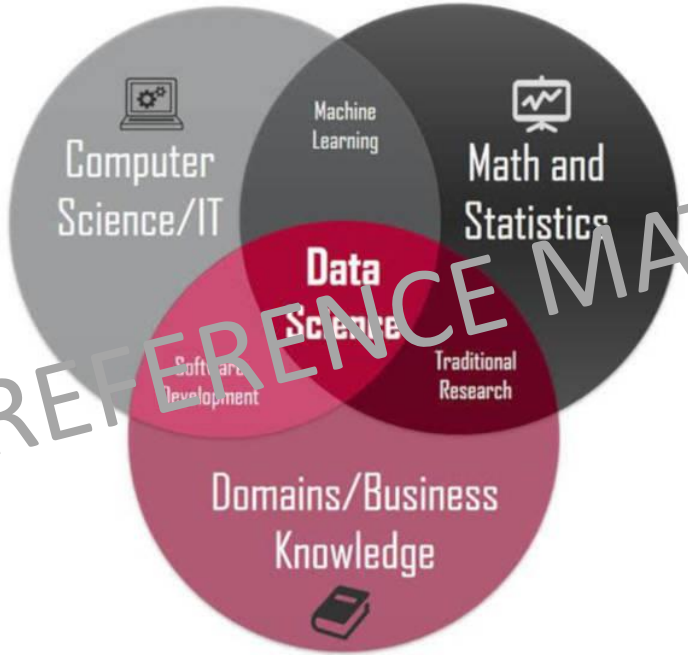
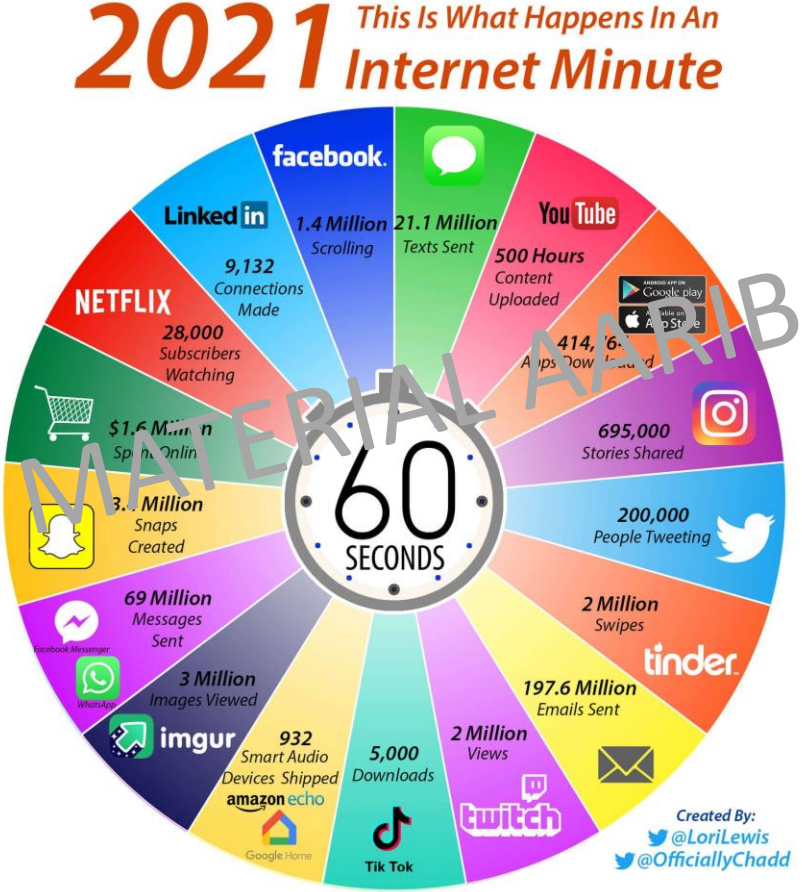

REFERENCE MATERIAL AARIB TRAINER
AI

Data Science

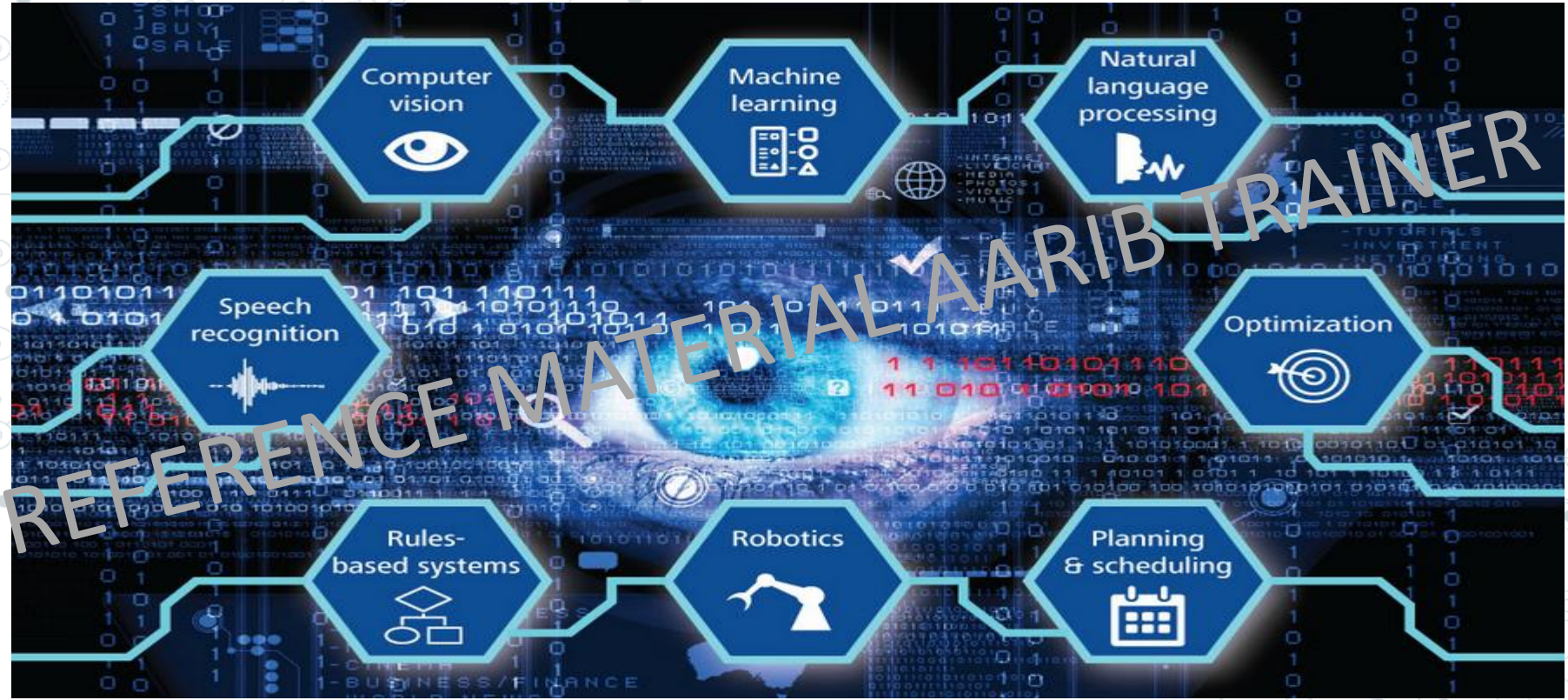


HUGE AMOUNTS OF DATA

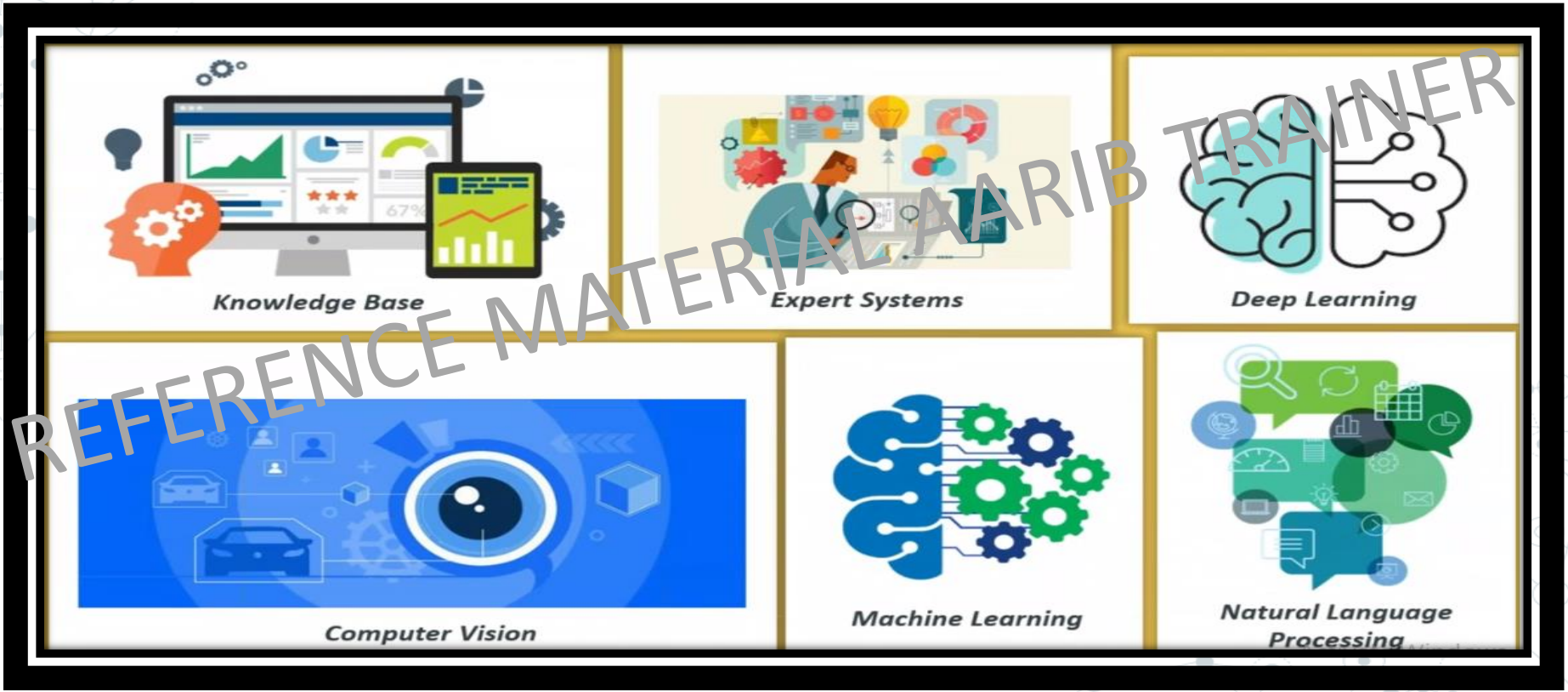
Big Data

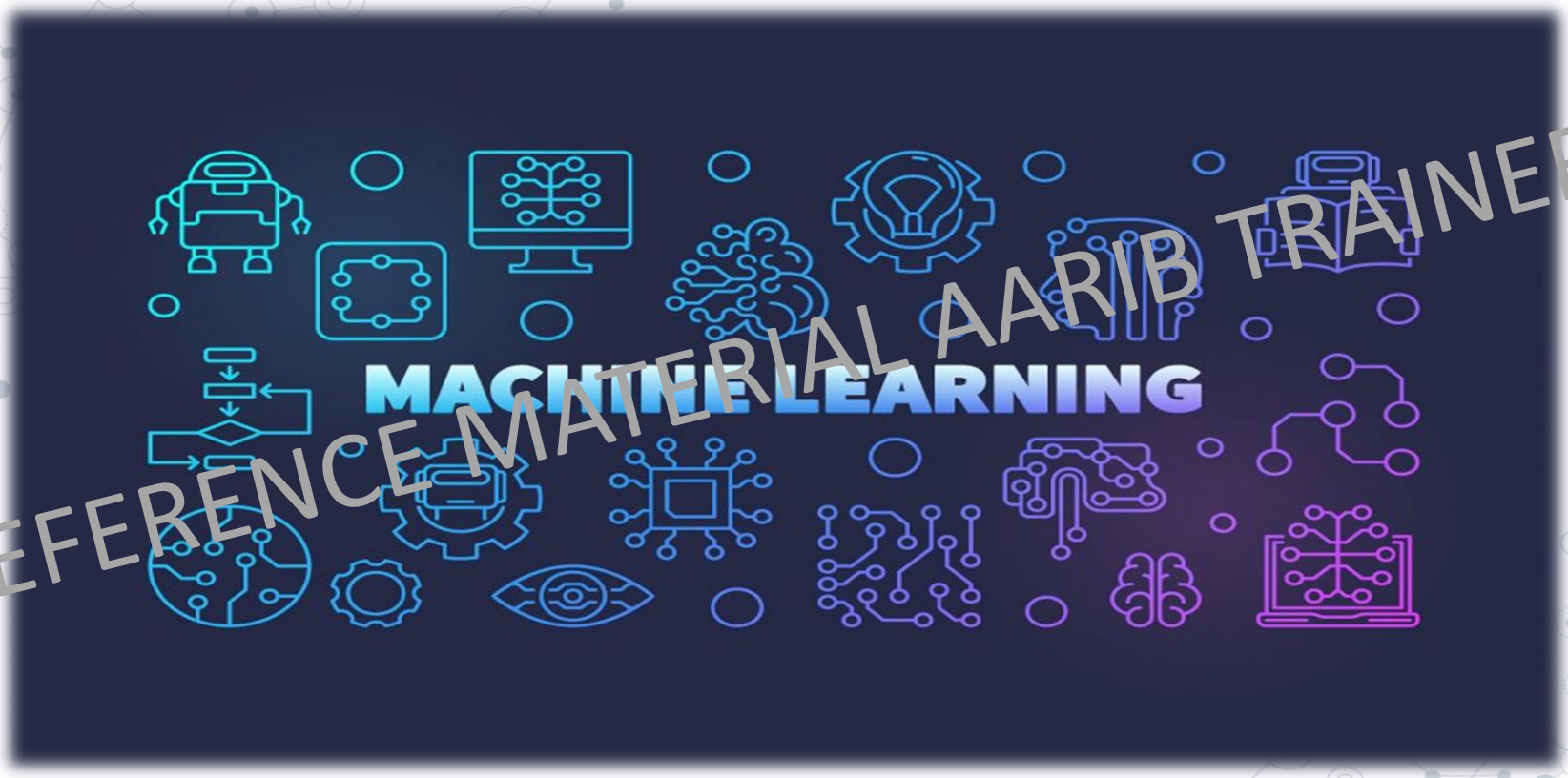


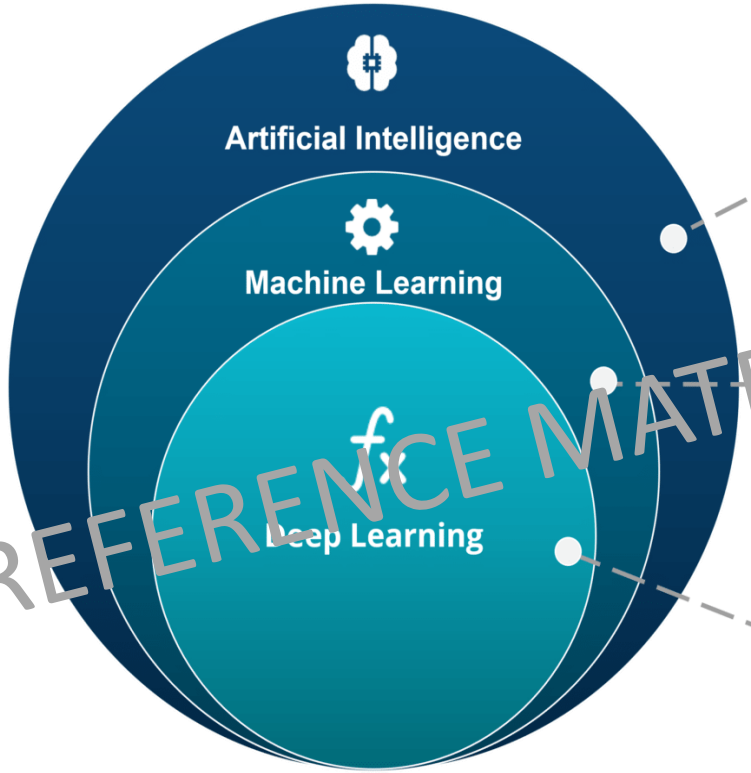
AI - Domains


Computer vision
Machine learning
Natural language processing
Speech recognition
Optimization
Rules-based systems
Robotics
Planning & scheduling
REFERENCE MATERIAL AARIB TRAINER

AI - Domains







ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour

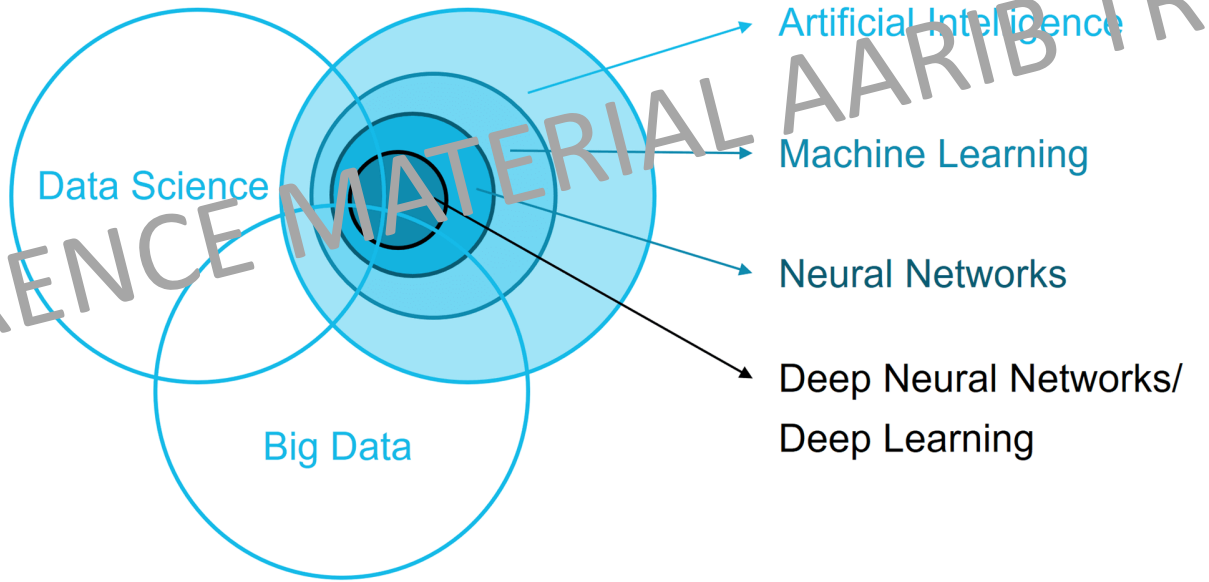
MACHINE LEARNING

Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING

Subset of ML which make the computation of multi-layer neural network feasible

AI vs ML vs DS



1.

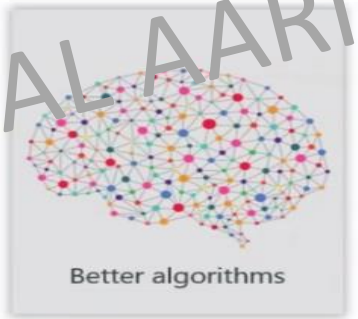
History of AI

John McCarthy - 1956

Science and Engineering of making Intelligent machines

2.5 quintillion bytes of data are created every single day

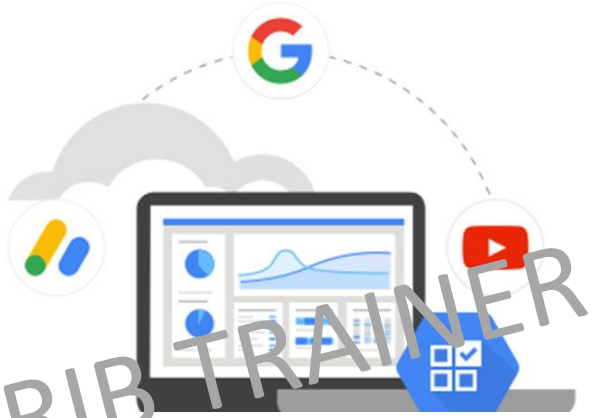
Why AI is famous now?



2.5 quintillion bytes(also called ExaBytes) of data is generated everyday

2.

Applications of AI



Google

Recommendation System

(Logic behind this is AI)

IBM Watson

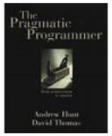
IBM Watson

A Supercomputer used in Health Care to cure rare Lukemia

https://www.asianscientist.com/2016/08/topnews/ibm-watson-rare-leukemia-university-tokyo-artificial-intelligence/

Amazon Recommendation System

Frequently Bought Together


+


Total price: $83.09

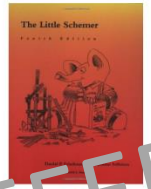
Add both to Cart

Add both to List

- **This item:** Structure and Interpretation of Computer Programs - 2nd Edition (MIT Electrical Engineering and… by Harold Abelson Paperback $50.50
- The Pragmatic Programmer: From Journeyman to Master by Andrew Hunt Paperback $32.59

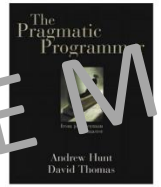
Customers Who Bought This Item Also Bought

Page 1 of 13

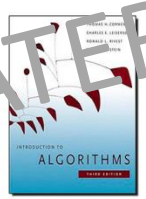


The Little Schemer - 4th Editio…
› Daniel P. Friedman
64
Paperback
$36.00 Prime

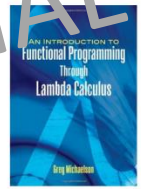
Instructor's Manual t/a Structure and Interpretation of Computer Programs…
› Gerald Jay Sussman
5
Paperback
$28.70 Prime



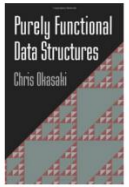
The Pragmatic Programmer: From Journeyman to Master
› Andrew Hunt
328
Paperback
$32.59 Prime



Introduction to Algorithms, 3rd Edition (MIT Press)
› Thomas H. Cormen
313
#1 Best Seller in Computer Algorithms
Hardcover
$66.32 Prime

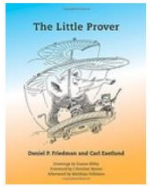


An Introduction to Functional Programming Through Lambda…
› Greg Michaelson
23
Paperback
$20.70 Prime



Purely Functional Data Structures
› Chris Okasaki
19
Paperback
$40.74 Prime

Code: The Hidden Language of Computer Hardware and Software
› Charles Petzold
334
#1 Best Seller in Machine Theory
Paperback
$17.99 Prime



The Little Prover (MIT Press)
Daniel P. Friedman
4
Paperback
$31.78 Prime

Facebook

Auto Tagging feature in FB

Self Driving Cars

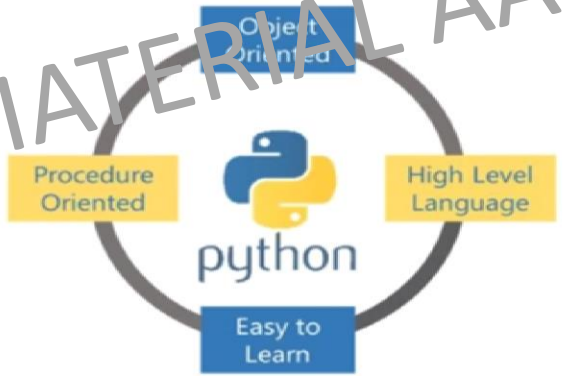
Uses Deep Learning

Languages for AI

Python
R Language
Java
Lisp

INTRODUCTION TO PYTHON

- Python is an interpreted, object-oriented, high-level programming language with dynamic semantics
- Python was created by Guido Rossum in 1989 and is very easy to learn



Version 3.9

Features of Python

- Simple and Easy to understand
- Free and Open Source
- High- Level Language
- Portable and Extensible
- Supports multi-paradigm

Softwares

- Python IDLE
- PyCharm
- Jupyter Notebook
- Google Colaboratory
- Kaggle Notebook

Libraries

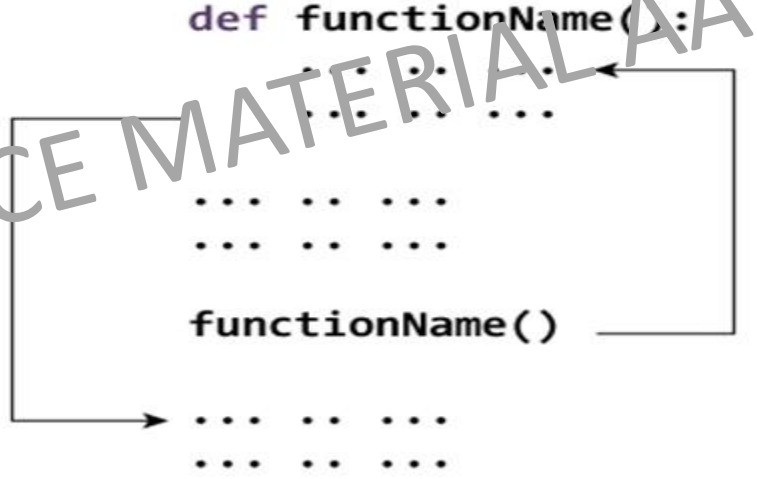
- Numpy
- Matplotlib
- Pandas
- Scikit-learn / Sklearn
- jupyter
- opencv-python

Python Core

1. Hello World
2. Normal Calculator
3. Data types – Int, Float, Char, Boolean
4. Variables
5. Functions
6. For loop
7. While loop
8. Break statement
9. If and else statements

Functions

How Function works in Python?



```
def weatherForecast(weather):
    if(weather=="Sunny"):
        print("it is hot day")
    elif(weather=="Rainy"):
        print("Its Raining")
    else:
        print("Cool Weather")

weatherForecast("Rainy")
```

Sequences in Python

1. Lists
2. Tuples
3. Dictionary

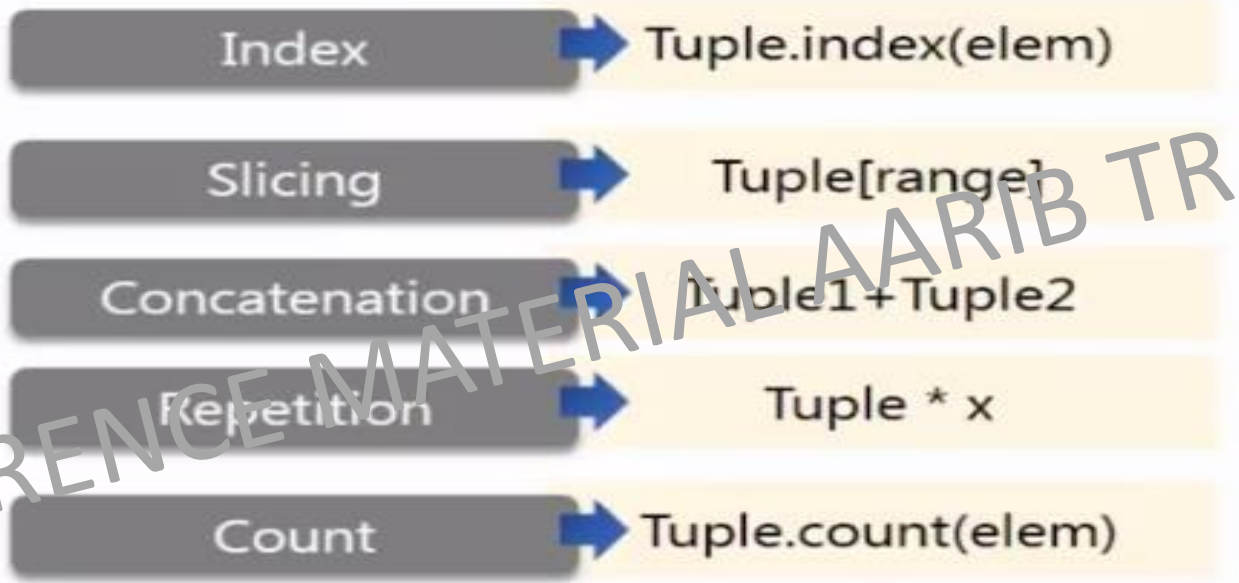
Lists

The list is a most versatile datatype available in Python which can be written as a list of comma-separated values (items) between square brackets. Important thing about a list is that items in a list need not be of the same type.

Fruits=[‘Mango’,’Apple’,’Orange’,]

Tuples

Fruits=('Mango','Apple','Orange')



NOTE: Tuple is immutable and is written in paranthesis

Try **Tuple_name.__len__()**

Why Tuples when we have lists?

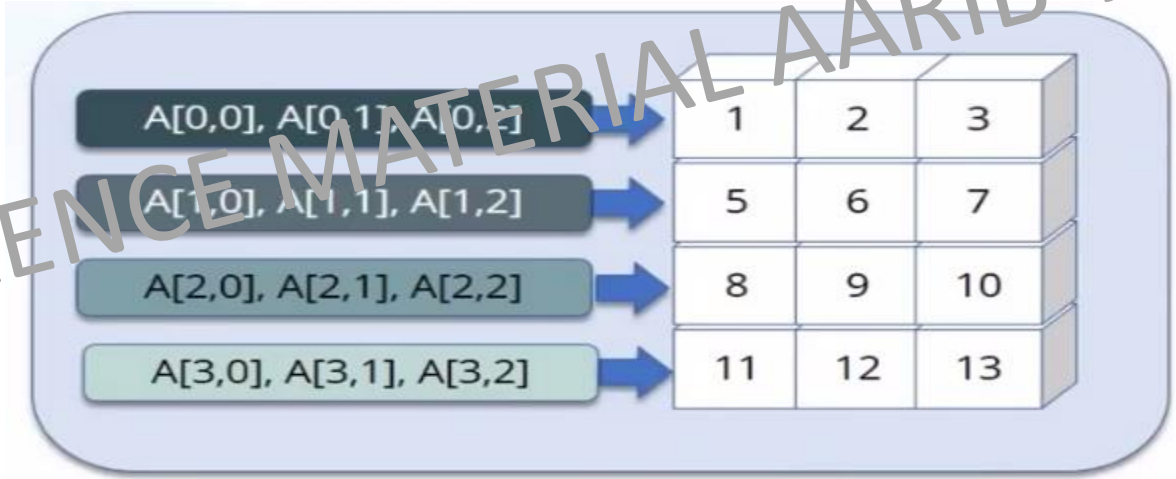
1. Iterating through tuples is easier and faster
2. Tuples are immutable and is secure

Dictionaries

Numpy

Library for scientific computing

Difference between List and Numpy



Numpy Functions

1. np.array()
2. np.ones()
3. np.zeros()
4. np.arange(10)
5. np.linspace(1,3,5)
6. np.sum()
7. np.max()
8. np.min()
9. np.mean()

1.a.size

2.a.ndim

3.a.itemsize

4.a.dtype

5.a.shape

matplotlib.pyplot

Plotting library used for 2D graphics

- Line plot
- Scatter plot
- Bar Graph
- Pie Plot

Matplotlib

```
import numpy as np
import matplotlib.pyplot as plt

x = np.arange(0,3*np.pi, 0.1)

y = np.sin(x)

plt.plot(x,y)

plt.show()
```

Pandas

```
Import pandas as pd
df = pd.read_csv('path.csv')
```

pandas

1. DataFrame – 2 Dimesions
2. Series – 1 Dimension

How to create DataFrames from different types of Data Files?

- pd.DataFrame() = List/Tuple/Dictionary/Numpy Arrays
- pd.read_csv() = CSV Files [Comma Separated Files]
- pd.read_excel() = Excel Spreadsheet
- pd.read_table() = Text File

Pandas Methods

1. head() – Prints top 5 rows
2. tail() – Prints top 5 columns
3. unique() – Unique values from a specific Column
4. nunique()- Tells the number of unique values
5. value_counts() – Tells the count of unique values
6. to_datetime() – Converts the value from object to datetime datatype

- df.shape – No.of values in each axis [rows*Columns]
- df.columns – All the column names will be printed
- df.info() – Gives the summary of the DataFrame
- df.describe() – Summary Statistics of DataFrame
- df.size – Prints the total number of values in dataframe

.iloc operator

Purely integer-location based indexing for selection by position.

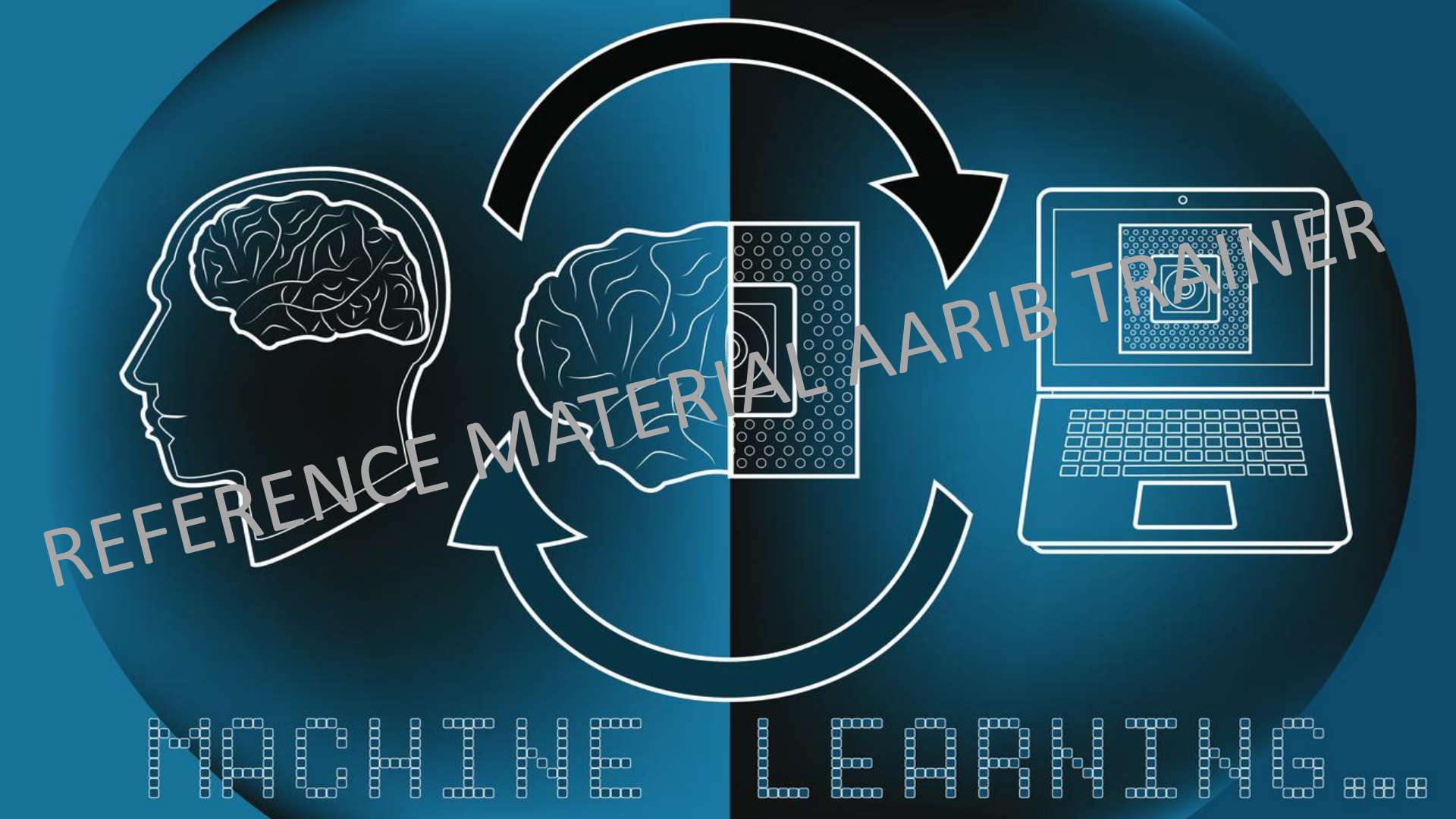
df.iloc[rows , columns]

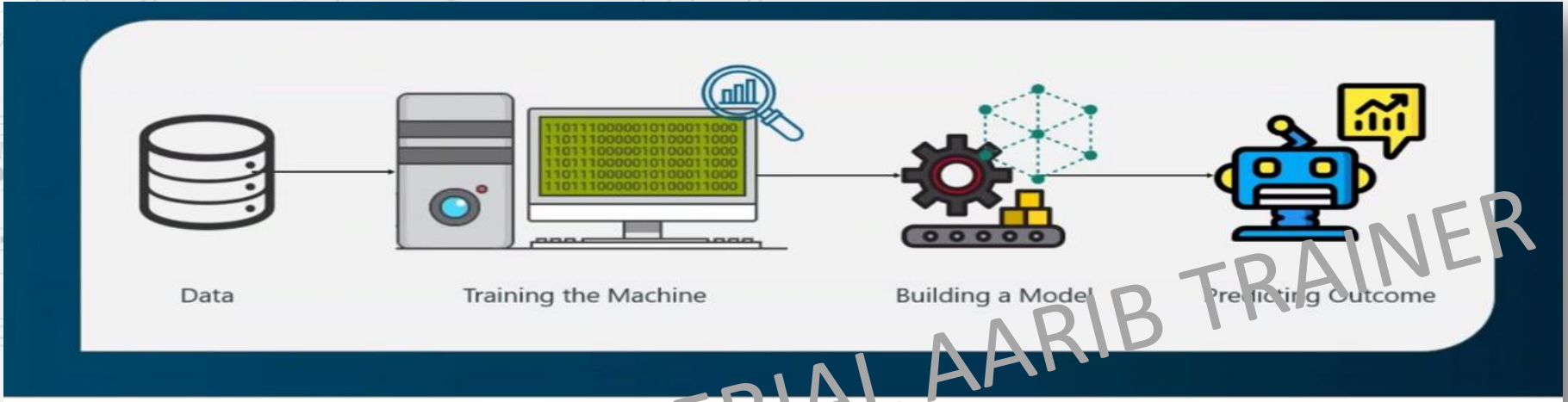
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html

Is Machine Learning same as Artificial Learning?

ML is used in AI

ML is way in which we feed in lot of data to a machine so that it makes its own decision


REFERENCE MATERIAL AARIB TRAINER
MACHINE LEARNING...



A simple definition of Machine Learning

Machine learning is a subset of Artificial Intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed.

Data can be either : 1. Numbers
2. Text
3. Image
4. Videos
5. Audios

Terminologies

- Algorithm - - Sets of procedures used to create a model from data
- Model – A trained machine
- Predictor Variable – Variable containing the Predicted output
- Training data – Data given to machine
- Testing data – Data used by machine to predict
- Features – Inputs Values of Data
- Labels/Targets – Output Values of Data

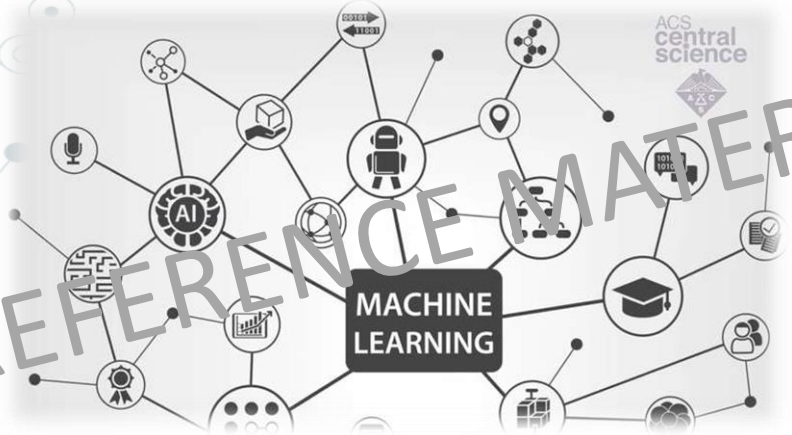
Pre-requisites to learn ML

- Linear Algebra
- Statistics and Probability
- Calculus
- Graph theory
- Programming Skills – Language such as Python, R, MATLAB, C++ or Octave

Features of ML

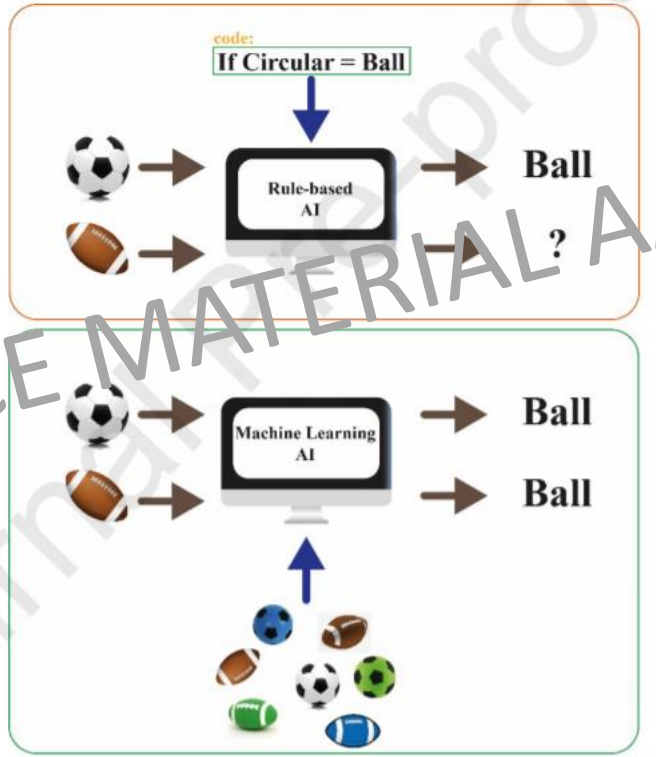
- Detects the pattern from huge dataset
- Finds hidden insights using Algorithms without being explicitly programmed
- Automates the analytical model

Real World ML use cases



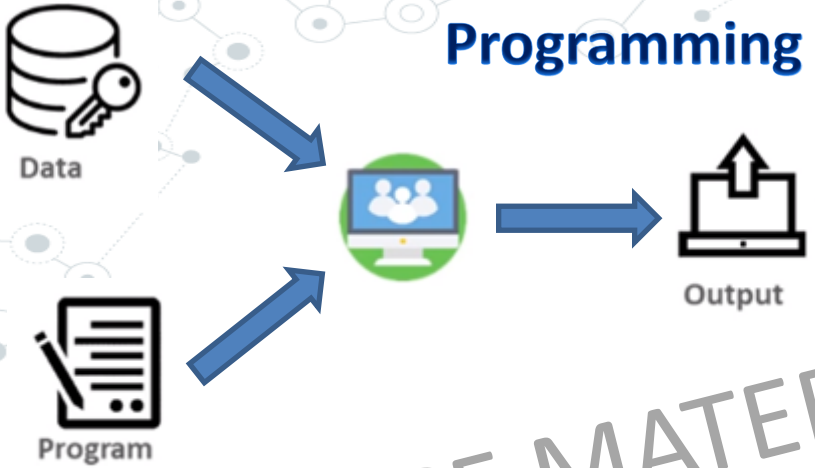
- SPEECH RECOGNITION
- CUSTOMER SERVICE
- COMPUTER VISION
- RECOMMENDATION ENGINES
- AUTOMATED STOCK TRADING

Rule Based Approach vs ML based Approach



How it works?

Traditional Programming

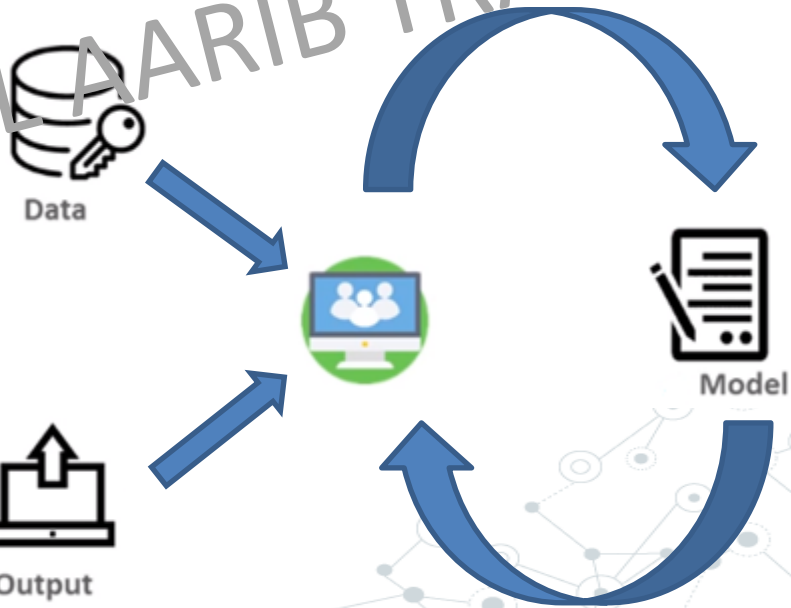


def sum(a,b):
    print(a+b) ⟵ Program

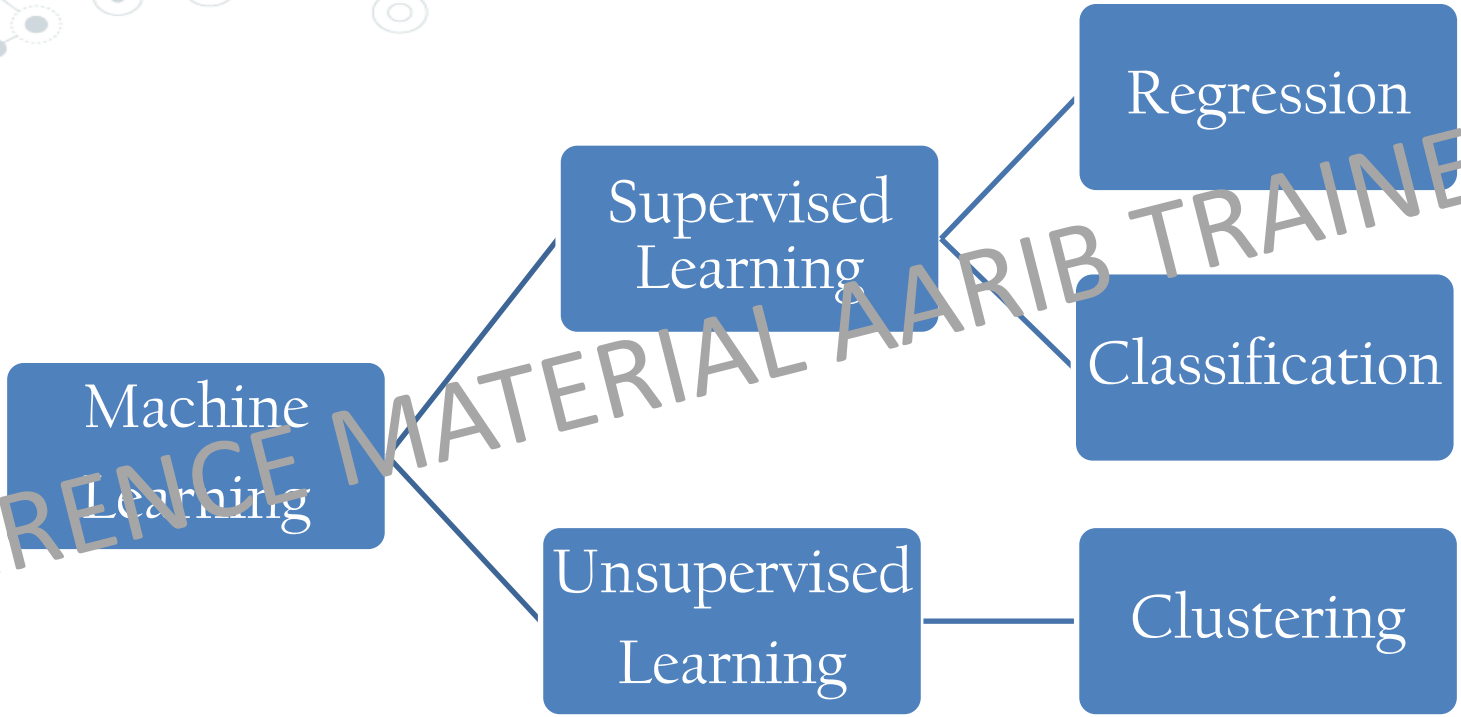
sum(2,4) ⟵ Data

>>>6 ⟵ Output

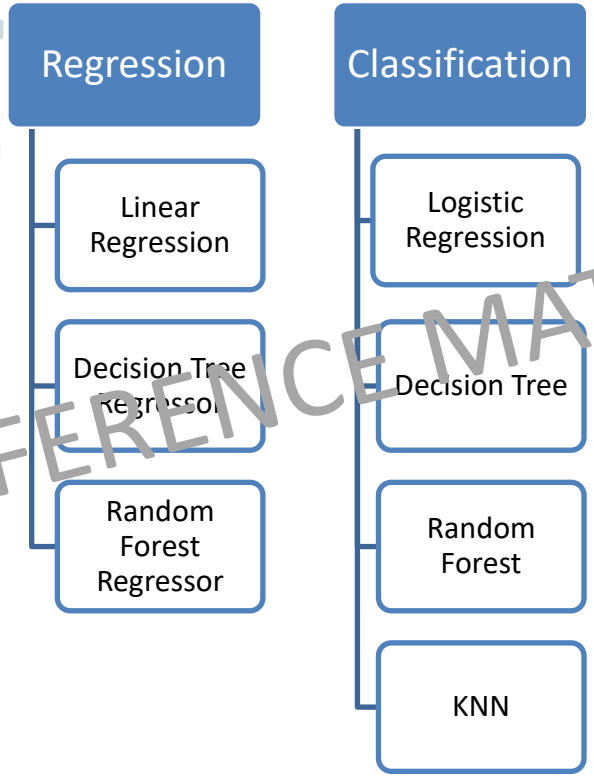
Machine Learning Programming



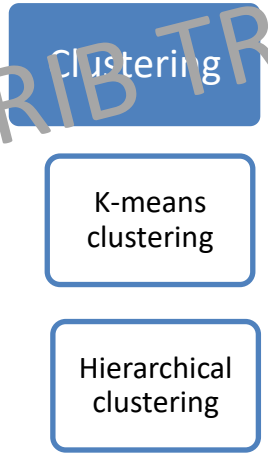
- Learns from data
- Finds insights
- Train and grow



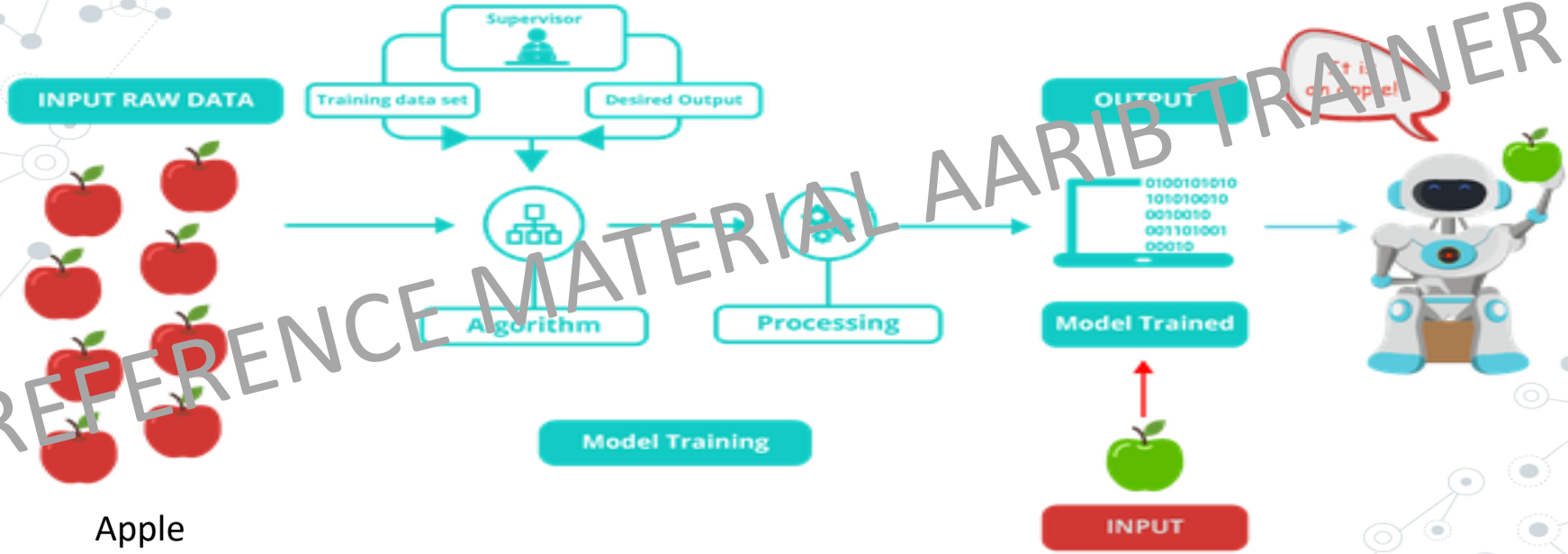
Supervised Learning



Unsupervised Learning



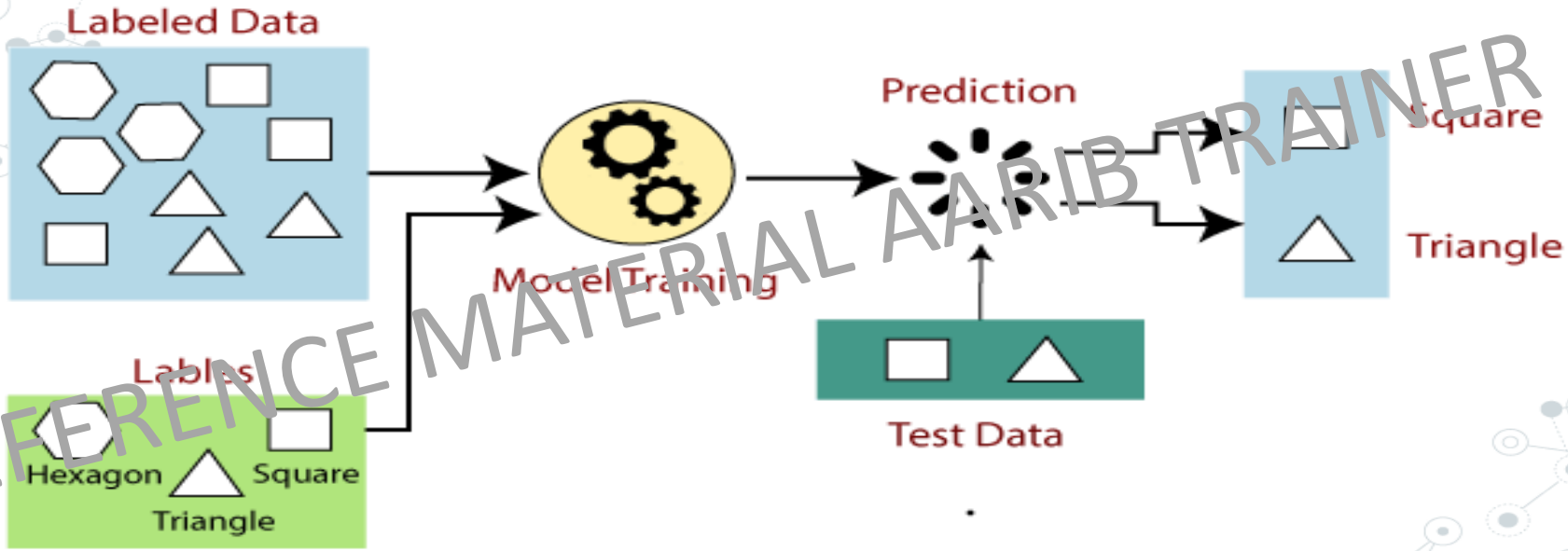
Supervised Learning



Apple

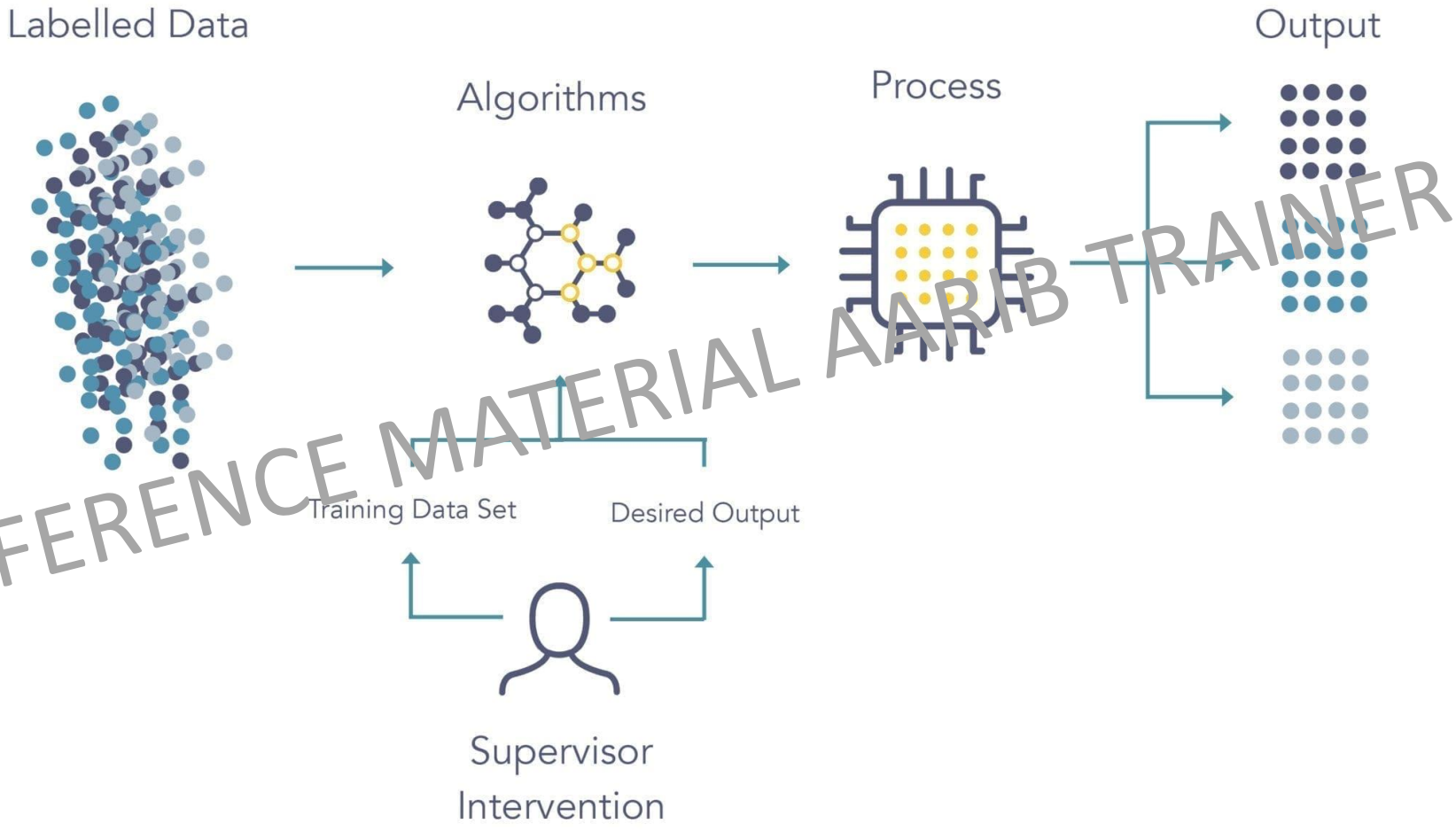
The data is well labeled

Supervised Learning

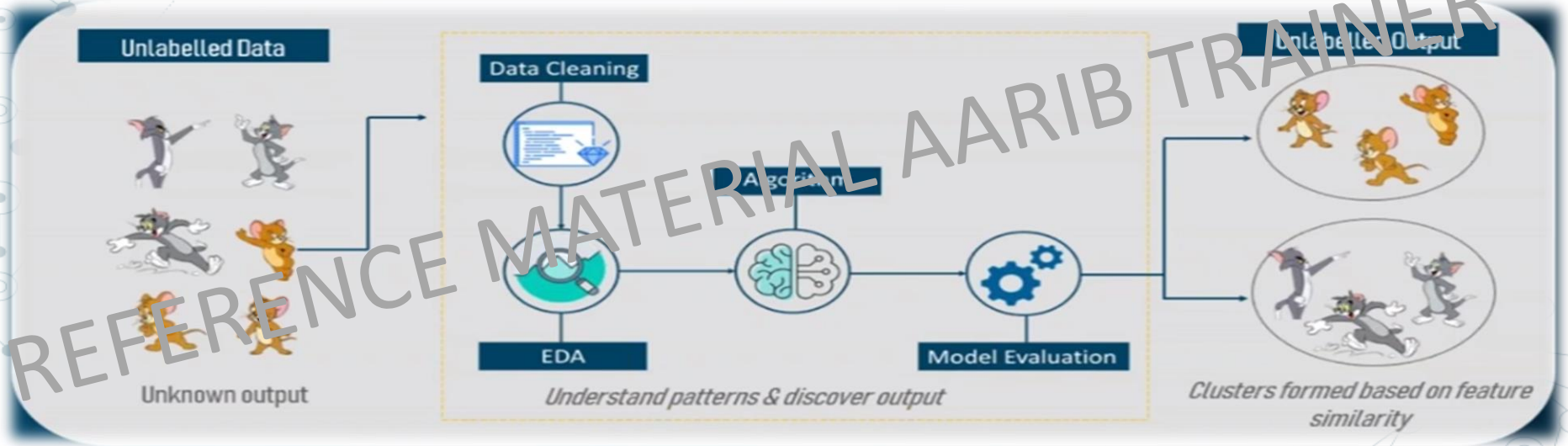


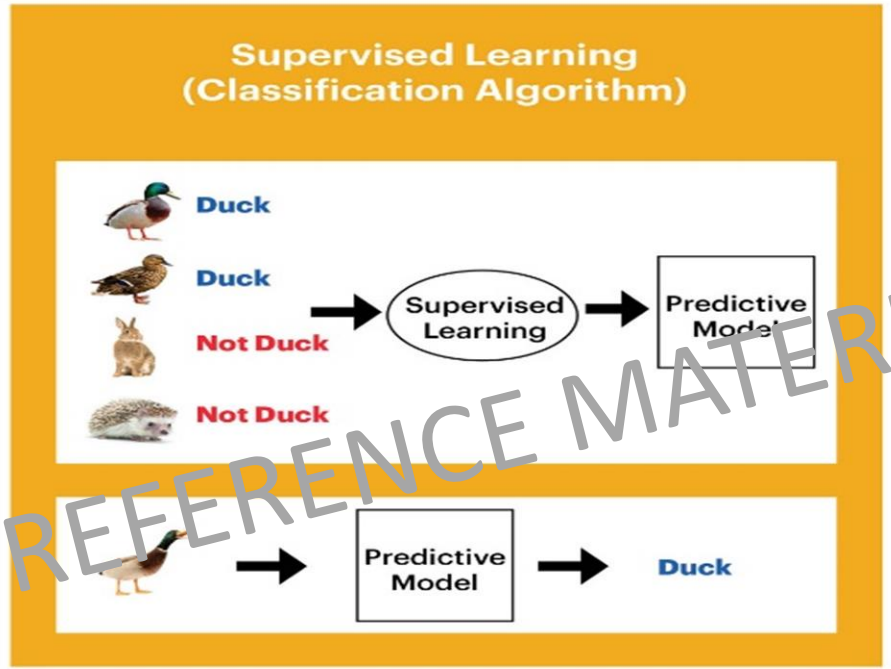
The data is well labeled

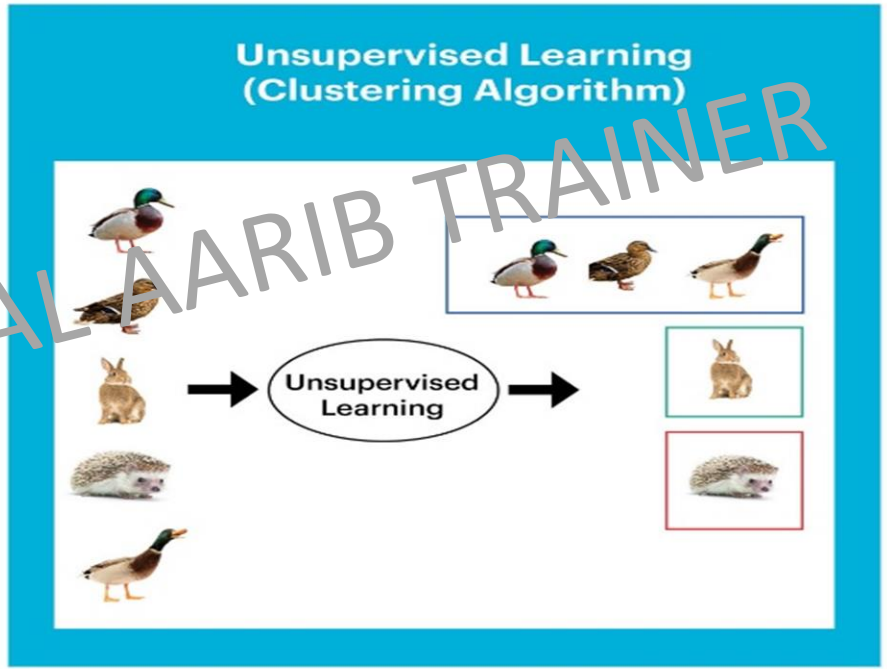
- Classification
- Regression



Unsupervised Learning

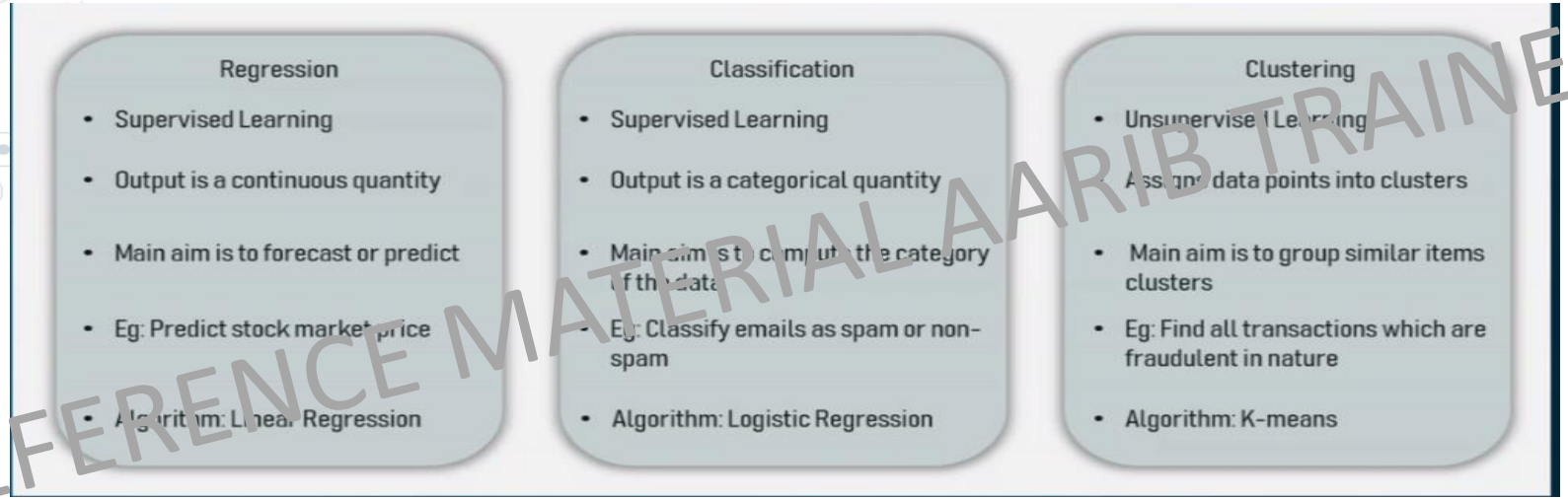






Western Digital.

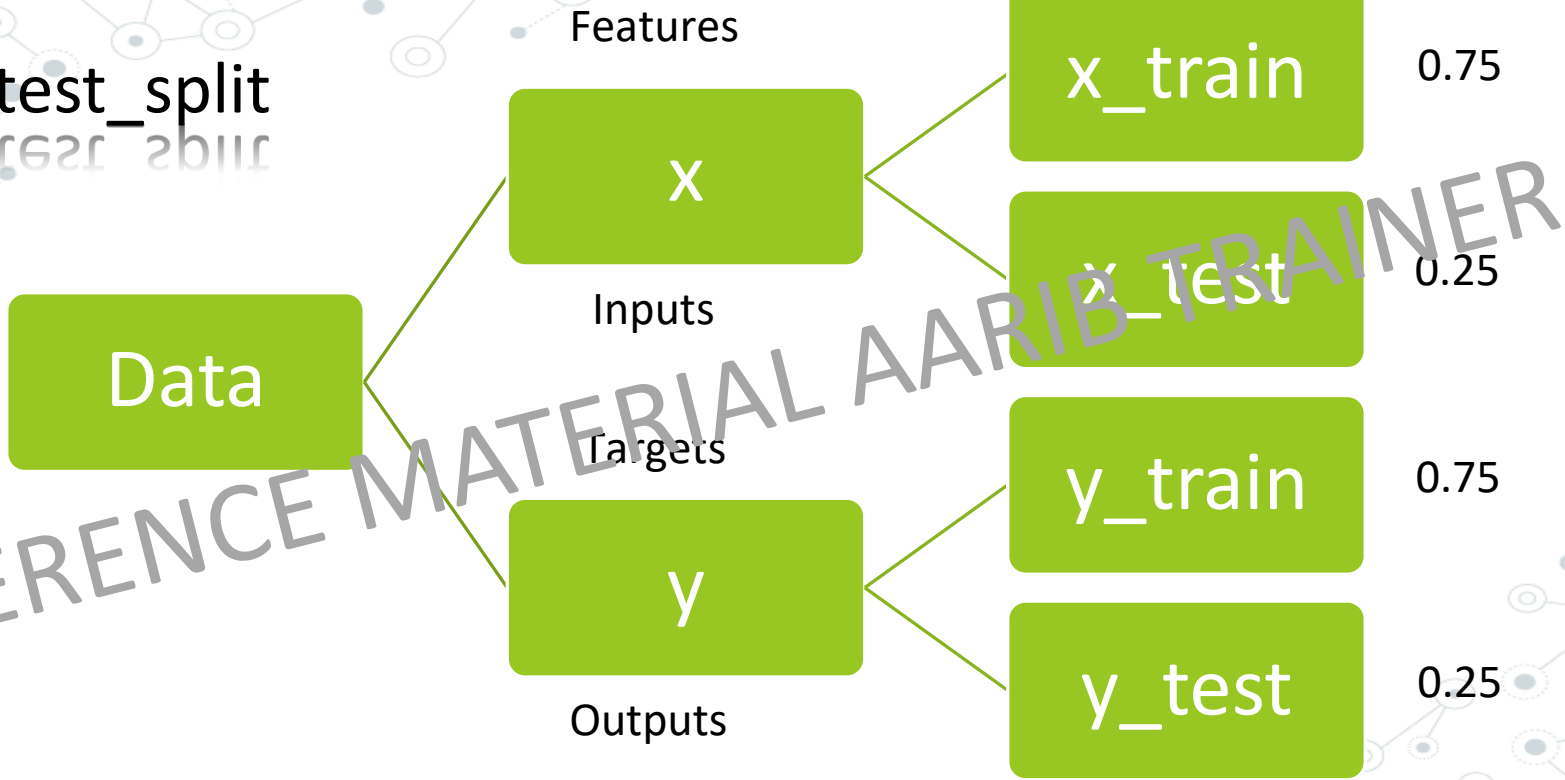
	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The machine learns by using labelled data	The machine is trained on unlabelled data without any guidance	An agent interacts with its environment by producing actions & discovers errors or rewards
Type of problems	Regression & Classification	Association & Clustering	Reward based
Type of data	Labelled data	Unlabelled data	No pre-defined data
Training	External supervision	No supervision	No supervision
Approach	Map labelled input to known output	Understand patterns and discover output	Follow trail and error method
Popular algorithms	Linear regression, Logistic regression, Support Vector Machine, KNN, etc	K-means, C-means, etc	Q-Learning, SARSA, etc



Steps in building ML model

1. Take the Data and create a dataframe
2. Preprocessing – Filtering of Data (Data Cleaning, Encoding, Dropping values, Missing values) [EDA]
3. Data Visualization
4. Divide into Input and Output (x - i/p , y-o/p)
5. Train and Test Variables
6. Normalize (Scaling) the data (Inputs only)
7. Run a Classifier/Regressor/Clusterer
8. Fit the model (Map inputs with output)
9. Predict the output
10. Evaluation : r2 score, accuracy score , Confusion Matrix

train_test_split



Available in model_selection library of sklearn package

LINEAR REGRESSION

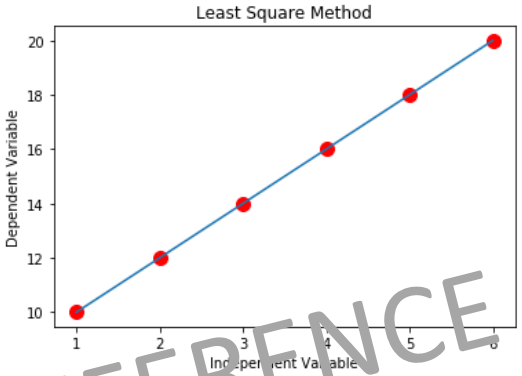
Linear relationship between **input(x)** and **output(y)**

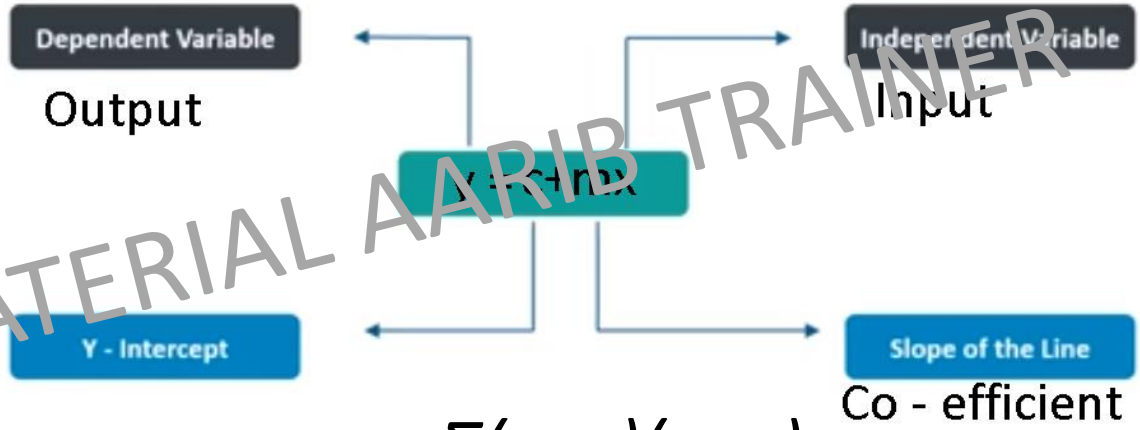
Equation of straight line [y = mx + c]

1. **X- Input**
2. **Y- Output**
3. **M- Slope**
4. **C- Intercept**

- **SIMPLE LINEAR REGRESSION (1 input column)**
- **MULTI LINEAR REGRESSION (multiple input columns)**
- **Method used – ordinary least square method**

Linear Regression – From Scratch





- $m = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

$\bar{x}$ - mean of x

$\bar{y}$ - mean of y

https://www.socscistatistics.com/tests/regression/default.aspx

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	10	-2.5	-5	6.25	12.5
2	12	-1.5	-3	2.25	4.5
3	14	-0.5	-1	0.25	0.5
4	16	0.5	1	0.25	0.5
5	18	1.5	3	2.25	4.5
6	20	2.5	5	6.25	12.5
Mean = 3.5	Mean = 15.0			$\sum$ = 17.5	$\sum$ = 35

$$m = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

$\bar{x}$ - mean of x

$\bar{y}$ - mean of y

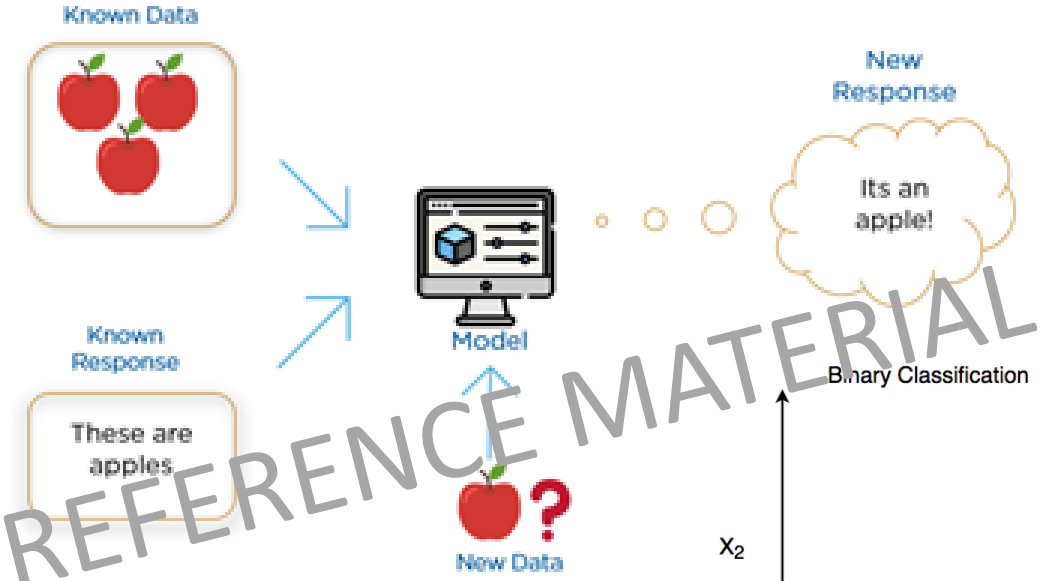
m = 2

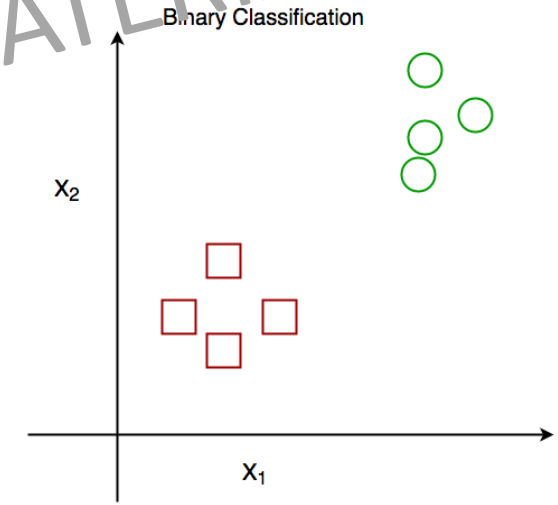
y = mx+c

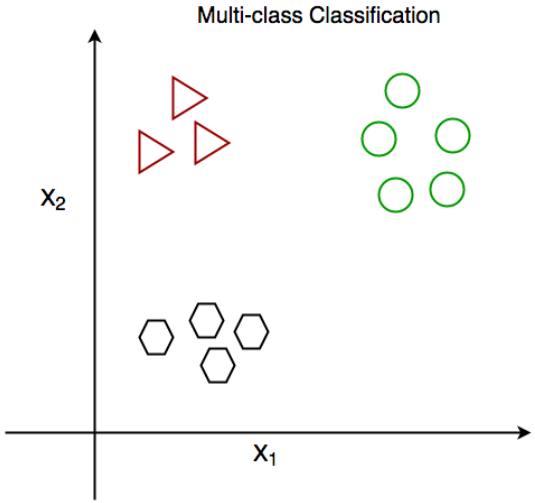
c= ymean-m*xmean

c =8

Classification







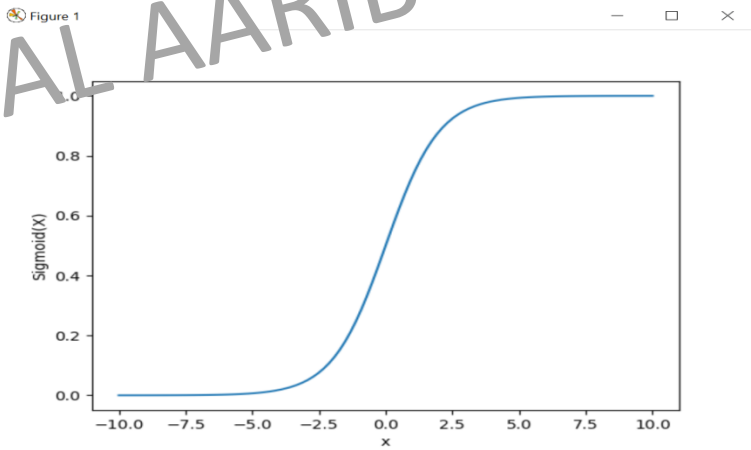
LOGISTIC REGRESSION

Logistic Regression

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

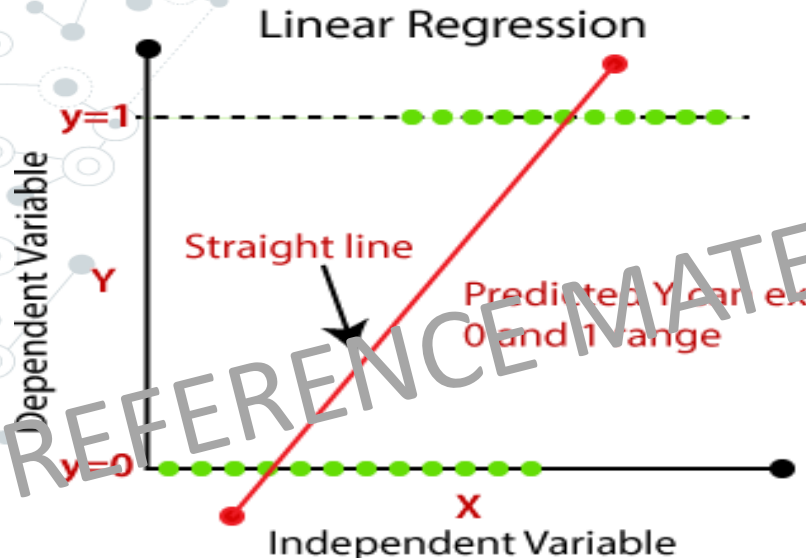
Sigmoid Function in Matplotlib

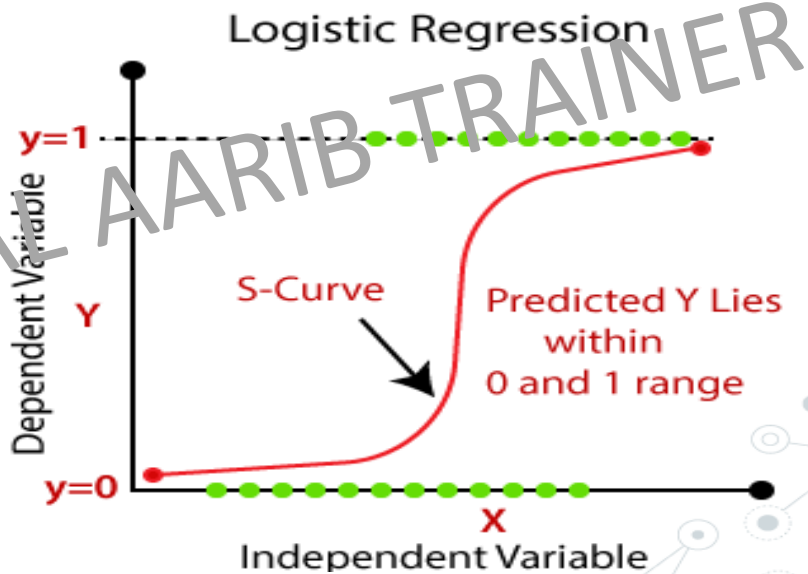
$$f(x) = \frac{1}{1 + e^{-(x)}}$$



Logic behind Logistic Regression

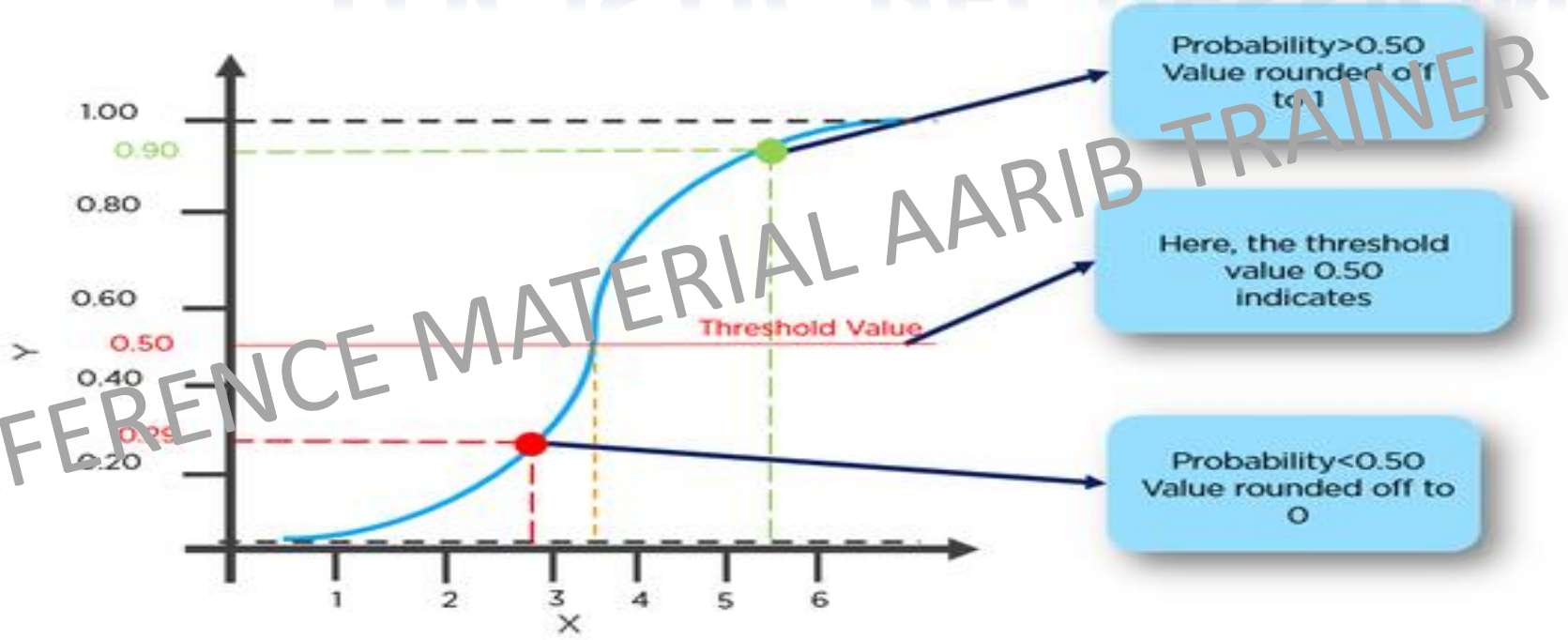
Linear vs Logistic Regression





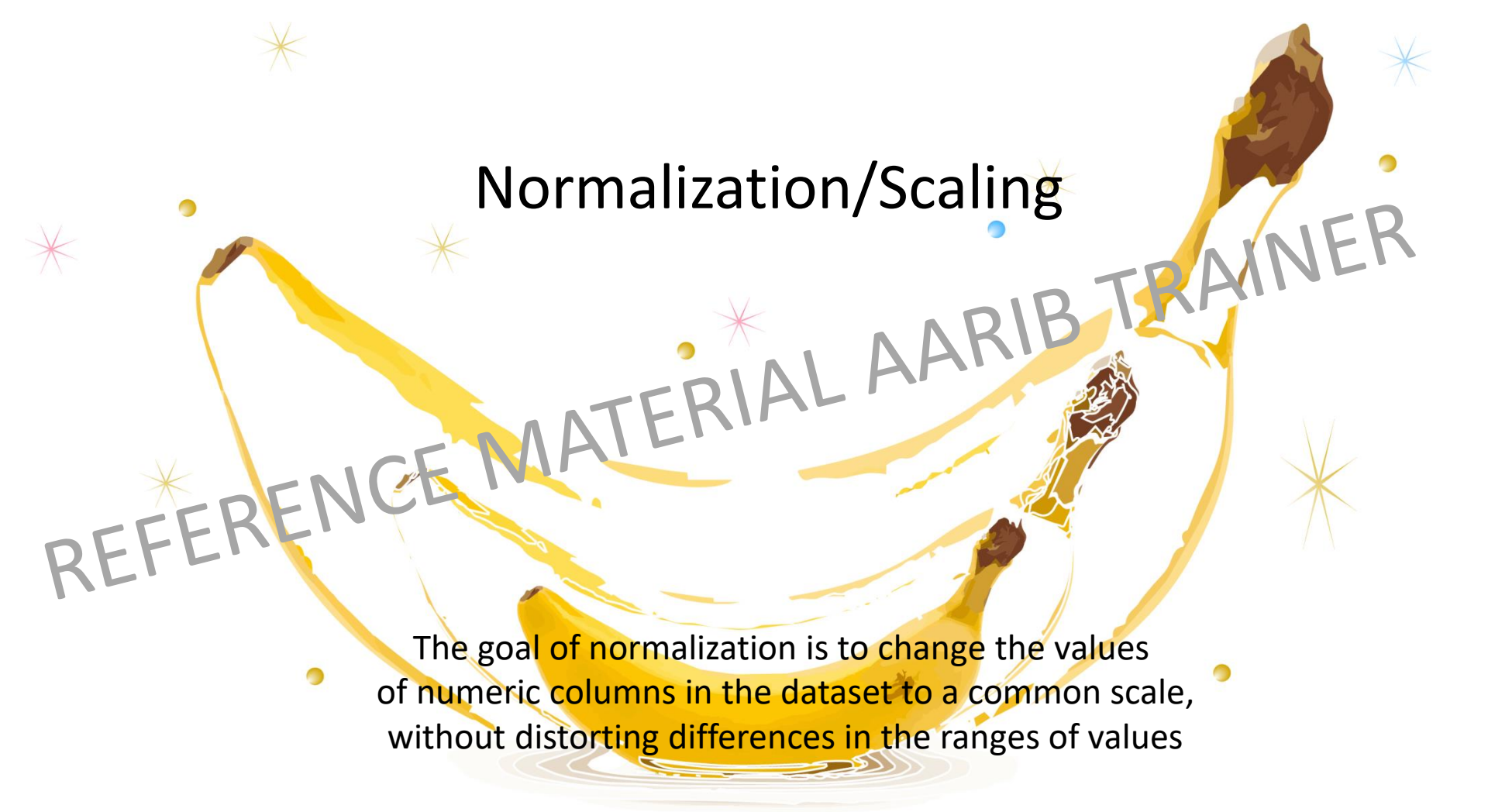
Red : Apple
Yellow : Not Apple

LOGISTIC REGRESSION



$$y = f(x) = \frac{1}{1 + e^{-(x)}}$$

Normalization/Scaling



The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values

Types

1. Min Max Scaler – (0,1)
2. Standard Scaler (0 as median)

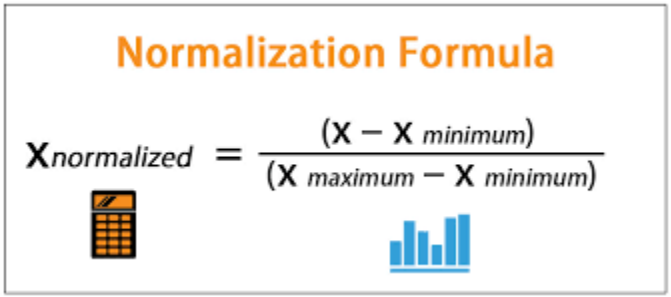
Standard Scaler

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

Min Max Scaler



How to use Min Max Scaler?

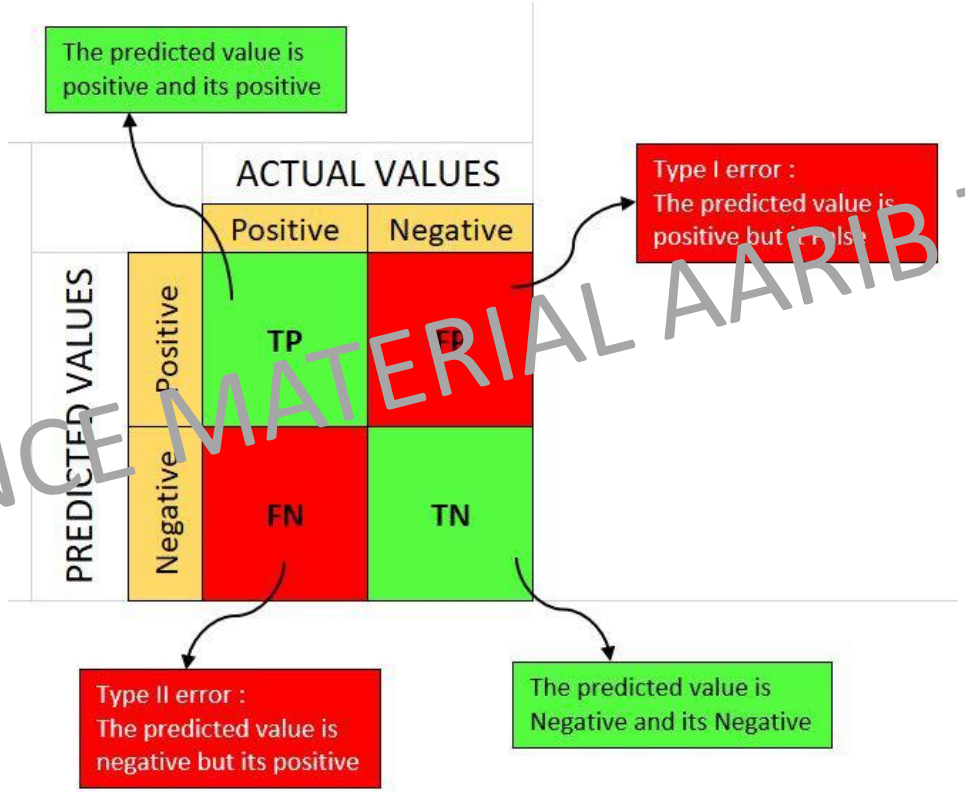
- For *X training set*, we do "fit_transform" because we need to compute features, and then use it to autoscale the data. For *X test set*, well, we already have the features, so we only do the "transform" part.

https://stackoverflow.com/questions/48692500/fit-transform-on-training-data-and-transform-on-test-data
https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Confusion Matrix

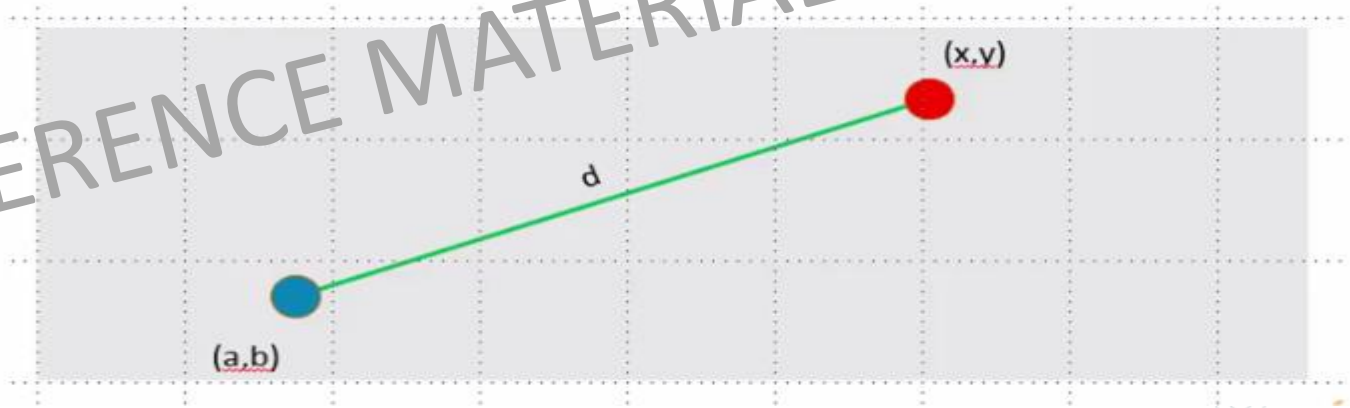


K NEAREST NEIGHBORS
Classifier

KNN

According to the **Euclidean distance** formula, the **distance** between two points in the plane with coordinates (x, y) and (a, b) is given by:

$$dist(d) = \sqrt{(x - a)^2 + (y - b)^2}$$

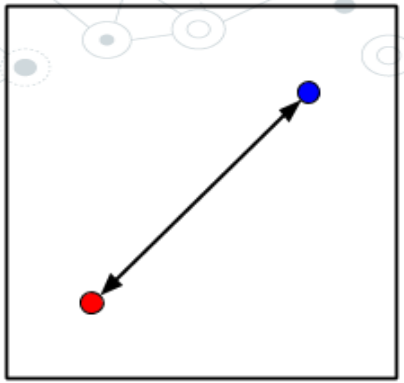


- **K in KNN** is a parameter that refers to the number of nearest neighbours to a particular data point that are to be included in the decision making process.

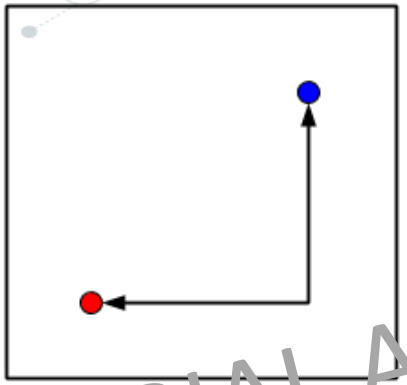
- This is the core deciding factor as the classifier output depends on the class to which the **majority** of these neighbouring points belongs.

Optimal value of K is the **square root** of the total number of samples that are present in the dataset.
Never take K = even number
Nor take K = 1 (doesn't give options to get clear majority)

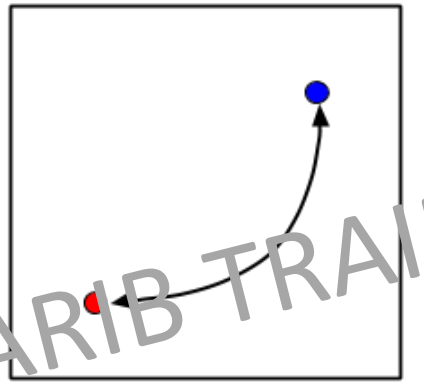
Euclidean



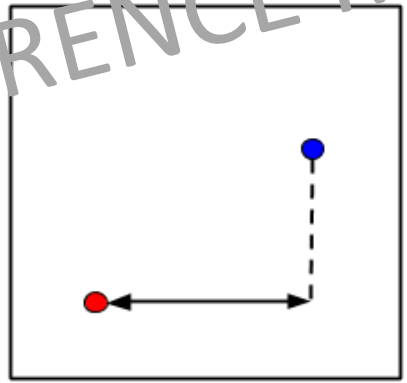
Manhattan



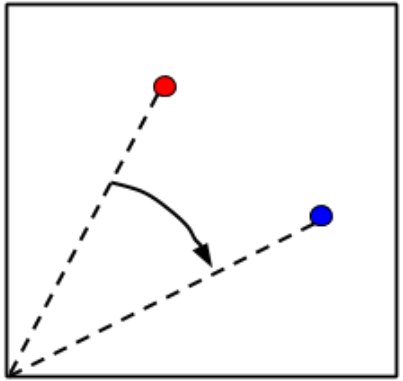
Minkowski



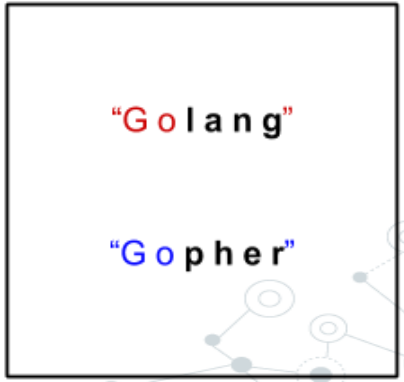
Chebychev



Cosine Similarity

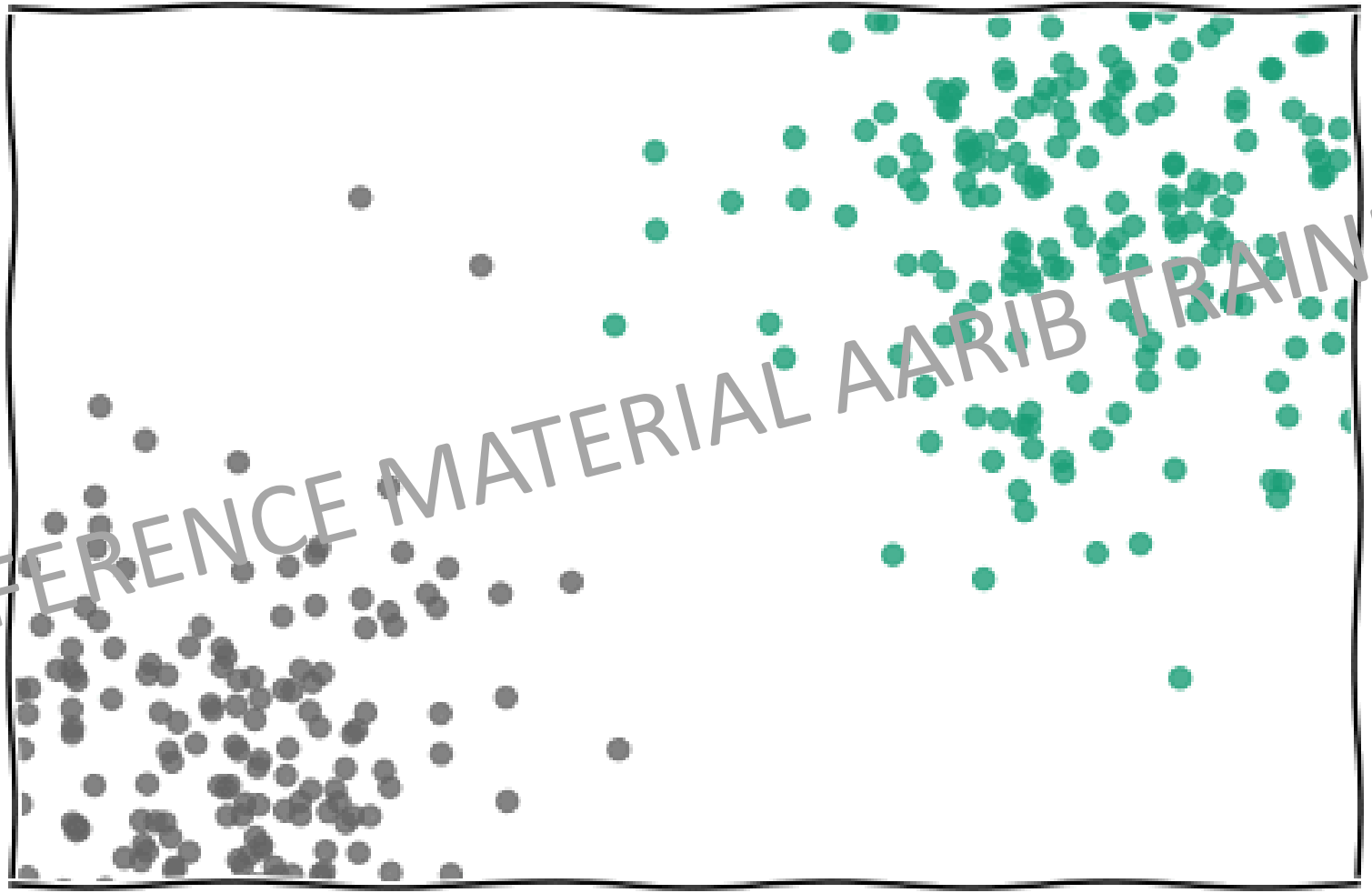


Hamming



LIMITATIONS OF KNN CLASSIFIER

- KNN is very sensitive to outliers.
- As dataset grows, the classification becomes slower
- KNN is not capable of dealing with missing values.
- It is computationally expensive due to high storage requirements.



KNN

$\Rightarrow x = (\text{Maths} = 6,\ CS = 8)$, $K = 3$

Pass

maths	Cs	Result
4	3	Fail
6	7	Pass
7	8	Pass
5	5	Fail
8	8	Pass

I $\sqrt{(6-4)^2 + (8-3)^2} = \sqrt{29} = 5.38$

II $\sqrt{(6-6)^2 + (8-7)^2} = 1$

III $\sqrt{(6-7)^2 + (8-8)^2} = 1$

IV $\sqrt{(6-5)^2 + (8-5)^2} = \sqrt{10} = 3.16$

V $\sqrt{(6-8)^2 + (8-8)^2} = 2$

Mean distance :-

$$\sqrt{|x_{01} - x_{A1}|^2 + |x_{02} - x_{A2}|^2}$$

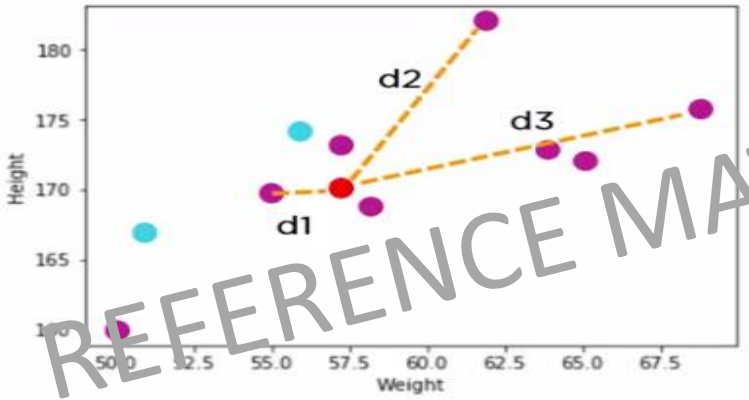
P P P $= 3P$

$3 > 0$ $\therefore$ OF

Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

57 kg	170 cm	?

Let's calculate it to understand clearly:



Unknown data point

dist(**d1**)= √(170-167)² + (57-61)² ~= 6.7

dist(**d2**)= √(170-182)² + (57-62)² ~= 13

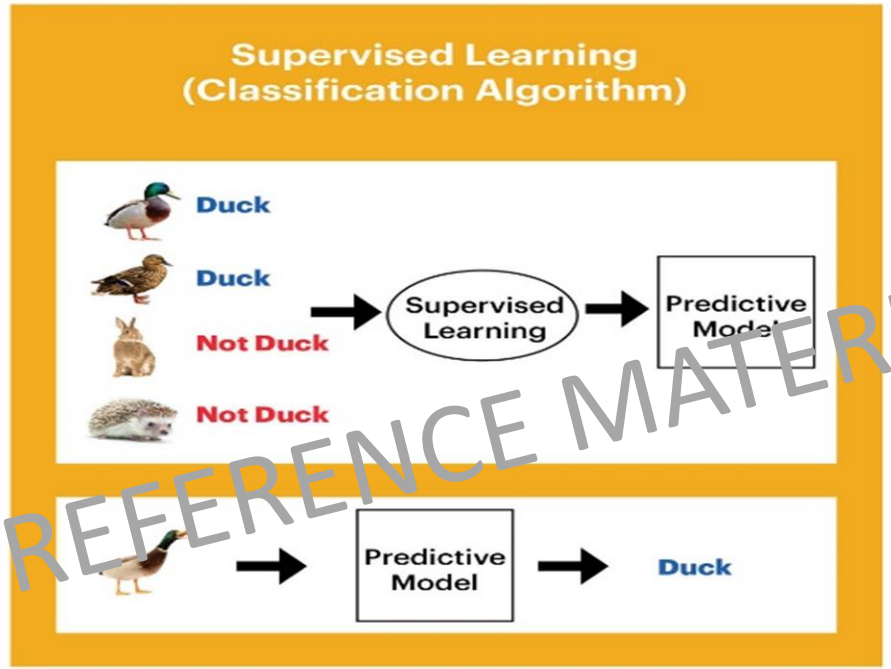
dist(**d3**)= √(170-176)² + (57-69)² ~= 13.4

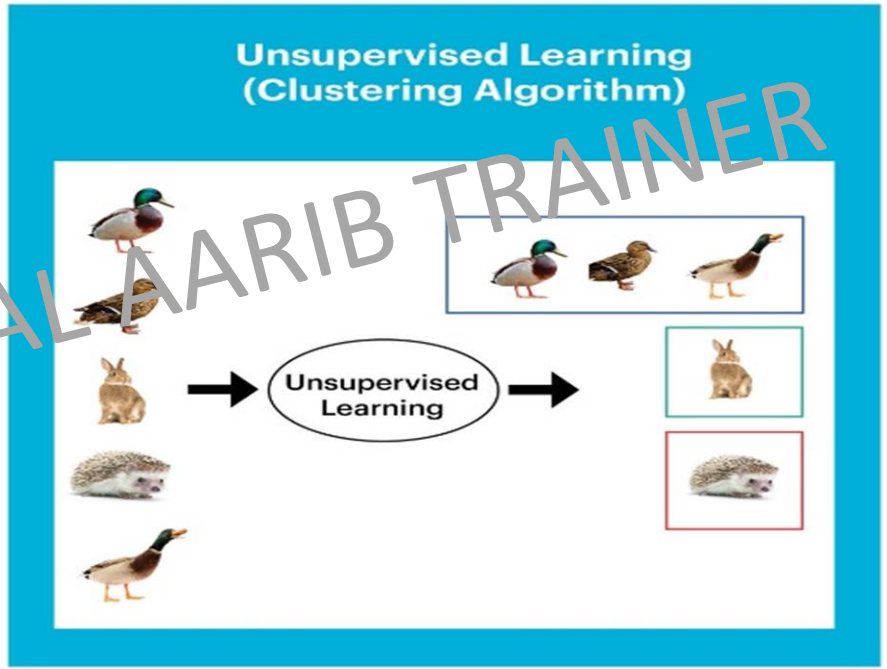
Similarly, we will calculate Euclidean distance of unknown data point from all the points in the dataset

57 kg	170 cm	?

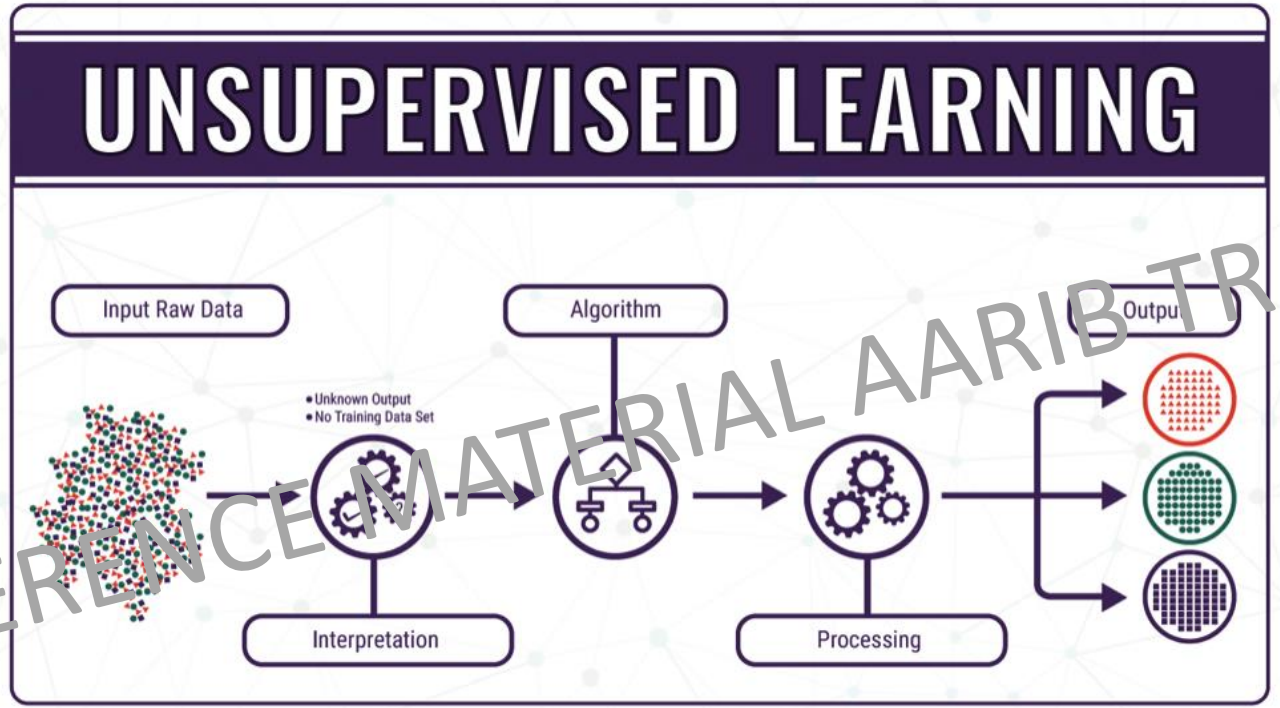
Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

K = 3





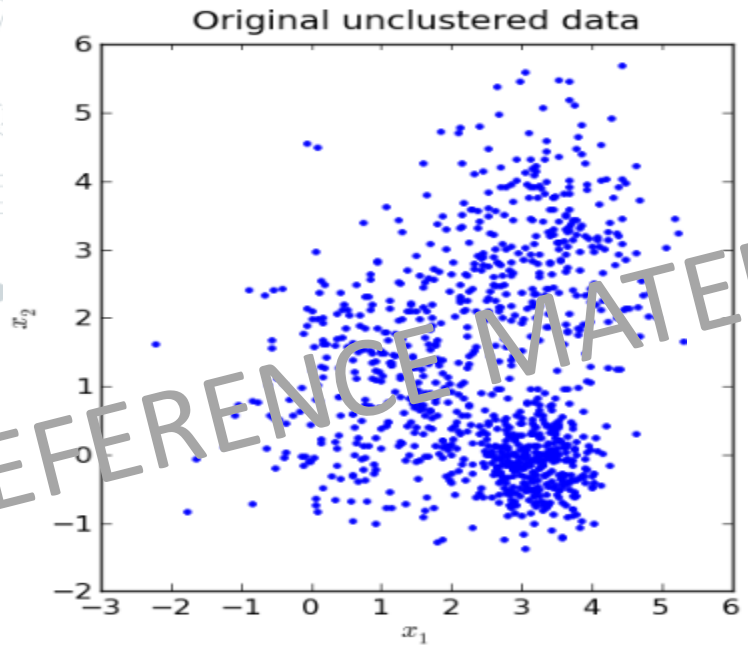
Western Digital.

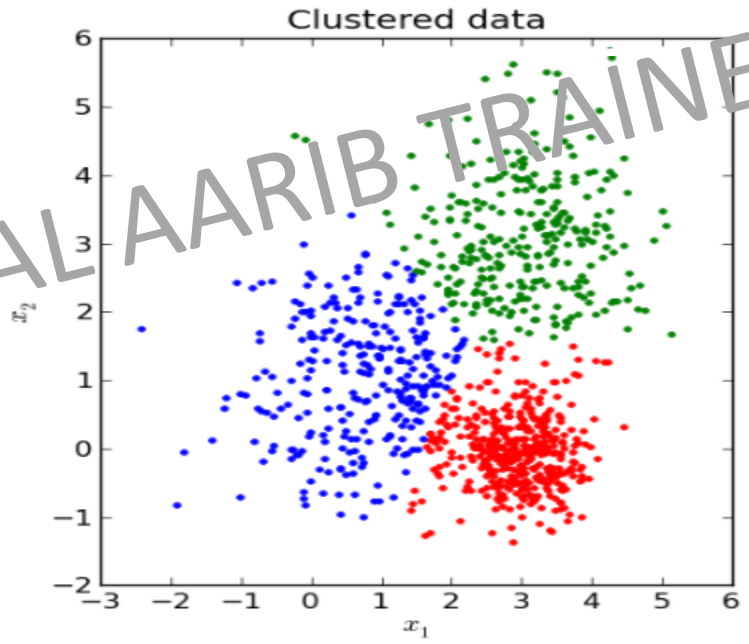


- Clustering

- Output isn’t mentioned explicitly

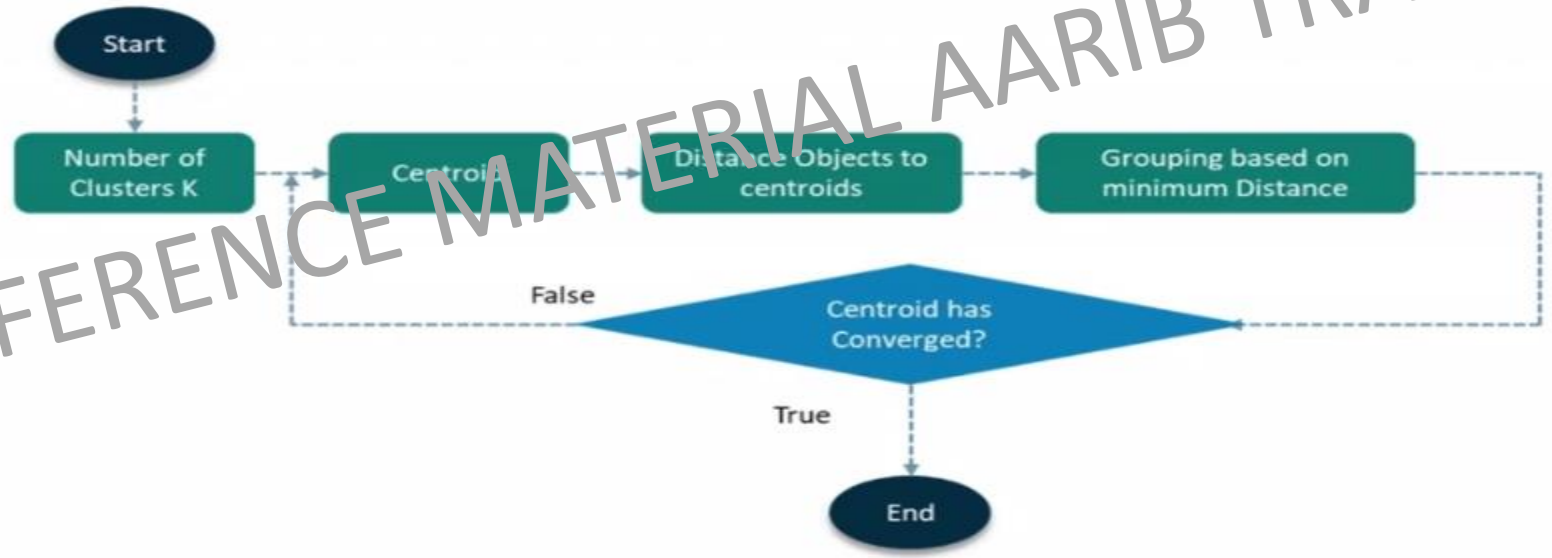
Unsupervised Learning - Clustering





Centroid based Clustering - KMeans

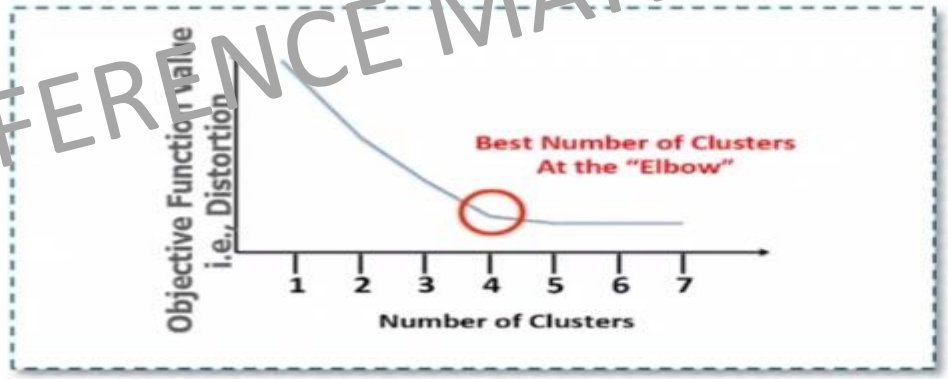
K-MEANS CLUSTERING



K-MEANS CLUSTERING

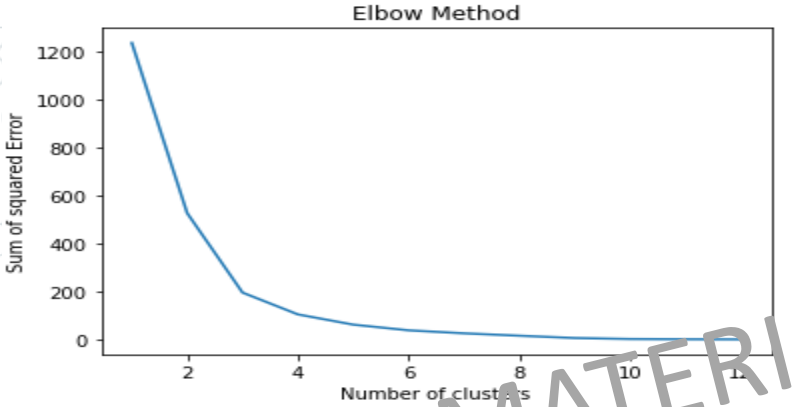
The Elbow Method:

First of all, compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid. Mathematically:



$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

SSE – Sum of Squared Errors



The formula for SSE is:

1.

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

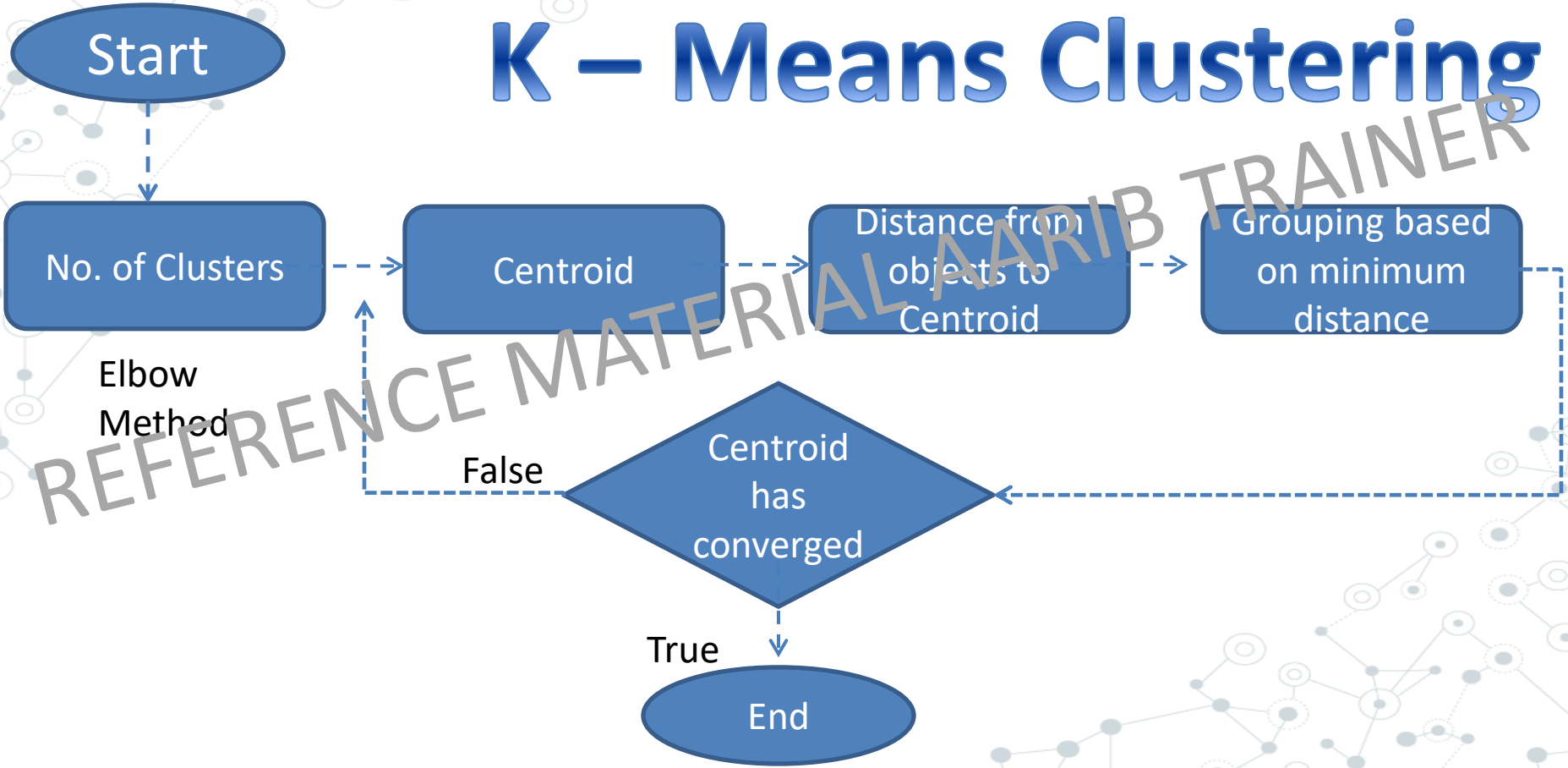
SSE is the sum of the squared differences between each observation and its group's mean.

It can be used as a measure of variation within a cluster.

No of Clusters

SSE

K – Means Clustering



Euclidean Distance

	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

$$d1 = \sqrt{(168-185)^2 + (60-72)^2}$$

$$d1 = 20.8$$

$$d2 = \sqrt{(168-170)^2 + (60-56)^2}$$

$$d2 = 4.48$$

CENTROIDS

CLUSTER 0 (K1)	185	72
CLUSTER 1 (K2)	169	58

$$K2 = \left(\frac{170+168}{2}, \frac{60+56}{2}\right)$$

K2 = (169,58)

Points in Clusters

CLUSTER 0	1,4,5,6,7,8,9, 10,11,12
CLUSTER 1	2,3

REFERENCE LINKS FOR DEMO AND WORKING

https://scistatcalc.blogspot.com/2014/01/k-means-clustering-calculator.html

https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/

https://stanford.edu/class/engr108/visualizations/kmeans/kmeans.html

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/