Study of Big Data Analytics Landscape:

Considerations for Market Entry of an E-commerce Analytics Vendor

By
Soumya Shukla

Bachelor of Technology, Production Engineering
National Institute of Technology, Gujarat, 2007
Post Graduate Programme in Management
Indian School of Business, Hyderabad, 2012

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN MANAGEMENT STUDIES
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUNE 2016

Signature of Author: ____ Signature redacted ____
MIT Sloan School of Management
May 13, 2016

Certified by: _Signature redacted____
Sinan K. Aral
David Austin Professor of Management
Thesis Supervisor

Accepted by: _________ Signature redacted ____
Rodrigo Verdi
Associate Professor of Accounting
Program Director, M.S. in Management Studies Program
MIT Sloan School of Management

Study of Big Data Analytics Landscape: Considerations for Market Entry of an E-commerce Analytics Vendor

By
Soumya Shukla

Submitted to MIT Sloan School of Management on May 13, 2016 in partial fulfillment of the requirements for the Degree of Master of Science in Management Studies.

Abstract

In today's age of 'Information Explosion' most companies are struggling with optimally utilizing all the data generated. The last two decades have seen tremendous progress in data collection, storage, processing and visualization technologies. The last decade has seen a remarkable growth in the types of user data: social, web and more recently mobile. As newer sources of data emerge, our ability to separate an individual from a segment improves. The data being analyzed is not just structured in nature. Several processing technologies are analyzing unstructured data for insights. Emerging technologies based on machine learning are improving our ability to migrate decision making from discovery & diagnostics to prediction & preemption. The rise of Internet of Things enhances the opportunity to collect further granular data. At the same time, as system efficiency increases, concerns about privacy loss and malpractices also increase. The world of big data is more complex and controversial than ever before.

This study focuses on creating a baseline of big data technologies and attempts to identify near term trends within the horizon of 5 years due to the pace of technological development. The study places special emphasis on E-commerce Sales & Marketing analytics to determine current challenges and develops a key considerations framework for a new entrant in that space.

Thesis Supervisor: Professor Sinan K. Aral
Title: David Austin Professor of Management

To my brother Shikhar, who is always the first to pick me up after a fall and the last to stop laughing about it

TABLE OF CONTENTS

1. INTRODUCTION

The Google search results for keywords 'Big Data Analytics' generate over 101 million results in 0.51 seconds. Experts and industry pundits agree on the utility of using data for generating business insights. Managerial decision making in the 21st century has been largely claimed to be data-driven and the role of a data scientist is presently considered the 'hottest' job in Silicon Valley.

Yet, time and again, it becomes evident that most large corporations are yet to optimally utilize the power of big data. One of the most interesting quotes on the use of big data is by Dan Ariely, Professor of Psychology and Behavioral Economics at Duke University – "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it"

The purpose of this thesis is to explore the field of big data analytics, assess current technological breakthroughs and forecast future outlook. One of the key focus areas for this thesis is unstructured data analytics. Various estimates indicate that upto ~70% (Stewart D., 2013) of all data generated and stored is unstructured. This includes images, videos, text and audio. The wealth of information stored in this data stands to improve insight generation significantly. This thesis examines current technologies, key players and users of unstructured data for insights. To investigate data analytics in depth, some of the important concepts covered in the scope of the thesis are Enterprise Resource Planning (ERP), Machine Learning & Artificial Intelligence, Rise of Mobile, Social Media, Internet of Things, Privacy Concerns and the Chinese Internet industry. The objective behind investigating these areas is to be able to create a reasonable view of the future of the data driven world.

At the outset, it is important to set the definitions of a few key elements under examination throughout this thesis. The objective is to ensure that the reader's understanding of these elements align with author's view while framing the narrative.

Business Analytics: The science of examining existing information for the purpose of drawing inferences and conclusions to aid business decision making. The process involves data collection, cleaning, processing, visualization and storage. In order to derive meaningful patterns in data, analytics relies on various application areas such as mathematics, statistics, computer programming, operations research and many others, to quantify insights. The objective of facilitating business decision-making is to enable better business performance. Predictive and prescriptive analytics are widely used for enterprise decision management. Additionally, analytics can be used for fraud prevention and risk management as well.

Machine Learning: An emerging field of artificial intelligence that allows computer algorithms to autonomously learn from newer data and use the learning for self-improvement. This has been one of the most important developments in the last 5 years, with ubiquitous application across e-commerce, ride sharing apps, advertising technology and social media. Newer applications are emerging rapidly such as self-driving cars, cancer research, and identifying terror suspects. It is a widely held notion that machine learning would be one of the key driving forces in analytics in the next decade.

Internet of Things: The concept of integrating our physical world with the invisible world of electronics. IOT aspires to embed everyday physical objects with electronic sensors that not only collects user data, but also enables communication between various other physical objects. IOT includes seemingly futuristic but achievable technologies such as smart grid, smart cities, intelligent transportation et cetera.

2. A Brief History of Tech-Time

The first decade of the 21st century, the period between 2000 and 2010, witnessed an exponential increase in the amount of new data generated. This 'Information Explosion' can be largely attributed to a rapid rise in the adoption of Internet around the world, introduction of social media, large-scale adoption of enterprise data management solutions. In the middle of 2016, data evolution seems remarkable in contrast with business intelligence and business statistics only 30 years ago.

Zooming in, the last 20 years have seen an unprecedented rise in data collection and analytics. The inventions and innovations in data storage, analytics and data intelligence come together in a unique combination that has lead to a revolution in the way businesses are managed.

It is worthwhile to review some of the important milestones in the field of data analytics, to gain an understanding of this data revolution. The timeline below captures major events in the Internet industry, and technological advancements that changed the course of data analytics.

Year 1997

- Google search is launched
    - Google's search engine managed to change the way data was handled on the Internet. It enabled access to relevant information and spurred the beginning of a technological revolution in the advertising industry
- IBM's Deep Blue (artificial intelligence) beats the world champion at Chess
- Formal attempts are made to quantify new data generation in light of internet adoption
    - Michael Cox and David Ellsworth publish a paper titled 'Application-controlled demand paging for out-of-core visualization'. The term 'big data' is coined in the paper that highlights the need for improving storage and management resources in light of the relentless increase in data generation
    - Michael Lesk published, "How much information is there in the world?" The paper accurately predicted the need for data analytics in the near future – "When we reach a world in which the average piece of

information is never looked at by a human, we will need to know how to evaluate everything automatically to decide what should get the precious resource of human attention."

Year 1998

- The term 'Big Data' gains widespread use
- Industry experts and academicians continue their focus on the quantum of data generated and the need to enhance current data storage capability of the world
  - Some of the notable papers were published by, John R. Masey ("Big Data… and the Next Wave of Infrastress."), K.G. Coffman and Andrew Odlyzko ("The Size and Growth Rate of the Internet.")

Year 1999

- The focus on the need to innovate data storage spurred the evolution of SSDs and HDDs. Inventions in Cloud Computing, portable storage devices and microprocessor chips pave the path for ensuring adequate and affordable storage
- The focus begins to shift from data storage to data visualization. Several papers are published debating the merits and demerits of various data visualization techniques

Year 2000 & 2001

- Dot-com collapse of 2000 saw several Internet businesses go bust. Many of the businesses that survived the crash are now giant corporations such as Amazon, eBay, Yahoo, Netflix etc.
- This period is also significant because of the September 11 attacks on USA. The Patriot Act was signed in October 2001. Many privacy activists claim that the signing of this act was the beginning of the end of privacy in the world
- The term 'Software-as-a-service' is coined in the article 'Strategic Backgrounder: Software as a Service by the Software and Information Industry Association'
- Ad Networks emerged as online advertising became more and more data driven

Year 2005

- Facebook.com is launched in August 2005. The service is extended to US high school students, Ivy league schools and a few universities outside US (such as Oxford)
- Hadoop – the open source framework for storing and analyzing Big Data sets, is launched. It offers great flexibility and is found to be particularly useful for the management of unstructured data (voice, video, images, text)
- Amazon launches Prime, a flat fee membership model with no additional shipping fee

Year 2008

- Facebook has 100 Million users worldwide
- iPhone 3G is launched, sparking the beginning of the age of touch based smartphones
- Android v1 is launched as open source operating system for touch based mobile and tablets
- AirBnB is founded. To be launched in the following year, sparking one of the first large scale applications of the concept of shared economy
- The definition of big and small data is observably changing. Microsoft excel is launched with the ability to carry 1 M rows; a significant jump from 65K rows in the previous version
- Estimates indicate that servers were processing 9.57 zettabytes (9.57 trillion gigabytes) of information worldwide. This indicates the amount of data processed is equivalent to 12 gigabytes per person, per day (Short J. E. et al, 2011)
    - It also estimated that 14.7 exabytes of new data was produced this year (Bounie, D. & Gille L., 2012)

Year 2010

- An average company in the US with over 1,000 employees is estimated to be storing more than 200 terabytes of data (Manyika, et al., 2011)
- Uber is launched in San Francisco Bay Area
- Microsoft Azure Data Marketplace launched
- Google exited China
    - Google forgoes an enormous market due to government censorship concerns

- This lead to strengthening of the Chinese social networks and search services. As a result, Baidu, Alibaba and Tencent are the grand trinity of China today

Year 2011

- Uber begins US-wide expansion
- Uber begins international expansion, starting with Paris
- AirBnB begins international expansion by setting up offices in Hamburg and London
- IBM's Watson beats human competitors in the game of Jeopardy

Year 2012

- Facebook crosses 1 B users worldwide
- Facebook acquires Instagram for $1 B, gaining access to images posted by millions of users
- Facebook launched its IPO on NYSE
- Mobile dating service, Tinder is launched, changing the face of online dating globally
- iOS and Android device adoption surpasses that of any consumer technology in history (Fargo P., 2012)

Year 2013

- Edward Snowden leaked US National Security Agency's snooping program details in June
- Twitter launched its IPO on NYSE
- Microsoft acquired Nokia's phone assets for $7.9 B
- Wearables market saw a few breakthroughs with Samsung's galaxy watch, Fitbit's improved activity tracker and Google glass
- IBM Watson's first commercial application is launched to assist in lung cancer treatment at Memorial Sloan Kettering Cancer Center
- Facebook turned profitable

Year 2014

- Facebook acquires Whatsapp for $19 B thereby gaining access to private conversations between millions of Whatsapp users

- Facebook acquires Oculus in July, becoming one of the major player in virtual and augmented reality space alongside Microsoft's Hololens and many others
- Tinder hits 1 B matches worldwide
- Google acquired Nest Labs for $3.2 B, a company specialized in designing home devices such as thermostat and smoke alarms, to strengthen IOT presence

Year 2015

- Facebook has 1.59 B users worldwide
- Uber announces investments in self-driving cars
- 2.8 B internet users worldwide with Asia (including China) accounting for over 50% (Meeker M., 2015)
- 5.2 B mobile phone users worldwide of which 40% use smartphones (Meeker M., 2015)
- Apple watch launched in April
- Amazon Machine Learning is launched by Amazon Web Services
    - Aimed towards developers for building predictive applications
- Uber's competitors Lyft (United States), Didi Kuaidi (China), Ola Cabs (India), and GrabTaxi (South-East Asia) join hands in a global technology and service alliance
- Microsoft's Distributed Machine Learning Toolkit is launched. It enables efficient distribution of machine learning problems across multiple computers, thereby increasing the scope and speed of machine learning
- Microsoft wrote off $7.6 B and stated that Nokia's acquisition was a strategic error

Year 2016

- Google's AI algorithm beats a professional at Chinese board game Go. Go is considered the world's most complex board game. It also cited as many times harder than chess. The AI algorithm, AlphaGo, is developed by Google DeepMind. It won five games out of five in the competition
- Whatsapp launches end-to-end data encryption of private texts. This is a fairly advanced step towards privacy protection

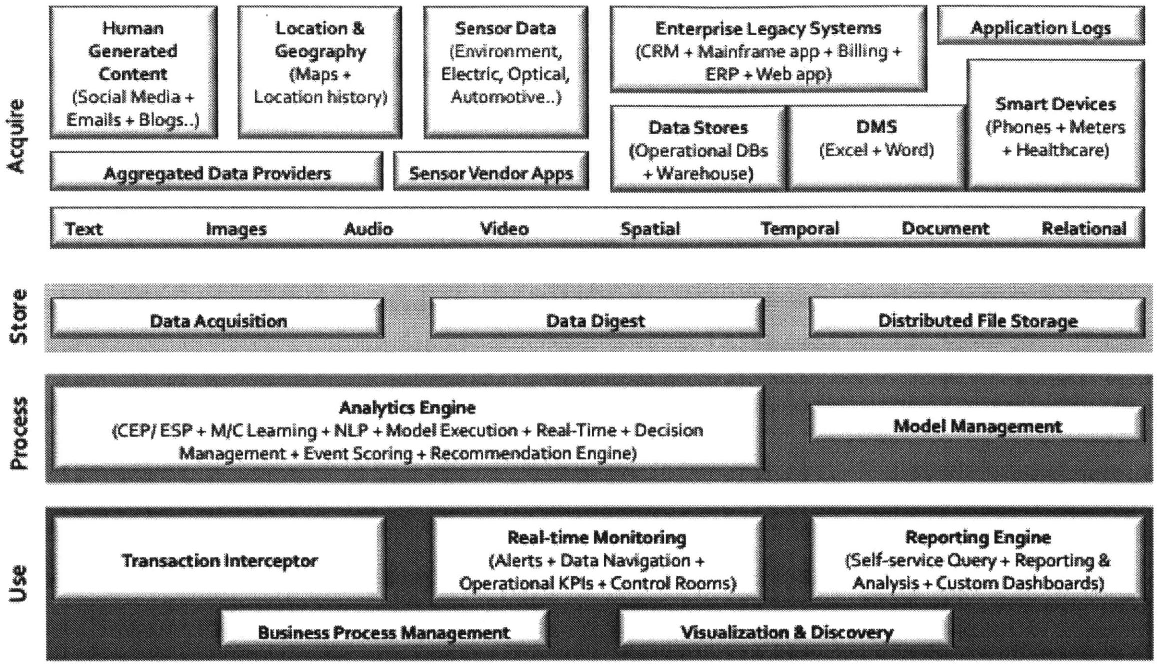
3. Building blocks of data analytics

As data is getting bigger and bigger, solutions to acquire, store, process and use data are getting more innovative. The last 15 years have seen some of the most important innovations in big data analytics. These innovations were brought forth equally by large players such as IBM, SAS, SAP and young companies.

Perhaps the most interesting transformations have come in the data acquisition phase. In the pre-internet era, data sources were fairly limited to log files maintained by companies for production records, sales data, supply chain and inventory records, human resources data, customer service logs and a few others. Enterprise software by companies such as IBM and SAP enabled business intelligence and several attempts were made to cross-categorize and use the data for insight generation. With the growing interconnectedness of the world, businesses grew more interconnected too. Data was observably generated in quantities higher than ever before. Several new sources of data emerged, especially as Internet adoption increased and newer devices such as mobile phones and wearables gained popularity. As was mentioned in the timeline above, the last few years of the last century were largely focused on data storage, in anticipation of the massive rise in data generation. The last decade was focused on understanding the new data sources as web 2.0 took shape and the present decade is dedicated towards better processing.

The figure below is a schematic view of a standard enterprise level big data analytics framework.

Figure 1: Big Data Infrastructure


Acquire
Human Generated Content (Social Media + Emails + Blogs..)
Location & Geography (Maps + Location history)
Sensor Data (Environment, Electric, Optical, Automotive..)
Enterprise Legacy Systems (CRM + Mainframe app + Billing + ERP + Web app)
Application Logs
Smart Devices (Phones + Meters + Healthcare)
Data Stores (Operational DBs + Warehouse)
DMS (Excel + Word)
Aggregated Data Providers
Sensor Vendor Apps
Text
Images
Audio
Video
Spatial
Temporal
Document
Relational
Store
Data Acquisition
Data Digest
Distributed File Storage
Process
Analytics Engine (CEP/ ESP + M/C Learning + NLP + Model Execution + Real-Time + Decision Management + Event Scoring + Recommendation Engine)
Model Management
Use
Transaction Interceptor
Real-time Monitoring (Alerts + Data Navigation + Operational KPIs + Control Rooms)
Reporting Engine (Self-service Query + Reporting & Analysis + Custom Dashboards)
Business Process Management
Visualization & Discovery

Source: IBM Big Data and Analytics

3.1 Data Acquisition

In this section, we will examine the various sources of data acquisition, by a business and the way these sources are evolving over time. The focus is largely on current and future sources of data while broadly attempting to classify these sources. The types of data may find high or low application based on the type of businesses that use it. However, irrespective of the type of business, application areas are rather consistent. For example, application of big data for better inventory management finds similar use in vastly different industries such as pharmaceutical manufacturing and online commerce. The underlying principle holds across industries and thus, for the sake of simplicity and ease of classification, business applications are categorized as follows

- **Research & Development**: In-house innovations lead by industry trends, user preferences, adjacent industry, new inventions, new technology adoption et cetera
- **Operations**: Factory operations such as inventory management, production planning, resource management, cost management in case of manufacturing firms
- **Sales and Distribution**: Function responsible for moving a product or service from its point of production to the hands of consumer. Uses sales and demand

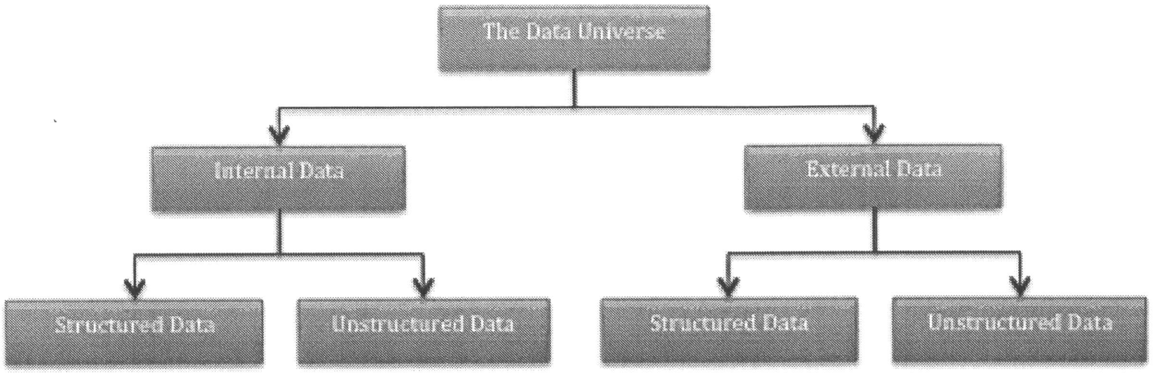
forecasting based on user preferences, sales volume, competition intelligence, supply chain inventory et cetera

- **Product Development**: New product development and improvement of existing product based on market feedback, user preferences, industry trends, innovations and inventions
- **Procurement**: Buying function that focuses on both direct and indirect materials or services; also manages vendor relationships. The primary objectives for this function are cost and quality. Vendor management is based on performance indicators such as defects data, timeliness, order history, supply history et cetera
- **Human Resource Management**: People management function focused on building employee motivation, improving productivity, enabling skilling; based on corporate goals, industry benchmarks and competition intelligence
- **Customer Service**: Responsible for customer's post-purchase experience. Key activities include issue resolution, feedback gathering, ongoing engagement, customer satisfaction
- **Fraud Prevention**: Especially relevant in Financial Services, also relevant in Treasury departments of corporations. Focused on deterring fraudulent activities in money movement. Also relevant for preventing accounting frauds and dissemination of mis-information
- **Risk Management**: Function dedicated towards the identification, assessment and prioritization of risks. Usually held in reference to financial risk but also covers operational and other business continuity risks.

3.1.1 TYPES OF DATA

The broad framework used for data categorization is as follows:

Figure 2



As we explore the various data sources, the abovementioned categorization helps in understanding the many facets of the data source.

Internal data refers to all the data owned by an enterprise. This includes sales and production logs, payroll information, customer reviews and feedback et cetera. This data is accessible only to the corporation and stands to be exploited by its own management. In some cases, companies may choose to sell or share their databases to other businesses for cross-utilization. These databases become 'External' data for the buyer firms. The decision to use external data for insight gathering is solely dependent on the sophistication of decision making at any corporation.

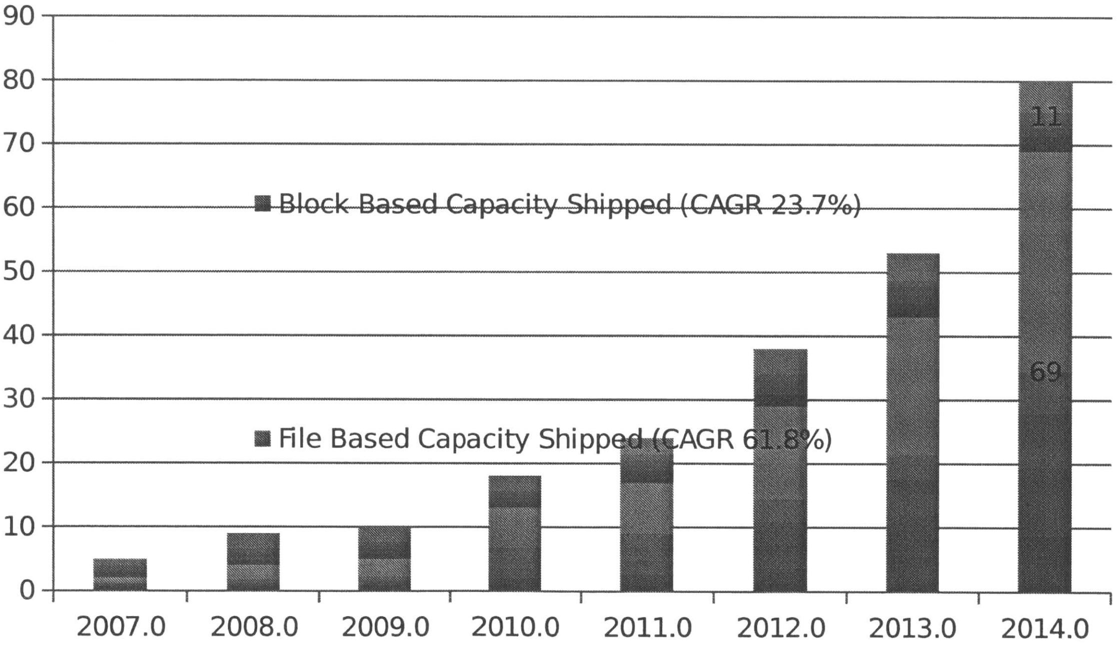
Both data categories, internal and external are further sub-categorized as structured and unstructured. Structured data refers to 'information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations' (Brightplanet, 2012). Many experts believe that structured data derives its name from Structured Query Language, more popularly known as SQL, a special-purpose programming language used for data management off relational database management systems (RDBMS). SQL consists of three languages, namely, a data definition language, a data manipulation language and a data control language. Structured data can be managed and queried using SQL codes.

Unstructured data, by definition, is unorganized data that cannot fit into an existing data model or is not organized in a pre-defined fashion. Historically, unstructured data referred to the wealth of information stacked neatly into a company's filing cabinets. In today's age, it refers to text, videos, images and other forms of unorganized data. Some of the most common techniques to handle unstructured data are Natural Language Processing, Text Analytics, Image Analytics et cetera. Unstructured data analytics is at a nascent stage with only one formal industry standard available, Unstructured Information Management Architecture. Some general frameworks such as General Architecture for Text Engineering (GATE) and the Natural Language Toolkit (NLTK) are also used for NLP and text analytics. This thesis explores analytics technologies and frameworks in the next section.

Structured versus Unstructured Data

One of the most remarkable distinctions between structured and unstructured data can be exhibited by reviewing the storage usage over time. The graph below marks structured data storage as 'Block' and unstructured data storage as 'File'. It indicates that the world is generating almost 6X unstructured data compared to structured data. The Y axis indicates capacity shipped in exabytes, where EB = 10^18 bytes

Figure 3



Source: IDC

Several agencies and industry experts have estimated the quantity of new unstructured data. Computer World states that 70-80% of all data owned by an organization is unstructured. As a rule of thumb, most industry analysts agree that the composition of data is heavily skewed towards unstructured and amounts to ~80%

The estimates seem fair when held in contrast with verifiable inputs into the world of unstructured data. Here are a few interesting few statistics:

- 400 hours of new video is uploaded to YouTube as of July 2015 (Statista, 2015)
- There are upwards of 101 M websites in the world (Internetlivestats, 2016)
- 300 M photos are uploaded on Facebook daily (Gizmodo, 2015)
- Over 500 M new tweets per day (Gizmodo, 2015)
- 70 M photos are shared on Instagram daily (Hootsuite, 2015)

It is fair to assume that unstructured data is significantly higher than structured data and is expected to remain so, in the coming years.

3.1.2 Data Sources

In this section, we will review the various kinds of structured and unstructured data available to a corporation through internal and external sources.

3.1.2 (a) **Internal Databases**

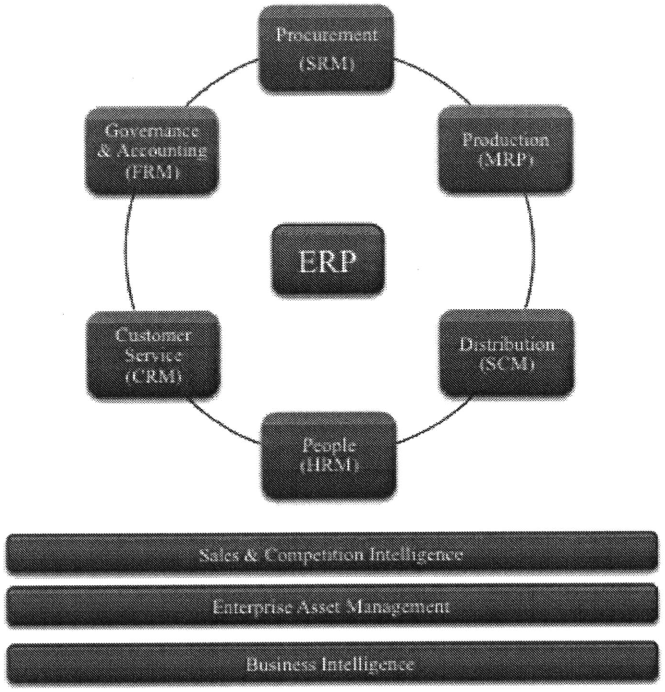
As the scale of corporations started growing exponentially in the post world war II era, the need to manage business processes became apparent. By the mid 1950s, several information technology companies such as IBM began to shift their focus towards resolving this problem. This period saw the birth of modern business intelligence software.

Business Intelligence and Business Analytics are used interchangeably in many contexts. There are several overlapping features between the two; however, a few distinctions set them apart. One of the key differences is that business intelligence refers to the process of collecting data and deriving insights through querying and online analytical processes while business analytics refers to the process of using statistics and quant for predictive and explanatory modeling. Broadly, both enable corporate decision making with different yet interconnected tools.

Business Intelligence tools saw an emergence in the 1980s. Adoption of Data Warehouse became more widespread and corporate decision-making started getting more and more data driven. The early 1990s saw the emergence of 'Enterprise Resource Planning'. The term ERP refers to a suite of business applications targeted towards integrating all business functions under a common umbrella in order to improve operational efficiencies and business management. Alongside BI and ERP, Customer relationship management (CRM) found large-scale adoption too. CRM refers to a host of applications that help manage a corporation's interaction with its customers. This includes organizing marketing efforts, managing sales pipeline, optimizing sales cycles, improving customer service, increasing lifetime value of a customer et cetera.

In today's times, ERP has gained a near 100% adoption. Its implementation is costly and time-consuming; therefore the market is currently an upgrade market.

Figure 4



Source: SAP

ERP vendors are increasingly observed to use BI to improve the attractiveness of their product. 'ERP with Intelligence' has become a trend that stands to gain popularity in the coming years. Convergence is the key to faster and better insights. BI tools enable pulling together data collected by ERP systems from across the organization, providing data visualization to spot trends and outliers.

In order to understand the extent of internal data sources that could potentially be available to a corporation, it is important to review the existing successful solutions in the market for ERP, CRM and BI and how these solutions play an intrinsic role in big data analytics.

Enterprise Resource Planning

One of the biggest strengths of using ERP for big data collection is that data from various functions within an organization can be consolidated, stored and analyzed together.

The following sources of internal data are available to a company through their ERP framework:

Table 1

Function	Data Collected	Standard Use Cases
Sales	• Intent expressed • Order placement • Order scheduling • Payment history	• Sales cycle time assessment • Sales-force performance assessment • Sales channel assessment
Customer Relationship Management	• Customer details (demographics, contacts) • Customer interaction/ engagement history • User journey • Customer satisfaction scores • Conversion funnel position	• Customer segmentation • Journey maps • Comparative features analysis • Willingness to pay assessment
Procurement	• Contract data • Supplier performance (issues reported, responsiveness rate, SLA adherence) • Cost savings	• Spend analysis • Comparative performance analytics • Performance optimization analytics • Risk analytics • Credit performance
Operations	• Production history • Cost of goods/ services • Capacity utilization	• Financial performance review (return on investment, EBITDA,

	• Demand-supply dynamics	Profitability) • Customer lifetime value assessment (average revenue per user) • Customer satisfaction index (demand fulfillment) • Operational performance analytics (unused capacity, peak demand management)
Supply Chain Management	• Ordering history • Delivery history • Delivery schedules • Nodes and channels • Bottlenecks	• Cycle time assessment • Risk assessment • Cost optimization • Operational efficiency assessment • Delivery times • Demand forecast
Human Resource Management	• Payroll • Productivity • Compensation & benefits • Employee engagement • Employee feedback • Learning & development	• ROI on training & development • Retention rate • Time to hire • Workload management • Workload forecast • Compensation & benefits trends analytics
Corporate Governance & Financial Accounting	• Accounting history • Shareholders data • Earnings trends • Stock market performance • Internal investment strategy & history • Business unit performance – profitability, costs, margins • Debt profile	• Financial health analysis • Investment strategy analysis • Competitive strategy development • Shareholder utility assessment

The above-mentioned analytics use cases are performed by most large organizations at varying degrees of sophistication. The level of sophistication is largely correlated with the IT infrastructure investments made by the company. It is worthwhile noting that according to a Gartner's CIO Agenda Survey 2015; big data analytics is the topmost strategic priority. Therefore, it is a fair assumption that most corporations would see an increased utilization of internal structured databases in the next 2-3 years.

ERP Service Providers

The most successful ERP vendor is SAP with the largest market share of 24%. SAP is followed by Oracle with a 12% market share. Sage and Infor are tied at number 3 with 6% market share closely followed by Microsoft with 5% (Columbus L., 2014)

The ERP market has had a flat growth rate for many years now. Currently, the focus is shifting from traditional ERP technologies to post-modern ERP. Post-modern ERP focuses on Agility, Flexibility and Scalability. The solutions are designed in "Software-as-a-service" model and are delivered over cloud (Hardcastle C., 2016)

Sources of Internal Unstructured Data

Large-scale adoption and implementation of ERP and BI tools has led to systematic collection of structured data. A large part of the previous decade was invested towards developing technologies to process structured data. Firms have always realized the potential of all the unstructured data stored in personnel's computers, corporate websites, social media pages, customer service portals et cetera. However, the common consensus is that a firm's internal unstructured data remains a largely untapped resource.

Examining the sources of internal unstructured data:

Based on primary interviews conducted with mid to senior level employees at large organizations such as General Motors, Google, Amazon, Microsoft, Deloitte, Shell, and many others; the following sources of internal unstructured data have been identified (Table 2). Unsurprisingly, a large percentage of unstructured and semi-structured data is not used for processing and insight generation.

Table 2

Function	Data Type	Utilization for Analytics

Sales & Marketing	• Social media chatter • User generated content – images, videos, text; in context of a campaign or product • Sales meeting minutes • Competition marketing collateral • Advertisements (digital & non-digital) • Website content • Documents & emails • Field reports • Journalistic reviews, industry forecast reports • Webinars, seminars	Medium
Customer Relationship Management	• Textual product reviews • Textual complaints log • Textual feedback • Call center audio reports • Call center notes/ transcripts • Customer meeting minutes	Medium
Procurement	• Contracts, agreements and amendments • Warranties • Annual reports • Marketing materials • Claims processing • Vendor performance feedback reports	Low to Medium
Operations	• Delivery notes • Scheduling notes • Claims processing • Quality assurance reports • Engineering changes	Low
Supply Chain Management	• Delivery directions • Delivery specifications (by customer) • Tracking information	Low
Human Resource Management	• Job descriptions • Email history • Hiring offers • Termination letters • Employee performance feedback reports • Skill level report • Employee manuals • Policies & procedures	Low

Corporate Governance & Financial Accounting	• Patents, Trademarks and Non-disclosures • Telephone transcripts • Documents for targets, forecasts, leadership review • Meeting minutes • Email history • Audit trails • Account description	Low

3.1.2 (b) **External Databases**

The general wisdom behind big data analytics is that a better understanding of one's customers enables businesses to target products and services that are most relevant to these customers. Relevant products and services increase the total utility of the customer, thereby increasing satisfaction and leading to repeat purchases and long term loyalty.

In present day, the combination of internally collected data and existing levels of technology have yielded results indicating towards a long road ahead. Here are a few statistics that are helpful in gaining perspective (Stec C., 2015)

1. Display ad viewability has plateaued and is in the ~46% range year on year
2. Average click-through rates of display ads across placements & formats is 0.06%
3. Ad-blocking has become common with a 41% rise in 2014-15 period
4. In a study conducted by Infolinks on banner blindness, 2.8% respondents indicated that ads targeted to them had any relevance
5. A 2014 study by Harris Interactive indicated that 18-34 year olds were more likely to ignore online ads, such as banners and those on social media and search engines, than they were traditional TV, radio and newspaper ads
6. ~50% clicks on mobile ads are accidental
7. ~33% users find display ads completely intolerable

It is clear that the present levels of analytics in identifying user preference to target products and services are not matching up. Tools and technologies built for the purpose of analytics are by and large using internal data sources. This thesis discusses these technologies in later sections.

However, there are additional sources of data, external to the company, which can provide valuable insight into user preferences. With existing technology, adding a new layer of data could presumably yield better results. In this section, we will explore the various sources of external data available, the data that is not available but could be extremely useful if made available and the ways in which external and internal data can be blended together.

Understanding Data Markets

The value of external data was most realized by businesses in the post-world war II era. As competition intensified and newer players entered the market, businesses felt the need to collect information outside the ambit of their organization. The most relevant data required by all businesses was about the customers. Fast Moving Consumer Goods (FMCG) sector was the biggest adopter of questionnaire-based market research. Collecting customer feedback became a norm across all sectors beyond FMCG by the 1980s. By the 2000s, customer data collection moved from telephone based surveys to web-surveys. The fundamentals behind data collection did not seem to change though. External data was still collected by surveys and clumsily merged with internal sales and production data. All external data was owned by the corporation and guarded. Strategy discussions required for the inclusion of macroeconomic data that was bought off a subscription based model by companies such as Factiva or Forrester Research or Capital IQ et cetera.

By the mid 2000s, the concept of 'Data Markets' started to emerge. The value of sharing data in exchange for data or money became clear. There are three key reasons behind the usefulness of Data Marketplaces

1. Single point of discoverability and data comparison
2. Data is clean and formatted, making it as close to ready-to-use as possible
3. Provide a much-needed economic platform for accessing and publishing data

The other important aspect of external big data is that data markets realize the benefit of hosting data on cloud. It is easier to move computation to the data than vice versa. The four major data market players are –

Microsoft Azure Data Marketplace

- Microsoft's data marketplace sits alongside Microsoft's applications marketplace. Azure offers a web interface for data access, including queries, using a standard data protocol, OData. This allows for programs such as Microsoft Excel and

PowerPoint to directly access the Azure marketplace data. This is a big strength of the offering as it enables access to external data using existing enterprise tooling. Additionally, OData supports a range of programming languages. Azure marketplace has maintained a strong reliability of the quality of data by ensuring a strict filtering mechanism. It has a respectable range of data published by reputed publishers such as Dun & Bradstreet and ESRI. Azure data marketplace is widely regarded as one of the best data publisher - data consumer platform

Infochimps

- Built on the concept of code-hosting sites such as SourceForge or GitHub, Infochimps attempts to bring together publically and commercially available data on platform. The concept is based on 'network effects' for data. The presence of more data on the same platform makes all data more valuable. The data sets sold off Infochimps is pre-cleaned and pre-integrated. The focus is largely on location accuracy and not on specializing in one category. Infochimps is not targeting a particular sector or solution category, instead is focusing on the entire breadth of customer data available. The overall idea is to tie together seemingly disparate set of data points in order to find tangible insights. So far, the scale of its vision is immense and yet to yield major monetary results

Factual

- An open data platform wherein a community of contributors improves data quality by leveraging its cloud based tools. In contrast with Infochimps, Factual had decided to specialize in one segment of the data market – geographical and location based data. Since the model is based on community development by sector, Factual is currently pursuing one community at a time, thereby optimizing the amount of marketing effort & dollars required for expansion. Another interesting business model innovation is that Factual does not incentivize organizations to publish data in exchange for money. The reason stated is that most organizations do not see much monetary value being generated by sharing data, and hence refrain from allocating resources towards it. On the contrary, sharing data in exchange for data has proven to be a stronger incentive and improves data quality for everybody.

DataMarket

- The above three data marketplaces are known for the value they bring to developers. Interestingly, the end user of data, an analyst or researcher remained largely ignored. DataMarket operates in the market catering to the end user of data. By combining strong data visualization techniques with accurate data sources, DataMarket has become a top choice for analysts and researches. Currently, the company is focusing on incentivizing publishers to signup with the platform. A significant portion of data is free of charge, with a fees for premium data

Other interesting data markets:

- Social media data streams are valuable for identifying emerging trends and for monitoring social chatter. Companies such as Gnip and Datasift are relatively well known in this space
- Wolfram Alpha is an interesting integrator of diverse databases and enables computation using some of its proprietary tools
- Jigsaw is an online central repository of contact information of individuals and organizations. It handles curation and distribution in exchange for money or data
- Kaggle is a platform that incentivizes companies to share data with a community of data scientists for the purpose of predictive analytics. So far this data does not seem to be exchanged between companies, however, that might be the logical next step in its evolution

Types of data available

In this section, we would explore the various types of data available, that a company could potentially leverage for better analytics results. It is important to note that various corporations use the available data in an unorganized fashion. Leading efforts are made towards leveraging social media data as an indicator of emerging trends in consumer behavior. Historically, companies did try to incorporate macroeconomic data into decision-making. However, this was done in an unscientific manner and usually relied on the ability of leadership to see apparent correlations. The evolution of big data would eventually lead to a world where trends and patterns between seemingly unrelated sources would emerge. This would be discussed in more detail in the next section.

Macroeconomic & Geopolitical

- For years, data published by The World Bank, International Monetary Fund (IMF), The CIA factbook, has been used by companies during crucial decisions around global expansion, new market entry, product development et cetera
- These data points are fairly accurate and contain a surprising level of detail about macroeconomic indicators, peace indicators and ease of business indicators. Most consulting companies use this data to publish reports about their assessment of the next phase of growth and promising markets. Scientifically incorporating this data into identifying shifts in user preferences by geography is yet to be done

Published Reports

- Market research and consulting companies conduct several data collection drives based on traditional market research methodology or primary interviews and surveys. This data is valuable as it collects information directly from the consumer and requires little to no interpretation based on consumer behavior. Survey based data collection remains one of the most powerful tools to hear straight from the horse's mouth
- Several companies such as Gartner publish their annual reports such as Hype Cycle, Magic Quadrant, which encapsulates industry trends in a succinct manner. Large corporations loosely use this data by buying into subscription to these services. However, a scientific tool that incorporates this data into other sources of data does not exist

Social

- A few years ago, the terms Facebook and Social Media were used interchangeably. However, as the world moved on from Web 2.0, many new avenues of social media emerged successfully. Twitter is experiencing flat user growth, yet it remains one of the most potent forces in shaping social trends.

Instagram has become a large repository of user-generated images. Snapchat and Vine are other emerging platforms

- As valuable as social media is, it is also the most challenging source of data for corporations to deal with. Unsurprisingly, user-generated content is almost entirely unstructured. So far, social media analytics focused on the relatively easy analytics of analyzing the structured bits of social content – hashtags, location, time of day, day of week, other meta data, user's past engagement history et cetera. Efforts were made to use Natural Language Processing and Sentiment Analytics to understand the context and content of text based user content. Several new companies are emerging in the space of image and video analytics to use the content of the image for the purpose of insight generation
- It is important to point out that the insights from structured social analytics are increasingly being merged with internal data around sales, to make better demand forecasts. The value of unstructured social analytics is widely recognized and yet to be harnessed

Reviews

- User reviews are by far one of the most powerful information sources available to any products or services company. The most challenging aspect of this information source is that, much like data from social media, user reviews are almost entirely unstructured. Reviews range from text-based feedback left on Yelp, IMDB, Amazon, Netflix, Spotify, Glassdoor and other feedback portals to images shared on the same portals plus platforms such as Instagram and Facebook. Tweets on product or service quality have known to generate a snowballing effect
- These reviews can be massively useful as an input into product development strategy and even positioning and launch strategy. For instance, targeting a new movie to connoisseurs of a particular taste of movie, music and books would be far more successful than relying on a simple 'like based' recommendation model

Behavioral & Lifestyle

- Identifying each user as a unique individual and not one point in the dataset of a 'segment' is the first step towards more accurate need identification. This requires for integrating data beyond the consumption of a product or service. User behavior on discussion forums, usage pattern on services such as Uber, AirBnB, engagement history on Netflix or HBO, reviews left on Good Reads, fitness data off Fitbit and many more

- Predictive analytics requires a micro understanding of a user's likes, dislikes, current needs and experiences. Seemingly unrelated data could very well be the key to better segmentation

IOT and the future of data

Internet of Things is slowly finding its way into everyday life. The current challenge is to install appliances and devices in homes and offices. The devices would carry basic sensors capturing basic data. At the moment, even the most simplistic data from IOT could prove invaluable.

The other upside of IOT data is that it would be structured and hence easy to integrate and analyze with present technologies. The infrastructure and data sharing norms governing IOT are yet to be determined. Several use cases such as a refrigerator recognizing depleting levels of milk and thus placing an order on Instacart, have been discussed. Although, the modus operandi between refrigerator manufacturer and Instacart (for example) remains yet to be established.

3.2 Data Storage

The most critical aspect of big data analytics is the infrastructure required to support the large quantity of data throughout the data lifecycle. The data lifecycle includes acquisition, preparation, integration and execution. Additionally, it is important that data must be stored in a way that it can be accessed promptly for analysis. For the data to be accessed quickly, it used to be stored on the main memory of a computer. However, as the scale of big data has increased beyond the capacity of a computer's dynamic random access memory (DRAM), other methods of storage were introduced. The concept of splitting a large dataset into smaller chunks to be stored across multiple hard-drives on several machines over the Ethernet network proved to be most effective. However, this storage architecture drastically increased the time to access data. Graduate students at

MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) developed a unique solution to this problem. This system is called the Flash storage system and as the name suggests, uses flash memory. Currently, flash storage solutions are the most expensive but perform remarkably on speed. Other storage options are Spanning Disk and Tape. Costs for both Spanning Disk and Tape are lower as compared to Flash, but speed of access is a serious trade off. Today, major players such as IBM, Amazon Web Services, Oracle, EMC and many others offer all kinds of big data storage solutions.

Some of the key considerations during the selection of a specialty storage system to accommodate big data applications are –

1. **Scalability**
    - The ability of the storage sub-systems to manage massive datasets
    - Scalability does not only refer to size of disc. Many storage solution vendors may insist on adding disc-space as a way to scale the solution. However, that adds considerably to both capital expense and operating costs
    - It is important that other parameters such as 'throughput' and 'speed of access' are also scalable
2. **Extensibility**
    - The ability of the storage system to grow with no artificial constraints
    - A tiered storage solution allows for extensibility by improving access to data in need and archiving data not needed into another layer
3. **Accessibility**
    - Access to a large user community without adversely affecting performance
    - Data storage is almost always shared between multiple applications and users. Storage systems must have the ability to make priority decisions through automation
4. **Fault-tolerance & Self-healing**
    - The ability of the system to handle faults and self-correct without loss of data
    - Automatic redirection of work to another server in the event of one server failing is by far the most important feature. Outages can prove catastrophic for business and self-healing properties of a solution would prevent that.
5. **High-speed I/O**
    - The ability of the system to meet challenging big data demands in short periods of time (especially relevant in today's real-time age)

- More often the need for data availability is across geographies as users are spread across different locations. The speed of input/output is best optimized in cloud based solutions

6. **Integrated**
   - The ability to seamlessly integrate with the production environment and possibly other legacy systems
   - It is best if the solution also supports workflow automation, that is, seamlessly transfer information between applications and users

3.2.1 Relational & Non-Relational Databases

Relational databases, as mentioned in section 3.1.1, are a mode of storing information in tables with rows and columns. The data stored in relational databases is strictly structured. SQL is used for storing and retrieving data in relational databases. The queries are generated in simple English, making adoption easier. Relational databases originated in the 1970s and have been in use since the adoption of Business Intelligence software became commonplace. The reason relational databases have been successfully adopted is that they meet all the above 6 criteria to a reasonable point.

Non-relational databases came into existence and widespread adoption due to the massive rise in unstructured data. The key difference between non-relational databases and relational databases is that NRDBMS does not require 'Referential Integrity' of data. Data can be de-normalized. Google (BigTable), Yahoo (HBase) and Facebook (Cassandra) are the creators (amongst others) of commonly used non-relational databases.

3.2.2 Data Management

Storage management and Data management are often assumed to be the same. However there are some key differences. Storage management is broadly refers to the technologies and processes used in order to improve the performance of data storage infrastructure or resources. Data management refers to technologies and processes used in order to manage the information lifecycle of data

The table below highlights the tasks and responsibilities of both the key functional areas.

Table 3

	Tasks	Responsibilities
Data Management	• Data deletion • Data archiving • File analysis • Copy data management • Archive product purchasing • Growth capacity and archive capacity planning and monitoring • Policy enforcement • Archive/backup administration • Asset management • Total cost of ownership reporting	• Acquisition cost tracking • Upgrade planning • Data management and backup/archive strategy • Policy enforcement • Data management product acquisition • Portfolio management and contract negotiation
Storage Management	• Storage capacity planning • Performance monitoring and reporting • SLA and operational level agreement tracking • Asset management • Backup and restore recovery point objective (RPO) and recovery time objective (RTO) agreements • Capacity planning	• Acquisition cost tracking • Upgrade planning • Storage strategy • Financial planning • Storage product acquisition • Portfolio management • Upgrade contract negotiation
Storage Administration	• Device configuration • Device installation • Storage provisioning • Problem determination • Upgrade and preventative maintenance • Dashboard monitoring • Backup and restore administration	• Storage device operation and administration • Dashboard implementation and maintenance

3.2.3 HADOOP AND DATA WAREHOUSE

Hadoop is an open source, Java based programming framework that uses a distributed computing environment for storing and processing large scales of data. Broadly, it works on the following principles

- Drew inspiration from Google's MapReduce software framework that increased processing speed by breaking down an application over thousands of nodes
- Key strengths include high computing power, fault tolerance and flexibility

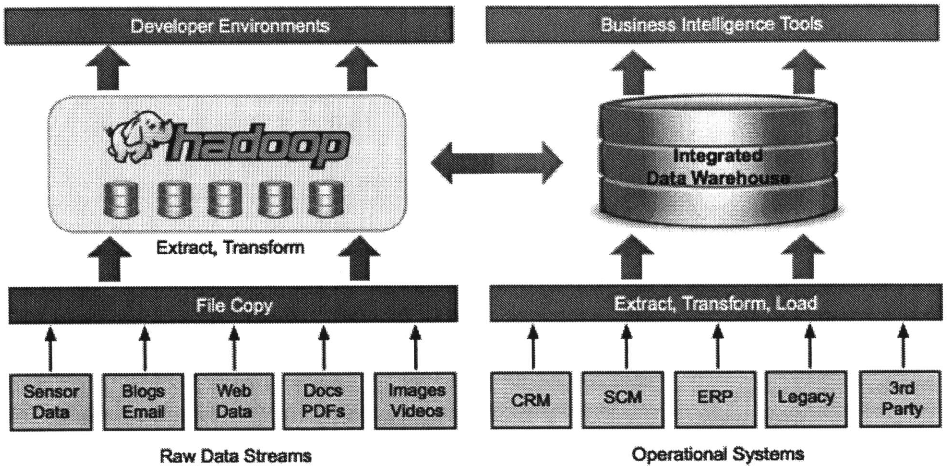
- Does not require pre-processing of data and can store both structured and unstructured data in large proportions. Offers the flexibility to store first and decide processing steps later
- Current Hadoop ecosystem consists of Hadoop kernel, MapReduce, Hadoop distributed file system and a few other related projects under the Apache umbrella

Data warehouse is an integrated repository of all the data collected within an organization. Data warehouse is an integral part of Business Intelligence. Broadly, it works on the following principles

- Typical data warehouses operate on ETL – extract, transform and load
- Uses staging, integration and access layers to facilitate analytics

As data gets bigger and bigger at a corporation, its movement over a network for the purpose of transformation and analytics becomes tougher and tougher. Attempts made to move terabytes of data over a network could lead to outages. It could also slow the programming effort. As a novel solution, instead of moving data across the network, moving processing applications towards the data is more feasible. Therefore, all big data cannot be stored over an ETL server or Storage Area Network (SAN). Assuming movement of data across the network is not a challenge, processing speed would still remain constrained by the limited bandwidth of SAN. Hadoop loads raw data directly onto low cost commodity servers. This action takes place one time. The higher value refined results are passed to other systems. Processing using ETL continues to run in parallel. The entire bandwidth is utilized effectively thereby making operations faster that the alternative of pulling data from SAN into ETL servers. Most often, Hadoop systems sit side by side with Data Warehouses.

Figure 5



Source: Teradata

3.3 DATA PROCESSING

Gartner defines big data as "high volume, high velocity, high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". All through the 1990s, the objective of leaders of business intelligence software was to improve analytics accuracy and speed. As the concept of big data took shape, business intelligence software continued to evolve.

One of the more commonly used categories for types of big data analytics is as follows –

1. Descriptive analytics
2. Diagnostic analytics
3. Discovery analytics
4. Predictive analytics
5. Prescriptive analytics

In the early stages of big data analytics, the focus was largely on Descriptive and Diagnostic analytics. These two types of analytics are about taking a hindsight view, after the fact. Discovery analytics attempted to move beyond hindsight into insight. This was allowed for, by technological improvements that reduced latency and enabled 'real-time' processing. Predictive and Prescriptive analytics are about foresight. Machine learning is a crucial enabler in the success of predictive and prescriptive analytics. Currently, we are transitioning from insight to foresight.

There are various tools and analytics technologies that can be used for each stage of analytics.

Table 4

Stages of Analytics	Structured	Unstructured
Descriptive	Statistical analysis	Search based applications
Diagnostic	+ Online analytical processing Dashboards and reporting	Search based applications
Discovery	+ Data mining Process mining Event processing Performance metrics	+ Search query Natural language processing
Predictive	+ Advanced machine learning using data mining and process mining AdHoc query and analysis Data discovery	+ Text/ Word analytics Sentiment analytics Audio analytics
Prescriptive	+ Machine learning and AI assisted analytics	+ Image analytics Video analytics

3.3.1 Structured Data Processing

As Business Intelligence gained widespread adoption through the 1990s, structured data processing became an industry norm. Business Intelligence comprises of the following steps –

1. Business process management & performance metrics
2. Analytics
3. Insight reporting
4. Data cross-pollination and collaboration
5. Knowledge management

In this section, we will explore the various kinds of analytics (step 2) performed and the technologies used to support them. A broad categorization of structured data analytics is as follows –

1. Sales analytics
2. Marketing analytics
3. Pricing analytics
4. Supply chain analytics
5. People analytics
6. Risk & credit analytics
7. Fraud detection
8. Cohort analysis
9. Behavioral analysis
10. Enterprise optimization, et al

Business intelligence applications for data analysis can perform all the above categories of analytics with minor modifications. A few crucial elements of business intelligence application are –

Online analytical processing (OLAP)

- OLAP enables its users to perform data analysis across multiple dimensions and use data interactively. It allows for three key operations – Consolidation, Drill-down and Slicing. Consolidation is the process of rolling up data into one unit while Drill-down is the process of viewing the details. Slicing allows for the viewing of data across different dimensions

Ad Hoc analysis & querying

- The ability of a tool to support need-based analysis is referred as Ad Hoc analysis. This allows the user to generate a query to further dig into the details of an observation made on the OLAP dashboard. It accords flexibility to the user and allows them to make special queries as the need arises

Mobile BI

- An alternate view of the application's user interface or reporting dashboard on mobile devices so as to enable users to view results and KPIs on the go

Real-time BI

- The ability of a system to provide performance parameters at a micro unit of time is referred as Rea-time BI. The definition of 'real-time' could vary by industry. In sectors with high volatility of sales, such as E-commerce, real-time is considered as small as one day

Operational BI

- Applications that analyze business performance on the go and create an alarm in case of deviation from previously set limits. The objective of these applications is to identify and fix a problem in real-time and not use the conventional after the fact BI model

Location intelligence

- An application's ability to present geographical context to data is referred as location intelligence. The application superimposes data from aerial maps, geopolitical maps and company databases in order to present an amalgamated view of a business situation

3.3.1 (a) **Analytics Techniques**

Prediction and Prescription analytics fall under advanced analytics. These are fairly new fields of analytics and require complex modeling. The underlying analytics techniques used for advanced data analytics are Statistical Regression Technique and Machine Learning Technique. This section will also touch upon Experimentation as an analytics technique as it is extremely relevant in Internet businesses.

Statistical regression technique is one of the fundamental analytics methodologies. Identifying a mathematical equation that is representative of the interactions between various elements or variables is the critical first step towards establishing a predictive model. Some industry veterans express skepticism towards the applicability of simpler regression models such as linear regression or multivariate regression for predictive analytics. However, for simpler models, regression has proved to be effective. The other types of regressions most commonly used in the development of predictive models, are

- **Logistic regression**

    - Used for cases with a binary outcome such as clinic trials and fraud detection
- **Ridge regression**
    - A better version of linear regression with constraints on regression coefficients to prevent possible over fitting
- **Lasso regression**
    - Another variation of Ridge regression that allows for regression coefficients to be zero thereby allowing for variable reduction
- **Logic regression**
    - Most useful in scoring algorithms with binary variables
- **Bayesian regression**
    - Stable version of regression without the constraints of a simple linear regression
- **Jackknife regression**
    - Used as a general clustering and data reduction technique. Considered ideal for black-box algorithms as it works under the circumstances when assumptions around linear regression are violated

There are many other forms of regression models that are used for different use cases. However, the ones mentioned above are by far the most commonly used and found to be the most effective.

Machine learning's origins lie in Artificial Intelligence. The principle behind machine learning was to enable a computer to learn from its interaction with data. Machine learning and data mining have a similar approach. Each system sifts through data extensively to identify patterns. The key difference lies in data mining's need for human comprehension upon pattern recognition, in contrast with machine learning's ability to adjust the program according to observed pattern to detect further useful insights. A further advancement of machine learning is un-supervised machine learning that can apply adjusted algorithms from one dataset onto a new dataset for draw inferences.

Some of the commonly used methods for designing predictive models using machine learning technology are –

- **Regression algorithms**
    - Statistical machine learning using standard regression models
- **Instance-based algorithms**

- Model builds up training databases and uses new data to compare to find the best match of instance-based database

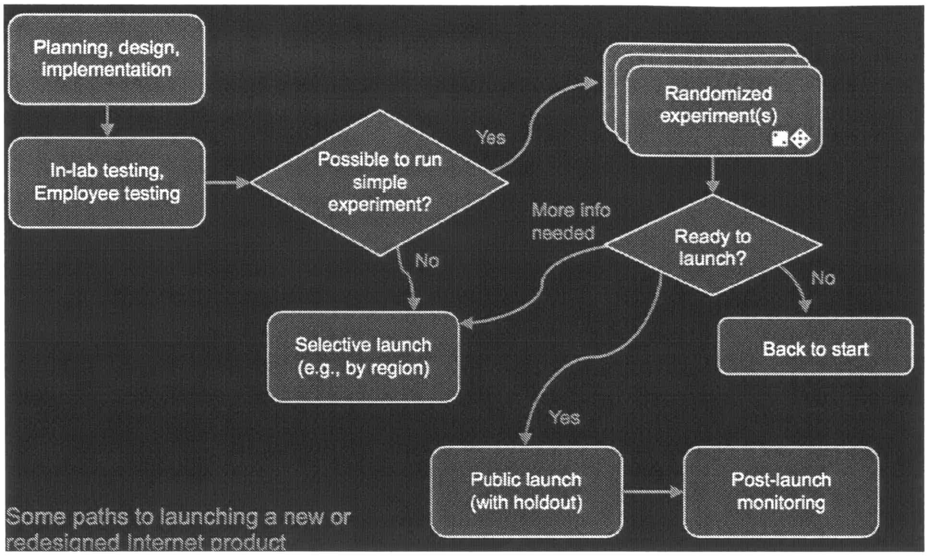
- **Regularization algorithms**
  - An extension of regression algorithms with an added feature of penalties for overly complex structures
- **Decision tree algorithms**
  - Uses standard decision tree principles and are the most popular algorithms for machine learning
- **Bayesian algorithms**
  - Use Bayes theorem for problems of classification and regression
- **Clustering algorithms**
  - Uses data's inherent structure and qualities to form naturally occurring buckets
- **Artificial neural network algorithms**
  - Classical methods of analytics that derive its principles from biological neural networks. Deep learning algorithms build more complex and larger neural networks
- **Ensemble algorithms**
  - Models that comprise of several smaller models that complete one part of the problem and assemble solutions to generate a larger result

Machine learning has proven to be the most effective tool in advanced analytics. There are multiple open source tools available such as Apache Mahout, OpenNN, R et cetera. Other tools are developed by large organizations such as IBM's SPSS, SAP, SAS, Stata, Tibco, Alpine Data labs and many others.

Experimentation is a technique used most commonly by product managers to collect feedback before shipping. Figure 6 below captures the typical decision algorithm used for new product launch using experimentation as a product testing methodology. The most commonly used methods of experimentation are – A/B testing and Multivariate testing.

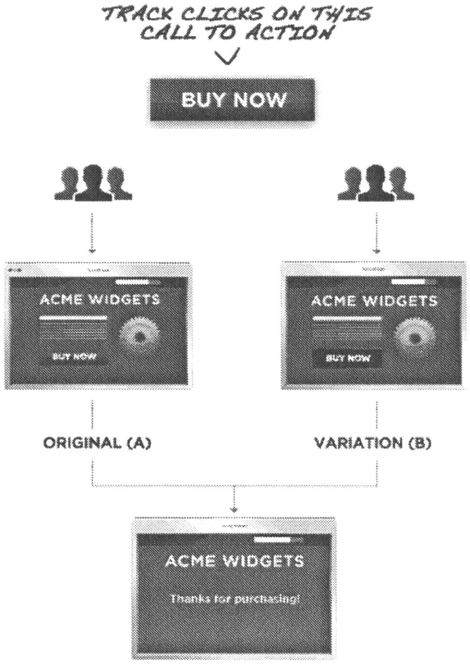
Figure 6

Source: Marketing Analytics, Dean Eckles, MIT Sloan


Planning, design, implementation
In-lab testing, Employee testing
Possible to run simple experiment?
Yes
No
Randomized experiment(s)
More info needed
Ready to launch?
No
Back to start
Selective launch (e.g., by region)
Yes
Public launch (with holdout)
Post-launch monitoring
Some paths to launching a new or redesigned Internet product

A/B testing is a method of split testing. Users belonging to the same customer segment are divided into two groups and further sub-divided into treatment and control. Each treatment sub-group is exposed to a different product option. User behavior data is collected and compared with the other group and cross-referenced with control groups. It is one of the most effective methods for web optimization and user interface design.

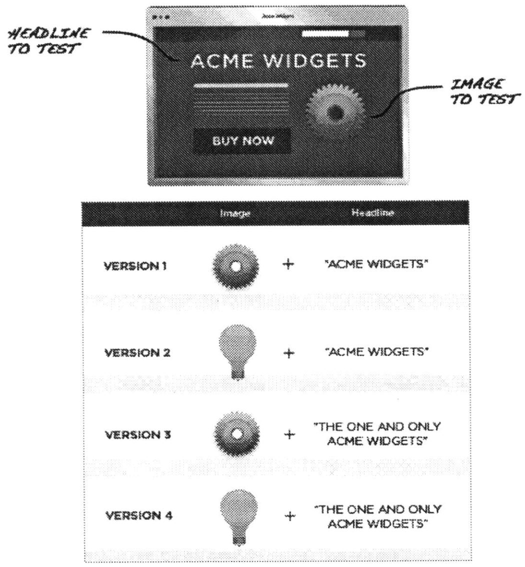
Figure 7



Source: Optimizely

Multivariate testing is an evolved version of A/B testing. Instead of comparing just two options as in A/B testing, multivariate testing compares a range of variables and provides observations about their interaction.

Figure 8



Source: Optimizely

Multivariate testing enables product managers to collect valuable information to enable product re-design efforts.

3.3.2 Unstructured Data Processing

The volume and value of unstructured data has been emphasized upon a few times in prior sections of the thesis. As valuable as unstructured data is, attempts to curate and analyze it are relatively recent. Processing technologies using relational databases do not fit well with the nature of unstructured data. To date, unstructured data is analyzed after converting it into structured data. The fundamental processing principles are not remarkably different from structured data analytics. The most important difference is that unstructured data requires pre-processing to convert into structured data. This principle can be illustrated by the following examples –

- In the world of social media, textual data has become omnipresent. As human beings, we are able to identify the 'sentiment' behind a text, appreciative, nonappreciative, sarcastic, et cetera. For the purpose of analyzing millions of such texts, most sentiment analytics engines rely on a simplistic algorithm. Words and phrases are flagged in the context of 'good', 'bad' and 'neutral'. Numeric values are assigned accordingly. For example, 'good' words and phrases would receive a value of +5, 'bad' words and phrases would receive a value of -5 and 'neutral' words would be assigned 0. The net score of a status or tweet is then used to analyze the net sentiment

- The most easily relatable form of image analytics, thanks to several crime dramas on television, is fingerprint analytics. Authorities have access to millions of fingerprints and those collected on the crime scene are run through a database to find a match. The processing principle used for fingerprint matching is that the software looks for unique points on a fingerprint to create a map or a polygon. Using a probabilistic method, it matches the map to the closest matches in its database. Amongst the shortlisted candidates, the final matching is done by human intervention based on rank ordered list provided by the software

It is clear that the pre-step to analysis of extracting structured information from unstructured information is where all the complications of unstructured data analytics lie. Once converted to structured, the volume and variety of data stored in unstructured information falls down remarkably. It becomes easy to manage and existing structured data tools and relational databases can be utilized. It is worthwhile mentioning that several attempts are being made to analyze unstructured data in its scrambled unstructured form. These technologies are in relatively nascent stage.

3.3.2 (a) **Key Technologies**

1. **Natural Language Processing**
    - NLP is by far one of the oldest forms of unstructured data processing technologies. The most significant progress came with the advancement of speech recognition. Present day NLP algorithms are based on machine learning. This is a remarkable progress from the 1980s when NLP algorithm required hand written rules that could run into several thousand pages. Semi-supervised or unsupervised machine learning is taking NLP to the next level. NLP is used most commonly for the following tasks

- Text summarization: Creates a short readable summary from a variety of textual sources
- Co-reference resolution: Resolves conflict between mentions and entities. Takes care of adverbs and adjectives
- Discourse analysis: Identifies discourse structure and classifies speech acts
- Machine translation: Language to language translation
- Name entity recognition: Sifts out proper nouns from rest of text
- Natural language generation: Converts computer databases to human language text
- Natural language understanding: Pick nuances and semantics in the in-built logic structures
- Optical character recognition: Identify text in an image
- Part of speech tagging
- Relationship extraction: Identify social relationships based on textual data
- Speech, topic and word segmentation

2. **Text Analytics**
    - Text analytics is very closely tied with Natural Language Processing. A variety of tasks between the two technologies overlap. It is fair to say that NLP is the bigger umbrella, while text analytics is largely covered under NLP with a few sub-tasks outside its ambit. The most important aspect of Text Analytics is pattern recognition. NLP allows for tagging and organizing of textual data. TA uses this organized data to identify patterns to generate insights. Some of the most common uses of Text Analytics are in the area of compliance and regulation. Financial services use TA extensively to ensure compliance. For example, TA is used on call center transcripts to identify possible non-compliance. Pro-active action is taken to prevent lawsuits and other liabilities
3. **Sentiment Analytics**
    - Sentiment analytics is closely tied with NLP and TA. It is a valuable technique to separate the opinion holder from the opinion object. As the name suggests, sentiment analytics aims to identify the mood or the emotion held by the author of the text. It is especially useful in customer service, feedback gathering, assessing support (to a campaign/ individual) et cetera. Sentiment

analytics is rarely performed independently. It is usually performed as a sub-task to NLP or TA

4. **Audio Analytics**
   - Audio analytics is in the least advanced stages of unstructured data analytics. The main concept behind audio analytics is to identify and categorize sounds. Industrial or enterprise level applications of audio analytics were limited. The most relevant audio files are call-center recordings that can be easily transcribed and analyzed using text and sentiment analytics. Sentiment identification is more accurate through audio analysis; however, the marginal benefit has not justified significant enterprise level investments. One of the more recent developments in the space of audio analytics is coming in the space of behavioral-modification. Sounds are analyzed for their impact on human psyche and ability to alter moods and behaviors. Possible use cases include enabling swifter movement of crowds in a public space (such as subway stations) using sound as a stimulus
5. **Image Analytics**
   - Image analytics technologies are moderately well developed but haven't found widespread adoption across all sectors. Healthcare has adopted medical image analytics for the identification of diseases such as detecting cancer in a mammogram. Facebook, with its humongous access to pictures uploaded by its users, has made some advancement in the field of image analytics. The 'tags' feature identifies user's friends based on facial recognition. Application of facial recognition transcends into national security, remote sensing, machine vision (in production plants) and many others. Lately, image analytics is finding its way into Augmented Reality (AR) and Virtual Reality (VR). AR and VR have not made a mass-market entry thus far, so we are yet to see how the image data collected through these technologies can be used. Another important application is in the recently introduced concept of self-driving cars. The cameras located at all sides of a car collect images and analyze them in real time to ensure safe transit
   - Image analytics is by far the most underserved area of unstructured data analytics. However, it is also the space where most innovations are taking place and startups are emerging

6. **Video Analytics**
    - Some of the most applications of video content analytics include facial recognition, object recognition, shape recognition, motion detection et cetera. So far, consumers of visual content analytics software are largely restricted in the area of law enforcement and terror prevention. Most governments have a huge repository of city level data captured by CCTV cameras. Video analytics software is used to identify potential criminal activity. Video analytics has also found some application in retail stores to identify consumer movement pattern to draw heat maps. These can be analyzed for better product placement, visual merchandising and planning other in-store promotions. Similar to applications of image analytics in Augmented Reality and Virtual Reality, video analytics would be useful in analyzing information collected through these emerging technologies

One of the important things to note about unstructured data is that it is so named because this category of data is not stored in a field in a relational database. This data is not accessed by SQL and does not follow its rigid structure. The overall principle of analyzing unstructured data in its original form still requires tagging and other data markers for retrieving data elements for use in software applications. The lost meaning of this data could be recovered by search engines that can index and search non-relational data types or semantic processors, thereby replacing SQL in the world of unstructured data. Several newer standards are emerging that could attribute ‘structure’ to unstructured data outside the rigidity of SQL. One such framework is JavaScript Object Notation (JSON). It uses readable text to transmit data objects with attribute or value pairs. It is basically used as an alternative to XML. It provides the desired flexibility by including all the data a developer wants to access. The applications are service-based and hosted on the cloud.

In most enterprises, data architectures are a hybrid of several database and file management systems. The time tested ones such as SQL and MapReduce are held in conjunction with NoSQL, NewSQL, Multivalue, Hierarchical, and Grid et al.

3.4 Data Visualization

The science of data analytics meets art and storytelling in its final step at data visualization. The motivation and objective behind developing complex tools and technologies for capturing big data, storing and processing it, is to be able to use insights generated for better decision-making. However, decision-making requires visualizing processed data in ways that insights emerge in contrast with the background.

The most challenging aspect of data visualization is that it caters to the band of corporate employees, 'leadership' that has an absolute scarcity of time-resource. The objective of data visualization tools has always been to maximize insight in minimal time.

There are two aspects of data visualization that are captured in this thesis:

1. Existing tools and visualization techniques
2. Thought-leaders & the future of visualization

3.4.1 Overview of Current Tools

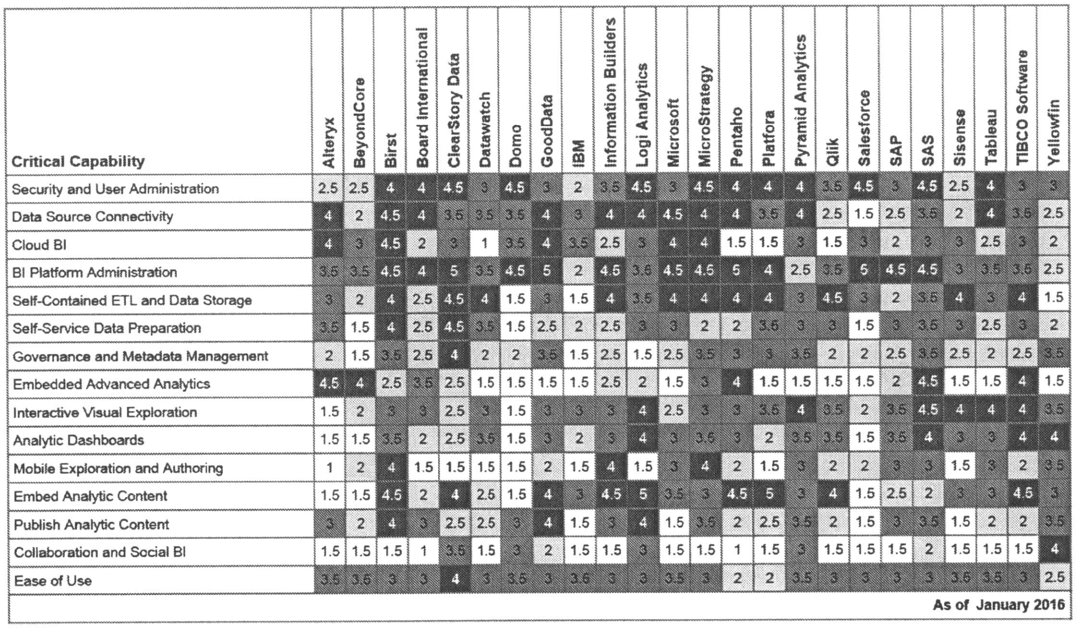
Business Intelligence tools or software have been extensively used for the past two decades. The key strength of business intelligence lies in its ability to create interactive dashboards and content. The main idea is data exploration using charts, graphs, and other visual objects in conjunction with colors, shapes and motion.

The key to good data visualization lies in three critical factors[39]

1. Understands the target audience
   - Simply based on the consumers of visualization tools
   - End goal of the analysis performed
2. Uses a clear framework
   - Should not leave room for syntax and semantics
   - Clear, concise and consistent interpretation of results
3. Uses the power of storytelling
   - Conveys insights to drive decision making
   - Used a dynamic form of persuasion

The balance between form and function is extremely hard to achieve. However, there are a few Business Intelligence service providers that are pushing the envelope and leading data visualization.

Figure 6



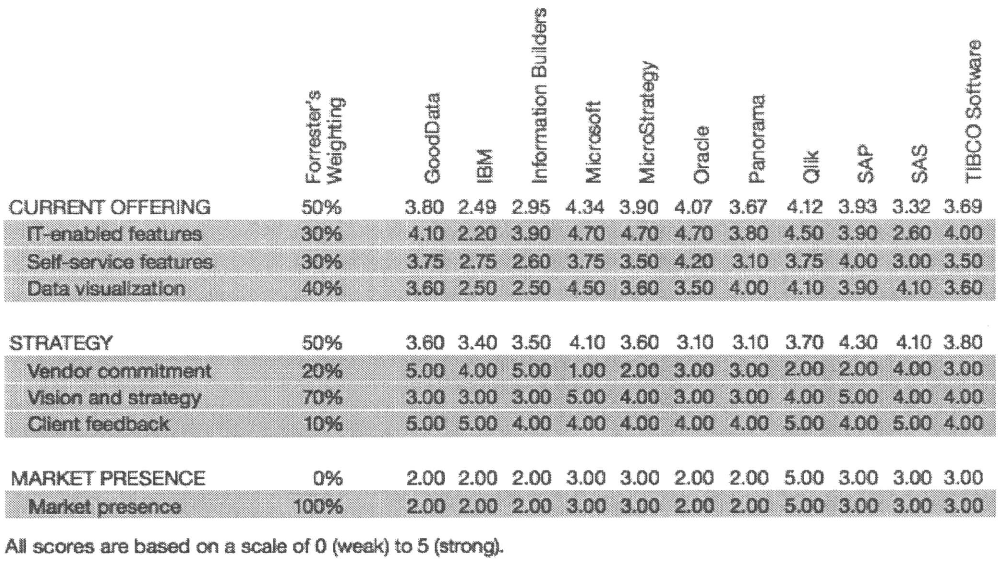
Critical Capability	Alteryx	BeyondCore	Birst	Board International	ClearStory Data	Datawatch	Domo	GoodData	IBM	Information Builders	Logi Analytics	Microsoft	MicroStrategy	Pentaho	Platfora	Pyramid Analytics	Qlik	Salesforce	SAP	SAS	Sisense	Tableau	TIBCO Software	Yellowfin
Security and User Administration	2.5	2.5	4	4	4.5	3	4.5	3	2	3.5	4.5	3	4.5	4	4	4	3.5	4.5	3	4.5	2.5	4	3	3
Data Source Connectivity	4	2	4.5	4	3.5	3.5	3.5	4	3	4	4	4.5	4	4	3.5	4	2.5	1.5	2.5	3.5	2	4	3.5	2.5
Cloud BI	4	3	4.5	2	3	1	3.5	4	3.5	2.5	3	4	4	1.5	1.5	3	1.5	3	2	3	3	2.5	3	2
BI Platform Administration	3.5	3.5	4.5	4	5	3.5	4.5	5	2	4.5	3.5	4.5	4.5	5	4	2.5	3.5	5	4.5	4.5	3	3.5	3.5	2.5
Self-Contained ETL and Data Storage	3	2	4	2.5	4.5	4	1.5	3	1.5	4	3.5	4	4	4	4	3	4.5	3	2	3.5	4	3	4	1.5
Self-Service Data Preparation	3.5	1.5	4	2.5	4.5	3.5	1.5	2.5	2	2.5	3	3	2	2	3.5	3	3	1.5	3	3.5	3	2.5	3	2
Governance and Metadata Management	2	1.5	3.5	2.5	4	2	2	3.5	1.5	2.5	1.5	2.5	3.5	3	3	3.5	2	2	2.5	3.5	2.5	2	2.5	3.5
Embedded Advanced Analytics	4.5	4	2.5	3.5	2.5	1.5	1.5	1.5	1.5	2.5	2	1.5	3	4	1.5	1.5	1.5	1.5	2	4.5	1.5	1.5	4	1.5
Interactive Visual Exploration	1.5	2	3	3	2.5	3	1.5	3	3	3	4	2.5	3	3	3.5	4	3.5	2	3.5	4.5	4	4	4	3.5
Analytic Dashboards	1.5	1.5	3.5	2	2.5	3.5	1.5	3	2	3	4	3	3.5	3	2	3.5	3.5	1.5	3.5	4	3	3	4	4
Mobile Exploration and Authoring	1	2	4	1.5	1.5	1.5	1.5	2	1.5	4	1.5	3	4	2	1.5	3	2	2	3	3	1.5	3	2	3.5
Embed Analytic Content	1.5	1.5	4.5	2	4	2.5	1.5	4	3	4.5	5	3.5	3	4.5	5	3	4	1.5	2.5	2	3	3	4.5	3
Publish Analytic Content	3	2	4	3	2.5	2.5	3	4	1.5	3	4	1.5	3.5	2	2.5	3.5	2	1.5	3	3.5	1.5	2	2	3.5
Collaboration and Social BI	1.5	1.5	1.5	1	3.5	1.5	3	2	1.5	1.5	3	1.5	1.5	1	1.5	3	1.5	1.5	1.5	2	1.5	1.5	1.5	4
Ease of Use	3.5	3.5	3	3	4	3	3.5	3	3.5	3	3	3.5	3	2	2	3.5	3	3	3	3	3.5	3.5	3	2.5

As of January 2016

Source : Gartner 2016

Evidently, Gartner's study indicates that Yellowfin, Tibo, SAS and Logi Analytics are outperforming others in the development of Analytical Dashboards. A similar study performed by Forrester indicates that Microsoft's data visualization is rated higher than competition (refer figure 7)

Figure 7



	Forrester's Weighting	GoodData	IBM	Information Builders	Microsoft	MicroStrategy	Oracle	Panorama	Qlik	SAP	SAS	TIBCO Software
CURRENT OFFERING	50%	3.80	2.49	2.95	4.34	3.90	4.07	3.67	4.12	3.93	3.32	3.69
IT-enabled features	30%	4.10	2.20	3.90	4.70	4.70	4.70	3.80	4.50	3.90	2.60	4.00
Self-service features	30%	3.75	2.75	2.60	3.75	3.50	4.20	3.10	3.75	4.00	3.00	3.50
Data visualization	40%	3.60	2.50	2.50	4.50	3.60	3.50	4.00	4.10	3.90	4.10	3.60
STRATEGY	50%	3.60	3.40	3.50	4.10	3.60	3.10	3.10	3.70	4.30	4.10	3.80
Vendor commitment	20%	5.00	4.00	5.00	1.00	2.00	3.00	3.00	2.00	2.00	4.00	3.00
Vision and strategy	70%	3.00	3.00	3.00	5.00	4.00	3.00	3.00	4.00	5.00	4.00	4.00
Client feedback	10%	5.00	5.00	4.00	4.00	4.00	4.00	4.00	5.00	4.00	5.00	4.00
MARKET PRESENCE	0%	2.00	2.00	2.00	3.00	3.00	2.00	2.00	5.00	3.00	3.00	3.00
Market presence	100%	2.00	2.00	2.00	3.00	3.00	2.00	2.00	5.00	3.00	3.00	3.00

All scores are based on a scale of 0 (weak) to 5 (strong).

Source: Forrester Research (Evelson B., 2015)

Some of the most common forms of data visualization are:

1. Bar graphs, pie charts, line charts, histograms
2. Scatterplots
3. Heat-maps
4. Gantt charts
5. Treemap
6. Network diagram
7. Decision trees

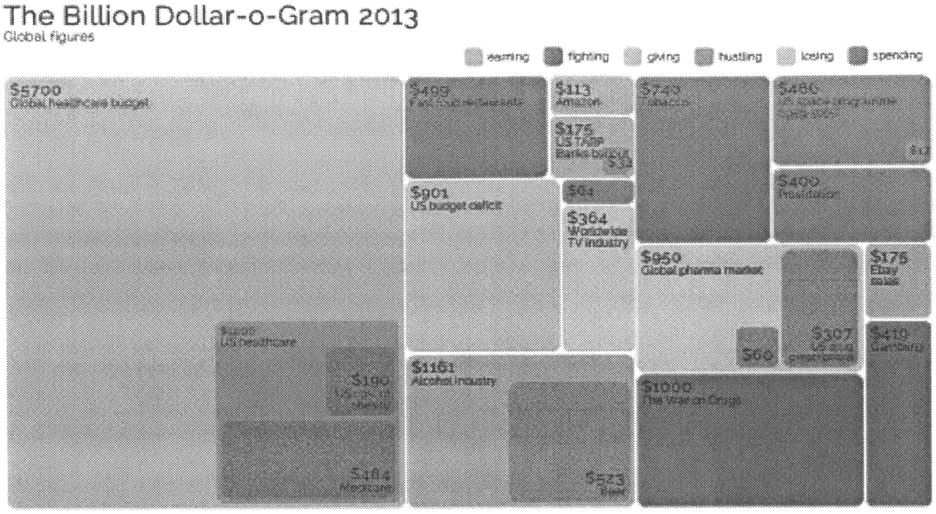
3.4.2 THOUGHT LEADERS

The emergence of new technologies has started the shift of data visualization away from art, towards science. The advances in multidimensional imaging combined with newer data sources, such as data from intelligent devices, is leading to the development of newer visualization methods.

Some of the noteworthy representations of data in recent years have come from thought leaders, who are responsible for using genius creative methods to arouse audience interest and memory. This section covers some of those visualization innovations.

1. **David McCandless's Billion Dollar O gram**

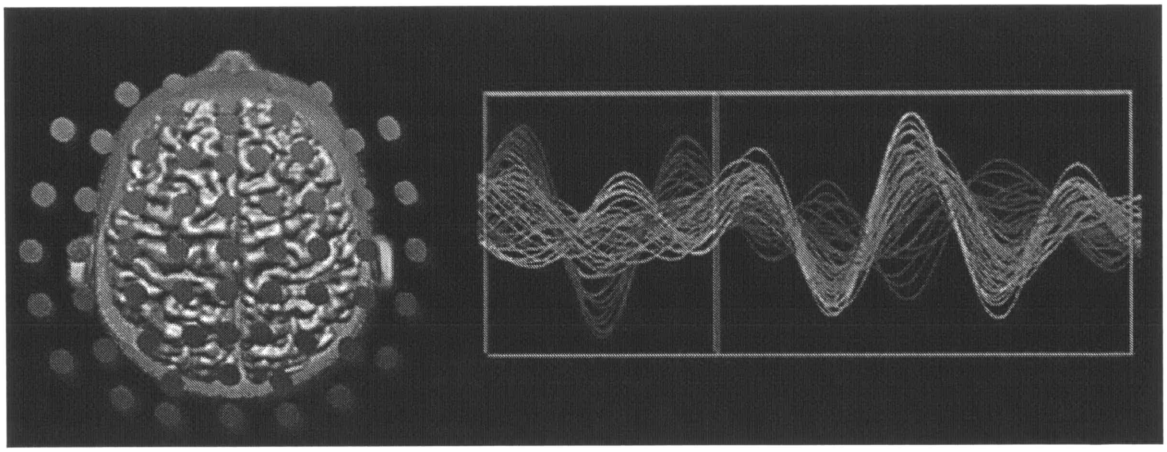
Figure 8



Source: Information Is Beautiful

2. 5D colorimetric technique

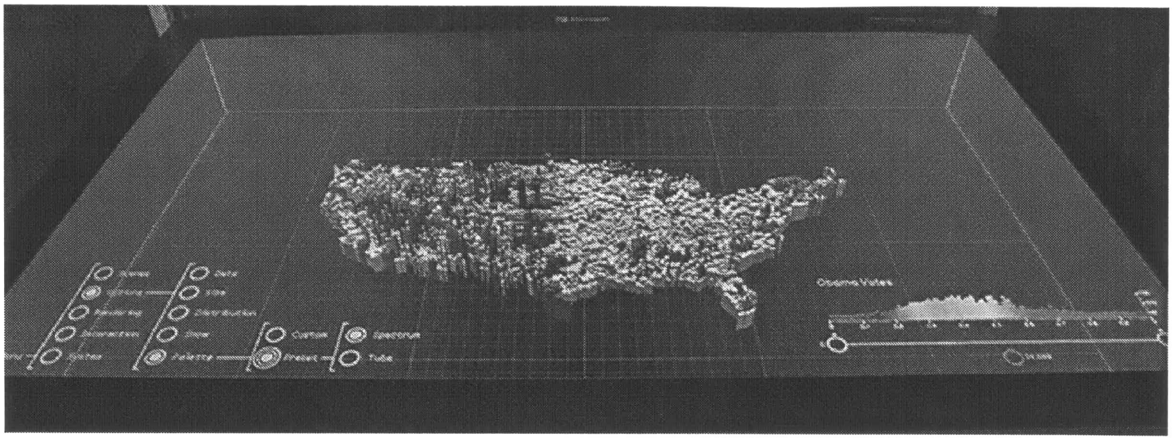
Figure 9



Source: Florida Atlantic University, Center for Complex Systems and Brain Sciences

3. Microsoft's Holograph

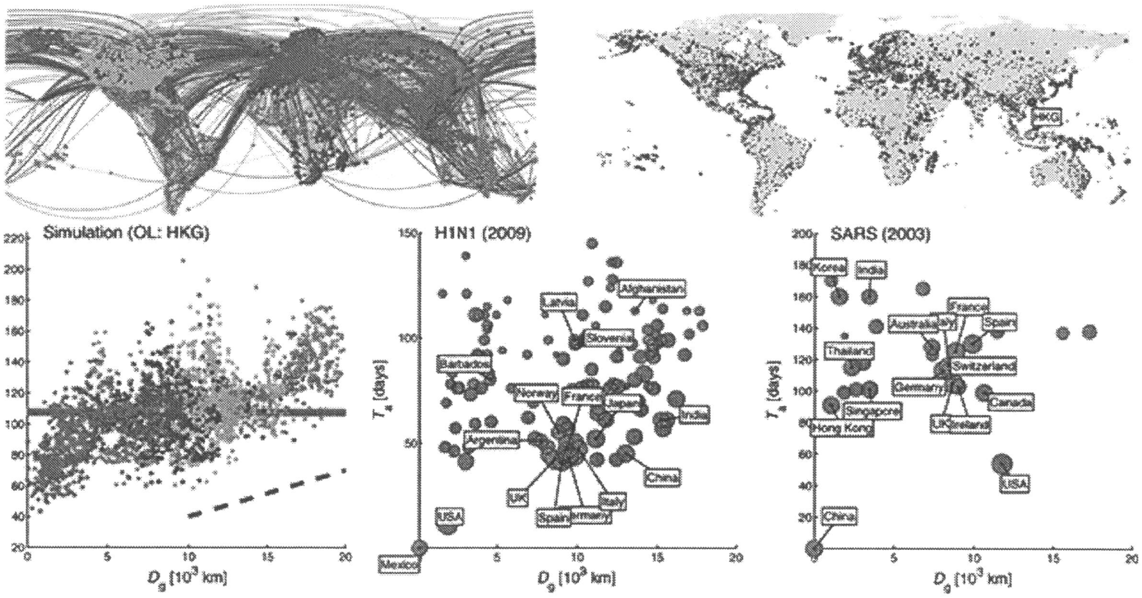
Figure 10



Source: Microsoft

4. Sociograms

Figure 11



Source: Science

Dashboarding of the future would provide powerful insights on the go. Mobile is expected to a promising channel of dashboard delivery. It is also expected that dashboards would be more interactive in nature and allow for the consumer to toggle with various levels to make further interpretations of their own.

Several companies are emerging in the field of data visualization. It is an easy prediction that the next 3 to 5 years would see strong innovations coming in the field of data visualization.

4. The Quest for Optimization

The objective of big data analytics was to assist decision-making. As time progresses, data analytics is getting more and more embedded into corporate decision-making, to a point where predictive analytics leads to prescriptive analytics, thereby eliminating the need for human intervention. As Artificial Intelligence develops further, technology could analyze, decide optimal course of action and proceed with implementation.

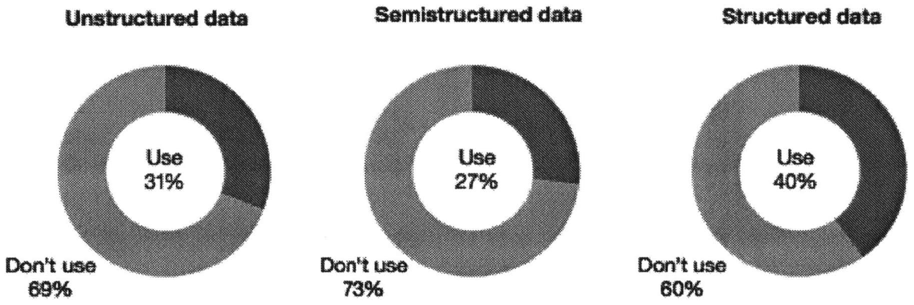
Technological Singularity is defined as the hypothetical event where artificial intelligence would autonomously build machines so powerfully intelligent that their functioning and abilities would reach beyond the scope of human comprehension. The timeline of the event is estimated to be around 2040 as the pace of technological evolution operates on the law of accelerating returns and humanity has experienced exponential technological growth, thus far. Whether machine learning can lead to technological singularity or whether singularity as a concept could exist, are all topics of lively debate. The one unanimous point of agreement is that the pace of technology is extraordinarily high. In this section, we will explore the many innovations and present a forecast of the expected state of the world in the next 3-5 years.

4.1 Ideal World

Despite roughly two decades of effort in developing technologies to leverage big data, most corporations are still underutilizing all the data at their disposal. There are two contrasting viewpoints on the subject of data utilization. One viewpoint states that all data is useful and it is technology's inability to extract useful information from it, which prevents us from using it. The other viewpoint states that not all data is useful and a part of technology's job is to identify useful data from junk data. With improvements in the field of computer science using advanced mathematics, we might arrive at a conclusion to this debate soon enough.

Nonetheless, the non-debatable aspect of big data analytics is that corporations are not at optimal levels of data utilization.

Figure 12: Percentage of the total size/ volume of internal data currently using BI



Base: 1,805 global technology decision-makers who know how much BI data their firm uses

Note: The percentages shown are estimates based on reported ranges; the values are not exact.
Source: Forrester's Business Technographics® Global Data And Analytics Survey, 2015

In this section we will explore a sector that leads the pack in terms of big data analytics – E-commerce. E-commerce companies invest heavily towards user-centric analytics. Some of the most important consumer behavior led retail innovations are coming from the fact that E-commerce data science teams invest most of their resources to identify user needs. Additionally, supply chain optimization and operational efficiency are other key areas of success for E-commerce. This does not necessarily mean that E-commerce lags in other forms of analytics such as People analytics, Fraud prevention or Procurement analytics. However, it is fair to assume that E-commerce is not an analytics leader in those areas. The next section focuses on E-commerce customer analytics in greater depth.

Besides E-commerce, two other sectors that stand to gain tremendously from enhancements in big data analytics are Healthcare and Financial Services.

Healthcare is a relatively new entrant in the field of analytics. Interestingly, most healthcare service providers (Hospitals) in the USA are yet to fully implement Electronic Health Records, a form of ERP system to collect user data on a centralized server

(Charles D., Gabriel M., & Searcy T., 2015, April). It can be considered a laggard in the field of enterprise software adoption. At the same time, the benefits of predictive analytics in disease prevention and epidemic containment are too many to ignore. Unstructured data analytics has one of the largest applications in Healthcare sector with IBM Watson contributing in cancer research.

Financial Services have always used data analytics. Currently, the focus is threefold – (1) Eliminate operational inefficiencies by uniting disparate systems to consolidate data and streamline its flow between trading desks. (2) Develop new products and target existing products to a wider audience by conducting customer analytics to identify preferences and market movement, (3) Ensure regulatory compliance by proactively identifying potential non-compliances through natural language processing and other tools.

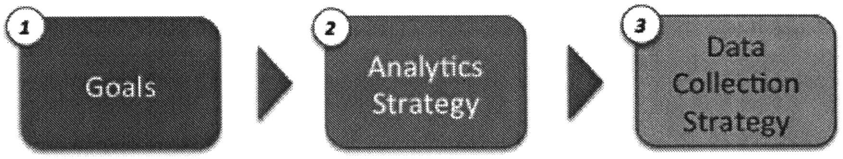
In addition to the above two sectors, online marketplaces, particularly services offered by Uber, Tinder, AirBnB conduct extensive analytics spanning user behavior, dynamic pricing, matching et cetera. The user analytics aspect of these services draws a lot of inspiration from E-commerce, however, there are other uniquely interesting elements that are worthy of consideration.

4.2 E-COMMERCE

E-commerce industry is considered one of the most advanced industries in the adoption of user-centric technologies. Due to the nature of this business, customer interaction is exclusively handled through an online portal, website or mobile app or both. Offline customer interaction is limited to product delivery at a customer's doorstep. This interaction is also recorded with a high degree of accuracy. Website and mobile analytics tools have allowed e-commerce companies to record every action by the customer. Each action is termed as an 'event' and used as an input into machine learning algorithms for predictive or prescriptive analytics.

Analytics is embedded at the core of an E-commerce business. To simplify the various elements involved, the thesis uses a simplistic framework exhibited in Figure 14. Corporate goals lead to the development of an analytics strategy, that includes decisions on analytics technology, infrastructure requirements, investment requirements et cetera. Analytics strategy in turn affects customer data collection strategy. By data collection strategy, I mean the decision regarding which data points should be collected as a necessary precursor to performing meaningful analytics.

Figure 13



Interestingly, the trend seems to have started to shift lately. As numbers of data sources have increased, the current industry focus is on finding better ways to utilize all data. As a matter of fact, the core methods of data collection have not changed much in the last decade. To analyze E-commerce analytics needs better, the thesis would take a bottom-up approach –

3. Data Collection Strategy

Most large online retailers reach their customers through two channels – Web and Mobile. Mobile is especially relevant in emerging markets such as India where smartphone penetration is higher than desktop penetration (Meeker M., 2015).

- **Website Data Collection**
    - Web traffic is divided into 4 groups – search, referral, campaign and direct
    - Search traffic refers to all the traffic that comes through search engines, Referral traffic is traffic linked through another website, Campaign traffic is traffic that comes with an advertiser's tag and Direct traffic is organic traffic. It also includes untagged or unidentifiable source traffic
    - The following key attributes are collected about the customer
        - Traffic category – prompt that led user to the website
        - Funnel position – user's last position in the sales funnel
        - User journey – browsing behavior such as number of page views pre checkout, number of items compared et cetera
    - Web data collection is primarily performed (Murdock K., 2006)
        - Log File Method: Log files are tracking files stored on a web host's server that record visitor behavior and can be analyzed
        - JavaScript Method: A JavaScript code is included on each webpage. Visitor activity is recorded by the JavaScript and

transmitted to a web server hosted by a web analytics service provider such as Google Analytics

- **Mobile Data Collection**
  - Two sets of activities are recorded – screen tracking and event tracking
  - Screen tracking refers to the user's movement between various screens of the mobile app and Event tracking refers to specific action taken by user such as button click, menu selections, ad clicks, video plays, swipes or other gestures such as touch and pinch
  - Additional data such as model, carrier, screen size, operating system, geographic location is also collected
  - Mobile data collection is done through integration of a third party SDK (software development kit) specifically designed for app tracking. SDK is integrated with mobile app by the developer and has no perceivable difference on the customer's end. Some of the leading SDKs are Mobile App Tracker (MAT), Appsflyer et al.
  - SDK collects data by creating a unique device fingerprint when a user downloads a mobile app. It records every action and saves it on a tracking server. The same SDK that collects in-app data can be used for advertising attribution, especially in cases of driving app installs or remarketing
- **Social Media Data**
  - Facebook page insights
  - Twitter data – interaction with tweets made E-commerce company, tweets made about E-commerce company, trending topics et cetera
- **Offline Retail Data**
  - For E-commerce companies with a brick & mortar presence, such as Walmart, Homedepot et al, offline retail data at point of sale is equally important in customer analytics
  - Usually captured by user with loyalty cards/ email IDs/ phone numbers
- **Marketing Engagement Data**
  - Data collected by customer's interaction with email promotions, blog links on social media, paid content, direct advertising

It is important to note that the data sources captured in Section 3.1.2 are entirely applicable to the E-commerce sector. The data unique to E-commerce sector has an impact on collection methodology; however, the fundamentals remain quite the same.

2. Analytics Strategy

With the wealth of data at its disposal, online retailers have set out to achieve higher customer satisfaction, increased sales and decreased costs. Several arms of the analytics and product management divisions are dedicated towards activities using analytics insights. E-commerce companies perform structured/ unstructured data analytics, employ several niche analytics service providers and invest heavily in development of better algorithms for machine learning. The focus of this section is to highlight current challenges with analytics strategies.

Table 5

Function	Sub-function	Analytics Strategy	Shortcomings
Customer Centric Enhancements	• Personalization • Loyalty Management • Cross-sell/ Up-sell • Campaign Management • Pricing & Offer Management	• Develop a deep understanding of every customer as a unit using web, mobile, social media and email analytics	• Incomplete view of customer behavior across devices and offline • Retailers are still using a larger segmentation model, not treating each customer unit uniquely
Operational Improvements	• Inventory Management • Catalog Management • Order Management • Payment Service Management	• Improve supply chain effectiveness, demand forecasting and operations nimbleness using order-delivery analytics, logistics analytics	• Logistics algorithms are not fully efficient, especially for small & medium size online retailers • Demand forecasts and inventory levels not optimal due to poor predictive models
Platform Improvements	• Content Management • User Interface Design • Platform	• Make user-experience based changes to platform using web/ mobile analytics, eye tracking, click tracking, AB testing &	• Incomplete view of customer behavior across devices and offline

	Extensions	user optimization testing • Introduce new product enhancements (such as Amazon dash button) based on customer feedback and need assessment	• Incremental changes made to platform, no major breakthroughs in user experience

Presently, in an attempt to utilize all the different sources of information, analytics strategy is extremely disparate. Companies such as Amazon use a common login ID for mobile and web platforms, hence, its possible to combine user behavior from both to identify one user's behavioral pattern across devices. Many online retailers do not have a mobile application and rely on mobile optimized sites for extending their web offering on mobile. In such cases, user behavior is not tracked across devices.

Social media analytics is an important component of E-commerce analytics. However, the view offered is not comprehensive. Let me illustrate how,

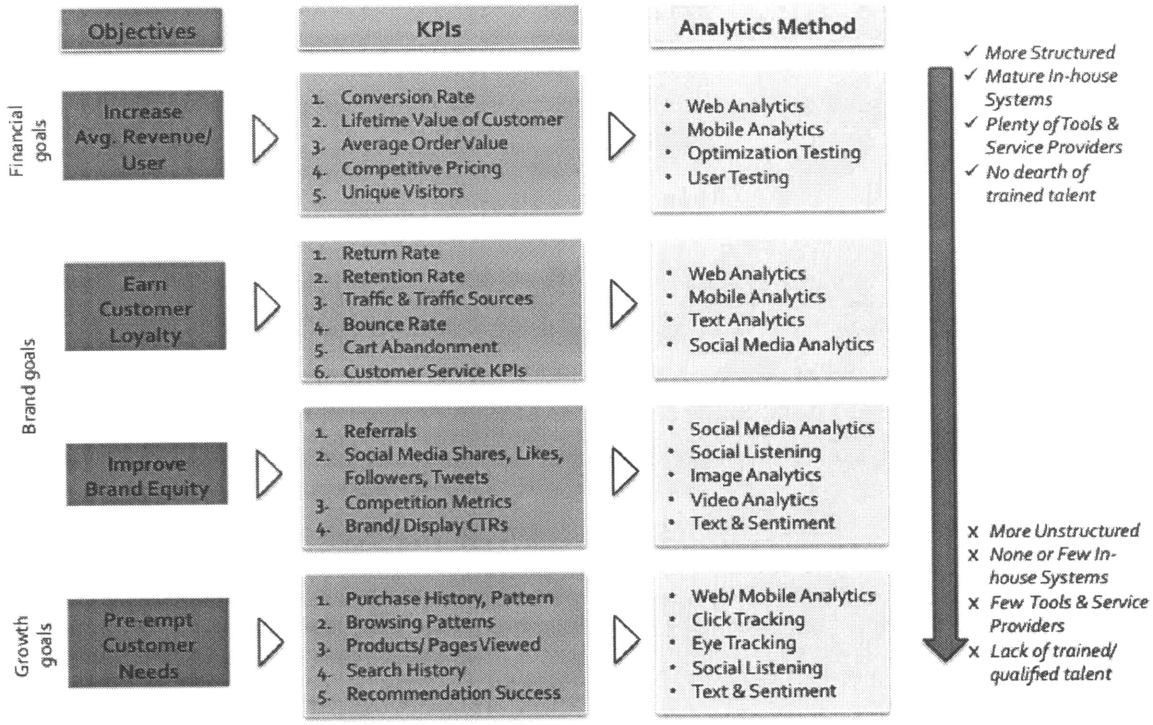
- Facebook's 'Page Insights' is used by in-house marketing for assessing user activity on Facebook page
- Twitter engagement is measured using hashtags and direct links to company's twitter handle
- Meta data analytics is performed on user generated content on YouTube and Instagram, for example, companies like Sephora could measure influencer popularity by analyzing number of views, likes and comments of makeup tutorial videos in addition to analyzing meta data such as location, time of post, file size
- Trend analytics is performed separately using social media trending topics and popular news of the day/ week

Sales & Marketing teams view the above analytics dashboards. The dashboards do not present a comprehensive view of the world but provide a disjointed yet interconnected story of the brand, user preferences and upcoming trends. Decision making is assisted, not yet automated.

1. Goals

The framework in Figure 14 exhibits the relationship between an e-commerce company's corporate goals with the Sales & Marketing performance metrics (or KPIs) that indicate the achievement of those goals. The measurement of performance metrics is achieved by analytics that in turn uses customer level data to achieve or exceed KPIs.

Figure 14*:



	Objectives	KPIs	Analytics Method
Financial goals	Increase Avg. Revenue/ User	1. Conversion Rate 2. Lifetime Value of Customer 3. Average Order Value 4. Competitive Pricing 5. Unique Visitors	• Web Analytics • Mobile Analytics • Optimization Testing • User Testing
Brand goals	Earn Customer Loyalty	1. Return Rate 2. Retention Rate 3. Traffic & Traffic Sources 4. Bounce Rate 5. Cart Abandonment 6. Customer Service KPIs	• Web Analytics • Mobile Analytics • Text Analytics • Social Media Analytics
	Improve Brand Equity	1. Referrals 2. Social Media Shares, Likes, Followers, Tweets 3. Competition Metrics 4. Brand/ Display CTRs	• Social Media Analytics • Social Listening • Image Analytics • Video Analytics • Text & Sentiment
Growth goals	Pre-empt Customer Needs	1. Purchase History, Pattern 2. Browsing Patterns 3. Products/ Pages Viewed 4. Search History 5. Recommendation Success	• Web/ Mobile Analytics • Click Tracking • Eye Tracking • Social Listening • Text & Sentiment

✓ *More Structured*
✓ *Mature In-house Systems*
✓ *Plenty of Tools & Service Providers*
✓ *No dearth of trained talent*

x *More Unstructured*
x *None or Few In-house Systems*
x *Few Tools & Service Providers*
x *Lack of trained/ qualified talent*

**'Analytics method' refers to a broad category of analytics folded up to a bigger category. For example, Web Analytics includes funnel analytics, referral channel analytics, email marketing funnels et al.*

It is evident from the framework that as companies re-adjust their focus from business sustenance (after having achieved Financial goals) to business continuity (growth goals), the expected outcome from big data analytics shifts from Diagnostic and Discovery analytics to Predictive and Prescriptive analytics. Predictive analytics is the hottest technological movement in online retail. E-commerce relies on the understanding of shopping behavior as a proxy for understanding customer segments, product popularity and service needs. Predictive analytics uses shopping behavior to preempt customer's upcoming shopping needs. It also allows for better service by indicating areas of platform improvement.

4.2.1 End Game: The Comprehensive Analytics Solution

An ideal comprehensive utopian analytics solution to meet E-commerce Sales & Marketing requirements in the next half a decade should have the following attributes:

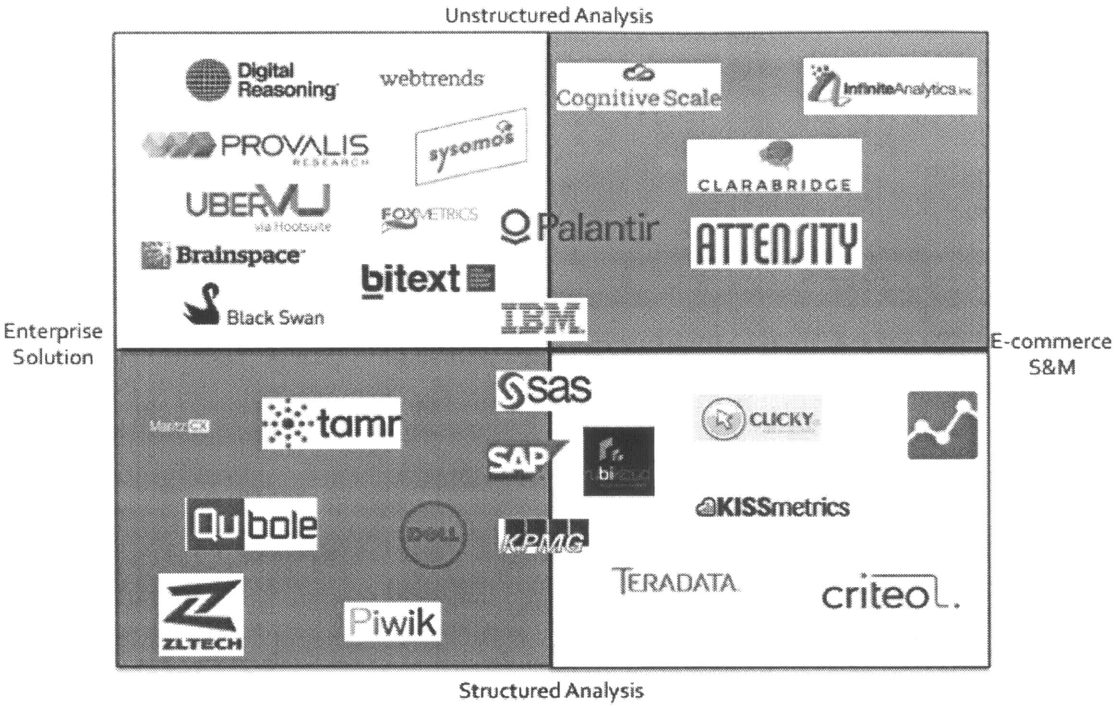
1. **Price**
   - That generates **measurable** ROI
2. **Data Visualization**
   - Single dashboard with **real time** metrics
   - Viewable **across devices** – desktop/ mobile/ on the go
   - High degree of **flexibility** and user **control** – multiple views, cross-references, sliced and diced, selectable at the push of a button
   - Insights **augmented** with implications and next step recommendations
3. **Data Storage & Processing**
   - Uses existing BI platform
   - No major infrastructure investment needed
   - Preferably a cloud based, software-as-a-service model
   - No additional headcount & training requirements
4. **Data Acquisition**
   - Internal data – structured & unstructured
     - Includes customer data across devices and channels
   - External data – structured & unstructured
     - Includes social media data
     - Includes review boards inputs

4.2.2 CURRENT SOLUTION PROVIDERS

The market for E-commerce analytics service is cluttered with solutions and service providers. Figure 16 gives a high-level view of the market. Companies such as Kissmetrics, Criteo, Infinite Analytics are focused on the E-commerce sector with solutions catered to solving E-commerce Sales & Marketing challenges.

Currently, none of the 75 companies analyzed (refer Appendix) present an end-to-end Sales & Marketing solution for E-commerce. Each solution provider solves a piece of the puzzle. Mobile app commerce has added an extra layer of complexity. Some solution providers are currently working on a holistic solution that combines external and internal unstructured data with social media (including online review boards such as Glassdoor and IMDB) to develop a more comprehensive view of the customer's likes and dislikes.

Figure 15



As cluttered as the market is, in my opinion, there is space for a truly revolutionary analytics technology that can provide a comprehensive solution.

4.2.3 Who To Target?

A Comparative Assessment

This section compares the relative analytics sophistication of three online retailers at three stages of evolution based on secondary research, primary interviews and online questionnaire. The assessment covered Amazon, Sears, Flipkart, Homedepot, eBay, Macy's, Walmart, Myntra and Target.

The table below summarizes the results of subjective relative evaluation. The idea is to compare and contrast the relative technological sophistication of the online retailers. It is important to note that some online retailers have strong offline presence and their business intelligence teams cater to both online and offline strategies. The focus for such retailers is to develop a strong omni-channel analytics strategy. Barring a few differences pertaining to offline channel, the analytics fundamentals between pure online retailers such as Amazon and combination retailers such as Walmart remain unaltered. Table 5 uses Harvey balls for relative comparison.

Low ○ ◔ ◑ ◕ ● ***High***

Table 5

	Amazon	Sears	Flipkart	Target	Myntra	Walmart	eBay
Use of internal and external data	◕	◑	◔	◕	◔	◕	◑
Unstructured data analytics	◕	◔	○	◑	○	◑	◔
Cross-device user identification	●	◔	●	◑	●	◑	◑
Advanced analytics infrastructure	●	◑	◔	●	◔	●	◑
In-house analytics capabilities	●	◑	◔	●	◔	●	◕
Analytics driven strategy	●	◑	◑	◕	◑	●	◑
Relative size of BI team	●	◕	◔	●	◔	●	◕

The retailers that present an interesting contrast are Amazon, Sears and Flipkart.

- Amazon is the most valuable retailer in USA (La Monica P., 2015) with a market capitalization of $280 B (Yahoo Finance, April 2016). It is the largest Internet based retailer in the US and has a worldwide presence. It was founded in 1994
- Sears is a 130-year-old American retail chain. It entered E-commerce in 1999 with sears.com. It does not figure in the top 10 online retailers in the US market (Wahba P., 2015)

- Flipkart is an Indian e-commerce website founded in 2007 and valued at $11 B (Choudhary S. R., 2016) It is the market leader in the vast and rapidly growing Indian market.

The beachhead target segment for an E-commerce analytics service provider would depend upon the following key factors:

- Sophistication of current analytics technology used by E-commerce customer
- Type of offering – cloud based/ dashboard service/ multiple sources et al
- Ease of implementation – includes time to implement and resources required

The above three E-commerce companies are at three levels of technological sophistication and have different degrees of purchasing power. An entrant must evaluate customer needs and solutions offered on a case-by-case basis.

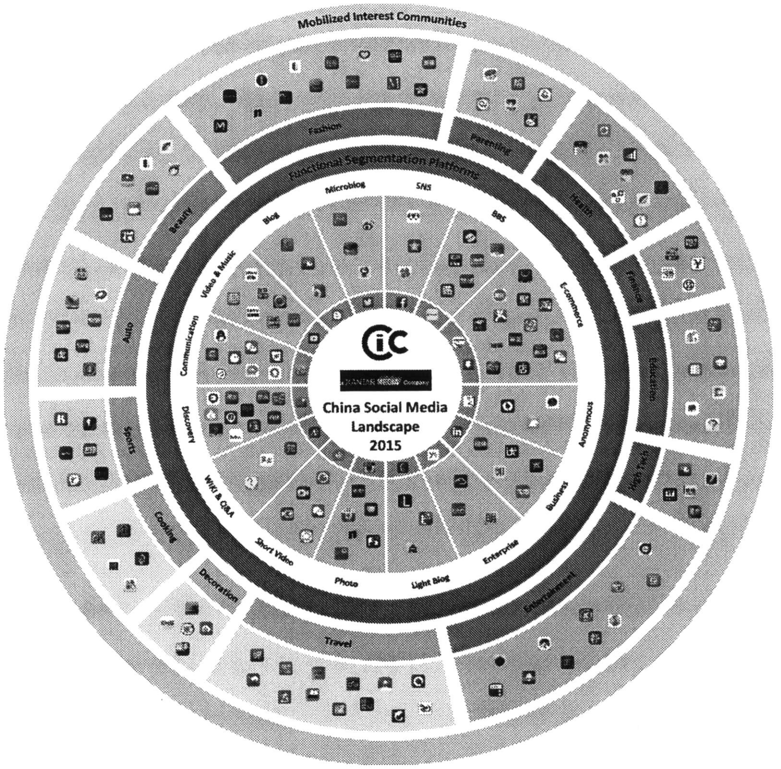
5. A PARALLEL UNIVERSE

There is a certain amount of uniformity in the data collected for running standard businesses across the world. The elements of Big Data Infrastructure (refer figure 1) focusing on data storage, processing and use are standard and replicable across geographies. The only major difference lies in data acquisition in the context of the Chinese market. It is rather remarkable that in this time and age, China has managed to wall off the world's leading Internet businesses by creating substitutes for them. Interestingly, some might argue that Chinese Internet businesses are run more effectively, provide superior products and generate higher profits. As we discuss data sources, it is essential to touch upon the Chinese Internet industry as the scale and volume of the market is too big to ignore.

The Chinese trinity is widely referred to as BAT. The Chinese counterparts of global businesses have a remarkable scale and volume, making them extremely formidable sources of data for any business or technology interested in the Chinese market. This section aims to understand the structure of the Chinese Internet industry and identify ways in which their data can be accessed for entering the market or creating products targeted at the Chinese consumer.

BAT comprises of Baidu (China's counterpart for Google), Alibaba (China's counterpart for Amazon) and Tencent (China's social media with services such as WeChat and JD.com). The Internet industry in China is extremely complicated with recurring ownership theme of one of the BAT members.

Figure 17



Source: CIC Chinese Social Media Landscape 2015

5.1 WeChat

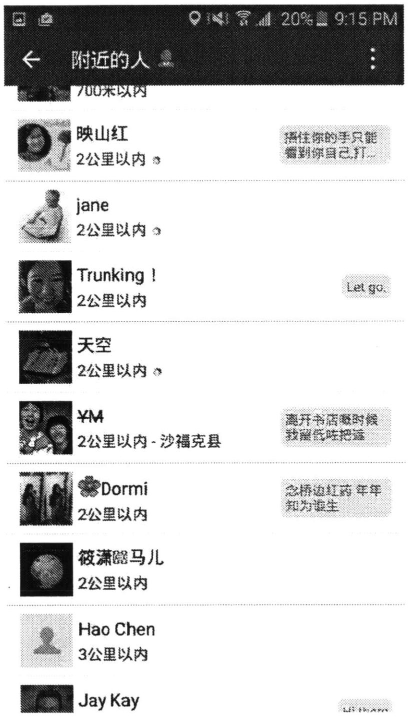
WeChat is a mobile text and voice messaging communication service with an active user base of 697 M per month (China Internet Watch, Oct-Dec 2015). Some of the interesting formats and characteristics of WeChat include

- Instant messaging: Text, voice and stickers. WeChat is credited with the most well executed introduction of stickers messaging
- Public Accounts: Enables interaction with subscribers to provide a service
- Moments: Similar to Facebook's 'wall'. Users can post images, text, share music or links to articles, as well as comment and "like" posts made by their friends

5.1.1 Users Characteristics

WeChat enjoys a young demographic with 86.2% of its users between the age of 18 and 36 (Statista, Jan 2016). The gender split is skewed towards male users with 64.3% of all users being male and only 35.7% being female (Grata, 2015). Users are largely educated urban, semi-urban dwellers. WeChat enjoys a high level of engagement with 55.2% of its users opening the WeChat app over 10 times per day and about 25% of its users opening the app over 30 times per day (Grata, 2015). This level of engagement far exceeds numbers reported by Facebook, Instagram or Snapchat.

Figure 18



700米以内

映山红
2公里以内
握住你的手只能看到你自己,打...

jane
2公里以内

Trunking !
2公里以内
Let go.

天空
2公里以内

¥M
2公里以内 - 沙福克县
离开书店嘅时候我留低咗把遮

Dormi
2公里以内
念桥边红药 年年知为谁生

筱潇囧马儿
2公里以内

Hao Chen
3公里以内

Jay Kay

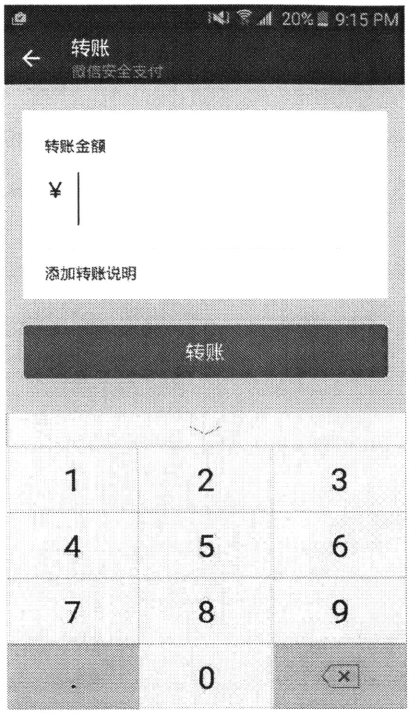
5.1.2 Unique Offerings – Beyond Social Media

WeChat financial products

1. **WeChat Pay**

- Enables payment of monthly household and utility bills such as internet and phone bills
- Also facilitates transaction between users. WeChat was the first messaging service to integrate a payments feature. This feature has been wildly successful in China

Figure 19



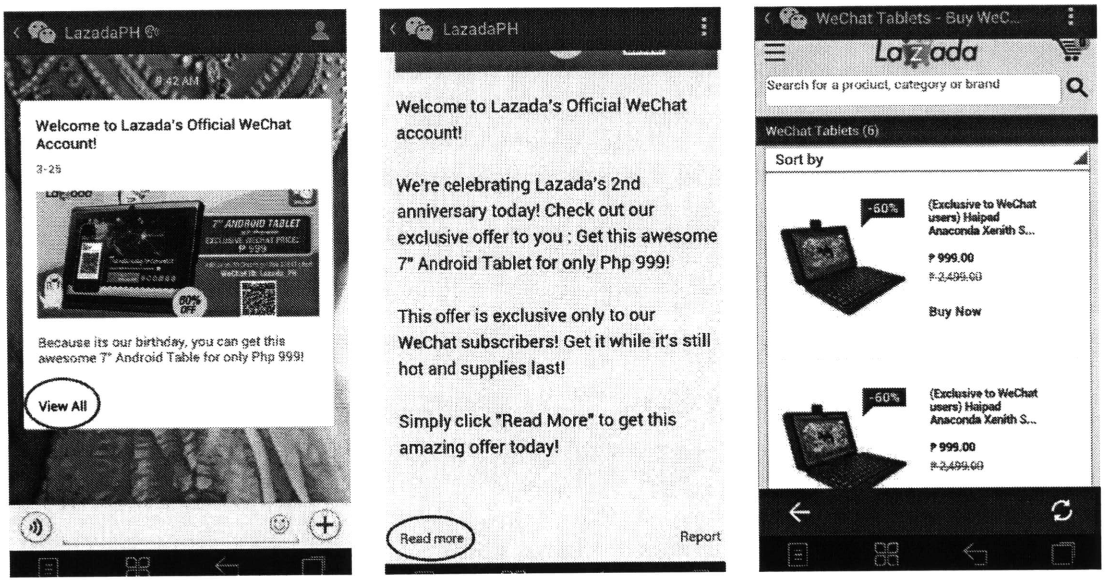
2. **Financial products**
   - Users can buy mutual funds, index fund and a few other financial products. WeChat is reportedly working on a feature that would facilitate stock trading. It is interesting to note that all major banks in China have a WeChat public account. Users can follow their bank's public accounts and carry out all online banking transactions through WeChat's interface. More specifically, users can check their account balance, pay credit card bills, receive bank account statements et cetera
   - Tencent is the biggest shareholder of WeBank with a 30% stake

3. **Small Loans**
   - A personal loan product Weilidai was launched by WeBank to allow users to borrow up to 200,000 yuan ($31,350) without guarantee or collateral
   - Using the new Weilidai feature, WeChat users can receive money in just a few minutes after submitting their applications. They apply for loans by giving their traditional bank account information and other basic personal data, and Weilidai assesses their credit history
   - In late July 2015, WeBank announced that its outstanding personal micro loans amounted to 800 million yuan, without disclosing the number of borrowers

- Currently, only users who receive an invitation can apply Welidai through WeChat. Reports suggest that the number of users who would receive an invitation in the first round is less than 10 million

4. **WeChat E-commerce**
   - Weidian is WeChat's e-commerce platform
   - It was launched on January 1st 2014 and has since grown to over 29M online stores
   - The platform is still in the early stages of development. According to a report by Tencent in January 2015, WeChat's influence on lifestyle consumption is of the order of $1.76B with shopping accounting for 13.2%

Figure 20



Welcome to Lazada's Official WeChat Account!

3-25

Because its our birthday, you can get this awesome 7" Android Table for only Php 999!

View All

Welcome to Lazada's Official WeChat account!

We're celebrating Lazada's 2nd anniversary today! Check out our exclusive offer to you : Get this awesome 7" Android Tablet for only Php 999!

This offer is exclusive only to our WeChat subscribers! Get it while it's still hot and supplies last!

Simply click "Read More" to get this amazing offer today!

Read more

Report

Search for a product, category or brand

WeChat Tablets (6)

Sort by

(Exclusive to WeChat users) Haipad Anaconda Xenith S...

₱ 999.00

~~₱ 2,499.00~~

Buy Now

(Exclusive to WeChat users) Haipad Anaconda Xenith S...

₱ 999.00

~~₱ 2,499.00~~

5.1.3 WeChat Database

WeChat's user database is by far the most comprehensive and extensive user database in the world. Facebook and Google are large players with access to micro-level details, however, nothing compares to the depth and width of data that WeChat carries for its users. Based on the services offered, one can make a fair assessment of the kind of data available at WeChat's disposal:

- Social network of friends – high interaction to low interaction
- Social activities – group payments, payment interaction between friends

- Utility consumption pattern – Internet, mobile, water, gas, electric bills
- Income and savings profile
- Debt profile
- Online purchase history
- Social behavior through 'moments'
- Product reviews and feedback

Suffice to say that in the presence of a central repository of interconnected data, WeChat's user profiling could become a lot superior to Facebook's profiling. Facebook relies on self-reported data and tries to draw inferences about income, purchasing power, and propensity to spend on discretionary products or services. At Facebook, data scientists make inferences using probabilistic models that use established high correlations between parameters such as 'school' and 'income' or 'city' and 'purchasing power'.

WeChat could take a more deterministic approach towards building a prediction engine. As a one stop shop for several critical activities of an individual's life, WeChat stands to improve the current industry standards of click-through-rates and relevance scores.

Currently, WeChat is focused upon developing its roots outside China. Their product is quite advanced, however, adoption in rest of the world is largely impacted by the prevalence of existing social networks such as Facebook, Instagram and Snapchat.

5.2 ALIBABA

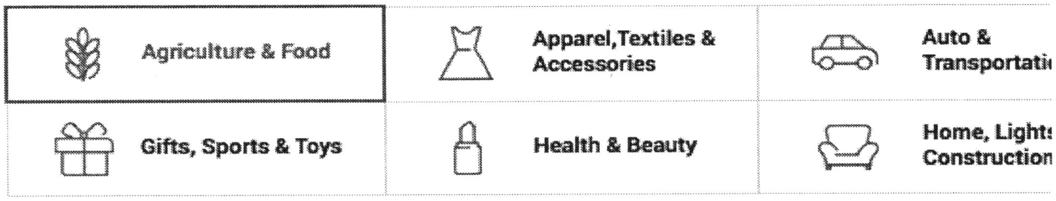
Alibaba was founded in 1999 as an e-commerce platform enabling business-to-business transaction. The main idea was that Internet would level the playing field and make the world flatter. Alibaba aimed to bring a platform for small manufacturers and entrepreneurs to expand their customer base beyond their geographic boundaries.

Alibaba Group Holding Limited has the following companies:

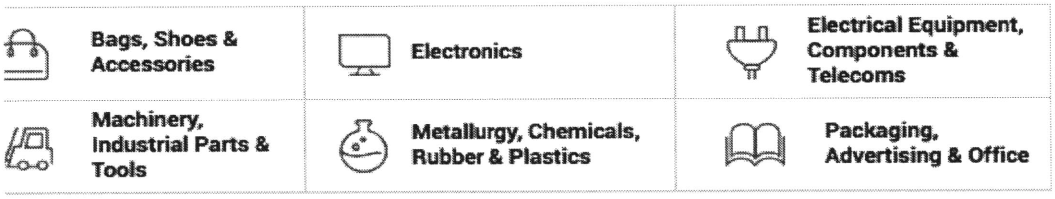
Alibaba.com

- The primary company of the Alibaba group, meant for enabling business to business transactions between small businesses around the world

Figure 21



Agriculture & Food	Apparel,Textiles & Accessories	Auto & Transportati
Gifts, Sports & Toys	Health & Beauty	Home, Lights Construction



Bags, Shoes & Accessories	Electronics	Electrical Equipment, Components & Telecoms
Machinery, Industrial Parts & Tools	Metallurgy, Chemicals, Rubber & Plastics	Packaging, Advertising & Office

Source: Alibaba.com

Aliexpress.com

- A smaller version of Alibaba.com, that allows smaller buyers to purchase goods at wholesale prices. Also operational around the world

Taobao Marketplace

- China's largest consumer-to-consumer selling platform with over 500 million registered users, 7 million merchants and a sale of 4800 items per minute (Steimle J., 2015)
- A quick glance at the website indicates that each product has anything between 10 to 1000 user comments
- Payment is handled by Alibaba's Alipay

Tmall

- Based on the concept of a brick and mortar mall, Tmall is a business-to-consumer format ecommerce website
- Tmall is home to 70,000 official stores ranging from the likes of Hugo Boss, Calvin Klein, Burberry, Costco, P&G and many more (Steimle J., 2015)
- Payment is handled by Alibaba's Alipay

The Alibaba group has diversified into various other businesses such as cloud computing and private equity. However, at the heart of the group is still the original e-commerce business. The revenue model is unique and not the same as Amazon. Alibaba is known to be the most efficient user of its online inventory and makes a large profit from reseller ads. A large contribution of this success comes from Alibaba's ability to analyze user data to identify needs and target relevant ads.

In terms of the value and volume of data available with Alibaba, it is fair to compare it with Amazon. Amazon carries data from around the world, as it is present in almost all large markets, while Alibaba's home turf remains its largest.

5.3 Baidu

Baidu was founded in January 2000 as a web services company with headquarters in Beijing. Baidu's most popular service is its Internet search engine much like Google's search service. In addition to search, Baidu offers a host of services that span across the spectrum. Table 6 attempts to map Baidu's services with the types of user data identified in section 3.1.2 Data Sources.

Table 6

Category	Service	Type of User Data
Search	Web, Image, Video, News, Web directory, Hao123.com, Dictionary, Top Searches & Search Index, Open Platform	Search and browsing history
Social	Post Bar, Space, Album	User generated content
UGC based knowledge products	Knows, Encyclopedia, Wenku, Experience	User generated content
Location based services	Maps, Group buy directory, Travel	Location history, travel and movement data
Music	Baidu music, Baidu FM, TT player	Behavioral & Lifestyle: Music preferences
PC client software	Browser, Input method editor, Toolbar and companion, Baidu Hi,	Not applicable

	Media player, Reader	
Mobile related	Mobile search and browser, Palm, Contacts, Photo Wonder, Wallpaper, Voice assistance, Cloud smart terminal platforms, Mobile phone input method editor, Netdisk, One click root	User behavior on mobile, applications used, usage patterns, search and browing history
Products & services for developers	Developer center, Personal cloud storage, App engine, T5 browsing engine, Mobile test center, LBS open platform, Webmaster platform, Statistics, Share	Not applicable
Others	Data research center, Patent search, Translation, Missing person search, Search for visually impaired, Senior citizen search, Ads manager, Application store, Search and store, Games, BaiduPay, BaiJob, Qunar et al.	Search history, browsing history, payment history, behavior and lifestyle (gaming preferences)

Baidu is especially relevant because of the size and scale that its service commands in the Chinese market. A few statistics from Baidu's first quarter 2016 results that grant perspective are:

- Monthly Active Mobile Search Users on Baidu: 663 M
- Monthly Active Map Users: 321 M
- Activated accounts on BaiduPay: 65 M

These figures are staggering considering Baidu's ~79% share of the Chinese market (China Internet Watch, Aug 2015). It is fair to say that Baidu has an almost unthreatened monopolistic position as far as the Chinese user data is concerned.

6. THE BIG BROTHER PHENOMENON

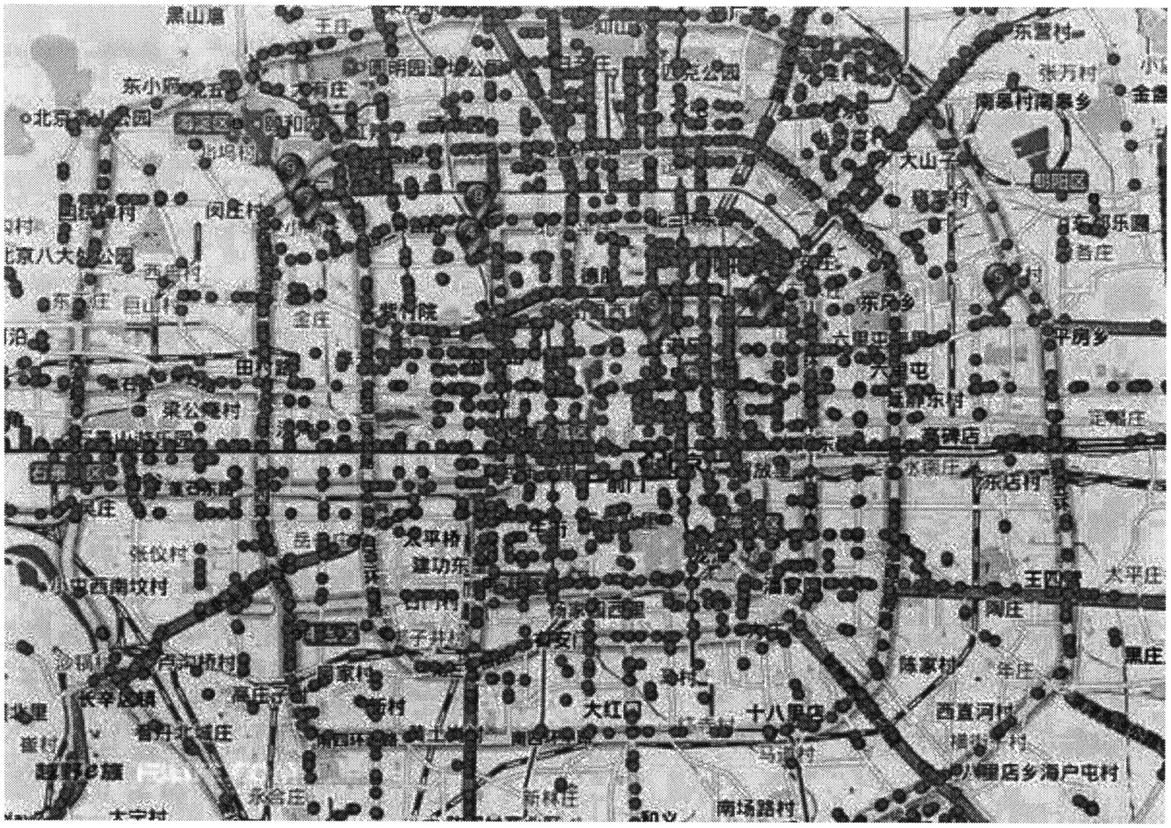
The privacy debate was prevalent in highbrow circles even back in the 1960s when technology was in its rudimentary stages. Messaging was as instant as a telegram could make it and the most sophisticated tech products in a household were a television set or a toaster. The more relevant debate of the time was about business confidentiality and trade secrets. Most workplaces did not allow reproduction of confidential documents. Product designs and drawing were stored in password-protected vaults and guarded by security personnel. Photography was not allowed in manufacturing plants.

Even as late as 2008, companies such as General Motors had not updated their no-camera policy. It was only when an in-built camera became a standard feature on mobile phones and the inconvenience of not carrying mobile phones outweighed the adherence to the legacy no-camera policy, did GM revise its old rule. That said; leaked images of a pre-launch product still remained a top concern for GM globally.

The workplace is undergoing a rapid transformation. Employees were required to check in by swiping their employee cards. Several workplaces have now implemented biometric scanners as time clocks. Communication is based on email and instant messaging. Meetings are held with participants in different parts of the world through conference calling technologies. The speed of communication has drastically reduced with the use of smartphones to answer emails on the fly, respond to texts and take calls any time any day. These technological enablers have improved productivity and increased collaboration. A study conducted by Ferrazi & Greenlight (Ferrazzi K., 2014) suggests that 79% of the employees at any organization work with virtual teams. These teams are geographically spread across various locations and communication is almost entirely driven by widely adopted office communication technologies. The concept of in-person meeting is fast dissolving. The result is that most communication between virtual teams is performed through virtual media. Most lines of communication are recorded and archived. This data is stored as unstructured data and is not extensively analyzed as of today.

The world around us is undergoing a rapid transformation too. Terror threats have caused governments to increase surveillance in cities. Beijing is world's most watched city with an estimated 46000 security cameras and a 100% coverage (Yin C., 2015). China is not the only country with increased video surveillance. Some estimates suggest that UK has about 5.9 M (BBC, 2015). CCTV cameras cover every street and park in London, and major streets and public places outside London.

Figure 22



Source: Weibo

The interconnectedness of our devices has eliminated the anonymity of an individual. A few companies such as Amazon, Uber, Facebook, Google, Netflix, Spotify and major banks carry personal information at a minute granular level. As households get connected with smarter devices and as Internet of Things finds its way into our lives, a few large corporations would have access to more of our personal data.

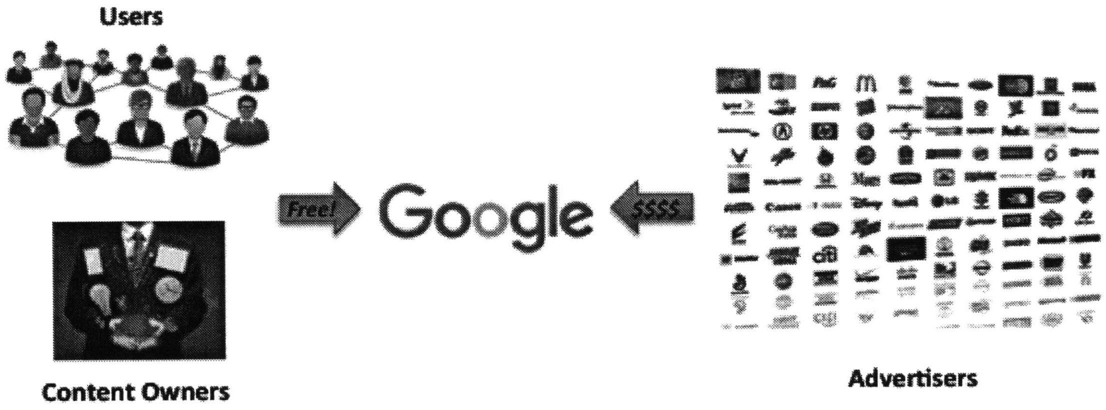
This section explores the extent of data captured, its current use as stated by corporations and the potential implications that micro-tracking could have on an individual.

6.1 Privacy & Ethics

In its early years, Internet businesses experimented with several revenue models. The one model that proved to be most successful is the 'two sided platform'. Some of the well-known examples of the two-sided platform in the pre-internet era are Yellow Pages and American Express credit cards. This model was adopted by search engines such as Google and later by social networks such as Facebook and Twitter. The success of this model depends on the strength of network effects.

Google's and Facebook's users derived maximum utility when the service is offered free-of-charge. With no upper limits on the number of searches one could conduct, users began to use Google's search engine more frequently. Similarly, with no restrictions placed on the number of emails sent or received, Google's email service gained adoption too. Facebook users not only cared about the service being free-of-charge, they also cared about adoption by their friends. Network effect plays a much stronger role here. The same concept applies to most Internet businesses today. The freemium model most recently popularized by LinkedIn, Spotify and a few video game publishers; is an extension of the same concept.

Figure 23



Images source: Business2community, allthestuffyouneedtoknow.com, emerging-advertising-media.wikispaces.com

The success of this revenue model depends on the value achieved by the sponsor-side contributing to the revenue that allows for the platform to remain free to the user-side.

Historically, the most efficient way to deliver value has been through online advertising. The Ad-Tech industry was virtually non-existent in the 1990s. Online advertising was handled in the traditional business model where marketers worked with publishers to purchase advertising space as media served as a proxy for audience (MIT Review, Evolution of Ad Tech, 2013)

As the Internet continued to explode, it became evident that user data could be used more effectively in order to target ads at the right audience. This was a breakthrough for the marketing industry that often relied on hunch and experimentation to reach its target audience. Ad Networks started emerging in early 2000s to facilitate the remnant publisher inventory. As data became the king and value of better targeting enabled prompter and stronger results for marketers, several innovations took place in the Ad-Tech world. Ad Exchanges and Real Time Bidding (RTB) were introduced around 2007. RTB allowed advertisers (marketers) to bid for ad inventory on webpages (publishers) in real-time. This was the first step towards personalized ad targeting. RTB enabled efficiency and effectiveness. Advertisers saw the benefits first hand and ROI was accurately measurable for the first time in advertising history.

Click-through-rate is a standard online advertising performance indicator. It is the ratio of number of clicks to the number of impressions served. It is used as a proxy for effective targeting. The more accurate the targeting of an ad, the higher is the probability of the ad getting clicked by its target audience and therefore, the higher the Click-through-rate. As growth in Ad-Tech has stalled in recent years (Tadena N., 2016) the focus has shifted toward optimizing revenues and improving profits by increasing the click-through-rate. This can be made by possible by improving targeting. Machine learning algorithms are now widely used to build prediction models and one of the key inputs into building a strong model is granular user data.

User data was majorly collected through browser cookies. As users toggle across different devices with mobile being the first device for most of the world, other formats of data collection emerged. Bringing together all the data about a unique user and offering the user a personalized product or service experience became the priority for every business. In the process, more and more data was collected, both online and offline, and brought together. Today user's (in this case, consumer) needs are pre-empted with varying degrees of accuracy. One of the famous examples of accurate customer need identification is Target's 2012 Minneapolis case (Duhigg C., 2012). Target's prediction model accurately predicted the pregnancy of a teenage girl and inundated her with baby product coupons, before she had broken the news to her father. As controversial as this case was, one must acknowledge the efficacy of Target's prediction model. In the world of user data analytics, there seems to be no place left for secrets.

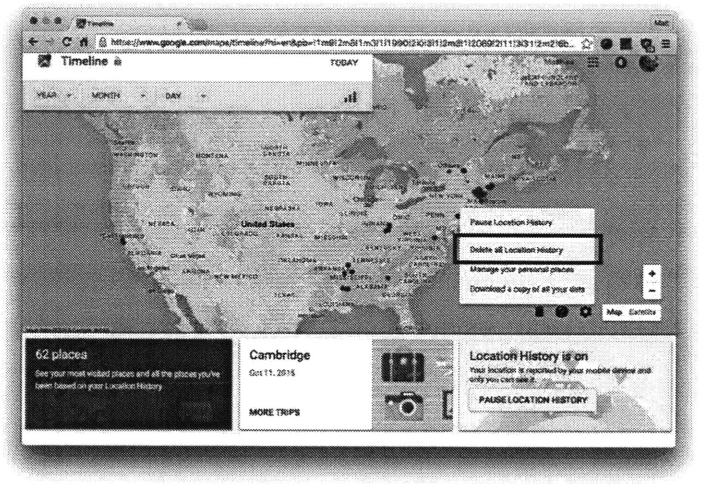
The world of Internet has come a long way from basic Ad-targeting. Today's world is about predictive analytics to preempt customer needs.

A broad classification of individual user data, at the hands of a few corporations:

1. **<u>Geo-location & travel history</u>**

In-built GPS is an extremely important feature in smartphones. It enables the functioning of applications such as maps, weather, local search et cetera. Google tracks location history with its 'Timeline' on Android phones and with any of the Google applications on an IOS device. There are options to turn the tracking off.

Figure 24



See your most visited places and all the places you've been based on your Location History

Your location is reported by your mobile device and only you can see it.

IOS has a default setting that enables tracking of location history through iPhones. There is a way to turn off this feature on iPhone, however, it leads to losing access to some crucial applications such as 'Find my iPhone'.

Additionally, nearly all installed applications seek access to one's location information for the purpose of better service. In the interest of better search results on Yelp, Tinder, AirBnB, Uber and many other location-based services, users tend to allow access to location.

Online ticketing companies such as Kayak.com, Booking.com, carry travel history and these details are passed onto email provider through emailed tickets. Users of Gmail have often experienced the seamless way ticket details merge with Calendar and ads are served for products and services at the upcoming travel destination.

2. Financial transaction history

A few large banks own all the financial information about an individual. This includes details about their income, sources of income, transactions, buying behavior, travel behavior, debts, savings et cetera.

Strong regulations in commercial banking sector prevent mishandling of customer data and does not allow for sharing this data to any outside entity. So far, banks have not invested heavily towards user analytics. Currently, user analytics is performed for the purpose of targeting the commercial banks' financial products such as mutual funds or insurance.

3. Past purchase history

Retailers, both online and offline, have access to their customer's buying history. All retailers perform analytics on customer's buying behavior, with varying degree of sophistication. Online retailers track every move of their customers through data recording tools on their website. Every click and page view is monitored to identify a user's funnel position in order to design a personalized marketing campaign. Purchase history is also used to preempt future purchase, identify look-alike customer segments, cross-sell and up-sell.

Originally, offline retail used loyalty member cards as a means to record purchase history and have unique customer identification. Overtime, newer methods such as credit card as an identifier have come into practice.

4. Personal preferences, political views, likes

Facebook is the single largest repository of user preferences through self-reported information regarding likes and dislikes. Users like pages of their favorite brands, follow artists, join events they intend to attend and discussion groups about topics they are passionate about. This data can be accessed through Facebook's search API. A lot of companies use this information for look-alike audience targeting. This has also proved to be a successful targeting means in political campaigns. Similar information can possibly be found on LinkedIn. The information on LinkedIn is relatively professional in nature and user groups and discussions are serious in nature. Event preferences are also recorded by some ticketing companies such as Eventbrite.

5. Social & professional network

A few years ago, the most potent force in the social network space was Facebook. It still remains in the top spot for social media, however several new players have emerged. LinkedIn's adoption in professional circles has made it the treasure trove of any user's professional network. LinkedIn encourages people to share more and more professional details, thereby enabling better access to newer positions, learning and development opportunities and old fashioned networking opportunities.

Services such as SoundCloud and to some extent Instagram allow for virtual strangers to follow one another based on shared interest.

6. Images, videos, personal content

Google through YouTube, Facebook (including Instagram) and Twitter have a dominant share of the user generated image and video content shared online.

Additionally, Google Photos provides a cloud-based service wherein users can store their personal pictures on cloud instead of a hard-drive. This has proven to be a successful service as most users prefer cloud storage over hard-drive in order to prevent loss of precious memories captured through pictures.

Google and Facebook have never categorically denied analyzing user images for better profiling. It is worthwhile noting that even after a user deletes their content, image, video or textual status, the data continues to reside in Google and Facebook's servers. The deletion of the content is only apparent on the surface layer and does not authorize permanent deletion.

Wordpress, Flicker and Mashable are smaller but significant players in the space of user generated content. The adoption of these services is not as large as that of Google, Facebook and Twitter, but critical for it's target demographic.

7. Personal communication

Email and instant messaging have become the communication norm of the 21st century. The largest market share again belongs to Google and Facebook. Google analyzes emails received by Gmail users to target better ads. Facebook has not identified a way to monetize Whatsapp instant messaging and is reportedly looking for a revenue model that does not depend on Ads. In the recent announcement on Facebook's first ever developer conference, F8 in April 2016, Mark Zuckerberg launched a platform named 'bots on messenger'. Whatsapp also announced end-to-end data encryption on its messenger service in April 2016.

Whether Facebook intends to directly use communication data remains to be seen.

6.2 Vested Interests

As is evident, a few large players in the Internet industry have a disproportionately high amount of access to individual data. Currently, these players are operating in silos. However, Google and Facebook, through their myriad application and acquisitions have access to data across several of the abovementioned categories.

Time and again, the major Internet companies have stated that they do not intent to use customer/ user data for malpractices. They have also stated that the data is stored securely and not shared with any third party without user's permission. Google's privacy policy has undergone 22 revisions over the years. It is interesting to note that 5 of these revisions were carried out in 2015. The latest revision is dated March 25th 2016.

Relative newbies in the Internet business such as Uber (Tassi K., 2015) have stated their intention of protecting user data. However, time and again reports about malpractices make it to the media. The most appalling example of misuse is by Uber's New York executive who reportedly tracked a Buzzfeed News reporter without her permission (Bhuiyan J., & Warzel C. 2014). The case dates back to late 2014 and brought forth Uber's internal tool called 'God View' that allows corporate staff to view any rider and any driver with unbridled accuracy. Any corporation is a sum of its parts and a violation by one miscreant employee stands to be counted as a violation by the corporation.

In the US, all major Internet companies invest heavily towards lobbying against strict privacy laws. In 2013, the same strategy was used in Europe. Unexpectedly, the EU parliament adhered to its original position of ensuring user privacy and implemented strict privacy rules in December 2015. News agencies report that the US government was in favor of the Internet companies and did not want the EU parliament to implement tighter control (Fox-Brewster T., 2015). This subtle underhanded collaboration between the US security agencies and Silicon Valley has been a subject of consternation amongst privacy advocates. Several instances of government intervention into private lives of its citizens have come up in recent years. Edward Snowden, an ex-consultant with US's National Security Agency (NSA), has made some of the most shocking claims regarding breach of private and violation of the fourth amendment of the American constitution.

<u>Some of the common concerns regarding Privacy and Data Security are:</u>

1. **Deliberate misuse by third-party**

- Several instances of data server hacking by self proclaimed watchdog groups have led to privacy loss and public embarrassment. Some recent examples are –
    - Ashley Madison, an online service that arranges extra-marital affairs for its users, was targeted in July 2015. The group calling itself 'The Impact Team' released 32 million user profiles to the public. The group claimed that its rationale behind the hacking was that it wanted to teach a lesson to users of an extra-marital affair service and bring about the end of Ashley Madison. Extortionists used the leaked data to extract ransom from users who wanted to avoid public shaming (Spector D., 2015)
    - iCloud leak of celebrity photos in August 2014: Apple's cloud backing up service iCloud was hacked with a targeted attack on usernames, passwords and security questions. 26 celebrities became victims of this attack (Remling A., 2014)
    - Image recognition for online re-identification: Researches from Carnegie Mellon University exhibited the ease with which facial recognition software can be used for user identification across different and unrelated

websites. The researchers were able to match 10% users of a dating site with corresponding Facebook profiles thereby dispelling the supposed anonymity these users maintained on the dating site (Acquisti A., Gross R., & Stutzman, F., 2014)

2. **"Nanny Statecraft"**

- The use of private and semi-private citizen data for pre-empting criminal behavior and ensuring stricter law adherence
    - Tim O'Reilly, a Silicon Valley publisher, aptly calls preemption as 'algorithmic regulation'. Some examples are – The Italian government uses an income meter to analyze citizen's spending patterns to identify tax frauds (Povoledo E., 2013), use of browsing history and spending patterns to pre-empt homicide (Galbraith R., 2015)
    - Active monitoring of social media for identifying 'possible' disruptors (Collins K., 2015)

3. **Misuse by data owners**

- The Uber case mentioned before is a classic example of how data can be misused by corporate staff even though the company policy states privacy and protection. Similar concerns were highlighted by whistleblower Edward Snowden in one of the disclosures about NSA personnel mishandling private data, especially in the context of sexually explicit images (Schmidt M. S., 2014)

4. **Overly personalized targeting**

- As companies work towards improving their algorithms to target customers better, the fine line between cool and creepy is often missed. With artificial intelligence technologies gaining wider adoption, personalized bots would soon be making product and service recommendations. Complex data models could potentially combine medical records, fitness history, eating habits, sleep pattern and other personal data to determine an individual's chances of using their health insurance. Companies could charge a higher premium to an individual with higher likelihood of needing medical aid. This level of granular targeting maybe optimal but holds against the functioning of a community at large

6.3 IOT & THE FUTURE OF PRIVACY

Internet of Things and Artificial Intelligence are two major technological waves coming towards the human race. As much as we believe in the benefits to be reaped from the implementation of IOT into households, cities and the individual; there remains a strong concern about complete loss of control over one's private information. One of the common use cases of IOT, exhibiting its benefits over current status is --

The pacemaker installed in an individual's heart would perform the additional functionality of recording heart rate and transmit that data real-time to a server that analyzes the user's likelihood to have a heart attack. In case of code red (recorded likelihood of heart attack), the server would send an SOS call to the nearest hospital (pacemaker would have an inbuilt GPS too!). The hospital would dispatch an ambulance in a timely manner. The server would also inform the user's family or friends through another SOS call. This would be the beginning of pre-emptive healthcare. The pacemaker and server company makes its revenue by monetizing the data available through all its users. Primary revenue model could be through the sale of pacemaker device and secondary revenue model could be an ad-supported platform.

The use case above is one of many such excellent possibilities that stand to improve our quality of life. In retail, companies would automatically ship products missing in our refrigerators, saving us the hassle of actual shopping.

The benefits of extensive data collection and analysis are obvious and too many to ignore. However, the downside is equally disturbing. If history has taught us anything, it is fair to say that no technology is foolproof and no amount of security is airtight. Cyber security and data protection are necessary, not only for privacy protection but also as crime deterrents. One of the recent cases about a stranger hacking into a family's baby monitor to talk to their 3-year-old child at night is only one of the many possible criminal scenarios (Owens C., 2016).

Regulatory agencies and crime prevention forces have not kept up with the pace of technology. It is also likely that the technology industry also does not realize the full potential of innovations. As much as I believe in creating a smarter future, the lack of adequate checks and balances remains a matter of deep concern.

7. References

Acquisti, A., Gross, R., & Stutzman, F. (2014). Face Recognition and Privacy in the Age of Augmented Reality. Retrieved April, 2016, from http://repository.cmu.edu/cgi/viewcontent.cgi?article=1122&context=jpc

AppsFlyer Mobile App Tracking | Campaign & Engagement Analytics. (2013, December 21). Retrieved April 28, 2016, from http://www.slideshare.net/AppsFlyer/appsflyer-mobile-ad-tracking-campaign-analytics

Bhuiyan, J., & Warzel, C. (2014, November 18). "God View": Uber Investigates Its Top New York Executive For Privacy Violations. Retrieved April, 2016, from http://www.buzzfeed.com/johanabhuiyan/uber-is-investigating-its-top-new-york-executive-for-privacy?utm_term=.fv1NLgV27#.xnkr5bqLB

Bounie, D., & Gille, L. (2012). International Journal of Communication. Retrieved April 27, 2016, from http://ijoc.org/index.php/ijoc/article/view/1389/744

Brockmann, D., & Helbing, D. (2014, January 6). The Hidden Geometry of Complex, Network-Driven Contagion Phenomenon. Retrieved April, 2016, from http://www.uvm.edu/~cdanfort/csc-reading-group/brockmann-science-2013.pdf

CCTV: Too many cameras useless, warns surveillance watchdog Tony Porter. (2015, January 26). Retrieved April, 2016, from http://www.bbc.com/news/uk-30978995

Charles, D., Gabriel, M., & Searcy, T. (2015, April). Adoption of Electronic Health Record Systems among U.S. NonFederal Acute Care Hospitals: 2008-2014. Retrieved April, 2016

Choudhary, S. R. (2016, February 29). Morgan Stanley slashes valuation of Flipkart. Retrieved April, 2016, from http://www.cnbc.com/2016/02/29/morgan-stanley-fund-marked-down-its-stake-in-flipkart-by-27-percent.html

Collins, K. (2015, June 8). Government pays companies to monitor you on social media (Wired UK). Retrieved April, 2016, from http://www.wired.co.uk/news/archive/2015-06/08/government-pays-companies-to-monitor-social-media-use

Columbus, L. (2014, May 12). Gartner's ERP Market Share Update Shows The Future Of Cloud ERP Is Now. Retrieved April, 2016, from http://www.forbes.com/sites/louiscolumbus/2014/05/12/gartners-erp-market-share-update-shows-the-future-of-cloud-erp-is-now/#6dcebd5c74a1

Duhigg, C. (2012, February 16). How Companies Learn Your Secrets. Retrieved April, 2016, from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1

Evelson, B. (2015, September 25). Agile Business Intelligence Platforms, Q3 2015. Retrieved April, 2016, from http://www.forrester.com/pimages/rws/reprints/document/116447/oid/1-SFDMEH

Fargo, P. (2012, August 27). IOS and Android Adoption Explodes Internationally. Retrieved April 27, 2016, from http://flurrymobile.tumblr.com/post/113379358945/ios-and-android-adoption-explodes-internationally

Ferrazzi, K. (2014, December). Getting Virtual Teams Right. Retrieved April, 2016, from https://hbr.org/2014/12/getting-virtual-teams-right

Fox-Brewster, T. (2015, December 15). Europe Stands Up To Amazon, Facebook Lobbyists -- And Privacy Will Never Be The Same Again. Retrieved April, 2016, from http://www.forbes.com/sites/thomasbrewster/2015/12/15/europe-data-protection-lobbying-amazon-facebook-fight/#4e0224d6ebaa

Galbraith, R. (2015, June 12). US anti-fraud law makes deleting browser history a crime punishable by 20yrs in jail. Retrieved April, 2016, from https://www.rt.com/usa/266389-browsing-history-obstruction-justice/

HBBL patent - 5D method to analyse big data. (2013, November). Retrieved April, 2016, from http://www.ccs.fau.edu/hbbl3/?p=1013

Hardcastle, C. (2016, January 11). Transforming ERP to Postmodern ERP Primer for 2016. Retrieved April, 2016, from https://www.gartner.com/doc/3184718/transforming-erp-postmodern-erp-primer

Knight, H. (2014, January 31). Storage system for 'big data' dramatically speeds access to information. Retrieved April, 2016, from http://news.mit.edu/2014/storage-system-for-big-data-dramatically-speeds-access-to-information-0131

La Monica, P. (2015, July 24). Amazon is now worth WAY more than Walmart. Retrieved April, 2016, from http://money.cnn.com/2015/07/24/investing/amazon-worth-more-than-walmart/

Manyika. (2011, May). Big data: The next frontier for innovation, competition, and productivity. Retrieved April 27, 2016, from http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

McCandless, D. (2013). Ideas, issues, knowledge, data - visualized! Retrieved April, 2016, from http://www.informationisbeautiful.net/visualizations/billion-dollar-o-gram-2013/

Meeker, M. (2015, May 27). 2015 Internet Trends Report. Retrieved April 27, 2016, from http://www.kpcb.com/internet-trends

Microsoft Research Talks About Holograph, An Interactive, 3-D Data-Visualization Research Platform - MSPoweruser. (2014, April 18). Retrieved April 27, 2016, from http://microsoft-news.com/microsoft-research-talks-about-holograph-an-interactive-3-d-data-visualization-research-platform/

Mobile App Tracking - How it Works. (2012, May 1). Retrieved April, 2016, from http://www.slideshare.net/MobileAppTracking/mobile-app-tracking-how-it-works

Murdoch, K. (2006, May 1). Web Analytics: Data Collection Methods. Retrieved April, 2016, from http://www.practicalecommerce.com/articles/196-Web-Analytics-Data-Collection-Methods

Ng, W. L. (2015, September 3). Baidu 101: An Overview of Baidu Webmaster Tools. Retrieved April, 2016, from https://www.semrush.com/blog/baidu-101-an-overview-of-baidu-webmaster-tools/

Number of Facebook users worldwide 2008-2015 | Statistic. (2016, January/February). Retrieved April, 2016, from http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Owens, C. (2016, January 7). Toddler tells parents someone is talking to him at night. Mom makes horrifying find. Retrieved April, 2016, from http://sfglobe.com/2016/01/06/stranger-hacks-familys-baby-monitor-and-talks-to-child-at-night/

Povoledo, E. (2013, January 27). Italians Have a New Tool to Unearth Tax Cheats. Retrieved April, 2016, from http://www.nytimes.com/2013/01/28/world/europe/italys-new-tool-for-tax-cheats-the-redditometro.html?pagewanted=all

Privacy Policy – Privacy & Terms – Google. (n.d.). Retrieved April, 2016, from https://www.google.com/policies/privacy/

Remling, A. (2014, September 21). ICloud Nude Leaks: 26 Celebrities Affected In The Nude Photo Scandal. Retrieved April, 2016, from http://www.ibtimes.com/icloud-nude-leaks-26-celebrities-affected-nude-photo-scandal-1692540

Schmidt, M. S. (2014, July 20). Racy Photos Were Often Shared at N.S.A., Snowden Says. Retrieved April, 2016, from http://www.nytimes.com/2014/07/21/us/politics/edward-snowden-at-nsa-sexually-explicit-photos-often-shared.html

Short, J. E., Bohn, R. E., & Baru, C. (2011, January). How much information? 2010. Retrieved April, 2016.

Spector, D. (2015, September 02). A 'cheating' husband reveals what it feels like to be exposed in the Ashley Madison hack. Retrieved April, 2016, from http://www.businessinsider.com/what-it-feels-like-to-be-exposed-in-ashley-madison-data-breach-2015-9?r=UK

Stec, C. (2015, September 25). 20 Display Advertising Stats That Demonstrate Digital Advertising's Evolution. Retrieved April, 2016, from http://blog.hubspot.com/marketing/horrifying-display-advertising-stats

Steimle, J. (2015, January 26). A Beginner's Guide To Alibaba Group. Retrieved April, 2016, from http://www.forbes.com/sites/joshsteimle/2015/01/26/a-beginners-guide-to-alibaba-group/#65be6c207fd7

Stewart, D. (2013). Big Content: The Unstructured Side of Big Data - Darin Stewart. Retrieved April 27, 2016, from http://blogs.gartner.com/darin-stewart/2013/05/01/big-content-the-unstructured-side-of-big-data/

Structured vs. Unstructured data - dataasaservice -BrightPlanet. (2012, June 28). Retrieved April 27, 2016, from http://www.brightplanet.com/2012/06/structured-vs-unstructured-data/

Tadena, N. (2016, January 4). Ad Tech Growth Hits Speed Bump. Retrieved April, 2016, from http://www.wsj.com/articles/ad-tech-growth-hits-speed-bump-1451936427

Tassi, K. (2015, May 28). An Update on Privacy at Uber. Retrieved April, 2016, from https://newsroom.uber.com/an-update-on-privacy-at-uber/

Tencent (2015, January 27). WeChat’s Impact: Inaugural Report on WeChat Platform Data

The Evolution of Ad Tech. (2013, September 05). Retrieved April, 2016, from https://www.technologyreview.com/s/518551/the-evolution-of-ad-tech/

Total number of Websites. (2016, April 24). Retrieved April, 2016, from http://www.internetlivestats.com/total-number-of-websites/#trend

Wahba, P. (2015, November 05). This Chart Shows Just How Dominant Amazon Is. Retrieved April, 2016, from http://fortune.com/2015/11/06/amazon-retailers-ecommerce/

Yin, C. (2015, October 5). More 'eyes' fight crime in crowds. Retrieved April, 2016, from http://www.chinadaily.com.cn/china/2015-10/05/content_22091634.htm

YouTube: Hours of video uploaded every minute 2015 | Statistic. (2015, July). Retrieved April, 2016, from http://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

APPENDIX

1. Unstructured Data Analytics Companies Profile:

InfiniteAnalytics, Inc.

Company Profile

- Founded in 2012 by MIT Alumni
- Based in Cambridge
- CEO: Akash Bhatia
- CTO: Purushottam Bolta
- Seed round: 1.1 M USD
- Undisclosed amount from Ratan Tata and Sixth Sense Ventures

Products & Technology

- Use machine learning, natural language processing, random forest modeling, image recognition, and semantic technologies for predictive analytics; cloud based
- Personalized Emails, Site Personalization, Predictive Analytics
- ***E-commerce Solutions***
    - ✓ Catalog Marketing Optimization
    - ✓ Structuring and unstructured product catalog
    - ✓ E-commerce Search
    - ✓ Content Recommendation
- ***Revenue Model***
    - ✓ "Partnership" -- Appears to be per transaction based
- ***Solution Deliver***
    - ✓ Plug ins for customization

Target Segments & Key Clients

- **Retail & E-commerce**
- ***Key Clients***
    - ✓ AirBnB, Comcast, B2W Digital, BabyOye, Croma Retail, eBay, Future Group, Infibeam, NBA, NDTV Retail, Trendin, Tata Marketplace

Sources
http://www.infiniteanalytics.com/

https://www.crunchbase.com/organization/infinite-analytics#/entity

ATTENSITY

Company Profile

- Founded in 2000
- 2009: merged with Empolis
- 2010: acquired Biz360
- 2012: divested from Empolis
- Investors: In-Q-Tel, CAPITAL (Swiss investment company)

Product & Technology

- Text data from public & private sources
- Amazing visualization and dashboard
- Patented NLP technology
- Monitors 150 million data sources
    - DiscoverCore, DiscoverNow, Analyze -- Cloud based
    - **Sales Acceleration**: Shorten sales cycles, sell, upsell and cross-sell
    - **Risk Management**: Early warning system alerting with red flags that enable action to limit client's exposure and costs, and to strengthen brand
    - **Product Feedback**: Analyze customer conversations and feedback Competitive Radar: Track competitor
    - **Customer Journey**: User journey
    - **Customer Success**: Analyze conversations to reduce churn rate
    - **Reputation Management**: Discover emerging themes and trends

Solution Delivery

- ✓ SAAS Model

Target Segment & Key Clients

- **Banking, Consumer electronics, eCommerce, Financial services, Hospitality, Insurance, Retail, Telecommunications**
- ***Key Clients***
    - ✓ Yahoo, Hermes, Lufthansa, Nintendo, Esprit, Lenovo

Sources
http://www.attensity.com/
https://www.crunchbase.com/organization/attensity#/entity

Cognitive Scale

Company Profile

- Founded in early 2013, came out of stealth mode in Oct '14
- Founded by ex-IBMers Austin based, with offices in UK and Hyderabad
- Seed fund by 'The Entrepreneur's fund' -- managed by Manoj Saxena

Products & Technology

- Cognitive Cloud Suite
- Cognitive Cloud Fabric
- Guided Service: Virtual service desk agents enable self service/ improve employee productivity
- Guided Commerce: Combines private and public+ structured and unstructured
- Guided Care: Patient's data for proactive identification, personalized intervention
- Guided Procurement: Optimize supply chain decisions
- ***Solution Deliver***
    - ✓ SAAS and Account Management
- ***Revenue Model***
    - ✓ Licensing fee + account management fee + service fee

Target Segment & Key Client

- Beachhead: Heathcare, self-health management tools (eg: Type I Diabetes mgmt)
- ***Key Clients***
    - ✓ Macy's
    - ✓ Deloitte
    - ✓ Two uses cases on company website for Retail and Healthcare

Sources
http://www.cognitivescale.com/
https://www.crunchbase.com/organization/cognitivescale#/entity

CLARABRIDGE

Company Profile

- Founded in 2005
- Founded by Sid Banerjee
- Leadership Team: ex MicroStrategy, brandAnalytics, SAP etc.
- Funding: $103.46M in 4 Rounds from 7 Investors

Products & Technology

- **CX Suite**
- Interpret customer feedbacks from internal & external sources (audio and text). Audio is CRM/ call center data – cloud based
- Net promoter score, Sentiment, Text, Speech Analytics, Social Engagement, Social Listening, VOC, Surveys, NLP, Customer Engagement, Customer Experience Engagement
- ***Revenue Model***
  - ✓ SAAS Subsciption fees + service fees + optional consulting/ training services
- ***Service Delivery***
  - ✓ Claim omni-channel support

Target Segment & Key Clients

- **Retail**, **Financial/ Insurance**, **Telecom**, **CPG**, **Consumer Tech**, **Travel/ Hospitality**
- ***Key Clients***
  - ✓ Amazon, eBay, United Airways, Walmart, Orbitz, Exxon Mobile, Whirlpool, Dell, GE Healthcare, GAP, Vodafone

Sources
http://www.clarabridge.com/
https://www.crunchbase.com/organization/clarabridge#/entity

sysomos

Company Profile

- Founded in 2007
- Founders -- Nilesh Bansal, Nick Koudas
- Acquired by Marketwired in 2010, split from Marketwired in 2015
- Current CEO -- Adnan Ahmed (ex Marketwired)
- Present in 10 cities, expanding to Shanghai and Singapore in 2016

Products and Technology

- Crawlers collect more than 8 million new posts every hour, from sources including blogs, forums, news sites, Twitter, YouTube, Facebook, Flickr, LinkedIn and numerous other social network services -- Cloud based
- ***E-commerce Solutions***
    - ✓ Map: Social Research Engine
    - ✓ Heartbeat: Social Listening
    - ✓ Expion: Content discovery, planning, publishing, moderation, and analytics
    - ✓ Reports: Standard dashboard
    - ✓ Gaze: Image monitoring
    - ✓ Scout: Social media analytics
- ***Revenue Model***
    - ✓ Per API and with a monthly/ annual subscription of data type and quantity
- ***Solution Deliver***
    - ✓ Plus Ins

Target Segments & Key Clients

- Typically brand managers across sectors, current focus on CPG
- ***Key Clients***
    - ✓ Nestle, Boeing, Coca-Cola, Marriott, Mondelez, Georgia State University, Midwestern University

Sources
https://sysomos.com/
https://www.crunchbase.com/organization/sysomos#/entity

bitext

when big data means big text

Company Profile

- Founded in 2007
- CEO: Antonio S. Valderrabanos
- Seed round: $900k on 2015 Jan from Inveready Technology Investment Group

Products & Technology

Specialize in linguistic technology. pick up deeper linguistic meaning, so it produces superior analytics results – with over 90% accuracy and precise topic detection

- **Bitext**

✓ Sentiment analysis

✓ Categorization,

✓ Entity & concept extraction

✓ Lead generation

Customized engine and additional linguistic consulting services, public API

Target Segment & Key Clients

- ***Key Clients***
  - ✓ TNS
  - ✓ Opentext
  - ✓ EY
  - ✓ Salesforce Marketing Cloud
  - ✓ Intel
  - ✓ Movistar

Source
https://www.bitext.com/
https://www.crunchbase.com/organization/bitext#/entity

UBERVU
via Hootsuite

Company Profile

- Founded in 2008
- CEO: Ryan Holmes,
- Founders: Dario Meli, Ryan Holmes, David Tedman
- Funding: In total $250million by Insight Venture Partners, Accel Partners and OMERS Ventures, and other unrevealed ones

Products & Technology

- Automatically analyzes all the brand's social media data to show insights like influencers, stories, and trends that can be leveraged.
- Client can then take that information and post updates to Facebook, Twitter, LinkedIn and Google+ business pages through the UBERvu platform.
    - Publish content to client social networks
    - Engage followers
    - Analytics reports on social media impact

Target Segments & Key Clients

- Individuals, businesses (social media marketer), consultants, agencies, large organizations and governments
- ***Key Clients***
    - ✓ Local World, Levis, ebay, Oakley, Orange, Siemens
    - ✓ Over 10 million users and the world's top brands

Source
https://hootsuite.com/products/insights
https://www.crunchbase.com/organization/ubervu#/entity

mention

Company Profile

- Founded in 2012
- Founder: Thibaud Elziere
- Funding:
    - ✓ $800k in 1 Round from 4 Investors
    - ✓ $800k Seed on March 14, 2013

Products & Technology

Media Monitoring for brands to obtain filtered and organized information from the web and social networks.

- Sentiment analysis
- Identify influencers
- Generate brand reports
- Monitor brand's online footprint
- Find buzz about clients' business
- Generate awareness for clients' companies

Target Segment & Key Clients

- More than 2,500 paying subscribers:
    - ✓ CrunchBase
    - ✓ GitHub
    - ✓ Microsoft

Sources
https://mention.com/en/
https://www.crunchbase.com/organization/mention

2. Amazon, Sears and Flipkart Interview Excerpts

Topic	Amazon	Sears	Flipkart
Use of External Data	• Yes, both internal and external data are used together • Most firms hire consultants but quality of service is low • Internal analytics teams act as bar raisers • Have not considered setting up data acquisition in-house because it is Cap-ex heavy and ROI is not proven	• Demographics, user characteristics data procured from 'Acxiom' (vendor) • Used extensively for online and offline • Online – use for targeted ads, product recommendations and homophily targeting • Do not use any end-to-end solution as tools are in-house • Extensive analytics performed on 'Shop your way'	• Do not use external data sources. It is part of product strategy, however currently lack tools and sources that would exhibit ROI • Better targeting and customer need identification is a challenge. Most feedback is on poor ad & recommendation relevance
Unstructured Data Analytics	• Most firms are prioritizing utilization of internal unstructured data, however, degree of maturity varies • See potential of external unstructured data	• Text analytics performed on customer reviews and feedback • Eg: Customers highlighted bugs on website before Sears spotted it • Extensive focus on Omni-channel so see potential of unstructured	• Do not perform unstructured data analytics at the moment • Exploring a few tools to introduce internal unstructured data analytics • Especially relevant for merchant services – due for launch soon
Social Media Analytics	• Has not proven to be a big revenue driver so limited analytics resources spent on social media • Largely managed by marketing teams with	• Only related to CRM – why customers are unhappy and how to bring them back • Not handled by analytics and hasn't been a revenue driver	• Perform social media analytics and retargeting on Facebook • Not very advanced and use Facebook's business solutions

	consultants		
Other Information	• Dashboarding and data reported should be relevant	• Org structure is extremely ‘messy’ and leads to inefficiencies	• Expanding rapidly and hiring to build strong analytics practice