecular units of structures (of macromolecules, such as proteins) as “agreed lexicon” to catalyze recognition and understanding of these structures between diverse systems to aid systems interoperability. However, in my biased view, elements of the linguistic system (Figure 8.11) seem to resonate with the proposal of molecular semantics. But, I shall be the first to recognize that this convergence, however attractive, may not be right, in the form proposed, and admit that it may be in error, if it is.

There may be other ways, but the concept of molecular semantics may be yet another tool to explore this scenario: MS analysis of serum from a patient with unidentified type of fever has identified a high concentration of a short peptide with some ambiguity about its sequence, but it appears that the peptide may be part of a common protein, myelin. First, a GP may rarely send a serum sample for MS analysis. Second, if the MS data are uploaded and an AGRMIA type system is available, is there a tool to compare the MS signature from this patient (MSpat) with other MS data in a database? The data (MSpat) entry point in the linguistic system is the conceptual structure, the EMRS part of the EMRS-AGRMIA system, in this context. The computational resources that may be needed to perform the comparative search, one MS data at a time in the database, may make this approach untenable. If there is a catalogue of unusual MS signatures (MScat) in the form of a (data) dictionary, then, perhaps, it may be feasible to perform this search and determine if a match or close relationship exists between MSpat and a pattern in MScat.

In the linguistic system, the dictionary equivalent may be the Lexicon (Figure 8.11). Is it really necessary to bring the Lexicon into this discussion? The MS data
and accompanying explanatory notes could also exist in a relational database, and it may suffice for this exploration. The introduction of the Lexicon in this scheme may seem, initially, on shaky grounds, but following the model of the linguistic system, it reasons that if the MS data catalogue (MScat) does exist in a Lexicon equivalent, it may also have links to syntactic structure and semantic structure (Figure 8.11). The wide variation in syntax needs no extra justification. Hence, the difficulty to match syntactic structures with respect to the Lexicon even if it contained a match to MSpat may not be unexpected. But the data (MScat) could point to descriptions that may be more specific in the semantic structures. It may identify MSpat as belonging to neural proteins (myelin is a neural protein). Once the MSpat data matched to data from the Lexicon (MScat) points to neural proteins in the semantic structure, the correspondence rules could point back to the syntactic structures and identify one or more descriptions of neural proteins that may offer matching segments to MSpat. The real finding is however the match between MSpat and the potential link to myelin that may happen if an extensive MScat is matched by protein sequence and classified in the semantic structure. It is quite possible that
an equivalent set of rules may work in place of the Rules of linguistic inference to relate the MS data with MSpart and point out the class of neural proteins and the ontological relationship: myelin is a neural protein. The Lexicon (data) and Rules of linguistic inference (rules) may be working in concert to indicate the semantic structure, and another set of rules, Correspondence Rules, could point to syntax or descriptions in the syntactic structure that builds on the identified semantic structure. The extent of the classification, precision, granularity, and other factors of the semantic and syntactic structures could be determined by rules; the equivalent in the linguistic system are the Semantic and Syntactic Well Formedness Rules (SemWFR and SynWFR, Figure 8.11), respectively. The presentation of the data may find some far-fetched relatedness to the phonetic representation and phonology domains in EMRS-AGRMIA, when visualization of the data is important.

Thus far, I have not addressed molecular semantics but attempted to explore how the linguistic system seems may offer a parallel in data analysis. The latter is expected because linguistics and artificial intelligence share common elements such as cognition, semantics, and ontology. The idea of molecular semantics was camouflaged in the discussion because the data (MSpart) attempted to find a match to an existing equivalence relationship to a description of relevance of the structure. Molecular semantics indeed performed its task in trying to find a match between MSpart and MScat. The location of the MScat in the lexicon may not be an optimal explanation, and the lexicon may have a database equivalent where the incoming data through EMRS, equivalent to the visual system that feeds the conceptual structures in Figure 8.11, are formatted by certain rules (equivalent to the Pragmatics in Figure 8.11) and channeled to MScat that operates under the AGRMIA umbrella, which could be a database with MS data, linked to the Lexicon (dictionary).

Molecular semantics, the definition and identification of structures, may contribute to the ability of the system to compare MSpart to MScat profiles. Although MS analysis does not deliver an actual structure, the spectral data offer a pattern, and for this discussion, this pattern or profile is referred to as structure. This comparison can be performed today only by special applications.

Current semantic web efforts are striving to stimulate global groups to contribute descriptive ontologies for chemical and biological systems that may be accessible by the software tools and standards promoted by the semantic web experts. That effort may not include the potential for considering ontological frameworks for chemical and biological structures or data patterns. Tools such as MS, echocardiogram, and magnetoencephalography do not generate syntax and thus may be excluded from ontological frameworks. This deficiency in the current practice is addressed by this proposal on molecular semantics.

Conservation of the amino acid arginine in LRP6 protein product from frogs to humans may help make it less surprising to understand, following the logic of evolution, why viruses and bacteria may also share homologies with sections of human DNA or why viral proteins [169] and bacterial proteins [170] may have some homology to amino acid sequences found in normal and common human
proteins, such as myosin. This apparently benign observation often produces some devastating medical consequences. Because small segments of proteins are involved, these segments or epitopes may have structures that may be distinct. Understanding the form (structure) is crucial to understanding the function in biological systems, hence the need for molecular structure in healthcare analysis, albeit only in special cases, and the potential importance of a mechanism that enables analysis of structure, that is, molecular semantics.

Disastrous medical consequences are observed due to the phenomenon commonly referred to as molecular mimicry. The situation is often caused by a foreign protein from a virus or bacteria that share a short stretch of amino acid sequence homology to a normal human protein. The immune system happens to target the foreign protein but for some reason specifically recognizes this homologous region that serves as the antigen. The body elicits an immune response and attacks the protein. Because this antigenic region (epitope) is also a part of a normal human protein, the antibodies produced by the body’s immune system also recognize the epitope that is present on normal “self” proteins. Unfortunately, the immune system begins to destroy the self protein (autoimmune) with severe consequences for the patient. Because the short amino acid sequence of the foreign protein, as short as only six amino acids [171], may mimic the corresponding sequence of the normal self protein, the phenomenon is referred to as molecular mimicry.

This autoimmune response is partly to blame in some cases where an individual in almost perfect health suddenly drops dead from cardiac arrest or succumbs to a heart attack. The etiology may be linked to common Streptococcal infection (strep throat) that most children and adults experience at some stage or the other. Segments of some proteins from Streptococcus share homology to proteins specifically found in heart tissue [172]. Various other cases of molecular mimicry are well known including T-cell-mediated autoimmunity [173] and ankylosing spondylitis caused by only six amino acids (QTDRED) found in Klebsiella pneumoniae nitrogenase enzyme (protein) that exactly matches the human leukocyte antigen (HLA) receptor protein (HLA-B27) antigenic epitope [174].

Do these very short stretches of amino acids create antigenic epitopes with certain structures that may form a class of “super antigens” responsible for eliciting the human immune response that leads to autoimmune diseases? Using basic rules of protein structure and conformation, these epitopes may reveal structural patterns that may influence “function” in biological systems, such as eliciting an immune response. The structure of a short sequence of amino acids may be almost identical to another structure of a short sequence of amino acid, but the structural homology may not indicate that the two short stretches of amino acids share sequence homology. Computer systems of the syntactic web and of the semantic web may help in the data analysis that involves sequence (words, such as QTDRED) homologies or differences but may be incapable of dealing with structures that may be homologous whereas sequences are not. The likely analytical result of a system presented with QTDRED versus QTDREG may suggest, erroneously, that the structures are
different. However, if the system, in future, could use the tools of molecular semantics and refer to the Lexicon (Figure 8.11) of structures, it might reveal that the two stretches of six amino acids analyzed are different in sequence but produce identical structures. Because antigen–antibody binding is structure dependent as long as the side chains and groups are similar, it is possible that two slightly different amino acid sequences may result in nearly identical structures that can still elicit the same autoimmune response. A relevant scenario often occurs in organic chemistry where the empirical formula of a chemical or compound is same but the properties of the resonant structures may be different. The current semantic web may be impotent to address these structural issues. In combination with the power of the semantic web, there may be a need to address structural issues in healthcare diagnostics, hence the value of the proposal of molecular semantics.

Why did I choose molecular mimicry and autoimmune diseases to make the case for the value of the structural approach in the proposal of molecular semantics? The partial answer is based on the fact that almost an infinite number of allergens and antigenic epitopes can share short homologies to self proteins. By exploiting molecular mimicry, it can lead to autoimmune diseases. Healthcare, in general, and clinical immunologists, in particular, may wish to understand at a greater depth the form and function relationship in biological systems and autoimmunity. It cannot be done without structure and structural comparison, for reasons already elaborated above. Globalization now provides wide access to food from all around the world. This window on world cuisine has the potential to spawn new and undocumented forms of allergies to various ingredients and antigens foreign to the body. The potential to cause some forms of autoimmune reaction may manifest as inflammatory bowel disease [175] or the generic irritable bowel syndrome [176].

Globalization and mobility will usher new domains of healthcare and seek knowledge about many more issues to diagnose complex diseases, for example, autoimmunity caused by antigens in food products. Early detection of antigens, which can pose the threat of molecular mimicry, may be of great significance. Research is needed to determine how to identify candidates for molecular mimicry and what type of assays, in vivo or in vitro, can detect these short segments. Only time can tell whether the growing demand for detection and need for analysis in healthcare may trigger an exploration to “productize” molecular semantics. In some cases, structural analysis may be necessary or even pivotal to complement numerical data and syntactic/semantic information for use in EMRS-AGRMIA (Figure 8.2) type systems that must be globally interoperable and locally responsible.

Molecular semantics or ideas that may originate from linguistics [177] may evolve when more people, on both sides of the aisle, medicine and engineering, can better appreciate the value, however subtle, of form versus function in biology. This proposal in its current format may become irrelevant, but it may provide some clues or may even serve as The Golden Key [178] to unlock creativity and innovative patterns in the mind [179] of young people to drive convergence of syntax, traditional semantics, numerical data, and structure, to improve analysis and benefit healthcare.
8.5 Auxiliary Space

8.5.1 Potential for Massive Growth of Service Industry in Healthcare

Adoption of this data-driven model may increase the diffusion of a new class of service industry. The growth of health services may create new markets and trigger new business and revenue models (pay-per-use) even in sectors not directly involved in healthcare, but it offers products that may be used by health services, for example, (1) software vendors deploying cloud and grid computing platforms, (2) telecommunications companies billing for data transmitted in real-time, (3) data routing or IP connectivity architects to ensure privacy, data confidentiality, and address verification using Internet Protocol version 6 (IPv6) and other security tools, and (4) data mining outfits that may create intelligent differential decision engines (IDDE) running on grids, cloud computing environments, in-network processing, embedded browser applications, or a host of other platforms yet to be determined or discovered.

The pivotal role of data mining in healthcare data analytics is expected to evolve in ways that are yet to be defined. Data mining as applied to so-called “business intelligence” applications may play a role but may be inadequate to address the service part of healthcare because the “service” of healthcare is about an individual or patient-centric data. On the other hand, the healthcare industry may have distinctive needs, but, in general, it is about business and operational efficiencies. Data, information, and knowledge hold the potential to improve both healthcare service and the healthcare industry, but the tools and applications are expected to differ in their pursuit of different goals and functions. This is where the prevalent view of data mining is expected to diverge.

The tools of data mining were enriched with a sea of changes when principles associated with complexity theory and swarm intelligence [180] emerged to offer practical business solutions [181] for a wide variety of routing and scheduling needs. A similar wave (see Ref. [10]) is imminent under the generic banner of data mining tools that may stem from reality mining (see Ref. [77]) and its link with social networking relationships (see Ref. [76]). Principles extracted from reality mining and social networking paradigms may yield tools applicable to a variety of fields including business services and healthcare analytics.

Data mining in healthcare, service, and industry perspectives taken together also offers a fertile ground for exploring whether the convergence of economic principles, tools, and techniques in healthcare data analytics may lead to further innovation. The healthcare ecosystem offers the opportunity to test at least four different economic principles as analytical tools. For healthcare service, the focus is on patient data, and these data are generated by the physiological system. Since human physiology is highly integrated, it may follow, naturally, that the physiological variables (e.g., blood pressure, heart rate, pulse rate) are likely to be co-integrated
(see Ref. [163]). In other words, because physiological systems strive to maintain homeostasis, it follows that the goal of physiology is to attain equilibrium. When one variable is affected, for example, pulse rate, its effect is “integrated” or reflected or related to another linked variable, for example, blood pressure. Physiological balancing mechanisms within the human body will attempt to rectify this situation and may try to restore the blood pressure of the individual to 120/80 mm Hg, the normal reading. Due to the innate physiological drive to restore equilibrium, data analysis in healthcare service may benefit from a potential exploration and application of the principles of Nash Equilibrium [182] to predict from a set of patient data what other parameters (co-integrated) are likely to change or may be influenced by the change documented (data at hand). It may provide clues to improve diagnosis.

Data mining tools for the healthcare industry may benefit from some traditional approaches coupled with a few emerging concepts. Like most businesses, healthcare industry suffers from systemic gaps of data and information in its complex supply chain. Consequently, the healthcare industry is prone to information asymmetry [183] and expected to benefit if information asymmetry could be reduced through appropriate acquisition of data including real-time data, for example, from RFID. Availability of high volume data from deployment of automatic identification technologies (see Ref. [131]) may help improve forecasting to better manage human resources and inventory planning in the healthcare industry. High volume data may be instrumental in improving the accuracy of forecasting using time-series data in combination with a host of forecasting tools including the econometric technique of generalized autoregressive conditional heteroskedasticity (see Ref. [46]).

Contrary to public opinion in business consulting, except in specifically designed business collaborations, the application of Nash equilibrium in business (see Ref. [65]) may be conceptually flawed. It is less useful for business decisions but better suited to healthcare analytics for healthcare service. In contrast, the concept of information asymmetry is foreign to human physiology but is almost a second nature to the competitive dynamics of business, which makes it useful as an analytical tool in the healthcare industry.

The global growth of the health services industry through the model illustrated in Figure 8.5 will dwarf the current revenues of $748 billion [184] from business services. The three major global giants that currently offer business services (IBM, HP, Microsoft), taken together, command less than 10% of the $748 billion in revenues. Therefore, by extrapolation, it seems reasonable to suggest that small companies, start-ups, and small and medium enterprises (SME) shall find the barrier to entry in the healthcare service industry to be low or nil. The healthcare service industry will be driven by innovation, which is best executed by small “skunk” works of talented individuals. Due to multiple convergences necessary to produce a complete product and/or health service, core competencies will be a driving factor. The latter may stimulate the need for collaboration and partnerships between a number of small or medium enterprises with local and global research institutions and medical facilities. Each group or alliance or SME may contribute its own
specific module or component but may find it essential to cooperate with multi-talented team made up of scientists, other companies, medical personnel, patient advocates, and strategists to act as an interface to catalyze implementation.

Stringent requirement for a higher level world-class advising and supervision to guarantee credibility of the process, products, and health services is necessary and essential. The noncommercial ad hoc supervisory team may begin their involvement from the conceptualization stage and continue through the cycle of planning, research, product development, service creation, testing, and implementation. Critical evaluation by a team of academic and commercial experts [185] may help define performance indicators (KPI) and determine the strength of this emerging vision of healthcare by exploring (1) quality of care improvements, (2) impact on human resources in terms of time savings for medical professional, (3) reduction in cost and potential for savings, (4) length of time required for return on investment, (5) profitability of businesses (SME) and growth of high potential start-ups, (6) economic benefits for the nation’s healthcare system, (7) reproducibility, portability, and sustainability of the services model as a global template, (8) business opportunities to implement similar services in other communities or nations, (9) creating market alliances in emerging economies to implement healthcare services, and (10) liaison with global organizations (World Health Organization, United Nations Development Fund, World Bank, Asian Development Bank, Bill & Melinda Gates Foundation) to help in the global diffusion and adoption of health services industry model.

8.5.2 Back to Basics Approach Is Key to Stimulate Convergence

The discussion in this chapter, in general, has continuously oscillated between medicine, engineering, and information technology in an attempt to emphasize convergence and suggest fruitful analogies between the fields. This chapter is about data, analytics, and tools from research that may improve healthcare. Hence, the preceding sections seek to harvest advances in systems engineering and information communication technologies (ICT) as well as translational medicine to improve healthcare through multidisciplinary confluence. Therefore, I shall be ethically remiss if I do not digress and fail to highlight in this section why the need for convergence is accepted but in reality organizations are sluggish to address the challenges in the clinical enterprise [186]. The problem has deeper implications, and unless reformed, the ramifications are bound to be increasingly disappointing.

In its simplest form, implementing convergence is often inhibited by the general biomedical illiteracy of technical experts and technical illiteracy of biomedical experts. Insightful degree programs in biological engineering [187] and health science technology [188] are key mechanisms to create the supply chain of talented individuals who have understanding of one field and depth in another, to
act as a knowledge bridge, which is key for the progress of convergence. The U.S. physician-scientist programs [189] that produce graduates with a PhD and MD are equally valuable and other countries are beginning to implement related strategies [190]. However, these programs only attract the crème de la crème of the nation, and in some countries the total number of these highly qualified individuals fails to reach a critical mass. Consequently, the few who succeed often move to other parts of the world where a critical mass of talent exists and where their multiple skills are valued, duly rewarded, and challenged to guide the nation or global groups.

What is sorely needed and missing in most countries is the focus on training programs for “middle level” workforce executing the bulk of the work yet remain firmly sequestered in one job or domain without the scope or the desire to become multifunctional. Programs with financial incentives, paid leave of absence, structured academic training, and practical internships are necessary to provide technical education for medical experts [191] and other healthcare professionals (consultants, GPs, nurses, physiotherapists, home-helpers, mental health workers) to understand (not necessarily gain expertise) the fundamentals of medical device engineering, sensors, remote monitoring, communication technologies, transmission protocols such as TCP/IP, software architecture, statistics, principles of artificial intelligence, basic principles of logic, and programming. Similarly, experts in engineering and technology should be offered the attractive opportunity to gain understanding of human physiology, pathology, pharmacology, anatomy, cellular and molecular biology, neurology and mental health, genetics, principles of internal medicine, nuclear medicine, medical imaging, biomedical data, inpatient and outpatient management in hospitals, hierarchy of decision-making, nutrition, social, and environmental factors in health, laboratory data reporting, and epidemiology. Cross-pollination of ideas is a key to innovation.

Implementing these parallel training programs may not pose an insurmountable barrier in most countries even if their vision of the future and commitment to financially invest in its people is modest, at best. What is likely to surface is the difficulty of attracting sustainable number of cohorts to the programs. The problem to attract mature mid-level working class for retraining or lifelong learning, at the tertiary level, is partially rooted in the primary and secondary education of the nation. The emphasis or lack thereof on mathematics and science education either due to (1) archaic policies, (2) compromised rigor to feign inclusion, (3) misguided teacher education programs that chooses process and dilutes content to serve the lowest common denominator, (4) emphasis on test preparatory teaching without room for problem-based learning, (5) inability to stimulate increasing number of female students to take up advanced mathematics and science or catalyze young women to pursue career paths in the hard sciences, or (6) shoddy and second grade teacher qualifications (especially in mathematics and science) masquerading as good enough [192].

In the United States, a seminal report [193] revealed that 51% of mathematics teachers in the U.S. public K-12 (primary and secondary) schools never took
mathematics as a part of their college curriculum. A third of the “education-school-certified” science teachers never took science as a major in college. A national survey [194] of high school physics found that 25% of students took “some” physics in high school and 1.2% of senior students (33,000 out of 2.8 million) enrolled in advanced physics. About 18% of certified teachers teaching high (secondary) school physics had degrees in physics while 11% certified teachers had “degree in physics education but not physics,” and 27% certified teachers teaching physics had neither a degree nor any relevant experience in the subject.

In the Third International Mathematics and Science Study (TIMSS), the United States ranked 28th in mathematics and 17th in science, lower than countries like Slovakia, Slovenia, Bulgaria, not to mention nations in Asia [195]. In 1998, U.S. high school students outperformed only two (Cyprus, S. Africa) of the participating countries. In addition, the TIMSS classroom study revealed 90% of U.S. middle school mathematics lessons are of low quality compared to Japan (10% low) and Germany (30% low). The quality of mathematics teaching was reflected in the poor performance of U.S. eighth graders (middle school) reported in the 1999 TIMSS analysis.

The declining effectiveness of mathematics and science education is reflected in the fact that U.S. colleges and universities awarded 24,405 bachelor degrees in computer science in 1996, 50% less than 1986 (30,963 in 1989), and engineering graduates dropped from 66,947 in 1989 to 63,066 in 1996 [196].

Despite these disappointing trends, the United States is still regarded as the “cradle of innovation” by global experts and organizations [197]. The enigma clears if one considers the actual number of qualified graduates: in thousands. It generates a critical mass of talent to innovate and contribute to economic growth. Each qualified individual contributes several magnitudes more than the average per capita contribution to the U.S. gross domestic product (GDP). As an example, by 1997, graduates of one U.S. institution, alone, had founded 4000 companies employing over 1.1 million people with annual sales close to $250 billion [198]. A recent analysis of the same institution indicates that innovation and inventions of this one institution, annually, create new companies that add 150,000 jobs and $20 billion in revenue to the U.S. economy, each year [199]. By extrapolation, this institution alone, therefore, thus far, has created companies that may directly employ over 2.5 million people and generate about half a trillion dollars in annual revenue. The dedication to research-based entrepreneurial spirit coupled with the freedom of some U.S. institutions to think out of the box as well as the strength of the U.S. investors to assume substantial risks are factors that continue to ignite innovation and profit even though investments, both academic and financial, are not immune from failure.

Attempts by other industrialized nations, with far smaller population, to partially mimic the U.S. strategy have produced mixed results. The striking visibility of the global success of the graduates and faculty from U.S. research institutions in creating innovative companies, products, and services is buoyed by investors willing to assume great risks. In addition to the favorable financial environment, the
numbers or critical mass necessary for innovation is a major determinant to spawn success, and it may not be available for countries with limited population. Equally, a public basic education system that lacks emphasis on rigorous mathematics and science education at the primary and secondary level reduces the supply chain of talent for the future MD or PhD pool. It may be one reason why mature mid-level professionals in one field prefer to cast a “blind eye” to convergence and stay in their comfort zone rather than acknowledge and take measures to improve their basic skills in mathematics and/or science. The latter prevents them from exploring training options to acquire new dimensions or pursue lifelong learning, as is necessary to create the type of multidisciplinary convergences important for healthcare and help build a knowledge economy.

8.6 Temporary Conclusion: Abundance of Data Yet Starved for Knowledge?

Patients want answers, not numbers. Evidence-based medicine must have numbers to generate answers. Therefore, analysis of numbers to provide answers is the Holy Grail of healthcare professionals and its future systems. Lack of action due to paralysis from analysis of risk associated with the complexities [200] in healthcare is no longer acceptable in view of spiraling costs. Generating data without improving the quality of healthcare service and extracting its value for business benefits [201] will not provide the return on investment (ROI). Distributed data and their relationships are dispersed in multiple network of systems or system of systems (SOS). The role of data analysis is central. The comatose stage of the Information Age due to data overload and information overdose is predicting its demise unless new ideas [202] emerge as its savior. The imminent death of the information age makes it imperative to better understand the systems age. The single most important system that deserves our attention in the twenty-first century is the healthcare ecosystem. The convergence of characteristics such as enterprise, innovation, research, and entrepreneurship (EIRE), often common in organizations with foresight in parallel with the vision to drive convergence of biomedical sciences, engineering, and information communication technologies, may act as the purveyor to advance healthcare for the progress of civilization [203].

Acknowledgment

Nothing in this chapter was invented by me. It is only original as a document because I have re-created ideas by borrowing from others. I am the combined effort of people who have inspired me. Hence, this chapter owes part of its origin to Bernard Lown, friend, cardiologist, and peace activist. The wisdom of Clive Granger, mathematician and economist, and the spirit of Glenn Seaborg, scientist and statesman, were
omnipresent, too. That I exist today is due to my wife, Rebecca Jane Austin Datta, chemist and educator, who forcibly took me to the hospital in May 2008, when we were in Europe. She “sensed” something was wrong. The majority of the content of this chapter was drafted while convalescing in the Anagh Ward of Kerry General Hospital in Tralee, Ireland, following an emergency cancer surgery on May 8, 2008, a few hours after arriving at the hospital, in a remote Irish hamlet. This chapter is dedicated to the exceptionally skilled surgeons Kevin Murray and Haris Shaikh as well as the nurses, aids, and staff of the surgical and the palliative care unit of the hospital in Ireland. The professional treatment and compassion of the Kerry General Hospital team during my recovery and subsequent chemotherapy is a shining example of the “art of healing” and the true ethos of global health care.

References

33. The Open University, www.open.ac.uk
34. The University of Phoenix, http://online.phoenix.edu/
35. Health Information and Quality Authority, www.hiqa.ie


57. UCLA Human Salivary Proteome Project, www.hspp.ucla.edu


60. Google, www.google.com


86. Cardiac Diagnostics, http://www.tastechip.com/cardiac/cardiac_diagnostics_research.html


121. Cardiovascular disease diagnostics kit, www.roche.com/prod_diag_cardio.htm


188. Harvard-MIT Health Sciences and Technology, http://hst.mit.edu
190. Molecular Medicine Ireland Clinician Scientist Fellowship Programme, http://www.dmmc.ie/MMI_Clinician_Scientist_Fellowship.htm
199. MIT Innovation, http://web.mit.edu/deshpandecenter


APPENDIX

I
THIS PAGE IS LEFT BLANK
Uniﬁed Theory of Relativistic Identiﬁcation of Information in a Systems Age:

Proposed Convergence of Unique Identiﬁcation with Syntax and Semantics through Internet Protocol version 6

Shoumen Palit Austin Datta

Research Scientist, Engineering Systems Division
Department of Civil and Environmental Engineering
Research Director & Co-Founder
MIT Forum for Supply Chain Innovation
School of Engineering
Massachusetts Institute of Technology
shoumen@mit.edu

March 2007

Shoumen Palit Austin Datta

Research Scientist, Engineering Systems Division, Department of Civil and Environmental Engineering and Co-Founder and Research Director, MIT Forum for Supply Chain Innovation, School of Engineering, Massachusetts Institute of Technology, Building 1, Room 1-179, Cambridge, Massachusetts 02139, USA

shoumen@mit.edu

Abstract. This paper proposes to utilize internet protocol version six (IPv6) to uniquely identify not only things (objects) but also processes, relationships (syntax, semantics) and interfaces (sensors). Convergence of identification with information using the 128-bit IPv6 structure offers $3.4 \times 10^{38}$ unique instances. It is not necessary that all instances must be connected to the internet or routed or transmitted simply because an IP addressing scheme is suggested. This is a structure for identification which (1) may improve revenue potential from data routing (P2P packet tracking) for telecommunication industries, (2) potential use in healthcare and in biomedical sciences, (3) scope of use in the semantic web structure by transitioning URIs used in RDF, (4) applications involving thousands of mobile ad hoc sensors (MANET) that demand dynamic adaptive auto-reconfiguration. This paper offers clues for innovation based on a confluence of ideas that may augment systems interoperability necessary for operational transparency in a global economy.

Keywords: Interoperability, Data, Information, IPv6, Semantics, Syntax, Security, MANET, Sensors, Healthcare, Biomedical Ontology, Decision Systems, RFID, EPC, Logistics, Adaptive Value Networks, Supply Chain, P2P, ZigBee, WiFi, WiMax, WiTriCity
1 Introduction

Data (bits) from unique identification of objects or things (atoms) are often helpful to the decision making process. Decisions, however, are often based on information that takes into account multiple factors. Physical objects and their unique identification may be one of many factors, as is the internet of things, from the perspective of a systems approach. Real-world decisions are often based on collective information gathered from multiple sources (or systems) that includes data (bits) about “things” (atoms) and processes associated with “things” which may be used in combination with a higher level domain that may eventually trigger a decision or execute an action, aided or unaided by a human. Currently, we do not have a globally unique mechanism to identify information derived from data originating from things (objects) and processes. Unique identification of information, hence, is an open question.

Information, to be of value, must be relative to the context of the process. In general, contextual information is of greater relevance in the decision making process or in decision systems. In this paper, I shall refer to such information as decisionable information.

Since information is key, one who holds information can use data, for profit, as a pay-per-use or pay-per-access service. Hence, unique identification of data has gained considerable momentum. Transmission of data is essentially the domain of routers. Routing data is an ubiquitous and essential function (real-time data, supply chain, emergency, medical results, networked entertainment, video-on-demand, energy-use optimization, any data) performed by products in boxes (for example, routers). Some corporations are likely to explore new revenue streams simply from use of raw data, for example, data-as-a-service or how to profit from data routing by providing access to data (pay-per-use hosted services). The next higher level for potentially higher revenue originates from processed data or information-as-a-service, in addition to physical data routing.

The transformation of data to information is made even more difficult by the inability of systems and software to comprehend or understand. Advances in systems interoperability [1], adoption of sophisticated analytical techniques [2] for forecasting
and risk analysis [4] and growth of the semantic web [4] infrastructure may stimulate in-network processing of data to boost the information-as-a-service business model by making sense of data and information relative to each other. These are new sources of revenue emerging from a function that is massive in scale but poorly regulated due to inadequate ability to document and charge for individual events and instances. The identification and identity structure necessary for a scale so massive calls for a system that is able to uniquely identify and assign identity to objects, process and decision layers. This paper explores the use of the structure and alphanumeric format [5] in Internet Protocol version 6 (IPv6), for this task.

2 Format of Internet Protocol version 6 (IPv6)

This paper advocates the most obvious distinguishing feature of IPv6 due to its use of much larger number of unique addresses. The size of an address in IPv6 is 128 bits, which is four times the larger than an IPv4 address. A 32-bit address space allows for $2^{32}$ or 4,294,967,296 unique addresses. IPv6 uses the 128-bit address space allowing for $2^{128}$ or 340,282,366,920,938,463,463,374,607,431,768,211,456 ($3.4 \times 10^{38}$) unique addresses. The relatively large size of the IPv6 address was designed to be subdivided into hierarchical routing domains that reflect the topology of the modern internet. The use of 128 bits allows for multiple levels of hierarchy and flexibility in designing hierarchical addressing and routing that is currently lacking in the IPv4-based internet (as well as 64-bit and 96-bit versions of the electronic product code or EPC).

IPv4 addresses are represented in the dotted-decimal format. This 32-bit address is divided along 8-bit boundaries. Each set of 8 bits is converted to its decimal equivalent and separated by periods. For IPv6, the 128-bit address is divided along 16-bit boundaries. Each 16-bit block is converted to a 4-digit hexadecimal, separated by colons (colon-hexadecimal).
The following is an IPv6 address in binary form:

0010000111011010000000001101001100000000000000000101110011101110110000010101010100000000111111111111101100010100010011100010110101111110000000010101010000000000000000001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
GPS, WiTriCity\(^1\)). Grid based ‘in-network’ processing functions may be pervasive with increasing diffusion of software as infrastructure, which may, in turn, offer systems the functionality to harvest and route distributed data from various sources to analytical engines running various hosted applications, analytics, predictive tools, forecasting algorithms, event management alerts, scheduling and planning updates. Output from online analytical processing (OLAP) engines may be made available to businesses on a pay-per-use and/or pay-per–access basis, with costs reflecting real-time versus near-real-time service delivery.

Hence, service may “mature” to provide “answers” and not only numbers (raw data). In this scenario the network is the business. With vast quantities of data (instances of data) that need accounting (if you are the service provider), it becomes critical to find a robust, globally feasible and easily adoptable modus operandi to contextualize and “number numbers” that does more than deliver data. Distributed data from multiple sources is dynamic, often sporadic and volatile yet must have unique identification, usually, to be valuable. The key to profitable service is to deliver information of value that can be uniquely identified, for example, in buckets or data cubes [7]. The ability to count and account for the identity of every instance of buckets on your network\(^2\) catalyses the profits if one can bill (charge) for every instance a new bucket is created.

---


2 Expanded routing and addressing capabilities that offer greater ability to control the path of traffic in order to direct the transmission of packets on the service provider’s network (hence billing for service or use) is a business driver for networks seeking new revenue streams. IPv6 offers this functionality by improving the scalability of multicast routing and by introducing a new type of address called “anycast” address. When used as a part of a route sequence, anycast permits a node to select which of the several internet service providers (ISP) is wants to carry its traffic (source selected policy). This is implemented by configuring anycast addresses to identify the set of routers assigned to the ISP (one anycast address per ISP). These anycast addresses can be used as intermediate addresses in an IPv6 routing header to direct a packet to be delivered via a particular provider or sequence of providers (alliances between ISPs). These new routing extensions in IPv6 are powerful tools for provider selection and mobility. If users are enabled to change routing, then it will increase competition between service providers because choices may be guided by cost factors and quality of service. [http://playground.sun.com/pub/ipng/html/IPNG-Paper.html](http://playground.sun.com/pub/ipng/html/IPNG-Paper.html).
It could be argued that distributed data identification, of the type referred above, may be achieved using distributed hash tables [8] commonly used in peer-to-peer (P2P) systems (e.g., Napster, Gnutella, Freenet, BitTorrent). However, the implications of using variants of consistent hashing to map key nodes may introduce the same type of nomenclature issues (disagreement) as with the date/format/EPC type keys. Node keys (identifier or ID) are linked to the keyspace between two keys, in other words, a node with ID \( i \) owns all the keys in the keyspace. The benefit of consistent hashing in distributed hash tables are due to the fact that alterations in the keyspace, i.e., addition or removal of nodes, changes one associated/adjacent key, leaving all other nodes unperturbed (in traditional hash tables changes require the entire keyspace to be re-mapped).

In the context of this idea, the change of nodes in the keyspace (addition or removal) may be viewed as independent events, for example “buckets” containing inventory data from a third-party supplier that is updated at regular intervals. In the pay-per-access business model, every instance or event of uploading or updating the data cube “bucket” with inventory data (for example hourly data after completion of batches of work-in-progress or WIP data) should incur a fee, payable by the user to the host network provider, as a fee for pay-per-access. It is unclear if event detection and link to pay-per-access for this type of itemization is possible through use of DHT.

In the same vein, the manufacturer (OEM) may periodically “ping” the bucket to determine inventory (supply) data. Each time the manufacturer “looks” in the bucket, there should be a mechanism to capture this event and the manufacturer (user) pays a fee to the service provider (network carrying the data and/or data cubes) for the service type: pay-per-use/access. The observation (inventory of specific item) could impact higher level WIP (work-in-progress) functions. For example, if 3 components are required for sub-assembly of a part then updating one SKU in the BOM (bill of materials) may not reflect a change of status of the sub-assembly inventory (this is no longer merely raw data, but information) because all 3 parts are required in a certain volume (ratio) to ensure that the sub-assembly can proceed and will produce the next higher level component, in question.
This is *contextualized* numbering of information where inventory of sub-assembly of component – *is the information* – dependent on inventory – *which is itemized raw data* – of individual components. While DHT is an advanced mapping tool that reduces bandwidth requirement to rapidly connect nodes in a vast network topology, it is unclear if DHT may be as useful as IPv6 to assign unique identification to every change in, or access to, data and information, that may profit from being *accounted*.

Hence, identification requires a unique way to *innumerate* packets or the data and the data holders (data cubes). Application of granular “innumeration” of data packets is in demand and likely to offer significant value for internet service providers (ISP) who are usually loathe to carry P2P (peer-to-peer) traffic on their networks because the P2P *modus operandi* (leaf to leaf) bypasses the current accounting practices of tracking packet flow (hence, revenue generation) which generally operates on the “trunk” of the system (charges for operations or content delivery that use bandwidth on the trunk of the ISP). Due to proliferation of P2P users who are taking advantage of the networks, ISPs are seeking tools to track and trace P2P traffic through the leaf nodes in order to charge P2P users for transactions (video, music, data) based on volume rather than the current flat-fee revenue model used by the telecommunication industries and ISPs because they do not yet have a granular mechanism to track how much data (packets) an user is downloading or uploading. A satisfactory revenue model based on tracking P2P usage in units of packets may further bolster revenue from legitimate P2P users and help reduce ISP bandwidth congestion by distributing payload to leaf nodes or mesh networks (away from trunk lines). To help with these accounting problems in P2P mesh networks, in one approach a file is reassembled in “slices” where a slice consists of the \( n \)th bit of every block. It is open to exploration whether in this context the numbering of files (the \( n \)th bit of every block) with an unique IPv6 id may offer business value through granularity of accountability.

---

3 Hui Zhang of Carnegie-Mellon University has founded Rinera Networks and Paul Francis of Cornell University has developed a system called Chunkyspread to partially address these problems (MIT Technology Review, March-April 2007).
3.2 Event Management

In event management and related supply chain operations, database tables with unique identification for each data cube may lend itself to re-use as data holders. Each event (that is, every time data or a data cube or a data holder is used) is uniquely numbered (companies subscribe to number domains) and itemized for pay-per-access service charges based on instances. Event data update and/or management may utilize online analytical and/or transaction processing (OLAP/OLTP) type systems for customer relationship and billing purposes, respectively. This proposal to use IPv6 may address the central dilemma: how do you number numbers? How does one number events? Merely “counting” the number of numbers is not enough. This will run into octillions of instances of data and even more instances of contextualized data that is of value (relativistic or decisionable information).

Consider problems faced by retailers if they double-count or re-count. Then imagine if a system-wide data duplication goes unnoticed and if services (finance, insurance) are duplicated or irregular. Individual pieces of data are often sterile as decisionable information but the collective analysis of data in the context of the process or transaction (where it is used) is far more critical and valuable as decisionable information. Object data identity using other formats, such as the electronic product code (EPC) or GUID (global unique id), are static identification formats that remain oblivious of context (process). Although EPC and other formats such as the UCR (unique consignment reference) offers unique id for things or objects, they lack structure to offer contextual information that may have unique identification.

The problem with “unique id” is addressed, grossly inadequately, by EPC Global and other related bodies (GS1) through the EPC-IS interface standard. Entities stored in EPC-IS are events, contextualized through “what/where/when/why/how” combination of parameters. Unique numbering for events rely on generating a combined key by using, for example, EPC+location+time. Pre-agreement of time, location coordinates and formats are core assumptions when creating such a key. It promotes the fallacy that such agreements are universal and universally accepted (explore how many
permutations and combinations are possible with different formats of times and dates used on different continents).

Therefore, this problem may benefit from innovation and further exploration of “numbering of numbers” to include EPC type information (if necessary) in a sub-layer of a multi-layer data and information aggregation model to provide unique identity for information. One practical example of data “layers” may be viewed as a substitute or replacement for EPC IS where a unique serial number (bottle of Aspirin with EPC) is combined with attributes (not designed for EPC association) such as [a] who checked the packaging, [b] where was it packed, [c] where was it manufactured, [d] what is the expiration date, [e] when was it shipped, [f] when was it received, [g] where was it shipped from, [h] where was it received. “Who” denotes a role and individuals with unique IPv6 id may be linked to role based authorization for accessibility purposes. Location, transportation vehicles, physical spaces (distribution center, retail store) may have unique IPv6 type id, too. The combined higher level information linked with the serial number of the bottle of aspirin (with its unique IPv6 id) with the “who-what-where-when” parameters is as simple as a “name” (John Smith) in a relational database linked to mailing address, phone number, date of birth, biometric data, social security number. If the “name” forms the skin of an IPv6 type crawl [9], it may extract the related data to generate information about John Smith. Alternatively, if the zip code or post code serves as the skin in a Web X.0 “search and discover” function, it may help track and trace all who are John Smith in Cambridge (02139 or CB2 1HQ). Thus, layers of unique IPv6 id representing object data and semantic data (who-what-when-where) collectively generates information when combined. The “simple sensed” semantic data (local distribution centre is a regional distribution centre) may not be mixed up with EPC type id. It may be worth the upheaval to forge a new direction by associating semantic data (such as, a retail store, that does not serve as an actuator) with IPv6 type identification to facilitate future search engines and “intelligent” agent-based systems to help track and trace not only the movement of objects but the linked data and processes in relation to information.
In this respect, it may be rather unfortunate that EPC has deviated from its insightful 128-bit structure that was proposed by Sanjay Sarma and Dan Engels (MIT, 1998). It may have successfully aligned EPC with IPv6 evolution. EPC 128-bit was a format for data (id) transmission but without routing yet it had the potential for convergence or structural amalgam with routing if translated to IPv6 format with substantial unique identification provision and transmission-routing capabilities for any RF mode (RFID, UWB, sensors) that can be plugged directly into the internet. Currently, sensor data cannot be uploaded directly to the internet because individual sensors cannot be assigned IP addresses due to limitations of the IPv4 unique addressing system. This proposal works with sensors (interfaces) because individual sensors can now have unique IPv6 address. However, the fundamental question is how to extract value from data and how can providing data as a network service evolve a robust revenue model?

Thus, data as a service model begs to find a mechanism that can number numbers in a rational manner to generate the higher order contextual and relativistic information. This mechanism must exist within the current framework of TCP/IP because it is aligned with the internet architecture. Hence, the suggestion to use the 128-bit IPv6, a scheme that is already at hand and being deployed, gradually but globally. This may be the “one-size-fits-all” hyper-id. IPv6 hyper-id may be a lucrative accounting tool in event recognition network delivered services to enable billing, for example, for high volume entertainment events, such as, iTunes from Apple. Management of iTunes billing, tracking and tracing, may be overwhelmed by sheer volume. It may also experience constraints not only for billing services but also up-selling/cross-selling types of customer relationship management (CRM) and marketing due to its inability to track and trace combinations of parameters that identify customer (nodes) choices.

3.3 Relativistic Information

In a similar vein, it may be necessary to identify people and process, in combination. How do we uniquely identify people and their linked or related context (process), for example in healthcare or in eGovernment functions (such as pension, voting).
EPC is object based identification. Think of IPv6 as a solution that is pre-agreed for global adoption. Consider contextual relevance of data and how to “number the numbers” but especially how to route numbers (data) which may be identical but with different identities. For example, a blood glucose result of 120 mg/dl may be identical for multiple individuals (same number, 120, with same units, mg/dl) but with different identities at the informational level because it belongs to different individuals. This may seem an esoteric brain teaser to some but *au contraire*. This is a fundamental information infrastructure issue that requires serious attention. It may have critical impact, for example, in a healthcare scenario where in-network data processing may be valuable with concomitant development of semantic tools for discovery and hosted analytics over semantic grid (Web X.0) services. However, in-network processing is granular but only for the application in question (per se it does not lead to higher order aggregation or network topology). Connecting data from, for example, from an emergency scenario (heart rate) with patient history may be viewed as a higher layer “abstraction” where not only the data (heart rate in bpm; corrected for age) but the information (heart rate and patient history) is identifiable and accessible through a “hyper-id” accessible via web services with linked information that is unique and uniquely identifiable. Thus, healthcare is a fertile ground for such applications. This suggestion may work with Agent-integrated security systems (to guarantee data security, confidentiality and privacy⁴) and monitoring systems to scan for errors (double counting, double dosage) as well as duplicate address detection (DAD).

---

⁴ Service providers can now offer secure services by using two integrated options in IPv6. One mechanism, termed “IPv6 Authentication Header” is an extension header which provides authentication and integrity without confidentiality. This may prevent a number of network attacks. The other mechanism, “IPv6 Encapsulating Security Header” provides data integrity and confidentiality. Authentication, data integrity and confidentiality are core elements of IPv6 and offers an improved option mechanism over IPv4 (not limited to 40 bytes). It permits IPv6 options to be used as functions which were not practical for IPv4 (authentication and security encapsulation for security and confidentiality, respectively). In addition, the use of anycast address enables service provider selection. This feature may be used for secure services (dedicated service provider) where the anycast address may not be altered and hence the data must be routed according to a source selected policy, eg: healthcare, customs, military and finance. (http://playground.sun.com/pub/ipv6/html/INET-IPng-Paper.html)
systems to prevent data binding irregularities that may have fatal consequences in healthcare. These applications may also develop to include “Medical Google” type web services for point-of-care reference where human experience or human-aided searches may be necessary to supplement diagnosis or diagnostics or automated information processing, distribution and execution systems.

The industry may not yet consider this a mature or profitable endeavour because it is erroneously concluded by some that the impact of this mechanism may not manifest soon enough. Japan and the Nordic countries have sufficient internet penetration and an increasing percentage of the population are e-savvy yet grey (over 65 year old with increasing demand on healthcare services, as one example). They may soon seek and need e-healthcare “sense then respond” systems (Figure 1) to stem down healthcare costs. The system must respect privacy, data identity and security. This problem is real and will require a solution if industrialized nations expect that their healthcare expenditure remain a reasonably small percentage of their gross domestic product.

Clearly a broad spectrum of applications will find it essential to “number numbers” in a manner where identity and relativistic identification lends itself to information and context including provision for quality of real-time\(^5\) service at the right-time, all the time, plus authentication, data integrity, confidentiality and privacy.

---

\(^5\) The buzz around “real-time” data is as intense as the inability of most businesses to extract value from real-time data (see [http://esd.mit.edu/WPS/esd-wp-2006-11.pdf](http://esd.mit.edu/WPS/esd-wp-2006-11.pdf)). However, the cost of network delays in data transmission can have hazardous or even fatal effects in emergencies or accidents. Thus, IPv6 offers a new capability by introducing \[a\] the 24-bit Flow Label and \[b\] the 4-bit Priority field in the IPv6 header to enable labeling of packets belonging to particular traffic “flows” for which the sender can request special handling for real-time applications. Currently some hosts and routers do not support this function.

\[a\] The 24-bit Flow Label in the IPv6 header is designed to be used by a host to identify packets for non-default quality of service (QoS) or “real-time” service. A flow is a sequence of packets sent from a particular source to a particular (unicast or multicast) destination for which the source requests special handling by intervening routers. The nature of the special handling may be conveyed to the routers by a control protocol (eg: resource reservation protocol) or through embedded information in the flow packet (eg: hop-by-hop option). There may be multiple active flows from a source to a destination, uniquely identified by combination of a source address and a non-zero flow label. Packets that do not belong to or not associated with any flow traffic, carry a flow label of zero. ([http://playground.sun.com/pub/ipng/html/INET-IPng-Paper.html](http://playground.sun.com/pub/ipng/html/INET-IPng-Paper.html))
3.4 Convergence with Semantic Web

The sub-title of this section is still really a question or a hypothesis, not a statement. The convergence is imminent because it is well nigh impossible to mention the word “information” without thinking about syntax, semantics, relations and other elements of interoperability, which are crucial for increasing the value of information to generate decisionable information.

[b] The 4-bit Priority field in the IPv6 header enables a source to identify the desired delivery priority of its packets relative to other packets from the same source. The Priority values are divided into 2 ranges: values 0 through 7 specify the priority of traffic for which the source is providing congestion control (traffic “backs off” in response to congestion). Values 8 through 15 are used to specify the priority of packets that does not back off in response to congestion (transmitted in real-time at a constant rate).
The ontological framework necessary for the semantic web uses a powerful tool, RDF (resource description framework), which works, in part, through use of URI (universal resource identifier). The latter is a higher layer abstraction that uniquely identifies a resource (even if the resource lacks a web address). These relationships must be pre-designed before RDF and RDF/XML can be applied.

These relationships, for example in the biomedical field, are referred to as “controlled ontological vocabulary” which must be created to plumb (search) the depths of the repertoire. This process generates even more syntax or labels to script the triples (subject-predicate-object) and introduces variability due to type of group, individual, nationality, mother tongue and a host of factors that may be difficult to standardize in an open environment (hence, routed through slew of committees for acceptance). The vast number of groups involved in creating ontologies uses a myriad of formats to specify application-specific domains. When these relation “trees” are consecutively added to higher level hierarchies (ontological soup ensues) the problem of keeping track of tags and meta data may create quagmire for which yet another standard may be sought! This problem is illustrated by the omics initiative [10] involving a handful of post genomic technologies (transcriptomics, proteomics, metabolomics). Several standardization groups are working on metadata and ontology in this sub-field.

If ‘omics’ may be defined as processes operating within a cell and the ‘cell’ subclass contains two related fields (intra-cellular and extra-cellular) and if cells are considered parts of ‘tissues’ and tissues are classified under ‘organs’ then, imagine the plethora of committees necessary to standardize and define biological molecules. Is there a feasible alternative to track and trace these ontology relationships that is amenable to machine intelligence and include use of metadata tags?

Hence, there are several deep issues involved in this simple example. Therefore, for the context of this argument, the focus is only on one sub-issue: the nature of the unique resource identifier (URI). Is there a benefit to transition the URI to a pre-agreed system that all groups may agree to use without further standardization? My naïveté (fools rush in where angels fear to tread) proposes the use of IPv6 to uniquely identify (map) URI layer from domains of pre-subscribed numbers which are assigned to classes and subclasses (illustrated in Figure 2 and Figure 3).
Table 1. An ever increasing number of standardization activities

<table>
<thead>
<tr>
<th>Community of effort and website</th>
<th>Standardization activities</th>
<th>Citation in this issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic research</td>
<td>Content (Minimal Information about a Genome Sequence: MIGS), syntax, and semantics</td>
<td>Field et al., 2006; Morrison et al., 2006a</td>
</tr>
<tr>
<td>International Nucleotide Sequence Database Collaboration (INSDC) (<a href="http://www.insdc.org">www.insdc.org</a>)</td>
<td>Content (INSDC Third-Party Annotation Submission Guidelines)</td>
<td>Cochrane et al., 2006</td>
</tr>
<tr>
<td>Genome Reviews: (<a href="http://www.ebi.ac.uk/GenomeReviews/">www.ebi.ac.uk/GenomeReviews/</a>)</td>
<td>Content (review of standardization within the Genome Reviews database)</td>
<td>Sterk et al., 2006</td>
</tr>
<tr>
<td>Post-genomic standardization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGED Society: (<a href="http://www.mged.org">www.mged.org</a>)</td>
<td>Content (Minimal Information about a Microarray Experiment: MIAME), syntax, and semantics</td>
<td>Ball and Brazma, 2006</td>
</tr>
<tr>
<td>HUPO-Protomics Standards Initiative (PSI) (<a href="http://psidev.sourceforge.net">http://psidev.sourceforge.net</a>)</td>
<td>Content (Minimal Information about a Protomics Experiment guidelines: MIAPE), syntax, and semantics</td>
<td>Taylor et al., 2006</td>
</tr>
<tr>
<td>Experimental Standards for Proteomics</td>
<td>Content (call for development of standard experimental mixtures of proteins)</td>
<td>Hogan et al., 2006</td>
</tr>
<tr>
<td>Metabolomics Society-MSI (Metabolomics Standards Initiative): (<a href="http://www.metabolomicsociety.org">www.metabolomicsociety.org</a>)</td>
<td>Content, syntax, and semantics</td>
<td>Feinberg et al., 2006</td>
</tr>
<tr>
<td>Integration activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting Structures for Biological Investigations Working Group (RSBI)</td>
<td>Contributions to content and semantics</td>
<td>Sansone et al., 2006</td>
</tr>
<tr>
<td>&quot;Env&quot; Community led by the Environmental Genomics Working Group (EGWG): (<a href="http://www.genome.gov/10010141">http://www.genome.gov/10010141</a>)</td>
<td>Content (MIAME/Env checklist), syntax, and semantics</td>
<td>Morrison et al., 2006b</td>
</tr>
<tr>
<td>National Center for BioMedical Ontology (NCBO): (<a href="http://www.bioontology.org">www.bioontology.org</a>)</td>
<td>Semantics</td>
<td>Robin et al., 2006</td>
</tr>
<tr>
<td>Other initiatives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MISFISHIE Working Group: (<a href="http://mgd.sourceforge.net/msfishe/">http://mgd.sourceforge.net/msfishe/</a>)</td>
<td>Content (MI Specification for In Situ Hybridization and Immunohistochemistry Experiments: MISFISHIE), syntax, and semantics</td>
<td>Deutsch et al., 2006</td>
</tr>
</tbody>
</table>
Class

Upperbody Outerwear

2007.db8.617.5ca.20a.95ff. 300d.2020

Subclass: Japanese Traditional Outer Garments
Range: 2007.db8.617.5ca.20a.95ff. 202a.7777 to 2007.db8.617.5ca.20a.95ff. 210f.9999

Uniquely Numbering URIs in the Ontological Framework
Extend (?) IP Addressing Strategies of Subnet & Subnet Masks to Label Class & Subclass Abstractions

Philosophy Reflection Noun Object Metaphor Observatory

2007.db8.617.5ca.20a.95ff. 999c
2007.db8.617.5ca.20a.95ff. 8ffe 888c
2007.db8.617.5ca.20a.95ff. 77x 95f.abcd. 999c
2007.db8.617.5ca.20a.95ff. BBAA abcd. 999c
2007.db8.617.5ca.20a.95ff. 889c

Fig. 2. Co-embedding unique id with syntax and semantics?

Fig. 3. Moving from descriptive or “analog” to unique or “digital” relationships?

Because IPv6 has a reasonable chance to be adopted globally, a mechanism to map URI abstractions to IPv6 structure offers synergistic convergence. It may reduce the gulf between business thinking and systems (ICT) specifications.

However, the creators of IPv6 pointed out that no matter how good a new protocol (or idea) may be, it may not matter if there isn’t a practical way to transition from one to another. There may be several known reasons why the suggestion in Figure 4 (below) of transitioning URI to an IPv6 format may not be feasible. On the other hand, unknown unknowns (reasons) may exist that may make this suggestion feasible or practical when and if the unknowns emerge as knowledge.

Fig. 4. Potential transition strategy?
3.5 Security

Another application area that may benefit from a vast library of detailed resources with unique identification that can be selected in a dynamic manner, is that of policy. Often, a “handbook” of policy may contain several thousand clauses and sub-clauses (assume that all these details exist as thousands of separate URIs). When a scenario surfaces (for example, US Department of Homeland Security certified Tier 2 business group authorized to import finished goods in containers, files a manifest that indicates additional inventory of sub-components) that calls for use of selected policies (to authorize a search of the container, in this case) then resources (that is, policies) that are applicable for that scenario can be selected (from a remote management location) by using the routing capability of IPv6 to choose sets of interfaces (with IPv6 addresses) from thousands of policies (from the policy domain) that may be involved in the analysis, assessment and management of security threats and risk (Figure 5).

![Risk Management through systems approach (convergence of data, policy, information)](image-url)
In a related security scenario, it may be significant to analyze the following: data acquired from sensors (data element 1) linked to a specific container (data element 2) transported on a vehicle (data element 3) driven by a credentialed driver (data element 4) who belongs to a logistics provider (data element 5) registered as a NVOCC (non-vessel operating common carrier) with a NVOCC code (data element 6). If customs and border protection (CBP) wants to analyze this sensor data at a local port of entry for targeting purposes, it is likely to assign a key, quite similar in concept and practice to EPC-IS (see section 3.2). In this scenario the key is a field or combination of fields used to “anchor” other fields. These methods may still be in use for “look data up” using key fields which are “arbitrarily created, for example, by creating some unique record ID or other unique count to distinguish otherwise unidentifiable data” [11].

It is, therefore, amply clear to understand why such arbitrary identification does not lend itself to portability and interoperability between systems. A clue uncovered and assigned a “key” at a port of entry by the US Coast Guard (USCG) remains “unique” in that system but may not be visible or meaningful in its association within the Automated Commercial Environment (ACE) or other targeting systems where non-obvious relationship analysis (NORA) is performed (for threat assessment and risk).

If NORA points to possible need for inspection, in a related scenario, confluence of policy becomes equally important to ensure execution of action is within legal limits. The laissez-faire concept of creating keys for supposedly unique identification needs a careful review, especially in threat assessment and risk analysis. Few “red handed” instances will be revealed through obvious relationships. Non-obvious relationship analysis is increasingly crucial to connect arcane associations that must be sorted from vast amount of essential and non-essential data, uniquely identified in a manner that can be accessed by any authorized system in any geographic location and analyzed repeatedly in different contexts in diverse domains to determine risk profile.

Data analysis for domestic (US) threat assessment in the transportation sector alone, could originate anywhere, in any form, in several instances, several times a minute or
hour, in the network connecting towns, cities, manufacturers and retailers, moving large volumes of goods and individuals through a system of approximately 4 million miles of roads and highways, 120,000 miles of rail road, 600,000 bridges, more than 300 tunnels and numerous sea ports, 2.2 million miles of pipeline, 500,000 train stations, 500 public airports operating more than 200,000 commercial aircrafts [12].

Table 2: Why IPv6 unique identification of information may be crucial for security and NORA
- 3.9 million miles of public roads
- 1.2 million trucking companies
- 15.5 million trucks
- 42,000 hazardous material (HAZMAT) trucks
- 10 million commercial vehicle drivers including 2.7 million HAZMAT drivers
- 2.2 million miles of hazardous liquid and natural gas pipeline
- 120,000 miles of major railroads
- 15 million daily riders on mass transit and passenger rail, nationwide
- 25,000 miles of commercial waterways
- 361 ports
- 250,000 containers per day
- 9.0 million containers through 51,000 port calls
- 11.2 million containers via Canada and Mexico
- 19,576 general aviation airports, heliports and landing strips
- 459 Federalized commercial airports
- 211,450 general aviation aircraft
- 77% of all flights are general aviation

3.6 Mobile ad hoc Sensor Networks (MANET)

By some estimates, the scope of unique identification using IPv6 is a staggering $6 \times 10^{19}$ addresses per mm$^2$ of the surface of the Earth, based on the fact that IPv6 supports $3.4 \times 10^{38}$ addresses and an approximation that $5 \times 10^{14}$ m$^2$ is the surface area of our planet. It may be argued that the number ($6 \times 10^{19}$ addresses per mm$^2$) represents a monolayer of interfaces (a node may have multiple interfaces, each with a globally unique IPv6 id) which is an unlikely scenario in internet devices, for example, an automobile or equipment such as HVAC (high voltage air conditioning) system. In a
vehicle or HVAC there may be a vast array of interfaces (nodes, sensors) packed in close proximity (multi-layer topology). Other internet devices or internet appliances (example: refrigerator, light switches) may have multi-layer topologies. Internet infrastructures, for example, buildings, may be densely packed with interfaces (for example, if every TV or light switch is an IP node, then think about the number of televisions and switches in Taipei 101, the tallest building in Taiwan).

Hypothetically, assume that the entire surface of the Earth may be organized as layers of interfaces 1 mm apart and that this hypothetical layer is 100 km in depth (consider sensors and actuators in deep sea drilling equipment and observatories on top of the Mount Everest). Even if we have such an improbable density of interfaces demanding globally unique id, the number of possible unique id is approximately $6 \times 10^{10}$ IPv6 addresses for $\text{mm}^2$ of Earth’s surface that is 100 km or $1 \times 10^9$ mm deep! In other words, 60 billion unique addresses per square mm of the Earth at a depth of 1 billion mm (Figure 6). This logic fuels the speculation that internet communication with Martians and objects in Mars may be within the scope of the design of IPv6.

![Fig. 6. Uniquely identifiable sensors](image-url)
This logic supports the extrapolation that IPv6 addresses may be used for unique id in mobile ad hoc sensor networks (MANET) in military applications where trillions of sensor nodes may be connected to logistics decision or monitoring systems of the war-fighter (Figure 7). These systems demand dynamic adaptive routing and auto-reconfiguration yet must retain critical data links, identity, coordination and interoperability with multiple commands or legacy (ERP) systems operating in vastly separated geographic locations.

Individual sensors may now be directly “plugged” in the internet using IPv6. Routing extensions in IPv6 support powerful functionalities like host mobility (route to current location) and auto-readdressing (route to new address) that may be useful for mobile ad hoc networks (MANET). The routing option is used by IPv6 source to list one or more intermediate nodes (in the network topology) to be “visited” on the path to the
destination of the packet. For dynamic and adaptive MANET, the management of host mobility and automatic readdressing functionalities may require convergence of IPv6 features in combination with mobile agents [13] or mobile cluster agents that can manage the “measure of distance” of routing protocols to a higher level of abstraction where granular data (packets) are aggregated (Cluster Agents) and decision or higher level information with unique IPv6 id is transmitted to one or more (unicast, anycast or multicast) superior node in the network (MANET) topology or linked to other decision systems (legacy, applications). Assignment of IPv6-based id (pre-subscribed domain) may be managed by higher order Cluster Agents. Because some of the activities may be closed loop or terminal, it is possible to reclaim underutilized assigned network id. The reclaiming process may follow a pattern of inactivity (loss of signal over extended period) and the recovery of id may be governed by elementary autonomic paradigms drawn from the AI (artificial intelligence) domain.

For consumer purposes, an example of MANET is a vehicle. Hence, transportation, logistics, supply chain and business, military, healthcare operations that need real-time data plus analysis of information may access granular data (unique data cubes) yet extract decisionable information in a manner that may be uniquely identified (traced) to its source. Track and trace identification is key to guarantee food-water safety and in customs operation where ad hoc assessment of threat or risk is crucial. Thus, IPv6 supported large hierarchical addresses will enable the functional growth of the internet and provide new routing capabilities that were not robust enough in IPv4.

4 Concluding Comments

The inclination to reap, with haste, the ‘low hanging fruits’ often derail convergences necessary for adaptive strategies [14] in business and in a broad spectrum of decision systems including logistics management [15]. Contextual numbering of numbers and contextual identification of decisionable information, using IPv6 represents, albeit only one element, but a potentially valuable confluence of identification, transmission and routing of data relative to information in systems to provide answers [16]. IPv6
may offer a plausible and simple mechanism not only to track global movement of bits (data) and atoms (physical goods) and to connect bits and atoms but also to connect \textit{bits to bits}. This suggestion has received some preliminary [17] support.
Acknowledgements. This idea germinated while attempting to explain to a class of international Master’s students the future potential for innovative tools in systems interoperability and visibility (www.student.chalmers.se/ka/hp?hp_id=3013). The author is indebted to the vibrant students in the ITR536 course for their reluctance to accept shoddy explanations of what constitutes and attributes necessary for, increasing visibility and in future, transparency, in geographically agnostic intelligent decision systems. This was during November 2006 while teaching the Master’s course in Supply and Demand Chain Design and Management (ITR536) at the Chalmers University of Technology in Gothenburg (Sweden), where the author was a visiting faculty in the Division of Logistics and Transportation in the School of Technology Management and Economics. The author is grateful to Professor Kenneth Lumsden of Chalmers University of Technology and Professor Gunnar Stefansson of Chalmers University of Technology and University of Iceland, for this rewarding invitation to teach and advise students at Chalmers University of Technology. Generous financial support for this period was provided by the Division of Logistics and Transportation, Chalmers University of Technology. The author’s teaching duties for ITR536 were particularly invigorating due to the kind support offered by Mr Ola Hultkrantz and Mr Per Medbo of the Division of Logistics and Transportation at Chalmers University of Technology (Goteborg, Sweden). In addition, the author has gained insights from discussions about this idea with Professor Gunnar Stefansson (University of Iceland and Chalmers University of Technology), Mr Andy Mulholland (Global Chief Technology Officer, CapGemini), Dr Joseph Salvo (Director, GE Global Research), Dr Robert Ciskowski (IBM Innovation Center), Professor Rodney Brooks (Director, Computer Science and Artificial Intelligence Laboratory, MIT), Professor John Williams (Director, Auto-ID Lab, MIT), Mr Micah Samuels (Senior Manager, Amazon), Dr Richard Swan (Chief Technology Officer, T3Ci), Dr Maria Azua (Vice President, IBM), Dr Ralph Droms (Chief Engineer, Office of the Chief Technology Officer, Cisco Systems) and Professor Joseph Sussman (J R East Professor, Professor of Engineering Systems and Civil and Environmental Engineering, MIT). This idea has received some support from Charles Simonyi of IntentSoft (formerly with Microsoft Corporation), Mark Greaves of
Vulcan (formerly with DARPA, US Department of Defense) and Professor Sanjay Sarma, MIT (Member of the Board of EPC Global and Co-Founder and former Research Director of MIT Auto-ID Center). Encouragement from my colleagues at the MIT Forum for Supply Chain Innovation (Professor David Simchi-Levi, Dr William Killingsworth, Ms Janet Kerrigan and Miss Megan Gately) is gratefully acknowledged. I am thankful to Ms Janice Hall for her help with uploading and updates on the MIT ESD website while this was a working paper.

References


(4) World Wide Web Consortium www.w3c.org


(9) Sarma, S.: personal communication


(12) Transportation Security Administration, Department of Homeland Security www.tsa.gov


(17) Simonyi, C.: personal communication
Management of supply chain: an alternative modelling technique for forecasting

S Datta\textsuperscript{1}, CWJ Granger\textsuperscript{2, **}, M Barari\textsuperscript{3} and T Gibbs\textsuperscript{4}

\textsuperscript{1}Massachusetts Institute of Technology, Cambridge, MA, USA; \textsuperscript{2}University of California at San Diego, La Jolla, CA, USA; \textsuperscript{3}Missouri State University, Springfield, MO, USA; and \textsuperscript{4}Intel Corporation, DuPont, WA, USA

Forecasting is a necessity almost in any operation. However, the tools of forecasting are still primitive in view of the great strides made by research and the increasing abundance of data made possible by automatic identification technologies, such as radio frequency identification (RFID). The relationship of various parameters that may change and impact decisions are so abundant that any credible attempt to drive meaningful associations are in demand to deliver the value from acquired data. This paper proposes some modifications to adapt an advanced forecasting technique (GARCH) with the aim to develop it as a decision support tool applicable to a wide variety of operations including supply chain management (SCM). We have made an attempt to coalesce a few different ideas toward a ‘solutions’ approach aimed to model volatility and in the process, perhaps, better manage risk. It is possible that industry, governments, corporations, businesses, security organizations, consulting firms and academics with deep knowledge in one or more fields, may spend the next few decades striving to synthesize one or more models of effective \textit{modus operandi} to combine these ideas with other emerging concepts, tools, technologies and standards to collectively better understand, analyse and respond to uncertainty. However, the inclination to reject deep-rooted ideas based on inconclusive results from pilot projects is a detrimental trend and begs to ask the question whether one can aspire to build an elephant using mouse as a model.

\textit{Journal of the Operational Research Society} advance online publication, 23 May 2007
doi:10.1057/palgrave.jors.2602419

**Keywords:** forecasting; supply chain management; multivariate GARCH; risk analysis; intelligent decision systems; RFID; sensor data

1. Background

Forecasting is an ancient activity and has become more sophisticated in recent years. For a long time, steady steps in a time series data set, such as simple trends or cycles (such as seasonals), were observed and extended into the future. However, now a mixture of time series, econometrics and economic theory models can be employed to produce several forecasts which can then be interpreted jointly or combined in sensible fashions to generate a superior value.

The variable being forecast is a random variable. Originally attention was largely directed towards the mean of this variable; later to the variance, and now to the whole marginal distribution. Pre-testing of the data to find its essential features has become important and that has produced modern techniques such as cointegration.

The horizon over which the forecast is attempted is also important, and longer-run forecasts are now being considered as well as forecasts of ‘breaks’ in the series.

\textsuperscript{*}Correspondence: S Datta, Engineering Systems Division, Department of Civil and Environmental Engineering, School of Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: shoumen@mit.edu

\textsuperscript{**}Co-author shared the 2003 Nobel Prize in Economics.

The question of evaluation of forecasts has also been greatly developed. Most forecasts are quite easy to evaluate but others, coming from the Global Models (which attempt to model the global economy), are more difficult. Business, commerce, global organizations and governments may find that these models, if explored, may offer valuable guidance for their forecasting activities or their attempts to improve accuracy of forecasts.

Thus, forecasting is a necessity almost in any operation. However, the tools of forecasting (software) in general business use are still primitive in view of the strides made by research. Hence, promoting advances in forecasting to aid predictive analytics is deemed a worthwhile endeavour and is the purpose of this paper. Such tools may further reduce uncertainty and volatility characteristic of global trade. The relationship of various business parameters that may change and impact decisions are so abundant that any credible attempt to drive meaningful associations are in demand by global businesses. This paper proposes some modifications to adapt an already advanced forecasting technique with the aim to develop it as a decision support tool applicable to a wide variety of operations including supply chain management (SCM).
2. Introduction

Total quality management (TQM) gained prominence in the 1970s by claiming to boost quality at a lower cost through proper management and operational design. Lean (manufacturing) was the euphoria in the 1980s following Toyota’s exemplary adoption of Just In Time (JIT) processes to enable flexible manufacturing and minimize costs by reducing inventory level. However, it soon became clear that manufacturing costs could not be reduced further by pursuing variations of TQM and JIT simply through classical operations research. Globalization of markets in the 1990s combined with improvements in ICT (information communication technologies) and short product life cycles shifted the focus on SCM that could adapt to demand or reduce costs through improvement in efficiency.

SCM is a set of approaches to efficiently integrate suppliers, manufacturers, distributors, warehouses and retail stores so that merchandise is produced and distributed in the right quantities, to the right locations, at the right time in order to minimize system wide costs while satisfying (customer) service-level requirements (Simchi-Levi et al., 2003). Viewed from this perspective, similarities exist between SCM practices and a competitive market economy. A market economy ensures that right mix of goods and services get produced (those that are most wanted by consumers) in the right way (ie least cost) and eventually distributed to the right people (those who are willing to pay the most). Therein lies the attractiveness of a market-based economy, that it gives rise to most efficient allocation of resources. Likewise, SCM, if appropriately designed and executed, may offer efficient business solutions, thereby minimizing costs and improving readiness or competitiveness.

Despite rapid advances in SCM and logistics, inefficiencies still persist and are reflected in related costs. During 2000, supply chain-related costs in the United States alone exceeded $1 trillion (10% of GDP), which is close to the 2005 GDP of Russia and Canada or the combined GDP of the 22 nations who are members of the oil-rich League of Arab Nations. A mere 10% savings of supply chain costs in the United States is close to the 2005 GDP of Ireland (Datta et al., 2004). More than US$3 trillion have been spent on global logistics in 2004 and this represents almost 5% of the global GDP or more than the GDP of Germany and just less than the GDP of India in 2005. Inefficiencies in the global logistics network estimated at an approximate of US$600 billion (close to the 2005 GDP of Australia) offers untapped opportunities for organizations to optimize or adapt to improve sustainable profitability. Hence, serious questions have been raised as to how to make decision systems more efficient in order to reduce cost of transaction (Coase, 1960, 1992). This, in turn, requires a thorough understanding of the factors that make design and operation of effective SCM strategy a challenging task due to the volatility of supply and demand.

It is challenging enough to design and operate a supply chain for one facility, in order that costs are minimized and service levels are maintained. The difficulty increases exponentially when the system as a whole is considered and system-wide costs must be minimized, that is, optimizing the interactions between various intermediaries, such as retailer, wholesaler, distributor, manufacturer and supplier of materials. This is mathematically equivalent to finding a global optimal solution as opposed to local optimization, the predominant business practise. Global optimization, involving several stages in the decision-making process, is far more complicated. It is also much broader in scope and encompasses local optimization, but only as a special case.

Some may argue, justifiably, that optimization itself is to be blamed for the inefficiencies in global SCM practices. Perhaps optimization suggests an innate assumption that operations are capable of reaching a steady state or equilibrium, once ‘optimal’ conditions are determined and executed. Global volatility, even in peaceful or stable political economies, may disprove this assumption. Hence, dynamic or recurrent real-time optimization is required and reflects the fundamental necessity of global supply chains to continuously adapt.

The task of global optimization is rendered difficult by the uncertainty of the business environment. First, businesses need to continually adapt to demand uncertainty and its impact on inventory management. In a market-based economy, production decision is primarily demand driven and must be made ex ante (before customer demand is realized). Furthermore, due to lack of information sharing ex post (after actual customer demand is realized) between partners, the variability in orders is amplified upstream, along the supply chain. This phenomenon is commonly referred to as the ‘Bullwhip Effect’ and it is a key driver of inefficiencies associated with SCM. It distorts the demand signals, resulting in costs in the form of excess capacity and inventory, need for increased storage, and transportation cost increases (due to less-than-truckload or LTL scenarios), to name a few (Lee et al., 1997).

The Bullwhip Effect and the resulting inefficiencies associated with traditional supply chains may be reduced, in theory, by centralizing information relating to supply and demand (Datta et al., 2004). In other words, a ‘centralized’ supply chain model is one where such information is made available to all participating businesses at various stages of the supply chain or network of partners. Advances in information and communication technologies in the past decade has made it easier to acquire, share, access and analyse data in a manner that is increasingly feasible for ‘sense and response’ systems. In the context of SCM, the idea is to enable intermediaries in the supply chain process to act as ‘infomediaries’ or serve as agents for sharing and accessing the real-time data flow through common interfaces, such as web-based services (Datta, 2006, 2007).

Acquisition of or access to data is not equivalent to use of decisionable information that can be extracted from data. Differences in forecasting methodologies applied (to the same
data) at different stages of SCM by the participants (process owners) may yield varying types of information that may further obscure the value of data or give rise to increased fluctuations, thus distorting the signals, such as demand (Lee et al., 1997). To rein in the Bullwhip Effect, one contribution may stem from a standardized forecasting model, that may be used by the supply chain partners, as an analytical tool to extract value from the data that is accessible to all the partners. Sharing of such an analytical engine by a group of businesses is possible through the use of grid computing (Datta, 2004). Although we have the tools and technologies at our disposable, the sluggish growth of collaborative systems such as CPFR (collaborative planning, forecasting and replenishment) may be indicative of only a mild interest in the benefits of collaborative information processing. It is quite possible that lack of trust between businesses and heightened data security risks may be slowing real-world implementations of valuable strategies such as CPFR.

While a standardized forecasting model applied to near real-time data in a shared or centralized database may try to tame the Bullwhip Effect, it may never be eliminated due to outlier events and inherent or unexplained variability. In SCM, innumerable sources of variability exist, including factors such as demand forecasting, variability in lead time, batch processing or bulk ordering to take advantage of transportation discounts, price variability due to product promotion or discount, to name a few.

Hence, the objective of this paper is to propose the potential use of an advanced statistical modelling technique for the purpose of forecasting (originally proposed by Datta, 2003, 2004). Based on the pioneering work on time series econometrics by Clive W. J. Granger and Robert F. Engle (Engle and Granger, 1987, 1991), this paper proposes a few modifications to the statistical model proposed by Robert F. Engle based on advances in time series econometrics. The modifications were introduced to make the model more amenable for use in decision support systems, such as, SCM. If the proposed modifications are indeed viable, it is expected to explicitly model the interactions between various intermediaries of SCM as well as the time varying (nonconstant) variability that manifests, at least in part, as the Bullwhip Effect. For example, our proposed model captures the cross-variable dynamics as reflected in interaction between supply chain nodes or stages (retailers, wholesalers, distributors) using vector auto regression (VAR) methodology which is essentially a model for the means of a vector process. The framework also explicitly models the time varying volatility (perhaps observed, in part, as the Bullwhip effect) by using a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) technique (Engle and Kroner, 1995). GARCH is a model for volatility of a single series, whereas multi-variate GARCH (MGARCH) is a model for volatility (variances and covariances) for a vector. Therefore, the proposed model may also be denoted as a VAR-MGARCH model.

While these techniques have been widely used in finance (and economics) in the past few decades (and also earned Engle and Granger the 2003 Nobel Prize in Economics), to our knowledge, they have not been applied or explored as decision support tools by supply chain planners or analysts in the area of SCM.

By using a dynamic model of volatility (defined as standard deviation of variance), a GARCH type model has the added advantage of providing a forecast of volatility in near term. Such a forecast may be useful in calculating value at risk (VaR). However, VaR is estimated with a simplifying assumption, such as (joint) normality. Consequently, ARCH technique has become an indispensable tool in risk assessment and management in the financial sector (Engle and Manganelli, 1999). Globalization of the supply chain has concomitantly increased the risk in the supply chain due to potential for loss of profits from over-capacity (cost of excess inventory) or opportunity lost due to out-of-stock situations. Hence, it is our contention that use of similar methodology in supply chain processes may enable businesses to better manage or even quantify the risk in the process.

VAR-GARCH type models require estimating a large number of parameters and hence cannot be used in practice unless a large sample of data is available. The lack of availability of high volume granular data may explain the scarce interest in applying this modelling technique as a forecasting tool in decision support systems. High volume accurate data is the single most important driver for forecasting accuracy. The recent rejuvenation of the use and adoption of automatic identification technologies (AIT) may partly ameliorate the lack of high volume data. The surge in the use of radio frequency identification (RFID) or ultrawideband (UWB) tags that may be embedded or attached with physical objects, will make it possible to track and locate objects along the entire supply chain, if the systems used by manufacturers, distributors, logistics providers and retailers are able to take advantage of middleware and hardware interoperability (software defined radio or SDR) to monitor, access and share near real-time data from RFID tags or sensors.

Thus, pervasive use of automatic identification may provide high volume object data in near real time with the maturing trend toward ubiquitous computing. Businesses may not have a clear understanding of how to use this data efficiently to extract decisionable and actionable information that offers business value not merely through savings but may actually increase profitability. We propose that businesses explore advanced techniques such as multivariate GARCH that requires high volume of data for estimation but offers the potential to deliver increasingly accurate forecasts along with a measure of VaR. Success in applying similar models to analyse financial market returns is well documented. Global SCM and any operation in need of planning for the future (healthcare, military, emergency) offers interesting applications for VAR-MGARCH techniques.
In the next section, we develop this statistical model for forecasting on a sequential, step-by-step basis, with the idea that independent variables represent operational ‘nodes’ (for example, the ‘stages’ or entities in supply chain). Section 4, discusses data requirements for model validation including the significance of automatic identification data. Risk in the global supply chain is qualitatively discussed in Section 5. Concluding thoughts are offered in Section 6 including comments about preliminary results obtained in (only) one study that explored the use of the modification proposed in this paper to simulate improvements in forecasting based on (only) one data source from an ongoing real-world operation.

3. Statistical model: VAR-MGARCH

Forecasting demand is a key tool in managing uncertainty. Forecast accuracy may depend on the understanding and coverage of parameters taken into account and the accuracy of historic data available for each variable that may have an impact on the prediction. In this section, we propose a statistical model that combines classical regression analysis with advanced time series techniques, hopefully to improve accuracy of forecasts.

3.1. CLRM

Classic linear regression models (CLRM) have been around for a century (Studenmund, 2000) and used for a variety of purposes including traditional supply chain planning software. CLRM may be expressed as follows:

\[ y_t = \beta_0 + \beta_1 x_t + \epsilon_t \]  

(1)

where \( y \) is the dependent variable of interest to be modelled for forecast (eg sales of a product, say aspirin); \( t \) the time period (frequency of observation, for example, \( t - 1 \) may indicate prior week 1, \( t - 2 \) week 2); \( \beta \) the coefficients to be estimated (based on values of \( y \) and \( x \)); \( x \) the explanatory variable that is used to ‘explain’ variations in the dependent variable \( y \) (for example, low sales of aspirin may be explained by low in-store inventory \( x \) of aspirin) and \( \epsilon \) the random (stochastic) error term.

This simple technique can model multiple independent or explanatory variables, that is, multiple \( x \)’s, since the variation in \( y \), say, sales of aspirin, is dependent on multiple parameters, such as inventory \( x_1 \), price \( x_2 \), expiration date \( x_3 \). The choice of \( x \)’s (number of explanatory variables) will drive the validity and accuracy of the model. \( x \)’s may be based on underlying economic principles (theoretical) or business logic (practical underpinnings). However, no matter how many \( x \)’s are included, there may be an inherent randomness in \( y \) that cannot be explained. Thus, the random error term \( \epsilon \) is included in the equation (admission of the fact that the dependent variable \( y \) cannot be modelled perfectly). The corresponding mathematical equation is given by Equation (2):

\[ y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \epsilon_t \]  

(2)

Objective of CLRM is to estimate the parameters \( (\beta_0, \beta_1, \ldots, \beta_K) \) of the model based on a sample of observations on \( y \) and \( x \), assuming that \( \epsilon \) is characterized by a normal distribution with mean \( 0 \) and variance \( \sigma^2 \) for all time periods \( t \). Normality assumption is needed for hypothesis testing with respect to \( \beta \)’s based on sample of data, unless the sample size is large.

\[ \epsilon_t \sim N(0, \sigma^2) \]

Given the multiple sets of \( (\beta_0, \beta_1, \ldots, \beta_K) \) may be estimated, the objective of CLRM is to choose that set of \( (\beta_0, \beta_1, \ldots, \beta_K) \) which minimizes the sum of squared residuals \( (\epsilon_1^2, \epsilon_2^2, \ldots, \epsilon_N^2) \):

\[ \sum_{t=1}^{N} \epsilon_t^2 \]

where \( \epsilon_t \) = empirical counterpart of \( \epsilon \) (and is estimated based on sample data). Intuitively, this amounts to finding a line that best fits the data points by minimizing the sum of squared vertical distances of the actual data points from the fitted line. Thus, residuals are essentially in-sample forecast errors as they measure the difference between actual \( y \) and fitted \( y \). This technique is commonly referred to as the principle of ordinary least squares (OLS) and widely used due to its simplicity. The attractiveness of CLRM-based OLS forecasting stems from the fact that we can model cross variable linkages. This feature is especially useful to carry out ‘what-if’ analysis. For example, what may happen to sales \( y \) (the dependent variable) of aspirin-based painkillers in retail sales if the in-store inventory of non-aspirin painkillers were increased by 10%? Clearly, ‘what if’ analysis is conditional upon assumptions we make about \( x \)’s in the model. Therefore, in building this model, the choice of \( x \) is a process decision based on the model builder’s knowledge about an operation or business or industry.

One may wonder if we are playing a ‘what if’ game or is 10% increase, cited above, a real-world scenario. The retail outlet surely knows what has happened in the past. This segues to the next phase in the development of our statistical model where it is no longer necessary to assume values of the explanatory variable to forecast \( y \) (the dependent variable). Instead of inserting arbitrary values for future \( x \)’s (such as a 10% increase), we start by forecasting the values of \( x \) based on its own past data to obtain an unconditional forecast for \( y \). In this stage of model development, the regression technique gets intertwined with time series techniques. By fitting a univariate autoregressive model to \( x \) where we use past (lagged) values of \( x \) to forecast \( x \), we obtain the following equations (for \( x_{1t}, \ldots, x_{Kt} \)):

\[ x_{1t} = \alpha_{10} + \alpha_{11} x_{1t-1} + \alpha_{12} x_{1t-2} + \cdots + \alpha_{1N} x_{1t-N} + u_{x_{1t}} \]

\[ x_{Kt} = \alpha_{K0} + \alpha_{K1} x_{Kt-1} + \alpha_{K2} x_{Kt-2} + \cdots + \alpha_{KN} x_{Kt-N} + u_{x_{Kt}} \]

(3)
Rewriting using general notation:

\[ y_t = \beta_0 + \sum_{i=1}^{N_{it}} z_{1i} x_{1t-i} + \cdots + \sum_{i=1}^{N_{it}} z_{ki} x_{kt-i} + \epsilon_t \]  

(4)

or

\[ y_t = \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{it}} z_{ki} x_{kt-i} + \epsilon_t \]  

(4a)

where \( x_{1t} \) is the variable \( x_1 \) at time \( t \) (for example, we used \( x_1 \) for inventory, thus \( x_{1t} \) is inventory at time \( t \)), \( x_{kt} \) the variable \( x_k \) at time \( t \) (up to \( k \) number of \( x \)'s), \( x_{1t-1} \) the value of \( x_1 \) at time \( t-1 \) (referred to as the lagged value by one period), \( N \) the period up to which the lagged values of \( x_1 \) will be used in the equation, \( u \) the random error term.

Note that \( \beta_0 \) in Equation (4) is a combination of constants from Equations (2) and (3), respectively.

In Equation (3), \( z_{11}, z_{12} \) are coefficients of \( x_{1t-1}, x_{1t-2} \) and are referred to as lagged weights. An important distinction is that instead of arbitrarily assigning weights, these coefficients are estimated using OLS. The error term in Equation (3) represented by \( u \) is analogous to \( \epsilon \) in Equation (1). Depending on the number of \( x \)'s \( (x_1, \ldots, x_K) \) that adequately represents the process being modelled in Equation (1), there will be \( K \) number of equations as given by (3) that must be estimated to forecast the \( x \)'s \( (x_1, \ldots, x_K) \) which will then be used to obtain an unconditional forecast of \( y \). Thus, to simplify the task, we can estimate all the parameters \( (z, \beta) \) simultaneously by re-writing Equation (1), the basic CLRM equation, as Equation (4) or its shortened version, as in Equation (4a).

Equation 4 is a step toward forecasting the dependent variable \( y \) with greater accuracy using forecasts of \( x \)'s based on historical data of \( x \)'s (lagged values). But, it is also clear that Equation (4) ignores the impact on \( y \) of the past values of \( y \) itself (lagged values). Consequently, a preferable model will include not only lagged values of \( x \) but also lagged values of \( y \), as shown in Equation (5).

\[ y_t = \beta_0 + \sum_{j=1}^{N_{yj}} \phi_j y_{t-j} + \sum_{k=1}^{K} \sum_{i=1}^{N_{yki}} z_{ki} x_{kt-i} + \epsilon_t \]  

(5)

In moving from conditional to unconditional forecasts of \( y \) using a time series model, we are increasing the number of parameters to be estimated. In Equation 2, we estimate \( K \) parameters \( (\beta_1, \ldots, \beta_K) \) excluding \( \beta_0 \). In Equation 3, we estimate \( n \) parameters \( (z_1, \ldots, z_N) \) excluding the intercept \( \beta_0 \) for each of the \( K \) number of \( x \)'s \( X_1, \ldots, X_K \). In Equation 5 we estimate \( j \) parameters for lagged values of \( y_{t-j} \) \( (\phi_1, \ldots, \phi_j) \) in addition to all the parameters for Equation 4. If we set \( K = 5 \) (five explanatory variables, the \( x \)'s), \( N_{yj} = 10 \) (number of lagged values to forecast the \( x \)'s) and \( N_{yki} = 10 \) (number of lagged values of \( y \)), then, we have increased the number of parameters to be estimated from 5 in Equation (2) to 50 in Equation (4) to 60 in Equation (5) (excluding intercept). To drive precision to the next (logical) step, equation 5 may be expanded further to include the important real-world observations regarding trend, seasonality and other cyclical dynamics. Businesses struggle to uncover ‘trends’ and once found, they are avidly pursued, often for short term gains but increasingly with less than stellar results due to fickle customer preferences.

3.2. GARCH

Thus far, our discussions have centred on CLRM in conjunction with time series techniques. CLRM is based on a set of assumptions mainly about \( \epsilon \), that, when satisfied, gives rise to desirable properties of the OLS estimates. Needless to emphasize, in the real world, these assumptions are almost always violated. Developments in time series, over the past couple of decades, have addressed the challenges that stem from the violation of some of these classical assumptions leading to inaccurate forecasts.

One of the assumptions often violated in practice relates to homoskedasticity (homoskedasticity≈variance≈mean squared deviation \( (\sigma^2) \), a measure of volatility) or constant variance for different observations of the error term. Forecast errors are frequently found to be heteroskedastic (unequal or non-constant variance). For example, in multi-stage supply chains, the error associated with manufacturer’s forecast of sales of finished goods may have a much larger variance than the error associated with retailer’s projections (the assumption being that the proximity of the retailer to the end consumer makes the retailer offer a better or more informed forecast of future sales through improved understanding of end-consumer preferences). The upstream variability reflected in the Bullwhip Effect violates the basic premise of CLRM, the assumption of homoskedasticity. CLRM ignores the real-world heteroskedastic behaviour of the error term \( \epsilon_t \) and generates forecasts which may provide a false sense of precision by underestimating the volatility of forecast error.

Homoskedastic and heteroskedastic error term distributions are illustrated in Figure 1. In a homoskedastic distribution, all the observations of the error term can be thought of as being drawn from the same distribution with mean \( \mu \) and variance \( \sigma^2 \) for all time periods \( (t) \). A distribution is described as heteroskedastic when the observations of the error term may be thought of as coming from different distributions with differing widths (measure of variance). In supply chains, the variance of orders is usually larger than that of sales. This distortion tends to increase as one move upstream from retailer to manufacturer to supplier. Therefore, the assumption of heteroskedasticity seems more appropriate as a characteristic that may be associated with the Bullwhip Effect.

While variance of error term may change across cross-sectional units at any point in time, it may also change over time. This notion of time varying volatility is frequently observed in financial markets and has been the driving force behind recent advancements in time series techniques. Robert
Engle is credited with the observation that not only is volatility non-constant (of financial asset returns), it also tends to appear in bursts or clusters. Instead of considering heteroskedasticity as a problem to be corrected (approach taken by CLRM practitioners in assuming homoskedasticity of error term), Robert Engle seized this opportunity to model this non-constant time-dependent variance (heteroskedasticity) using an autoregressive moving average (ARMA) technique.

ARMA has been in use for several decades and is a combination of AR (autoregression) and MA (moving average) techniques. We have already invoked autoregressive (AR) representation in Equations (4) and (5). AR links the present observation of a variable to its past history, for example:

$$y_t \rightarrow y_{t-1}, y_{t-2}, \ldots, y_{t-p}$$

where $p$ is the order of the autoregressive process AR$(p)$ or the period up to which the historical data will be used (a determination made by using other statistical tools).

Thus, AR is a technique by which a variable can be regressed on its own lagged values. For example, today’s sales $(y_t)$ may depend on sales from yesterday $(y_{t-1})$ and the day before $(y_{t-2})$. AR$(p)$ is appealing since it links the present to the past. MA expresses observations of a variable in terms of current and lagged values of squared random error terms $\epsilon_t$, $\epsilon_{t-1}$, $\epsilon_{t-2}$, $\ldots$, $\epsilon_{t-q}$ where $q$ is the order of the moving average process MA$(q)$. Combining AR$(p)$ and MA$(q)$, we get ARMA$(p,q)$ where $p$ and $q$ represent the lagging order of AR and MA, respectively.

Robert Engle used the MA technique to model the time varying volatility in a series and proposed the so-called Generalized AutoRegressive Conditional Heteroskedasticity model or GARCH. The GARCH technique represents a parsimonious model than ARCH, while allowing for an infinite number of past error terms to influence current conditional variance. Hence, GARCH is widely used than ARCH.

GARCH evolved when Tim Bollerslev extended the MA$(q)$ representation of $\sigma_t^2$ (the ARCH model) to include an AR$(p)$ process, that is, regressing a variable $(\sigma_t^2)$ on its own (past) lagged values $(\sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2)$ as well. Thus, variance of the random error term in a certain period $(\epsilon_t)$ can be modelled to depend not only on squared past errors $(\epsilon_{t-1}^2, \ldots, \epsilon_{t-q}^2)$ but also on the lagged value of the variance $(\sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2)$ as shown in Equation (7) below.

$$y_t = \beta_0 + \sum_{j=1}^{N_y} \theta_j y_{t-j} + \sum_{k=1}^{K} \sum_{i=1}^{N_{X_k}} \alpha_{ki} X_{kt-i} + \epsilon_t$$

$$\sigma_t^2 = \theta_0 + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \tau_j \sigma_{t-j}^2$$

Thus, GARCH may enable supply chain practitioners to model the volatility in the supply chain, a phenomenon documented by the Bullwhip Effect. How GARCH may help calculate the VaR for various supply chain stages deserves deeper investigation. Future research may reveal a mechanism to quantitatively determine the risk associated with various supply chains. The latter tool, when developed, may be of considerable value in general risk management in the globalized world of international commerce.

3.3. VAR-GARCH

In developing the GARCH model, Equation (7) takes into account the lagged values of the dependent variable (sales), the impact of multiple explanatory variables ($K$ number of $x$’s that influence sales such as inventory, price) and their respective lagged values, as well as time-dependent heteroskedasticity of the error term. But, thus far, we have not considered the fact that to predict sales $h$ periods ahead, it is also crucial to model the interaction between the entity level nodes (manufacturer, supplier, distributor in supply chain) which can impact sales.

In any operation, including supply chains, interaction between partners can influence any outcome (profit, service,
readiness, response). The strikingly different business ‘clockspeed’ and dynamics of the supply chains partners is what partly fuels the Bullwhip Effect. Thus, to incorporate the dynamics of interaction between players, it is essential to explicitly model the dynamics between the entities to be a useful real-world model. A combination of vector autoregression technique with GARCH captures this dynamics. Vector autoregression (VAR) is a model for the means of a vector process and was developed over a quarter century ago by Sims (1980). Previously, we discussed AR(p) with respect to Equation (5), which is a univariate model. In contrast, VAR(p) is a n-variante model where we estimate n different equations (for \(Y_1, Y_2, Y_3 \ldots Y_n\)). In each equation, we regress a variable on \(p\) lags of itself as well as \(p\) lags of every other variable in the system. Thus, the right-hand side variables are the same for every equation in the system.

The key advantage of VAR lies in its ability to capture cross-variable dynamics (vector process). For example, future sales (prediction) of Michelin brand tyres may not be precisely forecasted by Sears unless the store takes into consideration the events or sales (vector) at the distributor. Thus, there are at least two parties (vectors) in this example (interaction between retail store and distributor). To model this cross variable dynamics of \(n = 2\) using VAR(1), let us assume that \(p = 1\) (lagged by 1 period). Equation (7) may be extended to the VAR-GARCH type model for two entities with \((n = 2, p = 1, q = 1)\) as shown in Equation (8).

\[
\begin{align*}
\gamma_t &= \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{ai}} \epsilon_{ki} X_{kt-i} + \varphi_{11} Y_{t-1} + \varphi_{12} Y_{2t-1} + \epsilon_{1t} \\
y_{2t} &= \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{ai}} \epsilon_{ki} X_{kt-i} + \varphi_{21} Y_{t-1} + \varphi_{22} Y_{2t-1} + \epsilon_{2t} \\
\sigma_{1t}^2 &= \tau_{11} \epsilon_{1t}^2 + \tau_{12} \epsilon_{2t}^2 + \tau_{11} \sigma_{1t-1}^2 \\
\sigma_{2t}^2 &= \tau_{22} \epsilon_{2t}^2 + \tau_{22} \sigma_{2t-1}^2
\end{align*}
\]  

In the VAR-GARCH model represented by Equation (8), this dynamics is captured by estimating the coefficient \(\varphi_{ij}\) which refers to changes in \(y_i\) with respect to \(y_j\). For example, if \(\gamma_1\) represents Michelin tyre sales at Sears retail store and \(y_2\) represents Michelin tyre sales at the distributor, then the parameter \(\varphi_{12}\) refers to changes in sales at retail store (\(\gamma_1\)) with respect to sales at the distributor (\(y_2\)). If any one of the two random error terms (\(\epsilon_{1t}\) and \(\epsilon_{2t}\)) changes, it will impact both the dependent variables (\(\gamma_1\) and \(y_2\)). In terms of Equation (8) above, if \(\epsilon_{1t}\) changes, it will change \(\gamma_1\) and since \(\gamma_1\) also appears as one of the explanatory variables for \(y_2\) in the equation, the change in any error term impacts both dependent variables in this VAR representation. This cross variable dynamic interaction has thus far been ignored by current modelling practices for forecasting. The VAR component of the proposal in this paper is closer to the real-world scenario and VAR-GARCH may make it possible to quantify such cross-variable dynamics.

### 3.4. Multivariate GARCH (MGARCH)

To move beyond the realm of univariate autoregression to a vector autoregression system, for further precision of forecast, it is necessary to model time varying conditional covariance (measuring the degree of association between any two variables) between \(\epsilon_1\) and \(\epsilon_2\) in addition to time varying conditional variance of the error term. In other words, the error terms associated with the retailer’s sales forecast and the distributor’s inventory level may be correlated (Granger and Swanson, 1996). This type of multivariate interaction is not explicitly captured by the VAR-GARCH model (Section 3.3), yet in the business world the association between, say, sales forecast and inventory level, is crucial for the overall efficiency and profitability of the supply chain. Thus, the next task is to combine the VAR representation with a multivariate GARCH component. Assuming \(p = q = 1\), MGARCH specification can be expressed as follows:

\[
\begin{align*}
y_{mt} &= \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{ai}} \epsilon_{ki} X_{kt-i} + \sum_{l=1}^{2} \varphi_{ml} Y_{lt-1} + \epsilon_{mt} \quad \forall m = 1, 2 \\
\sigma_{1t}^2 &= C_{11} + \theta_{11} \epsilon_{1t-1}^2 + \tau_{11} \sigma_{1t-1}^2 \\
\sigma_{12t} &= C_{12} + \theta_{12} \epsilon_{1t-1} \epsilon_{2t-1} + \tau_{12} \sigma_{1t-1} \sigma_{2t-1} \\
\sigma_{2t}^2 &= C_{22} + \theta_{22} \epsilon_{2t-1}^2 + \tau_{22} \sigma_{2t-1}^2
\end{align*}
\]

where \(\sigma_{12t}\) indicates conditional covariance between \(\epsilon_1\) and \(\epsilon_2\) in time period \(t\), based on information set available up to period \((t-1)\).

Thus, the conditional variances and conditional covariances will depend on their respective lagged values, as well as the lagged squared errors and the error cross products. Clearly, estimating such a model may be a formidable task, even in a bi-nodal scenario, for example, a retailer and distributor. If the GARCH system is functional, it may be used to better analyse Impulse Response Function (IRF). At present, IRF values are of limited use because it is difficult to provide confidence intervals for the values. Confidence intervals are necessary for forecasts. GARCH values may provide these confidence intervals. IRF traces the impact of changes (‘shock’) in error terms on the dependent variable for several periods in the future. Applied to operational planning, IRF may offer insight about ‘sense and respond’ scenarios. IRF simulation may enable exploration of multi-component ‘what if’ scenarios by creating challenges and learning (from simulation) how to prepare (readiness) for such challenges (hurricane, fire, flood, earthquake, epidemics, pandemics, military escalations).

### 3.5. Is there a link between Bullwhip effect and GARCH processes?

We have often used ‘volatility’ to indicate the observation of fluctuation represented by the Bullwhip effect but it is unclear if there is an actual link between Bullwhip Effect and GARCH processes. Simply going ‘along’ the supply chain, there may
not be an use for GARCH but going over ‘time’ there might be, as explained below.

Consider a supply chain with a sequence of stages or locations: \( L_0 \) (origin), \( L_1 \) (stage one or first location), \( L_2 \) (second), \ldots, \( L_f \) (final stage or end). Goods moving along the chain are associated with a number of delivery times between these locations. Let us denote \( T_{j-k}(t) \) as the time taken to deliver a good from location \( j \) to location \( k \), the goods having started at time \( t \) at \( L_0 \) (origin). The 1st leg of the chain takes time \( T_{0,1}(t) \), the 2nd leg \( T_{1,2}(t) \) and so forth. These \( L \) values are positive random variables, possibly auto correlated, but initially considered as an independent sequence. Note that,

\[
T_{0,k}(t) = \sum_{j=0}^{K} T_{j,j+1}(t)
\]

is, essentially, a random walk, with an increasing mean and variance. If all the \( T_{j,k+1}(t) \) are uncorrelated with mean \( m > 0 \) and variance \( \nu \), then \( T_{0,k}(t) \) will have mean \( km \) and variance \( kv \). As \( k \) increases, volatility will increase, which is the Bullwhip Effect (there is no need to use GARCH models to fit this process).

The total time taken for the supply chain \( T_{0,f}(t) \) will, as \( t \) varies, generate a time series which can be analysed. In the unlikely event that the chain does not change, this will be a stationary series, but it is likely that volatility (Bullwhip Effect) will be experienced by the chain. Thus, an AR-GARCH model may be appropriate.

4. Data

The modelling technique proposed above, may represent an opportunity to apply advanced statistical and econometric tools to improve the quality of predictive analytics in general and supply chain forecasting, in particular. However, validating such a model requires high volume data and involves estimating a large number of parameters. It is possible that advanced organizations, such as the military establishments, may have considered using these techniques but could not substantiate the models due to fewer than necessary reliable data points (degrees of freedom).

However, data ‘points’ may no longer be the limiting factor if the increasing interest in adoption of AIT is transformed to reality. Widespread adoption of AIT (such as RFID or UWB tags and sensor data) may pave the way for use of real time data to validate a model such as the one proposed in this paper. The innovative convergence of fields as diverse as AIT and time series econometrics may improve decision support systems in domains beyond finance and economics (Datta, 2004).

AIT and progress toward embedding intelligence into physical objects may allow them to communicate with each other (thing-to-thing) as well as with business systems or users (consumers) in near real-time. Hence, businesses may soon be faced with ultra high volume multi-gigabit data streams that may be expressed succinctly only in terms of exabytes per second (1 exabyte = \( 10^{18} \) bytes or \( 10^9 \) gigabytes). Infrastructure necessary to acquire such data may not offer a satisfactory return on investment (ROI) unless decisional information derived from this data offers value or profitability. The question of value from high volume data may be considerably enhanced by using data in advanced statistical models (as proposed above) to yield useful analytics.

Availability of near real-time data at the right time may be especially useful for industries where historical data is an agonizing cliché due to short product life cycles, such as mobile phones, digital cameras and laptop computers, which are characteristic of high ‘clockspeed’ industries (Fine, 1998). For a product with a sales life cycle of 200 days (about 6 months), if we use data from the past 100 days (more than 3 months) in the time series model, it may be difficult to ‘change course’ and respond or adapt (based on forecasts or predictions from such models). This is where the granularity of high volume AIT data from RFID tags offers the potential to deliver real business value and ROI.

Re-consider the above example but assume the availability of high volume accurate AIT data (from RFID tags on high value products with rapid obsolescence). The data from RFID tags may be modelled with \( N = 100 \) where data is lagged every hour \( (N = 100 \text{ hours instead of } N = 100 \text{ days}) \). However, whether the quality of the information that may be extracted from such data, may change if \( N = 100 \) is in hours or days, is a business question, not a technology or analytics issue. Consequently, whether high volume data of a certain granularity is sufficient for reliable forecasts will depend on process. If the hourly data is used \( (N = 100) \), then predictive analysis can be made available within 5 days from launch of a product with 195 days (97.5%) of its sales life cycle still viable, in case it is necessary to re-engineer the product in order to respond to or meet customer preferences. If compared to daily batch data with \( N = 100 \), analytics may be available after 100 days or with only 50% of the product sales life cycle still viable.

Thus, use of high volume real-time data in these models may make it possible and feasible for sales, marketing, production or distribution to adapt in real-time or at the right-time. Changes can be initiated, based on forecasts, earlier in the (sales) cycle of the product or even at the production stage, by using delayed product differentiation strategies, if products were designed with modular architecture or if the product lifecycle was carefully optimized by balancing the demands of development versus fulfillment supply chain parameters.

Regarding estimation technique, the OLS technique, although simple, may not be preferred for use with GARCH. OLS technique proceeds by minimizing sum of squared residuals but residuals, by definition, do not depend on the parameters of the conditional variance equation. Thus, in the presence of GARCH specification, minimizing residual sum of squares is no longer an appropriate objective. Instead, to estimate models from the GARCH family, the maximum likelihood estimation (MLE) is the technique of choice. However,
under an assumption of normality, MLE is simply generalized OLS.

MLE works by finding the most likely values of the parameters given the actual data. Multivariate GARCH models are similar to their univariate counterparts and thus MLE technique can be used. However, due to explicit modelling of conditional covariances over time in MGARCH, the number of parameters to be estimated increases exponentially. A few different MGARCH specifications have been proposed, such as the VEC model proposed by Bollerslev et al. (1988) and the BEKK model proposed by Engle and Kroner (1995). This is an area that warrants deeper exploration keeping in mind increased data availability through use of AIT data acquisition tools.

5. Implications for risk management

Risk in SCM originates from two key areas: supply and demand. At the next level of equal importance are environmental, political, process and security risks. Political and environmental risks may always remain amorphous and refractory to adequate quantification. Security risks are even more volatile but on a higher priority level that demands advanced risk management tools and analysis for targeting operations in global trade.

Too often, risk is viewed as simplistic as merely the product of frequency and consequence. A high-frequency but low-consequence event (currency exchange rates) is viewed as similar to a low-frequency but high-consequence event (sinking of a cargo ship laden with spare parts). In reality, such apparently ‘similar risks’ may have vastly different effects. Sensational risks grab attention and beg for resource-consuming mitigation while risk managers tend to ignore the smaller risks that create the real friction in the supply chain. With the increasingly complex business environment that is the hallmark of globalization, today’s supply chain presents a myriad of specific risks ranging from external sources (such as terrorist strikes or vulnerability to political instability in developing countries due to global outsourcing) to internal sources (pressure to enhance productivity and reduce costs by eliminating waste, removing duplication through use of single source supplier). If accounted as parameters in traditional optimization equations, the sheer number of factors will exponentially increase the state space and as a result may grind the computation of the optimization algorithms to a pace that may become unacceptable for decision support systems to aid in the management of supply chain adaptability.

The VAR-MGARCH model proposed here may be well suited to take into account the details of the operational nodes (assuming we have data available from each of these nodes/processes). Recurring analysis performed in near real-time (assuming real-time data is available to the analytical engine) may offer results that predicts or detects risks in the operational model (supply chain) far in advance of what is possible at present. The validity of the proposed model as a tool for risk analysis may be tested by simulating a model of a real-world business operation and running the simulation with real-time data (observed or simulated).

Availability of abundant data from various supply chain nodes (supplier, distributor, logistics provider) will reduce risk, if the data is analysed and its impact sufficiently understood to deploy risk mitigation steps, at the right time. Operational transparency at or within supply chain nodes is likely to improve with the increase in object associated data acquisition that may be possible through pervasive adoption of automatic identification (RFID, UWB, GPS, sensors, GE VeriWise System). The use and analysis of this data in a model that captures the end-to-end business network (as well as links to other factors that may impact the function of a specific node) may help to reduce risk. It is in this context that a combination of MGARCH and VAR techniques may offer value hitherto unimaginable.

This model is also relevant to businesses increasingly using ‘lean’ principles and depends on global outsourcing practices which may compromise the visibility of the supply chain. Transparency of operations within the corporation (internal risk drivers) is as critical as data from business partners in ‘lean’ and ‘global’ operations to evaluate external risk drivers. In some cases, outlier events may be even more influential given that uncertainty is far greater than risk and it is very difficult to assign proper weights to distant elephants.

Use of GARCH in supply chain to estimate risk through VaR (value at risk) analysis may also help create a merger of financial and physical supply chains. The financial supply chain, which drives financial settlement, takes over where the physical supply chain ends. Exporters want rapid payment while importers demand accurate data on goods received to better manage inventory and cash-flow to optimize working capital management. Thus, capital efficiency (the traditional domain of the chief financial officer or CFO) depends on data and sharing of information (traditional domain of the chief information or chief technology officer, CIO or CTO) about cross-border movement of goods (customs and excise), transfer of title, risk mitigation and payment. Facilitation of the flow of (decisionable, actionable) information across physical and financial supply chains has a direct impact on working capital.

From a risk management perspective, the supply chain, therefore, appears to evolve as a component of the CFO’s responsibility. Adapting the GARCH model to serve as a tool in supply chain risk analysis may offer financial managers a familiar tool that may yield clues to effective supply chain risk mitigation strategies. In general, comprehensive solutions are necessary over the life of a transaction cycle that may integrate cash management, trade settlement, finance, logistics, supply nodes, procurement, demand projections, inventory, human resources, regulatory compliance and management of information across physical and financial supply chain boundaries. Creating one or more models that may work in synergy and integrating such real-world scenarios is a challenge.
However challenging, risk management may soon become a ‘household’ issue for business and industry. Cost of doing business with and in the US may soon have to figure in the cost necessary to implement transparency in order to mitigate risk. Businesses must share data with US Department of Homeland Security if their goods originate overseas. This model of data sharing may soon be adopted by other countries, determined to counter terrorism. The move toward global supply chain transparency is not a matter of if but a question of when, due to the great uncertainty posed by terrorists that heighten security risks. The lack of analytical tools to make sense of this data may create many more problems before it starts providing solutions. If even a tiny fraction of the 25,000 containers that arrive in US ports each day require inspection, then businesses will face costly delays in receiving customs clearance. In October 2002, a war game that mimicked this delay found that closing US ports for only 12 days created a 60-day container backlog and cost the economy roughly $58 billion (Worthen, 2006).

The proven success of GARCH in finance and the potential to adapt GARCH for business operations may be viewed as one of the promising solutions to offer a synergistic multi-faceted tool for risk-adjusted SCM by acting as a bridge for some of the interdependent issues in global business: finance, supply chain and security risk analysis (Figure 2).

6. Concluding remarks
In this paper, we propose a model for forecasting with potential broad spectrum applications that include SCM. The model is based on advances in time series econometrics. GARCH technique is used to explicitly model the volatility generally associated with supply chains. A VAR framework captures the dynamics of interactions that characterize multi-stage SCM. From a theoretical point of view, such a model is expected to yield an accurate forecast, thereby reducing some of the operational inefficiencies. In addition, businesses and security organizations may benefit from GARCH because it may
enable the quantification of VaR associated with a wide variety of processes that require better tools for management of risk.

The proposed model, by its very construction, requires high volume data to estimate a large number of coefficients. Availability of high volume data may not be the limiting factor in view of the renewed interest in AIT that may facilitate acquisition of real-time data from products or objects affixed with RFID tags. Although speculative, it stands to reason that use of a GARCH type model may enhance the ROI from AIT infrastructure by delivering value from acquired data. However, understanding the ‘meaning’ of the information from data is an area still steeped in quagmire but may soon begin to experience some clarity if the operational processes take advantage of the increasing diffusion of the semantic web and organic growth of ontological frameworks to support intelligent decision systems coupled to agent networks (Datta, 2006).

Rigorous validation of the proposed model with real-world data is the next step. In one isolated experiment, the model proposed in this paper was tested to compare forecasting accuracy. When simulated using real-world data and compared to traditional CLRM type techniques, the GARCH type model provided a forecast that was appreciably closer to the observed or realized value (Don Graham, personal communication). This observation is immature. Several more experiments with rigorous controls must be performed before this result may be even considered to offer ‘preliminary’ evidence that the GARCH type model proposed in this paper may represent an advanced tool.

In this paper, we have attempted to coalesce a few ideas toward a ‘solutions’ approach aimed to model volatility in supply chain and in the process, perhaps, better manage risk. It is possible that business, industry, governments, consultants and academics with deep knowledge in one or more fields, may spend the next few decades striving to synthesize one or more models of effective modus operandi to combine these ideas with other emerging concepts, tools, technologies and standards to collectively better understand, analyse and respond to uncertainty. However, the inclination to reject deep-rooted ideas based on inconclusive results from pilot projects is a detrimental trend and begs to ask the question whether one can aspire to build an elephant using a mouse as a model.

Acknowledgements—Part of this work was made possible by the financial support from the sponsors of the MIT Forum for Supply Chain Innovation. The corresponding author (S Datta) is the Co-Founder and Research Director of the MIT Forum for Supply Chain Innovation. The corresponding author benefited from numerous discussions with industry and government thought leaders, worldwide. Special thanks are due to Robert Engle, New York University.

References


Received August 2005; accepted September 2006 after two revisions.
ADVANCES IN SUPPLY CHAIN MANAGEMENT: POTENTIAL TO IMPROVE FORECASTING ACCURACY

Shoumen Palit Austin Datta¹ and Clive W. J. Granger²

¹Research Scientist, Engineering Systems Division
Department of Civil and Environmental Engineering
Research Director & Co-Founder, MIT Forum for Supply Chain Innovation
School of Engineering
Massachusetts Institute of Technology
shoumen@mit.edu

²Research Professor
Department of Economics
University of California
cgranger@ucsd.edu

July 2006
Advances in Supply Chain Management and Decision Support Systems:

Potential to Improve Forecasting Accuracy

by

Shoumen Palit Austin Datta †
Research Scientist, Engineering Systems Division
Department of Civil and Environmental Engineering
Research Director & Co-Founder, MIT Forum for Supply Chain Innovation
School of Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
shoumen@mit.edu

and

Clive W. J. Granger *
Research Professor
Department of Economics
University of California
9500 Gilman Drive
La Jolla, California 92093
cgranger@ucsd.edu

† Corresponding author

* Co-author shared the 2003 Nobel Prize in Economics

Publication version may include additional authors.
Abstract

Forecasting is a necessity almost in any operation. However, the tools of forecasting are still primitive in view of the great strides made by research and the increasing abundance of data made possible by automatic identification technologies, such as, radio frequency identification (RFID). The relationship of various parameters that may change and impact decisions are so abundant that any credible attempt to drive meaningful associations are in demand to deliver the value from acquired data. This paper proposes some modifications to adapt an advanced forecasting technique (GARCH) with the aim to develop it as a decision support tool applicable to a wide variety of operations including supply chain management. We have made an attempt to coalesce a few different ideas toward a "solutions" approach aimed to model volatility and in the process, perhaps, better manage risk. It is possible that industry, governments, corporations, businesses, security organizations, consulting firms and academics with deep knowledge in one or more fields, may spend the next few decades striving to synthesize one or more models of effective modus operandi to combine these ideas with other emerging concepts, tools, technologies and standards to collectively better understand, analyze and respond to uncertainty. However, the inclination to reject deep rooted ideas based on inconclusive results from pilot projects is a detrimental trend and begs to ask the question whether one can aspire to build an elephant using mouse as a model.

Key Words

1. **Background**

Forecasting is an ancient activity and has become more sophisticated in recent years. For a long time steady steps in a time series data set, such as simple trends or cycles (such as seasonals), were observed and extended into the future. However, now a mixture of time series, econometrics and economic theory models can be employed to produce several forecasts which can then be interpreted jointly or combined in sensible fashions to generate a superior value.

The variable being forecast is a random variable. Originally attention was largely directed towards the mean of this variable; later to the variance, and now to the whole marginal distribution. Pre-testing of the data to find its essential features has become important and that has produced modern techniques such as cointegration.

The horizon over which the forecast is attempted is also important, and longer-run forecasts are now being considered as well as forecasts of “breaks” in the series.

The question of evaluation of forecasts has also been greatly developed. Most forecasts are quite easy to evaluate but others, coming from the Global Models (which attempt to model the global economy), are more difficult. Business, commerce, global organizations and governments may find that these models, if explored, may offer valuable guidance for their forecasting activities or their attempts to improve accuracy of forecasts.

Thus, forecasting is a necessity almost in any operation. However, the tools of forecasting (software) in general business use are still primitive in view of the strides made by research. Hence, promoting advances in forecasting to aid predictive analytics is deemed a worthwhile endeavour and is the purpose of this paper. Such tools may further reduce uncertainty and volatility characteristic of global trade. The relationship of various business parameters that may change and impact decisions are so abundant that any credible attempt to drive meaningful associations are in demand by global businesses. This paper proposes some modifications to adapt an already advanced forecasting technique with the aim to develop it as a decision support tool applicable to a wide variety of operations including supply chain management.

2. **Introduction**

Total quality management (TQM) gained prominence in the 1970’s by claiming to boost quality at a lower cost through proper management and operational design. Lean (manufacturing) was the euphoria in the 1980’s following Toyota’s exemplary adoption of Just In Time (JIT) processes to enable flexible manufacturing and minimize costs by reducing inventory level. However, it soon became clear that manufacturing costs could not be reduced further by
pursuing variations of TQM and JIT simply through classical operations research. Globalization of markets in the
1990’s combined with improvements in ICT (information communication technologies) and short product life cycles
shifted the focus on supply chain management (SCM) that could adapt to demand or reduce costs through
improvement in efficiency.

SCM is a set of approaches to efficiently integrate suppliers, manufacturers, distributors, warehouses and retail
stores so that merchandise is produced and distributed in the right quantities, to the right locations, at the right
time in order to minimize system wide costs while satisfying (customer) service-level requirements (Simchi-Levi et
al, 2003). Viewed from this perspective, similarities exist between SCM practices and a competitive market
economy. A market economy ensures that right mix of goods and services get produced (those that are most
wanted by consumers) in the right way (i.e. least cost) and eventually distributed to the right people (those who are
willing to pay the most). Therein lies the attractiveness of a market based economy, that it gives rise to most
efficient allocation of resources. Likewise, SCM, if appropriately designed and executed, may offer efficient business
solutions, thereby minimizing costs and improving readiness or competitiveness.

Despite rapid advances in SCM and logistics, inefficiencies still persist and are reflected in related costs. During
2000, supply chain related costs in the United States alone exceeded $1 trillion (10% of GDP), which is close to the
2005 GDP of Russia and Canada or the combined GDP of the 22 nations who are members of the oil-rich League of
Arab Nations. A mere 10% savings of supply chain costs in the United States is close to the 2005 GDP of Ireland
(Datta et al, 2004). More than US$3 trillion have been spent on global logistics in 2004 and this represents almost
5% of the global GDP or more than the GDP of Germany and just less than the GDP of India in 2005. Inefficiencies
in the global logistics network estimated at an approximate of US$600 billion (close to the 2005 GDP of Australia)
offers untapped opportunities for organisations to optimise or adapt to improve sustainable profitability. Hence
serious questions have been raised as to how to make decision systems more efficient in order to reduce cost of
transaction (Coase, 1960 and Coase, 1992). This, in turn, requires a thorough understanding of the factors that
make design and operation of effective SCM strategy a challenging task due to the volatility of supply and demand.

It is challenging enough to design and operate a supply chain for one facility, in order that costs are minimized
and service levels are maintained. The difficulty increases exponentially when the system as a whole is considered
and system-wide costs must be minimized, i.e., optimizing the interactions between various intermediaries, such as
retailer, wholesaler, distributor, manufacturer and supplier of materials. This is mathematically equivalent to finding
a global optimal solution as opposed to local optimization, the predominant business practise. Global optimization,
involving several stages in the decision making process, is far more complicated. It is also much broader in scope
and encompasses local optimization, but only as a special case.
Some may argue, justifiably, that optimization itself is to be blamed for the inefficiencies in global SCM practices. Perhaps optimization suggests an innate assumption that operations are capable of reaching a steady state or equilibrium, once “optimal” conditions are determined and executed. Global volatility, even in peaceful or stable political economies, may disprove this assumption. Hence, dynamic or recurrent real-time optimization is required and reflects the fundamental necessity of global supply chains to continuously adapt.

The task of global optimization is rendered difficult by the uncertainty of the business environment. First, businesses need to continually adapt to demand uncertainty and its impact on inventory management. In a market based economy, production decision is primarily demand driven and must be made ex ante (before customer demand is realized). Furthermore, due to lack of information sharing ex post (after actual customer demand is realized) between partners, the variability in orders is amplified upstream, along the supply chain. This phenomenon is commonly referred to as the ‘Bullwhip Effect’ and it is a key driver of inefficiencies associated with SCM. It distorts the demand signals, resulting in costs in the form of excess capacity and inventory, need for increased storage, and transportation cost increases (due to less-than-truckload or LTL scenarios), to name a few (Lee et al, 1997).

The Bullwhip Effect and the resulting inefficiencies associated with traditional supply chains may be reduced, in theory, by centralizing information relating to supply and demand (Datta, S., et al, 2004). In other words, a “centralized” supply chain model is one where such information is made available to all participating businesses at various stages of the supply chain or network of partners. Advances in information and communication technologies in the past decade has made it easier to acquire, share, access and analyze data in a manner that is increasingly feasible for “sense and response” systems. In the context of SCM, the idea is to enable intermediaries in the supply chain process to act as “infomediaries” or serve as agents for sharing and accessing the real-time data flow through common interfaces, such as web-based services (Datta, 2006).

Acquisition of or access to data is not equivalent to use of decisionable information that can be extracted from data. Differences in forecasting methodologies applied (to the same data) at different stages of SCM by the participants (process owners) may yield varying types of information that may further obscure the value of data or give rise to increased fluctuations, thus distorting the signals, such as demand (Lee, 1997). To rein in the Bullwhip Effect, one contribution may stem from a standardized forecasting model, that may be used by the supply chain partners, as an analytical tool to extract value from the data that is accessible to all the partners. Sharing of such an analytical engine by a group of businesses is possible through the use of grid computing (Datta, 2004). Although we have the tools and technologies at our disposable, the sluggish growth of collaborative systems such as CPFR (collaborative planning, forecasting and replenishment) may be indicative of only a mild interest in the benefits of
collaborative information processing. It is quite possible that lack of trust between businesses and heightened data
security risks may be slo\-\-\-wing real-world implementations of valuable strategies such as CPFR.

While a standardized forecasting model applied to near real-time data in a shared or centralized database may
try to tame the Bullwhip Effect, it may never be eliminated due to outlier events and inherent or unexplained
variability. In SCM, innumerable sources of variability exist, including factors such as demand forecasting, variability
in lead time, batch processing or bulk ordering to take advantage of transportation discounts, price variability due to
product promotion or discount, to name a few.

Hence, the objective of this paper is to propose the potential use of an advanced statistical modeling technique
for the purpose of forecasting (originally proposed by Datta, 2003 and Datta et al, 2004). Based on the pioneering
work on time series econometrics by Clive W. J. Granger and Robert F. Engle (Engle and Granger, 1987 & Engle and
Granger, 1991), this paper proposes a few modifications to the statistical model proposed by Robert F. Engle based
on advances in time series econometrics. The modifications were introduced to make the model more amenable for
use in decision support systems, such as, supply chain management. If the proposed modifications are indeed
viable, it is expected to explicitly model the interactions between various intermediaries of SCM as well as the time
varying (non-constant) variability that manifests, at least in part, as the Bullwhip Effect. For example, our proposed
model captures the cross-variable dynamics as reflected in interaction between supply chain nodes or stages
(retailers, wholesalers, distributors) using vector auto regression (VAR) methodology which is essentially a model
for the means of a vector process. The framework also explicitly models the time varying volatility (perhaps
observed, in part, as the Bullwhip effect) by using a Generalized Autoregressive Conditional Heteroskedasticity
(GARCH) technique (Engle and Kroner, 1995). GARCH is a model for volatility of a single series whereas multi-
ivariate GARCH (MGARCH) is a model for volatility (variances and covariances) for a vector. Therefore, the proposed
model may also be denoted as a VAR-MGARCH model.

While these techniques have been widely used in finance (and economics) in the past few decades (and also
earned Engle and Granger the 2003 Nobel Prize in Economics), to our knowledge, they have not been applied or
explored as decision support tools by supply chain planners or analysts in the area of supply chain management.

By using a dynamic model of volatility (defined as standard deviation of variance), a GARCH type model has the
added advantage of providing a forecast of volatility in near term. Such a forecast may be useful in calculating value
at risk (VaR). However, VaR is estimated with a simplifying assumption, such as (joint) normality. Consequently,
ARCH technique has become an indispensable tool in risk assessment and management in the financial sector.
Globalization of the supply chain has concomitantly increased the risk in the supply chain due to potential for loss of
profits from over-capacity (cost of excess inventory) or opportunity lost due to out-of-stock (OOS) situations. Hence, it is our contention that use of similar methodology in supply chain processes may enable businesses to better manage or even quantify the risk in the process.

VAR-GARCH type models require estimating a large number of parameters and hence cannot be used in practice unless a large sample of data is available. The lack of availability of high volume granular data may explain the scarce interest in applying this modeling technique as a forecasting tool in decision support systems. High volume accurate data is the single most important driver for forecasting accuracy. The recent rejuvenation of the use and adoption of automatic identification technologies may partly ameliorate the lack of high volume data. The surge in the use of radio frequency identification (RFID) or ultrawideband (UWB) tags that may be embedded or attached with physical objects, will make it possible to track and locate objects along the entire supply chain, if the systems used by manufacturers, distributors, logistics providers and retailers are able to take advantage of middleware and hardware interoperability (software defined radio or SDR) to monitor, access and share near real-time RFID data.

Thus pervasive use of automatic identification tags may provide high volume object data in near real time with the maturing trend toward ubiquitous computing. Businesses may not have a clear understanding of how to use this data efficiently to extract decisionable and actionable information that offers business value not merely through savings but may actually increase profitability. We propose that businesses explore advanced techniques such as multivariate GARCH that requires high volume of data for estimation but offers the potential to deliver increasingly accurate forecasts along with a measure of Value at Risk. Success in applying similar models to analyze financial market returns is well documented. Global supply chain management and any operation in need of planning for the future (healthcare, military, emergency) offers interesting applications for VAR-MGARCH techniques.

In the next section, we develop this statistical model for forecasting on a sequential, step by step basis, with the idea that independent variables represent operational 'nodes' (for example, the 'stages' or entities in supply chain). Section 4, discusses data requirements for model validation including the significance of automatic identification (RFID) data. Risk in the global supply chain is qualitatively discussed in section 5. Concluding thoughts are offered in section 6 including comments about preliminary results obtained in (only) one study that explored the use of the modification proposed in this paper to simulate improvements in forecasting based on (only) one data source from an ongoing real-world operation.

3. Statistical Model: VAR-MGARCH

Forecasting demand is a key tool in managing uncertainty. Forecast accuracy may depend on the understanding and coverage of parameters taken into account and the accuracy of historic data available for each variable that
may have an impact on the prediction. In this section, we propose a statistical model that combines classical regression analysis with advanced time series techniques, hopefully to improve accuracy of forecasts.

3.1 CLRM

Classic linear regression models (CLRM) have been around for a century (Studenmund, 2000) and used for a variety of purposes including traditional supply chain planning software. CLRM may be expressed as follows:

\[ y_t = \beta_0 + \beta_1 x_t + \epsilon_t \]  \hspace{1cm} (1)

where,

- \( y \): dependent variable of interest to be modeled for forecast (e.g., sales of a product, say aspirin)
- \( t \): time period (frequency of observation, for example, t-1 may indicate prior week 1, t-2 \( \rightarrow \) week 2)
- \( \beta \): coefficients to be estimated (based on values of \( y \) and \( x \))
- \( x \): explanatory variable that is used to ‘explain’ variations in the dependent variable \( y \) (for example, low sales of aspirin may be explained by low in-store inventory \( \{x\} \) of aspirin)
- \( \epsilon \): random (stochastic) error term

This simple technique can model multiple independent or explanatory variables, that is, multiple \( x \)’s, since the variation in \( y \), say, sales of aspirin, is dependent on multiple parameters, such as inventory \( \{x_1\} \), price \( \{x_2\} \), expiration date \( \{x_3\} \). The choice of \( x \)’s (number of explanatory variables) will drive the validity and accuracy of the model. \( x \)’s may be based on underlying economic principles (theoretical) or business logic (practical underpinnings). However, no matter how many \( x \)’s are included, there may be an inherent randomness in \( y \) that cannot be explained. Thus, the random error term \( \{\epsilon\} \) is included in the equation (admission of the fact that the dependent variable \( y \) cannot be modeled perfectly). The corresponding mathematical equation is given by equation (2):

\[ y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + \epsilon_t \]  \hspace{1cm} (2)

Objective of CLRM is to estimate the parameters \( \{\beta_0, \beta_1, \ldots, \beta_k\} \) of the model based on a sample of observations on \( y \) and \( x \), assuming that \( \epsilon \) is characterized by a normal distribution with mean \( = 0 \) and variance \( = \sigma^2 \) for all time periods \( t \).\(^1\)

\(^1\) Normality assumption is needed for hypothesis testing with respect to \( \beta \)'s based on sample of data, unless the sample size is large.
\( \varepsilon_t \sim N(0, \sigma^2) \)

Given the multiple sets of \( (\beta_0, \beta_1, \ldots, \beta_k) \) may be estimated, the objective of CLRM is to choose that set of \( (\beta_0, \beta_1, \ldots, \beta_k) \) which minimizes the sum of squared residuals \( (\varepsilon_t^2, \varepsilon_t^2, \ldots, \varepsilon_n^2) \):

\[
\sum_{t=1}^{n} \varepsilon_t^2
\]

where \( \varepsilon_t \) = empirical counterpart of \( \varepsilon \) (and is estimated based on sample data). Intuitively, this amounts to finding a line that best fits the data points by minimizing the sum of squared vertical distances of the actual data points from the fitted line. Thus residuals are essentially in-sample forecast errors as they measure the difference between actual \( y \) and fitted \( y \). This technique is commonly referred to as the principle of ordinary least squares (OLS) and widely used due to its simplicity. The attractiveness of CLRM based OLS forecasting stems from the fact that we can model cross variable linkages. This feature is especially useful to carry out ’what-if’ analysis. For example, what may happen to sales \( (y, \) the dependent variable) of aspirin-based pain killers in retail sales if the in-store inventory of non-aspirin pain killers were increased by 10%? Clearly, ’what if’ analysis is conditional upon assumptions we make about \( x \)'s in the model. Therefore, in building this model, the choice of \( x \) is a process decision based on the model builder’s knowledge about an operation or business or industry.

One may wonder if we are playing a ’what if’ game or is 10% increase, cited above, a real-world scenario. The retail outlet surely knows what has happened in the past. This segues to the next phase in the development of our statistical model where it is no longer necessary to assume values of the explanatory variable to forecast \( y \) (the dependent variable). Instead of inserting arbitrary values for future \( x \)'s (such as a 10% increase), we start by forecasting the values of \( x \) based on its own past data to obtain an unconditional forecast for \( y \). In this stage of model development, the regression technique gets intertwined with time series techniques. By fitting a univariate autoregressive model to \( x \) where we use past (lagged) values of \( x \) to forecast \( x \), we obtain the following equations (for \( x_{1t}, \ldots, x_{kt} \)):

\[
X_{1t} = \alpha_{10} + \alpha_{11} x_{1t-1} + \alpha_{12} x_{1t-2} + \ldots + \alpha_{1N} x_{1t-N} + u_x
\]

\[
X_{kt} = \alpha_{k0} + \alpha_{k1} x_{kt-1} + \alpha_{k2} x_{kt-2} + \ldots + \alpha_{kN} x_{kt-N} + u_x
\] (3)
Rewriting using general notation:

\[ y_t = \beta_o + \sum_{i=1}^{N_{xt}} \alpha_{1i} x_{1t-i} + \ldots + \sum_{i=1}^{N_{kt}} \alpha_{ki} x_{kt-i} + \epsilon_t \]  

(4)

or

\[ y_t = \beta_o + \sum_{k=1}^{K} \sum_{i=1}^{N_{kt}} \alpha_{ki} x_{kt-i} + \epsilon_t \]  

(4a)

where,

- \( x_{1t} \) = variable \( x_1 \) at time \( t \) (for example, we used \( x_1 \) for inventory, thus \( x_{1t} \) is inventory at time \( t \))
- \( x_{kt} \) = variable \( x_k \) at time \( t \) (up to \( K \) number of \( x \)'s)
- \( x_{1t-1} \) = value of \( x_1 \) at time \( t-1 \) (referred to as the lagged value by one period)
- \( N \) = period up to which the lagged values of \( x_{1t} \) will be used in the equation
- \( U \) = random error term

Note that \( \beta_0 \) in equation (4) is a combination of constants from equation (2) and (3) respectively.

In equation 3, \( \alpha_{11} \), \( \alpha_{12} \) are coefficients of \( x_{1t-1}, x_{1t-2} \) and are referred to as lagged weights. An important distinction is that instead of arbitrarily assigning weights, these coefficients are estimated using OLS. The error term in equation 3 represented by \( u \) is analogous to \( \epsilon \) in equation 1. Depending on the number of \( x \)'s \( (x_1, \ldots, x_K) \) that adequately represents the process being modeled in equation 1, there will be \( K \) number of equations as given by (3) that must be estimated to forecast the \( x \)'s \( (x_1, \ldots, x_K) \) which will then be used to obtain an unconditional forecast of \( y \). Thus, to simplify the task, we can estimate all the parameters \( (\alpha, \beta) \) simultaneously by re-writing equation 1, the basic CLRM equation, as equation (4) or its shortened version, as in equation (4a).

Equation 4 is a step toward forecasting the dependent variable \( (y) \) with greater accuracy using forecasts of \( x \)'s based on historical data of \( x \)'s (lagged values). But, it is also clear that equation 4 ignores the impact on \( y \) of the
past values of y itself (lagged values). Consequently, a preferable model will include not only lagged values of x but also lagged values of y, as shown in equation (5).

\[
y_t = \beta_0 + \sum_{j=1}^{N_y} \varphi_j y_{t-j} + \sum_{K=1}^{K} \sum_{i=1}^{N_x} \alpha_{ki} x_{kt-i} + \varepsilon_t
\]  

In moving from conditional to unconditional forecasts of y using a time series model, we are increasing the number of parameters to be estimated. In equation 2, we estimate K parameters (\(\beta_1, \ldots, \beta_K\)) excluding (\(\beta_0\)). In equation 3, we estimate n parameters (\(\alpha_1, \ldots, \alpha_n\)) excluding the intercept (\(\alpha_0\)) for each of the K number of x’s (\(x_1, \ldots, x_K\)). In equation 5 we estimate j parameters for lagged values of \(y_{t-j}\) (\(\varphi_1, \ldots, \varphi_j\)) in addition to all the parameters for equation 4. If we set K=5 (5 explanatory variables, the x’s), \(N_x = 10\) (number of lagged values to forecast the x’s) and \(N_y = 10\) (number of lagged values of \(y_t\)), then, we have increased the number of parameters to be estimated from 5 in equation (2) to 50 in equation (4) to 60 in equation 5 (excluding intercept).²

3.2 GARCH

Thus far, our discussions have centered on CLRM in conjunction with time series techniques. CLRM is based on a set of assumptions mainly about \(\varepsilon\), that, when satisfied, gives rise to desirable properties of the OLS estimates. Needless to emphasize, in the real world, these assumptions are almost always violated. Developments in time series, over the past couple of decades, have addressed the challenges that stem from the violation of some of these classical assumptions leading to inaccurate forecasts.

One of the assumptions often violated in practice relates to homoskedasticity (homo = equal, skedasticity = variance or mean squared deviation (\(\sigma^2\), a measure of volatility) or constant variance for different observations of the error term. Forecast errors are frequently found to be heteroskedastic (unequal or non-constant variance). For example, in multi-stage supply chains, the error associated with manufacturer’s forecast of sales of finished goods may have a much larger variance than the error associated with retailer’s projections (the assumption being that

² To drive precision to the next (logical) step, equation 5 may be expanded further to include the important real-world observations regarding trend, seasonality and other cyclical dynamics. Businesses struggle to uncover ‘trends’ and once found, they are avidly pursued, often for short term gains but increasingly with less than stellar results due to fickle customer preferences.
the proximity of the retailer to the end consumer makes the retailer offer a better or more informed forecast of future sales through improved understanding of end-consumer preferences). The upstream variability reflected in the Bullwhip Effect violates the basic premise of CLRM, the assumption of homoskedasticity. CLRM ignores the real-world heteroskedastic behavior of the error term $\varepsilon_t$ and generates forecasts which may provide a false sense of precision by underestimating the volatility of forecast error.

Figure 1: Homoskedasticity, Heteroskedasticity and the Bullwhip Effect

Homoskedastic and heteroskedastic error term distributions are illustrated in Figure 1. In a homoskedastic distribution, all the observations of the error term can be thought of as being drawn from the same distribution with mean = 0 and variance = $\sigma^2$ for all time periods (t). A distribution is described as heteroskedastic when the observations of the error term may be thought of as coming from different distributions with differing widths (measure of variance). In supply chains, the variance of orders is usually larger than that of sales. This distortion tends to increase as one move upstream from retailer to manufacturer to supplier. Therefore, the assumption of heteroskedasticity seems more appropriate as a characteristic that may be associated with the Bullwhip Effect.

While variance of error term may change across cross sectional units at any point in time, it may also change over time. This notion of time varying volatility is frequently observed in financial markets and has been the driving force behind recent advancements in time series techniques. Robert Engle is credited with the observation that not only is volatility non-constant (of financial asset returns), it also tends to appear in bursts or clusters. Instead of considering heteroskedasticity as a problem to be corrected (approach taken by CLRM practitioners in assuming homoskedasticity of error term), Robert Engle seized this opportunity to model this non-constant time dependent variance (heteroskedasticity) using an autoregressive moving average (ARMA) technique.
ARMA has been in use for several decades and is a combination of AR (autoregression) and MA (moving average) techniques. We have already invoked autoregressive (AR) representation in equations (4) and (5). AR links the present observation of a variable to its past history, for example:

\[ y_t \text{ to } y_{t-1}, \ y_{t-2}, \ldots, \ y_{t-p} \]

where \( p \) = the order of the autoregressive process AR(\( p \)) or the period up to which the historical data will be used (a determination made by using other statistical tools).

Thus, AR is a technique by which a variable can be regressed on its own lagged values. For example, today’s sales \( (y_t) \) may depend on sales from yesterday \( (y_{t-1}) \) and the day before \( (y_{t-2}) \). AR(\( p \)) is appealing since it links the present to the past. MA expresses observations of a variable in terms of current and lagged values of squared random error terms \( \epsilon_t, \ \epsilon_{t-1}, \ \epsilon_{t-2}, \ldots, \epsilon_{t-q} \) where \( q \) is the order of the moving average process MA(\( q \)). Combining AR(\( p \)) and MA(\( q \)) we get ARMA(\( p,q \)) where \( p \) and \( q \) represent the lagging order of AR and MA, respectively.

Robert Engle used the MA technique to model the time varying volatility in a series and proposed the so called AutoRegressive Conditional Heteroskedasticity model or ARCH. The ‘conditional’ nature of non-constant variance (heteroskedasticity) refers to forecasting of variance conditional upon the information set available up to a time period \( t \). Using ARCH, the variance of the random error term \( (\epsilon_t) \) in equation (5) can be expanded in terms of current and lagged values of squared \( \epsilon \) \( (\epsilon^2_{t-1}, \ \epsilon^2_{t-2}, \ldots, \epsilon^2_{t-q}) \) as follows:

\[
\sigma^2_t = \theta_0 + \theta_1\epsilon^2_{t-1} + \theta_2\epsilon^2_{t-2} + \ldots + \theta_q\epsilon^2_{t-q} \tag{6}
\]

where \( \sigma^2_t \) = variance of \( \epsilon_t \) \( [\text{var}(\epsilon_t)] \).

This MA(\( q \)) representation of \( \sigma^2_t \) was later generalized to an ARMA representation of \( \sigma^2_t \) and is referred to as the Generalized AutoRegressive Conditional Heteroskedasticity model or GARCH.\(^3\)

\(^3\) The GARCH technique represents a parsimonious model than ARCH, while allowing for an infinite number of past error terms to influence current conditional variance. Hence GARCH is widely used than ARCH.
GARCH evolved when Tim Bollerslev extended the MA(q) representation of \( \sigma_t^2 \) (the ARCH model) to include an AR(p) process, that is, regressing a variable \( (\sigma_t^2) \) on its own (past) lagged values \( (\sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2) \) as well. Thus, variance of the random error term in a certain period \( (\varepsilon_t) \) can be modeled to depend not only on squared past errors \( (\varepsilon_{t-1}^2, \ldots, \varepsilon_{t-q}^2) \) but also on the lagged value of the variance \( (\sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2) \) as shown in equation (7) below.

\[
y_t = \beta_o + \sum_{j=1}^{N_{yt}} \phi_j y_{t-j} + \sum_{k=1}^{N_{xt}} \sum_{i=1}^{K} \alpha_{ki} X_{kt-i} + \varepsilon_t
\]

\[
\sigma_t^2 = \theta_o + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \tau_j \sigma_{t-j}^2
\]  

(7)

Thus, GARCH may enable supply chain practitioners to model the volatility in the supply chain, a phenomenon documented by the Bullwhip Effect. How GARCH may help calculate the value at risk for various supply chain stages deserves deeper investigation. Future research may reveal a mechanism to quantitatively determine the risk associated with various supply chains. The latter tool, when developed, may be of considerable value in general risk management in the globalized world of international commerce.

3.3 VAR-GARCH

In developing the GARCH model, equation (7) takes into account the lagged values of the dependent variable (sales), the impact of multiple explanatory variables (K number of x’s that influence sales such as inventory, price) and their respective lagged values, as well as time dependent heteroskedasticity of the error term. But, thus far, we have not considered the fact that to predict sales h periods ahead, it is also crucial to model the interaction between the entity level nodes (manufacturer, supplier, distributor in supply chain) which can impact sales.

In any operation, including supply chains, interaction between partners can influence any outcome (profit, service, readiness, response). The strikingly different business “clockspeed” and dynamics of the supply chains partners is what partly fuels the Bullwhip Effect. Thus, to incorporate the dynamics of interaction between players, it
is essential to explicitly model the dynamics between the entities to be a useful real-world model. A combination of vector autoregression technique with GARCH captures this dynamics. Vector autoregression (VAR) is a model for the means of a vector process and was developed over a quarter century ago by Sims (1980). Previously we discussed AR(p) with respect to equation (5), which is a univariate model. In contrast, VAR(p) is a n-variate (multivariate) model where we estimate n different equations (for \( Y_1, Y_2, Y_3 \ldots \ldots Y_n \)). In each equation, we regress a variable on p lags of itself as well as p lags of every other variable in the system. Thus, the right hand side variables are the same for every equation in the system.

The key advantage of VAR lies in its ability to capture cross-variable dynamics (vector process). For example, future sales (prediction) of Michelin brand tyres may not be precisely forecasted by Sears unless the store takes into consideration the events or sales (vector) at the distributor. Thus, there are at least two parties (vectors) in this example (interaction between retail store and distributor). To model this cross variable dynamics of \( n=2 \) using VAR(p), let us assume that \( p=1 \) (lagged by 1 period). Equation (7) may be extended to the VAR-GARCH type model for two entities with \( (n=2, p=1, q=1) \) as shown in equation (8).

\[
\begin{align*}
\mathbf{y}_{1t} &= \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{xkt}} \alpha_{ki} \mathbf{x}_{kt-i} + \varphi_{11} y_{1t-1} + \varphi_{12} y_{2t-1} + \varepsilon_{1t} \\
\mathbf{y}_{2t} &= \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{xkt}} \alpha_{ki} \mathbf{x}_{kt-i} + \varphi_{21} y_{1t-1} + \varphi_{22} y_{2t-1} + \varepsilon_{2t} \\
\sigma_{1t}^2 &= \mathbf{c}_{11} + \theta_{11} \varepsilon_{1t-1}^2 + \tau_{11} \sigma_{1t-1}^2 \\
\sigma_{2t}^2 &= \mathbf{c}_{22} + \theta_{22} \varepsilon_{2t-1}^2 + \tau_{22} \sigma_{2t-1}^2
\end{align*}
\] 

(8)
In the VAR-GARCH model represented by equation (8), this dynamics is captured by estimating the coefficient \( \Phi_{ij} \) which refers to changes in \( y_i \) with respect to \( y_j \). For example, if \( y_1 \) represents Michelin tyre sales at Sears retail store and \( y_2 \) represents Michelin tyre sales at the distributor, then the parameter \( \Phi_{12} \) refers to changes in sales at retail store \( (y_1) \) with respect to sales at the distributor \( (y_2) \). If any one of the two random error terms \( (\varepsilon_{1t} \text{ and } \varepsilon_{2t}) \) changes, it will impact both the dependent variables \( (y_1 \text{ and } y_2) \). In terms of equation (8) above, if \( \varepsilon_{1t} \) changes, it will change \( y_{1t} \) and since \( y_{1t} \) also appears as one of the explanatory variables for \( y_{2t} \) in the equation, the change in any error term impacts both dependent variables in this VAR representation. This cross variable dynamic interaction has thus far been ignored by current modeling practices for forecasting. The VAR component of the proposal in this paper is closer to the real-world scenario and VAR-GARCH may make it possible to quantify such cross-variable dynamics.

3.4 Multivariate GARCH (MGARCH)

To move beyond the realm of univariate autoregression to a vector autoregression system, for further precision of forecast, it is necessary to model time varying conditional covariance (measuring the degree of association between any two variables) between \( \varepsilon_1 \) and \( \varepsilon_2 \) in addition to time varying conditional variance of the error term. In other words, the error terms associated with the retailer’s sales forecast and the distributor’s inventory level may be correlated. This type of multivariate interaction is not explicitly captured by the VAR-GARCH model (section 3.3) yet in the business world the association between, say, sales forecast and inventory level, is crucial for the overall efficiency and profitability of the supply chain. Thus, the next task is to combine the VAR representation with a multivariate GARCH component. Assuming \( p=q=1 \), MGARCH specification can be expressed as follows:

\[
Y_{mt} = \beta_0 + \sum_{k=1}^{k} \sum_{i=1}^{N_{xt}} \alpha_{ki} X_{kt-1} + \sum_{l=1}^{2} \phi_{ml} Y_{lt-1} + \varepsilon_{mt}, \quad \forall m=1,2
\]

\[
\sigma^2_{1t} = C_{11} + \theta_{11} \varepsilon^2_{1t-1} + \tau_{11} \sigma^2_{1t-1}
\]

\[
\sigma^2_{12t} = C_{12} + \theta_{12} \varepsilon^2_{1t-1} \varepsilon_{2t-1} + \tau_{12} \sigma^2_{2t-1}
\]

\[
\sigma^2_{2t} = C_{22} + \theta_{22} \varepsilon^2_{2t-1} + \tau_{22} \sigma^2_{2t-1}
\]
where $\sigma_{12,t}$ indicates conditional covariance between $\varepsilon_1$ and $\varepsilon_2$ in time period $t$, based on information set available up to period $(t-1)$.

Thus, the conditional variances and conditional covariances will depend on their respective lagged values, as well as the lagged squared errors and the error cross products. Clearly, estimating such a model may be a formidable task, even in a bi-nodal scenario, for example, a retailer and distributor.$^4$

3.5 Is there a link between Bullwhip Effect and GARCH Processes?

We have often used "volatility" to indicate the observation of fluctuation represented by the Bullwhip effect but it is unclear if there is an actual link between Bullwhip Effect and GARCH processes. Simply going "along" the supply chain, there may not be an use for GARCH but going over "time" there might be, as explained below.

Consider a supply chain with a sequence of stages or locations: $L_0$ (origin), $L_1$ (stage one or first location), $L_2$ (second), ..., $L_t$ (final stage or end). Goods moving along the chain are associated with a number of delivery times between these locations. Let us denote $T_{j \rightarrow k} (t)$ as the time taken to deliver a good from location $j$ to location $k$, the goods having started at time $t$ at $L_0$ (origin). The $1^{st}$ leg of the chain takes time $T_{0, 1} (t)$, the $2^{nd}$ leg $T_{1, 2} (t)$ and so forth. These $L$ values are positive random variables, possibly auto correlated, but initially considered as an independent sequence. Note that,

$$T_{0,k} (t) = \sum_{j = 0}^{k} T_{j, j+1} (t)$$

is, essentially, a random walk, with an increasing mean and variance. If all the $T_{j, j+1} (t)$ are uncorrelated with mean $m > 0$ and variance $\nu$, then $T_{0,k}(t)$ will have mean $km$ and variance $k\nu$. As $k$ increases, volatility will increase, which is the Bullwhip Effect (there is no need to use GARCH models to fit this process).

$^4$ If the GARCH system is functional, it may be used to better analyze Impulse Response Function (IRF). At present, IRFs values are of limited use because it is difficult to provide confidence intervals for the values. Confidence intervals are necessary for forecasts. GARCH values may provide these confidence intervals. IRF traces the impact of changes ("shock") in error terms on the dependent variable for several periods in the future. Applied to operational planning, IRF may offer insight about 'sense and respond' scenarios. IRF simulation may enable exploration of multi-component "what if" scenarios by creating challenges and learning (from simulation) how to prepare (readiness) for such challenges (hurricane, fire, flood, earthquake, epidemics, pandemics, military escalations).
The total time taken for the supply chain $T_0, r(t)$ will, as $t$ varies, generate a time series which can be analyzed. In the unlikely event that the chain does not change, this will be a stationary series, but it is likely that volatility (Bullwhip Effect) will be experienced by the chain. Thus, an AR-GARCH model may be appropriate.

4. Data

The modeling technique proposed above, may represent an opportunity to apply advanced statistical and econometric tools to improve the quality of predictive analytics in general and supply chain forecasting, in particular. However, validating such a model requires high volume data and involves estimating a large number of parameters. It is possible that advanced organizations, such as the military establishments, may have considered using these techniques but could not substantiate the models due to fewer than necessary reliable data points (degrees of freedom).

However, data “points” may no longer be the limiting factor if the increasing interest in adoption of Automatic Identification Technology (AIT) is transformed to reality. Widespread adoption of AIT (such as radio frequency identification or ultrawideband tags and sensor data) may pave the way for use of real time data to validate a model such as the one proposed in this paper. The innovative convergence of fields as diverse as AIT and time series econometrics may improve decision support systems in domains beyond finance and economics (Datta, 2004).

AIT and progress toward embedding intelligence into physical objects may allow them to communicate with each other (thing-to-thing) as well as with business systems or users (consumers) in near real-time. Hence, businesses may soon be faced with ultra high volume multi-gigabit data streams that may be expressed succinctly only in terms of exabytes per second (1 exabyte = $10^{18}$ bytes or $10^9$ gigabytes). Infrastructure necessary to acquire such data may not offer a satisfactory return on investment (ROI) unless decisionable information derived from this data offers value or profitability. The question of value from high volume data may be considerably enhanced by using data in advanced statistical models (as proposed above) to yield useful analytics.

Availability of near real-time data at the right time may be especially useful for industries where historical data is an agonizing cliché due to short product life cycles, such as mobile phones, digital cameras and laptop computers, which are characteristic of high ‘clockspeed’ industries (Fine, 1998). For a product with a sales life cycle of 200 days (about 6 months), if we use data from the past 100 days (more than 3 months) in the time series model, it may be difficult to ‘change course’ and respond or adapt (based on forecasts or predictions from such models). This is where the granularity of high volume AIT data from RFID tags offers the potential to deliver real business value and ROI.
Re-consider the above example but assume the availability of high volume accurate AIT data (from RFID tags on high value products with rapid obsolescence). The data from RFID tags may be modeled with N=100 where data is lagged every hour (N=100 hours instead of N=100 days). However, whether the quality of the information that may be extracted from such data, may change if N=100 is in hours or days, is a business question, not a technology or analytics issue. Consequently, whether high volume data of a certain granularity is sufficient for reliable forecasts will depend on process. If the hourly data is used (N=100), then predictive analysis can be made available within 5 days from launch of a product with 195 days (97.5%) of its sales life cycle still viable, in case it is necessary to re-engineer the product in order to respond to or meet customer preferences. If compared to daily batch data with N=100, analytics may be available after 100 days or with only 50% of the product sales life cycle still viable.

Thus, use of high volume real-time data in these models may make it possible and feasible for sales, marketing, production or distribution to adapt in real-time or at the right-time. Changes can be initiated, based on forecasts, earlier in the (sales) cycle of the product or even at the production stage, by using delayed product differentiation strategies, if products were designed with modular architecture or if the product lifecycle was carefully optimized by balancing the demands of development vs fulfillment supply chain parameters.

Regarding estimation technique, the OLS technique, although simple, may not be preferred for use with GARCH. OLS technique proceeds by minimizing sum of squared residuals but residuals, by definition, do not depend on the parameters of the conditional variance equation. Thus, in the presence of GARCH specification, minimizing residual sum of squares is no longer an appropriate objective. Instead, to estimate models from the GARCH family, the Maximum Likelihood Estimation (MLE) is the technique of choice. However, under an assumption of normality, MLE is simply generalized OLS.

MLE works by finding the most likely values of the parameters given the actual data. Multivariate GARCH models are similar to their univariate counterparts and thus MLE technique can be used. However, due to explicit modeling of conditional covariances over time in MGARCH, the number of parameters to be estimated increases exponentially. A few different MGARCH specifications have been proposed, such as the VEC model proposed by Bollerslev, Engle and Woolridge (1988) and the BEKK model proposed by Engle and Kroner (1995). This is an area that warrants deeper exploration keeping in mind increased data availability through use of AIT data acquisition tools.

5. Implications for Risk Management

Risk in supply chain management originates from two key areas: supply and demand. At the next level of equal importance are environmental, political, process and security risks. Political and environmental risks may always
remain amorphous and refractory to adequate quantification. Security risks are even more volatile but on a higher priority level that demands advanced risk management tools and analysis for targeting operations in global trade.

Too often, risk is viewed as simplistic as merely the product of frequency and consequence. A high-frequency but low-consequence event (currency exchange rates) is viewed as similar to a low-frequency but high-consequence event (sinking of a cargo ship laden with spare parts). In reality, such apparently “similar risks” may have vastly different effects. Sensational risks grab attention and beg for resource-consuming mitigation while risk managers tend to ignore the smaller risks that create the real friction in the supply chain. With the increasingly complex business environment that is the hallmark of globalization, today’s supply chain presents a myriad of specific risks ranging from external sources (such as, terrorist strikes or vulnerability to political instability in developing countries due to global outsourcing) to internal sources (pressure to enhance productivity and reduce costs by eliminating waste, removing duplication through use of single source supplier). If accounted as parameters in traditional optimization equations, the sheer number of factors will exponentially increase the state space and as a result may grind the computation of the optimization algorithms to a pace that may become unacceptable for decision support systems to aid in the management of supply chain adaptability.

The VAR-MGARCH model proposed here may be well suited to take into account the details of the operational nodes (assuming we have data available from each of these nodes/processes). Recurring analysis performed in near real-time (assuming real-time data is available to the analytical engine) may offer results that predicts or detects risks in the operational model (supply chain) far in advance of what is possible at present. The validity of the proposed model as a tool for risk analysis may be tested by simulating a model of a real world business operation and running the simulation with real-time data (observed or simulated).

Availability of abundant data from various supply chain nodes (supplier, distributor, logistics provider) will reduce risk, if the data is analyzed and its impact sufficiently understood to deploy risk mitigation steps, at the right time. Operational transparency at or within supply chain nodes is likely to improve with the increase in object associated data acquisition that may be possible through pervasive adoption of automatic identification (RFID, UWB, GPS, sensors, GE VeriWise System). The use and analysis of this data in a model that captures the end-to-end business network (as well as links to other factors that may impact the function of a specific node) may help to reduce risk. It is in this context that a combination of MGARCH and VAR techniques may offer value hitherto unimaginable.

This model is also relevant to businesses increasingly using “lean” principles and depends on global outsourcing practices which may compromise the visibility of the supply chain. Transparency of operations within the corporation (internal risk drivers) is as critical as data from business partners in “lean” and “global” operations to evaluate
external risk drivers. In some cases, outlier events may be even more influential given that uncertainty is far greater than risk and it is very difficult to assign proper weights to distant elephants.

Use of GARCH in supply chain to estimate risk through VaR (value at risk) analysis may also help create a merger of financial and physical supply chains. The financial supply chain, which drives financial settlement, takes over where the physical supply chain ends. Exporters want rapid payment while importers demand accurate data on goods received to better manage inventory and cash-flow to optimize working capital management. Thus, capital efficiency (the traditional domain of the chief financial officer or CFO) depends on data and sharing of information (traditional domain of the chief information or chief technology officer, CIO or CTO) about cross-border movement of goods (customs and excise), transfer of title, risk mitigation and payment. Facilitation of the flow of (decisionable, actionable) information across physical and financial supply chains has a direct impact on working capital.

From a risk management perspective, the supply chain, therefore, appears to evolve as a component of the CFO’s responsibility. Adapting the GARCH model to serve as a tool in supply chain risk analysis may offer financial managers a familiar tool that may yield clues to effective supply chain risk mitigation strategies. In general, comprehensive solutions are necessary over the life of a transaction cycle that may integrate cash management, trade settlement, finance, logistics, supply nodes, procurement, demand projections, inventory, human resources, regulatory compliance and management of information across physical and financial supply chain boundaries. Creating one or more models that may work in synergy and integrating such real-world scenarios is a challenge.

However challenging, risk management may soon become a “household” issue for business and industry. Cost of doing business with and in the US may soon have to figure in the cost necessary to implement transparency in order to mitigate risk. Businesses must share data with US Department of Homeland Security if their goods originate overseas. This model of data sharing may soon be adopted by other countries, determined to counter terrorism. The move toward global supply chain transparency is not a matter of if but a question of when, due to the great uncertainty posed by terrorists that heighten security risks. The lack of analytical tools to make sense of this data may create many more problems before it starts providing solutions. If even a tiny fraction of the 25,000 containers that arrive in US ports each day require inspection, then businesses will face costly delays in receiving customs clearance. In October 2002, a war game that mimicked this delay found that closing US ports for only 12 days created a 60-day container backlog and cost the economy roughly $58 billion (Worthen, 2006).
The proven success of GARCH in finance and the potential to adapt GARCH for business operations may be viewed as one of the promising solutions to offer a synergistic multi-faceted tool for risk-adjusted supply chain management by acting as a bridge for some of the interdependent issues in global business: finance, supply chain and security risk analysis.

**Illustration 1: GARCH and Global Risk Management?**

Illustration outlines some of the pilot projects in progress in the US. There exists a possibility of a mandate by the US in the form of Customs-Trade Partnership Against Terrorism (C-TPAT). To qualify for C-TPAT Tier 3 certification, business must share data through the Advanced Trade Data Initiative (ATDI). Sharing sensitive data will add layers of data security. With data from ATDI, the customs "enterprise" system or Automated Commercial Environment (ACE) is expected to run analysis to spot anomalies, integrate biometric information (individuals, meat and agricultural products), perform non-obvious relationship analysis (NORA) and forecast risk profile associated with containers or shipments. Armed with this information, customs aims to selectively "target" cargo for inspections.
6. Concluding Remarks

In this paper, we propose a model for forecasting with potential broad spectrum applications that include supply chain management. The model is based on advances in time series econometrics. GARCH technique is used to explicitly model the volatility generally associated with supply chains. A VAR framework captures the dynamics of interactions that characterize multi-stage SCM. From a theoretical point of view, such a model is expected to yield an accurate forecast, thereby reducing some of the operational inefficiencies. In addition, businesses and security organizations may benefit from GARCH because it may enable the quantification of value at risk associated with a wide variety of processes that require better tools for management of risk.

The proposed model, by its very construction, requires high volume data to estimate a large number of coefficients. Availability of high volume data may not be the limiting factor in view of the renewed interest in automatic identification technologies (AIT) that may facilitate acquisition of real-time data from products or objects affixed with RFID tags. Although speculative, it stands to reason that use of a GARCH type model may enhance the ROI from AIT infrastructure by delivering value from acquired data. However, understanding the “meaning” of the information from data is an area still steeped in quagmire but may soon begin to experience some clarity if the operational processes take advantage of the increasing diffusion of the semantic web and organic growth of ontological frameworks to support intelligent decision systems coupled to agent networks (Datta, 2006).

Rigorous validation of the proposed model with real world data is the next step. In one isolated experiment, the model proposed in this paper was tested to compare forecasting accuracy. When simulated using real world data and compared to traditional CLRM type techniques, the GARCH type model provided a forecast that was appreciably closer to the observed or realized value (Don Graham, personal communication). This observation is immature. Several more experiments with rigorous controls must be performed before this result may be even considered to offer “preliminary” evidence that the GARCH type model proposed in this paper may represent an advanced tool.

In this paper, we have attempted to coalesce a few ideas toward a ‘solutions’ approach aimed to model volatility in supply chain and in the process, perhaps, better manage risk. It is possible that business, industry, governments, consultants and academics with deep knowledge in one or more fields, may spend the next few decades striving to synthesize one or more models of effective modus operandi to combine these ideas with other emerging concepts, tools, technologies and standards to collectively better understand, analyze and respond to uncertainty. However, the inclination to reject deep rooted ideas based on inconclusive results from pilot projects is a detrimental trend and begs to ask the question whether one can aspire to build an elephant using a mouse as a model.
Acknowledgements

This work was made possible by the financial support from the sponsors of the MIT Forum for Supply Chain Innovation. The corresponding author benefited from numerous discussions with industry and government thought leaders, worldwide. Special thanks are due to Robert Engle (New York University), Jon Stine (Intel Corporation), Bob Betts (MainStreet Applications), Joseph Salvo (GE Global Research), Franz Dill (Procter & Gamble), Andy Mulholland (Capgemini), Peter Koudal (Deloitte Research), Reuben Slone (OfficeMax), Jeff Wilke & Micah Samuels (Amazon), Jeffrey Tew (General Motors), Paul Kern (The Cohen Group), Greg Parlier (SAIC), Kevin Larson (AAR), Stan Horowitz (Institute for Defense Analysis), Robert Ciskowski, Grace Lin and Robert Mayberry (IBM), Alexander Renz (Microsoft - Copenhagen, Denmark), Franz Hero (SAP AG - Walldorf, Germany), Hideki Sakai (SAP - Tokyo, Japan), Satoshi Arizumi (Sony - Tokyo, Japan), Masami Kinugasa and Sam Nakane (PoweredCom - Tokyo, Japan), Susumu Takahashi (Agnie - Tokyo, Japan), Eelco de Jong (LogicaCMG - Hoofdorp, Netherlands), Michel Danet, Ray McDonagh & Dietmar Jost (World Customs Organization - Brussels, Belgium), Roongrote Rangsiyopash, Bhanumas Srisukh, Bunn Kasemsup & Pimchanok Chuengcharoen (Siam Cement Group - Bangkok, Thailand), Rajanish Dass (Indian Institute of Management - Ahmedabad, India), Louis Brennan, Kate Phillip Connolly & John Murray (Trinity College - Dublin, Ireland), Petri Kalliokoski & Sanju Toivonen (VTT - Espoo, Finland), Raimo Pulkkinen & Heidi Lindroth (TEKES - Helsinki, Finland), Vesa Salminen (Lappeenranta University of Technology - Lappeenranta, Finland), Tarja Pursiainen (Aker Kvaerner - Vantaa, Finland), Pekka Malinen (Helsinki University of Technology – Helsinki, Finland), Kenth Lumsden & Gunnar Stefansson (Chalmers University of Technology - Goteborg, Sweden), Yichen Lin (National University of Tainan, Taiwan), Jr Jung Lyu (National Cheng Kung University, Taiwan), Han Pang Hwang (National Taiwan University, Taipei, Taiwan), Dominique Estampe (ISLI - Bordeaux, France), Alain Lempereur & Michele Pekar (ESSEC - Paris, France), Gerry Frizelle (University of Cambridge, Cambridge, UK), Yoon Chang (Hankuk Aviation University, Goyang-City, Korea), Alexander Zier (The Hasso Plattner Institute - Berlin, Germany), Michael Barr (California Agricultural Leadership Foundation), Ozalp Ozer (Stanford University), Gita Swamy & Ananth Raman (Harvard Business School), Nicole DeHoratius (University of Chicago School of Business), Mahua Barari (Missouri State University), Sanjay Sarma, David Brock, Charles Fine, Hui Li, Mike Li, David Simchi-Levi & Joseph Sussman of Massachusetts Institute of Technology.

References


Chapter 10
Forecasting and Risk Analysis in Supply Chain Management: GARCH Proof of Concept

Shoumen Datta, Don P. Graham, Nikhil Sagar, Pat Doody, Reuben Slone and Olli-Pekka Hilmola

Abstract Forecasting is an underestimated field of research in supply chain management. Recently advanced methods are coming into use. Initial results presented in this chapter are encouraging, but may require changes in policies for collaboration and transparency. In this chapter we explore advanced forecasting tools for decision support in supply chain scenarios and provide preliminary simulation results from their impact on demand amplification. Preliminary results presented in this chapter, suggests that advanced methods may be useful to predict oscillated demand but their performance may be constrained by current structural and operating policies as well as limited availability of data. Improvements to reduce demand amplification, for example, may decrease the risk of out of stock but increase operating cost or risk of excess inventory.

Key words: Forecasting; SCM; demand amplification; risk management; intelligent decision systems; auto-id data; GARCH; RFID; operations management

10.1 Introduction

Uncertainty fuels the need for risk management although risk, if adequately measured, may be less than uncertainty, if measurable. Forecasting may be viewed as a bridge between uncertainty and risk, if a forecast peels away some degrees of uncertainty but on the other hand, for example, may increase the risk of inventory. Therefore, forecasting continues to present significant challenges. Boyle et al. (2008) presented findings from electronics industry, where original equipment manufacturers (OEM) could not predict demand beyond a 4 week horizon. Moon et al. (2000) presented demand forecasting from Lucent (Alcatel-Lucent), demonstrating improvement in forecasting accuracy (60% to 80–85%). Related observations (Datta, 2008a) resulted in inventory markdowns.
Availability of increasing volumes of data (Reyes et al., 2007; Reyes and Frazier 2008) demands tools that can extract value from data. Recent research has shown that advanced forecasting tools enable improvements in supply chain performance (Zhao et al., 2001; Zhao et al., 2002; Bayraktar et al., 2008; Wright and Yuan, 2008), if certain pre-requisites are optimized (ordering policies, inventory collaboration). Autoregressive models have been effective in macroeconomic inventory forecasts (Albertson and Aylen, 2003). Zhao et al. (2002) and Bayraktar et al. (2008) emphasize that the role of forecasting in supply chain is to indicate the right direction for the actors rather than being exactly right, at every moment. Choosing the correct forecasting method is often a complex issue (Chatfield and Yar, 1988).

The purpose of this work is to explore how advanced forecasting methods could be applied in global supply chain management and their requirements. We present real world results and use simulation of a 4-stage supply chain model, beer-game (Vensim simulation). We have also used SPSS statistical analysis software to construct autoregressive forecasting models. The problems may be described as: (1) how to construct autoregressive forecasting models for a supply chain environment and (2) what changes may be needed in supply chain design to apply these advanced forecasting models? In the next section, we introduce a few challenges and Sect. 10.3 discusses demand amplification in supply chain management. In Sect. 10.4 we discuss features of autoregressive models and generalized autoregressive conditional heteroskedasticity (GARCH). Data and analysis from a supply chain inventory model using GARCH is presented and although the results are preliminary, they are encouraging. Concluding thoughts and further research issues are proposed in Sect. 10.5.

### 10.2 Supply Chain Management and Demand Amplification

Despite rapid advances in SCM and logistics, inefficiencies still persist and are reflected in related costs (Datta et al., 2004). In developing nations the actual amounts are lower, but proportional share is higher (Barros and Hilmola, 2007). One of the logistically unfriendly country groups are oil producers (Arvis et al., 2007).

The high cost for operations offer prosperity for the service providers. In 2006, AP Moller-Maersk raked in US$ 46.8 billion in revenues (Hilmola and Szekely 2008). Deutsche Post reported revenues of € 63.5 billion in 2007 (Annual Report, 2007). Profitability and growth of these services are increasing, fueled by globalization. Global transportation growth exceeds global GDP growth (United Nations, 2005, 2007), since trade grows twice as fast as GDP. For decades companies emphasized lower inventories and streamlined supply chains but it has resulted in a situation (Chen et al., 2005; Kros et al., 2006) where material handling in distribution centers has increased (transportation growth combined with lower lot sizes). Management science and practice continues to explore ways to decrease transaction costs (Coase 1937, 1960, 1972, 1992) through real-time information arbitrage. Cooper and Tracey (2005) reported that in the 1990’s Wal-Mart had an informa-
tion exchange capacity of 24 terabytes. While massive investments in information technology (IT) may be prudent, the sheer volume of data begs to ask the question whether we have the tools to separate data from noise and if we have systems that can transform data into decisionable information.

In supply chain management, the issue of demand amplification or Bullwhip effect has been in the forefront for some time (Forrester, 1958; Lee et al., 1997) but it took decades before its importance was recognized. The development of supply chain management (Oliver and Webber, 1982; Houlihan, 1985) catalysed by globalization, highlighted the strategic importance of logistics and pivotal role of information technology. Small demand changes in the consumer phase resulted in situations, where factories and other value chain partners faced sudden peaks and down turns in demand, inventory holdings and a corresponding impact on production and delivery (delivery structure phenomenon due to Bullwhip effect is referred to as “reverse amplification” by Holweg and Bicheno (2000) and Hines et al. (2000) referred to it as “splash back”). Human intervention to tame the Bullwhip effect, compared to simple heuristics, leads to higher demand amplification (Sterman 1989).

It follows that demand amplification may have serious consequences due to increased uncertainty and increases the significance of risk management. During a down turn, Towill (2005) showed that amplification causes possible shortages on product volume (products are not ordered, even if demand is undiminished) and variety as well as on idle capacity in operations and involves potential layoff costs. In the case of positive demand, Towill (2005) identifies that stock deterioration and sales cannibalization produces lost income.

Consumers purchase products lured by discounts and that diminishes sales in the following time periods or seasons (Warburton and Stratton, 2002). During upswings, operations cost a premium for manufacturing and distribution (orders increase rapidly), but also decreases productivity development and increases waste levels. Recent emphasis on outsourcing and large-scale utilization of low cost sourcing has worsened demand amplification (Lowson, 2001; Warburton and Stratton, 2002; Stratton and Warburton, 2003; Hilletofth and Hilimola, 2008). Risks associated with production and transportation delays are considerably higher. To mitigate such risks, some corporations are using responsiveness as a strategic differentiator and have built their supply chains to react on market changes through more localized supply networks, for example, Benetton (Dapiran, 1992), Zara (Fraiman and Singh, 2002), Griffin (Warburton and Stratton, 2002; Stratton and Warburton, 2003), Obermeyer (Fisher et al., 1994) and NEXT (Towill, 2005). The carbon footprint of sourcing strategies will become increasingly relevant in view of future legislation. Logistics may ultimately benefit from a disruptive innovation (Datta, 2008b) in energy sourcing and management using wireless sensors networks (Datta, 2008e).

In recent decades, even macroeconomists are including inventory as a key indicator of economic decline of national economies (Ramey, 1989; Albertson and Aylen, 2003). Ramey (1989) argued that manufacturing input inventories, raw materials and work in process (WIP), fluctuated most in recession, while end-item or finished goods inventories fluctuate less (Table 10.1). However, the labour market volatility is also an issue in changing economic environments (Ramey, 1989). Although,
Table 10.1 All numbers billions of US-dollars (1972), annual rates of change

<table>
<thead>
<tr>
<th>Recessions</th>
<th>Retail</th>
<th>Wholesale</th>
<th>Manufact. Finished Inventories</th>
<th>Manufact. Input Inventories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960: 1–1960: 4</td>
<td>−6.3</td>
<td>−1.7</td>
<td>−3.1</td>
<td>−6.3</td>
</tr>
<tr>
<td>1969: 3–1970: 4</td>
<td>−8.2</td>
<td>1.2</td>
<td>−0.4</td>
<td>−5.2</td>
</tr>
<tr>
<td>1973: 4–1975: 1</td>
<td>−16.0</td>
<td>−5.8</td>
<td>2.4</td>
<td>−13.2</td>
</tr>
<tr>
<td>1980: 1–1980: 2</td>
<td>3.6</td>
<td>1.9</td>
<td>−0.3</td>
<td>−4.1</td>
</tr>
<tr>
<td>1981: 3–1982: 4</td>
<td>−7.6</td>
<td>−2.3</td>
<td>−7.8</td>
<td>−11.1</td>
</tr>
</tbody>
</table>

Fig. 10.1 Capacity addition change in US computer and semiconductor 1986–2008 (Federal Reserve 2008)

Inventory positions seem to fluctuate, Albertson and Aylen (2003) argue that autoregressive forecasting models are able to forecast next period situation with a 50% accuracy. While autoregressive techniques have been widely used in finance (and economics) in the past few decades, they may not have been applied or explored as decision support tools by supply chain planners or analysts in the area of supply chain management (Datta et al., 2007) or in other verticals (healthcare, energy).

Logic of demand amplification is evident in economic cycles (Forrester, 1976; Sterman, 1985). Order backlog, existing inventory holdings, amount of production, amount of employment and capacity additions were used in simulation trials to forecast different levels of economic cycles. In long-term changes, both Forrester (1976) and Sterman (1985) have emphasized the importance of capacity additions. A similar methodology has been used in maritime economics to estimate price level changes (Dikos et al., 2006) and investment cycle lengths in capital intensive industries (Berends and Romme, 2001). Hilmola (2007) has shown that capacity additions in US in semiconductors and computers industries may explain the behavior of stock market indices.
10.3 Beer Game and Role of Advanced Forecasting Methods

Forrester (1958) introduced a classical 4-stage beer-game simulation and revealed that demand information amplifies within the supply chain as we move further upstream. Figure 10.2 shows, customer demand is flat at 8 units per time period (it increased from 4 to 8 during time period of 100), but over-under reaction appears when supply chain is moved further with respect to time. Production orders spike to over 40 units per period, while waiting collapses to 0 units only 15 time units later. This occurs mostly due to time-delayed supply process in each stage, which is following make-to-stock (MTS) inventory principles (each phase has “target” for end-item inventory levels, which they try to reach with order algorithms).

![Fig. 10.2 Forrester effect (delay = 4, step-wise demand change from 4 to 8 units during time unit 100)](image)

It may be observed from previous research that conventional forecasting methods do not reduce negative impact from demand amplification. As shown in Fig. 10.3, classical forecasting techniques such as exponential smoothed moving average (EWMA) only heightens demand amplification (highest value reaches above 50 units per time period) due to the assumption that all previous values should be used to predict demand. Although, use of EWMA may be justified under certain circumstances, the non-discriminatory or mandatory use of past data to predict future demand may often generate an undesirable over-reaction.

Figure 10.4 reveals implementation problems for sharing of demand information. Often different supply chain phases use different competing suppliers to gain cost efficiency and true demand is often confidential. Lam and Pestle (2006) describe 60% of respondents in a survey (in China) indicating that their customers are not willing to exchange information. Sharing data about high demand periods could result in inflated purchase price if suppliers decide to form cartels. However, it has been shown in a signaling game theoretic approach that sharing of information increases total supply chain profit (Datta, 2004).
In previous and current work, we suggest that improving forecasting accuracy could be profitable by using advanced forecasting methods, such as autoregressive moving average (ARMA) models. Figure 10.5 shows that demand forecasting in an amplified environment may be completed by assigning a positive value for last observed demand and a negative coefficient for older observations (for simplicity we have used lag of one and two).

Table 10.2 shows coefficient of two lagging parameters are within the neighborhood of ARMA models built with a larger amount of data. However, the differences among different models are rather minimal.

Applying advanced forecasting models to tame the Bullwhip effect is challenging because it calls for process transformation (Zhao et al., 2002; Bayraktar et al., 2008). In Figure 10.6 the manufacturing unit reserves forecasted amount of inventory one

---

1 Juha Saranen, Lappeenranta University of Technology, Finland
Fig. 10.5 Partial autocorrelation of four stage supply chain data from production phase (2401 observations from 300 time units)

Table 10.2 ARMA models built with 0.125 interval data from original beer-game setting for production phase (number of observations in integer time units given in parenthesis)

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>Co-efficient $t-1$</th>
<th>Co-efficient $t-2$</th>
<th>Goodness of fit $R^2$ (%)</th>
<th>$R^2$ (%) whole sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 (12.5)</td>
<td>1.936</td>
<td>-0.938</td>
<td>99.8</td>
<td>99.97</td>
</tr>
<tr>
<td>200 (25)</td>
<td>1.937</td>
<td>-0.939</td>
<td>99.9</td>
<td>99.97</td>
</tr>
<tr>
<td>300 (37.5)</td>
<td>1.937</td>
<td>-0.939</td>
<td>99.9</td>
<td>99.97</td>
</tr>
<tr>
<td>2401 (300.1)</td>
<td>1.918</td>
<td>-0.919</td>
<td>100.0</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 10.6 Production orders as ARMA model of wholesaler demand is used in production orders with modifications on operating structure
period beforehand (in order to distribute knowledge from future demand into operative decisions). However, this is not enough. We have used another manufacturing unit, which serves as an emergency inventory, dedicated for sudden upswings of demand. This emergency inventory is served with local short response manufacturing, which continuously replenishes emergency inventory with low lot size (in this case lot size is 7 units, lead time to emergency inventory is 1 time unit, instead of 4). During simulation trials we explored how ARMA model may be built within dynamic environments with respect to time.

10.4 Advanced Statistical Models

Forecasting demand is a key tool in managing uncertainty. Forecast accuracy depends on the understanding and coverage of parameters as well as the accuracy of historic data available for each variable that may have an impact on the forecast or predictive analytics. The broad spectrum applicability of forecasting includes such diverse verticals as healthcare and energy utilization (Datta 2008e, 2008f).

One of the assumptions in the Classical Linear Regression Models relates to homoskedasticity (homo $\equiv$ equal, skedasticity $\equiv$ variance or mean squared deviation $(\sigma^2)$, a measure of volatility) or constant variance for different observations of the error term. Forecast errors are heteroskedastic (unequal or non-constant variance). For example, in multi-stage supply chains, the error associated with manufacturer’s forecast of sales of finished goods may have a much larger variance than the error associated with retailer’s projections (the assumption being that the proximity of the retailer to the end consumer makes the retailer offer a better or more informed forecast of future sales through improved understanding of end-consumer preferences). The upstream variability reflected in the Bullwhip Effect violates the basic premise of CLRM, the assumption of homoskedasticity. CLRM ignores the real-world heteroskedastic behavior of the error term $\epsilon_t$ and generates forecasts, which may provide a false sense of precision by underestimating the volatility of the error terms.

![Fig. 10.7 Illustrations of homoskedasticity, heteroskedasticity and the Bullwhip Effect](image)

---

2 See www.cids.ie/Research/DCSES.html and www.dcsenergysavings.com
In a homoskedastic distribution, observations of the error term can be thought of as being drawn from the same distribution with mean \( \mu = 0 \) and variance \( \sigma^2 \) for all time periods \( t \).

A distribution is described as heteroskedastic when observations of the error term originate from different distributions with differing widths (measure of variance). In supply chains, the variance of orders is usually larger than that of sales and the distortion increases as one moves upstream from retailer to manufacturer to supplier. Therefore, the assumption of heteroskedasticity, over time, seems appropriate as a characteristic that may be associated with demand amplification or the Bullwhip effect.

While variance of error term may change across cross sectional units at any point in time, it may also change over time. This notion of time varying volatility is frequently observed in financial markets and has been the driving force behind recent advancements in time series techniques. Instead of considering heteroskedasticity as a problem to be corrected (approach taken by CLRM practitioners in assuming homoskedasticity of error term), Robert Engle modelled non-constant time dependent variance (heteroskedasticity) using an autoregressive moving average (ARMA) technique.

ARMA consists of two components, an autoregressive (AR) component and a moving average (MA) component. AR is a technique by which a variable can be regressed on its own lagged values and thus link the present observation of a variable to its past history. For example, today’s sales may depend on sales from yesterday and the day before. MA expresses observations of a variable in terms of current and lagged values of squared random error terms using a moving average.

Robert Engle used the MA approach to propose AutoRegressive Conditional Heteroskedasticity or ARCH to model the time varying volatility in a series. The ‘conditional’ nature of non-constant variance (heteroskedasticity) refers to forecasting of variance conditional upon the information set available. This MA representation was later generalized to an ARMA representation referred to as Generalized AutoRegressive Conditional Heteroskedasticity model or GARCH. The GARCH technique represents a parsimonious model than ARCH, while allowing for an infinite number of past error terms to influence current conditional variance.

These advances from econometrics may be developed into tools for forecasting and risks analytics with a broad spectrum of applications in business, energy, industry, security and healthcare (Datta, 2008d), as well as decision support systems and operations management, for example, supply chain management (Datta et al., 2007).

**Impact of Real-Time High-Volume Data from Automatic Identification Technologies (AIT)**

Tracking technologies evolved from the discovery of the RADAR at MIT in the 1940’s. AIT is slowly beginning to impact information flow in the modern value chain network. Tracking products from manufacturers to retailers may have its ori-
gins in the 1970’s with the introduction of the bar code to identify stock-keeping units (SKU). Now it is embracing AIT or auto id, for example, use of radio frequency identification (RFID). AIT makes it possible to electronically log product movement in the “digital” supply chain. This information may be available in real-time. However, the standards, such as the electronic product code (EPC), to capture unique identification of physical objects and processes (Datta, 2007a) calls for a paradigm shift (Datta, 2008c).

Since RFID updates reports every time an individual item or SKU moves from one stage to another stage in the supply chain, or when the item is sold, it is possible to determine the demand for an item in real-time rather than wait for batch updates or weekly or monthly buckets to generate a forecast. The granularity of the data from auto id systems may result in very high volume data which may reveal peaks and troughs of demand for the product, hourly or daily. This volatility is lost when data is aggregated in buckets or batch.

Extracting the value from this high volume near real-time data and deciphering the meaning of the implicit volatility may be a boon to business intelligence and predictive analytics, including forecasting.

Indeed many warehouses adopt an inventory policy of ordering products when stock levels fall below a certain minimum amount $s$ and order up to a maximum amount $S$ [the $(s, S)$ policy]. With auto id it is possible to ascertain this at the instant the threshold is attained, thereby, eliminating the likelihood of out of stock (OOS). Hence, it follows that capitalizing on the increased volume of near real-time demand data from auto id may have profound impact on supply chain forecasting.

However, current software with its CLRM engines and clustering approach, does not improve forecasts even with high volume data. The assumption of error terms implicit in CLRM limits the gains in forecast accuracy from high volume data and fails to show return on investment (ROI) from adoption of (new) auto id tools.

It is our objective to justify why deployment of new tools, for example, auto id, calls for adoption of new techniques for data analytics, for example, advanced techniques from financial econometrics. It is safe to state that new streams of data emerging from a multitude of sources, for example, auto id and sensors, cannot yield value or ROI if used in conjunction with archaic software systems running ancient forms of analytical engines that are typically CLRM based, at least, in the forecasting domain.

To extract decisionable information from high volume near real-time auto id and sensor data, the use of techniques like GARCH deserves intense exploration. The fact that GARCH may be a clue to generating ROI from auto id and RFID data is no accident because GARCH requires high volume data to be effectively utilized and generate results with higher accuracy levels. Hence, this convergence of auto id data with tools from econometrics may be an innovative confluence that may be useful in any vertical in any operation including security and healthcare, as well as obvious and immediate use in supply chain management. Datta et al. (2007) and this chapter, has attempted to highlight how to extract the advances in econometrics from the world of finance and generalize their valuable use in decision systems, with a broad spectrum of general applications.
Evolution of such a tool may help analysts and planning managers since a key concern of any manager is the accuracy of the predictions on which their budget is based. The proper allocation of resources for acquisition of personnel and equipment has long been plagued by errors in traditional forecasting. A part of the answer to this problem may be latent in the potential for the combined use of VAR (vector autoregression) with other forecasting techniques. The primary VAR model best suited for the planning function in resource allocation is the standard GARCH model. It is well suited for pragmatic studies involving supply chain, army personnel requirements and defense equipment requirements. Any system that may be modeled using time-series data, may explore how to include GARCH based on the error correction innovation that may improve forecasts. Forecast accuracy of GARCH model may be quantified in a number of different ways. Traditional methods are:

1. Mean Square Error.
2. Mean Average Percentage Error.
3. Aikiki Information.

Once the forecast is developed, the accuracy can be measured by comparing the actual observed values with the predicted values. If the collected data falls within the confidence interval of the forecast model, then the model provides a good fit for the system. Although GARCH is useful in forecasting it is important to realize that it was designed to model volatility and can be applied to positive series but not to economic series.

**GARCH Proof of Concept Demonstrated Using an Example from Spare Parts Inventory Management**

Although GARCH models have been almost exclusively used for financial forecasting in the past, we propose (Datta et al. 2007b) that with appropriate modifications, it may be applicable to other areas. An operation exhibiting volatility may benefit from VAR-GARCH in addition to other techniques, either in isolation or in combination. It is known that in times of conflict military supply chains experience spikes in demand. These spikes and troughs occur over a short period of time and result in losses due to OOS (out of stock) or surplus. GARCH may minimize forecast errors and fiscal losses in some of the following domains:

1. Cost of personnel, supplies, support;
2. Planning, programming and budgeting;
3. Defense program and fiscal guidance development;
4. Force planning and financial program development.

In one preliminary study (supported by the US DoD, Institute for Defense Analysis, Washington DC and also mentioned in Datta et al., 2007) the spare parts supply chain of a military base was examined for inefficiencies. The data for a 9 month period was collected for a spare part for a military vehicle (HumV). Because the
US Department of Defense affixes RFID tags on some spare parts from some of its suppliers, the hourly auto id demand data was available for analysis. The historical data was used to develop CLRM and GARCH (1,1) model. In this case, the linear regression model was found to have Mean Square Error (MSE) of 0.20 (20%). By comparison, the GARCH (1,1) model produced a MSE of 0.06 (6.7%). These results are encouraging and the US DoD case lends credibility for exploration of GARCH in forecasting analytics. Although this finding is promising, it needs to be repeated with other forms of data and subjected to rigorous mathematical analysis.

Financial Profit from Application of GARCH Technique in Retail Inventory Management

A pilot implementation using GARCH in a commercial supply chain has been undertaken with real-world retail data from a major US retailer. Preliminary results reveal that using GARCH as a forecasting technique offers some advantages (even with limited data volume) compared to CLRM and ARMA. The retail data are from office supply products for business and home customers. Thousands of product lines are sourced from different suppliers, globally. In this study, nine different SKU’s from three different product classes were chosen for analysis. It is not known whether some of the products may suffer from seasonality effects. For each of the 9 products the historical demand data is available for 70 continuous weeks. 52 weeks of history was used to develop projections of demand variability for the subsequent 18 weeks. This projection was compared to the actual observations of demand variability over the 18 week test horizon. Three different techniques were used to forecast the standard deviation. The methods used were CLRM, ARMA and GARCH.

Figure 10.8 shows the performance of each of the three forecasting techniques based on retail data on the 9 products. The error of the forecasted standard deviation is calculated as an average over the final 16 weeks of forecast data. CLRM is outperformed by ARMA and GARCH models for almost all SKU’s.

The results suggest that GARCH may be better (or as good as) across different SKU’s. GARCH outperforms ARMA for a number of products. For the eight favorable tests, an average improvement of 800 basis points was observed. That translates to 1.1% in-stock improvement from a preliminary application of GARCH. In this real-world case, that amounts to about $13 million in additional revenue in terms of recovered lost sales. The product-dependent variability of GARCH performance may be linked to seasonality or other factors (accuracy of input data). Further testing with granular data (hourly or daily) and higher volume data per SKU may increase the accuracy and benefits from using the GARCH technique in forecasting. It may be apparent that systemic use of GARCH type techniques in forecasting, therefore, may substantially improve corporate profit.
Fig. 10.8 Classic Linear Regression Model (CLRM) is almost always outperformed by ARMA and GARCH.

Fig. 10.9 Comparing autoregressive techniques: ARMA vs. GARCH (same data set but excluding CLRM)

It is not difficult to extrapolate that high volume item level retail sales or inventory data (per minute or by the hour) may be available with the diffusion of item level tagging using RFID or radio frequency identification tags. The volume of the data may increase exponentially if embedded sensors are deployed to enhance security and/or detect movement of any physical object from any location. Businesses dealing with short life cycle products (electronics, semi-conductors industries) may explore how these advanced techniques may help to reduce the volatility of supply-
demand since the ability to re-address sales or marketing issues are often limited if the shelf-life of the product is merely a few months (laptops, MP3 players, cell phones).

10.5 Temporary Conclusion

Making sense of data may benefit from high volume data acquisition and analysis using GARCH and VAR-MGARCH (Datta et al., 2007) techniques in addition to and in combination with other tools for forecasting and risk analysis in diverse verticals that may span from healthcare to energy (Datta, 2008e). In this work, we explored the possibility of using advanced forecasting methods in context of supply chains and demonstrated financial profitability from use of the GARCH technique. It remains unexplored if concomitant business process transformation may be necessary to obtain even better results. The proposed advanced forecasting models, by their very construction require high volume data. Availability of high volume data may not be the limiting factor in view of the renewed interest in automatic identification technologies (AIT) that may facilitate acquisition of real-time data from products or objects with RFID tags or embedded sensors. It is no longer a speculation but based on proof that use of advanced forecasting methods may enhance profitability and ICT investments required to acquire real-time data may generate significant return on investment (ROI). However, understanding the “meaning” of the information from data is an area still steeped in quagmire but may soon begin to experience some clarity if the operational processes take advantage of the increasing diffusion of the semantic web and organic growth of ontological frameworks to support ambient intelligence in decision systems coupled to intelligent agent networks (Datta, 2006). To move ahead, we propose to bolster the GARCH proof of concepts through pilot implementations of analytical engines in diverse verticals and explore advanced forecasting models as an integrated part and parcel of real-world business processes and systems including the emerging field of carbonomics (Datta, 2008f).

Acknowledgements The corresponding author (SD) is deeply indebted to Prof. Clive Granger (UCSD) for his guidance, critical reading and review of the manuscript. The author (SD) is grateful to Professor Robert Engle (NYU) for his advice. The detailed editing of this document was accomplished by Brigid Crowley at the School of Science and Computing at ITT, Ireland. Comments from General Paul Kern of The Cohen Group (Washington DC) and Dr. Stan Horowitz of the Institute for Defense Analysis (Washington DC) were very helpful to the corresponding author (SD). The encouragement from Mr Kazunori Miyabayashi, Senior Vice President of Hitachi (Tokyo, Japan) and Professor Joseph Sussman, MIT is acknowledged. This work was partially supported by the members and sponsors of the MIT Forum for Supply Chain Innovation, Massachusetts Institute of Technology (Cambridge, Massachusetts) and the Centre for Innovation in Distributed Systems.
References


Abstract: If a typhoon in the South China Sea impacts the shipment and delivery of memory chips to an assembly plant in Mexico City, you can count on the ripple effect to impact financial service providers, manufacturers and suppliers, shippers in charge of logistics and of course, the end-consumer. Can we plan to reduce the risk arising from such uncertainties? Can businesses (semiconductor plants, banks, logistics providers) cooperate to minimize uncertainties? Conventional wisdom states that uncertainties are equivalent to accidents and hence by nature remain unpredictable. However, application of tools and technologies based on emerging standards may partially disprove such wisdom. Focus on demand management may be the guiding light for supply chain practitioners. Can we collapse information asymmetries (between manufacturers and their lending institutions, for example) and add far more value to networks or demand webs? Real-time operational adaptability is key, especially in fast ‘clockspeed’ industries. Confluence of emerging tools, technologies and standards are required to converge to catalyze the evolution of such adaptable enterprise. Can real-time distributed data, in-network processing, Agent-based autonomy, taken together, tame the Bullwhip Effect? Can the (semantic) web catalyze the “Nash Equilibrium” of people (games) and information (theory) in our quest for real time “predictive” decision support systems? We will explore a few of these issues and how they may coalesce to enable the adaptive value network of the future.

Key words: Adaptive value networks, Game theory, Bullwhip effect, Agent, Automatic Identification Technologies
New technologies for supply chain management and flexible manufacturing imply that businesses can perceive imbalances in inventories at an early stage — virtually in real time — and can cut production promptly in response to the developing signs of unintended inventory build up.

Alan Greenspan
Testimony to the U.S. Senate Committee on Banking, Housing and Urban Affairs (13 February 2001)

*Disclaimer: This article is an over-simplified and incomplete exploration of a few tools and technologies that may converge to influence decision support systems in a manner that may catalyze the transformation of current-day supply chains toward an adaptive value network. In addition to named contributions, the corresponding author has used several sources of information in an effort to ‘connect the dots’ to suggest how apparently distant disciplines, if coalesced, may be complementary. Although the list of references is seriously incomplete, it should be amply clear that the original research is not due to the corresponding author. However, quotes, opinions, comments expressed in this article are solely attributable to the corresponding author and do not represent the views of MIT as an institution or the co-authors or their institutions or organizations. Some terms, for example, information asymmetry, may be used in a generic sense to imply lack of information or information visibility. This article is not an original research document but rather a synthesis of a few ideas, which, if taken together, may be catalytic in the transformation of supply chain management to become adaptive or perhaps even predictive. In this chapter we suggest that adaptive may morph into ‘predictive’ through a confluence of principles. We advocate inclusion of ARCH (autoregressive conditional heteroskedasticity) and GARCH (generalized ARCH) in the context of supply chains to model high frequency (volume) real-time data (from RFID tags) that may also exhibit volatility (http://pages.stern.nyu.edu/~rengle/). All errors of content or coherence are solely due to the corresponding author.
1. INTRODUCTION

“At the science museum in Barcelona, I saw an exhibit that beautifully illustrated ‘chaos.’ A nonlinear version of a pendulum was set up so that a visitor could hold the bob and start out in a chosen position with a chosen velocity. One could then watch the subsequent motion, which was also recorded with a pen on a sheet of paper. The visitor was then invited to seize the bob again and try to imitate exactly the previous initial position and velocity. No matter how carefully it was done, the subsequent motion was quite different from what it was the first time. I asked the museum director what the two men were doing who were standing in a corner, watching us. He replied, “Oh, those are two Dutchmen waiting to take away the “chaos.” Apparently, the exhibit was about to be dismantled and taken to Amsterdam. I have wondered ever since whether the services of those two Dutchmen would not be in great demand across the globe, by organizations that wanted their chaos taken away.” (Gell-Mann 1994).

The holy grail of industry is to remove “chaos” from the supply chain in order to better adapt to demand fluctuations. Managing uncertainty is compounded by the increasing degree of information asymmetry\(^1\) between the supply “chain” (value network\(^2\)) partners (designers, suppliers, distributors, retailers, consumers) who have different and often conflicting objectives, that threaten to create barriers on the road to adaptive business networks of the future (Heinrich and Betts, 2003).

*Ampex pioneered the video recorder market in 1956. Each unit was priced at $50,000 and the only competitors, RCA and Toshiba, were way behind. Sony, JVC and Matsushita were mere observers. Masaru Ibuka, co-founder of Sony and Yuma Shiraishi at JVC, issued directives for their respective engineers to produce an unit that would cost $500, a mere 1\% of Ampex’s price. In the 1980’s, video recorder sales went from $17 million to $2 billion at Sony, $2 million to $2 billion at JVC, $6 million to $3 billion at Matsushita and $296 million to $480 million at Ampex. Failure to adapt eclipsed Ampex. (Tellis and Golder, 1996).*

One business objective of suppliers is to secure large volume purchase commitments (with delivery flexibility) from manufacturers. It conflicts with the manufacturer’s objective that must include rapid response to demand fluctuation yet the manufacturer must mass produce (to take advantage of economies of scale) yet production runs must adapt to fluctuations even though a certain run may have been planned based on demand forecast. Thus, manufacturers may need more or less raw materials and therefore seek flexibility in purchasing raw materials, which conflicts with the supplier’s objective. Manufacturer’s desire to run long production batches are in
conflict with the warehouse and distribution centers that aim to reduce inventory due to storage capacity constraints. The latter increases cost of transportation for all the players.

During 2000, supply chain related costs in USA alone exceeded $1 trillion (10% of GDP), which is close to the entire GDP of Russia, is more than the GDP of Canada or Spain or the combined GDP of all the 22 nations who are members of the League of Arab Nations. The combined GDP of all 22 Arab nations is less than that of Spain (www.wrmea.com/archives/sept-oct02/0209044-2.html; www.bea.doc.gov; www.cia.gov). A mere 10% savings of supply chain costs in USA is nearly equal to the GDP of Ireland. Therefore, tools and processes that may reduce supply chain inefficiencies and help it better adapt to demand changes, are valuable. We will briefly explore some of the tools that may catalyze the adaptive value network.

Some emerging technologies may take leading catalytic roles but technology is not the solution. Ability to adapt supply chains will depend on continuous business process innovation led by management capable of envisioning use of technology as a tool to reduce (1) inefficiencies, (2) uncertainties and (3) information asymmetry within the value network. In essence, decision making processes should respond to (dynamic) information such that the system (enterprise) is able to rapidly adapt and/or respond.

One driver of this transformation (from ‘push’ based supply chain management to ‘pull’ based adaptive value networks) is the potential use of real-time information to catalyze or trigger autonomous decision steps capable of re-planning and execution. By some estimates, business in 2003 generated more than 1 terabyte of data per second (excludes data gathered by automatic identification technologies). Is this equivalent to information? It is unlikely that this data, as is, can be considered as information. Even when we extract the information, will it offer a “transactional” value? The ability to extract intelligence from data to manage information may be the differentiator between companies who will profit from data (such as automatic identification or sensors) versus those who will not. Data that is stored in business systems (such as ERP) may suffer from problems that reduces the value of their information. ERP systems may also compromise the efficacy of dynamic data if the data feeds static systems unable to respond in near real-time. When such ERP data and/or information sources are used by strategic planners for forecasting and optimization, it leaves room for speculation about the validity of the outcome since the process may have been optimized, or forecast delivered, based on “noise” rather than robust dynamic data. Stemming from poor data quality and information asymmetry between supply chain partners, these errors (of optimization, forecasting) accumulate at successive stages of a supply chain and manifests
itself as the generic Bullwhip Effect (Forrester 1961, Sterman 1989, Lee et al., 1997) (Figure 1-1 and Figure 1-2).

Figure 1-1. The Bullwhip Effect (Source: Joshi 2000)

Figure 1-2. How do we taming the Bullwhip Effect? (Source: Joshi 2000)
Example from the semiconductor equipment supply chain shows demand forecast versus actual purchase of equipment by Intel Corporation (Figure 1-3).

![Graph showing demand forecast versus actual purchase of equipment by Intel Corporation]

*Figure 1-3. Intel Tool Order Data 1999-2001 (Source: Cohen, Ho, Ren and Terwiesch, 2003)*

2. **TOWARD ADAPTIVE VALUE NETWORKS: INFORMATION VISIBILITY (TRANSPARENCY)**

Tools and technologies that may be catalytic in taming the Bullwhip Effect may also have an impact on making supply chains more adaptive by ushering in adaptive decision support systems. However, both assume the success of business process innovation to improve intra- and inter-enterprise information exchange as well as efforts to break down data silos as a segue to distributed data infrastructure. In thinking about adaptive supply chain management, it is helpful to analyze how the tools and technology catalysts may help connect objects to processes and processes to decision systems. Some of these “catalysts” may be classified into two broad (albeit arbitrary) categories. We will make an attempt to show how a few of these catalysts may (converge to) transform current supply chains to become more adaptive (Table 1-1).
### 1. ADAPTIVE VALUE NETWORKS

<table>
<thead>
<tr>
<th>Tools and Concepts</th>
<th>Data Mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operations Research and Game Theory</td>
<td>Automatic identification technologies</td>
</tr>
<tr>
<td></td>
<td>(RFID, UWB, GPS)</td>
</tr>
<tr>
<td>Distributed Artificial Intelligence and Agents</td>
<td>Wireless protocols (802.11, 802.16)</td>
</tr>
<tr>
<td>‘Clockspeed’ as defined by Charles Fine, MIT</td>
<td>Sensor networks</td>
</tr>
<tr>
<td>Semantic Web</td>
<td>Simulation</td>
</tr>
</tbody>
</table>

3. **OPERATIONS RESEARCH AND GAME THEORY**

The workhorse of optimization (algorithms) is based on operations research. It is an area of intense research and innumerable sources of information are available. However, Game Theory was not such a “household” name until 1994 when John Nash, and later the movie about him, changed the public’s perception so much so that generic business journals began touting the virtues of Game Theory.


*Managers have much to learn from game theory - provided they use it to clarify their thinking, not as a substitute for business experience.* (The Economist, 15 June 1996).

Game theory helps us model, analyze and understand the behavior of multiple self-interested parties who interact while making decisions. As such, Game Theory deals with interactive optimization problems. In particular, it is a tool for analyzing situations where the parties strive to maximize their (expected) pay-offs while choosing their strategies. Each party’s final pay-off depends on the profile of strategies chosen by all parties. Most business situations can be modeled by a “game” since in any business interaction, involving two or more parties, the pay-off of each party depends on the other party’s actions.

For centuries economists have worked on various game-theoretic models but Neumann and Morgenstern (1944) are credited as the fathers of modern Game Theory. Game Theory has since enjoyed an explosion of developments, including the concept of equilibrium (Nash, 1950), games
with imperfect information (Kuhn, 1953), cooperative games (Aumann, 1959; Shubik, 1962) and auctions (Vickrey, 1961).

3.1 Let the Game Begin

The overarching theme in Game Theory is “interactions.” In business, each decision maker is a player making a decision or choosing a strategy that will be impacted by the competitor.

“A chip manufacturer slashed prices of its desktop and mobile processors just days after a similar move by a rival. We’re going to do what it takes to stay competitive’ on prices, said representative. The companies aggressive price-chopping means the company doesn’t want to give up market share gains, even at the cost of losses on the bottom line (CNet News.com, May 30, 2002)”

In this example, companies compete on price to gain market share. During Q1 of 2002, this semiconductor company increased processor shipments (compared to Q4 of 2001) worth $8 million but processor revenue declined by 3% (sold more chips for less money). Companies engaged in price wars rarely, if ever, benefit from such competition. Reducing prices slightly might increase the overall market potential but decreasing prices beyond a certain limit has a diminishing impact. Eventually the size of the “pie” does not increase anymore and firms fight harder to get a bigger “pie” by slashing prices and profits. Why do firms behave this way? In this situation and in some others, firms are caught in what is known in Game Theory as the “Prisoner’s Dilemma” where the rational response may not be the optimal.

3.1.1 Prisoner’s Dilemma

Two burglars, Alice and Bob, are arrested near the scene of a burglary and interrogated separately. Each suspect can either confess with a hope of a lighter sentence or may refuse to talk (does not confess). The police do not have sufficient information to convict the suspects, unless at least one of them confesses. Each must choose without knowing what the other will do. In other words, each has to choose whether or not to confess and implicate the other. If neither confesses, then both will serve one year on a charge of carrying a concealed weapon (Table 1-2). If both confess and implicate each other, both will go to prison for 10 years. However, if one burglar confesses and implicates the other but the other burglar does not confess, then the one who cooperates with the police will go free, while the other burglar will go
to prison for 20 years on the maximum charge. The “strategy space” in this case is simple: confess or don't confess (each chooses one of the two strategies). The payoffs (penalties) are the sentences served. In effect, Alice chooses a column and Bob chooses a row.

<table>
<thead>
<tr>
<th></th>
<th>Alice</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>Confess</td>
<td>10, 10</td>
</tr>
<tr>
<td>Bob</td>
<td>Does not</td>
<td>20, 0</td>
</tr>
</tbody>
</table>

The two numbers in each cell show the outcomes for the two prisoners when the corresponding pair of strategies are chosen. The number to the left shows the payoff to the person who chooses the rows (Bob) while the number to the right tells the payoff to the person who chooses the columns (Alice). Thus (reading down the first column) if they both confess, each gets 10 years, but if Alice confesses and Bob does not, Bob gets 20 and Alice goes free. Therefore, what strategies are "rational" in this game if both of them want to minimize their sentences? Alice might reason: "Two things can happen: Bob can confess or Bob can keep quiet. If Bob confesses, I get 20 years (if I don't confess) and 10 years if I do confess (cooperate), so in that case it is better to confess. On the other hand, if Bob doesn't confess and I don't either, I get a year but in that case, if I confess I can go free. Either way, it is better if I confess. Therefore, I will confess."

But Bob can and presumably will reason in the same way. So they both reason rationally to confess and go to prison for 10 years each. But, if they had acted "irrationally" and did not confess, they each could have gotten off with a year (http://william-king.www.drexel.edu/top/eco/game/game.html).

Prisoner’s Dilemma is an example of a non-cooperative static game where the players choose strategies simultaneously and are thereafter committed to their chosen strategies. The main issue of such games is the existence and uniqueness of Nash equilibrium (NE). NE is the point where no player has incentive to change her strategy since each player has chosen a strategy that maximizes his or her own payoff given the strategies of the other players. It may be prudent to point out that the fundamental distinction between cooperative and non-cooperative games is that cooperative games allow binding agreements while non-cooperative games do not. Study of cooperative games focuses on the outcome of the game in terms of the value.
created through cooperation of players, but does not specify the actions that each player will take. The study of non-cooperative games is more concerned with the specific actions of the players. Hence the former allows us to model outcomes of more complex business processes.

3.1.2 Dilemma in Prisoner’s Dilemma

A key concept not captured in “Prisoner’s Dilemma” is the repetition of interactions. In business, most players know they will be in the “game” for awhile. Hence, they may choose to cooperate, especially if they deem that cooperation today may increase the chances of cooperation, or even collaboration, in the future. With repeated actions, companies build a reputation, which influences the actions of others. For example, a restaurant might make a higher profit today by selling slightly less fresh food (prepared yesterday), but will it be worth the possible consequence of losing customers in the future? Thus, rationally speaking, companies aim to act strategically with competitors and partners. In practice the behemoths continually try to squeeze their suppliers (blinded by “cost” reduction) and could lose critical partners.

Intel uses its much envied supplier ranking and rating program - which tracks a supplier’s total cost, availability, service, supports responsiveness and quality – to keep its top suppliers on a course for better quality. ‘We reward suppliers who have the best rankings and ratings with more business,’ says Keith Erickson, Director of Purchasing.

As an added incentive, Intel occasionally plugs high-quality suppliers in magazine and newspaper advertisements. The company even lets its top performing suppliers publicize their relationship with Intel. That’s a major marketing coup, considering that Intel supplies 80% of chips used in PCs today and is one of the most recognized brand names in the world.

Given that each party in a supply chain acts entirely on self interest, individual choices collectively do not lead to an “optimal” outcome for the supply chain. Thus, supply chain profit of a “decentralized” supply chain composed of multiple, independently managed companies, is less than the total supply chain profit of the “centralized” version of the same chain where the partner interactions (suppliers, manufacturers, retailers) are managed by a single decision-maker (information symmetry) to optimize total supply chain profits. Sharing of information in centralized supply chains reduces inefficiencies that are obvious in decentralized supply chains due to “double marginalization” stemming from self-centered decision making.
3.1.3 Optimal Profit is Higher in Centralized Supply Chains with Information Sharing (Symmetry)

![Graph showing Centralized versus Decentralized Profit](image)

*Figure 1-4. Optimal Profit versus Forecast (personal communication; Ozalp Ozer, Stanford University)*

One strategy for reducing inefficiencies in decentralized supply chain (and consequent loss of profit) is “vertical integration” where a company owns every part of its supply chain. An example of vertical integration is Ford Motor Company. Earlier in the 20th century, in addition to automobile factories, Henry Ford owned a steel mill, glass factory, rubber plantation and iron mines. Ford’s focus was on (mass production) making the same car, the Model T, cheaper. This approach initially worked well. The price of a Ford Model T fell from $825 in 1908 to $290 in 1924. By 1914, Ford had a 48% share of the US market. By 1920, Ford was manufacturing half the cars made worldwide. Vertical integration allows a company to obtain materials at a low cost and control the entire supply chain.

In today’s economy, customer demand and preferences change rapidly. Companies that focus on core competencies are more likely to be nimble in order to stay ahead of competition and succeed. Hence, we see a trend towards “virtual integration” where supply chains are composed of independently managed but tightly partnered companies. Information sharing and vendor managed inventory (VMI) are strategies successfully used by some companies (such as Dell Corporation) for higher degree of
Chapter 1

virtual integration. However, most companies still find it difficult or face internal reluctance to usher changes in their supply chain practices. Similar issues apply to independently managed intra-company divisions, such as marketing, production, and sales.

Game Theory makes some assumptions in its study of the impact of interactions of multiple players and the resulting dynamics in a market environment. Two key assumptions are:

- Each player in the market acts on self-interest, but they pursue well-defined exogenous objectives, that is, they are rational;
- In choosing a strategy, a player considers the potential reaction of other players and takes into account his or her knowledge of other decision makers’ behavior, that is, he or she reasons strategically.

These assumptions rule out games of pure chance, such as lotteries and slot machines, where strategies do not matter and games without strategic interaction between players, such as Solitaire. Credibility is a central issue in games.

Coca-Cola is developing a vanilla-flavored version of its best-selling flagship cola, a report says, extending the company’s palette of flavorings from Cherry Coke and Diet Coke with lemon. But don’t expect to see a vanilla-flavored Pepsi anytime soon. ‘It’s not something we’re looking at,’ said spokesman Larry Jabbonsky of Pepsi. ‘We think it’s a bit vanilla.’ (USA Today, 1 April 2002).

PepsiCo is launching Pepsi Vanilla and its diet version in stores across the country this weekend. Coke came out with Vanilla Coke in May 2002 and it was a resounding success, selling 90 million cases. ‘We’re a little surprised that Pepsi decided to enter the vanilla segment,’ said Mart Martin of Coca-Cola. ‘When we came out with Vanilla Coke, Pepsi originally said the idea sounded ‘a bit vanilla.’ (CNN/Money August 8, 2003).

3.2 Game Theory in Quantity and Price Competition

Business decisions include what to produce/procure, sell, how much, and at what price. Study of competitive interactions around these issues can be addressed using game theoretic models that focus on price and quantity decisions (several excellent papers including that of Albeniz and Simchi-Levi, 2003).
Quantity competition (Cournot Game) is especially important for commodities where there is an inverse relationship between quantity and market price. Price competition (Bertrand Game), on the other hand, occurs in every market, as competing companies try to maintain or increase market share.

OPEC decided to slash its crude oil production by 1.5 million barrels a day (6%). The issue came to a head this autumn with weakening world economy, together with the uncertainty caused by the Sep 11 attacks on the US, dragged down prices some 30%. The cut is expected to lift OPEC’s benchmark price to $22 a barrel, the group’s minimum target price (CBS News, December 28, 2001).

Burger King will put its Whopper on sale for 99 cents. The move is likely to intensify and prolong the burger price wars that have been roiling the US fast-food industry in recent months. Burger King Officials had said earlier they had little choice given a $1 menu at McDonald’s that included a Whopper-like hamburger called the Big ’N Tasty.” (Chicago Sun-Times, January 3, 2003).

Tesco announced plans to slash £80 million from prices of more than 1,000 products, with some prices falling by more than 30%. The cuts came as rival Asda also said it was slashing selected prices. The cuts echo memories of the supermarket price wars in 1999 as stores fought to capture more customers and increased market share (Sunday Telegraph, January 5, 2003).

Cournot Game
A market with two competing firms, selling homogeneous goods, where the two firms choose production quantities simultaneously, is known as a Cournot Game. It is a static game where the player’s action sets are continuous (each player can produce any non-negative quantity). This is a tacit collusion to raise prices to a jointly optimal level and thus is a “cartel.” A cartel is defined as a combination of producers of any product joined together to control its production, sale and price, so as to obtain a monopoly and restrict competition in any particular industry or commodity (www.legal-database.com). Cartels can be quite unstable. At each stage, the players have a huge incentive to cheat.

On Tuesday, 23 September 2003, an agreement was submitted to the US District Court in Philadelphia for an out-of-court settlement for a suit filed by industrial purchasers in 1999. According to this agreement, International Paper, Weyerhaeuser and Georgia-Pacific will pay US$68
million to avoid litigation related to class-action lawsuits that accused them of conspiring to fix prices for container-board (packaging material).

The oil market is notoriously difficult to balance - demonstrated by the rollercoaster of prices over the last few years. Member states of OPEC do not have identical interests and find it difficult to reach consensus on strategy. Countries with relatively small oil reserves are often seen as ‘hawks’ pushing for higher prices. Meanwhile, producers with massive reserves and small populations fear that high prices will accelerate technological change and the development of new deposits, reducing the value of their oil in the ground (BBC News, February 12, 2003).

**Bertrand Game**
Models situations where firms choose prices rather than quantities. Assume two firms produce identical goods which are perfect substitutes from the consumers’ perspective (consumers will buy from the producer who charges the lowest price). If the firms charge the same price, they will split the market evenly. There are no fixed costs of production and the marginal costs are constant. As in the Cournot Game, the firms act simultaneously. Therefore, when the costs and the products are identical, there exists a unique equilibrium in which all output is sold at the price equal to the marginal cost. Thus, the Bertrand Game suggests that when firms compete on price, and the costs are symmetric, we obtain a perfectly competitive market even in a duopoly situation. However, in real life, customers do not choose based on price alone. For example, Wendy’s fast food chain decided to stay out of the Burger King and McDonald’s price war (for a while) by aiming to gain market share by offering high quality food.

**Stackelberg Game**
In most business situations, firms choose their actions sequentially rather than simultaneously. In price wars, one firm responds after it observes another firm’s actions. For our discussion, consider that firm 1 moves first and firm 2 responds. We call firm 1 the Stackelberg “leader,” and the “follower” is firm 2.

### 3.3 Game Theory in Inventory Optimization

In time-dependent multi-period games, the players’ payoff in each period depends on the actions in the previous, as well as, current periods. The payoff structure may not change from period to period (so called stationary payoffs). This resembles multi-period inventory models in which time periods are connected through the transfer of inventories and backlogs. Due
to this similarity, time-dependent games have applications in supply chain management, for example, Stochastic Games. (For detailed mathematical review, see Cachon and Netessine, 2003).

Stochastic Games may help in analyzing:
- two-echelon game with the wholesaler and retailer making stocking decisions;
- price and service competition;
- game with the retailer exerting sales effort and wholesaler stocking inventory and van;
- two-period game with capacity choice in 1st period and production decision under capacity constraint in 2nd period.

These games involve a sequence of decisions that are separated in time, but many supply chain models rely on continuous-time processes. Such applications of Differential Games are especially valuable in the area of dynamic pricing and in marketing-production games with manufacturer and distributor.

**Biform Games** have been successfully adopted in supply chain management (Anupindi et al., 2001). Consider a game where multiple retailers stock at their own locations, as well as, at several centralized warehouses. In the first (non-cooperative) stage, retailers make stocking decisions. In the second (cooperative) stage, retailers observe demand and decide how much inventory to trans-ship (cross-dock) among locations to better match supply and demand and how to appropriate the resulting additional profits. Variations on this theme are:
- allow retailers to hold back residual inventory. This model has three stages: inventory procurement, decision about how much inventory to share with others and, finally, the trans-shipment stage;
- option of pooling their capacity and investments to maximize the total value. In the first stage, firms choose investment into effort that affects market size. In the second stage, firms bargain over division of market and profits.

### 3.4 Game Theory in Contracts (Revenue Sharing)

This model is motivated by revenue sharing contracts implemented in practice. Blockbuster purchases movies from studios (suppliers) and rents them to customers. The supplier’s wholesale price impacts how many videos Blockbuster orders and hence, how many units are available for rent. Before 1998, purchase price of a video tape from the studio was around $65. Given that video rental fees are $3-$4 per tape, Blockbuster could purchase only a
limited number of videos and suffered lost demand during the initial release period (peak demand <10 weeks). About 20% of customers could not find the desired tape to rent. The studio’s high wholesale price impacted on the quantity purchased by Blockbuster and in turn, revenues and profitability of both firms. Thus, Blockbuster and the studios crafted a revenue sharing agreement such that Blockbuster pays only $8 per tape initially but then gives a portion (30 to 45%) of rental revenues to the studio (supplier). Since this agreement reduced Blockbuster’s initial investment, it could order more tapes to meet peak demand and generate more revenues even with contracted revenue sharing with the studio (supplier). Blockbuster increased its overall market share from 25% to 31% and improved its cash flow by 61% (CNet News.com, October 18, 2000).

3.5 Games with Incomplete Information (Game Theory and Information Asymmetry)

Ubiquitous knowledge about players and decisions or payoffs is rarely a reality in real world supply chains. It is common that one firm may have a better demand forecast than another or a firm may possess superior information regarding its own costs and operating procedures. If a firm knows that another firm may have better information, it may choose actions that take this into account. Game Theory provides tools to study cases with information asymmetry with increasing analytical complexity.

Signaling Game

In its simplest form, a Signaling Game has two players, one of which has better information than the other. The player with the better information makes the first move. For example, a supplier must build capacity for a key component for a manufacturer’s product. The manufacturer has a better demand forecast than the supplier. In an ideal world, the manufacturer may truthfully share his or her demand forecast with the supplier so that the supplier could build the appropriate capacity. But the manufacturer benefits from a larger capacity at the supplier in case of higher demand. Hence, the manufacturer has an incentive to inflate his or her forecast. However, the supplier will bear the cost of building capacity if it believes the manufacturer’s (inflated) forecast. The manufacturer hopes the supplier believes the (inflated) forecast and builds capacity. Fortunately, the supplier is aware of the manufacturer’s “game” to inflate (distort) forecast. What move (signal) from the manufacturer may induce the supplier to believe the forecast is credible? Consider the example below (from Ozalp Ozer of Stanford University).
S announces: 
(w_a, w)

Time realized S updates belief about \( \xi \) and builds capacity \( K \)

M places advance order \( y \) at price \( w_a \)
M orders \( (D-y)^* \) at price \( w \)

\( \epsilon \) realized

Where, \( D = \mu + \xi + \epsilon \)

---

**Figure 1-5. Signaling Game**

In its simplest form, in this example, Demand (D) is represented as a sum of three forecasts (Figure 1-5). A market forecast \( \mu \) is common information and published by commercial analysts. The manufacturer has sources and/or experience to derive private forecast information \( x_i \) which is unknown to the supplier in a decentralized system (information asymmetry). However, the supplier can categorize the manufacturer into certain “types” based on prior actions or credibility of the manufacturer. Thus, the supplier updates its “belief” about the “type” of the manufacturer’s forecast information and may select some value of \( \xi \) that is spread over a distribution (function). This introduces a random variable. The general market uncertainty is given by epsilon (\( \epsilon \)) and neither the manufacturer nor the supplier can control its value, although using appropriate tools, a closer to reality approximation of \( \epsilon \) is possible. This introduces another random variable which is also spread over a distribution (function).

The Signaling Game, shown here, commences with a price announcement by the supplier: \( w \) (regular) and \( w_a \) (advance purchase) price. The manufacturer creates a demand (D) forecast and based on the strength of forecast, reacts to the supplier’s price package by placing an advanced order \( y \) to be purchased at \( w_a \). The volume of \( y \) sends a “signal” to the supplier. The “signal” is used to update the supplier’s “belief” about the credibility of manufacturer’s forecast (D). Based on this, the supplier can determine how much capacity to build (K) to optimize his or her profits (inventory risk). Down the timeline, the market uncertainty is realized and using this value of \( \epsilon \) the manufacturer may update its forecast. The volume of the D minus \( y \) is
the remaining volume the manufacturer orders from the supplier at a price \( w \). While optimization based on Signaling Games may increase profits for manufacturer and supplier, it will still remain vulnerable to the value chosen for the variables \( \xi \) and \( \varepsilon \). But, this may be further reduced using near real-time data (from automatic identification technologies, as we shall discuss in a later section), which offers greater adaptability to demand.

If signaling favors optimization of the supplier’s capacity planning, then what is the manufacturer’s incentive to signal? Does the manufacturer incur a cost to signal? Is the manufacturer’s expected profit in the signaling equilibrium lower than what it would be if the manufacturer’s type were known to the supplier to update his or her “belief” with certainty?

An ideal action for a high demand manufacturer is one that sends the signal of his or her high demand forecast at no cost. If a costless signal does not exist, then the goal is to seek the lowest cost to signal. Whether or not a costless signal exists depends upon what commitments the manufacturer can impose on the supplier. Suppose the manufacturer dictates (contractually) to the supplier a particular capacity level and the supplier accepts the terms. By accepting the contract, the supplier has essentially no choice but to build that level of capacity (severe penalty for non-compliance). This is referred to as “forced compliance” and in this case many costless signals exist for the manufacturer. However, if the supplier could potentially deviate without penalty, referred to as voluntary compliance, then the manufacturer’s signaling task becomes more complex. One solution for a high demand manufacturer is to give a sufficiently large advance payment to the supplier. Since the high demand manufacturer’s profit is higher than the low demand manufacturer’s profit, only a high demand manufacturer could offer such an advance payment. This is referred to as signaling by “burning money” (who can afford to burn money?). A better signal is a contract that is costless to a high demand manufacturer, but expensive to a low demand manufacturer. An example of such a signal is a minimum commitment. The latter is costly only if realized demand is lower than the commitment and the manufacturer is forced by contract to purchase excess inventory. That scenario is less likely for a high demand manufacturer but a minimum commitment may be costly for a low demand manufacturer. Should a manufacturer agree to a minimum commitment if it possesses perfect information? Likely, because these contracts could be used solely for the purpose of signaling information.

**Screening Game**

In this game the player who lacks information makes the first move. For example, a supplier offers a menu of contracts with the intention of getting the manufacturer to reveal his or her type via the contract selected (in
1. ADAPTIVE VALUE NETWORKS

Economics this is referred to as mechanism design). The supplier is in charge of designing a mechanism to extrapolate the manufacturer’s information.

The space of potential contract menus may be large. How many contracts should be offered and what form should they take? Furthermore, for any given menu, the supplier needs to infer for each manufacturer type which contract that manufacturer will choose. The revelation principle begins with the presumption that a set of optimal mechanisms exists. Associated with each of these mechanisms is a Nash Equilibrium (NE) that specifies which contract each manufacturer type chooses and the supplier’s action given the chosen contract. (NE is the point where no player has incentive to change her strategy since each player has chosen a strategy that maximizes her own payoff given the strategies of the other players.) However, it is possible that some manufacturer type chooses a contract that is not designated for that type. For example, a high demand manufacturer chooses an option that the supplier had designed for the low demand manufacturer. Therefore, even though this game does not seem desirable, it is possible that this mechanism is still optimal in the sense that the supplier may not be able to do better on average because the supplier ultimately only cares about optimizing expected profit (not the means by which that profit is achieved). Auction design in the context of supplier procurement contracts and inventory contract design are two of the potential applications of the revelation principle in supply chain management.

Even though an optimal mechanism may exist for the supplier, this does not mean that the supplier earns as much profit as he or she would if he or she knew the manufacturer’s type. The gap between what a manufacturer earns with the menu of contracts and what the same manufacturer would earn if the supplier knew her type is called an information rent. The separation of manufacturer types goes hand in hand with a positive information rent, that is, a manufacturer’s private information allows the manufacturer to keep some rent that the manufacturer would not be able to keep if the supplier knew his or her type. Hence, even though there may not be any cost involved in information revelation with a Signaling Game, the same is not true with a Screening Game.

Bayesian Games

With a Signaling or Screening Game, actions occur sequentially, such that information may be revealed through observation of actions. There also exist games with private information that do not involve signaling or screening. Consider that a single supplier has a finite amount of capacity. There are multiple retailers and each knows his or her demand, but not the demand of other retailers. The supplier announces an allocation rule, the retailers submit their orders. Then, the supplier produces and allocates units.
If the retailer’s total order is less than the supplier’s capacity, then each retailer receives his or her entire order. If the retailer’s total order exceeds the supplier’s capacity, the supplier’s allocation rule is implemented to allocate the capacity. To what extent does the supplier’s allocation rule influence the supplier’s profit, retailer’s profit and the supply chain profit? In this setting the firms (retailers) that have the private information choose their actions simultaneously (no information exchange among retailers). If the supplier’s capacity is fixed before the game starts, the supplier is unable to use any information from retailers (demand) to adapt capacity planning. However, it is possible that correlation exists in the retailers demand information, that is, if a retailer observes his or her demand type to be high, then he or she might assess that other retailers may have high demand types as well (if there is a positive correlation). Thus, each player uses Bayes’ rule to update his or her belief regarding the types of the other players in a Bayesian Game. Bayesian Equilibrium is a set of strategies for each type that is optimal given the updated beliefs with that type and the actions of all other types. If a player deceptively inflates demand (high type) and other players use this information to update their “beliefs” then this effect may contribute to the observed Bullwhip Effect.

3.6 Temporary Conclusion

God definitely plays dice! Combined GT/OR may offer approaches to use (data) dynamic information for continuous optimization in terms of location and real-time availability (improve from visibility to transparency, among players) as a step toward an adaptive value network.

4. AGENTS

Linearization of real world conditions to fit mathematical models, such as Game Theory, may stifle real-time adaptability of value networks. As an example (see preceding section), a Bayesian Game potentially could contribute to the Bullwhip Effect representing wide fluctuations in supply chain. The discrete, dynamic and distributed nature of data and applications require that supply chain solutions do not merely respond to requests for information but anticipate, adapt and (support users to) predict. In that vein, ‘intelligent’ autonomous Agents are an essential tool for adaptive value networks to emerge.

The idea of Agent originated with John McCarthy in the 1950’s at MIT. The term “Agent” was coined by Oliver Selfridge, a colleague of McCarthy’s at MIT. Recent trends, beginning 1977, in Agent systems are
based on research in distributed artificial intelligence. Research from MIT, DARPA, Carnegie-Mellon University and University of Michigan at Ann Arbor has made significant contributions.

We define an autonomous Agent as a software entity that functions continuously in an environment, often inhabited by other Agents. Continuity and autonomy empower Agents to (plan) execute processes in response to changes in the environment without requiring constant human guidance, intervention or top-down control from a system operator. Thus, Agents offer the ability to rapidly adapt. An Agent that functions continuously in an environment over a period of time also learns from experience (patterns). In addition, Agents that inhabit an environment with other Agents in a Multi-Agent System (MAS) are able to communicate, cooperate and are mobile between environments. Agents work best for clearly discernible tasks or processes, such as, to monitor data from, for example, automatic identification technologies (radio frequency identification or RFID), ultrawideband (UWB) transponders, global positioning system (GPS), WiFi and sensors. Data Agents can share this data with high level Information Agents and offer real-time information to Process Agents (Inventory Agent, Purchasing Agent). The emergence of Multi-Agent Systems (MAS) may be slow to take-off unless the Semantic Web sufficiently permeates the environment for ubiquitous deployment of Agents.

Design of Agent-Based Modeling (ABM) draws clues from natural behavior of biological communities. Although it still remains a paradox, it is increasingly undeniable that simple individual behaviors of bugs like ants and wasps, collectively, may offer intelligent models of complicated overall behavior. In fact, this may have been known for centuries. One ancient observer, King Solomon, knew from his father, David, of the elaborate court organizations of oriental kings and preparations needed for military campaigns. He marveled that insects could accomplish both these tasks without any central control. Thinking of the complex systems needed to maintain the palace commissary, he wrote, “Go to the ant, consider her ways and be wise. Having no guide, overseer or ruler, she prepares her bread in summer and gathers her food at harvest time.” He knew the complexity of a military organization and was impressed that “locusts have no king, yet all of them go forth by companies.” Nearly 3000 years later, a participant in the NCMS Virtual Enterprise Workshop (1994) commented, “we used to think that bugs were the problem. Now we suspect they may be the solution!” (Parunak 1997)

Adaptability in biological systems is a fundamental characteristic of nature, and thus, models based on and inspired by such superior systems can contribute significantly to reduce key inefficiencies (and stem the loss of profit) between centralized and decentralized supply chains. Most software
is based on equations that link rates and flows (consumption, production). Variables (cost, rebates, transportation time, and out-of-stock) evaluate or integrate sets of ordinary differential equations (ODE) or partial differential equations (PDE) relating these variables. Operations research provides the framework to optimize for the “best” result. What if the “best” result is not necessarily the optimal “best” for that situation? Shortest lead time could plan a route through an area with a high probability of flash flood due to a brewing storm or threat of sniper attack on a portion of the highway. Planning software (today) fails to, or is incapable of, modeling such random events that may have profound implications for business, at that time. Thus, the “best” solution may not be adaptive to supply chain events at hand.

Even excluding random events or decisions that require integration with other models (weather, road construction), what is the half-life of ‘best’ solution in a fickle economy or high “clockspeed” industry? Compared to ABMs, a significant shortcoming of such Equation-based (ODE, PDE) models (EBM) is that EBM based software processes assume that these parameters are linear in nature and relevant data is available (for optimization). In the real world, events are non-linear, actions are discrete, information about data is distributed (CRM, PLM, SCM data silos) and data is corrupted with “noise” (according to a study by Ananth Raman of Harvard Business School and Nicole DeHoratius of the University of Chicago, for a global retailer, in some cases, 65% of SKUs (bar coded) were found to be inaccurately represented between system data, back-store and availability on store shelf, see Dehoratius, 2002).

Virtually all computer-based modeling, up to this point has used system dynamics, an EBM approach. But the struggle to adapt and respond in real-time will eventually and collectively fuel a paradigm shift that will make it imperative to model business software based both with Agents and equations. The question is no longer whether to select one or the other approach, but to establish a business-wise mix of both and develop criteria for selecting composition of software-based on one or the other approach that can offer valuable combinatorial solutions. The “balance” is subject to dynamic change (seek analogy with Screening Games). For traditionalists in supply chain management, the situation is analogous to a “push-pull” strategy where the dynamic push-pull boundary shifts with changing demand (pull).

ABM and EBM, both simulate the system by constructing a model and executing it on a computer. The differences are in the form of the model and how it is executed. In ABM, the model consists of a set of Agents that encapsulate the behaviors of the various individuals that make up the system, and execution consists of emulating these behaviors, which is essentially dynamic. In Equation-Based Modeling (EBM), the model is a set of
1. ADAPTIVE VALUE NETWORKS

Equations (pre-determined, static) and execution consists of evaluating them. Thus “simulation” is a generic term that applies to both methods, which are distinguished as Agent-based *emulation* and equation-based *evaluation*.

Thus, the need for supply chains to be adaptive should rapidly trigger demand for Agent integration with existing EBM systems. But the demand for Agents software is slow to materialize. One reason may be gleaned from the observation by Norman Poire, an economist (Figure 1-6, blue lines, http://www.smalltimes.com/document_display.cfm?document_id=2141). As shown in figure 1-6, it takes about a quarter of a century for a technology to gain acceptance. Then it fuels a period of rapid growth lasting an additional half a century. Almost after a century since “invention” or introduction, the innovation may become a commodity and grows in line with fluctuations in macroeconomic forces. We propose that Agents, in principle linked to some of the fundamentals from distributed artificial intelligence (DAI), may follow a similar trajectory which suggests increasing adoption beginning about 2005 (Figure 1-6, red line). These Agents are the types that are capable of machine learning and utilize learning algorithms, such as (ant-based) swarm intelligence, genetic algorithms, and neural networks (single and multilayer perceptions, Hopfield networks, Kohonen networks, radial basis function networks).

Continuity and autonomy of biology offer behavior patterns that are flexible, adaptive and responsive to change. Thus, the mobile, networked, autonomous, self-learning, adaptive Agent may have different principles compared to those that were developed for monolithic systems. Examination of naturally occurring Agent-based systems suggests design principles for Agents. While some circumstances may warrant deliberate exceptions, in general, Agents are aligned with the concepts listed below from Parunak (1997) and Parunak et al., (1998):

*Figure 1-6. How Conceptual Advances Lead to the Wealth of Nations*
1. Agents should correspond to “things” in the problem domain rather than to abstract functions;
2. Agents should be small in mass, time (able to forget), and scope (avoid global knowledge action);
3. Multi-Agent Systems should be decentralized (no single point of control/failure);
4. Agents should be neither homogeneous nor incompatible but diverse;
5. Agent communities should include a dissipative mechanism (entropy leak);
6. Agents should have ways of caching and sharing what they learn about their environment;
7. Agents should plan and execute concurrently rather than sequentially.

4.1 Agents versus Equations: Conceptual and Practical Considerations

The difference in representational focus between ABM and EBM has consequences for how models are modularized. EBM represents the system as a set of equations that relate observables to one another. The basic unit of the model, the equation, typically relates observables whose values are affected by the actions of multiple individuals. ABM represents the internal behavior of each individual. An Agent’s behavior may depend on observables generated by other (Agents) individuals, but does not directly access the representation of those individual behaviors, thus, maintains boundaries among individuals. This fundamental difference in model structure gives ABM a key advantage in commercial applications such as an adaptable value network where partners may interact over an e-marketplace.

First, in an ABM, each firm has its own set of Agents. An Agent’s internal behaviors are not required to be visible to the rest of the system, so firms can maintain proprietary information about their internal operations. Groups of firms can conduct joint modeling exercises (Public MarketPlaces) while keeping their individual Agents on their own computers, maintaining whatever controls are needed. Construction of EBM requires disclosure of the relationships that each firm maintains on observables so that the equations can be formulated and evaluated. Distributed execution of EBM is not impossible, but does not naturally respect commercially important boundaries (why the early wave of e-MarketPlaces failed to survive).

Second, in many cases, simulation of a system is part of a larger project whose desired outcome is a control scheme that more or less automatically regulates the behavior of the entire system. Agent systems may correspond 1-to-1 with the individuals (firms or divisions) in the system being modeled, and the behaviors are analogs of real behaviors. These characteristics make
Agents a natural locus for the application of adaptive techniques that can modify their behaviors as the Agents execute, so as to control the emergent behavior of the system. Migration from simulation model to adaptive control model is more straightforward in ABM than in EBM. One can imagine a member of adaptable business network using its simulation Agent as the basis for an automated control Agent that handles routine interactions with trading partners. It is unlikely that such a firm would submit aspects of its operation to an external “equation manager” that maintains specified relationships among observables from several firms.

EBM most naturally represents the process being analyzed as a set of flow rates and levels. ABM most naturally represents the process as a set of behaviors, which may include features difficult to represent as rates and levels, such as step-by-step processes and conditional decisions. ODEs are well-suited to represent purely physical processes. However, business processes are dominated by non-linear, discrete decision-making.

Both ABMs and EBMs can be validated at the system level by comparing model output with real system behavior. In addition, ABM’s can be validated at the individual level, since the behaviors encoded for each Agent can be compared with local observations on the actual behavior of the domain individuals. ABMs support direct experimentation. Managers playing “what-if” games with the model can think directly in terms of business processes, rather than translate them into equations relating observables. A purpose of what-if experiments is to identify improved business practices that can be implemented. If the model is expressed and modified in terms of behaviors, implementation of its recommendations is a matter of transcribing the modified behaviors of Agents into task descriptions for the underlying physical entities in the real world.

In many domains, ABM gives more realistic results than EBM, for manageable levels of representational detail. The qualification about the level of detail is important. Since PDEs are computationally complete, in principle, one can construct a set of PDEs that completely mimics the behavior of any ABM (thus produce the same results). However, the PDE model may be much too complex for reasonable manipulation and comprehension (for example what we observe in repetitive Stochastic Games with incomplete information). EBMs (like system dynamics) based on simpler formalisms than PDEs may yield less realistic results regardless of the level of detail in the representation. For example, the dynamics of traffic networks achieved more realistic results from traffic models that emulate the behaviors of individual drivers and vehicles, compared with the previous generation of models that simulate traffic as flow of a fluid through a network. The latter example bears strong similarities to the flow-and-stock approach to supply chain simulation.
Chapter 1

The disadvantages of EBM in this and other examples result largely from the use of averages of critical system variables over time and space. EBM assumes homogeneity among individuals but individuals in real systems are often highly heterogeneous. When the dynamics are non-linear, local variations from the averages can lead to significant deviations in overall system behavior (outcome). Refer back to the section on Game Theory and in light of ABM vs. EBM, re-consider the example of the Signaling Game: the choice of values (of $\xi$ and $\varepsilon$ from the distribution) can significantly impact capacity planning (inventory risk) and profit optimization (price risk). In such business applications, driven by “if-then” decisions, non-linearity is the rule. Because ABM’s are inherently local and can adapt to changes, it is beneficial to let each Agent monitor the value of system variables locally (for example, real-time data for $\varepsilon$, in the Signaling Game), without averaging over time and space.

Ant-based algorithms based on naturally occurring systems, enables the Agent to forget (ant pheromones evaporate and obsolete paths leading to depleted food sources disappear rather than misleading members of the colony). The mechanism of forgetting is an important supplement to the emphasis in conventional artificial intelligence (AI) systems on mechanisms for learning. In a discrete-event system, forgetting can be as complex as learning since both represent discrete state transitions. In a time-based system, forgetting can take place “automatically” through the attenuation of a state variable that is not explicitly reinforced. The Agents ability to “forget” is a boon to real-world adaptable business networks. EBM based demand forecasting generally uses a weighted-average of past consumption data. If there was a marked variation (for example, spike in sales, 20 weeks ago) the planning algorithm continues to consider that value because equation-based modeling cannot “forget” facts, although the weight will decrease successively in each planning cycle unless manual intervention or program insertion specifies a “forget” rule. The forecasting engine, therefore, may continue to reflect the effect in its subsequent forecast for weeks or months. Consider the cumulative error from such events, if aggregated over geographies prior to generating a global forecast that may guide procurement or production. Such events produce the Bullwhip Effect. Agents can improve forecasting and with real-time data, accuracy may be further enhanced. As a result, for example, the manufacturer may adjust production to manage inventory better and reduce waste. Reduced inventory decreases working capital charges which improves return on assets because manufacturing the cash cycle gets shorter.

In a traditional system, forecast determines production planning and subsequently, execution of the plan. Some manufacturers develop a schedule each night that optimizes manufacturing the next day, a process not much
different from grocery chains that order perishables the day before it is displayed in stores. Engineers in industries as diverse as auto, semiconductors, aerospace, and agricultural equipment will agree that a daily schedule is obsolete less than an hour after the day begins. But Agents seek to avoid the “plan then execute” mode of operation and instead responds dynamically to changes in the environment. In concurrent planning and execution, the actual time at which a job will execute may not be known until the job starts. The resource does not schedule a newly-arrived job at a fixed point in time but estimates probabilistically the job’s impact on its utilization over time, based on information from the customer about acceptable delivery times. The width of the window within which the job can be executed is incrementally reduced over time, as needed, to add other jobs (may be rated by priority, at that time) to the resource’s list of tasks. If the resource is heavily loaded, the jobs organize themselves into a linear sequence but if it is lightly loaded, the actual order in which jobs are executed is decided at the moment the resource becomes available, depending on the circumstances that exist at that time. Figure 1-7 shows simplified view of agent in system architecture.

Figure 1-7. A simplified View of Agents within the System Architecture
Chapter 1

4.2 Agents in Maintenance (US Air Force Case Study)

This example of a multi-Agent framework (and this case study) was developed by Shehory, Sycara and Sukthankar in 1999 (Agent aided aircraft maintenance) at the Carnegie-Mellon University, Pittsburgh (Shehory et al., 1999). It provides information retrieval and analysis in support of decision making for aircraft maintenance and repair for the US Air Force (USAF). Although the solution was developed for a specific type of aircraft, the Agents and interactions were designed to apply to a range of similar maintenance scenarios.

Aircraft maintenance in the USAF is performed at different levels. Basic and intermediate levels are usually performed at the base where the aircraft is deployed, whereas periodic, comprehensive maintenance is performed at special depots. Initially, mechanics inspect the aircraft for discrepancies (and may also receive such information from pilots). For each discrepancy, the mechanic consults the technical manuals for a standard repair procedure. In case such a repair procedure is found and the resources (parts) are available, the mechanic proceeds with the repair. In cases where parts are not available or they are too expensive or require too much time and additional machinery for replacement or in cases where a procedure is not provided in the technical manual, a mechanic needs to consult an expert engineer. The engineer, in turn, may consult external sources of information. These include manuals, historical maintenance data and may even include consultation with experts.

Inventory of parts is based on traditional data input from goods received. Locating spares, therefore, could be a time consuming and arduous undertaking that can be automated to a significant extent by use of automatic identification technologies (UWB, RFID) and to link inventory object related data with service/maintenance processes to offer transparency of the spares supply chain.

Until recently, no automation was introduced to the consultation processes, either, of this information-rich environment. Hard-copy repair manuals are used by mechanics and engineers. Search for relevant information may be time consuming and incomplete. Historical data (records of previous similar repairs) is scarcely used, since it is stored in paper format with no search mechanisms and usually only kept for short periods (distributed along remotely located service centers). Expert engineers may be located remotely and their advice is available by voice or fax messages, usually delayed for hours or days. All of these factors contribute to a slow, inefficient maintenance that compromises readiness.

The inspection, consultation and repair process consists of the following steps:
1. Aircraft arrives at a maintenance center, either at its home base or depot (depending on the type of maintenance required). In both cases, the maintenance procedures must be completed within a limited time period. This period varies. Basic and intermediate maintenance must be completed within hours or a few days, whereas depot maintenance may be scheduled for several weeks (depends on aircraft).

2. Mechanics inspect the aircraft and locate discrepancies. For each discrepancy a mechanic performs the following:
   a) browse the technical manual for repair procedures;
   b) in case an appropriate procedure is located, mechanic needs to verify whether it can be completed given limitations on repair time and parts availability. Mechanic may also need to consider the price of repair. For example, the technical manual may require replacing a whole wing if a crack in the wing is greater than some given threshold. This may take too long and become too expensive thereby causing delay or compromise operational activity or readiness;
   c) if the procedure found can be performed, the mechanic performs it. If not, mechanic proceeds to fill out form 202a, standard USAF form for reporting aircraft discrepancies and requesting advice. The mechanic may attach supporting information. The mechanic may consult Illustrated Part Breakdown (IPB) technical manuals and possibly other experienced mechanics. Form 202a is sent for advice and authorization for non-standard repair.

3. An engineer, upon receipt of a Form 202a, proceeds to:
   a) use experience, historical repair information and manuals to design appropriate repair;
   b) fill in a Form 202b, standard US Air Force form for discrepancy repair instructions. To this form the engineer may attach graphical illustration to clarify required repair procedure;
   c) file 202a and 202b for future use as historical repair information.

4. When a standard repair procedure is found or on receipt of Form 202b from engineer, the mechanic performs the repair as instructed. The current inspection, consultation and repair processes, as described above, have several problems. The multi-Agent system (MAS) implementation reported here attempts to address these problems. The majority of the information, both historical repair information and technical manuals, is found in hard-copy format as well as hand-written pieces. Mechanics and engineers spend precious time on:
   a) Browsing manuals and searching for historical repair information;
   b) Drawing graphical discrepancy and repair illustrations;
   c) Mechanics are idle, waiting for Form 202b to arrive from engineers in reply to their Form 202a;
d) Historical information is unused when stored remotely or local hard-copy is difficult to browse.

For information needs of mechanics, using manuals during inspection for diagnosis is inefficient and at times impossible due to the physical constraints of the inspection environment. Scribbled information both from historical forms and the current Form 202 may have limited comprehensibility. The problem intensifies due to deterioration in the quality of such information when it is transmitted via fax or photo-copied. Historical forms are kept only for two years. Time and effort spent on paperwork and filing should be used instead for diagnosis and repair. Technical manuals (IPB) are not consistently updated.

The problem consists of decision support in a physically distributed, dynamically changing environment, rich in multi-modal information, where users have diverse (varying over time) information needs. This is the type of problem for which RETSINA (REusable Task-based System of Intelligent Networked Agents) is a solution. It is a multi-Agent infrastructure that was developed for information gathering and integration from web-based sources and decision support tasks. It includes a distributed MAS organization, protocols for inter-Agent interaction as well as collaboration and a reusable set of software components for constructing Agents. Each Agent in RETSINA specializes in a special class of tasks. When Agents execute tasks or plan for task execution, they organize themselves to avoid processing bottlenecks and form teams to deal with dynamic changes in information, tasks, number of Agents and their capabilities.

In RETSINA, the Agents are distributed and execute on different machines. Based on models of users, an Agents and tasks, the Agents decide how to decompose tasks, whether to pass them to others, what information is needed at each decision point, and when to cooperate with other Agents. The Agents communicate with each other to delegate tasks, request or provide information, find information sources, filter or integrate information, negotiate to resolve inconsistencies in information and task models. The RETSINA infrastructure consists, by convention, of 3 broad types of Agents:
- Interface Agents;
- Task Agents;
- Information Agents.

In the RETSINA multi-Agent infrastructure, Interface Agents interact with users receiving specifications and delivering results. They acquire, model and utilize user preferences. The Interface Agents hide the underlying structural complexity of the Agent system. Main functions of an Interface Agent include:
Task Agents formulate and execute plans. They have knowledge of the task domain and which other Task Agents or Information Agents are relevant for performing various parts of a task. Task Agents have strategies for resolving conflicts and fusing information retrieved by Information Agents. A Task Agent:

- receives user-delegated task specifications from an Interface Agent;
- interprets the specifications and extracts problem-solving goals;
- forms plans to satisfy these goals;
- identifies information-seeking sub-goals that are present in its plans;
- decomposes plans and cooperates with appropriate Task Agents or Information Agents for planning, execution, monitoring, and results compilation.

Information agents provide intelligent access to heterogeneous collections of information sources. They have models of information resources and strategies for source selection, information access, conflict resolution, and information fusion. Information Agents can actively monitor information sources.

Middle agents collect and provide information about the location, availability, and capabilities of other Agents (possibly additional information about them). They may also serve as mediators, hiding the identities of either service requester Agents or service provider Agents or both. Middle Agents (Matchmakers) provide RETSINA-based MAS with openness. That is, Agents may leave and enter the system dynamically. When an Agent appears, it advertises itself with a Middle Agent. When it leaves, it informs the Middle Agent, as well. Agent disappearance as a result of Agent or network failure is detected by a Middle Agent via a pinging mechanism. The RETSINA internal Agent architecture is based on a multi-module, multi-thread design. It consists of two component types: functional units and data stores. Given its properties, we found the RETSINA infrastructure appropriate to solve the USAF maintenance problem. By developing and using Agent architecture, we gain the following advantages:

1. The RETSINA architecture can be used to wrap legacy software systems by equipping them with a Communicator module. Thus the resulting system remains backwardly compatible with the older systems, without restricting future software development to an obsolete model. For instance, in 1999 the Warner Robins Air Force Base (AFB) engineers were experimenting with entering some of the data into an Access
database format, as a temporary measure while waiting for (the ITL-ALC) another system to become available. With this design, separate Information Agents can easily be designed to accommodate both data sources. Since the maintenance personnel only interact with Interface Agents, they are shielded from internal data discontinuities;

2. The information required by the maintenance engineers is likely to be distributed among several computers in different geographic locations. RETSINA architecture provides built-in networking support useful for developing distributed systems, in the form of the Communicator. The Agent Name Sever (Matchmaker) allows service requesters to locate service providers. Although the current focus is on handling the repair operations described in Form 202A, which are performed locally in Warner Robins AFB, additional Agents can be added to the system to access collections of Form 00-107 (immediate repair requests), which can be filed from multiple locations. These Agents would be located on computers at the local Air Force base performing the repair and would communicate to agents at the central F-15 repair location (Warner Robins AFB);

3. The Warner-Robins Air Force Base is in a transitional phase of reorganizing their data and also training personnel. Rapid prototyping of a group of Agents are underway to address the current situation and slowly add to the “Agent population” as new information sources become available electronically. Since the Interface Agent is decoupled from the Information Agents, it is possible to replace older Information Agents without disruptions or disturbance to the users.

4.3 Agents in Manufacturing

Commercial aerospace industry makes fewer products and sells to a different set of customers than the retail industry (Figure 1-8 shows typical aerospace supply chains). Some (modular) parts and components are shared between different models (variants) of aircraft. Significant profit in this (and the automobile industry) is derived from the aftermarket sale of parts and service. The companies therefore have access to a large amount of usage data. Premature failure of two hydraulic pumps in different corners of the world prompts an Agent to explore the pattern. Both pumps came from the same manufacturing lot. The Agent prompts maintenance technicians to perform non-routine vibration analysis. Results indicated that the manufacturing lot had a defect. If vibration analyses data from manufacturer’s test results were available to the Agent in this value network, a pattern may have emerged even before a single pump failed. Comparative
analysis involves access to massive data processing that is beyond human reach in a reasonable time frame. Agents could accomplish such tasks rapidly and be able to predict, thereby avert, a potential catastrophe. The information required for such Agent operations to recognize a pattern from manufacturer data, lot information, date of installation and hours of usage are possible in value networks with integrated points of access to distributed data, but impossible in silos of supply “chains” which are common today.

*Figure 1-8. Commercial Aerospace Industry Supply Chain: Information Collection*
4.4 Future Agents at Work?

Transistor Titikaka Promethium (TTP), a small electronics retailer, starts selling a digital camera (named, CELC) and soon runs out of inventory due to the popularity of the new product. TTP places another order. A week later some customers returned the cameras and others call with questions. TTP is unable to determine the cause and loses time and revenues. Think different.

You are Must-See-Borgium Corporation, the bleeding-edge retailing behemoth. You started selling CELC and soon your return center in Moose Jaw is flooded with CELC from unhappy customers. Fortunately, your ex-VP (exiled to Timbuktu) had created a liaison with a tiny institute around Boston. She quietly integrated a system called MY-CAH that offered no satisfactory ROI to your bean counters. Within a week of mounting CELC returns to Moose Jaw, Must-See-Borgium’s MY-CAH Agent sends an e-mail alert (cc you) to N. E. Shee in Urawa (manufacturer’s headquarter) that many US customers who returned CELC to Moose Jaw also bought a certain brand of BELL notebooks with Dumb-Bell Mobile Bambino. In your in-box you also find a response from Shee-san that the camera’s software is incompatible with systems installed with Dumb-Bell Mobile Bambino without a special patch from MacroHard that can be downloaded from www.bosonic-hadrons.net and the CELC website will soon upload the link for customers. MY-CAH Agents already posted an update on the corporate website, informed Moose Jaw Center, CELC customers who registered or returned their products, sent e-mail to only those customers who bought CELC with Must-Have-Borgium credit-loyalty card and printed out an exact number of stickers (per inventory) with instructions to be affixed to CELC boxes in all your local mega-stores. You find a note of gratitude from Miss Fermionic Baryons at TTP who saw the notice about CELC on your website and could inform TTP’s customers by phone. You had no problem getting out of a mess and a bad PR wrap because MY-CAH actually works! Didn’t you vociferously object to the VP’s proposal to sponsor research at that tiny institute around Boston? Anyway, you solved the problem.

What really happened? Your store was running an Agent system that analyses data for trends. The Agent was able to identify this trend in minimal time. The missing patch could have been identified without the use of an Agent, but it would have taken much longer and resulted in many more unhappy customers, which would generate bad publicity. Why did an Agent work in this situation? Data and information derived from data is the key enabler for decision systems to be agile. In this example, the Agents were autonomously collecting product, customer, and service related information.
Customer purchases were compared for people who bought and returned this new product (SKU). How does a company know what information to collect? Easily enough, companies should collect the same information that was needed to find previous patterns if the company had data mining capabilities. In this case, real-time data over short time windows were constantly under analysis and random associations were easier to track by multi-Agent systems monitoring multiple operations both within the company and its interactions in the value network. Concurrently, it was analyzing legacy data (ERP) to learn or create analytic parameters from past data patterns.

In another scenario, consider an Agent system that operates in a services business area (only) charged with the analysis of returns. The Agent spots that the rate of return for a certain manufacturer’s products has risen above a certain level in recent weeks. Why? They are a relatively high value product, which weighs more than 15 pounds and the majority was shipped 500 miles or more. An alert from the Agent reaches the manager and she intuitively inspects the packaging and... Voila! It is different than the packaging for products that have a lower return rate. A phone call confirms the hunch that the manufacturer recently switched to a different packaging vendor in an effort to conserve costs. The Agent succeeded in creating the alert because the Agent system collects, processes, correlates and cross-references vendor data, shipping method data, shipping distance information data and other cradle-to-grave stages and any related ePC data that it can extract from the local data store connected with goods movement (RFID/UWB tags attached to this item). SKU information (only) still exists as a barcode on the outer packaging. The Agent also extracts the UPC code from the store master data (redundant information). If packaging type information was stored on RFID/UWB tags for each SKU sold, the Agent system may have been able to spot the trend without the aid of a human, the manager (Figure 1-9).

Agents can also help with marketing. Dell sells computers that consumers can configure. Bundling is a marketing technique that pairs two products together to sell at a single price, which is lower than the normal price of the two if sold individually. Single price gives a greater revenue and profit than if either item were sold alone. Dell stores exabytes of information on customer buying patterns. An “analytic” Agent is able to spot a pattern where 40% of customers who buy extra memory also buy a certain high-speed processor. A “marketing” Agent can “talk” to the “pricing” Agent to offer discounts if memory is bought together with the processor. As the trend of choices for combinations (memory vs. processor speed) changes or differs in demographics or geographies, the data from “analytic” Agent can be used for the “marketing” and “pricing” agent to adapt and offer new bundling
options (dynamic pricing). This can augment demand for the memory and increase total revenue and profit. Customers who are likely to buy a product may be targeted for marketing (may not buy without bundle discount).

The number of potential product combinations increases if three or more options are thrown into the mix, not to mention accessories like cameras, MP3 players or printers. It is simple for Agents to analyze gargantuan amounts of data and spot potential (multiple) bundling opportunities and adapt to the fluctuations in demand in near real-time much faster (by several orders of magnitude) than a human or software based on equation (EBM). Bundling strategies can be catalytic to sell slow moving inventory or end of life (EOL) product prior to new product introduction.

![Diagram of retail industry agents]

*Figure 1-9. Agents in Retail Industry (also shows where “returns” and “bundling” Agents may integrate)*
When I want to go out to the movies, rather than read reviews, I ask my sister-in-law. We all have an equivalent who is both an expert on movies and an expert on us. What we need to build is a digital sister-in-law (“Less Is More: Interface Agents as Digital Butlers” by Nicholas Negroponte, 1994).

4.5 WHY THINK DIFFERENTLY?

The approach to system design and management with Agents in the software landscape is at odds with the centralized top-down tradition in systems engineering. The question usually arises in terms of the contrast between local and global optimization. Decision-makers fear that by turning control of a system over to locally autonomous Agents without a central decision-making body, they will lose value that could have been captured by an integrated (enterprise) global approach.

Benefits of Agent-based architecture versus centralized ones are conditional, not absolute. In a stable environment, a centralized approach can be optimized to out-perform the efforts of an opportunistic distributed system of Agents. If the system has appropriate learning capabilities, it will eventually become as efficient. The appropriate comparison for systems designers of enterprise software is not between local and global optima but between static versus adaptable systems. Thus, evaluate the competing options (in any particular case) theoretically, strategically, tactically and practically.

Theoretically, there are decentralized mechanisms that can achieve global coordination. For example, economists have long studied how local decisions can yield globally reasonable effects. Recently these insights have been applied to a number of domains that were not traditionally considered as economic, such as network management, manufacturing scheduling and pollution control.

Strategically, managers must weigh the value of a system that is robust under continual change against one that can achieve a theoretical optimum in a steady-state equilibrium (that may never be realized). A company that anticipates a stable environment may well choose centralized optimization. One that also incorporates Agent-based software does so because it cannot afford to be taken by surprise.

Tactically, the life-cycle software costs may be lower for Agent-based systems than for centralized enterprise software. Agents can be modified and maintained individually at a fraction of the cost of opening up a complex enterprise software system. In systems that must be modified frequently, losses due to sub-optimal performance can be recovered in reduced system maintenance expenses.
Practically, Agent-based systems that follow these principles have been piloted or deployed operation (US Air Force case study by CMU). The Agents reflect the principles outlined rather than those of centralized systems. Growing acceptance of Agents in competitive business environments may be evidence of the benefit they bring to their adopters (Figure 1-10)

Figure 1-10. P&G’s Agent-enabled Supply Network in 2008

5. AUTOMATIC IDENTIFICATION TECHNOLOGIES

Automatic identification technologies offer tools to acquire data about objects (e.g., IV pumps, toothpaste, and ammunition). Innovation and leadership lies in the effective use of the data, not in its acquisition.

In 1894, Oliver Lodge demonstrated how to communicate (data) using radio waves. Half a century later, with the discovery of the RADAR at MIT, it was likely that the natural frequency spectrum was going to “make waves”
for quite some time. Near the tail end of the last century, with the establishment of the MIT Auto ID Center (which morphed into Auto ID Center), once again, more than a century later, a radio frequency-based identification (RFID) and communication protocol created waves whose impact will be inescapable in the future and for the future of most businesses that were present in the past.

Neither the technology nor concepts are new but the two thinker-founders of the MIT Auto ID Center (Sarma and Brock, 1998) created a “storm in a tea cup” by reversing the conventional thinking (kilobytes of data on RFID tags) in their proposed standardization of a format for minimal data on RFID tags, referred to as electronic product code (ePC) that will serve only as a reference to a physical object, data about which may be stored on the internet (Figure 1-11). The generic organization of ePC was to extend the Universal Product Code (UPC) format currently used in bar codes. Thus, ePC was re-using an ‘old bag of tricks’ yet ‘disruptive’ to the status quo since the business of RFID usage had been around for half of the 20th century. A ‘killer’ ePC application may be a simple way to connect bits (information) with atoms (physical objects) in a manner that may make it feasible for widespread business adoption by offering low cost tags and use of the internet as the ‘data’ store. Low cost passive tags suffer from some limitations (signals absorbed by metal, such as beverage cans) which can be circumvented by a combinatorial approach to include emerging technologies, such as the active ultra wideband (UWB) tags. UWB tags transmit data at distances 30-300 meters using low power levels and the signals can penetrate metal barriers as well as concrete walls.

![Figure 1-11. Evolution of ePC](image-url)
Chapter 1

The 96 bit electronic Product Code (ePC) as proposed by the Auto ID Center, is made up of Header, ePC manager (manufacturer’s information, also in bar code), object class (product category similar to bar code) and serial number space that is expected to be unique for each unit, such as an individual can of Coke. The ePC manager is defined by 28 bits that can uniquely represent more than 268 million companies. Similarly 16 million different product classes (object) can be defined by 24 bits. Coke Classic and Diet Coke belong to 2 different object classes. The 36 bit serial number space refers to the maximum number of individual items in a specific product class that may be assigned a unique number. Thus, more than 68 billion individual Coke Classic cans may be individually identified if each can had a RFID tag encoded with ePC (Figure 1-12). In 2000, The Coca Cola Corporation, the largest bottler, sold 3.8 billion ‘unit cases’ each containing 192 ounces. About 42%, or 1.6 billion, ‘unit cases’ were Coke Classic (19.2 billion individual cans, assuming that all Coke was sold in 16 ounce cans). If each 16 oz. can had a unique identifier (19.2 billion cans per year), even then the ePC serial number space, as defined by the Auto ID Center, will accommodate individual numbering of Coke Classic cans for many years!

If the company made a sensible business process decision that the granularity of information at the level of each can was unnecessary, it could still track and trace ‘unit cases’ affixed with RFID tags encoded with ePC. If we use 2000 sales figures for Coke Classic (1.6 billion unit cases), the ePC serial number space will accommodate unique numbering of each Coke Classic ‘unit case’ for about 40 years. The 96 bit ePC serial number space will be sufficient for nearly a century for Perrier, the French bottling plant that produces 3 million bottles of Perrier per day.

Figure 1-12. 96 bit Electronic Product Code (ePC) proposed by the Auto ID Center
But, these are only tools which may act as catalysts if thought leaders develop a vision to use this rich yet raw data. Businesses may manage uncertainty, reduce inefficiencies, and information asymmetry if corporate leaders are capable of utilizing real-time data to stimulate business process innovation. Can real-time information compress time between supply and demand? Auto identification technologies, as enablers of data acquisition about objects, to be valuable, must feed real-time information to update processes (maintenance, cross-docking) or decision systems (planning, execution) to trigger adaptive response(s). As an extension of adaptive decision support capabilities, real-time data can offer ‘transparency’ if pervasive and accessible via distributed data infrastructure among the value network. Transparency may be the key to further catalyze the practice of supply chain management to evolve toward an adaptive value network. Point A to B data visibility may augment a few operations and offer savings, but is far from the supply chain profits from real-time adaptability through pervasive real-time data (RFID) usage.

The impact of pervasive RFID (or UWB) deployment will create an avalanche of data, but can we extract the information from this data that will be valuable for business transactions? (Figure 1-13) In US alone, there are 1.5 million retail outlets, 160,000 grocery store chains, 400,000 factories and 115 million homes. The US consumer packaged goods (CPG) industry produces 1 billion items per year. If we read each item 10 times (in the supply chain) it translates to 300,000 “reads” per second. At 100 bytes to store each ‘read/event’ data, we will be faced with 1000 terabytes of static data storage each year, from CPG operation alone. The road to ubiquitous tagging of objects will dwarf the current internet that now holds about 1 billion web pages with only 10 petabytes of data. Current (year 2003) estimates suggest that we generate about 1 terabyte of data per second. The future requires a radically different mechanism of data and information handling, through Agents and use of semantic tags, to make sense of it all.

*Figure 1-13. The Ultra High Frequency (UHF) Range for which an ePC Standard is Now Available*
Given the potential impact, the ‘RFID’ market is, naturally, in quagmire, in part, spawned by unrealistic claims by some proponents of RFID who are focused on cost. Others discuss supply chain but understand less of its implication in terms of transparency in a value network. Still other contributors include vendors pushing products and consultants pushing services to offer you awe-inspiring ROI. Both groups want to “make hay while the sun shines.” Nay-sayers (with other commercial interests) are eager to point out shaky ROI because their methods still cannot prove the value. Another component has emerged in the form of individuals or groups (in search of media attention) who are quite vociferous about privacy of information but offers little substance to explain what constitutes violation of privacy in the context of an ePC alphanumeric string serving as a reference for Jiffy Peanut Butter or Wrigley’s Chewing Gum.

More than 5 billion bar codes are scanned each day, worldwide, but its potential for ubiquity may be cut short by the up-start ePC, but not anytime very soon. The inventors of the first linear bar code system were, naturally, decades ahead of their time. Bernard Silver and Norman Joseph Woodland applied to patent the system in 1949 and their patent was granted in 1952. Both were graduate students at the then Drexel Institute of Technology in Philadelphia and the idea was triggered by overhearing a conversation in 1948 between the President of a grocery chain store imploring the Dean at DIT to develop an automated checkout system. Woodland took a job at IBM after graduation but IBM expressed limited interest in this work for bar codes. Disappointed, the duo sold their patent to Philco. Bernard Silver died in 1962. In the late 1960’s when their patent expired, several new technologies converged to make product scanning commercially feasible. In 1970, ten grocery companies formed a committee to choose a standard for encoding product data (the present day universal product code, UPC, the predecessor to ePC). By then IBM wanted “in” on the action and brought in Norman Woodland, still an employee at IBM, to help launch the bar code research effort. In 1973, Woodland’s leadership may have persuaded the standards committee to choose IBM’s symbol over six other competitors. On 26 June 1974, in a Marsh Supermarket in Troy, Ohio (USA) a package of Wrigley’s Chewing Gum was the first item scanned using the (universal product code) bar code (Scanlon, 2003).

“ In contrast, at highly successful firms such as McKinsey and Company [...] Hundreds of new MBAs join the firm every year and almost as many leave. But the company is able to crank out high-quality work year after year because its core capabilities are rooted in its processes and values rather than in its resources (vision). I sense, however, that these
The capabilities of McKinsey also constitute its disabilities. The rigorously analytical, data-driven processes that help it create value for its clients in existing, relatively stable markets render it much less capable [...] in technology markets.” (Christensen, 2000).

Given the volume (some of dubious quality) of information already available on every facet of RFID and its applications in various industries, it is not necessary to add any technology or application review in this article. Our view of RFID deployment from a process perspective includes, albeit in stages, gradual integration with Agents in the system, for possible transition from real-time to adaptive to predictive states within the value network. The following figure outlines this convergence (Figure 1-14).

![Figure 1-14. Convergence: Near Real-time Predictive Model](image)

The version of the above illustration that may be forwarded as the “real-time model” can be viewed by removing the Agents and the ‘Pull’ signal from the above figure. Similarly, the “real-time adaptive model” may be visualized by excluding the “Pull” signal but including Agents. In general, the ability to identify any object in real-time (without errors and manual data entry), offers data that may be sieved through “intelligent” middleware to improve or adapt processes. High level or aggregated information and/or learnings may enable precision planning in the decision layers. Prior to execution, decision systems will be able to optimize how many objects may be distributed, displayed or destroyed. The ability of Agents to monitor and process vast surges of data in near real-time will enhance the adaptive
abilities implied in the model. However, what the customer “wants” to buy still remains the predominant market uncertainty, ε in the Signaling Game (see section on Game Theory and Operations Research). Are there mechanisms or innovative strategies that can “extract” this future demand signal to move the push-pull boundary? Actual “pull” data that is verifiably robust (value of ε) is at the core of the ‘predictive’ model since such customer “pull” data for future demand may be one pivotal factor in reducing supply chain inefficiencies by taming the Bullwhip Effect (Figure 1-15).

![Figure 1-15. Bullwhip Effect after RFID? (Source: Yogesh V. Joshi, 2000. MIT Thesis)](Figure 1-15.

The immense diversity of the “end” consumer makes it impossible to suggest any general mechanism. As an example, consider super-market type retailers who sell both dry goods and perishables. Retailers operating in single digit profit margins dream about improving accuracy of demand forecasting. Consider a down-to-earth scenario where a family of four living on San Silvestro in Venezia does not own an internet linked, Agent impregnated, ePC reader enabled refrigerator (from Being Digital Inc).

Instead, this family has a note pad on the refrigerator door. If Kathleen is using all the pesto, she writes Pesto (Butoni) on the super market shopping list, which keeps growing since the last shopping trip to Tesco. Charles wants fresh bananas and adds it to the list. Colin, manager of the Albertson’s Super Store, due to open next week near the Rialto Bridge, visits you. He is engaging and talks about his last job in Garden City. As a part of Albertson’s marketing campaign, Colin offers you a sleek tablet PC-like personal digital assistant (PDA). You are struck by the logo of Carleton urging all of us to
“invent” and it inspires you to think different. Colin explains that Albertson’s has teamed up with Moore Inc who bought Boingo Wireless from Sky Dayton. Colin is very convincing and you realize that this is not “a pie in the sky” scheme. You just may be on the road when the future arrives.

The PDA is wireless internet accessible. You can use it at a T-Mobile “Hot Spot” such as one in the McDonald’s in San Marco. However, Colin would like you to use the magnetic holder of the PDA and slip it on the refrigerator door. Every time Matt is close to emptying the shampoo or Elaine finishes the Barilla tortellini, they should add these to the shopping list, as usual, but on the PDA with the sensor pen. What’s that to Albertson’s? Well, if you wrote down Barilla Pasta and bought Barilla Pasta the next time you shopped at Albertson’s with your Club Card, you shall receive a 2% discount, which also applies to all the items you scribbled on the PDA, if you actually bought those items at the store. What happens if you shopped online at Albertson’s virtual store, A_Pea_in_the_Pod.com? Colin explains that the PDA is still going to save you money. If you can plan ahead such that you can wait 24 hours before home delivery, then you get a discount. If you wait 48 hours, you receive 2.5% off your total bill. What if I can wait for 5 days? Colin explains that any wait longer than 48 hours is rewarded with a massive discount of 3%. But if you did go to the store with your PDA, it will wirelessly guide you to find things on your list and offer other tips or alert you to manufacturers or competitors e-coupons for things on your list. The first 100 people to sign up for Albertson’s offer also gets an autographed copy of the book of poetry "Moy Sand and Gravel," by the Pulitzer Prize winning author Paul Muldoon of Princeton University. Kathleen loves “Daffodils” and you want “in” on the action. Does it matter if Albertson’s gets to know today what I want tomorrow?

Convergence of falling prices on PDAs, low cost wireless/wired access and some “intelligent” software is the infrastructure a retailer may need to capture the “pull” demand directly from some customers, as illustrated in the near real-time predictive model. Can this data from customers reduce your waste of perishables by 10% or adapt forecasting to reduce your purchasing capital by 1%? Real-time POS data from RFID tagged (ePC encoded) objects and the data flow from customers’ pre-shopping lists may be combined for accurate forecasting and planning, particularly in procurement of perishables with short half-lives. In case of the latter, a final purchase order is sent only 36-24 hours prior to expected store delivery from producers (farmers, poultry, dairy). You can model the metrics in this scenario and claim that there may not be sufficient ROI to justify investment in this “pull” signal. How do you model the behavior of customers, say, in an area where more than 50% of the adults are internet users?
In 1959, GE recruited the reputable consulting firm of Arthur D. Little Inc. in Boston to conduct a study to determine whether there was a market for portable TV sets that GE could now build using solid state transistors. Several months later in 1959, after spending a staggering amount of money (millions) in focus groups and discussions, Arthur D. Little Inc. sent their analysis to GE suggesting that they do not believe there is any market for such TV sets. GE management pushed aside the project proposed by its engineers. Just before Christmas in 1959, Sony introduced a small B&W television in the US market. Sony sold more than 4 million television sets within months (Tellis and Golder, 1996).

5.1 ULTRAWIDEBAND: THE NEXT GENERATION RFID?

Instead of the customer’s pre-shopping list in the retail scenario, what if that list was for spare parts at the Warner Robbins US Air Force Base (Agents case study) or US Army Aviation and Missile Command in Huntsville, Alabama? Can MRO (maintenance, repair and overhaul) improve its efficiency if the mechanics had visibility of the inventories of approved spare parts? In these and several other scenarios, it is likely that the benefits of using active ultrawideband tags will exceed low cost RFID tag usage. Only a brief overview of UWB is provided below since there is a mountain of original work in this area, especially from Dr. Gerry Ross and Dr. Robert Fontana (www.aetherwire.com/CDROM/Welcome.html).

The origin of ultra wideband technology stems from work in time-domain electromagnetics that began in 1962. At the Sperry Research Center, then part of the Sperry Rand Corporation, Dr. Gerry Ross, the father of baseband technology, applied these techniques to various applications in radar and communications. The experimental phases of these studies were aided by the development of the sampling oscilloscope by Dr. Bernard Oliver of the Hewlett-Packard Corporation (1962). In April 1973, Sperry Research Center was awarded the first UWB communications patent, due to Dr. Gerry Ross. Through the late 1980's, this technology was alternately referred to as baseband, carrier-free or impulse. The term "ultra wideband" was first applied by the US Department of Defense in 1989. By 1989, Sperry Research Center had been awarded over 50 patents including UWB applications such as communications, radar, collision avoidance, positioning systems, liquid level sensing and altimetry.

One recent application of UWB communications technology is the development of highly mobile, multi-node, ad hoc wireless communications networks for the US Department of Defense. The system is designed to be secure with low probability of intercept and detection. UWB ad hoc wireless
network supports encrypted voice/data (128 kbps) and high-speed video (1.544 mbps). A parallel effort, funded by the Office of Naval Research, under the Dual Use Science and Technology (DUST) program is developing a state-of-the-art, mobile *ad hoc* network (MANET) based upon Internet Protocol (IP) suite to provide a connectionless, multihop, packet switching solution for survivable communications in a high link failure environment. The thrust of DUST is toward commercialization of UWB technology for applications to high-speed (>20 mbps) wireless applications for the home office. The Hummingbird collision avoidance UWB sensor (originated from a US Marine Corps project) was created for an electronic license plate commissioned by the US National Academy of Science (Transportation Research Board). The UWB Electronic License Plate provides a dual function capability for both automobile collision avoidance and (RF) tagging for vehicle to roadside communications.

Therefore, UWB usage in tagging is a proven technology. A comparative analysis of RFID versus UWB shows that UHF RFID has a spatial capacity of 1 kbpspm$^2$, ([grouper.ieee.org/groups/802](grouper.ieee.org/groups/802)). Spatial capacity of UWB is 1000 kbpspm$^2$ or 1000-fold more than RFID. Spatial capacity focuses not only on bit rates for data transfer but on bit rates available in confined spaces (retail stores) defined by short transmission ranges. Measured in bits per second per square meter, spatial capacity is a gauge of "data intensity" that is analogous to the way lumens per square meter determine illumination intensity. Growing demand for greater wireless data capacity and crowding of regulated radio frequency may increasingly favor usage of spectrum that will offer appreciable bit rates that will function despite noise, multi-path interference and corruption when concentrated in smaller physical areas (grocery stores and warehouses). Spatial capacity limits may clog (like cholesterol in arteriosclerosis) "interrogation" systems when and if item level tags are a reality and readers in smart shelves continually emit electromagnetic signals to solicit tag data from objects. Part of this reasoning is evident in independent efforts by Hitachi and Sony who are exploring BlueTooth options with spatial capacity of 30 kbpspm$^2$ and others in asset management (Rockwell Automation) are exploring 802.11a compliant 5GHz with spatial capacity of 55 kbpspm$^2$ (spare parts). Unfortunately, 802.11a is non-compliant with 802.11b but 802.11g is compliant with both (.11a and b).

In the past couple decades, several companies have engaged in commercializing UWB. As implied by its name, UWB spans several gigahertz of spectrum at low power levels below the noise floor of existing signaling environment. Conventional narrow band technology relies on a base "carrier" wave modulated to embody a coded bit stream. Carrier waves are modified to incorporate digital data through amplitude, frequency or phase shift key modulation. These mechanisms are, therefore, susceptible to
interference and the coded bit stream (for example, electronic product code or ePC) may be decoded/intercepted. UWB wireless technology uses no underlying carrier wave (hence secure military use) but modulates individual pulses either as a bipolar modulation or amplitude modulation or pulse-position modulation, where it sends identical pulses but alters the transmission timing. UWB offers narrower pulse time (300 picoseconds) and covers a broader bandwidth extending up to several gigahertz. Because UWB operates in picosecond bursts, power requirement is as low as 200 mW (compare 802.11b at 500 mW or 802.11a at 2000 mW). High data rate (0.1 to 1.0 gbps\(^2\)) for UWB compares poorly with 802.11b (0.01 gbps\(^2\)) or 802.11g (0.05 gbps\(^2\)). Thus, UWB is used for wireless transmission of data, video as well as networked games, toys and appliances. There are robotic vacuum cleaners (from iRobots) and lawn mowers that may clean the living room or manicure the garden without touching the sofa or grazing the rose bush. Universal appeal for UWB is enhanced by its capability to accommodate several standards (ePC, GTAG). Without spectrum restrictions specific to country or region, UWB may become a global wireless medium.

After the events of 11 September 2001, UWB transmitters (like RFID readers) were mounted on robots for search missions at the World Trade Center. UWB is not hindered by metal or layers of concrete. On 14 February 2002, the FCC gave qualified approval to use UWB (www.fcc.gov/e-file/ecfs.html) in the range >960 MHz, 3.1-10.6 GHz and 22-29 GHz. Active UWB tags cost $1-$10 while the transmitters may be cheaper than RFID readers because they do not need many analog components to fix, send and receive on specific frequencies. However, software defined radio (SDR) based readers may soon arrive. UWB is not without its critics. Dispute stems
from claims that UWB transmissions could interfere with spectrum used by GPS, cell phones and air traffic control. FCC is investigating, but plans to open up more spectrum for UWB commercial applications. Without the burden of fees for spectrum usage, the commercial floodgates for UWB usage may be unstoppable. Telecommunication giants who rushed to buy spectrum seduced by the future of 3G services are fighting to keep UWB off the news after investing billions in auctions to buy spectrum. Perhaps worse affected are the GSM sponsors in EU and USA. UWB is a tool for data acquisition (healthcare, hazardous chemicals) and thus a contributor to the future of adaptive value networks. An added value is its dual ability to provide data about objects when tagged to objects and form a wireless network to upload the data (over distances of 30-300 meters through metal and/or concrete) to the data infrastructure in much the same way that WiFi (802.11b) wireless networks may be used to upload RFID data in warehouses, stores or hospitals. Figure 1-17 shows plot of data rate and range capabilities of UWB.

![Data Rate and Range Capabilities of UWB](www.multispectral.com/pdf/APPsVGs.pdf)

6. SENSOR NETWORKS

Wireless sensor networks may be the first example of pervasive computing. Its applications extend from sensing blood pressure in arteries and transmitting them to a patient monitoring device to suggesting trends of warehouse shelf occupancy to a plethora of uses in the security industry.
Sensors do not transmit identification data or ePC. Sensor data models cannot be used in the same manner as data models from UWB or RFID. Sensors are self-powered and form wireless ad hoc networks that upload through specific nodes which may be connected to data stores or the internet (Figure 1-18). However, each sensor has certain analytical abilities and due to in-network processing, the sensor network transmits analyses of the data rather than the raw bits of data to provide “answers” rather than “numbers” to the system. Embedded sensors are most likely to influence various forms of supply chain-related functions. For example, sensors attached to spindles in drilling machines may continually upload the status of the spindle such that it is serviced or replaced within a reasonable time to avoid breakdown of the machine and systemic downtime. Metrics like meantime between failure (MTBF) and other parameters may be helpful to determine when the service may be scheduled. Sensor data may require different thinking in terms of “adaptive flow” databases where the data (or analyses from sensor nodes) stream through the database where the query is stored. For example, embedded light emitting sensor network in a secure room sends positive light emission data on which the query (is anybody entering the room) need not act. Only when an obstruction causes a break in the ad hoc network or occludes the light signal from a sensor or group of sensors, then the query comes into effect. Service supply chains (such as heating, cooling companies) may benefit from sensor-based information to pre-dispatch technicians to stem problems before they reach break-points or require emergency attention. The key is to try to understand how to integrate sensor data to benefit supply chains functions. With the flood of nanosensors soon to arrive, the involvement of Agents may be absolutely imperative to harvest the benefits by extracting intelligence from such data.
7. THE SEMANTIC WEB (IS SPREADING)

The average user will never see this web but the buzz about the Semantic Web is as intense as the internet itself. Semantic metadata will let you do things with meaning. The massive amounts of data that we are likely to experience will be useless unless meaningful correlations and connections help us drive innovations, the profitable ones. But just because it is hidden from view does not mean that you can bypass the evolution of the Semantic Web, although it is intended for computers to improve searches, viewing data, interacting with services and sharing information. Taken together, it can offer process transparency across language and geographic boundaries to connect partners in a value network even if individual partners define or perform certain functions differently from others.

Tim Berners-Lee of MIT, the creator of the world wide web as we know it today (while at CERN, Geneva), had described the Semantic Web concepts perhaps as early as 1995 and certainly more clearly by 1998. Since that time, Tim Berners-Lee’s vision has matured and significant progress has taken place in research communities around the world to demonstrate that semantic web can solve a variety of today’s business problems. Semantics is a collection of Resource Description Framework (RDF) data (or any other semantic language) which describes the meaning of data through links to ontologies, which act as global decentralized vocabularies. In philosophy,
ontology is a theory about the nature of existence (of what types of things exist). Ontology as a discipline studies such theories. Artificial intelligence and semantic web researchers have co-opted the term to indicate a document or file that formally defines the relations among terms. Computers, empowered with this metadata, will be far more “meaningful” in their understanding of the data without human intervention provided data is in machine readable format.

Michael Dertouzos of MIT and James Hendler of the University of Maryland have authored books and articles which are excellent resources to understanding the concepts of semantic web (Dertouzos 2002). Human language thrives when using the same term to mean somewhat different things, but automation does not. The authors provide this example: Imagine that I hire a clown messenger to deliver balloons to my customers on their birthdays. Unfortunately, the service transfers the addresses from my database to its database, not knowing that the "addresses" in mine are where bills are sent and that many of them are post office boxes. My hired clowns end up entertaining a number of postal workers, not necessarily a bad thing, but certainly not the intention. An address that is a mailing address can be distinguished from one that is a street address and both can be distinguished from an address that is a speech, with the tools from the Semantic Web.

This is not the end of the story, because two databases may use different identifiers for what is, in fact, the same concept, such as zip code. A program that wants to compare or combine information across the two databases has to know that these two terms are being used to mean the same thing. Ideally, the program needs to discover such common meanings for whatever databases it encounters. For example, an address may be defined as a type of location and city codes may be defined to apply only to locations. Classes, subclasses and relations among entities are a very powerful tool for web use. We can express a large number of relations among entities by assigning properties to classes and allowing subclasses to inherit such properties. If city codes must be of type city and cities generally have web sites, we can discuss the web site associated with a city code even if no database links a city code directly to a web site.

Inference rules in ontologies supply further power. Ontology may express the rule "if a city code is associated with a state code, and an address uses that city code, then that address has the associated state code." A program could then readily deduce, for instance, that a Cornell University address, being in Ithaca, must be in New York State, which is in the US and therefore should be formatted to US standards. The computer doesn't truly "understand" any of this information, but it can now manipulate the terms much more effectively in ways that are useful and meaningful to the human user.
The real power of the Semantic Web will be realized when Agents collect web content from diverse sources (stock quotes from Bloomberg), process the information (in relation to your business) and exchange the results with other programs or data (demographic data from the US Census Bureau). The effectiveness of such Agents will increase exponentially as more machine-readable web content and automated information services (such as, real time-data) become available. The Semantic Web promotes the synergy between Agents that were not expressly designed to work together but can now transfer data among themselves if data comes with semantics (which levels the playing field in terms of the meaning of data, such as, your purchase order is the supplier’s sales order).

With ontology pages on the Web, solutions to terminology (and other) problems begin to emerge. The meaning of terms or XML codes used on a web page can be defined by pointers from the page to ontology. Of course, the same problems as before now arise if you point to an ontology that defines *addresses* as containing a *zip code* and one that uses *postal code*. This kind of confusion can be resolved if ontologies (or other web services) provide equivalence relations: one or both of our ontologies may contain the information that a zip code is equivalent to a postal code.

The scheme for sending in the clowns to entertain customers is partially solved when the two databases point to different definitions of *address*. The program, using distinct URIs (universal resource indicators) for different concepts of address, will not confuse them and in fact will need to discover that the concepts are related at all. The program could then use a service that takes a list of *postal addresses* (defined in the first ontology) and converts it into a list of physical *addresses* (the second ontology) by recognizing and removing post office boxes and other unsuitable addresses. The structure and semantics provided by ontologies makes it easier to provide such a service and can make its use completely transparent.

Ontologies can enhance the functioning of the web in many ways. They can be used in a simple fashion to improve the accuracy of web searches. Advanced applications will use ontologies to relate the information on a page to the associated knowledge structures and inference rules. An example of a page marked up for such use is www.cs.umd.edu/~hendler. If you send your Web browser to that page, you will see the normal web page entitled "Dr. James A. Hendler." As a human, you can readily find the link to a short biographical note and read there that James Hendler received his PhD from Brown University. A computer program trying to find such information, however, would have to be very complex to guess that this information might be in a biography and to understand the English.

For computers, the page is linked to an ontology page that defines information about computer science departments. For instance, professors
work at universities and they generally have doctorates. Further markup on the page (not displayed by the typical web browser) uses the ontology's concepts to specify that James Hendler received his PhD from the entity described at the URI http://www.brown.edu (the web page for Brown University in Rhode Island). Computers can also find that James Hendler is a member of a particular research project, has a particular e-mail address. All that information is readily processed by a computer and may be used to answer queries (where did Dr. Hendler receive his degree?) that currently would require a human to sift through the content turned up by a search engine.

In addition, this markup makes it much easier to develop programs that can tackle complicated questions whose answers do not reside on a single Web page. Suppose you wish to find the Miss Cook you met at a trade conference last year. You do not remember her first name, but you remember that she worked for one of your clients and that her son was a student at your alma mater. An intelligent search program can sift through all the pages of people whose name is "Cook" (sidestepping all the pages relating to cooks, cooking, the Cook Islands and so forth), find the ones that mention working for a company that's on your list of clients and follow links to Web pages of their children to track down if any are in school at the right place.

An important facet of (Agent) functioning will be exchange of "proofs" written in the Semantic Web's unifying language using rules and information such as those specified by ontologies. For example, suppose Miss Cook's contact information was located by an online service which places her in Johannesburg. Naturally, you want to check this, so your computer asks the service for a proof of its answer, which it promptly provides by translating its internal reasoning into the Semantic Web's unifying language. An inference engine in your computer readily verifies that this Miss Cook indeed matches the one you were seeking and it can show you the relevant Web pages if you still have doubts. Although they are still far from plumbing the depths of the Semantic Web's potential, some programs can already exchange proofs in this way, using the preliminary versions of the unifying language. Figure 1-20 shows Tim Berners-Lee’s Semantic Web layers.
Many automated web services already exist commercially without semantics and their claims may be doubtful. Even if these services had Agents, at present Agents have no way to locate a service that will perform a specific function. This process, called service discovery, can happen only when there is a common language to describe a service in a way that lets other Agents "understand" both the function offered and how to take advantage of it. Services and Agents can advertise their function by, for example, depositing such descriptions in directories analogous to the Yellow Pages. Some low-level service-discovery schemes are currently available, such as Microsoft's Universal Plug and Play, which focuses on connecting different types of devices. Sun Microsystems's Jini aims to connect services. These initiatives, however, attack the problem at a structural or syntactic level and rely heavily on standardization of a predetermined set of functionality descriptions. Standardization can only go so far because we cannot anticipate all possible future needs.

The Semantic Web, in contrast, is more flexible. The consumer and producer Agents can reach a shared understanding by exchanging ontologies, which provide the vocabulary needed for discussion. Agents can even "bootstrap" (learn) new reasoning capabilities when they discover new ontologies. Semantics also makes it easier to take advantage of a service that only partially matches a request. A typical process will involve the creation of a "value chain" in which sub-assemblies of information are passed from one Agent to another, each one "adding value" to construct the final product requested by the end user. To create complicated value chains automatically on demand, Agents will increasingly exploit more and more artificial
intelligence technologies in addition to the Semantic Web. But the Semantic Web will provide the foundations and the framework to make such technologies more feasible.

8. CONCLUSION

Scientists use models to represent the basic nature of the universe. Businesses use models to optimize profits, products and services. Models may even predict future action. But, as ubiquitous as models are, they are, for the most part, isolated from one another. In other words, a model from one domain, such as weather forecasting, does not interact with another, such as purchasing or customer behavior. Can we harness the power of multiple individual data models into larger aggregates? What if we could make predictions based on not a few parameters in an equation based model but billions of diverse facts and functions that Agent based models might be able to accommodate? The latter may result in an unprecedented increase in productivity through the optimal use of resources, ability to adapt and prepare for change according to prediction. We may dramatically reduce the cost of goods and services through the elimination of inefficiencies.

To build these models, individually and then test them in combinations may be a worthy endeavor for generations of engineering and business students, supported by businesses. However, the business community may wish to embrace the key elements (tools and technologies) mentioned in this paper and seek ways to bring about the convergence, repeatedly mentioned throughout this article. Principles from Game Theory empowered by real-time data from automatic id technologies may enhance your profit optimization. Reducing information asymmetry with partners in your value chain through secure Agents-based systems may exponentially eliminate inefficiencies. Deriving more meaning from data through the Semantic Web will allow you to enhance inter-operability between diverse environments of the partners in a value network. Convergence will determine, in part, the pace of your ability to adapt.

Translating convergence to create a merger between bits and atoms is an evolution and is underway. The ability to use it in your business processes to innovate or invent is only limited by your imagination. You cannot visualize the future if your imagination is out of focus.
The payoff from information technology is in making transactions and processes more effective and efficient, it’s not about creating a new economy or creating new models of industry. It is about taking a tool, powerful tool, and saying, “How can I make my supply chain more effective and efficient?” (Lou Gerstner, CEO, IBM, The New York Times, 10 March 2002).

ACKNOWLEDGEMENTS

For discussions, information and material in this article, the author wishes to gratefully acknowledge the contributors and help from:

Benson Adams, US Army
David Brock, MIT
Gerard Cachon, University of Pennsylvania
Nicole DeHoratius, University of Chicago
H. van Dyke Parunak, Altarum
Charles Fine, MIT
Gerry Frizelle, University of Cambridge
Neil Gershenfeld, MIT
Henry Lieberman, MIT
Grace Lin, IBM
Serguei Netessine, University of Pennsylvania
Ozalp Ozer, Stanford University
Greg Parlir, University of Alabama
Ananth Raman, Harvard Business School
Sanjay Sarma, MIT
David Simchi-Levi, MIT
Katia Sycara, Carnegie-Mellon University
Artificial Intelligence Laboratory, MIT

Support for the MIT Forum for Supply Chain Innovation provided by:
General Electric
General Motors
Intel Corporation
Michelin Corporation
MIT School of Engineering
MIT Sloan School of Management
SAP
The physicist Leo Szilard once announced to his friend Hans Bethe that he was thinking of keeping a diary: “I don’t intend to publish. I am merely going to record the facts for the information of God.” “Don’t you think God knows the facts?” Bethe asked. “Yes,” said Szilard. “He knows the facts, but He does not know this version of the facts.”

Hans Christian von Baeyer in *Taming the Atom*
APPENDIX

CLUES FOR BUSINESS PROCESS RELATED INNOVATION TO USHER IN ADAPTIVE SCM:

This article deals with the ideas and concepts that may converge for the future of adaptive value networks. Often the key question is where to get started. We have made the point that process is the key and technology is a catalyst. Here are some clues with respect to processes that may offer room for innovation (Simchi-Levi et al, 2002).

Table 1-3. What’s Good for Your Business: Optimize or adapt?

<table>
<thead>
<tr>
<th></th>
<th>Optimize?</th>
<th>Adapt?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distribution to Customer Assignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distribution Logistics Strategies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distribution Network Configuration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production-Distribution Schedule</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inventory Control by SKU and Nodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vendor Managed Inventory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supply Contracts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outsourcing &amp; Procurement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategic Partnerships</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product Design and Differentiation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plant-Product Assignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer Value/Profitability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Support Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Technology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1-4. SCM Models and Parameters

<table>
<thead>
<tr>
<th>SCM : Model Simple, Think Complex</th>
<th>Paradigm or Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic analytical models</td>
<td>Variables are known and specified</td>
</tr>
<tr>
<td>Stochastic analytical models</td>
<td>At least one variable is unknown and follows a probability distribution</td>
</tr>
<tr>
<td>Economic models</td>
<td>Game Theory models of buyer-supplier</td>
</tr>
<tr>
<td>Cost-based simulation CBS for material</td>
<td>Order quantities, response times, cost data</td>
</tr>
<tr>
<td>CBS of production control</td>
<td>Lot size, lead time, material response time</td>
</tr>
<tr>
<td>CBS of finished goods stockpile</td>
<td>EOQ, demand data, production lead times</td>
</tr>
<tr>
<td>CBS of distribution</td>
<td>Ordering policies, transportation time requirements, demand &amp; cost data, fill rate</td>
</tr>
</tbody>
</table>

Model sources of uncertainty (with certainty theory, Bayesian updating, fuzzy sets):
- Customer demand (Bullwhip Effect or “Hog-Cycle” Effect);
- Supply deliveries;
- External markets.

Quick Wins from Logistics Network Configuration:
- Storage at manufacturing plants (raw material, WIP, finished goods);
Chapter 1

- Pick, load, ship to warehouse or DC;
- Unload and store at warehouse or DC;
- Pick, load and ship for delivery to next node (customer);

**Savings may be possible from analytical optimization of (Table 1-5):**

<table>
<thead>
<tr>
<th>Process</th>
<th>Impact of Real-Time Data</th>
<th>Value from RFID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dock receiving capabilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage capabilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Receiving methodologies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order-generation capabilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery time constraints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pricing and Promotions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merchandising requirements</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Collection based on model:**
- Group products and/or product families (demand per product per customer);
- Group accounts by customer value and/or geography (zone) plus delivery frequency;
- Product shipment mapped by source warehouse vs. customer/zone;
- Demand by SKU per product (family) per zone;
- Production capacity (annual) at each plant;
- SKU storage capacity in warehouses (BOM for delayed differentiation products);
- Transportation mode, rates (TL/LTL by SKU) and cost (product/mile) between nodes;
- Service level (observed vs. expected vs. promised vs. industry best);
- Inventory carrying cost for safety stock to reach service level;
- Delivery time by customer/zone (map locations vs. transportation distances);
- Order processing cost (labor) and fixed operating cost (by nodes);
- Return and warranties (service cost);
- Wastage and shrinkage (costs).

**Estimate/quantify time lag between processes (consider total system benefit):**
- point of origin of data/information and data upload/update/accessibility;
- systems visibility of data from any single point of contact;
- data/information usage in systems (disruption management delay);
- information application/use to improve (adapt) decision support system (DSS).

**One Outcome:**
Industry “clockspeed” vs. “lag” may suggest process innovations for real-time adaptive SCM.

**Quick Wins from Inventory Management** (Raw Material, WIP, Finished Product)
The source of system-wide savings forecast can be based on near-actual customer demand or “pull” strategies, such as buy-back or revenue-sharing contracts. Inventory carrying cost is
about 20-40% of (turnover) value/year. Most (software) planners use Economic Lot Size Model (1915) to calculate economic order quantity (EOQ) that minimizes the cost function:

\[ Q^* = \sqrt{\frac{2KD}{h}} \]

Parameters of this model may still be valid, but assumptions in this formula are likely targets for (real time data catalyzed) improvements toward “adaptive” supply chain management. Model assumptions (Table 1-6):

<table>
<thead>
<tr>
<th>Assumptions: subject to change if process modified/benefits from</th>
<th>Real Time Data (RFID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demand is constant at a rate of D items per day</td>
<td></td>
</tr>
<tr>
<td>Order quantities fixed at Q items per order (safety stock)</td>
<td></td>
</tr>
<tr>
<td>Fixed cost K is incurred each time warehouse places an order</td>
<td></td>
</tr>
<tr>
<td>Inventory carrying/holding cost (h) accrues per unit per day</td>
<td></td>
</tr>
<tr>
<td>Lead time is zero</td>
<td></td>
</tr>
<tr>
<td>Initial inventory is zero (shift inventory cost to supplier)</td>
<td></td>
</tr>
<tr>
<td>Cycle Count Frequency (labour)</td>
<td></td>
</tr>
<tr>
<td>Infinite planning horizon (periodic review)</td>
<td></td>
</tr>
</tbody>
</table>

If one can factor in the “improvements” from real-time data that may help reduce variability (lead time heterogeneity, demand fluctuations) then, this formulation may still remain an effective model to indicate when orders should be received at warehouses (precisely when the inventory level drops to zero). Implementing ZIOP (zero inventory ordering property) involves precision real-time data synchronization across value chain partners that may make it possible to delay orders until inventory is zero (for whom?). In a centralized system, practice of ZIOP may approach near-reality but in a decentralized system concepts like CPFR along with real-time data sharing may be required as a precursor to practice of ZIOP.

![Figure 1-21. Steps in Model Building](image-url)
Traditional SCOR Model based on “Push” System: Is it still relevant for your need?

(Table 1-7)

<table>
<thead>
<tr>
<th>Perspectives</th>
<th>Metrics</th>
<th>Measure</th>
<th>Real-Time Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply Chain Reliability</td>
<td>On-time delivery</td>
<td>Percentage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Order fulfillment lead time</td>
<td>Days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fill Rate</td>
<td>Percentage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Perfect order fulfillment</td>
<td>Percentage</td>
<td></td>
</tr>
<tr>
<td>Flexibility &amp; Responsiveness</td>
<td>Supply chain response time</td>
<td>Days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Upside production flexibility</td>
<td>Days</td>
<td></td>
</tr>
<tr>
<td>Expenses</td>
<td>SCM cost</td>
<td>Percentage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Warranty as % of revenue</td>
<td>Percentage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Value added per employee</td>
<td>USD/EUR</td>
<td></td>
</tr>
<tr>
<td>Asset / Utilization</td>
<td>Total inventory days of supply</td>
<td>Days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cash-to-cash cycle time</td>
<td>Days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Net asset turns</td>
<td>Turns</td>
<td></td>
</tr>
</tbody>
</table>

(One solution fits all?) SCM Software: Can it help with strategic effects?

Top Line Revenue Growth:
- Reduced time from concept to production;
- Minimize engineering change orders after production release;
- Increased rate of innovation;
- Better on-time delivery (fewer canceled orders; fewer late penalties);
- Higher quality (fewer returns).

Reduced Requirements for Working Capital
- Raw material, WIP and finished goods inventory;
- Inventory obsolescence;

Higher Return on Fixed Assets

Higher Margins
- Lower shipping cost;
- Lower manufacturing costs;
- Lower wastage;
- Improved product mix;
- Reduced inventory carrying cost.

Above measures/metrics may be driven by the following applications:
1. **ADAPTIVE VALUE NETWORKS**

- Collaborative Product Design;
- Collaborative Planning and Forecasting;
- Optimized Manufacturing Planning;
- Inventory Planning and Optimization;
- Synchronized Planning;
- Detailed Scheduling;
- Accurate Order Promising;
- Optimized Transportation Routing.

Strategic “quick wins” categories likely to benefit from real-time information/data (RFID):

- Reduced requirements for working capital;
- Higher margins;

## NOTES

1. Information Asymmetry is a concept borrowed from economics. In 1776, in *The Wealth of Nations*, Adam Smith put forward the idea that markets by themselves lead to efficient outcomes. The mathematical proof specifying the conditions under which it was true, was provided in 1954 by Gerard Debreu (Nobel Prize 1983) of the University of California at Berkeley and Kenneth Arrow (Nobel Prize 1972) of Stanford University (Arrow, K. and G. Debreu (Existence of an equilibrium for a competitive economy. Econometrica 3 265–290). However, the latter result showing that when information is imperfect (information asymmetry) or markets are incomplete, competitive equilibrium is not efficient is due to B. Greenwald and J. Stiglitz in 1986 (Globalization and Its Discontents by Joseph E. Stiglitz).


3. It is beyond the scope of this chapter to delve into even a moderate level of discussion of Operations Research and Game Theory. Our intent is to offer some simple descriptions and indications about the possibilities of Game Theory applications in SCM. Game Theory applications, per se, are unlikely to make SCM more adaptive but these models can help the current processes by providing deeper insights. It is not uncommon to find businesses that are severely under-optimized in their current SCM practices. In such cases, it is speculative whether real-time information or efforts to be adaptive will meet with success. Optimization, including game theoretic tools, may be necessary to “tune the engine” before adaptive SCM can offer value.

4. Prisoner’s Dilemma was authored by A. W. Tucker of Princeton University [PhD advisor of John Nash]. Al Tucker was on leave at Stanford in the Spring of 1950 and, because of the shortage of offices, he was housed in the Psychology Department. One day a psychologist knocked on his door and asked what he was doing. Tucker replied: “I’m working on game theory,” and the psychologist asked if he would give a seminar on his work. For that seminar, Al Tucker authored the Prisoner’s Dilemma.

(www.nobel.se/economics/laureates/1994/nash-lecture.pdf)
REFERENCES

5. Christensen, C., The Innovators Dilemma: When New Technologies Cause Great Firms to Fail, Chapter 8, 168-169, 2000

Other Web Sites Referenced
http://ww.wrmca.com/archives/sept-oct02/0209044-2.html
http://ww.bea.doc.gov
http://ww.cia.gov
http://ww.legal-database.com
http://ww.aetherwire.com/CDROM/Welcome.html
http://grouper.ieee.org/groups/802
http://www.multispectral.com/pdf/APPsVGs.pdf
http://ww.technologyreview.com
http://www.csail.mit.edu/research/abstracts/abstracts03/web/02berners-lee.pdf
Les outils et procédés permettant de réduire les insuffisances de la chaîne logistique sont très précieux. La capacité à s’adapter peut en effet ne pas dépendre de la seule technologie mais d’une perpétuelle aptitude à l’innovation des processus dans la gestion de la Supply Chain. Pour cela, les dirigeants doivent être capables d’imaginer l’utilisation combinée de différents concepts, outils et technologies pour réduire les insuffisances, les incertitudes, et l’asymétrie de l’information dans le réseau de valeur. Cet article propose de combiner différentes idées dans un esprit d’approche « solutions » visant à améliorer les décisions.

Cette proposition est illustrée par la Figure 1.

En ce qui concerne les systèmes de décision, nous prendrons souvent le Supply Chain Management (SCM) ou la disponibilité militaire, la gestion de l’incertitude est essentielle.

Nous proposons une convergence raisonnable des concepts, des outils, des technologies et des normes existants ; utilisés conjointement, ils peuvent améliorer l’adaptabilité des systèmes de décision pour lutter contre l’incertitude dans des applications aussi diverses que l’optimisation des profits ou le temps de réaction dans la disponibilité hospitalière ou militaire. Ces améliorations doivent viser à réduire le « bruit » et favoriser l’adaptabilité. Cette proposition est illustrée par la Figure 1.

En ce qui concerne les systèmes de décision, nous prendrons souvent le Supply Chain Management comme exemple et nous discuterons de la façon dont les pratiques actuelles de la chaîne logistique pourraient s’améliorer si les praticiens (les décideurs) adoptaient et utilisaient réellement un nouveau schéma de pensée, de nouveaux outils analytiques, de nouvelles technologies, etc.

Avertissement
Cet article est vraiment très simplifié, incomplet et plein de digressions. L’auteur endosse la pleine responsabilité des éventuelles erreurs et incohérences, et il présente ses excuses au cas où la lecture en serait peu convaincante. Il espère que ce brassage d’idées (simple exploration) suscitera de nouvelles réflexions. Outre les collaborateurs cités en fin d’article, l’auteur s’est permis d’utiliser plusieurs sources d’information pour « nourrir les fils » et montrer comment des disciplines a priori éloignées peuvent, lorsqu’elles sont fusionnées, offrir de nouvelles directions de recherche. Enfin, la liste des références est tout à fait incomplète. Il est évident que la recherche initiale n’est pas due à l’auteur. En revanche, les opinions et commentaires exprimés ici sont bien les siens, et ne représentent en aucun cas la position du MIT. Enfin, la liste des collaborateurs ou de leurs organismes d’appartenance. Pour les spécialistes, il est possible que cet article n’amène rien de neuf. Mais c’est la synthèse, la convergence des idées telle que le suggère l’auteur qui peut précipiter la transformation de certains systèmes décisionnels dans le sens de l’adaptabilité ou peut-être, avec le temps, de la prédiction. Toute suggestion peut être adressée par courriel à l’adresse suivante : shoumen@mit.edu

Titre original
Adapting Decisions, Optimizing facts and Predicting Figures: Can Confluence of Concepts, Tools, Technologies and Standards Catalyze Innovations
Figure 1 : Proposition pour passer de l’optimisation du « bruit » à une meilleure adaptabilité à l’avenir.

Quand Napoléon affirmait (si tant est qu’il le fit réellement) :
« ceux qui ont besoin de se forger une représentation mentale de chaque chose sont inutiles au commandement », il faut comprendre cette dénonciation du défaut originel.

En effet, s’il aborde une bataille ou telle circonstance, le chef militaire réalisera dès les premiers contacts que quelque chose a mal tourné. Son Figure est détruit ; et il ne lui reste plus rien, sauf peut-être une autre représentation personnelle, qui elle-même ne lui servira pas longtemps.

Il est possible d’autre part, quand sa première prévision s’avère erronée, qu’il soit assailli par une telle masse d’images différentes que, là-aussi, il ne sait plus quelles adaptations pratiques faire. En fait, trop de références au passé peut être presque aussi gênant que de pas de référence du tout.

Pour répondre de manière appropriée aux exigences d’un environnement en constante évolution, nous devons non seulement sélectionner et extraire des éléments de leur cadre général, mais aussi savoir quelles les en sont les parties qui peuvent circuler et muter sans modifier leur sens et leurs fonctions générales. F.C. Bartlett

nouvelles technologies et les normes émergentes pour réduire l’incertitude. L’emploi de données temps réel est déterminant pour l’industrie (commerce de détail, santé) et la défense. Pourtant, ces quelques dernières années (1999-2003), on a constaté un intérêt disproportionné pour les outils d’acquisition de données type Technologies d’Identification Automatique (AIT), dont un exemple particulièrement marquant est l’identification par fréquence radio (RFID). À notre avis, le bon point de départ d’une réflexion sur la réduction de l’incertitude du système, c’est de voir comment les données temps réel peuvent influer sur le processus (opérationnel).

L’emploi des AIT pour identifier des objets avec le RFID est intéressant quand ces données (systémiques, locales) deviennent utiles au bon point de départ d’une réflexion sur l’identification (EBM) afin de connecter les données temps réel à un logiciel tradi-ondes (systémiques, locales) deviennent utiles au bon point de départ d’une réflexion sur l’identification (EBM) afin de connecter les données temps réel à un logiciel tradi-

Figure 2 : La connectivité des données temps réel et du processus (analytique temps réel) offre de meilleures informations pour les prises de décision.

Les données temps réel au « bon moment » (données right-time) peuvent modifier les processus opérationnels actuels et stimuler l’innovation. Cependant, il n’est pas suffisant de fournir des données temps réel aux « systèmes Legacy » ou aux ERP. Pour améliorer l’adaptabilité, les systèmes décisionnels doivent pouvoir accéder à l’information « right-time », elle-même fondée sur des données temps réel (analytique temps réel) provenant de différentes sources (marqueurs RFID, capteurs, GPS, codes à barres). Le débat sur le format (code électronique/EPC ou codification identifiant universel/UID) reste encore ouvert, mais cela ne devrait pas bloquer la réflexion, condition préalable pour que l’innovation des processus fasse bon usage des « données right-time. »

Il est possible que les architectures agnostiques de format (logiciels) détiennent la clé pour connecter les données temps réel à d’autres logiciels ou ERP. Mais il nous semble peu probable que les progiciels ERP puissent traiter ou analyser des flux de données temps réel de l’ordre du Gigabit. Certains avaient suggéré (et nous leur emboîtons le pas) de s’intéresser à l’utilisation d’un logiciel basé sur agent (ABM) combiné à un logiciel traditionnel basé sur équation (EBM) afin d’extraire l’information du flot croissant de données temps réel. La connectivité de ces données est tout aussi importante que le partage de l’information entre entités pour améliorer le processus de prise de décision. Et à son tour, ce partage de données ou d’information permet d’améliorer la performance de l’ensemble du réseau de valeur.

L’interaction entre entités exige une infrastructure sécurisée et une plate-forme diffuse totalement ouverte à la collaboration. Nous proposons une telle plate-forme (ubiqautiste, agnostique avec transpondeur et appartenant à l’infrastructure civile) dans laquelle les interrogateurs de données sont des systèmes de radio logicielle (SDR-SWR). L’accès et le contrôle du partage des données sont régulés et authentifiés via la couche Application logicielle (transmise par l’Internet) de la même manière que pour une Internet « appliance1 » avec régulation à distance (ex : allumer le four micro-ondes tout en rentrant chez soi en voiture).
Dans les 5, ou plus vraisemblablement les 25 à 50 prochaines années, quand nous pourrons passer d’un statut opérationnel auto-adaptif à un statut prédicatif, il nous faudra d’autres concepts et outils, probablement encore inconnus aujourd’hui. Cependant, en contribution à la phase prédicative des opérations, nous proposons une idée issue de l’économétrie des séries de temps. Tout simplement, il s’agit d’étudier la possibilité d’utiliser des données temps réel « brutes » (sans regroupement ni tri) pour mieux comprendre les changements et y réagir, quasiment en temps réel. Ceci peut aider à réduire encore plus les risques et l’incertitude, et peut-être même à atténuer l’effet Coup de Fouet (Bullwhip). De tels outils analytiques économétriques sont utilisés dans le domaine de la finance (volatilité des prix de la bourse). Si l’on peut modifier les outils économétriques (tels que le GARCH - Hétéroscédasticité conditionnelle autorégressive généralisée) pour les utiliser avec des données ODD (données dépendant d’objets) temps réel, cela non seulement nous aidera à prédir les paramètres clés de la chaîne logistique (prévision de demande, prix) basés sur les entrées (données temps réel) mais, en plus, nous pourra nous fournir une mesure du risque associée à la prédiction. Et pouvoir quantifier le risque dans la chaîne logistique est tout particulièrement intéressant.

En résumé, nous allons essayer dans cet article de combiner différentes idées dans un esprit d’approche « solutions » dont le but est de réduire l’incertitude et d’améliorer les décisions. Nous enchaînerons les notions, allant de la Théorie des Jeux à la simulation, via les Technologies d’Identification Automatique (AIT), l’Econométrie des séries de temps, le Grid Computing, les Agents et le Web Sémantique. Il est fort possible que des gouvernements, des sociétés, des firmes de consultants et des universitaires ayant une connaissance approfondie dans un ou plusieurs de ces domaines passent les quelques prochaines décennies à tenter de synthétiser un ou des modèles ou un modus operandi efficace pour combiner ces idées avec d’autres concepts, outils, technologies et normes émergents qui, utilisés conjointement, pourraient permettre de mieux saisir les incertitudes, les analyser, les réduire et y répondre (Figure n°4). Comprendre la convergence facilitera l’étude du paradigme entre adaptabilité et efficacité. Le cadre de gestion (tableau de bord des outils) pour diagnostiquer et déterminer l’équilibre dynamique (spéciﬁque à l’industrie) peut optimiser le « push-pull » entre adaptabilité et efficacité dans la recherche de l’Équilibre Dynamique de Gibbs.

« Au musée de la Science à Barcelone, j’ai vu une exposition qui illustrait parfaitement le chaos. Une version non linéaire d’un pendule était montée de telle manière qu’un visiteur pouvait saisir le poids et le faire se balancer comme il voulait, à la vitesse qu’il voulait. On observait alors le déplacement du balancier tout en l’enregistrant avec un crayon sur une
feuille de papier. Puis on demandait au visitateur de reprendre la lentille et de reproduire exactement le même mouvement (position et vitesse). Quel que fût le soin qu’il y mettait, le résultat obtenu était très différent de ce qu’il était la première fois.

Je demandais au directeur du musée ce que faisaient les deux hommes debout dans un coin, qui nous observaient. Il me répondit : « Oh, ce sont deux Hollandais qui attendent d’embarquer le « Chaos » ». Apparemment, cette exposition devait être démontée et transférée à Amsterdam. Depuis, je me suis toujours demandé si les services de ces deux Hollandais n’allaient pas être recherchés un peu partout dans le monde par des organisations qui voudraient se débarrasser de leur « chaos » [1].

Pour l’industrie, la quête du Graal, c’est de débarrasser la chaîne logistique du « chaos » pour mieux l’adapter aux fluctuations de la demande. La gestion de l’incertitude est obérée par le degré croissant d’asymétrie de l’information⁶ entre les partenaires de la chaîne logistique ou du réseau de valeur⁷ - designers, fournisseurs, distributeurs, détailants, consommateurs, dont les objectifs sont différents et souvent opposés, ce qui risque de parsumer d’embûches le chemin des réseaux d’affaires adaptifs de l’avenir [2].

Un des objectifs commerciaux des fournisseurs est d’obtenir des fabricants qu’ils s’engagent sur des achats fermes de grands volumes (mais avec flexibilité de livraison). C’est donc en contradiction avec l’objectif du producteur pour lequel une réaction rapide à une fluctuation de la demande conduit à un excédent des stocks de matières premières. Le fabricant doit produire en masse (pour profiter des économies d’échelle), mais les flux de production doivent s’adapter aux variations de la demande même si l’utilisation des ressources avait été planifiée sur la base de prévisions. En fait, le producteur peut avoir besoin de plus ou moins de matières premières et rechercher la flexibilité en jouant sur les quantités achetées, ce qui va à l’encontre de l’objectif du fournisseur. Le désir du fabricant de produire en grand nombre se heurte à son tour à celui des centres de stockage et de distribution qui cherchent à réduire les volumes stockés, leur coût augmentant les frais de transport pour l’ensemble des acteurs.

Pendant l’année 2000 et pour les USA seuls, les coûts relatifs à la chaîne logistique dépassèrent le trillion de USD (10% du PIB), c’est-à-dire à peu près le PIB russe et plus que le PIB du Canada ou même le PIB combiné des 22 pays membres de la Ligue Arabe (dont les riches nations pétrolières), lui-même étant inférieur à celui de l’Espagne. Une simple économie de 10% des coûts de la chaîne logistique aux USA équivaudrait pratiquement au PIB de l’Irlande[4].

En conséquence, les outils et procédés permettant de réduire les insuffisances de la chaîne logistique sont très précieux[5]. La capacité à s’adapter peut en effet ne pas dépendre de la seule technologie mais d’une perpétuelle aptitude à l’innovation des processus dans la gestion de la SCM. Pour cela, les dirigeants doivent être capables d’imaginer l’utilisation combinée de différents concepts, outils et technologies pour réduire 1/ les insuffisances, 2/ les incertitudes, et 3/ l’asymétrie de l’information dans le réseau de valeur.

Le but recherché, c’est de passer d’un management logistique global basé sur l’offre (« push ») à un réseau auto-adAPTé basé sur la demande (« pull »). Un des moteurs de cette transformation, c’est l’utilisation potentielle de données et d’informations temps réel comme déclencheur d’étapes de décision autonomes susceptibles d’effectuer simultanément re-planification et exécution. D’après différentes études, en 2003 les entreprises ont généré un flux de données par seconde supérieur à un téraoctet (sans compter les données rassemblées par les AIT). Doit-on considérer cela comme de l’information ? En l’état, c’est peu probable ! L’aptitude à extraire des renseignements à partir de données de manière à gérer l’information peut être le facteur de différenciation entre les entreprises qui sauront exploiter les données (telles que l’identification automatique ou les capteurs) et celles qui ne le sauront pas.

Les données stockées dans les ERP peuvent pâtrir de problèmes qui réduisent la valeur de

---

**a.** L’asymétrie de l’information est un concept emprunté aux économistes et utilisé de façon assez informelle ici pour bien souligner le manque de visibilité de l’information (données) entre organisations. En 1976, dans The Wealth of Nations, Adam Smith avançait l’idée que les marchés en eux-mêmes donnaient des résultats efficaces. La preuve mathématique spécifiant les conditions dans lesquelles ceci était vrai fut apportée en 1934 par Gérard Debreu (Prix Nobel 1983) et Kenneth Arrow (Prix Nobel 1972) dans leur ouvrage Existence of an equilibrium for a competitive economy, Econométrica.


---

Ampex lança le premier magnétoscope en 1956. Chaque unité valait 50 000 USD et seuls concurrents, RCA et Toshiba, étaient loin derrière. Sony, JVC et Matsushita étaient de simples observateurs. Masaru Ibuka, co-fondateur de Sony et Yuma Shiraishi, JVC, demandèrent à leurs ingénieurs respectifs de produire des appareils à 500 USD, soit à peine 1% du prix d’Ampex. Dans les années 1980, les ventes de magnétoscopes sont passées, chez Sony, de 17 millions à 2 milliards, chez JVC, de 2 millions à 2 milliards, chez Matsushita de 6 millions à 3 milliards de dollars US et pour Ampex [3], de 296 millions à 480 millions de USD.

S’adapter ou Disparaître !
leur information. De même, les systèmes ERP peuvent compromettre l’efficacité des données dynamiques si les systèmes statiques sont incapables de répondre en temps quasi réel. Aussi, quand les planificateurs utilisent ce type de données ERP ou de sources d’information dans leurs prévisions ou optimisations, on peut s’interroger sur la validité du résultat. En effet, l’optimisation du processus ou la prévision elle-même peuvent avoir été fondées sur des « bruits » plutôt que sur des données dynamiques solides (cf. Figure n°1). La piètre qualité des données et l’asymétrie de l’information entre les différents partenaires de la chaîne logistique sont en fin de compte des erreurs (d’optimisation, de prévision), et celles-ci s’accumulent aux différents niveaux de la chaîne logistique, ce qui se traduit par l’Effet Coup de Fouet/Bullwhip. La Figure n°5 montre la Figure n°5. La Figure n°6 présente l’Effet Bullwhip basé sur des données réelles provenant de l’industrie semi-conducteurs.

**Vers des réseaux auto-adaptifs ?**

Les outils et technologies servant à lutter contre l’Effet Bullwhip peuvent aussi être utilisés pour inciter les chaînes logistiques à devenir auto-adaptatives. Cela suppose que les organisations encouragent l’innovation des processus d’affaires visant à améliorer l’interaction entre entités (échanges d’information intra et inter-entreprises) et qu’elles recherchent aussi la disparition des silos de données. En réalité cependant, « culture » organisationnelle et gestion du changement sont nécessaires pour stimuler de nouveaux modes de pensées concernant la convergence des concepts, des outils, des technologies et des normes. Alors, quelques groupes pourront imaginer la manière de connecter des objets (atomes) chargés de données (bits) à des processus ou à l’analytique temps réel pour en extraire l’information temps réel destinée à des décisions auto-adaptatives - à informations auto-adaptives qui pourront, à leur tour, optimiser la nature (design, répartition) ou les caractéristiques (prix, risque) d’objets dans une chaîne télémétrique.

**Tableau 1 : Éléments de la convergence de concepts, d’outils, de technologies et de normes proposées.**

<table>
<thead>
<tr>
<th>Outils et Concepts</th>
<th>Sources de Données</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recherche Opérationnelle</td>
<td>Technologies d’Identification Automatique (RFID, ULB, GPS, RTLS)</td>
</tr>
<tr>
<td>Théorie des Jeux</td>
<td>Technologies d’identification (GPRS, Voix, Manuel, Code à barres 2e génération, Code à barres)</td>
</tr>
<tr>
<td>Agents Intelligence Artificielle Distribuée</td>
<td>Protocoles sans fil (802.11, 802.16)</td>
</tr>
<tr>
<td>Outils</td>
<td>Réseaux capteurs (technologie Zigbee - 802.15.4)</td>
</tr>
<tr>
<td>Économétriques</td>
<td>Systèmes de Radio Logicielle (SWR).</td>
</tr>
<tr>
<td>Web Sémantique</td>
<td>Grid Computing</td>
</tr>
</tbody>
</table>

**Recherche Opérationnelle et Théorie des Jeux.**

La locomotive de l’optimisation (les algorithmes), c’est la recherche opérationnelle. C’est un domaine de prospection intense et d’innombrables sources d’information sont disponibles. La Théorie des Jeux n’était pas un nom d’usage courant jusqu’en 1994, date à laquelle les travaux de John Nash, et le film dont il fut l’objet, incitèrent les joueurs économiques généralistes à en découvrir progressivement les vertus. Pendant des siècles, les économistes ont travaillé sur différents modèles de jeux théoriques, mais on considère que les pères de la Théorie des Jeux modernes sont les pères de la Théorie des Jeux moderne.

**Figure 5 : L’Effet Coup de Fouet/Bullwhip [10]**

**Figure 6 : L’Effet Coup de Fouet dans la chaîne logistique de l’équipement semi-conducteur montre les prévisions de demande par rapport à l’achat réel d’équipement [11]**

---

La Théorie des Jeux nous aide à modéliser, analyser et comprendre le comportement des nombreuses parties prenantes qui interagissent dans le processus de prise de décision. En tant que telle, cette théorie s’occupe des problèmes d’optimisation interactive. C’est, en particulier, un outil d’analyse de situations dans lesquelles les parties s’efforcent de maximiser leurs propres bénéfices (espérés) en même temps qu’elles choisissent leur stratégie. Au final, les dividendes de chaque partie dépendent du profil des stratégies choisies par l’ensemble des parties. La plupart des situations économiques peuvent être modélisées par un « jeu » car, dans toute interaction impliquant deux parties ou plus, le bénéfice de chacune dépend des actions de l’autre. Donc le thème essentiel dans la Théorie des Jeux est l’interaction. Dans le monde des affaires, chaque décideur est un acteur qui prend une décision ou choisit une stratégie qui sera influencée par le concurrent. Et on suppose que les hommes d’affaires font des choix rationnels pour optimiser leurs bénéfices.

Pourquoi les entreprises se comportent-elles ainsi ? Dans cette situation, et dans quelques autres, elles sont confrontées à ce qui, dans la Théorie des Jeux, est connu sous le nom de « Dilemme du Prisonnier », la réponse rationnelle n’étant pas forcément la plus sensée. 

## Le Dilemme du Prisonnier

Alice et Bob sont arrêtés près du lieu d’un cambriolage. Ils sont interrogés séparément. Chaque suspect peut soit avouer, avec l’espoir d’une peine plus légère, soit refuser de parler (ne pas avouer). La police n’a pas d’informations suffisantes pour les inculper, à moins que l’un d’eux n’avoue. Chacun doit donc choisir entre avouer et incriminer l’autre ou se taire, sans savoir ce que l’autre va faire. Si aucun n’avoue, tous les deux encourtent alors une peine d’un an de prison pour port d’armes prohibé. Si les deux avouent et s’accusent mutuellement, ils écopent de dix ans de prison. Mais, si l’un des voleurs avoue et accuse l’autre et que l’autre n’avoue pas, alors celui qui a coopéré avec la police ressortira libre tandis que l’autre sera condamné à la peine maximale, 20 ans de prison. Dans ce cas « l’espace stratégique » est simple : avouer ou ne pas avouer, et chacun choisit une des deux stratégies. Les dividendes (pénalités) sont les peines infligées.

### Tableau 2 : Le Dilemme du Prisonnier : Alice (colonne) opposée à Bob (ligne).

<table>
<thead>
<tr>
<th></th>
<th>Alice avoue</th>
<th>Alice n’avoue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob avoue</td>
<td>10 - 10</td>
<td>0 – 20</td>
</tr>
<tr>
<td>Bob n’avoue</td>
<td>20 - 0</td>
<td>1 - 1</td>
</tr>
</tbody>
</table>

Dans ce tableau, les nombres représentent les peines encourues par les prisonniers suivant les combinaisons de stratégies choisies, celui de gauche est pour la personne située dans les lignes (Bob), celui de droite pour celle figurant dans les colonnes (Alice). Ainsi, la lecture de la première colonne donne, de haut en bas : si tous les deux avouent, chacun est condamné à 10 ans, mais si Alice est la seule à avouer, Bob est alors condamné à 20 ans et Alice est libre.

En conséquence, quelles sont les stratégies « rationnelles » dans ce jeu si tous deux veulent avoir la peine la plus légère ? Alice va faire le raisonnement suivant : « De deux choses l’une : ou Bob avoue, ou il se tait. S’il avoue, je prends 20 ans (si moi, je n’avoue pas) et 10 ans si j’avoue (si je coopère) ; dans ce cas, il est préférable que j’avoue. D’un autre côté, si ni Bob ni moi n’avouons, j’écope d’une année ; mais là, si j’avoue, je ressors libre. Dans les deux cas, il s’avère préférable que j’avoue. Donc, je vais avouer. » Mais Bob peut, et vraisemblablement va, suivre le même cheminement intellectuel.

Les deux individus raisonnent de façon rationnelle, avouent et prennent chacun 10 ans. Or s’ils avaient agi de façon irrationnelle et n’avaient pas avoué, ils n’auraient écopé chacun que d’une année seulement. 

Le Dilemme du Prisonnier est un exemple simple de jeu statique non-coopératif dans lequel les joueurs choisissent en même temps leur propre stratégie, et doivent s’y tenir. L’élément principal de ces jeux, c’est l’existence et la spécificité de l’équilibre de Nash (NE). NE est le point où aucun joueur n’a intéré-

---


Un fabricant de puces casse les prix de ses concurrents. Il sait que le marché peut être représenté par les deux individus raisonnant de la même façon. C’est donc le même raisonnement qu’il peut faire. Or, dans ce type de situations, il est préférable de maximiser le profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximiser son profit. Donc, si le fabricant raisonne comme le psychologue, il va maximiser son profit. Il va donc conclure que le fabricant va maximise
rêt à changer sa stratégie puisque chacun a choisi celle qui maximise ses propres bénéfices compte tenu des stratégies des autres joueurs.

Un concept clé qui n’a pas été mis en exergue dans le Dilemme du Prisonnier, c’est la répétition des interactions. Or dans le monde des affaires, les joueurs savent qu’ils sont dans le « jeu » pour longtemps. Ils peuvent donc choisir de coopérer, surtout s’ils estiment que cette démarche a des chances de faciliter la coopération, voire la collaboration de demain. C’est par des actions répétées que les compagnies se bâissent une réputation, qui influence les actions des autres. Par exemple, Intel Corp. utilise son programme de notation « fournisseurs », qui évalue, pour chacun d’entre eux, les coûts, la disponibilité, le service, la rapidité de réaction et la qualité des prestations. Il maintient ainsi ses principaux fournisseurs dans une course permanente à l’excellence. « Nous récompensons les fournisseurs qui ont les meilleures notes en faisant plus d’affaires avec eux », dit Keith Erickson, Directeur des Achats. Et comme incitation supplémentaire, Intel Corp. englobe de temps en temps ses meilleurs fournisseurs dans ses publicités. La compagnie laisse même les plus performants se prévaloir de leurs relations avec Intel Corp dans leur propre publicité.

Dans la réalité, chaque partie de la chaîne logistique agit de façon tout à fait égoïste. Et donc la somme des choix individuels n’aboutit pas à un résultat optimal pour l’ensemble de la chaîne. Le bénéfice d’une chaîne « décentralisée » composée de multiples compagnies gérées indépendamment est en général inférieur au total des bénéfices de la version centralisée de la même chaîne, version dans laquelle les interactions entre partenaires (fournisseurs, fabricants, détaillants) sont gérées par un seul décideur (moins d’inconnues) de façon à optimiser les bénéfices au profit de l’ensemble. Le partage de l’information dans cette dernière configuration réduit les évidentes insuffisances de la version décentralisée, qui sont dues à la « double marginalisation » découlant d’un processus de décision égocentrique. Donc, le bénéfice optimum est plus important dans une chaîne centralisée où l’information est partagée.

Une stratégie de réduction des carences dans une chaîne décentralisée est « l’intégration verticale » : une entreprise possède chaque élément de sa chaîne logistique. Ford Motor Company en fut un bon exemple. De nos jours, les demandes et préférences des consommateurs changent rapidement. Les compagnies qui se concentrent sur leur métier de base sont généralement plus réactives pour rester en tête de la compétition. En conséquence, on constate une tendance à « l’intégration virtuelle » dans laquelle les chaînes logistiques sont constituées de compagnies gérées indépendamment mais étroitement associées. Des stratégies fondées sur le partage de l’information comme celle de la gestion partagée des approvisionnements (VMI – Vendor Managed Inventory) sont utilisées par quelques firmes (Dell, P&G, Wal Mart).

En dépit de progrès dans le partage de l’information, il est rare que les chaînes logistiques actuelles disposent d’une connaissance totale des acteurs et des décisions ou des résultats. Il est courant qu’une firme ait une meilleure prévision de la demande qu’une autre ou qu’elle ait une meilleure information sur ses propres coûts et procédures de fonctionnement. Si une entreprise sait qu’une autre est susceptible d’avoir une meilleure information, elle peut choisir des modes d’action intégrant cette donnée. La Théorie des Jeux donne des outils pour étudier des cas avec une asymétrie de l’information d’une complexité analytique croissante. Pour illustrer ceci, nous allons nous intéresser à un type particulier de jeu, le Jeu de Signaux [20].

**Jeu de Signaux**

Dans sa forme la plus simple, le Jeu de Signaux se joue à deux. Un joueur a de meilleures infor-
mations que l’autre, et celui qui est le mieux informé bouge le premier. Par exemple, un fournisseur doit se mettre en situation de fournir un composant clé de la production d’un fabricant. Celui-ci a une meilleure prévision de la demande que le fournisseur. Dans un monde idéal, il pourrait partager son information avec son fournisseur pour que celui-ci se dote de la capacité appropriée. Mais il a aussi intérêt à ce que son fournisseur dispose d’un volume de composants plus important au cas où la demande serait plus forte que prévu. Il a donc intérêt à gonfler ses prévisions. Ce sera alors le fournisseur, s’il se fie au fabricant, qui supporterà le coût de l’augmentation de ces volumes. Le fabricant espère donc que le fournisseur va le croire et augmenter ses réserves. Heureusement, ce dernier connaît le « jeu » du producteur. Quel geste (signal) venant du fabricant pourrait-il bien inciter le fournisseur à considérer comme crédibles les prévisions du fabricant ?

Dans cet exemple, la demande D est en fait la somme de trois prévisions. La prévision du marché (µ) est faite par des analystes. Le fabricant, à partir de ses sources ou de son expérience, établit ses propres prévisions xi (ξ) qui, dans un système décentralisé, ne sont pas portées à la connaissance du fournisseur (asymétrie de l’information). Cependant, celui-ci, par son expérience, s’est constitué une grille de classement des fabricants fondée sur leurs antécédents et leur crédibilité. Ainsi, le fournisseur confirme la confiance qu’il accorde au « type » d’information prévisionnelle du fabricant et peut choisir une valeur de ξ censée être représentée par une distribution normale. Ceci introduit une variable aléatoire (stochastique). L’incertitude du marché est représentée par epsilon (ε) dont ni le fabricant ni le fournisseur ne peuvent contrôler la valeur. C’est donc une autre variable aléatoire (terme d’erreur), censée appartenir elle aussi à une distribution normale.

Ces hypothèses, qui ne sont pas rigoureusement quantifiées et que l’on suppose appartenir à une fonction donnée par une distribution normale, introduisent une variabilité. De telles erreurs s’accumulent au fil de chacune des étapes des chaînes logistiques (voir Figure n°9), contribuant ainsi à l’apparition de l’Effet Bullwhip. Nous énoncerons plus loin une proposition pour étudier comment on pourrait réduire ces écarts en se servant de données temps réel avec des outils analytiques combinant méthodes statistiques et dernières avancées dans l’économétrie des séries de temps [21].

Le jeu de signaux tel que décrit ci-dessus démarre par l’annonce des prix par le fournisseur : un prix ordinaire (w) et un prix pour achat anticipé (wa). Le fabricant établit sa prévision de commande et, s’appuyant sur la fiabilité qu’il accorde à ses prévisions, il réagit à la proposition de prix global du fournisseur en émettant une commande anticipée (y) au prix convenu de wa. Le volume de y envoie un « signal » au fournisseur, qui s’en sert pour confirmer la crédibilité qu’il doit accorder à la prévision « D » du fabricant. Fort de cela, le fournisseur peut donc déterminer ce que doit être sa capacité de fourniture (K) de manière à optimiser ses profits (risque inventaire). Au fur et à mesure, l’incertitude du marché se précise et le fabricant met à jour ses prévisions en utilisant cette valeur (ε). Il commande alors à son fournisseur le volume D-y, à un prix plus élevé (w). Bien que l’optimisation fondée sur des signaux puisse accroître les profits du fabricant et du fournisseur, elle demeure cependant sujette aux erreurs commises dans la valeur attribuée aux variables ξ et ε.

Le Jeu de Signaux permet de penser qu’un bon moyen de réduire les incertitudes serait de mieux préciser la valeur des variables ξ et ε. On dispose déjà d’un large éventail d’outils de recherche et d’optimisation pour s’attaquer à ces valeurs ou à la façon d’obtenir des valeurs...
sûres. Cependant, la persistance d’amples variations dans les chaines logistiques montre que les outils existants pour lutter contre l’incertitude ne sont pas forcément adaptés. C’est, en partie, l’une des raisons pour les quelles nous proposons l’emploi de données temps-réel pour réduire les erreurs, notamment pour les valeurs des variables $\xi$ et $\varepsilon$.

**Odd-var-garch :**
**Un outil analytique pour mieux utiliser les données temps-réel.**

Les outils actuels de prévision peuvent être obérés par :
- des données de piètre qualité,
- des données agrégées,
- une visibilité réduite pour les décideurs,
- des modèles de flux de données médiocres,
- une architecture informatique incomplète.

La précision d’une prévision tient en partie à l’adaptation des modèles mathématiques aux processus des entreprises. Pour autant, même avec de meilleurs outils statistiques, le décrochage entre fonctionnement et stocks sera cause d’inadéquation. Seule la convergence des données, des processus et des décisions permettra d’obtenir une réelle visibilité.

Nous avons à notre portée une excellente occasion d’étudier les outils statistiques avec des flots à ultra haut-débit de multi-gigabits de données précises.

Un modèle proposé est le **Odd-var-garch**

(Données Liées à l’Objet – AutoRégression Vectorielle – Hétéroscédasticité AutoRégressive Conditionnelle Généralisée). Ce n’est que maintenant qu’arrive à maturité ce secteur des perturbations actuelles ou passées ou des termes d’erreur aléatoires. En ce qui nous concerne, pour étudier cette modélisation, examinons les prévisions de demande du marché $D$ :

$$D = \mu + \xi + \varepsilon$$

sachant que :

En 1959, GE confia à la firme de consultants Arthur D. Little Inc. de Boston une étude pour déterminer s’il y avait ou non un marché pour des téléviseurs portables que GE était maintenant en mesure de construire en utilisant des transistors solides. Au bout de plusieurs mois, toujours en 1959, et après avoir dépensé des sommes vertigineuses (des millions de dollars) en groupes d’études ad hoc et en discussions, Arthur D. Little Inc. transmittit ses conclusions à GE : il n’y avait pas de marché pour de tels téléviseurs. La direction de GE écarta donc ce projet proposé par ses ingénieurs… Juste avant Noël 1959, Sony introduisait sur le marché américain un petit téléviseur B&W et en vendit plus de 4 millions en quelques mois [3].

2/ étant données les multiples combinaisons de \((\beta_0, \beta_1)\) que l’on peut évaluer, le but du CLRM est de choisir le couple \((\beta_0, \beta_1)\) qui minimise la somme des restes au carré:

\[
\sum_{i=1}^{n} e_i^2
\]

dans laquelle : \(e_i\) est le terme d’erreur aléatoire pour les données « échantillon », \(e_t\) est le terme d’erreur aléatoire des données « population », prise comme un tout.

Cette technique (le principe des moindres carrés ordinaires - OLS) minimise les erreurs de prévision en sélectionnant la paire \((\beta_0, \beta_1)\) la plus pertinente.

Le CLRM combiné aux données temps réel (RFID) représente une avancée. Cette (combinaison) idée est tirée des modèles arch et garch. Odd-var-garch a besoin de grands volumes de données et donne des prévisions avec une bien plus grande précision que ne le pourrait n’importe lequel des composants particuliers pris séparément.

Odd-var-garch requiert des progressions de type séquentiel pour combiner le CLRM avec les techniques de séries de temps. Une équation CLRM élémentaire est (1):

\[
Y_t = \beta_0 + \beta_1 X_{it} + \beta_2 X_{i2} + \cdots + \beta_k X_{it-k} + \varepsilon_t
\]

Et, parce que la dépendance du processus est capitale pour la réussite, la fourniture de données d’identification automatique temps réel à de vieux modèles de processus ne peut produire que des avantages mineurs.

Une tâche pour l’innovation du processus est d’étudier comment les « X » changent avec le temps réel. Dans la prochaine étape du développement de Odd-var-garch, les valeurs des « X » dans l’équation 1, sont prédites en utilisant des données passées de Y lui-même, d’où de l’intégration dans les systèmes décisionnels. La plupart des planificateurs pratiquent l’analyse de régression en la supposant fixe, produisant ainsi « des régressions fausses ».

La valeur prévisionnelle de \(x_t\) est définie par le modèle multi-co-intégré (2):

\[
X_{it} = \alpha_{0i} + \alpha_{1i} X_{i1} + \alpha_{2i} X_{i2} + \cdots + \alpha_{Ni} X_{iN} + \nu_{it}
\]

(\(i = 1, \ldots, n\);
\(t = 1, \ldots, T\))

\[
\cdots X_{it} = \alpha_{0i} + \alpha_{1i} X_{i1} + \alpha_{2i} X_{i2} + \cdots + \alpha_{Ni} X_{iN} + \nu_{it}
\]

(3):

\[
Y_t = \beta_0 + \sum_{i=1}^{N_1} \alpha_{1i} X_{i1-t} + \cdots + \sum_{i=1}^{N_k} \alpha_{ki} X_{ki-t} + \varepsilon_t
\]

sachant que:

\[
X_{it} = \text{variable } X_i \text{ au temps } t,
\]

\[
X_{kt} = \text{variable } X_k \text{ au temps } t,
\]

\[
X_{it-1} = \text{valeur de } X_i \text{ au temps } t-1 \text{ (valeur précédente)},
\]

\(N \) = période jusqu’à laquelle les valeurs précédentes de \(X_t\) peuvent être utilisées dans l’équation, et

\(U\) = terme d’erreur aléatoire.

Dans l’équation 2, \(\alpha_{11}\) et \(\alpha_{12}\) sont coefficients de \(X_{i1} \) et \(X_{i2}\) et sont appelés coefficients de pondération décalés. Suivant le nombre de « X » dans l’équation 1, on aura un nombre K d’équations qu’il faudra évaluer pour prévoir les « X » \((X_1, \ldots, X_k)\) qui seront alors utilisés pour obtenir une prévision inconditionnelle de « Y ». Ainsi, nous pouvons évaluer tous les paramètres \((\alpha, \beta)\) simultanément en re-écrivant l’équation CLRM élémentaire (équation 3 ou 4). L’équation 5, qui prend en compte l’impact sur Y des valeurs passées de Y lui-même, est de préférence (5):

\[
Y_t = \beta_0 + \sum_{j=1}^{N_1} \varphi_{j} U_{j} + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \alpha_{ki} X_{ki-t} + \varepsilon_t
\]

sachant que \(N\) ~ nombre de valeurs précédentes pour chacun des « X » dans le modèle.

En passant aux prévisions inconditionnelles de « Y », on accroît énormément le nombre de paramètres à évaluer, ce qui nécessite un important volume de données précises (AIT).

Il se peut que les entreprises plus petites se plaignent qu’il leur est impossible d’évaluer 6000 paramètres ou plus pour chaque sku. Pourtant, cela deviendra faisable en recueillant, via le grid, la puissance de traitement non utilisée dans de nombreux domaines.

On peut élargir l’équation 5 de manière à englober le principe d’application de la co-intégration dans les systèmes décisionnels. La plupart des planificateurs pratiquent l’analyse de régression en la supposant fixe, produisant ainsi « des régressions fausses ». L’idée de variables multi-co-intégrées a été développée par l’introduction du concept de degré d’intégration d’une variable. Si une
variable peut être rendue plus ou moins fixe en la différenciant \( d \) fois, on dit qu’elle est intégrée de l’ordre \( d \), ou \( I(d) \). Des variables aléatoires faiblement fixes sont par conséquent \( I(0) \). Beaucoup de variables (macro-économiques) peuvent être considérées comme des variables \( I(1) \). Si \( Z_t \sim I(1) \), alors \( \Delta Z_t \sim I(0) \).

Considérant le CLRM : \( Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \) supposons que \( Y_t \sim I(1) \) et en même temps \( X_t \sim I(1) \). Alors, puisque les variables \( I(1) \) dominent les variables \( I(0) \) dans une combinaison linéaire, généralement \( Y_t - \beta_1 X_t \sim I(1) \). Cependant, si le terme d’erreur aléatoire \( \varepsilon_t \sim I(0) \), alors \( Y_t - \beta_1 X_t \sim I(0) \). Dans une combinaison où le coefficient \( \beta_1 \) est unique, les variables \( Y_t \) et \( X_t \) sont dites co-intégrées.

La somme cumulative des écarts de la relation co-intégrante \( Y_t - \beta_1 X_t = 0 \) est alors nécessairement une variable \( I(1) \). Si cette nouvelle variable \( W_t \) est co-intégrée avec les variables co-intégrées d’origine \( (Y_t \text{ ou } X_t) \), alors on dit que \( Y_t \) et \( X_t \) sont multi-co-intégrées.

Les insuffisances du CLRM découlent du fait que la valeur attendue du terme d’erreur prévu, \( \hat{\varepsilon}_t \), à l’étape de l’erreur carrée de plus, est supposée être la même en tout point, en tout temps et quelle que soit l’observation (données transversales).

Il y a souvent une présomption d’*homooscédasticité* quant aux données transversales. Quand cette supposition n’est pas respectée, on parle alors d’*hétérooscédasticité*.

Le CLRM ignore le comportement hétéroscédastique du terme d’erreur \( \varepsilon_t \) et il produit des estimations de variance qui peuvent donner un faux sentiment de précision parce que la volatilité de la prévision est liée à la volatilité du terme d’erreur \( \varepsilon_t \). Pour modéliser cette variance non-constante (hétéroscédasticité), on utilisait la technique du Déplacement Moyen AutoRégressif (AR-MA).

AR est une technique qui permet de ramener une variable à ses propres valeurs précédentes : \( Y_t \) à \( Y_t, Y_{t-1}, \ldots, Y_{t-p} \).

MA exprime les observations présentes d’une variable en fonction des valeurs actuelles et précédentes du terme d’erreur aléatoire \( \varepsilon_t \). En combinant AR \((p)\) et MA \((q)\), on obtient ARMA \((p,q)\). \( p \) et \( q \) représentant l’ordre précédent de AR et MA.

On utilisait cette technique ARMA pour modéliser la volatilité variant dans le temps, ce qui a conduit au modèle d’Hétérooscédasticité Conditionnelle Auto-Régressive appelé ARCH. ARCH nous permet de prévoir la volatilité du terme d’erreur aléatoire ; il nous offre ainsi une mesure de Valeur à Risque. La variance du terme d’erreur aléatoire \( \varepsilon_t \) dans l’équation 5 peut être développée en fonction des valeurs actuelles et précédentes \( (\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+q}) \) comme suit :

\[
\sigma_t^2 = \theta_0 + \theta_1 \varepsilon_{t-1}^2 + \theta_2 \varepsilon_{t-2}^2 + \ldots + \theta_q \varepsilon_{t-q}^2
\]

Dans laquelle \( \sigma_t^2 \) est la variance de \( \varepsilon_t \).

Cette représentation MA \((q)\) de \( \sigma_t^2 \) fut plus tard étendue à une représentation ARMA de \( \sigma_t^2 \) bande 1, que l’on appelle modèle « Hétérooscédasticité Conditionnelle Auto-Régressive Généralisée » ou GARCH. GARCH évolua par régression d’une variable \((\varepsilon_t^2)\) sur ses propres valeurs précédentes (passées) \( \sigma_{t-1}, \sigma_{t-2}, \ldots, \sigma_{t-p} \).

\[
Y_t = \beta_0 + \sum_{j=1}^{N_q} \varphi_j Y_{t-j} + \sum_{k=1}^{N_p} \sum_{i=1}^{N_q} \alpha_{ki} X_{ki-i} + \varepsilon_t
\]

La variance du terme d’erreur aléatoire \( \varepsilon_t \) dépend non seulement des valeurs précédentes de \( \varepsilon_t \), mais aussi des valeurs précédentes de la variance \( \sigma(t-1, t-2, \ldots, t-p) \).

\[
Y_t = \beta_0 + \sum_{j=1}^{N_q} \varphi_j Y_{t-j} + \sum_{k=1}^{N_p} \sum_{i=1}^{N_q} \alpha_{ki} X_{ki-i} + \varepsilon_t
\]

\[
\sigma_t^2 = \theta_0 + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \tau_j \varepsilon_{t-j}^2 \quad (6)
\]

**GARCH et Prévision.**

La technique odd-var-garch saisit aussi la dynamique entre entités dans le réseau de valeur. La dynamique transversale réelle captée par les modèles VAR relie chaque variable aux valeurs passées de toutes les autres variables dans le modèle.

Pour modéliser une dynamique multi-variante de \( n = 2 \) en utilisant \( VAR(p) \), supposons que \( p = 1 \) (décalé de 1 période). L’équation 6 peut être étendue au type odd-var-garch pour modéliser deux entités et considérer seulement une période de décalage \((n=2, p=1)\) comme figuré dans l’équation 7.

\[
Y_{1t} = \beta_0 + \sum_{k=1}^{N} \sum_{i=1}^{N} \alpha_{ki} X_{ki-i} + \varphi_{11} Y_{1t-1} + \varphi_{12} Y_{2t-1} + \varepsilon_{1t}
\]
Dans le modèle var-garch, les interactions entre domaines de décision sont prises en compte en évaluant le coefficient \( \phi_{ij} \) qui se rapporte aux changements en \( Y_i \) compte-tenu de \( Y_j \).

L’impact du changement dans le terme d’erreur sur la variable fiable est appelé Fonction de Réponse d’Impulsion.

Appliqué à un grand volume de données précises fournies par l’AIT grâce au grid computing (sémantique), Odd-var-garch pourrait très bien devenir un outil pour des organisations recherchant une grande précision et dans les prédictions de chiffres en combinant les modèles basés sur équation (EBM) et les modèles basés sur agents (ABM).

**Le GRID**

Le modèle de simulation peut être optimisé par l’analyse basée sur données, qui en améliore ainsi la précision. De plus, les modèles basés sur équation (EBM) peuvent imposer d’évaluer des milliers de paramètres comme des objets uniques. En conséquence, l’avenir des analyses de données d’affaires peut avoir des besoins informatiques radicalement différents.

L’analytique de la supply chain exige des applications créatives, un partage des ressources à grande échelle et une informatisation à haut-débit qui puisse fonctionner dans un environnement dual EBM et ABM. Le Grid computing\(^7\) annonce un monde dans lequel les ressources partagées seront intégrées dans des « organisations virtuelles » (VO) réalisées via des portails Web (sémantiques) et des applications très largement distribuées.

Le Grid computing a fait un pas de plus vers le monde commercial lorsque le Projet Globus a sorti de nouveaux outils logiciels qui fusionnent grid et Web services, lesquels reviennent alors sous l’appellation : *Web Services Grid sémantique*.

---

\( Y_{2t} = \beta_0 + \sum_{k=1}^{K} \alpha_{kt} X_{kt-1} + \phi_{21} Y_{t-1} + \phi_{22} Y_{2t-1} + \varepsilon_{2t} \)

\( \sigma_{1t}^2 = \theta_0 + \sum_{i=1}^{q} \theta_i \varepsilon_{1t-i}^2 + \sum_{j=1}^{p} \tau_j \varepsilon_{t-j}^2 \)

\( \sigma_{2t}^2 = \theta_0 + \sum_{i=1}^{q} \theta_i \varepsilon_{2t-i}^2 + \sum_{j=1}^{p} \tau_j \varepsilon_{2t-j}^2 \)

---

7 - Technologie de mutualisation des ressources en réseau. N.d.T
8 - ou Computing distribué.
individuelles de garder le contrôle ultime sur leurs propres ressources.

La Figure 11 illustre les différentes couches de l’architecture grid établie sur un modèle « de sablier » dont « l’étranglement » définit un petit ensemble d’abstractions et de protocoles centraux sur lesquels de nombreux comportements différents de haut niveau (en haut) et de nombreuses technologies sous-jacentes (en bas) peuvent être mis en correspondance.

- Le « grid Fabric » est l’interface pour le contrôle « local ».
- La couche « Interface Fabric » fournit les ressources pour lesquelles l’accès partagé est géré par les protocoles grid.
- La couche « Connectivité » définit les protocoles centraux de communication et d’authentification requis pour les transactions de réseau spécifique grid.
- La couche « Ressource » s’ajoute à la couche Connectivité pour définir les protocoles régissant la gestion de la sécurité, le contrôle, la comptabilité et le paiement des opérations de partage faites sur les ressources individuelles.
- La couche « Collection » est donc globale par nature et peut mettre en œuvre une grande variété de comportements de partage sans imposer de nouvelles exigences sur les ressources.
- La couche finale du grid englobe les Applications « utilisateur » qui fonctionnent au sein d’une VO. Ces applications sont élaborées par échange de messages de protocoles avec le service ad hoc de manière à réaliser les actions souhaitées.

Prise dans sa globalité, cette architecture ouverte offre une multitude de combinaisons souples. Mais le grid computing concerne le partage contrôlé. Et actuellement, il y a un débat pour savoir si oui ou non le logiciel grid devrait définir les services du système de fonctionnement (OS – operating system) à installer sur chaque système participant.

Une grande partie de ce qui doit être réalisé par l’architecture grid ouverte est appelée « grid middleware ». On attend un grid computing haute performance pour qu’il soit possible d’effectuer des simulations à partir de plusieurs organisations différentes, ce qui permettrait d’entreprendre une simulation de l’ensemble du système.

La technologie et l’architecture constituent les « Grid Middleware services » nécessaires au soutien d’un ensemble complet d’applications dans un environnement de réseau distri-

bué tel qu’une « supply chain » ou un réseau de valeur.

Pourquoi l’interopérabilité est-elle une préoccupation fondamentale ? Imaginez les conséquences pour votre supply chain si vous deviez abandonner un fournisseur aux normes RosettaNet et vous approvisionner chez un fabricant de Shinzen (en Chine) qui n’a jamais entendu parler de RosettaNet ni de sa norme PIP’s ! Le Grid doit faciliter les relations de partage qui peuvent être mises en place entre des parties arbitraires de manière à intégrer de nouveaux participants de façon dynamique et rapide, au-delà des différences de plate-formes, de langages, de localisations géographiques et d’environnements de programmation. Les mécanismes ne servent pas à grand chose s’ils ne sont pas définis et développés pour être interopérables quelles que soient les frontières organisationnelles, les politiques opérationnelles et les types de ressources.
En conséquence, les protocoles sont également cruciaux pour l’interopérabilité. Le Web a révolutionné le partage de l’information en fournissant une syntaxe (http, html) et un protocole universels pour les échanges d’information. De son côté, le Grid exige des protocoles et syntaxes standards pour le partage de ressources générales. Les protocoles Grid, par définition, précisent à la fois la façon dont les éléments du système distribué interagissent les uns avec les autres (pour parvenir à un fonctionnement déterminé), et la structure de l’information échangée pendant cette interaction. Dans les versions futures, les ontologies pourraient jouer un rôle de plus en plus important dans la « compréhension » de l’information échangée via l’utilisation du langage sémantique (Web Ontology Language - WOL). L’accent mis sur les interactions externes plutôt qu’internes (software, caractéristiques des ressources) a des avantages pratiques importants pour le SCM, système dans lequel la fédération un peu distendue des partenaires tend à être floue.

De là, les mécanismes utilisés pour découvrir les ressources, établir l’identité, déterminer l’autorisation et lancer le partage doivent être suffisamment souples pour être établis et changés rapidement. Du fait que les organisations virtuelles (VO), type réseau de valeur, complètent plutôt qu’elles ne remplacent les institutions existantes, le partage de mécanismes ne peut pas imposer de changements substantiels aux politiques locales et il doit permettre à chaque établissement de garder le contrôle ultime sur ses propres ressources. Et comme les protocoles régissent l’interaction entre composants (mais pas leur mise en œuvre), le contrôle local est bien préservé. Le partenariat commercial de Shinzen, en conséquence, peut ne pas avoir à transformer son système b2b ou s’abonner à RosettaNet pour s’engager avec vous dans une relation d’affaires en tant que nouveau membre de la chaîne de valeur.

Pour que les outils de découverte soient utiles, les Web services exigent une grille ontologique et l’intégration avec le grid, de manière à évoluer en services du grid Sémantique. Enfin, pour parvenir à une fonctionnalité complexe, il pourrait être nécessaire d’avoir une convergence des ABM à l’intérieur des services du grid sémantique.

Agents
Nous allons étudier ce que sont les Agents et quelle est leur importance dans les systèmes décisionnels. Le concept d’Agents a germé au sein du MIT et s’est développé à partir de l’étude du fonctionnement de l’esprit humain et de l’intelligence [38]. Les origines des Agents sont donc inspirées de la biologie et inscrites dans le mouvement de l’Intelligence Artificielle (AI), mouvement qui s’imposa autour des années 1950. Quelques uns des tout premiers concepts d’Agents servirent à créer des « différence engines », l’unité logique peut-être la plus ancienne du réseau neuronal d’aujourd’hui (software). L’aspect « intelligence » des Agents est un paradoxe conceptuel. Par lui-même un Agent n’est pas intelligent ; en revanche, les « Agences » le sont. On insuffle aux Agents, qui fonctionnent collectivement comme un « essaim », une dynamique supérieure et des capacités adaptatives basées sur l’apprentissage, ce dont les algorithmes statiques basés sur équation manquent.

Leur autonomie permet aux Agents d’exécuter des opérations en réponse à des changements sans qu’il y ait besoin d’une intervention humaine permanente. De plus, les Agents qui partagent un même environnement avec d’autres Agents (Agences) dans un Système Multi-Agents9 sont capables de mobilité entre les différents environnements. Un logiciel ebm est incapable de s’adapter aux événements non-linéaires en intégrant de multiples sources d’information. Les Agences « Intelligentes » pourraient donc devenir un outil essentiel pour servir d’interface avec des sources multiples de données de manière à en extraire une information en temps réel. Les Agents utiles doivent être alignés sur ces concepts :

- correspondre à des « choses » dans le domaine en question plutôt qu’à des fonctions abstraites,
- être petit en masse, en temps (capable d’oublier) et en étendue,
- respecter une grande diversité,
- inclure un mécanisme dispersif,
- avoir des moyens de saisir et de partager ce qu’ils apprennent sur leur environnement,
- planifier et exécuter de manière concourente plutôt que séquentielle, et
- les Systèmes Multi-Agents (MAS) devraient être décentralisés.

Agents vs Equations
L’unité élémentaire des modèles basés sur équation (EBM) relie des données observa-

9 - MAS = Multi-Agent System.
bles dont les valeurs sont affectées par les actions d’individus multiples, alors que le fonctionnement d’un Agent n’influence pas directement sur la représentation de ces comportements individuels ; ce qui maintient donc des frontières entre individus. Cette différence donne aux modèles basés sur Agent (ABM) un avantage en terme de sécurité des applications commerciales :

- Dans un ABM, chaque entreprise a son propre groupe d’Agents. Il n’est pas nécessaire que les comportements internes de l’Agent soient visibles au reste du système. De cette façon, les firmes peuvent conserver l’information relative à leur fonctionnement interne.

- Une supply chain multi-sections complexe peut être découpée au niveau de l’Agent, permettant une modification des Agents individuels pour refléter une optimisation locale. Puisque ces agents font partie d’une Agence supérieure, la chaîne logistique complexe reflétera alors une optimisation globale – la migration d’un modèle de simulation vers un modèle de contrôle adaptatif est simple et directe.

Les Agents sont aussi pourvus d’un mécanisme d’oubli. Les outils de prévision types basés sur ebm utilisent comme alternative une moyenne pondérée des données historiques. Les ebm ignorent aussi le fait que la planification et l’exécution ne sont pas concourantes. Les Agents, qui cherchent à éviter le mode de fonctionnement « planification puis exécution », évaluent de manière probabiliste l’impact du travail sur son utilisation en fonction du temps, d’après l’information en provenance du client sur des durées acceptables.

On demande aussi que les systèmes puissent calculer plusieurs paramètres et les pondérer. En conséquence, les Agents présents dans l’architecture du système doivent être en lien logique avec les différents systèmes distribués.

Le Système Multi-Agent (MAS) nous offre la possibilité de répondre avec la finesse qui requiert de nombreuses connexions à la mémoire en succession rapide.

Agents en maintenance

Même quand une opération n’a que peu de variables critiques de mission, l’emploi d’Agents est crucial. Par exemple, telles opérations de maintenance sur un aéronef donné révèlent que l’incapacité à réparer les appa- reils ou les moteurs était due à un manque de pièces de rechange. La modélisation basée sur Agent des processus de maintenance permettrait d’intégrer la demande en provenance des multiples postes, en anticipant sur les défaillances de pièces par la gestion des métriques comme le MTBF (Mean Time Between Failure - Temps moyen écoulé entre deux pannes).

Agents dans la Fabrication

L’industrie aérospatiale commerciale fabrique peu de produits et vend à un ensemble de clients différents de celui du commerce de détail. Quelques pièces (modulaires) et composants servent à plusieurs modèles différents (variantes). Ce secteur, comme celui de l’industrie automobile, tire un bénéfice significatif de la vente subséquente de pièces de rechange et de services. En conséquence, les compagnies ont accès à une masse importante de données d’utilisation.

Prenons un exemple : la panne un peu trop rapide de deux pompes hydrauliques aux deux bouts du monde incite un Agent à étudier cette récurrence. Elles venaient toutes deux du même lot de fabrication. L’Agent pousse alors les techniciens de la maintenance à réaliser une analyse de vibration particulière, qui met en évidence un défaut dans l’ensemble du lot. Si l’Agent avait disposé, dans ce réseau de valeur, des données d’analyses de vibration faites par le fabricant, il aurait décelé cette tendance avant même qu’une seule pompe ne tombe en panne. Une analyse comparative implique l’accès à un traitement de données massives en un temps raisonnable. Les Agents pourraient accomplir de telles tâches rapidement et prédire, donc prévenir, une catastrophe potentielle.

Or, les réseaux de valeur avec points d’accès intégrés aux données distribuées via le Grid peuvent rendre disponible l’information nécessaire pour qu’un Agent puisse déceler une prédisposition à partir des données fabriquant, lot, date d’installation et heures d’utilisation.

Dans un autre scénario, prenons un système Agent qui opère dans le domaine exclusif de l’analyse des retours. Cet Agent remarque que le taux de retour des produits d’un fabricant a dépassé un certain niveau ces dernières semaines. Pourquoi ? Ce produit a une valeur relativement élevée, il pèse plus de 08 kilos et la plupart ont été envoyés à 800 kilomètres ou plus. L’alerte lancée par l’Agent arrive au
manager qui inspecte par hasard un conditionnement. Bingo ! Ce packaging est différent de celui des produits qui ont un taux de retour moins élevé. Un coup de téléphone confirme que le fabricant a changé tout dernièrement de fournisseur d’emballage.

Les Agents peuvent aussi être utiles dans le domaine de la vente. Ainsi, Dell permet à ses clients de configurer eux-mêmes leurs ordinateurs. L’allotissement (ou bundling) est une technique de marketing : on associe deux produits que l’on vend à un prix unique inférieur au montant normal de la vente séparée des deux produits. Ce système procure des recettes et bénéfices plus importants que si chaque élément était vendu séparément. Dell emmagasine des exaoctets d’information sur les habitudes d’achat des clients. Un Agent « analytique » est capable de repérer une tendance du type : 40% des clients qui achètent un supplément de mémoire prennent aussi un processeur grande vitesse. Un Agent « marketing » peut « parler » à l’Agent « prix » pour offrir des ristournes si la mémoire est achetée avec le processeur. Comme la tendance des choix pour les combinaisons (mémoire contre vitesse de processeur) évolue en fonction du profil démographique ou géographique, les données fournies par les Agents analytiques peuvent être utilisées par les agents « marketing » et « prix » pour offrir de nouvelles options de bundling (fixation dynamique du prix). Ce qui peut augmenter la demande pour la mémoire et accroître recettes et bénéfices. Il est possible aussi de cibler les clients potentiels pour commercialiser l’option de discount groupée.

En proposant trois options ou plus, sans parler des accessoires type appareils photos, MP3 ou imprimantes, le nombre de combinaisons possibles de produits augmente. Pour les Agents, analyser des montagnes de données, déceler les possibilités (multiples) de bundling et s’adapter aux fluctuations de la demande quasiment en temps réel est très simple. Ils le font beaucoup plus vite (des milliards de fois plus vite) que ne pourrait le faire un être humain ou un logiciel basé sur équations. La technique du bundling permet d’améliorer les ventes d’un stock de produits peu demandé ou en fin de vie (EOL- End-of-Life) avant d’introduire de nouvelles versions ou de nouveaux produits.

Technologies d’identification automatique (AIT).

Les AIT offrent aux outils le pouvoir d’acquérir des données sur l’objet. Les Codes Produits Electroniques (EPC) introduisent la standardisation d’un format pour données minimales sur les étiquettes RFID (étiquettes ULB) qui servent de référencement à des objets physi-
ques, dont l’avalanche des données pourrait être stockée sur Internet.

**RFID : les questions de protection de la confidentialité.**

Les avocats de la protection de la confidentialité sont anti-RFID parce qu’ils ne comprennent pas que le RFID, en tant que technologie, ne peut pas envahir la vie privée. Que les données consommateurs soient liées ou non aux données inventaire est seulement une décision de processus.

**ULB**

Plusieurs scénarios semblent indiquer que les bénéfices tirés de l’utilisation d’étiquettes ULB dépasseraient le coût d’utilisation d’étiquettes RFID passives à faible coût. Le métal ou le béton n’arrêtent pas l’ULB, et ce système est sûr. La capacité spatiale de l’ULB est 1000 fois supérieure à celle de l’UHF. Elle offre d’impressionnantes taux de débit qui fonctionnent malgré des interférences multi-voies. Dépourvue de restrictions de spectre pays, l’ULB peut devenir une norme mondiale pour des communications courte portée.

**Les réseaux de capteurs**

Les réseaux de capteurs sans fil sont un bon exemple de grille informatique diffuse. Ils sont autonomes et forment des réseaux sans fil qui se chargent de données à travers des besoins spécifiques qui peuvent être connectés aux stocks de données. Du fait de certains pouvoirs analytiques, le réseau de capteurs transmet des analyses de données plutôt que des données brutes de manière à fournir des réponses.

**Le web sémantique**

Le Web sémantique est un recueil de données en langage rdf (Cadre de Description de Ressources) qui décrit la signification des données à travers les liens aux ontologies\(^{10}\), qui agissent comme des glossaires décentralisés. Le Web sémantique peut offrir la transparence du processus au-delà des frontières du langage et de la géographie, même si les partenaires individuels accomplissent certaines fonctions d’une manière différente des autres.

Les agents seront capables de collecter le contenu du Web à partir de différentes sources, de traiter l’information et d’échanger les résultats avec d’autres programmes ou données, rendant ainsi plus facile le développement de programmes qui peuvent s’attaquer à des questions compliquées dont les réponses ne se trouvent pas sur une seule et unique page du Web.

Les ontologies peuvent aussi être utilisées d’une manière toute simple pour améliorer la précision des recherches sur le Web, par exemple par l’utilisation des règles de déduction.

Une importante facette du fonctionnement (de l’Agent) sera l’échange de « preuves » écrites en langage unificateur du Web sémantique. Quelques programmes peuvent déjà échanger des preuves de cette manière, en utilisant des versions préliminaires. La Figure 14 montre les couches du Web sémantique de Tim Berners-Lee.

**Le web sémantique adaptif dans le domaine de la santé ?**

Le partage de la connaissance dans le Web Sémantique via la conception d’Agents peut trouver un emploi dans le domaine de la santé :

L’utilisation de micro-antennes permet d’examiner l’expression des gènes, mais elle produit en même temps une masse ahurissante de données. Pour identifier un état de maladie, on doit identifier la position d’un gène sur la carte du génome humain. Une combinaison de l’information qui nous dit quels sont les gènes actifs et où l’on peut trouver ce gène actif sur le chromosome est un gain extraordinaire pour la bonne intelligence du meilleur mode d’administration du traitement.

Pour réaliser les connections logiques à partir de ce vaste amas de données, nous avons besoin d’Agents dans le Web Sémantique.

\(^{10}\) Ontologie = Spécification formelle de la représentation des concepts, objets et entités et de leurs relations. N.d.T.
Conclusion

Les entreprises utilisent des modèles pour optimiser les profits ou même prévoir l’action future. Mais ces modèles ont beau être omniprésents, un modèle tiré d’un domaine particulier n’interagit pas avec un autre. Pouvons-nous exploiter la puissance de modèles à données multiples ? Nous pourrions faire des prédictions basées non plus sur quelques paramètres comme dans un modèle basé sur équation (EBM), mais bien sur des milliards de faits et de fonctions différents que les modèles basés sur Agent (ABM) permettraient de contenir.

Les décideurs peuvent souhaiter profiter des éléments clés (outils et technologies) mentionnés dans ce document et rechercher des moyens de réaliser la convergence. Les progrès dans cette direction peuvent être mesurés par le degré de fusionnement significatif entre bits et atomes en fonction du processus.

Bibliographie


[17] Shubik, M., « Incentives, Decentralized Control, the Assignment of joint Costs and Internal Pricing » ; Management Science n°8 (1962) pages 325-343


THIS PAGE IS LEFT BLANK
The concept of the internet of things (IoT), and its variations [1], data of things (DoT), network of things (NoT) and the industrial internet of things (IIoT), may have started in earnest, circa 1988 with Mark Weiser of Xerox Palo Alto Research Center, who suggested that computers may “weave themselves into the fabric of everyday life” and influence the future of business (Scientific American, 1991). Weiser was referring to the discussion by Herbert Simon in his 1987 paper “The Steam Engine and the Computer: What makes technology revolutionary” where Herbert Simon frames his thoughts about the computer, “you have to make friends with it, talk to it, let it talk to you.”

Contrary to the media hype, neither the vision nor the implication of the Internet of Things (IoT) were the outcome of a marketing manager from P&G [2]. In the seminal paper (2000) THE NETWORKED PHYSICAL WORLD (MIT-AUTOID-WH-001) provided clues to the concept of the IoT [3] and the evolution of the industrial internet. Facts about IoT as well as the name “internet of things” [4] was discussed at a symposium at the MIT Sloan School.

Connecting physical world objects (made of atoms) with information (packaged as bits) may segue to another revolution, predicted by many, among them, Neil Gershenfeld. The current wave is referred to as the 3rd industrial revolution, in relation to the Information Age (2nd) and the First Industrial Revolution. In the EU, the trend is also referred to as Industrie 4.0 but these versions are also cyber-physical systems (CPS).

Revolutions are supposed to reshape the future. The current wave will not be an exception. It will generate friction, both social and economic. The clash of status quo with business not as usual. The collision between the imaginative vs those whose imagination is out of focus. The asphyxiation from old world ideas, versus, geographically-agnostic unbridled innovation, unleashing the wizardry of technology to leak into our daily lives.
Pundits, market observers and industries, are divided over the transformational capabilities of technologies, and the ubiquity of connectivity that IoT may necessitate. Social friction is erupting from erosion of privacy in its conventional format and the redefinition of privacy which challenges old world beliefs. There is justifiable concern about security, yet measurable reluctance to give up the benefits associated with either. Financial friction is evident both in industrial nations and emerging economies whenever unskilled labor is a part of the target. Labor, in general, abhors automation and robotics. It shrinks the demand for unskilled labor. The latter conjures a negative impression, even if the actual impact on social economy, as a whole, may be positive, over time.

These observations are centuries old. It will be repeated, again, albeit in different shades. According to economic historian Norman Poire, “the five centuries that span the years 1440 to 1939 were among the most dynamic in all of history. Many technological advances surfaced during that time, but three inventions stand above the rest as turning points in the direction of technology that led to decisive social change. The invention of the printing press by Johannes Gutenberg in 1440 spurred the arrival of the Information Revolution that spread the Renaissance throughout Europe. In 1609, Galileo Galilei’s telescope ushered in the Scientific Revolution and the Age of Reason. The Industrial Revolution and Marxism arrived shortly after James Watt unveiled his steam engine in 1769. In 1939, a fourth technological revolution began. In that year, John Atanasoff and his graduate student Clifford Berry, invented the electronic digital computer, and unwittingly, with it, the Second Information Revolution.” A little less than century later, we are on the cusp of yet another sea of change, which has the potential to evolve over at least a century. Perhaps, the fruits of progress, may sufficiently unfold or culminate around 2105-2120 [1].

The current (3rd) Industrial Revolution may spur the grand convergence of the industrial revolution with the information revolution, and may include many unknowns. Erik Brynjolfsson and Andrew McAfee at the MIT Sloan School of Management (Center for Digital Business) pontificates about the frictions that may surface from the 3rd revolution, namely, higher unemployment and rising inequality (Race against the Machine and The Second Machine Age). The incisive insight about inequality may be also found in the works of Joseph Stiglitz (The Price of Inequality) and Robert Reich (Inequality for All).

---

2 Disequilibrium due to Digital Transformation by Dr Shoumen Palit Austin Datta • shoumen@mit.edu
Translations available in French, Spanish, Italian, Chinese • https://dspace.mit.edu/handle/1721.1/111021
Brynjolfsson and McAfee revisit the discussion of higher unemployment which John Maynard Keynes described as “technological unemployment” in the 1930’s. Robert Frank revisits the same topic as technology-catalyzed “winner takes all” labor markets in 1990’s and in *The Darwin Economy*. Brynjolfsson and McAfee expect “our world will prosper on the digital frontier” but what about the *path* to the frontier? The road ahead is fraught with feuding nations, malnutrition, dysfunctional sanitation, inadequate education and poverty of energy. These factors are fueling acrimonious socio-economic frictions which are exacerbated by the pseudo-promise of a resplendent future due to IoT. The latter are perpetrated through glib marketing campaigns sponsored by smug corporations and consulting firms. These collusions are drowning the real benefits of connectivity and IoT.

The design and fruits of IoT may mature in proportion to our ability to interoperate between systems, and platforms, connecting objects and devices, from diverse ecosystems and environments, supporting different standards of operations, tools to acquire data, protocols, applications and analytics. It is impossible to expect the world to support any one common standard.

Hence, not standardization, *per se*, but the *interoperability between* major standards, is key to systemic adoption through *systems* integration of IoT related products and services, including the industrial internet. Industry-enabled open standards for APIs may connect with the global connectivity bus. SMEs may use common *platforms* to offer value-added services, bundled with analytical engines or niche applications. The general purpose technical design, implicated by the IoT metaphor, may offer value only when *systemic* deployment of connectivity improves data acquisition and decision support.

Ultimately, the ability to make sense of the data, information arbitrage or extraction of “intelligence” from data will drive the value proposition of connectivity. In the process of transforming data into information, one must be careful about the tools of the analytical trade lest we should lose the knowledge in the process.

Long before information technology opened the gate to an information society, it was Thomas Stearns Eliot, who asked, “Where is the Life we have lost in living? Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in the information?” (*The Rock* in Collected Poems 1909–1935, Faber & Faber, London, 1951).
There aren’t any simple and straight paths between information, knowledge and 
wisdom [6]. Whoever does not recognize this, will lose knowledge and wisdom, only to find 
a new, and a dangerous, form of stupidity. This form of stupidity has finally arrived and is 
now often referred to as “low hanging fruit” in the vernacular of modern day business [7].

The emphasis on low hanging fruits, short term return, prevalent in the business 
world, may impact the extent of acquisition of data. Inadequate investment may limit the 
tools necessary to accumulate critical mass of data. However, without sufficient data, the 
analytical tools may bumble around to unlock hidden patterns in the data. Patterns may be 
necessary in real time, if dynamic analytical engines (at the edge and core) may be a path to 
monetization of IoT. New sources of revenue may be created from micro-payments, based 
on pay-per-analytics model of information arbitrage. The latter may depend on intelligent 
predictive analytics to augment decision support for various levels of autonomic activities.

One lesson in data acquisition and analysis may be cryptic in the classical 
experiment in quantum mechanics, described as Young’s double-slit experiment. A 
variation of the experiment was performed at Hitachi Central Research Labs by Dr Akira 
Tonomura (1942-2012) which revealed [8] the build-up of interference pattern from single 
electrons. But, it was not observed until sufficient electrons were allowed to pass through 
the slit. The lesson from this experiment for business is obvious – running pilots and 
experiments on small scale may not offer appropriate outcomes or provide wrong 
indications. One cannot construct an elephant using the mouse as a model. This work is 
insightful because it suggests large scale deployments may be the key to extracting the 
value and significance of the tools and technologies which, when combined and converged, 
may provide solutions.

Another lesson may be found in the history of general process technologies, in 
particular, the strategies which enabled the spread of electricity (The Economic Future in 
Historical Perspective edited by P. A. David and M. Thomas. British Academy Publication, 
2006). Clayton Christensen’s (Innovators Dilemma) over-use of the word ‘disruptive’ is a 
hype. He usurped the original concept of general process technologies introduced by Paul 
David and Gavin Wright. Christensen mis-used the word and using shoddy data analysis 
mis-led the business world through his consulting effort to suggest everything is disruptive.
We have observed for the past couple decades the lack of systemic integration of RFID. As a consequence, we may not have sufficiently profited from high volume data from RFID tags. As a result, we may have failed to deliver adequate transparency within supply chains and the savings from the value chain remains far below what was anticipated. The lessons from the abandoned RFID initiative at WalMart (Is RFID dead? Florian Michahelles [2010] Auto-ID Labs, ETH) is not a failure of the technology but inadequate use of data and data extraction tools, to utilize actionable information, in the context of business processes.

IoT by design may learn from the history of electrification, and RFID, in order to find better ways to progressively penetrate our daily reality through systemic use, integrated with connectivity and applications. IoT must evolve from things to internet of systems (IoS). Connectivity between the ecosystems of systems may create the next tsunami of profitability. It will generate clamor for security, privacy, trust and ethics, as well as, policy. IoT by design, connectivity, and communications with objects and processes, will morph the way we interact, and behave, in our personal and professional lives, in the systems era.

The pursuit of autonomy in healthcare, transportation and manufacturing will create new solutions, old headaches and germinate new business models. The prediction of cancer at least a decade before it affects you, is not an illusion. Autonomous vehicle that parks itself and freight trucks delivering cargo without humans in the loop is yesterday’s news. The death of inventory and birth of distributed manufacturing on demand, at the edge, is the digital foundry of the future catalyzed by 3D printing of metals and non-metals. From heart valves to nano-satellites, from NASA-guided soil moisture active passive (SMAP) guidance for precision farming and graphene-purified arsenic-free desalinated drinking water, and everything euphoric in between (neurosynaptic web, neuromorphic chips), we have begun another 100-year journey, and may reach the summit during 2120.

According to Jeff Immelt, former CEO of GE, “in the future one expects an open, global fabric of highly intelligent machines that connect, communicate and cooperate with us. The Industrial Internet is not about a world run by robots, it is about combining the world’s best technologies to solve our biggest challenges. It is about economically and environmentally sustainable, energy, it is about curing the incurable diseases, and preparing our infrastructure and cities for the next 100 years.”
Economic friction is evident from the loss of middle-income repetitive tasks which may be largely and partially automated or can use online tools for completion. Bank tellers, store check-out clerks and even segments of K-16 teachers may be eliminated from the workforce in favor of ATMs, self-check-out kiosks and MOOCs. This is not only due to IoT, but the integration of computation with our daily lives, as predicted by Herbert Simon.

When, where, what and how, digital transformation will proceed will be determined by cost versus value. Even the veneer of altruism will be guided by transaction cost analysis and transaction cost economics (The Nature of the Firm by Ronald Coase, 1937) will be at the front and center of the discussion on the structure of infrastructure. Without the latter, connectivity will languish, without the ecosystem of networks, the value of haphazard connections will ruin the return on investment by increasing costs and decreasing value. Without sustained value and profitability, adoption and diffusion will remain a mirage.

IoT connectivity with a greater cross-section of objects and processes in addition to exposure to greater degree of monitoring (for example, in healthcare) will induce changes in behavior. Whether ubiquitous connectivity modifies rational versus irrational activity remains to be observed and analyzed (Thinking Fast and Slow by Daniel Kahneman). The outcome of such analyses must be taken into account when designing future products and services, for example, the wireless hospital of the future or MRI machines in hydrogen refueling stops or portable x-rays in medical huts in the Amazon. The utility may depend on the Gini coefficient, the socio-economic ethos (Scarcity by Sendhil Mullainathan) and its stage in socio-political development (Development as Freedom by Amartya Sen).

The prediction that connectivity will change behavior is rooted in the fundamental principles of particle physics. The observer effect, as it is called, refers to changes that the act of observation will have on a phenomenon being observed (not to be confused with the uncertainty principle proposed by Werner Heisenberg). The former may explain why one can sing in the shower but perhaps not in public.

It may be noted that combined behavior, especially, time-centricity of cyberphysical systems (hardware and software integrated with physical objects), may often change, if any one of the components are changed, even if the components are almost near-identical. The tryst with time may be difficult at times and time-spoofing is a global cyber-threat.
In addition to slow changes in behavior, economic re-equilibration will be sluggish because massive changes in our education system are necessary to optimize social consumption of the fruits of technology. No amount of technology or online courses will deter the spread of the rupture in our financial fabric unless we retrofit public education, re-install respect for academia, re-focus on rigor, re-invest in basic science research, restore the dignity due to educators and help to re-ignite the passion, latent in teachers.

The emerging supply chain of talent must include an abundance of girls who excel in mathematics, who can code and write cohesively, and may be multi-lingual with affinity for music. The trinity of math, music and multi-lingualism in the early years may be key for the pursuit of multi-disciplinarity of imagination, invention and innovation. It is essential that women pursue tertiary science, engineering, mathematics, economics and philosophy. We must not accept that 50% of the brain power is left out of the workforce, currently [9].

Educated women will help educate boys who are respectful and girls who are dignified. Integrity and dignity may lift humanity and induce innovation from distant crevices of the world, which will usher tectonic shifts in society and commerce. Mistakes will be made but failure is the new road to success, the new mantra for prosperity.

Distributed innovation demands an entrepreneurial approach and an assault on multiple levels, concurrently, rather than the mythical silver bullet solution (Innovation: The Attacker's Advantage by Richard Foster). The taxi cab industry vs uber, the hospitality industry vs airbnb and temp agencies vs oDesk/Elance/Upwork are examples [10].

Explosion of engineering tools has dramatically reduced the cycle time necessary to introduce innovation by vastly compressing the time from conception (development supply chain) to realization (fulfillment supply chain). The digital foundry made possible by 3D printing is one example. Industry giants must harness this explosion by giving away platforms in order to aggregate the intelligence that can run on open source platforms. The flow of micro-revenue from billions of “pings” on your product will be the differentiator. Value-added data-service will be related to intelligent analytics of data and delivery of actionable information to the point of use before data perishes. Real-time data, in a pay per use model of micro-payments will reduce the barrier to entry, in any market in the world, even when the per capita GDP may be struggling to reach $2 per day.
The dynamics of perishability of data changes when the accumulation of time series data is far more critical for predictive analytics (for example, healthcare) rather than data with short half-life (for example, mean time between failure (MTBF) metric for spare parts). Data transport and data storage are important in this business but consumers may be willing to pay only for real-time analytics. Consumers may expect raw data to be free. Consumers in impoverished nations may only focus on essential supply chains for living.

However, all advantages are temporary. The financial wisdom from micro-revenue earnings from leasing the platform is one reason why Apple opened up its “bus” for anyone to hop on (create applications). Apps pour in from all over the world. The app creator is a part of the economic avalanche, allowing Apple, as the channel master, to aggregate micro-payments using crowd innovation. Small data from millions is the reason why Apple is laughing all the way to the bank with the world’s largest database of payments, to the tune of 99 cents at a time. PayPal’s success fueled Tesla, which may give away the car to sell swappable graphene based batteries, and on-board services using software defined networking (SDN). The automobile may be the mobile electricity grid of the future and the substitute for the smart immobile grid for off-grid distribution of power.

Free products with pay-per-use micro-revenue based services is a proven business strategy (printers vs ink, mobile phones vs services, water coolers vs bottled water) to amplify micro-earnings, which will last not only the life cycle of the product but the life cycle of the user, the human. Micro-payments can induce a paradigm shift by moving away from product lifecycle (1-10 years) to the user’s lifecycle (5-50 years). The micro-payment for service is paid by the (human) user as long as the person lives and uses the product. Even if the product lifecycle expires, the product can be replaced with a new version. The user continues to pay the micro-payment for the service, agnostic of the product version.

The diffusion of ubiquitous connectivity will accelerate the adoption of IoT by design, which is expected to give rise to new products and services. The consumption of such goods and improvements in efficiency may generate a magnitude of economic growth which is inducing CEOs to be euphoric. According to GE, Cisco and others, the IoT and the industrial internet (IIoT) may add about $14 trillion to $19 trillion to the global economy. One must not view this corporate marketing propaganda with blinders.
A concomitant explosion of consumerism is necessary for such numbers to materialize. The billions who are writing on the wall or posting photographs to buoy the software market cap of the social media bubble, are in an earnings group which cannot afford the talking car or the avatar to manage the morning bed-tea or robotic laparoscopy. The educated consumer is the best customer. The bubble of the twitter frenzied social media economy may be limited by the amount of “energy under the ROC curve” and the irrational exuberance may fuel the next global recession which may be around the corner. With 2008 as the last recession, we expect the next one around 2022 based on the 14 year “boom-bust cycle” according to economist Finn Kydland (Nobel Prize in Economics, 2004).

In the US, several initiatives throughout industry and academia are emerging to address the next generation of advances in IoT, industrial internet and exciting possibilities from research in cyber-physical systems (CPS). Several consortia were formed in 2014-2015 with backing from market-leading companies. The EU has funded a massive multi-year program called Horizon 2020 to the tune of more than $100 billion to explore, at least in part, the potential growth of IoT and how to harvest the associated economic windfall.

But, it will be an egregious error on the part of the global leadership to be blinded by the economic projections and continue to polish the chrome without paying attention to tune the engine of education. Humanity needs dreamers [11] and education [12] is the salt (Salt by Mark Kurlansky) which drives inspiration, imagination, invention, innovation and implementation of ideas [13]. The education of a boy may change the fate of a man. The education of a girl may change the destiny of a nation.

---

1 Datta, S (2018) Connecting Atoms to Bits (Chapter 1) in Haphazard Reality – IoT is a Metaphor. https://dspace.mit.edu/handle/1721.1/111021
3 http://tinyurl.com/Industrial-Internet
6 https://hbr.org/2010/02/data-is-to-info-as-info-is-not
10 https://www.forbes.com/sites/elainepofeldt/2015/05/05/elance-odesk-becomes-upwork-today-odesk-brand-gets-phased-out/#6560a1ae51f5
Dynamic socio-economic disequilibrium catalyzed by the Internet of Things

Dr Shoumen Palit Austin Datta
Research Affiliate, School of Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 and Senior Vice President, Industrial Internet Consortium
shoumen@mit.edu

Letter from Academia

The technology based conceptualization of the internet of things (IoT) and the industrial internet may have started circa 1988 with the work of Mark Weiser of Xerox Palo Alto Research Center who suggested that computers may “weave themselves into the fabric of everyday life” and influence the future of business, as a consequence (Scientific American, 1991). The knowledge base Weiser was referring to is the discussion by Herbert Simon in his 1987 paper “The Steam Engine and the Computer: What makes technology revolutionary” where Herbert Simon frames his thoughts about the computer, “you have to make friends with it, talk to it, let it talk to you.”

Hence, contrary to the media hype, in progress, the vision of the Internet of Things (IoT) and its meaning did not germinate from a presentation (Ashton, 2009) at a retail product manufacturer. In 2000, the seminal paper entitled THE NETWORKED PHYSICAL WORLD (MIT-AUTOID-WH-001) gave birth to the concept of the IoT (Manyika et al., 2011; Sarma et al., 2000) and the evolution of the industrial internet. Facts about IoT as well as the name “internet of things” (The MIT Sloan CIO Symposium, 2013) was discussed at a recent symposium at the MIT Sloan School of Management.

Connecting physical world objects (made of atoms) with information (packaged as bits) may segue to another revolution, predicted by many, among them, Neil Gershenfeld. The current wave is often referred to as the third industrial revolution, in relation to the Information Age (second) and the first Industrial Revolution. In some quarters, the present trend is (also referred to as Industrie 4.0) the age of cyber-physical systems (CPS).

Revolutions are supposed to reshape things to come. The third wave will be no exception. It will generate friction, both social and economic. The clash of status quo with business not as usual. The collision between the imaginative versus those whose imagination is out of focus. The asphyxiation from old world ideas versus geographically-agnostic unbridled innovation unleashing the wizardry of technology to leak into our lives.

Pundits, market observers and industry players are divided over their belief regarding the transformational capabilities of technologies and the ubiquitous connectivity IoT necessitates. Social friction is erupting from erosion of privacy in its conventional format and the redefinition of privacy which challenges old world beliefs. There is justifiable concern about security yet there is measurable reluctance to give up the benefits associated with either. Financial friction is evident both in industrial nations
and emerging economies whenever unskilled labor is a part of the workforce. Labor, in general, abhors automation, which shrinks the demand for unskilled labor and creates a negative impact on the economy and society, as a whole.

But these are not new observations, in fact these are centuries old and will be repeated over and over, again, albeit in different shades. According to economic historian Norman Poire, “the five centuries that span the years 1440 to 1939 were among the most dynamic in all of history. Many technological advances surfaced during that time, but three inventions stand above the rest as turning points in the direction of technology that led to decisive social change. The invention of the printing press by Johannes Gutenberg in 1440 spurred the arrival of the Information Revolution that spread the Renaissance throughout Europe. In 1609, Galileo Galilei’s telescope ushered in the Scientific Revolution and the Age of Reason. The Industrial Revolution and Marxism arrived shortly after James Watt unveiled his steam engine in 1769. In 1939, a fourth technological revolution began. In that year, John Atanasoff and his graduate student Clifford Berry invented the electronic digital computer and unwittingly with it the Second Information Revolution.” A little less than century later, we are on the cusp of yet another sea of change.

The Third Industrial Revolution may spur the grand convergence of the industrial revolution with the information revolution and other existing unknowns.

Erik Brynjolfsson and Andrew McAfee at the MIT Sloan School of Management (Center for Digital Business) talks about the frictions that may surface from the third revolution, namely, higher unemployment and rising inequality (Race against the Machine and The Second Machine Age). The incisive insight about inequality may be also found in the works of Joseph Stiglitz (The Price of Inequality) and Robert Reich (Inequality for All).

Brynjolfsson and McAfee revisit the discussion of higher unemployment which John Maynard Keynes described as “technological unemployment” in the 1930’s. Robert Frank revisits the same topic as technology-catalyzed “winner takes all” labor markets in 1990’s and also in his book The Darwin Economy. Brynjolfsson and McAfee expect “our world will prosper on the digital frontier” but what about the path to the frontier? The road ahead is fraught with feuding nations, malnutrition, dysfunctional sanitation, inadequate education and poverty of energy. Taken together, these factors are already fueling glaring socio-economic frictions which may be exacerbated by the attributes necessary for the global diffusion of IoT (internet of things).

The fruits of IoT will depend on our ability to interoperate between systems, objects and devices in different environments supporting different standards of operations, protocols and applications. It is impossible to expect that the world will strive to support one common standard. Hence, not standardization per se but the interoperability between major standards will be the key to diffusion of the products and services of the IoT and the industrial internet which reaches into the domain of all things mechanical. Industry leaders must enable open standards for interfaces (APIs) where products from SMEs can plug into a common global bus to access the connectivity and add their value added services, analytical engines or enhance niche applications. The systemic deployment of open connectivity backbone is central to data acquisition and the spread of IoT.

Ultimately, the ability to extract intelligence from data will drive the value proposition of the connectivity. Transaction cost economics (The Nature of the Firm by Ronald Coase, 1937) of connectivity will determine the return on investment which will influence business adoption.

The emphasis on low hanging fruits, short term return, prevalent in the business world, may impact the extent of acquisition of data. Inadequate investment may limit
the tools necessary to accumulate critical mass of data. However, without sufficient data, the analytical tools may stumble to unlock hidden patterns in the data. The latter is necessary if real time dynamic analytical engines (at the edge and core) may be one path to monetization of IoT. New sources of revenue may be created from micro-payments based on pay-per-analytics model of information arbitrage which will use intelligent predictive analytics to augment decision support for semi-autonomous activities.

One lesson in data acquisition and analysis may be cryptic in the classical experiment in quantum mechanics described as Young’s double-slit experiment. A variation of the experiment was performed at HCRL (Hitachi Central Research Labs) by Dr Akira Tonomura (1942-2012) which revealed (HITACHI, 2015) the build-up of interference pattern from single electrons but it was not observed until sufficient electrons were allowed to pass through the slit. The lesson from this experiment for business is obvious – running pilots and experiments on small scale may not offer appropriate outcomes or provide wrong indications because you cannot construct an elephant using the mouse as a model. This work is insightful because it suggests large scale deployments may be the key to extracting the value and significance of the tools and technologies which, when combined and converged, may provide solutions.

One lesson may be found in the history of general process technologies, in particular, the strategies which enabled the spread of electricity (The Economic Future in Historical Perspective edited by P. A. David and M. Thomas, Oxford University Press, 2003). Clayton Christensen’s (The Innovators Dilemma) ‘disruptive’ is a hype based on the original concept of general process technologies (GPT) introduced during the era of electrification to indicate systemic integration versus “slap-on” ad hoc usage. Christensen mis-used the word and mis-led the business world using poor data analysis to suggest everything is disruptive.

We have observed for the past 15 years the lack of systemic integration of RFID. As a consequence, we have not sufficiently profited from the ability of RFID tags to acquire sufficient high volume data from a systems approach. As a result, we may have failed to deliver adequate transparency within supply chains and the savings from the value chain remains far below what was anticipated. The lessons from the abandoned RFID initiative at WalMart (Is RFID dead? Florian Michahelles [2010] Auto-ID Labs St. Gallen, ETH Zurich) is not a failure of the technology but an inadequate use of data tools in the context of the business process.

IoT (internet of things) may learn from the history of electrification and RFID in order to find better ways to progressively penetrate our daily reality through systems integration, connectivity and applications. IoT must evolve from things to internet of systems (IoS). Connectivity between the ecosystems of systems may create the next tsunami of profitability. In turn, it will generate even more clamor for security, privacy, trust and ethics related issues on our social policy agendas. IoT connectivity and communications with objects and processes will change the way we interact and behave in our personal and professional lives in the IoS era.

The pursuit of autonomy in healthcare, transportation and manufacturing will create new solutions, old headaches and germinate new business models. The prediction of cancer at least a decade before it affects you, is not an illusion. The autonomous vehicle that parks itself and a freight truck that delivers cargo without humans in the loop is yesterday’s news. The death of inventory and birth of distributed manufacturing on demand (dMOD) at the edge (dMODE) is the embryonic Manufacturing 5.0 catalyzed by 3D printing. From heart valves to nano-satellites and from NASA-guided soil moisture active passive (SMAP) guidance for precision farming and graphene-purified arsenic-free desalinated drinking water and everything euphoric in between (neurosynaptic web and neuromorphic chips), we have already begun the next 100-year journey.
According to Jeff Immelt of GE, “in the future one expects an open, global fabric of highly intelligent machines that connect, communicate and cooperate with us. The Industrial Internet is not about a world run by robots, it is about combining the world’s best technologies to solve our biggest challenges. It is about economically and environmentally sustainable, energy, it is about curing the incurable diseases, and preparing our infrastructure and cities for the next 100 years.”

Economic friction is evident from the loss of middle-income repetitive tasks which may be largely automated or can use online tools for completion. Bank tellers, store check-out clerks and even K-16 teachers will be eliminated from the workforce in favor of ATMs, self-check-out kiosks and MOOCs. This is not only due to IoS but the integration of computation with our daily lives, as predicted by Herbert Simon and Mark Weiser.

IoS connectivity with a greater cross-section of objects and processes in addition to exposure to greater degree of monitoring (for example, in healthcare) will induce changes in behavior with increasing diffusion of the internet of things. Whether ubiquitous connectivity modifies rational versus irrational activity remains to be observed and analyzed (Thinking Fast and Slow by Daniel Kahneman). The outcome of such analyses must be taken into account when designing future products and services, for example, the wireless hospital of the future or MRI machines in hydrogen refueling stops or portable x-rays in medical huts in the Amazon. The utility of these advances may depend on the socio-economic ethos of the society (Scarcity by Sendhil Mullainathan) and its stage in socio-economic development (Development as Freedom by Amartya Sen).

The prediction that connectivity will change behavior is rooted in the fundamental principles of particle physics. The observer effect, as it is called, refers to changes that the act of observation will have on a phenomenon being observed (not to be confused with the uncertainty principle proposed by Werner Heisenberg). The former may explain why one can sing in the shower but not in public.

It may be noted that combined behavior, especially, time-centricity of cyberphysical systems (hardware and software integrated with physical objects) changes, if any one of the components are changed, even if the components are almost near-identical. The tryst with time may be difficult at times.

In addition to slow changes in behavior, economic re-equilibration will be sluggish because massive changes in our education system are necessary to optimize social consumption of the fruits of technology. No amount of technology or online courses will deter the spread of the rupture in our financial fabric unless we retrofit public education, re-install respect for academia, re-focus on rigor, rejuvenate all aspects of scientific research, restore the dignity due to a teacher and re-ignite the passion expected from a teacher.

The emerging supply chain of talent must include an abundance of girls who excel in math, who can code and write cohesively. It is essential that women pursue higher level of science, engineering, mathematics, economics and philosophy. How can we accept that about 50% of the brain power is left out of the workforce?

Educated women will help educate boys who are respectful and girls who are dignified. Taken together, they will accelerate the massively parallel innovation from distant crevices of the world. The latter is already ushering tectonic shifts even in the most traditional businesses. The analysis-paralysis approach of the behemoths may lead to their extinction if they continue to remain oblivious of the fact that failure is the new road to success, failure is the new key to success and failure is the mantra for those who wish to succeed.

Distributed innovation demands an entrepreneurial approach and an assault on multiple levels, concurrently, rather than the mythical silver bullet solution.
(Innovation: The Attacker’s Advantage by Richard Foster). The taxi cab industry vs Uber, the hospitality industry vs Airbnb and temp agencies vs oDesk are bright examples. Explosion of engineering tools has dramatically reduced the cycle time necessary to introduce innovation by vastly compressing the time from conception (development supply chain) to realization (fulfillment supply chain). Industry giants must harness this explosion by giving away platforms in order to aggregate the intelligence that can run on open source platforms. The flow of micro-revenue from billions of pings on your product will be the differentiator and that value-added data-service will be related to intelligent analytics of data and delivery of actionable information to the point of use before the data perishes.

However, the dynamics of perishability of data changes when the accumulation of time series data is far more critical for predictive analytics (for example, healthcare) rather than data with short half-life (for example, mean time between failure (MTBF) metric for spare parts). Data transport and data storage are important in this business but consumers may be willing to pay only for real-time analytics. Consumers expect raw data to be free.

However, all advantages are temporary. The financial wisdom from micro-revenue earnings from leasing the platform is one reason why Apple opened up its “bus” for anyone to hop on (create applications). Apps pour in from all over the world. The app creator is a part of the economic avalanche by allowing Apple, as the channel master, to aggregate micro-payments using open innovation. Small data from millions is the reason why Apple is laughing all the way to the bank with the world’s largest database of payments, to the tune of 99 cents at a time. PayPal’s success fueled Tesla which may give away the car to sell swappable graphene based batteries and on-board services using software defined networking (SDN). The automobile may be the mobile electricity grid of the future substituting for the smart immobile grid for off-grid distribution of power. Free products with pay-per-use micro-revenue based services is indeed a proven business strategy (printers vs ink, mobile phones vs services, water coolers vs bottled water) to amplify micro-earnings, which will enjoy a long life and substantiate the value of long tails.

The spread of IoT and IoS is expected to give rise to new (Datta, 2015b) products and services. The consumption of such goods and improvements in efficiency may generate a magnitude of economic growth which is inducing CEOs to be euphoric. According to GE, Cisco and others, the IoT and the industrial internet (IIoT) may add about $14 trillion to $19 trillion to the global economy, over the next decade. An explosion of consumerism is necessary for such numbers to materialize. The billions who are writing on the wall or posting photographs to buoy the software market cap of the social media bubble are in an earnings group which cannot afford the talking car or the avatar to manage the morning bed-tea or robotic laparoscopy. The educated consumer is the best customer. The bubble of the twitter frantically social media economy may be limited by the amount of “energy under the curve” and the irrational exuberance may fuel the next global recession which may be just around the corner (2020-2022). With 2008 as the year of the last recession, we expect the next one around 2022 if the “boom-bust cycle” has a 14 year periodicity according to Finn Kydland (Nobel Prize in Economics, 2004).

In the US, several initiatives throughout industry and academia are emerging to address the next generation of advances in the IoT space, industrial internet, internet of systems (IoS) and the exciting possibilities from research in cyber-physical systems (CPS). Several consortia were formed in 2014-2015 with backing from market-leading companies. Several academic groups are leading the way with new inventions and innovation. The EU has funded a massive multi-year program called Horizon 2020 to the tune of more than $100 billion to explore the growth of IoT and help harvest the associated economic windfall.
But, it will be an egregious error on the part of the global leadership to be blinded by the economic projections and continue to polish the chrome without paying attention to tune the engine of education. Is a smart city (Datta, 2015a) really smart without smart citizens?

Humanity needs dreamers (Datta, n.d.) and education (Datta, 2014) is the quintessential salt (Salt by Mark Kurlansky) which acts as the purveyor of inspiration, imagination, invention, innovation and drives implementation of ideas (Datta, 2015b). The education of a boy may change the fate of a man. The education of a girl may change the destiny of a nation.

References


L’Internet des Objets : catalyseur des tensions sociaux-économiques et de l’innovation technologique

Dynamic Socio-Economic Disequilibrium catalyzed by Internet of Things

Dr Shoumen Palit Austin Datta
Research Affiliate, School of Engineering, Massachusetts Institute of Technology, Cambridge, uSA


Mots clés : xxxxx

The technology based conceptualization of the internet of things (IoT) and the industrial internet may have started circa 1988 with the work of Mark Weiser of Xerox Palo Alto Research Center who suggested that computers may “weave themselves into the fabric of everyday life” and influence the future of business, as a consequence (Scientific American, 1991). The knowledge base Weiser was referring to is the discussion by Herbert Simon in his 1987 paper “The Steam Engine and the Computer: What makes technology revolutionary” where Herbert Simon frames his thoughts about the computer, “you have to make friends with it, talk to it, let it talk to you.”

Keywords: xxxxx

Ainsi, contrairement aux idées véhiculées, la vision de l'Internet des Objets (IoT) et de sa signification même n'est pas le fruit d'une quelconque présentation industrielle. En 2000, l'article fondateur The networked physical world a donné naissance au concept d'IoT et à l'évolution de l'Internet industriel. Certains faits relatifs à l'IoT et l'expression même d'Internet des Objets ont donné lieu à des débats lors d'un colloque qui s'est tenu récemment à la Sloan School of Management du MIT.

La connexion des objets physiques (faits d'atomes) aux informations numériques (exprimées en bits) pourrait aboutir à une nouvelle révolution, ainsi que le prédit certains experts dont Neil Gershenfeld. On parle souvent de troisième révolution industrielle, en allusion à l'ére – ou à la seconde ère – de l'information et à la première révolution industrielle. Certains préfèrent parler d'industrie 4.0 ou d'ère des systèmes cyber-physiques.

Les révolutions sont supposées refondre l’avenir. Cette troisième vague ne devrait pas faire exception. Elle devrait susciter des tensions sociales et économiques. La collision entre deux mondes : celui de l’imagination et de la créativité face à celui de la productivité. D’un côté, les idées conventionnelles ; de l’autre, une innovation effrénée qui apporte la technologie au cœur de nos vies.

Les experts, les observateurs du marché et les industriels sont divisés quant aux capacités de transformation prétées aux technologies, mais aussi concernant la connectivité généralisée nécessaire à l’IoT. Certaines tensions d’ordre social apparaissent en raison de l’érosion de la sphère privée sous son accension traditionnelle et de l’émergence d’une nouvelle définition remettant en cause d’anciennes convictions. On peut raisonnablement craindre qu’il soit porté atteinte à la sécurité, mais on note une certaine réticence à renoncer aux bénéfices asociés. Des dissensions d’ordre financier émergent logiquement, tant dans les pays industrialisés que dans les pays émergents, dès lors qu’une part de la main-d’œuvre est non qualifiée. En général, l’automatisation tend à réduire les besoins en main-d’œuvre non qualifiée, avec un impact négatif sur l’économie et la société dans leur ensemble.

Ces observations ne sont pas nouvelles et l’on peut gager qu’elles se répéteront à l’avenir sous différentes nuances. D’après l’historien de l’économie Norman Poire, « les cinq siècles courant les années de 1440 à 1939 furent parmi les plus dynamiques de l’histoire ». De nombreuses avancées technologiques se firent jour sur cette période, mais trois inventions se démarquent et font figure de tournants car elles provoquèrent des bouleversements sociaux majeurs. L’invention de l’imprimerie par Johannes Gutenberg en 1440 a impulsé une véritable révolution de l’information, favorisant l’avènement de la Renaissance dans toute l’Europe. En 1609, le télescope de Galilée a ouvert la voie à la révolution Scientifique et à l’âge de la Raison. La révolution industrielle et le marxisme suivent de peu l’invention de la machine à vapeur par James Watt (1769). 1939 marque les prémices de la quatrième révolution technologique. Cette année-là, John Atanasoff et Clifford Berry mettent au point le premier ordinateur numérique électronique qui aboutit incendiennement à la deuxième révolution de l’information. Un peu moins d’un siècle plus tard, nous voilà au seuil d’un nouveau bouleversement.

La troisième révolution industrielle devrait être le prélude à la grande convergence de la révolution industrielle et de la révolution de l’information, avec un certain nombre d’inconnues.

Erik Brynjolfsson et Andrew McAfee de la Sloan School of Management du MIT (Center for Digital Business) prédissent d’éventuels troubles liés à la troisième révolution, avec notamment une augmentation du chômage et un accroissement des inégalités (Race against the Machine and The Second Machine Age). Cette vision perspicace concernant les inégalités est également présente dans les travaux de Joseph Stiglitz

3 - http://tinyurl.com/Industrial-Internet
Brynjolfsson et McAfee ont réexaminé le sujet du chômage conjoncturel abordé par John M. Keynes, et qualifié dans les années 1930 de « chômage technologique ». Le phénomène est également abordé par Robert H. Frank qui décrit, dans les années 1990, un marché du travail et une société « des vainqueurs » (winners-take-all) impactés par le développement des technologies (The Darwin Economy).

Pour Brynjolfsson et McAfee, « notre monde va prospérer sur le front du numérique ». Certes, mais quelle voie emprunter ? De sérieux problèmes socio-économiques persistent : nations discordantes, malnutrition, hygiène publique inefficace, systèmes éducatifs défaillants, pauvreté énergétique. Additionnés, ces facteurs alimentent les tensions socio-économiques, que la diffusion de l'Internet des Objets (IoT) à l'échelle mondiale peut excéder encore.

Les retombées de l'IoT vont dépendre de notre faculté à interagir au sein de systèmes, objets et dispositifs, dans des environnements variés, régis par divers modes de fonctionnement, protocoles et applications. Il est difficilement envisageable que notre monde s'aligne sur un seul et unique standard. Ainsi, ce n'est pas la standardisation en soi, mais l'interopérabilité entre les principales normes, qui sera la clé de la diffusion des produits et services de l'IoT, allant jusqu'à intégrer le domaine de la mécanique. Les grandes entreprise industrielles doivent autoriser des normes ouvertes pour les interfaces de programmation (APIs = Application Programming Interfaces) pour permettre aux produits de PME de rallier un protocole global commun, d'être connectés et d'accéder à des services à valeur ajoutée, voire de développer certaines applications de niche. Le déploiement systémique d'une connectivité ouverte est essentiel à la collecte des données et à la diffusion de l'IoT.

Enfin, la capacité à extraire des informations exploitable à partir de multiples données brutes impulse la valeur recherchée via la connectivité. L'économie des coûts de transaction (The Nature of the Firm de Ronald Coase, 1937) détermine le retour sur investissement de la connectivité, lequel influence à son tour l'adoption de cette dernière par les entreprises.

L'accent sur les résultats faciles et le court-termisme, très courant dans le monde des affaires, pourraient affecter l'importance des données collectées. Des investissements inappropriés pourraient limiter les outils nécessaires pour recueillir une masse critique de données. Toutefois, faute de données suffisantes à traiter, les outils d'analyse pourraient ne pas être en mesure de mettre au jour certaines tendances cachées. C'est pourtant indispensable pour que les outils dynamiques d'analyse en temps réel constituent une étape vers la monétisation de l'IoT. De nouvelles sources de revenus peuvent apparaître via des micro-paiements basés sur un modèle d'arbitrage informationnel de type pay-per-analytics qui permettrait de recourir à des analyses prévisionnelles intelligentes facilitant la prise de décisions pour les activités semi-autonomes.

Les déploiements de grande envergure sont la clé pour conférer de la valeur et de l'importance aux outils et technologies, lesquels apportent des solutions satisfaisantes lorsqu'ils sont employés de manière convergente et coordonnée.

Un autre enseignement tiré de l'histoire des GPT (general process technologies) concerne les stratégies de diffusion de l'électricité (The economic future in historical perspective de P.A. David et M. Thomas, 2003). L'expression « de rupture » (disruptive), employée largement par Clayton Christensen (The Innovators Dilemma), peut découler du concept de GPT (general process technologies), apparu à l'époque de l'électrification pour indiquer une intégration systémique par opposition à un usage ponctuel.

Au cours des 15 dernières années, on constate une faible intégration systémique des systèmes RFID. Dès lors, la capacité des balises RFID à recueillir des volumes suffisamment importants de données a été négligée. D'où certainement un manque de transparence au sein des supply chain d'une part, mais aussi des économies inférieures au niveau prévu pour la chaîne de valeur. L'abandon du projet de RFID par WalMart (Is RFID dead? Florian Michahelles, 2010) ne révèle pas l'échec de la technologie en question, mais souligne plutôt une utilisation inappropriée des outils de gestion des données dans le contexte des processus métiers.

Peut-être faudrait-il s'inspirer de l'histoire de l'électrification et de la RFID pour faire
en sorte que l’IoT prenne de plus en plus de place dans notre quotidien, via l’intégration des systèmes, la connectivité et les applications. On doit passer d’un IoT à un Internet des systèmes (Internet of Systems – IoS en anglais). La connectivité entre « écosystèmes de systèmes » déclencheur peut-être la prochaine manne de profits. En revanche, devrait apparaître également une revendication pour plus de sécurité, de protection de la vie privée, plus de confiance et de déontologie, thèmes liés à la politique sociale. La connectivité propre à l’IoT ainsi que la communication avec les objets et processus devraient affecter nos comportements et notre façon d’interagir dans les sphères personnelle et professionnelle, à l’ère de l’Internet des systèmes.

L’autonomisation en cours dans les domaines de la santé, des transports et de l’industrie va faire resurgir de vieux casse-têtes, tout en donnant lieu à de nouvelles solutions et à de nouveaux modèles d’affaires. Prévoir les risques d’un cancer dix ans au moins avant son apparition ne constitue plus une vue de l’esprit. Les véhicules autonomes capables de se garer automatiquement, les camions de transport délivrant les marchandises sans être obligés de se garer automatiquement, les camions de transport délivrant les marchandises sans intervention humaine sont désormais une réalité. L’abandon des stocks au profit d’une fabrication répartie proche de la demande, préfigure la Fabrication 5.0, facilitée par le développement de l’impression 3D.

Les valves cardiaques, les nano-satellites, le satellite télé guidé SMAP (Soil Moisture Active Passive) de la NASA, le télé guidage pour l’agriculture de précision, l’eau dessalée et purifiée (sans graphène, ni arsenic), le web neuro-synthétique et les puces neuro-morphiques : autant d’exemples attestant que le futur est déjà là !

D’après Jeffrey R. Immelt, PDG de General Electric, « à l’avenir, on peut s’attendre à une économie planétaire ouverte de machines très intelligentes qui interagissent, communiquent et coopèrent avec nous. L’Internet industriel ne renvoie pas à un monde gouverné par des robots. Il s’agit plutôt d’associer les meilleures technologies mondiales pour relever nos plus grands défis. Il s’agit de promouvoir les sources d’énergie renouvelables, compétitives et respectueuses de l’environnement, de soigner les maladies incurables, mais aussi de préparer nos infrastructures et nos villes pour les cent prochaines années. »


L’innovation distribuée exige une démarche entrepreneuriale et une offensive simultanée sur plusieurs fronts, plutôt qu’une prophétique solution miracle (Innovation: The Attacker’s Advantage de Richard Foster). À titre d’exemples éclairants : l’industrie du taxi contre Uber, l’industrie hôtelière contre Airbnb, ou encore les agences de placement temporaire contre oDesk. Le récent foisonnement des outils d’ingénierie a considérablement réduit la durée du cycle nécessaire à l’introduction d’une innovation en compressant la durée entre conception et réalisation. Les géants de l’industrie doivent mettre à profit cette explosion en se donnant les moyens de regrouper les renseignements et informations utiles sur des plateformes ouvertes. Le flux de micro-revenus générés par les milliards de pings des produits fera la différence et le service de données à valeur ajoutée sera lié aux analyses de données intelligentes et à la transmission d’informations exploitables au point d’utilisation avant que ces données ne disparaissent. Le transfert et le stockage de données sont importants dans ce secteur, mais les consommateurs pourraient rechigner à payer, hormis pour les analyses en temps réel. En effet, tout le monde s’attend aujourd’hui à ce que les données soient gratuites.

Toutefois, les avantages ne sont que temporaires. Les micro-revenus générés par le leasing de sa plateforme ont incité Apple à ouvrir son « bus » et à permettre à tout un
chacun de créer ses propres applications, déclenchant un afflux d’applis en provenance du monde entier. Tout créateur d’applications participe de la manne économique, permettant à Apple, d’engranger des micro-payements (99 cents en moyenne) via sa plateforme d’innovation ouverte. C’est ainsi que la société engrange d’importants revenus grâce à la plus importante base de données sur les paiements au monde. Le succès du service de paiement PayPal a bénéficié au constructeur automobile Tesla qui pourrait mettre de côté son activité première pour commercialiser des batteries au graphène et des systèmes embarqués utilisant les technologies de réseau programmables SDN (Soft-ware Defined Networking). L’automobile pourrait devenir le réseau électrique mobile du futur, remplaçant le réseau statique classique pour une distribution hors réseau d’électricité. Les produits gratuits offrant des services facturés à l’usage (imprimantes / cartouches d’encre ; téléphones mobiles / applications ; fontaines d’eau fraîche / eau en bouteilles) constituent en fait une stratégie commerciale éprouvée destinée à accroître les micro-revenus sur le long terme.

Le développement de l’IoT et de l’IdS devrait favoriser l’apparition de nouveaux produits et services. La consommation de tels biens et les améliorations en matière d’efficacité devraient favoriser la croissance économique, ce qui ne peut que réjouir les dirigeants. D’après General Electric, Cisco et d’autres, l’IoT et l’Internet industriel des Objets (IoT en anglais) devraient injecter de quatorze à dix-neuf milliards de dollars dans l’économie mondiale au cours de la décennie à venir. Une forte accélération du consumérisme est toutefois nécessaire pour parvenir à de tels chiffres. Les milliards d’individus qui publient des messages ou postent des photographies sur les réseaux et maintenant ainsi le marché des logiciels à flot tout en entretenant la bulle financière autour des médias sociaux font partie d’une certaine tranche de revenus et n’ont pas les moyens de se payer une voiture robot téléguidée, un robot ménager à tout faire ou une laparoscopie robotisée. Le consommateur aguerri constitue le meilleur client. La bulle financière qui s’est constituée autour des médias sociaux (engouement frénétique pour Twitter…) peut être atténuée par la quantité « d’énergie en dessous de la courbe ».

Aux États-Unis, diverses initiatives se font jour dans les milieux industriels et universitaires qui s’intéressent aux prochaines étapes des avancées touchant à l’IoT, l’IoS, l’Internet industriel, mais également aux possibilités prometteuses de la recherche en matière de systèmes cyber-physiques (SCP). Plusieurs consortiums se sont constitués courant 2014-2015, avec le soutien de sociétés leaders du marché. Des groupes de chercheurs universitaires ouvrent la voie via diverses inventions et innovations. L’UE a financé un vaste programme pluriannuel, baptisé Horizon 2020, à hauteur de quelque 100 milliards de dollars, dans le but d’analyser la croissance de l’IoT et de mieux profiter des avantages économiques ainsi générés.

En termes de leadership mondial, ce serait une grave erreur de se laisser éblouir par les projections économiques en faisant fi du moteur que constitue l’éducation. Les smart cities ou « villes intelligentes » peuvent-elles se passer de citoyens instruits ? L’Humanité a besoin de rêveurs et l’éducation fournit un terreau propice à l’inspiration, à l’imagination, à l’inventivité, et à la mise en pratique des idées (cf. Salt by Mark Kurlansky).

---

Dr Shoumen Palit Austin Datta est....

Dr Shoumen Palit Austin Datta is Senior Vice President, Industrial Internet Consortium and Senior Scientist, Massachusetts General Hospital, Harvard Medical School.

---

El desequilibrio socioeconómico consecuente de la industria de internet de las cosas

The consequent socioeconomic imbalance of the Internet industry of Things

Resumen

La evolución de la tecnología hacia internet de las cosas y la industria de internet está provocando cambios muy importantes en nuestra sociedad. Los cambios que se producen en las relaciones humanas así como el uso de los objetos físicos conducen a nuevas formas de comportamiento y por tanto de aprendizaje. La intimidad y la privacidad se convierten en un retó difícil de mantener y la seguridad se ve amenazada en todas sus vertientes. Sin renunciar a las ventajas del desarrollo tecnológico y a las nuevas fuentes de información en este artículo se plantea un desarrollo social responsable y consciente. Este artículo realiza una reflexión desde la evolución histórica de la tecnología industrial al desarrollo del internet de las cosas reconociendo los problemas a los que el desequilibrio socioeconómico consecuente

Summary

The evolution of technology towards the Internet of Things and the Internet industry is causing major changes in our society. The changes that occur in human relationships and the use of physical objects lead to new ways of behavior and therefore of learning. The intimacy and privacy have become a difficult challenge to maintain and our security is threatened in all its aspects. Without giving up the advantages of technological development and new information sources, the present article tries to develop a responsible and conscious social development. This article makes a reflection from the historical evolution of industrial technology, to the development of the Internet of Things recognizing the problems that a new society faces and proposing as a best solution,
se enfrenta la nueva sociedad y proponiendo como mejor solución la educación desde la igualdad y el respeto a los valores establecidos.

**Palabras clave:** internet de las cosas, desarrollo tecnológico, tecnología industrial, economía, desarrollo.

the education based on equality and respect for the established values.

**Keywords:** Internet of Things, technological development, industrial technology, economy, development.
Introducción

La conceptualización basada en la tecnología de internet de las cosas (IoT) y la industria de internet dio comienzo alrededor de 1988 con la obra de Mark Weiser del Centro de Investigación Xerox Palo Alto, quien sugirió que los computadores pueden “tejerse a sí mismos en el tejido de la vida diaria” e influir de forma determinante en el futuro de los negocios (Scientific American, 1991). Mark Weiser se refería al artículo seminal “El motor de vapor y el ordenador: ¿qué hace revolucionar la tecnología?”, donde Herbert Simon (1987) enmarca sus pensamientos sobre la tecnología del futuro: “tienes que hacer amistad con él, hablar con él, dejar de hablar tú”.

Por tanto, contrariamente a la exageración de los medios, la visión de internet de las cosas (IoT) y su significado no nacen en una presentación (Ashton, 2009) de la venta de un producto. En 2000, un primer trabajo titulado “The networked physical world” dio a luz el concepto de internet de las cosas (Manyika et al., 2011; Sarma, Brock y Ashton, 2000) y la evolución de internet industrial. Tanto el concepto como el término internet de las cosas se discutieron en un simposio en la Escuela del MIT Sloan de Administración (The MIT Sloan CIO Symposium, 2013) donde Sarma acredita Ashton para el nombre de internet of the things.

La actual conexión de los objetos del mundo físico (átomos) y la información (bits) puede llevar a otra revolución, ya predicada por muchos, como Neil Gershenfeld. Esta corriente a menudo se denomina tercera revolución industrial en relación con la primera revolución industrial y la era de la información (segunda revolución).

En algunos sectores, la tendencia actual se denomina la edad de los sistemas ciberfísicos (cyber-physical system), aunque también aparece el término industria 4.0 en referencia al mismo fenómeno.

Evolución de la sociedad en el entorno tecnológico

A lo largo de la historia, las revoluciones han surgido para remodelar el futuro y esta tercera revolución no será una excepción. Como cualquier otra revolución, generará controversia, tanto social como económica. Sin embargo, el choque del statu quo no será como hasta ahora. La sociedad ya está empezando a sufrir la colisión entre la imaginación y las ideas del viejo mundo que ahora se ven asfixiadas ante una innovación desenfrenada. La magia de la tecnología se filtra en nuestra vida cotidiana sin apenas control. Los expertos, observadores de los mercados y los actores de la industria, tienen opiniones divididas respecto de la capacidad de transformación de las tecnologías y la conectividad que necesita el internet de las cosas.

La sociedad está cambiando el formato convencional de intimidad y el concepto de privacidad se redefine y desafía los valores establecidos. La preocupación por la seguridad está muy justificada, pero todavía hay personas reticentes a renunciar a los beneficios individuales asociados a su pérdida. Los individuos clamaron por su seguridad y su privacidad, pero al mismo tiempo se desprecian de estos valores cuando se les ofrece a cambio servicios gratuitos (commodities) como aplicaciones de mensajería, juegos o suscripciones a redes sociales.

En los países industrializados, es evidente la fricción financiera al surgir una fuerza de trabajo emergente de la que buena parte es trabajo no cualificado. Por otra
parte, la automatización de procesos reduce la demanda de mano de obra no cualificada, y crea un efecto negativo en la economía y en la sociedad en su conjunto. Sin embargo, la automatización no abolirá la necesidad de mano de obra en absoluto.

Estas observaciones no son nuevas. De hecho se repiten una y otra vez desde hace siglos, aunque en diferentes tonos. De acuerdo con el economista e historiador Norman Poiré, “los cinco siglos que abarcan los años 1440 hasta 1939 fueron los más dinámicos de toda la historia”. En dicho periodo, surgieron muchos avances tecnológicos pero algunos inventos destacan por encima del resto como puntos de inflexión en la dirección de la tecnología que llevaron a cabo un cambio social decisivo. La invención de la impresión por Johannes Gutenberg en 1440 estimuló la llegada de la revolución de la información que extendió el Renacimiento en Europa. En 1609, el telescopio de Galileo Galilei marcó el comienzo de la revolución científica y la edad de la razón. La revolución industrial y el marxismo llegaron poco después de que James Watt diera a conocer su máquina de vapor en 1769. En 1939, una cuarta revolución tecnológica dio comienzo. En ese año, el doctor John Atanasoff y su estudiante graduado Clifford Berry inventaron la computadora digital electrónica y sin querer con ello la segunda revolución de la información. Un poco menos de un siglo después estamos en la cúspide de un nuevo mar de cambios, donde nos preparamos con fervor para el próximo tsunami.

La tercera revolución industrial puede estimular la gran convergencia de los sectores de la revolución industrial con la revolución de la información y otras incógnitas existentes. En el Instituto Tecnológico de Manchester, Erik Brynjolfsson y Andrew McAfee de la Sloan School of Management del MIT (Center for Digital Business) hablan de dos enfrentamientos que pueden surgir de la tercera revolución: el aumento del desempleo y el aumento de la desigualdad (Race against the Machine y The second machine age). La visión incisiva acerca de la desigualdad también se encuentra en las obras de Joseph Stiglitz (El precio de la desigualdad) y Robert Reich (Desigualdad para todos).

Brynjolfsson y McAfee revisitan la discusión de un mayor desempleo que John Maynard Keynes describe como “desempleo tecnológico” en el 1930. Robert Frank vuelve a debatir el mismo tema en los mercados laborales en 1990 y los problemas causados por la tecnología (winner takes all) en su libro La economía Darwin.

Brynjolfsson y McAfee predicen “nuestro mundo va a prosperar en la frontera digital”, pero ¿qué pasa con la ruta de acceso a la frontera? El camino por delante está lleno de naciones enfrentadas, desnutrición, sanidad disfuncional, educación y pobreza de la energía. Tomados en conjunto, estos factores ya están alimentando fricciones socioeconómicas evidentes que pueden ser exacerbadas por los atributos necesarios para la difusión global de IoT (internet de las cosas).

Los frutos de IoT dependerán de nuestra capacidad de interoperación entre sistemas, objetos y dispositivos en diferentes ambientes así como diferentes estándares de operaciones, protocolos y aplicaciones. No cabe esperar que el mundo se esfuerce por apoyar un estándar común en un primer momento. La opinión generalizada es que las normas son una buena idea, pero son difíciles de aplicar, puesto que nadie quiere usar normativas escritas por otros. Un sistema de normalización internacional en IoT desarrollado desde la Organización Internacional de Normalización (ISO
‘International Organization for Standardization’), no parece viable a corto/medio plazo. Por tanto, no la normalización en sí, sino la interoperabilidad entre los principales estándares puede ser la clave para la difusión de los productos y servicios de IoT y la internet industrial que alcanza en el dominio de todas las cosas mecánicas. Los líderes de la industria deben permitir estándares abiertos para interfaces (application programming interface), donde los productos de las pymes pueden conectarse a un sistema global común para acceder a la conectividad y poner sus servicios de valor añadido, motores de análisis o mejorar aplicaciones. El despliegue sistemático de datos abiertos (open data) y su conectividad es fundamental para la adquisición de datos y la difusión de IoT (López, 2014, 2015).

En última instancia, la capacidad inteligente a partir de la extracción de datos impulsará la propuesta de valor de la conectividad. La economía de los costes de transacción (naturaleza de la firma de Ronald Coase, 1937) determinará el retorno de la inversión que puede influenciar la adopción por negocios y empresas.

Los procedimientos tradicionales de retorno a corto plazo, y muy frecuentes en el mundo de los negocios, pueden retardar la adquisición de datos. Una inversión inadecuada puede limitar las herramientas necesarias para acumular una masa crítica de datos. Así, sin datos suficientes, las herramientas analíticas pueden fallar en el descubrimiento de patrones ocultos en los datos. Sin embargo, los procesos de descubrimiento de patrones son muy necesarios especialmente cuando los motores de análisis dinámicos en tiempo real son un camino a la monetización de IoT. En este aspecto, se pueden crear nuevas fuentes de ingresos a partir de micropagos basados en el modelo de pago por análisis de arbitraje de la información (pay-per-analytics model), que utiliza análisis predictivo inteligente para mejorar el apoyo a la toma de decisiones semiautónoma.

Una lección de adquisición y análisis de datos puede extraerse del experimento clásico de la mecánica cuántica, que describe el experimento de la doble rendija de Young. Una variación de este experimento se realizó en el Hitachi Laboratorio Central de Investigación (HITACHI) en Kokubunji por el doctor Akira Tonomura (1942-2012). Este segundo experimento (HITACHI, 2015) reveló la existencia de un patrón de interferencia a partir de electrones individuales, pero no se observó hasta que se permitió el paso de un número de electrones suficientemente grande a través de la rendija. La lectura de este experimento para los negocios es obvia, porque la ejecución de experimentos a pequeña escala puede no ofrecer resultados adecuados e incluso podría proporcionar indicaciones equivocadas (puesto que no se puede construir un elefante tomando un ratón como modelo). En este sentido, este trabajo es intuitivo, porque sugiere despliegues a gran escala como clave para extraer el valor de los datos y la importancia de las herramientas y las tecnologías que, cuando se combinan y convergen, pueden aportar excelentes soluciones.

En la historia de los procesos tecnológicos, se pueden encontrar ejemplos como las estrategias que permitieron la difusión de la electricidad (The economic future in historical perspective, editado por P. A. David y M. Thomas, Oxford University Press, 2003). En el trabajo The innovators dilemma, de Clayton Christensen, el término disruptive es una exageración basada en el concepto original de las tecnologías de procesos generales (general process technologies) introducidas durante la era de la
electrificación para indicar integración sistémica frente al uso ad hoc del término slap-on. Christensen hizo un mal uso del término disruptive así como un erróneo análisis de datos que condujo a concluir que todo es perjudicial (Andrew King, MIT Sloan School of Management).

**Cambios en la organización**

Durante los últimos quince años se ha observado una falta de integración sistémica de etiquetas RFID (radio frequency identification) de identificación por radiofrecuencia, muy habituales en todo tipo de establecimientos para identificar los productos. A pesar de su uso muy extendido, la industria no se ha beneficiado suficientemente de la capacidad de las etiquetas RFID para adquirir grandes volúmenes de datos desde un enfoque de sistemas. Como resultado, es posible que hayamos fracasado para ofrecer una adecuada transparencia en las cadenas de suministro y el ahorro del valor de la cadena se mantiene muy por debajo de lo previsto. En Walmart (Is RFID dead? Florian Michahelles [2010] Auto-ID Labs St. Gallen, ETH Zurich), un trabajo sobre el abandono de la iniciativa de las tarjetas RFID, se sugiere que no es un fracaso de la tecnología, sino un uso inadecuado de la herramienta de datos en el contexto del proceso de negocio. Es un fracaso de cambio organizacional.

Internet de las cosas puede aprender de la historia de la electrificación y la tecnología por radiofrecuencia RFID, con el fin de encontrar mejores maneras de penetrar progresivamente nuestra realidad cotidiana a través de sistemas de integración, conectividad y aplicaciones. Internet de las cosas debe evolucionar a internet de los sistemas (IoS). La conectividad de sistemas en nuevos ecosistemas puede generar el siguiente tsunami de rentabilidad. A su vez, aumentarán aún más los problemas de seguridad, privacidad, confianza y ética relacionados con nuestros programas de política social. La conectividad y las comunicaciones con los objetos y procesos cambiarán nuestra forma de interactuar y nuestro comportamiento tanto en la vida personal como en el entorno profesional en la era de la internet de los sistemas (IoS).

**Nuevos modelos de negocio**

La búsqueda de la autonomía en el cuidado de la salud, el transporte y la fabricación creará nuevas soluciones y germinará con nuevos modelos de negocio. La predicción de enfermedades como el cáncer con una década de anticipación no es una ilusión. Los vehículos con sistemas autónomos de estacionamiento, los camiones de carga y la distribución de paquetería sin interacción humana son ya noticias del pasado. La muerte del inventario y el nacimiento de la fabricación y distribución por demanda (dMODE) es el embrión de la Manufacturación 5.0 con el desarrollo de los sistemas de impresión 3D. Con el desarrollo de válvulas cardiacas, nanosatélites, la medición de la humedad del suelo por la NASA (SMAP) para la agricultura de precisión, la fabricación de agua potable desalinizada libre de arsénico y otros avances tecnológicos (neurosynaptic web y neuromorphic chips), podemos decir que ya ha comenzado el próximo viaje de cien años.

Según Jeff Immelt, CEO de General Electric,

> En el futuro se espera un tejido abierto y global de máquinas altamente inteligentes que se conectan, se comunican y...
La fricción económica es evidente tras la automatización de tareas repetitivas que hasta ahora han supuesto puestos de trabajo de ingresos medios: el personal de los bancos, el personal que atiende las ventanillas e incluso los profesores de cualquier grado en el sistema educativo (K-16 Teachers) están siendo sustituidos por cajeros automáticos, quioscos de autoservicio y los cursos abiertos masivos (MOOC). Esto no solo se debe a internet de los sistemas (IoS), sino también a la integración de la computación en nuestra vida cotidiana, según lo predicho por Herbert Simon en 1987 y popularizado por Mark Weiser en 1991.

Internet de los sistemas (IoT)

La conectividad de internet de los sistemas (IoS) con una sección transversal mayor de los objetos y procesos además de la exposición a mayor grado de supervisión (por ejemplo en cuidado de la salud) inducirá cambios en el comportamiento con el aumento de la difusión de internet de las cosas. La evolución de la actividad racional frente a la irracional debe ser observada y analizada (Thinking fast and slow, Daniel Kahneman). El resultado de estos análisis debe tenerse en cuenta al diseñar futuros productos y servicios, por ejemplo, hospitales wireless del futuro o máquinas de resonancia magnética con repostaje automático de hidrógeno o máquinas de rayos X portátiles para ser usadas en las chozas en la Amazonia. La utilidad de estos avances puede depender de los valores socioeconómicos (Scarcity, Sendhil Mullainathan) y de la etapa de desarrollo del país o la región donde se aplica (Development as freedom, Amartya Sen).

La predicción de que la conectividad va a cambiar el comportamiento de las personas tiene sus raíces en el principio fundamental de la física de partículas. El “efecto del observador”, como se le llama, se refiere a los cambios que el acto de la observación produce cuando se observan fenómenos (y que no debe confundirse con el principio de incertidumbre propuesto por Werner Heisenberg). El primero puede explicar por qué cantamos en la ducha, pero no en público.

Cabe señalar que el comportamiento combinado de hardware y software integrados con objetos físicos cambiará si alguno de los componentes cambia, incluso si se sustituye por otro prácticamente idéntico. Además de los cambios en el comportamiento, el reequilibrio económico será lento, porque son necesarios cambios sustanciales en nuestro sistema de educación para optimizar el consumo social de los frutos de la tecnología. Ninguna cantidad de cursos de tecnología o cursos en línea evitará la propagación de la ruptura en nuestro tejido económico a menos que reformemos la educación pública y restuemos el respeto a las autoridades educativas, nos centremos en rejuvenecer todos los aspectos de la investigación científica, en restablecer la dignidad a los maestros y volver a encender la pasión que se espera de un profesor.
La cadena de suministro emergente de talento debe incluir la gran cantidad de niñas que sobresalen en matemáticas y que pueden programar y escribir coherente-mente. Es esencial que las mujeres persigan el nivel más alto de la ciencia, la ingeniería, las matemáticas, la economía y la filosofía. ¿Cómo podemos aceptar que alrededor de 50 % de la energía del cerebro se quede fuera de la mano de obra especializada?

Las mujeres educadas ayudarán a educar niñas y niños respetuosos y dignos. El trabajo conjunto acelerará la innovación masiva en paralelo de las lejanas grietas del mundo en el que ya se está marcando el comienzo de movimientos tectónicos, incluso en las tareas más tradicionales. El enfoque de análisis-parálisis de los gigantes puede conducir a su extinción si siguen siendo ajenos al hecho de que el fracaso es el nuevo camino hacia el éxito, el fracaso es la nueva clave para el éxito y el fracaso como mantra para el éxito de la economía emergente.

La distribución de la innovación exige un enfoque empresarial y un salto a múltiples niveles en lugar de soluciones mágicas (Innovation: The attacker’s advantage, Richard Foster). La industria del taxi frente a servicios como UBER (www.uber.com), la industria hotelera frente a aplicaciones como Airbnb (www.airbnb.com) y las agencias de trabajo temporal frente a oDesk (www.upwork.com) son buenos ejemplos. La explosión de herramientas de ingeniería ha reducido drásticamente el tiempo necesario para introducir el ciclo de innovación por lo que ha comprimido enormemente el tiempo desde la concepción (desarrollo de la cadena de suministro) a la realización (cumplimiento de la cadena de suministro). Los gigantes de la industria deben aprovechar esta explosión y facilitar la creación de plataformas de código abierto donde concurran las ideas y el desarrollo inteligente de aplicaciones. El flujo de microingresos de miles de millones de clics en la web de un producto será el elemento diferenciador y el servicio de datos de valor añadido estará relacionado con análisis inteligente de datos y la entrega de información útil antes que los datos perezcan. La inteligencia de negocio es el factor clave en la economía resultante.

Sin embargo, el carácter perecedero de los datos cambia cuando la acumulación de datos en el tiempo es mucho más crítica para un análisis predictivo (por ejemplo en medicina, en la evolución de algunas enfermedades) y no de datos con corta vida media (por ejemplo el tiempo medio entre fallos MTBF [mean time between failures] o la métrica de piezas en una turbina). El transporte y almacenamiento de datos son importantes, pero los consumidores pueden estar dispuestos a pagar solo por su análisis en tiempo real y los resultados de aplicar las técnicas inteligentes. Los consumidores esperan que los datos en bruto sean abiertos y libres (gratis).

Aun así, todas estas ventajas son temporales. La sabiduría financiera de microingresos de arrendamiento de las plataformas es una de las razones por las que Apple abrió su bus para que cualquiera puede subir sus aplicaciones. Las aplicaciones llegan de todas partes del mundo como una avalancha económica y le permite a Apple, como propietaria del canal, agregar micropagos mediante la innovación abierta. Small data de millones de fuentes es la razón por la que Apple sonríe con la base de datos más grande del mundo de pagos, con una suma de 99 centavos de dólar en cada pago. El éxito de PayPal se ve alimentado por firmas como Tesla, que es capaz de regalar el coche que fabrica para vender baterías intercambiables a base de grafito y servicios a
bordo (utilizando software definido en red, software defined networking). El futuro del automóvil Apple podría usar la red eléctrica móvil en lugar de la red inmóvil inteligente para distribución de energía off-grid. Los productos gratuitos con servicios basados en pago por el uso (pay per use) son ya una estrategia de negocio madura (impresoras frente a compra de tinta, teléfonos móviles frente a servicios, cafeteras frente a cápsulas de café) para ampliar las microganancias y garantizar una larga vida a los productos. Rolls Royce venderá horas de uso (thrust hours) en lugar de motores de avión; Google venderá planes de ahorro de energía solar (solar savings plans) en lugar de alertas y Bosch no venderá sensores físicos sino la inteligencia de los sistemas de sensores. Se espera que la propagación de IoT y IoS dé lugar a nuevos productos y servicios (Datta, 2015b). El consumo de estos bienes y las mejoras en la eficiencia puede generar una magnitud de crecimiento económico que está induciendo mucha euforia en los CEO (chief executive officer) de las empresas. Según General Electric, Cisco y otras compañías, IoT y la internet industrial (IIoT) pueden sumar entre US$14 000 millones y US$19 000 millones a la economía global en la próxima década. Para que estas cifras se materialicen, debe ocurrir una explosión en el consumo. Los cientos de millones de personas que están escribiendo en las redes o publicando fotografías para mantener a flote la burbuja de la capitalización del mercado de los medios sociales pertenecen a un grupo social con ingresos insuficientes para pagar coches que hablan o una laparoscopía robótica. El consumidor educado es el mejor cliente. La burbuja económica de Twitter como medio social puede estar limitada por la cantidad de “energía bajo la curva” y la exuberancia irracional puede alimentar la próxima recesión global que podría estar a la vuelta de la esquina (2020-2022). Con 2008 como el año de la última recesión, se espera la próxima alrededor de 2022 si el “ciclo de auge y caída” todavía exhibe la periodicidad de catorce años, como explica Finn Kydland (Premio Nobel de Economía 2004).

**Reflexiones finales**

En los Estados Unidos están surgiendo varias iniciativas en la industria y el mundo académico para hacer frente a la próxima generación de avances en el espacio de IoT, IIoT, IoS y las interesantes posibilidades de investigación en sistemas ciberfísicos. En el periodo 2014-2015, se formaron varios consorcios de líderes del mercado. En investigación, varios grupos académicos están liderando el camino con nuevos inventos e innovación. También la Unión Europea ha financiado un programa de varios años denominado Horizonte 2020 por una suma de más de US$100 000 millones para explorar el crecimiento de IoT y ayudar a la cosecha de la bonanza económica asociada. Pero será un error atroz por parte de los líderes globales cegarse por las proyecciones económicas sin prestar atención a un ajuste necesario del motor de la educación. ¿Una smartcity puede ser realmente inteligente sin ciudadanos inteligentes (Datta, 2015a)? ¿Serán las smartcity también ciudadades sostenibles (López, 2014, 2015; López et al., 2015)?

La humanidad necesita soñadores (Datta, 2009) y la educación (Datta, 2014, 2014b) es la quinta esencia (Salt: A world history, Mark Kurlansky) que actúa como el proveedor de inspiración, imaginación, invención, innovación y unidades de implementación de las ideas (Datta, 2015b). La educación de un niño puede cambiar el
destino de un hombre. La educación de una niña puede cambiar el destino de una nación.

Referencias

https://www.youtube.com/watch?v=44MLERLwxig
Come l’IoT potrà cambiare la società e l’economia

di Shoumen Palit Austin

La connessione degli oggetti alle informazioni apre la strada alla trasformazione digitale. Un cambiamento che porterà conseguenze rivoluzionarie per la società e per il business, e che non sarà indolore. Potrà essere guidato nel suo svolgimento positivo se l’educazione favorirà l’acquisizione della cultura necessaria

Ospitiamo su Industria Italiana un articolo del Dr Dr Shoumen Palit Austin Datta Senior Member, MIT Auto-ID Labs, Research Affiliate, Department of Mechanical Engineering, Senior Scientist, MDPnP Lab, Department of Anaesthesiology, Massachusetts General Hospital, Harvard Medical School. La traduzione dell’articolo intitolato “Dynamic socio-economic disequilibrium catalysed by the Internet of Things” già pubblicato nel Journal of Innovation Management vol. 3, n. 3, 2015 è stata elaborata dalla Dr. Marianna Marchesi Architetto e Ricercatore, Member of ISAAAC Association - Global Advisory Group e rivista dalla nostra redazione

Lettera dal mondo accademico

L’evoluzione della tecnologia basata sul concetto di Internet delle Cose (IoT, dall'inglese Internet of Things) e l’IoT industriale (IIoT) ha avuto inizio più o meno nel 1988 grazie al lavoro di Mark Weiser al Research Center di Xerox a Palo Alto. In questo lavoro l’autore avanzò l’ipotesi che i computer potessero intrecciarsi alla vita di ogni giorno e, conseguentemente, influenzare il futuro dell’imprenditoria (Scientific American, 1991). Le conoscenze a cui Weiser alludeva si riferivano alla discussione di Herbert Simon nel suo articolo del 1987 “The Steam Engine and the Computer: What makes technology revolutionary” [Il motore a vapore e il computer: cosa rende la tecnologia rivoluzionaria] nel quale Simon fotografa la sua riflessione sul computer in questo modo: “you have to make friends with it, talk to it, let it talk to you” [devi fare amicizia con esso, parlare ad esso, lasciare che esso parli a te].

Quindi, contrariamente a quanto sostenuto con una esagerata pubblicità dai media, l’idea dell’IoT e il suo significato non sono nati da una presentazione (Bernardi, Sarma and Traub, 2017) a un imprenditore di prodotti a dettaglio.E’ stato nel 2000, con l’articolo fondamentale intitolato “The Networked Physical World” (MIT-AUTOID-WH-001) che è stata data origine al concetto di IoT (Manyika et al., 2011; Sarma et al., 2000) e all’evoluzione dell’IoT. Le vicende
relative all’IoT - incluso il nome “internet of things” - sono state oggetto di discussione in un recente simposio presso la Sloan School of Management all’MIT (The MIT Sloan CIO Symposium, 2013).

**Una nuova rivoluzione**

La connessione degli oggetti fisici (composti di atomi) alle informazioni (confezionate in bit) potrebbe dare seguito a un ‘altra rivoluzione che in molti, tra cui Neil Gershenfeld, hanno predetto. L’attuale cambiamento viene spesso indicato come la terza rivoluzione industriale in relazione alla seconda, l’era dell’Informazione, e alla prima rivoluzione industriale. In alcuni settori, l’attuale tendenza (anche definita come Industria 4.0) è indicata come l’età dei sistemi ciber-fisici (CPS, dall’inglese cyber-physical system).

Si suppone che le rivoluzioni modifichino il futuro che ci aspetta e la terza rivoluzione non farà eccezione: genererà attriti sia sociali che economici. Un conflitto con l’attuale status quo che sarà fuori dall’ordinario. Determinerà uno scontro tra visionari e persone dotate di scarsa immaginazione, nonché tra vecchie idee asfittiche contrapposte all’innovazione prorompente e geograficamente delocalizzata, con il risultato che l’incantesimo tecnologico si riverserà nelle nostre vite.

Esperti, osservatori di mercato e figure di spicco del mondo industriale sono divisi nelle loro convinzioni relativamente alla capacità di trasformazione delle tecnologie e alla connettività diffusa di cui ha bisogno l’IoT. Sta emergendo con forza dirompente un attrito sociale dovuto al deterioramento della privacy per come viene intesa tradizionalmente e una sua ridefinizione che sfida le convinzioni vecchio stamato. C’è una preoccupazione giustificata che riguarda la sicurezza e, contemporaneamente, c’è riluttanza a rinunciare ai benefici ad essa associati. Un attrito di carattere finanziario si evidenzia sia nelle nazioni industrializzate sia nelle economie emergenti ogni qual volta manodopera non qualificata diventa parte della forza lavoro. La manodopera, in generale, aborre l’automazione, la quale restringe la domanda di lavoro non qualificato e crea un impatto negativo sull’economia e sulla società nel loro insieme.

Ma queste non sono osservazioni nuove, sono tendenze secolari che si ripeteranno ancora e ancora, anche se con differenti sfumature. Secondo lo storico economista Norman Poire, i cinque secoli compresi tra il 1440 e il 1939 furono i più dinamici di tutta la storia. Molti progressi tecnologici apparvero in quel periodo, ma tre invenzioni emersero rispetto al resto come punti di svolta nell’evoluzione della tecnologia, e condussero ad un decisivo mutamento sociale. L’invenzione della
macchina tipografica ad opera di Johannes Gutenberg nel 1440 spronò l’arrivo della Rivoluzione dell’Informazione che diffuse il Rinascimento attraverso l’Europa. Nel 1609 il telescopio di Galileo Galilei diede inizio alla Rivoluzione Scientifica e all’Epoca dell’Illuminismo. La Rivoluzione Industriale e il Marxismo arrivarono poco dopo l’invenzione della macchina a vapore da parte di James Watt nel 1769. Nel 1939 apparve una quarta innovazione tecnologica: in quell’anno John Atanasoff e il suo studente Clifford Berry inventarono il computer elettronico digitale e inconsapevolmente con esso avviarono la seconda Rivoluzione dell’Informazione. Poco meno di un secolo più tardi, siamo ancora alle soglie di un’altra fase di cambiamento. La terza rivoluzione industriale potrebbe incoraggiare decisamente la grande convergenza della rivoluzione industriale con la rivoluzione dell’informazione, con tutte le incognite esistenti.

Gli effetti sul mercato del lavoro

Erik Brynjolfsson e Andrew McAfee alla Sloan School of Management (Center for Digital Business) dell’MIT si interrogano sugli attriti che potrebbero manifestarsi in seguito alla terza rivoluzione: maggiore disoccupazione e una crescente disuguaglianza (Race against the Machine e The Second Machine Age). Una analisi incisiva relativamente alle disuguaglianze è presente anche nei lavori di Joseph Stiglitz (The Price of Inequality) e Robert Reich (Inequality for All). Brynjolfsson e McAfee ripercorrono la discussione sulla elevata disoccupazione che John Maynard Keynes aveva definito come “disoccupazione tecnologica” negli anni ’30. Robert Frank negli anni ’90 e anche nel suo libro The Darwin Economy rivisita lo stesso tema e lo individua come quello che, catalizzato dalla tecnologia, la fa da padrone sui mercati del lavoro.

Brynjolfsson e McAfee prevedono che il nostro mondo “prospererà sulla frontiera digitale”. Ma cosa si può dire riguardo il percorso per raggiungere questa frontiera digitale? La strada che si staglia di fronte a noi è piena di insidie, tra nazioni in lotta, malnutrizione, scarse condizioni igieniche, inadeguata istruzione e scarsità di energia. Considerati assieme, già di per sé questi fattori hanno la capacità di alimentare in maniera lampante attriti socio economici, che potrebbero essere ulteriormente esacerbati dalle caratteristiche ascritte alla diffusione globale dell’IoT.

L’impossibile definizione di un unico standard per l’IoT

I frutti che possono derivare dall’IoT dipenderanno dalla nostra abilità di scambiare e condividere informazioni tra sistemi, oggetti e apparecchi in differenti ambienti, che supportano diversi standard operativi, procedurali e
applicativi. Non è possibile coltivare l’aspettativa che il mondo cerchi a tutti i costi di supportare un unico comune standard. Quindi nessuna standardizzazione di per sé, ma lo scambio e la condivisione di informazioni tra i principali standard saranno la chiave per la diffusione dei prodotti e dei servizi dell’IoT e dell’IIoT, (Industrial IoT) che metteranno capo al dominio di tutte le cose meccaniche. Le aziende leader nel campo industriale devono abilitare standard aperti per le interfacce (APIs) attraverso le quali i prodotti delle piccole e medie imprese (SMEs) possono connettersi ad un bus, a un canale globale comune (un insieme di collegamenti fisici utilizzati in comune da più elementi per scambiare informazioni da elaborare ) in modo da accedere al collegamento e includere i loro servizi a valore aggiunto i loro computer o valorizzare le loro applicazioni di nicchia. Lo sviluppo sistematico di una dorsale open di connessione è decisivo nell’acquisizione di dati e nella diffusione dell’IoT.

In definitiva sarà la capacità di estrarre informazioni dai dati a guidare la value proposition della connessione. I dati economici fondamentali relativi ai costi di compravendita nella connessione (The Nature of the Firm by Ronald Coase, 1937) determineranno il ROI, la resa del capitale investito, che influenzerà l’adozione del business da parte delle aziende.

L’enfasi sulla ricerca di risultati che siano immediatamente conseguibili e di un ritorno a breve termine, prevalenti nel mondo aziendale, può avere un impatto sulla misura dell’acquisizione dei dati. Investimenti inadeguati possono limitare gli strumenti necessari per accumulare una massa critica di dati. In ogni caso, senza un numero sufficiente di dati, gli strumenti di analisi potrebbero fallire nell’ evidenziare dei modelli non evidenti. Un’azione di questo tipo risulta necessaria se i motori di analisi dinamica in tempo reale (all’edge e al core ) rappresentano una delle strade per monetizzare l’IoT. Nuove fonti di fatturato potrebbero essere create da micro pagamenti basati su modelli “pay-per-analytics di “information arbitrage”, che utilizzeranno analisi intelligenti predittive per implementare il supporto decisionale nel caso di attività semi autonome.

**Per l’IoT è necessario un dispiegamento su larga scala**

Troviamo una lezione sulla acquisizione e l’analisi dei dati occultata nel classico esperimento di meccanica quantistica descritto come il cosiddetto esperimento della doppia fessura di Young.

Una variazione dell’esperimento venne eseguita al HCRL (Hitachi Central Research Labs) dal Dr. Akira Tonomura (1942-2012). Si constatò lo sviluppo di modelli di interferenza derivanti da singoli elettroni, ma il fenomeno non si
manifestò se non quando un numero sufficiente di elettroni poté oltrepassare la fessura (HITACHI, 2015). La lezione che ne deriva per il mondo degli affari è ovvia: condurre test ed esperimenti a una piccola scala potrebbe non offrire adeguati risultati o potrebbe fornire indicazioni sbagliate perché non si può costruire un elefante utilizzando un topo come modello. Questo lavoro è rivelatorio, perché dimostra come il dispiegamento su larga scala può essere la chiave per estrarre valore significativo da strumenti e tecnologie che possano fornire soluzioni quando sono comбинate e convergenti.

Un’ altra lezione da tener presente può essere individuata nella storia delle tecnologie generali di processo, in particolare nelle strategie che resero possibile la diffusione dell’eletricità (The Economic Future in Historical Perspective edited by P. A. David and M. Thomas, Oxford University Press, 2003). Il termine ‘disruptive’ utilizzato da Clayton Christensen in “The Innovators Dilemma” è una iperbole basata sul concetto originale di tecnologie generali di processo (GPT, dall’inglese general process technologies) introdotto durante l’epoca dell’elettrificazione per indicare una integrazione sistemica, contrapposta a un uso ad hoc, di momento. Christensen usò impropriamente la parola e fuorviò il mondo delle aziende utilizzando un’analisi povera di dati per far pensare che ogni cosa è disruptive.

Negli ultimi 15 anni abbiamo osservato la mancanza di integrazione sistemica del RFID (dall’inglese Radio-Frequency Identification; in italiano identificazione a radiofrequenza). Di conseguenza, non abbiamo tratto sufficiente profitto dalla capacità dei tag RFID di acquisire un elevato volume di dati tramite un approccio di sistema. Come risultato, non siamo stati in grado di fornire adeguata trasparenza all’interno delle supply chains e i risparmi derivanti dalla catena di valore rimangono ben al di sotto delle attese. Le considerazioni che si possono trarre dall’ abbandono dall’iniziativa RFID a WalMart (Florian Michahelles [2010] Is RFID dead? Auto-ID Labs St. Gallen, ETH Zurich), non riguardano il fallimento di una tecnologia, ma un utilizzo inadeguato dei data tools nel contesto del processo di svolgimento del business.

**Dall’ IoT (Internet of things) all’ IoS (Internet of Systems)**

L’ IoT può imparare dalla storia dell’elettrificazione e da quella dell’ RFID così da individuare i modi migliori per farsi progressivamente spazio nella nostra realtà quotidiana attraverso l’integrazione dei sistemi, la connessione, e gli applicativi. L’IoT deve evolversi dall’internet delle cose all’internet dei sistemi (IoS). Dalla connessione tra ecosistemi di sistemi può arrivare il prossimo tsunami di redditività. A sua volta questo generarà ancora più dibattito nelle
nostre agende di politica sociale relativamente a sicurezza, privacy, fiducia e aspetti relativi all’etica. La connettività IoT e la comunicazione con gli oggetti e i processi cambierà la maniera in cui interagiremo e ci comporteremo nelle nostre vite personali e professionali, nell’era dell’IoS.

Il perseguimento dell’autonomia nell’assistenza sanitaria, nel trasporto e nella produzione creerà nuove soluzioni e vecchi grattacapi, e genererà nuovi modelli di business. La possibilità di diagnosticare un tumore almeno un decennio prima che si manifesti non è un’illusione. Il veicolo autonomo che si parcheggia da solo e un carico merci che arriva a destinazione senza la presenza di umani nel processo è ormai storia vecchia. La morte delle giacenze di magazzino e la nascita delle produzioni distribuite su richiesta (dMOD) marginali (dMODE) rappresenta l’embrione della Manifattura 5.0, catalizzata dalle stampanti 3D. Con le valvole cardiache fino ai nano-satelliti, con l’assistenza attiva e passiva (SMAP) per l’umidità del suolo messa a punto dalla NASA ai fini dell’agricoltura di precisione, con l’acqua potabile desalinizzata priva di arsenico e depurata dal grafene e ogni cosa esaltante che sta nel mezzo (web neurosinaptico e chip neuromorfici), abbiamo appena cominciato il viaggio dei prossimi 100 anni.

Secondo Jeff Immelt di GE “nel futuro ci si aspetta una fabbrica globale e aperta di macchine altamente intelligenti che si connettono, comunicano e cooperano con noi. Quando si parla di Internet industriale non si parla di un mondo fatto girare dai robot, ma si parla di mettere insieme le migliori tecnologie del mondo per risolvere le nostre sfide più importanti. Vale a dire quelle che riguardano l’energia economicamente e ambientalmente sostenibile, la cura di malattie incurabili e la predisposizione delle nostre infrastrutture e città in vista dei nostri prossimi 100 anni.

La connettività diffusa cambierà i modelli di comportamento

L’attrito economico risulta con evidenza quando si considera la perdita delle mansioni ripetitive a reddito medio, che possono essere largamente automatizzate o che possono disporre di strumenti online per il loro svolgimento. Cassieri di banca, addetti alla cassa del negozio e anche insegnanti di scuola superiore saranno eliminati dalla forza lavoro a favore di sportelli bancomat, casse self-service e MOOC (Massive Open Online Courses; in italiano, corsi online aperti su larga scala). Questo non è solo dovuto all’IoS,( Internetwork Operating System) ma all’integrazione della modellistica computazionale nella nostra vita di tutti i giorni, come previsto da Herbert Simon e Mark Weiser.

https://www.industriaitaliana.it/come-l-iot-potra-cambiare-la-societa-e-leconomia/
La connessione dell’IoS con una più vasta interdipendenza di oggetti e processi unita all’esposizione ad un maggiore grado di monitoraggio (per esempio nella salute) indurranno cambiamenti nel comportamento con una crescente diffusione dell’IoT. Che poi la connessione ubiquitaria modifichi l’attività razionale contrapposta a quella irrazionale, è tutto da vedere e da essere analizzato (Thinking Fast and Slow di Daniel Kahneman). I risultati di queste analisi devono essere presi in considerazione nella progettazione di oggetti e servizi futuri, per esempio l’ospedale wireless o le macchine per la risonanza magnetica (RM) situate in stazioni di rifornimento d’idrogeno, o quelle portatili per i raggi X nei presidi medici di Amazon. L’utilità di queste proposte potrebbe dipendere dall’etica socio economica della società (Scarcity di Sendhil Mullainathan) e dal livello raggiunto nell’ambito dello sviluppo socio-economico (Development as Freedom di Amartya Sen).

La previsione che la connettività cambierà i modi del comportamento è radicata nei fondamentali principi della fisica delle particelle. Il cosiddetto “effetto dell’osservatore”, si riferisce ai cambiamenti che l’azione dell’osservare avrà sul fenomeno osservato, da non confondere con il principio di indeterminazione proposto da Werner Heisenberg. Il primo potrebbe spiegare perché uno può cantare sotto la doccia e non in pubblico.

Si può osservare che il comportamento collegato, in particolare la centralità temporale nei sistemi ciber-fisici (hardware e software integrati con oggetti fisici), cambia se uno qualsiasi dei componenti viene modificato, anche se i componenti sono pressoché identici. Il rendez vous con il tempo può a volte risultare difficile.

**La necessità di ammodernamento dell’istruzione pubblica**

In aggiunta alla lentezza dei cambiamenti nel comportamento, pure il riequilibrio economico sarà lento, perché sono necessari cambiamenti di massa del nostro sistema educativo per ottimizzare il consumo sociale dei frutti derivati dalla tecnologia. Nessun quantitativo di tecnologia o corsi online inibirà la diffusione della frattura nel nostro sistema finanziario, a meno che non venga ammodernata l’educazione pubblica, ripristinato il rispetto per l’università, ci si rifocalizzi sul rigore, vengano rinnovati tutti gli aspetti della ricerca scientifica, restaurata la dignità dovuta all’insegnante e riaccesa la passione che ci si attende dall’insegnante stesso.

La filiera emergente del talento deve includere con dovizia ragazze che siano eccellenti in matematica, e che possano programmare e scrivere in maniera
logica. È essenziale che le donne abbiano come obbiettivo il raggiungimento di un elevato livello nelle scienze, in ingegneria, matematica economia e filosofia. Come possiamo accettare che il 50% circa del potere intellettuale sia escluso dalla forza lavoro?

Donne istruite contribuiranno all’educazione di ragazzi che siano rispettosi e ragazze che si possano distinguere. Tutto questo insieme accelererà la parallela innovazione di massa proveniente dai più distanti recessi del mondo e che sta preannunciando uno spostamento tettonico anche nei più tradizionali business. L’approccio analisi-paralisi dei colossi potrebbe portarli all’estinzione, qualora rimanessero noncuranti del fatto che l’insuccesso è la nuova strada del successo e l’insuccesso è il mantra per coloro che desiderano avere successo.

I nuovi modelli di business

L’innovazione distribuita richiede un approccio imprenditoriale e un’azione aggressiva che si svolge su livelli multipli simultanei piuttosto che la mitica soluzione semplice e diretta ad un problema complesso (Innovation: The Attacker’s Advantage di Richard Foster). L’industria dei tassisti contro Uber, l’industria alberghiera contro Airbnb e le agenzie interinali contro Upwork sono un chiaro esempio. L’esplosione di strumenti ingegneristici ha dramaticamente ridotto il ciclo di tempo necessario per introdurre innovazioni comprimendo enormemente il tempo necessario dalla concezione (sviluppo della filiera) alla realizzazione (realizzazione della filiera). I giganti dell’industria devono imbrigliare questa esplosione liberandosi delle piattaforme in modo da aggregare intelligenza che può operare su piattaforme open source. Il flusso di micro-entrate che deriva da miliardi di ping (Packet internet grouper) sul tuo prodotto sarà quello che fa la differenza e quel data service a valore aggiunto sarà messo in relazione con l’analisi intelligente dei dati e il possesso di una informazione operativa fino al punto d’utilizzo, prima che il dato si deteriori.

Ad ogni modo, la dinamica di deperibilità dei dati cambia quando la raccolta della loro serie storica è molto più determinante per l’analisi predittiva (ad esempio l’assistenza sanitaria) piuttosto che per i dati di breve vita (per esempio, la misura del tempo medio di guasto (MTBF) per i pezzi di ricambio). Il trasporto e l’archiviazione dei dati sono importanti in questo business, ma i consumatori potrebbero voler pagare solo per il real-time analytics (l’istantaneo accesso e uso dei dati di analisi); i consumatori si aspettano che i dati grezzi siano gratis.

Ad ogni modo tutti i vantaggi sono temporanei. L’esperienza finanziaria di guadagni derivati da micro-entrate per il noleggio della piattaforma è una delle
ragione per cui Apple ha aperto il suo “bus” affinché chiunque potesse salirvi (creando applicativi). Le App vi si riversano in massa da ogni parte del mondo. Il creatore di app è parte della valanga economica che ha consentito ad Apple, in qualità di gestore del canale, di aggregare micropagamenti utilizzando l’open innovation. Questo meccanismo di aggregazione da milioni di small data ha consentito ad Apple di realizzare il più grande database di pagamenti al mondo per la modesta somma di 99 centesimi alla volta. Il successo di PayPal ha alimentato Tesla, la quale potrebbe svendere sottocosto l’auto in modo da piazzare batterie estraibili al graphene e servizi a bordo basati sul SDN (Software Defined Networking). L’automobile può diventare la rete elettrica mobile del futuro per la distribuzione off-grid dell’energia sostituendo le reti smart fisse. Dare la disponibilità libera di prodotti in cambio di micro-entrate derivate dal pagamento dei servizi in base all’effettivo utilizzo (pay per use) è una strategia di business collaudata (stampanti contro inchiostro, telefoni cellulari contro servizi, distributore di acqua fredda contro acqua imbottigliata) per accrescere quei micro-guadagni che godranno di una lunga vita e comproveranno il valore della “long tail”.

Il consumatore istruito è il miglior consumatore

È previsto che la diffusione dell’IoT e dell’IoS dia luogo a nuovi prodotti e servizi (Datta, 2015 b). Il consumo di questi beni e i miglioramenti in efficienza potrebbe generare una crescita economica talmente grande da indurre all’euforia gli amministratori delegati. Secondo GE, Cisco e altri, l’IoT e l’IIoT potrebbero aggiungere circa da 14000 a 19000 miliardi di dollari all’economia globale nel corso del prossimo decennio. Per materializzare questi numeri è necessaria un’esplosione di consumerismo.

I milioni di persone che scrivono sui muri o che pubblicano fotografie per tenere a galla il tappo del mercato dei software della bolla dei social media, appartengono ad un gruppo i cui guadagni non permettono di comprarsi la macchina parlante o l’avatar per gestire il tè del mattino a letto o la laparoscopia robotica. Il consumatore istruito è il migliore consumatore. La bolla dell’economia dei social media del chiacchiericcio frenetico può essere limitata dalla quantità di “energia sottesa dalla curva” e l’irrazionale esuberanza potrebbe alimentare la prossima recessione globale, potenzialmente dietro l’angolo (2020-2022). Con l’ultimo anno di recessione nel 2008, attendiamo la prossima attorno al 2022, se è vero che il “ciclo di espansione e contrazione” ha una periodicità di 14 anni, stando a quanto dice Finn Kydland (Premio Nobel in Economia, 2004).
Negli Stati Uniti si stanno sviluppando numerose iniziative che riguardano l’industria e il mondo accademico per indirizzare la nuova generazione di innovazioni verso l’IoT spaziale, l’IoT industriale, l’IoS e le entusiasmanti possibilità derivate dalla ricerca nei sistemi ciberfisici (CPS). Numerosi consorzi sono stati creati nel 2014-2015 con il supporto dalle aziende leader nel mercato. Numerosi gruppi accademici fanno da battistrada con nuove invenzioni ed innovazioni. L’Unione Europea ha finanziato un massiccio programma pluriennale chiamato Horizon 2020 per un importo pari a più di 100 miliardi di dollari per esplorare la crescita dell’IoT e aiutare a raccogliere i frutti del guadagno economico inaspettato.

Ma sarà un flagrante errore da parte della leadership globale ignorare le previsioni economiche continuando a lucidare la cromatura senza prestare attenzione a calibrare il motore dell’educazione. Una città smart (Datta, 2015 a) è realmente smart senza cittadini smart?

L’umanità ha bisogno di sognatori (Datta, n.d.) e l’istruzione (Datta, 2014) è il sale per eccellenza (Salt di Mark Kurlansky), che agisce come fornitore di ispirazione, immaginazione, invenzione, innovazione guidando l’implementazione di idee (Datta, 2015 b). L’istruzione di un ragazzo potrebbe cambiare il destino di un uomo. L’istruzione di una ragazza potrebbe cambiare il destino di una nazione.

Bibliografia


https://www.industriaitaliana.it/come-l-iot-potra-cambiare-la-societa-e-leconomia/


https://www.industriaitaliana.it/come-l-iot-potra-cambiare-la-societa-e-leconomia/
INDUSTRIA ITALIANA
ANALISI & NEWS SU ECONOMIA REALE, INNOVAZIONI, DIGITAL TRANSFORMATION

DIRETTORE FILIPPO ASTONE

lot: una galassia di nuovi modelli

ISCRIVITI ALLA NOSTRA NEWSLETTER
di Shoumen Palit Austin Datta • L’Internet delle cose cambia la società e le aziende, e rende possibile digital transformation e creazione di valore. Un cambiamento che non è indolore e necessita di investimenti in cultura e formazione. Ecco come fare

Ospitiamo su Industria Italiana un articolo del Dr Shoumen Palit Austin Datta, Senior Member, MIT Auto-ID Labs, Research Affiliate, Department of Mechanical Engineering, Senior Scientist, MDPnP Lab, Department of Anaesthesiology, Massachusetts General Hospital, Harvard Medical School (vedi scheda a fondo pagina per la biografia completa e la bibliografia relativa all’articolo). La traduzione dell’articolo intitolato originariamente “Dynamic socio-economic disequilibrium catalysed by the Internet of Things” già pubblicato nel Journal of Innovation Management vol. 3, n. 3, 2015 è stata elaborata dalla Dr. Marianna Marchesi, Architetto e Ricercatore, Member of ISAAAAC Association – Global Advisory Group.
L’evoluzione della tecnologia basata sul concetto di Internet delle Cose (IoT, dall’inglese Internet of Things) e l’IoT industriale (IIoT) ha avuto inizio più o meno nel 1988 grazie al lavoro di Mark Weiser al Research Center di Xerox a Palo Alto. In questo lavoro l’autore avanzò l’ipotesi che i computer potessero intrecciarsi alla vita di ogni giorno e, conseguentemente, influenzare il futuro dell’imprenditoria (Scientific American, 1991). Le conoscenze a cui Weiser alludeva si riferivano alla discussione di Herbert Simon nel suo articolo del 1987 “The Steam Engine and the Computer: What makes technology revolutionary” [Il motore a vapore e il computer: cosa rende la tecnologia rivoluzionaria] nel quale Simon fotografa la sua riflessione sul computer in questo modo: “you have to make friends with it, talk to it, let it talk to you” [devi fare amicizia con esso, parlare ad esso, lasciare che esso parli a]

Una nuova rivoluzione

La connessione degli oggetti fisici (composti di atomi) alle informazioni (confezionate in bit) potrebbe dare seguito a un ‘altra rivoluzione che in molti, tra cui Neil Gershenfeld, hanno predetto. L’attuale cambiamento viene spesso indicato come la terza rivoluzione industriale in relazione alla seconda, l’era dell’Informazione, e alla prima rivoluzione industriale. In alcuni settori, l’attuale tendenza (anche definita come Industria 4.0) è indicata come l’età dei sistemi ciber-fisici (CPS, dall’inglese cyber-physical system).

Si suppone che le rivoluzioni modifichino il futuro che ci aspetta e la terza rivoluzione non farà eccezione: genererà attriti sia sociali che economici. Un conflitto con l’attuale status quo che sarà fuori dall’ordinario. Determinerà uno scontro tra visionari e persone dotate di scarsa immaginazione, nonché tra vecchie idee asfittiche contrapposte all’innovazione prorompente e geograficamente delocalizzata, con il risultato che l’incantesimo tecnologico si riverosera nelle nostre vite.
Sta emergendo con forza dirompente un attrito sociale dovuto al deterioramento della privacy per come viene intesa tradizionalmente e una sua ridefinitione che sfida le convinzioni vecchio stampo. C’è una preoccupazione giustificata che riguarda la sicurezza e, contemporaneamente, c’è riluttanza a rinunciare ai benefici ad essa associati. Un attrito di carattere finanziario si evidenzia sia nelle nazioni industrializzate sia nelle economie emergenti ogni volta manodopera non qualificata diventa parte della forza lavoro. La manodopera, in generale, aborre l’automazione, la quale restringe la domanda di lavoro non qualificato e crea un impatto negativo sull’economia e sulla società nel loro insieme.

Ma queste non sono osservazioni nuove, sono tendenze secolari che si ripeteranno ancora e ancora, anche se con differenti sfumature. Secondo lo storico economista Norman Poire, i cinque secoli compresi tra il 1440 e il 1939 furono i più dinamici di tutta la storia. Molti progressi tecnologici apparvero in quel periodo, ma tre invenzioni emersero rispetto al resto come punti di svolta nell’evoluzione della tecnologia, e condussero ad un decisivo mutamento sociale. L’invenzione della macchina tipografica ad opera di Johannes Gutenberg nel 1440 sprouò l’arrivo della Rivoluzione dell’Informazione che diffuse il Rinascimento attraverso l’Europa.

Nel 1609 il telescopio di Galileo Galilei diede inizio alla Rivoluzione Scientifica e all’Epoca dell’Illuminismo. La Rivoluzione Industriale e il Marxismo arrivarono poco dopo l’invenzione della macchina a vapore da parte di James Watt nel
dell’Informazione. Poco meno di un secolo più tardi, siamo ancora alle soglie di un’altra fase di cambiamento. La terza rivoluzione industriale potrebbe incoraggiare decisamente la grande convergenza della rivoluzione industriale con la rivoluzione dell’informazione, con tutte le incognite esistenti.

JOHN ATANASOFF E IL COMPUTER DIGITALE MESSO A PUNTO CON CLIFFORD BERRY

Gli effetti sul mercato del lavoro

ISCRIVITI ALLA NOSTRA NEWSLETTER
"Machine Age"). Una analisi incisiva relativamente alle disuguaglianze è presente anche nei lavori di Joseph Stiglitz (The Price of Inequality) e Robert Reich (Inequality for All). Brynjolfsson e McAfee ripercorrono la discussione sulla elevata disoccupazione che John Maynard Keynes aveva definito come "disoccupazione tecnologica" negli anni '30. Robert Frank negli anni '90 e anche nel suo libro The Darwin Economy rivista lo stesso tema e lo individua come quello che, catalizzato dalla tecnologia, la fa da padrone sui mercati del lavoro.

Brynjolfsson e McAfee prevedono che il nostro mondo “prospererà sulla frontiera digitale”. Ma cosa si può dire riguardo il percorso per raggiungere questa frontiera digitale? La strada che si staggia di fronte a noi è piena di insidie, tra nazioni in lotta, malnutrizione, scarse condizioni igieniche, inadeguata istruzione e scarsità di energia. Considerati assieme, già di per sé questi fattori hanno la capacità di alimentare in maniera lampante attribito socio economici, che potrebbero essere ulteriormente esacerbati dalle caratteristiche ascritte alla diffusione globale dell'IoT.

L'impossibile definizione di un unico standard globale per l’IoT

I frutti che possono derivare dall’IoT dipenderanno dalla nostra abilità di scambiare e condividere informazioni tra sistemi, oggetti e apparecchi in differenti ambienti, che supportano diversi standard operativi, procedurali e applicativi. Non è possibile coltivare l’aspettativa che il mondo cerchi a tutti i costi di supportare un unico comune standard. Quindi nessuna
Le aziende leader nel campo industriale devono abilitare standard aperti per le interfacce (APIs) attraverso le quali i prodotti delle piccole e medie imprese (SMEs) possono connettersi ad un bus, a un canale globale comune (un insieme di collegamenti fisici utilizzati da più elementi per scambiare informazioni da elaborare) in modo da accedere al collegamento e includere i loro servizi a valore aggiunto i loro computer o valorizzare le loro applicazioni di nicchia. Lo sviluppo sistematico di una dorsale open di connessione è decisivo nell’acquisizione di dati e nella diffusione dell’IoT.

In definitiva sarà la capacità di estrarre informazioni dai dati a guidare la value proposition della connessione. I dati economici fondamentali relativi ai costi di compravendita nella connessione (The Nature of the Firm di Ronald Coase, 1937) determineranno il ROI, la resa del capitale investito, che influenzerà l’adozione del business da parte delle aziende. L’enfasi sulla ricerca di risultati che siano immediatamente conseguibili e di un ritorno a breve termine, prevalenti nel mondo aziendale, può avere un impatto sulla misura dell’acquisizione dei dati. Investimenti inadeguati possono limitare gli strumenti necessari per accumulare una massa critica di dati. In ogni caso, senza un numero sufficiente di dati, gli strumenti di analisi potrebbero fallire nell’ evidenziare dei modelli non evidenti. Un’azione di questo tipo risulta necessaria se i motori di analisi dinamica in tempo reale (all’edge e al core) rappresentano una delle strade per monetizzare l’IoT. Nuove fonti di fatturato potrebbero essere
Per l’IoT è necessario un dispiegamento su larga scala

Troviamo una lezione sulla acquisizione e l’analisi dei dati occultata nel classico esperimento di meccanica quantistica descritto come il cosiddetto esperimento della doppia fessura di Young. Una variazione dell’esperimento venne eseguita al HCRL (Hitachi Central Research Labs) dal Dr. Akira Tonomura (1942-2012). Si constatò lo sviluppo di modelli di interferenza derivanti da singoli elettroni, ma il fenomeno non si manifestò
risultati o potrebbe fornire indicazioni sbagliate perché non si può costruire un elefante utilizzando un topo come modello. Questo lavoro è rivelatorio, perché dimostra come il dispiegamento su larga scala può essere la chiave per estrarre valore significativo da strumenti e tecnologie che possano fornire soluzioni quando sono combinate e convergenti.

Un’altra lezione da tener presente può essere individuata nella storia delle tecnologie generali di processo, in particolare nelle strategie che resero possibile la diffusione dell’elettricità (The Economic Future in Historical Perspective edited by P. A. David and M. Thomas, Oxford University Press, 2003). Il termine ‘disruptive’ utilizzato da Clayton Christensen in “The Innovators Dilemma” è una iperbole basata sul concetto originale di tecnologie generali di processo (GPT, dall’inglese general process technologies) introdotto durante l’epoca dell’elettrificazione per indicare una integrazione sistemica, contrapposta a un uso ad hoc, di momento. Christensen usò impropriamente la parola e fuori il mondo delle aziende utilizzando un’analisi povera di dati per far pensare che ogni cosa è disruptive.
Negli ultimi 15 anni abbiamo osservato la mancanza di integrazione sistemica del **RFID** (dall’inglese Radio-Frequency Identification; in italiano identificazione a radiofrequenza). Di conseguenza, non abbiamo tratto sufficiente profitto dalla capacità dei tag RFID di acquisire un elevato volume di dati tramite un approccio di sistema. Come risultato, non siamo stati in grado di fornire adeguata trasparenza all’interno delle supply chains e i risparmi derivanti dalla catena di valore rimangono ben al di sotto delle attese. Le considerazioni che si possono trarre dall’abbandono dell’iniziativa RFID a **WalMart** (Florian Michahelles [2010] *Is RFID dead?* Auto-ID Labs St. Gallen, ETH Zurich), non riguardano il fallimento di una tecnologia, ma un utilizzo inadeguato dei data tools nel contesto delle **ETICHETTE RFID**.

**ETICHETTE RFID**

**ISCRIVITI ALLA NOSTRA NEWSLETTER**
progressivamente spazio nella nostra realtà quotidiana attraverso l’integrazione dei sistemi, la connessione, e gli applicativi. L’IoT deve evolversi dall’internet delle cose all’internet dei sistemi (IoS). Dalla connessione tra ecosistemi di sistemi può arrivare il prossimo tsunami di redditività. A sua volta questo genererà ancora più dibattito nelle nostre agende di politica sociale relativamente a sicurezza, privacy, fiducia e aspetti relativi all’etica. La connettività IoT e la comunicazione con gli oggetti e i processi cambierà la maniera in cui interagiremo e ci comporteremo nelle nostre vite personali e professionali, nell’era dell’IoS.

Il perseguimento dell’autonomia nell’assistenza sanitaria, nel trasporto e nella produzione creerà nuove soluzioni e vecchi grattacapi, e genererà nuovi modelli di business. La possibilità di diagnosticare un tumore almeno un decennio prima che si manifesti non è un’illusione. Il veicolo autonomo che si parcheggia da solo e un carico merci che arriva a destinazione senza la presenza di umani nel processo è ormai storia vecchia. La morte delle giacenze di magazzino e la nascita delle produzioni distribuite su richiesta (dMOD) marginali (dMODE) rappresenta l’embrione della Manifattura 5.0, catalizzata dalle stampanti 3D. Con le valvole cardiache fino ai nano-satelliti, con l’assistenza attiva e passiva (SMAP) per l’umidità del suolo messa a punto dalla NASA ai fini dell’agricoltura di precisione, con l’acqua potabile desalinizzata priva di arsenico e depurata dal grafene e ogni cosa esaltante che sta nel mezzo (web neurosinaptico e chip neuromorfici), abbiamo appena cominciato il viaggio dei prossimi 100 anni.
Secondo Jeff Immelt di GE « nel futuro ci si aspetta una fabbrica globale e aperta di macchine altamente intelligenti che si connettono, comunicano e cooperano con noi. Quando si parla di Internet industriale non si parla di un mondo fatto girare dai robot, ma si parla di mettere insieme le migliori tecnologie del mondo per risolvere le nostre sfide più importanti. Vale a dire quelle che riguardano l'energia economicamente e ambientalmente sostenibile, la cura di malattie incurabili e la predisposizione delle nostre infrastrutture e città in vista dei nostri prossimi 100 anni.»

La connettività diffusa cambierà i modelli di comportamento
superiore saranno eliminati dalla forza lavoro a favore di sportelli bancomat, casse self-service e MOOC (Massive Open Online Courses; in italiano, corsi online aperti su larga scala). Questo non è solo dovuto all’IoS, (Internetwork Operating System) ma all’integrazione della modellistica computazionale nella nostra vita di tutti i giorni, come previsto da Herbert Simon e Mark Weiser.
dell’IoT. Che poi la connessione ubiquitaria modifichi l’attività razionale contrapposta a quella irrazionale, è tutto da vedere e da essere analizzato (*Thinking Fast and Slow* di Daniel Kahneman). I risultati di queste analisi devono essere presi in considerazione nella progettazione di oggetti e servizi futuri, per esempio l’ospedale wireless o le macchine per la risonanza magnetica (RM) situate in stazioni di rifornimento d’idrogeno, o quelle portatili per i raggi X nei presidi medici di Amazon. L’utilità di queste proposte potrebbe dipendere dall’etica socio economica della società (*Scarcity* di Sendhil Mullainathan) e dal livello raggiunto nell’ambito dello sviluppo socio-economico (*Development as Freedom* di Amartya Sen).

La previsione che la connettività cambierà i modi del comportamento è radicata nei fondamentali principi della fisica delle particelle. Il cosiddetto effetto dell’osservatore”, si riferisce ai cambiamenti che l’azione dell’osservare avrà sul fenomeno osservato, da non confondere con il principio di indeterminazione proposto da Werner Heisenberg. Il primo potrebbe spiegare perché uno può cantare sotto la doccia e non in pubblico. Si può osservare che il comportamento collegato, in particolare la centralità temporale nei sistemi ciber-fisici (hardware e software integrati con oggetti fisici), cambia se uno qualsiasi dei componenti viene modificato, anche se i componenti sono pressoché identici. Il rendez vous con il tempo può a volte risultare difficile.
La necessità di ammodernamento dell’istruzione pubblica

In aggiunta alla lentezza dei cambiamenti nel comportamento, pure il riequilibrio economico sarà lento, perché sono necessari cambiamenti di massa del nostro sistema educativo per ottimizzare il consumo sociale dei frutti derivati dalla tecnologia. Nessun quantitativo di tecnologia o corsi online inibirà la diffusione della frattura nel nostro sistema finanziario, a meno che non venga ammodernata l’educazione pubblica, ripristinato il rispetto per l’università, ci si rifocalizzi sul rigore, vengano rinnovati tutti gli aspetti della ricerca scientifica, restaurata la dignità dovuta all’insegnante e riaccesa la passione che ci si attende dall’insegnante stesso. La filiera emergente del talento deve includere con dovizia ragazze che siano eccellenti in matematica, e che possano
ingegneria, matematica economia e filosofia. Come possiamo accettare che il 50% circa del potere intellettivo sia escluso dalla forza lavoro? Donne istruite contribuiranno all’educazione di ragazzi che siano rispettosi e ragazze che si possano distinguere. Tutto questo insieme accelererà la parallela innovazione di massa proveniente dai più distanti recessi del mondo e che sta preannunciando uno spostamento tettonico anche nei più tradizionali business. L’approccio analisi-paralisi dei colossi potrebbe portarli all’estinzione, qualora rimanessero noncuranti del fatto che l’insuccesso è la nuova strada del successo e l’insuccesso è il mantra per coloro che desiderano avere successo.
L’innovazione distribuita richiede un approccio imprenditoriale e un’azione aggressiva che si svolge su livelli multipli simultanei piuttosto che la mitica soluzione semplice e diretta ad un problema complesso (Innovation: The Attacker’s Advantage di Richard Foster). L’industria dei tassisti contro Uber, l’industria alberghiera contro Airbnb e le agenzie interinali contro Upwork sono un chiaro esempio. L’esplosione di strumenti ingegneristici ha dramaticamente ridotto il ciclo di tempo necessario per introdurre innovazioni comprimendo enormemente il tempo necessario dalla concezione (sviluppo della filiera) alla realizzazione (funzionamento della filiera). I giganti dell’industria devono imbrigliare questa esplosione liberandosi delle piattaforme in modo da aggregare intelligenza che può operare su piattaforme open source. Il flusso di micro-entrate che deriva da miliardi di ping (Packet internet grouper) sul tuo prodotto sarà quello che fa la differenza e quel data service a valore aggiunto sarà messo in relazione con l’analisi intelligente dei dati e il possesso di una informazione operativa fino al punto d’utilizzo, prima che il dato si deteriori.

Ad ogni modo, la dinamica di deperibilità dei dati cambia quando la raccolta della loro serie storica è molto più determinante per l’analisi predittiva (ad esempio l’assistenza sanitaria) piuttosto che per i dati di breve vita (per esempio, la misura del tempo medio di guasto (MTBF) per i pezzi di ricambio). Il trasporto e l’archiviazione dei dati sono importanti in questo business, ma i consumatori potrebbero voler pagare solo per il real-time analytics (l’istantaneo accesso e uso dei dati di analisi); i consumatori si aspettano che i dati grezzi siano gratis.

L'automobile può diventare la rete elettrica mobile del futuro per la distribuzione off-grid dell'energia sostituendo le reti smart fisse. Dare la disponibilità libera di prodotti in cambio di micro-entrate derivate dal pagamento dei servizi in base all'effettivo utilizzo (pay per use) è una strategia di business collaudata (stampanti contro inchiostro, telefoni cellulari contro servizi, distributore di acqua fredda contro acqua imbottigliata) per accrescere quei micro-guadagni che godranno di una lunga vita e comerteranno il valore della "long tail".

**Il consumatore istruito è il miglior consumatore**

È previsto che la diffusione dell'IoT e dell'IoS dia luogo a nuovi prodotti e servizi (Datta, 2015 b). Il consumo di questi beni e i miglioramenti in efficienza potrebbe generare una crescita...
I milioni di persone che scrivono sui muri o che pubblicano fotografie per tenere a galla il tappo del mercato dei software della bolla dei social media, appartengono ad un gruppo i cui guadagni non permettono di comprarsi la macchina parlante o l’avatar per gestire il tè del mattino a letto, o la laparoscopia robotica. Il consumatore istruito è il migliore consumatore. La bolla dell’economia dei social media del chiacchiericcio frenetico può essere limitata dalla quantità di “energia sottesa dalla curva” e l’irrazionale esuberanza potrebbe alimentare la prossima recessione globale, potenzialmente dietro l’angolo (2020-2022). Con l’ultimo anno di recessione nel 2008, attendiamo la prossima attorno al 2022, se è vero che il “ciclo di espansione e contrazione” ha una periodicità di 14 anni, stando a quanto dice Finn Kydland (Premio Nobel in Economia, 2004).
DATTA: UNA SMART CITY È REALMENÉ SMART SENZA CITTADINI SMART

Negli Stati Uniti si stano sviluppando numerose iniziative che riguardano l’industria e il mondo accademico per indirizzare la nuova generazione di innovazioni verso l’IoT spaziale, l’IoT industriale, l’IoS e le entusiasmanti possibilità derivate dalla ricerca nei sistemi ciberfisici (CPS). Numerosi consorzi sono stati creati nel 2014-2015 con il supporto delle aziende leader nel mercato. Numerosi gruppi accademici fanno da battistrada con nuove invenzioni ed innovazioni. L’Unione Europea ha finanziato un massiccio programma pluriennale chiamato Horizon 2020 per un importo pari a più di 100 miliardi di dollari per esplorare la crescita dell’IoT e aiutare a raccogliere i frutti del guadagno economico inaspettato.

Ma sarà un flagrante errore da parte della leadership globale ignorare le previsioni economiche continuando a lucidare la cromatura senza prestare attenzione a calibrare il motore economico e realizzare che i vantaggi economici sono accessori. ISCRIVITI ALLA NOSTRA NEWSLETTER
guidando l’implementazione di idee (Datta, 2015 b). L’istruzione di un ragazzo potrebbe cambiare il destino di un uomo. L’istruzione di una ragazza potrebbe cambiare il destino di una nazione.


Come scienziato ricercatore per la Engineering Systems Division, MIT School of Engineering, ha analizzato nell’ambito dell’ innovazione tecnologica RFID, IoT, supply chain, data analytics e la convergenza delle soluzioni in verticale. Ha insegnato Supply Chain, Sistemi ed Imprese e il MIT Supply Chain of

ISCRIVITI ALLA NOSTRA NEWSLETTER
Senior Vice President dell’ Object Management Group (OMG). Il suo impegno con il MDPnP Programal Massachusetts General Hospital, Harvard Medical School riguarda la convergenza tra l’interoperabilità delle apparecchiature mediche con le attrezzature in via di sviluppo nell’ambito medicale, relativamente ai CPS e all’ IoT.
I suoi interessi spaziano dall’IoT/IIoT alla promozione della vision dell’internet of systems (IoS) che estende la connettività ubiquitaria, incluso l’industrial internet, per catalizzare la crescita economica globale. E autore di diversi articoli riguardanti l’evoluzione dell’industrial internet (2003), isoftware intelligenti (2001), gli ambienti per l’analisi predittiva, la gestione e le operazioni generali relative alla supply chain, l’utilizzo dei sensori nelle piattaforme sanitarie, le energie rinnovabili e tutte le varie forme della trasformazione digitale dell’economia globale. Ha svolto il ruolo di advisor per associazioni e governi incluso il Dipartimento della difesa statunitense, le Nazioni Unite (UNDP), la World Customs Organization (WCO) e altri ancora.

E’ stato Research Fellow in Medicina (Thyroid Lab, Neuro-Endocrine Lab, Molecular Oncology) al Massachusetts General Hospital and Instructor in Medicine all’Harvard Medical School. E’ stato Research Associate al Whitehead Institute del MIT e membro del MIT Human Genome Project, così come Research Scientist all’University of California UCSF School of Medicine, San Francisco, California. Il Dr Datta si è impegnato nel settore pubblico per migliorare l’educazione e la tecnologia come Special Assistant per City & County of San Francisco, California; Science Education Partnership alla UCSF School of Medicine;
Bibliografia


COMMENTARY [B]

Shoumen Datta

HAPHAZARD REALITY – IOT IS A METAPHOR
THE COMMENCEMENT ADDRESS

esse non videri

This “speech” was not delivered at any commencement, to the best of our knowledge. The first couple pages are modified from an inspiring manifesto by Sanjay Sarma. The remainder was authored by Shoumen Datta. The opinions in this article are personal views of the authors and do not represent the views of MIT or any other organizations.
Thank you Madam President, students, parents, faculty and friends – it is a great day for all of us to celebrate the achievements of the graduating class of 2014. Congratulations!

While preparing this address, I reminded myself that some people come in our lives and it is a blessing but others come in our lives as lessons. Today, I think, I shall serve as a lesson to some of my young friends in the audience. I propose to ask a few questions. It is not for the faint-hearted and business will not be as usual. I apologize if it offends a few. It may.

Today, especially today, as you graduate, is your 6th sense telling you that change is afoot? Are you afraid or excited that you do not have immunity from the risk of disruption?

Are you able to navigate through the explosion in engineering tools which is collapsing the time between conception and realization of an idea?

What an engineer took years to accomplish a decade ago, she might be able to accomplish it in a matter of a few weeks, today. Is that engineer among you and graduating today?

→ Need 20 plastic cases for a new Internet-controlled thermostat you just invented?
  • Design it with a CAD package in the cloud and 3D-print it in your garage.

→ Need a complex circuit to control the system?
  • Hire a designer in Pakistan for $11 an hour. Delivery time: 5 days.

→ Need that prototyped?
  • Send it off to your favorite PCB fab and have it expressed to you within a week.

→ Need customers?
  • Pretail on Kickstarter

→ Need to make a million devices now that you’ve pre-sold a million units?
  • Find a manufacturer on Alibaba or work with an emerging class of plug-in companies that can act as your virtual manufacturing arm.

→ Need a complex piece of web technology to run the above system?
  • Prototype it using “Ruby on Rails” instead of writing miles of code in C or Java.

→ Need to have it up and running for your first million customers?
  • Deploy it on the Amazon Cloud (AWS EC2).
What did I just describe?

I described an express version in the engineering and product development process cascading from each other in a chain-reaction.

- Did your curriculum, industry internships and peer pressure prepare you to imagine, undertake the challenges of innovation, manage technology explosion, expose you to entrepreneurship and the social business of science in society?
- Can you exude and embrace balanced impatience as a virtue?
- Did your education teach you to think?
- Did you question the *status quo*?

Conventional wisdom encourages patience, planning, deliberation, steadiness and a “look before you leap” attitude. All good virtues and all sadly outdated. The issue is not the thoughts themselves, but the fetishized weight we give them and the paralysis and risk-aversion they have come to represent. The competitive world values action over over-planning. It rewards impatience over stodgy patience. It values agility over static steadiness. It values deliberate aggression over passive deliberation. And here is the key: the new way of thinking is not about taking risks, but about controlling risks using quick, small actions and experiments followed by rapid and agile adaptation.

In this age of systems of ecosystems, the risk lies increasingly in not realizing that you are in a manic episode of the Hunger Games. This is about “ready” “shoot” “aim” – not ready, aim, shoot.

If you have only one bullet, it makes sense to follow the mantra: ready, aim (carefully) and then shoot. If you have many bullets, then, get ready, shoot first, then correct your aim and shoot again. And repeat. Or your opponents will shoot first and a bullet will likely hit you.

Innovation is transitioning from breach-loaded dueling pistols to machine guns in a sort of massively multiplayer parallel innovation.

I know you are rolling your eyes and thinking, hey, don’t you know that the first instinct of institutions to disruptive thoughts or to innovative people is to expel them, prevent them from entering or locking them in a cage of red tape if one happens to enter through the cracks?
That is where the cultural chasm deepens. Unlike old school institutions, innovative companies are ruthless in breaking rules to get the job done. Entrepreneurs do not fear failure. Entrepreneurs have fought battles and lost and lived to fight again. Failure is the new road to success. Learning is expensive and painful, but death is the nail on the coffin (just ask Nokia).

In 1997, one of my friends helped a major city and a venerable academic institution to start an online tutoring program in mathematics over a fixed point to point Ethernet connection using the “new” medium of the Internet.

Fifteen years later, in the Winter of 2012, MIT launched MITx, a MOOC-esque effort to offer learning using digital tools to transform the way students learn on campus and the way students might be able to learn around the world. MITx was launched almost overnight. The time from its announcement in the end of December 2011, to the launch of the first massive online class (6.002x, Circuits and Systems) was just 3 months. More than 150,000 students registered for the class but don’t conclude that all 150,000 students took the class or completed the class. But I think a few did and we may have changed a few lives.

In the summer of 2012, MIT and Harvard launched edX, an independent not-for-profit entity to offer free courses to the world. They each invested $30M into edX. Today, edX has over 2 million enrollees and courses from 30 universities offered to students in 195 countries. EdX is the only not-for-profit in this space (Coursera, Udacity and NovoEd are MOOCs-for-profit). EdX is the only one whose platform code has been open-sourced to enable massively parallel innovation. Anant Agarwal, an MIT professor who happens to be an entrepreneur, spearheaded edX. In short, MIT embraced innovation, it jumped on the engineering express, it leveraged an intrapreneur, it protected the idea from institutional antibodies and grafted on to itself the Office of Digital Learning at MIT (directed by another brilliant intrapreneur Sanjay Sarma of MIT) and placed it in the Office of the Provost at MIT. Leadership from the top, directly from the Provost and President of MIT, in order to create a new world order.

I see a hand waving in the audience, “Yes, do you have a question for me?”

“Yes, I do. My name is Aziz Boxwala. I am the Vice President for Human Resources in this University. How many applicants did you have when you advertised for the position of CEO of edX and Director of Digital Learning at MIT?”
“We often do not advertise for leadership positions that require vision and institutional knowledge.”

“How amazing! There was no process, charter, by-laws and committee approvals. You just did it by fiat?”

“Yes, Mr Aziz Boxwala. Things are not “as is” and we have moved beyond the box. Charters and by-laws are relics found in the old countries.”

Let me return to the students. The task that lies ahead of you, students, requires that you must break rules and must climb out of boxes in order to innovate toward a better karma.

That is your task, students, to unchain the shackles of solipsistic bliss and proactively help to create a new world order. The new order of things is not about the internet of things or polishing the chrome, it is about tuning the economic engines using a trinity of platforms which are inextricably linked and must be pursued using the same mantra of massively parallel innovation. They are as follows -

[1] education of girls and women in math, science, engineering, economics and philosophy

[2] access to energy for all nations at a cost which is feasible, at a level of sustainability which is balanced with respect to the dynamic socio-economic context and at a quality of service which acts as an egalitarian purveyor of civilization


For the remainder of my time, I will drive toward my conclusion, which will significantly digress from the “speech” and the rhetoric. I will present a tangible grand challenge that I continue to dream about and I wish to pursue, if the opportunity presents itself. May be you can help or if you are fired up, you can pursue it. It is neither a proprietary idea nor a recipe for a secret sauce. It is a confluence of ideas including platforms [1] & [2] outlined, earlier.

My idea is about the global business of carbon neutral energy manufacturing and distributed cooperative ownership of the energy enterprise by women, who know math.

Before you tune off from me and turn on your Pandora, please allow me to explain the out of the box thinking. Let us start in this state, locally, but it is applicable, globally, where ever the sun shines. Yes, this specific example is about renewable solar energy but there is more.
In California, by 2020, we have a mandate to generate about a third of the state-wide consumption of energy from carbon-neutral or renewable sources. This energy can be produced by private corporations, such as the recently completed Ivanpah project. The utility companies are mandated to buy the renewable energy from these sources.

Efficiency of solar panels and photo-voltaics are driving down the cost from a high of around 80 cents per kilowatt-hour (kWh) to less than a tenth of that price. Government guaranteed loans combined with 9 to 1 leveraging of private capital and free government land available in the US southwest may bring the cost down to about 6 cents per kWh or about twice the cost of coal generated electricity. We may save a few more cents with Nest-like devices or standard wireless sensor networks (WSN) to monitor, reduce waste and optimize usage.

The gigantic nature of energy may be gleaned from the fact that a three reactor nuclear power plant produces about 3 GW. Compare that to the US annual peak demand for 786 GW and another 117 GW to meet the NERC target reserve supply. The Ivanpah Solar Power Plant took six years to construct at a cost of about $1.6 billion to tax payers and another $168 million from BrightSource Energy Company and Google. It was inaugurated on February 13, 2014 by US Secretary of Energy Ernie Moniz. This solar plant will generate 392 MW (0.392 GW) of electricity. The total amount of solar installations in the US generate just above 2 GW while the US peak demand is creeping toward 1000 GW or 1 TW.

Solar energy is time dependent and electricity is perishable. The solar peak does not match the demand peak. Hence, the importance of storage to make the energy supply agnostic of insolation and add mobility to energy in order to divorce the distribution from the grid infrastructure. Transporting energy to serve geographies with local micro-grids or an absence of electricity grids (major parts of the world) is of paramount importance for developing nations.

The era of lithium nano-phosphate batteries may be reaching its zenith but graphene based storage systems may offer hope with nearly 50% more capacity and charge density. Ask Anne (Sastry) or Samir (Mayekar) or Bob (Robert F Curl) or Harry (Kroto) to explain.

The convergence of insolation-dependent production with high density storage is a far better solution that is implementable. The abundance of high insolation areas in the US, Africa, India, Brazil, Indonesia and China makes this combined solution a sustainable approach to reduce dependency on fossil fuels and greenhouse gas (GHG) emissions.
Building and operating solar farms require no new invention or imagination. Distributing the energy through conventional schemes (grid) and new storage (nano-tech or graphene based batteries) offers room for [1] technology innovation, [2] financial innovation in the use of price as a supply-demand control parameter and [3] social innovation to deliver mobile energy units (truck load of energy containers) to areas devoid of grid or any other source of electricity. The latter immediately enables the operation of cell towers and explodes the potential penetration of mobile telephony, smart phones and the connectivity to the internet of things or the industrial internet of intelligent machines or smart objects.

What stands between this simple idea and reality is the relative availability of investment to create and operate solar farms. OECD nations are less likely to struggle for investors but the emerging economies may remain immersed in the vicious cycle of cost versus profits or the length of time necessary for the return on investment based on the cost per unit which the local market can bear. It is the same argument that asphyxiated the bio-butanol effort which can help 80% of the world even if photo-butanol production remained sub-optimal.

This economic segregation is in no way less perverse than the “legal” social segregation in the US in the first half of the 20th Century. I have to digress to present you with the context.

In the fall of 1944, Soledad Vidalaurri took her children and those of her brother, Gonzalo Méndez, to enroll at the 17th Street School in Westminster, California, a part of the Garden Grove School District. Although they were cousins and shared a Mexican heritage, the Méndez and Vidalaurri children looked quite different: Sylvia, Gonzalo Jr. and Geronimo Méndez had dark skin, hair and eyes, while Alice and Virginia Vidalaurri had fair complexions and features. An administrator looked the five children. Alice and Virginia could stay, he said. But, their dark-skinned cousins would have to register at the Hoover School, the town’s "Mexican school" located a few blocks away.

In *Méndez v. Westminster*, James L. Kent, the superintendent of the Garden Grove School District, took the stand and under oath, Kent said he believed people of Mexican descent were intellectually, culturally and morally inferior to Euro-Americans. Even if a Latino child had the same academic qualifications as a white child, Kent stated, he would never allow the Latino child to enroll in an Anglo school. U.S. District Court Judge Paul J. McCormick was also appalled by Kent’s blatant bigotry. On February 18, 1946, he ruled in favor of the plaintiffs.

The Orange County school boards filed an appeal.
By now, *Méndez v. Westminster* was drawing national attention. Civil rights lawyers in other states were watching the proceedings closely. For half a century, they had been trying to strike down the "separate but equal" doctrine of *Plessy v. Ferguson*. Among those following the suit was a young African American attorney, Thurgood Marshall. Marshall and two of his colleagues from the NAACP submitted an *amicus curiae* brief in this case.

On April 14, 1947, the Ninth Circuit Court of Appeals in San Francisco upheld the lower court decision regarding *Méndez v. Westminster*. The court stopped short, however, of condemning the "separate but equal" doctrine of *Plessy v. Ferguson*.

Seven years later, the NAACP did find a successful test case to reverse *Plessy v. Ferguson*. Thurgood Marshall argued the landmark *Brown v. Board of Education of Topeka* before the U.S. Supreme Court, presenting the same social science and human rights theories he outlined in his *amicus curiae* brief for *Méndez v. Westminster*. Former California Governor Earl Warren, then Chief Justice of the U.S. Supreme Court, wrote the historic opinion finally ending the legal segregation of students on the basis of race in American schools in 1954.

In 1954, by an unanimous vote (9-0) the U.S. Supreme Court declared "separate educational facilities are inherently unequal." However, sixty years later, US public education may be still separate, still unequal, but what is even more alarming is the fact that the ethos appears to be infectious. It has infected the renewable energy economy.

The rich and the powerful are using energy to drive a wedge between nations. Access to energy and development of non-fossil alternatives are at the heart of this quagmire. Sixty years from now, the energy economy will dictate if nations are "still separate, still unequal" because of the reluctance of seed investors to invest in alternative forms of energy with lower return on investments. Thus, the power of investment is depriving the emerging nations of energy and creating more "have-nots" if measured by purchasing power parity.

A few months ago in Brussels, seated in the basement of 5 rue Duquesnoy, I heard a fine gentleman from the famed Tata Sons talk about the mantra of social responsibility that Mr Jamshed Tata expounded before the West claimed social responsibility as a buzz word.

That morning, in Brussels, it dawned on me that the obvious idea of the solar energy farm, may be an instrument to improve mathematics education among girls and women. Math to a basic level, for example, a US high school standard of math comparable to students graduating from Stuyvesant High School. Delivering math through MOOCs should be easy.
To inspire and reward women who are able to grasp the fundamental principles of high school mathematics, can we offer them a percent ownership of the solar farm energy cooperative?

If you are an investor in the audience, you may want to walk out. I can’t stop you. But, those of you who may have open minds, consider my suggestion before you begin to deride me.

If Mr Tata were to invest in creating the solar farm, I suggest that he retains 51% ownership of the enterprise and recovers his entire investment from the energy sales (which may take 10-20 years). After Mr Tata (or Mr Gates or Mr Khosla or Mr Slim) has recovered the initial investment plus interest, the “energy company” morphs to an “energy cooperative” which continues to provide 51% of the profit to the investors.

It is the remaining 49% which can ignite global productivity and usher in a new era of civilization if 49% of the company is owned by women who are proficient in high school mathematics. Hence, 49% of the profit from the energy cooperative will be distributed between millions of math-proficient women-owners. What if you owned 0.01% of Apple?

This plan lifts many boats. Natural resources belong to people and a right to share in the economy. Investors must obtain the ROI but is it so difficult to uphold ethical globalization?

You are the first group to hear of this plan, this radical idea, this dream, this fantasy. I acknowledge that energy is not a panacea to solve everything, but, I don’t think it is absurd or impossible or financially untenable. Women will apply for their ownership in droves. If they are not qualified in math, we will offer temporary ownership with the expectation that in about three years, with a stipend and access to Khan Academy, they will qualify for the mathematics proficiency test. Imagine the impact of millions of math proficient women.

Once they pass the test and are owner-members of the energy cooperative, visualize the quantum leap in their self-esteem, knowledge and confidence not to mention the dramatic rate of socio-economic evolution which will change the tapestry of life in every domain.

*Can you force a math proficient woman to accept domestic abuse?*
*Can you force a math proficient woman to become pregnant each year?*
*Can you force a math proficient woman to submit to female genital mutilation?*
*Can you force a math proficient woman to ignore math education for her own children?*
The education of a boy can change the fate of a man. The education of a woman will change the destiny of a nation. Your task, students, is to change the world, one solar cooperative at a time. Your task, students, is to create a cottage industry for energy. Your task, students, is to change the fate of freedom and education.

It is time, students, for your tryst with destiny.

This is the true joy in life, the being used for a purpose you consider a mighty one, the being a force of nature rather than a feverish, selfish clod of ailments and grievances complaining that the world will not devote itself to making you happy. (GBS)

The best people possess a feeling for beauty, the courage to take risks, the discipline to tell the truth, the capacity for sacrifice. Ironically, their virtues make them vulnerable; they are often wounded, sometimes destroyed. (EH)
THE COMMENCEMENT ADDRESS

This “speech” was not delivered at any commencement, to the best of my knowledge. Suggestions reflect the views of the author and do not represent the views of any other organizations. Any similarity to names, symbols and acronyms, are purely coincidental.
Welcome to the University of the Future Commencement ceremonies for 2018. I thank you for joining us today on this special day for our students at the University of the Future (UF). I am J. Fox Kenneth, President of UF. My friends call me JFK and most people from the South call me President Fox. Many years ago, when I was a faculty at Purdue University, the Dean of the School of Engineering, called me Ken, an abbreviated form of my last name. Ken has stuck with me in my professional world.

I am Ken. I am happy to be here with you at UF, today. Thank you for the privilege Madam Governor, students, parents, faculty and friends. It is a bright sunny day for all of us to celebrate the achievements of the graduating class of 2018. Congratulations!

While preparing this address, I reminded myself that the success we are celebrating here, today, is a reason for great joy, yet one may pause to reflect that success is ephemeral. The brevity of the feeling of success may not diminish our success, but it reinforces a fact, which we must have the courage to accept. Success is a rental. The rent is due every day.

Hence, you and I, and everybody else who may aspire to succeed in their chosen path, must prepare to pay the rent, every day. The chef who prepared a delicious lamb stew, yesterday, will not receive any praise, today, if she serves a soggy, runny and bland lasagna. The surgeon who performed a miracle, yesterday, separating conjoined twins, may be chastised tomorrow if the spina bifida surgery goes awry, today. Rent is due every day.

The job of the University, the mission of education and the role of academics, is to prepare you, in as many ways, directions and dimensions, as possible, to enable you to pay your rental dues, each day, every day, for the next 25,000 days or more, hopefully.

UF is committed to support you, and those who will pass through these doors, after you and after them. UF must prepare, every day, to garner the resources, to deliver the academic environment, to seed the network, which will allow our students and faculty and staff, to move forward the mission of UF, through caring, teaching, research and global entrepreneurship, to lift many boats.

Today, I will take this opportunity, to put forward our collective vision of the future. I wish to speak my mind and what we are thinking.

Rosalyn Sussman Yalow & Norman Ernest Borlaug Memorial Commencement Address by J. Fox Kenneth, President of the University of the Future (UF), on November 3, 2018
I chose to share ideas with you because now you are adults, and I can request you to help me, to help UF, and through UF, to help make the world a better place, to serve our fellow citizens, and civilization, with dignity.

Conventional wisdom encourages patience, planning, deliberation, steadiness and a “look before you leap” attitude. All good virtues but perhaps, outdated, because of the paralysis and risk-aversion they have come to represent. The competitive world values action, over, over-planning. It rewards impatience over stodgy patience. It values agility over steadiness. It values deliberate aggression, without malice, over passive deliberation. The new way of thinking is not a revolution but a higher threshold of assuming the risks of leadership and, hopefully, controlling the risks using quick, small actions and experiments. Rapid and agile adaptation, while pursuing goals and targets, needs no additional emphasis.

Hence, we must act, we must transform our vision into reality, even if we fail, at first. Failure is an option, in order to prepare for grand success. The safety of using statistics to show incremental progress, has lost its place on the leadership agenda. We must lead and shoulder the risks of leadership. Failure, or the sense of failure, never goes unrewarded, if we have the vision to visualize, beyond where the eye can see, if we have the vision to consult and use corrective lenses, to re-view, if our imagination appears to be out of focus.

In this age of systems of ecosystems, the risk lies increasingly in not realizing that you are in a manic episode of the Hunger Games. Today it is about “shoot” then “aim” – it is no longer sufficient to stick to the old adage: ready, aim, shoot, repeat (if you are still alive).

The educational institution, as mammoth as UF, must keep pace with the speed and trajectory of global dynamics. Innovation is transitioning from breach-loaded dueling pistols to machine guns in a sort of massively multiplayer parallel innovation. UF must approach the world on multiple fronts, in order to prepare you better to pay the rent, on success, every day. UF entrepreneurs must not fear failure. UF entrepreneurs include all UF faculty, staff and students. UF entrepreneurs must unite behind our entrepreneurial spirit. Entrepreneurs must fight battles, wage wars, and lose, and live, to fight again. Failure is the new road to success. Stagnation is not an option.

If we wish to use the fruits of science, engineering, medicine, mathematics and the arts to lift society to a higher quality of life, for everybody, locally and globally, then we must acknowledge the immense difficulties we may face. There are no short cuts in science and life. Only those who may not dread its fatiguing climb, may reach its luminous summit.

Rosalyn Sussman Yalow & Norman Ernest Borlaug Memorial Commencement Address by J. Fox Kenneth, President of the University of the Future (UF), on November 3, 2018
I will share with you the history, and the future, shaped, literally, by a couple French visionaries. I will start with one now and end my speech narrating the “future” by another.

In 1830, Ferdinand de Lesseps, a French diplomat in Cairo, dreamed of linking Europe and Asia (Mediterranean and Red Seas) by digging a canal through 118 miles of arid land at a cost of 200 million French Francs (FFR). On November 17, 1869, the Suez Canal opened. Fewer than 500 ships passed through in 1870, far below the projected return on investment. Dividends failed to materialize and Egyptians, desperate for cash, sold the shares to Great Britain for 4 million Sterling (GBP). In 2002, 15,000 ships passed through the Suez Canal generating US$2 billion in revenues for Egypt (Parting the Desert: The Creation of the Suez Canal by Zachary Karabell). In 2016 nearly 25,000 ships passed through the Suez Canal and generated nearly US$5 billion for Egypt. In 2016, Suez Canal II was completed. Now, each day 100 ships can cross the Canal or 36,000 each year.

In 1880, Ferdinand de Lesseps formed a company to replicate his feat in Panama. This travail was undone both by weather and tropical diseases (yellow fever, cholera, malaria) that killed 22,000 laborers. Ferdinand de Lesseps (1805 - 1894) could not complete the Panama Canal. He did not live to see the successful completion of the Panama Canal, in 1914, by the US Army Corps of Engineers.

In 1904, through the urging of US President Theodore Roosevelt, the US Army Corps of Engineers completed the Panama Canal. It would cost US$352 million and 5,609 lives. In 1918, after 4 years in operation, less than 5 ships passed through the Panama Canal, daily. Today, more than 14,000 ships pass each year, generating US$2 billion in revenue (The Path Between the Seas: The Creation of the Panama Canal by David McCullough).

In the 20th century, the 31 mile tunnel linking England to the Continent got started in 1987 fueled by GBP 5 billion from banks and 112,000 British investors. In its first 3 years since May 1994, The Chunnel saw fire and operating expenses around GBP 2 billion.
In 1997, the investors suffered losses when the banks seized shares in exchange for restructuring the crushing debt. By 2001, the winds changed, failure turned into success (*The Chunnel: The Amazing Story of the Undersea Crossing of the English Channel* by Drew Fetherston). In 2016, The Chunnel transported more than 10 million passengers, nearly 2 million vehicles, generated more than 1 billion Euro in revenue. In 2016, net profit was 200 million Euro.

Big ideas offer big dividends, but may have stormy beginnings. Society will regress if leaders are blinded by short-term ROI, resist change, are risk averse and choose *status quo*. I wish to present to you a few ideas that UF may wish to think, reflect and perhaps, pursue.

In addition to supporting our ongoing efforts, collaborations, multi-disciplinary institutes and cross-pollinating centers of convergence, we will embark on strategic advancements in manufacturing, healthcare, energy and global goods.

In every endeavor, we will abide by our code of moral ethics, ethical globalization and entrepreneurship which is commensurate with reasonably shared profitability. I must mention “profitability” in this context because sustaining the educational mission, and research, is resource intensive. We cannot continue to depend on state and federal governments, because fiscal cycles continuously clash with our need for continuity. The support necessary for research to be productive, must take into consideration the long road ahead. I quote, Albert Szent-Györgyi von Nagyrápolt, a Hungarian biochemist, who was awarded the Nobel Prize in Physiology or Medicine, in 1937, and is credited with discovering vitamin C. He said “research is four things: brains with which to think, eyes with which to see, machines with which to measure and, fourth, money.”

Entrepreneurial innovation, out-of-the-box confluence of ideas and unusual paths to social businesses, must be captured by UF. Segments may be ploughed back into UF accelerators to boost even more imagination, invention and convergence of innovation. The thought of profitability, however ethical, is an anathema for academics and a paradox for non-profit, land-grant, sea-grant and space-grant nature of major US state universities.

As a top 10 university, we may not lack in capacity, in any dimension, in any metric, except for the number of Nobel Prize winners on faculty, when compared to private institutions with endowments of $35 billion, Harvard, or $22 billion, Stanford, or $15 billion, MIT, to name a few of the prominent names.
Stanford inventions generated $65 million in income in 2009. Stanford’s total earnings from inventions were $1 billion during the past 4 decades, with more than half coming from 2 inventions: the hypertext search and recombinant DNA technology.

UF is not inexperienced in this domain but we must be more proactive. We must look for the next Gatorade. The mixture of common everyday ingredients, which packed an uncommon profitability, is made of water, sugar, lemon juice, sodium, potassium, and phosphate. It has yielded the University of Florida, more than $80 million, since 1973. The abstraction latent in this mixture – common yet unconventional convergence – is the paradigm which we must emulate, innovate and differentiate a non-formulaic approach.

Manufacturing, healthcare, energy and global goods, are a few of the domains in our connected view of the networked physical world. Appropriate integration, with digital trends, data analytics, operations and online education, will follow. I shall summarize the ideas in each of these domains.

I request that you try to abstract the systems approach, in each effort. The design metaphor – atoms to bits – was suggested by Herbert Simon at CMU in 1987, but the phrase was used by Hiroshi Ishii, at MIT in 1997. After a quarter century, perhaps, it is time to extend the concept and propose a new design metaphor – swappable atoms. I will mention the context of swappable atoms, near the end of my speech, today.

Manufacturing is no longer limited to your grandfather’s steel mill in Pittsburgh. Today, manufacturing is a platform to aggregate creative ideas, new tools and applications. Students and faculty from different departments can use this platform to cross-pollinate. By creating corporate partnerships, we can collectively add value to the manufacturing industry, in many sub-verticals, for example, aerospace, automobiles, devices, prosthetics, sensors, energy and tools for the telecommunication and semiconductor industries.

Manufacturing has grown to become a catch-all term. As I lay out the suggestion, one or more areas of interest will rise to the surface. It may span from nanotech to mega machines. Let us fast forward to a term that is now a part of our daily vernacular, 3D printing. We propose the creation of the UF “Digital Foundry” as a complement to the 3,000 year old industry of casting (investment casting, sand casting). The UF Digital Foundry will break new grounds for integrated manufacturing, from CAD models to finished products. It will involve multiple arenas, including distributed cyber-physical production systems.
3D printers are now sold by Amazon and can be bought for the price of a dinner at Harry's. UF alum Erin Winick gained entrepreneurial attention by printing toys and beads for sale. We are not talking about the convenience store 3D printed mugs or plastic jewelry. We propose a robust materials science approach, working with scientists, engineers and corporations, involved in creating parts and tools for the aerospace, auto and biomedical industry. In addition to metal and ceramic 3D additive manufacturing, we hope to explore far and wide, including, embedded printed electronics, soft sensors, nano-bio-sensor-array integrated packaging, prosthetics and sol-gel printing (for example, skin for burn victims).

The barrier to 3D additive manufacturing for metals and alloys is due to the fact that a majority of the more than 5,500 alloys in use, today, cannot be additively manufactured, because the melting and solidification dynamics, during this process, lead to cracks in the material. Hence, very high value components for heavy industries, rely exclusively on investment casting to create these parts. The latter creates cycle times of months to years, starting from CAD drawings to a final cast, which must be validated and certified prior to production. The process could become prohibitively expensive due to iterative design changes, testing, metrology and concurrent 3D engineering reconfiguration. High cost and reduced agility, handicaps the ability of manufacturers to compete with low wage nations. We are losing high paying jobs and transferring knowledge base to other countries.

We propose that UF takes the steps necessary to retain additive manufacturers in this country, create a cluster for additive manufacturing and bring new jobs to grow the economy through the UF Digital Foundry initiative. Academic-industry-government collaboration can boost additive manufacturing 3D printing by reducing the cycle time and cost by half, or even three quarters, by facilitating rapid prototyping and manufacturing for final product assembly, testing and certification. 3D printers can create parts directly from CAD drawings transmitted to 3D printer without the need for co-location (distributed systems). It will speed up iterative design changes and variant configuration of components by allowing (crowd-sourced) designers, researchers and industrial experts, located in different corners of the world, to collaborate and 3D print various prototypes, in a matter of hours, obtain feedback, refine designs and identify design errors. Using machine learning, the 3D printed product can be self-checked to make sure that the 3D printer is delivering according to specifications. These processes may be repeated until component design and quality expectations are satisfied, in terms of the structure and function of the component. The final design version can be mass 3D printed or the design used for investment casting in order to mass produce by conventional methods, depending on the total cost of operation.
Digital Foundry at UF may enable the manufacturing industry to create the tools necessary for this process. In one vein, the convergence of nano-particulate chemistry with material science of alloys. Recently, software analysis of over 4,500 different alloys and nano-particle combinations, uncovered two aluminum alloys (7075, 6061), when coated with hydrogen-stabilized zirconium nanoparticles, and additively manufactured using SLS (selective laser sintering), showed no signs of cracking and demonstrated strengths comparable to conventional wrought materials.

Digital Foundry at UF can take orders from any part of the world, to create anything, if amenable to the platform. Conversely, the platform can serve as a medium or a clearing house, for design orders, to be channeled to 3D printers in remote parts of the world, closer to where the component may be used. It calls for a revision of cost structures with respect to logistics and transportation of materials, components and spare parts. The use of the Digital Foundry platform will allow experts from UF, and any other part of the world (crowd sourcing), to influence concurrent 3D re-engineering, to improve the design or recommend changes to the structure, which may optimize the function. These pay-per-services *modus operandi* may be used by millions of small and medium businesses, which may not have in-house experts, yet engaged in manufacturing, eg, China, India, and Brazil.

UF Digital Foundry is expected to become a node in the value chain and global supply chain of manufacturing for high value items, with metal alloys and ceramics, for the aerospace industry, hyperloop, vehicles, military-industrial complex and machine tools.

The academic contribution of the Digital Foundry to students at UF will be immense. Students graduating with engineering degrees will be skilled in the complete end-to-end scenario, from design to manufacturing, applicable to any product. The core competency of UF may not be in the actual product (it can be, we don’t know) but UF will certainly lead the R&D and data analytics services platform. The value proposition of the UF Digital Foundry Collaborative will be rooted in its research credibility, unbiased analytical tools and trusted advisory services. We expect financial returns from these activities to fuel UF Accelerator Programs and feed forward to enhance faculty research, student projects and global dissemination, through online education, skills training and workforce development.

I hope that UF Digital Foundry, in about 5 years of its commencement, will grow to become involved in, and penetrate to capture, 0.1% of the global parts and components manufacturing market, in some form or the other, related to additive manufacturing (AM).
We wish to earn $100 per aviation component or part. Commercial aviation industry manufactures about 3,000 aircrafts each year, with 300,000 parts per aircraft but 3D-AM printed metal parts may constitute only 1% of parts. Thus, total number of new 3D-AM parts manufactured per year is 9 million. If UF claims 0.1% of 9 million at $100 per part, then Digital Foundry revenue from aviation related components may be $0.9 million a year.

What I just explained to you is the conventional wisdom of product sales, applied to the radically unconventional Digital Foundry concept, in the midst of a tsunami of digital transformation. Incongruent. Isn’t it?

Let us re-try the scenario and this time let us break the mold. Forget about products and one-time payments based on product sales. Let us try the service model where we seek payment for each event, for example, each successful flight by the aircraft. Rolls Royce no longer sells the aircraft engine (turbine) as a product but sells “thrust-hours” as a service to the buyer (airline manufacturer, such as Boeing). If the engine is not generating thrust, it is not working and the plane is not moving. Hence, each time the airline flies, thrust is key and Rolls Royce will charge for thrust-hours, which means, you pay-per-use.

The product concept of sales is likely to die a very slow death, at the hands of digital transformation. We are entering an era of ubiquitous service, where all products are simply the “Trojan Horse” to deliver the service. If you let the product in, you must pay for service.

Let us think different and adopt the service model for long term stream of micro-revenue. Let us change our mind set from products to services. From our example, let us still assume we will penetrate only 0.1% of the market. From a service point of view, there are about 100,000 flights per day. Hence, 0.1% market means only 100 of these flights may have UF DF 3D-AM printed parts on board. Instead of charging $100 per part as a product, we will charge $100 as the pay-per-flight fee, for each successful flight. For 100 flights per day, anywhere in the world, the revenue will be $10,000 per day or $3.65 million per year. By moving from product cost to services pricing, we increased our revenue from aviation market, alone, from $0.9 million to $3.65 million per year, every year. Think long term!

Sounds, good? Let me rain on this parade. The chasm, between grand ideas and successful monetization, is deep and wide. Orchestration of multiple domains, replete with their own dynamics, volatility and uncertainty, requires cohesive collaboration, to bring together diverse people, disciplines and cultures, to work together, towards a common goal.

Rosalyn Sussman Yalow & Norman Ernest Borlaug Memorial Commencement Address by J. Fox Kenneth, President of the University of the Future (UF), on November 3, 2018
The science and engineering research that must precede the 3D-AM printing is our core competency as a University. We need electrical engineering, computer science and chemical engineering to mesh and meld with material science, mechanical engineering and biomedical engineering. External collaboration is necessary to translate our technical expertise to an actual product. The latter is the outcome expected from creating the UF Digital Foundry initiative under the umbrella of manufacturing. We can create the part, test, evaluate and partner with certification agencies to validate the widget.

What must follow for successful monetization is the adoption of the product by aerospace industry. This is at the edge of our competency which we may not influence and domains we cannot control. Even if our product is sufficiently differentiated by quality, cost and value, how can we guarantee that it will used in the manufacturing process? This is where alliances will be critical yet these alliances will not come into effect unless we have the product. The critics will ask if the adage “if you build they will come” may still hold true.

The one-time sale of goods (aircraft and spare parts) may not be our preferred path. For the manufacturing behemoths in the aerospace industry, our potential clients, agility is not a household word. Hence, the pay-per-use service model, each time the product is in use, is a different concept. Most service models are associated with standard maintenance contracts (MRO) and based on servicing the aircraft. Here, again, the pay per flight model is a business innovation which UF can standardize and catalyze its adoption in the aviation industry. The road block is actually an opportunity. The latter may be a prerequisite if we expect to claim a fraction ($100) of the per flight revenue, if $100 is a reasonable amount, after the manufacturer completes the transaction cost analysis and revenue sharing model.

The revenue will be paid by the user (airline) to the manufacturer. The aircraft manufacturer must honor revenue sharing. What is our visibility into this process? To answer this question, in the context of digital transformation, we must address another enormous arena of multi-disciplinary R&D potential, referred to as **Digital Twins**.

Digital Twins evolved from NASA experiments, to reproduce on earth, exactly what may be happening in a spacecraft, millions of miles away, in space. Visibility of physical processes are no longer limited to space explorations. The quest for low cost (TQM/TCO) operations, made it imperative that manufacturing must be globally distributed. Development of semi-autonomous mass manufacturing further eliminated humans in the loop to improve quality control.
Remote visibility of distributed manufacturing and remote operations, for example, oil rigs, ushered in the need for a digital replica of the physical machines and operation of plants. Thus evolved the Digital Proxies or Digital Duplicates or Digital Twins, to capture the operations and their status in near real-time, through 3D model visualization on a remote screen. Today that screen is an iPhone or an iPad. Behemoths like GE can now monitor its machines operating in Dalian, China from GE headquarters in Boston or GE GRC in Niskayuna, NY, in near real-time, 24/7/365.

Digital Twin driven remote monitoring, in near real-time, offers performance optimization of machines, reduction of waste, predictive maintenance to reduce downtime. The data from machines, “swarms” of machines on the shop floor, plants and massive operations (refineries, oil rigs, water filtration plants) are associated with traditional operations, for example, enterprise resource planning (ERP) solutions. The data from the Digital Twins, therefore, can vastly improve the digitalization and/or workflow automation of supply chain, product life cycle management, spare parts inventory, transportation and logistics planning.

Digital Twins are the data proxies of the entire operational world, any operation, anywhere, can be, theoretically, captured and displayed on any screen. We can peer inside coal mines, to cobalt mines, submarines and space stations, your dog in the veterinary clinic and your Mom in the hospital, just after her gall-bladder surgery. With the right platforms, with the appropriate interoperable software architecture, with seamless handover between telecommunication protocols and user-friendly graphic user interfaces, we can visualize on a smartphone, the temperature inside the turbine in a 787 flying over the North Pole or query the end-tidal carbon dioxide level and respiratory rate of your Mum recovering in the post-operative care unit.

Digital Twins represent the tip of the visible iceberg, which is an outcome of digital transformation, sweeping every industry, every domain, every facet of modern life. Hence, digital transformation, albeit amorphous and evolving, may be a key bullet point on our list of UF innovation initiatives. Digital Foundry is not an independent separate concept. The journey from CAD models to 3D printed end product requires digital transformation, too.

The world needs an army of experts in computer science and math. We need engineering experts from every field and specialty, we need integration of cyber-security by design with software innovation, we need convergence of hardware and software in distributed cyber-physical production systems, perhaps even on nano-satellites in space.
Digital Twins, in practice, may be the tool to track and trace the part produced by the UF Digital Foundry. Unless we can uniquely identify the part and its status (sitting on a shelf or in a plane at the gate or in a plane which is in flight), we cannot get paid because payment is based on pay-per-use. Therefore, we must know when the product is in service.

To access the inventory of the manufacturer and the “in use” status, which will be controlled by the airline and the FAA, we will need multiple levels of secure access through a cacophony of systems, which must be interoperable to catalyze transparency. The latter in turn, will enable us to acquire the track and trace data from each node of the operation. It may be complex as a process but technically it is possible, by combining automatic identification systems, such as, RFID or radio frequency identification, plus sensors to monitor the status and link the visibility of the data, in real time, from the physical object to the Digital Twin.

This digital by design metaphor has taken the world by storm and is referred to as the Internet of Things or IoT. Digital Twins, digital transformation and various parts of other forms of connectivity we will discuss, including healthcare and medical IoT, may be collectively a result of the IoT era. IoT is not a tool or technology. IoT is a concept, it is an idea with origins in embedded systems, ubiquitous computing and connectivity. Diffusion of the IoT concepts may have started early in the 20th century. The term IoT was coined by the MIT Auto ID Center in 1999 and their first paper “Networked Physical World” by Sanjay Sarma, David Brock and Kevin Ashton, is generally viewed as the birthplace of IoT. Various other descriptions have emerged and industrial internet of things (IIoT) has gained quite a bit of momentum since the founding of the Industrial Internet Consortium on 3/27/2014.

At UF, we are proceeding, albeit gingerly, on the IoT trajectory, through engineering domains, to build a comprehensive momentum. In this amorphous zone of IoT, as a design metaphor, a clear direction is the practice of connectivity, almost, without boundaries. This is no longer only embedded systems, software or electronics but the convergence of all these areas with many others, such as natural language processing, neural networks, remote sensing, biomedical engineering, medical devices, automobiles and anything where sensing and connectivity may add value. The grand explosion has ushered the crucial need to secure the connections, data transmission and information dissemination. Cybersecurity by design, hence, is a foundational layer. Hence, “pay per cybersecurity as a service” has the vast potential to become a must use service, if any form of telecommunication is involved.
IIoT or the industrial version of this explosion, cannot even take its first step without cybersecurity. Data breach, or intrusion in data communication, can be fatal.

In the design of IIoT, data from active components are coupled with objects (things). When this data, for example, sensor data, is fed to the relevant physical object (sensor icon in the digital version) and the processes, taken together and displayed in the digital version, is referred to as a Digital Twin, that is, a functional digital replica of the physical object complete with the status of the physical components (turbine in the aircraft, parts, the landing gear complex, water pumps, HVAC). But, it is not a static model or a simulation of the physical structure. The Digital Twin is a visualization platform for near real-time status of the functions that physical structure is expected to execute. If an aircraft is flying, then the Digital Twin will display, upon query, the actual temperature inside the turbine.

Imagine if this temperature data is compromised during the communication. Algorithms in the optimization function can change the rate and flow of the fuel, which could over-heat the turbine and cause the insulation to melt away. Reduced burn rate could decrease thrust and plane can lose control, in flight. The role of cybersecurity is critical. It is creating next generation of Cognitive Digital Twins and it includes cognitive firewalls.

The “live” Digital Twin concept is central to our current thinking and the tsunami of digital transformation we shall experience. The manufacturer or operator or engineering systems designer, wants raw data, in 5 seconds interval, of the range of temperatures inside the combustion chamber, while the aircraft is flying at 37,000 feet over Polynesian Islands. The temperature sensors, behind the ceramic tiles that layer the combustion chamber of the turbine, is sensing the temperature. The sensor operating system (tinyOS) may transfer the data to the sensor database system (tinyDB) which uploads the MEMS data, in real-time, to the private manufacturer-specific cloud hosted on an Accion nano-satellite, in the low earth orbit (LEO). Accion transmits the data from the nano-satellite to an AWS secure private cloud server in Reykjavik, Iceland, which auto-updates the laptop computer at UF, where we have installed the Digital Twin repository. When I point to the image representing the turbine and drill down to the parameter, temperature value, the temperature data is displayed on my iPhone screen. Behind the scenes, solvers using various algorithms and machine learning tools are computing the data to optimize function.

This Digital Twin is alive. I observe the fluctuation of the temperature range while staring at the iPhone screen for a few minutes.
Margaret Hamilton, in the mathematics department at UF, and our engineering systems expert Gwilym Jenkins, demonstrates how the temperature data is used in a machine learning (ML) algorithm. The solver also takes into account the molecular composition of the fuel and real-time humidity data from the in-flight onboard hygrometer. The outcome of this complex calculation using ML is a dynamic optimization protocol to adjust the fuel pump pressure, which influences the rate and flow of fuel through multiple fuel injectors, to deliver the fuel to the turbine combustion chamber. Swati Patel, a student working with the faculty team, explains the target is environmental conservation by reducing direct and indirect greenhouse gases (GHG). Temperature monitoring is an indicator of energy efficiency of the aircraft turbine engine, while in flight. The goal is to optimize efficiency – maximize the aircraft motion while reducing NOx emissions (indirect GHG) without increasing carbon dioxide emissions.

I have digressed, intentionally, to capture a few strands of the fabric of the world, evolving constantly. The UF Digital Foundry initiative is one layer among several cascading layers of the ecosystem.

We have to deal with the ecosystem. We cannot simply create a widget and leave it there, because the rest of the process is not our domain. We have to get into these domains, through collaborations, alliances and associations. The world needs answers, not numbers.

But, we must advance research to get to the "numbers" and, in parallel, educate the citizens of the future, at the University of the Future, to sense the big picture, to contribute the details and develop the ability to continuously oscillate between these worlds, when necessary, with ease and confidence, to generate answers, provide global solutions and lift the plight of humanity.

This is not an easy task. But we will pursue these goals and "do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept, one we are unwilling to postpone, and one which we intend to win."

To recap, UF Digital Foundry initiative, using back of the envelope arithmetic, offers the potential to increase our revenue, from the aviation market alone, from $0.9 million to $3.65 million per year, every year, if we can move from product costing to services pricing.
60 million commercial automobiles are produced each year, with about 30,000 parts, each. Assume 3D printed metal parts may constitute 1% of the whole. Thus, total number of new 3D-AM parts manufactured per year is 18 billion. If UF claims 0.1% penetration into 18 billion parts, at $1 per part, UF Digital Foundry revenue from automobile parts may be $18 million per year, in the traditional product approach.

In the pay-per-service model, we charge $1 per mile, because consumers pay for service when they use the automobile. If the user, drives the vehicle, the user pays $1 per mile. We understand that vehicles may be driven by people or may be semi-autonomous or autonomous and may have internal combustion engines (ICE) or electric vehicles (EV) or fuel cell vehicles (FCV). What may not change is the fact that they will travel. Assuming it is driven 12,000 miles per year per vehicle, 60 million vehicles, taken together, will travel 0.7 trillion miles per year. If we charge $1 per mile then UF may earn $0.7 trillion per year.

The product vs service approach enables us to promote ethical profitability and lowers the barrier to entry, into economic zones of billions. Instead of $1 pay-per-mile service fee, let us charge $0.01 or 1 cent (1 penny). If we charge 1 penny per mile then UF may earn $7.2 billion per year, each year, every year, from UF Digital Foundry auto parts.

Hip and knee replacements may be made to exact specifications using MRI data feeding 3D-AM printers. Assume we wish to earn $1 per replacement. If 1% of the susceptible global population required hip or knee replacements, that will be about 30 million people, worldwide. If UF claims 1% of 30 million joint replacements at $1 per part then Digital Foundry revenue from 3D printed hip or knee joints is $0.3 million per year.

Alternatively, we could use the “Miss Moneypenny” strategy and charge 1 penny-per-day for each joint as long as the recipient is using the replacement joint. Let us assume that an average post-replacement longevity is 10 years. The recipient pays one penny every day for 365 days per year for a total of $3.65 per year. For UF, 1% of the 30 million market amounts to $1 million per year in revenue, at a rate of one penny-per-day per replacement.

In 2013, 767 million people (10% of the global population) lived on less than $1.90 a day. Our quest for one penny-per-day may be affordable for all individuals in the world. But, please don’t rush to gush with unbridled optimism. That 10% of the world may not even have access to clean water. About 50% of the world population may not have access to formal healthcare and more than 1 billion are without any healthcare, whatsoever.
The reports from a few august organizations claiming that only 400 million people lack access to essential health services is an example that may belong to the 21st Century edition of the seminal book “How to Lie with Statistics” by Darrell Huff.

Healthcare is a complicated problem. It will be arrogant and audacious of us even to think that we can alleviate the ills which has infected healthcare. We are only scratching the surface of a very tiny area. The biotech / bio-prosthetics sector in UF may benefit from 3D-AM (printed devices/skin/teeth, dental implants, implanted or ingestible nano-sensors).

The examples from the UF Digital Foundry initiative stimulates the core mission of education and research. Micro-revenue will boost UF development, provide the school of management with new challenges to help manage labyrinth of operations, globally, to sustain social businesses. Coupling rigorous research programs with entrepreneurial innovation, global education and ethical profitability, calls for new models that may not have been conceived, yet. Exploring ways to reduce cost of transaction (transaction cost economics) may benefit from the ecosystem of digital transformation, data and analytics. The central role of software, systems, data and analytics, will need robust cybersecurity.

Data analytics and cybersecurity demands depth of knowledge in mathematics, computer science, statistics and how to use tools from artificial intelligence (the truth, not the hype). The “ammunition” necessary may include knowledge representation, symbolics, representational learning, machine learning, deep learning and other AI techniques.

To raise that army we must increase enrollment of women in STEM fields in UF. But we can’t stop there. Convergence, cross-pollination and trans-disciplinary collaborations are quintessential to create solutions. Technologies, tools, papers and monographs are useful but the solutions approach require we dissolve silos of data, information and knowledge.

We need more women to pursue statistics, computer science, basic sciences, engineering, agriculture, medicine, public health, economics and econometrics. For all students, we need to emphasize depth but concomitantly an awareness for breadth. The academic expectation of students to learn more and more, about fewer and fewer things, is pivotal for depth and rigor. But, the new generation must also remain cognizant about the global need to connect and converge, more and more things, from wider and wider circles, in order to provide the end-to-end solutions, which improves quality of life and living.
This amorphous tapestry may not be a defined system but system of systems or even more pertinent, complex system of ecosystems. The principles and practice of connectivity, convergence and complex systems, are here to stay, and amplify.

The supply chain of talent, necessary for these transformation, are not unique to UF. We must tune the engine, if we hope to polish the chrome. UF outreach to middle schools and high schools, must be at the heart of the UF strategy to be a purveyor of civilization. UF outreach to teachers in the elementary and secondary education system, must rise far above the shoddy and second grade masquerading, as good enough. There is no room for good enough, if we aspire to educate, shape and transform the future. The UF promise to build, maintain and secure the Cathedral of Learning, cannot happen in a field of ignorance.

In summary, the manufacturing proposal of the future has been given a name, we refer to it as the “Digital Foundry” but if we elaborate, we may think of it as a decentralized cyber-physical production system for additive manufacturing using 3D printing tools. The scope of the UF Digital Foundry is not limited to materials science or computer engineering or embedded systems or MEMS or NEMS or cybersecurity but also relevant to medicine, business, and social sciences, as well as social re-engineering and social enterprises.

UF will need resources, collaborations, partnerships with companies, governments, VC’s. But, ultimately, success depends on the zeal of the faculty and talented students, to create not what has been created but create a resplendent sense of the future, using the Digital Foundry initiative as one tool, a catalyst, an enabler. It is a means to an end, but the end is not an end, literally, it is a metaphor for a new beginning, a continuum, in which we use science to help society, to lift many boats. It will be driven by you, the students, and you, the faculty, but only if you think differently, only if you are not afraid of failure.

Now, I shall turn to health and healthcare, as an equally important engagement, where UF is already a champion but there is always room to evolve. Health and healthcare are related but not synonymous, a distinction that I hope to illustrate, in a few minutes.

The evolution of technology has created an illusion of proximity. After the telegraph and the telephone, it was AOL’s Instant Messenger, or IM, and “you’ve got mail” which signaled the birth of a new era in communications for the public, at large. Today you cannot find a communications tool without an “instant” component, to advertise yourself. The selfie storm appears to be a category five hurricane.
Narcissism aside, we use this electronic space as a buffer, a justification to increase our emotional distance, perhaps for some, to protect the fragile nature of disequilibrium in our micro-space. Are we reducing our ability to engage in inter-personal interactions? It is a vicious social cycle which further fuels the nuclear ethos and helps to build walls. There is less and less “skin” in the game. It is more and more cyber, a declining preference for, even, ethical physical presence, a social discomfort for civilized touch. We are sporting colors of camouflage by embracing the aphorism that only words can touch places where touch doesn’t know how to enter. Hence, a veiled and calm, yet vociferous denial of all things personal, extending to personalized services.

Healthcare is under siege by these forces. Denial of personalization even wants to target the role of the primary care physician, citing reduction of cost of operation, as the benefit. In the October 17, 2017, issue, the august journal Nature chimed in to replace primary care using artificial intelligence. The stupidity of publishing may be excused by acknowledging Nature’s desire to exercise its First Amendment rights. In the article, the author’s lack of understanding of the role of primary care physician in health, healthcare and society, was paraded as the march of AI and cost-savings. The capitalistic pressure to focus on cost erodes the socio-economic concept of relative value. The short term gain far outweighs the long term vision. Socio-economic parameters of primary care are a small part of a complex network of factors which shape quality of life, not only morbidity and mortality.

The University of the Future initiative for health and healthcare shall take the long road. Improving global quality of life will serve as our target. It is a lofty target, a complex vision and a grand challenge. I urge you to reflect on this quote from Daniel Burnham, an architect. But, I want you to do more, in addition to the written words, I want you to reach out to the inaudible and the unwritten allegory and allusion these words may provoke or evoke. “Make no little plans; they have no magic to stir men’s blood and probably will not be realized. Make big plans, aim high in hope and work, a noble diagram, once recorded will never die, but long after we are gone will be a living thing.”

This commencement speech is the “record” I wish to share with you. It may stir new ideas and make you see old things with new eyes. It doesn’t matter if ideas in this speech never sees the light of the day. What matters is that these ideas may ignite new thoughts or shed some light how to transform your vision into reality. This is not my plan for a “me” “me” “me” modus operandi.
I request you think big, dream bigger and ACT, where T stands for “things” and C stands for “changes” and A stands for “action.” Remember Action Changes Things (ACT).

The internet of things (IoT) will remain an utopian design metaphor unless we ACT to embed this design through the principles and practice of connectivity. Ubiquitous connectivity, ubiquitous computing and computing at the edge are ideas to connect atoms to bits. All things physical, all processes, all decisions, may eventually be a part of this networked physical world. The data from these plethora of sources, the bits, will be connected to the things, events, processes and decisions, and that, as a whole is the practice of atoms to bits. Do you think it can happen without ACT?

The pursuit of ACT must touch all facets of life. For healthcare, it is the lack of action, lack of common sense, lack of standards, lack of leadership, lack of our zeal to fight greed, which are a few of the reasons why nearly half a million people, may die in the US, each year, due to errors. These errors arise due to lack of interoperability and data sharing between medical devices and patient data, in medical records, which cannot talk to each other. This is a gross injustice. But is it a remediable injustice? You must decide and ACT.

Imagine, today, you use your ATM card in the Student Center to take out cash, before your trip to Mongolia. Upon arrival, using the same ATM card, you take out money from an ATM machine at Sükhbaatar Square, in Ulaanbaatar, Mongolia. You can take out money in Gainesville, FL and Ulaanbaatar, Mongolia, using your globally interoperable ATM card.

But, here is a less known fact. Your electronic health/medical records (EHR/EMR) cannot be used in a different hospital across the street. Assume your records are in the UF Shands system and you are unfortunate enough to be in an accident, when visiting your friend Josh, in Jacksonville. You are rushed to the emergency room at the Baptist Health Center, in Jacksonville, FL. The ER staff desperately needs your medical data. But, your medical records cannot be accessed by the Baptist ER or Trauma Center, even if you consciously permitted the staff at Baptist to access your EHR/EMR.

The medical records software and system used by these two hospital systems, are not compatible or accessible from another system, due to lack of interoperability between systems, which was deliberately configured by design, by the manufacturer of the software systems, the vendor who sold the EHR/EMR. The same applies to vast number of devices.
You may ask, if this is the biggest problem in healthcare? I recommend that you ask a better and more specific question. Framing the correct question is important. I suggest you ask, is this the biggest problem in US healthcare? My answer will be quite incomplete.

But, I would reply, no, it is not the biggest problem in US healthcare, it is the THIRD biggest problem.

You can easily imagine that problems in healthcare are immensely complex. I am neither going to dissect the problems nor suggest that UF feels obligated to address the framework of all these problems. I suggest we view our role, contribution and use of our brains, to address segments of this system. Remember, our goal is to improve quality of life.

The dedication of the students, faculty, and staff at UF can find ways to get the world congregate behind this worthy cause. I think UF can improve the quality of life, for at least few of the nearly half million people in the US, who may die due to medical errors. There are many problems with healthcare but as Mother Teresa of Calcutta, often remarked, “charity begins at home.” Let us pledge to help decrease the number of deaths due to medical errors, in the US. Even if we can reduce the numbers by only 10%, we have saved 25,000 to 50,000 lives.

The published figures are somewhere between 250,000 and 440,000, deaths due to medical errors, per latest reports from the Institute of Medicine (IoM). According to the seminal study by Johns Hopkins University, published by the British Medical Journal, in 2016, if deaths due to medical errors per year, are taken at their lowest numbers, that is, 250,000, per year, then, deaths due to medical errors in the US are the THIRD leading cause of death after cancer (585,000) and heart disease (611,000).

Hence, my reply, earlier, it is not the biggest problem in US healthcare, it is the THIRD biggest problem.

My suggestions, plans and views may be incorrect. I wish to reiterate that this is not about me or the President of UF or institutional advancement. This is about the granularity of talent that is embedded in each one of you. This is about your eye on the target to help improve quality of life. This is your challenge to find ways and ACT to implement the tools in a manner that is safe, secure and sustainable. This is not going to be easy.
That is why I urge you to pursue this goal, because it is going to be hard, difficult and complex. You will meet insurmountable barriers that you must overcome, you will face political quagmire you must navigate with diplomacy, you will face the wrath of the medical-industrial complex, you will feel the burden of the world on your shoulders, yet you must rise to meet the challenge. When you think about the very hard road ahead, and whether you should make it your mission, or not, I recommend you reflect on the ethos, semantics and eschatological wisdom attributed to the words from Matthew 7:13 and 7:14.

My faith in the ability of UF students, faculty and staff to ACT, is rooted in the fact that death due to medical errors have solutions which can evolve from engineering, science and medicine. The attempt to alleviate this moral, ethical and inhuman disaster of death due to errors, may be held hostage only by our inaction, lack of intervention and failure of implementation. UF and its medical excellence in the US and affiliations we wish to create, globally, may strengthen UF’s leadership to prevent this march of unreason.

Let us begin with relatively simple things which may be the lowest common denominator. Medical devices are data harvesting tools. When data from medical devices are acquired, curated and appropriately integrated for analysis, in context of the person or patient, it offers us information. Part of this information may add to the person’s medical knowledge base (EHR/EMR), a part of this information may offer predictive analytics, to help in prevention, and yet another part of data analytics may trigger immediate action.

When you are in an automobile or aircraft, you don’t even think why you should integrate the data from the fuel consumption meter, the transmission fluid chamber, tire pressure, brake fluid level, the residual battery charge, wind velocity, speed and motion. We take systems integration for granted, and of course, structural integration for function.

Health and healthcare, is the outcome of systems integration, and an integrated structure, optimized for function. We wouldn’t be here, there would be no evolutionary development of organisms or homo sapiens, if biological systems lacked integration.

With the advances in medicine, there must be a concomitant advance in systems integration for medical devices, working outside the body, with the intent to help restore equilibrium, in some manner, inside the body. The medical devices outside the body cannot remain oblivious to the basic need for an integrated outcome, for the whole patient or the person, not only some for parts or organs.
The failure of such an integrated outcome, related to the external processes, is one reason for the nearly half million deaths, per year, due to preventable medical errors.

For example, a 49-year-old woman underwent an uneventful total abdominal hysterectomy (removal of uterus and cervix through an abdominal incision) and bilateral salpingo-oophorectomy (removal of both sets of fallopian tube and ovary). Post-surgical pain management was initiated with intravenous morphine sulfate, in small increments. She began receiving a continuous infusion of morphine via a patient controlled analgesia (PCA) pump. A few hours after leaving the PACU [post-anesthesia care unit] and arriving on the floor of the general ward, she was found pale with shallow breathing, a faint pulse and pinpoint pupils. The nursing staff called a “code” and the patient was resuscitated and transferred to the intensive care unit on a respirator [ventilator]. Based on family wishes, life support was withdrawn. The patient expired. A case review affirmed death due to PCA overdose, a medical systems error which lead to morphine overdose through PCA.

Medical devices connected to this patient did their job of collecting the data, but did not do their job of connecting the data. Making sense of data is fundamental to our sense of the future.

Morphine administered by the analgesic delivery pump delivers morphine and it can be pre-set or controlled by the patient (patient controlled analgesia or PCA). Morphine reduces respiratory rate (RR). RR may be monitored using an inductive or resistive belt. The plethysmogram produced by the pulse oximeter, a device to monitor oxygenation, however, does not act until oxygenation (pO2) is affected. Generally, a heart rate monitor also displays respiration rate. But when RR is depressed, it reduces oxygen saturation in the blood (↓pO2) due to reduced diffusion in the alveoli (decreased RR leads to shallow breathing). Capnography, measures end tidal carbon dioxide (etCO2), an indicator of carbon dioxide which was not exhaled (residual) due to shallow breathing, caused by the reduction in RR, triggered by morphine (↑etCO2). Analysis of plethysmogram by wavelet transforms would enable the determination of changes in respiratory rate (RR) at an earlier stage (predictive analytics). If corroborated with an increase in end tidal carbon dioxide ((↑etCO2) / (↓pO2)) due to shallow breathing, the combination of data from two different devices, may suggest that the negative effect of morphine, on reducing the respiratory rate of the patient, is a far higher risk, than the anticipated reward, from the administered dose of morphine, to reduce post-operative pain (increase patient comfort).
Convergence of the data from the heart rate monitor, respiratory rate, pulse oximeter and capnography, must be integrated, analyzed and the status synthesized, in real time, in the context of the patient. The data analytics output must determine the dose and duration of morphine administration (control of pump), when the patient requests more morphine through the PCA. The raw data, waveforms, analytics and decisions, must be communicated in real-time, to the nurse manager or attending physician’s smartphone. Data integrity and security must validate situational alarms and “code red” signals.

This is not in place, today, in most instances. The data from these devices are not integrated and not synthesized to reflect the real-time physiological status of the person or patient. Today, the nurse manager, in the observation station of the ward, may glance at several monitors, for several patients, and may arrive at an ad hoc determination, of the actions to follow, the alarms to ignore, or the situations to escalate.

In our example, the 49-year old woman, continued to deliver/receive morphine through the patient controlled analgesia (PCA) device, to reduce her post-operative pain. The lack of integration of medical device data, the lack of information interoperability between devices and the lack of a platform to integrate the clinical environment, in order make sense of the data, in context of the patient, lead to her un-timely demise. This is the nature of the problem, a relatively simple problem, we can solve at UF, in order to reduce the morbidity and mortality due to lack of real-time data integration, data analytics and data-dependent decisions in healthcare. If there was an intervention, an auto shut-off should have been triggered for the morphine PCA pump, based on the integrated data and information. Then, the woman may be celebrating her 50th birthday, with us, here today.

On closer inspection, you can identify the domains that must come together to solve this problem: [a] biomedical engineering of devices, [b] embedded electronics [c] data delivery interfaces (application programming interfaces, API) from medical devices for raw waveform data, [d] wireless transmission of data in multi-user dense clinical environments which must address and resolve errors due to radio frequency signal collision, bandwidth consumption, jitter, bounded latency, and time guarantee [e] interoperable platform for integration of data, [f] analytical engines using basic workflow or machine learning (ML) to make sense of the data, [g] feedback systems and control automation for rate limiting nodes [h] visualization on 3rd party monitors [i] real-time mobile data connectivity to the medical staff [j] authorization, privacy and cybersecurity of data, analytics and information, [k] back-end synchronization with EHR/EMR [l] de-identified data for systems and census.
Did we just transform one of the problems in healthcare as a solution we may expect from medical IoT? Collect the bits about atoms and connect the data to gain information.

This is a simple design metaphor and if you thoughtfully apply this metaphor to any domain or any industry or vertical, you will find, in most cases, we shall return to the principles of systems engineering, the need to address system of systems. We must address the whole, not just a piece of the problem, in a complex systems approach we must address the ecosystem. Most initiatives shy away due to the sheer complexity and lack of resources.

There are many tactical, practical and pragmatic problems in health and healthcare where UF can provide leadership. I would like to appeal to each one of you to undertake this complex challenge, albeit in a modular fashion, where modules can be connected to synthesize the outcome, the whole outcome. I request you to make it our mission to help create the systems and ecosystems to develop, demonstrate, disseminate and implement the processes, tools, technologies and platforms, necessary to reduce death due to errors of integration.

This is the true convergence of minds with machines. This is the convergence of principles with practice. This is the convergence of research and its translation to serve society, to lift many boats, not just a few yachts. Can we ACT (Action Changes Things) to improve healthcare?

Health is a prime example of how we can turn our thoughts into action through sense and sensibility. Health is human health, health of animals, environmental health, plants and organisms. Our sense of the future must be an informed sense of the future. Our sense of the future, in practice, may turn into an euphoria for sensors to predict the state of the future. The latter has created the utopia where trillions of sensors permeate every fabric of life. Somewhere between utopia and euphoria lies the panorama that we must foresee, anticipate and cultivate. The journey doesn’t have a roadmap. We will encounter hills and valleys, failures and successes, joys and sorrows, dismay and exhilaration.

But, what we may have for the journey, our tryst with destiny, is a compass. Data is the new compass. If we cannot sense, if we cannot acquire data, if we cannot make sense of data, if we cannot convert data to information, if we cannot extract knowledge from information, then, we may not be able to guide our future, our future decisions. Without data, our imagination will be out of focus, and we may not be able to visualize what lies ahead, we may not be able to see the future.
A sense of the future, as seen through the eyes of sensor data, is one view of the transformation of paradoxes to paradigms. Agreed in 1997, the Stockholm Convention on Persistent Organic Pollutants (POP) was tasked to detect and eliminate POP. Today we can use low cost, pervasive deployment of sensors and other tools for remote monitoring of aerosols and the Finnish Meteorological Institute (Ilmatieteen Laitos) is leading the way to monitor the health of the planet.

We wish to sense everything, we wish to acquire data from [a] temperature sensors inside the aircraft engine sending real-time data to the digital twin of the turbine, [b] gas-detecting sensors on the ocean floor transmitting data about chemo-organotrophs and chemo-lithotrophs [c] sensors inside the stomach of cows alerting us about complex microbiome changes due to lactation, food intake and pesticides/herbicides in the pasture [d] sensor-guided micro-robots in the large intestine of humans removing benign polyps [e] surface enhanced Raman spectroscopy based nano-sensors measuring parts per billion of Arsenic in water in Bangladesh or Mercury contamination in artisan gold mines of Colombia [f] nano-wire nano-array sensors screening for TB, leprosy, Hanta, Ebola, Flu, Measles, Plague and other communicable diseases, at immigration check points, in airports of the world [g] embedded nano-sensors on flash drives connected to smartphones which measures blood glucose and transmits data to your nurse practitioner [h] graphene nano-ribbon hydrogen sensor in your fuel cell (FCV) automobile [i] nitrate/phosphate sensor for precision agriculture [j] ad hoc self-organizing mesh network of sensors dropped from low flying aircrafts to detect landmines in hostile territory [k] implanted sensor-controlled time dependent chemotherapy dose which can be remotely monitored [l] swarm of sensors with nano-copper catalysts converting carbon dioxide to alcohols for renewable energy. The list can go on and on. Sensors are going to be a part of the fabric, more or less, sooner or later.

Deploying trillions ($10^{12}$) of sensors may not be wishful thinking. One could imagine bigger numbers, such as, decillion ($10^{33}$) or centillion ($10^{303}$) of sensors. To keep it in perspective, note that we may have about 7 quintillion ($10^{18}$) of sand in the world. Let us assume that we will “sense” every grain of “sand” in the world and send this “sand” castle of data directly to the internet, to feed distributed analytical engines. For 7 quintillion grains of sand we need 7 quintillion $(7 \times 10^{18})$ sensors. If each sensor must communicate with the internet, directly, then we need $7 \times 10^{18}$ unique IP addresses (internet protocol). Can we handle that volume of direct data routing to the internet? If I had to answer this question before June 6, 2012, I would say perhaps. But, now I can tell you, yes, we have the ability to provide unique IP addresses for each of the $7 \times 10^{18}$ sensors, under discussion.

———

Rosalyn Sussman Yalow & Norman Ernest Borlaug Memorial Commencement Address by J. Fox Kenneth, President of the University of the Future (UF), on November 3, 2018
The latter is possible because as of June 6, 2012, the world has moved from internet protocol version 4 (IPv4) to IPv6. One advantage of IPv6 is its larger address space. The length of an IPv6 address is 128 bits, compared with 32 bits in IPv4. Because every binary “bit” has 2 states, 0 and 1, the address space was limited to $2^{32}$ for IPv4. The address space for IPv6 has the potential to accommodate $2^{128}$ or about $3.4 \times 10^{38}$ unique addresses.

Yes, we can sense every grain of sand in the world, if there is a good reason. Yes, we can take the space elevator to the Moon on a nanowire frame, if necessary, in the future.

The value of $3.4 \times 10^{38}$ unique addresses is not in our theoretical discourse about attaching a sensor to every grain of sand in the world. The real value of this discussion is the granularity with which we can track and trace events, and how the ecosystem partners can monetize miniscule nano-fractions of micro-payments, from social business ventures.

To alleviate hunger, poverty and hopelessness, in the world, we must gently transform our efforts from dependence on governments, philanthropists and foundations to solutions which are low-cost, sustainable and socially structured business ventures peddling prosperity for all, not a few.

We are back to making money. Today, our penchant for progress must keep pace with the inaudible stampede to gather resources. For academic institutions, our money making ability is entirely in your hands, the outcome of your brains and your desire to share your fertile imagination with the institution, in terms of your output, contributions and patents. UF must seek financial resources from every angle, because every penny must be re-invested in you. UF wishes to better enable you and provide an environment which may trigger, catalyze and fertilize even more opportunities for cross-pollination, synthesis, convergence and random sparks to ignite your imagination, inventions and innovations.

IPv6 *per se* is a routing protocol co-invented by Bob Hinden. But, the conceptual implications of the $3.4 \times 10^{38}$ unique addresses, which are a result of the hexadecimal format with sub-divided domains, may spark new thinking about formats and domain subdivisions which may be essential to identifying every instance, of every activity, for every transaction, for every person, related to every object, every process, every decision, and every bit, connected to groups of atoms (objects). In ancient times, a related concept was the hash table or hash map and use of hash function to compute indices. A related hype, is the use of blockchain. Today, blockchain is on steroids, for all and every purpose.
Nothing is a panacea. Hash tables are still very much in use. Multi-disciplinary research is necessary to figure out [a] how can we decompose a systemic solution to its granular parts (think beyond BOM explosion, bill of materials) [b] assign a monetary value for that part [c] pay the entity who developed/created/manufactured that part, if that part was consumed by users (the core monetization function necessary for the service model).

This *modus operandi* is essential for pay-per-use business models using micro-payments, micro-earnings and micro-revenues, as tools for economic growth, globally, for more than 6 billion people in the world, who are much less fortunate than most of us in US, UK, Europe and Japan.

This research and its potential integration in global business is crucial for many, including UF. I wish to advocate this mechanism as one of the tools for UF to use and claim one penny from each instance of work, which our research can contribute to the global economy.

Why do I keep returning to this one penny model? Because, we can achieve miracles if we are less greedy, if we let the rest of the world find and share resources to do what is good for them, if we are a partner for global progress rather than a scavenger. We must help ourselves but we must help others, too. Collective growth fosters opportunities, equity and trust. There is enough in the world for human need but never enough for human greed.

The “Moneypenny” strategy, applied to “trillions of sensors” idea ($10^{12}$), generates the potential to earn 1 cent every day from every sensor or US$3.65 trillion per annum. If UF could transform the vision and find ways to *insert itself* in 100 million sensors ($10^8$) then UF may anticipate to earn $10^8$ cents or US$365 million per annum. If these ideas are only 10% successful, then, UF can earn US$36.5 million per year.

Critics will be justified to question the sanity of UF producing 100 million sensors. I admit that it is well-nigh impossible for an academic institution to manufacture 100 million sensors. That’s not what we do. True. But, don’t forget institution-supported start-ups which can grow in a garage to become a gigantic enterprise. Think outside the garage.

Think how many different massive economies of scale use cases we may address. If 10% of India and China may buy a sensor to check for blood glucose, that is a market for 250 million sensors and the service associated with the lifecycle of data from that sensor.
If UF had a remote hold even in 10% of that market, we can “insert” UF as a service in the lifecycle of 25 million sensors. This application may generate about US$91 million per year for UF at a rate of 1 cent per day per sensor. Critics will jump up to point out that why would anybody allow UF to capture 10% of a foreign market where we do not have any presence, experience, credibility, or relationships, whatsoever.

The critics will be correct. UF needs an international advancement strategy and create research relationships, before we engage in potential entrepreneurial enterprises. Let me remind you that our strategy is not to disrespect or deplete the people in these nations. Our strategy is to help improve quality of life, our strategy is to build long term relationships, our strategy is to co-opt the talent in these markets to improve the economy. Remember, success is a rental. The rent is due every day. For our role, we request 1 penny.

Our claim for 1 penny per sensor per day is an open source approach, to penetrate any market in the world, by lowering the barrier to entry, to its most basic denominator. The per capita income in Kenya is about US$3,000 per year, India’s per capita income is twice that of Kenya and China is more than twice compared to India. At US$3,000 per year, it is not impossible that a person may spend 10% on healthcare. Out of the US$300 budget for a person in Kenya, can UF claim $3.65 per year? I don’t know. Is it impossible?

What may appear to be impossible, may just take a bit longer. We have to think big, bold and out of the box. We have to invest to generate revenue. But, above all, we need the leadership to promote the vision, the economies of scale by seeking granularity, not greed. It is not about blockbusters which costs $9,850 per month, for example, the cancer drug, Ibrance, produced and priced by Pfizer. It is about population-busters at 1 penny per day.

Creating population-buster sensors may not be difficult for UF but creating the ecosystem to take the sensors from the lab to the million level manufacturing, distribution, sales and return on investment, is complicated. This is why UF needs the relationships at every node in that ecosystem. The global ecosystem is not going to accept proprietary tools.

Monetization of open source models, patents and technology licensing, deserves attention. Data acquisition from devices, the data model and interfaces, are some of the essential “open” elements. Safety, security and privacy must be demonstrated in the open model. Authorized access to raw data must be kept open, and almost free for patients, hospitals, healthcare workers and anybody else who may be legally permitted to use.
The sensor is the hardware, one part of the package, where software managed services are central to the lifecycle value of the sensor (hardware). It may be paradoxical, in the past, to contemplate that one may buy an object but may not own all of it. The case with John Deere and the software used in the tractor to obtain data was the milestone. The farmer owned the equipment but not the data since the software in the tractor was only licensed for use, not a product for purchase. However unfair, this is the paradigm shift where we may buy less and less “things” and pay per service, when used. If you use your washing machine, LG will record the use and charge you for its use. You may never have to buy another washer or dryer. The paradox is now a paradigm.

In another twist to this model, open source healthcare platform vendors may not be able to legally restrict you from obtaining your data (raw data, just the numbers) but may impose a micro-payment for accessing the vendor’s platform to access your data. It is a model that the banks have been using since the inception of banking. The outcome of the banking model is right in your face – if you visit any US city and look for the nicest tallest structure or the most opulent building – it is a bank.

Health and healthcare services will incur a cost to develop, implement, maintain and secure the infrastructure to enable the transactions. It will be a quantum leap forward if the health and healthcare industry is able to obtain raw data from devices and use an open platform model as a tool for convergence. If it is your data, then, you should be able to access your health data, anytime, anywhere. But, perhaps, there will be pay-per-access nano-payment for service. For health purposes, the information from this data could prevent you from a heart attack or alert you to the possibility of glaucoma due to type II diabetes. For healthcare, the information from this data, integrated for your specific physiological status, can spell the difference between life and death.

The data-driven service is the key. The discussion is no longer about the sensor, the physical product, but the lifecycle of the sensor which is inextricably linked to its purpose to sense, acquire and transmit data. The sensor, as a physical object, group of atoms, is like music, silent, unless performed. Data as a service is the yardstick for performance. Hence, the service must accompany the sensor in order for it to deliver its value. This is the next level where UF must “insert” itself – the interdisciplinary domain of software as a service. For global operations, security, interoperability between systems, data models and formats must be amenable to “discovery” through standards and open source standardization. If you cannot “discover” the availability of a function, then you cannot connect. In software parlance, you can’t initiate a remote function call (RFC) without an API or routing protocol.

Rosalyn Sussman Yalow & Norman Ernest Borlaug Memorial Commencement Address by J. Fox Kenneth, President of the University of the Future (UF), on November 3, 2018
UF must accelerate its role in global standardization bodies. UF faculty are integral part of IEEE, ITU, TOGAF, DoDAF and other organizations. Do we need to do more in this space? Instead of creating millions of sensors as hardware, UF can contribute to create interfaces, architecture and semantic interoperability standards. These standards may become integrated not with millions but billions of sensors and the sensor data ecosystem which is critical to make sense of the data through data analytics. The diffusion of CORBA sponsored by the Object Management Group may serve both as an example and a lesson.

A person using a blood glucose sensor on a smartphone observes that the blood glucose level is 139 mg/dl and wishes to know what it means. This query may trigger a different application, which may require the services of an analytical engine to process the data and respond to the query. This analytical engine may be provided by the platform vendor or may be sourced from a third party vendor peddling apps on Apple/Google. This is the engine that can evolve out of UF brains. UF can insert an algorithm in this analytical tool via an app. The query may cost the customer 10 cents each time. If 1% of the world’s population (7.5 billion people) asked one question, each month, then this one single app could generate $90 million per year. The monetization model must distribute fractional micro-payments to the app creator, platform host, telecom provider and other service vendors. If the system required 30 different vendors to contribute for systems integration in order to answer that question, then each vendor may receive 1/30th of the $90 million revenue or $3 million per year (if revenue was shared equally among the partners). If UF had 1% share of the revenue, from this app, then UF earns $0.9 million per year per app.

Dynamic pricing of queries may be based on the computational complexity of the question. Dynamic composability of micro-services, analytical tools, algorithm engines or portfolio of solvers, must be sourced, aggregated, activated and executed in sequence, in order to respond. The structural elements of this architecture must be modular, “drag and drop” blocks to be “discovered” if necessary, interoperable, agile, compatible, and perhaps capable of natural language processing and preferably deployed as Agent based software.

Platform analytics will evolve as a key tool for monetization for open platforms and internet of things including medical internet of things (MIoT). Security, privacy and data de-identification may play pivotal roles. The question of “intelligence” in analytics and how the platform will address context of input data will always remain a challenge. Reasoning (often erroneously and synonymously referred to as AI) must occur in a specific context for a specific goal(s). These are questions which must be rooted in research outcomes.
The hype around so-called medical or healthcare AI systems accepting list of symptoms as input and generating output (most likely diagnosis) ignores the basic tenets of medicine: hypothesis-directed gathering of information, complex task of interpreting sensory data (obtain uncertain indicators of symptoms) and the goal of curing the patient. The latter may involve the decision to treat less likely but potentially dangerous condition rather than dwell on the likely but harmless condition. Is that the role of AI in healthcare?

These various steps and processes are events and instances that are proceeding in near real-time through the fabric of connectivity. The plethora of hardware and software related components, cyberphysical systems and Digital Twin instances, must be stitched, synthesized and orchestrated in sequence, to deliver a seamless service output, for which the end customer will be eager and ready to pay. That payment, at the end of the tunnel, is the only incoming revenue. The customer will pay if the customer is satisfied.

The customer will not pay UF, if UF contributes one algorithm, inserted in one app, which is hosted by Amazon cloud (AWS), in the app store for UCSF School of Medicine, distributed either through Apple or Samsung Health. This is similar to General Motors advertising their North Star onboard service. It is an outsourced operation hosted by GM to connect emergency services for humans and automobiles. Other value added services from external parties (restaurants, movie theaters, shopping malls) used the platform to push their offers, discounts and updates, based on location, proximity or customer’s preferences.

Trans-disciplinary research is necessary to create unique identification tools that specifies what is requested, what is delivered, who serves as the medium of delivery, who is receiving the delivery and how payment is handled. Each sequence must be identified. An entity (corporation, person, crowd-sourced, agency) will have created and contributed that sequence, as a part of the seamless symphony, delivered to the user. The creator, vendor or supplier will want a share of that revenue. That fraction is critical. It must be accounted in a manner that is trustworthy, credible and resistant to cyber-theft.

For example, consider a MEMS in an automobile. Let us assume that it is by design integrated as a system on a chip (SoC) with on-board tinyOS and tinyDB. In an alternative scenario, the MEMS could be just a MEMS device (devoid of any microprocessor). If UF created this MEMS, in principle, and it was licensed by Bosch, in principle, we, UF, must make money each time the customer uses the MEMS (classic pay-per-use service model). The end user is completely unknown to us. We have licensed the MEMS to Bosch Inc.
In order to earn the micro-revenue from the service through the MEMS, we can pursue the Digital Twins route. However, Bosch, may not allow us to access the Digital Twin. I am unclear how the “politics” will be resolved but let us follow the engineering steps.

We need to have a record of every ping on the MEMS, which must have an unique id. This UID linked relational system log must capture each time instance of activity involving the MEMS. Using traditional systems, each time there is a state change, it may be logged in a hash table (hash map). Using this hash function, the series of state changes may be recorded in an array of buckets or slots (think tinyDB on the MEMS SoC). If the raw data from the MEMS is transported, through a gateway to a node in the automobile, then a central CPU may host the hash table for UID linked to the MEMS (if the MEMS SoC had a CPU/IC, then we could be looking at an ASIC).

For micro-payment purposes, event granularity \(<\text{UID} + \text{state change}\>\) tells us how many times the MEMS service was requested and delivered. If the transaction cost of the MEMS is billed as a pay-per-use service then this data is the key to the billing process. This data needs to be secured but using blockchain could be akin to killing a fly with a sledge hammer. The log must be secured and relevant only when the cost is challenged or if the security was breached with the intent to tamper with the log (increase or decrease events).

The output we need from this is just a simple number – the MEMS was used 45 times from 1104 hours (car ignition time stamp) through 1713 hours (last time stamp).

This number (45) is the number of events for which the end-user or customer or OEM has to pay, for operations on a given day or 24 hour period or any other bucket. This number is transmitted to the bank or credit card registered to the automobile. Hereafter, it is fintech as usual, which transfers the charged amount (let us say 45 events at 1 cent per event for a total of 45 cents). Cybersecurity in this financial operation is quintessential.

The innovation and challenge lies in identifying each instance, recording it in an irreversible digital ledger, securing the granularity of each state change, preferably using an unique ID system, in a hash like function, which can id each state change, the time stamp and its point of origin. The latter is key if the data must be audited or if the command origination is in question, if cyber-threat from an intruder is suspected. The concept of creating a cascade of “locks” is an idea, which has been proposed but yet to be explored.

The format of identification of each sequence, and logging each unit activity for payment, may find something important in the conceptual implication of \(3.4 \times 10^{38}\) unique addresses, generated using the hexadecimal routing structure of IPv6.
Convergence of IP routing format, with business models, standards, nomenclature, semantics and financial technologies, such as, blockchain, is likely to evolve. It provides ample room for innovation and leadership for adoption. The monetization tools, capable of converting the granularity of service delivery to currency, are missing. Batch processing may be the closest approximation and that leaves out the pay-per-use billions. For diffusion and broad spectrum adoption of IoT type services, each unit shall seek payment. Could we build such tools and proceed to nurture their progress to catalyze global standards?

These standards will enable widespread sensing, especially remote sensing, with the intent to detect and perhaps to prevent. Billions of users may be willing to use services if it is a pay-per-service model, which makes capital investment negligible, yet offers them data or information, when they want (pull), compared to fixed price subscription as a service (push). Consider the scope of new business if we can replace fixed insurance rates with pay-per-mile automobile insurance. To transform this idea into practice, we need several points of data to validate the model in terms of duration of drive and associated risks.

In other fields, remote sensing in agriculture, and environment, may be critical for food security and public health. Remote sensing at home may decrease emergency room visits and re-hospitalizations. Sensor data is indeed critical but without pay-per-use, the adoption may be sluggish, at best. Transparency, visibility and security are necessary but there aren’t any complete solutions because research and systems integration are lacking.

The example of radio frequency identification (RFID) showed us that when the cost of the RFID tags decreased, adoption increased to billions. The role of the MIT Auto ID Center, where, incidentally, the term IoT was introduced, in 1999, did not re-invent RFID. They simply took the data on the tag and uploaded it to a database accessible via the internet. The RFID tag offered a short (64 bit) alphanumeric identity, which MIT advocated through the academic-industry-government consortia. MIT helped in the standardization of the EPC or electronic product code, along with established standard bodies and industry associations in the US, EU and Japan, in addition to other countries.

Resurgence of RFID is not a perfect example and we need to find new roads. But, the idea of lowering the entry barrier, to gain market penetration, is not a new idea and an idea which is unlikely to change in the next 10 years or in the next 100 years. Cheaper, faster, quicker may be an anathema to academics but the swan song for those on Wall Street who may wish to use money as their preferred choice for wallpaper.
Billions of people who are likely to use these services are unable to invest in capital costs. Hence, the appeal for the “Miss Moneypenny” strategy and the growth of social business. Academics and institutions could play a much more involved role, to usher in a semblance of equity. In 2016, only 8 men owned as much as half of the world’s population.

Half of the world’s population is also suffering from energy poverty. It is the commodity that can spell success, simply by its presence, or disaster, by its absence. Energy will re-invent the automobile industry and Tesla is one example. At UF, I would like experts to engage in the automobile industry through energy and then use the knowledge gain from such energy advances to help the billions in the world gain access to energy.

From an energy R&D perspective, we need to explore nano-composites for high density energy storage, nano-materials to convert carbon dioxide to liquid fuel and nano-materials to re-shape automobile parts and body. I wish to predict the rise of the fuel cell vehicle (FCV) and its use of hydrogen as fuel. Scientists and engineers at UF may wish to think about the next automobile evolution as an energy revolution spreading from EV to FCV. I don’t see why UF cannot contemplate manufacturing FCV’s as a global collaboration.

In parallel, research on nano-crystalline photo-voltaic cells for solar energy, high density graphene batteries for storage, and nano-membrane applications (water-less toilets) are bursting at the seams with R&D potential as well as entrepreneurial potential. The UF credo, to make lives better through science and society, can take a giant leap with solar energy for the billions in the world without access to energy. It may offer us global credibility, visibility and a stream of development activities. In a previous year, I had focused on solar energy and connecting the activity with education for girls and women.

Thus far, I have presented to you R&D opportunities and suggested the need to translate some of our activities for long term investment in the future of UF. How do we start to practice what I am preaching? You have started it already. Schools, institutes and departments under the UF are engaged in multiple industry-academic partnerships. Our faculty are creating Centers, with small and large funding, from various agencies.

I think the practice of creating academic-industry partnerships may be enhanced by increasing the sphere of our creativity, expanding our network and finding innovative ways to embrace not only US companies, but corporations in Europe, South America, Asia and Africa. Our alumni and faculty and staff are from all over the world and they can help UF.
The research output of our faculty has great depth. The institutional reward system prefers those who are plumbing deeper and deeper into fewer and fewer domains. As a scientist and engineer, and as an academic, this is music to my ears. Respect for depth is second to none. As a part of a team to run, finance and sustain institutional advancement, I want the depth in our faculty to reach new heights, gain unparalleled excellence and enjoy global credibility. But, in addition, I want a few people at UF to help us practice the art of synthesis, convergence and connecting the dots. These people, the practitioners, must work with the depth, bring value to the depth, improve the academic environment to sustain those who are deeply embedded, and collectively, help UF reach for the luminous summit.

Practitioners may be generalists, people with diverse experience and breadth of expertise, who will serve as UF’s instrument of goodwill, to guide, to develop, to better utilize, to bring together, to create bridges and to cross-pollinate the wealth of depth, throughout UF and the external world. Through the work of these practitioners, we need to accelerate our proactive outreach to industry, to government agencies, to foundations, to foreign governments and external organizations, for example, standardization bodies. UF must convey and communicate the potential of our faculty, in order to co-opt these external groups to choose UF as their preferred partner. Practitioners must spread the word about the wealth of depth at UF, uncover areas of synergy, strive to find common grounds and then catalyze engagements between funding groups and faculty. Then, the magic will ensue.

This must be repeated, in principle, over and over again, with a variety of faculty, sub-disciplines, multiple domains and diverse groups. However, in each cycle the principle may remain constant but the practice may depart for any formulaic approach. It is advised that we do not create a formula but entertain new measures, avidly explore unchartered territories and build new roads for development, funding and ethical globalization.

In the global arena, US education is one of our strongest national exports. The private institutions are all over the world, with global campus and collaborations and research partnerships. Science and engineering executive education is in very high demand. The agile business models of the “new” world may not mimic the stoic and traditional US models, or find any interest in case studies of US conglomerates. But, the R&D advances from the US, are still the bread and butter, with jam, for global businesses. This is a new area for academics with depth. Few engineers and scientists, steeped in the pursuit of excellence, can conceive that they can teach and share advances from their fields with business people, through courses, workshops, digital education and online learning.
At the undergraduate and Master’s level, UF’s role in global education and digital learning needs to be re-defined, re-evaluated and re-structured, repeatedly. If we cannot adapt, we will die. Learning from MicroMaster’s and nano-degrees, I think we may wish to explore the Bachelorette, the online mini-Bachelor’s degree. We can identify the global supply chain of talent from the exceptional top 1% of Bachelorettes and invite them to UF.

In addition, the depth of our faculty could boost our online resources to explore partnerships abroad, research collaborations, potential joint degree programs and use of our online excellence to specifically empower women’s education. The education of a boy will change the fate of a man. The education of a girl may change the destiny of a nation. According to the World Bank, the global economy may experience a surge in the range of $12 trillion to $28 trillion if we had a modest level of gender parity and more women were allowed to join the workforce, especially, in India.

The ideas and suggestions I have presented are well within the reach of UF. We will attempt to create collaborations, to invest in our students and faculty and staff, to stimulate trans-disciplinary research and multi-pronged associations, to work towards our credo to improve quality of life, for all. In our pursuit, we must not get into a chest-thumping match about UF. We want to celebrate success, encourage failures but never to give up and inspire the young to aspire. But, we must remain cognizant of the problems that we didn’t address, we couldn’t address, problems we may not be able to solve. Hence, I also wish to suggest to keep your inner eye open, indulge in modesty and abide by another credo, *esse non videri.*

Before I conclude this commencement address, I wish to go on the record, and mention a few things that we, UF and the world, may wish to re-think. Perhaps one or more leaders will emerge, who will have the audacity and ability, to take on these ideas and move forward. The grand challenges, I have mentioned thus far, may pale by comparison.

Sahara presents us with unlimited but immobile solar energy. We must add mobility to energy. We must move beyond the metaphor, atoms to bits. We must embrace the paradox of swappable groups of atoms. We must design a new, inevitable, paradigm shift.

Based on transaction cost economics, is it feasible to capture solar energy, store it in a near non-perishable state, in high density modular (nanomaterial or graphene) credit card size batteries and distribute the batteries, worldwide? Slide in the credit card form, or even smaller USB sized, battery in a USB-like slot in any equipment, machine, automobile, smartphone, washer-dryer, tractor or laptop or cooking stove for grid-free mobile power.

---

Rosalyn Sussman Yalow & Norman Ernest Borlaug Memorial Commencement Address by J. Fox Kenneth, President of the University of the Future (UF), on November 3, 2018
EV’s carry 200kg – 600kg (400–1200lbs) of weight in batteries. This is absurd and wasteful and ripe for innovation. The battery in an automobile must a modular form factor similar to a credit card or USB flash drive or at the maximum, the size of a smartphone. For a rickshaw-taxi, in China or India, or a taxi cab in US or EU, the ‘smartphone-size’ battery may provide sufficient energy to travel 20-30 miles or about 10 trips. An average taxi ride in New York City is 2.3 miles. Why carry a 200kg battery for a 2.3 mile taxi ride? After ten rides, drop off the “empty” “smartphone” battery in any local convenience store and swap it for a fully charged battery.

Hence, transforming the paradox of swappable atoms to create a new paradigm shift. In the next decade, fuel cell batteries, may assume form factors similar to a smartphone. If you are a long distance lorry driver, carry a six-pack of batteries and that will get you at least a couple hundred miles, may be a lot more. It is a global ecosystem that must get involved in this paradigm shift. It is a complex challenge and I hope UF shall lead.

The wealth of the Sahara, in terms of insolation, is offset by its poverty of water and food. Sahara is both an asset and a liability, in advancing the world’s next frontier, the great continent of Africa. Malnutrition, disease and social disequilibrium are oozing out of every corner and crevice. The vast natural resources of the land may be better utilized if we had a bridge and/or tunnel between Puerto de Tarifa, Spain and Eurogate Tangier, Morocco. The shortest physical distance between Europe and Africa is less than 10 miles (about 14 km).

To partially address these issues, I think it is time to re-discover and invoke an idea suggested in the 19th century. Inspired by the successful completion of the 102 mile long Suez canal, in 1869, a few years later, in 1874, Captain Francois-Elie Roudaire, an army geographer from France, suggested the creation of a 120-mile-long canal to connect the Mediterranean Sea to Chott Melrhir, an endorheic salt lake in northeast Algeria, covering an area of 3,000 square miles (~7000 km²), which lies 40 meters (130 feet) below the sea level. It is usually longer than 80 miles from east to west.

If flooded, such a huge body of water in the interior of North Africa and the Sahara Desert, may present opportunities for agriculture, improve sanitation and may change the climate. Can a few degrees drop in the temperature of the Sahara reduce the number of Cape Verde hurricanes? In 2016, we encountered Hurricane Irma, a Cape Verde hurricane, the most powerful Atlantic hurricane, yet recorded, outside of the Caribbean Sea or Gulf of Mexico. The prevention of billions of dollars in damage, due to these Cape Verde hurricanes is sufficient financial return, on the investment necessary to flood the Sahara.
If flooded, can it help to prevent, at least a few, of the 11 million children who die each year, mostly in Sub-Saharan Africa, due to preventable causes? Improvement in sanitation and sewer systems can reduce the vector borne diseases and spread of infection. About a billion people in the world, today, have no access to a toilet and must defecate in the open. About half of a billion are in Africa, and 569 million people in India, defecate in the open. Bill Gates’ water-less nano-membrane toilets are innovative, expensive and may not be a long term solution. If flooded, it may help half a billion people in Africa to find a way out of defecating in the open.

If flooded, the water from the Mediterranean Sea may pass through giant nano-membrane desalination sieves, attached to water gates, to reduce the natural salinity from about 3.5% to 0.5% in the filtered water? The volume of fresh water in 3,000 square mile area, cannot be anything but a lifesaver, if used and distributed, appropriately.

The financial risks aside, the debate of the future flooding of the Sahara, and other similar projects, must weigh human benefits in light of potential ecological disturbances and environmental changes. With the world population reaching 10 billion around 2050, production of food must assume “Borlaug-ian” significance. Is Sahara a part of the solution?

Sahara’s abundance of untapped energy and lack of running tap water, are triggers for world changing events. Re-shaping the geography of the Sahara may improve quality of life, for billions. Sahara is a story of rags to riches. Can UF become a co-author of the new book “Peddling The Flood of Prosperity – The Sahara Solution” published by UF Press?

We don’t get what we wish for, we may get what we work for. I urge UF students, to pay heed to the words of Mahatma Gandhi, “be the change you wish to see in this world.”

May God bless you.

Please remain seated as the students exit the auditorium. The UF Marching Band will play the Commencement theme.

“When I’m Sixty-Four”

Thank you.
This is the true joy in life, the being used for a purpose you consider a mighty one, the being a force of nature rather than a feverish, selfish clod of ailments and grievances complaining that the world will not devote itself to making you happy. (GBS)

The best people possess a feeling for beauty, the courage to take risks, the discipline to tell the truth, the capacity for sacrifice. Ironically, their virtues make them vulnerable; they are often wounded, sometimes destroyed. (EH)
COMMENTARY [C]
PROPOSAL

To create the World Sensor Organization, WSO, a global collaborative entity where scientists and engineers may contribute their findings, and methods, for sensors. WSO will try to transform proof of concepts to field deployable sensor **systems**. Successful sensor **systems** may be converted to market ready phase and implemented for designated use, if feasible.

VISION

Thousands of publications, perhaps millions, are generated each year on topics related to sensors. Scientists and engineers are producing laboratory versions of sensors for humans and healthcare, veterinary medicine, environmental and industrial purposes. From this wealth of knowledge, (see [http://bit.ly/PLASMONICS](http://bit.ly/PLASMONICS)) few sensors make it to the market, if transaction costs spell profit. What happens to the outcome of sensor related research?
It may collect dust or loiter in cyberspace or found hidden on the second page of Google search. Yet, critical issues for detection, which could save lives or improve the environment, remain unsolved. Translation from vision to market reality, is the rate limiting factor, for many sensor related research results.

World Sensor Organization (WSO) aims to partially alleviate this problem through a global consortia (center). WSO will catalyze the convergence of expertise necessary for the sensor to evolve as an integrated sensor system including, but not limited to, signal transduction/capture, data collection software, mobile interface, analytics, visualization, secure access via smartphone apps, delivery of information to the user, cybersecurity and transmission for global connectivity.

RATIONALE

The infectious enthusiasm for entrepreneurial innovation calls for confluence of multiple factors to create the optimum ecosystem. Few locations in the world are able to capture the key factors. Even fewer scientists and engineers possesses the attributes necessary to shepherd their research results to a market deployable product, or service, through the entrepreneurial grind.

Grand imagination often generates great inventions but too many good inventions die without a proper path to innovation. The latter may not succeed without entrepreneurial skills. We are left with an embarrassment of riches, millions of published papers and unpublished ideas, but the fruits of labor are few, and far between, because the continuum is difficult to coalesce. What can we do to drive more ideas to fruition?
Inventors are keen to patent their findings and rush to create start-ups. Stand-alone start-ups ignore the fact that a sensor, no matter how profound in scope, is simply a tool for detection. The ecosystem and local/global connectivity is germane to function. Sensor data/information must trigger action, based on appropriate data analysis after sensor data is available to the system. If the sensor data does not contribute to the decision, its value may perish, rapidly. The integrated system of systems, and transaction costs associated with each event in this cascade, reflects a convergence of technical issues with economics, which must be navigating, in order to arrive at a feasible solution, affordable by the end user and the end-user community.

Solutions, for example, GE Pharmacia BiaCore (old) and Genalyte (new), are marketed as high cost products, suitable for affluent users in wealthy nations (about 10% of the world population). Investors and inventors of sensors, are keen to follow the patent protected profitable path, vying for the 10% of the global population in US, EU and JP (approx. 700 million), in their elusive quest for personal wealth, as a reward for their imagination, invention and innovation.

THE ROAD NOT TAKEN

WSO wishes to “think different” but as an initiative in the US, its germination may face hurdles. World Sensor Organization (WSO) wants to remind the women and men, involved in sensor research, that their work, and outcome, may not have to bite the dust. There is a path to immortality. By contributing their research to WSO and pursuing a collaborative “open source” approach, for translating their sensor ideas, to a deployable sensor system, they may reap the reward of gratification, the reward for enabling science to serve society, the reward for their labor to improve quality of life, the reward for contributing to global public goods.
Few efforts survive in the public domain, without economic incentives. WSO is cognizant of this fact and wishes to introduce the financial reward named “Miss Moneypenny.”

For every sensor event, triggered by every user, WSO will attempt to find a path to extract one penny, in profit, from every transaction. In other words, extracting 1 penny (US one cent) as net profit, by using the “pay-per-use” concept, we monetize “sensor-as-a-service” rather than the sensor-as-a-product, sold as an one-time product. By reducing the barrier to entry into world markets, WSO promotes social business entrepreneurship. WSO can penetrate the global market of 7 billion people. WSO envisions a paradigm shift from the product lifecycle mindset to arrive at the user lifecycle paradigm. The user does not pay for the product but pays for “service” as long as the user lives and uses the service based on this or related products. The product lifecycle may be 2-10 years but the user lifecycle may be 50 years. Hence, micro-payments and micro-revenue for services, provided as micro-payments, during the life of the user, ~50 years.

WSO “sensor-as-a-service” aims to rake in net profit of one penny per use, per person, per day. That translates to 700 million pennies per day, or US$7 million per day in revenue, if only 10% of the global population (700 million) uses a WSO certified sensor, once, each day.

ETHICAL GLOBALIZATION AND ETHICAL PROFITABILITY

In this example, the micro-revenue from ethical profitability, due to economies of scale, from only 10% market penetration, may exceed US$2.5 billion per year.

Development and modularization of product, segmentation of the sequence, work unit synthesis and structuring the alliances to aggregate the end-to-end ecosystem to capture the pay
per use billing granularity, are difficult tasks, but nothing that violates any laws of physics or thermodynamics. Hence, it is possible.

Micro-payment based pay-per-use service model can be applied to almost any product if analyzed and dissected with the tools of transaction cost economics. The billing granularity and evidence of use may be captured in a digital ledger. Unique id of each transaction may use the internet protocol version 6 (IPv6) format using a 128-bit hexadecimal system which is capable of generating $2^{128}$ or 340,282,366,920,938,463,463,374,607,431,768,211,456 or 3.4X10^{38} unique addresses [ref 13].

The technical convergence in combination with the business model paradigm shift has the immense potential to unlock the power of social business entrepreneurship. The benefits due to WSO may reach remote corners of the world. The barrier to entry is probably low enough to penetrate markets of billions where the per capita earning barely approaches US$1.90 per day.

One cent profit may be too much or too little, depending on the value delivered by the sensor data, demographics of the market and economic bandwidth of the end user community.

The micro-payment model is nothing new and originates from micro-loans and micro-investment practices. It is the dominant model for business wealth created by MacDonald’s and PayPal. Grameen Bank of Dhaka, Bangladesh pioneered micro-loans to women, only. In recognition, Dr Yunus, Chairman of the Grameen Bank, was awarded the 2006 Nobel Prize.

Micro-payments may lead to mega-profits and the model is not limited to sensors. It can be applied to 3D printed automobile or aviation industry spare parts, transplanted hip or knee joints, bidirectional energy arbitrage, autonomous vehicle services, energy storage, home
appliances, jet engines, and security-as-a-service for modular, mobile, retail cybersecurity offerings, delivered online or purchased from a corner grocery store or neighborhood Walmart.

For related ideas, please explore “18. Commencement” which is a PDF in the zipped folder ‘CHAPTERS’ available from MIT Library https://dspace.mit.edu/handle/1721.1/111021.

FINANCIAL INCENTIVES

WSO proposes that [a] a quarter of the earnings are distributed to scientists and engineers involved in the sensor system development [b] a quarter of the earnings are re-invested in WSO and its affiliated institutional R&D and [c] a quarter of the earnings are used in projects to spur education and economic growth in the less fortunate nations (our future markets, the emerging economies, for example, sub-Saharan Africa) and [d] a quarter of the earnings are used to strengthen the WSO organization, its foundation or endowment, to ensure uninhibited progress.

WSO is a solution for the vast majority of the global sensor scientists whose work may serve humanity. Aggregating sensor data on open source platforms, with adequate cybersecurity and data analytic tools, will allow sensor information arbitrage in near real-time, at the point of need. What happens when we can measure? We can establish metrics, generate quantitative performance indicators, inform decisions to direct policy, structure appropriations based on information rather than the pork barrel approach practiced by politicians. Policy measures, if pursued ethically, can lead to social re-engineering of communities through distribution of social services, prioritized restructuring of access to healthcare, education and workforce development.
EXTENDED MISSION

WSO is an idea, which, if implemented, may change the world through ubiquitous sensing, at a cost which will grow the global economy, rather than sap its economy by charging $100 for a sensor. It will lift many boats, not just a few yachts. It will empower social business.

WSO will serve as an educational platform, for inspiring undergraduate and graduate students, to train for the steps necessary to transform ideas into reality. The collaboration of global scientists may enable paths to include rigor, innovation and exposure to diverse problems. The potential for successful outcomes through WSO will remind students that completion of their degrees and graduation may lead to that proverbial “light” at the end of the tunnel. The impact of student motivation on graduation rates and research funding may be significant.

WSO may serve secondary education outreach. Almost every discipline, and their convergence, may be exemplified by sensors, sensor systems and sensor networks (similar to the significance of “energy” as a theme for science and engineering). In a manner similar to the Cisco Networking Academy (CNA), which commenced in 1996, as a high school networking lab project (I helped to start CNA at Thurgood Marshall High School in San Francisco, CA), WSO can help to start the Sensor Networking Academy Program (SNAP) in high schools, worldwide.

Hence, WSO is a tool for workforce development through SNAP for secondary and tertiary education. WSO may offer students a variety of opportunities for research in basic sciences, materials science, nanotech, biotech, medical instrumentation, electronics, ASICS, devices, systems engineering, embedded systems, sensors, time sensitive networks, signal transduction, networking, data, analytics, statistics, mathematics, simulation, computer science, programming, AI, telecommunications, cyberphysical systems, standards, cybersecurity and IoT.
BUILDING ORGANIZATIONS

World Sensor Organization is a grand challenge, an idea which will only grow if fertilized with credibility, cooperativity and collaboration. Global community of scientists and engineers, must work with industry, government and non-government organizations (NGOs). Support from the parent institution must be uncompromising and unequivocal.

Building the structure of the infrastructure, necessary to create this organization, in a sustainable manner, will require greed-free leadership. Individuals who are oblivious to not receiving credit, individuals who are dedicated to serve as an instrument, individuals who are committed to science as a tool for social progress, individuals who are not prioritizing personal wealth creation, individuals who believe in creating economic growth, individuals who are capable of excellence in communication across a very diverse spectrum of stakeholders, globally, at multiple levels, including science and engineering, engaging the public, as well as cultural respect, diplomatic finesse and political-context awareness. Building trust is the secret sauce.

As with any consortia, seed funding for organizational nodes may be quintessential for lift-off. The primary step is to build the foundational coalition with appropriate credibility and “hooks” to reflect the vision of the World Sensor Organization, at a global scale. Unlike traditional consortia, WSO may not operate on an “entry fee” mechanism, in order to uphold its claim for public goods. Hence, “joining” the consortia and “member finances” for the WSO, must be separate operations, without dependencies. This will make our task hard, financially. The leadership must remain cognizant of the steep path which we must scale for future success.
WSO members are expected to be individual engineers/scientists/students, institutions, NGOs, organizations, corporations, government agencies, governments, international standards bodies and associations.

WSO may be financed by gifts, seed funding, endowments, grants, foundations, and these categories (listed to show increasing amounts) “friends” ($1K+), “partners” ($10K+), “contributors” ($100K+), “patrons” ($1M+), “sponsors” ($10M+) and “donors” ($100M+).

Working groups (WG) and special interest groups (SIG) will be created to modularize and distribute the operations, starting with sourcing (when an idea is contributed) and extending to systems oversight.

Liaison with standards bodies may be required to ensure that the work and outcome from WSO is in compliance, or creates new standards, which must be interoperable with other existing standards, or possesses the requisite characteristics for pursuing global adoption.

WSO use cases for sensor research, development and applications must offer a broad portfolio, to attract membership from diverse industries and government agencies, for example, [a] machine, tools, 3D manufacturing [b] oil, gas, water, energy, desalination, natural resources [c] environment, communities, smart cities, sewer systems, emergency management systems [d] healthcare and medical IoT (https://dspace.mit.edu/handle/1721.1/107893) [e] Digital Twins in manufacturing systems, maintenance, transportation, aviation, (please see “04.Digital Twins” PDF in the “CHAPTERS” folder accessible from https://dspace.mit.edu/handle/1721.1/111021) and [f] agriculture including animal farming (micro-robots in cows to monitor metabolomics of lactation).
WSO may evolve in ways we may not anticipate due to the plethora of unknown unknowns. But, WSO must be capable of generating value from research, education, students and applications. The primary goal of the coalition is to advance research outcomes and promote education which can create new leaders, endowed equally with scholarship and leadership.

RISK

The central thesis of WSO is to “apply” the research outcome from scientists, and shape it through further R&D, to serve a purpose, which generates value through sensing. By connecting the sensor data with the sensor ecosystem, we make sense of the data and extract actionable information. The information may improve a function or performance, in near real-time. By connecting this data and/or information to real world situations or by connecting a “swarm” of sensors, we may optimize processes, workflows, automation or improve accuracy, and/or precision, of diagnostic, predictive or prescriptive analytics. Taken together, using control feedback loops or decision support, we deliver dynamic optimization, due to the sensor.

The central problem in this proposed concept of WSO is the assumption of the existence of trust and altruism, not only in sourcing the sensor idea or scheme, but also in working with the partners in the ecosystem, which is salient to the delivery of value from sensing to the end user.

[a] Why will an engineer or scientist trust WSO to contribute her/his research? [b] Why will the partners of the ecosystem (for example, manufacturing execution systems or MES) allow WSO to work with its systems? [c] Can we identify and define the transaction costs associated with each stage? [d] What degree of systems interoperability do we need, to uniquely track and trace every change of state, in order to bill the client for the service used, according to pay-per-use?
Question [a] is crucial. If WSO cannot be trusted serve as the global sensor research repository, then the central thesis is null and void. Explicit institutional support and the ability to bring together a team of pioneering global leaders, as co-founders, may be the first step to earn trust. For question [b], gaining agreement to create a few real world test beds with industry and corporations, may elevate the trust level. The latter may help to engage with governments and agencies, at home and abroad. Can we get a critical mass of pioneers to support the WSO vision?