Global Journal of Engineering Science and Research Management

BIG DATA ANALYTICS: REVIEW OF TOOLS AND TECHNIQUES

Dr Tesfaye Demissie (PhD)
*Correspondence Author: **Dr Tesfaye Demissie (PhD)**

Keywords: Big data, data science, insight analysis, business intelligence (bi), data analytics.

Abstract

Perhaps nothing will have as large an impact on advanced analytics in the coming years as the ongoing explosion of new and powerful data sources. When analyzing customers, for example, the days of relying exclusively on demographics and sales history are becoming the things of past. Virtually every industry has at least one completely new data source coming online soon, if it isn’t here already. Some of the data sources apply widely across industries; others are primarily relevant to a very small number of industries or niches. Many of these data sources fall under a new term that is receiving a lot of buzz: big data. Big data is sprouting up everywhere and using it appropriately will drive competitive advantage. Ignoring big data will put an organization at risk and cause it to fall behind the competition. To stay competitive, it is imperative that organizations aggressively pursue capturing and analyzing these new data sources to gain the insights that they offer.

Introduction

The concept of making analytics operational isn’t new. However, it was rarely achieved in practice in the past. The fact is that companies could get away with less, and so they did. The Analytics era began in the early 2000s to emerge and guide us into the world of big data. Big data is in many ways new. It encompasses data that is often more complex than, larger in volume than, and not necessarily as structured as the data used in traditionally. Big data can include anything from documents, to photos, to videos, to sensor data. A lot of big data used for analysis, such as social media data, is also external to an organization. Though externally created, data can still be very valuable. As technology has advanced and businesses have become more sophisticated, however, analytics is becoming an inevitable requirement. It just won’t be possible to compete in the future without analytics being at the heart of a wide range of daily decisions and actions [8], [9], and [12].

In the era of Analytics today, technologies such as Hadoop have gone from obscurity to being well known, and analytics processes have been updated to account for such new technologies [10], [21]. A major focus in the Analytics era is finding the cheapest way to collect and store data in its raw format. One strong trend has been the recent rise of the term “data science” to describe how analytics professionals analyze big data; and the associated “data scientist” term to describe the analytics professionals doing the analysis. A primary difference between data scientists and traditional analytics professionals is the choice of tools and platforms used for analytics. Traditional analytics professionals in large organizations tend to use tools like SAS and SQL to analyze data from a relational database environment. Data scientists tend to use tools like R and Python to analyze data in a Hadoop environment [15]. However, those differences are tactical and largely a matter of semantics. Anyone strong in one of those environments can easily make transition to the other. The underlying skill sets and mind sets are virtually identical across these analytics professionals even if the labels are different.

Research methodology

Big Data and Data Science has been buzz words floating around in recent years. Books, literatures, articles and even the wider internet have tried extremely to spread the words. However, there appear to be a gap in explaining and differentiating the two, rather presented intermingled. Data Science and Big Data are quite distinct and the author of this article tries to narrow the gap by presenting reviewed explanation and relationship among these two hot subjects of modern science concepts. Moreover, the aim of this publication is to present skills and knowledge necessary to carry out practical data analytics to gain insight from data in our ubiquitous information era. As part of the study, the author presents the concept of Big Data and Data Science in the context of gaining business, research or consummate advantage from data. In addition, strategies for implementing data Analytics in any walk of life is featured. The study method used is subjective descriptive analysis based on extensive literature review in the context of gaining advantage.

What is Big Data?

Big Data refers to data because of its size and format that is Volume, Velocity and Variety that can’t be easily stored, manipulated or analyzed by traditional methods like spreadsheets, relational databases and common statistical software’s. Practical definition of big data includes among others investigating how it relates to fields such as Coding, Statistics and Domain knowledge and varieties of people involved in big data itself. Also it gives clarity to its definition when we relate and see how big data is used in fields such as marketing and scientific research. Looking at how big data influences customer services like recommendation engines and the ethical issues raised by big data when analyzing anonymous public records such as voters list (name, address, date registered, party affiliation, date last voted, zip code, sex, etc.) and medical data (ethnicity, visit date, diagnosis, procedures, medication, total charge, contact details, etc.) justifies for its popularity and success [12]. The applications of big data have been extraordinary and its possibilities are immense and it will become important to understand its common methods such as measuring and capturing, storing and manipulating, analyzing and visualizing including data mining and predictive analytics.

Three V's of Big Data

Big data is an ambiguous and relative term. It may be better to find out what it's not. It is not regular data. It is not business as usual and it is not an experienced data analyst may be able to deal with. To put it in another term, big data is data that does not fit well into familiar paradigm. It does not fit in to rows and columns of excel spreadsheet. It cannot be analyzed with conventional multiple regression technique, and probably will not fit into the computer's hard drive anyhow. On the other hand, one way of describing it is by looking at the three V's – Volume, Velocity and Variety. This came from an article written by Doug Laney where he comments on the common characteristics of big data; and certainly are not the only once. (http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf)

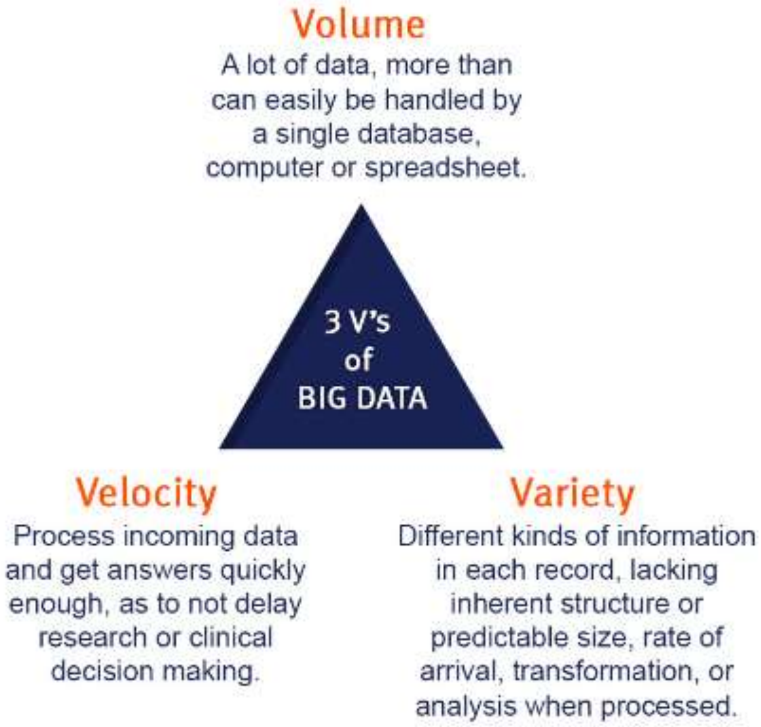

Volume
A lot of data, more than can easily be handled by a single database, computer or spreadsheet.
3 V's of BIG DATA
Velocity
Process incoming data and get answers quickly enough, as to not delay research or clinical decision making.
Variety
Different kinds of information in each record, lacking inherent structure or predictable size, rate of arrival, transformation, or analysis when processed.

Figure 1: 3 V's of Big Data

Volume

In the simplest possible definition, big data is data that is just too big to work on our computer. Obviously this is a relative definition. What is big for one system at one time may not be big for another system at another time. That reminds us of the general principles of Moore's law – a well-known observation in computer science which pronounces that physical capacity and processing power of computers double every two years (http://visual.ly/moores-law). For instance, the amount of rows that we can have in a single spreadsheet changed over time. Previously it was 65,536 in Excel 2003 and it is 1,048,576 in Excel 2007 which seems a lot. But if we login into Internet activity where something can occur hundred and thousands of times a second, we can reach million rows very quickly. On the other hand, if we look at photos and videos where we need to have information in memory at once, we have entirely different issue. An iPhone for example takes photos 2-3Mb per photo and video 18Mb per minute or 1GB per hour. If we have red epic video cameras, we could do up to 18 GB per minute and instantly we can end up having very big data. Some people call this lots of data, meaning it is the same idea of using data that we generally used to but it is just more of it. That takes us into the issue of velocity and variety.

Velocity

Velocity is when data is coming in very fast. In conventional scientific research, it takes months to gather data from say 100 cases; weeks to analyze data; and years to get the research published. This kind of data generally is static at one center and doesn't change, and the process time consuming. As an example, perhaps the most familiar dataset for teaching the statistical procedure cluster analysis is the iris dataset collected by Edgar Anderson and analyzed by Ronald Fisher. Both have published their paper in 1936 [8]. The dataset consists of four measurements – type, petal width (PW), petal length (PL), sepal width (SW), and sepal length (SL) for a sample of 150 irises. The lengths are measured in millimeters. Type is coded categorical (0 is Setosa; 1 is Verginica; and 2 is Versicolor). The dataset is used ever since until recently and it is one of the built in dataset in statistical programming language R and it has not changed for nearly 80 years.

At the other end of the scale, if we are interested in using data from social media platform such as twitter, we may have to deal with what is called fire hose. In fact if we manage to hook in right now, they are processing 6000 tweets per second globally and that is 500 million tweets per day and 200 billion tweets per year (https://about.twitter.com/company). The first tweet was sent on March 21, 2006 by Jack Dorsey, the creator of twitter. It took three years until the end of May 2009 to reach the billion tweets. Today it takes less than two days for one billion tweets to be sent. Even a simple temperature sensor hooked up to our domino microprocessor through a serial connection and sending just one bit of data at a time can overwhelm the computer if left running for long enough. Now this kind of constant influx of data better known as streaming data presents special challenges for analysis because the dataset itself is a moving target. If we are constrained in working on static dataset in a programming language like R or SPSS, the demanding complexity of streaming dataset can be very daunting to say the least.

Variety

The third aspect of big data is variety. What do we mean here is that dataset is not just rows and columns nicely formatted in spreadsheet for example. We can have many datasets in many different formats. We can have unstructured text like books, big posts, columns of news article, tweets, etc. Research has estimated that 80% of enterprise data is unstructured [17]. So for the majority it is a common case and can also include photos and videos. Similarly, datasets can also include networked graph data like social connection data. Or if we are dealing with datasets like what is called Not Only SQL (NoSQL) databases, we may have graphs of social connectivity such as hierarchical structures and documents. If we have any number of dataset formats that do not fit well into the rows and columns of a conventional relational database or spreadsheet, then we can have very serious analytical challenges. In fact recent study indicates variety is the biggest factor that is leading companies to big data solutions.

The final question here is that, do we have to have all the three V's – Volume, Velocity and Variety at once for one to have big data. It may be true that if we have all the three at once then we have big data but any one of them might be too much for standard approach to data. Really what big data means is that we cannot use our standard approach with it. As a result big data can present a number of special challenges. The fact is that big data approach has been used already and amazing things have been accomplished by using big data for research, for business and even for the casual consumer [15], [19].

Additional V's

At the beginning we defined big data is typically defined by three Vs.

Volume – very large amount of data; Velocity – it often comes quickly and can be streaming data; variety – a lot of different formats especially not in regularly structured rows and columns of a spreadsheet or a relational database. On the other hand, there have been a series of other V's proposed. The thing is whatever V we add to define big data, it has to have legitimacy and lots of factors to consider with big data research.

Veracity for example refers to the question: will the big data give us insight into the truth about our research question? This has to do with the reasoning out if the data contains enough information at a sufficiently micro level for us to be able to make accurate conclusion about larger group of cases.

Validity: - is the data clean? Is it well managed, does it meet the requirements, is it up to a set of standards of the discipline?

Value: - Does the data has any value? In a business settings that is specifically translated to ROI (Return on Investment). Is it worth our time to engage in a big data project, as it is not always going to be the case? Big data is still an expensive time consuming major undertaking and in most situations we need to think about whether that particular analysis is really going to further the organizational goal.

Variability: - the data can change overtime and we can usually analyze. But it can also change or replaced. There are a lot of uncontrolled factors that may introduce noise into our data and must be measured and account for them.

Venue: - this refers as to where the data is located and how does that affect access to the data. How does it affect its formatting?

Vocabulary: - this refers to the meta-data that is used to describe the data especially when we combine the data from very different sources. When we are talking about the same variable, the same kind of information, others may use different terms to describe it. It may not be clear but what is recorded by others might be the same thing. So that becomes one of the challenges in combining data sources in a big data project.

Vagueness: - this really refers if we actually know what we are talking about. What do we mean by big data? What are the goals? What are we trying to accomplish? We have to have clarity of purpose or we have the potential of wasting an enormous amount of time and energy chasing the wrong thing. Big data is there to serve a purpose and to give us an insight that we cannot get otherwise and gives us an ability to function much more efficiently and intelligently. We need to be clear about what we are doing or our time and resource may be wasted.

Big data sources

Big data is everywhere and it can help organizations and industry in many different ways. Nowadays there is so much data that existing hardware and software are not able to deal with the vast amount of different types of data that is created at such a high speed. Big data has become too complex and too dynamic to be able to process, store, analyze and manage with traditional data tools. But what are the various sources of Big Data?

Human Generated Data

Big data can come from several different sources. One way to think about is whether the data is produced by human or whether it is produced by machines. We humans generate a lot of data whether we meant it or not [13].

Intentional data

This is data that we know we are creating. For instance if we take a photo, video or audio; or if we put text on a social network we know we are creating it. We can also click 'Like', use Facebook, do web searches, a record of webpages that we viewed, bookmarked webpages, emails and text messages, mobile phone calls, eBook we read, highlights, notes, online purchases, etc. – all of these are kinds of data that do not exist until the person deliberately makes them happen. So these are records of human actions. What is interesting about it is in addition to these intentional pieces of information, there is also metadata.

Metadata

Metadata is data about data. We might call this 'second order' human generated data. The idea here is that these are data that accompanies the things we do and we may not be aware of it. The metadata first of all can be enormous sometimes larger than the actual piece of data we created and most significantly for the big data world, metadata because it is computer generated it is already machine readable and searchable. If we take for instance when we take picture with our phone, we are not only getting the picture but also there is what is called 'EXIF' data – that is for Exchangeable Image File format. And this is the metadata that comes from the picture of Phone. If we look at the individual piece of information of the metadata, there is an enormous amount of information attached to the particular image. Another interesting thing about mobile phone metadata is that information is not normally publicly available. This really is about a sort of academic interest there but the idea here is that there is a lot of information that accompanies our phone call without even knowing we are calling and what we said on the phone call. Two pieces of information from such a phone call are the time of the phone call and the location of the call. With four separate pieces of such data or rather four phone calls where we know the time and position in an anonymized dataset that is enough information to identify 95% of individuals. This is to emphasize that this is the information that is not publicly available but the point here is that it is possible to tell a lot about people from the metadata. (http://www.digicamhelp.com/glossary/exif-data/)

Email metadata

There are a lot of information that accompanies each email message. Four very common pieces are who is it from, who is it going to, did the message copied to somebody (CC) and the time it was sent (From, To, CC, and Timestamp). Now what an interesting thing we can do to ourselves is that MIT has created a web App called 'Immersion' that allows us to do a quick analysis of our own social network via our own email account (Immersion.media.mit.edu).What Immersion does is that it takes four pieces of information about our emails and it puts together in an image of the network. It takes few minutes and we have to refresh it to do an update. The result shows the number of email contacts we made and the number of emails communicated during particular time span. The information only comes from the four common pieces of a typical email message (http://www.thewire.com/technology/2013/06/email-metadata-nsa/66657/).

Twitter metadata

Twitter is very popular and one of the most useful social network platform. Tweets are very small, limited to 140 characters. What is interesting about twitter from a research point of view is that there is an enormous amount of metadata that accompanies each tweets. What is, technically, included in our tweets? There are not less than 38 different pieces of information in single tweet (http://www.slaw.ca/wp-content/uploads/2011/11/map-of-a-tweet-copy.pdf).

The metadata in this case is several times greater than the actual content. This is why Twitter in particular is a very rich dataset for people who are doing marketing research or social connection research. The point of all these is that we all are sources of big data. The metadata in particular does not have to be processed. It is already computer readable, it is searchable and we can start to get information about it immediately to reach our big data analysis purposes.

Machine generated data

It has been estimated that as much as 95% of the world's data will never be seen by human eyes. Much of these unseen data is called machine – to - machine (M2M) data. The machines are talking to each other not to be heard by humans. The sources of these machine generated data has very long list although difficult to give comprehensive list. For example, when mobile phones are pinned to the mobile towers to check where they are, when the satellite radio and GPS are connecting to locate the carrier of our mobile phone, when the RFID (Radio Frequency Identification) readings tags on a billion of small objects, recordings from medical devices, web crawlers and spam bolts to especially find out about things like electric spam bolts being served by electronic spam filters, each working against another.

Perhaps the most interesting part of the M2M communication falls under the real brick of the Internet of Things (IoT) sometimes. It is estimated that by 2020 which is only few years away from now, as much as 30 billion uniquely identifiable devices may be connected to the Internet [3]. This actually requires the major change how the Internet works as addressing system for all these things to fit. But basically everything will have a chip and will be connected to Internet and will be talking to each other sharing information. So it will not be long before we see or hear people talking about smart sensors in our home, in our city, our air conditioning, or the smart home which turns the lights on, changes the temperature, and the smart grids where the city generates and send out electricity or the smart city itself which coordinates the traffic and utilities as a way of being more efficient, more economical in providing better services. All of these can be enabled by small objects communicating with one another in the Internet of Things that communicates directly not with the human intermediary.

Some of the uses for these can include putting centers and production lines to monitor systems if they need maintenance, or the smart meters for utility systems to shut them out at peak times if they can do without interrupting service, identifying pets and farm animals and check on them through systems, thermostats and light bulbs (we can actually get our Phone controlled light bulbs already) or just environmental monitoring that can include things like air and water quality, atmospheric soil conditions, movements of wildlife's, earth quakes and tsunami early warning systems. Also things like infrastructure management that talks about controlling and monitoring of bridges, railways, wind farm traffic system; industrial applications like manufacturing process control, supply chain network, having predictive maintenance or integration with smart gate energy consumption; energy management that is switches and outlets, bulbs, TV's, screens, heating systems, controlling ovens, changing lighting conditions; and medical and healthcare systems, building and home automation systems, transport systems, etc.

Basically anything that is mechanical can be eventually hooked up and communicate through the Internet of things. All of these generate an enormous amount of data as they talk to one another and coordinate their own activities. In looking at differences between machine generated data and human generated data, the most obvious difference is that the machines do not generally post selfies on Facebook, they do not make silly videos on You tube, they do not write job applications and put them on LinkedIn, etc. The content is different but what is interesting is that content may not be the most important difference. Instead the most important distinguishing feature of machine generated data is almost all of the machine generated data is machine readable. They can be immediately searched and read and mined. It is high in volume and velocity, two of the characteristics of big data, but low in variety makes it easy for machines to deal with it. That brings up the important distinction between what is called structured, unstructured and semi-structured data.

Reasons to adopt big data

Having big data does not automatically solve organizational questions or overcome research challenges. We can have Hadoop but that does not mean we understand what is going into it. We may have the most sophisticated predictive analytic equation but if it is based on something that is irrelevant, we have wasted our time. In introducing big data, there are more things to be aware of because there are more places for things to go wrong and will be confusing to trace and apply remedy. Responsibility tend to be spread out in a big data project. There involve usually different groups of people who prepare the data, who analyze, who visualize, who apply it, who form the new sets of questions and find other information and bring it in. Because we can have hundreds or even thousands of people involved in a single big data project, no one person is overlooking everything and as a result responsibilities to answer these questions such as things like value, vagueness become incumbent in everybody's understanding of what is going on. There is a greater need to think about quality because there are so many opportunities to let things slip through the crack. Also there is a greater need to think about meaning of the project as we go through it. If we do that, we are going to be in a much better situation. And if we assume the quality of the data in the project has been verified, then our next step is the actual analysis of the big data.

Big Data for Consumers, Business and Research

According to [19], by adopting big data techniques into their operations, Consumers, Business and Research organizations will be able to:

1. **Stop wasting data exhaust:** - An organization's ability to produce data greatly exceeds its ability to store and manage it.
2. **Save time and money:** -Even when data is captured and saved, it is typically stored in disconnected silos. This increases the amount of time it takes employees to find information.
3. **Improve performance:** -Big data enables information from both inside and outside the organization to be combined, so that key performance indicators can be developed and acted upon.
4. **Improve Product Offerings: -** Data from a variety of sources can be combined to improve existing products.
5. **Segment Groups within Larger Populations:** - Big data techniques can combine and analyze data from a number of sources and segment specific sub-groups within larger populations.
6. **Improve decision making:** - Big data is a powerful means for making informed decisions. However, as big data technologies continually improve, there remains a shortage of skilled professionals who can take full advantage of these technologies.
7. **Innovate:** - Big Data diminishes the need to rely on preconceived ideas or assumptions, by enabling innovators to analyze and experiment using real data, in real time.

Big data for consumers

Most of the time when people talk about big data, they are talking about it in the commercial settings such as how big data can be used in advertising and marketing strategies. But one really important place that big data is also used is consumers. What is interesting about this is while data and the algorithm are available, and once incredibly complicated processing involved, data is nearly invisible. The result is so clean, small and is exactly what we need in most cases. There are some common applications of big data for consumers that we may be using them already without being aware of the sophistication of the big data analysis that is going into it (Siri of iPhone, YELP, Netflix, Google Now) [14]. Therefore, for consumers, big data is providing valuable services but again with the irony that it operates invisibly by taking a huge amount of information from several different sources and distilling it into just two or three things that gives us what we need.

Big data for business

We saw how big data can provide important conveniences and functionalities for consumers. But for the business, big data is revolutionizing the way people do commerce. As an example, let's look at where most people have encountered big data in commerce, the result of Google Ad searches. Whenever we search for something on Google or any other search engine, we type in our term and then we are not going to find only the result of our search but also ads. Those ads are not placed on random. They are placed based on not only on the term we are currently searching for but also on what the search engine knows about us. The search engine draws all the information about things we have been searching for and information about us together and tries to place ads that we would be most likely interested with and respond to. That is something to guess about having a very large amount of data available to tailor thing to be most applicable to the consumer. Another interesting place to see is what is known as predictive marketing. This is when big data is used to help decide who the audience would be for something before they actually get there. These include things like trying to predict major life events, getting married, graduation, new job or having a child or any number of events that are often associated with the whole series of commercial transactions. To do so, those companies are looking at consumer behavior. They look at how often we log on into their website, look at what credit card we use, how often we look at particular item before moving into something else. They look at whether we have applied for an account from an organization. They can make use of a huge amount of information already available to them. Similarly, they can use demographic information. This can include things like our age, marital status, number of children we have, home address, how far we live from their store, our estimated salary, are we moved recently, what credit card we have been using, which website we visit, etc. All of these information are potentially available to them in one form or another to the company trying to make the prediction. Similarly they can rely on additional purchase data. It is possible for the company to get information about our ethnicity, job history, magazine subscriptions, whether we are declared bankrupt, whether we are divorced, college attended, and the kind of things we talk about online and so on. There is an enormous amount of information out there that they can be potentially important. Now this is going to lead into an important discussion about ethics and big data and similar issues that needs addressing. Using these kind of information it is possible for companies to predict what products we could potentially purchase when we are about to buy a new house, for example. In most cases if we buy a new house we make an enormous amount of purchases so that they can wink into us before we can make commitment to them.

Another area we can look about big data is trying to predict trends. One of fascinating places for this is in fashion. The company called "EDITED" is already making use of big data in predicting fashion trends. What they does is that they gather information from customers and turn it in to value so that retailors have the right products, at the right price, at the right time. They usually advise retailors what colors, brands and styles are most popular and are going to win the market. Obviously these kind of information is enormously important to the companies that are going to sell these products. And EDITED is able to do this through their reliance on big data. Another area worth considering when using big dat in commerce is in the area of fraud detection. It turned out that fraud is an enormous industry; that online retailors lose about 3.5 billion dollars each year to online fraud. And insurance fraud and health insurance loses are enormous per year. So fraud is big issue in commerce. There are a number of things that companies can do to lesser the prevalence of fraud especially about online transactions. They can look at the 'Point of sale' that specifically needs to know how we are going to make the purchase. They can gather information like - are we making the purchase online, in store, or what website. They can also use geolocation and IP Address to learn where we are physically located when browsing the website. They can look at log in time, are we somehow making purchases at 4PM when we have not done after 11PM before. Interestingly they can also look at biometrics. It is evidenced that the way people move their mouse or the time they take between pressing keys on the computer are distinctive measurements of people. Similarly, when we hold our mobile phone and look at it, people have different heights, hold their phones at different angles as measured by the accelerometer on the phone. All these can be used to determine whether the person who is making the purchase is who they say they are.

Big data for research

Big Data is influencing scientific progress. Let's see some examples of how big data has been revolutionizing aspects of scholarship and research.

As discussed in [4] and [8], Google Flu Trends (GFT) is an example which was built to predict center for disease control and prediction's reports (CDC) using big data, an organization responsible to find search patterns for flu related illness. GFT were actually able to identify outbreaks of the flu in the United States much faster than the center for disease control do. Similarly most recent research found out Wikipedia searches could identify them with even greater accuracy. The national institute of health created the "BRAIN" initiative as a way of taking enormous number of brain scans to create a full map of brain functionality (http://en.wikipedia.org/wiki/BRAIN_Initiative). Additionally, NASA's Kepler Space Telescope has been on a mission to find exoplanets or planets outside our solar system. So far it has discovered nearly 1000 confirmed planets and worth over 4000 candidates (http://www.nasa.gov/sites/default/files/exoplanetdiscoverieshistogram.jpg).

Closer look also reveals psychological research is also been influenced through big data. In the United States, a recently published research reveals people in the north-central Great Plains and south tend to be conventional and friendly, those in the western and eastern seaboards lean towards being mostly relaxed and creative, while new Englanders and mid-Atlantic residents are prone to being more temperamental and uninhabited, according to API's online journal. (Journal of Personality and Social Psychology)

Similarly another group of researchers created another application on Facebook that used a scientifically valid measure of personality. They get data from several hundred respondents and by combining those with the patterns of likes each of those people has on Facebook, they finally were able to create a single question App that really just has access for our lives. It is able to give a surprisingly accurate evaluation of what our personality would be if we took the entire questionnaire.
(https://apps.facebook.com/mypersonalitytest/quiz.php?quizid=5220&defaultcookieset=1)

The Google books project (http://www.google.co.uk/googlebooks/about/) has been scanning books that have been published over the last few hundred years. They currently have over 30 million books that they scanned and makes them digitally accessible. And now there is what people call "digital humanity" to look at changes in word usage over time. There are also some interesting things for instance, from the last 280 years of the prevalence of the words math, arithmetic and algebra, where arithmetic shows a spike in the 20’s and 30’s but decreases over time; whereas the word math increases over the last 50-60 years. The idea here is that big data has brought the volume of information that is available, the variety of information that can be combined and the velocity and especially with the things like flu trends that is constantly changing. All of the areas mentioned above are able to make good use of big data for scientific research and advancement, and it is an exciting time to learn from what has happened and to see what will happen in the future.

Big data and data science - The 3 Facts

When people talk about big data, they nearly always talk about data science and data scientists as well. Just as the definition of big data is debated, the same is true with the definition of data science. To some people the term is simply the fancy way of saying Statistics and Statisticians. On the other hand others argue that Data Science is a distinctive field. It has different trainings, techniques, tools and goals than Statistics typically has. Let’s start by looking at the Data Science Venn Diagram.

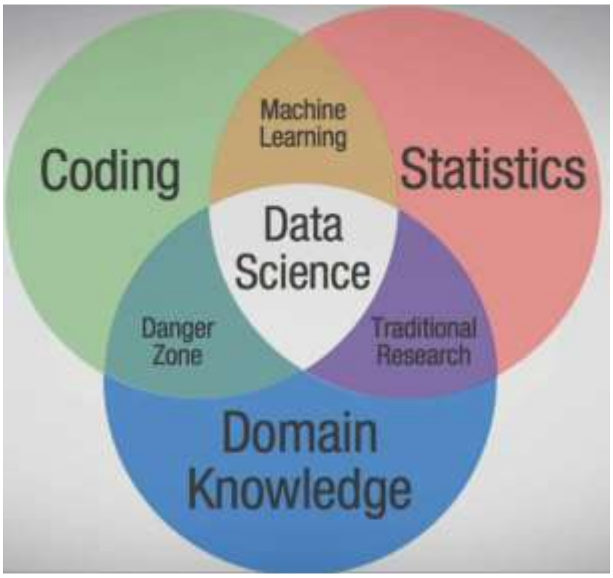


Figure 2: Data Science Venn diagram

This is the chart created by Drew Conway in 2010 (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram). What Conway is arguing here is that Data Science involves a combination of three different skills. The first is Statistics at the top right. The second at the bottom is Domain Knowledge to refer to what we actually know about the subject for instance management, or advertising or support recruiting. And the one at the top left is Coding, the ability to be able to program on computers. Conway is arguing that a Data Scientist is a person who needs to be able to do all three of these.

The first is Statistics. The trick is that a lot of things are going to use Statistics and Mathematics and can be counted twice. If we do not have the specific formal training, we can make some real big mistakes. One of the best example could be the birth day problem and probability. Let’s consider a situation that “What is the probability of two people in the room to have the same birth day? – Month and day” Intuition suggests that to have a 50% chance of a match, we should have over 180 people in the group because that is about half of the number of people as are days in a year. On the other hand the correct answer is a lot smaller than that. It is in the 20’s. That is what we need to have a 50% chance of a match. Because Data Scientists often have been looking for matches and associations, skill to get this probability correct is really very important part. That is why Statistical training is an important part of Data Science.

The second element of Data Science is Domain Knowledge. The idea here is that researchers should know about the topic area that they are working on. So if we are working say in marketing, we need to understand how marketing works. That makes easy that we can make more insight and better direct our analysis and procedures to match the question we might have. There is a quote from Sicular of Gartner that says:

“Organizations already have people who know their own data better than mystical Data Scientist – this is a key. The internal people already gained experience and ability to model, research, and analyze. Learning Hadoop is easier than learning company’s business.” – Svetlana Sicular of Gartner, Inc. That really underscores the importance of Domain Knowledge in Data Science.

The third element in the diagram is Coding. This refers to computer programming ability. It does not need to be complicated, or we do not need to have a PhD in Computer science either. A little bit of Java, Python, R, SAS or SPSS programming for example can go very long way. Because this allows for the creative exploration and manipulation of dataset especially when we consider the variety of data that is part of big data. The ability to combine data that comes in different format can be really an important thing and that often requires some coding ability. It also helps to develop algorithmic thinking – thinking in linear steps, steps-by-step to get through a problem.

The question now is - Can Data Science still work and make sense with lesser combination? Let’s see the combination of two of these elements at a time. If we look Statistics together with Domain Knowledge without Coding, anything we do this way is traditionally called Traditional research. This is where researchers work within the field of their expertise, use common tools for working with familiar data format. It is extremely productive and nearly all existing researches have been conducted this way. Organizations usually require researchers to use the simplest methods possible that can adequately address their research questions. That is what they call a minimally sufficient analysis. So this traditional methods are extremely important but they are not just sufficient for working with big data.

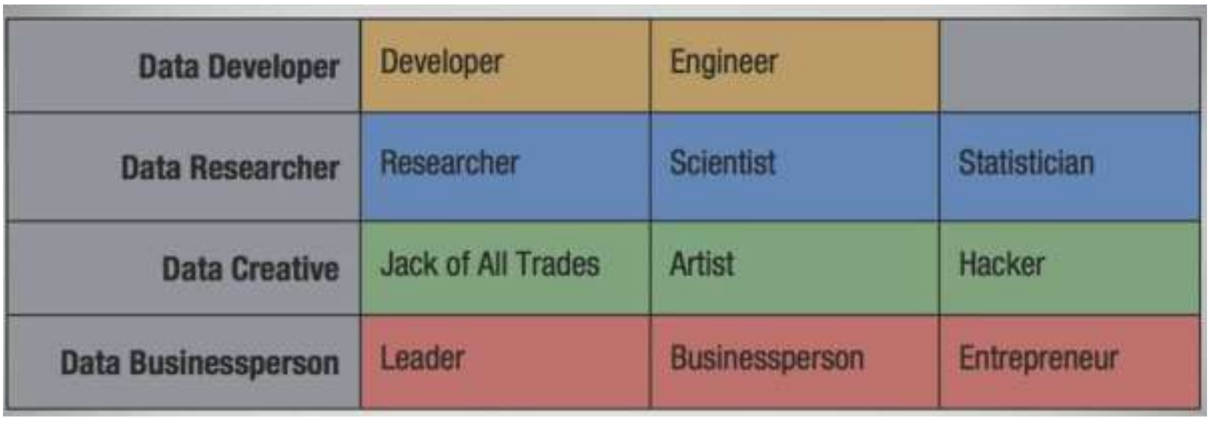
The second combination is Statistics with Coding without substantive expertise. This is what is commonly referred as Machine learning. Now it should not be confused with data mining. Machine learning is when an algorithm or a program updates itself in a variable in a specific analytical task. The most familiar example of this is Spam filter in email messages. The user or a whole large group of users identify a message as spam or not spam. And the formula that a program uses to determine whether something is spam or not updates with each new piece of information to have increased accuracy the more we use it. In machine learning, we do not actually know how the program is doing what it’s doing. On the other hand, if what we are looking is for prediction only, this can be a very effective method. However it is not always enough to constitute Data Science without an important element of substantive data knowledge.

The third combination is Coding with Domain knowledge without Statistics. Can we label this as danger zone? With the idea being that we have enough knowledge to be dangerous. There are promises with this. First, it seems very unlikely that a person can develop both programming expertise and substantive knowledge without also learning some Mathematics and Statistics. This is not populated category. On the other hand, there is really important Data Science concept and contributions that comes out of this marriage including for instance word count. It is simple stuff. These are procedures that do not require sophisticated Statistics and we just count how often things occur and we get important insight out of that. However, it is less likely that we develop expertise from Coding and Domain knowledge without learning Statistics. Finally, all these – Coding, Statistics and Domain knowledge at once- makes the most correct combination and hence definition of Data Science [16].

Types and skills in data science

When people talk about big data in newspaper articles or professional conferences, it is easy to get the idea that Data Scientists are not just people who have done major expertise and who understands Statistics and computer programs. Instead they often start to sound like omnificent or omnipotent super humans who can do anything and do it instantly and effortlessly. Of course that is not the real picture. As with any other domain there is a large range of skills involved in Data Science beyond the three elements mentioned earlier.

A great report called “Analyzing the Analyzers” – an introspective survey of Data Scientists and their work goes over this detail (http://www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf). The authors surveyed about 250 Data Science practitioners and asked them how they identify themselves and cluster themselves on the relevant skills to Data Science. Each of these classifications was subjected to cluster analysis and cross-classification. What they found out is as expected that there is high level of heterogeneity of skills among people in big data; not everyone is the same. Accordingly the respondents placed themselves into eleven different professional identities.

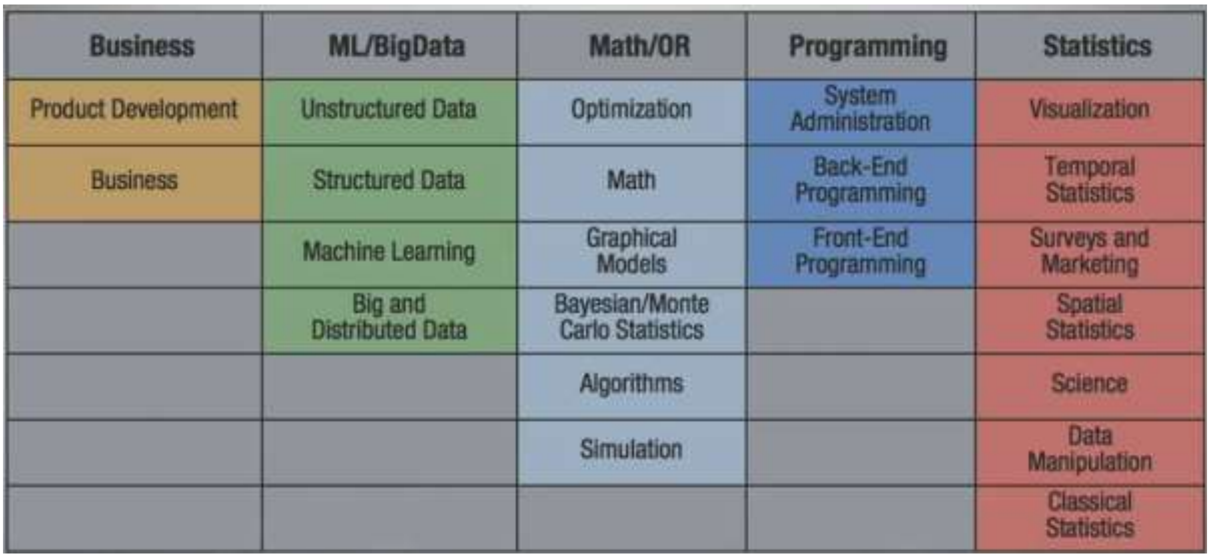


Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepreneur

Table 1: Data Science relevant skill clusters

The table shows how these eleven identities clustered based on the individual responses. They fall into four basic categories: Data Developer (Developer and Engineer), The Data Researchers (Researcher, Scientist, and Statistician), The Data Creator (Jack of All Trades, Artist, Hacker) and The Data Business Person (Leader, Business person, Entrepreneur).

The respondents also ranked themselves into twenty two possible skills including Algorithms, Visualization, Product development and Systems Administration. These skills are clustered or sorted into five general categories relevant to Data Science: Business, Machine Learning (ML)/Big Data, Mathematics/Operations research (OR), Programming, and Statistics. What is important here is that we can see the skills are not the same for all of them. For example, in the Business category, we have product development and business whereas in programming we have systems administration, back-end programming and front-end programming skills.



Business	ML/BigData	Math/OR	Programming	Statistics
Product Development	Unstructured Data	Optimization	System Administration	Visualization
Business	Structured Data	Math	Back-End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front-End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian/Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Table 2: Data Science skills

Not everybody needs to be able to do all of the same thing. In fact when the researchers crossed the self-identification category with the skills, they got rough profiles of the skills associated with each category of Data Science practitioner. Not surprisingly enough, the data business person is a person who has skill in working with data but views himself primarily as a business person and later on as entrepreneur. The Data creative person is interesting for being the one whose skills are evenly distributed. For example the Data researchers sees their skills lie primarily on Statistical Analysis. The most obvious thing here is again not everybody is the same. Each group has some skill in some area but the distribution differ dramatically. What this makes clear is that there is room for substantial variation and personal interest and skill settings in Data Science by extension within big data. It is helpful for everybody to know at least a little bit about each of the five skills that the researchers mentioned but diversity is really the name of the game here.

Data Science without Big Data

If we argue that big data requires all of the three V's – Volume, Velocity and Variety at once to count as big data proper, then it is entirely possible to be a Data Scientist without touching big data , i.e. a person with Domain expertise, Statistical knowledge and Coding skills. Let's look at the previous Venn-Diagram of both Data Science and big data at the same time. We are talking here about Statistics, Domain Knowledge and Coding of Data Science; and Volume, Velocity, and Variety of Big data at a time. Example:-

Data Science + Volume: - This means very large static dataset with a consistent format. The data would generally be structured as well so we can have free text. A good example for this can be genetic data. Genetic data is huge but it follows a structure and we have an enormous amount of task to work through it but it is consistent in that way.

Data Science + Velocity: - This primarily refers to streaming data with consistent structure. By streaming data we mean data is consistently coming in and very often we are not holding on to the data; we just keep small window as it opened.

Data Science + Variety: - This is a case where we have a complex but small and relatively static dataset. Example could include facial recognition and personal photo collection so we do not have enormous amount of photo but we can have a variety of photos and it is static that we do not add to it constantly.

Despite the strong association between big data and data science, the skills of Data Science that Statistical knowledge, Domain expertise and knowledge of Coding skills; those apply even when the three major aspects of big data are not all present at the same time.

Big data without data science

Here we will be looking at situations where a person who is working with big data but does not require the full Data Science skill set. As a way of reminding ourselves, big data usually involve unusual volume, velocity and variety in the data. Data science on the other hand involves Statistical skills, Domain knowledge and coding capability. Three of them together gets Data Science. Can we do big data with just two of the Data science skills, say just Statistics and Coding? The answer of course is yes. That is why we have machine learning. Machine learning is very important area of Data science. This is where a computer program learns to adopt to new information that comes in. The two most familiar examples are spam filters where the computer program learns that a particular kind of email is spam or not based on our own individual responses and based on the responses of millions of people who use the same email like Gmail; or facial recognition and photographs where the computer program learns what face belongs to whom. Therefore, machine learning is a good example of working with big data without Data science because it can have volume, velocity and variety but we do not necessarily need to have the main knowledge of Data science. Because the computer is working without any knowledge what so ever, the issue as far as machine learning is concerned simply is did it got it or did it not.

Another possibility is the region that is referred as danger zone in the Data Science Venn-diagram which omitted Statistics and considered Coding and Domain knowledge for big data. There are a good examples of Data Science that do not involve Statistical knowledge. The most common of those are word count and parsing of natural language toolkit (NLTK). This is a package in Python programming language that allows people to do all sorts of amazing things. There are amazing things that can be done with natural language by simply counting how often the word occur without requiring any Statistical knowledge. This is because there is no influential procedure that goes into it.

Those are the two combinations of Data Science skills where we have two at a time without big data. The third combination is Statistics with domain knowledge of traditional research. Unfortunately, as much as we all value traditional research one cannot do big data without Coding skills. Without the Coding, it is particularly impossible to deal with the volume, velocity and variety of data that characterizes big data. Tremendous things are accomplished with traditional research but big data is not one of them. And it also goes to say that as far as we all can tell, unless we have at least two of these, it is just a non-start – we simply cannot work with big data if we have just statistical knowledge, or just Coding skills or just Domain knowledge. Instead what we have to do in that situation is to collaborate with people who have that information. In fact collaboration is the rule in Data Science because there exist such a broad range of skills that are necessary. No body usually will be able to bring all of them to it but they have to work together. This is one of the wonderful things because things in most interesting development are come about through collaboration and that is something Data Science encourages strongly. So when we look at the relationship between Data science and big data, the situation is not exactly even. It is possible to do Data Science with incomplete version of big data. But it is much more difficult to big data work without triumph of Data Science skills.

Big data analysis

Structured Data

Data is said to be structured when it is possible to place the file with fixed filled variables. The most familiar example of this kind of structured database is spreadsheet, where every column is a variable and a row is a case of an observation. In the business world however a large dataset are usually stored in databases, relational databases have to be specific which shares some characteristics with spreadsheets such as rows and columns and allows more datasets, more flexibility and more consistency. Research shows that nearly 80% businesses use some form of relational database with Microsoft SQL server, MySQL and Oracle as the most common options. As a background information to the founding principle for query language, in the early seventieth researchers at IBM (Donald D. Chamberlin and Raymond F. Boyce) wrote a paper describing the standard query language or SEQUEL (Structured English Query Language) [6]. Then later it has changed to SQL because of a copy right issue but still is generally pronounced as SEQUEL. IBM however did not commercially launched SQL. That happened in the late seventieth (1979) by relational software which become Oracle which still is the biggest provider of database software in the world. Oracle is also well known for making one of its co-founder, Larry Ellison one of the richest people in the world (fifth-wealthiest person in the world, with a fortune of $52 billion, according to Wikipedia). What SQL does is that it makes it easy to extract, count and sort data, create unions and intersections between sets. It is also used to add, update and delete data. It does all these in a language that is much easy to manage than most of the selects and clicks of spreadsheet and there are a whole lot that can be said about structured data in SQL databases.

Unstructured and semi-structured data

While standard row-by-column table is easy for a person to understand, if unstructured that is data not fixed in fields, text documents, presentations, images, video, audio, pdf, etc. is much more difficult for a machine to understand. It is not easy to sort this kind of data, hard to re-arrange it, count the values, and add more observations. The majority of data in business settings that are unstructured are estimated as much as 80% [7]. It is a little hard to deal with. We might have to convert it to text and use text-mining program to try to get structure out of the sentences of data. But that is difficult and time consuming to do. On the other hand data has not to be either structured in spreadsheet or unstructured in text. A third option is possible and that is semi-structured data. Semi-structured data is data that is not in fixed fields. It is not in rows and columns in spreadsheet but the fields are still marked and data are still identifiable. The two common formats for semi-structured data, not the only once, are XML (Extensible Mark-up Language) and JSON (JavaScript Object Notation).

Once we have the dataset, the next thing we need to have is the database in which we are going to store the information, As opposed to SQL databases that are used for structured data in rows and columns, semi-structured and unstructured data usually go in to what is called NoSQL databases. It is used to mean Not SQL, but now it is termed as Not Only SQL, because NoSQL databases are extremely flexible and can handle a wide range of data format. So NoSQL databases most of them uses a semi-structured format. For instance, the most common NoSQL database is MongoDB that uses JSON. It is nice because it has flexible structure. For certain tasks NoSQL database can be faster than SQL databases. On the other hand there have not been adopted as widely as relational databases. For example, research made in [1] and [20] shows that about 80% of companies have used relational databases, whereas only about 16% have adopted NoSQL databases. If we look the sheer volume, because Hadoop is a NoSQL database, and almost all big data is installed in Hadoop or MongoDB, even though there is fewer by hand count, an enormous amount of data is installed in that format [5, 11, and 18]. One of the big problem is that whereas the SQL databases all use at least relatively standardized version of the SQL query language, there is no standardized query language for NoSQL databases. And this means as we switch from one to the other, we may have to learn all over again how to work with it. That is the problem, but on the other hand NoSQL databases are an area of huge development and no doubt that will get resolved very quickly.

Implementing Data Analytics: counting the values

Analytics as the Goal, not a by-product: - A big trend that is connected to analytics is that a large number of products now collect data. In many cases, the analytics executed against that data are actually a primary, if not *the primary*, purpose of the product. In other words, a physical product often is simply a mechanism for collecting data today. Historically speaking, companies have always developed new products, whether it was a toy, a calling plan, or a type of bank account. The goal was obviously to have that product succeed, but the success of the product didn't depend much on data or analytics. Companies would collect data over time about the sales performance of a product, who was buying it, and what defects or issues were commonly identified. This would lead to ideas

on how to improve the product, but the data was a by-product of the efforts to sell the product rather an inherent property of the product.
What has changed today is that products are being released whose entire purpose is the data it is collecting and the analytics that it enables. The physical product itself is actually secondary and is nothing more than a channel for the collection and analysis of data. In some cases the value of the product to customers will be the analytics provided; in other cases customers get value another way while the company gets value from the analytics. When the analytics are for the benefit of customers, the product that can provide the most valuable data and analytics, rather than more traditional features, will beat the competition [14], [16].

Examples are starting to abound. A lot of the free services available on the web fall into this category. Consider free email services. The companies providing free email aren't giving people free email service because they want to perform a community service. The companies give away free email service because they can learn a lot about subscribers as they use the email service. The provider has opportunities to serve advertisements based on users' behaviors, and it gets paid when they respond. In some cases, a free email service actually reads through users' email texts and analyzes it to generate offers. If we frequently email our friends about sports, we can bet that we'll be getting a lot of offers focused on sports. In addition, the email provider may sell its knowledge of our interest in sports to other organizations that are willing to pay to find sports fans. It all comes down to reading privacy policies very carefully before agreeing to them.

The marketplace also now has analytics processes that have been directly turned into products. One example is Netflix's well known movie recommendation engine we mentioned before. It uses the data collected from customers as they navigate the Netflix site to identify other movies that the customers might enjoy. The movie recommendation system is actually considered a formal product at Netflix. Netflix looks for opportunities to add features and functions to the recommendation engine and to improve how it is presented to customers. The recommendation engine is credited with being a huge factor in Netflix's success. But this product called a recommendation engine is really just analytics and the use of data. The engine is also a fully operational process that runs its algorithms and presents results to customers millions of times per day with no human intervention.

Analytics products are blurring industry lines: - Let's now explore an interesting example that illustrates how products focused on analytics are starting to blur industry lines by discussing the new wave of personal fitness monitoring devices that are worn on a wrist or waist. While there are a number of products on the market from Nike, Jawbone, and FitBit, we focus here on Nike. If we go out, survey 100 people on the street, and asked them what Nike does, probably at least 98 to 99 percent would respond that Nike is a clothing manufacturer, a sportswear manufacturer, or something very similar. None of those statements is untrue. After all, to a large extent, that's what Nike has been known for over the years. However, some changes at Nike necessitate a re-examination of what industry the company actually is in. The same type of change is happening for many other businesses as well. In 2012, Nike released a product called the FuelBand. The FuelBand is a device that is worn on the wrist like a watch, and it measures things like the number of steps taken each day and several facts about sleep patterns. The device and other products like it are very popular. Let's examine what the FuelBand does to challenge Nike's industry classification and how it alters Nike's traditional business model (http://www.nike.com/gb/en_gb/c/nikeplus-fuelband).

Although most people still think of Nike as a clothing or sportswear manufacturer, the FuelBand breaks this assumption. To start with, the FuelBand is actually a piece of high-tech equipment complete with sensors, a transmitter, and more. Nike is now in the high-tech manufacturing business. What's the first thing customers have to do after buying a FuelBand if they are going to make effective use of it? They must download software to their desktop, tablet, or mobile device. Nike's now in the software business. And why do customers need the software? So their mobile device or computer can interact with the FuelBand and upload the data it collects to Nike. Nike is now in the data collection and storage business. The reason for all of this activity is to enable Nike to provide analytics and trends about customers' sleeping and activity patterns. Nike is now in the analytics as a service business. It is even possible to argue that Nike is in the health business too if over time the company finds ways to correlate the data a FuelBand collects with health issues. As a result of the FuelBand, Nike has entered a lot of business lines that truly have nothing to do with fashion or clothing. Perhaps the most important point is that the choice of buying a FuelBand or a similar competitive item really doesn't come down to how nice it looks or how fashionable it is. Those factors are important for traditional Nike items, but with a product like the FuelBand, it comes down to which device customers believe will collect the best data and which device will provide the best analytics. The data and analytics drive the purchase of the product. There may be a physical product involved, but what Nike is really selling, and what customers are really buying, is data and analytics.

Nike is transforming into a wearable technology and analytics consumer goods organization. Eventually, sensors will be found in shoes, gloves, shirts, and other Nike products. These products will work together to form a richer set of analytics for customers as well as for Nike. This is an important and fundamental shift. We now have a physical product that isn't purchased based on the attributes of the physical product itself. Nike recognizes this, and it is pivoting its business to embrace products of this nature. As traditional manufacturers suddenly find themselves embedding sensors, collecting data, and producing analytics for their customers, industry lines blur. Not only are new competencies needed, but the reason customers choose a product may have less to do with traditional selection criteria than with the data and analytics offered with the product.

It is focused here on a personal fitness product, but the same concept is playing out in other industries as well. Cars, airplanes, tractors, wind turbines, and trucks are all being embedded with sensors. Customers are beginning to use the data collected by those sensors for more and more purposes. As people decide which car model to buy, it may be a close race between two options. The final choice today could well depend on the data and analytics that are available from one automobile versus another. There is opportunity and there is risk in this shift to having analytics and data become the focus of a product rather than the physical product itself. But we can't view business as we have in the past, given the state of the world today. Data and analytics are most likely going to change a lot of things about our business.

Analytics will be transformative: - Some industries will be fundamentally transformed by all of the new data and new analytics generated. This is especially true for industries that historically have severely lacked both. While there are many possible examples to focus on, we focus here on one industry that is ripe for change: the education industry. We're still following a decades- or centuries-old model in education. We take children who just happen to be born around the same time and regardless of their background and skill level (with rare exceptions), we throw them all into a classroom together. Nine –year-olds in year four are going to cover a certain curriculum regardless of how well or poorly they are doing in school. Instead of moving away from this model, the education migrating toward enforcing ever more rigid rules about what kids learn during each year of school.

But in the age of big data and analytics, why don't we allow self-paced learning? Wouldn't school be more engaging if teachers became enablers who are there to answer questions and help students when they're stuck rather than reciters of mandated material? As students proceed through lessons at their own pace, they can ask the teacher for guidance at any time. There are already organizations, such as Khan Academy and Coursera, working to enable this approach. The way it works is that educational material is posted online for viewing. Then users watch the videos and take tests to verify that they have grasped the material. Why can't we use data and analytics to allow students to learn at their own pace all the time? Why can't students learn material from different grade levels every day? To complete year four, a student still will have to grasp the entire year four curriculum, but why can't a student be at a year five level in science coursework while still completing some of the year four classes for his or her history requirements? If a student learns all the required material at his or her own pace and can demonstrate to progress further, why should anyone care what route he or she takes or when the student was born?

Expect analytics to transform business models: - Some industries have already embraced analytics and changed how business is done, but others still look much as they did decades ago. The farther behind an industry is, the greater the potential for disruptive (but positive!) change to be achieved through the use of analytics. The key here is that data and analytics will enable this transition. From the above example, it is possible to monitor exactly which instructional videos each student watches, exactly which exercises each student completes, and how the student performs on each and every exercise and test question. Which areas does a student need to revisit? It is easy to tell because the analytics generated from the exercises can identify not just that a student is struggling in calculus but that he or she is struggling on topics related to one specific underlying concept. Since it is possible to quickly analyze every question a student has answered and identify the pattern that led to his or her performance on the test, the student can be guided to the right support material immediately. By collecting and analyzing data at a very detailed level, the analytics behind the scenes will help a student navigate the material in a way that provides freedom while still ensuring that all the necessary material is covered. The use of analytics to track and analyze student performance and progress at a new level may lead to education being one of the industries most disrupted by data and analytics in the coming years.

Big data for monitoring and anomaly detection

Big data can be helpful for people to know when unusual things happen or possibly when they are about to happen. This kind of notification can fall into two general categories although there are other systems for describing notification. These are monitoring and anomaly detection. The risk of making the differences between these two procedures sounds bigger than it is.

Monitoring can be helpful when we know what we are looking for and need a notification when that things occurs. It detects when specific event occurs. So we need to be able to specify the criterion or criteria in advance. For example a manufacturer needs to know when one of their machines needs maintenance. They may look at temperatures, vibration levels, or a number of factors that let them know that breakdown is imminent, and hence take care of it now. A doctor or nurse may need to know when one of their patients is sick. They may be monitoring for instance for temperature and pulses that it may be possible for white cells indicate infection. Similarly a credit card company may need to know when a charge is potentially fraudulent. In this case it may be possible for user to specify a specific criterion that they need in order to trigger the event. With the monitoring, we can be very specific and it even may be possible in certain situations to set up an automatic response that says for instance, if X occurs then Y results and we take care of it automatically. So monitoring is a specific thing, we know what we are looking for or waiting to happen and possibly even an automatic response.

Anomaly detection on the other hand can describe a situation in which the user wants to know when something unusual happens. We are looking here for notification of unusual activity without necessarily knowing in advance what that something might be. As a result it needs to be based on flexible criteria. It is similar to saying 'let me know when something that is out of ordinary happens', it may not just one factor but on a combination of several different factors. The flexible criteria usually exists to draw a person's attribute to something. For instance in security cameras, we do not know what is going on but we do know that something is out of the ordinary or in a stock trading situation they might say we do not know what is going on but it needs to be examined. It does not really trigger an automatic response but invites inspection. Similarly anomaly detection can notice patterns for instance too far spread apart in big data or may be too far for humans to notice on their own.

Both of these approaches, monitoring and anomaly detection are common practices in pre-date big data and computers as well. What big data adds to them though is the possibility to watch the extremely rare events or combination of factors. For instance if we have an event that only occurs one-in-million event, that can be really hard to spot if we are doing it by hand, even if we do it a hundred cases at a time. But if we have 10 billion cases, the one-in-million event is going to occur 10,000 times. And certainly that is not a small number. It is not so rare in fact 10,000 is pretty large number and allows us to do statistical modelling, or do sub-categories, figure out exactly what is associated with it and what is causing it. As to anomaly detection, big data's advantage is similar, especially when we look for every common notion of events. It may be possible to measure a thousand different things at once instead of just 10 or 12. This allows the machine learning algorithm that identifies anomaly in most of these cases to become much more specific and have better chance of detecting it and avoid false positive and false negatives.

To take a relatively trivial example, let's take email spam filters. Spam is a tricky situation because spam is a very fast evolving sort of virus like mechanism. It is never the same. It has to change all the time because there is these arms race between spam and spam filters. We cannot just write a single rule that says this is spam because spam will adopt to circumvent that rule. It is a very quick evolution. We cannot give clear rule for spam and it is very hard to do. What we find is that if we have a spam filter that looks only at our email and we say this is spam and this is not, we get a lot of false results – false positive and false negative. On the other hand, when we hook-up to a big data collection, when we're not categorizing spam just on our own, but when we combine data from many millions or hundreds of millions of users like for instance if we use Gmail, or Hotmail, or Yahoo; then it combines the collective wisdom of the crowd to determine whether it is spam or not.

So big data makes it possible to form these two kinds of watching – monitoring and more flexible anomaly detection with much greater power by being able to search for much larger datasets and to look for more diagnostic science at each point.

Data mining and text analytics

One of the most powerful and common application of big data is data mining and its close cousin text analytics. Data mining covers a large and diverse field of activities but the most basic idea is to use Statistical procedures to find unexpected patterns in data. Those patterns might include unexpected associations between variables or people who are clustered together in unanticipated ways [2]. For example, managers in a supermarket chain may find people who visited their stores in a particular region on a particular night of the week are generally different from people who came on other times and places. The market can then change to where coupons are displayed, or if at all, it can change where certain items are found from day to day to build on those differences. Or an investment company may find that when certain stocks move up together or certain others go down. Then a particular stock will generally fall low and that allows them to invest now and make a profit half way. Or medical researcher may find patients who exhibit a very particular pattern of symptoms of one kind. Even if they do not meet the criteria for a diagnosed illness, are more likely to check in to the hospital in the next six weeks for example. Because the database is so large and it is so easy to adopt the results for each specific viewer, perhaps the most common application of these kind of data mining is with online advertising. In fact that is one of the big promises of data mining – the ability to tailor services to the preference and behavior of each individual person once enough data has been gathered.

Text analytics is closely related to the standard kind of data mining that deals exclusively with numbers. However text analytics is sufficiently distinct to be its own field. The goal is to take the actual content of the text data such as twitter customer views and find meaning and pattern in the words. It is different from the meta-data example we have seen earlier. Because that prototype which can be shocking by informative, dealt with just numerical information that the computers created on their own, and do not even need to deal with the content of the information. When researchers look at the text itself, the interpretive and computational problem becomes really enormous. That is because human language is so flexible and solo where phrases that sound very similar can have different meaning. If it is that difficult for humans to understand immediately, it is nearly impossible for computers to understand and explains why there is a field called Natural Language processing (NLP). It has had so many challenges to overcome and evidence why NLP is such an active area of research.

Perhaps in text analytics the most common task is probably what is called sentiment analysis. It is to determine how people feel about something. That makes sense when we think from advertising or marketing point of view. The most basic task in sentiment analysis is to determine whether a person’s feelings are positive or negative. We definitely would like to know if people feel good or bad about our particular product. This is referred to as polarity in the text analytics world (positive or negative polarity). Fortunately, because this distinction is such a common task, there are many programs and packages that we can use such as Python and R that have been developed to help with text analytics.

Of course sentiment analysis and text analytics are generally much more sophisticated than just good or bad, but that is the basic idea. There is much more that can be said, but data mining and text analytics work best when they have very large and diverse dataset to work with and that is what big data does. As researchers continue to develop and refine methods for data mining and text analytics, the ability to find patterns in numerical data and meaning in textual data will become faster and simpler.

Big data predictive analytics

Predictive Analysis is the crystal ball of big data. It represents a range of techniques that are adopted to work with data to try to predict future events based on past observations. Ever since there has been people what have been tried to predict the future, the raw resources of big data and the sophistication of modern predictive modelling have fundamentally changed the way we look into the future. In the popular world there are few well known examples of predictive analytics.

The first is in the baseball that is shown in the book money ball where statistical analysis is used for instance to identify an effective player’s scoring ability [14]. And the standard criteria that has been set by people for hundred years in baseball is to look at things like batting average, RBIs (Run Batted In), and stolen bases. And what happens is baseball has an enormous states and it is very easy to count the discrete events that occur. We can go back and deal with an extraordinary large dataset for sport. Researchers found out that now the batting average in the RBIs are not the best predictors but on Pythagorean expectation which is a formula invented by Bill James to estimate how many games a baseball team "should" have won based on the number of runs they scored and allowed. Comparing a team's actual and Pythagorean winning percentage can be used to evaluate how lucky that team was (by examining the variation between the two winning percentages). The name comes from the formula's resemblance to the Pythagorean Theorem.

The second example is from Note Silver (of Five Thirty Eight) remarkable accuracy predicting results for every single state in the 2012 US presidential elections (http://fivethirtyeight.com/). Now what Note did was that he had the blog called ‘FiveThirtyEight’ which has to do with the number of representatives in US Congress. He took data from wide range of polls and combined that and weighed them by the reliability and he was able to come up with an accurate prediction for every single state in the election. It was remarkable.

The next example to look at is Netflix prize. This is the case where Netflix provided a million dollar prize to anybody who could improve the quality of their recommendations by 10% using anonymized dataset that they provided. What happened was that there was a really remarkable statistical analysis that came out of it. Perhaps the biggest that came out from the Netflix prize was the efficacy of what is called ‘Ensemble Models’. The idea is we do not need to build a single predictive model and do not have to try to say this is our regression equation or this is our random force model to predict. We build as many predictive models as we possibly can and basically average the result of all of them. It turns out that when it comes to predictions, the average prediction is usually more accurate than any one individual prediction. It actually comes about similar approach to guessing the number of jellybeans in a large jar. If we take everybody’s prediction and average them that is usually going to be closer to the real number than any one’s individual actual guess.

Predictive analytics has an enormous areas of interest because especially if we are in a business and trying to predict what is going to happen, then having a little bit of formal knowledge can get us a huge competitive advantage. It is an area of incredible growth and it really is one of the most fascinating thing about statistics because there is always a very clear criterion which is something that is often lacking. We can tell whether our model is good or not and the progress in the field makes it possible to learn more and more. Especially with the raw material from big data, there is so much more to work with refining models and get better predictive ability for better competitive advantage.

Data visualization

Up to this point we have been talking about big data and the things that the computers are able to do for the humans. On the other hand it turns out that there are certain things that humans still do better than computers and visualization is one of them. Humans are visual animals. We work on sight and get a huge amount of information in that way. Computers are very good in spotting certain patterns. They are also good at calculating and predictive models and doing data mining in a way that human's would have hard time doing in their thousand life times. But humans perceive and interpret patterns much better than computers do. Hence human vision still play an important role in big data. Humans can see the patterns and they can see the exceptions to the patterns of the anomalies very quickly. They can also see those patterns across multiple variables and groups. And also humans are much better at interpreting the content of images than computers are.

The most important things to be warned about visualization are that prettier graphs are not always better. Also in many situations animated or interactive graphs can be distracting even though they can be more informative. And the goal of data visualization in any kind of graphics is insight. We want to get to the insight as clearly as quickly as possible, and anything that distracts from that or gives a wrong impression is a mistake and has to be illuminated. Therefore, data visualization is still an area where humans make an important contribution into big data analysis and computers can contribute all of the other models discussed. It is important to remember human element when planning big data project. There is still a need for the human perception and interpretation to make sense of the data in addition to what the computer is able to provide.

Summary

While analytics have become more pervasive, many people have not realized the extent to which analytics are now fundamentally changing business models. An organization will quite possibly have to think bigger and more boldly about how analytics fits within its future.

Organization needs to start taming big data now. As of today, we've only missed the chance to be on the bleeding edge if we've ignored big data. Today, we can still get ahead of the pack. In a few years, we'll be left behind if we are still sitting on the side-lines. If our organization is already committed to capturing data and using analysis to make decisions, then going after big data isn't a stretch. It is simply an extension of what we are already doing today. Enabling innovative analytics requires effort. It will take a concerted, focused effort to do so. Putting effort toward innovation in analytics needs to get similar attention to developing product and service innovations. Analytics should be considered a pillar of a business, and not an optional add-on.

References

1. *Alex Holmes (2012). Hadoop in Practice. Manning Publications Co.*
2. *Carothers, B. J., & Reis, H. T. (2012). Men and Women Are From Earth: Examining the Latent Structure of Gender. Journal of Personality and Social Psychology. Advance online publication.*
3. *Daniel Minoli (2013). Building the internet of things (IoT) with IPv6 and MIPv6. John Wiley & Sons, Inc.*
4. *David Lazer, et.al. (2013). The Parable of Google Flu: Traps in Big Data Analysis.*
5. *Donald Miner and Adam Shook (2012). MapReduce Design Patterns, O'Reilly Media, Inc.*
6. *Donald D. Chamberlin & Raymond F. Boyce, SEQUEL: Structured English Query Language, IBM Research Laboratory, California.*
7. *Evan Stubbs (2014). Big data, big innovation: enabling competitive differentiation through business analytics. John Wiley & Sons, Inc., Hoboken, New Jersey.*
8. *Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics,* ***7****, Part II, 179–188.*
9. *Franks, Bill (2012). Taming the big data tidal wave: finding opportunities in huge data streams with advanced analytics. John Wiley & Sons, Inc., Hoboken, New Jersey.*
10. *Garry Tarkington (2013). Hadoop Beginner's Guide. Packet Publishing.*
11. *George Reese (2009). Cloud Application Architectures. O'Reilly Media, Inc.*
12. *Jared Dean (2014). Big data, data mining, and machine learning: value creation for business leaders and practitioners, John Wiley & Sons, Inc., Hoboken, New Jersey.*

13. *Matthew A. Russell (2014). Mining the Social Web. O'Reilly Media, Inc., USA.*
14. *Michael Lewis, Money ball: The Art of Winning an Unfair Game.*
15. *Michael Minnelli, Michele Chambers, Ambiga Dhiraj (2013). Big data, big analytics: emerging business intelligence and analytic trends for today's businesses John Wiley & Sons, Inc., Hoboken, New Jersey.*
16. *Nina Zumel & John Mount (2014). Practical Data Science with R. Manning Publications Co.*
17. *Reilly Media (2012). Big Data Now. O'Reilly Media, Inc.*
18. *Robert Elsenpeter, Anthony T. Velte, Toby J. Velte (2010). Cloud Computing: A Practical Approach. The McGraw-Hill Companies.*
19. *Rodney Heisterberg & Alakh Verma (2014) Creating Business Agility. John Wiley & Sons, Inc., Hoboken, New Jersey*
20. *Syed A. Ahson and Mohammad Ilyas (2011) Cloud Computing and Software Services Theory and Techniques. Taylor and Francis Group, LLC.*
21. *Tom White (2012). Hadoop: The Definitive Guide. O'Reilly Media, Inc.*