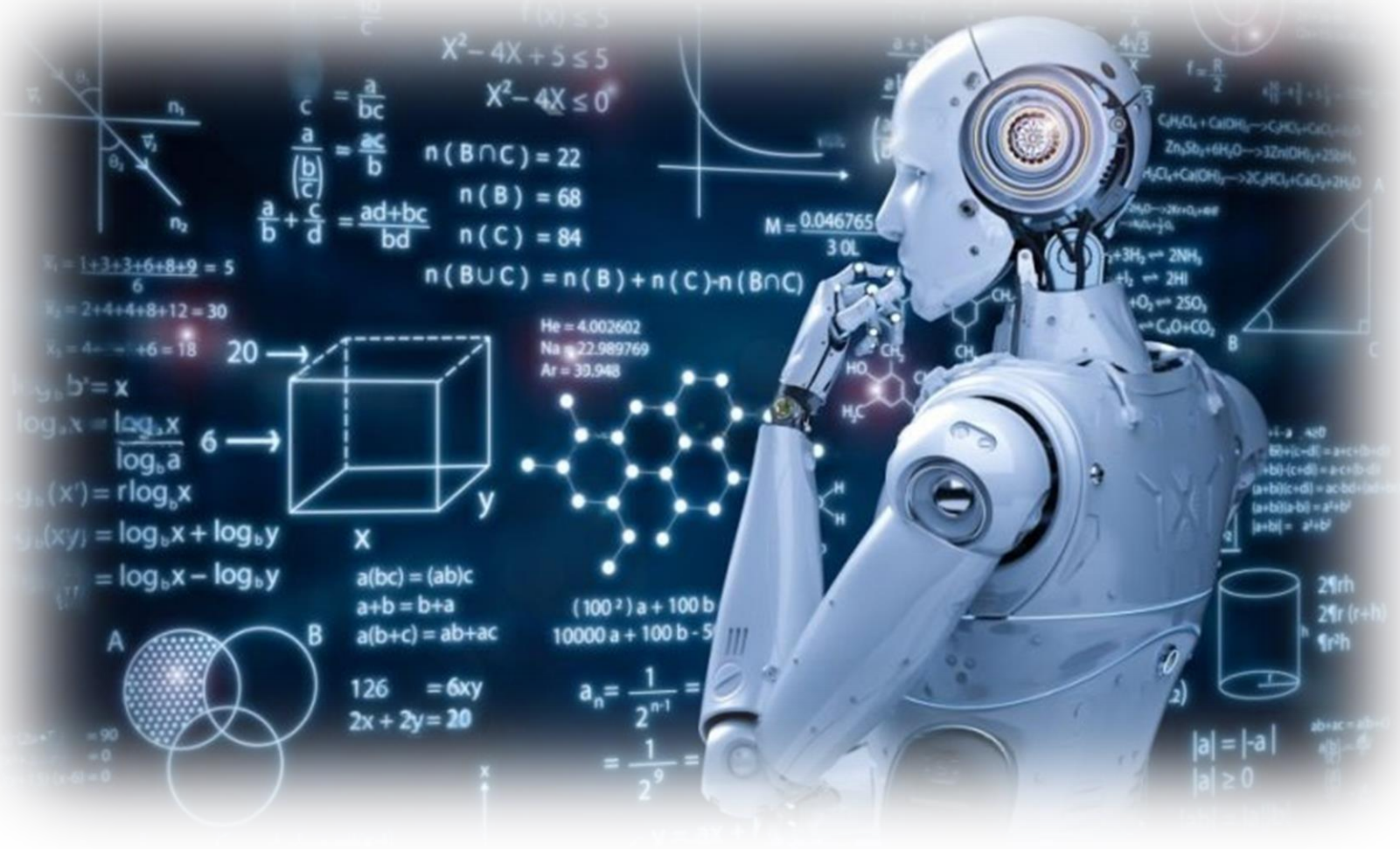
Unit-5: Big Data Application

2CEIT702: BIG DATA ANALYTICS



Difference between

Conventional / Traditional / Normal Programming

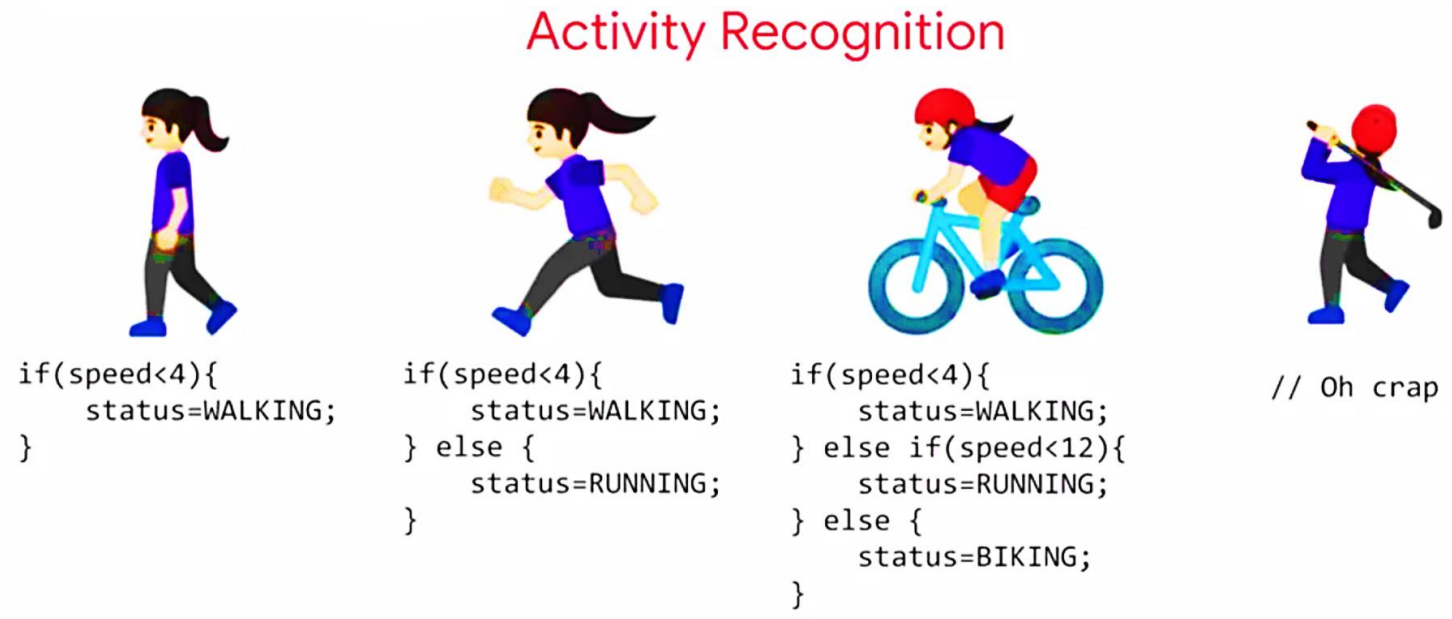
Vs

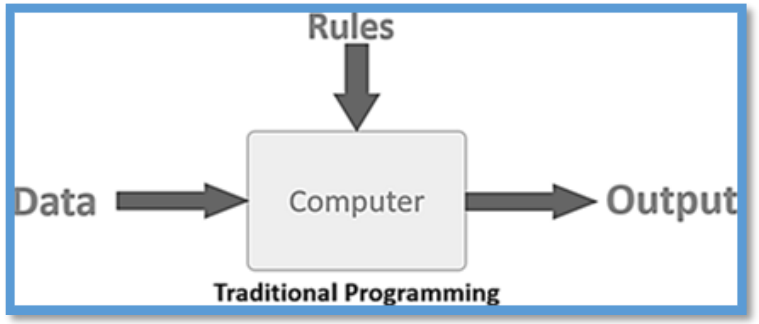
Machine Learning

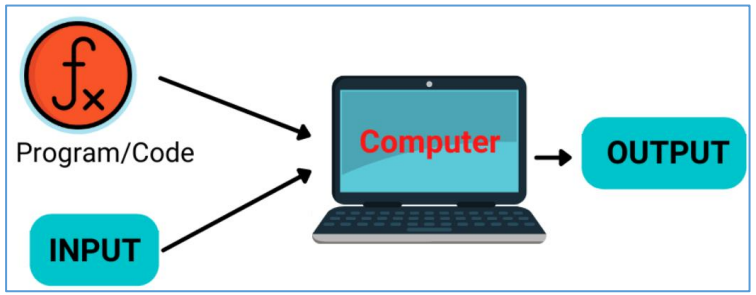
	Traditional programming	Machine Learning
Definition	Just writing the code and giving instructions to a computer to perform specific tasks	Uses data-driven approach, typically trained on historical data and used to make predictions on new data.
Approach	developers write explicit instructions for the computer to follow. The solution is typically handcrafted and designed to handle specific scenarios.	developers create models which learn from data. Instead of explicitly programming the solution, feed the model with data to train it and learn patterns within the data to make predictions.
Data Dependency	generally doesn't depend on large amounts of data. The logic and rules are explicitly provided in the code.	algorithms heavily depend on data for training. The quality and quantity of data play a crucial role in the performance of the model.
Adaptability	typically static and don't adapt or improve on their own. Changes or improvements require manual intervention and reprogramming.	Designed to be adaptive. They can continuously learn from new data and improve their performance over time without the need for manual reprogramming.
Application	used for a wide range of tasks, including software development, web development, system administration, and more.	Used in fields like, image and speech recognition, natural language processing, recommendation systems, autonomous vehicles, and more.
Skillset	Programming languages, algorithms, and data structures are essential for programming tasks.	In addition to programming skills, machine learning requires knowledge of statistics, linear algebra, and ML algorithms and frameworks.

Conventional / Traditional / Normal Programming Vs Machine Learning

- **Disadvantage of Traditional Programming:** When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain. The expert's knowledge may not be comprehensive enough to provide accurate rules.

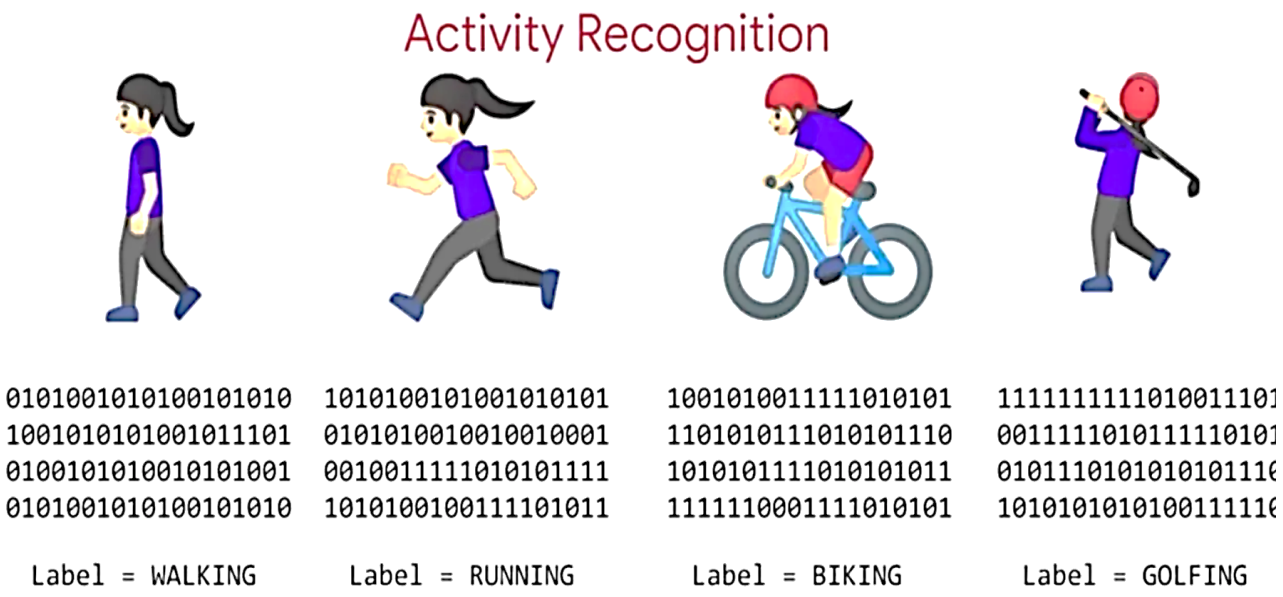


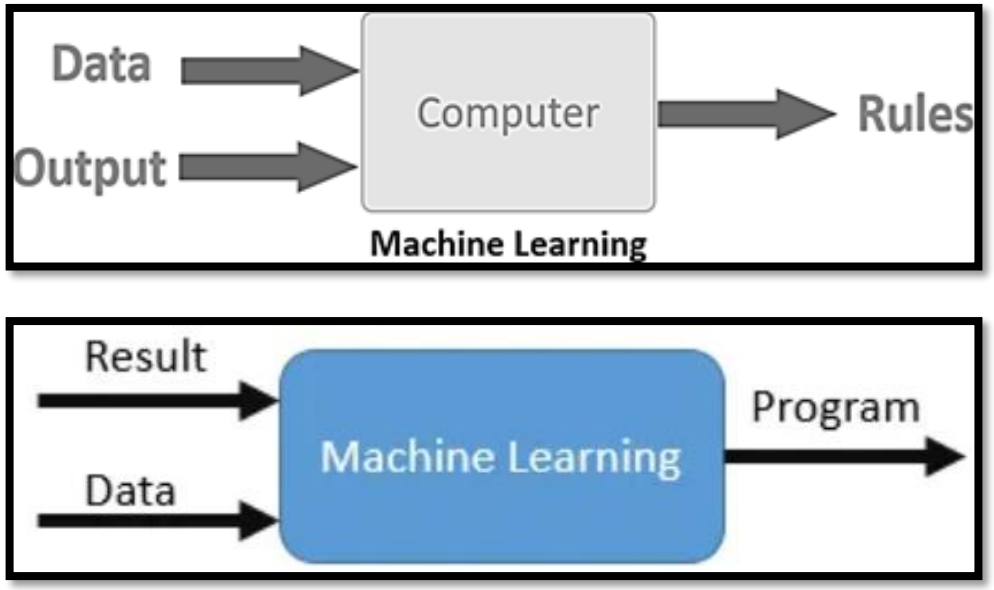


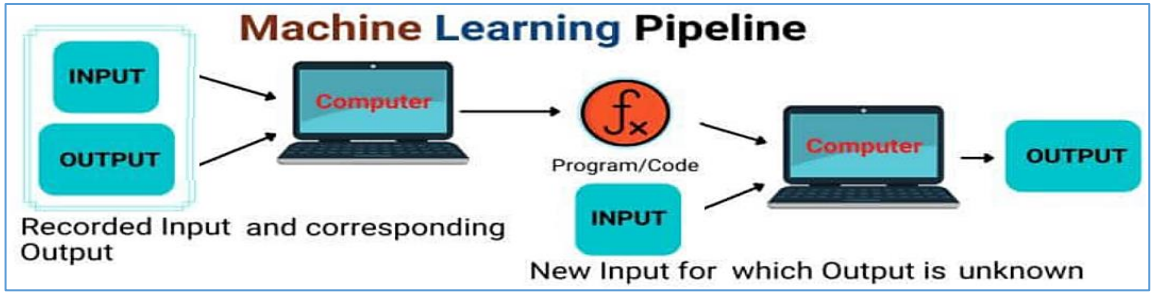


Conventional / Traditional / Normal Programming Vs Machine Learning

- **Machine Learning** is a system that can learn from example to produce accurate results through self-improvement and without being explicitly coded by programmer.



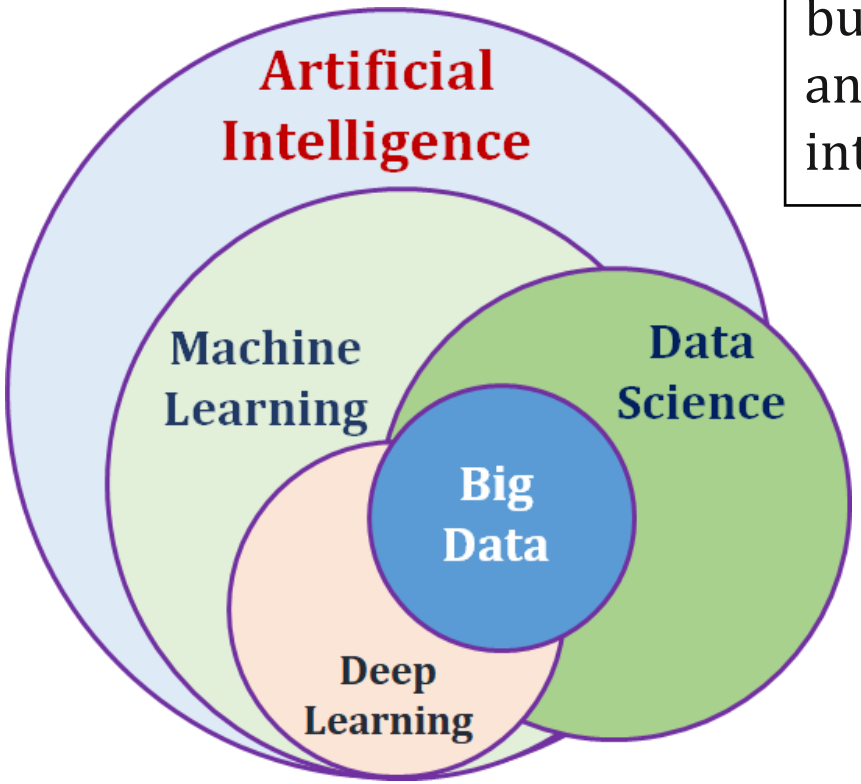




Relation of AI, ML, DL, Big Data and Data Science

Artificial Intelligence involves making the machine as much capable, So that it can perform the tasks that typically require human intelligence.

Machine Learning is a subset of AI that focus on learning from data to develop an algorithm that can be used to make a prediction.

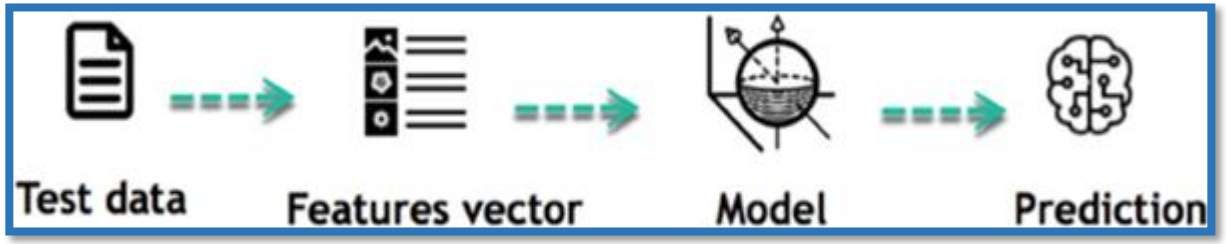


Data Science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, and artificial intelligence to analyze large amounts of data.

Deep Learning is a subset of Machine Learning that involves the use of neural networks to model and solve complex problems.

Machine Learning Terminology

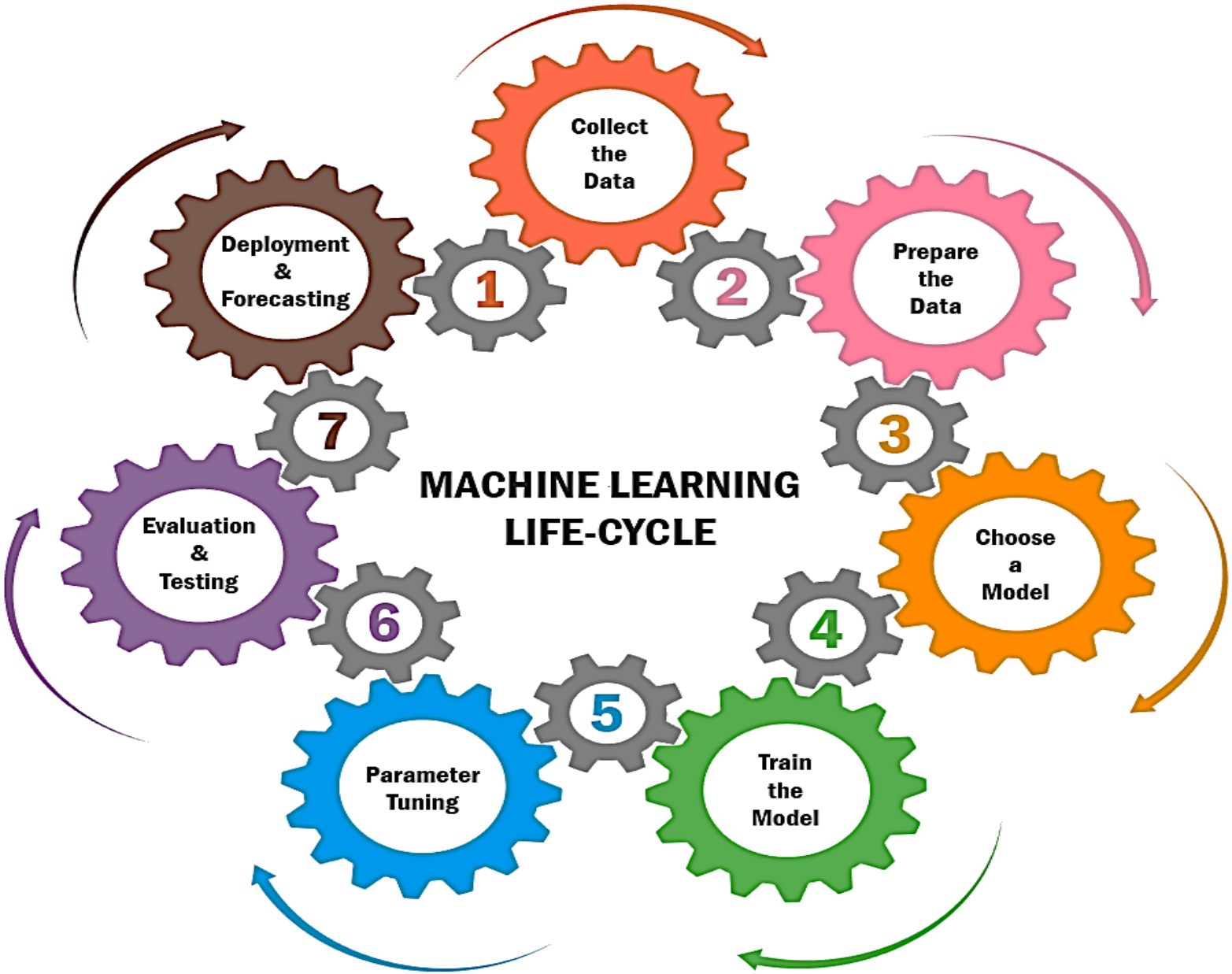
- **Feature vector:** List of attributes used to solve a problem. E.g. Gender and Age.
- **Label:** The target we want to predict. E.g. Willing to buy or not, True / False.
- **Model:** The rules used to predict.
- **Algorithm:** The program that is used to generate the model.
- The process to generate a model a called **Training**, and the set of data used in this process is called **training data**.
- The process to use a trained model to predict is called **Testing or Inference**, and the data used in this process is called **Test data.**
- **Sample:** A single data instance is called a**.**



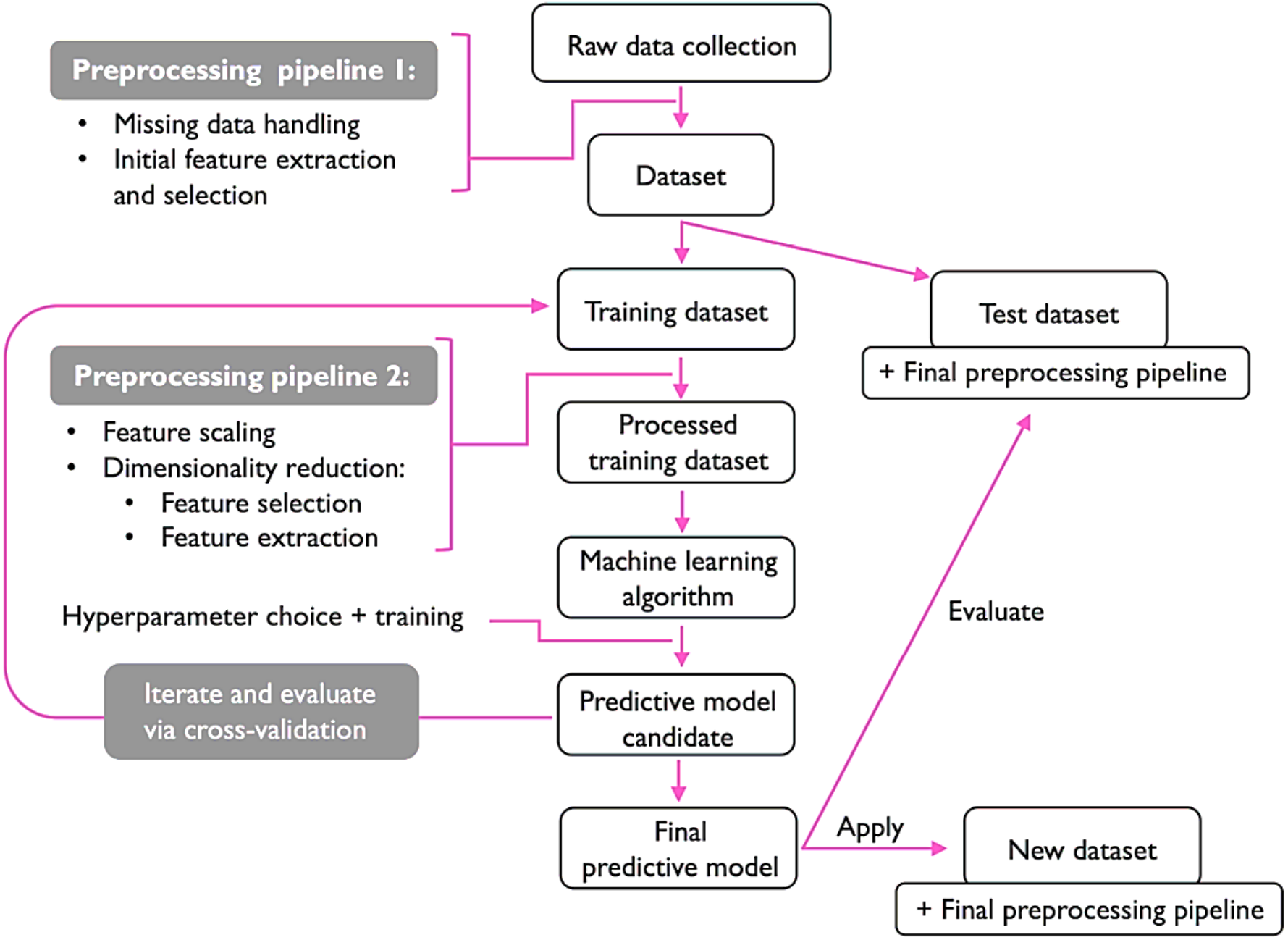
Machine Learning Process

- **Data collection:** Identify the data sources and collect the quality and quantity data. The more will be the data, the more accurate will be the prediction.
- **Data preparation:** Data preparation is defined as a gathering, combining, cleaning, and transforming raw data to make accurate predictions in Machine learning projects. Also known as data "pre-processing," "data cleaning," and "feature engineering.“
- **Choose Model:** It is the process of choosing the best ML model for a given task.
- **Train the Model:** Use datasets to train the model using various machine learning algorithms.
- **Test the Model:** Once ML model has been trained on a given dataset, then test the model. In this step, we check the accuracy of model by providing a test dataset to it.
- **Model Evaluation:** It is a process of assessing the model’s performance on a chosen evaluation setup. It is done by calculating performance metrics like F1-Score, RMSE etc.
- **Deployment:** Deploy the model in the real-world system.

Machine Learning Process



ML Process


Raw data collection
Dataset
Preprocessing pipeline 1:
• Missing data handling
• Initial feature extraction and selection
Training dataset
Test dataset
+ Final preprocessing pipeline
Preprocessing pipeline 2:
• Feature scaling
• Dimensionality reduction:
• Feature selection
• Feature extraction
Processed training dataset
Machine learning algorithm
Hyperparameter choice + training
Iterate and evaluate via cross-validation
Predictive model candidate
Evaluate
Final predictive model
Apply
New dataset
+ Final preprocessing pipeline

Types of Machine Learning Algorithms

TYPES OF MACHINE LEARNING

Supervised Learning

Training with Labeled Data
(input + output)

(Learning by examples)

Classification

Disease detection
Email spam detection
Image classification
Bank loan prediction

Regression

House price prediction
Stock price prediction

Unsupervised Learning

Training with Unlabeled Data

(Learning by observation)

Clustering

Search engines
Face recognition
Targetted marketing
Recommender system

Association Rule Mining

Market basket analysis
Medical diagnosis
Census data

Semi-Supervised Learning

Combination of Labeled and Unlabeled Data

Reinforcement Learning

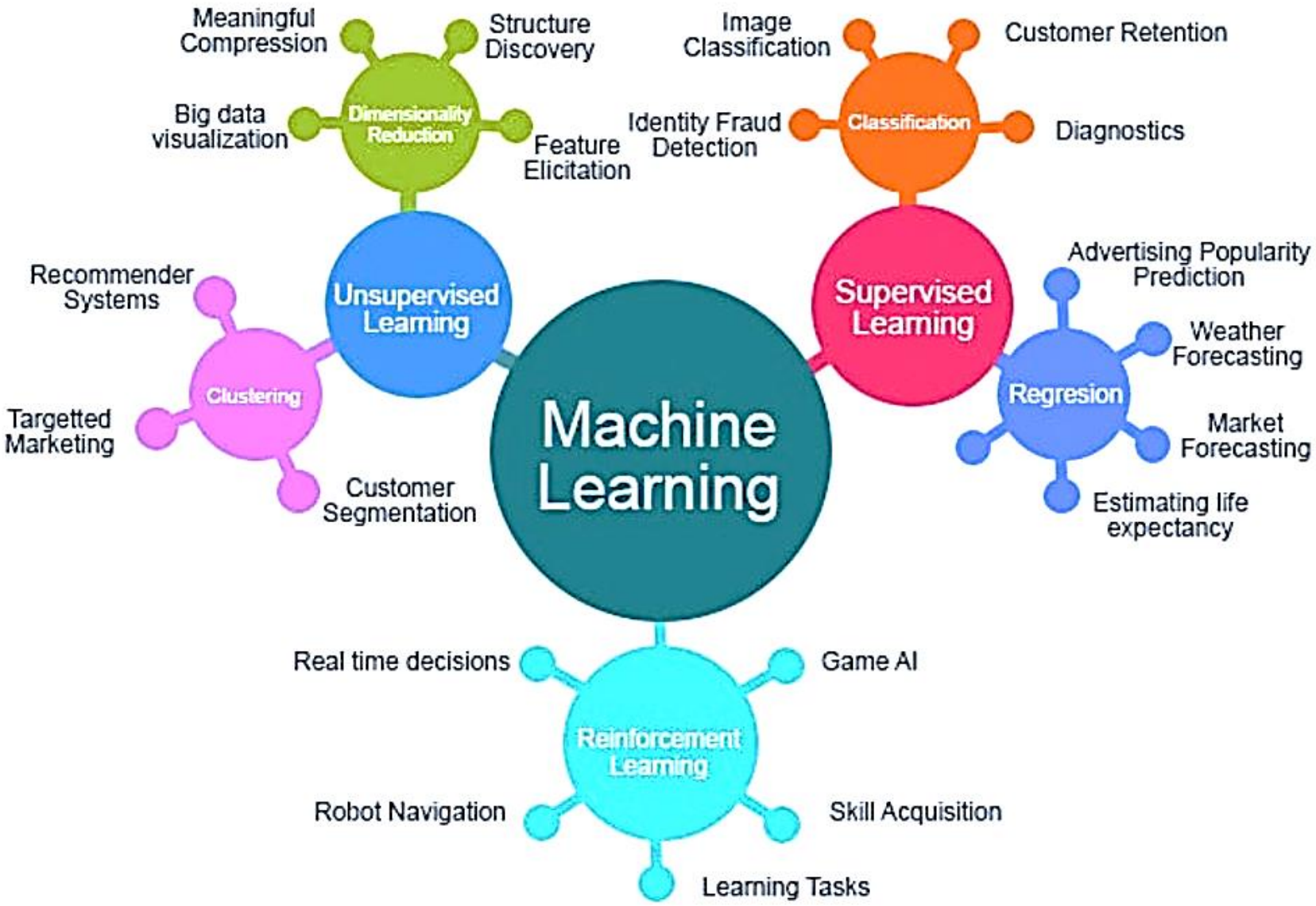
Learning through Interaction with the Environment and Feedback
(Reward/Punishment)

(Learning from mistakes)

Example

Self-driving car
Gaming AI
Robot navigation
Inventory management
Finance sector

Types of Machine Learning Algorithms



Machine Learning Algorithms- Supervised Learning

- **Definition:** Supervised learning is a type of machine learning where the algorithm or model is trained on a labeled dataset.

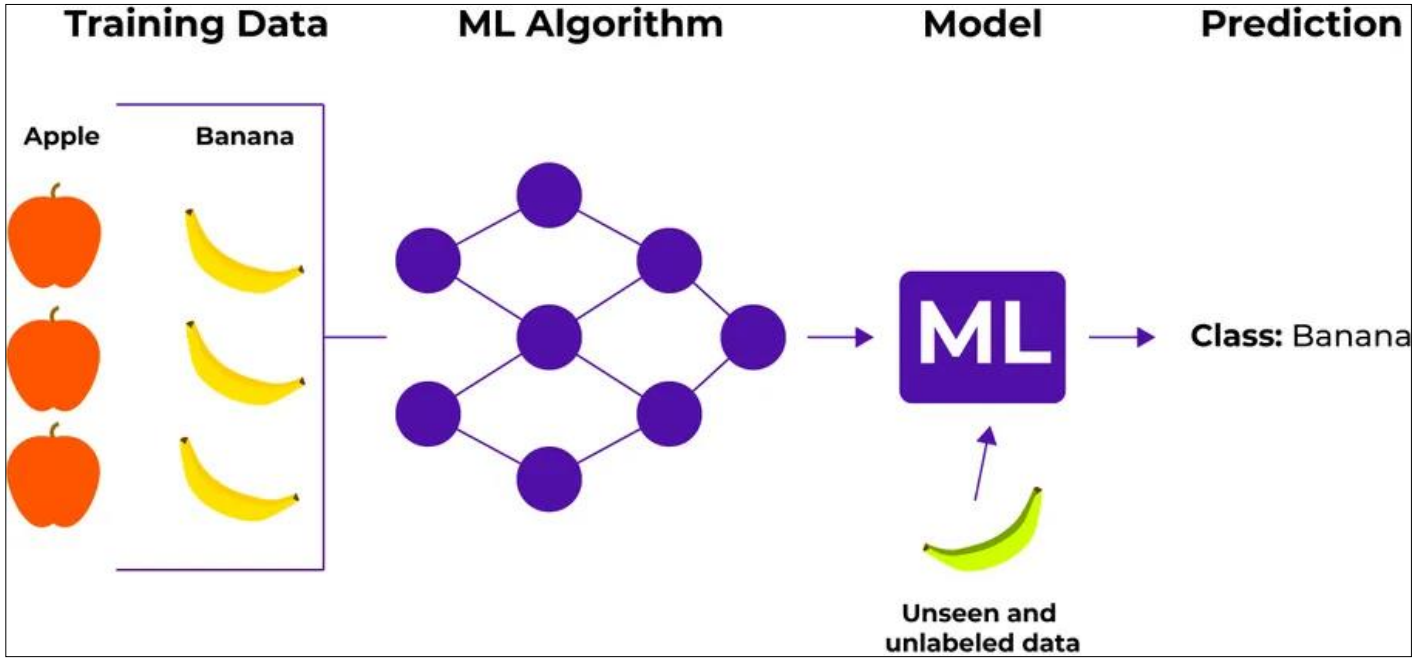
- Once the training and processing are done, the model is tested by providing a sample test data (or new data) to check whether it predicts the correct output.

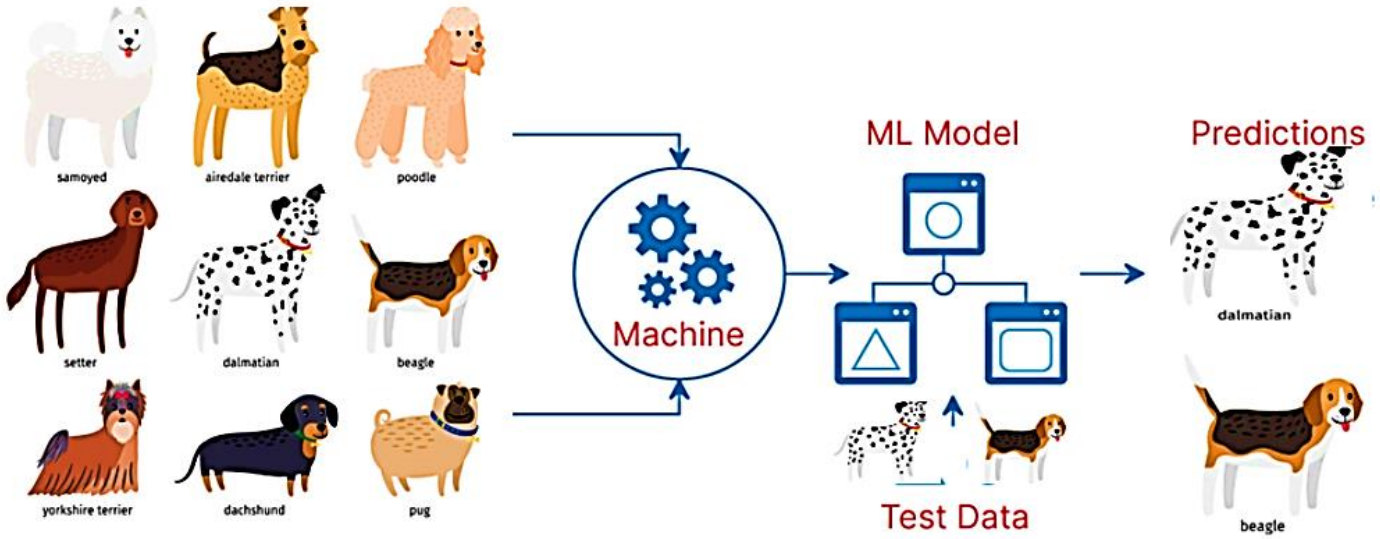
- The algorithm tries to learn the relationship between the input and output data so that it can make accurate predictions on new, unseen data.

- Believe it or not, over ninety percent of machine learning used in real-world business technology today is actually using supervised learning.

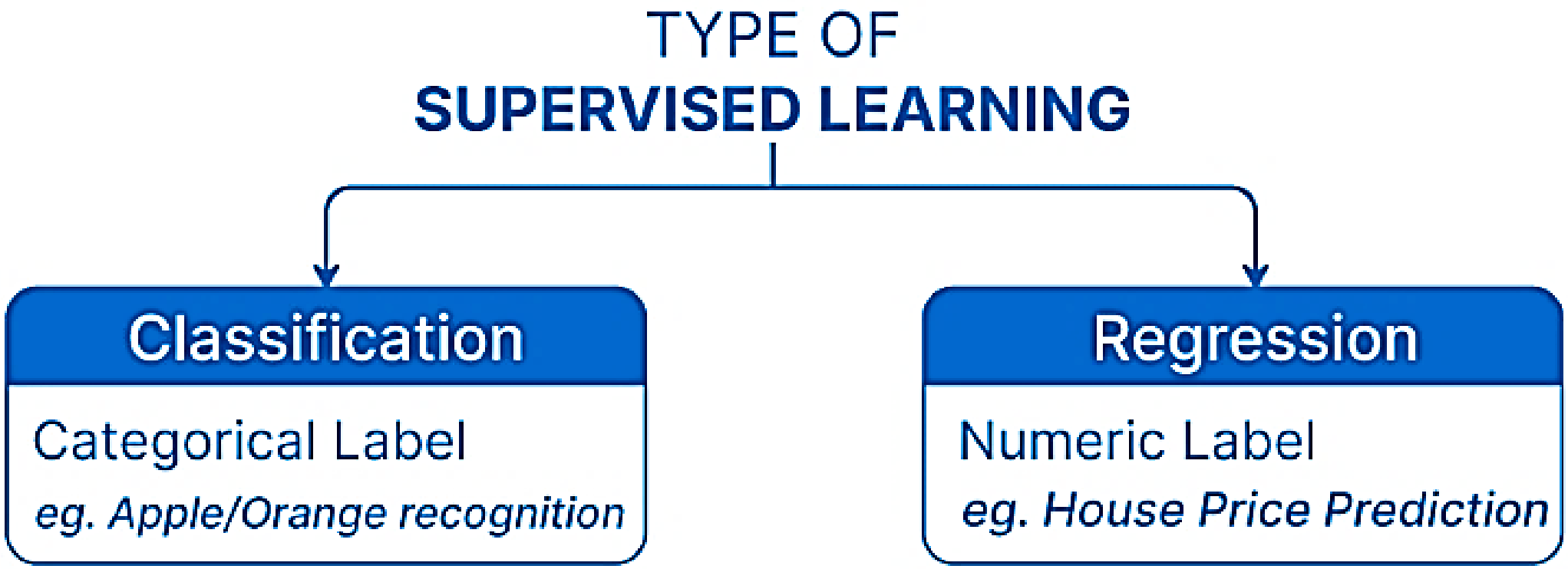
- Field: Finance, Healthcare, Marketing, and more.
- Example: Email Filtering, Credit Scoring, Image Classification

Machine Learning Algorithms- Supervised Learning



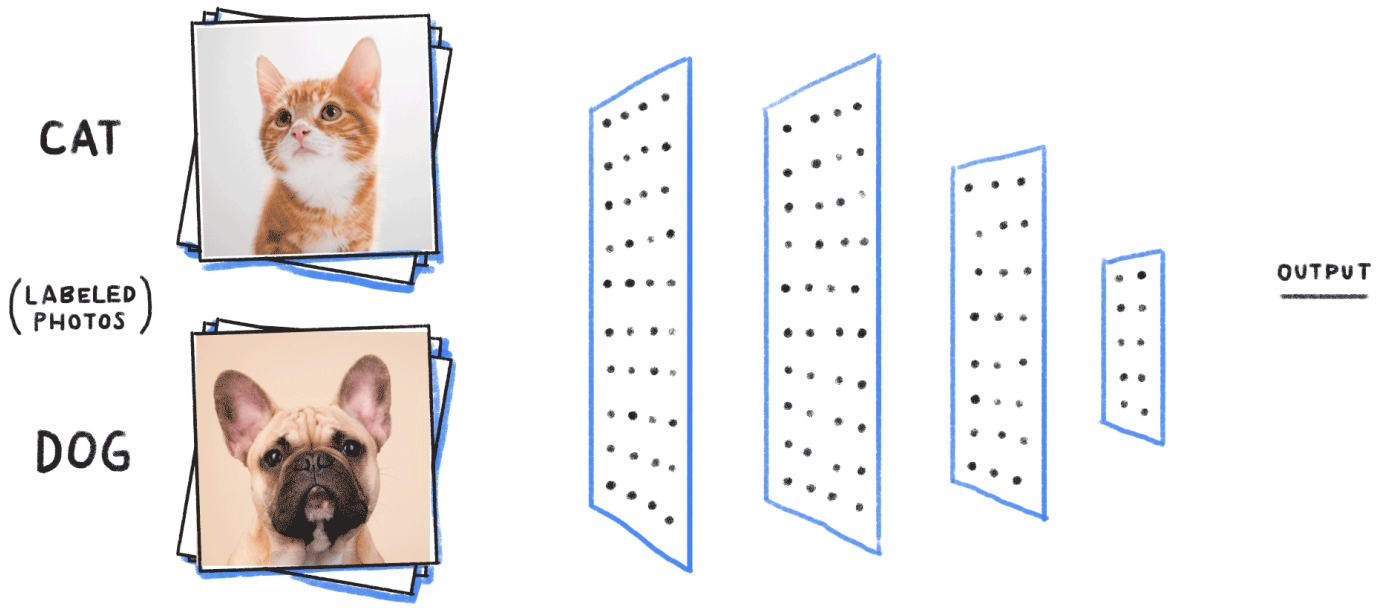


Types of Supervised Learning

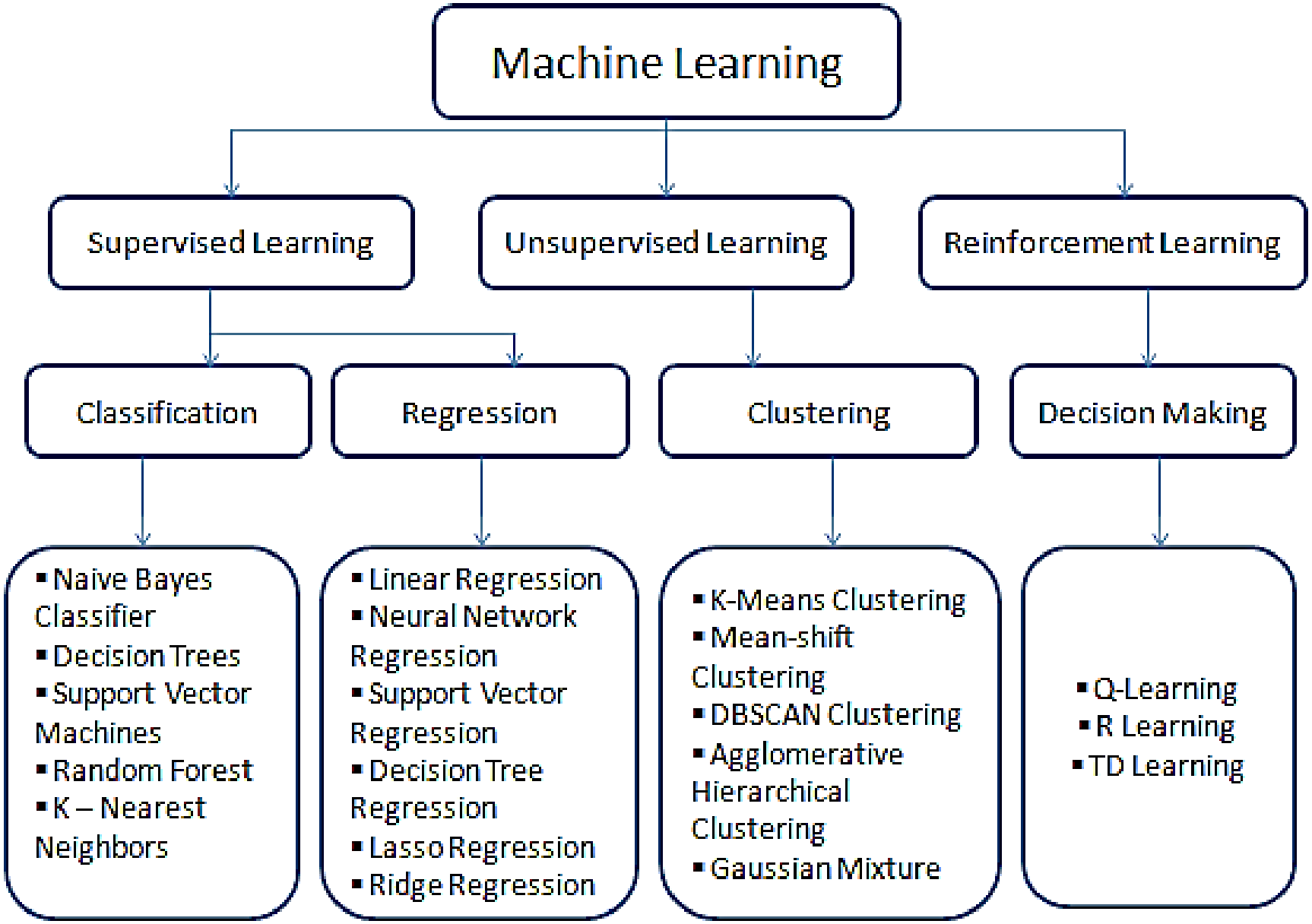


Classification- Supervised Learning Algorithm

- A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”, “Cat” or “Dog”, “Yes” or “No”, “Male” or “Female”.
- Classification is a type of machine learning task that involves predicting a discrete label based on input data.
- So in a simple way, supervised learning involves an input and output. The image is the input and the label is the output. “cat” or “dog”.



ML Algorithms



Supervised Learning – Linear Regression

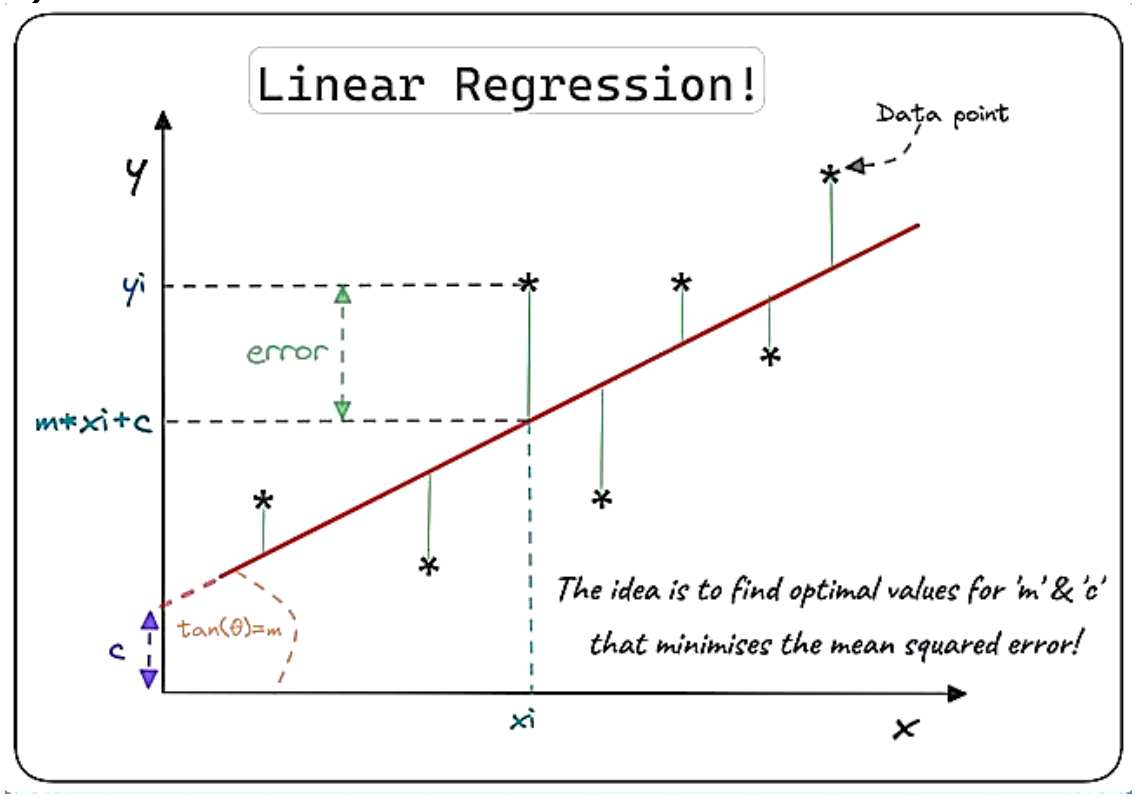
- **Regression:** It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.
- **Linear Regression**
  - Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent variables (features) by fitting a linear equation to observed data.
  - When there is only one independent variable, it is known as **Simple Linear Regression**, and when there are more than one independent variable, it is known as **Multiple Linear Regression**.
  - Similarly, when there is only one dependent variable, it is considered **Univariate Linear Regression**, while when there are more than one dependent variables, it is known as **Multivariate Regression**.

Supervised Learning - Regression

- Example:
  - Predicting the price of a house based on its size, location, and other features.
  - Predicting the demand for a product based on historical sales data
  - Forecasting the stock price of a company based on financial data
  - Predicting the likelihood of a customer defaulting on a loan based on their credit history
  - Estimating the life expectancy of a patient based on their medical history and other factors
  - Predicting the fuel efficiency of a car based on its engine size and other features
  - Determining how much a customer is willing to pay for a particular product based on age.

Supervised Learning – Linear Regression

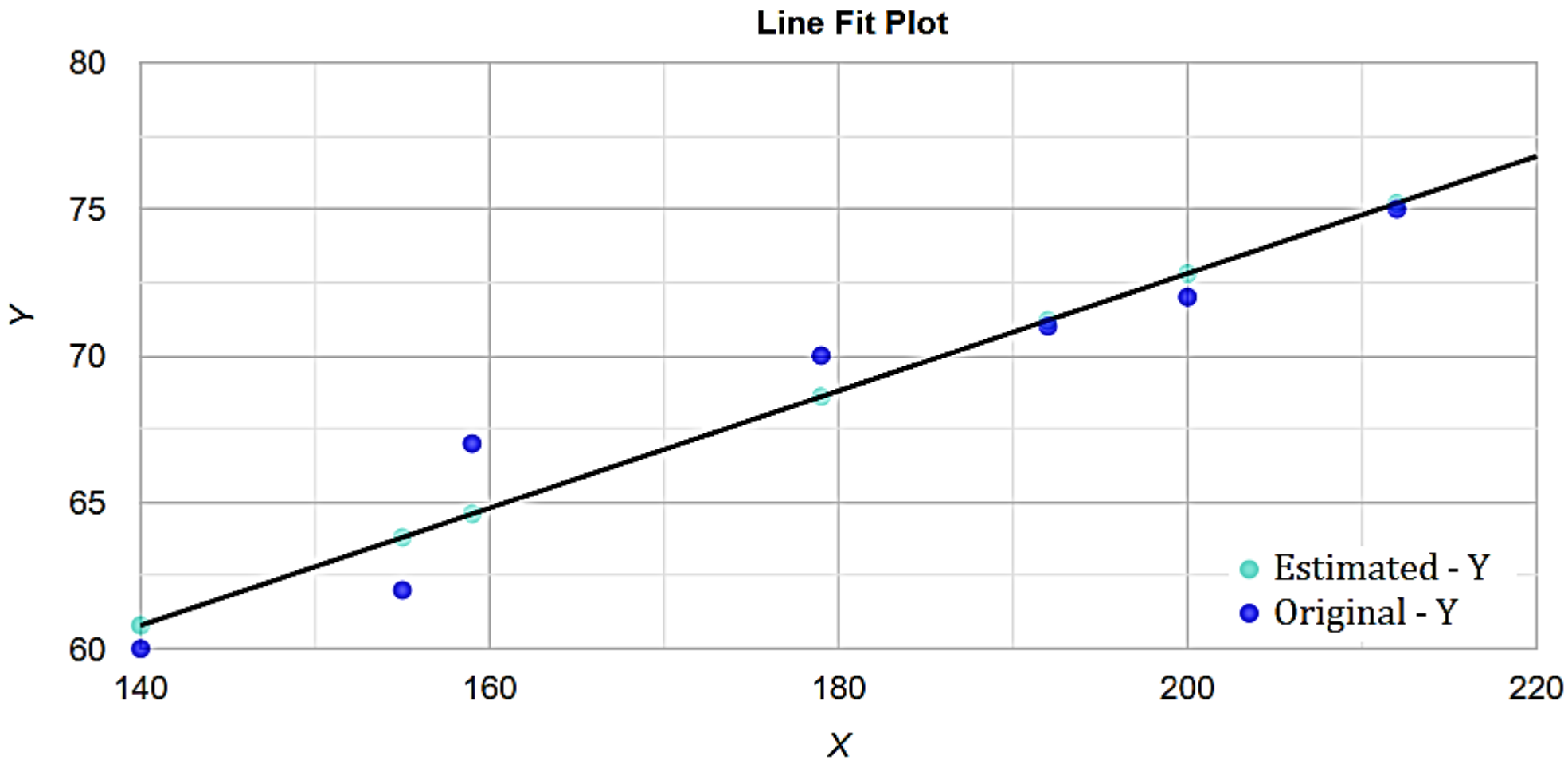
- Equation of Linear Regression is: **y = mx + b** (or y = mx + c)
- **y =** Dependent Variable
- **x =** Independent Variable
- **m =** slope of the line (Regression line) (How much “y” changes for unit change in “x”)
- **b =** intercept (it show the value of “y” when x=0)



Supervised Learning – Linear Regression Example

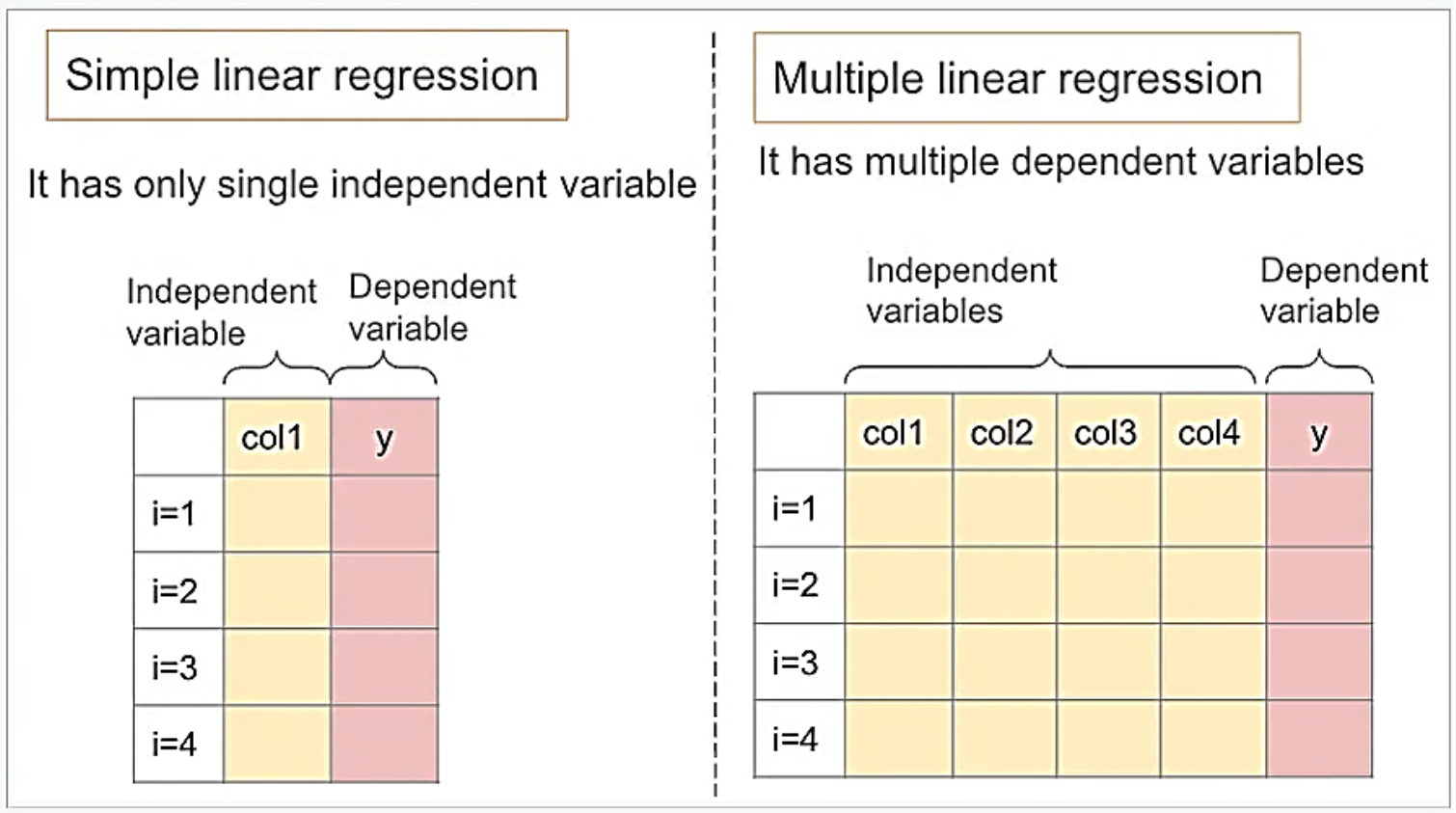
Weight(LBS) X (Independent)	Hight(Inches) Y (Dependent)	Mean (X)	Mean (Y)	Deviations (X)	Deviations (Y)	Product of Deviations	Sum of product of Deviations	Square of Deviations of X	y = mx + b
140	60	176.71	68.14	-36.71	-8.14	298.82	832.29	1347.62	60.7943
155	62			-21.71	-6.14	133.3		471.32	63.7958
159	67			-17.71	-1.14	20.19		313.64	64.5962
179	70			2.29	1.86	4.26		5.24	68.5982
192	71			15.29	2.86	43.73		233.78	71.1995
200	72			23.29	3.86	89.9		542.42	72.8003
212	75			35.29	6.86	242.09		1245.38	75.2015
								4159.4	
y = mx + b									
Calculate **m** = (Sum of product of Deviations) / (Sum of Square of Devia						**0.2001**			
Calculate **b** = (Mean of Y) - (m * Mean of X)						**32.7803**			

- Linear Regression: $Y = 0.2001*X + 32.78$



Multiple Linear Regression

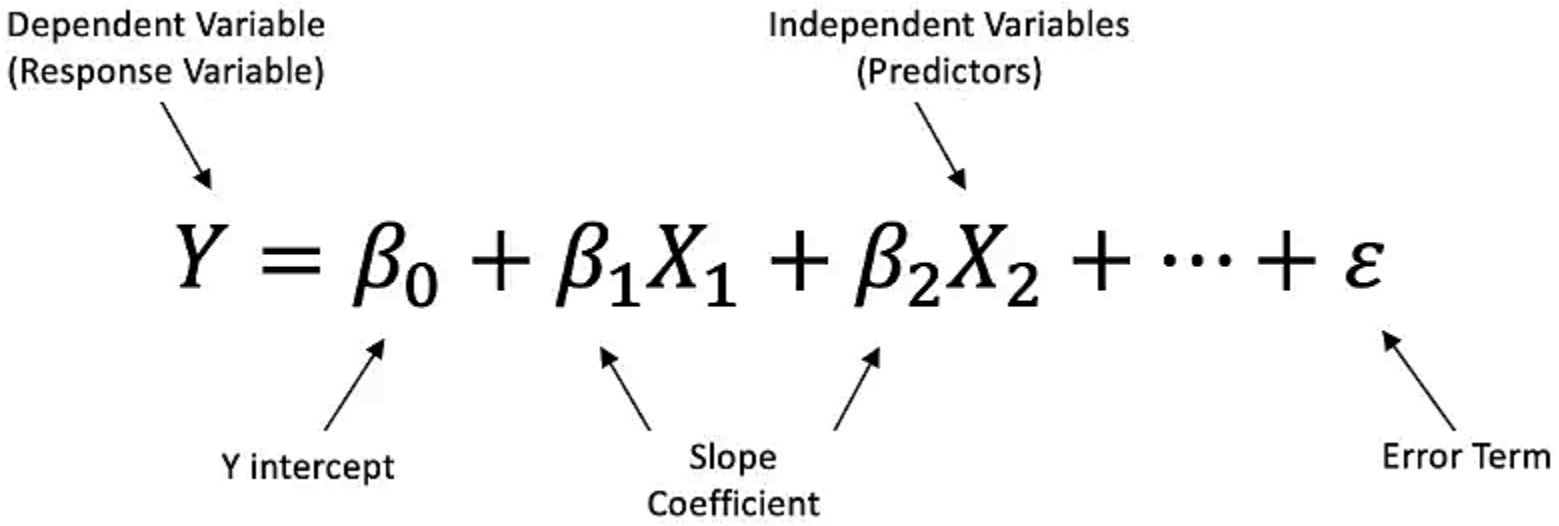
- Multiple linear regression is the most common and most important form of regression analysis and is used to predict the outcome of a variable based on two or more independent variables.



- Example: Housing prices (i.e. Type, Location, View, Neighborhood, Area).

Multiple Linear Regression

- Equation of Multiple Linear Regression



Multiple Linear Regression – Example (Using Matrix Approch)

- Equation of Multiple Linear Regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- Dataset:

X_1	X_2	Y
1	5	2
2	6	7
3	7	9
4	3	12

- Matrix and vector are defined: $\mathbf{X} = \begin{bmatrix} 1 & 1 & 5 \\ 1 & 2 & 6 \\ 1 & 3 & 7 \\ 1 & 4 & 3 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 2 \\ 7 \\ 9 \\ 12 \end{bmatrix}$
- Now, the vector with the estimated regression coefficients **β** (β_0, β_1, β_2) is computed through the following matrix operation:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- (Where, **X'** = Transpose Matrix X, $\mathbf{(X'X)^{-1}}$ = Inverse matrix of X'X))

Multiple Linear Regression – Example (Using Matrix Approch)

- Estimate regression coefficients is computed as follows:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 5 \\ 1 & 2 & 6 \\ 1 & 3 & 7 \\ 1 & 4 & 3 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 7 \\ 9 \\ 12 \end{bmatrix} \right) = \begin{bmatrix} 4 & 10 & 21 \\ 10 & 30 & 50 \\ 21 & 50 & 119 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 30 \\ 91 \\ 151 \end{bmatrix}$$

$$= \begin{bmatrix} 7.133333 & -0.933333 & -0.866667 \\ -0.933333 & 0.233333 & 0.066667 \\ -0.866667 & 0.066667 & 0.133333 \end{bmatrix} \cdot \begin{bmatrix} 30 \\ 91 \\ 151 \end{bmatrix} = \begin{bmatrix} -1.8 \\ 3.3 \\ 0.2 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

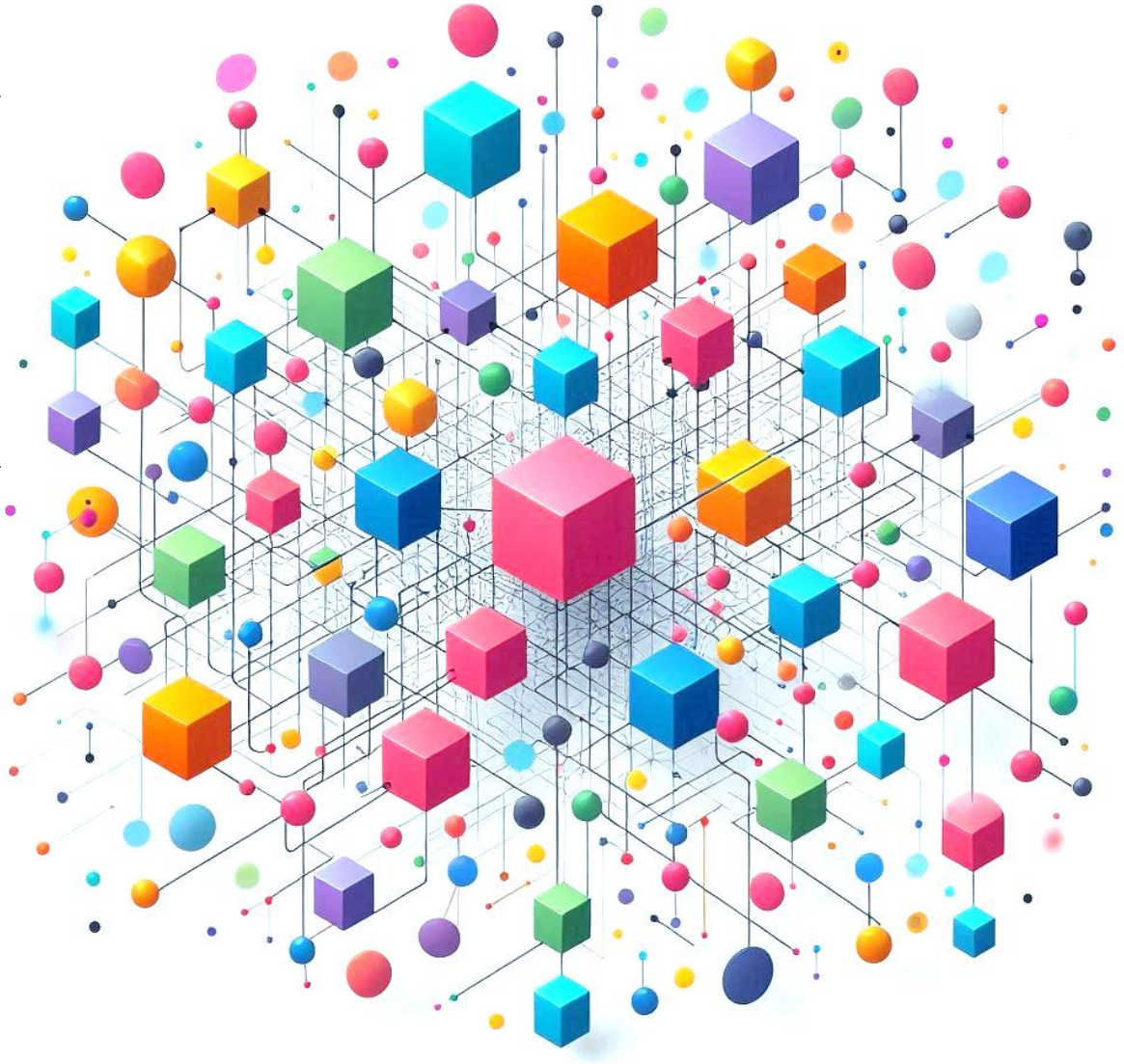
Based on the data provided, the estimated multiple linear regression equation is:

$$Y = -1.8 + 3.3X_1 + 0.2X_2$$

K-Means Algorithm (Unsupervised Learning)

Unsupervised learning, uses machine learning (ML) algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction.



K-Means Algorithm (Unsupervised Learning)

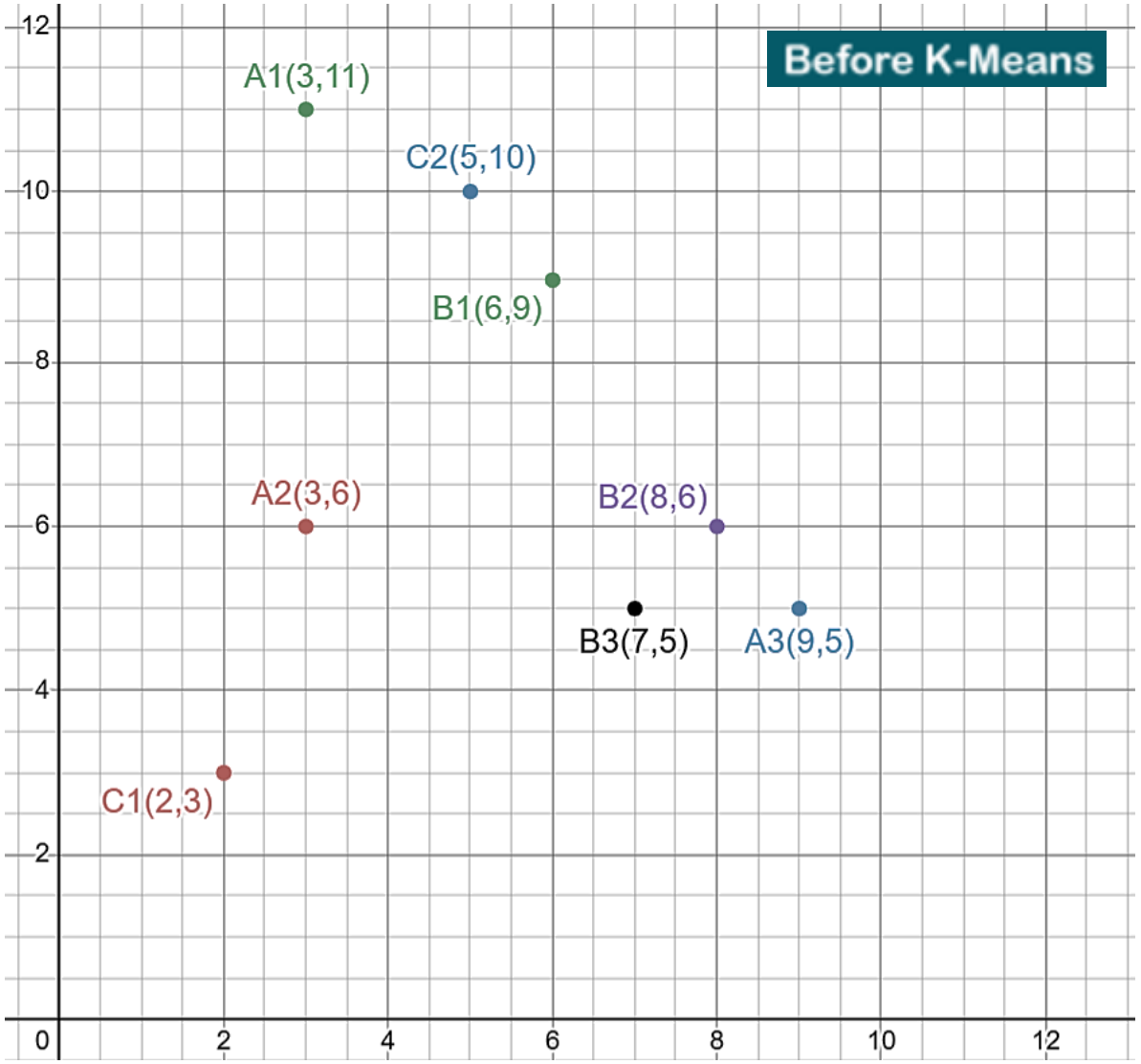
- K-means clustering is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centroid. (Note: Exclusive clustering is as the name suggests and stipulates that each data object can only exist in one cluster)

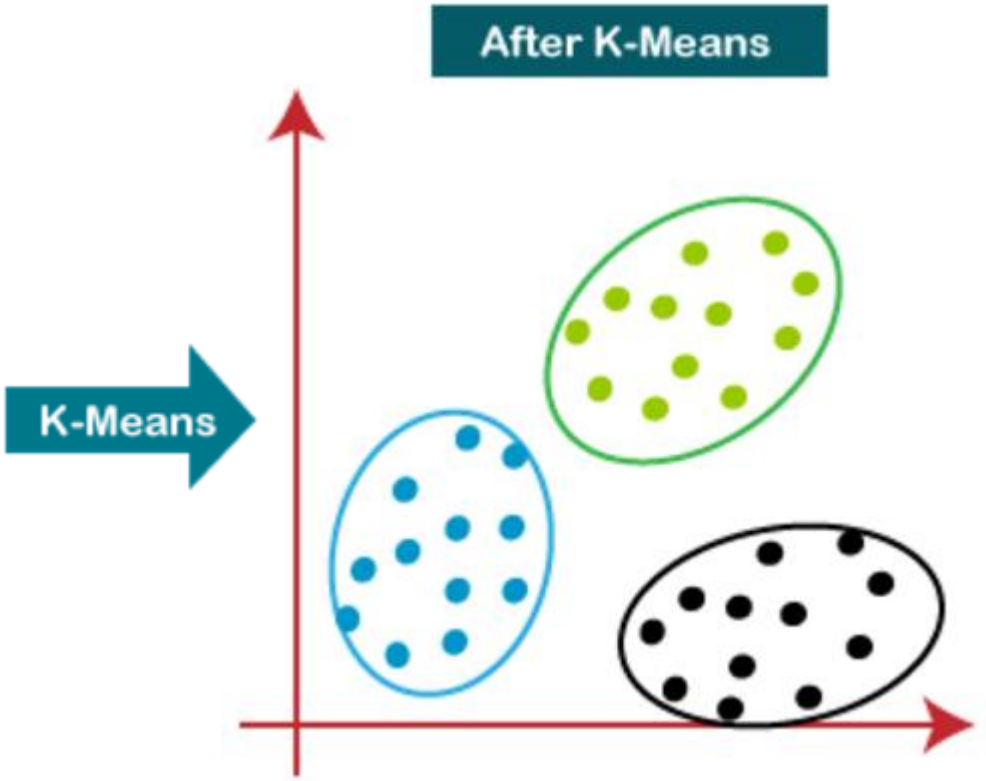
- The data points closest to a given centroid will be clustered under the same category.

- K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.

Example: Clustering the points into three cluster using K-Means

- Data points are: A1(3,11),A2(3,6),A3(9,5),B1(6,9),B2(8,6),B3(7,5),C1(2,3),C2(5,10)



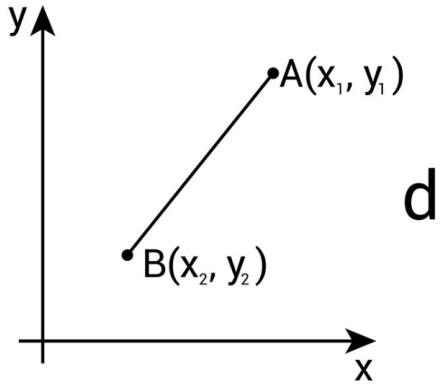


Example: Clustering the points into three cluster using K-Means

- **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).
  - Suppose initially we assign A1, B1, and C1 as the center of each cluster respectively. (K=3)

- **Step-3:** Calculate the distance between each data point and cluster centers.
  - Use Euclidean Distance function.



$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Example: Clustering the points into three cluster using K-Means

- **Step-4:** Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

- **Step-5:** Recalculate the new cluster center using:
  - (Sum of points in same cluster)/(number of points in same cluster)

- **Step-6:** Recalculate the distance between each data point and new obtained cluster centers.

- **Step-7:** If no data point was reassigned (same data points assigned in same cluster) then stop, otherwise repeat from **step-3**.

Example: Clustering the points into three cluster using K-Means

	Data Points			Distance to						Cluster
				C1		C2		C3		
				3	11	6	9	2	3	
Iteration-1	**A1**	3	11	0		4		8		1
Intial Centroids	**A2**	3	6	5		4		3		3
C1 = A1(3,11)	**A3**	9	5	8		5		7		2
C2 = B1(6,9)	**B1**	6	9	4		0		7		2
C3 = C1(2,3)	**B2**	8	6	7		4		7		2
	B3	7	5	7		4		5		2
	C1	2	3	8		7		0		3
	C2	5	10	2		1		8		2
Cetroid for for FIRST Cluster(C1)	3	11								
C1 = **(3,11)**										
Cetroid for for SECOND Cluster(C2)	7	7								
C2 = A3(9,5),B1(6,9),B2(8,6),B3(7,5),C2(5,10) = **(7,7)**										
Cetroid for for THIRD Cluster(C3)	3	5								
C3 = A2(3,6),C1(2,3) = **(3,5)**										

Example: Clustering the points into three cluster using K-Means

	Data Points			Distance to						Cluster
				C1		C2		C3		
				3	11	7	7	3	5	
Iteration-2	**A1**	3	11	0		6		6		1
Intial Centroids	**A2**	3	6	5		4		1		3
C1 = (3,11)	**A3**	9	5	8		3		6		2
C2 = (7,7)	**B1**	6	9	4		2		5		2
C3 = (3,5)	**B2**	8	6	7		1		5		2
	B3	7	5	7		2		4		2
	C1	2	3	8		6		2		3
	C2	5	10	2		4		5		1
Cetroid for for FIRST Cluster(C1)	4	11								
C1 = A1(3,11),C2(5,10) = **(4,11)**										
Cetroid for for SECOND Cluster(C2)	8	6								
C2 = A3(9,5),B1(6,9),B2(8,6),B3(7,5) = **(8,6)**										
Cetroid for for THIRD Cluster(C3)	3	5								
C3 = A2(3,6),C1(2,3) = **(3,5)**										

Example: Clustering the points into three cluster using K-Means

	Data Points			Distance to						Cluster
				C1		C2		C3		
				4	11	8	6	3	5	
Iteration-3	**A1**	3	11	1		7		6		1
Intial Centroids	**A2**	3	6	5		5		1		3
C1 = (4,11)	**A3**	9	5	8		1		6		2
C2 = (8,6)	**B1**	6	9	3		4		5		1
C3 = (3,5)	**B2**	8	6	6		0		5		2
	B3	7	5	7		1		4		2
	C1	2	3	8		7		2		3
	C2	5	10	1		5		5		1
Cetroid for for FIRST Cluster(C1)	5	10								
C1 = A1(3,11),B1(6,9),C2(5,10) = **(5,10)**										
Cetroid for for SECOND Cluster(C2)	8	5								
C2 = A3(9,5),B2(8,6),B3(7,5) = **(8,5)**										
Cetroid for for THIRD Cluster(C3)	3	5								
C3 = A2(3,6),C1(2,3) = **(3,5)**										

Example: Clustering the points into three cluster using K-Means

	Data Points			Distance to C1 (5, 10)	Distance to C2 (8, 5)	Distance to C3 (3, 5)	Cluster
Iteration-4	**A1**	3	11	2	8	6	1
Intial Centroids	**A2**	3	6	4	5	1	3
C1 = (5,10)	**A3**	9	5	6	1	6	2
C2 = (8,5)	**B1**	6	9	1	4	5	1
C3 = (3,5)	**B2**	8	6	5	1	5	2
	B3	7	5	5	1	4	2
	C1	2	3	8	6	2	3
	C2	5	10	0	6	5	1

Cetroid for for FIRST Cluster(C1)	5	10
C1 = A1(3,11),B1(6,9),C2(5,10) = **(5,10)**		
Cetroid for for SECOND Cluster(C2)	8	5
C2 = A3(9,5),B2(8,6),B3(7,5) = **(8,5)**		
Cetroid for for THIRD Cluster(C3)	3	5
C3 = A2(3,6),C1(2,3) = **(3,5)**		

Note: After Iteration-4, No change in clustering so, stop the Algorithm