Senga textlar va ohirida savollar beraman shunga javob berasa. savol bergunimcha textlarni o'rganib tur:

Textlarni o'zini alohida alohida qilib manga ber

Course 1 Module 2

**🔹 Text 1: 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.

**🔹 Text 2: 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.

**🔹 Text 3: Neural Networks and Deep Learning (Interview Story)**

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.

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.

Course 1 Module 3

**📄 Text 1: Measuring and Capturing Data**

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.

**📄 Text 2: 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.

**📄 Text 3: Data Science in Business and Consumer Use**

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 will 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, your 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 and 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? Lets 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 of 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 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.

**📄 Text 4: 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.

**📄 Text 5: The Final Deliverable**

The ultimate purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy. Analytics summarize findings in tables and plots. The data scientist should then use the insights to build the narrative to communicate the findings. In academia, the final deliverable is in the form of essays and reports. Such deliverables are usually 1,000 to 7,000 words in length. In consulting and business, the final deliverable takes on several forms. It can be a small document of fewer than 1,500 words illustrated with tables and plots, or it could be a comprehensive document comprising several hundred pages. Large consulting firms, such as McKinsey and Deloitte, routinely generate analytics-driven reports to communicate their findings and, in the process, establish their expertise in specific knowledge domains. Let's review the "United States Economic Forecast", a publication by the Deloitte University Press. This document serves as a good example for a deliverable that builds narrative from data and analytics. The 24-page report focuses on the state of the U.S. economy as observed in December 2014. The report opens with a grabber highlighting the fact that contrary to popular perception, the economic and job growth has been quite robust in the United States. The report is not merely a statement of facts. In fact, it is a carefully crafted report that cites Voltaire and follows a distinct theme. The report focuses on the good news about the U.S. economy. These include the increased investment in manufacturing equipment in the U.S. and the likelihood of higher consumer consumption resulting from lower oil prices. The Deloitte report uses time series plots to illustrate trends in markets. The GDP growth chart shows how the economy contracted during the Great Recession and has rebounded since then. The graphic presents four likely scenarios for the future. Another plot shows the changes in consumer spending. The accompanying narrative focuses on income inequality in the U.S. and refers to Thomas Pikkety's book on the same. The Deloitte report mentions many consumers did not experience an increase in their real incomes over the years, while they still maintained their level of spending. Other graphics focused on housing, business, and government sectors, international trade, labor, and financial markets, and prices. The appendix carries four tables documenting data for the four scenarios discussed in the report. Deloitte's "United States Economic Forecast" serves the very purpose that its authors intended. The report uses data and analytics to generate the likely economic scenarios. It builds a powerful narrative in support of the thesis statement that the U.S. economy is doing much better than 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.

**📄 Text 6: Lesson Summary Video Script**

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 of Key Terms**

| **Term** | **Definition** | **Video where the term is introduced** |
| --- | --- | --- |
| **Arithmetic Models** | Data science often uses mathematical models to analyze data and predict outcomes. | Old problems, new data science solutions |
| **Case Study** | In-depth analysis of an instance of a chosen subject to draw insights that inform theory or practice. | Old problems, new data science solutions |
| **Data Mining** | Extracting information from raw data, such as making decisions, predicting trends, or understanding it. | How Data Science is Saving Lives |
| **Data Science** | The field involves collecting, analyzing, and interpreting data to extract valuable insights. | Old problems, new data science solutions |
| **Data Strategy** | A plan that outlines how an organization will collect, manage, and use data to achieve its goals. | Old problems, new data science solutions |
| **Predictive Analytics** | Using data, algorithms, models, and machine learning to make predictions. | How Data Science is Saving Lives |

Module 4

**📘 Text 1 – What Makes a Real 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 programming 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.

**📘 Text 2 – Learning by Doing and Motivation**

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.

**📘 Text 3 – Where Does Data Science Fit in a Company?**

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.

**📘 Text 4 – Hiring a Data Scientist: What to Look For**

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.

**📘 Text 5 – Data Science: Job Market and Growth**

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.

**📘 Text 6 – Getting Started as a Young Data Scientist**

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.

**📘 Text 7 – 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 and illustrative graphics, you will move toward formally testing your hypothesis.

In case you need to run statistical models, you might turn to regression models or categorical analysis. If you are working with time-series data, you can analyze trends over time. You can also report results from other empirical techniques that fall under the general rubric of data mining. 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.

**📘 Text 8 – 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.

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