BTech (Elements of AIML)

Overview of Machine Learning Life Cycle

SOURBH KUMAR

Content

- About me
- Introduction
- Machine Learning Life Cycle.
    - Data Collection
    - Enforcement Data Analysis
    - Data Cleaning
- Basic of Linear Regression
- Cost function
- Optimisation
- Regularization
- Testing (Confusion Matrix) (1-2 Min)
- Coding Example (2-3 Min)

Introduction

Definition: A subset of AI that enables systems to learn from data and improve over time without being explicitly programmed.

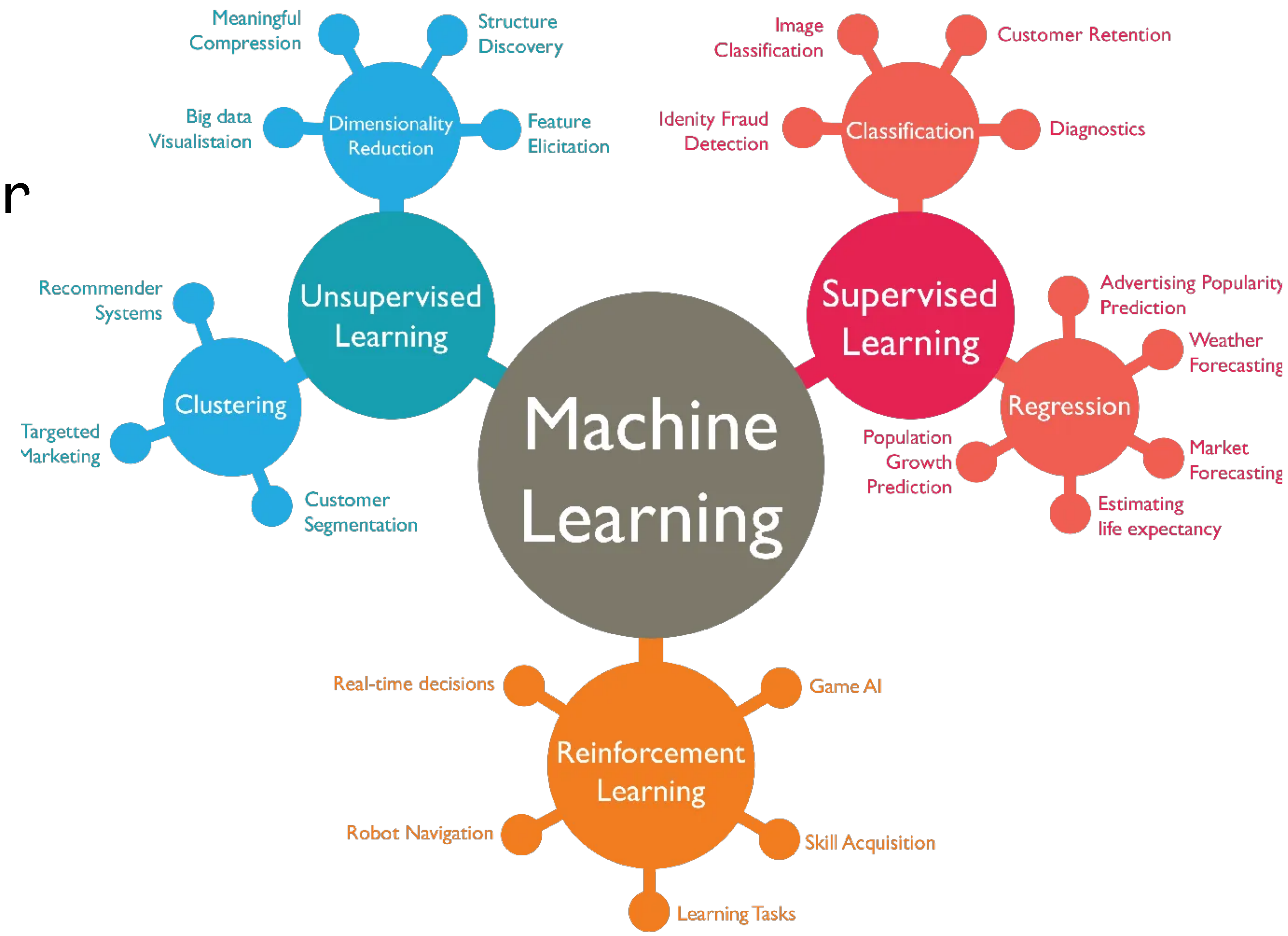
Type of Machine learning:

1. Supervised : a. Regression Model

b. Classification Model

2. Unsupervised Model:

3. Reinforcement Learning



Overview of Machine Life Cycle.

1. Collection of Data
2. Exploratory Data Analysis:

    b. Data Visualisation

    c. Hypothesis Testing(Statistics Analysis)

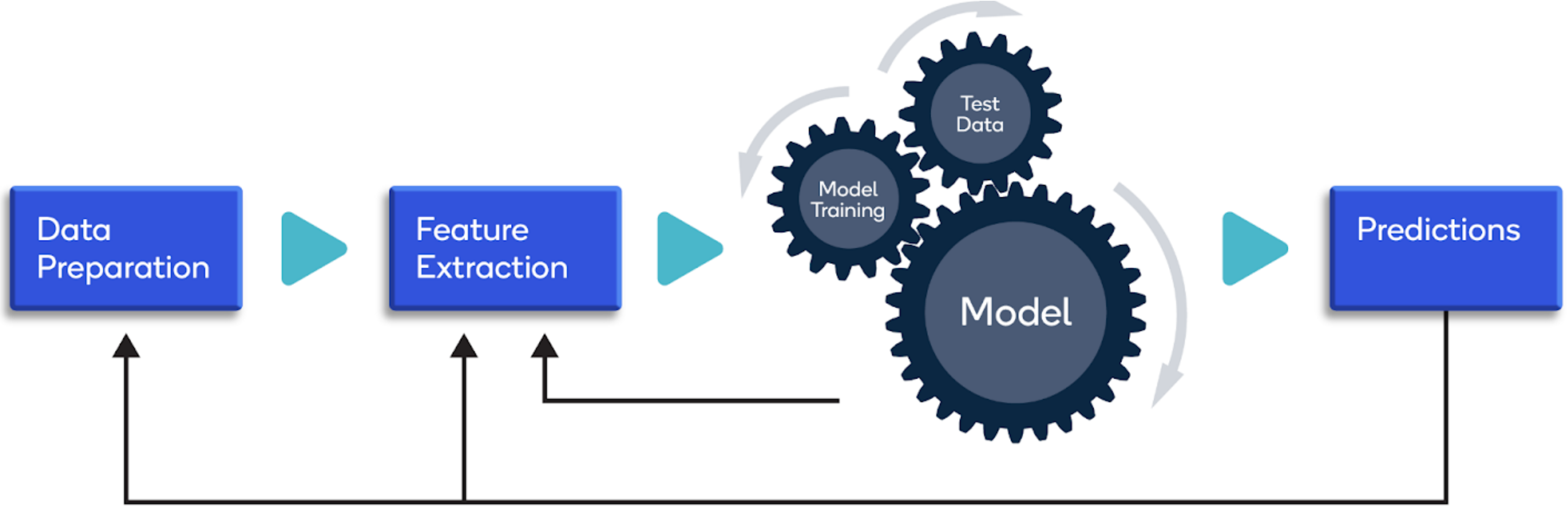
    d. Data Cleaning: Handling Missing Value, Outlier

    e. Feature Engineering

3. Model Building: a. Trained the model

    B. Regularisation(L1, L2)

4. Model Evaluation: Testing and Selection
5. Deploy the model: Server or Cloud.



ML Life cycle

Data Collection

- Data: Gathering relevant data from various sources. Such: APIs, Web Scraping, Surveys, Sensor Data, Public Datasets, Databases.

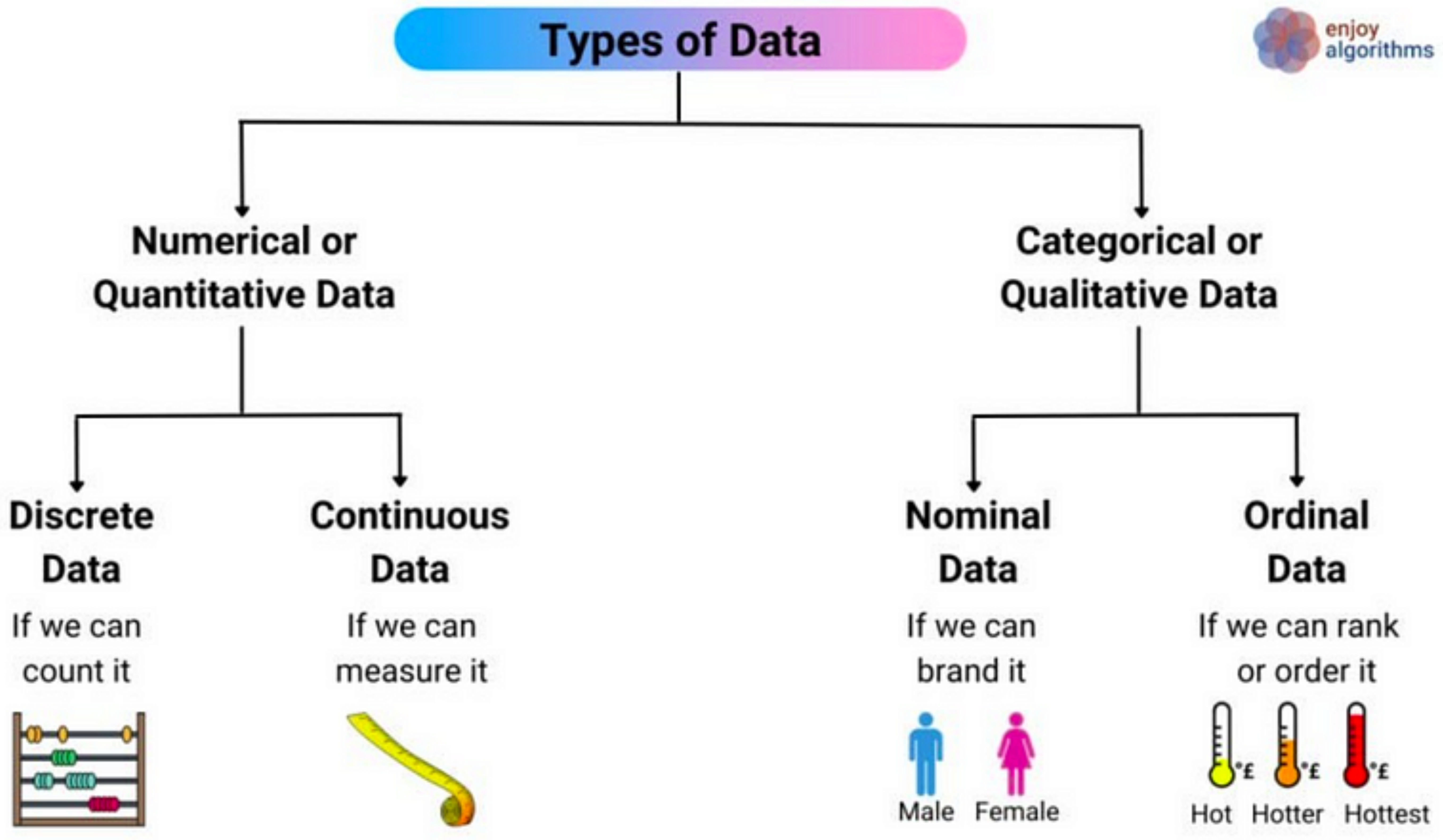
- Types of Database:

a. Structured: Relational Data base(SQL)

- Numerical, Categorical data

b. Unstructured: No SQL(Image, Audio)

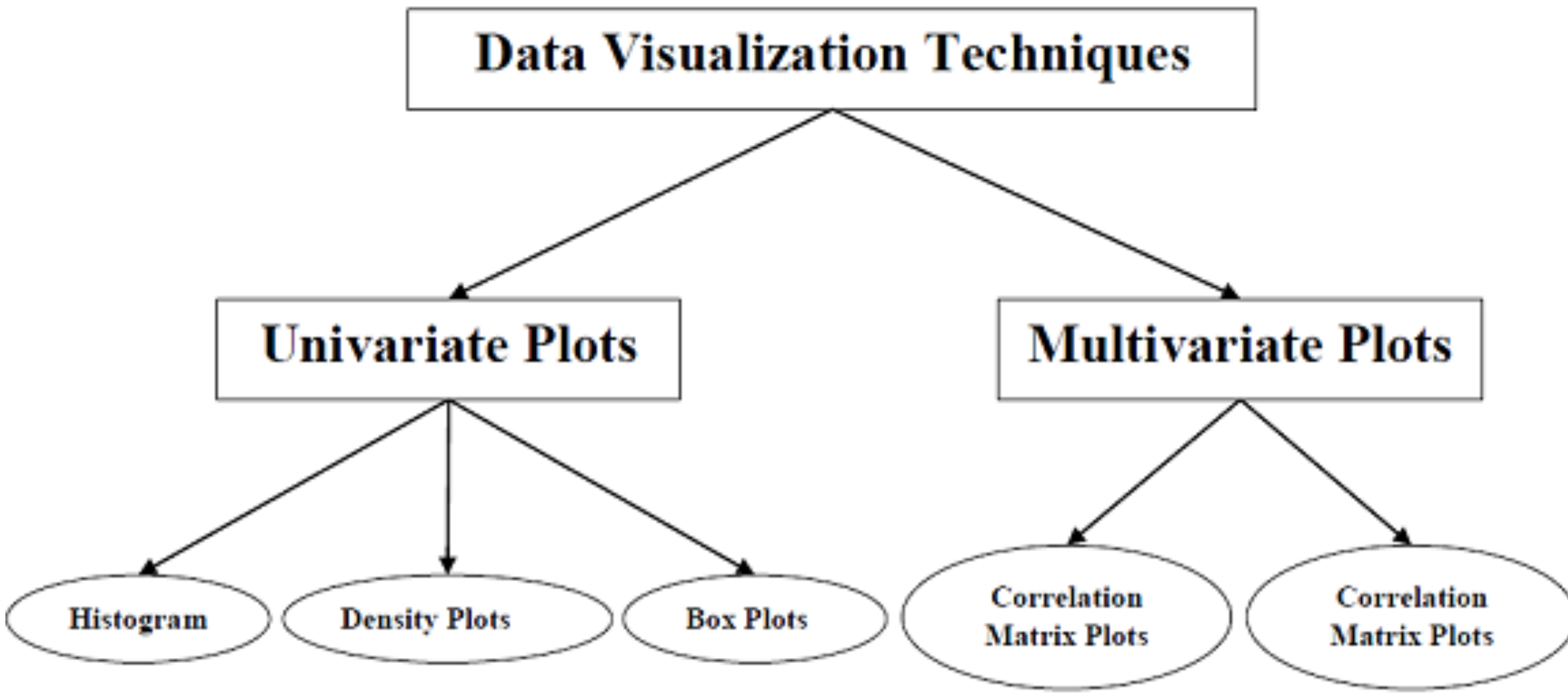
c. Semi-structured: (Xml, Json)



Caption

Exploratory Data Analysis

- Understanding the dataset and uncovering patterns.
- Tools: Visualisation, summary statistics.



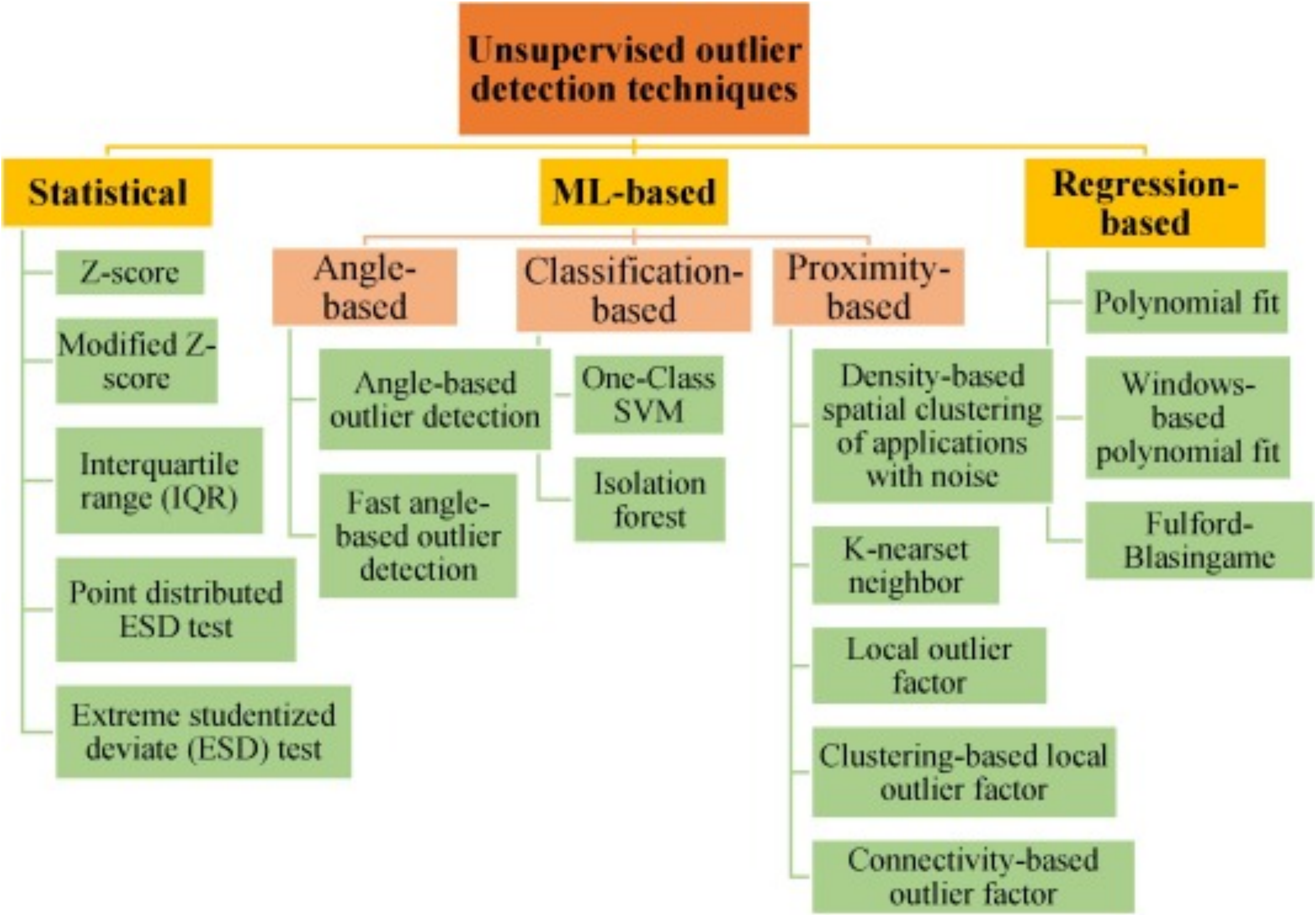
Caption

Data Cleaning

- In Data Cleaning there are two major Parts:

1. Outlier Detection

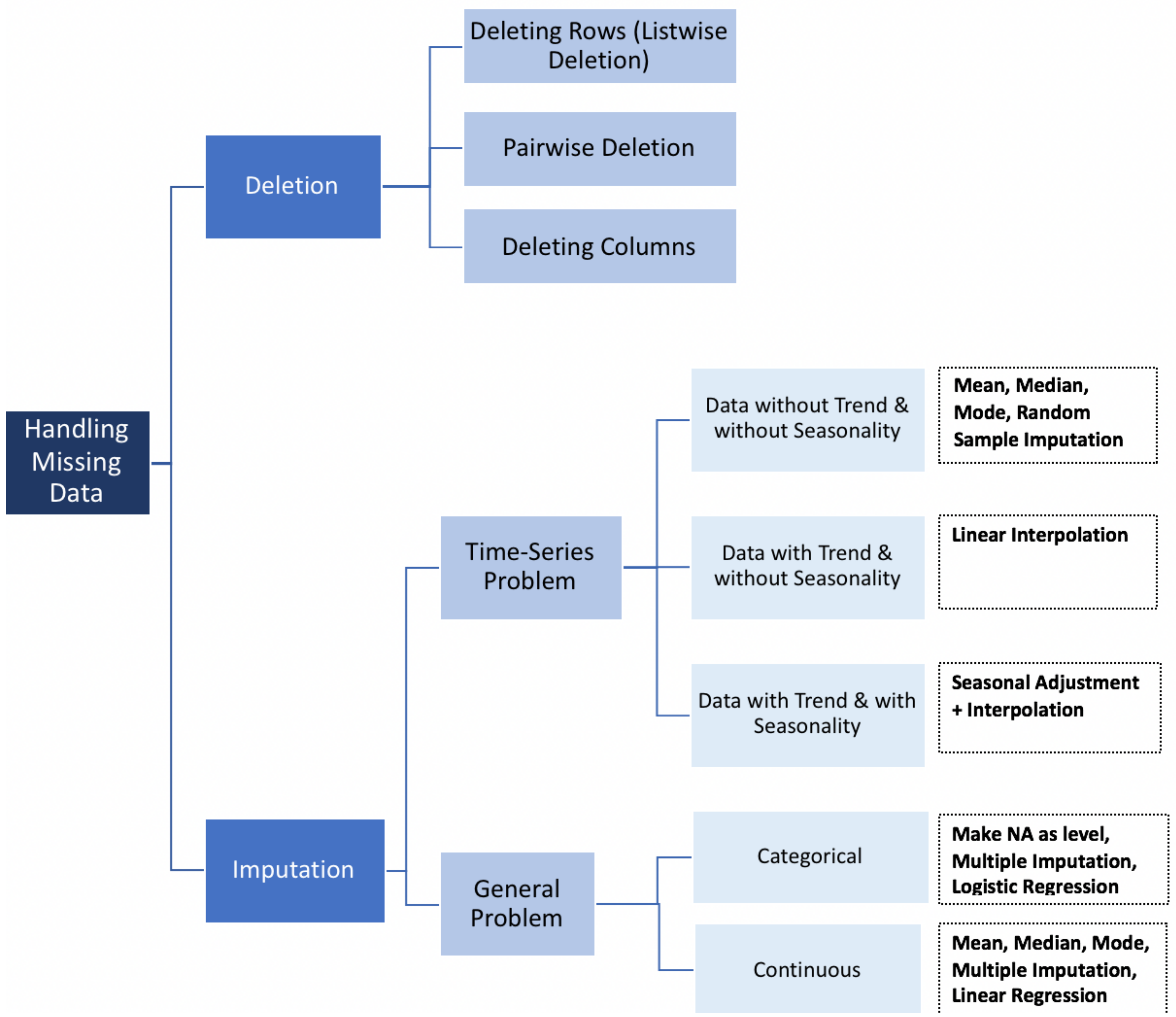
2. Handling Missing Value



Caption

Handling Missing Value:

Handling Missing Value:



Linear regression Model

- **Definition**: Linear approach to modelling the relationship between a dependent variable and one or more independent variables.
- Mathematical Equation:

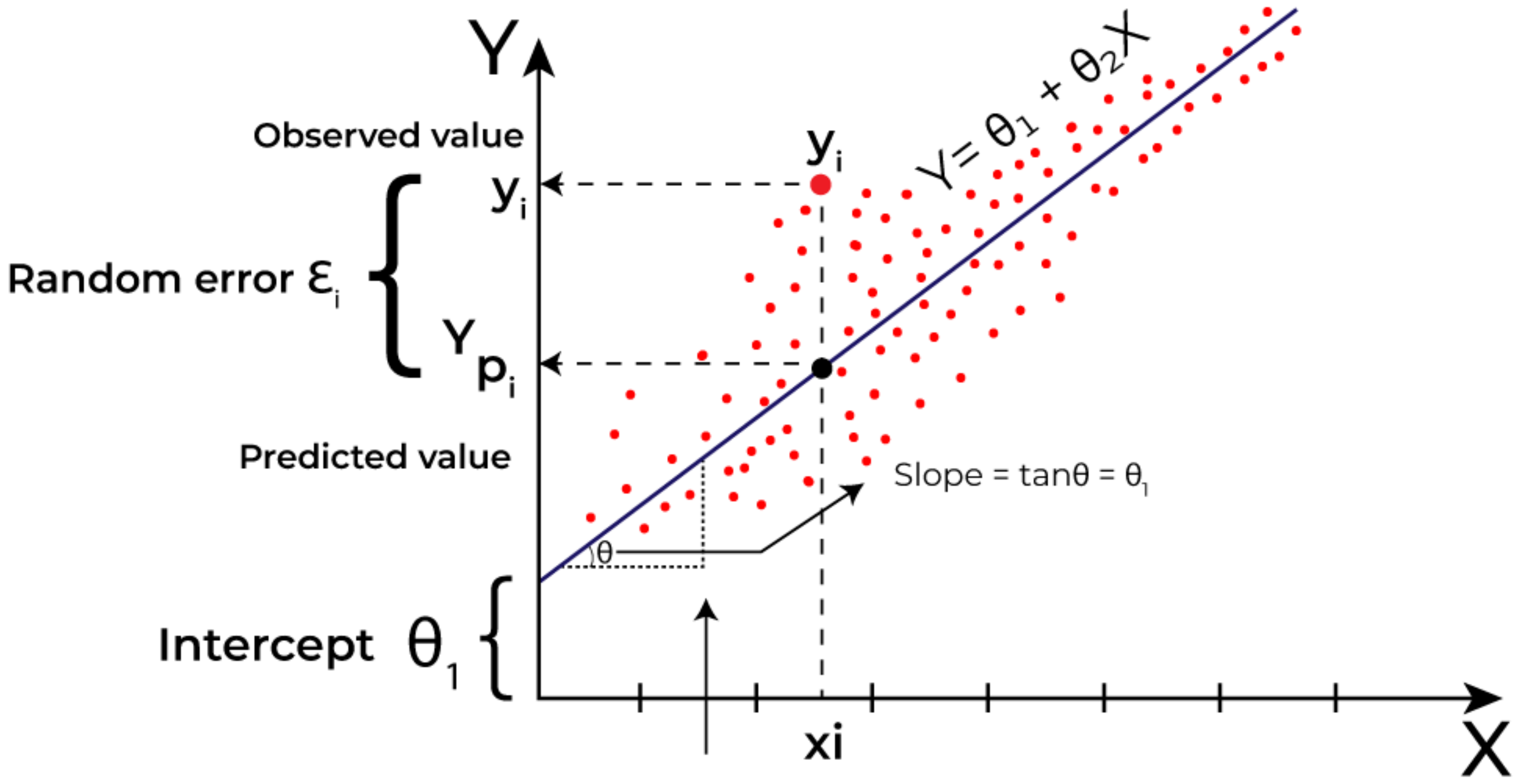
Constant/Intercept

Independent Variable

$$Y_i = \beta_0 + \beta_1 X_i$$

Dependent Variable

Slope/Coefficient



Cost function(Loss or Error Function)

- The cost function measures how well the model's predictions match the actual data. Various way:

1. Mean Absolute Error (MAE)
2. Mean Square Error(MSE)
3, Root Mean Square Error (RMSE)

- For linear regression, the most commonly used cost function is the Mean Squared Error
- Ratio= Sum of Square error for Own Model/ Sum of Square error for Base Model
- R squared= 1- Ratio

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,
$\hat{y}$ – predicted value of y
$\bar{y}$ – mean value of y

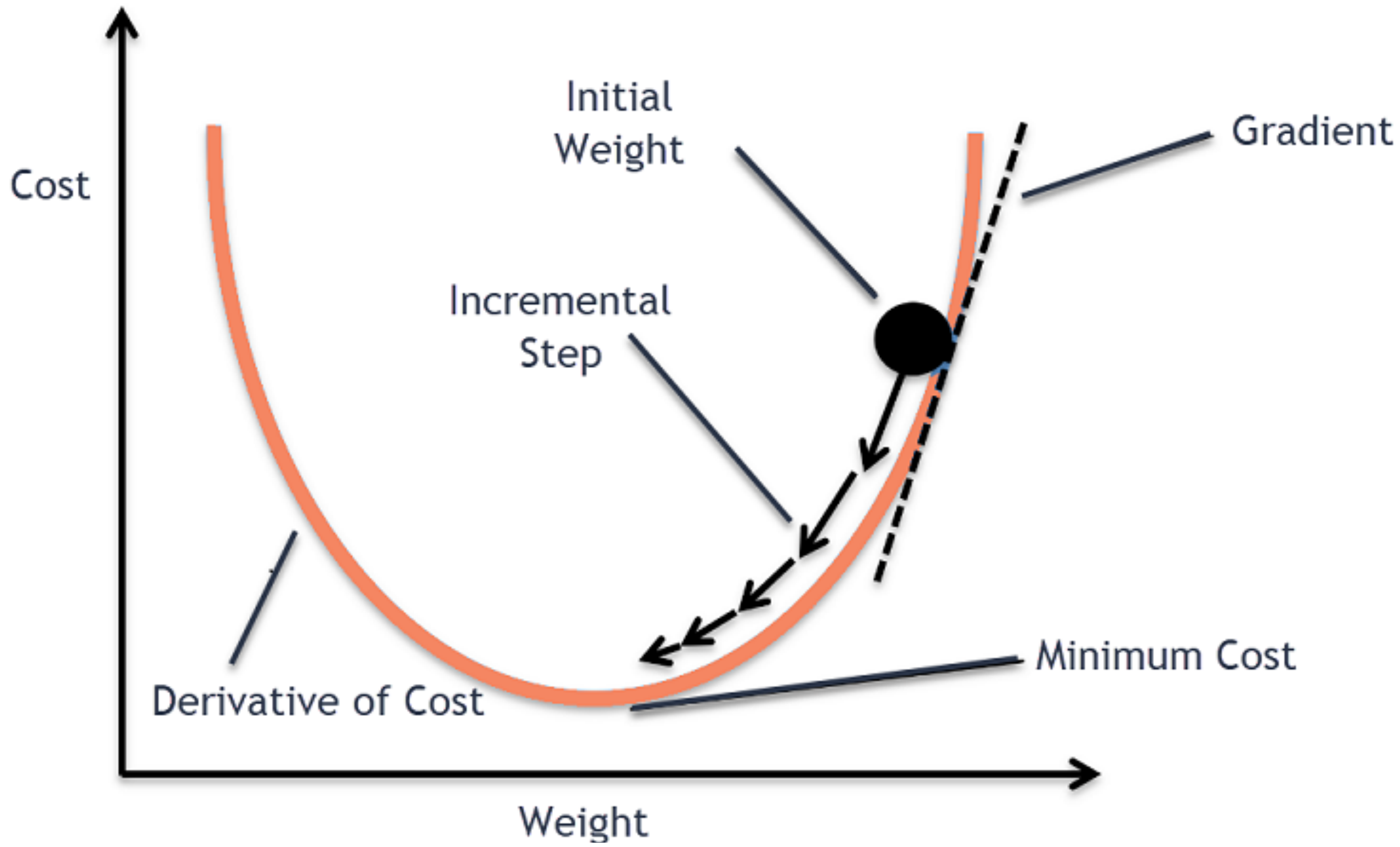
Optimisation

- **Gradient Descent:** An iterative optimisation algorithm used to minimise the cost function.

$$\text{Repeat until convergence } \{$$

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$$\}$$



Various types of Gradient Descent

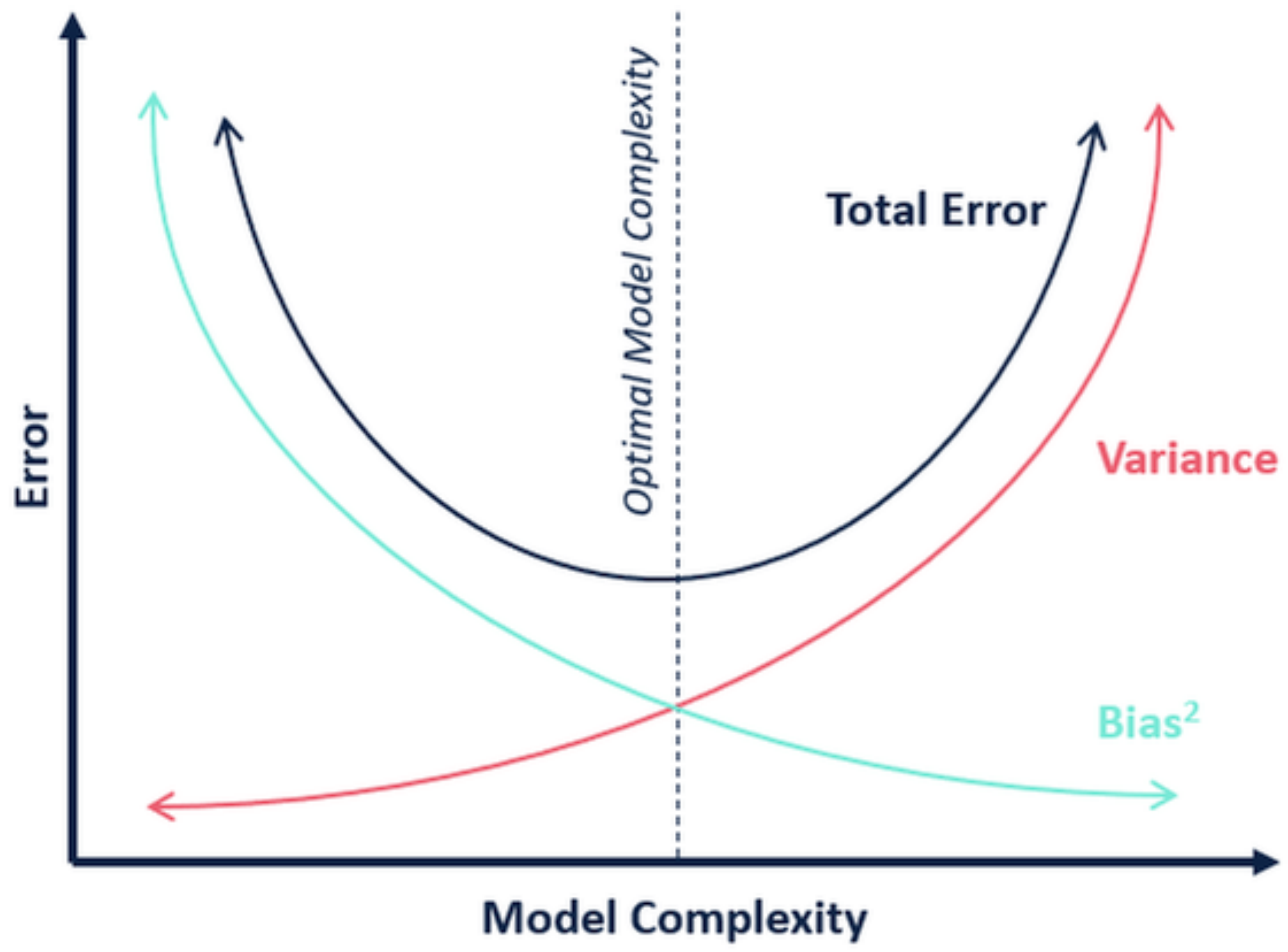
Batch Gradient Descent	Stochastic Gradient Descent (SGD)	Mini-Batch Gradient Descent
• Entire dataset for updation	• Single observation for updation	• Subset of data for updation
• Cost function reduces smoothly	• Lot of variations in cost function	• Smoother cost function as compared to SGD
• Computation cost is very high	• Computation time is more	• Computation time is lesser than SGD
		• Computation cost is lesser than Batch Gradient Descent

Regularization

- Regularisation use for balancing the model from overfit and under fit.

- **L1 Regularization (Lasso):** Adds the absolute value of magnitude of coefficient as penalty term to the loss function.

- **L2 Regularization (Ridge):** Adds the squared magnitude of coefficient as penalty term to the loss function.



Caption

L1 Regularization

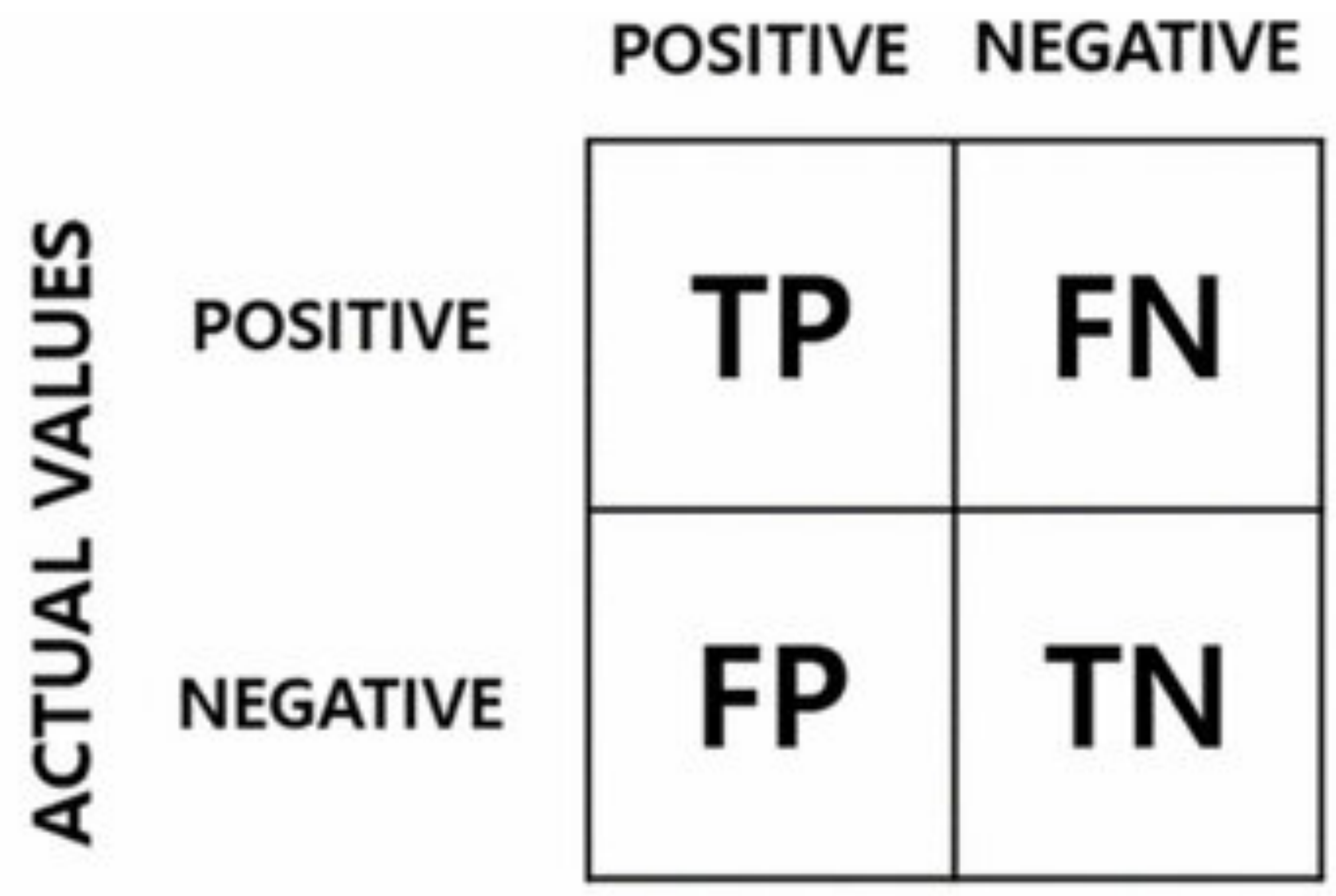
$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \underbrace{\sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2}_{\text{Loss function}} + \lambda \underbrace{\sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

Testing

- R squared: Used for Regression
- **Confusion Matrix:** A table used to evaluate the performance of a classification algorithm.
- **Metrics:**
    - Accuracy
    - Precision
    - Recall
    - F1 Score



$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Model Deployment

- Integrating the model into a production environment. Such as Cloud(AWS, Azure)
- Methods: APIs, embedded systems, cloud services.

Monitoring and Maintenance:

- Continuous monitoring of the model’s performance.
- Handling model drift and updating the model as necessary.

Conclusion

Summary:

- Recap the stages of the machine learning life cycle.
- Emphasise the importance of each stage.
- **Challenges**:Data quality issues, model interpretability, scalability

Final Thoughts:

- Continuous learning and adaptation are key to successful machine learning projects.
- **Best Practices**: Regular updates, thorough validation, comprehensive documentation.

Q&A

- Invitation for Questions:
  - Open the floor for questions and discussions.

Thank You