رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Microsoft Machine Learning Engineer

Course

Ministry of Communications and Information Technology

Digital Egypt Pioneers

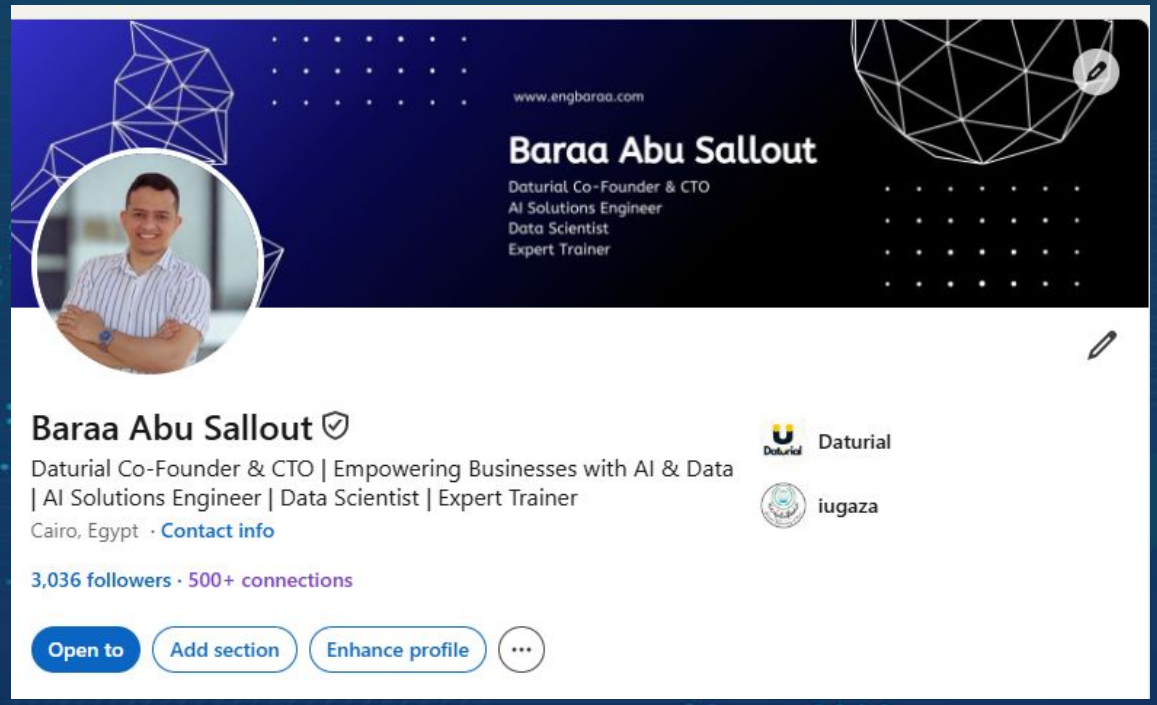
Linkedin Link

Who Am I ?

Eng. Baraa Abu Sallout - Palestinian From Gaza

- Bachelor's and Master's in Computer Engineering
- AI Solutions Engineer & Business Empowerment Specialist
- CTO at Daturial
- AI Development Consultant
- Data Scientist
- Expert Trainer & Consultant

  (Work in DEPI R1,R2, R3 with 11 Groups)
- Human & Mentor
- Lifelong Learner



Ministry of Communications and Information Technology

Digital Egypt Pioneers

NOW,,

Who are you?

Why are you here?

What’s your university major?

Are you working right now or not?

If you're working, what do you do?

GIZ4_AIS2_S2

الخريجين

DEPI / Graduates / Round 4																									
Friday 28 Nov 2025 - Thursday 21 May 2026 (24 Weeks Training & Coaching)																									
Training Week #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17		18	19	20	21	22	23	24	End
Week Start Day Friday	28-Nov	5-Dec	12-Dec	19-Dec	26-Dec	2-Jan	9-Jan	16-Jan	23-Jan	30-Jan	6-Feb	13-Feb	20-Feb	27-Feb	6-Mar	13-Mar	20-Mar	27-Mar	3-Apr	10-Apr	17-Apr	24-Apr	1-May	8-May	15-May	21-May
Session 1	Technical Sessions (17 WKs)																	Technical Sessions (6 WKs)							Project Discussion	
Session 2	Technical Sessions (17 WKs)																	English for FreeLancing (6 WKs)								
Session 3	Skills for Freelancing (6 WKs)						FreeLancing (6 WKs)						Coaching for Hunting Opertunities (10 WKs)													

الطلاب

DEPI / Students / Round 4																														
Friday 28 Nov 2025 - Thursday 23 July 2026 (24 Weeks Training & Coaching)																														
Training Week #	1	2	3	4					5	6	7	8	9	10	11	12	13		14	15	16	17	18	19					20	21	22	23	24	End
Week Start Day Friday	28-Nov	5-Dec	12-Dec	19-Dec	26-Dec	2-Jan	9-Jan	16-Jan	23-Jan	30-Jan	6-Feb	13-Feb	20-Feb	27-Feb	6-Mar	13-Mar	20-Mar	27-Mar	3-Apr	10-Apr	17-Apr	24-Apr	1-May	8-May	15-May	22-May	29-May	5-Jun	12-Jun	19-Jun	26-Jun	3-Jul	10-Jul	17-Jul	23-Jul
Session 1	Technical Sessions (17 WKs)																						Technical Sessions (6 WKs)										Project Discussion	
Session 2	Technical Sessions (17 WKs)																						English for FreeLancing (6 WKs)											
Session 3	Skills for Freelancing (6 WKs)										FreeLancing (6 WKs)								Coaching for Hunting Opertunities (10 WKs)															

Ministry of Communications and Information Technology

Digital Egypt Pioneers

Time is divided between

- **Group sessions:** we explain new concepts (aka 'theory')
- **Practise sessions:** you work on exercises or case studies or projects
- **Interactive sessions:** encompass a range of activities aimed at hands-on

In case of questions, remarks, suggestions, you can always interrupt us and just ask.

Feel lost?

Please stop me
Just ask,
We are here to help you.

Ministry of Communications and Information Technology

Digital Egypt Pioneers

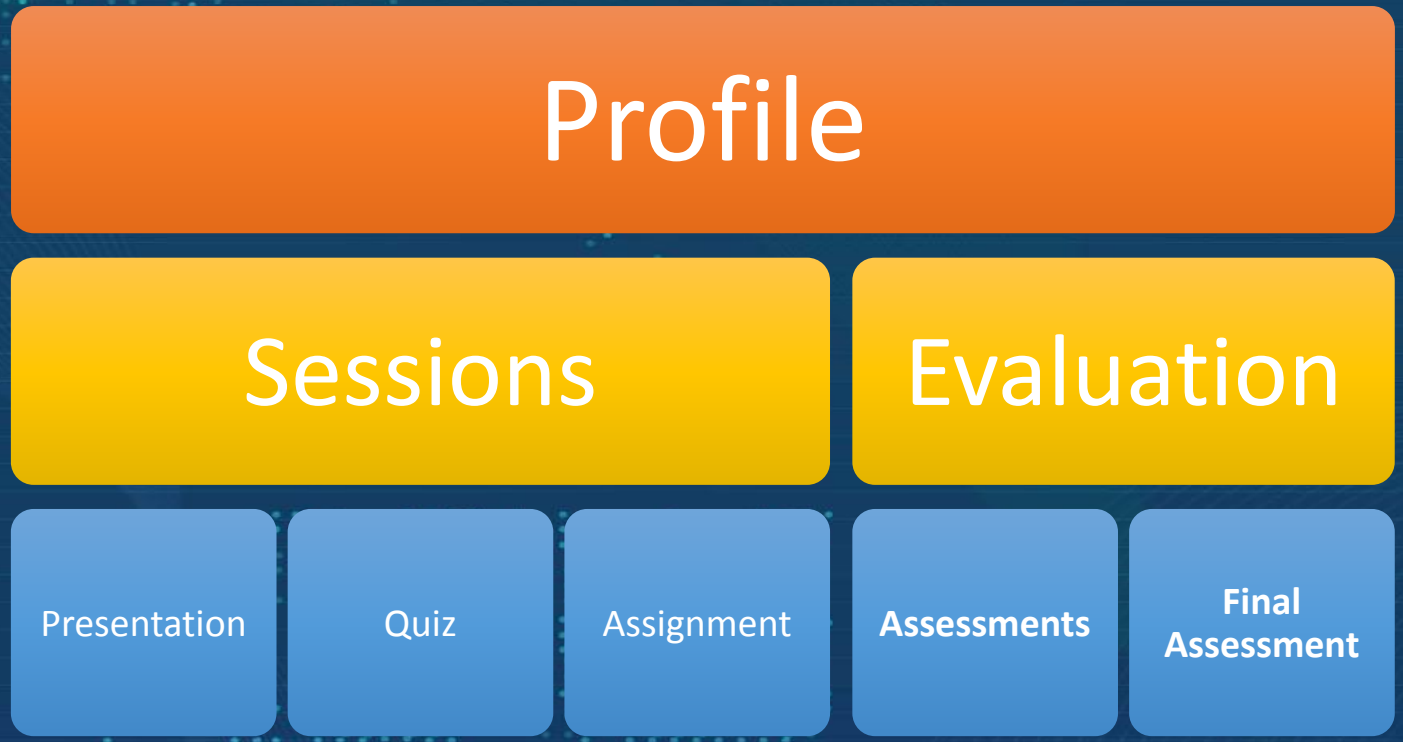
Our Rule of sessions

- **Students are expected to join sessions on time.**
- **Interactive, Interactive, Interactive.**
- **Assignments and small projects must be submitted on time.**
- **Random surprise questions during the session 😁 (don't answer be marked as absent.)**

Ministry of Communications and Information Technology

Academic Model

Digital Egypt Pioneers



Ministry of Communications and Information Technology

Microsoft Machine Learning Engineer

Digital Egypt Pioneers

Microsoft Machine Learning	
Course Name	**Hours**
Math	6
Python	12
Preprocessing & Visualization	15
Machine Learning	24
Deep Learning	9
NLP	18
Computer Vision	15
Azure	18
MLFlow	3
Total	**120**

Ministry of Communications and Information Technology

Microsoft Machine Learning Engineer

Digital Egypt Pioneers

1 - Math	Statistics Fundamentals	Week 1	2
	Linear Algebra Fundamentals		
2 - Python	Data Structures and Control Flow	Week 2	4
	Loops with Practical Application		
	Functions	Week 3	
	Object-Oriented Programming		
3 - Preprocessing & Visualization	Numpy	Week 4	5
	Pandas		
	Matplotlib & Seaborn	Week 5	
	Titanic Project		
	California housing Project	Week 6	
4 - Machine Learning	Basics of Machine Learning – Linear and Polynomial Regression	Week 7	8
	Data Preparation and Overfitting Control in Machine Learning		
	Logistic Regression and Model Evaluation	Week 8	
	Decision Trees and Naive Bayes		
	K-Nearest Neighbors and Support Vector Machines	Week 9	
	Feature Selection and Dimensionality Reduction		
	Ensemble Learning Techniques: Bagging and Boosting	Week 10	
	Introduction to Unsupervised Learning: Clustering Methods		

Ministry of Communications and Information Technology

Microsoft Machine Learning Engineer

Digital Egypt Pioneers

5 - Deep Learning	Introduction to Neural Networks and Deep Learning	Week 11	3
	Deep Learning: Overfitting, Regularization, and Network Optimization		
	TensorFlow and Model Fine-Tuning	Week 12	
6 - NLP	Introduction and Core Concepts and Techniques in NLP	Week 13	6
	Text Representation & Encoding Techniques in NLP		
	Sequential Data Modeling with RNNs and LSTMs	Week 14	
	Sequence Modeling with RNNs and Attention Mechanisms		
	Transformers	Week 15	
7 - Computer Vision	Introduction to Image Processing and Computer Vision	Week 16	5
	From MLPs to CNNs		
	CNN Family and Transfer Learning Techniques	Week 17	
	Object Detection: RCNN, Fast-RCNN, and YOLO		
	Generative Adversarial Networks (GANs) - Concepts and Applications		
8 - Azure	Microsoft Azure Fundamentals	Week 18	6
	Microsoft Azure AI Fundamentals		
	Introduction to Machine Learning	Week 19	
	Computer Vision		
	Natural Language processing Fundamentals		
9 - MLFlow	MLOPS Tools	Week 20	1

‘’Every positive jump for humanity has been fuelled by intelligence.“

What is intelligence?

The ability for solving problems

What is artificial intelligence (AI) ?

Artificial Intelligence is a set of computer science techniques that allows computer software

to learn from experience,

adapt to new inputs

And complete tasks that resemble human intelligence.

Why talk AI now?

Computing power is AI's engine
Data is AI's fuel
Algorithms are AI's design

Applications of AI in Various Industries

Machine learning & AI being used everyday in more ways than ever

INDiGROWTH insights driven growth

ON YOUR SMARTPHONE...

Ok Google

What channel does GoT Air On?

Hey Siri

Hey Cortana

Translate .

Que voulez-vous dire...

Maps

Way from the airport to home

WHEN YOU'RE...

FB Moments

Pics of you & I at Anna's party

Shopping

Customers who bought This utem also ..

Videos

Other movies you might...

Music

Recommened

Email

Primary inbox, smart reply

MAKING BUSINESS HAPPEN...

Robo-advisor

Your Investment Portfolio

Scoring Engine

Writing Proficiency

Marketing & Advertising

Brining it all together in Real-Time

Fraud Detection

Machine Learning at play

Machine Learning Use Cases in Finance

Financial Monitoring

Making Investment Predictions

Process Automation

Secure Transactions

Risk Management

Algorithmic Trading

Financial Advisory

Customer Data Management

Decision Making

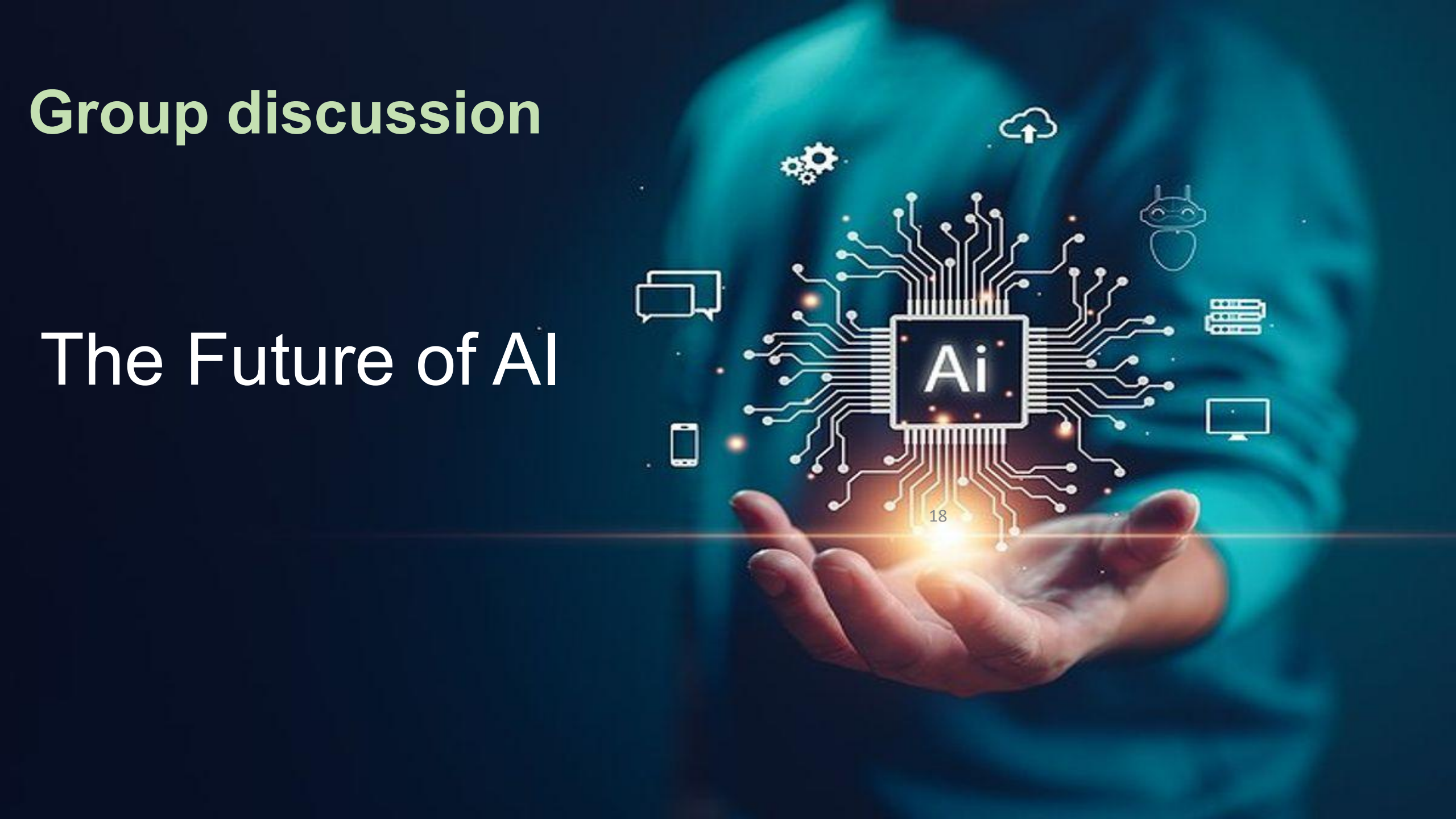
Customer Service Level Improvement

Customer Retention Program

Marketing

Group discussion

The Future of AI



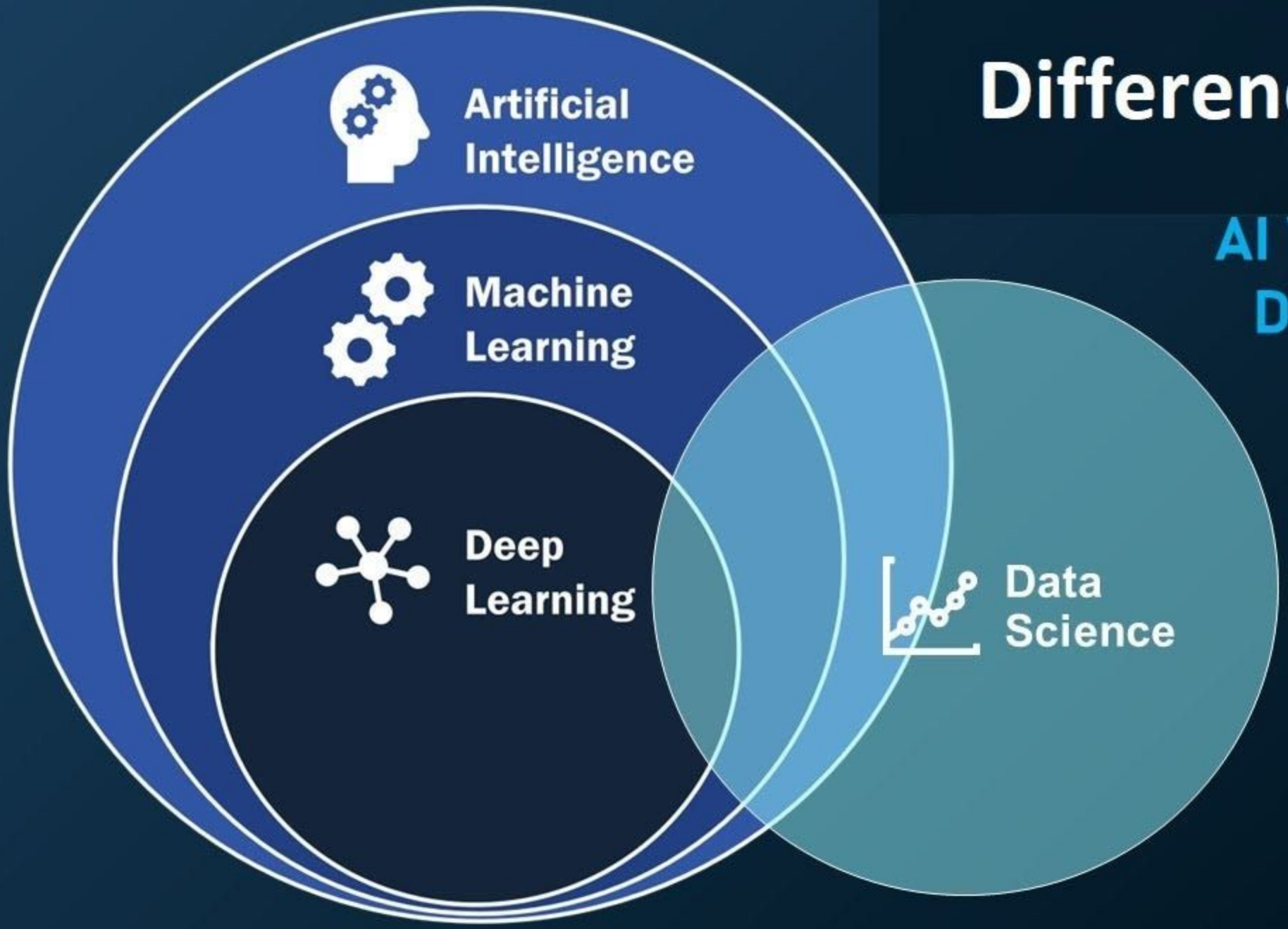
رواد مصر الرقمية

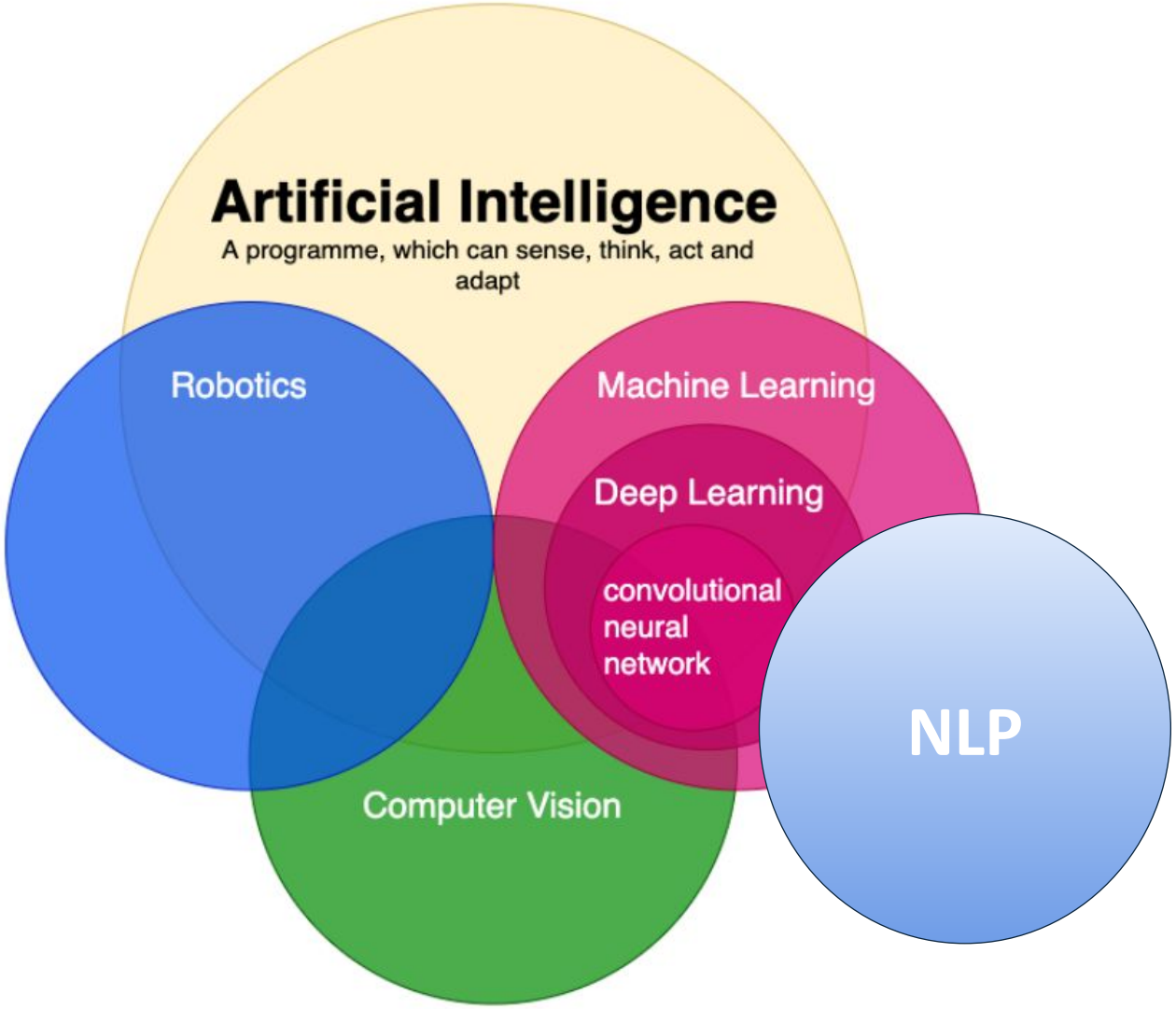
وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Microsoft Python Pre Assessment - Microsoft

Difference Between

AI VS ML VS DL VS DS




Artificial Intelligence
A programme, which can sense, think, act and adapt
Robotics
Machine Learning
Deep Learning
convolutional neural network
NLP
Computer Vision

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Fundamentals & Paths to a Career of AI and Data Science

1. **Learn Programming**: Start with Python or R.
2. **Understand Math and Statistics**: Basics of linear algebra and statistics are essential.
3. **Data Understanding (Data Analysis, Visualization and Exploration)**
4. **Data Preprocessing and Cleaning**
5. **Machine Learning**: Learn basic algorithms like Linear Regression and Decision Trees.
6. **Deployment**
7. **Real Projects**: Participate in projects or competitions like Kaggle.
8. **Continuous Learning**: Keep up with new courses and research.
9. **Choose a Specialization**: Focus on a specific field like financial analysis or healthcare
10. **Communication and Interpretation.**

Data Scientist Skills

01

Data Analysis

Ability to analyze large datasets using statistical methods and machine learning algorithms.

02

Programming Skills

Proficiency in languages like Python, R, or SQL for data manipulation and analysis.

03

Data Visualization

Creating visual representations of data to communicate insights effectively.

04

Domain Knowledge

Understanding of the specific industry or domain to interpret data in context.

05

Problem-Solving

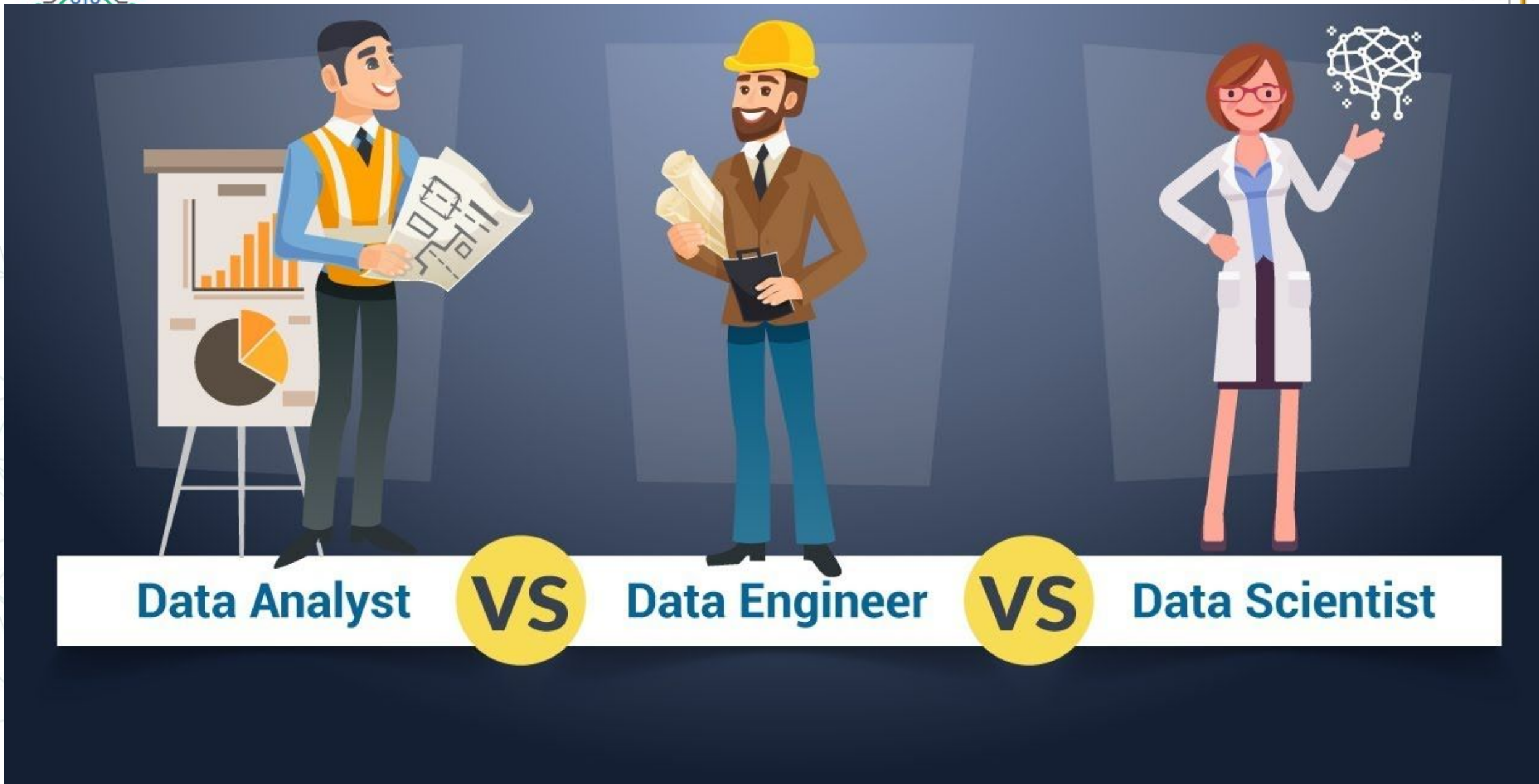
Applying analytical skills to solve complex business problems using data-driven approaches.

06

Communication

Effectively communicating findings and insights to stakeholders through reports and presentations.

Data Different Jobs



Data Scientist VS Data Analyst VS Data Engineer VS Machine Learning Engineer

- **Data Analyst:**
  - Think of a Data Analyst as a detective who is always investigating, "What happened, and how can we improve things?"

- **Data Engineer:**
  - Think of a Data Engineer as the engineer who builds the pipelines that deliver clean water (data) to everyone.

- **Data Scientist:**
  - Think of a Data Scientist as the explorer guiding the ship into the unknown, always asking, "What can I discover in this data?"

- **Machine Learning Engineer:**
  - Think of a Machine Learning Engineer as the engineer who turns theoretical designs into reality, making machines work intelligently in the real world.

رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

TOP 50 AI & ML JOBS

Research & Development

- AI Research Scientist
- Machine Learning Researcher
- Deep Learning Scientist
- Computer Vision Research Scientist
- Natural Language Processing (NLP) Scientist
- AI Algorithm Engineer
- Reinforcement Learning Scientist

AI Implementation & Support

- AI Implementation Engineer
- AI Systems Integrator
- AI Support Specialist
- AI Deployment Engineer
- AI Technical Support Engineer

Business & Strategy

- AI Strategist
- AI Consultant
- AI Business Analyst
- AI Project Manager

AI Product Development

- AI Product Manager
- AI Product Owner
- Technical Product Manager (AI/ML Focus)
- AI Application Developer
- AI Solutions Architect

Niche AI Roles

- Robotics Engineer (AI Focus)
- Autonomous Systems Engineer
- AI Healthcare Specialist
- AI Financial Analyst
- AI Content Strategist

ML Engineering

- Machine Learning Engineer
- Machine Learning Infrastructure Engineer
- Deep Learning Engineer
- AI Software Developer
- Data Scientist (Machine Learning Focus)
- Machine Learning Operations (MLOps) Engineer
- AI/ML DevOps Engineer

AI Ethics & Governance

- AI Ethics Officer
- AI Governance Specialist
- AI Policy Advisor

Emerging Technologies

- Quantum Machine Learning Specialist
- AI Edge Computing Engineer
- AI for IoT (Internet of Things) Engineer
- Generative AI Specialist

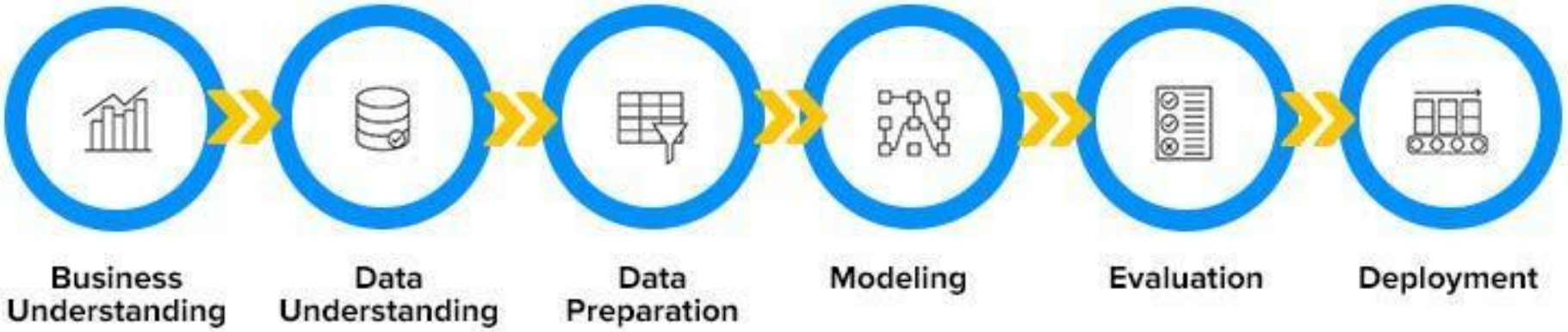
Leadership & Executive

- Chief AI Officer (CAIO)
- AI Engineering Lead
- Head of AI/ML
- Director of AI Research

Data Roles

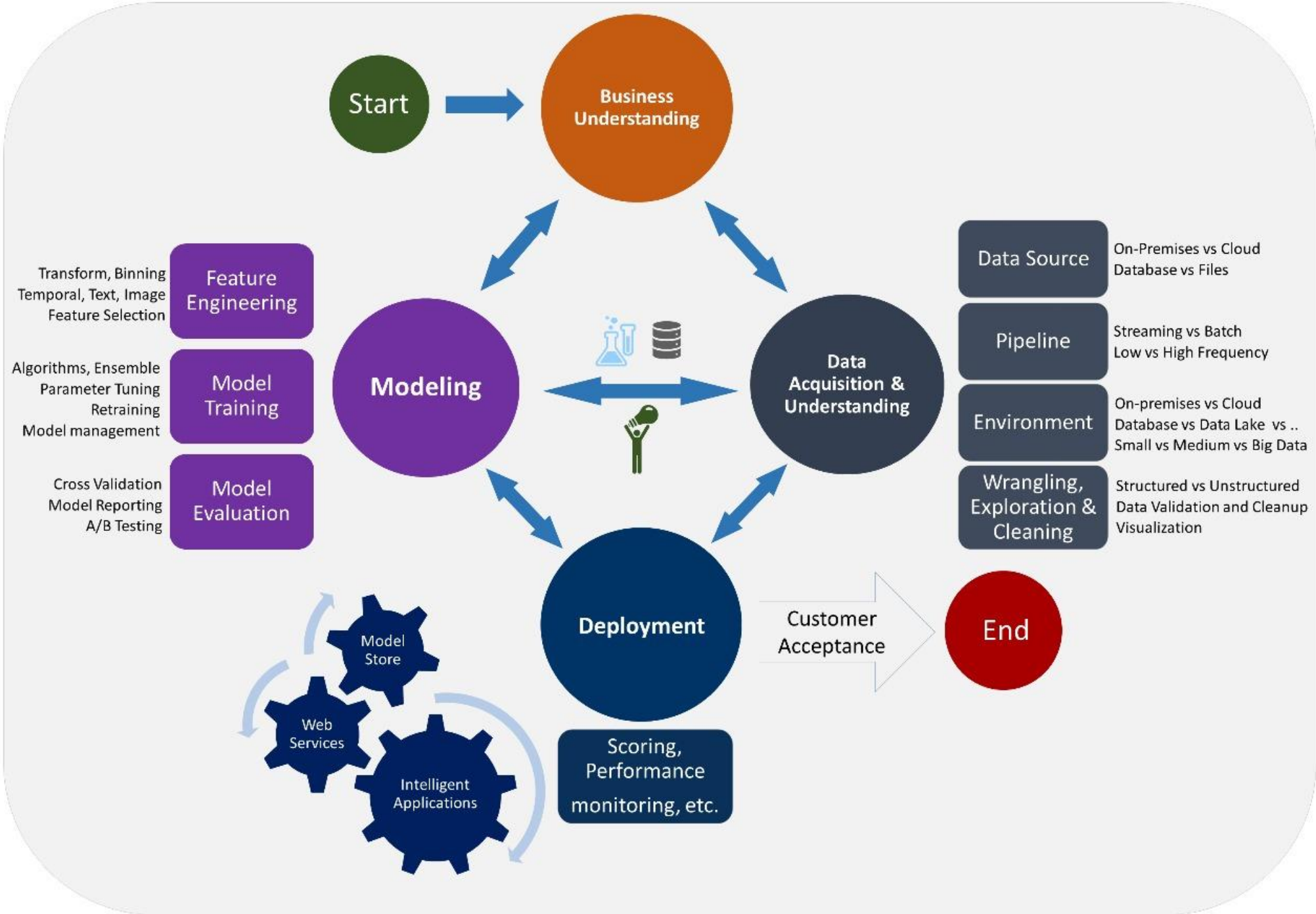
- Data Engineer
- Data Scientist
- Data Analyst
- Big Data Engineer
- Data Architect

The Data Science Process



رواد مصر الرقمية

Data Science Process



AI in production: expectation

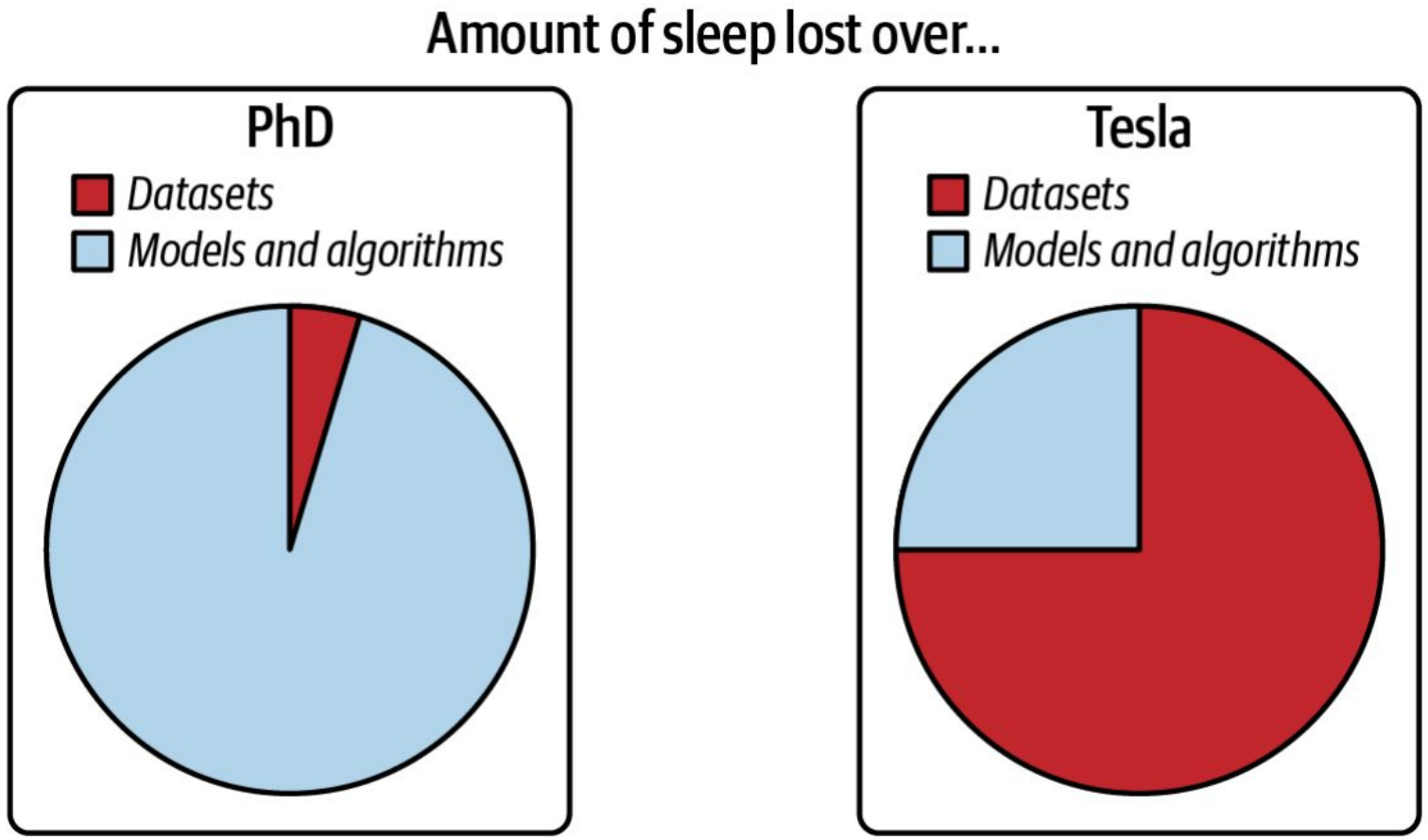
1. Collect data
2. Work in Data
3. Train model
4. Deploy model
5.

AI in production: reality

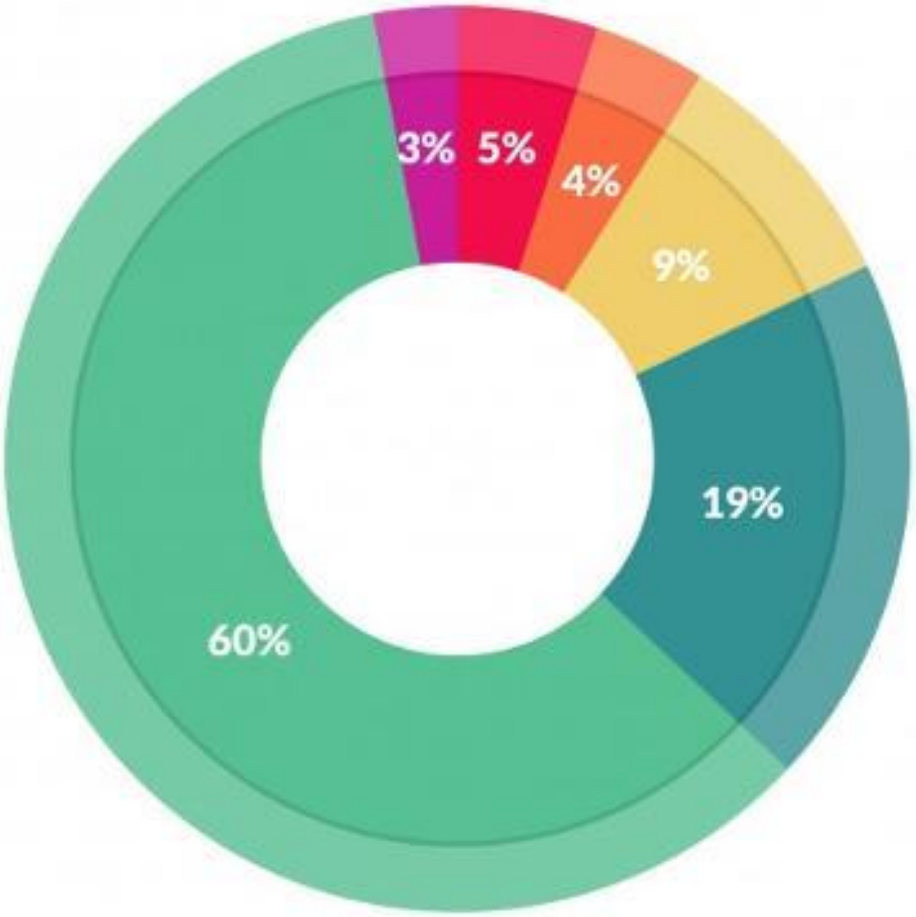
1. Choose a metric to optimize
2. Collect data
3. Work in Data
4. Train model
5. Realize many labels are wrong -> relabel data
6. Train model
7. Model performs poorly on one class -> collect more data for that class
8. Train model
9. Model performs poorly on most recent data -> collect more recent data
10. Train model
11. Deploy model
12. Dream about $$$
13. Wake up at 2am to complaints that model biases against one group -> revert to older version
14. Get more data, train more, do more testing
15. Deploy model
16. Pray
17. Model performs well but revenue decreasing
18. Cry
19. Choose a different metric
20. Start over

Step 16 and 18 are essential

Data in research vs. data in production

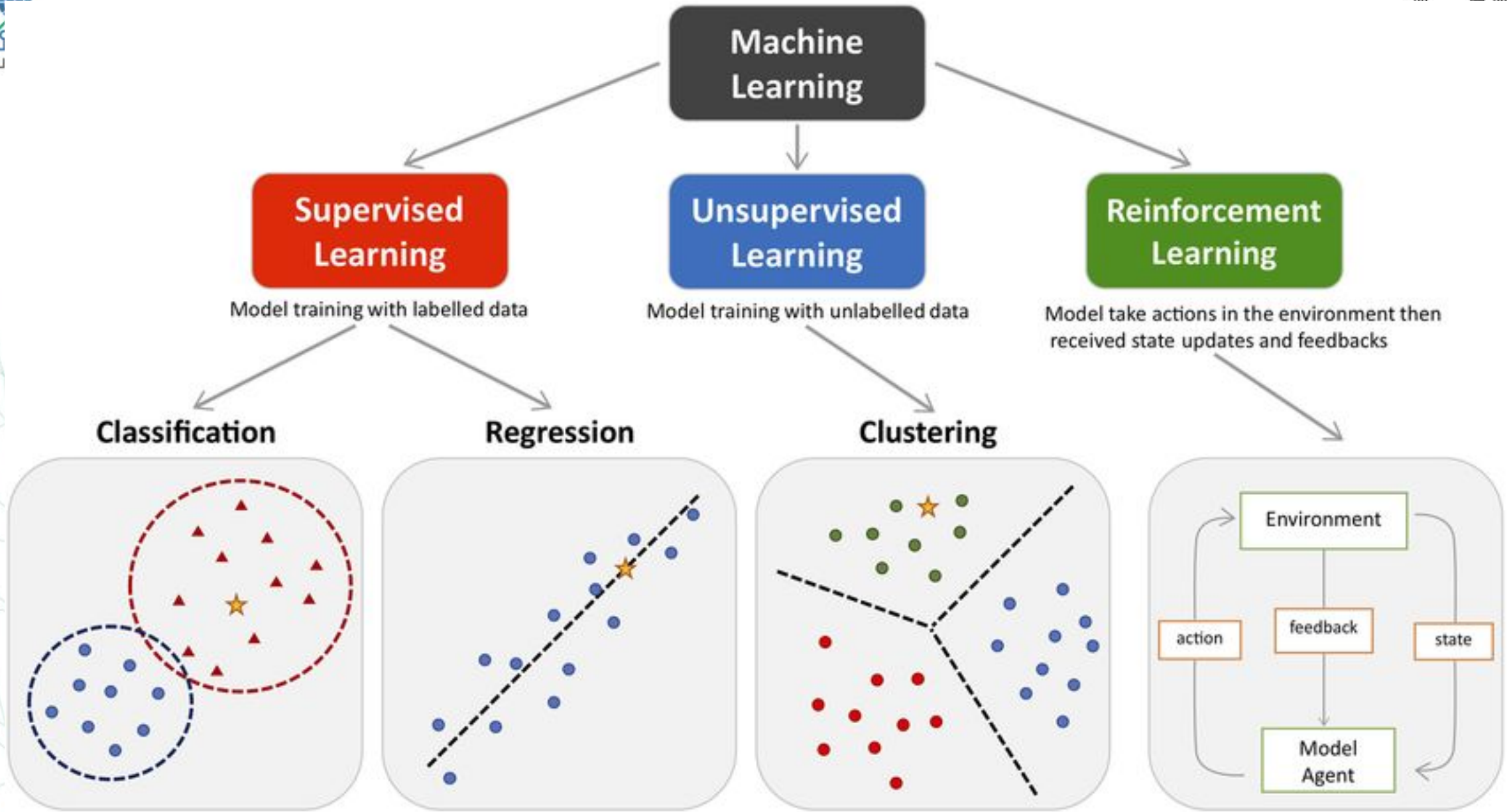


Andrej Karpathy (Former Director of AI at Tesla)



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Model training with labelled data
Model training with unlabelled data
Model take actions in the environment then received state updates and feedbacks
Classification
Regression
Clustering
Environment
action
feedback
state
Model Agent

Label Data

Name	Age	Experience	Martial_Status	Salary
John	25	4	YES	1000 USD
Nathan	26	5	NO	1200 USD
Garima	27	6	YES	1500 USD
Alice	26	5	NO	1200 USD
Mark	32	10	YES	2000 USD
Saurabh	35	13	NO	3000 USD

Labelled data

Name	Age	Experience	Martial_Status	Salary
Eric	27	2	NO	??

Unseen data

Hence, we learned from past histrorically availaible data (prior knowledge) for all the employees and predict it for any unseen or future data (Eric in our case).

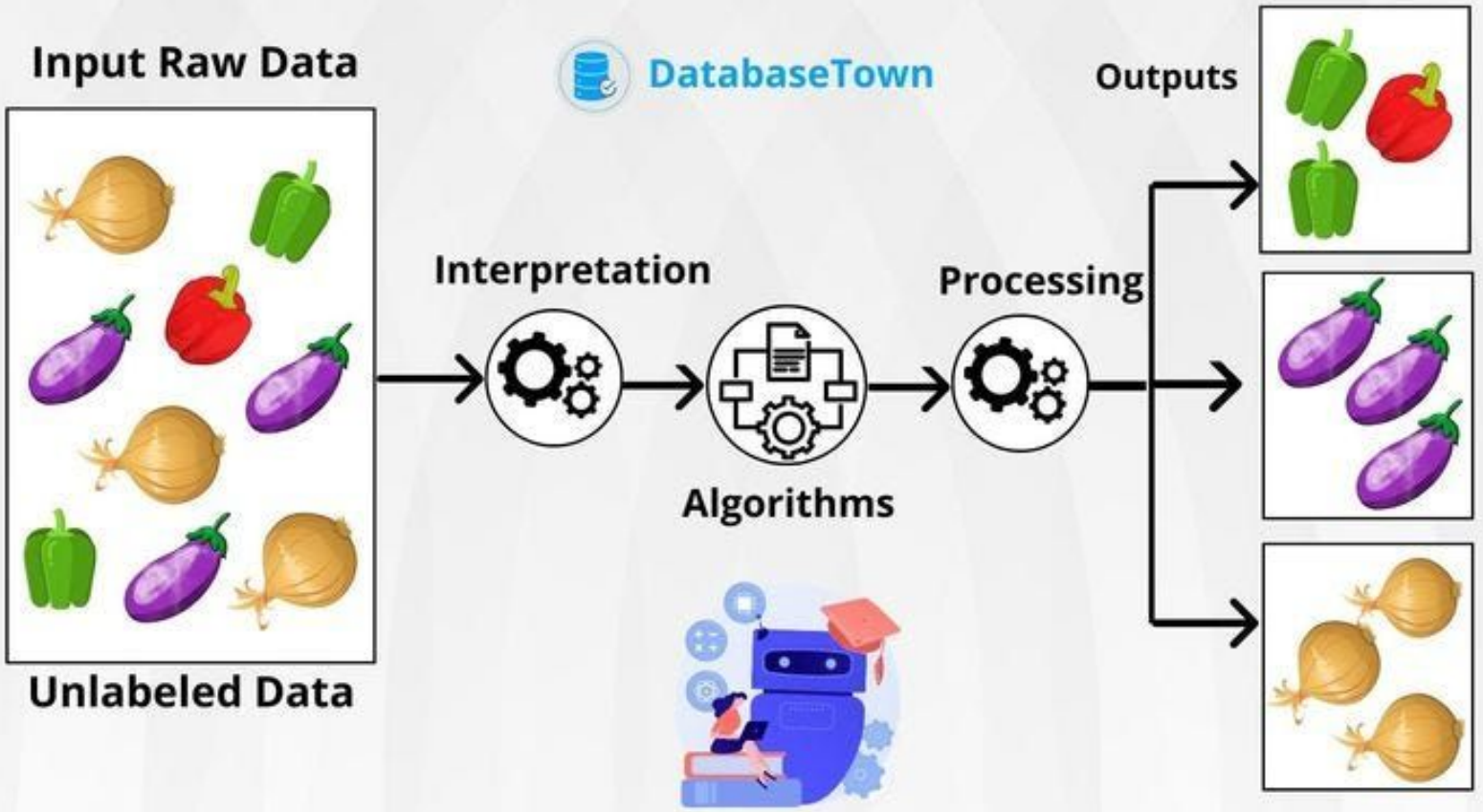
UnLabel Data

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

unlabeled

UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data without any predefined outputs or target variables.



Why Do We Need Data Preprocessing?

Real-World Data Is Messy

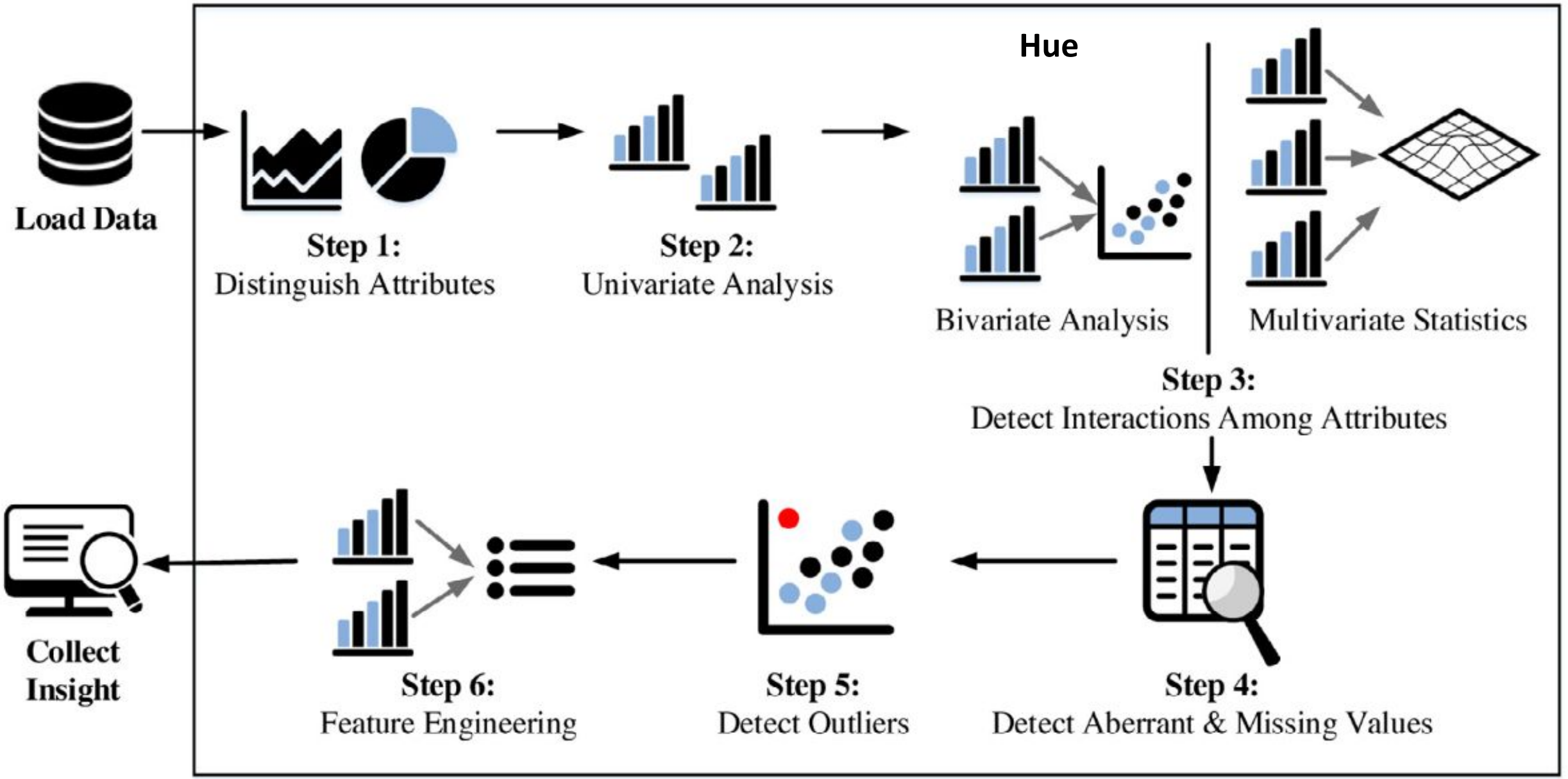
- Incomplete
- Inconsistent
- Noisy
- Contains errors
- Not ready for machine learning

Raw data ≠ usable data.

ML performance depends **more on data quality** than on the chosen model.

“Better Data → Better Features → Better Model.”

Exploratory Data Analysis overview



Why EDA is important in the machine learning process?

In general, EDA is a critical step in the machine learning process as it helps in :

- Understanding the data
- Assessing data quality
- Guiding feature engineering
- Informing model selection
- Effectively communicating results

Data Preprocessing

- Data Cleaning or Cleansing
- Work with Missing data
- Detect and Handle Outliers

>>> Data analysis (to u, to technical, to)

>>> Feature Engineering

- Deal with Imbalanced classes
- Work with Categorical data
- Feature Scaling
- Split data to Train and Test Sets

Feature Engineering and Extraction

- Domain knowledge features (age , wh, **BMI**)
- Date and Time features
- String operations
- Web Data
- Geospatial features

“You don’t choose algorithms by memorizing them.

You choose them by understanding the data: its size, shape, noise, and what problem you want to solve. Every algorithm has a personality—strengths, weaknesses, and situations where it shines.”

Data Cleaning (First Step in Preprocessing)

"Garbage in → Garbage out"

Always **diagnose your dataset** before training

→ “Bad data in → bad model out.”

→ "Garbage in → Garbage out"

Why Cleaning First?

- Raw data is always messy.
- ML models are **sensitive** to errors.
- Cleaning builds the **foundation** for all later steps.
- Example: You can’t cook with dirty vegetables → first wash & clean.

Quick Data Inspection

- **`df.shape`** **`# rows & columns`**
  - Tells you (rows, columns).
  - Example: (2000, 10) → 2000 records, 10 features.
- **`df.info()`** **`# data types`**
  - Shows column names, data types (int, float, object, datetime).
  - Detects mismatches:
    - Age stored as “object” instead of “int”.
    - Date stored as string instead of datetime.
- **`df.describe()`** **`# summary stats`**
  - Mean, min, max, std for each numeric column.
  - Quickly spot impossible values (e.g., Height min = -10).

Like a **doctor’s first check-up**: weight, temperature, blood pressure.

Before deep analysis, you need to know the “overall health” of the dataset.

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

Math Basics

Session 1: (Statistics Fundamentals)

Agenda:

1	Introduction to Statistics (30min)
2	Descriptive Statistics (1 hour)
3	Probability Basics (1 hour)
4	Introduction to Inferential Statistics (30 mins)

Introduction to Statistics (30min)

Importance of Statistics in AI and Machine Learning

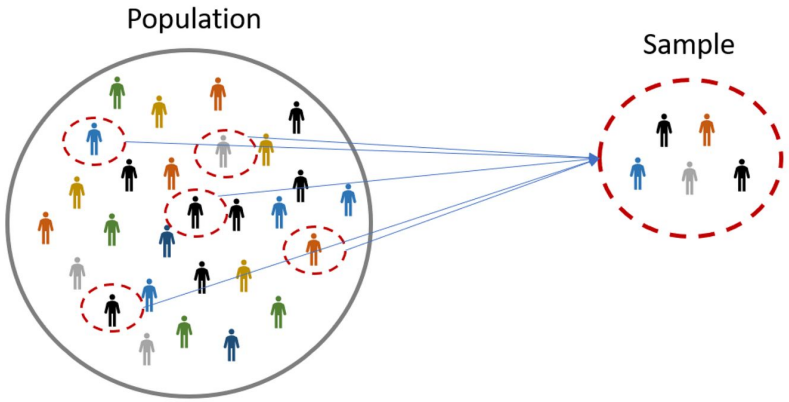
Why Learn Statistics for AI?

- **Data Analysis**: Statistics provide the tools to analyze and interpret data, which is the backbone of AI. Understanding data distributions, variability, and patterns is crucial for building robust models.
- **Model Evaluation**: In AI, especially in machine learning, evaluating the performance of models (e.g., accuracy, precision, recall) relies on statistical measures. Without a strong understanding of statistics, interpreting these metrics can be challenging.
- **Decision Making**: AI models often make decisions based on probabilities. For example, a classification model might predict the likelihood of an email being spam. Understanding probability and statistical inference is key to trusting and explaining these decisions.
- **Real-World Applications**: From predicting customer behavior in marketing to diagnosing diseases in healthcare, statistics provide the foundation for analyzing data and making informed decisions.

Basic Statistical Definitions

Population vs. Sample

- **Population**: The entire group of individuals or items that you're interested in studying. For example, if you want to study the average height of adult males in a country, the population would be **all** adult males in that country.
- **Sample**: A subset of the population that is used to represent the whole. For instance, measuring the height of **1,000 adult males** from different regions could be a sample representing the entire population. Sampling is often done because it is impractical or impossible to study the whole population.

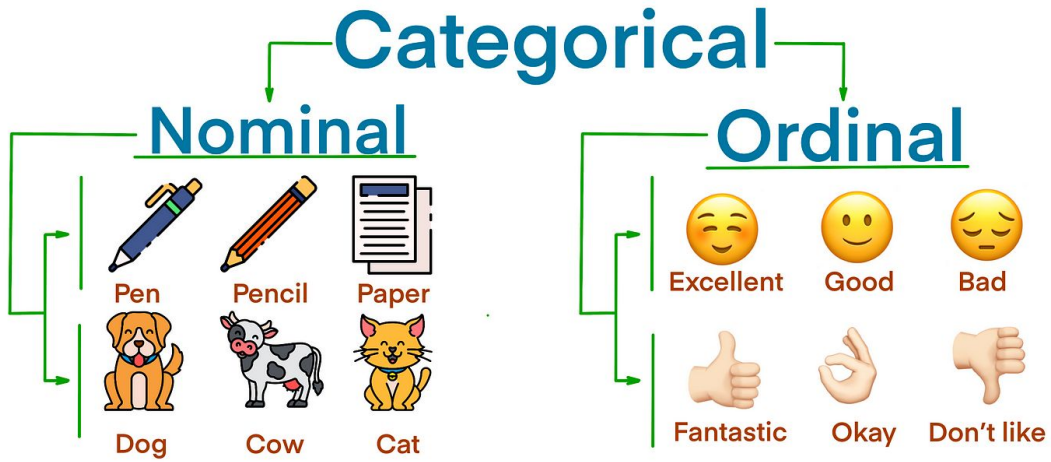


Types of Variables

Categorical (Qualitative) Variables

These variables represent categories or groups and can be either:

- **Nominal**: **No inherent order among categories**. Examples include colors (red, blue, green), gender (male, female).
- **Ordinal**: There is **an inherent order**, but the difference between levels is not quantifiable. Examples include customer satisfaction ratings (poor, fair, good, excellent), education levels (high school, bachelor's, master's).



Types of Variables

Numerical (Quantitative) Variables

These variables represent measurable quantities and can be either:

- **Discrete**: Countable, finite values. Examples include the number of students in a class, the number of cars in a parking lot.
- **Continuous**: Infinite possibilities within a range. Examples include height, weight, temperature.

Real-World Relevance and Applications

Examples of Statistics in AI Applications

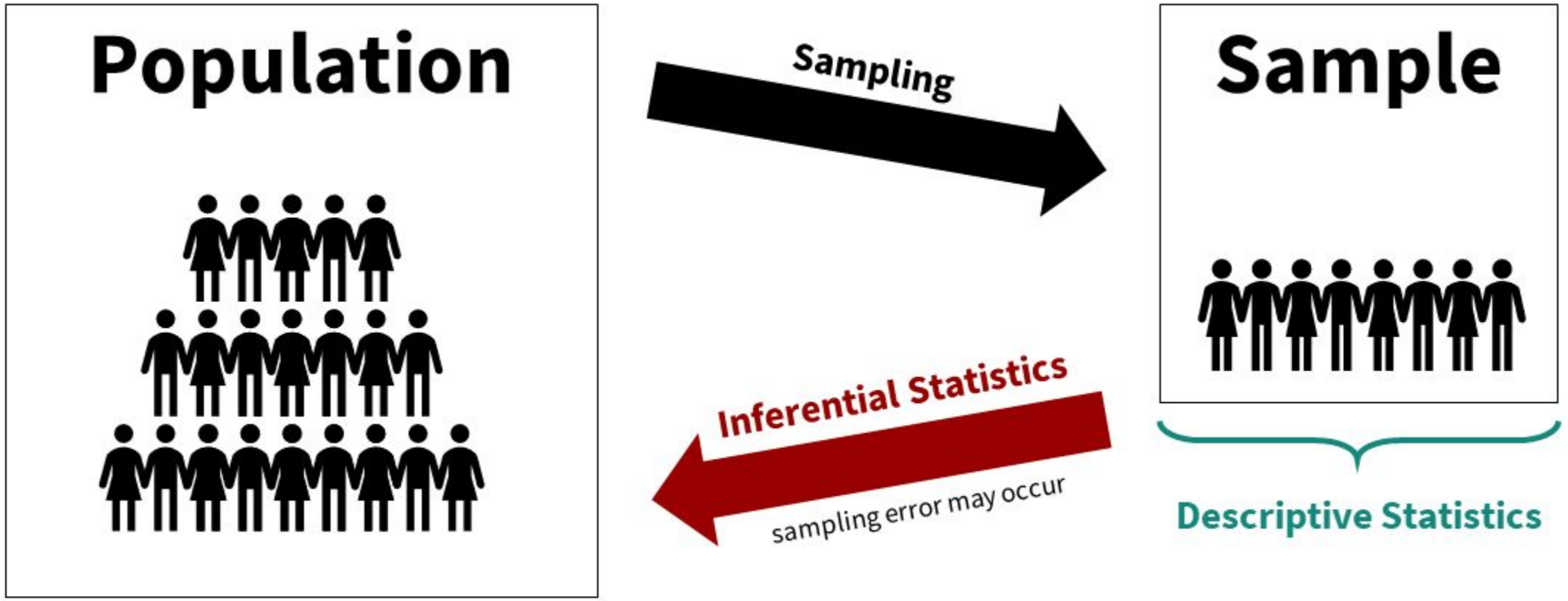
- **Fraud Detection**: Analyzing transaction data to identify patterns that indicate fraudulent activity. Here, statistics help to detect anomalies that deviate from normal behavior.
- **Recommendation Systems**: E-commerce platforms like Amazon use statistics to analyze customer preferences and recommend products based on what similar customers have purchased.
- **Healthcare Analytics**: Predicting the likelihood of diseases based on patient data. Statistical models can assess risk factors and provide personalized healthcare recommendations.

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Descriptive Statistics
(1 hour)

Descriptive statistics vs Inferential statistics



Descriptive Statistics

Descriptive statistics help provide a clear understanding of data through numerical calculations, graphs, and tables. This is a crucial step before conducting any further statistical analysis or building machine learning models.

Cases vs. Variables

Cases

- A case represents an individual entity or subject in a dataset on which measurements or observations are made. It can be a person, a country, an event, or any other subject of study.

Variables

- A variable is a characteristic or attribute that can be measured or observed for each case. Variables are features or properties that describe the cases

Variables

	Store	Sales	Customers	Refunds
Cases	1	14	9	3
	2	19	13	4
	3	22	19	4
	4	24	20	3
	5	29	26	8
	6	40	34	6

Data Matrix vs. Frequency Table

Data Matrix

A data matrix is a structured table where each row represents a single case (individual data point), and each column represents a variable (characteristic or feature of the cases).

Variables →

Cases ↓

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
6.5	3.2	5.1	2	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3	5.5	2.1	Iris-virginica
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3	4.5	1.5	Iris-versicolor
5.8	2.7	4.1	1	Iris-versicolor

Data Matrix vs. Frequency Table

Data Matrix

Purpose

The data matrix is useful for raw data representation where each case's specific details are important. It's a comprehensive way to store all the information but not necessarily the best for summarizing data.

When to Use

Data matrices are used in scenarios where you need to keep all details about each individual case, such as during data collection or when you need to perform case-by-case analysis.

Data Matrix vs. Frequency Table

Frequency Table

A frequency table is a summary of the data that shows how often each value of a variable occurs. It can display frequencies, percentages, and cumulative percentages.

Score	Frequency
50-59	2
60-69	2
70-79	6
80-89	7
90-99	3

Data Matrix vs. Frequency Table

Frequency Table

Purpose

Frequency tables are used to summarize and visualize the distribution of a variable, making it easier to see patterns and understand the data's structure.

When to Use

When you want to quickly grasp how data is distributed across different categories or ranges, especially when dealing with categorical or quantitative variables.

Descriptive Statistics

2 Types of Descriptive Statistics

Measures of central tendency

- Mean
- Median
- Mode

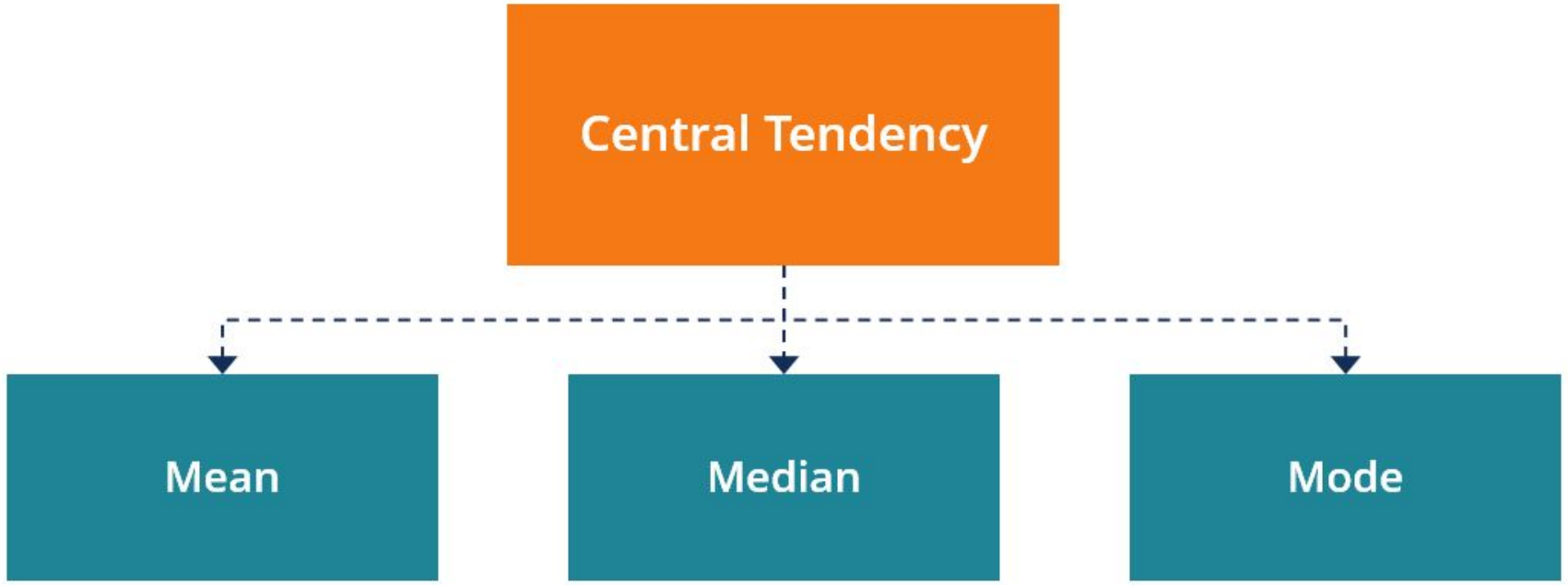
Measures of dispersion

- Range
- Interquartile range
- Variance
- Standard Deviation

Descriptive Statistics

Measures of Central Tendency

Understand how to describe the center of a dataset using different measures and learn how to calculate and interpret these values.



Measures of Central Tendency

Mean (Average)

The mean is the sum of all values divided by the number of values. It is a commonly used measure of central tendency.

Formula

For a dataset with values $x_1, x_2, \ldots, x_n$, the mean μ is calculated as:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example

For the dataset $[4, 8, 6, 5, 3, 4]$, the mean is

$$\mu = \frac{4 + 8 + 6 + 5 + 3 + 4}{6} = 5$$

The mean is sensitive to outliers. If the dataset has extreme values (very high or very low), the mean can be misleading

Measures of Central Tendency

Median

The median is the middle value of a dataset when it is ordered from smallest to largest. If the dataset has an even number of values, the median is the average of the two middle values.

Example

For the dataset [4,8,6,5,3,4], first sort the data:[3,4,4,5,6,8]. The median is the average of the two middle numbers, 4 and 5, whi$\frac{4+5}{2} = 4.5$

The median is not affected by **outliers** and provides a **better measure** of central tendency for **skewed distributions**.

Measures of Central Tendency

Mode

The mode is the value that appears most frequently in a dataset.

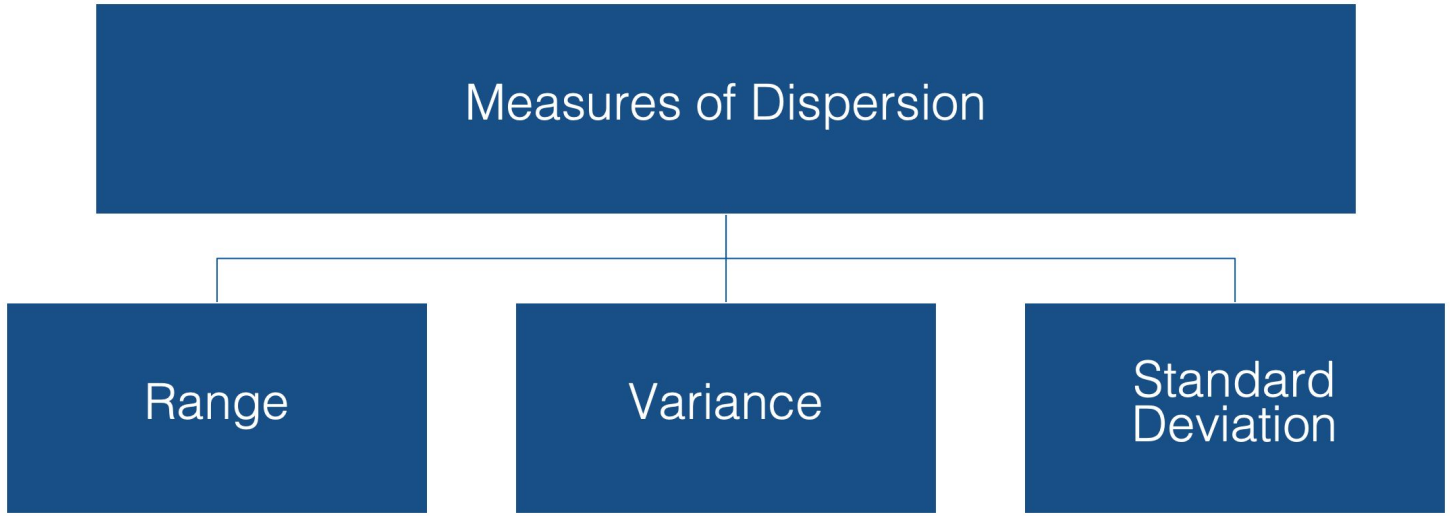
Example

In the dataset [4,8,6,5,3,4], the mode is 4 because it appears twice, more than any other number.

The mode is useful for **categorical data** and for identifying the most common item in a dataset. A dataset can have more than one mode (bimodal or multimodal) or no mode at all.

Measures of Dispersion

Learn how to describe the **spread** or **variability** of a dataset using different measures.



Measures of Dispersion

Range

The range is the difference between the **maximum** and **minimum** values in a dataset.

Formula

Range=Max value−Min valu
e.

Example

For the dataset [4,8,6,5,3,4], the range is
8−3=5

The range gives a quick sense of the spread of the data but is **highly** sensitive to outliers.

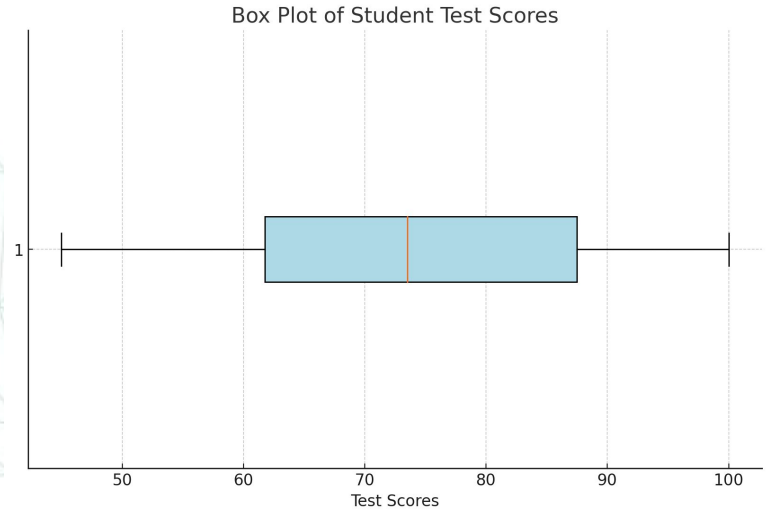
Data Distribution and Visualization

Box Plots (Box-and-Whisker Plots)

A box plot displays the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Example

Create a box plot for a dataset of test scores to identify the median score, quartiles, and potential outliers.



- Box plots are useful for detecting outliers and understanding the spread and symmetry of the data.

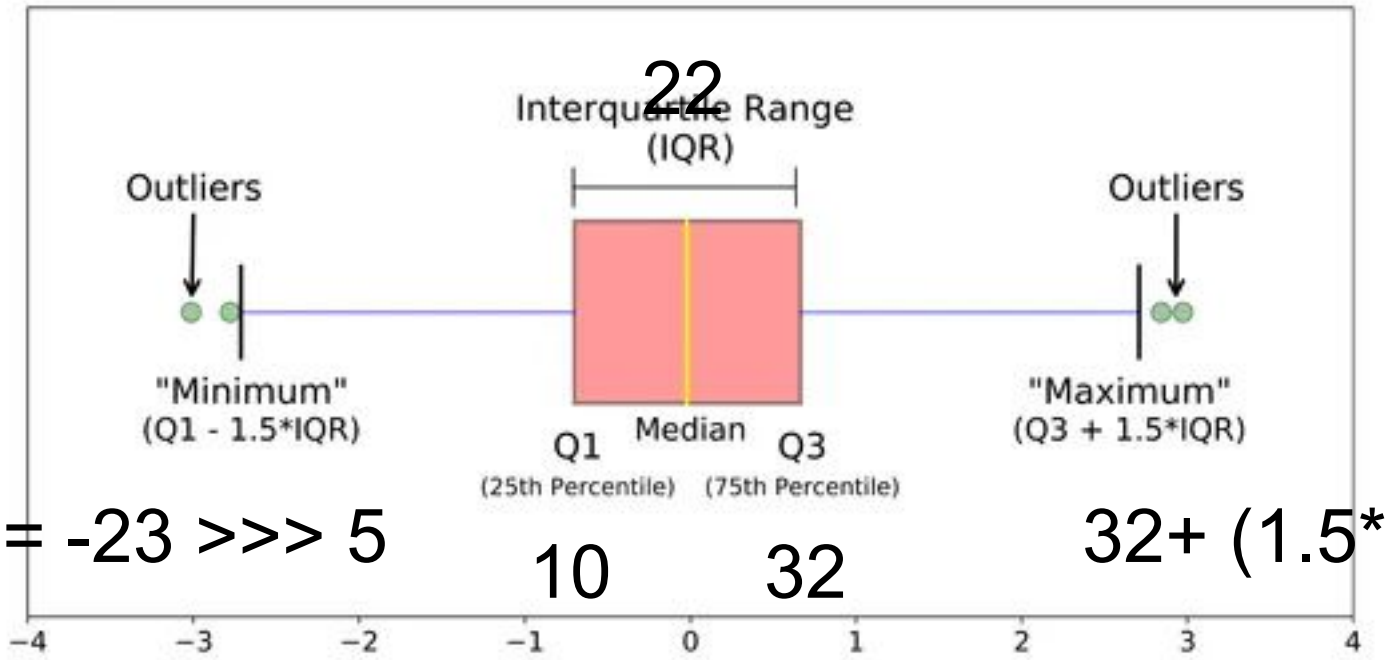
Quartiles

First Quartile (Q1, 25th percentile): The left edge of the box represents the 25th percentile, indicating that 25% of the students scored below this value.

Third Quartile (Q3, 75th percentile): The right edge of the box represents the 75th percentile, showing that 75% of the students scored below this value.

Box Plot

When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.



10- (1.5*22) = -23 >>> 5

10

32

32+ (1.5*22) = 65

Measures of Dispersion

Interquartile Range (IQR)

The IQR is the range between the first quartile (Q1) and the third quartile (Q3). It measures the spread of the middle 50% of the data, effectively capturing the central tendency without being influenced by extreme values or outliers. The IQR is calculated as

$$\text{IQR} = Q3 - Q1$$

- **Q1 (First Quartile):** The value below which 25% of the data falls.
- **Q3 (Third Quartile):** The value below which 75% of the data falls.

Interquartile Range (IQR)

Why Use IQR?

- **Robustness Against Outliers**: Unlike the range, which considers only the minimum and maximum values, the IQR focuses on the middle 50% of the data. This makes it less sensitive to outliers and extreme values, providing a more reliable measure of dispersion for skewed distributions.

- **Understanding Data Spread**: The IQR helps understand the variability of the central portion of the data. A larger IQR indicates more spread in the middle 50% of the dataset, while a smaller IQR suggests that the data points are closer to the median.

Interquartile Range (IQR)

Example Calculation of IQR

Let's use a small dataset to illustrate:
Dataset: [2,4,4,5,6,8,9]

1. **Arrange the data in ascending order** (already sorted in this case).
2. **Find Q1 and Q3:**
   - **Q1 (First Quartile):** The median of the first half of the dataset (excluding the median if the number of data points is odd). For [2,4,4], Q1 = 4
   - **Q3 (Third Quartile):** The median of the second half of the dataset. For [6,8,9], Q3 = 8
3. **Calculate IQR**: $\mathrm{IQR} = Q3 - Q1 = 8 - 4 = 4$

Measures of Dispersion

Variance

Variance measures the **average squared deviation** of each number from the **mean**. It provides insight into the spread of all data points around the mean.

Formula

For a dataset with values $x_1, x_2, \ldots, x_n$, the variance σ^2 is calculated as $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$

Example

For the dataset [4,8,6,5,3,4], calculate the mean (μ=5), then compute the variance:

$$\sigma^2 = \frac{(4-5)^2 + (8-5)^2 + (6-5)^2 + (5-5)^2 + (3-5)^2 + (4-5)^2}{6} = \frac{1+9+1+0+4+1}{6} = 2.67$$

Variance is in squared units, making it less interpretable in the original scale. However, it is fundamental in statistical theory.

Measures of Dispersion

Standard Deviation

The standard deviation is the **square root** of the **variance** and provides a measure of the **average distance** from the mean. It is in the same units as the data, making it more **interpretable**.

Formula: $\sigma = \sqrt{\sigma^2}$

Example

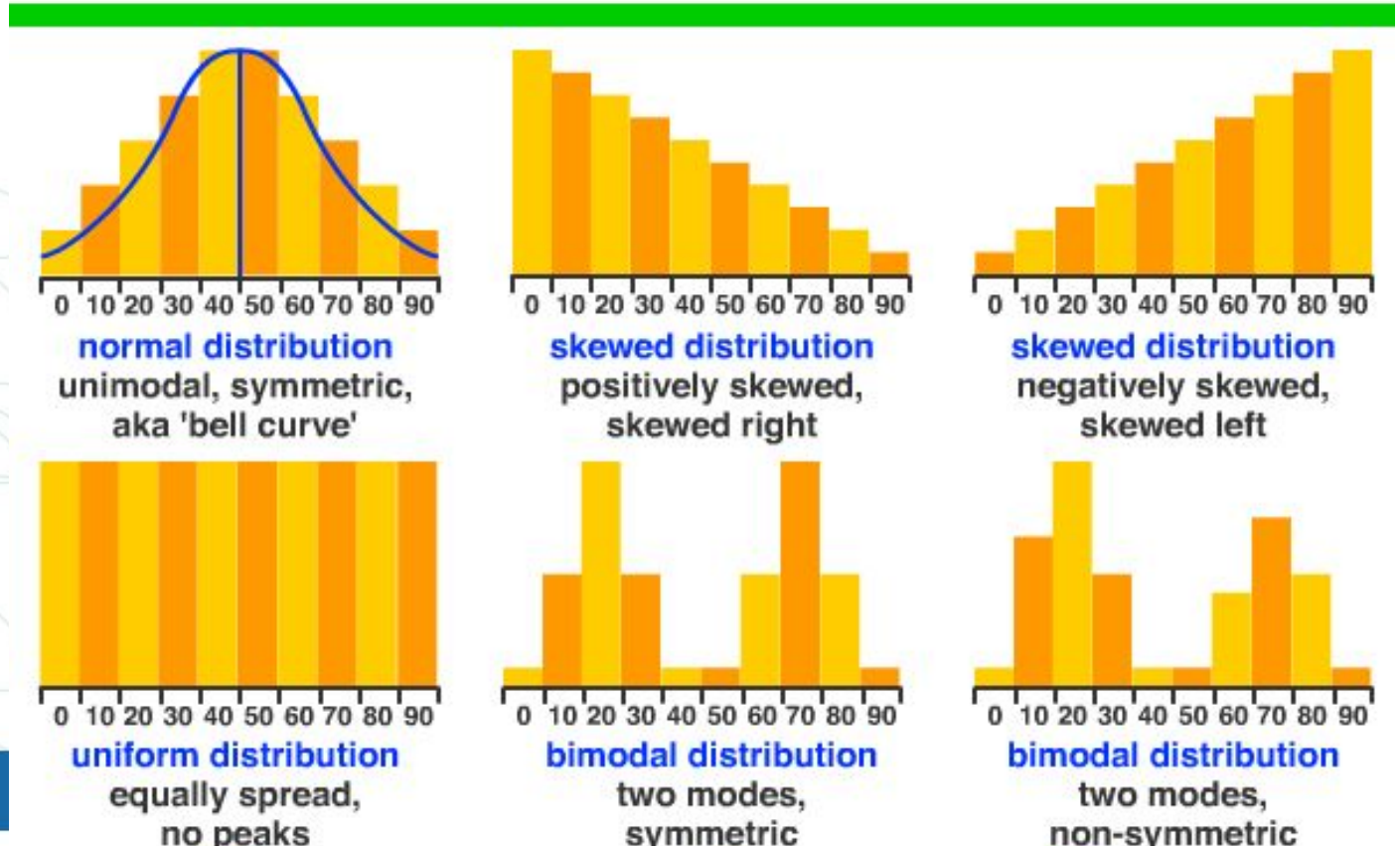
For the dataset [4,8,6,5,3,4], the standard deviation i $\sigma = \sqrt{2.67} \approx 1.63$

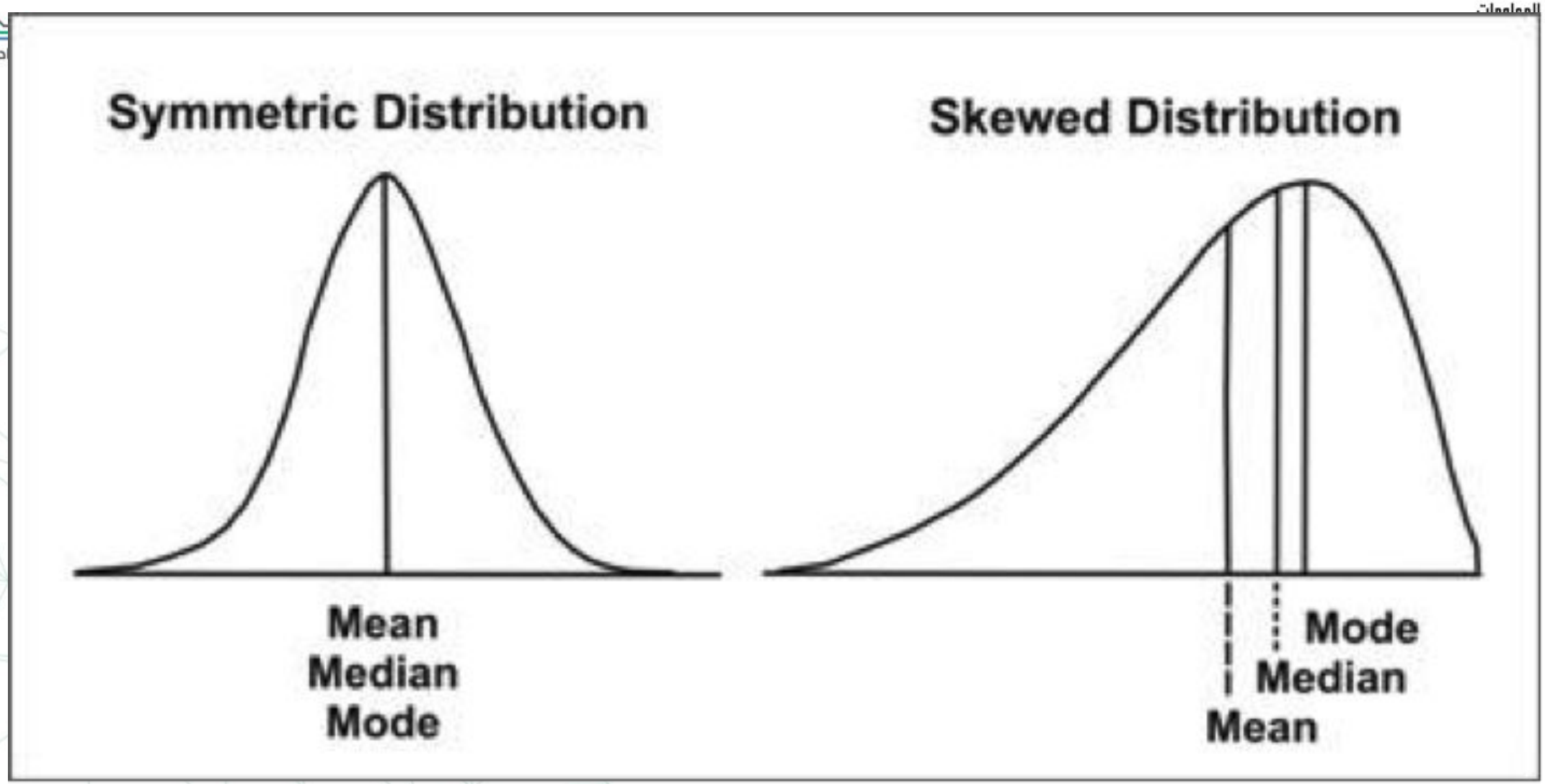
A small standard deviation indicates that the values are close to the mean, while a large standard deviation indicates that the values are spread out over a wider range.

Descriptive Statistics

Data Distribution and Visualization

Learn how to visualize data distribution and understand its shape using graphical representations.




Symmetric Distribution
Mean
Median
Mode
Skewed Distribution
Mode
Median
Mean

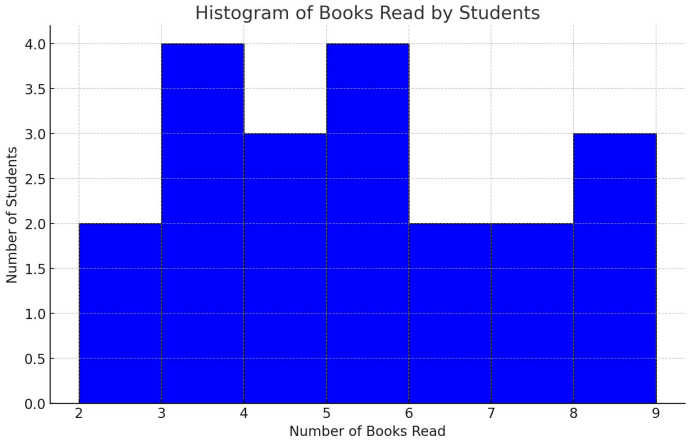
Data Distribution and Visualization

Histograms

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. It shows the frequency distribution of a dataset.

Example

Create a histogram for a dataset representing the number of books read by students in a class.



- Histograms help visualize the shape of the data distribution (e.g., normal, skewed) and identify patterns like bimodality or skewness.

Probability Basics (1 hour)

Probability Basics

fundamental probability concepts, including the rules of probability, conditional probability, and Bayes' theorem. These concepts are widely used in AI and machine learning, particularly in classification problems, predictive modeling, and decision-making processes.

Basic Probability Concepts

Definition of Probability

Probability measures the likelihood of an event occurring. It ranges from 0 to 1, where 0 indicates an impossible event and 1 indicates a certain event.

Example

The probability of getting heads in a fair coin toss is 0.5.

Random Events and Outcomes

Random Event

- An event whose occurrence cannot be predicted with certainty. Examples include rolling a die, flipping a coin, or picking a card from a deck.

Outcomes

- The possible results of a random event. For a coin toss, the possible outcomes are heads or tails. For rolling a six-sided die, the outcomes are 1, 2, 3, 4, 5, and 6.

Basic Probability Concepts

Probability Formula

For an event A, the probability P(A) is calculated as:

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

Example

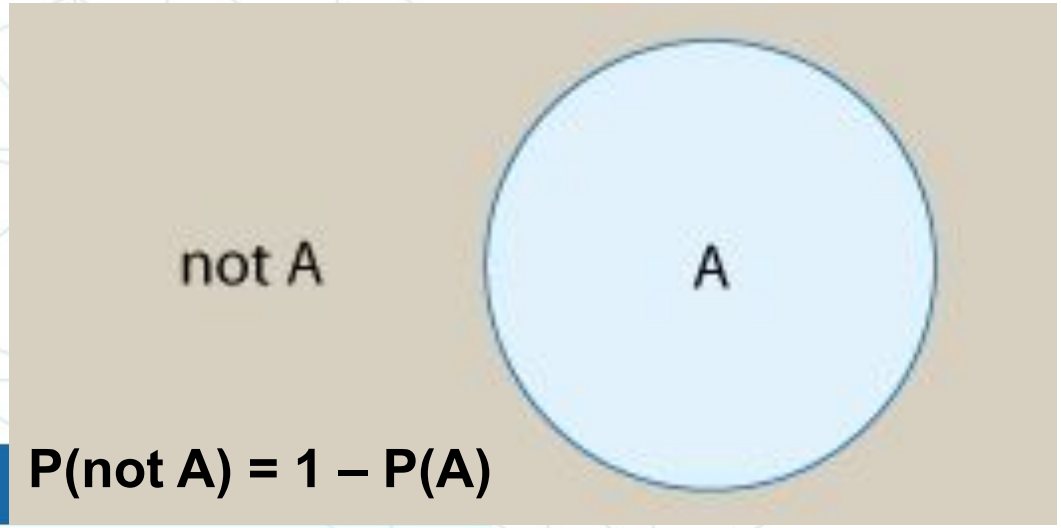
The probability of drawing a heart from a standard deck of cards:

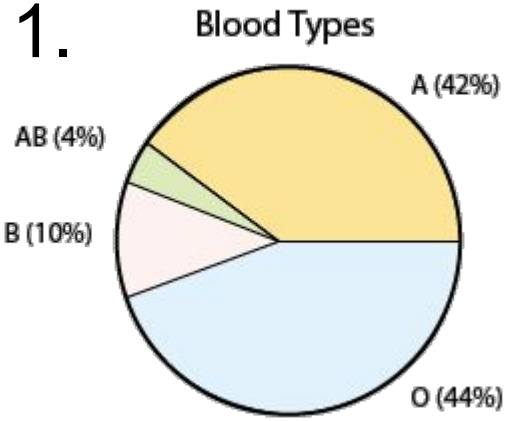
$$P(\text{heart}) = \frac{13}{52} = \frac{1}{4} = 0.25$$

Elementary Probability

A probability gives the likelihood that a defined event will occur. It is quantified as a positive number between 0 (the event is impossible) and 1 (the event is certain). Thus, the higher the probability of a given event, the more likely it is to occur.

- For any event A, $0 \leq P(A) \leq 1$.
- The sum of the probabilities of all possible outcomes is 1.
-





Probability Rules

Addition Rule (For Mutually Exclusive Events)

If two events A and B cannot happen at the same time (i.e., they are mutually exclusive), the probability that either A or B will occur is:

$$P(A \cup B) = P(A) + P(B)$$

Example

Example of the Addition Rule for Mutually Exclusive Events:

Let's take the example of rolling a six-sided die.

- **Event A** is rolling a 2.
- **Event B** is rolling a 5.

Since these two events **cannot happen at the same time** (you can't roll both a 2 and a 5 on a single roll of the die), they are **mutually exclusive events**.

According to the **addition rule** for mutually exclusive events, the probability of rolling either a 2 **or** a 5 is the sum of the probabilities of each event:

$$P(A \text{ or } B) = P(A) + P(B)$$

The probability of rolling a 2 is:

$$P(A) = \frac{1}{6}$$

The probability of rolling a 5 is:

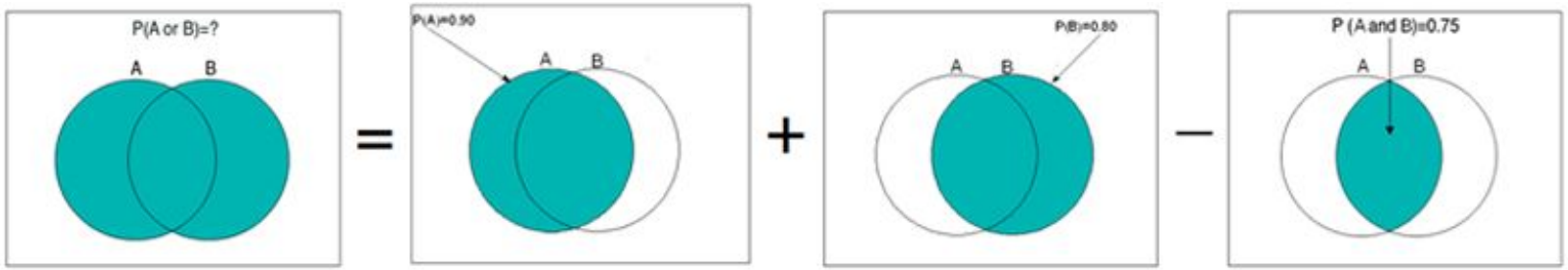
$$P(B) = \frac{1}{6}$$

So, the probability of rolling either a 2 or a 5 is:

$$P(A \text{ or } B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Thus, the probability of rolling a 2 or a 5 on a six-sided die is **1/3** or approximately **0.333**.

Addition Rule



We therefore get:

P(A or B) = P(A) + P(B) − P(A and B)

addition rules (union of events)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

mutually exclusive $P(A \cup B) = P(A) + P(B)$

Probability Rules

Multiplication Rule (For Independent Events)

If two events A and B are independent (the occurrence of one does not affect the occurrence of the other), the probability that both A and B will occur is:

$$P(A \cap B) = P(A) \times P(B)$$

Example

The probability of getting heads when flipping a coin and rolling a 4 on a six-sided die:

$$P(\text{heads and } 4) = P(\text{heads}) \times P(4) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

multiplication rules (joint probability)

dependent $P(A \cap B) = P(A) * P(B|A)$

independent $P(A \cap B) = P(A) * P(B)$

mutually exclusive $P(A \cap B) = 0$

P(B\A) "the probability of B happening *given* A has occurred

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$P(A \mid B)$: Probability of *A* given *B*

$P(A \cap B)$: Probability of *A* and *B*

$P(B)$: Probability of *B*

A math teacher gave her class two tests. 80% of the class passed the first test. 60% of the class passed both tests. What percent of those who passed the first test also passed the second test?

P(second given first) = P(first and second) / P(first) = 0.6/0.8 = 75%

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

Probability of B

A math teacher gave her class two tests. 80% of the class passed the first test. 60% of the class passed both tests. What percent of those who passed the first test also passed the second test?

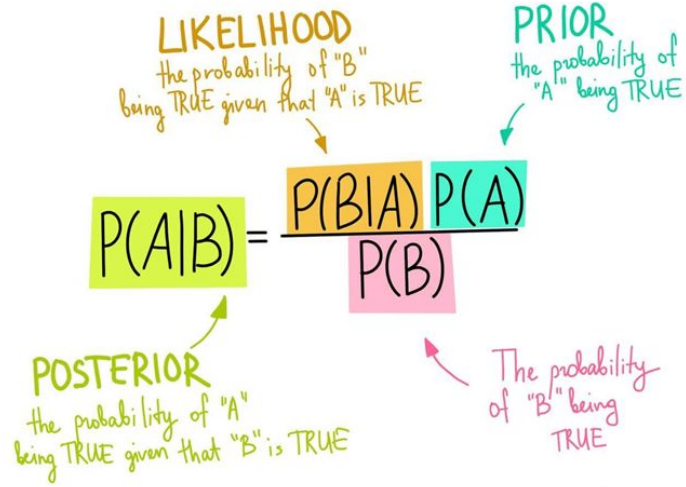
P(second given first) = P(first and second) / P(first) = 0.6/0.8 = 75%

Bayes' Theorem

Introduction to Bayes' Theorem: Bayes' theorem allows us to update our probability estimates based on new **evidence** or **information**. It is particularly useful in AI for making predictions based on **prior knowledge** and new data.

Bayes' Theorem Formula:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$



Where:

- $P(A|B)$ is the posterior probability: the probability of event A given B has occurr
- $P(B|A)$ is the likelihood: the probability of event B given A is true.
- $P(A)$ is the prior probability: the probability of event A before observing B.
- $P(B)$ is the marginal likelihood: the total probability of event B.

Bayes' Theorem

Example

Given Information:

- The probability that a person has the disease is 1% of the population:

  $$P(\text{Disease}) = 0.01$$

- The test correctly identifies the disease 99% of the time (test sensitivity):

  $$P(\text{Positive Test} \mid \text{Disease}) = 0.99$$

- The test produces false positives 5% of the time (false positive rate):

  $$P(\text{Positive Test} \mid \text{No Disease}) = 0.05$$

- The probability that a person does not have the disease is 99%:

  $$P(\text{No Disease}) = 0.99$$

Problem:

We want to calculate the probability that a person has the disease given that they tested positive, i.e., $P(\text{Disease} \mid \text{Positive Test})$.

Bayes' Theorem

Example

Using Bayes' Theorem:

Bayes' Theorem gives us the following formula:

$$P(\text{Disease} \mid \text{Positive Test}) = \frac{P(\text{Positive Test} \mid \text{Disease}) \times P(\text{Disease})}{P(\text{Positive Test})}$$

Calculating $P(\text{Positive Test})$:

$$P(\text{Positive Test}) = P(\text{Positive Test} \mid \text{Disease}) \times P(\text{Disease}) + P(\text{Positive Test} \mid \text{No Disease}) \times P(\text{No Disease})$$

Substituting the values:

$$P(\text{Positive Test}) = (0.99 \times 0.01) + (0.05 \times 0.99) = 0.0099 + 0.0495 = 0.0594$$

Apply Bayes' Theorem: $$P(\text{Disease}|\text{Positive Test}) = \frac{0.99 \times 0.01}{0.0594} \approx 0.1667$$

This means there is approximately a 16.67% chance that the person actually has the disease, despite testing positive. This highlights how Bayes' theorem can provide more accurate probability estimates when additional information is available.

Introduction to Inferential Statistics

Introduction to Inferential Statistics

What is Inferential Statistics?

Inferential statistics involve using data from a sample to make generalizations about a larger population. **Unlike** descriptive statistics, which summarize data, inferential statistics allow us to draw conclusions and make predictions.

Purpose in AI and Data Science

- Helps estimate population parameters (e.g., the mean, proportion).
- Allows for hypothesis testing to determine if observed patterns are likely to be genuine or if they occurred by chance.
- Facilitates predictions and decision-making based on sample data.

Key Concepts of Inferential Statistics

Sampling Distributions

A sampling distribution is the probability distribution of a given statistic (e.g., mean, proportion) based on a random sample.

Importance

- Sampling distributions help understand how the sample statistic (like the sample mean) varies from sample to sample.
- They form the basis for making inferences about the population.

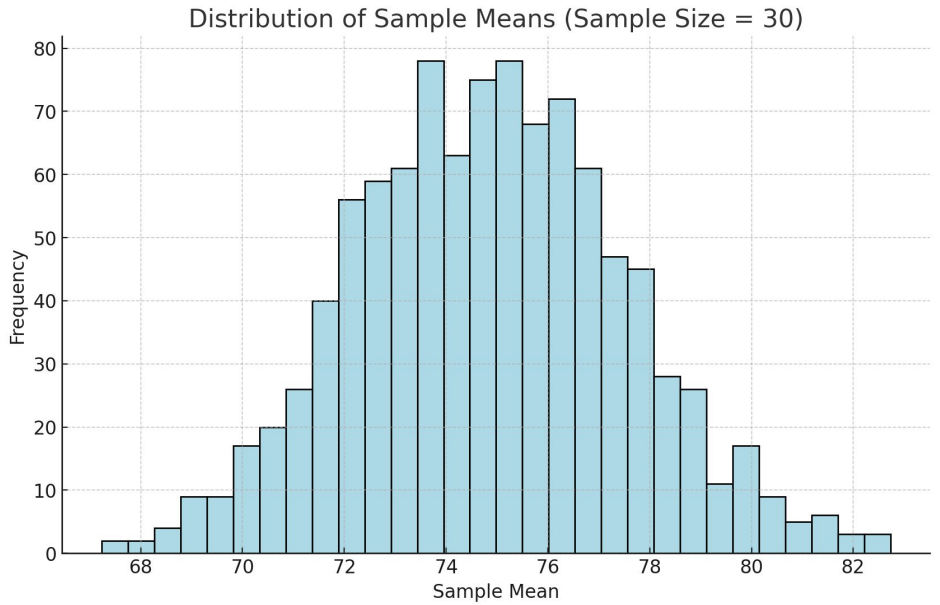
Example

If we take multiple random samples of students' test scores and calculate the mean for each sample, the distribution of those sample means would form a sampling distribution. This distribution tends to be normal, especially as the sample size increases, according to the Central Limit Theorem.

Key Concepts of Inferential Statistics

Example

If we take multiple random samples of students' test scores and calculate the mean for each sample, the distribution of those sample means would form a sampling distribution. This distribution tends to be normal, especially as the sample size increases, according to the Central Limit Theorem.



Key Concepts of Inferential Statistics

Central Limit Theorem (CLT)

The CLT states that the sampling distribution of the sample mean will be approximately normally distributed, regardless of the population's distribution, provided the sample size is sufficiently large (usually $n>30$).

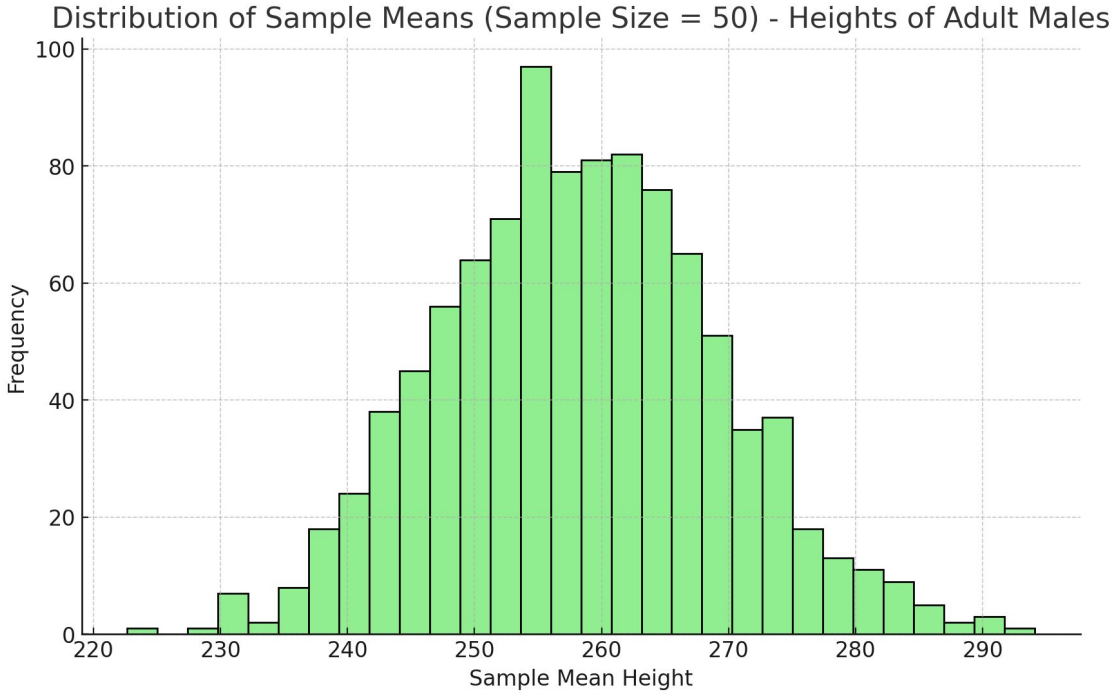
Significance

The CLT enables us to use the normal distribution to make inferences about population parameters even when the population distribution is not normal.

Key Concepts of Inferential Statistics

Example

Suppose we are analyzing the average height of adult males. Even if the population distribution of heights is skewed, the mean of sufficiently large samples will follow a normal distribution.



Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. It helps us decide whether there is enough evidence to reject a null hypothesis.

Key Terms

- **Null Hypothesis (H0):** A statement that there is no effect or no difference. It serves as the default assumption. For example, "The mean score of students is equal to 75.“
- **Alternative Hypothesis (H1 or Ha):** The statement that contradicts the null hypothesis. For example, "The mean score of students is not equal to 75.“
- **P-value**: The probability of observing the sample data, or something more extreme, assuming the null hypothesis is true. A low p-value (typically < 0.05) suggests that the observed data is unlikely under the null hypothesis, leading to its rejection.
- **Significance Level (α):** A threshold chosen before conducting the test, often set at 0.05, indicating a 5% risk of concluding that a difference exists when there is no actual difference.

Hypothesis Testing

Steps in Hypothesis Testing

1. **State the Hypotheses**: Define the null and alternative hypotheses.
2. **Select a Significance Level (α)**: Commonly set at 0.05.
3. **Collect Data and Calculate a Test Statistic**: Based on the sample data, calculate a statistic (e.g., t-statistic, z-statistic) that measures the difference.
4. **Find the P-value**: Compare the test statistic to the theoretical distribution to find the p-value.
5. **Make a Decision**:
   - If the p-value < α, reject the null hypothesis (evidence suggests an effect).
   - If the p-value ≥ α, do not reject the null hypothesis (insufficient evidence to suggest an effect).

Hypothesis Testing

Example

A company claims that their employees work an average of 8 hours per day. A random sample of 30 employees shows an average of 7.5 hours with a standard deviation of 1 hour. We could conduct a hypothesis test to determine if this sample provides enough evidence to reject the company's claim.

Confidence Intervals

A confidence interval provides a range of values that is likely to contain the population parameter (e.g., mean) with a certain level of confidence (e.g., 95%).

Formula : $$\text{Confidence Interval} = \bar{x} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

- $\bar{x}$ = Sample mean
- Z = Z-value (from standard normal distribution for a given confidence level)
- σ = Population standard deviation (or sample standard deviation if population standard deviation is unknown)
- n = Sample size

Interpretation

A 95% confidence interval means that if we took 100 different samples and computed a confidence interval for each sample, approximately 95 of the intervals would contain the true population mean.

Confidence Intervals

Example

If a sample mean score on a test is 70 with a standard deviation of 5 and sample size of 30, a 95% confidence interval for the mean score might be calculated. This interval provides a range within which we expect the true mean score for the entire population to fall.