HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

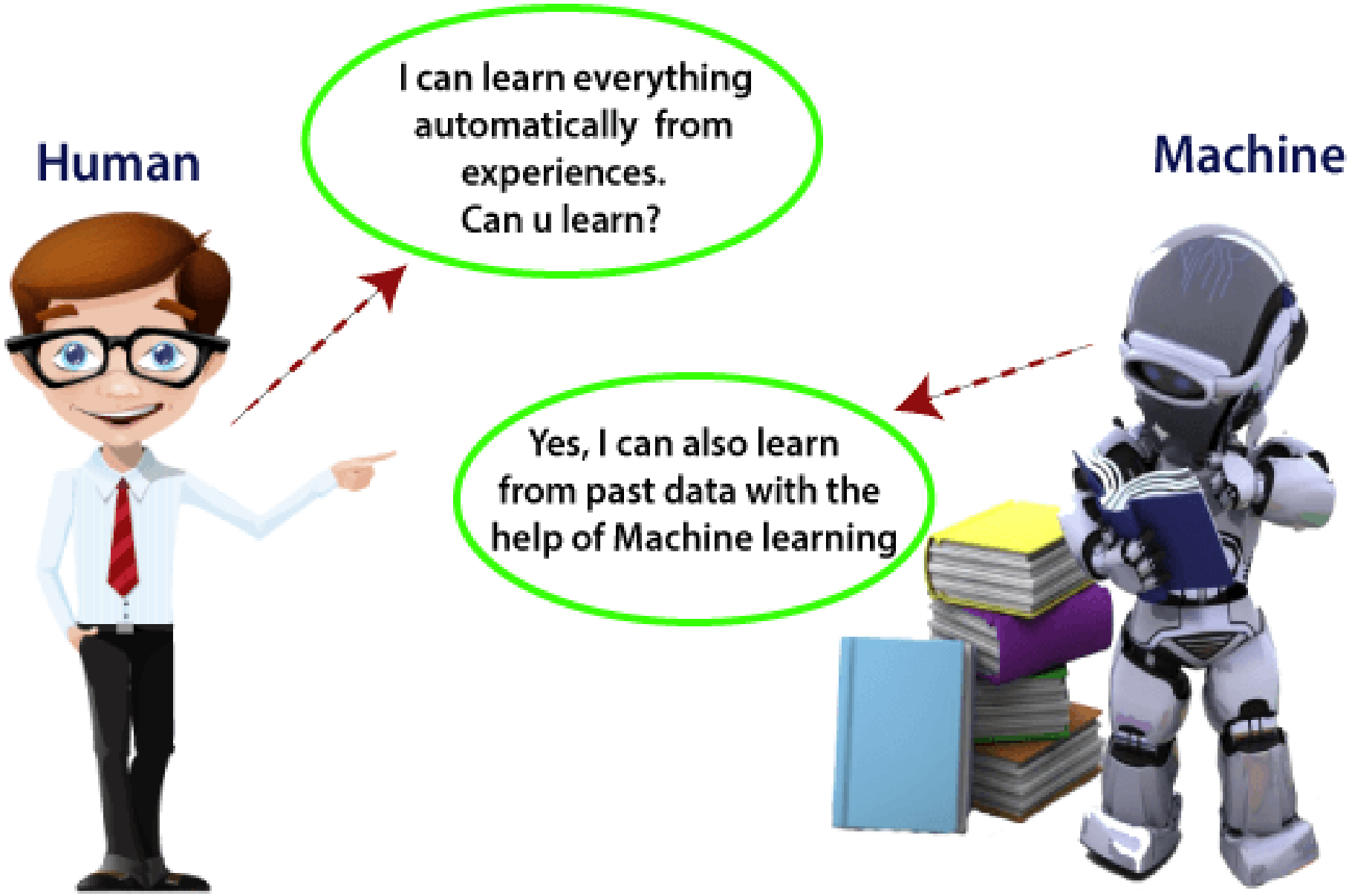
EAL51501 – ARTIFICIAL INTELLIGENCE
B.Tech[AIML] – III Semester

K.Kowsalya
Assistant Professor (SS)
School of Computing Sciences,
Department of Computer Science and Engineering

UNIT-III

- Motivation for Machine Learning, Applications, Machine Learning, Learning associations, Classification, Regression, The Origin of machine learning, Uses and abuses of machine learning, Success cases, How do machines learn, Abstraction and knowledge representation, Generalization, Factors to be considered, Assessing the success of learning, Metrics for evaluation of classification method, Steps to apply machine learning to data, Machine learning process, Input data and ML algorithm, Classification of machine learning algorithms, General ML architecture, Group of algorithms, Reinforcement learning, Supervised learning, Unsupervised learning, Semi-Supervised learning, Algorithms, Ensemble learning, Matching data to an appropriate algorithm.

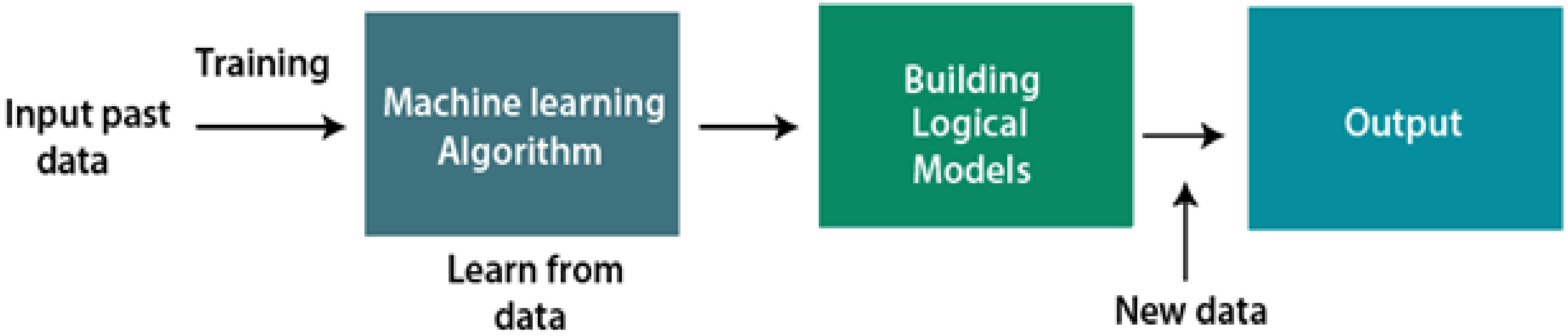
- **MACHINE LEARNING**
- It is a growing technology which enables computers to learn automatically from past data.
- Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**.
- Currently, it is being used for various tasks such as **image recognition**, **speech recognition**, **email filtering**, **Facebook auto-tagging**, **recommender system**, and many more.


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

- *"Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. "*
- The term machine learning was first introduced by **Arthur Samuel** in **1959**.
- "Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed."

- With the help of sample historical data, which is known as **training data**, machine learning algorithms build a **mathematical model** that helps in making predictions or decisions without being explicitly programmed.
- Machine learning brings computer science and statistics together for creating predictive models.
- Machine learning constructs or uses the algorithms that learn from historical data.
- **A machine has the ability to learn if it can improve its performance by gaining more data.**

- **How does Machine Learning work**
- **A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it.** The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.
- Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output.


Training
Input past data
Machine learning Algorithm
Learn from data
Building Logical Models
New data
Output

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

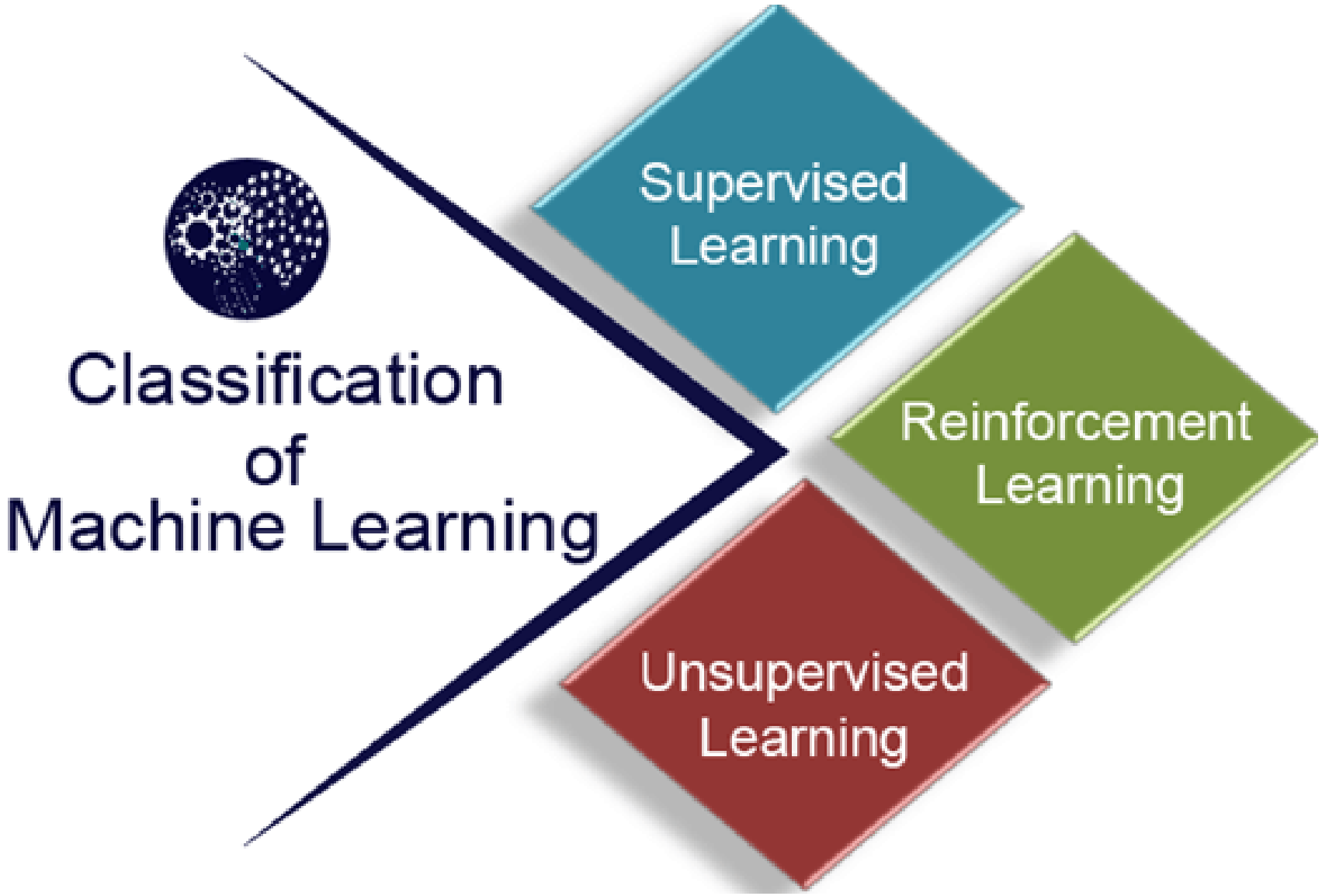
- The importance of machine learning can be easily understood by its uses cases, **Currently, machine learning is used in self-driving cars, cyber fraud detection, face recognition, and friend suggestion by Facebook, etc.**
- Various top companies such as **Netflix and Amazon have build machine learning models** that are using a vast amount of data to analyze the user interest and recommend product accordingly.
- *Netflix using **Subscription video-on-demand** Model.*
- *Amazon using* **ecommerce market platform.**

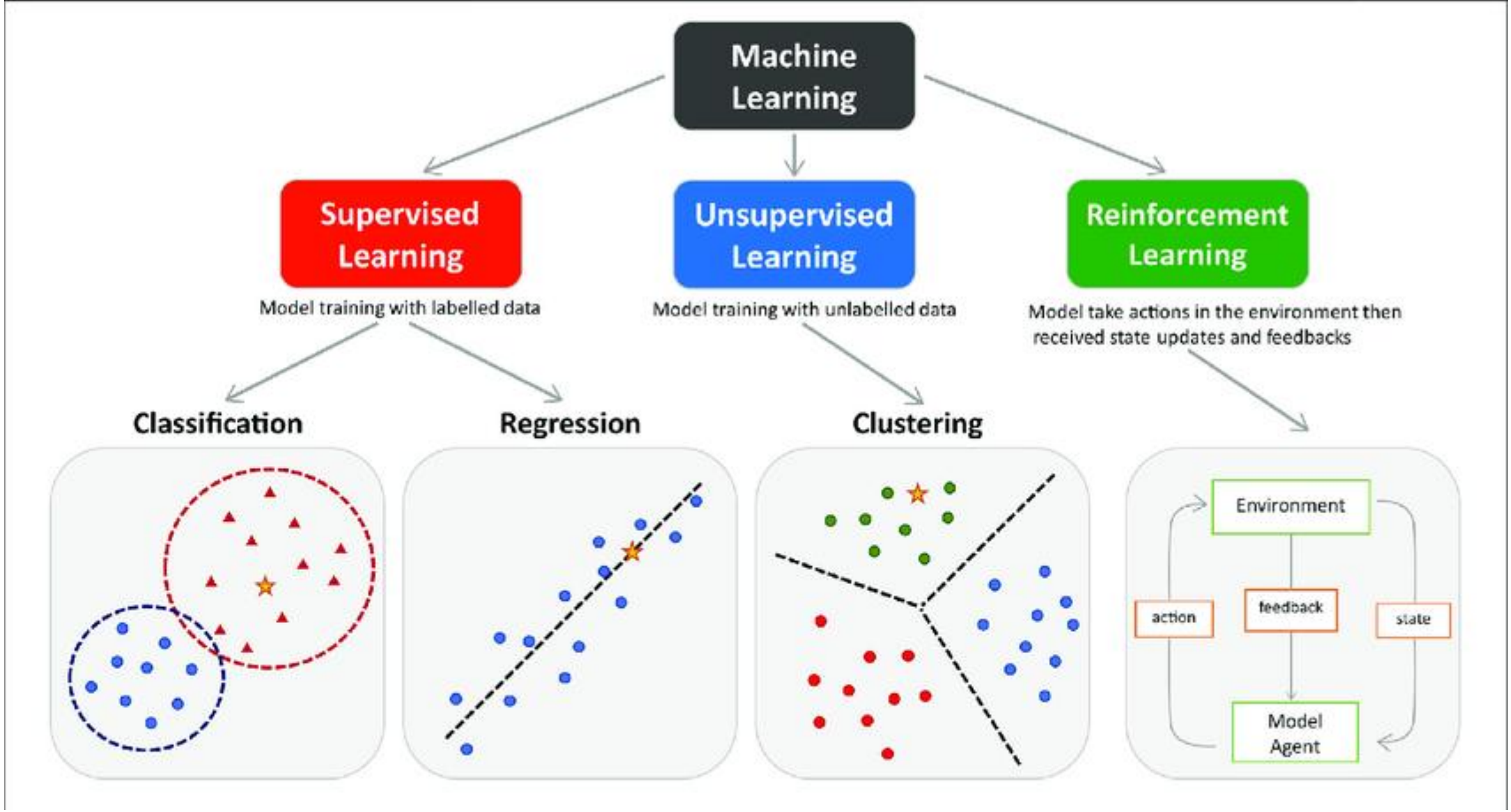
Following are some key points which show the importance of Machine Learning:

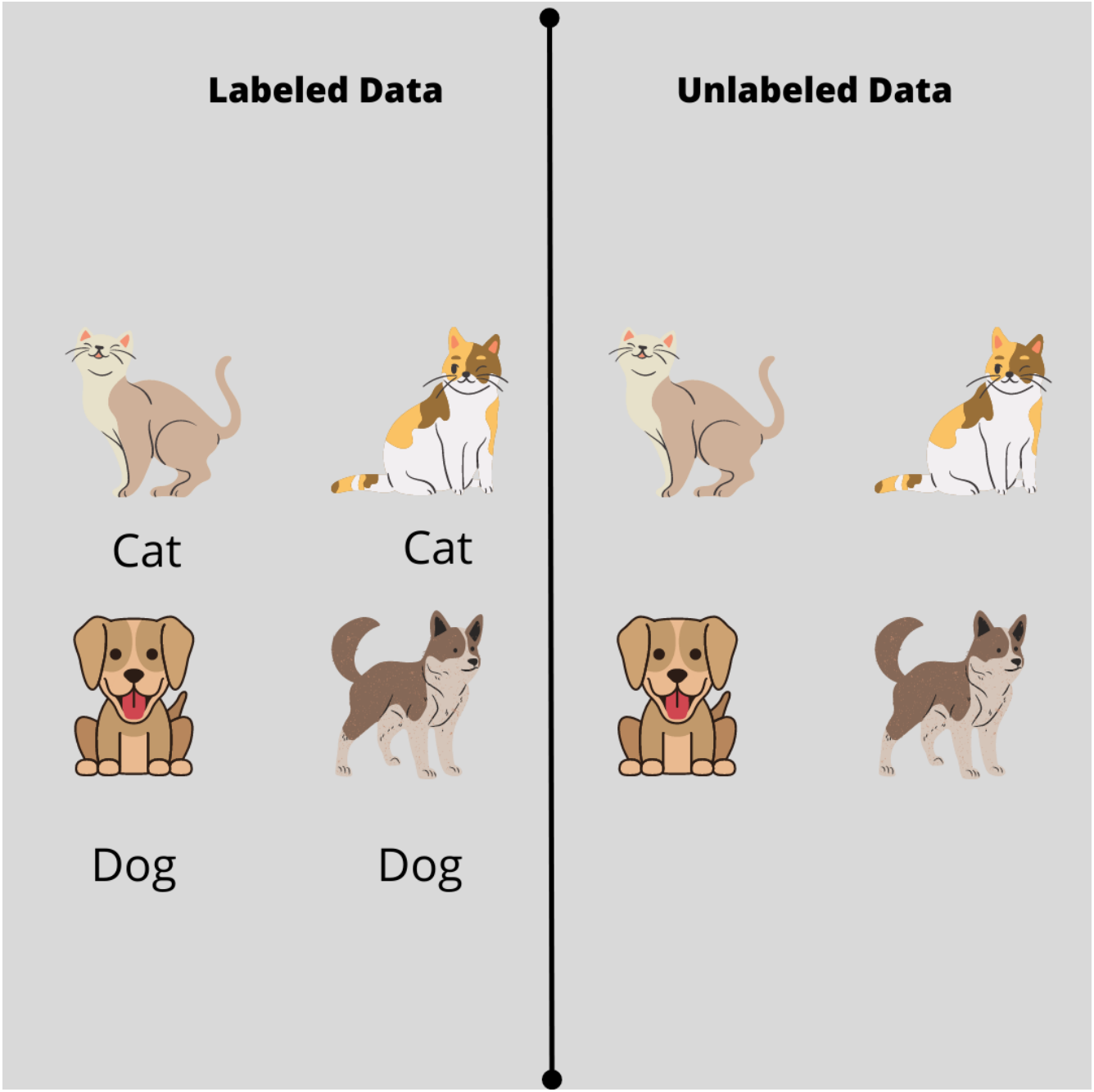
- Rapid increment in the production of data.
- Solving complex problems, which are difficult for a human.
- Decision making in various sector including finance.
- Finding hidden patterns and extracting useful information from data.

Classification of Machine Learning

- At a broad level, machine learning can be classified into three types:
- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement learning**


Classification
of
Machine Learning
Supervised Learning
Reinforcement Learning
Unsupervised Learning

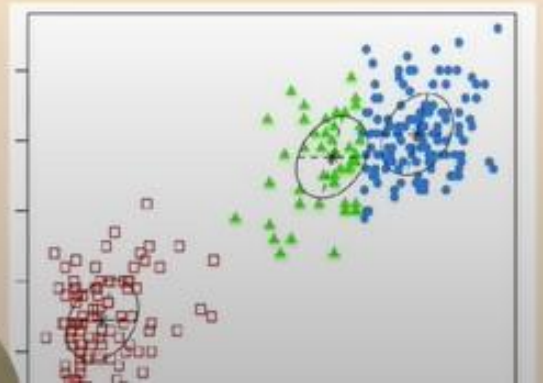

Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Model training with labelled data
Model training with unlabelled data
Model take actions in the environment then received state updates and feedbacks
Classification
Regression
Clustering
Environment
action
feedback
state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


Labeled Data
Unlabeled Data
Cat
Cat
Dog
Dog

Unsupervised Learning

- ➔ Unlabeled data X
- ➔ Learn X
- ➔ Generate fakes, insights

"This product does what it is supposed to. I always keep three of these in my kitchen just in case ever I need a replacement cord."



Supervised Learning

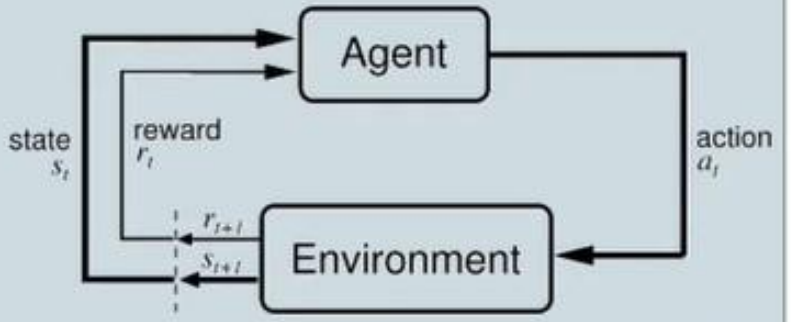
- ➔ Labeled data X and Y
- ➔ Learn X -> Y
- ➔ Make Predictions

→ *cat*

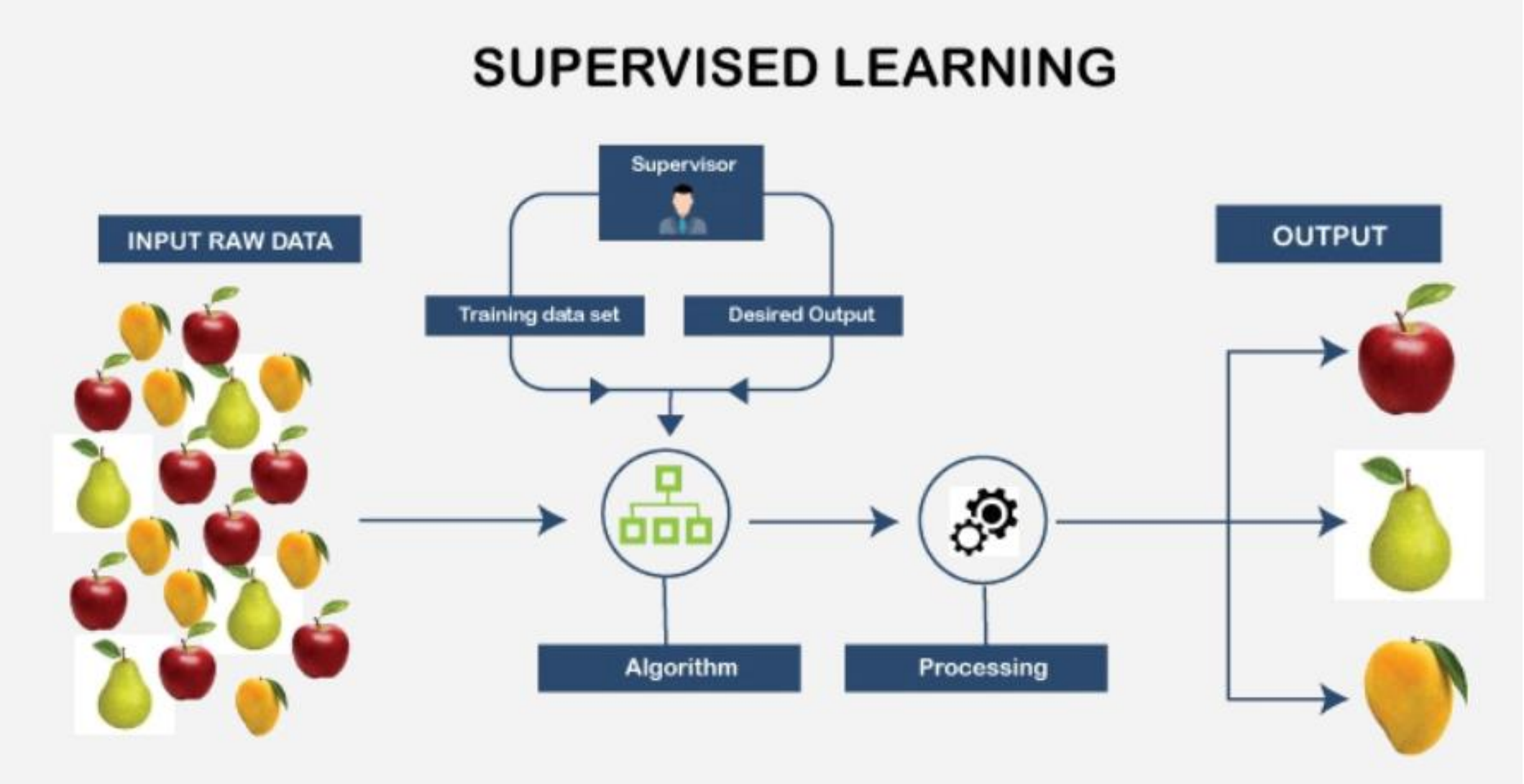
→ *"Hey Siri"*

Reinforcement Learning

- ➔ Learn how to take Actions in an Environment

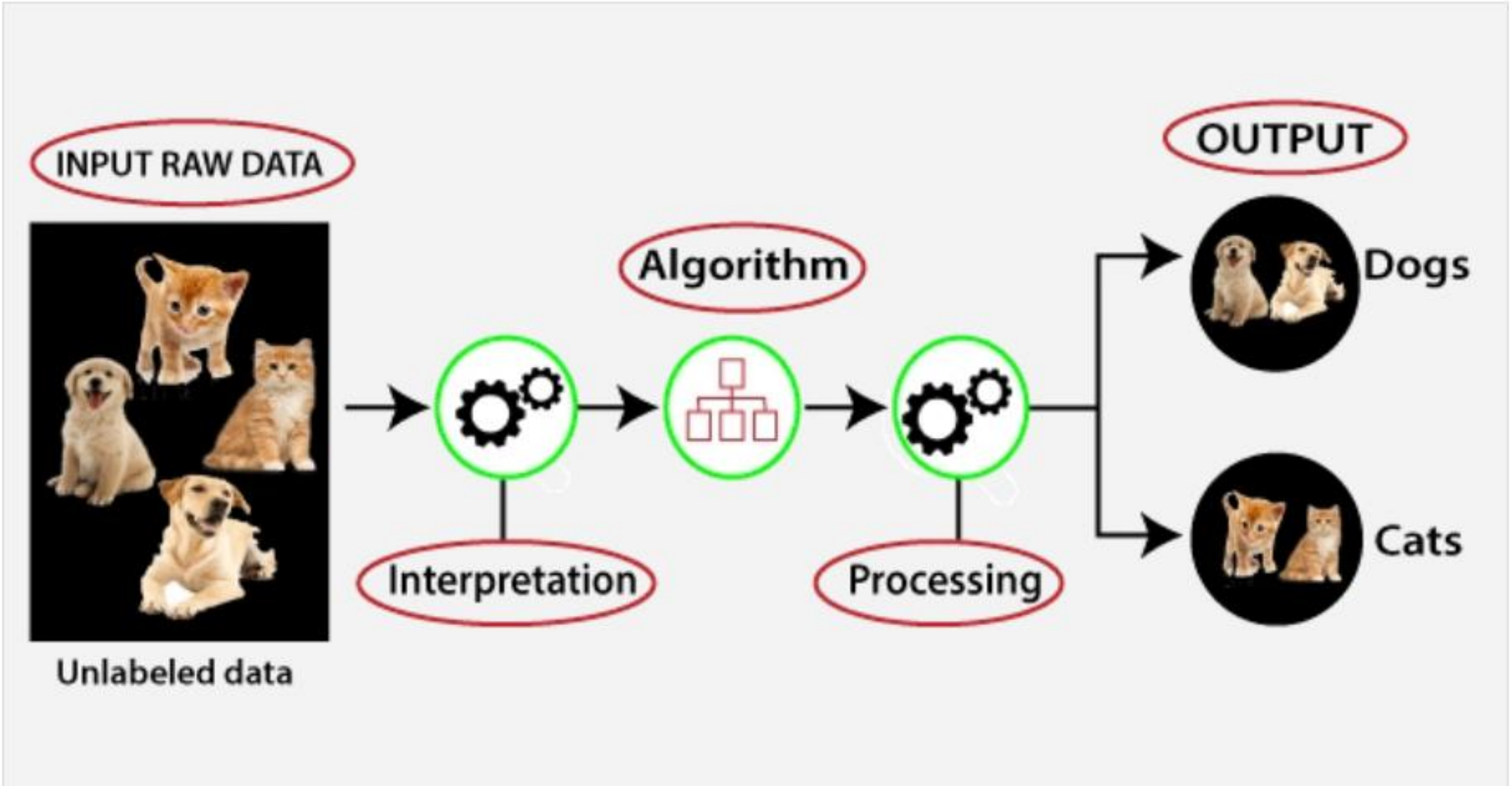


Supervised Learning

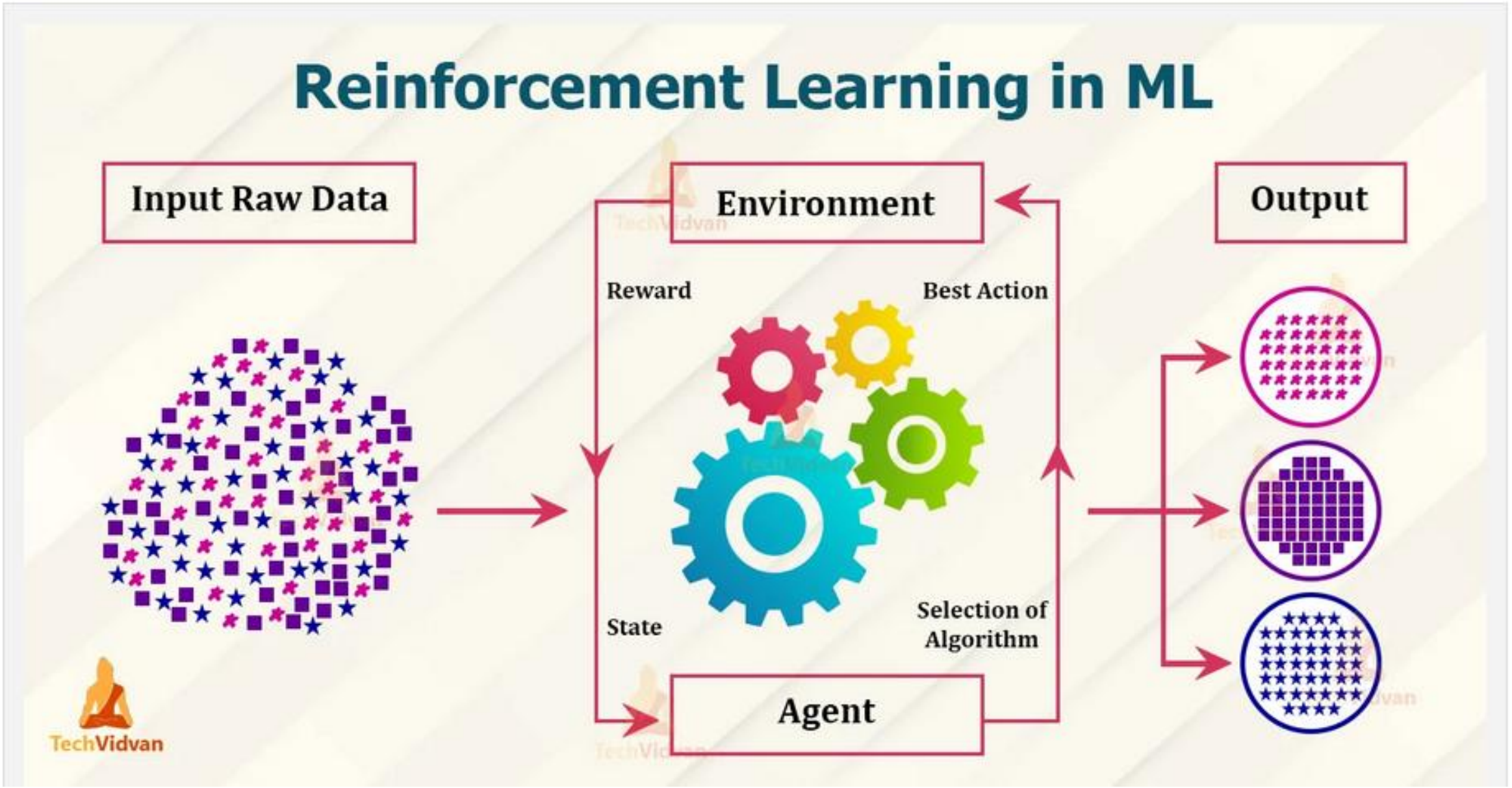


- In supervised learning, we feed the algorithm's output into the system so that the machine knows the patterns before working on them.
- In other words, **the algorithm gets trained on input data that has been labeled for a particular output.**
- The model undergoes training until it can detect the underlying patterns and relationships between the input data and the output labels, enabling it to yield accurate labeling results when presented with never-before-seen data.

Unsupervised learning

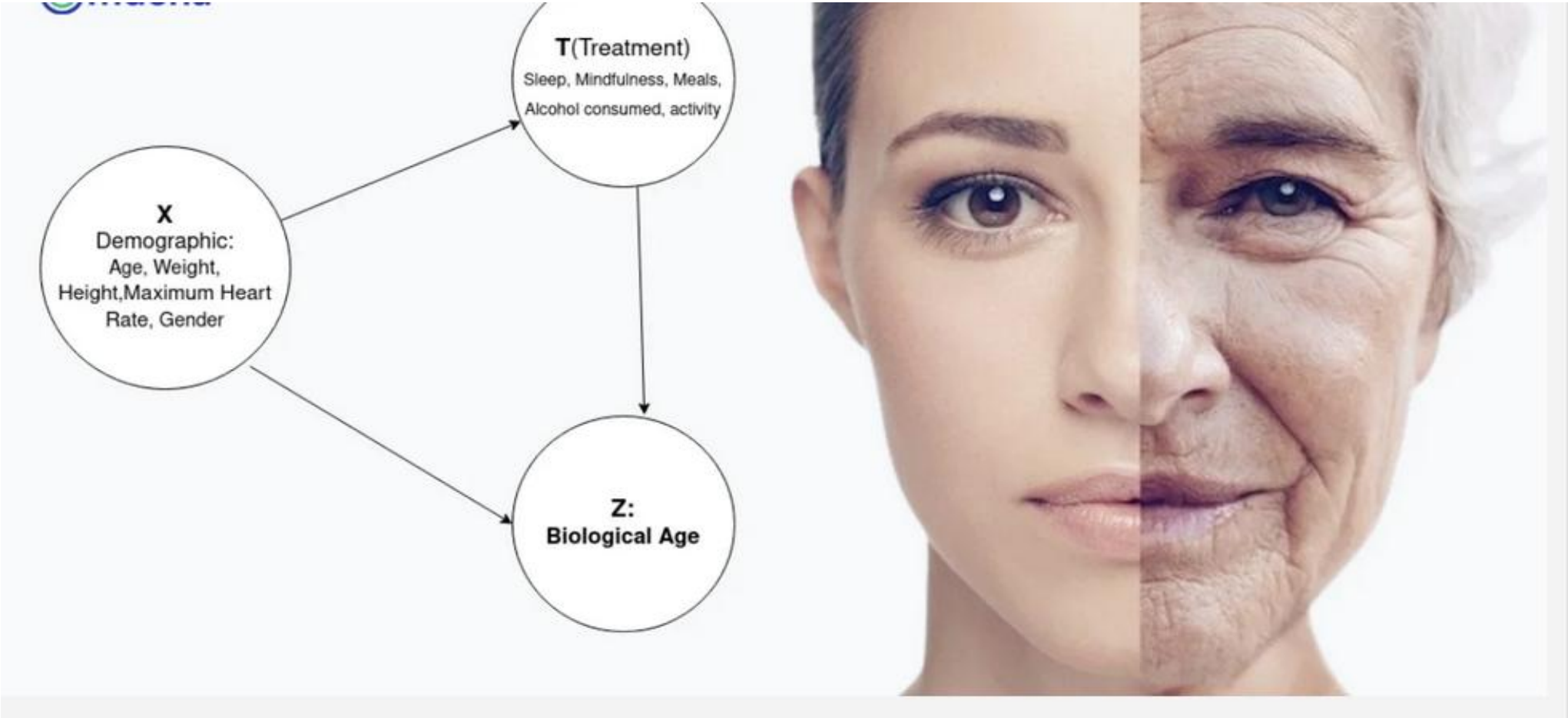


- The unsupervised learning approach is fantastic for uncovering relationships and insights in unlabeled datasets.
- Models feed input data with unknown desirable outcomes.
- So, inferences are made based on circumstantial evidence without training or guidance.
- Machine learning clustering examples fall under this learning algorithm.


Reinforcement Learning in ML
Input Raw Data
Environment
Output
Reward
Best Action
State
Selection of Algorithm
Agent
TechVidvan

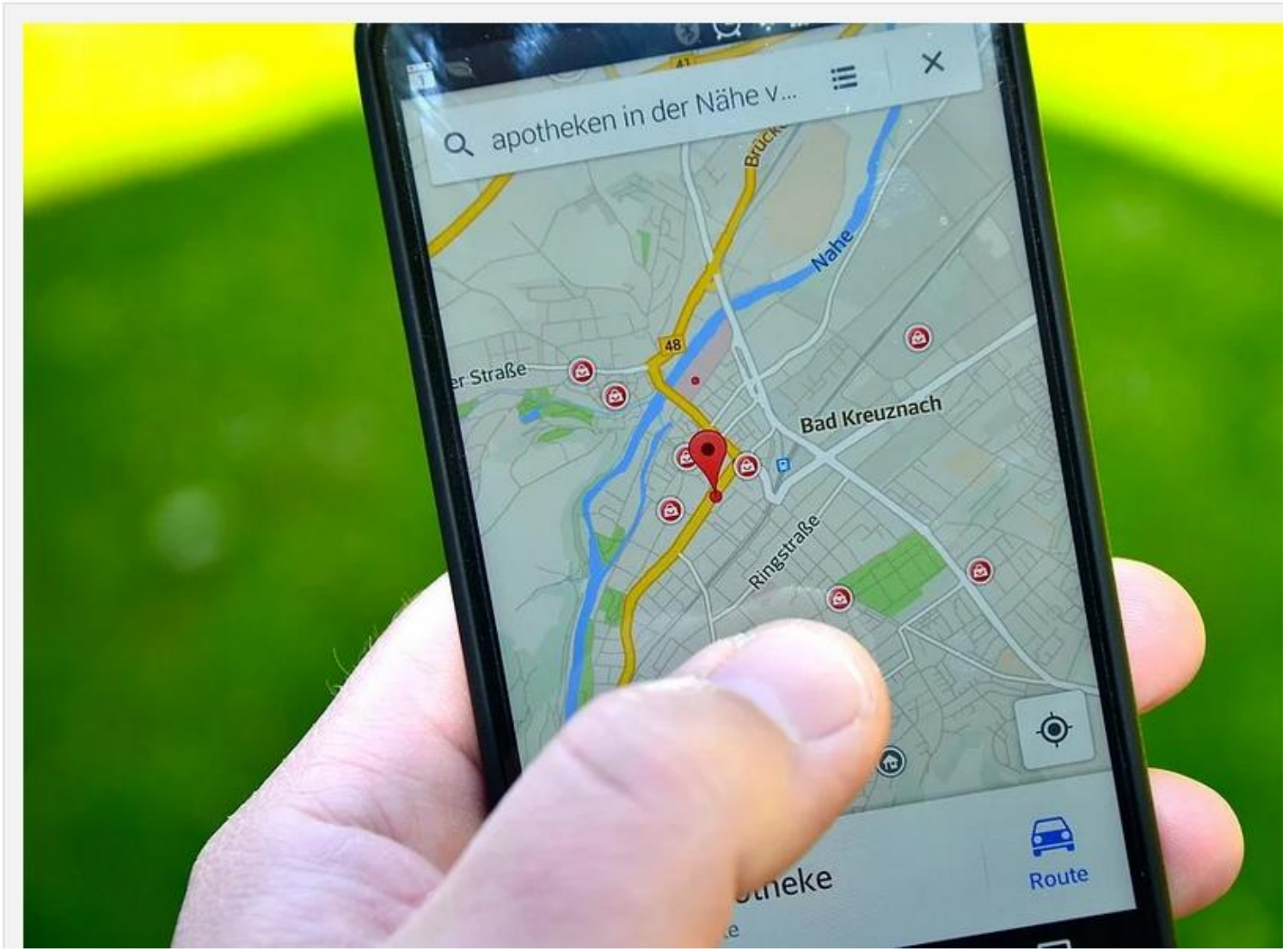
- The reinforcement learning approach in machine learning determines the best path or option to select in situations to maximize the reward.
- Key machine learning examples in daily life like video games, utilize this approach.
- Apart from video games, robotics also uses reinforcement models and algorithms.

- **Top 10 examples of machine learning in real life (which make the world a better place)**
- **1. Healthcare and medical diagnosis**



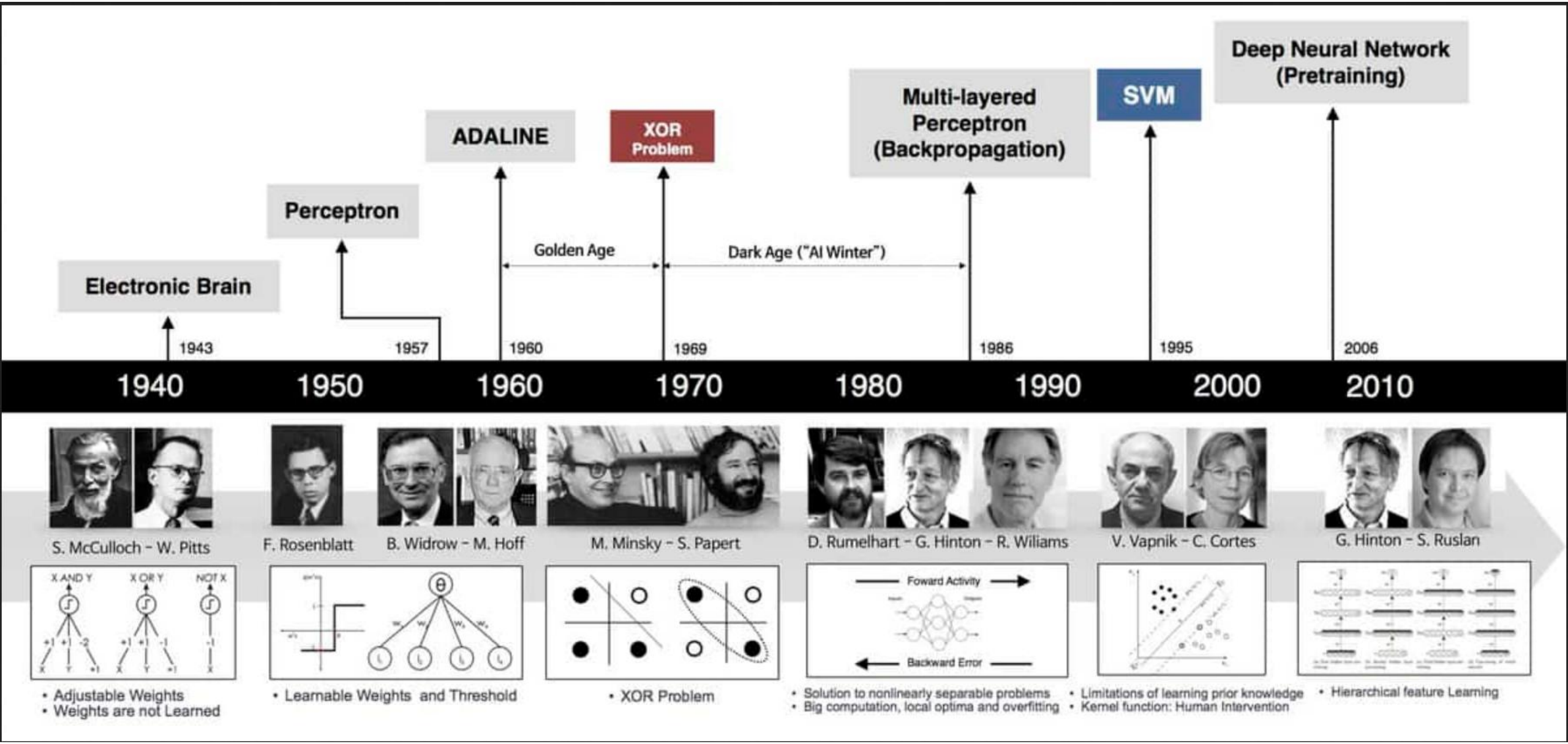
2. Face detection in images

3. Commute predictions

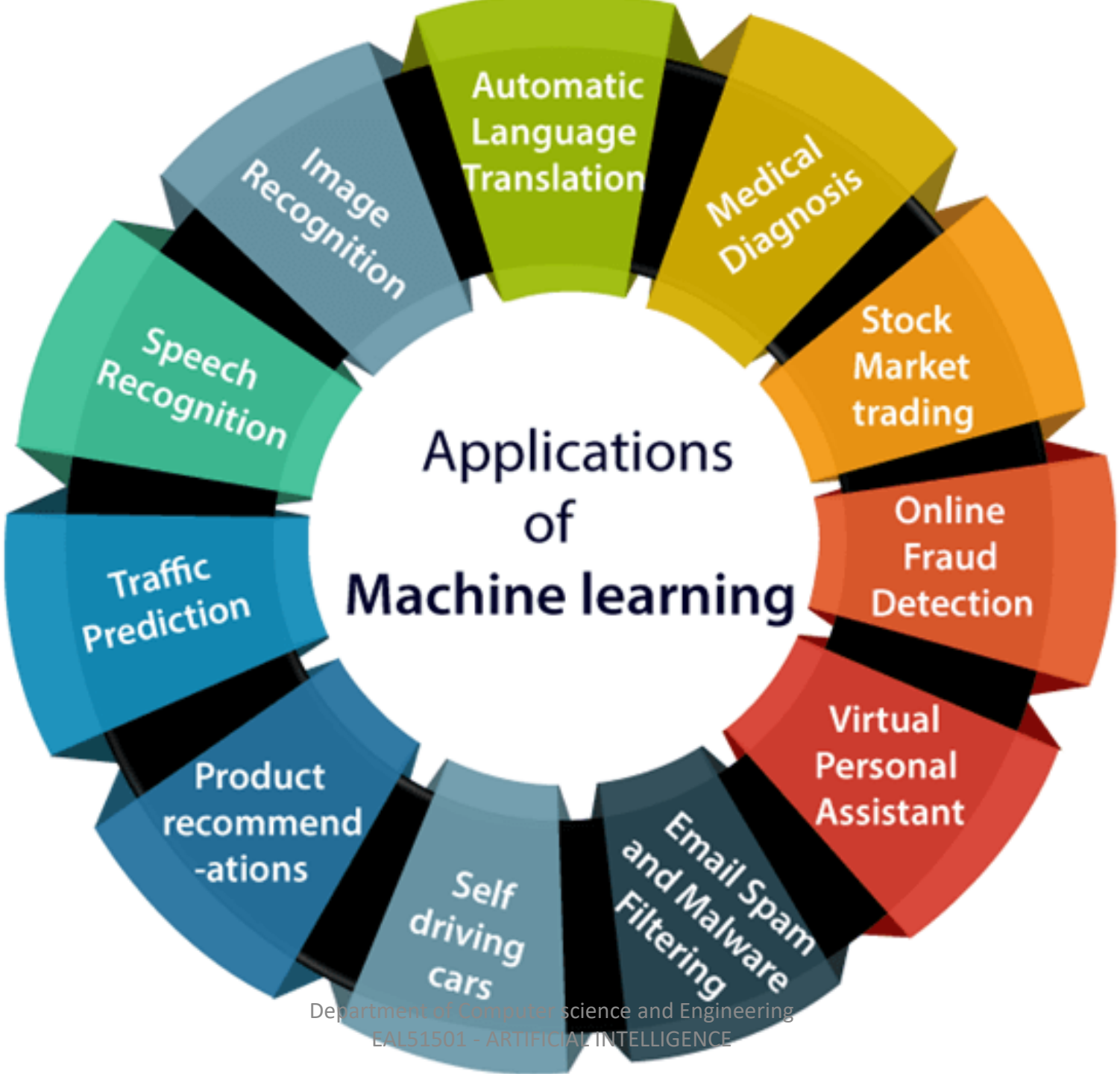


4. Agriculture

5. Cyber security


Deep Neural Network (Pretraining)
SVM
Multi-layered Perceptron (Backpropagation)
XOR Problem
ADALINE
Perceptron
Electronic Brain
Golden Age
Dark Age ("AI Winter")
1943
1957
1960
1969
1986
1995
2006
1940
1950
1960
1970
1980
1990
2000
2010
S. McCulloch - W. Pitts
F. Rosenblatt
B. Widrow - M. Hoff
M. Minsky - S. Papert
D. Rumelhart - G. Hinton - R. Wiliams
V. Vapnik - C. Cortes
G. Hinton - S. Ruslan
X AND Y
X OR Y
NOT X
Foward Activity
Backward Error
• Adjustable Weights
• Weights are not Learned
• Learnable Weights and Threshold
• XOR Problem
• Solution to nonlinearly separable problems
• Big computation, local optima and overfitting
• Limitations of learning prior knowledge
• Kernel function: Human Intervention
• Hierarchical feature Learning

- **Machine Learning at present:**
- Now machine learning has got a great advancement in its research, and it is present everywhere around us, such as **self-driving cars**, **Amazon Alexa**, **Chatbots**, **recommender system**, and many more.
- It includes **Supervised**, **unsupervised**, and **reinforcement learning with clustering**, **classification**, **decision tree**, **SVM algorithms**, etc.
- Modern machine learning models can be used for making various predictions, including **weather prediction**, **disease prediction**, **stock market analysis**, etc.


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend-ations
Traffic Prediction
Speech Recognition
Image Recognition

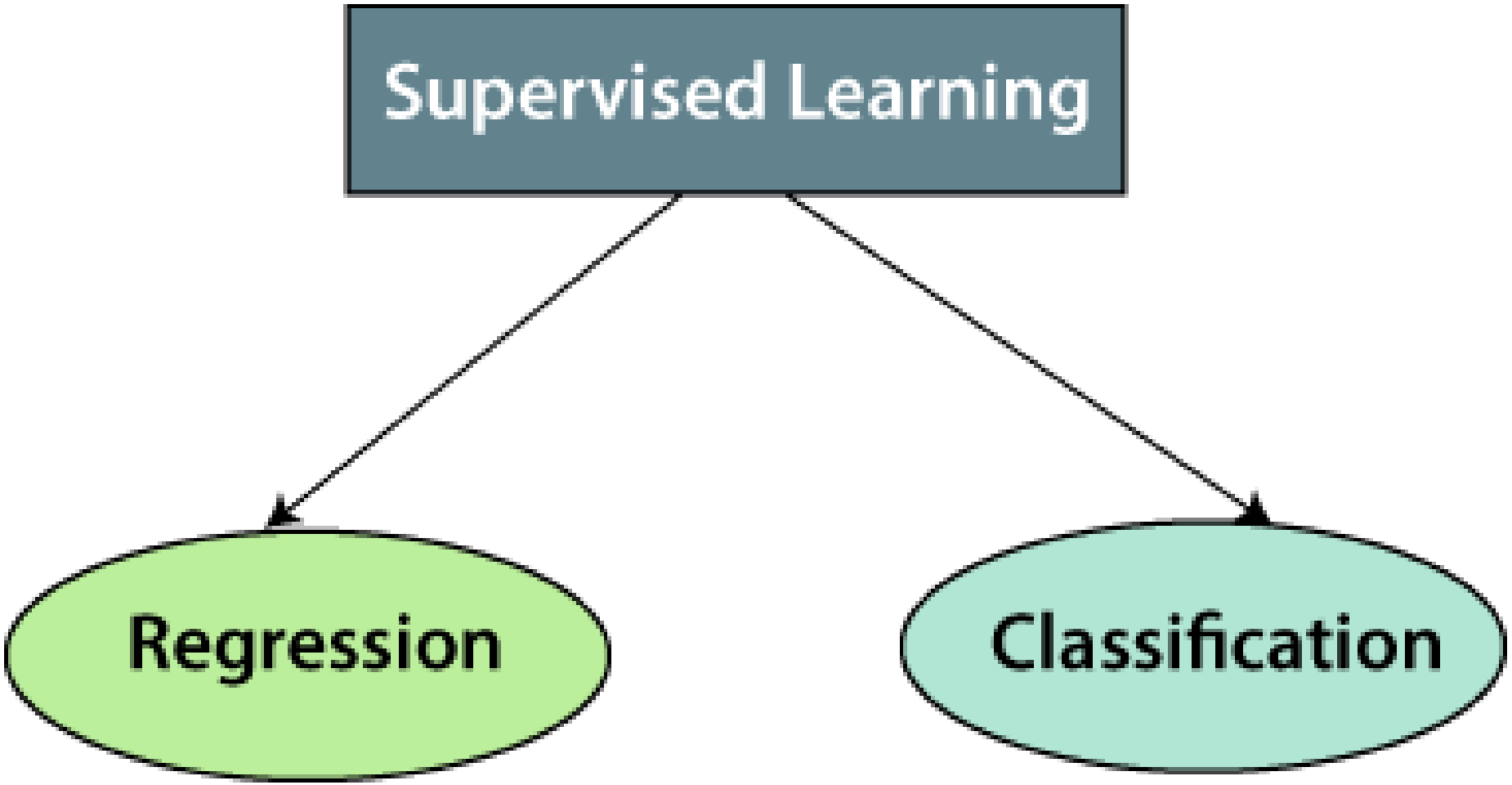
HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

EAL51501 – ARTIFICIAL INTELLIGENCE
B.Tech[AIML] – III Semester

K.Kowsalya
Assistant Professor (SS)
School of Computing Sciences,
Department of Computer Science and Engineering

MODULE-III

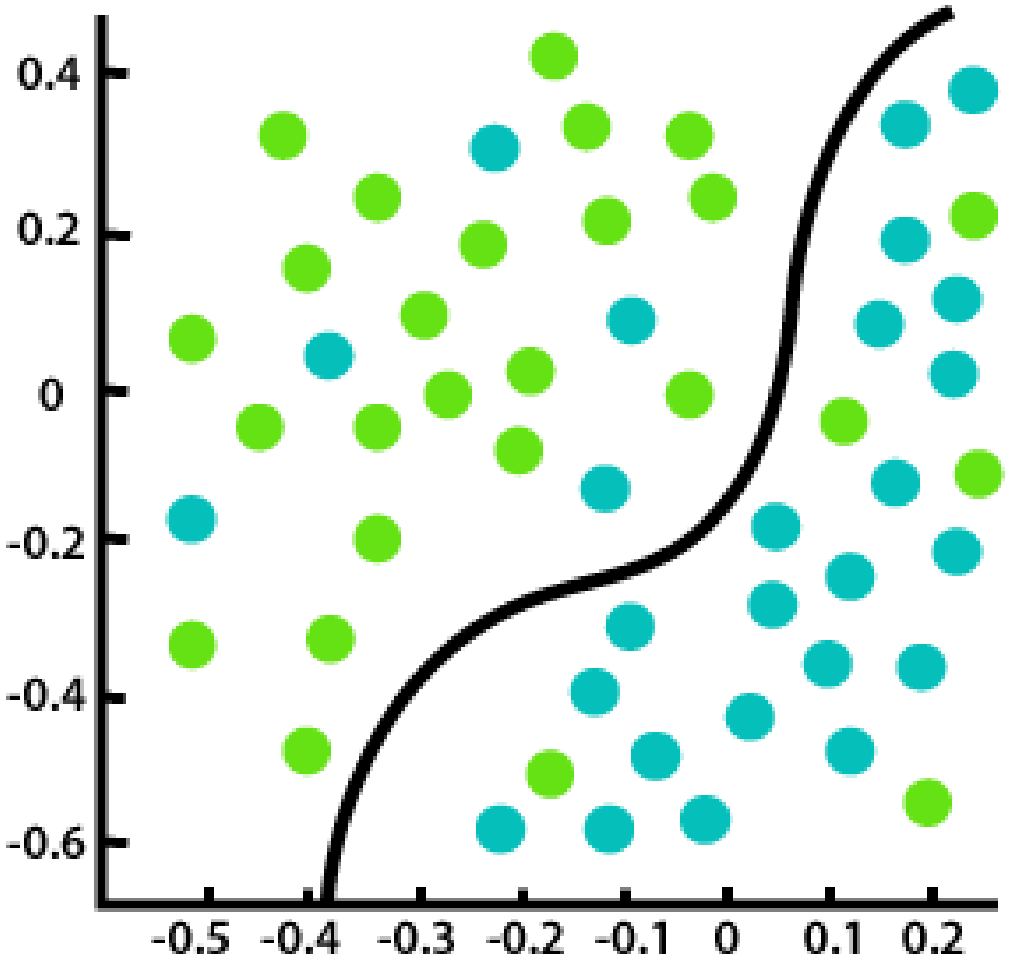
- Motivation for Machine Learning, Applications, Machine Learning, Learning associations, **Classification, Regression,** The Origin of machine learning, Uses and abuses of machine learning, Success cases, How do machines learn, Abstraction and knowledge representation, Generalization, Factors to be considered, Assessing the success of learning, Metrics for evaluation of classification method, Steps to apply machine learning to data, Machine learning process, Input data and ML algorithm, Classification of machine learning algorithms, General ML architecture, Group of algorithms, Reinforcement learning, Supervised learning, Unsupervised learning, Semi-Supervised learning, Algorithms, Ensemble learning, Matching data to an appropriate algorithm.

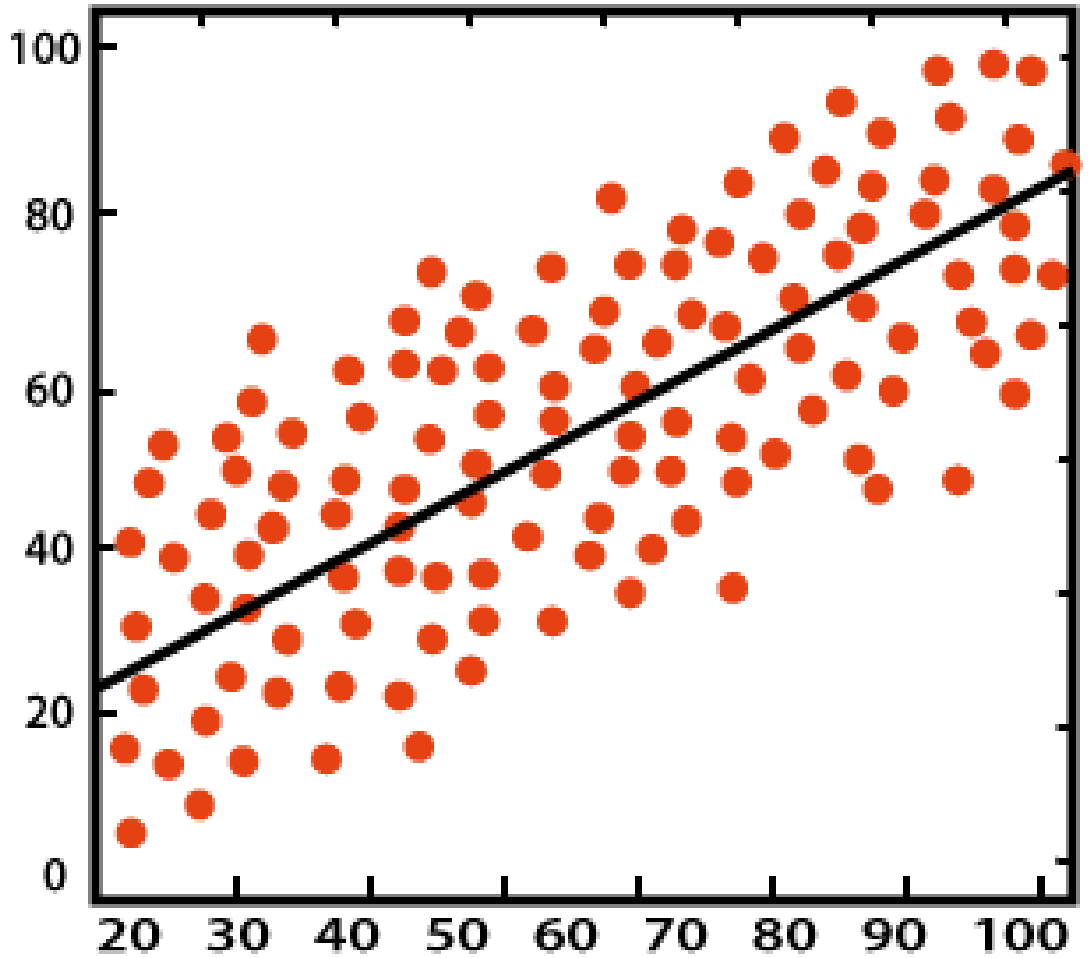

Supervised Learning
Regression
Classification

- **Regression and Classification algorithms**

- Regression algorithms are used to **predict the continuous** values such as price, salary, age, etc.

- Classification algorithms are used to **predict/Classify the discrete values** such as Male or Female, True or False, Spam or Not Spam, etc.



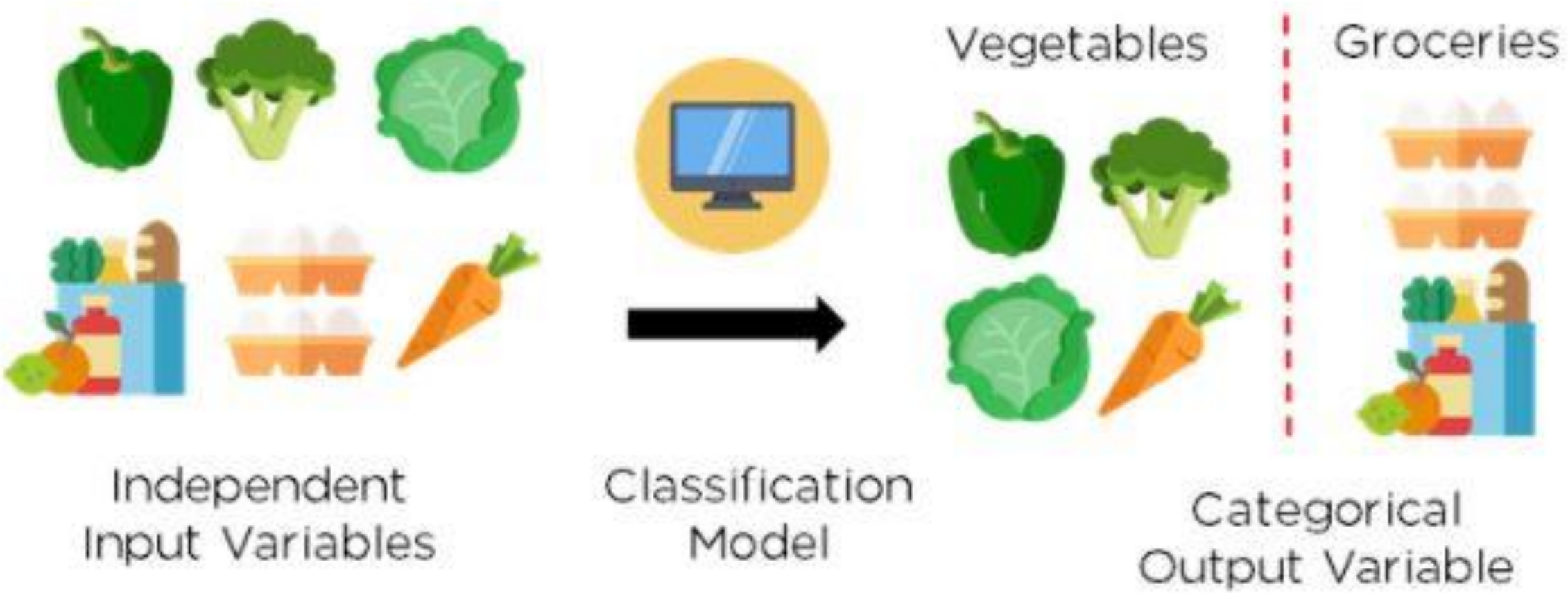


Classification

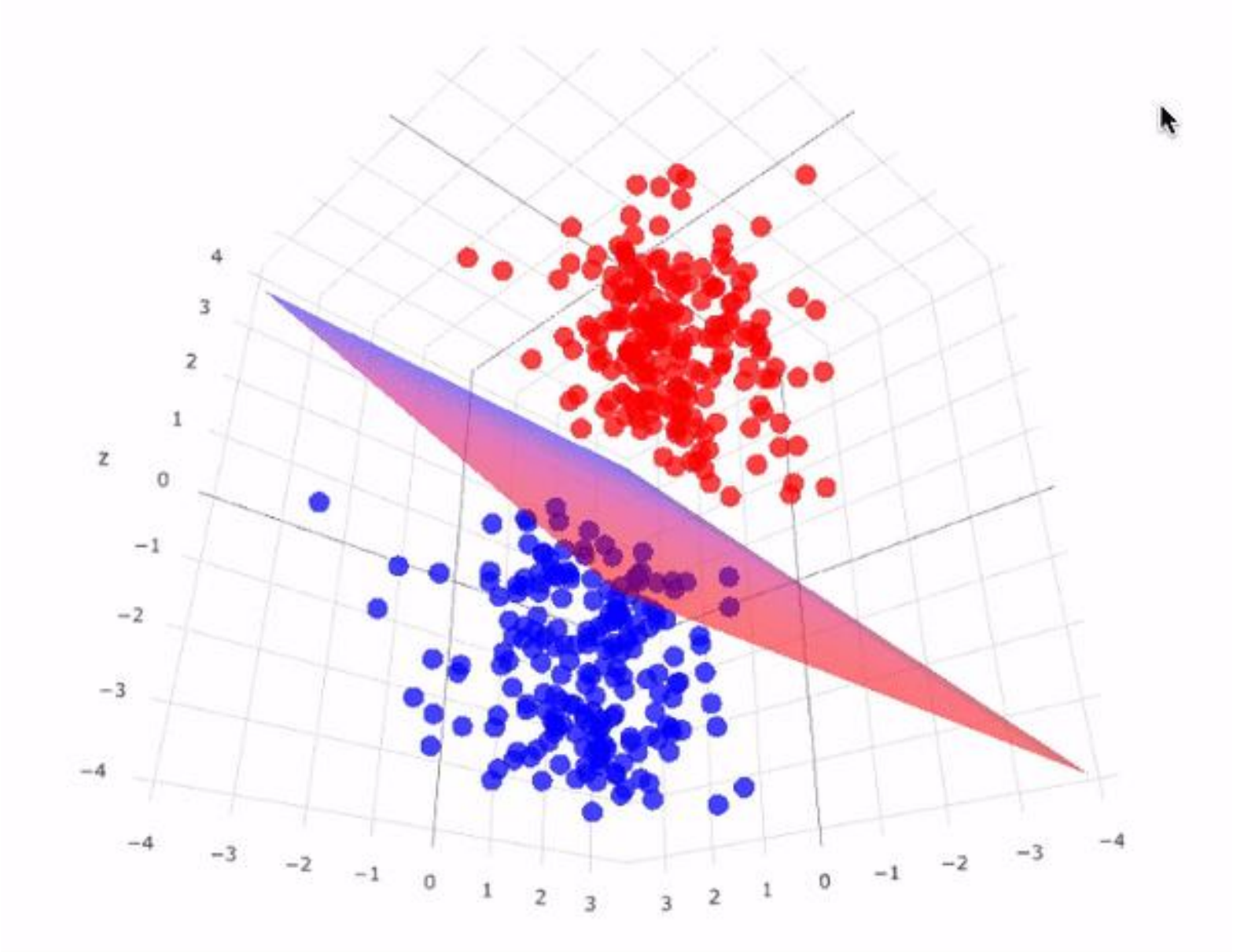
Regression

CLASSIFICATION

- The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
- In Classification, **a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups.**
- Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc.
- Classes can be called as targets/labels or categories.
- It takes labeled input data, which means it contains input with the corresponding output.
- **In classification algorithm, a discrete output function(y) is mapped to input variable(x).**



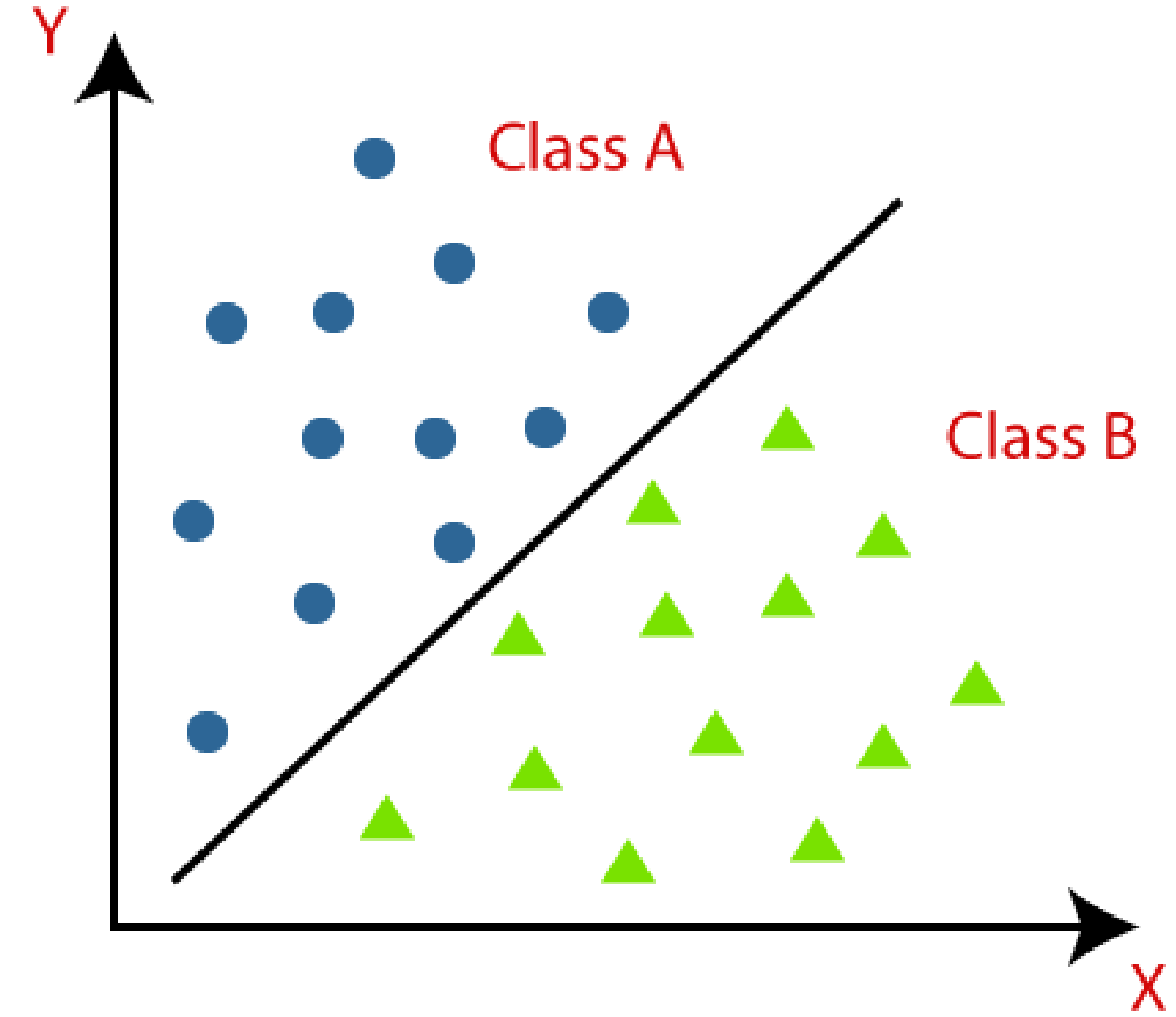
Classification for Groceries and Vegetables


z
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
3
2
1
0
−1
−2
−3
−4

y=f(x), where y = categorical output

- The best example of an ML classification algorithm is **Email Spam Detector**.
- The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to **predict the output for the categorical data.**
- In the below diagram, there are two classes, class A and Class B.
- These classes have features that are similar to each other and dissimilar to other classes.

Classification



Data Set Sample

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa

- The algorithm which implements the classification on a dataset is known as a classifier. There are two types of Classifications:

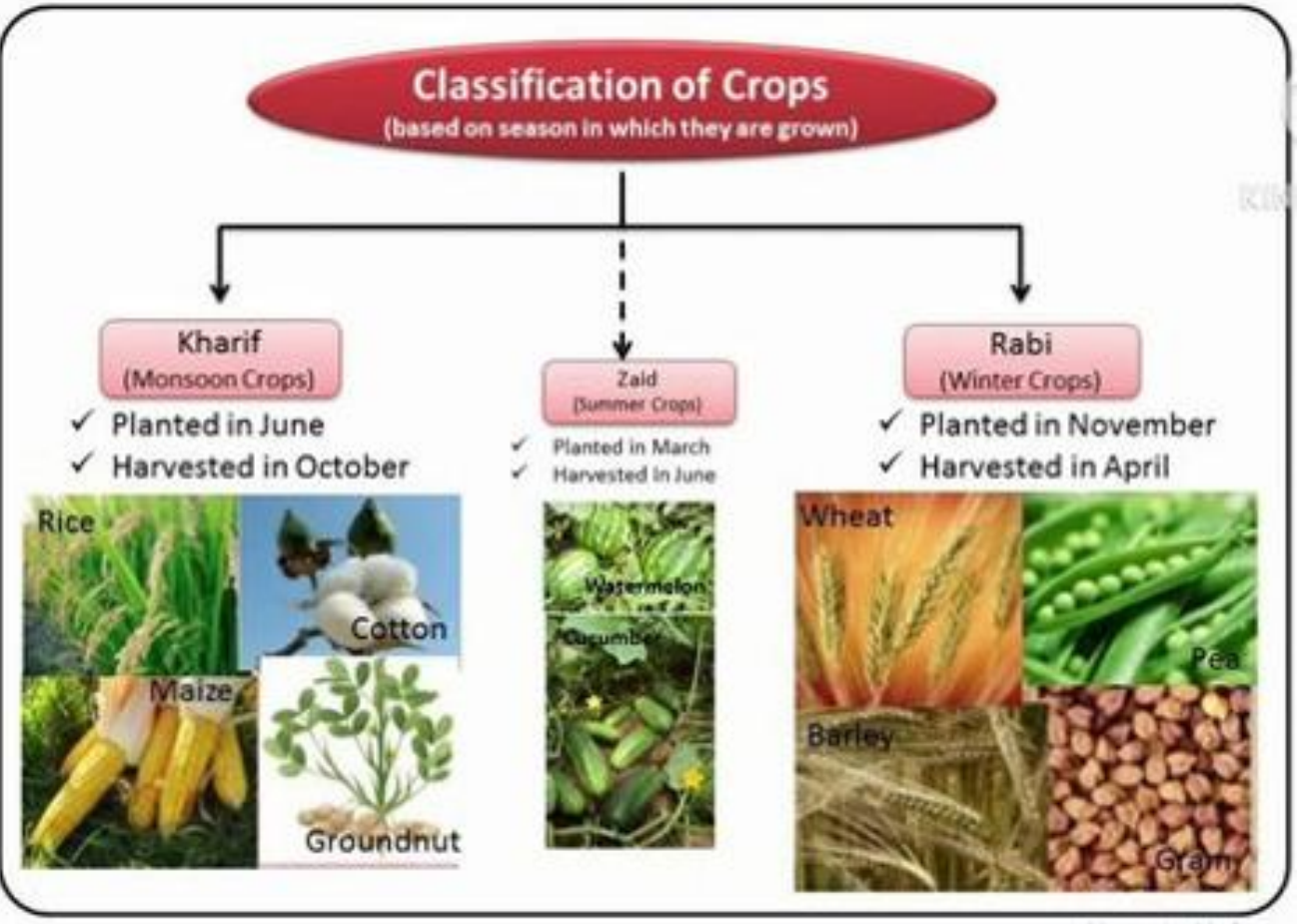
Binary Classifier:

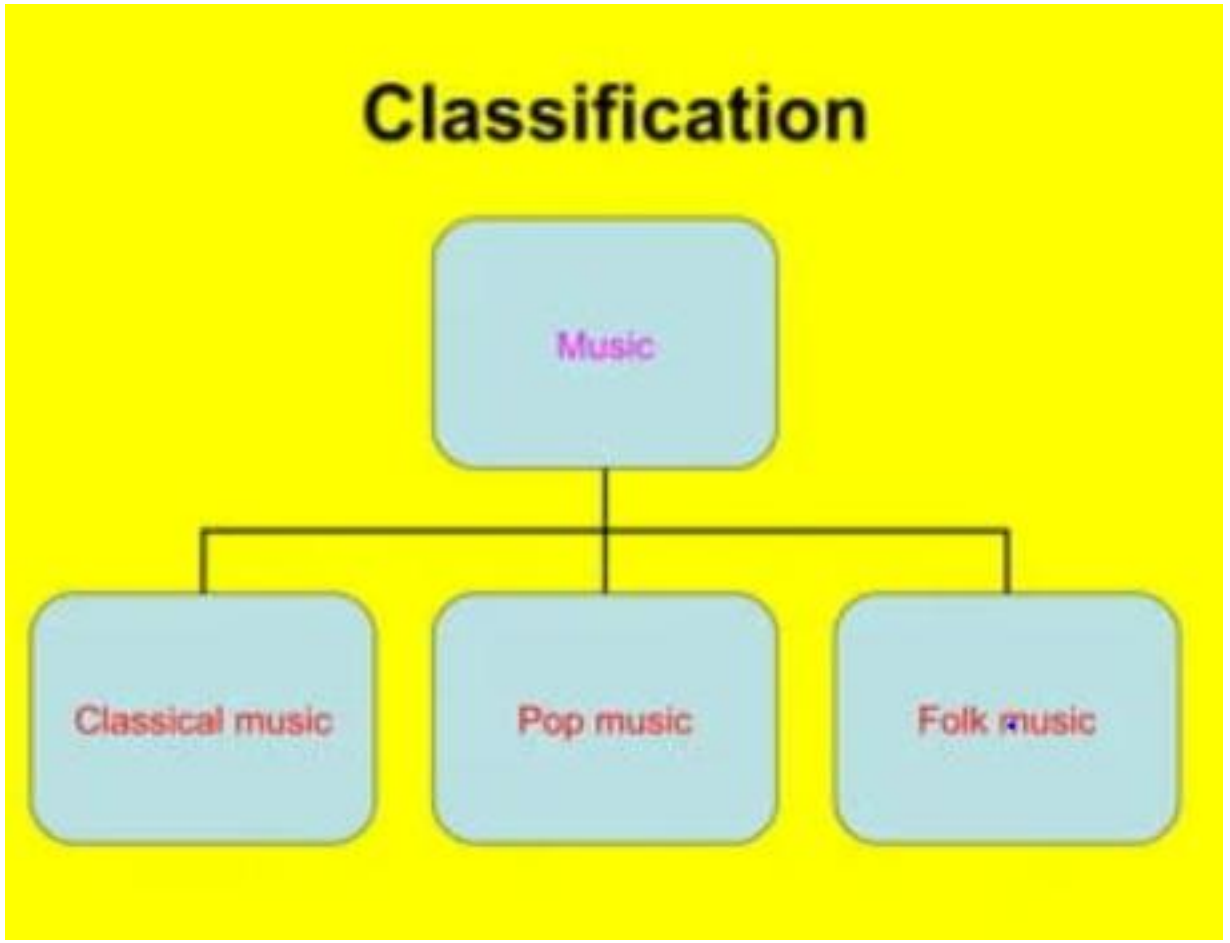
- If the classification problem has only two possible outcomes, then it is called as Binary Classifier.

  Eg: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.

- **Multi-class Classifier:** If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.

  Eg: Classifications of types of crops, Classification of types of music.

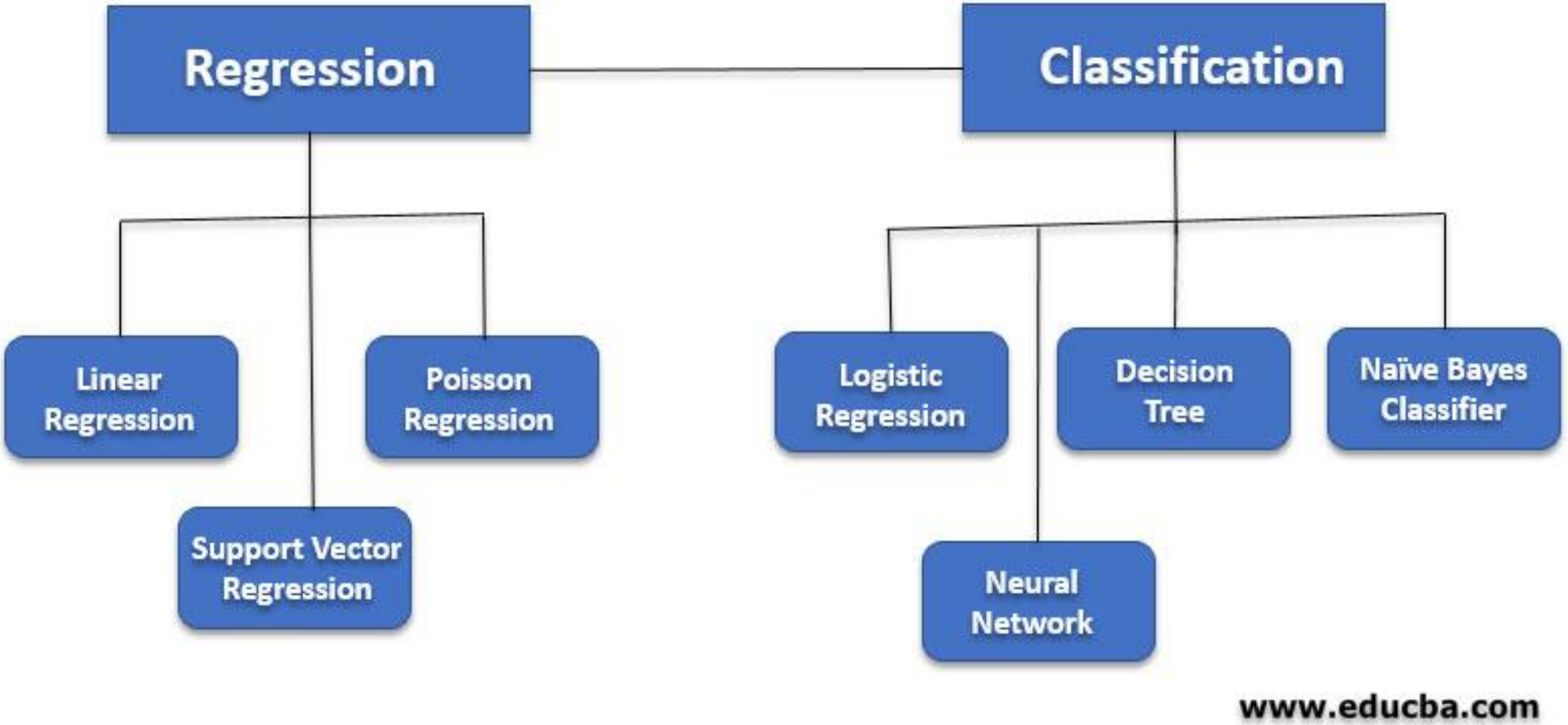

Classification of Crops
(based on season in which they are grown)
Kharif
(Monsoon Crops)
✓ Planted in June
✓ Harvested in October
Zaid
(Summer Crops)
✓ Planted in March
✓ Harvested in June
Rabi
(Winter Crops)
✓ Planted in November
✓ Harvested in April
Rice
Cotton
Maize
Groundnut
Watermelon
Cucumber
Wheat
Pea
Barley
Gram


Classification
Music
Classical music
Pop music
Folk music

Learners in Classification Problems:

- In the classification problems, there are two types of learners:
- **Lazy Learners:** Lazy Learner firstly stores the training dataset and wait until it receives the test dataset. In Lazy learner case, classification is done on the basis of the most related data stored in the training dataset. It takes less time in training but more time for predictions. **Example:** K-NN algorithm, Case-based reasoning

- **Eager Learners:**Eager Learners develop a classification model based on a training dataset before receiving a test dataset. Opposite to Lazy learners, Eager Learner takes more time in learning, and less time in prediction. **Example:** Decision Trees, Naïve Bayes, ANN.


Regression
Classification
Linear Regression
Poisson Regression
Support Vector Regression
Logistic Regression
Decision Tree
Naïve Bayes Classifier
Neural Network
www.educba.com

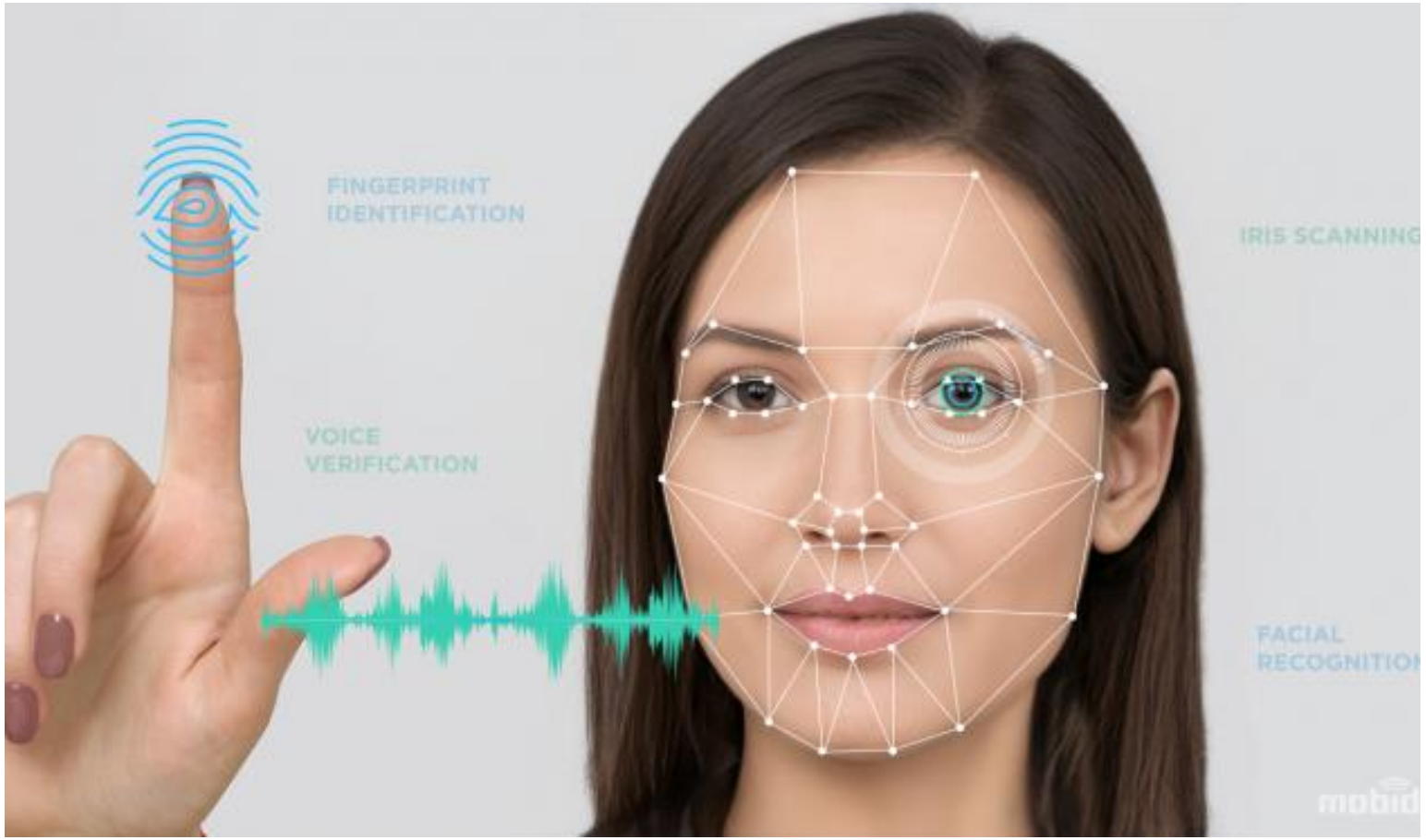
- **Types of ML Classification Algorithms:**
- Classification Algorithms can be further divided into the Mainly two category:
- **Linear Models**
  - Logistic Regression
  - Support Vector Machines
- **Non-linear Models**
  - K-Nearest Neighbours
  - Kernel SVM
  - Naïve Bayes
  - Decision Tree Classification
  - Random Forest Classification

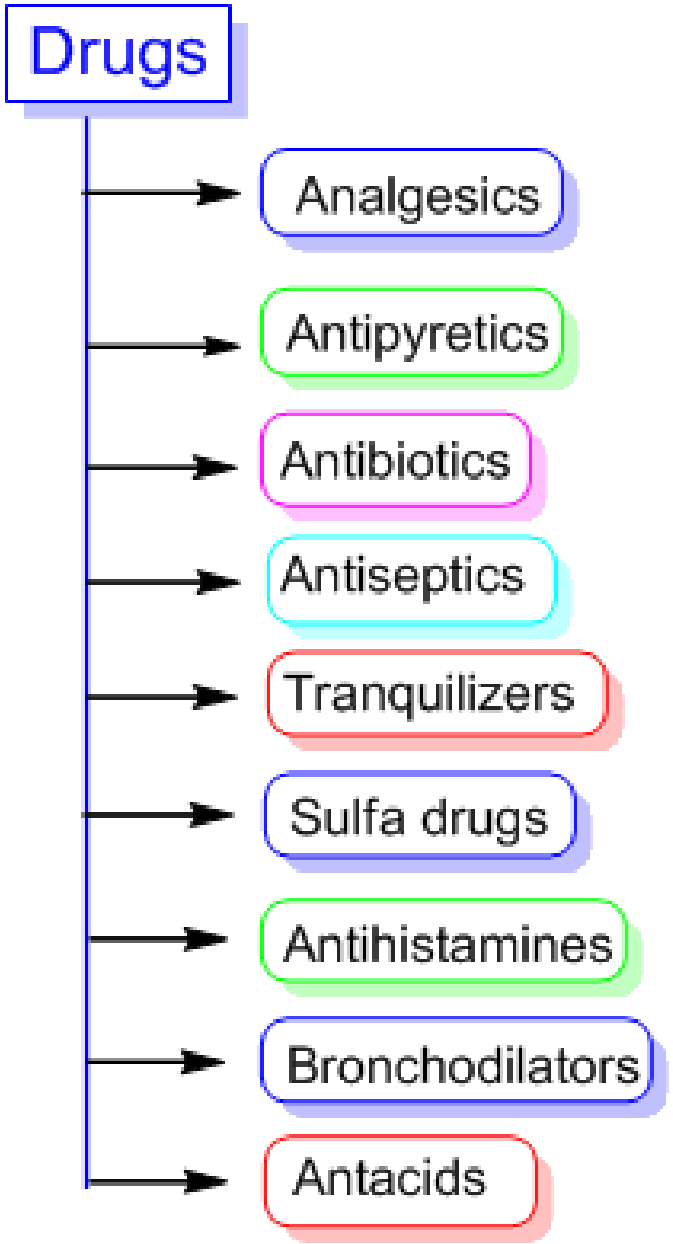
- Linearity refers to the property of a system or model where the output is directly proportional to the input, while nonlinearity implies that the relationship between input and output is more complex and cannot be expressed as a simple linear function

Use cases of Classification Algorithms

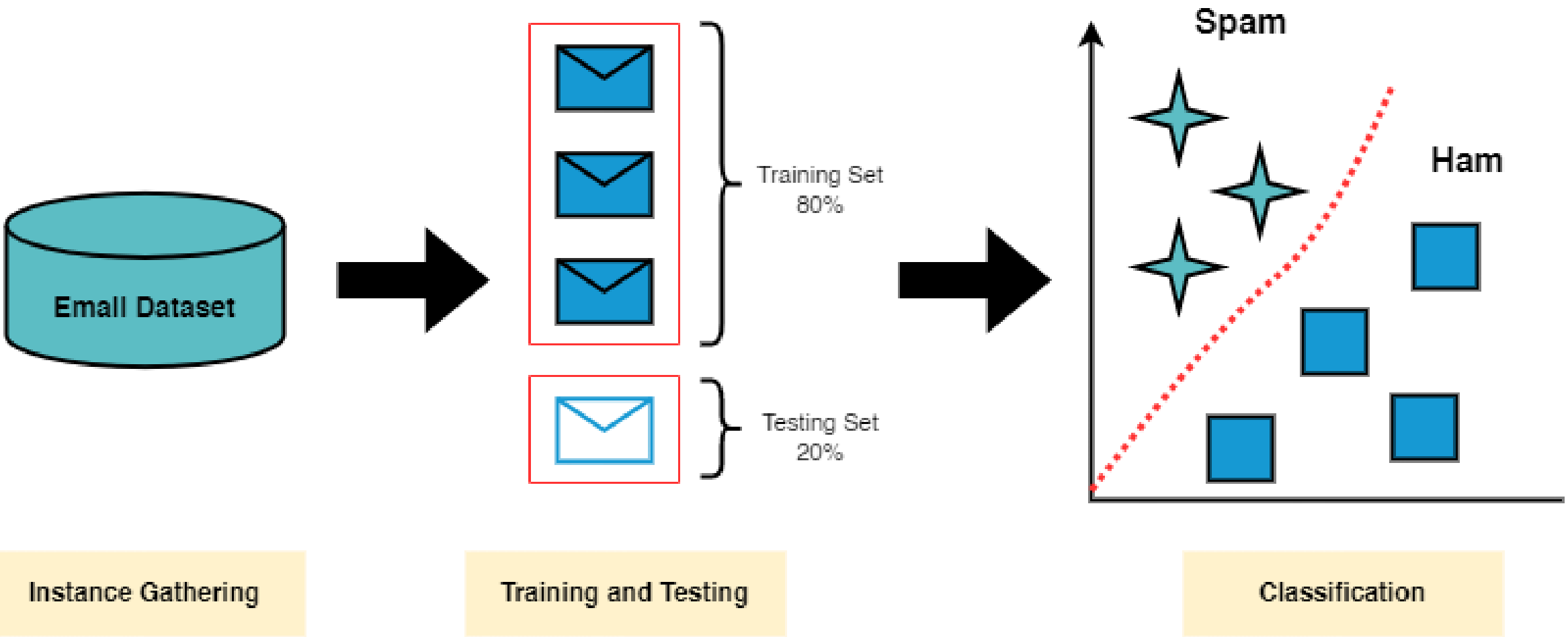
- Classification algorithms can be used in different places. Below are some popular use cases of Classification Algorithms:
- Email Spam Detection
- Speech Recognition
- Identifications of Cancer tumor cells.
- Drugs Classification
- Biometric Identification, etc.

Biometric Identification





Email Spam Detection

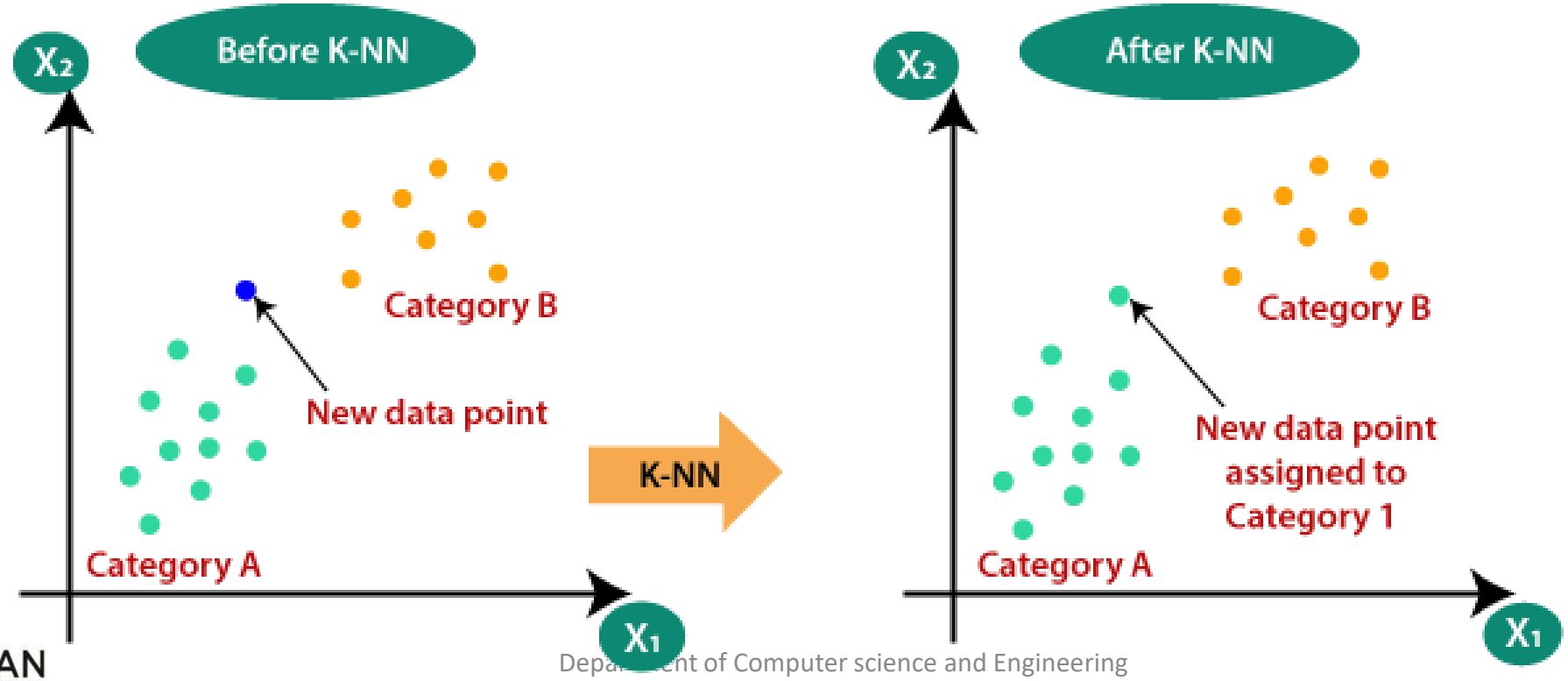


K-NN Algorithm

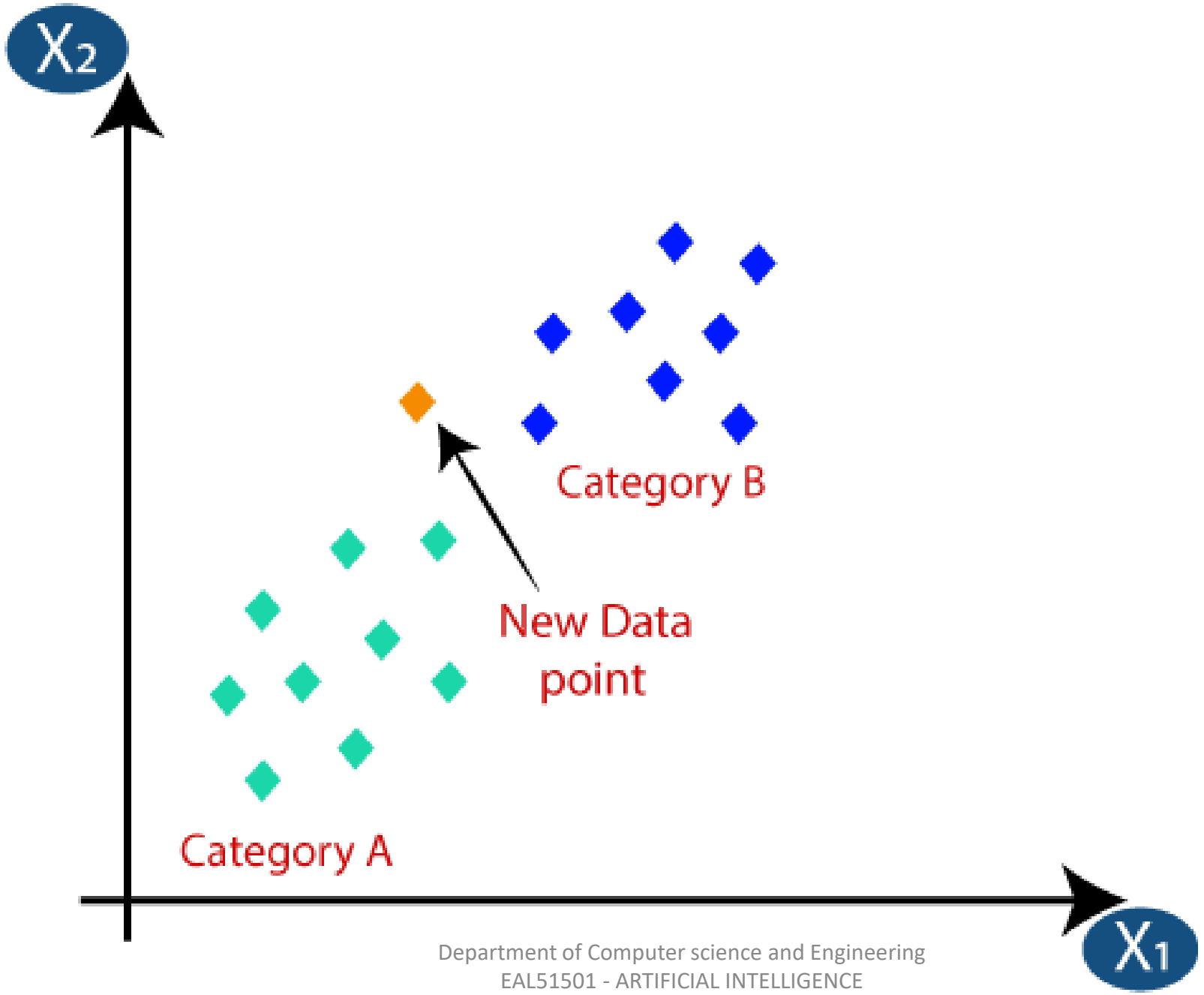
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

- **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

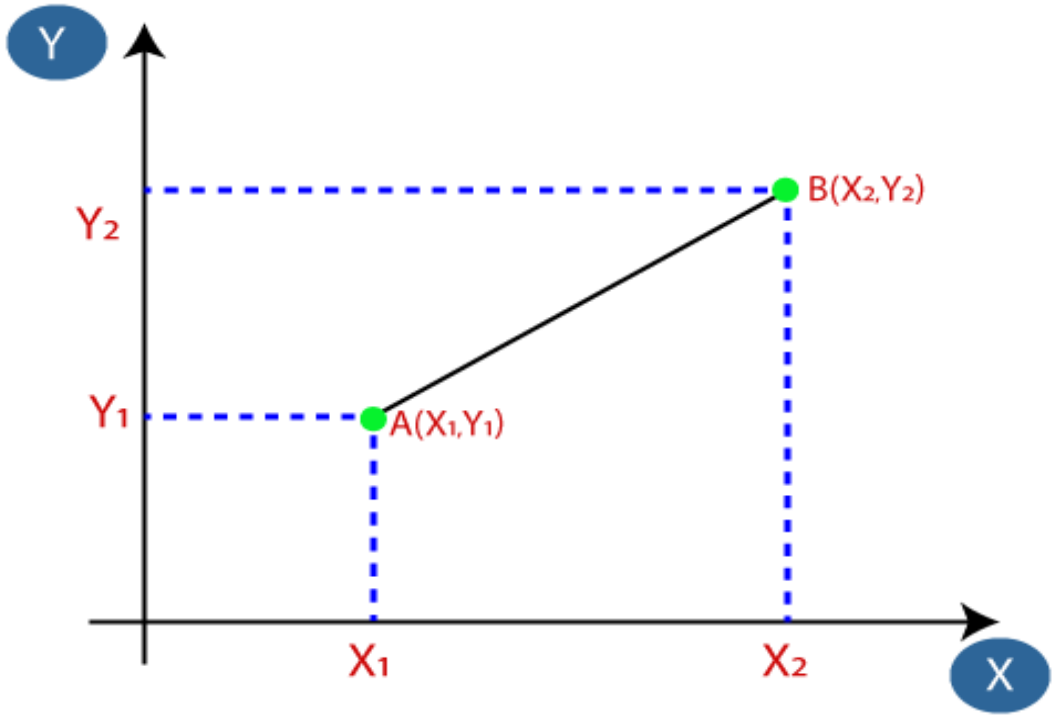
- **Why do we need a K-NN Algorithm?**
- Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



- **How does K-NN work?**
- The K-NN working can be explained on the basis of the below algorithm:
- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.
- Suppose we have a new data point and we need to put it in the required category. Consider the below image:

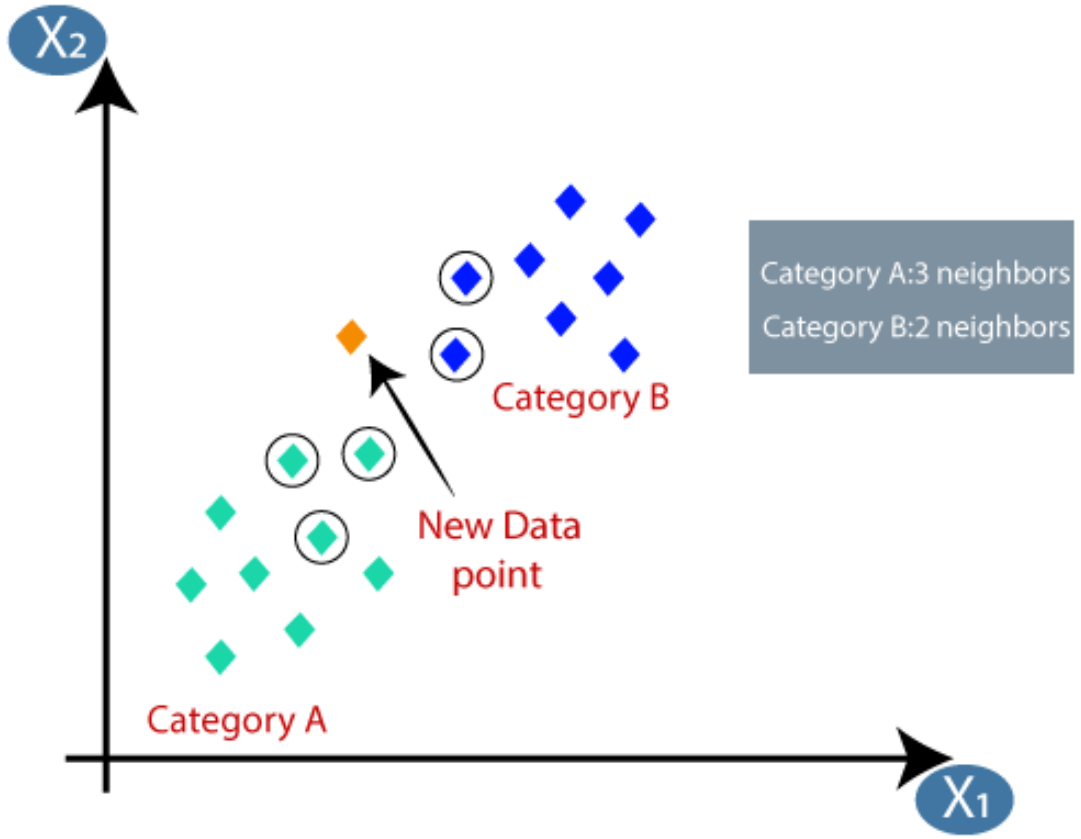

X2
Category B
New Data
point
Category A
X1

- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



Euclidean Distance between A_1 and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



- As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Regression

- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
- It predicts continuous/real values such as **temperature, age, salary, price,** etc.

- **Example:** Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

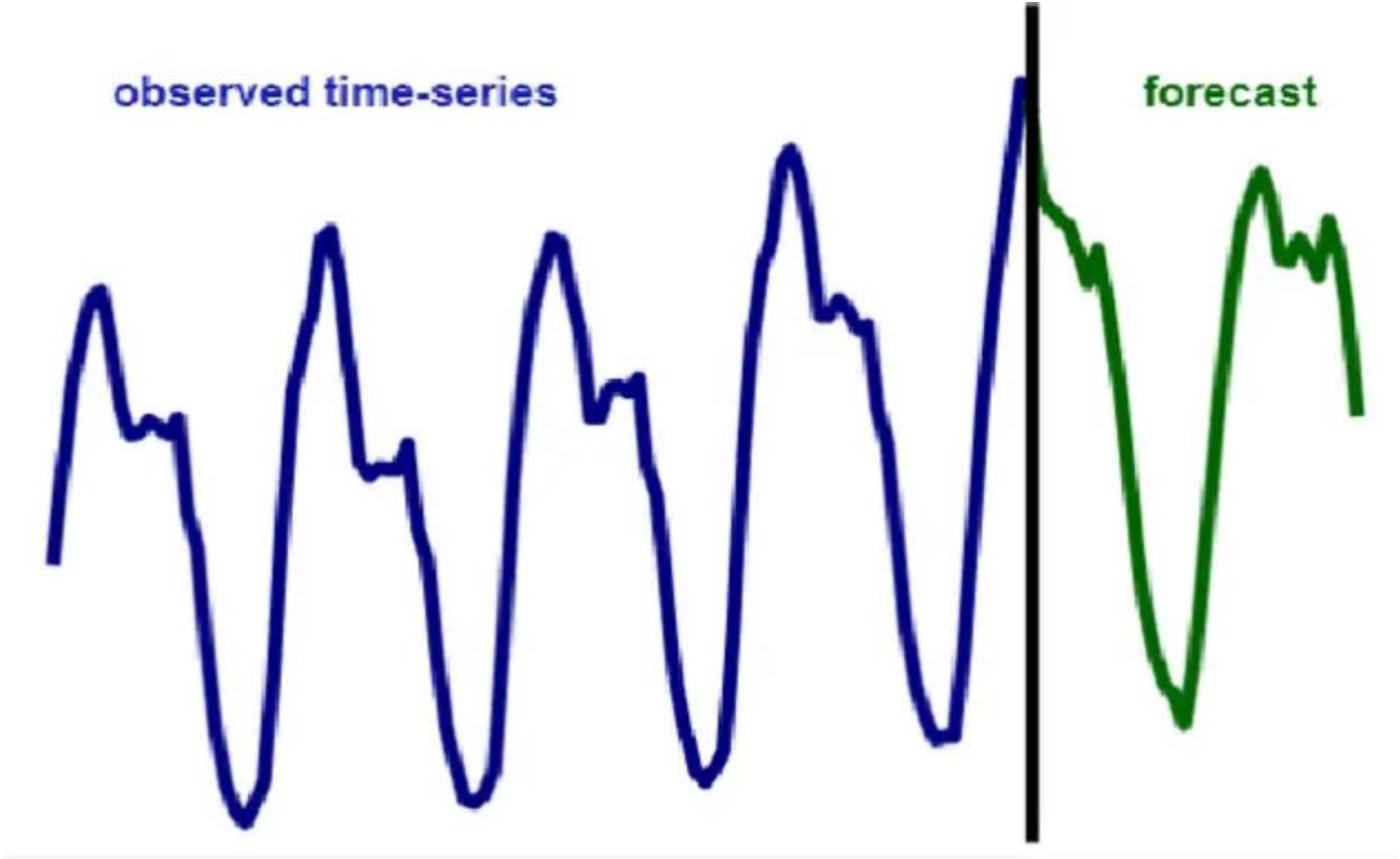
Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

- Now, the company wants to do the advertisement of $200 in the year 2019 **and wants to know the prediction about the sales for this year**. So to solve such type of prediction problems in machine learning, we need regression analysis.
- Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
- It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

- In simple words, ***"Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum."***
- The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

- **Prediction of rain using temperature and other factors**
- **Determining Market trends**
- **Prediction of road accidents due to rash driving.**


observed time-series
forecast

- **Terminologies Related to the Regression Analysis:**

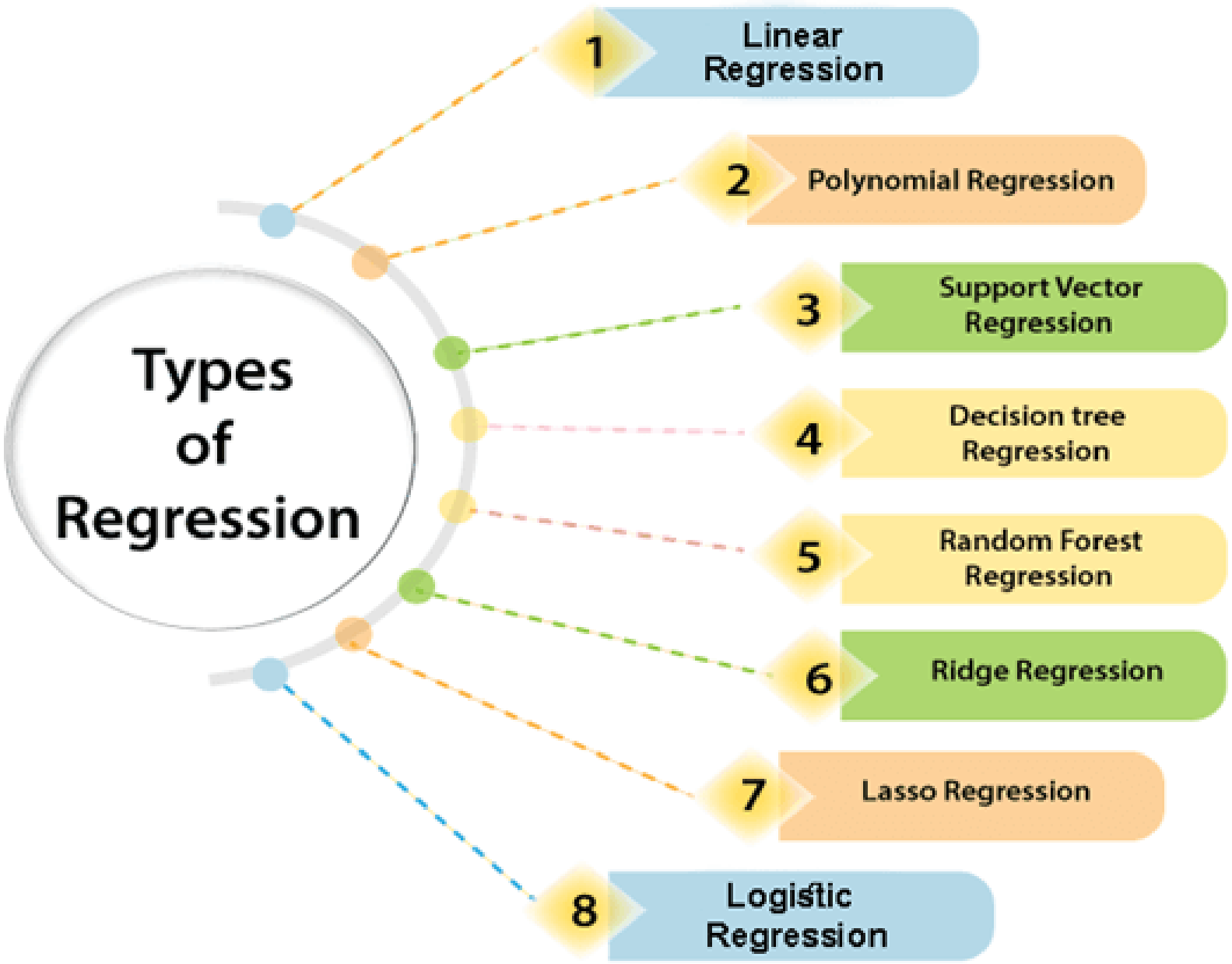
- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.

- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.

- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

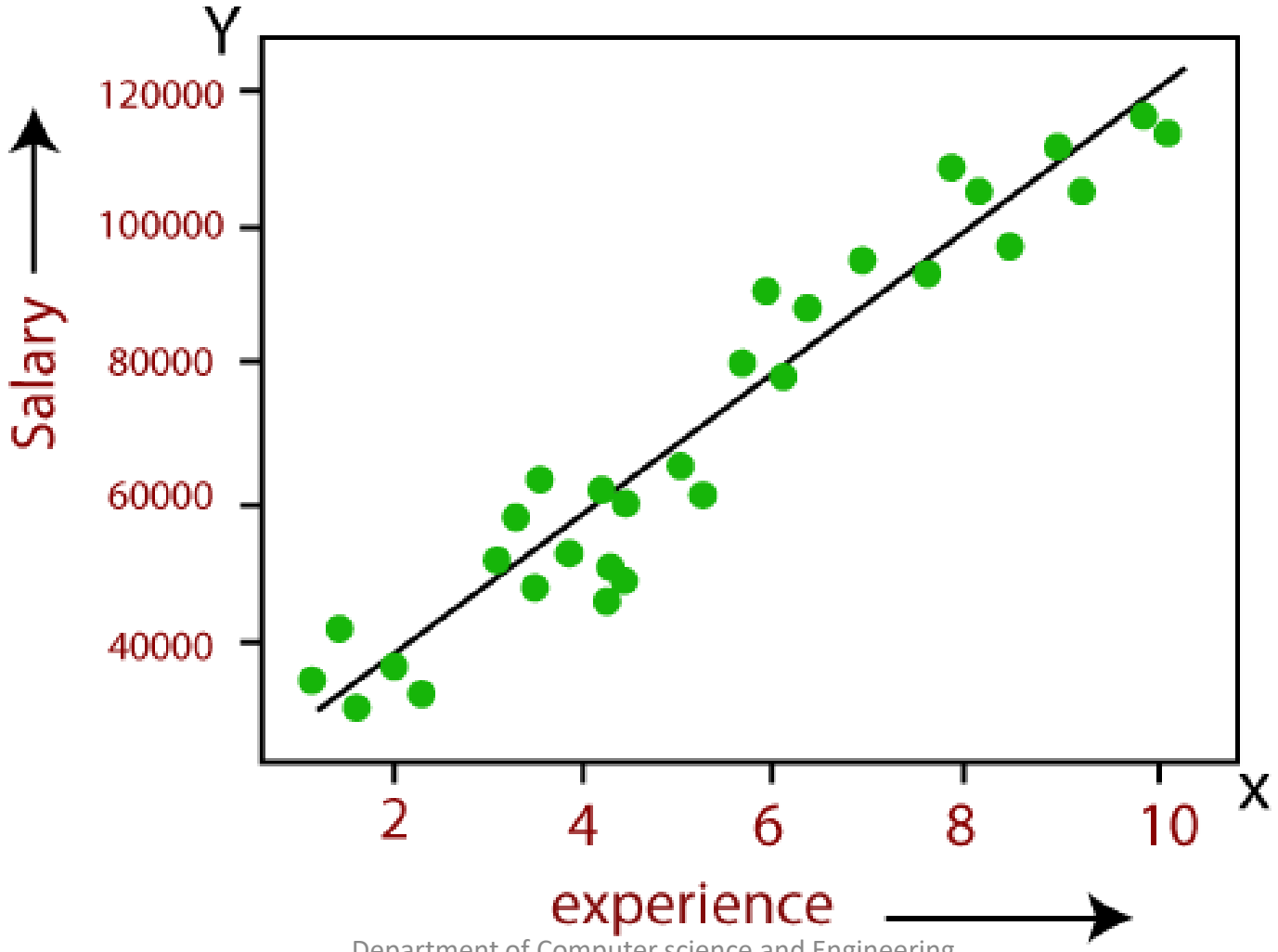
- **Why do we use Regression Analysis?**
- It helps in the prediction of a continuous variable.
- Various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately. So for such case we need Regression analysis which is a statistical method and used in machine learning and data science. Below are some other reasons for using Regression analysis:
- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

- **Types of Regression**
- Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:
- **Linear Regression**
- **Logistic Regression**
- **Polynomial Regression**
- **Support Vector Regression**
- **Decision Tree Regression**
- **Random Forest Regression**
- **Ridge Regression**
- **Lasso Regression**


Types of Regression
1 Linear Regression
2 Polynomial Regression
3 Support Vector Regression
4 Decision tree Regression
5 Random Forest Regression
6 Ridge Regression
7 Lasso Regression
8 Logistic Regression

- **Linear Regression:**
- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.

The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



- Below is the mathematical equation for Linear regression:

$$Y = aX + b$$

- **Here, Y = dependent variables (target variables),**
  X= Independent variables (predictor variables),
  a and b are the linear coefficients

Some popular applications of linear regression are:

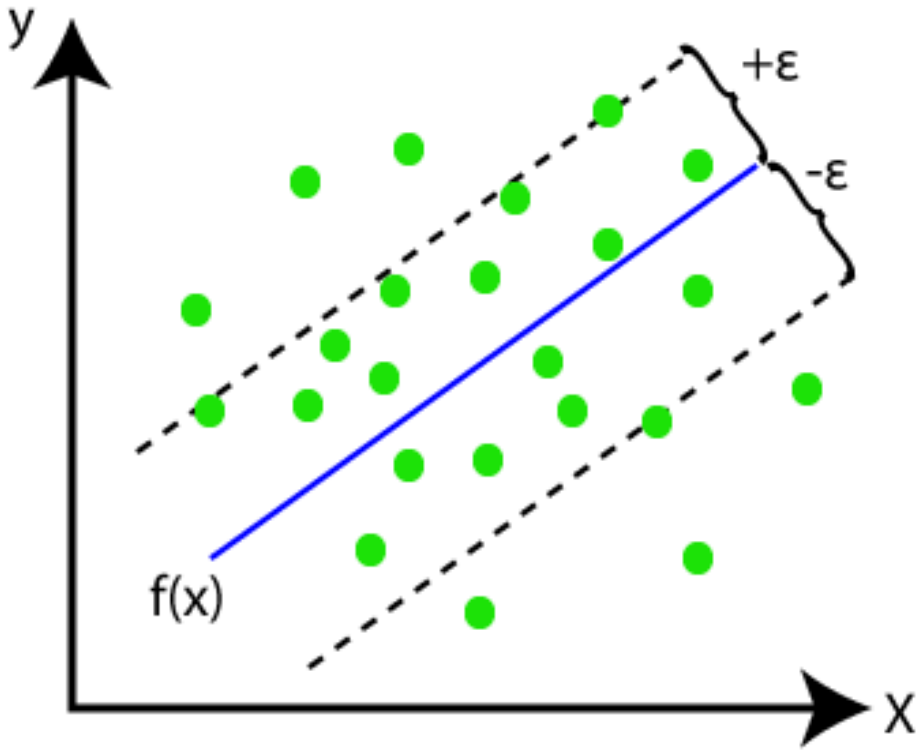
- **Analyzing trends and sales estimates**
- **Salary forecasting**
- **Real estate prediction**
- **Arriving at ETAs in traffic.**

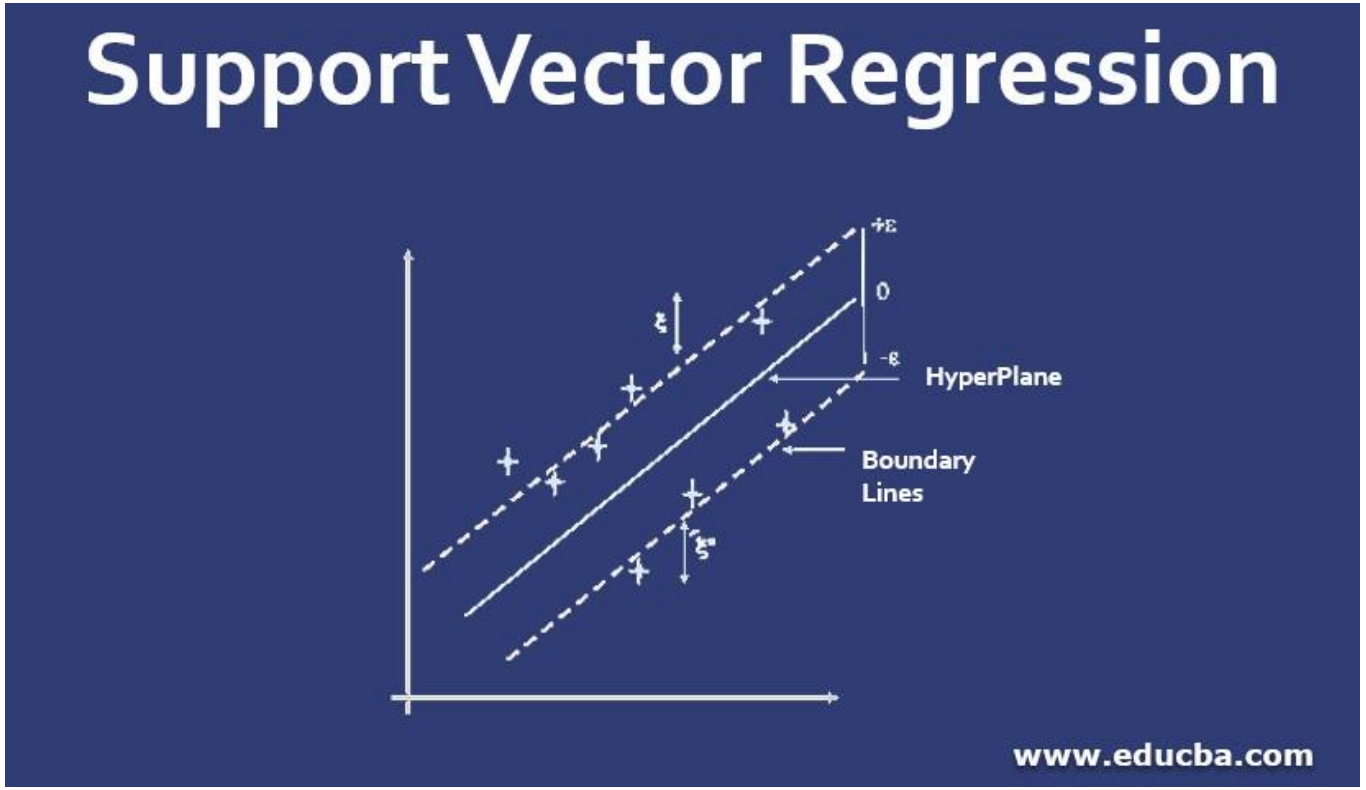
- **Support Vector Regression:**
- Support Vector Machine is a supervised learning algorithm which can be used for regression as well as classification problems. So if we use it for regression problems, then it is termed as Support Vector Regression.
- Support Vector Regression is a regression algorithm which works for continuous variables. Below are some keywords which are used in **Support Vector Regression**:
- **Kernel:** It is a function used to map a lower-dimensional data into higher dimensional data.
- **Hyperplane:** In general SVM, it is a separation line between two classes, but in SVR, it is a line which helps to predict the continuous variables and cover most of the datapoints.

- **Boundary line:** Boundary lines are the two lines apart from hyperplane, which creates a margin for datapoints.
- **Support vectors:** Support vectors are the datapoints which are nearest to the hyperplane and opposite class.

- In SVR, we always try to determine a hyperplane with a maximum margin, so that maximum number of datapoints are covered in that margin.

- ***The main goal of SVR is to consider the maximum datapoints within the boundary lines and the hyperplane (best-fit line) must contain a maximum number of datapoints***.





Here, the blue line is called hyperplane, and the other two lines are known as boundary lines.

- **Decision Tree Regression:**
- Decision Tree is a supervised learning algorithm which can be used for solving both classification and regression problems.
- It can solve problems for both categorical and numerical data
- Decision Tree regression builds a tree-like structure in which each internal node represents the "test" for an attribute, each branch represent the result of the test, and each leaf node represents the final decision or result.
- A decision tree is constructed starting from the root node/parent node (dataset), which splits into left and right child nodes (subsets of dataset). These child nodes are further divided into their children node, and themselves become the parent node of those nodes.

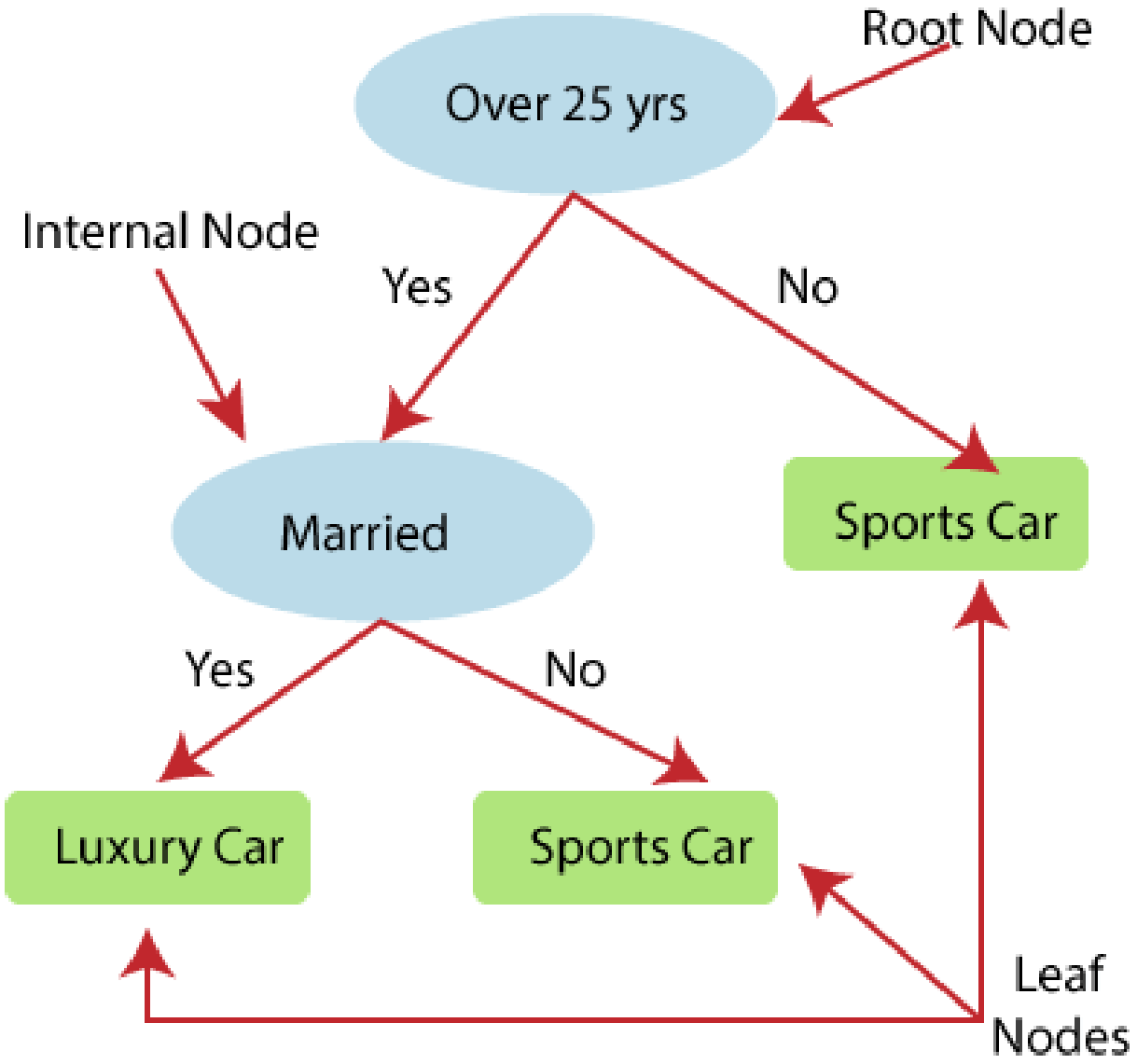
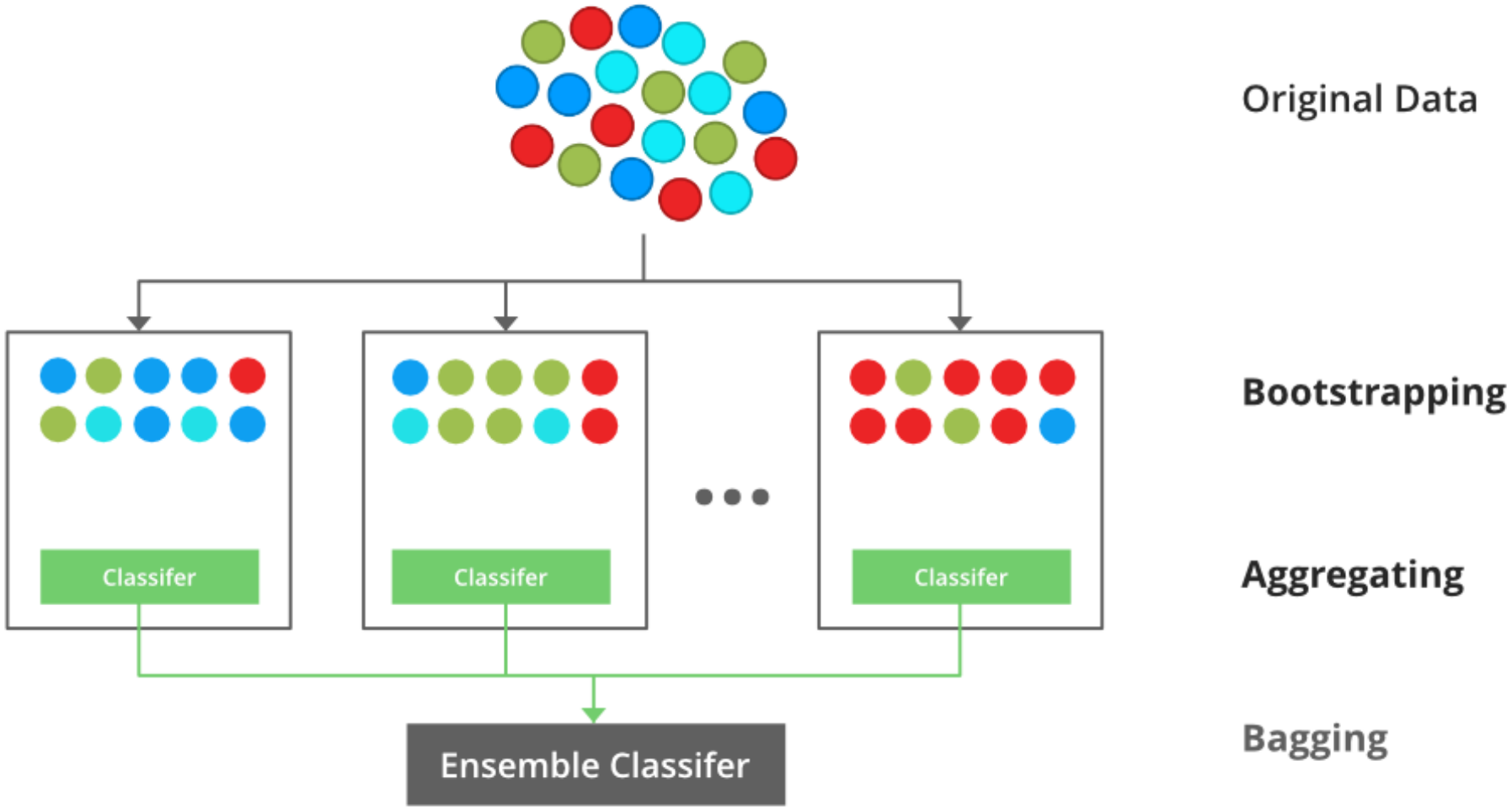


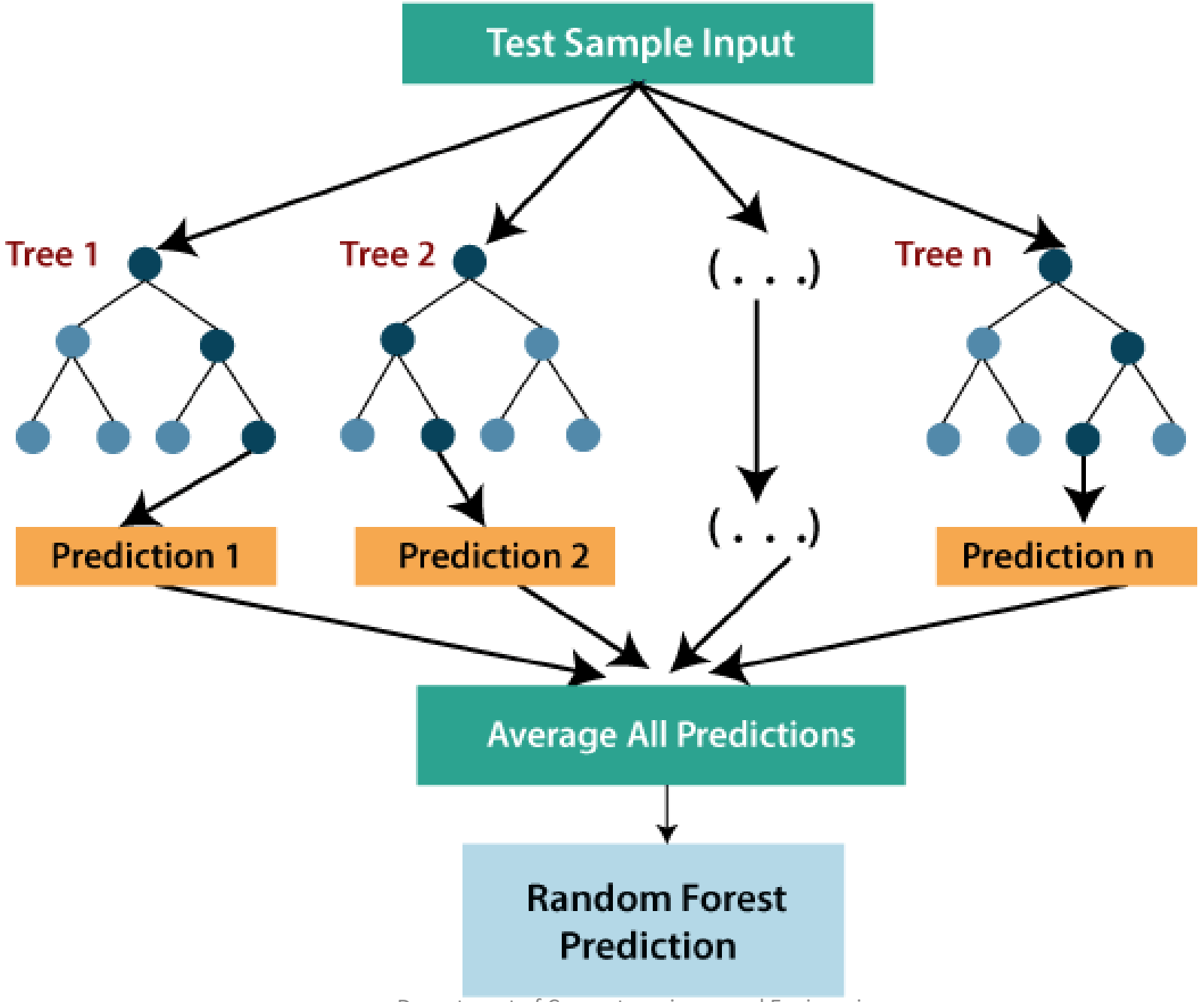
Image showing the example of Decision Tee regression,
here, the model is trying to predict the choice of a person between Sports cars or Luxury car.

- **Random Forest**
- Random forest is one of the most powerful supervised learning algorithms which is capable of performing regression as well as classification tasks.
- The Random Forest regression is an ensemble learning method which combines multiple decision trees and predicts the final output based on the average of each tree output. The combined decision trees are called as base models, and it can be represented more formally as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots$$

- Random forest uses **Bagging or Bootstrap Aggregation** technique of ensemble learning in which aggregated decision tree runs in parallel and do not interact with each other.[To resolve overfitting]
- With the help of Random Forest regression, we can prevent Overfitting in the model by creating random subsets of the dataset.


Original Data
Bootstrapping
Classifer
Classifer
Classifer
Aggregating
Ensemble Classifer
Bagging


Test Sample Input
Tree 1
Tree 2
(. . .)
Tree n
Prediction 1
Prediction 2
(. . .)
Prediction n
Average All Predictions
Random Forest Prediction

Regression Algorithm	Classification Algorithm
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
The regression Algorithm can be further divided into Linear and Non-linear Regression.	The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.

HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

EAL51501 – ARTIFICIAL INTELLIGENCE
B.Tech[AIML] – III Semester

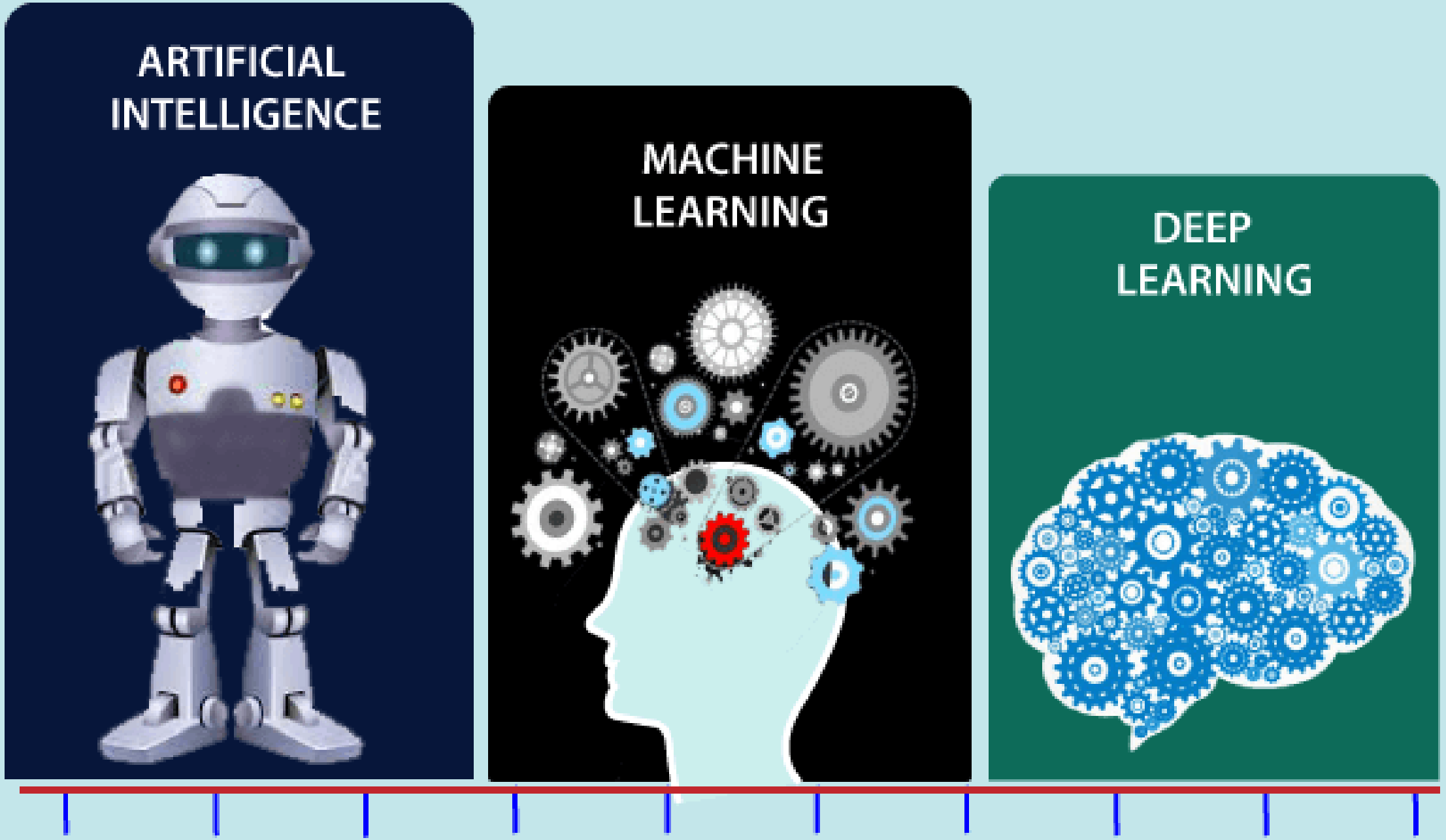
K.Kowsalya
Assistant Professor (SS)
School of Computing Sciences,
Department of Computer Science and Engineering

UNIT-III

- Motivation for Machine Learning, Applications, Machine Learning, Learning associations, Classification, Regression, **The Origin of machine learning, Uses and abuses of machine learning, Success cases, How do machines learn[INTRO], Abstraction and knowledge representation, Generalization, Factors to be considered,** Assessing the success of learning, Metrics for evaluation of classification method, Steps to apply machine learning to data, Machine learning process, Input data and ML algorithm, Classification of machine learning algorithms, General ML architecture, Group of algorithms, Reinforcement learning, Supervised learning, Unsupervised learning, Semi-Supervised learning, Algorithms, Ensemble learning, Matching data to an appropriate algorithm.

- **History of Machine Learning**

- Before some years (about 40-50 years), machine learning was science fiction, but today it is the part of our daily life.
- Machine learning is making our day to day life easy from **self-driving cars** to **Amazon virtual assistant "Alexa"**.
- However, the idea behind machine learning is so old and has a long history.
- Below some milestones are given which have occurred in the history of machine learning:


ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
DEEP LEARNING
1940's
1950's
1960's
1970's
1980's
1990's
2000's
2010's

- **The early history of Machine Learning (Pre-1940):**
- **1834:** In 1834, Charles Babbage, the father of the computer, conceived a device that could be programmed with punch cards. However, the machine was never built, but all modern computers rely on its logical structure.
- **1936:** In 1936, Alan Turing gave a theory that how a machine can determine and execute a set of instructions.
- **The era of stored program computers:**
- **1940:** In 1940, the first manually operated computer, "ENIAC" was invented, which was the first electronic general-purpose computer. After that stored program computer such as EDSAC in 1949 and EDVAC in 1951 were invented.
- **1943:** In 1943, a human neural network was modeled with an electrical circuit. In 1950, the scientists started applying their idea to work and analyzed how human neurons might work.

- **Computer machinery and intelligence:**
- **1950:** In 1950, Alan Turing published a seminal paper, "**Computer Machinery and Intelligence**," on the topic of artificial intelligence. **In his paper, he asked, "Can machines think?"**
- **Machine intelligence in Games:**
- **1952:** Arthur Samuel, who was the pioneer of machine learning, created a program that helped an IBM computer to play a checkers game. It performed better more it played.
- **1959:** In 1959, the term "Machine Learning" was first coined by **Arthur Samuel**.

- **The first "AI" winter:**
- The duration of 1974 to 1980 was the tough time for AI and ML researchers, and this duration was called as **AI winter**.
- In this duration, failure of machine translation occurred, and people had reduced their interest from AI, which led to reduced funding by the government to the researches.
- **Machine Learning from theory to reality**
- **1959:** In 1959, the first neural network was applied to a real-world problem to remove echoes over phone lines using an adaptive filter.

- **1985:** In 1985, Terry Sejnowski and Charles Rosenberg invented a neural network **NETtalk**, which was able to teach itself how to correctly pronounce 20,000 words in one week.
- **1997:** The IBM's **Deep blue** intelligent computer won the chess game against the chess expert Garry Kasparov, and it became the first computer which had beaten a human chess expert.
- **Machine Learning at 21st century**
- **2006:** In the year 2006, computer scientist Geoffrey Hinton has given a new name to neural net research as "**deep learning**," and nowadays, it has become one of the most trending technologies.

- **2012:** In 2012, Google created a deep neural network which learned to recognize the image of humans and cats in YouTube videos
- **2014:** In 2014, the Chabot "**Eugen Goostman**" cleared the Turing Test. It was the first Chabot who convinced the 33% of human judges that it was not a machine.
- **2016: AlphaGo** beat the world's number second player **Lee sedol** at **Go game**. In 2017 it beat the number one player of this game **Ke Jie**.
- **2017:** In 2017, the Alphabet's Jigsaw team built an intelligent system that was able to learn the **online trolling**. It used to read millions of comments of different websites to learn to stop online trolling

- **Machine Learning at present:**
- Now machine learning has got a great advancement in its research, and it is present everywhere around us, such as **self-driving cars**, **Amazon Alexa**, **Chatbots**, **recommender system**, and many more. It includes **Supervised**, **unsupervised**, and **reinforcement learning with clustering**, **classification**, **decision tree**, **SVM algorithms**, etc.
- Modern machine learning models can be used for making various predictions, including **weather prediction**, **disease prediction**, **stock market analysis**, etc.

- **Uses and abuses of machine learning**

- Predict the outcomes of elections
- Identify and filter spam messages from e-mail
- Foresee criminal activity
- Automate traffic signals according to road conditions
- Produce financial estimates of storms and natural disasters
- Create auto-piloting planes and auto-driving cars
- Identify individuals with the capacity to donate
- Target advertising to specific types of consumers

- Identification of unwanted spam messages in email
- Segmentation of customer behavior for targeted advertising
- Forecasts of weather behavior and long-term climate changes
- Reduction of fraudulent credit card transactions
- Actuarial estimates of financial damage of storms and natural disasters
- Prediction of popular election outcomes
- Development of algorithms for auto-piloting drones and self-driving cars
- Optimization of energy use in homes and office buildings
- Projection of areas where criminal activity is most likely
- Discovery of genetic sequences linked to diseases

Success cases

Making Your Advertising Dollars Go (Much) Further: RedBalloon

Machine learning can be used in many ways in the world of advertising. From Lexus' first ad written by IBM's Watson to Red Balloon's discovery of a whole new audience, AI can touch nearly every process of the advertising experience.

RedBalloon, a gift and experience company in Australia, managed to reduce spending per lead acquisition by 30%, with a 3,000% return on some campaigns. How was it possible? The AI model was able to discover micro-audiences that humans at an ad agency simply wouldn't have time to target (for example, men over 65 in Adelaide who love flying). Advertising dollars went a lot further than the $50 per lead acquired in the past and marketing teams had more free time to make more strategic decisions, while the AI model did the grunt work.

Quantitative Fund Analysis: Morningstar

Morningstar, the company that assigns benchmark ratings for investments, has a large team of analysts to examine fund performance. As big as this team is, they weren't large enough to provide Analyst Ratings for every fund in the U.S.

When Morningstar turned to machine learning to create Quantitative Ratings, the company could review fund performance for six times as many funds. This Quantitative Rating harnesses the power of the wisdom-of-crowds effect. The machine learning algorithm is fed current and historical analyst ratings, and the data used to arrive at those ratings. The final model is a collection of thousands of models that together can create a picture of how an analyst would likely rate a fund.

Finding Medicines, Faster: Bristol Myers Squibb

In the quest to cure cancer, machine learning — combined with biology, statistics, and engineering — saves critical time for medical researchers. Using these combined disciplines, they can sift through billions of DNA sequences and massive amounts of clinical trial data to synthesize and focus on hypotheses that have greater chances of succeeding. Researchers at Bristol Myers Squibb have been using machine learning to generate models that show how therapies interact with the body to predict how medicines could behave in clinical trials. The time savings is huge.

In the past, researchers would study these drug combinations one by one. With next-generation models, they can weed out incompatible pairings, which could save decades of testing. These processes not only generate new data sets, they unlock insights that have never been analyzed before.

Understanding And Preventing Downtime In Manufacturing: Georgia-Pacific

When data science meets manufacturing, machine learning models can save money and, in this case, create safer work environments. At Georgia-Pacific, the giant paper manufacturer, data scientists identified a way to reduce paper tears by 40%. They combined data on paper roll quality with the time it takes for paper to tear to create precise schedules for the company's converting lines. This knowledge allowed Georgia-Pacific to save millions on a single production line, which they can now apply to 150 other lines.

They've also learned how to predict equipment failure up to three months in advance. By analyzing machinery performance across time, they have been able to anticipate unplanned downtime, improving asset utilization and safety in their paper mills.

Preventing Fraud In Banking: Wells Fargo

AI and machine learning have transformed the way banks detect — and now predict — fraudulent transactions and money laundering. By combining automated processes with multilayered, deep-learning analysis, companies spend less money to prevent fraud.

Where binary transaction monitoring systems generated warnings with high rates of false positives, AI has reduced those false positives by half. The cost savings also come from focusing human efforts on leads that are less likely to be fraudulent. Where artificial neural networks are used, companies can detect criminals who would otherwise know their way around the binary, rule-based security systems. The result? Banks can predict criminals' next moves.

Harvard Business Review says AI could become a requirement for fraud detection and prevention for larger businesses because there is "no other way to rapidly detect and interpret patterns across billions of pieces of data."

For Wells Fargo, it's not about examining individual transaction points, it's about observing behaviors along a continuum of transactions. They're using AI to scour vast amounts of online data, including on the dark web, to identify anti-money laundering signals to make connections humans wouldn't consider.

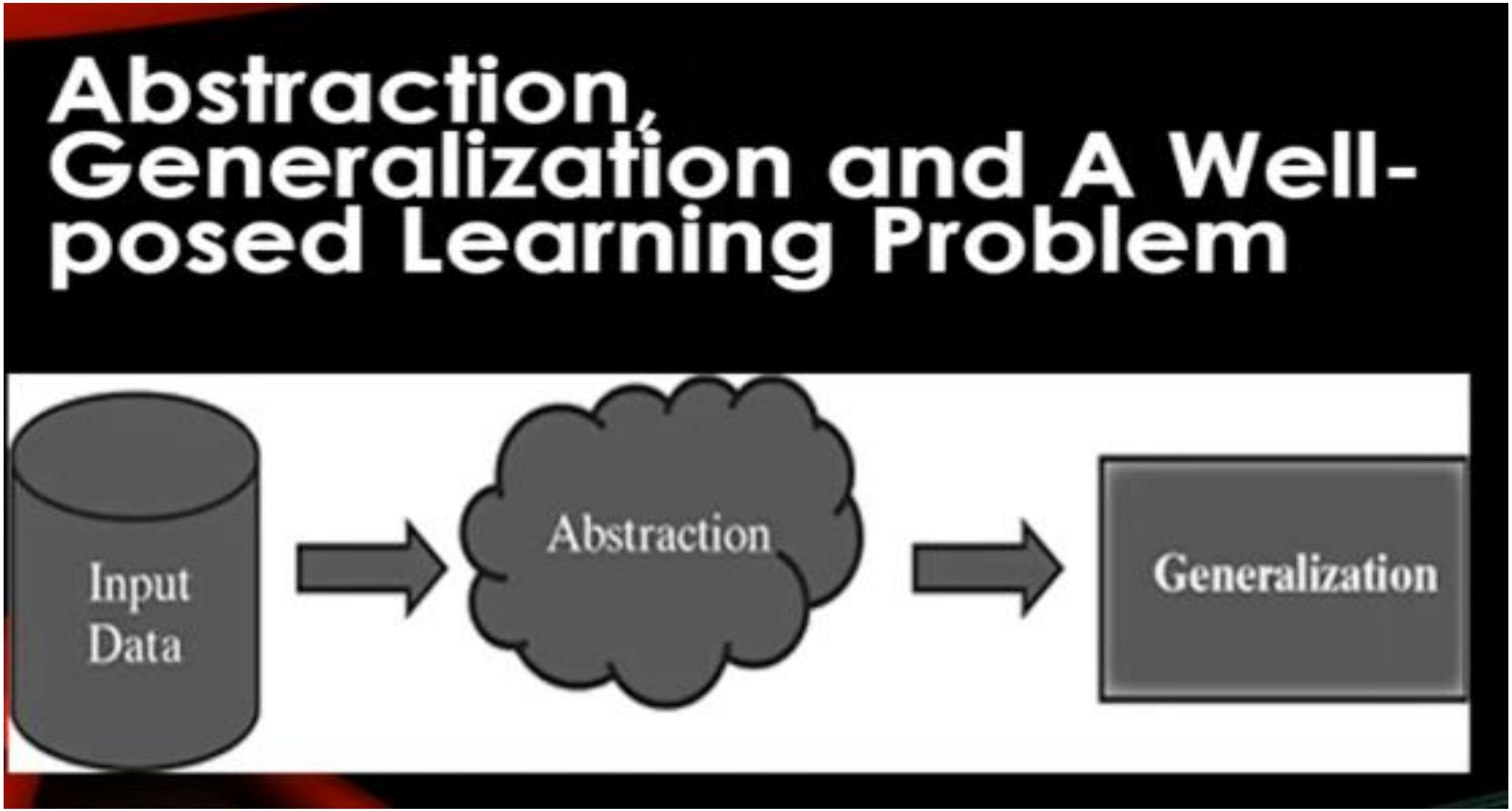
Personalize Experiences, Disrupt An Industry: Netflix

The next time you watch via a friend's Netflix account, compare the imagery on the screen to your own. You might notice the same titles displayed, but in a different order, and with different imagery. *Stranger Things*, for example, could be displayed with an image of teenagers, a spooky landscape with an ominous creature, or a picture of a character's bloody nose.

Why the difference? Netflix employs user data to personalize the way content is displayed for 140 million products. Machine learning guides each unique view to display titles and images with the goal to anticipate which title you'd like to watch based on your watch history.

As you know, Netflix turned the television industry upside down. They did this, in part, by curating user experiences. Netflix employed machine learning to assign user to data clusters, which considered when programs were watched, users' ages, genders, times spent watching, and how often they paused and resumed programs.

- **Abstraction and Knowledge representation**



ABSTRACTION

- During the machine learning process, knowledge is fed in the form of input data.
- However, the data cannot be used in the original shape and form.
- Abstraction helps in deriving a conceptual map based on the input data.
- This map, or a model as it is known in the machine learning paradigm, is summarized knowledge representation of the raw data.

ABSTRACTION

- The model may be in any one of the following forms
  - Computational blocks like if/else rules
  - Mathematical equations
  - Specific data structures like trees or graphs
  - Logical groupings of similar observations

ABSTRACTION

- The choice of the model used to solve a specific learning problem is a human task. The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:
  - The type of problem to be solved: Whether the problem is related to forecast or prediction, analysis of trend, understanding the different segments or groups of objects, etc.
  - Nature of the input data: How exhaustive the input data is, whether the data has no values for many fields, the data types, etc.
  - Domain of the problem: If the problem is in a business critical domain with a high rate of data input and need for immediate inference, e.g. fraud detection problem in banking domain.

ABSTRACTION

- Once the model is chosen, the next task is to fit the model based on the input data. Let's understand this with an example.
- In a case where the model is represented by a mathematical equation, say '$y = c_1 + c_2x$' (the model is known as simple linear regression which we will study in a later chapter), based on the input data, we have to find out the values of c_1 and c_2.

ABSTRACTION

- . So, fitting the model, in this case, means finding the values of the unknown coefficients or constants of the equation or the model.
- This process of fitting the model based on the input data is known as training.
- Also, the input data based on which the model is being finalized is known as training data.

GENERALIZATION

- The first part of machine learning process is abstraction i.e. abstract the knowledge which comes as input data in the form of a model.
- However, this abstraction process, or more popularly training the model, is just one part of machine learning.
- The other key part is to tune up the abstracted knowledge to a form which can be used to take future decisions.
- This is achieved as a part of generalization. This part is quite difficult to achieve.
- This is because the model is trained based on a finite set of data, which may possess a limited set of characteristics.

11:28 / 22:57 • Generalization

GENERALIZATION

- But when we want to apply the model to take decision on a set of unknown data, usually termed as test data, we may encounter two problems:
  1. The trained model is aligned with the training data too much, hence may not portray the actual trend.
  2. The test data possess certain characteristics apparently unknown to the training data.

GENERALIZATION

- Hence, a precise approach of decision-making will not work.
- An approximate or heuristic approach, much like gut-feeling-based decision-making in human beings, has to be adopted.
- This approach has the risk of not making a correct decision – quite obviously because certain assumptions that are made may not be true in reality.
- But just like machines, same mistakes can be made by humans too when a decision is made based on intuition or gut-feeling – in a situation where exact reason-based decision-making is not possible.

Well posed problems

A problem is well posed if:

- A solution exists.
- The solution is unique.
- The solution depends continuously on the data (boundary and/or initial conditions).
- Problems which do not fulfill these criteria are ill-posed.

Well posed problems have a good chance to be solved numerically with a stable algorithm.

1865-1963

Ill-posed problems

Ill-posed problems play an important role in some areas, for example for inverse problems like tomography.

Problem needs to be reformulated for numerical treatment. Add additional constraints, for example smoothness of the solution.

Input data need to be regularized / preprocessed.

Ill-conditioned problems

- Even well posed problems can be ill-conditioned.
- Small changes (errors, noise) in data lead to large errors in the solution.
- Can occur if continuous problems are solved approximately on a numerical grid.
- PDE => algebraic equation in form $Ax = b$
- Condition number of matrix A:

$$\kappa(A) = \left| \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \right|$$

$\lambda_{\max}(A)$, $\lambda_{\min}(A)$ are maximal and minimal Eigen values of A.

- Well conditioned problems have a low condition number.

Well-Posed Learning Problem

Problem: A task which need to be solved.

Learning problem: The problems that can be effectively solved using **machine learning**

Well-posed Learning Problem: A learning problem which have three features:

- **Task (T) –** Objective.
- **Experience (E)-** Dataset from which algorithm can learn.
- **Performance (P) -** Metric used to assess the effectiveness of the algorithm at the task.

A robot driving learning problem:

- Task (T): driving on public four-lane highways using vision sensors
- Experience (E): a sequence of images and steering commands recorded while observing a human driver.
- Performance measure (P): average distance traveled before an error (as judged by human overseer).

A handwriting recognition learning problem:

- Task (T): recognizing and classifying handwritten words within images.
- Experience (E):a database of handwritten words with given classifications .
- Performance measure (P): percent of words correctly classified.

Factors to be considered

1. Understanding Data

- Generally, it is recommended to gather a huge amount of data to get better accuracy.
- However, it is not always possible to gather such a massive amount of data.
- Hence choosing an algorithm that has high bias and low variance such as Linear regression, Naïve Bayes can be very effective.
- On the other hand, if the training data is sufficiently large then choosing an algorithm that has low bias and high variance such as KNN, Decision trees can be a smart choice.

2. Accuracy

- Technically the definition of accuracy is “the degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard”.
- It gives a measure of how a model is able to truly predict a response value for a given input.
- Often interpretability of the model decreases with an increase in the efficiency of the model.
- This is due to the change in flexibility of the model and thus complex models can generate and map a wider range of possible input values.
- For example, KNN with k=1 is highly flexible when compared to a KNN with K=5.
- The selection of K is a highly subjective matter and is to be addressed as per the business application. Decreasing the value of “K” may give better accuracy but will decrease the interpretability of the model drastically.

3. Speed or Training time

- Realistically, algorithms require more time to train on large training data. Higher the accuracy, the higher the training time.
- Also, In real-world applications, the choice of algorithm is driven by these two factors predominantly.
- **ML Algorithms like Linear regression and Logistic regression take less time when compared to algorithms like SVM, Neural networks, random forests.**

4. Linearity

- Algorithms such as logistic regression and support vector machines assume that types can be separated by a straight line.
- Thus if the data is linear, then these algorithms have good performance.
- For nonlinear data algorithms such as SVM, random forest, neural nets work well, as these ML Algorithms can handle nonlinearity and high dimensional complex data structures.
- The best way to find out the linearity is to try different algorithms.

5. Number of features

- A dataset may have many parameters and for a certain business application, certain types of fields may be able to address it.
- Other features may not be relevant to the application and can create imbalance leading to inefficiency.
- When many features are important to the business and need to be considered, ML Algorithms such as SVM[Support Vector Machine] are better suited.

- Some of the main aspects to consider when trying to solve a new problem are:
- **The Objective of the problem,Categorize the problem.**
- **Understand Your Data, Find the appropriate ML Algorithms.**
- **Implementation of different machine learning algorithms.**
- Optimizing of hyperparameters.

HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

EAL51501 – ARTIFICIAL INTELLIGENCE
B.Tech[AIML] – III Semester

K.Kowsalya
Assistant Professor (SS)
School of Computing Sciences,
Department of Computer Science and Engineering

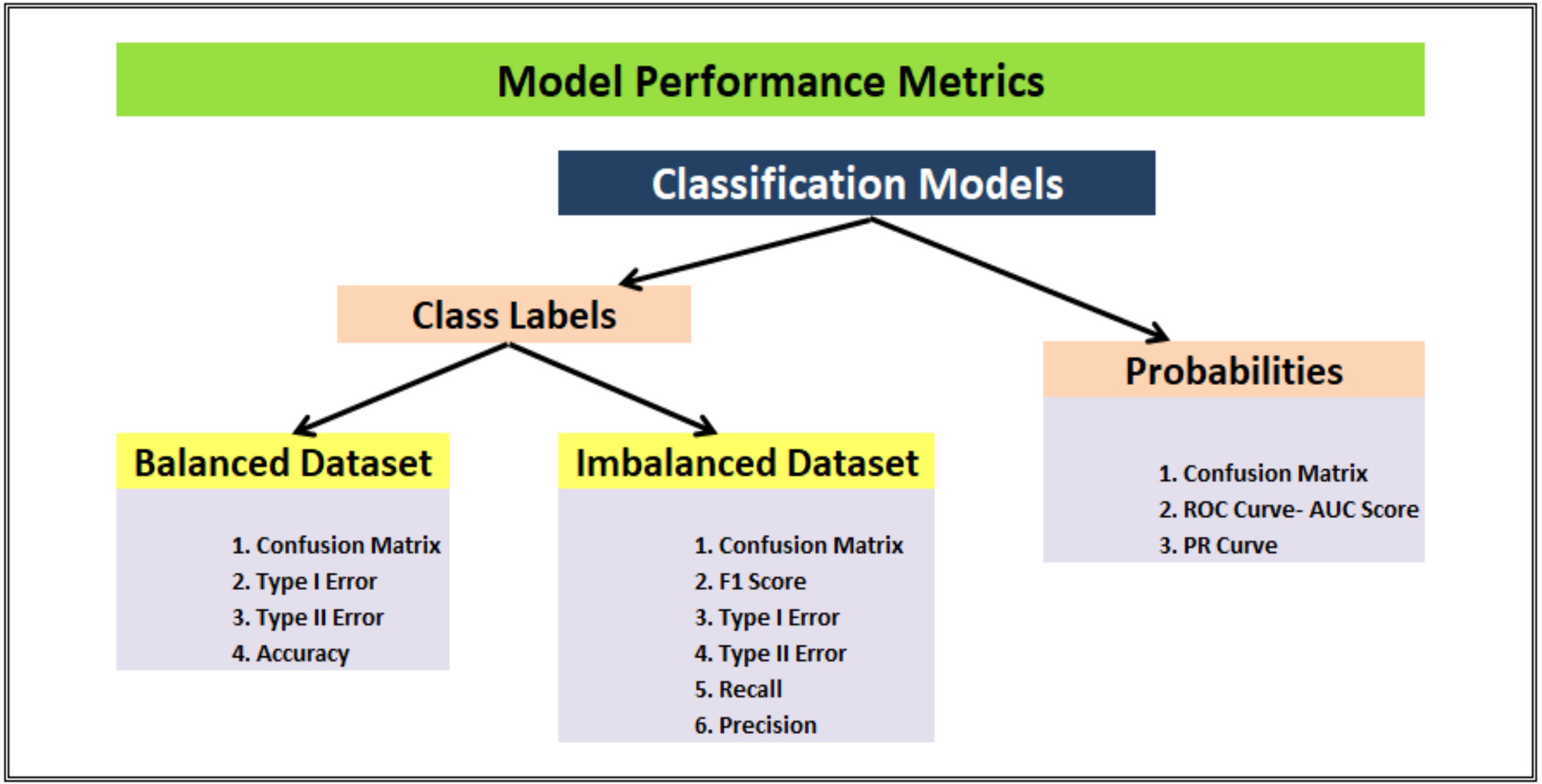
UNIT-III

- Motivation for Machine Learning, Applications, Machine Learning, Learning associations, Classification, Regression, The Origin of machine learning, Uses and abuses of machine learning, Success cases, How do machines learn[INTRO], Abstraction and knowledge representation, Generalization, Factors to be considered, **Assessing the success of learning, Metrics for evaluation of classification method,** Steps to apply machine learning to data, Machine learning process, Input data and ML algorithm, Classification of machine learning algorithms, General ML architecture, Group of algorithms, Reinforcement learning, Supervised learning, Unsupervised learning, Semi-Supervised learning, Algorithms, Ensemble learning, Matching data to an appropriate algorithm.


AUC
Precision score
Recall
Model Evaluation Metrics
Confusion matrix
F1 Score
Loss

- ***To evaluate the performance or quality of the model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics.***
- These performance metrics help us understand how well our model has performed for the given data.
- In this way, we can improve the model's performance by tuning the hyper-parameters.
- Each ML model aims to generalize well on unseen/new data, and performance metrics help determine how well the model generalizes on the new dataset.

- In machine learning, each task or problem is divided into **classification** and **Regression**.
- Not all metrics can be used for all types of problems; hence, it is important to know and understand which metrics should be used.
- Different evaluation metrics are used for both Regression and Classification tasks.


Model Performance Metrics
Classification Models
Class Labels
Probabilities
Balanced Dataset
1. Confusion Matrix
2. Type I Error
3. Type II Error
4. Accuracy
Imbalanced Dataset
1. Confusion Matrix
2. F1 Score
3. Type I Error
4. Type II Error
5. Recall
6. Precision
1. Confusion Matrix
2. ROC Curve- AUC Score
3. PR Curve

1. Performance Metrics for Classification

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

2. Performance Metrics for Regression

- **Mean Absolute Error**
- **Mean Squared Error**
- **R2 Score**

Classification Accuracy

It is most common performance metric for classification algorithms. It may be defined as the number of correct predictions made as a ratio of all predictions made. We can easily calculate it by confusion matrix with the help of following formula –

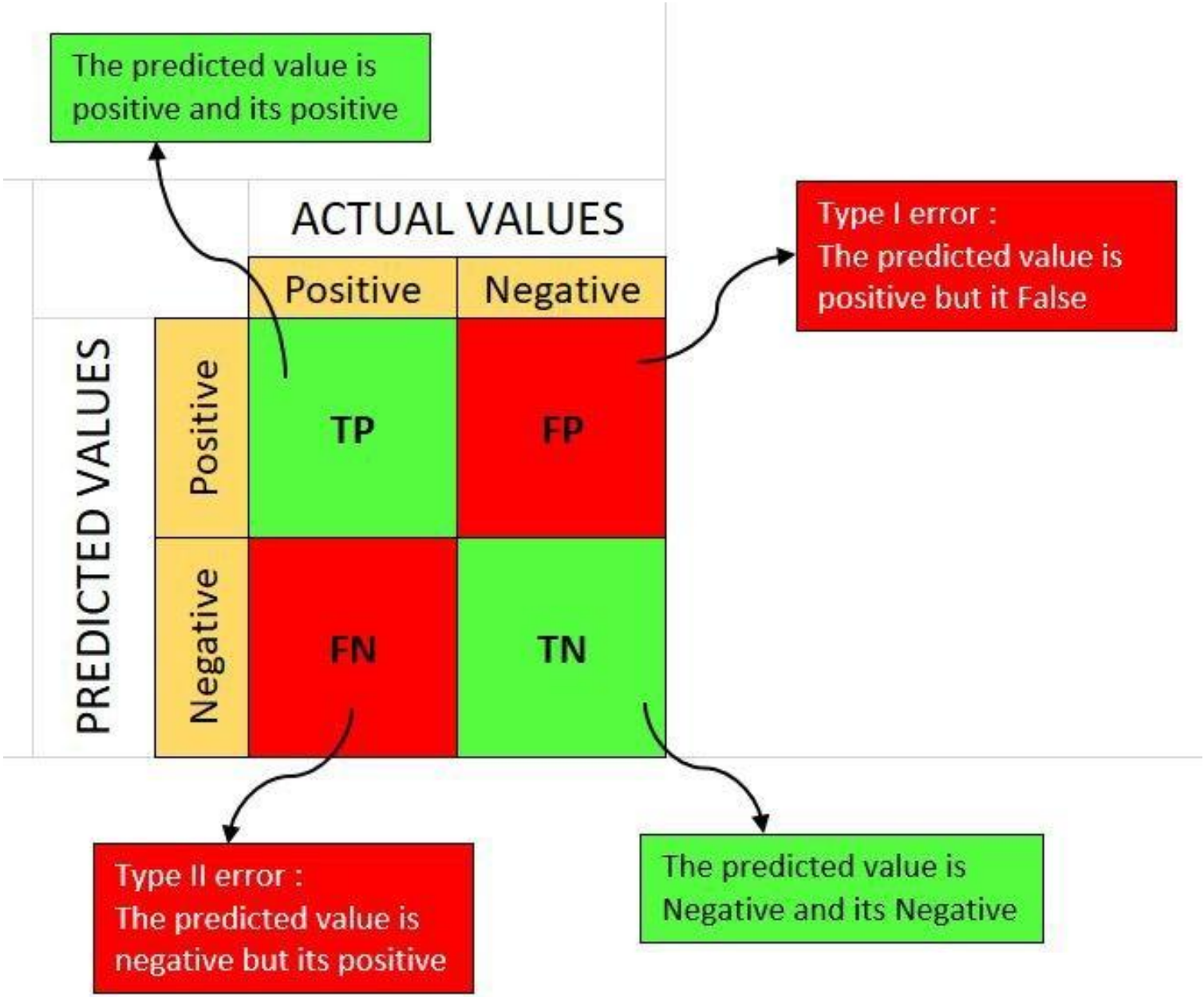
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **When to Use Accuracy?**
- It is good to use the Accuracy metric when the target variable classes in data are approximately balanced.
- For example, if 60% of classes in a fruit image dataset are of Apple, 40% are Mango.
- In this case, if the model is asked to predict whether the image is of Apple or Mango, it will give a prediction with 97% of accuracy.

- **When not to use Accuracy?**
- It is recommended not to use the Accuracy measure when the target variable majorly belongs to one class.
- For example, Suppose there is a model for a disease prediction in which, out of 100 people, only five people have a disease, and 95 people don't have one.
- In this case, if our model predicts every person with no disease (which means a bad prediction), the Accuracy measure will be 95%, which is not correct.

Confusion Matrix

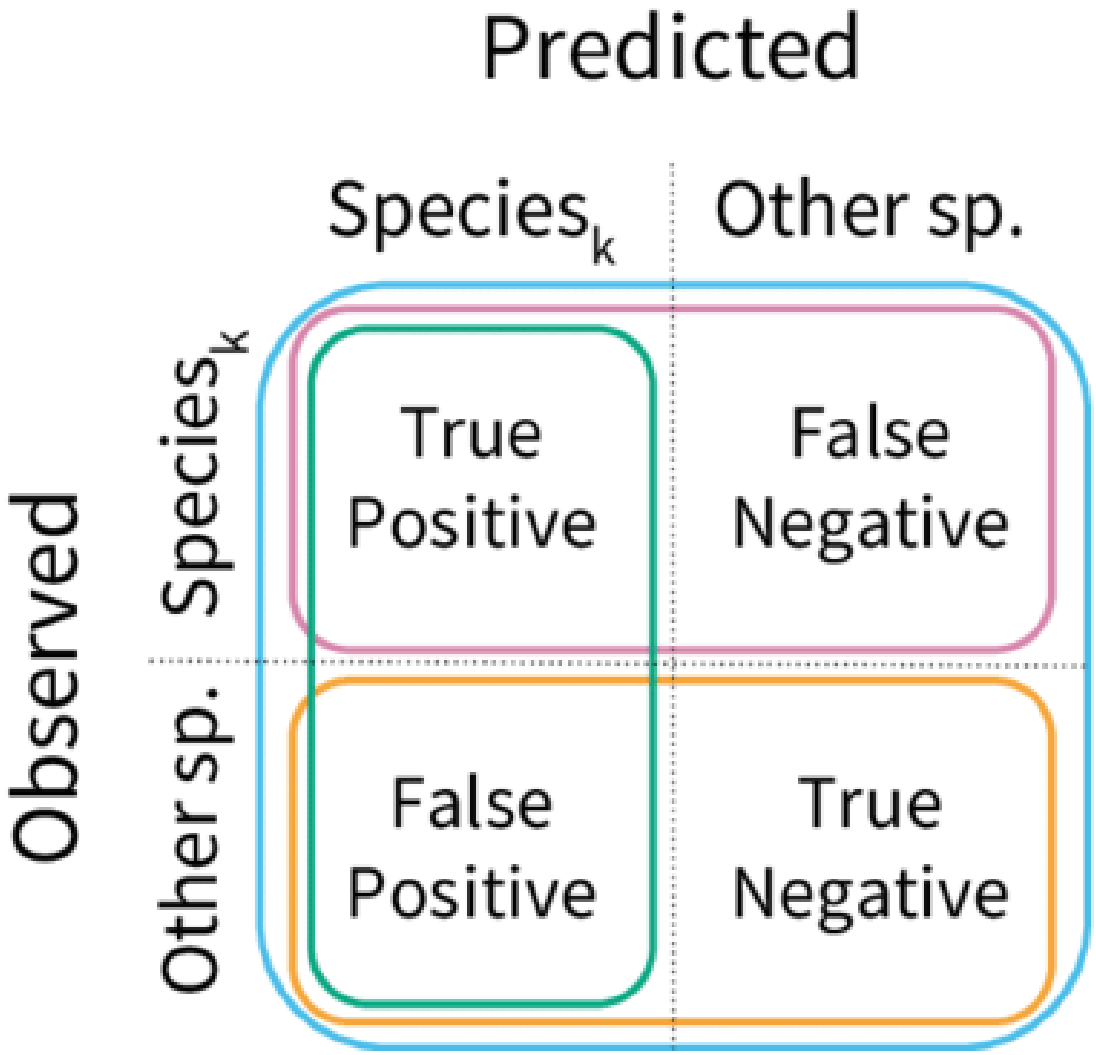
It is the easiest way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is nothing but a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, “False Negatives (FN)” as shown below –


The predicted value is positive and its positive
ACTUAL VALUES
Positive
Negative
PREDICTED VALUES
Positive
Negative
TP
FP
FN
TN
Type I error :
The predicted value is positive but it False
Type II error :
The predicted value is negative but its positive
The predicted value is Negative and its Negative

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

- We can determine the following from the above matrix:
- In the matrix, columns are for the prediction values, and rows specify the Actual values.
- Here Actual and prediction give two possible classes, Yes or No.
- So, if we are predicting the presence of a disease in a patient, the Prediction column with Yes means, Patient has the disease, and for NO, the Patient doesn't have the disease.
- In this example, the total number of predictions are 165, out of which 110 time predicted yes, whereas 55 times predicted No.
- However, in reality, 60 cases in which patients don't have the disease, whereas 105 cases in which patients have the disease.

- In general, the table is divided into four terminologies, which are as follows:
- **True Positive(TP):** In this case, the prediction outcome is true, and it is true in reality, also.
- True Negative(TN): in this case, the prediction outcome is false, and it is false in reality, also.
- False Positive(FP): In this case, prediction outcomes are true, but they are false in actuality.
- False Negative(FN): In this case, predictions are false, and they are true in actuality.

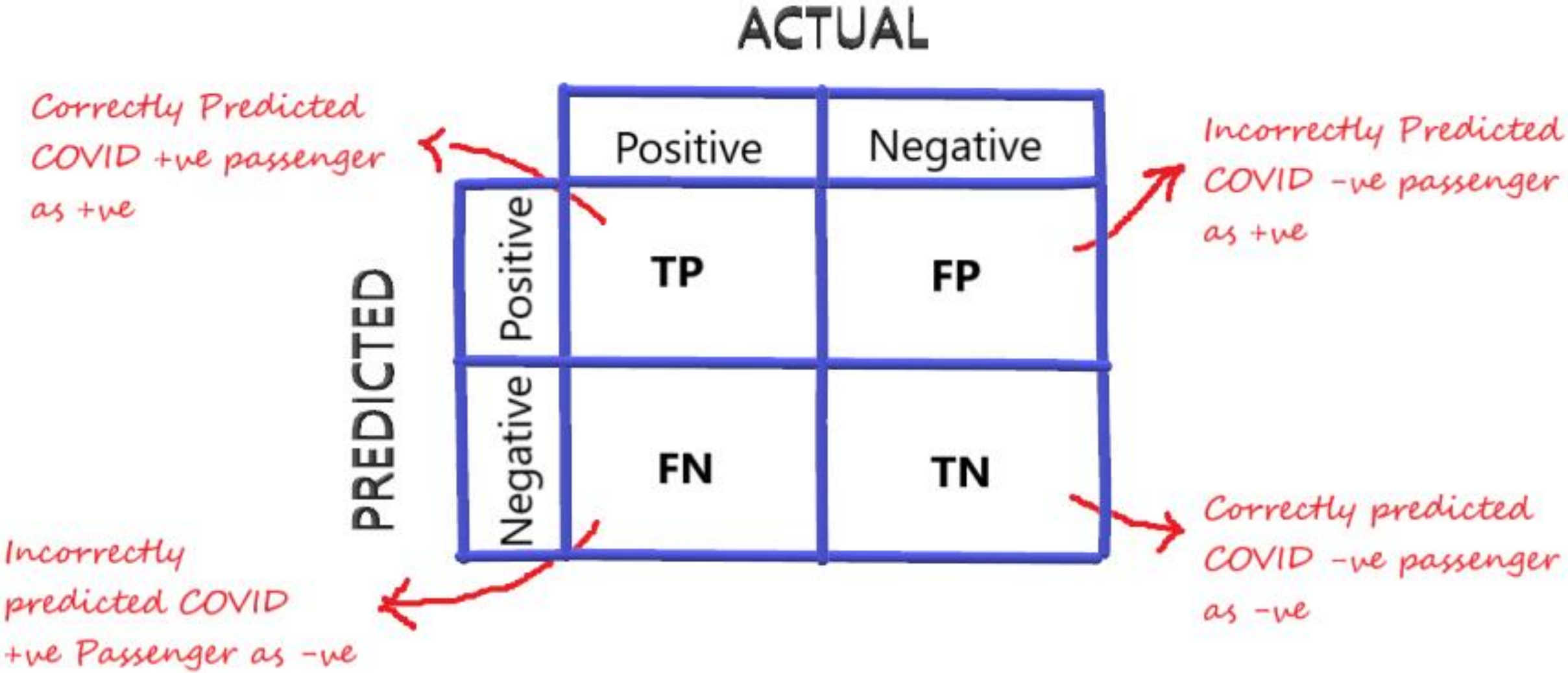


$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$


ACTUAL
Positive
Negative
PREDICTED
Positive
Negative
TP
FP
FN
TN
Correctly Predicted COVID +ve passenger as +ve
Incorrectly Predicted COVID -ve passenger as +ve
Incorrectly predicted COVID +ve Passenger as -ve
Correctly predicted COVID -ve passenger as -ve

Explanation of the terms associated with confusion matrix are as follows –

- **True Positives (TP)** – It is the case when both actual class & predicted class of data point is 1.
- **True Negatives (TN)** – It is the case when both actual class & predicted class of data point is 0.
- **False Positives (FP)** – It is the case when actual class of data point is 0 & predicted class of data point is 1.
- **False Negatives (FN)** – It is the case when actual class of data point is 1 & predicted class of data point is 0.

- **Precision**
- The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct.
- It can be calculated as the True Positive or predictions that are actually true to the total positive predictions (True Positive and False Positive).

$$Precision = \frac{TP}{(TP + FP)}$$

- **Recall or Sensitivity**
- It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly.
- It can be calculated as True Positive or predictions that are actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).
- The formula for calculating Recall is given below:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **When to use Precision and Recall?**

- *if we maximize precision, it will minimize the FP errors, and if we maximize recall, it will minimize the FN error.*

Precision

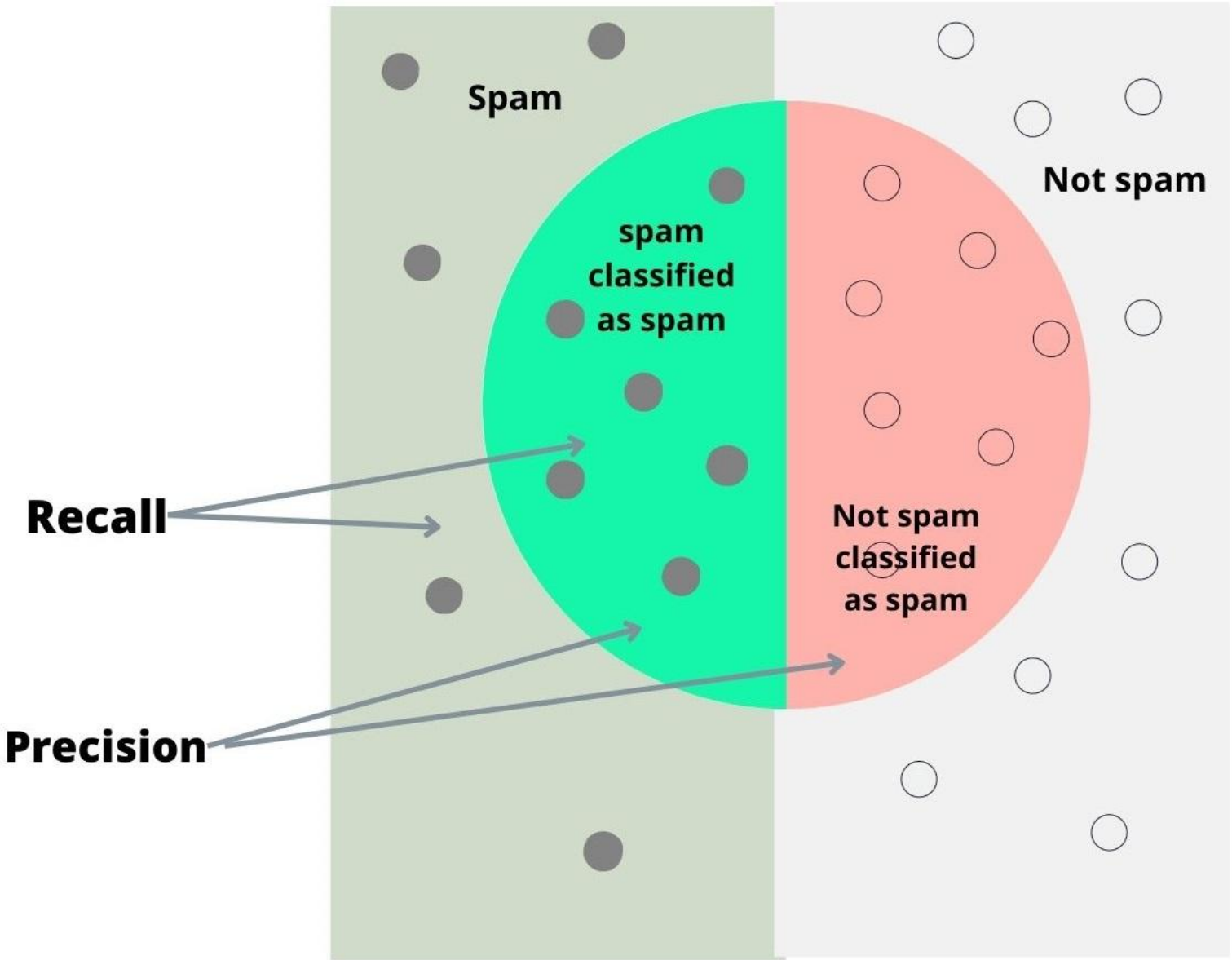
Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula –

$$Precision = \frac{TP}{TP + FP}$$

Recall or Sensitivity

Recall may be defined as the number of positives returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula –

$$Recall = \frac{TP}{TP + FN}$$


Spam
Not spam
spam
classified
as spam
Not spam
classified
as spam
Recall
Precision

F1 Score

This score will give us the harmonic mean of precision and recall. Mathematically, F1 score is the weighted average of the precision and recall. The best value of F1 would be 1 and worst would be 0. We can calculate F1 score with the help of following formula −

$$\boldsymbol{F1 = 2 * (precision * recall) / (precision + recall)}$$

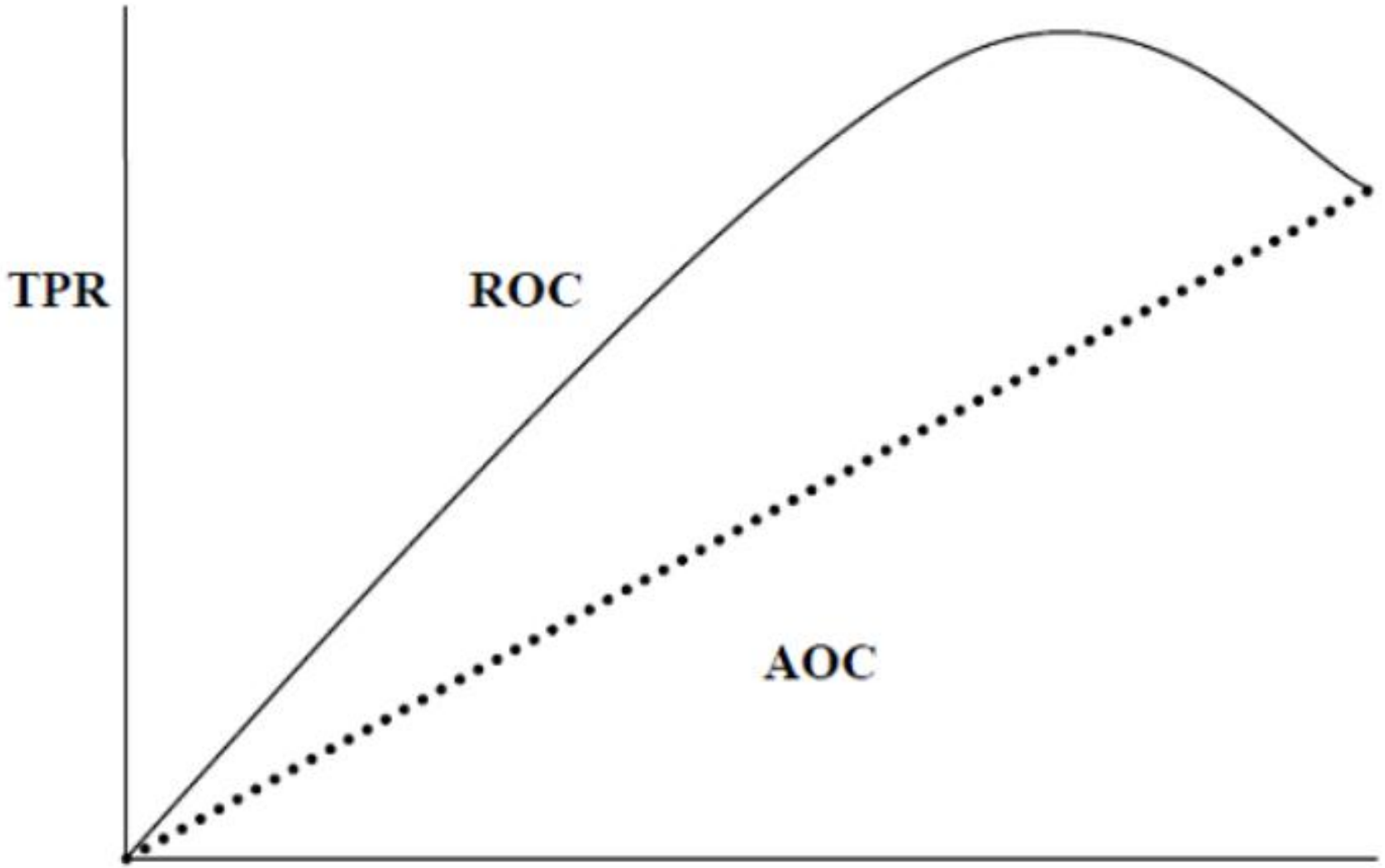
F1 score is having equal relative contribution of precision and recall.

We can use classification_report function of sklearn.metrics to get the classification report of our classification model.

AUC (Area Under ROC curve)

AUC (Area Under Curve)-ROC (Receiver Operating Characteristic) is a performance metric, based on varying threshold values, for classification problems. As name suggests, ROC is a probability curve and AUC measure the separability. In simple words, AUC-ROC metric will tell us about the capability of model in distinguishing the classes. Higher the AUC, better the model.

Mathematically, it can be created by plotting TPR (True Positive Rate) i.e. Sensitivity or recall vs FPR (False Positive Rate) i.e. 1-Specificity, at various threshold values. Following is the graph showing ROC, AUC having TPR at y-axis and FPR at x-axis –


TPR
ROC
AOC

Performance Metrics for Regression

- **Mean Absolute Error**
- **Mean Squared Error**
- **R2 Score**

Mean Absolute Error (MAE)

It is the simplest error metric used in regression problems. It is basically the sum of average of the absolute difference between the predicted and actual values. In simple words, with MAE, we can get an idea of how wrong the predictions were. MAE does not indicate the direction of the model i.e. no indication about underperformance or overperformance of the model. The following is the formula to calculate MAE –

$$MAE = \frac{1}{n}\sum |Y - \hat{Y}|$$

Here, Y=Actual Output Values

And $\hat{Y}$= Predicted Output Values.

Mean Square Error (MSE)

MSE is like the MAE, but the only difference is that the it squares the difference of actual and predicted output values before summing them all instead of using the absolute value. The difference can be noticed in the following equation –

$$MSE = \frac{1}{n}\sum(Y - \hat{Y})$$

Here, Y=Actual Output Values

And $\hat{Y}$ = Predicted Output Values.

We can use mean_squared_error function of sklearn.metrics to compute MSE.

R Squared (R^2)

R Squared metric is generally used for explanatory purpose and provides an indication of the goodness or fit of a set of predicted output values to the actual output values. The following formula will help us understanding it –

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y_i})^2}$$

In the above equation, numerator is MSE and the denominator is the variance in Y values.

HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

EAL51501 – ARTIFICIAL INTELLIGENCE
B.Tech[AIML] – III Semester

K.Kowsalya
Assistant Professor (SS)
School of Computing Sciences,
Department of Computer Science and Engineering

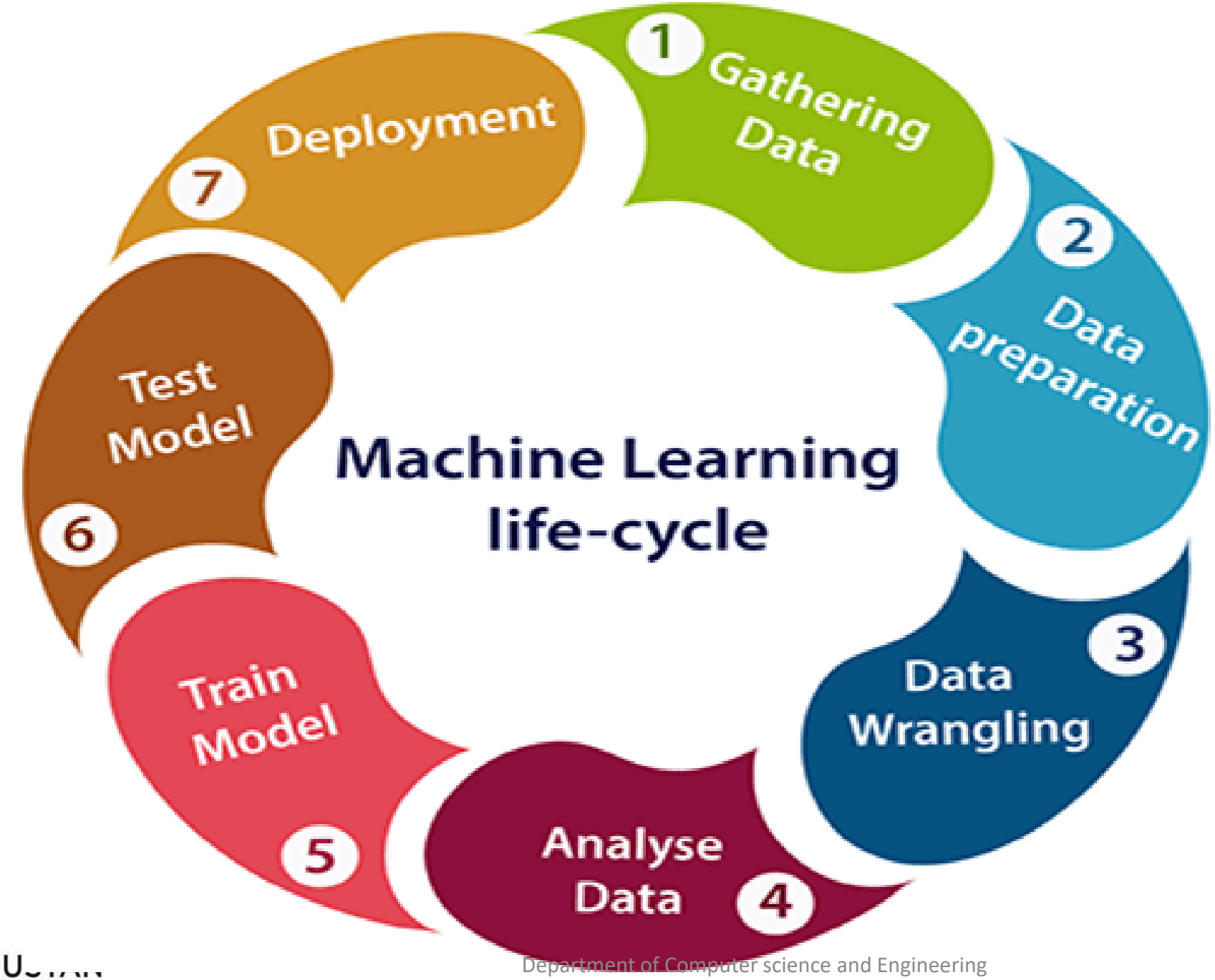
UNIT-III

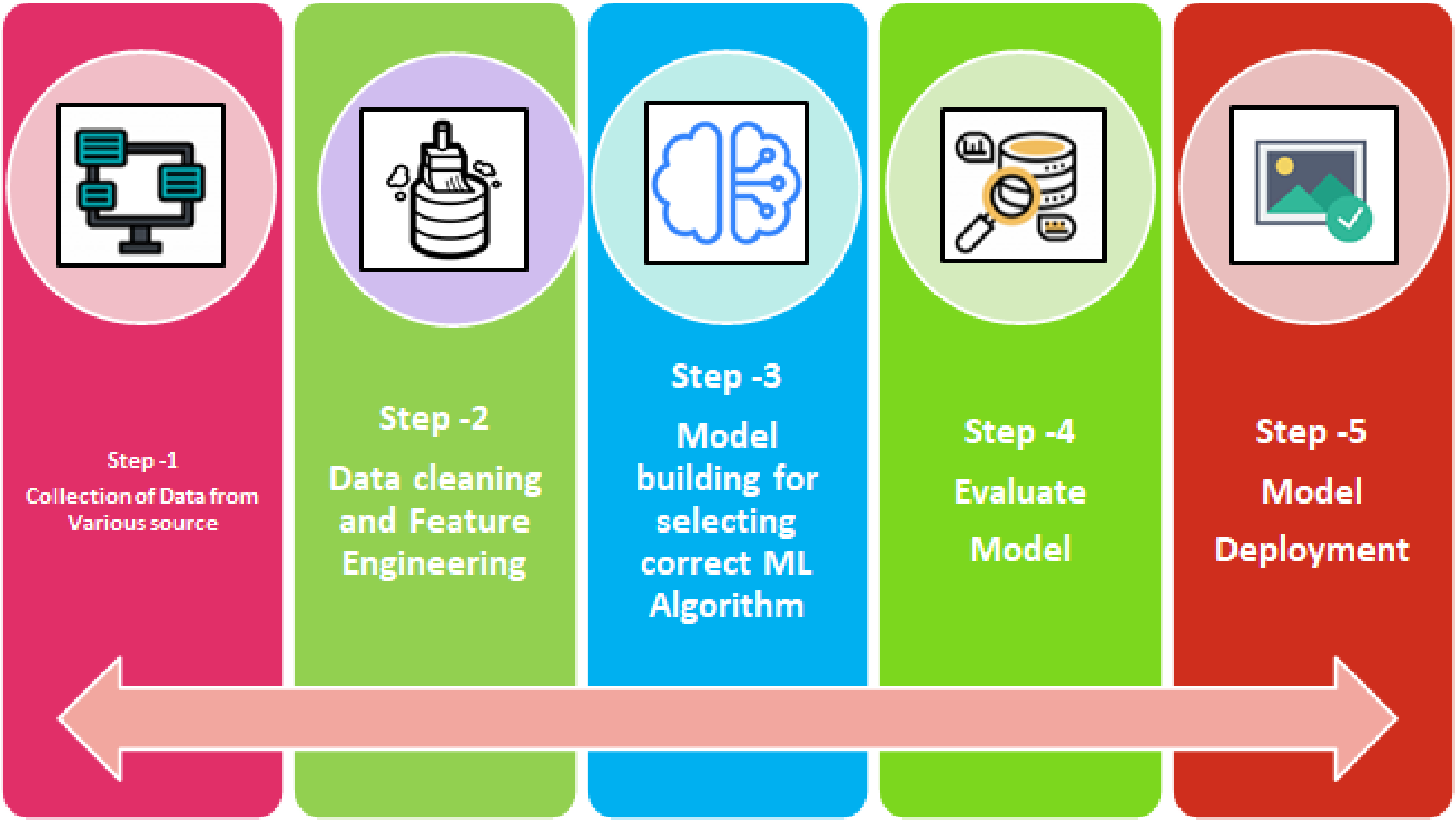
- Motivation for Machine Learning, Applications, Machine Learning, Learning associations, Classification, Regression, The Origin of machine learning, Uses and abuses of machine learning, Success cases, How do machines learn[INTRO], Abstraction and knowledge representation, Generalization, Factors to be considered, Assessing the success of learning, Metrics for evaluation of classification method, **Steps to apply machine learning to data, Machine learning process, Input data and ML algorithm, Classification of machine learning algorithms, General ML architecture,** Group of algorithms, Reinforcement learning, Supervised learning, Unsupervised learning, Semi-Supervised learning, Algorithms, Ensemble learning, Matching data to an appropriate algorithm.

- **Steps to apply machine learning to data,Machine Learning Process**
- Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project.

Machine learning life cycle involves seven major steps, which are given below:

1. **Gathering Data**
2. **Data Preparation**
3. **Data Wrangling**
4. **Analyse Data**
5. **Train the model**
6. **Test the model**
7. **Deployment**


Machine Learning life-cycle
1 Gathering Data
2 Data preparation
3 Data Wrangling
4 Analyse Data
5 Train Model
6 Test Model
7 Deployment


Step -1
Collection of Data from Various source
Step -2
Data cleaning and Feature Engineering
Step -3
Model building for selecting correct ML Algorithm
Step -4
Evaluate Model
Step -5
Model Deployment

- In the complete life cycle process, to solve a problem, we create a machine learning system called "model", and this model is created by providing "training".
- But to train a model, we need data, hence, life cycle starts by collecting data.

1. Gathering Data:

- Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.
- In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output.

- The more will be the data, the more accurate will be the prediction.
- This step includes the below tasks:
- **Identify various data sources**
- **Collect data**
- **Integrate the data obtained from different sources**

2. Data preparation

- After collecting the data, we need to prepare it for further steps.
- Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
- In this step, first, we put all data together, and then randomize the ordering of data.

- This step can be further divided into two processes:
- **Data exploration:**

  It is used to understand the nature of data that we have to work with.
- We need to understand the characteristics, format, and quality of data.

  A better understanding of data leads to an effective outcome.
- In this, we find Correlations, general trends, and outliers.
- **Data pre-processing:**

  Now the next step is preprocessing of data for its analysis.

- **Data Wrangling**
- Data wrangling is the process of cleaning and converting raw data into a useable format.
- It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step.
- It is one of the most important steps of the complete process.
- Cleaning of data is required to address the quality issues.
- It is not necessary that data we have collected is always of our use as some of the data may not be useful.

- In real-world applications, collected data may have various issues, including:
- **Missing Values**
- **Duplicate data**
- **Invalid data**
- **Noise**
- So, we use various filtering techniques[Neural networks ,SVM etc.] to clean the data.
- It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

4. Data Analysis

- Now the cleaned and prepared data is passed on to the analysis step. This step involves:
- **Selection of analytical techniques**
- **Building models**
- **Review the result**
- The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.
- It starts with the determination of the type of the problems, where we select the machine learning techniques such as **Classification**, **Regression**, **Cluster analysis**, **Association**, etc. then build the model using prepared data, and evaluate the model.
- Hence, in this step, we take the data and use machine learning algorithms to build the model.

5. Train Model

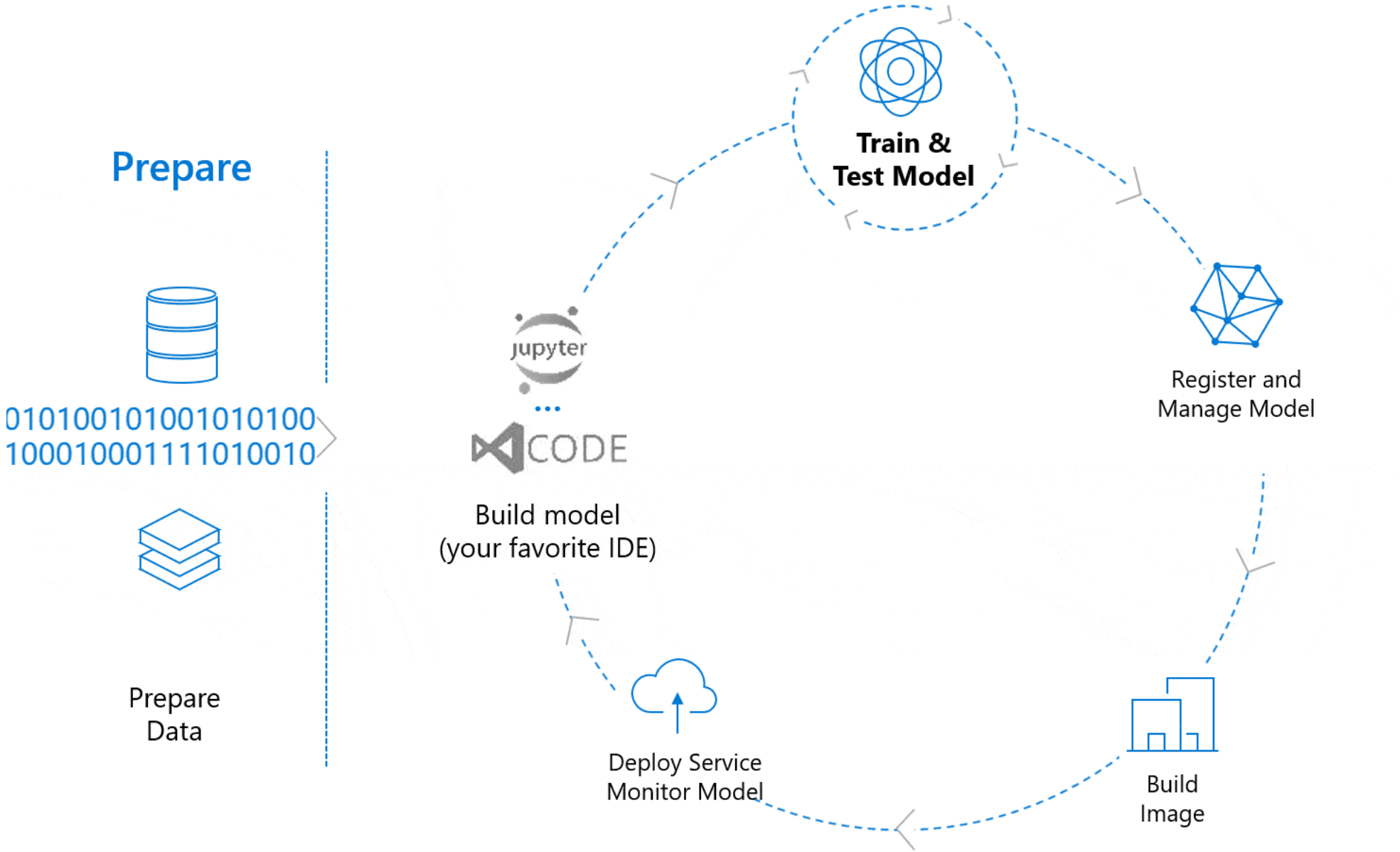
- Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.
- We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

6. Test Model

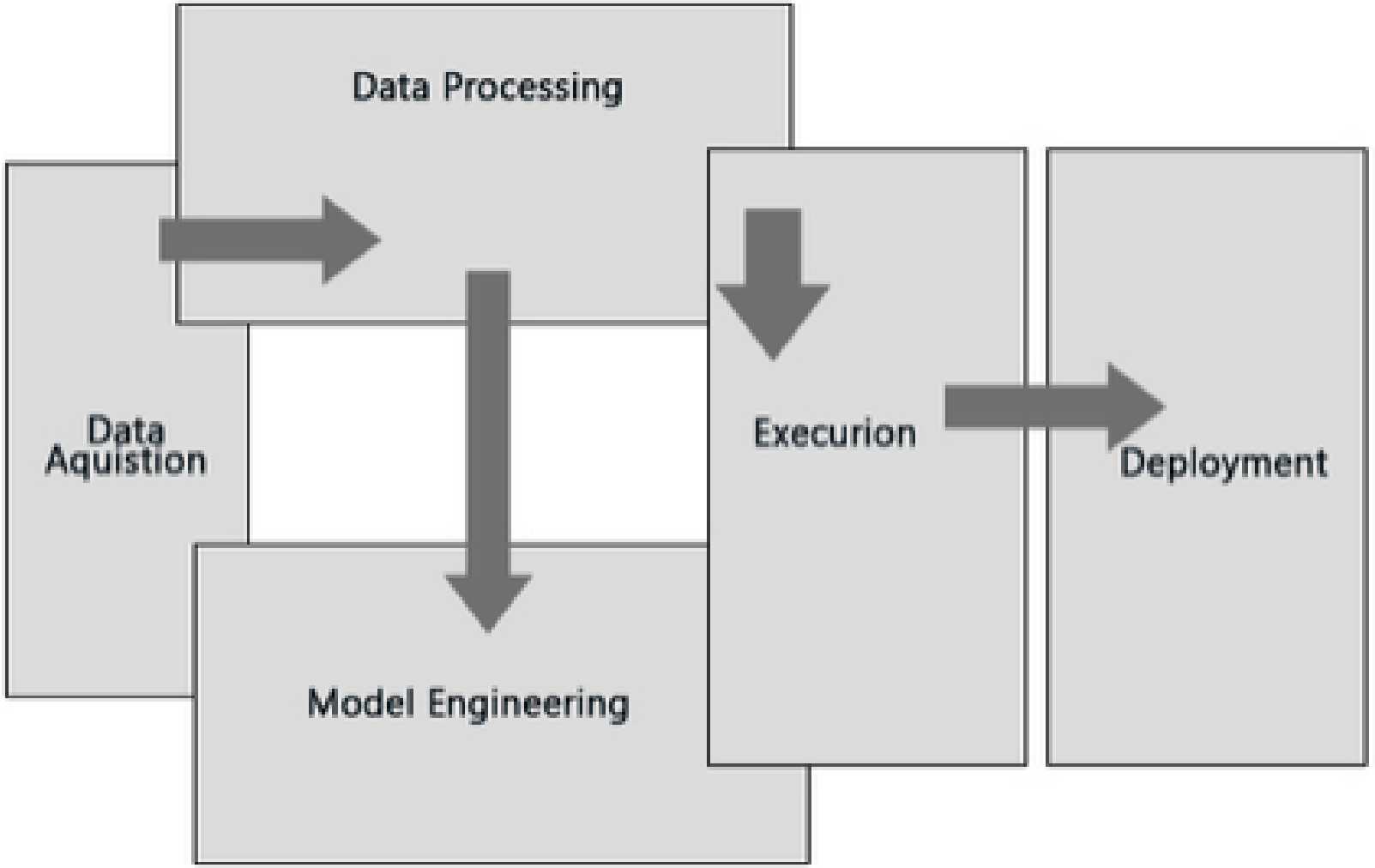
- Once our machine learning model has been trained on a given dataset, then we test the model.
- In this step, we check for the accuracy of our model by providing a test dataset to it.
- Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7. Deployment

- The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.
- If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system.
- But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is similar to making the final report for a project.


Prepare
01010010100101010010001000111101001 0
Prepare Data
jupyter
CODE
Build model (your favorite IDE)
Train & Test Model
Register and Manage Model
Build Image
Deploy Service Monitor Model

Machine Learning Process



1. Data Acquisition

- This involves data collection, preparing and segregating the case scenarios based on certain features involved with the decision making cycle and forwarding the data to the processing unit for carrying out further categorization.
- This stage is sometimes called the data preprocessing stage. The data model expects reliable, fast and elastic data which may be discrete or continuous in nature.
- The data is then passed into stream processing systems (for continuous data) and stored in batch data warehouses (for discrete data) before being passed on to data modeling or processing stages.

2. Data Processing

- The received data in the data acquisition layer is then sent forward to the data processing layer where it is subjected to advanced integration and processing and involves normalization of the data, data cleaning, transformation, and encoding.
- The data processing is also dependent on the type of learning being used.
- For e.g., if supervised learning is being used the data shall be needed to be segregated into multiple steps of sample data required for training of the system and the data thus created is called training sample data or simply training data.
- Also, the data processing is dependent upon the kind of processing required and may involve choices ranging from action upon continuous data which will involve the use of specific function-based architecture, for example, lambda architecture, Also it might involve action upon discrete data which may require memory-bound processing. The data processing layer defines if the memory processing shall be done to data in transit or in rest.

3. Data Modeling

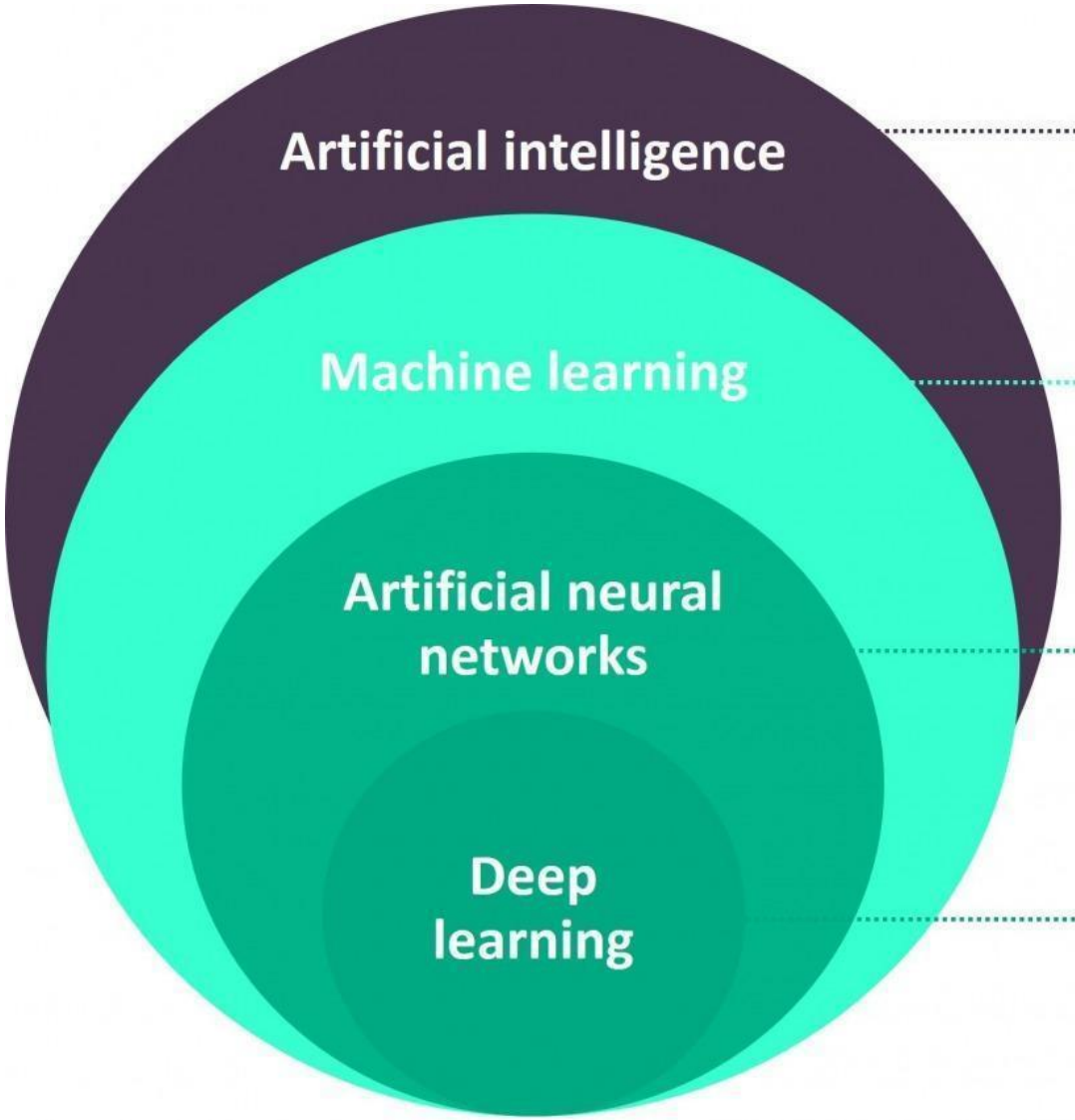
- This layer of the architecture involves the selection of different algorithms that might adapt the system to address the problem for which the learning is being devised, These algorithms are being evolved or being inherited from a set of libraries.
- The algorithms are used to model the data accordingly, this makes the system ready for the execution step.

4. Execution

- This stage in machine learning is where the experimentation is done, testing is involved and tunings are performed. The general goal behind being to optimize the algorithm in order to extract the required machine outcome and maximize the system performance, The output of the step is a refined solution capable of providing the required data for the machine to make decisions.

5. Deployment

- Like any other software output, ML outputs need to be operationalized or be forwarded for further exploratory processing. The output can be considered as a non-deterministic query which needs to be further deployed into the decision-making system.
- It is advised to seamlessly move the ML output directly to production where it will enable the machine to directly make decisions based on the output and reduce the dependency on the further exploratory steps.



Artificial intelligence (AI)

Any techniques that enable machines to solve a task in a way like humans do

Machine learning (ML)

Algorithms that allow computers to learn from examples without being explicitly programmed

Artificial neural networks (ANN)

Brain-inspired machine learning models

Deep learning (DL)

A subset of ML which uses deep artificial neural networks as models and automatically builds a hierarchy of data representations

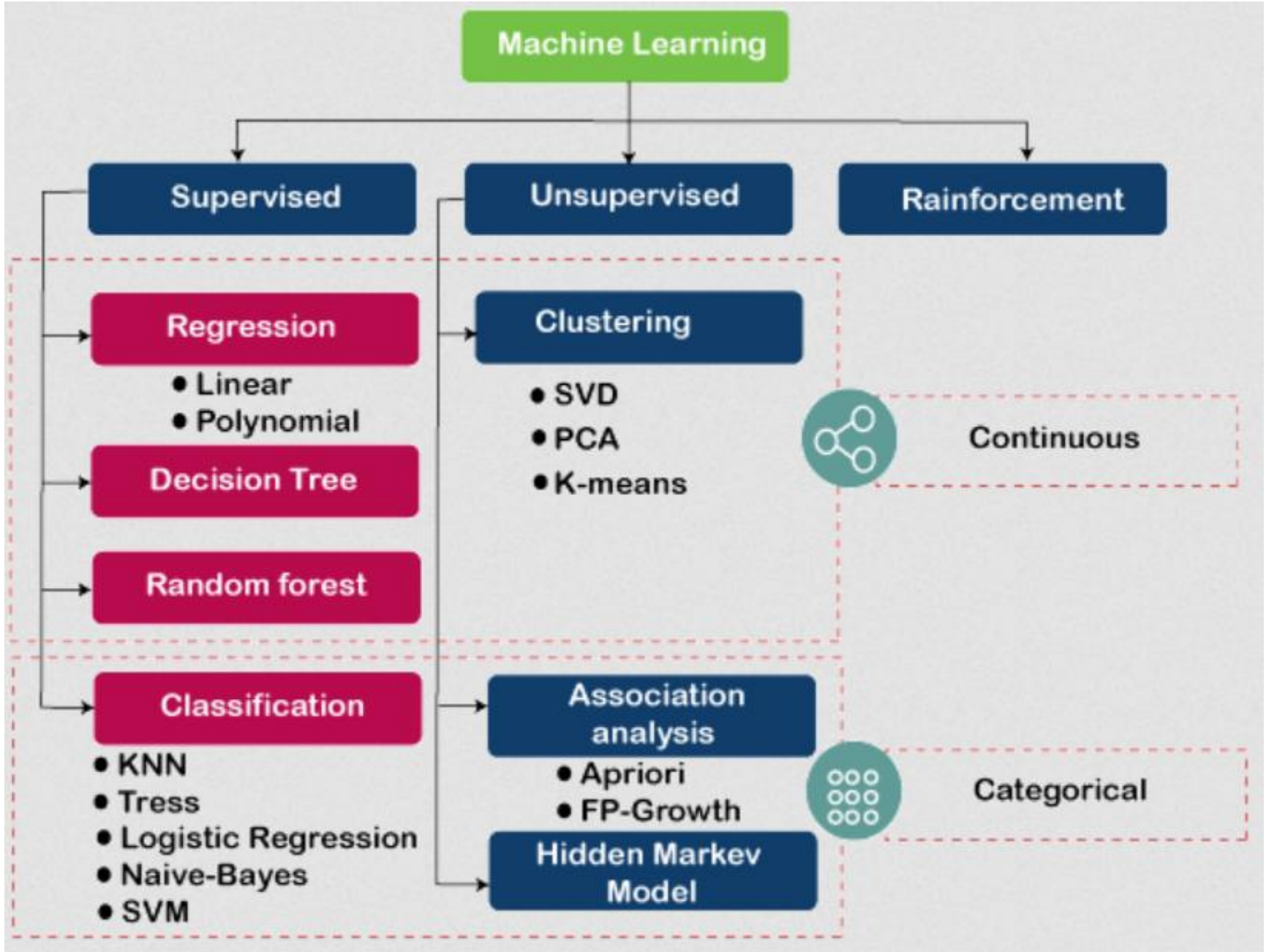
HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

EAL51501 – ARTIFICIAL INTELLIGENCE
B.Tech[AIML] – III Semester

K.Kowsalya
Assistant Professor (SS)
School of Computing Sciences,
Department of Computer Science and Engineering

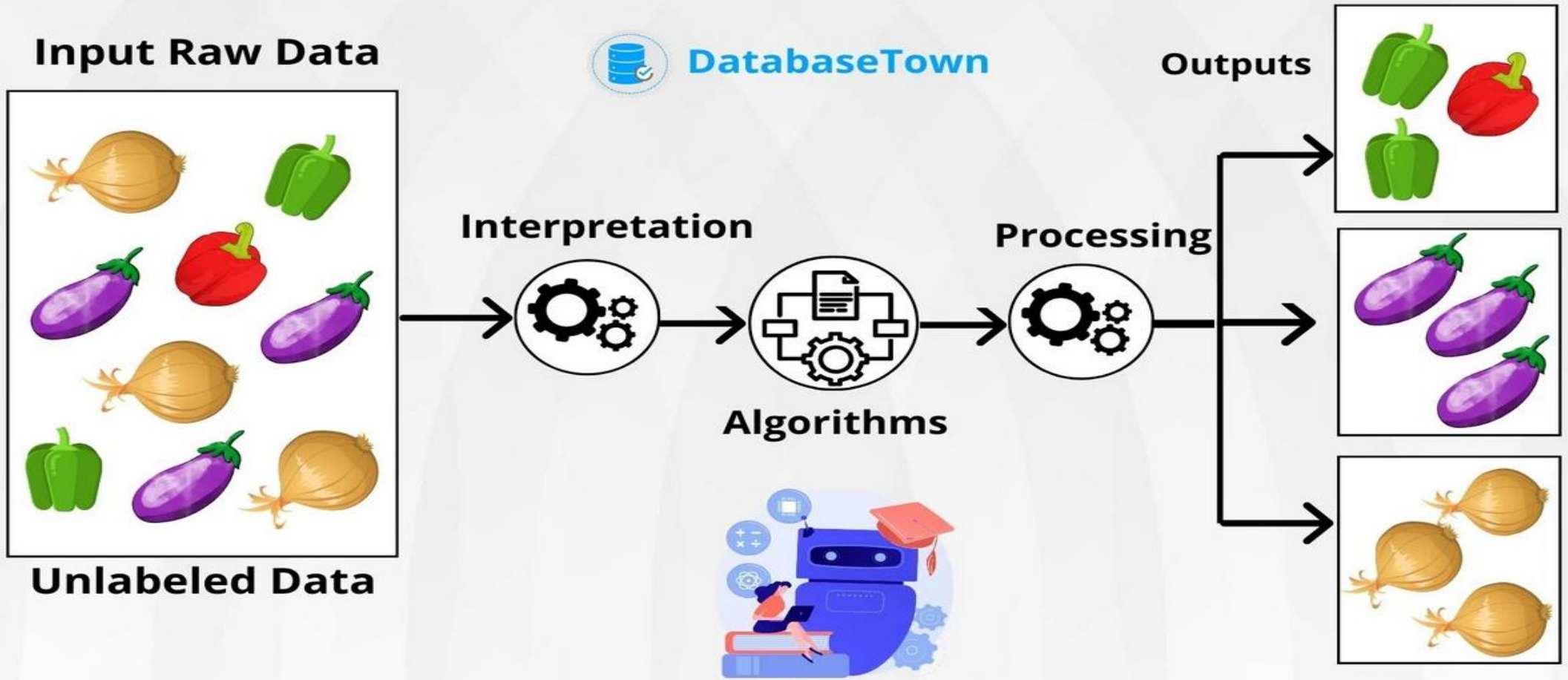
UNIT-III

- Motivation for Machine Learning, Applications, Machine Learning, Learning associations, Classification, Regression, The Origin of machine learning, Uses and abuses of machine learning, Success cases, How do machines learn[INTRO], Abstraction and knowledge representation, Generalization, Factors to be considered, Assessing the success of learning, Metrics for evaluation of classification method, Steps to apply machine learning to data, Machine learning process, Input data and ML algorithm, Classification of machine learning algorithms, General ML architecture, **Group of algorithms, Reinforcement learning, Supervised learning, Unsupervised learning**, Semi-Supervised learning, Algorithms, Ensemble learning, Matching data to an appropriate algorithm.


Machine Learning
Supervised
Unsupervised
Rainforcement
Regression
• Linear
• Polynomial
Decision Tree
Random forest
Clustering
• SVD
• PCA
• K-means
Continuous
Classification
• KNN
• Tress
• Logistic Regression
• Naive-Bayes
• SVM
Association analysis
• Apriori
• FP-Growth
Hidden Markev Model
Categorical

UNSUPERVISED LEARNING

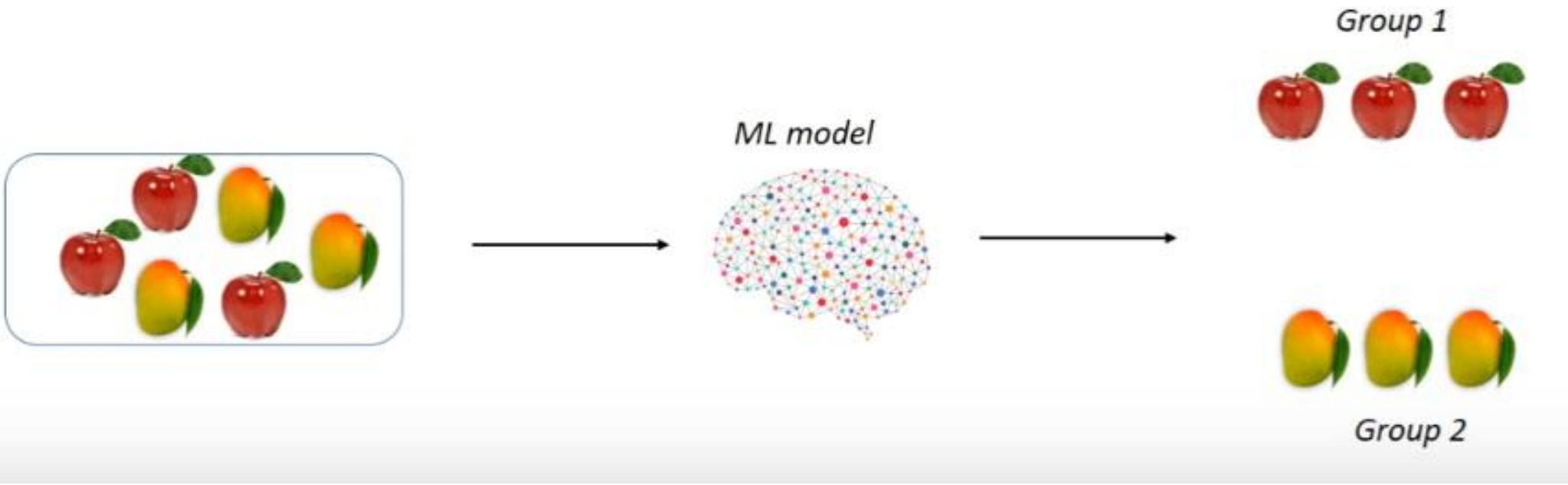
Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data without any predefined outputs or target variables.



- UNSUPERVISED LEARNING

Unsupervised Learning

In Unsupervised Learning, the Machine Learning algorithm learns from ***Unlabelled Data***



For
Example

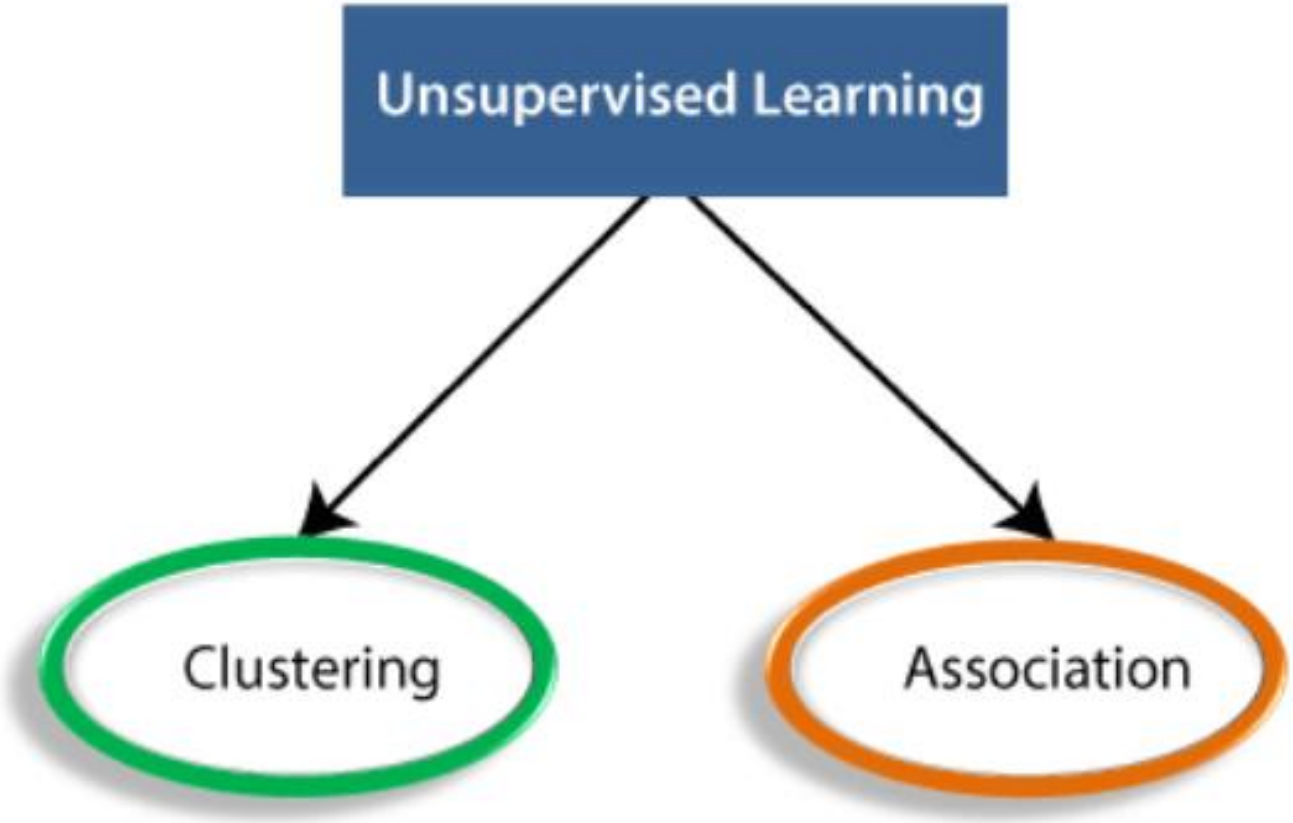
Suppose you and your friends want to watch the cricket match but you do not know what cricket is. But for your friends, you say yes. You reach home and start watching the match.

Unsupervised Learning - Example

You come to certain conclusions.

- There are 2 teams, India and Australia
- Different kinds of players such as batters, bowlers and so forth
- Ball hits wickets or is caught, batsman is out
- Cheer when India score a 4 or a 6

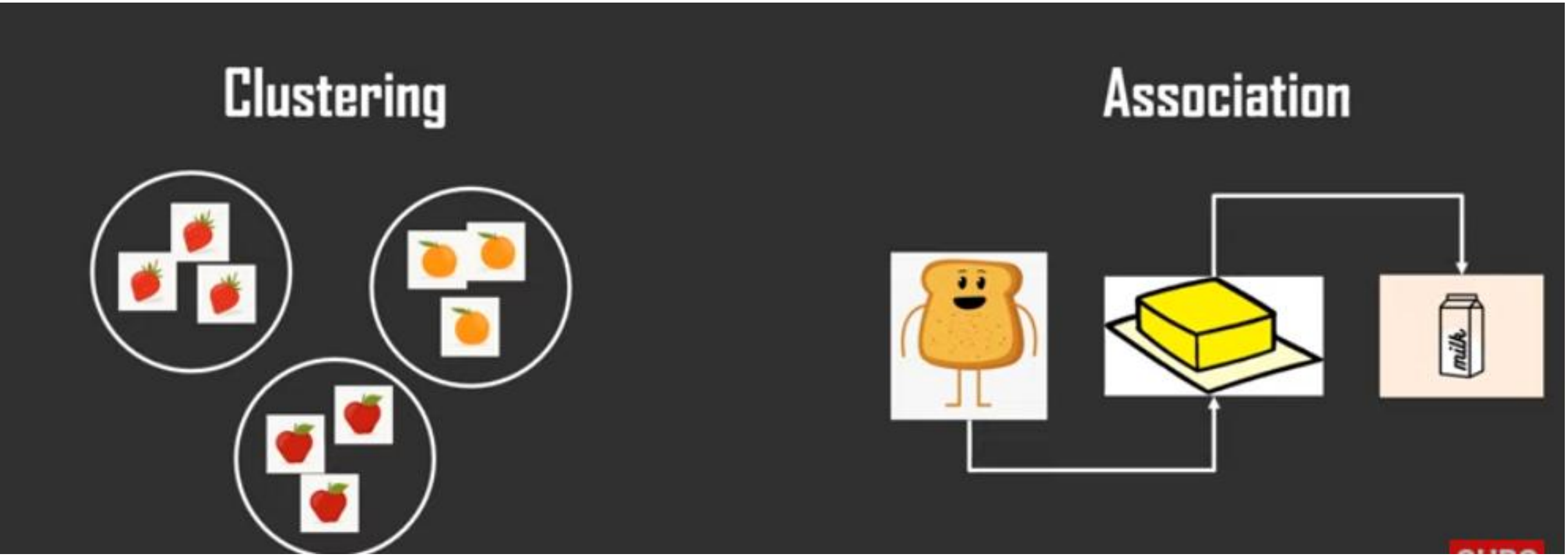
- **Types of Unsupervised Learning Algorithm:**
- The unsupervised learning algorithm can be further categorized into two types of problems:



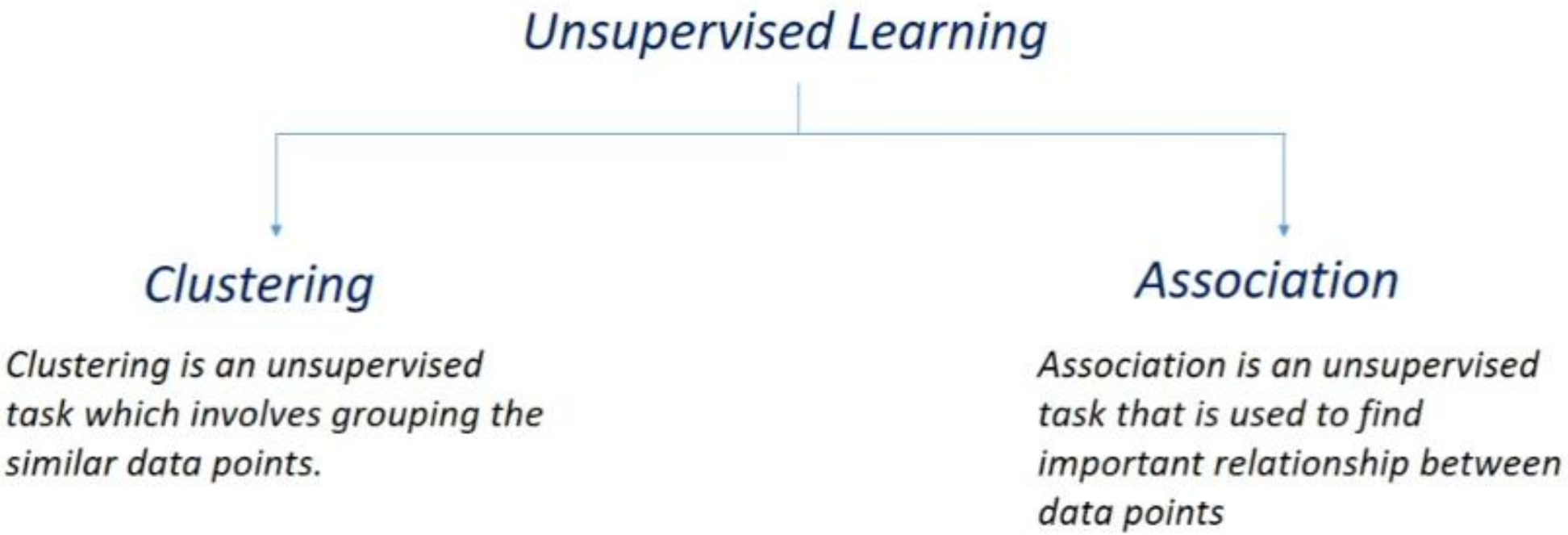
- **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.
- **Association**: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. **Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.**

Why is it Important?

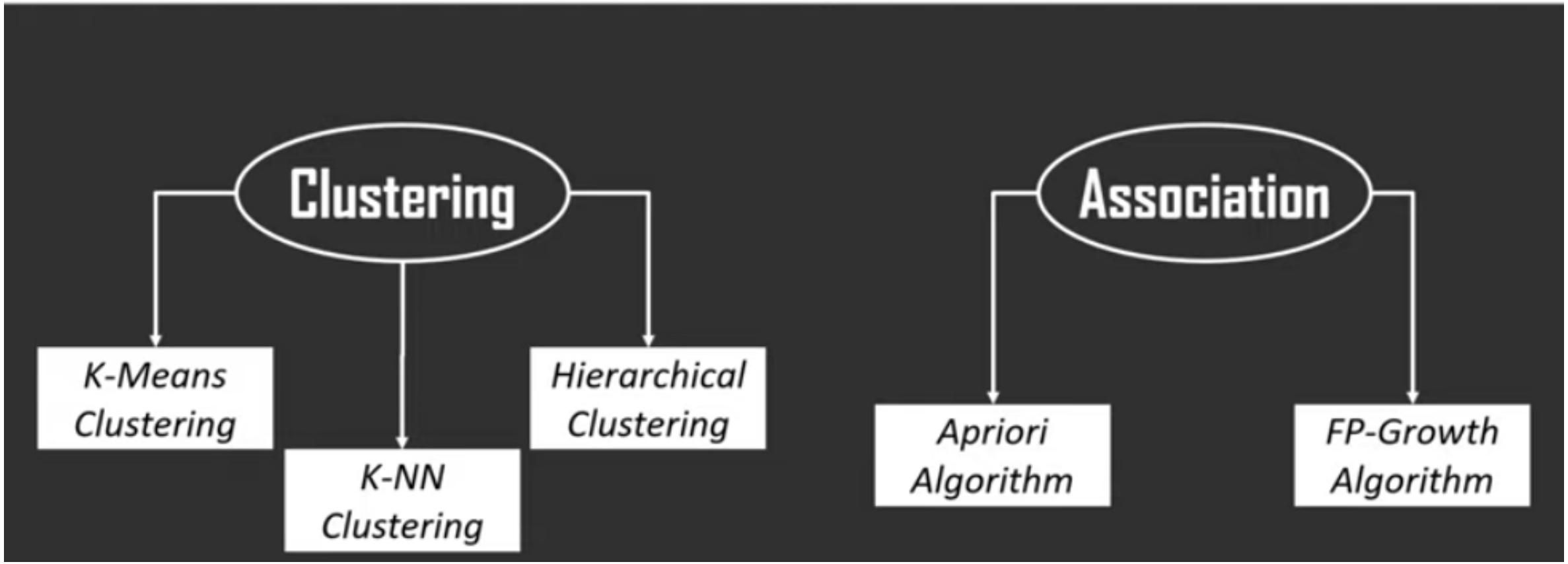
- They find patterns which were previously unknown
- Patterns help in categorization or finding association
- They can detect anomalies and defects in the data
- They work on unlabeled data which makes our work easier


Clustering
Association
milk

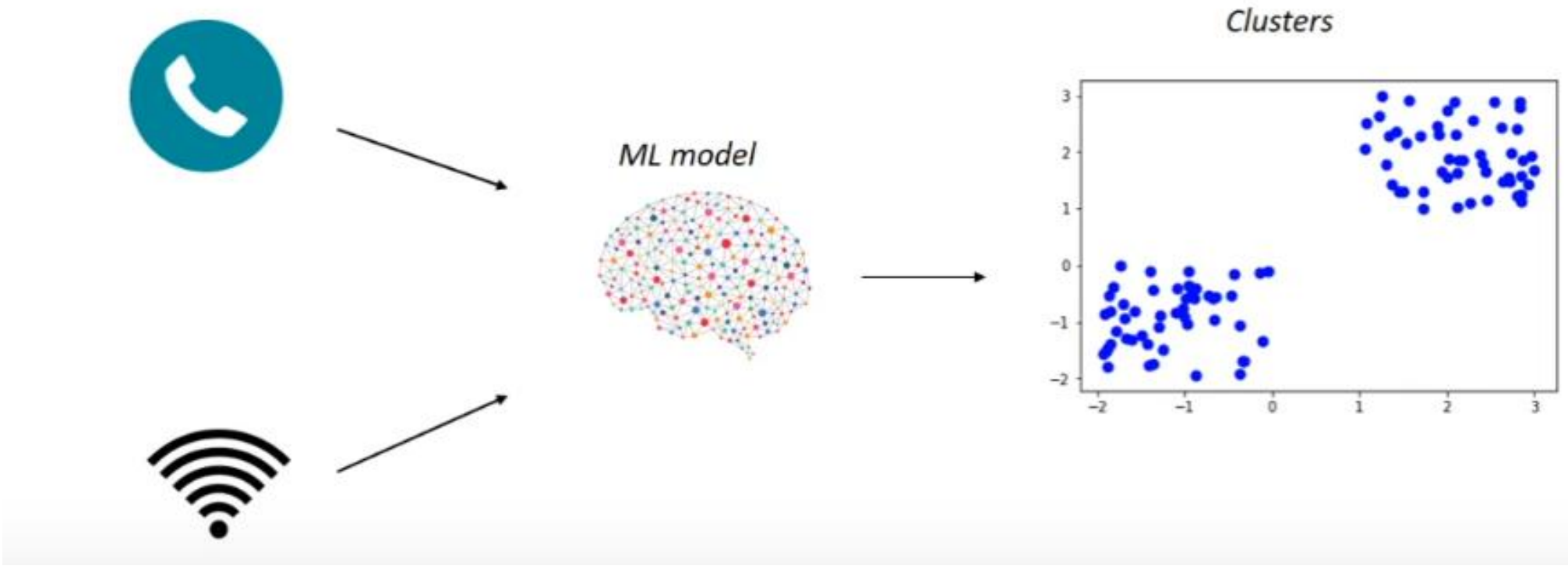
Types of Unsupervised Learning



Types of Unsupervised Learning



Clustering



Association

Customer 1

- Bread
- Milk
- Fruits
- wheat

Customer 2

- Bread
- Milk
- Rice
- Butter

Customer 3

Now, when customer 3 goes and buys bread, it is highly likely that he will also buy milk.

Advantages of Unsupervised Learning

- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning

- Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

Types of Unsupervised Learning

Clustering Algorithms

Clustering involves grouping similar data points together based on their inherent characteristics.

Clustering Algorithms

1.K-Means Clustering: In this algorithm, data is divided into **a specific number of groups or clusters**. It is achieved by minimizing the total squared distances between the data points and the **centers of each cluster.**

2.Hierarchical Clustering: It develops a hierarchy of clusters by **merging or splitting them depending on their similarity.**

3.DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN identifies clusters as **dense regions of data points separated by sparser regions.**

- **Dimensionality Reduction Algorithms**
- These are used to reduce the number of input variables or features while retaining meaningful information. Popular dimensionality reduction algorithms include:

-> Principal Component Analysis (PCA): PCA transforms the original features into a lower-dimensional space while preserving the maximum amount of information.

-> t-SNE (t-Distributed Stochastic Neighbor Embedding): t-SNE is a technique that **visualizes high-dimensional data by reducing it to a lower-dimensional space while preserving local structure**

- **Association Rule Mining**
- It focuses on discovering **interesting relationships or patterns in transactional data.** It is commonly used in **market basket analysis** and **recommendation systems**. The widely used algorithm for association rule mining is the Apriori algorithm.
- A real-life example of this is market basket analysis, where retailers analyze customer purchase data to identify relationships between products frequently bought together. **For instance, this analysis might reveal that customers who purchase bread also tend to buy jam.**

Applications of Unsupervised Learning

- Unsupervised learning finds applications across various domains. Some notable applications include:
- **Customer Segmentation:** Unsupervised learning algorithms can group customers based on their **purchasing behavior**, allowing businesses to tailor marketing strategies.
- **Anomaly Detection:** By identifying abnormal patterns or outliers, unsupervised learning can help **detect fraud, network intrusions, or manufacturing defects.**

Applications of Unsupervised Learning

- **Image and Text Clustering:** Unsupervised learning can **automatically group similar images or texts,** aiding in tasks like image organization, document clustering, or content recommendation.
- **Genome Analysis:** Unsupervised learning algorithms can analyze **genetic data to identify patterns and relationships,** leading to insights in personalized medicine and genetic research.
- **Social Network Analysis:** Unsupervised learning can be used to identify communities or influential individuals within social networks, **enabling targeted marketing or detecting online communities.**

- Clustering means grouping of objects based on the information found in the data describing the objects or their relationship.
- The goal is that objects in one group should be similar to each other but different from objects in another group.
- It deals with finding a structure in a collection of unlabeled data.

Some Examples of clustering methods are:

- K-means Clustering
- Fuzzy/ C-means Clustering
- Hierarchical Clustering

Clustering Use Cases

Marketing
Discovering distinct groups in customer databases, such as customers who make lot of long-distance calls.

Insurance
Identifying groups of crop insurance policy holders with a high average claim rate. Farmers crash crops when it is "profitable".

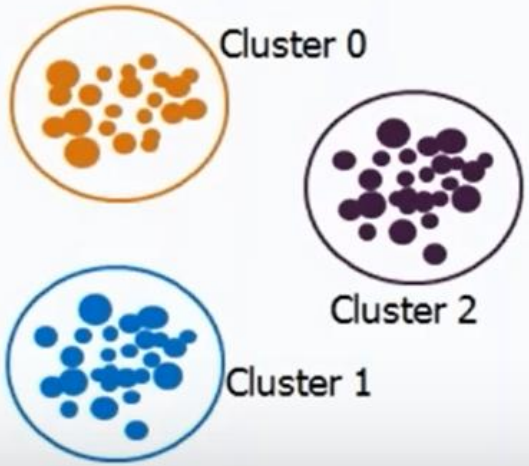
Land use
Identification of areas of similar land use in a GIS database.

Seismic studies
Identifying probable areas for oil/gas exploration based on seismic data

Types of Clustering

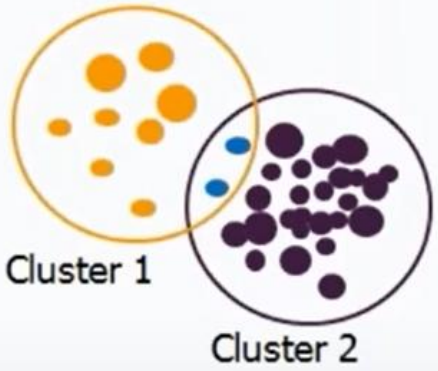
Exclusive Clustering

- An item belongs exclusively to one cluster, not several.
- K-means does this sort of exclusive clustering.



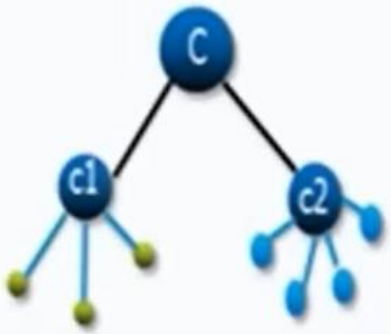
Overlapping Clustering

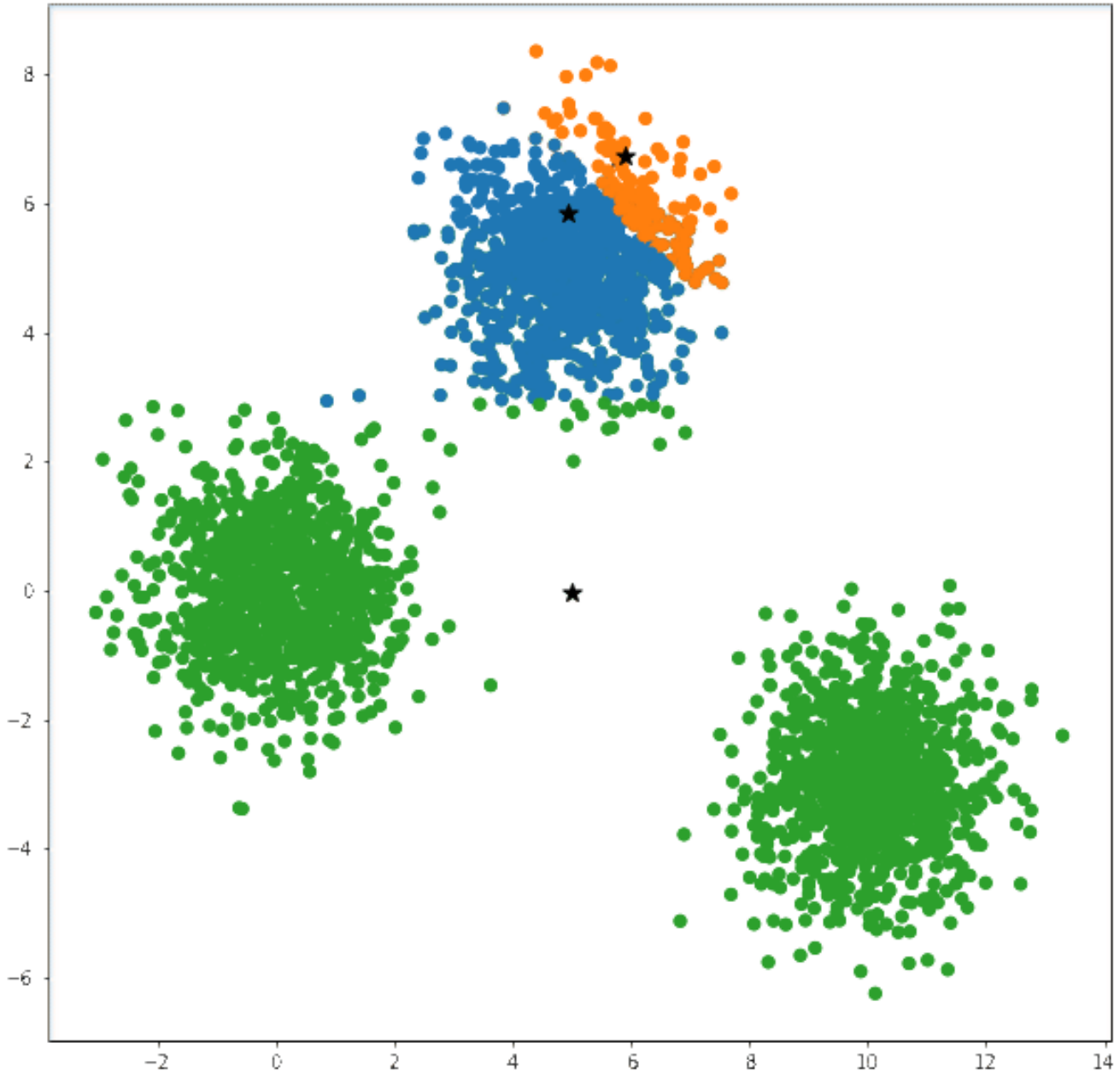
- An item can belong to multiple clusters
- Its degree of association with each cluster is known
- Fuzzy/ C-means does this sort of exclusive clustering.



Hierarchial Clustering

- When two cluster have a parent-child relationship or a tree-like structure then it is Hierarchical clustering




8
6
4
2
0
−2
−4
−6
−2
0
2
4
6
8
10
12
14

K-Means Clustering

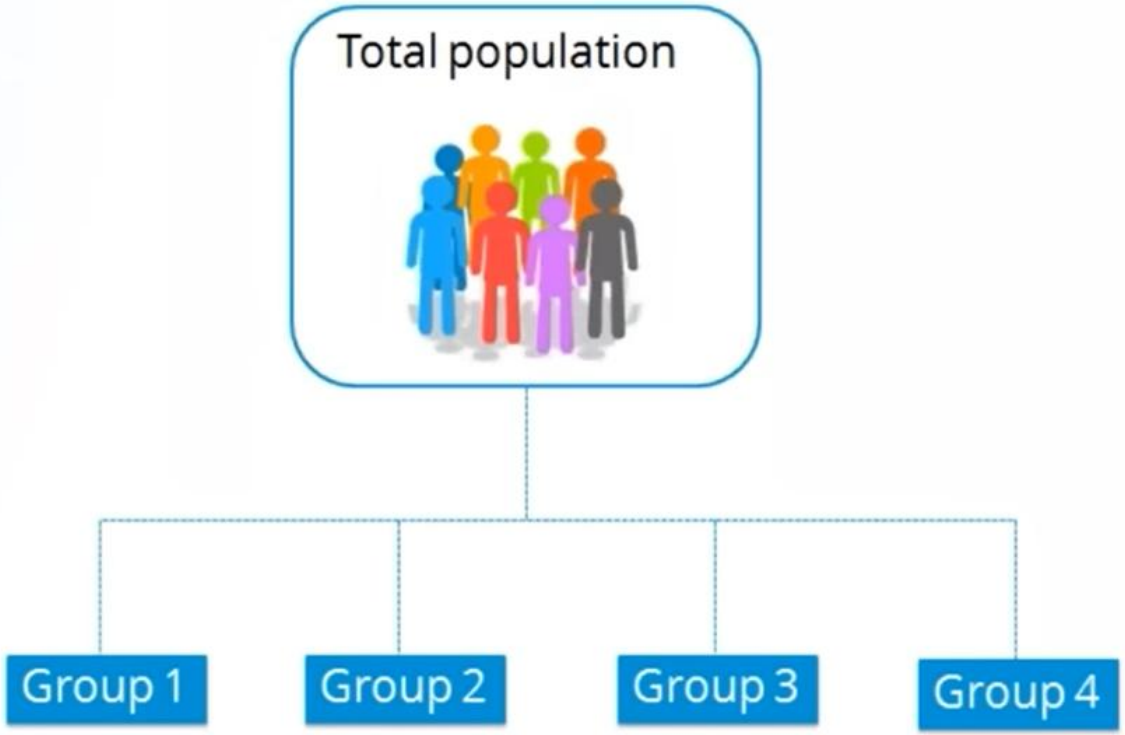
k-means clustering

k-means clustering is one of the simplest algorithms which uses unsupervised learning method to solve known clustering issues.

Divides entire dataset into k clusters.

k-means clustering require following two inputs.

1. K = number of clusters
2. Training set(m) = {x1, x2, x3,....., xm}



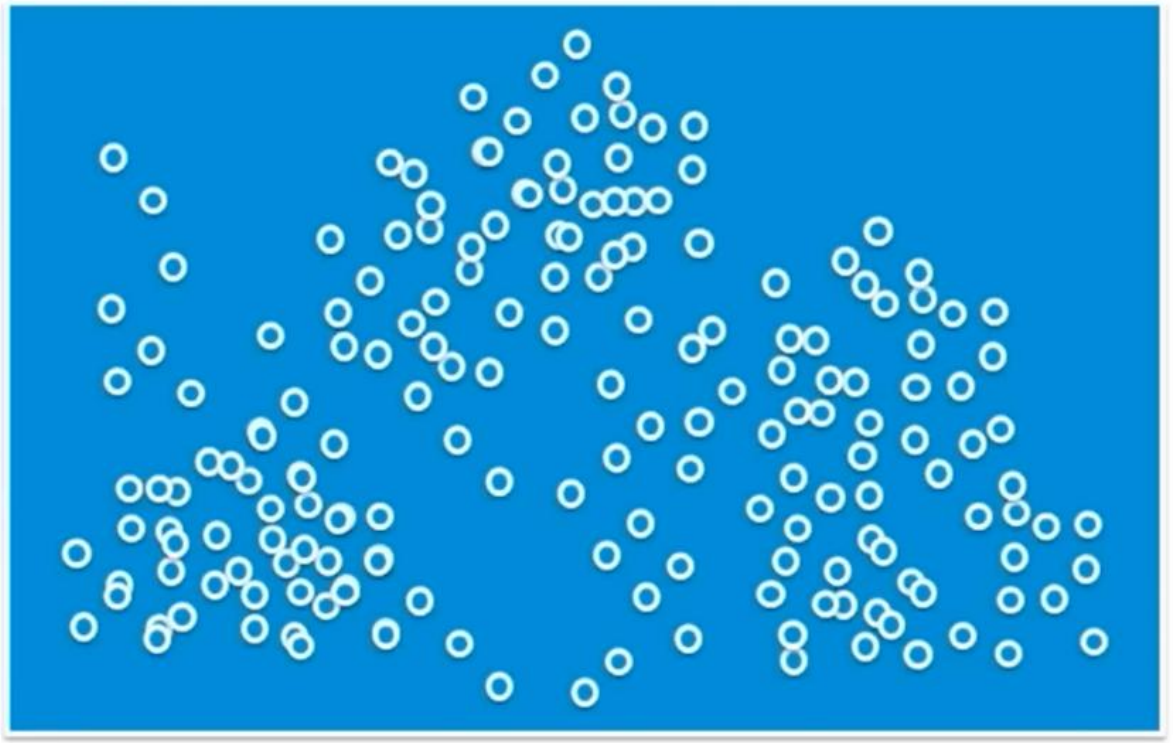
Example-Google News

- Various news URLs related to Trump and Modi are grouped under one section.
- K-means clustering automatically clusters new stories about the same topic into pre-defined clusters.

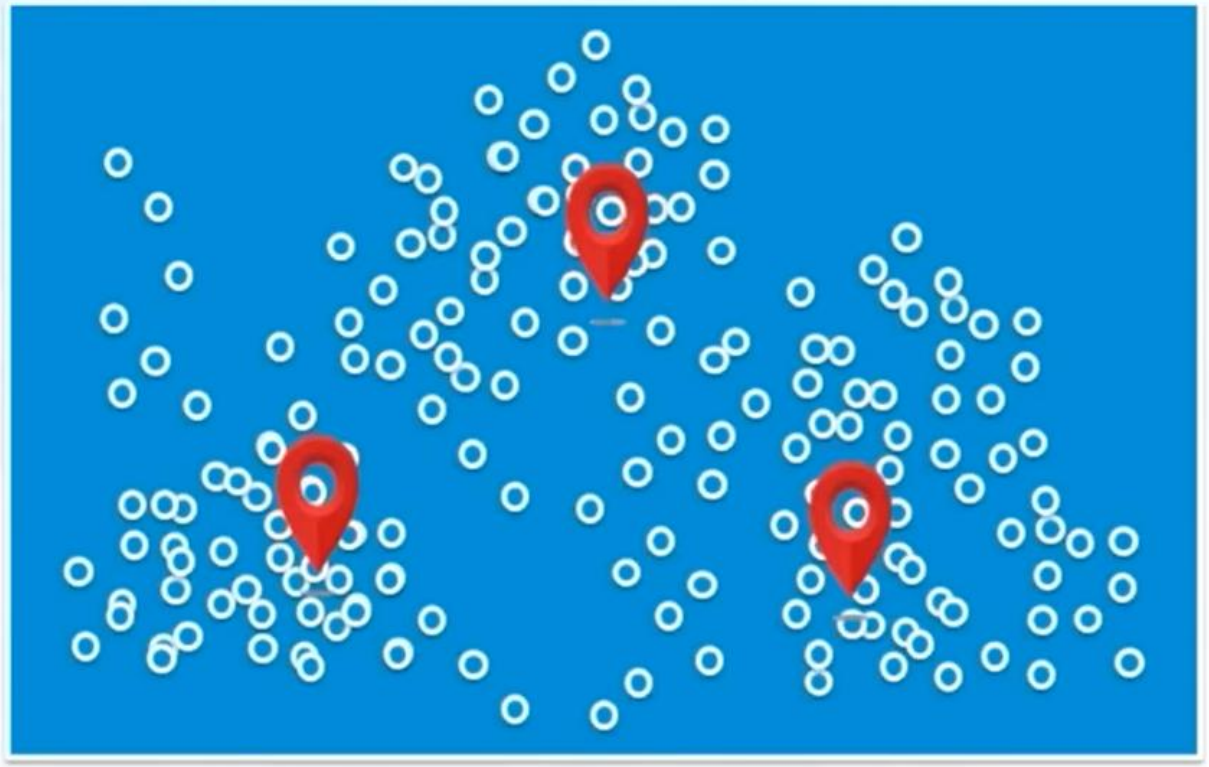
Example:

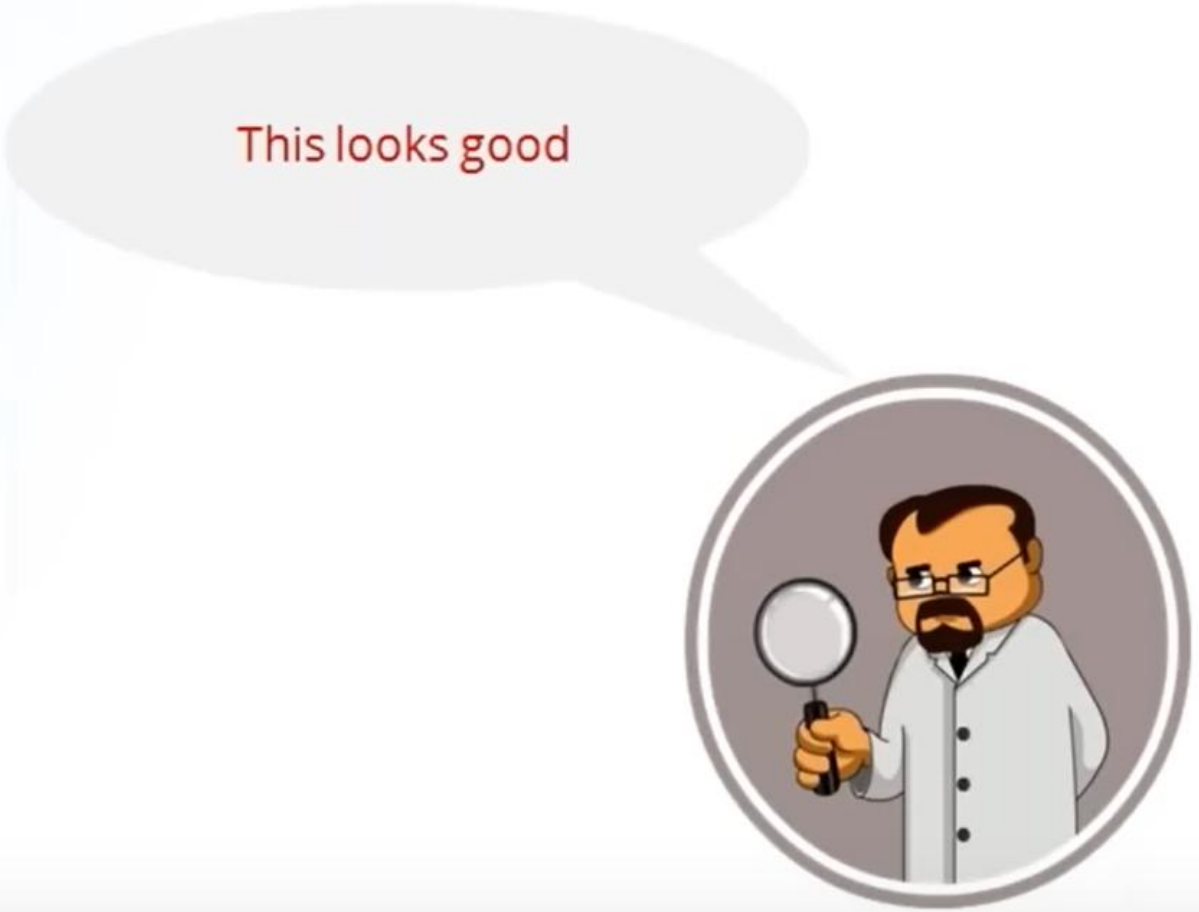
The plot of students in an area is as given below,

I need to find specific locations to build schools in this area so that the students doesn't have to travel much



Example-Solution





- **K-Means Clustering** is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.
- **What is K-Means Algorithm?**
- K-Means Clustering is an Unsupervised Learning algorithm, **which groups the unlabeled dataset into different clusters.**
- Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

- It allows us to cluster the data into different groups and a convenient way to discover the **categories of groups in the unlabeled dataset** on its own **without the need for any training**.
- It is a **centroid-based algorithm**, where each cluster is associated with a centroid. The main aim of this algorithm is to **minimize the sum of distances between the data point and their corresponding clusters.**
- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

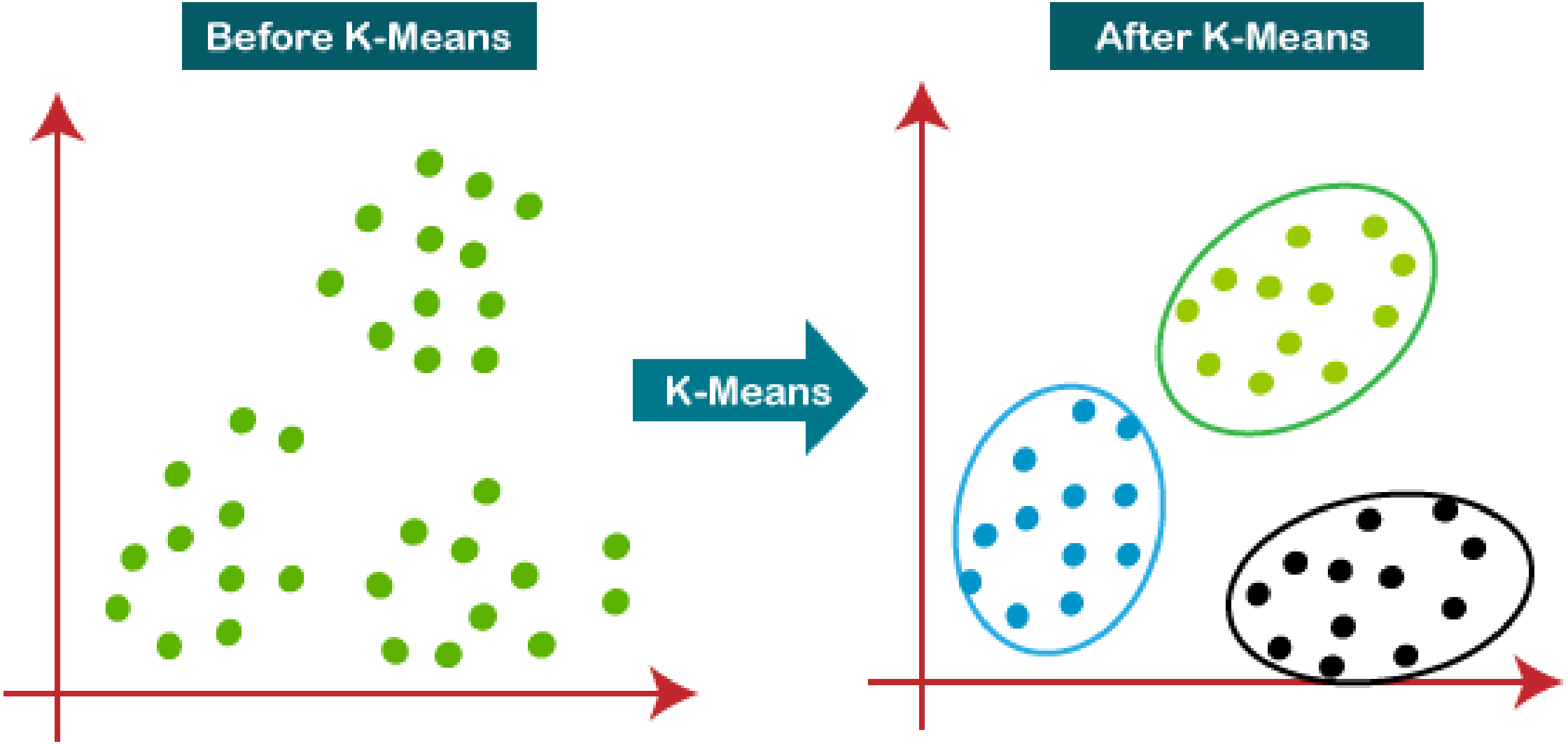
The k-means clustering algorithm mainly performs **two tasks:**

->Determines the best value for K center points or centroids by an iterative process.

->Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

->Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

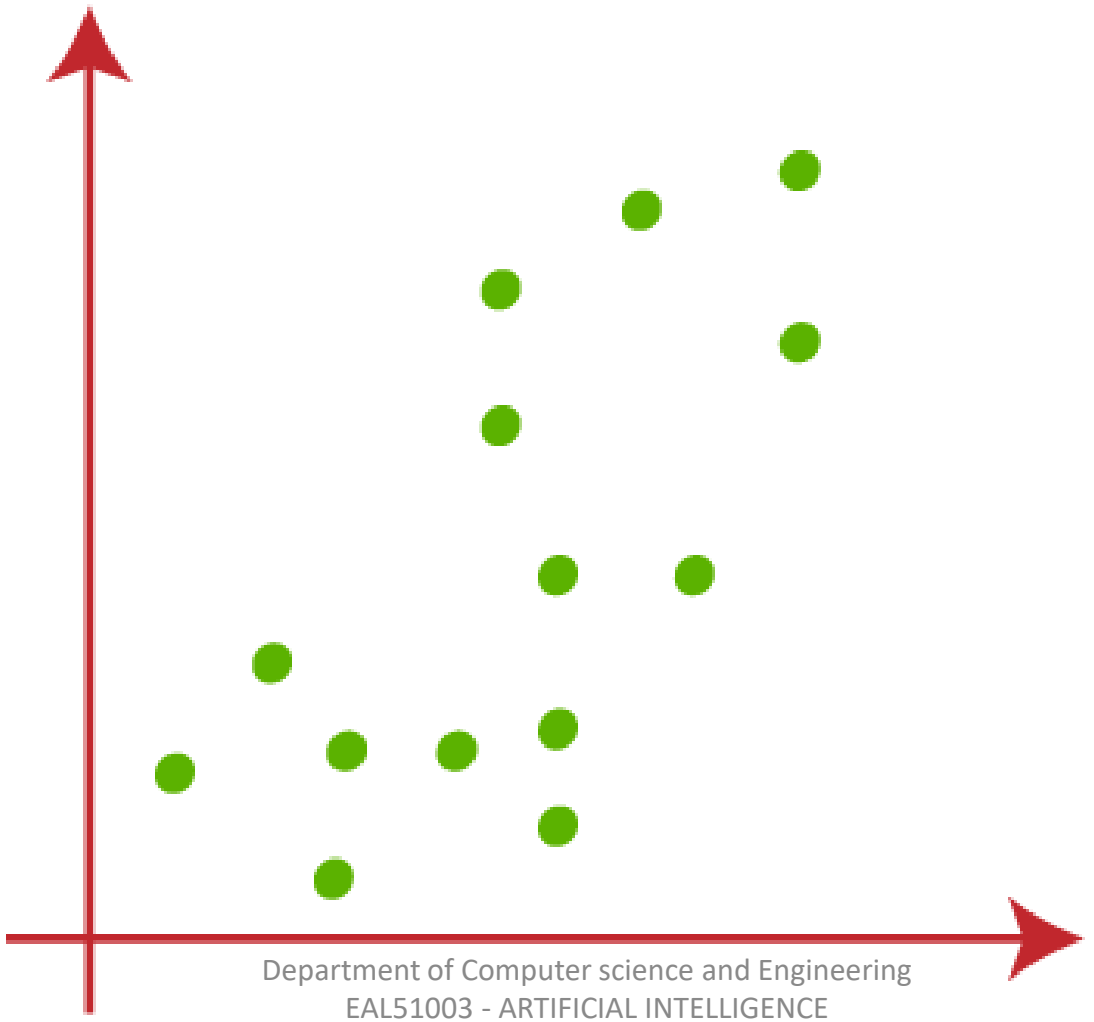
- The below diagram explains the working of the K-means Clustering Algorithm:


Before K-Means
K-Means
After K-Means

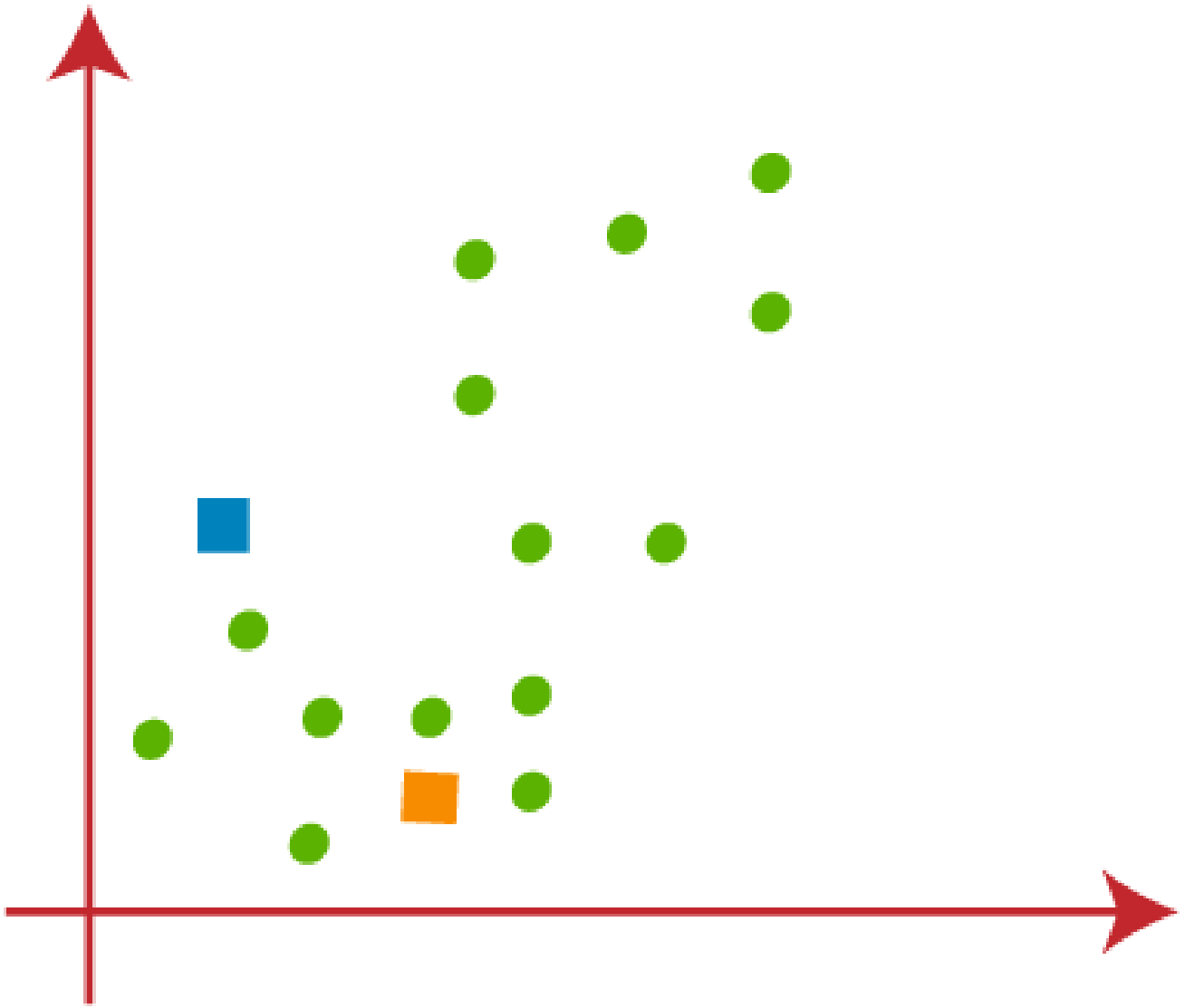
How does the K-Means Algorithm Work?

- The working of the K-Means algorithm is explained in the below steps:
- **Step-1:** Select the **number K to decide the number of clusters**.
- **Step-2:** Select **random K points or centroids**. (It can be other from the input dataset).
- **Step-3:** Assign **each data point to their closest centroid**, which will form the predefined K clusters.
- **Step-4: Calculate the variance** and place a new centroid of each cluster.
- **Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.
- **Step-7**: The model is ready.

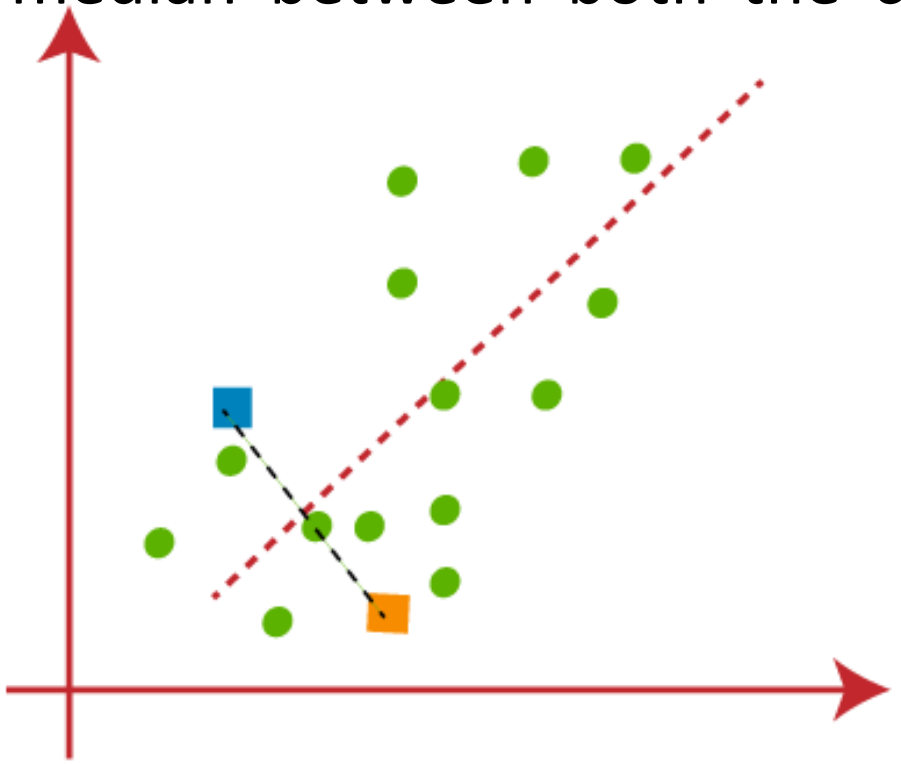
- Let's understand the above steps by considering the visual plots:
- Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables is given below:



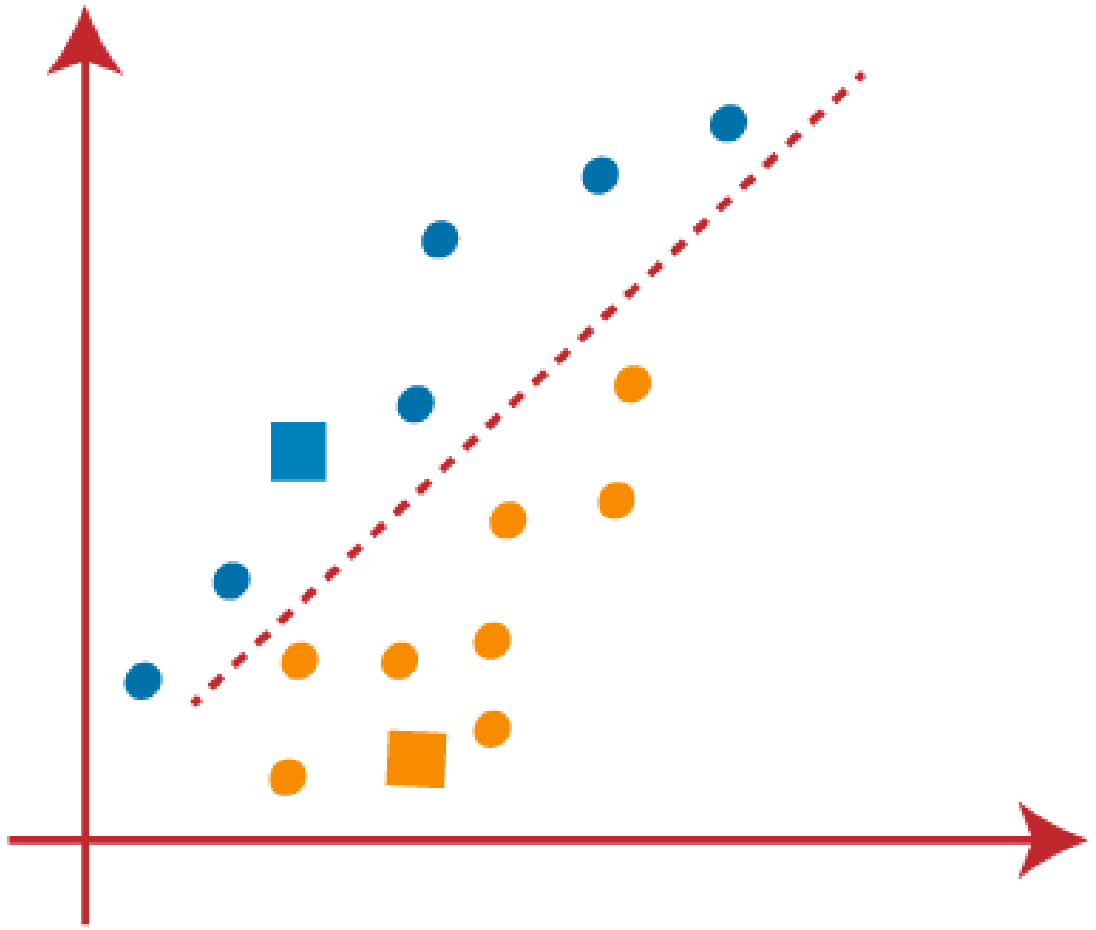
- Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters. It means here we will try to group these datasets into two different clusters.
- We need to choose some random k points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as k points, which are not the part of our dataset.
- Consider the below image:



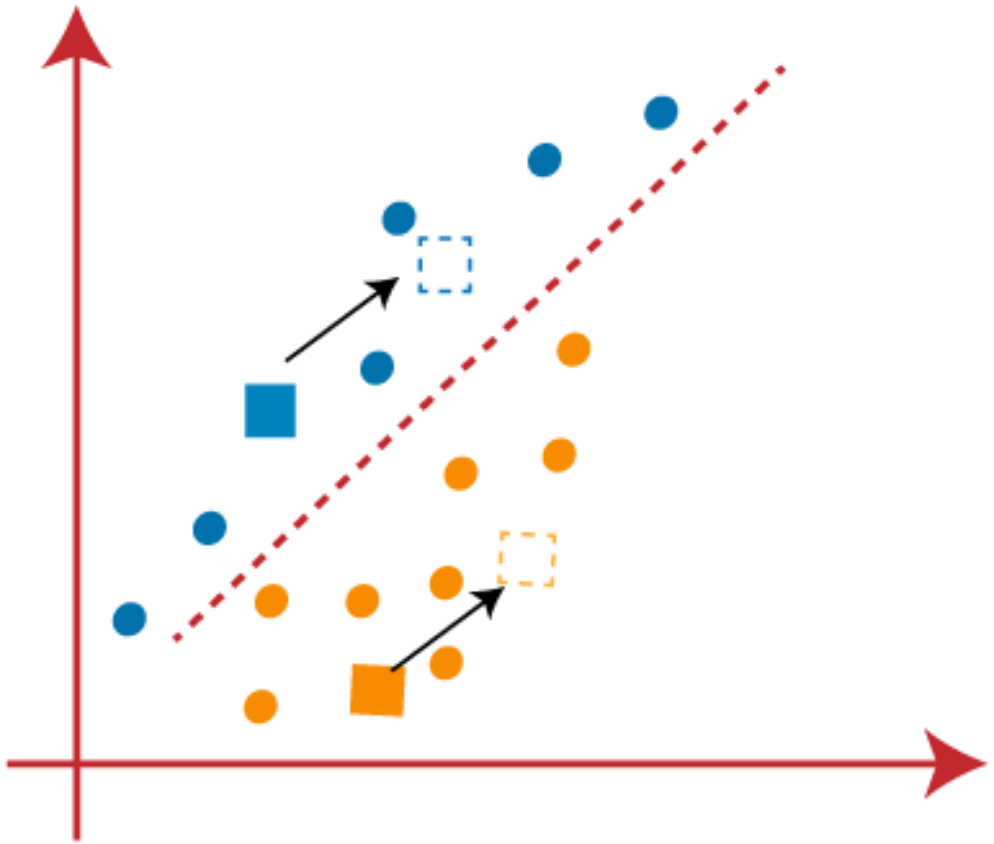
- Now we will assign each data point of the scatter plot to its closest K-point or centroid.
- We will compute it by applying some mathematics that we have studied to calculate the distance between two points.
- So, we will draw a median between both the centroids. Consider the below image:



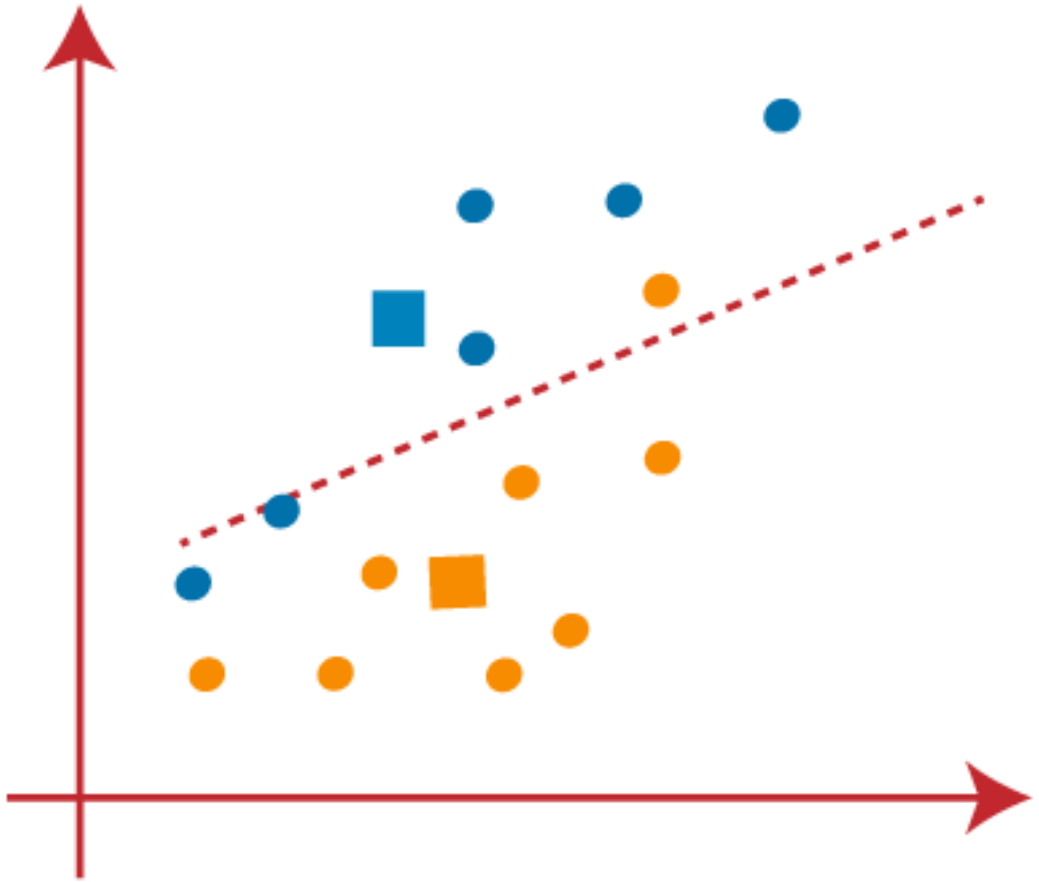
- From the above image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid. Let's color them as blue and yellow for clear visualization.



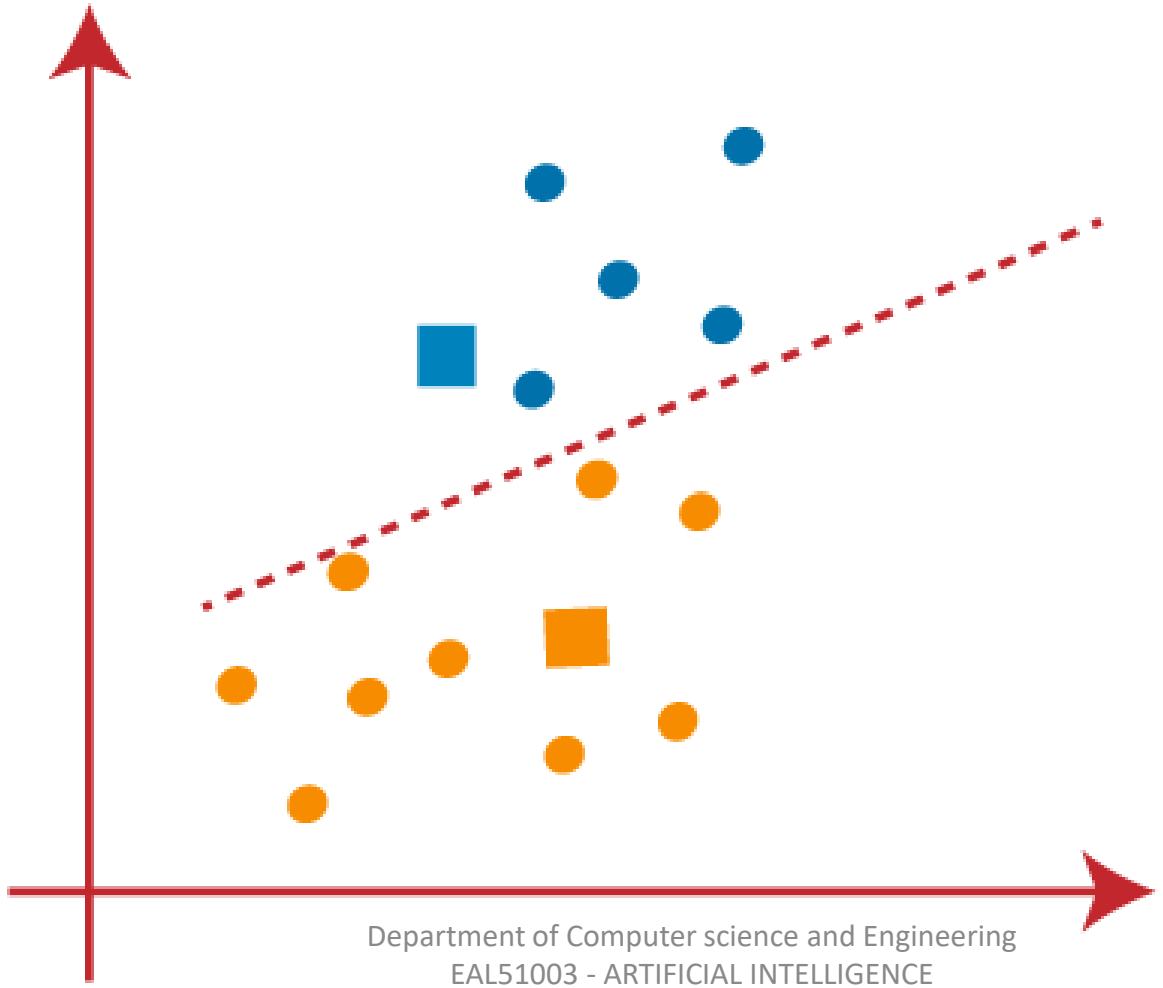
- As we need to find the closest cluster, so we will repeat the process by choosing **a new centroid**. To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as below:



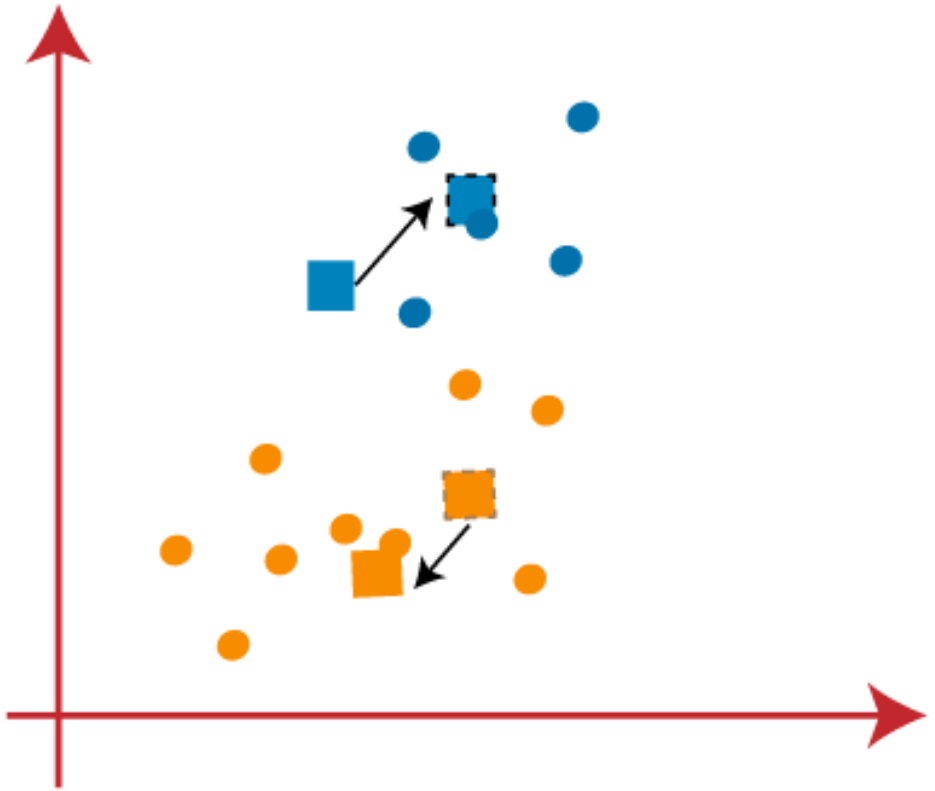
- Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line. The median will be like below image:



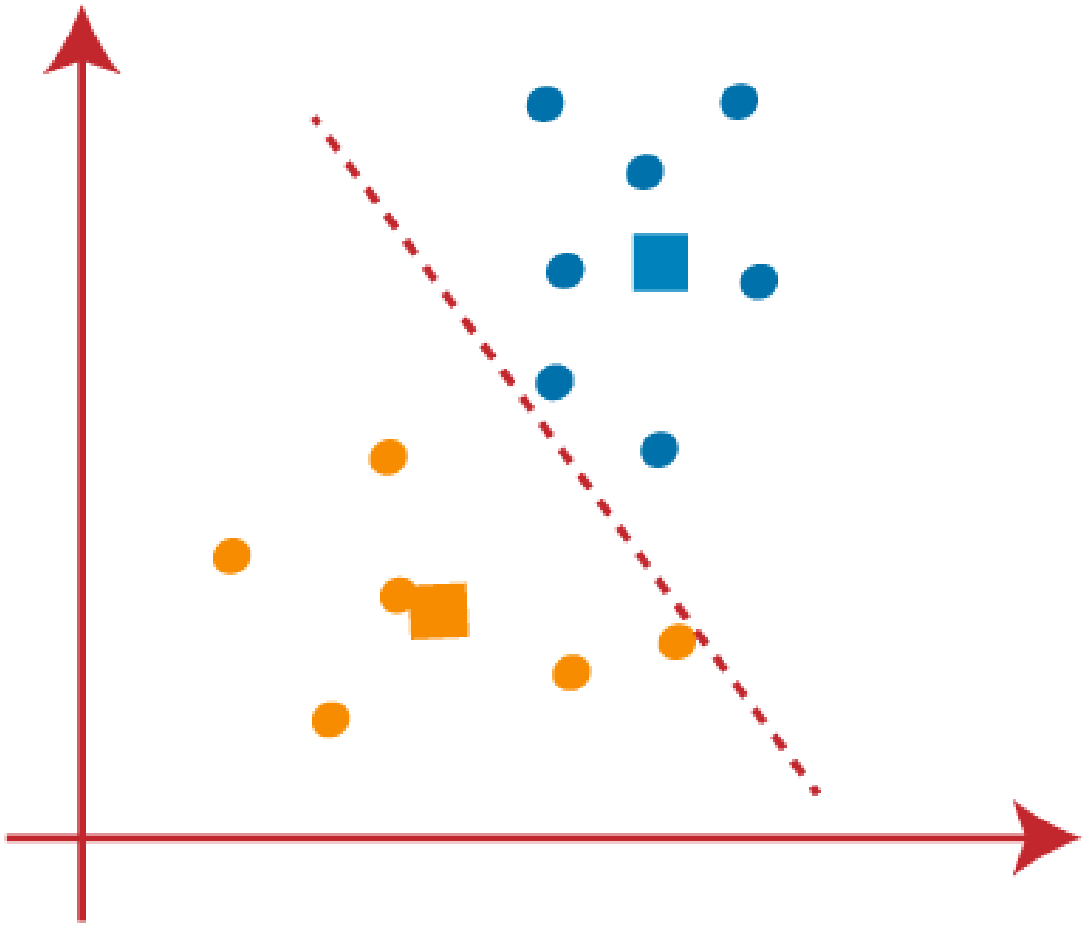
- From the above image, we can see, one yellow point is on the left side of the line, and two blue points are right to the line. So, these three points will be assigned to new centroids.



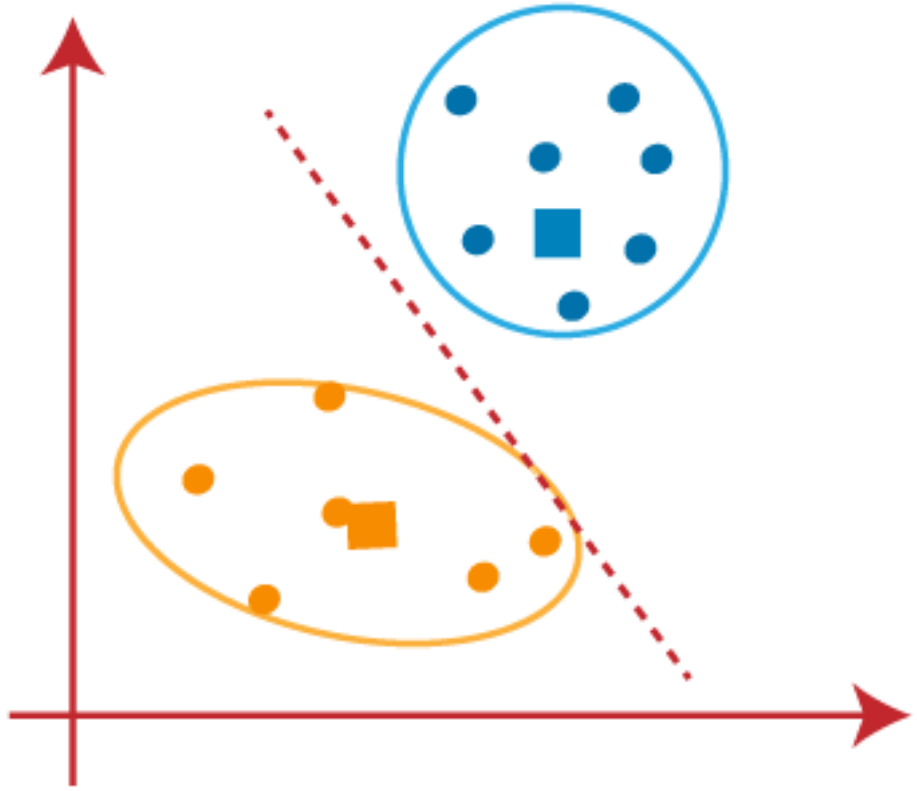
- Reassignment has taken place, so we will again go to the step-4, which is finding new centroids or K-points.
- We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below image:



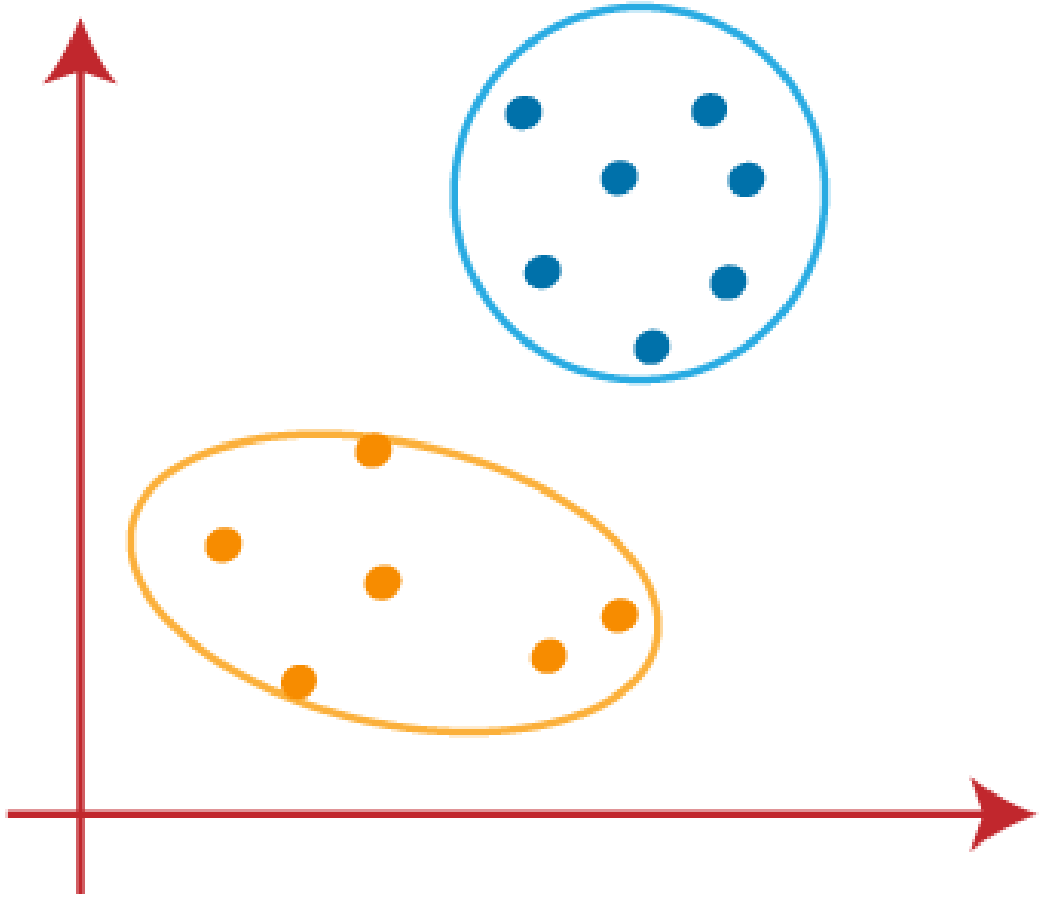
- As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:



- We can see in the above image; there are no dissimilar data points on either side of the line, which means our model is formed. Consider the below image:



- As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be as shown in the below image:



- **Elbow Method**
- The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster.
- The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$WCSS = \sum_{P_i \text{ in Cluster1}} distance(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} distance(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} distance(P_i\ C_3)^2$$

In the above formula of WCSS,

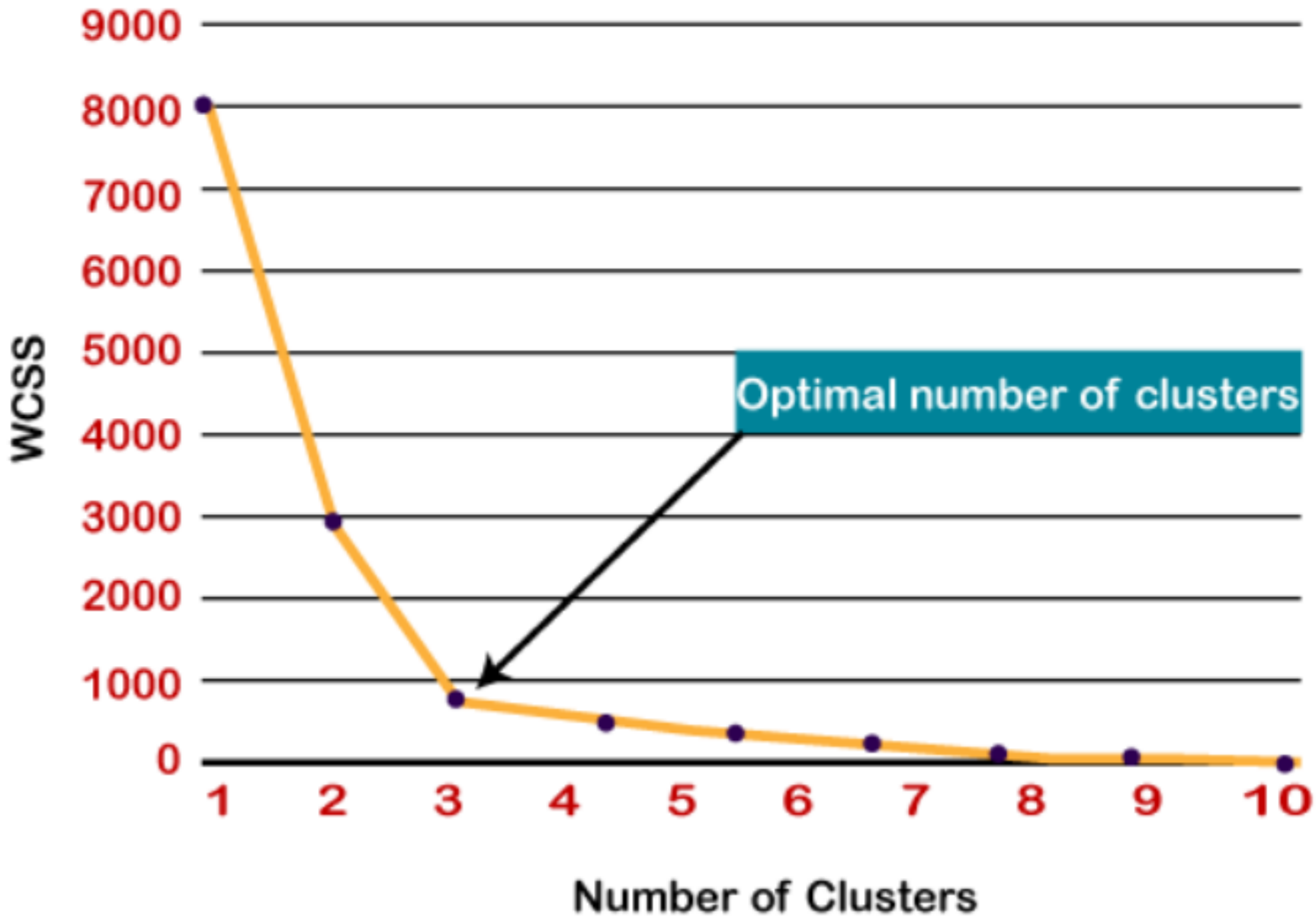
$\sum P_{i \text{ in Cluster1}}$ distance$(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method. The graph for the elbow method looks like the below image:


9000
8000
7000
6000
5000
4000
3000
2000
1000
0
WCSS
Optimal number of clusters
1
2
3
4
5
6
7
8
9
10
Number of Clusters

HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

EAL51501 – ARTIFICIAL INTELLIGENCE
B.Tech[AIML] – III Semester

K.Kowsalya
Assistant Professor (SS)
School of Computing Sciences,
Department of Computer Science and Engineering

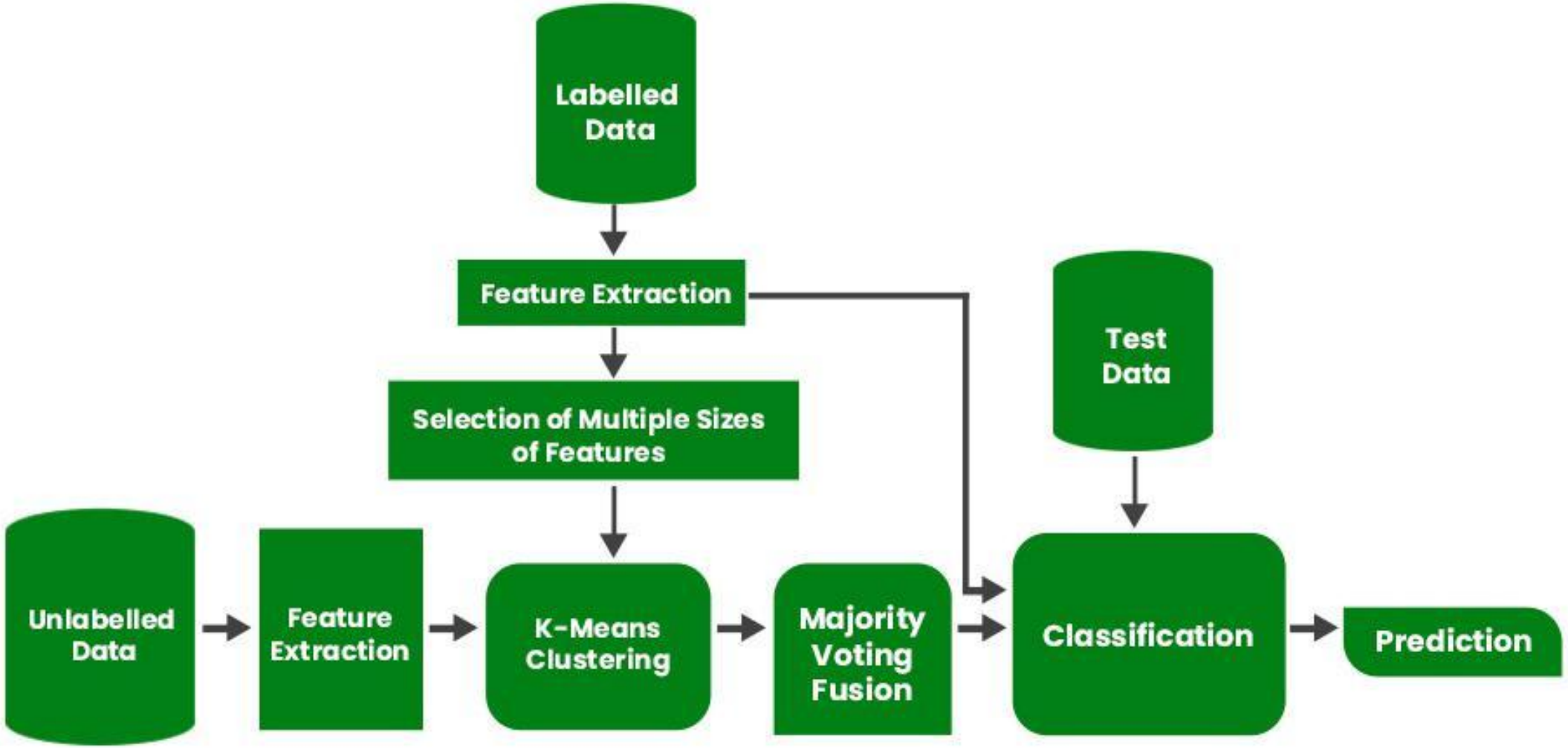
UNIT-III

- Motivation for Machine Learning, Applications, Machine Learning, Learning associations, Classification, Regression, The Origin of machine learning, Uses and abuses of machine learning, Success cases, How do machines learn[INTRO], Abstraction and knowledge representation, Generalization, Factors to be considered, Assessing the success of learning, Metrics for evaluation of classification method, Steps to apply machine learning to data, Machine learning process, Input data and ML algorithm, Classification of machine learning algorithms, General ML architecture, Group of algorithms, Reinforcement learning, Supervised learning, Unsupervised learning, **Semi-Supervised learning, Algorithms, Ensemble learning, Matching data to an appropriate algorithm.**

- ***Semi-Supervised learning is a type of Machine Learning algorithm that represents the intermediate ground between Supervised and Unsupervised learning algorithms.***
- ***It uses the combination of labeled and unlabeled datasets during the training period.***

- It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model.
- The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning.
- However, unlike supervised learning, the algorithm is trained on a dataset that contains both labeled and unlabeled data.

- Semi-supervised learning is particularly useful when there is a **large amount of unlabeled data available, but it's too expensive or difficult to label all of it.**

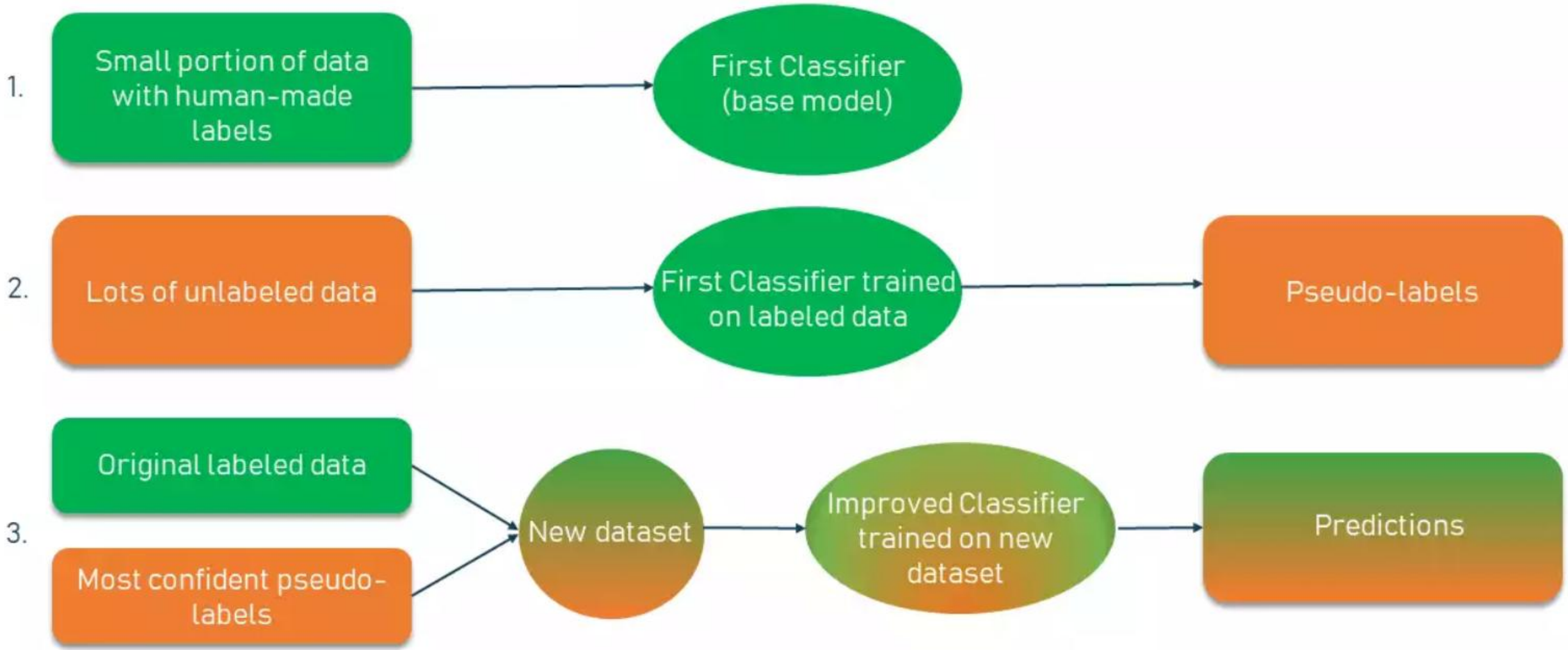

Labelled Data
Feature Extraction
Selection of Multiple Sizes of Features
Test Data
Unlabelled Data
Feature Extraction
K-Means Clustering
Majority Voting Fusion
Classification
Prediction

- Intuitively, one may imagine the three types of learning algorithms as Supervised learning where a student is under the supervision of a teacher at both home and school, Unsupervised learning where a student has to figure out a concept himself and Semi-Supervised learning where a teacher teaches a few concepts in class and gives questions as homework which are based on similar concepts.

- **Working of Semi-Supervised Learning**
- Semi-supervised learning uses pseudo labeling to train the model with less labeled training data than supervised learning. The process can combine various neural network models and training ways. The whole working of semi-supervised learning is explained in the below points:
- Firstly, it trains the model with less amount of training data similar to the supervised learning models. The training continues until the model gives accurate results.
- The algorithms use the unlabeled dataset with pseudo labels in the next step, and now the result may not be accurate.
- Now, the labels from labeled training data and pseudo labels data are linked together.
- The input data in labeled training data and unlabeled training data are also linked.
- In the end, again train the model with the new combined input as did in the first step. It will reduce errors and improve the accuracy of the model.

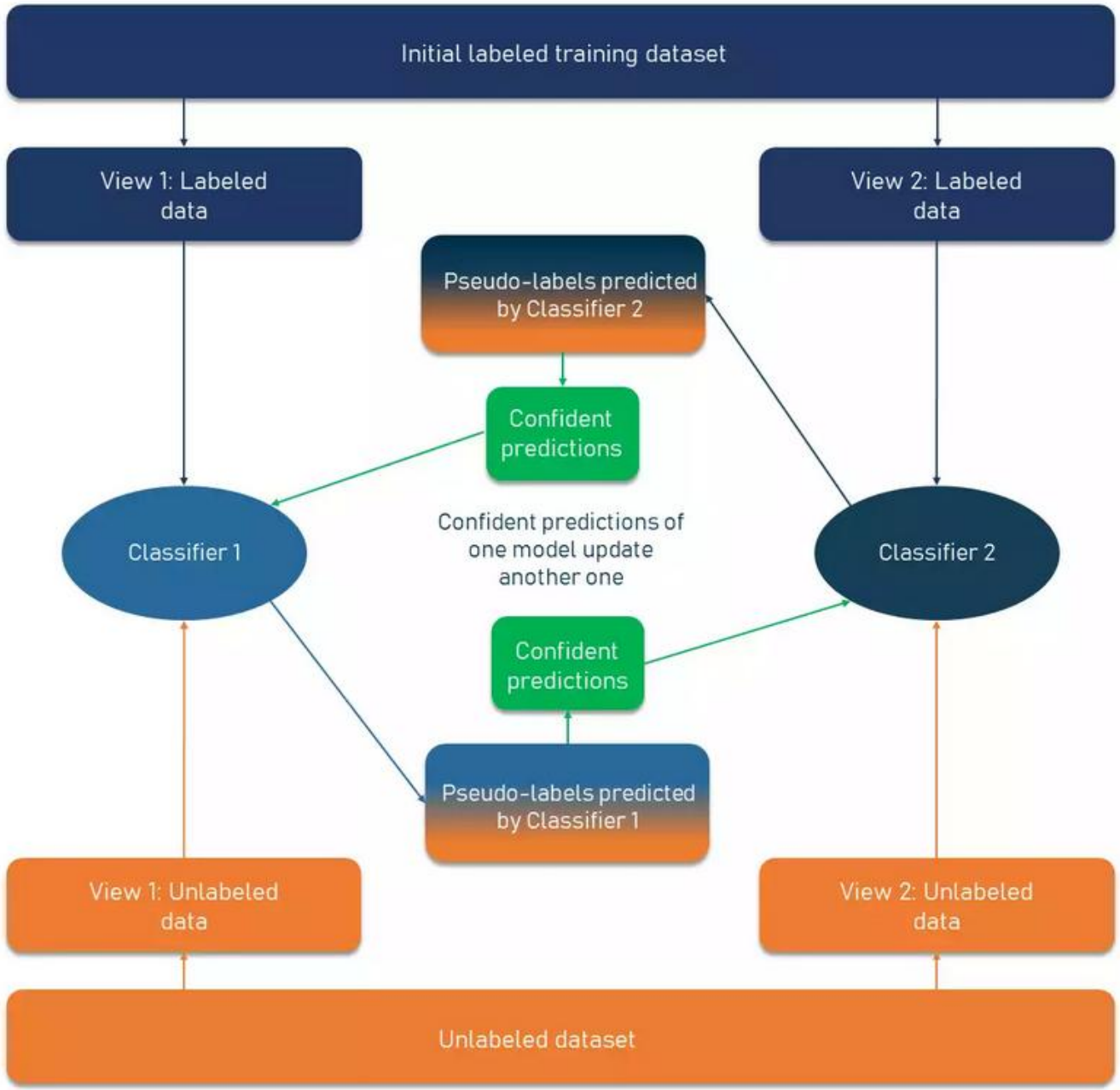
- **Self-training**
- One of the simplest examples of semi-supervised learning, in general, is self-training.
- **Self-training** is the procedure in which you can take any supervised method for classification or regression and modify it to work in a semi-supervised manner, taking advantage of labeled and unlabeled data. The standard workflow is as follows.

SEMI-SUPERVISED SELF-TRAINING METHOD



- **Co-training**
- Derived from the self-training approach and being its improved version, **co-training** is another semi-supervised learning technique used when only a small portion of labeled data is available. Unlike the typical process, co-training trains two individual classifiers based on two *views* of data.
- The views are basically different sets of features that provide additional information about each instance, meaning they are independent given the class. Also, each view is sufficient — the class of sample data can be accurately predicted from each set of features alone.
- The original co-training research paper claims that the approach can be successfully used, for example, for web content classification tasks. The description of each web page can be divided into two views: one with words occurring on that page and the other with anchor words in the link leading to it.

SEMI-SUPERVISED CO-TRAINING METHOD



- So, here is how co-training works in simple terms.
- First, you train a separate classifier (model) for each view with the help of a small amount of labeled data.
- Then the large unlabeled data is added to receive pseudo-labels.
- Classifiers co-train one another using pseudo-labels with the highest confidence level.
- If the first classifier confidently predicts the genuine label for a data sample while the other one makes a prediction error, then the data with the confident pseudo-labels assigned by the first classifier updates the second classifier and vice-versa.
- The final step involves the combining of the predictions from the two updated classifiers to get one classification result.

- As with self-training, co-training goes through many iterations to construct an additional training labeled dataset from the vast amounts of unlabeled data.

- **Examples of Semi-Supervised Learning**
- **Text classification**: In text classification, the goal is to classify a given text into one or more predefined categories. Semi-supervised learning can be used to train a text classification model using a small amount of labeled data and a large amount of unlabeled text data.
- **Image classification**: In image classification, the goal is to classify a given image into one or more predefined categories. Semi-supervised learning can be used to train an image classification model using a small amount of labeled data and a large amount of unlabeled image data.

- **Anomaly detection**: In anomaly detection, the goal is to detect patterns or observations that are unusual or different from the norm.

Applications of Semi-Supervised Learning

- **Speech Analysis:** Since labeling audio files is a very intensive task, Semi-Supervised learning is a very natural approach to solve this problem.
- **Internet Content Classification:** Labeling each webpage is an impractical and unfeasible process and thus uses Semi-Supervised learning algorithms. Even the Google search algorithm uses a variant of Semi-Supervised learning to rank the relevance of a webpage for a given query.
- **Protein Sequence Classification:** Since DNA strands are typically very large in size, the rise of Semi-Supervised learning has been imminent in this field.

Unit-3

Unsupervised Learning

Apriori Algorithm

The Apriori algorithm uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rule, it determines how strongly or how weakly two objects are connected. This algorithm uses a **breadth-first search** and **Hash Tree** to calculate the itemset associations efficiently. It is the iterative process for finding the frequent itemsets from the large dataset.

This algorithm was given by the **R. Agrawal** and **Srikant** in the year **1994**. It is mainly used for *market basket analysis* and helps to find those products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

What is Frequent Itemset?

Frequent itemsets are those items whose support is greater than the threshold value or user-specified minimum support. It means if A & B are the frequent itemsets together, then individually A and B should also be the frequent itemset.

Suppose there are the two transactions: A= {1,2,3,4,5}, and B= {2,3,7}, in these two transactions, 2 and 3 are the frequent itemsets.

Steps for Apriori Algorithm

Below are the steps for the apriori algorithm:

Step-1: Determine the support of itemsets in the transactional database, and select the minimum support and confidence.

Step-2: Take all supports in the transaction with higher support value than the minimum or selected support value.

Step-3: Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.

Step-4: Sort the rules as the decreasing order of lift.

Apriori Algorithm Working

We will understand the apriori algorithm using an example and mathematical calculation:

Example: Suppose we have the following dataset that has various transactions, and from this dataset, we need to find the frequent itemsets and generate the association rules using the Apriori algorithm:

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Given: Minimum Support= 2, Minimum Confidence= 50%

Solution:

Step-1: Calculating C1 and L1:

- In the first step, we will create a table that contains support count (The frequency of each itemset individually in the dataset) of each itemset in the given dataset. This table is called the **Candidate set or C1.**

Itemset	Support_Count
A	6
B	7
C	5
D	2
E	1

- Now, we will take out all the itemsets that have the greater support count that the Minimum Support (2). It will give us the table for the **frequent itemset L1.** Since all the itemsets have greater or equal support count than the minimum

support, except the E, so E itemset will be removed.

Itemset	Support_Count
A	6
B	7
C	5
D	2

Step-2: Candidate Generation C2, and L2:

- In this step, we will generate C2 with the help of L1. In C2, we will create the pair of the itemsets of L1 in the form of subsets.
- After creating the subsets, we will again find the support count from the main transaction table of datasets, i.e., how many times these pairs have occurred together in the given dataset. So, we will get the below table for C2:

Itemset	Support_Count
{A, B}	4
{A,C}	4
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	0

- Again, we need to compare the C2 Support count with the minimum support count, and after comparing, the itemset with less support count will be eliminated from the table C2. It will give us the below table for L2

Itemset	Support_Count
{A, B}	4
{A, C}	4
{B, C}	4
{B, D}	2

A, B, C, D

Step-3: Candidate generation C3, and L3:

- For C3, we will repeat the same two processes, but now we will form the C3 table with subsets of three itemsets together, and will calculate the support count from the dataset. It will give the below table:

Itemset	Support_Count
{A, B, C}	2
{B, C, D}	1
{A, C, D}	0
{A, B, D}	0

- Now we will create the L3 table. As we can see from the above C3 table, there is only one combination of itemset that has support count equal to the minimum support count. So, the L3 will have only one combination, i.e., **{A, B, C}.**

Step-4: Finding the association rules for the subsets:

To generate the association rules, first, we will create a new table with the possible rules from the occurred combination {A, B.C}. For all the rules, we will calculate the Confidence using formula **sup(A ^B)/A.** After calculating the confidence value for all rules, we will exclude the rules that have less confidence than the minimum threshold(50%).

Consider the below table:

Rules	**Support**	**Confidence**
A ^B → C	2	Sup{(A ^B) ^C}/sup(A ^B)= 2/4=0.5=50%
B^C → A	2	Sup{(B^C) ^A}/sup(B ^C)= 2/4=0.5=50%
A^C → B	2	Sup{(A ^C) ^B}/sup(A ^C)= 2/4=0.5=50%
C→ A ^B	2	Sup{(C^(A ^B)}/sup(C)= 2/5=0.4=40%
A→ B^C	2	Sup{(A^(B ^C)}/sup(A)= 2/6=0.33=33.33%
B→ B^C	2	Sup{(B^(B ^C)}/sup(B)= 2/7=0.28=28%

As the given threshold or minimum confidence is 50%, so the first three rules **A ^B → C, B^C → A, and A^C → B** can be considered as the strong association rules for the given problem.

Advantages of Apriori Algorithm

- This is easy to understand algorithm
- The join and prune steps of the algorithm can be easily implemented on large datasets.

Disadvantages of Apriori Algorithm

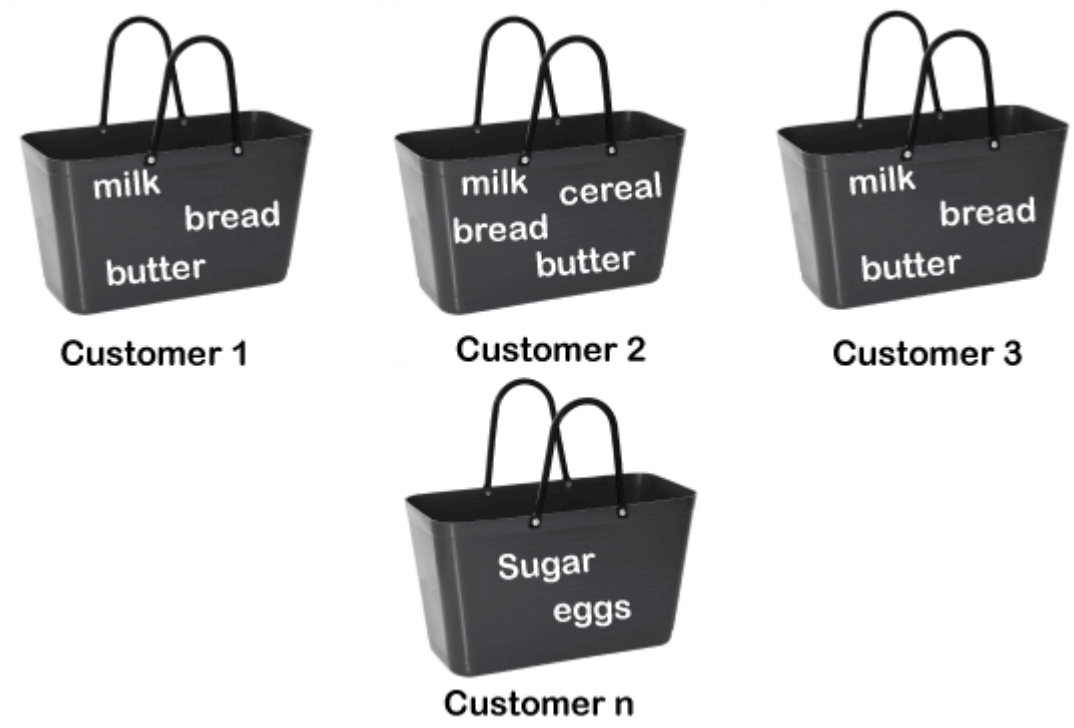
- The apriori algorithm works slow compared to other algorithms.
- The overall performance can be reduced as it scans the database for multiple times.
- The time complexity and space complexity of the apriori algorithm is $O(2^D)$, which is very high. Here D represents the horizontal width present in the database.

Association Rule Learning

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

The association rule learning is one of the very important concepts of machine learning, and it is employed in **Market Basket analysis, Web usage mining, continuous production, etc.** Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:



Association rule learning can be divided into three types of algorithms:

1. **Apriori**
2. **Eclat**
3. **F-P Growth Algorithm**

How does Association Rule Learning work?

Association rule learning works on the concept of If and Else Statement, such as if A then B.

Here the If element is called **antecedent**, and then statement is called as **Consequent**. These types of relationships where we can find out some association or relation between two items is known *as single cardinality*. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:

- **Support**
- **Confidence**
- **Lift**

Let's understand each of them:

Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{Freq(X)}{T}$$

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

$$\text{Confidence} = \frac{Freq(X,Y)}{Freq(X)}$$

Lift

It is the strength of any rule, which can be defined as below formula:

$$\text{Lift}= \frac{Supp(X,Y)}{Supp(X)\times Supp(Y)}$$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

- If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1**: It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

Types of Association Rule Learning

Association rule learning can be divided into three algorithms:

Apriori Algorithm

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Eclat Algorithm

Eclat algorithm stands for **Equivalence Class Transformation**. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

F-P Growth Algorithm

The F-P growth algorithm stands for **Frequent Pattern**, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

Applications of Association Rule Learning

It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

- **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
- **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
- **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
- It is also used for the **Catalog Design** and **Loss-leader Analysis** and many more other applications.