Introduction to Data Science Specialization

What I'll learn

1. Describe what data science and machine learning are, their applications & use cases, and various types of tasks performed by data scientists
2. Gain hands-on familiarity with everyday data science tools, including JupyterLab, R Studio, GitHub and Watson Studio.
3. Develop the mindset to work like a data scientist and follow a methodology to tackle different data science problems.
4. Write SQL statements and query Cloud databases using Python from Jupyter notebooks.

Specialisation Link:-https://www.coursera.org/specializations/introduction-data-science

Skills I will gain:-

1. Data Analysis
2. Machine Learning
3. Python Programming

Course Link:- https://www.coursera.org/learn/what-is-datascience

- Also, a Free Career Support
  Ex. Resume Building from Alumni

Course-1

What is Data Science?

What You Will Learn

1. Define data science and its importance in today's data-driven world.
2. Describe the various paths that can lead to a career in data science.
3. Summarize advice seasoned data science professionals give to data scientists starting out.
4. Explain why data science is considered the most in-demand job in the 21st century.

Modules

1. Defining Data Science and What Data Scientists Do
2. Data Science Topics
3. Applications and Careers in Data Science
4. Data Literacy for Data Science (Optional)

Module-1 Defining Data Science and What Data Scientists Do

Topics Covered

1.1 Defining Data Science

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

1.2 What do Data Scientists Do?

- A Day in the Life of a Data Scientist
- Data Science Skills & Big Data
- Working on Different File Formats

- Data Science Topics and Algorithms
- Discussion Prompt: Introduce Yourself
- Reading: What Makes Someone a Data Scientist?

1.1.1 What is Data Science?

Definitions
“Term Data Science is Coined by DJ Patil and Marjorie Lee Brown”

“ Data science is the art of uncovering the insights and trends hiding behind data.**”**

“Study of Data”

“Data science is a field about processes and systems that extract data from various forms, whether in unstructured or structured form.”

1.1.2 Fundamentals of Data Science

- ❖ Data science can help organisations **understand their environments, analyse existing issues, and reveal previously hidden opportunities.**
- ❖ Data scientists use data analysis to add to the organisation's knowledge by investigating data and exploring the best way to use it to provide value to the business.
- ❖ Many organisations will use data science to focus on a specific problem, so it's essential to clarify the question the organisation wants answered.
- ❖ **Data scientists can analyse structured and unstructured data from many sources, and depending on the nature of the problem**, they can choose to analyse the data in different ways.
- ❖ The role of the data scientist becomes that of a storyteller, communicating the results to the project stakeholders.
- ❖ Data scientists can **use powerful visualisation tools** to help the project stakeholders.

Day-2 Jul 14, 2024

1.1.3 The Many Paths to Data Science

“Speakers shared their experience of landing in their career in Data Science”.Most of them have not purposefully taken Data Science.

1.1.4 Data Science: The Sexiest Job in the 21st Century
Link:- https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century

Course Text Book: 'Getting Started with Data Science' Publisher: IBM Press; 1 edition (Dec13 2015) Print.

Author: Murtaza Haider

Prescribed Reading: Chapter 1 Pg. 4

“ The report by the **McKinsey Global Institute** warns of massive talent **shortages for data and analytics**. By 2018, the United States could face a shortage of 140,000 to 190,000 people with deep analytical skills and 1.5 million managers and analysts with the know-how to analyse big data to make effective decisions.”

“Bernard reported on **Walmart**, which turned to **crowd-sourcing for its analytics needs**. Walmart approached **Kaggle** to host a **competition to analyse its proprietary data**. The retailer provided sales data from a shortlist of stores and asked the competitors **to develop better forecasts of sales based on promotion schemes.”**

“SAP, a leader in data and analytics, reported from a survey that **92% of the responding firms in its sample experienced a significant increase in their data holdings**. At the same time, **three-quarters** identified the need for new **data science skills** in their firms.”
“Accenture believes that the demand for data scientists **may outstrip supply by 250,000 in 2015 alone.** A similar survey of 150 executives by KPMG in 2014 found that 8**5% of the respondents** did not know **how to analyse data**.

• The New York Times reported **$100,000** as the average base salary of a **software engineer a**nd **$112,000** for data scientists.

1.1.5 Advice for New Data Scientists

Be Argumentative, Judgmental, and Curiosity

If you want to be a data scientist and **work for an IT firm or a web-based or Internet-based firm**, you need different **skills**.

For example, in the health industry, you need different skill sets. So figure out first what you're interested in and **your competitive advantage**. Your competitive advantage is going to be something other than your analytical skills. Your competitive advantage is your understanding of some aspect of life where you exceed others in understanding that. Maybe it's film, perhaps it's retail, perhaps it's health, perhaps it's computers.
Once you've figured out where your expertise lies, you start acquiring analytical skills, what platforms to learn, and tools specific to the industry you're interested in.

1.1.6 Summary of Module

1. **What is Data Science?**

- Data science is the **study of data.**
- Data science uses data to understand the world around us.
- Some consider data science to be the **art of uncovering the insights and trends hiding behind data.**

2. **Data scientist's role in an organisation**

- Data scientists **translate the data into stories to generate insights**, which aid in **strategic decision-making for companies or institutions**. Like biological or physical sciences,
- Data science deals with structured and unstructured data.
- The process of gleaning insights from data includes clarifying the problem, data collection, analysis, pattern recognition, storytelling, and visualisation.

3. what makes a skilled data scientist and expert advice on how to acquire these skills?

- **Curiosity, argumentation, and judgment are vital for data scientists**.
- Curiosity helps you explore data and come up with meaningful questions. Good, sound, reasonable arguments help you explain your findings after learning from the data, compelling the listener to adjust their ideas based on the new information. Good judgment guides you to start in the right direction**.**
- Skilled data scientists go beyond just being statisticians or computer experts. Companies are looking for **versatile individuals who know a lot about a particular subject, have some experience in programming and analysing data, and can communicate well**.
- Generally, data scientists are **comfortable with math, curious, and good at telling stories.** Their backgrounds can come from various fields like economics, engineering, medicine, and more. Once you understand your strengths and interests, focus on mastering data analysis in that field and select suitable tools for your industry

4. what does the future look like for you as a skilled data scientist?

- Data scientist jobs will also change as technology changes and data roles develop. Companies will require certification to ensure their employees are qualified.
- Data scientists must always think logically, use algorithms, and follow a systematic approach.
- Most importantly, they must gather data correctly and carefully analyse the models used, aiming toachieve successful business results.

Glossary

1. **Algorithms:-** A set of step-by-step instructions to solve a problem or complete a task.
2. **Model:-** A representation of the relationships and patterns found in data to make predictions or analyse complex systems retaining essential elements needed for analysis.
3. **Outliers:-.**When a data point or points occur significantly outside most of the other data set, it indicates anomalies, errors, or unique phenomena that could impact statistical analysis or modelling.
4. **A systematic quantitative approac**h using mathematical and statistical analysis is used to interpret numerical data.
5. **Structured data** is organised and formatted into a predictable schema, usually related tables with rows and columns.
6. **unstructured data:**- Unorganized data that lacks a predefined data model or organisation makes it harder to analyse using traditional methods. This data type often includes text, images, videos and other content that doesn't fit neatly into rows and columns like structured data.

Day 3 Jul 16, 2024

1.2 What do Data scientists do?

1.2.1 A Day in the Life of a Data Scientist

Speakers discussed the projects that they have done in data science.

1.2.2 Data Science Skills & Big Data

Summary:- Students in the course use **Jupyter Notebooks** and have their **virtual machine on Amazon Web Services.**

1.2.3 Understanding Different Types of File Formats

Standard file formats:

1. Delimited text file formats, or .CSV
2. Microsoft Excel Open .XML Spreadsheet, or .XLSX
3. Extensible Markup Language, or .XML
4. Portable Document Format, or .PDF
5. JavaScript Object Notation, or .JSON

Delimited text file formats, or .CSV

- Files used to store data as text.
- A delimiter separates each value.
- Delimiter - A sequence of one or more characters for specifying the boundary between independent entities or values.
- Comma, Tab, Colon, Vertical Bar, Space
- Delimited text file formats, or .CSV
  Ex:- CSV, TSV

JavaScript Object Notation, or .JSON

JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web.

• Language-independent data format
• Can be read in any programming language.
• Easy to use
• Compatible with a wide range of browsers.
Considered as one of the best tools for sharing data.

Portable Document Format, or .PDF

Portable Document Format, or PDF, Is a format **developed by Adobe** to present documents **independent of application software, hardware, and operating systems**.
• Can be viewed the same way on any device
• Is frequently used in legal and financial documents
• Can also be used to fill in data for forms.

Extensible Markup Language, or .XML

Extensible Markup Language, or XML, is a **markup language** with **set rules for encoding data.**
• Readable by both humans and machines
• Self-descriptive language
• Similar to .HTML in some respects
• Does not use predefined tags like .HTML does
• Platform independent
• Programming language independent

Microsoft Excel Open .XML Spreadsheet, or .XLSX

Is a Microsoft Excel Open XML file format that falls under the **spreadsheet file format**. It is an **XML-based file** format created by **Microsoft.**

• Open file format, accessible to most other applications
• Can use and save all functions available in Excel
• Is a secure file format as it cannot save malicious code

What Makes Someone a Data Scientist?

- I define **data science as something that data scientists do**.

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

Summary

1. Data scientists investigate and find explanations for many problems.

Ex.Toronto public transit customer complaints.
Transport-> Bad Weather
2. Environmental problem investigation
Ex:- Predicting algae blooms to prevent water toxicity.
Helping water treatment companies safeguard the water.
3. **Education**
   - ➢ Python notebooks
   - ➢ Linux
   - ➢ Databases
   - ➢ Pandas
4. **Mathematical sciences**
   - ➢ Algebra
   - ➢ Probability
   - ➢ Calculus
   - ➢ Statistics

5. **A data scientist uses statistics but is not only a statistician.**
6. Models and algorithms

Regression:- Relationship between two variables

7. About the data

Data sources	**Data structure**	**Data formats**
• Video	• Structured	• Text files
• Audio	• Unstructured	• Spreadsheets
• Text		• XML
		• PDFs
		• JSON

- An exceptional data scientist will be **a Computer scientist, Curiosity, Software engineer and Statistician.**

- It's a journey of **Exploration, Innovation, Story Telling.**

Glossary

1. Comma-separated values (CSV) / Tab-separated values (TSV): These are commonly used formats for storing tabular data as plain text where either the comma or the tab separates each value.

2. **Data file types:-** A computer file configuration is designed to store data in a specific way.

3. **Data format:-** How data is encoded to be stored within a data file.

4. **Data visualisation:-** A visual way, such as a graph, of representing data in a readily understandable way makes it easier to see trends in the data.

5. **Delimited text file:-** A plain text file where a specific character separates the data values.

6. **Extensible Markup Language (XML)**:- A language designed to structure, store, and enable data exchange between various technologies.

7. **Nearest neighbour:-** A machine learning algorithm that predicts a target variable based on its similarity to other values in the dataset.

8. **neural networks:-** A computational model used in deep learning mimics the structure and functioning of the human brain's neural pathways. It takes an input, processes it using previous knowledge, and produces an output.

Summary 2:-

What Do Data Scientists Do?

- ➢ Data science studies large quantities of data, which can reveal insights that help organisations make strategic choices.
- ➢ There are many paths to a career in data science; most, but not all, involve math, programming, and curiosity) about data.
- ➢ New data scientists need to be curious, judgemental and argumentative.
- ➢ Knowledgeable data scientists are in high demand. Jobs in data science pay high salaries for skilled workers:
- ➢ The typical work day for a Data Scientist varies depending on the project they are working on.
- ➢ Many algorithms are used to bring out insights from data.
- ➢ Some fundamental data science-related terms you learned in this lesson include outliers, models, algorithms, JSON,
- ➢ XML. CSV, and regression.

Module-2

Data Science Topics

2.1 How Big Data is Driving Digital Transformation

Digital Transformation affects business operations, updating existing processes and operations and creating new ones to harness the benefits of new technologies.

This digital change integrates digital technology into all areas of an organisation, resulting in fundamental changes in how it operates and delivers value to customers. It is an organisational and cultural change driven by Data Science, especially Big Data. The availability of vast amounts of data and the competitive advantage that analysing it brings have triggered digital transformation throughout many industries.

- ➢ Netflix moved from being a postal DVD lending system to one of the world's foremost video-streaming providers.
- ➢ The Houston Rockets NBA team used data from overhead cameras to analyse the most productive plays.
- ➢ Lufthansa analysed customer data to improve its service.

2.2 Introduction to Cloud

1. **Cloud Computing Overview**:
    - Cloud computing **delivers on-demand resources** (networks, servers, storage, applications, services) **over the Internet on a pay-for-use basis.**
    - Users access applications and data over the Internet instead of local computers (e.g., Google Drive, OneDrive, Dropbox).
    - Benefits include cost-effectiveness, accessing the latest versions,
    - saving local storage space,
    - Enabling collaborative work.
    - Use Online Versions of applications and pay a monthly subscription.

2. **Five Essential Characteristics of Cloud Computing**:
    - **On-demand self-service**: Access to resources (processing power, storage) without human interaction.
    - **Broad network access**: Resources are accessible via standard devices (mobile phones, laptops, etc.).
    - **Resource pooling**: Resources are pooled and dynamically assigned to meet demand, offering cost-efficiency.
    - **Rapid elasticity**: Resources can be scaled up or down according to need.
    - **Measured service**: You only pay for what you use, and resource usage is monitored.
3. **Cloud Deployment Models**:
    - **Public cloud**: Cloud services over the internet, shared with other companies, owned by the provider.
    - **Private cloud**: Exclusive infrastructure for a single organisation, can be on-premises or externally managed.
    - **Hybrid cloud**: A combination of public and private clouds that work together seamlessly.
4. **Cloud Service Models**:
    - **Infrastructure as a Service (IaaS)**: Access to infrastructure resources (servers, storage) without managing them.
    - **Platform as a Service (PaaS)**: Access to platforms (hardware and software) for application development and deployment.
    - **Software as a Service (SaaS)**: Centrally hosted software licensed on a subscription basis, also called "on-demand software."
5. **Conclusion**:
    - Cloud computing changes how computing services are consumed, making them cost-efficient and improving organisational agility.
    - It operates based on three deployment models (public, private, hybrid) and three service models (IaaS, PaaS, SaaS).

2.3 Cloud For Data Science

Cloud Advantages for Data Scientists:

- Cloud computing helps bypass the physical limitations of local machines.
- Allows the use of advanced storage and computing capacities without owning the machines.
- Enables high-performance computing with large datasets.
- Provides access to powerful algorithms that aren't available locally.

Data Storage & Collaboration:

- Cloud stores large amounts of data remotely and makes it accessible globally.
- Allows multiple entities or teams to work with the same data simultaneously from different locations.

Access to Tools & Libraries:

- Cloud platforms provide instant access to open-source **technologies like Apache Spark.**
- Up-to-date tools and libraries are available without local installation or maintenance.

Accessibility:

- Cloud is accessible from any device (laptop, tablet, phone) and supports global collaboration.

Cloud Platforms:

- **Major cloud providers include IBM Cloud, Amazon Web Services (AWS), and Google Cloud Platform**.
- IBM also **offers Skills Network labs** for learners to practice with tools like **Jupyter Notebooks and Spark clusters.**

Productivity Boost:

- The Cloud dramatically enhances productivity for data scientists through accessibility and collaboration.

2.4 Foundations of Big Data

1. **Big Data Definition**:
    - According to E&Y,
    - Big Data refers to large, dynamic, and diverse volumes of data generated by people, tools, and machines.
    - It requires innovative and scalable collection, hosting, and analysis technology to derive real-time business insights.
2. **5 V's of Big Data**:
    - **Velocity**: Data is generated at an extremely fast rate, and technologies (e.g., real-time streaming, local, cloud) are used to process it.
    - **Volume**: The sheer amount of data created is massive, driven by increased data sources, higher resolution sensors, and scalable infrastructure.
    - **Variety**: Data comes in many forms (structured and unstructured) from different sources like machines, people, and processes.
    - **Veracity**: Ensuring data accuracy, quality, and traceability is a challenge due to inconsistencies and the origins of the data.
    - **Value**: The goal of Big Data is to derive value, including business profit, social benefits, or personal satisfaction.
3. **Examples of the V's in Action**:
    - **Velocity**: Large volumes of data, such as hours of YouTube footage, are uploaded every minute, demonstrating rapid data generation.
    - **Volume**: Billions of digital devices generate around 2.5 quintillion bytes of data daily.
    - **Variety**: Data comes in many types—text, pictures, videos, health data, and Internet of Things (IoT) devices.
    - **Veracity**: 80% of data is unstructured, requiring advanced methods to derive reliable insights.
4. **Big Data Tools**:
    - Traditional data analysis tools need to be improved for the scale of Big Data.
    - Distributed computing technologies like Apache Spark and Hadoop can handle large datasets, enabling efficient extraction, loading, and data processing.
5. **Impact of Big Data**:
    - Big Data analysis offers organisations new ways to connect with customers and improve their services.
    - Everyday digital interactions, such as using smartwatches or smartphones,

contribute to the global flow of Big Data.

2.5 Data Science and Big Data

1. **Programming Background:**
- Many people have at least some programming knowledge.
- Some individuals are highly skilled, such as a Master of Science in Computer Science, while others have essential exposure, like taking a course years ago.
- The critical requirement is the ability to think computationally.
2. **Rise of Data Science and Business Analytics:**
- These fields have become very popular in the last four or five years.
- New tools, approaches, and vast amounts of data are available that traditional methods couldn't handle.
3. **Growing Importance for Employers:**
- Initially, companies had small efforts in handling big data, but now they have significantly expanded their capacity.
- **Example:** A significant bank went from a tiny big data cluster to five or six major ones, using various data science techniques on their credit card data.
4. **Rising Popularity of Data Science in Education:**
- The number of undergraduate data science course students has dramatically increased, reflecting the rising awareness and demand.
- Parents advise their children to pursue data science and analytics instead of traditional fields like accounting or marketing.
5. **Definition of Big Data:**
- Big data refers to **data with such high volume and velocity that traditional database systems cannot manage it.**
- Statisticians often define big data a**s something too large to fit on a thumb drive.**
- Google played a crucial role in developing big data techniques with their PageRank algorithm, which led to tools like Hadoop.
6. **Big Data Analysis:**
- New analytical and statistical techniques have been developed to manage large data sets.
- Deep learning will likely become part of the conversation as big data evolves.

2.6 What is Hadoop?

Traditional Computation vs. Big Data:

- Traditionally, data was brought to the computer to process it.
- Larry Page and Sergey Brin (Google founders) innovated by slicing data into pieces, distributing, and replicating it across many computers (initially hundreds, now tens of thousands).
- A program is sent to each computer in the cluster, where it processes its piece of data, and results are sent back.
- This method involves **two fundamental processes**: "map" (mapper process) and "reduce."
- This approach enables handling very large datasets and is highly scalable.

Scalability of Big Data Clusters:

- Big data clusters scale linearly – more servers lead to increased performance and data handling capacity.
- This breakthrough benefitted significant social media companies.

Development of Hadoop:

- Yahoo hired Doug Cutting, who developed a clone of Google's significant data architecture called Hadoop.
- Hadoop has become very popular with many companies in the extensive data ecosystem.

Components of Data Science:

- Data science includes long-standing areas like probability, statistics, algebra, linear algebra, programming, and databases.
- Machine learning is now an integral part of the process, allowing the analysis of large datasets to find patterns and generate hypotheses.

Evolution of Data Science:

- Traditional statisticians may be bothered by the shift from hypothesis-driven research to pattern recognition in large datasets.
- **Decision Sciences emerged from computer science, probability, statistics, and mathematics.**

Role of NYU Stern School in Data Science:

- NYU Stern was well-positioned to embrace data science, with departments in statistics, operations management, and information systems.
- Foster Provost was the first director of the NYU Center for Data Science.

Rise of Data Science:

- Data science has grown rapidly, similar to big data, with notable increases in interest over the past few years.
- Five years ago, the term "data science" was hardly known.

Business Analytics and Data Science:

- Faculty are still defining the scope of business analytics and data science as the fields evolve.
- Deep learning, added in the last few years, has revolutionised neural networks.
- Multi-layer neural networks, particularly research in Toronto, have become widely adopted by companies like Google and Facebook.

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

Big Data Technologies Overview:

- Big Data involves handling structured, semi-structured, and unstructured data to derive value.
- Key technologies discussed include NoSQL databases, Data Lakes, Apache Hadoop, Apache Hive, and Apache Spark.

Apache Hadoop:

- A Java-based, open-source framework for distributed storage and processing of large datasets.
- Hadoop works across clusters of computers, with each node offering local storage and computation.
- **Key benefits:**
  - Reliable, scalable, and cost-effective data storage without format requirements.
  - It supports emerging data formats (audio, video, social media, and clickstream).
  - Enables real-time, self-service access and cost optimisation by moving

infrequently used data ("cold data") to Hadoop.

Hadoop Distributed File System (HDFS):

- HDFS is a **scalable, reliable storage system that partitions files across multiple nodes.**
- It allows parallel access to data, ensuring fault tolerance by replicating file blocks on multiple servers.
- Benefits include fast recovery from hardware failures, support for large datasets, and portability across platforms.
- Benefits:-
  - ➢ **Fault Tolerance**: HDFS replicates data across multiple nodes, ensuring data availability and protection against data loss if a server fails.
  - ➢ **Scalability**: HDFS can scale to hundreds of nodes in a single cluster, allowing it to handle very large datasets.
  - ➢ **Fast Recovery**: HDFS is built to detect and recover hardware failures quickly and automatically.
  - ➢ **Parallel Data Access**: By partitioning large files across multiple nodes, HDFS allows computations to run in parallel, improving performance.
  - ➢ **Data Locality**: HDFS moves computation closer to where the data resides, minimising network congestion and increasing throughput.
  - ➢ **High Throughput**: HDFS is optimised for high data throughput, making it suitable for applications requiring access to large datasets.
  - ➢ **Portability**: HDFS is compatible with various hardware

Apache Hive:

- Hive is an open-source data warehouse software built on top of Hadoop for **reading, writing, and managing large datasets**.
- It's suitable for data warehousing tasks like ETL (Extract, Transform, Load), reporting, and data analysis.
- Hive is **read-based**, making it less ideal for transaction processing or applications needing fast response times.

Apache Spark:

- Spark is a general-purpose data processing engine that handles large data volumes and complex real-time analytics.
- Critical use cases include interactive analytics, stream processing, machine learning, and ETL.

- It **uses in-memory processing for faster computations** and **supports multiple programming languages (Java, Scala, Python, R, SQL).**
- Spark can integrate with Hadoop and access data from HDFS and Hive.

Data Mining Lab

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

Author: Murtaza Haider

Prescribed Reading: Chapter 12, Pg. 529-531

1. **Establishing Data Mining Goals:**
    - Set clear goals and critical questions for data mining.
    - Consider cost-benefit trade-offs, expected accuracy, and the diminishing returns of increasing accuracy.
2. **Selecting Data:**
    - The quality of data affects the output.
    - Identify suitable data sources or plan new data collection if necessary.
    - Data type, size, and frequency impact costs.
3. **Preprocessing Data:**
    - Clean the data by removing irrelevant and erroneous data.
    - Address missing data and ensure the data's integrity.
    - Handle systematic biases due to missing data.
4. **Transforming Data:**
    - Reduce the number of data attributes using techniques like Principal Component Analysis.
    - Transform variables (e.g., converting continuous to categorical).
    - Aggregate data where necessary for better analysis.
5. **Storing Data:**
    - Store data in formats that allow easy read/write access.
    - Ensure data security, privacy, and storage efficiency for easy mining retrieval.
6. **Mining Data:**
    - Data analysis methods (parametric, non-parametric, machine learning) should be used for mining.
    - Start with data visualisation to understand trends.
7. **Evaluating Mining Results:**
    - Test predictive models through in-sample forecasts.

    - Share results with stakeholders and iterate the process based on feedback to improve future analysis.

2.8 Lesson Summary: Data Mining

Impact of Big Data:

- Big data is transforming society, business operations, and industries such as sports. Organizations need to change their approach to business to adapt to big data.

Big Data Characteristics:

- Commonly accepted characteristics of big data are:
    - **Value**: Big data creates value for organisations.
    - **Volume**: The large scale of data is driven by increasing data sources and infrastructure.
    - **Velocity**: Data is **generated quickly and continuously from many sources**.
    - **Variety**: Data comes from different sources, both structured and unstructured.
    - **Veracity**: Refers to the data's quality, accuracy, and reliability.

Role of Cloud Computing:

- Cloud computing enables the processing and storage of big data with five essential characteristics:
    - **On-demand access**: Access to processing, storage and network.
    - **Network access**: Resources are accessible through the Internet.
    - **Resource pooling**: Multiple users can share resources efficiently.
    - **Elasticity**: Ability to scale resources up or down as needed.
    - **Measured service**: You only pay for what you use.

Cloud Tools:

- **Hadoop**: Provides distributed storage and processing for large data sets.
- **Hive**: A data warehouse built on Hadoop for querying and analysing large datasets.
- **Spark**: A data processing engine for handling large volumes of data across various

applications.

Data Mining Process:

- **Goal setting**: Define critical questions and concerns about cost and benefits.
- **Selecting data sources**: Identify where data will come from.
- **Preprocessing**: Clean the data by removing irrelevant information.
- **Transformation**: Format the data appropriately for analysis.
- **Mining**: Apply analysis methods and machine learning algorithms to extract insights.
- **Evaluation**: Test the results and share insights with stakeholders.

Iterative Process:

- Data mining is iterative, with each round of analysis informing future efforts.

Summary:

- Big data's characteristics (value, volume, velocity, variety, veracity) transform industries. Cloud technologies enable handling big data, and tools like Hadoop, Hive, and Spark make it efficient to process and mine big data.

Glossary

https://drive.google.com/file/d/1uid-e9b_20Rmo0cHwi3P5ARcMQSFVx4I/view?usp=sharing

Deep Learning and Machine Learning

Lesson Overview: Deep Learning and Machine Learning

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**:
   - Refers to massive, fast-growing, and varied data sets that cannot be analyzed using traditional methods.
   - Defined by the five V's: **velocity**, **volume**, **variety**, **veracity**, and **value**.
   - With advancements in computing power and tools, analyzing vast data sets has become feasible, providing new knowledge and insights.
2. **Data Mining**:
   - The process of automatically searching and analyzing data to discover hidden patterns.
   - Involves data preprocessing, transforming it for analysis, and extracting insights using tools like visualisation, machine learning, and statistical models.
3. **Machine Learning**:
   - A subset of AI that uses algorithms to analyse data and make decisions without being explicitly programmed.
   - Algorithms are trained on large data sets and learn from examples instead of following rules-based programming.
   - Enables machines to solve problems and make accurate predictions based on provided data.
4. **Deep Learning**:
   - A specialised subset of machine learning that uses layered neural networks to simulate human decision-making.
   - Deep learning algorithms improve over time, continuously learning and enhancing the quality and accuracy of results.
5. **Artificial Neural Networks**:
   - Comprise small computing units called neurons that learn to make decisions over time.
   - Neural networks are essential to deep learning, becoming more efficient with increasing data volumes, unlike other machine learning algorithms that may plateau.
6. **Artificial Intelligence (AI) vs. Data Science**:
   - **Data Science**: An interdisciplinary field that **extracts knowledge and**

insights from large data sets using techniques like statistics, machine learning, and data visualization.

- **AI**: Involves enabling computers to learn, solve problems, and make intelligent decisions.
- Data Science can employ AI techniques like machine learning and deep learning to draw insights from data, but AI and Data Science are not subsets of one another.

7. **Overlap of AI and Data Science**:
   - Both can involve **big data** and use similar techniques. Still, they serve different purposes: Data Science focuses on extracting meaning from data, while AI focuses on intelligent decision-making by machines.

Generative AI and Data Science

1. **Generative AI Overview**:
   - Generative AI is a subset of artificial intelligence focused on producing new data, unlike traditional AI, which analyses existing data.
   - It allows machines to create content such as images, text, music, and more, mimicking human creations.
2. **How Generative AI Works**:
   - Generative AI relies on deep learning models like **Generative Adversarial Networks (GANs)** and **Variational Autoencoders (VAEs)**.
   - These models learn from large datasets to create new instances replicating the original data's underlying distribution.
3. **Applications of Generative AI**:
   - **Natural Language Processing**: Tools like GPT-3 generate human-like text for content creation and chatbots.
   - **Healthcare**: Generating synthetic medical images for training healthcare professionals.
   - **Art**: Creating unique and stunning visual artworks.
   - **Gaming**: Developing realistic environments, characters, and game levels.
   - **Fashion**: Designing new styles and offering personalized shopping

recommendations.

4. **Use of Generative AI by Data Scientists**:
    - **Data Augmentation**: Data scientists use generative AI to create synthetic data, mimicking real datasets, which helps when limited data is available.
    - **Model Training and Testing**: Synthetic data complements accurate data to train and test models more effectively.
    - **Coding Automation**: Generative AI can generate and test software code, allowing data scientists to focus on higher-level tasks like problem-solving and hypothesis evaluation.
    - **Business Insights**: AI tools like **IBM's Cognos Analytics** help generate and update business insights, uncover hidden patterns, and support better decision-making.
5. **Benefits of Data Science**:
    - Generative AI helps overcome data limitations and provides more robust data-driven insights.
    - It enhances efficiency by automating tasks like coding and report generation, allowing data scientists to work on more strategic aspects.

Neural Networks and Deep Learning

- **Neural Networks**:
    - Initially designed to mimic brain neurons and synapses.
    - 20-23 years ago, I used to recognise simple patterns like handwritten digits.
    - **Basic process:** Inputs are fed into processing nodes, which transform, aggregate, and pass them through layers, resulting in an output.
    - Training involves repeatedly feeding data into the network until it converges, producing desired outputs.
    - **Problem:** Computational intensity limited their early use.
- **Decline and Revival**:
    - Neural networks became impractical for more significant problems due to high computational demands.
    - Went out of favour around 15 years ago.

    - Resurgence with **deep learning**, a more advanced form of neural networks with many layers, powered by significant computing resources.
- **Deep Learning**:
    - Popularised in the last 4-5 years.
    - Essentially, "neural networks on steroids" have multiple layers.
    - It requires enormous computational power, particularly **graphics processing units (GPUs)**, for complex matrix and linear algebra calculations.
    - Neural networks and deep learning are now **used for tasks** like:
        - Speech recognition
        - Image recognition
        - Classifying objects (e.g., distinguishing between cats and dogs).
    - Deep learning models **"learn" on their own without explicit programming.**
- **Technological Advances**:
    - **The importance of GPUs:** GPUs excel at handling the vast amount of parallel computation needed for neural networks.
    - **Example:** A university research centre had 200 servers equipped with GPUs, each GPU offering the equivalent of 600 processing cores.
- **Applications**:
    - Speech recognition and generation, as seen in class demonstrations where a model learns to recognise speech in real time.
    - Image classification on a large scale.
    - Continuous learning capabilities – the model improves performance through repeated data exposure.
- **Practical Considerations to Start Learning Deep Learning**
    - Linear algebra is a **core component of the transformations in deep learning models.**
    - Deep learning requires specialized computational resources (high-powered servers, GPUs).
    - While many packages are available that simplify deep learning processes, it's crucial to have a basic understanding of the underlying math, like matrix operations.

Applications of Machine Learning

Machine Learning Applications:

- **Recommender systems** are a significant application of machine learning.
- **Classification and cluster analysis are standard techniques used in marketing.**
- Market basket analysis (finding items bought together) was once computationally tricky but is now familiar with machine learning.

Predictive Analytics:

- Machine learning techniques are used for predictive analytics.
- Methods such as decision trees, Bayesian analysis, and naive Bayes are widely employed.
- While the detailed workings of these methods may be optional, understanding their **trade-offs, such as precision vs. recall and challenges like overfitting, is important.**

Recommender Systems in Fintech:

- Like Netflix or Facebook recommendations, fintech uses machine learning to suggest similar investment ideas to users.
- Recommendations can help professionals explore related assets, companies, or investment techniques.

Fraud Detection in Finance:

- Machine learning plays a critical role in real-time fraud detection.
- Models are built using historical transaction data to flag suspicious credit card transactions, which are then routed for further review.

Machine Learning Packages:

- Tools like R have made it easier to apply complex machine-learning techniques, even for users with basic data science knowledge.

Regression

The key points from Chapter 7 (pg. 235-236) of *Getting Started with Data Science* by Murtaza Haider:

1. **Introduction to Regression Models**:
    - Regression models are statistical techniques widely used across various fields, such as medicine and business.
    - These models help understand relationships between variables, such as predicting outcomes or understanding consumer behaviour.
2. **Tall Parents and Height of Children**:
    - Taller parents often have tall children, but the children are not necessarily taller than their parents. This observation relates to a statistical phenomenon studied by **Sir Francis Galton in 1886, which led to the development of regression models.**
3. **Author's Personal Story on Regression Analysis**:
    - The author narrates an anecdote about his **Master's thesis on hedonic price models for real estate.** His wife initially dismissed his findings as apparent (e.g., "larger homes sell for more than smaller ones").
    - However, the critical value of his research lies in **quantifying these relationships, such as how much more a house sells for with additional features like bedrooms or washrooms.**
4. **Key Findings of Author's Thesis**:
    - Larger homes sell for more, but certain features, like additional washrooms, contribute more to the price than others, like additional bedrooms.
    - Proximity to transport infrastructure, like subways, increases property prices, while proximity to highways decreases them.
    - Proximity to shopping centres has a nonlinear impact: houses very close to shopping centres are less valuable, while those slightly farther away (within a 2.5-5 km range) tend to sell for more.
5. **Utility of Regression Models**:
    - Regression models allow researchers and professionals to measure the

magnitude of relationships between variables, providing deeper insights beyond what is obvious.
    - These models can answer a range of questions, such as the impact of house features (lot size, number of bedrooms) or location factors (proximity to power lines or transportation) on housing prices.

The overall message is that regression models are powerful tools that provide measurable, quantifiable insights into relationships between variables that may appear obvious at first glance.

Lab: Exploring Data using IBM Cloud Gallery

Summary of IBM Cloud Resource Hub Exploration

In this lab, you learned how to explore different datasets and notebooks available on IBM Cloud Resource Hub. The exercise involved working with three samples: a numeric dataset, a non-numeric dataset, and a Jupyter Notebook that used decision optimisation. Below are the key takeaways:

Exercise 1: Numeric Dataset (Forest Fires)

- **Dataset**: UCI Forest Fires Dataset.
- **Objective**: Predict the burned area of forest fires in northeast Portugal based on meteorological and other factors.
- **Attributes**:
    - Spatial coordinates (X, Y), month, day, FFMC, DMC, DC, ISI (indexes from the FWI system), temperature, humidity, wind speed, rain, and burned area.
- **Exploration**: The data is skewed towards small burned areas, suggesting that transforming the area attribute (e.g., using a logarithm) could be helpful for modelling.

Exercise 2: Non-Numeric Dataset (Airbnb Reviews)

- **Dataset**: Airbnb Data for Analytics: Trentino Reviews.
- **Objective**: Analyze various attributes related to Airbnb listings, including reviews,

ratings, and booking data.

- **Attributes**:
  - **Listings**: 96 attributes, such as price, longitude, latitude, listing type, superhost status, neighbourhood, and ratings.
  - **Reviews**: Includes date, listing ID, reviewer ID, and textual comments.
  - **Calendar**: Booking details for the upcoming year, including availability and price.
- **Exploration**: This dataset can be used for sentiment analysis, occupancy rate prediction, star rating analysis, and guest location preferences. Potential uses include improving guest satisfaction and optimizing listings.

Exercise 3: Jupyter Notebook (Optimal Store Locations)

- **Notebook**: Finding optimal locations of new stores using Decision Optimization with Python.
- **Objective**: Determine the optimal location of a coffee shop to minimize the distance from local libraries, enhancing accessibility for book readers.
- **Features**:
  - Displays libraries on a map.
  - Uses an optimization model to identify ideal coffee shop locations.
- **Exploration**: The optimization model solves the problem of minimizing distance, providing a data-driven solution to determine store placement.

Conclusion

The lab demonstrated how to explore and use different datasets on the IBM Cloud Resource Hub, ranging from purely numeric data to datasets with text attributes and Jupyter Notebooks for decision optimization. This gives data scientists various tools and data types, depending on the nature of their analysis or modelling task.

Lesson Summary: Deep Learning and Machine Learning

Key Points: Deep Learning and Machine Learning Summary

1. **Overview of Artificial Intelligence (AI)**:
    - AI mimics tasks associated with human intelligence.
    - AI has become more accessible and used frequently by data scientists.
2. **Machine Learning (ML)**:
    - Subset of AI that uses algorithms to analyze data and make predictions.
    - ML learns from data without requiring explicit programming.
    - It is commonly used for tasks like predictive analytics and fraud detection.
3. **Deep Learning (DL)**:
    - Subset of ML that uses layered neural networks to simulate decision-making.
    - Neural networks consist of neurons that process data and learn over time.
    - DL becomes more efficient as data volume increases.
4. **Neural Networks**:
    - It comprises small units (neurons) that learn to differentiate data patterns (e.g., distinguishing between a cat and a dog).
5. **Generative AI**:
    - Focuses on producing new data rather than analyzing existing data.
    - Can generate content such as images, music, languages, and code.
    - It helps create synthetic data when there's a lack of real data.
6. **Regression in Machine Learning**:
    - Statistical technique to identify correlations between inputs and outputs.
    - Commonly used in ML to predict outcomes (e.g., house prices based on features like size and bedrooms).
7. **AI, ML, and DL Relationship**:
    - AI is a broad field encompassing machine learning.
    - Machine learning is a subset of AI.
    - Deep learning is a subset of ML.

8. **Application of AI in Data Science**:
    - Data scientists use AI, ML, and DL to make predictions and derive insights from big data.

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.

Applications and Careers in Data Science

Learning Objectives

- Describe the contents of a data science job posting.
- Describe the application of data science in healthcare.
- Explain how companies can start on their data science journey.
- Describe some of how consumers generate data.
- Describe how businesses such as Netflix, Amazon, UPS, Google, and Apple use data generated by their consumers and employees.
- Compare some of the qualities that differentiate data scientists from the qualities of other data professionals.
- Articulate the purpose of the final deliverable of a data science project and the role of storytelling in the final deliverable.
- Describe what the final report of a Data Science project should cover and how it should be structured for best results.
- Demonstrate your understanding of data science by articulating what data scientists do and what a data science report contains.

Data Science Application Domains

Lesson Overview: Data Science Application Domains

This lesson, titled **"Data Science Application Domains"**, explores how data science is applied across various sectors, helping to drive innovation, solve complex problems, and even save lives. You will engage with a series of multimedia resources designed to give you a holistic understanding of the powerful impact of data science on industries and society.

Here's a breakdown of the critical assets included in this lesson:

Asset Breakdown:

1. **"How Should Companies Get Started in Data Science?" (Video)**
    - Learn how organisations can begin their data science journey effectively,

focusing on the essential first steps and strategies for success.

2. **"Old Problems with New Data Science Solutions" (Video)**
    - Explore how data science offers modern, innovative solutions to long-standing, real-world problems, helping industries overcome challenges more efficiently.
3. **"Applications of Data Science" (Video)**
    - This video dives into the diverse applications of data science, from finance and marketing to healthcare and agriculture, showcasing how it transforms various industries.
4. **"How Data Science is Saving Lives" (Video)**
    - Learn about the life-saving potential of data science, particularly in healthcare, where it aids in diagnostics, treatment plans, and disease prevention.
5. **"The Final Deliverable" (Reading)**
    - This reading material focuses on the end product of data science projects, detailing what constitutes a successful final deliverable, including the presentation of actionable insights.
6. **Practice Quiz**
    - After the reading, you can take a quiz to test your understanding of key concepts related to the final deliverable in data science.
7. **"Lesson Summary" (Video)**
    - Recap the core takeaways from the lesson, reinforcing your understanding of how data science is applied across various domains.
8. **Practice Quiz**
    - Another quiz to help you evaluate your grasp of the material and check your overall understanding of the concepts presented in the lesson.
9. **Glossary**
    - A handy glossary of terms is provided to help you review and solidify the key terminology introduced throughout the lesson.

Key Takeaways:

By the end of this lesson, you will understand how data science is applied in real-world scenarios, how companies can integrate data science into their operations, and how it is transforming sectors like healthcare, business, and more. Additionally, you'll be equipped with knowledge on how to create successful data science project deliverables and understand the terminology used in the field.

How Should Companies Get Started in Data Science?

1. **Measurement is Key to Improvement**:
   - If a business cannot measure something, it cannot improve it.
   - Companies need to measure costs to reduce them and profits to increase them.
2. **Data Collection is Crucial**:
   - Companies must start capturing cost-related data like labour and material costs.
   - Record sales costs per product, total costs, and revenue sources.
3. **Analyse Revenue Sources**:
   - Evaluate where revenue comes from. For example, does 80% of revenue come from 20% of customers?
4. **Proper Data Handling**:
   - Archive data instead of overwriting it. All data, even if it's decades old, remains relevant.
   - Data should be appropriately documented and archived for future reference.
5. **Consistency and Best Practices**:
   - Implement consistent data archiving and documentation practices as early as possible.
   - If you haven't done so, start now to ensure data preservation and usefulness over time.
6. **Data Science Relies on Quality Data**:
   - Data science is only as good as the quality of the data collected—garbage in, garbage out.
   - For practical analysis, companies must ensure they have clean and relevant data.
7. **Building a Data Science Team**:
   - Hire a team of data scientists, not just one individual. Each team member should have strengths in different areas.
   - Employees must be interested in data science for engagement and success in data-related tasks.
8. **Start Measuring Early**:
   - If a company hasn't measured things properly in the past, it's crucial to start now to avoid poor data-driven decisions.

Old Problems, New Data Science Solutions

1. **Data Science in Problem-Solving:**
   - Organisations use data science to discover optimal solutions to problems, leveraging vast amounts of available data.
2. **Uber's Use of Data in Transportation:**
   - Uber collects real-time data to determine driver availability, adjust pricing, and place drivers in optimal locations at the correct times.
3. **Toronto Transportation Commission:**
   - Toronto uses data science to analyze traffic flow and streetcar operations.
   - Efforts include gathering data, analyzing customer complaints, and using traffic performance data to restructure traffic flow.
   - By focusing on peak-hour traffic congestion, they reduced commuter delays from 4.75 hours in 2010 to 3 hours by 2014.
4. **Data Science in Environmental Issues:**
   - Freshwater lakes are at risk due to harmful cyanobacterial blooms.
   - U.S. scientists use robotic boats, drones, and data collection tools to monitor lakes, **build predictive models, and improve water safety.**
   - This proactive data approach helps protect public health and ecological needs by predicting harmful bloom occurrences.
5. **Developing Efficient Data Solutions:**
   - To create better solutions, organizations must:
     - Identify and understand the problem.
     - Gather and analyse data.
     - Choose the right data tools and develop a strategy.
   - Case studies can help tailor potential solutions, and machine learning models can be developed from the extracted data.
6. **Refining Data Strategy:**
   - Organisations need time to refine their data strategies, but the benefits of using data science are significant in enhancing efficiency.

Applications of Data Science

1. **Impact of Data Science on Business:**

- Data science and big data transform businesses, particularly in day-to-day operations, financial analytics, and customer interactions.
- Businesses can gain significant value from the insights provided by data science.

2. **Mass Data Generation by Individuals:**

- People generate large amounts of data daily, often unknowingly, through digital activity, creating online behaviour patterns.

3. **Recommendation Engines:**

- Sites like Amazon, Netflix, and Spotify use data science **to create recommendation engines based on customer preferences and historical behaviour.**
- Personal assistants like Siri use data science to provide answers to user questions.

4. **Google's Data Use:**

- Google tracks user habits (e.g., online shopping, social media) and uses location data to recommend restaurants, shops, and other attractions.

5. **Wearable Devices and Data Collection:**

- Wearable devices like **Fitbits and smartwatches add personal data** (e.g., activity levels, sleep patterns, heart rate) to the **growing pool of user-generated data.**

6. **Business Adoption of Data Science:**

- In 2011, McKinsey & Company predicted data science would become the key competitive tool for businesses, driving productivity, growth, and innovation.
- In 2013, UPS utilised data from drivers and vehicles in a route guidance system to optimise time, money, and fuel use.

7. **Data Science Driving Competitive Advantage:**

- Data science is changing how businesses compete and operate, providing a

competitive edge.

- Netflix is an example of **collecting massive data on users' preferences** (e.g., viewing habits, searches for actors/directors) to predict the success of content.

8. **Netflix's Use of Data Science:**

- Netflix analyzes data from millions of users to predict which shows will succeed before production begins.

Example: Netflix knew "House of Cards" would be a hit based on data showing user preference for director **David Fincher, actress Robin Wright, and the success of the original British series.**

9. **Predictive Power of Data Science:**

- Data science allows Netflix to understand user preferences and predict what they will want to watch, often before users know their desires.

How Data is Saving Lives

Data Science in Healthcare:

- **Predictive Analytics:**
  - It helps healthcare professionals give tailored treatments.
  - Combines data mining, modelling, statistics, and machine learning.
  - Examines gene markers, environmental factors, associated conditions, and more.
  - Recommends tests, trials, and treatments specific to each patient.
- **Knowledge Accessibility:**
  - Physicians have personal knowledge, but predictive analytics ensure they can access the latest disease information, tests, and treatment options.
  - Standardising the knowledge physicians use improves consistency and patient outcomes.
- **Study Example (Boston Consulting Group & AdvaMedDx):**
  - The patient's oncologist may not always know the best diagnostic tests for cancer with specific gene markers.
  - Data science tools can bridge this gap by providing the necessary information.
- **Use of Electronic Medical Records (EMR):**
  - Institutions like NorthShore University HealthSystem lead the **use of EMR systems.**

    - These systems provide anonymized data for advanced analytics and medical research.
    - Moving from basic descriptive analytics to predictive insights through data science.

2. Data Science in Disaster Preparedness:

- **Predictive Analytics for Natural Disasters:**
    - It enhances the ability to predict natural disasters like earthquakes, hurricanes, floods, and volcanic eruptions.
    - Early prediction can save lives by alerting populations to danger in advance.
- **University of Warwick Research:**
    - Social media data (photos, keywords) was used to track and predict weather events like floods and hurricanes.
    - Combined with traditional data sources from scientists and weather stations to improve localized predictions.
- **Education Focus:**
    - Institutions like the **University of Chicago** offer programs focused on threat and response management, integrating data science for disaster preparedness.

3. Overall Impact of Data Science:

- Enables the analysis of vast data sets from various sources.
- Provides insights that can improve healthcare treatments and disaster response.
- Saves lives by offering early warnings and the best medical interventions.

The Final Deliverable IBM Lab

Notes from Chapter 3 (Pages 52-53) – *Getting Started with Data Science* by Murtaza Haider:

The Final Deliverable in Analytics

- The **primary goal** of analytics is **to communicate insights and findings to stakeholders who will use the information for decision-making** (e.g., policy formulation or strategy development).
- **Essays and reports** are typical deliverables in academia, ranging from 1,000 to 7,000 words, summarizing insights through tables and plots.
- In **consulting and business**, reports vary in size:

    - **Short documents**: Less than 1,500 words, illustrated with visuals like tables and plots.
    - **Comprehensive reports**: Can extend to several hundred pages.

Example: Deloitte's "United States Economic Forecast"

- **A 24-page report** published by Deloitte University Press in December 2014.
- It focuses on the **U.S. economy** with a **positive outlook,** countering the perception that it was sluggish.
- The report begins with a **grabber** highlighting **robust economic and job growth**.
- Cites **Voltaire** and presents a **narrative** that emphasizes:
    - Investment in U.S. manufacturing equipment.
    - Potential rise in consumer consumption due to lower oil prices.

Use of Data and Analytics in the Report

- **Time series plots** illustrate economic trends, such as:
    - GDP growth showed contraction during the Great Recession and subsequent rebound.
    - Projections for **four future economic scenarios**.
    - Changes in **consumer spending** tied to income inequality, referencing **Thomas Piketty's** work.
- Graphics cover a range of sectors:
    - **Housing, business, government**, and **international trade**.
    - **Labor markets, financial markets**, and **prices**.
- The appendix contains **four tables** showing data for the four projected scenarios.

Impact of the Narrative

- **The narrative** is crucial for making the report impactful:
    - Without the storytelling element, a PowerPoint with just graphics and tables would fail to convey the same message.
    - References to **Piketty and Voltaire** strengthen the message.
- **Deloitte's report** illustrates data and showcases the firm's ability to provide actionable insights, building credibility and expertise in the economic domain.

Importance of Planning

- **Planning the final deliverable** is critical to creating a solid document.
    - The authors of the Deloitte report likely started by defining the scope and the **key message**.
    - Data and analytics were then used to **support the narrative**.
- Failing to conceptualise the final deliverable beforehand often results in

poor-quality documents where the **analytics and narrative** don’t align well.

Lesson Summary: Data Science Application Domain

Purpose of Data Science in Organizations:

- **Drive Business Goals:** Improve efficiency, make predictions, and save lives.
- **Problem-Solving:** Data science helps find optimal solutions to existing problems, requiring a clear understanding of the problem.

Critical Steps in the Data Science Process:

1. **Measurement:** Capturing and gathering data is crucial. If data isn't measured, it can't be improved. If the organisation isn't capturing data, help them find ways to do so.
2. **Data Integrity:** Never overwrite old data – it remains relevant.
3. **Data Cleaning:** Once collected, the data must be cleaned for analysis.

Developing Analytical Solutions:

- Identify tools and analysis strategies.
- Use case studies to customize potential solutions.
- Develop machine learning and statistical models.
- Refine data strategies over time.

Applications of Data Science in Business:

- **Recommendation Engines:** Amazon uses machine learning to make personalized product recommendations.
- **Logistics Optimization:** UPS uses data to optimize driver routes, saving time, fuel, and money.
- **Demand Prediction:** Uber uses data to match drivers to demand, optimizing pricing and availability.

Gaining Competitive Advantage:

- Streaming services, like Netflix, analyze massive amounts of user data to predict the success of shows, detect viewing habits, and improve content recommendations.

Impact on Healthcare:

- Data scientists use **predictive analytics** (data mining, modelling, machine learning) to improve patient outcomes, identify disease markers, and recommend treatments.
- Healthcare organizations can shift from **descriptive** to **predictive analytics** with better data analytics capabilities.

Data Science in Natural Disasters:

- Predicting disasters (earthquakes, hurricanes, floods, etc.) by analyzing large datasets, potentially saving lives.

Delivering Findings:

- The final deliverables in academia are research papers/reports.
- Data scientists present analytics-driven reports with tables and plots in business, creating clear, evidence-based narratives.

These notes summarise how data science powers businesses, improves efficiency, and saves lives, emphasizing the process of data collection, analysis, and delivery of actionable insights.

Careers and Recruiting in Data Science

Lesson Overview: Careers and Recruiting in Data Science

In this lesson, "Data Science Application Domains," you'll explore the many real-world applications of data science and how it impacts different industries. The key assets in this lesson offer insights into careers, recruitment, and the skills needed to succeed in the data science field.

Asset Summaries

1. **"How Can Someone Become a Data Scientist?" Video**
    - This video outlines the steps to becoming a proficient data scientist, highlighting key skills, education, and tools required for success.
    - Emphasises the importance of programming (Python, R), mathematics, statistics, and real-world experience through projects.
2. **"Recruiting for Data Science" Video**
    - Learn how organizations approach recruiting for data science positions.
    - Key recruitment considerations include technical skills, problem-solving abilities, and business acumen.
3. **"Careers in Data Science" Video**
    - This video explores the wide range of career opportunities in data science, from data analyst roles to more specialized positions like machine learning engineers and data engineers.
    - Different industries have varying needs for data science professionals, offering diverse pathways.
4. **"Importance of Mathematics and Statistics for Data Science" Video**
    - Mathematics and statistics form the foundation of data science.
    - This video emphasizes their critical role in data analysis, modeling, and interpreting results, helping data scientists make accurate predictions and informed decisions.
5. **"The Report Structure" Reading**
    - Understanding the structure of reports in data science projects is essential for clear communication of insights.
    - This reading provides a guide on how to organize and present findings effectively, focusing on clarity, coherence, and conciseness.
6. **Practice Quiz**
    - A quick test to assess comprehension of the "Report Structure" reading material.
7. **"Infographics on Roadmap" Reading**

    - The infographic provides a roadmap for building a successful career in data science, detailing key milestones like learning essential programming languages, gaining experience, and acquiring domain-specific knowledge.
8. **"Lesson Summary" Video**
    - A video recap of the key concepts covered in the module, reinforcing the important takeaways about careers, recruitment, and the role of foundational knowledge in data science.
9. **Practice Quiz**
    - Another quiz to test your knowledge and retention of the lesson content.
10. **Glossary**
    - A glossary of terms used throughout the lesson, which serves as a helpful reference for key data science terminology.
11. **Grade Quiz**
    - A graded quiz to evaluate your understanding of the applications of data science in a business context.
12. **"Data Science in Business" Reading**
    - This reading provides a summary of the module, highlighting the application of data science in business, career paths, recruiting practices, and the importance of mathematical skills.

Key Takeaways

- **Career Pathways**: Data science offers various career opportunities, ranging from data analysts to more specialized roles like machine learning engineers.
- **Recruitment**: Organizations value both technical skills and the ability to solve complex problems when recruiting data science talent.
- **Importance of Math & Stats**: A solid foundation in mathematics and statistics is essential for success in data science.
- **Communication**: Structuring reports and effectively communicating insights is critical to data science work.

How Can Someone Become a Data Scientist?

High-end Data Scientists:

- Mostly PhDs are from physics, statistics, and computer science.
- Require strong mathematics, computer science, databases, statistics, and probability backgrounds.

Skills for New Data Scientists:

- Programming skills and computational thinking are essential.
- **Knowledge of algebra (up to analytics), geometry, basic calculus, probability, and statistics is required.**
- Understanding statistical distributions and relational databases is essential.
- Relational databases are a starting point for understanding **how to store and access data, which can extend to big data clusters.**

Advanced-Data Science Knowledge:

- Higher-level data scientists require **deep computer science theory, statistics, and probability knowledge.**
- Skills in **intersection areas of computer science and statistics** are crucial for advanced roles.

Self-Learning and Building:

- Self-learning is a vital part of becoming a data scientist.
- Building things and experimenting is an effective learning method, as experiences with **projects like HPC (high-performance computing) clusters demonstrate**.
- Online learning platforms like **IPython, Jupyter Notebooks, and Zeppelin offer practical learning opportunities.**

Motivation and Learning Platforms:

- Motivation is key to sustained learning.
- Badging systems like those used in Big Data University can help encourage continued learning.
- Success depends on personal drive and understanding one's goals.

Data Science and Corporate Structure:

- Data science teams should not be under CIOs (Chief Information Officers), who often have accounting/finance backgrounds and lack research understanding.
- Data science tends to thrive in companies with research agendas (e.g., pharmaceuticals, finance, technology companies).

Demand for PhD Data Scientists:

- High demand for **Ph.D.-level data scientists, with companies like Facebook, Linkedin, Uber, and Lyft hiring them for complex problem-solving tasks.**
- These roles offer lucrative pay and engaging challenges, such as optimizing Uber's scheduling systems.

Recruiting for Data Science

General Insights on Hiring Data Scientists:

1. **Unicorn Skill Set is Rare**:
   - Ideal candidates with all technical, domain-specific, storytelling, and presentation skills are rare.
   - Focus on finding candidates that **resonate with the company's DNA**.
2. **Prioritize Passion and Domain Interest**:
   - Look for candidates excited about the specific data and domain you work in.
   - Passion for the domain **increases productivity more than purely technical skills.**
3. **Curiosity, Humor, and Soft Skills Matter**:
   - Look for **curiosity** about everything, not just data.
   - Candidates should have a **sense of humor** to handle challenges with a lighthearted approach.
   - **Social skills** and the ability to tell a story from data are critical.
   - **Technical skills can be taught, but curiosity and storytelling cannot.**
4. **Technical Skills Shouldn't be the Primary Focus**:
   - Prioritize curiosity, sense of humor, and the ability to communicate findings.
   - If technical skills are present, they can be further developed.
5. **No One-Size-Fits-All Rule**:
   - Hiring varies case by case. The role and company needs dictate what specific

skills and personality traits are important.

Skills to Look for in Data Science Candidates:

1. **Problem Solving and Analytical Thinking**:
   - **Strong math and statistics background.**
   - Ability to approach and analyze problems effectively.
2. **Data Manipulation and Communication**:
   - Technical competence with data manipulation and analysis.
   - Ability to explain data findings in a compelling and relatable way.
3. **Cross-Department Collaboration**:
   - Data scientists must build relationships and communicate across different company departments.

Role-Specific Skills and Tools:

1. **Structured vs. Unstructured Data**:
   - Market research and structured data require knowledge of **basic statistical algorithms, machine learning, and tools like R, Stata, or Python.**
   - **Unstructured data is better suited for Python**.
2. **Big Data Environments**:
   - For big data roles, expertise in storing and manipulating large data sets is essential.
   - Tools include **Hadoop** and **Spark** for big data environments.
3. **Technical Platforms**:
   - For structured predictive analytics: R, Stata, Python.
   - For unstructured data: Python.
   - For big data: Hadoop, Spark.

Communication and Presentation Skills:

1. **Storytelling with Data**:
   - Candidates should be able to turn analysis into a compelling story.
   - Presentations and documents should synthesize data findings clearly and engagingly.
2. **Creating Impact**:
   - Findings should be presented in a way that surprises and enlightens the

audience, giving them new insights and ideas for action.

Careers in Data Science

The Growth of Data Science:

1. **Rise of Data Science Due to IoT and Distributed Computing**:
   - The Internet of Things (IoT) and advances in distributed computing have led to vast amounts of data and the capability to analyze it effectively.
   - Data science focuses on extracting insights from this data and using it to drive actionable outcomes.
2. **Top Career Choice**:
   - Data science has been consistently ranked as a top career since 2016, reaching the number one spot and remaining in the top three by 2020.
   - Job postings for data scientists span across various industries, not just tech, showing its broad applicability.
3. **Market Growth**:
   - The data science platform market **is predicted to grow by $314.8 billion by 2025, with a compounded growth rate of 38.2%**, according to Global Industry Analysts Incorporated.
4. **Talent Shortage**:
   - McKinsey Global Institute warned of **talent shortages in data and analytics by 2018.**
   - The demand for data scientists continues to rise, making it difficult for recruiters to fill positions.

Motivation and Skills for Entering Data Science:

1. **Applicability Across Disciplines**:
   - Data science can be applied to almost any discipline, making it an attractive field for those who enjoy working with data, coding, and statistics.
2. **Essential Skills**:
   - Required skills include data analysis, coding, math, statistics, and storytelling.
   - Continuous learning is crucial, as new tools and techniques are always emerging in the field.
3. **Women in Data Science Initiative**:
   - The Stanford Institute for Computational and Mathematical Engineering

launched an initiative to inspire and educate women in data science globally, supporting gender diversity.

Preparing for a Career in Data Science:

1. **Tailoring Skill Set**:
   - Aspiring data scientists should **match their skill set to the specific role they are targeting.**
   - **Missing skills can be acquired through online training resources.**
2. **Educational Resources**:
   - The availability of diverse educational platforms makes entering the data science field more accessible than ever.
3. **Rewarding Career**:
   - Data science offers a fascinating and rewarding career with broad applications and plenty of opportunities for growth.

Importance of Maths and Statistics for Data Science

Developing Skills for Data Science:

1. **Start Early with Programming and Math**:
   - Learn **programming, math, probability, and statistics.**
   - Hands-on experience (building systems, writing code) helps identify knowledge gaps and drive learning.
   - Early exposure to these skills, even in high school, provides a head start in college and future careers.
2. **Fostering Curiosity and Creativity**:
   - Curiosity is key in data science—engaging in activities **like detective games or treasure hunts helps build these traits.**
   - Encourage experimentation, similar to science fairs, but with data sets instead of physical projects (e.g., polls, surveys).
   - Asking questions and applying the scientific method to data enhances problem-solving skills.
3. **Learn Databases and SQL**:
   - High school students should familiarize themselves with databases and SQL.

- Taking computer science courses early on is helpful for building foundational skills.

Encouraging a Career in Data Science:

1. **High Demand for Data Scientists**:
   - Data science is a knowledge profession in high demand across industries, as companies need **help with efficiency and smart decision-making.**
   - Data scientists play a crucial role in **business growth and innovation, ensuring a future demand for these skills.**
2. **Math Skills Can Improve with Application**:
   - Even if math isn't a strength, data science allows for practical application, making math easier to grasp.
   - Understanding the real-world impact of math on problems and people makes it more meaningful and engaging.
3. **The Fun Aspect of Data Science**:
   - Data science is enjoyable for those with a natural curiosity and interest in solving puzzles and answering complex questions.

Career Outlook and Advice:

1. **Promising Career Path**:
   - Data science offers a rewarding and enjoyable career with opportunities for personal growth and high earnings.
   - **Encourage anyone with an interest in data science to pursue it, as it will be a valuable and sought-after profession in the future.**
2. **Hands-On Learning**:
   - Building systems, experimenting with real data, and continuously learning are essential to developing expertise in data science.

The Report Structure IBM LAB

Report Length and Purpose:

- A brief report (5 pages or fewer) presents key findings concisely.
- A detailed report (100+ pages) includes thorough data analysis, research

methodology, and references.
    - Reports from firms like Deloitte or McKinsey vary depending on the purpose—commentaries are shorter, while in-depth analyses are more comprehensive.
2. **Essential Components of Any Report**:
    - **Cover Page**: Must include the title, authors, affiliations, contact information, publisher, and date of publication. It is often neglected but essential for professionalism.
    - **Table of Contents (ToC)**: Provides a "map" of the document for the reader, listing headings, tables, and figures, especially important for longer reports.
3. **Executive Summary/Abstract**:
    - Even in short documents, an abstract or executive summary summarizes key arguments in a few paragraphs.
    - For larger reports, the executive summary can be more detailed, providing a snapshot of the main findings.
4. **Introduction and Literature Review**:
    - The **introduction** frames the problem for readers new to the subject, providing context.
    - The **literature review** highlights relevant research and identifies knowledge gaps that the report aims to address. Its length depends on the complexity of the subject matter.
5. **Methodology Section**:
    - Details the research methods and data sources used.
    - Justify your choice of variables, data, and methods by referring to the literature.
6. **Results and Analysis**:
    - Present findings with **descriptive statistics** and **illustrative graphics** (e.g., plots, maps).
    - If using statistical models, such as **regression** or **categorical analysis**, refer to relevant chapters in the textbook for guidance.
    - Results should be presented clearly, sometimes using graphics to simplify complex statistical findings.
7. **Discussion Section**:
    - Build your main arguments based on the presented results.
    - Address the research question and knowledge gaps, showing how your findings contribute to solving the problem.
    - Acknowledge limitations and any caveats in the data or results.
8. **Conclusion**:
    - Summarize the key findings and market them to the reader, emphasizing their importance.

    - Identify potential future research areas or applications stemming from your findings.
9. **Final Sections**:
    - Include **references**, **acknowledgments** (to thank contributors), and **appendices** (for supplementary material).

Checklist for Effective Writing:

1. **Reader Engagement**: Have you explained the purpose and significance of your work from the beginning?
2. **Clarity of Aim**: Is the aim of your report clear?
3. **Contribution Significance**: Have you explained the value of your contribution?
4. **Context**: Have you provided sufficient background and references?
5. **Practicality**: Have you addressed the usefulness and practicality of your findings?
6. **Future Developments**: Have you suggested future developments based on your research?
7. **Logical Structure**: Is your report structured logically and clearly?

By following this structured approach, your data science reports will be more effective and professional.

Lesson Summary: Careers and Recruiting in Data Science

1. **Skills vs. Team Dynamics**:
    - Companies often seek a broad set of skills in a single data scientist, including domain-specific knowledge, data analysis, storytelling, and proficiency with both structured and unstructured data.
    - It's rare to find one person with all these skills. Instead, companies should build teams with diverse skills and potential for growth.
2. **Passion and Curiosity**:
    - Passion for the industry and curiosity about the data are crucial traits. Data scientists should be excited about the field they work in and motivated to ask meaningful questions.
3. **Self-Learning and Tinkering**:
    - Data scientists should enjoy experimenting with data, creating visualizations, and continually learning new skills.
4. **Technical Skills**:
    - **Strong background in mathematics, statistics, and probability is essential for data analysis.**
    - Proficiency in programming languages like Python and R is important.

    - Understanding data storage, retrieval systems, and machine learning algorithms is necessary.
5. **Communication and Storytelling**:
    - Data scientists need strong communication skills to present findings clearly and engagingly.
    - Reports should be well-organized and provide clear goals, significance, context, and future implications.
    - Effective storytelling helps make the data's insights impactful and memorable.
6. **Team Composition**:
    - Teams should include individuals with expertise i**n statistics, mathematics, machine learning, and programming, as well as those with the ability to tell compelling stories with data.**

By focusing on these aspects, companies can build effective data science teams and enhance their data-driven decision-making processes.

Summary: Careers and Recruiting in Data Science

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

- Data Science helps physicians provide the best treatment for their patients , helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
- Companies can start on their data science journey by capturing data. Once they have data, they can begin analyzing it.
- Everyone who uses the Internet generates mass amounts of data daily.
- Amazon and Netflix use recommendation engines, and UPS uses data from customers, drivers, and vehicles to use the drivers' time and fuel efficiently.
- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.
- The report should present a thorough analysis of the data and communicate the project findings.
- Companies should look for someone excited about working with the data in their particular industry. They should seek out someone curious who can ask interesting, meaningful questions about the types of data they intend to collect. They should hire people who love working with data, are fluent in statistics, and are competent in applying machine learning algorithms.
- A clearly organized and logical report should communicate the following to the reader:

- What they gain by reading the report
- Clearly defined goals
- The significance of your contribution
- Appropriate context by giving sufficient background
- Why this work is practical and useful
- Conjecture plausible future developments that might result from your work

A Roadmap to your Data Science Journey

Case Study: Final Assignment

1. Education and Skill Acquisition:

- **Background:** Lila has an undergraduate degree in economics and a background in data analysis.
- **Additional Learning:** To transition into data science, she enrolled in the IBM Data Science Professional Certificate online program. This included coursework in statistics, machine learning, data analysis, and programming languages like Python and SQL.

2. Building a Strong Foundation:

- **Skills Development:** Lila gained proficiency in data manipulation and visualization using Python libraries (NumPy, Pandas, Matplotlib).

3. Visualization for Storytelling:

- **Effective Communication:** She learned to create informative visualizations to represent data such as sales trends and customer segmentation, which aids in storytelling and decision-making.

4. Hands-On Experience:

- **Practical Application:** Lila participated in Kaggle competitions, worked on personal data projects, and built her GitHub profile to showcase her skills.

5. Data Wrangling and Preprocessing:

- **Data Cleaning:** She spent time cleaning and preprocessing data, handling missing values, outliers, and feature engineering to improve model performance.

6. Communication and Storytelling:

- **Presentation Skills:** Lila honed her data storytelling skills, using tools like Matplotlib and Plotly to present insights clearly and compellingly.

7. Networking and Collaboration:

- **Community Engagement:** She participated in data science communities, attended conferences, and collaborated on open-source projects.

8. Domain Expertise:

- **Choosing a Niche:** Lila decided to focus on e-commerce due to her background in economics and interest in the field.

9. Landing the First Job:

- **Job Application:** She tailored her resume, highlighted relevant skills and projects, and showcased her work through an online portfolio.

10. Approach to First Task:

- **Dataset Selection and Sourcing:**
  - Lila selected and integrated datasets from various sources, seeking input from product professionals and data engineers.
- **Data Understanding and Cleaning:**
  - Imported data into Python and SQL, cleaned the data by addressing missing values and outliers.
- **Exploratory Data Analysis (EDA):**
  - Conducted EDA to understand customer behavior and sales trends using summary statistics and visualizations.
- **Feature Engineering:**
  - Created new features to enhance the dataset's utility.
- **Statistical Analysis and Machine Learning:**
  - Employed regression analysis and explored machine learning models for tasks like demand forecasting and customer segmentation.
- **Presentation and Reporting:**
  - Compiled findings into a comprehensive report and presentation using Jupyter Notebook, providing actionable insights to stakeholders.

11. Continuous Learning:

- **Further Education:** Lila plans to further develop her machine learning skills by pursuing the IBM Machine Learning Professional Certificate and experimenting with various algorithms.

This case study outlines Lila's educational journey, skill development, and approach to her first task as a data scientist, emphasizing continuous learning and practical experience.

Review and Evaluate a Data Scientist Job Post

Assignment Overview

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

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

Identify the following aspects of data science job post:

1. What is the company name that is advertising the job?
2. What is the job title?
3. Where is the role located?
4. What is the expected salary or salary range?
5. What is the total number of results from the search for the job post?
6. What is one technical responsibility from the job post related to something you learned about in this course?
7. What are two required technical skills from the job post?
8. What are at least two ideas or concepts you learned about in this course relevant to these job posts?

Congrats & Next Steps

Congratulations on completing this course! We hope you enjoyed it.
As you have learned, data science is an emerging field in high demand. To meet these, data scientists are expected to combine soft skills like curiosity and the knowledge of data science tools.
If you aspire to become a data scientist, we highly encourage you to complete the optional Honors project at the end of the course and explore jobs in this field.
If you want to know more about data science, we encourage you to take the next step by pursuing either Tools for Data Science course or IBM Data Science Professional Certificate. The introductory course for both of these programs is this: What is Data Science? course, which you have now successfully completed.
We also encourage you to leave your feedback and rate this course so that we can continue to improve our content.
Good luck!