Unit - 1

What is Data Science

What is Data Sciecne

- Data science is the study of data to extract meaningful insights for business.
- It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.
- This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

Why is data science important?

- Data science is important because it combines tools, methods, and technology to generate meaning from data.
- Modern organizations are inundated with data; there is a proliferation of devices that can automatically collect and store information.
- Online systems and payment portals capture more data in the fields of e-commerce, medicine, finance, and every other aspect of human life.
- We have text, audio, video, and image data available in vast quantities.

History of data science

- While the term data science is not new, the meanings and connotations have changed over time.

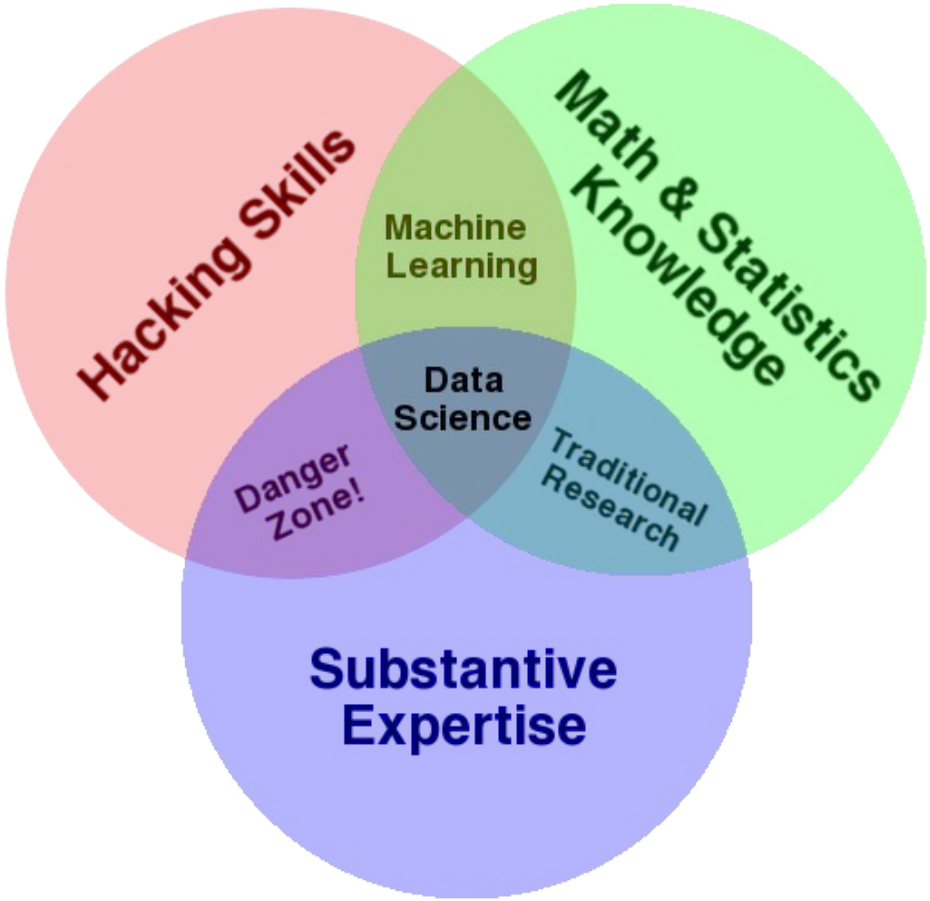
- The word first appeared in the ’60s as an alternative name for statistics. In the late ’90s, computer science professionals formalized the term.
- A proposed definition for data science saw it as a separate field with three aspects: data design, collection, and analysis. It still took another decade for the term to be used outside of academia.

What is Data Science?

- Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

- Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence.

The Data Science Venn diagram



The Data Science Venn Diagram

- According to the Data Science Venn Diagram, Machine learning involves the knowledge of Computer programming and Math but without any domain expertise. This means that you just need to throw your data into the model without necessarily knowing about the details of the data such as what data is, what it means etc.
- Hacking requires great coding skills. Coding is important because it helps you to gather and prepare the data because a lot of data is unstructured or present in unusual formats. You also require programming skills to apply statistics to your problems, handle the database, etc. One with hacking skills can apply very complex algorithms by computer programming.
- After collecting and preparing the data, now comes the part of extracting the insights from it. Mathematics is important for analyzing the data.For analyzing the data, you will require several tools from mathematics such as probability, algebra, etc. It helps in the diagnosis of the problem by applying various mathematical and statistical approaches to your data.

Terminology

- Data Science is the theory and practice powering the data-driven transformations we are seeing across industry and society today.
- **Algorithm**
- **Artificial Intelligence**
- **Bayes Theorem**
- **Behavioural analytics**
- **Big Data**
- **Citizen Data Scientist**
- **Classification**
- **Clickstream analytics**
- **Clustering**
- **Data Mining**
- **Data Set**
- **Data Governance**

Case Studies

- **Case Study 1: Text Emotions Detection**

If you are one of them who is having an interest in natural language processing then this use case is for you. The idea is to train a machine learning model to generate emojis based on input text. Then this machine learning model can be used in training Artificial Intelligent Chatbots.

- ***Use Case:*** A human can express his emotions in any form, such as the face, gestures, speech and text. The detection of text emotions is a content-based classification problem. Detecting a person's emotions is a difficult task, but detecting the emotions using text written by a person is even more difficult as a human can express his emotions in any form.

Case Studies

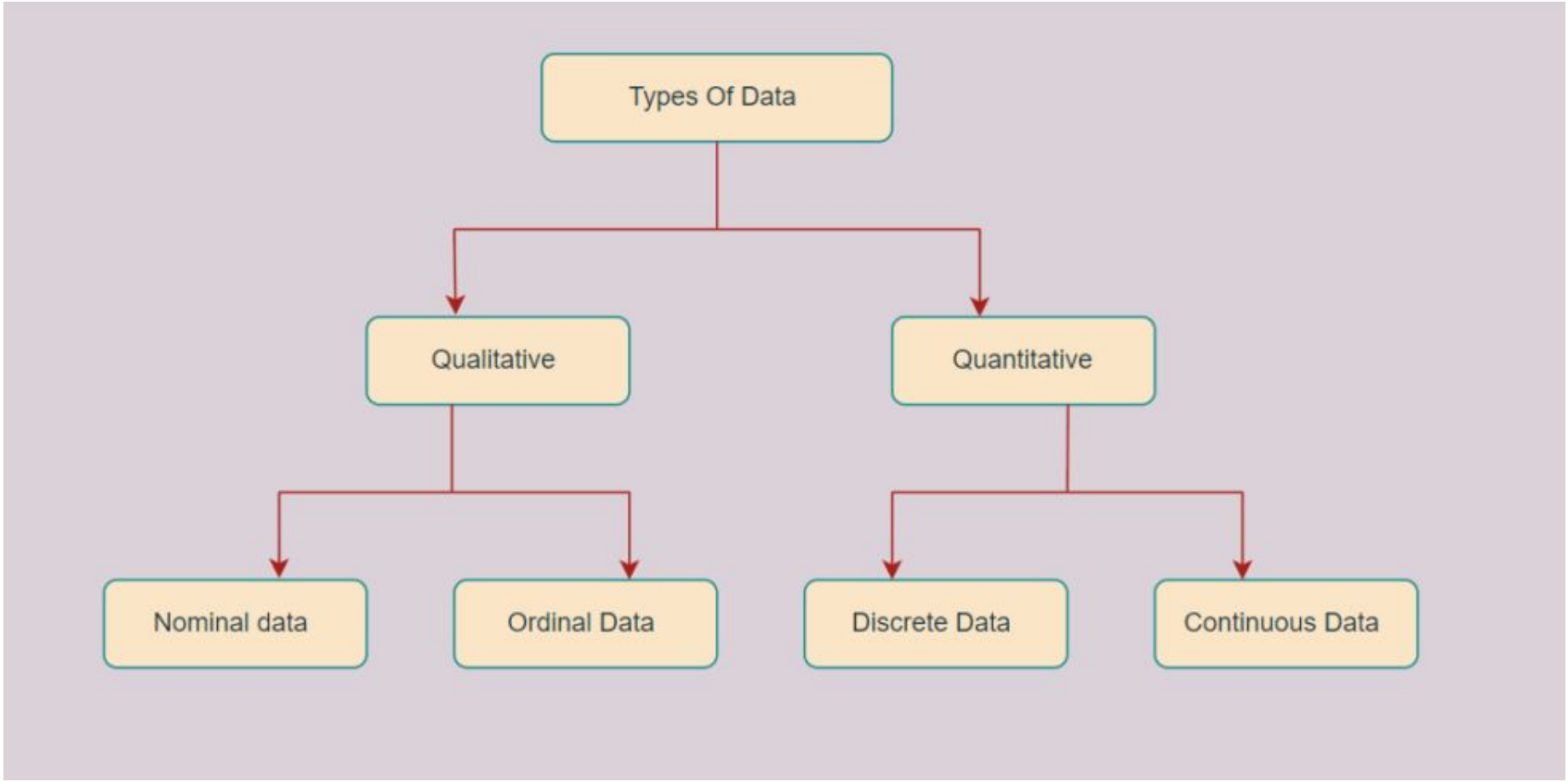
- **Case Study 2: Hotel Recommendation System**
- A hotel recommendation system typically works on collaborative filtering that makes recommendations based on ratings given by other customers in the same category as the user looking for a product.
- ***Use Case:*** We all plan trips and the first thing to do when planning a trip is finding a hotel. There are so many websites recommending the best hotel for our trip. A hotel recommendation system aims to predict which hotel a user is most likely to choose from among all hotels. So to build this type of system which will help the user to book the best hotel out of all the other hotels. We can do this using customer reviews.

Types of Data

- **4 Types Of Data – Nominal, Ordinal, Discrete and Continuous**
- **Importance of Data**
- "Data is the new oil." Today data is everywhere in every field. Whether you are a data scientist, marketer, businessman, data analyst, researcher, or you are in any other profession, you need to play or experiment with raw or structured data. This data is so important for us that it becomes important to handle and store it properly, without any error. While working on these data, it is important to know the types of data to process them and get the right results.

Types of Data

- **There are two types of data: Qualitative and Quantitative data**, which are further classified into **four types of data: nominal, ordinal, discrete, and Continuous.**



Types of data

Qualitative or Categorical Data

- Qualitative or Categorical Data is data that can’t be measured or counted in the form of numbers.
- These types of data are sorted by category, not by number. That’s why it is also known as Categorical Data.
- These data consist of audio, images, symbols, or text.
- The gender of a person, i.e., male, female, or others, is qualitative data.

The other examples of qualitative data are :

- What language do you speak
- Favourite holiday destination
- Opinion on something (agree, disagree, or neutral)
- Colours

Types of data

- **The Qualitative data are further classified into two parts :**
- **Nominal Data**
- Nominal Data is used to label variables without any order or quantitative value. The colour of hair can be considered nominal data, as one colour can't be compared with another colour.
- The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed to distinct categories.

- **Examples of Nominal Data :**
- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)

Ordinal Data

- Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.
- The ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered as "in-between" the qualitative data and quantitative data. The ordinal data only shows the sequences and cannot use for statistical analysis. Compared to the nominal data, ordinal data have some kind of order that is not present in nominal data.
- **Examples of Ordinal Data :**
- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of peoples in a competition (First, Second, Third, etc.)

Difference between Nominal and Ordinal Data

Nominal Data	Ordinal Data
Nominal data can't be quantified, neither they have any intrinsic ordering	Ordinal data gives some kind of sequential order by their position on the scale
Nominal data is qualitative data or categorical data	Ordinal data is said to be "in-between" qualitative data and quantitative data
They don't provide any quantitative value, neither we can perform any arithmetical operation	They provide sequence and can assign numbers to ordinal data but cannot perform the arithmetical operation
Nominal data cannot be used to compare with one another	Ordinal data can help to compare one item with another by ranking or ordering
Examples: Eye colour, housing style, gender, hair colour, religion, marital status, ethnicity, etc	**Examples:** Economic status, customer satisfaction, education level, letter grades, etc

Types of Data

Quantitative Data

- Quantitative data can be expressed in numerical values, which makes it countable and includes statistical data analysis.
- These kinds of data are also known as Numerical data.
- It answers the questions like, “how much,” “how many,” and “how often.”
- For example, the price of a phone, the computer’s ram, the height or weight of a person, etc., falls under the quantitative data.
- Quantitative data can be used for statistical manipulation and these data can be represented on a wide variety of graphs and charts such as bar graphs, histograms, scatter plots, boxplot, pie charts, line graphs, etc.

Examples of Quantitative Data :

- Height or weight of a person or object
- Room Temperature
- Scores and Marks (Ex: 59, 80, 60, etc.)
- Time

The Quantitative data are further classified into two parts :

Discrete Data

- The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.
- The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.
- **Examples of Discrete Data :**
- Total numbers of students present in a class
- Cost of a cell phone
- Numbers of employees in a company
- The total number of players who participated in a competition

Continuous Data

- Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.
- The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.
- **Examples of Continuous Data :**
- Height of a person
- Speed of a vehicle
- “Time-taken” to finish the work
- Wi-Fi Frequency

Difference between Discrete and Continuous Data

Discrete Data	Continuous Data
Discrete data are countable and finite; they are whole numbers or integers	Continuous data are measurable; they are in the form of fraction or decimal
Discrete data are represented mainly by bar graphs	Continuous data are represented in the form of a histogram
The values cannot be divided into subdivisions into smaller pieces	The values can be divided into subdivisions into smaller pieces
Discrete data have spaces between the values	Continuous data are in the form of a continuous sequence
Examples: Total students in a class, number of days in a week, size of a shoe, etc	**Example:** Temperature of room, the weight of a person, length of an object, etc

The four levels of Data

There are 4 levels of measurement:

- **Nominal**: the data can only be categorized
- **Ordinal**: the data can be categorized and ranked
- **Interval**: the data can be categorized, ranked, and evenly spaced
- **Ratio**: the data can be categorized, ranked, evenly spaced, and has a natural zero.

The four levels of Data

Nominal level	Examples of nominal scales
You can categorize your data by labelling them in mutually exclusive groups, but there is no order between the categories.	• City of birth • Gender • Ethnicity • Car brands • Marital status
Ordinal level	**Examples of ordinal scales**
You can categorize and rank your data in an order, but you cannot say anything about the intervals between the rankings. Although you can rank the top 5 Olympic medallists, this scale does not tell you how close or far apart they are in number of wins.	• Top 5 Olympic medallists • Language ability (e.g., beginner, intermediate, fluent) • Likert-type questions (e.g., very dissatisfied to very satisfied)

The four levels of Data

Interval level	Examples of interval scales
You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point. The difference between any two adjacent temperatures is the same: one degree. But zero degrees is defined differently depending on the scale – it doesn't mean an absolute absence of temperature. The same is true for test scores and personality inventories. A zero on a test is arbitrary; it does not mean that the test-taker has an absolute lack of the trait being measured.	• Test scores (e.g., IQ or exams) • Personality inventories • Temperature in Fahrenheit or Celsius
Ratio level	**Examples of ratio scales**
You can categorize, rank, and infer equal intervals between neighboring data points, and there is a true zero point. A true zero means there is an absence of the variable of interest. In ratio scales, zero does mean an absolute lack of the variable. For example, in the Kelvin temperature scale, there are no negative degrees of temperature – zero means an absolute lack of thermal energy.	• Height • Age • Weight • Temperature in Kelvin