The justification of Napoleon's statement – if indeed he ever actually made it – that those who form a picture of everything are unfit to command, is to be found as the first defect. A commander who approaches a battle with a picture before him of how such and such a fight went on such and such occasion, will find, two minutes after the forces have joined, that something has gone awry. Then his picture is destroyed. He has nothing in reserve except another individual picture and this too shall not serve him for long. Or it may be that when his first forecast is found to be inapplicable, he has so multifarious and pressing collection of pictures that, equally, he is at a loss what practical adjustments to make. Too great individuality of past reference may be very nearly as embarrassing as no individuality of past reference. To serve adequately the demands of a constantly changing environment, we have not only to pick items out of their general setting, but we must know what parts of them may flow and alter without disturbing their general significance and functions.

F. C. BARTLETT

Adapting Decisions, Optimizing Facts and Predicting Figures

Can Confluence of Concepts, Tools, Technologies and Standards Catalyze Innovation ?

by

Shoumen Datta
Massachusetts Institute of Technology

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 US Senate Committee on Banking, Housing and Urban Affairs (13 February 2001)

Working Paper (First Draft AUGUST 2003)

Adapting Decisions, Optimizing Facts and Predicting Figures

Can Confluence of Concepts, Tools, Technologies and Standards Catalyze Innovation ?

by

Shoumen Datta

Research Scientist, Engineering Systems Division, School of Engineering and Executive Director, Forum for Supply Chain Innovation

Massachusetts Institute of Technology

Contributors:

Benson Adams
Executive Deputy Commanding General, Office of the Commanding General, US Army Materiel Command, US Department of Defense

Mohua Barari
Professor of Economics, Missouri State University

Bob Betts
President and Founder, Mainstreet Applications

Mark Dinning
Supply Chain Strategy Group, Dell Corporation

Tom Gibbs
Director, Intel Corporation

Hui Li
Graduate Student, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology

Mike Li
Research Scientist, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology

Yichen Lin
Director, Institute for Technology Management, National Tainan University, Taiwan

Greg Parlier
Office of Economic Development, University of Alabama and Scientist, Institute for Defense Analysis

Micah Samuels
Senior Operations Manager, Amazon.com

Santtu Toivonen
Graduate Student, VTT Information Technology, Finland

Disclaimer

This article is over-simplified, incomplete and plagued with digressions. All errors of content or coherence are due to the author. The author apologizes for an unsatisfactory reading experience but hopes that the amalgam of ideas may spark new thinking. This is a mere exploration. In addition to named contributors, the author has freely used several sources of information to 'connect the dots' and show how distant disciplines, if coalesced, may offer new directions. The list of references is seriously incomplete. It may be amply clear that the original research is not due to the author. Opinions and comments expressed here are attributable to the author and do not represent the views of MIT as an institution or the contributors or their organizations. For experts, there may be nothing 'new' in this article. But, it is the synthesis of ideas from a variety of sources, when presented in confluence, as suggested by the author, may be catalytic in the transformation of some types of decision support systems to adapt or perhaps, with time, to predict. Please e-mail suggestions to Dr. Shoumen Datta, Research Scientist, Engineering Systems Division, School of Engineering and Executive Director, MIT Forum for Supply Chain Innovation, Room 1-179, MIT, Cambridge, MA 02139 (Phone 1.617.452.3211) shoumen@mit.edu

CONTENTS

CONTENTS (continued)

CENTRAL THESIS (Executive Summary)

This conceptual article is aimed to provoke a broad spectrum of decision makers who wish to make even better decisions based on deeper insight from process innovation as well as right-time analysis of real-time data. It is not a panacea to rid of all poor decision steps nor can it function without appropriate and in some cases, adequate, help from the 'enablers' that we shall discuss. Managing uncertainty is key in decision systems, such as supply chain management or military readiness. We propose a reasonable confluence of existing concepts, tools, technologies and standards that may, collectively, improve adaptability of decision systems to combat uncertainty in such diverse applications as profit optimization, response time in hospitals or military readiness. Improvements must be directed to reduce noise and optimize to adapt. This proposal is illustrated in Figure 1.

While thinking about the variety of suggestions in this article, readers are encouraged to consider and evaluate these suggestions in view of their organization from the perspective of [1] efficiency, [2] time compression and [3] transaction cost economics or TCE proposed in 1932 by Ronald Coase (Coase received the 1991 Nobel Prize in Economics for his concept of TCE). Reduction of transaction costs may be the most important ***value*** from real-time data, if used at the right-time to execute the right decision.

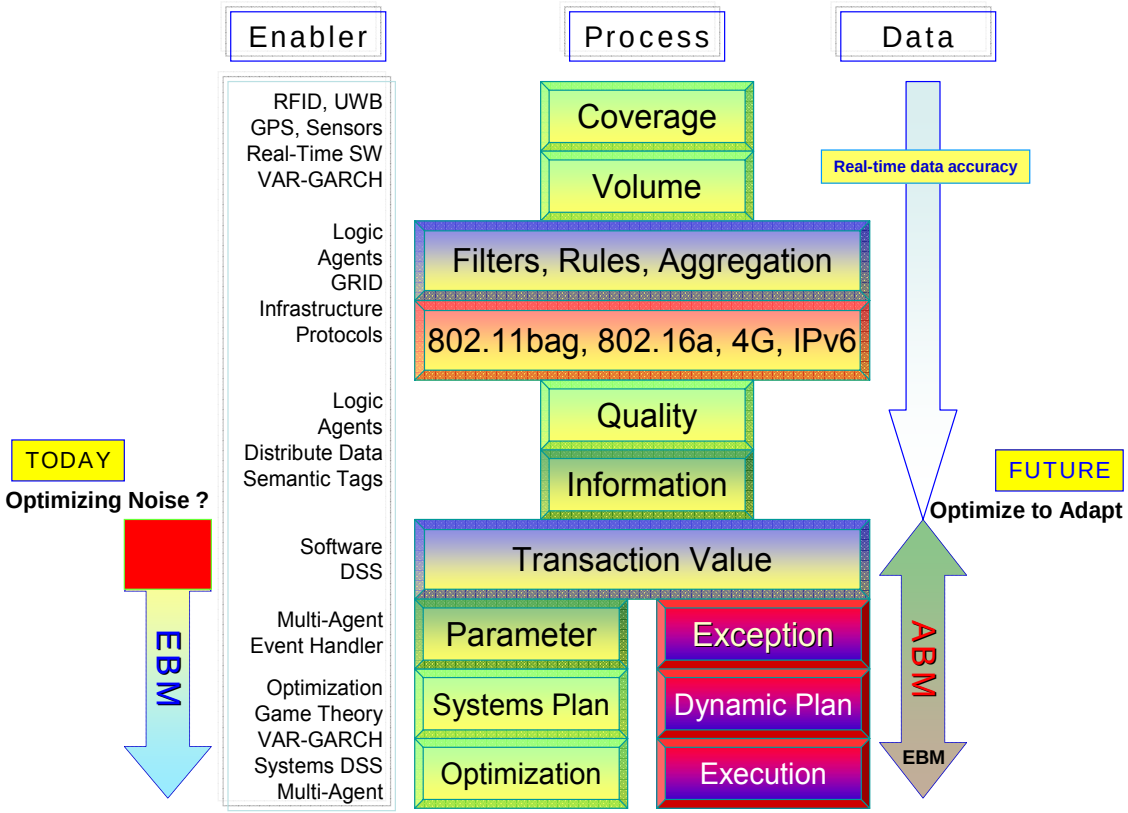


Figure 1: A proposal to migrate from optimizing 'noise' to better adaptability in the future.

In the context of decision systems, we will often refer to supply chain management as an example and discuss how current supply chain practices may change for the better if practitioners (decision makers) adopt and ***effectively*** use new thinking, analytical tools, technologies and emerging standards to reduce uncertainty, hence, reduce transaction costs. The ***use*** of real-time data is crucial for industries (retail, healthcare) and military, yet the past few years (1999-2005) have witnessed a disproportionate focus on data acquisition tools, such as, automatic identification technologies (AIT) aggressively represented by radio frequency identification (RFID). In our opinion, the impact of real-time data on transaction cost economics (the operational process) may be the key parameter for businesses aiming to reduce volatility and/or uncertainty. Use of AIT to identify objects with RFID (UWB) is beneficial when data (systemic, local) is used at the ***right-time*** with respect to the operational process and if such processes, then, yield, ***decisionable information*** to shape decisions, rapidly, to adapt, if necessary, and to respond through action or preparation. This connectivity that follows from the proposal outlined in Figure 1 is illustrated in Figure 2.

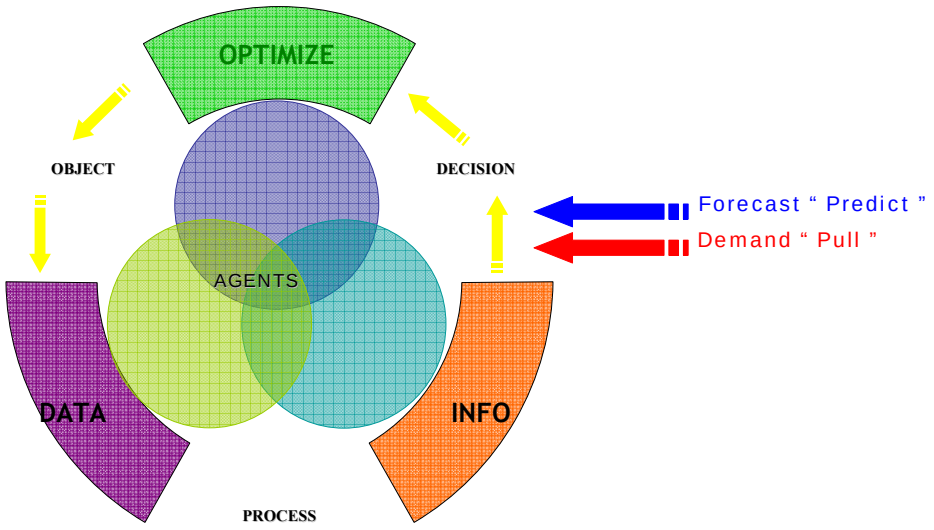


Figure 2: Connectivity of real-time data to process (real-time analytics) may improve decisionable information.

Real-time data at the right time (***right-time data***) may change operational processes and stimulate process innovation. However, real-time data feeds to legacy systems or ERP may not be productive. Adaptability may be enhanced if decision systems can access information at the right time based on real-time data (real-time analytics) which may be acquired from diverse sources (RFID, UWB, sensors, GPS, barcodes). The argument over format (electronic produce code or EPC, universal identifier code or UID) may continue but that should not inhibit the thinking pre-requisite for process innovation to make adequate use of right-time data.

Format agnostic architectures (software) may hold the key to connect real-time data to other software or ERP. In our view, enterprise resource planning (ERP) software packages are unlikely either to handle or analyse the exabytes of real-time data. We echo earlier proposals to explore the use of Agent-based software (ABM or Agent-based models) in ***combination*** with traditional equation-based software (EBM or equation-based models) to extract information at the right-time from the emerging abundance of real-time data. Semantic connectivity of this data is as important as sharing of information ***between entities***, to improve decision making.

Sharing data (implied by concentric rings, used in a Forrester illustration) or information to improve decisions, then, improves the performance of the entire value network. Interaction between entities demand secure infrastructure and a pervasive open platform for collaboration. We propose such a platform where data interrogators are software defined radio (SDR-SWR) (see note L on page 100) that is ubiquitous, transponder agnostic and part of the civil infrastructure. Access and control of data sharing is regulated and authenticated via the software application layer (delivered over the internet) as would be for an internet appliance (turning on the microwave while driving toward home). The use of semantic software as infrastructure is the next step.

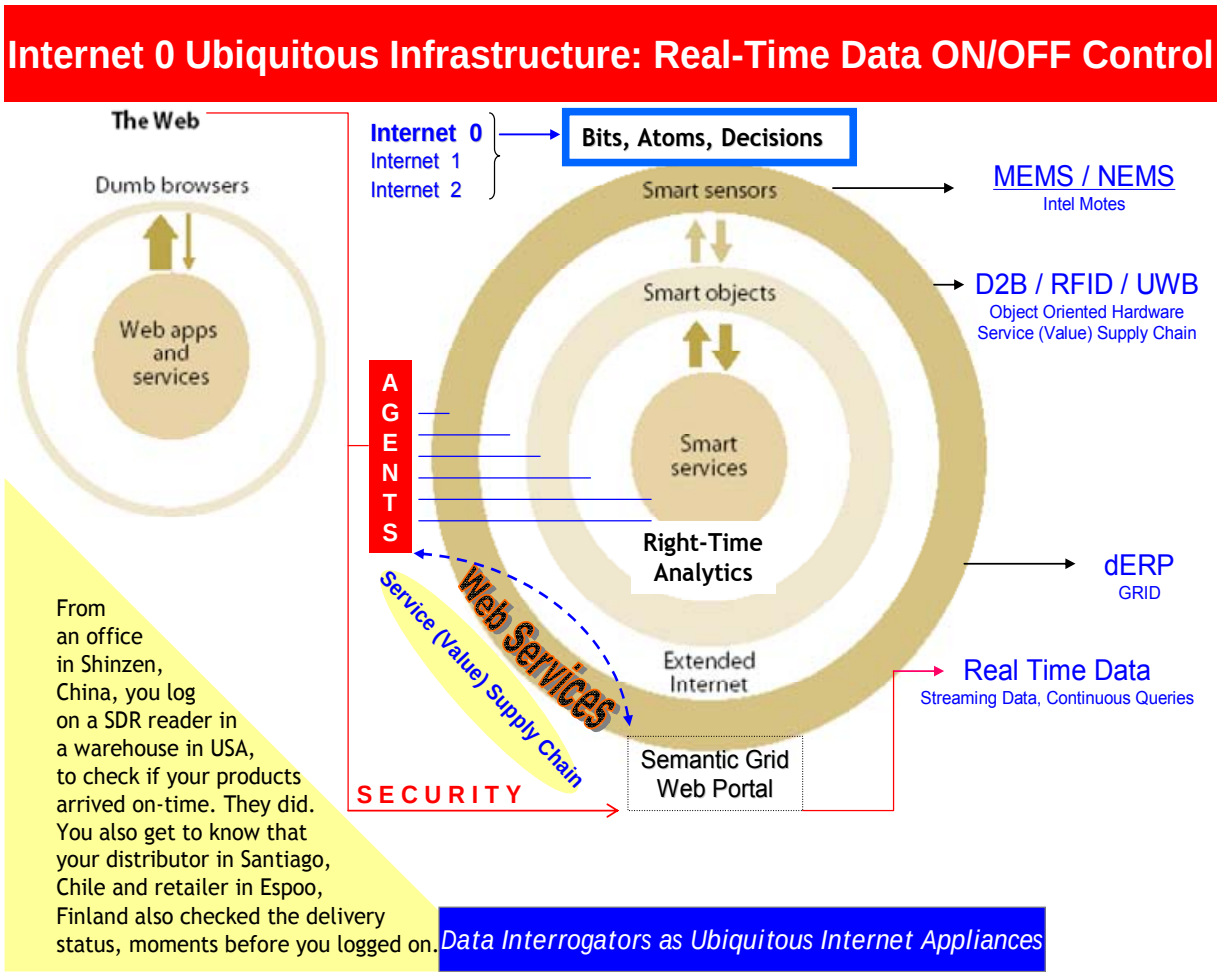


Figure 3: Connecting Bits, Atoms and Decisions is necessary for the emergence of adaptive decision systems.

In the next 5 years or, more likely, in the next 25-50 years, when we may migrate from adaptive to ***predictive*** status of operations, we will require other concepts and tools that may be unknown, today. However, as a contribution to the predictive phase of operations, we propose an idea that draws from the field of time series econometrics. In its simplest form, the proposal explores the possible use of 'raw' real-time data (without clustering or classification) to better understand and respond to changes, in near real-time. It may help further reduce risk and uncertainty, perhaps, even, may tame the Bullwhip Effect. Such econometric analytical tools are used in finance (stock price volatility). If econometric tools (such as, GARCH or generalized autoregressive conditional heteroskedasticity, first proposed by Robert Engle) can be modified for use with real-time object-dependent data (ODD), it will not only help predict key supply chain parameters (demand forecast, price) based on input (real-time data) but can also provide a measure of ***risk***, associated with the prediction.

In this article, therefore, we will try to coalesce different ideas from a variety of sources to offer a 'solutions' approach aimed to reduce uncertainty and improve decisions. This article weaves in ideas from Game Theory, automatic identification technologies (AIT), time series econometrics, Grid computing, Agents, Semantic Web and simulation. It is quite possible that governments, corporations, 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 or effective *modus operandi* to combine these ideas with other emerging concepts, tools, technologies and standards to collectively better understand, analyse, reduce and respond to uncertainty (see Figure 4). Understanding confluence will help explore the paradigm between adaptability and efficiency. Management framework (tools dashboard) to diagnose and determine the dynamic equilibrium (industry specific) may optimize the 'push-pull' between adaptability and efficiency (see 'concluding comments' on page 92-93).

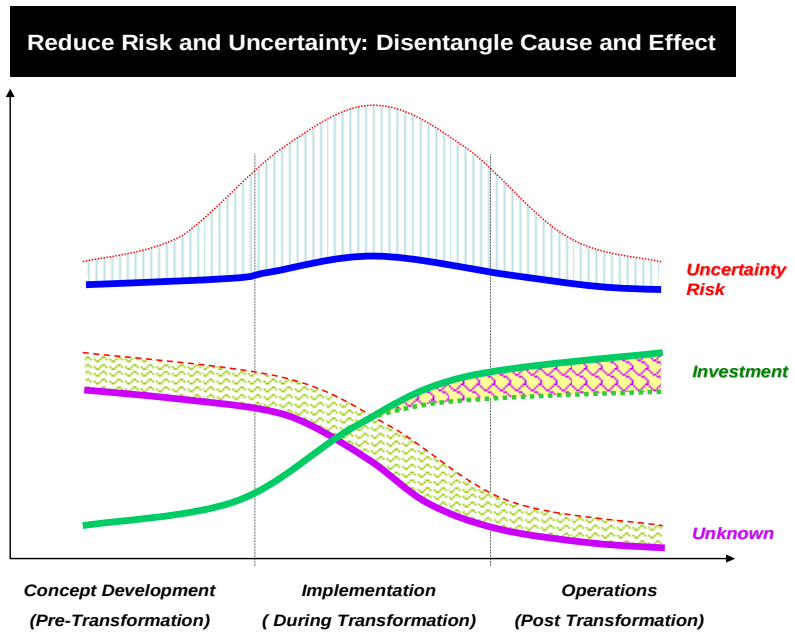


Figure 4: Confluence of ideas to reduce risk and uncertainty as a step toward Gibbs Equilibrium (see page 92)

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." (1)

The holy grail of industry is to remove 'chaos' from the supply chain to better adapt to demand fluctuations. Managing uncertainty is compounded by the increasing degree of information asymmetry (see note a, page 95) between the supply "chain" or value network (see note b, page 95) 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 (2).

	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, JVC, issued directives for their respective engineers to produce units 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 (3). Adapt or die!

One business objective of suppliers is to secure large volume purchase commitments (with delivery flexibility) from manufacturers. It conflicts with the manufacturer's objective if rapid response to demand fluctuation leads to excess raw material inventory. The manufacturer must mass produce (to take advantage of economies of scale) yet production runs must adapt to fluctuations even though resource utilization plans were based on demand forecast. Thus, manufacturers may need more or less raw materials and seek flexibility in purchasing raw materials, which conflicts with the supplier's objective. The manufacturer's desire to run long production batches are in conflict with the warehouse and distribution centers that aim to reduce storage capacity. The latter increases cost of transportation for all the players (5).

During 2000, supply chain related costs in USA, alone, exceeded $1 trillion (10% of GDP), which is close to the GDP of Russia, more than the GDP of Canada or the combined GDP of the 22 nations who are members of the League of Arab Nations. The combined GDP of all 22 Arab nations, including the oil opulent nations, is less than that of Spain. A mere 10% savings of supply chain costs in USA is nearly equal to the GDP of Ireland (4).

Therefore, tools and processes that may reduce supply chain inefficiencies are valuable. Ability to adapt may not depend on technology but may depend on continuous business process innovation in supply chain practice if the management is capable of envisioning use of various, concepts, tools and technologies to reduce (a) inefficiencies, (b) uncertainties and (c) information asymmetry within the value network.

One driver of this transformation (from 'push' based supply chain management to 'pull' based adaptive value networks) is the potential use of real-time data and information to trigger autonomous decision steps capable of concurrent re-planning and execution. According to Forrester Research, businesses 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. 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 (ERP) may suffer from problems that reduce the value of their information. ERP systems may also compromise the efficacy of dynamic data if the static systems are unable to respond in near real-time. When such ERP data and/or information sources are used by planners for forecasting or 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, as illustrated in Figure 1. Stemming from poor data quality and information asymmetry between supply chain partners, errors (of optimization, forecasting) accumulate at successive stages of a supply chain and manifests as the Bullwhip Effect (6-9), as illustrated in Figure 5. The Bullwhip Effect based on actual data from the semiconductor industry, is shown in Figure 6.

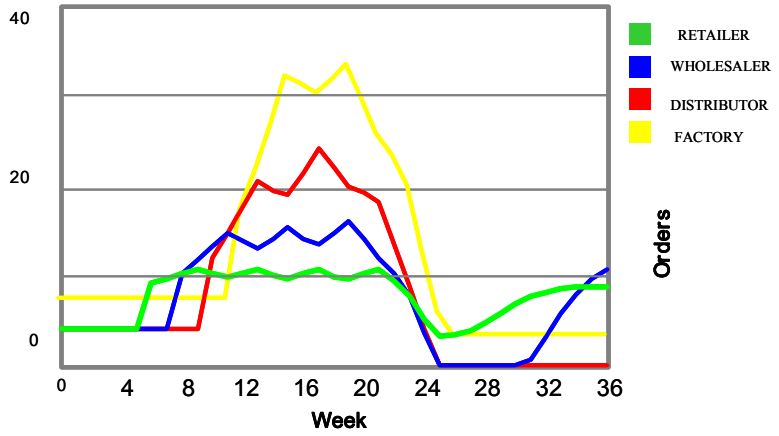


Figure 5: The Bullwhip Effect (10)

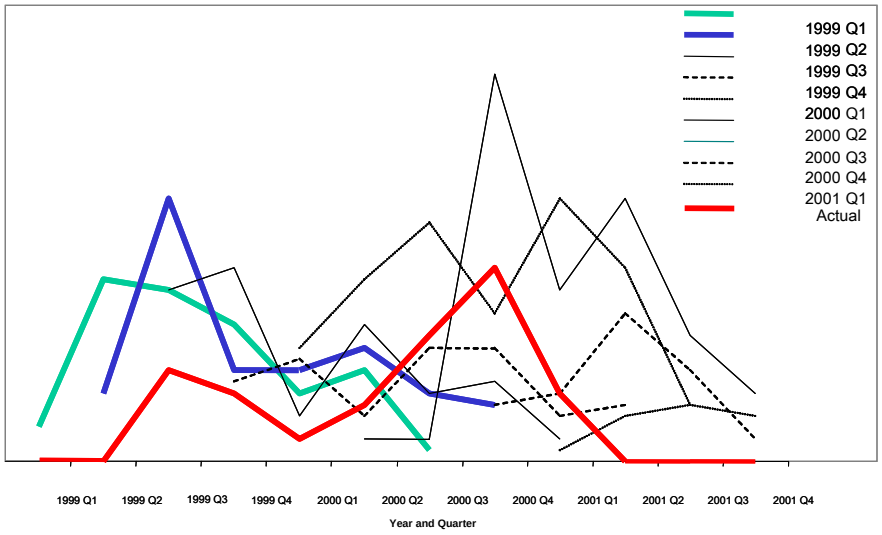


Figure 6: The Bullwhip Effect in the semiconductor equipment supply chain shows demand forecast versus actual purchase of equipment (11).

TOWARD ADAPTIVE VALUE NETWORKS ?

Tools and technologies that may be catalytic in taming the Bullwhip Effect (6-9) may also be a driver for supply chains to be more adaptive. The latter assumes that organizations will promote business process innovation aimed at improving interaction between entities (intra- and inter-enterprise information exchange) and target extinction of data silos by investing in semantic connectivity. Organizational 'culture' and change management are necessary to stimulate new thinking with respect to confluence of concepts, tools, technologies and standards. Some organizations may appreciate the vision of how to connect objects (atoms) with data (bits) to processes or real-time analytics to extract real-time information for adaptive decisions, that may, in turn, optimize the nature of objects (design, distribution) or characteristics of objects (price, risk) in a value chain.

Tools and Concepts	Data Sources
Operations Research	Automatic Identification Technologies
Game Theory	(RFID, UWB, GPS, RTLS)
Agents (distributed artificial intelligence)	Identification Technologies
Econometric Tools (ODD-VAR-GARCH)	(GPRS, Voice, Manual, 2d-Barcode, Barcode)
Semantic Web	Wireless Protocols (802.11, 802.16)
Grid Computing	Sensor Networks (802.15.4 / ZigBee)
Tangible User Interfaces	Software Defined Radio (SDR-SWR)

Table1: Elements of the Proposed Confluence of Concepts, Tools, Technologies and Standards

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 (see note c, pages 95-96). Game Theory (GT) was not a "household" name until 1994 when John Nash, and later the movie about him, changed the public perception so much so that generic business journals began touting the virtues of GT (25). For centuries economists have worked on various game-theoretic models but John von Neumann and Oskar Morgenstern (1944) are credited as the fathers of modern Game Theory (12). GT has since enjoyed an explosion of developments, including the concept of equilibrium, games with imperfect information, cooperative games and auctions (13-18, 25).

Game theory helps us model, analyze and understand the behavior of multiple self-interested parties who interact for decision making. As such, Game Theory deals with interactive optimization problems. In particular, it is a tool to analyze 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 interaction, involving two or more parties, the pay-off of each party depends on the other party's actions. Thus, 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. We ***assume*** that businesses make rational choices to optimize its profits. Do they?

	A chip manufacturer slashed prices of its desktop and mobile processors days after a similar move by a rival. We're going to do what it takes to stay competitive on prices, said a representative. The company's 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, May 30, 2002)

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 (see note d, page 96).

Prisoner's Dilemma

Alice and Bob are arrested near the scene of a burglary and interrogated separately (19). Each suspect can either confess with a hope of a lighter sentence or 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. 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.

		Alice	Alice
		Confess	Does not
Bob	Confess	10, 10	0, 20
Bob	Does not	20, 0	1, 1

Table 2: Prisoner's Dilemma: Alice (column) versus Bob (row).

The numbers in each cell show the outcomes for the prisoners when the corresponding pair of strategies are chosen. The number to the left is the payoff to the person who chooses the rows (Bob) while the number to the right is 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 only 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 only a year (19).

Prisoner's Dilemma is a simple example of a non-cooperative static game where the players choose strategies simultaneously and are thereafter committed to their chosen strategies (25). 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. A key concept not captured in "Prisoner's Dilemma" is the repetition of interactions. In business, players know they will be in the 'game' for a while. 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, Intel uses its supplier ranking and rating program, which tracks a supplier's 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.

In the real world, each party in a supply chain acts entirely on self interest. Thus, individual choices collectively do not lead to an "optimal" outcome for the supply chain. Supply chain profit of a "decentralized" supply chain composed of multiple, independently managed companies, is usually 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 (reduced unknowns) to optimize total supply chain profit. 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. Thus, optimal profit is higher in centralized supply chains with information sharing.

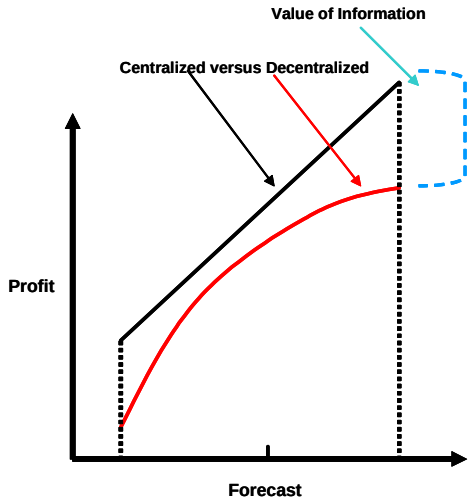


Figure 7: Value of Information Sharing – Increase in Total Supply Chain Profit and Performance (20)

One strategy for reducing inefficiencies in decentralized supply chain is 'vertical integration' where a company owns every part of its supply chain. A good example of vertical integration was Ford Motor Company. In today's economy, customer demand and preferences change rapidly. Companies that focus on core competencies are likely to be nimble to stay ahead of competition. Hence, we see a trend towards "virtual integration" where supply chains are composed of independently managed but tightly partnered companies. Information sharing based strategies, such as, vendor managed inventory (VMI) are used by some (Dell, P&G, Wal*Mart) industries.

Despite progress in information sharing, 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. To illustrate the ideas relevant to this article, we focus on one particular type of game, a Signaling Game (20).

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 share her demand forecast with the supplier so that the supplier may 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 her forecast. However, the supplier bears the cost of building capacity if it believes the manufacturer's (inflated) forecast. The manufacturer hopes the supplier ***will*** believe the (inflated) forecast and build 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 that the manufacturer's forecast is indeed credible?

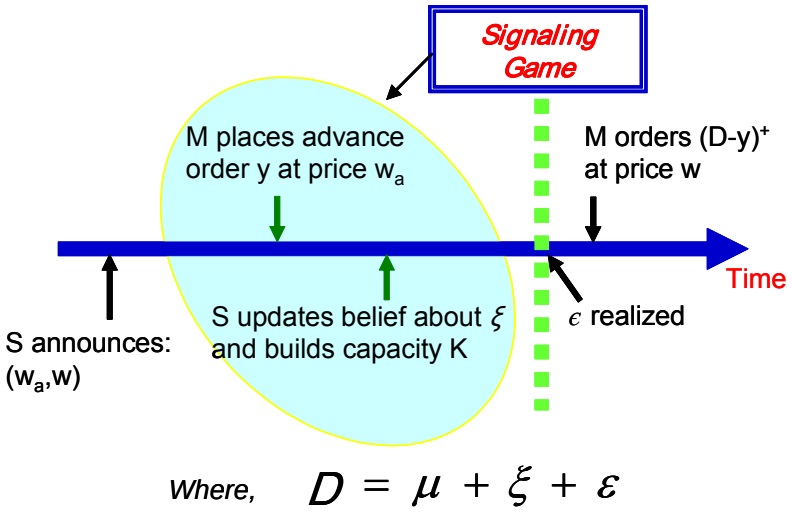


Figure 8: Signaling Game (20)

In this example, Demand (D) is represented as a sum of three forecasts. A market forecast mu (μ) is predicted by analysts. The manufacturer has sources and/or experience to derive private forecast information xi (ξ) which is not known 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 a value of ξ ***assumed*** to be represented by a normal distribution. This introduces a random (stochastic) variable. Market uncertainty is given by epsilon (ε) and neither the manufacturer nor the supplier can control its value. This introduces another random variable (error term) which is also ***assumed*** to belong to a normal distribution.

These assumptions introduces variability, that are not rigorously quantified, hence, the assumption that they belong to a function given by a normal distribution. Such errors successively accumulate from each stage of multi-stage supply chains (see Figure 9) and collectively contribute to the Bullwhip Effect. We shall advance a proposal, in a later section, to explore how these errors may be reduced through the use of real-time data in analytical tools that combine statistical methods with advances in time series econometrics (21, 26, 27).

The signaling game (20), shown above, commences with a price announcement by the supplier: w (regular) and w_a (advance purchase) price. The manufacturer creates a demand 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 her profit (inventory risk). Moving down the timeline, the market uncertainty is realized and using this value of ε the manufacturer updates its forecast. The volume D-y is ordered by the manufacturer from the supplier at a higher price (w). While optimization based on signaling may increase profits for manufacturer and supplier, it remains vulnerable to errors in the value chosen for the variables ξ and ε.

– Stage *i* places order q^i to stage *i+1*.
– L_i is lead time between stage *i* and *i+1*.

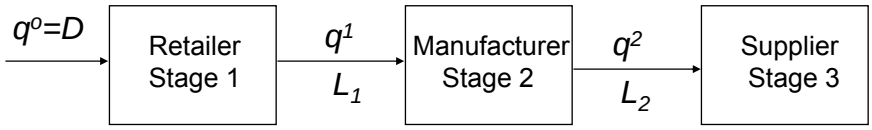


Figure 9: Can real-time data impact the traditional supply chain (5) at the right time?

The signaling game suggests that to reduce uncertainties, improving the values of the variables ξ and ε may be one right step forward. A vast array of research and optimization tools is already aimed at tackling these values or the ability to obtain dependable values. However, persistence of wide fluctuations in supply chains makes it unclear whether existing tools are adequate to stem uncertainty. The latter, in part, is one reason why we propose the use of real-time data to reduce errors, for example, for values of the variables ξ and ε. In addition, developments and techniques from AI may be helpful (see pages 68-69 and see notes M on page 104).

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 (3).	*"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 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." (22).*

ODD-VAR-GARCH: An Analytical Tool to Better Use Real-Time Data?

Forecasts made by electronics companies are often inflated. Now, Solectron has $4.7 billion in inventory. *Business Week, March 19, 2001*	*Cisco is stuck with chips, circuit boards and other components -- $2.5 billion worth of inventory that it believes it won't be able to sell within the next year.* *San Jose Mercury News, April 27, 2001*

Forecasts influence decisions and inaccurate forecasts can debilitate even the savviest corporations. Current state-of-the-art forecasting tools may be woefully inadequate and appears to be plagued by:

- poor data quality due to data acquisition errors or system-driven inaccuracies
- aggregated data punctuated by long intervals that dwell in static repositories
- restricted availability and/or visibility for planners or decision makers
- weak or inappropriate stochastic algorithms (poor data flow model)
- incomplete computing architectures to handle data or provide real-time decision support

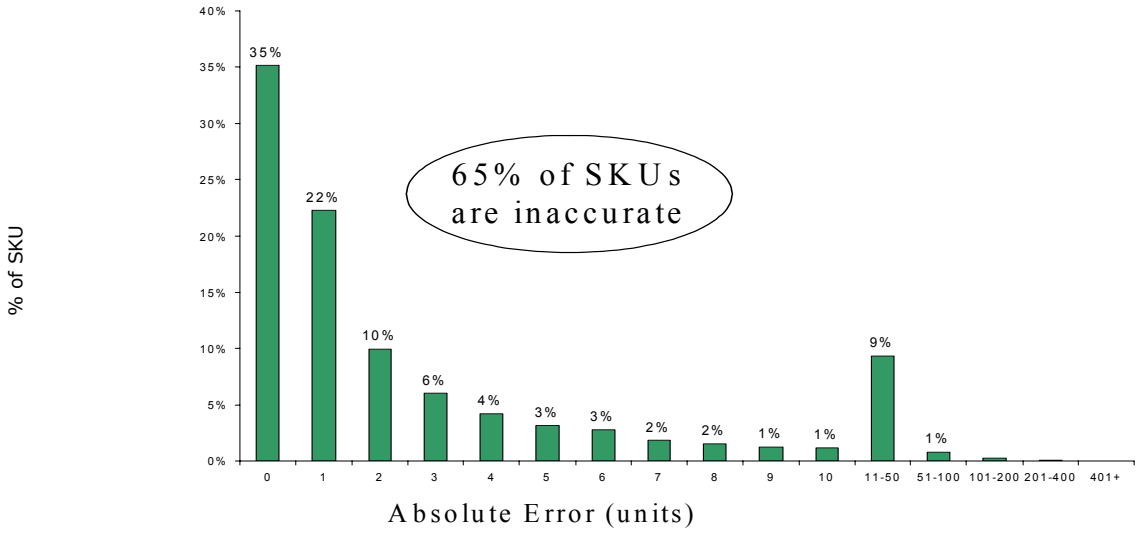


Figure 10: Inventory Record Inaccuracies (23)

Forecasting accuracy, then, is a function of data accuracy as well as the fit of mathematical models to business processes. Issues regarding data quality, granularity and stage-specific visibility may be addressed, albeit in part, by use of improved data acquisition through automatic identification technologies (RFID, ultrawideband tags, GPS, sensors). Whether or not the abundance of data is acquired and shared depends on AIT infrastructure investments to acquire the data and business strategy decisions to share the data between value network partners. Such practices, if materialized, shall bolster collaborative planning, forecasting and replenishment (CPFR), for example.

Data inaccuracies and infrequent data availability may have shaped current business process models that represent and/or analyze seemingly complex interactions by stripping away variables (using constants, assuming normal distribution) in a 'reductionism' approach. Perhaps most planners use weighted mean of historical data or classical linear regression models or simple smoothing technique for forecasting (24). In some corporations analytical teams may be reluctant even to use standard deviation and/or variance of historical data.

Even with better statistical tools that may be used in mission critical operations, such as supply chain management in the military establishment, the inadequacies stem from disconnect between operations and inventory. The spare parts inventory may not be coordinated with the ***process*** of demand and consumption of spare parts. Because the approved items, overall, are stocked at a certain level, it follows, that the metrics or key performance indicators (KPI) that monitor the inventory situation (divorced from operation) may not reflect the operational discontinuity. In reality, the operation (repairs) suffers since some of the spare parts or unique parts required to complete the job may not be available, as revealed from data summarized in Figure 11 (also, see Figure 27 on page 58).

Forecasting inventory levels or requirements to match the goal, common sense dictates, may benefit from better cross-sectional data visibility as well as integration with demand (field operations) and consumption (repairs). In an earlier book chapter (25) the author referred to the use of Agents in the repair process at Warner-Robbins Air Force base (44). Process-defining Agent based models (42a, 42b, 43) in conjunction with real-time data from AIT, if incorporated with the tools to be described in this section, may be begin to address and resolve some of the operational discrepancies that, in this case, affects readiness. To attain such readiness levels and respond to challenges in near real time, key decision makers may wish to ***implement the vision***, albeit in sequential steps, to bring about the ***confluence of semantic connectivity to data, process and decisions***. Tools discussed here and elsewhere are essential but use of any one tool (AIT-RFID) is only a *means* to an end, hence, a part of the vision.

AIT (especially RFID) proponents emphasize that automated data capture will eliminate a significant amount of errors (Figure 10). Improved data accuracy may be useful ***if*** data accuracy can be successfully married to process innovation to improve decisions. If Procter & Gamble suffers from a 10% out-of-stock (OOS) situation for its popular brands, the company's forecasting measures must be crying out for new blood. Similarly, if the military finds that lack of spare parts prevents it from certifying "mission capable status" for its aviation equipment, surely the element of 'readiness' is compromised due to mismatch between inventory and use of the unique spare parts.

Larry Kellam, director (…) at P&G, notes that reducing out-of-stock products by 10% to 20% could boost its (P&G) sales from $400 million to $1.2 billion.
Fortune, November 17, 2003

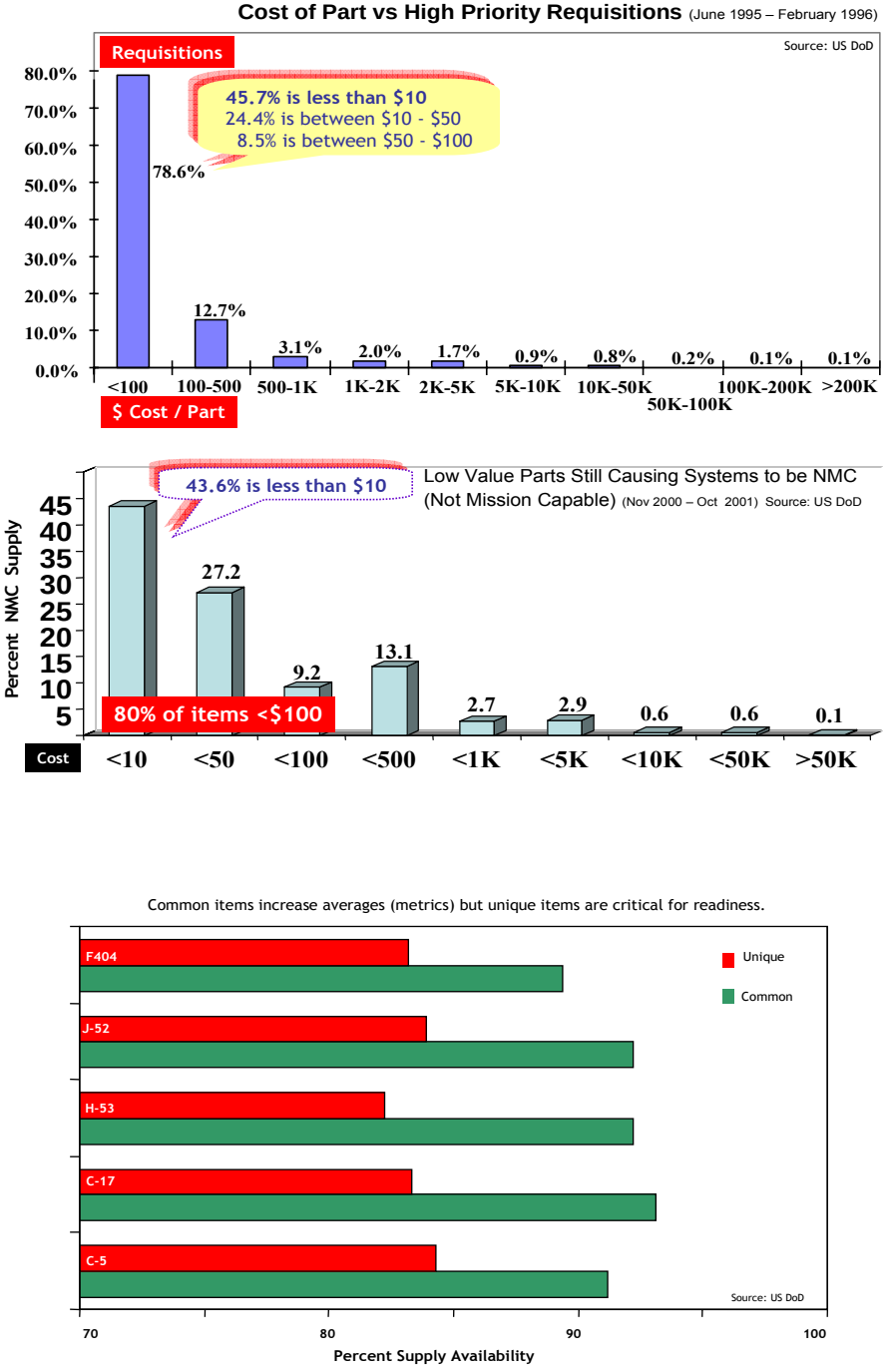


Figure 11: Inventory and Operations: Persistent Disconnect?

In parallel, with the surge of interest in AIT (RFID), a similar, but until recently, unrelated, level of interest has been emerging in Grid or utility computing architectures that allows flexible access to virtually any distributed device that can be connected to the internet. Grid was initially developed in the 1990's are now being extended to embrace the use of semantic language in its architecture (middleware). These approaches may usher in a new paradigm for computing, storage and communications to enable a more productive way to deal with business process, data handling, decision support and execution. We will briefly discuss Grid computing in the next section. This new paradigm may help implement some of the tools we are proposing for improved planning and forecasting.

Within the scope of process innovation, the component of forecasting (prediction), in our opinion, has much room for improvement. To utilize such improvements, organizations, first, must assimilate the vision. Second, if they wish to remain competitive, equip themselves to invest, explore, pilot and deploy advanced forecasting tools. It is well nigh impossible to deploy the existing financial econometric tools, as is, in the context of situations we are discussing in this article (supply chain, manufacturing, military readiness). The economic or financial models are significantly different from the models that these decision systems may require. Hence, a great ***opportunity*** is at hand to explore statistical and econometrics tools that may be modified for applications relevant to business scenarios soon to be faced with ultra high volume data. Surge in data may soon demand expression in terms of exabytes per second (1 exabyte = 10^{18} bytes or 10^{9} gigabytes). To extract value from the accurate high volume data, it may be shared through advances in open Grid services architecture (Grid computing will be briefly discussed later). In the medium term, by channeling data through appropriate middleware to feed pioneering forecasting tools, we may catapult the ROI from AIT (ROI from RFID) and make profitability gains replicable across industries and sustainable over several economic cycles. In the long term, expect semantic connectivity of data.

This proposal was outlined (26) as an interest-provoking point and also noted in the author's book chapter (25). Since its initial introduction, we decided to change the model name (abbreviation) to **ODD-VAR-GARCH** to reflect **O**bject **D**ata **D**ependent - **V**ector **A**uto**R**egression - **G**eneralized **A**uto**R**egressive **C**onditional **H**eteroskedasticity.

ROI from AIT may be only partially realized unless practitioners invest in deeper thinking about the ***processes*** that are likely to evolve. Process innovation between entities is key as well as data (semantics) availability maturing to visibility and transparency between stages in the supply chain, then, extending to the extra-enterprise or value network. The value of this data may be considerably improved by using analytical tools that combine advances in statistical and econometric modeling techniques. Thus far, to the best of our limited knowledge, the combined techniques, that we will propose, are not in use by supply chain planners or decision system analysts. It is quite possible that advanced corporations or organizations (military establishments) may have considered using these techniques but could not substantiate the models due to fewer than necessary reliable data points. Data 'points' may no longer be a limiting factor if AIP adoption increases. Thus, only now, the field may be gradually maturing to entertain the possibility of exploring time series and econometric tools in supply chain and decision systems.

Accurate model building that can be dynamically altered (hence, responsive and adaptive) is at the heart of this discussion. Evaluation, simulation and refinement of these process models for forecasting depend on very high volume accurate real-time data on the critical variable. For example, when forecasting sales or demand, the critical variable in the equation is 'sales' or 'demand' data. Time series models can relate 'current' values of a critical variable to its past (lagged) values and to the ***values*** of current and past disturbances or random error terms (rather than the assumptions discussed in the signaling game section).

Time series models, in contrast to econometrics models, may not be limited by its economic roots (hence, the scope to modify the tool, for the purposes of forecasting, in other areas). To explore model building for our purposes, let us re-visit the signaling game (20) and review what determines the market demand forecast (D):

$D = \mu + \xi + \varepsilon$

where,
- μ = market forecast
- ξ = manufacturers information
- ε = market uncertainty

To determine μ, planners and analysts may use one or more statistical tools (24) that may include:

[1] smoothing techniques (see note e, page 96)
[2] classic linear regression models (CLRM)
[3] autoregression (AR)
[4] moving averages (MA)
[5] ARMA (AR+MA)
[6] vector autoregression (VAR)

Classic linear regression models (CLRM) have been around for a century and widely used for a variety of purposes including some supply chain management software. CLRM may be expressed as an equation for a straight line:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \qquad \textbf{(0)}$$

where,
- **y** = dependent variable of interest to be modeled for forecast (for example, sales of a product, say aspirin)
- **t** = time period (frequency of observation, for example, t-1 may indicate prior week 1, t-2 → week 2)
- **β** = 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)
- **ε** = random (stochastic) error term

This simple technique can model multiple 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. Therefore, x's may be based on underlying economic principles (theoretical) and/or business logic (practical underpinnings). However, no matter how many x's are included, there may be an inherent randomness that cannot be explained by the model. Thus, the random error term (ε) is included in the equation (admission of the fact that the dependent variable (y) cannot be modeled perfectly). To solve for y, a bold ***assumption*** is made that ε is characterized by a ***normal distribution*** with a mean = 0 and variance = σ^2 for all time periods (t):

$$\varepsilon_t \sim N(0, \sigma^2)$$

The objective of CLRM is to estimate the parameters (β_0,β_1) of the model (from data on y and x), depending on the *sample* of observations on y and x. Therefore, there can be multiple sets of (β_0,β_1) that can, when plotted, produce straight lines with varying slopes (gradient). This statistical procedure introduces two sources of error.

First, taking *sample* data from a large number of observations inherits sampling errors. To eliminate this error, can we use raw AIT data instead of sample data? One reason for use of sample data (as practiced by the US Bureau of Census) may stem from lack of granular data acquisition tools. Another reason may be a lack of computing power. With low cost yet powerful microprocessors and the emergence of Grid computing, we may be increasingly better prepared to process exabytes of raw data. Second, given the multiple sets of (β_0,β_1) that may be estimated, the objective of CLRM is to choose that pair of (β_0,β_1) which minimizes the sum of squared residuals $(e_1^2, e_2^2, \ldots, e_n^2)$:

$$\sum_{t=1}^{n} e_t^2$$

where, e_t is the random error term for the *sample* and ε_t represents the random error term of the 'population' data. This technique is known as the principle of ordinary least squares (OLS). The sole objective of OLS is to minimize forecast errors by selecting the most suitable (β_0,β_1), thus ignoring the volatility of the sample.

The attractiveness of CLRM based forecasting stems from the fact that we can model cross variable linkages. The regression model is an explicit multi-variate model. Hence, forecasts are made not only on the basis of the variable's own historical data (for example, sales of aspirin, y, the dependent variable) but also takes into account the historical data of other related and relevant explanatory variables, x_1 through x_K, that is, any number of x's (inventory (x_1), price (x_2), expiration date (x_3)). In our example, the sales of a specific SKU, Bayer's Aspirin, may be modeled by the analysts of a retail outlet not only based on the history of its own inventory (x_1), price (x_2) and expiration date (x_3) but also taking into account the historical data with respect to inventory (x_4), price (x_5) and expiration date (x_6) of its competitor products $(x_{4t}, x_{4t-1}, x_{4t-2}, \ldots, x_{4t-n})$ sold in the same store.

To what extent is CLRM used by practitioners, today? Even this simplistic CLRM model, if combined with real-time (RFID) object-dependent data, may represent a step forward both in terms of accuracy of forecasting as well as determining the ROI from RFID. Given the emerging availability and abundance of real-time high volume data, we will extend this simple CLRM model to take advantage of the recent developments in time series techniques that has garnered a Nobel Prize in 2003 (30). It is this (combination or confluence of) idea that was first noted in the author's book chapter (25). It draws on ARCH (31) and GARCH models (32) hence, **ODD-VAR-GARCH**.

ODD-VAR-GARCH will require very high volumes of data and may deliver forecasts (or predictions) with far greater accuracy than any one of the individual components (CLRM, VAR, GARCH), separately. Abundance of data (RFID, UWB, RTLS, GPS, sensors) makes it ***possible*** to use ODD-VAR-GARCH. In addition to forecasts or predictions, it may be worth exploring in the future, how to predict the ***risk*** associated with the forecast (value at risk measure). Because these recent developments in time series techniques (ARCH) also offers a measure of risk (for financial analysis), it may be possible to use these tools to deliver a similar measure associated with forecasts for supply chain scenarios. For example, Pirelli may use these tools to predict how many tires it must manufacture for 2005. With this forecast at hand, it may apply to Bank of America for a loan to invest in production. Bank of America may wish to estimate the risk or validity of this forecast (number of tires) based on which the Bank may choose to modify the amount of the loan or interest on the loan or both or may even reject the loan application.

The yet untested concept of ODD-VAR-GARCH requires a few sequential stepwise progressions to combine CLRM with time series techniques. Let us develop the concept by starting with a basic CLRM equation:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_K x_{Kt} + \varepsilon_t \qquad \textbf{(1)}$$

The model given by equation 1 may be used to carry out 'what if' analysis. For example, what may happen to sales (y, dependent variable) of aspirin in the retail store if the in-store inventory of non-aspirin products were increased by 10%? The usefulness of this "what if" analysis is ***conditional*** upon assumptions we make about x's in the model. The effect of change in only one explanatory variable (x_1, … , x_K) may be analysed at a time (all else remains constant). Therefore, in building this model, the choice of x is a ***process*** decision based on the model builder's knowledge about the operation or business.

Because ***process dependency*** is critical for success and accuracy of predictions based on such models, it follows, that, feeding real-time automatic identification data to old process models may yield only minor benefits, if at all. Thus, focusing on the technology (RFID) to acquire the high volume accurate data may offer little value because ***use of the data*** requires far greater insight and process innovation. Since the latter is more involved and difficult, the market digresses to focus on the technology rather than ***use*** of the technology as a ***tool to improve decisions***.

Processes are reflected in the type of models given by equation 1. A task for process innovation is to explore how these x's (or their relevance) may change with real-time data to make the model adaptable enough to respond in near real-time. In the past, when supply chain tools were created for forecasting (assuming they were multi-variate models) they may have used explanatory variables (x's) for which (some) data was available (volume and quality of data remains questionable). Old process models may have excluded some highly relevant explanatory variables simply due to lack of data or data visibility with respect to stage-specificity (pallets in the distributor's warehouse versus in-transit). With the deployment of automatic identification technologies, there will be a surge of stage-specific data (work-in-progress, transit, theft) that presents a great opportunity for effective use of data. The opportunity in the area of forecasting and prediction may be closely linked to our ability to use this real-time data to adapt and respond in a manner that was not possible, in the past. To harvest data-dependent opportunities we must use the data in models that includes variables of the processes that may have surfaced, brimming with data. In other words, we may need to add new x's that now offer data and eliminate other x's that may have been made redundant by the abundance of AIT data. Process innovation, reflected in the choice of x's in new CLRM models should better define the operation and maximize the effective use of AIT data.

Note that data that may not be 'connected' to reflect the entire operational scenario (above). The latter may be made possible with the increasing diffusion of the semantic web. In future, improved CLRM models and accurate stage-specific RFID data may be monitored by Agents and shared through OGSA or open Grid services architecture (28). Applications in an OGSA environment may simulate scenarios based on observed data. These operations can take advantage of Grid computing to access (via semantic web services portal) applications hosted on a remote computer yet run the computation at the local site by harvesting unused processing power in its local domain (for example, powerful Pentium microprocessors in point-of-sale terminals). Thus, local data can help local as well as global optimization necessary to keep pace with the global economy.

Let us re-visit the 'what if' scenario. What happens to sales (y, dependent variable) of aspirin in the retail store if the in-store inventory of non-aspirin products increased by 10%? Are we playing a 'what if' game or is 10% increase 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 ODD-VAR-GARCH) where it is not necessary to assume values of the explanatory variable x (in this case 10% increase in inventory of non-aspirin products) to forecast y (the dependent variable). We start by forecasting the values of x's to obtain an ***unconditional*** forecast for y. Instead of inserting arbitrary values for future x's (such as, 10%), we use forecasted values based on historical data. To forecast x, we fit an univariate model to x where we use past (lagged) values of x to forecast x, as given in equation 2 (for x_{1t}, …, x_{Kt}):

$$x_{1t} = \alpha_{01} + \alpha_{11}x_{1t-1} + \alpha_{12}x_{1t-2} + \ldots + \alpha_{1N}\underset{x_{1t}}{x_{1t-N}} + u_{x_{1t}}$$

$$\vdots$$

$$x_{Kt} = \alpha_{01} + \alpha_{11}x_{kt-1} + \alpha_{12}x_{kt-2} + \ldots + \alpha_{1N}\underset{x_{kt}}{x_{1t-N}} + u_{x_{kt}} \quad (2)$$

$$y_t = \beta_0 + \sum_{i=1}^{N_{x_1}}\alpha_{1i}x_{1t-i} + \ldots + \sum_{i=1}^{N_{x_{kt}}}\alpha_{Ki}x_{Kt-i} + \varepsilon_t \quad (3)$$

$$y_t = \beta_0 + \sum_{k=1}^{K}\sum_{i=1}^{N_{x_{kt}}}\alpha_{ki}x_{kt-i} + \varepsilon_t \quad (4)$$

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

In equation 2, α_{11}, α_{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 technique. The error term in equation 2 represented by u is analogous to ε in equation 1. Depending on the number of x's (x_1, … ,x_K) that adequately represents the process being modeled in equation 1, there will be K number of equations (of the type equation 2) that must be estimated to forecast the x's (x_1, … ,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 (α, β) simultaneously by re-writing equation 1, the basic CLRM equation, as equation 3 or its shortened version, as in equation 4 (above).

Equation 4 is another 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 no sooner, we have moved a step ahead, it is 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 (below).

$$y_t = \beta_0 + \sum_{j=1}^{N_y} \varphi_j y_{t-j} + \sum_{k=1}^{K} \sum_{i=1}^{N_{x_{kt}}} \alpha_{ki}\, x_{kt-i} + \varepsilon_t \qquad (5)$$

Moving from conditional to unconditional forecasts of y using CLRM, it is evident that we are vastly increasing the number of parameters to be estimated. The latter necessitates high ***volume*** data. Precision in forecasting, in turn, demands ***accurate*** high volume data. AIT enables the acquisition of accurate high volume data. In equation 1, we estimate K parameters (β_1, … , β_K) excluding (β_0). In equation 2, we estimate n parameters (α_1, … , α_N) excluding the intercept (α_0) for each of the K number of x's (x_1, … , x_K). In equation 5 we estimate j parameters for lagged values of y_{t-j} (φ_1, … , φ_j) in addition to all the parameters for equation 2. If we set K=5 (only 5 explanatory variables, the x's), N=10 (number of lagged values to forecast the x's) and j=10 (number of lagged values of y_t), then, we have increased the number of parameters to be estimated from 5 in equation 1 to 50 in equation 4 to 50+10 = 60 in equation 5.

What is N? In the example above, it is the number of lagged values for each of the x's in the model. If N = 10, it could refer to daily data from past 10 days (N=1 indicates data from day 1). Is that sufficient? Let us assume that supply chain planners currently choose to use data from the immediate past 100 days. Then, using the traditional model, N = 100 days. With AIT, the granularity of data, say, from RFID, is expected to be far greater than using daily aggregates. Let us assume that real-time data is available in hourly buckets. Now, if we consider N = 100, we have used 100 hours or only 5 days worth of 'historical' data (assuming 20 hours as the 'active' handling time per day). If planners choose hourly aggregates of RFID data from the immediate past 50 days, then, N=1000. Similarly, let us say j=1000 and still keep K=5. Now, equation 5 now requires 6,000 parameters to be estimated!

Since 'historical' data is an agonizing cliché for products that have a tryst with obsolescence (high 'clockspeed' industries (58), such as mobile phones, digital cameras, laptop computers), the availability of high volume data may unleash analytical opportunities inconceivable in the past. It may be one more reason why the return on investment (ROI) in AIT (such as, RFID) may hold promise for industries that must meet immediate demand from customers (retail, consumer goods) or where the product lifecycle is short (electronics). For a product with a sales life cycle of 200 days (about 6 months), if we use the example of N=100 (past 100 days of data), it may be difficult to 'change course' and respond or adapt (based on forecasts or predictions after half the life cycle is over). The granularity of RFID type high volume accurate data, if available, may be modeled with N=100 where it is lagged every hour and the volume of data (item level) may be sufficient for reliable forecasts. For example, if hourly data is used and N=100, then predictive analysis can be made available within 5 days from launch of a product with 195 days or 97.5% of its sales life cycle still intact (if necessary to adapt and respond). Compare that to data flow on a daily basis. For the same number of parameters to be estimated, that is, N=100, suitable forecasts may be available only after 100 days or with 50% of the product sales life cycle still remaining. Thus, use of high volume real-time data in these models makes it not only possible but also feasible for sales, marketing, production or distribution to adapt. Changes can be initiated, based on forecasts, earlier in the (sales) life cycle of the product.

Increase in data volume made possible, by AIT, therefore, is ***necessary*** if reliability and accuracy of estimation is desired from the model given by equation 5 to forecast y. This "necessity" is rooted in the concept of 'degrees of freedom' which, by definition, is the excess of number of observations or data (EPC or GTIN product code) over the number of parameters to be estimated. Hence, the greater the volume of data, the higher is the degree of freedom. The precision of the forecast or prediction is directly proportional to the degrees of freedom.

Small and medium enterprises may vociferously complain that estimating 6,000 parameters or more (as we shall soon see) for each SKU is not feasible no matter how precise the outcome (forecast) may be. Such complaints may fade away when the power of Grid computing bears on the issue. In our opinion, the 'invention' of fire by our ancestors, *Homo australopithecus* may be analogous to what the Grid may be for precision forecasting. Estimating thousands or even millions of parameters to dynamically adapt and respond to changes and challenges may be a reality sooner that one might expect, due to semantic Grid services (discussed briefly, later). It is worth repeating that high throughput computation can be made feasible by harvesting unused processing power within many domains via the Grid. Thus, local autonomy or decision making is preserved, yet the system is not limited to local optimization. Decision makers and supply chain planners at subsequent higher echelons (local, regional, national, international) can harvest the local data or optimized data in near real-time to update or adapt the global coordinates and take advantage of economies of scale or risk pooling strategies. Thus, concurrent planning and execution through local and global optimization is a feasible scenario that relies on the ***confluence*** of real-time data (RFID), analytical models (for example, equation 5) and Grid computing. In a later section, we will discuss how Agents may occupy a prominent role in this confluence and continue to point out the need for semantic web.

To drive precision to the next (logical) step, equation 5 may be expanded 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. Because of the dominant role of trends and seasonality in some industries (and promotions linked to such events), the principle of cointegration (29) and its application in decision systems (other than econometrics) deserves a deeper analysis. (In a future version of this article we will explore, in further details, co-integration as well as risk with respect to value at risk (VaR) measure and use of extreme value theory.)

The concept of cointegration (29) began with the study of non-stationarity (a variable is non-stationary when it has no clear tendency to return to a constant value or trend). Ignoring the impact of non-stationarity, most practitioners perform regression analysis ***assuming*** stationarity. These results often produce "spurious regressions" that suggests a ***statistically significant*** relationship between variables (example, manufacturer offers discount and observed increased sales at a retail outlet) where none in fact exist (manufacturer may not know that people bought the lower priced item because the retail store's inventory was depleted of other equivalent products). Because of the complexity introduced by non-stationary variables, as late as the 1980's, most econometricians assumed that variables were stationary although they had full knowledge of the problem. Such assumptions may still linger (dominate) in the business world which tends to abhor academic complexity. Since promotions and promotion linked pricing may be related to seasonality and trends (promotions on cameras during Christmas), it may be profitable to explore the impact of cointegrated or multi-cointegrated variables.

This idea was developed by introducing the concept of degree of integration of a variable (29). If a variable can be made approximately stationary by differencing it ***d*** times, it is called integrated of order ***d*** or I(***d***). Weakly stationary random variables are thus I(0) or stationary without differencing. Many (macroeconomic) variables can be regarded as I(**1**) variables. If $z_t \sim I(\mathbf{1})$, then $\Delta z_t \sim I(\mathbf{0})$, in other words, the variable z_t can be made stationary by differencing z_t only once. Now consider the CLRM equation 0 (equation for a straight line):

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Assume that both $y_t \sim I(\mathbf{1})$ and $x_t \sim I(\mathbf{1})$. Then, generally $y_t - \beta_1 x_t \sim I(\mathbf{1})$, since I(**1**) variables dominate I(**0**) variables in a linear combination. There is an important exception. If the random error term, $\varepsilon_t \sim I(\mathbf{0})$, then $y_t - \beta_1 x_t \sim I(\mathbf{0})$. There exists one such combination so that coefficient β_1 is unique. It is in this special case that variables y_t and x_t are said to be ***cointegrated***. Then, the cumulative sum of deviations from the cointegrating relation $y_t - \beta_1 x_t = \mathbf{0}$ is necessarily an I(**1**) variable. If this new variable, say, w_t, is cointegrated with one of the original cointegrated variables (y_t or x_t) then y_t and x_t are said to be ***multi-cointegrated*** (29). For example, production (y_t) and sales (x_t) may be $y_t - \beta_1 x_t \sim I(\mathbf{1})$ and cointegrated, in which case their difference (change in inventory) is $y_t - \beta_1 x_t \sim I(\mathbf{0})$ variable. Then, the level of inventory (initial level plus cumulated changes) will be an I(**1**) variable (if y_t and x_t are cointegrated then $y_t - \beta_1 x_t = \mathbf{0}$**).** With a target level of inventory, defined as a fixed proportion of sales, inventory (w_t) and sales (x_t) would be cointegrated. This, in turn, makes the original variables, production (y_t) and sales (x_t), multi-cointegrated.

The concept of multi-cointegration may become an useful tool for equation based modeling (EBM) of stock-flow relationships, critical for forecasting in manufacturing and other industries. Although, a stochastic (random) trend may be removed by differencing (as discussed here), the impact of the concept of cointegration is far more profound when estimating vector autoregressive (VAR) equations that contain non-stationary variables (error correction). We will explain VAR, a component of ODD-VAR-GARCH, later in this section.

Thus far, discussions have centered on CLRM. Despite recent advances, CLRM is based on a set of assumptions mainly about ε, the random error (population disturbance term). Needless to emphasize, in the real world, these assumptions are almost always violated. Because CLRM is an old tool that many forecasters may be "comfortable" with, the 'deadweight' of old ideas could jeopardize the benefits that can be gleaned from new developments.

Developments in time series, over the past couple decades, are a welcome change and address the challenges that stem from the violation of these assumptions leading to inaccurate forecasts. The desirable properties of the OLS principle may cease to be useful if raw (AIT) RFID data can be used rather than *sample* data from a 'population' of (AIT) RFID data. Since ε_t is interpreted as the forecast error, when CLRM is used for forecasting, the assumptions made about ε_t the error term is especially important. Inadequacies of CLRM resulting in inaccuracies of forecast stems from the fact that expected value of the forecast error term, ε_t, when squared (thus, to homogenize negative or positive changes), is assumed to be the same at any given point or over time and across observations (cross-sectional data).

Review of cross-sectional data (conceptually illustrated in Figure 12) reveals that the ranges between highest and lowest values of the variables are often quite large. Consider observations for the same stock keeping unit (SKU), say, Bayer's Aspirin, from different entities (Albertson's in Tucson vs San Diego). The difference in weather, demographics, income may impact demand, sales, distribution and inventory. Yet, to simplify analytical models, some planners may choose to use simple regression techniques to find that 'best fit' line for (sales of) Bayer's Aspirin in Southwestern USA (bold dashed line in cross-sectional data, Figure 12, top) and assume that the errors will be represented by a normal distribution (Figure 12, bottom left). This is the assumption of ***homoskedasticity*** (homo ≈ equal, skedasticity ≈ variance or mean squared deviation). When this assumption is violated, it is referred to as ***heteroskedasticity*** (hetero ≈ unequal). Isn't it easy to grasp why heteroskedasticity may be the business rule rather than exception? Is there any link between assumptions about homoskedasticity and the Bullwhip Effect?

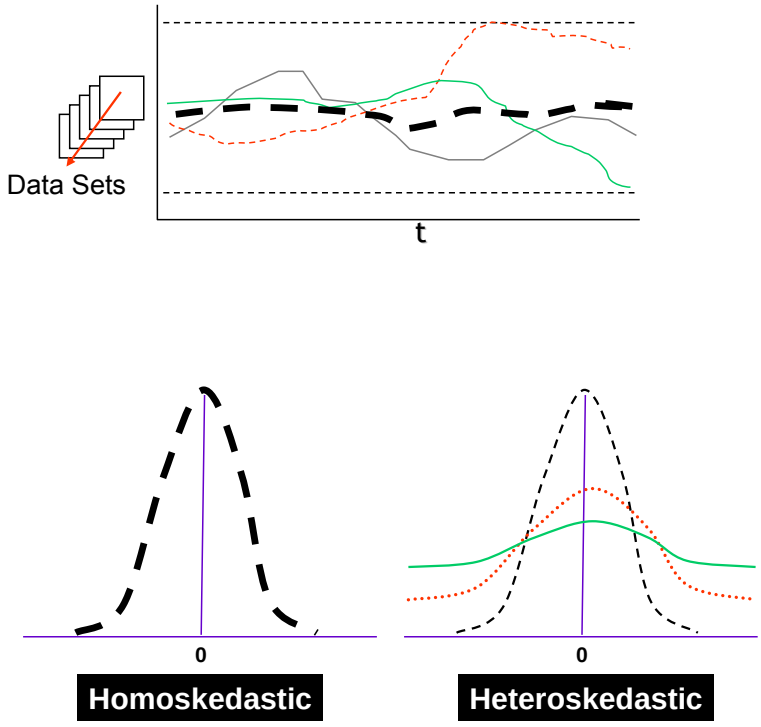


Figure 12: Cross-sectional data sets and homoskedasticity versus heteroskedasticity

Homoskedastic and heteroskedastic error term distributions are illustrated in Figure 12. All the observations of the error term can be thought of as being drawn from the same distribution with a mean = 0 and variance = σ^2 for all time periods (t) in homoskedasticity. With heteroskedasticity, the distribution of the error term depends on the observation (graph) and wide disparities may exist between the largest and smallest observed values as illustrated by the graph with differing widths (measure of variance). The degree of disparity of data increases the likelihood that error term observations associated with them will have different variances and hence will be heteroskedastic.

CLRM ignores the heteroskedastic behaviour of the error term ε_t and generates forecasts which may provide a false sense of precision because the volatility of the forecast is linked to the volatility of the error term ε_t (variance (σ^2) is a measure of volatility). The notion of time varying volatility is not unique for financial markets. In context of this discussion, one link between uncertainty and volatility (time- and/or stage-dependent volatility) is perhaps expressed as the Bullwhip Effect. Robert Engle shared the 2003 Nobel Prize in Economics (30) for his observation that not only is the volatility non-constant (of financial asset returns), it 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 the error term), Robert Engle seized this opportunity to model this non-constant variance (heteroskedasticity) using an autoregressive moving average (ARMA) technique (see note f, page 97).

ARMA has been in use for several decades and is a combination of AR (autoregression) and MA (moving average) techniques (24). We have already come across autoregressive (AR) representation in equation 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 indicates 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 to forecasters because a real-world model must link the present to the past (yet remain dynamic). MA expresses current observations of a variable in terms of current and lagged values of the random error ε_t , ε_{t-1} , ε_{t-2} ,…, ε_{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* represents the lagging order of AR and MA (24).

Engle used this ARMA technique to model the time varying volatility and proposed the AutoRegressive Conditional Heteroskedasticity model or ARCH (31). The 'conditional' nature of non-constant variance (heteroskedasticity) refers to the forecasting of variance *conditional* upon the information set available up to a time period (t). Modeling variance in this fashion allows us to forecast the volatility of the random error term (ε). Thus, ARCH also offers a measure of Value at Risk, for example, the risk associated with a forecast (see earlier example citing Pirelli and Bank of America). Using ARCH technique, the variance of the random error term (ε_t) in equation 5 can be expanded in terms of current and lagged values (ε_t , ε_{t-1} , ε_{t-2} ,…, ε_{t-q}), as follows:

$$\sigma_t^2 = \theta_0 + \theta_1\varepsilon^2_{t-1} + \theta_2\varepsilon^2_{t-2} + \ldots. + \theta_q\varepsilon^2_{t-q}$$

where,

σ_t^2 is the variant of ε_t [var(ε_t)].

This MA(q) representation of σ_t^2 was later generalized to an ARMA representation of σ_t^2 by Tim Bollerslev (then, graduate student with Robert Engle) and is referred to as GARCH (32). GARCH evolved when Bollerslev extended Engle's MA(q) representation of σ_t^2 (the ARCH model) by combining the existing MA(q) with an AR(p) process, that is, regressing a variable (σ_t^2) on its own (past) lagged values (σ_{t-1}^2 , σ_{t-2}^2 , … , σ_{t-p}^2). Thus, variance of the random error term (ε) in a certain period (ε_t) depends not only on previous errors (ε_{t-1}, ε_{t-2},….,ε_{t-q}) but also on the lagged value of the variance (σ_{t-1}, σ_{t-2}, … , σ_{t-p}). Thus, GARCH may be represented by equation 6:

$$y_t = \beta_0 + \sum_{j=1}^{N_y} \varphi_j y_{t-j} + \sum_{k=1}^{K} \sum_{i=1}^{N_{X_{kt}}} \alpha_{ki} x_{kt-i} + \varepsilon_t$$

$$\sigma_t^2 = \theta_0 + \theta_1 \varepsilon^2_{t-1} + \theta_2 \varepsilon^2_{t-2} + \ldots + \theta_q \varepsilon^2_{t-q}$$

Variance of the random error term depends not only on lagged values of ε (t-1, t-2, …., t-q) but also on lagged values of the variance σ² (t-1, t-2, …, t-p)

$$y_t = \beta_0 + \sum_{j=1}^{N_y} \varphi_j y_{t-j} + \sum_{k=1}^{K} \sum_{i=1}^{N_{X_{kt}}} \alpha_{ki} x_{kt-i} + \varepsilon_t$$

$$\sigma_t^2 = \theta_0 + \sum_{i=1}^{q} \theta_i \varepsilon^2_{t-i} + \sum_{j=1}^{p} \tau_j \sigma^2_{t-j} \quad (6)$$

Is it necessary to model σ_t^2 using GARCH to improve supply chain performance? The answer may not seem obvious unless one considers the real-world time and stage dependent volatility (orders, inventory) in their organization and the degree of uncertainty in its extended supply chain or value network.

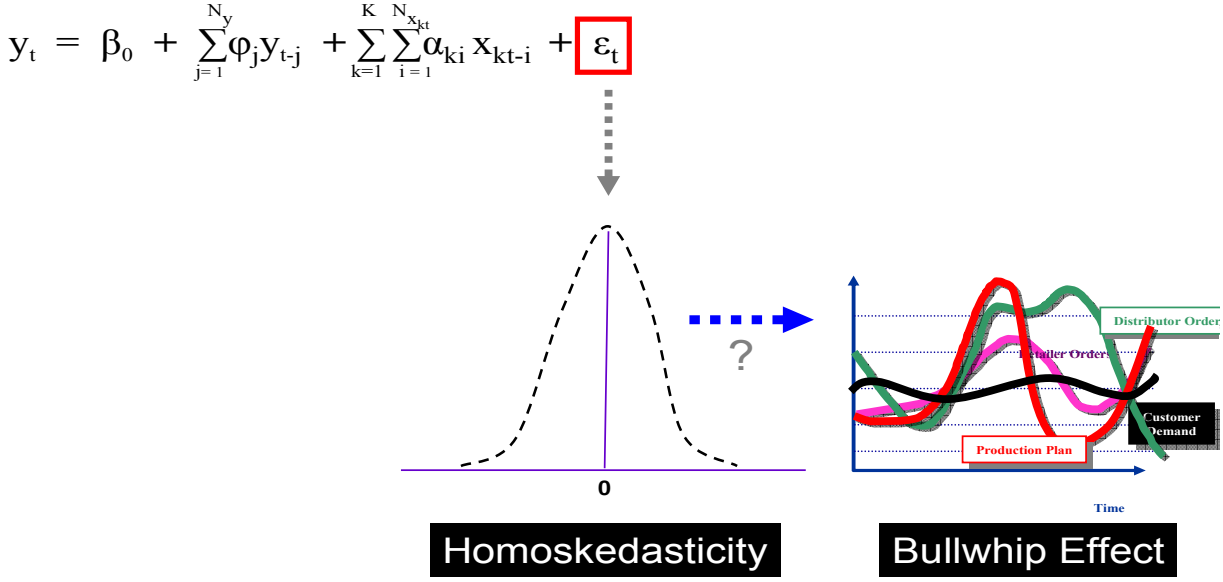


Figure 13: Does homoskedastic assumptions contribute to the Bullwhip Effect?

Since the lesson from GARCH is based on observations from the financial markets, it may be worthwhile to look at one example of stock price from the New York Stock Exchange (33). The author, ***without*** theoretical or practical proof, at hand, at this time, ***extrapolates*** that the nature of the data (reflecting volatility) shown in the stock price fluctuation (Figure 14, top), may be, in some ways, what we might expect from object-dependent data (EPC, GTIN from RFID) over time, between stages or geographies. Thus, GARCH may be useful for supply chain ***if*** high volume automatic identification (RFID, UWB) data is available. The latter is necessary to have enough degrees of freedom to use GARCH since the number of parameters to be estimated is even higher (compare equation 5 vs 6). The GARCH technique, consequently, may help to better substantiate the ROI from RFID.

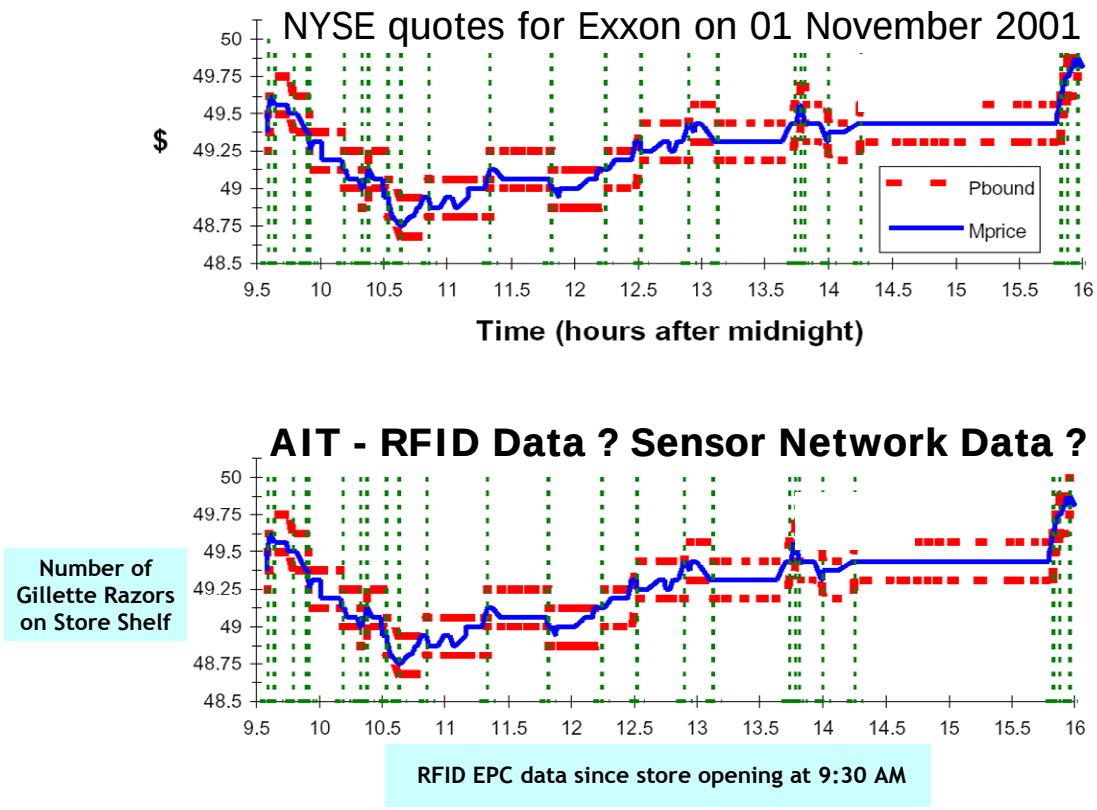


Figure 14: Extrapolation (33): Where Natural Stupidity meets Artificial Intelligence?

GARCH in Forecasting

Given the data up to time period t, we can use equation 6 to predict (say, sales of Gillette razors), h periods ahead (1, 2, 3, … , h) where h may be hours or days or weeks. Since forecasting using GARCH is already in practice in the financial markets, can we modify existing financial software (34), to aid in supply chain forecasting? The ***process*** model is a key to precision forecasting, hence, tools to create and validate such models may be a pre-requisite.

In developing the GARCH model, equation 6 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), the heteroskedasticity of the error term and the lagged values of the variance of the error term. But, we have not considered the fact that to predict sales h periods ahead, it is also crucial to model the interaction ***between*** the entities (manufacturer, supplier, distributor) in the value network which affects sales.

In supply chains, interaction between partners can impact any outcome (profit, service, readiness). Even if RFID improves data visibility between entities, collaborative strategies such as CPFR (collaborative planning forecasting and replenishment) may be still plagued by lack of trust between entities and efforts to share data or information, may be, sluggish, at best. The strikingly different dynamics of the supply chains partners fuels the Bullwhip Effect. To tame the Bullwhip Effect, it may be essential to model the dynamics between entities. The ODD-VAR-GARCH technique captures this dynamics through incorporation of VAR or vector autoregression (24) in addition to GARCH.

Vector AutoRegression (VAR) was developed a quarter century ago by Sims (24). Previously we discussed AR(p), which is an univariate model. In contrast, VAR(p) is a n-variate (multi-variate) model where we can estimate n different equations and in each equation we regress a variable on p lags of itself ***as well as p lags of every other variable***. The real-world cross-variable dynamics captured by VAR models enables each variable to be related not only to its own past but also to the past values of all other variables in the model. Univariate autoregression AR(p), cannot capture this multi-variate real-world dynamics that may be at the heart of business processes, such as supply chain management or CPFR and VMI.

For example, future sales (prediction) of Michelin brand tires may not be precisely forecasted by Sears unless the store takes into consideration the events (sales) at the distributor (to simplify the example). Thus, there are at least two parties in this example (interaction between store and distributor). To model this multi-variate dynamics of n=2 using VAR(p), let us assume that p=1 (lagged by 1 period). Equation 6 can be extended to the VAR-GARCH type to model two entities and consider only one lag period (n=2, p=1) as shown in equation 7.

$$y_{1t} = \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{x_{kt}}} \alpha_{ki} x_{kt-i} + \varphi_{11} y_{1t-1} + \varphi_{12} y_{2t-1} + \varepsilon_{1t}$$

$$y_{2t} = \beta_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{x_{kt}}} \alpha_{ki} x_{kt-i} + \varphi_{21} y_{1t-1} + \varphi_{22} y_{2t-1} + \varepsilon_{2t}$$

$$\sigma_{1t}^2 = \theta_0 + \sum_{i=1}^{q} \theta_i \varepsilon^2_{1t-i} + \sum_{j=1}^{p} \tau_j \sigma^2_{1t-j} \qquad (7)$$

$$\sigma_{2t}^2 = \theta_0 + \sum_{i=1}^{q} \theta_i \varepsilon^2_{2t-i} + \sum_{j=1}^{p} \tau_j \sigma^2_{2t-j}$$

Real world results or outcomes are generally influenced by events or interactions between decision domains. In the VAR-GARCH model represented by equation 7 (above), this dynamics is captured by estimating the coefficient φ_{ij} which refers to changes in y_i with respect to y_j. For example, if y_1 represents Michelin tire sales at Sears retail store and y_2 represents Michelin tire sales at the distributor, Merisol, then the parameter φ_{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 (ε_{1t} and ε_{2t}) changes, it will impact both the dependent variables (y_1 and y_2). For example, changes in the sales at the retail store may impact sales at the distributor. Thus, the VAR component, in ODD-VAR-GARCH, brings us closer to the real world scenario by making it possible to quantify cross-variable dynamics. For example, if ε_{1t} changes, it will change y_{1t} and since y_{1t} also appears as one of the regressors (explanatory variable) for y_{2t} in the equation, the change in any error term impacts both dependent variables in this VAR representation. These changes have been thus far ignored by current practices and may continue to fuel the Bullwhip Effect.

In the economics literature, the latter effect (impact of change in error term on dependent variable), is referred to as the Impulse Response Function (21). Impulse Response Function (IRF) may trigger new thinking about tools to explore "sense and respond" scenarios. IRF, as used by econometricians, can trace the impact of changes ('shock') in the error terms on the dependent variable for several periods in the future. Thus, the IRF concept (underlying process) may find considerable value if materialized into a simulation tool to explore ***multi-component*** "what if" scenarios by creating challenges and learning from the simulation how to prepare (readiness) for such challenges (fire, earthquake, epidemics, military escalations). What is the impact of one unit of change (shock) to ε_i on y_j and combinations of i and j, in the future? Such a tool may aid military planners or emergency agencies to prepare for containment of biohazard (Sandia Labs has created a somewhat related model to simulate spread of diseases). Although the complexity of equation 7 (n=2, p=1) may discourage a few, apparently estimating the coefficients in equation 7 can still use OLS (ordinary least squares method). Because OLS introduces errors, we will explore use of raw data. Since OLS and VAR already exists, one wonders, why isn't VAR a staple forecasting tool for business?

The author's assumption that VAR may not be widely used in businesses (non-financial) as a forecasting tool is not based on any direct knowledge. Since equation based models (EBM) have been in existence for centuries and VAR (ARCH, GARCH) have been in existence for decades, before the current surge to use RFID (wherever and however), then, why, a company as insightful and sophisticated as Procter & Gamble publicly claims (page 18) that simply reducing out-of-stocks (OOS is one process) of its products in retail stores is pregnant with the potential to add anywhere from $2 billion to $10 billion to its current sales revenue? Such claims are fueling and tempting for ROI analysts on behalf of RFID proponents. The claims, however, may not be irrational but the *confluence of tools and* ***vision*** that may enable to reap ***that ROI*** from RFID are unlikely to be a part of the 'bag of tricks' of those (vendors and consultants) who are generally focused on the "low hanging fruit" and eager to equate the financial rewards mentioned by P&G to be reflective of ROI from RFID deployment. In the opinion of the author, the opportunities are far greater when one considers value network performance (rather than individual processes, such as, OOS).

Forecast accuracy from VAR-GARCH models may be compromised in a number of ways, such as, the use of OLS. But, perhaps, degrees of freedom with respect to the number of parameters to be estimated, is a more important factor. We explored one scenario where we were required to estimate 6,000 parameters. Now consider a VAR model of multi-stage interaction with ten stages (n=10) from raw material supplier to retailer ($y_{1t}, \ldots, y_{10t}$). If we choose p=1000 (number of lags of y in hours), we must estimate φ for $y_{1t-1}, \ldots, y_{10t-1000}$ (or 10,000 φ coefficients for each equation) plus 5,000 for x's (K=5, N=1000) or 15,000 coefficients for each stage. The number of N however must be carefully chosen to limit 'historical' overemphasis. Nevertheless, for a ten stage VAR model we need to estimate 150,000 parameters (excluding constants and GARCH coefficients θ and τ). There must be, therefore, enough volume of data (sufficient degrees of freedom required for reliability and accuracy of forecast) to estimate over 150,000 parameters (excluding constants) to model this interaction.

The tsunami of data from RFID, therefore, is welcome. The lack of it, thus far, may be one reason why advanced organizations like the military may not have successfully used VAR-GARCH type models in the supply chain area. Perhaps, the timing is right, to explore ODD-VAR-GARCH, as a prediction tool. Estimating 150,000 parameters per SKU for the sake of accuracy of a future prediction may find it rather difficult to develop its ardent following (see note g, page 97). It is with the same degree of incredulity most of us reflect on the extent of absurdity of Thomas Watson, former chairman of IBM, who, in 1943, said, "I think there is a world market for may be five computers."

An important factor that may impact the value from ODD-VAR-GARCH, is the location of entities. Because these tools also analyse the impact between locations (equation 7), it is imperative that physical locations (retail store, distribution center) are optimized for performance (lead time, inventory management) by use of classical supply network planning (SNP) and operations research (OR) tools. ODD-VAR-GARCH will help predict values of the dependent variable h periods ahead but the precision from this tool will be of less value if the basic SNP is flawed. This consideration with respect to basic SNP optimization also has bearing on a more strategic question concerning the investment of technology, viewed by some, as a solution. The ROI from investing in infrastructure to acquire real-time RFID data from multiple nodes in an under-performing supply network is likely to be very poor. Under-performing SNP feeding RFID data to a sophisticated ODD-VAR-GARCH analytic tool may also fail to utilize the tool.

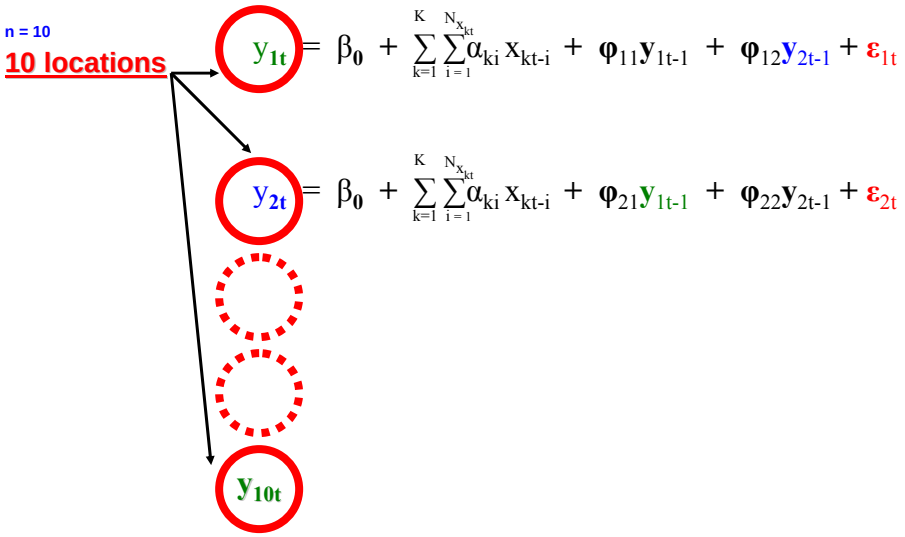


Figure 15: Optimizing locations by using supply network planning tools may be a pre-requisite for ODD-VAR-GARCH

The proposed analytical tool, ODD-VAR-GARCH, outlined in equation 7, is essentially modular. Businesses that use arithmetic mean as their sole guiding light, may prefer "not to boil the ocean" by jumping into ODD-VAR-GARCH or invest in the effort required to successfully use this tool. Tracing back the equation to its components, one could start with CLRM and then move to AR or MA before engaging with VAR and then refine it with GARCH. Thus, ODD-VAR-GARCH when successfully married to high volume accurate data from automatic identification technologies, with the help of (semantic) Grid computing, can, potentially evolve as a tool for organizations seeking accuracy in forecasting demand ($D = \mu + \xi + \varepsilon$) and predicting figures by combing equation based models (EBM) with Agent-based models (ABM). The combination is crucial. EBM, alone, no matter how sophisticated, may be less dynamic, hence, less adaptive or responsive, than EBM plus ABM, when used appropriately. In a later section we will re-visit ODD-VAR-GARCH and suggest how it may be integrated to function in Agents-based systems (see pages 67-69).

Finally, a word about non-linearity in decision systems. Essentially most systems are non-linear (webs, not chains). It is imperative that ODD-VAR-GARCH accommodates non-linearity. By definition, GARCH models are estimated using Maximum Likelihood Estimation (MLE) technique which is more general than OLS (ordinary least squares) and can be used for linear and non-linear models. In fact, a multivariate GARCH model may seem more appropriate for our consideration. It is similar to the model described above but includes equations that specify how covariances are changing over time, across series (y_1, y_2, …, y_n). We will explore multivariate GARCH in a future version since it may better capture cross-market volatility that is common across value networks.

The Grid

Demand for increasingly better results (profit, customer service, readiness) may make simulation an indispensable planning tool for decision makers. Data-dependent analysis may optimize the simulation model and the process may be repeated by feeding yet more data to such 'optimized' model to further refine or improve accuracy. Even without simulation (what if), based on our discussion about ultra high volume of object-dependent data (RFID), equation based models, such as ODD-VAR-GARCH, may require thousands of parameters to be estimated for single objects (SKU). Therefore, the future of business data analyses may have radically different computation needs. It may not be surprising if business computation demands approach or exceed the requirements for collaborative scientific research, such as, data analyses from high energy particle physics experiments.

Assuming 'positive' collaboration and data sharing trends that may continue to increase, plus the appreciation of ontologies (semantic web), one may need to re-think the data flow model to feed analytical tools in near real-time (applications). Data may originate from highly dispersed sources during the lifecycle of a product. That is, it may be acquired in several locations (as objects, components, spares, migrate from one supply chain stage to another) over widely dispersed geographies (manufacturing, military). Locally acquired and stored data may be globally necessary in near real-time for end-to-end transparency. Decision tools may drive real-time data to feed applications that may not be co-located (de-coupled processes). But, the results of the analysis may be required 'centrally' (global optimization) to concurrently re-plan and execute. However, such execution may also require local flavours (local optimization). The ***heterogeneity of processes*** for business to business exchanges (between businesses and geographies) may be an initial fledgling barrier that may succumb to spread of the semantic web once the context and 'meaning' of the terms (words) are 'understood' by the system (computer, software). Thus, the transformation from the syntactic to the semantic web, through use of ontologies, ontological mapping and semantic tags, may enable businesses to maintain their "lingo" yet be "understood" when interacting with other businesses or geographies, where the processes may be defined differently. The spread of the semantic web and its integration with grid computing and web services, may improve security and eliminate, some of the business to business process standardization attempts, for example, the commercial consortium referred to as the RosettaNet.

Supply chain analytics, thus, demand innovative applications, large-scale resource sharing, high-performance and high-throughput computing that can function in a dual environment, that of, equation based modeling (EBM) and Agents based modeling (ABM). This implies the use of the Grid that is different from conventional distributed computing (see note h, page 97). Grid computing promises a world where shared resources, data, tools, Agents, people, are coordinated and integrated in "virtual organizations" accomplished through (semantic) web-based portals together with wide-area distributed applications for highly productive collaboration (35-37). Distributed computing was proposed as a tool for this collaboration. However, distributed computing, in its current form, may find it difficult to pave the path for the emergence of ***value networks*** since such multi-dimensional interactive collaboration demands data and process visibility across operational domains and organizational boundaries.

Grid computing and its properties, we suggest, offers advantages necessary for robust value networks to emerge. The barrier to Grid computing in certain environments, such as, business, is not the inability to implement or invest in the technology but the inability to collaborate at the level of process, data and information sharing. Large scale scientific research (human genome project, large hadron collider) is based on collaboration, where resources (computing power, applications) are also shared. However, the business mindset views data, resource or application sharing with justifiable mistrust and skepticism. The ability to attach rules or sharing policies to data or information will be made possible by the semantic web and may partially alleviate the widespread mistrust.

Irrespective of the business culture, it is almost an undeniable truth that, in a couple decades, Grid computing will enable any person with a portable computer or even a mobile phone, to have the power of a supercomputer at her fingertips. The tools that will unleash that power of Grid computing may be comparable to the explosion of the internet catalyzed by the world wide web. Imagine, for example, a handful of concerned citizens, eager to get away from the grasp of the pseudoscience peddled by the environmental militia. They can now run their own simulation of the benefits of lower energy cost, greater productivity and less pollution in their environment and on their standard of living, from use of safe (portable) nuclear energy. To conduct the impact of a proposed energy development in the community they neither need their own data center or consultants from 'Red Peas' or other disenfranchised groups. The citizens can describe what they want and intelligent semantic software will find the relevant data as well as summon the computing resources needed for the simulation. Small or large, groups or businesses, with shared interests may find dependable answers to very complex problems that may have been, computationally, a challenge, only a decade ago.

There has been a flurry of Grid projects in the last few years in the US, EU and Japan. More recently, computer companies including IBM, Sun, HP and Microsoft have become increasingly interested and invested in Grid tools and technology, as some of the early commercial applications emerge. In July 2003, Grid computing moved further toward the commercial mainstream when the Globus Project released new software tools that blend the Grid with web services. However, web services, in its current state, still remains a merely automated computer-to-computer communication medium. Yet, it harbours the potential to be an useful platform for the future and re-emerge as the ***Semantic Grid Web Services*** (Figure 21, page 49). Though Grid technologies are distinct from the internet, distributed and peer-to-peer computing, these areas may also reap benefits from growing into the problem space addressed by the Grid.

Due to the impact of the Grid in the near future, we will briefly discuss Grid computing and extrapolate some uses in the business environment that may be relevant to use of AIT (RFID). Readers may wish to note that most of the Grid computing related material presented here is from the work by the creators of the basic software (Ian Foster and Steve Tuecke, Argonne National Laboratory, University of Chicago and Carl Kesselman of Information Sciences Institute, University of Southern California) and NASA (28, 35, 36, 37).

The Grid is defined as flexible, secure, coordinated ***resource sharing*** among dynamic collections of individuals, institutions and resources referred to as VO or virtual organizations (28, 35, 36, 37). It encounters authentication, authorization, resource access, resource discovery and other unique challenges. These issues are important since business to business (B2B) collaboration is crippled by the lack of trust that segues into security concerns when businesses start interacting, virtually. Thus, this class of problem must be addressed by Grid technologies in order to begin to persuade businesses to invest in Grid computing. Within an extensible and open Grid architecture, there exist protocols, services, application programming interfaces (API) and software development kits (SDK) which are categorized according to their roles in enabling ***secure resource sharing***. A set of interGrid protocols and inclusion of ontological mapping in such protocols may enable interoperability among different Grid systems. The latter may impact how Grid technologies relate to other contemporary technologies, including enterprise integration, application service provider, storage service provider and peer-to-peer computing.

Use of Grid technologies for enterprise integration may offer a different perspective with respect to enterprise resource planning software that claims to provide "connectivity" of operations (procurement, sales, distribution). ERP is generally limited to large businesses yet key suppliers are often small businesses (SME's). They need to interact with behemoths but cannot afford EDI or expensive ERP packages. This introduces system inefficiencies. For example, a small business that supplies knitted gloves to match Patagonia brand ski jackets, may need a sliver of information (updates) from Patagonia's supply chain demand planning module or inventory management system (part of ERP) to better synchronize their production of gloves with demand (anticipated demand) for ski jackets. The author has proposed that resource sharing enabled by telco networks or the Grid may help explore the concept of "distributed" ERP where the channel master or lead business can share resources, virtually, with small and medium partners in its value network (virtual organization of value network or VOVN). A simple use of the Grid may be 'vanilla' visibility of objects with respect to stages. Such 'vanilla' visibility was explored by NTTDoCoMo in Japan around 2001. Through its i-Mode mobile phones, NTTDoCoMo seemed to enable access to real-time data regarding location of objects, if they were affixed with a RFID tag (www.ntt-east.co.jp/tmmall/rf.html).

The real and specific problem that underlies the Grid concept is coordinated resource sharing and problem solving in dynamic multi-institutional virtual organizations. The sharing is not about primary file exchange but rather ***direct access*** to computers, software, data and other resources, as required by a range of collaborative problem-solving and emergent resource brokering strategies. This sharing must be highly controlled, with resource providers and consumers defining clearly and carefully just what is shared (rules, business logic), who is allowed to share (policy) and the conditions under which sharing occurs (policy, goals). Set of individuals and/or institutions defined by such sharing guidelines, form, what is referred to as a virtual organization (VO). The use of Agents as security guards is a distinct possibility and perhaps a necessity for proliferation of such VO's or virtual B2B environments.

VO's may include application service providers, storage service providers, cycle providers, group of consultants engaged by a manufacturer to perform evaluation, members of an industrial consortium bidding on purchases, crisis management team and the databases and simulation systems they use to plan a response to an emergency situation, industry-academic associations like IEEE or scientific communities (members of an international high energy physics collaboration). Each of these examples represents an approach to computing and problem solving based on collaboration in computation in data-rich environments. VO's vary in their purpose, scope, size, duration, structure, community and most significantly, in their sociology. Yet they all share the need for:

[1] highly flexible sharing relationships (ranging from client-server to peer-to-peer, for sophisticated and precise levels of control over how shared resources are used, including fine-grained and multi-stakeholder access control, delegation and application of local-global policies)
[2] sharing of varied resources (ranging from programs, files and data to computers, sensors and networks)
[3] diverse usage modes (ranging from single user to multi-user and from performance sensitive to cost-sensitive, thus, embracing issues of quality of service, scheduling, co-allocation and accounting).

Distributed computing technologies, currently, do not address these concerns and requirements (28). Hence, the need for the Grid. For example, the 'markets' euphoria (Ariba, SAP Markets, Commerce One) of the recent past and current wave of 'web services' claim to address communication and information exchange among computers but do not provide integrated approaches to decision making or coordinate use of resources at multiple sites for computation. B2B exchanges focus on information sharing via centralized servers but the Grid shall extend it to applications and physical devices (the concept of emerging device-to-business or D2B capabilities). Enterprise distributed computing technologies (CORBA and Enterprise Java) enable resource sharing ***within*** an organization.

The OAG's Distributed Computing Environment (DCE) supports secure resource sharing across sites but may be too inflexible. Storage service providers (SSPs) and application service providers (ASPs) allow groups to outsource storage and computing requirements to other parties under quite rigid constraints. SSP resources are linked to a customer via a virtual private network.

Emerging 'distributed computing' companies seek to harness idle computers on an international scale but, to date, can support only highly centralized access, albeit limited, to such resources. Therefore, current technology cannot accommodate the range of resource types or does not provide the flexibility and control on sharing relationships needed to establish functional and dynamic VO's. Ability to remain dynamic is the key to adaptability. The latter is important in the context of adaptive value networks where business partners and interactions can change often and suddenly. It is here that Grid technologies can offer value. Over the past decade, research and development efforts within the Grid community have produced protocols, services and tools that aim to address the challenges that arise when we seek to build and scale virtual organizations. These technologies include:

[1] security solutions (management of credentials and policies when computations span multiple organizations)

[2] resource management protocols (secure remote access to computing, data, co-allocation of multiple resources)

[3] information query protocols & services (configuration, status, information resources, organizations, services)

[4] data management services (locate and transport datasets between storage systems and applications).

Because of their focus on dynamic, cross-organizational sharing, rather than intra-organizational connectivity, Grid technologies ***complement*** rather than compete with existing distributed computing technologies. For example, enterprise distributed computing systems can use Grids to achieve resource sharing ***across*** organizations. ASP's and SSP's can overcome the limitations due to their static configurations by using Grid technologies to establish dynamic markets for computing and storage resources.

Thus, supply chains, if still embedded in processes that result only in *sequential* optimization, may feel adequately served by distributed computing. The undeniable necessity for *global* optimization and business growth through value networks, may stimulate the practitioners to embrace the Grid, if they wish to remain viable. However, current business analytics and sophistication of business model simulations may not be advanced enough, yet, to outpace the available advantages from simple distributed computing.

Gradual implementation of the generalized concept of VO's, for example, in the form of value networks, harbours the potential to dramatically change use of computers to solve problems, in the same way that the world wide web has changed the dimensions of information arbitrage. VO concepts are not limited to business, science or engineering. But the thinking in terms of Grid is still in its infancy. Even scientists and engineers must begin to consider that the Grid is ***their computer***, rather than simply a disparate collection. The latter will gradually stimulate new thinking with respect to algorithms and application paradigms to take advantage of the advanced capabilities, such as ODD-VAR-GARCH. The Grid will enable process simulation with a completely new level of realism that could integrate experimental and theoretical work between diverse groups and organizations, in real time, in ways never before possible. The broad spectrum applicability of the concept of VO adds to the importance of Grid technology as a potent catalytic tool for decision makers in an uncertain and unconnected world.

Consider the following scenarios (from 36):

[1] A supplier's Agent detects the threat of a looming hurricane that triggers the transport Agent to alert the recipient that the supplier's raw materials may not reach the manufacturer's plant in San Juan. The manufacturer must decide whether to re-direct the shipment for delivery to its Miami flexi-plant, in order to make its deliveries on time. Thus, it invokes a sophisticated forecasting model, ODD-VAR-GARCH, from an ASP and provides it with access to appropriate proprietary historical data from a corporate database on storage systems operated by an SSP. During the decision-making meeting, what-if scenarios are run collaboratively and interactively between other members of the value network, even though the decision makers participating in the decision are located on different continents. The ASP contracts with a cycle provider for additional computation power during particularly demanding scenarios or calculation of analytical coefficients. High volume data driven, flexible manufacturing and cross-domain decision making enables the manufacturer to manage the uncertainty and reduce the risk to its product delivery schedule despite a sudden change in raw material availability (see figure 31 on page 66).

[2] A crisis management team responds to a chemical spill by using local weather, hydrology and soil models to estimate the spread of the spill, determine the impact based on population location, service availability (hospital inventory) as well as geographic features such as vegetation, rivers and drinking water reservoirs, creating a short term mitigation plan (based on chemical reaction models that determines reactivity of spilled chemical). The plan then incorporates human resource skills (fire, police, hospital emergency response personnel) and delegates tasks to an *ad hoc* emergency response team for coordination, treatment and evacuation.

[3] A gravitational wave detector (GEO600) in Hanover, Germany, detects a cosmic catastrophe (black hole or neutron star collision). Astrophysicists and astronomers around the world are alerted to turn their telescopes to view the ephemeral event but determining the location in the sky requires a time-critical analysis. In Berlin, an astrophysicist accesses the GEO600 portal and using the performance tool, finds the resources required for cross-correlating the raw data with the available templates. The brokering tool finds the fastest affordable machines around the world. Merely clicking to accept the portal's choice initiates a complex process by which executables and data files are automatically moved to these machines by the scheduling and data management tools necessary for the analysis to commence. Twenty minutes later, on his way home, the astrophysicist's mobile phone receives a message from the portal's notification unit, informing her that more templates are required (must be generated by a full-scale numerical simulation). She immediately contacts an international collaboration of experts for such simulations. Using a tool in their simulation portal, the experts assemble a simulation code with appropriate physics modules suggested by present analysis. The portal's performance prediction tool indicates that the required simulation cannot run on any single machine to which they have access. The brokering tool recommends that the simulation run across two machines, one in USA and the other in Germany (connected to form a virtual supercomputer) to accomplish the job within the required time limit. The simulation begins and after querying a Grid Information Server, decides autonomously to spawn off a number of time-critical template generating routines to run asynchronously on various other machines around the world. An hour later, the network between the 2 machines degrade and the simulation again queries the server, this time deciding to migrate to computers in Finland and Taiwan while still maintaining connections to the various template generators at other sites. All the while, the international teams of collaborators are monitoring the progress from their workstations and wireless devices (some members of the virtual team are in an airport). In a few hours the template data are assembled and sent to the GEO600 scientist in Hanover, Germany.

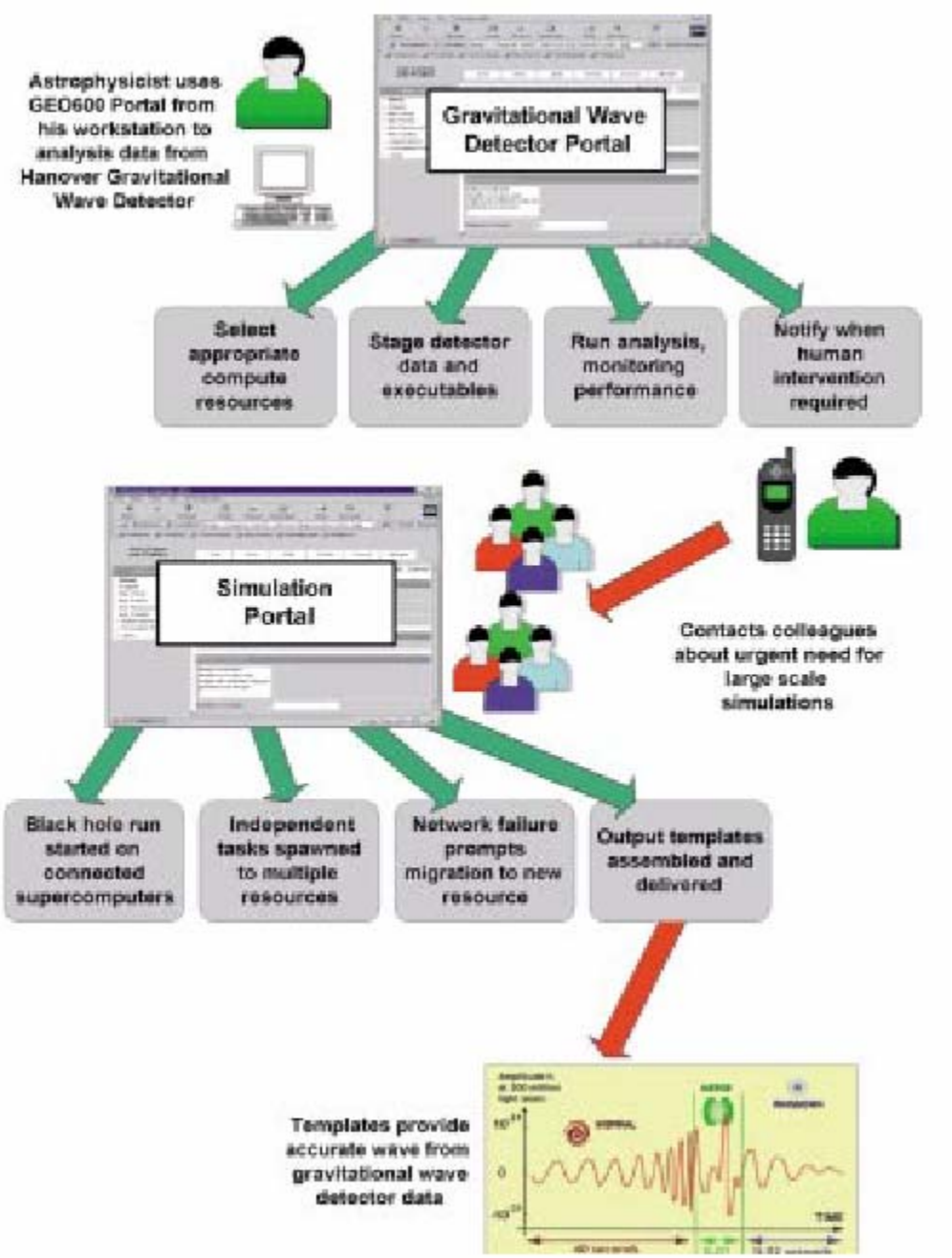


Figure 16: A Grid Application Environment (from 36)

This scenario heralds a future that supports the seamless connection of people, resources, simulations and devices. Such an environment depends on the construction of an infrastructure of fundamental services for Grid computing. This infrastructure must eliminate discontinuities (between machines, policies, file storage, operating systems), must enforce trusted security and tolerate extreme faults (recover from failure of any single component).

The establishment, management and exploitation of dynamic, cross-organizational sharing relationships require the Grid architecture to identify fundamental system components, specify the purpose and function of these components and indicate how these components interact with one another. An effective VO must be able to share relationships among any potential participants. Interoperability in a networked environment stems from common protocols. Hence, Grid architecture is first and foremost a ***protocol*** architecture, with protocols defining the basic mechanisms by which users and resources negotiate, establish, manage and exploit sharing relationships. A standards-based open architecture facilitates extensibility, interoperability, portability and code sharing. Standard protocols make it easy to define standard services to provide enhanced capabilities. Application programming interfaces (API) and software development kits (SDK) provide the programming abstractions required to create a usable Grid. The technology and architecture constitute the ***Grid middleware*** services needed to support a common set of applications in a distributed network environment, such as, a supply chain or value network.

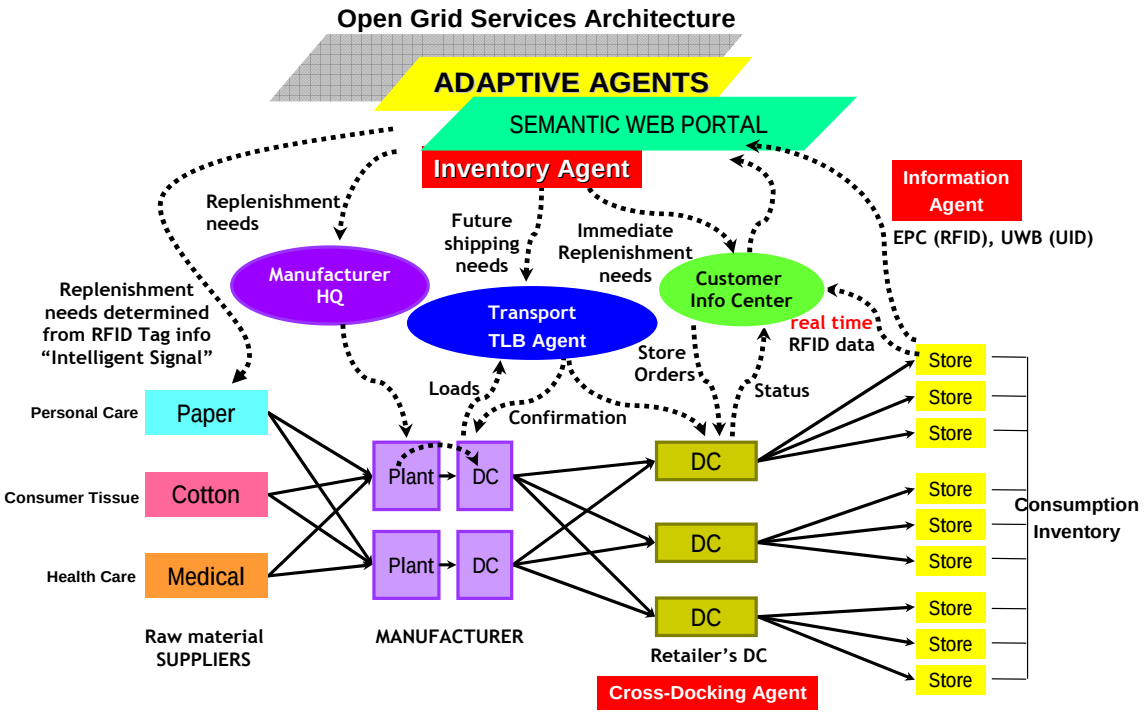


Figure 17: Adaptive Value Network?

Why is interoperability a fundamental concern? Consider what happens to your supply chain when you must change your RosettaNet-compliant supplier and source from a manufacturer in Shinzen (China) who has never heard of RosettaNet or its standard PIP's? The Grid must enable sharing relationships that can be initiated among ***arbitrary parties***, accommodating new participants dynamically, rapidly, across different platforms, languages, geographies and programming environments. Mechanisms serve little purpose if they are not defined and implemented to be interoperable across organizational boundaries, operational policies and resource types. The syntactic web further aggravates these inefficiencies, hence, the anticipation that the semantic web may improve context connectivity.

It follows, therefore, that protocols are equally critical to interoperability. The web revolutionized information sharing by providing universal protocol and syntax (HTTP, HTML) for information exchange. Grid requires standard protocols and semantics for general resource sharing. Grid protocols, by definition, will specify how distributed system elements interact with one another (in order to achieve a specified behaviour) and the structure of the information exchanged during this interaction. In future versions, ontologies could play an increasingly important role in 'understanding' the information exchanged through the use of semantic language (Web Ontology Language).

The focus on external rather than internal interactions (software, resource characteristics) may offers important benefits for supply chain management where the loose federation of partners tends to be fluid. Hence, the mechanisms used to discover resources, establish identity, determine authorization and initiate sharing must be flexible enough to be established and changed quickly. Because virtual organizations, such as a value network, complement rather than replace existing institutions, sharing mechanisms cannot require substantial changes to local policies and must allow individual institutions to maintain ultimate control over their own resources. Since protocols govern the interaction between components (not the implementation of the components), local control is well preserved. The business partner in Shinzen, China, therefore, may not have to re-invent its B2B system or subscribe to RosettaNet in order to engage with you in a business relationship as a new member of the value chain.

Figure 18 illustrates the various layers of the Grid architecture (28) that follows a 'hourglass' model where the 'narrow neck' of the hourglass defines a small set of core abstractions and protocols (TCP and HTTP for internet) onto which many different high level behaviours (top of the hourglass) and underlying technologies (base of the hourglass) can be mapped. In Grid architecture, the neck of the hourglass consists of ***Resource*** and ***Connectivity*** protocols, which facilitate the sharing of individual resources. Protocols at these layers, defined in the **Globus Toolkit**, are designed to be implemented on top of a diverse range of resource types, defined at the ***Fabric*** layer and can, in turn, be used to construct a wide range of global services and application-specific behaviours at the ***Collective*** layer, so called because it involves the coordinated or "collective" use of multiple resources.

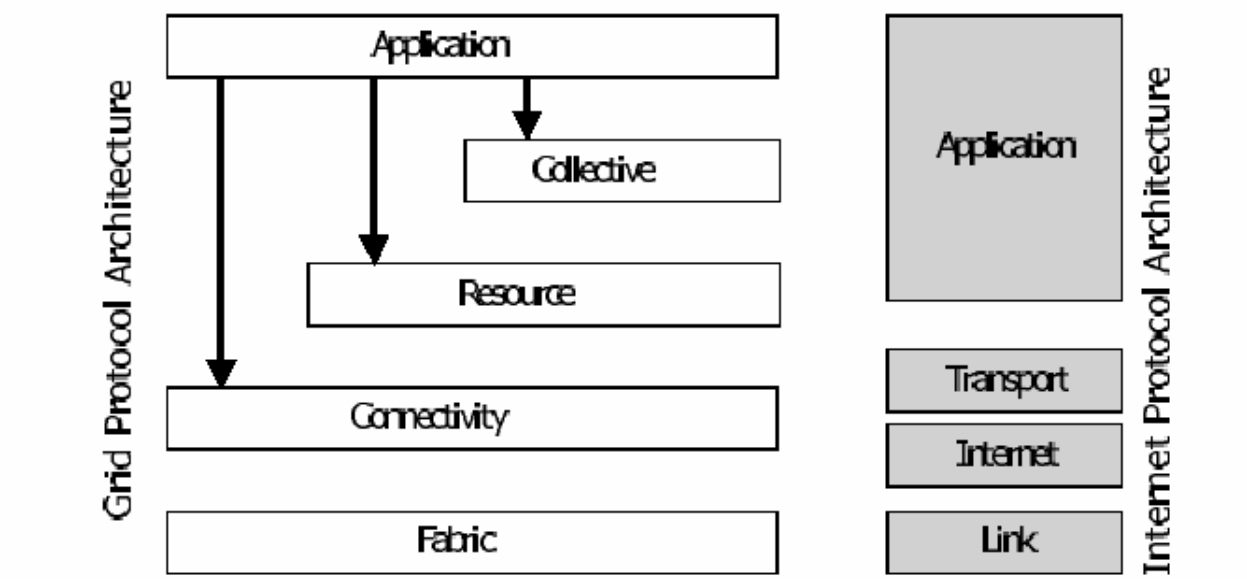


Figure 18: Grid architecture in relationship to the internet protocol architecture. Because the internet protocol architecture extends from network to application, there is a mapping from Grid to internet layers (28).

The Grid Fabric (28) is the interface to 'local' control. Fabric layer provides the resources to which shared access is mediated by Grid protocols, for example, computational resources, storage systems, catalogs, network resources, sensors and other automatic data acquisition sources (for example, local RFID data instances). A "resource" may be a logical entity, such as a distributed file system, computer cluster or distributed computer pool (alternatively one may think manufacturing floor terminals or point-of-sale (POS) terminals). Here, a resource implementation may involve internal protocols (storage access protocol or a cluster resource management system's process management protocol) but these may not be an integral part of the Grid. Fabric components implement the local, resource-specific operations that occur on specific resources (physical or logical) as a result of sharing operations at higher levels. There is, thus, interdependence between the functions implemented at the Fabric level and the sharing operations supported by the Grid.

The Globus Toolkit (37) is designed to use (primarily) existing Fabric components, including vendor-supplied protocols and interfaces. If a vendor does not provide the necessary Fabric-level behaviour, the Globus Toolkit may include missing functionality. For example, enquiry software is provided for discovering structure and state information for various common resource types, such as computers (OS version, hardware configuration, load, scheduler queue status), storage systems (available space) and networks (current and predicted future load). The Globus Toolkit packages this information in a form that facilitates the implementation of higher-level protocols, specifically at the Resource layer. Software to complement the Globus Toolkit may be used in the near future as an interface to link the astronomical number of physical entities (RFID readers, SDR, sensor node) to the Fabric layer to feed real-time data to higher level protocols (including reliable error reporting when operations fail) and applications.

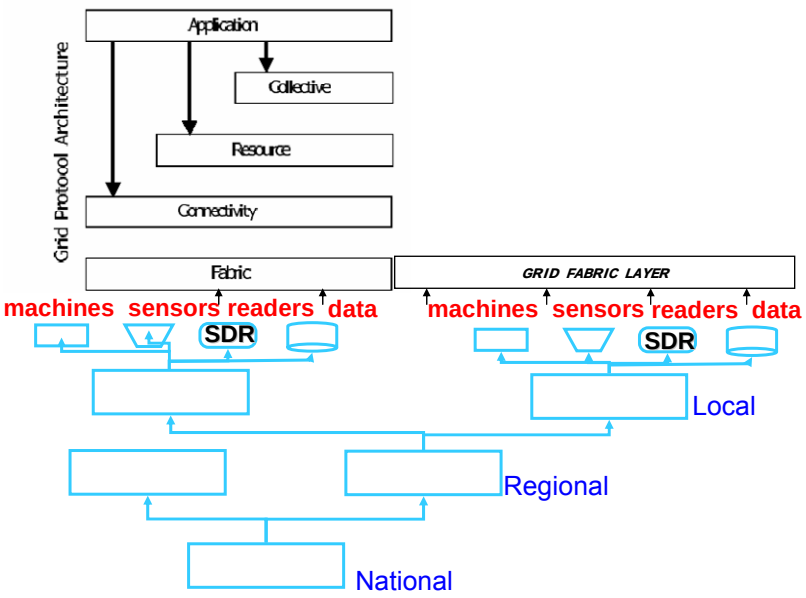


Figure 19: How AIT data may interface with the Grid

Connectivity layer (28) defines communication and authentication protocols required for Grid-specific network transactions. Communication protocols enable exchange of data between Fabric layer resources. Authentication protocols build on communication services and are expected to provide cryptographically secure mechanisms for verifying the identity of users and resources. Communication requirements include transport, routing and naming. Resource layer builds on the Connectivity layer communication and authentication protocols to define protocols (as well as APIs and SDKs) for the secure negotiation, initiation, monitoring, control, accounting and ***payment of sharing operations*** on individual resources. Resource layer implementations of these protocols call Fabric layer functions to access and control local resources (balance of global and local interactions).

The Resource layer (28) is focused on interactions with a single resource, but the next layer contains protocols and services (and APIs and SDKs) that are not associated with any one specific resource but rather are global in nature and capture interactions across collections of resources. For this reason, it is referred to as the Collective layer. Collective components can implement a wide variety of sharing behaviours without placing new requirements on the resources being shared. While Resource layer protocols must be general in nature and are widely deployed, Collective layer protocols span the spectrum from general purpose to highly application or domain specific, with the latter existing perhaps only within specific organizations or value networks to accommodate diverse needs. Collective functions can be implemented as persistent services, with associated protocols or may be designed to be linked with applications (as SDKs with associated APIs). In both cases, their implementation can build on Resource layer (or other Collective layer) protocols and APIs.

The final layer of the Grid comprises the user Applications that operate within a virtual organization environment. Applications are constructed in terms of, and by calling upon, services defined at any layer. At each layer, we have well-defined protocols that provide access to some useful service (resource management, data access, resource discovery). At each layer, APIs may also be defined whose implementation (ideally provided by third-party SDKs) exchange protocol messages with the appropriate service(s) to perform desired actions.

Taken together, this ***open*** architecture offers a plethora of flexible options to maximize the benefits from use of Grid computing. However, the Grid is ***not*** a next-generation internet or an alternative to the internet. It is a set of additional protocols and services that ***build on*** internet protocols and services to support the creation and use of computation- and data-enriched environments. Any resource that is 'on the Grid' is, by definition, on the internet. But, the Grid is ***not*** a source of free computational cycles. Grid computing does not imply unrestricted access to resources although it has the potential to enable such access. Grid computing is about controlled sharing. Resource owners (say, Ford Motor Company) typically want to enforce policies that may constrain access according to group member's ability to pay. Hence, ***accounting*** is important. Therefore, Grid architecture must incorporate resource and collective protocols for exchanging usage and cost information, as well as for exploiting this information when making a decision (preferably automated) whether to enable sharing (policy).

Sharing brings to the forefront the debate whether Grid software should define the operating system (OS) services to be installed on ***every*** participating system to provide for the Grid what an OS provides for a single computer. Microsoft and Sun Microsystems may be strong advocates of the latter perspective since it allows OS vendors (by analogy, Microsoft XP OS) to dominate in the Grid space (market share) by advocating that the Grid requires a distributed operating system where the role of the Grid OS is to define a virtual machine. This perspective is inconsistent with the 'open source' notion of the Grid and its goals for broad deployment and interoperability.

The appropriate (sharing) model may follow the globally well established internet protocol (IP) suite which serves the networked world. The tremendous physical and administrative heterogeneities likely to be encountered in Grid environments mean that the traditional transparencies are unobtainable. Yet, it is feasible, as demonstrated in the past, to obtain agreement on standard protocols (TCP/IP). Thus the Grid architecture must be deliberately open rather than prescriptive. The open Grid may define a compact and minimal set of protocols that a resource must speak (equivalent to TCP/IP for the internet) to be on the Grid, beyond that, it will only provide a framework within which many behaviours can be specified. The open Grid architecture is essential to harness the thousands or even millions of processors that may be accessible within a supply chain, value network or virtual organization. It represents a significant source of computational power that may be used, through proper ***business process innovation***, to sense, re-plan and execute (respond) in near real-time based on real-time data (AIT, RFID), even when re-calculating MIPS-devouring applications based on complex equations, such as ODD-VAR-GARCH.

Much of what needs to be accomplished by the open Grid architecture is referred to as the Grid middleware. We have briefly mentioned (above) the essential and basic functions for resource access and management. We expect that high performance Grid computing, based on real-time data analysis, will enable simulations from several different organizations (de-coupled supply chain stages; see figure 32 on page 67) to exchange data and cooperate in order to undertake a whole system simulation, as is increasingly needed in near real-time, to respond to real, complex challenges (eg: military operations). Commercial Grid usage may spur software demand to design and develop higher level services that can be assembled in minimal time from modular components composed of different software functions so that specific business process related software systems with complex higher level functionality may be rapidly deployed from an array of "plug-and-play" modules (tangible user interfaces).

The Grid and the semantic web, therefore, may spawn a new generation of software vendors. Currently, such services are being approached by leveraging industry efforts in XML based Web Service by integrating Web Services and Grid services. Web services are a set of industry standards being developed and pushed by the major IT houses (IBM, Microsoft, Sun, HP). They provide a standard way to describe and discover, connect and interoperate Web accessible application components.

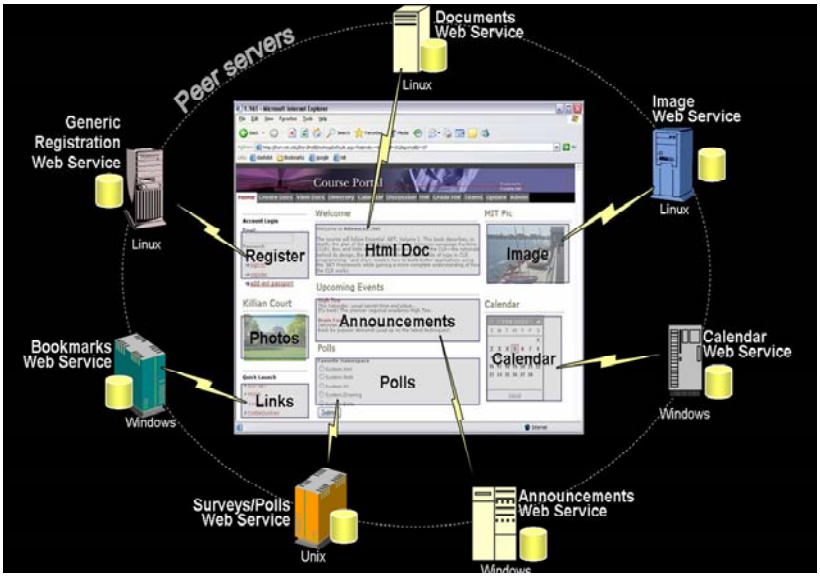


Figure 20: Example of web services in use at MIT

What is missing in current web services, among other things, is the lack of context or meaning of the business process. Hence, efforts such as RosettaNet, to define process related standards (PIPs), are in vogue. For discovery tools to be useful, the web services will require an ontological framework and integration with the Grid to evolve as Semantic Grid Services, or something similar. Yet, at least for now, web services allows the use of commercial and public domain tools (such as web interface builders, problem solving environment framework builders) to start building toward the complex application systems that may eventually provide some of the desired functionality. This Web-Grid integration is under exploration by the Open Grid Services Interface Working Group at the Global Grid Forum (37). Evolution toward 'intelligent' systems offering complex functionality may require combination of Agent based models within Semantic Grid Services. An idea of the likely confluence is illustrated below.

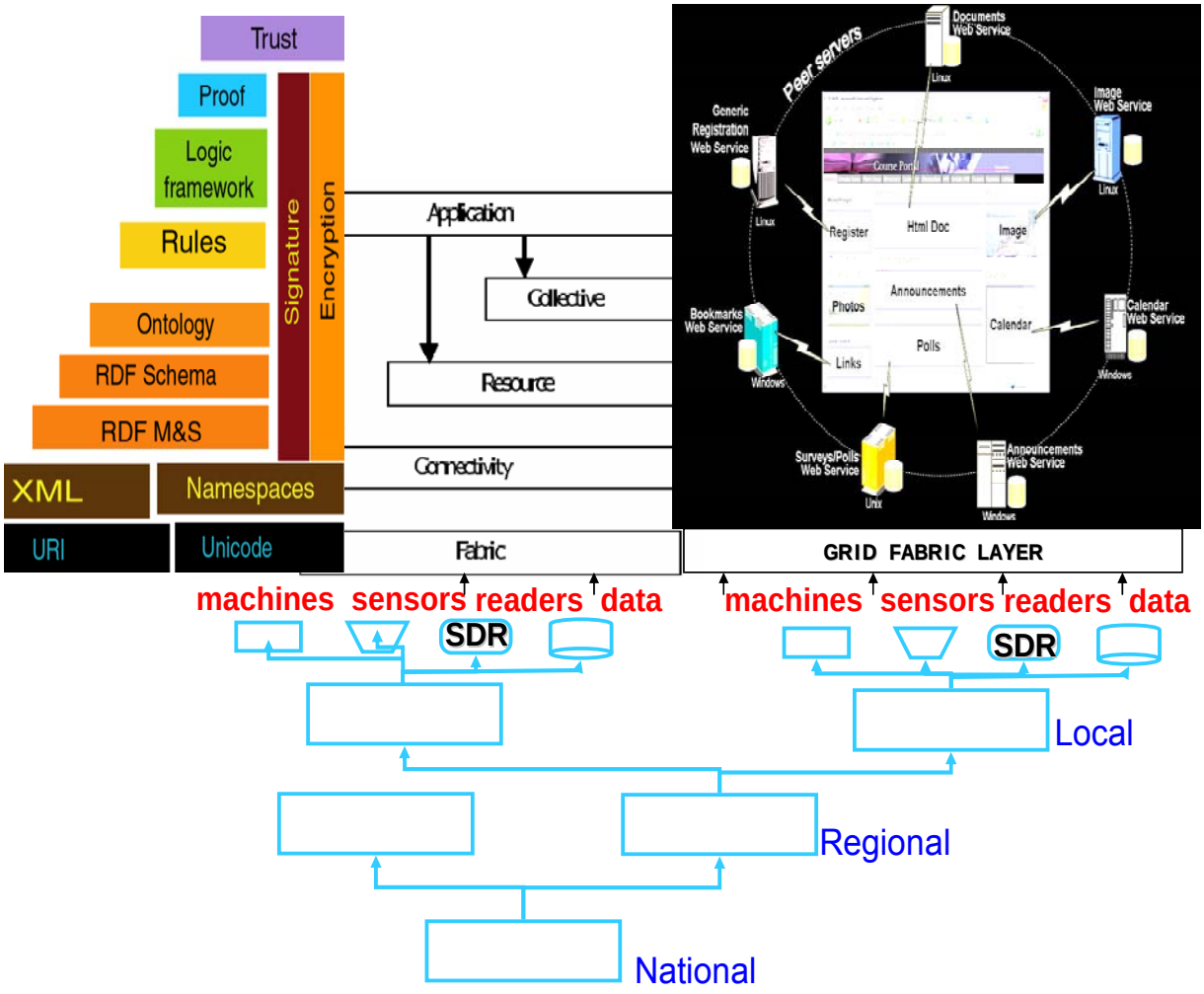


Figure 21: Semantic Grid Web Services: Confluence of Semantic Web layers with the Grid, Web Service Portals and Data (object-dependent data from AIT-RFID, sensor data, device-linked data from machines, appliances, hospitals)

AGENTS

We shall explore what are Agents and their importance in decision systems. The concept of Agents germinated at MIT and evolved from the study of how the human mind works and what is intelligence (38). The roots of Agents, therefore, are bio-inspired and embedded in the Artificial Intelligence (AI) movement that gained recognition around 1950's. Some of the earliest Agent concepts were used to create difference engines, perhaps the earliest logical precursor unit of the present day neural network (software).

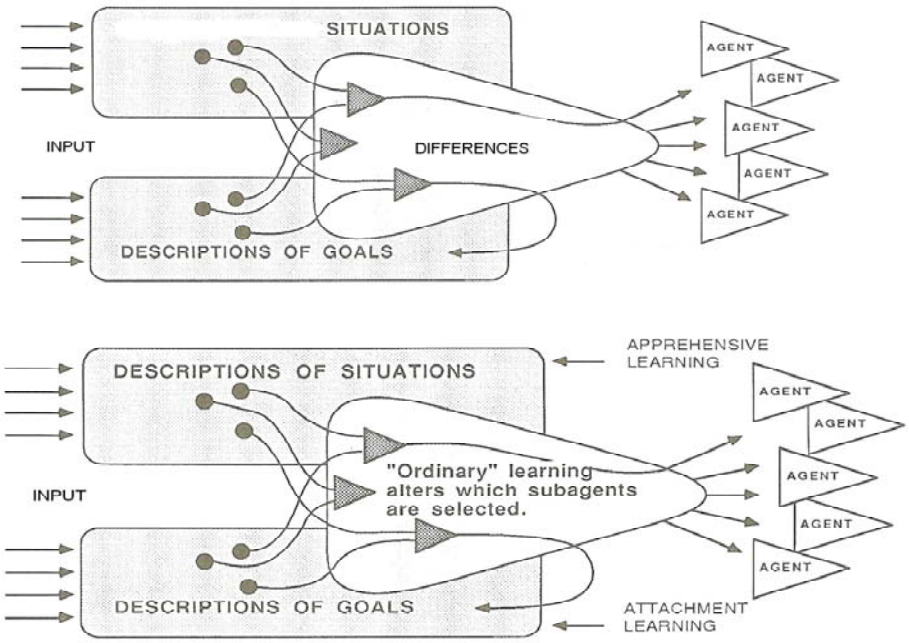


Figure 22: Difference Engine (from 38)

It is a slightly arresting notion that if you were to pick yourself apart with tweezers, one atom at a time, you would produce a mound of fine atomic dust, none of which had ever been alive but all of which had once been you. Yet somehow for the period of your existence they will answer to a single rigid impulse: to keep you, you. Why atoms take this trouble is a bit of a puzzle. Being you is not a gratifying experience at the atomic level. For all their devoted attention, your atoms don't actually care about you – indeed, don't even know that you are there. They don't even know that they are there. They are mindless particles, after all, and not even themselves alive.

Bill Bryson
A Short History of Nearly Everything

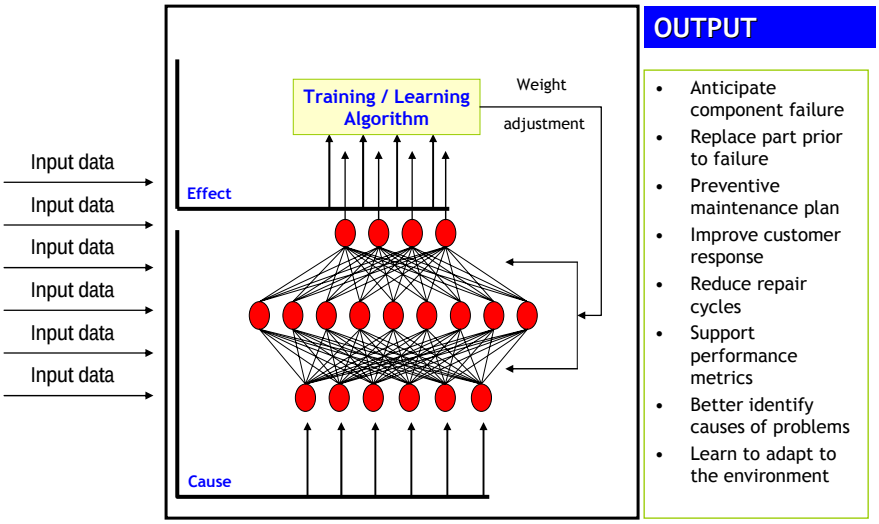


Figure 23: Basic Principle of Neural Network

The 'intelligence' aspect of Agents, like atoms, is a conceptual paradox. Yet, this paradox is vital and empowers Agents to make superior decisions by drawing from widely distributed and divergent sources of information (input, goal, cause, effect). Businesses eager to exploit the power of Agent-based systems may wish to understand that one Agent by itself is neither critical nor intelligent. According to Marvin Minsky, intelligence is a combination of simpler things. Thus, an Agent is not 'intelligent' but 'Agencies' are intelligent. We shall refer to the concept of Agencies in its simplified form as Multi-Agent Systems but may casually use Agents to imply Agencies, as well.

By analogy, another deceptive pair of words that conceals countless different skills is common sense. It may seem obvious and natural but common sense is intricate (and not as 'common' as it is made out to be). It is an immense 'society' of hard-earned practical ideas, multitudes of life-learned rules, exceptions, dispositions, tendencies, checks and balances (38). When humans face situations where one must respond with 'common sense' the mind (consciousness) integrates views from a 'sub-society' of simpler elements, draws upon learned rules or exceptions or tendencies. Taken together, it generates a response that, then, may seem obvious and natural. Similarly, when an Agent has to execute a complicated task, it will morph into an 'Agency' collectively drawing upon a sub-society of Agents performing simpler tasks. As an ***Agency*** the system will ***know*** its job and offer superior performance but individual Agents may not know about the job. This 'distributed' aspect of Agents, functioning collectively as a 'swarm' infuses the Agents (Multi-Agent Systems or Agencies) with superior dynamic as well as learning-dependent adaptive capabilities than static equation-based algorithms used in almost all current commercial software.

Most current planning or decision support systems software, such as enterprise resource planning (ERP) or supply chain management software is based on equation-based models (EBM) that link rates and flows (consumption, production). Variables (cost, rebates, transportation time, out-of-stock) evaluate or integrate sets of ordinary differential equations (ODE) or partial differential equations (PDE) relating the variables. Operations research provides the framework to optimize for the 'best' result but what if 'best' is not the optimal for that situation? Shortest lead time could plan a route through an area with a high probability of flash flood and perhaps a storm is predicted. Planning software is incapable of modeling such events by integrating multiple information sources. This inability generates inefficiencies because the solution may not be adaptive to supply chain events, at hand. It follows, therefore, linearization of real world conditions to fit mathematical models, such as EBM, classical linear regression model (CLRM) or game theoretic models, may disregard the dynamic changes and stifle real-time adaptability. The discrete, dynamic and distributed nature of data and applications require that solutions do not merely respond to requests for data or information but anticipate, adapt and (support users to) predict. Agents, collectively functioning as Agencies, may be better suited to help anticipate, adapt or predict.

However, before we can plan Agent-based simulation of future scenarios (for example, truck route and path of storm plus road conditions after rainfall), it is imperative that independent models (scenarios) can interact. Currently such simulation may not be possible because models (weather, highway, road construction) cannot 'talk' to each other (we may include a brief discussion of simulation in a future version of this article).

Excluding random events or decisions requiring integration with other models, what is the half-life of an optimized solution in a high 'clockspeed' industry (58) or fickle economy? Shortcomings of equation-based (ODE, PDE, CLRM) models include assumptions that parameters are linear, accurate relevant data are available (for optimization) and error terms are homoskedastic. In the real world, events are non-linear, actions are discrete, information about data is difficult to obtain (CRM, PLM, SCM data silos) and data is often corrupted with "noise" to a significant extent. According to a study, of a global retailer, 65% of barcoded SKU's were inaccurately represented (23).

Automatic identification technologies may help decrease such inaccuracies, in part stemming from human errors (scanning barcodes). In addition, 'intelligent' autonomous Agencies (Agents) potentially could become an essential tool to interface with multiple sources of data (including automatic identification data, say, from RFID tags affixed to objects) to extract information and feed processes or parameters necessary for informed and dynamic decision making. Our ability to transform this vision into reality is one key element of confluence necessary for adaptive value networks to emerge.

For practical purposes, for now, we will steer away from the richness of the AI complexities that created the concepts of Agents and Agencies based on how the mind works and connectivity due to neural trajectories. For our discussion at hand, we will refer to Parunak (42a, 42b) and choose to define an autonomous Agent as a software entity that functions continuously in an environment, often inhabited by other Agents. Continuity and autonomy empowers Agents to (plan) execute processes in response to changes in 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 (Agencies) in a Multi-Agent System (MAS) are able to communicate, cooperate and are likely to be mobile between environments.

Agents work best for clearly discernible tasks or processes, such as, to monitor data from, for example, RFID tags, ultrawideband (UWB) transponders, global positioning system, WiFi and sensors (temperature, light, vibration, sound, acceleration). Data Agents can share acquired data with the next higher level (Agent hierarchy) Information Agents, that may offer real-time information to Process Agents (Inventory Agent, Purchasing Agent). Thus multi-Agent environments (Agencies) can monitor data, update information and instruct processes to perform. MAS are capable of more complex operations and its decisions may result in an 'intelligent' output.

Task-specific deployment of Agents by corporations (GM, Southwest Airlines, Deutsche Post) are increasing as are claims by companies (SAP, CA) touting Agents based software. Although hyped up claims regarding off-the-shelf Agents software may reach a feverish pitch and parallel that of RFID, in the near future, it may be prudent for corporations to seek guidance from informed sources should their plans call for adaptive supply chain management through Agent integration. We will later see how distributed cognition may impact future Agents.

Although preparations for Agent integration should commence for medium to large businesses, the emergence of multi-Agent systems may be slow to materialize unless the semantic web (39) increases its rate of diffusion. Agent Modeling Languages and Web Ontology Languages may enable Agents to sense, understand and sufficiently interact across distributed systems, such as the Grid. Then, we shall migrate from incremental or geometric gains toward that elusive 'quantum' gain in productivity that distributed artificial intelligence, in some forms, promised, nearly half a century ago. In reality, despite the use of computers for nearly half a century, we have not realized such mythical gains in productivity because human input is not machine-readable. Hence 'commands' instruct programs how and what to perform based on yet another level of human specification. When programs (computers) are able to understand human input (words, meanings and context of use or 'semantics'), then the ability for computers to actually offer pro-active help (machine-learning) without human intervention will be possible for basic tasks (invoicing, payment of bills, preventive maintenance). The emerging semantic web is a step in that direction. It may contribute toward such scenarios where computers actually make us exponentially productive (40, 41). In a forthcoming article, we plan to explore (demystify) the semantic web and suggest how it may aid the usefulness of web portals and web services.

To appreciate why Agent integrated decision systems may be worthwhile, it may be useful to understand that design of Agent-based modeling (ABM) draws clues from natural behavior of biological communities (42). According to Parunak, although it still remains a paradox, it is increasingly undeniable that simple individual behaviours of insects (ants, wasps), collectively (in ***swarms***) offer 'intelligent' models of complicated overall behavior. Adaptability in biological systems is an evolutionary necessity. Models based on and inspired by such superior systems can contribute significantly to reduce inefficiencies that plague supply chains and value networks.

Virtually all computer-based modeling, up to this point has used system dynamics, which is an equation-based approach (EBM). But the struggle to adapt and respond in real-time may fuel a paradigm shift that will make it imperative to model business software based both with Agents and equations (ABM and EBM). 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 that can offer combinatorial solutions. The balance itself is subject to dynamic change. For traditionalists in supply chain management, the situation is analogous to a "push-pull" strategy where the dynamic push-pull boundary must shift with changing demand (pull).

ABM and EBM, both simulate systems by constructing models 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 ***behaviours*** of the various individuals that make up the system and execution consists of *emulating* these behaviours, which are essentially dynamic. In EBM, the model is a set of ***equations*** (pre-determined, static) and execution consists of *evaluating* them. Although "simulation" in a generic sense applies to both methods, in this case they should be distinguished as Agent-based emulation versus equation-based evaluation (42a, 42b).

The success of Agent-based systems depends on the continuity and autonomy of biology where behaviour patterns must be flexible as well as adaptive in order to respond to change. Learning how to respond effectively is key to survival. Examination of naturally occurring Agent-based systems (ant colonies) suggests design principles for Agents. Thus, bio-inspired, autonomous, self-learning, adaptive, mobile, networked Agents are likely to use very different principles when dealing with cause (input, data) and effect (output, response) compared to equation-based programs in monolithic software systems. While some circumstances may warrant deliberate exceptions, in general, practical and useful Agents must be aligned with these concepts from Parunak (42a, 42b):

- Agents should correspond to "things" in the problem domain rather than abstract functions
- Agents should be small in mass, time (able to forget) and scope (avoid global knowledge action)
- Agents should be neither homogeneous nor incompatible but diverse
- Agent communities should include a dissipative mechanism (entropy leak)
- Agents should have ways of caching and sharing what they learn about their environment
- Agents should plan and execute concurrently rather than sequentially
- Multi-Agent Systems should be decentralized (no single point of control/failure for an Agency)

Agents versus Equations

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 behaviours, thus, maintains boundaries among individuals (42a, 42b). This fundamental difference in model structure gives ABM a key advantage in security of commercial applications such as an adaptable value network where partners may interact over an e-marketplace or use semantic Grid web services, in the future.

First, in an ABM, each firm has its own set of Agents. An Agent's internal behaviours are not required to be visible to the rest of the system, so firms can maintain proprietary information about their internal operations (42a, 42b). 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 auto-regulates the behaviour of the entire system (42a, 42b). Agent systems may correspond one-to-one with individuals (firms, divisions) in the system being modeled and the behaviours are analogs of real behaviours. These characteristics make Agents a natural locus for the application of adaptive techniques that can modify their behaviours as the Agents execute, so as to control the emergent behavior of the system (Agency). In other words, a complex multi-stage supply chain can be decoupled at the Agent level, enabling modification of individual Agents to reflect stage-specific changes or local optimization. Since these modified Agents are a part of a higher Agency, by virtue of the 'emergent behaviour' that is characteristic of Agent systems, when viewed as a whole, the complex supply chain will reflect global optimization by integrating the stage-specific modifications or local optimization.

Migration from simulation model to adaptive control model is more straightforward in ABM than in EBM. One can imagine a member of the value network using its simulation Agent as the basis for an automated control Agent that handles routine interactions with trading partners. It is unlikely that a firm may 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 behaviours, which may include features difficult to represent as rates and levels, such as step-by-step processes and conditional decisions. EBMs are well-suited to represent physical processes. However, business processes are dominated by non-linear, discrete, decision-making (42a, 42b).

Both ABMs and EBMs can be validated at the system level by comparing model output with real system behavior. In addition, ABMs can be validated at the individual level, since the behaviours encoded for each Agent can be compared with local observations on the actual behaviour of the domain (42a, 42b). ABMs support direct experimentation. Managers playing what-if games can think directly in terms of business process, rather than translate them into equations relating observables (VAR-GARCH). One purpose of what-if games is to identify best practices. If a model is expressed and modified in terms of behaviours, implementation of its recommendations is a matter of transcribing the modified behaviours of Agents into tasks for the 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. For example, in principle, PDE's are computationally complete. Hence, one can construct a set of PDE's 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. 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 behaviours 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 often used in traditional supply chain simulation.

Disadvantages of EBM in this and other examples result largely from the use of assumptions, averages of critical system variables over time and space. EBM also assumes homogeneity among individuals but individuals in real systems are often highly heterogeneous. When the dynamics are non-linear, local variations from the averages (errors) can lead to significant deviations in overall system behavior (outcome), such as, the Bullwhip Effect. Because ABMs are inherently local and adapt to changes, it is beneficial to let each Agent monitor the value of system variables without averaging over time and space, as we have discussed in the section on ODD-VAR-GARCH.

Ant-based algorithms may form a core of some Agent architectures (42a, 42b, 59). Such 'swarm intelligence' 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 the ants). 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.

Typical EBM based demand forecasting tools generally use a weighted-average of historical 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 subsequent forecast for weeks or months. Consider the cumulative error from such events for a global forecast that guides procurement or production. Such errors contribute to the Bullwhip Effect. Agents can improve forecasting and accurate real-time data with tools such as ODD-VAR-GARCH to enhance their precision (see page 69, last paragraph). Armed with precision, a manufacturer may adjust production to better manage inventory and reduce waste. Reduced inventory decreases working capital charges, which, in turn, improves return on assets (shorter cash cycle).

Generally, forecast determines production planning and subsequently the plan is executed. However, planning and execution are not concurrent. Some manufacturers develop a schedule each night that optimizes resource utilization for the next day, a process not much different from grocery chains that may order perishables the day before. Engineers in industries as diverse as auto, semiconductors, aerospace and agricultural equipment will agree that a daily schedule may be obsolete less than an hour after the day begins. 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 (42a, 42b). 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 (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 order in which jobs are executed is decided at the moment the resource becomes available, depending on the circumstances that exist, at that time.

Concurrent planning and execution, targeted (micro) changes within an environment and several other rapid response abilities demand that systems can compute several parameters and adjust weights. Thus, Agents in a system architecture must have connectivity (and the environment) to distributed and diverse systems. The myriad of software systems and data or operational silos even in smart companies fails to offer the required visibility of goods, products, inventory and customer service levels across various divisions of a company. The latter became painfully obvious to the author in a recent transaction with Dell Corporation, an organization considered to be almost at par with Amazon.com in its supply chain management practices. Connectivity is an aspect that can be most readily addressed by the very nature of the Agent infrastructure in Agencies (Multi-Agent Systems). Figure 24 (below) shows a simple view of how Agents may be connected to various distributed data sources and systems (43). Within the Grid architecture, Agents will play an increasingly important role.

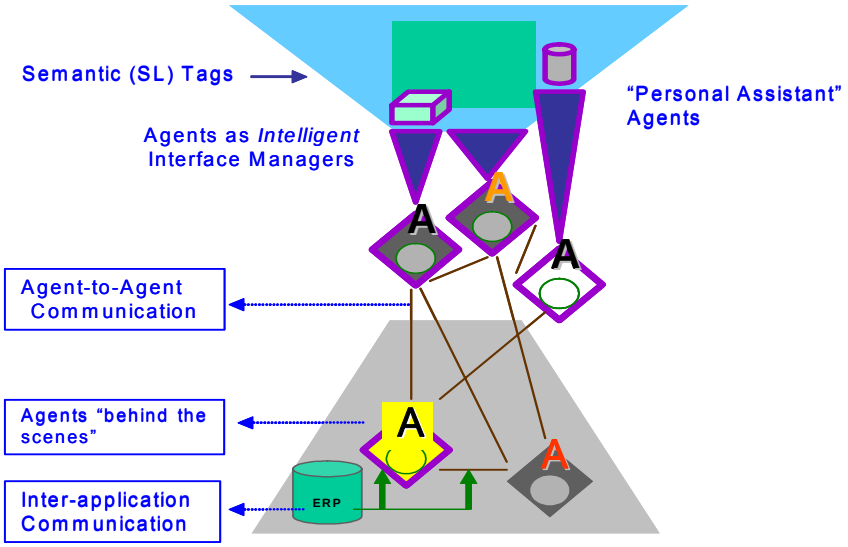


Figure 24: Agents in a Simple System Architecture (43)

The vision of integrating Agents and Agencies in systems architecture is far more profound than shown in this 'practical' illustration (Figure 24). The strength of the claim that Agents can offer unprecedented levels of connectivity is based, in part, on exploiting the principles extracted from our rudimentary understanding as to how neurons connect to various regions in the brain. Heredity and environment shapes the neural tracks that enable us, for example, to respond with 'common sense' or associate gestures and emotions (love, joy, anger) with trajectories (smile, scowl, slap, snarl, sigh, caress) that are captured by sensors (tactile, auditory, visual). These *n to m* connections offers us the ability to respond with accuracy and finesse (expression of sympathy versus empathy) that demand many connectivities to memory (sources of information, past experience) in a limited time or rapid succession (processing).

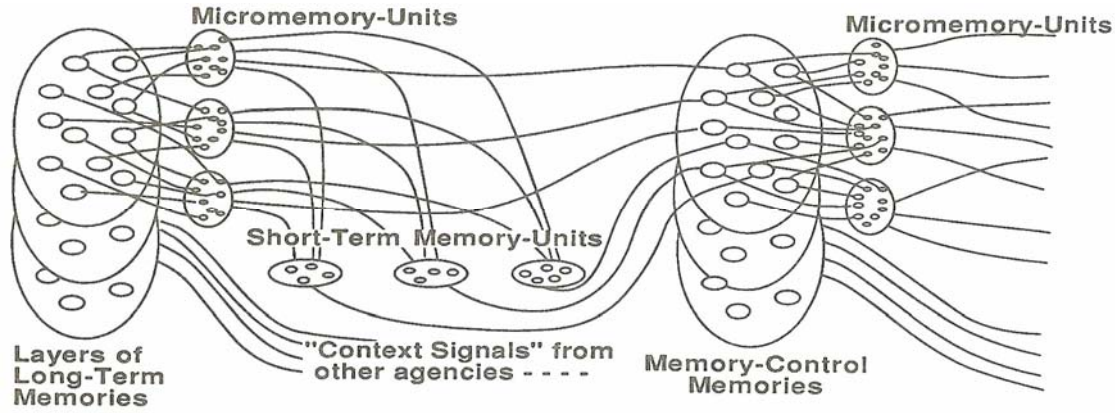


Figure 25: Basic Neural Circuitry (38)

Marvin Minsky (38) provides an idealized representation (below) based on such neural pathways. This illustrates the profound connectivity that can be achieved by Multi-Agent Systems (Agencies) to enable greater systems integration even when dealing with vast number of variables (dependents) or innumerable data or information sources, no matter how complex the organization, supply chain, value network or decision system, may be.

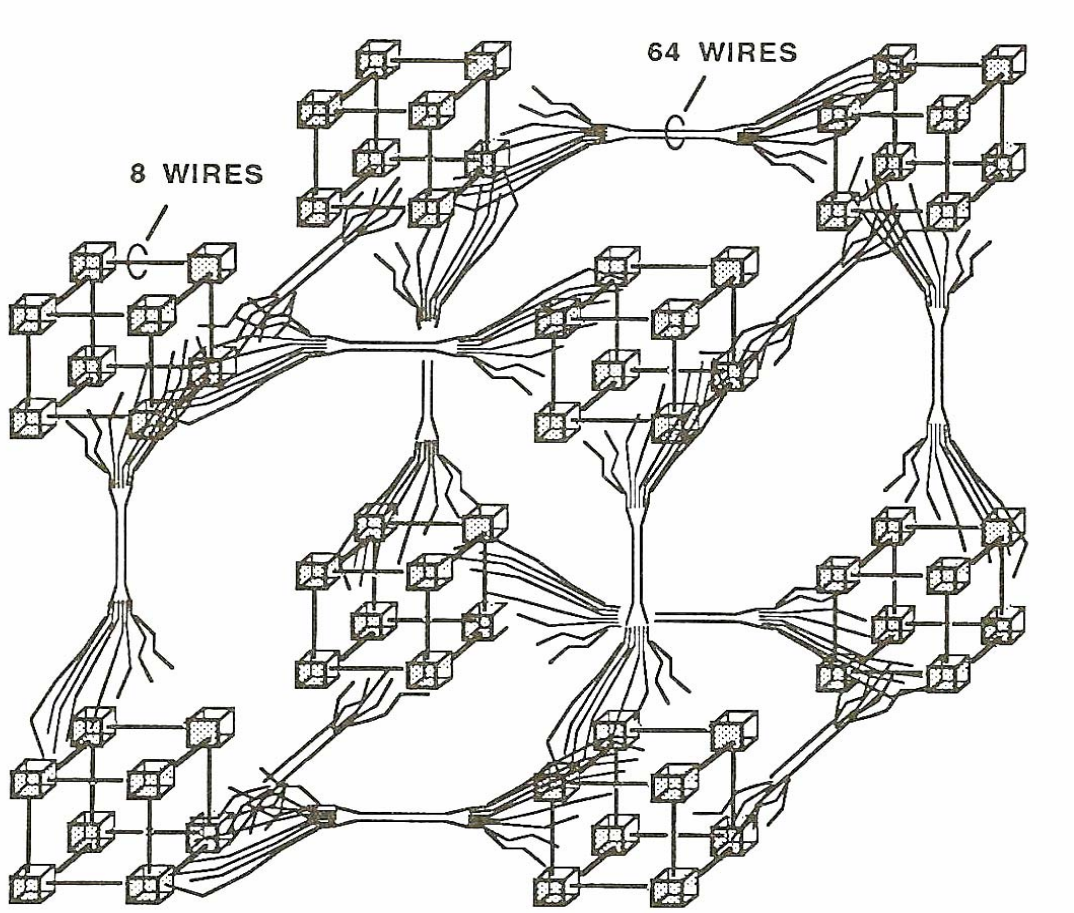


Figure 26: Cube-on-Cube (38)

Imagine that the smallest cube (above) is made of 8 Agents (8 corners) connected to each other. 8 such cubes form the corners of another cube, making a 64-Agent cube (8x8). If this 64-Agent cube joins to form the cube above, it will make a 512-Agent cube (8x8x8). If we repeat this cube-on-cube pattern 10 times (10 steps), the supercube (8^{10} = 1, 073,741,824) will contain over 1 billion Agents. Each Agent in the original smallest cube (of 8 Agents) could communicate with 1 billion Agents (sources, variables) in 10 steps. If we link each Agent to 50 other Agents, then each Agent could communicate with >15 billion Agents in 6 steps (50^6 = 15,625,000,000). In other words, CocaCola can monitor ***nearly all RFID tagged unit cases*** of its product and real-time data can be collected by an Agent (Agency) in 6 steps for analysis (inventory, distribution, storage, transit, temperature) and optimized decision. In 2004, CocaCola produced 19.8 billion unit cases (each unit case =192 ounces).

Agents in Maintenance

A soft-drink manufacturer may argue that use of Agents in its architecture and infrastructure is an excessive investment given that the operation may have a few mission critical variables. In addition, case-specific granularity of data (for 19.8 billion unit cases) is simply unnecessary for effective decision making. The latter may be partially true for a bottling operation (eg: Pepsi Bottling Group) with only a few SKUs. However, in the example below, over a period of five years, certain aircraft maintenance operations reveal that failure to repair aircrafts or engines (to certify as 'capable' to serve) is due to lack of spares. It is especially interesting to note that about 80% of the unavailable requisitioned parts cost less than $100 (an amount that appears to be insignificant given the scope and importance of this operation).

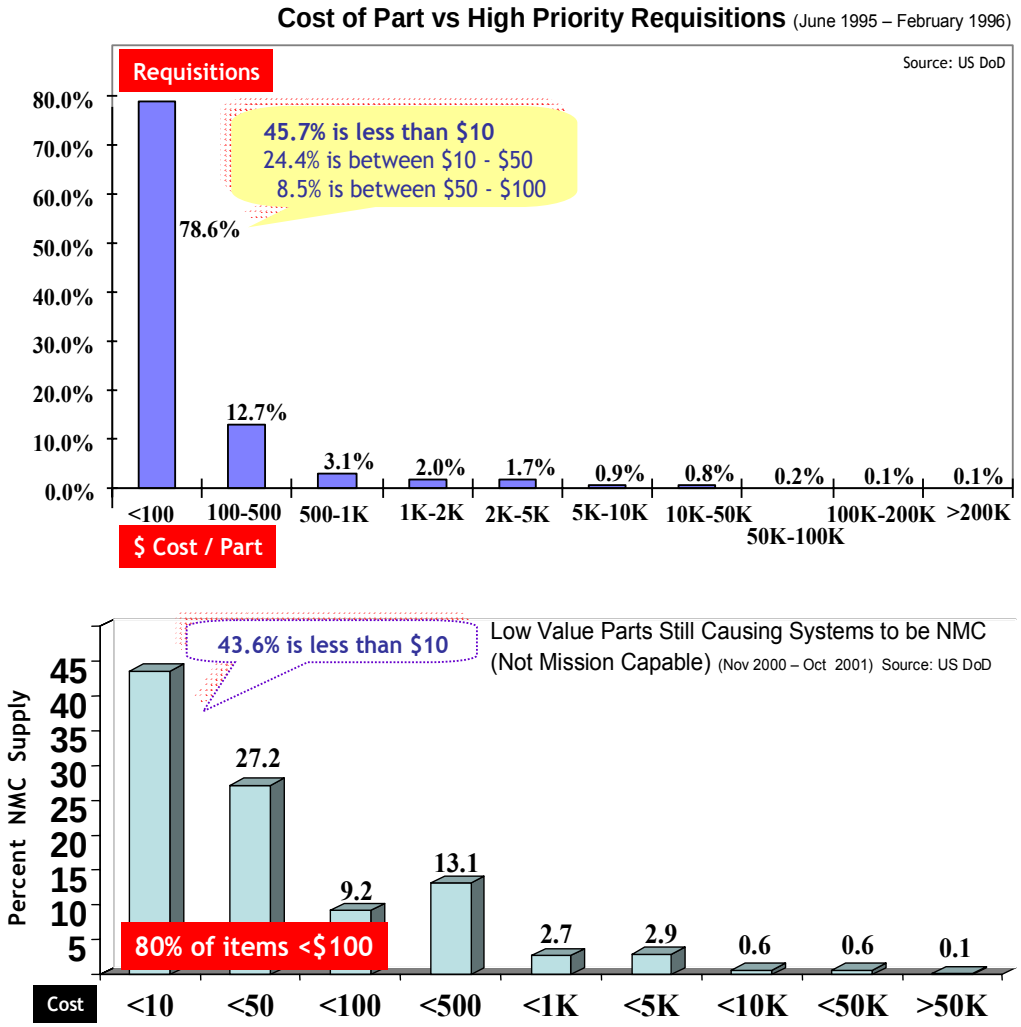


Figure 27: About 80% of the out-of-stock (OOS) spare parts cost less than $100 per piece.

Assuming that there are no deliberate bureaucratic obstacles, one quick solution may be to create a workflow that automatically orders parts that cost less than $100. But is that an effective solution? The ordering process via a workflow does not 'anticipate' need for spare parts, it simply executes in response to human intervention (requisition process). The fulfillment process needs monitoring and EDI links, often proposed, are not an effective solution. Are the systems integrated to update the requisition request with available data from the supplier? Are purchase orders, invoices and bills payable synchronized with advance shipping notice, receiving and quality control? Is the work order or job scheduling for the maintenance engineer updated online to reflect the priority, delivery status and availability of the spare parts requisitioned?

Figure 27 suggests that perhaps one solution, to mitigate the out-of-stock (low value) spare parts situation, is to order more parts so that the availability of common parts were higher and the 'percent' of inventory was increased to improve KPI. In reality, the goal (mission capable status) was left unmet due to a shortage of unique parts. Ordering or receiving of spares is divorced from its association with the task. While the engineers on the ground fail to accomplish the task due to shortage of necessary unique spare parts, the system reflects improved metrics due to increase in ***average*** availability (bolstered by the inventory of common parts).

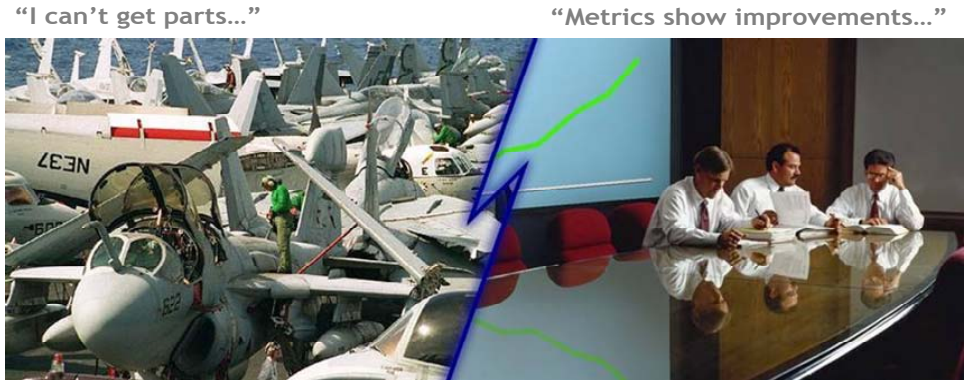


Figure 28: Disconnect

Agent-based modeling of repair or maintenance (MRO) may help integrate demand from multiple stages, each with its own series of variables. Preventive maintenance schemes using ABM can anticipate parts failure from monitoring metrics such as mean time between failures (MTBF). If equipment failure necessitates a repair, the location of the failed equipment may communicate with the repair operation. Agent based updates from field operations, if connected to MRO to schedule jobs, can match requisition of appropriate spares with resources (tools, engineers) to accomplish 'mission capable' status. It may be possible with Agents integrated systems to plan or project equipment failure and the resource planning necessary to respond in crisis (note k on page 99).

Agents in Manufacturing

Commercial aerospace industry makes fewer products and sells to a different set of customers than the retail industry (Figure 29 shows a typical aerospace supply chain). Some (modular) parts and components are shared between different models (variants). 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 in a reasonable time. Agents could accomplish such tasks rapidly and may predict, thereby avert a potential catastrophe. The information required for Agents 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 through the Grid and to make sense of the data through the use of semantic web.

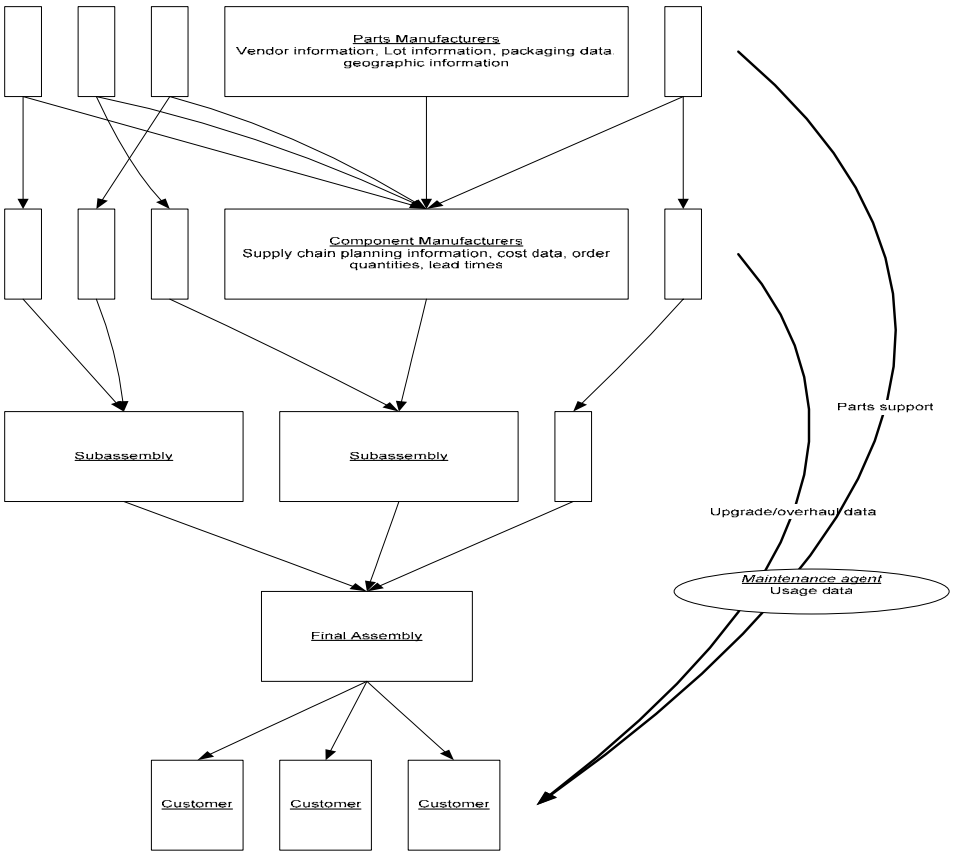


Figure 29: Commercial Aerospace Industry Supply Chain: Information Collection

Future Agents at Work?

Transistor Titicaca Promethium (TTP), a small 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 revenue.

You are Must-See-Borgium Corporation, the bleeding-edge retailing behemoth. You start selling CELC and soon your return center in Moose Jaw is flooded with CELC from unhappy customers. Fortunately, your ex-VP (self-exiled to Myanmar) 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 alert (cc you) to N E She in Urawa (manufacturer's headquarter) indicating 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 She-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 (CELC website will soon upload the link). 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 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 local 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?

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. 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 autonomously collected product, customer and service data. Customer purchases were compared for people who bought and returned this new product. 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 (non-obvious) associations were easier to track by multi-Agent systems monitoring multiple operations both in the company and its interactions with partners. 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 manufacturer's products has risen above a certain level in recent weeks. Why? It is 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 happens to inspect the packaging. Voila! It is different than the packaging for products that have a lower return rate. A phone call confirms that the manufacturer recently switched to a different packaging vendor.

The Agent succeeded in creating the alert because the Agent system collects, processes, correlates and cross-references vendor data, shipping method, shipping distance and other cradle-to-grave stage-specific data it can extract from the local data store connected with goods movement. 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 tags for each SKU sold, the Agent system may have been able to spot the trend without human intervention (see Figure 30).

Agents can help with marketing. Dell allows consumers to configure their computers. 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 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 Agents can be used by 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 the bundle discount option.

The number of potential product combinations increases if three or more options are thrown into the mix, not to mention accessories (cameras, MP3 players, printers). It is simpler for Agents to analyze gargantuan amounts of data and spot potential (multiple) bundling opportunities as well as adapt to demand fluctuations in near real-time much faster (orders of magnitude) than a human or software based on equations. Bundling improves sale of slow moving inventory or near end-of-life products prior to introduction of new versions or products.

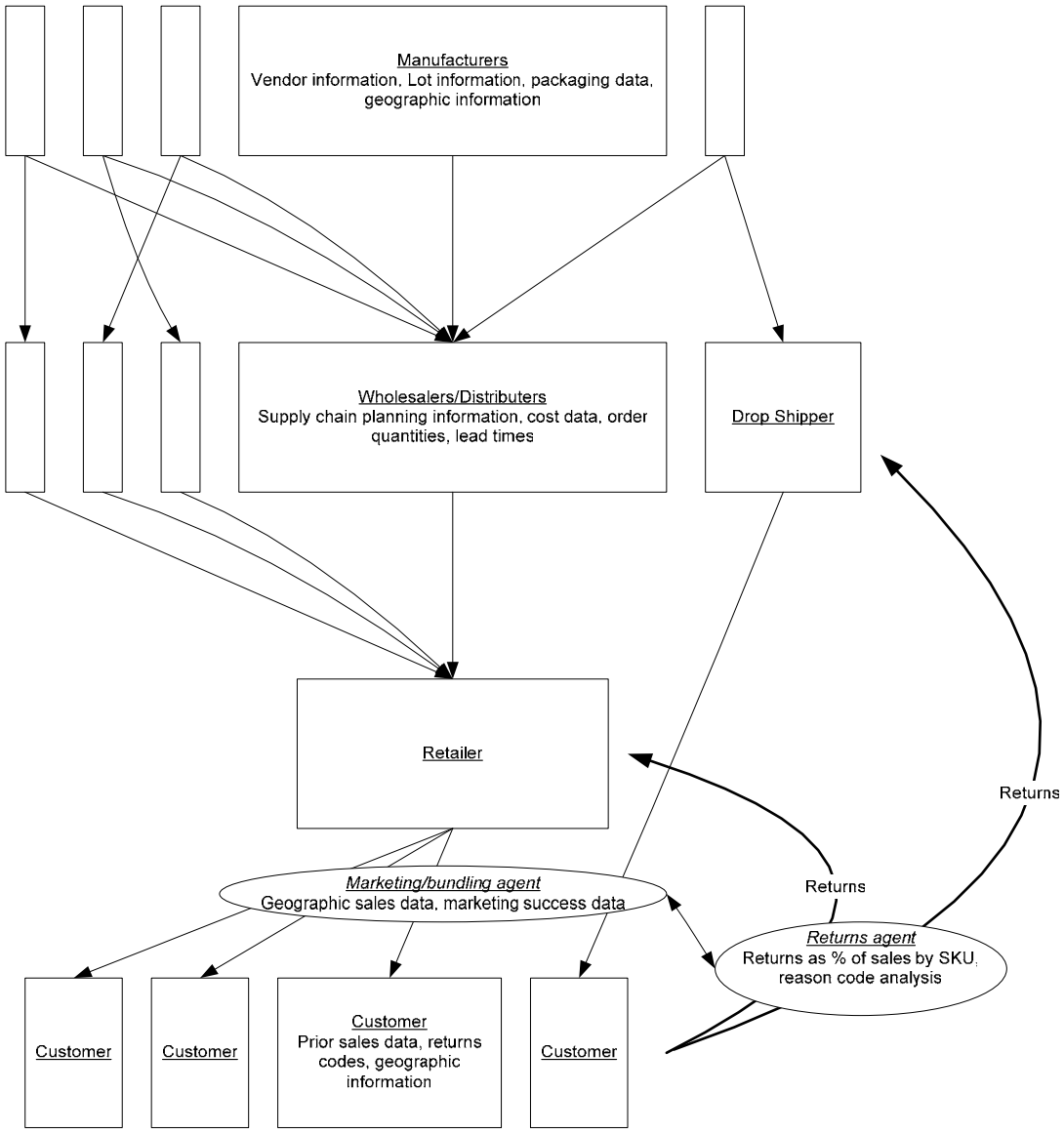


Figure 30: Agents in Retail Industry (shows where "returns" and "bundling" Agents may integrate)

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 (42a, 42b). 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. In a very 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 be as efficient. The appropriate comparison for systems designers of enterprise software is not between local and global optima but between static versus adaptable systems. Let us evaluate the competing options (in any particular case) theoretically, strategically, tactically and practically (42a, 42b).

Theoretically, there are decentralized mechanisms that can achieve global coordination. Economists have long studied how local decisions can yield globally reasonable effects. Recently these insights have been applied to domains beyond economics, 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 ERP. 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 (44). 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 31).

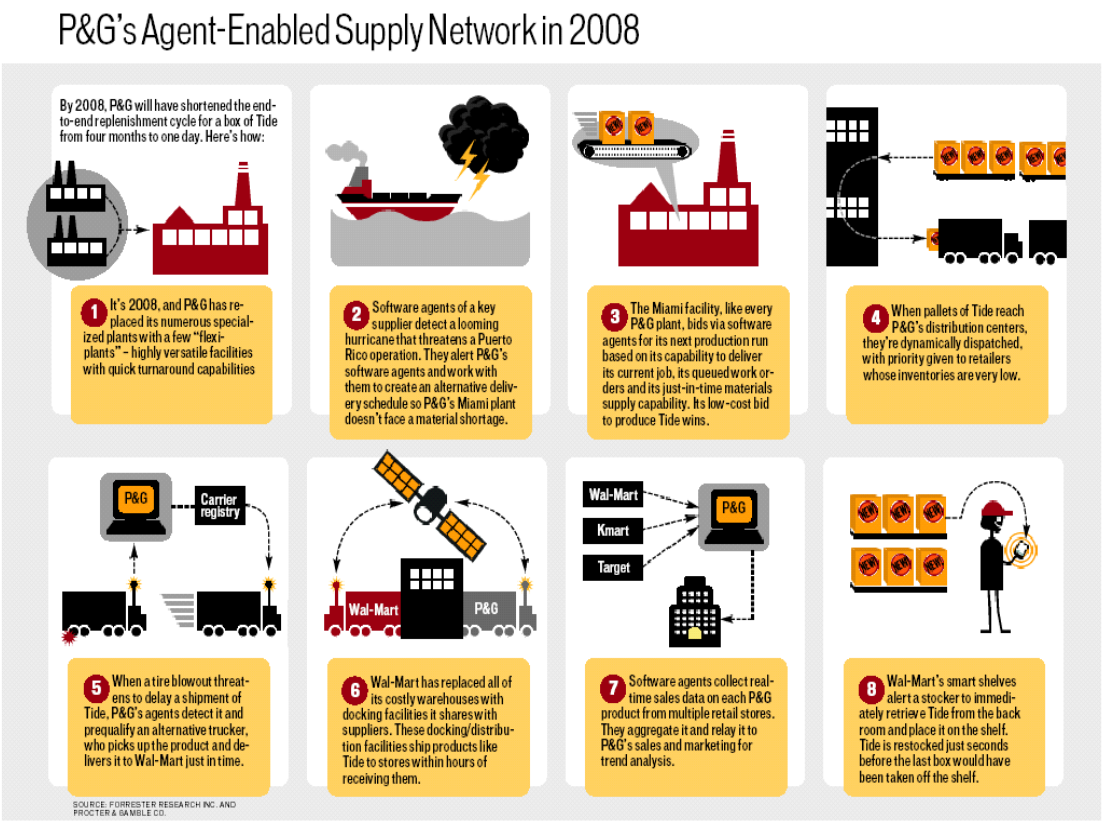


Figure 31: Agent-enabled Supply Network proposed by P&G

The Agents-based scenario illustrated above, hypothetically, treats various stages of this multi-stage supply network as independent Agent models that communicates and interacts with other stages to optimize or adapt. At the granular level, a similar argument can be used for ODD-VAR-GARCH to suggest how equation-based ODD-VAR-GARCH may be transformed to an Agent integrated model (Figure 32). Using the CLRM construct (equation 1; page 23), the explanatory variables (x_1, x_2, …, x_k) may represent inventory, price or expiration related Agents that feed the equation with necessary data for inventory, price and expiration. The ability to decouple a multi-stage supply chain using ABM makes EBM dynamic and responsive to changes. The EBM+ABM approach may rapidly accommodate changing business models if partners in a value network drop in or out (Figure 33). Adaptability may also demand changes or reshuffling of parameters – changing the entire equation (EBM) may be problematic. The latter is not necessary if Agents act as independent entities (inventory agent, price agent, expiration agent) in an EBM framework. For example, if price is not a consideration, in a specific case, it can be excluded or if a new parameter (that was not a part of the model) is now important, it can be included (for collaborative efforts). Thus, 'cross-docking' of Agents and variables may hold the potential to rapidly provide analytical support even in volatile business scenarios in order to better adapt to demand fluctuations. Progress and convergence of research from related fields in artificial intelligence will segue to systems that effectively combine EBM and ABM to stem the impact from uncertainty.

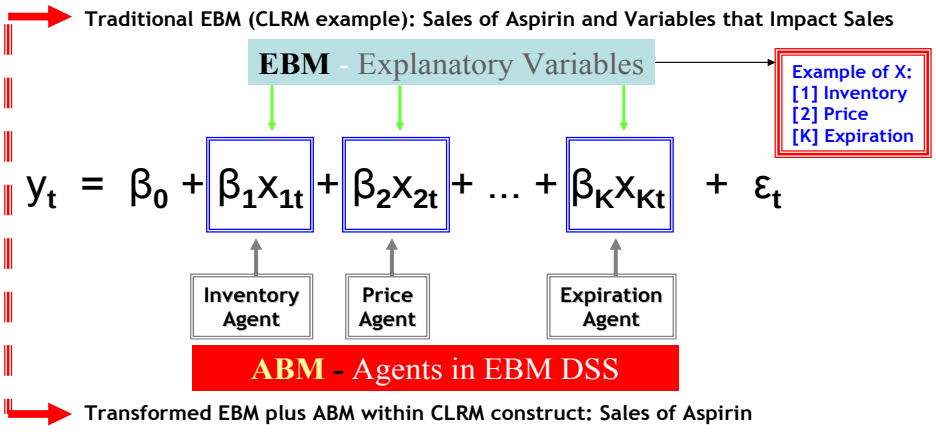


Figure 32: Decoupling Supply Chains: Can it Integrate Local Changes in Global Optimization ? (note m, page 96)

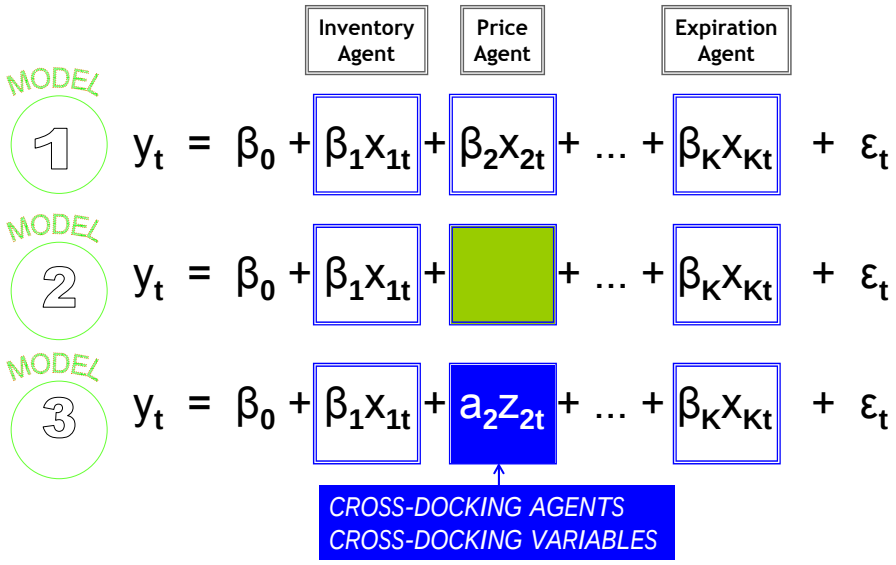


Figure 33: Cross-Docking Agents and Variables in Dynamic Business Scenarios (see note m, on page 96)

Illustrations in figures 32 and 33 (page 67) are proposed by the author to indicate one possibility in the evolution from equation-based model (EBM) to a profitable mix of EBM plus agents-based model (ABM) systems (software) to improve decision support. Various versions of so-called 'adaptive' EBM exist but their efficiency vanishes in face of increasing levels of uncertainty and variables. It is still unclear how the EBM-ABM synergy may be better exploited but a part of that vision may evolve from convergence of ongoing research in artificial intelligence. The following discussion is based on combining excerpts of various threads of research developments from a recent proceeding (see notes M on page 104 and references therein).

It may be of interest to note that, for example, Markov decision processes (MDP) form the foundations of decision-theoretic planning but classical solution techniques for fully observable MDPs, generally, rely on explicit state and action space enumeration, making it useful in 'toy' problems. In a MDP model, the system is in one of a finite set of states at any time point. In each state there are a number of actions to choose from. Execution of an action may offer a 'reward' and cause a stochastic change in the system state. The difficulty arises in finding a policy that maps from states to actions so that the total reward over an indefinite number of action executions is maximized (optimized). One problem is an assumption that a description of the system and actions are available (based on which a policy is mapped from states to actions). Another problem surfaces because domains are often factored – state space consists of assignments of values to a set of variables – thus, domains have state spaces that grow exponentially with the number of relevant variables. State explosion limits use of the MDP framework. Intuitively, the above situation can be extrapolated in terms of a supply chain decision support system where a parameter, say price, is influenced by several variables but data or observations may be available for only a subset of the variables. Hence, decision scenarios where observations do not offer a complete description of the state, uses an extension of MDP, known as partially observable MDP or POMDP. Thus, MDPs are special cases of POMDPs in which everything is observable (rare in real world). Other approaches to reduce the effective state space include limiting computation to states that are reachable from the starting state(s) and feature-based representations to create state abstractions.

Such abstractions may help conceptualize suggestions presented in figures 32 and 33 on page 67. Each explanatory variable may be linked to (observable?) data from a plethora of other variables that may have an impact on the explanatory variable (eg: historical values of the variable, in equation 2, page 24). However, the relevance of all possible impacts on the explanatory variable may not be the same. Similarly, while conceptually 'cross-docking agents and variables' (figure 33), it may be important to use those explanatory variables that are relevant to the dependent variable. When Agents help to make such decisions, the Agents may 'learn' what is relevant in the 'context' of the decision.

In AI, as in humans, ability to make decisions based on only relevant features is a critical aspect of intelligence. According to Andre and Russell (see notes M on page 104 and references therein), for example, if one is driving a taxi from A to B, decisions about which street to take should not depend on the current price of tea in China. Similarly, when changing lanes, the traffic conditions matter most but not the name of a street. State abstraction, in AI, is a process of eliminating features to reduce the effective state space. This concept is, in part, the basis for the illustration in figure 33 on page 67 where business model 2 eliminates the price variable (agent) if in that particular model the impact of price is not relevant for determining the value of the dependent variable (y). In AI, such reductions can speed up dynamic programming and reinforcement learning (RL) algorithms and improvements based on RL.

Without state abstraction, in the taxi driving example offered by Andre and Russell, every trip from A to B is a new trip, every lane change is a new task to be learned by an Agent, from scratch. Further, it has been noted that a variable can be irrelevant to the optimal decision in a state even if it affects the *value* of that state. Suppose the taxi is driving from A to B to pick up a passenger whose destination is C. Hence C is a part of the state but is not relevant to navigation decisions between A and B. The value (sum of future rewards or costs) of each state between A and B can be decomposed into a part dealing with the cost of getting to B (unaffected by the choice of C) and a part dealing with the cost from B to C (unaffected by the choice of A).

This concept of decomposition is applied, at least in principle, to illustrate the 'decoupling' ideas of the author presented in figure 32 and 33 on page 67 in the context of SCM. Further evidence of general applicability of decomposition is found in work motivated by navigational problems arising in mobile robotics domain. Lane and Kaelbling (see notes M on page 104 and references therein) offers this example: a package delivery problem in which an Agent navigates through a building with stochastic movement commands and attempts to deliver packages to fixed (arbitrary) locations. In principle, this is a traveling salesman or 'salesdroid' problem where it is possible to decouple the stochastic local-navigation problem from the deterministic global-routing problem and to solve each with dedicated methods (yielding a net exponential improvement).

Taken together, state abstractions (taxi driving example) and nearly deterministic abstractions of MDP (mobile robotics example) are to be viewed as principles or scaffolds with respect to the issues addressed in this article regarding decoupling 'chains' to improve overall decisions using Agents in the framework. In AI, overall 'world utility' is an area of research where processes with individual goals (sales of aspirin) must interact to maximize 'payoff' in the 'world utility' function (total store sales). One such framework is referred to as COIN or 'Collective Intelligence' and may be relevant, in principle, to the discussion of including local optimization within global optimization in an Agent integrated decision system.

When we talk about 'state' in the context of planning and execution of the plan that alters the 'state' model, we are referring to assumptions made in AI regarding the complementary approaches of knowledge-based and domain-independent planning. Handcrafted user-provided control information is one part of the knowledge-based planning. Classical planning assumes that the initial state of the system is known and that state transitions are deterministic. When these assumptions are removed (in real world scenarios), the state is no longer known. Thus planning with uncertainty requires us to define how uncertainty is modeled and how sensing or feedback are taken into account to reduce uncertainty. Of course, sensing makes sense in a state of uncertainty (ie, the real world). If there is no uncertainty, sensing provides no useful information and can be ignored. Planning under uncertainty without sensing or feedback reduces the problem to a deterministic search problem in a belief space. We have discussed this 'belief' space in context of a Game Theory (Signaling Game) application in figure 8 on page 15 and observed that it holds the potential to fuel the Bullwhip Effect. Data (sensing) and information derived from data may help reduce uncertainty. Hence, automatic identification technologies (AIT) such as RFID may be helpful to improve the degree to which we 'know' the state of a system (observations). In another vein, we have discussed that high volume data from AIT (RFID) may be used to reduce model inaccuracies that are treated as noise or error in order to improve forecasts or prediction, for example, by using ODD-VAR-GARCH, essentially a time series tool. Research by Wah and Qian (see notes M on page 104 and references therein) raises the hope that the ODD-VAR-GARCH tool proposed by the author may improve predictions. Wah and Qian applied constrained artificial neural network (ANN) formulations and learning algorithms on time series models (related, in principle, to ODD-VAR-GARCH) to demonstrate prediction accuracy of future stock prices over a 10-day horizon.

AUTOMATIC IDENTIFICATION TECHNOLOGIES

Automatic identification technologies offer tools to acquire data about objects (pumps, toothpaste, cameras, bullets, insulin). Innovation and leadership lies in the effective use of the data (61), not in its acquisition. In 1894, a 20-year-old Guglielmo Marconi and Oliver Lodge, independently, demonstrated how to communicate (data) using radio waves. Half a century later, with the discovery of the RADAR at MIT, it was evident that the RF spectrum was going to "make waves" for quite some time (see notes o, on page 107). Near the end of the last century, with the establishment of the MIT Auto ID Center (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.

The two thinker-founders of the MIT Auto ID Center (45) created a "storm in a tea cup" by simply reversing the conventional thinking (kilobytes of data on RFID tags) and proposing minimal reference data (number) on RFID tags to be the electronic product code (EPC) which serves as a reference to physical objects, data about which is stored on the internet. Legendary Vinton Cerf once commented that "the latest technologies often produce opportunities to reapply earlier ideas more effectively." The internet allowed Sanjay Sarma and David Brock of MIT to merely store a reference number on the RFID tag and link it to data on the internet. For example, type 151.193.204.72 into Tim Berners-Lee's innovation, the web browser, and arrive at www.usairways.com where you can transact your airline travel needs. A simple string (151.193.204.72) leads you to the information.

The generic organization of EPC was to extend the Universal Product Code (UPC) format currently used in bar codes (46). Thus, EPC is re-using an 'old bag of tricks' yet may be 'disruptive' to the *status quo*. The '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 a '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 (and passive) ultra wideband (UWB) tags. UWB tags can transmit data to distances of 30-300 meters using low power levels. UWB signals can penetrate metal barriers as well as concrete to an appreciable extent.

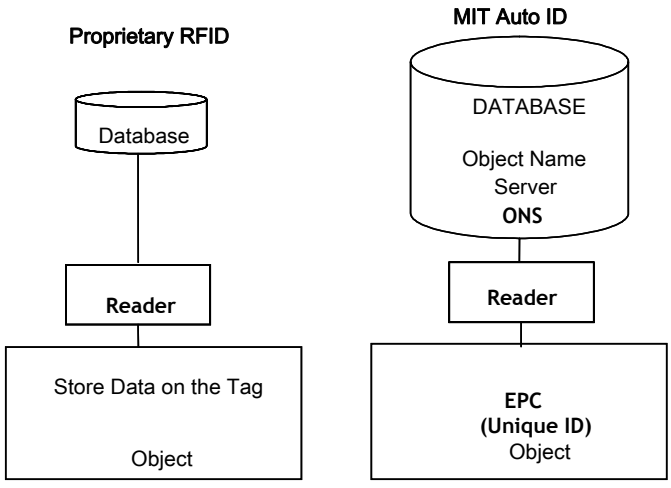


Figure 34: Example of Confluence: Evolution of the EPC made possible by the diffusion of the Internet

The 96 bit electronic product code (EPC) as proposed by the Auto ID Center (46), is made up of header, EPC manager (manufacturer's information, also in bar codes), 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. In one version (Figure 35), 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 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 soft drink cans may be individually identified if each had a RFID tag. CocaCola Corporation, the world's largest bottler, produced only 19.8 billion unit cases in 2004.

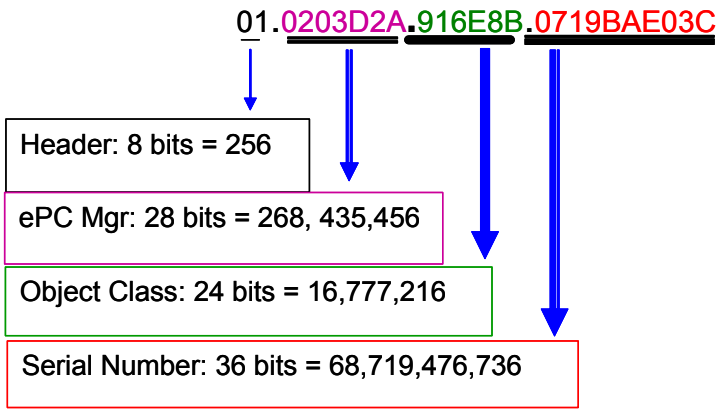


Figure 35. One version of the 96 bit Electronic Product Code (EPC)

The impact of pervasive RFID (UWB) deployment will create an avalanche of data, but can we extract valuable information from this data? Mechanisms to grow ubiquitous decentralized data infrastructure compliant with evolving ontological frameworks are still missing. In US, there are 1.5 million retail outlets, 160,000 grocery store chains, 400,000 factories and 115 million homes. US consumer packaged goods (CPG) industry produces one 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 the CPG industry, alone. The road to ubiquitous tagging of objects will dwarf the current internet that now holds about 1 billion web pages with about 10 petabytes of data. In 2003, estimates suggest that businesses generated about 1 terabyte of data per second, excluding AIT data. The future requires a radically different mechanism of data and information handling. In principle, each object may have its own IP address since the organization of IPv6 and 128 bit EPC are compatible formats to help catalyse this goal.

Given the potential impact, the 'RFID' market is, naturally, in quagmire, in part, spawned by unrealistic claims by some proponents of RFID and others who are disproportionately focused on the cost to acquire data but not high on the value (from connected data). Another component has emerged in the form of individuals or advocacy groups who are quite vociferous about privacy of information yet offers no substance to explain what constitutes violation of privacy if an alphanumeric string serves as a reference for Wrigley's Chewing Gum. How is the 'reference' through EPC different from UPC or bar code information?

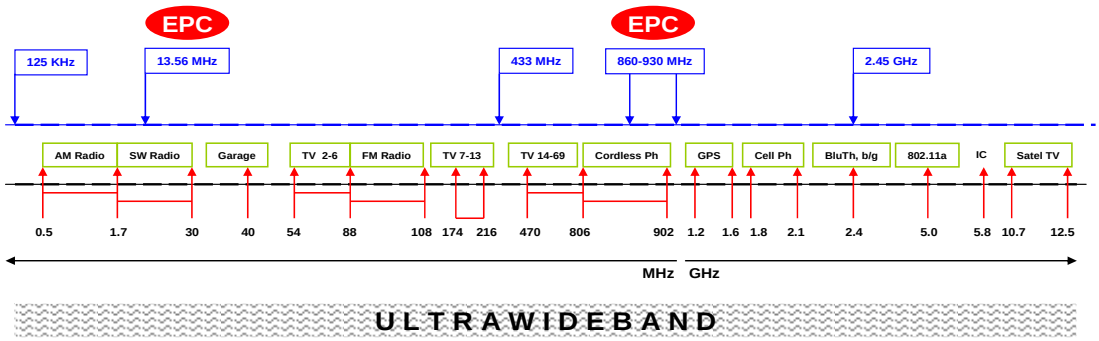


Figure 36: EPC specifications are available for 13.56 MHz and UHF (860-930 MHz)

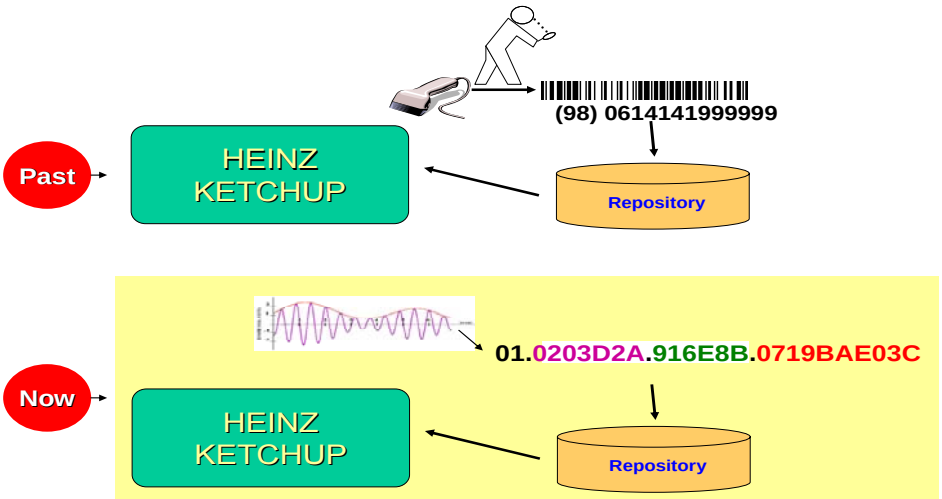


Figure 37: Privacy Issues?

According to Lisa Scanlon (47), 5 billion bar codes are scanned each day, worldwide, but change may be imminent. The inventors of the first linear bar code system, naturally, were 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 (47). Both were graduate students at the then Drexel Institute of Technology in Philadelphia and the idea was triggered by over-hearing a conversation in 1948 between the President of a grocery 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, 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 (UPC or universal product code). Then, IBM wanted "in" on the action and brought in Norman Woodland, still an employee at IBM, to help launch the bar code effort. In 1973, Woodland's leadership persuaded the standards committee to choose IBM's symbol over six other competitors. On 26 June 1974, in a Marsh Supermarket in Troy, Ohio, a pack of Wrigley's Chewing Gum was the first item scanned using the (universal product code) bar code (47).

Given the volume of information available on every facet of RFID, it is unnecessary to add a technology review in this article. It is the opinion of the author that short sightedness and hype has engulfed the *applications* market. Do businesses expect to ride the RFID bandwagon to nirvana? Unfounded statements regarding price of tags and bogus assumptions in business cases peddled by disingenuous cartels continue to fuel this ludicrous mayhem. Disillusionment from this RFID panacea is inevitable. The backlash from the disenfranchised may leave a scar on the sensible use and potential future benefits from use of improved RFID systems (see notes n, on page 106).

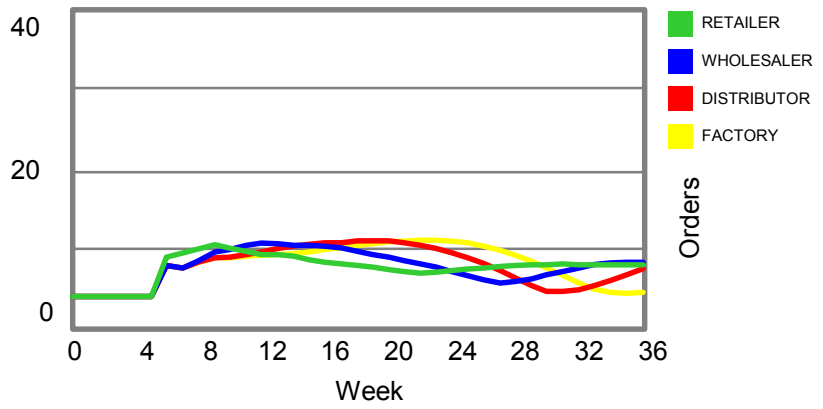


Figure 38. Bullwhip Effect after RFID (10) ?

There aren't any quick solutions. It is unlikely that one solution will serve all industries in an evolving scenario where even engineering principles of RFID systems will evolve (page 100, 106). Review of current processes may promote process innovation: pre-requisite to realize gains from right-time data and analytics. If businesses continue to use old processes and system and/or analytical software tools using obsolete algorithms, static mathematical frameworks or off-the-shelf optimization, then, gains from 'slapping on' RFID systems may be negligible.

The euphoria catalysed by the (MIT) Auto ID Center may have had its virtue in enabling the market to think about potential applications of a global unique identification (EPC) standard and a feasible tool (RFID) which could save operational costs if data accuracy exceeded that of barcodes. Combining accuracy of data acquisition with data sharing to optimize supply chain practices surfaced as a key benefit. However, the euphoria was soon followed by a near-hysteria of profiteering that abandoned the difficult questions regarding systemic gains from adoption and pervasive deployment of auto id (RFID). Respectable companies rushed to produce RFID kits as turnkey solutions!

There is an immediate need for the 'next-generation' infrastructure thinking and changes necessary for adoption of automatic identification technologies. RFID usage is at a critical point in its evolution to deliver the benefits of real-time supply chain management, military readiness and security applications. Some of the challenges are:

- Current RFID technology, theoretically, can process a maximum of 80 tags per second. The actual numbers of tags that can communicate with the readers are much less. The limitation is in the back-scattering method the tags use to communicate with readers and the number of packet collisions (that increases nulls) that reduces the packet rate significantly.
- RFID tags cannot be simply placed on liquid or metal containers, because of RF absorption and reflection.
- Tags only transmit their ID but many services require location information that cannot be provided since the communication is based only on narrowband RF.
- Fixed frequency readers need analog components. Hence, multi-frequency readers are too expensive due to 'doubling' of components. Therefore, it is unlikely to help augment large scale adoption.
- Absence of hardware/software infrastructure as an ubiquitous data infrastructure for intelligent information.

Despite its trials and tribulations, RFID remains a very important tool and benefits from its use are significant, but realization of such 'fruits' are less likely to materialize under the current practices, in terms of RFID hardware and software as well as business systems where the information from RFID data is used (see notes L on page 100 and notes N on page 106). RFID hardware appears in considerable variety (tags, readers) due to different [1] frequency [2] emitted radiated power [3] standards compliance [4] air interfaces and [5] immediate data handling and data transfer after initial acquisition. The latter is the crucial link that enables ***use of data*** to improve processes or systems (business software, legacy system, decision support, analytical tools), for example, inventory, spare parts optimization. Current heterogeneity of platforms and middleware to handle and transfer data, poses significant systems and information integration problems that are fueling inter-operability nightmares, thus, retarding the progression from 'pilots' to system-wide deployment or adoption.

The current thinking to use 'readers' specific to one or more RF modes may not be a sustainable approach for the infrastructure necessary for object identification to become pervasive. In 2004, heterodyne readers that can interrogate MHF (13.56 MHz) and UHF (902-956 MHz) tags cost about US$4000 or more. Consider commonly used frequencies, passive vs active tags, many standards (EPC, GTIN, GTAG) and regional regulations (RF spectrum, emitted radiated power). Taken in combination, it spawns several types of transponders and to read the tags we will need a variety of readers. Multi-frequency tags will not stem the problem. According to the current model, businesses dealing with objects from global partners, therefore, must possess infrastructure (several types of readers) compatible to read a plethora of tags. Current reader vendors, present hype and lack of foresight, thus, may deliver a debilitating blow to the real benefits of object identification and sharing of that data to improve processes, such as supply chain management or military readiness. Readers must be ubiquitous and part of the *civil engineering infrastructure* similar to electrical outlets or switches, evolving to form the internet of devices (Interdev). Such a framework may make pervasive data acquisition and sharing a real possibility. Control, security, updates and hardware improvements are delivered via software which itself becomes a part of the infrastructure.

It is this scenario that is outlined in Figure 3 (page 7) where the reader in the warehouse is always 'on' but the ability to read certain objects (or not) is controlled through the software layer by the authorized user and the authorizations allowed by the principal user. The 'views' of the contents of the warehouse is limited to goods that the user can 'read' by virtue of the preamble that must be exchanged and validated between the reader and tag (similar to current architecture that is embedded in EPC specifications and can be adopted elsewhere).

MIT's Vanu Bose created software defined radio (SDR) which has evolved over the past 15 years (see note L, page 100) is probably the best solution at hand to deliver this ubiquitous infrastructure in a manner that is (transponder) hardware agnostic with all modulations effected through the SDR OS. This view, that of, using SDR (in some form) as ubiquitous RFID interrogators (in your refrigerator or in a warehouse) is the author's proposal based on the current understanding of SWR (software radio) and its ability to read RFID and UWB.

Because SDR is intrinsically linked to the future of global mobile telephony, an area of convergence between SWR infrastructure for real-time data and *delivery of real-time data as a service*, may evolve as a business for telecom providers to serve small and medium enterprises. In 2002, a relevant service model was explored by NTT (www.ntt-east.co.jp/tmmall/rf.html). Telecoms could become even more innovative to explore combining such data service model with the concept of distributed ERP, proposed by the author, earlier.

Our view of RFID ***deployment*** is from a process perspective, that of a tool and an element of the confluence that may improve decision making. Most importantly, can real-time data from RFID or UWB, when used at the right time, help reduce supply chain inefficiencies? Can RFID data tame the Bullwhip Effect? Moving the 'push-pull' boundary can reduce some uncertainties evident in push-only systems.

The diversity of the end consumer makes it impossible to suggest any general mechanism to get a better handle on how to improve the 'pull' signal (demand). The hypothetical mock scenario below is one suggestion of 'pull' signal:

Retailers in single digit profits dream about improving accuracy of demand forecasting, especially for perishables, to reduce waste. Consider a scenario for super-market retailer: A down-to-earth family of four living on San Silvestro in Venezia does not own an internet linked, EPC enabled, Agent impregnated, 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 at San Stae 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 Emma is close to emptying the shampoo or CMC finishes the Barilla tortellini, they should add these to the shopping list, as usual, but instead of the writing pad, they should write it 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 and wait 24 hours before home delivery, then you get a discount. If you wait 48 hours, you receive 2.5% off your bill. What if you 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 tips (recipes) 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 Mr. Hurd gets to know, today, that Becks may want to buy Bolognese sauce, tomorrow?

Convergence of falling prices on PDAs, low cost of wireless (WiFi, Wimax) 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 objects and the data flow from customers' ***pre-shopping list*** 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 behaviour of customers in an area where 50% of the adults are internet users?

RFID Privacy Issues: Where's the Beef?

When beef from Canada was alleged to contain meat from cows affected with bovine spongiform encephalitis (BSE), commonly known as mad cow disease, one super market in the midwest USA, who sold Canadian beef, tracked customers who bought packages from that specific beef shipment, by reviewing purchases of customers who used their supermarket loyalty cards. The system used only barcodes and the action may have saved lives.

The unreasonable claims about RFID by privacy advocates stems from a poor understanding of the technology ***versus*** the processes that may be linked to the technology. Whether customer data will be linked to inventory data is a ***process decision*** not a function of technology. RFID technology, ***per se***, cannot even begin to invade privacy. A 24-bit software instruction programmed on the transponder (see Figure 39) can inactivate, block or destroy its ability to transmit any string of binaries (48). The information encoded by the binary data (EPC, for example) is merely a reference number, almost identical to the barcode that we have been using since 1974.

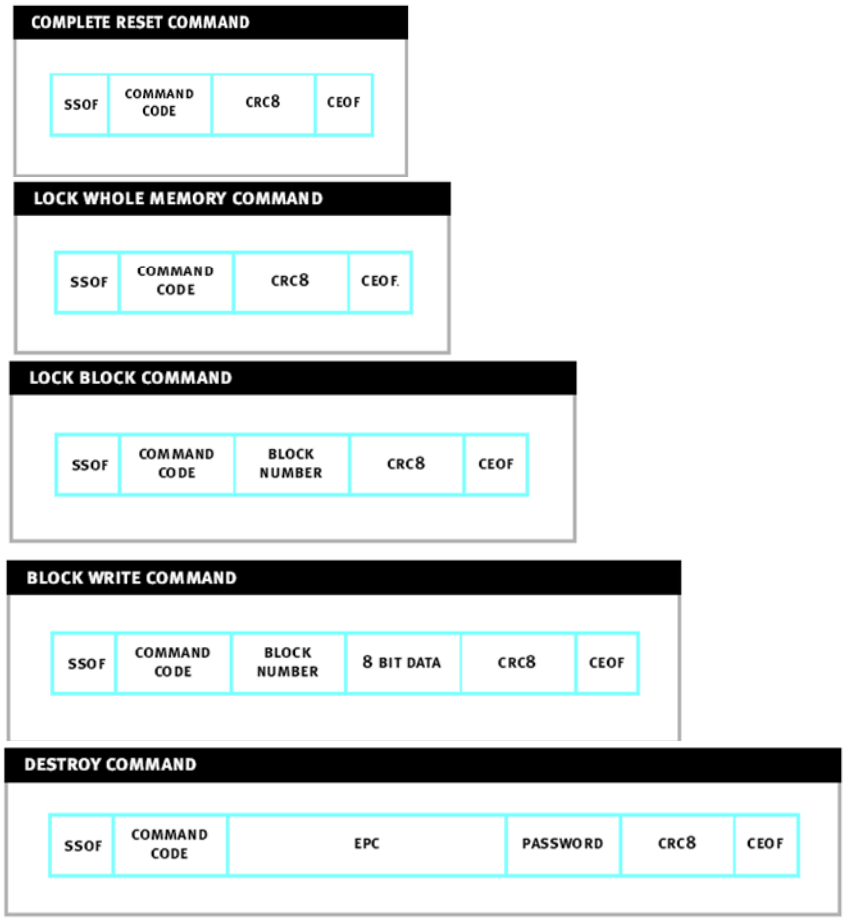


Figure 39: Steps in the Right Direction: Why Privacy May Not be a Real Issue

Ultrawideband: RFID Made Useful

Instead of the customer's grocery pre-shopping list, what if it was for spare parts at the Redstone Arsenal 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 or passive ultrawideband (UWB) tags will far exceed low cost passive RFID tag usage.

UWB stems from work in time-domain electromagnetics that began in 1962 (49). At Sperry Research Center, then part of Sperry Rand Corporation, 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 Bernard Oliver of HP (1962). In April 1973, Sperry Research Center was awarded the first UWB communications patent. Through the 1980's, this technology was alternately referred to as baseband, carrier-free or impulse. The term "ultra wideband" was applied by the US Department of Defense in 1989. Sperry Research holds 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 connection-less, multi-hop, 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 (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 when and if the automated highway becomes a reality.

Comparative analysis of UHF versus UWB shows that UHF RFID has a spatial capacity of 1 kbpspm^2 (50). Spatial capacity of UWB is 1000 kbpspm^2 or 1000-fold more. Most RFID types (125KHz, 13.56MHz, 915MHz) possess a spatial capacity of 1 kbpspm^2 (IEEE). Spatial capacity focuses not only on bit rates for data transfer but on bit rates available in confined spaces (grocery 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 lumens per square meter that determines the illumination intensity of a light source. Growing demand for greater wireless data capacity and crowding of regulated radio frequency (approved ISM spectra) will increasingly favour systems (spectrum) that offers appreciable bit rates and will function despite noise, multipath interference and corruption when concentrated in smaller physical areas (stores, warehouses). Will spatial capacity limitations clog the 'interrogation' system when item level tagging becomes a reality? Some are exploring BlueTooth with spatial capacity of 30 kbpspm^2 while asset management may use 802.11a (5.15-5.35 GHz) with spatial capacity of 55 kbpspm^2 (spare parts inventory in an air force base). Part of this reasoning is evident in independent efforts by Hitachi and Sony who are exploring BlueTooth options. Unfortunately, 802.11a is non-compliant with 802.11b but 802.11g is compliant with both 802.11a and 802.11b. It is obvious to most that the properties of UWB could make 802.11, like BlueTooth, a thing of the past. Already ultrawideband gadgets are available and the market is looking forward to integration with 802.16a or WiMax.

Quite a few companies are exploring ultra wideband since its appearance on the scene. UWB spans several gigahertz of spectrum (fig 36, pg 72) at low power level below the noise floor of existing signaling environment. Conventional narrow band technology (802.11a-b) rely on a base "carrier" wave that is modulated to embody a coded bit stream. Carrier waves are modified to incorporate digital data through amplitude, frequency or phase modulation (IEEE). These mechanisms are, therefore, susceptible to interference and the coded bit stream (for example, electronic product code or EPC) may be decoded or intercepted (defense/security). UWB wireless technology uses no underlying carrier wave but modulate individual pulses either as bipolar or amplitude or pulse-position modulation (sends identical pulses but alters transmission timing). UWB offers pulse time of 300 picoseconds and covers a broad bandwidth extending to several gigahertz. UWB operates in picosecond bursts, hence, power requirements are drastically lower (200 mW) compared to 802.11b (500 mW) or 802.11a (2000 mW). Equally staggering is the data rate for UWB (0.1 – 1.0 $gbps^2$) when compared to 802.11b protocols (0.006 $gbps^2$). Sony and Intel are leading this research for wireless transmission of data, video, networked games, toys and appliances. Today we have robotic vacuum cleaners and lawn mowers that could clean the living room or the garden without ever touching the sofa or grazing by the rose bush. Universal appeal for UWB is latent in its capability to offer a global standard. Without FCC-like country-specific restrictions, an old technology like UWB still remains virgin for many possible applications and may be the only global wireless communication medium that may claim, someday, a truly global standard to help create the 'internet of things' as proposed (45). After September 11, UWB transmitters were mounted on robots for search missions at the World Trade Center since UWB is less hindered by metal (Coke cans or turbine spare parts) or concrete (buildings and warehouses). 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.

UWB-RFID active transponders are not cost prohibitive while transmitters are cheaper than 802.11b RFID readers because they do not need many analog components to fix, send, receive specific frequencies. The combination of UWB plus narrowband technology to produce a passive UWB transponder may be a reality (by combining UWB communication with narrowband RFID tag). Combining a narrowband receiver and a wideband transmitter in the tag optimizes collecting RF energy on the receive channel combined with ultra low power on transmit channel. At the MAC layer, optimized conflict resolution algorithms allow multiple tags to communicate efficiently and effectively with the reader. Due to this algorithm the channel is used efficiently (OFDM or orthogonal frequency division multiplexing), resulting in an increased effective bandwidth that allows more tags to communicate with the reader. (Similar use of OFDM to enhance fixed frequency RFID readers is possible. Use of OFDM in powerline communication may reduce communication infrastructure cost). Utilization of narrow band downlink and wide band uplink communication enables wholly (passive) or partially battery-less UWB tags to be manufactured at low cost. Unlike passive RFID tags, passive UWB tags use accumulated power to transmit UWB impulses to the reader. UWB communication is resilient to selective RF absorption, since the data can be recovered by the reader by relying on the message content in the 'not-absorbed' frequency bands. Due to the broad frequency content of the transmitted UWB impulse, it is resilient to path fading and enables readers to determine location of tags. Thus, UWB tags identify and also locate. For example [a] movement of objects in a warehouse as well as the storage location, [b] wearable cardiac monitors with UWB transmitters could alert hospital workers to an emergency and pinpoint patient's location and [c] smart highway might be equipped with UWB transmitters to communicate with UWB-equipped vehicles. UWB tags can be rewriteable and programmed for 64, 96 or 128 bits. The 128 bit architecture is compliant with IPv6 structure as well as EPC, GTIN and other UPC-EAN schemes. UWB not only provides data and location of objects when tagged to objects but can also form a wireless network to upload the data (over distances of 30-300 meters through metal and/or concrete) in much the same way that WiFi (802.11b) wireless networks may be created (in addition) to upload data from traditional RFID tags.

Despite advantages of UWB, in general, the disputes stem from claims that UWB transmission may interfere with spectrum used by cell phones and air traffic control. FCC is investigating but it is poised to open up even more of spectrum for UWB commercial applications. However, to ensure that there is no interference with fire and police radio systems, the FCC suggests UWB usage in the 3.1-10.6 GHz band, which may limit UWB effective range (10 to 20 meters). FCC licensed the first chipset based on UWB in August 2004. The UWB standard under contention is dubbed 802.15.3a by IEEE. Backed by Motorola and 60 other companies, the UWB Forum is pushing a variant that transmits data in a continuous sequence (DS, Direct Sequence). The rival MultiBand OFDM Alliance (MBOA) is being promoted by Intel and a longer list of companies including HP, TI, Philips, Samsung. MBOA proposes use of OFDM that hops between multiple frequency bands to improve performance (OFDM is used already in DSL). Without the burden of license fees for spectrum, the commercial floodgates for UWB usage may be unstoppable much to the chagrin of the telecom industry. Telecommunication giants who rushed to buy spectrum seduced by the future of 3G services are fighting to keep UWB off the news. Devoid of country-specific spectrum restrictions, UWB may become a global standard for short range communications. MSSI is charting new territories in commercial use of UWB and PulseLink has shown that SDR readers work with UWB chip sets.

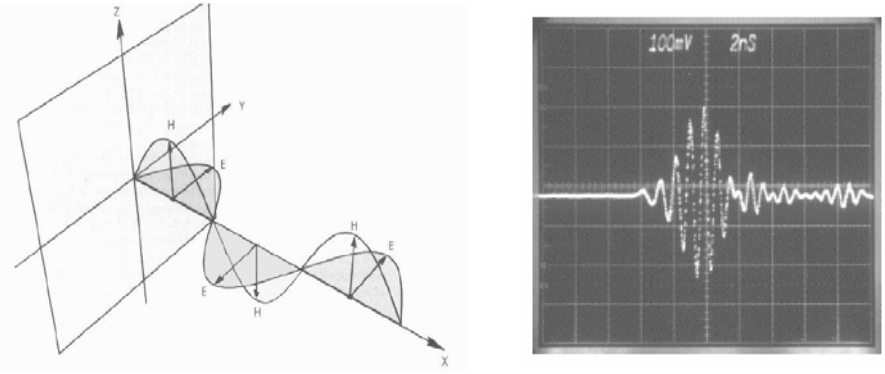


Figure 40: UHF RFID (left) and "Pulse" Transmission of UWB (right)

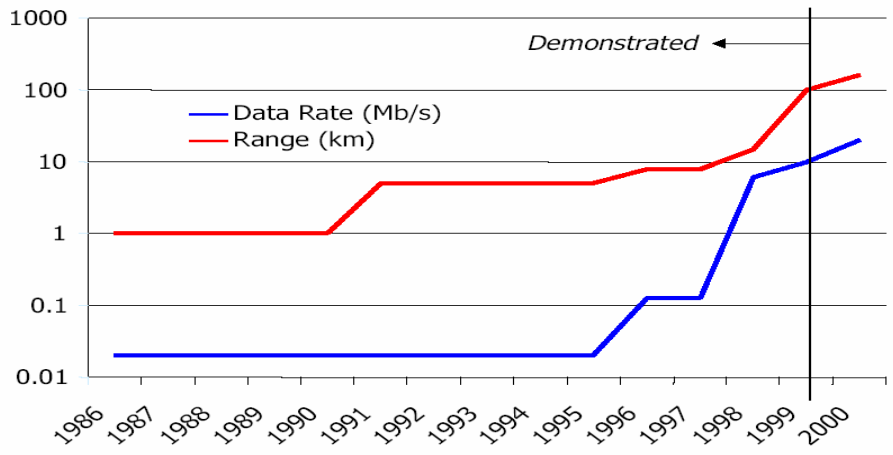


Figure 41: Demonstrated capabilities of UWB as of 2000: Data range of 100km and data rate of 10 Mbps (50)

Sensor Networks

Wireless sensor networks may be a good example of pervasive computing (52). Applications extend from sensing blood pressure and transmitting them to monitoring devices to suggesting trends of warehouse shelf occupancy or 'smelling' hydrogen leaks (63). Sensors do not transmit identification data, such as EPC. Sensor data, therefore, cannot be used in the same manner as RFID. Sensors are self-powered and may form wireless *ad hoc* networks that upload through specific nodes which may be connected to data stores or the internet (see Figure 42). However, each sensor has certain analytical abilities and due to in-network processing, some sensor networks transmit analyses of the data rather than the raw bits of data to provide "answers" instead of only "numbers" to the system. Sensor data may require different thinking in terms of "adaptive flow" databases. The data (analyses from sensor nodes) may stream through databases 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 from a sensor or group of sensors, then, the query comes into effect. Embedded sensors are likely to influence fields as diverse as healthcare and supply chain. Sensors attached to spindles in drilling machines may upload the status of the spindle such that it is serviced or replaced within a reasonable time to avoid breakdown and downtime. Metrics like meantime between failure (MTBF) and other parameters may be helpful to schedule preventive maintenance. Service supply chains (such as heating, cooling) may benefit from sensor-linked monitoring to determine when to send technicians to stem problems before they require emergency attention. The key is to integrate sensor data to improve performance (see note i, page 98). The flood of data from nanosensors (see figure 43) may require Agent integrated systems to extract intelligent information. Bio-nanosensors may evolve as an influential component of healthcare services and management.

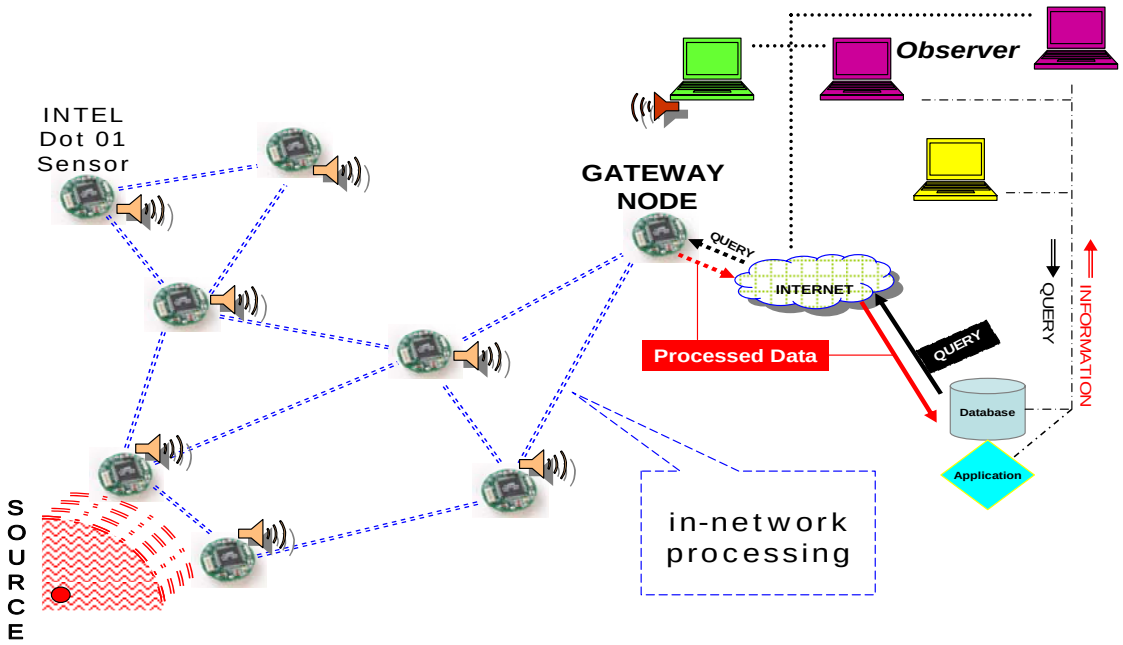


Figure 42: Unwired Sensor Nets can communicate via 802.15.4 (ZigBee)

Information from glucose nano-sensors, through semantic web services, may improve diagnosis and prognosis of type I diabetes that results from insulin-dependent disequilibrium of blood glucose. Diabetics currently use a host of methods to detect blood glucose but lack an automated continuous monitoring tool. Catalysing such convergence may spawn efficient healthcare monitoring, in real-time, delivery and management. Can the semantic web and glucose nano-sensors (see Figure 43) fuel a confluence to reduce diabetic retinopathies?

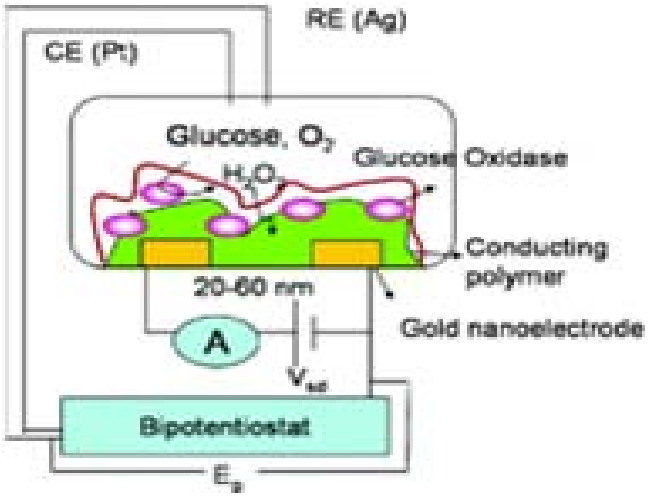


Figure 43: Glucose Nano-Sensors in Healthcare Supply Chain Management? (NanoLetters 2004 **4** 1785-1788)

Let us explore a simple healthcare scenario where a local hospital in US zip code 08544 is the first point of contact for seniors aged 65 and above. About 200 patients are identified to have some form of diabetes (type I and late onset). It is necessary for these individuals to have a monthly check-up and when the blood glucose is elevated, the physician may prescribe a period of insulin therapy. Family history reveals that some patients may have parents or grandparents who had glaucoma, suggestive of diabetes-dependent glaucoma. At present, the 200 diabetic patients must travel to the local hospital for monthly tests in the outpatient clinic. Perhaps most will have normal blood glucose (120 mg/dl) or levels within limits or that which may be controlled by modification of carbohydrate intake. However, a few patients (assume 5) with a family history of glaucoma, require more careful monitoring yet only monthly visits are covered by the insurance. If this check-up happens to reveal elevated blood glucose in this 'risk' group, then, the outpatient clinic may schedule a physician appointment for check-up and intravenous insulin administration.

Process analysis of this scenario will reflect the costs associated with screening and monitoring the patients, who may be otherwise normal. The "supply chain" perspective will reveal volatility of inventory (insulin) with short half-life and other supplier-dependent (location of Genentech on US west coast) factors that makes it necessary to stock enough insulin (inventory carrying cost) and risk expiry. The quality of life analysis may reveal that 195 patients with diabetes but without known family history of glaucoma are perfectly fine but the less than optimum screening for five at risk patients may have reduced their quality of life as well as increased their risk of glaucoma.

Thus, poor exception management in the preventive phase may eventually precipitate glaucoma in one or more of these five at risk patients. Glaucoma induced poor vision or lack of vision creates far greater needs on the system (insurance, social fabric) and drastically reduces the patient's quality of life.

Can glucose nano-sensors alter the outcome? Patients with sub-cutaneous (implanted) glucose nano-sensors may monitor blood glucose levels in real-time (not merely one reading each month). The data may be updated to a domestic node (in-home WiFi 802.11b) and transmitted through web-based service (via 802.16a or WiMax MAN) to a monitoring portal at the local hospital. Hospital policy kernels and patient authorizations will be linked to patient-specific data (made possible by the semantic web). The data is analysed in applications with rules or Agent impregnated monitoring systems which, if necessary, alerts nurse practitioners or doctors should any aberrant blood glucose fluctuation is detected in 'otherwise normal' or 'at risk' patients. The frequency of necessary insulin administration for the 'at risk' group will be synchronized in near real-time. The latter has the potential to appreciably decrease the onset of glaucoma or perhaps keep the patient glaucoma-free for life. The ability to record, monitor and analyse the variation of blood glucose levels over 24 hour period in real-time in the 'otherwise normal' group may uncover subgroups with patterns of glucose utilization that may offer clues to variations in glucose-insulin interactions indicative of other anomalies (insulinoma, autoimmune, insulin receptor dysfunction).

The petabytes of real-time data, if appropriately available through medical semantic web (semantic metrics), may reveal population genetic traits, which could guide future healthcare policy and funding for biomedical research based on the need of the population or self-service communities (independent living for Alzheimer's patients). For very low birth rate nations with ageing population and healthcare costs around 10-15% GDP, the ability to ask right questions may help convert the acquired real-time data into revealing valuable information. Mining patterns for insulin consumption may help hospital procurement managers adapt the 'supply chain' of insulin to diminish inventory carrying cost and stem wastage without decreasing service level (quality of service). Elimination of the need for monthly blood glucose test will decrease costs for out-patient services and improve resource utilization. Taken together, over a number of services and drugs, real-time connected data may offer risk pooling of services and products by geography to reduce costs for managed healthcare and national health services (eg: the hemorrhaging NHS scheme in UK).

Nano bio-sensors for blood glucose, cartilage degeneration, blood pressure and others are already available or will soon become available. Wireless data transmission protocols are in existence along with the ubiquitous internet. All the tools necessary to reduce costs, now, are at hand, as is, profuse skepticism. If innovative leadership is allowed to function, the scenario above may be implemented, in a few short months. For the next phase, this data may be expected to help with better, improved and accurate diagnosis-prognosis through connected thinking via medical semantic web. The latter is an important emerging confluence and a task for creative organizations that may be modeled on Doug Lenat's Cycorp. The diffusion and incorporation of the semantic web is still a slow process but it is gaining momentum almost in proportion to the increasing honours being bestowed on its founder and his penchant for open source idealism.

SEMANTIC WEB (Is Spreading)

The average user may 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 one 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. It can offer process transparency across language and geographic boundaries to connect partners even if individual partners define or perform certain functions differently from others.

Sir 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 as early as 1995 and more clearly by 1998. His vision has matured and progress has taken place in research communities around the world to demonstrate that semantic web may 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 decentralized vocabularies. In philosophy, ontology is a theory about the nature of existence (of what types of things exist). 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 in future, empowered with this metadata, may be far more "meaningful" and "contextual" in their understanding of the data without human intervention, provided the data is in machine readable format (53).

Human language thrives when using the same term to mean somewhat different things, but automation does not. Michael Dertouzos (40) and James Hendler (54) offer 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. Our current syntactic web is incapable of such distinctions (meaning and context). This is not the end of the clown 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 (53, 54).

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). 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. In other words, transformational mapping functions between ontologies will be necessary as ontological frameworks begin to evolve, perhaps in the same 'organic' manner that characterized the explosive growth of websites from 0 to 25 million in the two decades since 1980. There may not be any one 'standard' ontological format even for very closely related topics because the same format can be framed differently in a different language. Therefore, for the semantic web to be globally useful it will be necessary to have layer(s) of mapping functions (analogous to adaptors and transformers (up/down) that are necessary to use select electrical appliances across geographic boundaries).

Ontologies can enhance the functioning of the web in many ways (54). 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.

For computers, the page is linked to an ontology page that defines information about computer science departments (54). 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, 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 (from where did Dr. Hendler receive his degree) that currently would require a human to sift through the content turned up by a search engine (54).

In addition, this markup makes it easier to develop programs that can tackle complicated questions whose answers do not reside on a single Web page (54). 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 Dublin. 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 44 (below) shows Tim Berners-Lee's Semantic Web layers (53) that we included in the confluence to illustrate the Semantic Grid Web Services (see Figure 21, page 48).

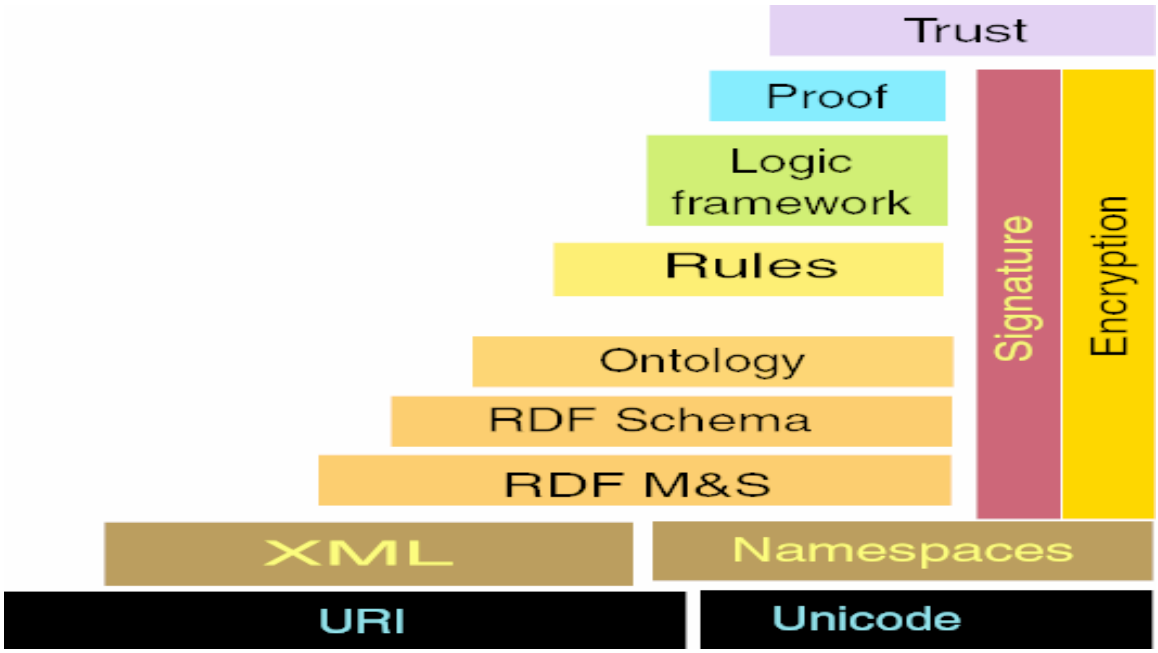


Figure 44: Semantic Web Layers from Tim Berners-Lee (53)

Automated web services claim to discover and connect to various services. Even if these services had Agents, at present Agents have no way to locate a service that will perform a specific function (54). 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 (hence the information box in Windows XP: Found New Hardware). 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 flexible. 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 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 an 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 may increasingly exploit artificial intelligence techniques. Semantic web will provide the foundation and framework to make such technologies more feasible. Its use will be ubiquitous and pervasive as context-dependent-communication evolves successfully to deftly deal with the idiosyncrasies of the human language dependent ontological frameworks through intelligent mapping functions.

Semantic Web in Global Security ?

Neutering threats before terror can strike may be the holy grail of governments and their security czars. The ability to detect threats, therefore, is of paramount importance. Since we are no longer dealing with nation states as primary sources of threats, the World War II treatment of “intelligence” and actions based on such information are practices in absurdity. In the 21st century we are faced with “global public bads” manifested by small groups or individuals that use the internet to communicate or transform commonplace objects to serve as destructive tools. Traditional approaches to uncover such ‘plots’ are practically impotent and hence the US National Security Agency has given considerable considerations to non-obvious relationship analysis. For example, a retail fertilizer store in Norman, Oklahoma, sells 1000 kg pallet of nitro-phosphate to an individual who pays cash $2000 at the store. Days prior to this transaction, an individual (we do not know who) uses an automatic teller machine (ATM) to take out $700 (maximum daily limit) each day for 4 consecutive days. One day prior to the transaction, a small moving van is rented from U-haul and a local warehouse store (Costco) sells 10 large duffle bags to one customer. Two days after the transaction, an individual boards a flight to EWR (Newark, NJ) with two large duffle bags as checked-in baggage and the individual returns to Tulsa, Oklahoma. A few days later, a garbage collector in Brooklyn, NY finds two relatively new bags in the dumpster and takes them home. After a few days, the individual from Oklahoma drives to Springfield, Illinois and catches a flight to Philadelphia, Pennsylvania, with large duffle bags. A few days later, a BMI Baby airline check-in clerk has a scuffle with a passenger who wants to check-in oversized duffle bags for a flight to Hahn, Germany. On a nice April day, a bomb blast in the Deutsche Post mail sorting center at the Frankfurt Main Airport kills 831 people. Days before the blast, bell-boys at the Sheraton Hotel at Frankfurt Main Airport had noticed and reported an individual who was sitting on the connecting walk-way and reading a book. What they reported was that this individual was there for several hours each day. The police checked him out to be an US citizen but the US Embassy was instructed by the US Department of Justice that the rights of the individual cannot be violated by ordering a search because the ‘suspected’ individual did not break any rules by sitting and reading a book.

The intersection of policy, rights and predictive clues when combined with investigation offers a very complex scenario when policy and rights must be evaluated by law enforcement prior to any action, either exploratory or investigative. Similar situations are common in healthcare and insurance industries. The policy framework enabled by the semantic web may help. Semantic webs ability to connect diverse decentralized information in order to improve the efficiency of searches may have produced relevant clues for a non-obvious relationship analysis of information pertaining to the sale of nitro-phosphates in Oklahoma and the bomb blast in Frankfurt.

Semantic Web in Healthcare ?

Since the semantic web is a virtual space, it makes informational structures more relevant. In one view, the semantic web is a collection of knowledge, which, by definition, has machine-accessible meaning (in contrast to the 'dumb' collection of information on today's syntactic web). This notion of the semantic web is currently being challenged in favour of the view of the semantic web as an action-enabling space (55). Who will enable such action? As we have discussed earlier, Agents may play a central role in the semantic web. However, we discussed that Agents are better suited to specific tasks while Agencies may be endowed with intelligence. In other words, the theory of distributed cognition may apply to Agencies (not to Agents). Theory of distributed cognition departs from the mainstream of cognitive sciences in that it emphasizes the participation of external elements (manuals, database) in the Agent's thinking process (56). Interaction of Agents in the physical space can be facilitated by using mobile Agents linked to devices. Such interactions call for introducing ontologies for describing interaction protocols and Agents that can recognize ontologies. For example, inform(door_open) and request(door_open) are different messages even though content(door_open) is the same. The former is about the state-of-affairs but the latter transmits an intention. These can be grouped as interaction protocols (conversation patterns). If Agents are aware of this ontology, Agents can learn new ways of communicating and by extrapolation, adapt to the needs of the system.

Distributing cognition in the semantic web through design of Agents that can understand ontologies describing new interaction or communication protocols may find, in the opinion of the author, significant immediate use in healthcare, today, and in the emerging confluence of nano-bio medicine, in the future. For example, when a paramedic attends an accident victim on the street, she would like to communicate to the nearest hospital trauma unit the status (answers) of the crash victim, not numbers. For example, report(heart_rate) should say 'normal' rather than 80/120 mm Hg. If we consider involvement of Agencies with distributed cognition, the 80/120 mm Hg blood pressure may be linked to context, age of victim. If it is a child, surely report(heart_rate) should not transmit status 'normal' to the trauma center!

One of the recent advances in biotechnology enables screening of gene expression using microarrays (57). The volume of data from such a screen is staggering. Genes rarely interact alone in creating a disease state. We can now identify the location of a gene on the human genome map (and determine its neighbours). Thus, a ***combination*** of information that tells us which genes are active (revealed by microarray) and where on the chromosome that active gene may be located (genome map). This is an exponential gain in understanding that could guide healthcare delivery and manage risk of future complications.

To mine this data and make logical connections, we need Agents in the semantic web with advanced abilities to look beyond the obvious (not only genes but ***genetic circuits***). The type of interaction protocols mentioned above and the theory of distributed cognition, taken together, may be one necessary tool that deserves exploration from the point of view of molecular medicine and bioinformatics with respect to (many) diseases that stems from perturbation of gene-protein networks (for example, p54).

However, the "success" of this confluence in terms of applications must wait for diffusion of the semantic web within IT and software infrastructures. Bio-medicine may reap the early harvest from nanoscale sensors but sensors are sterile unless data can be used for decisions to improve or aid diagnosis. The focus on opportunities in healthcare may yield rapid results but the real value from data may still remain disconnected unless the industry invests and reaps the benefits from the development of the semantic web to connect the dots.

Research and advances in nanoscale diagnostics can save millions of women from the morbidity caused by breast cancer and linked mortality. Two genes, BRCA1 and BRCA2, identified nearly a decade ago, appear to be closely linked with familial early-onset breast cancer that constitutes about 5% of all cases. In addition to BRCA1 and BRCA2, at least eight other genes have been identified as contributors to breast cancer either directly or indirectly [AKT2, BWSCR1A, CDH1, ESR1, FKHR, PAX7, PIK3CA, ST8]. BRCA1 and BRCA2 are also linked to ovarian cancer and rhabdomyosarcoma (cell types are significantly different from mammary epithelium). Among the familial early onset group of women with breast cancer, 40% of them have mutations in BRCA1 and BRCA2. About 120 different type of mutations have been identified in these two genes and twice as many are actually thought to exist in the 10,000 base pairs of A, T, G, C that make up these two genes. Any attempt at patient care management for this subset of afflicted women will be incomplete [or even wrong] without an understanding of the nature of changes [which one of the hundreds of possible mutations] that caused the carcinoma in the first place. The situation is analogous to repairing a leak in a water pipe that is a mile long. If you cannot identify the site of the leak, what are the chances of repairing the pipe? If you do not know 'where' or 'how' to look, what are the chances of identifying the site of the leak? Hence, the value of the semantic web in medicine and healthcare.

In 1998, micro-array was used to stratify patient populations with AML (acute myeloid leukemia). Cytarabine produced remission in 78% of patients (even after a 5-year period) in patients selected by genotype profile [pharmacogenomics, toxicogenomics]. In parallel studies with other AML patients, who were genotypically grouped and predicted to be less responsive, Cytarabine showed remission rates of 21%. In the pre-genomic era when genotypic stratification was not available, this drug could have produced a widely variable remission rate depending on the mix of patients and may not have received FDA approval. Stratification is not yet the norm due to high cost but nano diagnostics may make it routine. Stratification segues toward materializing the 'designer drug' concept that aims to treat patients with genotypically matched drugs to maximize efficacy. This is another example that is in need of ontologies and translational mapping *between* ontologies to enable semantic web users (in medicine) to make the connections from vast resources of decentralized information. Research data on AML, clinical data on AML, pharmaceutical data on chemotherapy, epidemiological data, patient demographics, research advances (RNAi), FDA policy, human trial rules, safety compliance, trial related resource availability and other important data are not and will not be organized in any one search space, in any one nation, in any one format or in any one specific ontological schema.

Creating global consensus to pursue ontological mapping in the biomedical domain may enable the future use of the semantic web to resolve as simple a problem as a single mother in Accra (Ghana) learning from the village kiosk to add iodized salt to her baby's diet to prevent mental retardation. The semantic web of the future may also reveal to a medical resident that the seemingly intractable pain from an apparent cervical spondylitis may not be an orthopedic case. The resident may uncover that such pain sensation can be triggered by local inflammation reflecting autoimmune reaction with roots in the patient's teen-age years when she had a strep throat (*S. aureus*) but did not complete the course of the prescribed antibiotics. Epitopes presented by *S. aureus* are also present on human leukocyte antigens (HLA-B2) in this tale of bio-mimicry. The resulting pain is caused by inflammation. Connected thinking from distributed and decentralized information sources, thus, offers application guidance at the point of contact (POC) to help the medical resident consider shifting the focus from conventional thinking that relates spinal pain to osteopathy or neurology to immunology. Based on this improved understanding, facilitated through the medium of semantic web integrated application at the point of contact, the medical resident offers accurate diagnosis and prescribes therapy to sufficiently relieve the discomfort from inflammation and augment the patient's quality of life.

Collectively, therefore, figures 45 and 46 illustrate real-time data use in healthcare. The computation needs in this arena and benefits are far greater than asset management in hospitals or tracking pharmaceuticals (work-in-progress, distribution, counterfeit). The latter are, however, necessary to improve efficiencies that remain poorly addressed in several medical organizations involved in patient care or in the pharmaceutical industry.

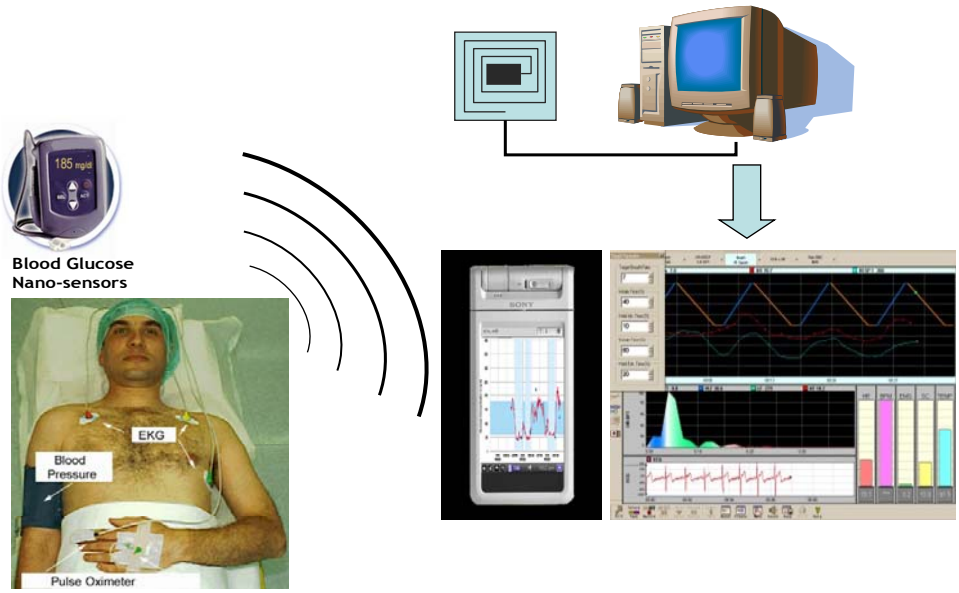


Figure 45: Sensing Technologies (RFID-linked telemetry; nano-sensor nets) for Right-Time Healthcare

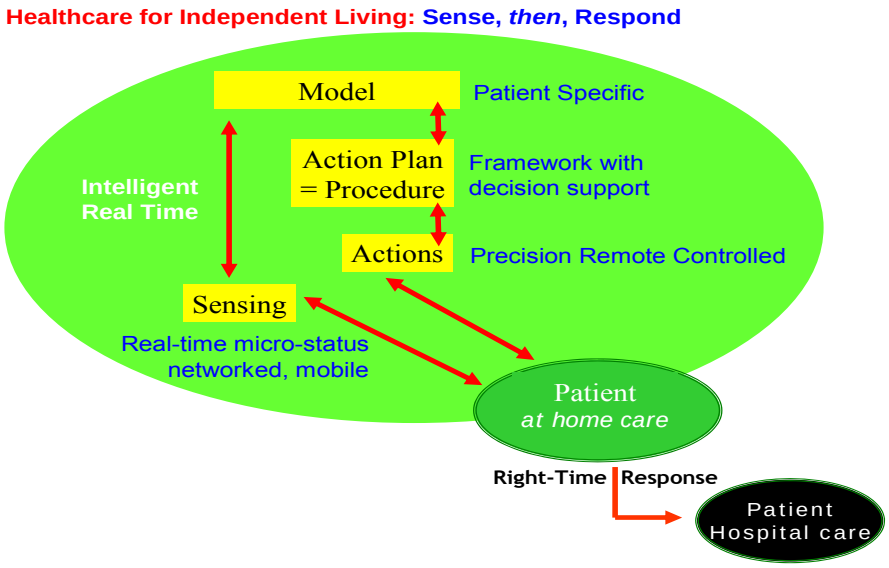


Figure 46: Cost of Old Age: Can we reduce the cost of healthcare through better 'Sense, *then*, Respond' ?

To improve healthcare delivery, the information technology community (bits) may wish to better understand some of the healthcare tools that stem from recent advances in biotechnology, biomedical engineering and molecular medicine. In parallel, the biomedical community that interacts with the patient (atoms) may wish to understand the tools for delivery of 'bits' to aid in making better decisions. Thus, by abstraction, this is a 'bits to atoms' endeavour that fits well with the recursive concept of real-time data linked to processes at the right-time to help optimize decisions (see notes p, on page 102).

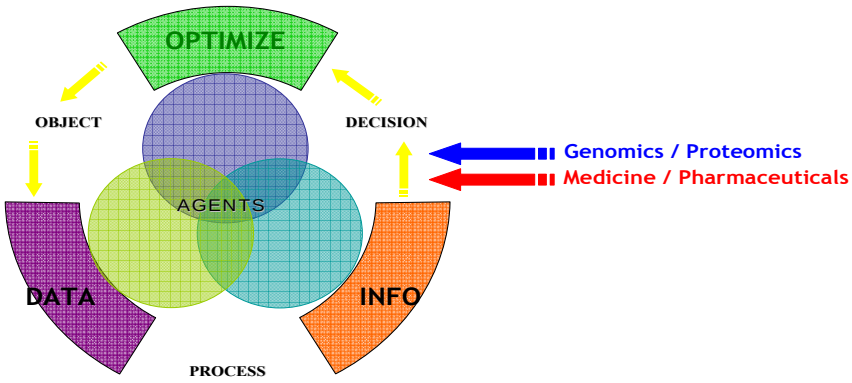


Figure 47: Can healthcare decisions improve by connecting the bits to atoms?

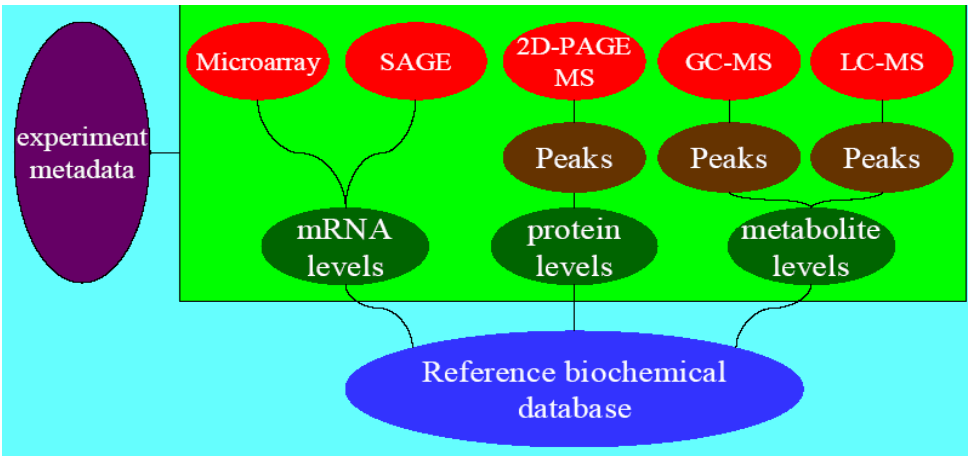


Figure 48: Have Data, Extract Links: Requires Distributed Learning Algorithms for Dynamic Data Agents (62)

CONCLUDING COMMENTS :

Are all Advantages only Temporary in the Adaptability versus Efficiency Paradigm ?

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, weather forecasting, does not interact with another, such as purchasing or customer behaviour. Can we harness the power of multiple individual data models? 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 exponential gains in healthcare delivery or unprecedented increase in productivity through the optimal use of decentralized resources, ability to adapt and prepare for change. We may reduce the cost of goods and services through the elimination of inefficiencies and reduction of transaction costs (TCE).

A model framework to explore the dynamic equilibrium between adaptability and efficiency may prove to be a powerful tool. We refer to this as ***Gibbs Dynamic Equilibrium*** (see page 8 and note j on page 98). The issues discussed in this article may offer clues how to create the necessary framework to address the equilibrium concept illustrated in Figure 49. This may produce a diverse collection of models because the equilibrium may also vary depending on the 'clockspeed' of the industry (58). The balance of adaptability to change (say, in demand) versus operational efficiency (say, in manufacturing or procurement) in order to profitably respond to change may be different for the cell phone industry when compared with the automobile or the retail industry.

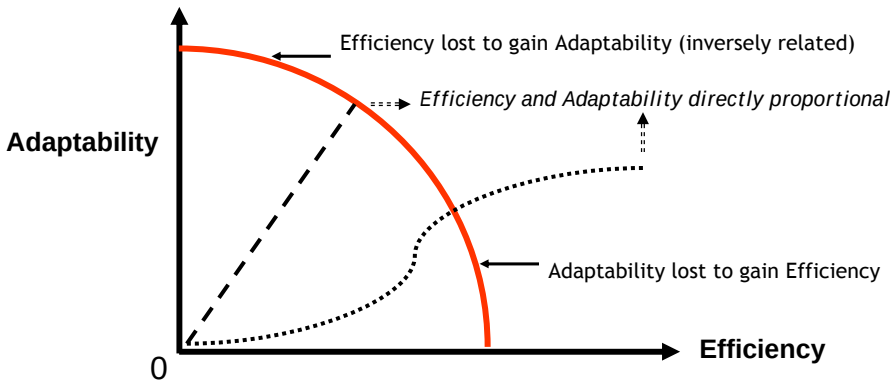


Figure 49: Gibbs Dynamic Equilibrium: Will 'clockspeed' impact proportionality of adaptability vs efficiency ?

Gibbs Dynamic Equilibrium implies that adaptability and efficiency may be inversely (solid line) or directly (broken lines – linear or sigmoid) proportional, depending on the industry. The auto industry may exemplify an inverse relationship. Economies of efficiency (scale) from mass production (make-to-stock) per manufacturer's plan may be compromised in order to adapt to increasing demand for customer-specific preferences (make-to-order or order-to-delivery). For distribution businesses, such as, Amazon.com, their KPI's (key performance indicators) are directly related to the efficiency with which they can adapt to service the volatility of demand.

To build these models, individually and test them in combinations may be a worthy endeavour for generations of engineering-business students and consultants. Practitioners may wish to embrace key elements (concepts, tools, technologies) mentioned here and seek ways to bring about the confluence, emphasized in this article, in part, to validate concepts such as ODD-VAR-GARCH and Gibbs Dynamic Equilibrium. Principles from Game Theory and tools from econometrics empowered by real-time data from automatic identification technologies may optimize adaptability and improve efficiencies. Reducing information asymmetry between entities through secure Agents-based systems may enhance total value network performance. Deriving meaning from and meaningful data through the Semantic Grid Web Services may enable responses based on real-time analytics in real-time or right-time to improve inter-operability between diverse environments. The positive return on investment from catalyzing the growth, diffusion and adoption of the semantic web is at a fundamental level that is deep enough to bubble up improved connectivity which can help global security as well as healthcare. Such tangible progress from confluence may still be measured by the degree of meaningful merger between bits and atoms with respect to process, any process.

Likewise, in the past few decades most of the companies that have created truly extraordinary amounts of wealth have done so by inventing great processes, not great products. Dell, Toyota and Wal*Mart, for example, have risen to the top of their respective industries by coming up with amazingly efficient ways of getting quite ordinary products into the hands of consumers more cheaply than their rivals.

The Economist, 24 April, 2004

For discussions, information and material in this article, the author specifically acknowledges:

David Brock, MIT
Gerard Cachon, University of Pennsylvania
David Culler, (Intel Research Lab), University of California at Berkeley
Nicole DeHoratius, University of Chicago
Charles Fine, MIT
Gerry Frizelle, University of Cambridge
Neil Gershenfeld, MIT
Ajit Kambil, Deloitte Research
Henry Lieberman, MIT
Andy Mulholland, Capgemini
Serguei Netessine, University of Pennsylvania
Ozalp Ozer, Stanford University
H. van dyke Parunak, University of Michigan at Ann Arbor
Ananth Raman, Harvard Business School
Sanjay Sarma, MIT
David Simchi-Levi, MIT
Katia Sycara, Carnegie-Mellon University
Computer Science and Artificial Intelligence Lab, MIT

Sponsors of the MIT Forum for Supply Chain Innovation (are not linked to this article):

AAR Corporation
Capgemini
Deloitte Research
General Electric
General Motors
Government of Finland (TEKES)
IBM
ITRI (Industrial and Technical Research Institute, Taiwan)
Intel Corporation
LogicaCMG
Michelin Corporation
National Science Foundation
Pepsi
Siam Cement
SAP
University of Alabama
US Department of Defense
Xerox Corporation

NOTES

a Information Asymmetry is a concept borrowed from economics and used somewhat casually in this article to imply lack of information (data) visibility between organizations. 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 is true, was provided in 1954 by Gerard Debreu (Nobel Prize 1983) and Kenneth Arrow (Nobel Prize 1972) (Arrow, K. and Debreu, G., Existence of an equilibrium for a competitive economy. Econometrica **3** 265–290). In 1986, B. Greenwald and J. Stiglitz offered proof that when information is imperfect (information asymmetry) or markets are incomplete, competitive equilibrium is not efficient (*Globalization and Its Discontents* by Joseph E. Stiglitz).

b Value Networks refers to concepts forwarded by Clayton Christensen in *The Innovator's Dilemma (*by Clayton Christensen, Harvard Business School Press, 1997). It builds on the concepts of Giovanni Dosi and Richard Rosenbloom. We may often use supply chain management and value networks interchangeably.

c It is beyond the scope of this article to delve into an adequate 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 any more adaptive but these models may offer deeper insights. Most businesses are severely under-optimized. In such cases, it is speculative whether real-time information (at the right time) may offer any substantial value. Thus optimization, including game theoretic tools, may be necessary to "tune the engine" before real-time information can help adapt to SCM events. The following may be considered a brief summary of emerging Game Theory (GT) applications in supply chain management from [25) (www.wkap.nl/prod/b/1-4020-7812-9?a=1).

Pragmatic applications of Game Theory (GT) in SCM are emerging, albeit slowly. The latter is in part due to the lack of 'bridges' that can help translate the benefits from considering GT frameworks and then mobilizing the systems necessary to develop software that may offer 'solutions' relevant to business scenarios. In sharp contrast to auto-id related technologies, where data acquisition overshadows the business process, GT is firmly rooted in process, so much so, that it assumes that information (data) is available for decision making. If these two extremes, RFID vs GT can be balanced and astutely embedded in a business vision, then it will deliver better decisions. With GT and RFID tempered by OR and ABM, businesses can extract deeper insight and can expect profitability gains. The confluence, thus, is the essential (and rate-limiting) factor.

GT based dynamic models, where decisions are made over time, are probably more representative of SCM than static games. But, the important caveat in GT models of this type is the lack of consideration of information asymmetry. In other words, the mathematical GT constructs assume 'complete information' at the hands of the players – a situation that may be rare in a business relationship. Nevertheless, one widely used abstraction is the Stackelberg Game that provides insight into 'leader – follower' relationships. The practical use most commonly referred to is the airline price wars but of late, the same principle is being applied to parameters such as service level, lead time, risk, contracts (manufacturer or supplier). Since SCM models offer an upstream firm (wholesaler) with power over a downstream firm (retailer), the Stacklberg equilibrium concept is creeping into OR and thus the models are closer to reality than GT *per se*.

When considering repeated actions over time, such as inventory replenishment decisions, the insights from multi-period games may be invaluable. Broadly speaking these games take into account trigger strategies, threat and implicit collusion. However, only recently the real-world inventory management issues (back-order, salvage, inventory transfer) were considered in an attempt to explore supply chain coordination. An example of differential GT in a stochastic problem may be a manufacturing scenario where two companies are engaged in production and sales of similar products, where, production level may affect price adjustments (feedback dependent continuous-time processes).

Cooperative games are perhaps over-simplified in the Nash Equilibrium (NE) strategy space as well as sub-game perfect equilibrium but their apparent conceptual simplicity masquerades the grave difficulties faced in the business world when true cooperation, collaboration or risk pooling strategies, can be effectively practised only when implicit or explicit revenue sharing issues are acknowledged. Blockbuster (video rentals) is the poster child for successful use of this GT concept and significantly increased their profits (www.wkap.nl/prod/b/1-4020-7812-9?a=1). The retailer W. H. Smith may be pursuing similar strategies with Sanford, known for their brand of 'Cross' products (pens).

Non-cooperative games with cooperative outcomes are covered in GT by Biform Games and according to some experts are probably most likely on their way to adoption due to their value-based strategy. For example, multiple retailers stock at their own locations as well as several centralized warehouses. In the first (non-cooperative) stage, retailers make stocking decisions. In the second (cooperative) stage, retailers observe demand (RFID data and demand aggregation vs volatility) and decide how much inventory to trans-ship (cross-dock) for cost/profit optimization (price elasticity). Another stage can be added in this GT model, that of, inventory procurement based on demand feedback.

For further discussion on Game Theory in supply chain management, please refer to (25) and:
(1) Albeniz, V. and Simchi-Levi, D. Competition in the Supply Option Market. Working paper, MIT, 2003
(2) Cachon, G. P. and Netessine, S. Game Theory in Supply Chain Analysis in *Supply Chain Analysis in the eBusiness Era* eds Simchi-Levi, D., Wu D. and Shen, Z. (2004) Kluwer Academic Publisher

d Prisoner's Dilemma was authored by A. W. Tucker of Princeton University [PhD advisor of John Nash]. Al Tucker was on leave at Stanford University in Spring 1950. 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 he authored *Prisoner's Dilemma*. (www.nobel.se/economics/laureates/1994/nash-lecture.pdf)

e Smoothing offers a different flavour than the model based techniques. Smoothing does not require best fitting models and do not generally produce optimal forecasts. Rather, they are simply a way to instruct a computer to draw a smooth line through data, just as we would do with a pencil. Smoothing makes no attempt to find the model that best fits the data. It forces a pre-specified model on the data. Smoothing is used if models cannot or should not be used, for example, when sample of data is very small (forecast based on four observations). Such uses are common when new products are introduced (or short life cycle). In a diametrically opposite scenario, smoothing is used when excessive data is available, for example, weekly forecasting of prices of 10,000 inputs to a manufacturing process (24). Isn't it time to put smoothing techniques to rest?

NOTES

f

Prior to 1980, the economic literature is devoid of such observations. It is interesting to speculate on the process of arousal of thought and what prepares us to think what we think what nobody has thought. Why did Robert Engle deduce or observe the "clustering" of volatilities? Engle's primary training was in physics at Cornell University. The concept of 'bursts' or 'clusters' may have strings attached to Planck's black body radiation (discontinuous, bursts) which can be explained only if light is emitted or absorbed by atoms in discrete quanta (in other words, clusters). Alternatively, Stephen Jay Gould, the Harvard evolutionary biologist, proposed the idea of "punctuated equilibria" in which he argued that evolution consisted of relatively rapid bursts of species evolution (rather than gradual, continuous transformations). Engle observed 'bursts' in financial markets: large (small) changes in asset prices are likely be followed by large (small) changes. Andy Grove, co-founder of Intel Corporation, capitalized on the idea of "punctuated equilibria" when he commented that, "economy is not a continuous process but rather a series of flash or strategic inflection points when the operative procedures may change suddenly and without warning" (Financial Times, 20 August 2003). It may pique the reader to explore the works of the Pulitzer Prize winning poet, Paul Muldoon of Princeton University. His book of poems, entitled, "Moy Sand and Gravel" provides an eclectic collection of works that 'connects the dots' from diverse observations of life and living that keeps us wondering about how we think and why?

g

To derive a frame of reference for the 'thousands' of parameter estimation, consider this: the human genome (DNA) consists of 3,000,000,000 (3 billion) units called base pairs of molecules known as Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). The order of these four molecules (A, G, C, T) is responsible for the diversity and complexity of every living organism including the human race. Thus, efforts to uncover the sequence of human DNA has been a holy grail. Around 1975, Frederick Sanger at University of Cambridge (UK), successfully deciphered the sequence of DNA for a small virus (ΦX174) that had about 5000 base pairs (A, G. C, T). It took him 4 years to sequence 5000 base pairs (bp). At that rate, it would have taken 2.4 million years to decipher the human genome. In 1988, leading scientists were convened by James Watson at the Cold Spring Harbor Lab on Long Island, NY, to discuss the human genome sequencing endeavour. Walter Gilbert of Harvard University estimated that given the advances in DNA sequencing technology, one person can sequence 100,000 bp per year and 3,000 individuals may accomplish the task of sequencing the human genome in 10 years. The most extensive ***robotic*** DNA sequencing operation was set up by Eric Lander at MIT. On 26 June 2001, President Clinton announced the completion of the sequencing of the human genome.

h

Distributed simulations harnessing the resources of multiple supercomputers have been performed in experimental environments for several years. Running these simulations still require a large amount of human investment, simultaneously reserving machines and networks, setting up the executables, parameter files and initial data in different locations. Early on, such experiments showed that such distributed simulations were possible, but the computational overhead of the wide-area network (WAN) typically lead to large performance degradations during simulation. Advances, both in infrastructure and simulation codes, are leading to improvements in coordinating simulations and performance. With such techniques applied adaptively to the changing environment, simulation performance can be dramatically increased in extremely dynamic grid environments with little or no user intervention. Hence, the paradigm shifts from 'distributed' to 'Grid' computing.

NOTES

i

Wireless network of sensors will add "eyes, ears and fingers" to the internet's core infrastructure. In addition to a host of other applications, it will impact supply chains and value networks in ways that are beyond conception, for now. Millennial Net in Cambridge, Massachusetts (USA), has developed low-power wireless transmission system, i-Bean, a stamp-size computer with digital and analog input-output ports and a wireless communications link. It may be attached to any kind of sensor — thermometer, accelerometer, pressure gauge — to transmit sensor data over a range of 30 to 100 meters. To extend the reach of i-Beans, credit-card-sized routers equipped with small antennas pick up their signals and relay them to other routers and eventually, to a gateway node (see Figure 39). This is not the internet but 'interdev' or some version of interdev (internet of devices). For example, to improve temperature control, thermometers attached to i-Beans may be scattered in a building, enabling a central computer to track and control temperatures in individual rooms without the expense of wiring thermostats. Attach an i-Bean to a intravenous (IV) pump in a hospital -- by tracking the path of its signal across routers it is possible to determine whether someone has left the IV pump in a third-floor patient area. The food and pharmaceutical industries could use i-Beans to track products through the distribution process, as well as keep tabs on data such as the temperature and humidity.

The impacts of these applications are rooted in the fact that wireless sensor networks marry communication with data gathering and control. In addition, the sensor networks may be "self-healing" or in other words may find the best path for sending data to the gateway even when sensors move or a node drops out of the network (think adaptive and ant-based algorithms). If i-Beans move out of network range, it shifts to sleep mode to conserve power (a cell phone, for example, can drain its battery quickly when it cannot find a good connection to a cellular tower). What about cost? As of January 2004, it would cost about $30 to produce a batch of 100,000 i-Beans. Millennial Net expects future development and volume manufacturing to reduce the cost to less than $10 per batch or 100 i-Beans per penny – even less than the 'penny a tag' bumper sticker prevalent in some circles!

Other noteworthy start-ups in this space are Ember (Cambridge, Massachusetts), Crossbow and Dust (Berkeley, California). Millennial and Ember are spin-offs from MIT whereas Crossbow and Dust are spin-offs from the University of California at Berkeley, CA and the Intel Research Lab at the University of California at Berkeley, CA.
http://ilp-www.mit.edu/download_file.a4d?file=P6_Library/Tech_Insider/ILPMITTI_2004_01.pdf

j

Gibbs Equilibrium was suggested by Tom Gibbs (Director, Intel Corporation) at the Board meeting of the MIT Forum for Supply Chain Innovation at MIT on 27 March 2004. It is therefore unrelated to Gibbs Free Energy (ΔG).

NOTES

k

Where rubber meets the road: Information asymmetry between planners and operators in addition to lack of data, at the right time, created problems for the US Army engaged in Iraq (partial source: Office of the Commanding General, US Army Materiel Command, Fort Belvoir, Virginia, USA). The following example reveals the disconnect and hence scope of decision system improvements necessary for military readiness.

Armoured vehicles such as the Abrams Tanks and Bradley Armoured Personnel Carriers run on 'tracks' that are layered on the exterior with heavy rubber pads to reduce wear both to the metal component (belt) of the track and to paved roads. Because the wear and tear on the rubber pads affects the mission capable status of the armoured vehicles, this operational detail is critical. Based on estimates that armoured vehicles may travel about 800 miles per year (peace-time planning), the rubber pads are ordered from the supplier, stocked and distributed for maintenance and repair (MRO). The early decision to use armoured vehicles in 2003 to storm Baghdad by crossing 1300 miles from Basra in 3 weeks, supposedly, did not stir any response from 'optempo' (operations tempo) support systems responsible for maintaining readiness and mission capable status of equipment. In this engagement, most vehicles clock a years worth of estimated mileage in a month and combat systems are running a years worth of optempo every two months. Consequently, combat systems were going through a set of track every two months that would last two years, according to planners. War reserve stocks are designed to buffer such *ad hoc* demands but funds for war reserves (in the MRO category) is not a priority in budgets. This reveals a typical information asymmetry scenario where war planners and resource planners remain disconnected. The latter increases supply chain lead time for the rubber pads and the situation holds 'readiness' as a hostage. Lack of this information caused a crisis. Ground forces abandoned otherwise perfectly capable armoured vehicles while requisitioning an increasing number of armored vehicles that had to be air-lifted. Currently, hundreds or even thousands of armoured vehicles may be 'missing' in Iraq (not official, anecdotal only).

Decisionable information for track usage was squandered and the demand signal escaped from planning. The US military industrial complex has one designated supplier of rubber pads with a production rate of a few thousand per month. This flawed supply network planning further increases the time required to reset the tracked vehicles to mission capable status, compromising readiness. A major source of rubber (raw material) for this sole source manufacturer of rubber pads is from a country mired in uncertainty due to an ongoing civil war. The secondary raw material supplier is an European firm. Taken together, the supply chain has, once again, run awry. This scenario resonates with the dependence of US scud missiles on Japanese semiconductor firms, as revealed during the Gulf War in 1991 (*The Rising Sun* by James Fallow). The repeating nature of these inadequacies in the US military (one example discussed in the article), begs to ask one simple question: are we telling a hair-raising story to a bald-headed man?

NOTES

L

EXCERPT FROM: Software-defined radio comes of age by Jeffrey Steinheider

Jeff is a member of the technical staff at Vanu Inc. He received his M. Eng. and B.S. in Electrical Engineering and Computer Science from MIT (jlstein@vanu.com). Vanu Inc was created by Dr. Vanu Bose (MIT), inventor of Vanu software radio.

In 1991, the term "software defined radio" was coined to describe radio devices implemented in software and running on generic hardware. This idea of software-programmable radio technology was a departure from traditional, hardware-specific radio architectures, and offered improvements in flexibility and upgradability. Before these multi-band, multi-mode radios could be fully implemented, however, advancements were necessary in the processing, A/D and RF hardware. These radios were not targeted for development until the year 2000.

In the beginning

Early cellular base stations were analog, with a shift to base stations built around digital hardware in the late 1980s. The evolution of SDR continued in the 1990s mostly through military interoperability endeavors. Commercial applications of SDR remained neither practical nor cost effective until the past few years. Advances in A/D converters, RF technology and processing hardware have allowed SDR to finally achieve commercial viability. SDR encompasses a wide range of communications systems, from reconfigurable hardware-based digital radios to fully programmable software radios, but can be grouped into three categories (see table 1).

The first tier

The simplest example of a SDR is a dual-mode cell phone. This is the modal SDR. A dual-mode cell phone has two hardware radios in it, one for each supported standard. Software determines which standard needs to run and activates the correct radio. Though the phone allows switching between the modes built into the radio, the user is limited only to those modes and lacks the ability to upgrade the system with new waveforms.

The next generation

Reconfigurable SDR is the category of software radio that has been built for defense applications over the last decade, typically involving a combination of processing technologies such as application-specific integrated circuits (ASIC), field progammable gate arrays (FPGA) and digital signal processors (DSP). Despite good performance of current systems, software investment for these specialized systems is high and they rapidly become obsolete as technology accelerates. One such example is the SpeakEasy system. It was built around a combination of FPGAs and 40 MHz TI C40 DSPs. By the time the first prototype was demonstrated, COTS DSPs were available at 166 MHz. As the SpeakEasy software was tied not only to the C40, but also to a specific layout of C40s and FPGAs, the new DSPs could not be exploited.

The final frontier

The most advanced type of SDR, Software Radio (SWR), maximizes software reuse across platforms and hardware generations. SWR implements the signal processing software as an application-level program running on top of a standard operating system (OS) (whether on general purpose (GP), central processing units (CPUs), DSP or other processing engines), giving it the flexibility lacking in other SDR types. The use of application-level software and an OS both reduces software development costs and allows the underlying hardware components to be upgraded without incurring the high cost of redeveloping the software. As a result, SWR systems can track the Moore's Law performance curve over time at a much lower cost than other types of SDRs.

As SDR technology progresses, the flexibility and performance of SWR will give it a clear advantage over not only traditional radio architecture, but also other SDR types. Its unique ability to add features through software upgrades and to enable a single radio to support multiple standards has drawn interest to this technology from many markets, from cellular providers to public safety agencies.

Hardware architecture

The architecture of a software defined radio can be divided into three distinct elements: a digital signal processing section, a section responsible for the conversion between RF and digital and the antenna. While the performance of the antenna and RF to digital conversion plays a key part in determining the capabilities of an SDR platform, the flexible digital signal processing is what qualifies it as a software defined radio. FPGAs, DSPs and GP processors are the three leading technologies that can provide the flexibility and processing power needed for SDR systems.

The hardware architecture groups the hardware components into three blocks representing the antenna, RF-to-digital and processing subsystems. No hardware component in the architecture is specialized to any particular waveform. While the architecture places no limitation on the achievable waveforms, any given implementation of the architecture can only support some waveforms. Each implementation supports a limited range of RF frequencies, bandwidths and amount of computational power. For example, in order for a platform to be software upgradeable from 2G to 3G cellular standards, the implementation must be able to receive a 5 MHz wide band in the appropriate frequency ranges and have enough computational power to perform the 3G processing.

The interfaces to the antenna block are RF transmit and receive analog lines and a digital control interface. With these interfaces, the architecture can accommodate traditional passive antennas (for which the digital interface has no function) as well as advanced systems such as electrically controllable antenna arrays. The architecture does not specify a particular type of digital connection (eg: RS-232), as this is a detail of the implementation.

RF-to-digital, is the only layer of the system that contains radio-specific analog components. On the receive side, its sole function is to generate a digitized representation of a down-converted slice of the radio spectrum. On the transmit side, it generates an up-converted radio signal from a digitized representation. This block does not perform waveform specific processing such as demodulation or equalization.

The third block, motherboard, is borrowed from the PC because software radios look much more like computers than like legacy radios. Like a PC, this layer contains memory and processor components and provides I/O to a network, to the user, timing support and similar functions.

Applications

SDR technology can be used in any device that uses RF for communication, which encompasses a wide range of products including cellular base stations, military communications systems and public safety radios.

Technology in cellular base stations

Cellular standards evolve slowly, from analog in the 1980s to digital in the 1990s and possibly to 3G sometime this decade. While the underlying processing, communications and DSP technology evolves rapidly, cellular service is limited to once-a-decade upgrades because the high capital costs of infrastructure upgrades are prohibitive.

For example, AT&T and Cingular are upgrading their networks from time-division multiple access to (TDMA) to the global system for mobile communications (GSM). This “upgrade” actually involves building out a new GSM network in parallel to their existing TDMA network, an initiative that costs each carrier upwards of $4 billion and requires a 10-year deployment to achieve a reasonable return on investment.

A wireless network infrastructure using software radio technology can be software upgraded to new standards, thus deploying new standards more quickly and at lower cost than today's approach. Carriers can then increase revenue by rapidly implementing new revenue generating services as well as new systems that use spectrum more efficiently. A further benefit of SDR is reduced operating expenses — many of the maintenance and upgrades today that require truck convoys to tower sites could be serviced as remote software changes in an SDR system.

The architecture for a SDR base station is essentially a basic SDR with an array of processing elements that can be scaled to handle more capacity or more complex waveforms. Using current x86 general purpose processors as an example, it is now possible to provide one GSM channel with 8 time slots for every 1 GHz of processing. Standard networking equipment such as gigabit Ethernet now has the bandwidth to supply digitized spectrum and allows the use of standard PC servers with x86 processors to act as the cluster of processing units. The radio section of a software radio base station is responsible for converting a wide band of radio spectrum to a digital IF. This equipment is available today in the form of multi-carrier power amplifiers, wide-band up-converters and down-converters and high-speed A/D and D/A converters. This provides a digital interface that is completely independent of the air standard and able to support multiple channels of different standards in a band. When coupled with a SDR backend, it is possible to change air standards simply through a software upgrade.

SDR eases standards woes

SDR can also mitigate problems carriers face when switching to a new standard. Generally, capacity is moved to the new standard slowly, so that customers are not forced to immediately upgrade their phones. Limited spectrum availability means the carrier must decide at some point to take away capacity from the old standard in order to add capacity to the new standard. A SDR base station can run two different air standards simultaneously, operating a control channel for each standard and saving an operator from having to make this decision at each tower. Additional capacity then can be added to each standard on an as-needed basis, changing the number of channels used by each standard dynamically, depending on the number of users requiring voice channels for each standard.

SDR and frequency allocation

Base station hotelling is a new architecture for deploying cellular systems that takes advantage of SDR's flexibility to lower capital costs and make more efficient use of the spectrum. Companies are now separating the base station from the antennas in order to improve coverage in urban areas and add coverage to tunnels, stadiums and within buildings by putting the antennas where they are most needed. These remote antennas provide the RF spectrum over a fiber optic cable back to a central location where all of the base station processing resides. This method also better utilizes base station resources, as channels can be allocated to different locations to match the load as it varies over the course of a day. For example, at rush hour, more resources can be applied to towers on the highway, whereas these same processing resources could be allocated to the downtown office area at other times of the day. It is no longer necessary to outfit towers with capacity for the peak load — capacity that will sit idle during off-peak hours. Adding capacity to the entire system is now as easy as adding a server to a rack in the central location, eliminating a trip to the tower.

Additionally, the benefits of a SDR base station still apply. It is possible to run multiple standards simultaneously from a single hotel site using the same hardware, even supporting multiple wireless services providers from the same infrastructure base. This ability to share infrastructure between standards or carriers greatly reduces capital costs for the providers.

Military

Interoperability problems are also an obstacle in joint operations, where each nation typically has its own radio systems. Recently, emphasis on peacekeeping, disaster relief, homeland security and other non-combat military operations has created further problems. In these roles, military units must communicate with public safety agencies, humanitarian organizations and civilians. Single SDR with the ability to support multiple waveforms significantly reduces the number of devices needed in the field. For military users, who must maintain, transport, power and manage each device under challenging conditions, the benefit of a streamlined system is substantial.

SDR promises to reduce military radio development and acquisition costs. Without SDR, new device development requires investing anew in the implementation of each supported communication standard. With SDR, the bulk of the implementation knowledge for a communication standard is captured in portable software, which can then be reused at low cost in new or different platforms. This software reuse holds the potential to revolutionize radio procurement by significantly increasing competition among platform vendors, leading to reduced per-unit costs.

US DOD recognizes the potential cost reduction of SDR and has established the Joint Tactical Radio System Joint Program Office (JTRS JPO) to achieve that goal. The JPO has begun to acquire software implementations of a first set of 33 communication standards. The linchpin of the JTRS effort is a standard - the software communications architecture (SCA) - intended to ensure portability of the implementations across platforms from many vendors. SCA standardizes the software's operating environment and the control and communication mechanisms for both the hardware and the external interfaces of the radio. Many NATO allies have signed agreements to apply the SCA in their future acquisitions. JPO expects SCA to become the basis for commercial SDR software standards as well.

The Federal Emergency Management Agency has identified radio interoperability as the one item that could have made the most significant difference in the rescue and cleanup effort after the 9/11 disaster. The unplanned nature of an emergency requires extremely flexible radio systems that are able to adapt to the situation and diverse communications needs, making SDR the ideal technology for public safety radio systems.

Conclusion

Software Defined Radio has been in development for many years, but is only now achieving commercial viability thanks to advances in integrated circuits. Initial SDR systems will likely appear in specialized areas such as military and public safety applications; visibility in consumer markets will follow. Eventually, as SDR is deployed in infrastructure and consumer devices, consumers will benefit from improved wireless coverage and new services. Digital processing exists today with enough capacity to handle many different waveforms but SDR is limited by specialized RF chipsets that are optimized for particular frequency bands and waveforms. When RF designs are developed that are tunable over broad frequency ranges and can handle several bandwidths, SDR will be able to demonstrate all of its advantages.

NOTES

M

Convergence of principles from artificial intelligence (AI) research to address problems associated with decision systems, such as supply chain management, may yield some spectacular gains in efficiency, hence, profitability. It is the objective of this note to merely point out some related research from AI that may have such an impact. Proceedings of the 18th National Conference in AI and 14th Conference on Innovative Applications of AI sponsored by the American Association for Artificial Intelligence, July 28 – August 1, 2002, Edmonton, Alberta, Canada (AAAI Press / MIT Press; ISBN 0-262-51129-0)]. The following papers are a few of many that may be of relevant interest:

Representing and Reasoning about Mapping between Domain Models (pages 80-86)
Jayant Madhavan jayant@cs.washington.edu, Philip A. Bernstein philbe@microsoft.com,
Pedro Domingos pedrod@cs.washington.edu, Alon Y. Halevy alon@cs.washington.edu

Regression Based Adaptation Strategy for Case-Based Reasoning (pages 87-92)
David Patterson wd.patterson@ulster.ac.uk, Niall Rooney nf.rooney@ulster.ac.uk
Mykola Galushka mg.galushka@ulster.ac.uk

State Abstraction for Programmable Reinforcement Learning Agents (pages 119-125)
David Andre dander@cs.berkeley.edu, Stuart J. Russell russell@cs.berkeley.edu

Data Perturbation for Escaping Local Maxima in Learning (pages 132-139)
Gal Elidan galel@cs.huji.ac.il, Matan Ninio@cs.huji.ac.il, Nir Friedman nir@cs.huji.ac.il
Dale Schuurmans dale@cs.uwaterloo.ca

Pruning and Dynamic Scheduling of Cost-Sensitive Ensembles (pages 146-151)
Wei Fan weifan@us.ibm.com, Fang Chu fchu@cs.ucla.edu, Haixun Wang haixun@us.ibm.com
Philip S. Yu psyu@us.ibm.com

Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers (pages 167-173)
Russel Greiner greiner@cs.ualberta.ca, Wei Zhou w2zhou@math.uwaterloo.ca

Reinforcement Learning for POMDPs based on Action Values and Stochastic Optimization (pages 199-204)
Theodore J. Perkins perkins@cs.umass.edu

Constrained Formulations and Algorithms for Stock-Price Predictions Using Recurrent FIR Neural Networks (pages 211-216)
Benjamin W. Wah wah@manip.crhc.uiuc.edu, Minglun Qian m-qian@manip.crhc.uiuc.edu

POMDP Formulation of Preference Elicitation Problems (pages 239-246)
Craig Boutilier cebly@cs.toronto.edu

Context-Specific Multi-agent Coordination and Planning with Factored MDPs (pages 253-259)
Carlos Guestrin guestrin@cs.stanford.edu, Shobha Venkataraman shobha@cs.stanford.edu
Daphne Koller koller@cs.stanford.edu

Continued **-** References for **Notes M** from Proceedings of the 18th National Conference in Artificial Intelligence

Nearly Determinisic Abstractions of Markov Decision Processes (pages 260-266)
Terran Lane terran@ai.mit.edu, Leslie Pack Kaelbling lpk@ai.mit.edu

Size of MDP Factored Policies (pages 267-272)
Paolo Liberatore paolo@liberatore.org

On Policy Iteration as a Newon's Method and Polynomial Policy Iteration Algorithms (pages 273-278)
Omid Madani madani@cs.ualberta.ca

Piecewise Linear Value Function Approximation for Factored MDPs (pages 292-299)
Pascal Poupart ppoupart@cs.toronto.edu, Craig Boutilier cebly@cs.toronto.edu
Relu Patrascu rpatrasc@cs.uwaterloo.ca, Dale Schuurmans dale@cs.uwaterloo.ca

Value Iteration Working with Belief Subset (pages 307-312)
Weixiong Zhang, Nevin L. Zhang

Design of Collectives of Agents to Control Non-Markovian Systems (pages 332-337)
John W. Lawson Lawson@ptolemy.arc.nasa.gov, David H. Wolpert dhw@ptolemy.arc.nasa.gov

Planning with a Language for Extended Goals (pages 447-454)
Ugo Dal Lago dallago@irst.itc.it, Marco Pistore pistore@irst.itc.it, Paolo Traverso traverse@irst.itc.it

Symbolic Heuristic Search for Factored Markov Decision Processes (pages 455-460)
Zhengzhu Feng fengzz@cs.umass.edu, Eric A. Hansen hansen@cs.msstate.edu

Plan Evaluation with Incomplete Action Descriptions (pages 461-467)
Andrew Garland garland@merl.com, Neal Lesh lesh@merl.com

Algorithms for a Temporal Decoupling Problem in Multi-Agent Planning (pages 468-475)
Luke Hunsberger luke@eecs.harvard.edu

Searching for Backbones and Fat: A Limit-Crossing Approach with Applications (pages 707-712)
Sharlee Climer sclimer@cs.wustl.edu, Weixiong Zhang zhang@cs.wustl.edu

Stochastic Link and Group Detection (pages 794-804)
Jeremy Kubica jkubica@cs.cmu.edu, Andrew Moore awm@cs.cmu.edu, Jeff Schneider schneide@cs.cmu.edu
Yiming Yang yiming@cs.cmu.edu

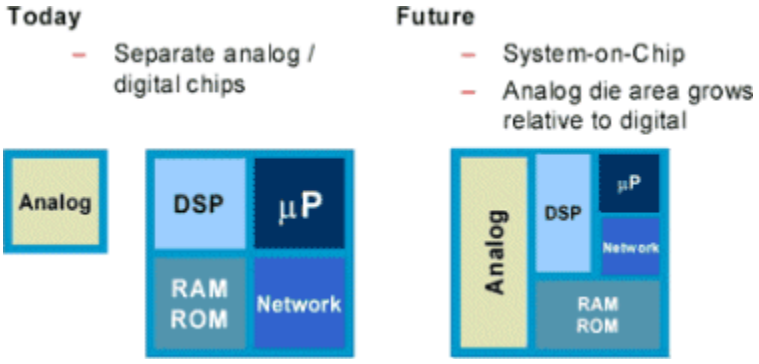
Learning in Open-Ended Dynamic Distributed Environments (page 980)
Doina Caragea dcaragea@cs.iastate.edu

Perspectives on Artificial Intelligence Planning (pages 1013-1023)
Hector Geffner hector.geffner@tecn.upf.es

NOTES
N

How evolving engineering principles may impact RFID tags: **Self-Adaptive Silicon** (summary from www.impinj.com)

While we await availability of printed plastic RFID tags from Richard Friend's Plastic Logic (www.plasticlogic.com and www-oe.phy.cam.ac.uk/PEOPLE/OESTAFF/rhf10.htm), at hand we have Chris Diorio's Self-Adaptive Silicon that uses transistor physics in a new way, enabling precision analog and wideband RF in low cost, high density logic CMOS. This technology enables mixed-signal System-On-Chip (SoC) products.



Self-Adaptive Silicon originates by rethinking the physics of floating-gate p-channel MOSFETs (pFETs). pFETs are one of the two types of transistors in Complementary MOS (CMOS) processing; the other is the n-channel MOSFET (nFET). Floating gates are typically associated with FLASH or EEPROM nonvolatile memory (NVM) technology, which adjusts the electronic charge on an nFET floating gate to store one of two digital values. Self-Adaptive Silicon rethinks both the pFET physics and the floating-gate physics, to enable local adaptation in silicon. This technology differs from traditional floating-gate technology, in two ways:

1. Impinj fabricates floating-gate devices in standard logic CMOS (with no additional process masks)
2. Impinj's floating-gate MOSFET remains a fully functional transistor during memory updates, allowing it to store precise analog values on the floating gate.

The end result is a tunable transistor with a nonvolatile (permanent) analog memory that is used to implement adjustable voltage or current sources, timing delay elements and a host of other analog circuit blocks. Self-Adaptive Silicon delivers two key benefits for analog technology:

1. Precision analog design using Impinj's technology is simpler than traditional methods because the circuits electrically tune themselves after fabrication
2. Impinj's circuits can continually adapt over their lifetime, maintaining performance despite component aging and variations in temperature and supply voltage.

www.impinj.com

Notes
O

Modified from: Han Pang Huang, National Taiwan University

What is *'new'* about RFID ? Evolution of RFID

1940	1950	1960	1970	1980	1990	2000
RFID born out of Radar effort (WWII)	**RFID crawls out**	**Theory of RFID, field trials planned**	**Early adopters implement RFID**	**Commercial RFID endeavors sprout**	**Many RFID standards emerge**	**RFID hype, peaks**
1948 Harry Stockman invents RFID. Publishes paper, "Communication by Means of Reflected Power"	1950 D.B. Harris patents RFID. "Radio transmission systems with modulatable passive responder" 1952 F.L. Vernon "Application of the microwave homodyne" 1959 Identification of Friend or Foe (IFF) long-range transponder system reaches breadboard demonstration stage.	1963-1964 R.F. Harrington advances theory with "Field measurements using active scatterers" and "Theory of loaded scatterers" 1966 Commercialization of EAS, 1-bit Electronic Article Surveillance	1973 Raytheon's "Raytag" 1977 RCA develops "Electronic identification system" 1975 Los Alamos National Lab (LANL) releases RFID research to public sector, publishes "Short-range radio-telemetery for electronic identification using modulated backscatter" 1976-1977 LANL RFID spin-offs Indentronix and Amtech 1975-1978 Raytheon, Fairchild & RCA develop RFID	1982 Mikron founded; bought by Philips 1987 First RFID road toll collection implemented in Norway	1991 TI creates TIRIS to develop and market RFID 1992-1995 Multi-protocol traffic control and toll collection implemented in Texas, Oklahoma, and Georgia (USA) 1998 David Brock and Sanjay Sarma of MIT publishes an idea: 'Internet of Things' 1999 Auto ID Center created at MIT. Retailers drive to standardize EPC	2003 UPC and EAN forced by US retailers to promote EPC 2005 Wal-Mart and US DoD fuels the hype curve by demanding suppliers use passive RFID and EPC.

Vast number of RFID companies and 'short-sight' enters the market.

Partial Source: Shrouds of Time – The History of RFID

REFERENCES

1. Gell-Mann, M. The Quark and The Jaguar. W. H. Freeman and Company, New York, 1994.
2. Heinrich, C. and Betts, B. Adapt or Die: Transforming Your Supply Chain into an Adaptive Business Network. John Wiley and Sons, 2003.
3. Tellis G.J. and Golder P.N. First to Market First to Fail? Real Causes of Enduring Market Leadership. Sloan Management Review (1996) 37(2) 65-75
4. www.wrmea.com/archives/sept-oct02/0209044-2.html; www.bea.doc.gov
5. Simchi-Levi, D., Kaminsky, P. and Simchi-Levi, E. Designing and Managing the Supply Chain. 2nd ed, 2002
6. Forrester, J. Industrial Dynamics, MIT Press, 1961
7. Sterman, J.D. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Management Science (1989) 35 321-339
8. Lee, H., Padmanabhan, P. and Whang, S. The Bullwhip Effect in Supply Chains. Sloan Management Review (1997) 38 93-102
9. Bramel, J. and Simchi-Levi, D. The Logic of Logistics. Springer, 1997.
10. Joshi, Y. V. Information visibility and its effect on supply chain dynamics. MS Thesis, MIT, 2000.
11. Cohen, M., Ho, T., Ren J. and Terwiesch, C. Measuring imputed cost in the semiconductor equipment supply chain. Working Papers. Wharton School, 2003
12. Neumann, J. V. and Morgenstern, O. Theory of Games and Economic Behavior, Princeton Univ Press, 1947
13. Nash, J. F. Equilibrium points in N-Person Games. PNAS, 1950
14. Kuhn, H. W. Extensive games and the problem of information in H. W. Kuhn and A. W. Tucker, eds, Contributions to the Theory of Games. Princeton University Press, 1953
15. R Aumann. Acceptable points in general cooperative n-person games. Contributions to the Theory of Games (IV). Princeton University Press, 1959
16. Shubik, M. Some Experimental Non Zero Sum Games with Lack of Information about the Rules. Management Science (1962) 8 215-234
17. Shubik, M. Incentives, Decentralized Control, the Assignment of Joint Costs and Internal Pricing. Management Science (1962) 8 325-343
18. Vickrey, W. Counterspeculation and Competitive Sealed Tenders. Journal of Finance (1961) 16 8-37
19. http://william-king.www.drexel.edu/top/eco/game/ game.html
20. Ozalp Ozer, Stanford University (personal communication)
21. Enders, W. Applied Econometric Time Series. Wiley, 2nd edn, 2003
22. Christensen, C. The Innovators Dilemma (pp 168-169) 2000
23. DeHoratius, N. http://gsbwww.uchicago.edu/news/capideas/summer02/measuremanage.html
24. Diebold, F. Elements of Forecasting. 2nd edn, 2001
25. Datta, S. *et al* www.wkap.nl/prod/b/1-4020-7812-9?a=1
26. Datta, S. http://supplychain.mit.edu/innovation/research-starch.htm
27. Engle, R.F. New Frontiers for ARCH Models. Journal of Applied Econometrics (2002) 17 425-446
28. Foster, I., Kesselman, C. and Tuecke, S. International Journal of Supercomputer Applications (2001) 15
29. Granger, C.W.J. and Swanson, N.R. Further development in the study of cointegrated variables. Oxford Bulletin of Economics and Statistics (1996) 58 374-386

REFERENCES (continued)

30. http://www.nobel.se/economics/laureates/2003
31. Engle, R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica (1982) 50 987-1007
32. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics (1986) 31 307-327
33. Engle, R.F. and Lange, J. Measuring, Forecasting and Explaining Time Varying Liquidity in the Stock Market. 1997. Discussion Paper 97-12R. University of California, San Diego.
34. www.eviews.com
35. Johnston, W. Computational and Data Grids in Large-Scale Science and Engineering. Future Generation Computer Systems, 2002.
36. Allen, G., Seidel, E. and Shalf, J. Scientific Computing on the Grid. Byte (2002) 24-32
37. www.ipg.nasa.gov; www.Gridforum.org
38. Minsky, M. The Society of the Mind. Simon & Schuster. 1999.
39. http://www.w3.org/2001/sw/
40. Dertouzos, M. The Unfinished Revolution. Harper Collins, 2001.
41. Gershenfeld, N. When Things Start To Think. Owl Books, 1999.
42a. Parunak, H. Go to the Ant: Engineering Principles from Natural Multi-Agent Systems. Annals of Operations Research (1997) 75 69-101
42b. Parunak, H., Savit, R. and Riolo, R. Agent-Based Modeling versus Equation-Based Modeling. Proceedings of Multi-agent Systems and Agent-based Simulation, 1998.
43. Bradshaw, J., eds. Software Agents. MIT Press, 1997.
44. Shehory, O., Sycara, K. and Sukthankar, G., Agent-aided Aircraft Maintenance. Proceedings of Autonomous Agents (1999) pp 306-312
45. Sarma, S. and Brock. D. The Internet of Things. White Paper. Auto-ID Center, MIT, 1998.
46. www.autoidcenter.org
47. Scanlon, L. Behind Bars, MIT Technology Review. April, 2003.
48. Auto ID Center, Technical Report. 2002.
49. www.aetherwire.com/CDROM/Welcome.html
50. www.multispectral.com/pdf/APPsVGs.pdf
51. http://grouper.ieee.org/groups/802
52. http://nms.lcs.mit.edu/~hari/; www.cs.berkeley.edu/~culler/
53. www.csail.mit.edu/research/abstracts/abstracts03/web/02berners-lee.pdf
54. http://www.cs.umd.edu/~hendler/
55. Chandrasekharan, S. (2002) Semantic Web: a distributed cognition view. Technical Report 2002-13, Ottawa.
56. Toivonen, S. (2004) Distributing Cognition in the Semantic Web (personal communication)
57. Microarray http://supplychain.mit.edu/innovation/shoumen.htm (#2)
58. Fine, Charles. 1998. Clockspeed. Perseus Books.
59. Agents http://supplychain.mit.edu/innovation/shoumen.htm (#9)
60. Radio Frequency Identification http://supplychain.mit.edu/innovation/shoumen.htm (#7)
61. RFID Made Easy … an incomplete story http://supplychain.mit.edu/innovation/shoumen.htm (#25)
62. Metabolomics www.vbi.vt.edu/~mendes
63. Global Public Goods (#29) & Hydrogen Economy (#24) http://supplychain.mit.edu/innovation/shoumen.htm

In praise of confluence ?

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.

Jeff Wilke

Senior Vice President
Amazon.com

---

17 June 2004