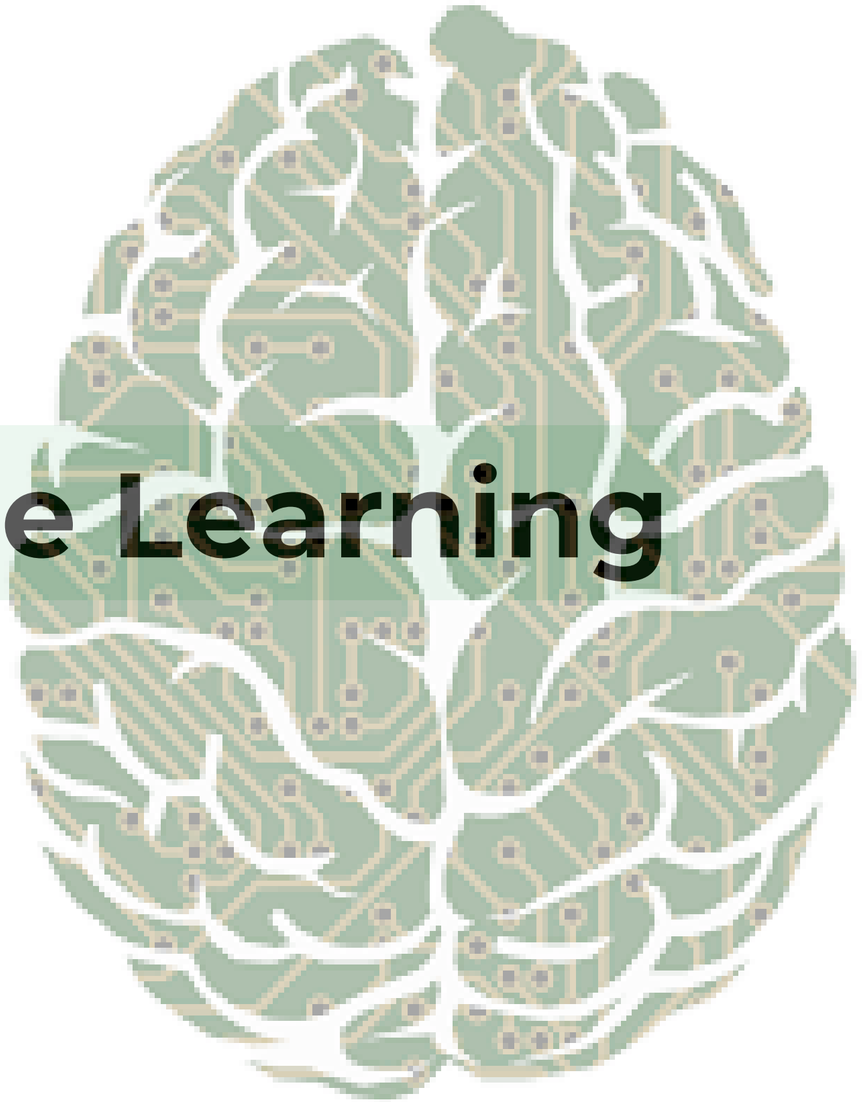
Introduction to Machine Learning

Pattern Recognition

PART 1



Mohammad Obeidat

PART 1

- Machine learning is all about finding patterns in data
- Looking for patterns
- Patterns in numerical data
- Patterns in images
- Patterns in text

COMPLETE THE PATTERNS

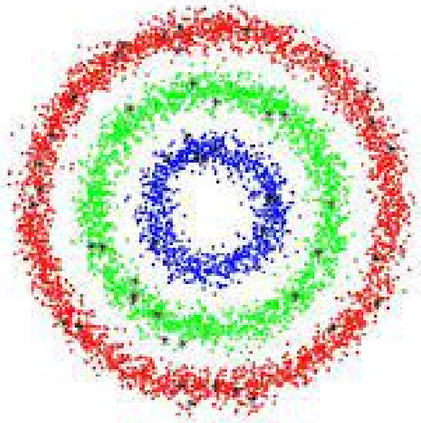
1, 5, 9, 13, 17, 21, ...

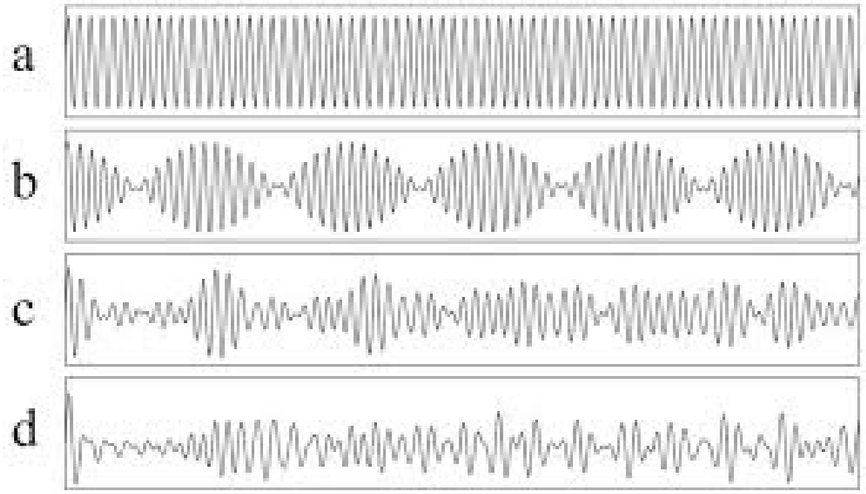
1, 2, 4, 8, 16, ...

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, ...

WHAT THIS IS ALL ABOUT

Machine Learning is all about finding pattens in the dataset.





PATTERNS IN NUMERICAL DATA

x: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

y: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20

if x = 10, y=?

if x = 11, y=?

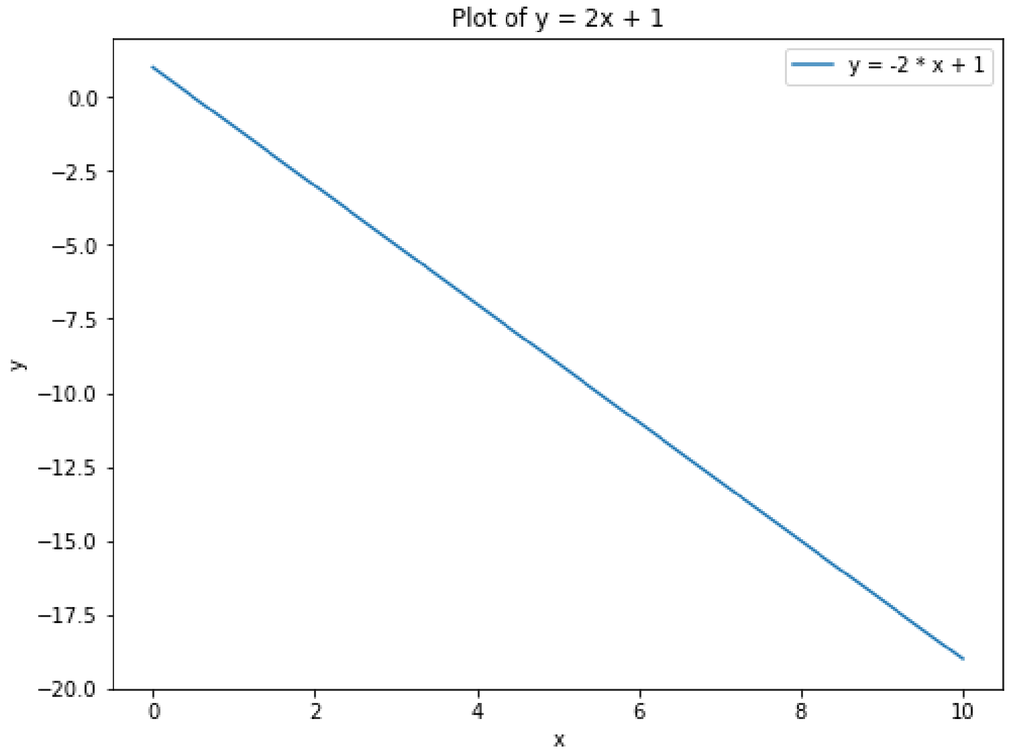
PATTERNS IN NUMERICAL DATA

x: 0, 1, 2, 3, 4, 5, 6, 7

y: 1, -1, -3, -5, -7, -9, -11, -13,

if x = 8, y=?

if x = 9, y=?



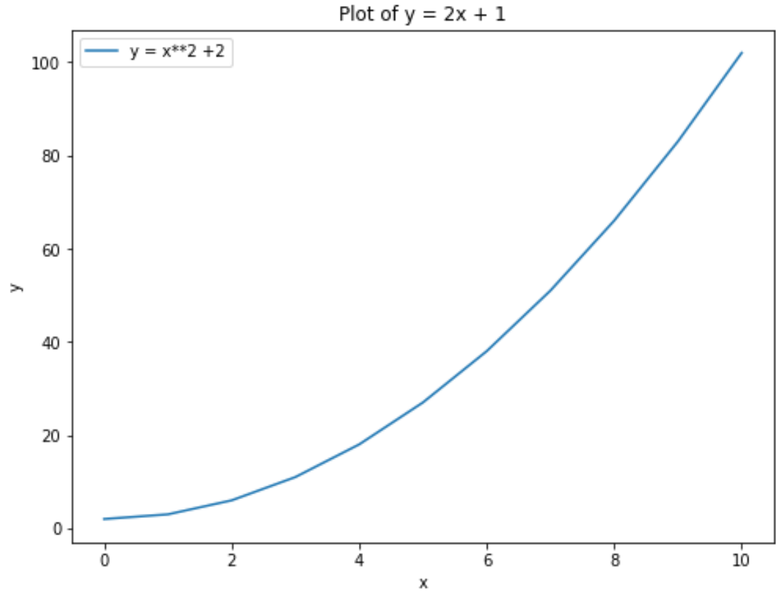
PATTERNS IN NUMERICAL DATA

x: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

y: 2, 3, 6, 11, 18, 27, 38, 51, 66, 83

if x = 10, y=?

if x = 11, y=?



PATTERNS IN VISUAL DATA

People

Sad People

Happy People

PATTERNS IN TEXTUAL DATA

English (Subject-Verb-Object):

- She (subject) reads (verb) a book (object).

French (Subject-Object-Verb):

- Elle (subject) lit (object) un livre (verb).
    - She reads a book

IT’S ALL NUMBERS

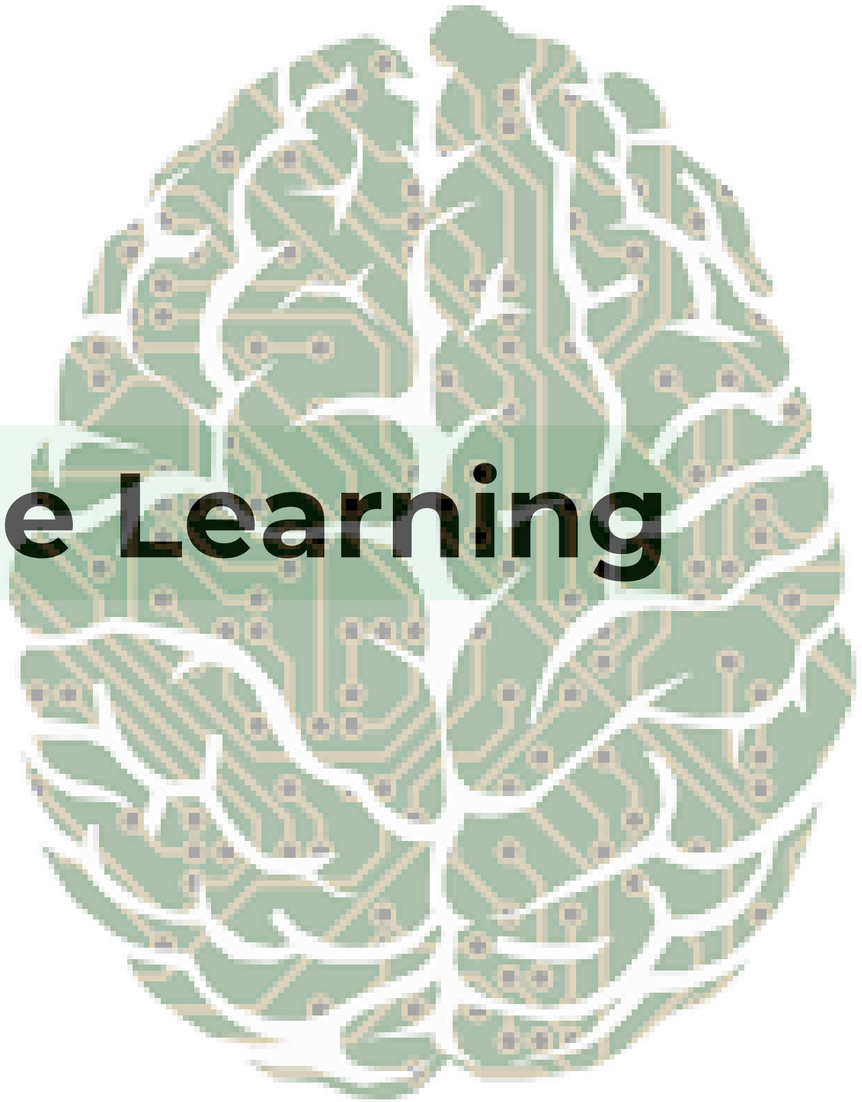
For the machine, all data types are numbers and everything is in binary code.

Texts, images, audio and videos must be represented as numbers before being fed into a machine learning model

Introduction to Machine Learning

It’s Just Statistics!

PART 2



Mohammad Obeidat

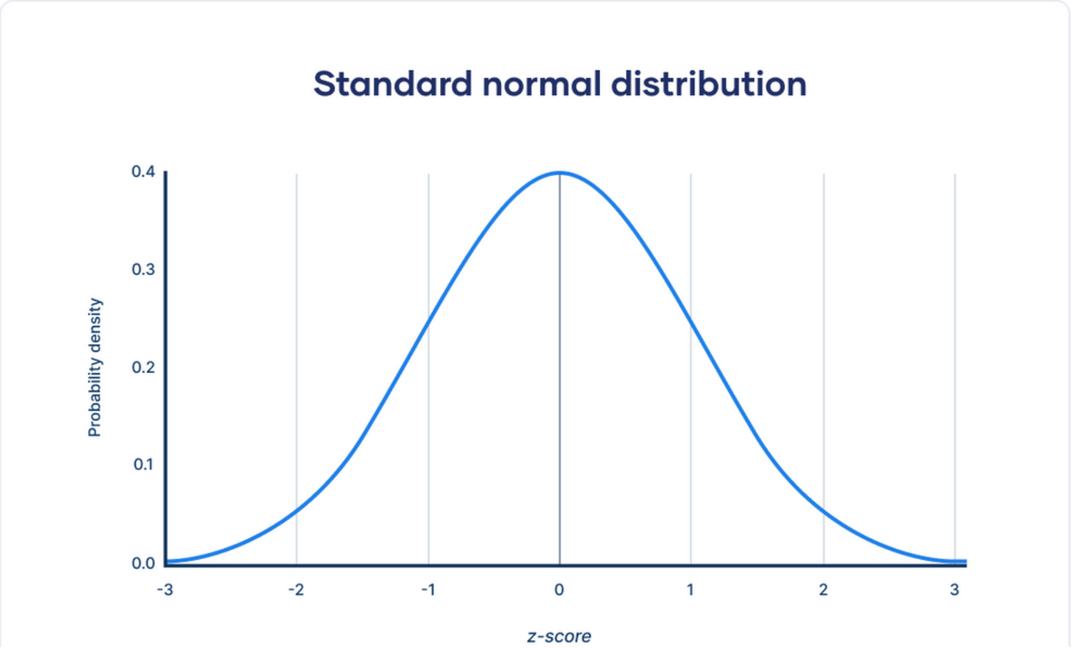
PART 1

- Statistics as the Foundation
- Probability Distributions
- Stats (mean, std, mode, variance) as patterns
- Machine Learning is just fancy statistics?

WHAT IS A PROBABILTY DISTRIBUTION

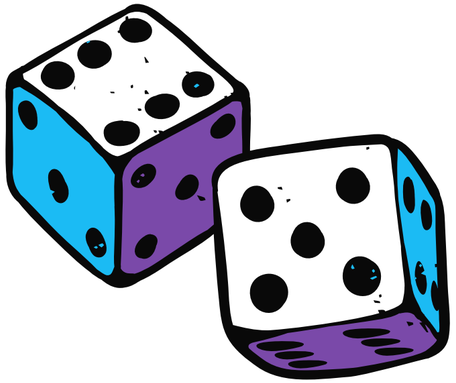
- A probability distribution assigns probabilities to possible outcomes in a random event or experiment.

- It provides a way to quantify uncertainty and understand the likelihood of different results.



EXAMPLES

$$\text{Probability of Event} = \frac{\text{No of Possibility of Event}}{\text{No of Total Possibility}}$$



Probability of getting any side on the dice is :
1/6

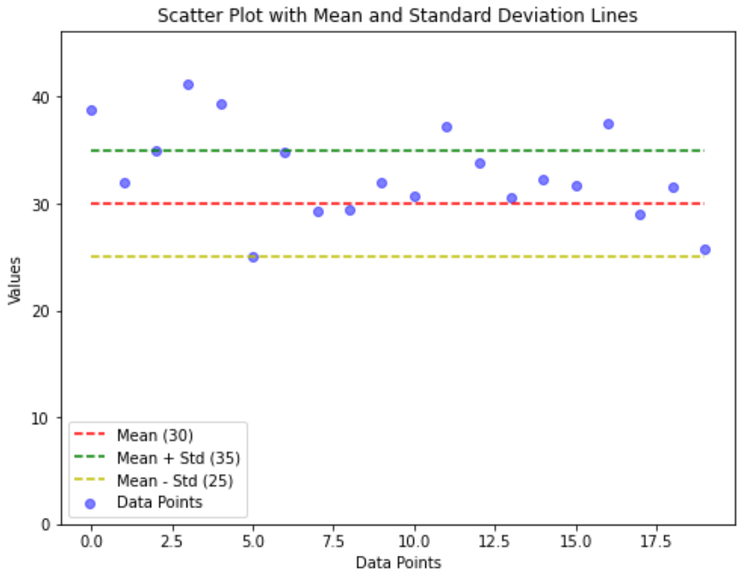
Probability of getting any face on the coin is :
1/2

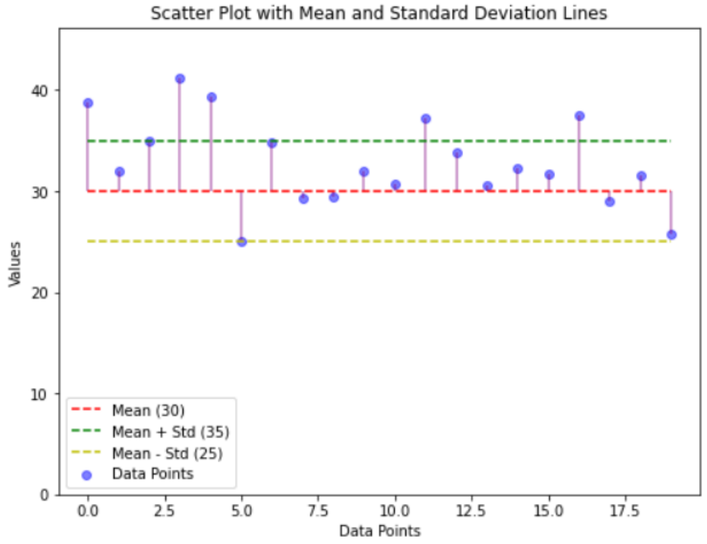
BASIC STATISTICS

- **Mean**: The average of all data points

- **Standard Deviation (Std)**: The standard deviation measures the spread or dispersion of data points in a dataset.

- **Mode**: The mode is the value or values in a dataset that appear most frequently.

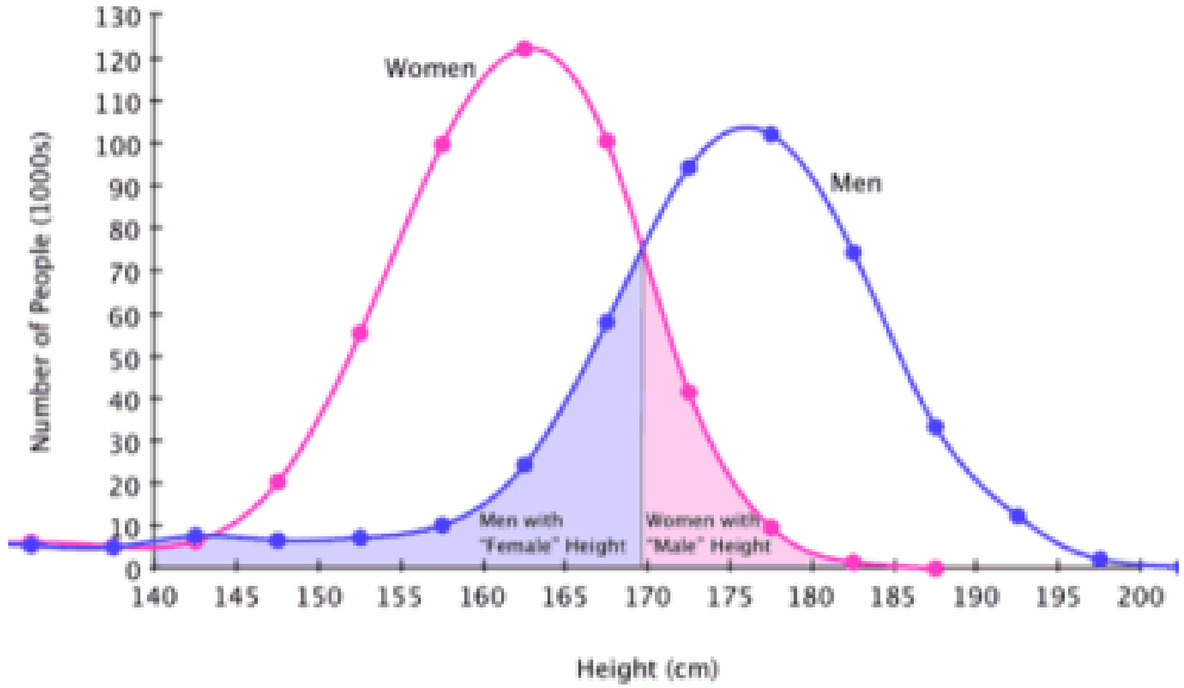




BASIC STATISTICS

If we take a sample of people from both genders and measure their heights, we can conclude that **on average** men are taller than women.

This does not provide any information about the height of every individual in the sample but rather general insight about each group.

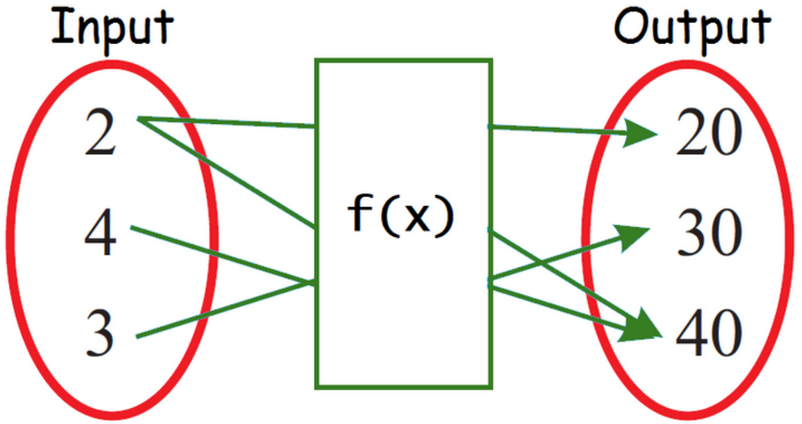


PROBABILITY ESTIMATION

- Training an ML model is to learn the conditional probability distribution of the **Target** variable given the **Input** variable(s).

- This probability distribution represents the relationship between the **Input** and the **Target** variables.

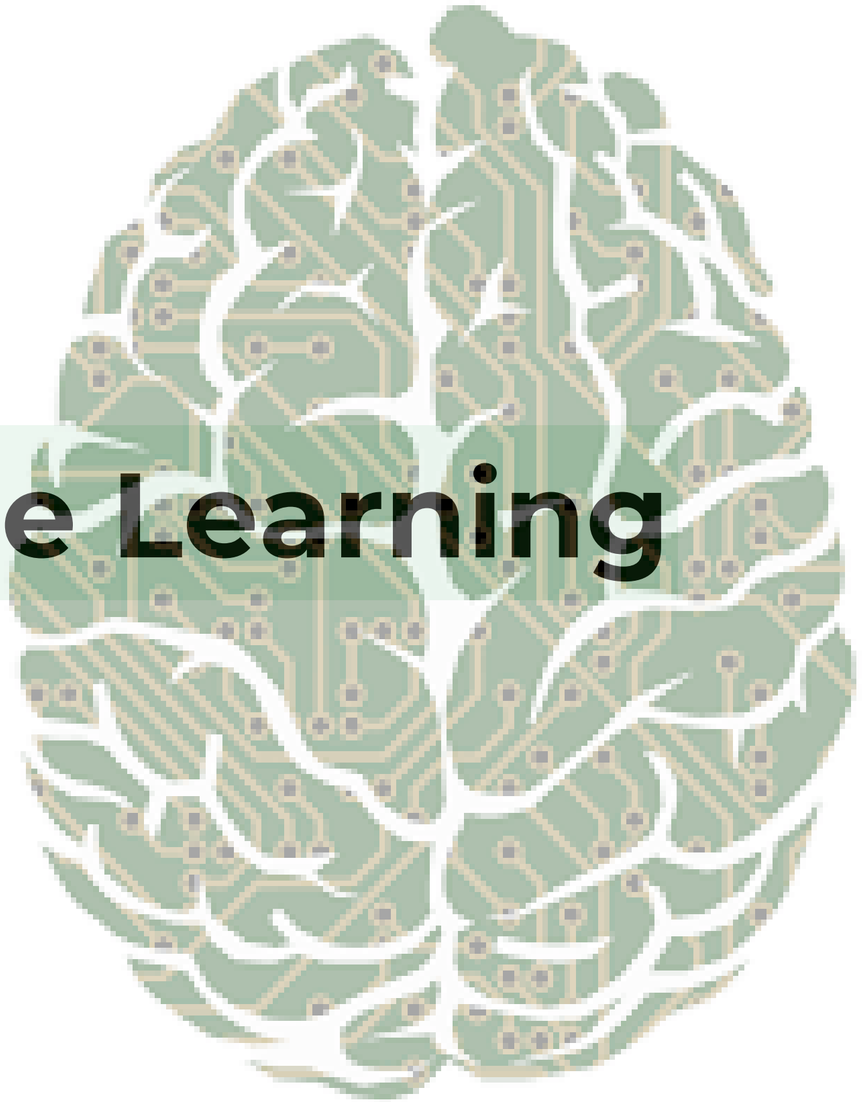
- $P(y = 1 | X = \theta) = f(x)$: given that the input variables X equals to some value θ, what is the probability of Y = 1.



Introduction to Machine Learning

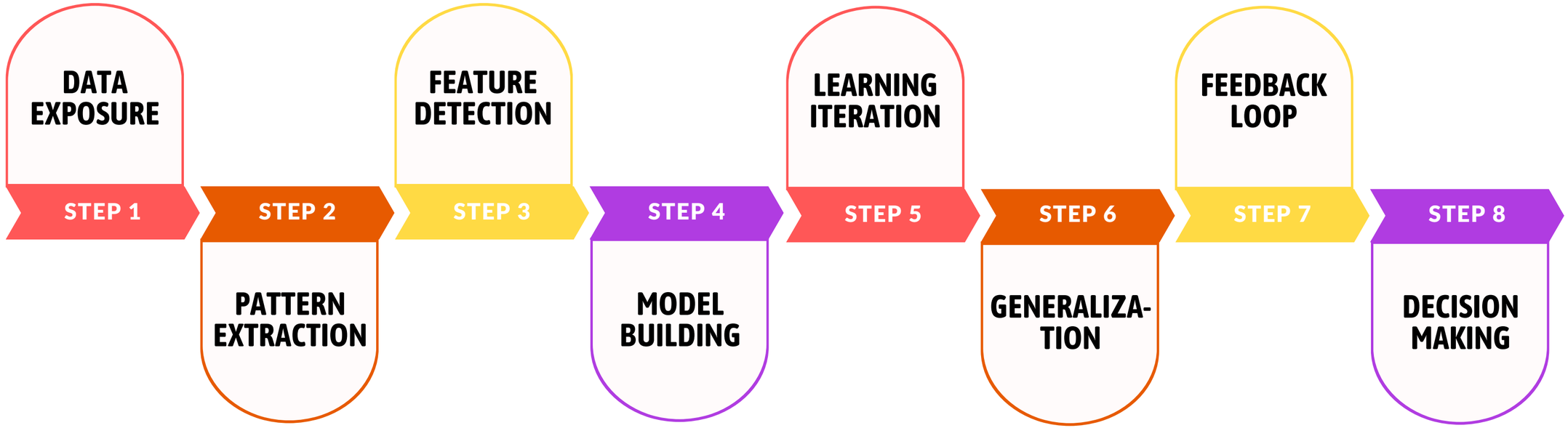
How It's Done

PART 3



Mohammad Obeidat

HOW DOES THE MACHINE LEARN



DATA EXPOSURE

- Machines learn by being exposed to large volumes of data, which serve as their source of knowledge. This data can encompass various types, such as text, images, numbers, or any information relevant to the problem at hand.

PATTERN EXTRACTION

- When exposed to this data, machines employ mathematical algorithms to identify patterns, relationships, and regularities within the information. This is akin to humans recognizing trends or correlations in data through observation.

FEATURE DETECTION

- Within the data, machines detect features or attributes that are indicative of the desired outcome or target. For instance, in image recognition, features might be edges, colors, shapes, or textures.

MODEL BUILDING

- Machines construct models or mathematical representations of the patterns and features they've discerned. These models are used to make predictions, classify data, or perform other tasks based on new, unseen data.

LEARNING ITERATION

- The learning process is iterative. Machines continually refine their models as they encounter more data and gather additional experiences. This ongoing refinement allows them to improve their accuracy and adapt to changing circumstances.

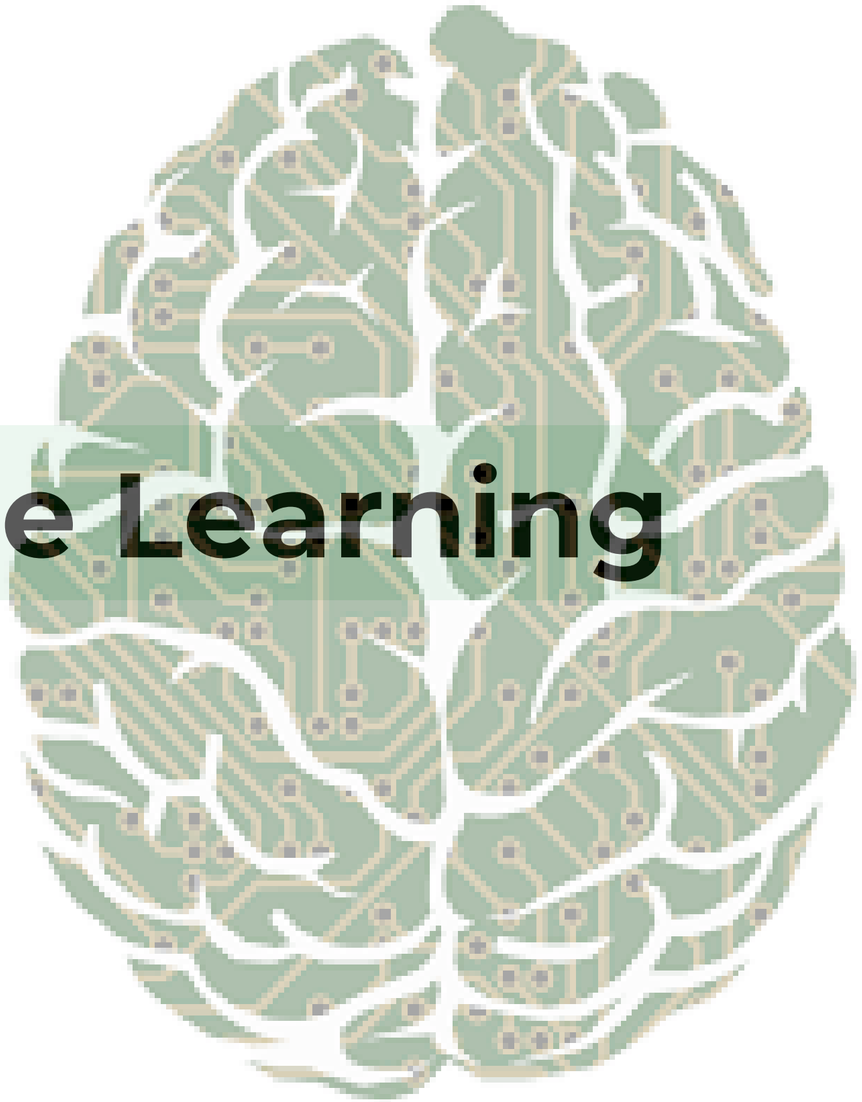
GENERALIZATION

- One of the key objectives of machine learning is to generalize from the data. In other words, machines aim to learn not just from the specific examples they've seen but to extract underlying principles that can be applied to new, unseen data.

DECISION-MAKING

- Once trained, machines can use their learned models to make predictions or decisions autonomously when presented with new, real-world data. This is where the true application of machine learning comes into play.

Introduction to Machine Learning



What problems can be solved with it

PART 4

Mohammad Obeidat

TYPES OF MACHINE LEARNING

- Supervised ML
- Unsupervised ML
- Semi-Supervised ML
- Reinforcement ML

SUPERVISED ML

- In supervised learning, the algorithm learns from a labeled dataset, where each input data point is associated with a corresponding target or label.

- The goal is to learn a mapping from **inputs** to **outputs**, enabling the algorithm to make predictions on new, unseen data.

- Common algorithms: Linear Regression, Decision Trees, Support Vector Machines, Neural Networks, etc.

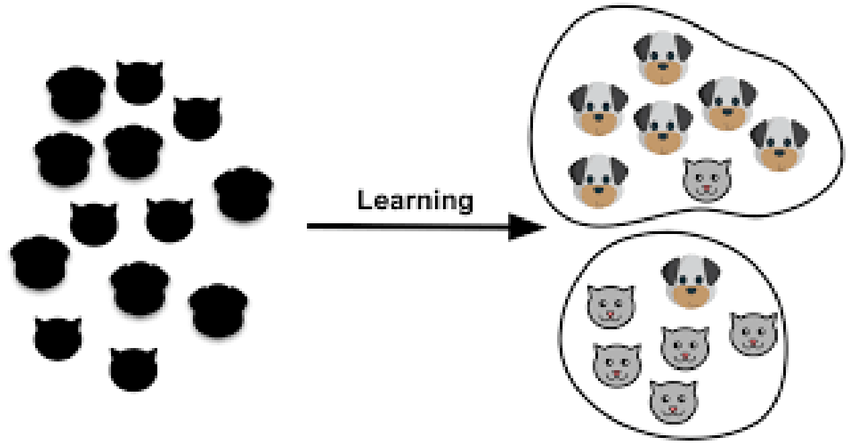
SUPERVISED ML - APPLICATIONS

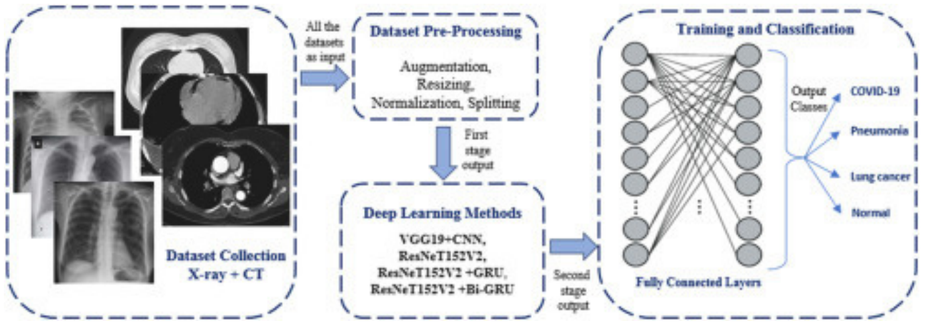
- Regression
- Classification

SUPERVISED ML - USE CASE

Image Classification

Supervised learning is widely used in image classification tasks, where algorithms are trained on labeled images to recognize objects, scenes, or patterns within the images.

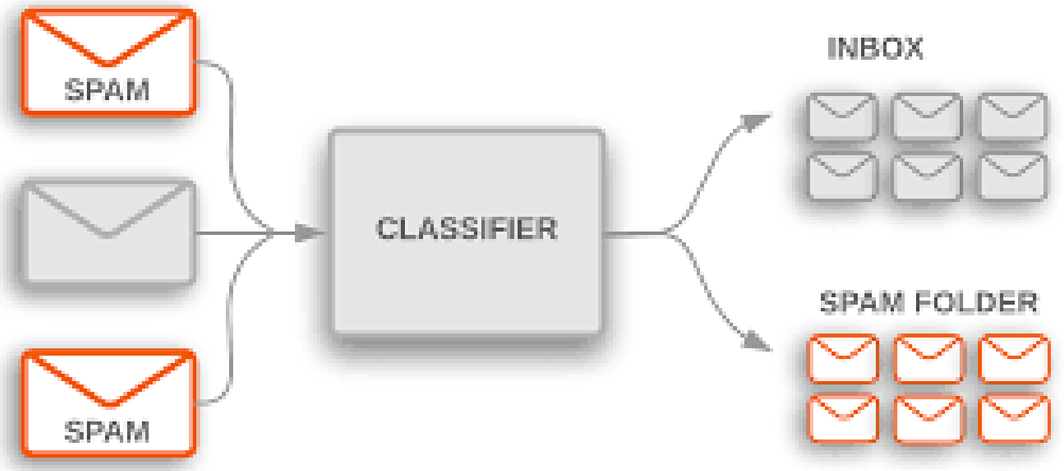




SUPERVISED ML - USE CASE

Spam Email Detection:

In email classification, supervised learning algorithms are trained on a dataset of emails that are labeled as spam or not spam. They learn to distinguish between the two categories and can filter out unwanted emails.



SUPERVISED ML - USE CASE

Predicting House Prices

Supervised regression is used to predict continuous numeric values, such as house prices. The algorithm learns from historical data where each house is associated with a sale price.

UN-SUPERVISED ML

- Unsupervised learning involves training the algorithm on an unlabeled dataset, where there are no predefined output labels.

- The algorithm explores the data to discover patterns, relationships, or structures within it.

- Common algorithms: K-Means Clustering, Principal Component Analysis (PCA), Autoencoders, etc.

UN-SUPERVISED ML - APPLICATIONS

- Clustering
- Anomaly Detection
- Topic Modeling

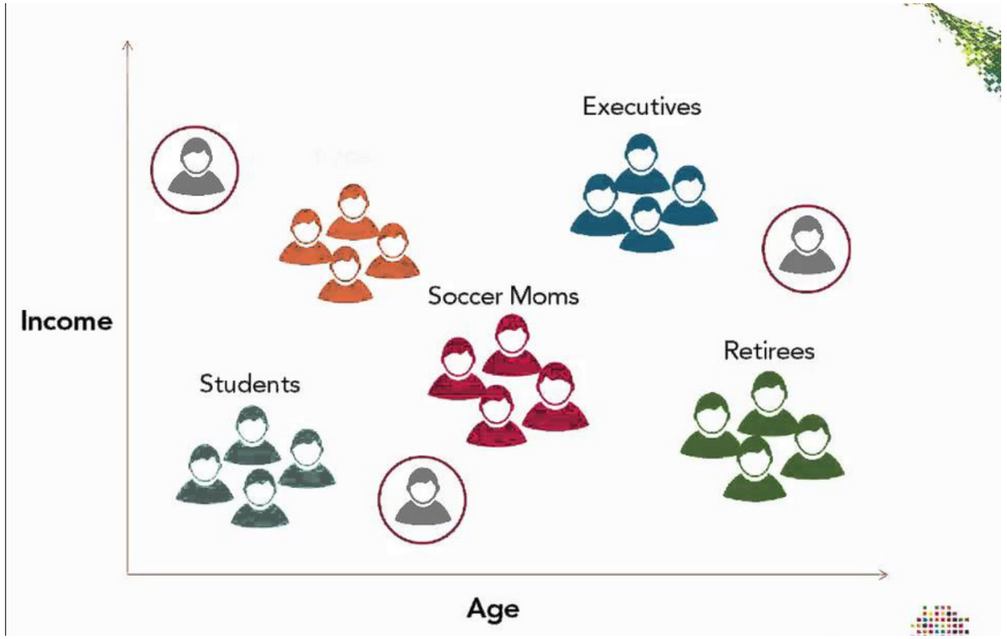
UN-SUPERVISED ML - USE CASE

K-Means Clustering for Customer Segmentation:

Imagine you work for a retail company, and you have a dataset of customer purchase histories. Each row in the dataset represents a customer, and the columns represent various features such as:

- Total amount spent
- Number of items purchased
- Frequency of visits to the store
- Average purchase amount
- etc.

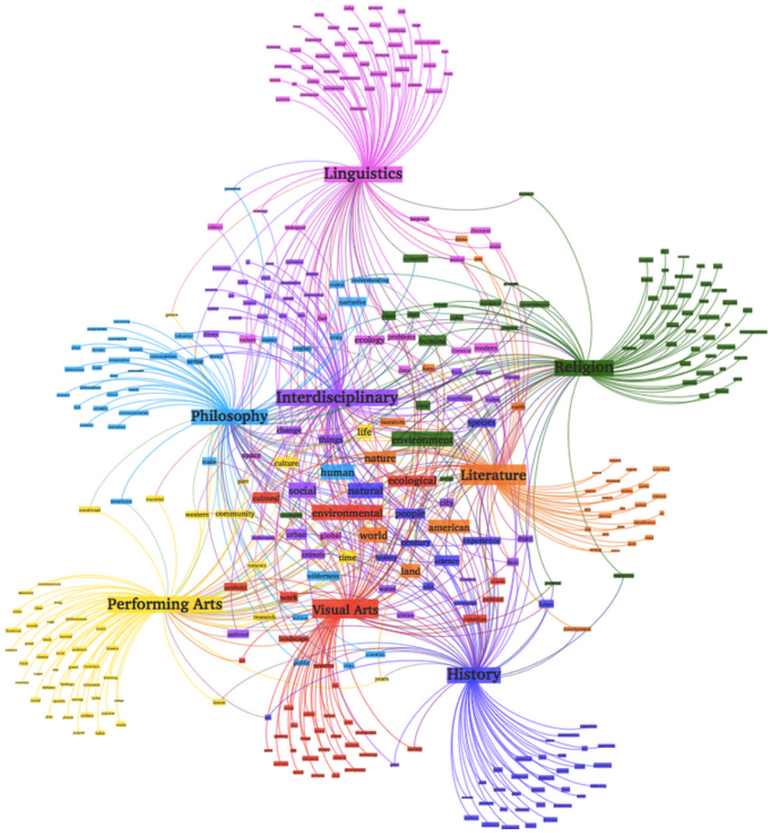
The goal is to segment these customers into distinct groups based on their purchasing behavior without any predefined categories or labels.



UN-SUPERVISED ML - USE CASE

Document Clustering for Topic Modeling

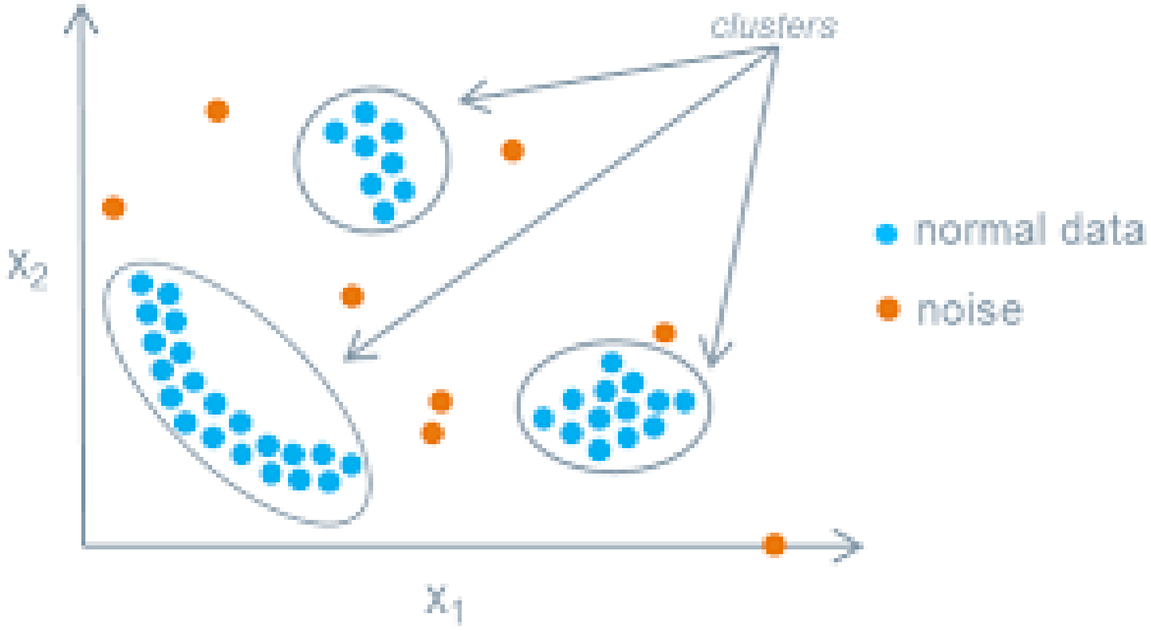
Imagine you have a large collection of text documents, such as news articles or customer reviews, and you want to uncover the underlying topics present in the these articles. You can group similar articles together and then assign labels for what might to common topic among each group be.



UN-SUPERVISED ML - USE CASE

Anomaly Detection: Credit Card Fraud Detection

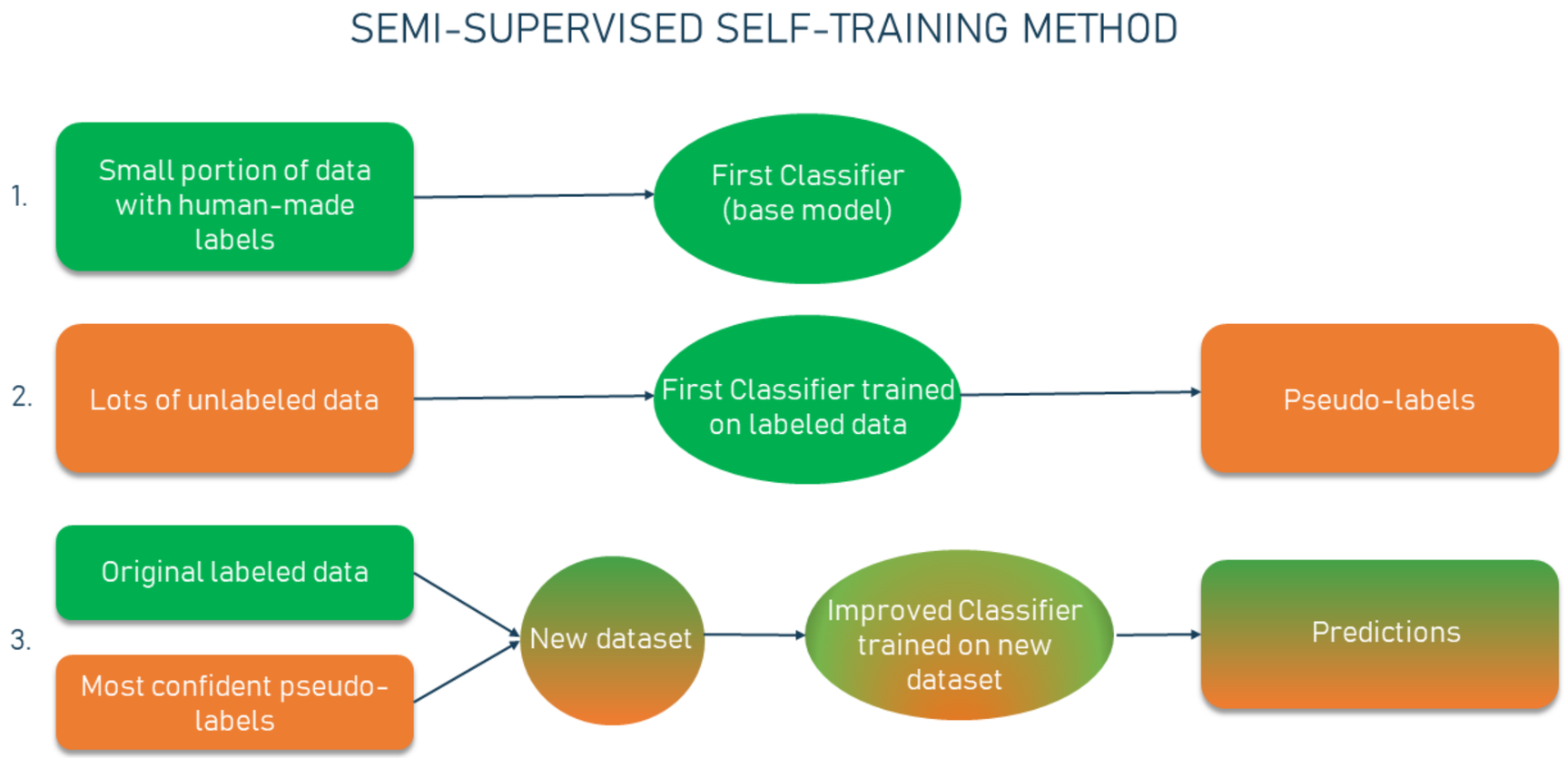
In the financial industry, credit card fraud is a significant concern. Unsupervised learning techniques can be employed to identify fraudulent transactions without the need for labeled examples of fraud.



SEMI-SUPERVISED LEARNING

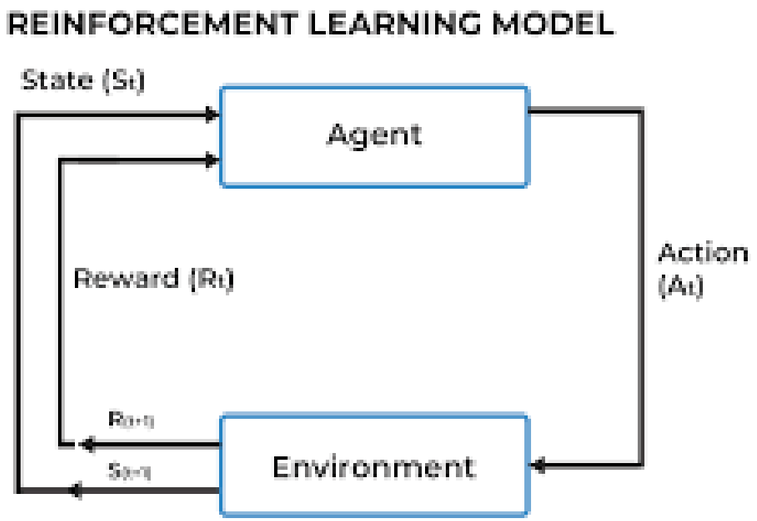
- Semi-supervised learning combines elements of both supervised and unsupervised learning.
- It uses a dataset that contains a mix of labeled and unlabeled data points.
- The goal is to leverage the limited labeled data to improve model performance on unlabeled data.
- Applications: Document classification with limited labeled data, sentiment analysis, etc.

SEMI-SUPERVISED LEARNING



REINFORCEMENT LEARNING

Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment. The agent takes actions to maximize a cumulative reward signal over time. It is a powerful approach used in various applications, including robotics, game playing, recommendation systems, and autonomous vehicles.



REINFORCEMENT LEARNING - COMPONENTS

- **Agent**: The agent is the learner or decision-maker that interacts with the environment. It takes actions based on its current state and seeks to maximize its long-term reward.
- **Environment**: The environment is the external system or process with which the agent interacts. It is the context in which the agent operates and where it receives feedback in the form of rewards.
- **State (S)**: The state represents the current situation or configuration of the environment, which the agent observes. The state is essential for the agent to make informed decisions.
- **Action (A)**: Actions are the choices or decisions that the agent can take at each time step. The set of possible actions depends on the specific problem.
- **Reward (R)**: The reward is a scalar signal that the agent receives from the environment after taking an action in a particular state. It indicates the immediate benefit or cost of the action.

ASSIGNMENT

Based on your understanding of today's topic, try to find a problem from your field of knowledge that can be solved with machine learning.

Submit your answers to following questions:

1. Short description of the problem
2. Why to use machine learning instead of conventional methods that are used in the industry?
3. What are the **input** and **output** variables?
4. Which type of machine learning should be used?