## Data Science Topics

In the first lesson in this module, you gain insight into the impact of big data on various aspects of society, from business operations to sports, and develop an understanding of key attributes and challenges associated with big data. You will learn about the big data fundamentals, how data scientists use the cloud to handle big data, and the data mining process. Lesson two delves into machine learning and deep learning and the relationship of artificial intelligence to data science.

**Learning Objectives**

* Define Big Data and its distinguishing characteristics, such as velocity, volume, veracity, and value
* Describe how Hadoop and other big data tools, combined with distributed computing power, are triggering digital transformation.
* List some of the skills required to be a data scientist and analyze big data.
* Describe the five essential cloud computing characteristics
* Explain what data mining is.
* Summarize the importance of establishing goals, data selection, preprocessing, transformation, and storage of data in preparation for data mining.
* Explain the difference between deep learning and machine learning.
* Describe regression and how it might be used to predict market behavior and trend analysis.
* Describe generative AI

# **Big Data and Data Mining**

## Lesson Overview: Big Data and Data Mining

In the 'Big Data and Data Mining' lesson, delve into the world of digital transformation driven by Big Data. Explore Cloud Computing's role, foundational Big Data concepts, tools like Hadoop and Spark, and gain insights into Data Mining techniques for informed decision-making.

|  |  |
| --- | --- |
| **Asset name and type** | **Description** |
| "How Big Data is Driving Digital Transformation" video | Explore the impact of Big Data on digital transformation. |
| "Introduction to Cloud" video | Get introduced to the fundamentals of cloud computing. |
| "Cloud for Data Science" video | Learn how the cloud is relevant in the field of data science. |
| "Foundations of Big Data" video | Build your understanding of key Big Data concepts. |
| "Data Scientists at New York University" video | Discover the work of data scientists at New York University. |
| "What is Hadoop?" video | Understand the significance of Hadoop in Big Data processing. |
| "Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark" video | Dive into the tools used for Big Data processing. |
| "Data Mining" reading | Read an excerpt on data mining explore the principles and concepts of data mining. |
| Practice quiz | Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson. |
| Glossary | Use this glossary of terms to review the terminology presented in this lesson. |
| Graded quiz | Test your knowledge from this lesson by taking the graded quiz. |

## **How Big Data is Driving Digital Transformation**

Digital Transformation: A Simple Explanation

Digital transformation is like giving a business a makeover using technology. It means changing how a company operates by using new digital tools and data to improve its processes and better serve its customers. Imagine a traditional bookstore that starts selling books online and uses customer data to recommend books based on what people like. This shift not only helps the bookstore reach more customers but also makes the shopping experience more personalized and enjoyable.

For example, the Houston Rockets, a basketball team, used data from video tracking to discover that certain types of shots were more successful. By analyzing this data, they changed their game strategy, leading to more wins. This shows how using data can transform not just one organization but an entire industry, making it more efficient and effective.

**Summary of Digital Transformation:**

* **Definition**: Digital transformation integrates digital technology into all areas of a business, leading to fundamental changes in how it operates and delivers value to customers.
* **Impact of Data**: The rise of Big Data has driven digital transformations across various industries, allowing organizations to analyze vast amounts of data for competitive advantage.
* **Example**: The Houston Rockets NBA team used video tracking data to identify the most effective plays, changing their strategy and leading to more wins.
* **Organizational Change**: Successful digital transformation requires support from top executives and a shift in organizational culture, affecting data management, employee roles, and customer interactions.

## **Introduction to Cloud**

**cloud computing** in simple terms.

Cloud computing is like using electricity from a power grid instead of having your own generator at home. Instead of buying and installing software or hardware on your computer, you can access everything you need over the Internet. This means you can use applications, store files, and even run programs without having to worry about where they are physically located. For example, when you use Google Drive to save your photos, you're using cloud computing because those files are stored on the Internet, not just on your computer.

Imagine you want to watch a movie. Instead of buying a DVD and keeping it at home, you can stream it online whenever you want. This is similar to how cloud computing works. You pay for what you use, and you can access it from anywhere, whether you're on your phone, tablet, or laptop. Plus, you don’t have to worry about updates or storage space because everything is managed for you in the cloud!

summary of **cloud computing**:

* **Definition**: Cloud computing is the delivery of computing resources (like servers, storage, applications) over the Internet on a pay-for-use basis.
* **Key Characteristics**:
  + **On-demand self-service**: Access resources without human interaction.
  + **Broad network access**: Available through various devices (phones, tablets, etc.).
  + **Resource pooling**: Multiple users share resources, making it cost-effective.
  + **Rapid elasticity**: Easily scale resources up or down as needed.
  + **Measured service**: Pay only for what you use.
* **Deployment Models**:
  + **Public Cloud**: Services shared over the Internet.
  + **Private Cloud**: Exclusive use by a single organization.
  + **Hybrid Cloud**: Combination of public and private clouds.
* **Service Models**:
  + **IaaS (Infrastructure as a Service)**: Access to physical computing resources.
  + **PaaS (Platform as a Service)**: Tools for developing and deploying applications.
  + **SaaS (Software as a Service)**: Software hosted online and accessed via subscription.

This technology has transformed how we use computing resources, making them more accessible and cost-efficient.

## **Cloud for Data Science**

"Cloud Computing" in simple terms.

Cloud Computing: A Simple Explanation  
Cloud computing is like having a magical storage box in the sky where you can keep all your important data and access it anytime, anywhere. Instead of relying on your own computer's memory, which can be limited, you can store your files, photos, and even run complex programs on powerful computers that are located far away. This means you can do things like analyze large amounts of data without needing a super-fast computer at home.

Imagine you have a big puzzle, but you don't have enough space on your table to work on it. Instead, you send the puzzle pieces to a friend who has a huge table. While your friend works on the puzzle, you can also look at it and help from your own home. That's how cloud computing works! It allows many people to work together on the same data from different places, making teamwork easier and faster.

summary of the key points about Cloud Computing from the course material:

* **Central Storage**: Cloud computing allows data to be stored in a central location, enabling access from anywhere, bypassing physical limitations of personal computers.
* **Advanced Computing**: Users can deploy complex algorithms and perform high-performance computing without needing powerful local machines.
* **Collaboration**: Multiple teams can work on the same data simultaneously, facilitating collaboration across different locations.
* **Access to Tools**: The Cloud provides instant access to open-source technologies and up-to-date tools without the need for local installation.
* **Examples of Cloud Platforms**: Major companies like IBM, Amazon (AWS), and Google offer cloud services, providing environments for data science projects.

## **Foundations of Big Data**

Big Data refers to the enormous amounts of information generated every day from various sources like our smartphones, social media, and even smartwatches. Imagine a giant library filled with millions of books, but instead of books, it's filled with data about our daily activities, preferences, and behaviors. This data is so vast and diverse that it requires special tools and technologies to collect, store, and analyze it effectively.

To make sense of Big Data, we often talk about the "V's" of Big Data: velocity (how fast data is created), volume (the sheer amount of data), variety (the different types of data), veracity (the accuracy of the data), and value (the insights we can gain from it). For example, every minute, countless videos are uploaded to platforms like YouTube, creating a massive flow of data (velocity). With billions of people using digital devices, we generate about 2.5 quintillion bytes of data daily (volume). Understanding these concepts helps businesses and organizations make better decisions and improve their services.

**Summary of Big Data:**

* **Definition**: Big Data refers to the large and diverse volumes of data generated by people, tools, and machines that require innovative technology for collection and analysis.
* **Key Elements (The V's)**:
  + **Velocity**: The speed at which data is generated and processed.
  + **Volume**: The vast amount of data created daily, estimated at 2.5 quintillion bytes.
  + **Variety**: The different types of data, including structured (organized) and unstructured (like social media posts).
  + **Veracity**: The quality and accuracy of the data.
  + **Value**: The insights and benefits derived from analyzing the data.
* **Examples**:
  + **Velocity**: Rapid data accumulation from platforms like YouTube.
  + **Volume**: Billions of digital devices generating data daily.

Understanding Big Data helps organizations make informed decisions and enhance their services.

## **Data Science and Big Data**

Big data refers to extremely large sets of data that are so vast and complex that traditional methods of processing and analyzing them just can't keep up. Imagine trying to fill a small bucket with water from a fire hose. The bucket can’t hold all that water, just like traditional databases can’t handle the massive amounts of information generated today. Companies like Google had to come up with new ways to store and analyze this data, leading to the development of technologies like Hadoop, which helps manage and make sense of big data.

To illustrate, think of big data like a huge library filled with millions of books. If you wanted to find a specific book using a simple card catalog, it would take forever! But with advanced search tools, you can quickly locate the book you need. Similarly, big data technologies allow businesses to sift through vast amounts of information quickly and efficiently, helping them make better decisions based on insights drawn from the data.

summary of the topic discussed:

* **Big Data**: Refers to large and complex data sets that traditional database systems cannot handle effectively.
* **Importance**: Companies need to analyze big data to gain insights and make informed decisions.
* **Technological Development**: Innovations like Hadoop were created to store and analyze big data efficiently.
* **Analogy**: Think of big data like a huge library; advanced tools help quickly find information, just as big data technologies help businesses analyze vast amounts of data.

## **What is Hadoop?**

Big Data Clusters are like a team of computers working together to handle huge amounts of information. Imagine you have a giant puzzle, and instead of trying to solve it all by yourself, you invite many friends to help. Each friend takes a small section of the puzzle, works on it, and then shares their progress with the group. This way, the puzzle gets solved much faster! In the same way, big data clusters break down large data sets into smaller pieces, distribute them to many computers, and then gather the results to make sense of the entire data set.

This method is efficient because if you add more computers to the team, you can handle even more data without slowing down. It's like having more friends to help with the puzzle; the more, the merrier! This approach has been a game-changer for companies that deal with a lot of data, like social media platforms, allowing them to analyze and understand trends quickly.

summary of the key points about Big Data Clusters:

* **Definition**: Big Data Clusters are groups of computers that work together to process and analyze large data sets efficiently.
* **Process**: Data is divided into smaller pieces, distributed to multiple computers, which run the same program on their respective pieces, and then send the results back for aggregation.
* **Scalability**: Adding more computers increases performance and allows for handling larger data sets, making the system scalable.
* **Impact**: This approach has transformed how companies, especially in social media, analyze vast amounts of data quickly and effectively.

## [**Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark**](https://www.coursera.org/learn/what-is-datascience/lecture/VULo5/big-data-processing-tools-hadoop-hdfs-hive-and-spark)

Hadoop is like a big library that helps us store and manage huge amounts of information. Imagine you have a giant book that is too big to fit on one shelf. Instead of trying to cram it all into one place, Hadoop spreads the book across many shelves (which we call nodes) in a library (the cluster). Each shelf holds a part of the book, and when you want to read it, Hadoop knows how to gather all the pieces together quickly. This way, even if one shelf is busy or has a problem, you can still access the rest of the book without any hassle!

In addition to storing data, Hadoop also helps us work with different types of information, like videos, social media posts, and traditional data. It’s designed to be flexible and can grow as needed, just like adding more shelves to a library when you get more books. This makes it a powerful tool for businesses that want to make sense of large amounts of data.

summary of the key points about Hadoop from the course material:

* **Hadoop** is an open-source framework for storing and processing large datasets across clusters of computers.
* It allows for **distributed storage** and processing, meaning data is spread out over multiple computers (nodes) for efficiency.
* **Hadoop Distributed File System (HDFS)** is a key component that stores data in a fault-tolerant way by splitting files into smaller pieces and replicating them across different nodes.
* Benefits of using Hadoop include:
  + **Scalability**: It can grow from a single computer to many nodes.
  + **Cost-effectiveness**: It uses commodity hardware.
  + **Flexibility**: It can handle various data formats (structured, semi-structured, and unstructured).
  + **Data locality**: It processes data where it is stored, reducing network congestion.
* Hadoop is suitable for tasks like data warehousing, ETL (Extract, Transform, Load), and reporting.

## **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**

In our lesson, we learned about big data, which refers to the enormous amounts of information generated every day from various sources like social media, sensors, and online transactions. Imagine a giant library filled with countless books, but instead of just books, it's filled with data from all over the world! This data can help businesses make better decisions, improve services, and even predict trends.

To make sense of this vast amount of data, we use a process called data mining. Think of it like digging for treasure in a huge pile of sand. We set goals, choose the right data, clean it up, and then analyze it to find valuable insights. Just like a treasure hunter shares their findings, data scientists share their results with others to help organizations grow and succeed.

summary of the key points from the lesson on big data and data mining:

* **Big Data**: Refers to vast amounts of data generated from various sources, impacting business and daily life.
* **Five Characteristics**:
  + **Value**: The usefulness of the data.
  + **Volume**: The scale of data collected.
  + **Velocity**: The speed at which data is generated.
  + **Variety**: The different types of data (structured and unstructured).
  + **Veracity**: The quality and accuracy of the data.
* **Cloud Computing**: Provides on-demand access to computing resources, making it easier to handle big data. Key features include elasticity, resource pooling, and measured service.
* **Data Mining Process**: A six-step process that includes:
  + **Goal Setting**: Identify key questions.
  + **Selecting Data Sources**: Choose where to gather data.
  + **Preprocessing**: Clean and prepare the data.
  + **Transforming**: Format the data for analysis.
  + **Mining**: Analyze the data using algorithms.
  + **Evaluation**: Assess the results and share findings.

This process helps organizations derive insights and make informed decisions based on data.

## **Glossary: Big Data and Data Mining**

|  |  |  |
| --- | --- | --- |
| **Term** | **Definition** | **Video where the term is introduced** |
| Analytics | The process of examining data to draw conclusions and make informed decisions is a fundamental aspect of data science, involving statistical analysis and data-driven insights. | Data Scientists at New York University |
| Big Data | Vast amounts of structured, semi-structured, and unstructured data are characterized by its volume, velocity, variety, and value, which, when analyzed, can provide competitive advantages and drive digital transformations. | How Big Data is Driving Digital Transformation |
| Big Data Cluster | A distributed computing environment comprising thousands or tens of thousands of interconnected computers that collectively store and process large datasets. | What is Hadoop? |
| Broad Network Access | The ability to access cloud resources via standard mechanisms and platforms such as mobile devices, laptops, and workstations over networks. | Introduction to Cloud |
| Chief Data Officer (CDO) | An emerging role responsible for overseeing data-related initiatives, governance, and strategies, ensuring that data plays a central role in digital transformation efforts. | How Big Data is Driving Digital Transformation |
| Chief Information Officer (CIO) | An executive is responsible for managing an organization's information technology and computer systems, contributing to technology-related aspects of digital transformation. | How Big Data is Driving Digital Transformation |
| Cloud Computing | The delivery of on-demand computing resources, including networks, servers, storage, applications, services, and data centers, over the Internet on a pay-for-use basis. | Introduction to Cloud |
| Cloud Deployment Models | Categories that indicate where cloud infrastructure resides, who manages it, and how cloud resources and services are made available to users, including public, private, and hybrid models. | Introduction to Cloud |
| Cloud Service Models | Models based on the layers of a computing stack, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), represent different cloud computing offerings. | Introduction to Cloud |
| Commodity Hardware | Standard, off-the-shelf hardware components are used in a big data cluster, offering cost-effective solutions for storage and processing without relying on specialized hardware. | What is Hadoop? |
| Data Algorithms | Computational procedures and mathematical models used to process and analyze data made accessible in the cloud for data scientists to deploy on large datasets efficiently. | Cloud for Data Science |
| Data Replication | A strategy in which data is duplicated across multiple nodes in a cluster to ensure data durability and availability, reducing the risk of data loss due to hardware failures. | What is Hadoop? |
| Data Science | An interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including programming, statistics, and analytical tools. | Data Scientists at New York University |
| Deep Learning | A subset of machine learning that involves artificial neural networks inspired by the human brain, capable of learning and making complex decisions from data on their own. | Data Scientists at New York University |
| Digital Change | The integration of digital technology into business processes and operations leads to improvements and innovations in how organizations operate and deliver value to customers. | How Big Data is Driving Digital Transformation |
| Digital Transformation | A strategic and cultural organizational change driven by data science, especially Big Data, to integrate digital technology across all areas of the organization, resulting in fundamental operational and value delivery changes. | How Big Data is Driving Digital Transformation |
| Distributed Data | The practice of dividing data into smaller chunks and distributing them across multiple computers within a cluster enables parallel processing for data analysis. | What is Hadoop? |
| Hadoop | A distributed storage and processing framework used for handling and analyzing large datasets, particularly well-suited for big data analytics and data science applications. | Data Scientists at New York University |
| Hadoop Distributed File System (HDFS) | A storage system within the Hadoop framework that partitions and distributes files across multiple nodes, facilitating parallel data access and fault tolerance. | What is Hadoop? |
| Infrastructure as a Service (IaaS) | A cloud service model that provides access to computing infrastructure, including servers, storage, and networking, without the need for users to manage or operate them. | Introduction to Cloud |
| Java-Based Framework | Hadoop is implemented in Java, an open-source, high-level programming language, providing the foundation for building distributed storage and processing solutions. | Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark |
| Map Process | The initial step in Hadoop’s MapReduce programming model, where data is processed in parallel on individual cluster nodes, often used for data transformation tasks. | What is Hadoop? |
| Measured Service | A characteristic where users are billed for cloud resources based on their actual usage, with resource utilization transparently monitored, measured, and reported. | Introduction to Cloud |
| On-Demand Self-Service | The capability for users to access and provision cloud resources such as processing power, storage, and networking using simple interfaces without human interaction with service providers. | Introduction to Cloud |
| Rapid Elasticity | The ability to quickly scale cloud resources up or down based on demand, allowing users to access more resources when needed and release them when not in use. | Introduction to Cloud |
| Reduce Process | The second step in Hadoop's MapReduce model is where results from the mapping process are aggregated and processed further to produce the final output, typically used for analysis. | What is Hadoop? |
| Replication | The act of creating copies of data pieces within a big data cluster enhances fault tolerance and ensures data availability in case of hardware or node failures. | What is Hadoop? |
| Resource Pooling | A cloud characteristic where computing resources are shared and dynamically assigned to multiple consumers, promoting economies of scale and cost-efficiency. | Introduction to Cloud |
| Skills Network Labs (SN Labs) | Learning resources provided by IBM, including tools like Jupyter Notebooks and Spark clusters, are available to learners for cloud data science projects and skill development. | Cloud for Data Science |
| Spilling to Disk | A technique used in memory-constrained situations where data is temporarily written to disk storage when memory resources are exhausted, ensuring uninterrupted processing. | Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark |
| STEM Classes | Science, Technology, Engineering, and Mathematics (STEM) courses typically taught in high schools prepare students for technical careers, including data science. | Data Scientists at New York University |
| Variety | The diversity of data types, including structured and unstructured data from various sources such as text, images, video, and more, posing data management challenges. | Foundations of Big Data |
| Velocity | The speed at which data accumulates and is generated, often in real-time or near-real-time, drives the need for rapid data processing and analytics. | Foundations of Big Data |
| Veracity | The quality and accuracy of data, ensuring that it conforms to facts and is consistent, complete, and free from ambiguity, impacts data reliability and trustworthiness. | Foundations of Big Data |
| Video Tracking System | A system used to capture and analyze video data from games, enabling in-depth analysis of player movements and game dynamics, contributing to data-driven decision-making in sports. | How Big Data is Driving Digital Transformation |
| Volume | The scale of data generated and stored is driven by increased data sources, higher-resolution sensors, and scalable infrastructure. | Foundations of Big Data |
| V's of Big Data | A set of characteristics common across Big Data definitions, including Velocity, Volume, Variety, Veracity, and Value, highlighting the rapid generation, scale, diversity, quality, and value of data. | Foundations of Big Data |

# **Deep Learning and Machine Learning**

## **Lesson Overview: Deep Learning and Machine Learning**

In this lesson, "Deep Learning and Machine Learning," you'll dive into the exciting concepts of artificial intelligence and data science. Throughout this module, you will explore various machine learning and deep learning aspects, gaining valuable insights and skills.

|  |  |
| --- | --- |
| **Asset name and type** | **Description** |
| “Artificial Intelligence and Data Science” video | Get introduced to the captivating field of artificial intelligence and its role in data science. |
| “Generative AI and Data Science” video | Discover the exciting realm of generative artificial intelligence and its applications in data science. |
| “Neural Networks and Deep Learning” video | Explore the fundamentals of neural networks and delve into the depths of deep learning. |
| “Applications of Machine Learning” video | Uncover the real-world applications of machine learning and its impact on various industries. |
| “Regression” reading | Learn about regression analysis, a fundamental statistical technique used in machine learning. |
| Practice quiz | Test your understanding of the previous reading. |
| “Exploring Data using IBM Cloud Gallery” lab | Engage in hands-on exploration of data using the IBM Cloud Gallery, gaining practical experience in data analysis. |
| “Lesson Summary” video | Sum up your learning from this module with a concise lesson summary. |
| Practice quiz | Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson. |
| Glossary | Use this glossary of terms to review the terminology presented in this lesson. |
| Graded quiz | Test your knowledge from this lesson by taking the graded quiz. |

## **Artificial Intelligence and Data Science**

1. **Big Data**: This is a term for very large sets of data that are too big, fast, or varied for traditional methods to analyze. Think of it as a huge amount of information that needs special tools to understand.
2. **Data Mining**: This is the process of searching through data to find patterns or useful information. It’s like digging for treasure in a big pile of sand to find valuable gems.
3. **Machine Learning**: This is a part of artificial intelligence (AI) where computers learn from data and make decisions without being specifically programmed. Imagine teaching a child to recognize animals by showing them many pictures; they learn to identify animals on their own.
4. **Deep Learning**: This is a more advanced type of machine learning that uses layered networks (like a brain) to analyze data. It helps computers understand complex patterns, similar to how humans learn from experience.
5. **Neural Networks**: These are computer systems inspired by the human brain. They consist of small units (neurons) that work together to process information and learn over time.
6. **Data Science**: This is the field that combines various methods (like statistics and machine learning) to extract knowledge from large amounts of data. It helps organizations make informed decisions based on data analysis.
7. **AI vs. Data Science**: While data science focuses on extracting insights from data, AI is about creating systems that can learn and make decisions. They are related but not the same.

## **Generative AI and Data Science**

**Generative AI** in simple terms.

Generative AI is like a creative artist that can produce new things instead of just looking at what already exists. Imagine a painter who can create beautiful paintings from scratch, or a musician who can compose original songs. In the world of technology, generative AI uses deep learning models to learn from a lot of data and then create new content, such as images, music, or even text that sounds like it was written by a human. It’s like having a smart assistant that can help you come up with fresh ideas!

For example, think of a chef who has learned recipes from all over the world. Instead of just following a recipe, this chef can invent a brand-new dish by mixing different ingredients in unique ways. Similarly, generative AI takes patterns from existing data and combines them to create something new and exciting. This technology is used in various fields, from creating art and music to helping doctors analyze medical images.

summary of **Generative AI**:

* **Definition**: Generative AI is a type of artificial intelligence that creates new data instead of just analyzing existing data.
* **How it Works**: It uses deep learning models, like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to learn patterns from large datasets and generate new content.
* **Applications**:
  + **Natural Language Processing**: Tools like OpenAI’s GPT-3 can produce human-like text.
  + **Healthcare**: It can synthesize medical images for training purposes.
  + **Art and Design**: Generates unique artworks and fashion designs.
  + **Gaming**: Creates realistic environments and characters.
* **Benefits for Data Scientists**: Helps in generating synthetic data to augment datasets, automates coding tasks, and uncovers hidden insights in data.

## **Neural Networks and Deep Learning**

neural networks and deep learning in simple terms.

Neural networks are like a computer's way of trying to think and learn, similar to how our brains work. Imagine your brain as a big network of tiny lights (neurons) that light up when you learn something new. In a neural network, we have inputs (like pictures or sounds) that go into these lights, which then process the information and produce an output (like recognizing a face or understanding a word). The more you practice, the better the network gets at making the right connections, just like how you get better at a game the more you play it!

Now, deep learning is like taking that neural network and making it even more powerful by adding more layers of these lights. Think of it as building a multi-story building where each floor (layer) helps the building learn more complex things. For example, a deep learning system can learn to tell the difference between a cat and a dog just by looking at lots of pictures, without needing someone to teach it every detail. It learns on its own, which is why it's called "deep learning."

summary of neural networks and deep learning:

* **Neural Networks**: These are computer systems designed to mimic how our brains work. They take inputs (like images or sounds), process them through interconnected nodes (like neurons), and produce outputs (like recognizing a face). They improve through practice, similar to how we learn.
* **Deep Learning**: This is an advanced form of neural networks that uses multiple layers of nodes. Each layer helps the system learn more complex patterns. For example, deep learning can distinguish between different animals in images without explicit instructions, as it learns autonomously.

## **Applications of Machine Learning**

**recommender systems** in machine learning.

Recommender systems are like your personal shopping assistant. Imagine you walk into a store, and the assistant says, "Since you liked this shirt, you might also like these pants!" They use your past choices to suggest new items that you might enjoy. In the world of technology, platforms like Netflix and Facebook do the same thing. They analyze what you've watched or liked and recommend similar shows or friends to follow. This helps you discover new content that matches your interests.

In finance, recommender systems can suggest investment opportunities based on your previous choices. For example, if you've shown interest in a particular tech company, the system might recommend another tech company that has similar characteristics. This way, you can explore options that you might not have considered otherwise!

summary of **recommender systems** in machine learning:

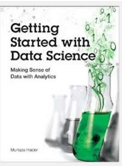
* **Definition**: Recommender systems analyze user behavior to suggest products, services, or content that align with individual preferences.
* **Applications**:
  + **Entertainment**: Platforms like Netflix recommend shows based on your viewing history.
  + **Social Media**: Facebook suggests friends based on your interactions.
  + **Finance**: In fintech, these systems recommend investment opportunities based on your past choices.

Recommender systems enhance user experience by helping individuals discover new options tailored to their interests.

## **Regression**

**Course Text Book: 'Getting Started with Data Science' Publisher: IBM Press; 1 edition (Dec 13 2015) Print.**

**Author: Murtaza Haider**



Prescribed Reading: Chapter 7 Pg. 235-236

**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?

## **Lesson Summary: Deep Learning and Machine Learning**

**Machine Learning** in simple terms.

Machine Learning is like teaching a computer to learn from experience, just like how we learn from our own experiences. Imagine you have a pet dog. When you train your dog to sit, you give it treats when it does the right thing. Over time, the dog learns to sit on command because it remembers that sitting gets it a reward. Similarly, in machine learning, we feed a computer lots of data (like pictures of dogs and cats) and let it learn patterns from that data. The computer then uses what it learned to make predictions or decisions, like identifying whether a new picture is of a dog or a cat.

To illustrate, think of a machine learning algorithm as a chef learning to bake a cake. At first, the chef might make mistakes, but with each attempt and feedback (like tasting the cake), the chef improves and eventually bakes a delicious cake. In the same way, machine learning algorithms improve their predictions as they process more data and receive feedback.

Is there any other concept within the content that you would like me to explain?

**A**

**Abhishek**

Give me a summary

**Coach**

Here's a brief summary of the key concepts related to **Deep Learning and Machine Learning**:

* **Artificial Intelligence (AI)**: A branch of computer science that develops systems to mimic human intelligence tasks.
* **Machine Learning**: A subset of AI that uses algorithms to learn from data and make predictions without being explicitly programmed.
* **Deep Learning**: A further subset of machine learning that uses layered neural networks to simulate human decision-making. It becomes more effective with larger amounts of data.
* **Neural Networks**: Collections of small computing units (neurons) that learn to make decisions over time, such as distinguishing between different animals.
* **Generative AI**: Focuses on creating new data (like images or music) rather than just analyzing existing data.
* **Regression**: A statistical technique used in machine learning to identify relationships between inputs and outputs, helping to make predictions (e.g., predicting house prices based on size and number of bedrooms).

This summary encapsulates the main ideas discussed in the lesson.

## **Deep Learning and Machine Learning Lesson Glossary**

|  |  |  |
| --- | --- | --- |
| **Term** | **Definition** | **Video where the term is introduced** |
| Artificial Neural Networks | Collections of small computing units (neurons) that process data and learn to make decisions over time. | Artificial Intelligence and Data Science |
| Bayesian Analysis | A statistical technique that uses Bayes' theorem to update probabilities based on new evidence. | Applications of Machine Learning |
| Business Insights | Accurate insights and reports generated by generative AI can be updated as data evolves, enhancing decision-making and uncovering hidden patterns. | Generative AI and Data Science |
| Cluster Analysis | The process of grouping similar data points together based on certain features or attributes. | Neural Networks and Deep Learning |
| Coding Automation | Using generative AI to automatically generate and test software code for constructing analytical models, freeing data scientists to focus on higher-level tasks. | Generative AI and Data Science |
| Data Mining | The process of automatically searching and analyzing data to discover patterns and insights that were previously unknown. | Artificial Intelligence and Data Science |
| Decision Trees | A type of machine learning algorithm used for decision-making by creating a tree-like structure of decisions. | Applications of Machine Learning |
| Deep Learning Models | Includes Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) that create new data instances by learning patterns from large datasets. | Generative AI and Data Science |
| Five V's of Big Data | Characteristics used to describe big data: Velocity, volume, variety, veracity, and value. | Neural Networks and Deep Learning |
| Generative AI | A subset of AI that focuses on creating new data, such as images, music, text, or code, rather than just analyzing existing data. | Generative AI and Data Science |
| Market Basket Analysis | Analyzing which goods tend to be bought together is often used for marketing insights. | Neural Networks and Deep Learning |
| Naive Bayes | A simple probabilistic classification algorithm based on Bayes' theorem. | Applications of Machine Learning |
| Natural Language Processing (NLP) | A field of AI that enables machines to understand, generate, and interact with human language, revolutionizing content creation and chatbots. | Generative AI and Data Science |
| Precision vs. Recall | Metrics are used to evaluate the performance of classification models. | Applications of Machine Learning |
| Predictive Analytics | Using machine learning techniques to predict future outcomes or events. | Neural Networks and Deep Learning |
| Synthetic Data | Artificially generated data with properties similar to real data, used by data scientists to augment their datasets and improve model training. | Generative AI and Data Science |

## **Deep Learning and Machine Learning Lesson Glossary**

Welcome! This alphabetized glossary contains many of the terms in this course. These terms are important for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

|  |  |  |
| --- | --- | --- |
| **Term** | **Definition** | **Video where the term is introduced** |
| Artificial Neural Networks | Collections of small computing units (neurons) that process data and learn to make decisions over time. | Artificial Intelligence and Data Science |
| Bayesian Analysis | A statistical technique that uses Bayes' theorem to update probabilities based on new evidence. | Applications of Machine Learning |
| Business Insights | Accurate insights and reports generated by generative AI can be updated as data evolves, enhancing decision-making and uncovering hidden patterns. | Generative AI and Data Science |
| Cluster Analysis | The process of grouping similar data points together based on certain features or attributes. | Neural Networks and Deep Learning |
| Coding Automation | Using generative AI to automatically generate and test software code for constructing analytical models, freeing data scientists to focus on higher-level tasks. | Generative AI and Data Science |
| Data Mining | The process of automatically searching and analyzing data to discover patterns and insights that were previously unknown. | Artificial Intelligence and Data Science |
| Decision Trees | A type of machine learning algorithm used for decision-making by creating a tree-like structure of decisions. | Applications of Machine Learning |
| Deep Learning Models | Includes Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) that create new data instances by learning patterns from large datasets. | Generative AI and Data Science |
| Five V's of Big Data | Characteristics used to describe big data: Velocity, volume, variety, veracity, and value. | Neural Networks and Deep Learning |
| Generative AI | A subset of AI that focuses on creating new data, such as images, music, text, or code, rather than just analyzing existing data. | Generative AI and Data Science |
| Market Basket Analysis | Analyzing which goods tend to be bought together is often used for marketing insights. | Neural Networks and Deep Learning |
| Naive Bayes | A simple probabilistic classification algorithm based on Bayes' theorem. | Applications of Machine Learning |
| Natural Language Processing (NLP) | A field of AI that enables machines to understand, generate, and interact with human language, revolutionizing content creation and chatbots. | Generative AI and Data Science |
| Precision vs. Recall | Metrics are used to evaluate the performance of classification models. | Applications of Machine Learning |
| Predictive Analytics | Using machine learning techniques to predict future outcomes or events. | Neural Networks and Deep Learning |
| Synthetic Data | Artificially generated data with properties similar to real data, used by data scientists to augment their datasets and improve model training. | Generative AI and Data Science |

## **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.
* Skills involved in processing Big Data include the application of statistics, machine learning models, and some computer programming.
* Generative AI, a subset of artificial intelligence, 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.