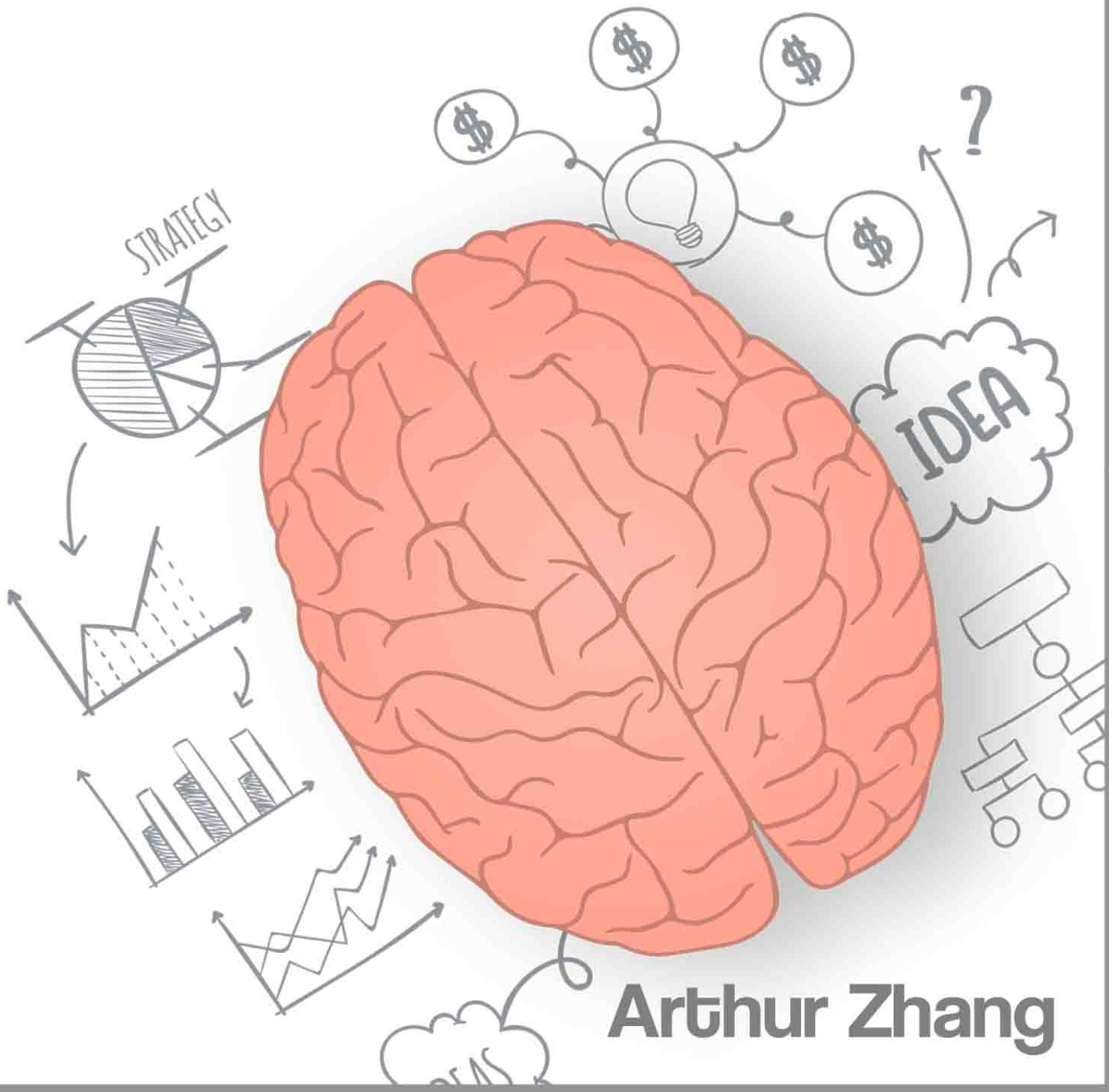
Data Analytics

Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life



Arthur Zhang

Chapter 1: Why Data is Important to Your Business

Have you ever been fascinated with ancient languages, perhaps those now known as “dead” languages? The complexity of these languages can be mesmerizing, and the best part about them is the extent to which ancient peoples went to preserve them. They used very monotonous methods to preserve texts that are anywhere from a few hundred years old to some that are several thousands of years old. Scribes would copy these texts several times to ensure they were preserved, a process that could take years.

Using ink made from burned wood, water, and oil they copied the text to papyrus paper. Some used tools to chisel the text into pottery or stone. While these processes were tedious and probably mind-numbing, the people of the time determined this information was so valuable and worth preserving that certain members of a society dedicated their entire lives to copying the information. What is the commonality between dead languages and business analytics?

The answer is data. Data is everywhere and flows through every channel of our lives. Think about social media platforms and how they help shape the marketing landscape for companies. Social media can provide companies with analytics that help them measure how successful – or unsuccessful – company content may be. Many platforms provide this data for free, yet there are other platforms that charge high prices to provide a company with high-quality data about what does or doesn’t work on their website.

When it comes to business, product and market data can provide an edge over the competition. That makes this data worth its weight in gold. Important data can include weather, trends, customer tendencies, historical events, outliers, products, and anything else relevant to an aspect of business. What is different about today is how data can be stored. It no longer has to be hand-copied to papyrus or chiseled into stone. It is an automatic process that requires very little human involvement and can be done on a massive scale.

Sensors are connected to today's modern scribes. This is the Internet of Things. Most of today's devices are connected, constantly collecting, recording, and transmitting usage and performance data. Sensors collect environmental data. Cities are connected to record data relevant to traffic and infrastructure information to ensure they are operating efficiently. Delivery vehicles are connected to monitor their location and functionality, and if mechanical problems arise they can usually be addressed early. Buildings and homes are connected to monitor energy usage and costs. Manufacturing facilities are connected in ways that allow automatic communication of critical data sets. This is the present – and the future – state of "things."

The fact that data is important isn't a new concept, but the way in which we collect the data is. We no longer need scribes; they have been replaced with microprocessors. The ways to collect data, as well as the types of data to be collected, is an ever-changing field itself. To be ahead of the game when it comes to business, you've got to be up-to-date about how you collect and use data. The product or service provided can establish a company in the market, but data will play the critical role in sustaining the success of the business.

The technology-driven world in which we live can make or break a business. There are large companies that have disappeared in a short amount of time because they failed to monitor their customer base or progress. In contrast, there are smaller startup businesses that have flourished because of the importance they've placed on customer expectations and their numbers.

Data Sources

Sources of data for a business can range from customer feedback to sales figures to product or service demands. Here are a few sources of data a business may utilize:

- Social media: LinkedIn, Twitter, and Facebook can provide insight into the kind of customer traffic your web page receives. These platforms also provide cost-effective ways to conduct surveys about customer satisfaction with products or services and customer preferences.
- Online Engagement Reporting: Using tools such as Google Analytics or Crazy Egg can provide you with data about how customers interact with your website.
- Transactional Data: This kind of data will include information collected from sales reports, ledgers, and web payment transactions. With a customer relationship management system, you will also be able to collect data about how customers spend their money on your products.

How Data Can Improve Your Business

By now you've realized that proper and efficient use of data can improve your business in many ways. Here are just a few examples of data playing an important role in business success.

Improving Marketing Strategies: Based on the types of data collected, it can be easier to find attractive and innovative marketing strategies. If a company knows how customers are reacting to current marketing techniques, it will allow them to make changes that will fall in line with trends and expectations of their customers.

Identifying Pain Points: If a business is driven by predetermined processes or patterns, data can help identify points of deviation. Small deviations from the norm can be the reason behind increased customer complaints, decreased sales, or a decrease in productivity. By collecting and analyzing data regularly, you will be able to catch a mishap early enough to prevent irreversible damages.

Detecting Fraud: In the absence of proper data management, fraud can run rampant and seriously affect business success. With access to sales numbers in hand, it will be easy to detect when and where fraud may be occurring. For instance, if you have a purchase invoice for 100 units, but your sales reports only show that 90 units have been sold, you know that ten units are missing from inventory and you will know where to look. Many companies are silent victims of fraud because they fail to utilize the data to realize that fraud is even occurring.

Identifying Data Breaches: With the availability of data streams ever-increasing, it creates another problem when it comes to fraudulent practices. Although comprehensive yet subtle, the impacts of data breaches can negatively affect accounting, payroll, retail, and other company systems. Data hackers are becoming more sneaky and devious in their attacks on data systems. Data analytics will allow a company to see a possible data breach and prevent further data compromises which might completely cripple the business. Tools for data analytics can help a company to develop and implement data tests that will detect early signs of fraudulent activity.

Sometimes standard fraud testing is not possible for certain circumstances, and tailored tests may be a necessity for detecting fraud in specific systems.

In the past, it was common for companies to wait to investigate possible fraudulent activity and implement breach safeguards until the financial impacts became too large to ignore. With the amount of data available today this is no longer a wise – or necessary – method to prevent data breaches. The speed at which data is dispersed throughout the world can mean a breach could happen from one point to the next, crippling a company from the inside out on a worldwide scale. Data analytics testing can prevent data destruction by revealing certain characteristics or parameters that may indicate fraud has entered the system. Regular testing can give companies the insight they need to protect the data they are entrusted to keep secure.

Improving Customer Experience: Data can also be gathered from customers in the form of feedback about certain business aspects. This information will allow a company to alter business practices, services, or products to better satisfy the customer. By maintaining a bank of customer feedback and continually asking for feedback you are better able to customize your product or service as the customers' needs change. Some companies send customized emails to their customers, creating the feeling that they genuinely care about their customers. They do this most likely because of effective data management.

Making Decisions: Many important decisions about a business require data about market trends, customer bases, and prices offered by competitors for the same or similar products or services. If data does not influence the decision-making process, it could cost the company immensely. For example, launching a new product in the market without considering the price of a competitor's product might cause your product to be overpriced – therefore creating problems when trying to increase sales. Data should not only apply to decisions about products or services, but also to other areas of business management. Certain datasets will provide information on how many employees it will take to foster the efficient functioning of a department. This will allow you determine where you are understaffed or overstaffed.

Hiring Process: Using data to select the right personnel seems to be neglected by many corporations. For effective business operation, it is crucial to put the right candidate in the right position. Using data to hire the most qualified person for a position will ensure the business will remain highly successful. Large companies with even larger budgets use big data to seek out and choose skilled people for their open positions. Smaller companies would benefit from using big data from the beginning to staff appropriately to further the successes of a startup or small business. This method of using gathered data during hiring has been proven to be a lucrative practice for various sizes of organizations. Data scientists can extract and interpret specific data needed from the human resources department for hiring the right person.

Job Previews: By providing an accurate description of an open position, a job seeker will be better prepared about what to expect should they be hired for the position. Pre-planning the hiring process utilizing data about the open position is critical in appealing to the right candidate. Trial and error are no doubt a part of learning a new job, but it slows down the learning process. It will take the new employee longer to catch up to acceptable business standards which also slows their ability to become a valuable company resource. By incorporating job preview data into the hiring process, the learning curve is reduced, and the employee will become more efficient faster.

Innovative Methods for Gathering Data for Hiring: Using new methods of data collection in the hiring process can prove to be beneficial in hiring the right professional. Social sites that collect data, such as Google+, Twitter, Facebook, and LinkedIn can give you additional resources for recruiting potential candidates. A company can search these sites for relevant data from posts made by the users to connect to qualified applicants. Keywords are the driving force for online searches. Using the most visible keywords in a job description will increase the number of views your job posting will receive.

Traditionally, software and computers have been used to determine if an employee would be better suited for another position within the company or to terminate employment. However, using this type of resource can also

help to find the right candidate for a job outside of the company. Basic standards such as IQ or skills tests can be limiting, but focusing on personality traits may open the field of potential candidates. By identifying personality characteristics, it will help to filter out candidates based on traits that will not be beneficial to the company. If a person is argumentative or prefers to be isolated, they certainly wouldn't thrive in a team-oriented environment. By eliminating mismatches between candidates and job expectations, it will save the company time, training materials, and other resources. By utilizing this type of data collection, it would not only find candidates with the right skills but also with the right personalities to align with current company culture. Being sociable and engaging will foster the new employee as they learn their new role. It's important that new candidates fit well with seasoned employees to reinforce working relationships. The health of the working environment greatly influences how productive the company is overall.

Using Social Media to Recruit: Social media platforms are chock full of data sources for finding highly qualified individuals to fill positions within a company. On Twitter, recruiters can follow people who tweet about a certain industry. A company can then find and recruit ideal candidates based on their interest and knowledge of an industry or a specific position within that industry. If someone is constantly tweeting about new ideas or innovations about an industry aspect, they could make a valuable contribution to your company. Facebook is also valuable for this kind of public recruitment. It's a cost-effective way to collect social networking data for companies who are seeking to expand their employee base or fill a position. By "liking" and following certain groups or individuals a company can establish an online presence. Then when the company posts a job ad, it is likely to be seen by many people. It is also possible to promote ads for a small fee on Facebook. This means your ad will be placed more often in more places, increasing your reach among potential candidates. It's a geometrical equation – furthering your reach with highly effective job data posts increases the number of skilled job seekers who will see your ad, resulting in a higher engagement of people who will be a great fit for your company.

Niche Social Groups: By joining certain groups on social media platforms recruiters will have access to a pool of candidates who most likely already possess certain specific skills. For instance, if you need to hire a human resources manager, joining a group comprised of human resource professionals can potentially connect you with your next hire. Within this group, you can post engaging and descriptive job openings your company has. Even if your potential candidate isn't in the group, other members will most likely have referrals. Engaging in these kinds of groups is a very cost-effective method to advertise open positions.

Gamification: This is an underused data tool but can be effective if the hiring process requires multiple steps or processes. By rewarding candidates with virtual badges or other goods, it will motivate candidates to put forth effort during the selection process. This will allow their relevant skills in performing the job to be highlighted and is a fun experience when applying for a job which is typically a rather boring process.

These are only a few of the ways in which data can help companies and human resource departments streamline the hiring process and save resources. As you can see, data can be very important for effective business functioning, and you've also seen the multitude of uses it has for *just* the hiring process. This is why proper data utilization is critical in business decision making for all other aspects of your business.

Chapter 2: Big Data

Across the globe, data and technology are interwoven into society and the things we do. Like other production factors – such as human capital and hard assets – there are many parts of the modern economic activity that couldn't happen without data. Big data is, in short, the large amounts of data that are gathered in order to be analyzed. From this data, we can find patterns that will better inform future decisions.

This data and what can be learned from it will become how companies compete and grow in the near future. Productivity will be greatly improved, as well. Significant value will be created in the economy of the world because of increase in the quality of services and products while reducing waste. While this data has been around, it has only really excited people that are already interested in data. As times have changed, we are getting more and more excited by the amount of data that we're generating, mining and storing. This data is now one of the most important economic factors for so many different people.

In the present, we can look back at trends in IT innovation and investment. We can also see the impact on productivity and competitiveness that have resulted from those trends and how big data can make large changes in our modern lives.

Like the previous IT-enabled innovations, big data has the same requirements to move productivity further. For example, if you see innovations in current technology, then there will need to be a close following after of complementary management innovations. Big data technology supplies and analytic capabilities are so advanced now that it will have just as much of an impact on productivity as suppliers of other technologies. Businesses around the world will need to start taking big data seriously because of the potential it has to create some real value. There are already retail companies that are putting big data to work because of the potential it has to increase the operating margins.

Big Data – A New Advantage

Since it has come to light, big data is becoming an incredibly important way that companies are outperforming each other. Even new entrants into the market are going to be able to leverage strategies that data has found in order to compete, innovate, and attain real value. This will be the way that all the different companies, new and established, will compete on the same level.

There are already examples of this competition everywhere. In the healthcare industry, data pioneers are looking at the outcomes of some pharmaceuticals that are widely prescribed. From the analysis of the results, they learned that there were risks and benefits that had not been seen in the limited trials that companies had run with the pharmaceuticals.

There are other industries that are using the sensors in their products to gain data that they can use. This can be seen in children's toys, large-scale industrial goods, and so many others. The data that they gather show how the products are used in real life. With this data, companies can make improvements on the products based on how people are really using them. This will make these products so much better for the future users.

Big data is going to help create new growth opportunities and create new companies that specialize in aggregating and analyzing data. There's a good proportion of companies that will sit right in the middle of flowing information. They'll be receiving information and data that comes from many sources just to analyze it. Managers and company leaders that are thinking ahead need to start creating and finding new ways to make their companies capable of dealing with big data. People that do so will need to be especially aggressive about it.

It's important to realize that not only the amount of big data but the high frequency and real-time nature of data as well. There's the idea of "nowcasting" around right now. This process is estimating metrics right away. These metrics can be things like consumer confidence. Knowing that information so soon used to be impossible and only something that could be done after a while. "Nowcasting" is being used more and more, adding a lot of potential to the ways that companies predict things.

The high frequency of the data will allow users to try to test theories and analyze the results in ways that they were incapable of before. There have been studies of major industries that have found ways that big data can be used:

1. Big data can unlock serious value for industries because it makes information transparent. There is a lot of data that isn't being recorded and stored. There is still a lot of information that cannot be found as well. There are people that are spending a quarter of their time looking for extremely specific data and then storing it, sometimes in a digital space. There's a lot of inefficiency in this work right now. More and more companies are storing data from transactions online, these people are able to collect tons of accurate and detailed information about everything. They can find out inventory and even the number of sick days that people are taking.

Some companies are already using this data collection and analysis to do experiments and see how they can make better-informed management decisions. Big data allows companies to put their customers into smaller groups. This will allow them to tailor the services and products that they are offering. More sophisticated analytics are also allowing for better decision making to happen. There are fewer risks and bring light to information and insights that might not have seen the light of day.

Big data can be used to create a brand new generation of services and products that wouldn't have been otherwise possible. Some manufacturers are already using the data that has been collected from their sensors to figure out more efficient and useful after-sales services.

Big Data Creates Value

Using the US healthcare system as an example, we can look at ways that big data can really create good value. If the healthcare system used big data to use the efficient and quality of their services, they would actually create $300 billion of value every year. 70% of that value would have been seen from a cut in expenditures. These expenditures that would be cut are only 8% of the current expenditures.

If you look at European developed economies instead, you can see a different way that big data creates value. The government administrations could use big data in the right way to improve operational efficiency. That would result in about €100 billion worth of value every year. This is just one area. If the governments used advanced analytics and boosted tax revenue collection, they would create ever more value just from cutting down on errors and fraud in the system.

Even though we've been looking at companies and governments so far, they aren't the only ones that are going to benefit from using big data. A consumer will benefit from this system as well. Using location data in specific services, people could find a consumer surplus of up to $600 billion. This can be seen especially in systems and apps that use real-time traffic information to make smart routing. These systems are some of the most used on the market and they use location data. There are more and more people using smartphones. Those that have smartphones are taking advantage of the free map apps that are available. With an increase in demand, it's likely that the nmber of apps that use smart routing are going to increase.

By the year 2020, more than 70% of mobile phones are going to have GPS capabilities built into them. In 2010, this number was only 20%. Because of the increase in GPS capable devices, we can expect that smart routing will have the potential to create savings of around $500 billion in fuel and time that people will spend on the road. That amount of money is equal to around 20 billion driving hours. It's like saving a driver 15 hours a year on the road. This would save them $150 billion dollars in fuel.

While we have seen specific pools of data in the examples listed above, but big data has a huge potential in combined pools of data. The US healthcare system is a great way to look at the potential future of big data. The healthcare system has four distinct data pools: clinical, medical, pharmaceutical products; research and development; activity and cost; and patient data. Each data pool is captured and managed by a different portion of the healthcare system.

If big data was used to its full potential, then the annual productivity of the healthcare system could be improved around 0.7%. But it would take the combination of data from all these different sources to create that improved efficiency. The unfortunate part is that some of the data would need to come from places that do not share their data at scale right now. Data like clinical claims and patient records would need to somehow be integrated into the system.

The patient, in turn, would have better access to more of their healthcare information and would be able to compare physicians, treatments, and drugs. This would allow patients to pick out their medications and treatments based on the statistics that are available to them. However, in order to get these kinds of benefits, patients would have to accept a trade for some of their privacy.

Data security and privacy are two of the biggest roadblocks in the way of this. We must find a way around them if we really ever want to see the true benefits of using big data. The most prevalent challenge right now is the fact that there is a shortage of people that are skilled in analyzing big data properly. By 2018, the US will be facing a shortage of 140,000 and 190,000 people with training in deep analysis. They'll also be facing a shortage of roughly 1.5 million people that have the quantitative skills and managerial experience needed to interpret the analyses correctly. These people will be basing their decisions off of the data.

There are many technological issues in the way as well that will need to be resolved before big data can be used effectively by more companies. There are so many incompatible formats and standards that are floating around as well as legacy systems that are stopping people from integrating data and from using sophisticated analytical tools to really look at the data sets.

Ultimately, there will have to be technology made for computing and storage through to the application of visualization and analytical software. All this technology will have to be available in a stack so that it is more effective. In order to take true advantage of big data, there has to be better access to data, and that means all of it. There are going to be so many organizations that will need to have access to data stores and maintained by third parties to add that data in with their own. These third parties could be customers or business partners.

This need for data will mean that companies that really need data will have to be able to come up with interesting proposals for suppliers, consumers, and possibly even competitors in order to get their hands on that data. As long as big data is understood by governments and companies, the potential it has to deliver better productivity will ensure that there will be some incentive for companies to take the actions that they have to get over the barriers that are standing in the way. In getting around these barriers, companies will find new ways to be competitive in their industries and against individual companies. There will be greater productivity and efficiency all around which will result in better services, even when money is tight.

Big Data Brings Value to Businesses Worldwide

Big data has been bringing value to business internationally for a while. The amount of value that it will continue to bring is almost immeasurable. There are several ways that the big data has impacted the world so far. It has created a brand new career field in Data Science. Data interpretation has been changed drastically because of big data. The healthcare industry has been improving quickly and considerably since they added predictive analytics into part of their business. Laser scanning technology is changing and has changed the way that law enforcement officers reconstruct crime scenes. Predictive analytics are changing how caregivers and patients interact. There are even data models that are being built now to look at business problems and help find solutions. Predictive analytics has had an impact on the way that the real estate industry conducts business.

Big Data is a Big Deal

Besides the fact that data is bringing so much value to so many different companies and industries, it is also opening up a whole new path of management principles that companies can use. Early on in professional management, corporate leaders discovered that one of the key factors for competitive success was a minimum scale of efficiency.

Comparatively, one of the modern factors for competitive success is going to be capturing higher quality data and using that data with more efficiency at scale. For the current company executives that might be doubting how much big data is going to help them, there are these five questions that will really help them figure out how big data is going to benefit them and their organizations.

What can we expect to happen in a world that is “transparent” meaning that data is readily available?

Over time information is becoming more accessible in all sectors. The fact that that data is coming out of the shadows means that organizations, which have relied heavily on data as a competitive asset, are potentially going to feel threatened. This can be seen especially in the real-estate industry.

The real-estate industry has typically provided a gateway to transaction data and a knowledge of bids and buyer behaviors that haven’t been available elsewhere. Gaining access to all of that requires quite a bit of money and even more effort. In recent years, online specialists are bypassing the agents to create a parallel resource for real-estate data. This data is gotten directly from buyers and sellers, and available to those same groups.

Pricing and cost data has also seen a spike in availability for several industries. There are even companies using satellite imagery that is available at their fingertips. They’re using processing and analysis to look at the physical facilities of their competitors. That information can provide insights into what expansion plans or physical constraints that their competitors are facing. But with all that data there comes a challenge. The data is being kept within departments. Engineering, R&D, service operations, and manufacturing will have their different information and it will be stored in different ways depending on the department.

However, the fact that all this information is kept in these little pockets means that the data cannot be used and analyzed in a timely manner. This can cause all sorts of problems for companies. For example, financial institutions don't share data across departments like money management, financial markets, or lending. This segmentation means that the customers have been compartmentalized. They don't see the customer across all of these different areas, but just as separate images.

Some companies in the manufacturing business are trying to stop this separation of data. They're integrating data from their different systems and asking their smaller units to collaborate in order to help their data flow. They're even looking for data and information outside of their groups to see if there's anything else out there that might help them figure out better products and services.

The automotive industry has suppliers all around the world making components that are then used in the cars that they're making. Integrating data across all of these would allow the companies and their supply chain partners to work together at the design stage instead of later on.

Can testing decisions change the way that companies compete?

Gaining the ability to really test decisions would cut down on costs and improve a company's competitiveness. These automotive companies would be able to test and experiment with the different components. By going through this process, they'll be able to gain results and data that will guide their decisions about operational changes and investments. Really, experimentation will allow companies and their managers to really see the difference between correlation and causation while also boosting financial performance and producing more effective products.

The experiments that companies will use to collect data can take several forms. Some online companies are always testing and running experiments. In particular cases, there will be a set of their web page views that they are using to test the factors that drive sales and higher usage. Companies with physical products will use tests to help make decisions, however, big data can make these experiments go even further.

McDonald's put devices in some of their stores that track customer interaction, traffic, and ordering patterns. The data gained through these devices can help them make decisions about their menus, the design of their restaurants, as well as many other things.

Companies that can't use controlled experiments may turn to natural experiments to figure out which variables are in play. A government sector collected data on different groups of employees that were working in various places but doing similar jobs. This data was made available and the workers that were lagging were pushed to improve their performance.

What effect will big data have on business if it is used for real-time customization?

Companies that deal with the public have been dividing and targeting specific customers for quite a while now. Big data is taking that further than it ever by making it possible for real-time personalization to become part of these companies. Retailers may become able to track individual customers and their behaviors by monitoring their internet click streams. Knowing this, they will be able to make small changes on websites that will help move the customer in a direction to buy. They will be able to see when a customer is making a decision on something they might purchase. From here, they will be able to "nudge" the customer towards buying. They could offer bundled products, benefits, and reward programs. This is real-time targeting.

Real-time targeting also brings in data from loyalty groups. This can help increase higher-end purchases made by the most valuable customers. The retail industry is likely to be the most driven by data. Because they're keeping track of internet purchases, conversations taking place on social media, and location data pulled from smartphones, they've got tons of data at their fingertips. Besides the data, they have better analytical tools now that can divide customers into smaller segments for even bettering targeting.

Will big data just help management or will it eventually replace it?

Big data opens up new ways for algorithms and analysis, mediated by machines, to be used. Manufacturers are using algorithms to analyze the

data that's being collected from sensors on the production line. This data and analysis help the manufacturers regulate the processes, reduce their waste, increase outfit, and even cut down on potentially expensive and dangerous human intervention.

There are "digital oilfields," where sensors monitor the wellheads, pipelines, and mechanical systems all the time. The data is fed into computers where the data is turned into results that are given to the operation centers where the oil flows are adjusted to post production and reduce the amount of downtime for the whole process. One of the largest oil companies has managed to increase oil production by five percent, while also reducing staff and operating costs by ten and twenty-five percent.

Products ranging from photocopiers to jet engines are now tracking data that helps people understand their usage. Manufacturers are able to analyze data and fix the problems, whether they're just simply fixing glitches in software or needing to send out a repair representative. The data is even predicting when products will fail and being used to schedule repairs before they're likely to fail. It's obvious that big data can create huge improvements in performance and help make risk management easier. The data could be used to even find errors that would otherwise unseen.

Because of the increasing demand for analytics software, communication devices, and sensors; prices for these things are falling fast. More and more companies will be able to find the time and money to get involved in collecting data.

Will big data be used for the creation of brand new business models?

Big data has already been responsible for the creation of new industries surrounding the analysis and use of the information it has. But the company categories that are also being produced big data have business models that are driven entirely by data. Many of these companies are intermediaries in a value chain. They are generating valuable "exhaust" data from transactions.

A major transport company was keeping data about their own business, but they were also collecting vast amounts of data about what products were

being shipped where. They took the opportunity and began selling the data that they were collecting to supplement economic and business forecasts.

There was another global company that was learning a lot by looking at their own data. From doing the analysis for themselves, they eventually decided to branch out and create a business that analyzes data for other organizations. The business aggregates supply chain and chop floor data for manufacturers. It also sells relevant software tools that a company will need to improve their own performance. This side business that the company opened is outperforming the manufacturing business, and that is because of the value of big data.

Big data is creating a whole new support model for the markets that already exist today. Companies have all sorts of new data needs, and they need qualified people to support that data. As a result, if you own a business, then you may need an outside firm to analyze and interpret any data you're producing for you.

These specialized firms can take large amounts of data in various forms and break it down for you. These firms exist because there is a need for support for larger companies in many different industries. The employees they hire are trained to locate and capture data in systems and processes. They're allowing larger companies to focus on their work and doing the data aggregation for the company. They assimilate, analyze, and interpret trends in the data, and then they report to the company about any notifications that they have.

For a company that doesn't want to hire out a firm, they have the option to create a data support department within their own company. This would be more cost-effective than hiring an entire outside firm, but it does require very specific and specialized skills within the company. The department would focus on taking the data flow and analyzing, interpreting, and finding new ways to use the data. These new applications and the new data department would monitor existing data for fraud, triggers, or issues.

Big data has created a whole new field of studies in colleges and higher institutions of learning. People are training in the latest methods of big data gathering, analyzing, and interpreting. This path will lead people to critical positions in the newly trending data support companies.

Big data has created all sorts of changes, and it will continue to make even more. In education areas, big data will influence and change the way that teachers are hired. The data will be able to look at recruiting processes and predictive analytics will be able to look at the traits that most effective teachers are going to need to most properly maximize the learning experience.

Chapter 3: Development of Big Data

While most of our data collection and analysis has only happened in the last couple of years, the term “big data” has been in our vocabulary since 2005. Analysis of data has been around for as long as we could count. Accounting in ancient Mesopotamia tracked the increases and decreases of herds and crops and even then we were trying to find patterns in that data. In the 17th century, John Graunt published a book, “Natural and Political Observations Made upon the Bills of Mortality,” that was the first large-scale example of data analysis. It provided insight into the causes of death at the time, and the book was meant to help stop the Bubonic plague.

Graunt’s book and the way he approached the data was a revolution. Statistics, as it is now, was invented at that time, even though we couldn’t use it fully before the invention of computers. Data analysis came in in the 20th century when the information age really began. There were many examples of early data analysis and collection even in the beginning. There was the machine invented Herman Hollerith that could analyze data in 1887; it was used to organize census data. Roosevelt’s administration used big data for the first time to keep track of the social security contributions for millions of Americans.

The first real data processing machine came during World War 2. The British intelligence wanted to decipher Nazi codes. The machine, Colossus, processed 5,000 characters per second to find the patterns in coded messages. The task of deciphering went from weeks to just hours. This was a huge victory for technology and a massive improvement for statistical analysis.

In 1965, the electronic storage of information started, as another idea of the American government. The system was put in place to store tax return claims and fingerprints. However, the project went unfinished because of the worries of the American people. They thought of that as something similar to “Big Brother,” but the electronic storage of information was already starting. It would be impossible to stop the flow of information.

The invention of the Internet was really what sparked the true revolution in data storage and analysis. Tim Berners-Lee couldn’t have known what he

had really started in the world. However, it was really in the 90s that his system was turned into the monster that it is today. In 1995, the first supercomputer was made. The machine was capable of doing in a single second what a human with a calculator could do in 30,000 years. This was the next great stride in data analysis.

In 2005, Roger Mougalas mentioned the term big data as a way of saying that traditional tools could not deal with the amount of data that was being collected. In that year, Hadoop was invented to index the internet. Today, this tool is used by companies to go through their own data.

Eric Schmidt said in 2010 at a conference that the amount of information created between the dawn of the time and 2003 (roughly 10 exabytes) was equal to the amount of information that had been created in just two days in 2010. Data had become so ingrained in our everyday lives. Hundreds of upstarts are attempting to take on big data. Thousands of business are using the data to optimize their business models. Almost every industry is using the inferences made by analyzing big data. That information has become the most valuable currency in the world; the second most valuable things are the people that are able to properly use it.

As the world becomes more caught in up in the digital world, it will bring us closer to each other. It was also brought more of our lives into the public eye. Data collection will become more and more important. Companies will be using all of this data to find new ways to sell people products and services. There's no doubt that the government will also be using it to improve the environment, get votes, and keep the people in check.

Even with big data, the future is still a mystery since it may go either way. The future could be changed for the better by big data; the future could also be hurt by the ways that private corporations are using that data now. Having more data out in the open gives more and more power to the governments. And one day it may lead to the realization of the people's "Big Brother" fears.

Chapter 4: Considering the Pros and Cons of Big Data

Back in the day, whenever a crisis such as a recession or a bubble burst hit, no one could truly understand it, and all anyone could come up with would be: “something went wrong somewhere.” Nowadays, however, due to the rise of big data, it is much easier to precisely describe socio-economic, political, and other types of factors. Though many might think that this quantum leap in our capabilities to measure and analyze data would be nothing but positive, there may be some negative implications that we must remember to keep in mind. The discussion on the existence of “big data” and how it shapes and will continue to shape our future has been never-ending, ever since the very concept of “big data” was introduced. Very few would dispute that big data has proved to be the catalyst of many positive changes and developments in our everyday lives. What many don’t see, however, are the various harms that big data has been introducing into our lives as well. Some economic and social studies experts have posited that the reduction in personal privacy thanks to the often unfettered access to our personal information that public and private corporations have is only one of the least of the drawbacks. Even national politicians in Washington have become aware of this growing unease that big data may be negatively impacting the lives of the average Joe. The White House has addressed the issue, stating that big data must seek balance between its socio-economic value and the privacy it may have been violating in order to become one of the strongest catalysts for socio-economic progress. Here we examine some of the pros and cons of big data in our modern society.

The Pros

As was earlier stated, big data may be an immense help to both the private and public sectors, but how exactly? Here we can find some of the more common ones.

New methods of generating profit

This one may not be directly beneficial to most, as the ones who benefit the most are company owners and employers, but greater revenues lead to a stronger economy, which means that more people can keep their jobs. It is crucial for companies to be able to turn a profit in order for them to be able to employ people.

Big data may open new opportunities for companies in any sector. Companies that directly use big data gather valuable information that is desired by other companies as well. The raw data as well as the analyzed and interpreted form of such data may be sold to other companies, generating even more revenue.

Improving Public Health

Going into more specific examples, healthcare is a component of the public sector that is a beneficiary of big data. The improved ability to gather and analyze massive amounts of data about hospital staff, patients, and even the wants and needs of the public has allowed experts to better develop methods and policies that will be more responsive to public needs. Perhaps even more important is the merger of big data and the science of genetics. This merger is one of the things that will revolutionize the world. It may someday be possible to include a patient's genetic code in their health records. It may be possible to analyze these genetic maps in order to discover the genetic bases of certain illnesses. The possibilities are endless, and no one knows just how much the partnership between big data and the health sector will be able to benefit us. Unlike most other industries, healthcare has been lax in following the trend of personalized services, but the arrival of big data will help in picking up the slack, and will indubitably shift the trend and bring us closer to the age of truly personalized medicine.

Improving Our Daily Environment

How many trash cans are needed on the street? What amount of street lamps is needed? At what point in the day do traffic jams occur? These are all questions that are easily answered through the use of big data. Thanks to the development of modern data gathering systems, it has never been easier to find out what happens in our public spaces. This data can be used not only to save vast amounts of money, but to create significant and concrete impacts in our daily lives. The city of Oslo in Norway has been able to greatly reduce the amount of energy used in their street lighting. Portland, Oregon has used big data to reduce their carbon dioxide emissions. Even the police department of Memphis, Tennessee has reduced the serious crime rate in their area by 30% through big data. Big data revolutionizes how we run our cities, and this is only the beginning. In the future, it may be possible that a central mainframe could gather and analyze the data in real time, and use this data to tweak the performance of the city's services. Imagine the improvements that this may bring to our cities. More and more cities are beginning to incorporate big data into their systems, and eventually, every city will be using this to improve our daily lives.

Improving Decisions: Speed and Accuracy

Regardless of the industry, and no matter what the final target may be, may it be increased security, revenue, or healthcare, the existence of big data lets us respond faster. Big data affords anyone using it the ability to make more informed decisions, from how to market to individual customers to providing adequate healthcare for everyone. As the big data industry evolves, we become better and better at being able to analyze it in real-time, which allows us rapid results and helps assist our decision-making.

Personalized Products and Services

Companies develop products based on what they think customers may buy. Now, with the advent of big data, companies are better able to find out about people's interests and preferences. One of the services that sees great use today is Google Trends, allowing companies to find what people are searching for on the World Wide Web. This data allows companies to develop personalized products and services that are even more responsive to consumer needs.

The Cons

As was mentioned earlier, while big data has many benefits, it is not without its negative side. The positive aspects are extremely helpful in developing society and moving progress forward, but there are certain aspects to it that give people legitimate cause for concern. Like anything too good to be true, big data can be a double-edged sword.

Privacy

The greatest critics of big data have been civil rights activists and people who maintain the belief that privacy is more valuable than any advantage that big data grants us. Big data collects personal data, and this allows companies to learn numerous things about any individual user. This enables marketers to use this knowledge to sell products by manipulating the subconscious of unsuspecting users. There are numerous methods used by marketers that allow them to convince us to buy products we would probably not have bought, and most of these methods make good use of what big data says about us. Detractors of big data say that this constitutes an unjustifiable invasion of our privacy, especially when carried out by the private sector. This argument carries a lot of weight, and should be considered. Big data tells marketers so much about us, companies can even tell a what color a product should be so people would be better incentivized to buy it. While this may sound like a magic trick, it’s quite real, and shows one of the dark sides of big data being commercialized.

Big Brother

Ever since its introduction into the mainstream culture by Orson Welles, the concept of a “Big Brother” has been a constant specter looming over everyone. We know that our governments observe us and carry out certain activities to ensure that we are kept “in check”. Some believe this more strongly than others, with certain conspiracy theorists positing that a cabal of men and women run the world from the shadows. Though many dismiss this, even the most moderate of us know and understand that governments really do collect a lot of personal data, some of which they may not have any business collecting. “Big Brother” as a concept has been a specter, but with the advent of modern technologies, this specter seems more and more likely to turn into reality. In most American cities, one cannot walk more than a few streets without being caught on camera. Most, if not all of our devices such as phones or even cars have GPS signals that an unscrupulous entity may be able to take advantage of. Even satellite footage has become more accessible to governments, begging the question: “are we ever really alone?”.

Over the years, the concept of Big Brother has been a very present argument for everyone, from the average, everyday people to conspiracy theorists. For quite some time now, we have known that our governments have been watching us and doing all sorts of things to keep us “in check.” Some conspiracy theorists go so far as to say that a small group of extremely powerful men and women are now running the entire world, but even the most moderate of us still understand that governments do collect a huge amount of data that they might not really have the right to collect. The fear of the Big Brother is something that is very prominent in Western societies, but as time progresses, it seems that this is becoming more and more real. For instance, it is now nearly impossible to walk several blocks of any American city without being filmed by numerous cameras. There is also the topic of the GPS devices that are on our phones and vehicles. Satellite footages are becoming more and more available to governments, and the question we have to ask ourselves is: Are we ever alone?

The sheer amount of data collected has done some good in the world, as we earlier saw, but naturally, people desire some measure of privacy, even

when they have nothing serious to hide. There have been recent leaks that revealed the existence of phone tapping, social media monitoring, and other such forms of government surveillance. This leads to a sense of distrust and unease, even for citizens not up to anything malicious. There has to be a balance, and people are afraid that if the government knows too much about the personal lives of its citizens, it will be holding too much power, as information, especially in our modern age, is power. This is why regulations are key to limiting the access of public agencies to big data. Even given a democratic system, a government with so much information holds a lot of power over its citizens, and citizens should at least have the right to be asked what information they want accessible.

Stifling Entrepreneurship

Small businesses are not banned from using big data; far from it. However, with the sheer amount of resources and capacity large corporations can bring to bear, it is well-nigh impossible for a small business to compete. One of the methods that a small business has always had access to in order to compete is the personalization of their services and products. With the dozens or even hundreds of data scientists large corporations have at their disposal, they can easily sift through the extensive amounts of data to better target their market. This neutralizes any comparative advantage a small business may have once had, and there will be virtually no way for a small business to offer something a big corporation can’t.

Data Safekeeping

Given the massive amount of data gathered, there is no feasible way to store it in traditional physical media. All this data is stored on computers, all accessible through the internet. It may already be bad enough that a corporation has all of your personal information, but how much worse does it become once they decide to sell it? What if hackers decide to steal it from them? These are all legitimate concerns held by many people. In fact, it is well known that companies do share user data among themselves, regardless of its legality. This leads to the very real possibility of someone's personal information landing in the hands of a company that they have never even interacted with. The possibilities get even worse when one accounts for hackers who may be able to access user data such as pictures, addresses, or even credit card numbers. Companies are constantly upgrading their systems to protect against this possibility, but hackers also constantly find ways around these defenses. There is no surefire way to protect the data from unwanted hands, which means that perhaps some sort of limit should be imposed on this type of data collection.

Erroneous Data Sets and Flawed Analyses

We already know that big data has major potential in shaping the directions that businesses, corporations, and the public sector take. This, however, is assuming that the proper data is collected, and it is properly analyzed. Nothing is perfect, however, and neither data collection systems nor the analysis systems are even near perfect. These flaws mean that businesses, corporations, or anyone intending to use big data should use their data judiciously, as over-reliance on this may prove to be counterproductive.

There are many opportunities for flaws to interfere with the process. Beginning with data collection, collection systems may not collect sufficient data, or perhaps the data gathered is skewed or biased, therefore intrinsically flawed. Even if the analysis of such data would be perfect, you would end up with incorrect conclusions and inferences, which may cost millions per error. That is assuming the data analysis is flawless. It is also possible, even likely, that the analysis would be flawed, leading to imperfect conclusions. In fact, it has been found that over 15% of data analysis projects are flawed enough that the very existence of the companies using them may be in jeopardy. In addition, around 30% of other projects may end up with a net loss. This just goes to show that while big data is useful and potentially game-changing, it must be used with caution, otherwise it may do more harm than good.

Conclusions

At the end of the day, big data has both its pros and its cons. Much like any other important issue, deliberate observation is needed, and both its benefits and drawbacks must be considered. There have been countless ways that big data has improved our lives, but there have also been causes for concern. Big data will most likely be the subject of numerous discussions and political debates, and the extent of its regulation or lack thereof still remains to be seen. We can be sure, however, that our lives are becoming more and more open every day, no matter what we do. We are being monitored more than ever, but we must keep in mind that the bulk of this data, especially those collected by the public sector, has been used to improve our lives and make them safer. The private sector's data collection tends to be through voluntary means, and it is up to the government to regulate their use of it.

Chapter 5: Big Data for Small Businesses? Why not?

Small businesses sometimes lag behind large companies when it comes to cutting-edge technologies, mainly because they simply cannot afford it. When it comes to big data, many large corporations have embraced it, and this has led people to believe that this is one of the cutting-edge technologies that small business will have trouble adapting. The truth is, to employ data analytics; there is no need for complex systems that are resource hogs. Small business may employ big data, as the most important components are based on developing human resources: how skillful the data analysts are, how much attention to detail they are, and how well-organized they are. Any company that has the capability to gather relevant data and analyze it critically will be able to create and seize new opportunities. Big data may be used by small businesses to improve the quality of their products and services, better tailor their marketing strategies, and even foster better customer relations. These are all impactful on a businesses' bottom line, and the use of big data to achieve this does not have to be prohibitively expensive.

The Cost Effectiveness of Data Analytics

Why is cost an issue, anyway? The reason that costs are always accounted for before a vital business decision is made is that the cost has a direct effect on the decision's profit potential, and therefore viability. If a small business increases their net revenue by 30% through new techniques, but their costs have gone up by 35%, the net profit decreases, and the innovations introduced will have turned out to be detrimental. However, it will spell doom for a business if they do not innovate, while their competitors adapt new techniques and technologies. Most companies have begun to use data analytics, and it would be folly to ignore the opportunity to make use of this field as well. As said earlier, the use of big data helps create opportunities, and oftentimes these opportunities reduce costs and increase revenues. Small businesses should consider using it, as they grow much as large corporations do: by taking advantage of new business opportunities.

Big Data can be for Small Businesses Too

There are certain traditions that come with running a business, no matter the size. Oftentimes, the inertia proves too much for a new technology to effectively gain a foothold. Developers try to get around this by pushing their ideas on large companies, as these have deeper pockets and a larger profit margin, and can therefore afford to test new concepts. This is what happened when big data was first introduced. The software solutions marketed to businesses were based on the advantages of an economy of scale. Small businesses have problems with creating economies of scale, however, as they often struggle to build base capacity to begin with. This leads many to think that big data isn't for them, especially as many of the existing software solutions require a large capital outlay in order to properly operate. However, innovators have developed data solutions that are viable for small companies as well. These solutions allow small businesses the capability to use the appropriate tools for their operations. They can then begin to catch up with the bigger companies, as big data increases efficiency and productivity, allowing a company to expand its market share. Unless these smaller companies start to adopt this, however, larger companies will eventually run them out of town.

Where can Big Data improve the Cost Effectiveness of Small Businesses?

Social media is a new development that allows for greater and greater connectivity between people. Small businesses can use social media sites such as Facebook, Instagram, Twitter, and other similar websites in order to gauge consumer trends. This can lead businesses to get ahead of the curve and rapidly capitalize on emerging trends. A strong social media presence can also contribute to rapid client base growth, especially with effective use of the site's consumer analytics and ad platforms. This moves away from traditional marketing strategies, and requires implementation of different strategies, such as incorporating targeted ad strategies to display ads to those who would most likely buy your companies' product.

Another method would be the launch of online services. Any website can be setup so the administrator can analyze the customer's tendency to visit the site, as well as study their habits, such as their most frequented pages. These little details can assist a company in developing a better website designed to better appeal to their customer base. These are all cost-effective solutions when done properly, and can easily be as beneficial to small businesses as they are to large corporations.

What to consider when preparing for a New Big Data Solution

As a small business, a mass produced data solution will not properly serve your needs. A small business needs a data solution tailored for their specific needs, which will help gather relevant data, as well as help in its analysis. This solution should also be streamlined, in order to do away with superfluous functions and to reduce costs. A good data solution should also be able to take advantage of and be able to work alongside the pre-existing solutions of a business, as well as its systems. At the end of the day, a small business needs a solution that is integrated and conveniently packaged. A solution requiring a complete overhaul of the business' pre-existing system may not be the best choice, as it would be very costly and would take much time and resources in order to return to a "business as usual" state.

One way to address this is to bypass the massive integrated solutions, and go for the smaller modular data solutions. These data solutions could be developed for specific departments, as all they have to do would be to lay out their system requirements, justify their costs, and then be able to conduct their own research as to which solution best addresses their needs, rather than going for a costly data solution that will apply to the whole organization. This would allow a small business to make use of multiple solutions, each tailor-fit to the relevant department. This allows a department to increase its efficiency while minimizing interference with other departments, as well as keeping costs down.

Another thing to remember is to obtain a data solution that can be easily deployed. This also means that there should be a relative ease of use for the end users. A data solution that takes more than a few weeks to deploy, from testing stage to final integration, may not suit the needs of a small business. Small businesses cannot afford the loss in productivity that a prolonged period of downtime would bring. Small businesses will also have a hard time using a data solution that requires too much specialization, as they may not have the resources to train or hire personnel who can properly make use of the system. In addition, acquiring a new big data solution will be

pointless if it is too complicated to properly use, as the entire point of obtaining a data solution is to improve efficiency and effectiveness, and to create new opportunities to be taken advantage of. The best data solution would be one easy enough to use that a new user can learn it quickly, rather than having to undergo specific training just to employ it.

As a small business with limited resources, one of the most important things to consider is the overall cost. A versatile data solution would be best, as the use of its capabilities can be increased as the company begins to expand. The data solution offered should be designed such that a company only has to pay for the capabilities that they actually use, rather than paying for all the bells and whistles. However, as the company expands, a greater amount of capabilities of the data solution will need to be employed, and as such, there should be a licensing option to allow access to these greater capabilities as the company grows.

It is possible, and very practical to make the switch from a traditional intuition-based business to an analytics based business, even when the business in question is small. All that is needed is the proper identification of an IT solution that will suit the size and purpose of the business involved. Once this is done, the business will be poised to reap the benefits associated with the use of big data, the chance to create and take advantage of new opportunities.

Chapter 6: Important training for the management of big data

Just as teachers impart education and police uphold law and order, the systems of an organization direct its operating processes. In the same way that the various departments of government are upgraded, organizations also make changes to keep up with modern trends and technologies. New systems are important to ensure optimal output of the organization and they must be fully understood for effective management. Army personnel attend refresher courses to update their skills and companies hold conventions to introduce new remedies to medical practitioners. Therefore talented people are selected and trained to handle the management of big data. Management must be fully cognizant of the implications and impact new systems have on the overall functioning of operations.

From time to time, new systems replace or update processes in order to meet fresh demands. No matter how large or small an organization is, the ability to use big data has immense benefits. Supporting innovation insures increased efficiency and productivity while accelerating growth and reducing costs. When a system change arrives, former processes become out of date and handlers have to be trained to manage the new functions. Big data supports the distribution of information even when departments work independently of each other. Departments interact more efficiently due to the availability of swiftly shared data. Updated facts improve collaboration and help department heads make smart, informed, on the spot decisions.

A lot of importance used to be placed on the acquirement of data, but since the internet has made information so accessible, data is no longer protected. Important data is now only useful when it is gathered and arranged specifically, and classified. It can then be used where it is most pertinent. This kind of information sorting takes place without any rules or regulations

and depends on the sorter’s intuition and judgement. This creates a deficit in information and impedes people’s capacity to make best use of such incomplete data. It is vitally essential that only those individuals in an organization who are able to use big data productively are trained in handling this information. Some criteria must be formalized for training for information gathering and sorting. Those employees who handle big data may need to use Hadoop, a software library in open source form that helps distributed data processing. They should get maximum precedence when it comes to training.

The Human Resources department usually selects those employees who have an aptitude for training. Their qualifications, skills and motivation are checked. Department heads are consulted since they understand their people best and can make informed recommendations. These officers can identify the individuals that handle data well during heavy traffic and they know whether the training of these persons will benefit the organization. The head of the IT division make the final selection since they have frequent dealings with the shortlisted candidates and know their caliber. The head has to be thorough in his final assessment since he must choose only those candidates who will bring dynamism to the department, drive the new technology, and help employees to become more efficient.

Present level of skill in managing data

Data management is becoming increasingly important to organizations, and yet, surveys continue to reveal how below par operators are when dealing with information handling. It is noted in these reports that a large number of IT staff are deficient in data management and analysis skills. They all struggle with similar issues with big data, having long term storage issues, problems with data organization and management, knowledge of data management plans, and need for consultation and instruction. Most employees lack of information about data management facilities. Few departments have little or no representation. The survey showed that employees showed an interest in streamlined systems and IT staff indicated that they may not be getting the appropriate guidance and support and requested more training in their field. Findings point to the need for improved training for staff managing big data. A conscious effort must be made to raise the level of data collection, management, and distribution, so that the installation of super systems benefits the organization.

Where big data training is necessary

The IT department

Technology changes constantly. If it is incorporated in the workplace, it is necessary that staff handling this area remain current. This should be a matter of course since trained employees lead to the success of an organization and move it forward. The IT department in particular should receive all the training they can get since it is they who suffer the brunt of complaints from the rest of the organization when employees fail to manage their systems. It is members of the IT division that are beset by irate colleagues with unending complaints about the inefficiency of the department and system. When connections between departments stop working correctly, it is these poor individual who are held accountable for the glitches whether it is their fault or not.

Demands to restore and update systems stresses the IT department since it is generally the user who is at fault. The blame for under-functioning systems must lie with the organization and department heads, who must realize the problems and ensure all departments are kept up-to-date. Proper training, policies, and working practices should be in place to maximize return on systems. It is vitally important that the right persons are selected for training so that organizational efforts are not wasted. Well trained IT staff fully understand the value of their system and optimize its usage to benefit the organization’s productivity.

IT security is compromised when systems are affected by theft or damage to its hardware, software, or the data on it. The security of systems has become a major concern in spite of all the new guidelines on the correct handling of data. The IT department must be aware of any threat to the security of their system and ensure its protection from any attack. Hacking and breaching applications are a constant threat and the department must

know how to preempt such hazards. Organizations suffer significant losses due to the theft of data and some establishments have created security systems supported by legal measures to insure their data isn’t compromised.

The Product Development department

Training is essential to the product development staff since they not only create new products but also re-engineer existing merchandise. They have to be proactive and innovative and the whole department is involved in research and development which requires fresh, original thinking. This department must come up with new products that will be readily accepted by competitive markets, and the team must also interact between departments to get their feedback. This is a huge responsibility since the future of the company is directed by this department. Failure of a product can lead to the failure of the organization.

Product development staff must have a complete understanding of product development issues. Reengineering products and inventing new ones requires the support of and big data and this must be well analyzed, critically evaluated, and properly utilized. For this reason staff in this department must be trained well. They must be able to recognize and take advantage of the benefits big data provides, principally in areas of data integration across touch points in the course of product development such as design, software applications, manufacture, and quality control. Training will promotes the productive use of big data which will assist the generation of new ideas and solutions while creating a greater level of interdepartmental cooperation that will benefit the entire organization.

The Finance department

This department is almost as vital as the IT department, some would argue it is the most vital area of any organization. The finance department is all about the money. This is understandable because business is about making money and organizations would collapse without substantial profits. Anyone who cannot deliver the money will be considered a liability for the organization. Staff in this department must know whether there is financial value in deploying certain assets. Clearly they must be well trained in using big data in order to make important monetary decisions. With the influx of information, financial functions have become intricate making training in big data a necessity.

Employees in the finance department must be trained to make use of big data platforms in order to construct financial prototypes that will sustain the organization. Money will be depleted without controlled spending and maximum project funding, and the financial team must keep the money coming in. Big data technologies must be understood and harnessed to support the goals of the organization. Training prepares these employees to make use of big data in their central roles of business planning, auditing, accounting, and overall regulation of finances. When the work is managed successfully, the organization’s finances are significantly improved. Well trained staff generate accurate statements regarding cash flow, compliance with finance standards, cost modeling, and prize realization, etc.

The Human Resources department

In today's fast changing world, the way in which a human resources department operates goes a long way in defining the success of the organization. The human resources department must be able to analyze information in order of relevance in order to make strategic decisions regarding functioning. Training in big data brings skills like data analysis, visualization and problem solving right from operational reporting to strategic analytics. Skillful application of big data improves personnel quality by evaluating the capabilities of current workers and determining competencies required in future employment.

Employment is no longer the main concern of the human resources department. Its scope has widened and is now understands and analyze all matters concerning its department. Human resources staff members must be able to use available tools for data analysis because such capabilities will eventually help in resolving issues which comprise staff retention, staff-customer relations that affect sales, skill and deficiency within the organization, quality of candidates for employment, and a host of other matters related to the department and organization. Intuition, experience, and a sound data driven approach benefits the human resources department and in turn the whole organization as well.

The supply and logistics department

Business will come to a standstill if supplies are delivered late or not at all due to stock and logistical failures. Late deliveries and breakages in transit will cause irate customers to cancel their orders spreading complaints about poor services. The aims of this department are speed and agility, saving costs, and improving performance. They capture and track different forms of data to improve operational efficiency, improve customer experience, and create new and better business models. These factors help organizations to conserve resources, build a better brand name, and create a systematic process for their supply chain and logistics.

Logistics and supply staff must also have the benefit of training in big data in order to realize and employ tactics to attain departmental objectives. Planning and scheduling are perhaps the most vital part of any supply chain. So much money can be lost or expended in scheduling and planning, and with big data this process can be optimized. Big data can help predict demands so that money, space, and time are not wasted. Levels of inventory can be maintained and market trends observed. Logistics and supply departments have transaction based possesses that generate an enormous amount of data. Big data has applications at all levels of business and the supply and logistic department must appreciate its advantages.

The Operations department

Operations are the organized daily activities of business. Operations departments administer business practices to create the highest level of efficiency within an organization. Operations divisions ensure that companies can effectively meet customer requirements using the least amount of resources necessary, and manage customer support. Undertakings also include product creation, development, production and distribution. Operations play a distinctive role that contributes towards the overall success of the organization.

Organizations have been using data in various ways to improve their operations. With the emergence of new technologies, big data is becoming an inherent part of modern operations where challenges are overcome by using big data analytics. Concepts and applications help businesses predict product and customer behavior. Organizations realize the utility of big data brings value through continuous improvements in their operations departments.

The Marketing department

Numbers are critical in marketing whether it concerns the number of customers, the percentage of market shares, or sales revenue. There are many more figures and statistics in the marketing department and all these numbers keep the team busy. There is a huge amount of data generated by market activities. Some of it is positive and useful while other data can possibly damage your brand. The marketing department must collect data that is relevant, analyze this information, discard what is unimportant, and preserve details key details that can be utilized to benefit the brand.

Without the skills to handle big data, it is difficult to understand the massive amount of information on the internet, particularly in an age when digital marketing is part of the business process. Training in big data enables marketing staff to measure with relative accuracy the response advertisements have on the market, as well as the volume and impact of click-through-rate impressions, return on investment, and other factors that impact marketing. There are plenty of these statistics on social media and while they may look unimportant to the casual eye, this information is gold for the marketing department.

The large amount of data that is available through the various levels of customer activity, social media, as well as evidence generated during sales, is useful to trained marketing employees who are able to make good use of it. Platforms such as Hadoop are very useful for framing marketing indicators. Big data training also helps in retrospection so that marketing staff can judge the performance of their brand as compared with other competing products. Big data even has a function that allows marketing teams to check out competitor’s websites and do a bit of helpful text mining. All this information will be used to improve their own brand.

The Data Integrity, Integration and Data Warehouse department

It is important for organizations to have teams of experts to monitor and sort through the immense amount of data that collects on various company systems. This team must be well trained and current, and be aware of the possible dangers that may be linked with accessing this data. They must also be able to warehouse this data and make structured sense of it. These teams need to know the rules that apply to customer protection and privacy laws if they are to work with and interpret this data.

The Legal and Compliance department

It has become necessary for the legal department in an organization to be aware of existing privacy and retention procedures since new legal and compliance rules that cover data are enacted regularly. Companies that do not monitor and report their data can run into problems with the law and must have policies in place to safeguard themselves against possible litigation and breaches. These departments must work together to understand data privacy laws and make sure that their files are secure, warehoused, and handled correctly.

Chapter 7: Steps Taken in Data Analysis

In this chapter, we'll break down the process of data modeling into steps and look at each one separately, but before that, we'll be defining it.

Defining Data Analysis

We need to know exactly what data analysis is before we can understand the process. Analysis of data is the procedure of first of all setting goals as to what data you need and what questions you’re hoping it will answer, then collecting the information, then inspecting and interpreting the data, with the aim of sorting out the bits that are useful, in order to suggest conclusions and help with decision making by various users.

It focuses on knowledge discovery for predictive and descriptive purposes, sometimes discovering new trends, and sometimes to confirm or disprove existing ideas.

Actions Taken in the Data Analysis Process

Business intelligence requirements may be different for every business, but the majority of the underlined steps are similar for most:

Phase 1: Setting of Goals

This is the first step in the data modeling procedure. It's vital that understandable, simple, short, and measurable goals are defined before any data collection begins. These objectives might be set out in question format, for example, if your business is struggling to sell its products, some relevant questions may be, "Are we overpricing our goods?" and "How is the competition's product different to ours?"

Asking these kinds of questions at the outset is vital because your collection of data will depend on the type of questions you have. So, to answer your question, "How is the competition's product different to ours?" you will need to gather information from customers regarding what it is they prefer about the other company's product, and also launch an investigation into their product's specs. To answer your question, "Are we overpricing our goods?" you will have to gather data regarding your production costs, as well as details about the price of similar goods on the market.

As you can appreciate, the type of data you'll be collecting will differ hugely depending on what questions you need answered. Data analysis is a lengthy and sometimes costly procedure, so it's essential that you don't waste time and money by gathering data that isn't relevant. It's vital to ask the right questions so the data modeling team knows what information you need.

Phase 2: Clearly Setting Priorities for Measurement

Once your goals have been defined, your next step is to decide what it is you're going to be measuring, and what methods you'll use to measure it.

Determine What You're Going to be Measuring

At this point, you'll need to determine exactly what type of data you'll be needing in order to answer your questions. Let's say you want to answer the question, "How can we cut down on the number of people we employ without a reduction in the quality of our product?" The data you'll need will be along these lines: the number of people the business is currently employing; how much the business pays these employees each month; other benefits the employees receive that are a cost to the company, such as meals or transport; the amount of time these employees are currently spending on actually making the product; whether or not there are any redundant posts that have may have been taken over by technology or mechanization.

As soon as the data surrounding the main question has been obtained, you'll need to ask other, secondary, questions pertaining to the main one, such as, "Is every employee's potential being used to the maximum?" and "Are there perhaps ways to increase productivity?"

All the data that's gathered to answer the main questions and these secondary questions can be converted into useful information that will assist your company in its decision making. For instance, you may in the light of what is found decide to cut a few posts and replace some workers with machines.

Choose a Measurement Method

It’s vital that you choose the criteria that’ll be utilized in the measurement of the data you’re going to collect. The reason being that the way in which the data is collected will determine how it gets analyzed later.

You need to be asking how much time you want to take for the analysis project. You also need to know the units of measurement you’ll be using. For example, if you market your company’s product overseas, will your money measurements be in dollars or Japanese yen? In terms of the employee question we discussed earlier, you would, for example, need to decide if you’re going to take the employees’ bonuses or their safety equipment costs into the picture or not.

Phase 3: Data Gathering

The next phase of the data modeling procedure is the actual gathering of data. Now that you know your priorities and what it is that you're going to be measuring, it'll be much simpler to collect the information in an organized way.

There are a few things to bear in mind before gathering the data: Check if there already is any data available regarding the questions you have asked. There's no point in duplicating work if there already is a record of, say, the number of employees the company has. You will also need to find a way of combining all the information you have.

Perhaps you've decided to gather employee information by using a survey. Think very carefully about what questions you put onto the survey before sending it out. It's preferable not to send out lots of different surveys to your employees, but to gather all the necessary details the first time around. Also, decide if you're going to offer incentives for filling out the questionnaires to ensure you get the maximum amount of cooperation.

Data preparation involves gathering the data in, checking it for accuracy, and entering it into a computer to develop your database. You'll need to ensure that you set up a proper procedure for logging the data that's going to be coming in and for keeping tabs on it before you can do the actual analysis.

You might have data coming in from different places, such as from your survey, from employee interviews, or from observational studies, and perhaps from past records like payrolls.

Remember to screen the information for accuracy as soon as it comes in, *before* logging it. You may need to go back to some of the employees for clarification. For instance, some of the replies on the questionnaires may not be legible, or some may not be complete.

If you've gathered data to analyze if your product is overpriced, for instance, check that the dates have been included, as prices and spending habits tend to fluctuate seasonally.

Remember to ascertain what budget your company sets aside for data collection and analysis, as this will help you choose the most cost-efficient methods of collection to use. For example, if the budget is small, you may decide to use a free online census, or use social media, rather than printed questionnaires. If the budget for data collection is generous, however, you could arrange online competitions with prizes as incentives to encourage customers to give out information, or use colorful printed survey forms.

Phase 4: Data Scrubbing

Data scrubbing, or cleansing, is the process where you'll find, then amend or remove any incorrect or superfluous data. Some of the information that you've gathered may have been duplicated, it may be incomplete, or it may be redundant.

Because computers cannot reason as humans can, the data input needs to be of a high quality. For instance, a human will pick up that a zip code on a customer survey is incorrect by one digit, but a computer will not.

It helps to know the main sources of so called "dirty data". Poor data capture such as typos are one, lack of companywide standards, missing data, different departments within the company each having their own separate databases, and old systems containing obsolete data, are a few others.

There are data scrubbing software tools available, and if you're dealing with large amounts of incoming information, they can save your database administrator a lot of time. For instance, because data has come in from many different sources like surveys and interviews, there is often no consistent format. As an example, there needs to be a common unit of measurement in place such as feet or meters, dollars or yen.

The process involves identifying which data sources are not authoritative, measuring the quality of the data, checking for incompleteness or inconsistency, and cleaning up and formatting the data. The final stage in the process will be loading the cleaned information into the log or "data warehouse" as it's sometimes called.

It's vital that this process is done, as "junk data" will affect your decision making in the end. For instance, if half of your employees didn't respond to your survey, these figures need to be taken into account.

Finally, remember that data scrubbing is no substitute for getting good quality data in the first place.

Phase 5: Analysis of Data

Now that you have collected the data you need, it is time to analyze it. There are several methods you can use for this, for instance, data mining, business intelligence, data visualization, or exploratory data analysis. The latter is a way in which sets of information are analyzed to determine their distinct characteristics. In this way, the data can finally be used to test your original hypothesis.

Descriptive statistics is another method of analyzing your information. The data is examined to find what the major features are. An attempt is made to summarize the information that has been gathered. Under descriptive statistics, analysts will generally use some basic tools to help them make sense of what sometimes amounts to mountains of information. The mean, or average of a set of numbers can be used. This helps to determine the overall trend, and is easy and quick to calculate. It won't provide you with much accuracy when gauging the overall picture, though, so other tools are also used. Sample size determination, for instance. When you're measuring information that has been gathered from a large workforce, for example, you may not need to use the information from every single member to get an accurate idea.

Data visualization is when the information is presented in visual form, such as graphs, charts, and tables or pictures. The main reason for this is to communicate the information in an easily understandable manner. Even very complicated data can be simplified and understood by most people when represented visually. It also becomes easier to compare the data when it's in this format. For example, if you need to see how your product is performing compared to your competitor's product, all the information such as price, specs, how many were sold in the last year can be put into graph or picture form so that the data can be easily assessed and decisions made. You will quickly see that your prices are higher overall than those of the competition, and this will help you identify the source of the problem.

Basically, any method can be used, as long as it will help the researcher to examine the information that has been collected, with the goals in mind of

making some kind of sense out of it, to look for patterns and relationships, and help answer your original questions.

The data analysis part of the overall process is very labor intensive. Statistics need to be compared and contrasted, looking for similarities and differences. Different researchers prefer different methods. Some prefer to use software as the main way of analyzing the data, while others use software merely as a tool to organize and manage the information.

There is a great deal of data analysis software on the market, among the currently most popular are Minitab, Stata, and Visio. Of course, Excel is always useful too.

Phase 6: Result Interpretation

Once the data has been sorted and analyzed, it can be interpreted. You will now be able to see if what has been collected is helpful in answering your original question. Does it help you with any objections that may have been raised initially? Are any of the results limiting, or inconclusive? If this is the case, you may have to conduct further research. Have any new questions been revealed that weren't obvious before?

If all your questions are dealt with by the data currently available, then your research can be considered complete and the data final. It may now be utilized for the purpose for which it was gathered- to help you make good decisions.

Interpret the Data Precisely

It is of paramount importance that the data you have gathered is meticulously and carefully interpreted. It's extremely vital that your company has access to experts who can give you the correct results.

For instance, perhaps your business needs to interpret data from social media such as Twitter and Instagram. An untrained person will not be able to correctly analyze the significance of all the communication regarding your product that happens on these sites. It is for this reason that most businesses nowadays have a social media manager to deal with such information. These managers know how the social platforms function, the demographic that uses them, and they know how to portray your company in a good light on them as well as extract data from the users.

For every company to be successful, it needs people who can analyze incoming data correctly. The amount of information available today is bigger than it has ever been, so companies need to employ professionals to help stay on top of it all. This is particularly true if the founders of a company don't have much knowledge of data. It would then be a great idea to bring an analyst onto the team early. There is so much strategic information to be found in the data that a company accumulates. An analyst can help you decide what parts of the information to focus on, show you where you are losing customers, or suggest how to improve your product. They will be able to suggest to management which parts of the data need to be looked at for decisions to be made.

For instance, a trained data analyst will be able to see that a customer initially "liked" your product on Facebook. He then googled your product and found out more about it. He then ordered it online and gave a positive review on your website. The analyst can trace this pattern and see how many other customers do the same. This information can then perhaps help your business with advertising, or with expansion into other markets. For instance, the analyst can collect data regarding whether putting graphics with "tweets" increases interest, and can tell what age group it appeals to

more. They’ll be able to tell you what marketing techniques work best on the different platforms.

It is hoped that from this you can see how vital data collection and analysis are for the well-being of your company, and how it can help in all departments of your business, from customer care, to employee relations, to product manufacture and marketing.

Chapter 8: Descriptive Analytics

Descriptive Analytics- What is It?

Businesses use descriptive analytics all the time, whether they are aware of it or not. It's often called *business intelligence*. Companies these days have vast amounts of data available to them, so they would do well to use analytics to interpret this data to help them with decision making. It helps them to learn from what happened in the past, and enables them to try to accurately predict what may happen in the future. In other words, analytics helps companies anticipate trends. For instance, if sales increased in November for the past five years, and declined in January, after the Christmas rush, then one could predict that the same thing is likely to happen in year six and prepare for it. Companies could use this to perhaps increase their marketing in January, offering special offers and other incentives to customers.

Descriptive analytics give insight into what happened in the past (this may include the distant past, or the recent past, like sales figures for last week.) They summarize data that describes these past events and make it simple for people to understand. In a business, for example, we may look at how sales trends have changed from year to year, or how production costs have escalated.

Descriptive statistics, basically then, is the name given to the analysis of data that helps to show trends and patterns that emerge from collected information. They are a way of describing the data in a way that helps us to visualize what the data shows.

This is especially helpful where there has been a lot of information collected. Descriptive statistics are the actual numbers that are used to describe the information that has been collected from, say, a survey.

Let's say, for instance, that we issued proficiency tests to all 200 of our company employees. From their results, we could work out the mean and the standard deviation. The group of data, which includes all the information of interest to management, is called a *population*. It may be big or small, provided that it includes all the information we're interested in. In our example, we're examining 200 employees, so they are our population.

The properties of the population, such as the mean test result, and the median, are called *parameters*.

Perhaps we're actually interested in the proficiency test results of everyone employed in the same sector of the industry across the world. It's not really possible to obtain such data, so we'd have to use a *sample* of our 200 employees to represent the entire industry. The properties of this sample would then not be called *parameters*, but rather, *statistics*.

By using what is called inferential statistics, we can use the sample to infer things about the entire group of employees across the world. The sample must be an accurate representation of the whole group for this to work. Obviously, mistakes will be made as a sample never perfectly represents the population.

How Can Descriptive Analysis Be Used?

Descriptive analysts are able to make information easier to understand and therefore to use. They do this by turning it into graphs, charts, or pictorial representations of what has been happening. This way, management and employees alike can see what has been happening within the company in the past, and make useful predictions and therefore good decisions for the future.

External data related to the company may also be used, such as stock market trends, or international events that may affect this particular business- for instance, an oil crisis in OPEC at the end of last year will have an indirect but tangible effect on a trucking transport company in the US early the following year. The company can use this information to perhaps stockpile fuel, or take on less labor in the new year.

Based on probabilities, the company takes existing data and fills in what's missing with educated guesses. The historical data will be combined, patterns identified, and then algorithms applied to find the relationships between the sets of data. The company may then be able to predict customer behavior such as spending patterns. This will then help them ensure that the supply chain can keep up with the demand.

Measures in Descriptive Statistics

Descriptive statistics are so-called because they help to *describe* the data which has been collected. They are a way of summarizing big groups of numerical information, by summarizing a sample. (In this way, descriptive statistics are different from inferential statistics, which uses data to find out about the population that the data is supposed to represent.)

Two types of statistics are most often used in this descriptive process, and these are *measures of central tendency*, and *measures of dispersion*.

Central tendency involves the idea that there's one figure that's in a way *central* to the whole set of figures. Measures of central tendency include the mean, the median, and the mode. They summarize a whole lot of figures with one single number. This will obviously make it easier for people to understand and use the data.

The *mean* is the average of all the figures. In other words, the sum of them all divided by how many there are. If the distribution of measurements is random, the mean represents the value that's expected.

The *median* is the figure in the middle once all the figures have been put into numerical order. In other words, half of the numbers we have will be smaller than this median number, and half will be bigger.

The *mode* is the figure that occurs most often. It's the most frequently appearing number in the number set.

Dispersion refers to how spread out the figures are from a central number. **Measures of dispersion** include the range, the variance, and the standard deviation. They help us see how the scores are spread out, and if they are close together or widely spread apart.

The *range* is the spread of figures from the lowest one to the highest. In other words, it describes the distance from the lowest to the highest score.

The *variance* measures how far our set of random numbers are spread out from the mean. Are they clustered together around this point, or are there wide variations in value? It's calculated by first subtracting the mean from

each of the numbers in our set. (This shows the distance of that figure from the mean.) Each of these answers are then squared, the sum of the squares found, then divided by how many numbers there are in the set.

The *standard deviation* refers to the average shift of scores from the mean. To find it, we'd measure how far all our figures are from the mean, then square each of them. We'd then find the sum, thus getting a result known as the variance. When we then calculate the square root of the variance, we will have the standard deviation.

Inferential Statistics

When research is done on groups of people, usually both descriptive and inferential statistics are used to analyze results and arrive at conclusions.

Inferential statistics are useful when we just want to use a small sample group to *infer* things about a larger population. So, we are trying to come to conclusions that reach past the immediate data that we actually have on hand.

They can help to assess the impact that various inputs may have on our objectives. For instance, if we introduce a bonus system for our workers, what might the productivity outcome be?

Inferential statistics can only be used if we have a complete list of the population members from which we have taken our sample. Also, the sample needs to be big enough.

There are different types of inferential statistics, some of which are fairly easy to interpret. An example is the confidence level. If say, our confidence interval is 98%, it means that we are 98% confident that we can infer the score of a population based on the score of our sample.

Inferential statistics therefore allow us to apply the conclusions from small experimental studies to larger populations that have actually never been tested experimentally. This means then, that inferential statistics can only speak in terms of probability, but it is very reliable probability, and an estimate with a certain measurable confidence level.

Chapter 9: Predictive Analytics

We are now aware of how data and the analysis of it are vital for a company to be able to function optimally. We'll now examine another branch of data mining that can help grow a company. In this chapter, we'll be taking a look at predictive analytics, starting with what it is and how it can help a company function more efficiently.

Defining Predictive Analytics

Predictive analytics is used to make predictions about unknown future events. It uses many techniques, such as statistical algorithms, data mining, statistics, modeling, machine learning and artificial intelligence, to analyze current data and make predictions about the future. It aims to identify the likelihood of future outcomes based on the available historical data. The goal is therefore to go beyond what *has* happened to provide the best assessment of what *will* happen.

More and more companies are beginning to use predictive analytics to gain an advantage over their competitors. As economic conditions worsen, it provides a way of getting an edge. Predictive analysis has become more accessible today for smaller companies, healthcare facilities, or smaller, low-budget organizations. This is because the volume of easily available data has grown hugely, computing has become more powerful and more affordable, and the software has become simpler to use. Therefore, one doesn’t need to be a mathematician to be able to take advantage of the available technologies.

Predictive analysis is extremely useful for a number of reasons: Firstly, it can help predict fraud and other criminal activity in places from businesses to government departments. Secondly, it can help companies optimize marketing by monitoring the responses and buying trends of customers. Thirdly, it can help businesses and organizations improve their way of managing resources by predicting busy times and ensuring that stock and staff are available during those times. For instance, hospitals can predict when their busy times of the year are likely to be, and ensure there will be enough doctors and medicines available over that time.

Thus, overall efficiency can be increased for whatever organization utilizes the data effectively.

Different Kinds of Predictive Analytics

Predictive analytics can also be called predictive modeling. Basically, it is a way of matching data with predictive models and defining a likely outcome. Let’s examine three models of predictive analytics:

Predictive Models

Predictive models are representations of the relationship between how a member of a sample performs and some of the known characteristics of the sample. The aim is to assess how likely a similar member from another sample is to behave in the same manner.

This model is used a lot in marketing. It helps identify implied patterns which indicate customers' preferences. This model can even perform calculations at the exact time that a customer performs a transaction.

The model can be used at crime scenes to predict who the suspects may be, based on data collected at the scene. Data can even be collected without investigators having to tamper with the crime scene, by use of, for example, 3D laser scanners, which make crime scene reconstruction easier and faster. The investigators don't even have to be at the actual scene, but can examine it from their office or home. Nothing at the actual crime scene is disturbed, and all the images and other data are stored for future reference. The scene can later be reconstructed in a courtroom as evidence.

Descriptive Modeling

Descriptive modeling describes events and the relationship between the factors that have caused them. It’s used by organizations to help them target their marketing and advertising attempts.

In descriptive modeling, customers are grouped according to several factors, for example, their purchasing behavior. Statistics then show where the groups are similar or different. Attention is then focused on the most active customers. Customers are actually given a “value”, based on how much they use the products or services and on their buying patterns. Descriptive modeling finds ways to take advantage of factors that drive customers to purchase.

It’s worth bearing in mind that while descriptive modeling can help a business to understand its customers, it still needs predictive modeling to help bring about desired results.

Decision Modeling

The rapidly growing popularity of decision models has enjoyed much attention recently. Modeling combines huge quantities of data and complex algorithms to improve corporate performance. Decision models can be extremely useful, helping managers to make accurate predictions and guiding them through difficult decisions, unhindered by bias and human judgement alone. By using models, data can be used more objectively.

Managers need to be able to use data to make decisions. A decision-centered approach is quickly becoming the central analytics focus for most businesses. The ability to model and find solutions for complex issues allows for better decision making.

Decision models can help with such problems as to how to optimize online advertising on websites, how to build a portfolio of stocks to get maximum profit with minimum risk, or how online retailers can deliver products to customers more cheaply and quickly. It can also help product developers decide which new products to develop in the light of market trends.

When at the decision-making stage, it’s best to focus on action-oriented decisions that are repeatable. Decisions should be based solidly on the available data and be non-trivial, and should also have a measurable business impact.

A good model will have well-defined questions and possible answers. It will be able to be easily shared among team members.

While managers should use models to the maximum, they should also bear their limitations in mind. For instance, models can predict how many days of sunshine and rain a farming operation in a certain area may receive that year, but they cannot actually influence the weather. It may predict the likely amount that customers may spend on a product in a given year, but it will never be able to directly control their spending habits.

Chapter 10: Predictive Analysis Methods

In this chapter, we're going to examine the various techniques that are used to conduct predictive analysis. The two main pigeonholes into which these methods may be grouped are *machine learning techniques* and *regression techniques*. They'll be discussed here in more detail:

Machine Learning Techniques

Machine learning is a method of data analysis that automates the building of analytical models. It uses algorithms that continuously adapt and learn from data and from previous computations, thereby allowing them to find information without having to be directly programmed where to search.

Growing volumes of available data, together with cheaper and more powerful computational processing have created an unprecedented interest in the use of machine learning. More affordable data storage has also increased its use.

When it comes to modeling, humans can maybe make a couple of models a week, but machine learning can create thousands in the same amount of time.

Using this technique, after you make a purchase, online retailers can send you offers almost instantaneously for other products that may be of interest to you. Banks can give answers regarding your loan requests almost at once. Insurance companies can deal with your claims as soon as you submit them. These actions are all driven by machine learning algorithms, as are more common everyday activities such as web search results and email spam filtering.

Regression Techniques

These techniques form the basis of predictive analytics. They seek to create a mathematical equation, which will serve as a model to represent the interactions among the different variables. Depending on the circumstances, different regression techniques can be used for performing predictive analysis. It can be difficult to select the right one from the vast array available. It’s important to pick the most suitable one based on the type of independent and dependent variables, and depending on the characteristics of the available data.

Linear Regression

Linear regression is the most well-known modeling approach. It calculates the relationship between the dependent and independent variables using a straight line (regression line). It's normally in equation form. Linear regression can be used where the dependent variable has an unlimited range.

If the dependent variable is discrete, another type of model will have to be used. Discrete, or qualitative, choice models are mainly used in economics. They are models which describe, and predict choices between different alternatives. For example, whether to export goods to China or not; whether to use shipping or air travel to export goods. Unlike other models, which examine, "how much", qualitative choice models look at "which one?"

The techniques *logistic regression* and *probit regression* may be used for analyzing discrete choice.

Logistic Regression

Logistic regression is another much-used modeling approach. It’s used to calculate the probability of event success and failure. It’s mainly used for classification problems, and needs large sample sizes.

The Probit Model

The word *probit* is formed from the roots *probability* and *unit*. It's a kind of regression where the dependent variable can only have two values, for instance, employed or unemployed. Its purpose is to appraise the likelihood that a certain observation will fall into one or other of the categories. In other words, it is used to model binary outcome variables, such as whether or not a candidate will win an election. Here, the outcome variable is binary: zero or one, win or lose. The predictor variables may, for example, include how much money was spent on the campaign, or how much time the candidate spent campaigning.

Probit regression is used a great deal in the field of economics.

Neural Networks

Neural networks are powerful data models capable of capturing and representing complicated input/ output relationships. They are widely used in medical and psychological contexts, as well as in the financial and engineering worlds. Neural networks are normally used when one is not aware of the exact relationship between the inputs and the output. These networks are capable of learning the underlying relationship through training. (This may also be called supervised training, unsupervised training, and reinforcement learning.)

Neural networks are based on the performance of “intelligent” functions similar to those performed by the human brain. They’re similar in that, like the brain amasses knowledge through learning, a neural network stores knowledge (data) inside inter-neuron connections called synaptic weights.

Their advantage is that they can model both linear and non-linear relationships, whereas other modeling techniques are better with just linear relationships.

An example of their use would be in an optical character recognition application. The document is scanned, saved as an image, and broken down into single characters. It’s then translated from image format into binary format, with each 0 and 1 representing a pixel of the single character. The binary data is then fed into a neural network that can make the association between the image data and the corresponding numerical value. The output from the neural network is then translated into text and stored as a file.

Radial Basis Function Networks

Radial basis function networks are artificial neural networks that use radial basis functions as activation functions. Radial functions are a class of functions that can be used in any type of model, whether linear or non-linear, and in any type of network, whether single or multi-layer. However, they are usually associated with radial functions in single-layer networks.

They’re used in many applications, for example, time series prediction and function control.

Support Vector Machines

Support vector machines are mainly a classifier method that carries out classification tasks by the construction of hyperplanes in multidimensional space. Support vector machines are based on the concept of decision planes that define decision boundaries. They are used to detect and use elaborate patterns in data. This is done by grouping, classifying, and ranking the data.

Support vector machines can support both classification and regression tasks and can cope with multiple variables. In order to construct the best hyperplane, they use interactive training algorithms.

Naive Bayes

This technique is based on the Bayes' conditional probability rule. It is a simple to use and interpret technique that is used to classify various tasks. This technique assumes that the predictors are statistically independent, thus making it a classification technique. It's a good method to use where there are a high number of predictors.

Instance-Based Learning

Instance-based learning, or the K-nearest neighbor (k-NN) algorithm, is one of the pattern recognition statistical methods. It's a method used for classification and regression in pattern recognition. It helps with looking up and matching new patterns during prediction.

Instance-based learning is not a new method, and is sometimes known as case-based learning, lazy learning, or non-parametric learning, depending on the application.

It works best with a small number of input variables, but not so well with a large number of inputs. K-nearest neighbor stores the whole training data set and uses it as its representation. It does not learn any model. K-NN is often used in search applications when looking for items that are similar. In other words, when your search involves the command, "find elements similar to this". This is known as a *k-NN search*.

Geospatial Predictive Modeling

Geospatial predictive modeling analyzes historical and present-day events through a geographic filter, so that future events may be predicted. Geospatial modeling provides an understanding of how events interact because of geographic proximity and common geographic indicators. It finds trends and patterns, which are finally presented in an easily accessible visual way.

The main idea behind this method is that events being modeled have a limited distribution, and are neither randomly nor uniformly distributed, and that events occur depending on their location.

Geospatial models are used in various applications. For instance, in the prediction of wildfires, and for natural resource management.

Hitachi's Predictive Analytic Model

This is a predictive crime analytics platform which, if successful, will enable us to basically predict the future, specifically in the application of crime. It's part of Hitachi's existing Hitachi Visualization Platform. The goal is to provide real time insights to enhance police investigative capabilities when a crime occurs, and even to prevent it from occurring in the first place.

At the moment, police basically operate using damage control after an event such as a burglary or murder has already happened. The system uses machine learning, utilizing social media and public and private data feeds to get information. It uses a variety of data sets: police station and streetlight locations, parole registers, gunshot events, historical crime stats, and, as mentioned, social media. It uses natural language processing to search for words that may be significant on social media.

The data sets are fed into the system and, over a two-week period, the system will pick up whether there is any interconnection among the sets of data.

Hitachi plan to test the system by making it available to various law enforcement agencies in unknown locations. The system will run in the background. At the end of the testing period Hitachi will analyze the results of the predictions and compare them with the actual daily crime incidents that occurred over the same time period.

If Hitachi's system is successful, the benefits would be enormous. Police departments would know where to deploy officers before crimes occurred. Officers would be less at risk, and incidents of robbery, rape, and domestic violence would hopefully decrease.

Predictive Analytics in the Insurance Industry

Many insurance companies are now using predictive analysis in their day to day business practices. For example, almost half of personal auto insurance companies use predictive analytics to help increase profit, reduce risk, and increase revenue growth.

Predictive analytics are changing the way insurance companies interact with their customers. They will now try to give their customers the product they want at right price, and there is increased sensitivity to what the clients' main issues actually are. Because of an increase in competition, insurance companies are now having to look after their customers better.

Here's how analytics helps them do this: the company builds a risk management profile of the client using predictive analytics. For example, the chances of the person being involved in an accident, the chances of their vehicle being stolen based on the model of vehicle and where the person lives are all taken into account. This information is compared to information from many other profiles, and an accurate assessment is made. An affordable premium package can then be put together for that specific client.

Once the client has purchased the package, the claim process can be made faster by using analytics as well. The paperwork can be processed faster, and damage assessed easier by uploading 3D images of the vehicle. This fact of the customer's claim being able to be processed quicker will then be able to be used in marketing pitches later on, enabling the company to further expand its existing client base.

As you can hopefully see from this chapter, predictive analytics are already playing a major role in many industries, and will continue to play a bigger and bigger role in our lives as technology improves.

Chapter 11: R - The Future In Data Analysis Software

Technology is essential when it comes to analyzing data, so much so that it is standard to use computers for data analysis. To have a computer is a great start but it is no use without being coupled with the appropriate software. Not only do you need special software but you also have to be able to know how to use it in order to be able to carry out great data analysis. The most widely used piece of data analysis software is R, it is one of the most powerful software programs available for data analysis. R is not only entirely free to use but it is also open-source meaning that the code can be changed by programmers if they feel it is appropriate. R operates differently to many of the software programs we use on a regular basis, this is because R is command-line oriented meaning it requires lines of complex code to be entered in order to instruct and control the software to complete the jobs. This software requires the user to have knowledge in the field before they can use it to its full potential, Microsoft Excel seems like child's play compared to R which is a much more powerful software. R is a programming language, meaning it is essential that you understand your data before you can write code to analyze it. It takes time to learn the complete syntax and commands of the R software but once you had got to grips with it you will find your data analysis is a great deal more thorough than previous data analysis. It takes a few simple steps to have R downloaded and installed onto a computer, however it is not currently available for tablets or smartphones. Downloading and installing is straightforward, it is learning how to use the software that will take some time but once you've mastered it you'll notice the difference.

When you open R, it will have an empty screen (excluding the about information, etc.) with a command prompt. This means the software is ready for instructions. It begins with the command prompt being introduced by the > sign, and then you enter the commands needed followed by the enter key and the code will be sent to R for processing. If the instructions are typed correctly you will receive the result of the entered instructions in

the next line. You can test it by typing simple instructions such as 2+2 and when you press enter R will respond with 4 in answer to your request. This is a very basic example, R is capable of analyzing extremely complex data and it has a lot of potential, especially regarding big data analysis.

Is R A Good Choice?

R combines data analysis with a programming language, unlike many of the currently available software which are applicative software. At first glance there may seem to be little difference between R and other online analytical software but when you look closer there is a significant distinctness between them. With the standard data analysis package you will be able to perform only certain fixed actions whereas when you are using programming language such as R you can continuously specify new tasks to complete. This means the sky's the limit and you are able to process your data and write a code in the way you want to. This is ideal for an experienced programmer who has knowledge of coding but there are also packages available for the less experienced programmer that enables pre-determined analysis of data without the need for writing complex code. These packages allow you to analyze data, make adjustments and run various possibilities. For programmers, R is a very enjoyable piece of software as it allows you to use your knowledge and skills. Especially due to it being an open-source programming language so you can alter and even reprogram it all together. It is complex but it is doable and many companies do make adjustments to R based on their needs.

Types of Data Analysis Available with R

R allows you to run any type of data analysis, you can write code to analyze data in a unique way if that is what you want to do. The openness and opportunities of R are what make it so brilliant. If you have the knowledge to do it, you can write whatever code you want. For the common data analysis there are packages available that allow you to do it without needing to write the code yourself, this is great for saving time rather than coding something that is already available. You can run very simple functions using R such as basic addition or subtraction, you can also find statistical values such as median or mean just as easily. Using R you will be able to easily access these values, regardless of the amount of data you have it will be straightforward and quick to access the information, this is essential for any statistical analysis. You can build up from the basic data to more complex data analysis (R has been used to complete some of the most complex data analysis). The process can be simplified by running many functions at once to crunch the data in ways that are significant and detailed in order to give you a large amount of information in a very short amount of time.

Is There Other Programming Language Available?

R is not the only available programming language, there are others such as Python, that can also be used for data analysis.

The majority of the other available languages are easier to learn meaning you can take advantage of them more quickly, however this comes with less flexibility which is an area that R is so renowned in. The time that goes into learning to use R pays off with the flexibility of statistical analysis, meaning R is, for many, the best analytical programming language available.

Chapter 12: Predictive Analytics & Who Uses It

Predictive analytics has continued to grow in importance over recent time and there is good reason for this. The positive impact of predictive analysis has not been limited to one field or applications, it has been beneficial to many areas. The main applications the use predictive analytics are listed and discussed in this chapter.

Analytical Customer Relationship Management (CRM)

Customer Relationship Management is one of the most popular applications that use predictive analytics. Varying methods are used on customer data in order to achieve CRM goals. The whole idea of CRM is to give a complete view of the customer, unconcerned with where the information about the customer lies.

Further uses of predictive analytics include marketing, sales and customer service. As there are different methods that can be used it helps in many different areas over a large customer base, allowing you to ensure their satisfaction with ease thanks to predictive analytics. You can use predictive analytics in many areas when it comes to CRM, these include; analysis of product demand - identifying and analysing the products that are in the most demand and the products that are in increasing demand, looking at current buying trends in order to predict future buying habits, analysing areas of customer dissatisfaction or loss in order to make improvements. Such analysis can help in improving the company and increasing product promotion for that company. This type of analysis can be used throughout the customer lifecycle, analysing from the very beginning through relationship growth, retention and win-back. Such detailed analysis gives a holistic and beneficial overview for companies.

The Use Of Predictive Analytics In Healthcare

The health care district use predictive analytics a lot to help calculate disease or disorder contraction risks. Using such data analytics can assist in determining a patient's' risk of developing health conditions, these conditions can be anything from asthma to heart disease. Although diagnostics is not the only use of predictive analytics when it comes to clinical use, doctors also use it for making decisions regarding some medical care patients. This works by connecting the information regarding the health of the patient with the information about health. When you combine the information it gives further, clear details to clinicians in order to aid making decisions to best benefit the health of the patient. These are not the only ways that predictive analysis is aiding healthcare, the revolutionary influence that predictive analytics can have is incredible. Healthcare can be positively impacted due to; predictions on insurance product cost for hospitals and employers, the ability to produce predictions without needing to study endless patient cases, the influence of medicine development - helping to develop the best and most effective medicines, overall providing better outcomes healthwise for patients. Another fantastic thing with this is that the models will increase in accuracy over time.

Predictive analytics could literally change the industry of healthcare, it has the ability to greatly improve accuracy resulting in lives saved and less medical expenses for patients. There would be less cases of malpractice, and there is great potential for a decrease in healthcare costs. Using data analysis doctors can be more accurate with diagnosis, this has the potential to save huge amounts of money for many individuals and insurance companies. If a doctor is able to correctly identify and diagnose a sickness first time this would have a large impact on every other aspect of healthcare, as a result decreasing healthcare costs overall. With such a domino effect in place, more people will be able to have healthcare insurance and doctors can charge less without the worry of being sued for malpractice. Taking away a patient's unnecessary spending on prescription medication that does not actually help their illness would save incredible

amounts of money. The whole system will become more accurate, streamlined and improved as a result of using predictive analytics.

It is all too often that patients are unable to afford medication or insurance at the moment but this could be changed by incorporating analysis of data. The medicines could be manufactured to better suit the public's needs, and the amount of drugs that currently exist could be reduced to only leave only the necessary and effective. The pharmaceutical companies would have the ability to produce medication to deal directly with the patients' conditions, the analytics can find accurate data regarding the patient's' health issues and this can be used by the companies. Done on a large scale, this would wipe out unnecessary and ineffective medicine causing their eventual removal from the market. This would leave only the required medications. If a patient comes in and they have a history of heart attacks in the family making them more prone to them, and the patient is around the same age range of the family members who have suffered heart attacks. The doctor would be able to use predictive analytics in order to anticipate when a heart attack is most likely to occur, this knowledge can be shared with the patient and a plan can be put in place to reduce or even eliminate the risk of the heart attack. This information could help extend the life of a patient as a result of the analysis and prediction. At the very least the current moment has been positively impacted. These changes are truly life changing and can be brought about by predictive analytics. The health data that is provided by predictive analytics would most likely alter the doctor and patient roles as a patient would be able to be more aware of options and therefore make better, more informed decisions for their own benefit. A doctor would be able to provide advice and assist the patient in deciding the right course of action for them. More options are available and more patients are aware of issues regarding their own health. As a result, we are seeing more patients being involved in the decision making and this really personalizes the whole healthcare system. Predictive analytics is innovative and will positively alter the healthcare industry thanks to the clinical decision support system.

The Use Of Predictive Analytics In The Financial Sector

Many industries have payment risks where their customers do not make payments on schedule. For example in the back and financial sector, in such situations the institutions bring in collection teams who have the task of recovering the funds from the customers in question. Out of these customers, some of them will not pay the money back, even with the assistance of collection teams. In these cases, the collection teams and financial departments have wasted time and effort, could this be prevented? Using predictive analytics could provide some beneficial factors in this respect as they can optimize the collection efforts. The allocation of resources that are being used for collection could be optimized, the agencies can be analysed in order to identify the most effective collection companies, collection strategies can be formulated, the customers who have failed to pay back and now require legal action can be quickly and easily identified and the collection strategies can be adapted to suit individual situations. The use of predictive analytics causes a concentration and simplification of efforts, making them more effective and efficient. Collections can be made with little stress or risk and the collection costs will also be reduced for the financial companies.

Predictive Analytics & Business

If an organization is selling more than one product it can use predictive analytics in order to promote their products. The customer base details are essential and the use of this information can greatly benefit a business. Available predictions include; determining a customer's spend capacity and a customer's usage and purchasing behaviour. The analysis of these areas results in the ability for company to build up the relationship they have with their customers, adapting their current model to suit their customers and therefore improving their business and profits.

Keeping Customers Happy

For an organization to work really well long term they need two things: satisfied customers and loyal customers. The competition between businesses has continues to increase meaning these two objectives continuously gain importance. A third factor has begun to emerge as important in recent years as a result of the large amount of competition, this is customer attrition. If new customers purchase a product from an organization it is not so important, the more important factor is if existing customers return, as when you hold on to customers and they continue to buy your products you can increase your profits with very little effort. Due to this, the customers must be satisfied and the existing customers needs must be met. A lot of businesses tend to react rather than prevent and be proactive when it comes to the needs of their customers. This method could easily lose them customers as their competition could be more understanding of the needs of the customer, causing the customer to go to them instead. If this occurs, you cannot change that customers mind. In order for an organization to hold on to a customer they have to put a lot of time and money into it. However, the use of predictive analytics means organizations can have a more proactive position when it comes to customer service rather than being reactive. The analysis works by taking the customer's spending habits and service usage history and using this to create predictive models that can identify which of the customers are most likely to stop using the organization. The organization can use this predictive information in order to act in a proactive manner and figure out why there is a high chance of their service being terminated. This is likely to decrease how many customers do walk away from the organization overall. The kind of things a business can do when they are at risk of losing customers is to offer them deals and discounts. Another way customers leave organizations is by slowly reducing the amount that they are purchasing until they stop purchasing altogether. This happens over a period of time which means they often go unnoticed by the company until it is too late and they have stopped buying. Using predictive analytics an organization can monitor a customer's habits and identify the customers

who are behaving in the way previously described. Once these customers have been picked up on it is possible to strategize in a way that may cause the customer to rethink and remain with the organization.

There is a lot of “churning” in businesses, this means when a customer discontinues buying at one store in order to buy at another retailer. Using predictive analytics can prevent that from happening. Here are some examples:

The Windsor Circle’s Predictive Analytical Suite produces predictive data in order to assist email marketing campaigns. They can predict order dates based on the buying habits of customers, be able to predict when the stock will need to be replenished, recommend products tailored to the buying history of the customer and recommend products that are typically bought together. Predictive analytics can be used to combine all of this information in order to produce product recommendations that would hopefully keep a customer buying from that business. The customers can be enticed using price reductions and predicting the sell out items. The business will be able to use the analysis provided by the analytics company in order to create effective models that will keep their customer base interested and buying from them. The business can use the information to ensure customers get what they need and return to them time and time again.

Another angle the predictive models take is being able to anticipate when a customer will purchase certain items together and which items that would include. The business will also be able to use the analysis in order to determine how long to keep products available based on the past customer demand. Windsor Suite provides predictive data for Coffee for Less.com for when they send out reminder emails to ensure customers don’t run out of coffee. Windsor uses predictive analytics to predict the date a customer who has made more than three purchases in the past will order again. The replenishment email will then be sent to the customer based on the buying patterns of that specific customer. The email would include a product image based on the product the email is about and the rest of the email would include products other customers like. The email is created using predicted

products that the customer will be most interested in based on their history. There are two strategies in play here in order for Coffee for Less to retain their existing customers. The strategies are recommendations and predicted ordering, and both of these stem from Windsor’s predictive analytic models. This is the way the predicted order field is usually used but an Australian retailer called Surf Stitch used it a little bit differently. They used the predicted order date field to locate customers who had the potential of “churning” and sent out win-back emails in cycles since the last purchase by that customer (60, 90 and 120 day cycles).

They used their customer's purchase history combined with the predicted order data in order to reduce churning customers by 72% in six months. This shows how effective the predictive analytics can be for a business. If you use the information in an innovative and thoughtful way it could really have a positive impact on the customer experience and retention. The information is standard data but when you apply to these fields and use it correctly you end up keeping customers and saving money. The great thing about predictive analytics is its flexibility, you can literally apply it and use it in countless ways. Companies are always coming up with new ways of using the data and it is paying off. Not everyone has realised the benefits and advantages of predictive analysis yet but it won't take long before they realise the positive impact a small amount of data analysis can have. Traditional predictive analytic methods are occasionally still used by some businesses despite the inefficiency and ineffectiveness. It takes a lot of time and a lot of work for the old methods to produce any results, causing a negative impact rather than a positive one when you take into account all of the time going into it. The newer methods are far more effective and user friendly. Eventually the older methods will be phased out completely as companies look towards the newer methods for better service. The switch will be easy for a company to make when they compare the traditional to the newer analytical methods, especially as the older methods will cost them more and more. With the newer methods they can streamline processes and increase their efficiency.

Marketing Strategies

Marketing is an essential part of business for many organizations. When marketing there are many considerations to make including the strategies and products of competitors and the pros and cons of their own products. This can be a time consuming and sometimes difficult role but when you use predictive analytics the jobs is straightforward and simple. You can use predictive analytics in order to: identify potential customers, identify which product combinations are most effective, establish the best marketing material, and most effective ways of communication, determine marketing strategy timing, reduce cost and calculate the marketing cost by the number of orders. All of these factors can be easily worked out and this information can be used in order to improve marketing strategies for your business. Analyzing this data helps to find out what the customers like and don't like making marketing a lot easier. Insurance companies take advantage of this information in order to attract potential new customers. They will analyze a large amount of customer feedback data and scope social media in order to discover how the customer feels about things.

The insurance company will also analyze the amount of time an individual spend on the frequently asked question section of their website as well as message boards. This is all analysed in order for the insurance company to make a custom profile specifically for that customer. All of this data is available to them so the insurance company is taking full advantage of that to give themselves the best chance of reaching new customers. The industry is competitive so they have to focus on meeting the needs of their customers in order to stay in the game. The new data analysis makes this easier and gives them a wider, more flexible reach. They are using it to figure out whether their customers are happy with their service or not. They also use the analysis to work out if a customer is likely to stop using them, this can be identified by tracking behaviour and comparing this to the customers who have actually cancelled their policies. If a customer is likely to cancel then they will be flagged up in order for the company to make an effort to keep the customer with the company. They can do this by using special offers and services to convince the customer to stay. All of this is made

possible by the analysis of data that is always available. The more time the business spends on resolving customer issues before they actually happen, the better for the business and that is how predictive analytics help. Without the analysis of data the customers go unnoticed until they have stopped buying products and by then it is too late. This shows how crucial and effective predictive analytics is for marketing and how personalized and adapted the marketing can be to suit the specific customer.

To summarize, data analytics is being used to help insurance companies make many improvements to customer service, reduce losses through fraud and lower premium prices. The FBI reports there is $400 to $700 more spent by customers on premium costs due to fraud losses. Predictive analytics can have a positive impact on this.

*Fraud Detection

Fraud is a huge threat to every organization, it can be very difficult for a company to detect fraud especially when data is lacking. Fraud includes fraudulent transactions (online and offline), identity theft, inaccurate credit applications and false claims of insurance. No matter whether the businesses is big or small they can still become affected by fraud. Any kind of company is at risk of fraud including retailers, insurance companies, credit card companies and service providers. If a company uses predictive analytics they can come up with models that can detect data that is incorrect meaning the risk of fraud is lessened. The public and private sectors can both have fraud risks reduced by using predictive analytics. There has been a risk scoring method developed by Mark Nigrini that identifies audit targets. This method has been used in franchise sales reports that have been linked to a large international food chain. The franchisee starts with a score of 10 predictors and this score increase with weights to end up with the overall risk score. This is used to identify the franchisees that are more at risk of being affected by fraud. Once a franchisee is aware of the risk of fraud they can take the necessary steps to help prevent this from occurring. Using this method can help to identify fraudulent travel agents, vendors and accounts. The more complex models that are used can send monthly reports on fraud that has been committed. The Internal Revenue Service of the United States of America uses predictive analytics in order to monitor fraud and help to prevent and stop fraud from happening. Cyber security is also an increasing problem but the use of predictive analytics can assist in creating a model that can identify abnormalities in specific networks.

Fraud is a huge problem and can have very serious impacts on businesses, companies are always trying to battle fraud and predictive analytics is one of the best ways to do this. Claims can be compared to those that have been previously found to be fraudulent, allowing the company to more accurately determine applications that are fake and fraudulent. By looking at certain variables it becomes quickly clear which claims are real and which are not, if the claims match up they can be further investigated in order to determine what is more likely fraud. Generally speaking there is a pattern to fraud and

computer analysis picks up on these patterns a lot easier and quicker than a human will. A computer will pick up on things a human may miss which is why computer analysis of data is so effective as well as time saving. The computer can flag items up that the person can then look into in more detail. The company can also investigate into the people the claimant associates with meaning looking at their social media activities, claim partners or more. The company may find dishonest behaviour from these associates in the past and will also check of credit reference agencies to ensure a thorough investigation of fraud as a result of the initial information provided by predictive analytics.

Processes

Analytic processes can be used to observe the institution, its goods or offerings, its portfolio, and even the finances. The reason for this is to pinpoint areas of profit, such as predicting how much inventory to purchase and how to best maintain factory resources. A good example of this is how airlines use analytics to determine how many flight tickets they need to sell at a certain price in order to make a profit. Hotels also use analytics to determine how to modify their prices in order to fill their rooms in a profitable manner at a certain time of the year. Predictive analytics can also be used by the Federal Reserve Board to foresee the unemployment rate over the next year.

Insurance Industry

The current-day big insurance companies actually use a rather outdated method of collecting predictive analytic data. Numerous data analysts must be employed, and with the enormous amount of information flowing into the company's gates, it's hard to manage, takes a longer amount of time, and is generally inefficient. With so much data coming in from so many directions, proficient data analysis is more difficult to achieve, especially with an ever-changing marketplace. In order for companies to stay up to date, they need to implement predictive analytic models that are relevant to their specific business factors in order to maximize overall company success. Results need to be constantly reevaluated through thousands of repetitions in order to remain accurate. Insurance companies in particular can make use of the following predictive analytic tactics to advance business progress:

- Insurance-specific predictive analysis and modeling
- Fully operational predictive analytics
- Big data analytics
- Prescriptive analytics
- Personalized acumen and well-informed solutions

While insurance companies need to understand that predictive analysis will advance total business progress, the old methods of gathering data are still prevalent in the industry. Although in the past they have been proven to be proficient, with the constantly modernizing marketplace, they are becoming increasingly outdated and of minimal use, which in the long run can actually do more damage than good to the business. Inaccurate data analysis can result in a detriment in revenue sources, not getting the proper full use out of resources, and above all, sales.

There is an overwhelming amount of data constantly flowing into the company systems, and it is vital to have an active, automated predictive model to keep up with the pace. As technology is progressing and businesses are growing, it is far too expensive to employ people to carry out manual predictive analysis operations. Although the implementation of an

automated system would still require the employment of knowledgeable technical personnel, the overall advancement and improvement in business and profits will cancel out the cost, which saves both money and time. The fact that analysts and statisticians are skeptical about implementing automated predictive analysis models is in the long run doing more harm to the business in a number of aspects, especially company costs and decreased profits. Updated methods of gathering analytic data not only cuts costs in manpower, but also provides the information necessary for these companies to generate more sales, reduce fraud, and achieve an expansion in clientele rather than a decline. Automated analytics also allows for thousands of operations to be carried out simultaneously, with an incredibly increased operational speed. With the saved time and more accurate data, the insurance companies will achieve a general growth in productivity, maximized everyday processes, efficiently discovering tactical business investments, and foreseeing fluctuations in the marketplace. In short, it can revolutionize business success. Business processes are taken to full capacity, internal processes are enhanced, and companies who implement automated analytic systems are ahead of the game against competitor companies. Contemporary analytics allows for more intimate workings with the clients, and have proven accommodating to a broad range of customers in all sorts of industries. With innovative algorithms and modernized technology, analytic data is acquired and put together in an immeasurably quicker manner, and results can be better customized for every client.

Shipping Business

A new database has been implemented by UPS by the name of On-Road Integrated Optimization and Navigation (ORION). This program dictates that predictive analysis methods should be applied throughout all of the company's operations, even with the employees themselves. UPS is installing this system of analytic data gathering to all 55,000 of its supervisors and transporters, in accordance with the following factors:

- Employee Regulations
- Business Regulations
- Map Data
- Client Data

The main purpose of these data variables is to determine which routes are most effective for the delivery drivers to take, in order to save time and mileage. UPS is using inventive methods in order to get the full use out of their predictive analysis program, training their supervisors to be able to comprehend the analytic results and use them to capitalize on business productivity. ORION was first introduced and tried out by UPS in 2008, but at the time there were too many factors with unclear connections for the managers to handle and work with. ORION has since been developed and improved upon, with only the most vital factors being considered and examined by the managers, who then instruct the drivers on how to understand the results. The renovation and simplification of the program proved highly successful, with one driver diminishing his route by 30 miles over the course of a few months by taking into consideration his route's personalized predictive data.

Controlling Risk Factors

In addition to optimizing business productivity and increasing profits, predictive analytics also calculates a business' risk factors, in the same way as how it can identify fraud. The cornerstone of determining a business' risk factors is that of credit scoring. A customer's credit score is a very useful factor in predicting their likelihood of non-payment. Credit scoring itself is a method of predictive analysis, made up of all credit information of a specific person.

Staff Risk

Not only is customer-associated risk an issue in business, but there is also the factor of employee risk. The same predictive analytics can be used to consider any employees who may be a liability to the company. There are two main approaches in deciphering which members of staff may pose a risk to the company: “blanketed management” programs and the “squeaky wheel” method. Blanketed management concentrates mainly on employees that are non-risk or have a possibility of going unnoticed. Conversely, the squeaky wheel method concentrates on employees that exhibit regular disconcerting actions. Analytic data is especially useful in very large companies that have thousands of employees, where manual collection of data is simply impossible. Predictive analytics comes into play here by employing risk management programs, in which thousands of pieces of data are gathered and made perceptible. This data keeps account of small but significant alterations in an employee’s actions that could in the future turn into major detrimental conduct. Also considered are preceding tendencies amongst employees, which assist in foreseeing what could result from a certain employee’s misconduct or strange actions. With manual data collection by risk management staff, many of these issues can slip through the cracks, but with modernized systems, company managers have enough advance notice to be able to interfere before an employee causes serious damage to the business.

Risky behaviors of staff members that are not picked up on could have devastating outcomes for the company, so early detection by managers greatly cuts down the chance of this. Individual troublesome employees can be confronted at the time of issue and while the problem is the most clear. These intercessions can help a company avoid worker compensation claims and reduce the turnover rate. The blanket management methods being exchanged for automated, modernized, and customized analytic models diminishes overall company risk, as well as saves time and money. The analytic models can take into account all past analysis of employees’ conduct and mannerisms and make it easier to pick out people that could end up being risk factors. It is important to note that employees are never

considered a risk factor upon the moment of their hire. Many times, the cause of employees becoming high-risk is that of personal issues irrelevant to their work, such as medical, familial, or financial problems, which can overflow into their work performance. Analysis of employees can therefore note the differences in behavior, no matter how subtle, and determine early on if a member of staff is showing decreased performance or worrisome behavior.

A fine example of where predictive analysis for employee risk holds key importance is that of the transportation industry. Since transportation staff is generally in charge of human lives, employees exhibiting high-risk behavior need to be confronted before serious issues are caused. In addition to the possibility of passengers being harmed or even killed, employees who exhibit evidence of dangerous driving habits can cost the company money through fines and traffic violations. Predictive analysis of drivers tracks and keeps account of their driving habits. Aside from irresponsible driving such as talking on their phone while on the road, non-work-related problems could be distracting to drivers, such as an ill family member. Besides simply distracting thoughts, it is possible for drivers to try to multitask and take care of these problems while on duty, such as calling to make doctor appointments for their family member while driving. Comparing with the driver’s specific past analytic data of their driving patterns, managers can be made aware of any changes or irregularities in their habits. This could be anything, such as braking too hard, speeding, or standing idle for too long. The manager can then confront the employee in question to work together to rectify whatever issue is causing these discrepancies before a serious accident occurs. An example of a compromise would be to alter the driver’s schedule or lessen their workload. Predictive analysis of employee behavior can help managers identify not only the problem itself, but also what is causing it. Usually, managers tend to concentrate on the issue itself and not the underlying cause, but with this added consideration, it can greatly decrease the likelihood of the problem being repeated in the future.

Underwriting and Accepting Liability

It is important for a lot of businesses to sign insurance policies to accept liability in case of loss or damage, also known as underwriting. This varies as to the type of services the company offers, and potential costs need to be tailored to their specific risk factors. Different insurance companies need to account for different things when signing liability agreements. For example, automobile insurance companies must define the precise premium to be billed in order to insure every vehicle and driver. Financial companies need to determine prospects and compensation abilities of a borrower before approving a loan. Health insurance companies can use predictive analysis to decipher a client's medical history, pharmacy records, and past medical bills. With this data, the insurance company can roughly calculate how much money in bills that the client may submit in claims over the coming years, which can assist the company in determining an appropriate plan and premium.

Predictive analytics assists in the underwriting of all of these risk factors through examining a specific customer's habits upon their application. In addition, it provides a quicker rate of transaction when processing loans and other financial matters upon credit score analysis, which has proven especially useful in the mortgage area. Settlements that previously took days or even weeks with the previous methods can now be completed within a few hours. The likelihood of defaulting customers can also be greatly reduced when assigning prices of financial products and services. The most popular cause of customer default is that interest or premium rates are too high, but with the modernized analytic methods, insurance companies have been able to more accurately avoid non-paying customers. Credit scores are more easily evaluated, and the companies can use this information to predict loss-ratio performance and financial habits of their customers. Insurance companies today are using predictive analytics to calculate how successful prospective policies will be. Predictive analytics are especially useful to home insurance companies, because there are so many factors that can determine a house's value and risk factors. Perhaps a house does not have foreseeable market value increase, or is in a location

where it is vulnerable to natural disasters such as earthquakes or floods. Predictive analytics can provide all of this data to determine an appropriate insurance premium, without which they could potentially lose millions of dollars in damages due to natural circumstances.

Freedom Specialty Insurance: An Observation of Predictive Analytics Used in Underwriting

There was a case study done by the D&O (Directors and Officers Liability) insurance industry, in which the executives of Scottsdale Insurance Company were proposed a precarious underwriting submission following the recession in 2008. The proposal stated that the liability insurance (compensation for damages or defense fee loans, given a scenario in which an insured customer was to suffer a loss as a result of a legal settlement) was to be paid to the institution and/or its executives and administrators. The Scottsdale Insurance Company approved this proposition, and thus Freedom Specialty Insurance Company was formed. This new company placed the industry as the top priority, using external predictive analytic data to calculate risk, on the basis that D&O claims could be foreseen from class action lawsuit data. An exclusive, multi-million dollar underwriting policy was created, the disbursements of which have proven profitable to Freedom in the amount of $300 million in annual direct written premiums. Losses have been kept at a minimum with a rate below 49% in 2012, which is the industry's average loss percentage. The model has proven successful in all areas, with satisfied and assured employees in all levels of the company, as well as the reinsured being contented. This case study is a great example of how predictive analytics helped a company like Freedom soar with a revamped and modernized underwriting model. Many teams took part in developing the new policy: the predictive model itself was constructed and assessed by an actuarial firm; the user interface was crafted by an external technology supplier, who also formed the assimilation with company systems; and technology from SAS supplied components such as data repositories, statistical analytics engines, and reporting and conception utilities.

The refurbished system that Freedom employed consists of the following parts, which are elaborated upon below:

- Data Sources
- Data Scrubbing

- Back-testing
- Predictive Model
- Risk Selection Analysis
- Interface with Company Systems

Data Sources: The system assimilates with six external sources of data (such as class action lawsuits and other financial material), and the data is acquired through executive company applications. The external sources are frequently utilized by the D&O industry. In the case of Freedom Specialty Insurance Company, they have exclusive sources to contribute to their predictive model. Freedom in particular occupies a lot of time with acquiring and retaining valuable information regarding merchant activities. Classification and back-testing expose merchant flaws, as well as inconsistencies in their information. Freedom exerts extra time and energy into collaborating with merchants to catalogue information and maintain it at a high worth. Freedom also keeps a close watch on their external data, applying stringent inspections to develop policy and claims information. The company upholds data liberty from merchants' identification schemes. Although it takes more effort to decipher values, the process safeguards that Freedom can promptly terminate business with certain merchants if necessary.

Data Scrubbing: Upon its delivery, information undergoes many "cleaning" processes, which assures that the information is able to be used to its maximum ability. For example, there is a review of 20,000 separate class action law suits per month to observe if any variations have occurred. They were originally classified by different factors, but now they are gone through monthly. In the past, before the automated system was put into place and the process needed to be carried out manually, the process took weeks to finish. Now, with the modernized methods and information cataloguing devices, everything can be completed within hours.

Back-testing: This is one of the most important processes, and determines the potential risk upon receiving a claim. The system will use the predictive model to run the claim and analyze the selection criterion, altering tolerances as required. Upon being used numerous times, the positive feedback loop polishes the system.

Predictive Model: Information is consolidated and run through a model, which defines the wisest range of appraisal and limits through the use of multivariate analysis. Algorithms assess the submission through numerous programmed thresholds.

Risk Selection Analysis: This provides the underwriter with a brief analytical report of recommendations. Similar risks are shown and contrasted alongside various other risk factors, such as the industry, size, monetary configuration, and considerations. The platform's fundamental standard is compelled by the underwriter's rationality, with the assistance of technology. In other words, the system is made to support, but not replace, the physical human underwriter.

Interface with Company Systems: Once a conclusion is made, designated data is delivered to the executives of the company. The policy distribution procedure is still generally done by hand, but is likely to be replaced by automated systems later on. The policy is distributed, and the statistical data is rerun through the data source element. More information is contributed to the platform as claims are filed through loss.

As is evident in all D&O processes, underwriters are required to have thorough understanding of technical insurance. While in the past underwriters put a great deal of effort into acquiring, organizing, and evaluating information, they now have to adapt to a system in which enormous quantities of data are condensed onto a number of analytical pages. Predictive analytics have greatly altered the responsibilities of a customary underwriter, who now cross over with policyholders and negotiators in book and risk control. Although the technology has simplified and cut down a lot of the manual work, additional experienced technical personnel also needed to be employed who have legal and numerical awareness that allow them to construct predictive models in the financial area. Integrating this model has enabled Freedom to advance proficiency in the process across many zones. Processes involved in managing information such as data scrubbing, back-testing, and classification were all discovered and learned by the people themselves and were originally carried out by hand. However, they have been increasingly mechanized since they were first conceived. Also, there is an ever-growing

quantity of external sources. Freedom is currently undergoing processes to assess the implementation of cyber security and intellectual property lawsuits, with the predictive model continuously being enhanced and improved.

The D&O industry has adopted many of the processes related to the upkeep, feeding, and preservation of the predictive model that are utilized by other industries. One situation in particular is that following the actuarial originally constructing the predictive model, Freedom achieved full fluency in the program’s complex processes over the course of many months. Operations were implemented to efficiently oversee all of the external merchants together. A number of external assemblies (including the actuarial firm, the IT firm, data vendors, reinsurers, and internal IT) came together to refine and organize the predictive model together, all of them in close collaboration with each other. It was a great feat for Freedom to unite all of these individuals to take advantage of their distinct expertise and understanding all together simultaneously.

Positive Results from the Model

Freedom ended up having very positive results from the implementation of their predictive analytics model, with many new opportunities and insights provided for the company. Communication and correspondence with brokers and policyholders on the topic of risk management was boosted as a result of the highly detailed analytic results. The model could even be expanded to cover other areas of liability, like property and indemnity. Back-testing and cataloguing mechanisms can also now be implemented to foresee other data components in the future. The updated and automated model highlights Freedom as a tough contender amongst competitor companies, and has opened up windows to uncover even more possible data sources.

The Effects of Predictive Analytics on Real Estate

Predictive analytics can also have an effect on real estate. There was one instance where a real estate company assisted a law firm in choosing whether or not to relocate to a different office space through the usage of data devices. The law firm wished to bring in and keep the most suitable employees, so the first factor to be evaluated as personnel retention. This was the main issue for the client, so the real estate company used its resources to map out where the employees were most often. The real estate company assisted the client by utilizing different location-conscious mechanisms to keep track of the whereabouts of the company's personnel, in which the data was accumulated based on employee partialities and activities. The end result was that the law firm decided concluded to relocate from the high-rise office into a more affordable space based on the location habits of its personnel. This saved on costs for the law firm as well as improved employee retention because of the new and more convenient location.

The National Association of Realtors (NAR) and Its Use of Predictive Analytics

Predictive analytics is having a huge effect on the NAR and revolutionizing how they carry out business transactions. The NAR consists of around 1 million affiliates, which makes it America's biggest trade association, so the data attained through predictive analytics can have a huge effect on the real estate business. The NAR has come up with a new analytics group that will assist them in adding value to their client affairs. The new analytics group is aimed to improve on the following points:

- Evaluating data tendencies of both affiliates and clients
- Adding worth to its realtors
- Using incongruent data models to construct analytical models
- Using models to resolve intricate issues in the real estate business
- Assist real estate agents, domestic associations, and others reach more accurate data-based conclusions

The new data group is meant to initiate in three stages, the first one being the experimentation stage. In this stage, the data group will decipher all of the data that has been collected and recognize tendencies and patterns. The following two stages will consist of building business relationships with other data users, which will assist in innovating and creating products to support their agents. The end goal of implementing the predictive analytic data models is to aid the NAR in discovering behavioral trends and construct specifically aimed consultations with prospective home buyers, which will ultimately improve functioning procedures for the NAR and raise their financial standards. The customary data-analyzing methods used by the NAR are to accumulate all-inclusive patented data, which showed what has occurred in the group's past transactions. The other component of data analysis is gathering data from the consumer relationship administrative system to define present happenings that are having an effect on the group. The new, modernized method will take both patented and public information systems to distinguish patterns that can provide a forecast for forthcoming real estate business transactions.

In addition to the real estate industry, insurance companies are also updating the way they collect data to improve business operations. Predictive analytics improves the relationships and communication between doctors and patients to achieve quicker and more effective healthcare solutions, which ultimately results in patients staying healthier because they are receiving the appropriate and most effective treatment. It is true that insurance companies will lose millions of dollars in premium charges as a result of patients not needed as many extreme and excessive treatments, and the price of insurance will drop dramatically, not to mention that extra claims will be eliminated. Although this is true, insurance companies can still improve their proceeds in new and different manners as the healthcare field continues to become more individually focused on the patients and their personal care plans. Hospitalization of patients will be individualized, but also shortened in length. Insurance companies can also follow the trend of specialization, and add new niches to their packages to cover certain areas, which in turn will generate more revenue. The same goes for medical apparatus companies. There will not be such a prevalent need for many of the common medical devices that they have been producing, which would imaginably result in a loss of revenue as well. However, similar to the insurance companies, the medical device manufacturers can also join in the game by producing more specialized devices that will be indispensable in the modernized healthcare setting, which will actually increase profits and raise their financial standards. The use of predictive analytics is revolutionizing many areas of the healthcare field, and these are just some of the changes that are to be seen in the future are predictive analysis becomes more sophisticated and advanced.

The Revolution of Predictive Analysis across a Variety of Industries

Predictive analysis has undoubtedly had a huge effect on many industries regarding the gathering, evaluation, and prediction of data. In the section, “Why Predictive Analytics,” it can be seen how predictive analytics are transforming a variety of industries all across the globe. There are numerous industries that are on the modern verge of the improvements that predictive analytics can provide in a number of areas. Freedom Insurance, for example, has established a predictive model that is being adopted by other branches of the insurance industry as well, and they have shot ahead of them all. Predictive analysis has proven to be a very profitable market, achieving between $5.2-6.5 billion by 2018/2019.

The healthcare industry in particular has been greatly influenced and improved through the use of predictive analytics, even down to doctor-patient relationships and care methods. In the past, it was customary for a doctor to assign a generic treatment regimen to a patient to become healthy again, but nowadays doctors are more individualized with their patients and focusing on what is best for their specific case. The patient is now much more involved in their own healthcare, and can access their medical records through their own particular doctor. Patients are no longer kept in the dark about their medical history or disease susceptibility related to heredities or other factors. Predictive analytics allows patients to evaluate their personal needs along with their doctor and mutually contribute to deciding what the best health care plan is for them. Whereas the patient and doctor roles were always separated by a clear line, those roles are now blending together into a cooperative relationship. All of these drastic and positive changes are a result of simply five to ten years of gathering predictive analytic data.

In every case study that was analyzed for the purpose of this book regarding predictive analytics, the result was consistently positive. All of the research cases exhibited improvements in all of the industries’ business operations that used modernized predictive analytics devices. Predictive analysis can

be applied to countless industries; the most prevalent that have seen positive results are listed below:

- **Law Enforcement:** Improvement in the recreation of crime scenes
- **Insurance:** Clarified and improved risk management in regards to both personnel and clientele
- **Transportation:** Easier identification of concerning behavior of drivers, resulting in the avoidance of thousands of citations and arrests, as well as the saving of millions of dollars in accident-related damages
- **Healthcare:** More accurate and effective diagnoses and assignments of surgeries or treatments by doctors; more intimate doctor-patient relationships; more lives saved as a result of more accurate and appropriate treatment plans
- **Retail:** Significant increase in number of returning customers
- **Real Estate:** Easier identification of popular property packages amongst potential homebuyers; pinpointing most profitable rental properties based on renters' monetary, geographical, and family needs

It is evident that predictive analytics has revolutionized operations across many industries. In the coming years, the prospective market will dramatically grow as more and more positive results are achieved and recognized through innovative companies making inventive use of it.

Chapter 13: Descriptive and predictive analysis

Every company will be at an advantage if they have a powerful intelligence department. With the massive amounts of data available to businesses, organizations are turning to analytics solutions to make sense of all this information to help improve decision making. It is an enormous benefit for any organization to have skilled personnel who can assemble and classify large amounts of data since this information can have critical importance to business. This class of analysis examines business on a macro level to scrutinize quantities of data and identify values that could be used for future reference. Companies need capabilities to analyze data to be able to forecast future trends. The intelligence gathered will only have significance if the team understands and applies the data to make calculations through methods of descriptive or predictive analysis.

Descriptive and predictive analysis are two approaches of data analysis in business. Descriptive analytics can tell you what has happened or what is happening now. Predictive analysis asks questions about the future. This type of analytic forecasting constructs investigative outlines on the various stages of industry. These forecasts may comprise of intelligence regarding a consumer or commodity that will make business decisions clear. Using information such as customer and employer behavior, the highs and lows of markets, and other such related data, accurate judgements can be made for prospective business direction.

When there is a need to get a comprehensive picture of what is going on in the organization, descriptive analysis is applied. It describes the past and summarizes basic data that describes different aspects of the business. Reports provide information about sales, customers, operations, finance, and correlations between the data. Questions about previous business dealings, details of product users, the status of certain outlets, and sales

indicators of supplies may be answered. Through this system, the company will be aware of bestselling products, average salary of customers, annual spending on niche products, and sales of seasonal commodities. All information is constructed from data of former activity.

Predictive analytics is used when information about forthcoming commercial activity is required. It provides organizations with actionable insights based on data, and estimates future outcomes. It predicts customer behavior in sales and marketing, demands for operation, or determines risk factors for finance. Predictive analytics will show appropriate charges for a certain product, the upgrades customers want, and whether customers will buy a certain product. Such information allows focus on the creation of products that will influence future growth and raise of business revenues. Predictive analytics has become an important aspect part of big data since it allows organizations to predict future outcomes.

Both descriptive and predictive approaches to data analytics are important. Descriptive analytics is used to understand data and tells you what has happened whereas predictive analytics forecasts techniques and tells you what will happen. Neither form of analysis may be more important because they are both codependent and interlinked, each dependent on the other. Without feedback from descriptive statistics, predictions would be impossible. Without these predictions, descriptive information would be useless. Both types of analytics are useful business supports because they allow learning from past behaviors and understanding about influencing future outcomes.

Chapter 14: Crucial factors for data analysis

The use of big data analysis is becoming standard procedure in organizations as more companies include this practice as a part of business activities. Companies that are progressive are investing funds in big data analysis because this business intelligence has proved to create amazing outcomes by rationalizing information to increase productivity and revenues. However, in many cases, organizations have shown no benefits from installing this system. Questions arise about whether big data still holds the promise that was expected. In principle it seems beneficial, while in practice, results were not as precise as anticipated. In cases where the information delivered was incorrect and useless, the system has been abandoned.

A 2007 study revealed that a third of companies did not have budgets for intelligence projects. 31% of these projects were discontinued by companies, while 17% functioned so poorly that the judgement and existence of these companies came into question. Certain companies have made 100% return on investment by using business intelligence. They were able to reduce overheads dramatically and customer and revenues increased. Other firms had only a slight return on investment while the majority have seen nothing but losses caused by the system.

Little research had been done about the factors that contributes towards the success or failure factors of business intelligence. The success of business intelligence projects depends on the ability to assimilate large amounts of information into a system and have the analytical facility required to examine this information correctly. Important measures to determine the success of any BI venture are information assimilation, value of data, ease of accessibility, diagnostic ability, application of analytics in corporate activities, and making systematic judgements.

Before investing in business intelligence systems, questions about important achievement aspects that indicate project success must be answered. Key drivers towards accomplishment include communication, project management, vendor and customer relationships, and user participation, to name a few categories. It has been observed that smooth system usage, organization elements, technical features, and organization procedures are all significant issues that relate to the success of business intelligence projects. Since it is difficult for businesses to concentrate completely on these many aspects, it is necessary to find out which of these is most vital and then focus on these major features.

Support by top management

Experts are of the opinion that the success of business intelligence projects depends on committed organization support, pliant and interactive technical structure, change management practices, solid IT and BI control, and a united BI and business system. Consistent support from top executives makes it easier to secure necessary operating resources such as funding, human skills, training, and other necessary requisites throughout the implementation process. The success of any business intelligence system depends largely on the support it receives from top management in the organization. This has been noted as a vital aspect that effects success or failure of such projects. Without forward thinking vision and backing, the project would have little or no chance of succeeding.

It is the organizational department heads must support this sector by providing necessary resources for data analytics to function at full capacity. Without dedicated support from top management, the project might not get the recognition it needs to be a success. This happens because users tend to follow to the attitude of top management and will usually accept a system backed by superiors. Sufficient financial allocation and quality human resources will go a long way in establishing an efficient big data system and here top level support will make a huge difference to the BI system and its success.

Resources and flexible technical structure

During the initial installation of a big data system it is important to ensure that all the components are able to adapt to changes. This is an expensive time consuming project so hardware and software required for analyzing the huge volume of information should be flexible enough to adapt to current requirements and expand with future growth of the system. Initially it takes much time and money to support such systems, but with sufficient investment the benefits are sure to follow. Having access to big data empowers the users and organization to make accurate decisions. While cutting out conflict, errors and incorrect feedback, and false management information, the system creates improved situational awareness which helps to direct business. Setting up a successful business intelligence operation can transform an organization.

When not enough time or money is invested in the big data system, problems could arise very quickly. It is not uncommon to hear of data warehouses not able to function properly after being installed. The main reason why many business intelligence applications fail to deliver expectations is simply because the technical structure put in place is inadequate due to insufficient investment. Exorbitant costs, lengthy timescales, and cumbersome user requirements which keep changing, make it imperative that sufficient resources are made available from the start because this is an expensive project that will expand, requiring even more time and funding as it grows. One way of keeping costs down is to select a specific area of business to focus on first, and then add another important area and another. This is one way of ensuring your BI project’s success while gaining momentum and credibility.

Change management and effective involvement

Change management guides the manner in which individuals are prepared, equipped, and supported in order to adopt changes that will propel organizational success and productive outcomes. The success of any BI project is directed by the rate of executive involvement at the beginning of its implementation. Business intelligence operations work better and faster with a highly motivated level of management involvement. It often becomes essential to change company processes when business intelligence projects are installed. This is because BI solutions must unite with the organization's strategic vision and the project must be scoped and prioritized to concentrate on the organization's best opportunities.

New innovative ways of working are imperative for enterprises that require an effective BI system. It is imperative for the execution team to understand the processes of the system. Employees must be encouraged to accept new, advanced ways of doing things. If there is resistance to change by any people who are responsible for BI functions, they may have to be changed as well. Change management is required for data users to accept software into their workflow. It delivers a support structure for employees to move from their present state to a future state which is more business friendly.

For change management to be consistently and effectively applied to initiatives, leaders should have the ability to guide their teams through changes. Employees should be able to embrace change more easily, quickly, and efficiently so that organizations are then able to respond speedily to market changes and employ strategic initiatives and technology with less effect on productivity. This capability does not take place by accident. It requires a strategic methodology to implant change management throughout the organization. Change is good as it creates new opportunities and widens perspectives and horizons.

Strong IT and BI governance

The substandard quality of the data supplied to the analysts system is among the one of the main reasons for the failure of business intelligence. This data is gathered by the operatives, and systems to ensure quality of this collected data must be installed in the system to guard against inferior input. Only then can the operator produce a satisfactory study of the data. Reliable business rules and processes must cover an organization's information assets. IT and BI data sources, increasing complexities, operational intelligence, and information diversity creates an environment that requires consistent and thorough data management strategies.

Surveys have revealed that corporations who have a good strategy to govern their information also succeed in governing their business intelligence. Data security and governance at every level is crucial. Strong data governance should encompass all data assets across the organization to create a cohesive view of information and provide a way to manage inconsistencies and potential data quality issues as they arise. Strong IT and BI governance generates efficient data management and efficient information, supporting automated analytics and broader business insights, leading to superior analytics.

Alignment of BI with business strategy

Successful business intelligence initiatives are always aligned with the organization's business strategy. Business intelligence initiatives that are not aligned with organizational business objectives fail. Where there is a struggle to align technology approach to BI with specific business goals and objectives, solutions are delivered that fail to meet business needs. It is important that the organization's general business strategy and BI are aligned in order for the data system to be effective. When the BI department does not have a transparent, understandable business policy, it becomes hard to control and drive big data analysis towards the company's amplified productivity and increased overall returns.

High level alignment is achieved by understanding the company's business strategy and integrating this plan with data usage elements, metrics, dimensions, lists and values, and patterns of use. Business intelligence alignment with business strategy supports organizational processes and enables analytics efforts to meet business needs. It enables agreement on business vocabulary and definitions, manages costs and risk in reporting and BI methods. Solutions are designed that meet customer needs and deliver value, and partnerships may be created to advance shared information. Alignment of BI with business strategy leads to improved user reception and intensifies the benefits received from business intelligence investment.

Chapter 15: Expectations of business intelligence

Big data is expected to evolve and expand rapidly. Observing the opportunity big data offers for businesses through social media, we understand why this incentive could become the main feature round which companies focus their future business. Customers continuously create new technologies by the very nature of their modes of communications and new applications will evolve to take advantage of this. Some businesses wait for the development of improved systems while others invest billions on big data investigation, spending significant amounts in search of the best apps to corner the market. Self-service platforms will be popular tools for harnessing big data assets.

Advances in technologies

Technological processes of business intelligence are being significantly simplified to making it more available to the public. The development of software for use in business intelligence projects such as programming languages or initiatives that combine the force of algorithms, are making these systems easier to operate. The process of big data analysis has been streamlined to a stage that can now be managed on one personal computer whereas, in the past it was handled by a group of supercomputers.

Advances in technology now enables anyone with basic computer skills to utilize big data at home. Programs can evaluate the databases of millions of users. Billions of social media posts can be analyzed on a single personal computer within minutes. It once too one thousand powerful computers over six hours to perform the same undertaking. This clearly shows that advancements have revolutionized how data is analyzed. Business intelligence projects can pay huge dividends, making it a perfect field for innovators.

Hyper targeting

One question often asked regarding big data is about what its real its significance is. The answer is quite obvious and simple. With hundreds of millions of accounts on social networks such as Facebook, Twitter, and LinkedIn, it is easy to see why companies are investing millions of dollars in engagement with these users. Our whole world nowadays is linked through social media. With so many users and available data online, customer and product engagement can be turned into billions of dollars in brand worth and sales. Hyper targeting allows marketers to access user information and advertise to those individuals who show interest in their product or service.

Hyper targeting delivers advertising content to specific segments in a network. This super effective marketing tactic aims at targeting advertisements on social network sites. The advantage of hyper targeting is getting potential consumer’s personal details along with their lifestyle preferences. This valuable, abundant source of customer information leads to precision performance marketing. By collecting data and directly contacting with prospective customers, operators can hyper target certain sections of people and reach out to consumers in ways that were never previously imagined. Benefits can go beyond delivering advertisements to the right audience and saving on budget. Hyper targeting often helps with audience research and can lead to partnerships, sponsorships, and other co-branding opportunities.

The possibility of big data getting out of hand

Some experts caution that just like customer relationship management turned into a very costly and unsatisfactory exercise, the same may happen with big data in the future if it is not handled correctly. They believe the time of predicting human behavior is over. This may not be totally accurate because big social data is now important, valuable information for businesses to gather. Because this information takes account of all categories of social activities online, such personal information, regions and areas of interest, what we tweet, when we publish posts, and how many messages we send, businesses are able to get a better picture of market requirements.

Companies use such data to determine likes and dislikes and what we are buying and want to buy, thus potentially increasing their revenues. It is up to users to control their systems and the possibility of big data getting out of hand is up to their efficiency or lack thereof. Whether or not big data will translate into gold field for business or prove hazardous for organizations to base their prospects is yet to be observed. At present however, the outlook is positive.

Making forecasts without enough information

According to some experts, we are, in fact, past the golden time in predicting human behavior. Some believe this was easiest to do about five decades ago, when consumer information was not easily accessible, but few simple factors, such as frequency and monetary value, were able to turn direct marketing into a huge success. Knowing how recently the last purchase was made by a consumer was apparently pretty much the only thing you needed to make them repeat this process by directly marketing more products to them.

Sources of information for data management

At one time only traditional business information could be tracked. With the advancement of technology, we now have regular assess to a growing amount of information networks. Channels from which we can collect statistics on consumers that includes social media, search engine, sensor, and semantic data. Information about what people are looking for online, what they are buying and their buying habits, is readily available through data sources While some of these sources are company generated, others are externally produced. This system provides a complete understanding into the nature of data through which accurate prediction of customer and product behavior. Attaining information from available network sources and examining it combined with company data, delivers a complete strategy for data management. There are five main sources of data that we can access for valuable information.

Sensor data is one of our five main social sources of information. It is a class of data that includes information collected from meters that register utility usage. It also gathers data from supermarket cash registers, listing products purchased and prices paid. This type of information helps to improve practices in numerous sections of day to day transactional outcomes.

Social media data is another of our five main social sources. Here data from social media posts is collected by social listening tools. When we use Facebook or Twitter, there is an enormous quantity of information about ourselves we are putting out there. All of this information is being gathered by various bodies, and is used by companies to promote business.

Search engine data is also one of our five sources of data. This information source is used by businesses as exterior data channels, which enables them to develop a variety of conclusions. Google permits us to make in depth

investigations into what we are looking for on the internet, and information recorded is utilized for search engine data aimed at specific sections of people.

Enterprise application data is yet another of the five data sources and is used to note buying habits of consumers along with additional social behaviors. This type of information is internal. It is the property of an organization and gathered by their own systems. This material may comprise of the complete analytics of the business's website, and any other information that supports the structure of the organization's BI system.

Mobile data completes the group of five sources of information. This works in the same way as social media. You dish out tons of information when you use your mobile apps is in the case when social media is used. This data details your location, sex, and age. It is gathered and sold to businesses to provide them with valuable consumer feedback.

Combining these five sources of big data analysis provides companies with complete picture of the market they cater to. Analytics merchants try to use all these sources to create a data filled package that a company must purchase to include in their information base. Organizations are going through extreme measures to acquire and apply the information collected from the five data sources in order to be ahead of their competitors and boost revenues.

Chapter 16: What is Data Science?

Data science is the deep knowledge discovery that can be obtained through exploration and data inference. Data scientists have experience with both mathematical and algorithmic techniques, and they use these to turn raw information into useful insights. It's based on analytical rigor, evidence, and building decision-making capabilities, and it gives companies the information needed to make intelligent, well-strategized decisions about their operations. Data science is about learning from data, and using it to add value to information. Data scientists can work on many different projects, including:

Predictive analytics: Predicting things such as future demand.

Recommendation engines: Software that recommends items, such as Amazon offering a customer similar products or Netflix offering similar television shows.

Tactical optimization: This helps businesses improve their processes, such as a marketing campaign.

Automated decision engines: These are things like automated fraud detection and self-driving cars.

Nuanced learning: The goal of this field is to develop a better understanding of consumer behavior.

All of these categories are very straightforward, but the problems involved in them are anything but. Solving them requires a deep knowledge of machine-learning algorithms and very good technical abilities. These are skills that take years to develop, not days, and data scientists require a certain skill set in order to be qualified for their jobs.

Skills Required for Data Science

Data science is a field that requires skills from many different disciplines, but there are three main areas you must be competent in.

Mathematics

In order to get any meaning from data, a scientist needs to be able to see it in a quantitative way. Data has textures, patterns, dimensions, and correlations expressed in its numbers, and finding those and their meaning requires a knowledge of mathematical techniques. Business models often need analytic models to solve, and these come from the theories of hard math. However, you don't only have to know how to build the models, you also need to have a good understanding of how they work. People commonly believe that data science only involves statistics, but it is just one of the many important math topics that need to be understood by scientists.

Statistics itself has two different branches, called classical and Bayesian. Most people talking about statistics are referring to the classical type, but data scientists need to have an understanding of both in order to do their job. They also need to know about matrix mathematics and linear algebra. Overall, data scientists need a deep and wide understanding of math.

Technology and Hacking

Before we get into this, I need to clarify that the hacking we are talking about it not the top-secret spy stuff where you hack into people's computers and steal classified information. The hacking we are talking about is the creativity and ingenuity needed for building models with learned technical skills, then finding solutions to problems.

These skills are vital because data scientists leverage technology in order to acquire huge data sets and use complex algorithms. Basically, they need to have a far greater knowledge than just Excel. They need to know about SAS, SQL, and R, all of which require the ability to code. These tools allow data scientists to look at data, organize the information within it, and use that information to come to useful insights, which would otherwise simply stay hidden in the vast amount of data.

Hackers are algorithmic thinkers who can take a difficult problem and find a solution to it. It is a very important skill for data scientists, since their job requires that they work within existing algorithmic frameworks and create their own to help solve problems.

Business Acumen

Data scientists are, first and foremost, strategy consultants. They are a valuable resource to a business because they have unique ways to add value to the company. However, this also means that they need to have the knowledge to approach and analyze business problems in addition to handling algorithmic problems. Their value isn’t thanks to a number, it is thanks to the strategic thinking they provide based on that number. A data scientist also has to be able to use data to tell a story, complete with a problem and solution. They use insights from their data analysis to craft this story, and it is a core skill for their job.

What does it take to be a data scientist?

Data scientists need to be intensely curious, and they also need to be able to think critically and deeply. Data science is based in curiosity, questions, discovery, and constant learning. True data scientists will tell you that money has never been a motive for them - it’s all about being able to satisfy their curiosity and use their skills to solve problems.

Turning data into meaningful information isn’t just about finding the answer, it’s also about finding what’s hidden. Solving problems is a journey that gives data scientists the intellectual stimulation they crave while bringing them to a solution. They love being challenged and they are passionate about what they do.

Data Science, Analytics, and Machine Learning

Analytics is a commonly-used word when it comes to the business world, and often it is used a little loosely. Analytics describes a quantitative method of critical thinking. It is the science of analysis, the process of gathering information gleaned from data and using it to make informed decisions.

An analyst can do many things or cover many roles, as it is a rather ambiguous term: it can refer to a market analyst, an operations analyst, a financial analyst, and many more. This brings us to the question of whether data scientists and analysts are the same thing. They're not, but analysts are basically training to be data scientists, and they certainly have the heart of one. Here are some examples of how an analyst can become a data scientist:

- An analyst who has mastered Excel learns how to use SQL and R to access raw warehouse data.
- An analyst who is knowledgeable about stats and able to report on the results of an A/B test learns how to build predictive models with cross validation and latent variable analysis.

The point here is that it takes motivation to move up from being an analyst and become a data scientist. You'll have to learn many new skills. However, many organizations have been successful by providing the necessary training and resources to their analysts.

When talking about data science, the term machine learning comes up fairly often. Machine learning is the ability to train systems or algorithms to gain information from a data set. There are many different types of machine learning, and they range from neural nets to the regression model. However, they all have one central goal: teach a computer to recognize a pattern. Possible uses for machine learning include:

- Predictive models that can anticipate a user's behavior.
- Classification models that can recognize and filter out spam.
- Neural nets that learn how to recognize patterns.
- Clustering algorithms that can mine and help find natural similarities between customers.

- Recommendation engines that can learn about individually-based preferences.

As you can see, data scientists use machine learning very often. They use it to build algorithms that automate and simplify certain parts of problem solving, which is essential when it comes to complex, data-driven projects.

Data Munging

Raw data has no structure and it can get very messy, and the process by which it is cleaned is called data munging. The cleaned data can then be analyzed and used in machine learning algorithms. Data munging requires excellent hacking skills and the ability to notice patterns - this helps when it comes time to merge the raw information and transform it. Dirty data often obscures the truth, and its results can be misleading if it isn't cleaned first. A data scientist must have excellent data munging skills in order to make sure they are working with accurate data.

Chapter 17: Deeper Insights about a Data Scientist's Skills

When you're wondering what skills a data scientist must have, one of the best people to ask if the person who hires them. Recruiters know exactly what they are looking for when they review potential consultants. For example, Burtch Works recruits senior data scientists to work in the industry and business fields. In 2014 they published information regarding the skills a data scientist is expected to have mastered. This information is very credible because it comes from an expert-led firm that has climbed all the way up into the Fortune 50, which is a sign of their huge success. Let's take a look at the attributes that Burtch Works felt were most relevant.

1. Solid education: Research shows us that most data scientists had have a solid and serious education. 88% of data scientists have a Master's degree, 46% had a Ph.D., and the rest have all been through extensive schooling as well. This is because a data scientist needs to have an incredible depth of knowledge to do their job.

   No matter their level, all data scientists must have experience in fields that require calculations and skills with formulas and the analysis of numbers. Burtch Works has started that 32% of data scientists have an extensive background in mathematics, 32% have a background in statistics, 19% are computer scientists, and 16% come from an engineering field.

2. Competent in R or SAS plus: R programming language is very useful when it comes to creating important functions. SAS is Statistical Analysis Software, so any potential data scientists should be familiar with it because it will help them analyze data at an advanced level, such as when it comes to predictive analytics or data management. Overall, any useful analytical tools are good for a data scientist.

3. Skilled in Python Coding: Burtch Works states that many employers want their data scientists to know how to use Python

Coding, a popular and common coding language. They also prefer a data scientist who understands how to use Perl, Java, or C/C++.

4. Background in Hadoop: Data scientists should be very knowledge when it comes to the Hadoop platform. These skills aren't mandatory, but being able to easily derive statistical data from such an open-source library is helpful. It's also good to know about Apache Hive, which is a kind of data warehouses software. It uses HiveQL, which is similar to SQL and is very helpful when it comes to querying data.

Apache Pig also gives a data scientist a boost when it comes to qualifying for a job. This platform helps data analysis. Also, if you intend to make a career out of being a data scientist, it is best that you learn about current cloud tools such as Amazon 83 - this also requires that you stay updated with any new additions to this topic.

5. SQL Database skills: Data scientists must be able to work with SQL, which is Structured Query Language. This is a programming language often used in stream processing and data management. A data scientist must know how to write and execute complex SQL queries. Other modeling and architecting data tools and software such as Erwin, DataSpy, and TOAD are also useful when it comes to compiling and analyzing data.

6. Ability to manage unstructured data: A solid educational background is useful when it comes to a data scientist's ability to work with unstructured data. Not every part of their work requires calculations, but even then their knowledge must go deeper than that. After the calculations, a data scientist must be able to comprehend which data is relevant in which equation, otherwise their results just won't make sense in the business world. This requires excellent critical thinking skills in order to make good use of the large amounts of information from sources such as video feeds and social media.

7. Intellectual curiosity: A data scientist needs to have more than an excellent memory when it comes to facts, formulas, and other textbook information. They also need to be naturally

curious and eager to learn new things about the world and its events, which have an effect on the efficiency of businesses, the cost of doing business, and overall profitability. Frank Lo founded DataJobs.com, and in 2014 he was a guest blogger for Burtch Works. He stated that curiosity is one of the most important soft skills a data scientist can have. It's not a technical ability, but it will still make a data scientist stand out from all the rest when it comes to analytics.

8. A good understanding of business: Companies want data scientists who are knowledgeable about the business world so that they can be useful when it comes to improving business practices. Business acumen is a non-technical skill, but it is very useful when a data scientist is trying to solve business problems by analyzing data.

A data scientist with this experience will be able to learn about the strengths and weaknesses of a business, which can be essential information for business leaders to have when it comes to prioritizing tasks, protecting the business, and helping it advance. A good data scientist can help a business leverage the data it has.

9. Effective communication skills: You've likely heard somebody talking just for the sake of sounding smart. They're annoying, right? This isn't a data scientist's goal. Sales and marketing managers likely won't care whether they use VLOOKUP or IFFEROR or INDEX+MATCH or IF. What they really care about it the quality of the information the data scientist can give them based on those formulas. Managers want to know about the current state of the market, which the data scientist will deduce from the possible scenarios and mass of data.

Their job is to provide insights into the business, and they need to communicate effectively and clearly with the relevant decision makers in order for that information to make any difference. If a data scientist can't communicate, their analysis and efforts will come to nothing.

Demystifying Data Science

In the Harvard Business Review magazine, expert contributors Thomas Davenport and D.J. Patil labeled being a data scientist the sexiest job of the 21st century. Generally, there are some skills that set you ahead as a data scientist. However, different employers want different things out of their data scientists, so you need to make sure to give any job postings or advertisements a thorough read-through before deciding whether or not to apply. The job title “data scientist” covers a wide range of skills that can be narrowed down to four categories of engagement:

1. Data Analyst: Some job advertisements ask for data scientists, but what they really need is a person who is skilled in Excel with a knowledge of MySQL database. In this case, you should apply even you do not meet all of the qualifications we discussed above. Sometimes people are hired to fill a data scientist's job, but as long as they know how to spool reports and data from MySQL and work with Excel pivot tables, they will be all right. Sometimes they are wanted to work with teammates on a Google Analytics account.

   None of these involve the deep skills of a data scientist, but that doesn’t mean they aren’t helpful. Instead, they give you a platform to practice data scientist-level skills, which can be helpful if you’re new to the field. In this case, the job will allow you to further explore the field of data analytics, and you can slowly build your repertoire of skills.

2. Data Engineer: Some job postings list the skills typical of a data engineer under a data scientist label. If you are a data engineer and knowledgeable about the skills they list, take a chance and apply. Machine learning expertise, statistics, and other software engineering skills are all useful.

   Also, the job might ask for somebody who can build data infrastructure for their organization. This is a common move when the company has too much data, but they don’t feel comfortable getting rid of it in case it becomes valuable in the

future. Or maybe they feel like the data is important, but it is unstructured and they don’t know how to use it.

3. Statistician: Companies who handle data-based services sometimes need employees who can handle intense machine learning activity and consumer-focused data analysis - in fact, many even run on a data analysis platform. These jobs require somebody who is skilled in statistics or math, but somebody with a knowledge of physics could probably pull it off as well. Generally these people want to advance their education in the field.

If this describes you, a job posted for a data scientist by a company who produces data-driven products might be a good fit for you. Your chances are higher if you have specialized in statistics, mathematics, or something else related to the analysis of figures and calculations. Your skills could be just what they need.

4. General Analyst: Some companies ask for data scientists but are more focused on finding somebody with machine learning or data visualization skills. This means that the company already has a team of data scientists and just needs somebody to pick up the lighter tasks, which means it would be a great learning experience for you. It’s all right to apply for this, and it will give you some great insights into what a data scientist's job looks like. If you are comfortable with Pig, Hive, and other big data tools, you will be able to succeed at this job, despite its title.

5. Data Architects and Modelers: Companies are required to keep and organize more and more data, so they need people who can collect all of it from their various systems and structure it in a meaningful way. This data can then be used to find alert triggers, risk, and fraud. Architects and modelers work alongside development project teams to ensure that all system changes will be put into a format that can be sent to the data repository and then used for functions and reports.

Data Scientists in the Future

Today, experts are trying to figure out what role data scientists will play in the future. Will the job become irrelevant? Different fields of the data world believe that data scientists will become obsolete in five to ten years. Some think that it will take 50 years. However, some think that data scientists will always have an active role in our world.

These people who believe in the longevity of data scientists base their predictions on “export-level tasks”. The basic idea is that certain data science tasks are simply too complicated to be completed by an automation or robot, so humans and their natural innovation and creativity will always be needed. Robots can’t think outside the box when new data model methods are required to be built or applied for data interpretations when solving unknown business issues.

The other side of the debate believes that it is possible to automate all expert-level tasks, no matter how complex they are, within 5 to 10 years. Software will take over the tasks that data scientists currently perform. For example, look at Tableau - it is a software tool that can visualize using an application. Many second-generation data science companies are busy creating software tools that will automate data interpretation and help their overall workflow. Today, the issue is still up in the air. These software programs are far from being complete, so for now the data scientist is safe, and only time will tell whether they will stay that way.

Chapter 18: Big Data and the Future

You may have noticed by now that companies not only advertise their products or services online. Companies are also using the internet marketplace to communicate directly with you, the consumer. Thinking about what you use the internet for, you might realize it's for more than subconsciously being fed advertisements. You get online to research a company and what they offer, find detailed information about a product or service, and even find competitors and how their products and/or prices compare. You will also find more entertainment than you can possibly explore in a day and if you're looking for videos or shows – it's probably much cheaper online than paying for a traditional provider. When it comes to connecting people with people or people with companies, public or private, many companies are using the web to reach a bigger audience in less time and for less money.

Online Activities and Big Data

With online accessibility being easier than ever, it also means the amount of data being generated every day is growing exponentially – thrusting big data to the forefront and making it hard to ignore. Before anyone can even consider analyzing all this data, though, efficient methods of simply tracking the data need to be evaluated. There are massive amounts of data generated from automated feedback such as weather trackers, traffic monitors, and other transactions.

With all this data flying around there are several questions that must be addressed. Which sources are reliable and which are not? What data is relevant to a certain scenario and what can be disregarded? Sorting through piles of data can be overwhelming, but big data has opened the field for a new kind of business for people specializing in how to use big data to your advantage. The potential power of big data is what these companies sort through for you and advise on how to use it to your organization's benefit.

This has also opened the door for specialized programs that help to manage data. Apache™ Hadoop® is an advanced database management technology that moves beyond just consolidating the information to increase your profits and improve efficiency. The potential of this kind of database management promises giant leaps for businesses – assuming they are open to using the tools.

Instead of simply accumulating heaps of data, they will be able to actually *use* the data before them. For a business involved in education, it may drastically reduce the cost of education. For scientific purposes, such as meteorology, the comprehension of natural phenomena will allow us to have a better understanding of weather patterns. Businesses can increase productivity by cutting out repetitive processes, and the job market will create correlated datasets to help match job seekers more efficiently to jobs and employers that suit them based on skills of the seeker and required skills for the job.

Big data doesn't only help education or business become more efficient or more productive or increase profitability. As research continues to develop, there is potential for understanding and reducing crime, web security

improvements, as well as the ability to predict natural or economic disasters with some accuracy well before they happen. As long as the research is allowed to continue into how big data can be sorted and applied to the variety of fields it impacts, the world is on the brink of witnessing radical change – most of which will be for the better.

The Value of Big Data

As the amount of data continues to grow there will be innovators who want in on the party who will work to find ways to develop data management tools that will convert big data into economic advantages. In 2012 the market value of big data was worth $5 billion – in 2017 that's estimated to reach nearly $50 billion if we continue down the exponential growth path we've seen so far.

Security Risks Today

Based on information from 2012, of all the websites hacked during the year, 63% of website owners didn’t even know they had been hacked. Of those hacked, 90% didn’t notice anything suspicious happening on their websites. Half of the web owners learned about their sites being hacked from warning messages in the search engines they were using or from the browser itself.

Big Data and Impacts on Everyday Life

Improved Technological Functioning: Websites and applications will be able to function better in terms of usability, becoming easier to navigate and at the same time becoming much safer to use. Big data can be used to track fraudulent activity in real-time. Big data will clean up the visibility of an organization's website and open the windows to be able to predict attacks. There are these types of programs already underway to safeguard data from intrusions. The machine learning program MLSec (available at www.MLSec.org) uses algorithms – with supervision – to find networks that harbor malicious programs. This machine learning program has proven itself to be accurate 92-95% of the time in cases it has tested.

Higher Education: Big data can make higher education accessible to more people by decreasing its costs. Although the United States is considered to be the land of opportunity, higher education is difficult for the average person to pursue – since tuition costs rise at double the rate of health care. When compared to the country's Consumer Price Index, the price of college comes in at four times higher. With the innovations to make websites easier to navigate and safer to do so, the availability of online educational resources is also growing and becoming available to more people. The Khan Academy (www.khanacademy.org), Big Data University (www.BigDataUniversity.com), Venture Labs (www.venture-labs.org), and Coursera (www.coursera.org) are only a few institutions who offer online higher education at reduced prices – sometimes even free. What's great about these institutions is that the student is tested on how well they can implement the skills taught because the courses offer material that can be applied to the high-tech environment we all live in. For instance, Big Data University teaches Hadoop® and other technologies related to data.

Employment: Big data can also make it easier for job seekers who turn to the online environment to search for employment opportunities. On average, the number of job searches conducted online reaches nearly 1.5 billion per month. There are already websites that collect information about employers which can be used to match them with job seekers, like www.indeed.com.

Road Safety: For teens aged 16-19 in the United States, the leading cause of death is traffic-related incidents. It should be noted that 75% of these deaths are not related to driving under the influence of drugs or alcohol. This leaves the majority of accidents resulting from poor judgment. What does this mean when it comes to big data? As scientists continue to make advances in how computers work, big data will be able to 'learn' behavior patterns of drivers and even predict the moves a car will make at given points. The end goal of this research is to allow vehicles to communicate with one another and up to a distance of three cars around them in any direction. It is thought that advanced versions of this type of data collection will allow one driver to see small details such as the posture and focus of the driver in another vehicle near them. As outlandish as it may seem, Google's self-driven cars have already revolutionized the industry.

Predictive Opportunities: Hadoop® and similar technologies will provide businesses with the opportunity to analyze data with such speed and accuracy that they will be able to act quickly on opportunities, damage control, and other business-critical issues. Some successful Hadoop® users include Disney, Facebook, Twitter, and eBay and its demand is on the rise. A renowned market research company, International Data Corporation (IDC), predicted Hadoop's® software to be conservatively worth $813 million by 2016. Recorded Future is a technology company that provides intelligence to data analysts to secure their information. It allows businesses to capitalize on opportunities and anticipate risks using algorithms to define predictive signals. There are other companies that offer similar services and as technology in this field continues to develop, leveraging data for strategic use will become more prevalent in business operations.

Weather Prediction: Weather-related disasters can cripple environments and economies. The ability to better predict weather can allow people and governments to position themselves to sustain less damage – environmentally and economically – in the event of a natural disaster. In this case, a natural disaster refers to wildfires, droughts, flooding, damaging storms, and other instances of Mother Nature's wrath that are out of our control. However, data technology will soon be able to help predict certain events. For example, set to launch in 2018 the JPSS (Joint Polar Satellite System) will use sensor technology and data to predict the path of a

hurricane or a storm front long before it occurs. This gives the population that lies in the path of the storm an early warning system, and they'll be able to protect their property and evacuate the area in a timely manner. The future of predictive science has already been mentioned by major news outlets that recognize the guesswork of meteorology may soon be a thing of the past.

Streamlining Healthcare: Not only can big data bring improvements to the healthcare sector by making services more efficient but it may also provide capabilities to customize services tailored to the individual. McKinsey & Company is a management advisor to many big-name businesses. They estimate between 50% and 70% of business innovations rely on the capacity to capture customer data rather than analyze the data. McKinsey & Company depends on quantitative and qualitative data to be able to provide constructive advice to management. They also estimate that nearly 80% of healthcare data is unstructured. With the recent trend of the medical field using big data in creative ways, there is no doubt that improvements in services will be positively impacted – to the tune of nearly $300 billion in value added annually. The other side of adding value is that medical expenditures are expected to be reduced by at least 8%. When patients are treated effectively, and there is data available about a patient, big data allows caregivers to give medical advice based on evidence. Beth Israel Deaconess in Boston is implementing an app for smartphones for medical caregivers that will allow them to access nearly 200 million data points – which equates to the medical records of about two million people. Rise Health utilizes data that has been analyzed from multiple dimensions, then aligns it with provider goals to improve healthcare. For the medical field, big data can bring speed to innovation. The Human Genome Project – which took thirteen years to complete – would merely take a few hours to complete today using previously unavailable technology.

Chapter 19: Finance and Big Data

There is definitely no short supply of data – in fact, there is probably an excess of data available today when accounting for traffic running through social media, transactions, real-time market feeds, and other places. This means there are explosive amounts of data available for the financial sector. The variety of data available and the speed of accessing data have also grown astoundingly. This can create two scenarios for organizations – they either harness the data and use it for innovation or stand by in awe at the massive amounts of data they are presented with. Since businesses are in business to succeed, they have taken the approach of hiring data scientists to help them sort through data. A data scientist takes a data set, analyzes it from all angles, and uses it to make inferences or predictions that can possibly lead to beneficial discoveries.

How a Data Scientist Works

A data scientist's goal is to identify fresh data sources for analyzation and then build predictive models. They may even run simulations of possible events to see the outcomes. All of these factors help the scientist to see a possible reality of a situation before implementing new innovations or discoveries. This way of using data allows organizations to foresee trouble and prepare for it before it becomes uncontrollable and may even show future opportunities as events play out in the real world.

Many data scientists use software such as NoSQL, Apache™ Storm, and Hadoop® to sort through and identify non-traditional data like geo-locations or sentiment data, and then correlate that data with traditional data, such as trade data.

Data scientists also want to ensure the future of their work, so they take precautions to ensure there is an available backlog of relevant data for future reference and find a way to store it cost effectively and safely. Their expertise and ease in storing data are further enhanced by the development of technology-based storage "facilities" – for example, cloud-based data storage. There are also tools of analysis that are quite sophisticated for their cost-effectiveness, some even being offered for free online as open-source tools. Data scientists have a wealth of financial tools at their disposal which are being used to improve and transform how business is conducted.

Understanding More Than Numbers

It might seem that a data scientist is concerned with concrete numbers and figures, solidly objective data. However, there are services and tools available that help them analyze people's sentiments – also known as opinion mining. A few examples are Think Big Analytics, MarketPsy Capital, and MarketPsych Data. These firms and programs analyze text, language processing, and computational linguistics to consolidate the information into usable material to improve business.

Applying Sentiment Analysis

Sentiment analyzing firms build and use algorithms to compile relevant data from the online marketplace, for instance, from Twitter feeds. These feeds provide a massive amount of data when concerned with specific impactful events such as disastrous weather or terrorist attacks. These feeds can also be mined by organizations to find trends when monitoring new products or services or responding to widespread issues that might affect the image of their brand overall.

For call centers, sentiment analysis can examine recorded phone calls to find ways to reduce customer turnover and recommend ways to improve customer retention. Many of today's businesses are customer-focused so having a service to analyze data about how or what customers feel toward a brand is a key factor to the success of the business. The demand for this is so great that companies have emerged that focus on providing this kind of service – gathering data, identifying sentiment indicators, and selling their findings to retail establishments.

Risk Evaluation and the Data Scientist

Data scientists have found ways to use the variety, frequency, and amount of data available in the online marketplace to enable finance companies to offer credit online with very little risk. Sometimes investors won't or don't access credit because of the lack of a way to give them a credit rating. However, it is essential for lenders and financiers to be able to measure the risk when considering handing out credit.

Internet finance companies have emerged thanks to big data and the scientists that analyze the data, developing ways of managing risk to be able to approve online loans. A good example of online lending enabled by big data analysis is Alibaba AliLoan, an automated online bank that offers small and flexible loans to online entrepreneurs. Recipients are typically creative individuals with no collateral, which makes securing loans from traditional banks nearly impossible.

Reduced Online Lending Risk

We'll continue to focus on more detail about AliLoan and how they're able to manage risk through an online forum. Alibaba monitors e-commerce and payment platforms to understand customer behavior and their financial strength. They will analyze a customer's ratings, transactions, shipping records, as well as other information, and can generate a loan cap for the customer and the associated level of risk. Alibaba also uses external third-party verifiers to cross-check their own findings. Other resources might include tax records or utility bills. Once the loan is granted, Alibaba continues to track customer behavior and spending patterns of the loan provided to them and will also monitor business development.

Lenddo and Kreditech are other companies that utilize data scientists' abilities to manage risk for recipients of automated loans. These companies have developed credit scoring techniques that are helpful and innovative in determining a customer's creditworthiness. Sometimes even data from online social networks is used to gauge a customer's creditworthiness.

The Finance Industry and Real-Time Analytics

The finance industry can't just rely on having a compilation of data that is easily accessible. What really matters when it comes to data is *when* it's analyzed. It's just not possible to rely on the data itself to make an informed decision – it must be analyzed at regular intervals. Sitting on compiled data for extended lengths of time decreases the likelihood to seize critical opportunities and increases the possibility of things going awry. However, by using the skills of data scientists the lag in time that once was an issue in the finance sector is no longer worrisome. There are several benefits to having real-time analytics as a function of your business.

Reducing the occurrences of fraud may be one of the most important contributions of real-time data analytics when it comes to protecting the consumer's information. Banks and credit card companies, along with others, have made it common practice to prioritize securing fundamental account information with the use of big data. They want to ensure your employment status is accurate and your location is up to date in the case that something occurs that is out of the "normal" behavior pattern. "Normal" behavior patterns include characteristics such as spending patterns, account balances, credit history analysis, and other general – yet important – details. With real-time analytics, the company is able to notice almost instantaneously when something seems out of character for a customer and can trigger a flag on the account. Usually, the account will be suspended in case the activity is fraudulent and the account holder is notified of the activity and suspension.

Big data real-time analysis is also beneficial for improving credit ratings. An accurate credit rating literally requires amassing an ample amount of current data, placing less focus on historical data. Since credit ratings are such an important part of securing almost any kind of loan or asset, it's important that the applicant has an accurate and up-to-date credit rating to determine the level of risk of an individual. Being available in real-time gives a reasonable assumption about the financial capacity of a customer. With the online marketplace readily available, many categories used for calculating credit ratings are easily accessible. These can include

transaction histories, business operations a customer may be engaged in, and other assets that are held by the customer.

Another benefit to customers is the ability to provide accurate pricing for various products and services. For example, in the case of a loan, the customer may be able to secure a better interest rate on borrowed funds if their credit rating is higher than when they previously applied for financing. Tools offered by car insurance companies are another service that uses real-time data analysis to warn drivers of accidents or traffic jams in their path, weather conditions, and other factors that may help reduce accidents. Customers who choose to use these tools are typically offered benefits on their insurance policy such as discounts or other rewards.

When data analytics is utilized, the cost of business is reduced across the board. In today's business arena, it is mostly attributable to real-time analytics. Many major financial institutions turn to PriceStats, an online firm that monitors prices for more than twenty-two countries to provide real-time inflation rates for those economies. PriceStats was originally founded to monitor only a few countries in South America but now can monitor the international market. The information provided by PriceStats allows businesses to tailor their actions in accordance with inflation rates around the world.

How Big Data is Beneficial to the Customer

Many banks and other financial institutions pay to acquire data from retailers and service providers. This is the foundation of data importance, especially when the capacity to analyze and use the data exists. In reality, all data is important depending on the reason someone wants it. Storing data for the sake of having it is a waste; it becomes an unnecessary distraction when it simply sits in a system, and no one accesses it. Even the influx of massive amounts of data is useless if there is no program in place to process and analyze it. That's why having a system in place that can sort data among and between departments is critical to share relevant data with appropriate departments rather than disregarding and throwing it aside. When necessary, irrelevant data can be discarded if it serves no purpose to someone in particular or an entire department.

This chapter places great emphasis on the data scientist because knowing how to handle and utilize data is a skill that not everyone innately possesses. Data scientists have learned and refined their skills to be able to help businesses deal with the massive amounts of data available to them. Big data is especially helpful when it comes to customer-based industries.

The companies that pay for data acquisition aim to gain a full and well-rounded perspective of their customer. Many businesses use the term KYC – Know Your Customer – to delineate which customers may pose a risk to engage in business transactions with, known as customer segmentation. The information obtained to make this term relevant is credible because it is acquired from big data to form an understanding of the customer. This concept is supported by a leading figure with IBM Analytics, Sushil Pramanick, who is also the founder of TBDI – The Big Data Institute.

Customer Segmentation is Good for Business

By segmenting the customer base, it is easier to meet their needs. Categorizing customers based on similar financial capacity and age, similar consumer tastes in the same income bracket and from the same cultural background narrows the focus to meet their needs accordingly. You can design products targeted at their preferences, you can tailor advertising techniques to appeal to them and retain them as customers, and you will be better able to re-engineer or invent products focusing on specific groups.

Chapter 20: Marketers profit by using data science

The marketing industry has been involved in gathering trade statistics for a long time and the amount of data collected to date is very substantial. This data has not only increased in capacity but also in range, quality, and accuracy. While there is an acute awareness of the significance of this data by retailers, they are not so clear about its many applications and how to get the best usage from the system. Most retailers are unable to handle this data correctly. Basically, they do not know how to collect, analyze, and apply it appropriately because they are not familiar with the science behind data gathering and usage, and the various interactions that make it invaluable for higher profitability in their areas of business.

Specialists in data science can convert haphazard, unstructured data clutter into a well-organized system which the marketer can operate with ease. Because of the inability to apply data correctly there has been a rising demand for data scientists. The retail industry has recognized the advantages of employing individuals who are skilled in data analytics, and have seen their profits rise exponentially through expertise in data handling. This has convinced retailers to work with scientists whether it is internally or externally. A huge competitive advantage could be gained by the smart use of data analytics in the retail business. Retailers who are able to organize, assimilate, and apply their data will maximize their marketing tactics and optimize their chances of success.

Reducing costs to increasing revenue

Sales are the most vital aspect of any business. Pricing must be realistic and yet profitable, and appeal to consumer requirements. It is always a priority to improve customer experience online or offline and this can be done accurately by using predictive analytics. Retailer requirements should be adapted to suit market demands and insightful data from social sites, call centers, product reviews, surveys, and any other means of customer feedback will be helpful in outlining best retail practices. Improved displays of promotional presentations and other marketing resources can be accomplished through the use of heat sensors or image analysis for retailers to obtain a greater appreciation of consumer behaviors. Scrutiny of video data can also be used to ascertain shopper tendencies and this analytical data will help identify cross-selling possibilities.

Data from internal and external sources can create consistent profits for marketers. Information may include raw material costs and competitor's prices, economic, weather, and traffic reports, seasonal products and high or low purchasing periods. Sales must be founded on the conception of a systematized, planned strategy and the detailed statistics of this data guides retailer's to create a perfect marketing plan. Without sufficient, accurate data there can be no foundation for determining and evaluating supply and demand. Revenues can be improved quicker due to detailed investigative market reports, and by making use of product sensor devices to convey significant statistics regarding product and purchase in real time.

Marketing practices may cause retailers to cut costs by making special offers through mobile messaging. They can deliver these deals which is more economical than general sales. Since retailers are able to communicate in real time they are able to modify pricing structures which change from second to second. Retailers can maintain an edge by comparing competitor pricing data which is available to them through sensors. Being aware of how a product compares, positions the seller and

buyer to make an advantageous deal. Segment consumers markets can be spotted by looking at analytics data on any number of information systems. Making use of behavior analytics helps to improve retailer's return on investment policy and shape effective marketing proposals as a result.

Data science helps when retailers need to track and manage inventory in real time. Supply and logistics sectors also profit from data science which permits retailers to improve distribution means through GPS enabled big data. This allows the choice of the most inexpensive, safest, and swiftest routes. Supply chain logistics also benefit from structured and unstructured data as dealers are able to prefigure prospective consumer needs well in time. The analysis of accessible market information on big data creates a chance for retailers to bargain with suppliers for better deals. Gathering and applying big data for insightful retail practice will meaningfully define the bottom line in all transactions.

Chapter 21: Use of big data benefits in marketing

One part of human endeavor that benefits greatly from big data usage is marketing. There is a constant need to know more about this area and as such, both marketers and data scientists keep on investigating this subject that has real value when applied properly. A few years ago, it would have been hard to describe consumers in such detail as is now possible with the assistance of big data. Information about every aspect of human life is expanding rapidly and when this huge stockpile is assimilated by big data, all our actions can be logically and accurately quantified. The effects of big data analysis on marketing practice are more extensive now than can be imagined. Each aspect of this subject can be determined and analyzed from the preferences of consumers, availability of products, pricing, supply, logistics, and every other conceivable aspect connected with marketing.

Big Data can define targeted consumers with a detailed accurateness that was impossible only ten years ago, when the best that marketers could do was study results of mailing promotions, or check out subscriber details from newssheets randomly distributed. Today's marketers can have any amount of information about people's behaviors. Usage of social devices is of great interest to marketers as they can observe time spent on several sites, buying history, individual inclinations, founded on social media posts. Digital clicking habits allow marketers to position advertisements for better visibility to increase hits on these ads and, subsequently, raise revenues. Now that big data analytics is widely used by marketers it is difficult to do without the help of this brilliant system.

Google Trends does all the hard work

The use of Google Trends allows individuals and organizations to optimize big data without any financial investment in the analysis. It has emerges as one of the most extensively used and efficient facilities that offers beneficial methods of incorporating big data into marketing procedures. This public web facility displays worldwide rates of searches as well as the areas of investigation. It also shows comparative data, and how breaking news effects search status. It is the most dependable way to find out what is trending in the world in general, or in an industry or product in particular. This tool is invaluable to marketers as it can show who is interested in what, and why, and when. Marketers can be quick to pick up latest trends and produce what customers want. Google Trends has now the development of promotional strategies and thinking up innovative design concepts for products easier than ever before. In fact, Google Trends does all the hard work and marketers simply reap the rewards.

The profile of a perfect customer

Marketers used to have limited or no proper information about their prospective customers and had to calculate product demands using a minimum of facts and a whole lot of guesswork. Sometimes it worked and most times it didn't and many traders went out of business before they had a chance to recover their investment. Today, assisted by big data, we are able to know most details about customers as well as market positions. Having this kind of information makes it possible to participate in a very competitive arena. The better we know our customers the more our business will prosper because this person is all important to the sales process.

Leads must be qualified in order to identify prospective customers. Make sure you look at products from a consumer's perspective Ask yourself if the product would be of interest to the customer and why would they buy from you instead of your competitor? Check out locations and seasonal sales prospects of the area of residence. What is your customer's buying habits? If you can answer all these questions you now have a customer profile. How you work with this customer will be determined by your services and motivation to build up your customer base. Studies show that adding customer profiles to big data can raise the success of marketing by about thirty percent.

Personalizing real-time

Being current is vital for marketing since all business activity requires immediate action. Quality merchandise, prompt service, and speedy delivery is vital for success and the use of big data will help realize all these objectives. Big data makes it possible to register consumer habits in real time which means instant facts are available to enable instant decisions. Customer wants, needs, likes and dislikes become known so that they can be attended to on the spot. Having relevant information about specific

groups makes it possible to design successful promotional programs aimed at engaging these consumers.

The greater the understanding marketers have of their customers, the better their services will become. The can find out wat is important to each customer and adapt their responses accordingly. Special offers will appeal to customers because these are pertinent deals that have been developed from their personalized data. Such data leads to personalized service founded on real-time customer performance. This real-time activity will promote the popularity of brands, build up interactive associations, produce competent leads, and settle lucrative dealings every time.

Ascertaining correct big data content

During its early days it was never definite whether certain segments of content would functioning correctly or not. A lot of guesswork was used to arrange content that was thought to drive customers to buy up the market. That didn’t happen and many and a loss was suffered due to incorrect or inefficient content. Today, big data affords businesses the most operative methods to interact with consumers. It is essential to understand data handling necessities, and choose the correct big data drivers. Big data content should also be chosen to suit the system. Accurate and informative data can be turned into great content. Big data helps us realize what succeeds and what does not. It can also assist us in finding applicable solutions and making educated decisions.

Marketers can use data content to predict the future requirements of their customers and develop campaigns based on this information. Creating content inspired by this information will impact sales productively and there will be no conjecture when using strategic applications. Content and strategies may be tested in advance to see how consumers respond, before an actual promotion is launched. Big data has an important and useful role in supporting content procedures because it delivers useful statistics to accomplish a more expansive interaction and provision of services to geolocations that would otherwise not have this access.

Lead scoring in predictive analysis

Marketers grade data in order of priority to gain insights into a consumer’s intentions and determine productive leads. Lead scoring processes use numbers to classify customers, based on available information. Predictive intelligence utilizing lead scoring numbers make the work of marketers and sales crews much easier and more profitable. Lead scoring calculates the effects and outcomes of consumer habits and this data makes predictive analysis is one of the most powerful tactics dealers can apply to boost market opportunities.

Whereas acquiring new leads can be an expensive undertaking, by using various big data sources and solutions, there can be up to a fifty percent decrease in lead costs. By using internal and external data sources, marketers can cut costs of acquiring leads and define and maximize a superior lead scoring plan. Charges and exchanges will achieve better rates by giving precedence to data from lead streams. Quality leads will be established by concentrating on best projections that will determine where focus and effort must be applied. This will save time and money and keep you ahead of the competition.

Geolocations are no longer an issue

The geographic whereabouts of consumers has been made easy by the accessibility of big data. Now marketers can pin point areas from which direct and indirect shopper activity emanates and create leads to manage these various customers and locations. Connections can be personalized to ensure service excellence. Technological advancements permits marketers to compete and by improving consumer satisfaction, they can raise their own business status. Retailers are using location sensing technology which focus on navigational information, location based promotional offers, and reviews of nearby products.

Marketers have come to depend on innovative and upgraded digital devices that gather on-the- spot data from customers in any geolocation. This instant information influences the way in which consumers are served and heightens their shopping experience. Getting instant feedback from customer localities can rapidly check any irregularities, deliver a more personalized service, and improve interaction. Data regarding customer satisfaction can be used by marketers in various forms to improve their services. Marketers already possess the tools to get the utmost advantage from their analytics setup, which will enable the formation of winning strategies regardless of geographical locations.

Evaluating the worth of lifetime value

It may seem callous, but marketers see consumers a mere numbers and each number has a market worth. This impersonal approach is required in the assimilation of business data when the customer is just another cog in the wheel of commerce. Big data doesn't differentiate when information is being measured to find the best marketing practice. Marketers have to know the significance of a customer or a demographic as a whole, to calculate lifetime value. Businesses depend on revenues to sustain them so this feedback is import to their survival and progress. Such information helps decisions about the advantage of special offers and discounts and whether such incentives will attract additional business. Unsuccessful decisions can be avoided when there is a knowledge of lifetime value.

Having a clear picture of lifetime value is important especially in the early stage of business when these numbers are so important for all planning. Big data can calculate the worth of consumers over a specific period of time. This estimate will include the amount of money a customer will spend, and this is a very important number. Quantities must work for business or else be changed to factor in to an effective strategy. Innovative variations can be introduced that will remedy unsatisfactory value and this could work to change number worth. This knowledge will assist in creating new possibilities to drive business up and increase a revised lifetime value. It is these numbers that will determine business success so careful attention must be paid to lifetime value numbers

Big data advantages and disadvantages

Marketers spend huge amounts on technology to improve their business opportunities. High tech tools scrutinize consumers for any marketing advantages while and advertising aims at getting their attention. All these strategies come at a high price and many marketers sometimes wonder if information analytics is worth the expense. No doubt there are some advantages and disadvantages when big data is used to assist marketing processes. One big advantage is the enormous amount of information this system harvests. On the downside, some of this data may be inaccurate and misleading, and if used could cause erroneous solutions and business losses.

Surveys may answer questions but on the other hand, solutions to a problem may not evolve. New information and updates are added every second but these may not tie in with previous data. Big data is an enormous system. Being able to extract relevant material from this colossal accumulation of information could prove difficult, and unsuitable content may inadvertently corrupt statistics. This could result in the loss of old and new customers, the failure of promotion efforts, and waste of funds. Skilled usage of big data will help create powerful systems that identify business and market opportunities that increase numbers and revenues.

Making comparisons with competitors

In the past, market information was difficult to gather and was kept secret from others in the business. Statistical data was protected and there was no way an outsider could see this information and make advantageous comparisons with competing markets. Data analytics now allows us to see our own indicators as well as the figures of competitors. By using social devices and analytics implements any enterprise can see their position in the market along with that of their business rivals. Information from search engines can keep a marketer ahead of the competition. The openness of this display helps rather than hinders by showing a clear picture of the market position and consumer requirements, as well as the state of services of everyone in a particular commercial enterprise.

Armed with this detailed market information, business persons can participate in a transparent manner, using smart strategies to make their product special and convince consumers to accept this fact. Visiting websites and scrutinizing activity on social sites can keep you abreast of trends and human nuances, which a mechanical process might miss. Remember, praising your product is not the way to win over customers. Through data analytics, what the customer really wants becomes obvious and markets demand their wishes are catered to on their terms. Business can be successful by distinguishing your product and services by valuing your consumers.

Patience is important when using big data

Collecting data is time consuming and analyzing data can also be labor intensive and expensive. Big data can assist marketers but its benefits may not be obvious in minutes, or hours, or even days. In spite of using outside help like Google Trends, more time, money, and effort will have to be spent before the point of perfect use can be expected. This is why a lot of patience is initially required in handling this system. Information gathered has to be from a reliable source and must clearly represent the area to be analyzed. Data must have certain features to meet the requirements of most investigations. Levels of information must be recognized and comprehended, should not fluctuate in similarity, and be evenly dispersed. To realize its true worth, big data must be used carefully, skillfully, and patiently, observing all the essential rules of statistics. Do not expect that analytics will suddenly drive your business to instant success. Rather, use the information to help grow your enterprise steadily, and gradually learn how to work the system to your best advantage.

Chapter 22: The Way That Data Science Improves Travel

One of the travel industry's favorite things is data. Well, it might not be exactly its favorite, but the world of travel has always collected lots of data, even when stakeholders aren't actually looking for that information. Storing all of that information has been a challenge, but now getting the data analyzed and improving how that information is used is another major challenge.

The amount of data that can actually benefit the travel sector is already huge because all facets of life are involved in travel. Culture, air fares, security all over the world, hotel pricing, and the weather are all taken into account when people travel. Because of how much information is involved, the entire sector cannot ignore data science and how it can help them deal with all that data and optimize its use.

Data Science in the Travel Sector

Big data has had so many different uses in other sectors, but there are some very specific ways that it can be optimized for the travel sector. This optimization is entirely possible for those that have a data scientist working with them. The list of those benefits for the travel sector is quite long:

Ability to track delivery of goods.

This ability is entirely possible, regardless of how the shipment is moving. It could be anything from a freight shipment, someone on the road, or even a voyage. Major online companies, like Amazon and Etsy, use this tracking for their benefit. The tracking can be used for not only the customer but the seller as well. The customer being able to track shipments usually inspires more confidence in the company and how they get their deliveries out to their customers. For online companies, this means that the sellers are more likely to come back, which is great for everyone.

Analysis done at each data point.

Doing analysis often is a way of increasing business. Different stakeholders will have the ability to share information that's gained from the analyzed data. This will ensure that nobody is getting information that is irrelevant to what they're doing or full of redundant values.

Improving access to travel booking records.

Mobile phones are making inquiries, bookings, and payments even easier than before. People can schedule travel and stays even easier to get organized.

Ease of access to information.

There's easy access to information like profiles, itineraries, feedback, and more data from internal sources. There's even easier access to information from social media reviews, weather, traffic, and everything else that would come into play when working on creating a trip or working out what you want to do. The travel sector works well with big data since it merges data from internal and external sources in order to find solutions to problems happening right now. That data helps everyone involved in the process anticipate what's going to happen based on the past.

The travel sector has been able to drastically cut their costs of operations because they're better able to make good choices based on the data that they're analyzing in a timely manner. Data science is making this entirely possible. Big data is interesting to the travel sector because that information can make them even more lucrative, which it is projected to become even more rewarding in the future.

People are traveling even more. There are projections that the growth will be so sharp that by 2022, the value of the sector will be 10% of the world's GDP. The travel sector realizes that the value of big data and optimization of it is important because it will help them get the best business intelligence to work with. This knowledge will help them make the most money that they can.

Travel Offers Can be personalized because of Big Data

A few years ago, travel was based on general classifications of customers. If you were in the high-income bracket, then you were given recommendations for certain facilities, but if you traveled regularly with kids, then you were targeted for different offers. This was true for almost every group of people. This method did increase revenue slightly, but the impact of that method was smaller than what is possible now with big data.

Big data is becoming more and more significant in the travel industry today because it enables travel companies and agencies to tailor offers for their individual customers. They'll be using complete views of customers to look at their different customers. Offering ultra-personalized packages and facilities to individuals is not that hard when big data is at play.

The data sets that can help create a complete view of the customer are many:

Data based on customer's reading behavior

How they look at this is really based on how often people go online and what websites they're looking at often.

The customer's posts on social media

An analyst could establish what travel people are talking about. They can also look at who is talking to other people about travel, whether it's through social media messages or through the reviews that they (or their friends) have posted.

Location tracking data

Itineraries and connected data

Customers' shopping patterns in the past

How people are using their mobile devices

Information and data gathered from image processing.

This list is rather short, but that's because there's just too many places that information come from to create a complete customer profile. Anything

that is even remotely related to travel would be added to this list. Depending on how new or old the customer is will also change what data sets are taken into account. If you're looking at buying habits and other historical patterns, then the target you're going to be able to work with best is one that has already bought from you. But if you're looking at a new customer, you will want to focus on the data sets about their behavior. There are some data sets that overlap most groups and others that are much more useful for smaller groups. It just depends on who the company is looking at.

Safety Enhancements Thanks to Big Data

It's quite easy to ignore how much data also helps with safety in travel. The information goes from those who are coordinating the travel to the pilots, to the control tower who speaks back to the pilot, then to the driver, then to the fleet headquarters, then to traffic headquarters, then to all travelers. Data passing through so many hands is actually one of the life savers of the travel sector.

Vehicles of all kinds, including planes, have sensors attached that detect, capture, and send information on a real-time basis. This information can be used to see all sorts of things about the travel experience. It may be specifically about airmanship, the behavior of a driver, the state of parts of the vehicle, the weather, and much more.

A data scientist can design algorithms that would allow travel institutions to be able to tell when a problem might happen based on the information that the sensors are gathering. Sometimes, a data scientist can even create something that will prevent the problem before it actually happens. There are many common problems that big data is helping companies address:

Whether or not there is a faulty part on the vehicle.

Being able to detect this information helps the company replace parts before the damage inflicted by it is too hard to salvage. Replacement and repairs are also the best way to prevent accidents.

To tell how well drivers or pilots are doing.

If the sensors detect that steps are being skipped or done out of order, then the company can pull the person and send them through some additional training. A company will probably do this to maintain their high standards for performance.

Identifying problems midflight.

There are some problems that crop up while a plane is in the air. There are some problems that can be fixed while the plane is the air. However, if the problem is hard to address in the air, the company can bring in a maintenance team for when the plane lands.

The ability to deal with problems so quickly shows just how crucial a role big data plays in the travel sector right now. It doesn't matter how you're traveling, the exchange of information allows you to get your destination as safely and quickly as possible. If everyone is in the loop, then all parts of your travel plans will move smoothly, weather permitting.

How Up-Selling and Cross-Selling Use Big Data

Big data can be used in not only up-selling but cross-selling. Vendors up-sell customers when they try to convince you of buying a pricier version of a product that you were looking for in the first place. Cross-selling, on the other hand, is when vendors try to convince you to buy something in addition to what you were looking for.

If you start looking at flying somewhere, you're going to start seeing offers that complement the one that you were originally looking for. There are many ways that this could happen:

There might be an offer for a hotel (especially ones that are partnered with your airline of choice) in the area that you are going. This would be cross-selling.

You could be booked as Economy Plus or the equivalent for other airlines. This would be up-selling. They are offering you extras like legroom and room to recline, but you have to give them even more money.

You might also get an offer for food, courtesy of a steward, and receive a coupon to use.

During the flight, you'll likely see ads during the in-flight entertainment that are trying to push you towards specific companies.

Companies, especially those in tourism, are using big data to cross-sell with each other. Most often you will see this with airlines, hotels, tour van companies, and taxi services.

Chapter 23: How Big Data and Agriculture Feed People

While agriculture may seem old fashioned, big data can have a huge impact on the industry if we put them together. Agriculture is one of the most important and biggest industries there are in the world, and still helps tons of people. We have come a long way from farmers just plowing their fields, and technology can really help even more in the process of growing crops.

Agricultural machines have been used for decades, but using data analysis for improved efficiency in the industry is still a new idea. Right now, we are able to collect so much more information about agriculture. You can take into fact the soil elevation and look at final yield along with so many other factors and data and analyze the information to find certain metrics that can help predict facts about future crops.

While much of this possible and this information has been known by farmers, it's hard for a farmer to take the raw data they're collecting and make it useful. Agricultural companies are now hiring teams of data analysts that are able to help not only the company but the farmers make educated decisions about how they deal with their crops for increased yields and final revenue.

How to Improve the Value of Every Acre

Revenue is one factor that every business is interested in. Everyone's doing what they do for the money, in one or another. Farmers are no exception to this. They need to make a living and big data analysis can help them with this.

There are tools, like Climate Pro, that can give farmers information that they will in turn use to increase the revenue that they gain per acre of land. Part of Climate Pro called the Nitrogen Advisor, allows farmers to track the nitrogen in their soil and advise them on how to move forward. The people that created Climate Pro believe that just Nitrogen Advisor can increase a framer's revenue by $100 per acre.

There's another tool called FieldScripts that allows farmers to increase their yield of corn. The tool breaks up the field into small sections and tells farmers how much corn to plant in each part of the field. This is based on how the soil is. With this tool, farmers can reach optimal amounts of corn being planted without having to guess. It can even be integrated with other tools that are used for planting seeds. This technology is revolutionary and would have been unthinkable in the recent past.

One of the Best Uses of Big Data

While the uses of big data may seem sketchy and only involved with money, there's way to use this information to feed starving populations. India has millions of acres that could be worked on, but the land is worked on by individual farmers that cannot use or collect vast amounts of data.

There are companies, like CropIn, that think big data analysis can help find ways to optimize the agricultural process. Indian farmers could especially benefit from that kind of data analysis that would allow them to make better decisions when it came to their crops.

How Trustworthy is Big Data?

Farmers put their entire business at stake every year. If their yield is low or crops are destroyed by natural disasters, then the farmer and their family are in trouble. This makes farmers slow to trust anyone when it comes to their crops. Trusting big data analysis might be too hard for some farmers because of how vague it is.

Companies that provide big data analysis, like Monsanto, are hated by much of the world for scandals that could be responsible for damaging and poisoning our food. This leads to even less trust of big data. Because of all the mistrust, it will take a long time before big data and agriculture are fully integrated together. Big companies will have to spend a long time convincing farmers to trust them, especially since the companies are so demonized right now.

Sooner or later, the world of agriculture will be improved by big data. It will happen eventually, but it may take a lot of time.

Can the Colombian Rice Fields be saved by Big Data?

There is one specific place in the world where big data could improve the agriculture of a whole nation. In Columbia, there are rice fields that have been facing inexplicable decreases in yields over the overs. They had actually been having increases in yields until 2007 where the decline began. Regardless of what they tried, they could not stop the decline.

For a while, climate change was declared the main culprit of the decline, but no one could find evidence that it was that. In fact, no one could figure out what exactly was the cause. Although there were concerns about sharing the data on the rice fields' yields, government agencies began surveys and had data collected.

From the analysis of the data, they found several ways to consider what was happening to the field. These inferences varied from region to region, but they could now at least take a guess at the cause. The city of Saldana appeared to have less daily sunlight and that was decreasing yields. However, the city of Espinal wasn't keeping up with the increasing nightly temperatures. With this information, farmers could adjust to the needs of their individual regions. What was going to help in Saldana wasn't going to help Espinal and vice versa.

The way that data analysis was used here could revolutionize the industry. If this method was used in the entire world, then all fields could be optimized and we could feed more people more efficiently. Using big data analysis for crops is likely to feed more of the world with less cost. It's a win for everyone.

Up-Scaling

Big data has been shown through small, successful projects to be able to help farmers. However, that data needs to be taken to agriculture in a larger way. In future years, more and more companies will be analyzing agricultural data. In turn, more farmers will be able to use the information to increase their yields and profits.

Chapter 24: Big Data and Law Enforcement

The reduction of financial resources due to changing priorities is among the many challenges that our law enforcement agencies face. This reduction of resources translates into the reduction of manpower available to the various police departments. However, though their resources are being reduced, law enforcement agencies are still expected to be able to maintain a similar level of service, forcing innovation when finding new methods of fighting crime.

Big data signifies a possible turning point in forensic investigations. However, many law enforcement agencies are not making full use of the potential of these data analysis tools. Many agencies work independently from one another, and they maintain their own separate data systems. This method is not an economical way of utilizing the resources available to them, as oftentimes, law enforcement agencies are not given enough resources to begin with. In fact, along with the rising crime rate in certain areas, the manpower of certain agencies have been reduced. In order to make the best use of their limited resources, these agencies should consider pooling their resources into a data networking system. This system would allow law enforcement agencies to share data between themselves, allowing this to help solve crimes and assisting them in being better able to protect the people. Big data helps these agencies find ways to reduce crime rates, even at a time that many agencies' resources are being reduced. Big data allows agencies to streamline their operations while still producing the needed results. Data platforms can do this in many ways, in almost any sphere that they are placed. Data platforms can help with human resource development, in criminal database management, or even with plate identification systems. No matter what platform is used, as long as the data system provides a foundation which helps with fighting crime, be it through improving the internal workings of the police department or through dissecting, analyzing, and developing crime-fighting techniques, big data can be a huge help in making our world safer.

We must remember that law enforcement is the foundation of the safety of our society. Without proper law enforcement, a person may not feel safe when walking at night, or in extreme scenarios, even when walking along the street in broad daylight. However, recent developments have shifted the public’s perception of law enforcement to the negative side. There is this idea of rising police brutality, along with an increase in crime rate. Now, more than ever, there is a need for our police forces to innovate, as the demands upon them increase, while the support given to them continues to decrease.

Data Analytics, Software Companies, and Police Departments: A solution?

As mentioned earlier, once the proposed data platforms are put into place, data analysis may be applied at a much more detailed level. This will allow law enforcement to improve the rate and percentage of crimes solved, allowing them to curtail criminal activity more easily. These data analytics may even be able to predict potential criminal behavior, and allow police to stop some criminals from breaking the law or harming others to begin with.

Microsoft is an example of a company that works closely with some police departments to prevent and solve crimes. There is even one police department that uses multiple streams of data to receive live camera feeds, 911 calls, and police reports in order to better fight crime. They employ these methods at any level of crime, allowing them to react quickly and develop strategies to prevent crimes from even occurring. They use certain technologies to assist them in doing this, such as Microsoft Azure Data Lake, Windows Speech Recognition, and Microsoft Cortana Analytics. Microsoft Azure is software that allows forensic specialists to store data regarding criminals and their activities in a central database. This data may be used for whatever forensic analysis or data processing method may be appropriate. This central database would allow police agencies around the world to create a network of shared data. This would greatly assist in tracking criminals, especially terrorists. This data may even be used, through predictive analytics carried out on historical data, to extrapolate the next target of terrorists, and may allow law enforcement agencies to prevent attacks from ever happening in the first place. Microsoft Cortana Analytics allows police departments to use machine-learning algorithms to rapidly forecast certain criminal scenarios. These specialists can also employ this method to enhance the decision-making process by assisting in choosing the optimal solution when a crime takes place. This tool also allows crime departments to automate certain decisions, which allow them to better account for rapidly changing criminal factors, improving their capabilities when it comes to preventing crimes, or halting a crime in progress. Windows voice recognition allows the police to dictate commands to the computer, enabling them to rapidly find pertinent information such as

warrant histories, DUI violations, or even ticket violations. Having this information at their fingertips allows police officers to prevent crime or even just traffic violations. The ability of an officer to reduce threats or diffuse situations or dangerous events is greatly enhanced through this software.

Analytics Decrypting Criminal Activities

Law enforcement agencies are able to make use of advanced analytics to interpret certain criminal patterns, such as homicides, domestic violence incidents, or theft. The availability of these data tools allows these agencies to better sift through massive amounts of data. This even allows police to identify the patterns of a burglar. It allows them to develop a better idea of the criminal's modus operandi, allowing them to answer questions such as when, where, and how a crime is likely to take place. This allows officers to position themselves in such a way that they may be able to prevent further crimes from taking place. Certain crimes tend to happen in specific areas and even timeframes. If the police can predict the most likely times and locations for these crimes to happen, they can send officers to the designated locations during the most likely time period. This may preemptively stop crimes, and will lead to greater safety for everyone.

Enabling Rapid Police Response to Terrorist Attacks

In situations such as the Boston Marathon bombing, where there is a massive amount of data, the police can use the data to help them quickly find the perpetrators. The Boston Marathon bombing was a very public attack, and there were huge amounts of video taken of the occurrence, both from personal phone videos and security footage. The police reached out through social media in order to retrieve all this video footage. They thoroughly examined the videos, and thanks to the images found on the data sources, they were able to identify two suspects. These images led to the successful capture of the two suspects, allowing for a quick and successful response to the attack.

Chapter 25: The Use of Big Data in the Public Sector

Analysis of big data benefits almost every sphere of life. We have already tackled numerous industries' methods of using big data, such as in the travel sector, finance sector, and even law enforcement. We can now expand that discussion to include how other public sector agencies may be able to benefit from big data.

One public sector that we could consider is the health sector. Hospitals and Emergency Rooms all over tend to have unpredictability as one of their major concerns. In order to deal with unpredictability, they have large amounts of staff and resources on hand to be able to cope with any situation that arises. Certain things may be predictable; given enough time and data, a pattern may arise. One pattern that has been seen is the patient inflow into an emergency room. Hospitals in the United States have been able to apply modern methods of data analysis to the patient intake data that they have, and they have been able to predict, with a 93% accuracy rate, the amount of daily emergency admissions. Everyone is able to benefit from this type of data analysis, as bed allocation is improved, with hospitals better being able to predict the amount of beds needed for emergency cases, making it less likely for non-emergency cases to be deprived of beds. In addition to this, the prediction allows hospitals to save money, lessens the need for overtime hours, and shifts the environment from a reactive to a proactive one. This improves staff morale, as people can have reasonable expectations and ready themselves rather than being on edge all day, not knowing what to expect.

The health sector is not the only public sector that has used big data to improve the efficiency of their work and improve their results. One example of this would be in Australia, with the Australian Taxation Office using modern methods of data analysis to sift through large amounts of tax data and find probable tax fraud cases, such as certain online retailers who do not meet their obligations. There are numerous methods of analyzing a country's tax data in order to discover various varieties of fraud, and to even discover certain trends among consumers.

Education is another public sector that greatly benefits from the proper utilization of big data. In our modern society, many schools have begun to collect big data, and have been starting to analyze it in order to make predictions for the future. This data includes various aspects on student life, such as scores, anxiety levels, or even a student's interest or lack thereof for certain subjects. This can be used in improving the school systems all over. When this data gets collected at a national level, and if subjected to a rigorous analysis, this may allow the Department of Education to find solutions to many problems in our schools, allowing for the improvement of our education system as well as reducing costs. There have been concerns, however, from both parents and children regarding possible violations of privacy if this data is used in an extensive analysis.

United States Government Applications of Big Data

Big data is not entirely new, especially when it comes to its role in the United States of America's government. The U.S government utilizes six of the ten most powerful supercomputers in existence for data analysis. Even President Barack Obama used big data when running his re-election campaign in 2012, and he announced the Big Data Research and Development Initiative soon after his re-election. This initiative's goal is to be able to discover various applications of big data research when it comes to finding solutions to the nation's current problems. Even the National Security Agency is building what may just be the world's largest data center. This data center is being built in Utah, and will be able to store multiple exabytes of data, and one of its major functions is to analyze what may possibly be some of the most delicate information in the world today.

Data Security Issues

The sheer amount of data collected by public agencies, as well as the security and possible risks of gathering such data as previously discussed, has understandably turned this into a major issue, and has instigated many discussions regarding it. Once the public found out that government agencies such as the National Security Agency and the Central Intelligence Agency have been collecting immense amounts of data on almost everyone, there was a general uproar. This furor was mainly due to the general lack of understanding as to why the government was collecting this information, as well as fear that this data would be misused. While it is entirely possible that big data may be misused, the concept of a “Big Brother” scenario tends to rise from lack of knowledge rather than from reality. However, such data being misused tends to be illegal in origin, such as through surveillance and other similar methods. Edward Snowden exposed certain methods that the National Security Agency used in order to gather massive amounts of personal data. Representatives of these public agencies as well as noted experts have in fact mentioned that far from the government having the most personal data, it is the private sector that tracks most people’s every move, as private companies have more data analysis capabilities and tend to collect more information than the public sector. On the flip side, however, much of the private sector’s data collection comes legally, from people voluntarily entering it through agreements, terms of use, and various online sign up methods.

The Data Problems of the Public Sector

There are many issues with the government's capability in carrying out big data analysis, however, as there are substantial amounts of data gathered, with the government tending to have less resources and capabilities devoted to big data analysis. As many public sector employees are less technologically savvy than their private sector counterparts, governments have to outsource their work. However, governments can rarely match the competitive rates that private companies offer, so they are less likely to be able to get industry leaders to work for them. This means that governments lack the human capital to properly utilize the data available to them. Even when it comes to sharing their data and the analysis, governments tend to lag behind the private sector, which has found numerous ways of sharing big data and analysis.

Chapter 26: Big Data and Gaming

When one speaks of the gaming industry, they may be referring to a plethora of activities. Gaming includes everything from mobile games built on social media such as Candy Crush, to casino-style games such as blackjack and poker that allows a player to gamble with real money. No matter which games are being referred to, however, from a children's game to a high-stakes casino game, the gaming industry employs the use of big data to improve their profit margin. Casino gaming companies have been using the numbers and data they gather from the very beginning, as casino games are number games that ensure that the house will always win in the end, as the statistics prove. Other gaming companies, however, such as Electronic Arts, have begun to adopt this data based approach in the modern era. These numbers are used to improve the gaming experience, improve platform personalization, and to find ways to keep players hooked, all through extensive data analysis.

Big Data and Improving Gaming Experience

Once upon a time, game manufacturers and publishers could only find out what players thought about a game by developing one, marketing it, selling it, then finding out what ideas players had to improve their game. This lack of information no longer holds true today. Much of the gaming currently happening today happens online. From social network games such as Mafia Wars and Farmville, to massive multiplayer roleplaying games such as World of Warcraft, to First Person Shooters such as CounterStrike, and so many other game types and genres, players log countless hours online and generate so much information that game developers can use. Such data includes what peak hours are, how long people tend to play, and other similar metrics. Game manufacturers can also find out what type of device was used to play, who was played with, and even what goal a player was out to complete in the game. The list is almost endless, and there is no cap to the information a developer can gather. The developers then eventually convert this data into conclusions and theories. Based on these metrics, game developers may change the experience by adding more relevant content, or even creating in-game rewards. They can even change it to suit the device it is most played in. Game developers can gather almost any sort of information about their players in this day and age, and have thus used this to improve their games. For example, in theory, a game may have an overly steep learning curve, leading many players to give up early. Game developers could find this out, and adjust their game to suit beginners more, allowing it to be friendlier to newbies. Even microtransactions are driven by big data, as many social games sell in-game items for real cash. Game developers can find out the how, when, and why people make purchases, and they can use this to tweak their game to improve in-game item sales.

Big Data in the Gambling Industry

The gambling industry has been one of the longest users of big data, especially when it comes to manipulating people to buy their product. This has led to many people viewing gambling companies as evil, when they are just like any other company, using tactics to get people to buy the products they sell. In fact, one way of looking at it is that there is barely any difference between a person losing money gambling and a person being manipulated to buy a useless product. Among the first adopters of big data were the bookmakers. These people, commonly referred to as bookies, develop odds for sporting events designed to be appealing to users, yet low enough that they would consistently turn a profit. Eventually, software was developed to do this, and online and live betting boomed, but bookies could still stay ahead thanks to their greater access to data. The use of big data has allowed bookies to be far more accurate in predicting the outcomes of sports events when compared to someone doing it without data. The data from hundreds or even thousands of matches played by the subject enables bookies to come to more accurate predictions, allowing them to stay one step ahead of your average customer. They can even process live matches in order to automatically adjust the odds of events, allowing them to ensure that nothing would be missed. However, other entities thought that if bookies and gambling companies could do this, why couldn't they? Large internet companies such as the search giants Google and Yahoo used big data to predict match outcomes, and were able to do so with a viable amount of accuracy. Famously, Google was able to predict 14 of 16, and Yahoo 15 of 16 match outcomes correctly at the 2014 Football World Cup. While admittedly, these were usually just the favorites winning, their accuracy still remains impressive. The use of big data is not only limited to predicting the odds, but has many other uses to improve a company's bottom line. Many casinos and gambling houses, such as Harrah's and Caesar's Palace, have been using big data for quite some time in order to gather important information regarding their clients, and they use this data to nudge clients to keep gambling. An example would be slot machines: casinos know how much time the average player will spend at a slot

machine, how long it takes for a player to get frustrated and walk away, and what type of slot machines they prefer. This data is used to improve the casino's marketing strategies, make the casino floor's setup more appealing to clients, and even find ways to minimize negative outcomes such as clients not returning after losses. Caesar's Palace famously increased their rate of return clients by giving losing players free dinner coupons before they exited the building. This method was backed up by big data, and was able to generate many returns. Every casino uses big data analysis to improve the placement of their products on the casino floor, and every casino floor has been carefully planned to optimize revenue.

Gaming the System

Big data is not restricted solely for company use, but the players may use them as well. The rise of the use of big data on the demand side of the transaction has created methods for individuals to make use of big data. There are numerous websites that use data analysis to improve the odds for gamblers. Poker is one example of a game ruled by numbers, and any player with knowledge in math and statistics has an edge over a player without. Poker also happens to be a type of game that is played exclusively between players, with the house taking no direct profit, allowing a significant amount of players to consistently win. There are software on the market such as Sharkscope and Hold ‘em Manager that assist poker players with their games. Sharkscope allows poker players access to the results of thousands of tournaments, allowing them to research data points such as buy-in, average return of investment, and similar data. This tool can allow savvy poker players to spot the best target at the table, and has proved to be very popular among poker players. Hold ‘em manager and Poker Tracker tracks the statistics of a player’s opponents, examining how they played certain hands. These apps store the data from hands played by opponents against the user, and allow the user to make accurate conclusions about the opponent’s playstyle based on the data, letting users of these apps dominate many of the games they participate in.

Another popular form of gambling is betting on sports. Bookies have always used a form of big data analysis to get their information and form their odds. Nowadays, there are many websites, for example, Betegy, where they claim to be more accurate than the bookies, and they promise to give users a way to beat the odds. Betegy even claims that 90% of English Premier League games can be predicted by their algorithm, and if this is true, then it will change the world of sports betting. Whether or not these websites can back up the truth of their claims is still unknown, but

regardless, big data analysis will change the gambling world, as more and more software is developed to beat the odds.

The Expansion of Gaming

The gambling industry has long been a profitable one, but the video game industry is a rising star in the world of gaming. Various online games such as Defense of the Ancients and League of Legends have drawn crowds, with thousands paying for the privilege of attending gaming events and tournaments. These games have developed a large following, with tournaments improving their production values and having prize pools reaching millions of dollars. The video game industry has been valued at over one hundred billion dollars, and is still rapidly expanding. This is the reason why major game developers such as EA have spent large amounts of cash on collecting and analyzing player data. The competition is fierce, and as players have only limited time and resources to spend on games, gamers look for the best in game design, personalization, and services. This demand makes big data analysis invaluable to keeping and expanding market share. Game developers have realized that the best way to create a game that sells well is to listen to the gamers. When done right, big data analysis helps create a major blockbuster game, but when done poorly, it may lead to a flop. There is a huge amount of pressure on the data scientists behind each videogame due to this. Videogame companies even use the internet connectivity of gaming consoles to collect data. Data about a player's gaming habits is collected whether or not the device is connected to the internet, and is sent once the device goes online. All of this is recorded and used to develop games. Social media games are an even tougher challenge, as they tend to rely on microtransactions rather than up-front payment to turn a profit. These gaming companies use complex processes to monetize these games, incorporating numerous factors. At the end of the day, the gaming industry is one of the industries that has seen an exponential growth rate due to the rise of big data, and we can see that as more data is analyzed, games will continue to move towards what the users really want.

Chapter 27: Prescriptive Analytics

Prescriptive Analytics- What is It?

Prescriptive analytics is the newest and most advanced analytics system since descriptive and predictive analytics. It's becoming more and more common in the business world. The three different existing analytic models may be defined as follows:

Descriptive analytics is the analysis of data to help show trends and patterns that emerged in the past. It tells us *what happened*.

Predictive analytics uses data, statistical algorithms, and machine learning techniques to make predictions about future events. It tells us *what will happen*.

Prescriptive analytics uses optimization and simulation algorithms to show companies the best actions to take in order to maximize profit and growth and determine options for the future. It tells us *what we should do*.

Prescriptive analytics reaches past what the future holds and tells companies what they should be doing to prepare for it. It provides different options for action, and suggests which ones would be best. There is a certain measure of artificial intelligence built into its processes, in the sense that it can analyze optimization and simulation algorithms and then determine what the options are, moving forward. It can even recommend which would be the best option to take.

Compared to the other two models, prescriptive analytics is more difficult to administer than the other two analytical models. With this model, machine thinking and computational modeling are applied to several different data sets, such as historical data, transactional data, and real-time data feeds.

Prescriptive analytics uses algorithms that work on data with very few parameters. The algorithms are specially designed to conform to the changes in established parameters and are not subject to external human controls. The algorithms are free to be optimized automatically. As time progresses, their "learning" ability helps them to better predict future events.

What Are its Benefits?

Prescriptive analytics is still fairly new, so many companies aren't yet using it in their daily processes, although some of the larger corporations are already using it on a daily basis. The benefits are already beginning to manifest in a number of industries, especially supply chains, insurance, credit risk management, and healthcare.

Because it learns from transactions the customer made in the past, predictive analytics can, for example, prescribe the best time of day to contact the client, which marketing channel is likely to be more successful, and what product will be most appropriate. Much of this can be automated, such as by using email services and text messages.

Thus, a benefit is that the knowledge gained from the analysis of big data can be applied to a vast number of decisions that would usually take up too much time for people to manually decide upon.

What is its Future?

The future of prescriptive analytics looks bright. Figures show that in 2014, about 3% of companies were using prescriptive analytics. It was predicted that, by 2016, so called “streaming analytics” would become mainstream. Streaming analytics is a form of analytics that gets applied in real-time as transactions occur.

Prescriptive analytics will become more and more important for cyber security, analyzing suspicious events as they happen, having great application in preventing, for example, terrorism events.

Prescriptive analytics is also set to become very big in lifestyle activities in 2016 onwards. This includes activities such as online shopping and home security.

The predictions for 2016 haven’t been entirely accurate, as it’s not yet used much in the home environment, but its use in the corporate world is taking off. However, even in businesses that use it, it seems that it’s used in some departments but not in others.

Google’s “Self-Driving Car”

Google made extensive use of prescriptive analytics when designing its self-driving car. Self-driving cars utilize machine learning to develop smarter ways of driving on the roads. In the vehicle, the machine, as opposed to a human driver, analyzes the real-time incoming and stored data to make decisions.

The vehicles house sensors and software that detects surrounding vehicles, other road users such as cyclists, and obstacles such as roadworks. It detects small movements such as hand signals given by other drivers, and uses these to predict what the other driver is probably going to do. Based on this, the software takes action. It can even adjust to unexpected events, such as a child running across the road.

It started in 2009, when the challenge for Google was to drive over ten uninterrupted 100 mile routes completely without human driver intervention. By 2012, they had branched out onto the highways and onto busy city streets. By 2014, new prototype vehicles were being designed with no steering wheels or foot pedals. These prototypes hit the roads (safely) in 2015, with Google employees as testers. In 2016, the Google self-driving car project became an independent company dedicated to making self-driving technology a safe and affordable option for all drivers. (The company’s name is Waymo.)

The cars are a perfect metaphor for what prescriptive analytics can do for a business: the computer tells management what route to navigate based on its analysis of all the data.

Prescriptive Analytics in the Oil and Gas Industry

The oil and gas industry uses predictive analytics in many different ways to ensure efficient, safe, and clean extraction, processing, and delivery of their product. While shale oil and gas are abundant in the US, they are difficult to find and extract safely. Horizontal drilling and fracking are expensive and possibly cause environmental damage. They are also relatively inefficient. As a result, some of the biggest oil and gas corporations are using prescriptive analytics to help deal with and minimize these problems.

The processes surrounding oil and gas exploration and extraction generate huge amounts of data, which are set to double in the next couple of years. It's easy to see how this will be possible when one considers that there are about 1 million oil wells currently in production in the US alone. The focus needs to be on ways of using all this data to automate small decisions and guide big ones, thus reducing risk, improving productivity, and lessening the environmental impact.

Companies can now look at a combination of structured and unstructured data, giving a better picture of problems and opportunities that may arise, and providing ideas for the best actions to take to solve these. This combination will mix machine learning, pattern recognition, computer vision, and image processing. The blend of all of these results in the ability to produce better recommendations of where and how to drill, and of how to solve problems that may crop up.

Types of data that are looked at include graphics from well logs and seismic reports, videos from cameras in the actual wells, fiber optic sensor sound recordings of fracking, and production figures. Geologists take data from existing wells to give information about the rocks forming the area and the nature of the ground below the surface. Prescriptive analytics is then used to interpret this information and predict what the ground may be like between wells, enabling the rest of the ground to be visualized with some accuracy. The best course of action is then suggested, using these observations.

Prescriptive analytics should enable oil and gas companies to predict the future of the wells in a given oil field and know where to drill and where

not to. Data is not only collected regarding the actual oil field and wells, but also about drilling equipment and other machinery. This is useful as it can be predicted when maintenance will be necessary, and the correct intervention can be prescribed. It can also be suggested when the old machinery will be likely to need replacing. Predictive analysis may predict corrosion in pipelines, using data collected by robotic devices in the pipelines, and then suggest preventative measures.

In this way, the technology helps the oil companies to extract the oil and gas safely and efficiently, as well as deliver it to market in the most environmentally friendly manner. The US is quickly becoming an energy superpower, and was set to overtake Saudi Arabia in 2016 as the world’s biggest oil producer. Prescriptive analytics can only help in this endeavor.

Prescriptive Analytics and the Travel Industry

Predicting the future in any industry is mainly to do with finding patterns in the large quantities of available data, so that we can gain insights from it. By looking at what customers have done in the past, industries can make predictions about what they're likely to do in the future and prescribe what to do about it. They can suggest the perfect product, specifically tailored to the needs of the customer, such as holiday destinations, hotel recommendations, or the best flight routes, all within a fraction of a second.

As a traveler, this would likely work for you as follows: say you're travelling to Denver for a four-day work conference, plus you want to spend a few days in the mountains straight after. You'd begin with an online search for flights into Denver using an online travel agency. Because of predictive analysis, you'd straight away receive a special offer from the airline you fly with most often, for the type of route you'd prefer (perhaps early morning with no stopovers.) You'd receive some information for some good restaurants in the city near to where you've booked your hotel, as well as some offers for mountain cabins and guided hikes. All this would be possible because, using big data specialists to help them, travel companies are getting really good at working out their customers' needs based on previous travel patterns.

The travel industry has been gathering data much faster than it could use in recent decades from airlines, hotels, and car rental companies. Now that analytics tools and computer storage capacity are bigger, more powerful, and more affordable than ever before, this data can now be made sense of and utilized.

In the travel industry, predictive analytics promises to bring greater profits for suppliers and a better travel experience for customers.

Prescriptive Analytics in the Healthcare Industry

The healthcare industry is also becoming aware of how the cutting-edge technology that prescriptive analytics offers can improve their complex operational activities. It’s not good enough to just have piles and piles of information. This on its own will not give the industry insight. It’s only when data is put into context that it provides usable knowledge and can be translated into clinical action. The goal must always be to use historical patient data to improve current patient outcomes.

An example is in the context of patient readmissions. Predictive analytics can accurately forecast which patients are likely to return in the next month, and can provide suggestions regarding associated costs, what medications are likely to be needed, and what patient education will probably be needed at the time of discharge.

It’s important to remember that even when clinical event prediction is specific and accurate, this information will only be useful if the proper infrastructure, staff, and other resources are available when the predicted events occur. If clinical intervention doesn’t happen, no matter how good the predictors were, they will not be utilized to the full. Decision makers must therefore not be too far removed from the point of decision. However, when it is used correctly, predictive analytics can help control costs and improve patient care, which are probably the main goals of any healthcare organization.

Prescriptive analytics can suggest which treatment would be the best for a specific patient’s needs. This helps streamline the diagnostic process for medical practitioners. Hours can be saved because the many options for any medical condition can be narrowed down by the software for the doctor to then make a final decision.

Prescriptive analytics can help the pharmaceutical industry as well by streamlining new drug development and minimizing the time it takes to get the medicines to the market. This would reduce expenditure on drug research. Drug simulations could possibly shorten the time it takes to perfect new drugs.

As we can see from looking at the use of prescriptive analytics by the industries above, among the many benefits provided by prescriptive analysis, the most important seem to be reduced risk as decisions are data-driven and therefore more accurate, increased revenue because processes are optimized, thereby minimizing cost and maximizing profit; increased efficiency as processes are streamlined and improved.

Data Analysis and Big Data Glossary

The following list are terms and words that you would find when reading about topics like data analysis, big data, and data science. Not all of the terms have been used in this book, but are extremely common in these fields. While a more in depth list would take up too much space, this list is of the most common words that are used.

A

ACID Test – this test checks for consistency, durability, isolation, and atomicity. It used specifically with data to check that everything is working the way it needs to.

Aggregation – collection of data from various places. Data collected in an aggregation is often used for processing or analysis.

Ad hoc Reporting – these reports are created for one specific use only.

Ad Targeting – trying to get a specific message to a specific audience. Businesses use this to try and get people interested. It is most often in the form of a relevant ad on the internet or by direct contact.

Algorithm – a formula that is put into software to analyze specific sets of data.

Analytics – determining the meaning of data through algorithms and statistics.

Analytics Platform – a combination of software and hardware or just software that works for you to use computer power as well as tools that you need to perform queries.

Anomaly Detection – a process which reveals unexpected or rare events in a dataset. The anomalies do not conform with other information in the same set of data.

Anonymization – the process by which links are severed between people and their records in a given database. This process is done to prevent people from discovering the person behind the records.

Application Program Interface (API) – a standard and set of instructions that allow people to be able to build web-based software applications.

Application – software designed to perform certain jobs or just a singular job.

Artificial Intelligence – the ability of a computer to use gained experience to situations in the same manner that human beings can.

Automatic Identification and Capture (AIDC) – in a computer system, this will refer to any method through which data is identified, collected, and stored. For example, a scanner that collects data through RFID chips about items that are being shipped.

B

Behavioral Analytics – the use of collected data about individual's or group's behavior to help understand and predict future actions done by a user or group.

Big Data – although big data has been defined over and over, every definition is roughly the same. The first definition originated from Doug Laney, who worked for META Group as an analyst. He used the term and defined it in a report titled, "3D Data Management: Controlling Data Volume, Velocity, and Variety." Volume is the size of a dataset. A McKinsey report, "Big Data: The Next Frontier for Innovation Competition and Productivity," went further on this subject. The report stated, "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." Velocity is the speed at which data is acquired and used. Companies are gathering data more and more quickly; companies are also working on ways to derive meaning from that data in much more quicker ways, aiming for real time. Variety is in reference to how many different data types are able to be used for collection and analysis. This is in addition to the data that would be found in a normal database.

There are four major categories of information that are considered in big data:

Machine generated – this includes RFID data, location data gathered from mobile devices, and the data that comes from different monitoring sensors.

Computer log – clickstreams from some websites as well as a couple more sources.

Textual social media – information from websites like Twitter, Facebook, LinkedIn, etc.

Multimedia social media – other information from websites like Flickr, YouTube, or Instagram.

Biometrics – the use of technology to identify people through fingerprints, iris scans, etc.

Brand Monitoring – the way that a brand's reputation could be monitored online. This usually requires software to analyze available data.

Brontobye – a vast amount of bytes. This term is not officially on the scale; it was proposed for a unit measure that goes beyond the typically used yottabyte scale.

Business Intelligence (BI) – this term is used to describe the identification of data. The term also covers extraction and analysis of the data as well.

C

Call Detail Record Analysis (CDR) – data gathered by telecommunications companies. This data can contain the time and length of phone calls. The data is used in various analytical operations.

Cassandra – a popular columnar database that is normally used for big data applications. The database is open-source and managed by The Apache Software Foundation.

Cell Phone Data – mobile devices are capable of generating tons of data at this point. Most said data can be used in analytical applications.

Classification Analysis – this specific type of data analysis allows data to be assigned to specific groups or classes.

Clickstream Analysis – this analysis is done by looking at web activity by gathering information on what users are clicking on the page.

Clojure – this programming language is based on LISP and uses JVM (Java Virtual Machine; the dynamic language is great for parallel data processing.

Cloud – this term encompasses any and all services and web-based applications that are hosted somewhere besides the machine they are being accessed from. A common use may be cloud file storage in something like Dropbox or Google Drive.

Clustering Analysis – using data analysis to locate differences and similarities in datasets. This allows for datasets with similarities to be clustered together.

Columnar Database/Column-Oriented Database – this database houses data in columns instead of the typical rows of row databases. Row databases usually contain information like names, addresses, phone numbers, etc. One row is one person in row databases. Instead, a column database houses all names in one column, addresses in the next, and so on and so forth. This type of database has the advantage of faster disk acess.

Comparative Analysis – this kind of analysis compares at least two sets of data or even processes to look for patterns. This is commonly used in larger sets of data.

Competitive Monitoring – companies may use software that will allow them to track the web activity of any competitors that they have.

Complex Event Monitoring (CEP) – this process watches the system's events, analyzing them as they happen. It will act if necessary.

Complex Structured Data – this set of data consists of at least two inter-related parts. This kind of data is hard for structured tools and query languages to process.

Comprehensive Large Array-Data Stewardship System (CLASS) – a digital library that houses data gained from NOAA (US National Oceanic and Atmospheric Association). This includes historical data.

Computer-Generated Data – data created by a computer and not a person. This could be something like a log file that the computer creates.

Concurrency – the ability of a system or program to be able to execute several processes at once.

Confabulation – making an intuitive decision as if were a data-based decision.

Content Management System (CMS) – software that is used for the publication and management of web-based content.

Correlation Analysis – the way that statistical relationships between two or more variables are determined. This process is sometimes used to look for predictive factors in the data.

Correlation – dependent statistical relationships such as the correlation between parents and their children, or the demand and cost of a specific product.

Cross-Channel Analytics – a specific kind of analysis that shows lifetime value, attribute sales, or average orders.

Crowdsourcing – a creator (for video games, board games, specific products, almost anything) asks the public to help them finish a project. Most people think of crowdsourcing in terms to Kickstarter, but it can also refer to forums where a programmer may post problematic code that they are dealing with.

Customer Relationship Management (CRM) – software that allows a company to more easily manage customer service and sales.

D

Dashboard – a report, usually graphical, showing real-time data or static data. This can be on a mobile device or a desktop. The data is collected and analyzed to allow managers access to quick reports about performance.

Data – a qualitative or quantitative value. Data can be simple like results from market research, sales figures, readings taken from monitoring equipment, projections for market growth user actions on websites, and demographic information.

Data Access – the ability and process to retrieve and view stored data.

Digital Accountability and Transparency Act 2014 (DATA Act) – a US law that intended to make it easier for people to get into the federal government expenditure information. This act required the White House of Management and Budget as well as the Treasury to standardize the data on federal spending as well publish it.

Data Aggregation – a collection of data that comes from a number of sources. Aggregated data is often used for analysis and reporting.

Data Analytics – the use of software to derive meaningful information from data. Results may be status indications, reports, or automatic actions that are based on the information found in the data.

Data Analyst – a person who prepares, models, and cleans data so that information can be gained from it.

Data Architecture and Design – the structure of enterprise data. The real design or structure could vary because the design is dependent on the result. There are three stages to data architecture:

A conceptual representation of the entities in the business.

The representations of the relationships between each entity.

A constructed system that supports functionality.

Data Center – a physical space that contains data storage devices as well as servers. The data center may house devices and servers for multiple organizations.

Data Cleansing – revising and reviewing data to get rid of repeated information, to correct any spelling errors, to fill in missing data, and to provide consistency across the board.

Data Collection – a process that captures data of any type.

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

Data Democratization – the idea to allow all workers in an organization to have direct access to data. This would be instead of waiting for data to make its way through other departments within the business to get delivered to them.

Data-Directed Decision Making – using data to support crucial decisions made by individuals.

Data Exhaust – this data is created as a byproduct of other activities being performed. Call logs and web search histories are some of the simplest examples.

Data Feed – how data streams are received. Twitter's dashboard or an RSS feed are great eamples of this.

Data Governance – rules or process that ensure data integrity. They guarantee that best practices (set by management) are followed and met.

Data Integration – a process where data from two or more sources is combined and presented in one view.

Data Integrity – a measure for how much an organization trusts the completeness, accuracy, and validity of data.

Data Management – a set of practices that were set by the Data Management Association. These practices would ensure that data is managed from when it is created to when it is deleted:

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

Data Management Association (DAMA) – an international non-profit organization for business and technical professionals. They are “dedicated to advancing the concepts and practices of information and data management.”

Data Marketplace – a location online where individuals and businesses can purchase and sell data.

Data Mart – this is the access layer of a data warehouse. It provides users with data.

Data Migration – a process where data is moved from one system to another, moved to a new format, or physically moved to a new location.

Data Mining – a process through which knowledge or patterns can be derived from large sets of data.

Data Model/Modeling – these models define the data structure that is necessary for communication between technical and functional people. The models show what data is needed for business, as well as communicating development plans for data storage and access for specific team members.

Data Point – an individual item on a chart or graph.

Data Profiling – a collection of information and statistics about data.

Data Quality – a measurement of data used to determine if it can be used for planning, decision making, and operations.

Data Replication – a process that ensures that redundant sources are actually consistent.

Data Repository – a location where persistently stored data is kept.

Data Science – a newer term with several definitions. It is commonly thought of as the discipline that uses computer programming, data visualization statistics, data mining database engineering, and machine learning to solve complex problems.

Data Scientist – someone qualified to practice data science.

Data Security – ensuring that data isn't accessed by unauthorized users or accounts. Also ensuring that data isn't destroyed.

Data Set – a collection of data stored in a tabular form.

Data Source – the source of data, like a data stream or database.

Data Steward – someone in charge of the data that is stored in a data field.

Data Structure – a method for storing and organizing data.

Data Visualization – the visual abstraction of data that can help in finding the meaning of data or communicating that information in a more efficient manner.

Data Virtualization – the way in which data is abstracted through one access layer.

Data Warehouse – a location where data is stored for analysis and reporting purposes.

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

Database Administrator (DBA) – someone, usually a certified individual, in charge of supporting and maintaining the integrity of content and the structure of a database.

Database as a Service (DaaS) – a cloud-hosted database that is sold on a metered basis. Amazon's Relational Database Service and Heroku Postgres are examples of this system.

Database Management System (DBMA) – a software that is often utilized for collecting and providing structured access to data.

De-Identification – removing data linking information to specific individuals.

Demographic Data – data that shows characteristics about the human population. This can be data such as geographical areas, age, sex, etc.

Deep Thunder – IBM's weather prediction service. It provides organizations like utility companies with weather data. This information will allow companies to optimize their distribution of energy.

Distributed Cache – a data cache that is spread over a number of systems. However, the data cache acts as one system and allows performance to be improved.

Distributed File System – a file system that spans several serves at the same time. This allows for data and file sharing across the servers.

Distributed Object – this software module was designed to work with other distributed objects that are house on different computers.

Distributed Processing – this is the use of several computers connected to the same network to perform a specific process. Using more than one computer may speed up the efficiency of the process.

Document Management – tracking electronic documents and scanned paper images, as well as storing them.

Drill – an open-source system distributed for carrying out analysis on extremely large datasets.

E

Elastic Search – an open-source search engine built on Apache Lucene.

Electronic Health Records (HER) – digital health record that is accessible and usable across different healthcare settings.

Enterprise Resource Planning (ERP) – a software system that allows a business to manage resources, business functions, as well as information.

Event Analytics – an analysis method that shows the specific steps that were taken to get to a specific action.

Exabyte – 1 billion gigabytes or one million terabytes.

Exploratory Data Analysis – data analysis with focus on finding general data patterns.

External Data – data outside of a system.

Extract, Transform, and Load (ETL) – a process that is used specifically in data warehousing to prepare data for analysis and reporting.

F

Failover – a process that automatically switches to another node or computer in the case of a failure.

Federal Information Security Management Act (FISMA) – a US federal law stating that all federal agencies have to meet specific security standards across all their systems.

Fog Computing – a computing architecture that enables users to better access data and data services by putting cloud services (analytics, storage, communication, etc.) closer to them through geographically distributed device networks.

G

Gamification – utilizing gaming techniques for applications that are not games. This can motivate employees and encourage specific behaviors from customers. Data analytics is often required for this in order to personalize the rewards and really get the best result.

Graph Database – NoSWL database that uses graph structures for semantic queries that have edges, nodes, and properties. These semantic queries can store, query, and map data relationships.

Grid Computing – improving the performance of computing functions by making use of resources within multi-distributed systems. These systems within the grid network don't have to be similarly designed, and they don't have to be in the same physical geographic location.

H

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

Hadoop Distributed File System (HDF) – a file system that is created to be fault-tolerant as well as work on low-cost commodity hardware. This system is written for the Hadoop framework and is written in the Java language.

HANA – this hardware and software in-memory platform comes from SAP. The design is meant to be used for real-time analytics and high volume transactions.

HBase – a distributed NoSQL database in columnar format.

High Performance Computing (HPC) – also known as super computers. These are usually created from state of the art technology. These custom computers maximize computing performance, throughput, storage capacity, and data transfer speeds.

Hive – a data and query warehouse engine similar to SQL.

I

Impala – an open-source SQL query engine distributed specifically for Hadoop.

In-Database Analytics – this process integrates data analytics into a data warehouse.

Information Management – this is the collection, management, and distribution of all kinds of information. This can include paper, digital, structured, and unstructured data.

In-Memory Database – a database system that uses only memory for storing data.

In-Memory Data Grid (IMDG) – a data storage that is within the memory and across a number of servers. The spread allows for faster access, analytics, and bigger scalability.

Internet of Things (IoT) – the network of physical objects full of software, electronics, connectivity, and sensors that enable better value and service through exchanging information with the operator, manufacturer, or another connected device. Each of these things, or objects, is identified through its unique system for computing; however, each object can interoperate within the internet infrastructure that already exists.

K

Kafka – this open-source messaging system is used by LinkedIn. It monitors events on the web.

L

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

Legacy System – an application, computer system, or a technology that, while obsolete, is still used because it adequately serves a purpose.

Linked Data – as described by Tim Berners Less, 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 – distributing a workload across a network or even a cluster in order to improve performance.

Location Analytics – using mapping and analytics that are map-driven. Enterprise business systems as well as data warehouses will use geospatial information as a way to associate location information with datasets it

Location Data – this data describes a specific geographic location.

Log File – these files are created automatically by a number of different objects (applications, networks, computers) to record what happens during specific operations. An example of this might the log that is created when you connect to the internet.

Long Data – this term was coined by Samuel Arbesman, a mathematician and network scientist. It refers to “datasets that have a massive historical sweep.”

M

Machine-Generated Data – data created from a process, application or other source that is not human. This data is usually generated automatically.

Machine Learning – using algorithms to allow a computer to do data analysis. The purpose of this is to allow the computer to learn what needs to be done when specific events or patterns occur.

Map Reduce – this general term refers the process of splitting apart problem into small bits. Each bit is distributed among several computers on the same network, cluster, or map (grid of geographically separated or disparate systems). From this, the results are gathered from the different computers to bring together into a cohesive report.

Mashup – a process by which different datasets are combined to enhance an output. Combining demographic data with real estate listings is an example of this, but any data can be mashed together.

Massively Parallel Processing (MPP) – this processing will break a single program up into bits and execute each part separately on its own memory, operating system, and processor.

Master Data Management (MDM) – any non-transactional data that is critical to business operations (supplier data, customer data, employee data, and product information). MDM ensures availability, quality, and consistency of this data.

Metadata – data that describes other data. The listed date of creation and the size of data files are metadata.

MongoDB – open-source NoSQL database that has login to keep the management under control.

MPP Database – a database optimized to work in an MPP processing environment.

Multidimensional Database – this database is used to stare data in cubes or multidimensional arrays instead of the typical columns and rows used in relational databases. Storing data like this allows for the data to be

analyzed from various angles for analytical processing. This allows for the complex quries on OLAP applications.

Multi-Threading – this process breaks up an operation in a single computer system into multiple threads so that it can be executed faster.

N

Natural Language Processing – the ability of a computer system or program to understand the human language. This allows for automated translation, as well as interacting with the computer through speech. This processing also makes it easy for computers and programs to determine the meaning of speech or text data.

NoSQL – a database management system that avoids the relational model. NoSQL handles large volumes of data that do not require the relational model.

O

Online Analytical Processing (OLAP) – A process of using three operations to analyze multidimensional data:

-Consolidation – aggregating available factors

-Drill-down – allowing users to see underlying details to the main data

-Slice and Dice – allowing users to pick specific subsets and view them from different perspectives

Online Transactional Processing (OLTP) – this process gives users to large amounts of transactional data so that they can derive meaning from the data.

Open Data Center Alliance (ODCA) – an international group of IT organizations that have the single goal. They wish to hasten the speed at which cloud computing is migrated.

Operational Data Store (ODS) – this location is used to store data from various sources so that more operations can be performed on the data before it is sent for reporting in the data warehouse.

P

Parallel Data Analysis – this process breaks up analytical problems into smaller parts. Algorithms are run on each individual part at the same time. Parallel data analysis happens in both single systems and multiple systems.

Parallel Method Invocation (PMI) – this process allows programmed code to call multiple functions in parallel.

Parallel Processing – executing several tasks at one time.

Parallel Query – executing a query over several system threads in order to improve and speed up performance.

Pattern Recognition – labeling or clarifying a pattern identified in a machine learning process.

Performance Management – the process of monitoring the performance of a business of a system. It will use goals that are predefined to better locate areas that need to be monitored and improved.

Petabyte – 1024 terabytes or one million gigabytes.

Pig – a language framework and data flow execution that are used for parallel computation.

Predictive Analytics – analytics that use statistical functions on at least one dataset to predict future events and trends.

Predictive Modeling – a model developed to better predict an outcome or trend and the process that creates this model.

Prescriptive Analytics – a model is created to “think” of the possible options for the future based on current data. This analytic process will suggest the best option to be taken.

Q

Query Analysis – a search query is analyzed in order to optimize the results that it provides a user.

R

R – this open-source software environment is often used for statistical computing.

Radio Frequency Identification (RFID) – a technology that uses wireless communication to send information about specific objects from one point to another.

Real Time – often used as a descriptor for data streams, events, and processes that are acted upon as soon as they occur.

Recommendation Engine – this algorithm is used to analyze purchases by customers and their actions on specific websites. This data is used to recommend products other than the ones they were looking, this include complementary products.

Records Management – managing a business's records from the date of creation to the date of disposal.

Reference Data – data describes a particular object as well as its properties. This object can be physical or virtual.

Report – this information is gained from querying a dataset. It is presented in a predetermined format.

Risk Analysis – using statistical methods on datasets to determine the risk value of a decision, project, or action.

Root-Cause Analysis – how the main, or root, cause of a problem or even can be found in the data.

S

Scalability – the ability of a process or a system to remain working at an acceptable level of performance even as the workload experienced by the system or process increases.

Schema – the defining structure of data organization in a database system.

Search – a process that uses a search tool to find specific content or data.

Search Data –a process that uses a search tool to find content and data among a file system, website, etc.

Semi Structured Data – data that has not been structured with a formal data model, but has an alternative way of describing hierarchies and data.

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

Solid State Drive (SSD) – also known as a solid state. A device that persistently stores data by using memory ICs.

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

Storage – any way of persistently storing data.

Structured Data – data that is organized according to a predetermined structure.

Structured Query Language (SQL) – a programming language designed to manage data and retrieve it from relational databases.

T

Terabyte – 100 gigabytes.

Text Analytics – combing linguistic, statistical, and machine learning techniques on text-bases sources to discover insight or meaning.

Transactional Data – unpredictable data. Some examples are data that relates to product shipments or accounts payable.

U

Unstructured Data – data that has no identifiable structure. The most common examples are emails and text messages.

V

Variable Pricing – this is used to respond to supply and demand. If consumption and supply are monitored in real time, then the prices can be changed to match the supply and demand of a product or service.

Conclusion

By now, you have realized the importance of a secure system for storing and managing data. In order to manage your data effectively, your organization might need to involve people skilled in analyzing and interpreting the information that you are bringing. However, with more effective data management, it will be easier to analyze the data. As competition increases, predictive analytics will also gain more importance.

I have talked about several case studies with large organizations that are using their data to expand and improve their operations. This book's information will hopefully provide you with some new insight into the field of predictive analytics.

Using big data analysis, which has been covered extensively in multiple chapters of this book, you should be able to see how industries ranging from gaming to agriculture will be able to increase their revenue, improve and maintain customer satisfaction, and increase their final product yield. I also discussed the potential dangers and pitfalls of big data. This included the dangers of privacy intrusion and the possibility of failure in business intelligence projects.

These dangers are only a part of the equation, however, they have a major role to play in the big data game. If you want to get involved, then you'll need to pay close attention to these parts of the equation. While big data is certainly the future of business, if the dangers and pitfalls are not considered now, then it might be too late to include them in later considerations.