AN INITIATIVE OF:

science, technology & innovation
Department:
Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

CSIR | 80th anniversary
Touching lives through innovation

nicis
NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM
DIRISA

# Introduction to Statistics & Algorithms

By Simanga Mchunu & Ntlharhi Baloyi

# Agenda

- Part 1: Introduction to Statistics
- Part 2: Introduction to Algorithms
- Focus on Decision Trees
- Mathematical Reasoning
- Q&A and Quizzes

01

# Introduction to Statistics

# What is Statistics?

- Science of collecting, analyzing, interpreting, presenting, and organizing data.
- Used in business, healthcare, sports, and machine learning.
- It is **making sense** of numbers.
- Making **informed decisions** in the presence of uncertainty and variation.
- Organising and summarizing data to **draw conclusions** based on the information contained in the data.

# What is Data?

**Data** is **information collected through observation, measurement,or facts**, used for **reference, analysis, or research**.

It represents raw inputs that, when structured and interpreted, form the basis of knowledge and decision-making

# Types of Statistics

- *“The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning". — Nate Silver, The Signal and the Noise*

**Descriptive Statistics:**

Summarize data (mean, median, variance).

**Inferential Statistics:**

Draw conclusions (hypothesis testing, confidence intervals).

# Descriptive Statistics Essentials

Mean: Average = $\mu = (1/n) \sum x_i$

Median: Middle value

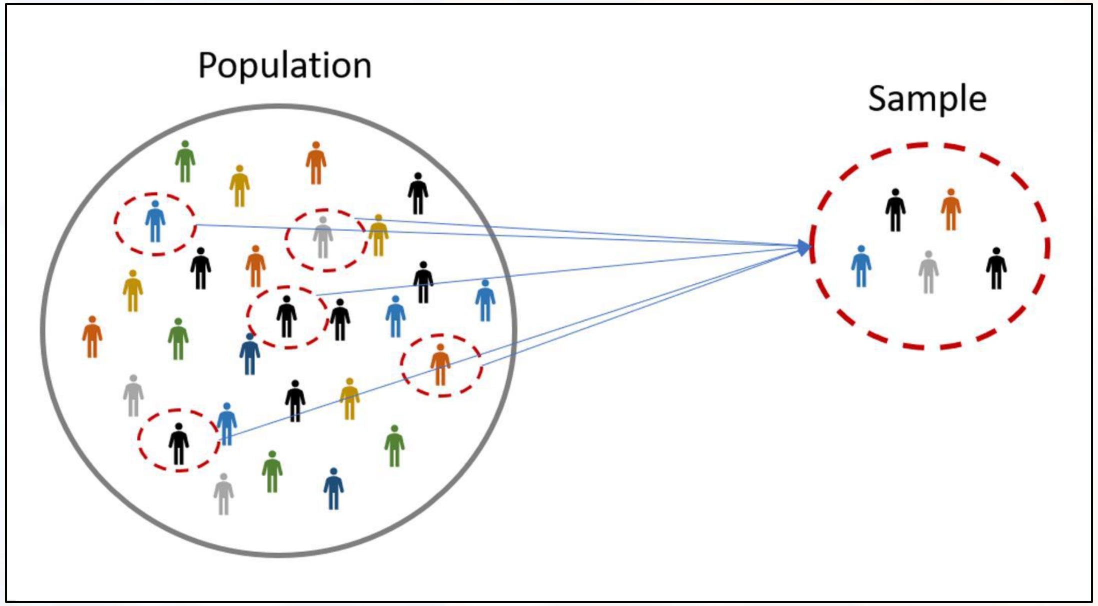
Mode: Most frequent value

Variance: Spread = $\sigma^2 = (1/n) \sum (x_i - \mu)^2$

Standard Deviation = $\sqrt{\text{Variance}}$

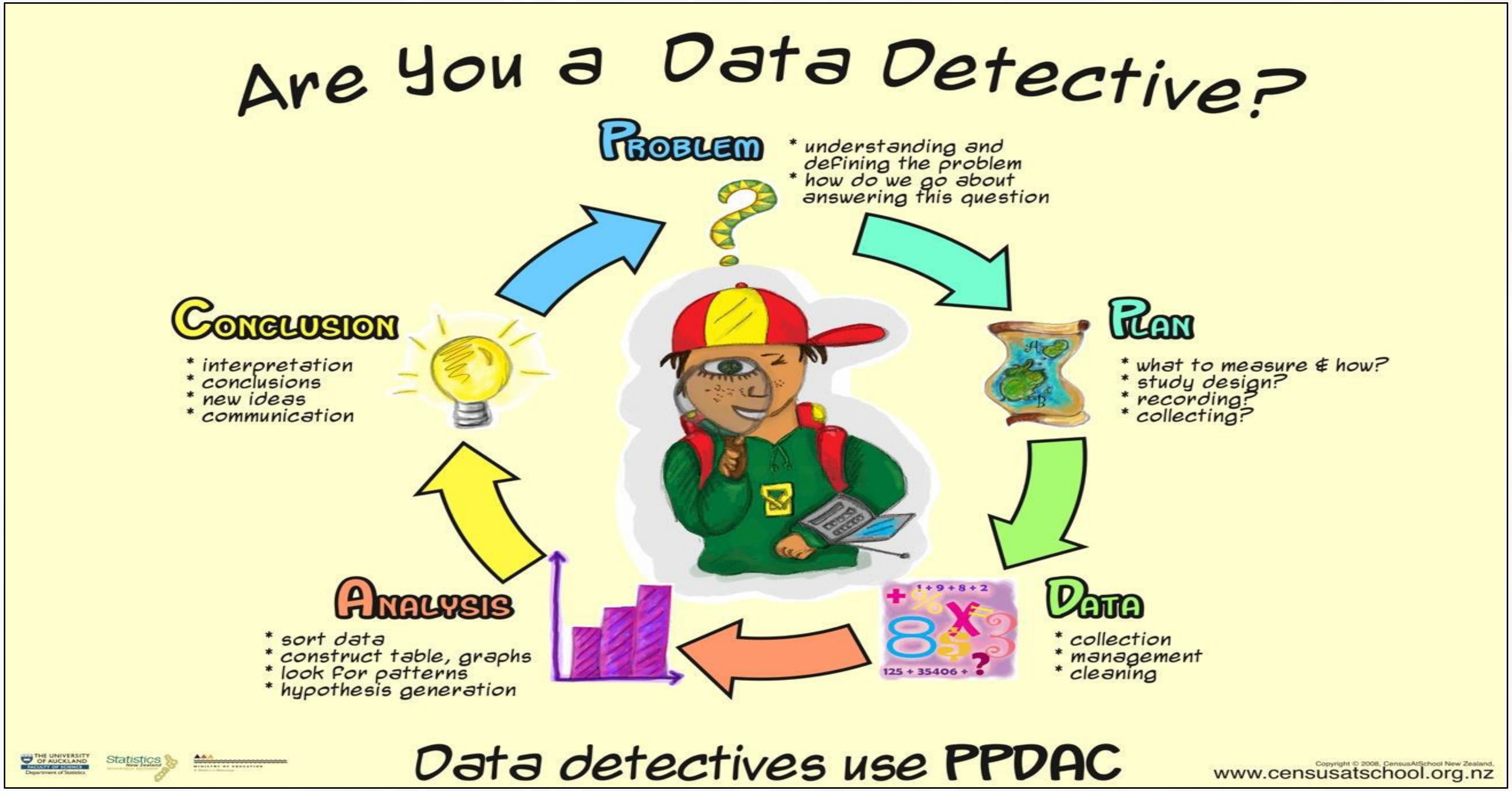
# Inferential Statistics Concepts

- **Sample** vs. **Population**
- Confidence Intervals
- Hypothesis Testing: Null & Alternative Hypotheses, p-values
- **Population:** Collection of objects about which information is sought.
- **Sample:** Part of the population that is observed.



**Source [1]**: https://medium.com/analytics vidhya/population sample parameter statistic biased unbiased ead2021d93d7

# Data Distributions



**Source [2]**: https://www.lse.ac.uk/Events/2019/03/20190327t1830vHKT/Learning-from-Data

# Shipman murdered more than 200 patients, inquiry finds

The British former GP Harold Shipman murdered at least 215 of his patients, the first phase of the public inquiry into the serial killings concluded last week. There is a “real suspicion” that he claimed the lives of another 45 victims, according to the judge leading the inquiry.

High court judge Dame Janet Smith said Shipman may have been hoping to get caught when he altered the will of his last victim, Kathleen Grundy, 81. The “crude forgery” of the will “made detection inevitable,” she said, concluding: “It is hard to resist the inference that Shipman was driven by a need to draw attention to himself and his crimes.”

The inquiry examined a total of 888 cases in its 2000 page report, *Death Disguised*. Shipman was found not responsible for 604 deaths, and no conclusion was reached in a further 38 cases. The 215 people that the inquiry determined were killed by Shipman comprised 171 women and 44 men.

**Source [3]**: https://pmc.ncbi.nlm.nih.gov/articles/PMC1123718/

# Looking at Data

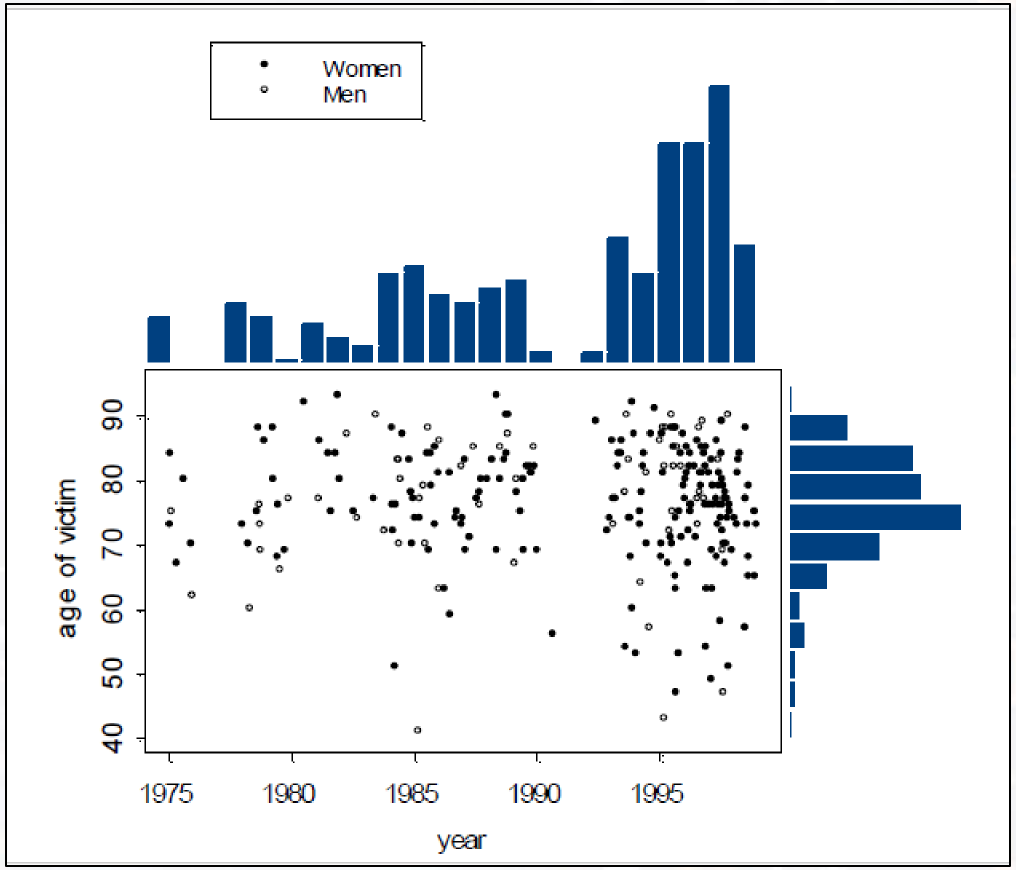
- What was the pattern of Harold Shipman’s murders?
- **Problem**: can more detail tell us more about what Shipman did?
- **Plan**: compare actual times at which his patients died with the times of deaths recorded by other local GPs
- **Data**: a huge exercise requiring examination of death certificates
- **Analysis**: simple plotting…..

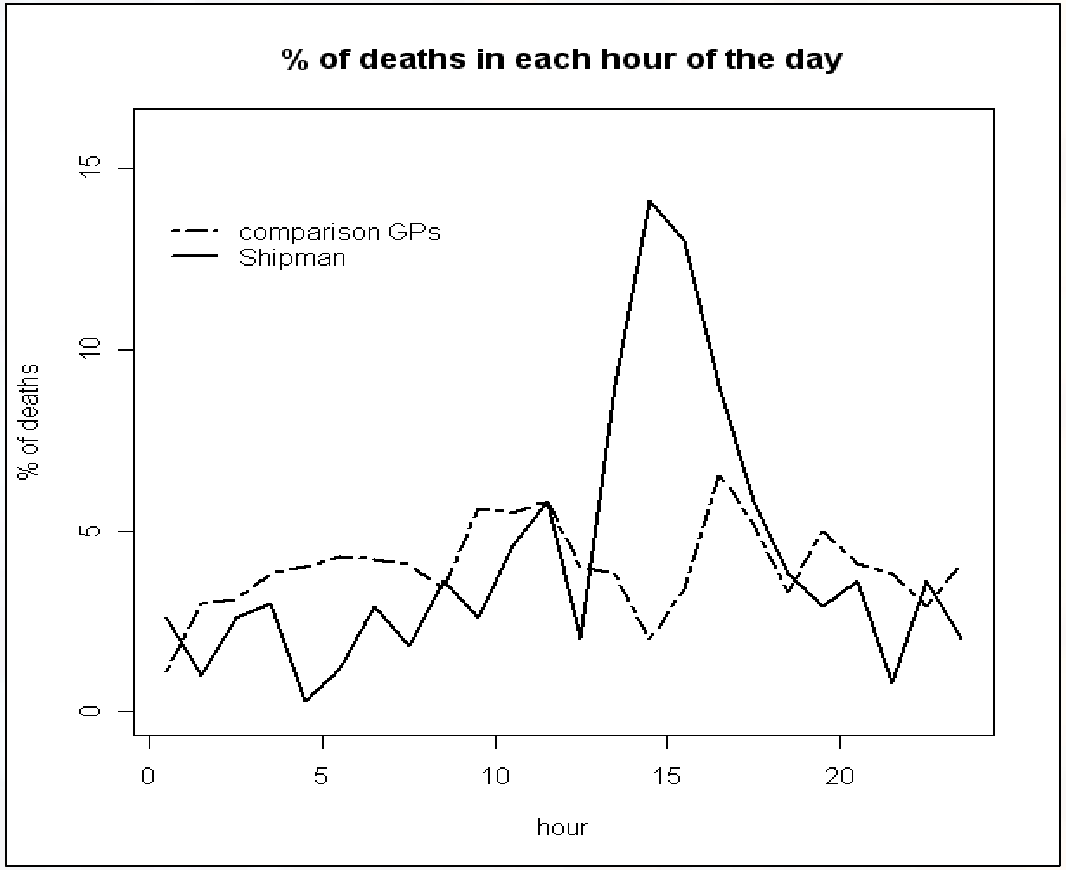
*‘I have nothing to hide’*

Dr Harold Shipman, general practitioner, on his arrest in September 1998

Source [4]: The art of Statistics by David Spiegelhalter’s book

# Looking at Data





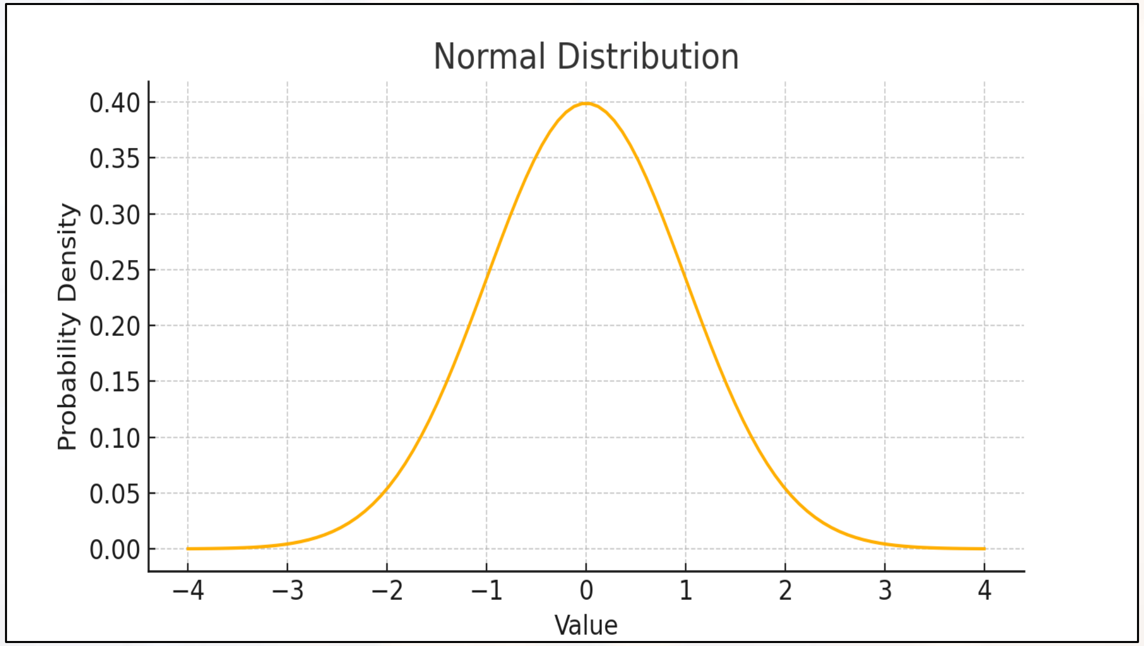
**Source: [4]** https://www.lse.ac.uk/Events/2019/03/20190327t1830vHKT/Learning-from-Data

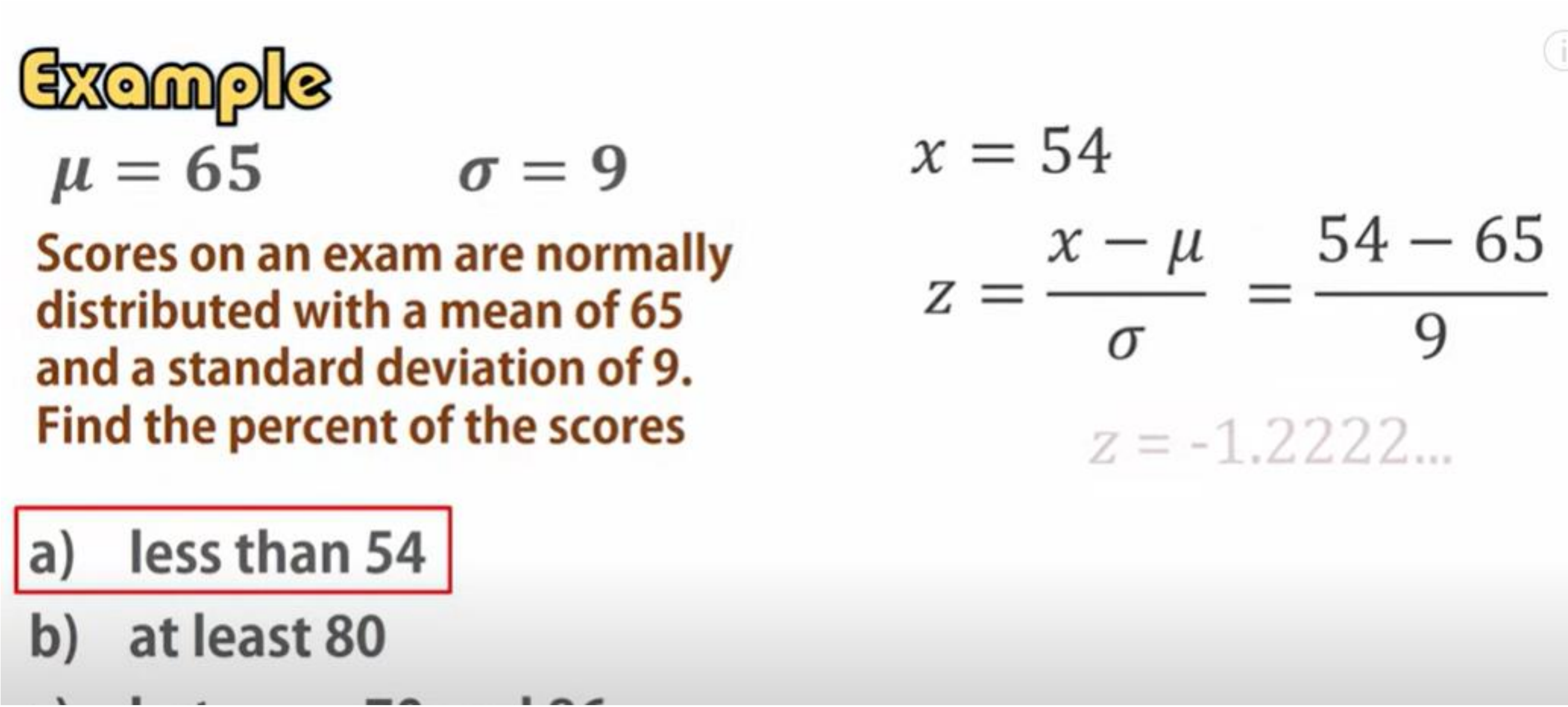
# What is a Data Distribution?

- **Data distribution** refers to how the values in a dataset are spread out or arranged. It tells you:
  - What values are most common
  - How much variation there is
  - Where the data is centered (e.g. average)
  - Whether it's symmetrical, skewed, or has outliers
- There are 2x main types of Data Distribution
  - **Normal Distribution**
  - **Skewed Distribution**

# What is a Normal Distribution?

- Normal Distribution (Known as the **Gaussian Distribution**) is a probability distribution that appears as a "bell curve" when graphed. Also called the **Bell Curve** because of its shape.
- **Characteristics**:
  - **Symmetrical** around the mean
  - Mean = Median = Mode
  - Most data falls near the center
  - Tails taper off equally on both sides
  - Defined mathematically by the **Gaussian function**





## Example

$\mu = 65$ $\sigma = 9$

Scores on an exam are normally distributed with a mean of 65 and a standard deviation of 9. Find the percent of the scores

a) less than 54

b) at least 80

$$x = 54$$

$$z = \frac{x - \mu}{\sigma} = \frac{54 - 65}{9}$$

z = -1.2222...

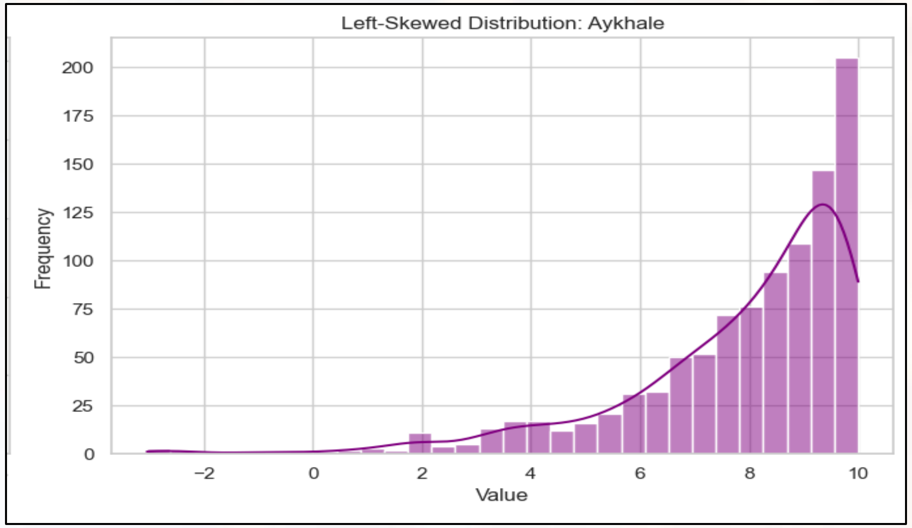
# What is Skewed Distribution?

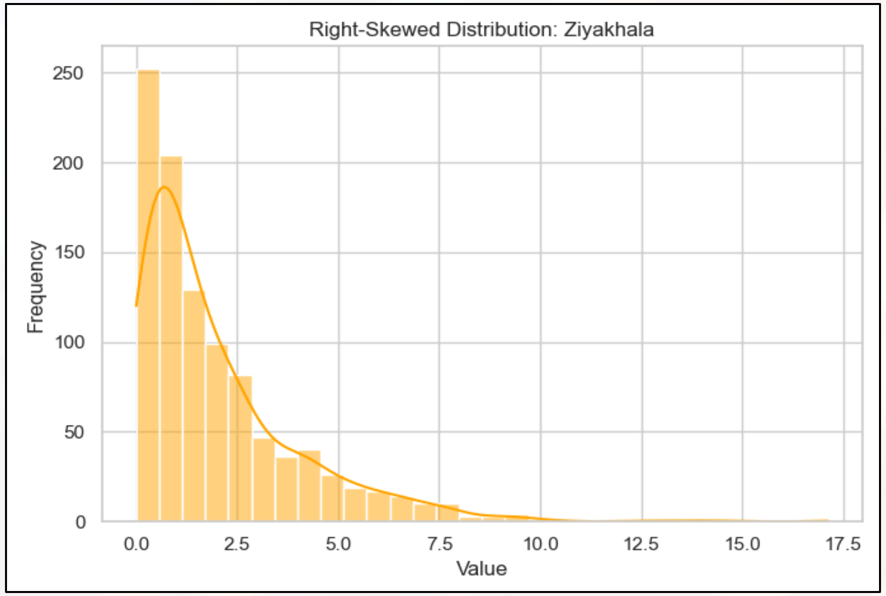
A **skewed distribution** occurs when the **data is not evenly distributed** around the mean. The values tend to **lean more to one side**, forming a **tail** on either the right or left.

**Characteristics:**

- **Asymmetrical** shape
- The **mean, median, and mode are not equal**
- **Two types** of skewness:
  - **Right-Skewed (Positive Skew)**: Tail on the **right**, Mean > Median > Mode
  - **Left-Skewed (Negative Skew)**:
  - Tail on the **left**, Mean < Median < Mode

Affects **statistical analysis and interpretation**

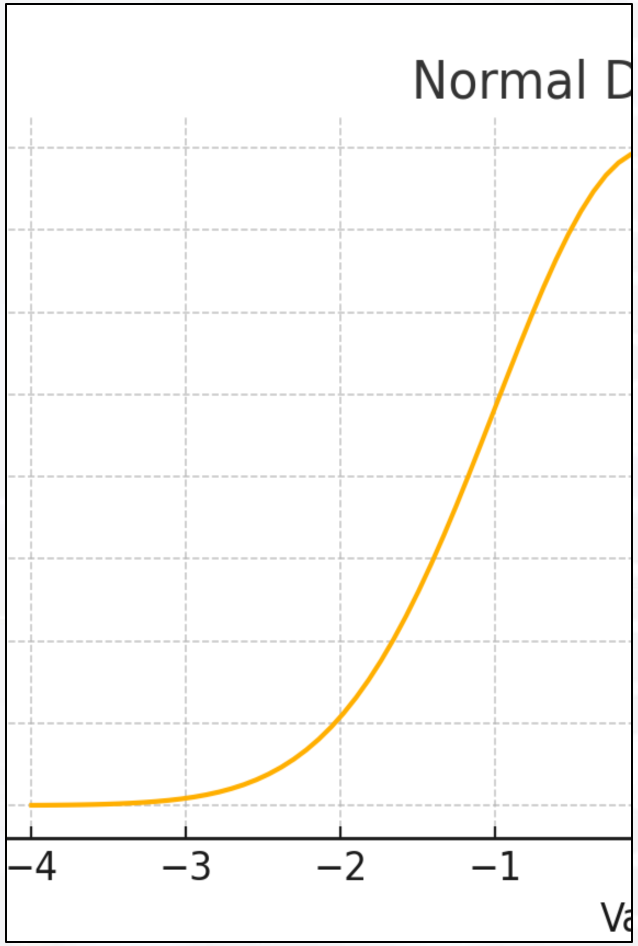




# Which Central Tendency Measure to Choose?

| Measure | Best When | Avoid When |
|---|---|---|
| **Mean** | Data is evenly distributed (normal distribution).<br>No outliers | Sensitive to outliers |
| **Median** | Data has outliers or skewed distribution | Less informative for symmetric, well-behaved data |
| **Mode** | Data is categorical or you need the most frequent value | Not very useful for continuous data or when no clear mode exists |

# What Distribution means in Data Science?



Understanding if your data is **normally distributed** matters because:

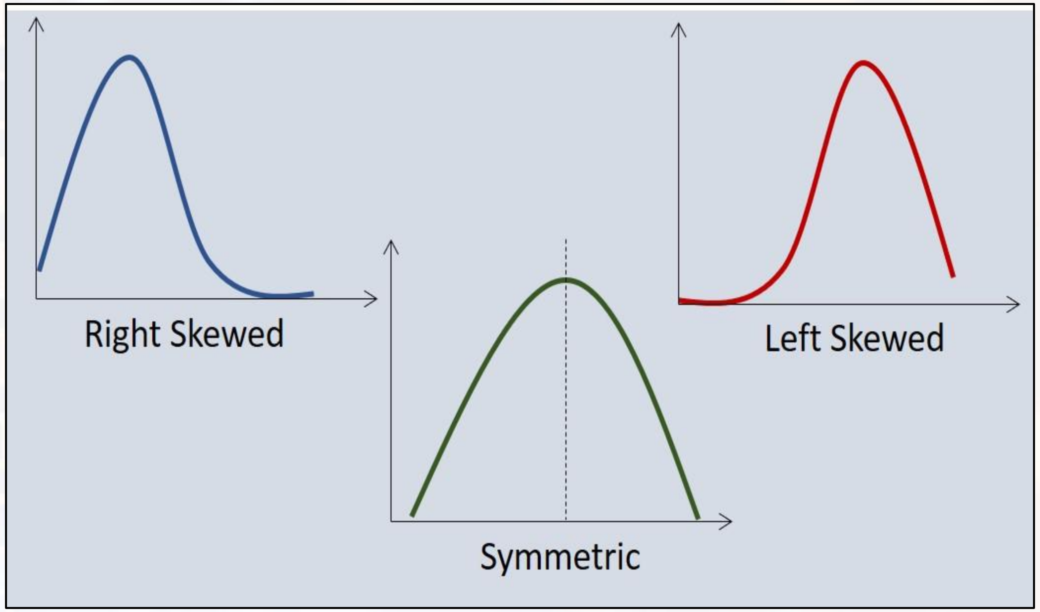
**Many algorithms assume it:**

- Linear regression
- Logistic regression
- Naive Bayes
- T-tests and ANOVA

**Standardized models:**

- Let you apply **z-scores** (how many std deviations from mean)
- Make it easier to detect **outliers**

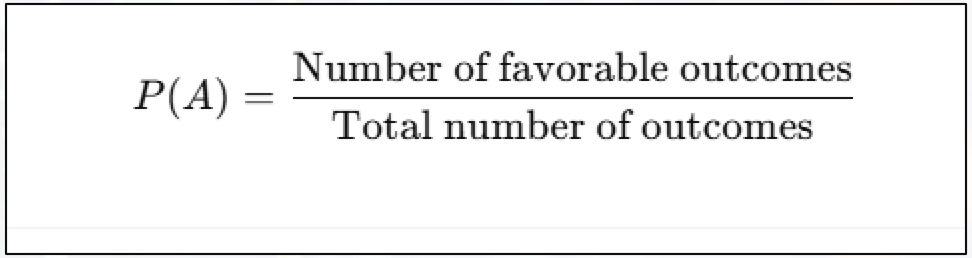
# What Distribution means in Data Science?



- **Helps with:**
  - Feature scaling
  - Probability estimates
  - Statistical inference
- **Real-World Example:**
  - If you're analyzing **student test scores**, a normal distribution means:
    - Most students score near the average
    - Fewer students score very high or very low
    - The grading curve might be centered on the class average

# Probability – The Language of Uncertainty

- **Formula for an event A:**



$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

- **Why it matters in Data Science:**
  - Handles **uncertainty** in predictions (e.g., spam detection, fraud detection)
  - Used in probabilistic models like **Naive Bayes**, **Bayesian networks**, **Hidden Markov Models**
  - Foundation of **sampling**, **A/B testing**, and **predictive modeling**

# Probability Example

- If 3 out of 5 emails are spam:

$$P(\text{Spam}) = \frac{3}{5} = 0.6$$

# Conditional Probability – Knowing Changes the Odds

- **Formula:**

$$P(A|B) = P(A \cap B) / P(B)$$

- **Data Science Example:**
  - What's the chance a user **buys** given they **clicked**?
  - Let A = Buy, B = Click
  - Crucial for **recommendation systems** and **conversion analysis**

# Bayes' Theorem – Update Your Beliefs

**Formula:**
P(A|B) = P(B|A) * P(A) / P(B)

**Logic:**
Combine prior knowledge P(A) and observed data P(B|A) to **update belief**

Core to **Bayesian reasoning**, **probabilistic classifiers**, and **medical diagnostics**

# Data Distributions

- Data Science Use Case – Naive Bayes Classifier:
  - Let’s classify if a message is **spam** or **not spam**:
    - A: Message is spam
    - B: Message contains the word "free“

$$P(Spam|"free") = \frac{P("free"|Spam) \cdot P(Spam)}{P("free")}$$

- Trained on labelled data to estimate all components.

# Data Distributions

**Algorithmic Flow:**

1.Collect labeled data: text + labels (spam/not spam)

2. Tokenize messages (features: words)

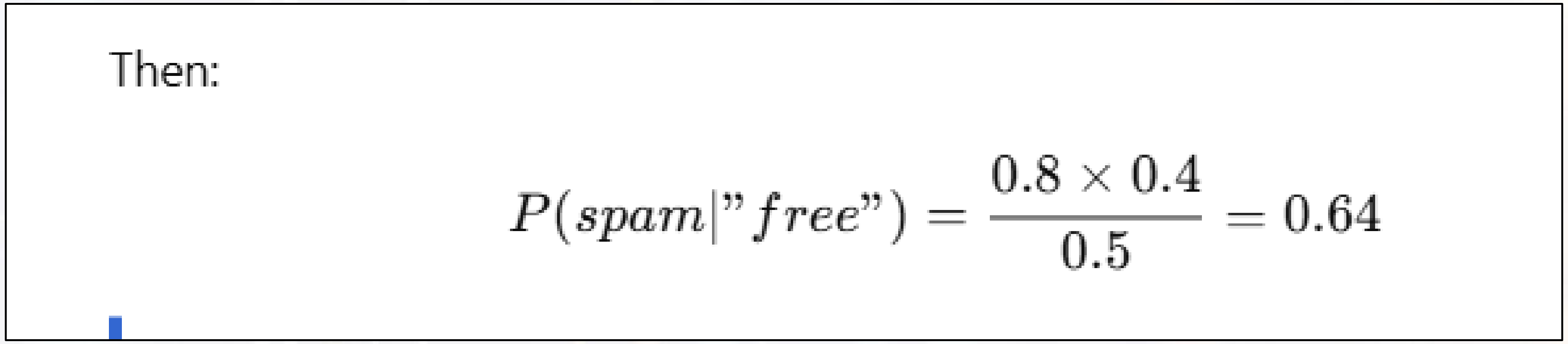
3. Compute probabilities: Prior, Likelihood

4. Apply Bayes' Theorem

5. Classify message by highest posterior probability

# Output Example

- - P(spam) = 0.4,
- - P("free"|spam) = 0.8,
- - P("free") = 0.5

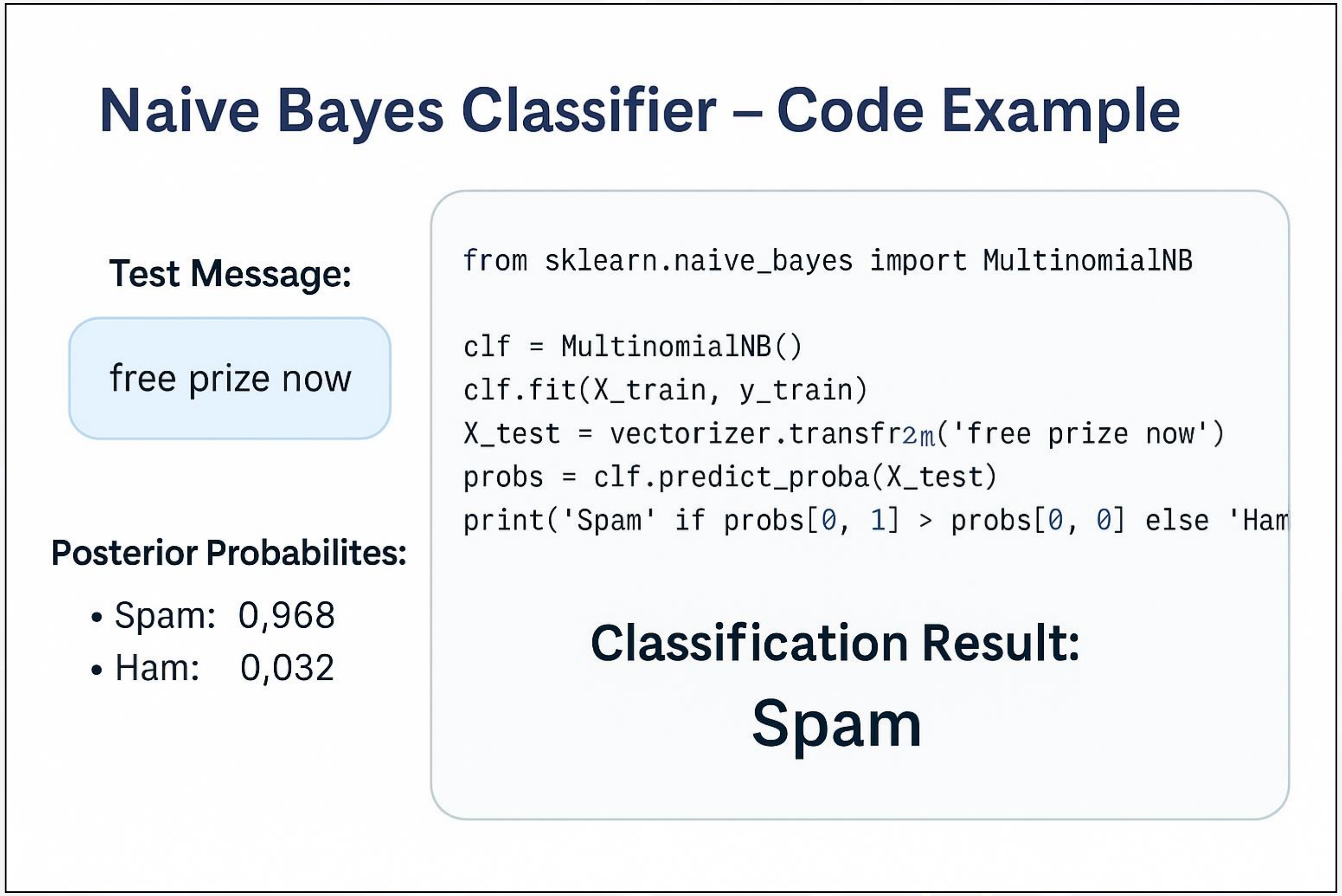


Then:

$$P(spam|\text{”}free\text{”}) = \frac{0.8 \times 0.4}{0.5} = 0.64$$

- → High probability → Label as spam

# Output Example



## Naive Bayes Classifier – Code Example

**Test Message:**

free prize now

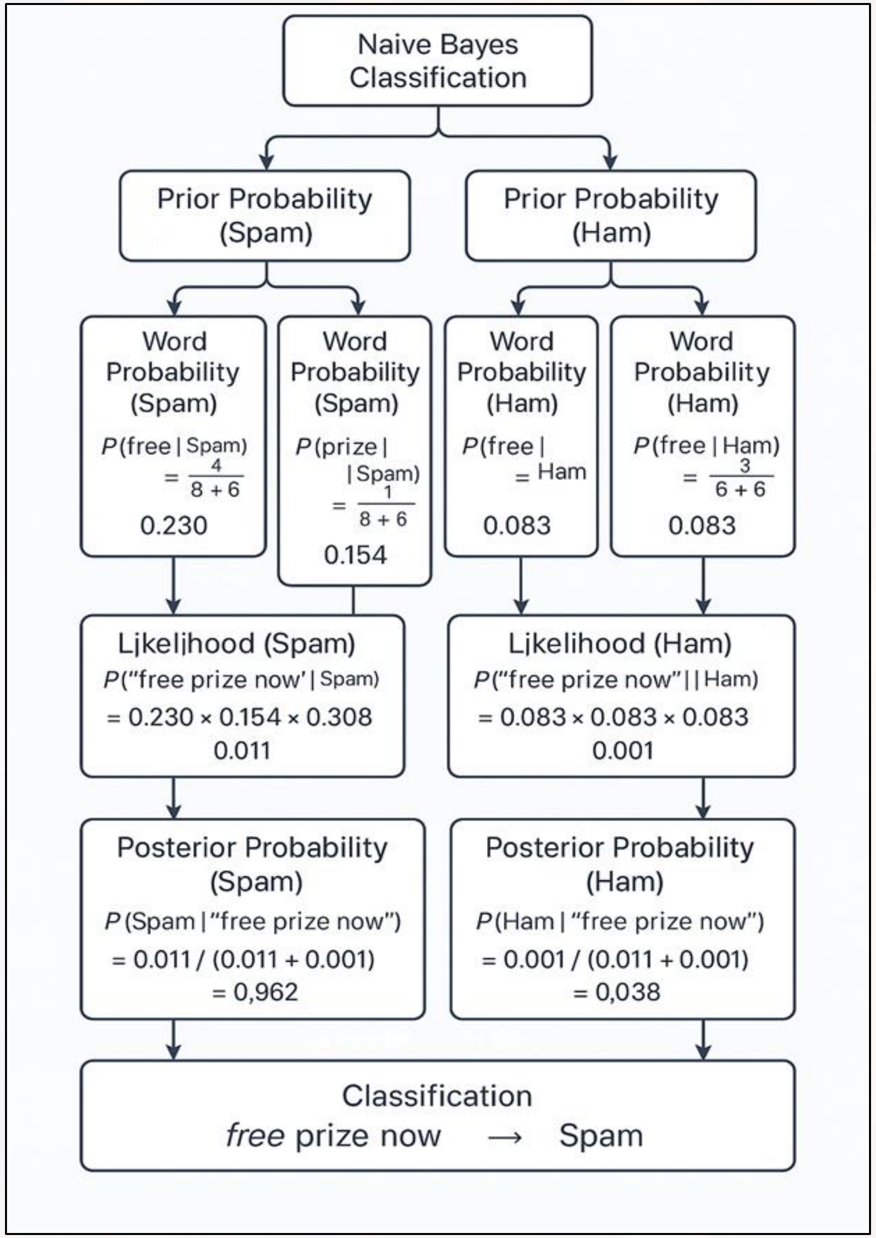
```
from sklearn.naive_bayes import MultinomialNB

clf = MultinomialNB()
clf.fit(X_train, y_train)
X_test = vectorizer.transfr2m('free prize now')
probs = clf.predict_proba(X_test)
print('Spam' if probs[0, 1] > probs[0, 0] else 'Ham
```

**Posterior Probabilites:**

- Spam: 0,968
- Ham: 0,032

**Classification Result:**

**Spam**



Naive Bayes Classification

Prior Probability (Spam)

Prior Probability (Ham)

Word Probability (Spam)
$P$(free | Spam) $= \frac{4}{8+6}$
0.230

Word Probability (Spam)
$P$(prize | | Spam) $= \frac{1}{8+6}$
0.154

Word Probability (Ham)
$P$(free | = Ham
0.083

Word Probability (Ham)
$P$(free | Ham) $= \frac{3}{6+6}$
0.083

Ljkeljhood (Spam)
$P$(“free prize now’ | Spam)
= 0.230 × 0.154 × 0.308
0.011

Ljkelihood (Ham)
$P$(“free prize now” | | Ham)
= 0.083 × 0.083 × 0.083
0.001

Posterior Probability (Spam)
$P$(Spam | “free prize now”)
= 0.011 / (0.011 + 0.001)
= 0,962

Posterior Probability (Ham)
$P$(Ham | “free prize now”)
= 0.001 / (0.011 + 0.001)
= 0,038

Classification
*free* prize now ⟶ Spam

**Source: [5]** by Author

# Key Takeaways

- **Probability and statistics** form the **foundation** for organizing, analyzing, and interpreting data. They enable us to:
- **Understand patterns** in historical data
- **Predict future outcomes**
- **Make informed decisions** in uncertain conditions

# QUIZ TIME!!!

- **Where the mean, mode, and median are equal?  Which graph distribution is:**

A. Negative Skewed

B. Normal Distribution

C. Data Distribution

D. None of the Above

Answer: **B. Normal Distribution**

- **What is Statistics?**

A. Information collected through observation, measurement

B. The state of the country.

C. Science of collecting, analyzing, interpreting, presenting, and organizing data

D. All of the above

Answer: **C. Science of collecting, analyzing, interpreting, presenting, and organizing data**

# QUIZ TIME!!!

- **What does a skewed distribution indicate?**

  A. Data has multiple peaks

  B. Data tails off more on one side
  C. Data is symmetrical
  D. Data is uniformly distributed

- **Ans: Correct Answer:** B
- **What does standard deviation measure in a dataset?**

  A. The average value of the data
  B. The most frequent value
  C. The median of the data

  D. The spread or variability of the data around the mean

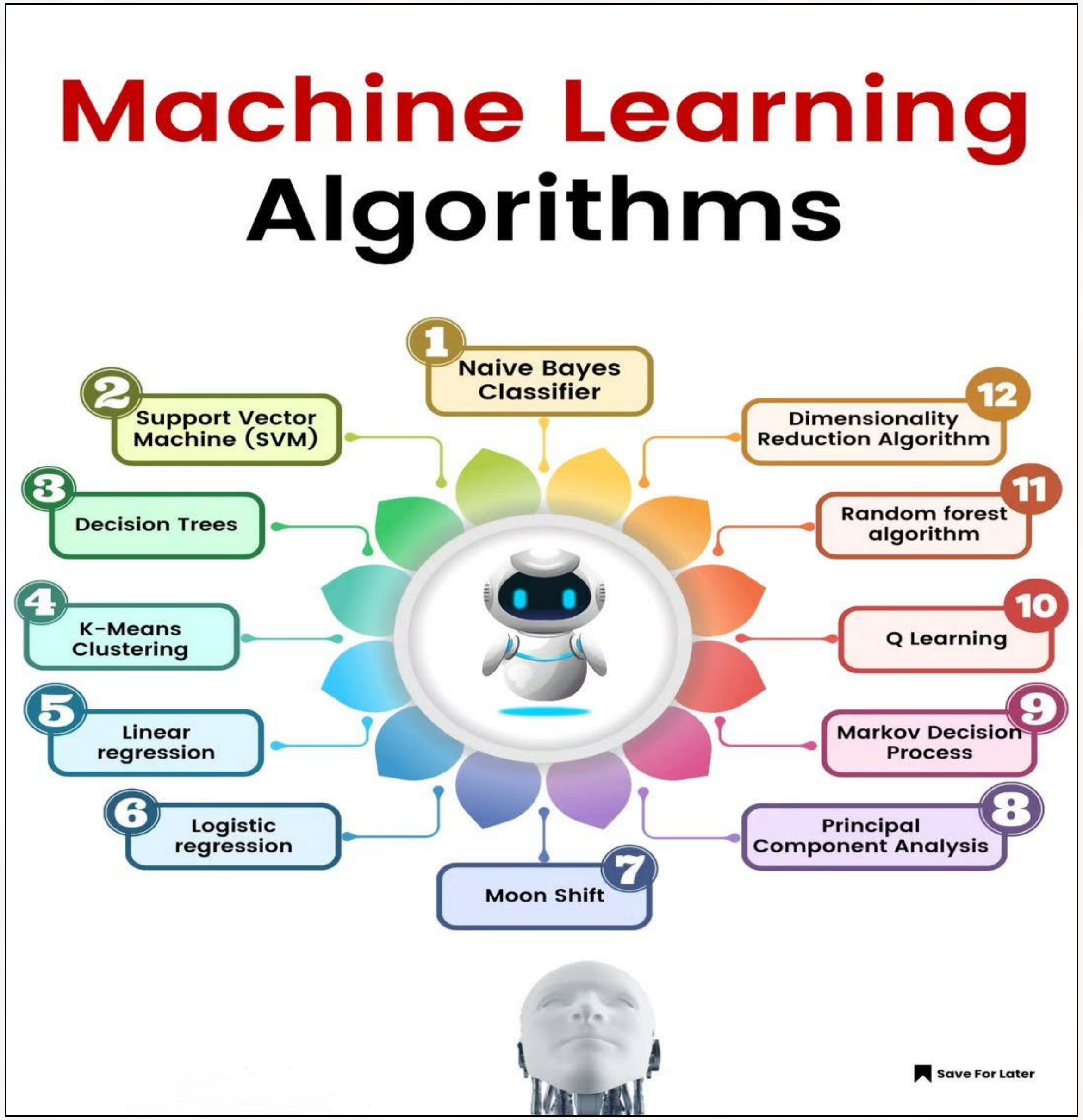
- Ans: **Correct Answer:** D

# 02

# Introduction to Algorithms

# What is an Algorithm?

- Step-by-Step procedure to solve a problem OR
- A procedure to accomplish a specific task
- In ML, used to find patterns and make predictions from data.
- e.g: What is the size of the set S = {x ∈ Z : x^2 < 23}?



Source: [6] TechGeekBuzz

# Types of Algorithms in ML

- Supervised Learning: Decision Trees, Linear Regression, SVM
- Unsupervised Learning: Clustering, PCA
- Reinforcement Learning: Q-Learning

# Introduction to Decision Trees

Decision trees are intuitive, flowchart-like structures used for decision-making and predictive modeling. They start at a root node and branch out based on feature-based questions, leading to leaf nodes that represent outcomes or predictions

Tree-like structure for classification/regression.

Each node: a feature; leaves: outcomes.

# Why Use Decision Trees?

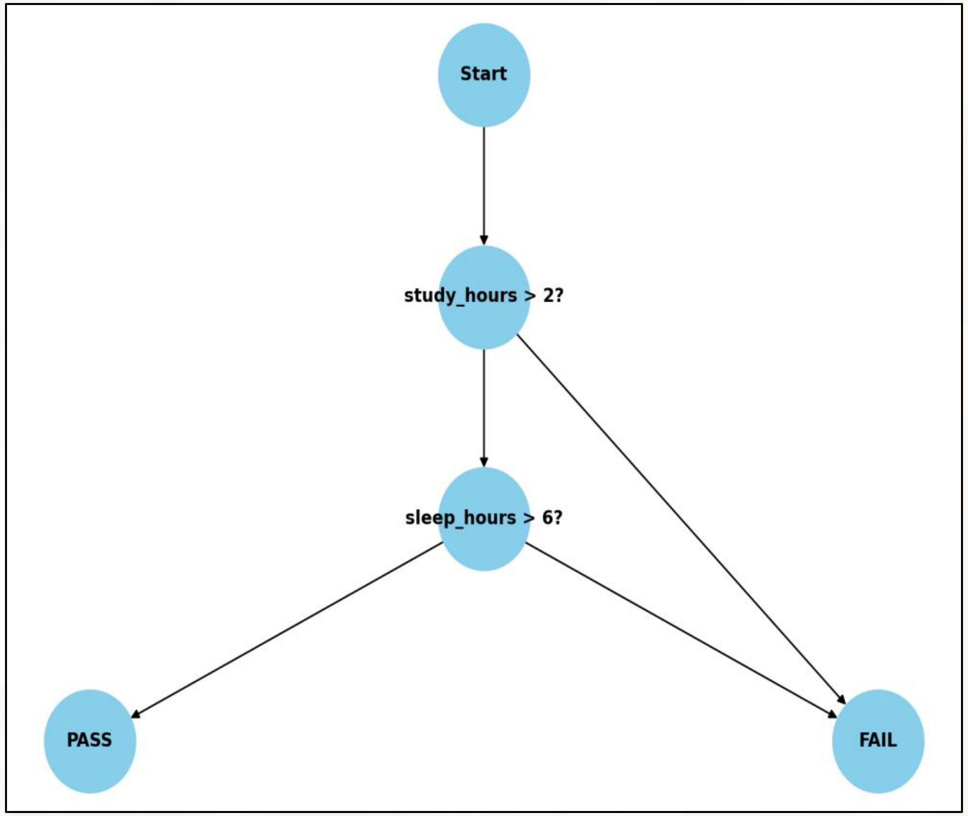
**Interpretability**: They provide clear, step-by-step reasoning behind decisions, making them easy to understand.

**Versatility**: Capable of handling both numerical and categorical data.

**Foundation for Advanced Models**: Serve as the basis for ensemble methods like Random Forests and Gradient Boosted Trees.

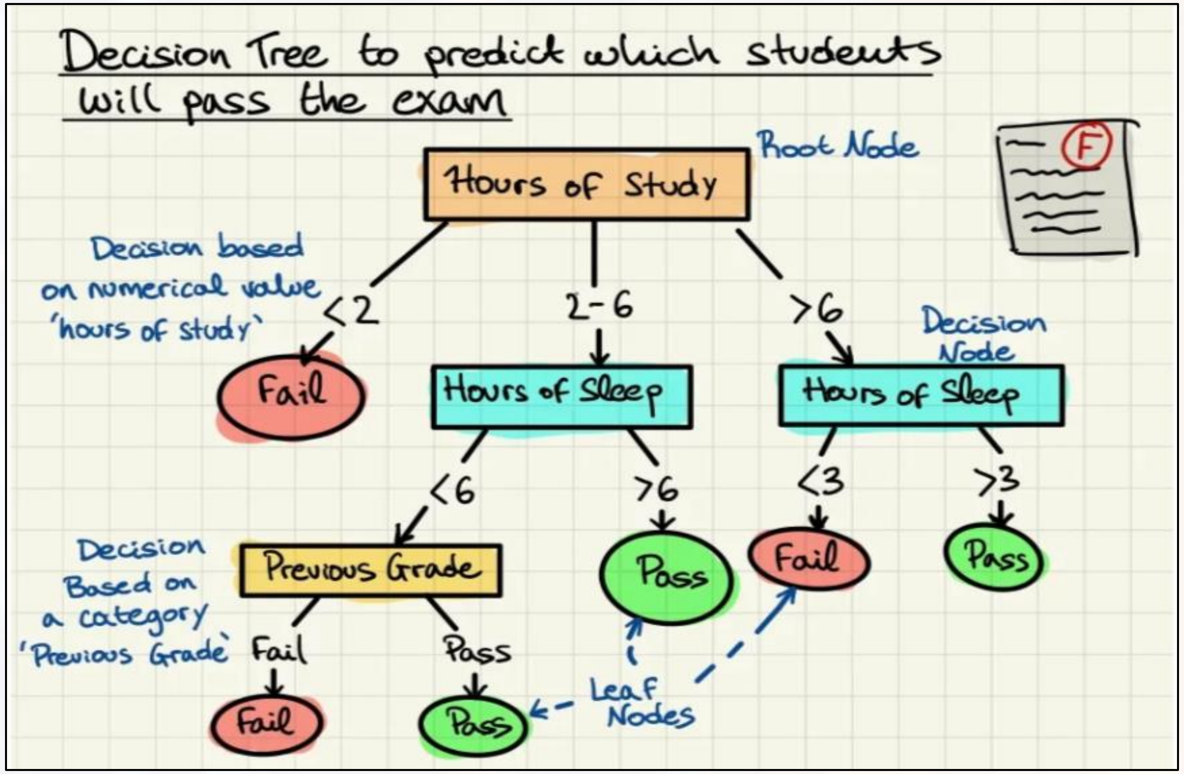
# Example: Predicting Exam Outcomes

- Consider a scenario where we predict whether students will pass an exam based on:
  - Hours studied
  - Hours slept before the exam
  - Previous grades
- The decision tree evaluates these features at each node:
  - **Root Node**: "Did the student study more than 2 hours?”
  - **Branching**:
  1. If **Yes**: "Did the student sleep more than 6 hours?"
     1. If **Yes**: Predict **Pass**
     2. If **No**: Predict **Fail**
  2. If **No**: Predict **Fail**
- This structure mirrors human decision-making, breaking down complex evaluations into simpler, sequential questions.



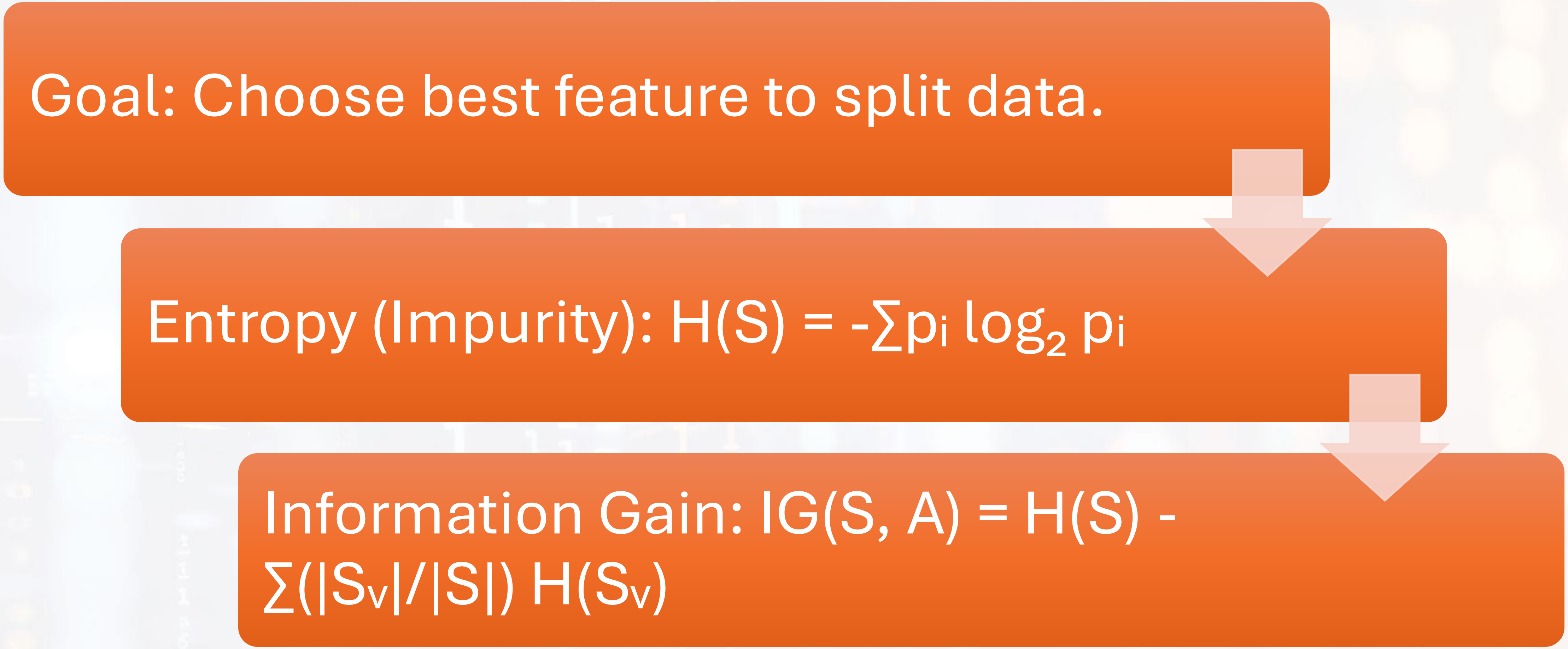
# Example: Predicting Exam Outcomes

- **Each leaf node represents a group of data points that have similar characteristics** and therefore are given the same prediction (Pass or Fail).
- For example, students who have studied between 2 to 6 hours, and have slept more than 6, are a similar group of students (from what was seen in the training data), and therefore the decision tree predicts they'll pass the exam.


Decision Tree to predict which students will pass the exam
Root Node
Hours of Study
F
Decision based on numerical value 'hours of study'
<2
2-6
>6
Decision Node
Fail
Hours of Sleep
Hours of Sleep
<6
>6
<3
>3
Decision Based on a category 'Previous Grade'
Previous Grade
Pass
Fail
Pass
Fail
Pass
Fail
Pass
Leaf Nodes

**Source: [7]**https://pub.towardsai.net/mastering-the-basics-how-decision-trees-simplify-complex-choices-7d2bd7dd35ba

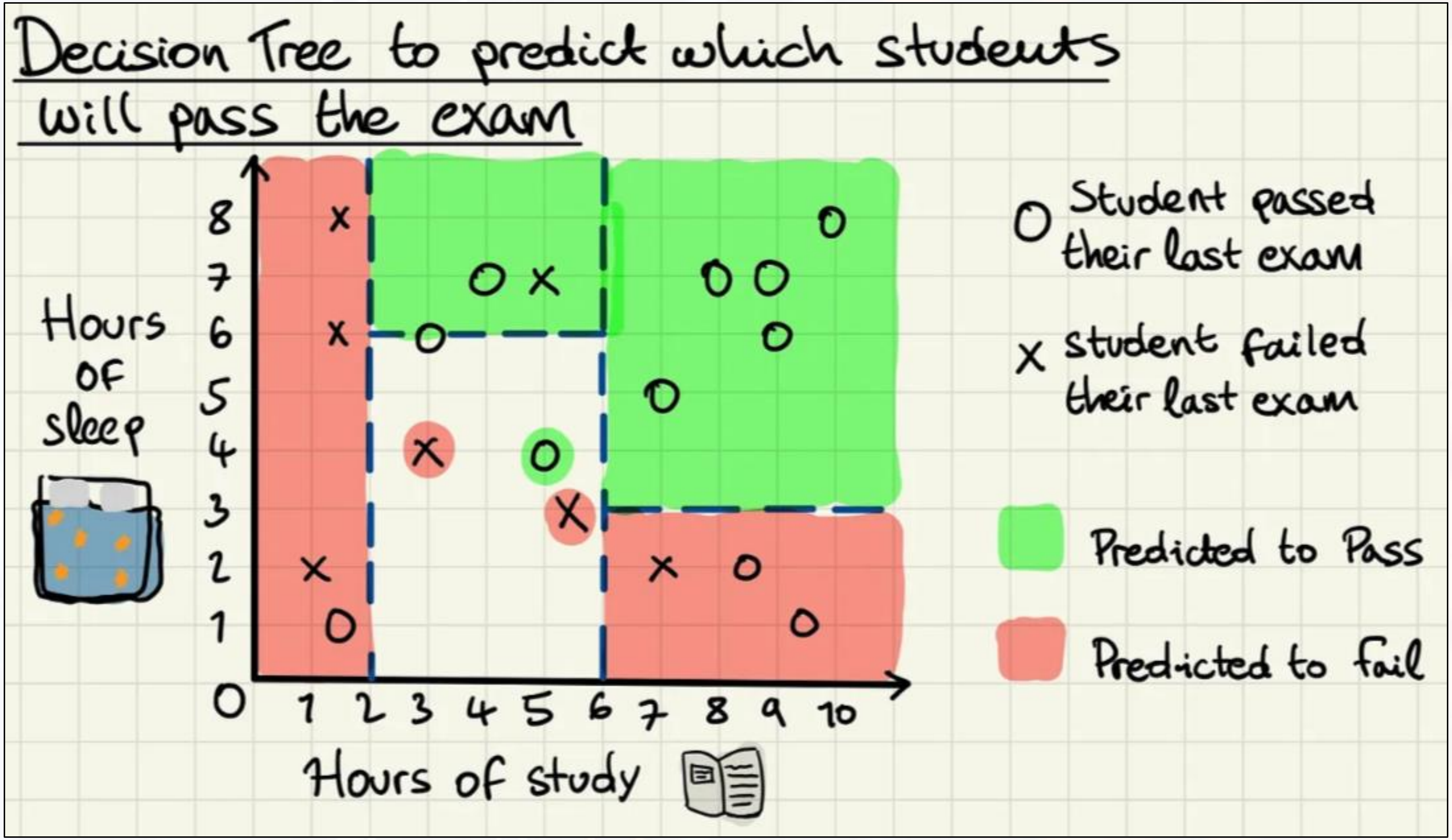
# Mathematical Reasoning Behind Decision Trees



Goal: Choose best feature to split data.

Entropy (Impurity): $H(S) = -\sum p_i \log_2 p_i$

Information Gain: $IG(S, A) = H(S) - \sum(|S_v|/|S|) H(S_v)$



**Source: [8]** https://pub.towardsai.net/mastering-the-basics-how-decision-trees-simplify-complex-choices-7d2bd7dd35ba

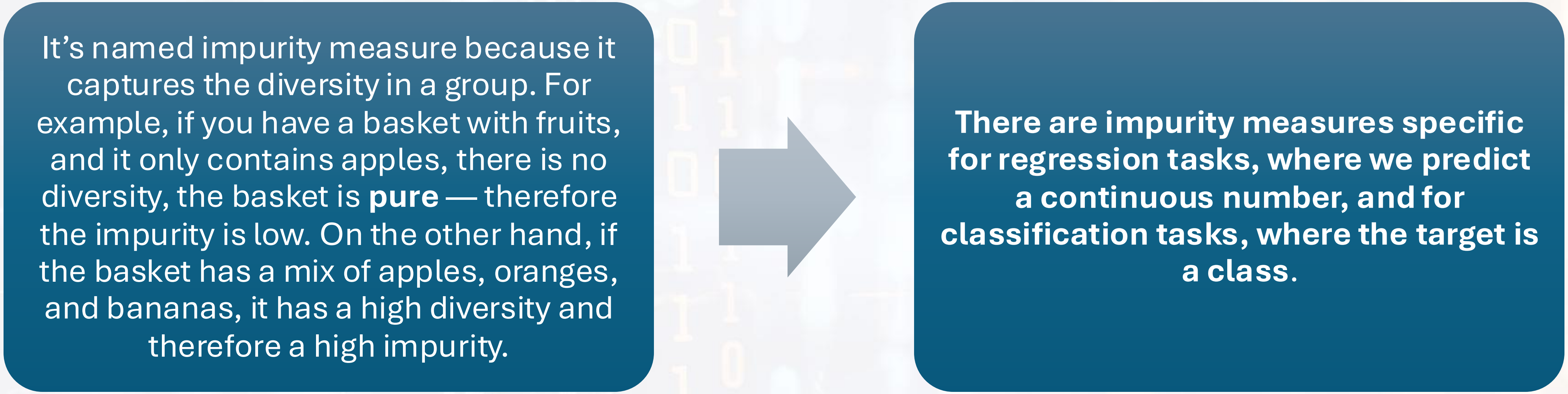
# How Decision Trees Learn

Decision trees do this by asking questions and using thresholds (numbers or categories) on the training data.

A **split** in a decision tree is a point where the data is divided based on a specific feature and threshold, creating branches. For example, in the case discussed earlier, one feature was the number of hours a student slept, with a threshold of ‘less than 2 hours.’ This split created a branch grouping students who slept less than two hours. These are predicted to fail their next exam.

To choose the best split decision trees **attempt all possible splits** (features and thresholds) and pick the one with the lowest **impurity**, a value that indicates how mixed or diverse the data in a group is. Lower impurity means the group has similar data, which is the aim of the learning process.

# Impurity Measure

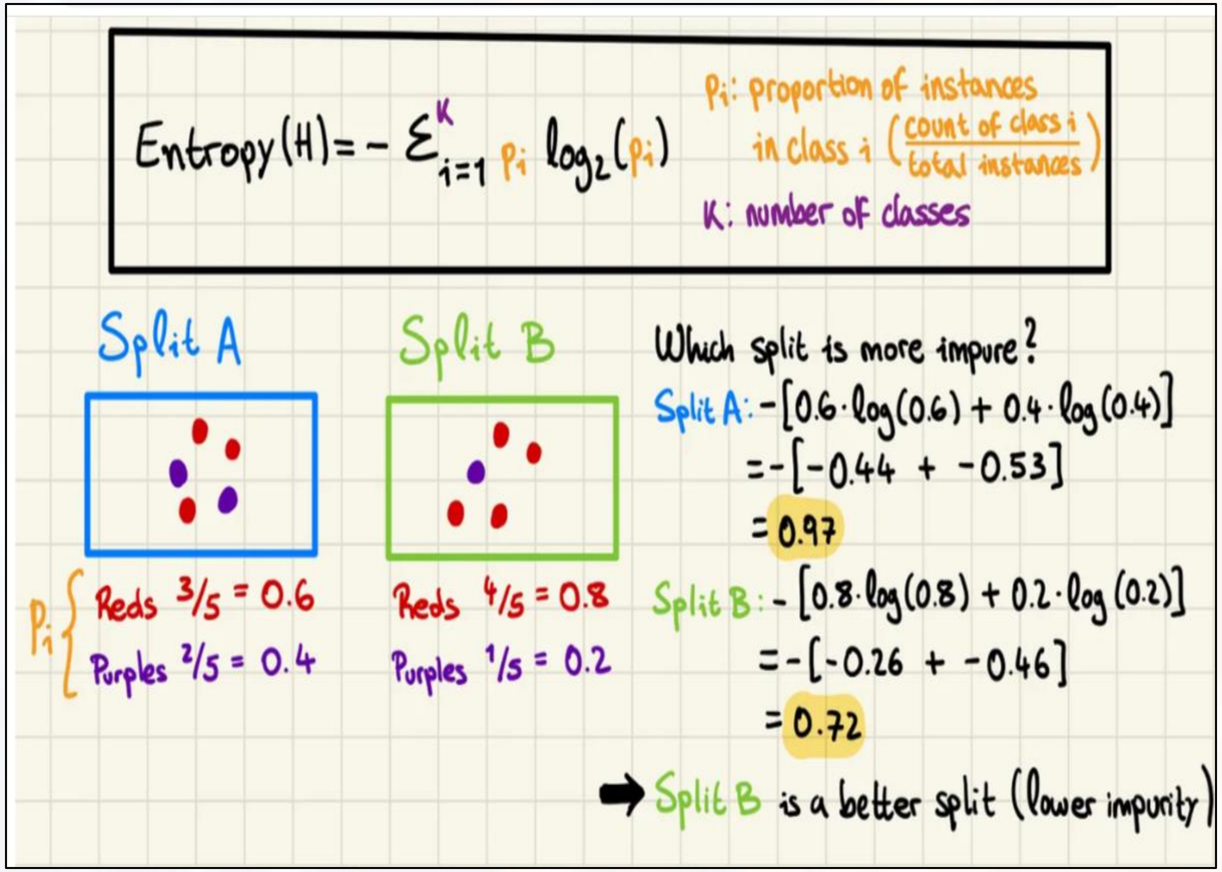


It’s named impurity measure because it captures the diversity in a group. For example, if you have a basket with fruits, and it only contains apples, there is no diversity, the basket is **pure** — therefore the impurity is low. On the other hand, if the basket has a mix of apples, oranges, and bananas, it has a high diversity and therefore a high impurity.

**There are impurity measures specific for regression tasks, where we predict a continuous number, and for classification tasks, where the target is a class**.

# Impurity Measure

- Let’s build intuition on the Entropy formula by understanding how different splits yield higher or lower Entropy values.
- Consider two splits, the first one, Split A, has 3 reds and 2 purple balls.
- The second split, Split B, has 4 reds and 1 purple.
- **Which one has a lower impurity?**



$$\text{Entropy}(H) = -\sum_{i=1}^{K} p_i \log_2(p_i)$$

$p_i$: proportion of instances in class $i$ $\left(\frac{\text{count of class } i}{\text{total instances}}\right)$

$K$: number of classes

$p_i$ { Reds $3/5 = 0.6$, Purples $2/5 = 0.4$

Reds $4/5 = 0.8$, Purples $1/5 = 0.2$

Which split is more impure?

Split A: $-[0.6 \cdot \log(0.6) + 0.4 \cdot \log(0.4)]$
$= -[-0.44 + -0.53]$
$= 0.97$

Split B: $-[0.8 \cdot \log(0.8) + 0.2 \cdot \log(0.2)]$
$= -[-0.26 + -0.46]$
$= 0.72$

➡ Split B is a better split (lower impurity)

**Source: [9]** https://pub.towardsai.net/mastering-the-basics-how-decision-trees-simplify-complex-choices-7d2bd7dd35ba

# Impurity Measure

The winner is **Split B**. The 5 balls are more similar to each other in Split B as there is a better division between red and purple balls.

# Pros & Cons of Decision Trees

Pros: Easy to understand, no need for feature scaling

Cons: Prone to overfitting, sensitive to small data changes

# How do you prevent overfitting?

You might have noticed that if you keep making splits looking to minimise impurity you will end up splitting the data so much, that it will isolate every single data point. This will create a massive tree that branches into leaf nodes that each represent just one sample from the training data.

To prevent this sort of overfitting we can introduce a **stopping criteria** that stops the decision tree from growing given certain conditions.

The stopping criteria is a set of rules that prevent the tree from growing too large and overfit the data.

**Maximum Depth Reached**
The tree stops growing when it reaches a set maximum depth (number of splits from root to leaf), it prevents overly complex trees.

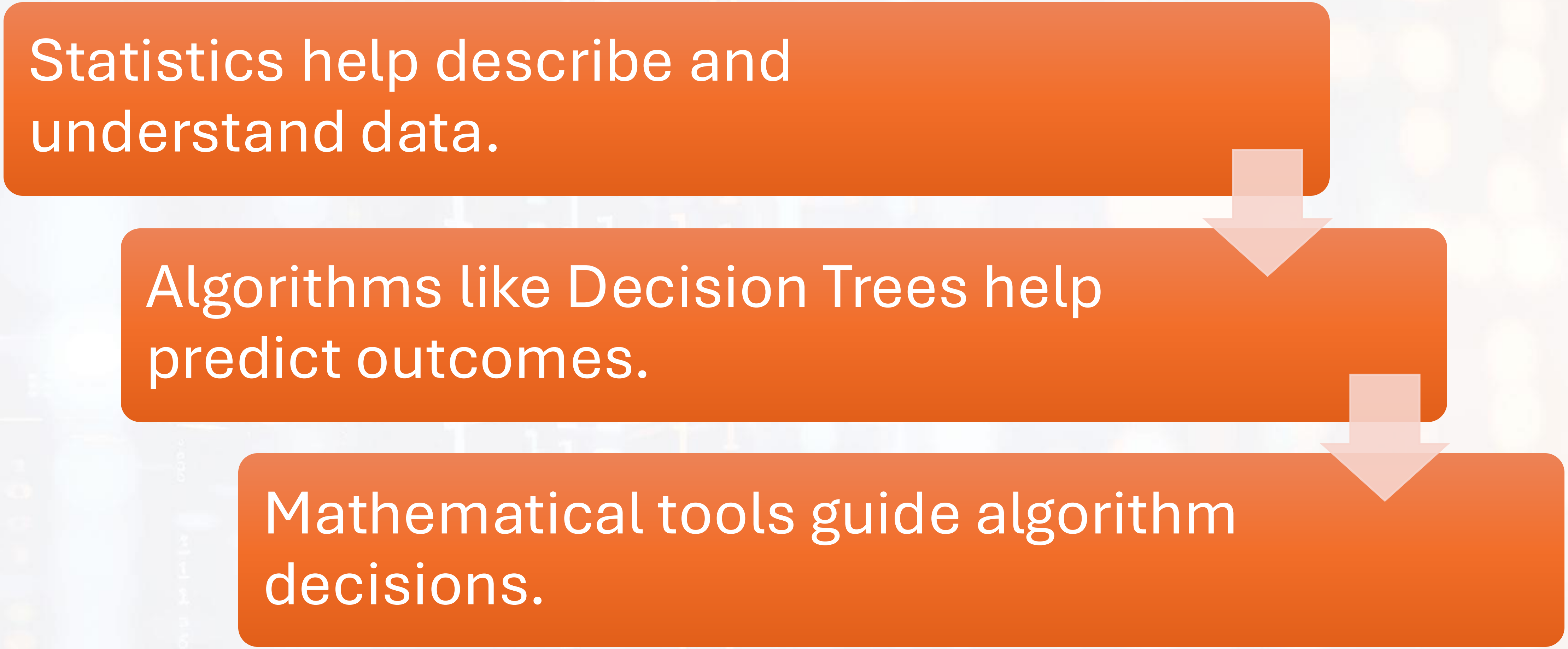
**Minimum Samples to Split**
A node must have at least a certain number of samples to be split further. This prevents splitting small, unreliable groups.

# From Tree to Ensemble

Combine multiple trees: Random Forest

Improves accuracy and reduces variance

# Key Takeaways



Statistics help describe and understand data.

Algorithms like Decision Trees help predict outcomes.

Mathematical tools guide algorithm decisions.

# Quiz Time!!

- **What is the key idea behind a decision tree?**

A. To train multiple models in parallel
B. To use linear regression for predictions
C. To split the data based on feature values to reduce uncertainty
D. To convert numerical features into categorical features

- **Ans: Correct Answer:** C

- **What metric is commonly used to decide splits in a decision tree?**
- A. Accuracy
  B. Information Gain or Gini Index
  C. Mean Squared Error
  D. R-squared
- **Ans: Correct Answer:** B

# Quiz

- **What happens if a decision tree is not pruned?**

A. It becomes faster
B. It generalizes better
C. It overfits the training data
D. It underfits the data

- Ans: **Correct Answer:** C

- **In a decision tree, what is a leaf node?**

A. A node with maximum information gain
B. A node used for pruning
C. A terminal node that gives the prediction
D. A node where all data points are numeric

- **Ans: Correct Answer:** C

# Any questions??

nicis

NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM

DIRISA