جامعة الامير سلطان
PRINCE SULTAN UNIVERSITY

FOR SPECIALIZATION

Robotics & Internet-of-Things

Lecture 1
Introduction to Data Science

Prof. Anis Koubaa

Prince Sultan University
August 2024

Lead Instructor

Name	Prof. Anis Koubaa
Academic Title	Professor in Computer Science
Admin Titles	Director of the Research and Initiatives Center Leader of Robotics and Internet-of-Things Lab
Research Interest	Deep Learning Mobile Robots Unmanned Aerial Systems Internet-of-Things
Email	akoubaa@psu.edu.sa
Phone	0114948851
Office Hours	Per Appointment
Location	Building 101 | RIOTU Lab

The Robotics & Internet-of-Things Lab: *The Talents' Incubator*

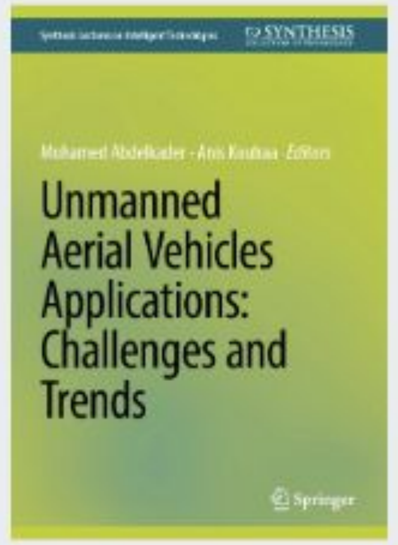
Wonderful Team

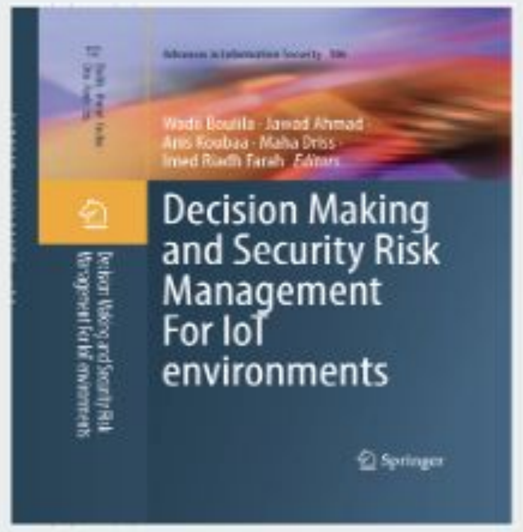
- TEAM
  - Total Members: 22
  - PhD Holders: 11
  - Research Assistants: 3
  - Postdoc: 1
- RESEARCH
  - Generative AI/AI
  - UAVs and Robotics
  - IoT and ITS

مركز الأمير سلطان للدراسات والبحوث الدفاعية
PSDSARC Prince Sultan Defense Studies & Research Center

Stanford University's Top 2% Scientists

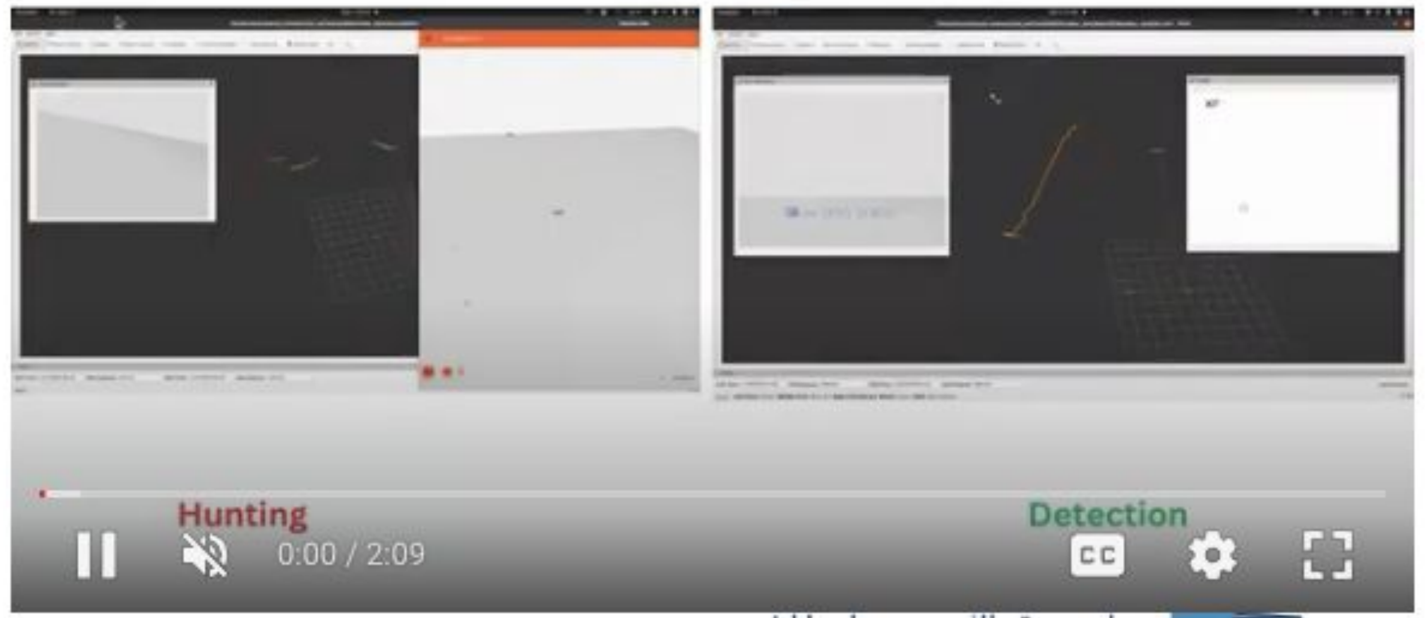
- Prof. Anis Koubaa
- Dr. Basit Qureshi
- Dr. Wadii Boulila





Robotics & Internet-of-Things
iHUNTER
جامعة الأمير سلطان PRINCE SULTAN UNIVERSITY

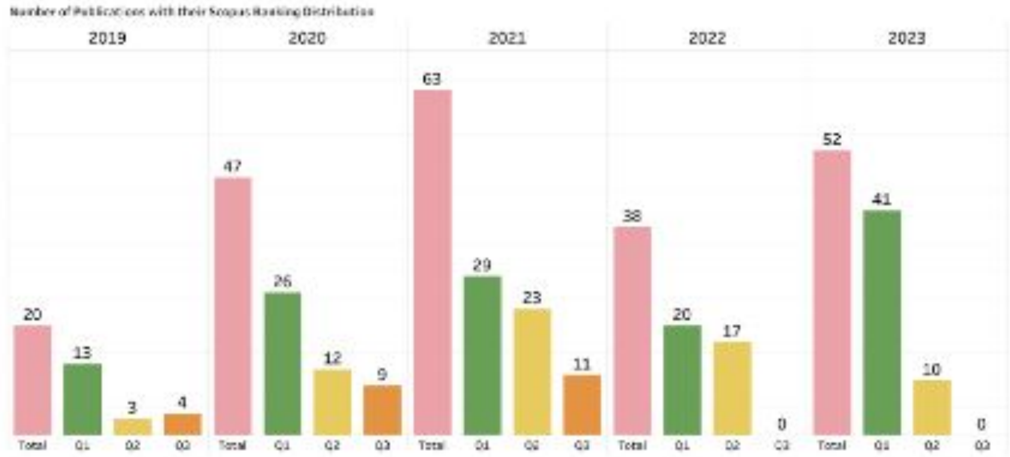
Drone Hunter

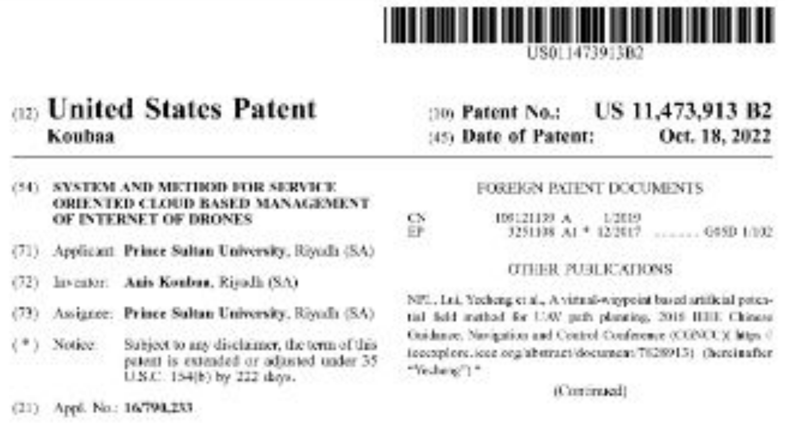


Patents

جامعة الأمير سلطان PRINCE SULTAN UNIVERSITY

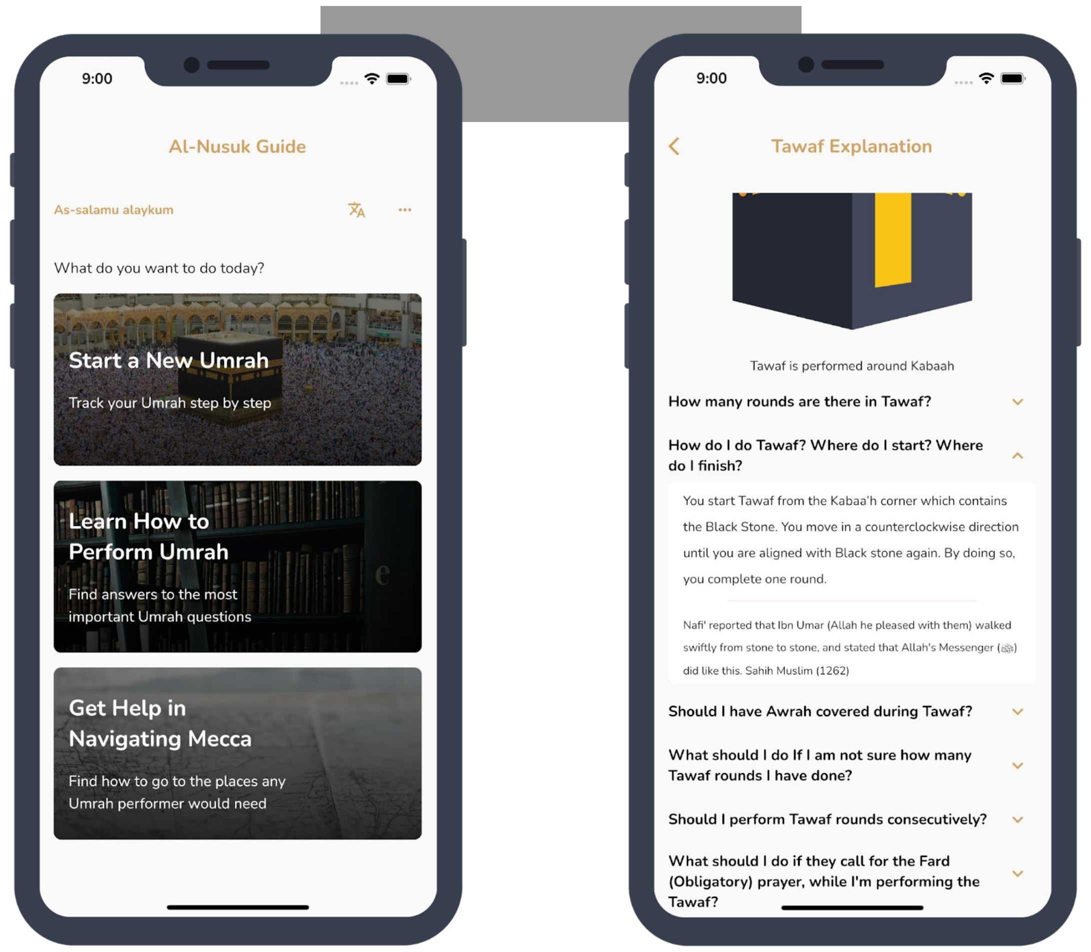
2. *FASEEH*

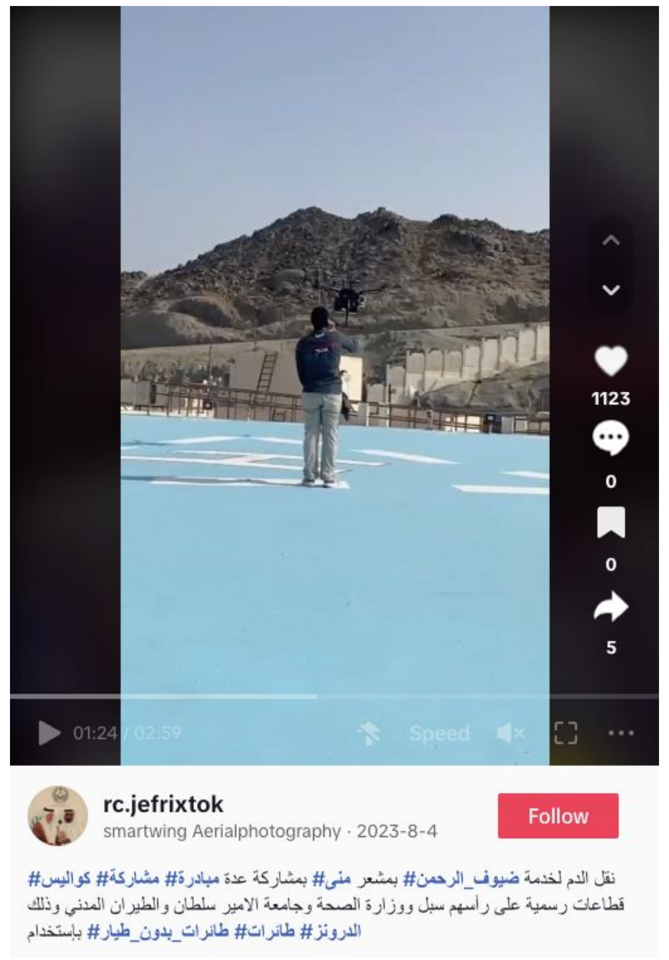


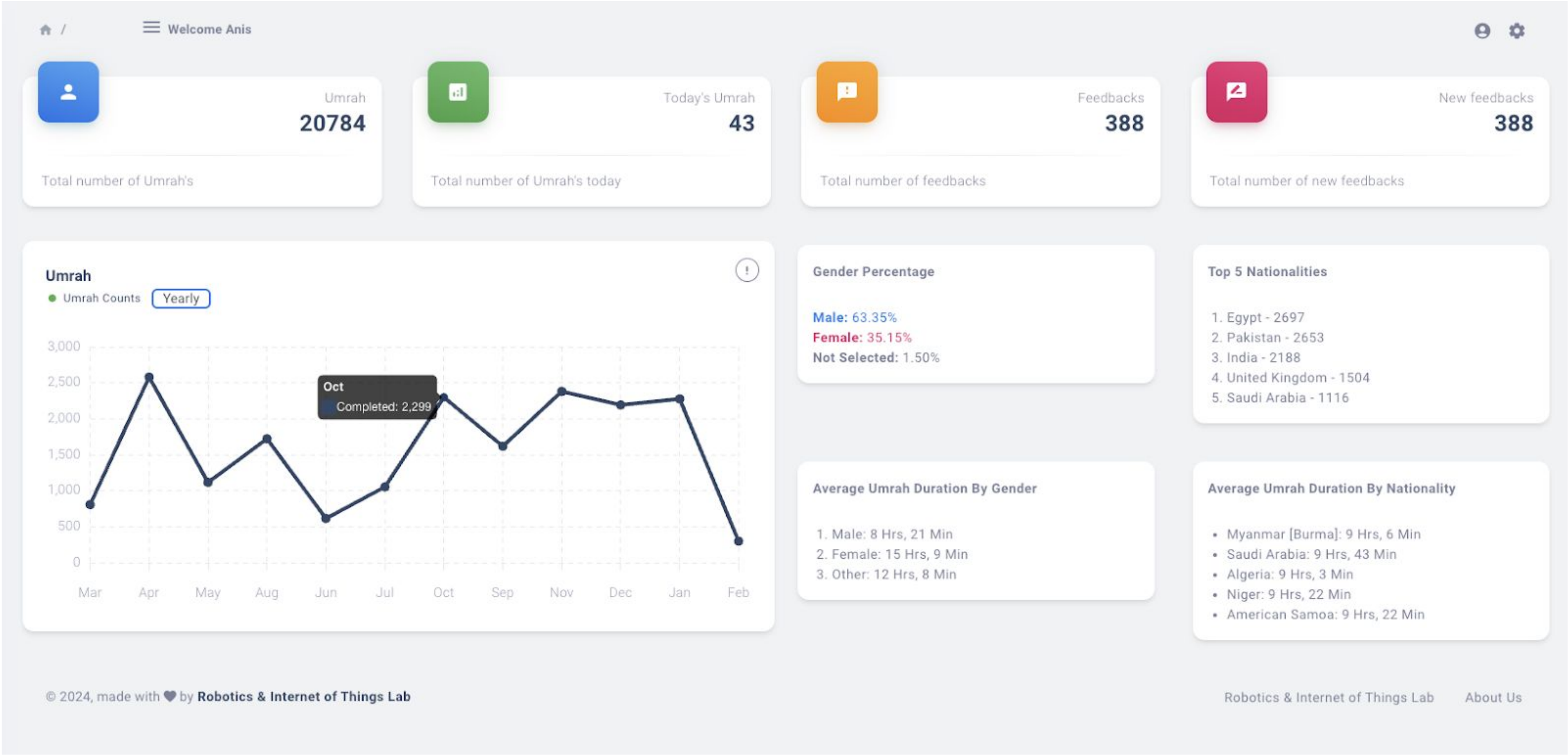

United States Patent
Koubaa
Patent No.: US 11,473,913 B2
Date of Patent: Oct. 18, 2022
(54) SYSTEM AND METHOD FOR SERVICE ORIENTED CLOUD BASED MANAGEMENT OF INTERNET OF DRONES
(71) Applicant: Prince Sultan University, Riyadh (SA)
(72) Inventor: Anis Koubaa, Riyadh (SA)
(73) Assignee: Prince Sultan University, Riyadh (SA)
(21) Appl. No.: 16/790,233
FOREIGN PATENT DOCUMENTS
OTHER PUBLICATIONS

Impact Beyond Academia

خدمة ضيوف الرحمن







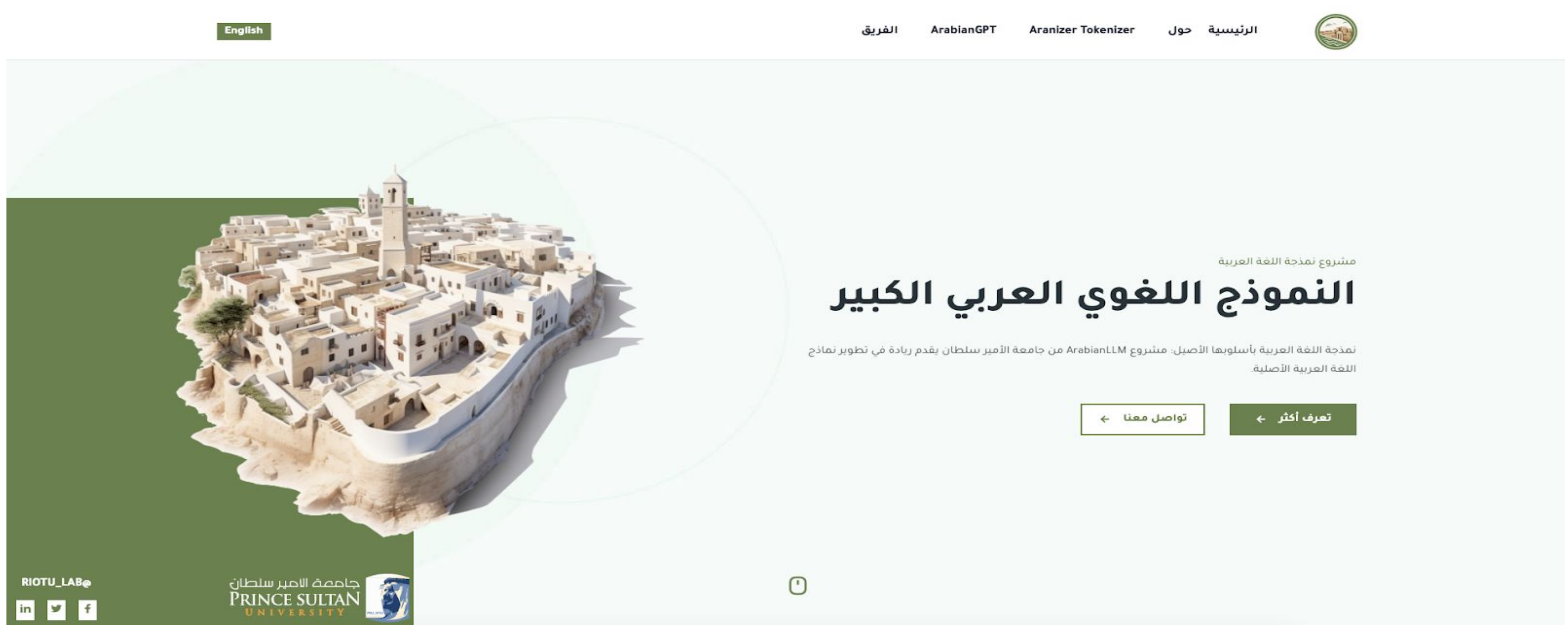
https://nusuk-guide.net/

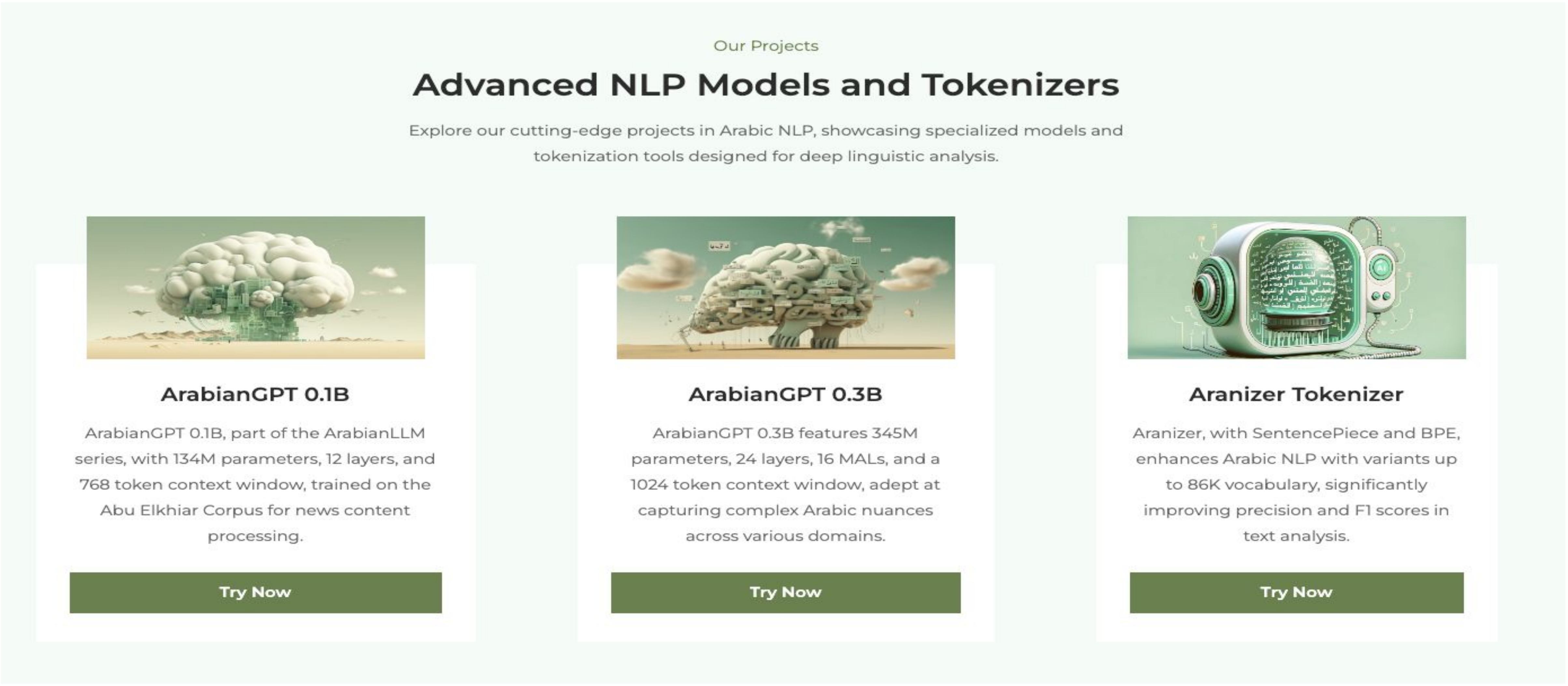
https://psugpt.ai/

https://tibyan-ai.com/

https://llm.riotu-lab.org/

A100 GPU Server





Course Overview

FOR

SPECIALIZATION

Lecture 1

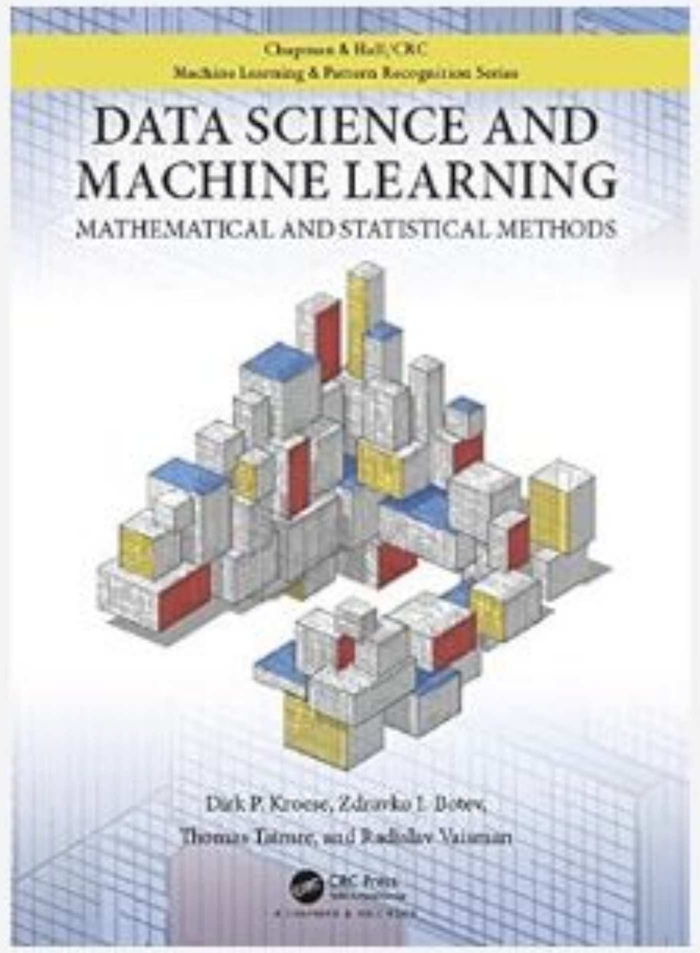
Introduction to Data Science

Prof. Anis Koubaa

Course Learning Outcome

- **CLO1. Apply** the fundamentals of Python programming for AI and data science and visualization.
- **CLO2. Demonstrate** a thorough knowledge of AI, Statistical Learning, and fundamental Machine Learning models to build supervised and unsupervised predictive models.
- **CLO3. Develop** predictive models using convex optimization techniques and evaluate their performance.
- **CLO4. Execute** a team capstone project applying theoretical knowledge to solve a significant AI and Data Science problem.
- **CLO5. Reflect** on the ethical implications, safety, societal impact, and professional responsibilities involved in AI and Data Science.

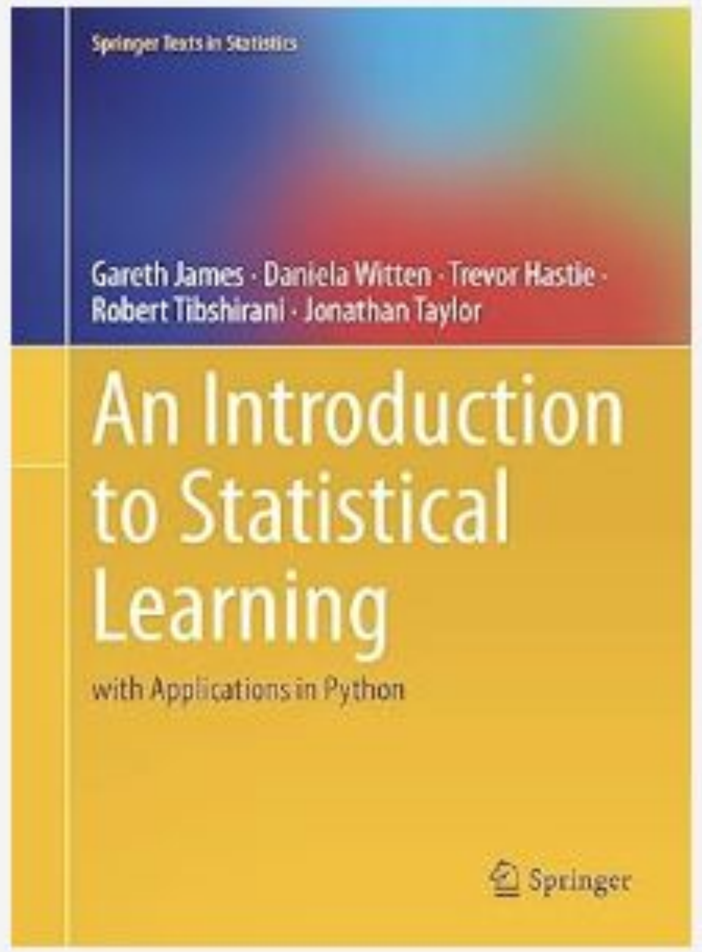
Textbooks



Data Science and Machine Learning: Mathematical and Statistical Methods 1st Edition

Authors: Dirk P. Kroese, Zdravko Botev, Thomas Taimre and Radislav Vaisman

https://github.com/DSML-book/



An Introduction to Statistical Learning: with Applications in Python 1st Edition

Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor

https://www.statlearning.com/online-courses

Textbooks

CS316 Introduction to AI and Data Science

Chapter 1

First Edition

Anis Koubaa
Riyadh, Saudi Arabia

Table of Contents

Chapter 1

Introduction to Data Science

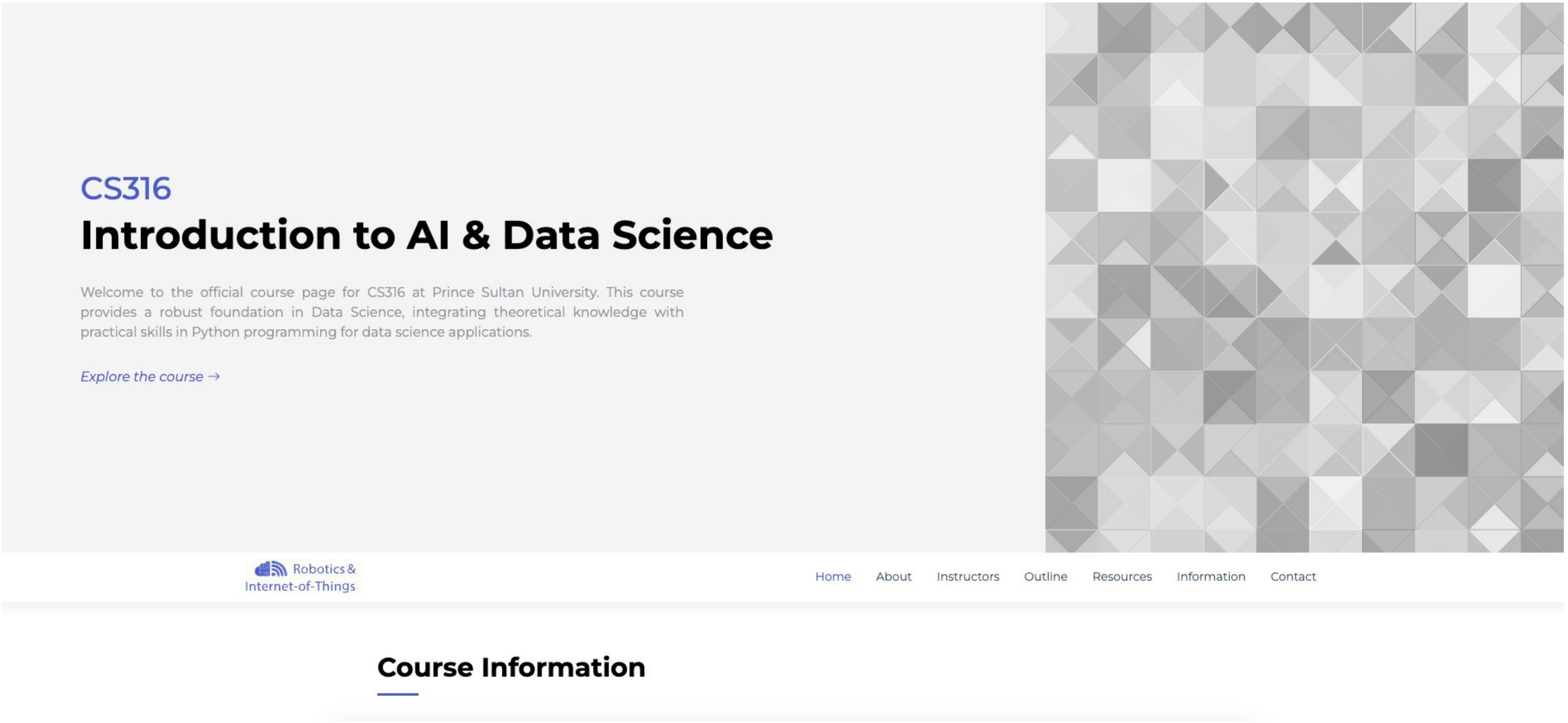
1.1 Introduction

In the current digital era, data science emerges as a crucial driver for innovation and operational efficiency, paralleling the significant roles oil played during the Industrial Revolution and electricity in the 20th century. This comparison, frequently cited by scholar Andrew Ng, serves to underline the extensive influence of data science across various industries. It emphasizes the significant role that skilled data analysis and application play in enhancing outcomes and addressing complex challenges across diverse fields.

Figure 1.1: Data is the New Oil

Data science utilizes extensive datasets to empower organizations to make

Course Website

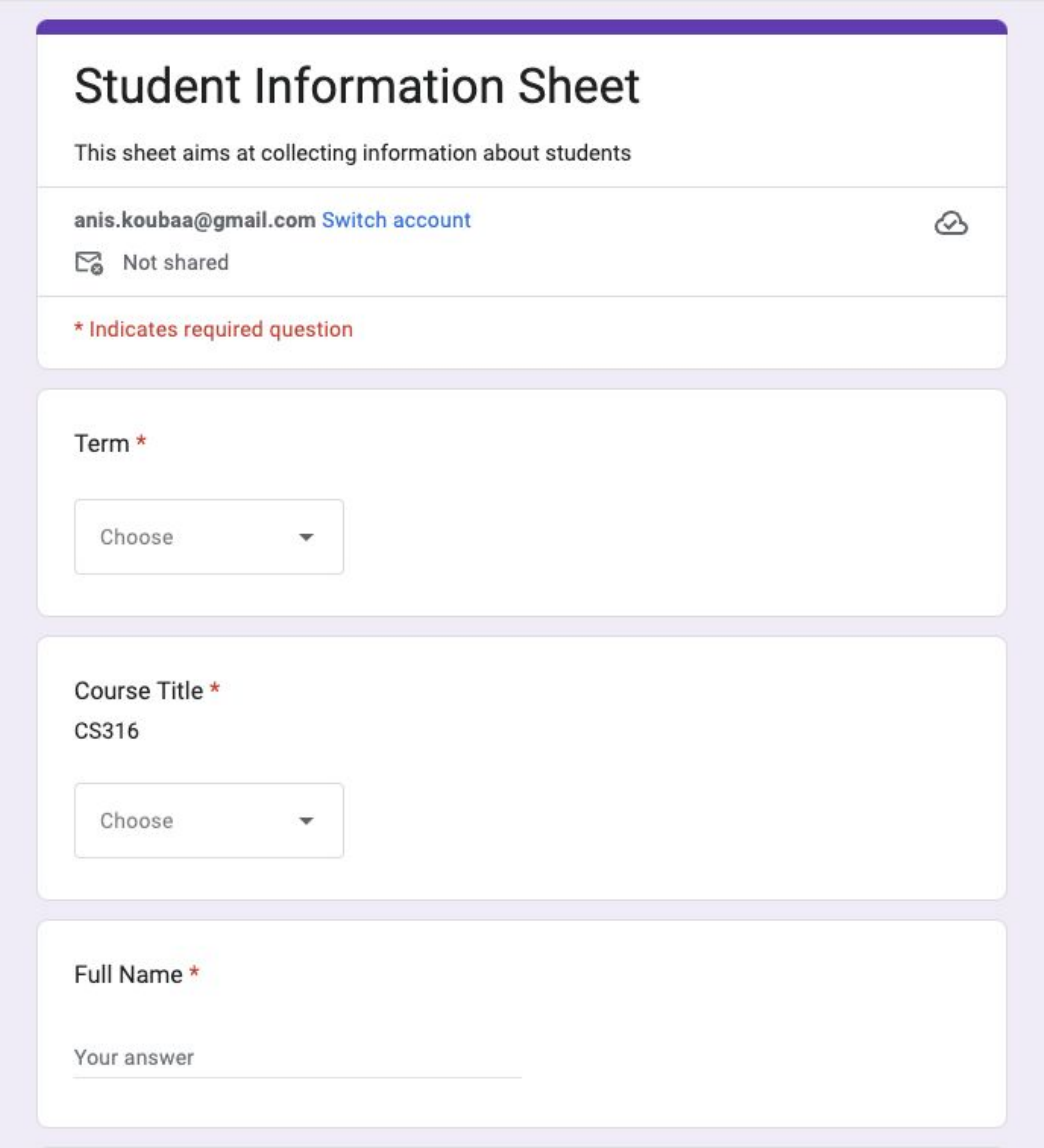


https://ds.riotu-lab.org/

ExamGPT

ExamGPT
Revolutionizing Exam Grading with Generative AI
Learn more
Log in
Why ExamGPT?
ExamGPT Capabilities & Features
AI
Powered by Generative AI
Experience the accuracy and efficiency of grading brought by Generative AI models. Say goodbye to manual grading and embrace automation.
Integrated with ChatGPT & OpenAI LLM
Leverage the power of ChatGPT and OpenAI's state-of-the-art language models for precise and fair grading.
Easy Integration
Integrate ExamGPT seamlessly into your academic workflow. Its user-friendly interface ensures minimal onboarding time for instructors.

Student Information Form



Student Information Sheet

This sheet aims at collecting information about students

anis.koubaa@gmail.com Switch account

Not shared

* Indicates required question

Term *

Choose

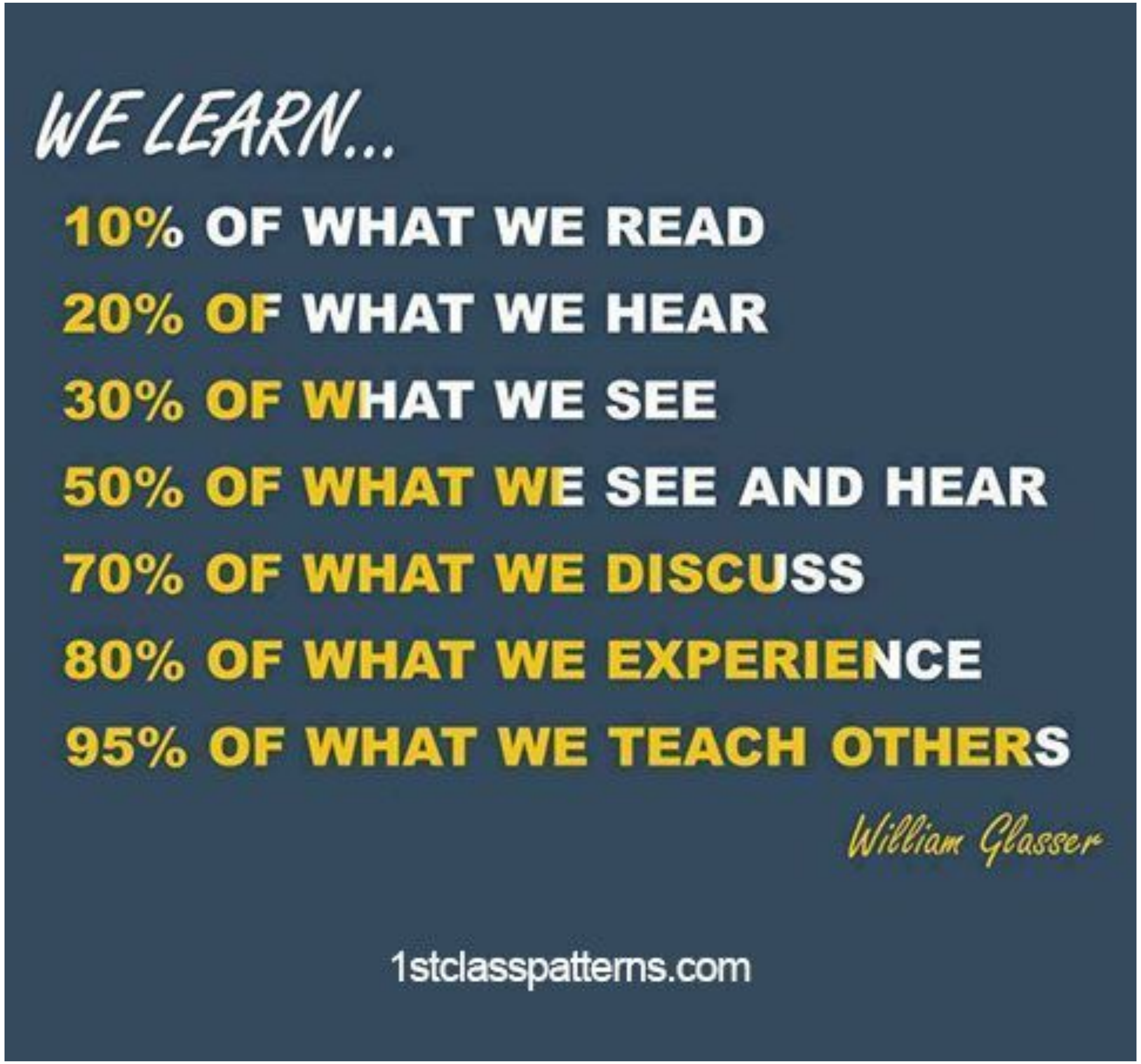
Course Title *

CS316

Choose

Full Name *

Your answer


WE LEARN...
10% OF WHAT WE READ
20% OF WHAT WE HEAR
30% OF WHAT WE SEE
50% OF WHAT WE SEE AND HEAR
70% OF WHAT WE DISCUSS
80% OF WHAT WE EXPERIENCE
95% OF WHAT WE TEACH OTHERS
William Glasser
1stclasspatterns.com

Lecture 1
Introduction to Data Science

Prof. Anis Koubaa

Prince Sultan University
August 2024

What is Data Science?

FOR

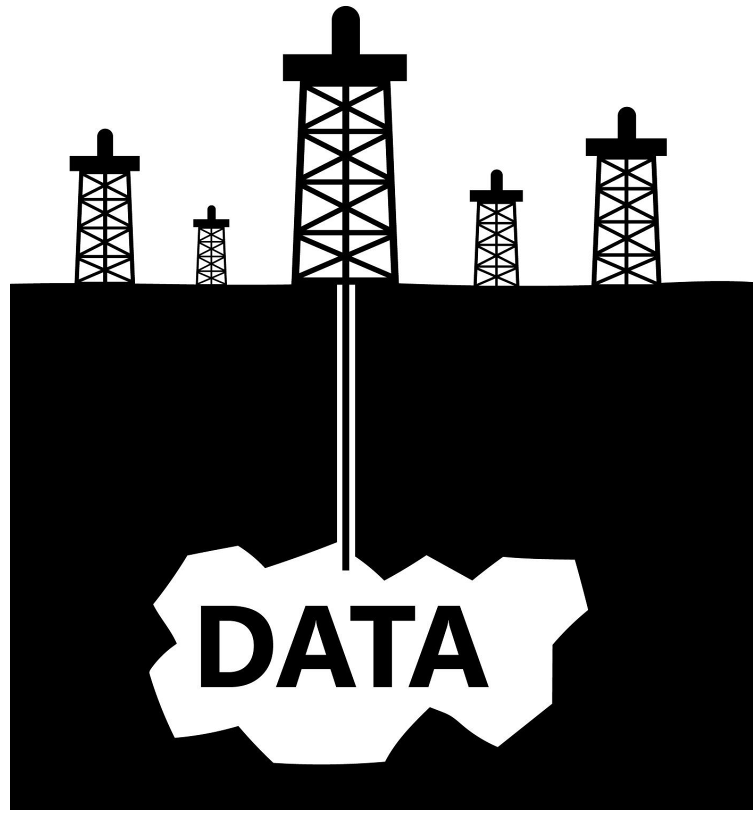
SPECIALIZATION

Lecture 1

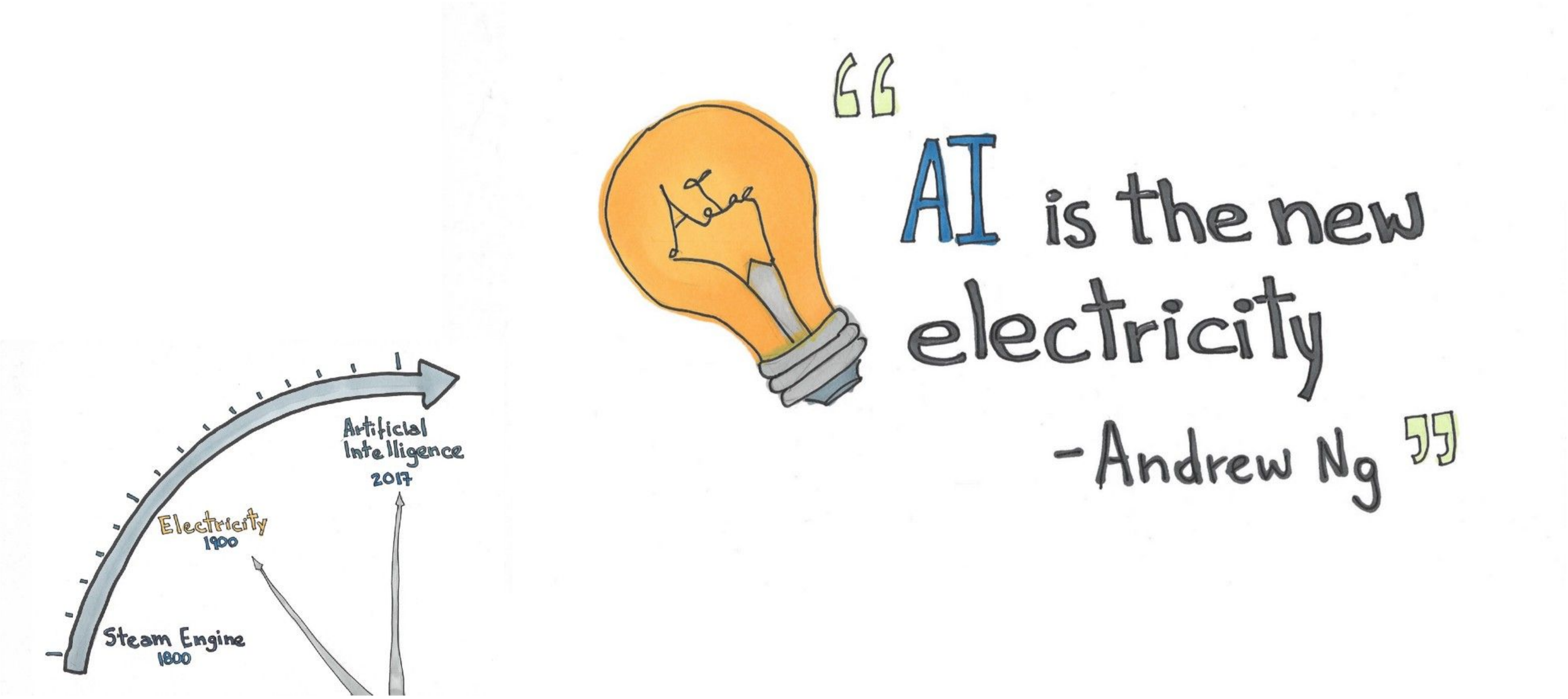
Introduction to Data Science

Prof. Anis Koubaa

DATA IS THE NEW OIL

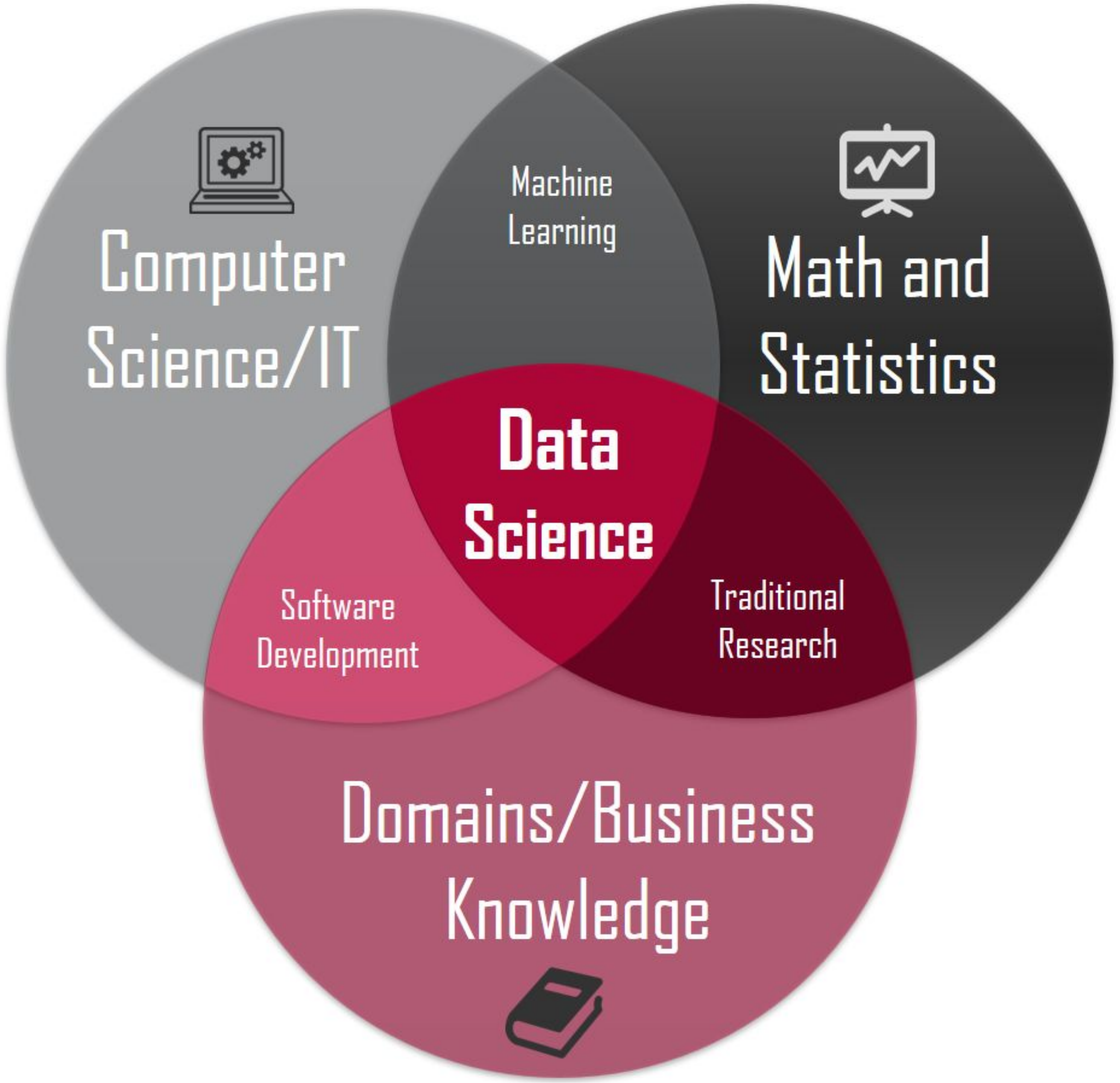


AI IS THE NEW ELECTRICITY



What is Data Science

- **Definition:** Data Science is an **interdisciplinary** field that uses scientific methods, algorithms, and systems to extract knowledge and insights from **structured** and **unstructured data**.
- **Core Components:**
  - Statistics,
  - Machine Learning,
  - Data Engineering,
  - Domain Expertise, and
  - Data Visualization.



https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

DATA SCIENCE TOOLS

Tools you will need in data science

MATH
Probability
Statistics
Linear Algebra
Calculus

PROGRAMMING
Python
Matplotlib
Sipy
Pandas
Numpy

MACHINE LEARNING
Regression
Classification
Clustering

DEEP LEARNING
Neural Networks
Segmentation
NLP

ANALYTICS
PowerBI
Tableau
Qlik Sense

DATA ANALYTICS

TYPES

DESCRIPTIVE ANALYTICS

What happened?

Use historical data to identify trends and relationships

BUSINESS INTELLIGENCE

PREDICTIVE ANALYTICS

What will happen?

use of statistical models and and machine learning techniques to predict the trends in the future

MACHINE LEARNING

PRESCRIPTIVE ANALYTICS

What should we do next?

process that analyzes data and provides instant recommendations on how to optimize business practices to suit multiple predicted

Decision Science

Data Science Process

FOR

SPECIALIZATION

Lecture 1

Introduction to Data Science

Prof. Anis Koubaa

DATA SCIENCE WORKFLOW

WHAT IS DATA SCIENCE?

01 BUSINESS UNDERSTANDING

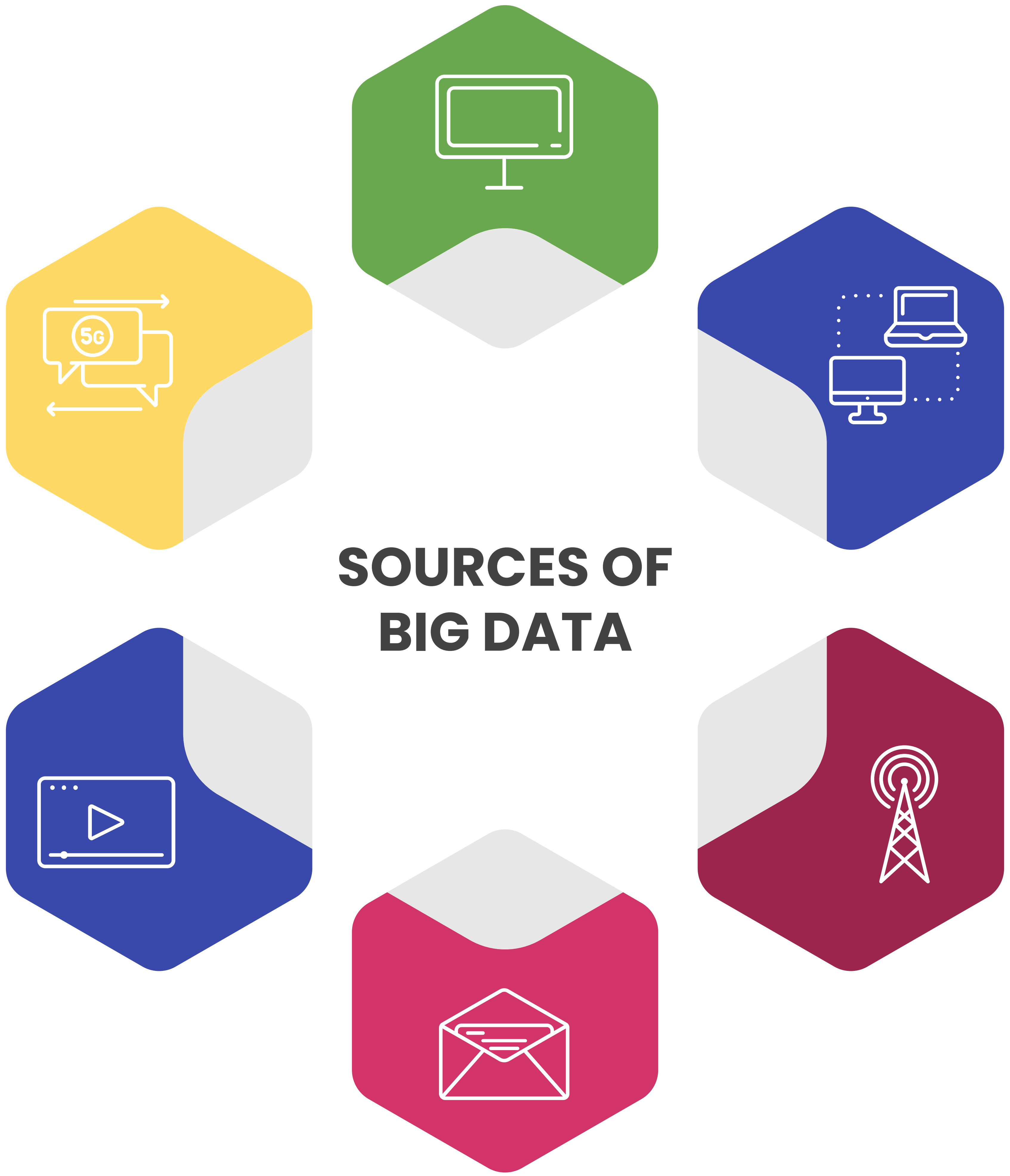
Identify the problem that must be considered in the study

02 DATA COLLECTION

Collect data that serves your study's objectives

03 DATA CLEANING

Fix the consistency in the data and handle missing values



04 FEATURE ENGINEERING

Transform your raw data into relevant and meaningful features

05 PREDICTIVE MODELS

Build Models
Train machine/deep learning models, and evaluate their performance and use them to make predictions

06 DATA VISUALIZATION

Communicate the finding with stakeholders and illustrate them with interactive visualization

Data Science Process

1. **Data Collection:** Gathering raw data from various sources (databases, APIs, sensors, etc.).
2. **Data Cleaning:** Handling missing data, outliers, and formatting issues.
3. **Data Exploration:** Analyzing data patterns, distributions, and anomalies through descriptive statistics and visualizations.
4. **Feature Engineering:** Transforming raw data into meaningful inputs for machine learning models.
5. **Model Building:** Developing predictive or inferential models using statistical learning methods.
6. **Model Evaluation:** Measuring model performance through metrics like accuracy, precision, recall, etc.
7. **Deployment & Monitoring:** Integrating the model into production and continuously monitoring performance.

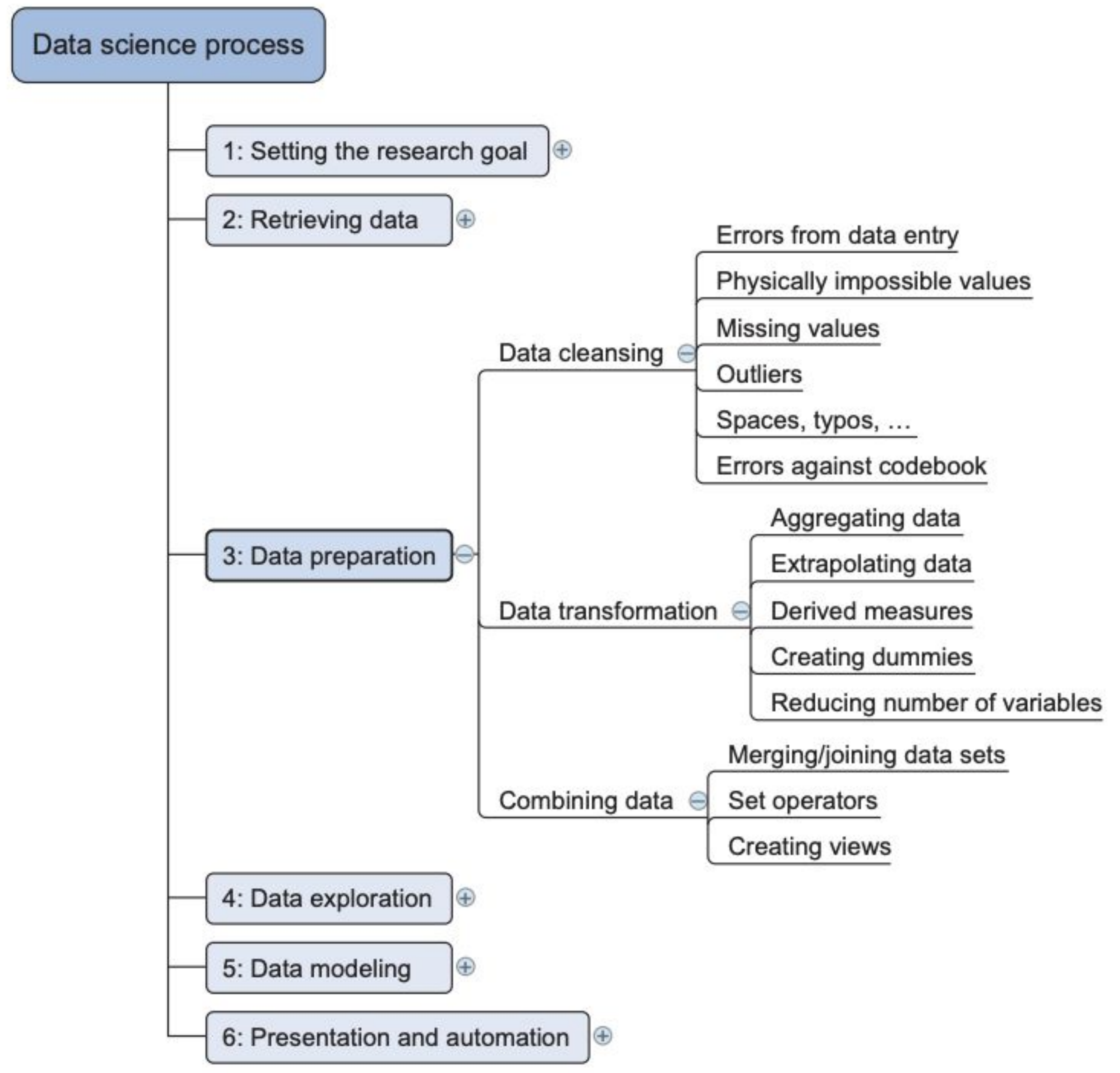


Figure 2.4 Step 3: Data preparation

Reference: https://livebook.manning.com/book/introducing-data-science/chapter-2/59

Course Data Analytics

Missing Data

Quiz1	Assignment:1	Quiz: 2	Assignment: 2	Assignment: 3	Programming Assignment total (Real)	Assignment: 4	Assignment: 5	Project total (Real)	Assignment: 6	Major Exam total (Real)	Course total (Real)	Final Exam: OpenCV	Final Exam TF	Final Exam ROS	Final Total	Total
7.65	10	9.11	10	10	18.71	18.5	10	19.25	19.5	19.5	58	15	14	9	38	96
8.77	10	8.67		10	18.97	20	10	20	19.75	20	59	11	10	5	26	85
8.89	10	9.33	10	10	19.29	18.5	10	19.25	17	17	56	12	12	5	29	85
2.41	10	-	10	10	16.2	20	10	20	11	11	48	10	8	4	22	70
7.16	10	10	10	10	18.86		10	20	17	17	56	14	13	7	34	90
7.53	10	7.56	10	10	18.03	20	10	20	16	16	55	11	11	6	28	83

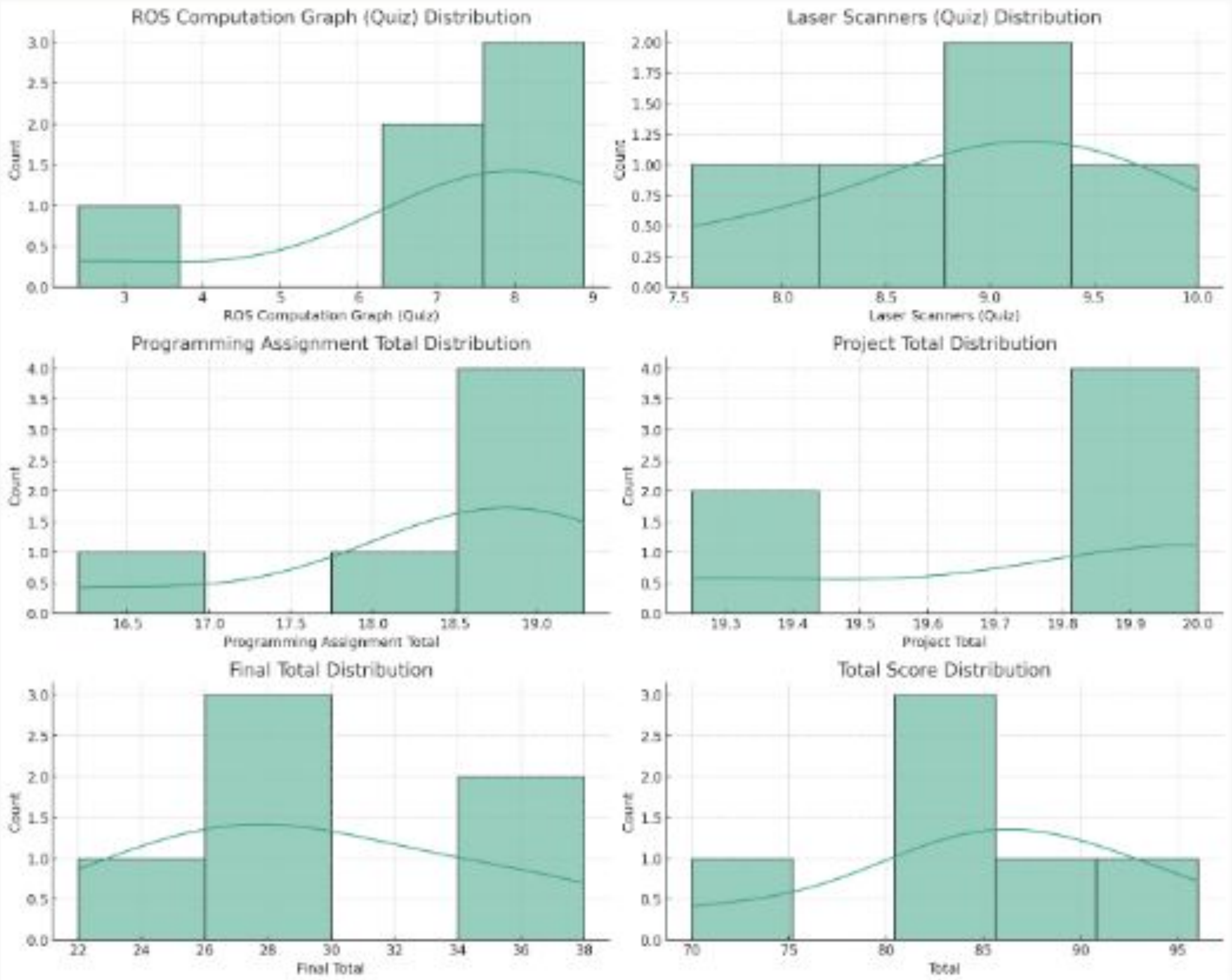
Exploratory Data Analysis (EDA)

From the initial statistics:

- The "Laser Scanners (Quiz)" column has a missing value since its count is 5 instead of 6.
- Most assignments have a consistent score of 10 for every student, indicating either they were easy or everyone did well.
- The total scores vary between 70 and 96, indicating a wide range of performances among the students.

Next, let's visualize the data distribution for significant columns using histograms and box plots.

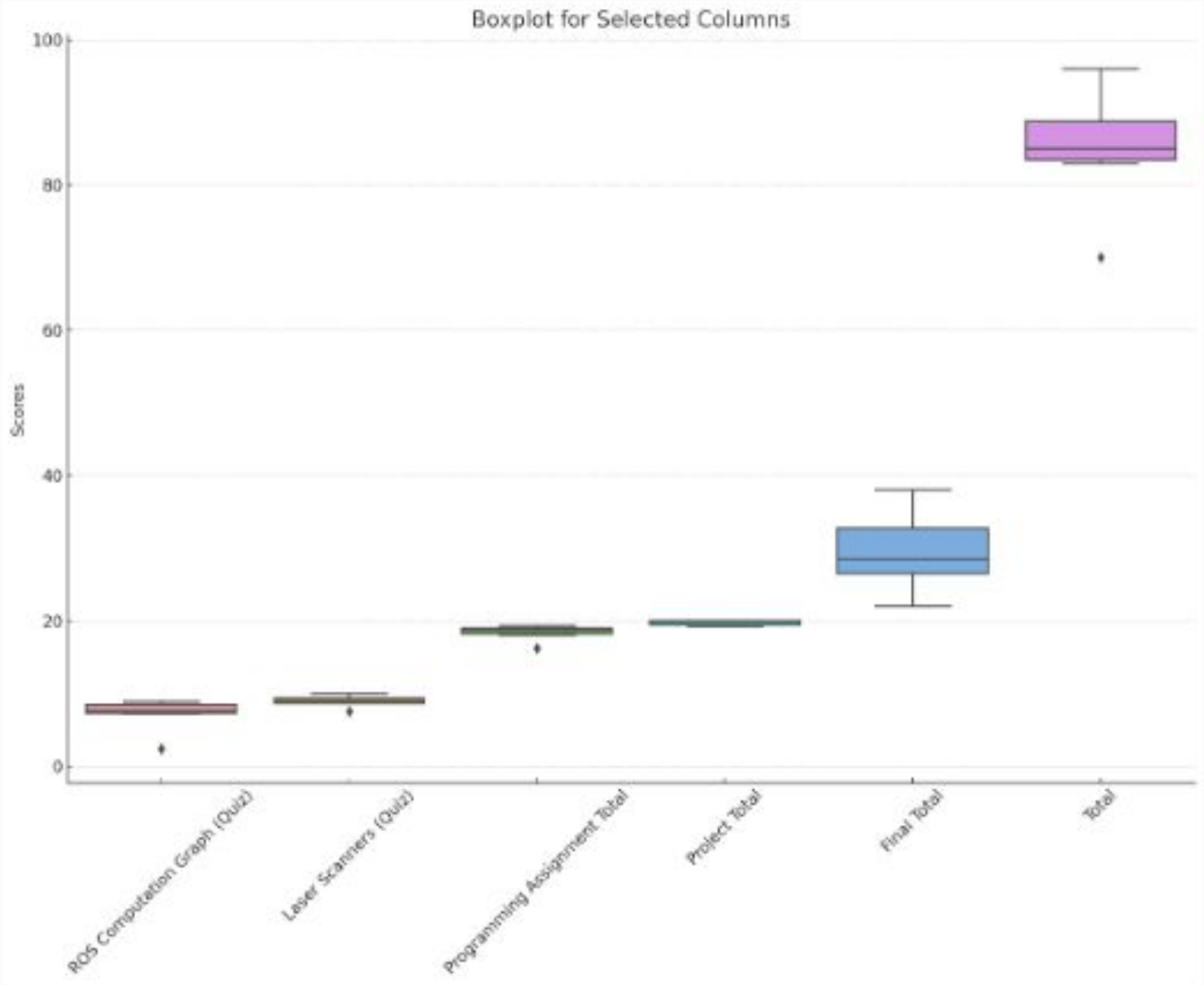
Finished working Show work



Here are the histograms showcasing the distribution of scores for selected significant columns:

Next, we'll use box plots to visualize the spread of scores and potentially identify any outliers.

Finished working Show work



Here are the box plots for the selected significant columns:

1. **ROS Computation Graph (Quiz)**: There's an outlier on the lower end, likely corresponding to the low score we observed in the histogram.
2. **Laser Scanners (Quiz)**: The scores are mostly between 8 and 10, but there's a lower whisker that might indicate an outlier or just a lower score.
3. **Programming Assignment Total**: The interquartile range (IQR) is narrow, indicating that most students scored similarly, close to the maximum.
4. **Project Total**: The scores are very concentrated at the upper limit. This reaffirms our observation that most students achieved the maximum score.
5. **Final Total**: The distribution is wider, with scores ranging from the low 20s to the upper 30s.
6. **Total**: There's a spread from 70 to 96, which means students' overall performance varied considerably.

There was an error generating a response

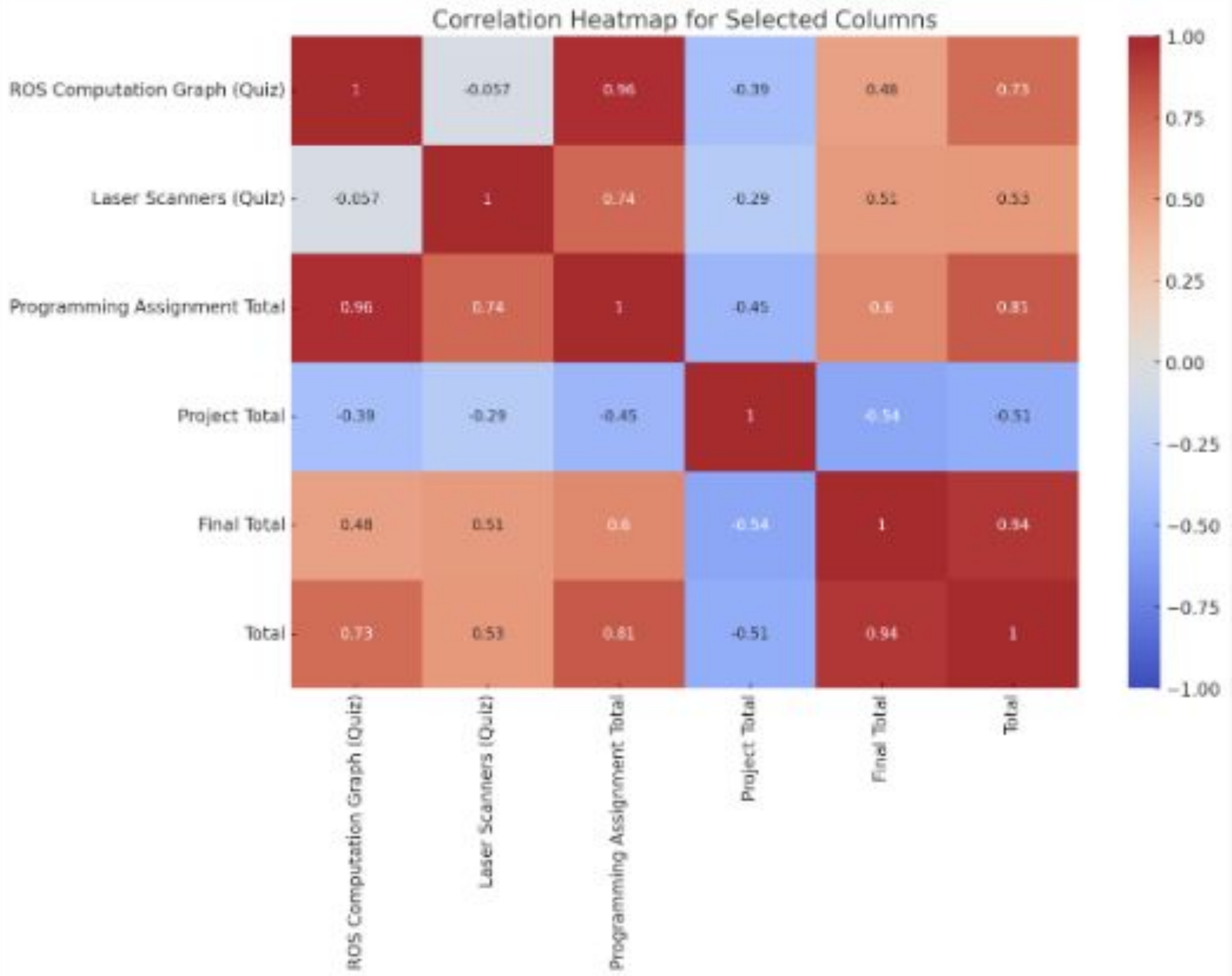
Regenerate

From the above results:

- Most students performed above average in the ROS Computation Graph Quiz and the Programming Assignments.
- Only half of the students scored above the average in the Laser Scanners Quiz.
- Only a third of the students performed above average in the final total.

Lastly, let's proceed with the Correlation Analysis to determine if there's a strong correlation between any two significant columns.

Finished working Show work



Exploratory Data Analysis (EDA)

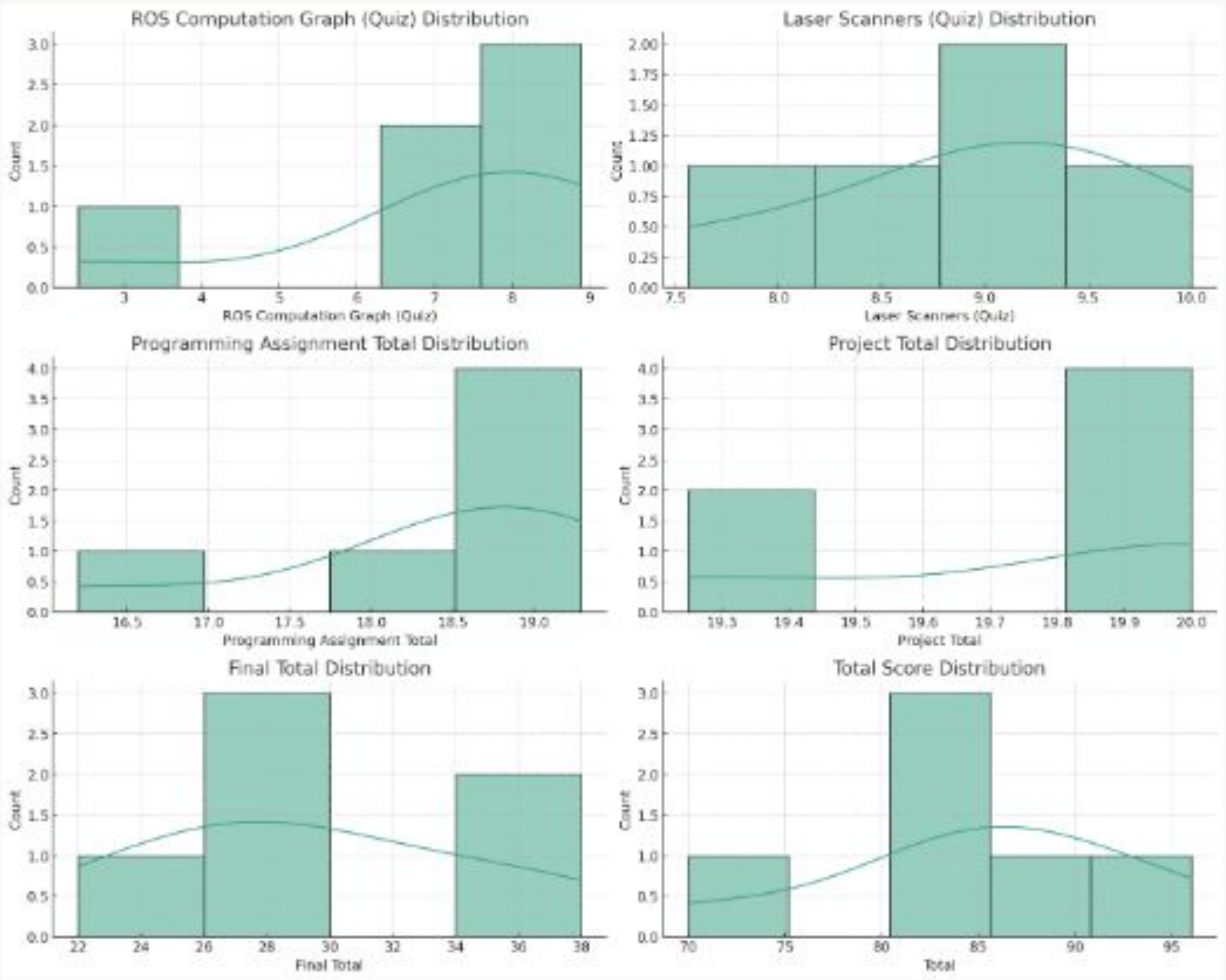
- **Purpose**
  - To understand data characteristics before modeling.
- **Techniques**
  - **Descriptive Statistics:** Mean, median, variance, etc.
  - **Data Visualization:** Histograms, scatter plots, box plots, etc.
  - **Correlation Analysis:** Identifying relationships between variables.
- **Tools**
  - Pandas, Matplotlib, Seaborn (Python); ggplot2, dplyr (R).

From the initial statistics:

- The "Laser Scanners (Quiz)" column has a missing value since its count is 5 instead of 6.
- Most assignments have a consistent score of 10 for every student, indicating either they were easy or everyone did well.
- The total scores vary between 70 and 96, indicating a wide range of performances among the students.

Next, let's visualize the data distribution for significant columns using histograms and box plots.

Finished working Show work

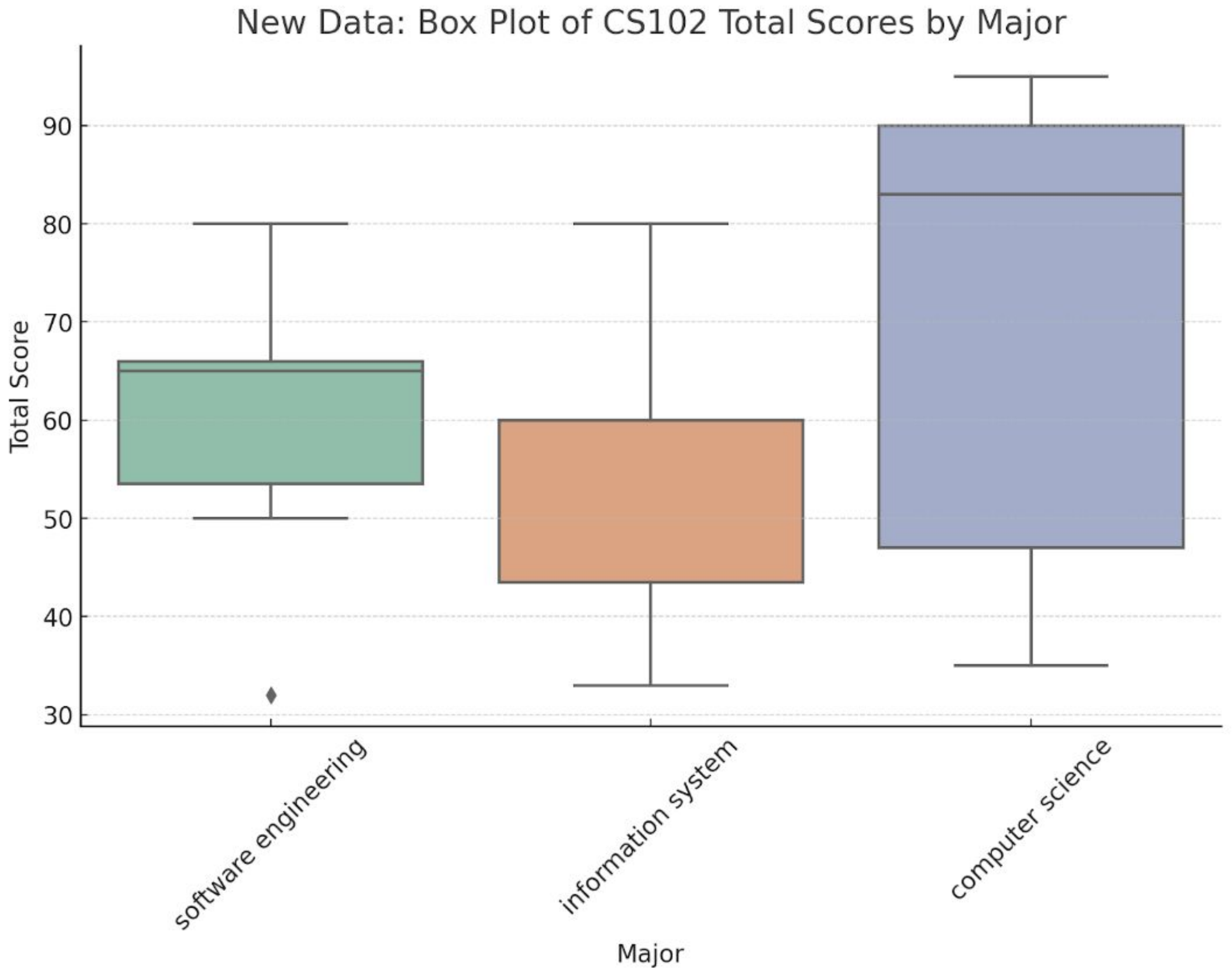


Here are the histograms showcasing the distribution of scores for selected significant columns:

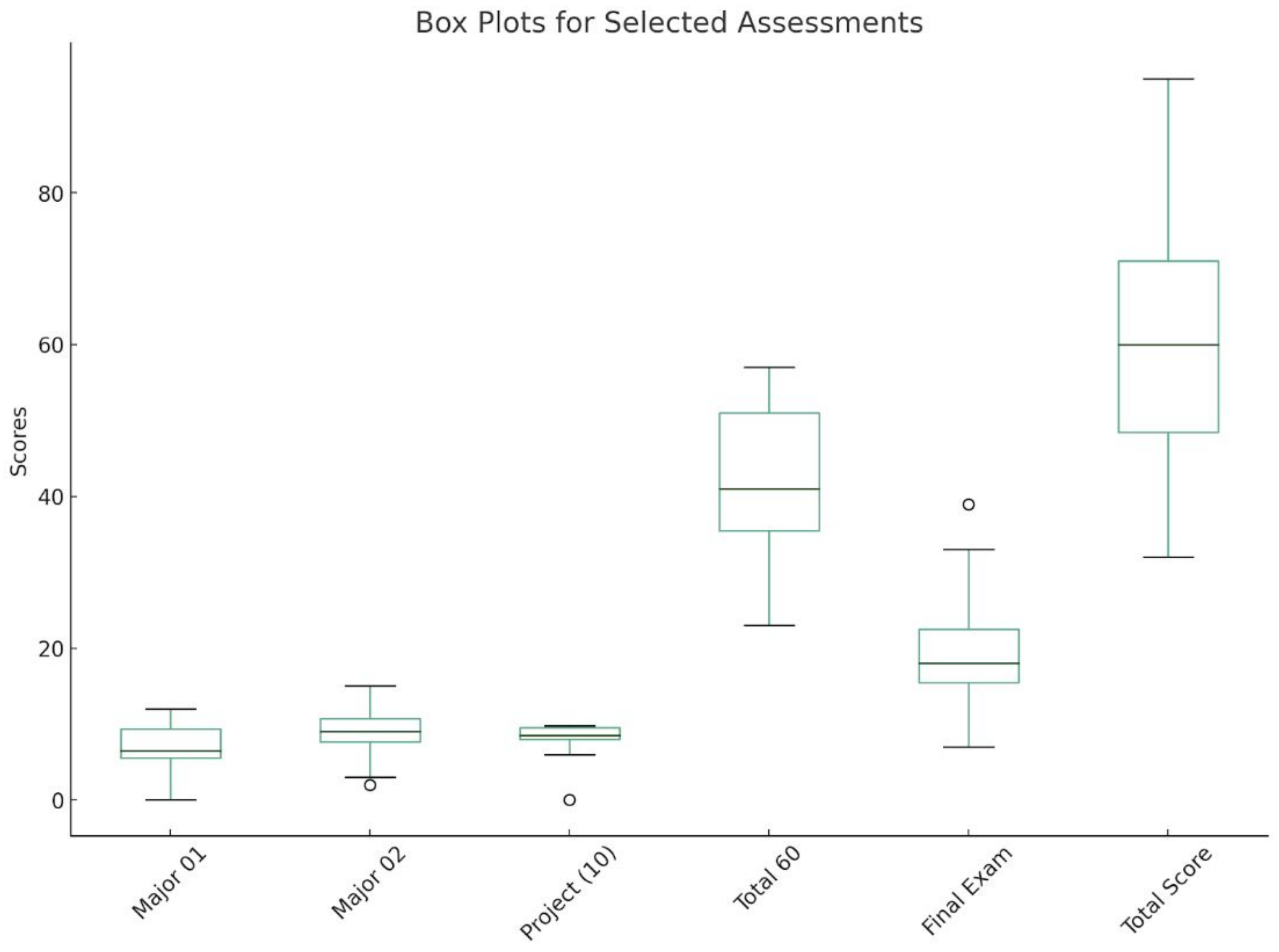
Analysis of Student Performance in CS102

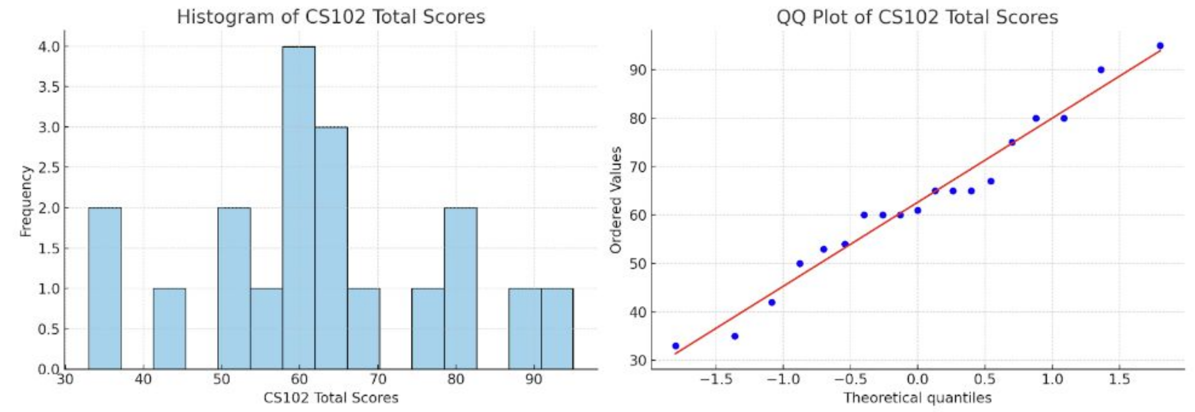
Summary Report:

Analysis of Student Performance in CS102 - Section 780 Fall 2023

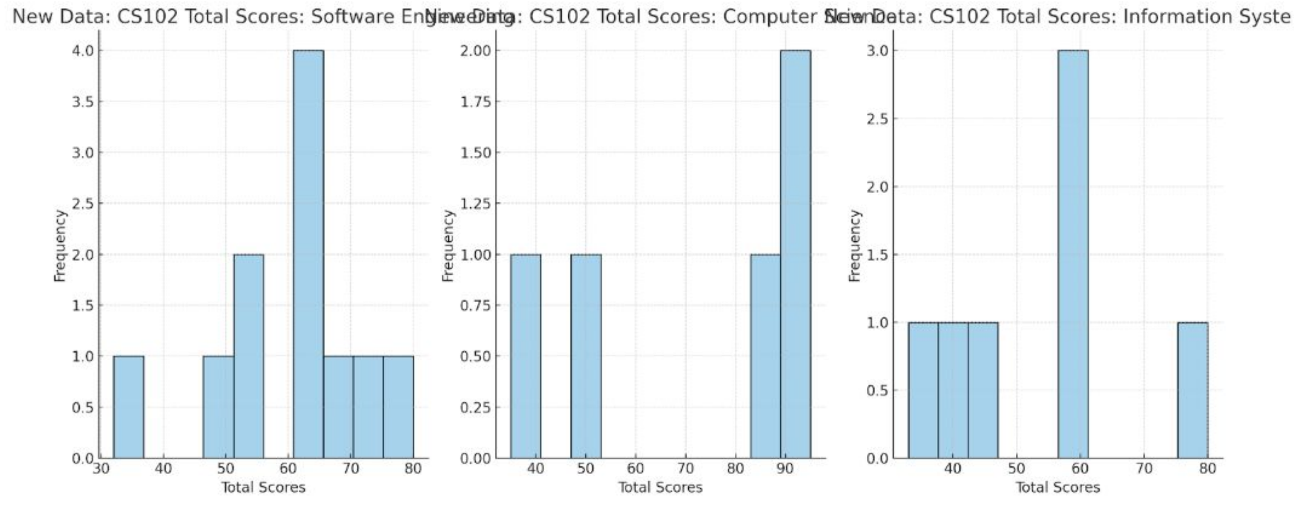


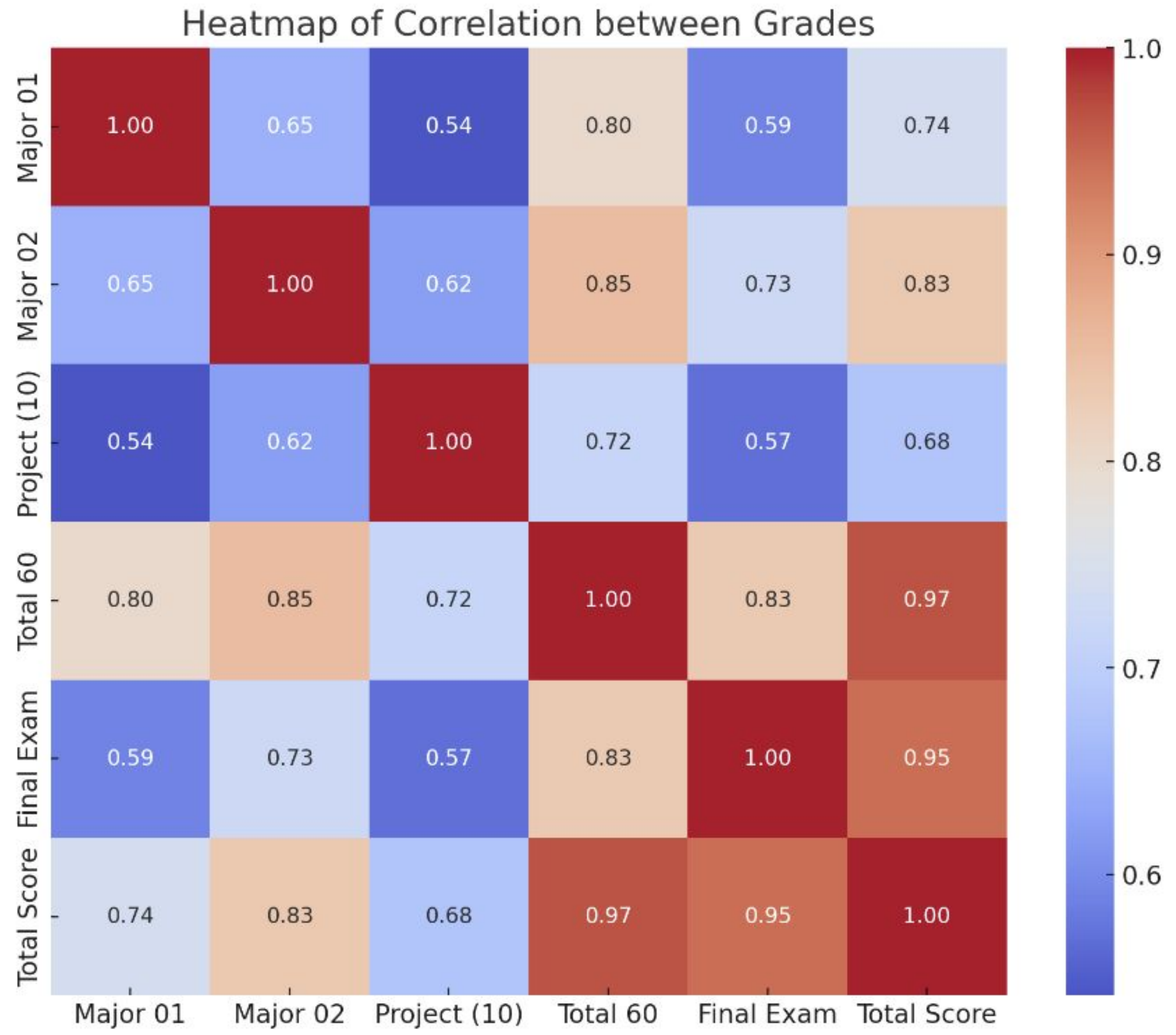
the considerable range and variance across assessments highlight the need for targeted support and interventions for students at different performance levels.

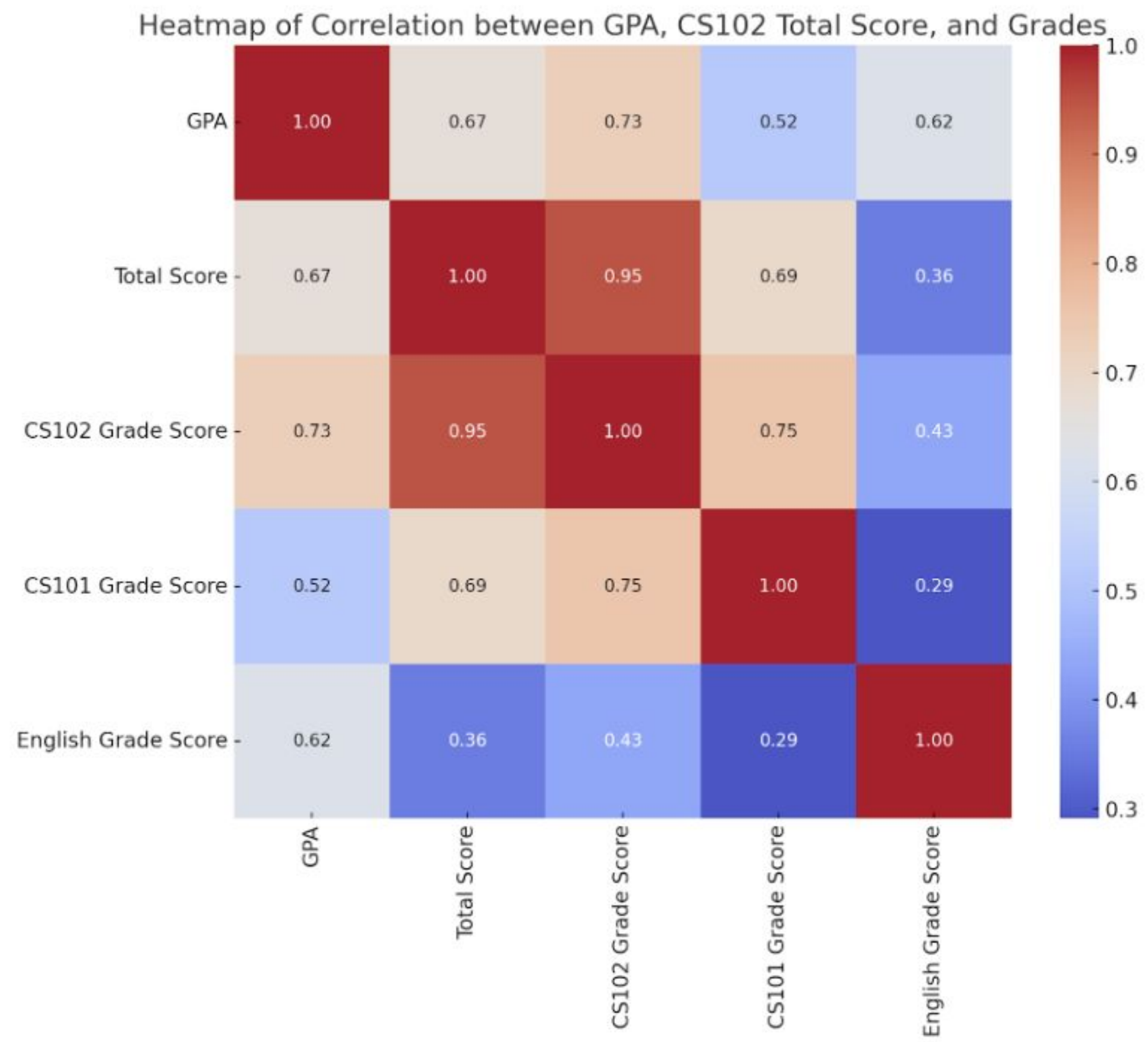




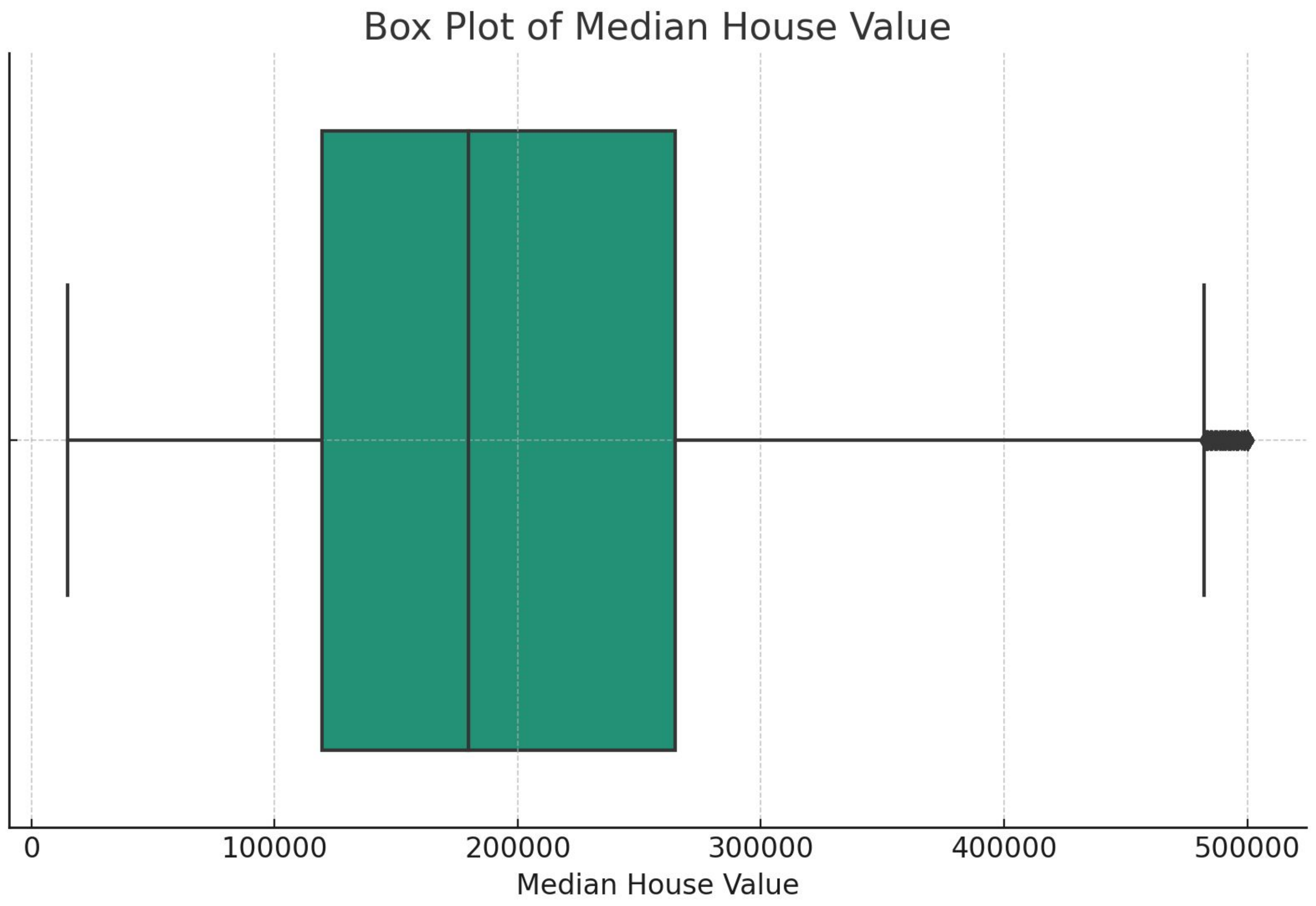
6. Major-Specific Observations:







California Housing Dataset – Price Ranges



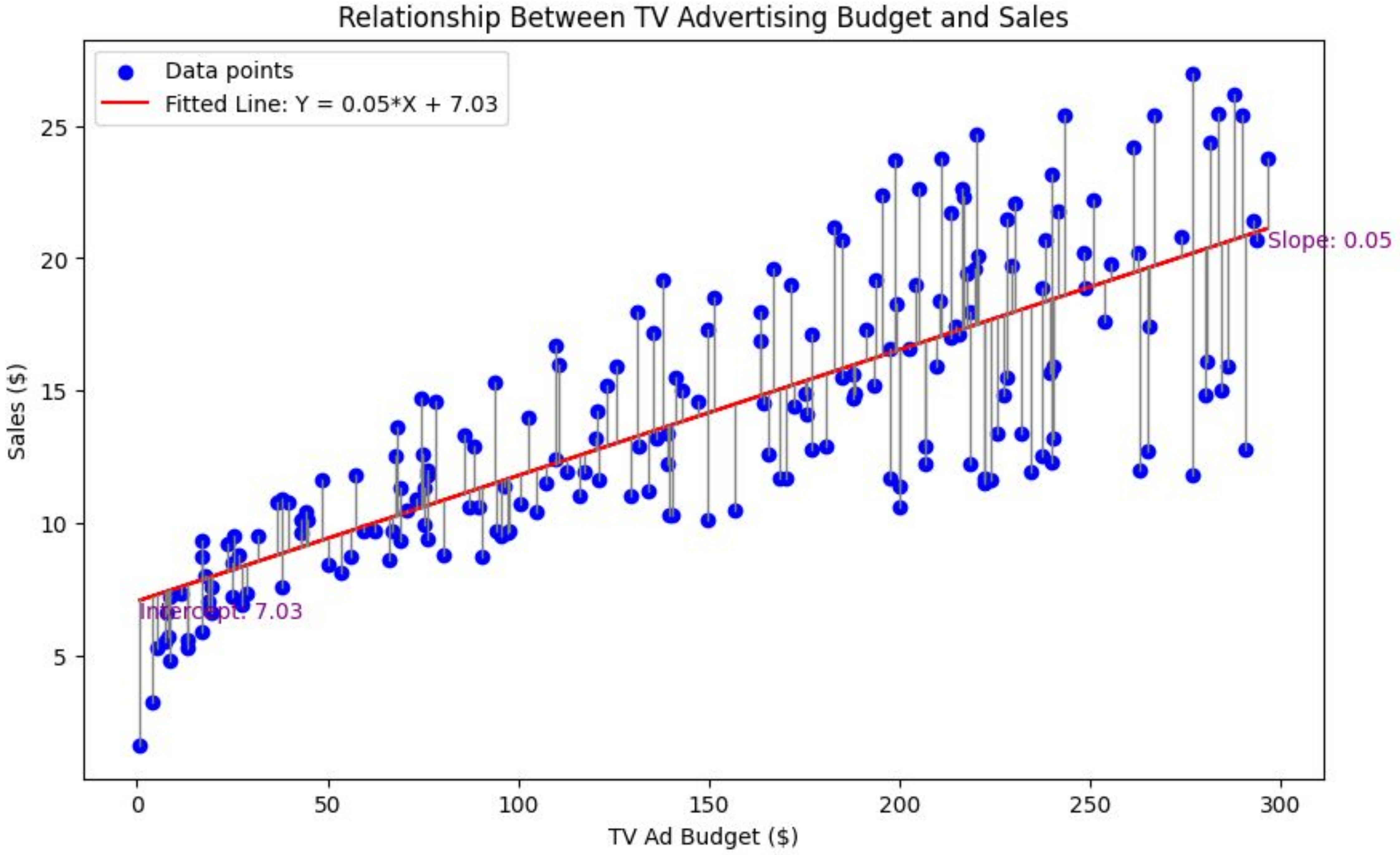
The descriptive statistics for the median house value in the dataset are as follows:

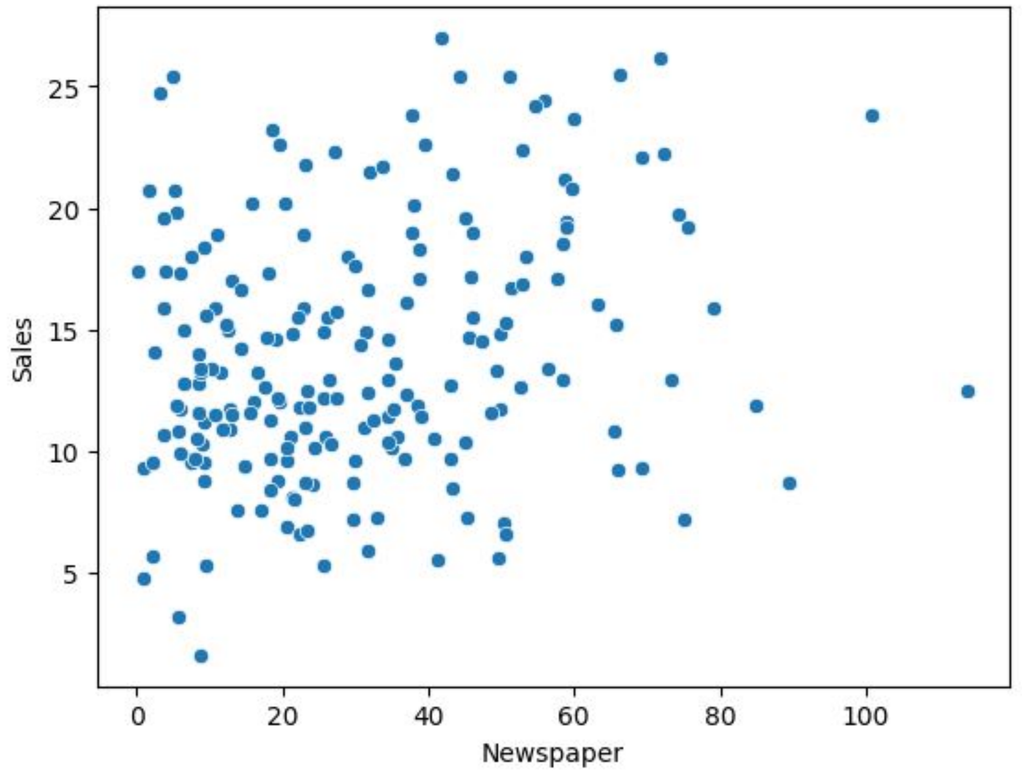
- **Count**: 20,640 (total number of entries)
- **Mean**: $206,855.82 (average median house value)
- **Standard Deviation**: $115,395.62 (variation in house values)
- **Minimum**: $14,999 (lowest median house value)
- **25th Percentile**: $119,600 (25% of houses have a median value below this)
- **Median (50th Percentile)**: $179,700 (half of the houses have a median value below this and half above)
- **75th Percentile**: $264,725 (75% of houses have a median value below this)
- **Maximum**: $500,001 (highest median house value)

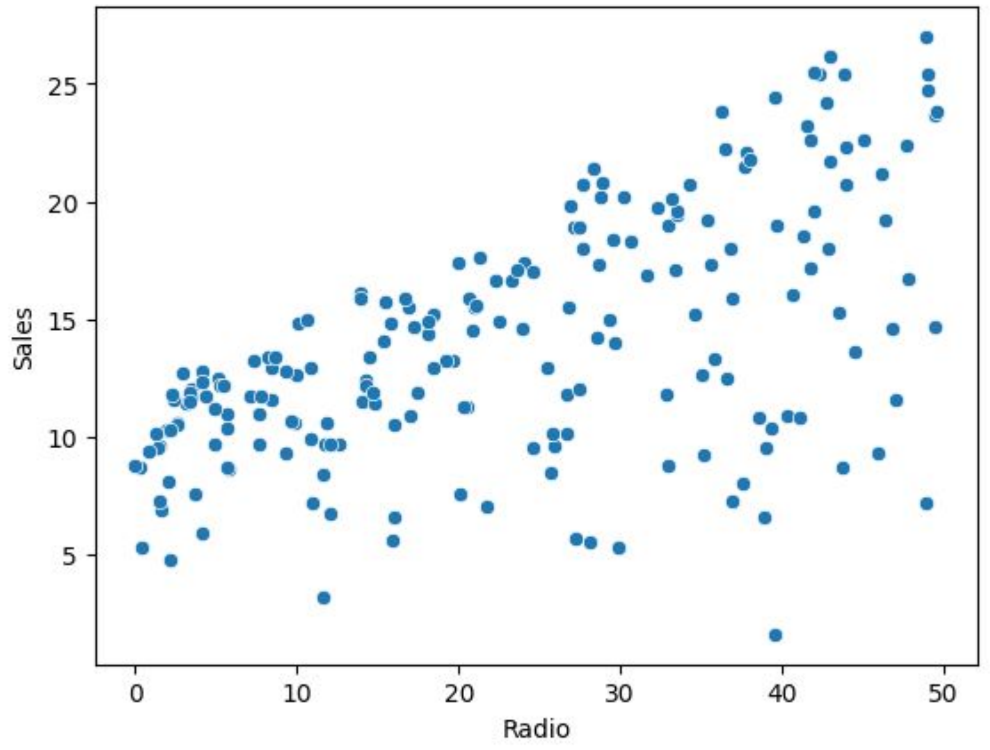
	A	B	C	D	E	F	G	H	I	J
1	longitude	latitude	housing_me	total_rooms	total_bedroc	population	households	median_inco	median_hou	ocean_proximity
2	-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
3	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
4	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
5	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
6	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
7	-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
8	-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
9	-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY
10	-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY
11	-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY
12	-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR BAY
13	-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR BAY
14	-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR BAY
15	-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR BAY
16	-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR BAY
17	-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR BAY
18	-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR BAY
19	-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR BAY
20	-122.26	37.84	50	2239	455	990	419	1.9911	158700	NEAR BAY

Advertising vs Sales – Predictive Modeling

	ID	TV	Radio	Newspaper	Sales
0	1	230.1	37.8	69.2	22.1
1	2	44.5	39.3	45.1	10.4
2	3	17.2	45.9	69.3	9.3
3	4	151.5	41.3	58.5	18.5
4	5	180.8	10.8	58.4	12.9







AI, Machine Learning, Deep Learning

FOR

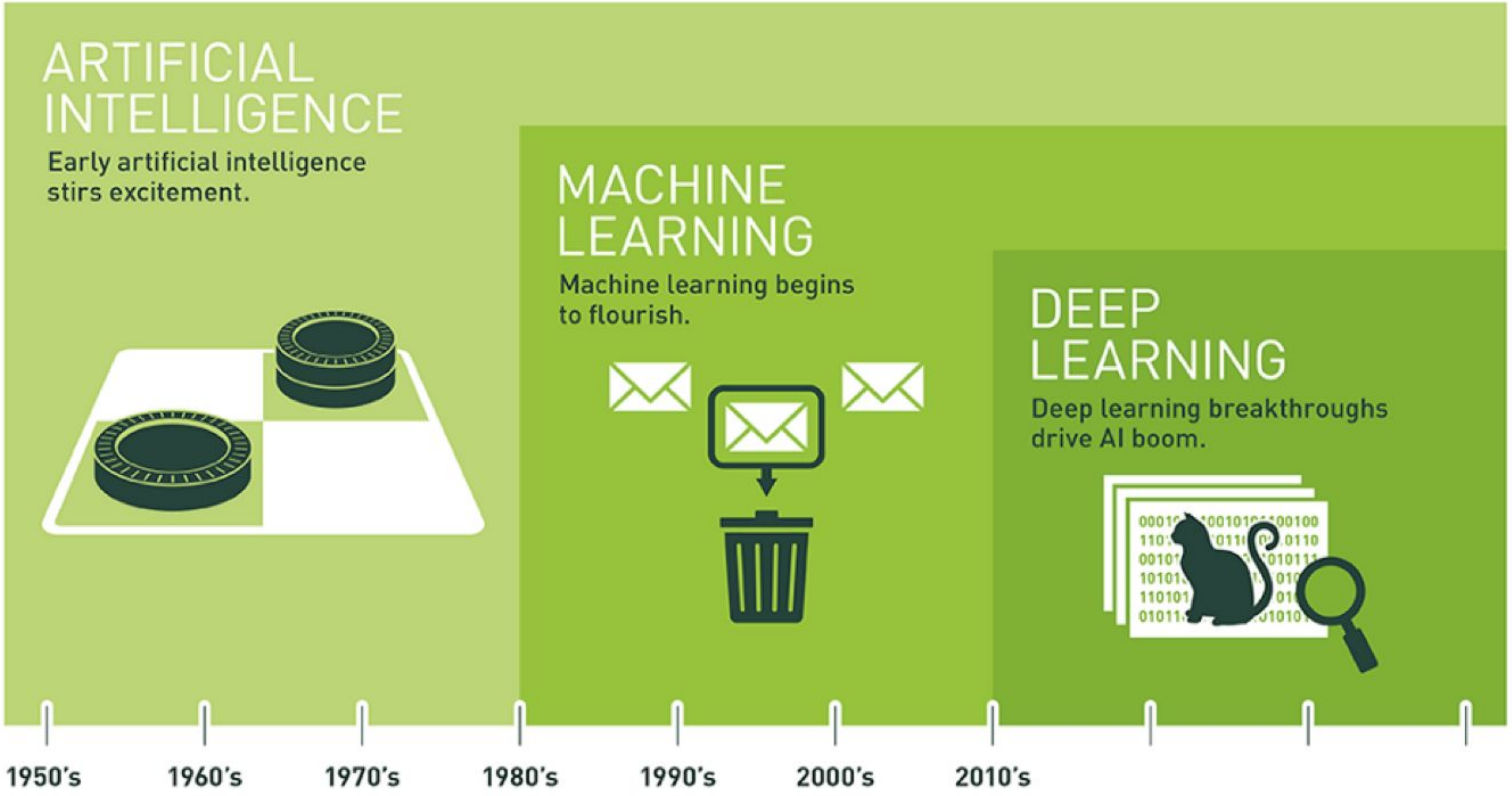
SPECIALIZATION

Lecture 1

Introduction to Data Science

Prof. Anis Koubaa

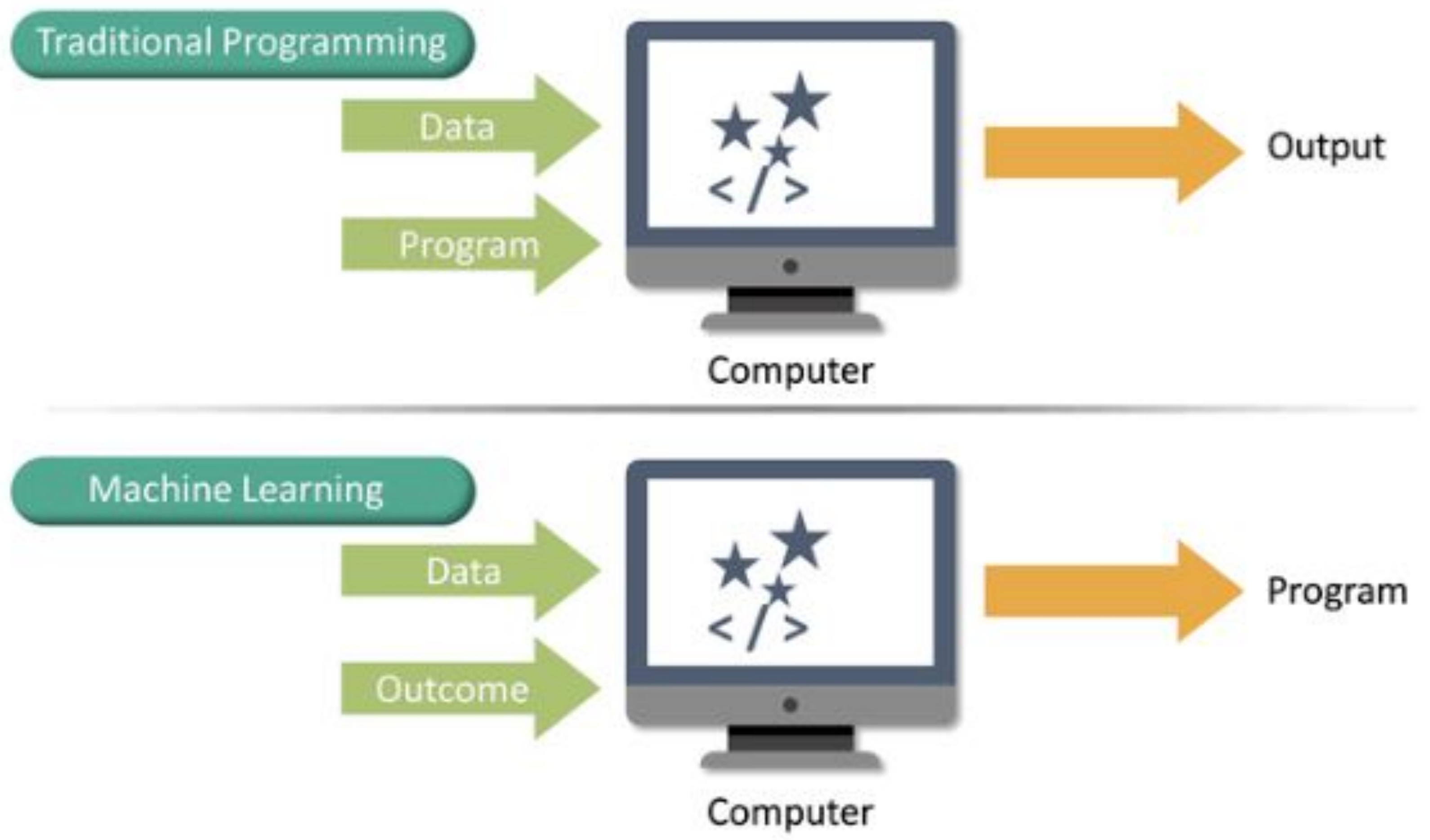
AI, MACHINE LEARNING, DEEP LEARNING



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

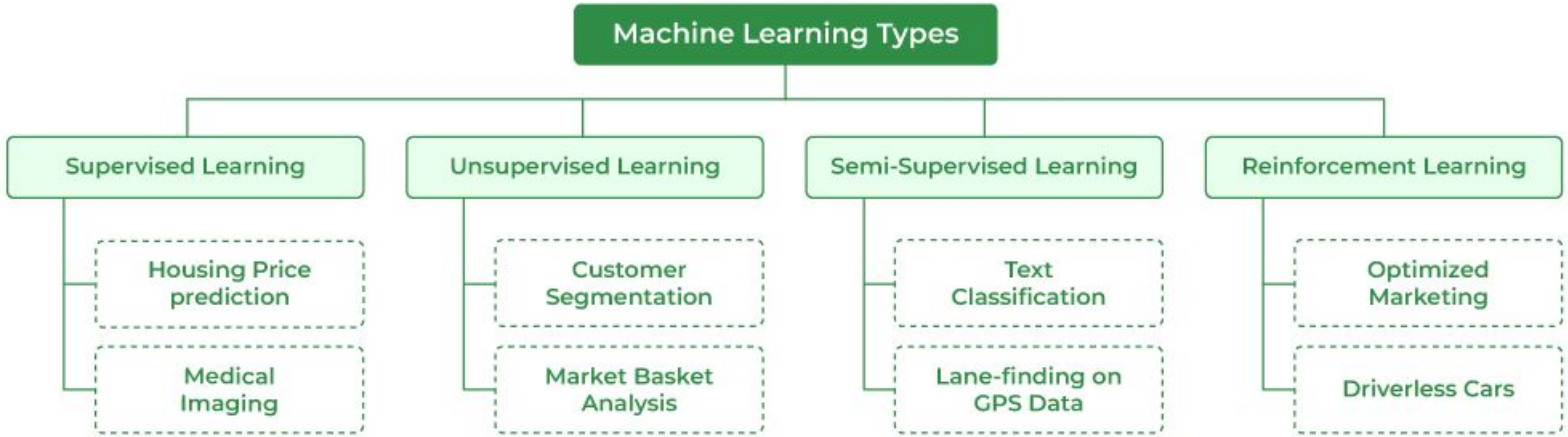
Machine Learning Overview

- **Supervised Learning:** Learning from labeled data (e.g., classification, regression).
- **Unsupervised Learning:** Finding patterns in unlabeled data (e.g., clustering, dimensionality reduction).
- **Reinforcement Learning:** Learning through interaction with an environment to maximize cumulative reward.
- **Popular Algorithms:** Linear Regression, Decision Trees, K-Means Clustering, Neural Networks.



https://www.sketchbubble.com/en/presentation-machine-learning.html

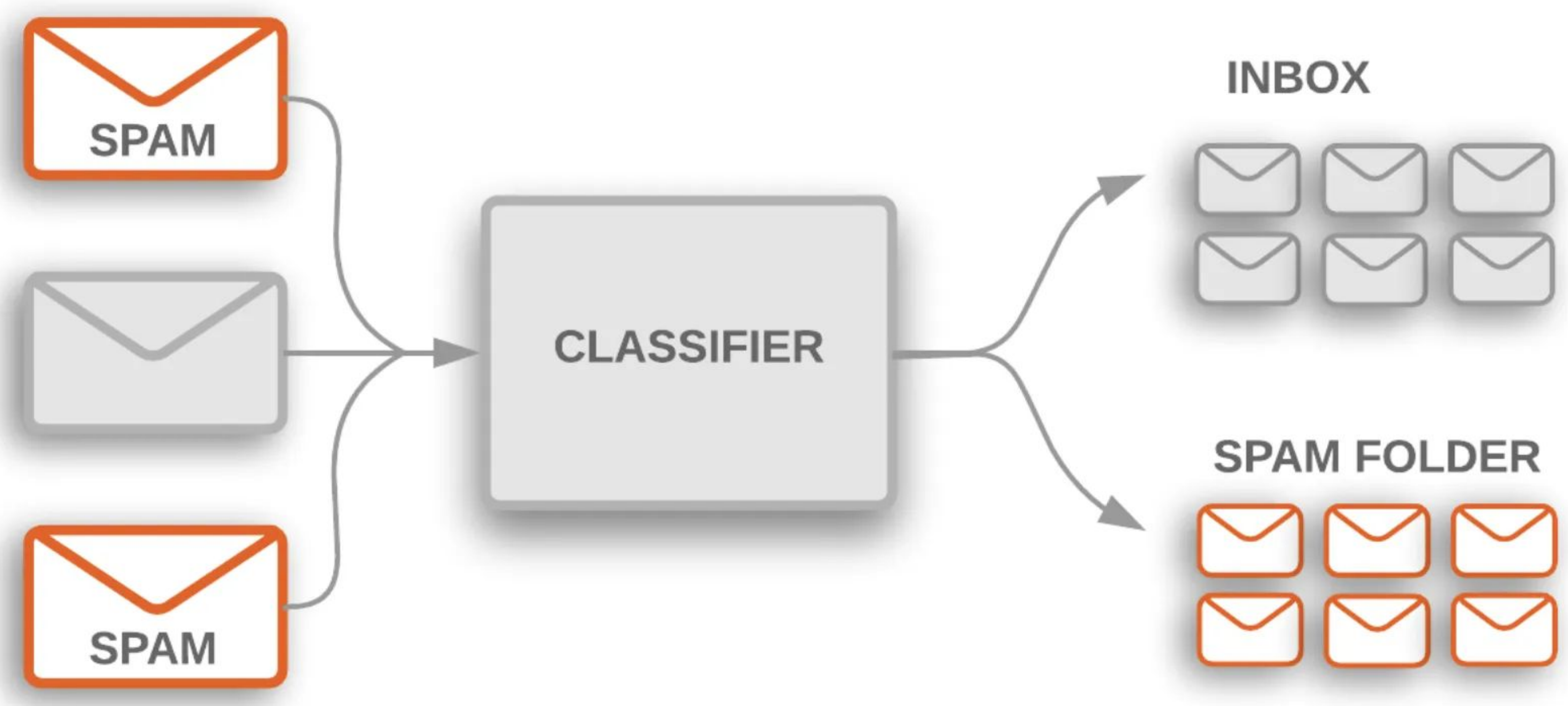
Machine Learning Types



Reference: https://www.geeksforgeeks.org/machine-learning-algorithms/

Supervised Learning (Classification) Example

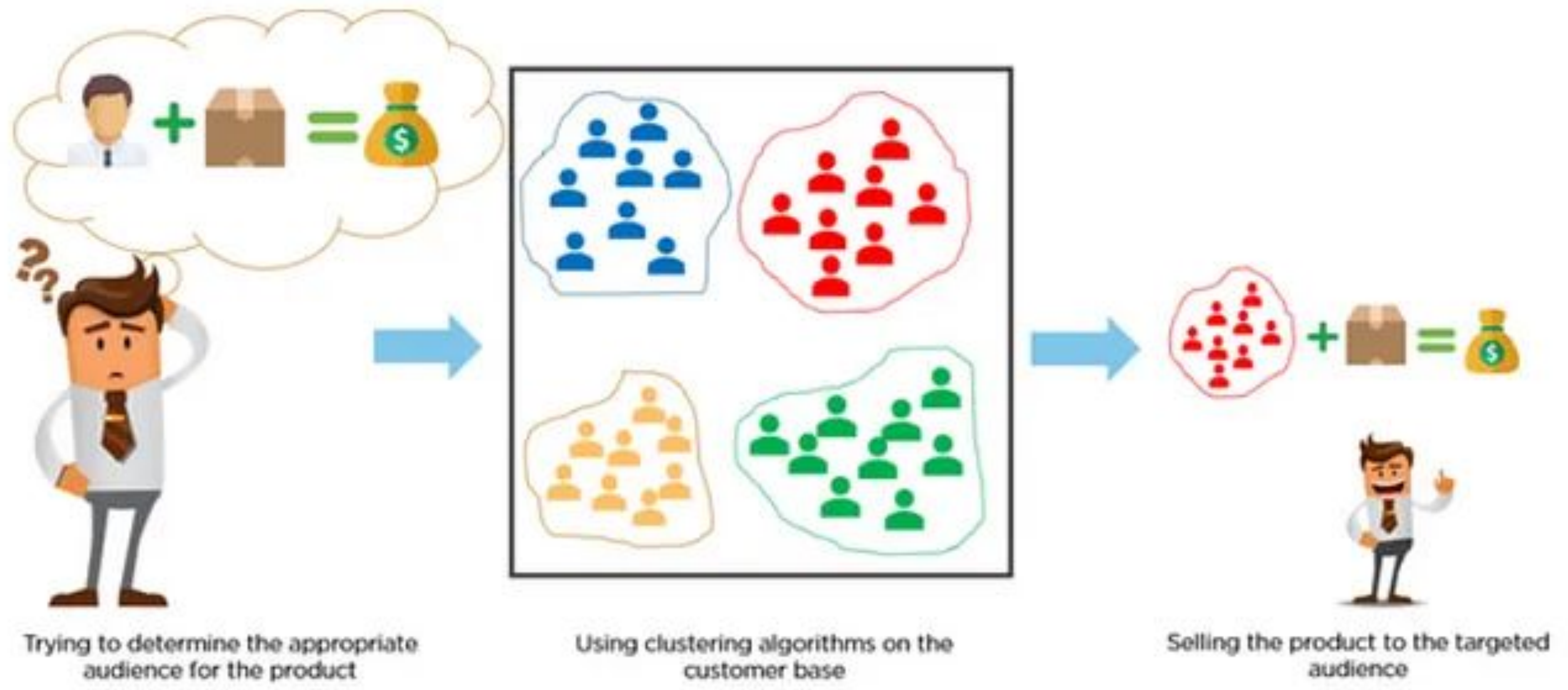
- **Objective:** Automatically classify incoming emails as either Inbox or Spam.
- **Technique:** A machine learning Classifier (e.g., Naive Bayes, SVM) predicts if an email is spam based on features like content and sender information.
- **Outcome:**
  - **Inbox:** Legitimate emails are delivered to the user's inbox.
  - **Spam Folder:** Unwanted emails are filtered into the spam folder, improving user experience and security.



Unsupervised Learning (Clustering) Example

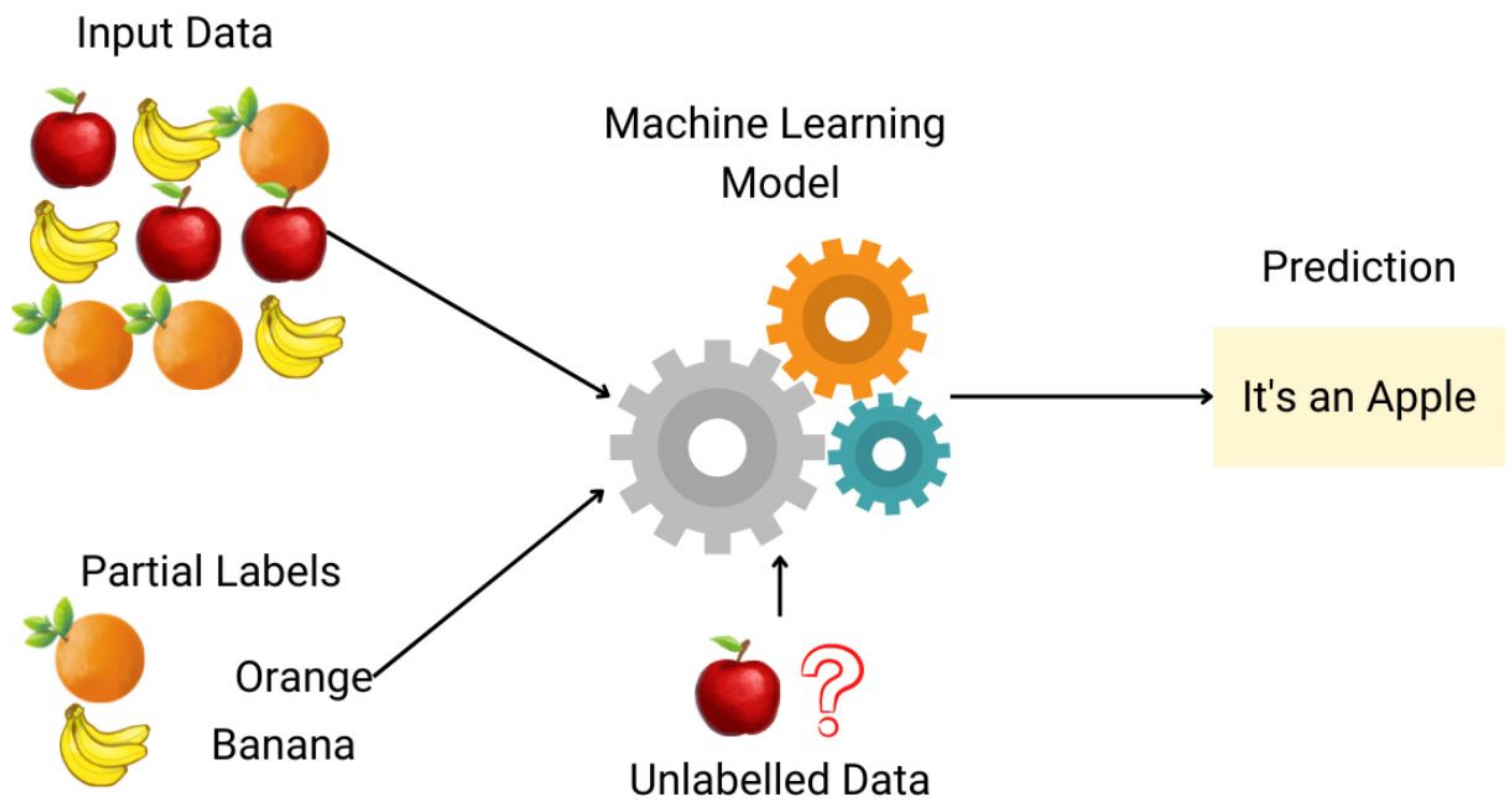
- **Objective:** Group customers into distinct segments based on behavior and characteristics.
- **Solution:** Apply clustering algorithms to segment the customer base into distinct groups.
- **Outcome:** Identify targeted audience clusters to optimize product marketing and sales strategies.
- **Benefits:**
  - Tailored marketing strategies for each customer segment.
  - Improved customer engagement and retention.
  - Enhanced product recommendations and personalized experiences.

Customer Segmentation



Semi-Supervised Learning Example

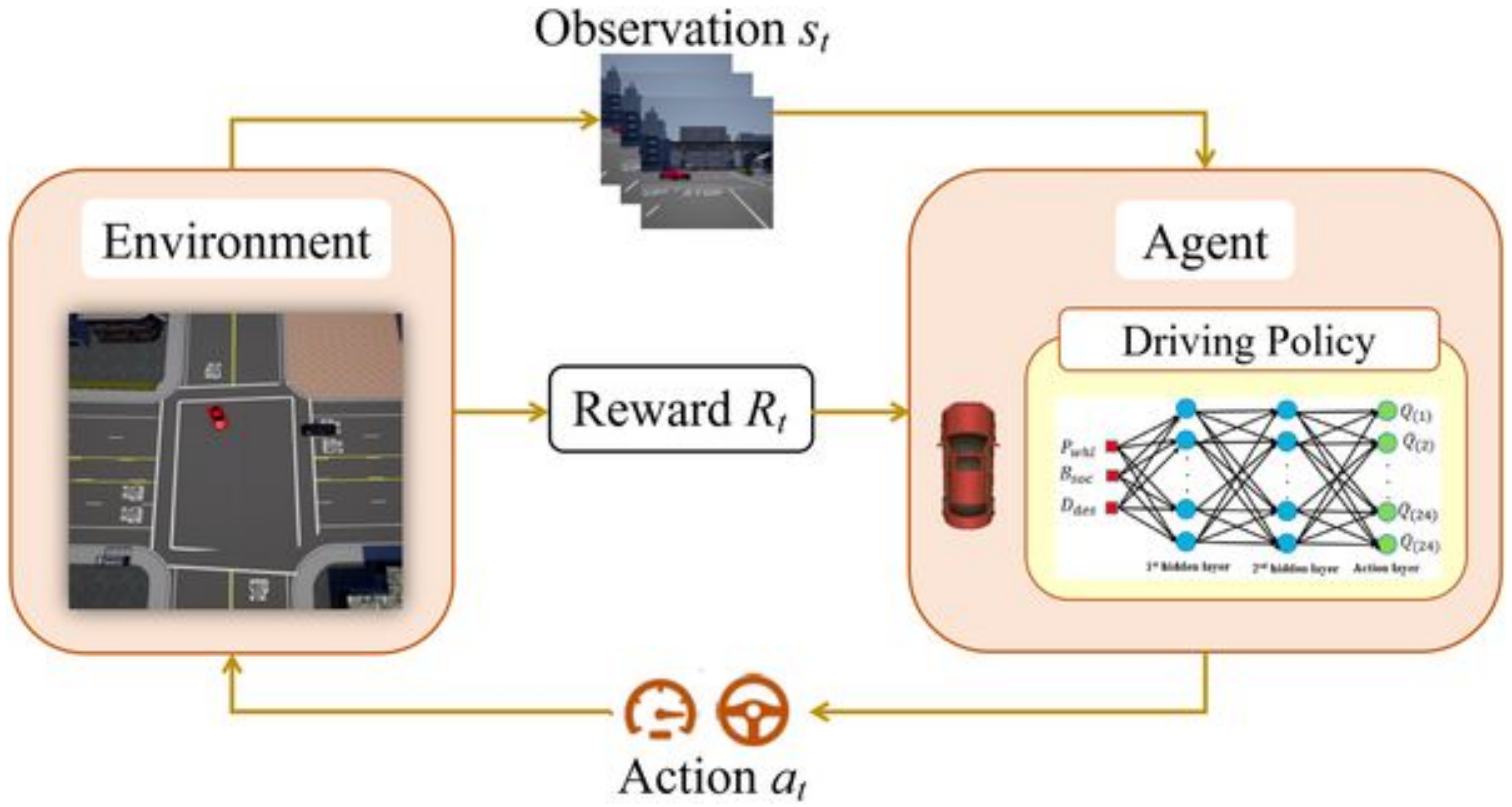
- **Objective:** Train machine learning models with limited labeled data and a large amount of unlabeled data.
- **Input:** Combination of labeled and unlabeled data points from various sources.
- **Technique**
  - Model leverages labeled data to learn key features.
  - Uses patterns in labeled data to infer labels for unlabeled data.
- **Applications**
  - **Text Classification:** Classify documents/emails with few labeled examples.
  - **Image Recognition:** Identify objects using small labeled datasets combined with large unlabeled datasets.
- **Outcome:** Improved model performance by utilizing all available data efficiently, reducing dependency on labeled data.



Objective: Classify fruits with partial labeled data using a machine learning model.

Reinforcement Learning Example

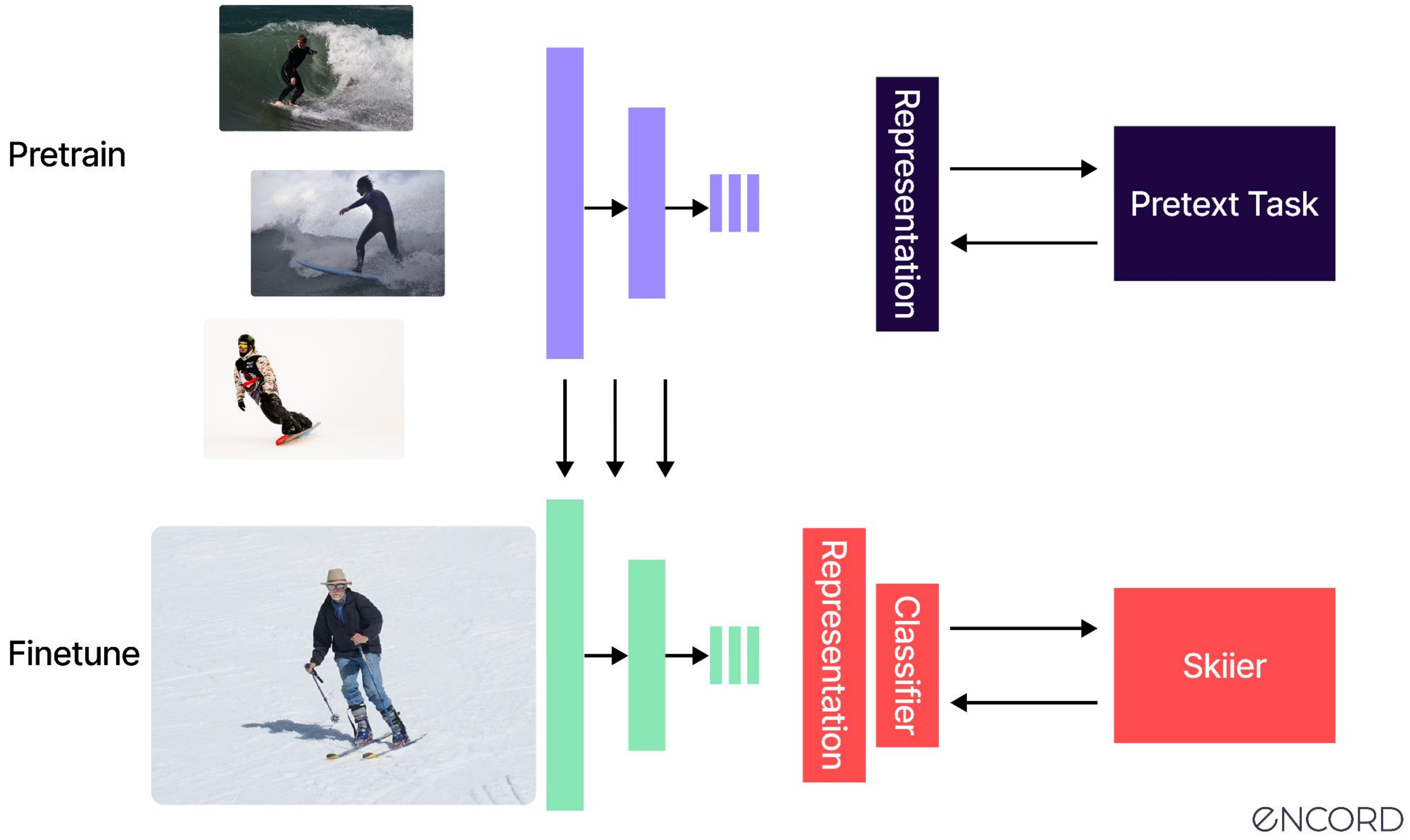
- **Objective**: Train an **Agent** (autonomous vehicle) to navigate an environment safely and efficiently.
- **Components**:
  - **Agent**: Learns a **Driving Policy** using neural networks to make decisions (actions).
  - **Environment**: Simulated driving scenario providing real-time feedback.
- **Process**:
  - **Observation** (s_t): The agent perceives the environment (e.g., road, obstacles).
  - **Action** (a_t): The agent takes actions (e.g., steering, acceleration) based on the policy.
  - **Reward** (R_t): The agent receives rewards or penalties based on the outcomes of its actions, learning to maximize long-term success.



Reference: https://link.springer.com/article/10.1007/s42154-020-00113-1

Self-Supervised Learning Example

- **Objective:** Enable models to learn useful representations from unlabeled data without explicit supervision.
- **Approach:** Create **pretext** tasks where the model generates labels from the data itself (e.g., predicting the next word in a sentence or missing parts of an image).
- **Key Characteristics:**
  - No need for large labeled datasets.
  - Data provides its own supervision through structured tasks.
- **Examples:**
  - **NLP:** Models like GPT and BERT predict missing words or sentence structure.
  - **Computer Vision:** Predicting the orientation of an image or filling in missing pixels.
- **Outcome:** Self-supervised learning helps in pretraining models that can be fine-tuned for downstream tasks with limited labeled data.

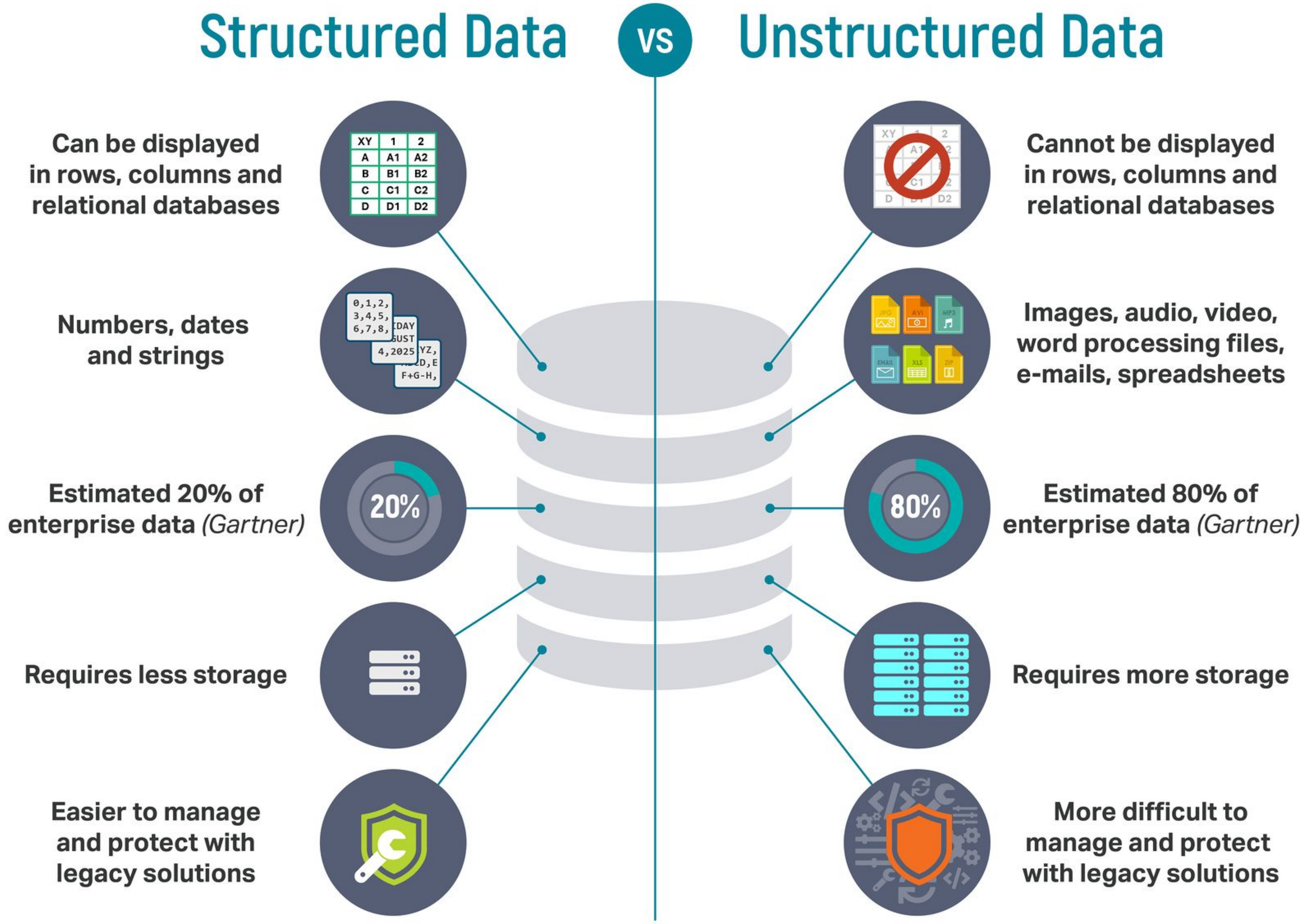


Type of Data

- **Structured Data**: Data that follows a predefined format, typically stored in tables (e.g., databases, spreadsheets).
- **Unstructured Data**: Data without a formal structure (e.g., text, images, videos).
- **Semi-Structured Data**: Has some organizational properties but lacks a rigid structure (e.g., JSON, XML).
- **Big Data**: Large-scale data that cannot be processed efficiently with traditional tools (e.g., Hadoop, Spark).

STRUCTURED vs. UNSTRUCTURED DATA

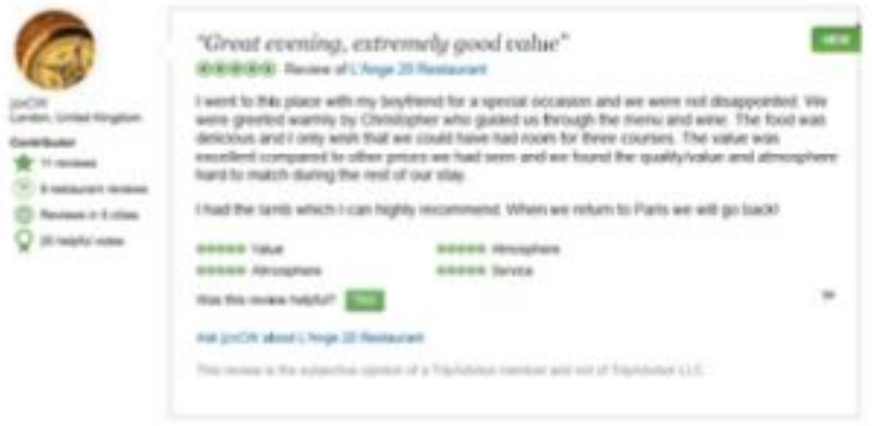
Data	Number of timepoints	Number of nodes
Antibiotics	56	386
Preterm births	216	318
Housing prices	254	944
Bikesharing	24	819
Global patterns	500	51



Audio

Image

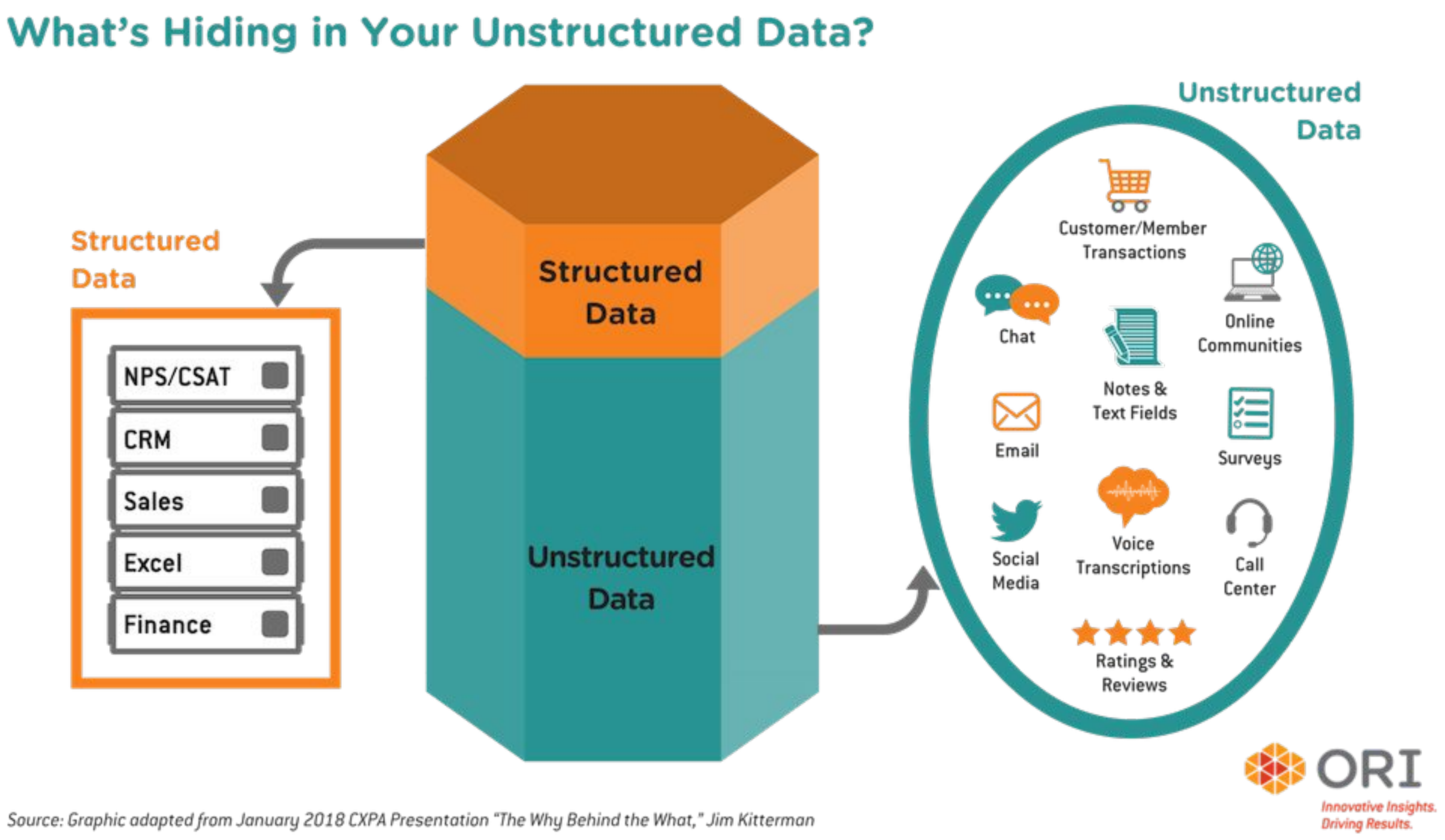
Text

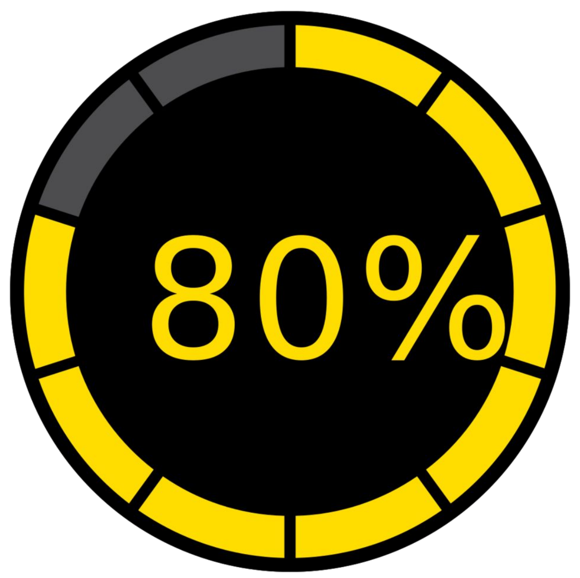


https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

STRUCTURED vs. UNSTRUCTURED DATA

EXAMPLE





https://www.oriresults.com/articles/blog-posts/whats-hiding-in-your-unstructured-data/

Semi-Structured Data

EXAMPLE

{JSON}

json   Copy code

```json
{
  "student": {
    "first_name": "Ahmad",
    "last_name": "Al-Farhan",
    "nationality": "Saudi",
    "university": "Prince Sultan University",
    "student_id": "PSU123456",
    "major": "Computer Science",
    "year": 3,
    "courses": [
      {
        "course_code": "CS101",
        "course_name": "Introduction to Computer Science",
        "grade": "A"
      },
      {
        "course_code": "MATH203",
        "course_name": "Calculus II",
        "grade": "B+"
      }
    ]
  }
}
```

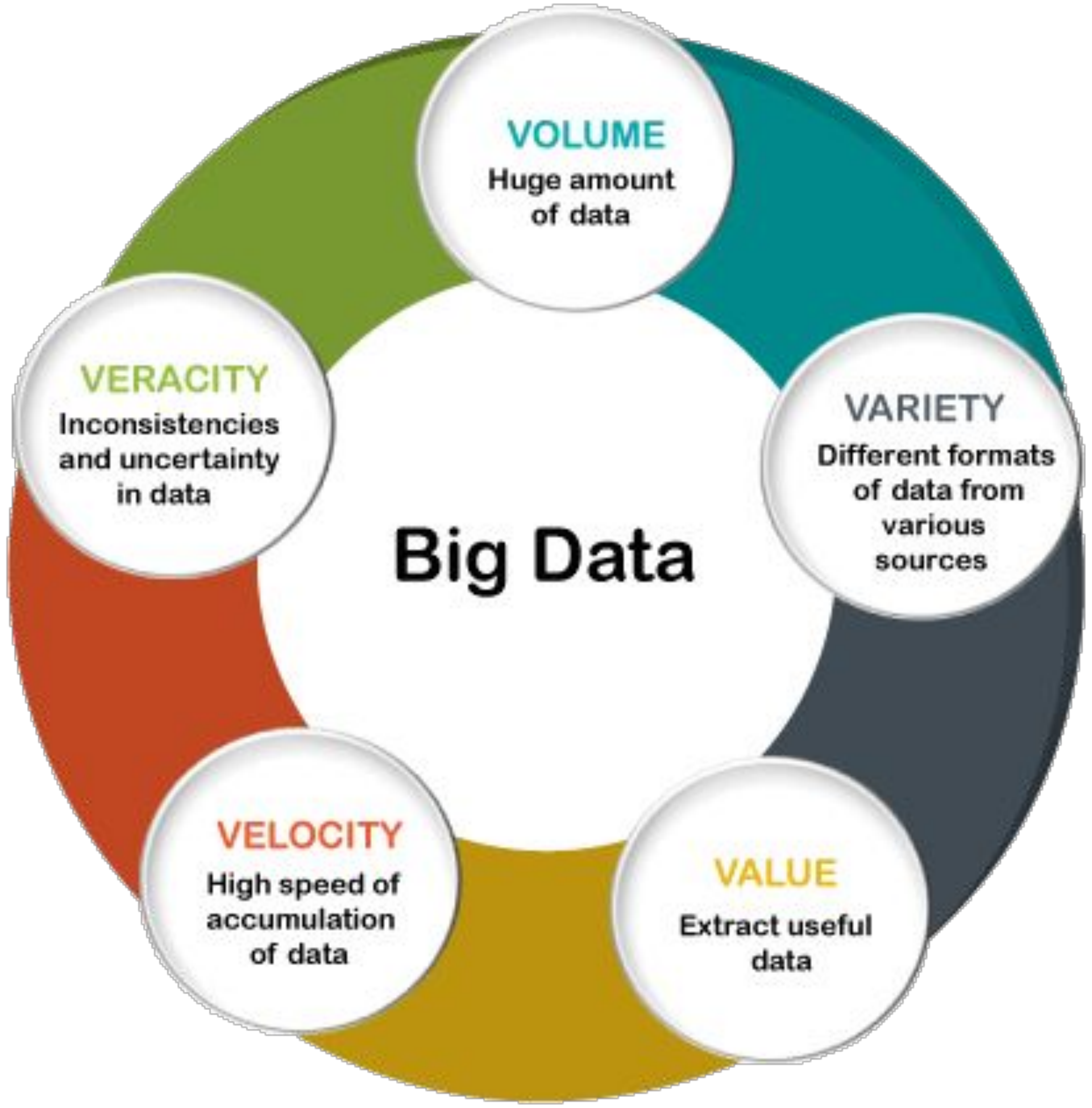
XML

xml   Copy code

```xml
<student>
    <first_name>Ahmad</first_name>
    <last_name>Al-Farhan</last_name>
    <nationality>Saudi</nationality>
    <university>Prince Sultan University</university>
    <student_id>PSU123456</student_id>
    <major>Computer Science</major>
    <year>3</year>
    <courses>
        <course>
            <course_code>CS101</course_code>
            <course_name>Introduction to Computer Science</course_name>
            <grade>A</grade>
        </course>
        <course>
            <course_code>MATH203</course_code>
            <course_name>Calculus II</course_name>
            <grade>B+</grade>
        </course>
    </courses>
</student>
```

Big Data

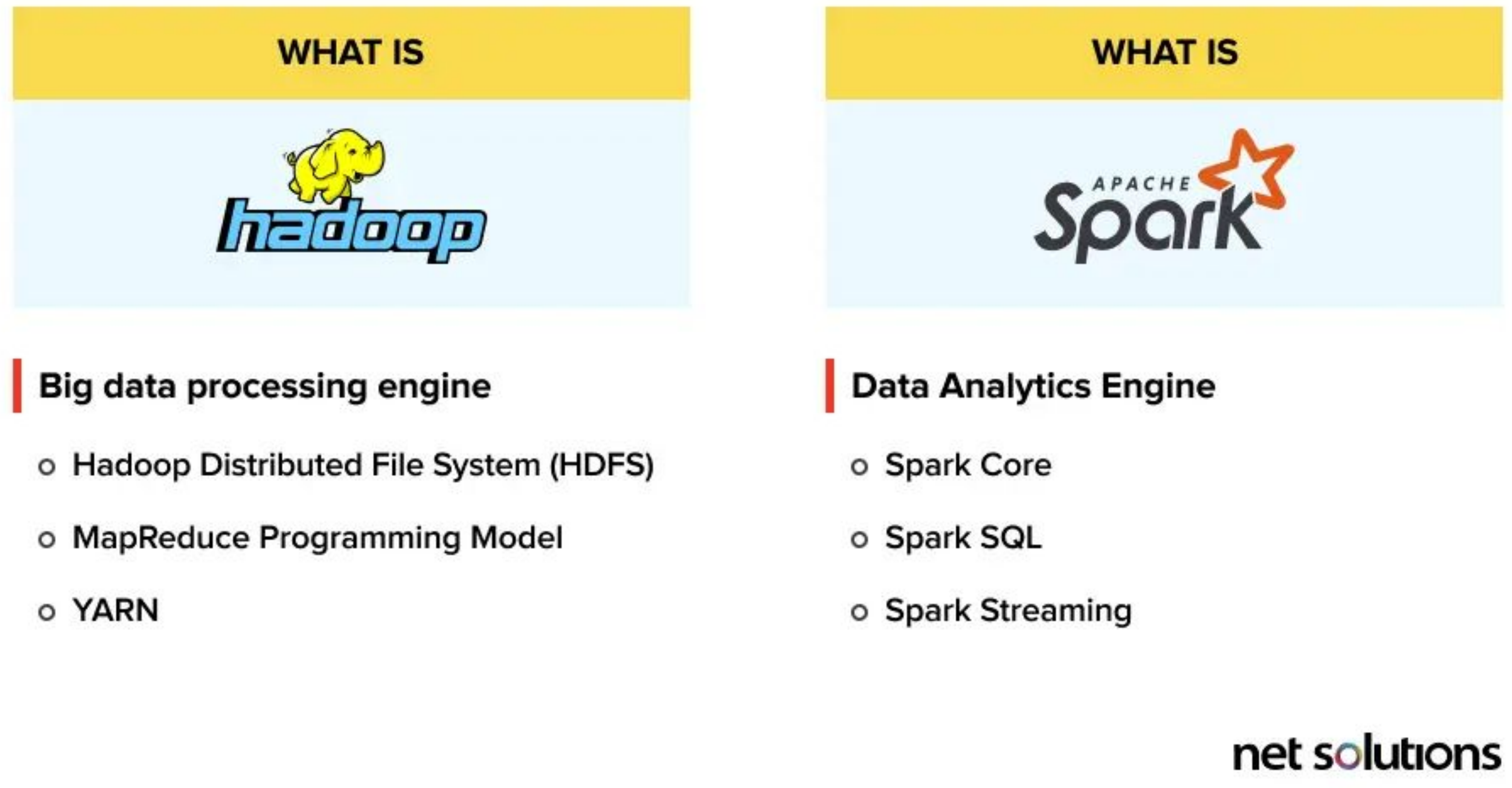
- **Definition:**
  - Big Data refers to large-scale datasets that are too complex, large, and dynamic for traditional data processing tools to handle efficiently.
- **Challenges:**
  - **Volume:** Massive amounts of data.
  - **Velocity:** High speed at which data is generated.
  - **Variety:** Different types of data (structured, unstructured, semi-structured).
  - **Veracity:** Uncertainty of data quality.



Reference: https://www.javatpoint.com/big-data-characteristics

Big Data

- **Solutions**
  - **Hadoop**
    - Framework for distributed storage and processing.
    - Enables batch processing across clusters.
  - **Spark**
    - In-memory processing for faster analytics.
    - Supports real-time and iterative processing.



Reference: https://www.netsolutions.com/insights/hadoop-vs-spark/