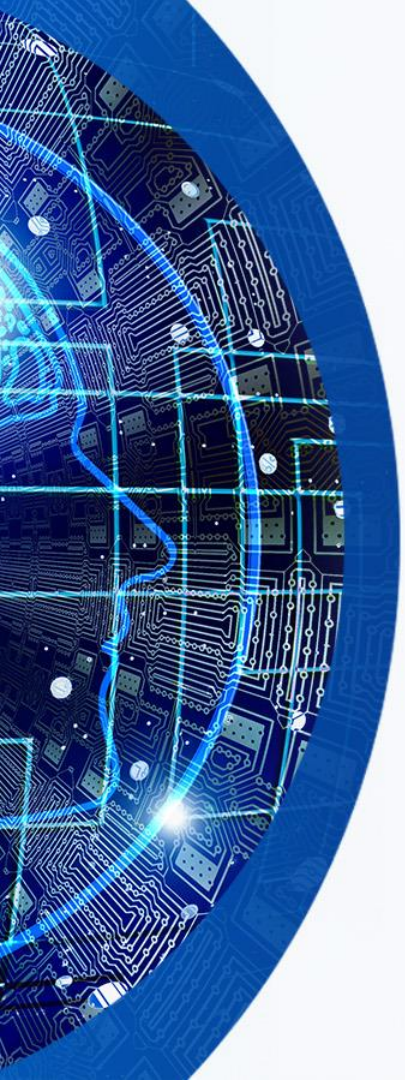
Data Science
Lecture 1

Prof. Dr. Rasha Ismail

Dr. Eman Amin

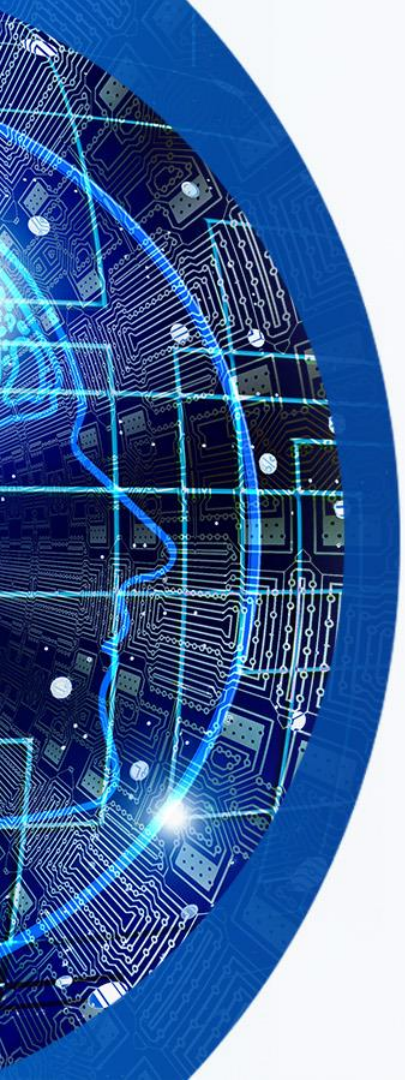


Agenda

- What is Data Science?
- Why We Need Data Science
- Examples of Data Science User Cases
- Who Uses Data Science Today?
- Aspects and Key Features of Data Science
- Data Science Associated Fields
- Data Scientist vs Data Analyst.
- Data Science Problems Classification
- Data Science Process
- Applications of Data science

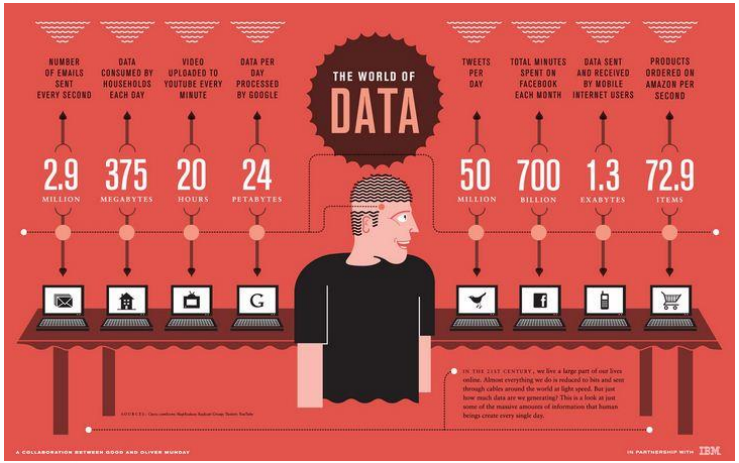
Data All Around

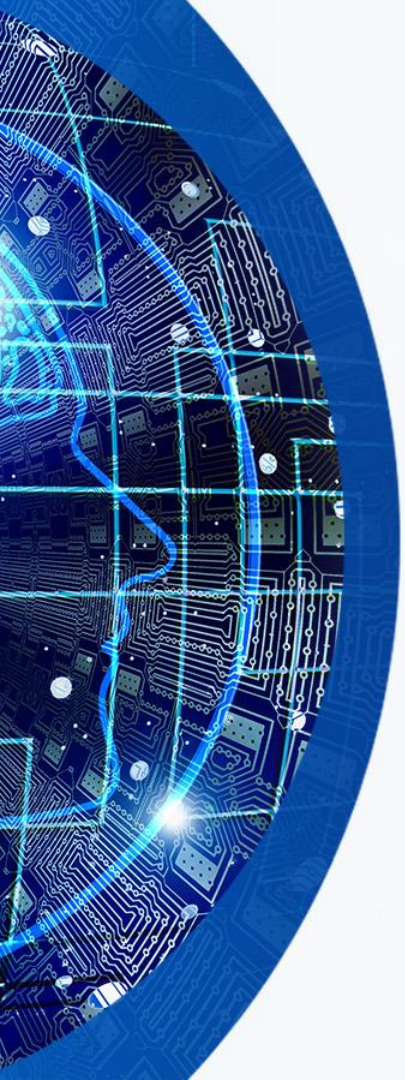
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network



How Much Data Do We have?

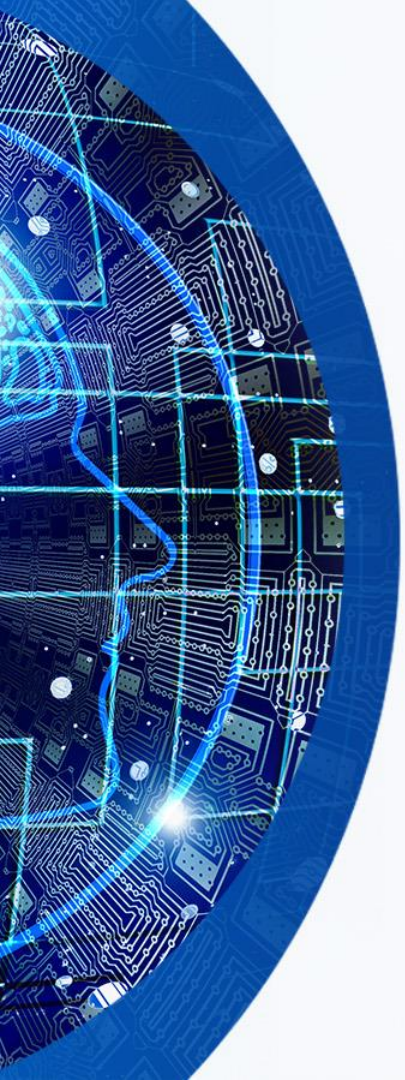
- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB





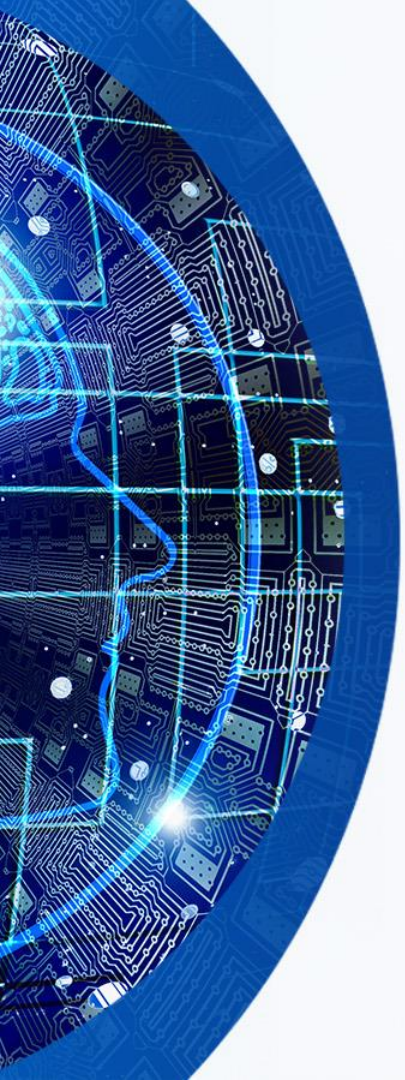
Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Graph Data
- Social Network,
- Streaming Data



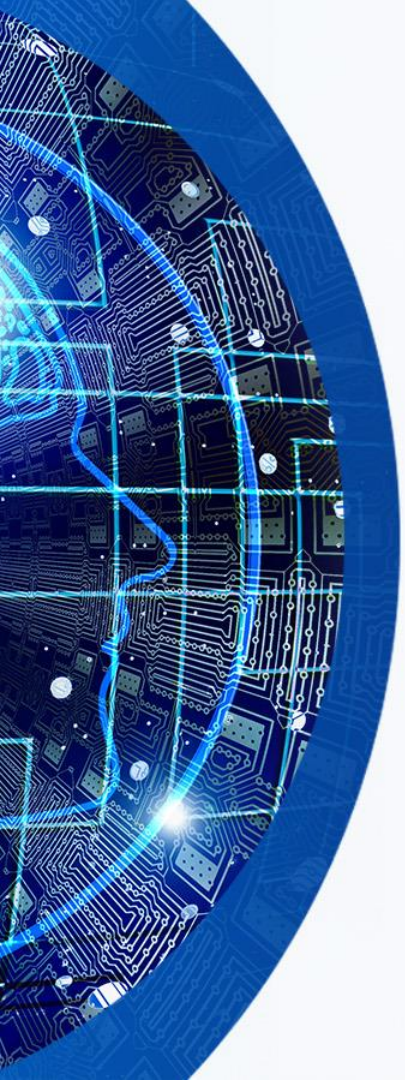
What To Do With These Data?

- Learn how to use data
  - **Explore**: identify patterns
  - **Predict**: make informed guesses
  - **Infer**: quantify what you know



What is Data Science?

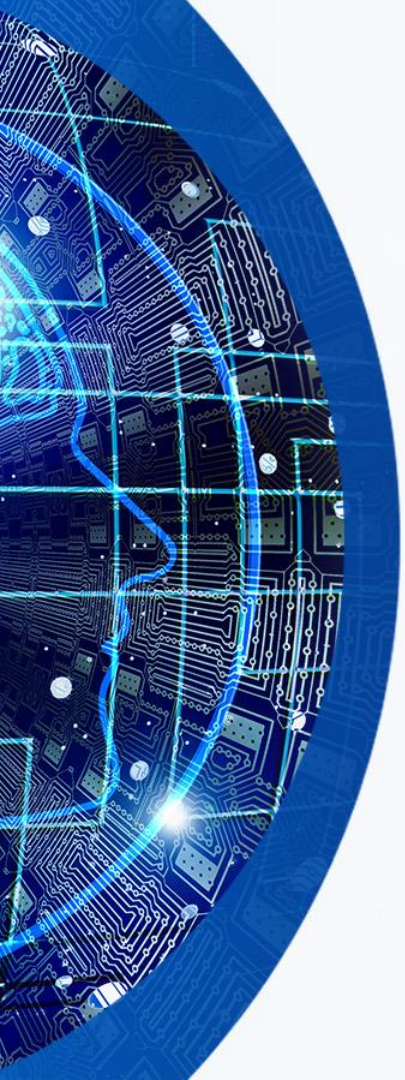
- Data science is a collection of techniques used to extract value from data.
- It has become an essential tool for any organization that collects, stores, and processes data as part of its operations.
- Data science techniques rely on finding useful patterns, connections, and relationships within data.



What is Data Science?

- Data science deals with processes and systems, that are used to **extract knowledge** or insights from **large amounts of data.**

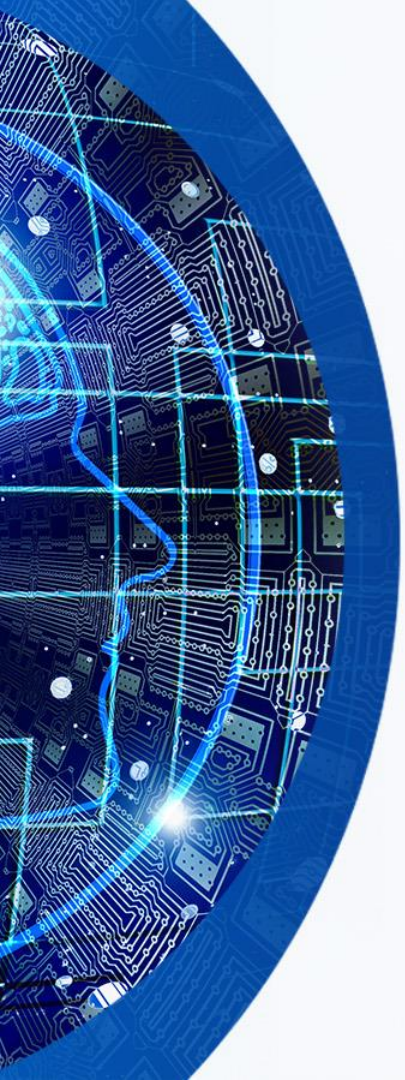
- Data science invokes methods from
  - Probability models
  - Machine learning
  - Data mining
  - Databases
  - Data visualization
  - Pattern recognition and learning,
  - Computer programming



What is Data Science?

- There is a wide variety of definitions and criteria for what constitutes data science.

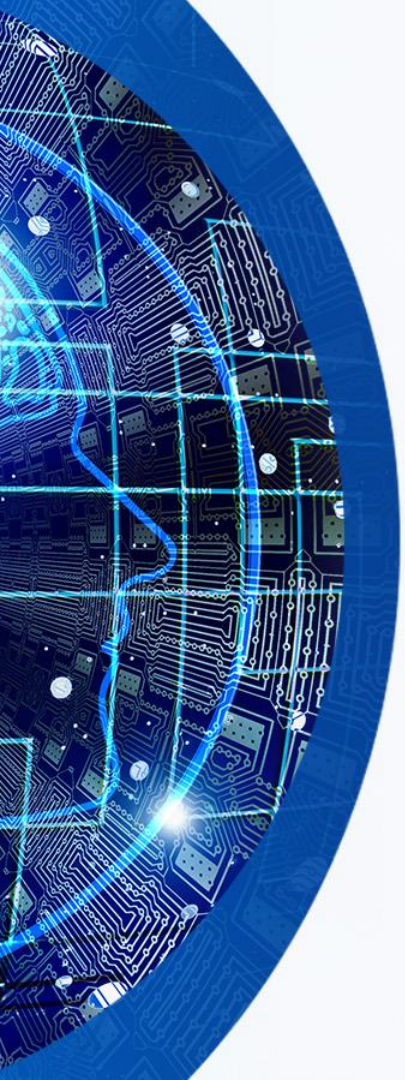
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining.
  - However, each term has a slightly different connotation depending on the context.



What is Data Science?

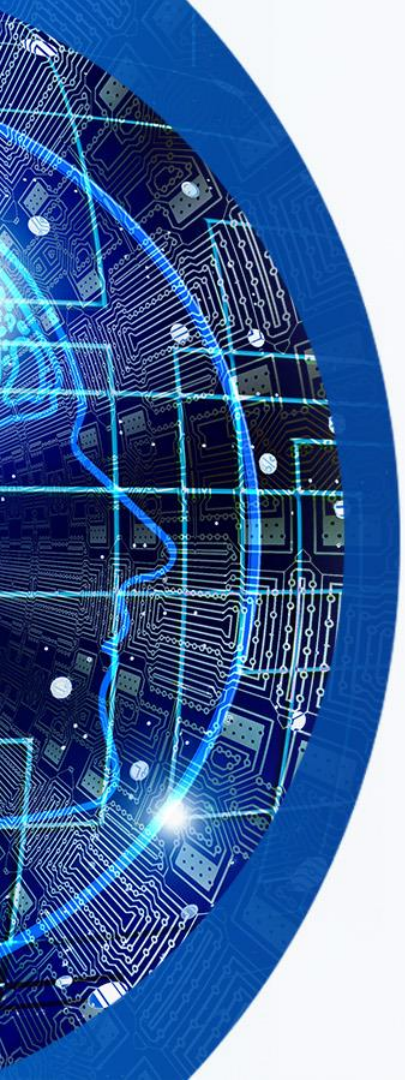
- Data science is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics.

- In the context of how data science is used today, it relies heavily on machine learning and is sometimes called data mining.



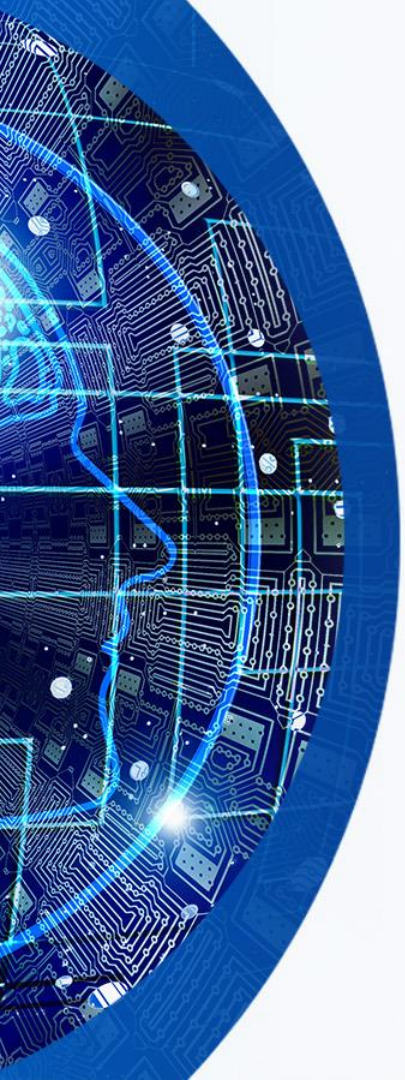
Why We Need Data Science

- Increased need to make data-driven decisions
- Better decisions increase quality of life, productivity and profitability



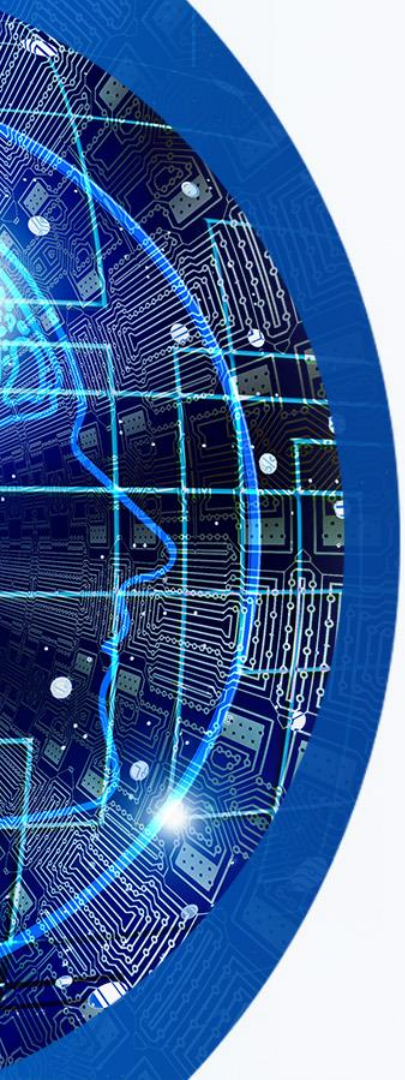
Examples of Data Science User Cases

- **Examples of data science user cases are**:
  - recommendation engines that can recommend movies for a particular user,
  - a fraud alert model that detects fraudulent credit card transactions,
  - find customers who will most likely churn next month,
  - or predict revenue for the next quarter.



Examples of Data Science User Cases

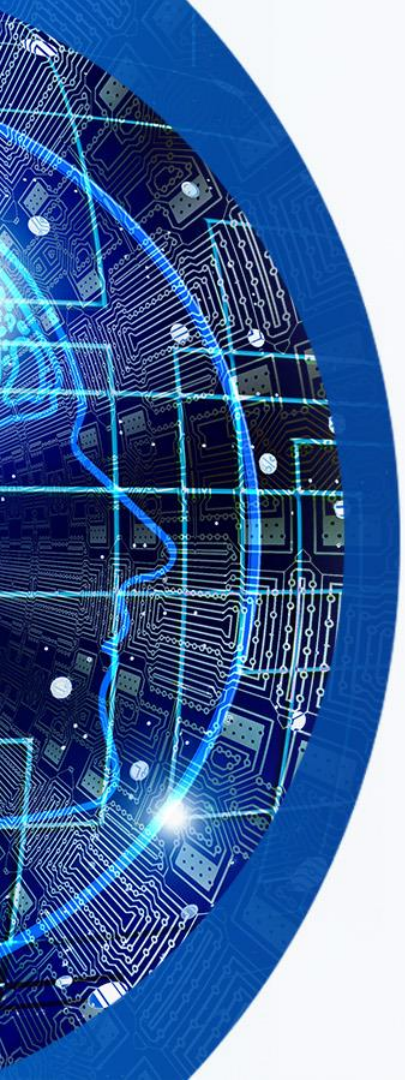
- **Examples of data science user cases are**:
  - many organizations like social media platforms, review sites, or forums are required to moderate posts and remove abusive content.
  - How can machines be taught to automate the removal of abusive content? The machines need to be shown examples of both abusive and non-abusive posts with a clear indication of which one is abusive.



Examples of Data Science User Cases

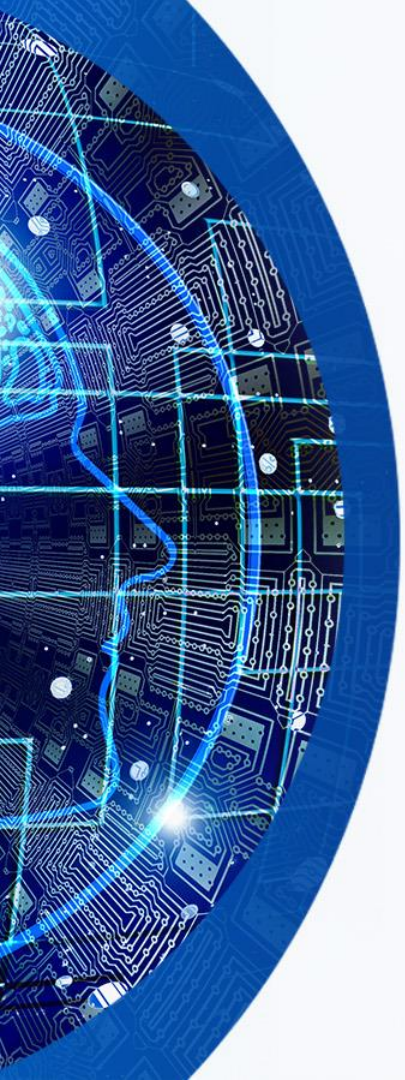
- The learners will generalize a pattern based on certain words or sequences of words in order to conclude whether the overall post is abusive or not.

- The model can take the form of a set of “if-then” rules. Once **the data science rules** or model is developed, machines can start categorizing the disposition of any new posts.



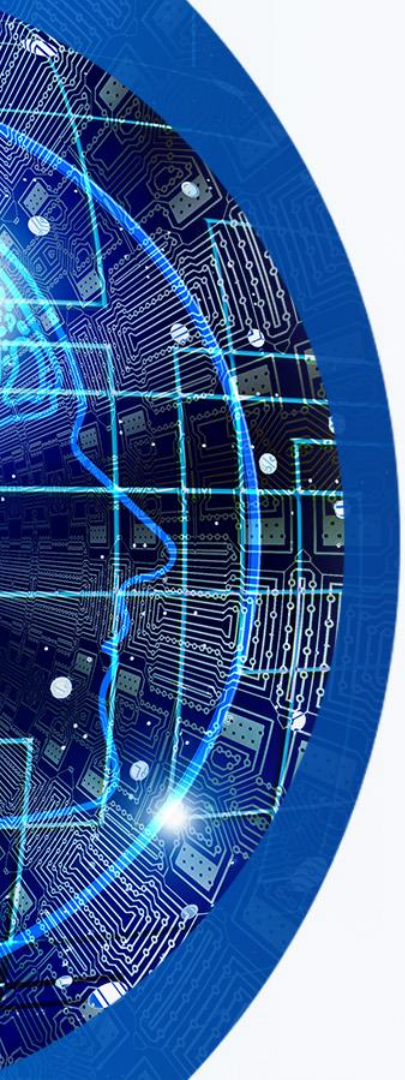
Who Uses Data Science Today?

- Almost every organization and business.
- The use of the term science in data science indicates that the methods are evidence based, and are built on empirical knowledge, more specifically historical observations.



Who Uses Data Science Today?

- As the ability to collect, store, and process data has increased and computing hardware capabilities double every two years, data science has found increasing applications in many diverse fields.
- Just decades ago, building a production quality regression model took about several dozen hours.
- Today, sophisticated machine learning models can be run, involving hundreds of predictors with millions of records in a matter of a few seconds on a laptop computer.

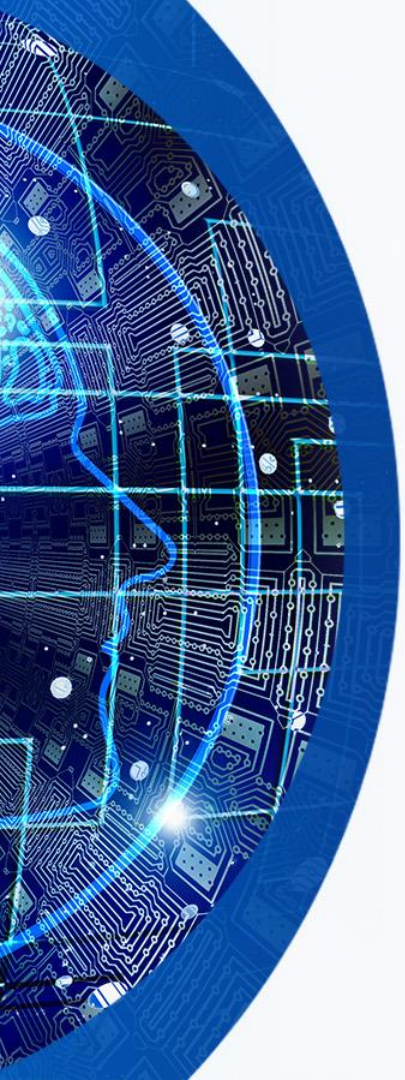


Aspects of Data Science

- The most important aspect of data science: interpreting the results of the analysis in order to make decisions.

- When it comes to the data science techniques, are there a core set of procedures and principles one must master?

  - It turns out that a vast majority of data scientists today use a handful of very powerful techniques to accomplish their objectives: decision trees, regression models, **deep learning**, and clustering

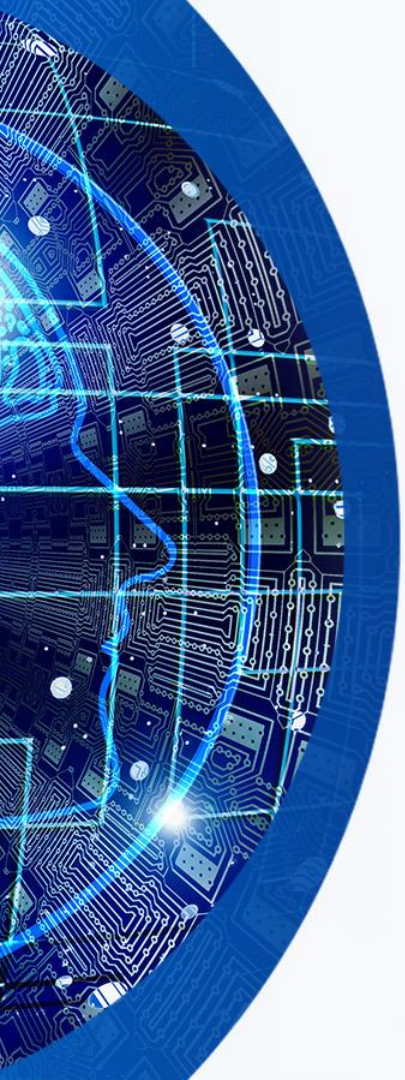


Aspects of Data Science

- Data science starts with data, which **can range from a simple array** of a few numeric observations to a complex matrix of **millions of observations** with thousands of variables.

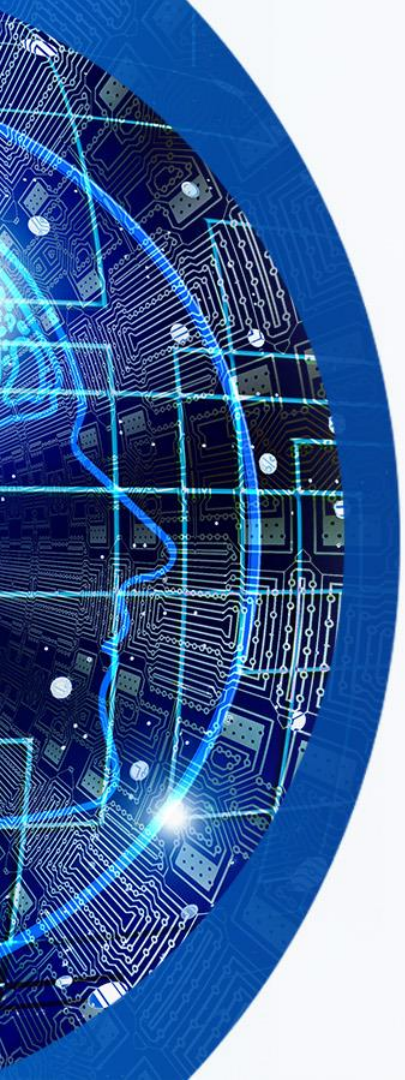
- Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset.

- The discipline of data science coexists and is closely associated with a number of related areas **such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI).**



Key Features of Data Science:
1. Extracting Meaningful Patterns

- Data science involves inference and iteration of many different hypotheses. One of the key features of data science is the process of **generalization of patterns from a dataset.**
  - The generalization should be valid, not just for the dataset used to observe the pattern, but also for new unseen data.

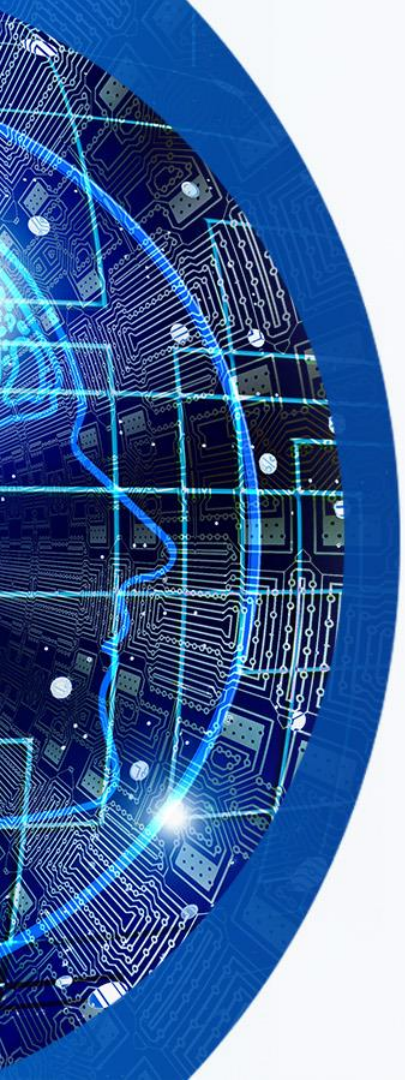


Key Features of Data Science:

2.Building Representative Models

- In statistics, a model is the representation of a relationship between variables in a dataset. It describes how one or more variables in the data are related to other variables.

- Modelling is a process in which a representative abstraction is built from the observed dataset.



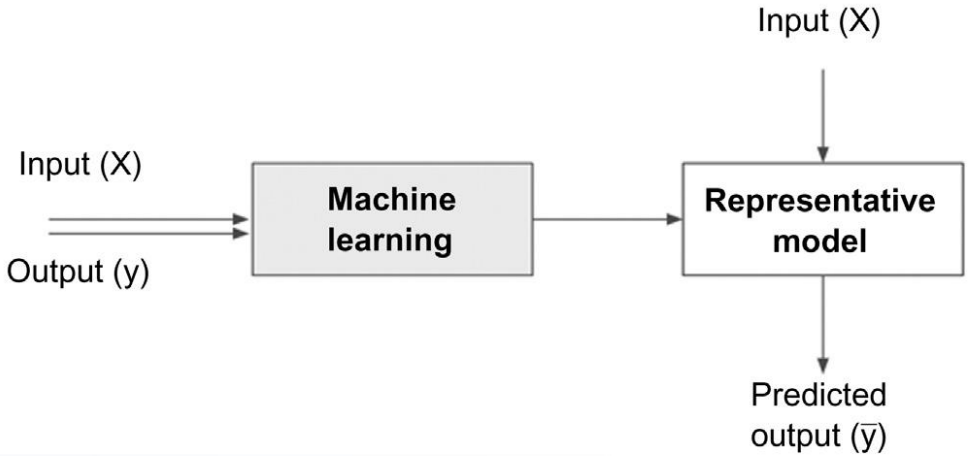
Key Features of Data Science:
2.Building Representative Models

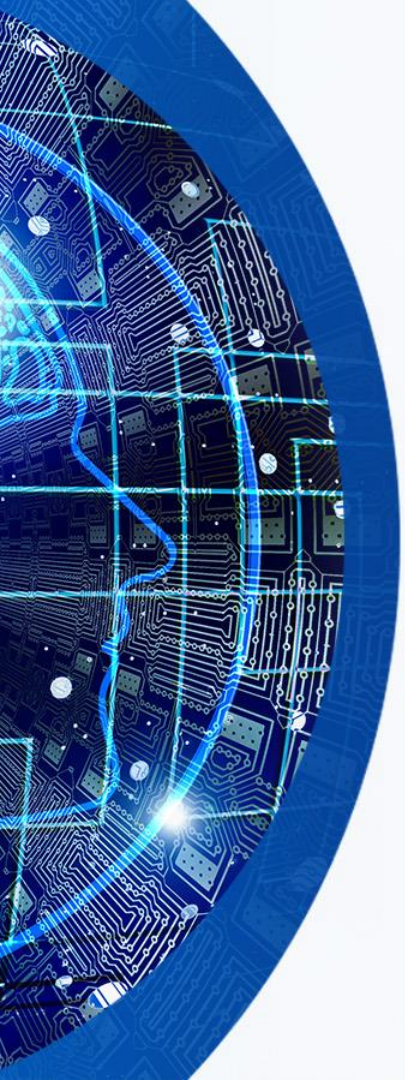
- For example, based on credit score, income level, and requested loan amount, a model can be developed to determine the interest rate of a loan.

- For this task, previously known observational data including credit score, income level, loan amount, and interest rate are needed.

Key Features of Data Science: 2.Building Representative Models

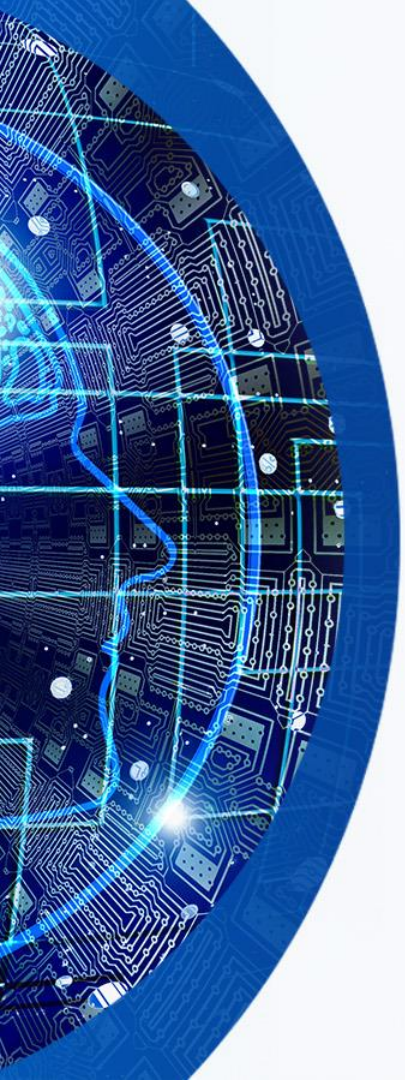
- This figure shows the process of generating a model. Once the representative model is created, it can be used to predict the value of the interest rate, based on all the input variables.





Key Features of Data Science: 2.Building Representative Models

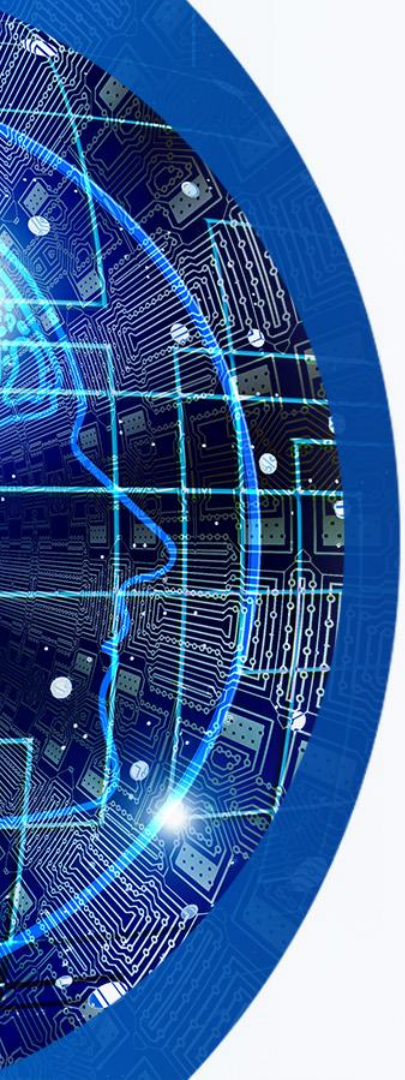
- The representative model serves two purposes:
  - it predicts the output (interest rate) based on the new and unseen set of input variables (credit score, income level, and loan amount),
  - the model can be used to understand the relationship between the output variable and all the input variables.



Key Features of Data Science: 2.Building Representative Models

- For example, does income level really matter in determining the interest rate of a loan? Does income level matter more than credit score? What happens when income levels double or if credit score drops by 10 points?

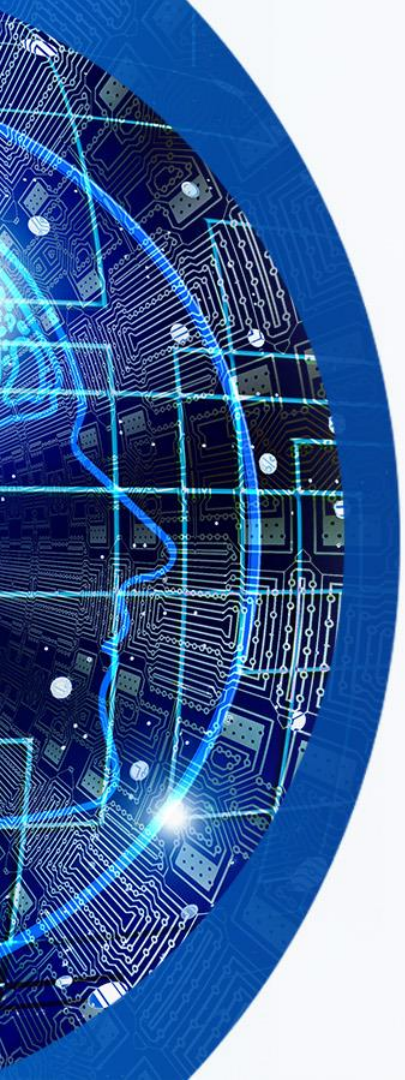
- A Model can be used for both predictive and explanatory applications.



Key Features of Data Science:

3.Combination of Statistics, Machine Learning, and Computing

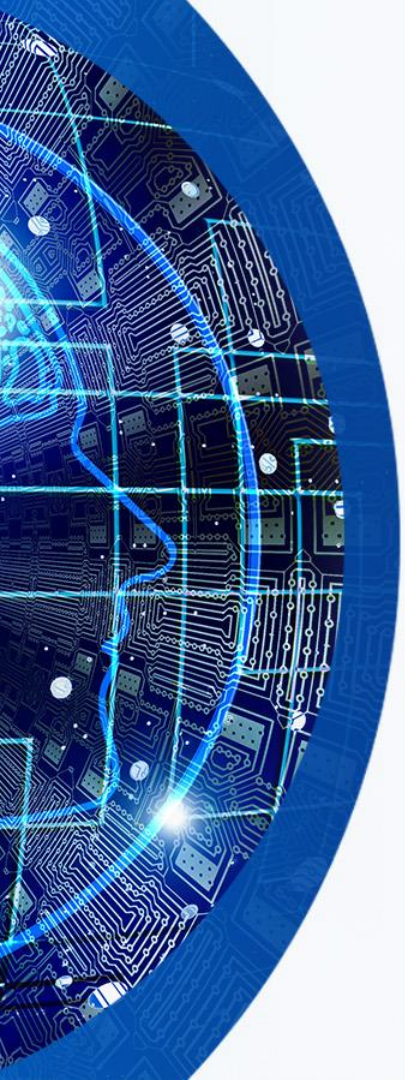
- Data science borrows computational techniques from the disciplines of statistics, machine learning, experimentation, and database theories.
- Data science also typically operates on large datasets that need to be stored, processed, and computed.
- This is where database techniques along with parallel and distributed computing techniques play an important role in data science.



Key Features of Data Science:

4.Learning Algorithms

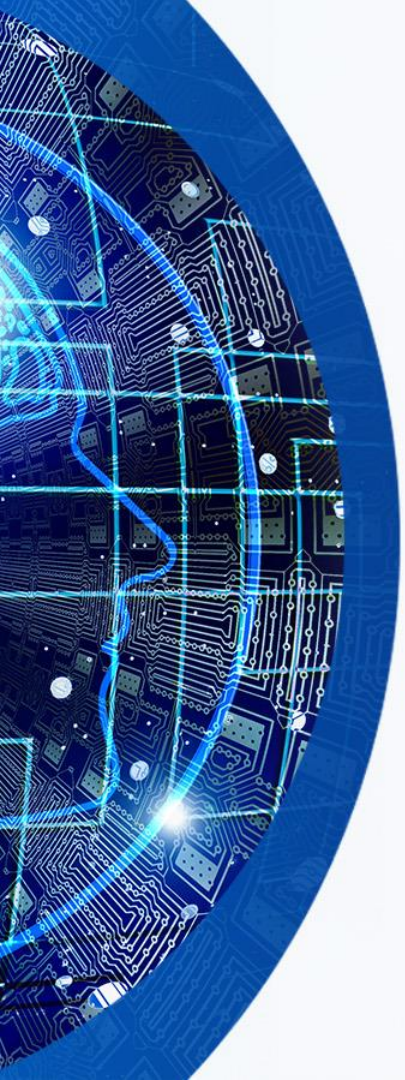
- We can also define data science as a process of discovering previously unknown patterns in data using automatic **iterative methods.**
- These iterative methods automate the process **of searching for an optimal solution for a given data problem.**
- Based on the problem, data science is classified into tasks such as classification, association analysis, clustering, and regression.



Key Features of Data Science: 4.Learning Algorithms

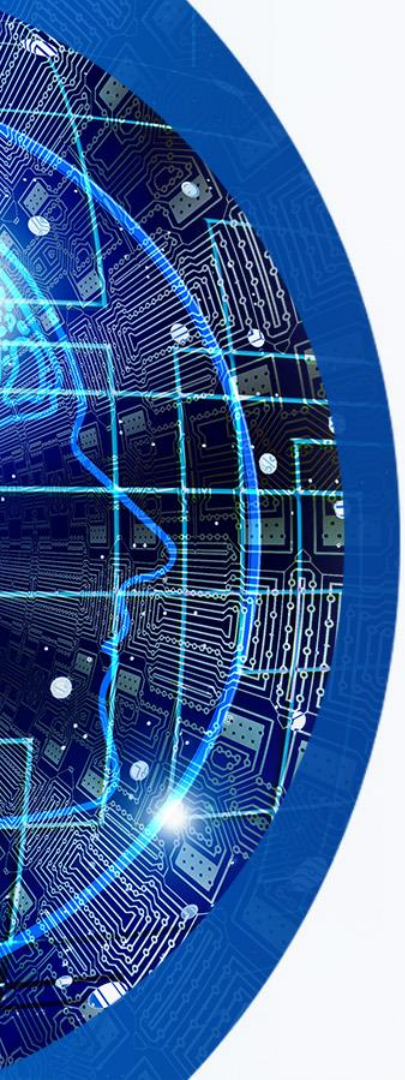
- Each data science task uses specific learning algorithms like decision trees, **neural networks**, k-nearest neighbors (k-NN), and k-means clustering, among others.

- With increased research on data science, such algorithms are increasing, but **a few classic algorithms remain foundational to many data science applications.**



Data Science Associated Fields

- While data science covers a wide set of techniques, applications, and disciplines, there a few associated fields that data science heavily relies on. The techniques used in the steps of a data science process and in conjunction with the term “data science” are:
  - **Descriptive statistics:** Computing mean, standard deviation, correlation, and other descriptive statistics.

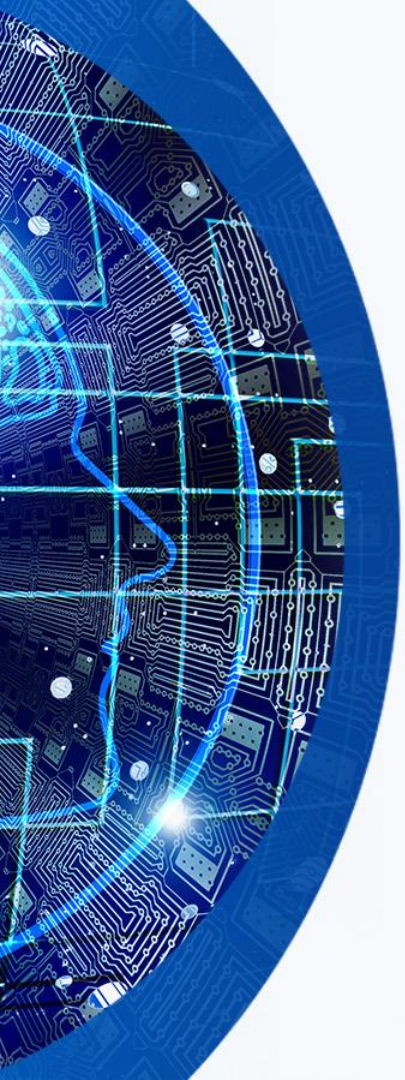


Data Science Associated Fields

- **Exploratory visualization:** The process of expressing data in visual coordinates enables users to find patterns and relationships in the data and to comprehend large datasets.

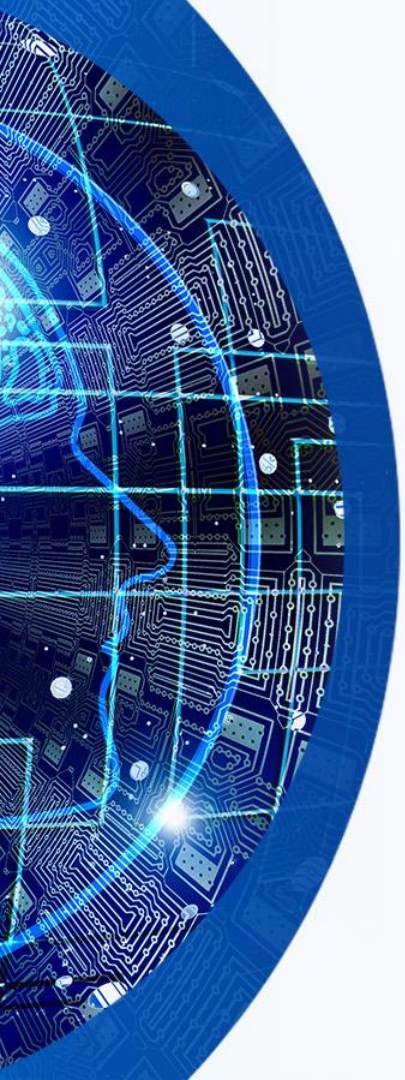
- **Dimensional slicing:** Online analytical processing (OLAP) applications, which are prevalent in organizations, mainly provide information on the data through dimensional slicing, filtering, and pivoting.

- **Data engineering:** Data engineering is the process of sourcing, organizing, assembling, storing, and distributing data for effective analysis and usage and preparing for data science learning algorithms.



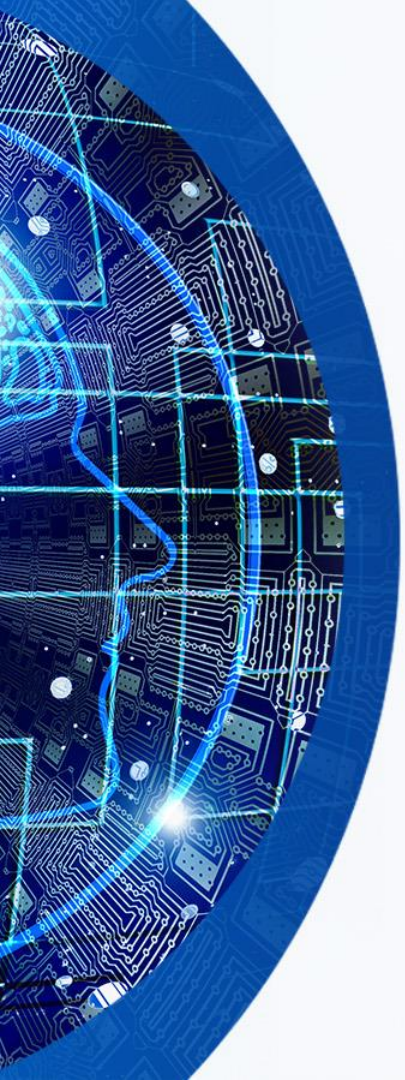
Data Science Associated Fields

- **Hypothesis testing:** In confirmatory data analysis, experimental data are collected to evaluate whether a hypothesis has enough evidence to be supported or not.
  - In general, data science is a process where many hypotheses are generated and tested based on observational data. Since the data science algorithms are iterative, solutions can be refined in each step.



Data Science Associated Fields

- **Business intelligence:** Business intelligence helps organizations consume data effectively. It helps query the ad hoc data without the need to write the technical query command or use dashboards or visualizations to communicate the facts and trends.
  - Historical trends are usually reported, but in combination with data science, both the past and the predicted future data can be combined. BI can hold and distribute the results of data science.



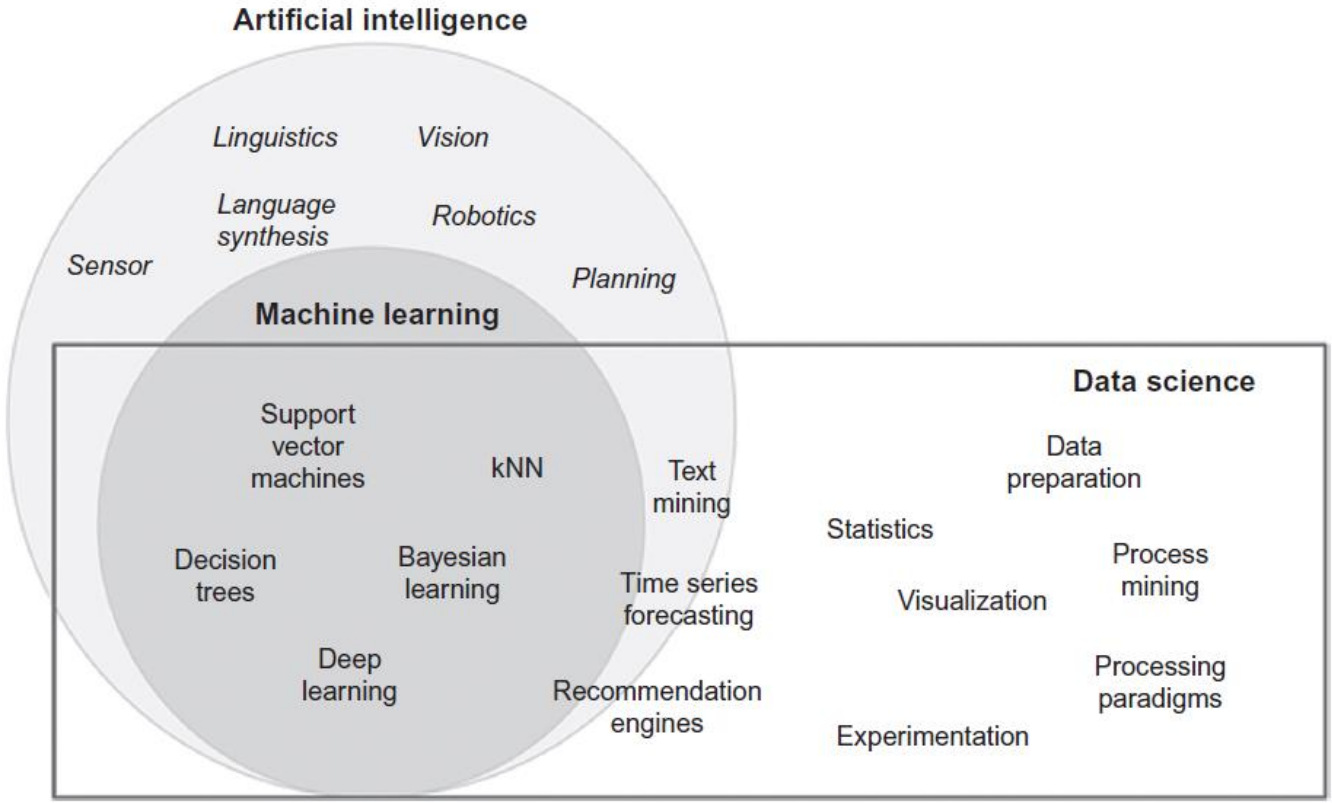
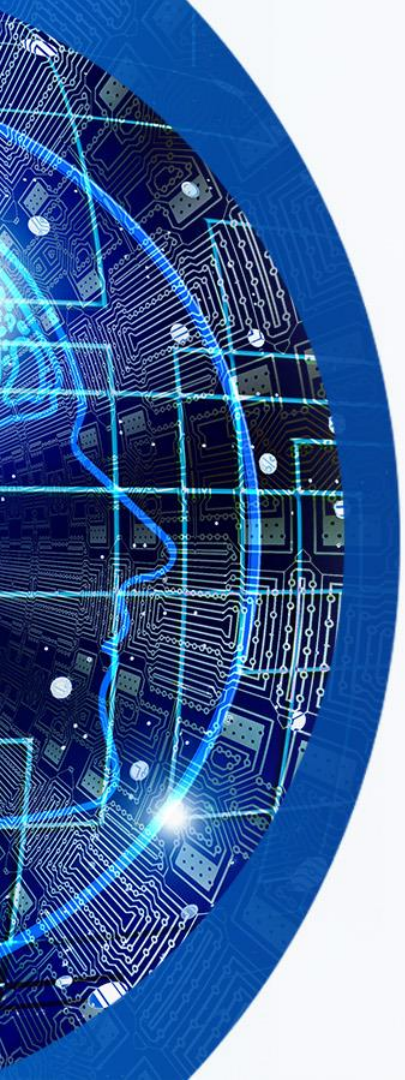


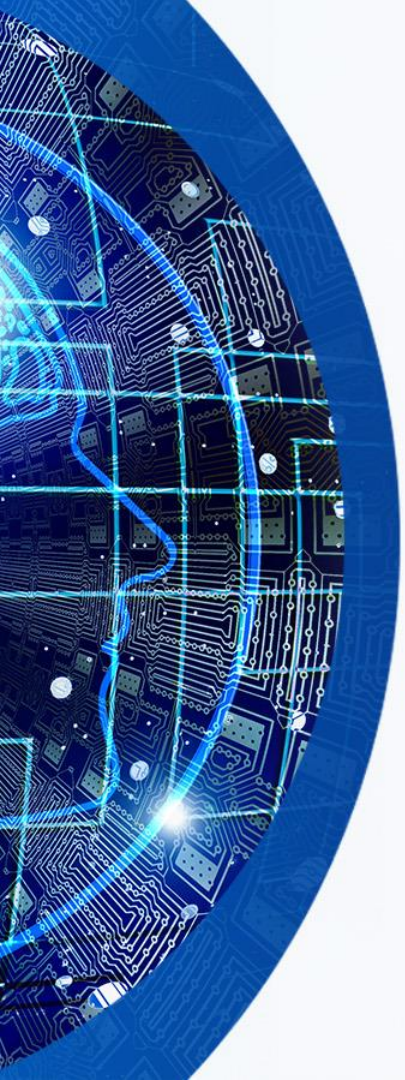
FIGURE 1.1
Artificial intelligence, machine learning, and data science.



Difference Between Data Scientist and Data Analyst

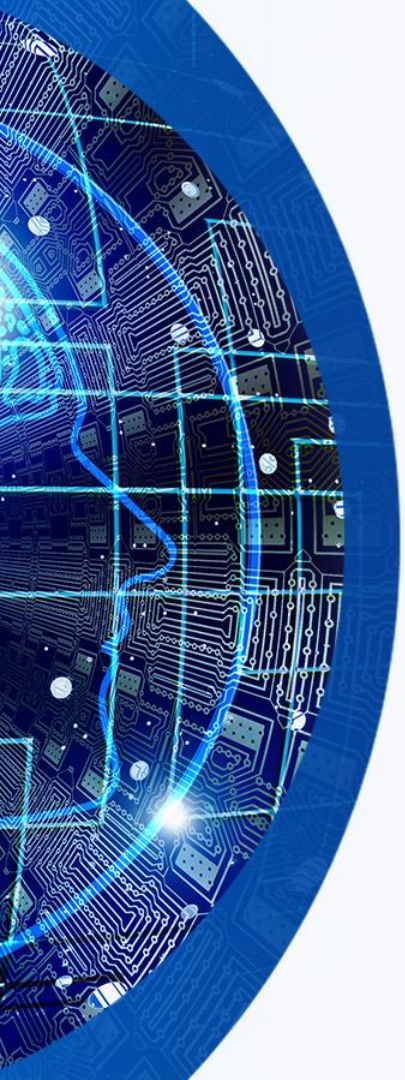
- While a data scientist is expected to forecast the future based on past patterns, data analysts extract meaningful insights from various data sources.

- A data scientist creates questions, while a data analyst finds answers to the existing set of questions.



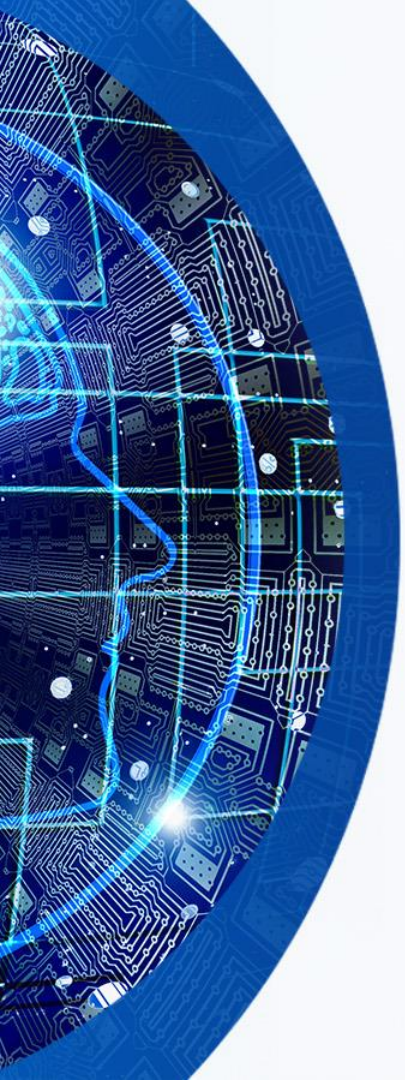
Difference Between Data Scientist and Data Analyst

- Data Science seeks to discover new and unique questions that can drive business innovation.
  - In contrast, Data Analysis aims to find solutions to these questions and determine how they can be implemented within an organization to foster data-driven innovation.



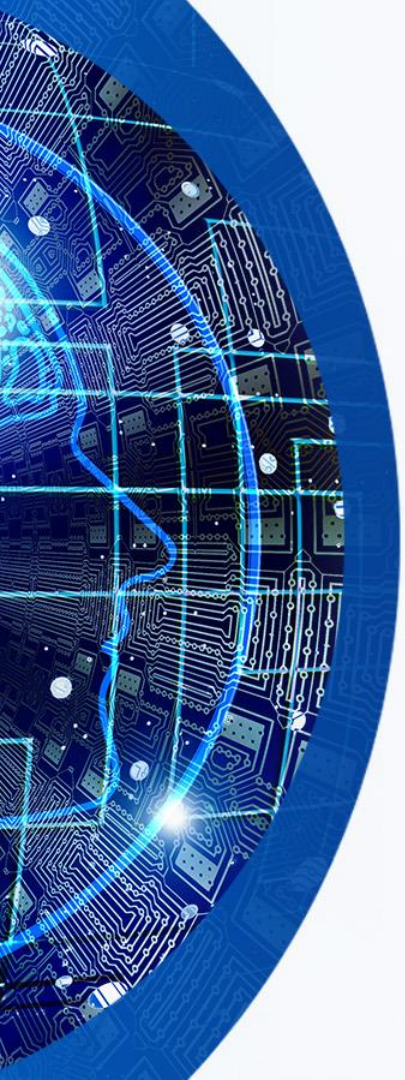
Data Science Problems Classification

- Data science problems can be broadly categorized into
  - **Supervised Learning Models** or
  - **Unsupervised Learning Models**.



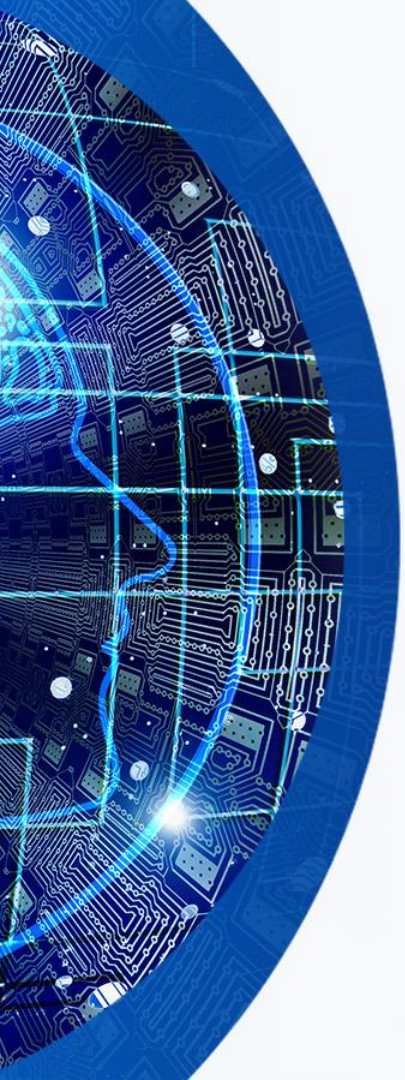
Data Science Problems Classification

- **Supervised Learning Models**
  - Supervised or directed data science tries to infer a function or relationship based on labeled training data and uses this function to map new unlabeled data.
  - Supervised techniques predict the value of the output variables based on a set of input variables. To do this, a model is developed from a **training dataset** where the values of input and output are previously known.



Data Science Problems Classification

- **Supervised Learning Models**
  - The model generalizes the relationship between the input and output variables and uses it to predict for a dataset where only input variables are known. **The output variable** that is being predicted is also called **a class label or target variable.**
  - Supervised data science needs a sufficient number of labeled records to learn the model from the data.

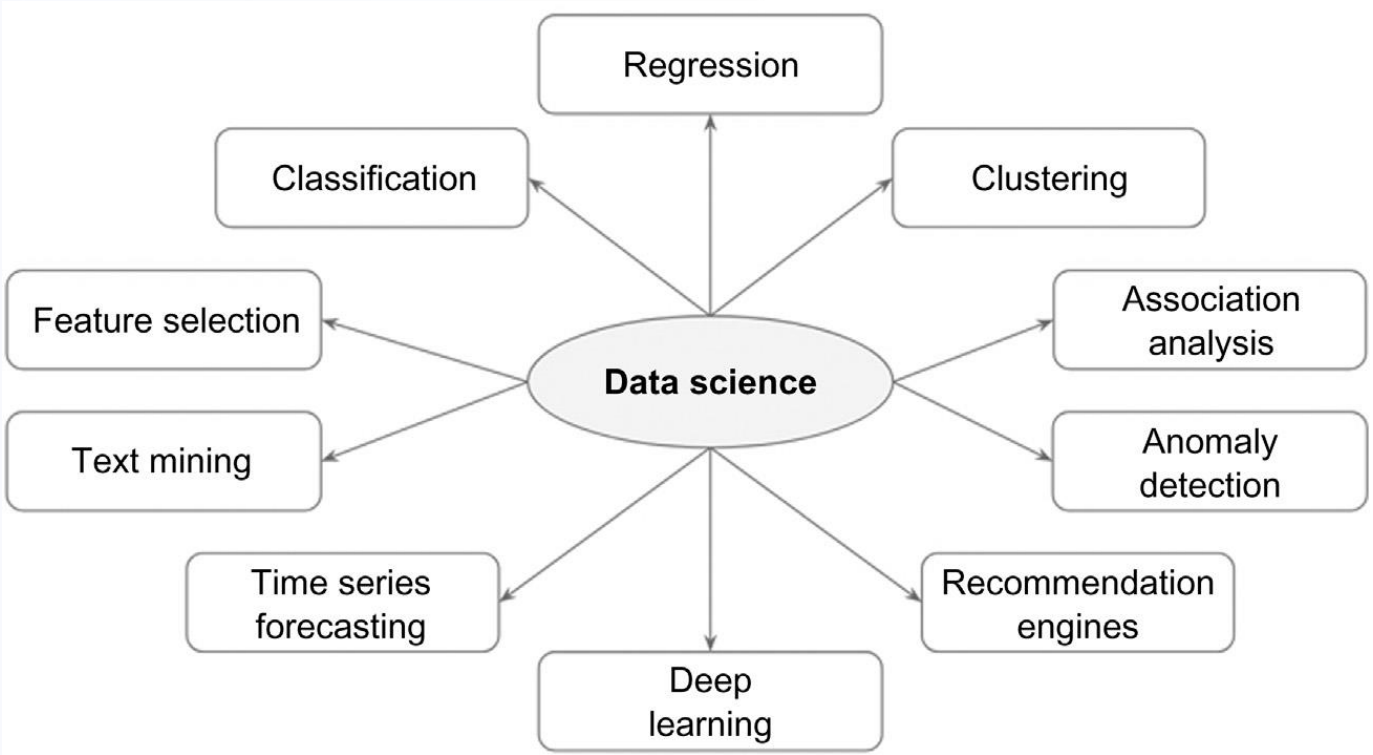


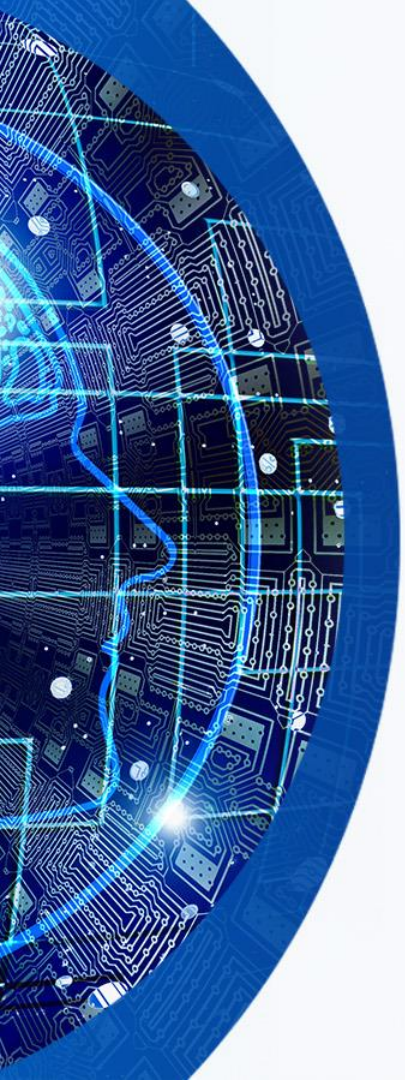
Data Science Problems Classification

- **Unsupervised Learning Models**
  - Unsupervised or undirected data science uncovers hidden patterns in unlabeled data. In unsupervised data science, **there are no output variables to predict.**
  - The objective of this class of data science techniques, is **to find patterns in data based on the relationship between data points themselves**. An application can employ both supervised and unsupervised learners.

Data Science Problems Classification

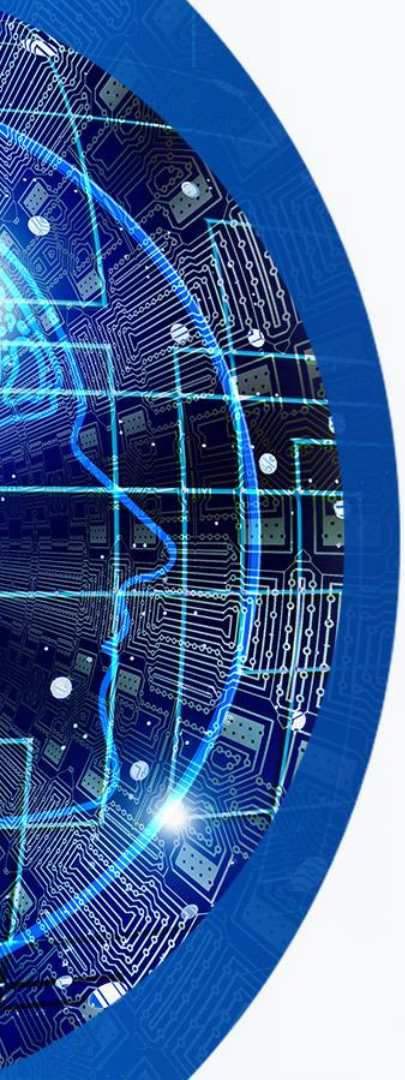
- Data science problems can also be classified into tasks.



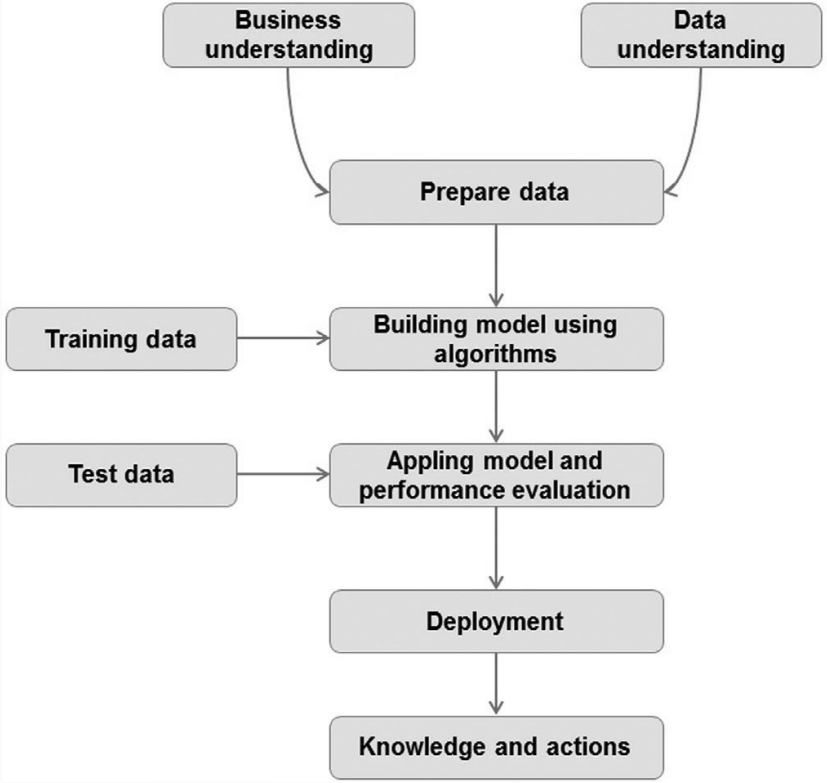


Data Science Process

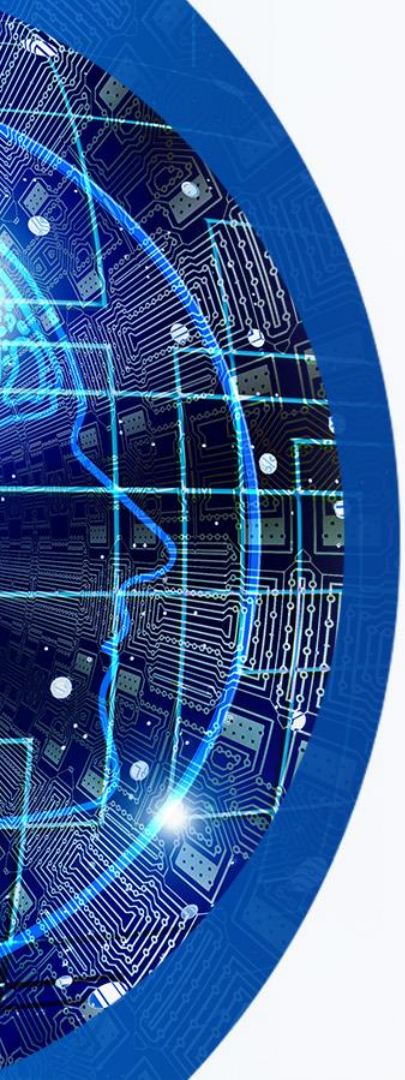
- The methodical discovery of useful relationships and patterns in data is enabled by a set of iterative activities collectively known as the data science process.
- The standard data science process involves:
  1. **Understanding the problem,**
  2. **Preparing the data samples,**
  3. **Developing the model,**
  4. **Applying the model on a dataset to see how the model may work in the real world,**
  5. **Deploying and maintaining the models**



Data Science Process

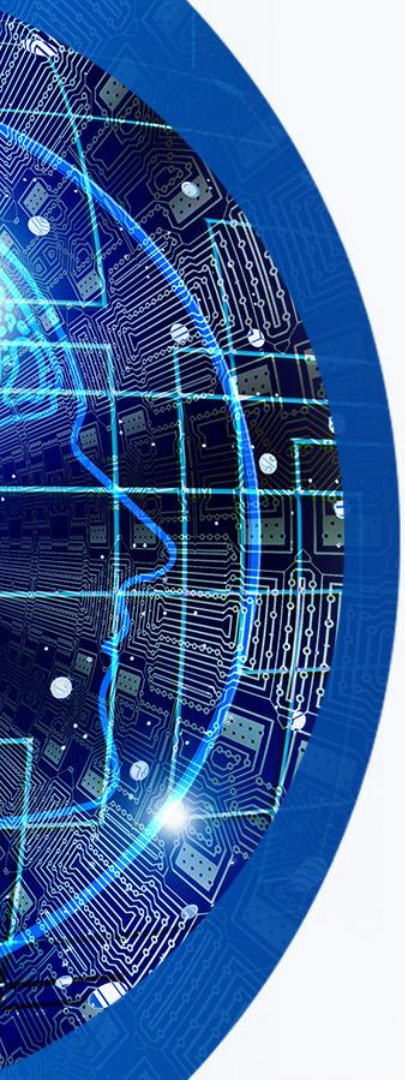


1. *Prior Knowledge*
2. *Preparation*
3. *Modeling*
4. *Application*
5. *Knowledge*



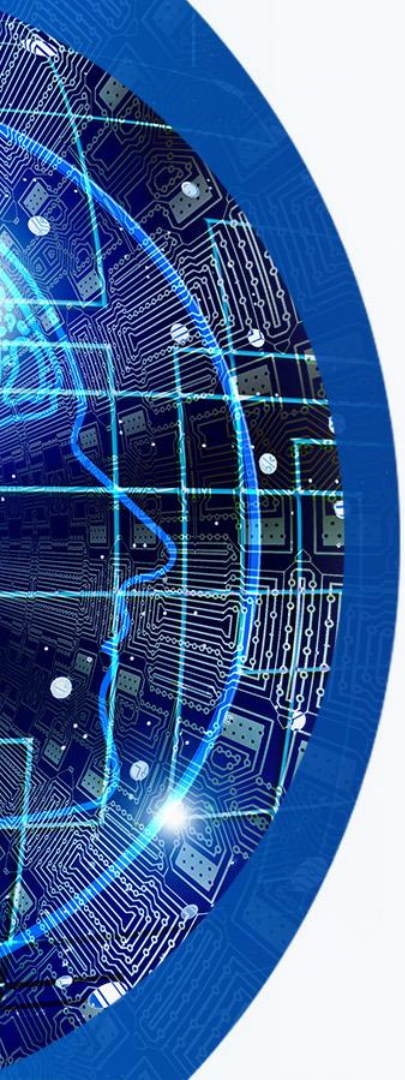
Data Science Process: Understanding the Problem

- Without a well-defined statement of the problem, it is impossible to come up with the right dataset and pick the right data science algorithm



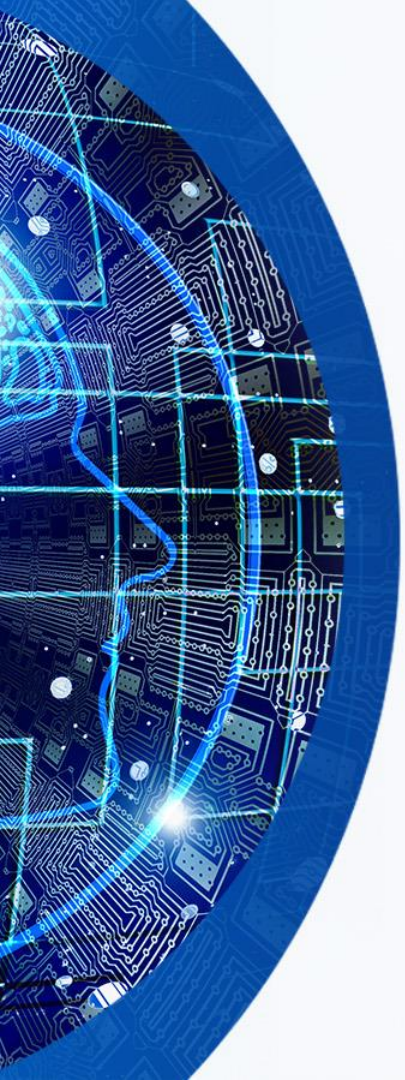
Data Science Process: Preparing the Data Samples

- Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process.
- There are quite a range of factors to consider: **quality of the data, quantity of data, availability of data, gaps in data.**
- The objective of this step is to come up with a dataset to answer the business question through the data science process



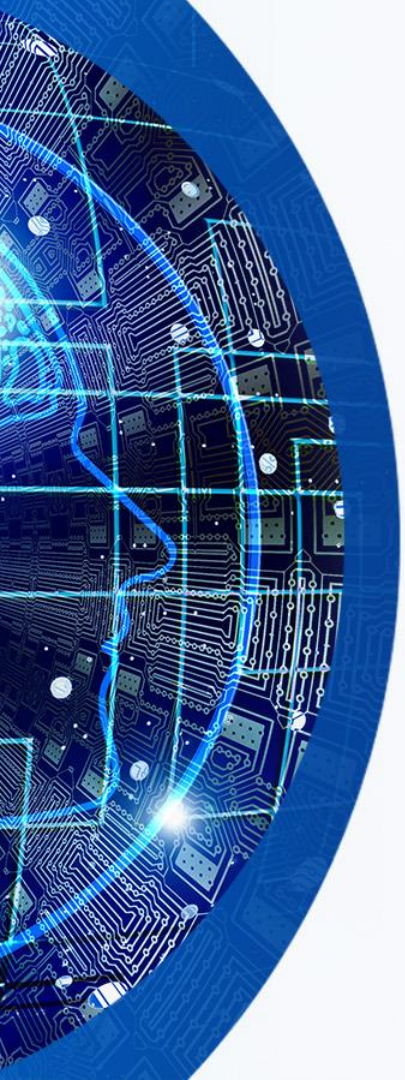
Data Science Process: Developing the Model

- A model is the abstract representation of the data and the relationships in a given dataset.
- As a data scientist, it is **sufficient to an overview of the learning algorithm**, **how it works, and determining what parameters** need to be configured based on the understanding of the business and data.
- There are a few hundred data science algorithms in use today, derived from statistics, machine learning, pattern recognition, and the body of knowledge related to computer science.



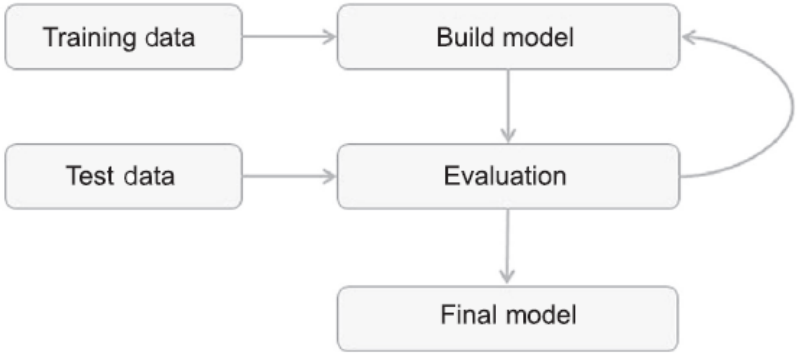
Data Science Process: Developing the Model

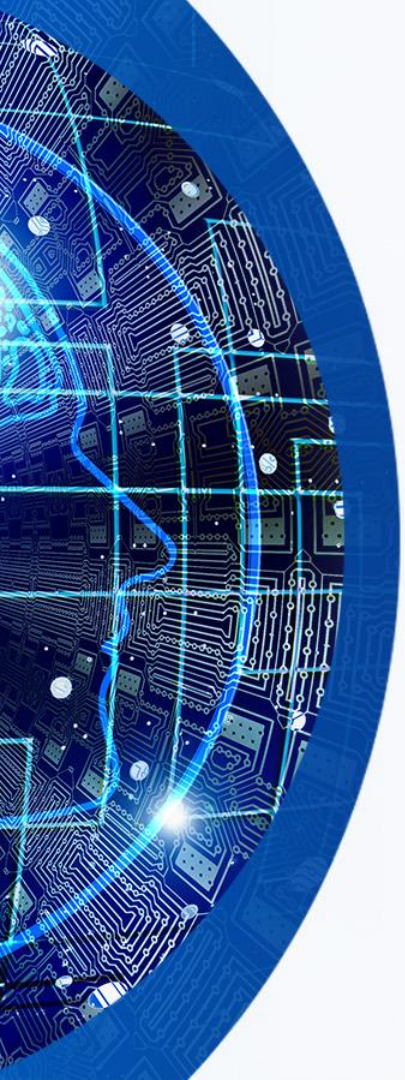
- The choice of which algorithm to use **depends on the type of dataset, objective, structure of the data, presence of outliers, available computational power, number of records, number of attributes**, and so on.
  - **It is up to the data scientist to decide which algorithm (s)** to use by evaluating the performance of multiple algorithms. There have been hundreds of algorithms developed in the last few decades to solve data science problems
  - For example, within a classification task many algorithms can be chosen from such as (Decision trees, Rule induction, Bayesian models, k-NN, **Neural Networks**)



Data Science Process: Developing the Model

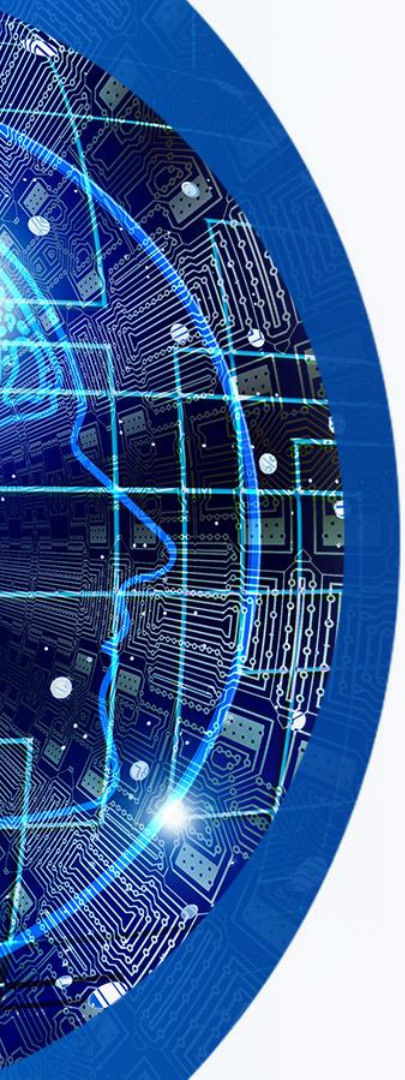
- **Classification and regression tasks** are predictive techniques because they predict an outcome variable based on one or more input variables. Predictive algorithms require a prior known dataset to learn the model.
- This figure shows the steps in the modelling phase of predictive data science.





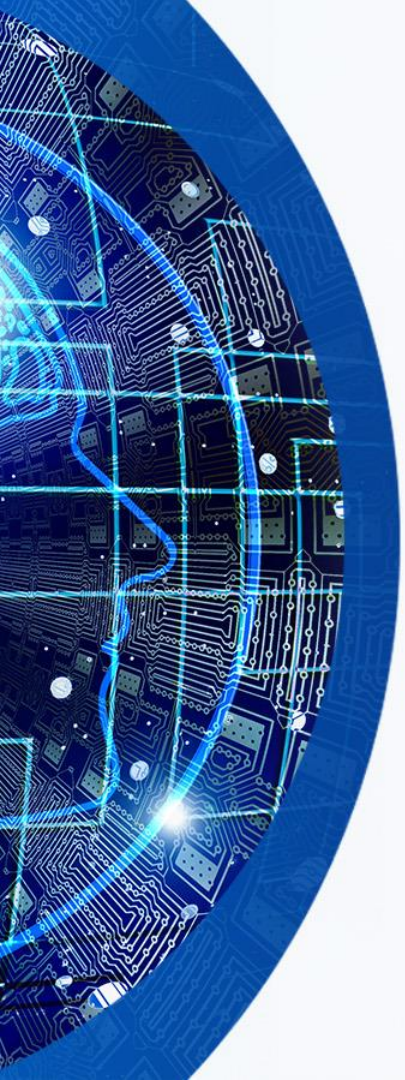
Data Science Process
Developing the Model

- **Association analysis and clustering** are descriptive data science techniques where there is no target variable to predict;
  - hence, there is no test dataset. However, both predictive and descriptive models have an evaluation step.



Data Science Process: Evaluating the Model

- To evaluate the model test dataset, which was not previously used in building the model, is used for evaluation,
- The actual value can be compared against the predicted value using the model, and thus, the prediction error can be calculated.
- As long as the error is acceptable, this model is ready for deployment. The error rate can be used to compare this model with other models developed using different algorithms.

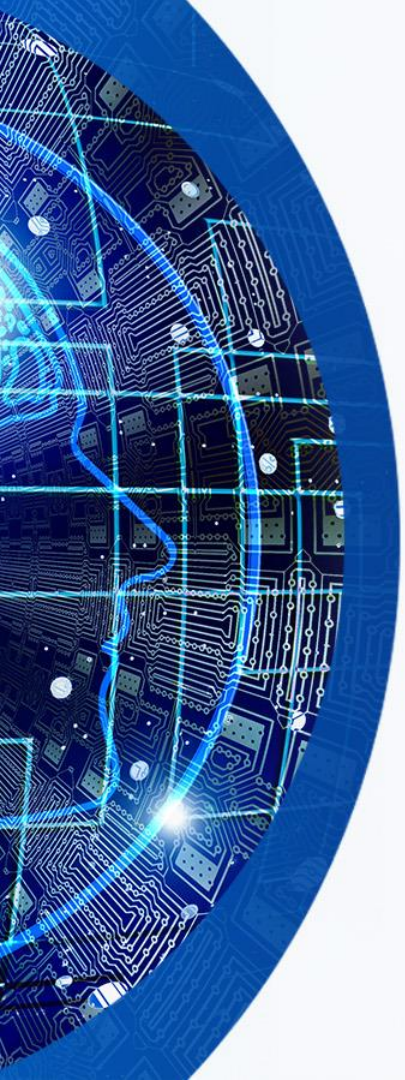


Data Science Process: Model Deployment

- Deployment is the stage at which the model **becomes production ready or live.**

- In business applications, the results of the data science process have to be assimilated into the business process—usually in software applications.

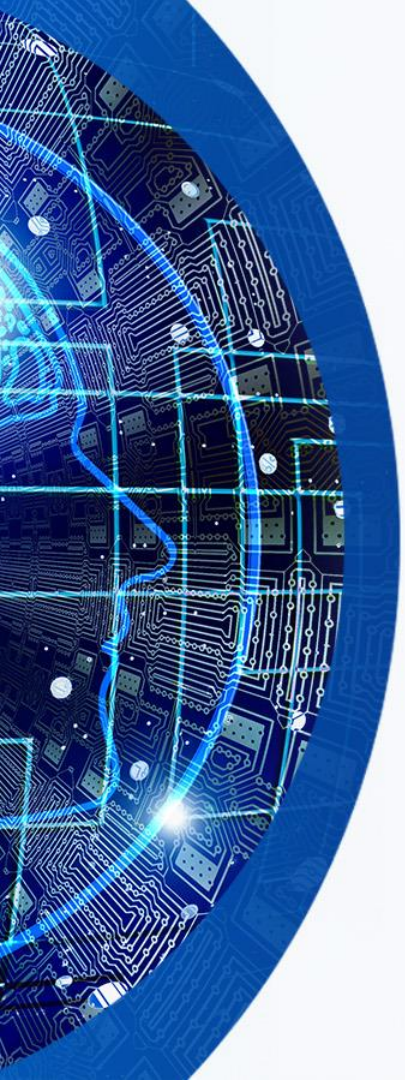
- The quality of prediction, accessibility of input data, and the **response time of the prediction** remain the critical quality factors in business application.



Data Science Process: Knowledge

- The data science process starts with prior knowledge and ends with posterior knowledge, which is the incremental insight gained.

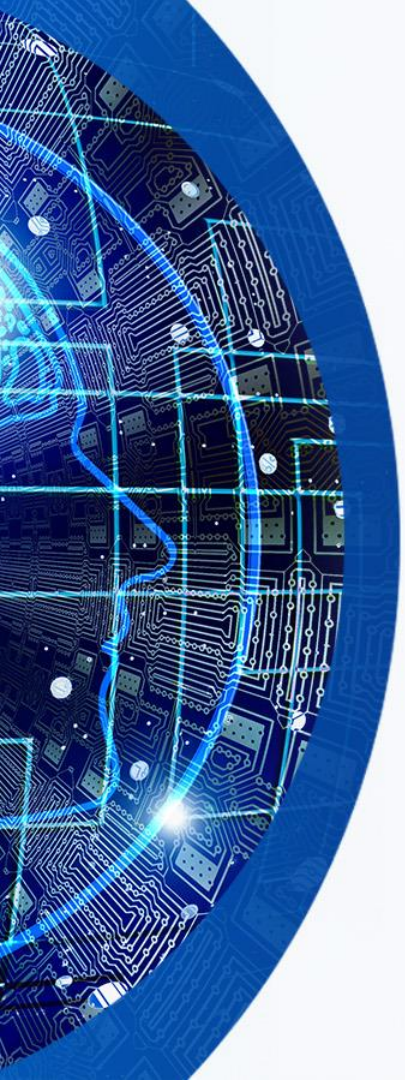
- Not all discovered patterns lead to incremental knowledge. Again, **it is up to the data scientist to invalidate the irrelevant patterns and identify the meaningful information.**



Data Science Process

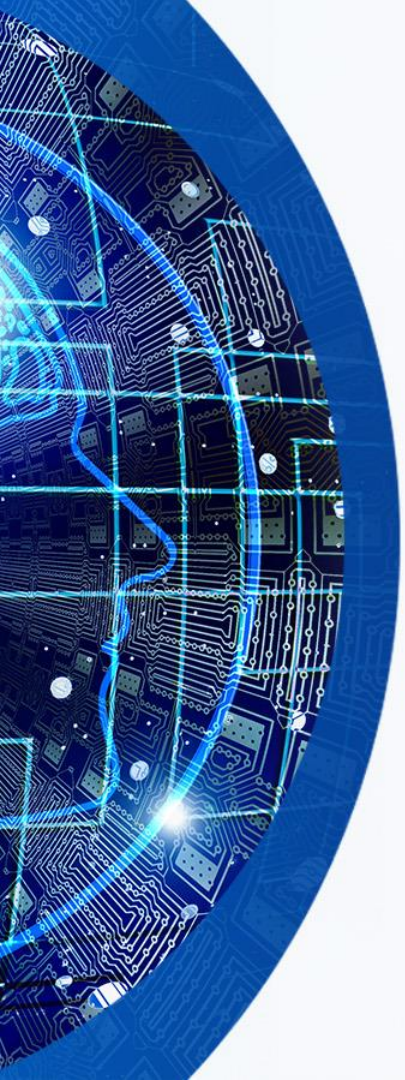
- Finally, the whole data science process is a **framework** to invoke the **right questions** and provide **guidance**, through the **right approaches**, to **solve a problem.**

- It is not meant to be used as a set of rigid rules, but as a set of **iterative**, distinct steps that aid in knowledge discovery.



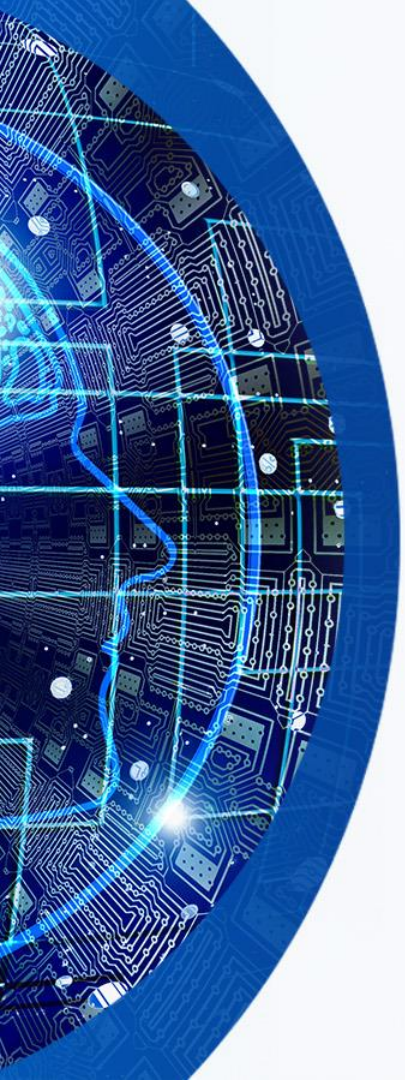
Application of Data science in Health analytics

- As attempts in the industry are being to provide quality health care at reasonable costs.
- Use cases of data science in medicine and healthcare.
  - Medical image analysis
  - Genetics and genomics
  - Predictive medicine: prognosis & diagnostic accuracy



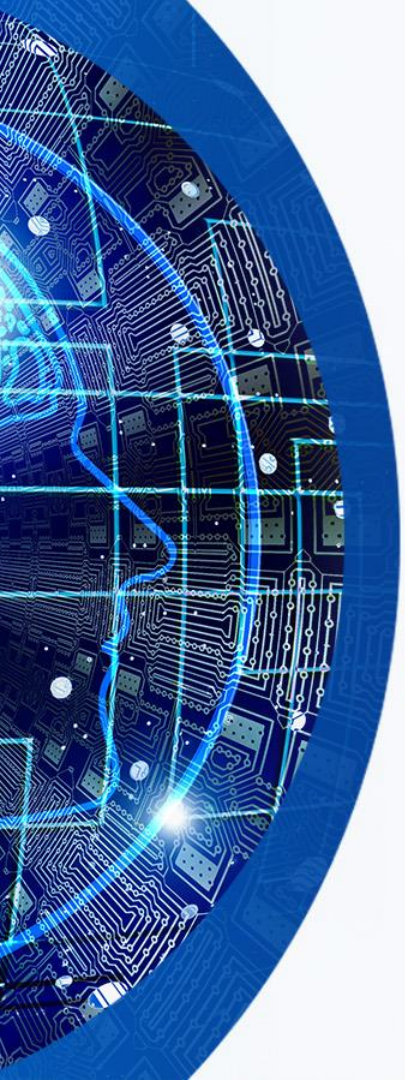
Application of Data science in Finance

- Embracing the ability of data science to cope with a number of principal financial tasks.
- Use cases of data science in finance:
  - Managing customer data
  - Predictive analysis
  - Fraud detection
  - Consumer analytics
  - Product development and targeted marketing



Application of Data science in Education

- Aims to improve learning outcomes, student performance, and teacher effectiveness
  - Predict/focus performance of students in class
    - Using statistical models
  - Student recruitment
  - Predictive models to evaluate the risks of student dropouts
  - Better ways of assessing teachers



Thank You