Defining Data Science and What Data Scientists Do

- Defining Data Science
- What is Data Science?
- Fundamentals of Data Science
- The Many Paths to Data Science
- Advice for New Data Scientists
- Data Science: The Sexiest Job in the 21st Century

What Do Data Scientists Do?

- A day in the Life of a Data Scientist
- Old problems, new problems, Data Science solutions
- Data Science Topics and Algorithms
- What is the cloud?
- What Makes Someone a Data Scientist?

Data Science Topics

- Foundations of Big Data
- How Big Data is Driving Digital Transformation
- What is Hadoop?
- Data Science Skills & Big Data
- Data Scientists at New York University
- Data Mining
- Quiz: Data Mining

Deep Learning and Machine Learning

- What's the difference?
- Neural Networks and Deep Learning
- Applications of Machine Learning
- Regression
- Quiz: Regression

Data Science in Business

- Applications of Data Science
- How Data Science is Saving Lives
- How Should Companies Get Started in Data Science?
- Applications of Data Science
- The Final Deliverable

- Quiz: The Final Deliverable

Careers and Recruiting in Data Science

- How Can Someone Become a Data Scientist?
- Recruiting for Data Science
- Careers in Data Science
- High School Students and Data Science Careers

The Report Structure

- The Report Structure
- Quiz: The Report Structure
- Final Assignment

Coursera Community and Career Support

As a Data Science learner on Coursera, you have access to networking opportunities in Coursera's Professional Certificate Community and Data Science forums. Talk about what you're learning, ask questions, find peers to work with on projects, and share your career goals.

Post-Completion Career Support Services for Professional Certificates

Completing a Professional Certificate on Coursera unlocks access to a private Professional Certificate Alumni Resources community, which provides exclusive career support resources, including:

- Step-by-step **guide to ensure your success** at every stage of your job search.
- 1 year of **free access** to Big Interview's expert video lessons, resume builder, and interactive interview practice tools (a $79/month value).
- **A network** and support of fellow completers of Coursera Professional Certificates.
- A variety of **special offers** such as career coaching, webinars, and more.

After completing your Professional Certificate, you'll get an email telling you how to access these career support resources.

Questions? You can also always contact Coursera Careers Team by emailing career-support@coursera.org.

Defining Data Science Week-1

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 biological sciences 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 organizations understand the world.

The Many Paths to Data Science

Data science didn't really exist when I was growing up. It's not something that I ever woke up and said, I want to be a data scientist when I grow up. No, it didn't exist. I didn't know I would be working in data science.

When I grew up, there isn't that field called data science. And I think it's really new.

Data science didn't exist until 2009, 2011. Someone like DJ Patil or Andrew Gelman coined the term. Before that, there was statistics. And I didn't want to be any of those. I wanted to be in business. And then I found data science a heck of a lot more interesting.

I studied statistics, that's how I started. I went through many different stages in my life where I wanted to be a singer and then a doctor. And then I realized that I was good at math. So, I chose an area that was focused on quantitative analysis. And from then I do think that I wanted to work with data. Not necessarily data science as it's known today.

The first time that I had contact with data science, when I was my first year as a mechanical engineering. And strategic consulting firms, they use data science to make decisions. So, it was my first contact with data science.

I had a complicated problem that I needed to solve, and the usual techniques that we had at the time couldn't help with that problem.

I graduated with a math degree in the worst possible time, right after the economic crisis, and you actually had to be useful to get a job. So, I went and got a degree in statistics. And then I worked enough jobs that were called data scientist that I suddenly became one.

My undergraduate degree was in business, and I majored in politics, philosophy, and economics. And then I did a master's in business analytics at New York University at the Stern School of Business. When I left my undergrad, the first company I joined, it turned out that they were analyzing electronic point of sale data for retail manufacturers. And what we were doing was data science. But we only really started using that term much later. In fact, I'd say four or five years ago is when we started calling it analytics and data science.

I had several options for my internship here in Canada. And one of the options was to work with data science. I used to work with project development. But I think that was a good choice. And then I start my internship with data science.

I'm a civil engineer by training, so all engineers work with data. I would say the conventional use of data science in my life started with transportation research. I started building large models trying to forecast traffic on streets, trying to determine congestion and greenhouse gas emissions or tailpipe emissions. So, I think that's where my start was. And I started building these models when I was a graduate student at the University of Toronto. Started working with very large data sets, looking at household samples of, say, 150,000 households from half a million trips. And that, too, I'm speaking from mid 90s when this was supposed to be a very large data set, but not in today's terms. But that's how I started. I continued working with it. And then I moved to McGill University where I was a professor of transportation engineering. And I built even bigger data models that involved data and analytics. And so, I would say, yes, transportation research brought me to data science.

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 might 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 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. But 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 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: The Sexiest Job in the 21st Century

In the data-driven world, data scientists have emerged as a hot commodity. The chase is on to find the best talent in data science. Already, experts estimate that millions of jobs in data science might remain vacant for the lack of readily available talent. The global search for skilled data scientists is not merely a search for statisticians or computer scientists. In fact, the firms are searching for well-rounded individuals who possess the subject matter expertise, some experience in software programming and analytics, and exceptional communication skills.

Our digital footprint has expanded rapidly over the past 10 years. The size of the digital universe was roughly 130 billion gigabytes in 1995. By 2020, this number will swell to 40 trillion gigabytes. Companies will compete for hundreds of thousands, if not millions, of new workers needed to navigate the digital world. No wonder the prestigious *Harvard Business Review* called data science ***"the sexiest job in the 21st century."***

A report by the McKinsey Global Institute warns of huge talent shortages for data and analytics. "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."

Because the digital revolution has touched every aspect of our lives, the opportunity to benefit from learning about our behaviors is more so now than ever before. Given the right data, marketers can take sneak peeks into our habit formation. Research in neurology and psychology is revealing how habits and preferences are formed and retailers like Target are out to profit from it. However, the retailers can only do so if they have data scientists working for them. For this reason, it is "like an arms race to hire statisticians nowadays," said Andreas Weigend, the former chief scientist at Amazon.com."

There is still the need to convince the C-suite executives of the benefits of data and analytics. It appears that the senior management might be a step or two behind the middle management in being informed of the potential of analytics-driven planning. Professor Peter Fader, who manages at the Customer Analytics Initiative at Wharton, knows that executives reach the C-suite without having to interact with data. He believes that the real change will happen when executives are well-versed in data and analytics.

SAP, a leader in data and analytics, reported from a survey that 92% of the responding firms in its sample experienced a significant increase in their data holdings. At the same time, three-quarters identified the need for new data science skills in their firms. Accenture believes that the demand for data scientists may outstrip supply by 250,000 in 2015 alone. A similar survey of 150 executives by KPMG in 2014 found that 85% of the respondents did not know how to analyze data. "Most organizations are unable to connect the dots because they do not fully understand how data and analytics can transform their business," Alwin Magimay, head of digital and analytics for KPMG UK, said in an interview in May 2015.

Bernard Marr writing for *Forbes* also raises concerns about the insufficient analytics talent. "There just aren't enough people with the required skills to analyze and interpret this information-transforming it from raw numerical (or other) data into actionable insights-the ultimate aim of any Big Data-driven initiative," he wrote. Bernard quotes a survey by

Gartner of business leaders of whom more than 50% reported the lack of in-house expertise in data science.

Bernard reported on Walmart, which turned to crowdsourcing for its analytics need. Walmart approached Kaggle to host a competition for analyzing its proprietary data. The retailer provided sales data from a shortlist of stores and asked the competitors to develop better forecasts of sales based on promotion schemes.

Given the shortage of data scientists, the employers are willing to pay top dollars for the talent. Michael Chui, a principal at McKinsey, knows this too well. Data science "has become relevant to every company .. . There's a war for this type of talent," he said in an interview. Take Paul Minton, for example. He was making $20,000 serving tables at a restaurant. He had majored in math at college. Mr. Minton took a three-month programming course that changed everything. He made over $100,000 in 2014 as a data scientist for a web startup in San Francisco. "Six figures, right off the bat ... To me, it was astonishing," said Mr.Minton.

Could Mr. Minton be exceptionally fortunate, or are such high salaries the norm? Luck had little to do with it; the *New York Times* reported $100,000 as the average base salary of a software engineer and $ 112,000 for data scientists.

In this lesson, you have learned:

- Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices.
- There are many paths to a career in data science; most, but not all, involve a little math, a little science, and a lot of curiosity about data.
- New data scientists need to be curious, judgmental and argumentative.
- Why data science is considered the sexiest job in the 20th century, paying high salaries for skilled workers.

What do Data Scientists Do?

A day in the Life of a Data Scientist

I've built a recommendation engine before as part of a large organization and worked through all types of engineers and accounting for different parts of the problem. It's one of the ones I'm most happy with because ultimately, I came up with a very simple solution that was easy to understand from all levels, from the executives to the engineers and developers. Ultimately, it was just as efficient as something really complex, and they could have spent a lot more time on.

Back in the university, we have a problem that we wanted to predict algae blooms. This algae blooms could cause a rise in toxicity of the water and it could cause problems through the water treatment company. We couldn't like predict with our chemical engineering background. So, we use artificial neural networks to predict when these blooms will occur. So, the water treatment companies could better handle this problem.

In Toronto, the public transit is operated by Toronto Transit Commission. We call them TTC. It's one of the largest transit authorities in the region, in North America. And one day they contacted me and said, "We have a problem." And I said, "Okay, what's the problem?" They said, "Well, we have complaints data, and we would like to analyze it, and we need your help." I said, "Fine I would be very happy to help." So, I said, "How many complaints do you have?" They said, "A few." I said, "How many?" Maybe half a million. I said, "Well, let's start working with it."

So, I got the data and I started analyzing it. So, basically, they have done a great job of keeping some data in tabular format that was unstructured data. And in that case, tabular data was when the complaint arrived, who received it, what was the type of the complaint, was it resolved, whose fault was it. And the unstructured part of it was the exchange of e-mails and faxes.

So, imagine looking at how half a million exchanges of e-mails and trying to get some answers from it. So, I started working with it. The first thing I wanted to know is why would people complain and is there a pattern or is there some days when there are more complaints than others? And I had looked at the data and I analyzed it in all different formats, and I couldn't find the impetus for complaints being higher on a certain day and lower on others.

And it continued for maybe a month or so. And then, one day I was getting off the bus in Toronto, and I was still thinking about it. And I stepped out without looking on the ground, and I stepped into a puddle, puddle of water. And now, I was sort of ankle deep into water, and it was just one foot wet and the other dry. And I was extremely annoyed. And I was walking back and then it hit me, and I said, "Well, wait a second. Today it rained unexpectedly, and I wasn't prepared for it. That's why I'm wet, and I wasn't looking forward." What if there was a relationship between extreme weather and the type of complaints TTC receives?

So, I went to the environment Canada's website, and I got data on rain and precipitation, wind and the light. And there, I found something very interesting. The 10 most excessive days for complaints. The 10 days where people complain the most were the days when the weather was bad. It was unexpected rain, an extreme drop in temperature, too much snow, very windy day.

So, I went back to the TTC's executives and I said, "I've got good news and bad news." And the good news is, I know why people would complain excessively on certain days. I know the reason for it. The bad news is, there's nothing you can do about it.

Old problems, new problems, Data Science solutions

Organizations can leverage the almost unlimited amount of data now available to them in a growing number of ways. However, all organizations ultimately use data science for the same reason—to discover optimum solutions to existing problems. Let's take a look at three examples of data science providing innovative solutions for old problems.

In transport, Uber collects real-time user data to discover how many drivers are available, if more are needed, and if they should allow a surge charge to attract more drivers. Uber uses data to put the right number of drivers in the right place, at the right time, for a cost the rider is willing to pay.

In a different transport related data science effort, the Toronto Transportation Commission has made great strides in solving an old problem with traffic flows, restructuring those flows in and around the city. Using data science tools and analysis, they have: Gathered data to better understand streetcar operations, and identify areas for interventions Analyzed customer complaints data Used probe data to better understand traffic performance on main routes Created a team to better capitalize on big data for both planning, operations and evaluation By focusing on peak hour clearances and identifying the most congested routes, monthly hours lost for commuters due to traffic congestion dropped from 4.75 hrs. in 2010 to 3 hrs. in mid-2014.

In facing issues in our environment, data science can also play a proactive role. Freshwater lakes supply a variety of human and ecological needs, such as providing drinking water and producing food. But lakes across the world are threatened by increasing incidences of harmful cyanobacterial blooms. There are many projects and studies to solve this long-existing dilemma.

In the US, a team of scientists from research centers stretching from Maine to South Carolina is developing and deploying high-tech tools to explore cyanobacteria in lakes across the east coast.

The team is using robotic boats, buoys, and camera-equipped drones to measure physical, chemical, and biological data in lakes where cyanobacteria are detected, collecting large volumes of data related to the lakes and the development of the harmful blooms. The project is also building new algorithmic models to assess the findings. The information collected will lead to better predictions of when and where cyanobacterial blooms take place, enabling proactive approaches to protect public health in recreational lakes and in those that supply drinking water.

Such interdisciplinary training prepares the next generation of scientists to address societal issues with the proper modernized data science tools. It takes gathering a lot of data, cleaning and preparing it, and then analyzing it to gain the insight needed to develop better solutions for today's enterprises.

How do you get a better solution that is efficient?

You must: Identify the problem and establish a clear understanding of it. Gather the data for analysis. Identify the right tools to use. Develop a data strategy. Case studies are also helpful in customizing a potential solution.

Once these conditions exist and available data is extracted, you can develop a machine learning model. It will take time for an organization to refine best practices for data strategy using data science, but the benefits are worth 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 a 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 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.

Let me explain regression in the simplest possible terms. If you have ever taken a cab ride a taxi ride you understand regression. Here's how it works. The moment you sit in a cab ride in a cab you see that there's a fixed amount there it's is $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 it was 2.50 and it would compute the relationship between the fare 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 take your data, take your information, and put it in the Cloud, put it in the 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.

So, 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 data sets and you're able to do it even though your own systems, your own machines, your own computing environment would not allow you to do so. So, Cloud is beautiful.

And the other thing Cloud is beautiful for is that it allows multiple entities to work with same data at the same time. So, 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. So, Cloud is beautiful.

At the Big Data University which is an IBM initiative, we have these courses people can take and learn about data science. But at the same time, we provide these Cloud-based environments for not only analytics but also for working with big and small data. So, one of the products that is integrated with Big Data University is Data Scientist Workbench.

Data Scientist Workbench is an internet-based solution. You log in, and the moment you log in, you now have access to some very advanced computing environments. As simple as R in RStudio, and data and algorithms to define the data set using OpenRefine, but also the ability to work with very large data sets using technologies like Spark.

So, the advantage of working with Data Scientist Workbench is not only that you have the ability to work with these advanced algorithms into computing platforms, but you also have the ability to work with very large data set, because Spark is integrated and it's all in the Cloud. You don't have to maintain it. You don't have to download it. You don't have to worry about updating it. All is being done for you in the Cloud by the Data Scientist Workbench.

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 *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 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 much different definitions. In September 2015, a co-panelist at a meet up organized by BigDataUniyersity.com in Toronto confined data science to machine learning. There you have it. If you are not using the black boxes that make up 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 50-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 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."

In this lesson, you have learned:

- The typical work day for a Data Scientist varies depending on what type of project they are working on.
- Many algorithms are used to bring out insights from data.
- Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.

Big Data and Data Mining 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, while relational databases and 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 era. 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 are 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 want to 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 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 breaking 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.

How does Data Science differ from traditional subject like statistics?

Most of the components of data science have been around for 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 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 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.

Do you recall a time when no one spoke about 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.

What did you think when you first heard the term 'Data Science'?

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.

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.

What did you study in your undergrad?

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 going to learn two things while I was there. One, I was going to learn how to program. I had taken one programming course in college. And I was going to 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. 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).

What does your typical Monday look like?

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 home works. 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.

Do you use Jupyter Notebooks?

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 Scientists at New York University

Everybody knows how to program, at least a little bit. They all have a little bit of programming background at least, and some of them have a lot. Some of them are Masters of Science and Computer Science, some of them are MBA students who've come in from technical fields and programmed every day. And others are ones who maybe took a programming course in college four or five years ago but at least they can think computationally, which I think is the most important thing that they need.

Are Data Science skills becoming more important in the work place?

Data science and business analytics have become very hot subjects in the last four or five years. We have new tools, we have new approaches, and we have lots and lots of data that traditional techniques just couldn't really store and handle. I think the word is out. I think at this point, at first, companies and employers understood the need, especially in certain fields. I can remember talking to a major bank three years ago about big data and there was one little group in the bank where one person had a little effort in putting a little cluster together. Now that same bank has five or six major big data clusters and they're putting all of their credit card data in it and they're grinding it upside down and sideways, using all sorts of data science kinds of techniques.

Two years ago, or was it last year, I think, our undergraduate dealing with data course had 28 students in it. This year it has 140. So that means that the parents are now beginning to get the word, because one thing we understand with our undergrads is the parents who are paying very hefty tuitions, they, you know, they tell their sons and daughters, "You know, you should be an accountant," right? Or, "You should go into financial services, "or into marketing, because this is where the money is." Now, they're getting the word that maybe you should take some more STEM classes in high school and be ready to go into data science or go into fields where analytics has become more and more important.

What is Big Data?

It depends on who you are (laughs). I have my own definition of big data. My definition of big data is data that is large enough and has enough volume and velocity that you cannot handle it with traditional database systems.

Some of our statisticians think big data is something you can't fit on a thumb drive. Big data, to me, was started by Google. When Google tried to figure out how they were, when Larry Page and Sergey Brin wanted to, basically, figure out how to solve their page rank algorithm, there was nothing out there. They were trying to store all of the web pages in the world, and there was no technology, there was no way to do this, and so they went out and developed this approach, which has now become, Hadoop has copied it, but this is where all these large, big data clusters are found. But big data has now also expanded into, how do you analyze? There are new analytical techniques and statistical techniques for handling these really, really, really large data sets. We'll probably get to deep learning at some point along here.

Establishing Data Mining Goals

The first step in data mining requires you to set up goals for the exercise. Obviously, you must identify the key questions that need to be answered. However, going beyond identifying the key questions are the concerns about the costs and benefits of the exercise. Furthermore, you must determine, in advance, the expected level of accuracy and usefulness of the results obtained from data mining. If money were no object, you could throw as many funds as necessary to get the answers required. However, the cost benefit trade-off is always instrumental in determining the goals and scope of the data mining exercise. The level of the accuracy expected from the results also influences the costs. High levels of accuracy from data mining would cost more and vice versa. Furthermore, beyond a certain level of accuracy, you do not gain much from the exercise, given the diminishing returns. Thus, 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. At times, data are readily available for further processing. For instance, retailers often possess large databases of customer purchases and demographics. 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

Preprocessing data is an important step in data mining. 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. For instance, human error might lead to inadvertent merging or incorrect parsing of information between columns. 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. For instance, a particular subset of individuals in a large data set may have refused to disclose their income. Findings relying on an individual's income as an input would exclude details of those individuals whose income was not reported. This would lead to systematic biases in the analysis. 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. For instance, an individual's income may be recorded in the data set as wage income; income from other sources, such as rental properties; support payments from the government, and the like. Aggregating income from all sources will develop a representative indicator for the individual income.

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.

Later sections in this chapter detail data mining algorithms and methods.

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.

In this lesson, you have learned:

- How Big Data is defined by the Vs: Velocity, Volume, Variety, Veracity, and Value.
- How Hadoop and other tools, combined with distributed computing power, are used to handle the demands of Big Data.
- What skills are required to analyze Big Data.
- About the process of Data Mining, and how it produces results.

Deep Learning and Machine Learning

What's the difference?

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 is 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. It is what enables AI systems to continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct.

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 layer-deep and are 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.

Now that you have a broad understanding of the differences between some key AI concepts, there is one more differentiation that is important to understand that between Artificial Intelligence and Data Science.

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.

Neural Networks and Deep Learning

It's, I guess, Computer Sciences attempt to mimic real, the neurons, in how our brain actually functions. So, 20-23 years ago, a neural network would have some inputs that would come in. They would be fed into different processing nodes that would then do some transformation on them and aggregate them or something, and then maybe go to another level of nodes. And finally, there would some output would come out, and I can remember training a neural network to recognize digits, handwritten digits and stuff.

How does a Neural Network work?

So, a neural network is trying to use computer, a computer program that will mimic how neurons, how our brains use neurons to process thing, neurons and synapses and building these complex networks that can be trained. So this neural network starts out with some inputs and some outputs, and you keep feeding these inputs in to try to see

what kinds of transformations will get to these outputs? And you keep doing this over, and over, and over again in a way that this network should converge. So, these inputs, the transformations will eventually get these outputs. Problem with neural networks was that even though the theory was there and they did work on small problems like recognizing handwritten digits and things like that. They were computationally very intensive and so they went on a favor and I stopped teaching them probably 15 years ago.

And then all of a sudden, we started hearing about deep learning, heard the term deep learning. This is another term, when did you first hear it? Four years ago, five years ago? And so, I finally said, what the hell is deep learning? It's really doing all this great stuff, what is it? And I Google, I was like, this is neural networks on steroids. What they did was they just had multiple layers of neural networks, and they use lots, and lots, and lots of computing power to solve them.

Just before this interview, I had a young faculty member in the marketing department whose research is partially based on deep learning. And so, she needs a computer that has a Graphics Processing Unit in it, because it takes enormous amount of matrix and linear algebra calculations to actually do all of the mathematics that you need in neural networks.

But they've been they are now quite capable. We now have neural networks and deep learning that can recognize speech, can recognize people, you got there, getting your face recognized. I guarantee that NSA has a lot of work going on in neural networks. The university right now, as director of research computing, I have some small set of machines down at our south data center, and I went in there last week and there were just piles, and piles, and piles of cardboard boxes all from Dell with a GPU on the side.

Well, the GPU is a Graphics Processing Unit. There's only one application in this University that needs two hundred servers each with Graphics Processing Units in it, and each Graphics Processing Unit, it has like the equivalent of 600 cores of processing. So, this is tens of thousands of processing cores that is for deep learning, I guarantee.

What are some of the Use Cases of Deep Learning?

Some of the first ones are speech recognition, who teaches the deep learning class at NYU, and is also the head data scientist at Facebook comes into class with a notebook, and it's a pretty thick notebook. It looks a little odd, because it's like this and it's that thick because it has a couple of Graphics Processing Units in it, and then he will ask the class to start to speak to this thing. And it will train while he's in class, he will train a neural network to recognize speech. So, recognizing speech, recognizing people, images, classifying images, almost all of the traditional tasks that neural nets used to work on in little tiny things. Now, they can do really, really, really large things. It will learn on its own, the difference between a cat and a dog, and different kinds of objects, it doesn't have to be taught. It doesn't, it just learns that's why they call it deep learning, and if you hear, he plays this, if you hear how it recognizes speech and generate speech. It sounds like a baby who learning to talk. You can just, you're like really do about, all of a sudden, this stupid machine is talking to you and learned how to talk. That's cool.

How can one get started with neural networks?

I need to learn some linear algebra, a lot of this a lot of this stuff is based on matrix and linear algebra. So, you need to know how to do use linear algebra do transformations. Now, on the other hand, there's now lots of packages out there that will do deep learning and they'll do all the linear algebra for you, but you should have some idea of what is happening underneath. Deep learning, particularly needs really high-powered computational power. So, it's not something that you're going to go out and do on your notebook for it. You could play with it. But if you really want to do it, seriously, you have to have some special computational resources.

Applications of Machine Learning

Everybody now deals with machine learning. Drives my wife crazy sometimes because recommender systems, right, this is one of the first applications. Netflix, my wife gets really upset because if I go in and sign in as her and watch a movie all of a sudden, she's watching action movies and not watching some of the deep philosophical movies she likes to watch. 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.

And I'll give a plug for Foster Provost's book where he's actually written a whole book basically telling practitioners how to use all of these new machine learning techniques.

I have two sprinkler systems running right now, one in Maine and one in New Jersey, that talk to me and tell me what's happening and I can turn them on and off and program them and whatever. Yes, we already have refrigerators that will tell you, scan things that tell you what's in them and stuff. Kettles, there's a little problem with heat there, I'm not so sure about kettles but many, many, many devices now furnaces, I mean any kind of device you put in the home is going to generate data. I have my thermostat, I mean I can tell you what the temperature is in my house right now if I wanted to, I can turn it up and down, I can turn lights on and off and I'm just, my wife won't let me go very far with this because it drives her totally crazy. Totally crazy.

I could right now connect to the camera in my living room and talk to my wife but if I did, I would probably have the door locked when I get home. But there are people out there that have their total lives totally generating data about them. My Fitbit, which I haven't had very long, this is collecting lots of data about me. And it could collect a lot more that I can feed to my doctor.

One of the cool things that's going on now is these peer-to-peer networks, what we call ZigBee, these ZigBee networks that are fairly very low frequency but a ZigBee device can run on a battery for a couple of years, they're only this big, and they can talk to each other so they can form a mesh network. Just throw them around the building and they'll talk to each other and one of them will connect to the internet and push its data out.

Regression

Why Tall Parents Don't Have Even Taller Children?

You might have noticed that taller parents often have tall children who are not necessarily taller than their parents and that's a good thing. 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 couple of millennia, human beings would become uncomfortably tall for their own good, requiring even bigger furniture, cars, and planes.

Sir Frances Galton in 1886 studied the same question and landed upon a statistical technique we today know as *regression models.* This chapter explores 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.

I would like to introduce regression models by narrating a story about my Master's thesis. I believe that this story can help explain the utility of regression models.

The Department of Obvious Conclusions

In 1999, I finished my Masters' research on developing hedonic price models for residential real estate properties. It took me three years to complete the project involving 500,000 real estate transactions. As I was getting ready for the defense, my wife generously offered to drive me to the university. While we were on our way, she asked, "Tell me, what have you found in your research?" I was delighted to be finally asked to explain what I have been up to for the past three years. "Well, I have been studying the determinants of housing prices. I have found that larger homes sell for more than smaller homes," I told my wife with a triumphant look on my face as I held the draft of the thesis in my hands.

We were approaching the on-ramp for a highway. As soon as I finished the sentence, my wife suddenly turned the car to the shoulder, and applied brakes. As the car stopped, she turned to me and said: "I can't believe that they are giving you a Master's degree for finding just that. I could have told you that larger homes sell for more than smaller homes."

At that very moment, I felt like a professor who taught at the department of obvious conclusions. How can I blame her for being shocked that what is commonly known about housing prices will earn me a Master's degree from a university of high repute?

I requested my wife to resume driving so that I could take the next ten minutes to explain her the intricacies of my research. She gave me five minutes instead, thinking this may not require even that. I settled for five and spent the next minute collecting my thoughts. I explained to her that 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.

As I explained my contributions to the study of housing markets, I noticed that my wife was mildly impressed. The likely reason for her lukewarm reception was that my findings confirmed what we already knew from our everyday experience. However, the real value added by the research rested in quantifying the magnitude of those relationships.

Why Regress?

A whole host of questions could be put to regression analysis. Some examples of questions that regression (hedonic) models could address include:

- How much more can a house sell for an additional bedroom?
- What is the impact of lot size on housing price?
- Do homes with brick exterior sell for less than homes with stone exterior?
- How much does a finished basement contribute to the price of a housing unit?
- Do houses located near high-voltage power lines sell for more or less than the rest?

In this lesson, you have learned:

- The differences between some common Data Science terms, including Deep Learning and Machine Learning.
- Deep Learning is a type of Machine Learning that simulates human decision-making using neural networks.
- Machine Learning has many applications, from recommender systems that provide relevant choices for customers on commercial websites, to detailed analysis of financial markets.
- How to use regression to analyze data.

Data Science in Business Week-3

How Data Science is saving lives

Using Data Science techniques to understand and analyze the large data sets available today has a huge impact on human lives. It can provide 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 lifesaving 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 localized 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 analyze 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.

How Should Companies Get Started in Data Science?

At the end of the day, for businesses, they know one thing, that if they are unable to measure something, they are unable to improve it. And if they are unable to measure their costs, they are unable to reduce them. If they're unable to measure their profits, they are unable to increase them. So, the first thing a company has to do is to start recording information, start capturing data, data about costs. And the differentiate it by labor costs and material cost, the cost to how much it cost to sell one product and the total cost.

And then you look at the revenue, where's your revenue coming from?

Is 80% of your revenue coming from 20% of your customers?

Or is it the other way around?

So, first thing first, start capturing data. Once you have data, then you can apply algorithms and analytics to it. So, the first thing to do would be to capture data. If you're not capturing it, start capturing it. If you're capturing it, archive it. Do not overwrite on your old data thinking you don't need it anymore. Data never gets old. Data is always relevant, even if it's 100 years old, 200 years old. It is relevant to you and your firm and your success. So, keep data, capture it, archive it, make sure nothing goes to waste. Make sure there's a consistency. So, someone 20 years later trying to understand that data should be able to do so, so have proper documentation. Do it now.

Put the best practices for data archiving in place the moment you start a business. And if you're already in business and you haven't done it, do it now.

>> Start measuring things.

Too many companies haven't measured things properly for a decade and, then they decide they want data science. Data science inside a company is only going to be as valuable as the data collected. Garbage in, garbage out is a rule in any sort of analysis.

>> If something is not measured, it's very difficult to improve it or to change it. So, the very first step is measurement. If companies have existing data, then they should start looking at it and cleaning it. If they don't have existing data, then they need to start collecting it.

>> I think to look for a team who love to work as a data scientist.

>> The first stop is to have employees that they are interested on data science. because if you don't have interest in your company, you will not have engagement.

>> Companies should remember that it's key to have a team. So, it's not one data scientist, but a team of them, that each of them have strengths in different areas of data science.

Applications of Data Science

- I think one of the good new applications of data science is in the medical field. Like in drug delivery or cancer treatment.

- I think a very interesting one is how now companies can use all the information they're gathering from their customers to actually develop new products that respond to the needs of the customers.

- A good new application of data science was the high trending news of Pokémon Go. So, they used Ingress. They used data of the Ingress app. The last app of the same company and they choose the locations for Pokémons and gyms according to data from the last app. So, they learned with their errors.

- Google Search is an application of data science. The Google Search, whenever we want to search anything. So, I think it's all because of data science. Whatever Google is now, it's all because of data science.

- Augmented reality is my favorite new implementation of data science. I think you can't look at a new technology and not see data science in there but augmented reality is the one I'm just the most excited about. The ability to walk around and see things on walls or around us that aren't really there. Pokémon's just the start.

- So, what has happened is that now the tools are available and datasets are available, people are applying them with not much diligence and I think one of the strange cases which got reported in the newspapers is about the story of a father walking into a Target store in the US and complaining about the fact that the Target was sending mails to his teenage daughter about diapers and milk, baby formula. He was angry with them. He said, "Why would you like for my teenage daughter to have a baby?" And he was obviously disturbed by this mail or the ad campaign. And they obviously apologized but then the father returned two weeks later and he apologized to them saying he didn't know his daughter was pregnant.

Now the question is, how did Target know this thing before the father knew. And what has happened is that they would look at the purchasing behavior of individuals. So, if you're buying some sort of supplements or vitamins then you know that this is the first trimester of pregnancy. So, they know what products to send to you assuming that the person who bought those supplements were pregnant.

Now this is a great story about data science and how data science can forecast and predict these consumer behaviors even before the family would find out. And I find it disturbing and strange and odd for a variety of reasons.

First of all, for every correct prediction, you have hundreds of incorrect predictions which we call the false positives and no data scientist actually advertises his or her false positives. We only advertise and promote what we got it right. But when we got it wrong hundreds of times, we don't tell it.

Second thing is, that's an abuse of data. That's basically not really not giving you much insight. You've just found a correlation but someone could be purchasing the same material for someone else. So, and then the odds of getting it wrong and the odds of getting false positives is much higher.

So, I find it strange and I think it gives a false sense of our ability to predict the future. The reality is about data science and the most important thing for the budding data scientist to know that all forecasts are wrong. They're useful but they're wrong. And so one should not put their faith into the fact that now that we can do predictive analytics that we can solve all problems.

I think a good example is the Google Search. Google published a paper saying they can predict flu epidemics before the Center for Disease Control. And what they did was they were looking at what people were searching on Google so flu symptoms. So, Google saw the flu symptom searches before anybody else and they were able to predict it.

The thing is these searches are good and they are correlated with some outcomes but not necessarily all the time. So, at that time, when Google announced, it was a big thing and everybody really like it and well that's a new era of predictive analytics.

Only that a few years later they realized that Google started to predict false positives. That they were predicting things that were not really there or the predictions were not that accurate for a variety of reasons. They changed probably their algorithms and the datasets were not really correlated with the outcomes.

So, what's the lesson to learn here?

One has to avoid what we call the data hubris. That you should not believe in your models too much because they can lead you astray.

Data science has tremendous potential to bring change in parts of the world, in parts of our society that have been disenfranchised for years.

One sees great examples of data science especially in the developing countries where they are targeting relief efforts. They're targeting food and other aid to individuals, to places that have not been targeted in the past. And the reason it is happening now is the greater availability of data and models and analytics to be able to pinpoint where the greatest needs are.

The ability to design and conduct experiments to see if one were to give micro-credits, small loans to very poor households in developing parts of the world, to see how they affect the individual household's ability to get out a poverty and also the local community's ability to collectively improve their economic well-being by just very small infusions of cash or credit.

So, these experiments happening all over the world are allowing that is a direct result of our ability to analyze data and be able to design experiments and then roll out humongous efforts in providing relief, providing credit, providing an opportunity to those who have been disenfranchised in the past an opportunity to join the rest of the world in prosperity and happiness and health.

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, 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. It builds a powerful narrative in support of the thesis statement that the U.S. economy is doing much better than what most would like to believe. 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.

Now consider if we were to exclude the narrative from this report and presented the findings as a deck of PowerPoint slides with eight graphics and four tables. The PowerPoint slides would have failed to communicate the message that the authors carefully crafted in the report citing Piketty and Voltaire. I consider the 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.

Now let us work backward from the Deloitte report. Before the authors started their analysis, they must have discussed the scope of the final deliverable. They would have deliberated the key message of the report and then looked for the data and analytics they needed to make their case. 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.

In this lesson, you have learned:

- Data Science helps physicians provide the best treatment for their patients, and helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
- That companies can start on their data science journey by capturing data. Once they have data, they can begin analyzing it.
- Some ways that data is generated by consumers.
- How businesses like Netflix, Amazon, UPs, Google, and Apple use the data generated by their consumers and employees.
- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.

Careers and Recruiting in Data Science

How Can Someone Become a Data Scientist?

A real data scientist, the high-end data scientists, are mostly PhDs. They often come out of physics, out of statistics, they have to have a computer science background, they have to have a math background, they have to know about databases and statistics and probability and all that stuff.

However, if you're coming into a data science team, I think the first skills you need is you need to know how to program, at least have some computational thinking, so having taken a programing course, you need to know some algebra, at least up to analytics, geometry, and hopefully some calculus, some basic probability, some basic statistics, I mean really have to understand the difference and different statistical distributions, and database.

I mean, one of the easiest places to start is relational databases, which stores lots and lots of our data so people can first walk before they can run by at least understanding about computers and databases and how we store things and if you understand relational databases nowadays you can still, just with that understanding, use big data clusters as if they were just a big relational database.

You don't have to really have understand the whole MapReduce programming model. But then, as you go further up in the field, then you have to know a lot of computer science theory and statistics, it's really, and probability, it's really the intersection of them that the high-end data scientists, the PhD data scientists work with.

I do a lot of self-learning. I think everybody these days, I mean, I learned about Hadoop all by myself, I read some articles, I watched some videos, I thought, I played, although I'm a builder, I'm a tinkerer, so if I wanna figure out how to do something, I build it. I mean, my first HPC cluster I heard about this term a Beowulf cluster, I mean, yeah, what the hell's that? So, I looked it up and said, oh, it's just a bunch of computers hooked together with a TCP/IP network, that's pretty easy, so we get a grant from Citi Bank and we built a five-thing cluster and I said, oh, well, that's HPC. I said, I had one of the first HPC clusters at the university, it was tiny but a lot of our researchers loved it because they could run stuff 40 and 50 times faster.

So, I think one of the ways you learn things is you do them, you have to do them, and these online learning platforms especially now that we have things like IPython and Jupyter Notebooks and I guess Zeppelin means that you can actually go in and take some of these courses and you can do things right then and you can see them and feel them and play with them and, at that point, you know, you'll start to get your head around what is actually happening.

Motivation is the key problem in all of these, is how to keep people motivated and I think the badge system that the, what was it, Big Data University has, is one of the ways is how do you get people to keep going through. But if they want to, they can. It's up to the individual to. So, they have to understand what the goal is.

Where should Data Science fit in the Org Structure?

The place it can't sit is probably under the CIO, the Chief Information Officer. CIOs current chief information officers in many companies got there from an accounting background or a finance background, they're clueless. Sorry.

But they really, it has to come out of the research side. So, you'll find data scientists primarily in companies that have some research agenda, pharmaceuticals, finance, all of, any technology company.

If you look at, we can't keep some of our PhD data scientists in our program, they are now at Facebook, they're at Linkedin, they're at Uber, they're at Lyft, because the demand out there for the PhD level data scientist is just unbelievable. They make large amounts of money and they're playing with problems that are really, really neat. How do you schedule the Uber cars? You have enormous amounts of data.

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.

Careers in Data Science

The emergence of Internet of things and advances in distributed computing have brought vast amounts of data and the technological capability to analyze it. Now that we can extract useful insights and new knowledge, we need to know how to shape that data to focus on what to do with it and what it can do for us.

Enter data science.

Companies like LinkedIn, Glassdoor, Indeed, and Dice track employment trends which show a career in data science moving up the list of most promising jobs to become number one since 2016. It remains one of the top three career choices for 2020. Dice noted that job postings are from companies in a wide variety of industries, not just tech.

Global Industry Analysts Incorporated predicts that the data science platform market will grow by $314.8 billion US by 2025, driven by a compounded growth of 38.2%. McKinsey Global Institute warned of huge talent shortages for data and analytics by 2018.

Forrester Research Analyst Brandon Purcell said, in January of 2019, the demand for data scientists will only grow as organizations increasingly rely on data-driven insights. We're now well into that period, and recruiters are finding it difficult to fill the growing need for talented data scientists.

What motivates someone going into a data science?

For one thing, data science applies to almost any discipline. So, if you have the aptitude and desire to work with data, enjoy coding, have no problem learning math and statistics, and you are a good storyteller, then you can certainly enter a data science field and excel.

For most people, this means acquiring additional tools and skills and continuously learning about new tools and techniques in the field. The women in data science initiative spearheaded by the Stanford Institute for Computational and Mathematical Computing have committed to inspire and educate data scientists worldwide, regardless of gender and to support women in the field.

When you are seeking a career in data science, you need to make sure your skill set matches the role you are targeting. You can tailor your skill set to the specific area you want to enter, adding missing skills via one of the many excellent online training resources. Then you'll be prepared for a fascinating and rewarding career. So now it is time to move into this field when there are such diverse choices available and education resources that make it a reality.

High School Students and Data Science Careers

What would you say to High School Students about Data Science Careers?

- Learn how to program. Learn some math. Take a course in probability. Learn a little bit of statistics. And then, play. Build something, write something. I mean, when I say build, programming and building systems, building things isn't just physical, right?

You can build computer systems, statistical systems, whatever. But once you try to do something, then you'll know what tools you need, right? And you'll say, "Oh, oh my god, what? " There's this expression there, "what does an inner product mean?" What's that? "How do I, oh, okay, I can learn that." And then when they get to college, they will have a big jump on many of the other college students. And so, when they get out of college, they'll have an even bigger jump, and then make a lot of money. And they'll be happy, too. This stuff is fun, right? It's fun.

What would you say to parents about Data Science Careers?

How about we talk about women in data science, right? I mean, this is the, I happen to have a granddaughter, and I keep trying to steer her gently towards STEM kinds of things.

I think parents need to understand that their children are going into a new world, that this is a world that's gonna be, it's different than the world they're living in. And their children have to be prepared for it, the jobs, you can just look, I mean just in yesterday's paper, this morning's paper, where the jobs are.

Manufacturing down, service is going up, but if you know some data science and some basic STEM kinds of things, I mean, besides data science, the rest of the sciences are also taking off, often based on computation. The ability to simulate things that, we don't have to build anything anymore, right?

I have a 3D printer. I don't have to print something to know what it's gonna look like, right? I can design something, or I can get a design off the internet, flip it around, look at it and say hmm, that sorta does what I want but let me pull this a little bit this way, and then I'll send it to my 3D printer.

And if I don't, I have a really junky 3D printer, but, if I say oh, I want to make a really good version of that, I can push a button and send it out over the internet and, you know, for 15, 20, 30, 40, 50 dollars have a really nice thing printed for me, and today, they can print things in wood, rubber, metal, it's incredible.

General Electric is, Jeff Immelt is betting part of the future of General Electric on their new platform, which is one they built for designing and building jet engines, but they're now rolling it out across and selling it to other manufacturing companies.

Manufacturing is becoming digital. And the good news is that it's more productive, you can do things faster, the bad news is it doesn't need anywhere near as many workers. So, you've gotta be on the right side of this trend.

In this lesson, you have learned:

- Data Scientists need programming, mathematics, and database skills, many of which can be gained through self-learning.
- Companies recruiting for a Data Science team need to understand the variety of different roles Data Scientists can play, and look for soft skills like storytelling and relationship building as well as technical skills.
- High school students considering a career in Data Science should learn programming, math, databases, and, most importantly practice their skills.

Report Structure

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, acknowledgements, 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, 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 the adequate context before you embark on your analysis. You might use 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 5, "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 *acknowledgement* section (acknowledging the support of those who have enabled your work is always good), and *appendices,* if needed.

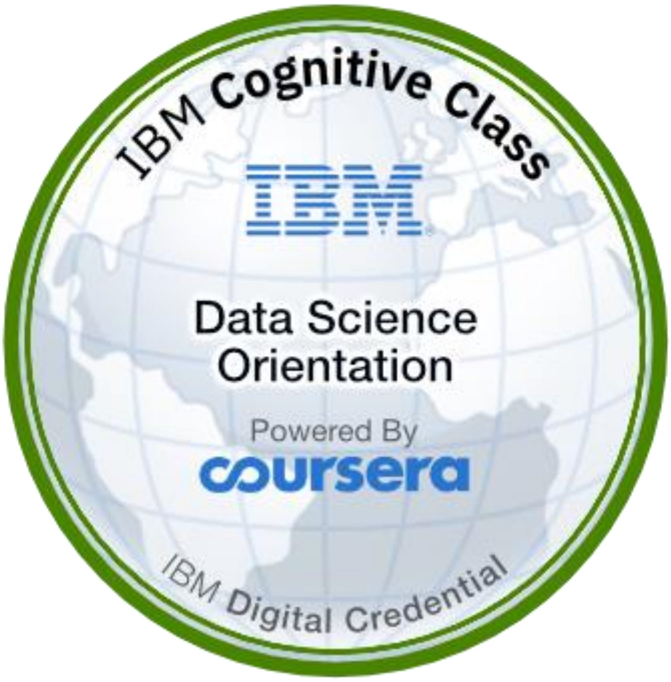
Have You Done Your Job as a Writer?

As a data scientist, you are expected to do a 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 manuscript to the journal. I believe the checklist is useful for budding data scientists and, therefore, I have reproduced it verbatim for their benefit.

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

In this lesson, you have learned:

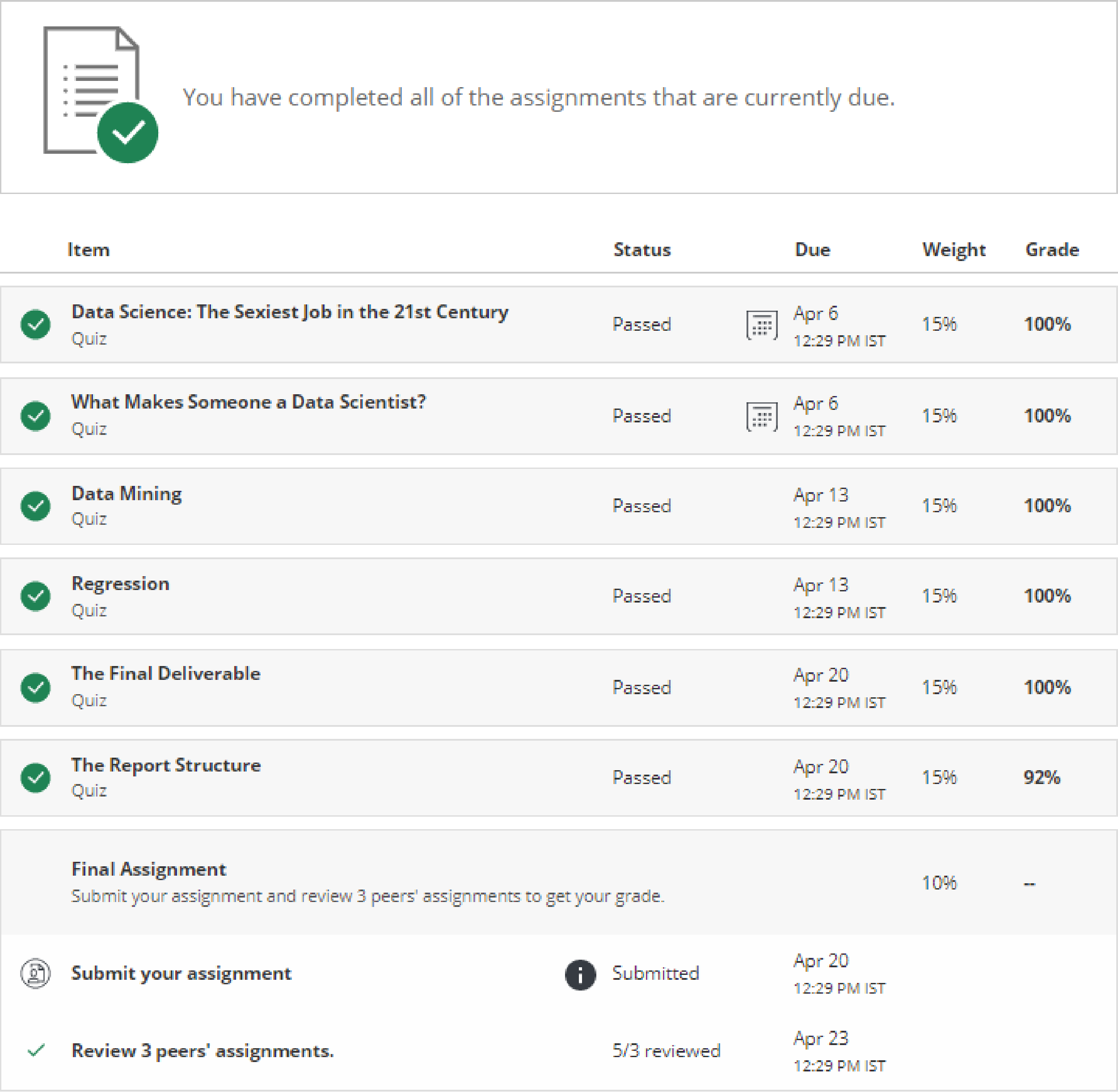
- The length and content of the final report will vary depending on the needs of the project.
- The structure of the final report for a Data Science project should include a cover page, table of contents, executive summary, detailed contents, acknowledgements, references and appendices.
- The report should present a thorough analysis of the data and communicate the project findings.



Earning the Coursera certificate for this course will entitle you to receive an IBM digital badge without any additional charge. (Note: payment is required to have full access to the course and to be eligible to qualify for the course certificate.) For this course you will earn Data Science Orientation Badge. Full details can be seen here.

https://www.youracclaim.com/org/ibm/badge/data-science-orientation

IBM digital badges are an on-line credential that validate the skills you acquired passing this course. You can share IBM digital badges on popular social media sites, such as Linked-In, Twitter or Facebook. Each badge you earn has a unique URL that you can embed in a website, email or CV, so it could not be easier to share your badges and your achievements. IBM digital badges adhere to the global Open Badges Standard managed by the IMS Global Learning Consortium, so you can also share IBM digital badges with any OBS-compliant badge site, such as Mozilla Backpack. IBM has partnered with Credly Acclaim to issue and manage IBM digital badges. If you pass this course and earn the Coursera certificate, you will be provided instructions for how to accept and claim your IBM digital badge.



You have completed all of the assignments that are currently due.

	Item	Status	Due	Weight	Grade
✓	**Data Science: The Sexiest Job in the 21st Century** Quiz	Passed	Apr 6 12:29 PM IST	15%	**100%**
✓	**What Makes Someone a Data Scientist?** Quiz	Passed	Apr 6 12:29 PM IST	15%	**100%**
✓	**Data Mining** Quiz	Passed	Apr 13 12:29 PM IST	15%	**100%**
✓	**Regression** Quiz	Passed	Apr 13 12:29 PM IST	15%	**100%**
✓	**The Final Deliverable** Quiz	Passed	Apr 20 12:29 PM IST	15%	**100%**
✓	**The Report Structure** Quiz	Passed	Apr 20 12:29 PM IST	15%	**92%**
	Final Assignment Submit your assignment and review 3 peers' assignments to get your grade.			10%	--
	Submit your assignment	Submitted	Apr 20 12:29 PM IST		
✓	**Review 3 peers' assignments.**	5/3 reviewed	Apr 23 12:29 PM IST		