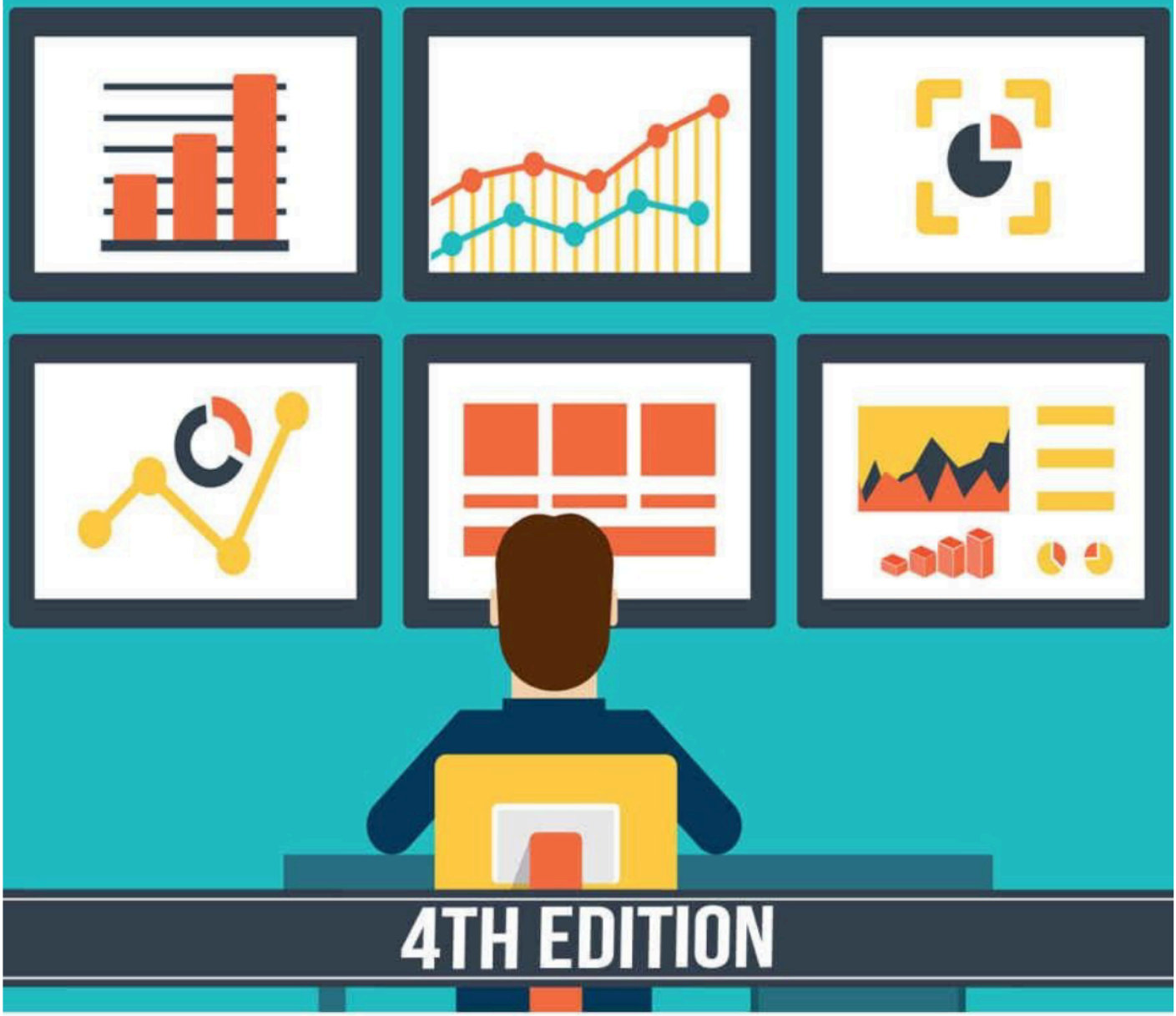
ANALYTICS

DATA SCIENCE, DATA ANALYSIS AND PREDICTIVE ANALYTICS FOR BUSINESS



4TH EDITION

Daniel Covington

Analytics

Data Science, Data Analysis and Predictive Analytics for Business 4th Edition

By Daniel Covington

Table of Contents

Introduction

What defines the success of a business? Is it the number of people employed by the firm? Is it the sales turnover of the business? Or is it the strength of the customer base of the firm? Does employee satisfaction play a role in the success of business operations? How does management factor into the overall operational success? How critical is the role of the data scientist in this process? Finally does fiscal responsibility play any role in the success of any business?

To answer any of the aforesaid questions, it is important that you have the required data in hand. For instance, you need to know how many staff you have employed first to assess the value contributed by them to the growth of your business. Similarly, you need to have a repository of all the customers along with details of their transactions to understand if your customer base is contributing to the success of your firm.

Why is data important? Data is important for the sustenance of a business. The most preliminary reason why it is important is because you need information to be aware of the state of affairs. For instance, if you don't know how many units your company sells, in a month, you will never be able to state if your business is doing well or not. There are several other reasons as to why data is important. I have dealt in detail in the upcoming chapters of this book regarding the importance of data.

Just collecting data is not enough. Analyzing and putting it to use is important. Of course, if you belong to the class of people who are not worried if their business lost a customer or not, you don't have to spend time on analyzing data. However, if this attitude persists, you will soon see the end of your business, with the growing competitors who care about the expectations of their customers. Hence, this is where predictive analytics comes into play. How you employ predictive analytics in your firm is what distinguishes your firm from the others in the market. Predictive analytics has the capacity to change the way you play your game in the market. It is capable of giving you that little edge over your competitor.

In the first chapter of this book, I have highlighted the importance of data in business. I have also highlighted how data plays an important role in increasing the efficiency of a business. In the second chapter of this book, I have mentioned the different steps involved in the process of data analysis. In the third chapter of this book, I will be taking you through the basics of predictive analytics and the various methods involved. I have mentioned the different techniques used for conducting predictive analytics in the fourth chapter of this book. In the final chapter of this book, you will see how predictive analytics is being employed in different fields. You will truly appreciate its flexibility,

seeing how it can be used in finance as well as in medicine, in operations as well as in marketing.

The fields of data analysis and predictive analytics are so vast and there are so many sub branches to these fields, which are even more extensive. One of which is prescriptive analysis, which, I will briefly cover in the last chapter. I have covered only the fundamentals of these fields in this book. It is being used by large industries to determine what will happen in future and using that data to prevent or make things happen in the future.

I hope you truly enjoy this book. I am sure you will be motivated to manage your data better and employ predictive analytics in your business and reap the maximum benefits. Thank you for purchasing this book.

Chapter 1: The Importance of Data in Business

Were you ever involved or interested in the study of the classic languages? Languages that are otherwise known as "dead" languages? While the sheer complexity of these languages is fascinating, the best part about them, the part that will simply blow your mind is that the ancients used some extremely tedious methods to preserve sacred texts that range from just a few hundred years old to several thousand. To preserve these texts, scribes would painstakingly copy them, not just once but several times and this could take years to complete. They copied the texts onto papyrus paper and used ink made from oil, burnt wood and water. Another common method was to chisel the characters onto shards of pottery or onto stone. While these were exceedingly time intensive and, I have no doubt extremely mind numbing at times, the information that has been preserved was deemed to be so valuable to the ancient civilizations, and certain people dedicated their whole lives to doing it.

So, I can hear you asking, what on earth do ancient texts and dead languages have to do with data analysis and business analytics? Data, everything is about the data; it is something that is all around us and we simply can't get enough of it. Think about today; think about social media platforms, platforms that are revolutionizing the landscape for marketing because they are providing companies with a whole set of analytics that allow them to measure how successful, or not as the case may be, their company content is and many of these platforms provide these analytic tools for free.

On the other hand, there are platforms that charge a high price to provide you with high quality data, telling you what does and doesn't work on your website. In the world of business, market and product data gives a business the edge over their competitors and that makes it worth its weight in gold. Important data includes historical events, weather, products, trends, customer tendencies, outliers and anything else that is relevant to business verticals.

What has changed is the way that data is stored. It is no longer a long and cumbersome task; it is automatic, requires little in the way of human intervention and it is done on a huge scale. Today's modern day scribes are connected sensors.

If you think about it, if you were to collect every single piece of information that you could, every piece of data that society generated, it would be nothing compared to what you will face in the next few years. This is the Internet of Things. This is because most of our devices are connected; they record performance and usage data, they transmit that data. Sensors record

environmental data. Cities are completely connected to ensure traffic and infrastructures are running at top level. Delivery vehicles are connected so their location is monitored, their efficiency, even to the extent that a problem with the vehicle can be identified early. Buildings and homes are connected to improve the cost and control of energy, while manufacturing premises are connected so that the communication of critical data is automatic. This is the present and ever more so, it is the future. We are connected; everything is connected to the internet.

The fact that data is important really isn't anything new. It's just that we have moved past a scribe and a chisel to using microprocessors. The way we capture data and the type of data we capture is ever hanging and it is vital that you stay up to date and ahead of the game if you want to be in the game to win.

Even though it is the service provided or the goods manufactured by a company, which helps it in establishing a niche for itself in the market, data plays a crucial role in sustaining the success. In today's technology driven world, information can make or break a business. For instance, there are businesses that have disappeared in such a short time, because they failed to gauge their progress or customer base. On the other hand, we also have start ups that have been doing extremely well because of the increased importance they show towards numbers and the expectations of their customer base.

What Is The Source Of Data?

By data, it could refer to the sales figures or the feedback from the customers or the demand for the product or service. Some of the sources of data for a company are as follows:

Transactional data:

This could be pulled out from your ledger, sales reports and web payment transactions. If you have a customer relationship management system in place, you will also be able to take stock of how your customers are spending on your products.

Online engagement reporting:

This is data pulled out based on the interaction of customers on your website. There are tools available such as Crazy egg and Google analytics, which can help you to collect data from your website.

Social media:

Social networking sites such as Twitter, Facebook and LinkedIn also provide insights on the customer traffic on your page. You can also use these platforms to conduct a cost effective survey on the tastes and preferences of the customers and use it to improve their products or services.

How Can Data Improve Your Business?

Data can improve the efficiency of your business in many ways. Here is a taste of how data can play an important role in upping your game.

Improving your marketing strategies:

Based on the data collected, it is easier for the company to come up with innovative and attractive marketing strategies. It is easier for the company to alter existing marketing strategies and policies in such a fashion that it is in line with the current trends and customer expectations.

Identifying pain points

If your business is driven by predetermined processes and patterns, then data can help you identify any deviations from the usual. These small deviations could be the reason behind the sudden decrease in sales or increase in customer complaints or decrease in productivity. With the help of data, you will be able to catch these little mishaps early and take corrective actions.

Detecting fraud

When you have the numbers in hand, it will be easier for you to detect any fraud that is being committed. For instance, when you have the purchase invoice of 100 units of tins and when you see, from your sales reports, that only 90 tins have been sold out and you are missing ten tins from your inventory, you know where to look. Most companies are being silent victims of fraud because they are not aware of the fraud being committed in the first place. One important reason for that is the absence of proper data management, which could have helped them detect fraud easily in the early stages.

Identify Data Breaches

The explosion of complex data streams in the past few years has brought on a new set of problems where fraudulent practices are concerned. They have become more subtle and comprehensive. Their negative effects can be widespread and impact your company's retail, accounting, payroll and other business systems adversely. In other words, data hackers will become more devious in their attack on company data systems.

Using data analytics and triggers, your company can prevent fraudulent data compromises in your systems, which can severely cripple your

business. Data analytics tools allow your company to develop data testing processes to detect early signs of fraudulent activity in your data systems. Standard fraud testing may not be feasible in certain circumstances. If this is the case with your company, then special, tailored tests can be developed and used to trace any possible fraudulent activity in your data processes.

Traditionally companies have waited until their operations have been impacted financially to investigate fraud and breach prevention strategies. This is no longer feasible in today's rapidly changing data-saturated world. With information being disseminated worldwide so quickly, undetected fraudulent activity can cripple a company and its subsidiaries in no time globally.

Conversely, data analytics testing can equally stop potential fraudulent data destruction by revealing indicators that fraud has begun to seep into data systems. Fraud can be stopped quickly for a company and its partners worldwide if these data analytic tests are put into application periodically.

Improving customer experience

As I mentioned before, data also includes the feedback provided by customers. Based on their feedback, you will be able to work on areas, which can help you improve the quality of your product or service and thereby satisfy the customer. Similarly, when you have a repository of customer feedback, you will be able to customize your product or service in a better fashion. For instance, there are companies, which send out customized personal emails to the customers. This just sends out a message that the company genuinely cares about its customers and would like to satisfy them. This is possible solely because of the effective data management.

Decision making

Data is very crucial for making important business decisions. For instance, if you want to launch a new product in the market, it is important that you first collect data about the current trends in the market, the size of the consumer base, the pricing of the competitors etc. If the decisions taken by the company are not driven by data, then it could cost the company a lot. For instance, if your company decides to launch a product, without taking into consideration the price of the competitor's product, then there is a possibility that your product might be overpriced. As is the case with most overpriced products, the company would have trouble in increasing the sales figures.

When I say decisions, I don't really just refer to the decisions pertaining to the product or service offered by the company. Data can also be useful in taking decisions with respect to the functioning of the departments, manpower management etc. For instance, data can help you assess how many personnel are required for the effective functioning of a department, in line with the business requirements. This information can help you to decide if a certain department is overstaffed or understaffed.

Hiring Process

Using data in selecting the right personnel seems to be a neglected practice in corporate life. It's critical to position the most qualified person in the right job in your company. You want your business to be highly successful in every facet of operation. Using data to hire the right person is a sure way to put the best person in the job. What kind of data would you use to hire a professional?

Big companies, with astronomical budgets, use big data to locate and select the most skilled individuals for the right jobs. Start-ups and small companies would benefit immensely from using big data to hire the right group of people to make their recruiting successful from the start. This avenue for gathering data for hiring purposes has proved to be a successful avenue for hiring the right fit for organizations of all sizes. Again companies can use their data scientists to extract and interpret the precise data for human resource departments.

Using Social Media platforms to recruit

Social Media platforms (Facebook, Twitter, and LinkedIn to name some) are hidden gold mines as data sources for finding high profile candidates for the right positions within companies. Take Twitter, for instance, company job recruiters can follow people who tweet specifically about their industry. Through this process, a company can find out and recruit the ideal candidates by their knowledge of a particular industry or job within that industry. Do their "tweets" inspire new thoughts, and possibly new innovations for their industry? If so, you have a whole pool of potential job applicants.

Facebook is another option for data gathering for potential job candidates. Remember, these avenues are virtually free which corporations could use as major cost effective strategies. Facebook is all about collecting social networking data for companies looking to expand their workforce or replace an existing open position(s). Company recruiters can join industry niche groups or niche job groups." Liking" and following group

member's comments will establish the company's presence within the group allowing highly focused job ads to be posted within the group. The company can increase views widening the pool of potential job candidates.

It is easy to establish a timeline to brand the company as an innovative and cutting-edge place to work. You establish your presence by engaging with friends/followers who are in the same industry as your company. For a minimal advertising fee promote your job ad. By doing this you geometrically increase your reach to potential job seekers. If your company puts out highly effective job data posts you will reel in a higher yield of highly skilled job searchers, therefore greatly increasing the percentages of getting the perfect fit for your job.

Niche social groups

Niche socials groups are specialized groups that you can join on social and web platforms that can help you find specific skillsets. For example, if you are looking to hire a Human Resources manager. What better place to find a potential recruit than by finding a specific Human Resources social group? Locate social connections within that group and post descriptive but alluring job data posts and you may find the right person to fit into your Human Resources position. Even if your company doesn't find the right person there, those group members surely have referrals. Again, approaching these groups is a very cost effective way to advertise your job data posts.

Innovative data gathering methods for the hiring process

Why not think outside the hiring process box and try new methods of data collection to hire the right professional? Use social collecting data sites that collect data from sites like Facebook, Google+, LinkedIn and Twitter. Your company can search on these sites extracting pertinent data from posts made by potential job candidates. This data can be used to help your company connect with highly efficient job applicants.

An overlooked, but very good data pool to use is keywords. Keywords are used on the internet for every type of search imaginable Why not use the most visible keywords in your online job descriptions? By doing this your company can widely increase the number of views your job posting will attract.

You can also use computers and software to find the right job candidate for your company. Traditionally, these data sources have been used to

either terminate a company's employee or analyze if an existing employee is a right fit for another job within the company.

Why not try a whole new data collecting system? This system would be rooted in a different set of standards than the usual IQ, tests, skill testing or physical exams. These are still valuable tools to measure candidates by but they are limiting. Another focus can be the strong personality traits a candidate may possess. Is the person negative? Is the person argumentative? Is the person an isolationist that doesn't get along with other people? These types of people can be identified in this personality trait database and be filtered out as possible team members. This type of data properly extrapolated will save the company time, resources and training materials. By eliminating a mismatch between the expectations the company requires for the job and the person who would potentially fill the job.

Another benefit of this data gathering system is the results will not only identify skilled people for the right jobs, but people with the right personality to fit in with the current company culture. It's imperative that a person is sociable and will be able to engage other employees to produce the most effective working relationships. The healthier the working environment the more effective company production is over all.

Gamification

This is a unique data tool that isn't currently in widespread use. It does motivate candidates to press in and put forth their best effort in the job selection process. You provide the person with "badges" and other virtual goods that will motivate them to persevere through the process. In turn their skills in being able to perform the job requirements will be readily obvious. This also makes the job application a fun experience instead of a typical tedious task.

Job Previews

Pre-planning the job hiring process with accurate data about what the job requirements are will prepare the job seeker to know what to expect if he is hired for the position. It is said a lot of learning on the job is by trial and error, which prolongs the learning process. This takes more time to get the employee up to speed to efficiently function as a valuable resource within the company. Incorporating the job preview data into the hiring process reduces the learning curve and helps the employee become efficient that much quicker.

These are some innovative data gathering methods companies can use to

streamline the hiring process. They also help Human Resource departments pick the most skilled people to fill their employment needs.

Hence, data is very crucial in aiding businesses to take effective decisions.

These are some of the reasons why data is crucial for the effective functioning of a business. Now that we have had a glance at the importance of data, let us get into the other aspects of data analysis in the upcoming chapters.

Chapter 2: Big Data

Data is firmly woven into the fabric of society, across the entire globe and, like every other important production factor, like human capital and hard assets, much of our modern economic activity could not happen without it. Big data is, in a nutshell, large amounts of data that can be gathered up and analyzed to see if any patterns emerge and to make better decisions. In the very near future, big data will become the base on which companies compete and grow, the base on which productivity will be strongly enhanced and significant value will be created for the global economy by increasing the quality of services and products and reducing the amount of waste.

Until now, the river of data that has been flooding the world was something that most likely only grabbed the excitement of some data geeks. Now, we are all excited by it simply because the sheer amount of data that is generated, mined and stored has become one of the most relevant factors, in an economic sense, for consumers, governments and businesses alike.

Looking back, we can see now that trends in IT innovation and investment, and their impact on productivity and competitiveness suggest that big data has the ability to make sweeping changes to our lives. Big data has the same preconditions that allowed previous IT-enabled innovation to power productivity (for example, innovations in technology, followed closed by complementary management innovations being adopted). We expect that big data technology suppliers and analytic capabilities that are now far advanced will have as much, if not more impact on productivity as do suppliers of different technologies.

Every business in the world needs to take big data very seriously as it has a huge potential to create real value. Some retail companies that are wholly embracing big data are seeing the potential for a 0% increase in operating margins

Big Data – The New Advantage

Big data is fast becoming the most important way for those companies at the top to seriously outperform their competition. In most industries, both new entrants to the market and those that are firmly established will leverage strategies that are driven by data to compete, to innovate and to capture real value. Examples of this exist everywhere. Take the healthcare industry; data pioneers are examining the outcomes of certain pharmaceuticals that are widely prescribed. During their analysis, they discovered that there were risks and benefits that were not seen during limited trials.

Other industries that have taken on big data and run with it are using sensors that are embedded into products to gain data. These products range from a child's toy to large-scale industrial goods and the data tells them how the products are used in the real world. This kind of knowledge allows for better design of these products in the future.

Big data will assist with creating new opportunities for growth and it will also help to create entirely new categories of companies, i.e. those that aggregate industry data and analyze it. A good proportion of these companies will be situated in the center of large flows of information, where the data that come from products, buyers, services and suppliers can be analyzed. Now is the time for forward thinking managers and company leaders to build up their company capabilities for big data and they need to be aggressive about it.

The scale of big data isn't the only important factor; we must also take into account the high frequency and the real-time nature of the data as well. Take "nowcasting" for example. This is the process of estimating metrics, like consumer confidence, straight away – something that at one time could only be done after the fact. This is being used more extensively, adding a considerable potential to prediction. In a similar way, high frequency allows users to analyze and test theories to a level that have before been possible.

Studies of major industries have shown that there are a few ways that big data can be leveraged:

1. **Big data has the potential to unlock some serious value for industries by making all information transparent. There is still a lot of data that has not yet been captured and stored in digital form or that cannot easily be found when searched for. Knowledge workers are expending up to 25% of their time and effort in searching for specific data and then transferring it to another location, sometimes a virtual one. This percentage represents a vast amount of inefficiency.**
2. **As more and more companies store their transactional data in a digital format, they are able to collect detailed and highly accurate performance information on just about everything – from inventory**

right down to the amount of ick days being taken – thus giving themselves the ability to boost performance and root out variabilities. Some of the leading companies use this ability to collect big data and analyze it to carry out experiments to see how they can make more informed management decisions.

3. **Big data allows companies to divide their customers down into smaller segments, allowing them to better and more precisely tailor the services and products that they offer.**
4. **More sophisticated analytic allows for much better decision making. It can also cut down the risks, and bring to light some very valuable insights that might otherwise never see the light of day.**
5. **We can use big data to create the next generation of services and products. For example, manufacturers are already using the data that they get from their embedded sensors to come up with more innovative after-sales service.**

Big Data Creates Value

Let's use the US healthcare system as an example here. If they were to make effective and creative use of big data to improve quality and efficiency, they could actually create in excess of $300 billion of value every single year. Around 70% of that would be from a cut in healthcare expenditure of around 8%.

Moving on to the European developed economies; if government administrates used big data in the right way, they could create improvements in operational efficiency of around €100 billion every year. And that is just in one area; we haven't looked at what they could achieve if they used advanced analytic tools to boost tax revenue collection and cut down on errors and fraud.

It isn't just organizations or companies that and gain benefit from using big data. The consumer can benefit significantly, i.e. those who use services that are enabled by location data can realize consumer surplus of up to $600 billion.

Take smart routing that uses real-time traffic information, for example. This is one of the most used of all the applications that use location data and, as more and more people use smartphones and as more and more of those people take advantage of the free maps apps provided, smart routing use is likely to grow significantly. By the year 2020, it is expected that more than 70% of all mobile phones will have GPS capability built in, way more than the 20% recorded in 2010. As such, we can estimate that, by 2020, the global value of smart routing has the potential to create savings of around $500 billion in fuel and time savings. This is the equivalent of cutting 20 billion driving hours or saving a driver around 15 hours a year on the road and savings of around $150 billion in fuel.

The most potential for value from big data is going to come from the combination of data pools. Again, the US healthcare system has four significant data pools – activity and cost, clinical, medical and pharmaceutical products R & D and patient data. Each of these data pools is captured and then managed by a different part of the healthcare system but if big data were used to its full potential, annual productivity could be increased by around 0.7%. This would require the combination of data from all the different sources, including from those organizations who don't share data at scale. Data sets would need to be integrated, for example clinical claims and patient records.

Doing this would realize benefits for everyone, from the industry payers to the patients. The patients themselves would have better access to more healthcare information and they would be able to compare process of physicians, treatment and drugs. They could compare effectiveness, allowing them to pick the medication and treatment that suited them the best. To be able to take advantage of these benefits though, they would have to accept a tradeoff between the

benefits and their privacy.

Data security and privacy are two of the biggest hurdles that must be faced if the benefits of big data are to be truly realized. The most pressing challenge is the huge shortage of people who have the skills needed to be able to analyze big data properly. By 2018, the US is facing a shortage of between 140,000 to 190,000 people with the right training in deep analysis and another shortage of around 1.5 million people with the quantative and managerial skills needed to interpret the analyses correctly so that they can base their decisions on them.

There are also a whole slew of technological issues, which will have to be resolved. Incompatible formats and standards as well as legacy systems often stand in the way of data integration and stop sophisticated analytical tools being used. Ultimately, in order to make full use of the larger digital datasets is going to require a technology stack being assembled from computing and storage right through to the application of visualization and analytical software.

Above all, to take true advantage of big data, access to all data has to be widened. More and more, organizations will need to have access to data stored with third parties, for example with customers or business partners, and then integrate that data with theirs. One of the most important competencies in the future for data-driven companies will be the ability to come up with compelling value propositions for other parties, like suppliers, consumers and possibly even competitors to a certain extent.

For as long as the true power of big data is understood by governments and companies, the power it has to deliver better productivity, more value for the consumer and the power it has to fuel the net wave of global economy growth, there should be some incentive for them to take the necessary actions to overcome the barrier that stand in their way. By doing this, they will open up new avenues of competitiveness among industries and individual companies. They will create a much higher level of efficiency in public sectors that will allow for better services, even when money is short and they will enable more productivity across the board.

Big Data Brings Value To Businesses Worldwide

The value Big Data has brought to businesses worldwide and will is immeasurable..

Here is just a brief summary of ways that it has impacted our world:

- Created a whole new career field-Data Science
- Big data has revolutionized the way data interpretation is applied
- The healthcare industry has been improved considerably by the

application of predictive analytics

- Laser scanning technology has changed the way Law Enforcement reconstructs crime scenes
- Predictive analytics is changing the roles of caregivers and patients
- Data models can now be built to investigate and solve many business problems
- Predictive analytics is changing the way the Real Estate industry is conducting business

Big Data = Big Deal

One more thing that big data could do is open the way for new management principles. In the early days of professional management, corporate leaders found that one of the key determining factors of competitive success was a minimum scale of efficiency. In the same way, the competitive benefits if the futures are likely to build up with companies that can capture more data that is of higher quality and use that data at scale with more efficiency.

The following five questions are designed to help company executives determine and recognize just how big data can benefit them and their organizations:

- **What will happen in a "transparent" world, with data available so widely?**

Information is becoming more and more accessible across all sectors and, as such, it has the potential to threaten those organizations that rely heavily on data as a competitive asset. Take the real-estate industry, for example. They trade on access to transaction data and a secretive knowledge of bids and buyer behaviors. To gain both requires a great deal of expense and a lot of effort. However, recent years have shown that online specialists have begun bypassing the agents and allowing buyers and sellers to exchange their own perspectives on property values and have created a parallel resource for real estate data to be garnered from.

Pricing and cost data have also become more widely accessible across a whole slew of industries. Some companies are now assembling satellite imagery that is readily available; imagery that, when it's processed and then analyzed can provide clues about the physical facilities of their competitors. This gives them ideas about the expansion plans or any constraints that their competitors are coming up against.

One of the biggest challenges is that much of the data that is being amassed is being kept in departmental "silos", such as engineering, R & D, service operations or manufacturing. This stop the data being exploited in a timely manner and can also cause other problems. Financial institutions, for example, suffer because they don't share data among the diverse business lines, such as money management, financial markets or lending. This can stop these companies from coming up with a coherent view of their customers, on an individual basis, or of having an understanding of the links between the financial markets.

Some manufacturing companies are trying their hardest to get into these enclaves. They integrate data from a number of different systems, inviting

formerly closed off units to collaborate and looking for information from external parties, such as customers and suppliers, to help co-create future products. In the automotive industry, global suppliers make hundreds of thousands of different components. Integrating their data better would allow the companies, along with their supply chain partners, to collaborate at the design stage, something that is crucial to the final cost of manufacturing.

- **Would being able to test your decisions change the way a company competes?**

Big data brings in the possibility of a different style of decision-making. Through the use of controlled experiments, organizations will be able to test out hypotheses and then analyze the results. This should give them results that they can use to guide their decisions about operational changes and investments. Effectively, experimentation is going to allow managers to distinguish between correlation and causation, cutting down on the wide variations in outcomes while boosting products and financial performance.

These experiments can take a number of different forms. Some of the top companies online are always testing and running experiments in some cases, they will set aside a specific part of their web page views in order to test which factors drive sales or higher usage. Companies who sell actual physical products use tests to help them make decisions but using big data can take these experiments to a whole new level. McDonalds has fitted devices into some of its stores to track customer interaction, the amount of traffic and patterns in ordering. Using the data gained, they can make decisions on changes to their menus, changes to the design of their restaurants and training on sales and productivity, among other things.

Where it isn't possible to carry out controlled experiments, a company could use a natural experiment to find where the variables in performance are. One government sector collected data on a number of different groups of employees who were all working at different sites but doing similar jobs. Just by making that data available, the workers who were lagging behind were pushed to improve performance.

- **If big data were used for real-time customization, what effect would it have on the business?**

Companies who face the public, as it were, have been using data to divide and target specific customers for a long time. Big data takes that much further that what always used to be considered as top of the range targeting, by making it possible to use real-time personalization. In the future, retailers will be able to keep track of individual customers and their behaviors, by monitoring internet

click streams, and will be able to make changes to preferences and model the potential behavior in real-time. By doing this, it will be easier for them to know when a customer is heading towards making a decision on a purchase; they will be able to "nudge" the decision and take the transaction through to completion by bundling together products, and offering benefits and reward programs. This s called real-time targeting and it will also bring in data from loyalty programs and this, in turn, can help to increase the potential for higher-end purchases by the most valuable customers.

Retailing is probably the most likely of all industries to be driven by data. They have vast amounts of data at their fingertips – from purchases made on the internet, conversations on social network sites and from location data from smartphone interactions, alongside the birth of new, better analytical tools that can be used to divide customers down into even smaller segments for better targeting.

- **How will big data help management, or even replace it?**

Big data opens up more avenues for the application of algorithms and analysis that is mediated by machines. Some manufacturers use algorithms to analyze data garnered from sensors on the production line. This helps them to regulate their processes, reducing waste, avoiding human intervention that can be expensive and, in some cases, dangerous, and increasing their output. In the more advanced "digital oilfields", sensors are used to constantly monitor the conditions of the wellheads, the pipelines and all the mechanical systems. That data is then analyzed by computers and the results are fed to operation centers where, in real-time, they adjust the oil flows to boost production and reduce downtime. One of the biggest oil companies in the world has managed to increase their oil production by 5% while reducing staffing and operating cost by between 10% and 25%.

Products that run from a simple photocopier to a complex jet engine are now able to generate streams of data that track usage. Manufacturers are able to analyze the data and, in a few cases, fix glitches in software or send out repair representatives automatically. In some cases, the data is being used to preempt failure by scheduling repairs to take place before systems can go down.

The bottom line is big data can be responsible for huge improvements in performance, much better risk management and uncover insights that would most likely never have been found. On top of that, the price of analytic software, communications devices and sensors are falling fast and that means more companies will be able to afford to join in.

- **Can big data be used to create a brand new business model?**

Big data is responsible for coming up with brand new company categories that embrace business models driven by data. A good many of these companies are intermediates in a value chain where they can generate valuable “exhaust” data that is produced by transactions. A major transport company saw that, while they were going about their transport business, they were gathering in large amounts of data on product shipments across the globe Sensing a top opportunity, they came up with a nit that now sells the data gained to supplement economic and business forecasts.

Another major global company learned a great deal by analyzing their own data and decided that they should create a new business to the same sort of work for other organizations. They now aggregate supply chain and shop floor data for a large number of manufacturers and sells the relevant software tools needed to improve performance. This part of the business is now outperforming their main manufacturing business and that is all thanks to big data.

Big Data has created a whole new support model for existing markets. With all the new data needs for companies the increased demand for qualified people to support that data is increasing.

As a result your business may need an outside firm to analyze and interpret the data for you. These are specialized firms that that can help interpret and analyze the data. These companies would specialize in assimilating large amounts of data both in structured and unstructured models. These companies will exist solely for the purpose of supporting leading companies in whatever industry. Their employee base would be trained in looking for and gathering data in systems and processes. These data analysis companies support your business in doing this function for you and charge for their expertise.

They would have the resources and training to assimilate, analyze and interpret new trends in the data and be required to report back to you any alerts.

A viable alternative would be for existing companies to employ data support departments within their existing infrastructure. This method would be much more cost effective then using an outside firm, but will require specialized skillsets within your company. This could be the next step from your current information technology group. The data support analysts only focus on data in the existing information systems. Their focus would be deciphering data flow by analyzing, interpreting and actually finding new applications for this data. These new applications and support team would monitor existing data for any fraud, triggers or issues that it might present.

Big data has created a whole new field of study for colleges and higher institutions of learning to offer in their degree programs. People are being trained in the latest methods of big data gathering, analyzing and interpreting. This career path will lead them to critical positions in the newly trending data support companies. Big data has not only created a new industry but a whole

new field(s) of study. Along educational lines, big data will change the way teachers are hired. Big data recruiting processes combined with predictive analytics can stimulate interactive models between the traits the most effective teachers should have to maximize the learning experience for students.

Chapter 3: Benefits of Big Data to Small Businesses

One reason small businesses watch from a distance as certain things happen, is that often the cost involved is prohibitive. As for big data, a good number of big size companies have embraced it, and that has created the notion that small businesses are yet to qualify for big data. Nothing could be further from the reality.

With data analytics, there is no need for complex systems that come with huge monetary demands if the enterprise in question is small. Much that is needed is based on proper organization, human attention to detail as well as how well people in the organization are equipped in the skills of data analytics. Any organization that can gather ample data that is relevant to its operations, and analyze it critically, has great chances of seizing business opportunities that they would otherwise have missed.

A small business can, therefore, utilize data analytics to improve the quality of its products or services; modify or fully alter its marketing strategies; and even improve on customer relations. All those areas have a great impact on the overall productivity of the business and hence its sustainability and profitability. Yet such improvement does not necessarily come with a massive price tag.

Examples Of Cost-Effective-Techniques Of Data Analytics:

Why is cost an issue anyway? The reason cost is always factored in before a crucial business decision is made is that cost has a direct impact on profitability. If, for instance, your small business increases sales revenue by 30% at a cost that has risen at 35%, the net profit will not have increased and the newly introduced techniques will have backfired. Still, you cannot stay put when your competitors are reaping increased profits from improved systems, whether those systems have anything to do with data analytics or not.

If you realize your competitors are using data analytics and before you know it they are eating into your market share, it would be suicidal not to consider the use of data analytics too. Since at the end of the day big data brings to the fore any business opportunities that exist; and those opportunities come with reduced costs and increased revenues; it is wise for small entrepreneurs to consider using it. After all, small businesses grow by making use of new business opportunities just the same way big businesses do.

This is what small businesses ordinarily rely on for growth:

- Personal intuition
- Commitment and ability to provide high quality service

Why Small Businesses Do Not Often Give Priority To Big Data

Well, all businesses, big or small, have, for a long time run their businesses traditionally. However, whenever developers come up with innovative ideas, they first of all target big businesses as those are likely to jump at new ideas simply because they can afford to implement them, cost notwithstanding. The scenario is not much different for big data. Many vendors who market enterprise-software-solution put a lot of emphasis on the advantages of economies of scale. Now, how can a small business even think of economies of scale, when it's actually struggling to build capacity?

With such kind of marketing, therefore, small enterprises consider themselves below par; like big data is not for them. This may have made business sense when the idea of big data had not become very popular, especially because a relatively big capital outlay was required upfront in preparation to roll out the application to the thousands of end users involved. And this is, definitely, where economies of scale come in.

However, in today's heightened level of business competition, small businesses need to borrow a leaf from the world leading companies, so that if big data is working for huge enterprises, small businesses may try it too. Luckily, innovators have come up with data solutions that are suitable for small businesses – solutions that equip the small businesses with the appropriate tools to be used by the type of end users small businesses engage. These solutions are thus designed to help the small business accomplish its work faster and more efficiently.

Unless small businesses begin to adopt the concept of big data, big competitors are likely to expand their market and run the small entrepreneurs out of town. After all, when big companies use big data, they increase their efficiency, and that gives them the drive and ability to embrace a bigger market share at the expense of small companies. As a consequence, it becomes increasingly difficult for the small business to break even after resigning themselves to the fate of continually declining profits.

Areas Where Big Data Can Help A Small Business Cost Effectively:

- Social media

Here the small business could analyze the relationship that exists between the

use of social media – like Twitter; Pinterest; Facebook; and others – and users' tendency to consume your product or service. With this kind of analysis, you can then proceed to capitalize on emerging sales opportunities. By setting up a business presence on these social platforms and targeting potential clients/customers using the site's consumer analytics and ad platforms you can grow your client base faster than using traditional marketing strategies. Setting special targeting ad strategies can place your ads in front of the clients and customers that are the most interested in your products and services. This is a cost effective marketing strategy when done correctly, however some specialized knowledge of each platform is needed to make this effective.

- Launch of an online service

You can take the opportunity to analyze people's tendency to visit your site. Is the visitor traffic high? If so, how high is it? While at it, you could study the users' habits; like which pages they click next after the landing page. This is the time you analyze the details that seem to capture the attention of these users. In the same breath, you can tell the details that turn off the users.

Aren't such details great at helping you identify the best pages to place your promotional content? Don't they also point at the pages and sites where cross selling opportunities abound? In short, there are cost effective ways of utilizing big data and small businesses can benefit just as much as big ones in this regard.

Factors To Consider When Preparing For A Big Data Solution:

1) A customized solution

When you are a small business, you don't go acquiring a big data solution that has been mass produced, or one that is suitable for a big business. You need one that your internal users will find applicable and helpful. If your big data solution is not tailor made for your business, then it needs, at least, to have capabilities that the users can take as options, without being encumbered by useless capabilities that have nothing to do with the needs of the business. It also needs to leverage the solutions your business has already put in place as well as your entire system.

Ultimately, what you want for your small business is a solution that has everything you need, integrated and packaged conveniently for use. Beware of vendors who may wish to have you overhaul your system, removing capabilities you already have in place; capabilities you have already adopted and even implemented. Remember to justify every cost you bring upon your business.

In case the recommendations for big businesses seem different from those of small ones, it is only because small businesses do not go for a massive

integrated solution for the entire organization. Rather, they work with smaller cost centers such as departments. For example, the marketing department could requisition for a marketing automation system without relying heavily on the views of the IT department. All that the marketing department is called upon to do is lay out its unique requirements, like its system requirements; spell out its cost justifications; and then do appropriate research to identify the solution that suits the department best. At the end of the day, you could have your small business with a varied range of solutions, each of them befitting its relevant department. Each department collects the type of data it deems relevant to its operations, analyzes it, and acts on it to increase the department's efficiency.

2) Ease of deployment

You need to acquire a big data solution that is simple to deploy. You also need to ensure that you are paying for a solution that the users find easy to use. In short, in terms of putting the solution in place, it need not take more than a couple of days of weeks, including testing. Any big data solution that is bound to take you months or even years to put in place in preparation for use is not suitable for your small business. For starters, time is money and a complex solution that will cause a department heavy downtime is not worth it. In any case, as you struggle to deploy your new solution, business opportunities might be passing you by.

As you ascertain simplicity, you also need to ensure that the big data solution you are contemplating acquiring can work with your other applications. A solution will be worth it only if it can work with your existing applications and systems seamlessly. Otherwise, you could be introducing a liability into your business. You need a solution that users will be able to use without pushing the organization to outsource highly paid specialists to help out.

By the same token, the big data solution you identify for your small business needs to be one that your staffers can use without the need to undergo expensive and time consuming training. In fact, you need to look for a solution with self service capabilities, so that even when you change users, new ones can still make use of it without a hitch; and without the need to always call on the IT department to help out.

3) The cost involved

Nothing in business makes sense until you know the cost involved. In the case of big data solutions, it is a good idea to buy one that is versatile so that you can increase the use of its capabilities as your business grows. When the solution is new, you may need to use only a few capabilities; and your solution should allow for that. There should also be room for you to pay for only the capabilities

you are prepared to use. However, as the business grows, you may need to use more of the big data solution's capabilities.

So, not only should the solution be priced reasonably, it should also come with a licensing strategy that allows the business to progressively increase the capabilities used as its need for data analytics goes up. Business owners always look forward to their small businesses growing fast, and so the big data capabilities you are able to bring on board should match the business rate of expansion as well as growth.

Clearly, it is possible and practical to switch from being intuition driven, even when yours is a small business, to being analytics driven. All you need to do is identify an IT solution that is suitable for the size and nature of your business. And from then on, your business will be in a position to enjoy the benefits that come with big data, mainly, the chance to identify viable business opportunities.

Chapter 4: Key Training for Big Data Handling

For the continued success of an organization, management needs to focus on the impact new systems are likely to have on overall operations, and ultimately the bottom line. Taking a country as an example, the reason that some countries choose to have small armies while others choose to have big ones is that some countries consider the existing threat to their security high while others consider it low. In short, countries do not enroll their youth into the military just to have them draw a salary. That is the same thinking that runs through the minds of management as they seek to establish the people in their organization that require training in big data.

Irrespective of whether an organization is big or small, the ability to employ big data is a worthwhile challenge. Even if various departments may work independently when deploying their applications, big data enables the departments to relate better as far as sharing of information is concerned, especially because each of them is able to feed the rest with credible information and relatively faster at that. Still, it is important that training in big data is limited to people who can actually use it productively in the organization. In short, there needs to be some eligibility criteria when it comes to training.

The highest priority needs to go to employees who handle massive data, and who may need to make use of Hadoop. As has been explained elsewhere in this book, Hadoop is a software library that is in open-source form, and which helps in distributed massive data processing. It is the role of the Chief Training Officer (CTO) to determine who in the organization needs training, if big data is to benefit the organization. The officer may consult with the various department heads because they understand what each employee under them does, and they know the individuals who handle data in heavy traffic. It is the same way a country's Army General liaises with government to determine the caliber of citizens that qualifies for training and recruitment into the army.

However, the CTO is ultimately responsible for enlisting those persons for big data training, being, as is expected, the topmost person as far as the IT infrastructure in the organization is concerned. The CTO needs to have in mind that whatever recommendations are taken, they will help to establish legacy systems to drive the new technology, which will help the employees to become more efficient in their respective roles.

Traditionally, organizations have tended to put a lot of emphasis on acquisition of data. Other times is just thrown around especially now that the web has a whole world of data from every imaginable quarter. However, it has become increasingly clear that massive data is only helpful if there is a way of sorting it

out so that specific data is classified and utilized where it is most relevant. Unfortunately, this sorting has been left to be done mainly as per individual's intuition and judgment, without any structured system. This means there is deficiency in people's ability to utilize data to its optimum.

Current Level Of Competence In Data Management

A report provided by CIO, a popular site that deals with issues of technology, indicates that the Computing Technology Industry Association (CompTIA) has done a survey that shows how poorly equipped staff in organizations leveraging data are, in the aspects of data management. A good number of them are also deficient in the skills of data analysis. This survey, which was based on information derived from 500 executives from the business as well as IT sectors, indicates that half of the organizations in the lead as far as leveraging data goes have staff that is not sufficiently skilled in data management and analysis. When it comes to organizations that only leverage data moderately, 71% of them see their staff as poorly equipped or just moderately skilled in the area of data management and analysis.

What these results point to is the need to make a conscious effort to train staff when it comes to big data, so that the availability of data makes business sense to the organization.

Where, Exactly, Is Big Data Training Required?

1. The IT Department

This is a no-brainer. If there are people who are bugged, raveled and nagged, by the rest of the organization when it comes to matters of IT, it is members of the IT department. Even when liaison of data between departments fails to work properly, these are the people who are called upon to help streamline the systems so that information flow is better coordinated. It is imperative, therefore, that they are well versed in matters of big data, to be able to give appropriate support to members of other departments. In fact, when the IT department is well trained, they save the organization the expenses it would have incurred seeking outside help in supporting the rolled out systems. The role of IT Security has become a new player given all of the new rules and regulations on how data should be treated and stored. This department also needs to be aware of the latest hacking and breaching applications being used and get in front of those before they happen. There are also departments being built around Data Privacy and Compliance within the IT Department that will need to work heavily with the Legal and Compliance departments to make sure that the data isn't compromised and cause significant losses for the company.

2. The Department of Product Development.

Not only is this department tasked with creating new products, it is also tasked with re-engineering existing products. Everyone involved is called upon to have a fresh way of thinking, what can be aptly termed, re-thinking innovation. Everyone in the department is deeply involved in every step of the Research and Development process.

For re-engineering, there is a massive responsibility on the part of the staff, and it involves a full understanding of the product as it currently is, and the product as it is envisaged to be after the innovative process of re-engineering. For this understanding to be comprehensive and helpful, massive data is involved, and it needs to be well analyzed, critically evaluated, and properly utilized. That is why the staff in the department of product development needs to be well trained. They should be in a position to capitalize on the advantages provided by big data, particularly when it comes to data integration across existing touch points in the process of product development. Some of those important touch points include product design; software applications; diagnostics; manufacturing; quality determination; among others.

3. The Department of Finance

Ever heard something like *show me the money?* Nobody wants to say that more than the staff in the department of finance. And if you cannot show them the money, they will want to find out independently if you are really worth funding as a department or you are actually a liability to the organization. It is, therefore, clear why the staff in this department needs to be well trained in handling big data. They need to be able to say in monetary terms if there is value in deploying certain resources or not.

In fact, the department of finance is almost as central to the organization as that of IT, and its employees need to be well trained to make use of big data platforms as far as financial modeling is concerned. Only then can a business be sustained because without conservative spending and optimum funding of projects, money, sometimes referred to as the actual blood of the business, is likely to dry out; meaning the business goes under. The essence of training personnel in the finance department is to prepare them to make use of big data in their core roles of business planning; auditing; accounting; and overall controlling of the company's finances. Conversely, when the department does these roles well, it ends up creating massive finances for the organization.

Some of the ways members of the finance staff contribute to the success of the organization when they are well trained in big data is by managing to generate reliable cash flow statements; compliance in finance standards; cost modeling;

prize realization; and so on. With the influx of data, the finance function has become more complex, especially when the staff wants to get insights into the future of the organization. As such, training of finance staff in big data is not a luxury but a necessity.

4. The Human Resource Department (HR)

Though it may not appear obvious, the Department of HR can utilize skills in application of big data to improve on the quality of personnel hired. Competence in big data analysis also comes in handy in assessing the capabilities of existing employees and in determining the capabilities required in future engagements.

As such, members of the HR department need to be well trained too so that they can analyze data according to relevance, with a view to engaging in a more strategic manner when it comes to its functions. With that demand, it becomes imperative that this staff is able to use available tools for data analysis, because such are the competencies that will ultimately help in solving issues of staff retention; staff-customer relations that affect sales; talent gaps within the organization; quality of candidates to shortlist; and so on. In short, HR is no longer about the number of employees engaged, but also about predictive analysis of matters that have to do with the human resource.

5. The Department of Supply and Logistics

No business can thrive when customers perennially complain about late deliveries; breakages on transit; and other factors that are irritating to customers. It is, therefore, important that the staff in the department of supply and logistics be well trained to handle big data, so that they can utilize it in implementing the department's strategies and in achieving its goals. Once the members of staff in this department are trained in big data, they will be in a position to improve performance and save on cost, primarily by being faster and more agile in their activities. On the overall, their training would greatly improve on the department's efficiency in service delivery.

What is the import of that? Well, it is massive improvement in customer experience as well as emergence of a fresh business model. Needless to say, this achievement comes with significant conservation of resources, in terms of reduced downtime; reduced overtime; reduced waste; and so on. Ultimately, the smooth supply chain and the great logistics put in place end up promoting your brand name markedly. And that is precisely how a business pulls in more customers, increasing the business market share, without planned and costly marketing.

6. The Department of Operations

In many organizations, this department also encompasses customer support. As such, it is important to train the employees in this department, so that whatever an employee does in the course of duty, there is customer satisfaction in consideration. Besides, it is important that the employees understand the impact of providing great customer support even after they have made a sale. Big data training enlightens the employees very well in this regard. Being able to analyze big data shows the staff clearly that great customer service and support in key to the success of the organization because it improves customer retention; expansion of customer base; and such other success.

7. Department of Marketing

Do you realize how critical numbers are in marketing? Whether it is the number of customers you have brought on board in the last one year; the percentage of market share you have captured within a given time; the sales revenue your product has earned from a certain demographic; this and more keep the marketing department busy. There is massive data in the market, some positive and some potentially damaging to your brand, and it is imperative that members of staff in the marketing department are able to amass the data that appears relevant; analyze it and sieve it to remain with what can be organized and acted upon to the benefit of your brand. Without the skills to handle big data, it is difficult to make head or tail of the mass of data making the traffic through social media, in this era when digital marketing is the order of the day.

However, with proper training in big data, the department of marketing can measure with relative accuracy the response your advertisements have on the market; the volume and impact of click-through-rate; impressions; Return on Investment (ROI); and such other factors. While plenty of such data from social media may look irrelevant to the general web visitors, it is as valuable as gold when it comes to the trained eye of a marketer. Besides, the trained marketer is able to make good use of the large volume of data generated through all stages of customer interaction; social media; as well as during the sales process. For the marketing team, platforms such as Hadoop are very helpful.

Big data training helps in retrospection too where the marketing staff can gauge how well their brand has been doing compared to competing brands; what it is that other similar brands offer; what varying features competing brands have; and such other information that the business can use to improve on its brand. There is even a function in big data's domain that the marketing team can use to crawl competitors' websites and do a bit of helpful text mining.

8. Department of Data Integrity, Integration and Data Warehouse

With the increased amount of data there is to monitor, it is crucial that you have a team of specialists that can take in data from your various company systems. This team will need to be current and trained on the various data in the systems, as well as potential risks associated with them. There also needs to be folks who know how to warehouse all this data and make structured sense of it. From customer protection and privacy laws, the teams that work with and interpret the data will need to know the appropriate rules for the handling of it.

9. Department of Legal and Compliance

With the new legal and compliance rules surrounding data in large organizations, it is necessary for the Legal Department to stay aware of new privacy and retention policies as it relates to certain pieces of data. The Legal team and Compliance departments should work together to interpret data privacy laws to make sure that the data is protected, stored and treated appropriately. Businesses that don't monitor and report their data can suffer significant legal issues and should have policies in place to protect themselves against potential lawsuits and breaches.

Chapter 5: Process of Data Analysis

In this chapter, I have highlighted the steps involved in the process of data analysis. Let us look at the steps one at a time.

What Is Data Analysis?

Before we get on with the steps involved in the process of data analysis, let us look at what data analysis is first. Data analysis is the process by which raw data is collected from various sources and gets converted into meaningful information, which can be used by various stakeholders for making better decisions.

To put it in the words of John Tukey, a famous Statistician, data analysis is defined as follows:

"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

Steps Involved In Data Analysis

Even though the data requirements may not be the same for every company, most of the below steps are common for all companies.

Step 1: Decide on the objectives

The first step in the data analysis process is the setting of objectives. It is important that you set clear, measurable and concise objectives. These objectives can be in the form of questions. For instance, your company's products are finding it difficult to get off the shelves because of a competitor's products. The questions that you might ask are, "Is my product overpriced?" "What is unique about the competitor's product?" "What is the target audience for the competitor's product?" "Is my process or technology redundant?"

Why is asking these questions upfront important? This is because your data collection depends on the kind of questions you ask. For instance, if your question is, "What is unique about the competitor's product?" you will have to collect feedback from the consumers about what they like in the product as well as do an analysis on the specifications of the product. On the other hand, if your question is, "Is my process or technology redundant?" you will have to do an audit of the existing processes and technologies used at your establishment as well as do a survey about the technology used by the others in the same industry. As you can see from this, the nature of data collected differs vastly based on the kind of questions you ask. Since data analysis is a tedious process, it is necessary that you do not waste the time of your data science team in collecting useless data. Ask your questions right!

Step 2: Set measurement priorities

Now that you have decided your objectives, you need to establish measurement priorities next. This is done through the following two stages:

Decide what to measure

This is when you have to decide what kind of data you need to answer your question. For example, if your question pertains to reducing the number of jobs, without compromising the quality of your product or service, then the data that you need in hand right now are:

- The number of staff employed by the company.
- The cost of employing the present number of staff.
- The percentage of time and efforts spent by the current staff on the

existing processes.

Once you have the above data, you will have to ask other questions ancillary to the primary question such as, “Are my staff not being utilized to their fullest potential?” “Is there any process that can be altered to improve the productivity of the staff?” “Will the company be in a position to meet increased demands in the future despite the downsizing of manpower?”

These ancillary questions are as important as the primary objective/question. The data collected in connection with these ancillary questions will help you in taking better decisions.

Decide how to measure

It is highly important that you decide the parameters to measure your data before you begin collecting it. This is because how you measure your data plays an important role in analyzing the collected data in the later stages. Some of the questions you need to ask yourself at this stage are:

- What is the time frame within which I should complete the analysis?
- What is the unit of measure? For instance, if your product has international markets and you are required to determine the pricing of your product, you need to arrive at the base price using a certain currency and extrapolate it accordingly. In this case, choosing that base currency is the solution.
- What are the factors that you need to include? This could again depend on the question you have asked in stage 1. In the case of staff downsizing question, you need to decide on what factors you need to take into consideration with respect to the cost of employment. You need to decide whether you will be taking the gross salary package into consideration or the net annual salary drawn by the employee.

Step 3: Collection of data

The next step in the data analysis process is the collection of data. Now that you have already set your priorities and measurement parameters, it will be easier for you to collect data in a phased manner. Here are a few pointers that you need to bear in mind before you collect data:

We already saw the different sources of data in the previous chapter. Before you collect data, take stock of the data available. For example, in the case of the downsizing of the staff case, to know the number of

employees available, you can just look at the payroll and get the numbers. This could save you the time of collecting this particular data again. Similarly, collate all available information.

If you intend to collect information from external sources in the form of a questionnaire, then spend a good amount of time in deciding the questions that you want to ask the others. Only when you are satisfied that the questionnaire looks satisfactory and serves your primary objective, should you circulate it. If you keep circulating different questionnaires, then you will have heterogeneous data in hand, which will not be possible to compare.

Ensure that you have proper logs as and when you enter the data collected. This could help you analyze the trends in the market. For instance, let us assume you are conducting a survey regarding your product over a period of two months. You will note that the shopping habits of people change drastically during holiday seasons than any other period of the year. When you don't include the date and time in your data log, you will end up with superfluous figures and this will affect your decisions in a grave fashion.

Check how much budget is allocated for the purpose of data collection. Based on the available budget, you will be able to identify the methods of data collection that are cost effective. For instance, if you have a tight budget and you still have to do a survey to gauge the preferences of your customers, you can opt for free online survey tools as opposed to printed questionnaires that are included in the packages. Similarly, you can make the best use of social networking sites to conduct mini surveys and collect the data required. On the other hand, if you have enough budget, you can go for printed and attractive questionnaires that can be circulated along with the package or can be distributed at retail outlets. You can set up drop boxes at nearby cafes and malls for the customers to drop these filled in questionnaires. You can also organize contests to collect data as well as market your product in one go.

Step 4: Data cleaning

Now, it is not necessary that the data you have collected will be readily usable. This is why data cleaning is very crucial in this process to ensure that meaningless data does not find its way into the analysis stage. For example, when you correct the spelling mistakes in the collected questionnaires and feed it into your system, it is nothing but data cleaning. When you have junk data in the system, it will affect the quality of your decision. For instance, let us assume 50 out of 100 people responded to your questionnaires. However, you get ten incomplete questionnaire forms. You cannot count these ten forms for the

purpose of analysis. In reality, you have gotten only 40% response to your questions and not 50%. These numbers make a big difference for the management to make decisions.

Similarly, if you are doing a region wise survey, you need to be extra careful at this stage. This is because most people have a tendency to not fill in their correct addresses in the questionnaires. Hence, unless you have a fair idea about the population of each region, you will never be able to catch these slip-ups. Why is it important to catch these mistakes? Let's assume that your survey results show that a majority (say 70%) of your customer base is from X region. In reality, the population of the region is not even close to 30% of your customer base. Now, let us assume that you solely decide to make a decision based on your survey results. You decide to launch an exclusive marketing drive in this region. Of course, the marketing drive will not improve your sales because even if all the citizens in the region buy your product, they have contributed to only 30% of your customer base and not 70% as you imagined. Hence, these little numbers play an important role when it comes to big and expensive decisions.

As you can see, improving the quality of data is highly important for taking better decisions. Since this involves a lot of time, you should automate this process. For instance, to detect fake addresses, you can get the computer to detect those entries that have incorrect or incomplete Zip code. This could be easily done if you are using an Excel to store your data. Alternatively, if you have customized software for feeding and storing data, you can get in touch with the software designer to put certain algorithms in place to take care of such data.

Step 5: Analysis of data

Now that you have collected the requisite data, it is time to process it. You may resort to different techniques to analyze your data. Some of the techniques are as follows:

Exploratory data analysis:

This is a method by which data sets are analyzed with a view to summarize their distinct characteristics. This method was developed by John W. Tukey. According to him, there was too much importance that was being shown towards statistical hypothesis testing, which is nothing but confirmatory data analysis. He felt the need to use data for the purpose of testing hypotheses. The key objectives of exploratory data analysis are as follows:

(i) Suggestion of hypotheses in connection with the causes of the phenomena under question.

(ii) Assessing the assumptions on which the statistical inference will be based.

(iii) Supporting the choosing of appropriate statistical techniques and tools.

(iv) Providing a basis for further data collection through modes such as surveys or experiments.

Several techniques prescribed by the exploratory data analysis have been put to extensive use in the fields of data mining and data analysis. These techniques also form part of certain curriculum to induce statistical thinking in students. As you perform explorative data analysis, you will be required to clean up more data. In some cases, you will be required to collect more data to complete the analysis. This is to ensure that the analysis is backed by meaningful and complete data.

Descriptive statistics:

This is another method of data analysis. By this method, data is analyzed to identify and describe the main features or characteristics of the data collected. This is different from inferential statistics. Under inferential statistics, the data collected is analyzed to learn more about the sample. These findings are then extrapolated to the general population based on the sample. On the other hand, descriptive statistics aims to only summarize and describe the data collected. These observations about the data collected can be either quantitative or visual. These summaries could just be the beginning of your data analysis process. These could form the basis on which further analysis is done to process the data. To understand better, let us look at an example. The shooting percentage in the game of basketball is nothing but a descriptive statistic. This shooting percentage indicates the performance of the team. It is calculated by dividing the number of shots made, by the number of shots taken. For instance, if a basketball player's shooting percentage is 50%, it means that he makes one shot in every two. Other tools used under descriptive statistics include mean, median, mode, range, variance, standard deviation etc.

Data visualization:

As the name suggests, data visualization is nothing but the representation of data in a visual form. This can be done with the help of tools such as plots, informational graphics, statistical graphics, charts and tables. The objective of data visualization is to communicate the data in an effective fashion. When you are able to represent data effectively in a visual form, it helps in the analysis of data and also to reason about data and evidence. Even complex data can be

understood and analyzed by people, when put in visual form. These visual representations also facilitate easy comparison. For instance, if you are vested with the job of reviewing the performance of your product and your competitor's product, you will be able to do so easily if all the related data are represented in visual form. All your data team needs to do is data pertaining to the parameters such as price; number of units sold, specifications etc, and put it in pictorial form. This way, you will be able to assess the raw data easily. You will also be able to establish correlation between the different parameters and make decisions accordingly. For instance, if you notice that your price is higher than your competitor's and your sales are lower than your competitor's, then you know where the problem lies. The decreased sales can be attributed to the increase in price. This can be set aside easily by reworking your prices.

Apart from these three major methods, you can also take the help of software available in the market. Some of the prominent software available in the market for the purpose of data analysis are Minitab, Stata and Visio. Let us not forget the multipurpose Excel.

Step 6: Interpreting the results

Once you have analyzed your data, it is now time to interpret your results. Once you have analyzed the data, here are a few questions that you need to ask:

Does the analyzed data answer your key question? If yes, how so?

If there were any objections to begin with, did your data help you defend them? If yes, how so?

Do you think there are any limitations to your results? Are there any angles that you haven't considered while setting priorities?

Do you have trained people to properly interpret data?

If the analyzed data satisfies all the above questions, then your analyzed data is final. This information can now be used for the purpose of decision-making.

Importance Of Interpreting The Data Effectively & Accurately

The importance of accurately interpreting the data cannot be emphasized enough. Your company, website, etc. must have experienced professionals who know how to take the organic data and interpret the results properly. For example, let's say your company finds it necessary to analyze data from the two of the most popular social media platforms, Facebook and Twitter.

Your company cannot depend on an untrained professional to effectively respond to your "likes" or "tweets" on a minute-by- minute basis. Most

companies today employ a Social Media Manager to manage their social platforms. These individuals are trained to know the ins and outs of each social platform and effectively be able to respond to your customers in a way that represents your brand.

At the core of every successful business is accurate interpretation of vital data. It's necessary to hire trained professionals who have the training to take the unstructured and otherwise random data and put it into an understandable structure. This will change the dynamics of how your company operates and what decisions need to be made based on the data.

People who can take the all of your customers "likes" on your corporate Facebook page and trace the consumer's behavior regarding the use of your product. Follow the decision-making process of these consumers. The consumers like your product, then what? Do they read the product description? Do they reap the benefits of using your product? Is your product reasonably priced in comparison to your competitor's prices? What makes your company's product better than the rest of the competition?

Trained data analysts will be able to trace these questions and analyze the pattern your consumers will take. They will follow the trace the consumers take. They can analyze the data from the original "like" from your consumer all the way to the purchase on your website.

The right people who have the training to follow and analyze this process can help your company generate increased product sales by taking this information and disseminate throughout the company to the appropriate team members. Actually having properly interpreted meaningful data may be the difference between your company expanding its influence or shutting down because of misinterpreted data.

An example of how you can analyze tweets is interpreting historical tweets, knowing the substantial "tweet" from the casual "tweet."

The data interpreters are able to analyze historical data from previous company "tweets" and their influence on consumer buying habits. These experts can translate which "tweets" are substantial and which "tweets" are just social. From the initial root message texted on twitter the analyst is able to trace the impact on the consumer's initial mindset as to whether they will follow the company's core goal to buy their product or not. Which text is more convincing than the other one? Why is it more successful? Do images with the "tweets" tend to convince your consumer base to buy your product? Which "tweets" work best with what regions in the world? Which "tweets" work best with what age group?

These are important questions that can be answered by the data and why it is important to have analysts review and show you what marketing strategies are

working the best on each platform. Analysts can interpret large amounts of data with the use of visual graphs of numerical data statistics. These can be given to the appropriate departments to make decisions to improve the overall sales experience for your customers.

Chapter 6: Predictive Analytics

We have seen now how data and data analysis is crucial for the effective functioning of a business. Let us look at another wing of data mining, which plays a vital role in the growth of a business. In this chapter, I will be taking you through the various aspects of Predictive analytics and help you appreciate the role played by it in facilitating the effective functioning of a business.

What Is Predictive Analytics?

In simple terms, predictive analytics is nothing but the art of obtaining information from the collected data and utilizing it for predicting behavior patterns and trends. With the help of predictive analytics, you can predict unknown factors not just in the future but also the present and past. For example, predictive analytics can be used to even identify the suspects of a crime that has already been committed. It can also be used to detect a fraud as it is being committed.

What Are The Types Of Predictive Analytics?

Predictive analytics can be referred to as predictive modeling. In simpler terms, it is pairing data with predictive models and arriving at a conclusion. Let us look at the three models of predictive analytics.

Predictive models

Predictive models are nothing but the models of the relationship between the specific performance of certain element in a sample and few known attributes of the sample. This model aims at assessing the likelihood that a similar element from a different sample might exhibit the specific performance. It is widely used in marketing. In marketing, predictive models are used to identify subtle patterns, which are then used to identify the customers' preference. These models are capable of performing calculations as and when a certain transaction is happening i.e., live transactions. For instance, it is capable of evaluating the opportunity or risk associated with a certain transaction for a given customer, thereby helping the customer decide if he wants to enter into the transaction or not. Given the advancements in the speed of computing, individual agent modeling systems have been designed to simulate human reactions or behavior for certain scenarios.

Now, let us look at some more aspects of these models in detail. The terminology 'training sample' refers to the sample units/elements that are available, whose attributes and performances are known. The units/elements that are present in other samples, whose attributes are known but whose performances are unknown, are referred to as 'out of training sample'. There is no chronological relation between the training sample and the out of training sample. For example, the blood splatters in a simulated environment are training samples. On the other hand, the blood splatter from an actual crime scene is the out of training sample. Predictive models can help in identifying the probable suspects and modus operandi of the murder based on these samples. As mentioned earlier, these samples are not required to be from the same time frame. Either of the samples could be from a different time frame.

Predictive Models in relation to Crime Scenes

3D technology has brought big data to crime scene investigations to help police departments, and criminal investigators reconstruct crime scenes without violating the integrity of the evidence. There are two types of laser scanners used by crime scene experts:

- Time-of–flight laser scanner

The scanner shoots out a beam of light. It bounces off the targeted object, and different data points are measured as the light returns to the sensor. It's capable of measuring 50,000 points per second.

- Phase shift 3D laser scanners

These scanners are much more expensive but also much more effective. They measure 976,000 data points per second. These scanners use infrared laser technology.

These data laser scanners make crime scene reconstruction much easier. Needless to say, the process takes a lot less time than traditional crime scene reconstruction took. The advantage of 3D technology is that the investigators can re-visit the crime scene anywhere. Investigators can now do this while they are at home, in their offices or out in the field. This makes their job more mobile, and they can visit the crime scenes virtually anywhere. They no longer have to depend on notes or their memories in calling up the details of the crime scene. Also, they visit the crime scene once, and that's it. They have all the data images recorded on their scanners.

Investigators are able to re-visit the crime scenes viewing the images on computers or iPads. Distance between objects will be reviewed (like weapons.) The beauty of this technology is crime experts don't have to second guess their information gathered from crime scenes. The original crime scene is constructed right there on the scanner images. It's as if the crime was committed right there on the scanned images. The images tell the story about who the perpetrators were and how they carried out the crime.

Investigators can look at the crime scenes long after the crime scenes are released. Nothing in the crime scene will be disturbed or compromised. Any evidence that is compromised is inadmissible in court and cannot be used. All evidence must be left in its original state and not tampered with. This is no problem when the evidence is recorded in the data scanners.

Law Enforcement engineers are able to reconstruct the whole crime scene in the courtroom. Forensic evidence is untouched and left intact guaranteeing a higher rate of convictions.

Forensic Mapping in Crime Scene Reconstruction

The purpose of 3D forensic mapping is to reconstruct every detail of the crime scene holistically. This is a very pure way of reconstructing crime scene evidence. None of the evidence is touched or accidentally thrown away. Investigators don't have to walk in or around the evidence avoiding the possibility of anything being accidentally dropped or kicked.

3D data mapping allows for greater understanding and insight into the motive and the method of the crime committed. This helps law officials to present a convincing understanding in court so evidence will convince the jury beyond a reasonable doubt that the crime was or wasn't committed.

3D forensic mapping is also invaluable in training new investigators in the process of criminal reconstruction.

Descriptive models:

Descriptive models are used to ascertain relationships to the data collected. This is similar to how a company would classify and group its customers or its products into certain categories. When it comes to the predictive models, it focuses on predicting only a single customer behavior, as in the case of the computation of the credit risk. On the other hand, descriptive models focus on identifying different relationships between products or customers. Descriptive models do not seek to rank customers based on their attributes or based on the actions taken by them, as in the case of predictive models. Descriptive models tend to categorize customers based on their product preferences. These models can be used to build more models, which can be used to make more predictions.

Decision models:

Decision model is nothing but a system, which contains at least one action axiom. An action axiom is nothing but the action that follows the satisfaction of a certain condition. A model action axiom is as follows:

If <a certain fact> is true, then do <this certain action>.

In simpler words, the action axiom is used to test a certain condition. The fulfillment of the certain condition necessitates the completion of a certain action.

The decision model also describes the relationship between all the elements that form part of a decision, such as the decision, the known data and also the forecast results associated with the decision. This model aims to predict the results of those decisions, which have many variables involved in it. Decision models are also used to achieve optimization and maximize certain outcomes, while minimizing certain other outcomes. These models can be used to come up with a set of rules for a business, which will be capable of producing the expected and desired action for every customer who avails the service of the business.

Chapter 7: Predictive Analysis Techniques

In this chapter, let us look at the different techniques used for the purpose of conducting predictive analytics. The two major categories into which these techniques can be grouped into are machine learning techniques and regression techniques. Let us look at these techniques in detail now.

Regression Techniques

These techniques form the foundation of predictive analytics. These techniques aim to establish a mathematical equation, which will in turn serve as a model for representing the interactions among the different variables in question. Based on the circumstance, different models can be applied for performing predictive analysis. Let us look at some of them in detail now:

Linear regression model:

This model assesses the relationship between the dependent variable in a given situation and the set of independent variables associated with it. This is usually expressed in the form of an equation. The dependent variable is expressed as a linear function of the different parameters. These parameters can be adjusted in such a way that it leads to the optimization of measure of fit. Model fitting is required to minimize the size of residual. This importance shown towards model fitting is required to ensure that it is distributed randomly in connection with the model predictions.

The objective of this model is the selection of the parameters of the model, with a view to minimize the sum of the squared residuals. This is known as the ordinary least squares estimation. Once you have estimated the model, you are now required to check the statistical significance of the different coefficients used in the model. This is where the t-statistic comes into play, whereby you test if the coefficient is different from zero. The ability of the model to predict the dependent variable depending on the value of the other independent variables involved can be tested by using the R^2 statistic.

Discrete choice models:

Linear regression models can be used in those cases, where the dependent variable has an unbounded range and is continuous. But there are certain cases where the dependent variable is not continuous. In these cases, the dependent variable is discrete. Since the assumptions related to the linear regression model do not hold good completely in the case of discrete variables, you will have to go for another model to conduct predictive analytics.

Logistic regression:

This model is used in those cases where the dependent variable is categorical. A categorical variable is one, which has a fixed number of values. For instance, if a variable can take two values at a time, it is called as a binary variable. Categorical variables that have more than two values are referred to as

polytomous variables. For example, blood type of a person is an example of polytomous variable.

Logistic regression is used to determine and measure the relationship between the categorical variable in the equation and the other independent variables associated with the model. This is done so by utilizing the logistic function to estimate the probabilities. It is similar to the linear regression model. However, it has different assumptions associated with it. There are two major differences between the two models. They are as follows:

The Linear regression model uses a Gaussian distribution as the conditional distribution whereas the logistic regression uses a Bernoulli distribution as the conditional distribution.

The predicted values arrived at in the logistic regression model are probabilities and are restricted to 0 and 1. This is because the logistic regression model is capable of predicting the probability of certain outcomes.

Probit regression

Probit models are used in place of logistic regression for coming up with models for categorical variables. This is used in the cases of binary variables i.e. categorical variables, which can take only two values. This model is popular in economics. This method is used in economics to predict those models, which use variables that are not only continuous but are also binary in nature. Two important reasons why the probit regression method is often opted over logistic regression method are as follows:

1. This is because the underlying distribution in probit regression is normal.
2. If the actual event is a binary proportion and not a binary outcome, then the probit method is more accurate.

Machine Learning Techniques

Machine learning is nothing but a field in artificial intelligence, which was used initially for the purpose of developing techniques that will facilitate computers to learn. Since, machine learning consists of an array of statistical methods for classification and regression, it is now being employed in different fields such as credit card fraud detection, medical diagnostics, speech recognition, face recognition and for analyzing the stock market. In certain applications, these techniques can be used to predict the dependent variable, without taking into consideration the independent variables or the underlying relationships between these variables. In other words, machine-learning techniques can be readily employed in those cases where the underlying relationships between the dependent and independent variables is complex or unknown. Some of the techniques used under this method are as follows:

Neural networks

Neural networks are nothing but sophisticated and non-linear modeling techniques, which are capable of modeling complex functions. These techniques are widely used in the fields of cognitive psychology, neuroscience, finance, physics, engineering and medicine.

Neural networks are generally employed when you are not aware of the exact relationship that exists between the inputs and output. However, they learn the underlying relationship by means of training. There are three kinds of trainings that form part of neural networks namely supervised training, unsupervised training and reinforcement learning. Of the three, supervised training is the most commonly used.

A few examples of training techniques used in neural networks are as follows:

- Quick propagation
- Conjugate gradient descent
- Backpropagation
- Projection operator
- Delta bar delta

Multilayer perceptron

The multilayer perceptron is made of an input layer and output layer. Over and above these, there are one or more hidden layers made up of sigmoid nodes or nonlinearly activating nodes. The weight vector plays an important role in adjusting the weights of the network. The backpropagation technique is

employed to minimize the squared error, which usually arises between the network output values and the expected values for the output.

Radial basis functions

A radial basis function is a function that has an inbuilt distance criterion in connection to a center. These functions are typically used for the interpolation and smoothing of data. These functions have also been used as part of the neural networks instead of the sigmoid functions. In such cases, the network has three layers namely, the input layer, the hidden layer with the radial basis functions and the output layer.

Support vector machines

Support vector machines are employed for the purpose of detecting and exploiting complex patterns in data. This is achieved by clustering, classifying and ranking of data. These are nothing but learning machines, which can be used for the purpose of performing regression estimations and binary classifications. There are several types of support vector machines such as polynomial, linear, sigmoid, to name a few.

Naïve bayes

This is based on the Bayes conditional probability rule. This is an easy technique using which classifiers are constructed. In other words, they are used for the purpose of classifying various tasks. This technique is based on the assumption that the predictors are statistically independent. It is this independence that makes it a great tool for the purpose of classification. Moreover, this technique is also easier to interpret. In those cases when the number of predictors is very high, it is wise to use this method.

K-nearest neighbors

The nearest neighbor algorithm forms part of the pattern recognition statistical methods. Under this method, there are no underlying assumptions associated with the distribution from which the sample is drawn. It consists of a training set and has both positive and negative values. When a new sample is drawn, it is classified based on its distance from the nearest neighboring training set. The sign associated with that point also plays an important role in classifying the new sample. While using the k nearest neighbor classifier, all the k- nearest points are considered. The sign of majority of these points are then used for the purpose of classifying the new sample.

The performance of this algorithm is determined by the following factors:

1. The distance measure that is used for the purpose of locating the nearest neighbors.
2. The decision rule that is used for the purpose of deriving a classification from the k- nearest neighbors.
3. The number of neighbors, which are being considered for the purpose of classifying the new sample.

Geospatial predictive modeling

The underlying principle of this technique is the assumption that the occurrences of events, which are being modeled, are limited in terms of distribution. In other words, the occurrences of events are neither random nor uniform in distribution. Instead, there are other spatial environment factors involved, such as socio cultural, infrastructure, topographic etc, which are capable of constraining and influencing the locations of the occurrences of these events. This is a process for assessing events using a geographic filter, with a view to come up with statements of likelihood for the occurrence of a certain event.

There are two kinds of geospatial predictive models. They are as follows:

1. Deductive method
2. Inductive method

Let us look at these methods in detail now.

1. Deductive method:

   This method is based on subject matter expert or qualitative data. This data is then used for the purpose of describing the relationship that exists between the occurrence of an event and the factors associated with the environment. In other words, this method relies majorly on subjective information. The limitation of this model lies in the fact that the number of factors that are being keyed in is completely dependent on the modeler.

2. Inductive method:

   This method is based on the spatial relationship (which is empirically calculated) that exists between the occurrences of events and the factors that are associated with the environment. Every occurrence of an event is first plotted in the geographic space. Following the plotting, a quantitative

relationship is then defined between the occurrence and the factors associated with the environment. An important reason why this method is highly effective is that you can develop software to empirically discover. This is crucial when you have hundreds of factors involved and there are both known and unknown correlations existing between events and factors. The quantitative relationship values, as defined before, are then processed, with the help of a statistical function. These values are processed to figure out spatial patterns, which are capable of defining areas that are not only highly probable for the occurrence of an event but also those areas which are less probable locations for the occurrence of an event.

Hitachi's Predictive Analytic System

How would you like to take a look into the future? This is what Hitachi's Predictive Analytic System (PCA) will achieve if it's successful in its unveiling. Their technology will revolutionize the way police departments handle crimes. The present method is labeled the "break/fix model" when police arrive on the scene after a crime has been committed. A death has taken place, rape, domestic violence or a burglary. Then the police have to administer damage control. Sometimes the police can get to the scene before the crime is committed or while the crime is being committed. Then they can stop it.

What if they could arrive on the scene before even the crime is committed? How many lives would be saved? People could be spared bodily pain from being abused or beaten up. Costly damages from property loss could be reduced or stopped. There are so many benefits that would be derived if the PCA is successful in its intended purpose.

The Pac would be able to predict crimes before they happened. This would enable police departments to arrive on the scene before the crime is committed. Yes, the police would arrive on the scene before the crime is committed. Instead of taking the quirks of traditional crime predicting methods where investigators use their own experiences (the product of their personal variables) to project the committing of crimes. PCA will eliminate countless variables to accurately analyze a multitude of factors that will affect crime. Here are the components PAC analyzes and will predict in real time the committing of a crime before it happens:

- Weather Patterns

- Public Transit Movements
- Social Media Actions
- Gun Shot Sensors
- Many other Factors

The above data sets are fed into the system and over a two-week period the Pac will determine if there is a correlation between the said datasets. Hitachi will allow different Law Enforcement agencies to test their systems in the real world working environments. These "experiments" will be held in different agency locations in unknown cities. The agencies will participate in what is known as "Double-blind trial." The predictive system will be running in the background but criminal experts will not be able to see the predictive results as they happen.

Summation of the testing period will be when Hitachi will compare and analyze the results of the predictions in relationship to the actual daily police activities over the same period.

Benefits of PCA

The benefits of the PCA being successful would be countless:

- Obviously, countless lives would be saved
- Law Enforcement Agencies could still work efficiently in the light of continual manpower cuts.
- Officers can be placed strategically in locations to stop potential crimes
- Officers would exponentially improve their crime-fighting time exponentially
- Law Enforcement Training costs would dramatically drop
- Law enforcement experts would in less life-threatening circumstances
- Homicides, rapes, domestic violence cases would drop
- Stolen property losses would be noticeably reduced
- Crime prevention success would dramatically increase

Predictive Analytics influencing many Insurance industries

It has been researched that across the many parts of the insurance industry they have switched to applying predictive analytics to their daily business practices. 49% of the Personal Auto Insurance industry now uses predictive analytics to:

- Profit increase
- Risk reduction
- Revenue growth
- Improve overall company operational efficiency

Big data and specifically predictive analytics are changing the way insurance companies conduct business and relate to their customers. The trend now is to give their customers the product they want at a price they want. Also, the insurance companies have become more sensitive to their customer's core issues. Competition has increased in this key industry and companies can no longer say, "You don't like our prices? Too bad." Insurance companies can no longer ignore the wishes of their clientele.

If the insured doesn't like the prices, their company is offering them they will go to the competition who will offer them a better insurance rate. How does predictive analytics help in these causes? Well, insurance companies are now offering telemetry-based packages to consumers. These packages are based on predictive analytics. The insurance company uses predictive modeling to form a risk management profile on the customer. They take all the risk factors connected with this particular person:

- Likelihood of the person getting in an accident
- Will their car be stolen?

By comparing this behavioral information with thousands of other profiles the insurance company can compile an accurate assessment on the risk factor of this individual. In turn, the insurer can offer the insured a reasonably priced premium. The data is transmitted from a box located in the car or from the person's smart phone app. This process helps insurance companies to bring value to their customers and stay in touch with their needs. The whole point is for the companies to offer better services to their customer base.

One insurance company, US insurers Progressive, is an example of a company using data to enhance their customer services. They have developed what they call "Business Innovation Garage" where their technicians will road test these

latest innovations. One example is they render images of damaged vehicles from computer graphics. The images are sent from cameras and constructed into 3D models showing the condition and extent of damage to the vehicle. This will speed up the claim process. It will enable the insurance company to process the paperwork quickly. They will have the condition of the car in real time. The customer will get the estimate quicker, and he will be able to get his car into the body shop that much quicker. The customer's car will back on the road in good working condition in no time. This will definitely help this insurance company increase their customer base.

There are still other methods of conducting predictive analysis, which I have not mentioned in this chapter. With these techniques, it is possible to reap the multiple benefits of predictive analytics.

Chapter 8: Why Predictive Analytics?

Now that you have a fair idea about predictive analytics, let us now understand why it has gained increased importance over the years. Predictive analytics has found its way into several fields and applications and has had a positive impact so far. In this chapter, I have highlighted the key applications, which employ Predictive analytics.

Analytical Customer Relationship Management (CRM)

This is one of the famous applications that employ predictive analytics. Different methods of predictive analytics are used on the customer data that is available with the company, with a view to achieve the CRM objectives. The CRM aims to create a holistic view of the customer, irrespective of where the information about the customer lies. Predictive analytics is also used in CRM for sales, customer services and marketing campaigns. The different methods of predictive analytics help the company in meeting the requirements of its huge customer base and making them feel satisfied. Some areas how predictive analytics is used in CRM are as below:

- Analyzing and identifying those products of the company that have the maximum demand.
- Analyzing and identifying those products of the company, which will have increased demand in the future.
- Predicting the buying habits of customers. This will help the company in the promotion of several of its other products across multiple areas.
- Proactive identification of pain points, which may result in the loss of a customer and mitigating such instances.

CRM can be used throughout the lifecycle of a customer, starting from acquisition and leading to relationship growth, retention and win back.

Clinical Decision Support Systems

Predictive analytics is extensively used in health care for calculating the risk of contracting certain diseases or disorders. It is primarily used to determine the patient's risk of developing health conditions such as heart diseases, asthma, diabetes and other chronic diseases. If you were under the impression that predictive analytics is used only for the purpose of diagnostic reasons, then you are wrong. Predictive analytics is also employed by doctors for making decisions regarding a certain patient who is under medical care at a certain point of time. How does it work? Clinical decision support systems aim at linking the observations of a certain patient's health with the knowledge about health. This relationship will aid clinicians in taking the right decisions with respect to the patient's health.

Here are some other ways predictive analysis is helping the healthcare industry:

Predictive analytics is influencing the healthcare industry. It could revolutionize Healthcare worldwide. Here are seven ways of how it will impact healthcare:

- Predictive analytics increases medical diagnosis accuracy
- Predictive analytics will support public health and preventive care.
- Predictive analytics gives doctor's solutions for their patients.
- Predictive analytics can provide predictions for hospitals and employers about insurance product costs.
- Predictive analytics lets researchers make prediction models without studying thousands of patient cases. The models will get more accurate over time.
- Predictive analytics help pharmaceutical companies develop the best medicines for the welfare of the public.
- Predictive analytics provide patients with potentially better health outcomes.

These seven points can have industry changing practices for the healthcare industry. With the ability to predict medical conditions with greater accuracy could save lives and medical expenses for patients. Malpractice suits will decrease, and potentially healthcare costs could decrease. Doctors being able to accurately diagnose illnesses can save patients and insurance companies thousands of dollars in individual cases. Potentially insurance companies could save untold millions in medical claims worldwide.

Imagine doctor's interpreting patient's sickness right the first time. There would be a domino effect. This should drop healthcare costs dramatically. People will

be able to afford healthcare insurance. Doctors can lower their rates and live in peace knowing they won't be sued for malpractice. Patients would save money because they wouldn't be spending money needlessly on prescriptions that weren't connected to their illnesses. Healthcare would be streamlined in the patient care process if doctors have the ability to use predictive analytics to give patients better care based on accurately researched information.

Elderly patients would be able to afford their medications and in some cases they are administered many prescription that could cost them thousands of dollars monthly. Pharmaceutical companies would use predictive analytics to research and manufacture medicines to better meet the needs of the public. It could be possible to reduce the inventory of existing drugs worldwide. How? Pharmaceutical companies can produce necessary medicines that would directly deal with particular conditions patients may have. By using predictive analytics to find accurate data about patient's health issues the companies can produce the exact medicines that will help cure those issues. If this practice were done on a wide-scale, eventually, unnecessary medicine thought to help treat conditions but didn't, could be removed from the market. Pharmacies, in turn, would be left with only medicines that were needed to treat various diseases, etc.

Back on the accurate diagnosis trend, this could be huge in medical care. Imagine a patient has a family history of having heart attacks, and their ancestors were prone to having heart conditions. The patient's doctor uses predictive analytics to predict the chances of the patient having a heart attack in the same age range as other family members have had. More to the point what if the patient is at the age where the risk of heart attack is imminent.

The patient's physician is aware of all these facts. The doctor can use predictive analytics to predict when the event of a heart attack could occur. Armed with this life-threatening knowledge the doctor shares the information with the patient. The doctor and the patient develop a health plan to reduce the risk of the heart attack or eliminate the risk all together. The patient may have extended their life for several years because of this health prediction. Better yet the patient may have saved their life in the here and now. These are the kind of radical changes predictive analytics can bring to the health care industry.

It is reasonable to assume the roles of the doctor and patient will change based on the new health data predictive analytics has/will bring to medicine. The patient being more aware of their care options will increasingly make health care decisions for themselves. Doctors will become more of a consultant in advising the patient on the best course of action regarding their health choices. It seems as if that trend is already taking place.

Today more than ever patients are aware of their health issues and options more than ever before. They are getting more involved in the direct decision-making process about their personal healthcare. Physicians seem to rely more on the

patient's awareness about health care plans than before.

These are just the beginning of wide sweeping changes that have to occur in the healthcare industry because of the innovation of predictive analytics.

Collection Analysis

There are several industries that have the risk of the customers not paying the payments on time. A classic example of this case would be banks and other financial institutions. In such cases, the financial institutions have no choice but to engage the services of collection teams to recover the payments from the defaulting customers. It is not surprising that there will be certain customers who will never pay despite the efforts of the collection teams. This is nothing but a waste of collection efforts for the financial institutions. Where does predictive analytics come into this? Predictive analytics will help the institution in optimizing the collection efforts. Predictive analytics play an important role in collection activities in the following manner:

- Optimizing the allocation of resources for the purpose of collection.
- Identification of collection agencies, which are highly effective.
- Formulation of collection strategies.
- Identifying those customers, against whom legal action is required to be taken.
- Formulation of customized collection strategies.

With the help of predictive analytics, it is easier for the financial institutions to collect their dues in an effective and efficient manner. This also reduces the collection costs for the financial institutions.

Cross Sell

This is applicable for those organizations, which sell multiple products. As I mentioned before, it is important that every organization has details of its customer base. When an organization has a comprehensive database of the customers, predictive analytics will help in the promotion of the other products of the organization. Predictive analytics helps in determining the spending capacity of each customer, their usage and other behavior, which makes it easy for the organization to promote other related products. This cannot only just lead to the increase in sales for the organization but also helps in building customer relationship.

Customer Retention

Customer satisfaction and customer loyalty are two of the major objectives of any organization. These objectives have gained more importance in the recent past, given the increase in the competition. Apart from these two major goals, another goal has emerged in the light of such heavy competition in the market, which is nothing but reducing customer attrition. It is not important if new customers purchase an organization's product or services. What is more important is that existing customers continue buying their products or services. When you are capable of retaining your customers, you are bound to increase your profits without much difficulty. Hence, it is highly important that each organization pays attention to the needs of the existing customers.

Most businesses tend to react to the customers' needs and not proactively address them. This attitude could easily break them, should their competitor be more empathetic towards the needs of the customers. Moreover, it is too late to change a customer's decision once he or she has decided to discontinue the service of an organization. No matter what the organization says or does, may not have that much of an impact on the customer. Ultimately, the organization will have to bend backwards to retain the customer, which is not only an added cost to the company but also an embarrassment.

With predictive analytics, it will be possible for organizations to be more proactive as opposed to being reactive. How does it work? With just a quick glance at the customer's service usage history and spending habits, on a frequent basis, it will be possible for the organizations to come up with predictive models. These predictive models will help in the identification of those customers who are most likely to discontinue the service of the organization. This gives the organization an opportunity to act proactively and figure out the reason behind the possible termination of service. It can also decrease the number of customers terminating its service in this fashion. Some strategies used by organizations to retain the customers who are on the verge of leaving include but are not limited to lucrative offers, discounts etc.

Another kind of customer attrition that is equally detrimental to the interests of a business is silent attrition. This is when a customer reduces his purchase of the organization's products or services, gradually over a period of time and discontinues eventually. Most companies are not aware of customers who have shifted loyalties this way. However, with the help of predictive analytics, it will be possible to pick out such customers who are exhibiting this behavior. As I mentioned before, predictive analytics mandates the need to monitor the customer's service history and expenditure regularly. This will help the companies in spotting those customers who have reduced their spending over a period of time. With this information in hand, it will be possible for the company to come up with a customized marketing strategy for such customers

and ensure that they continue availing the services of the organization.

Let's introduce some case studies to support the claim that predictive analytics can help increase customer retention and reduce "churning." Churning is where the customer stops buying products from a store or business. They begin shopping at another retailer's website or physical store.

Here are the case studies: Windsor Circle's Predictive Analytical Suite produces many predictive data fields to help vendors in their e-mail marketing campaigns:

- Predicted Order date–based on customers personal purchasing habits
- Replenishment Re-Order date, when the product they purchased will run out
- Product Recommendation–related to products customers have purchased historically
- Hot Combo-Products that are typically bought together.

The goal in predictive analytics in direct e-mail marketing is to combine Predicted Order Dates, Replenishment Re-Order Dates with product Recommendations and Hot Combo sales to retain customers by appealing to their predicted spending patterns. The vendors entice customers with future Hot Combo sells by reducing prices and predicting when those Hot Combo sells will sell out.

By using Windsor Circle's Predictive Analytics Suite, vendors can create models predicting which customers will buy which products in what particular time frames. Based on historical data showing previous product purchases; vendors can piece together effective direct e-mail marketing strategies to induce customers to keep coming back.

These predictive models can also predict when customers will purchase certain products together and which products they most likely would buy together, another component of these predictive models is they can foretell the right timeframes to keep products in inventory based on customer demand in the past.

Coffee for Less.com uses Windsor Suite's predictive data to run their replenishment e-mails to remind customers to stock up on their coffee purchases. Windsor automatically pulls up a predicted order date on customers who have made three or more purchases. They then trigger replenishment e-mails and send it to the customer. The e-mail is based on the customer's historical buying patterns. In this e-mail example, they would put a smart image to the left. This would then populate a particular product image based on the products the customer was getting the replenishment e-mail about. Shrewdly

enough Coffee for Less will use the rest of the e-mail page to suggest other products the customer will like. This e-mail is generated on predicted customer products they will most likely buy. So here Coffee for Less is using two data marketing strategies to retain this particular customer. They are Product Recommendations + Predicted Order dates all based on Windsor's predictive analytic models.

Normally, vendors use Windsor's predicted order date field to trigger product recommendations and replenishment e-mails. Australia's premiere surf and apparel retailer Surf Stitch used the predicted order date field in an innovative way. They use it to locate potential "churning" customers. (Defined previously as customers who will stop buying products from a particular vendor.) The usual timeframe for win-back e-mails to be sent out is in 60, 90 and 120-day cycles since the last customer purchase.

Surf Stitch has access to their customer's purchase history and their predicted order data. By using the replenishment e-mails in an innovative way (win-backs) they were able to reduce churning customers by a whopping 72% over a six-month period.

The case studies above show that companies (in many industries) can use predictive analytics in an innovative way. They can take standard data fields and use them in new ways to retain their customer base and win-back lost customers. Predictive analytics are so flexible in their data applications they can be used in thousands, potentially millions of ways. There are applications waiting to be discovered that no company has thought about using.

The possibilities that predictive analytics can/will have on company's business practices are unlimited. Sadly, though there are still companies in various industries that are lagging behind in using predictive analysis models. They are still using traditional predictive analytic methods that are no longer effective. These old methods are slow, cumbersome and not cost effective. They take too much manpower and time to facilitate all their processes. They slow down revenue sources, production and overall operational efficiency of these companies.

The old predictive analytical models aren't user-friendly and can't predict as many different variables as the new methods can. In the long run, the newer predictive systems will eventually be implemented in all the companies that are currently not using them. What will motivate the change? When the companies see that there traditional predictive analytical models are increasingly costing them their market share to their competitors. Internal processes will not be streamlined and run more efficiently. Their bottom line will not increase but decrease because they aren't using the newer predictive analytical models to their fullest capacity. They aren't using them at all!

Direct Marketing

For organizations that manufacture consumer products or provide consumer services, marketing is always an important matter of discussion. This is because, when marketing, the team should not just consider the pros and cons of their organization's products or services, they also have to take into consideration the competitor's marketing strategies and products.

With predictive analytics, this job is made easy. It can improve your direct marketing strategies in the following manner:

- Identification of prospective customers.
- Identification of the combination of product versions, which are the most effective.
- Identification of effective marketing material.
- Identification of communication channels, which are highly effective.
- Determining the timing of marketing strategies to ensure that it reaches the maximum audience.
- Reducing the cost per purchase. Cost per purchase is calculated by dividing the total marketing cost incurred by the company divided by the number of orders.

Insurance companies are using predictive analytics to develop personal marketing strategies to attract potential customers. They will analyze many sources of data to find out what customers like. They will analyze customer feedback data, which involves scouring their own websites and social media sites. Algorithms scope through unstructured data in phone calls, e-mails, and information revealed on social media about what customers like or dislike. Also, the insurance company will analyze how much time a person will spend on the FAQ section of their website. Even forums and message boards are scanned to see what a customer's preferences are. All is this done so the Insurance Company can make a unique profile for the potential customer.

I mentioned this in another section of this book that insurance companies are now using all data available to them. They need to focus individually on each customer's needs to stay competitive actually to stay in business. Insurance companies have always focused in on whether they are meeting the customer's needs or not in the past. But now they are doing this more aggressively through the new data analysis techniques available to them. Clearly they are attempting to determine if they are having a positive impact on their customers or not.

Marketing departments of the Insurance Companies are also using predictive analytics to track if customers will terminate their policies or not. Like other tracing processes the company will investigate a customer's behavior and

compare that behavior to other customer profiles who have actually canceled their insurance policies. This will help determine or predict whether a customer will cancel in the near future. One pattern data source that is analyzed is whether a customer has made a high or low amount of calls to the helpline. If the numbers are especially high or low, it will be flagged.

The flagging will alert the company to try and change the customers mind about canceling his policy. They might offer lower premiums or discounts on other services to get the customer to stay and having this data to make those calls before they cancel is critical. More importantly, they can spend more time solving the customer's issues before they happen, so predictive analytics is playing a crucial role in marketing. Insurance companies are now zeroing in on more customized customer profiles.

Summing it up big data is helping the insurance industry to improve its customer service, fraud losses and lower the price of its premiums. A side notes the FBI reports on average that customers pay $400-$700 more in premium costs because of fraud losses.

This is how predictive analytics plays an important role in direct marketing.

Fraud Detection

As I mentioned earlier in this book, fraud is an inherent threat for every organization. Lack of data makes it difficult for a company to even detect fraud. Fraud can be of the following types:

- Fraudulent transactions. This is not just restricted to online transactions but also involves offline transactions.
- Identity thefts.
- Credit applications, which are inaccurate.
- False insurance claims.

As I said before, the size of the organization does not shield it from fraud. Whether it is a small business or big corporate, both stand the risk of being victims to fraud. Classic examples of organizations, which are often victims of fraud, are retail merchants, insurance companies, suppliers, service providers and credit card companies. Predictive analytics can help these kinds of organizations to come up with models, which will be capable of detecting incorrect data or applicants and thereby reduce the company's risk of being a victim to fraud.

Predictive analytics plays an important role in detecting fraud, even in the private and public sector. Mark Nigrini developed a risk scoring method that is capable of identifying audit targets. He employed this method to detect fraud in the franchisee sales reports associated with an international food chain. Each franchisee is scored 10 predictors at the outset. Weights are then added to these initial scores, which will give you the overall risk score for every franchisee. These scores will help in identifying those franchisees that are more prone to fraud than the others.

This approach can be used to identify travel agents, who are fraudulent, questionable vendors and other accounts. There are certain complex models, which are developed using predictive analytics, which are capable of submitting monthly reports on the fraud committed. In fact, the Internal Revenue Service of the United States of America makes use of predictive analytics to keep a tab on tax returns and identify any fraudulent entries or evasion and also catch tax fraud early.

Another growing concern when we speak about fraud is cyber security. Predictive analytics can help you come up with models based on behaviors, which are capable of examining actions on a specified network, continuously, and spot any deviations or abnormalities.

Widespread fraud is a serious challenge for the insurance industry. But companies are fighting fraud on the front lines with profiling and predictive

analytics. The companies will compare current claims with claims that have been historically found to be fraudulent. Certain variables will be picked up by computer analysis that indicates a claim that is being filed is a fake. The fake historical claim has certain markings to show it is a fraudulent document. The claims are matched up, and the claim in question has the same markings as the historical fraudulent claim then it will be investigated further. The matches could be the behavior of the claimant. Criminals of fraud typically display certain patterns that can be detected by computer analysis. These might not get flagged by a human who manually these process claims, but computers can look at a volume of claims for triggers and alert the appropriate people of the occurrence for further review.

The other factors the company will look at are the people the claimant associates with. These are investigated through open sources, like their activity on social media, or they will check the other partners in the claim like an auto body shop. The insurance company will be looking at dishonest behavior the auto body shop may have engaged in the past. Credit reference agencies are checked out as well.

Operations

Sometimes, analysis can be employed for the purpose of studying the organization, its products or services, the portfolio and even the economy. This is done to identify potential areas of business. Businesses use these predictive models to even manage factory resources as well as for forecasting inventory requirements. For example, airline companies use predictive analytics to decide the number of tickets that they are required to sell at a specified price range for every flight. Similarly, hotels may use predictive models to arrive at the number of guests they may be expecting during a given night. This way, they can alter their price range accordingly and increase their revenue.

For instance, if the Federal Reserve Board is keen on forecasting the rate of unemployment for the upcoming year, it can be done so using predictive analytics.

Insurance Operations

Today's Insurance giants seem to be behind the times in their approach on how they use predictive analytic methods. The current methods employ many data analysts to run it. The method is too time consuming and it cannot handle the massive amount of data flying through insurance companies data portals to be effective systems. When you have data funneling through many sources, it makes it improbable for Insurance Companies to analyze data in a quick and efficient manner. These two factors are a must if the company wants to stay abreast a highly changing market place.

Companies to be relevant must adopt predictive analytic models that will turn business value-specific decisions and actions which will maximize operations across the board. Solutions must continuously improve and refine needing thousands of iterations. Analytic output must connect into business operations to optimize revenues and be relevant. Modern Analytics provides the following predictive analytic features to Insurance Companies who want improve business operations:

- Predictive analytics specifically for insurance
- Predictive Modeling for Insurance
- Fully operational predictive analytics
- Big Data analytics
- Prescriptive Analytics
- Customized insight and knowledgeable answers

Insurance Companies must realize that proper predictive analysis will improve overall business operations. Unfortunately, there is this mindset within the insurance industry that current predictive models are not efficient. They worked in former times but in today's market they've become obsolete and will eventually hurt the overall business operations revenue sources, resource optimization and sales nurturing of the Insurance Companies.

They need to have an automated predictive model in place to keep up with the influx of data constantly bombarding their systems now. The manpower labor is too costly to continue to use manual predictive analytic methods. These resources can be channeled into automated systems saving the companies a lot of time and money. True with the advent of automated predictive models you would have to hire additional technical support people who understand the complexities of running these automated systems. But the revenue increase in all the other areas will more than justify the manpower expense. Reluctance on the part of analysts and statisticians to fully automate predictive processes is too costly for Insurance Companies to continue operating in.

The predictive analytic model that Modern Analytics offers will be the very model Insurance Companies need to increase sales, decrease fraud and increase their policyholder's market instead of losing them. Modern Analytics fully automated predictive systems allows them to run thousands of models at one time. Their methodologies run much faster than traditional analytic systems can. Plus, they will assist insurance companies to increase overall productivity, increase daily operations, effectively locate strategic business investments and predict any changes in the current and future marketplace.

Their predictive analytic models will be game changers for any Insurance Company that chooses to implement their services. How? By optimizing business operations, improve internal processes and surpass competitors in predictive analytic processes. Modern Analytics works very closely with their clients, and they have helped a wide base of customers across many industries. They gather and structure data with their cutting-edge algorithms and cutting edge technology. They also rapidly supply top-notch data solutions uniquely tailored for each client.

Shipping Industry

UPS has adopted a new program called (On-Road Integrated Optimization and Navigation) Orion for short. UPS believes that Predictive Analytics needs to be implemented all the way down to the front line workforce. UPS is deploying this Predictive Analytics model to its front-line supervisors and drivers totaling 55,000 in all. The predictive model factors in the following variables:

- Employee Work Rules

- Business Rules
- Map Data
- Customer Information

These are used to predict the optimal delivery routes for the UPS drivers. The data sources will answer such questions like “

What is better to make a premium delivery 15 minutes earlier or is it better to save a mile of driving, to reach peak performance objectives. UPS is taking innovative steps to bring their predictive model to the execution. The front-line supervisors will learn to interpret predictive analytics data to maximize their delivery times on their routes, that is the key behind this move. Within UPS, these data skills are no longer just used by analysts.

UPS originally distributed the ORION system in 2008. The problem was, there were too many variables for the managers to manipulate. It confused them because they didn’t understand how these variables applied to them. UPS refined ORION and now only the essential variables have to be input by managers. The managers teach the drivers how to interpret the ORION predictive data. The data was put into terms the front-line could understand. Understand they did one driver over a few months cut 30 miles off his route interpreting the particular predictive data for his route.

Risk Management

Predictive analytics is employed to assess the risk associated with a business, in a fashion similar to how it is used to detect fraud. Credit scoring is the crux of ascertaining the risks associated with the business. Credit scores can determine the likelihood of a customer to default in payments. How is this done so? The credit score is nothing but a number, which is generated by a predictive model. This model consists of all data associated with the creditworthiness of the customer in question.

Employee Risk

There is another side of Risk Management that is typically overlooked. The risk associated with a company’s employees. Companies can use predictive analysis to gauge employees who may be a risk. There are two traditional methods in place to detect employees who could become troublesome. They are the “blanketed management” programs that focus more on non-risk employees and the employees at risk slip into by without being noticed. The other method is the “squeaky wheel” approach when company management focuses in on

employees that display active troublesome behaviors.

Another issue company's face is if they have thousands of employees how can they effectively monitor for employees at risk? They can't use traditional methods. So here is where predictive analytics comes into focus. Predictive analytics will provide specific based risk management programs. They use thousands of data points that are turned into tangible data. This data tracks subtle but profound changes in an employee's behavior that could lead to major behavioral issues down the road. Often these changes otherwise go undetected even with the staff of Risk Management employed. Previous trends are also analyzed. These components are interpreted to predict future outcomes of an employee's unusual behavior. This information arms the managers and gives them the ability to intervene in the employee's issues before their issues impact the business.

Dramatically undetected risk could cause employees to engage in tragic behavior. So, the intervention of the manager early can sometimes avoid serious negative behavior down the road. The managers can intervene with the right person, at the right time and on the right subject. These interventions can stop worker compensation claims and voluntary employee turnover.

Their blanket management methods are replaced with these streamlined task-specific models, which will save time, money and most importantly reduce risk. The value of these predictive models is they can historically analyze employees past behaviors to identify present and future behavioral patterns.

I need to point out here that when employees are hired in the beginning, they aren't at risk. It's events after their hiring that develop that put them at risk. Problems in the employee's lives can surface outside of the workplace. They could be having medical or personal issues that carry over into their performance at work. Again, as stated earlier predictive analytics will spot subtle changes in an employee's behavior from historical analyzing of their records.

Let's briefly talk about the transportation industry in relation to Risk Management. This industry is highly vulnerable to risk with its employees. The transportation uses predictive analytics to track unsafe driving patterns of its employees. The data reveals that the employee may be driving unsafely which could lead to ample violations, fines possibly costly losses due to accidents. People may even be killed in an auto or other type of vehicle accident. Specifically, one type of hazardous driving pattern could be the employee not paying attention to his driving because he has a sick elderly parent. He is too busy talking on the phone while driving, setting up doctor's appointments and other activities for this sick parent.

Based on historical data patterns the transportation manager can detect if the employee is driving abnormally. They could be braking extra hard or spending

too much time in idle. The manager can communicate with the employee in question to work out some type of solution to help the employee deal with their issues to avoid a costly accident, such as a reduced workload or adjusted schedule.

Predictive analytics can help a manager focus on why an incident happened, instead of what happened. Managers tend to focus on what happened, instead of why it happened. The manager's focus is directed on the correcting the issue and not necessarily the root cause. Changing focus on the incident can solve tie employee's personal issues much quicker.

Underwriting

Many businesses will have to account for their exposure to risk. This is because of the different services offered by the organization. They will have to estimate the cost associated for covering these risks. For instance, insurance companies that provide insurance for automobiles need to determine the exact premium to be charged for the purpose of covering every automobile and driver. In the case of a financial company, they will need to analyze the potential and the repaying capacity of a borrower, before they grant the loan. In the case of a health insurance provider, predictive analytics can be employed to figure out the following information – past medical bills of the person, pharmacy records etc. This information can be used to predict the quantum of bills that the person may submit in the years to come. The health insurance provider can arrive at a suitable plan and premium.

Predictive analytics can help you underwrite these risks. With the help of predictive analytics, it is possible to determine the risk behavior of a certain customer as and when they make an application. This can help these companies take a calculated decision.

Predictive analytics have also reduced the turnaround time for processing loans and other claims, based on the credit scores. The reduced turnaround time has made a big difference in the case of mortgages. These decisions are now taken in a matter of few hours. This is different from the erstwhile procedure whereby it used to take weeks for the financial institutions to offer a mortgage. Predictive analytics also helps in the pricing of these financial products, which can reduce the chances of default, for often the most common reason for default is the fact that the interest rates are high or the premium is high.

Recently the insurance industry has turned to predictive analysis to help solve the issues they face. They are using this technology to analyze credit scores to foretell loss-ratio performance and the behavior of the insured. Now insurance carriers are using predictive analytics to see into the future during underwriting. They are creating models to see how future policies will perform. They face a very serious issue where they really have no reliable data on homes they insure. There is a big risk that the houses they insure could decrease in value in the market. Or what if the house burns down? Is the house vulnerable to earthquake damage? Is the house located in a potential flood zone? These are the types of questions that predictive analysis will help them answer.

With unknown data about houses they insure, they are at risk of losing millions of dollars in natural disaster damage or weather damage. There is also the insurance/deficiency value and the potential for liability risk. Here is where predictive analysis comes into play. Predictive analysis is cheaper, faster and more efficient than traditional analysis methods. Time will tell if predictive analysis will be a real game changer or not.

Predictive Analytics A Case Study In Underwriting-Freedom Specialty Insurance

In the aftermath of the major economic slowdown of 2008, a risky underwriting proposal was submitted to the executive management team of Scottsdale Insurance Company. This case study is taken from the D&O insurance industry (D&O is Directors and Officers Liability Insurance. This liability insurance is paid to directors and officers or the organization itself. The money is reimbursement for losses or a loan of defense fees in case an insured experiences a loss because of a legal settlement.)

The Scottsdale Insurance Company accepted the proposal and Freedom Specialty Insurance Company was born. Freedom developed an industry first method, feeding external data into a predictive model to support risk intuition. The theory was class action lawsuit data could predict D&O claims. A multi-million-dollar unique underwriting platform was born. The dividends are beginning to pay off as Freedom has grown to $300 million in annual direct written premium. They have consistently kept their losses under control. In 2012, they were under the average industry loss rate of 49%.

Their model delivers a high level of honesty in their dealings. This has caused great trust among personnel from the underwriter to the executive branch in the parent company. The reinsured are confident as well. The point of this case study is to show how Freedom built, maintained and refined their underwriting model. In hopes that other companies can take this model and incorporate it into their underwriting operations. Using lessons Freedom learned along the way. The platform was a multi-team effort in its development. An actuarial firm built and tested the predictive model. The external technology supplier built the user interface and constructed the integration with corporate systems. Next, technology from SAS was used for many parts such as many data repositories, statistical analytics engines and reporting and visualization tools.

Their platform involves the following components:

- Data Sources
- Data Scrubbing
- Predictive Model
- Risk selection analysis
- Interface with corporate systems

These are described as follow:

Data Sources-The system integrates with six external data sources and receives data from corporate administrative applications. External sources, like class action lawsuit and financial information, are commonly used in the D&O industry. Other sources are specific to Freedom and drive their predictive model. Data is central to the platform's function, so Freedom spends a lot of time maintaining quality data and vendor activities. The back-testing and categorization processes illuminate vendor weaknesses and incongruities in vendor data. Freedom goes to great lengths to work with vendors to classify data and keep its quality high.

Freedom has realized they need to keep an eye on their eternal data as well. They apply strict checks to improve policy and claims data. Freedom absolutely maintains data freedom from vendor's identification schemes. It takes extra work to translate values but this insures Freedom can get rid of vendors quickly if they have to.

Data Scrubbing-The data goes through many cleansing cycles when it's received. This guarantees maximum usefulness. One example of this is 20,000 individual class action lawsuits are reviewed monthly. Initially, they were categorized by different parameters, but they are reviewed monthly to see if any changes have taken place. In the beginning, this process took weeks to complete when done manually. Now it's done in hours using advanced data categorizing tools.

Back-testing-This key process calculates the risk involved when an initial claim is received. The system will funnel the claim back through the predictive model; testing the selection criterion and changing tolerances as needed. The positive feedback loop refines the system through many uses.

Predictive model- Data is condensed and fed through a model, which uses multivariate analysis to determine the best range of pricing and limits. Algorithms evaluate the submission across many predetermined thresholds.

Risk Selection Analysis- This produces a one-page analytical report of suggestions for the underwriter. Comparisons with the same risks are displayed along many risk factors- industry, size, financial structure and other parameters. The underlying principle of the platform is that it's driven by underwriter logic helped by technology. It will never replace the human underwriter.

Interface with corporate systems- Upon a decision being made selected information is sent over to corporate administration. The policy issuance process is still largely done manually. There is room for complete automation in the future. The policy is issued, and the statistical information is looped back through the data source component. As claims file through loss, data is added to the platform.

Like any D&O operation, underwriters possess deep technical insurance

knowledge. They have had to adjust and become comfortable with a system that reduces mass amounts of information into several analytical pages, where the underwriters traditionally spent time collecting, collating and analyzing data. They now interface with policyholders and brokers in book and risk management. Here again, predictive analytics has streamlined the data process changing the traditional role of the underwriter.

Of course with this revised underwriting process in place, skilled technical personnel were added. These people have legal and statistical knowledge that enable them to be able to build predictive models in the financial field.

Incorporating this model has allowed Freedom to develop process expertise in many areas. The data management tasks-data scrubbing, back-testing, and categorization were learned from scratch. Initially, these processes were handled manually, but they have been continuously automated since their inception. In addition, the number of external sources is always expanding. Freedom is in the process of evaluating the use of cyber security and intellectual property lawsuits. The predictive model is constantly undergoing refinement. The tasks involved in the care, feeding and maintenance of the predictive model used in other industries has been developed in the D&O industry. Of special interest is after the actuarial initially built the predictive model, it took many months for Freedom to gain a full understanding of its intricate operations.

Processes were put in place to effectively manage all these external vendors together. Several external groups came together to pull the predictive model project together each group having to work in tandem with the other groups. Again a list of those groups:

- Actuarial Firm
- IT Firm
- Data vendors
- Reinsurers
- Internal IT

I mention, again, the groups involved in this project because it took a very skilled and concerted effort, by Freedom, to bring these entities together to work their skills in the same place at the same time.

Success of the Platform

This predictive analytic model was very successful for Freedom. It has opened up new horizons for the Insurance Company. First, the analysis was so specific

it now opens up dialog with brokers and policyholders in risk management discussions. Secondly, this model may be extended beyond liability lines to property lines. Expanding this model will make it applicable to other lines of liability such as surety. This predictive analytical model makes Freedom stand out from its competitors. Finally, the methods used in the back-testing and categorization elements can be used to forecast other data element. Freedom has positioned itself to discover other potential information sources.

Predictive Analytics is Impacting Real Estate

How can Predictive Analytics impact Real Estate?

One example is a Real Estate corporation that used data devices to help a law firm decide if it needed to move into a different office space.. They started out by addressing employee retention factors. The law firm wanted to recruit and maintain the best prospects. As this was the major concern for the client, the Real Estate took this initiative instead of approaching it about a direct real estate suggestion. The firm used various location-aware devices to track the movements of the employees. The tracking data was implemented into data mapping based on employee preferences and movements.

The interpretation of the data resulted in the law firm moving out of the high rise office space into a cheaper office space. The data saved the law firm money and positioned them into a new location. The improved location solved the employee retention problem.

Predictive Analytics is changing the Way of the National Association of Realtors (NAR)

Predictive analytics is helping NAR re-think the way they conduct business. They are America's largest trade association at 1million members. So their results with predictive analytics could influence the Real Estate industry dramatically. The NAR is looking how to add value to their client relationships. They established a new analytics group who will:

- Analyze member and customer data trends
- The NAR will use the data results to add value to its realtors
- Use disparate data models to build analytical models
- The models will solve complex problems in the housing industry
- Help Real Estate agents, state and local associations and others make better data driven decisions

The data group will step forward in three phases. The first phase is the experimentation phase where the data group will interpret the accumulative data and identify trends and patterns. The last two phases will involve forming partnerships with other data users; the partnerships will design and develop products to help their agents. The golden result of applying predictive data models will help the NAR to uncover behavioral patterns and build targeted meetings with potential home buyers. This will result in increased operational processes for NAR. Ultimately, the NAR will increase their financial bottom line.

NAR's traditional process of analyzing data is to gather holistic big picture proprietary data. The data extracted show what had happened historically in the group's business dealings. The other part of analyzing data is gleaning information from the customer relationship management system to determine current events affecting the group.

The new and improved big data analyzing process will interpret focus on the big picture and the finer details in one source. The new process will analyze both proprietary and public data systems to discern trends that will predict the future in real estate business practices.

On the other side of the coin Insurance companies will be changing the way they conduct business as well. As predictive analytics helps doctors and patients consult to create better and faster health care solutions. Patients will stay healthier longer and will heal faster because the right treatment options are being administered to them to meet their physical needs. Insurance companies will have reduced premiums and claims to pay on. Yes, the insurance companies will lose millions of dollars in premium costs because patients won't need extensive and unnecessary medical procedures. This will surely drive down health insurance costs.

But insurance companies can increase revenues in new ways as the medical profession becomes more focused and singular in its individual patient health care plans. This will reflect in specialized but shorter hospital admissions. Insurance companies may become more specialized in what they cover. The result will be new avenues of revenue.

Medical device companies will be impacted as well. The familiar medical devices they have been manufacturing to take care of traditional healthcare issues will diminish. The obvious conclusion lost revenues for these companies. Here too though, the device companies can adjust to the new playing field. They can too shift gears and start manufacturing the specialized devices that will undoubtedly be needed in the new healthcare environment. Thereby, bringing increased revenues to their financial bottom line as well. Predictive analytics is bringing widespread change to every facet of the healthcare system. There is yet to come unseen changes as well as predictive analytics processes become more

refined and increase in accuracy.

Predictive Analysis is Changing Industries

It is obvious that Predictive Analysis is changing the game drastically in how data is collected, interpreted and predicted. This section on "Why Predictive Analytics" shows that Predictive Analytics is changing the face of so many industries internationally. Many industries are on the cutting edge of the advancements Predictive Analytics is making in various applications in specific phases of those industries. Freedom Insurance who has developed a predictive model that is spreading to other types of applications in the Insurance industry has placed themselves as the leaders in their respective industries.

Predictive analysis is reported to be a very lucrative market reaching 5.2billion-6.5 billion by 2018/2019. Mind you not millions but millions will be spent in this market.

Take the healthcare industry for example predictive analytics is changing the way physicians take care of their patients. Traditionally, the doctor would just give the patient a medical treatment to follow. So the patient could recover and return to his/her full health. Now physicians have become more like counselors to advice patients on the best course of treatment to take. The patient is intricately involved in his own care now because of predictive analytics. The patient has access to his medical records through his personal physician.

Gone are the days where the patient was unaware of his medical history or vulnerability to different diseases because of genetic factors and family history. Now because of Predictive Analytics patients view their records and witness the multiple predictive models that will steer them into the best care for their future health. They now have a strong influence in which predictive treatment will best bring about the desired result-good health. Here the role of doctors and their patients are gravitating to definite role changes. Who would have thought 5-10 years ago Predictive Analytics would change healthcare so much? (Maybe someone somewhere created a predictive model that anticipated the different changes that are affecting the health industry now).

In all the research done for this book concerning Predictive Analytics, every time the predictive analysis applications have been used, they have improved the processes of companies. Every single case researched for predictive analytics has shown that these newer models can help industries improve their overall operational capabilities. Some industries have been radically changed by Predictive Analytics:

- Law enforcement (Crime Scene Reconstruction)
- Insurance (Risk Management concerning employees and customers)

- Transportation (Predicting troublesome behavior and averting thousands of citations, arrests, millions of dollars lost in property damage due to accidents)
- Healthcare. This is a big one because Doctors can now effectively diagnose illnesses, conditions surgeries based on predictive models. Patient and Physician's roles are overlapping at integrating together. It's more of a democratic process where patient, and doctor jointly decide the best treatment options. Many more lives will be saved because of updated and accurate health care treatments based on healthcare predictive models.
- Retail Industry. Predictive analytics are helping retailers to increase customer retention rates. In some cases, by large percentages.
- Real Estate. The NAR for example, is using Predictive Analytics to predict what preferential property packages consumers will likely buy according to historical home buying practices. Rental property investors are using predictive analytics to predict where the most lucrative rental properties may be found based on renter's economic, geographical and family needs.

It's not unrealistic to say that Predictive Analytics has caused industrial revolutions in some industries. The potential market will exponentially increase in the next few years as more and more success stories are detected about innovative ways Predictive Analytics are being used by visionary companies.

Chapter 9: What is Data Science?

Data science is defined as "the deep knowledge discovery through data inference and exploration". Data scientists need to have experience at algorithmic and mathematical techniques so that they can solve some of the most complex of problems, using pools of raw information to find the insights hidden beneath the surface. Data science is centered on analytical rigor, based on evidence and on building decision-making capabilities.

Data science is important because it allows companies to operate more intelligently and to strategize better. It is about adding significant value by learning from data. A data scientist may be involved in a wide variety of different projects, including:

- Tactical optimization – improving business processes marketing campaigns etc.
- Predictive analytics – anticipating future demand, etc.
- Nuanced learning – development of a deep understanding of consumer behavior
- Recommendation engines – recommendations for Netflix movies, or Amazon products, etc.
- Automated decision engines – self-driving cars, automated fraud detection, etc.

While the object of all of these may be clear, the problems that arise require extensive expertise to solve them. It may be that a number of models need to be built, such as predictive, attribution, segmentation, etc. and this requires an extensive knowledge of machine-learning algorithms and a very sharp technical ability. These are not skills that you can pick up in a couple of days learning, they can take years. Below we look at the skill set required to become a data scientist:

Data Science Skill Set:

Data science is a multidisciplinary job and there are three main competencies required:

Mathematics

The very core of determining meaning from data is the ability to be able to see that data through in a quantative way. Data contains patterns, textures, correlations and dimensions that are expressed numerically; determining any meaning becomes a kind of a brainteaser requiring mathematical techniques to solve it. Finding the solution to any number of business models will often require that analytic models are built, models that are grounded in the theory of hard math. It is just as important to understand how the models work as it is to understand the process of building them.

One of the biggest misconceptions about data science is that it is all about statistics. While these may be important, they are not the only math that has to be understood by the scientist. There are two branches to statistics – classical and Bayesian. Most people who talk about that are talking about the classical ones but a data scientist needs to understand both types. As well as that, they have to have a deep understanding of linear algebra and matrix mathematics. In short, a data scientist has to have a very wide and very deep knowledge of math.

Hacking and Technology

Before we go any further, let me just clarify something – I am NOT talking about breaking into computers and stealing sensitive data when I talk about hacking. I am talking about the ingenuity and creativity required in using learned technical skills to build models and then find the right, clever, solution to a problem.

The ability to hack is vital because a data scientist needs to be able to leverage technology to get the vast amounts of data sets and to work with algorithms that, for the most part, are complex. Just being able to use Excel is not going to cut it in the world of a data scientist. They need to be able to use tools like R, SAS and SQL and for that they have to have the ability to code. With these tools, a data scientist is able to piece together data and information that is not structured and bring out the insights that would otherwise remain hidden.

Hackers are also algorithmic thinkers – they are able to break down a messy problem and turn it into something that can be solved. This is a vital skill for a data scientist especially as they work very closely with algorithmic frameworks

that already exist, as well as building their own, in order to solve an otherwise complex problem.

Business Acumen

One of the most important things to recognize is that a data scientist is a strategy consultant, before anything else. Data scientists are valuable resources in companies because they and they alone are in the position to be able to add significant value to the business. However, this means that they have to know how to approach a business problem, how to dissect it and this is just as important as knowing how to approach an algorithmic problem. Ultimately, value doesn't come from a number; it comes from the strategic thinking that is based on that number. In addition, one of the core competences of a data scientist is the ability to tell a story using data. This means they have to be able to present a narrative that contains the problem, and the solution, derived from insights gained from the data analysis.

What Is A Data Scientist?

One of the defining traits of a data scientist is their ability to thin deeply coupled with an intense curiosity. Data science is about being nosy, about asking questions, finding new things and learning. Ask any true data scientist what the driving factor is in their job and they will not tell you it's money. Instead, they will tell you that it is all about being able to use creativity, to use ingenuity to solve problems and to be able to indulge curiosity on a constant basis. Finding a meaning in data is not just about getting the answer, it is about uncovering what is hidden. Solving problems is not a task; it is a journey, an intellectually stimulating one that takes them to the solution. Data scientists are passionate about their work and they take great satisfaction in meeting a challenge head on.

How Analytics And Machine Learning Are Linked To Data Science

Analytics is now one of the most-used words in business talk and while it is used quite loosely in some cases, it is meant as a way of describing critical thinking of a quantative nature. Technically, analytics is defined as the "science of analysis", or, in easier terms, the process of making decisions based on information gained from data.

The word "analyst" is somewhat ambiguous as it covers a range of roles – operations analyst, market analyst, financial analyst, etc. Are analysts and data scientists the same? Not quite but it I fair to say that any analysts are data scientists at heart and in training. The following are a couple of examples of how an analyst can grow to be a data scientist:

- An analyst who is a master at Excel learns how to use R and SQL to get into raw warehouse data
- An analyst who has enough knowledge of stats to report on the results of an A/B test goes ahead and learns the expertise needed to build predictive models with cross validation and latent variably analysis.

The point I am trying to make is that to move from analyst to data scientist requires a great deal of motivation. You have to want to learn a lot of new skills. Many organizations have found a great deal of success in cultivating their own data scientist by providing the necessary resources and training to their analysts.

Machine learning is a term that is always used when we talk about data science. Put simply, machine learning is the art of training algorithms or systems to gain insight from a set of data. The type of machine earning is wide ranging, from a regression model to neural nets but it all centers on one thing – teaching the computer to recognize patterns and recognize them well. Examples include:

- Predictive models that are able to anticipate the behavior of a user
- Clustering algorithms that can mine and find natural similarities between customers
- Classification models that can recognize spam and filter it out
- Recommendation engines that can learn, at an individual level, about preferences
- Neural nets that learn what a pattern looks like

Data scientists and machine learning are tied closely together. The scientist will use machine learning to build algorithms that are able to automate some elements of problem solving, vital for complex projects that are data driven.

Data Munging

Raw data is often very messy with no real structure and data munging is a term that we use to describe the process of cleaning the data. This is so that it can be analyzed and used in machine earning algorithms. Data munging requires very clever skills in hacking and the ability to recognize patterns so that vast amounts of raw information can be merged and then transformed. Dirty data hides the truth that may be hidden beneath the surface and, if it isn't cleaned, it can be misleading. As such, a data scientist has to be good at data munging so that they have accurate data to work with.

Chapter 10: Further Analysis of a Data Scientist's Skills

One of the best people from whom to seek a worthwhile opinion regarding the skills necessary for a data scientist is someone whose job is to hire data scientists. The reason is that recruiters know precisely what skills they are looking for in the potential employee or consultant. Burtch Works is one such firm that deals with recruitment of senior personnel in the field of business and industry. They have some work that was published in 2014, and it spells out the skills a data scientist should have. Such information is credible, not just because it comes from a renowned firm of experts, but also because the firm itself has managed to walk the talk and climbed up the ladder of success to join the Fortune 50.

Why then not borrow a leaf from Burtch Works and implement their recommendations in regards to equipping potential data scientists with relevant skills?

Further Skills Recommend For A Data Scientist

1) Solid educational background

From research, it is evident that people who manage to specialize as data scientists have a solid educational background. In fact, a respectable 88% of them have a Masters degree; a good 46% are PhD holders; while others are generally well educated. Being a data scientist demands that the person is capable of developing the required depth of knowledge, and one can only do that with a good level of education.

Yet, even with a good level of education, a potential data scientist needs to have an inclination towards fields that declare you are not afraid to do calculations; fields that deal with analysis of numbers and formulas and so on. According to Burtch Works, 32% of data scientists have a great background in Mathematics as well as Statistics; 19% in Computer Science; and 16% in Engineering.

2) Competent in SAS plus/or R

SAS stands for Statistical Analysis Software, and it is only natural that a potential data scientist should be comfortable using software that helps in data analytics at an advanced level; data management; predictive analytics; and such. It is even better if the person knows something about the R programming language helpful in creating important functions. Generally speaking, any good analytical tool available is good for a data scientist.

3) Skills in Python Coding

According to Burtch Works, employers who seek to employ data scientists want someone who is has the technical skills in the use of Python, which is a popular coding language. Often they are thrilled if the person is comfortable in the use of Java; Perl; or even C/C++.

4) Knowledge of Hadoop

It is important that a data scientist be skilled in the use of the Hadoop platform. These technical skills may not be mandatory, but it helps to be able to derive statistical data with ease from such an open source library. Other technical skills that are preferred include competence in the use of the Apache Hive, which is some data warehouse software. The Hive helps in querying data using a language almost similar to SQL, which goes by the name HiveQL.

And whoever is skilled in the use of Apache Pig on top of Hadoop stands a good chance of securing a job as a data scientist. Apache Pig is another platform that helps in data analysis. Also, as long as a person is interested in making a career as a data scientist, it is advisable to get familiar with existing cloud tools like Amazon S3, and any other that may be developed over time.

5) Skills in SQL Database

It is important that a potential data scientist be able to work with Structured Query Language (SQL), which is a programming language used in data management and stream processing of data. A data scientist needs to be skilled at writing complex SQL queries as well as executing them. It is helpful for the data analysts to know tools that will help with analyzing and compiling data. Some of these tools are TOAD, DataSpy, Erwin, and many more effective modeling and architecting data tools and software.

6) Ability to work with unstructured data

This is one area where a solid background in education helps a lot. As a data scientist, not every aspect of the work involves calculations. And even if it does, one needs to be able to determine which data is relevant in what equation for the outcome to make business sense. A high level of competence in critical thinking is, therefore, required, to be able to maximize the benefits of massive data from the social media; video feeds; and other sources.

7) Having Intellectual Curiosity

You surely can't afford to have someone for a data scientist, if they rely solely on memory, like in remembering formulas and facts, and other stuff in the books. A data scientist needs to be someone eager to learn more about things that are happening in the world, which can impact the cost of doing business, improve efficiency in doing business, or generally impact profitability. Intellectual curiosity is one of the important soft skills Frank Lo, DataJobs.com's founder, wrote about back in 2014, as a guest blogger for Burtch Works. It is a non-technical skill, but it makes a data scientist stand a cut above others in the field of analytics

8) Competence in understanding a business situation

You can bank on someone with great business acumen to help your organization better as a data scientist, than one who is least interested in matters of business. Business acumen is another of those non-technical skills that are, nevertheless,

very helpful when a data scientist is analyzing data with a view to solving business problems. A data scientist with business acumen is able to analyze data and deduce what the lead problem for your business is, and what factors are escalating it. This is very important because it then helps you in prioritizing business decisions in a manner that allows you to protect the business, even as you capture fresh business opportunities. On the overall, your organization can leverage whatever data it has better than otherwise.

9) Skills in Effective Communication

Ever listened to learned people who addressed a group as if to impress them? That shouldn't happen with a data scientist. The marketing or sales manager, for example, would care less if you used the IFFEROR or VLOOKUP formula; the INDEX+MATCH formulas; or even went for the popular IF formula in your data analysis. All they want is information that tells them the current market situation as deduced from available data; and possible scenarios from that same mass of data. Data analysis is meant to provide quantified insights into the business situation, and unless those insights are communicated clearly and effectively to the relevant decision makers by the data scientist, all resources spent on data analysis will have been wasted.

Further Demystification Of Data Science

Have you heard someone describe data science as the sexiest job of this 21st century? Well, those were the sentiments of expert contributors of the magazine, Harvard Business Review, Thomas Davenport and his counterpart, D.J. Patil. Although there are some skills that would make you better placed as a data scientist, different employers often have different needs in mind when they advertise for a data scientist. So, it is important that you don't give advertisements a casual look but a deep one; reading the details mentioned therein. In short, when it comes to daily practice, the term data scientist is a blanket term often covering four categories of engagements as follows:

1. Data Analyst

What we are saying here is that although you do not have the qualifications mentioned earlier on of a data scientist, you would be doing yourself disservice if you are great at Excel and have some knowledge of MySQL database and dismiss a job advertisement just because its banner is data scientist.

At times people are hired to fill in the vacancy of a data scientist yet they end up spending their days spooling reports and data from MySQL or doing work on Excel pivot tables. Other times they are called upon to work with a team on some Google Analytics account. Whereas you may not be doing the deep work expected of a data scientist, the tasks just mentioned are not irrelevant or useless. Rather, they accord you a platform to practice the basics of a data scientist's job if you are new in the field. In fact, if you were ever hired for such a post, you can take the opportunity to explore additional use of data analytics beyond what the company demands, hence expanding your skill set.

2. Data Engineer

When they advertise for a data scientist and somewhere in there you find the mention of duties you are comfortable with as a data engineer, don't let the job slip by. If the job description includes things to do with statistics; machine learning expert; and such other roles that you can do as software engineer, it is fine to go for it. Alternatively, you may find the job description being one of building data infrastructure for the organization. Companies often make such advertisements when they have too much data around and they do not want to discard it because they feel they are likely to need it sometime in the future. Other times they see the data as being important, but since it is unstructured they don't know how to put it to good use.

3. Statistician

Sometimes companies dealing with data based services need someone who can handle consumer focused data analysis or even to deal with intense machine learning activity. Most of such companies actually run a data analysis platform. The tasks involved call for someone who is great at mathematics or statistics; or even someone with a background in physics; presumably people who are looking forward to advancing their academics along those lines. So, if you see a data scientist's job for a company of this kind that produces data driven products, do not hesitate to apply if you are specializing in mathematics, statistics or something relating to calculations and analysis of figures – you could be just the person with the skills they require at that time.

4. General Analyst

You may find a big company seeking to employ a data scientist, but when you look at the details you find that the tasks the person is to handle deal more with data visualization or even machine learning. It is alright to apply for the job because it means mostly the successful candidate is going to join an experienced team of data scientists and just want someone to help out in the lighter chores. What a great learning experience it would be for you especially if you haven't had much experience in a position of a data scientist! In short, if you are skilled in some big data tools such as Hive or Pig, you should be comfortable applying for this job irrespective of the heading, 'Data Scientist'.

5. Data Architects & Modelers

With the increasing amount of data that companies are required to keep and maintain, it is necessary to have people that can transform data from all the various systems into databases that can make more structured sense of it. This newly structured data is used to identify risk, fraud and alert triggers the data might need to have. The Modelers and Architects role is to work with development project teams to make sure any system changes get sent down to the data repository and will be in a format that can be used for reports and functions needed to use that data.

The Future Role of the Data Scientist

Experts are debating the future role of data scientists. Will they become extinct by 2025? In the different regions of the data world, there is an emerging opinion that data scientists could become unnecessary personnel anywhere from 5-10

years. Others say this could become the reality in 50 years.

There are those who say the need for a data scientist will always exist. The core group who say data scientists will never become extinct base this on what is called "expert level tasks." The basic assumption is that there are some data science tasks that are too complicated for robots or automation to be able to perform. The need for human creativity and innovation in data expert level tasks will always be needed. Robots won't be able to think outside the box when data interpretation calls for new data model methods to be applied or built to solve unknown or hidden business issues arise.

The other side says that all data expert level tasks no matter what their level of complexity will be automated within 5-10 years. Software tools will be able to complete complex tasks that data scientists now perform. One example that was given is that with software tools like Tableau the very difficult task of visualization has been mastered with an application. Second generation data science companies are making software tools that will improve company workflow and automate data interpretation. Currently, this issue is far from solved, so only time will tell if the Data Scientist's job is in jeopardy or not.

Chapter 11: Big Data Impact Envisaged by 2020

Have you realized that it is not just advertising and marketing that organizations are taking online these days? If you consciously think about what takes you online on a daily basis, you will realize that a good part of it is business, sometimes because you feel you can find a wider choice of items there, plenty of information about the product or service that you want, and even a wide range of pricing. The web also seems to provide much more entertainment than you could physically reach in a short time, and most probably at a relatively smaller price. And when it comes to connecting with others, individuals, private and public institutions, and all manner of organizations are taking their communication online, where they can reach a wider audience much faster and more cheaply.

How Does Moving Online Impact Data?

Well, it means that the amount of data being generated on a daily basis is growing exponentially, and we can no longer ignore big data. Secondly, even before anyone can speak of the ability to analyze and organize the massive data, tracking it is itself a mammoth challenge. Would you believe that internet users are generating data in excess of 2½ quintillion bytes each day? This includes the automated feedback like traffic monitors, weather related trackers, and all manner of transactions. Is this a good thing?

Well, potentially, it is. But there is the question of what sources are reliable and which ones are not; what data is relevant to your scenario and which one is not; and so on. As you may already know, having massive data before you can also be overwhelming. That is why big data has created room for unique business, where you get people specially training to make good use of big data. That potential power of big data is what specialized companies seek to help you unravel and take advantage of, by handling for you big data that affects your organization.

In fact, specialized technology has come up to make big data management convenient. A good example of this is Apache™ Hadoop®, an advanced database management technology that takes you beyond consolidation of information to improved efficiency of your business and increased profits. As long as organizations are open to the use of advanced technology that makes big data analysis and management convenient, the world is headed for better times.

Instead of being overwhelmed by high data traffic and accumulated mass of data, organizations are going to unleash the power of that data and if they are in education they are going to bring down drastically the cost of education. If they are in meteorology, they are going to have better comprehension of certain complex phenomena like the weather. Those in business will be able to raise their productivity by drastically cutting on all manner of redundancies; and the job market is going to have better correlated data sets that will help to match job seekers against potential employers as per the skills offered and needed respectively. Some of these things are already happening and they can only become better.

And the story of big data does not end with reduced cost of doing business and acquiring education; increased efficiency and productivity; as well as increased profits. Research that is already going on points to the potential of improving the fight against crime; significant improvement of web security; and ability to foretell early enough when there is likelihood of an economic or natural disaster sometime in the future. In fact, as long as there is work going on regarding big

data and its intricacies, the world is likely to witness bigger and more radical changes, much of it for the better.

What's The Market Like For Big Data?

As already explained, there are many innovators trying to create analytical and data management tools that will turn the massive data generated into an economic advantage. In 2012, the big data market was worth $5 billion. If you remember what we said earlier that big data has been growing exponentially, you'll not be surprised to know that going by the situation today, market for big data is projected to be $50 billion by the time we get to 2017.

Specific Ways Big Data Can Significantly Impact Life – Soon

1. Websites and applications functioning better

In this regard, not only is it expected that it will be much easier and convenient to navigate websites and make sense of data availed there, but also that the sites will be much safer than they are today.

This is particularly so because big data can be handy in identifying and tracking fraudulent activity on a real time basis. Big data can introduce fresh and clearer visibility on an organization's website, and also make it possible to foretell when attacks are imminent. Innovators have actually already begun designing programs geared towards safeguarding data against destructive intrusion. For example, there is the machine learning program known as MLSec that you can find at MLSec.org. and it uses algorithms under supervision, to locate networks harboring malicious programs. It must be very encouraging to web users to learn that this machine learning program has been proven to be accurate to the rate of 92% - 95% for every case tested.

How Bad Is The Security Situation Today?

Well, as per 2012 statistics:

- Of all the websites hacked in the year, 63% of the owners did not realize they had been invaded.
- In fact, 90% of the web owners did not even seem to notice anything strange at all going on within the site.
- 50% of the web owners learnt that their websites had been hacked from their browser warnings or even warnings from the search engine they were using.

2. Higher education becoming more accessible

Why is it so exciting that cost of education would fall? Well, accessing higher education is generally a problem for the average person in most countries. In the US, where it looks from the outside like the land of plenty, the cost of tuition rises at a double rate compared to that of healthcare. And when you compare the hike in cost of education to that of the country's Consumer Price Index, it goes four times as high.

Luckily, just as websites are becoming better designed and protected, something is happening to make education more easily accessible to more and more people. A good example is the emergence of online sites offering great courses. The Khan Academy found at khanacademy.org; Big Data University found at BigDataUniversity.com; Venture Labs found at venture-labs.org; Coursera found at courser.org; are just some examples of institutions that are offering higher education online at a much lower cost than conventional tertiary institutions; and sometimes free of charge.

The good thing about many of the courses offered in this manner is that students are tested on how well they have clinched the skills taught, particularly because they the skills taught are applicable in the current high-tech environment. For example, Big Data University teaches Hadoop plus some other big data related technologies.

3. Relative ease in getting a job

Have you ever given thought to the number of job seekers there are worldwide? It must be a staggering figure. In fact, on a monthly basis, the number of job searches on the web alone has hit 1.5 billion. On the positive side, there are websites that have already begun to match job seekers with potential employers by amassing valuable information on various employers as well as information on interested job seekers. One such website is indeed.com.

4. Improved road safety

Are you aware that car accidents are the main cause of death in the category of youth aged between 16yrs and 19yrs in America? And while it is dangerous to drive while under the influence of either drugs or alcohol, ¾ of the deaths in this category are not alcohol or drug related. What this means is that there are very many accidents that occur because of merely poor judgment. It is envisaged that as scientists continue to work on advancing the technology being used on computers, big data is going to help them predict the behavior of drivers on the road, as well as the move a vehicle is about to make at a certain point.

This route is geared towards reaching a point where cars on the road can exchange data, so that drivers in different cars can see up to three cars that are ahead of them; three that are immediately following them; and three on either side of them at any one time. In fact, it is said that big data will enable the drivers in data swapping vehicles to see the posture and focus of a driver in a car near theirs. This may sound a little far-fetched, but think about it: Haven't Google's self-drive cars taken the auto industry to a totally new level?

5. Ability to predict business future

What this means is that organizations will be able to use software like Hadoop to analyze the data at their disposal in a way that will bring to the fore future possibilities. The speed and accuracy with which data will processed will enable organizations to take prompt action where there are business opportunities emerging; damage control needed; and such other actions that require accurate assessment. There are already big organizations using Hadoop with great success, including eBay; Twitter; FaceBook; Disney and others. The demand for Hadoop is rising rapidly. IDC, a renowned market research firm has predicted that by 2016, the conservative worth of this software will be $813 million.

Another good example is Recorded Future, a technology company based on the web. This one provides security intelligence mostly to data analysts, which they use to keep their information safe. It puts businesses in a situation where they can anticipate risks and also capitalize on business opportunities by unlocking predictive signals using clever algorithms. There are other examples already helping businesses prepare for eventualities, but suffice it to say, as technological development continues, it will become all the more possible to leverage data, hence avoiding surprises.

6. Ability to predict matters of weather

This ability to predict the weather brings the advantage of being in a position to protect the environment better. Such protection in turn brings in huge savings to the country and the world at large considering the massive expenses that are brought about by weather related disasters. Just for example, weather and climate related disasters cost the US losses in staggering figures; actually in excess of $1 billion. Here, we are talking of disasters such as drought; wild fires; incidences of flooding and storm; and such other unfortunate events.

Things are already looking bright on this front of data technology. There is the Joint Polar Satellite System (JPSS), for instance, that is set to be launched in 2018. The JPSS, which will have the capability to utilize sensor technology, will also use data to a hurricane's path or a storm's path, well before these disastrous events occur. This will then give everyone concerned time to plan what to do to safeguard life and property. As CNBC News has already noted, situations that relied on guesswork some years back are slowly but surely becoming something to predict with precision using predictive science.

7. Healthcare becoming all the more efficient

Big data is expected to bring improvement to the health sector, not just by raising efficiency levels in service delivery, but also by customizing services to suit respective consumers. McKinsey & Company, an advisor to the management of many renowned businesses, says that between 50% and 70% of business innovations depend to a good extent on the capacity to capture customer's data and not as much on external analytics. McKinsey & Company relies heavily on qualitative as well as quantitative analysis, to be able to give helpful advice to management.

It is said that 80% of data in the medical sector is unstructured. However, with the trend the health sector in the US has taken in using big data creatively, great improvement in service delivery is anticipated. It is actually envisaged that big data is going to help the sector increase value through efficiency, adding value in excess of $300 billion each year. By the same token, expenditure is anticipated to reduce by a good 8%.

In treating patients effectively, caregivers benefit a lot from available patient data. The reason is that the data helps the caregivers provide evidence based advice. Currently, a medical center within Boston by the name of Beth Israel Deaconess, is putting a Smartphone App in the market, meant to help medical care givers access 200 million data points. These data points are expected to avail data concerning around 2 million patients. There is also Rise Health that utilizes the accessible mass of patient data, analyzing it from all dimensions, and aligning it to the health providers' goals with the aim of improving healthcare through fresh insights. On the overall, big data brings speed to

innovation. A fitting example is the project on Human Genome that took a whole 13yrs to complete, yet today it would only take a couple of hours to accomplish it.

Chapter 12: Benefits of Data Science in the Field of Finance

As far as data is concerned, there is no longer a shortage. There may even be excess of it all around, if you consider the traffic going through social media; real time market feeds; transaction details; and elsewhere. The volume of data available for use in the finance sector is almost explosive. Its variety has also expanded, and even the velocity at which the data becomes accessible has sharply risen. That scenario can either take organizations to heights unknown before, or leave them dumbfounded from the feeling of overwhelm caused by the data influx.

Since organizations are in business to succeed, they have learnt that the best way to utilize this flood of data is to engage data scientists. A data scientist is that guru who takes the data, explores it from all possible angles, and makes inferences that ultimately help him or her to make very informed discoveries.

A data scientist utilizing data:

- Identifies and captures fresh data sources, analyses them, and then builds predictive models. The data scientist also runs live simulations of various market events. All this makes it possible to visualize the reality of possible situations even before any measures have been implemented. Data science, therefore, helps the organization to foresee trouble well in advance and prepare accordingly, and foretell future opportunities as different factors play out in the business arena and the environment in general.
- Utilizes software like Hadoop; NoSQL; and even Storm, to optimize data sets of a non-traditional nature like geo-location and things like sentiment data. After that the data scientist integrates the data sets with that which is more traditional, like trade data.
- Takes the precautionary move of ensuring there is ample raw data in storage for future reference and analysis. In that regard, the data scientist finds the relevant data in its raw form and selects the safest and most cost effective way of storing it.

The expertise of data scientists in utilizing big data is being made even more convenient by the emergence of other technology based storage facilities. There is, for example, the cloud based data storage; analytical tools that are not only sophisticated in the things they can accomplish, but also cost effective. Some are tools that you can access online free of charge; presented as open source

tools. In short, there is a whole range of financial tools that are at the disposal of data scientists, and they are being put to use to transform the way of doing business.

A data scientist understands people's sentiments

Is it strange that a data scientist should be able to analyze people's sentiments simply from data? Well, there are fitting tools for that. In doing sentiment analysis, or what you can aptly call opinion mining, a data scientist makes use of natural language processing; does text analysis; and makes use of computational linguistics to consolidate the relevant material required for the process. There are already firms using sentiment analysis to improve business. Some good examples include MarketPsy Capital; MarketPsych Data; and Think Big Analytics.

How firms work through sentiment analysis:

- By building algorithms surrounding sentiment data from the market, for example, by use of twitter feeds. These ones give ample data when incidences with big impact occur, like a terrorist attack or even a big storm.
- By tracking trends, monitoring as new products get introduced into the market; responding to issues affecting your brand; and in general, improving the brand perception.
- Analyzing voice recordings right from call centers, when those recordings are unstructured; then recommending ways of reducing customer churn, or in other words, recommending ways of raising customer retention.

Considering that most of today's business is on customer focus, the need for the service of analyzing data in order to have a vivid picture what customers feel about a brand cannot be overstated. In fact, data companies have emerged to fulfill this need, and their role is that of intermediaries where they gather data, identify sentiment indicators, and then sell that crucial information to retail businesses.

Data Scientists in Risk Credit Management

The manner in which data scientists utilize the amount, the frequency and variety of data available online, has enabled firms to offer online credit with minimal risk. In some places, potential investors simply fail to access credit because there is no way of giving them a credit rating. Yet a lender or financier

needs to know the extent of risk involved, any time lending is about to take place. Luckily, with big data and expertise from data scientists, internet finance companies have emerged, and they have found ways of approving loans and managing risk. Alibaba Aliloan is a good example of online lending that is enabled by use of big data.

Aliloan is not a conventional bank but an automated online system, which offers flexible small size loans to online entrepreneurs. The recipients are often creative and innovative persons who find it difficult to get credit from traditional lenders like banks simply because they have no collateral.

Big Data Reducing Risk of Online Lending

Using Aliloan as an example of online lending:

i Alibaba monitors its e-commerce platforms as well as the ones it uses for payments, with a view to understanding customer behavior and financial strength. After analyzing the customer's transaction records and customer ratings, and also analyzing related shipping records as well as other related information, Alibaba is able to determine the loan ceiling to put for the customer; considering the level of risk learnt after the comprehensive data analysis.

ii Alibaba also gets the online findings confirmed by 3rd party verifiers, even as it seeks other external data sets to cross check the online findings against. Such helpful external data sets include customs and other tax records, electricity records and such other utility bills.

iii After granting the loan, Alibaba keeps tabs on the customer's activities, monitoring how the customer is utilizing the funds provided for investment. The lender generally monitors the customer's business strategic development.

Other companies that are offering loan facilities by relying heavily on data scientists' expertise on big data are Kreditech and also Lenddo, both of which offer small loans on an automated basis. These ones have come up with credit scoring techniques that are very innovative yet very helpful in determining a customer's creditworthiness. There are also cases where much of the data used to assess a customer's position is from online social networks.

Real Time Analytics Improving the Finance Industry

Any decision maker in the finance industry will tell you it's not enough to have data somewhere within reach – it matters when you analyze it. As many people dealing in critical thinking will confirm, it is not possible to make an informed decision before you analyze the data before you. So, the longer you delay the

process of data analysis, the more you risk business opportunities passing you by; and the higher the chance of other things going wrong in your business. However, with the skills that data scientists have relating to big data, time lags are no longer a handicap in the finance sector.

How real time analytics help:

1) Fighting fraud

It is possible today to detect attempts at fraud through data analytics relating to people's accounts. Institutions like banks, credit card companies and others, have gotten into the trend of keeping on top of things as far as fundamental account details are concerned, courtesy of big data. They want to ensure they know if your employment details are up to date and your physical location too; analyze and see the trend of your account balances; learn about your spending patterns; analyze your credit history; and have such other important information at their fingertips. Since data is analyzed on a real time basis, data scientists ensure that there is a red flag triggered whenever there is suspicious activity taking place; or even when an attempt is detected. What happens then is that the account gets suspended so that the suspected fraudster cannot continue operating, and also the owner receives an alert to that effect instantly.

2) Improving credit ratings

Can anyone give a credit rating without ample data available? Needless to say, the rating is only credible if it factors in current data and not just historical data. That's why big data is so important in this era when credit ratings play a crucial role in determining the level of risk you carry, and the amount of credit you can be allowed to enjoy by lending institutions. The fact that data analytics takes place on a real time basis means that customers' credit ratings are up to date and they provide a reasonable picture of the customer's financial capacity. In any case, most of the categories of data necessary are already covered online, including assets in the customer's name; various business operations the customer is engaged in; as well as relevant transaction history.

3) Providing reasonably accurate pricing

This pricing factor cuts across products and services. In financing, it may means a customer can get a better rate of interest levied on money borrowed is the current rating is better than before. For insurance, a policy holder can enjoy benefits derived from data analysis, issuing timely warnings over accidents ahead; traffic jams that may affect the driver; weather conditions; and such other information that can help in reducing the rate of accidents. With a policy holder having a clean driving record – or at least an improved one – it is possible to win a discount on the price of insurance policy. It also means that

insurance companies will have less to pay out as compensation.

On the overall, cost of business goes down for everyone whenever data analytics are used, mostly today that is mostly because of the benefits accruing from real time analytics. In fact, major financial institutions like banks have made PriceStats the order of the day. This online firm that began as a collector of daily inflation rates for a few countries in South America now monitors prices for around 22 countries (as at 2015), with a view to providing daily inflation rates for those economies. This means you can easily follow the fluctuating or steady trend of inflation rates and tailor your business actions accordingly. PriceStats has a lot of useful data based information for the US too, the largest world economy.

Big Data Also Great for the Customer

For those who may not be aware, many institutions, including banks and a good number of other financial institutions do pay to acquire data from a myriad of retailers as well as service providers. This underlines the importance of data, particularly when you have the capacity to analyze it and use it where it matters most. In fact, all data is important depending on the reasons someone wants it for. You don't want to store data just for the sake of it. Data held within your system with nobody accessing it is an unnecessary distraction. Even when data comes your way without you requisitioning for it, unless you have a good processing plan in place, you can easily get overwhelmed.

That is why it is important to have a data strategy that is inter-departmental, so that you can identify the category of data to shove to another department instead of discarding it, and which portions of data to get rid of straightaway. And the reason this chapter puts emphasis on the contribution of an analyst in data handling and utilization is that not everyone can make good use of data from scratch. But a data scientist has the necessary skills to handle big data from A to Z.

That is why big data is helpful particularly when there is customer segmentation.

In the case of institutions that pay to receive data, their aim is to use the data to create a 360° visual image of their customer. As such, when they speak of KYC (Know Your Customer), they are speaking from a point of credible information; and that reduces the risk of doing business with those individuals. Can you see predictive analytics coming into play right there? This aspect of using big data to have an overall understanding of the customer has been emphasized by Sushil Pramanick, a leading figure with IBM analytics. Pramanick also happens to be the founder of The Big Data Institute (TBDI).

Improving business through customer segmentation

Once you can put together customers who have the same needs and probably the same financial capacity; customers who have similar consumer tastes and are in the same income bracket; customers who are in the same age bracket and are from a similar cultural background; and other who match in various ways; it becomes relatively easy to meet their needs.

a) You can conveniently design customized products and services with them as a target group

b) You can adjust the manner of relating with them with a view to retaining them as customers; avoiding customer churn

c) You can tailor your advertising and marketing to appeal to target groups

d) You can re-engineer products or develop new ones with specific groups in mind

Chapter 13: How Data Science Benefits Retail

As already mentioned elsewhere in the book, data in circulation is increasing in volume each year at a rate that can only be described as exponential; and its variety and velocity is also shooting. Retailers who are smart in business know there is value in data. Yet they may not be certain how to trap and analyze data at every possible interaction for use; where use in this case implies more business profitability.

This is one reason the demand for data scientists has risen over recent years. Experts like McKinsey have weighed in on the benefits of employing data analytics in the retail business. In a report they released in 2011, McKinsey projected that the retailers going for big data analytics are likely to have their operating margins rise by a good 60%. That was a way of reassuring retailers that data scientists have the capacity to turn big data upside down and inside out – whether it is structured or not; internal or external – organizing it in a way that makes business sense, and ultimately helping the retailer create gold out of the mound of apparent clutter.

Retailer's way of reducing cost and increasing revenue

- Receiving recommendations that are customized to the needs of the individual retailer
- Taking advantage of sentiment analysis on data from social media or call centers; or even the one displayed through product reviews. All this data is important in giving a clear picture through customer feedback and gives depth in market insights.
- Improving customer experience through predictive analytics; that improvement being both online and offline.
- Improving how retailers lay out their promotional displays and such other merchandizing resources, by utilizing heat sensors and even image analysis to get a better understanding of customer behavior patterns.
- Utilizing video data analysis to identify customers' shopping trends. This analysis is also helpful in identifying cross selling opportunities.
- Making consistent daily profits because of the ability to respond to the information derived from internal as well as external data. Would there be any benefit, for instance, in taking advantage of reduced cost of raw materials if the weather did not allow for deliveries of the final product? So, data analysis puts the retailers in good stead because they can plan for business with prior knowledge of the weather; economic forecast; traffic reports; and even knowledge about whether it is low or High Holiday season, among other details.

- Growing revenues at a faster rate simply because it is possible to do a detailed market analysis
- Making use of product sensors to communicate on a real time basis, important information concerning post purchase use.
- As far as marketing goes, retailers are able to cut cost by offering personalized offers via mobile devices. They are also able to deliver location based offers, which in itself is much cheaper than generalized marketing.
- Retailers can communicate real time pricing; meaning it is possible to issue prices based on metrics of second by second. In this manner, retailers can also access data on their competitors; data regarding consumer behavior; and other factors close to business.
- Ability to identify appropriate marketing channels going by the indicators from analytics on segment consumers. This helps the retailers to optimize their Return on Investment (ROI).
- Making use of web analytics as well as online behavioral analysis to tailor marketing offers accordingly.
- Retailers also benefit in the area of supply chain logistics. In this regard, data science helps when data scientists use it to track inventory and do general inventory management on a real time basis.
- Retailers are also in a position to optimize on their routes of product delivery through the telematics of GPS enabled big data. Of course with well analyzed data, retailers are bound to select the cheapest, safest and fastest route.
- The area of supply chain logistics benefits also from both structured as well as unstructured data, where retailers are able to foretell likely customer demand well in advance.
- Retailers also get the opportunity to negotiate with suppliers once they read the scenario from analysis of available data is done.

Chapter 14: Data Science Improving Travel

The world of travel has always had an appetite for lots of data. And it has always encountered lots of data too even when stakeholders are not deliberately searching. The biggest challenge has mostly been how to store the data as well as how to analyze it and optimize its use. The data that can be of benefit to the travel sector is massive considering that all facets of life are involved, including people's cultures; air fares; security in different geographical areas; hotel classes and pricing; the weather; and a lot more. For that reason, this sector cannot afford to ignore data science, being the discipline that will help in the handling of the large data sets involved in travel.

How The Travel Sector Uses Data Science

There are lots of benefits that come with the use of big data, which is possible when you have a data scientist working with you. Some of the benefits that come with the use of big data in the travel sector include:

- Ability to track delivery of goods. This is possible whether it is a freight shipment involved; a traveler on road; or a voyage. Major online firms like Amazon and Etsy benefit a lot from such tracking, especially because it is not just the seller who is able to track the shipment but the customer too. This is likely to give customers more confidence in the company and their mode of delivery, thus making them repeat customers – and that is, obviously, good for business.
- Analysis done at each data points. This is one way of increasing business because then it means that different stakeholders will be able to share information from well analyzed data. This ensures that nobody receives data that is irrelevant or redundant.
- Improving access to travel booking records; the use of mobile phones to make inquiries, bookings as well as payments and other simplification of transactions
- Easy access to customer profiles; itineraries; positive as well as negative feedback; and a lot of other data from internal sources like sensor data.
- Easy access of external data such as reviews from social media; the weather; traffic reports; and so on.

In short, the travel sector is a great beneficiary of big data, and it merges the data from both its internal and external sources to find solutions to existing problems. It also gets the same data analyzed in a way that gives helps the stakeholders anticipate with relative precision the position of future events. The sector is also able to cut drastically on cost of operation because they can make better choices when they have massive data analyzed in a timely manner as happens when data science is employed.

One reason there is a lot of interest in big data in the travel sector is that the sector is lucrative and it is projected to become even more rewarding as far as revenues are concerned. Global travel is projected to grow sharply so that by 2022, its value will have reached 10% of the world's Gross Domestic Product (GDP). Of course, the main players involved realize the value of big data and want to optimize its use so that they get the best business intelligence to work with. And that makes them see big bucks ahead.

Big Data Enabling Personalization Of Travel Offers

Some years back, travel considerations were given as per general classification of customers. Those in the high income bracket were sent recommendations for certain facilities; those who regularly travel with children were targeted for different offers; and so on. While that may have increased revenues a little bit from the normal, its impact is still peanuts compared to today's use of big data.

The reason big data is very significant in the travel industry today, is that it enables the Travel Company or agency to tailor offers as per individual customers, by relying on a 360° view of the customer. So the potential of offering ultra personalized packages or facilities to individuals is not far-fetched where big data is at play.

Some data sets that help produce a 360° view

- Data directed at reading behavior.

A good example of this is the regularity with which a person goes online; and probably the websites the person is a regular visitor

- The posts that one posts on social media

The analyst can establish if the person writes issues about travel on social media; whether friends speak about travel with this person; or even if there are any travel reviews contributed by friends.

- Data from location tracking
- Data on itineraries on various occasions
- Historical data about a person's shopping pattern
- Data reflecting how the person uses mobile devices
- Data regarding image processing

This list is by no means exhaustive. In actual fact, any data or information relating to an individual's preferences on travel would make the list. As you will realize different data sets suit different people while others are common to all. For example, the data sets targeting a person's behavior pattern will mostly be useful for a potential travel customer; or one who is new to the sector. However, when you speak of historical patterns, then you are targeting old customers.

Big Data Enhancing Safety

If you consider the information transmitted between coordinators of travel from the pilot to the control tower and vice-versa; driver and the fleet headquarters and vice-versa; from traffic headquarters to all travelers; from electronic media to travelers at large; and so on, you will easily acknowledge that big data is a real life saver in the sector of travel. In fact, in present travel, the vehicles and planes concerned are fitted with different types of sensors that detect, capture and relay certain information on real time basis. The information relayed varies greatly and may include airmanship; the behavior of the driver; mechanical state of the vehicle or plane; the weather; and so on.

This is where a data scientist comes in to design complex algorithms that enable the travel institution to foretell when a problem is imminent. Better still, the work of the data scientist helps to prevent the problem before it crops up.

Here are some of the common problems that big data helps address:

- If there is a part of the vehicle or plane that is not working well and is detected, it is replaced before the damage on that part becomes too serious for the part to be salvaged. Replacement or repair is also done to prevent an accident.
- In case the drive is missing one step or more when traveling, you need to pull him or her out of active duty and take him or her through some further training. This is necessary in order to maintain high standards of performance.
- For flights, big data helps to identify problems that can be tackled when the plane is mid air and those that are possible to tackle in such a position. For those that are found to be difficult to tackle when the plane is in the air, the company puts a maintenance team on standby so that they can dash to check the plane immediately on arrival.

From the facts provided, it is safe to say that big data plays a crucial role in averting accidents during travel, whether by air; by sea; or even by road. In fact, it is easy to appreciate that when you think about the transmission of weather related data and information on real time basis.

Up-Selling And Cross-Selling

Sometimes big data is used in up-selling and other times in cross selling. Just to re-cap, vendors are up-selling when they are trying to woo you into buying something that is pricier than what you were seeking in the first place – great marketing attempt. As for cross selling, vendors just try to get you to buy something different from what you were searching for initially.

In case you want to travel by air, you are likely to receive many other offers together with the one you inquired about in the first instance.

Here are examples of up-selling and cross selling:

- You may find in your inbox an offer that is pretty personalized, and that is a form of cross-selling
- You may find yourself being booked for the Economy Plus, which means you are being up-sold. With the advantages of additional leg room and an opportunity to recline further on your seat, it is tempting to take up the offer even if it means forking an extra amount of dollars.
- You are likely to find discounted offers of hotels partnering with the airline you are using for your travel.
- You could also receive an offer for dining, courtesy of the steward, and this is on a complimentary basis – so you get a coupon
- You may find yourself looking at an ad, during an in-flight entertainment session, trying to attract you into using a certain city tour on arrival or during your local travel.

Companies particularly in the tourism sector are partnering and using big data to cross sell each other. The usual suspects include airlines; hotels; tour van companies; and companies offering taxi services.

Chapter 15: Big Data and Law Enforcement

Big data is making critical inroads in criminal investigations. Today some law enforcement agencies aren't using data tools to their fullest potential. They will work independently of one another and maintain individual standalone systems. This is impractical and not a cost effective way of running these agencies. Law enforcement is expected to do more with less. Manpower has been cut back but crime increases in some areas.

It is to the benefit of these agencies to pool their shrinking resources into a data networking system where law enforcement agencies network data across the board to help solve crimes and protect the public more efficiently.

Big Data is the answer to finding effective solutions to stopping crime in a time when agencies are being asked to perform more tasks with less help. This is where big data is the go to for these agencies. They are being forced to streamline operations and still produce the desired results. Data has to be interpreted holistically to be able to streamline their operations Data platforms make these objectives a reality.

The data platform can be placed in the Human Resources system or even in the plate identification system. It doesn't matter where the platform is integrated it just has to be inserted into one of the data systems. So there will be a foundation from which criminal activity can be dissected, analyzed and translated into effective crime fighting techniques. Challenges facing Law Enforcement:

- Law Enforcement is being straddled continually with shrinking financial resources
- Limited money equates to less manpower
- Even with less manpower Law Enforcement is expected to maintain the same level of services
- Law enforcement must find innovative ways to fight crime through data systems
- Big Data can meet all of the above requirements if use to its fullest potential

Law enforcement is the foundation upon which our society is kept safe, where people can feel secure and safe walking down the street in broad daylight. Where a young woman can walk safely in remote areas at night and not feel threatened.

Lately, the public view of law enforcement has turned negative. Police brutality is perceived to be on the rise. Crime seems to be escalating around the nation. Where is law enforcement when you need them most? We must realize law enforcement is in a catch 22 situation here in our 21st century. How can they be expected to do more with fewer resources? Are demands being put on them improbable for them to realize?

They have fewer personnel to cover the same amount of ground as before when law enforcement agencies had full rosters to meet all the demands of the criminal activity.

Data Analytics is the solution to law enforcement's current dilemma of shrinking resources

Once the previously mentioned data platform is put into place, then data analytics can be applied at a very detailed level. Police departments can now solve crimes faster than ever before. They can stop criminal activity with a greater success rate. They can even now use data analytics to set in potential criminal behavioral tendencies and stop these lawbreakers from committing the crime before they ever get the chance. Here are some technologies Microsoft uses to help in this quest:

- Microsoft Azure Data Lake
- Microsoft Cortana Analytics
- Windows speech recognition

There is one police department working with Microsoft using their data analytic tools. Some of these tools are Microsoft Azure, which allows crime analysts to store criminal data in any size, shape or speed. The beauty of this technology is that it can be used for any analytics or processing across platforms and languages. No matter how much data they have, it can be stored all in one data central system. Do you catch the significance of this?

Law enforcement agencies can network with each other worldwide. This would be helpful in tracking terrorist activities worldwide. Using predictive analytics data can be extracted and used to project where terrorists will strike next. They can decipher the most likely form of attack the group or individuals will use. What will be the extent of the operation be based on historical data stored from previous attacks? They can stop the attacks before they materialize. Microsoft Cortana Analytics: This analytic application allows crime analysts to predict crime scenarios using machine-learning algorithms at lightening quick rates of speed. The analysts can also use this tool to support and enhance decision-

making processes by recommending the best course of action to take in the event of a real-time crime or predictive crime taking place. It also helps them to automate and simplify critical decisions, which include many criminal variables that constantly change in real time, giving them a greater capability to make the right decisions quicker to stop crimes before they happen and while they are happening.

Windows voice recognition: assists police officers in dictating critical commands pertaining to a driver's outstanding warrant history, DUI violations and ticket violation patterns. Knowing this information quickly in real time allows the police officer to position himself in a place to stop potential crimes or traffic violations from occurring. Essentially the police officer could reduce the threat or actually diffuse a life-threatening event before it occurs.

Microsoft A Central Player In Police Department Crime Prevention

Microsoft is using their data applications in assisting police departments to prevent crime before it happens. There is one police department they are working with, to measure public safety in real time using varied sources. This particular police department uses a plethora of data streams to analyze live 911 calls, live feeds from cameras and police reports to fight terrorism and other types of violent crimes. Crimes at all levels can be measured, which allows for quicker reaction time and the ability to develop strategies to fight these crimes in real time to proactively stop them before they happen.

Advanced Analytics Used For Deciphering Across The Board Criminal Activities

Police departments are also using advanced analytics to automatically interpret crime patterns such as burglaries, homicides, and domestic violence incidents. This data tool helps police departments quickly sort through large amounts of data quickly. Police can now identify a burglar's crime patterns. When will the burglary be committed and what will be the point of entry into the house or building? What area will the burglar next try to steal from? What objects will he steal? Armed with this information police officers can be positioned to stop the burglaries from happening.

Another facet is to preventing burglaries from happening in certain geographical areas and in certain timeframes. Let's say a certain area of the city is vulnerable to burglaries in certain times of the year. For example, a row of apartment complexes will be hit in the summer months between 8: 00a.m-12:00p.m.

Knowing this, police departments can dispatch police officers to the designated areas during the crucial time period. Therefore, they can stop a series of robberies before they start.

Police Departments Can Respond To Major Terrorist Crimes Successfully After They Occur

Remember the Boston Marathon bombing? Police used a few data sources to respond to the crime promptly and make arrests quickly. The Boston police used Social Media outlets to gather huge amounts of citizen data. It was a huge task to filter through the data, but it sure paid off. The police were able to identify and track down the 2 suspects.

Cell phone videos and security camera footage were painstakingly pored through but the police's task paid off. The police were able to identify the two suspects in the video. They tracked them down from the images caught on these two data sources.

Chapter 16: What about Prescriptive Analytics?

What is Prescriptive Analytics?

The latest technical data buzz word-"Prescriptive Analytics" is the third branch of the big three data branches. It is predicted this analytic model will be the last of the three. Is it the next groundbreaking data model in technology? This analytic system is a summation of the previous two-Descriptive Analytics and Predictive Analytics. Definition of the three Analytic models:

- **Descriptive Analytics**-uses Data Aggregation and Data mining techniques, to give enlightenment to the past. Answering this question "What has happened?"
- **Predictive Analytics**-uses statistical models and forecasting techniques to giving insight into the future. Answers this question "What will happen?
- **Prescriptive Analytics**- uses optimization and simulation algorithms to options for the future "What should we do?"

The prescriptive analytics model is the newest branch in the data world. Some are already arguing it is the only way to go in the analytical world in the coming future. It doesn't seem like Prescriptive Analytics hasn't caught fire in the business world as of now. It will become more widespread in different industries. Companies will realize the benefits of having a model that will suggest answers to future questions which is the most advanced component of this relatively new technology.

It takes predictive analytics one-step further. It goes beyond predicting what will happen in the future. It gives options for the future then it suggests the best option for that course of action. It has some "artificial intelligence" involved in its processes. It analyzes the optimization and simulation algorithms then "thinks" of all the options based on those systems of algorithms. It will even suggest the best option to take out of all the options analyzed. Prescriptive Analytics are comparatively speaking difficult to administer than the other two analytical models. It also uses machine thinking and computational modeling.

These various techniques are used against several many data sets historical data, transactional data, real-time data feeds and big data. It also incorporates algorithms that are released on data with minimal parameters to tell them what to do. The algorithms are programmed to conform according to the changes in established parameters instead of being controlled externally by humans. The algorithms are allowed to optimize automatically. Over time, their ability to predict future event improves.

What Are The Benefits Of Prescriptive Analytics?

Many companies aren't currently using Prescriptive Analytics in their daily operations. Some of the huge corporations are using them in their daily operations. These big companies are effectively using prescriptive analytics to optimize production and successfully schedule and guide inventory, in their respective supply chains. To be sure they are delivering the right products at the right times and to the right places. This also optimizes the customer experience. You know it is advised to use prescriptive analytics to tell customers what to do. This can be a very effective management tool for end-users.

The Future Of Prescriptive Analytics

What does the future hold for prescriptive analytics in the international business world? In 2014, there were approximately 3% of the business world was using prescriptive analytics. Scott Zoldi, Chief Analytics Officer, at FICO, made the following predictions concerning prescriptive analytics for 2016:

- Cornerstone of prescriptive analytics "streaming analytics" will become dominant in 2016
- Streaming Analytics is applying transaction-level logic to real time events (They have to occur in a prescribed window-last 5 seconds, last 10,000 observations, etc.)
- Prescriptive analytics will be a must have cyber security application (This process will analyze suspicious behavior in real time)
- Prescriptive analytics will go mainstream in lifestyle activities in 2016 (From home appliances to automated shopping it is poised for explosive growth)

Analytical experts are predicting the next major breakthrough for prescriptive analytics will be it will go mainstream spreading across all industries. In 2015, it was predicted that prescriptive analytics would increase in consultancy partnering with academics. As of 2016, there still was no off-the-shelf general purchase for prescriptive analytics, which seems to be a long way off.

It seems to be a hit and miss preposition for prescriptive analytics within departments of the same company. One department will be using prescriptive analytics while the other department won't be.

Google Using Prescriptive Analytics For Their "Self-Driving Car"

The search engine giant is using prescriptive analytics in their five-year-old self-driving cars. The cars drive automatically without a human guide. They make lane changes, turn right or left, slow down for pedestrians and stop at lights. Generally, they drive like any other car but without the human driver.

The car uses prescriptive analytics to make every decision it will need to make during the driving experience. It makes decisions based on future events. For instance, the car approaches a traffic light intersection it has to decide whether to turn right or left. It will take immediate action on future possibilities concerning which way to turn. The factors it considers before it makes that decision are many. It considers what is coming towards it in terms of traffic, pedestrians, etc. before it determines which way to turn. It even considers the effect of that decision before it makes the decision to turn right or left at the intersection. Google being such a high profile multi-billion-dollar company will have a major impact on prescriptive analytics moving forward across the business world.

The Oil and Gas Industry is using Prescriptive Analytics

The oil and gas industry is currently using prescriptive analysis to interpret structured and unstructured data. It also uses this analytical model to maximize fracking (is the process of injecting liquids at high pressure to subterranean rock, boreholes, etc. to force open existing fissures to extract out oil and gas) Here are applications the oil and gas industry is using prescriptive analytics in:

- Maximize scheduling, production and tune the supply chain process, so the right products are shipped to the right person, at the right time and to the right location.
- Maximize customer experience
- Locate functioning and non-functioning oil wells
- Optimize the equipment materials needed to pump oil out of the ground
- The above tasks are to deliver the right products at the right time to the right customers at the perfect time

Most of the world's data (about 80%) are unstructured as videos, texts, images, and sounds. The oil and gas industry used to look at images and numbers but in separate silos. The advent of prescriptive analytics changed this. Now, the industry has the ability to analyze hybrid data- the combination of structured and unstructured gives the industry a much clearer picture and more complete scenario of future opportunities and problems. This, in turn, gives them the best actions to determine more favorable outcomes. For example, to improve the hydraulic fracking process the following datasets must be analyzed simultaneously:

- Images from well logs, mud logs, and seismic reports
- Sounds from fracking from fiber optic sensors
- Texts from drillers and frack pumpers' notes
- Numbers from production and artificial lift data

Along these lines, hybrid data is essential in analyzing because of the multi-billion-dollar investment and drilling decisions made by oil companies. They must know where to drill, where to frack and of course how to frack. Any one of these steps is skipped or done incorrectly it could have disastrous results for these energy companies. The cost could be astronomical. It would be unimaginable and a waste of time, energy and manpower if they fracked or

drilled in the wrong way or place.

There are more to the components involved in completing the prescriptive analytics. Scientific and computational disciplines must be combined to interpret different types of data. For instance, to algorithmically interpret images like (log wells) machine thinking must be connected with pattern recognition, computer vision, and image processing. Mixing these various disciplines enables energy companies to visualize a more holistic system of recommendations: where to drill and frack, while minimizing issues that might come along the way. By forming detailed traces-using data from production, subsurface, completion, and other sources–energy companies are able to predict functioning and non-functioning wells in any field.

This process is anchored by prescriptive analytics technology's ability to digitize and understand well logs to produce dispositional maps of the subsurface, knowing where to drill oil companies save untold millions in resources. This process helps them to pinpoint where non-functional wells are located, obviously saving them optimum drilling time. It needs to be said that the environmental impact is minimized on the particular landscape. Prescriptive analytics should be used in other areas of oil and gas production. In both traditional and non-traditional wells, simply by using data from pumps, production, completion and subsurface characteristics, oil companies are able to predict failure of submersible pumps and reduce production loss. (Again saving untold millions of dollars.)

The Apache Corporation is using prescriptive analytics to predict potential failures in pumps that extract oil from the subsurface, thereby minimizing the associated costly production loss due to the failure of these pumps. Another potential application of prescriptive analytics is too hypothetically predicting corrosion development on existing cracked oil pipelines. The prescriptive model could describe preemptive and preventive actions, by analyzing video data from cameras. It could also analyze additional data from robotic devices known as "smart pigs" that are positioned within these damaged pipelines.

Using prescriptive analytics in their daily operations will help oil companies make wiser decisions. This will result in fewer manpower accidents, reduce production loss, maximize resources and increase financial profits. Another valuable benefit will be to reduce the environmental impact in the particular landscapes and surrounding communities. Oil Companies will put themselves in a favorable light with environmentalists. By showing concern and taking steps to minimize or reduce environmental hazards around their operating facilities.

Hopefully, the oil companies will see fit to pass on the cost-savings to the consumers in the form of reduced gas prices, from their reduced costs using the prescriptive analytics model. Further investigation into the oil and gas industry reveals they use prescriptive analytics in deeper ways. They use it to locate the

oil fields with the richest concentrations of oil and gas. It helps track down oil pipes with leaks in it. When caught early these leaks can be repaired, therefore, a major environmental hazard from an oil spill. Prescriptive analysis also helps refine the fracking process to avoid endangering the environment and to improve the output of oil and gas.

Conversely, the oil and gas industry will have to increase expenses to hire the proper experts to run the prescriptive analytics models. The training involved in getting these technicians up to speed with the company operations will further inflate company costs. But streamlining all the processes outlined earlier in this chapter will outweigh the increased costs to increase manpower. To further justify the hiring of analysts, statisticians and of course data scientists is a necessary expense. The prescriptive analytics model is more complicated to implement than the descriptive and predictive analytical models. But the potential to change industries is a given as the prescriptive analytics system has already drastically changed the oil and gas industries.

Prescriptive Analytics Is Making Significant Inroads Into The Travel Industry

The Travel Industry has come on board and is using Prescriptive Analytics in their daily activities as well as in their overall operational functions. Prescriptive analytics calls for many large data sets. Because of this factor the travel industry sees great potential in this latest round of data analytics. Online traveling websites like airline ticketing sites, car rental sites or hotel websites have seen the tremendous benefits of predictive analytics in their local business functions.

They are implementing prescriptive analytics to filter through many multiple and complex phases of travel factors, purchase and customer factors such as demographics and sociographics, demand levels and other data sources to maximize pricing and sales. Predictive analytics may turn the travel industry upside down as it did the gas and oil industry. It could be said that predictive analytics is like reinventing the wheel in data analytics. It is more complicated in its daily inter-facing than its two earlier predecessors have been.

Other applications the travel industry is finding applicable for prescriptive analytics are as follows: Segmenting through potential customers predicated on multiple data sets on how to spend marketing dollars. (By doing this the travel industry can optimize every dollar to attract a wider customer base. They will be able to predict where customers preferred traveling locations are based on past traveling preferences. This will result in marketing and advertising strategies that will appeal to the particular target age groups.)

An example of this: The Intercontinental Hotel Group currently utilizes 650 variables on how to spend their marketing dollars to increase their customer base.

Other Industries Using Prescriptive Analytics

The Healthcare Industry is waking up to the cutting edge technology that prescriptive analytics can provide for their intricate overall operational activities. Using so many variables offers doctor, nurses and physician assistants can choose the optimal options for the best treatment programs for their patients. Also, the prescriptive analytics model can suggest which treatment would be the best fit for the illness or physical condition of the patient. This will streamline the diagnostic process for the medical profession. Imagine the hours saved because the many options for any medical condition will be populated through the many data sets used in the prescriptive analytics model.

One very successful example of a medical company saving a large amount of money is noted as follows: The Aurora Health Care Centre was able to improve healthcare and reduce re-admission rates by a healthy 10% resulting in a significant savings of $ 6 million dollars this is the type of potential prescriptive analytics has in impacting industries worldwide.

The Pharmaceutical industry will not be left out of the loop either. Prescriptive analytics can help the industry across the board by helping it to reduce drug development and minimize the time it takes to get the medicines to the market. This would definitely reduce drug research expenditures and greatly reduce manpower hours and resources. Also, drug simulations will quicken the time it takes to improve the drugs and patients will be easier to find to trial the medications on.

We can see that prescriptive analytics is the wave of the future in the data analytics world. The model does require massive amounts of data sets, though. Currently, only the largest size corporations have the amount of data sets to run the prescriptive analytics system feasibly. If the data sets are not available in great quantities, the system absolutely will not work. This is a drawback for smaller companies who don't have the large quantities of data sets to effectively run the prescriptive analytics in their operations. This could be one reason why this model has not taken off to this point. But the model is starting to make inroads into the larger corporate settings. As they realize the great potential this analytical model has. It could take 5-10 years for prescriptive analytics to become a household system in businesses at the International level.

The measurable impact that prescriptive analytics has had in the gas and oil and travel industries just cannot be ignored. Truly, Prescriptive Analytics could be the dominant data system of the near future. It has made noteworthy inroads thus far.

Data Analysis And Big Data Glossary

The following words and terms are related to data analysis, big data and data science. While not all of them have been used in this book, they are words and terms that you will come across in your studies. The whole list of terms is quite exhaustive and would take forever to list so I have chosen to give you definitions for the more common terms

A

ACID Test - this is a test that is applied to data to test for consistency, durability, isolation and atomicity – in other words, to make sure it all works as it should

Aggregation – the collection of data from a number of different databases for analysis or data processing

Ad hoc Reporting – reports that are generated for one-off use

Ad Targeting – an attempt made by business to use a specific message to reach a specific audience. This usually takes the form of a relevant ad on the internet or by direct contact.

Algorithm – mathematical formula that s inserted into software to analyze a certain set of data

Analytics – the art of using those algorithms and certain statistics to determine the meaning of the data

Analytics Platform – either software or a combination of software and hardware that provides you with the computer power and tools that you need to perform different queries

Anomaly Detection – The process by which unexpected or rare events are identified in dataset, these events will not conform to any other events in the same dataset

Anonymization - the process of severing the links between people and their records in a database. This is done to prevent the source of the records being discovered

API – acronym for Application Program Interface. These are programming standards and a set of instructions that enable you to access or build software applications that are web-based

Application - a piece of software designed to perform a certain job or a whole suite of jobs

Artificial Intelligence – the apparent ability of a computer to apply previously gained experience to a situation in the way that a human being would

AIDC – Automatic Identification and Capture – any method by which data is identified and collected on items and then stored in a computer system. An example of this would be a scanner that collects data through an RFID chip about an item that is being shipped

B

Behavioral Analytics – the process of using data collected about people's behavior to understand that intent and then predict their future actions

Big Data – there are many definitions for the term but all of them are pretty similar. The first definition came from Doug Laney in 2001, then working for META Group as an analyst. The definition came in the form of a report called "3-D Data Management: Controlling Data Volume, Velocity and Variety". Volume is referring to the size of the datasets and a McKinsey report, called "Big Data: The Next Frontier for Innovation, Competition and Productivity" goes further on this subject saying "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze".

Velocity is referring to the speed the data is acquired at and used. Companies are not only collecting the data at a faster speed, they also want to determine the meaning of it at a faster rate as well possibly in real-time.

Variety is referring to the different data types available for collection and analysis as well as the structured data that would be found in a normal database. There are four categories of information that make up big data:

- Machine generated – including RFID data, data that comes from monitoring devices and geolocation data from a mobile device
- Computer log – including clickstreams from some websites
- Textual social media – from Twitter, Facebook, LinkedIn, etc.
- Multimedia social media – and any other information gleaned from place like YouTube, Flickr and other sites that are similar

Biometrics – the process of using technology to use physical traits to identify a person, i.e. fingerprints, iris scans, etc.

Brand Monitoring - the process of monitoring the reputation of your brand online, normally by the use of software

Brontobyte – a unit representing a vast amount of bytes. Although it is not yet officially recognized as a unit, a brontobyte has been proposed as a unit measure

for data that goes further than the yottabyte sale

Business Intelligence – BI - term that is used to describe the identification of data, and encompasses extraction and analysis as well

C

CDR – Call Detail Record Analysis - **a** CDR contains data collected by telecommunications companies and includes the time and the length of a phone call. The data is used in a variety of different analytical operations

Cassandra – One of the most popular of the columnar databases normally used in big data applications. Cassandra is an open source database that is managed by The Apache Software Foundation

Cell Phone Data - a mobile device is capable of generating a vast amount of data and most of it can be used with analytical applications

Classification Analysis – data analysis used for assigning data to specific groups or classes

Clickstream Analysis - the process of analyzing web activity by collecting information about what a person clicks on a page

Clojure – this is a dynamic program language that is based on LISP and uses JVM (Java Virtual Machine); suited for use in parallel data processing

Cloud - a wide term covering any service or web based application hosted remotely

Clustering Analysis – the process of analyzing data and using it to identify any differences or similarities in data sets; this is so that any that are similar can be clustered together

Columnar Database/Column-Oriented Database – a database that stores data in columns, not rows. Row databases could contain information like name, address, phone number, age, etc., all stored on one row for one person. In column databases, all of the names would be stored in one column, addresses in another, ages and telephone numbers in their own columns and so on. The biggest advantage to this type of database is that had disk access is faster

Comparative Analysis – analysis that compares at least two sets of data or processes to see if there are any patterns in large sets of data

Competitive Monitoring - using software to automate the process of monitoring the web activity of competitors

CEP – Complex Event Monitoring - this is the process of monitoring the events in the systems of any organization, analyzing them and then acting when necessary

Complex Structured Data - structured data made up of at least two inter-related parts which is therefore difficult for structured tools and query language to process

CLASS – Comprehensive Large Array-Data Stewardship System – digital library of historical data gained from NOAA (US National Oceanic and Atmospheric Association

Computer-Generated Data - data that is generated through a computer instead of a human being, for example, a log-file

Concurrency – the ability to execute several different processes at once

Confabulation – the process of making a decision based on intuition look as if it were based on data instead

CMS – Content Management System – software that allows for the publication and the management of web-based content

Correlation Analysis - A process of determining the statistical relationship between two or more variables, usually to try to identify if there are any predictive factors present

Correlation – refers to multiple classes of dependent statistical relationships. Examples would be the correlation of parents and their children or the demand for a specific product and its price

Cross-Channel Analytics - analysis that can show lifetime value, attribute sales or how an average order

Crowdsourcing – the process of asking the public to help complete a project or find a solution to a problem

CRM – Customer Relationship Management - software used to help manage customer service and sales

D

Dashboard – graphical report of either static or real-time data on a mobile device or a desktop. The data is usually high level in order to give managers access to quick reports of performance

Data – a qualitative or quantative value. Examples of the more common data includes results from market research, sales figures, readings taken from monitoring equipment, projections for market growth, user actions on websites customer lists and demographic information

Data Access – the process of retrieving or viewing data that has been stored

DATA Act – Digital Accountability and Transparency Act 2014 - A

relatively new US law that is intended to make it easier to gain access to federal government expenditure information by requiring the White House Office of Management and Budget and the Treasury to standardize data on federal spending and to publish it

Data Aggregation - the collection of data from a number of sources for the purpose of analysis or reporting

Data Analytics – the use of software to determine meaning of or information from data. The result may be a status indication, a report or an automatic action, based on what information is received

Data Analyst – the person who is responsible for preparing, modeling and cleaning data so that actionable information can be gained from it

Data Architecture and Design - This is the structure of enterprise data. The actual design or structure will vary, as it is dependent on the result that is required. There are three stages to data architecture:

- The conceptual representation of the entities in the business
- The representation, logically, of the relationship between each of the entities
- The construction of the system that supports the functionality

Data Center - physical place that is home to data storage devices and servers that may belong to one or more organizations

Data Cleansing – the review and revision of data to eliminate any duplicate information, to correct spelling errors to add in any missing data and to provide consistency across the board.

Data Collection – a process that captures any data types

Data Custodian – The person who is responsible for the structure of the database and the technical environment, including data storage

Data Democratization – the concept of ensuring that data is available directly to all workers in an organization, instead of them having to wait for the data to be delivered to them by another party, usually the IT department, with the business

Data-Directed Decision Making – The use of data to support the need to make crucial decisions

Data Exhaust - the data that is created by a person as a byproduct of another activity, for example, call logs or web search histories

Data Feed - the means by which data streams are received, for example, Twitter, or an RSS feed

Data Governance - the rule or processes that ensure data integrity and that management best practices are followed and met

Data Integration - The process by which data from various sources is combined and presented in one view

Data Integrity – a measure of trust that an organization places in the completeness, accuracy, validity and timeliness of data

Data Management – The Data Management Association says that data management should include this practice to ensure the full lifecycle of data is managed:

- Data governance
- Data design, analysis and architecture
- Database management
- Data quality management
- Data security management
- Master data management and reference
- Business intelligence management
- Data warehousing
- Content, document and record management
- Metadata management
- Contact data management

DAMA – Data Management Association – non-profit international organization for business and technical professionals that is "dedicated to advancing the concepts and practices of information and data management"

Data Marketplace – an online location where people can purchase and sell data

Data Mart - access layer of a data warehouse that provides users with data

Data Migration – the process by which data is moved between different formats, computer systems or storage places

Data Mining – the process of determining knowledge or patters from large sets of data

Data Model/Modeling - data models are used to define the data structure needed for communication between technical and functional people to show which data is needed for the business. Also used for communication of development plans for data storage and access among specific team members, usually those in application development

Data Point – individual item on a chart or graph

Data Profiling – the collection of information and statistics about data

Data Quality - the measurement of data to determine if it can be used in planning, decision making and operations

Data Replication – the process by which information is shared to ensure consistency between sources that are redundant

Data Repository - the location where persistently stored data is kept

Data Science – this is a relatively recent term that has several definitions. It is accepted as a discipline that incorporates computer programming, data visualization statistics, data mining database engineering and machine learning in order to solve complex problems

Data Scientist - a person who is qualified to practice data science

Data Security – the practice of ensuring that data is safe from unauthorized access or from destruction

Data Set – a collection of data stored in tabular form

Data Source – a source from which data is provided, such as a data stream or database

Data Steward - the person who is responsible for the data that is stored in a data field

Data Structure – a certain method for the storage and organization of data

Data Visualization – Visual abstraction of data that is used for determining the meaning of the data or for communication the information in a more effective manner

Data Virtualization – process by which different sources of data are abstracted through one access layer

Data Warehouse – a place where date is stored for analysis and reporting purpose

Database - digital collection of data and the structure that the data is organized around.

Database Administrator – DBA – a person who is usually certified for being responsible for the support and maintenance of the integrity of content and structure of a database

DaaS – Database as a Service - a database that is hosted in the cloud and is sold on a metered basis. Examples of this include Amazon Relational Database Service and Heroku Postgres

DBMA – Database management System - software used for collecting and providing structured access to data

De-Identification – the process of removing data that links specific information to a specific person

Demographic Data - data that relates to the characteristics of the human population, i.e. in a specific area, age, sex, etc.

Deep Thunder – weather prediction service from IBM that provides specific organizations, such as utility companies, with weather data, allowing them to use the information to optimize the distribution of energy

Distributed Cache – data cache that spreads over a number of systems but works as a single system; usually used as a way of improving performance

Distributed File System - a file system that is mounted on several servers at the same time to enable data and file sharing

Distributed Object - software module that has been designed to work with other distributed objects that are stored on different computers

Distributed Processing - execution of a specific process over several computers that are connected to the same computer network

Document Management - the process of tracking electronic documents and paper images that have been scanned in, and then storing them

Drill - this is an open source system, distributed for carrying our interactive analysis on very large datasets.

E

Elastic search – open source search engine that is built on Apache Lucene

HER – Electronic Health Records - a digital health record that should be accessible and usable across a number of different health care settings

ERP – Enterprise Resource Planning - software system that enables a business to manage resources, business functions and information

Event Analytics – shows the steps taken to lead to a specific action

Exabyte - 1 billion gigabytes or one million terabytes of information

Exploratory Data Analysis - data analysis approach that focuses on identifying general data patterns

External Data - data that lives outside a system

ETL – Extract, Transform and Load – process used in data warehousing to prepare data for analysis or reporting

F

Failover – the process of automatically switching to a different node or computer when one fails

FISMA – Federal Information Security Management Act - US federal law that says all federal agencies have to meet specific information security standard across all systems

Fog Computing - computing architecture that enables users to have better access to data and data services by moving cloud services, like analytics, storage and communication closer to them though a device network that is distributed geographically

G

Gamification – the process of using gaming techniques in applications that are not games. This is used to motivate employees and encourage specific behaviors from customers. Data analytics often is applied to this in order to personalize rewards and encourage specific behaviors to get the best result

Graph Database – a NoSQL database that makes use of graph structures for semantic queries with edges, nodes and properties that store, query and map data relationships

Grid Computing - performance of computing functions making use of resources from systems that are multiple distributed The process involves large files and is mostly used for multiple applications The systems that make up grid network do not need to be designed similarly, neither do they have to be in the same location geographically

H

Hadoop - open-source software library that is administered by Apache Software Foundation. Hadoop is defined as "a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model".

HDF – Hadoop Distributed File System - fault-tolerant distributed file system that is designed to work on low-cost commodity hardware. It is written for the Hadoop framework and written in Java

HANA - a hardware/software in-memory platform that comes from SAP. It is designed for real-time analytic and high volume transactions

HBase – distributed NoSQL database in columnar format

HPC – High Performance Computing – Also known as supercomputers. Normally these are custom-built using state of the art technology, as a way of maximizing computing performance, throughput, storage capacity and data transfer speeds.

Hive - a data and query warehouse engine similar to SQL

I

Impala - open source SQL query engine distributed for Hadoop

In-Database Analytics – the process of integrating data analytics into the data warehouse

Information Management - the collection, management and distribution of all different types of information, include paper, digital, structured and unstructured.

In-Memory Database – a database system that relies solely on memory for storing data

IMDG – In-Memory Data Grid – data storage in memory across a number of servers for faster access, analytic and bigger scalability.

Internet of Thing – IoT - this is the network of physical things that are full of software, electronics, connectivity and sensors to enable better value and service through the exchange of information with the operator, manufacturer and/or another connected device/s. Each of these things is identified through its own unique computing system but is also able to interoperate within the internet infrastructure that already exists.

K

Kafka - open source message system from LinkedIn that is used to monitor events on the web

L

Latency - a delay in the response from or delivery of data to or from one point to another one

Legacy System - an application, computer system of a technology that is now obsolete but is still used because it adequately serves a purpose

Linked Data - this is described by Time Berners Lee, the inventor of the World Wide Web as "cherry-picking common attributes or languages to identify connections or relationships between disparate sources of data".

Load Balancing - the process of the distribution of a workload across a network or cluster as a way of optimizing performance

Location Analytics - this enables enterprise business systems and data warehouses to use mapping and analytics that is map-driven. It enables the use of geospatial information and a way to associate it with datasets.

Location Data - data used to describe a particular geographic location

Log File - files created automatically by applications, networks or computers to record what happens during an operation. An example of this would be the time that a specific file is accessed

Long Data - this term was created by Samuel Arbesman, a network scientist and mathematician, referring to "datasets that have a massive historical sweep"

M

Machine-Generated Data - data that is created automatically from a process, application or any other source that is not human

Machine Learning - a process that uses algorithms to enable data analysis by a computer. The purpose of this is to learn what need to be done when a specific event or pattern occurs

Map/Reduce - General term that is referring to the processing of splitting a problem down into bits, each of which is then distributed among several computers that are on the same cluster or networks or a map, a grid of geographically separated or disparate systems. The results are collected in from each computer and combined into a single report.

Mashup – a process by which different datasets from within one application are combined to enhance the output. An example of this would be a combination of demographic data with a real estate listing

MPP – Massively Parallel Processing - the process of breaking up a program into bits and executing each part separately on its own memory, operating system and processer

MDM – Master Data Management - master data is any data that is non-transactional and is critical to business operation, i.e. supplier or customer data, employee data, product information. MDM is the process of master data management to ensure availability, quality and consistency

Metadata – data that is used to escribe other data, i.e. date of creation, size of a data file

MongoDB - NoSQL database that is open source under the management of 10gen

MPP Database - database that is optimized to work in an MPP processing environment

Multidimensional Database - a database that is used to store data in cubes or multidimensional arrays instead of the usual columns and rows used in relational databases. This allows the data to be analyzed from a number of angles for analytical processing and complex queries on OLAP applications

Multi-Threading - the process of breaking an operation in a single computer system up into several threads so it can be executed faster

N

Natural Language Processing - the ability for a computer system or program to understand human language Application of this include the ability for humans to interact with a computer by way of automated translation, speech and by determine the meaning of data that is unstructured, such as speech or text data.

NoSQL – database management system that avoids using the relational model. NoSQL can handle large volume of data that don't follow a fixed plan. It is best suited to use with large volumes of data that do not need the relational model

O

OLAP – Online Analytical Processing - the process of using three operations to analyze multidimensional data:

- Consolidation – aggregation of available
- Drill-down – allow users to see details that are underlying the main data
- Slice and Dice – allows users to choose subsets and see them from different perspective

OLTP -Online Transactional Processing - process by which users are given access to vast amounts of transactional data in a way that they are able to determine a meaning from it

ODCA – Open Data center Alliance - a consortium of IT organizations from around the globe who have a single goal – to hasten the speed at which cloud computing is migrated

ODS – Operational Data Store – location to store data from various sources so that a higher number of operations can be performed on the data before it is sent for reporting to the data warehouse

P

Parallel Data Analysis – the process of breaking an analytical problem up into several smaller parts. Algorithms are run on each part at the same time. Parallel data analysis can happen in a single system or in multiple systems

Parallel Method Invocation – PMI - a process that allows programming code to call many different functions in parallel

Parallel Processing - the ability to execute several tasks at once

Parallel Query - the execution of a query over several system threads to speed up performance

Pattern Recognition – labeling or clarification of a pattern that is identified in the machine learning process

Performance management – Process of monitoring the performance of a business or system against goals that are predefined to identify any specific areas that need looking at

Petabyte - 1024 terabytes or one million gigabytes

Pig - data flow execution and language framework used for parallel computation

Predictive Analytics – the use of statistical functions on at least one dataset to predict future events or trends

Predictive Modeling - the process by which a model is developed to predict an outcome or trend

Prescriptive Analytics- the process by which a model is created to "think" of all the possible options for the future. It will suggest the best option to take.

Q

Query Analysis – the process by which a search query is analyzed to optimize it to bring in the best results

R

R - Open source software environment that is used for statistical computing

RFID – Radio Frequency Identification – Technology that makes use of wireless communication to send information about a specific object from one point to another point

Real Time - descriptor for data streams, events and processes that have actions performed on them as soon as they occur

Recommendation Engine - an algorithm that is used to analyze purchase by a customer and their actions on a specific e-commerce site; the data is used to

recommend other product, including complementary ones

Records Management - the process of management of a business's records though their whole lifecycle. From the data of creation to the date of disposal

Reference Data - this is data that describes a particular object and its properties. This object can be a physical one or t may be virtual

Report – information gained from a dataset query and presented in a predetermined format

Risk Analysis - the use of statistical methods on at least one dataset to determine the risk value of a decision, project or action

Root-Cause Analysis – the process by which the main cause of a problem or event is determined

S

Scalability – ability of a process or a system to maintain acceptable levels in performance as scope and/or workload increases

Schema – Defining structure of data organization in a database system

Search - the process that uses a search tool to find specific content or data

Search Data - the aggregation of data about specified search times over a specified time period

Semi Structured Data - data that has not been structured with the use of a formal data model but provides alternative means of describing the hierarchies and data

Server – virtual or physical computer that serves software application requests and sends them over the network

SSD – Solid State Drive – sometimes called a Solid State and is a device that persistently stores data by sing memory ICs

SaaS – Software as a Service – application software used by a web browser or thin client over the web.

Storage - any means that can be used to persistently store data

Structured Data – data that is organized with the use of a predetermined structure

SQL – Structured Query Language – programming language that is specifically designed to manage data and retrieve it from a relational database

T

Terabyte – 1000 gigabytes

Text Analytics - application of linguistic, statistical and machine learning techniques on sources that are text-based, to try to derive insight or meaning

Transactional Data – Data with an unpredictable nature, i.e. accounts receivables data, accounts payable, data or data that relates to product shipments

U

Unstructured Data - data that does not have any identifiable structure, i.e. emails or text messages

V

Variable Pricing - the practice of changing a price on the fly, as a response to supply and demand. Consumption and supply have to be monitored in real-time

Conclusion

I am sure that by now, you would have realized the importance of having a sound system in place to manage your data. In order to effectively manage that data, you might need to expand your organization to include people that are skilled in analyzing and interpreting it effectively. With effective data management, you will find it easier to analyze data.

With the increasing competition, predictive analytics is also gaining more and more importance, as the days go by. I have shown you several case studies of large organizations that are using data to further expand and improve their operations. I hope that the information mentioned in this book gave you an insight into the field of predictive analytics.

I hope you enjoyed this book and apply the techniques mentioned in this book in your business.

Lastly, I'd like to ask you a favor. If you found this book helpful or enjoyed this book, then I'd really appreciate you leaving a review and your feedback on Amazon by clicking the link below.

Click Here To Leave a Review for This Book on Amazon

Thank you again.

Daniel Covington