Big Data: Evolution, Components, Challenges and Opportunities

by

Alejandro Zarate Santovena

B. S. Chemical Engineering
Universidad Iberoamericana, 1995

Master of Business Administration
Carnegie Mellon University, 2002

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

JUNE 2013

Signature of Author: ______________________________
MIT Sloan School of Management
May 10, 2013

Certified by: ______________________________
Michael A. Cusumano
SMR Distinguished Professor of Management
Thesis Supervisor

Accepted by: ______________________________
Stephen Sacca
Director, MIT Sloan Fellows Program in Innovation and Global Leadership
MIT Sloan School of Management

Big Data: Evolution, Components, Challenges and Opportunities

by

Alejandro Zarate Santovena

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT
ON MAY 10, 2013 IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN MANAGEMENT OF TECHNOLOGY

ABSTRACT

This work reviews the evolution and current state of the "Big Data" industry, and to understand the key components, challenges and opportunities of Big Data and analytics face in today business environment, this is analyzed in seven dimensions:

Historical Background. The historical evolution and milestones in data management that eventually led to what we know today as Big Data.

What is Big Data? Reviews the key concepts around big data, including Volume, Variety, and Velocity, and the key components of successful Big Data initiatives.

Data Collection. The most important issue to consider before any big data initiative is to identify the "Business Case" or "Question" we want to answer, no "big data" initiative should be launched without clearly identify the business problem we want to tackle. Data collection strategy has to be closely defined taking in consideration the business case in question.

Data Analysis. This section explores the techniques available to create value by aggregate, manipulate, analyze and visualize big data. Including predictive modeling, data mining, and statistical inference models.

Data Visualization. Visualization of data is one of the most powerful and appealing techniques for data exploration. This section explores the main techniques for data visualization so that the characteristics of the data and the relationships among data items can be reported and analyzed.

Impact. This section explores the potential impact and implications of big data in value creation in five domains: Insurance, Healthcare, Politics, Education and Marketing.

Human Capital. This chapter explores the way big data will influence business processes and human capital, explore the role of the "Data Scientist" and analyze a potential shortage of data experts in coming years.

Infrastructure and Solutions. This chapter explores the current professional services and infrastructure offering and how this industry and makes a review of vendors available in different specialties around big data.

Thesis Supervisor: Michael A. Cusumano
Title: SMR Distinguished Professor of Management

To Emilie, Santiago, Daniela

Table of Contents

1. Historical Background

Big data has been around us since much longer than we think, and in order to better understanding the potential of this great new jump in technology and analytics it is essential to understand the challenging problems that motivated the now so called "data scientists" to conceive and develop solutions and technologies that now drive this revolution.

I still remember when I was 14 year old; in my junior high school it was the first day of classes and Mrs. Elizabeth Laing, my new history professor, started to explain the class that year program content, I do not remember much from that course except from one thing; almost every single class Mrs. Laing reinforced the point that in order to understand the present and its implications we need to understand the past, this is a learning that has been with me since then.

In this chapter, I would like to make a review of the key milestones that have led to what we know to the concept and technology around big data. This review does not aims to be exhaustive but reflect key milestones in recent history that on hand he way society produces data, and on the other the way technology has allowed to deal with this.

As we will see, the census, a major change in a public policy, a war, and even a librarian concerns have triggered ideas and the development of technologies that have allowed human race to be able to process data in fast and efficient ways, resulting in a new discipline now known as "data science". In the early days of this discipline; and until very recently, the most important challenge seemed to lay on the available storage solutions, now this concerns seem to be trivial, as solutions and technology allow the storage of massive amounts of data.

Today is not anymore a problem of storage space, but a problem of developing and implementing solutions that will allow collecting, process and obtaining insights from data it in the most efficient and fast way.

The 1890 census – the first technology enabled "snapshot" data collection

Probably one of the first big data problems in North America is the 1890 census [1], this census was the first to be compiled using methods and technology invented by Herman Hollerith. In this technology, data was entered on a machine-readable medium, punched cards, and tabulated by machine. [2]

In 1881 Herman Hollerith began designing a machine to compile census data more efficiently than traditional hand methods, and by the late 1880s, he had built a punched card tabulating machine that could be read by electrical sensing[3]. The way this technology worked was simple, when a stylus was inserted into a hole on a template; a corresponding hole was punched in a card at the other end. Each card represented one person and each hole a different statistic, such. The cards were sorted and later read electronically by a press containing pins that penetrated the card only at its holes. Each pin that passed through a hole made electrical contact with a small cup of mercury, closing a circuit and advancing a dial counter by one [3].

This technology demonstrated to considerably reduce the time required to tabulate the census from an estimate of more that eight years (as it was the case for the 1880 census) to one year for the 1890 census[2]. A total population of 62,947,714 was announced after six weeks of processing.

Hollerith's machines acquired considerable exposure and reputation, leading to the establishment of his own company, The Tabulating Machine Company, being probably the first big data start up company in the history. In 1911, Hollerith merged his company with two other companies to create The Computing Tabulating Recording Company.

In 1914, Thomas J. Watson was hired to merge the new company; his remarkable technical and business skills resulted in the leading computing company of its time; that in 1924 changed its name to International Business Machines Corporation (IBM).[3] For decades, Hollerith's card system was used in a variety of applications and industries, most notably by IBM to program early computers.[3]

The 1935 Social Security Act

In 1935 President Franklin D. Roosevelt's Social Security Act launches the U.S. government in one of the most ambitious data-gathering efforts ever. IBM was the winning contractor responsible to gather and keep employment records on 26 million working Americans and 3 million employers.

The Social Security program inspired its basic principles on those of "social insurance" an intellectual practice born in Europe in the 19th century that represented the European social welfare tradition[5]. Social Security it was first implemented in 1889 in Germany by Chancellor, Otto von Bismarck[5].

In 1935, when the United States decided to implement a social security system there where at least 34 nations already operating some form of social insurance program (about 20 of these were contributory programs like Social Security).[5]

On December 27, 1936 Luther A. Houston [6] a reporter from the New York Times published that at the time it was estimated that at least 26 million people were entitled to social security in the United States and the roles that different government agencies would have in this process.

According to Houston chronicle [6],the distribution and collection of the applications for the social security was entrusted to the Post Office Department. For this purpose the Post Office established 1,072 typing centers throughout the country. Here, all applications were copied in triplicate, as were the account number cards, which were issued to the workers and distributed to government agencies as follows: [6]

1) One copy was sent to the Internal Revenue Bureau (which would collect the tax and disburse payments);
2) One copy was sent to the Wage Record Office;
3) One copy was retained for the post office records.

Houston's chronicle continues explaining that applications were filled in numerical order and then punch cards prepared containing the name and other important information of each worker. From the punch cards an index of employees eligible to the benefits was prepared and a ledger account for each eligible employee was opened. Upon the punch card, a reports were received, was the recorded wage of every worker.

The operating center of this machine was the Wage Record Office, initially located in Baltimore because there was no sufficient office space available in Washington to house the "gigantic" files and the expected 5000 employees to keep these records. Another 2,500 employees would be distributed in 108 field offices all over the country. Today the Social Security Administration employs more than 65,000 people in 1900 field offices all over the country. [7]

The 1943 "Colossus"

At Bletchley Park, a British facility dedicated to breaking Nazi codes during World War II, engineers developed a series of groundbreaking mass data-processing machines, culminating in the first programmable electronic computer: The Colossus. [4]

The Colossus computers were developed and used by the British during World War II for the cryptanalysis of the Lorenz cipher. These allowed the British army (and the allies) to decode the vast quantity of encrypted high-level telegraphic messages between the German High Command and their army commands throughout occupied Europe.

Engineer Thomas "Tommy" Harold Flowers, designed Colossus at the Government Code and Cypher School at Bletchley Park. The first prototype, the Colossus Mark 1 was operational by February of 1944, and an a second version, the improved Colossus Mark 2 was operational by June of 1944. A total of ten Colossus computers were in use by the end of the war. [8]

Colossus used vacuum tubes to perform Boolean operations and calculations and was the first of the electronic digital machine with programmability: [8]

- It had no internally stored programs. To set it up for a new task, the operator had to set up plugs and switches to alter the wiring;
- Colossus was not a general-purpose machine, being designed for a specific cryptanalytic task involving counting and Boolean operations.

The technology of Colossus, had a significant influence on the development of early computers as helped to confirm the feasibility of reliable high-speed electronic digital computing devices. [8]

A librarian storage and organization problem

Arthur Fremont Rider (May 25, 1885 – October 26, 1962) was an American writer, poet, editor, inventor, genealogist, and librarian. Throughout his life he wrote in several genres including plays, poetry, short stories, non-fiction and an autobiography. In 1933 he became a librarian at Wesleyan University in Middletown, Connecticut. [9]

In 1944, Fremont Rider published *The Scholar and the Future of the Research Library*. In this work, estimates that American university libraries were doubling in size every sixteen years. Given this growth rate and assuming that data storage would continue to be made in books, he speculated that by 2040 the library at Yale University would have approximately 200,000,000 volumes, would occupy over 6,000 miles of and would require a cataloging staff of over six thousand people; [4] in order to put these numbers in context, as of today the Library of Congress has a printed collection of 23 million volumes (that could be easily stored in 10 terabytes of memory) and employs about 3,600 people.

In order to solve this problem -and inspired by developments in the production and printing of micro-text- Fremont Rider e proposed his own invention a 7.5 by 12.5-centimeter opaque card (microform) which in the front side would have the catalogue information, and on the reverse as many as 250 pages of an ordinary book. These microforms would have a dual purpose, as the catalogue and as the collection, with the benefits of saving shelving space and physically integrating the manuscripts and the catalogue collection. [4]

Fremont Rider envisioned that Researchers would search the catalogue for the entry they wished, and then having selected it would take the card to a reader machine no bigger than a briefcase. [9]

Rider's prediction that microforms would be employed to solve the issue of data storage (the collection) and growth was prescient though at the time he could not have known that microform would in turn be superseded by the digital revolution. [9] His idea that a catalogue and collection could be integrated in a single "record" or "form" presaged the possibilities opened by digital media storage and organization capabilities. [9]

The 1961 U.S. National Security Agency backlog

As of the 1950's, it seems that most of the challenges and concerns with data seemed to be related to volume, however in the 1960's the developments in automated data collection unveiled a second key component in big data: Velocity.

In 1961, the U.S. National Security Agency (NSA) faced a massive information overload due to automatized intelligence collection and processing during the espionage-saturated cold war. Despite having more than 12,000 cryptologists the agency continuously struggled to digitize a backlog of records stored on analog magnetic tapes, just in July 1961, the agency received about 17,000 reels of tape. [4]

1961 and the growth of scientific knowledge

In 1961 Derek Price publishes *Science Since Babylon*, in which he charts the growth of scientific knowledge by looking at the growth in the number of scientific journals and papers. He concludes that the number of new journals has grown exponentially rather than linearly, doubling every fifteen years and increasing by a factor of ten during every half-century. Price calls this the "law of exponential increase," explaining that *"each scientific advance generates a new series of advances at a reasonably constant birth rate, so that the number of births is strictly proportional to the size of the population of discoveries at any given time."* [10]

This work puts light again on the fact that data is not only about Volume but also about Velocity.

The 1964 recipe on how to cope with the information explosion

April 1964 Harry J. Gray and Henry Ruston publishes *"Techniques for Coping with the Information Explosion,"* in which they offer the following advice in order to cope with the increasing volume of information produced [10]:

1) No one should publish any new papers. [10]
2) If 1) is not feasible, only short papers should be published. "Short" means not more than 2500 characters counting "space," punctuation marks, etc. as characters. [10]

3) If 2) is adopted the following restriction should apply: "Only those papers should be published which delete one or more existing papers which combined length is 2,501 characters or more." [10]

According to Gray and Ruston an important byproduct of implementing these practices would be the reduction of the burden on personnel selection committees. This will happen because the person's list of publications will be replaced by a single negative number denoting the net number of papers he has deleted from the present information store. [10]

1965: First Data Center is conceived

In 1965, faced with the growing problem of where to keep more than 742 million tax returns and 175 million sets of fingerprints, the US federal government considered a plan to consolidate its data centers into a single mega-center.[11]

That year, a report was submitted to the Office of Statistical Standards of the Bureau of the Budget entitled *"A Review of Proposals for a National Data Center."* That report analyzed the challenges that prevented the most effective use of the resources of the Federal Statistical System in the establishment of public policy, the management of public affairs, and the conduct of research. It recommended changes in the mission of the Federal Statistical System that could transform it into a more effective source of information services for today's needs.[12]

The report suggested that the cross-agency sharing of information could a more efficient government operation — for example, states could share data on suspended drivers licenses.

During the time that this report was under review by the Administration it became "caught up" in a substantial public controversy over raising concerns on potential threat to personal privacy embodied in its recommendations, and the government using computers to keep records on all citizens, from cradle to grave. [11]

The report and the Administration's intentions was object of hearings in the Senate and the House. Through extensive comment in the public press, the report acquired the image of a design to establish a "gargantuan" centralized national data center calculated to bring Orson Well's "1984" at least as close as 1970. [12]

While the plan was dropped, it would later be remembered as one that heralded the dawn of the electronic data storage. [11]

1967: Data compression

Maybe to that point it was clear that alternatives had to be found in order to cope with the increasing volume of data available, and instead of limiting the amount of data that could be stored (due to limited storage capabilities of the time) new technologies had to be developed in order to make better use of the space available.

In November 1967, B. A. Marron and P. A. D. de Maine published *"Automatic data compression"* in the Communications of the ACM, stating that:

"The "information explosion" noted in recent years makes it essential that storage requirements for all information be kept to a minimum. A fully automatic and rapid three-part compressor which can be used with "any" body of information to greatly reduce slow

external storage requirements and to increase the rate of information transmission through a computer"

"The system will also automatically decode the compressed information on an item-by- item basis when it is required."

1974: Data Science is born [10]

In 1974 Peter Naur (a Danish pioneer in computer science and Turing award winner) publishes "*Concise Survey of Computer Methods in Sweden and the United States.*" The book is a survey of contemporary data processing methods that are used in a wide range of applications. It is organized around the concept of data as defined in The International Federation for Information Processing (IFIP) Guide to Concepts and Terms in Data Processing, which defines data as "*a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process.*" [10]

The Preface to the book tells the reader that a course plan was presented at the IFIP Congress in 1968, titled "*Datalogy, the science of data and of data processes and its place in education,*" and that in the text of the book, "*the term 'data science' has been used freely.*"

Naur offers the following definition of data science: "*The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.*" [10]

1977: The International Association for Statistical Computing [18]

In 1977 The International Association for Statistical Computing (IASC) is founded with the mission of linking traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.

1989 The Internet is born [19]

In 1989, British computer scientist Tim Berners-Lee invents the World Wide Weba as a solution to facilitate the sharing of information via a 'hypertext.'

1996 The Data Science evolution [10]

In 1996, during the International Federation of Classification Societies (IFCS) conference in Tokyo the term "data science" is included in the title of the conference: "*Data science, classification, and related methods*", this one of the first times the term is used, and with this recognized the future importance the data science will coming years.

The IFCS was founded in 1985 by six country and language-specific classification societies, one of which, The Classification Society, was founded in 1964. The aim of these classification societies is to support the study of "the principle and practice of classification in a wide range of disciplines"(CS), "research in problems of classification, data analysis, and systems for ordering knowledge"(IFCS), and the "study of classification and clustering (including systematic methods of creating classifications from data) and related statistical and data analytic methods" (CSNA bylaws). The classification societies have variously used the terms data analysis, data mining, and data science in their publications. [10]

1997 the year for Data Mining and Big Data

In March 1997, the journal: *Knowledge Discovery and Data Mining* is launched. This publication focuses on the theory, techniques and practice for extracting information from large databases

Publishes original technical papers in both the research and practice of data mining and knowledge discovery, surveys and tutorials of important areas and techniques, and detailed descriptions of significant applications.

In the first editorial of this journal, Usama Fayyad, Senior Researcher at Microsoft Research pointed out the following questions, many of them start to get answers with today's technology:

"New approaches, techniques, and solutions have to be developed to enable analysis of large databases. Faced with massive data sets, traditional approaches in statistics and pattern recognition collapse. For example, a statistical analysis package (e.g. K-means clustering in your favorite Fortran library) assumes data can be "loaded" into memory and then manipulated. What happens when the data set will not fit in main memory? What happens if the database is on a remote server and will never permit a naıve scan of the data? How do I sample effectively if I am not permitted to query for a stratified sample because the relevant fields are not indexed? What if the data set is in a multitude of tables (relations) and can only be accessed via some hierarchically structured set of fields? What if the relations are sparse (not all fields are defined or even applicable to any fixed subset of the data)? How do I fit a statistical model with a large number of variables?"[13]

In this same year, in their paper titled Application-controlled demand paging for out-of-core visualization, Michael Cox and David Ellsworth are among those who acknowledge the problems that big data will present as information overload continues on its relentless path:

"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources."

Cox and Ellesworth's use of the term "big data" is generally accepted as the first time anyone has done so, although the honor of actually defining the term must go to one Doug Laney, who in 2001 described it as being a *"3-dimensional data challenge of increasing data volume, velocity and variety"*, a definition that has since become almost pervasive among industry experts today.[1]

2004: Let's play with Hadoop

Hadoop was originally developed to support distribution for the Nutch search engine project by Michael J. Cafarella and Doug Cutting, who it after his son's toy elephant. [14]

Hadoop is one of the driving forces behind the big data revolution, enabling the distributed processing of large data sets across clusters of commodity servers and it is designed to scale up from a single server to thousands of machines with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures. [15]

Hadoop has two main subprojects: [15]

- MapReduce - The framework that understands and assigns work to the nodes in a cluster.
- HDFS - A file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system. HDFS assumes nodes will fail, so replicates data across multiple nodes in order to achieve reliability.

Hadoop has changed the economics and the dynamics of large scale computing as allows to build high computing capabilities a very low cost, according to discussions with experts in this field a Hadoop cluster can cost as low as $1 million and provided cloud services to several users.

Hadoop has four important characteristics: [15]

- *Scalable.* New nodes can be added as needed without needing to change data formats, how data is loaded, how jobs are written, or the applications on top. [15]
- *Cost effective.* Hadoop brings massively parallel computing to commodity servers. The result is a significant saving in the cost per petabyte of storage, which in turn makes it affordable to model complete datasets. [15]
- *Flexible.* Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide. [15]

- *Fault tolerant.* In case the cluster loses a node, the system is capable to redirect work to another location of the data and continues processing. [15]

The 2011 Jeopardy champion: Watson

But perhaps one of the most important milestones in the evolution of technology and big data is Watson, an artificial intelligence computer system capable of answering questions posed in natural language.

Watson was developed as part of the IBM's DeepQA project by a research team led by David Ferrucci. Watson was named after IBM's first president, Thomas J. Watson. The machine was specifically developed to answer questions on the quiz show Jeopardy! In 2011, Watson competed on Jeopardy against former winners Brad Rutter, and Ken Jennings, winning the first prize of $1 million. [16]

According to IBM, Jeopardy! was selected as the ultimate test of the machine's capabilities because it relies on many human cognitive abilities traditionally seen beyond the capability of computers, these include: [17]

- The ability to discern double meanings of words, jokes, verses, poems, and inferred hints.
- It provides extremely rapid responses.
- Is capable to process vast amounts of information to make complex and subtle logical connections.

In a human, these capabilities come from a lifetime of human interaction and decision-making along with an immersion in culture. [17]

For the Watson team, replicating these capabilities was an enormous challenge, moving beyond keyword searches and queries of structured data to asking questions and accessing and assessing a vast amount of unstructured data to find the best answer. [17]

To meet this grand challenge, the Watson team focused on three key capabilities: [17]

- Natural language processing
- Hypothesis generation
- Evidence-based learning

Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia, but was not connected to the Internet during the game. For each clue, Watson's three most probable responses were displayed on the television screen. Consistently outperforming its human opponents. However, it had trouble responding to a few categories, notably those containing clues with only a few words. [16]

Watson's software was written in various languages, including Java, C++, and Prolog and uses Apache Hadoop framework for distributed computing, Apache UIMA (Unstructured Information Management Architecture) framework, IBM's DeepQA software and SUSE Linux Enterprise Server 11 operating system.

According to IBM, "more than 100 different techniques are used to analyze natural language, identify sources, generate hypotheses, find and score evidence, and merge and rank hypotheses." [16]

In February 2013, IBM announced that Watson software system's first commercial application would be for utilization management decisions in lung cancer treatment at Memorial Sloan-Kettering Cancer Center in conjunction with health insurance company WellPoint.[16]

2. What is Big Data?

Big data has become of the most widely used terms in information technology worldwide but the implications and major benefits for companies investing in "big data" are yet to be realized in the coming years. According to IBM[20] every day we create 2.5 quintillion (10^{18}) bytes of data and as much as 90 percent of the world's data today has been created in the last two years.

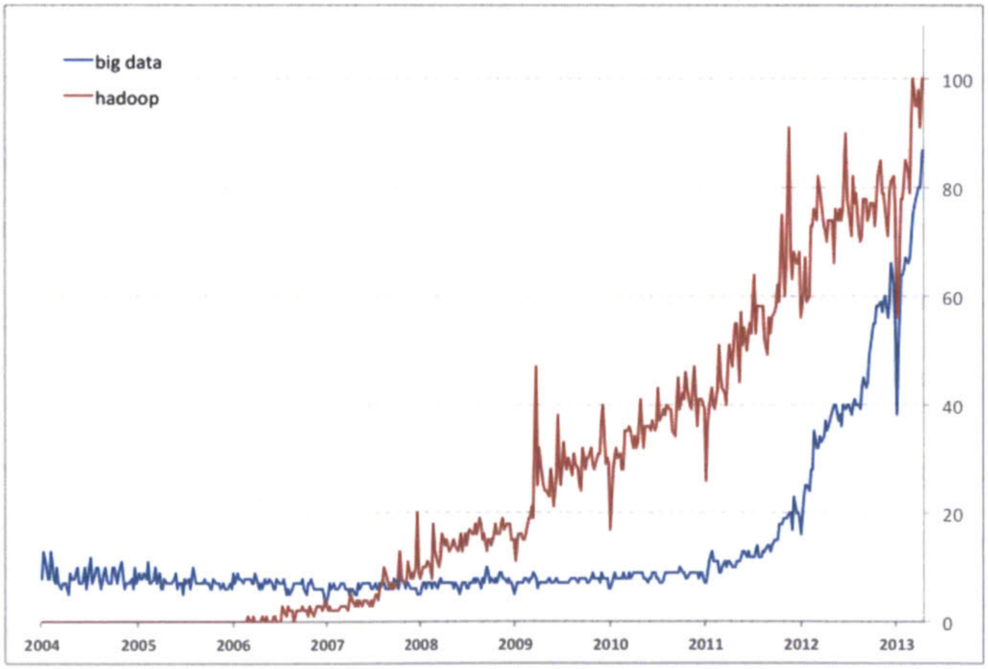
The global data explosion is highly driven by technologies including digital video and music, smartphones, and the Internet[27]. This data has its' origins in a variety of sources including web searches, sensors, commercial transactions, social media interactions, audio and video uploads, and mobile phone GPS signals. [20]

It is estimated[23] that every minute email users send more than 204 million messages, Google receives over 2,000,000 search queries, Facebook users share more than 684,000 pieces of content, consumers spend more than US $270,000 on web shopping, Twitter users send over 100,000 tweets, more than 570 new websites are created, and Instagram users share 3,600 new photos. All these sources of information will contribute to reach 35 Zettabytes of data stored by 2020[24].

These increasingly diverse data sets complement each other and allow businesses to fill in missing gaps and unveil new insights. Filling these gaps improves operations, decision-making and gives elements to improve business processes. [21]

The amount and variety of sources of information have created significant storage, processing and analysis challenges for organizations worldwide. However, the tremendous size and complexity of "big data" are only one side of the issue. The other aspect of the big data issue is the demand for cost effective forms of capture, storage, analytics and visualization. Ultimately, data management through the life cycle, information has different quality, security and access requirements at different points of its lifecycle and these differences form much of the complexity aspects of big data[21].

In order to get an idea on the relevance that big data has taken over the last three years, let's take a look to the frequency of google searches of the term "big data," as can be seen in the figure below, the term has considerably "taken off" since 2011 reaching almost similar search frequency as Hadoop, a technology launched in 2004 and that is closely tied to any big data initiative; however, as we can see the term Hadoop started to appear in google web searches since mid 2006.



Google search frequency for Hadoop and Big Data

Source: Google Trends

The 3V's of Big Data

As discussed in the "Historical Background" chapter; historically, data management has evolved mainly around two key problems: volume and processing capacity. However, challenges have shifted, it is no more an issue of storage or even processing capacity but how data has become more complex with a variety of sources not seen before, and how is data now collected at record speeds; this creates a tree dimension challenge which according to the research firm Gartner[21] can be described as follows: *"Big data is high-volume, -velocity and –variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."*

High-volume[21] -From Terabytes to Zettabytes-

As Steve Lohr from the New York Times pointed out in his article "The Age of Big Data," advancing trends in technology are opening the door to a new approach to understanding the world and making decisions. Today there is a lot more data than we can process, and keeps growing at 50% a year, or more than doubling every two years.

According to Gardner[21], high data volume can be defined as follows: *"Big data is high volume when the processing capacity of the native data-capture technology and processes is insufficient for delivering business value to subsequent use cases. High volume also occurs when the existing technology was specifically engineered for addressing such volumes – a successful big data solution".*

Velocity[21] -From Batch to Streaming Data-

Data has evolved into a continuous flow. The velocity by which organizations are collecting data is constantly growing with the arrival of streaming technologies and the constant increase of sources of information. What in the past use to be snapshots of a business analytics have evolved in real time dashboards.

However, it is not just a bigger stream of data, but entirely new ones. For example, there are now countless digital sensors in telephones, automobiles, utilities meters, and industrial equipment, which measure and communicate location, movement, vibration, temperature, and humidity, among many other variables.

According to Gardner[21], high velocity can be defined as follows: *"High velocity refers to a higher rate of data arrival and/or consumption, but is focused on the variable velocity within a dataset or the shifting rates of data arrival between two or more data sets."*

Variety[21] - From Structured to Unstructured -

However, the biggest challenge is not about the size or speed data be collected, but how this information can be used to address specific business goals and how organizations will eventually adapt their business processes to better profit from this opportunity.

The biggest benefit of Big Data for organizations will be the ability not only to ask but also to answer highly complex questions[6], collect relevant data, and to use the right technologies to translate and process these information assets into analytics that lead to bold insights, enabling real time decision-making.

According to Gardner[21], high data variety can be defined as follows: *"Highly variant information assets include a mix of multiple forms, types and structures without regards to that structure at the time of write or read. And previously under-utilized data, which is enabled for use cases under new innovative forms of processing, also represents variability."*

Big data creates value in several ways[30]

According t McKinsey Global Institute (MGI) there are five applicable ways to leverage big data that offer transformational potential to create value and have implications on how organizations will have to be designed, organized and managed[30].

- *Creating transparency[30]*

By making big data more easily accessible to relevant stakeholders in a timely manner.

- *Enabling experimentation, expose variability, and improve performance[30]*

As organizations collect more detailed data (in real or near real time) on several dimensions (e.g. product inventories, personnel sick days), IT will allow instrumenting processes and setting up controlled experiments.

Using data will also enable to analyze variability in performance, (thru natural or controlled experiments) and to understand its root causes, in order to enable leaders to manage performance to the higher levels.

- *Segmenting populations to customize actions[30]*

Big data will allow more organizations to create highly specific segmentations and to tailor products and services precisely to meet specific needs. Despite, this approach is well known in marketing and risk management, even consumer goods and service companies that have used segmentation for many years are beginning to deploy more sophisticated big data techniques such as the real time segmentation of customers to target promotions and advertising.

- *Replacing/supporting human decision making with automated algorithms*[30]

Sophisticated analytics can substantially improve decision-making, minimize risks, and unveil valuable insights. Such analytics have applications for organizations from credit risk assessment in financial institutions to automatic price fine-tuning in retail.

As discussed in the Watson case, in some cases decisions will not necessarily be automated but augmented by analyzing huge, entire datasets using big data techniques and technologies. Decision making may never be the same; some organizations are already making better decisions by analyzing entire datasets from customers, employees, or even sensors embedded in products.

- *Innovating new business models, products and services*[30]

Big data can enable companies to create new products and services, to enhance the existing ones, or invent entirely new business models. Manufacturers are using data obtained from the use of actual products to improve the development of the next generation of products and to create after-sales service offerings. The emergence of real-time location data has created an entirely new set of location-based services from pricing property to casualty insurance based on where, and how, people drive their cars.

Key elements of a successful Big Data initiative

Once understood the characteristics, implications and possibilities of bid data is key to understand which factors contribute to have successful big data initiatives in organizations, these can help to create a preliminary framework to identify potential gaps and develop action plans aimed to close these gaps.

The following chapters of this work explore four key aspects that will allow organizations to make successful use of big data initiatives:

- Data collection
- Analytics
- Human capital
- Infrastructure

3. Data Collection

In this chapter, we will review some basic definitions and challenges involved in data collection, and how taking in consideration these aspects can significantly contribute to have successful big data initiatives in organizations.

To understand the three essential elements that define Big Data, High-Volume, -Velocity and –Variety becomes critical to make the right decision on how data should be collected and stored for use in the organization in order to be processed and analyzed.

In order to take advantage of big data, organizations will have to deploy the right processes, technology and skills to collect, process and make available real-time information in order that the right people in the organization can analyze their businesses environments and make the right decisions. This decision-making process can be supported by automation that can save time and improve efficiency.

A good example of this can be Tax Auditing, in this process data mining techniques, such as regression trees, can be applied to big data sets in order to analyze and identify potential irregularities in tax returns, in order to support decision making on whether or not auditing and individual, automated decision making can make a first selection of potential candidates to be audited, leaving final decision making to humans.

As more and more data is produced and collected another key point to take in consideration is "Data Shelf Life"[22], so organizations must be able to must be able to understand that specific sets of data may be useful for very short periods of time.

A good example of this is the *"Billion Prices Project"* led by Roberto Rigobon and Roberto Cavallo from MIT, in this project they automatically collect information on prices of thousands of products in almost 100 countries allowing them to develop accurate predictions on inflation up to three months in before any official data is released.

To understand the impact that the ability to analyze "big data" will have in business and organizations it is key to go deeper not only in the sources and collection techniques but also to be well aware of the important of collecting the right kind of data according to the "business question or case" we are trying to solve. Initiatives like the *"Billion Prices Project"* had first to clearly define the information and data structure requirements before developing the right data collection techniques.

Current data collection technologies represent a great temptation for businesses and it is important to keep in mind that collecting data for sake of data collection will probe technical capability but poor business judgment. In order, to identify which data is worth to collect will require close collaboration between technical and business teams in order to make the right data sourcing and collection decisions.

In the near future, data availability and analysis capabilities will transform business processes, as we become more and more competent to perform analytics not available in the past, managers will redefine (or in some cases reinvent) business processes to take advantage of this great opportunity, moving from intuition to "data driven" wil become an strategic imperative that will redefine business processes as we know the today.

As recently described by Thomas Davenport, Paul Bart and Randy Bean in Sloan Management Review,[26] *"organizations are swimming in an expanding sea of data that is* *either too voluminous or too structured to be analyzed thru traditional means. Among its burgeoning sources are the clickstream data from the web, social media content (tweets, blogs, Facebook wall postings, etc.) and video data from retail and other settings and from vide entertainment. But big data also encompasses everything from call center voice data to genomic and proteomic data from biological research and medicine" " Yet very little of the information is formatted in the traditional rows and columns of conventional databases."* Perhaps as data collection capabilities develop; structured data (as we know it today) will become scarcer, leaving the stage to more complex forms of data.

The starting point: The "Business Case"

The most important issue to consider before any big data initiative is to identify the "Business Case" or "Question" we want to answer, no "big data" initiative should be launched without clearly identify the business problem we want to tackle.

According to Gartner[25], *"a significant barrier to succeeding with big data will be the ability to ask the right question and therefore use the right technologies to get the answers."* Per se, big data it is just a mean and not a goal. And it is important to keep in mind that the only reason to pursue any Big Data initiative is to deliver against a business objective.[25]

According to Gartner[25], many of these efforts fail because the business outcome (business case or question) has not been adequately defined. This doesn't mean that organizations should presuppose or anticipate the answer, but they should first focus on defining the questions and understanding how they will use the insight an answers from this process in order to deliver the expected results to the business.

It is also important to remember that many organizations lack the skills required to exploit big data, so organizations will have to make significant investments not only to hire the talent with advance data management and analytic skills, but also train subject matter experts to become more effective "data and analytics consumers," we will talk more about this topic in the Human Capital chapter in this work.

Gartner[25] report continues, stressing that a focus on Big Data is not a substitute for the fundamentals of information management.

It should also kept in mind, however, that some of the fundamental assumptions that drive existing information quality and governance programs are significantly challenged by the volume, velocity and variety of big data assets.

If organizations do not want to suffer from the negative effects caused by poor information management, they should keep close attention to the basic principles of data quality, information governance and metadata management and should not try to apply them as-is, as they will likely cause existing governance structure and quality programs to collapse under the additional weight of the information being processed. Instead, organizations should rethink their expectations of quality and governance in the context of their control over the information and the quality requirements of the use case. For example, the quality requirements for clickstream analysis and data requirements will be very different if you are trying to figure out why people are abandoning their shopping carts than they are if you are doing fraud detection.

But probably one of the most important consequences of Big Data will be the way these capabilities will redefine and shape business processes around data. Once a critical business process has been defined and the appropriate data sources and technology identified, Managers will have to question current processes and in many cases to develop the right internal processes to capture, process and maximize the value from new insights and analysis.

Data and Databases

In order to further understand the way big data solutions are related and organized is key to understand that data may reside in databases in three basic forms: unstructured data, semi-structured data, structured data.

Also it is key to understand the kind of databases normally available in organizations: Non-relational databases and Structured databases.

Unstructured data[30]*.*

Data that do not reside in fixed fields. Examples include free-form text (e.g., books, articles, body of e-mail messages), untagged audio, image and video data. Contrast with structured data and semi-structured data.

Semi-structured data. [30]

Data that do not conform to fixed fields but contain tags and other markers to separate data elements. Examples of semi-structured data include XML or HTML-tagged text. Contrast with structured data and unstructured data.

Structured data. [30]

Data that reside in fixed fields. Examples of structured data include relational databases or data in spreadsheets. Contrast with semi-structured data and unstructured data.

Relational database. [30]

A relational database is composed of a collection of tables (relations); in this, data is stored in rows and columns. Relational Database Management Systems (RDBMS) store a type of structured data.

Initially developed at IBM, SQL (Structured Query Language) is the most widely used language for managing relational databases. SQL consists of a data definition and a data manipulation languages that include data insert, query, update and delete, schema creation and modification, and data access control. [72]*Non-relational database.*

RDBMSs (relational database management system) are an excellent alternative to handle high volumes of transactions and to manage several dimensions of information management (policy, integrity and security). [72] A Non-relational database is a database that does not store data in tables (rows and columns). [30]

New use cases, such as interactions with websites, social networks, and multi-structured data and machine-generated files, are examples of data sources requiring no relational database capabilities. [72]

NoSQL technologies allow a more flexible use in high volume, Web-scale applications in clustered environments. However, NoSQL products still have a long way to go in terms of technology development and maturity, and most vendors are small startups. [72]

NoSQL includes four solutions: [72]

- Key-value stores;
- Document-style stores;

- Table-style databases;
- Graph databases.

Each solution has characteristics and capabilities suited for particular use cases. However, it is important to keep in mind that NoSQL products are not designed to support classic transactional applications; and may lack one the atomicity, consistency, isolation and durability (ACID) of RDBMS.[72]

The use of NoSQL solutions may contribute to improve performance in a a number of important applications:[72]

- Internet commerce
- Machine-to-machine communications
- Mobile computing
- Social networking
- The integration of IT and operational technologies

The requirement to capture, process and respond to data faster, and the variety of information types employed often are not suited to an RDBMS architectures, driving a need for alternative approaches. Examples of massive new data that is increasingly targeted by NoSQL data stores include machine-generated data are created from sources such as tweets, blogs, archives of text, and audio and video content. Moreover, the volume of data often is beyond the processing capacity of traditional DBMSs.[72]

Vendors in the NoSQL space include: 10gen, Amazon, Basho Technologies, Couchbase, DataStax, Neo Technology, Objectivity, Oracle, Redis.[72]

Data intensity

According to McKinsey Global Institute (MGI),[30] The growth of big data is a phenomenon that we have observed in every sector. More important, data intensity—for example, the amount of data stored across sectors is sufficient for companies to use techniques enabled by large datasets to drive value (although some sectors had significantly higher data intensity than other).

In order to get an idea of the magnitude data intensity has exploited recently, a recent report from The Wall Street Journal reports that Facebook user data content makes up more than 100 petabytes of pictures, videos and other entries, and the analysis of this information generates another 500 terabytes of new information everyday.

MGI estimates that enterprises around the world used more than 7 exabytes of incremental disk drive data storage capacity in 2010; nearly 80 percent of that total appeared to duplicate data that had been stored elsewhere. By 2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data per company (for companies with more than 1,000 employees) and that many sectors had more than 1 petabyte in mean stored data per company.[30] Some individual companies have far higher stored data than their sector average, potentially giving them more potential to analize and derive value from big data.

According to MGI, some sectors exhibit higher levels of data intensity than others.[30] be Transaction-intensive financial services sectors, including securities and investment services and banking, have the most digital data stored per firm on average. For example, only the New York Stock Exchange, for instance, boasts about half a trillion trades a month.

Communications and media firms, utilities, and government also have significant digital data stored per enterprise or organization, which appears to reflect the fact that such entities have a high volume of operations and multimedia data.

Discrete and process manufacturing have the highest aggregate data stored in bytes. However, these sectors rank much lower in intensity terms, since they are fragmented into a large number of firms. Because individual firms often do not share data, the value they can obtain from big data could be constrained by the degree to which they can pool data across manufacturing supply chains.

Looking to a geographic profile of where big data are stored, North America and Europe together combine 70 percent of the global total currently[30]. However, both developed and emerging markets are expected to experience strong growth in data storage and, by extension, data generation at rates of anywhere between 35 and 45 percent a year. An effort to profile the distribution of data around the world needs to take into account that data are not always stored in the country where they are generated; data centers in one region can store and analyze data generated in another.

Types of data[30]

In addition to the differences in the amount of data stored in the different sectors, the types of data generated and stored—for example, whether the data encodes video, images, audio, or text/numeric information—also varies by industry sector: [30]

- *Multimedia*
    - *Video:* Communications and media, government, education, healthcare.

  - *Audio:* Communications and media, government, education.

  - *Image:* Health care, media.

- *Text / Numbers:* Banking, insurance, securities and investment services, retail, wholesale, professional services, health care, transportation, communication and media, utilities, government.

According to the McKinsey Global Institute[30] *s*ectors such as manufacturing, health care, and communications and media are responsible for higher percentages of multimedia data. R&D and engineering functions in many manufacturing subsectors are heavy users of image data used in design.

Image data in the form of X-rays, CT, and other scans represent most of health care data storage volumes. While a single page of records can total a kilobyte, a single image can require 20 to 200 megabytes or more to store[30].

If we analyze pure data generation (rather than storage), despite communications and media industries images and audio dominate storage volumes; some subsectors such as health care generate more multimedia data in the form of real-time procedures and surveillance video, but this is rarely stored for long periods of time[30].

Financial services, administrative parts of government, and retail and wholesale all generate significant amounts of text and numerical data including customer data, transaction information, and mathematical modeling and simulations[30].

Manufacturing generates a great deal of text and numerical data in its production processes[30].

The power of linked data

So far we have discussed the importance of big data and the way different industries will benefit from this, however, it may appear we are talking about different "islands" of different kinds of data barely connected among each other, it is key to keep the vision that big data will evolve in more complex ways and gains considerable value when linked.

Linked data is a Web-oriented approach to information sharing that is having a transformational impact on the information architecture of an enterprise[31]. Linked data is a data management and mathematical principle, which holds that anytime two data points are used together it creates a weighted link[30]. As more valid use cases for the link emerge, the link grows in weight. *"Linked data is the principle that this "weight" is further increased by the use of datasets in a pervasive linking pattern and the uses of the sets of data begin to alter business processes to the positive"*[21]. Currently, the most common emerging practice for using linked data concepts is found in Web-oriented approaches[21].

A study from Gartner[31] estimates that linked data will reach mainstream adoption in five to seven years and reveals two key findings of great relevance for the integration of linked data to any big data initiative:

- Linked data lowers the barriers to the reuse, integration and application of data from multiple, distributed and heterogeneous sources. By combining data across silos, linked data is an effective way, enabling organizations to make more rapid discoveries about business environments.
- Business benefits offered by the use of linked data include revenue generation, deeper connections with customers and easier compliance to transparency mandates.

According to Gartner[31], Linking internal corporate data with external data from the Web allows organizations to increase the value of their information assets with relatively low effort. Linked data allows us to see the bigger picture and establish meaningful relationships.

For internal IT, linked data is considered a disruptive trend in information architecture, information management and data integration[31]. In the Web Age of ubiquitous data flows, information structures are no longer static and cannot be under the control of a single organization[31]. Consequently, IT leaders shift their focus from the storage and compression of data to mastering the flow and shareability of information using linked data approaches[31].

Key Challenges

According to Deloitte[33] there are three key challenges when it comes to Data Management, these are: Quality, Governance, Privacy

Data Quality

Initiative considering big data has to make sure not to compromise quality of data due to volume and variety of data. It is key to maintain sight in all costs of maintaining all data quality dimensions: Completeness, validity, integrity, consistency, timeliness, and accuracy

Governance

Identifying relevant data protection requirements and developing and appropriate governance strategy. This may require to re-evaluate internal and external data policies and the regulatory environment

Privacy

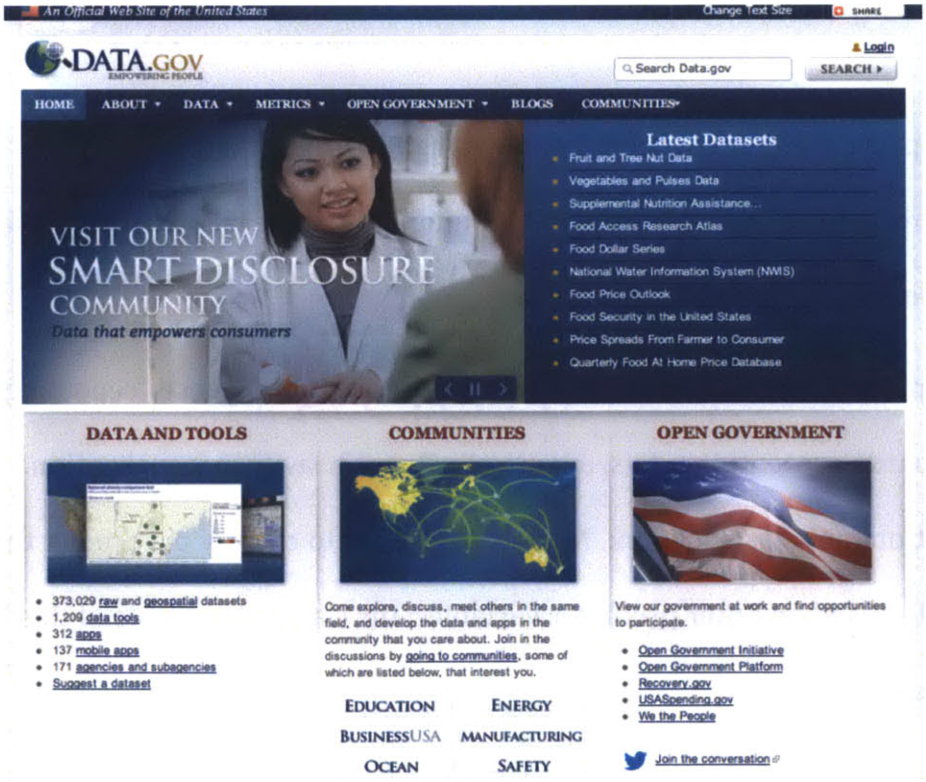
As data collection technologies and sources increase, privacy issues related to direct and indirect use of big data assets, in the future this might imply an increased regulatory environment and evolving security implications of big data.

Public Sources of Data - Data.gov

Data.gov is a public depository of high value, machine-readable datasets generated by the U.S. Government.

Data.gov increases the ability of the public to easily find, download, and use datasets, providing descriptions (metadata) and tools that leverage government datasets.

A primary goal of Data.gov is to expand creative use of those data beyond the walls of government by encouraging innovative ideas and applications.



Latest Datasets

- Fruit and Tree Nut Data
- Vegetables and Pulses Data
- Supplemental Nutrition Assistance...
- Food Access Research Atlas
- Food Dollar Series
- National Water Information System (NWIS)
- Food Price Outlook
- Food Security in the United States
- Price Spreads From Farmer to Consumer
- Quarterly Food At Home Price Database

DATA AND TOOLS

- 373,029 raw and geospatial datasets
- 1,209 data tools
- 312 apps
- 137 mobile apps
- 171 agencies and subagencies
- Suggest a dataset

COMMUNITIES

Come explore, discuss, meet others in the same field, and develop the data and apps in the community that you care about. Join in the discussions by going to communities, some of which are listed below, that interest you.

EDUCATION ENERGY
BUSINESSUSA MANUFACTURING
OCEAN SAFETY

OPEN GOVERNMENT

View our government at work and find opportunities to participate.

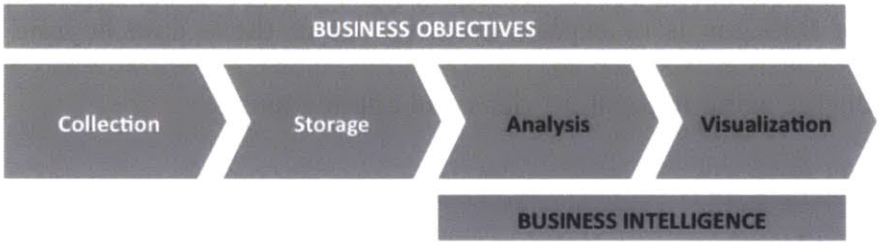
- Open Government Initiative
- Open Government Platform
- Recovery.gov
- USASpending.gov
- We the People

Join the conversation

4. Data Analysis

As discussed in previous chapters, the only reason to pursue a big data initiative should be to deliver against a specific business objective[25].

So far we have discussed the challenges and opportunities associated with data collection and storage of information assets; however, this is just one side of the issue on the other side we have the two key components of business intelligence: Analysis and Visualization.



These two aspects will allow organizations not only to get the right insights but also to understand and widely communicate these findings to relevant decision makers. Data analysis can be split in two main applications:

- Descriptive analysis: The "classical" business intelligence reporting, including dashboards presenting totals, means, etc. Relies heavily in visualization.
- Prediction and Classification: Models that allow to classify or predict the value of a variable based on other predictor variables, the classical example of this is the linear regression.

These techniques are applied in different fields in order to develop specific insights, for example micro-targeting, a technique based cluster analysis and widely used in marketing, now has become a key tool in political campaigns, in order to identify potential voters.

Why is important to understand these techniques? Many vendors offer black box solutions, we risk fall to the temptation to blindly trust on solutions that do not necessarily are adapted to our analysis needs and opportunities. Big data offers the potential of increased personalization and this concept applies also to select the right technique or solution to answer the "business case" we want to solve.

Most techniques draw on disciplines such as statistics and data mining that can be used to analyze datasets. In the following pages, we provide a review of techniques and I provide examples on how these can be applied in business applications. This list is by no means exhaustive.

SEMMA Methodology[37]

Any technique used in big data analysis has to follow a development and validation process before roll-out, the SEMMA methodology is a good alternative for the development of predictive and classification models, this methodology was developed by SAS corporation, and it is conceived around beginning with a statistically representative sample of data, and apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy[37]. This methodology is composed by five components:

1. **Sample** from data sets, Partition into Training, Validation and Test datasets
    - *Training set* is the data used to construct the model.
    - *Validation set* is the data used to validate the predictive ability of the model; adjustments to the model may result from the model validation.

- o *Test set* is the data used to confirm the predictive ability of the model

2. **Explore** data set statistically and graphically in order to get a sense of the relationships between variables. This includes the use of descriptive statistics on the dataset.
3. **Modify**: Transform variables and impute missing values in order to have data in the right format before modeling.
4. **Model**: fit models are created (e.g. regression, classification tree, neural net.) this

   is a process itself that can be divided in four basic steps: *Built* the Model, *Evaluate the Model, Re-evaluate the model and adjust*
5. **Assess**: Compare models using Partition, Test datasets are used for this purpose.

Techniques for analyzing Big Data

A number of techniques have allowed the development of data science; statistics and data mining are the biggest contributors to the development of data science. In the following pages I present the most common techniques for analysis, and despite this review it is not exhaustive it provides a good sense of techniques currently available.

It is important to mention that despite most of these techniques do not necessarily require big data sets (e.g. regression), all of them can be applied to big data sets. Larger and more diverse datasets can be used to generate more diverse and insightful results than smaller, less diverse ones[30].

K-Nearest Neighbors[34, 36]

The k-nearest neighbors method is a non-parametric classification and prediction technique that relies on identify "k" records in the training dataset that are similar to a new record we want to classify[36].

This method allows breaking up data or population in smaller groups composed of records similar to each other. The k-nearest record neighbors of a record "x" are data points that have the k smallest distance to "x." This requires a measure of distance between records; the most popular measure of distance between records is the Euclidean Distance between the two records[36]. This measure is very sensitive to scale, so normally it is recommended to standardize or normalize the data before computing distances.

The general idea of the method can be explained in two simple steps[36]:

1. Find the "k" nearest neighbors (1…n records) to the record to be classified.
2. Use a majority decision rule to classify the record, where the record is classified as a member of the majority class of the k neighbors.

Practical applications of the K-Nearest Neighbors method include recommendation machines similar to those developed by amazon.com or Pandora.

Naïve Bayes[34, 36]

Naïve Bayes classifier is a probabilistic framework for solving classification problems by estimating the probability that a record belongs to a class of interest. This method works only with categorical variables and in those cases were numerical predictors are available these should be converted to categorical predictors.

Practical applications of this method include predict voting behavior based on demographic characteristics of individuals, fraud detection and spam filtering

Classification and Regression Trees[34, 36]

The tree methodology can be used for two purposes, classification and prediction (called regression tree). This method is based in in separating records or observations in subgroups by creating splits (recursive partitioning) on predictors.

This technique is considered to work well in across a wide range of applications as requires reasonable computing effort, it provides easily understandable classification rule, it is quite easy to understand to by the consumer of the analysis, and in some cases can handle categorical predictors.

Discriminant analysis[34, 36]

Discriminant analysis is a classification technique that can be used for classification and profiling. It is based on using continuous predictors (variables) measurements on different classes to classify new items into one of those classes (classification).

This method tends to be considered more of a statistical classification method than a data mining method, and the use and performance of it are similar to those of linear regression, by searching for the optimal weighting of predictors. However, in discriminant analysis the weighting is it in relation to separating the classes. Both techniques use least squares as estimation methods, and the resulting estimates are robust to local optima.

This technique can be used to highlight aspects that distinguish classes in the dataset (profiling), to classify applications for credit cards, loans and insurance; classifying potential customers of new products as early adopters, early majority, late majority or laggards.

Logistic Regression[34, 36]

Logistic regression is a classification technique that extends the ideas of linear regression to the situation where the dependent variable (Y) is categorical. The goal is or simply to classify the observation into one of the classes.

This technique can be used to classify a new record that the class is unknown into one of the categories, based on the value of the predictor variable or to predict which class a new observation will belong. It can be also used for profiling.

The most common use of logistic regression is to deal with binary dependent variables with two classes (0,1), getting the probability of belonging to class 1, after that we use a cutoff value on this probability in order to classify the record in one of the classes, if it goes below the cutoff value then the record belong to class 0, if it goes above the cutoff value the record belongs to class 1.

Neural Networks[34, 36]

Neural networks (also called artificial neural networks) can be used for classification or prediction. This technique is based on the way neurons are interconnected in the human brain and are capable to learn from experience, trying to mimic the way in which human experts learn. This technique combines information in a very flexible way that captures complicated relationships among these variables and between them and the response variable leading to a good predictive performance where other methods fail.

Multilayer feed-forward networks are the most widely used and successful neural networks include tree layers: Input layer, hidden layer, and output layer. Information flow is one-way and processing is performed one observation at a time.

Despite neural networks are considered to have good predictive performance and are able to model complex relationships, they are computationally intensive, require much data and may get stuck in a local optimum.

Applications in financial industry include bankruptcy predictions, currency market trading and commodity trading.

Regression[30, 36]

The multiple linear regression model is probably the most popular technique to analyze data, it allows to determine the value of the dependent variable when one or more independent variables are modified.

This model is used to fit a linear relationship between a quantitative dependent variable Y (outcome or response variable) and a set of predictors (X_1, X_2,... X_P) also knew as independent variables. It assume that the population holds the following relationship

$$Y = \beta_0 + \beta x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

Where β_0 ... β_p are the coefficients and ε is the noise or unexplained part. The most common reasons of fitting a linear model to the dataset include understanding the underlying relationships in the dataset, and second, to predict the outcome of new cases.

This method is widely used in sales forecasting, manufacturing, marketing, and measurement of customer satisfaction.

Association Rules

Association rules also known as affinity analysis or market basket analysis, has its origins the interest of manager o know if certain groups of items are consistently bought together evolving in the study of customer transactions databases in order to identify the dependences between the purchasing these group[36].

Practical applications of association rules are commonly encountered in online recommendation engines, such as that of amazon.com, in which consumer reviewing an specific articles are presented other articles commonly purchased in conjunction with the first item[36].

The concept behind the association rules technique is to examine possible associations between items in an *if-then* format and select those with highest possibility of reflecting a true dependence. Producing dependency rules that will predict occurrence of an item based on occurrences of other items[36].

The process of creating association rules has two steps. First, generate a set of candidate rules based on frequent item sets, (apriori algorithm). Second, from these candidates, measures of support and confidence are used to evaluate the uncertainty in a rule[36]. The minimal support and confidence values to be used in rule selection process have to be defined[36]. The lift ratio, allows comparing the efficiency of the rule to differentiate a real association from a random combination[36].

Time series analysis

Time series analysis is one of the most common techniques used in business, nearly every organization produces time series data that eventually can be used in forecasting; this includes retail businesses forecasting sales, utilities companies forecasting demand and prices, and governments forecasting tax collections[36].

Modeling time series can be done for explanatory or predictive purposes. When used for explanatory purposes, time series modeling allow to determine the components such as seasonal patterns, trends, and relations to external factors these can be used for to support decision making or policy formulation[36]. When used for forecasting, uses the information in time series and other complementary information to predict the future values in the series. The differences between the results of time series analysis and time series forecasting lead to the selection of different modeling alternatives. For example, when used for explanatory purposes, time series analysis priority is given to techniques that produce explainable results; explaining is done normally is retrospective terms. Time series forecasting models are prospective in nature and they aim to predict the future values in the time series[36].

A/B Testing

This technique is also known as split testing or bucket testing. In A/B testing a control group is compared with a variety of test groups in order to determine what treatments or changes influence a given objective variable. This technique is frequently used to determining what text, layouts, images, or colors contribute to improve conversion rates in e-commerce[30].

Big data technology allows to execute and analyze hundreds o experiments, ensuring that groups are of sufficient size to detect statistically significant differences between the control and treatment groups[30]. If one variable is simultaneously manipulated, the multivariate generalization of this technique is often called "A/B/N" testing[30].

Spatial Analysis[30]

A combination of statistical techniques to incorporate topological, geometric, or geographic properties encoded in a data set into a model. Often the data for spatial analysis come from geographic information systems (GIS) that capture data including location information, such as addresses or latitude/longitude coordinates[30].

Examples of applications include the incorporation of spatial data into regressions to understand consumer willingness to purchase a product correlated with location or understand how would a manufacturing supply chain network performs with sites in different locations[30].

Hierarchical Cluster Analysis[36]

This technique is based on the unsupervised learning task of clustering where the goal is to segment the data into a set of homogeneous clusters, each having a set of attributes, and a similarity measure among them, this includes:

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

In hierarchical clustering, observations are sequentially grouped in cluster, based on distances between observations and distances between clusters.

To measure the distance between records there are measuring techniques such as Euclidean Distance if attributes are continuous, but there are other distance measures that can be used, such as correlation-based similarity, statistical distance (also called Mahalanobis distance), Manhattan distance and maximum coordinate distance.

To measure the distance among cluster there are several options, these include Single Linkage which measurers the distance between the pair of records that are closest; Complete Linkage which measures the distance between the pair of records that are the farthest; Average Linkage which averages the distance of al possible distances between records in one cluster and records in the other cluster; and Centroid Distance which measures the clusters centroids.

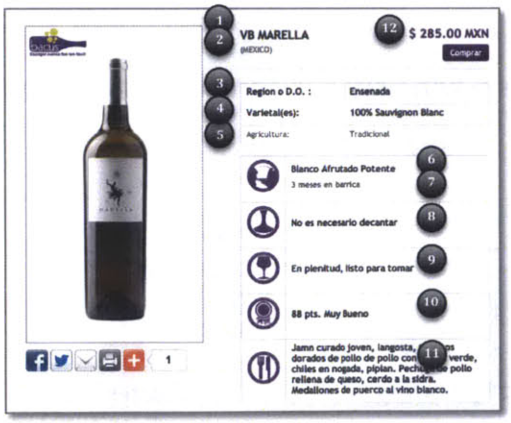
Applications of this technique include recommendation engines, market segmentation, market structure analysis, creation of balanced portfolios and even defining new uniform sizes in for the U.S. army.

Case Study: Wine Recommendation using Hierarchical Clustering

An example of a practical application of hierarchical clustering method wine classification and recommendation tool I developed. For this purpose I used a database of 313 wines from 8 countries: Argentina, Chile, Slovenia, Spain, France, Italy, Mexico and Uruguay. All of these were classified by at least two professional sommeliers. Using a classification system based continuous and categorical variables; it is worth to mention that many variables are the result of the subjective appreciation conducted by a sommelier in order to classify flavor, when to drink, recommended service among others.

The goal is to be able to recommend the two closest records in the dataset to a specific record selected by a hypothetical consumer.

In 2008, I launched a small wine retail business in my hometown; Mexico City, as part of our value proposition we developed a simple product classification system in order to help our customers to compare and select the wine of their choice, we have called this the "Sensorial Guide" the sensorial guide offers the following information.



VB MARELLA
(MEXICO)

$ 285.00 MXN

Region o D.O. : Ensenada
Varietal(es): 100% Sauvignon Blanc
Agricultura: Tradicional

Blanco Afrutado Potente
3 meses en barrica

No es necesario decantar

En plenitud, listo para tomar

88 pts. Muy Bueno

Jamn curado joven, langosta, dorados de pollo de pollo con verde, chiles en nogada, pipian. Pech pollo rellena de queso, cerdo a la sidra. Medallones de puerco al vino blanco.

1. *Name.*
2. *Country of origin.*
3. *Region*
4. *Blend or varietal*
5. *Agriculture*
6. *Style Classification (taste)*
7. *Months in oak barrel*
8. *Recommended decanting or "aeration"*
9. *"When to drink" recommendation*
10. *Rating*
11. *Food Pairing*
12. *Price*

As of today we have totally or partially classified more than 520 wines; of these, 440 wines from 15 countries have been fully classified and this data is available. Due to computing limitations I finally used only 313 wines from 8 countries and 30 origin denominations.

The "Sensorial Guide" includes variables measured by expert wine tasters (Sommeliers) and all classification has been made in Spanish. The details of each attribute are:

1. Name. Name of the wine in its original language
2. Country of origin. Name of the country in which the specified wine has been produced.
3. Region. Name of the specific region on "appellation of origin" in which the wine has been produced. Good examples of this are Bordeaux in France,or Mendoza in Argentina.
4. Blend or varietal. Wine can be produced either 100% from a single varietal (something very common in the United States) or can be produced from a blend of different varietals.
5. Agriculture. Four agriculture classifications: Traditional, Organic, Biologic, and Biodynamic.
6. Style Classification (taste). This defines a wine style is done by sommeliers in our team in three steps:

1. Color	**2. Style**	**3. Intensity**
1. Red	**Red**	**1.** Light*
2. White	1. Fruity*	**2.** Medium*
3. Rose	2. Tannic*	**3.** Strong*
4. Sparkling	3. Sweet*	
	White, Rose and Sparkling	* Indicates a variable analyzed and rated by a Sommelier
	4. Fruity*	
	5. Dry*	
	6. Semi-Dry*	
	7. Sweet*	

7. Months in oak barrel. Information provided by the wine maker.

8. Recommended decanting or "aeration." This is done by sommeliers and classifies as follows:

- No need of decanting or aeration
- Pour in a pitcher before serving
- Oxygenate or aerate for 15 minutes
- Oxygenate or aerate for 30 minutes
- Oxygenate or aerate for 45 minutes
- Oxygenate or aerate for 60 minutes
- Separate solids

9. When to drink" recommendation

- Ready to Drink
- Drink in 1 to 2 years
- Drink in 3 to 4 years
- Drink in 5 years or more

10. Rating. This is done by sommeliers and classifies as up to 100 points

11. Food Pairing. This is done by sommeliers and includes suggestions of food that harmonize with this specific wine.

12. Price. Current retail price in our stores. Prices are presented in Mexican pesos

To conduct this analysis I decided to use the described before variables except for the Varietals used for every wine and the food paring recommendation leaving the following variables:

Categorical Variables	• Country of Origin • Region • Agriculture • Style • Recommended decanting • When to drink recommendation	These variables were converted to dummy variables
Continuous Variables	• Months in oak barrel • Price • Rating	These variables were normalized

In order to recommend a similar wine to the one selected by the customer (reference record), we need to take in consideration that we need to have those records with the lowest distance from the reference cluster. Resulting in 30 different clusters the Average Distance (Average Linkage) method resulted as the first choice for a recommendation engine.

Cluster	# of records	Cluster #	Cluster	# of records	Cluster #
Aguacalientes Mexico	4	9	Italian Sicilia	2	5
Argentinian Cachapoal	2	22	Italian Veneto - Montepulciano	5	3
Biodinamic Toro	1	30	Mainstream	213	1
Biological Argentina and Mexico	2	21	Mexican Clairete	1	18
			Mexica Demi-sec	1	26
Canelones Uruguay	10	8	Organic France	1	29
Chilean Limari	8	20	Premium Venetian Italy	2	28
Chilean Maipo	4	6	Slovenia	4	14
Chilean Maule	4	10	Slovenia Demi-sec	1	27
Alsacian France	5	7	Spanish Penedes	1	23
French Beaujolais	6	15	Spanish Priorat	5	17
French Bourgogne	7	25	Spanish Rioja	7	12
French Cahors	3	19	Spanish Almansa	7	2
French Cotes du Rhone	4	11	Sweet France	1	24
Italian Abruzzo	1	16			
Italian Rubicone	1	13			

5. Visualization

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported[34].

Visualization of data is one of the most powerful and appealing techniques for data exploration [35] for three main reasons:

- Visualization allows the graphical representation of data that demonstrates complex concepts.
- Humans have a well-developed ability to analyze large amounts of information that is presented visually[34].
- Humans can detect general patterns and trends – Visualization allows finding relationships, and detecting outliers and unusual patterns[35].

However, data visualization implies several challenges for organizations, in a recent white paper SAS institute[38] identifies the following challenges::

- *Speed.* To integrate the right technological solutions to deliver these visualizations in fast speed, according to SAS Institute[38], the delivery speed challenge will increase as granularity and the complexity of relationships increases in business intelligence
- *Understanding the data.* To have good domain expertise in place will allow finding the right visualization tools in order to make sure decision-makers in organizations have the right elements to react to their business environment.

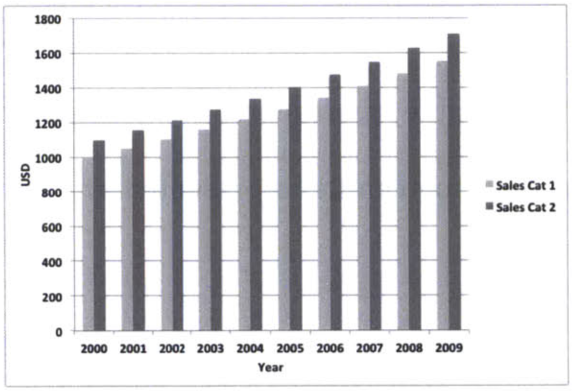
- *Displaying meaningful results.* Displaying massive amounts of data in different categories can significantly complicate processing and delivery. Subject matter experts in conjunction with business leaders should analyze and define the appropriate levels of granularity required for decision-making.

The data visualization process, is probably one of the most challenging issues in business intelligence as it has to convey the right levels of speed, processing capabilities, granularity and graphic communication to deliver the right message.

The following pages present the most common forms of data visualization, despite it would be possible to find visualization techniques far more complex than the ones presented here, we have to keep in mind that these new forms frequently end up being combinations or modifications derived from these basic forms.

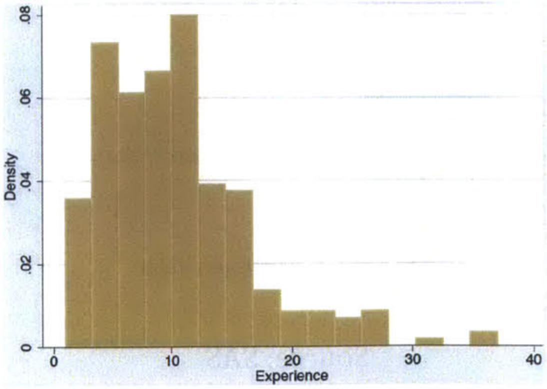
Bar Charts[38]

Probably the most widely used graphic in business, bar charts consist of a grid and a number of vertical or horizontal columns or bars. Each column represents quantitative data. Bar charts are commonly used to compare the quantities of different categories or groups.



Histogram

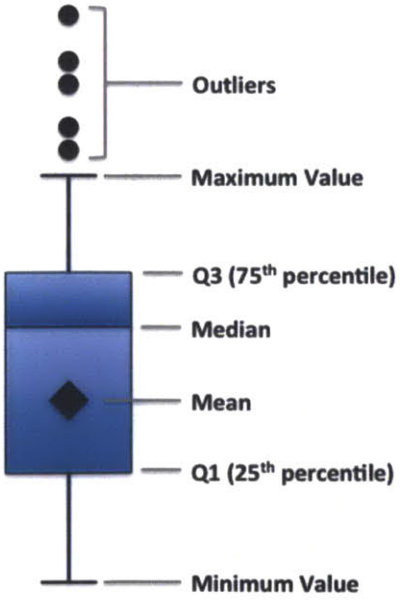
Histograms are variations of bar charts that show the frequency of data in successive numerical intervals of equal size, the sum of this density equals 100% and is quite useful for to visualize the distribution in a data set. The bar height can also represent the number of observations for each value range[42].



Box Plots[34, 38]

Box plots can be used to compare attributes and represent the distribution of data values by using a rectangular box and lines called whiskers. They display five statistics that summarize the distribution of a set of data:

- *The minimum value,* represented by the whisker that extends out from the lower edge of the box
- *The lower quartile* (25th percentile) represented by the lower edge of the box
- *The median,* The median (50th percentile) represented by a central line that divides the box into sections
- *The upper quartile* (75th percentile) represented by the upper edge of the box
- *The maximum value*, represented by the whisker that extends out from the higher edge of the box.



Source: SAS

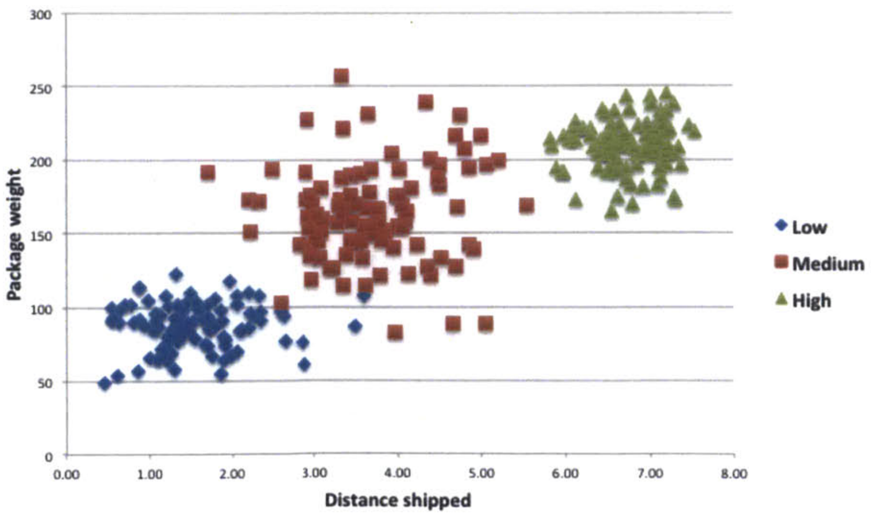
Scatter Plots[34, 38]

Scatter plots are two-dimensional plots that show the joint variation of two (or more) variables. They are useful for:

- Examining the relationships, or correlations, between numeric data items,
- Get a sense of how spread out the data is and how closely related data points are.
- Identifying patterns present in the distribution of data.

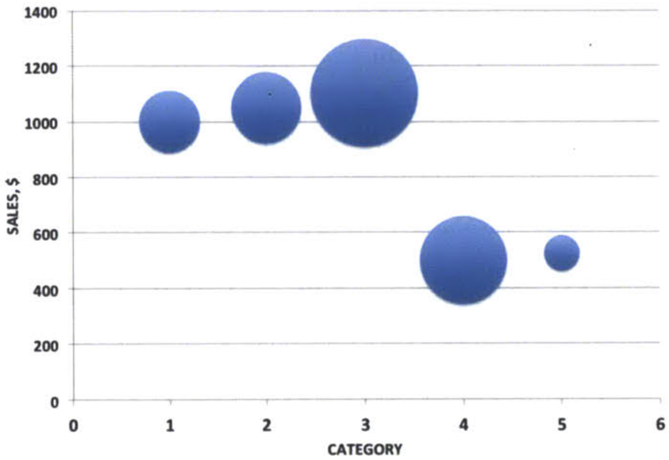
The most relevant characteristics of scatterplots include:

- Attributes values determine the position
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects, for example, bubble charts are a form of scatter plot in which color and size of the bubble are additional attributes for visualization.
- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes.



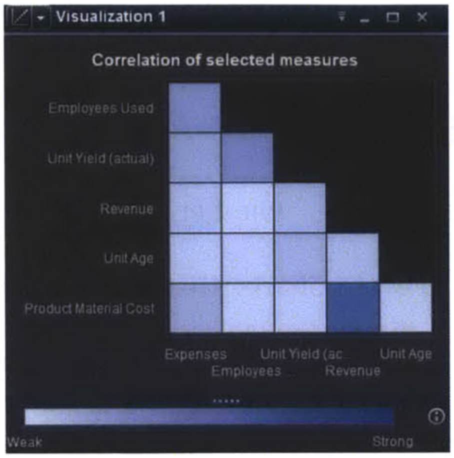
Bubble Plot[38]

The bubble plot is variation of the scatter plot in which the data markers are replaced with bubbles; each bubble represents an observation or a group of observations. The values of two measures are represented by the position on the graph axes, and the value of the third measure is represented by the "bubble" (marker) size. These are useful for data sets with additional measures of magnitude. Animated bubble plots are a good way to display changing data over time.



Correlation Matrix[42]

Correlation matrices allow to quickly identifying which variables are related and the strength of the relationship between them. The correlation matrix displays the degree of correlation as a matrix of rectangular cells, in which each cell represents the intersection of the two measures, and the color of the cell indicates the degree of correlation between those two measures[38] (it can include also the actual correlation index for the categories or variables analyzed). A positive correlation value means that as one variable increases, the second variable increases. A negative correlation means that as one variable increases, the second variable decreases[38].

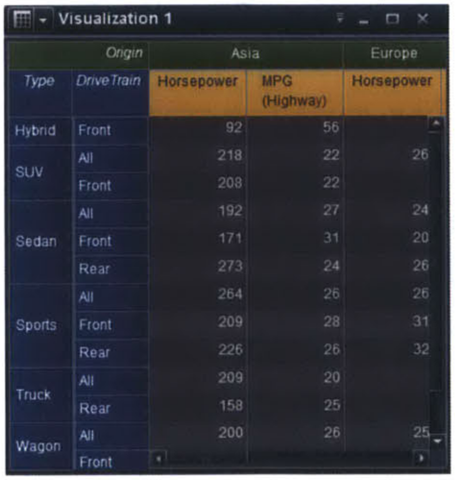


Source: SAS[38]

Cross-Tabulation[42]

Cross-tabulation tables show frequency distributions or other aggregate statistics for the intersections of two or more categories or variables in the data set[38].

Crosstabs allows examining data for intersections of hierarchy nodes or category values. These tables have the flexibility to rearrange the rows and columns and apply sorting.



	Origin	Asia		Europe
Type	*DriveTrain*	Horsepower	MPG (Highway)	Horsepower
Hybrid	Front	92	56	
SUV	All	218	22	26
	Front	208	22	
Sedan	All	192	27	24
	Front	171	31	20
	Rear	273	24	26
Sports	All	264	26	26
	Front	209	28	31
	Rear	226	26	32
Truck	All	209	20	
	Rear	158	25	
Wagon	All	200	26	25
	Front			

Source: SAS[38]

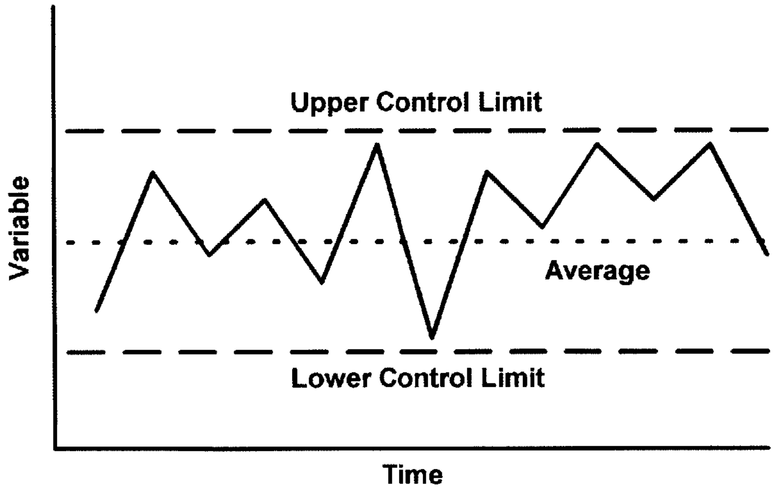
Run Chart[34]

A Run Chart is used to display how a process performs during a specific time period. It is a line graph of data points plotted in chronological order. Data points represent measurements, counts, or percentages of process output.

Using Run Charts allow determining whether a process is stable, consistent, and predictable. Unlike other tools (Pareto Charts, Histograms), Run Charts display data in the sequence in which they occurred (chronological order), helping to detect signals of causes of variation and deviations from pre-established limits.

A Run Chart also allows you to present some simple statistics related to the process:

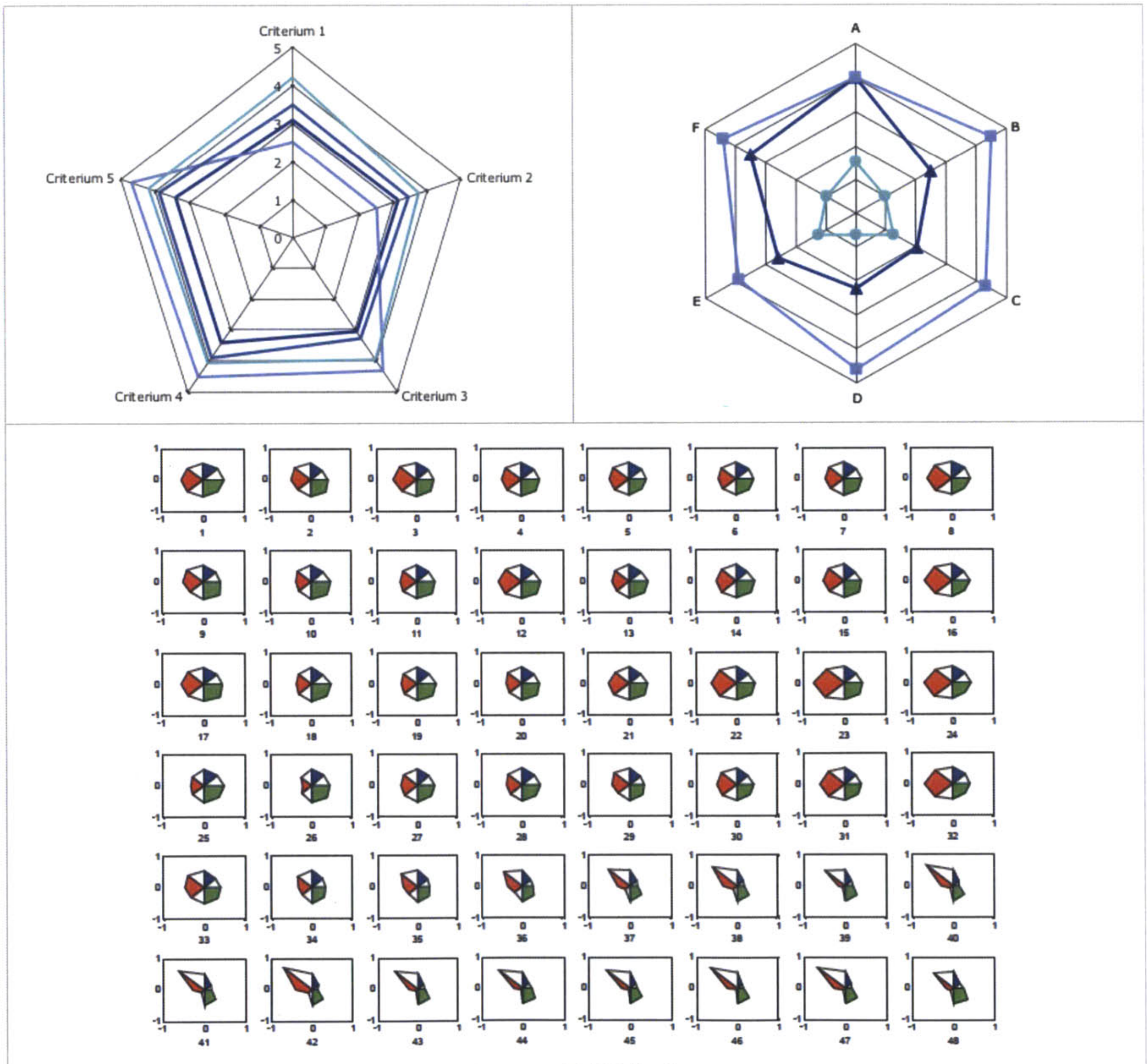
- Median / Average: Centerline on the Run Chart.
- Range: The difference between the upper control limit and the lower control limit values in the data.



Star Plots[34]

In this technique axes radiate from a central point, each axe can be category or variable in the data; scale can be adjusted or not in order to maintain the symmetry of the plot.

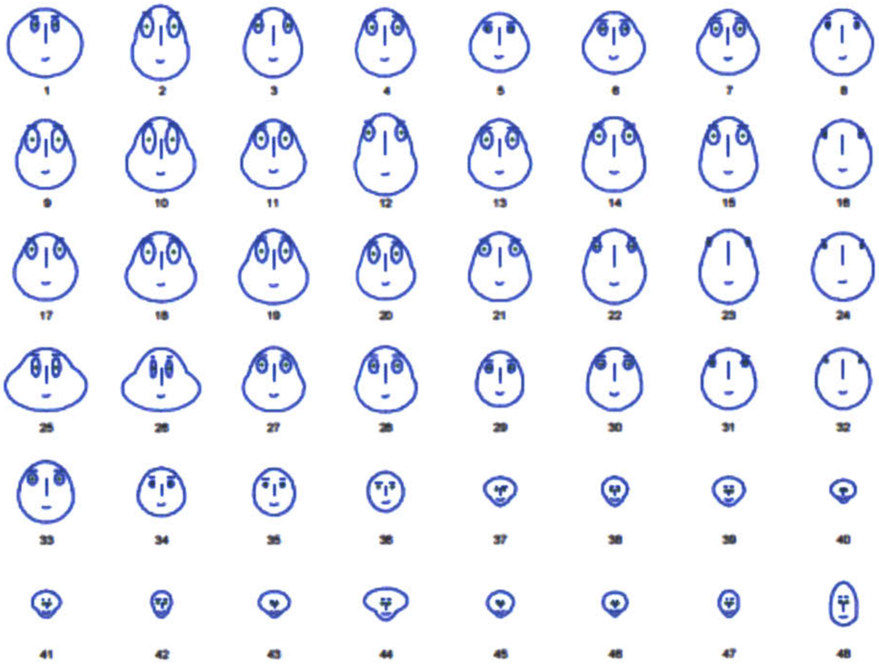
Data points are connected by a line creating a polygon, this allows to create multidimensional comparison in two planes.



Source: R. Welsch

Chernoff Faces[34]

A Chernoff face is a data visualization technique created by Herman Chernoff. This technique relies on human's ability to distinguish faces. Each attribute is associated with a characteristic of a face and the value of each attribute determine the appearance of the corresponding facial characteristic resulting in that each object becomes a separate face.



Source: R. Welsch

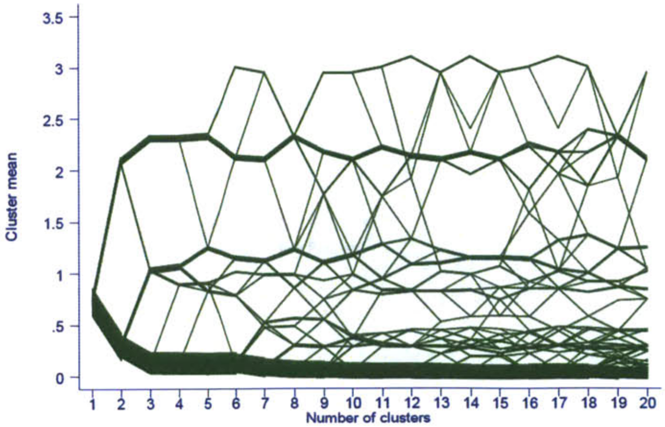
Word cloud

Word cloud is a visualization technique in which words that appear most frequently are larger and words that appear less frequently smaller. This type of visualization helps the reader to quickly perceive the most salient concepts in a large body of text. [30]

development
business
sales
new
Led
marketing
financial
resulting
product
model
Data
strategy
Management
leading
customer
Program
insurance
MELLON
Magazine
increasing
major
application
Contributed
acquisition
CARNEGIE
teams
President
plan
FULBRIGHT
Project
two
EUR
MEXICAN
TECHNOLOGY
SCIENCE
NPV
TV
market
MBA
proposals
Graduate
Market
North

Clustergram

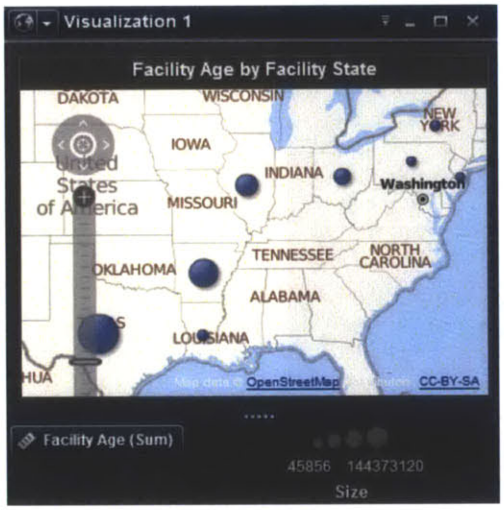
A clustergram displays how individual members of a dataset are assigned to clusters as the number of clusters increases. This technique enable to better understanding of how the results of clustering vary with different numbers of clusters. [30]



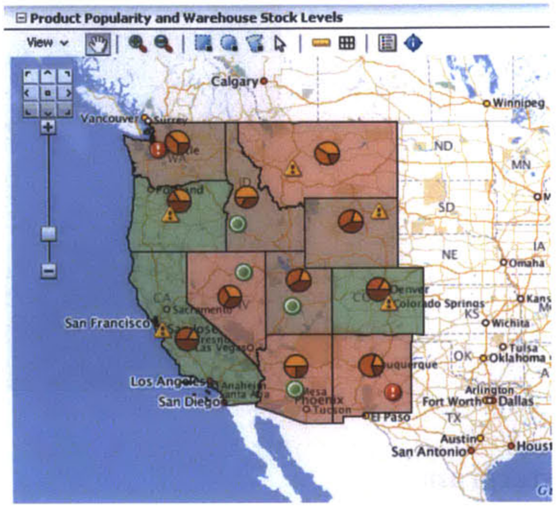
Source: MGI[30]

Geo Map[42]

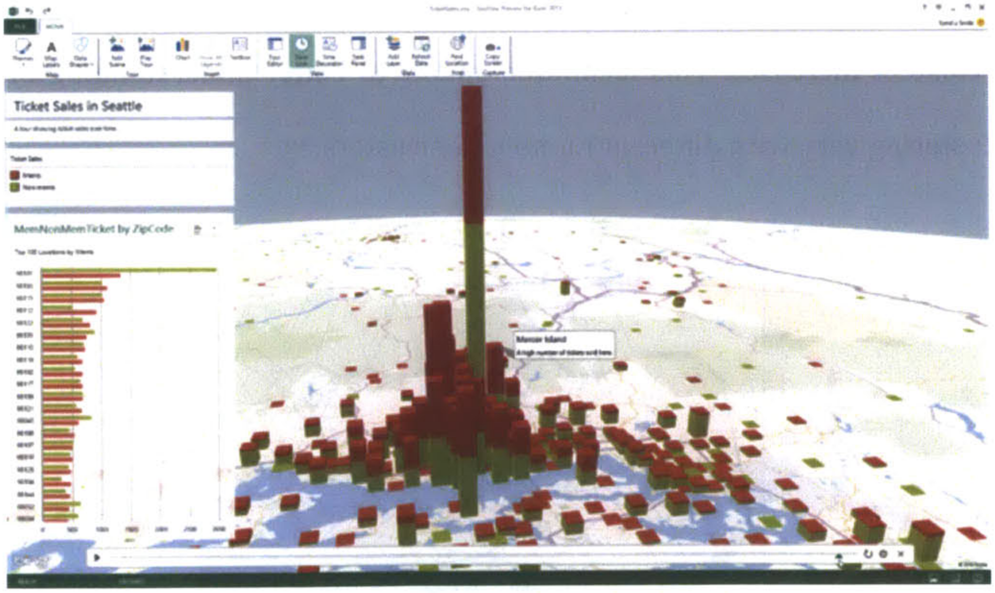
Geo maps display data overlaid on a geographic map by superimpose multiple layers of information. These layers can include: bar graph, pie graph, color, point, bubbles etc. To display a geo map requires defining one or more categories as a geography data item[38].



Source: SAS[38]



Source: Oracle[39]

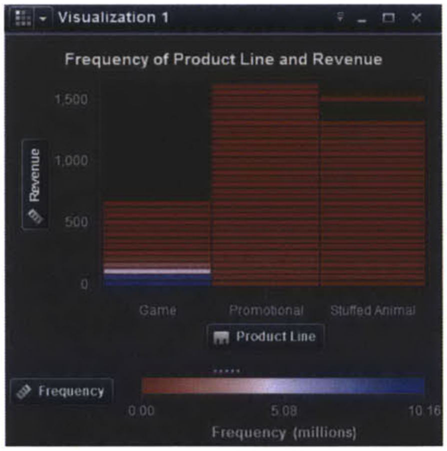


Source: Microsoft[40]

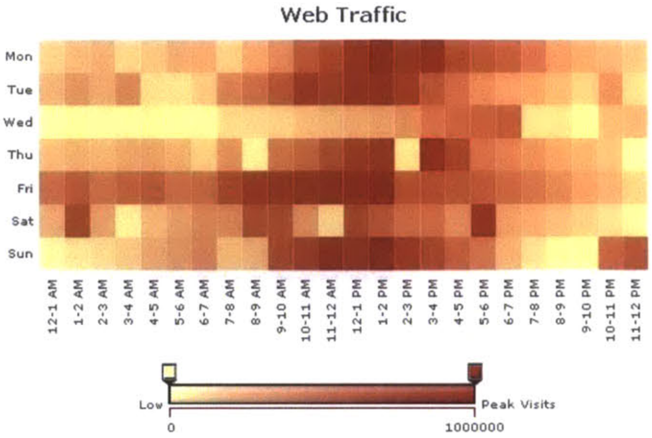
Heat Map[41,42]

Heat maps display the distribution of values data items using a table with colored cells. Colors are used to communicate relationships between data values that would be harder to understand if presented in a spreadsheet. Some of the characteristics of heat maps include:

- Data is arranged in a tabular format containing finite number of rows and columns.
- Both numeric and non-numeric data can be provided to plot data.
- A solid color or a gradient color can be set to display a range of values.
- Interactive legend is used to show or hide data plots.



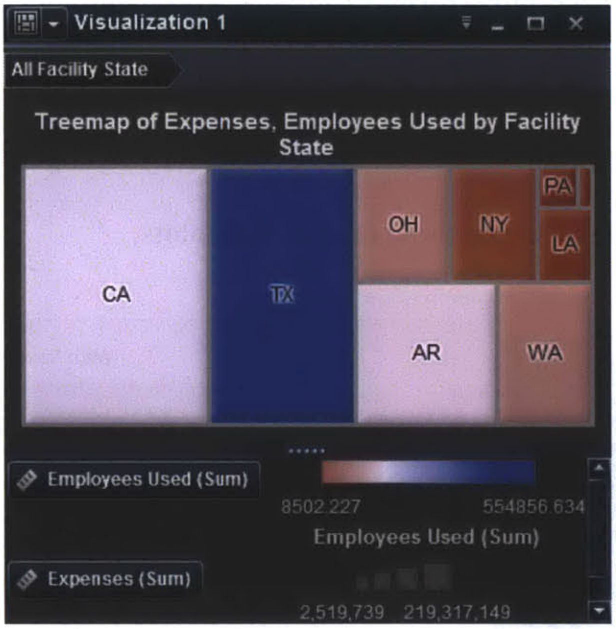
Source: SAS[38]



Source: Fusion Charts[41]

Tree Map[42]

Tree maps use rectangles (tiles) to represent data components. The largest rectangle represents the dominant division of the data and smaller rectangles represent subdivisions. The color of each rectangle can indicate the value of an additional measure[42].



Source: SAS[38]

6. Impact

Big data will transform many industries will evolve in the future, the extensions of big data in decision-making environment are several, the following table presents some of the implications of this evolution in business intelligence moving from traditional decision making to data-driven decision making:

Traditional Decision-Making Environment	**Big Data Extensions**
Determine and analyze current business situation	Provide complete answers, predict future business situations, investigate new business opportunities
Integrated data sources	Virtualized and blended data sources
Structured data	Multi-structured data
Aggregated and detailed data (with limits)	Large volumes of detailed data (no limits)
One size fits all data management	Flexible and optimized data management
Reporting and OLAP	Advanced analytics functions and predictive models
Dashboard and scorecards	Sophisticated visualization of large result sets
Structured navigation	Flexible exploration of large results sets
Humans interpret results, patterns and trends	Sophisticated trend and pattern analysis
Manual analyses, decisions and actions	Model/rules-driven decision and actions

Source: Vertica

In this section we will review five domains in which big data can play a transformative role

- Political campaigns
- Education
- Customer Marketing
- Insurance
- Health

Political Campaigns

It was in 2000 when George W. Bush advisers started the process of using commercial and consumer data to tag and eventually persuade potential supporters for the 2004 re-election process[46]. Since then, the power of increased computing and big data storage capabilities has changed the way political campaigns should be run in the United States.

More recently, the 2012 presidential campaign seems to open our eyes to how political campaigns will be fight in the future; and how this new era of "linked" information may transform privacy in a scarce asset.

Big data has the power of transforming our lives and understand our environment, and that is precisely what a team of data scientists in the Democrat party has been doing in favor of their candidates.

They understand that the use and combination of different data sources can create significant competitive advantage when running for public posts. In his book "The Victory Lab" author Sacha Issengber[25] gives a glance of the techniques used to apply data science to win political campaigns. He describes the basic logic behind this new savoir-faire in political marketing; *"When Home Depot wanted to know where to put a new store, it made sense to buy a list of people in a given ZIP code who had swimming pools. But political campaigns were less interested in a single consumer variable that in combining many of them with election-specific information, such as polls or phone bank IDs, to find patterns."*

Based on these principles Catalist [48] (catalist.us) a consulting company focused on supporting Democratic campaigns and liberal causes, has developed methodologies and algorithms to better develop micro-targeting techniques: [47]

- Everything starts with a system in which each individual is assigned unique ID number (similar to that of social security) that allows tracking her movements and keeping contact with over her life.
- Whenever is possible, information on old voters and phone banks, is collected.
- This information is blended with information available in large commercial data warehouses like infoUSA, Acxiom and Experian.
- This new blended information is used in predictive algorithms, that estimate how likely someone will vote and how willing is he or she is to support a certain candidate. For this purpose a ten-point scale is calculated to evaluate four dimensions:
    1. The likelihood that an individual would support a candidate. A person with a 7 support score was more likely to back this candidate than anyone with a support score of 6.
    2. The likelihood that a prospect would show up at the polls.
    3. The odds that an inconsistent supporter (voter) could be mobilized to the polls.
    4. How persuadable someone was by a conversation on a particular issue. A predictive model is also used to determine the issue to be discussed.

Based on these predictive models different strategies are developed depending on the scores and characteristics of the specific voter group to be targeted. However, the biggest challenge lies in being capable to design a model that accounted for which personal variables play a consistently predictive role in a particular election.

Campaign teams, like that of Barak Obama, have used these tools to redefine campaign advertising and messages in order to reach and attract specific groups of interest.

For example, according to a report by the New York Times[49]; during the Obama campaign the advertising team identified the sorts of programs that "likely" and "undecided" voters would watch and when; after combining set-top cable detailed information and traditional Nielsen Media Research, strategist were capable to identify the networks and channels this voters would watch. *"This resulted in increased late-night advertising than it otherwise would have on shows like "Late Night With Jimmy Fallon," "Jimmy Kimmel Live," ESPN and most surprisingly, TV Land, the basic cable network devoted to reruns of old programs".*

However, this kind of techniques has brought serious concerns regarding ethics and privacy. According to Zeyned Tufekci from the New York Times[45],, as Social Scientists increasingly understand that much of our decision-making is irrational and emotional, they can influence the "likability" factor of a specific candidate, making these new methods potentially more effective in manipulating people than spreading an specific political offering.

From the privacy point of view, is enough to see the magnitude of data already owned by commercial data warehouses, only Catalyst it is estimated to have about half a petabyte of voter information, and according to their website[24,48]:

"Catalist maintains and constantly updates a complete national (50 States, plus the District of Columbia) database of over 265 million persons (more than 180 million registered voters and 85 million unregistered adults). This sets us apart from other voter file vendors who assemble their database only immediately prior to election cycles or when customer demand for the data is high."

Another key player in this market is Acxiom, according to another recent report by the New York Times[45], this company headquartered in Conway, Ark. Currently owns more than 23,000 servers that collect, collate and analyze consumer data of about 500 million active consumers worldwide, resulting in more than 50 trillion data "transactions" per year. Of these, 190 million individuals are located in the United States (126 million households). Acxiom has a proprietary classification system called PercsonicX, which classifies consumers to one of 70 detailed socioeconomic clusters.

Education

Education at all levels will benefit from Big Data in two ways; predictive models and analytics, these two application areas will allow institutions to deliver more personalized and higher quality education, increasing collaboration between students, teachers, administrators and parents, and resulting in increased satisfaction and improved academic performance.

Predictive Models

Predictive models and specialized adaptive algorithms will allow delivering personalized classroom materials, homework and tests. Students will have access to relevant materials in line with their capabilities, areas of opportunity and interests, allowing them to reach their full potential.

In this context, the use of this new concept in knowledge delivery and evaluation will result in fast feedback in the learning process. Specially developed software as a service solutions will guide and "hint" the students in order to put extra emphasis in their areas of opportunity and reinforce those areas in which the student has proved proficiency.

People learn in different ways, some students approach problem solving step-by-step and analyze material in a linear manner.[49] Predictive models will allow also teachers, administrators and parents to identify the right pedagogic approaches most effective for every particular student.[50] To identify these learning styles will be crucial to be able to deliver personalized education.

As of today, schools in sixteen states employ data science techniques to identify at-risk students. [50] Predictive models have been developed to incorporate academic, social and non-academic variables[49] to predict and evaluate a student engagement with coursework, and take corrective actions in real time.

Analytics

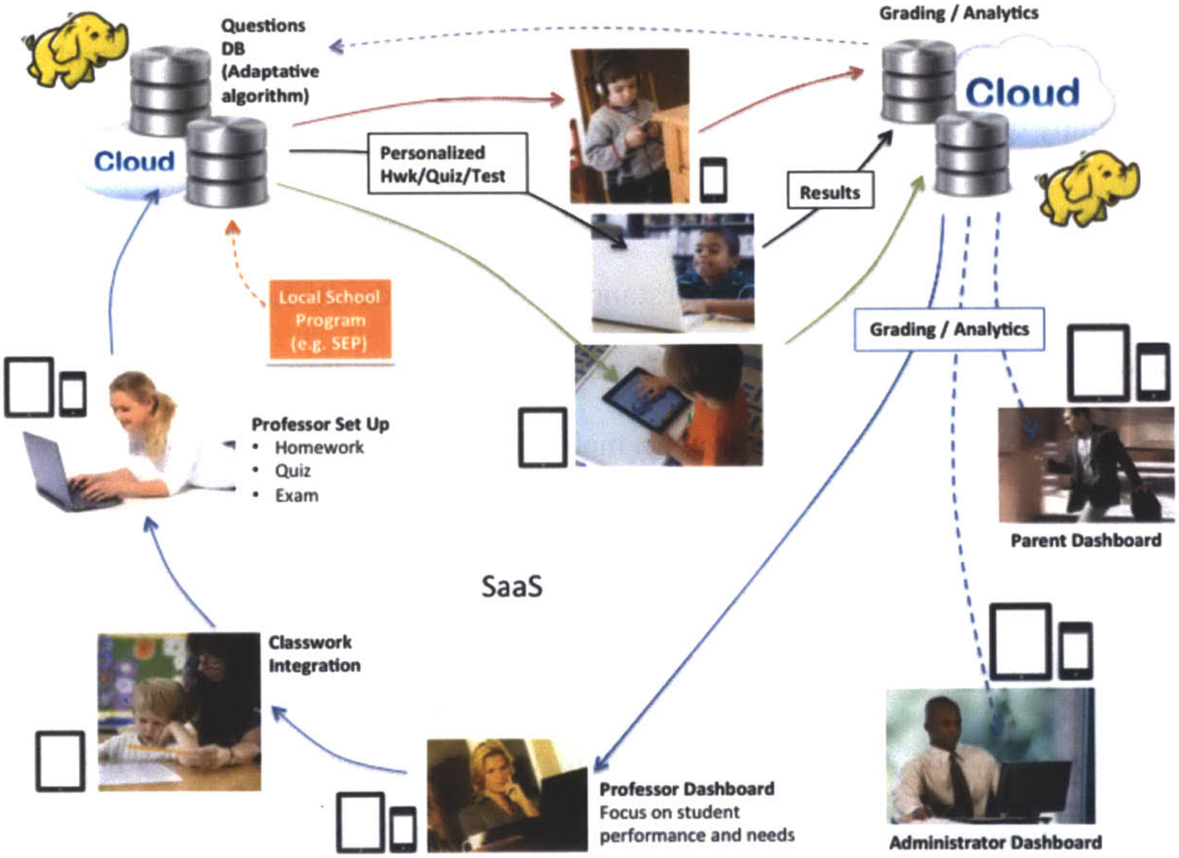
Analytics and visualization tools will allow parents, teachers and administrators to have much more efficient and fast follow of students' performance, and focus on how to support their learning effort.

Apps installed in mobile phones and tablets will allow parents to have real-time insight and a clearer view of the academic performance, strengths and areas of opportunity and areas of opportunity of their children.

Automated homework and exam scoring, will allow teachers to reallocate time from grading to education planning and understanding students educational needs. Dashboards will allow professors to have a much better picture of children evolution over their academic history and the connectivity with mobile devices will allow them to have a much better communication with parents and educators. Educators will also have the possibility to create their own materials and contribute to the overall community with their experience and knowledge.

Administrators will be able to visualize, follow-up and benchmark in real time and at different granularity the levels the performance of schools, from national, state, district, school, grade, group, level all the way to a student file. These enhanced "business intelligence tools" will allow administrators to have a much faster response to support students, schools and faculty.

Academic history will safely reside in the cloud and in case a student moves from one school to another, the transfer of academic records will be as easy as reassigning access rights to the new school. The cloud can help also to solve the challenge of massive storage requirements as student populations incorporate to this new ecosystem. However, privacy concerns may eventually arise. All these tools will be available for students, professors, parents and administrator in a variety of platforms including PC's, tablets, iPods, iPads, and mobile telephones.



Conceptual use of a SaaS platform to enable personalized homework delivery

Challenges

However, several challenges lay ahead before big data can bring such benefits to education. The model will have to be adapted according to the education and specific level in which it wants to applied; for example, children in basic school may still require the high personal contact, interaction and socialization only offered by traditional education models, in this cases technology will be a complement of the class room experience. In other cases, such as college education, MOOCs will have pivotal role in allowing more students to access high level education. A very good example of this is Edx, a non for profit enterprise created by MIT and Harvard to allows students all over the world to access thru the internet and without cost, classes from some of the best graduate programs in the United States

A cultural change from all stakeholders in this process will be essential, as all of them will have to adapt and change their attitudes and behavior towards the use of technology in their corresponding environments. Cost of talent may increase, as teachers and administrators will have to be comfortable with statistics and new virtual environments. It is probable that in the short future, the way teachers are trained will have to evolve at all levels, specially for basic and mid-school, basic knowledge of statistics and analytics will be required. Different technological alternatives are in development by entrepreneurs and governments around the world, this eventually may create issues in the interoperability of systems[50].

Finally, an exceptional user experience will become essential in order to capture students attention in an environment already saturated of messages, HD images and sounds; elements of gamification will have to be incorporated in order to offer a user experience in line with what new generations are used and expect to find in mobile and web applications.

Customer Marketing

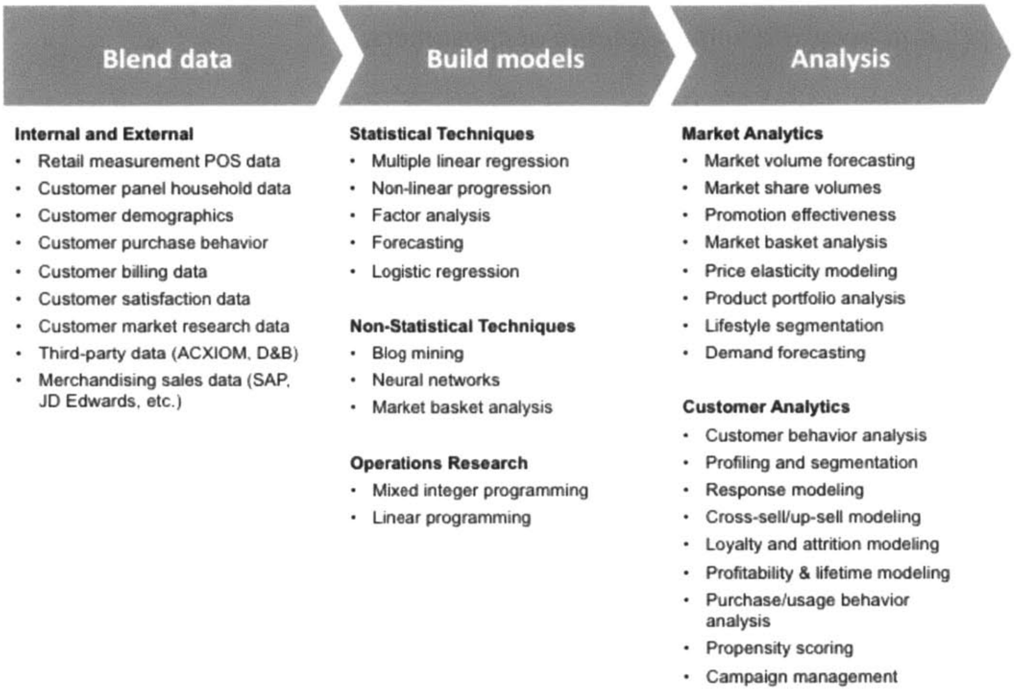
For decades marketers have used data to better understand consumers and fine-tune their campaigns and pitches. Today, every second, an abundance of data streams from point-of-sale transaction, social networks, web traffic, and other digital sources feed the servers of retailers and ecommerce enterprises worldwide. The way this data is distributed, mined, combined, tracked, and connected will play a critical role in the design, promotion and delivery of many of the products and services we enjoy today.

However, big data still has a long way to go in marketing, according to a recent study by Columbia Business School in which 253 corporate marketing decision makers where interviewed, analytics penetration has not yet reached its peak: [54]

- *91% of senior corporate marketers believe that successful brands use customer data to drive marketing decisions.*
- *Yet, 39% say their own company's data is collected too infrequently or not real time enough.*
- *The universal desire to be data-driven is not yet matched by a consistent effort to collect the data necessary to make these real-time decisions. 29% report that their marketing department have too little or no customer/consumer data.*

New start-ups and established technology companies will play a key role to help retailers, manufacturers and service providers to harness these challenges by developing solutions enabling data collection, consolidation, mining, analysis and visualization.

According to Vertica; an important player in the big data infrastructure and consulting space, Big data should allow businesses to improve decision making a product development from two vertices: Market Analytics and Customer Analytics.[56]



Blend data

Internal and External

- Retail measurement POS data
- Customer panel household data
- Customer demographics
- Customer purchase behavior
- Customer billing data
- Customer satisfaction data
- Customer market research data
- Third-party data (ACXIOM, D&B)
- Merchandising sales data (SAP, JD Edwards, etc.)

Build models

Statistical Techniques

- Multiple linear regression
- Non-linear progression
- Factor analysis
- Forecasting
- Logistic regression

Non-Statistical Techniques

- Blog mining
- Neural networks
- Market basket analysis

Operations Research

- Mixed integer programming
- Linear programming

Analysis

Market Analytics

- Market volume forecasting
- Market share volumes
- Promotion effectiveness
- Market basket analysis
- Price elasticity modeling
- Product portfolio analysis
- Lifestyle segmentation
- Demand forecasting

Customer Analytics

- Customer behavior analysis
- Profiling and segmentation
- Response modeling
- Cross-sell/up-sell modeling
- Loyalty and attrition modeling
- Profitability & lifetime modeling
- Purchase/usage behavior analysis
- Propensity scoring
- Campaign management

Market analytics will allow understanding specific dynamic trends and aggregate behaviors such as forecasting, price elasticity or market basket analysis.

For example, in retail and consumer products[57] market analytics will allow retailer to analyze vast quantities of sales transaction data to unearth patterns in user behavior, as well as monitor brand awareness and sentiment with social networking data

Customer analytics, will allow developing more personalized marketing campaigns and offerings.

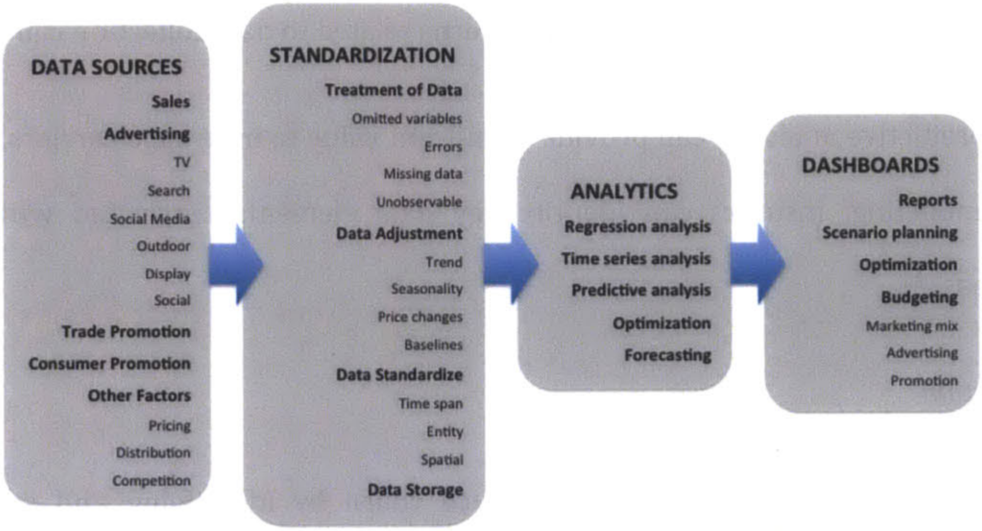
According to Euromonitor[58], consumer analytics provides a quantitative basis for decisions and helps marketing managers develop innovative and winning strategies. With the following benefits:

- *In-depth and accurate understanding of consumers;*
- *Measure and monitor marketing performance;*
- *Develop creative targeting and cross-selling strategies;*
- *Determine at-risk consumer groups and improve loyalty;*
- *Identify emerging consumer satisfaction trends;*
- *Target next generation consumers;*
- *Overcome diversity challenges;*
- *Strategies for different marketing cultures.*

This will translate in better insights on: [58]

- Who to target and why;
- Identifying, key target consumer groups;
- Identifying who is buying, what, when, how and why;
- Quantifying segments and rank targets by attractiveness;
- Understanding spending profiles of target consumers;
- Identifying new opportunities for new product development.

A good example in the way entrepreneurs are creating value for organizations interested in using big data is Marketshare.com, a company that uses advanced econometrics to connect marketing investments with revenue and profit. Using historical data relating to sales and advertising expenditures and information about the brand and its consumers, Marketshare[55] develop quantitative based recommendations of the optimal marketing budget, mix and spend by media vehicle in order to maximize revenue or profit. The figure below exemplifies a generic view how this can be achieved using big data analytic tools.



Insurance

Insurers are gradually recognizing the importance of big data to help them define future strategies and competitiveness. [61] However, there are yet several challenges to overcome. Insurers infrequent interaction with customers, reflects in limited transactional data to work from and potential customer privacy concerns related to data collection can arise.[61]

The use of predictive analytics can provide significant value to insurance carriers, by using predictive modeling; insurers can identify key data elements correlated with risk in multiple areas: [25]

Churn reduction

Predictive models can help insurers to reduce churn by identifying and segmenting customers "at risk" in order to define customer retention actions in advance.[25]

Marketing and Customer Acquisition

Classification models can help insurers to identify customers and segments, which would have a greater response rate for new marketing campaigns. This will also contribute to reduce the Cost of Customer Acquisition by allowing insurers to have a closer look to potential customers and focus better customer marketing efforts and investment. This includes answers questions such as, what is the likelihood this prospect changes insurance carrier? What is the likelihood this prospect buys insurance thru this specific channel? (e.g. online quoting and sales) or what is the likelihood of this prospect to signup with us as carriers?

Fraud detection

Insurers can make use of data mining techniques and historical data (internal and external) related to claims, fraud patterns, accidents, social networks, and medical and criminal records (whenever is possible) in order to identify potentially fraudulent claims.[25]

This will reflect in lower claim related expenses, improved customer satisfaction (reduction in unjustified examinations of legitimate claims), and lower claims-handling expenses.[61]

Claims mitigation

In health insurance, predictive algorithms can be used to analyze medical history, demographic profile, and lifestyle in order to estimate the probability of development of long-term health problems potentially linked previous accidents or health problems. On the basis of this analysis, additional prevention or mitigation treatment can be prescribed.[61]

Pricing

Insurers can use predictive modeling tools to help determine appropriate pricing model for consumers or segments based on profitability, behavior, pricing tolerance or other relevant factors.[25]

Telematics, a growing phenomenon worldwide, has the potential to significantly transform auto insurance, helping insurers to optimize pricing models, capture a greater share of low-risk drivers, cut claims management costs, and enhance customer relationships. [61]

Telematics is based on the idea of the "connected vehicle," in which the use of GPS devices allows transmission of real time data (location, speed, miles driven, location, maintenance, and driving behavior) back to an insurer. [61] This is expected to become the platform for new insurance schemes, such as usage-based insurance, pay-per-use insurance, pay-as-you-drive insurance, pay-how-you-drive. [25]

However, the use of telematics involves challenges such as the need to install after factory devices, the inherent cost of real-time data collection, and the systems and skills required to maintain and to analyze data.[25]

An interesting way the industry is dealing with these challenges is Progressive Snapshot®. a device that plugs into the vehicle and logs key metrics such as: how hard break, mileage, and how often the user drive between midnight and 4 AM.The device doesn't use GPS to track where the drivers go. [59]

After 30 days, Progressive gives incremental discounts to customers who travel less than the average driver in their state. After six months, the driver returns the device and Progressive will lock in a rate that reflects that driver's overall risk profile. [59]

Healthcare

Today, health care is one of the biggest sectors in U.S. economy representing more than 17% of GDP and 11% of the country workforce. However, it is also a major contributor to the high national debt.

According to a study by McKinsey Global Institute[25], to reform the U.S. healthcare system will require to find the right way to reduce the rate at which costs have been increasing while sustaining its current strength. In order to respond to this strategic imperative, healthcare has to improve operational performance and make an even higher adoption of technology to improve processes and costs. [25]

Big data analytics can contribute to achieve these goals by helping to understand population health and trends better. The wide availability of machine-readable data, the low cost related to link data from different sources and the wide and easy availability of computing resources can significantly enable this process.

However, the frequency of data collection remains one of the biggest challenges and the development of new sensors can significantly contribute to solve this problem.

A first example of how sensors can help to improve the speed in data collection is the Health eHeart Study[62], an initiative by the University of California in San Francisco. This study plans to gather more data about heart health from more people than any research study has done before and use this information to develop strategies to prevent and treat all aspects of heart disease. The Health eHeart Study has three key goals: [63]

- *Develop new and more accurate ways to predict heart disease based on measurements, behavior patterns, genetics, and family and medical history.*

- *Understand the causes of heart disease (including heart attack, stroke, heart failure, atrial fibrillation, and diabetes) and find new ways to prevent it.*
- *Create personalized tools to forecast heart disease or, when it might be getting worse.*

The project team plans to combine information from a variety of data sources. Some participants will wear special sensors to monitor heart health. The sensors will link up to a smartphone to gather data on blood pressure, heart rate, ECG, sleep, arrhythmias, activity, and more.[63] Other tools for this study include GPS-enabled phones that can record whether a person is at a fast-food restaurant or in a hospital bed, and a photo app can estimate a meal calories and the served portion size[62].

Blood pressure will be measured several times a day with the aid of a Bluetooth-enabled blood pressure cuff that sends reading to the study database using a smartphone app. [62]

Mobile Apps in smartphones will be also used to apply surveys, get real-life measurements, track behaviors, send medical information, receive reminders, and invite people to the study. [63] This will allow to reduce doctor visits and to empower patients to track their own health and compliance with good life and exercise habits.[62]

With all this information researchers expect to better understand risk factors and propose potential interventions in order to reduce risk.

A second example is Watson, the once unbeatable winner of Jeopardy has now turned for a much nobler career: Support oncologists in Cancer diagnosis and treatment.

In March 2012, the Memorial Sloan-Kettering Cancer Center in New York and IBM announced a collaboration on the development of a decision support tool built upon the natural language capabilities, hypothesis generation, and evidence-based learning of Watson[65]. The goals of this initiative include: [64]

- To provide medical professionals with improved access to current and comprehensive cancer data and practices.
- To define individualized cancer diagnostic and treatment recommendations.

According to IBM, Watson capabilities can be used in diagnosing and treating patients as follows: [64]

- *The physician poses a query to the system, describing symptoms and other related factors.*
- *Watson parses the input to identify the key pieces of information.*
- *Watson then mines the patient data to find relevant facts about family history, current medications and other existing conditions.*
- *Watson combines this information with current findings from tests and instruments and then examines all available data sources to form hypotheses and test them.*
- *Watson can incorporate treatment guidelines, electronic medical record data, doctor's and nurse's notes, research, clinical studies, journal articles, and patient information into the data available for analysis.*
- *Watson provides a list of potential diagnoses along with a score that indicates the level of confidence for each hypothesis.*

Finally, this information is used by the doctor to decide a final diagnostic and treatment for the patient. The ability to take context into account during the hypothesis generation and scoring phases allows Watson to address these complex problems, helping the doctor and patient to make more informed and accurate decisions.

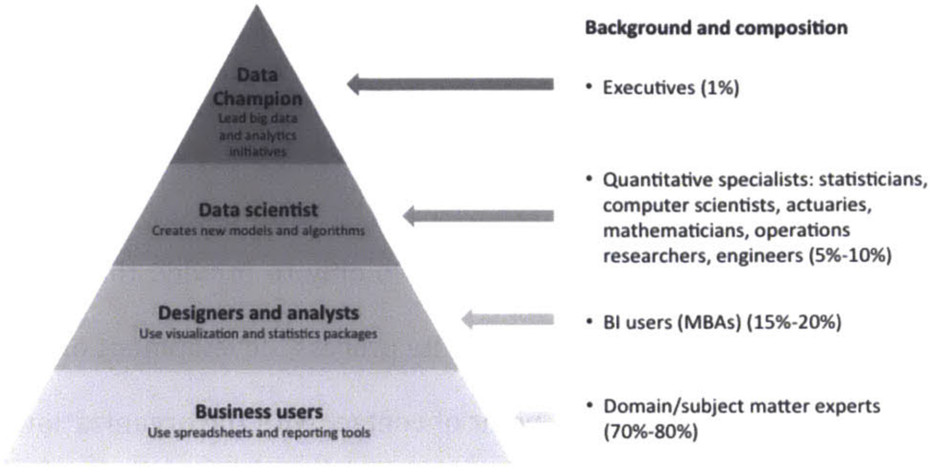
7. Human Capital

As technology continues to develop and organizations will increasingly rely on exploiting data for innovation, customer insight, social analytics or predictive modeling, not only their business processes will evolve but also the concepts of human capital and change management will become crucial for success in this new market reality.

To have the right human capital to respond to the increasing demands in data processing and analysis is perhaps this is one of the most important aspects in value creation derived from Big Data. Managers will have to become highly "literate" in analyzing and take data supported decisions, representing great challenges for senior executives. According to Andrew McAfee and Erik Brynjolfsson from *MIT "a new culture of decision making will arise impacting executive roles in two ways: The mutation of the HiPPOs (The Highly Paid Person Opinion) and the development of new roles; in this new culture of decision making executives will fully support decision making on data and analytics."*

Despite it has always requirements for professionals with advance analytical skills to support the organization's capabilities; the requirements for personnel are different with Big Data[7]. Because interacting with the data itself – obtaining, extracting, manipulating, and structuring – is critical to any analysis, the people who work with big data need substantial and creative IT skills[7]. This includes identifying the role and responsibilities not only of the Data Scientists but also many other professionals involved in the data value chain. According to Deloitte[33], less than 10% of executives will require a data scientist background and less that 1% of executives will require the necessary skills to lead big data and analytics initiatives.

Below we may find the different levels of data literacy as recently proposed by Deloitte:


Data Champion
Lead big data and analytics initiatives
Data scientist
Creates new models and algorithms
Designers and analysts
Use visualization and statistics packages
Business users
Use spreadsheets and reporting tools
Background and composition
• Executives (1%)
• Quantitative specialists: statisticians, computer scientists, actuaries, mathematicians, operations researchers, engineers (5%-10%)
• BI users (MBAs) (15%-20%)
• Domain/subject matter experts (70%-80%)

Source: Deloitte[33]

People with developed analytical skills and education (such as those with MBAs) will be key users of business intelligence will comprise 15% to 20% of data and business intelligence users. Subject matter experts will comprise the other 70% to 80% of users.

The Big Data initiative team

The overall project from the data collection process to the data mining activities and analysis can combine a variety of technologies[25]. This means that enterprises will need to create a multi-skilled team, including data consumers, within the business, analytics professionals, motivated IT staff with the potential to learn big data approaches, and external consultants. [25]

According to Deloitte big data teams will require at least six complementary skills: [33]

Industry Sector[33]

Professionals with extensive experience and a deep knowledge of the industry or sector in which the organization wants to play, for example: Technology, Media and entertainment, Telecommunications.

Those involved as industry experts should expect no only to provide their knowledge to identify sources of information, or contribute to the process development or redesign, but recognize that they will become the main point of contact with the organization and active users of the information and analytics finally produced by these initiatives, requiring them to become highly skilled in understanding the underlying statistical analysis performed in the process, and interpreting and communicating the results to the rest of the organization.

Business Process Domain[33]

Professional in this domain should master specific business process skills in which it is planned to implement the big data initiative, for example: Risk management, Finance, Customer service, marketing, supply chain.

This will require to incorporate team members who do not necessarily are highly skilled in technology or big data analysis, but who understand the core business requirements or the specific business process. As in the case of the industry sector representatives they will eventually become users of the information and analytics derived from the initiative, and specific training and change management actions should be taken in order to make them comfortable with their new data driven processes.

Technology[33]

In this area we will require professionals with the right skills in programming, infrastructure management and support, distributed systems, cloud management and systems integration. Professionals with the right skills in data management and Hadoop integration will be essential.

Design[33]

Graphic Design skills will be required in order to develop the right user experience, user interphase, and dashboard design.

This is probably one of the most important aspects of a big data development team, as designers should collaborate with subject matter experts and business process domain experts to make sure the content and visualization techniques used in the project are aligned with the business objectives and the business case they are trying to solve.

Startups like gooddata.com are successfully contributing to this space by focusing their efforts in developing intuitive data visualization and reporting tools in order to support decision making with clarity of information.

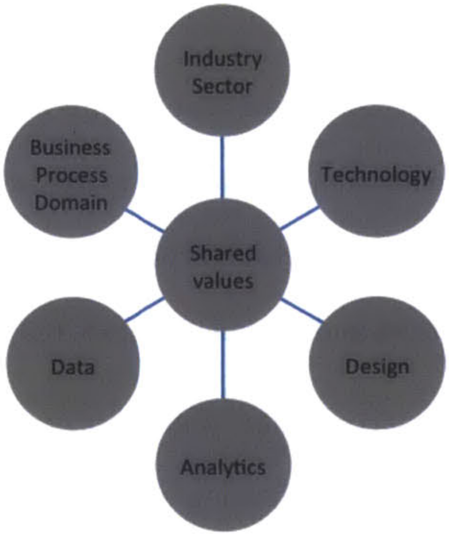
Analytics[33]

Analytics talent will allow developing the right business intelligence and reporting content, to perform advance analytics -such as regression, time-series, classification, clustering, optimization- and graphical and text mining. This position will require professionals with data science skills.

Data[33]

Data management talent will allow developing: Data architecture, data modeling, data extraction, transformation and loading and the right data governance and quality control.

According to my own professional experience, this kind of team structure is not new in technology projects; however, it is key to keep in mind the importance that different from traditional projects, the development and permanence of data analysis and understanding skills



Source: Deloitte[33]

The mutation of the HiPPOs

According to McAfee and Brynjolfsson, a critical aspect of big data is its impact on how decisions are made and who gets to make them. *"When data are scarce, expensive to obtain, or not available in digital form, it makes sense to let well-placed people make decisions."*[28]

McAfee and Brynjolfsson, sustain that today many organizations trust decision making on seasoned executives who rely on their experience and (when available) limited data, their opinions are known as HiPPO – the highest-paid person opinion[28], and for years, they have been the essence of executive decision making. However, the evolution to a big data culture will have a significant impact in the way executive and operational teams support and perform decision-making. McAfee and Brynjolfsson suggest that executives interested in leading big data transitions in their organizations should start applying two simple techniques.

- *First - get in the habit of asking, "What the data say?" when faced with an important decision and following up with more-specific questions such as "where did the data come from?" "What kind of analyses were conducted?" and "how confident are the results?"*[28]
- *Second - allow themselves to be overruled by the data; few things are more powerful for changing a decision-making culture than seeing a senior executive concede to data driven decision-making.* [28]

However, is key not to lose sight that domain expertise will remain key in order to identify what are the right questions to ask to identify the right sources of right information and to anticipate potential scenarios consequence of data driven decision making.

The Data Scientist

In 2008 D.J. Patil and Jeff Hammenbacher coined the term "data scientist," and defined it as a high-ranking professional with the training and curiosity to make discoveries in the world of big data.

The data scientist role is gaining importance for organizations looking to extract insight from "big dataset." However, these initiatives requires a broad combination of skills that may be fulfilled as a team[25]:

- Collaboration and teamwork is required for working with business stakeholders to understand business issues[25].
- Analytical and decision modeling skills are required for discovering relationships within data and detecting patterns[25].
- Data management skills are required to built the relevant dataset used for analytics[25]

Today thousands of data scientists are already working in all kinds of organizations; from both start-ups to well-established companies, and their mission is to contribute to maximize value extraction from information that comes in varieties and volumes never seen before[29].

Big data initiatives for most organizations are still in the exploratory phase[25]. There are two main factors slowing down organizations[25]:

- The ability to articulate a business case justifying the investment is difficult, as the business doesn't always perceive the benefits of such initiatives[25].

- The lack of skills, both internally and externally makes it difficult to get started[25].

Technologies such as the Hadoop framework (the most widely used framework for distributed file system processing), cloud computing and data visualization tools have allowed people with the right skill-set to cope with the technical challenges processing multiple petabytes of data in forms other than rows and columns of numbers, and give answer to questions involving several analytical efforts[29].

The role of the Data Scientist is to effectively integrate the different skills, tools and methodologies available in order to make the best use of the data assets of the organization, this skills and tools include:

  - Mathematics and Operations research: C++, C#, Phyton
  - Statisticians/ analysts: SAS Base, SAS EM

Talent supply

According to McKinsey Global Institute (MGI) a significant[30] constraint on realizing the value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data[30].

According to MGI[30], by 2018 only in the United States would be a demand of 440-490 thousand professionals with the data analysis right; resulting in about 140-190 thousand positions not filled, this type of talent is difficult to produce, and normally takes years of training to develop someone with intrinsic mathematical abilities[30].

In addition, MGI estimates a need for 1.5 million additional managers and analysts in the United States that can ask the right questions and consume the results of the analysis of bid data effectively.[30] This will require retraining a significant amount of the talent in place; fortunately, this level of training does not requires years of study.[30]

However, governments can significantly contribute to increase the supply of talent necessary to leverage big data in different ways:

- *Governments can put in place initiatives to increase the pipeline of graduates with the right skills, this can be done by supporting programs focused on science, technology, engineering, mathematics and engineering, including advanced training in statistics and machine learning[30]. This should include incentives to train existing managers and analysts in analytics[30].*
- *To reduce barriers to access these talent pools in other regions, this can be trough international outsourcing, remote work or the immigration of skilled workers[30].*

In a recent report, the Wall Street Journal[35] reports that training of data scientists has not caught up with demand resulting in that technology workers with big data and analytics skills are paid on average 11.5% more than professionals without those skills.

8. Infrastructure and Solutions

As enterprises come under more pressure from the business to provide access to and analysis of increasingly complex information sets, vendors continue to attach the term "big data" to an ever widening set of products[25].

According to Gartner[25] some of the factors influencing this focus on big data include:

- Innovative processing approaches have emerged that leverage low-cost servers/CPUs and an increasingly robust, wide array of open source of freeware technology, thereby fundamentally changing the cost benefit equation[25].
- Increased availability of scalable elastic resources in the cloud have allowed organizations begin big data project without investing in infrastructure[6].

To understand the complex world of suppliers and solutions in the big data space will be paramount for any organization aiming for a successful Big Data strategy. According to a recent study by Gartner,[66] IT spending driven by Big Data will total $28 billion in 2012 and may reach $55 billion by 2016. According to this same report, Big Data requirements in some cases is forcing the early retirement of solutions and impacts investments mostly on social media, social network analysis and content analytics (up to 45% of new spending pr year). [66]

According to Gartner[66], by 2018, big data solution will cease representing a competitive advantage for organizations; this creates a strategic imperative with three dimensions:

- Engage early in Big Data and Analytics initiatives.
- Identify and pilot business cases in line with the organization strategy and goals.

- Develop the right human resources at all levels of the organization.

Organizations will have to develop technical and human resources that allow them to be flexible enough to respond to continuous market conditions and requirements; this response has to be leveraged by solid data supported decision-making.

In the following pages we will make a review on the way big data can contribute to improve productivity, then we will review a general framework for big data initiatives in any enterprise, the idea is to review a general methodology that allows to cover the key activities and challenges regarding big data, including data governance and storage. Finally we will make a review of a list of selected vendors and solutions available in this increasingly competitive space.

Impact of Big Data on Productivity

IT has always considered and important contributor to improve productivity, and according to McKinsey Global Institute (MGI) there have been four waves of IT adoption with different degrees of impact on productivity growth in the United States: [11]

- *The first era (1959 to 1973).* The "mainframe era," during this period IT's contribution to productivity was rather modest, (annual US productivity growth overall was very high at 2.82 percent); at that stage, IT's share of overall capital expenditure was relatively low. [11]
- *The second era (1973 to 1995).* The "minicomputers and PCs era," Companies began to boost their spending on more distributed types of computers, and these computers became more powerful as their quality increased. However, experienced much lower

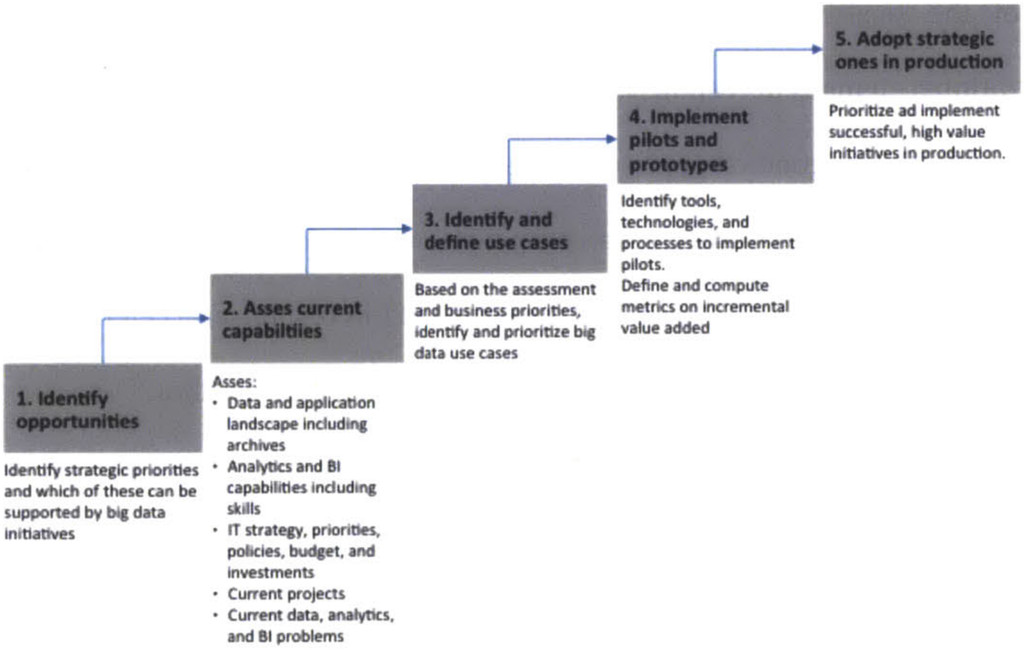
growth in overall productivity, but we can attribute a greater share of that growth to the impact of IT. [11]

- *The third era (1995 to 2000).* The era of the "Internet and Web 1.0." In this period, US productivity growth returned to high rates, underpinned by significant IT capital deepening, an intensification of improvements in quality, and also the diffusion of significant managerial innovations that took advantage of previous IT capital investment. McKinsey suggests, there was a lag between IT investment and the managerial innovation necessary to accelerate productivity growth. During this period, most of the positive impact on productivity in IT came from managerial and organizational changes in response to investments in previous eras. [11]
- According to MGI, a good example of this is Wal-Mart's implementation of IT-intensive business processes; this allowed the company to outperform competitors in the retail industry. Eventually those competitors invested in IT in response, accelerated their own productivity growth, and boosted the productivity of the entire sector. [11]
- The fourth and final era (2000 to 2006), The period of the "mobile devices and Web 2.0." During this period, the contribution from IT capital deepening and that from IT-producing sectors dropped. However, the contribution from managerial innovations increased—again, this wave of organizational change looks like a lagged response to the investments in Internet and Web 1.0 from the preceding five years. [11]

What do these previous eras say about the potential impact of big data on productivity? Like previous eras, obtaining the benefits derived from big data will require not only significant IT investments but also business process and managerial innovation. These will allow not only to boost productivity but also to revamp innovation at all organizational levels.

Big Data Roadmap

According to a recent publication by Deloitte[56] there are 5 key steps that any big data roadmap has to include in order to develop robust solutions that can respond to business environment, and allow organizations to develop solutions that can increasingly develop the necessary skills to succeed in the use of information as an strategic competitive asset.



1. Identify opportunities

Identify strategic priorities and which of these can be supported by big data initiatives

2. Asses current capabiltiies

Asses:
- Data and application landscape including archives
- Analytics and BI capabilities including skills
- IT strategy, priorities, policies, budget, and investments
- Current projects
- Current data, analytics, and BI problems

3. Identify and define use cases

Based on the assessment and business priorities, identify and prioritize big data use cases

4. Implement pilots and prototypes

Identify tools, technologies, and processes to implement pilots.
Define and compute metrics on incremental value added

5. Adopt strategic ones in production

Prioritize ad implement successful, high value initiatives in production.

Source: Deloitte[56]

- *Identify opportunities.* This step requires identifying strategic priorities of the organization that can be supported by big data and analytics.
- *Assess current capabilities.* This will require to evaluate the analytics and BI technical capabilities, skills, and current projects in order to identify potential gaps in Data and BI strategy and priorities.

- *Identify use cases.* Based on the analysis and understanding of short and long term business strategy and priorities, identify the key big data use cases for the organization.
- *Implement pilots and prototypes.* This will allow the organization not only to develop hands on experience but also to identify the tools and processes required to successfully implement business cases. This will also allow defining metrics that allow measuring potential improvements on productivity in the organization.
- *Adopt strategic ones in production.* Once it has been confirmed the value of big data initiatives for the organization; successful business cases should be prioritized and moved rolled out in the rest of the organization.

Big data technology drivers

According to Deloitte[56] in order to implement successful big data initiatives it is also key o define clear technology drives that respond to the specific needs of every organization. This will allow defining clear criteria to be used in the technology selection process:

- *Scalable and efficient.* Identify technologies that can operate at peta-byte scale with daily or even real time uploads.
- Schema-*agnostic.* Take in consideration the multi-structured, dyamic and flexible nature of big data.
- *Fault tolerant and reliable.* Systems need the ability to recover from failure to assure reliability of processing and data integrity.
- *Cost.* The cost associated with storing and processing data should be justify by robust cost-benefit analysis, this includes the implementation and maintenance of infrastructure

- *Integration and Security.* Integration with internal sources sources and systems in a secure manner.

Data processing

The diverse nature of big data will require a new approach to data processing selection, depending on data variety (structured, semi-structured, unstructured) and processing speed requirements (batch, near real-time, real time) architecture choices should be considered.

		Turnaround time / Processing Speed			
		Batch		**Near Real Time – Real Time**	
Variety	**Structured**	Traditional	MPP	MPP	MPP + in-memory
	Semi-structured	Traditional + MPP	Distributed Cluster	Specialized MPP + Distributed Cluster	
	Unstructured	Distributed Cluster		Specialized System	
		Low → *Volume* → High			
		Lower → *Cost* → Higher			

Source: Deloitte

Massive Parallel Processing (MPP)[67]. In a parallel processing configuration, the workload for each job is distributed across several processors on one or more computers, called compute nodes. These nodes work concurrently to complete each job quickly and efficiently.

In-memory processing [69]. Data is loaded into Random Access Memory (RAM) instead of hard disks. The source database is queried only once instead of accessing the database every time a query is run thereby eliminating repetitive processing.

Distributed cluster [68]. Consists of a cluster of connected computers that work together as a single system. The components of this cluster are usually connected to each other, with each node running its own instance of an operating system.

Big data governance

In the same line of data processing, the diverse nature of big data will require data governance to play an important role in its management. If implemented early on and in the right organizational process, enterprises will reap rewards and reduce any associated risks. There are seven key data governance concepts to take in consideration in any big data initiative:

- *Stewardship.* User have to take stewardship of data is available for their needs. From inception to delivery, they know better than anyone else the data that best respond to their needs. [70]
- *Information governance.* Architects, system administrators, and data center administrators are critical stakeholders in information governance; this includes landing, processing requirements, infrastructure management and security of big data.[70]
- *Data definition and usage standards.* Data definition and how will be consumed, will require clear standards and structure on how to format and process information for

every business are that will use this information. This may include Master Data Management and metadata management programs.[70]

- *Master data management.* Big data will be consumer of MDM outputs. MDM can be used to parse the content and provide it to analytic applications.[70]
- *Metadata management.* The appropriate context for metadata processing has to be defined, this include taxonomies, semantic libraries and anthologies.[70]
- *Data lifecycle management.* Enterprises and stakeholder should define the right the length of time data is required in the system define the right storage of data after consumption.[70]
- *Risk and cost containment.* It is key that organizations understand risk associated with big data initiatives and the potential hidden costs associated. The implementation of robust infrastructure governance, data processing and consumption will contribute to mitigate these risks. [70]

Big data solutions landscape

Despite Big data is a market still in development, solutions can be grouped on four main groups: Infrastructure, Analytics, Applications and Data Sources. Each of these groups is focused in responding

Infrastructure	Analytics	Applications	Data Sources
• Hadoop • NoSQL Databases • SQL Databases • MPP Databases • Management/Monitoring • Crowdsourcing • Cluster Services • Security • Collection/Transport	• Analytics Solutions • Data Visualization • Statistical computing • Social Media • Sentiment Analysis • Analytics Services • Location/People/Events • Big Data Search • IT Analytics • Real-time • Crowd sourced analytics • SMB Analytics	• Ad optimization • Publisher tools • Marketing • Industry Applications • Application Service Providers	• Data Marketplace • Data sources • Personal Data

Infrastructure

Big data infrastructure comprises the integration of several solutions habilitating the deployment of solutions; this is probably one of the most competitive markets in the big data ecosystem with several players in each specialty. There are nine key application areas in infrastructure as described in the table below, this same table presents key suppliers in each application.

Solution	Key Vendors
Hadoop	Cloudera, Talend, Hadapt, Nortonworks, Infochimps, MAP, Hstream, ZettaSet, IBM, Mortar, Microsoft, Greenplum, Amazon, QBole, Sqrl
NoSQL Databases	10Gen, Datastax, Basho, Cloudant, Couchbase, Hypertable, Neo4j, Sones
New SQL Databases	Marklogic, Paradigm4, Drawnscale, Memsql, SQLFire, VoltDB, NuoDB,
MPP Databases	Vertica, Kognitio, Paraccel, Greenplum, Teradata, Netezza, InfiniDB, Microsoft SQL Server
Management/Monitoring	OuterThought, OceanSync, StackIQ, Datadog
Crowdsourcing	Crowdcomputing Systems, CrowdFlower, Amazon, Mechanicalturk
Cluster Services	LexisNexis, HPCC Systems, Acunu
Security	Stormpath, Imperva, TraceVector, Codefortytwo, Dataguise
Collection/Transport	Aspera, Nodeable

Analytics

Analytics is comprised 12 key applications segments, in this space are considered all vendors providing analytics and predictive models solutions; from visualization all the way to localization analysis. It is interesting to note the high intensity of competition in the analytics solutions and data visualization segments.

Solution	Key Vendors
Analytics Solutions	Palantir, Platfora, Pervasive, Datameer, Karmaspere, Datahero, Digitalreasoning, Datasphora, Precog
Data Visualization	Quid, Gooddata, Visua.ly, Actuate, Kitenga, Centrifuge, Ayasdi, Clearstory, Tableau, ISS, Quantum4D
Statistical computing	Skytree, PriorKnowledge, Revolution Analytics, Matlab, SAS, SPSS
Social Media	Bit.ly, Bluefin, Tracx, Simplereach, Dataminr
Sentiment Analysis	CrimsonHexagon, General Sentiment
Analytics Services	ThinkBig, MuSigma, McKinsey & Co., Accenture, Opera
Location/People/Events	Rapleaf, Fliptop, Recorded Future, Place IQ, Radius
Big Data Search	Elastic Search, Autonomy
IT Analytics	Splunk, Sumologic
Real-time	Continuity, ParStream, Feedzai
Crowd sourced analytics	Dataking, Kaggle,
SMB Analytics	Sumall, RJMetrics, Custora

Applications

In the applications space we can find 5 key segment, three of them (Ad Optimization, Publisher tools and Marketing) focused on Marketing and the other two segment to industry applications and application system, where we have identified only one player Collective9 a SaaS company focused on raw data treatment.

Solution	Key Vendors
Ad optimization	DataXu, Aggregate Knowledge, m6d, aiMatch, BlueKai, RocketFuel, The Trade Desk, Turn, 33Across
Publisher tools	Visual, Yieldex,
Marketing	LatticeEngines, Sailthru, Retention Science, BloomReach, ClickFox
Industry Applications	Next Big Sound, Knewton, ZestCash, Wonga, NumberFire, Mile Sense, Bill$ Guard, Climate Solutions, Bloomberg
Application Service Providers	Collective9

Data Sources

Data sources have three main applications segments of which the data marketplace is of special interest as companies such as Factual and DataMarket are actually betting high on aggregating and commercializing significant amounts of data.

Solution	Key Vendors
Data Marketplace	Factual, DataMarket, Windows Azure Marketplace
Data sources	Premise, DataSift, Knoema, GNIP, SpaceCurve
Personal Data	Withings, Jawbone, Runkeeper, Basis, Nike, Fitbit

For example, Factual is a location platform that enables development of personalized, contextually relevant experiences in a mobile world. Factual builds and maintains data on a global scale covering over 65 million local businesses and points of interest in 50 countries.

Infrastructure/Analytics

Several vendors offer infracstructure/analysis cloud-based solutions including Google (BigQuery), IBM (BigTable) and 1010data is the leading provider of Cloud-based analytics for Big Data.

The table below presents some key vendors in this space:

Key Vendors
SAP,Sas, IBM, Google, Oracle, Microsoft, VMware, Amazon, 1010data, Metamarkets, Teradata, Autonomy, NetApp

Case Study: Google Big Query

A relevant player in this field is Google with Big Query, a solution that allows running SQL-like queries against very large datasets. Big Query has been built using Google's Dremel technology. According to Google[52], BigQuery works best for the analysis of datasets, typically using a small number of very large, append-only tables. For more traditional relational database scenarios, Google also offer as Cloud SQL solution.

This application is currently used in applications such as game and social media analysis, infrastructure monitoring, advertising campaign optimization, or sensor data analysis.

BigQuery can be used through a web UI called the BigQuery browser tool, the bq command-line tool, or by making calls to the REST API using various client libraries in multiple languages, such as Java, Python, etc. [71] BigQuery offers the following features: [71]

- Speed - Analyze billions of rows in seconds.
- Scale - Terabytes of data, trillions of records.
- Simplicity - SQL-like query language, hosted on Google infrastructure.
- Sharing - Powerful group- and user-based permissions using Google accounts.
- Security - Secure SSL access.
- Multiple access methods - Connect to BigQuery using the BigQuery browser, the bq command-line tool, the REST API, or Google Apps Script.

BigQuery is an OLAP (online analytical processing) system. So, is good for analyzing data, but not for modifying it.

9. Conclusions

As have all over this document the use of big data and analytics will have a transformative effect not only on businesses but also in our every day life. I have no doubt that virtually all fields of human activity will be touched by this revolution.

New technologies are not only allowing us to collect data in the most unimaginable ways but also to constantly collect almost every byte of knowledge and activity on real time, all the time. In a few years, what we call big data today tomorrow will grow further in complexity and volume and probably will be just called "data" again.

We are just seeing the tip of iceberg that will change the fundamentals of all human activity, decision maker will evolve from HiPPOs to data-driven HiIPOs (Highest Informed Person Opinion), probably this will be the same people who was taking decisions in the past, but their rich experience and knowledge will be complemented with strong analytical tools. On this respect, the biggest challenge will be to give this people the elements to evolve and adapt to this new information reality, fortunately universities such as MIT, NYU, Northwestern or Carnegie Mellon, are taking the right direction by strengthening their program offering in analytics and predictive models in order to provide highly skilled professionals.

The 3V's of Big Data Volume, Variety and Velocity will constantly grow in complexity, data will arrive faster in much more complex form and in increasingly higher volumes. However, an increasing amount of "perishable" data, which not only will be collected and processed at "light-speed" but also, will lose relevance much faster than before. Much of this data will be useful in some cases just a few seconds or even less.

Sensor will keep evolving and probably will become the most important form of data collection as do not require any kind of specific interaction and are already embedded in multiple devices used today, this will be further fueled by the advent of "the internet of things." However, our privacy will be increasingly compromised if we do not find good formulas to incorporate this new omnipresence form of data collection to our everyday life. As we could see in the review of the use of big data in politics, we also risk the use of this information to manipulate instead of supporting good decision-making.

Analysis tools combined with new and faster visualization tools will give early competitive advantage to those organizations capable to embrace big data. However, today these powerful tools seem to be mostly available to big organizations, the next step will be to develop solutions that allow medium and small size enterprises to enter this revolution and grow faster. They surely do not have datasets comparable to those of Wal-Mart, Amazon or the United States Government, however, if the entrepreneurial community takes the leadership on making these tools available to smaller enterprises, the collection, analysis and visualization of data can be of great use for medium and small enterprises. Additionally, the aggregated value of analyzing these data sets represents a huge opportunity to entrepreneurs and businesses. These data assets do have not only commercial value but also will allow a better understanding how our society evolves and responds to innovation and events.

As we have seen in the industry review of this volume, every year new startups will appear offering more sophisticated solutions for data analysis and predictive modeling; the availability of cloud computing is not only reducing the barrier to entry to this market but also will enable to offer products and services never seen before.

Entrepreneurs are taking a vital role in the evolution and wider use of big data tools. I find in MIT the perfect ecosystem to foster the development of these entrepreneurs, initiatives such as the 100K competition, the Martin Trust Entrepreneurship Center and the MIT Media lab are just tokens on the way MIT is actively contributing to stimulate new generations to create the leading data-driven organizations of tomorrow.

10. References

1. Mike Wheatley 2012, *Five Biggest Milestones In The History of Big Data,* Silicon Angle, <http://siliconangle.com/blog/2012/10/09/five-biggest-milestones-in-the-history-of-big-data/>
2. Wikipedia 2013, *1890, United States Census,* Wikipedia <http://en.wikipedia.org/wiki/1890_United_States_Census>
3. Leonard C. Bruno, 2012, *Plate, punch card, and instructions for Herman Hollerith's Electric Sorting and Tabulating Machine, ca. 1895.* Library of Congress < http://memory.loc.gov/cgi-bin/query/r?ammem/mcc,gottscho,detr,nfor,wpa,aap,cwar,bbpix,cowellbib,calbkbib,consrvbib,bdsbib,dag,fsaall,gmd,pan,vv,presp,varstg,suffrg,nawbib,horyd,wtc,toddbib,mgw,ncr,ngp,musdibib,hlaw,papr,lhbumbib,rbpebib,lbcoll,alad,hh,aaodyssey,magbell,bbc,dcm,raelbib,runyon,dukesm,lomaxbib,mtj,gottlieb,aep,qlt,coolbib,fpnas,aasm,denn,relpet,amss,aaeo,mff,afc911bib,mjm,mnwp,rbcmillerbib,molden,ww2map,mfdipbib,afcnyebib,klpmap,hawp,omhbib,rbaapcbib,mal,ncpsbib,ncpm,lhbprbib,ftvbib,afcreed,aipn,cwband,flwpabib,wpapos,cmns,psbib,pin,coplandbib,cola,tccc,curt,mharendt,lhbcbbib,eaa,haybib,mesnbib,fine,cwnyhs,svybib,mmorse,afcwwgbib,mymhiwebib,uncall,afcwip,mtaft,manz,llstbib,fawbib,berl,fmuever,cdn,upboverbib,mussm,cic,afcpearl,awh,awhbib,sgp,wright,lhbtnbib,afcesnbib,hurstonbib,mreynoldsbib,spaldingbib,sgproto,scsmbib,afccalbib,mamcol,:@field(SUBJ+@band(+Hollerith,+Herman++1860+1929++))>
4. Uri Friedman, 2012, *Big Data: A Short History,* Foreign Policy <http://www.foreignpolicy.com/articles/2012/10/08/big_data?print=yes&wp_login_redirect=0>
5. The Official Website of the U.S. Social Security Administration 2012, *The Social Insurance Movement,* <http://www.ssa.gov/history/briefhistory3.html>
6. Luther A. Huston, December 27, 1936, *Huge Machine Busy on Social Security,* The New York Times. <http://query.nytimes.com/gst/abstract.html?res=F50C11FE385D13728DDDAE0A94DA415B868FF1D3#>
7. The Official Website of the U.S. Social Security Administration, 2012, *Monthly Statistical Snapshot, March 2013* <http://www.ssa.gov/policy/docs/quickfacts/stat_snapshot/>
8. Wikipedia, 2013, *Colossus computer* <http://en.wikipedia.org/wiki/Colossus_computer>
9. Wikipedia, 2013, *Fremont Rider* <http://en.wikipedia.org/wiki/Fremont_Rider>
10. Gil Press, June-September 2012, *A Very Short History of Big Data*, What's the Big Data? <http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/>
11. Joe McKendrick, June 2012, *The Federal Government's First Foray into Cloud Computing, Circa 1965*, Forbes < http://www.forbes.com/sites/joemckendrick/2012/06/29/the-federal-governments-first-foray-into-cloud-computing-circa-1966/>
12. Edgar S. Dunn, Jr., Feb. 1967, *The Idea of a National Data Center and the Issue of Personal Privacy*, The American Statistician, <http://www.jstor.org/stable/2681910>
13. *Data Mining and Knowledge Discovery,* 1, 5–10 (1997) 1997 Kluwer Academic Publishers, Boston.
14. Wikipedia, 2013, *Apache Hadoop*, <http://en.wikipedia.org/wiki/Apache_Hadoop#History>

15. IBM, 2013, *What is Hadoop?* <http://www-01.ibm.com/software/data/infosphere/hadoop/>

16. Wikipedia, 2013, *Watson (Computer)* <http://en.wikipedia.org/wiki/Watson_(computer)>

17. IBM, 2013, *The Science Behind Watson* <http://www-03.ibm.com/innovation/us/watson/science-behind_watson.shtml>

18. Gil Press, April 2012, *A Very Short History of Data Science*, What's the Big Data? <http://whatsthebigdata.com/2012/04/26/a-very-short-history-of-data-science/>

19. Mark van Rijmenam, January 2013, *Big Data History* < http://www.bigdata-startups.com/big-data-history/>

20. IBM, 2013, *What is Big Data?* < http://www-01.ibm.com/software/data/bigdata/>

21. Gartner Research, July 2012 *The importance of "Big Data": A definition*

22. C. Eaton, D. Deroos, T. Deusch, G. Lapiz, P. Zikopoulos *, 2012, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, The McGraw-Hill Companies < http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF >

23. Josh James, June 2012, *How Much Data is Created Every Minute?* Domo <http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/>

24. C. Eaton, D. Deroos, T. Deusch, G. Lapiz, P. Zikopoulos *, 2012, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, The McGraw-Hill Companies < http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF >

25. Gartner Research, July 2012, *Hype Cycle for Big Data, 2012*

26. Th. Davenport, P. Barth, R. Bean, July 2012, *How 'Big Data' is different*, Sloan Management Review. Pp 43-46

27. Oracle, 2012, *Meeting the Challenge of Big Data* *<http://www.oracle.com/us/technologies/big-data/big-data-ebook-1866219.pdf>*

28. A. McAffe, E. Brynjolfsson *Big Data: The management revolution*, Harvard Business Review, October 2012, pp 61-68

29. Th. Davenport, D.J. Patil *Data Scientist: The sexiest job of the 21st century.* Harvard Business Review, October 2012, pp 70-76

30. McKinsey Global Institute, 2011, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey & Co.

31. David Newman, *Innovation Insight: Linked Data Drives Innovation Through Information-Sharing Network Effects* Gartner, 15 December 2011

32. Deloitte Development LLP, 2012, *Big data matters- except when it doesn't*

33. D.Steier, D. Mathias, November 27, 2012, *The Backstory on Big Data: What TMT Executives Should Know"* Deloitte

34. Lecture Notes *15.062 / ESD.754J Data Mining: Finding the Data and Models that Create Value*, R. Welsch MIT Fall 2012

35. S. Rosenbush, M. Totty, March 11, 201, *How Big Data is Changing the Equation For Business*, The Wall Street Journal. <http://online.wsj.com/article/SB10001424127887324178904578340071261396666.html>

36. G. Shmueli, N. Patel, P. Bruce, 2011, *Data Mining for Business Intelligence*, Wiley

37. SAS Institute Inc., 2012, *SAS Enterprise Miner: SEMMA* <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>

38. SAS Institute Inc., 2012, *Five Big Data Challenges* <http://www.sas.com/resources/asset/106008_5BigData_FINAL.pdf>

39. Oracle, 2012, *Fusion Middleware Fusion Developer's Guide for Oracle Application Development Framework 11g Release 1 (11.1.1): Creating Databound ADF Data Visualization Components* <http://docs.oracle.com/cd/E12839_01/web.1111/b31974/graphs_charts.htm>

40. Ari Schorr, April 2011, *Public preview of project codename "GeoFlow" for Excel delivers 3D data visualization and storytelling,* Microsoft <http://blogs.office.com/b/microsoft-excel/archive/2013/04/11/public-preview-of-geoflow-for-excel-delivers-3d-data-visualization-and-storytelling.aspx>

41. Fusion Charts, 2011, *How do I create a Heat Map Chart?* <http://kb.fusioncharts.com/questions/464/How+do+I+create+a+Heat+Map+Chart%3F>

42. SAS Institute Inc., 2012, *Data Visualization: Common Charts and Graphs* <http://www.sas.com/data-visualization/common-charts.html>

43. Catalist, 2013, *About Us* < http://www.catalist.us/about>

44. Sasha Issenberg, 2012, *The Victory Lab: The Secret Science of Winning Campaigns,* Crown Publishers, New York.

45. Natasha Singer, June 16, 2012, *You for Sale: Mapping, and Sharing, the Consumer Genome,* The New York Times. <http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html?pagewanted=all>

46. Peter Hamby, October 26, 2012, *Micro-targeting offers clues to early vote leads,* CNN Politics <http://politicalticker.blogs.cnn.com/2012/10/26/micro-targeting-offers-clues-to-early-vote-leads/>

47. Zeynep Tufekci, November 12, 2012, *Beware the Smart Campaign* <http://www.nytimes.com/2012/11/17/opinion/beware-the-big-data-campaign.html>

48. Catalist, 2013, *Products* <http://www.catalist.us/product>

49. Jim Rutenberg, November 12, 2012, *Secret of the Obama Victory? Rerun Watchers, for One Thing* <http://www.nytimes.com/2012/11/13/us/politics/obama-data-system-targeted-tv-viewers-for-support.html>

50. Darrel M. West, 2012, *Big Data for Education: Data Mining, Data Analytics, Web Dashboards,* Governance Studies at Brookings <http://www.insidepolitics.org/brookingsreports/education%20big%20data.pdf >

51. Greg Satell, March 3, 2013, *The Limits Of Big Data Marketing,* Forbes http://www.forbes.com/sites/gregsatell/2013/03/06/the-limits-of-big-data-marketing/2/

52. Christine Hall, March 1, 2013, *The Most Disruptive Marketing Trends Of 2013* <http://www.forbes.com/sites/bmoharrisbank/2013/03/01/the-most-disruptive-marketing-trends-of-2013/>

53. S. Rosenbush, M. Totty, March 11, 201, *How Big Data is Changing the Equation For Business*, The Wall Street Journal. <http://online.wsj.com/article/SB10001424127887324178904578340071261396666.html>

54. D. Rogers, D. Sexton, 2012, *Marketing ROI in the Era of Big Data*, Columbia Business School <http://www.iab.net/media/file/2012-BRITE-NYAMA-Marketing-ROI-Study.pdf >

55. Marketshare.com, May 18, 2010, *Understanding Television's role in driving sales effects* <http://marketshare.com/dev/images/documents/white-papers/ms_fox_broadcasting_whitepaper.pdf >

56. C. White, October 24, 2012, *Completing the Big Data Picture: Understanding Why and Not Just What,* TDWI Research

57. N. Desai, D. Steier, A. Verma, 2012, *Data: A growing problem,* Deloitte Development LLC

58. Euromonitor International, 2012, *Consumer Analytics* <http://www.euromonitor.com/consumer-analytics/consulting>

59. Cheapinsurance.com Blog, August 6, 2012, *What's the idea behind Progressive's Snapshot®?* <http://www.cheapinsurance.com/whats-the-idea-behind-progressives-snapshot/>

60. Liane Yvkoff, March 21, 2011, *Gadget helps Progressive offer insurance discount,* CNET <http://reviews.cnet.com/8301-13746_7-20045433-48.html>

61. E. Brat, S. Heydorn, M. Stover, M. Ziegler, March 25, 2013, *Big Data: The Next Big Thing for Insurers?* BCG Perspectives by Boston Consulting Group <https://www.bcgperspectives.com/content/articles/insurance_it_performance_big_data_next_big_thing_for_insurers/

62. Ron Winslow, March 19, 2013, *Heart beat: a plan to chart heart risk in 1 million adults in real time,* The Wall Street Journal <http://online.wsj.com/article/SB10001424127887324323904578368572640617966.html>

63. The Health eHeart Study, 2013, *Technology* <www.health-eheartstudy.org/technology>

64. IBM, March 22, 2012, *Memorial Sloan-Kettering Cancer Center, IBM to Collaborate in Applying Watson Technology to Help Oncologists (Press Release)* <http://www-03.ibm.com/press/us/en/pressrelease/37235.wss>

65. IBM, 2012, *Watson in Healthcare* <http://www-03.ibm.com/innovation/us/watson/watson_in_healthcare.shtml>

66. M. A. Beyer, J. Lovelock, D. Sommer, M. Adrian, October 12, 2012, *Big Data Drives Rapid Changes in Infrastructure and $232 Billion in IT Spending Through 2016*, Gartner Research

67. IBM, August 10, 2012, *IBM InfoSphere Foundation Tools, IBM InfoSphere Information Server, Version 8.5: Parallel processing configurations* <http://pic.dhe.ibm.com/infocenter/iisinfsv/v8r5/index.jsp?topic=%2Fcom.ibm.swg.im.iis.productization.iisinfsv.install.doc%2Ftopics%2Fwsisinst_pln_engscalabilityparallel.html>

68. Wikipedia, 2012, *Computer Cluster* <http://en.wikipedia.org/wiki/Distributed_cluster>

69. Wikipedia, 2012, *In-Memory Processing* <http://en.wikipedia.org/wiki/In-Memory_Processing>

70. By Krish Krishnan, December 20, 2012, *Data Governance for Big Data* <http://www.information-management.com/news/data-governance-for-big-data-10023716-1.html>

71. Google Developers, 2013, *Analyze terabytes of data with just a click of a button Use Google BigQuery to interactively analyze massive datasets* <https://developers.google.com/bigquery/>

72. Merv Adrian, June 7, 2012, *Who's Who in NoSQL DBMSs,* Gartner, Inc.