بسم الله الرحمن الرحيم

Workshop on Data Science

Instructor: Mr. Mubahir Iqbal

Thursday, 12 Dec 2024

Lecture outline

- Data Science
- Datafication, Data, & it’s types, Big Data
- Exploratory Data Analysis
- Artificial Intelligence, Machine Learning & it’s Types
- Data Imbalance & Data Splitting
- Model Training
- Evaluation Matrix

Artificial Intelligence (AI)

A field of computer science that focuses on **creating systems that can perform tasks that would normally require human intelligence**.

These tasks include understanding language, pattern recognition, making decisions, and solving problems.

AI aims to build machines or programs that can think, learn, and adapt to new situations, making it possible for computers to mimic human abilities and cognitive function.

Types of AI

AI can be divided into **three main types** based on capabilities:

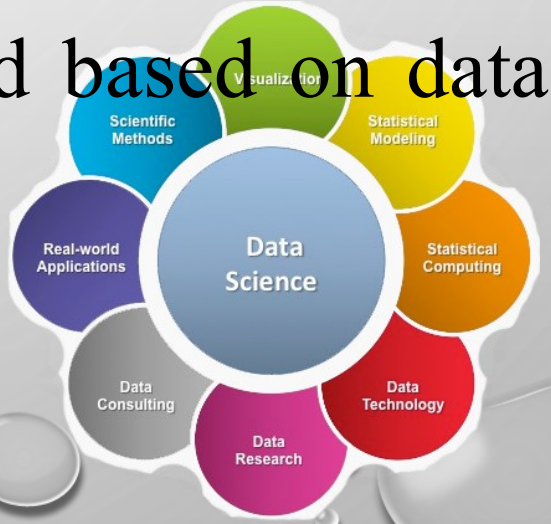
- **Narrow AI (Weak AI):**
  Designed to perform a single task **or a narrow set of tasks**.
  Examples: Voice assistants, recommendation systems, and image recognition.

- **General AI (Strong AI):**
  AI with human-like intelligence that can understand, learn, and decide.
  This type **doesn't yet exist** but remains a goal in AI research.

- **Superintelligent AI:**
  Hypothetical AI that surpasses human intelligence and capabilities.

What is Data Science?

- Data science is a field that **aims to find useful insights from big data**.
- It combines **statistical mathematics**, **computer science**, and **subject knowledge to analyze**, **interpret**, and **visualize data**.
- This helps in solving complex problems and guiding decisions.
- Artificial intelligence models are implemented based on data to make decisions and predictions.



What is Big Data?

What is Datafication?

- The datafication term was introduced in 2013.
- A process of "**Taking all aspects of life and turning them into Data**" is called datafication.
- This includes **converting** **actions**, **emotions**, and **interactions** into data points.

Examples:

- Want to purchase a product or use a service (**Rating**)
- Want to download a mobile app - (**Reviews**)
- Want to Watch a movie on YouTube (**Watch time/Likes, Comments**)
- Social media posts

Properties of data by Structure

1. **Structured Data:** Data that is organized in a clear, fixed format, typically in tables with rows and columns (like a spreadsheet)

An Excel sheet with columns for "Name," "Age," and "Salary" is structured data, as each column has a specific type and format

2. **Unstructured Data:** Data that does not have a specific format, making it harder to store in a structured database.

Text documents, emails, images, videos, and social media posts are unstructured

Properties of data by Structure

3. **Semi-Structured Data:** Data that isn't as organized as structured data but has some tags or markers that make it easier to categorize. It is also called mix-structured data.

JSON or XML files are semi-structured, as they use tags (like <name>John</name>) to give some structure to the data, even though it's not in a table format.

Properties of data by Type

1. **Quantitative Data (Numerical Data):** Data that can be measured and expressed as numbers

The height of students in centimeters (like 150 cm, 160 cm) or the number of items sold (like 30, 50)

i. Discrete Data: Numeric data with specific, countable values.

Number of children in a family (1, 2, 3, ...)

ii. Continuous Data: Numeric data that can take any value within a range.

Temperature (20.5°C, 22.7°C) or height (170.2 cm, 172.5 cm)

Properties of data by Type

2. **Categorical Data (Qualitative Data):** Data representing categories or groups rather than numerical values.

Colours of cars (like "red," "blue," "black") or types of products in a store (like "electronics," "furniture").

i. **Nominal Data:** Data without any order.

Types of pets (dog, cat, bird)

ii. **Ordinal Data:** Data with an order but without specific numeric differences.

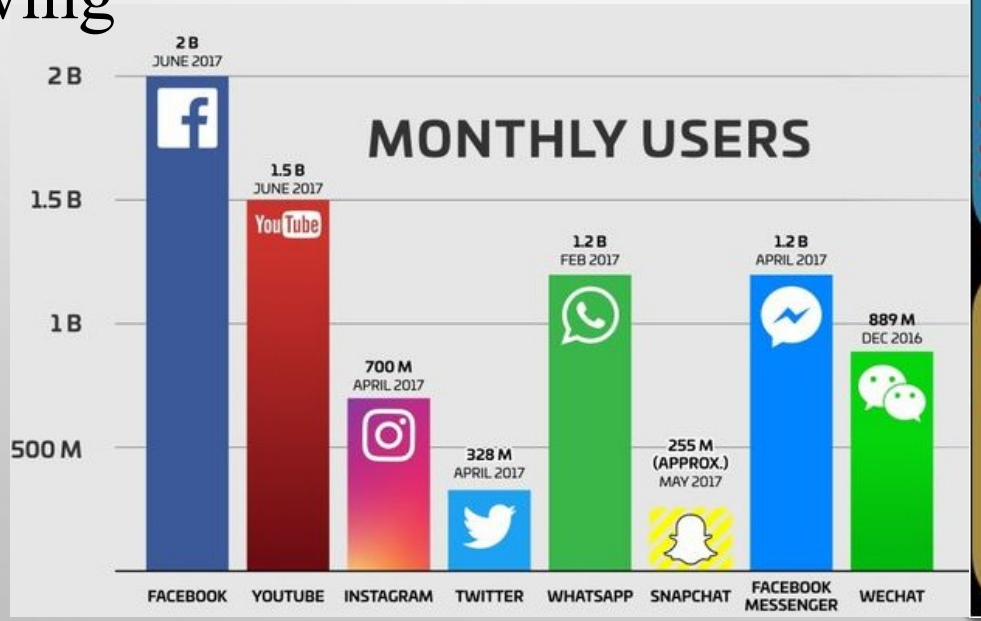
Rating levels (poor, fair, good, excellent)

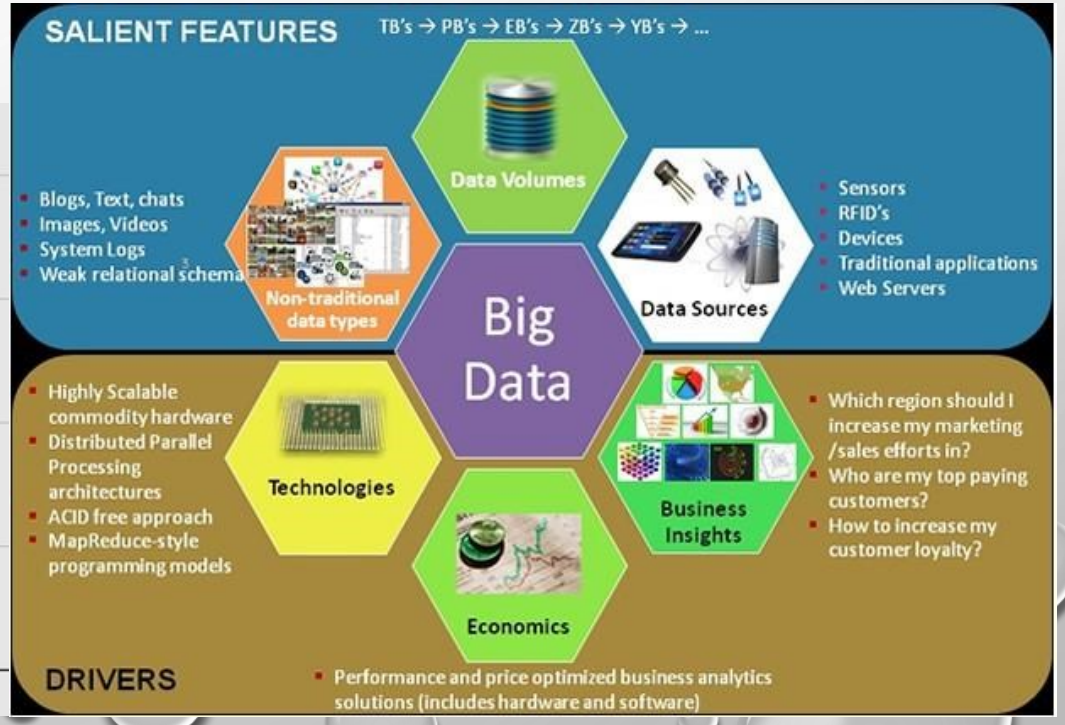
What is Big Data?

Big Data is the **outcome of datafication**. As more actions and objects are datafied, they contribute to Big Data.

Big Data refers to datasets that are

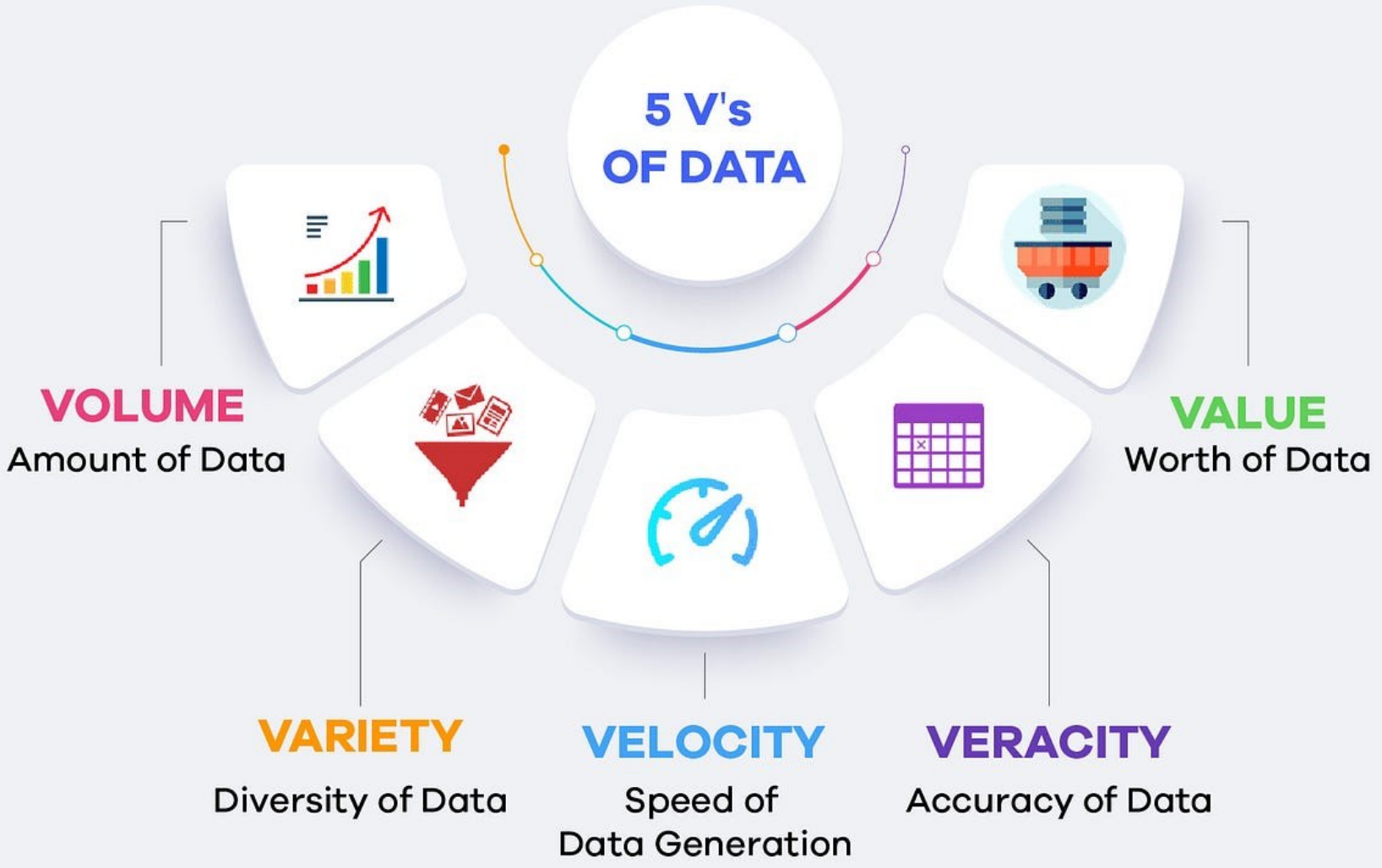
- Large in size
- Have complex relations
- Fast-growing





Characteristics or Properties of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value



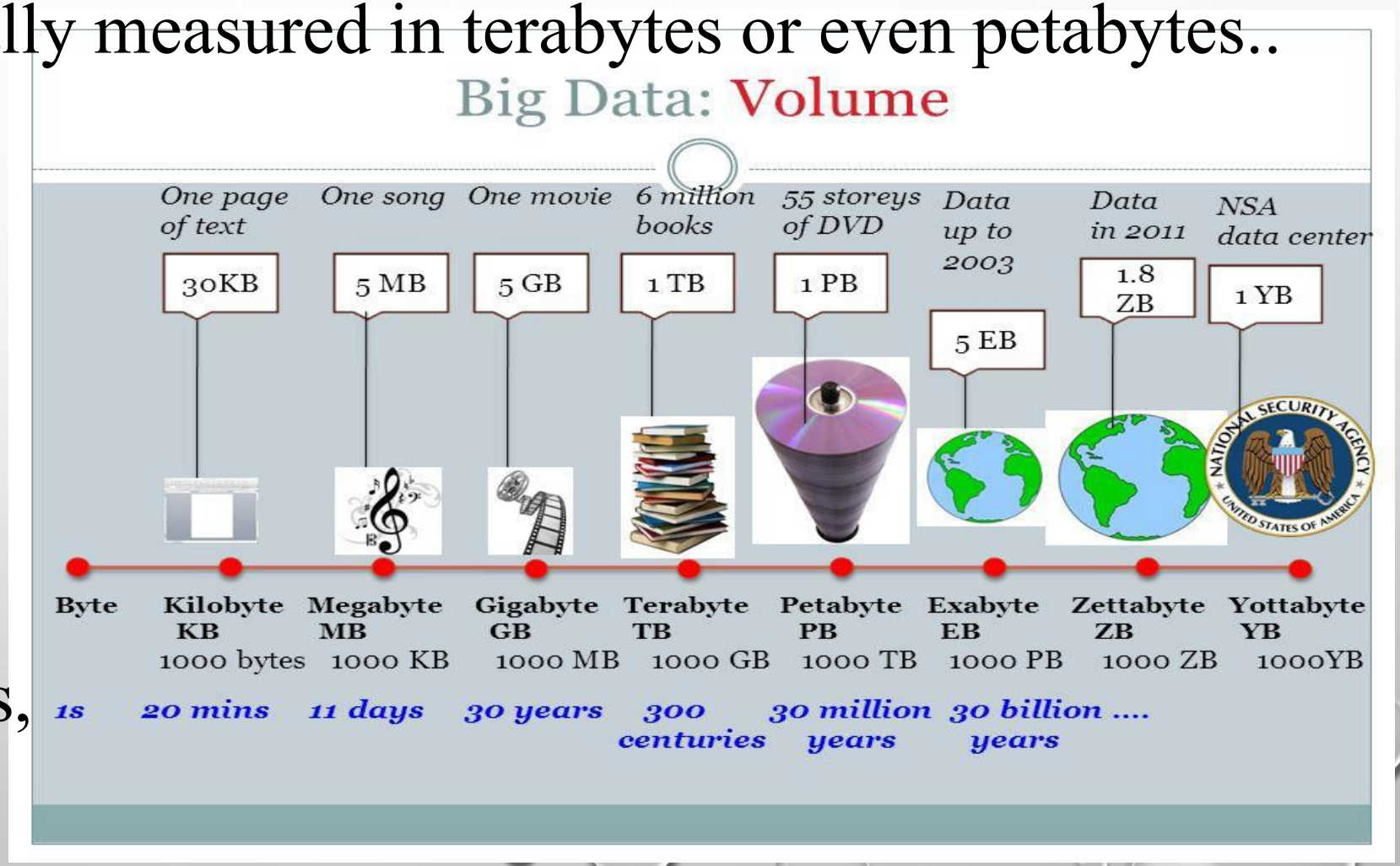
Big Data: Volume

This is about the amount of data. Big Data refers to huge amounts of data, usually measured in terabytes or even petabytes..

Example:

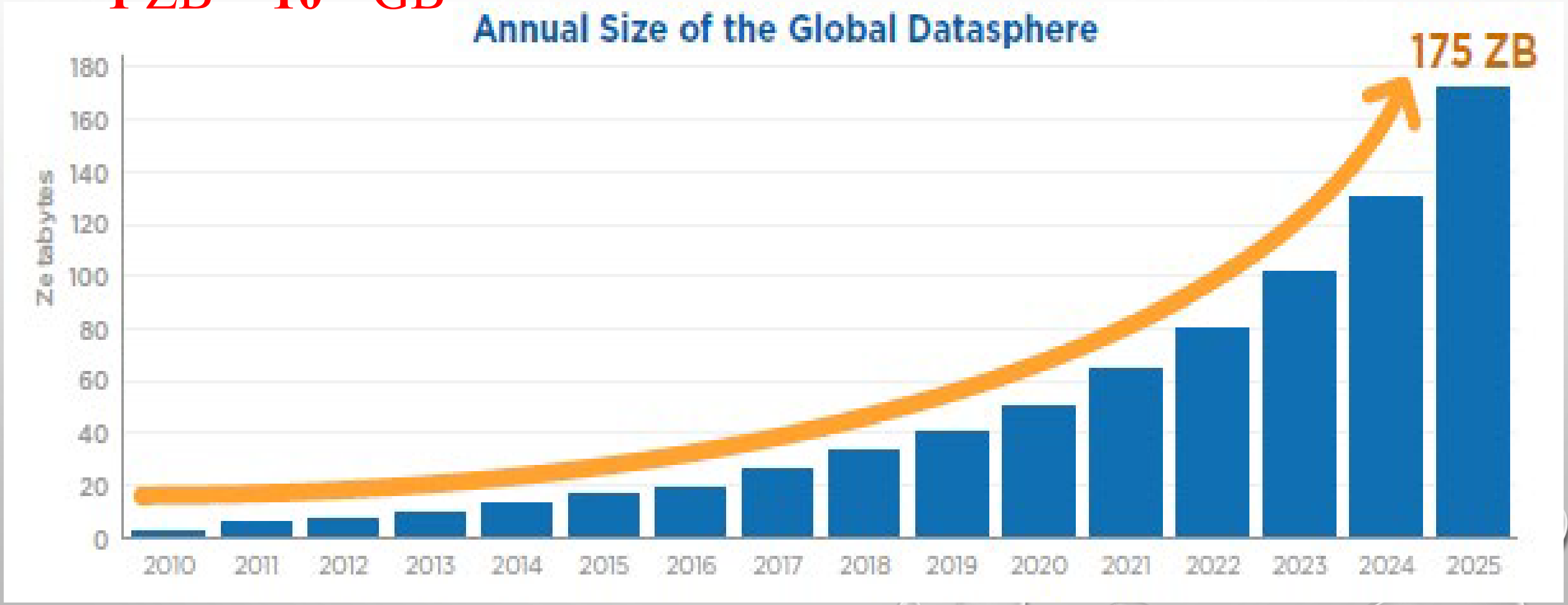
Social media platforms store

- Billions of posts
- Photos, and videos
- Manage Comments, likes



Big Data: Volume

$1\ ZB = 10^{12}\ GB$



Big Data: Variety

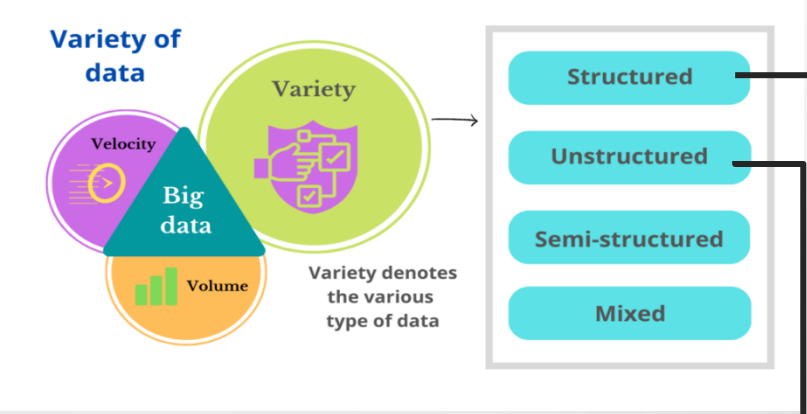
This refers to the *different types* of data.

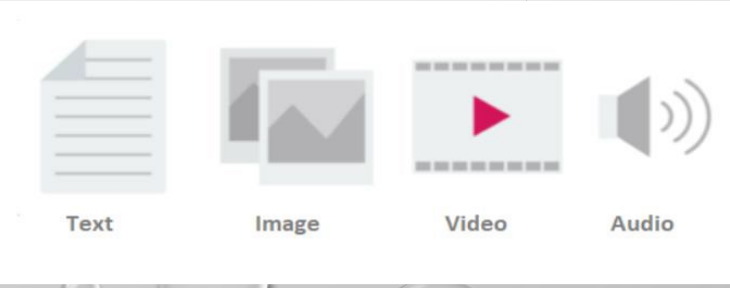
Types of data

- Structured
- Un-Structured
- Semi-Structured
- Mixed

Format of data

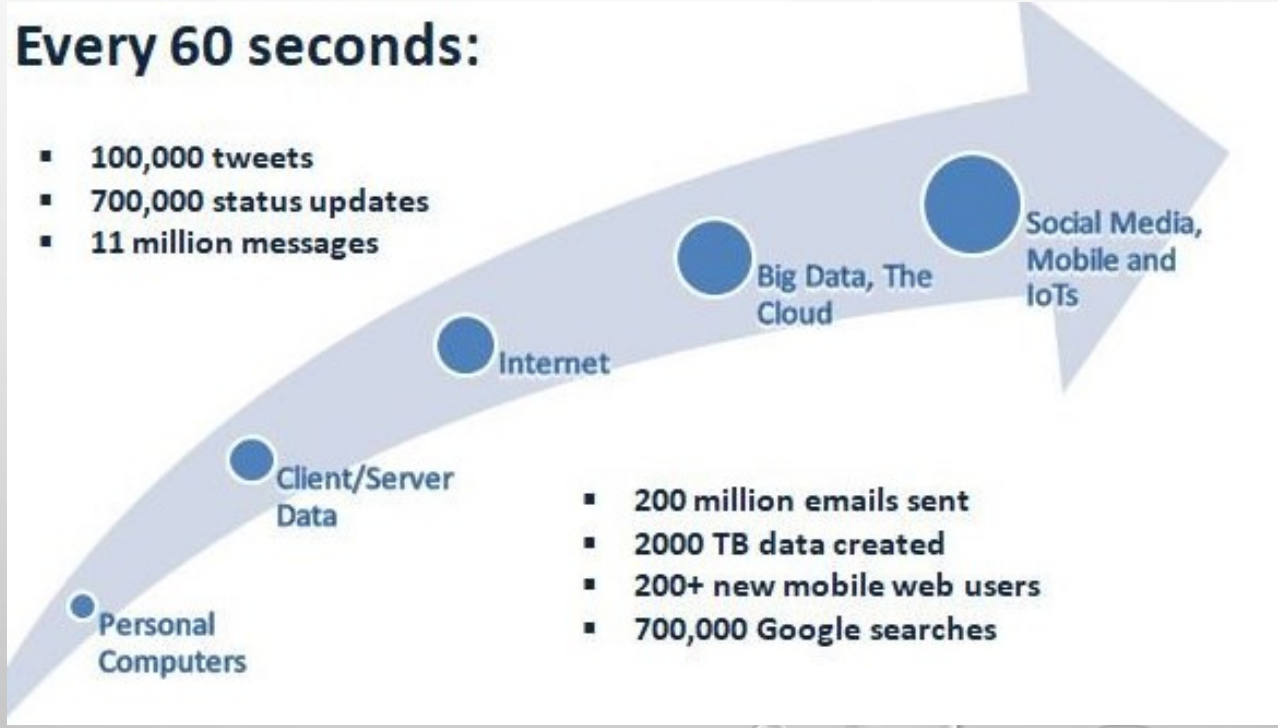
- Numerical (Sensory data)
- Text (in many different languages)
- Images (Posts)
- Videos (YouTube)
- Audios (iTunes,





Big Data: Velocity

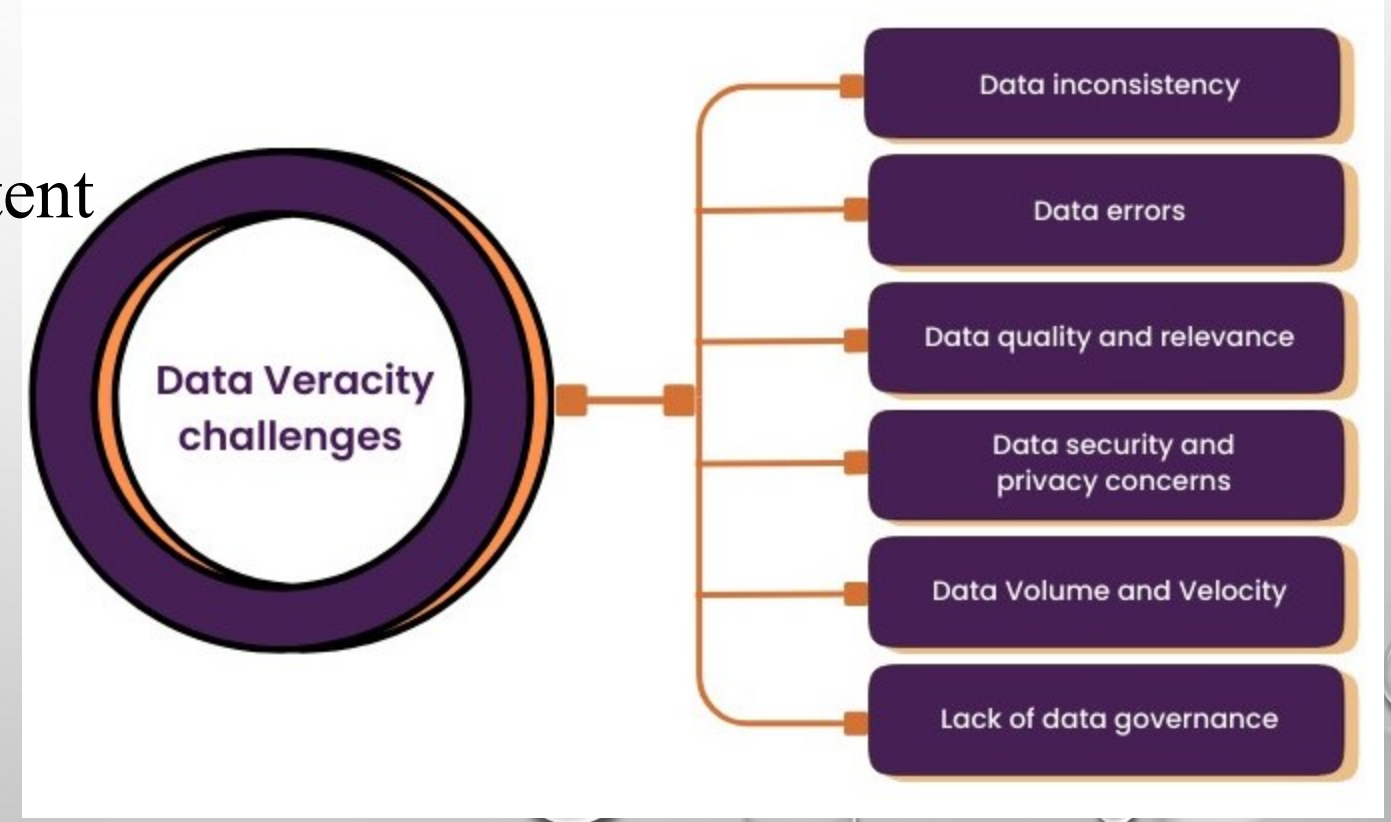
- This is the **speed at which data is created and uploaded to the internet**.
- Big Data often needs to be processed quickly, sometimes in real-time.



Big Data: Veracity

This is about the quality and accuracy of the data.
With Big Data, there can be errors, inconsistencies, fake, rumours, or "noise" that make it challenging to trust and make decisions.

Example: User-generated content
- Duplicate posts
- Fake accounts
- Spam
- Biased opinions



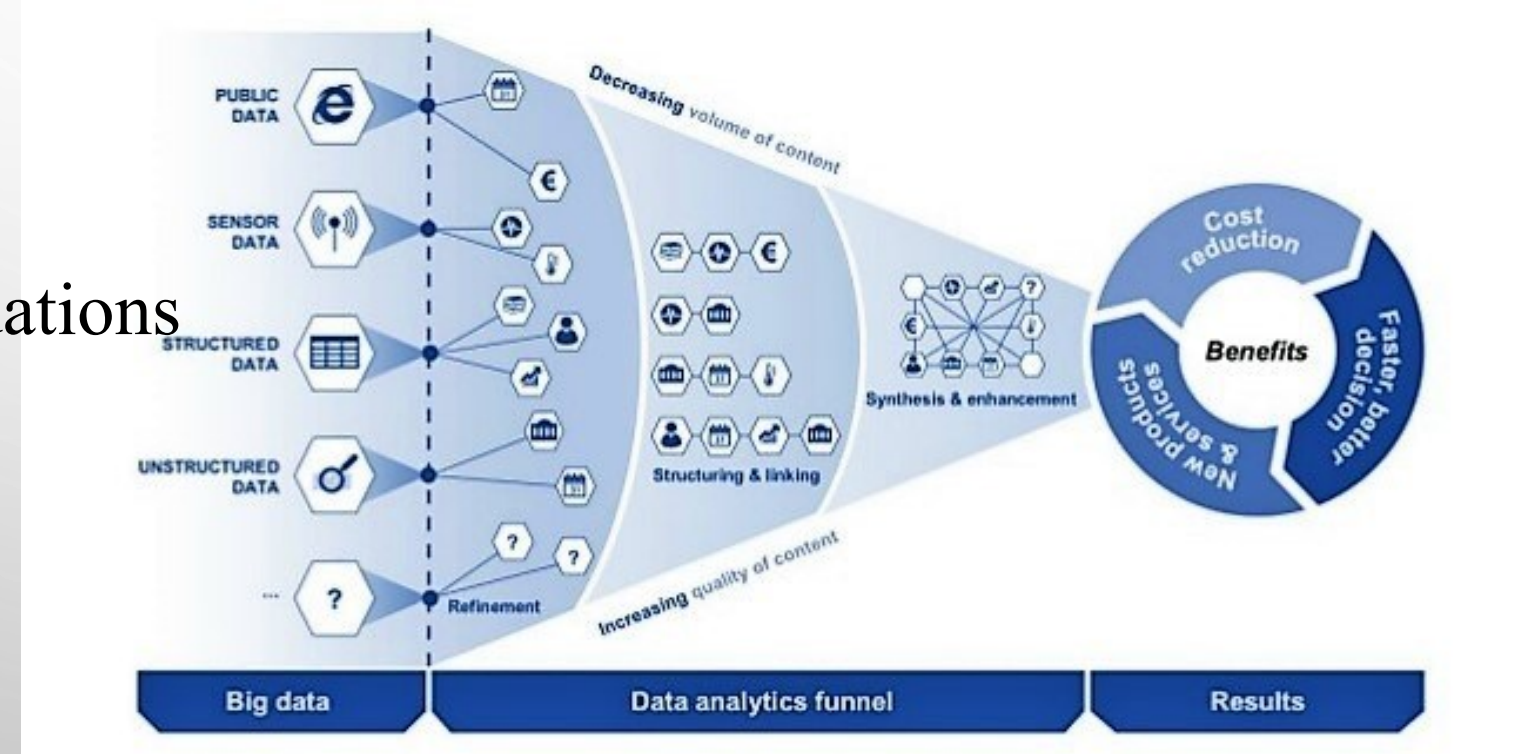
Big Data: Value

The usefulness of the data.

Big Data is only valuable if it provides insights or helps make better decisions

How it is useful

- Personalized Recommendations
- Targeted Ads
- Traffic Navigation
- Predictive Maintenance
- Healthcare Diagnostics



What is Datasets?

What is the term "Dataset"?

A dataset is **a collection of related data** used to train and evaluate Artificial Intelligence models.

Data Science often consists of **tabular data** (CSV files) with rows (examples/samples) and columns (features), such as sales records, weather data, or survey responses.

A good dataset is crucial for AI success, as it teaches the model about the specific problem it aims to solve.

Finding Datasets for practice

1. Kaggle: Your Home for Data Science
2. UCI Machine Learning Repository
3. Data.gov Home - Data.gov
4. World Bank Open Data | Data
5. GitHub: awesomedata/awesome-public-datasets: A topic-centric list of HQ open datasets.
6. OpenML
7. DataHub

Datasets we use in this workshop

1. **Iris Dataset**
   Classifies **flowers** into three species based on **sepal and petal** dimensions.
   Link: iris.csv
2. **Zoo Dataset**
   Classifies animals based on features like fur, feathers, and aquatic.
   Link: Zoo - UCI Machine Learning Repository
3. **Titanic Dataset**
   Predicts survival on the Titanic based on passenger details like age, sex, and class.
   Link: datasets/titanic.csv at master · datasciencedojo/datasets

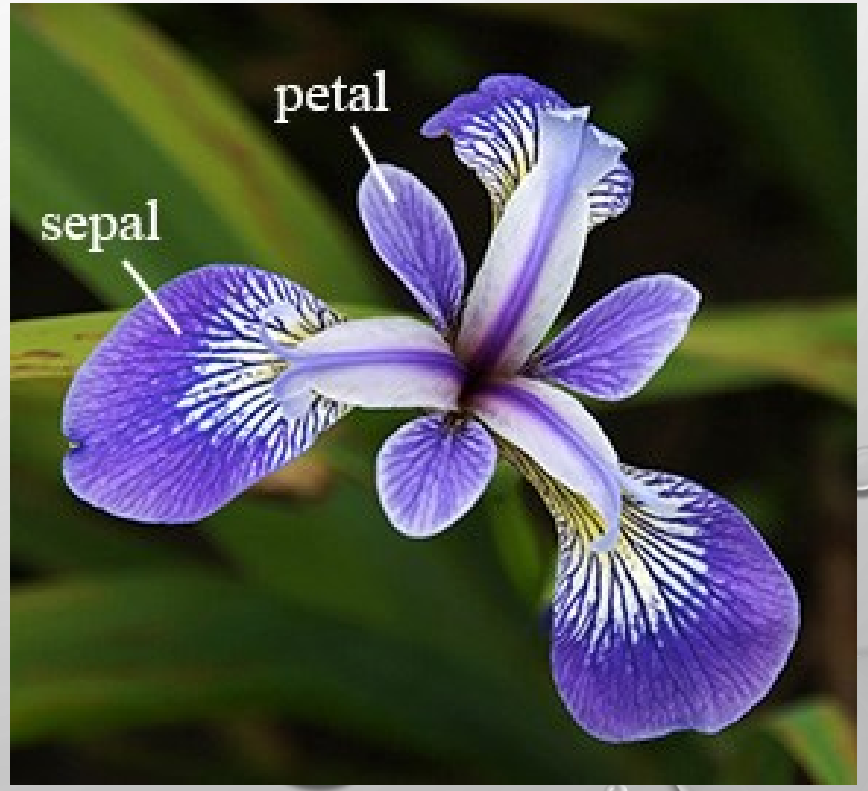
What is Iris Dataset?

- The Iris dataset was introduced by Ronald Fisher in 1936.
- Dataset contains 150 records/samples.

The dataset is used to classify different species of the Iris flower based on their physical measurements.

Features:

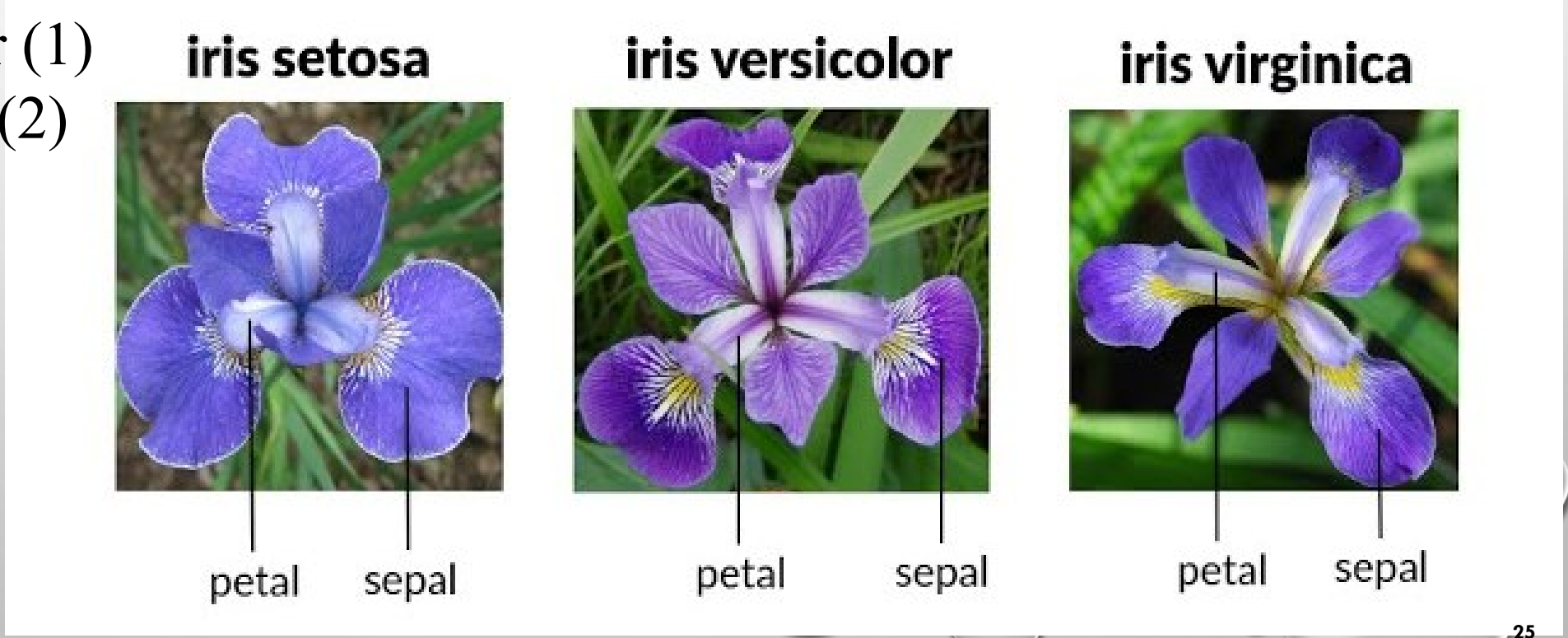
- Sepal length (in cm)
- Sepal width (in cm)
- Petal length (in cm)
- Petal width (in cm)



Example (Iris Dataset)

Target Variable: The species of the flower, which can be one of three classes:

- Iris Setosa (0)
- Iris Versicolor (1)
- Iris Virginica (2)



Example (Iris Dataset)

sepal.length	sepal.width	petal.length	petal.width	variety
5.1	3.5	1.4	0.2	Setosa
4.9	3	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
7	3.2	4.7	1.4	Versicolor
6.4	3.2	4.5	1.5	Versicolor
6.9	3.1	4.9	1.5	Versicolor
6.5	3	5.8	2.2	Virginica
7.6	3	6.6	2.1	Virginica
4.9	2.5	4.5	1.7	Virginica

What is Zoo Dataset?

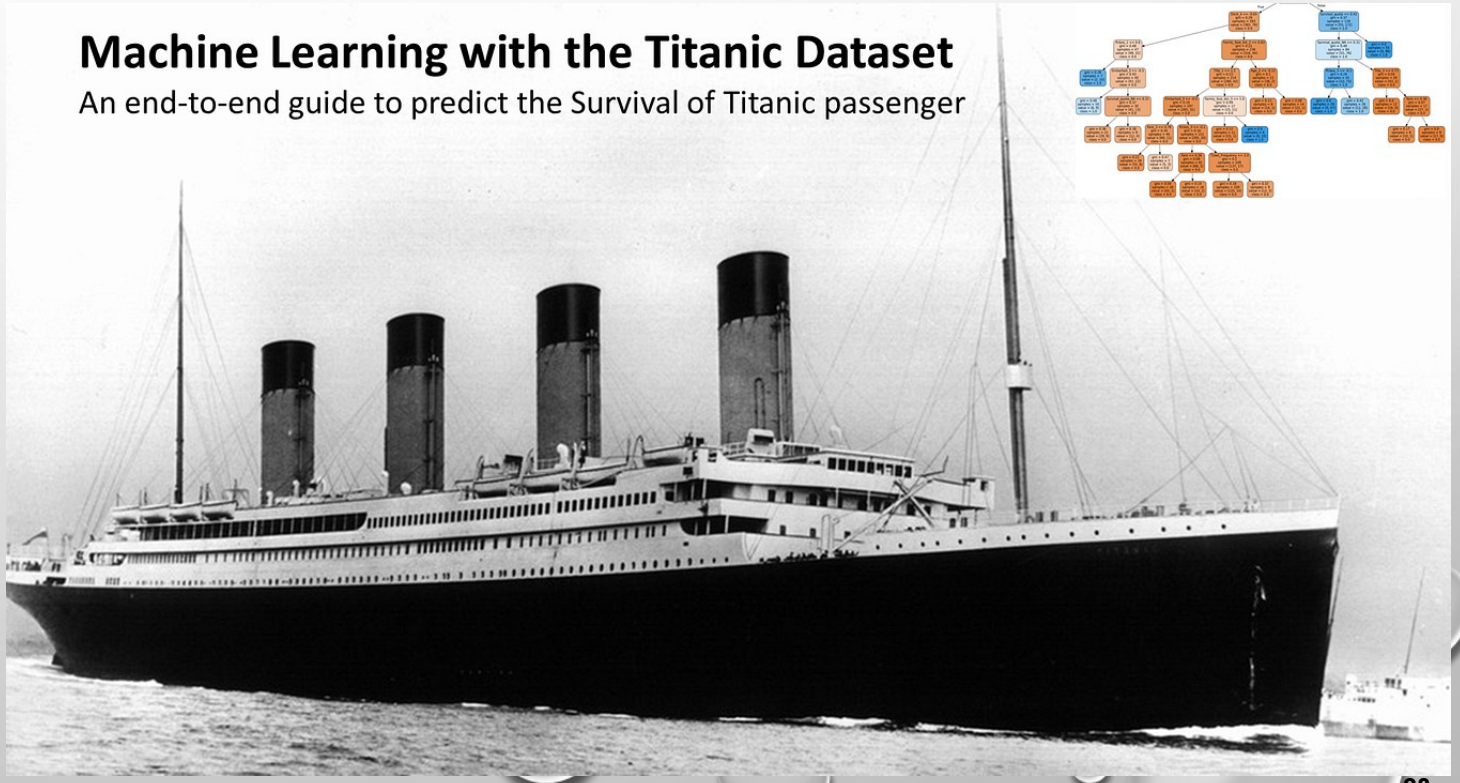
The ZOO dataset contains 59 records/samples and 16 features. There are 7 target classes in it and all the data is in discrete form.

List of Features		
Hair	Feathers	Eggs
Milk	Airborne	Aquatic
Predator	Toothed	Backbone
Breathes	Venomous	Fins
Legs	Tail	Domestic
Catsize		

What is Titanic Dataset?

The Titanic dataset contains 891 records/samples and 11 features. It is the binary classification (Survived or Not-Survived). The dataset needs the label-encoding because some features are in categorical form.

List of Features		
Passenger ID	Survived	Pclass
Name	Sex	Age
SibSp	Parch	Ticket
Fare	Cabin	Embarked



Tools

- OS: Windows 10 or 11
- IDE: VS code + Extension (python + Jupiter Notebook)
- Programming Language: Python
  - List of Libraries (main)
    - Pandas & Numpy
    - Matplotlib & Seaborn
    - SciKit-Learn

Exploratory Data Analysis

Exploratory Data Analysis

- EDA is a crucial first step in the data science process.
- It involves examining your dataset to **understand its structure**, detect patterns, **spot anomalies**, and **generate insights**.
- EDA typically helps you make decisions about data preprocessing and the types of analysis to apply later on.

```
import pandas as pd
dataFrame = pd.read_csv(r"your_dataset_file_path.csv")
```

Steps of EDA

- **Understanding Data Structure**

print(dataFrame.shape)
print(dataFrame.head())
print(dataFrame.info())

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 9 columns):
Name        object
Team        object
Number      float64
Position    object
Age         float64
Height      object
Weight      float64
College     object
Salary      float64
dtypes: float64(4), object(5)
memory usage: 32.3+ KB
```

- **Descriptive Statistics**

print(dataFrame.describe())
print(dataFrame ['column_name'].value_counts())

s = pd.Series ([2, 3, 4])
s.describe ( )

count	3.0	3 numbers
mean	3.0	mean or average
std	1.0	Standard Deviation
min	2.0	minimum value
25%	2.5	25th percentiles
50%	3.0	50th percentiles
75%	3.5	75th percentiles
max	4.0	maximum value

Steps of EDA

- **Data Cleaning**

1. ***Handling Missing Values***: *Identify and handle missing data, either by filling or dropping.*

```
data = data.dropna(inplace=True)
```

2. ***Removing Duplicates***: *If there are duplicate rows, consider removing them.*

```
data = data.drop_duplicates()
```

Steps of EDA

- **Encode Categorical Variables**

1. **Label Encoding**: For ordinal categories, assign a unique integer to each category

2. **One-Hot Encoding**: For nominal (non-ordered) categories, create dummy variables (one column per category).

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

→

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

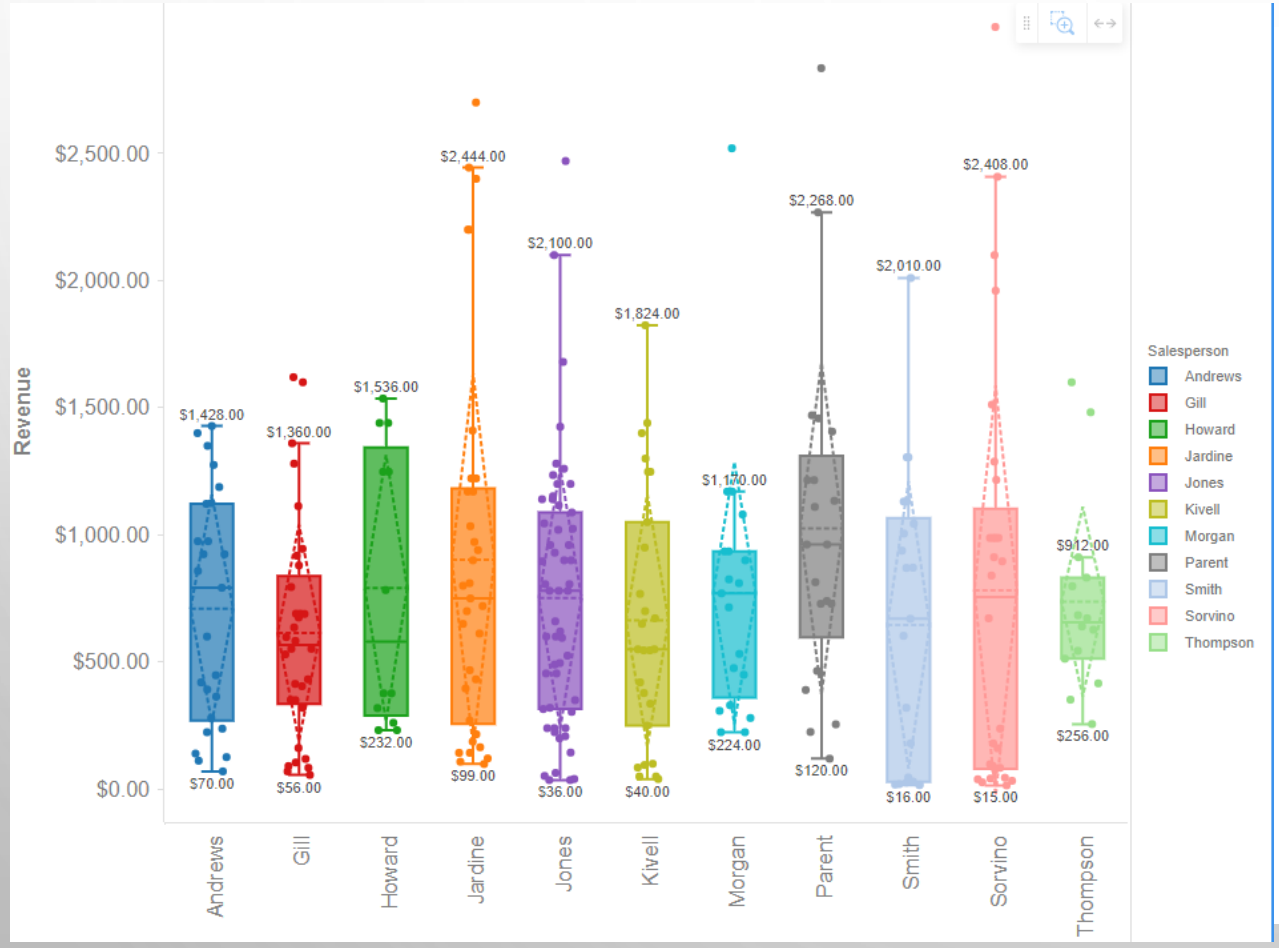
→

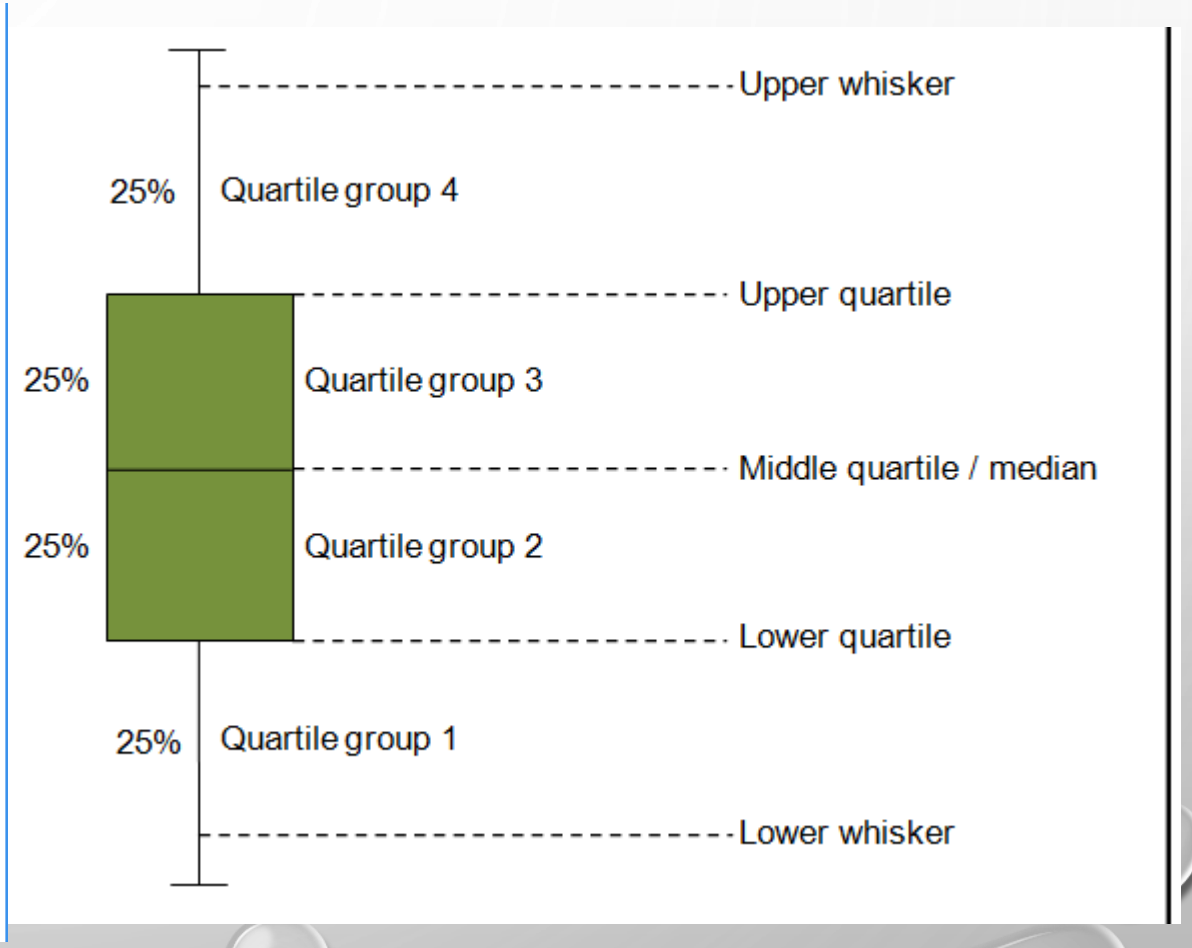
One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Steps of EDA

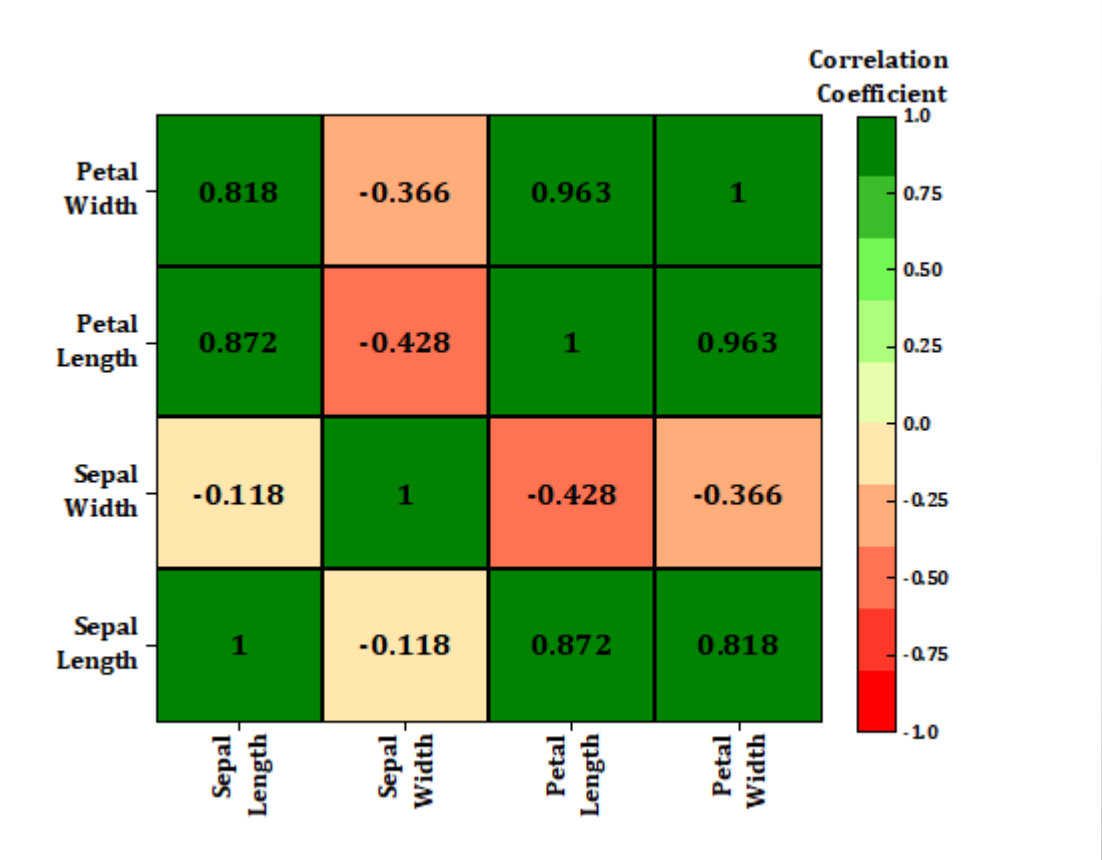
- Identifying Outliers (Box-Plot)

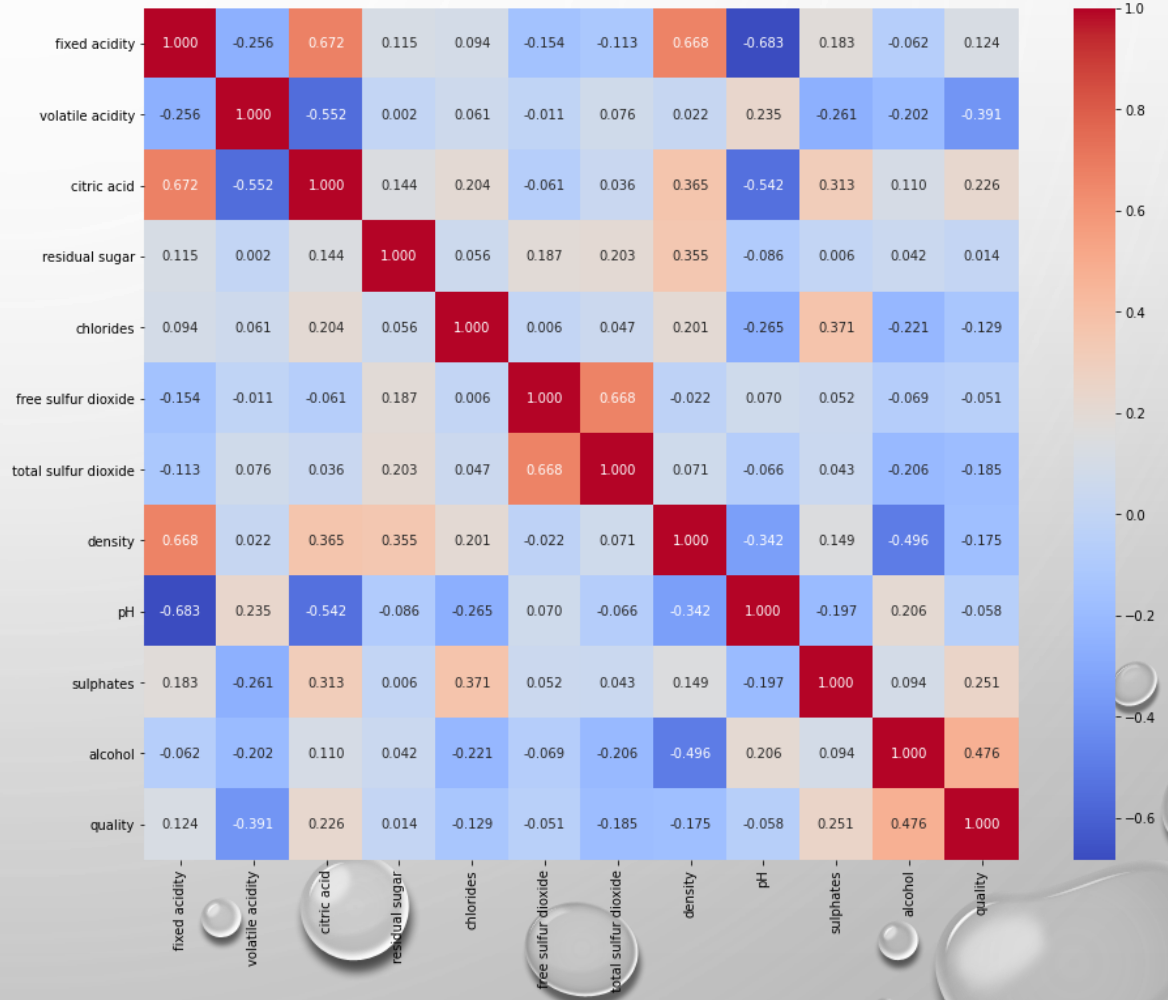




Steps of EDA

- Heat Map (you already know and draw)





Data Imbalance

Data Imbalance

In data science, **Data imbalance** happens when the classes in a dataset are not evenly distributed.

For example:

- In **medical datasets**, most patients might be healthy (90%), and only a few might have a rare disease (10%).
- In **fraud detection**, fraud transactions are far fewer than normal transactions.

Why is Data Imbalance Important?

Model behaviour which is trained on imbalanced data:

- **Bias in Predictions:** The model might only predict the majority class and ignore the minority class.
- **Fairness (**Recall, Precision**):** Models trained on imbalanced data may not provide fair or accurate results for all categories, reducing trust in their predictions.
- **Real-world Impact:** It's crucial in areas like health care, fraud detection, and fault prediction, where identifying minority classes can be lifesaving or cost-effective.

Solutions for Data Imbalance

- **Resampling Techniques:**
  1. **Oversampling**: Duplicate examples from the minority class
  2. **Undersampling**: Remove some examples from the majority class

- **Synthetic Data Generation:**
  1. **SMOTE** (Synthetic Minority Oversampling Technique)
  2. **ADAYSN** (Adaptive Synthetic Sampling)
  3. **Science-Kit Learn** library

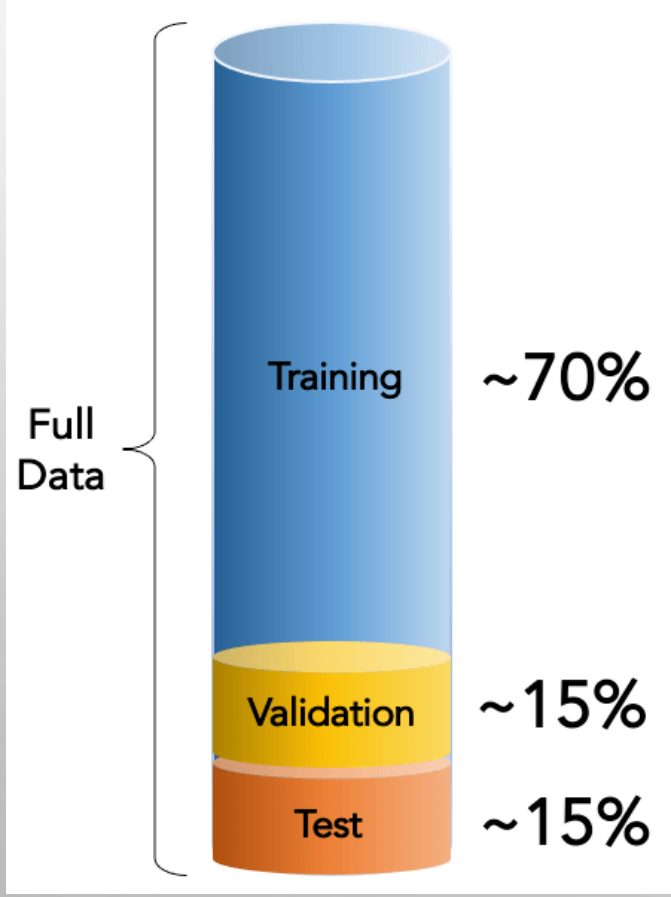
- **Algorithmic Solutions:**

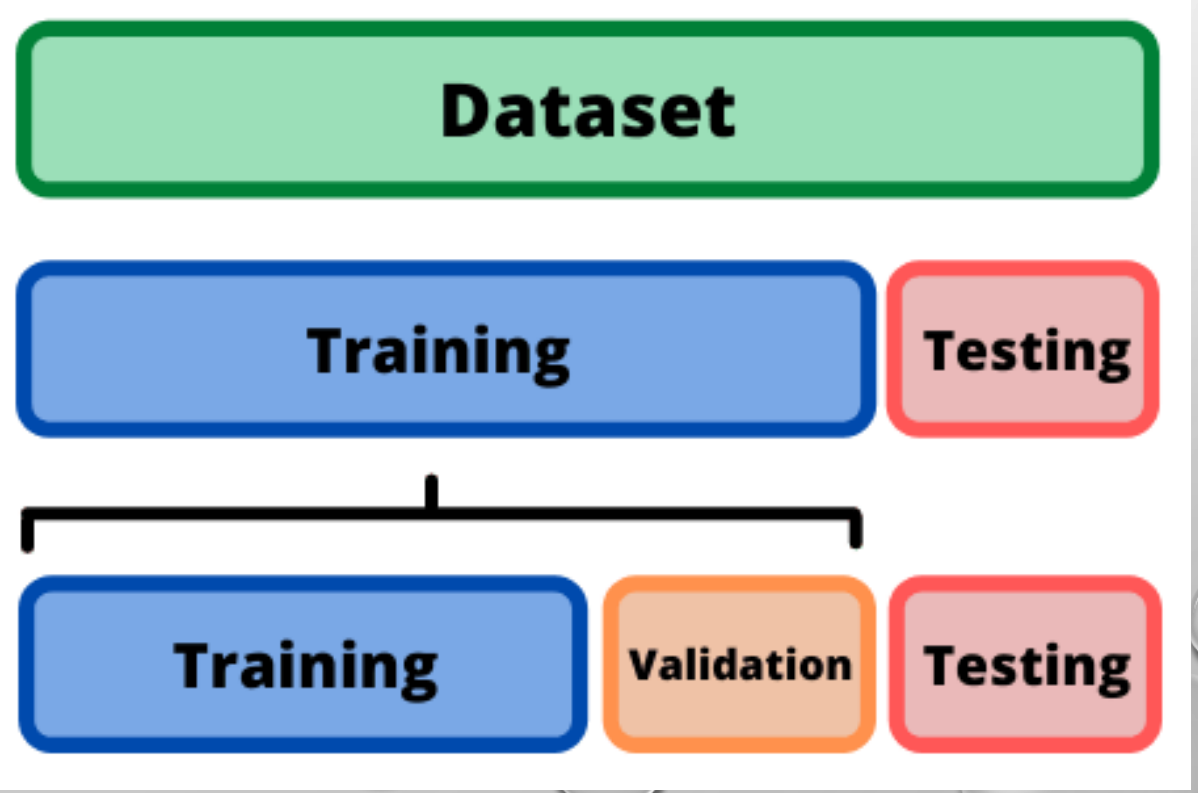
  Use algorithms designed to handle imbalance, like decision trees or **ensemble** methods

Data Splitting

Data Splitting

Data splitting divides a dataset into parts to **train**, **validate**, and **test** artificial intelligence models. It ensures that the model is evaluated on unseen data, preventing **overfitting** and improving its generalization.





The sub-datasets ratios?

The dataset is commonly divided into three parts:

- **Training Data:** Typically, 60-70% of the dataset.
- **Validation Data:** Used to tune hyperparameters and evaluate the model during training. 10-20% of the dataset.
- **Testing Data:** Used for the final evaluation of the model after training. 20-30% of the dataset.

Example Ratios:

- 60-20-20: Training (60%), Validation (20%), Testing (20%)
- 70-15-15: Training (70%), Validation (15%), Testing (15%)

- 80-20: Training (80%), Testing (20%)

Machine Learning

Machine Learning

- **Definition:** Machine Learning (ML) algorithms used to create models that allow systems to **learn from data and make decisions or predictions.**

- ML algorithms are classified into **three main types** based on how they learn from the data:

  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning.

Supervised Learning

Supervised Learning algorithms **learn from labeled data**, meaning each training example includes input data as well as the correct output (label).

The goal is to learn a **mapping from inputs to outputs** so that the model can predict the label for new, unseen data

Used for: Classification (categorizing data) and regression (predicting continuous values).

Algorithms of Supervised Learning

- **Linear Regression:** Predicts continuous outcomes by fitting a line to the data.
- **Logistic Regression:** Used for binary classification problems (e.g., spam or not spam).
- **Decision Trees:** A flowchart-like structure where each internal node represents a feature, and each leaf node represents an outcome.
- **Support Vector Machine (SVM):** Classifies data by finding the hyperplane that best separates different classes.
- **k-Nearest Neighbors (k-NN):** Classifies data points based on the closest data points in the training set.
- **Naive Bayes:** A probabilistic classifier based on Bayes' theorem, often used in text classification.
- **Neural Networks:** Models that mimic the human brain, especially useful for complex tasks like image recognition

Unsupervised Learning

- Unsupervised Learning algorithms work with **data that is unlabeled**.
- The model tries to identify patterns, groupings, or structures within the data **on its own**.
- Unsupervised learning is often used for **exploratory data analysis (EDA)** and data segmentation.

Used for: Clustering (grouping similar data) and association (finding relationships in data).

Algorithms of Unsupervised Learning

- **K-Means Clustering:** Divides data into **k clusters**, where each data point belongs to the cluster with the **nearest mean**.
- **Hierarchical Clustering:** Builds a hierarchy of clusters, either by merging or splitting clusters iteratively.
- **Principal Component Analysis (PCA):** Reduces the dimensionality of data by transforming it into a smaller number of principal components.
- **Association Rules (e.g., Apriori)**: Finds relationships between variables in large datasets, often used in **market basket analysis**.
- **Autoencoders:** A type of neural network used to **learn efficient data codings**, primarily for dimensionality reduction.

Reinforcement Learning

Reinforcement Learning (RL) is a type of ML where an agent/model **learns by interacting with an environment and receiving feedback through rewards or penalties**.

The agent/model aims to **maximize the cumulative reward over time** by choosing the best actions.

RL is widely used in games, robotics, decision-making tasks, and sequential tasks

Algorithms of Reinforcement Learning

- Q-Learning
- Deep Q-Networks (DQN)
- Policy Gradient Methods
- Actor-Critic Methods
- Proximal Policy Optimization (PPO)

List of models include in this workshop

1. Decision Tree
2. Random Forest
3. Support Vector Machine
4. K Nearest Neighbour
5. Naïve Bayes

Evaluation Matrix

Need of Evaluation

The evaluation process is the way to check **how well a model or program works**. This process uses different methods **to see if the model's predictions (what it thinks will happen) are close to the actual results**.

Evaluating a model is important because **it shows the model has learned useful patterns or just memorized the data**. A good evaluation process helps to pick the best model, find any mistakes, and improve its performance.

Confusion Matrix

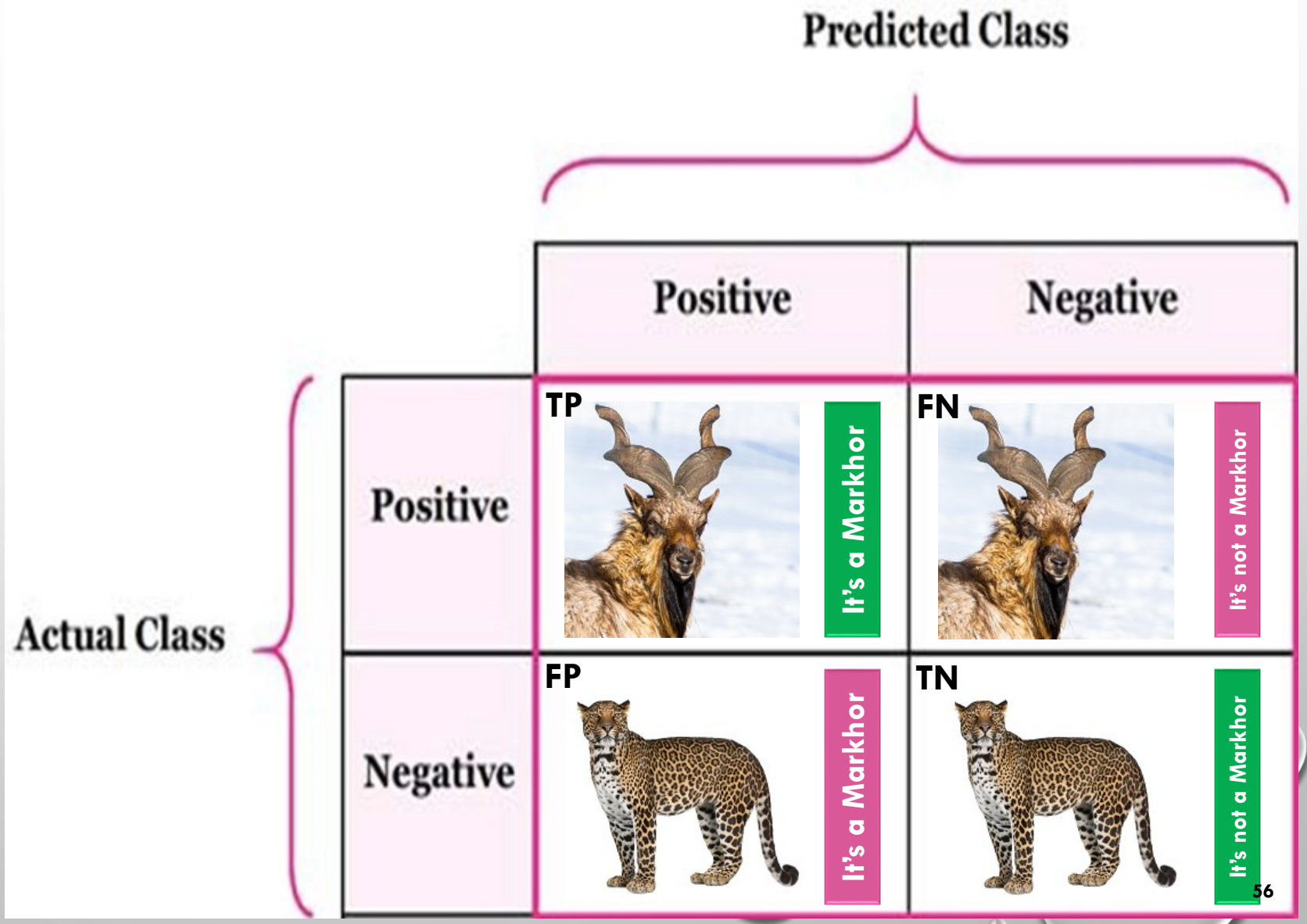
- A confusion matrix is a table that helps visualize the performance of a classification model.

- The number of correct and incorrect predictions are summarized with count values and broken down by each class.

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)

Confusion Matrix

- **True Positive**
- **False Negative:**
  (Type 1 Error)
- **False Positive:**
  (Type 2 Error)
- **True Negative:**



Confusion Matrix

		Predicted values		
		Positive	**Negative**	Totals
Actual Values	**Positive**	TP	FN	**P = (TP + FN) = Actual Total Positives**
	Negative	FP	TN	**N = (FP + TN) = Actual Total Negatives**
	Totals	Predicted Total Positives	Predicted Total Negatives	

Example of CM

Actual class\Predicted class	buy_computer (yes)	buy_computer (no)	Total
buy_computer (yes)	6954	46	7000
buy_computer (no)	412	2588	3000
Total	7366	2634	10000

Can you answer these questions:

a) How many computers were bought? = 7000

b) What is the value of **prediction** about the sale of computers? = 7366

c) What is the numbers of computers which were **actually bought** and the algorithm also predicted it correctly? = 6954

d) What is the total number of samples in the data? = 10000

1. Accuracy

The accuracy evaluation metric is one of the **simplest and most widely used evaluation measures** for the performance of classification models.

Accuracy is **the ratio of correctly predicted records/ samples** (both positive and negative) to the total number of testing records/ samples.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Range: Accuracy values range from 0 to 1, where:

0 means the model is always wrong.

1 (or 100%) means the model is always correct.

Example of Accuracy

	buy_computer (yes)	buy_computer (no)	Total
buy_computer (yes)	6954	46	7000
buy_computer (no)	412	2588	3000
Total	7366	2634	10000

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} = \frac{(6954 + 2588)}{(6954 + 2588 + 412 + 46)} = \frac{(?)}{(10{,}000)} = ?$$

- This algorithm has ___ % accuracy.

2. Recall

Recall is used to check **how well a model finds all the true positive cases** in a dataset.
It's especially **useful when the model missing positive cases** (like failing to detect a disease) is a bigger problem than having a few extra false positives

$$Recall = \frac{True\ Positives}{True\ Positive + False\ Negative}$$

Range: Recall values range from 0 to 1, where:

0 means the model missed all the positive cases.

1 (or 100%) means the model correctly found every positive case.

Example of Recall

	buy_computer (yes)	buy_computer (no)	Total
buy_computer (yes)	6954	46	7000
buy_computer (no)	412	2588	3000
Total	7366	2634	10000

$$recall = \frac{(TP)}{(TP + FN)} = \frac{(6954)}{(6954 + 46)} = \frac{(6954)}{(7000)} = ?$$

- This algorithm has ___ % recall.
- Recall is valuable when the dataset is **imbalanced**.
- In such cases, accuracy can be misleading, but recall focuses on finding all positive samples.

3. Precision

It shows how **accurate the positive predictions of a model are**.
It is useful when more important to avoid false positives (wrongly marking something as positive) than it is to find every positive case.

$$Recall = \frac{True\ Positives}{True\ Positive + False\ Positive}$$

Range: Precision values range from 0 to 1, where:

0 means that all positive predictions are incorrect.

1 (or 100%) means that all positive predictions are correct.

Example of Precision

	buy_computer (yes)	buy_computer (no)	Total
buy_computer (yes)	6954	46	7000
buy_computer (no)	412	2588	3000
Total	7366	2634	10000

$$Precision = \frac{(TP)}{(TP + FP)} = \frac{(6954)}{(6954 + 412)} = \frac{(6954)}{(7{,}366)} = ?$$

- This algorithm has ___ % precision.
- Recall is valuable when the dataset is **imbalanced**.
- In such cases, accuracy can be misleading, but recall focuses on finding all positive samples.

4. F1-Score

The F1 Score is the **harmonic mean of Precision and Recall**.

$$F1\ Score = \frac{2 \times (Precison \times Recall)}{Precison + Recall}$$

Range: The F1-Score ranges from 0 to 1, where:

0 means either precision or recall is zero (perform poorly).

1 means both precision and recall are perfect (perform very well).

Calculating F1-Score

	buy_computer (yes)	buy_computer (no)	Total
buy_computer (yes)	6954	46	7000
buy_computer (no)	412	2588	3000
Total	7366	2634	10000

$$F1\ Score = \frac{2\times(Precison\times Recall)}{Precison+Recall} = \frac{2\times(0.9934\times 0.9440)}{0.9934 + 0.9440} = ?$$

- This algorithm has ___ F1-Score.

5. ROC Curve

- It is the **Receive Operating Characteristics** curve.
- It is a plot of the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)**.

- True positive rate:

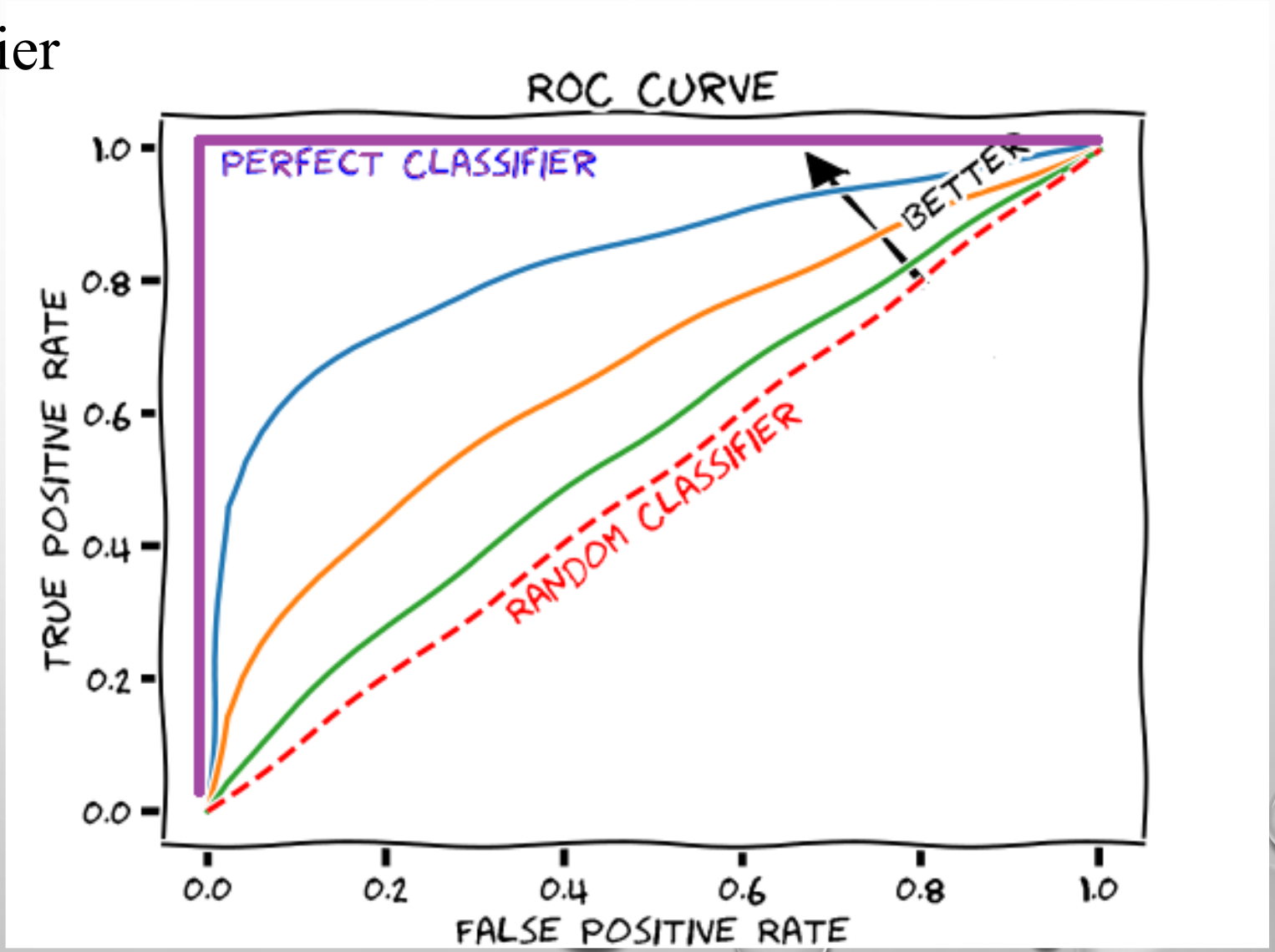
$$TPR = \frac{TP}{TP + FN}$$

- False positive rate (or Fallout):

$$FPR = \frac{FP}{TN + FP}$$

ROC Explain

- Go for the perfect classifier




QUESTIONS

Why Me? | Understanding Qadar with Dr. Omar Suleiman

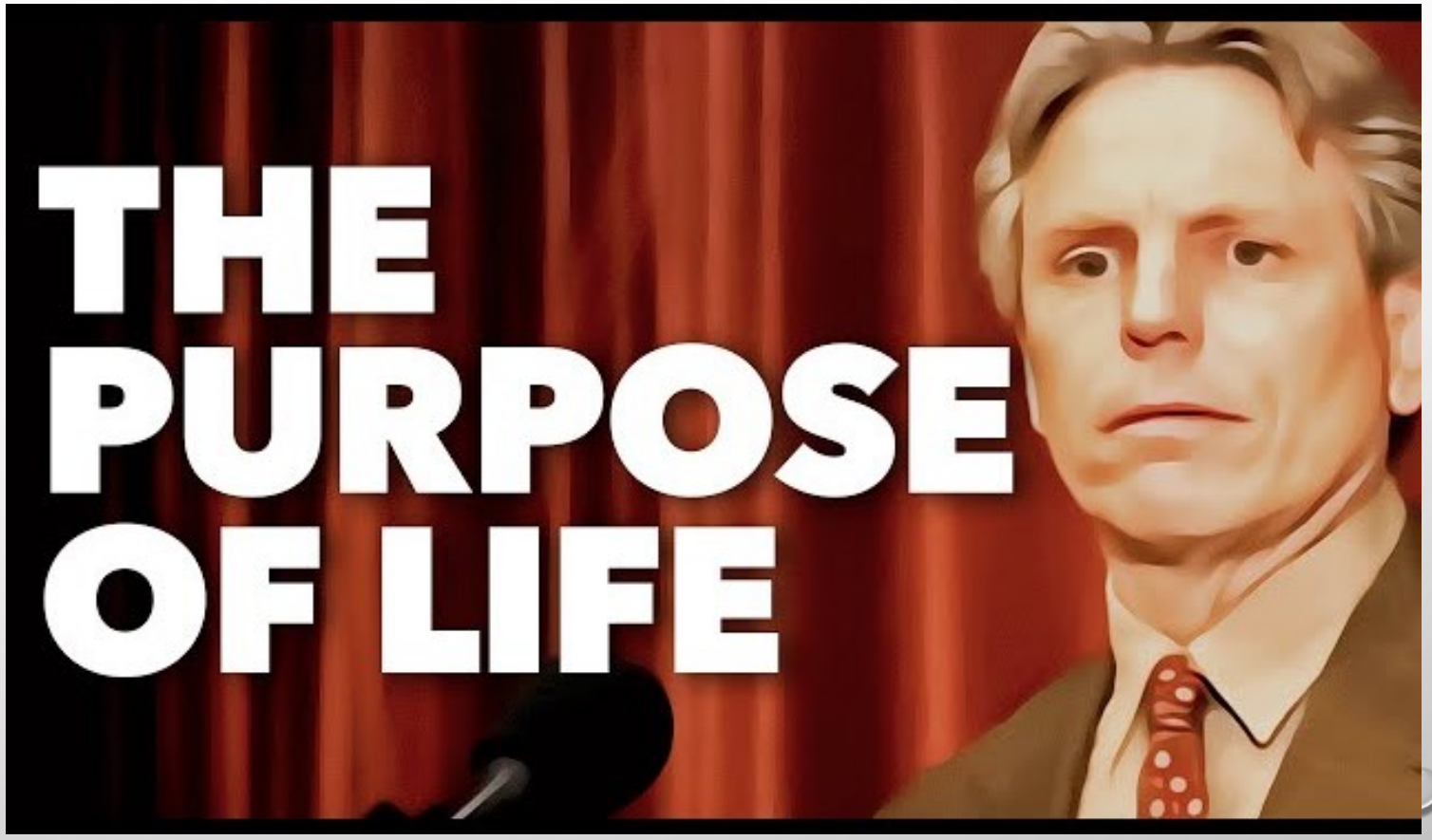
URL: [Why Me? | Understanding Qadar with Dr. Omar Suleiman | Ramadan Series 2024 TRAILER]

Number of episodes: 30

Total Duration: 05 ~ 06 Hrs

The Purpose of Life – Prof. Jeffrey Lang

- **Concept of the first test of Hazrat Adam (A.S) in Jannah.**
- **Cycle of life in this temporary world.**
- **Which qualities do we need to build in ourselves?**



URL: [The Purpose of Life - Jeffrey Lang]
Total Duration: 01:31 Hrs