

# COMP4131: Data Modelling and Analysis

## Lecture 4: Analysis and Modelling

Kian Ming Lim

University of Nottingham Ningbo China

*kian-ming.lim@nottingham.edu.cn*

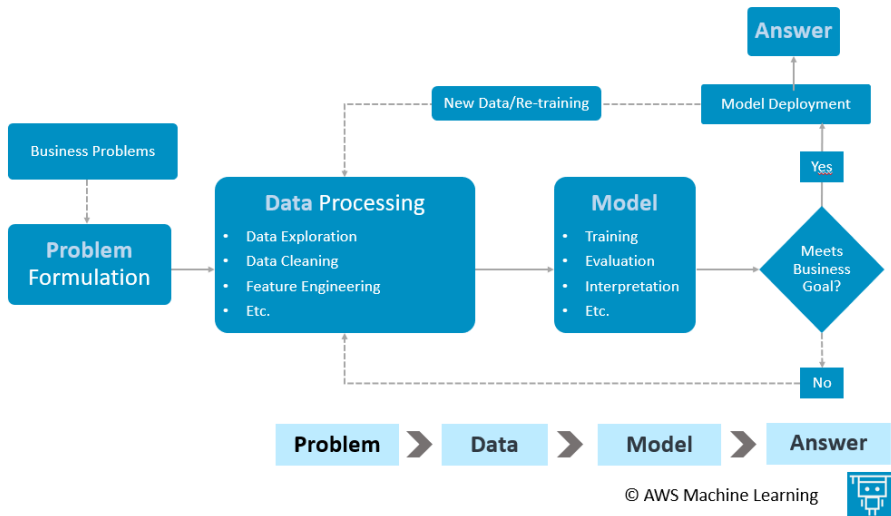
March 6, 2025

# Outline

- 1 Introduction to Analysis and Modelling
- 2 Data
- 3 Data Analysis
- 4 Data Modelling
- 5 Model Evaluation and Validation

# Introduction to Analysis and Modelling

# The Data Pipeline



# Definition of Data Analysis and Modelling

**Data Analysis:** The process of examining, cleaning, transforming, and interpreting data to extract useful insights and support decision-making.

**Data Modelling:** The creation of abstract representations (models) to describe, analyze, and predict real-world phenomena.

## Relationship Between Analysis and Modelling:

- Data analysis provides the foundation for modelling by identifying patterns, trends, and relationships in the data.
- Modelling uses these insights to create predictive or descriptive models.

# Research Questions and Negotiation

- The analysis start with the needs and objectives from stakeholders.
- You need to understand the requirements of the analysis:
  - Who is it for (at what level are you pitching it)? Professors, manager?
  - What resources do you have? Existing dataset/new data?
  - How much time is available?
  - How will the analysis be validated?
  - What level of uncertainty is acceptable?
  - Does it need to be reproducible? (Less obvious than you think!)
- Those things need to be negotiated with the stakeholders
  - Do not assume stakeholders know what they need.

# Characteristics of a Good Analysis Question

## The SPAIN Characteristics:

- **Specific:**
  - "Does eating vegetables improve health?" (Too general)
  - "Does increasing vegetable consumption improve blood sugar in diabetics?" (Specific)
- **Plausible:**
  - "Does eating vegetables increase your bench press?" (No plausible link)
  - "Does your credit limit influence your salary?" (Causality link is reversed)
- **Answerable:**
  - Is it possible to collect the data to answer the question?
  - Does the data even exist?
- **Interesting to a Specific Audience:**
  - Who is your audience? (Researcher, boss, friends)
- **Novel:**
  - Has someone else recently answered it?
  - Nobody likes to do work for no reason.

# The Win-Win Strategy of Formulating Questions

- Do not get trapped into questions which are conditionally interesting
- The goal of a question is to learn something.
- Start with undirected questions, then follow with directional hypotheses.
- Example:
  - **Boring:** "Does algorithm X outperform algorithm Y?"
  - **Interesting:** "What is the interaction of algorithms X and Y and parameters A and B?"
- Sometimes, you may only have conditionally interesting questions, but it's worth checking if that's the case.



# ACTIVITY . THINK-PAIR-SHARE

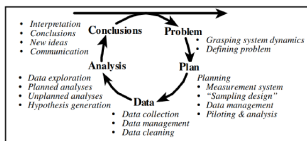
- Pair yourself with the person next to you and make up a good question you could answer from the dataset we collected in class.

ID	<b>CHARACTERISTICS OF A GOOD QUESTION</b>  Specific Plausible Answerable Interesting Novel
Start time	
Completion time	
Email	
Name	
I love	
Age Band	
Do you love Marmite	
Main reason for taking DMA	
I am optimistic about the future of the world	
My favourite movie of all time is	
Today I am feeling	
My favourite way to relax after a stressful day	
On average, how many hours of sleep do you get on a weekday night?	
How many sweets do you think there are in this tin?	

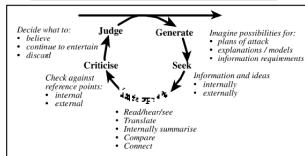
# Many attempts to formalise data analysis

## (a) DIMENSION 1: THE INVESTIGATIVE CYCLE

(PPDAC)



## (c) DIMENSION 3: THE INTERROGATIVE CYCLE



## (b) DIMENSION 2: TYPES OF THINKING

### GENERAL TYPES

- Strategic**
  - planning, anticipating problems
  - awareness of practical constraints
- Seeking Explanations**
- Modelling**
  - construction followed by use
- Applying Techniques**
  - following precedents
  - recognition and use of archetypes
  - use of problem solving tools

### TYPES FUNDAMENTAL TO STATISTICAL THINKING (Foundations)

- Recognition of need for data**
- Transnumeration**

(Changing representations to engender understanding)

  - capturing "measures" from real system
  - changing data representations
  - communicating messages in data
- Consideration of variation**
  - noticing and acknowledging
  - measuring and modelling for the purposes of prediction, explanation, or control
  - explaining and dealing with
  - investigative strategies
- Reasoning with statistical models**
  - aggregate-based reasoning
- Integrating the statistical and contextual**
  - information, knowledge, conceptions

## (d) DIMENSION 4: DISPOSITIONS

- Scepticism**
- Imagination**
- Curiosity and awareness**
  - observant, noticing
- Openness**
  - to ideas that challenge preconceptions
- A propensity to seek deeper meaning**
- Being Logical**
- Engagement**
- Perseverance**

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International statistical review*, 67(3), 223-248.

# The Epicycle of Analysis

## Main Activities of Data Analysis:

- Stating the question
- Exploring and analysing the data
- Building a model
- Interpreting the results
- Communicating the results

## Sub-Activities for Each:

- Setting expectations
- Collecting data/information and comparing it to expectations
- Revising expectations

Source: Peng, R.D., & Matsui, E. (2015). *The Art of Data Science: A Guide for Anyone Who Works with Data*.

# Data

- **Stevens' four level scale:**

- Nominal
- Ordinal
- Interval
- Ratio

- **Mosteller and Tukey's scale:**

- Names
- Grades
- Ranks
- Fractions
- Counts
- Amounts
- Balances

- **Chrisman's expanded scale:**

- Nominal
- Gradation of membership
- Ordinal
- Interval
- Log-interval
- Extensive ratio
- Cyclical ratio
- Derived ratio
- Counts
- Absolute

# Describing the Data

- Nominal
  - Names, labels, categories, etc.
  - Unordered and distinct
  - Relevant statistics: **mode**
  - Bar charts as far as the eye can see
- Ordinal
  - Scales, ratings, etc.
  - Ordered
  - Distinct (no overlap)
  - Relevant statistics: **median, percentile**
- Interval
  - Numeric
  - Ordered, but differences are meaningful
  - Distinct, ordered and **additive**
  - Interval data has **no meaningful 0**
  - Relevant statistics: **mean, standard deviation, correlation, regression**
- Ratio
  - Numeric
  - Ordered and can be used with any arithmetic operator
  - Distinct, ordered, additive, multiplicative
  - Ratio data **has a meaningful, absolute 0**
  - **All statistics apply**

# ACTIVITY . THINK-PAIR-SHARE

- Pair yourself with the person next to you and look at the data we collected in our survey. What type of attributes are those?

ID	154
Start time	2/15/22 15:11:58
Completion time	2/15/22 15:13:18
Email	anonymous
Name	
I love	Both
Age Band	31 to 40
Do you love Marmite	Yes
Main reason for taking DMA	Increase knowledge of subject area
I am optimistic about the future of the world	Agree
My favourite movie of all time is	None of the above
Today I am feeling:	Happy
My favourite way to relax after a stressful day is:	Cooking
On average, how many hours of sleep do you get on a weekday night?	7
How many sweets do you think there are in this tin?	75

Nominal?  
Ordinal?  
Interval?  
Ratio?

## Tricky ones:

- Date / time?
- Age range?
- Amount of sleep?
- Number of sweets in the tin?

- We use numbers and drawings to get a sense of a large dataset.
- The process is called **Transnumeration**.
- **Location:**
  - What's the most common value?
- **Variability/Dispersion:**
  - How spread out is the data? What's a deviant value?
- **Shape:**
  - How is the data distributed?
- **Relationship:**
  - How are two variables linked?



# Transnumerative Techniques

Technique	Description	Example
Sorting	The data are sorted on some criterion. No new variables arise.	The data are sorted by hours of exercise, from lowest to highest.
Grouping	The data are grouped according to some criterion. This creates a new variable. This may involve the change variable type transnumeration beforehand.	A new variable "level of consumption" is created using the fast food data, with values "low" (0-1 fast food meals/week), "medium" (2-3 meals/week), and "high" (4 fast food meals/week).
Subset selection	A subset of the data is selected for further transnumeration.	Data associated with "low" and "high" levels of consumption are considered ("medium" is not).
Change variable type	A numerical variable is thought of in categorical terms or a categorical variable is thought of in numerical or ordinal terms.	Favourite activity (a categorical variable) can be given ordinal status, by ordering activities from most to least active.
Frequency calculation	The frequencies of occurrence of values of a categorical variable are determined. Creates new variable.	The numbers of people in each of the "level of consumption" categories are determined.
Proportion calculation	Proportions (e.g., fractions) are determined in relation to a whole. This creates a new variable	The percentage of people in each of the activity categories is determined.
Graphing	Some or all of the variables in the data (in their current form) are graphed or tabulated.	A scatter graph of hours of exercise v number of fast food meals consumed is constructed.
Summary statistics	A measure of central tendency (e.g., mean), variability, shape, correlation, etc. is determined for a variable. May create new variable.	The average number of fast food meals consumed per week is determined, or parameters of line of best fit between number of hours of exercises and fast food meals consumed.

Partially reproduced and adapted from: Chick, Helen. "Tools for transnumeration: Early stages in the art of data representation."

- **Mode:**

- Most frequent attribute in the sample.
- Not necessarily unique.

- **Median:**

- Middle value in a sorted dataset.
- If  $n$  is odd:  $\text{Median} = x_{\frac{n+1}{2}}$ .
- If  $n$  is even:  $\text{Median} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$ .

- **Arithmetic Mean:**

- $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ .

# Summary Statistics - Variability

## Range:

- Difference between minimum and maximum value.
- Allows you to quickly see the bounds of the data.

## Inter-Quartile Range (IQR):

- The  $p$ -quantile of data is the value relative to which the fraction  $p$  of the data is smaller.
- The  $\frac{1}{2}$ -quantile value is the median.
- The  $p$ -inter-quantile range is the difference between the  $(1 - p)$ -quantile value and the  $p$ -quantile value.
- It is common to use  $p = \frac{1}{4}$  (inter-quartile range).

- Average absolute deviation

$$d_{\bar{x}} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

- Variance

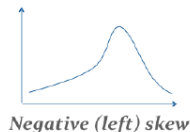
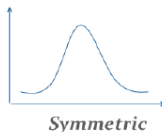
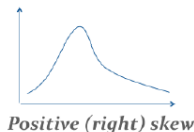
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

## Skewness:

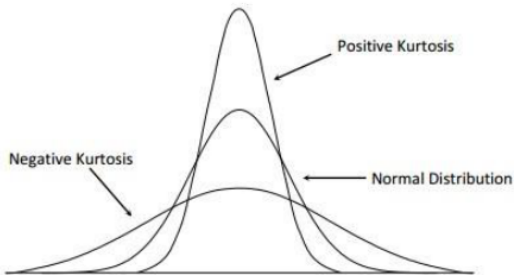
- Measures the asymmetry of the probability distribution of a real-valued random variable about its mean.
- **Positive (Right) Skew:** Tail is longer on the right.
- **Symmetric:** No skew.
- **Negative (Left) Skew:** Tail is longer on the left.



# Summary Statistics - Shape

## Kurtosis:

- Measures how peaked a distribution is.
- **Mesokurtic:** Looks like a Gaussian (normal) distribution.
- **Platykurtic:** Wider and flatter than a Gaussian (negative kurtosis).
- **Leptokurtic:** Thinner and peakier than a Gaussian (positive kurtosis).
- Measure is usually computed relative to the Gaussian distribution.



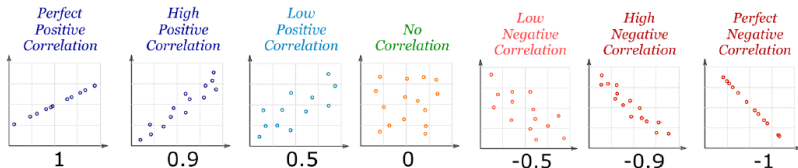
# Summary Statistics - Relationships

## Pearson Correlation Coefficient:

- Measures the strength and direction of the linear relationship between two variables.
- Ranges between  $[-1, +1]$ :
  - 1: Perfect negative linear relationship.
  - 0: No linear relationship.
  - +1: Perfect positive linear relationship.
- Formula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Many other correlation measures exist for different use cases.



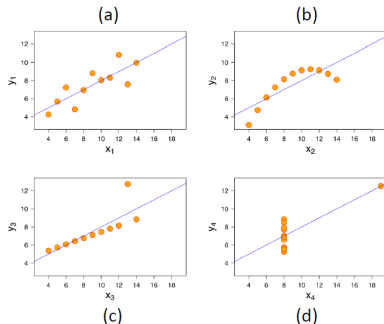
# The Power of a Good Plot

- Don't overestimate numbers.
- Which plot (a, b, c, or d) is represented by this table?

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation(x,y)	0.816
Linear regression line	$y = 3 + 0.5x$

→ Anscombe's quartet

Anscombe, F. J. (1973). Graphs in statistical analysis. *The american statistician*, 27(1), 17-21.



# Diagnosing Data Problems

## Check for Boundary Violations:

- **Impossible Values:**

- If a field contains "Age," negative values are impossible.

- **Outliers:**

- Different from impossible values!
  - Example: An "Age" value of 185.

- Knowing how the data is generated helps in identifying faulty data.

## Check the Packaging:

- Look at the first few rows and the last few rows. Does it all look as expected?
- Check the dimensions of your dataset (number of observations, number of columns). Does it all match up?



# Diagnosing Data Problems - Imputation

## Imputation:

- Replacement of missing data with substituted values.
  - **Unit Imputation:**
    - Impute a data point (e.g., remove a person with missing Age).
  - **Item Imputation:**
    - Impute a value of a data point (e.g., replace missing Age value).
- **Hot Deck Imputation:**
  - Replace value by value of previous observation.
- **Cold Deck Imputation:**
  - Replace value by value randomly sampled from an external dataset.
- **Mean Substitution:**
  - Replace value by central value of all other cases (mean/mode/median).
- **Regression Imputation:**
  - Use a basic machine learning model to predict the missing value.
- **Multiple Imputations:**
  - Sometimes you need more than one method! A lot of design choices go into imputation.

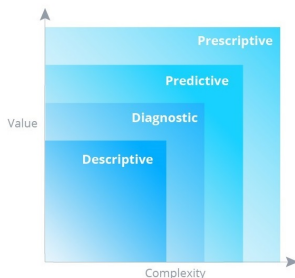
# Data Analysis

# Overview of Data Analysis Techniques

**Purpose:** To transform raw data into actionable insights.

## Types of Data Analysis:

- Descriptive Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis
- Exploratory Data Analysis (EDA)



**Descriptive** analytics addresses the issue of what happened.

**Diagnostic** analytics answers the question of why something happened.

**Predictive** analytics describes what is likely to happen.

**Prescriptive** analytics prescribes what step to take to avoid a future problem.

Source: Things You Should Know About Types Of Data Analysis

# Descriptive Analysis

## Definition:

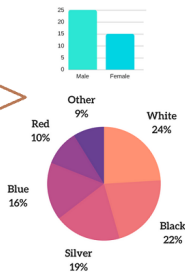
- Summarizes historical data to understand what has happened.

## Examples:

- Mean, median, mode.
- Sales reports, customer demographics.

	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

**RAW DATA**



**Descriptive Statistics**

Source: Descriptive Statistics Examples, Types and Definition

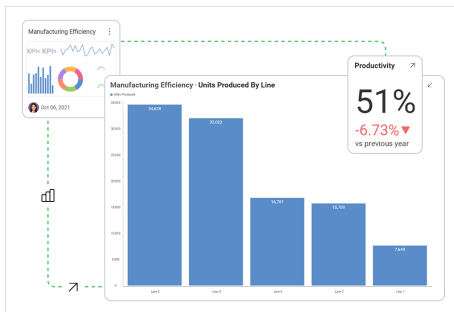
# Diagnostic Analysis

## Definition:

- Identifies the causes of past events or behaviors.

## Examples:

- Root cause analysis.
- Correlation analysis.



Source: Types of Data Analysis, Benefits & Examples

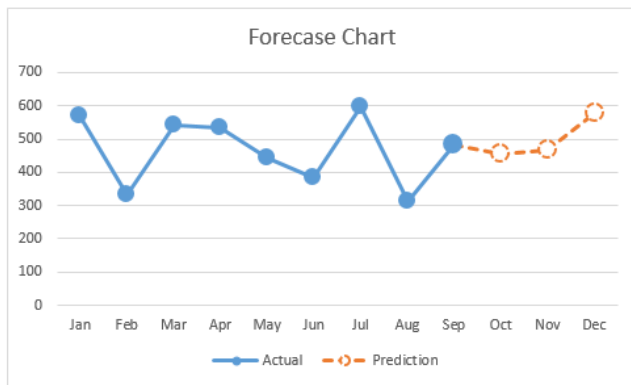
# Predictive Analysis

## Definition:

- Uses historical data to predict future outcomes.

## Examples:

- Sales forecasting, customer churn prediction.



Source: Forecast Chart.

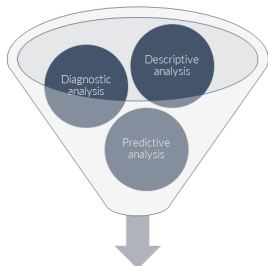
# Prescriptive Analysis

## Definition:

- Recommends actions based on data insights.

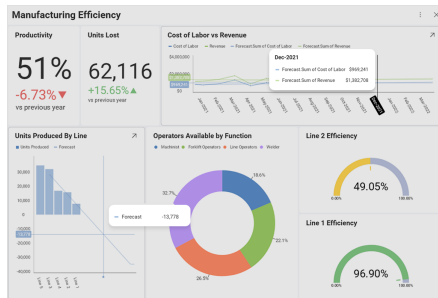
## Examples:

- Optimization models, decision trees.



**Prescriptive analysis** [www.lido.app](http://www.lido.app)

Source: Data Analysis 101



Source: Types of Data Analysis, Benefits  
& Examples

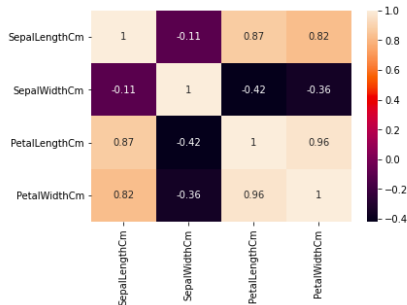
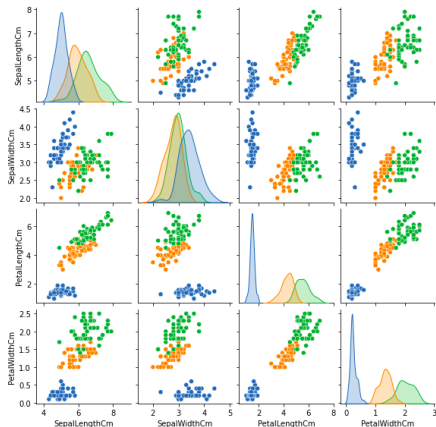
# Exploratory Data Analysis (EDA)

## Definition:

- Investigates data to discover patterns, trends, and anomalies.

## Examples:

- Pairplots, heatmaps, summary statistics.



Source: Exploratory Data Analysis on Iris



# Data Modelling

# What is a Model?

## Definition:

- A simplified representation of a real-world system.

## Purpose:

- To understand, predict, or simulate real-world phenomena.

# Types of Models

**1. Statistical Models:** Uses probability and statistical techniques to make inferences from data.

- Examples: Regression models, time-series models.

**2. Machine Learning Models:** Uses data-driven algorithms to learn patterns and make predictions.

- Examples: Decision Trees, Neural Networks.

**3. Simulation Models:** Mimics real-world processes using computational techniques.

- Examples: Monte Carlo simulations, agent-based models.

**Definition:** A statistical model is a mathematical representation of relationships between variables in data.

- Used for inference, prediction, and hypothesis testing.
- Helps in understanding patterns and making data-driven decisions.
- Examples: Regression models, time-series models, probabilistic models.

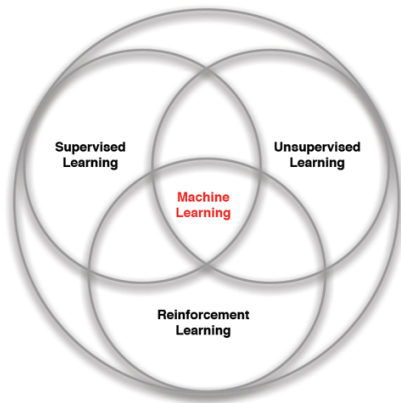
# Introduction to Machine Learning Models

## Definition:

- Models that learn patterns from data.

## Types:

- Supervised learning, unsupervised learning, reinforcement learning.

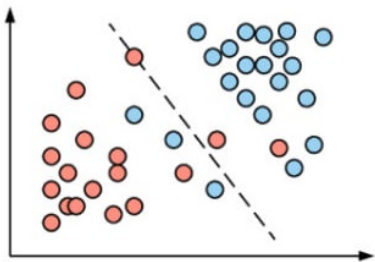


Source: [Link](#)

# Supervised Learning

**Definition:** A machine learning approach where the model is trained using labeled data.

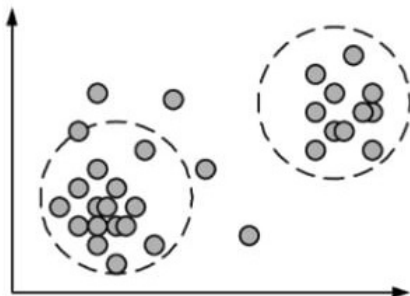
- Input (features) and output (labels) are known.
- Data:  $(x, y)$   $x$  is data,  $y$  is label
- Learn a function to map  $x \rightarrow y$
- Classification, regression, object detection, semantic segmentation, image captioning, etc.
- Examples: Linear regression, logistic regression, decision trees, support vector machines, and many more.



# Unsupervised Learning

**Definition:** A machine learning approach where the model learns patterns without labeled data.

- Identifies hidden structures and relationships in data.
- $\times$  Just data, no labels!
- Examples: K-means clustering, hierarchical clustering, principal component analysis (PCA), and many more.



Source: SuperAnnotate

# Reinforcement Learning

**Definition:** A learning paradigm where an agent learns by interacting with an environment and receiving rewards or penalties.

- Used in robotics, gaming, and autonomous systems.
- Learn how to take actions in order to maximize reward.
- Key components: Agent, Environment, Reward, Policy.



Source: Reinforcement Learning



# Simulation Models

**Definition:** A model that uses computational techniques to simulate real-world processes and assess outcomes.

- Example: Monte Carlo simulation for risk assessment.
- Used in finance, healthcare, and engineering.



Source: An Introduction and Step-by-Step Guide to Monte Carlo Simulations

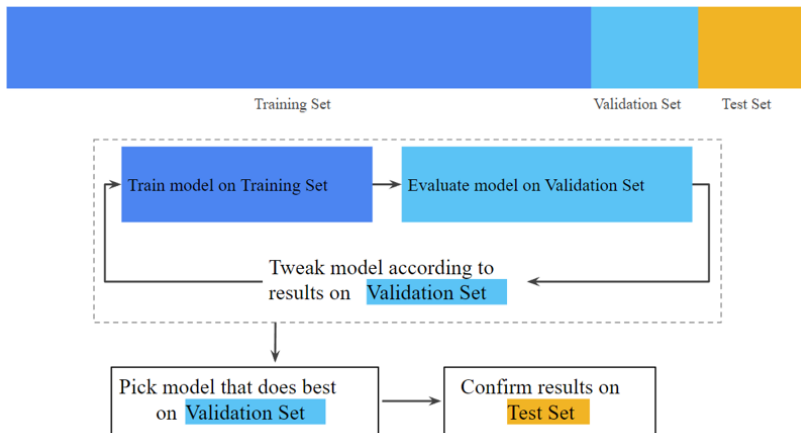
# Model Evaluation and Validation

## Why Model Evaluation Matters:

- Ensures model reliability and robustness.
- Helps detect overfitting and underfitting.
- Assists in selecting the best model for a given task.
- Provides insights into model performance on unseen data.

# Train-Test Split/Holdout Method

**Train-Test Split:** A simple validation technique where data is split into training, validation, and testing sets.

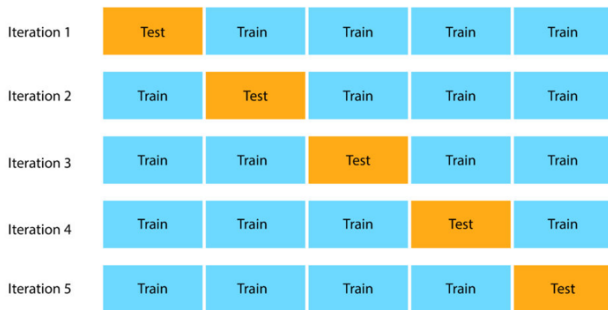


Source: Pi.Exchange

# Cross-Validation

**Cross-Validation:** A more robust method where the dataset is divided into multiple folds.

- $k$ -Fold Cross-Validation: Data is split into  $k$  subsets, and each subset is used for testing once.
- Helps improve model generalization.



Source: Medium

# Model Evaluation Metrics

## Key Metrics for Regression:

- $R^2$  (Coefficient of Determination)
- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)

## Key Metrics for Classification:

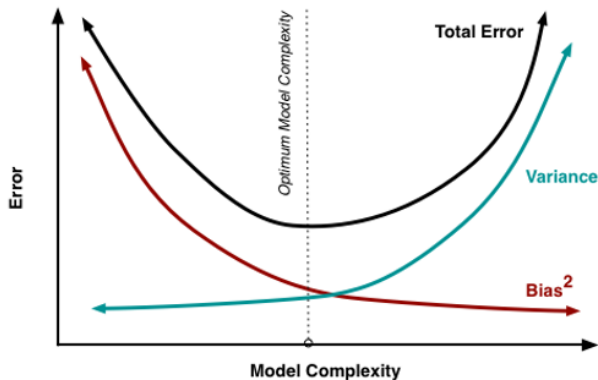
- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- Receiver Operating Characteristic (ROC) Curve
- Area Under the Curve (AUC)

# Bias-Variance Tradeoff

**Bias:** Error due to overly simplistic assumptions.

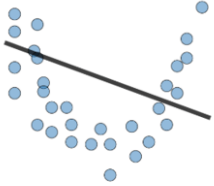
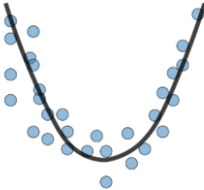
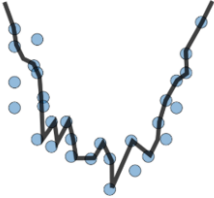
**Variance:** Error due to excessive sensitivity to training data.

- High bias leads to underfitting.
- High variance leads to overfitting.
- The goal is to find an optimal balance between bias and variance.



Source: Dataquest

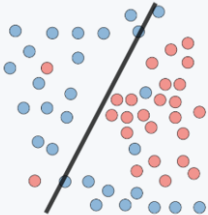
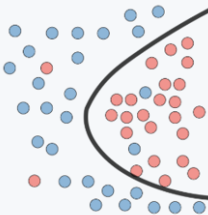
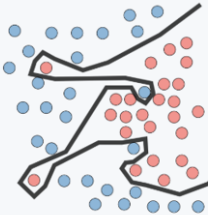
# Underfitting vs Overfitting

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"><li>• High training error</li><li>• Training error close to test error</li><li>• High bias</li></ul>	<ul style="list-style-type: none"><li>• Training error slightly lower than test error</li></ul>	<ul style="list-style-type: none"><li>• Very low training error</li><li>• Training error much lower than test error</li><li>• High variance</li></ul>
Regression illustration			

Source: Machine Learning tips and tricks cheatsheet






# Underfitting vs Overfitting

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"><li>• High training error</li><li>• Training error close to test error</li><li>• High bias</li></ul>	<ul style="list-style-type: none"><li>• Training error slightly lower than test error</li></ul>	<ul style="list-style-type: none"><li>• Very low training error</li><li>• Training error much lower than test error</li><li>• High variance</li></ul>
Classification illustration			

Source: Machine Learning tips and tricks cheatsheet

# Underfitting vs Overfitting

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"><li>• High training error</li><li>• Training error close to test error</li><li>• High bias</li></ul>	<ul style="list-style-type: none"><li>• Training error slightly lower than test error</li></ul>	<ul style="list-style-type: none"><li>• Very low training error</li><li>• Training error much lower than test error</li><li>• High variance</li></ul>
Deep learning illustration			

Source: Machine Learning tips and tricks cheatsheet

# Underfitting vs Overfitting

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"><li>• High training error</li><li>• Training error close to test error</li><li>• High bias</li></ul>	<ul style="list-style-type: none"><li>• Training error slightly lower than test error</li></ul>	<ul style="list-style-type: none"><li>• Very low training error</li><li>• Training error much lower than test error</li><li>• High variance</li></ul>
Possible remedies	<ul style="list-style-type: none"><li>• Complexify model</li><li>• Add more features</li><li>• Train longer</li></ul>		<ul style="list-style-type: none"><li>• Perform regularization</li><li>• Get more data</li></ul>

Source: Machine Learning tips and tricks cheatsheet

# Characteristics of a Good Model

- **Accuracy:** Produces results that match real-world observations.
- **Generalizability:** Works well with unseen data.
- **Interpretability:** Can be easily understood and explained.
- **Efficiency:** Uses computational resources effectively.
- **Scalability:** Performs well with increasing data.

# Model Selection Criteria

- **Performance Metrics:** Accuracy, RMSE, Precision, Recall.
- **Interpretability:** How easily the model's decisions can be explained.
- **Computational Efficiency:** Resource usage and execution time.
- **Generalization Ability:** How well the model performs on unseen data.
- **Data Availability:** The amount and quality of data required for training.