Problems from the Past, Problems from the Future, and Data Science Solutions

Mahesh Babu Pasupuleti

Student, Masters of Science in Computer Science, Northwestern Polytechnic University, 47671 Westinghouse Dr, Fremont, CA 94539, **USA**

Corresponding Contact:
Email: maheshbp.gs@gmail.com

ABSTRACT

According to the findings of this study, the usual workday for a Data Scientist varies based on the sort of project on which they are working at the time. In order to extract insights from data, a variety of algorithms are employed. Because Data Scientists can access algorithms, tools, and data over the Cloud, they can keep up to date and collaborate more readily than ever before.

Keywords: Data Science, Data Scientist, Data Science Solution

12/31/2015 **Source of Support:** Nil, **Conflict of Interest:** None Declared

INTRODUCTION

There are an increasing number of ways in which organizations may take use of the virtually limitless quantity of data that is now available to them. However, data science is eventually used by all businesses for the same reason: to uncover the most optimal solutions to existing issues. Examine three cases in which data science has provided novel answers to age-old challenges. In the transportation industry, Uber analyzes real-time customer data to determine how many drivers are available, if more are required, and whether a surge fee should be implemented to entice additional drivers. Uber uses data to locate the appropriate amount of drivers in the appropriate location at the appropriate time for a price that the rider is prepared to pay. As part of a distinct transportation-related data science initiative, the Toronto Transportation Commission has made significant progress in resolving an age-old problem with traffic patterns, reorganizing those movements inside and around the city. We have discovered the following using data science methods and analysis: collected information to better understand streetcar operations and to identify potential intervention areas data on consumer complaints was evaluated, Probing traffic performance on major routes to gain a better understanding of how to improve it, and putting up a team to better capitalize on big data for planning, operations, and assessment In part because of a focus on peak-hour clearances and identification of the most congested roads, the number of monthly hours wasted by commuters owing to traffic congestion has decreased from 4.75 hours in 2010 to 3 hours in the middle of 2014. When it comes to dealing with environmental challenges, data science may be both proactive and reactive. Freshwater lakes provide a range of functions for humans and the environment, including supplying drinking water and generating food, among others. Lakes all over the world, however, are being endangered by an increase in the occurrence of dangerous cyanobacterial blooms. There are several efforts and research underway to address this long-standing problem. In the United States, a team of scientists from research sites across the east coast, from Maine to South Carolina, is developing and deploying high-tech techniques to investigate cyanobacteria in lakes throughout the region. Physical, chemical, and biological data in lakes where cyanobacteria have been found are being collected by the team utilizing robotic boats, buoys, and camera-equipped drones. The team is gathering massive amounts of data on the lakes and the growth of dangerous blooms. In addition, the research involves developing new computational models to evaluate the findings. The information gathered will allow for more accurate forecasting of when and where cyanobacterial blooms will occur, allowing for more proactive ways to safeguard public health in recreational lakes as well as those that supply drinking water to communities. Multidisciplinary training like this allows the future generation of scientists to solve social concerns with the appropriate updated data science methods, which are becoming increasingly common (Pasupuleti, 2015). It requires a great deal of data to be collected, cleaned and prepared, and then analyzed in order to acquire the knowledge necessary to produce better solutions for today's businesses. How can you come up with a more effective solution that is also more cost-effective? You must do the following: Determine the nature of the problem and come to a clear grasp of it. Obtain the information needed for analysis. Identify the appropriate tools for the job and create a data strategy to go along with them. Case studies can also be useful in tailoring a proposed solution to a specific situation. It is possible to create a machine learning model after all of these requirements are met and all of the accessible data has been retrieved. It will take time for a company to establish best practices for data strategy through the use of data science, but the rewards will be well worth it in the long run.

Cloud for Data Science

For data scientists, the cloud is a lifesaver. The ability to take [your] data, take your information, and store it in the Cloud, in a centralized storage system, is the primary reason for using the Cloud. It enables you to go over the physical restrictions of the computers and systems that you are currently using, and it enables you to take use of the analytics and storage capabilities of modern machines that do not necessarily have to be your own or your company's equipment. In addition to storing large amounts of data on servers in California or Nevada, cloud computing enables the deployment of extremely advanced computing algorithms as well as the ability to perform high-performance computing on machines that are not owned by the company using cloud computing services. Consider the following scenario: you have some information that you can't store, so you transmit it to a storage space (let's call it the Cloud), but you don't have the algorithms you need to apply since you don't have them with you. However, in the Cloud, you have access to those algorithms at your disposal. Consequently, what you do is you deploy those techniques on really huge datasets, and you are able to do so even if your own systems, machines, and computing environments were not capable of doing so. Cloud, on the other hand, is stunning. The other advantage of the Cloud is that it allows several entities to interact with the same data at the same time, which is really convenient. We call this a cloud because it allows you to work with the same data that your colleagues in Germany and another team in India and another team in Ghana are working with. They are able to do so because the information, the algorithms, and the tools, and the answers, and the results, and whatever else they require are all available in a centralized location that we call a cloud. The cloud is a sight to behold. With the Cloud, you may receive immediate access to open source technologies such as Apache Spark without having to install and setup them on your own machine. Using the Cloud also allows you to have access to the latest up-to-current tools and libraries without having to worry about maintaining them or ensuring that they are up to date yourself. The Cloud is available from any location and at any time of day or night. Using cloud-based technologies from your laptop, tablet, or even your phone makes collaboration easier than ever before, since it allows you to work from any location at any time. Multiple collaborators or teams can view the data at the same time, collaborating on the development of a solution to a problem. A cloud platform is a pre-built environment that is provided by some large technology businesses that allows you to become comfortable with cloud-based technologies. There are three cloud computing platforms available: IBM Cloud, Amazon Web Services (often known as AWS), and Google Cloud Platform. IBM also offers Skills Network labs, also known as SN labs, to learners who have registered with any of the learning portals on the IBM Developer Skills Network. In these labs, you will have access to tools such as Jupyter Notebooks and Spark clusters, allowing you to create your own data science project and develop solutions. Practicing and being more comfortable with the Cloud will allow you to see how it may significantly increase the productivity of data scientists.

FOUNDATIONS OF BIG DATA

In this digital age, everyone leaves a digital footprint. Increasingly, the expanding number of internet-connected gadgets with which we interact on a daily basis collect large quantities of information about us, including anything from our travel patterns to our exercises and entertainment. Even a term has been coined to describe it: Big Data. According to Ernst and Young, the following is an example of a definition: People, tools, and machines are creating vast and varied amounts of data in a dynamic and unpredictable manner. This is referred to as Big Data. For real-time business insights that relate to consumers, risk management and profit maximization, performance management, productivity management, and enhanced shareholder value, new and innovative technology must be developed and implemented that is scalable and able to handle the massive amounts of data that are being collected. Although there is no one definition of Big Data, there are several characteristics that are shared by all definitions, including as velocity, volume, diversity, truthfulness, and value, that are important to consider. The V's of Big Data are as follows: The rate at which data accumulates is referred to as its velocity. Data is created at breakneck speed, as part of a never-ending process that never ends. Near-real-time streaming, local, and cloud-based systems are all capable of processing large amounts of data extremely fast. The magnitude of the data, or the rise in the amount of data saved, is referred to as volume. The rise in data sources, greater resolution sensors, and scalable infrastructure are all contributing to the increase in volume. The diversity of the data is represented by the term "variety." Data in relational databases is structured and fits neatly into rows and columns; unstructured data, on the other hand, is not organized in a pre-defined fashion and includes things like tweets and blog entries, photos, statistics, and video. Data originates from a wide range of sources, including equipment, people, and processes, both within and outside of companies, which contributes to its diversity. Mobile technologies, social media, wearable technologies, geotechnologies, video, and a slew of other factors are driving the market. The quality and origin of data, as well as its conformance to facts and correctness, are all indicators of veracity. Consistency, completeness, integrity, and ambiguity are some of the characteristics. Cost and the necessity for traceability are two of the primary motivators. The argument about the accuracy of data in the digital age continues to flare as a result of the vast amount of information available. Is the information accurate or is it a fabrication? Worth is defined by our capacity and desire to convert facts into something of value. Profit isn't the only thing that adds value. In addition to consumer, employee, and personal happiness, it may provide health-related or societal advantages. Obtaining value from Big Data is the primary reason for which individuals devote their effort to learning about it. Let's take a look at some real-world instances of the V's at work. The rate of uploading hours of film to YouTube is 60 seconds every 60 seconds, which generates a lot of data. Consider how much data may gather in a matter of hours, days, or even years. There are roughly seven billion people on the planet, and the great majority of them use digital devices, such as mobile phones, desktop and laptop computers, wearable gadgets, and so on. Every one of these devices generates, captures, and stores data, generating around 2.5 quintillion bytes of data every day. In terms of Blu-ray DVDs, this is the equal of 10 million. Consider the many various sorts of data available: text, images, film, music, health data from wearable devices, and a plethora of other types of data from devices connected to the Internet of Things (IoT). Veracity: Because unstructured data constitutes 80 percent of all data, we must design methods for producing trustworthy and accurate insights from this data. It is necessary to categorize, analyze, and visualize the information. In today's world, data scientists are responsible for extracting insights from large data sets and dealing with the issues that these vast data sets provide. It is not possible to employ typical data analysis methods because of the sheer volume of information being gathered and stored. Alternative tools, such as those that make use of distributed computer capacity, can, however, circumvent this issue. Tools like as Apache Spark, Hadoop, and its ecosystem provide methods for extracting, loading, analyzing, and processing data across distributed computing resources, resulting in the generation of new insights and information. This provides enterprises with more opportunities to engage with their consumers and to improve the quality of the services they provide. As a result, the next time you put on your smartwatch or unlock your smartphone.

WHAT IS HADOOP?

Traditional computing and data processing involved bringing the data to the computer and processing it. You'd want to program, and you'd want to include the information in the program. Larry Page and Sergey Brin created a big data cluster by doing something very simple: they took the data and sliced it into pieces, then distributed each piece and replicated or tripled each piece, and then sent the pieces of these files to thousands of computers, first hundreds, then thousands, and now tens of thousands, as the data grew in size. Afterward, they would distribute the identical program to all of the machines within the cluster. And then each computer would run the software on its own small section of the file and relay the findings back to the other computers in turn. A sorting process would then take place, with the results of that sorting process being reallocated to another process. The first phase is referred to as a map or a mapper process, whereas the second process is referred to as a reduction process. Although the principles were very straightforward, it turned out that you could perform a plethora of various types of tasks and manage a plethora of different types of issues, as well as handle extremely, extremely, extremely enormous data sets. So the one advantage of these massive data clusters is that they increase linearly in size, which is a wonderful feature. Due to the fact that you have double the number of servers, you receive twice the performance and can manage twice the quantity of data. As a result, all of the main social media businesses were able to break through a bottleneck. After then, Yahoo joined the party. It was via the hiring of someone named Doug Cutting that Yahoo discovered that someone had been developing a duplicate of the Google big data architecture, which is now known as Hadoop. In fact, if you look up Hadoop on Google you'll see that it's become a very popular term and that there are hundreds of thousands of companies out there that have some kind of footprint in the big data world. If you look at the big data ecology, there are hundreds of thousands of companies out there that have some kind of footprint in the big data world.

WHAT MAKES SOMEONE A DATA SCIENTIST?

Now that you are aware of what is contained within the book, it is time to establish some definitions. Despite their widespread use, the concepts of big data and data science continue to elude understanding. Who is a data scientist, is a legitimate question. The debate is still very much alive, and it is being challenged by individuals, some of whom are only concerned with maintaining their own disciplines or academic turfs. In this part, I will make an attempt to answer and clarify these controversies. When we define Big data or Data science too narrowly, we risk dismissing hundreds of thousands of people who have just lately become interested in the nascent subject of data science.

According to Simon Rogers (2012), writing in the Guardian, "everyone likes a data scientist." Mr. Rogers also attributed the rekindled interest in number crunching to a comment from Google's Hal Varian, who predicted that statisticians will be the sexiest job in the next 10 years, according to Mr. Rogers.

While Hal Varian referred to statisticians as "sexy," it is usually assumed that he was referring to data scientists when he said this. This raises a number of crucial concerns, including:

What is data science and how does it work?

What is the difference between it and statistics?

What exactly qualifies someone as a data scientist?

In this age of big data, even a question as simple as "what is data science?" can be difficult to answer. It is possible to get a variety of responses. In other instances, the disparity in viewpoints on these responses is bordering on animosity.

A data scientist, in my opinion, is someone who finds answers to issues by analyzing large or tiny amounts of data using proper tools, and then develops stories to communicate her findings to the right stakeholders and audiences. I do not utilize the data size as a constraint in my programming. A data set that falls below a specific arbitrary threshold does not imply that one is a less qualified data scientist. My concept of a data scientist is not limited to certain analytic technologies, such as machine learning, and neither is it limited to a particular field of study. I consider someone to be a data scientist if they have an inquisitive mind, are fluent in analytics, and are able to effectively convey their discoveries.

Data science is something that data scientists do, according to my definition. Years ago, when I was a student at the University of Toronto studying engineering, I was perplexed by the question: What exactly is engineering? For my master's thesis, I looked at the predicting of housing prices, and for my doctorate research, I looked at the forecasting of homebuilders' decisions about what they construct, when they build it, and where they build it in the future. Others in the civil engineering department were working on the design of buildings, bridges, and tunnels, as well as worrying about the stability of slopes, among other things. This was not your typical engineering project, nor was it the product of my supervisor's team. It goes without saying that I was frequently asked by others if my study was indeed in the field of engineering.

DATA SCIENCE SOLUTIONS

This paper suggested that data science is what data scientists do, to put it succinctly.

Others have a slew of other interpretations. According to some professionals in the area, if you are not employing the black boxes that make up machine learning, you are not considered a data scientist at all. Even if you were to discover a treatment for a disease that was endangering the lives of millions of people, your turf-protecting colleagues would still bar you from joining the data science club because of your discovery.

Dr. Vincent Granville (2014), an author on data science, suggests that there are some requirements to achieve in order to become a data scientist. The new data science professor, according to Dr. Granville's description on pages 8 and 9 of Developing Analytic talent, is a non-tenured instructor at a non-traditional university who publishes research results in online blogs, does not waste time writing grants, works from home, and earns significantly more money than traditional tenured professors. It is sufficient to suggest that the vibrant academic community of data scientists may be at odds with Dr Granville's position.

Dr. Granville defines data science by imposing limitations on the amount of data and the techniques used to collect it. An individual who is capable of processing a So-million-row data collection in a few of hours and who does not believe in (statistical) models is defined as a data scientist by him. He makes a distinction between data science and statistics. Nonetheless, he considers algebra, calculus, and expertise in probability and statistics to be crucial prerequisites for understanding data science and statistics.

Some people assume that big data is just about exceeding a given threshold in terms of data size or the number of observations, or that it is about the usage of a specific tool, such as Hadoop, to achieve this goal. Such artificial limits on data size are troublesome since, because to technological advancements, even ordinary computers and off-the-shelf software are now capable of manipulating extremely massive data sets. Stata, a data science and statistics software package, has revealed that its desktop solutions can now process between 2 billion and 24.4 billion rows, an increase over the previous limit of 1 billion rows. The capacity of Stata to handle 24.4 billion rows under specific constraints has just crashed the big data party, if Hadoop is the secret password to the big data club.

It is critical to recognize that anybody who attempts to create arbitrary standards in order to exclude others will almost certainly encounter discrepancies. A more exclusive, discipline- and platform-independent, and size-free definition of data science should be sought, in which data-centric problem solving and the capacity to weave compelling narratives take center stage.

Given the controversies surrounding data scientists, I would want to engage with others to see how they define the term. Why don't we confer with the Chief Data Scientist of the

United States of America once more? Remember that Dr. Patil told the Guardian newspaper in 2012 that a data scientist is a person who possesses a unique combination of talents that allows them to both uncover the insights contained inside data and create a spectacular story using the data. That Dr. Patil's definition is inclusive of individuals with a variety of academic backgrounds and training is commendable; it does not limit the definition of a data scientist to the use of a particular tool or subject it to a certain arbitrary minimum threshold of data size, which is also commendable.

Curiosity is the other essential characteristic of a great data scientist, and it is a behavioral feature. A data scientist must have a very inquisitive mind and be prepared to put in substantial time and effort to test her hypotheses and hypothesis. It is referred regarded as "having a nose for news" in the journalistic industry. Not all journalists are aware of the latest developments in the news. The Story is only available to individuals who have a keen sense of curiosity. Curiosity is just as crucial for data scientists as it is for journalists when it comes to their work.

Rachel Schutt works as the Chief Data Scientist for News Corporation in New York. She is also an instructor at Columbia University, where she teaches a data science course. She is also the author of an outstanding book titled Doing Data Science, which is available on Amazon. In an interview with the New York Times, Dr. Schutt characterized a data scientist as someone who is a combination of computer scientist, software engineer, and statistician, among other qualifications (Miller, 2013). The average data scientist, on the other hand, is defined as follows: Her contention was that "the greatest tend to be highly inquisitive individuals, thinkers who ask smart questions, and are comfortable dealing with unstructured situations and attempting to find order in them."

CONCLUSION

It has been several decades since the majority of the components of data science have been in existence in some form or another. However, I believe that they are all coming together now with some additional subtleties. Probability and statistics are found at the bottom of the data science hierarchy. You'll see algebra, linear algebra, programming, and databases among the other things. They've all come to this place. As a result of recent technological advances, we now have the computer power to use certain novel approaches, such as machine learning. Now, instead of taking a sample and trying to test a theory, we may take extremely huge data sets and seek patterns, rather than taking a sample and trying to test a hypothesis. To do this, take a step back from hypothesis testing and look for patterns that could lead to the generation of hypotheses. Now, this can irritate some very conventional statisticians, who become really irritated when they realize that they are meant to have a hypothesis that is not reliant on the data, and then they must test that hypothesis. So, as some of these machine learning approaches became available, they were really the only thing that could be used to examine some of the really big social media data sets that were becoming available. The confluence of classic [technique] disciplines such as computer science, probability, statistics, and mathematics has resulted in the development of what we call Decision Sciences, which is an umbrella term for the study of decision making. Stern's department of linguistics As a small pitch, I'd want to point out that we've been fortunate in that we've been well-positioned among business schools since we're one of the few business schools that has a true statistics department with real Ph.D. level statisticians on staff. Operations management and information systems are separate departments within our organization. As a result, we have a diverse group of professionals, from computer scientists to statisticians to operations researchers. Because of this, we were ideally positioned when a handful of other business schools decided to jump on the bandwagon and declare that, indeed, Decision Sciences was the way to go. The NYU Center for Data Science was founded by Foster Provost, who works in my department and was the first head of the center.

REFERENCES

Granville, V. (2014). Developing Analytic Talent: Becoming a Data Scientist, John Wiley and Sons, Incorporated, US.

Miller, F. D. (2013). Aristotle on Belief and Knowledge. In Anagnostopoulos and Miller (Eds.), 285-307.

Pasupuleti, M. B. (2015). Data Science: The Sexiest Job in this Century. *International Journal of Reciprocal Symmetry and Physical Sciences*, 2, 8–11. Retrieved from https://upright.pub/index.php/ijrsps/article/view/56

Rogers, S. (2012). What is a data scientist? The Guardian. https://www.theguardian.com/news/datablog/2012/mar/02/data-scientist

--0--