**IBM DATA SCIENCE**

**MODULE 1 - WHAT IS DATA SCIENCE?**

**1.1 - DEFINING DATA SCIENCE**

**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 organisations 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 want 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 that 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 masters 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.

**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 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 crowd-sourcing 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, 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.

**ADVICE FOR NEW DATA SCIENTISTS**

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

**LESSON SUMMARY – DEFINING DATA SCIENCE**

Welcome to defining data science, lesson summary. In this video, we'll review important points you learned from the videos and readings in this lesson, we'll also link together the ideas from the resources in this lesson. We will quickly recap what data science is, data scientists role in an organization, what makes a skilled data a scientist, and expert advice on how to acquire these skills? Let's begin, so what really is data science? Simply put, data science is the study of data, data science uses data to understand the world around us, some consider data science an art of uncovering the insights and trends hiding behind data. Data analysis isn't new, recent data access and enhanced computing power drives new insights and knowledge through analysis. We also have the computing power needed to analyze this data to reveal new knowledge, with the power of computers, we can dig through this information to reveal valuable insights. Just like a detective uncovering secrets, data scientists translate the data into stories to generate insights, these insights aid strategic decision-making for companies or institutions. Similar to biological or physical sciences, data science deals with structured and unstructured data. The process of gleaning insights from data includes clarifying the problem, data collection, analysis, pattern recognition, storytelling, and visualization. According to Professor Murtez Aheter from the Ted Rogers School of Management, curiosity, argumentation, and judgment are vital for data scientists. Curiosity helps you explore data and come up with meaningful questions. Good, sound, reasonable arguments help you explain your findings after learning from the data, compelling the listener to adjust their ideas based on the new information. Good judgment guides you to start in the right direction. Skilled data scientists go beyond just being statisticians or computer experts. Companies are looking for versatile individuals who know a lot about a particular subject, have some experience in programming and analyzing data, and can communicate well. Generally, data scientists are comfortable with math, they're curious, they're good at telling stories. Their backgrounds can come from various fields like economics, engineering, medicine, and more. Once you understand your strengths and interests, focus on mastering data analysis in that field and select suitable tools for your industry. As you become proficient, apply your expertise to solve real world issues using data, similar to solving mysteries by deciphering clues. So, what does the future look like for you as a skilled data scientist? Data scientist jobs will also change as technology changes and data roles develop, to ensure their employees are qualified, companies will require certification. Data scientists will always need to think logically, use algorithms, and follow a methodical approach. Most importantly, they must gather data correctly and carefully analyze the models being used, all aiming to achieve successful business results.

**GLOSSARY: DEFINING DATA SCIENCE**

**Uma imagem contendo Aplicativo

Descrição gerada automaticamente**

**1.2 – WHAT DO DATA SCIENTIS 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 accounted 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 reoccur. 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 [what] 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 for it." 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.

**DATA SCIENCE SKILLS & BIG DATA**

I'm Norman White, I'm a Clinical Faculty Member in the IOMS Department, Information, Operations and Management Science Department here at Stern. I've been here for a long time (laughs), since I got out of college, pretty much. I'm sort of a techy, geeky kind of person. I really like to play with technology in my spare time. I'm currently Faculty Director of the Stern Center for Research Computing, in which we have a private cloud that runs lots of different kinds of systems. Many of our faculty or PhD students who need specialized hardware and software will come to us, we'll spin up a machine for them, configure it, I'll help them and advise on them. A lot of the data scientists, or virtually all the data scientists at Stern use our facilities. And their PhD students use them a lot.

I have an undergraduate degree in Applied Physics and while I was an undergrad I took a number of economics courses, so I ended up deciding to go to business school, but I had, this was in the early days of computers (laughs) and I had gotten interested in computers. I came to Stern, which was then NYU Business School downtown and they had a little computer center, and I decided that I was gonna learn two things while I was there. One, I was gonna learn how to program. I had taken one programming course in college. And I was gonna learn how to touch type. I never did learn how to touch type (laughs). Or maybe I did but I've forgotten now, and back to two finger pecking. But I became a self taught programmer, and then I took a number of courses at IBM because I eventually became the director of the computer center, while I was getting my PhD in Economics and Statistics at Stern.

In 1973, the school formed a department called Computer Applications and Information Systems and I was one of the first faculty members in the department and I've been here ever since (laughs).

My typical Monday is, I usually get in around 11 o'clock and I do my email at home first, but I come in and I have two classes on Monday. I have a class on design and development of web based systems at six o'clock. Two o'clock, I have a dealing with data class. The class is based on Python notebooks, so we start with the basics of Unix and Linux, just to get the students used to that. We move onto some Python, some regular expressions, a lot of relational databases, some Python Pandas, which is sort of like R for Python, lets you do mathematical and statistical calculations in Python. And then I end up with big data, for which, as you probably know, I'm an evangelist. The students I have, weekly homeworks. I put them in teams and they have to do a big project at the end of the term, and they do some really cool things. (music) 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.

**UNDERSTANDING DIFFERENT TYPES OF FILE FORMATS**

As a data professional, you will be working with a variety of data file types, and formats. It is important to understand the underlying structure of file formats along with their benefits and limitations. This understanding will support you to make the right decisions on the formats best suited for your data and performance needs. Some of the standard file formats that we will cover in this video include: Delimited text file formats, Microsoft Excel Open XML Spreadsheet, or XLSX Extensible Markup Language, or XML, Portable Document Format, or PDF, JavaScript Object Notation, or JSON, Delimited text files are text files used to store data as text in which each line, or row, has values separated by a delimiter; where a delimiter is a sequence of one or more characters for specifying the boundary between independent entities or values. Any character can be used to separate the values, but most common delimiters are the comma, tab, colon, vertical bar, and space. Comma-separated values (or CSVs) and tab-separated values (or TSVs) are the most commonly used file types in this category. In CSVs, the delimiter is a comma while in TSVs, the delimiter is a tab. When literal commas are present in text data and therefore cannot be used as delimiters, TSVs serve as an alternative to CSV format. Tab stops are infrequent in running text. Each row, or horizontal line, in the text file has a set of values separated by the delimiter, and represents a record. The first row works as a column header, where each column can have a different type of data. For example, a column can be of date type, while another can be a string or integer type data. Delimited files allow field values of any length and are considered a standard format for providing straightforward information schema. They can be processed by almost all existing applications. Delimiters also represent one of various means to specify boundaries in a data stream.

Microsoft Excel Open XML Spreadsheet, or XLSX, is a Microsoft Excel Open XML file format that falls under the spreadsheet file format. It is an XML-based file format created by Microsoft. In an .XLSX, also known as a workbook, there can be multiple worksheets. And each worksheet is organized into rows and columns, at the intersection of which is the cell. Each cell contains data. XLSX uses the open file format, which means it is generally accessible to most other applications. It can use and save all functions available in Excel and is also known to be one of the more secure file formats as it cannot save malicious code.

Extensible Markup Language, or XML, is a markup language with set rules for encoding data. The XML file format is both readable by humans and machines. It is a self-descriptive language designed for sending information over the internet. XML is similar to HTML in some respects, but also has differences. For example, an .XML does not use predefined tags like .HTML does. XML is platform independent and programming language independent and therefore simplifies data sharing between various systems.

Portable Document Format, or PDF, is a file format developed by Adobe to present documents independent of application software, hardware, and operating systems, which means it can be viewed the same way on any device. This format is frequently used in legal and financial documents and can also be used to fill in data such as for forms.

JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web. The file format is a language-independent data format that can be read in any programming language. JSON is easy to use, is compatible with a wide range of browsers, and is considered as one of the best tools for sharing data of any size and type, even audio and video. That is one reason, many APIs and Web Services return data as JSON. In this video, we looked at some popular file and data formats. In the next video, we will learn about the different sources of data.

**DATA SCIENCE TOPICS AND ALGORITHMS**

I really enjoy regression. I'd say regression was maybe one of the first concepts that I, that really helped me understand data so I enjoy regression. I really like data visualization. I think it's a key element for people to get across their message to people that don't understand that well what data science is.

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

Data visualization with R. I love to do this. Nearest neighbor. It's the simplest but it just gets the best results so many more times than some overblown, overworked algorithm that's just as likely to overfit as it is to make a good fit. So structured data is more like tabular data things that you’re familiar with in Microsoft Excel format. You've got rows and columns and that's called structured data. Unstructured data is basically data that is coming from mostly from web where it's not tabular. It is not, it's not in rows and columns. It's text. It's sometimes it's video and audio, so you would have to deploy more sophisticated algorithms to extract data. And in fact, a lot of times we take unstructured data and spend a great deal of time and effort to get some structure out of it and then analyze it. So if you have something which fits nicely into tables and columns and rows, go ahead. That's your structured data. But if you see if it's a weblog or if you're trying to get information out of webpages and you've got a gazillion web pages, that's unstructured data that would require a little bit more effort to get information out of it.

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

**WHAT MAKES SOMEONE A DATA SCIENTIST?**

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

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

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

What is data science?

How does it differ from statistics?

What makes someone a data scientist?

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

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

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

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

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

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

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

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

Dr Vincent Granville (2014), an author on data science, offers certain thresholds to meet to be a data scientist. On pages 8 and 9 in Developing Analytic talent, Dr Granville describes the new data science professor as a non-tenured instructor at a non-traditional university, who publishes research results in online blogs, does not waste time writing grants, works from home, and earns more money than the traditional tenured professors. Suffice it to say that the thriving academic community of data scientists might disagree with Dr Granville.

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

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

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

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

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

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

**LESSON SUMMARY: WHAT DO DATA SCIENTISTS DO?**

Welcome to a realm where data isn't just numbers. It's the gateway to innovation, discovery, and the endless possibilities that lie ahead. To understand the heart and soul of a data scientist world, let's review what data scientists do as discussed in the readings and videos from this lesson.

Data scientist investigate and find explanations for many problems. For example, Dr. Murtaza Haider found an explanation for why half a million customers complained about public transit in Toronto. After much investigation, he found a relationship between unexpected bad weather events and the number of complaints on that particular day.

Data scientist may tackle environmental challenges such as predicting algae blooms to prevent water toxicity. By harnessing data in the prowess of artificial neural networks, data scientist can help water treatment companies safeguard the ecosystem.

Norman White, a clinical faculty member at the Stern School of Business, also discussed the journey that led him to build a recommendation engine that simplified intricate problems across departments.

This underscores the essence of data science, solving real world issues with innovative solutions. Education serves as a cornerstone, equipping future data scientists with essential skills. Dr. White's classroom comes alive with Python notebooks, revealing the secrets of Unix, Linux, relational databases, and powerful tools like Pandas.

Dr. Vincent Granville, an author on data science, lists algebra, calculus, and training, and probability and statistics as necessary educational backgrounds to be a data scientist. He distinguishes between a statistician and a data scientist. A data scientist uses statistics, but is not only a statistician. However, data scientists do use a lot of statistical models, such as statistical regression. Regression shows the probable relationship between two variables, such as the distance you drive and the amount of gas you use.

Data scientists also utilize machine learning algorithms such as nearest neighbor to process what much of the media refers to as big data. The term should be used with caution. What was once considered big data is continually reshaped by innovation. Tools like Hadoop and software advancements have shattered traditional limits, ushering in a new era of possibilities.

Neither does Dr. Patel restrict data scientists to dealing with datasets of arbitrary size, nor does he restrict them to using particular tools. His definition includes individuals of various academic backgrounds and training. A data scientist not only unlocks the insights within a dataset, but conveys a compelling narrative to stakeholders. It's this blend of technical acumen and communication finesse that sets them apart.

The data that data scientists use comes from a wide variety of sources. Sometimes video, sometimes audio, and often unformatted text. Text-based data can also be structured, such as in tables with rows and columns, or unstructured like emails or logs.

The data they work with comes in a wide variety of formats, such as delimited text files, spreadsheets, XML, PDFs, and Javascript Object Notation, or JSON. What makes a data scientist truly exceptional? The answer, according to Rachel Schutt, chief data scientist at News Corp, lies in curiosity.

A successful data scientist is a blend of computer scientist, software engineer, and statistician. Their ability to transform unstructured solutions into structured insights defines their prowess. As we reflect on the insights from a day in the life of a data scientist, we realize that data science isn't just a profession. It's a journey of exploration, innovation, and story telling. The world of data is vast and the data scientist, armed with skills, curiosity and determination, navigates it to unravel the extraordinary.

**GLOSSARY: WHAT DO DATA SCIENTISTS DO?**

**Uma imagem contendo Tabela

Descrição gerada automaticamente**

**Uma imagem contendo Tabela

Descrição gerada automaticamente**

**Interface gráfica do usuário, Texto, Aplicativo, Email, Site

Descrição gerada automaticamente**

**Summary: What Do Data Scientists Do?**

Congratulations! You have completed this lesson. At this point in the course, you know:

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 math, programming, and curiosity about data.

New data scientists need to be curious, judgmental and argumentative.

Knowledgeable data scientists are in high demand. Jobs in data science pays high salaries for skilled workers.

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.

Some key data science related terms you learned in this lesson include: outliers, model, algorithms, JSON, XML. CSV, and regression.

**MODULE 2 – BIG DATA AND DATA MINING**

**2.1 - BIG DATA AND DATA MINING**

**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.

**INTRODUCTION TO CLOUD**

Welcome to Introduction to Cloud Computing and Cloud Deployment and Service Models. After watching this video, you will be able to: Describe cloud computing concepts, define cloud deployment models and cloud service models, and identify the characteristics of cloud computing.

Cloud computing, also referred to as the cloud, is the delivery of on-demand computing resources such as networks, servers, storage, applications, services, and data centers over the Internet on a pay-for-use basis. The term “cloud computing” can be used to describe applications and data that users access over the Internet rather than on their local computer. Examples of cloud computing include users using online web apps, employees using secure online business applications to conduct their work, and users storing personal files on cloud-based storage platforms such as Google Drive, OneDrive, and Dropbox.

One of the main user benefits of cloud computing is that instead of users needing to purchase their own applications and install them locally on their computer, they can use online versions of those applications and pay a monthly subscription. Not only is this typically more cost-effective initially, but users can also access the latest version of the application without having to purchase a full retail copy of the newer version. A side advantage of this is that the user also saves lots of local storage space as the application is hosted online. And, the beauty of most cloud-based applications is that they also enable users to work collaboratively with their colleagues, working on the same files in real time and being able to see each other’s edits and updates.

Cloud computing is composed of five essential characteristics, three deployment models, and three service models. Let’s start with understanding the five essential characteristics of the cloud.

On-demand self-service means that you get access to cloud resources such as the processing power, storage, and network you need, using a simple interface, without requiring human interaction with each service provider. Broad network access means that cloud computing resources can be accessed via the network through standard mechanisms and platforms such as mobile phones, tablets, laptops, and workstations.

Resource pooling is what gives cloud providers economies of scale, which they pass on to their customers, making cloud cost-efficient. Using a multitenant model, computing resources are pooled to serve multiple consumers, and cloud resources are dynamically assigned and reassigned according to demand, without customers needing to know the physical location of these resources.

Rapid elasticity implies that you can access more resources when you need them, and scale back when you don’t, because resources are elastically provisioned and released. And measured service means that you only pay for what you use or reserve as you go. If you’re not using resources, you’re not paying. Resource usage is monitored, measured, and reported transparently based on consumer utilization. As you have seen, cloud computing is really about using technology “as a service,” leveraging remote systems on-demand over the open Internet, scaling up and scaling back, and paying for what you use. And it has changed the way the world consumes compute services, by making them more cost-efficient while also making organizations more agile in response to changes in their markets.

Cloud deployment models indicate where the infrastructure resides, who owns and manages it, and how cloud resources and services are made available to users. There are three types of cloud deployment models: public, private, and hybrid.

Public cloud is when you leverage cloud services over the open internet on hardware owned by the cloud provider, but its usage is shared by other companies.

Private cloud means that the cloud infrastructure is provisioned for exclusive use by a single organization. It could run on-premises or it could be owned, managed, and operated by a service provider.

And when you use a mix of both public and private clouds, working together seamlessly, that is classified as the hybrid cloud model.

Now, let’s look at the three cloud service models that are based on the three layers in a computing stack: infrastructure, platform, and application. These cloud computing models are aptly referred to as Infrastructure as a Service (or IaaS), Platform as a Service (or PaaS), and Software as a Service (or SaaS).

In an IaaS model, you can access the infrastructure and physical computing resources such as servers, networking, storage, and data center space without the need to manage or operate them.

In a PaaS model, you can access the platform that comprises the hardware and software tools that are usually needed to develop and deploy applications to users over the Internet.

And an SaaS is a software licensing and delivery model in which software and applications are centrally hosted and licensed on a subscription basis. It is sometimes referred to as “on-demand software.”

In this video, you learned that:

Cloud computing is the delivery of on-demand computing resources over the Internet on a pay-for-use basis.

Cloud computing is composed of five essential characteristics, three deployment models, and three service models.

The five essential characteristics of cloud computing are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.

There are three types of cloud deployment models: public, private, and hybrid. And the three cloud service models are based on the three layers in a computing stack (infrastructure, platform, and application), and they are referred to as Infrastructure as a Service (or IaaS), Platform as a Service (or PaaS), and Software as a Service (or SaaS).

**CLOUD FOR DATA SCIENCE**

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

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

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

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

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

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

**FOUNDATIONS OF BIG DATA**

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

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

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

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

Volume is the scale of the data, or the increase in the amount of data stored. Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.

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

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

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

Let's look at some examples of the V's in action. **Velocity:** Every 60 seconds, hours of footage 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.

**DATA SCIENCE AND BIG DATA**

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.

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, 'cause 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.

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.

**WHAT IS HADOOP?**

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

Most of the components of data science have been around for many, many, many, many decades. But they're all coming together now with some new nuances I guess. At the bottom of data science you see probability and statistics. You see algebra, linear algebra you see programming and you see databases. They've all been here. But what's happened now is we now have the computational capabilities to apply some new techniques - machine learning. Where now we can take really large data sets and instead of taking a sample and trying to test some hypothesis we can take really, really large data sets and look for patterns. And so back off one level from hypothesis testing to finding patterns that maybe will generate hypotheses. Now this can bother some very traditional statisticians and gets them really annoyed sometimes that you know you're supposed to have a hypothesis that is not that is independent of the data and then you test it. So once some of these machine learning techniques started were really the only thing the only way you can analyze some of these really large social media data sets. So what we've seen is that the combination of traditional [technique] areas computer science probability, statistics, mathematics all coming together in this thing that we call Decision Sciences. Our department at Stern I'll give a little plug here we happen to have been very well situated among business schools because we're one of the few business schools that has a real statistics department with real PhD level statisticians in it. We have an operations management department and an information systems department. So we have a wide range of computer scientists to statisticians, to operations researchers. And so we were like perfectly positioned as a couple of other business schools were to jump on this bandwagon and say; okay this is Decision Sciences. And Foster Provost who's in my department was the first director of the NYU Center for Data Science.

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

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

**Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark**

The Big Data processing technologies provide ways to work with large sets of structured, semi-structured, and unstructured data so that value can be derived from big data. In some of the other videos, we discussed Big Data technologies such as NoSQL databases and Data Lakes.

In this video, we are going to talk about three open source technologies and the role they play in big data analytics — ApacheHadoop, Apache Hive, and Apache Spark.

Hadoop is a collection of tools that provides distributed storage and processing of big data. Hive is a data warehouse for data query and analysis built on top of Hadoop. Spark is a distributed data analytics framework designed to perform complex data analytics in real-time.

Hadoop, a java-based open-source framework, allows distributed storage and processing of large datasets across clusters of computers. In Hadoop distributed system, a node is a single computer, and a collection of nodes forms a cluster.

Hadoop can scale up from a single node to any number of nodes, each offering local storage and computation. Hadoop provides a reliable, scalable, and cost-effective solution for storing data with no format requirements.

Using Hadoop, you can: Incorporate emerging data formats, such as streaming audio, video, social media sentiment, and clickstream data, along with structured, semi-structured, and unstructured data not traditionally used in a data warehouse. Provide real-time, self-service access for all stakeholders. Optimize and streamline costs in your enterprise data warehouse by consolidating data across the organization and moving “cold” data, that is, data that is not in frequent use, to a Hadoop-based system.

One of the four main components of Hadoop is Hadoop Distributed File System, or HDFS, which is a storage system for big data that runs on multiple commodity hardware connected through a network. HDFS provides scalable and reliable big data storage by partitioning files over multiple nodes. It splits large files across multiple computers, allowing parallel access to them. Computations can, therefore, run in parallel on each node where data is stored. It also replicates file blocks on different nodes to prevent data loss, making it fault-tolerant. Let’s understand this through an example.

Consider a file that includes phone numbers for everyone in the United States; the numbers for people with last name starting with A might be stored on server 1, B on server 2, and so on. With Hadoop, pieces of this phonebook would be stored across the cluster. To reconstruct the entire phonebook, your program would need the blocks from every server in the cluster. HDFS also replicates these smaller pieces onto two additional servers by default, ensuring availability when a server fails, In addition to higher availability, this offers multiple benefits. It allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all servers in the cluster for better scalability.

Finally, you gain the benefit of data locality, which is the process of moving the computation closer to the node on which the data resides. This is critical when working with large data sets because it minimizes network congestion and increases throughput. Some of the other benefits that come from using HDFS include: Fast recovery from hardware failures, because HDFS is built to detect faults and automatically recover. Access to streaming data, because HDFS supports high data throughput rates. Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers, in a single cluster. Portability, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems.

Hive is an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase. Hadoop is intended for long sequential scans and, because Hive is based on Hadoop, queries have very high latency—which means Hive is less appropriate for applications that need very fast response times. Also, Hive is read-based, and therefore not suitable for transaction processing that typically involves a high percentage of write operations.Hive is better suited for data warehousing tasks such as ETL, reporting, and data analysis and includes tools that enable easy access to data via SQL.

This brings us to Spark, a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications, including Interactive Analytics, Streams Processing, Machine Learning, Data Integration, and ETL. It takes advantage of in-memory processing to significantly increase the speed of computations and spilling to disk only when memory is constrained. Spark has interfaces for major programming languages, including Java, Scala, Python, R, and SQL. It can run using its standalone clustering technology as well as on top of other infrastructures such as Hadoop. And it can access data in a large variety of data sources, including HDFS and Hive, making it highly versatile. The ability to process streaming data fast and perform complex analytics in real-time is the key use case for Apache Spark.

**DATA MINING**

**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 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 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.

**Lesson Summary: Big Data and Data Mining**

Welcome to the big data and data mining lesson summary video. In this lesson, you gained insights into the impact of big data on various aspects of society, from business operations to sports. And developed an understanding of key attributes and challenges associated with big data.

In this video, we will recap fundamentals of big data and how big data drives digital transformation. How data scientists leverage the essential characteristics of the Cloud to gain insights from big data, the data mining process, and common tools used to process big data.

The availability of vast amounts of data, resulting in what we now call big data, is driving transformation in business and industry and consequently, how we live our daily lives.

Organizations realize that we require fundamental changes to their approach to business, impacting every aspect of the organization. The availability of so many disparate amounts of data created by people, tools and machines requires new, innovative, and scalable technology.

Big data drives us to derive real time business insights relating to consumers risk, profit, performance productivity management, and ultimately enhancing business values. Not everyone agrees on the definition of big data, but people generally agree on the five characteristics of this data value, volume, velocity, variety and veracity.

People expect investing time in studying big data will create value. Volume refers to the scale of the data, drivers of volume include increasing collectible data sources and scalable infrastructure. Velocity indicates ever increasing sources of nonstop processes that generate data quickly. Variety reflects that related data comes from different sources, both structured and unstructured. Veracity refers to the quality and origin of data and that it accurately conforms to facts.

The development of cloud and cloud technologies enables us to work with big data. Cloud refers to the delivery of on-demand computing resources on a pay-for-use basis. Cloud computing has five essential characteristics, on-demand, network access, resource pooling, elasticity, and measured service. On-demand means access to processing, power, storage and network that you need. These computing resources can be accessed via a network with Internet access. Resource pooling allows providers to service multiple consumers with the resources dynamically assigned according to demand, making cloud computing cost efficient. Elasticity means that you can access resources as you need them and automatically scale back when you don't. With measured service, you only pay for what you use or reserve as you go.

You also gain an understanding of how cloud computing addresses challenges related to scalability, collaboration, accessibility. And software maintenance, making it a valuable resource for data analysis and other computational tasks.

The Cloud gives you instant access to technologies without needing to install or configure them, and provides updated versions of these tools as they get released. Popular open source tools to compute using big data include Apache Hadoop, Apache Hive, and Apache Spark.

Hadoop provides distributed storage and processing tools across clusters of computers. Hive is a data warehouse for data query and analysis built on top of Hadoop. Hive allows you to read, write and manage large datasets directly in the Hadoop File system or HDFS or Apache HBase. Spark provides a general purpose data processing engine designed to extract and process large volumes of data for a wide range of applications.

Big data requires a process called data mining to make use of. This six step process includes goal setting, selecting data sources, preprocessing, transforming, mining, and evaluation.

In the first step, goal setting, you identify key questions you want to answer, concerns about cost and benefits should inform this step. Once you identify the questions, select the data by identifying sources or planning data collection initiatives. In the next step, preprocessing, you identify irrelevant attributes of data and enormous aspects of the data by flagging them as necessary. After preprocessing, you transform the data by determining the appropriate format to store the data. Now you get to mine the data, which includes determining analysis methods and the machine learning algorithms you will use to process the data. Once the data has been mined, you finally must evaluate your outcomes. By testing the predictive capabilities of the models on the observed data to find effectiveness and efficiency of your algorithms. In addition, you share your results with stakeholders. This entire process should be conducted iteratively as your results from this iteration will inform further data mining efforts. In summary, big data characteristics that data scientists agree on, even though they might not agree on the exact definition, include value, volume, velocity, and veracity.

Data with these qualities is driving transformation across industries and in our daily lives. In large part, cloud technologies enable us to handle big data because they provide ubiquitous access to computational power and storage capacity. Open source cloud tools such as Hadoop, Hive, and Spark leverage these advantages, allowing us to effectively and efficiently mine big data.

**Glossary: Big Data and Data Mining**

**Texto

Descrição gerada automaticamente com confiança média**

**Uma imagem contendo Tabela

Descrição gerada automaticamente**

**Uma imagem contendo Texto

Descrição gerada automaticamente**

**Uma imagem contendo Tabela

Descrição gerada automaticamente**

**Uma imagem contendo Tabela

Descrição gerada automaticamente**

**Uma imagem contendo Aplicativo

Descrição gerada automaticamente**

**DEEP LEARNING AND MACHINE LEARNING**

**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

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.

**GENERATIVE AI AND DATA SCIENCE**

Welcome to Generative AI and Data Science. After watching this video, you will be able to describe generative AI and explain how data scientists use generative AI in data science.

Generative AI is a subset of artificial intelligence that focuses on producing new data rather than just analyzing existing data. It allows machines to create content, including images, music, language, computer code, and more, mimicking creations by people.

How does generative AI operate, though? Deep learning models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are at the foundation of this technique. These models create new instances that replicate the underlying distribution of the original data by learning patterns from enormous volumes of data.

Generative AI has found diverse applications across various industries. Let’s look at some fascinating examples! Natural language processing like OpenAI’s GPT-3 can generate human-like text, revolutionizing content creation and chatbots. In healthcare, Generative AI can synthesize medical images, aiding in the training of medical professionals. Generative AI can create unique and visually stunning artworks, generating endless creative visual compositions. Game developers use Generative AI to generate realistic environments, characters, and game levels. Generative AI assists in fashion by designing new styles and creating personalized shopping recommendations.

Now, let's discuss how data scientists use Generative AI. Training and testing a model takes lots of data. Sometimes, the data you want to study doesn’t have enough observations to build a model. Interest in synthetic data as a tool for analysis and model creation has increased due to recent developments in generative AI.

Data scientists can augment their data sets using generative AI to create synthetic data. It creates this data with similar properties as the real data, such as its distribution, clustering, and many other factors the AI learned about the real data set. Data scientists can then use the synthetic data along with the real data for model training and testing.

Data scientists, researchers, and analysts frequently find themselves confined by time when examining data patterns and insights. Due to this restriction, they can only conceive, develop, and evaluate a small number of hypotheses, leaving many other possibilities untested.

With generative AI, data scientists can leverage generative AI to generate and test software code for constructing analytical models. Coding automation has the potential to revolutionize the field of analytics, allowing the data scientist to focus on higher-level tasks such as identifying and clarifying the problem the models intend to solve and evaluating hypotheses from a wider range of data sources.

Generative AI can generate accurate business insights and comprehensive reports, making it possible to update these insights as the data evolves. Furthermore, it can autonomously explore data to uncover hidden patterns and insights that might go unnoticed during manual analysis and enhance decision-making.

For instance, IBM’s Cognos Analytics, which provides AI-powered automation, enables you to unlock the full potential of your data with the help of its natural language AI assistant. You describe the question you are looking to answer or the hypothesis you want to test, and it helps generate insights you need.

In this video, you learned that: Generative AI, a subset of artificial intelligence, focuses on producing new data rather than analyzing existing data. Generative AI augments data science efforts, enabling deeper insights, addressing data limitations, and improving the overall quality of data-driven outcomes.

**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.

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 input, 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 out of 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.

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 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.

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

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

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

So, some applications of machine learning in fintech are probably the - couple of different things I could talk about there. One of them is recommendations. Right, so, when you use Netflix, or you use Facebook, or a lot of different software services, the recommendations are served to you. Meaning, "Hey, you're a user, you've watched this show, so maybe you'd like to see this other show." Right, or, you follow this person, so maybe you should follow this other person. It's actually kind of the same thing in fintech, right. Because you've looked at - if you're an investment professional, right, and because you've looked at this investment idea, it might be really cool for you to look at this other investment idea, which is kind of similar. Right, it's a similar kind of asset, it's a similar kind of company. Or it's a similar kind of technique for doing the investment. So, We can apply recommendations using machine learning throughout a lot of different parts of fintech. Another one that people talk about, and is important especially on retail, in the retail aspects of banking and finance is fraud detection.

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

**REGRESSION**

**Chapter 7. 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 a 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 to 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 exteriors sell for less than homes with stone exteriors?

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?

**LAB: EXPLORING DATA USING IBM CLOUD GALLERY**

**IBM Cloud Gallery**

Estimated Time (45 min)

IBM Cloud Resource hub is a growing collection of data sets, notebooks, and project templates. In this lab, you will use *IBM Cloud Resource hub* to explore different datasets. As you learned in the course, data can be more than just numbers. Data can be numeric, text, images, videos, audios and more. You will look at three samples.

**Sample 1** contains data with only numeric attributes.

**Sample 2** contains data with numeric & text attributes.

**Sample 3** cantains a Jupyter Notebook, a tool which data scientists use to create models.

Let's take a look at how data scientists use different datasets.

**Objectives :**

You will learn to:

* Explore the IBM Cloud Resource hub
* Examine a numeric dataset
* Examine a dataset with non-numeric attributes
* Examine a Jupyter Notebook

**Exercise 1: Examine a numeric dataset**

1. Click on the link: <https://dataplatform.cloud.ibm.com/gallery>
2. Click the filter button in the top right of the window:

Interface gráfica do usuário

Descrição gerada automaticamente com confiança baixa

1. In the dropdown menu that appears, select the *Data* checkbox under *Sample type*. Then click on the *Tags* dropdown, and select the *Environment* checkbox.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

1. In the search results, click on *UCI: Forest Fires*.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

1. Preview the data using the *Preview* option.

Tabela

Descrição gerada automaticamente com confiança média

**Explore the data**

The data is related to forest fires where the aim is to predict the burned area of forest fires, in the northeast region of Portugal, by using meterological and other data.

**Attribute Information:**

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84  
    (this output variable is very skewed towards 0.0, thus it may make  
    sense to model with the logarithm transform).

**Exercise 2: Evaluate a non-numeric dataset**

The data doesn't have to be only based on numbers. Data can be text, images and other types as well. Let's look at a dataset which has text values.

1. At the top of the page, select the *Resource hub* option.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

1. Type *Airbnb* into the search bar.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

1. Select the *Airbnb Data for Analytics: Trentino Reviews* option. You may need to scroll to find it.

Calendário

Descrição gerada automaticamente

1. Preview the data using the *Preview* option.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

**Explore the data**

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Airbnb guests may leave a review after their stay, and these can be used as an indicator of airbnb activity. The minimum stay, price and number of reviews have been used to estimate the occupancy rate, the number of nights per year and the income per month for each listing.

You could use this data in multitude of ways - to analyze the star ratings of places, to analyze the location preferences of the customers, to analyze the tone and sentiment of customer reviews and many more. Airbnb uses location data to improve guest satisfaction.

💡 What else might you use this data for?

The dataset comprises of three main tables:

* listings - Detailed listings data showing 96 attributes for each of the listings. Some of the attributes used in the analysis are price(continuous), longitude (continuous), latitude (continuous), listing\_type (categorical), is\_superhost (categorical), neighbourhood (categorical), ratings (continuous) among others.
* reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date (datetime), listing\_id (discrete), reviewer\_id (discrete) and comment (textual).
* calendar - Provides details about booking for the next year by listing. Four attributes in total including listing\_id (discrete), date(datetime), available (categorical) and price (continuous).

**Exercise 3: Evaluate Jupyter Notebook**

Return to the Resource hub. Select *Notebook* from the *Sample type* menu that appears after clicking on the filter button. In the search bar type *Finding optimal locations* Select the card that says *Finding optimal locations of new stores using…*

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

This Jupyter notebook uses *Decision Optimization* with Python to help determine the optimal location of a new store.

This Notebook aims to identify where to place a coffee shop that minimizes the total distance from libraries in the area to the shop so that a book reader can get to the shop easily.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Part of the Python code in the notebook displays the locations of the libraries on a map.

Mapa

Descrição gerada automaticamente

But with this data, you cannot determine the ideal location of the coffee shops by just looking at the map.

The code then solves this with an optimization model that will help determine possible locations for the coffee shops with the stipulation of minimizing the distance between the libraries and the shop.

Diagrama

Descrição gerada automaticamente

**Summary**

In this lab, you have learnt about to explore datasets and notebooks in IBM cloud Resource hub.

**Author(s)**

[**Malika Singla**](https://www.linkedin.com/in/malika-goyal-04798622/)

**Other Contributor(s)**

[**Lavanya**](https://www.linkedin.com/in/lavanya-sunderarajan-199a445/)

**Change log**

| **Date** | **Version** | **Changed by** | **Change Description** |
| --- | --- | --- | --- |
| 2024-05-02 | 1.3 | Sathya Priya | Updated Screenshots and instructions |
| 2023-10-09 | 1.3 | Bethany Hudnutt | Clarified Language and updated images |
| 2022-10-27 | 1.3 | Lakshmi Holla | Updated Instructions |
| 2022-07-22 | 1.2 | Appalabhaktula Hema | Updated Screenshots and instructions |
| 2022-02-16 | 1.1 | Niveditha | Updated watson Screenshot |
| 2021-06-010 | 1.0 | Malika Singla | Initial Version |

**LESSUN SUMMARY: DEEP LEARNING AND MACHINE LEARNING**

Welcome to the Deep Learning and Machine Learning lesson summary. In this video, we’ll review what you learned about Deep Learning and Machine Learning from the videos and reading in this lesson.

We will recap: Many of the terms commonly used in artificial intelligence and how to differentiate them, How data scientists use Artificial intelligence OR AI, The relationship between machine learning and data science, AND The regression model used to show relationships between things.

Artificial intelligence (AI) has boomed and become accessible to almost everyone. Data scientists use AI regularly when analyzing data. Let’s discuss some of the terms used in this lesson related to AI, such as machine learning, deep learning, neural networks, and generative AI.

AI is the branch of computer science that includes the development of systems that can mimic many of the tasks associated with human intelligence.

Machine learning is a subset of AI that uses computer algorithms to learn about data and make predictions with it without needing to explicitly program the analysis methods into the system.

Deep learning is a subset of machine learning that uses layered neural networks to simulate human decision-making.

A neural network is a collection of small computing units, called neurons, that take incoming data and learn to make decisions over time, such as the difference between a dog and a cat. Deep learning algorithms become more efficient as the amount of data increases in volume, unlike other machine learning algorithms, which tend to plateau.

Another subset of AI, generative AI, focuses on producing new data rather than just analyzing existing data. It allows machines to create content, including images, music, languages, and computer code, mimicking human creations. Generative AI also can make data sets with similar traits to a raw data set.

Sometimes, data scientists, when they do not have enough data, can create synthetic data and use it to train and test their models. As a data scientist, you will use machine learning algorithms to derive insights from data. You will frequently apply machine learning algorithms for predictive analytics or make recommendations. For example, you may also use these algorithms for fraud detection to identify fraudulent credit card purchases based on previous purchasing habits.

Machine learning algorithms rely heavily on statistical technique called regression. Regression identifies the strength and amount of the correlation between one or more inputs and an output. For instance, how much does the price of a house increase based on its square footage and number of bedrooms, and how confident can you be of this relationship?

In summary, Generative AI produces new data like a raw data set. deep learning is a subset of machine learning, and machine learning is a subset of artificial intelligence. Deep learning utilizes neural networks to teach itself patterns in inputs and outputs. Using big data, data scientists use all of these areas of AI to make predictions.

**GLOSSARY: DEEP LEARNING AND MACHINE LEARNING**

**Uma imagem contendo Texto

Descrição gerada automaticamente**

**Uma imagem contendo Gráfico

Descrição gerada automaticamente**

**Summary: Deep Learning and Machine Learning**

Congratulations! You have completed this lesson. At this point in the course, you know:

Big Data has five characteristics: velocity, volume, variety, veracity, and value.

The five cloud computing characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.

Data mining has a six-step process: goal setting, selecting data sources, preprocessing, transforming, mining, and evaluation.

The availability of so many disparate amounts of data created by people, tools, and machines requires new, innovative, and scalable technology to drive transformation.

Deep learning utilizes neural networks to teach itself patterns in inputs and outputs. Machine learning is a subset of AI that uses computer algorithms to learn about data and make predictions without explicitly programming the analysis methods into the system.

Regression identifies the strength and amount of the correlation between one or more inputs and an output.

**MODULE 3: DATA SCIENCE APPLICATION DOMAINS**

**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 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.

**OLD PROBLEMS, NEW 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 and 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, and 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.

**APPLICATIONS OF DATA SCIENCE**

Data science and big data are making an undeniable impact on businesses, changing day-to-day operations, financial analytics, and especially interactions with customers. It's clear that businesses can gain enormous value from the insights data science can provide. But sometimes it's hard to see exactly how. So let's look at some examples.

In this era of big data, almost everyone generates masses of data every day, often without being aware of it. This digital trace reveals the patterns of our online lives. If you have ever searched for or bought a product on a site like Amazon, you'll notice that it starts making recommendations related to your search. This type of system known as a recommendation engine is a common application of data science. Companies like Amazon, Netflix, and Spotify use algorithms to make specific recommendations derived from customer preferences and historical behavior.

Personal assistants like Siri on Apple devices use data science to devise answers to the infinite number of questions end users may ask. Google watches your every move in the world, you're online shopping habits, and your social media. Then it analyzes that data to create recommendations for restaurants, bars, shops, and other attractions based on the data collected from your device and your current location.

Wearable devices like Fitbits, Apple watches, and Android watches add information about your activity levels, sleep patterns, and heart rate to the data you generate. Now that we know how consumers generate data, let's take a look at how data science is impacting business.

In 2011, McKinsey & Company said that data science was going to become the key basis of competition. Supporting new waves of productivity, growth, and innovation. In 2013, UPS announced that it was using data from customers, drivers, and vehicles, in a new route guidance system aimed to save time, money, and fuel. Initiatives like this support the statement that data science will fundamentally change the way businesses compete and operate. How does a firm gain a competitive advantage? Let's take Netflix as an example.

Netflix collects and analyzes massive amounts of data from millions of users, including which shows people are watching at what time a day when people pause, rewind, and fast-forward, and which shows directors and actors they search for. Netflix can be confident that a show will be a hit before filming even begins by analyzing users preference for certain directors and acting talent, and discovering which combinations people enjoy.

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

**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 localised weather events.

Because the real benefit of this knowledge is so important, schools are starting to include this type of data science education in their curriculum. For instance, the University of Chicago Graham School offers a Master of Science course in Threat and Response Management. Data science tools enable organizations to analyse vast quantities of data from widely different sources, and present that information in a way that allows data scientists to gain new knowledge, in some cases, saving hundreds of lives.

**THE FINAL DELIVERABLE**

The Final Deliverable

The ultimate purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy. Analytics summarize findings in tables and plots. The data scientist should then use the insights to build the narrative to communicate the findings. In academia, the final deliverable is in the form of essays and reports. Such deliverables are usually 1,000 to 7,000 words in length.

In consulting and business, the final deliverable takes on several forms. It can be a small document of fewer than 1,500 words illustrated with tables and plots, or it could be a comprehensive document comprising several hundred pages. Large consulting firms, such as McKinsey and Deloitte, I routinely generate analytics-driven reports to communicate their findings and, in the process, establish their expertise in specific knowledge domains.

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

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

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

Deloitte's "United States Economic Forecast" serves the very purpose that its authors intended. The report uses data and analytics to generate the likely economic scenarios. It builds a powerful narrative in support of the thesis statement that the U.S. economy is doing much better than 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 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.

**LESSON SUMMARY: DATA SCIENCE APPLICATIONS DOMAIN**

Welcome to the data science application lesson summary video. In this video, let’s review what you’ve learned in this lesson about the power of data science applications and how organizations leverage this power to drive business goals, improve efficiency, make predictions, and even save lives. You also reviewed the process that you will follow as a data scientist to help your organization accomplish these ends.

All organizations ultimately use data science for the same reason— to discover optimal solutions to existing problems. But to discover these solutions, your organization should identify the problem and establish a clear understanding of it.

Measurement is the first step for an organization to solve its problems using data. You need to capture and gather your data. If something is not measured, it’s challenging to improve or change it. If your organization isn’t capturing the data, help them to figure out how to capture it.

Never overwrite old data – it’s always relevant and never gets old. Once you have the data, you can start looking at it and cleaning it. As a data scientist, it’s your job to help your organization identify tools and develop an analysis strategy. Consider case studies when customizing a potential solution. Identify your analysis tools, and then develop your machine learning and statistical models. It will take time for your organization to refine data strategy best practices, but the benefits are worth it.

Everyone who uses the Internet generates mass amounts of data daily, often without being aware of it. Your company can leverage that data to reveal patterns. Take, for instance, a company like Amazon. You’ll notice that it starts making recommendations related to your search. That’s called a recommendation engine; data scientists build those engines using machine learning and statistical models.

Companies like UPS use data science to inform their business decisions. They use data from customers, drivers, and vehicles to create routes for their drivers, making more efficient use of the drivers’ time, money, and fuel.  Uber uses data to put the correct number of drivers in the right place at the right time, for a price the rider is willing to pay.

Initiatives like this show how data science fundamentally changes how businesses compete and operate. Businesses also use data science to gain competitive advantages. Take streaming companies as an example. They collect and analyze massive amounts of data from millions of users, including which shows they watch and at what time of day. They detect when people pause, rewind, and fast-forward. They collect information such as which directors and actors they search for. A streaming company can confidently predict the success of a show before filming begins by thoroughly analyzing its users’ preferences and behaviors.

Beyond helping businesses with their bottom line, data science helps companies and organizations save lives. 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 factors for a disease, including  gene markers, associated conditions, and environmental factors.

Data scientists use their models to help physicians, recommend appropriate tests, suitable trials, and suggest treatments. Developing more extensive data analytics capabilities supports healthcare organizations move from basic descriptive analytics towards predictive insights, thanks to data science. Data science tools enable organizations to analyze vast amounts of data from various sources.

Large, high-quality data sets can help predict the occurrence of natural disasters such as earthquakes, hurricanes, tornados, floods, and volcanic eruptions. The data scientists present that information in a way that allows organizations to gain new understandings and, in some cases, save lives.

When presenting your findings, you will provide your organizations with a final deliverable that explains and summarizes their conclusions. In academia, the final deliverable comprises research papers and reports. You often present the final deliverable illustrated with tables and plots in consulting and business. You will generate analytics-driven reports to communicate your findings to your organization and, in the process, establish powerful, convincing, and evidence-based narratives.

**GLOSSARY: DATA SCIENCE APPLICATION DOMAINS**

**Texto

Descrição gerada automaticamente com confiança média**

**CARRERS 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.

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.

**CARRERS 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.

**IMPORTANCE OF MATHEMATICS AND STATISTICS FOR DATA SCIENCE**

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.

If you're in high school and you're considering data science, I would say get familiar with data bases, start learning SQL, start thinking about, you know, computer science, if that's interesting. If you have a computer science course in your school, you know, take it, and that's a good part of being a data scientist. Beyond that there are probably ways to foster your creativity, right, your curiosity. If you like detective games, that's kind of cool, right. And if you like treasure hunts or whatever, right, if you're into that stuff, I think you'll, and you get the opportunity to do that stuff, that will help you as a data scientist because it's a really a good way to kind of make sure that you can be curious as you go about your daily life.

Encourage the curiosity, encourage the experimentation. It's kind of like science fairs, science fairs are great, they encourage that experimentation, that learning from, asking a question and answering it through a scientific method, but doing that with data sets rather than vinegar volcanoes. It's kind of the same thing, but learning from data and we're going through an election season right now, there's a lot of stuff in the news about polls and survey results and that's a great way to start a conversation and talk about how do the people who ran the polls, how do they know, how can they predict what's going to happen in the election. So that's another cool way to start a conversation about data science.

I would say encourage the person who is interested in data science because to pursue to, because it's a great career and it is something that is definitely going to be in need in the future. It's one of those highly sought after knowledge professions that are really important to businesses around the world. So being a data scientist and being able to help companies as they grow and learn how do to things more efficiently or how to do things smarter, there will always be a need for people like that. And data scientists are those people.

I would say that I understand what you're talking about because I was never a great mathematics student as well. And I think there's actually a bunch of data scientist, who are really successful and popular, who are in the same boat. You know there's kind of arithmetic and math in school is not necessarily everybody's best subject. But when you combine it with, you know these aren't just hypothetical numbers, these aren't just, problem statements that you have no connection to. When you have a connection to the problem, it suddenly becomes much easier to use math to help understand it, I found. And so you know, knowing the people who will benefit from the math that you do I think is really cool.

**THE 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, acknowledgments, references, and appendices (if needed).

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

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

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

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

An "introductory section" is always helpful in setting up the problem for the reader who might be new to the topic and who might need to be gently introduced to the subject matter before being immersed in intricate details. A good follow-up to the introductory section is a review of available relevant research on the subject matter. The length of the literature review section depends upon how contested the subject matter is. In instances where the vast majority of researchers have concluded in one direction, the literature review could be brief with citations for only the most influential authors on the subject. On the other hand, if the arguments are more nuanced with caveats aplenty, then you must cite the relevant research to offer adequate context before you embark on your analysis. You might use the literature review to highlight gaps in the existing knowledge, which your analysis will try to fill. This is where you formally introduce your research questions and hypothesis.

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

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

In case you need to run statistical models, you might turn to regression models (see Chapter 7, "Why Tall Parents Don't Have Even Taller Children") or categorical analysis (see Chapters 8, "To Be or Not to Be" and 2., "Categorically Speaking About Categorical Data"). If you are working with time-series data, you can turn to Chapter 11, Doing Serious Time with Time Series. You can also report results from other empirical techniques that fall under the general rubric of data mining (see Chapter 12, "Data Mining for Gold"). Note that many reports in the business sector present results in a more palatable fashion by holding back the statistical details and relying on illustrative graphics to summarize the results.

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

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

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

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

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

**Have You Done Your Job as a Writer?**

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

Have you told readers, at the outset, what they might gain by reading your paper?

Have you made the aim of your work clear?

Have you explained the significance of your contribution?

Have you set your work in the appropriate context by giving sufficient background (including a complete set of relevant references) to your work?

Have you addressed the question of practicality and usefulness?

Have you identified future developments that might result from your work?

Have you structured your paper in a clear and logical fashion?

**LESSON SUMMARY: CAREERS AND RECRUITING IN DATA SCIENCE**

Welcome to the Careers and Recruiting and Data Science, Lesson Summary video. In this lesson, you investigated what companies seek in a competent, experienced data scientist. You learned how to position yourself to get hired as a data scientist. Amidst the diverse backgrounds from which data scientists emerge, you also find they share qualities and skills that consistently set them apart from other data related roles.

Companies recruiting data scientists may seem to want data scientists to have it all. They may seek a person who has all the desired skills, ranging from domain specific knowledge to proficiency in analyzing both structured and unstructured data, as well as skills in presenting and storytelling. Realistically, this person is a rare find.

Instead, companies need to develop teams of people who work together who have these skills rather than seeking out individuals who have all of these qualities. They also may need to seek out individuals with potential and help them develop the skills they need. What really matters is they find someone passionate about the kind of business that they do.

Companies should look for someone excited about working with the data in their particular industry. They should seek out someone curious who can ask interesting, meaningful questions about the types of data they intend to collect. For example, a person who could be a great data scientist in the retail environment may not be excited about working in IT related firms or working with gigabytes of health related web logs.

Companies will find that excitement results in high productivity. That excitement also relates back to curiosity, which we discussed in another lesson as an important trait for a data scientist. Someone excited about their field is likely to be curious and find the right questions to ask. Constantly wondering why things are the way that they are helps you stay motivated and engaged.

Similarly, self-learning and tinkering are also helpful traits. As a data scientist, you should love to play with data and create data visualizations. You need to think analytically and computationally. You need a strong background in mathematics, statistics, and probability to reach valid conclusions with that data. You also need computer programming skills. Tools may vary with the type of data that you'll work with.

Though common languages include Python and R, an open source programming language developed for statistical analysis. Because data scientists work with large quantities of stored data, you must also understand data storage and retrieval systems with structured and unstructured data. With artificial intelligence, you also need to know common machine learning algorithms to gain insights from their data.

One or more persons from the data science team at an organization will write a report as the final step of the analysis, so communication, instructional and presentational skills, and storytelling skills are also important. The report synthesizes these large amounts of data to create a narrative that engages and surprises the reader. A clearly organized and logical report should communicate the following to the reader, what they gain by reading the report, clearly defined goals, the significance of your contribution.

Appropriate context by giving sufficient background, why this work is practical and useful, and conjugate plausible future developments that might result from your work. As Dr. Hader explained, imagine your reader is driving on a mountain. There's a sharp turn ahead, and they can't see what's beyond the turn. They go around the curve and suddenly they see a tremendous valley in front of them, they experience this great sense of awe. So, when you were presenting this great finding and communicating it well, this is what people feel because they were not expecting it.

In summary, companies need to look for individuals to create a data science team and be cautious about trying to find all the desired skills in a single individual. On that team, they should have curious people who understand the subject matter well, people who love working with the data and those with statistics, mathematics, machine learning and computer programming expertise. Lastly and possibly most importantly, they need a skilled storyteller who can present the team's findings in a creative and engaging way.

**GLOSSARY: CARRERS AND RECRUITING IN DATA SCIENCE**

Texto

Descrição gerada automaticamente

Texto, Aplicativo

Descrição gerada automaticamente com confiança média

**SUMMARY: CARRERS AND RECRUITING IN DATA SCIENCE**

Congratulations! You have completed this module. At this point, you know that:

Data Science helps physicians provide the best treatment for their patients, helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.

Companies can start on their data science journey by capturing data. Once they have data, they can begin analyzing it.

Everyone who uses the Internet generates mass amounts of data daily.

Amazon and Netflix use recommendation engines, and UPS uses data from customers, drivers, and vehicles to use the drivers’ time and fuel efficiently.

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.

The report should present a thorough analysis of the data and communicate the project findings.

Companies should look for someone excited about working with the data in their particular industry. They should seek out someone curious who can ask interesting, meaningful questions about the types of data they intend to collect. They should hire people who love working with data, are fluent in statistics, and are competent in applying machine learning algorithms.

A clearly organized and logical report should communicate the following to the reader:

What they gain by reading the report

Clearly defined goals

The significance of your contribution

Appropriate context by giving sufficient background

Why this work is practical and useful

Conjecture plausible future developments that might result from your work

**CASE STUDY: FINAL ASSIGNMENT**

**Case Study: Lila's Journey to Becoming a Data Scientist: Her Working Approach on the First Task**

This case study explores the data scientist's career path and key attributes, highlighting the skills, education, and experiences required to excel in this dynamic field. We'll follow the story of Lila, a fictional individual who aspires to become a successful data scientist.

There will be a quiz after this reading based on the contents of this case study.

**Education and Skill Acquisition**

With an economics undergraduate degree and a substantial data analysis background, Lila finds data science and its potential to drive meaningful change captivating. Inspired by her experiences, she makes a determined decision to transition her career and step into the role of a data scientist.

Lila realizes that to embark on her data science journey, she needs to enhance her skills and knowledge. She enrolled in the IBM Data Science Professional Certificate online program that covers key topics like statistics, machine learning, data analysis, and programming languages like Python and SQL. She diligently completes coursework and practices her coding skills on real datasets.

**Building a Strong Foundation**

As she progresses in her studies, Lila gains a deep understanding of data science fundamentals such as data manipulation and visualization with Python libraries like NumPy, Pandas, and Matplotlib. This strong foundation equips her with essential skills for data analysis.

**Visualization for Storytelling**

Lila knows she must communicate her findings effectively, so she learns which types of data visualizations will be most informative. She learns to create charts and graphs that visually represent data like sales trends, customer segmentation, and product popularity, allowing stakeholders to grasp the data's significance. These visualizations help in storytelling and decision-making.

**Hands-On Experience**

Lila understands that practical experience is invaluable in data science. She started participating in Kaggle competitions and working on personal data projects. These experiences expose her to real-world data problems and help her develop problem-solving skills. Furthermore, she created her GitHub account and uploaded her projects to build her profile.

**Data Wrangling and Preprocessing**

Lila learns that data scientists spend a significant portion of their time on data cleaning and preprocessing. She worked on various datasets, learned data preprocessing as she used sed NumPy and pandas Python libraries, and became skilled in handling missing data, outlier detection, and feature engineering to improve model performance.

**Communication and Storytelling**

Recognizing that data scientists must communicate their findings effectively, Lila honed her data storytelling skills. She learned various tools like matplotlib and plotly while she pursued her IBM Data Science Professional Certificate. She learned how to create compelling visualizations and present her insights in a clear and understandable manner.

**Networking and Collaboration**

Lila actively participates in data science communities and attends meetups and conferences. She collaborates on open-source projects, connects with fellow data scientists, and gains exposure to various industries when she attended the IBM TechXchange Conference.

**Domain Expertise**

Understanding that domain knowledge is crucial, Lila chooses a niche that aligns with her interests. She looks deeply into several domains, including e-commerce, healthcare, finance, and several other fields to which she could apply her data science skills effectively. Since her master's in economics, she chose e-commerce as her core domain to land herself a data science career.

**Landing the First Job**

After months of preparation, Lila started applying for data scientist positions. She tailors her resume to highlight her relevant skills and projects. Her online portfolio showcases her capabilities and demonstrates her commitment to the field.

**Lila's Approach to Working on Her First Task as a Data Scientist**

As a newly hired junior data scientist at a retail company, Lila uses data insights to improve customer service. Her first assignment involves diving into customer data to identify patterns and anomalies that could impact customer service. She uses data analysis to enhance the overall customer experience.

**Dataset Selection and Sourcing**

In the initial phase of her data science journey, Lila faced the challenge of selecting a suitable dataset and procuring it from different sources. Apart from the historical data available for the organizations for the past four years, she scoured various repositories, websites, and databases to find the right datasets for her project. Upon collecting data from diverse sources, Lila encountered another crucial decision point. She had to decide how to harmonize and integrate these disparate datasets into a cohesive whole. She reached out to product professionals, data engineers, and domain specialists, seeking their input and expertise in merging datasets.

**Data Understanding and Cleaning**

Lila begins by importing the dataset into her data analysis environment using Python and SQL. She loads the data and examines the first few rows to understand its structure and contents. Upon acquiring the dataset, Lila encounters her first challenge: data cleaning. Lila checks for missing values, duplicates, and outliers in the dataset. She addresses missing data by imputing or removing rows or columns with missing values. Outliers are identified and treated appropriately based on their impact on the analysis.

**Exploratory Data Analysis (EDA)**

As she delves into exploratory data analysis, Lila faces numerous choices. She must determine which summary statistics, visualizations, and distribution analyses will best reveal insights into customer behavior and sales trends. Each choice she makes during EDA influences the story the data will tell. Lila conducts EDA to gain insights into the dataset. She generates summary statistics and visualizations (histograms, scatter plots) and explores the distribution of variables. EDA helps her understand customer behavior, popular products, and sales trends.

**Feature Engineering**

Lila recognizes the potential for feature engineering to enhance her analysis. She assesses whether creating new features, such as calculating total purchase amounts, will improve the dataset's utility for her project.

**Statistical Analysis, Machine Learning**

Lila evaluates whether statistical tests or machine learning algorithms are necessary. She employs regression analysis to understand relationships between variables and explore machine learning models for demand forecasting or customer segmentation tasks. Lila also performs statistical tests to uncover patterns in the data. She uses regression analysis to understand the impact of unit price on sales.

**Presentation and Reporting**

At the culmination of her analysis, Lila faces the challenge of presenting her findings. Lila compiles her analysis and findings using a Jupyter Notebook into a comprehensive report and presentation. She highlights actionable insights and recommendations for the e-commerce platform's stakeholders.

**Continuous Learning**

After completing her first project, Lila continues to refine her skills, explores more complex datasets, and tackles increasingly challenging data science tasks.

**Machine Learning Skills**

Although Lila took an introductory course on Machine Learning as part of the IBM Data Science Professional Certificate, the field intrigues her, and she wants to develop her skills further by taking the IBM Machine Learning Professional Certificate. She identified Machine Learning Repository datasets in the course and experimented with various algorithms. Lila dives into machine learning to excel as a data scientist, wherein she studies various algorithms, such as linear regression, decision trees, and deep learning models. She continues to gain expertise in selecting and fine-tuning algorithms based on specific data problems.

**EXPLORE DATA SCIENCE JOB LISTINGS**

Review and evaluate a data science job post.

Assignment overview

For this project, you should find a data science job posting on a job board of your choice, such as LinkedIn, Indeed, Zip Recruiter, Glassdoor, Monster, Naukri, or USAjobs.gov, that interests you.

Analyze the posting by responding to the following questions and statements. You do not need to submit your responses. This is an exercise to familiarize yourself with actual data science related jobs.

Identify the following aspects of data science job post:

What is the company name that is advertising the job?

What is the job title?

Where is the role located?

What is the expected salary or salary range?

What is the total number of results from the search for the job post?

What is one technical responsibility from the job post related to something you learned about in this course?

What are two required technical skills from the job post?

What are at least two ideas or concepts you learned about in this course relevant to these job posts?

**COURSE SUMMARY**

Congratulations! You have completed this course. At this point, you know that:

Data science is the practice of extracting valuable insights from vast datasets to guide strategic decision-making.

Data science careers offer diverse paths, often involving mathematics, programming, and a curiosity for data exploration.

Successful data scientists exhibit qualities like curiosity, critical judgment, and an aptitude for constructive argumentation.

The data science field is characterized by high demand, resulting in attractive remuneration for skilled professionals.

A Data Scientist's daily routine can vary significantly depending on the project's nature.

A wide array of algorithms is available for extracting insights from data.

Big Data plays a pivotal role in driving digital transformation across industries.

Cloud computing is a fundamental technology in modern data science.

Data mining techniques are essential for uncovering patterns and knowledge from data.

Tools like Hadoop, HDFS, Hive, and Spark are employed for processing Big Data.

Deep learning, machine learning, and regression are critical data science topics.

Data science applications span diverse domains, solving complex problems.

Companies can harness data science to address age-old challenges with innovative solutions.

Data science contributes significantly to saving lives and improving various aspects of society.

Careers in data science offer exciting opportunities, with mathematics and statistics being essential foundations.

Reports in data science adhere to specific structures, and career roadmaps provide guidance.

Case studies and projects offered practical application of the knowledge acquired during the course.

**MODULE 4: UNDERSTANDING DATA**

**UNDERSTANDING DATA**

Data is unorganized information that is processed to make it meaningful. Generally, data comprises of facts, observations, perceptions, numbers, characters, symbols, and images that can be interpreted to derive meaning. One of the ways in which data can be categorized is by its structure. Data can be: Structured; Semi-structured, or Unstructured.

Structured data has a well-defined structure or adheres to a specified data model, can be stored in well-defined schemas such as databases, and in many cases can be represented in a tabular manner with rows and columns. Structured data is objective facts and numbers that can be collected, exported, stored, and organized in typical databases. Some of the sources of structured data could include: SQL Databases and Online Transaction Processing (or OLTP) Systems that focus on business transactions, Spreadsheets such as Excel and Google Spreadsheets, Online forms, Sensors such as Global Positioning Systems (or GPS) and Radio Frequency Identification (or RFID) tags; and Network and Web server logs. You can typically store structured data in relational or SQL databases. You can also easily examine structured data with standard data analysis methods and tools.

Semi-structured data is data that has some organizational properties but lacks a fixed or rigid schema. Semi-structured data cannot be stored in the form of rows and columns as in databases. It contains tags and elements, or metadata, which is used to group data and organize it in a hierarchy. Some of the sources of semi-structured data could include: E-mails, XML, and other markup languages, Binary executables, TCP/IP packets, Zipped files, Integration of data from different sources. XML and JSON allow users to define tags and attributes to store data in a hierarchical form and are used widely to store and exchange semi-structured data.

Unstructured data is data that does not have an easily identifiable structure and, therefore, cannot be organized in a mainstream relational database in the form of rows and columns. It does not follow any particular format, sequence, semantics, or rules. Unstructured data can deal with the heterogeneity of sources and has a variety of business intelligence and analytics applications. Some of the sources of unstructured data could include: Web pages, Social media feeds, Images in varied file formats (such as JPEG, GIF, and PNG), video and audio files, documents and PDF files, PowerPoint presentations, media logs; and surveys. Unstructured data can be stored in files and documents (such as a Word doc) for manual analysis or in NoSQL databases that have their own analysis tools for examining this type of data.

To summarize, structured data is data that is well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools; Semi-structured data is data that is somewhat organized and relies on meta tags for grouping and hierarchy; and Unstructured data is data that is not conventionally organized in the form of rows and columns in a particular format In the next video, we will learn about the different types of file structures.

**DATA SOURCES**

As we touched upon in one of our previous videos, data sources have never been as dynamic and diverse as they are today. In this video, we will look at some common sources such as Relational Databases; Flatfiles and XML Datasets APIs and Web Services; Web Scraping; Data Streams and Feeds.

Typically, organizations have internal applications to support them in managing their day to day business activities, customer transactions, human resource activities, and their workflows. These systems use relational databases such as SQL Server, Oracle, MySQL, and IBM DB2, to store data in a structured way.

Data stored in databases and data warehouses can be used as a source for analysis. For example, data from a retail transactions system can be used to analyze sales in different regions, and data from a customer relationship management system can be used for making sales projections.

External to the organization, there are other publicly and privately available datasets. For example, government organizations releasing demographic and economic datasets on an ongoing basis. Then there are companies that sell specific data, for example, Point-of-Sale data or Financial data, or Weather data, which businesses can use to define strategy, predict demand, and make decisions related to distribution or marketing promotions, among other things. Such data sets are typically made available as flat files, spreadsheet files, or XML documents.

Flat files, store data in plain text format, with one record or row per line, and each value separated by delimiters such as commas, semi-colons, or tabs. Data in a flat file maps to a single table, unlike relational databases that contain multiple tables. One of the most common flat-file format is CSV in which values are separated by commas. Spreadsheet files are a special type of flat files, that also organize data in a tabular format–rows and columns. But a spreadsheet can contain multiple worksheets, and each worksheet can map to a different table. Although data in spreadsheets is in plain text, the files can be stored in custom formats and include additional information such as formatting, formulas, etc. Microsoft Excel, which stores data in .XLS or .XLSX format is probably the most common spreadsheet. Others include Google sheets, Apple Numbers, and LibreOffice.

XML files, contain data values that are identified or marked up using tags. While data in flat files is “flat” or maps to a single table, XML files can support more complex data structures, such as hierarchical. Some common uses of XML include data from online surveys, bank statements, and other unstructured data sets.

Many data providers and websites provide APIs, or Application Program Interfaces, and Web Services, which multiple users or applications can interact with and obtain data for processing or analysis. APIs and Web Services typically listen for incoming requests, which can be in the form of web requests from users or network requests from applications, and return data in plain text, XML, HTML, JSON, or media files.

Let’s look at some popular examples of APIs being used as a data source for data analytics: The use of Twitter and Facebook APIs to source data from tweets and posts for performing tasks such as opinion mining or sentiment analysis—which is to summarize the amount of appreciation and criticism on a given subject, such as policies of a government, a product, a service, or customer satisfaction in general.

Stock Market APIs used for pulling data such as share and commodity prices, earnings per share, and historical prices, for trading and analysis. Data Lookup and Validation APIs, which can be very useful for Data Analysts for cleaning and preparing data, as well as for co-relating data—for example, to check which city or state a postal or zip code belongs to. APIs are also used for pulling data from database sources, within and external to the organization.

Web ScrapingWeb scraping is used to extract relevant data from unstructured sources. Also known as screen scraping, web harvesting, and web data extraction, web scraping makes it possible to download specific data from web pages based on defined parameters. Web scrapers can, among other things, extract text, contact information, images, videos, product items, and much more from a website. Some popular uses of web scraping include collecting product details from retailers, manufacturers, and eCommerce websites to provide price comparisons; generating sales leads through public data sources; extracting data from posts and authors on various forums and communities; and collecting training and testing datasets for machine learning models Some of the popular web scraping tools include BeautifulSoup, Scrapy, Pandas, and Selenium.

Data streams are another widely used source for aggregating constant streams of data flowing from sources such as instruments, IoT devices, and applications, GPS data from cars, computer programs, websites, and social media posts. This data is generally timestamped and also geo-tagged for geographical identification. Some of the data streams and ways in which they can be leveraged include: stock and market tickers for financial trading; retail transaction streams for predicting demand and supply chain management; surveillance and video feeds for threat detection; social media feeds for sentiment analysis;sensor data feeds for monitoring industrial or farming machinery; web click feeds for monitoring web performance and improving design; and real-time flight events for rebooking and rescheduling. Some popular applications used to process data streams include Apache Kafka, Apache Spark Streaming, and Apache Storm.

RSS (or Really Simple Syndication) feeds, are another popular data source. These are typically used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis. Using a feed reader, which is an interface that converts RSS text files into a stream of updated data, updates are streamed to user devices.

**VIEWPOINTS: WORKING WITH VARIED DATA SOURCES AND TYPES**

In this video, we will listen to several data professionals talk about their experience of working with very data sources and types of data. So you'd be surprised the different ways that data can come at you.

I tend to be a relational database fan. And so I spend a lot of time with SQL and using the power of SQL to deal with moving data from one place to another, to deal with structuring of data, to deal with all the security details around data. But obviously that doesn't apply to every scenario and even when we're dealing entirely in relational databases we're often moving data from one relational database to another. And especially when we're talking about one vendor to another that can be challenging. The things that also get in the way tend to be the versioning. So sometimes the feature of something that you want is in a version two levels above where you are, or it doesn't work the same way as it did two versions ago. So working with multiple data sources is about flexibility. It's about finding the function that works and works with the performance you need. Moving data one time is usually not all that hard as long as we're sub-terabyte. But moving data consistently and continually, and in a performant way can cause us to evaluate a lot of different solutions. So we really need to be open to new ideas and looking for new solutions that meet the requirements that we have. Mostly I work with relational databases. They are extremely flexible and withstood the test of time. However, with the evolution of unstructured data such as logs, documents, XML, and JSON, their reputation as a cure for all of your data problems came under intense scrutiny. And most of data intensive applications such as IoT on social media applications started to look elsewhere. For example, Google released a white paper back in 2006 called Google BigTable.

That idea quickly caught fire. For example, Cassandra and HBase came out of the same architectural model as the Google BigTable. And they became widely popular databases to solve some of the problems that relational databases failed to solve. For example, relational databases struggle a little bit with heavy write intensive applications such as IoT or sensor data, social media data because the B-tree data structures that drive, or power, these relational databases slows down due to their nature of the random reads and random writes for the heavy write applications. It's an inevitable part of a data engineer's job to work with a variety of data. You will need to work with standard formats like CSV, JSON, XML, but also you'll need to work with proprietary formats. And you will need to get data in different sources, whether it be relational databases, NoSQL or big data repositories.

You will need to work with data at rest, streaming data, or data in motion. And you might not have the skills to work with all of these different types of data sources from day one. But you need to be able to learn as you go and pick up the skills required for the project to work with different datasets, different data formats, and different data sources.

When it comes to the data formats, log data, XML data, JSON, etc., each of them comes with their own challenges. For example, log data is extremely challenging because it's unstructured and you may need to write your own custom tools to pass the data depending on what you want to look at.

Whereas XML was widely popular like a decade ago, especially with the SOAP protocol of the web applications. However, soon the web developers and corporations discovered that it can be a resource intensive, especially memory, because it has both the starting and ending tags. So then JSON came into the picture. They got it off the ending tags and just looked like a key-value pairs and it saved some resources. And it is now widely used as part of the RESTful APIs. And then even newer versions of the data format such as Apache Avro are gaining wide popularity because of the efficiency on how they store the data.

One particular situation where we were converting data from a Db2 database into a SQL Server database and it was challenging because the way that each of those expect imports and exports to happen is a little bit different. The data was particularly challenging, and that's where a lot of your challenge might come from in these projects, is from the data itself. In this particular case, the data had a lot of different characters in it. So usually we're looking for a character we can use as a delimiter. Oftentimes that's comma delimited, so we can separate our fields using commas, but we also have to think about situations where we have data that has commas in it. How do we properly separate that data? How do we properly define our fields? And in this particular case we had to use different separators for different tables, because every single special character that we could think of was in one of those tables. And the special characters that weren't were sometimes ones we couldn't use for separation, such as the Bell character.

**READING: METADATA**

**Metadata and Metadata Management**

**Objectives**

After completing this reading, you will be able to:

Define what metadata is

Describe what metadata management is

Explain the importance of metadata management

List popular tools for metadata management

**What is metadata?**

Metadata is data that provides information about other data.

This is a very broad definintion. Here we will consider the concept of metadata within the context of databases, data warehousing, business intelligence systems, and all kinds of data repositories and platforms.

We'll consider the following three main types of metadata:

Technical metadata

Process metadata, and

Business metadata

**Technical metadata**

Technical metadata is metadata which defines the data structures in data repositories or platforms, primarily from a technical perspective.

For example, technical metadata in a data warehouse includes assets such as:

Tables that record information about the tables stored in a database, like:

each table's name

the number of columns and rows each table has

A data catalog, which is an inventory of tables that contain information, like:

the name of each database in the enteprise data warehouse

the name of each column present in each database

the names of every table that each column is contained in

the type of data that each column contains

The technical metadata for relational databases is typically stored in specialized tables in the database called the System Catalog.

**Process metadata**

Process metadata describes the processes that operate behind business systems such as data warehouses, accounting systems, or customer relationship management tools.

Many important enterprise systems are responsible for collecting and processing data from various sources. Such critical systems need to be monitored for failures and any performance anomalies that arise. Process metadata for such sytems includes tracking things like:

process start and end times

disk usage

where data was moved from and to, and

how many users access the system at any given time

This sort of data is invaluable for troubleshooting and optimizing workflows and ad hoc queries.

**Business metadata**

Users who want to explore and analyze data within and outside the enterprise are typically interested in data discovery. They need to be able to find data which is meaningful and valuable to them and know where that data can be accessed from. These business-minded users are thus interested in business metadata, which is information about the data described in readily interpretable ways, such as:

how the data is acquired

what the data is measuring or describing

the connection between the data and other data sources

Business metadata also serves as documentation for the entire data warehouse system.

**Managing metadata**

Managing metadata includes developing and administering policies and processes to ensure information can be accessed and integrated from various sources and appropriately shared across the entire enterprise.

Creation of a reliable, user-friendly data catalog is a primary objective of a metadata management model. The data catalog is a core component of a modern metadata management system, serving as the main asset around which metadata management is administered. It serves as the basis by which companies can inventory and efficiently organize their data systems. A modern metadata management model will include a web-based user interface that enables engineers and business users to easily search for and find information on key attributes such as CustomerName or ProductType. This kind of model is central to any Data Governance initiative.

**Why is metadata management important?**

Good metadata management has many valuable benefits. Having access to a well implemented data catalog greatly enhances data discovery, repeatability, governance, and can also facilitate access to data.

Well managed metadata helps you to understand both the business context associated with the enterprise data and the data lineage, which helps to improve data governance. Data lineage provides information about the origin of the data and how it gets transformed and moved, and thus it facilitates tracing of data errors back to their root cause. Data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data, and data controls are implemented that support business objectives.

The key focus areas of data governance include availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organization.

**Popular tools for metadata management**

Popular metadata management tools include:

IBM InfoSphere Information Server

CA Erwin Data Modeler

Oracle Warehouse Builder

SAS Data Integration Server

Talend Data Fabric

Alation Data Catalog

SAP Information Steward

Microsoft Azure Data Catalog

IBM Watson Knowledge Catalog

Oracle Enterprise Metadata Management (OEMM)

Adaptive Metadata Manager

Unifi Data Catalog

data.world

Informatica Enterprise Data Catalog

**Summary**

In this reading, you learned that:

Metadata is data that provides information about other data, and includes three main types: technical, process, and business metadata

The technical metadata for relational databases is typically stored in specialized tables in the database called the system catalog

A primary objective of business metadata management modelling is the creation and maintenance of a reliable, user-friendly data catalog

Having access to a well-implemented data catalog greatly enhances data discovery, repeatability, governance, and can also facilitate access to data

Metadata management tools from IBM include InfoSphere Information Server and Watson Knowledge Catalog

**LESSON SUMMARY: UNDERSTANDING DATA**

Welcome to the understanding data lesson summary video. Data is the foundation of data science. To be a successful data scientist, you must understand data in different forms, whether structured, semi-structured, or unstructured. In the videos and reading in this lesson, you explored how data can be generated, stored, and accessed. For data to be useful, it must be interpreted to derive meaning.

Structured data has a well-defined structure or adheres to a specified data model, can be stored in well-defined schemas such as databases, and in many cases, can be represented in multiple tables with rows and columns. The schema defines the relationships between tables.

Semi-structured has some organizational properties, but lacks a fixed or rigid schema. It cannot be stored in rows and columns. It contains tags and elements, or metadata, which is used to organize data in a hierarchy. The metadata provides useful information about this data.

Metadata can be categorized as technical, process, or business. Metadata also needs managing to ensure information access and integration from various sources and appropriately shared across an organization, usually in a data catalog. A well-implemented data catalog enhances data discovery, repeatability, governance, and access.

Unstructured data is heterogenous. It comes from a broad range of sources and has a variety of business intelligence and analytics applications. Analyzing unstructured data often requires artificial intelligence to gain insights from it. It is important to know that the data for analysis can be sourced from anything with an electronic footprint, whether it’s generated and stored automatically or through manual efforts. While older records may exist in analog formats like paper, they typically require conversion to electronic formats for effective mining and processing. Typically, organizations use internal applications to support their day-to-day management of business activities and workflows.

You can use data stored in databases and data warehouses as a source for analysis. You can also find publicly and privately held data sets for analysis. You can even purchase proprietary data sets. These data sets are usually made available as a single flat file such as CSV format or spreadsheets. For a number of years, flat files stored information about their structure in a format called XML. More recently, you see text data exchanged in JSON format, a straightforward format readable by humans and machines. JSON data does not have a predefined schema, so you can easily transfer it between data structures as they evolve. You can usually access the data in modern cloud-based software applications through an application programming interface or API.

When dealing with JSON data, applications use what is called a “RESTful API” for data transfer. Many data providers and websites provide APIs for accessing their data. Twitter and Facebook both provide APIs to source data from posts for performing tasks such as opinion mining. This allows you to analyze sentiments on a given subject, such as customer satisfaction. Usually data gathering and management are typically assigned to data engineers, although as a data scientist, you most often need to transfer data for analysis.

This process demands flexibility. As a data scientist, you will frequently work with extensive data sets, sometimes reaching terabytes in size. These continually updating massive data sets stem from sources like IoT applications, sensor data, and the constant influx of social media data collected from millions of users. Data scientists have a unique relationship with data. You specialize in data, so you must understand it intimately. But before you can analyze it, you’ll need to know the modern-day ecosystem for organizing, storing, manipulating, and retrieving data. This lesson introduced you to some of these ideas.

**GLOSSARY: UNDERSTANDING DATA**

**Texto

Descrição gerada automaticamente**

**DATA LITERACY**

**DATA COLLECTION AND ORGANIZATION**

A data repository is a general term used to refer to data that has been collected, organized, and isolated so that it can be used for business operations or mined for reporting and data analysis. It can be a small or large database infrastructure with one or more databases that collect, manage, and store data sets. In this video, we will provide an overview of the different types of repositories your data might reside in, such as databases, data warehouses, and big data stores, and examine them in greater detail in further videos.

Let’s begin with databases. A database is a collection of data, or information, designed for the input, storage, search and retrieval, and modification of data. And a Database Management System, or DBMS, is a set of programs that creates and maintains the database. It allows you to store, modify, and extract information from the database using a function called querying. For example, if you want to find customers who have been inactive for six months or more, using the query function, the database management system will retrieve data of all customers from the database that have been inactive for six months and more.

Even though a database and DBMS mean different things the terms are often used interchangeably. There are different types of databases. Several factors influence the choice of database, such as the data type and structure, querying mechanisms, latency requirements, transaction speeds, and intended use of the data. It’s important to mention two main types of databases here—relational and non-relational databases.

Relational databases, also referred to as RDBMSes, build on the organizational principles of flat files, with data organized into a tabular format with rows and columns following a well-defined structure and schema. However, unlike flat files, RDBMSes are optimized for data operations and querying involving many tables and much larger data volumes. Structured Query Language, or SQL, is the standard querying language for relational databases.

Then we have non-relational databases, also known as NoSQL, or “Not Only SQL”. Non-relational databases emerged in response to the volume, diversity, and speed at which data is being generated today, mainly influenced by advances in cloud computing, the Internet of Things, and social media proliferation. Built for speed, flexibility, and scale, non-relational databases made it possible to store data in a schema-less or free-form fashion. NoSQL is widely used for processing big data.

A data warehouse works as a central repository that merges information coming from disparate sources and consolidates it through the extract, transform, and load process, also known as the ETL process, into one comprehensive database for analytics and business intelligence. At a very high-level, the ETL process helps you to extract data from different data sources, transform the data into a clean and usable state, and load the data into the enterprise’s data repository.

Related to Data Warehouses are the concepts of Data Marts and Data Lakes, which we will cover later. Data Marts and Data Warehouses have historically been relational, since much of the traditional enterprise data has resided in RDBMSes. However, with the emergence of NoSQL technologies and new sources of data, non-relational data repositories are also now being used for Data Warehousing.

Another category of data repositories are Big Data Stores, that include distributed computational and storage infrastructure to store, scale, and process very large data sets. Overall, data repositories help to isolate data and make reporting and analytics more efficient and credible while also serving as a data archive.

**NOSQL**

NoSQL, which stands for “not only SQL,” or sometimes “non SQL” is a non-relational database design that provides flexible schemas for the storage and retrieval of data. NoSQL databases have existed for many years but have only recently become more popular in the era of cloud, big data, and high-volume web and mobile applications. They are chosen today for their attributes around scale, performance, and ease of use. It's important to emphasize that the "No" in "NoSQL" is an abbreviation for "not only" and not the actual word "No." NoSQL databases are built for specific data models and have flexible schemas that allow programmers to create and manage modern applications. They do not use a traditional row/column/table database design with fixed schemas, and typically not use the structured query language (or SQL) to query data, although some may support SQL or SQL-like interfaces. NoSQL allows data to be stored in a schema-less or free-form fashion. Any data, be It structured, semi-structured, or unstructured, can be stored in any record. Based on the model being used for storing data, there are four common types of NoSQL databases: Key-value store, Document-based, Column-based, and graph-based.

Key-value store: Data in a key-value database is stored as a collection of key-value pairs. The key represents an attribute of the data and is a unique identifier. Both keys and values can be anything from simple integers or strings to complex JSON documents. Key-value stores are great for storing user session data and user preferences, making real-time recommendations and targeted advertising, and in-memory data caching. However, if you want to be able to query the data on specific data value, need relationships between data values, or need to have multiple unique keys, a key-value store may not be the best fit. Redis, Memcached, and DynamoDB are some well-known examples in this category.

Document-based: Document databases store each record and its associated data within a single document. They enable flexible indexing, powerful ad hoc queries, and analytics over collections of documents. Document databases are preferable for eCommerce platforms, medical records storage, CRM platforms, and analytics platforms. However, if you’re looking to run complex search queries and multi-operation transactions, a document-based database may not be the best option for you. MongoDB, DocumentDB, CouchDB, and Cloudant are some of the popular document-based databases.

Column-based: Column-based models store data in cells grouped as columns of data instead of rows. A logical grouping of columns, that is, columns that are usually accessed together, is called a column family. For example, a customer’s name and profile information will most likely be accessed together but not their purchase history. So, customer name and profile information data can be grouped into a column family. Since column databases store all cells corresponding to a column as a continuous disk entry, accessing and searching the data becomes very fast. Column databases can be great for systems that require heavy write requests, storing time-series data, weather data, and IoT data. But if you need to use complex queries or change your querying patterns frequently, this may not be the best option for you. The most popular column databases are Cassandra and HBase.

Graph-based: Graph-based databases use a graphical model to represent and store data. They are particularly useful for visualizing, analyzing, and finding connections between different pieces of data. The circles are nodes, and they contain the data. The arrows represent relationships. Graph databases are an excellent choice for working with connected data, which is data that contains lots of interconnected relationships. Graph databases are great for social networks, real-time product recommendations, network diagrams, fraud detection, and access management. But if you want to process high volumes of transactions, it may not be the best choice for you, because graph databases are not optimized for large-volume analytics queries. Neo4J and CosmosDB are some of the more popular graph databases.

Advantages of NoSQL was created in response to the limitations of traditional relational database technology. The primary advantage of NoSQL is its ability to handle large volumes of structured, semi-structured, and unstructured data. Some of its other advantages include: The ability to run as distributed systems scaled across multiple data centers, which enables them to take advantage of cloud computing infrastructure; An efficient and cost-effective scale-out architecture that provides additional capacity and performance with the addition of new nodes; and Simpler design, better control over availability, and improved scalability that enables you to be more agile, more flexible, and to iterate more quickly.

To summarize the key differences between relational and non-relational databases: RDBMS schemas rigidly define how all data inserted into the database must be typed and composed, whereas NoSQL databases can be schema-agnostic, allowing unstructured and semi-structured data to be stored and manipulated. Maintaining high-end, commercial relational database management systems is expensive whereas NoSQL databases are specifically designed for low-cost commodity hardware Relational databases, unlike most NoSQL, support ACID-compliance, which ensures reliability of transactions and crash recovery. RDBMS is a mature and well-documented technology, which means the risks are more or less perceivable as compared to NoSQL, which is a relatively newer technology. Nonetheless, NoSQL databases are here to stay, and are increasingly being used for mission critical applications.

**DATA MARTS, DATA LAKES, ETL, AND DATA PIPELINES**

Earlier in the course, we examined databases, data warehouses, and big data stores. Now we’ll go a little deeper in our exploration of data warehouses, data marts, and data lakes; and also learn about the ETL process and data pipelines. A data warehouse works like a multi-purpose storage for different use cases. By the time the data comes into the warehouse, it has already been modeled and structured for a specific purpose, meaning it is analysis ready. As an organization, you would opt for a data warehouse when you have massive amounts of data from your operational systems that needs to be readily available for reporting and analysis.

Data warehouses serve as the single source of truth—storing current and historical data that has been cleansed, conformed, and categorized. A data warehouse is a multi-purpose enabler of operational and performance analytics.

A data mart is a sub-section of the data warehouse, built specifically for a particular business function, purpose, or community of users. The idea is to provide stakeholders data that is most relevant to them, when they need it. For example, the sales or finance teams accessing data for their quarterly reporting and projections. Since a data mart offers analytical capabilities for a restricted area of the data warehouse, it offers isolated security and isolated performance. The most important role of a data mart is business-specific reporting and analytics.

A Data Lake is a storage repository that can store large amounts of structured, semi-structured, and unstructured data in their native format, classified and tagged with metadata. So, while a data warehouse stores data processed for a specific need, a data lake is a pool of raw data where each data element is given a unique identifier and is tagged with metatags for further use. You would opt for a data lake if you generate, or have access to, large volumes of data on an ongoing basis, but don’t want to be restricted to specific or pre-defined use cases. Unlike data warehouses, a data lake would retain all source data, without any exclusions. And the data could include all types of data sources and types. Data lakes are sometimes also used as a staging area of a data warehouse. The most important role of a data lake is in predictive and advanced analytics.

Now we come to the process that is at the heart of gaining value from data—the Extract, Transform, and Load process, or ETL. ETL is how raw data is converted into analysis-ready data. It is an automated process in which you gather raw data from identified sources, extract the information that aligns with your reporting and analysis needs, clean, standardize, and transform that data into a format that is usable in the context of your organization; and load it into a data repository.

While ETL is a generic process, the actual job can be very different in usage, utility, and complexity. Extract is the step where data from source locations is collected for transformation. Data extraction could be through: Batch processing, meaning source data, is moved in large chunks from the source to the target system at scheduled intervals. Tools for batch processing include Stitch and Blendo. Stream processing, which means source data is pulled in real-time from the source and transformed while it is in transit and before it is loaded into the data repository. Tools for stream processing include Apache Samza, Apache Storm, and Apache Kafka.

Transform involves the execution of rules and functions that converts raw data into data that can be used for analysis. For example, making date formats and units of measurement consistent across all sourced data, removing duplicate data, filtering out data that you do not need, enriching data, for example, splitting full name to first, middle, and last names, establishing key relationships across tables, applying business rules and data validations.

Load is the step where processed data is transported to a destination system or data repository. It could be: Initial loading, that is, populating all the data in the repository, Incremental loading, that is, applying ongoing updates and modifications as needed periodically; or Full refresh, that is, erasing contents of one or more tables and reloading with fresh data. Load verification, which includes data checks for missing or null values, server performance, and monitoring load failures, are important parts of this process step. It is vital to keep an eye on load failures and ensure the right recovery mechanisms are in place.

ETL has historically been used for batch workloads on a large scale. However, with the emergence of streaming ETL tools, they are increasingly being used for real-time streaming event data as well. It’s common to see the terms ETL and data pipelines used interchangeably. And although both move data from source to destination, data pipeline is a broader term that encompasses the entire journey of moving data from one system to another, of which ETL is a subset.

Data pipelines can be architected for batch processing, for streaming data, and a combination of batch and streaming data. In the case of streaming data, data processing or transformation, happens in a continuous flow. This is particularly useful for data that needs constant updating, such as data from a sensor monitoring traffic.

A data pipeline is a high performing system that supports both long-running batch queries and smaller interactive queries. The destination for a data pipeline is typically a data lake, although the data may also be loaded to different target destinations, such as another application or a visualization tool. There are a number of data pipeline solutions available, most popular among them being Apache Beam and DataFlow.

**VIEWPOINTS: CONSIDERATIONS FOR CHOICE OF DATA REPOSITORY**

In this video, we will listen to several data professionals talk about some of the factors they consider while deciding on the most appropriate data repository for their organizations. There's a number of factors to keep in mind while picking the right database for the job.

You need to look at the use case. What is the data repository going to be used for? Is it going to be used for storing structured information, semi- structured or unstructured information. Or do you know beforehand what the schema of the data is? Is there performance requirements? Are you working with data at rest, or streaming data, or data in motion? Does the data need to be encrypted? Does there... is there, you know, what's the volume of data that you're working with? Do you need a big data system? And what are the storage requirements? Does the data need to be updated frequently and accessed frequently, that it just needs to be stored and kept in a vault for a long time and is needed for backup purposes for example?

And then your organization might have certain standards that might have put in place of which databases or which data repositories you're allowed to use for different kinds of tasks. So all of these factors need to be kept in mind. So when we consider what data repository we want to choose, we look at these factors. We look at what are the kind of capacities that this data repository is supposed to handle. And then we also look at the type of access that we need this for. Do we access it in short intervals or do we run long running queries on it? Am I using it more for transaction processing or am I using it for analytics or archival purposes, or for data warehousing purpose? We also look for compatibility. How compatible this new data repository is with my existing ecosystem of programming languages, tools, and any processes that we have. We also consider the security features this repository gives us. And the most important thing is scalability. We may be happy with its performance today, but is it scalable enough? Can it scale along with the organization?

I don't often get to choose the type of data repository that my organization uses, and very few organizations use one data repository these days. On my team that I work on these days, we have a set of preferred solutions. We have a preferred enterprise relational database. We have a preferred open-source relational database for some of the smaller projects and for the microservices. And then we also have a preferred unstructured data source. So those are three main ones. The important thing is to think about the skills that you have within your organization or that you want to foster within your organization. And consider the costs of the various solutions. In our case, we have some experts on Db2, so our enterprise database of choice is Db2. However, there are other projects that use different ones. For open source, we've changed that a couple of times. We've got a couple of different directions with where we really want to be there.

And all of these... the hosting platform makes a difference as well, because now it's not just do I want to use IBM Db2 or do I want to use some other vendors, Microsoft SQL Server or whatever. It's not between those two choices. It's when I do those, do I want to do them on AWS RDS? Maybe I should consider Amazon's Aurora. Maybe I should consider Googles relational offerings. There's so many different choices there that you have to consider. There's the decision of how should the data be stored. There's the decision of how should the data be retrieved, and there's also the decision of where. Those are all very important questions when you're deciding on data storage.

I would say the structure of the data, the nature of the application, and the volume at which the data is getting ingested into your database, all these factors determine the nature of the data source that you should pick. In most cases a relational database should be enough, however, there will be edge cases where relational databases such as IBM Db2, Oracle or Postgres won't necessarily do the job for you. In those cases, so depending on the use case, for example, if you are ingesting gigabytes or terabytes of data per day. Then document stores such as MongoDB, or wide column stores such as Cassandra might be a good fit for you. At the same time, if you're trying to build a product recommendation engines or trying to show the network of relationships between different people on the social media, then graph data structures such as Neo4J or Apache TinkerPop would be an ideal fit for you. At the same time, if you are mining through terabytes or petabytes of data for analytics, Hadoop engine with MapReduce would be a good fit for you. So it really boils down to the nature of the application and the volume of the data, and the structure of the data, before you can pick the right database or data source whatever the use case.

**DATA INTEGRATION PLATFORMS**

Gartner defines data integration as a discipline comprising the practices, architectural techniques, and tools that allow organizations to ingest, transform, combine, and provision data across various data types. The report further explains that data integration has several usage scenarios, such as data consistency across applications, master data management, data sharing between enterprises, and data migration and consolidation.

In the field of analytics and data science, data integration includes accessing, queueing, or extracting data from operational systems transforming and merging extracted data either logically or physically data quality and governance, and delivering data through an integrated approach for analytics purposes.

For example, to make customer data available for analytics, you would need to extract individual customers' information from operational systems such as sales, marketing, and finance. You would then need to provide a unified view of the combined data so that your users can access, query, and manipulate this data from a single interface to derive statistics, analytics, and visualizations.

How does a data integration platform relate to ETL and data pipelines? While data integration combines disparate data into a unified view of the data, a data pipeline covers the entire data movement journey from source to destination systems. In that sense, you use a data pipeline to perform data integration, while ETL is a process within data integration. There is no one approach to data integration. However, modern data integration solutions typically support the following capabilities: An extensive catalog of pre-built connectors and adopters that help you connect and build integration flows with a wide variety of data sources such as databases, flat files, social media data, APIs, CRM and ERP applications. Open-source architecture that provides greater flexibility and avoids vendor lock-in. Optimization for both batch processing of large-scale data and continuous data streams, or both. Integration with Big Data sources. Support for big data is increasingly driving the decision regarding choice of integration platforms. Additional functionalities.

For example, specific demands around data quality and governance, compliance, and security. Portability, which ensures that as businesses move to cloud models, they should be able to run their data integration platforms anywhere. And data integration tools are able to work natively in a single cloud, multi-cloud, or hybrid cloud environment.

There are many data integration platforms and tools available in the market, ranging from commercial off-the-shelf tools to open-source frameworks. IBM offers a host of data integration tools targeting a range of enterprise integration scenarios, such as Information Server for IBM, Cloud Pak for Data, IBM Cloud Pak for Integration, IBM Data Replication, IBM Data Virtualization Manager, IBM InfoSphere Information Server on Cloud, and IBM InfoSphere DataStage all target a range of enterprise data integration scenarios.

Talend's data integration tools include Talend Data Fabric, Talend Cloud, Talend Data Catalog, Talend Data Management, Talend Big Data, Talend Data Services, and Talend Open Studio. SAP, Oracle, Denodo, SAS, Microsoft, Qlik, and TIBCO are some of the other vendors that offer data integration tools and platforms.

Examples of open-source frameworks include Dell Boomi, Jitterbit, and SnapLogic. There are a significant number of vendors who are offering cloud-based Integration Platform as a Service, or iPaaS, as a hosted service via virtual private cloud or hybrid cloud. Such as the Adeptia Integration Suite, Google Cloud's Cooperation 534, IBM's Application Integration Suite on Cloud, and Informatica's Integration Cloud. The data integration space continues to evolve as businesses embrace newer technologies and as data grows, be it in the variety of sources or its use in business decision-making.

**LESSON SUMMARY: WELCOME TO DATA LITERACY**

Welcome to the data literacy lesson summary. As a data scientist, you need an awareness of data storage possibilities for its organization and management, and options for retrieval. These systems enable you to find and analyze the data you need to make great discoveries hidden in that data.

In this video, we'll summarize what you learned in this lesson about the technologies and tools to handle large amounts of data. Let's consider data repositories. These repositories need the ability to find the data you want and return it to you in a usable format. Your data type helps determine the type of repository you need. You can store structured, semi structured or unstructured data. Depending on the organization, you may need a relational or no SQL database. For big data stores, your needs may call for a data warehouse, a data mart, or a data lake.

Relational databases store structured data. These are the oldest types of repositories. The most conventional and frequently used relational database management systems, often abbreviated as RDBMSs, are based on the foundational concept of structuring data in a tabular format with data arranged in rows and columns. Each table usually relates to a topic, and the columns of data in the table contain a specific type of information related to that topic. Then the database contains a defined schema that describes the table to each other. Relational databases usually rely on structured query language or SQL, to search for and retrieve the data you need. You use SQL to manipulate the data. Relational databases are beneficial for visualizing, analyzing, and finding connections between different pieces of data. You link tables together by creating schemas, you can restrict database fields to specific data types and values which minimizes irregularities and leads to greater consistency and data integrity. They offer easy export and import options, making back up and restoration easy. However, RBMS is do not work well with semi structured or unstructured data.

They are also slow to query with enormous datasets. Since RDBMSs use predefined structures for data to reside in, it becomes problematic when the data evolves and no longer conforms to that structure. Relational databases also limit field length, which means that sometimes they cannot accommodate the information you need. Because of these limitations and the quantities and diversity of data collected, many organizations have turned to not only SQL databases, or no SQ L for short.

Built for speed, flexibility, and scale, non relational databases allow storing data without stringent schemas. They can house semi structured and unstructured data. No SQL databases include document based, key value, columnar and graph. Document based databases store semi structured documents, such as Jason files. You group documents into collections, and each document has its structure. Key value stores each piece of data as a key value pair, so you retrieve and update the data using the key. Columnar databases store data and columns rather than rows, enabling storage of large volumes of data suitable for analytical workloads. Graph databases store data in nodes. Nodes have relationships and properties, and can manage and query complex relationships between them.

You can use technologies such as data warehouses, data marts, and data lakes for high volumes of data. A data warehouse works like a multipurpose storage for different use cases. The data has already been modeled and structured for a specific purpose. As an organization, you would opt for a data warehouse when you have a massive amount of data from your operational systems that must be readily available for reporting and analysis.

A data mart is a subsection of the data warehouse built specifically for a particular business function, purpose, or community of users. A data mart offers analytical capabilities for restricted data warehouse area, offering isolated security and performance.

A data lake is a storage repository that can store large amounts of structured, semi structured, and unstructured data in their native format, classified and tagged with meta data.

Let's review storage options. Data pipelines address an organization's need to collect, transform, and move data. Data pipelines have multiple steps, providing a systematic process to handle massive amounts of data as it is continually collected, processed, and made available.

ETL, which stands for extract, transform, and load, is a subset of a data pipeline, referring to an automated process where an organization converts its raw data into data ready for analysis.

Now as a future data scientist, you are aware of many technologies needed to handle big data before analysis can begin. These include data storage, organization and management and retrieval. Data storage options depend on the type of data, its volume, and how you intend to organize it. Using a data pipeline such as ETL, provides a process to manage and retrieve the data so you can analyze it as a data scientist.

**GLOSSARY: DATA LITERACY FOR DATA SCIENCE**

Uma imagem contendo Tabela

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

Uma imagem contendo Tabela

Descrição gerada automaticamente

**SUMMARY: DATA LITERACY FOR DATA SCIENCE**

Congratulations! You have completed this lesson. At this point in the course, you know:

The basics of data collection and organization methods.

What RDBMS is and its significance.

NoSQL databases and their flexible schema.

Types of data storage and the ways to process data.

The factors influencing data repository selection.

The various data integration tools and the solutions they provide.