201A FOUNDATIONS OF DATA SCIENCE NOTES:

Table of Contents

[Introduction 2](#_Toc129277417)

[What is Data Science 2](#_Toc129277418)

[Job Opportunities 2](#_Toc129277419)

[Assessments 3](#_Toc129277420)

[AI Correlations: 4](#_Toc129277421)

[Skills required for DATA SCIENCE 5](#_Toc129277422)

[Data Science Process: 5](#_Toc129277423)

[Cloud Services 6](#_Toc129277424)

[Data 7](#_Toc129277425)

[Data Types 7](#_Toc129277426)

[Data Representation 7](#_Toc129277427)

[Statistical Data Extraction 7](#_Toc129277428)

[Big Data 7](#_Toc129277429)

[Computational resources 8](#_Toc129277430)

[Digital Transformation 8](#_Toc129277431)

# Introduction

## What is Data Science

Data science is an interdisciplinary field that involves the extraction of knowledge and insights from structured and unstructured data using statistical, computational, and machine learning techniques. It combines elements of statistics, mathematics, computer science, and domain-specific knowledge to uncover patterns and insights that can help businesses make better decisions, identify opportunities, and solve complex problems.

Data science involves several stages, including data collection, data cleaning, data transformation, data analysis, and data visualization. The process often requires working with large and complex datasets, and the use of various tools such as programming languages (e.g., Python, R), statistical packages, databases, and data visualization software.

The applications of data science are diverse and widespread, ranging from predicting customer behaviour and detecting fraud, to improving healthcare outcomes and optimizing supply chain operations. As the amount of data generated by individuals and businesses continues to grow, the importance of data science in many industries is likely to increase.

In conclusion Data science involves the collection, analysis, and organization of structured, unstructured, and semi-structured data. This is done using statistical, computational, and machine learning techniques, which the goal is to extract knowledgeable insights from the data, which can be used to solve complex problems. By applying these insights, organizations can make better decisions and create predictive models.

## Job Opportunities

* Data Analyst - Analyses data using statistical methods to extract insights and inform decision-making.
* Data Scientist - Applies statistical and machine learning techniques to analyse complex data sets and develop predictive models.
* Data Engineer - Designs and builds data pipelines, databases, and other infrastructure to support data analysis.
* Big Data Engineer - Focuses on managing and analysing large volumes of data using distributed systems like Hadoop and Spark.
* Database Administrator - Manages and maintains databases to ensure that data is stored securely and efficiently.
* Data Visualization Engineer - Creates visualizations and dashboards to help people understand complex data sets.
* Market Research Analyst - Conducts research and analyses data to understand market trends and inform business decisions.
* ETC. - There are many other roles in the data science field, including machine learning engineer, business intelligence analyst, data architect, and more.

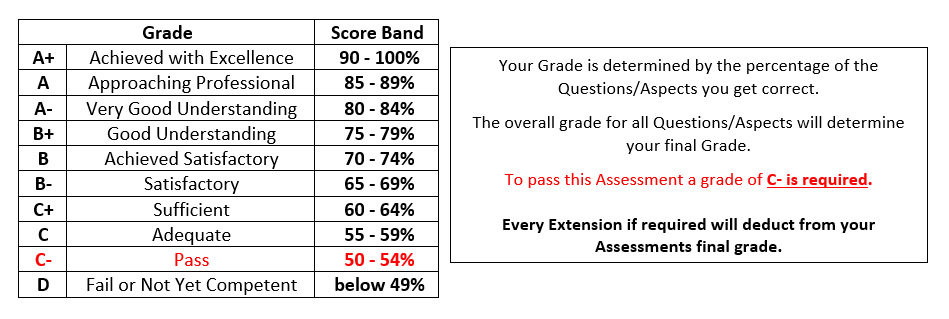
## Assessments

Two assessments:

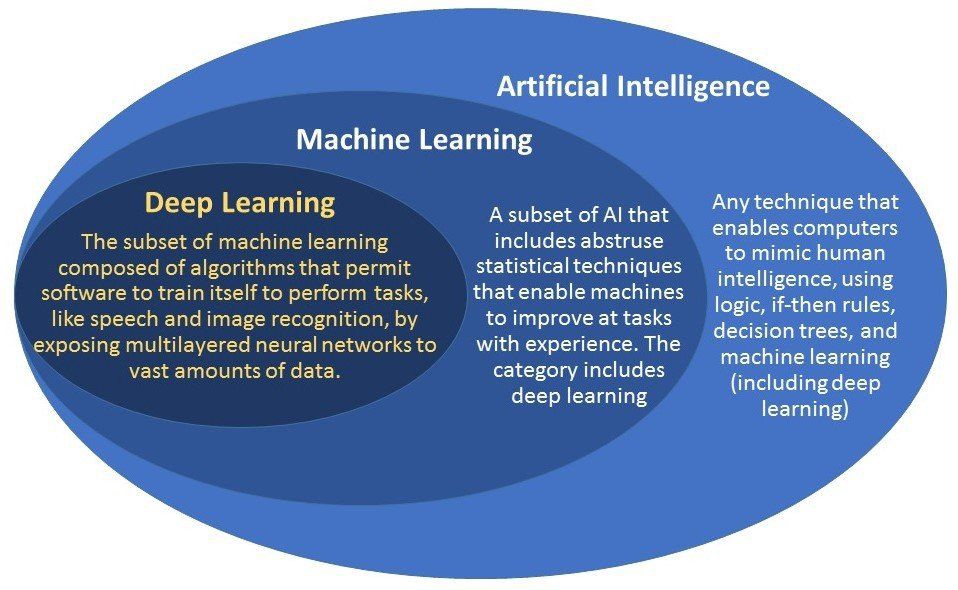
1. Theory test: Data related theory, e.g. Means, deviation, algorithms. (60%)
2. Math (60%)

Data analysis and modelling (40%)

1. Practical assignment: solving a business problem using data science tools (40%)



## AI Correlations:



Artificial Intelligence (AI) refers to the use of computer techniques that enable machines to mimic human intelligence. These techniques may include logic-based rules, decision trees, and machine learning (including deep learning). By using these methods, machines can learn from data and make decisions or predictions without being explicitly programmed to do so. The goal of AI is to create intelligent machines that can perform tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, and language translation.

Machine learning is a subset of AI that involves using statistical techniques to enable machines to learn from data and improve their performance on a specific task over time. This can be done using various algorithms and methods, such as decision trees, neural networks, and support vector machines. Deep learning is a specific type of machine learning that uses artificial neural networks to analyse and learn from complex data sets. By using these methods, machines can automatically identify patterns and make predictions or decisions without being explicitly programmed to do so. The ultimate goal of machine learning is to create intelligent machines that can learn and adapt on their own.

Deep learning is a subset of machine learning that involves using algorithms to train artificial neural networks with multiple layers. By exposing these networks to vast amounts of data, deep learning enables software to perform complex tasks such as speech and image recognition. Deep learning algorithms can automatically identify patterns and features within the data, making them highly effective at tasks that would be difficult or impossible to program manually. The ultimate goal of deep learning is to create software that can learn and improve on its own, without the need for human intervention.

In conclusion, AI is a broad field of computer science that involves creating machines that can perform tasks that traditionally require human intelligence. Machine learning is a subset of AI that focuses on the development of algorithms that can learn from data, while deep learning is a specific type of machine learning that uses artificial neural networks to analyse and learn from complex data sets. Together, AI, machine learning, and deep learning represent different approaches to creating intelligent machines that can perform tasks that are challenging for humans. These fields are closely related and build upon one another to achieve the common goal of creating intelligent machines.

### Skills required for DATA SCIENCE

1. Curious
2. Argumentative
3. Judgemental
4. Analytica platform
5. Communicate

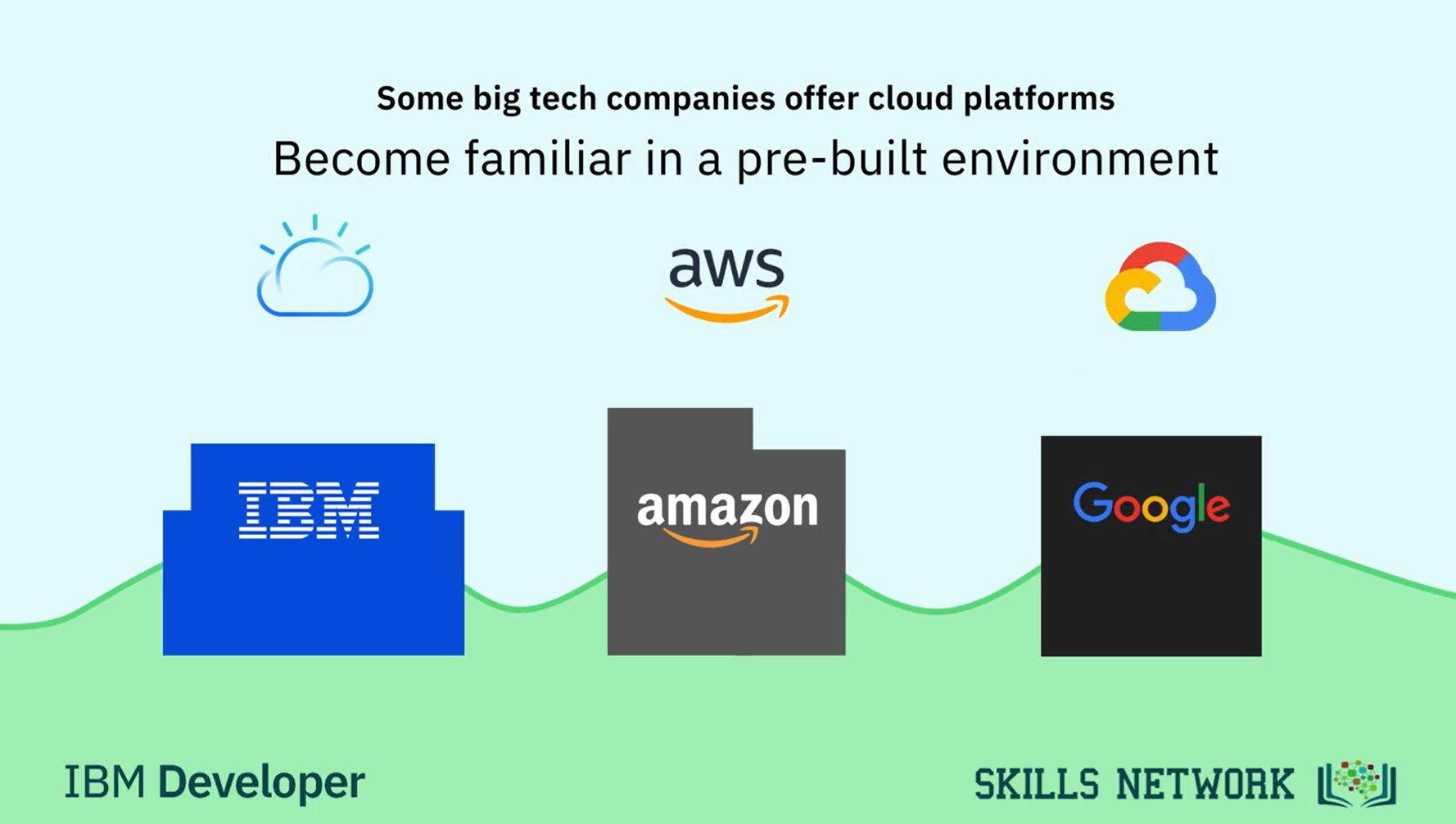
## Data Science Process:

Diagram

Description automatically generated

1. Define the problem - Identify the problem you want to solve or the question you want to answer.
2. Collect data - Gather data relevant to the problem you are trying to solve.
3. Explore the data - Analyse the data to identify patterns, trends, and relationships. This may involve using visualization tools to help understand the data.
4. Clean and pre-process data - Clean the data and remove any irrelevant or inconsistent data points.
5. Analyse the data - Apply statistical and machine learning techniques to the data to identify patterns and trends. If the data does not show any clear trends, consider reprocessing the data or collecting additional data.
6. Build models and algorithms - Develop models and algorithms that can make predictions or classify new data points based on the patterns identified in the data.
7. Visualize and report - Present the results of the analysis in a clear and understandable way using visualizations and reports.
8. Draw conclusions and make recommendations - Use the results of the analysis to draw conclusions about the problem you are trying to solve and make recommendations for future action.
9. Deploy the model or solution - Implement the solution or model in real-world applications to solve the problem.

## Cloud Services



Cloud services such as Google Cloud, Amazon Web Services (AWS), and IBM Cloud are important in data science for several reasons:

Scalability - Cloud services can provide access to virtually unlimited computing power and storage, allowing data scientists to work with large and complex data sets that might otherwise be difficult or impossible to process on local hardware.

Flexibility - Cloud services allow data scientists to experiment with different tools, technologies, and configurations without the need for large upfront investments in hardware or software.

Collaboration - Cloud services provide a platform for collaboration and data sharing across teams and organizations, allowing data scientists to work together more effectively.

Cost-effectiveness - Cloud services offer a cost-effective alternative to building and maintaining local computing infrastructure. This can be particularly important for smaller organizations or startups that may not have the resources to invest in large-scale hardware.

Security - Cloud service providers typically have robust security measures in place to protect data from unauthorized access or loss.

Overall, cloud services are important in data science because they provide a powerful and flexible platform for data processing, analysis, and collaboration, while also offering cost savings and improved security.

# Data

## Data Types

Raw Data could be any of the following…

* Numeric (continuous vs discrete)
* Categorical
* Ordinal

## Data Representation

* Tabular
* Samples and Features
* Variables (dependant or independent)
* (<https://archive.ics.uci.edu/ml/datasets/Multiple+Features>)

## Statistical Data Extraction

* Very useful in extracting initial data structure.
* Could be basic or very complex.
* Average: Mean, median, mode
* Variance
* Percentile
* Covariance (Cov(X, Y) = Σ(Xi-µ)(Yj-v) / n) (<https://byjus.com/maths/covariance/>)

## Big Data

Big data refers to extremely large and complex datasets that are difficult to process and analyse using traditional data processing tools and techniques. These datasets can come from a variety of sources, such as social media, IoT devices, scientific research, and business transactions.

* Volume - Big data refers to datasets that are so large that traditional data processing tools and techniques are insufficient to handle them.
* Velocity - Big data is often generated and updated rapidly, requiring fast and efficient processing techniques to keep up.
* Variety - Big data can come in many different formats and structures, including structured, semi-structured, and unstructured data, making it difficult to process and analyse using traditional methods.
* Veracity - Big data may include data that is incomplete, inaccurate, or inconsistent, making it difficult to trust and analyse.
* Value - Big data can provide valuable insights and enable new applications and services, making it an important resource for organizations in many fields.

To manage and analyse big data, new technologies and tools have emerged, such as Hadoop, Spark, and NoSQL databases. These technologies are designed to handle the large volume, high velocity, and varied structure of big data.

Overall, big data represents a major challenge and opportunity for organizations in many fields, as it can provide valuable insights and enable new applications and services. By using advanced technologies and techniques to analyse big data, organizations can gain a competitive edge and make better decisions.

## Computational resources

Computational resources refer to the computing power and hardware needed to process large and complex datasets. In the context of big data, which often involves processing large volumes of data in real-time, computational resources are a critical component of the data processing infrastructure.

Two popular tools for managing and processing big data are Apache Spark and Hadoop. Apache Spark is an open-source, distributed computing system that provides a fast and efficient way to process large datasets. It is designed to work with a variety of data sources, including Hadoop Distributed File System (HDFS), Amazon S3, and Cassandra. Spark provides a flexible and scalable platform for processing data in parallel across clusters of computers, making it ideal for handling big data workloads.

Hadoop is another open-source software framework designed for distributed storage and processing of large datasets. It includes two main components: Hadoop Distributed File System (HDFS), which provides a distributed file system for storing and managing large datasets, and MapReduce, which is a programming model for processing large data sets in parallel across clusters of computers. Hadoop is designed to be fault-tolerant and scalable, making it an ideal platform for processing big data.

Both Apache Spark and Hadoop provide powerful tools for managing and processing big data, but they have different strengths and use cases. Spark is often used for real-time data processing and machine learning, while Hadoop is well-suited for batch processing of large datasets. Ultimately, the choice of which tool to use will depend on the specific needs of the data processing project.

## Digital Transformation

Digital transformation refers to the process of using digital technologies to transform business operations and create new value propositions. This includes using data and analytics to gain insights into customer behaviour and business operations, and using digital technologies to improve efficiency and create new products and services.

Big data and data science are important components of digital transformation, as they enable organizations to collect and analyse vast amounts of data in order to gain insights and make data-driven decisions. Here are some examples of how big data and data science relate to digital transformation:

Customer analytics: By analysing customer data, such as purchasing history and online behaviour, organizations can gain insights into customer preferences and behaviour, allowing them to better understand and target their customers.

Supply chain optimization: By analysing data on suppliers, inventory levels, and transportation routes, organizations can optimize their supply chain operations, reducing costs and improving efficiency.

Predictive maintenance: By analysing data from sensors and other sources, organizations can predict when equipment will require maintenance or repair, allowing them to proactively address issues before they become critical.

Fraud detection: By analysing transaction data and other information, organizations can detect patterns and anomalies that may indicate fraudulent activity, helping to prevent financial losses.

Overall, big data and data science are essential components of digital transformation, as they enable organizations to make data-driven decisions and create new value propositions through the use of digital technologies.

# Probability Review

## Sample Spaces

To calculate the probability of an event occurring *P(event) = outcomes / total outcome.* But to fully understand how to do this, we have to first understand sample spaces.

Sample Space is