What is Data Science?

Data Science is a process, not an event. It is the process of using data to understand different things, to understand the world. For me is when you have a model or hypothesis of a problem, and you try to validate that hypothesis or model with your data.

Data science is the art of uncovering the insights and trends that are hiding behind data. It's when you translate data into a story. So use storytelling to generate insight. And with these insights, you can make strategic choices for a company or an institution.

Data science is a field about processes and systems to extract data from various forms of whether it is unstructured or structured form.

Data science is the study of data. Like a biological science is a study of biology, physical sciences, it's the study of physical reactions. Data is real, data has real properties, and we need to study them if we're going to work on them.

Data Science involves data and some science. The definition or the name came up in the 80s and 90s when some professors were looking into the statistics curriculum, and they thought it would be better to call it data science. But what is Data Science? I'd see data science as one's attempt to work with data, to find answers to questions that they are exploring. In a nutshell, it's more about data than it is about science. If you have data, and you have curiosity, and you're working with data, and you're manipulating it, you're exploring it, the very exercise of going through analyzing data, trying to get some answers from it is data science. Data science is relevant today because we have tons of data available. We used to worry about lack of data. Now we have a data deluge. In the past, we didn't have algorithms, now we have algorithms. In the past, the software was expensive, now it's open source and free. In the past, we couldn't store large amounts of data, now for a fraction of the cost, we can have gazillions of datasets for a very low cost. So, the tools to work with data, the very availability of data, and the ability to store and analyze data, it's all cheap, it's all available, it's all ubiquitous, it's here. There's never been a better time to be a data scientist.

Fundamentals of Data science

Everyone you ask will give you a slightly different description of what Data Science is, but most people agree that it has a significant data analysis component.

Data analysis isn't new.

What is new is the vast quantity of data available from massively varied sources: from log files, email, social media, sales data, patient information files, sports performance data, sensor data, security cameras, and many more besides.

At the same time that there is more data available than ever, we have the computing power needed to make a useful analysis and reveal new knowledge.

Data science can help organizations understand their environments, analyze existing issues, and reveal previously hidden opportunities.

Data scientists use data analysis to add to the knowledge of the organization by investigating data, exploring the best way to use it to provide value to the business.

So, what is the process of data science? Many organizations will use data science to focus on a specific problem, and so it's essential to clarify the question that the organization wants answered.

This first and most crucial step defines how the data science project progresses. Good data scientists are curious people who ask questions to clarify the business need. The next questions are: "what data do we need to solve the problem, and where will that data come from?".

Data scientists can analyze structured and unstructured data from many sources, and depending on the nature of the problem, they can choose to analyze the data in different ways.

Using multiple models to explore the data reveals patterns and outliers; sometimes, this will confirm what the organization suspects, but sometimes it will be completely new knowledge, leading the organization to a new approach.

When the data has revealed its insights, the role of the data scientist becomes that of a storyteller, communicating the results to the project stakeholders.

Data scientists can use powerful data visualization tools to help stakeholders understand the nature of the results, and the recommended action to take.

Data Science is changing the way we work; it's changing the way we use data and it’s changing the way organisations understand the world.

Advice for new Data scientists

My advice to an aspiring data scientist is to be curious, extremely argumentative and judgmental. Curiosity is absolute must. If you're not curious, you would not know what to do with the data. Judgmental because if you do not have preconceived notions about things you wouldn't know where to begin with. Argumentative because if you can argument and if you can plead a case, at least you can start somewhere and then you learn from data and then you modify your assumptions and hypotheses and your data would help you learn. And you may start at the wrong point. You may say that I thought I believed this, but now with data I know this. So, this allows you a learning process. So, curiosity being able to take a position, strong position, and then moving forward with it. The other thing that the data scientist [should] would need is some comfort and flexibility with analytics platforms: some software, some computing platform, but that's secondary. The most important thing is curiosity and the ability to take positions. Once you have done that, once you've analyzed, then you've got some answers. And that's the last thing that a data scientist need, and that is the ability to tell a story. That once you have your analytics, once you have your tabulations, now you should be able to tell a great story from it. Because if you don't tell a great story from it, your findings will remain hidden, remain buried, nobody would know. Your rise to prominence is pretty much relying on your ability to tell great stories. A starting point would be to see what is your competitive advantage. Do you want to be a data scientist in any field or a specific field? Because, let's say you want to be a data scientist and work for an IT firm or a web-based or Internet based firm, then you need a different set of skills. And if you want to be a data scientist, for lets say, in the health industry, then you need different sets of skills. So figure out first what you're interested, and what is your competitive advantage. Your competitive advantage is not necessarily going to be your analytical skills. Your competitive advantage is your understanding of some aspect of life where you exceed beyond others in understanding that. Maybe it's film, maybe it's retail, maybe it's health, maybe it's computers. Once you've figured out where your expertise lies, then you start acquiring analytical skills. What platforms to learn and those platforms, those tools would be specific to the industry that you're interested in. And then once you have got some proficiency in the tools, the next thing would be to apply your skills to real problems, and then tell the rest of the world what you can do with it.

Data Science Topics and Algorithms

I really enjoy regression. I'd say regression was maybe one of the first concepts that I, that really helped me understand data so I enjoy regression.

I really like data visualization. I think it's a key element for people to get across their message to people that don't understand that well what data science is.

Artificial neural networks. I'm really passionate about neural networks because we have a lot to learn with nature so when we are trying to mimic our, our brain I think that we can do some applications with this behavior with this biological behavior in algorithms.

Data visualization with R. I love to do this. Nearest neighbor. It's the simplest but it just gets the best results so many more times than some overblown, overworked algorithm that's just as likely to overfit as it is to make a good fit.

So structured data is more like tabular data things that you’re familiar with in Microsoft Excel format. You've got rows and columns and that's called structured data.

Unstructured data is basically data that is coming from mostly from web where it's not tabular. It is not, it's not in rows and columns. It's text. It's sometimes it's video and audio, so you would have to deploy more sophisticated algorithms to extract data. And in fact, a lot of times we take unstructured data and spend a great deal of time and effort to get some structure out of it and then analyze it. So if you have something which fits nicely into tables and columns and rows, go ahead. That's your structured data. But if you see if it's a weblog or if you're trying to get information out of webpages and you've got a gazillion web pages, that's unstructured data that would require a little bit more effort to get information out of it.

There are thousands of books written on regression and millions of lectures delivered on regression. And I always feel that they don’t do a good job of explaining regression because they get into data and models and statistical distributions. Let's forget about it.

Let me explain regression in the simplest possible terms. If you have ever taken a cab ride, a taxi ride, you understand regression. Here is how it works. The moment you sit in a cab ride, in a cab, you see that there's a fixed amount there. It says $2.50. You, rather the cab, moves or you get off. This is what you owe to the driver the moment you step into a cab. That's a constant. You have to pay that amount if you have stepped into a cab. Then as it starts moving for every meter or hundred meters the fare increases by certain amount. So there's a... there's a fraction, there's a relationship between distance and the amount you would pay above and beyond that constant. And if you're not moving and you're stuck in traffic, then every additional minute you have to pay more. So as the minutes increase, your fare increases. As the distance increases, your fare increases. And while all this is happening you've already paid a base fare which is the constant. This is what regression is. Regression tells you what the base fare is and what is the relationship between time and the fare you have paid, and the distance you have traveled and the fare you've paid. Because in the absence of knowing those relationships, and just knowing how much people traveled for and how much they paid, regression allows you to compute that constant that you didn't know. That it was $2.50, and it would compute the relationship between the fare and and the distance and the fare and the time. That is regression.

Cloud for Data Science

Cloud is a godsend for data scientists. Primarily because you're able to take [the] your data, take your information and put it in the Cloud, put it in a central storage system. It allows you to bypass the physical limitations of the computers and the systems you're using and it allows you to deploy the analytics and storage capacities of advanced machines that do not necessarily have to be your machine or your company's machine.

Cloud allows you not just to store large amounts of data on servers somewhere in California or in Nevada, but it also allows you to deploy very advanced computing algorithms and the ability to do high-performance computing using machines that are not yours.

Think of it as you have some information, you can't store it, so you send it to storage space, let's call it Cloud, and the algorithms that you need to use, you don't have them with you. But then on the Cloud, you have those algorithms available. So What you do is you deploy those algorithms on very large datasets and you're able to do it even though your own systems, your own machines, your own computing environments were not allowing you to do so.

So Cloud is beautiful. The other thing that Cloud is beautiful for is that it allows multiple entities to work with same data at the same time. You can be working with the same data that your colleagues in say Germany and another team in India and another team in Ghana, they are collectively working and they're able to do so because the information, and the algorithms, and the tools, and the answers, and the results, whatever they needed is available at a central place, which we call Cloud.

Cloud is beautiful. Using the Cloud enables you to get instant access to open source technologies like Apache Spark without the need to install and configure them locally. Using the Cloud also gives you access to the most up-to-date tools and libraries without the worry of maintaining them and ensuring that they are up to date.

The Cloud is accessible from everywhere and in every time zone.

You can use cloud-based technologies from your laptop, from your tablet, and even from your phone, enabling collaboration more easily than ever before. Multiple collaborators or teams can access the data simultaneously, working together on producing a solution. Some big tech companies offer Cloud platforms, allowing you to become familiar with cloud-based technologies in a pre-built environment. IBM offers the IBM Cloud, Amazon offers Amazon Web Services or AWS, and Google offers Google Cloud platform. IBM also provides Skills Network labs or SN labs to learners registered at any of the learning portals on the IBM Developer Skills Network, where you have access to tools like Jupyter Notebooks and Spark clusters so you can create your own data science project and develop solutions. With practice and familiarity, you will discover how the Cloud dramatically enhances productivity for data scientists.

What Makes Someone a Data Scientist?

Now that you know what is in the book, it is time to put down some definitions. Despite their ubiquitous use, consensus evades the notions of Big data and Data Science. The question, Who is a data scientist? is very much alive and being contested by individuals, some of whom are merely interested in protecting their discipline or academic turfs. In this section, I attempt to address these controversies and explain Why a narrowly construed definition of either Big data or Data science will result in excluding hundreds of thousands of individuals who have recently turned to the emerging field.

Everybody loves a data scientist, wrote Simon Rogers (2012) in the Guardian. Mr. Rogers also traced the newfound love for number crunching to a quote by Google's Hal Varian, who declared that the sexy job in the next ten years will be statisticians.

Whereas Hal Varian named statisticians sexy, it is widely believed that what he really meant were data scientists. This raises several important questions:

What is data science?

How does it differ from statistics?

What makes someone a data scientist?

In the times of big data, a question as simple as, What is data science? can result in many answers. In some cases, the diversity of opinion on these answers borders on hostility.

I define a data scientist as someone who finds solutions to problems by analyzing Big or small data using appropriate tools and then tells stories to communicate her findings to the relevant stakeholders. I do not use the data size as a restrictive clause. A data below a certain arbitrary threshold does not make one less of a data scientist. Nor is my definition of a data scientist restricted to particular analytic tools, such as machine learning. As long as one has a curious mind, fluency in analytics, and the ability to communicate the findings, I consider the person a data scientist.

I define data science as something that data scientists do. Years ago, as an engineering student at the University of Toronto, I was stuck With the question: What is engineering? I wrote my master's thesis on forecasting housing prices and my doctoral dissertation on forecasting homebuilders' choices related to What they build, when they build, and where they build new housing. In the civil engineering department, Others were working on designing buildings, bridges, tunnels, and worrying about the stability of slopes. My work, and that of my supervisor, was not your traditional garden-variety engineering. Obviously, I was repeatedly asked by others whether my research was indeed engineering.

When I shared these concerns with my doctoral supervisor, Professor Eric Miller, he had a laugh. Dr Miller spent a lifetime researching urban land use and transportation and had earlier earned a doctorate from MIT. "Engineering is what engineers do," he responded. Over the next 17 years, I realized the wisdom in his statement. You first become an engineer by obtaining a degree and then registering with the local professional body that regulates the engineering profession. Now you are an engineer. You can dig tunnels; write software codes; design components of an iPhone or a supersonic jet. You are an engineer. And when you are leading the global response to a financial crisis in your role as the chief economist of the International Monetary Fund (IMF), as Dr Raghuram Rajan did, you are an engineer.

Professor Raghuram Rajan did his first degree in electrical engineering from the Indian Institute of Technology. He pursued economics in graduate studies, later became a professor at a prestigious university, and eventually landed at the IMF. He is currently serving as the 23rd Governor of the Reserve Bank of India. Could someone argue that his intellectual prowess is rooted only in his training as an economist and that the fundamentals he learned as an engineering student played no role in developing his problem-solving abilities?

Professor Rajan is an engineer. So are Xi Jinping, the President of the People's Republic of China, and Alexis Tsipras, the Greek Prime Minister who is forcing the world to rethink the fundamentals of global economics. They might not be designing new circuitry, distillation equipment, or bridges, but they are helping build better societies and economies and there can be no better definition of engineering and engineers—that is, individuals dedicated to building better economies and societies.

So briefly, I would argue that data science is what data scientists do.

Others have many different definitions. In September 2015, a co-panelist at a meetup organized by BigDataUniversity.com in Toronto confined data science to machine learning. There you have it. If you are not using the black boxes that makeup machine learning, as per some experts in the field, you are not a data scientist. Even if you were to discover the cure to a disease threatening the lives of millions, turf-protecting colleagues will exclude you from the data science club.

Dr Vincent Granville (2014), an author on data science, offers certain thresholds to meet to be a data scientist. On pages 8 and 9 in Developing Analytic talent, Dr Granville describes the new data science professor as 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 more money than the traditional tenured professors. Suffice it to say that the thriving academic community of data scientists might disagree with Dr Granville.

Dr Granville uses restrictions on data size and methods to define what data science is. He defines a data scientist as one who can easily process a So-million-row data set in a couple of hours, and who distrusts (statistical) models. He distinguishes data science from statistics. Yet he lists algebra, calculus, and training in probability and statistics as necessary background to understand data science (page 4).

Some believe that big data is merely about crossing a certain threshold on data size or the number of observations, or is about the use of a particular tool, such as Hadoop. Such arbitrary thresholds on data size are problematic because, with innovation, even regular computers and off-the-shelf software have begun to manipulate very large data sets. Stata, a commonly used software by data scientists and statisticians, announced that one could now process between 2 billion to 24.4 billion rows using its desktop solutions. If Hadoop is the password to the big data club, Stata's ability to process 24.4 billion rows, under certain limitations, has just gatecrashed that big data party.

It is important to realize that one who tries to set arbitrary thresholds to exclude others is likely to run into inconsistencies. The goal should be to define data science in a more exclusive, discipline- and platform-independent, size-free context where data-centric problem solving and the ability to weave strong narratives take center stage.

Given the controversy, I would rather consult others to see how they describe a data scientist. Why don't we again consult the Chief Data Scientist of the United States? Recall Dr Patil told the Guardian newspaper in 2012 that a data scientist is that unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data. What is admirable about Dr Patil's definition is that it is inclusive of individuals of various academic backgrounds and training, and does not restrict the definition of a data scientist to a particular tool or subject it to a certain arbitrary minimum threshold of data size.

The other key ingredient for a successful data scientist is a behavioral trait: curiosity. A data scientist has to be one with a very curious mind, willing to spend significant time and effort to explore her hunches. In journalism, the editors call it having the nose for news. Not all reporters know where the news lies. Only those Who have the nose for news get the Story. Curiosity is equally important for data scientists as it is for journalists.

Rachel Schutt is the Chief Data Scientist at News Corp. She teaches a data science course at Columbia University. She is also the author of an excellent book, Doing Data Science. In an interview With the New York Times, Dr Schutt defined a data scientist as someone who is a part computer scientist, part software engineer, and part statistician (Miller, 2013). But that's the definition of an average data scientist. "The best", she contended, "tend to be really curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them."

**Week 2**

Foundations of Big Data

In this digital world, everyone leaves a trace. From our travel habits to our workouts and entertainment, the increasing number of internet connected devices that we interact with on a daily basis record vast amounts of data about us. There’s even a name for it: Big Data.

Ernst and Young offers the following definition: “Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.”

There is no one definition of Big Data, but there are certain elements that are common across the different definitions, such as velocity, volume, variety, veracity, and value. These are the V's of Big Data.

Velocity is the speed at which data accumulates. Data is being generated extremely fast, in a process that never stops. Near or real-time streaming, local, and cloud-based technologies can process information very quickly.

Volume is the scale of the data, or the increase in the amount of data stored. Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure. Variety is the diversity of the data. Structured data fits neatly into rows and columns, in relational databases while unstructured data is not organized in a pre-defined way, like Tweets, blog posts, pictures, numbers, and video. Variety also reflects that data comes from different sources, machines, people, and processes, both internal and external to organizations. Drivers are mobile technologies, social media, wearable technologies, geo technologies, video, and many, many more.

Veracity is the quality and origin of data, and its conformity to facts and accuracy. Attributes include consistency, completeness, integrity, and ambiguity. Drivers include cost and the need for traceability. With the large amount of data available, the debate rages on about the accuracy of data in the digital age. Is the information real, or is it false?

Value is our ability and need to turn data into value. Value isn't just profit. It may have medical or social benefits, as well as customer, employee, or personal satisfaction. The main reason that people invest time to understand Big Data is to derive value from it.

Let's look at some examples of the V's in action.

Velocity: Every 60 seconds, hours of footage is uploaded to YouTube which is generating data. Think about how quickly data accumulates over hours, days, and years.

Volume: The world population is approximately seven billion people and the vast majority are now using digital devices; mobile phones, desktop and laptop computers, wearable devices, and so on. These devices all generate, capture, and store data -- approximately 2.5 quintillion bytes every day. That's the equivalent of 10 million Blu-ray DVD's.

Variety: Let's think about the different types of data; text, pictures, film, sound, health data from wearable devices, and many different types of data from devices connected to the Internet of Things.

Veracity: 80% of data is considered to be unstructured and we must devise ways to produce reliable and accurate insights. The data must be categorized, analyzed, and visualized. Data Scientists today derive insights from Big Data and cope with the challenges that these massive data sets present.

The scale of the data being collected means that it’s not feasible to use conventional data analysis tools. However, alternative tools that leverage distributed computing power can overcome this problem.

Tools such as ***Apache Spark, Hadoop*** and its ecosystem provide ways to extract, load, analyze, and process the data across distributed compute resources, providing new insights and knowledge. This gives organizations more ways to connect with their customers and enrich the services they offer. So next time you strap on your smartwatch, unlock your smartphone, or track your workout, remember your data is starting a journey that might take it all the way around the world, through big data analysis, and back to you.

What is Hadoop?

Traditionally in computation and processing data we would bring the data to the computer. You'd wanna program and you'd bring the data into the program.

In a big data cluster what Larry Page and Sergey Brin came up with is very pretty simple is they took the data and they sliced it into pieces and they distributed each and they replicated each piece or triplicated each piece and they would send it the pieces of these files to thousands of computers first it was hundreds but then now it's thousands now it's tens of thousands. And then they would send the same program to all these computers in the cluster. And each computer would run the program on its little piece of the file and send the results back. The results would then be sorted and those results would then be redistributed back to another process. The first process is called a map or a mapper process and the second one was called a reduce process. Fairly simple concepts but turned out that you could do lots and lots of different kinds of handle lots and lots of different kinds of problems and very, very, very large data sets. So the one thing that's nice about these big data clusters is they scale linearly. You had twice as many servers and you get twice the performance and you can handle twice the amount of data. So this was just broke a bottleneck for all the major social media companies. Yahoo then got on board. Yahoo hired someone named Doug Cutting who had been working on a clone or a copy of the Google big data architecture and now that's called Hadoop. And if you google Hadoop you'll see that it's now a very popular term and there are many, many, many 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.

Most of the components of data science have been around for many, many, many, many decades. But they're all coming together now with some new nuances I guess. ***At the bottom of data science you see probability and statistics. You see algebra, linear algebra you see programming and you see databases.*** They've all been here. But what's happened now is we now have the computational capabilities to apply some new techniques - machine learning. Where now we can take really large data sets and instead of taking a sample and trying to test some hypothesis we can take really, really large data sets and look for patterns. And so back off one level from hypothesis testing to finding patterns that maybe will generate hypotheses. Now this can bother some very traditional statisticians and gets them really annoyed sometimes that you know you're supposed to have a hypothesis that is not that is independent of the data and then you test it. *So once some of these machine learning techniques started were really the only thing the only way you can analyze some of these really large social media data sets. So what we've seen is that the combination of traditional [technique] areas computer science probability, statistics, mathematics all coming together in this thing that we call Decision Sciences.*

Our department at Stern I'll give a little plug here we happen to have been very well situated among business schools because we're one of the few business schools that has a real statistics department with real PhD level statisticians in it. We have an operations management department and an information systems department. So we have a wide range of computer scientists to statisticians, to operations researchers. And so we were like perfectly positioned as a couple of other business schools were to jump on this bandwagon and say; okay this is Decision Sciences. And Foster Provost who's in my department was the first director of the NYU Center for Data Science.

Four years ago maybe five years ago. I mean, I feel this is one of those cases where you can just to Google and search for data science and see how often it occurred and you'll see almost nothing and then just a spike. The same thing you would see with big data about seven or eight years ago. So data science is a term I haven't heard of probably five years ago.

The first question is what is it? And I think faculty and everybody is still trying to get their hands around exactly what is business analytics and what is data science. We certainly know the components of it. But it's morphing and changing and growing. I mean the last three years deep learning has just been added into the mix. Neural networks have been around for 20 or 30 years. 20 years ago, I would teach neural networks in a class and you really couldn't do very much with them. And now some researchers have come up with multi-layer neural networks in Toronto in particular the University of Toronto. And that technology is now rapidly expanding it's being used by Google, by Facebook, by lots of companies.

***Hadoop*** is an **open-source software framework for storing data and running applications on clusters of** commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

***Apache Hadoop*** is **an open source framework that is used to efficiently store and process large datasets** ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

How Big Data is Driving Digital Transformation

Digital Transformation affects business operations, updating existing processes and operations and creating new ones to harness the benefits of new technologies. This digital change integrates digital technology into all areas of an organization resulting in fundamental changes to how it operates and delivers value to customers. It is an organizational and cultural change driven by Data Science, and especially Big Data. The availability of vast amounts of data, and the competitive advantage that analyzing it brings, has triggered digital transformations throughout many industries.

Netflix moved from being a postal DVD lending system to one of the world’s foremost video streaming providers, the Houston Rockets NBA team used data gathered by overhead cameras to analyze the most productive plays, and Lufthansa analyzed customer data to improve its service.

Organizations all around us are changing to their very core. Let’s take a look at an example, to see how Big Data can trigger a digital transformation, not just in one organization, but in an entire industry. In 2018, the Houston Rockets, a National Basketball Association, or NBA team, raised their game using Big Data. The Rockets were one of four NBA teams to install a video tracking system which mined raw data from games. They analyzed video tracking data to investigate which plays provided the best opportunities for high scores, and discovered something surprising. Data analysis revealed that the shots that provide the best opportunities for high scores are two-point dunks from inside the two-point zone, and three-point shots from outside the three-point line, not long-range two-point shots from inside it. This discovery entirely changed the way the team approached each game, increasing the number of three-point shots attempted. In the 2017-18 season, the Rockets made more three-point shots than any other team in NBA history, and this was a major reason they won more games than any of their rivals. In basketball, Big Data changed the way teams try to win, transforming the approach to the game. Digital transformation is not simply duplicating existing processes in digital form; the in-depth analysis of how the business operates helps organizations discover how to improve their processes and operations, and harness the benefits of integrating data science into their workflows.

Most organizations realize that digital transformation will require fundamental changes to their approach towards data, employees, and customers, and it will affect their organizational culture. *Digital transformation impacts every aspect of the organization, so it is handled by decision makers at the very top levels to ensure success.* The support of the Chief Executive Officer is crucial to the digital transformation process, as is the support of the Chief Information Officer, and the emerging role of Chief Data Officer. But they also require support from the executives who control budgets, personnel decisions, and day-to-day priorities. This is a whole organization process. Everyone must support it for it to succeed. There is no doubt dealing with all the issues that arise in this effort requires a new mindset, but Digital Transformation is the way to succeed now and in the future.

Data Science Skills & Big Data

I'm Norman White, I'm a Clinical Faculty Member in the IOMS Department, Information, Operations and Management Science Department here at Stern. I've been here for a long time (laughs), since I got out of college, pretty much. I'm sort of a techy, geeky kind of person. I really like to play with technology in my spare time. I'm currently Faculty Director of the Stern Center for Research Computing, in which we have a private cloud that runs lots of different kinds of systems. Many of our faculty or PhD students who need specialized hardware and software will come to us, we'll spin up a machine for them, configure it, I'll help them and advise on them. A lot of the data scientists, or virtually all the data scientists at Stern use our facilities. And their PhD students use them a lot. I have an undergraduate degree in Applied Physics and while I was an undergrad I took a number of economics courses, so I ended up deciding to go to business school, but I had, this was in the early days of computers (laughs) and I had gotten interested in computers. I came to Stern, which was then NYU Business School downtown and they had a little computer center, and I decided that I was gonna learn two things while I was there. One, I was gonna learn how to program. I had taken one programming course in college. And I was gonna learn how to touch type. I never did learn how to touch type (laughs). Or maybe I did but I've forgotten now, and back to two finger pecking. But I became a self taught programmer, and then I took a number of courses at IBM because I eventually came the director of the computer center, while I was getting my PhD in Economics and Statistics at Stern. In 1973, the school formed a department called Computer Applications and Information Systems and I was one of the first faculty members in the department and I've been here ever since (laughs). My typical Monday is, I usually get in around 11 o'clock and I do my email at home first, but I come in and I have two classes on Monday. I have a class on design and development of web based systems at six o'clock. Two o'clock, I have a dealing with data class. The class is based on Python notebooks, so we start with the basics of Unix and Linux, just to get the students used to that. We move onto some Python, some regular expressions, a lot of relational databases, some Python Pandas, which is sort of like R for Python, lets you do mathematical and statistical calculations in Python. And then I end up with big data, for which, as you probably know, I'm an evangelist. The students I have, weekly homeworks. I put them in teams and they have to do a big project at the end of the term, and they do some really cool things. Yes, in fact, the whole course is taught using Jupyter notebooks. Every student has their own virtual machine on Amazon Web Services, so we pre configure all the machines and they get a standard image that has all of the materials for the course either loaded on it or in a Jupyter notebook, there are the commands to download it or update the server with the right software. So everybody is in the same environment, it doesn't matter what kind of, whether they have a Mac or a Windows machine or how old it is, everybody can do everything in the class.

Data Mining

Establishing Data Mining Goals

* The cost-benefit trade-offs for the desired level of accuracy are important considerations for data mining goals.

Selecting Data

* The output of a data-mining exercise largely depends upon the quality of data being used.
* On the other hand, data may not be readily available for data mining. In such cases, you must identify other sources of data or even plan new data collection initiatives, including surveys.
* The type of data, its size, and frequency of collection have a direct bearing on the cost of data mining exercise.
* Therefore, identifying the right kind of data needed for data mining that could answer the questions at reasonable costs is critical.

Preprocessing Data

* Often raw data are messy, containing erroneous or irrelevant data. In addition, even with relevant data, information is sometimes missing.
* In the preprocessing stage, you identify the irrelevant attributes of data and expunge such attributes from further consideration.
* At the same time, identifying the erroneous aspects of the data set and flagging them as such is necessary.
* Data should be subject to checks to ensure integrity.
* Lastly, you must develop a formal method of dealing with missing data and determine whether the data are missing randomly or systematically.
* If the data were missing randomly, a simple set of solutions would suffice.
* However, when data are missing in a systematic way, you must determine the impact of missing data on the results.
* Therefore, you must consider in advance if observations or variables containing missing data be excluded from the entire analysis or parts of it.

Transforming Data

* After the relevant attributes of data have been retained, the next step is to determine the appropriate format in which data must be stored.
* An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena.
* This may require transforming data *Data reduction algorithms,* such as *Principal Component Analysis* (demonstrated and explained later in the chapter), can reduce the number of attributes without a significant loss in information.
* In addition, variables may need to be transformed to help explain the phenomenon being studied.
* Often you need to transform variables from one type to another.
* It may be prudent to transform the continuous variable for income into a categorical variable where each record in the database is identified as low, medium, and high-income individual.
* This could help capture the non-linearities in the underlying behaviors.

Storing Data

* The transformed data must be stored in a format that makes it conducive for data mining.
* The data must be stored in a format that gives unrestricted and immediate read/write privileges to the data scientist.
* During data mining, new variables are created, which are written back to the original database, which is why the data storage scheme should facilitate efficiently reading from and writing to the database.
* It is also important to store data on servers or storage media that keeps the data secure and also prevents the data mining algorithm from unnecessarily searching for pieces of data scattered on different servers or storage media.
* Data safety and privacy should be a prime concern for storing data.

Mining Data

* After data is appropriately processed, transformed, and stored, it is subject to data mining.
* This step covers ***data analysis methods, including parametric and non-parametric methods, and machine-learning algorithms.***
* A good starting point for data mining is data visualization.
* Multidimensional views of the data using the advanced graphing capabilities of data mining software are very helpful in developing a preliminary understanding of the trends hidden in the data set.

Evaluating Mining Results

* After results have been extracted from data mining, you do a *formal evaluation* of the results.
* **Formal evaluation** could include testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data. This is known as an *"in-sample forecast".*
* In addition, the results are shared with the key stakeholders for feedback, which is then incorporated in the later iterations of data mining to improve the process.
* Data mining and evaluating the results becomes an iterative process such that the analysts use better and improved algorithms to improve the quality of results generated in light of the feedback received from the key stakeholders.

Deep Learning and Machine Learning

In data science, there are many terms that are used interchangeably, so let's explore the most common ones. The term big data refers to data sets that are so massive, so quickly built, and so varied that they defy traditional analysis methods such as you might perform with a relational database. The concurrent development of enormous compute power in distributed networks and new tools and techniques for data analysis means that organizations now have the power to analyze these vast data sets. A new knowledge and insights are becoming available to everyone.

Big data are often described in terms of five V's; velocity, volume, variety, veracity, and value.

***Data mining*** is the process of automatically searching and analyzing data, discovering previously unrevealed patterns. It involves preprocessing the data to prepare it and transforming it into an appropriate format. Once this is done, insights and patterns are mined and extracted using various tools and techniques ranging from simple data visualization tools to machine learning and statistical models.

***Machine learning*** is a subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it is learned without being explicitly programmed. Machine learning algorithms are trained with large sets of data and they learn from examples. They do not follow rules-based algorithms. Machine learning is what enables machines to solve problems on their own and make accurate predictions using the provided data.

***Deep learning*** is a specialized subset of machine learning that uses layered neural networks to simulate human decision-making. Deep learning algorithms can label and categorize information and identify patterns.

***Artificial neural networks***, often referred to simply as neural networks, take inspiration from biological neural networks, although they work quite a bit differently. A neural network in AI is a collection of small computing units called neurons that take incoming data and learn to make decisions over time.

***Neural networks*** are often layered - deep and is the reason deep learning algorithms become more efficient as the data sets increase in volume, as opposed to other machine learning algorithms that may plateau as data increases.

***Data Science*** is the process and method for extracting knowledge and insights from large volumes of disparate data. It's an interdisciplinary field involving mathematics, statistical analysis, data visualization, machine learning, and more. It's what makes it possible for us to appropriate information, see patterns, find meaning from large volumes of data and use it to make decisions that drive business.

Data Science can use many of the AI techniques to derive insight from data. For example, it could use machine learning algorithms and even deep learning models to extract meaning and draw inferences from data.

There is some interaction between AI and Data Science, but one is not a subset of the other. Rather, Data Science is a broad term that encompasses the entire data processing methodology while AI includes everything that allows computers to learn how to solve problems and make intelligent decisions.

Both AI and Data Science can involve the use of big data. That is, significantly large volumes of data.

Applications of Machine Learning

Everybody now deals with machine learning. But recommender systems are certainly one of the major applications. Classifications, cluster analysis, trying to find some of the marketing questions from 20 years ago, market basket analysis, what goods tend to be bought together. That was computationally a very difficult problem, I mean we're now doing that all the time with machine learning.

So predictive analytics is another area of machine learning. We're using new techniques to predict things that statisticians don't particularly like. Decision trees, Bayesian Analysis, naive Bayes, lots of different techniques. The nice thing about them is that in packages like R now, you really have to understand how these techniques can be used and you don't have to know exactly how to do them but you have to understand what their meanings are.

Precision versus recall and the problems of over sampling and over fitting so you can, someone who knows a little about data science can apply these techniques but they really need to know, maybe not the details of the technique as much as how, what the trade-offs are.

So, some applications of machine learning in fintech are probably the - couple of different things I could talk about there. One of them is recommendations. Right, so, when you use Netflix, or you use Facebook, or a lot of different software services, the recommendations are served to you. Meaning, "Hey, you're a user, you've watched this show, so maybe you'd like to see this other show." Right, or, you follow this person, so maybe you should follow this other person.

It's actually kind of the same thing in fintech, right. Because you've looked at - if you're an investment professional, right, and because you've looked at this investment idea, it might be really cool for you to look at this other investment idea, which is kind of similar. Right, it's a similar kind of asset, it's a similar kind of company. Or it's a similar kind of technique for doing the investment. So, We can apply recommendations using machine learning throughout a lot of different parts of fintech.

Another one that people talk about, and is important especially on retail, in the retail aspects of banking and finance is fraud detection. Trying to determine whether a charge that comes a credit card is fraudulent or not, in real time, is a machine learning problem. Right, you have to learn from all of the transactions that have happened previously and build a model, and when the charge comes through you have to compute all this stuff and say, "Yeah we think that's ok," or "hmm, that's not so good. Let's route it to, you know, our fraud peope to check."

This is not to suggest that children born to tall parents are not necessarily taller than the rest. That may be the case, but they are not necessarily taller than their own "tall" parents. Why I think this to be a good thing requires a simple mental simulation. Imagine if every successive generation born to tall parents were taller than their parents, in a matter of a couple of millennia, human beings would become uncomfortably tall for their own good, requiring even bigger furniture, cars, and planes.

The workings of regression models, which have become the workhorse of statistical analysis. In almost all empirical pursuits of research, either in the academic or professional fields, the use of regression models, or their variants, is ubiquitous. In medical science, regression models are being used to develop more effective medicines, improve the methods for operations, and optimize resources for small and large hospitals. In the business world, regression models are at the forefront of analyzing consumer behavior, firm productivity, and competitiveness of public and private­ sector entities.

my research has not just found the correlation between housing prices and the size of housing units, but I have also discovered the magnitude of those relationships. For instance, I found that all else being equal, a term that I explain later in this chapter, an additional washroom adds more to the housing price than an additional bedroom. Stated otherwise, the marginal increase in the price of a house is higher for an additional washroom than for an additional bedroom. I found later that the real estate brokers in Toronto indeed appreciated this finding. I also explained to my wife that proximity to transport infrastructure, such as subways, resulted in higher housing prices.

For instance, houses situated closer to subways sold for more than did those situated farther away. However, houses near freeways or highways sold for less than others did. Similarly, I also discovered that proximity to large shopping centers had a nonlinear impact on housing prices.

Houses located very close (less than 2.5 km) to the shopping centers sold for less than the rest. However, houses located closer (less than 5 km, but more than 2.5 km) to the shopping center sold for more than did those located farther away.

I also found that the housing values in Toronto declined with distance from downtown.

**Use decision optimization**

Step 1: Import the docplex package

Step 2: Model the data

Step 3: Prepare the data

* Define how to compute the earth distance between 2 points
* To easily compute distance between 2 points, we use the Python package [geopy](https://pypi.python.org/pypi/geopy)
* Declare the list of libraries
* Parse the JSON file to get the list of libraries and store them as Python elements.
* Define number of shops to open[¶](https://eu-gb.dataplatform.cloud.ibm.com/exchange/public/entry/preview?url=https://raw.githubusercontent.com/IBMDataScience/sample-notebooks/master/Cloud/HTML/Finding%20optimal%20locations%20of%20new%20stores%20using%20DO.html#Define-number-of-shops-to-open)
* Create a constant that indicates how many coffee shops we would like to open.
* Validate the data by displaying them[¶](https://eu-gb.dataplatform.cloud.ibm.com/exchange/public/entry/preview?url=https://raw.githubusercontent.com/IBMDataScience/sample-notebooks/master/Cloud/HTML/Finding%20optimal%20locations%20of%20new%20stores%20using%20DO.html#Validate-the-data-by-displaying-them)
* We will use the [folium](https://folium.readthedocs.org/en/latest/quickstart.html#getting-started) library to display a map with markers.

**Week 3**

How Data Science is saving lives

Data Science provides targeted information to help healthcare professionals give the best treatment to patients, or help predict natural disasters so that people can prepare early, and much more besides. In healthcare, data scientists use predictive analytics developed from data mining, data modeling, statistics, and machine learning to find the best options for patients. This type of predictive analytics examines all known factors for a disease, including gene markers, associated conditions, and environmental factors. It then recommends appropriate tests, suitable trials, and any suggested treatments.

Every individual physician has their own store of knowledge gained from their studies, interests, and experiences. Data science systems that use predictive analytics ensure that all physicians can also access the latest information about the disease, tests, and treatment plans, tailored to their specific patient. With this type of system, every physician has access to the same knowledge, and the best options can be consistently offered, improving patient outcomes.

For example, a study by the Boston Consulting Group and AdvaMedDx, an industry association of medical diagnostics companies, examined the barriers to the adoption of potentially life saving diagnostic tests for patients with a specific cancer and a particular gene marker. The study discovered that the biggest factor in the patient being offered a specific test was the patient’s oncologist, who may or may not have known about the test and its relationship to the gene marker. By providing extra information through data science tools, physicians can be made aware of the most helpful tests and treatments for a specific patient.

There are many opportunities to explore other ways to mine data, such as from electronic medical records for different types of medical research. Schools such as the NorthShore University HealthSystem in suburban Chicago, a leader in the implementation of Electronic Medical Records (EMR) systems, now offer guidance on data mining. It is the first healthcare provider in America to be awarded the highest level of EMR deployment for both inpatient and outpatient care. This remarkable effort has generated much-anonymized data available for innovative analytics research. Developing more sophisticated big data analytics capabilities helps healthcare organizations move from basic descriptive analytics towards predictive insights, thanks to data science.

In the field of Disaster Preparedness, the ability to save lives using Data Science tools has been under development for many years. The use of predictive analytics tools is improving and providing new data analysis in a multitude of ways, alerting populations to danger faster than ever before.

Large, high-quality data sets can be used to predict the occurrence of numerous types of natural disasters, which can be the difference between life and death for thousands of people. Earthquakes, hurricanes & tornados, floods, and volcanic eruptions can be predicted with the help of data science.

Recent research at the University of Warwick in the UK used social media content such as photos and keywords to track the development of floods, hurricanes and other weather events. When added to the information recorded by scientists and weather stations, this type of data can be used to improve the predictions for localised weather events. Because the real benefit of this knowledge is so important, schools are starting to include this type of data science education in their curriculum. For instance, the University of Chicago Graham School offers a Master of Science course in Threat and Response Management.

Data science tools enable organizations to analyse vast quantities of data from widely different sources, and present that information in a way that allows data scientists to gain new knowledge, in some cases, saving hundreds of lives.

Applications of Data Science

Data science and big data are making an undeniable impact on businesses, changing day-to-day operations,financial analytics, and especially interactions with customers. It's clear that businesses can gain enormous value from the insights data science can provide. But sometimes it's hard to see exactly how. So let's look at some examples. In this era of big data, almost everyone generates masses of data every day, often without being aware of it. This digital trace reveals the patterns of our online lives.

If you have ever searched for or bought a product on a site like Amazon, you'll notice that it starts making recommendations related to your search. *This type of system known as a* ***recommendation engine*** *is a common application of data science.*

Companies like Amazon, Netflix, and Spotify use algorithms to make specific recommendations derived from customer preferences and historical behavior.

**Personal assistants like Siri** on Apple devices use data science to devise answers to the infinite number of questions end users may ask.

Google watches your every move in the world, you're online shopping habits, and your social media. Then it analyzes that data to create recommendations for restaurants, bars, shops, and other attractions based on the data collected from your device and your current location. Wearable devices like Fitbits, Apple watches, and Android watches add information about your activity levels, sleep patterns, and heart rate to the data you generate.

Now that we know how consumers generate data, let's take a look at how data science is impacting business.

In 2011, McKinsey & Company said that data science was going to become the key basis of competition. Supporting new waves of productivity, growth, and innovation.

In 2013, UPS announced that it was using data from customers, drivers, and vehicles, in a new route guidance system aimed to save time, money, and fuel. Initiatives like this support the statement that data science will fundamentally change the way businesses compete and operate.

How does a firm gain a competitive advantage?

Let's take Netflix as an example. Netflix collects and analyzes massive amounts of data from millions of users, including which shows people are watching at what time a day when people pause, rewind, and fast-forward, and which shows directors and actors they search for. Netflix can be confident that a show will be a hit before filming even begins by analyzing users preference for certain directors and acting talent, and discovering which combinations people enjoy. Add this to the success of earlier versions of a show and you have a hit.

For example, Netflix knew many of its users had streamed to the work of David Fincher. They also knew that films featuring Robin Wright had always done well, and that the British version of House of Cards was very successful. Netflix knew that significant numbers of people who liked Fincher also liked Wright. All this information combined to suggest that buying the series would be a good investment for the company. They were right. It was a huge hit. Thanks to data science, Netflix knows what people want before they do.

The Final Deliverable

The ultimate purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy. Analytics summarize findings in tables and plots. The data scientist should then use the insights to build the narrative to communicate the findings.

In academia, the final deliverable is in the form of essays and reports. Such deliverables are usually 1,000 to 7,000 words in length.

In consulting and business, the final deliverable takes on several forms. It can be a small document of fewer than 1,500 words illustrated with tables and plots, or it could be a comprehensive document comprising several hundred pages.

Large consulting firms, such as McKinsey and Deloitte,I routinely generate analytics-driven reports to communicate their findings and, in the process, establish their expertise in specific knowledge domains.

Let's review the "United States Economic Forecast", a publication by the Deloitte University Press. This document serves as a good example for a deliverable that builds narrative from data and analytics. The 24-page report focuses on the state of the U.S. economy as observed in December 2014. The report opens with a **grabber** highlighting the fact that contrary to popular perception, the economic and job growth has been quite robust in the United States. The report is not merely a statement of facts.

In fact, it is a carefully crafted report that cites Voltaire and follows a distinct theme. The report focuses on the **good news** about the U.S. economy. These include the increased investment in manufacturing equipment in the U.S. and the likelihood of higher consumer consumption resulting from lower oil prices.

The Deloitte report uses time series plots to illustrate trends in markets. The GDP growth chart shows how the economy contracted during the Great Recession and has rebounded since then. The graphic presents four likely scenarios for the future. Another plot shows the changes in consumer spending. The accompanying narrative focuses on income inequality in the U.S. and refers to Thomas Pikkety's book on the same. The Deloitte report mentions many consumers did not experience an increase in their real incomes over the years, while they still maintained their level of spending. Other graphics focused on housing, business, and government sectors, international trade, labor, and financial markets, and prices. The appendix carries four tables documenting data for the four scenarios discussed in the report.

Deloitte's "United States Economic Forecast" serves the very purpose that its authors intended. The report uses data and analytics to generate the likely economic scenarios. At the same time, the report shows Deloitte to be a competent firm capable of analyzing economic data and prescribing strategies to cope with the economic challenges.

Deloitte's report a good example of storytelling with data and encourage you to read the report to decide for yourself whether the deliverable would have been equally powerful without the narrative.

The initial planning and conceptualizing of the final deliverable is therefore extremely important for producing a compelling document. Embarking on analytics, without due consideration to the final deliverable, is likely to result in a poor-quality document where the analytics and narrative would struggle to blend.

Recruiting for Data Science

When the companies are hiring people for a data science team, maybe a data scientist or an analyst, or a chief data scientist, the tendency would be to find the person who has all the skills, that they know the domain-specific knowledge. They're excellent in analyzing structured and unstructured data. And they're great at presenting and they've got great storytelling skills. So if you put all this together, you will realize you're looking for a unicorn. And your odds of finding a unicorn are pretty rare. I think what you need to do to is to see, given the pool of applicants you have, who has the most resonance with your firm's DNA. Because you can teach analytics skills, anyone can learn analytics skills if they dedicate time and effort to it. But what really matters is who's passionate about the kind of business that you do. Someone could be a great data scientist in the retail environment, but they may not be that excited about working in IT related firms or working with gigabytes of weblogs. But if someone is excited about those weblogs, if someone is excited about health-related data then they would be able to contribute to your productivity much more so. And I would say if I'm looking for someone, if I have to put together a data science team, I would first look for curiosity. Is that person curious about things not just for data science but anything like, are they curious about why this room is painted a certain way, why do the bookshelves have books, and what kinds of books? They have to have a certain degree of curiosity about everything that is in their vision, that they look at. The second thing is do they have a sense of humor because, you see, you have to have a lighthearted about it. If someone is too serious about it, they probably would take it too seriously, and would not be able to look at the lighter elements. The third thing I think, and I think the last thing that I would look for if I had to have a hierarchy, the last thing I would look for are technical skills. I would go through the social skills, curiosity, and sense of humor. The ability to tell a story. The ability to know that there is a story there. And then once all is there then I would say, well, can you do the technical side of it? And if there is some hope or some sign of some technical skills, I would take them because I can train them in whatever skills they need. But I cannot teach curiosity. I cannot teach storytelling. I cannot certainly, instill sense of humor in anyone. >> I think there's no hard and fast rule for hiring data scientists. I think it's going to be a case by case thing. I would say there has to be some sort of technical component, somebody should be able to work with and manipulate the data. They should be able to communicate what they find in the data. I find quite often nobody really cares about the r-square or the confidence interval. So you have to be able to introduce those things and explain something in a compelling way. And they also have to find somebody who is relatable, because data science, it been typically new means that the person in that role has to make relationships and they have to work across different departments. >> If these data scientist has a good mathematics and statistics background.>> They have to consider like problem solving abilities and analysis. The scientist needs to be good in analyzing problems.>> The persons they are hiring, they should love to play with data. And then they know how to play with the data visualization. They have analytical thinking.>> When a company is hiring anyone to work on a data science team, they need to think about what role that person is going to take. Before a company begins, they need to understand what they want out of their data science team. And then they need to hire to begin it. As they grow a data science team, they need to understand whether they need engineers, architects, designers to work on visualization. Or whether they just need more people who can multiply large matrices. >> From a skills point of view, let's focus on the technical skills and in that case, first thing would be what kind of a technical platform would you like to adopt? Let's say you want to work in a structured data environment and let's say you want to work in market research. Then the type of skills you need are slightly different than someone who would like to work in big data environments. If you want to work in the traditional market research data, structure data environment, your skills should be some statistical knowledge and some knowledge of basic statistical algorithms, maybe some machine learning algorithms. And these are the tools that you would like to develop. If you want to work in big data, then there's the other aspect of it and that is to be able to store data. So you start with the expertise in storing large amounts of data. And then you look into platforms that allow you to do that. The next step would be to be able to manipulate large amounts of data, and the final step would be to apply algorithms to those large sets of data. So it's a three-step process. But most likely it starts, most importantly, it starts with where you would like to be, in what field, in what domain. In terms of platforms, let's you want to be in the traditional predictive analytics environment, and you're not working with big data, then R or Stata, or Python would be your tools. If you're working mostly with unstructured data, then Python is most suitable than R. If you're working with big data, then Hadoop and Spark are the environments that you will be working with. So it all depends upon where you would like to be and what kind of work excites you and then you pick your tools. In addition to technical skills, the second aspect of the data science is to have the ability to communicate. The communication skills or presentation skills. I call them story telling skills, that is that you have your analysis done, now can you tell a great story from it? If you have a very large table, can you synthesize this and make it more appealing that when it goes on the screen, or is it part of a document that it just speaks? It sings the findings and the reader just gets it right there. So the ability to present your findings, either verbally, or in a presentation, or in a document. So those communication and presentation skills are equally important as the technical skills are. When you have a grading side, when you're presenting your results, imagine you're driving on a mountain and then there's a sharp turn. And you can't see what's beyond the turn. And then you make that turn and then suddenly, you see a tremendous valley in front of you. And this great sense of awe, that I didn't know that, right? So when you present your findings and you have this great finding and you communicate it well, this is what people feel because they were not expecting it. They were not aware of it, and then this great sense of happiness that now I know. And I didn't know this, now I know. And then it empowers them, it gives them ideas, what they can do with this knowledge, this new insight. It's a great sense of joy. And you are able as a data scientist, you are able to share with your clients because you enabled it.

**The Report Structure**

Before starting the analysis, think about the structure of the report. Will it be a brief report of five or fewer pages, or will it be a longer document running more than 100 pages in length?

The structure of the report depends on the length of the document. A brief report is more to the point and presents a summary of key findings. A detailed report incrementally builds the argument and contains details about other relevant works, research methodology, data sources, and intermediate findings along with the main results.

I have reviewed reports by leading consultants, including Deloitte and McKinsey. I found that the length of the reports varied depending largely on the purpose of the report. Brief reports were drafted as commentaries on current trends and developments that attracted public or media attention. Detailed and comprehensive reports offered a critical review of the subject matter with extensive data analysis and commentary. Often, detailed reports collected new data or interviewed industry experts to answer the research questions.

Even if you expect the report to be brief, sporting five or fewer pages, I recommend that the deliverable follow a prescribed format including *the cover page, table of contents, executive summary, detailed contents, acknowledgments, references, and appendices (if needed).*

I often find the cover page to be missing in documents. It is not the inexperience of undergraduate students that is reflected in submissions that usually miss the cover page. In fact, doctoral candidates also require an explicit reminder to include an informative cover page. I hasten to mention that the business world sleuths are hardly any better. Just search the Internet for reports and you will find plenty of reports from reputed firms that are missing the cover page.

At a minimum, *the cover page should include the title of the report, names of authors, their affiliations, and contacts, the name of the institutional publisher (if any), and the date of publication.* I have seen numerous reports missing the date of publication, making it impossible to cite them without the year and month of publication. Also, from a business point of view, authors should make it easier for the reader to reach out to them. Having contact details at the front makes the task easier.

*"A table of contents (ToC)" is like a map needed for a trip never taken before.* You need to have a sense of the journey before embarking on it. A map provides a visual proxy for the actual travel with details about the landmarks that you will pass by in your trip. The ToC with main headings and lists of tables and figures offers a glimpse of what lies ahead in the document. *Never shy away from including a ToC, especially if your document, excluding cover page, table of contents, and references, is five or more pages in length.*

*Even for a short document, I recommend an "abstract" or an "executive summary".* Nothing is more powerful than explaining the crux of your arguments in three paragraphs or less. Of course, for larger documents running a few hundred pages, the executive summary could be longer.

*An "introductory section"* is always helpful in setting up the problem for the reader who might be new to the topic and who might need to be gently introduced to the subject matter before being immersed in intricate details.

A *good follow-up to the introductory* section is a review of available relevant research on the subject matter.

The *length of the literature review* section depends upon how contested the subject matter is. In instances where the vast majority of researchers have concluded in one direction, *the literature review could be brief with citations for only the most influential authors on the subject.* On the other hand, if the arguments are more nuanced with caveats aplenty, then you must cite the relevant research to offer adequate context before you embark on your analysis. You might use the literature review to highlight gaps in the existing knowledge, which your analysis will try to fill. This is where you formally introduce your research questions and hypothesis.

In *the "methodology" section*, you introduce the research methods and data sources you used for the analysis. If you have collected new data, explain the data collection exercise in some detail. You will refer to the literature review to bolster your choice for variables, data, and methods and how they will help you answer your research questions.

The *results section* is where you present your empirical findings. Starting with descriptive statistics (**see Chapter 4, "Serving Tables"**) and illustrative graphics (**see Chapter S, "Graphic Details" for plots and Chapter 10, "Spatial Data Analytics" for maps**), you will move toward formally testing your hypothesis (**see Chapter 6, "Hypothetically Speaking"**).

In case you need to *run statistical models,* you might turn to *regression models* (**see Chapter 7, "Why Tall Parents Don't Have Even Taller Children"**) or categorical analysis (**see Chapters 8, "To Be or Not to Be" and 2., "Categorically Speaking About Categorical Data"**).

If you are *working with time-series data*, you can turn to Chapter 11, **Doing Serious Time with Time Series.** You can also report results from other empirical techniques that fall under the general rubric of data mining (**see Chapter 12, "Data Mining for Gold"**). Note that many reports in the business sector present results in a more palatable fashion by holding back the statistical details and relying on illustrative graphics to summarize the results.

The *results section is followed by the discussion section*, where you craft your main arguments by building on the results you have presented earlier.

The *"discussion section"* is where you rely on the power of narrative to enable numbers to communicate your thesis to your readers. You refer the reader to the research question and the knowledge gaps you identified earlier. You highlight how your findings provide the ultimate missing piece to the puzzle.

Of course, not all analytics return a smoking gun. At times, more frequently than I would like to acknowledge, the results provide only a partial answer to the question and that, too, with a long list of caveats.

In the *"conclusion"* section, you generalize your specific findings and take on a rather marketing approach to promote your findings so that the reader does not remain stuck in the caveats that you have voluntarily outlined earlier. You might also identify future possible developments in research and applications that could result from your research.

What remains is housekeeping, including a *list of references, the acknowledgment section* (**acknowledging the support of those who have enabled your work is always good**), and "appendices", if needed.

Have You Done Your Job as a Writer?

As a data scientist, you are expected to do thorough analysis with the appropriate data, deploying the appropriate tools. As a writer, you are responsible for communicating your findings to the readers. Transport Policy, a leading research publication in transportation planning, offers a checklist for authors interested in publishing with the journal. The checklist is a series of questions authors are expected to consider before submitting their manuscripts to the journal. I believe the checklist is useful for budding data scientists and, therefore, I have reproduced it verbatim for their benefit.

* Have you told readers, at the outset, what they might gain by reading your paper?
* Have you made the aim of your work clear?
* Have you explained the significance of your contribution?
* Have you set your work in the appropriate context by giving sufficient background (including a complete set of relevant references) to your work?
* Have you addressed the question of practicality and usefulness?
* Have you identified future developments that might result from your work?
* Have you structured your paper in a clear and logical fashion?