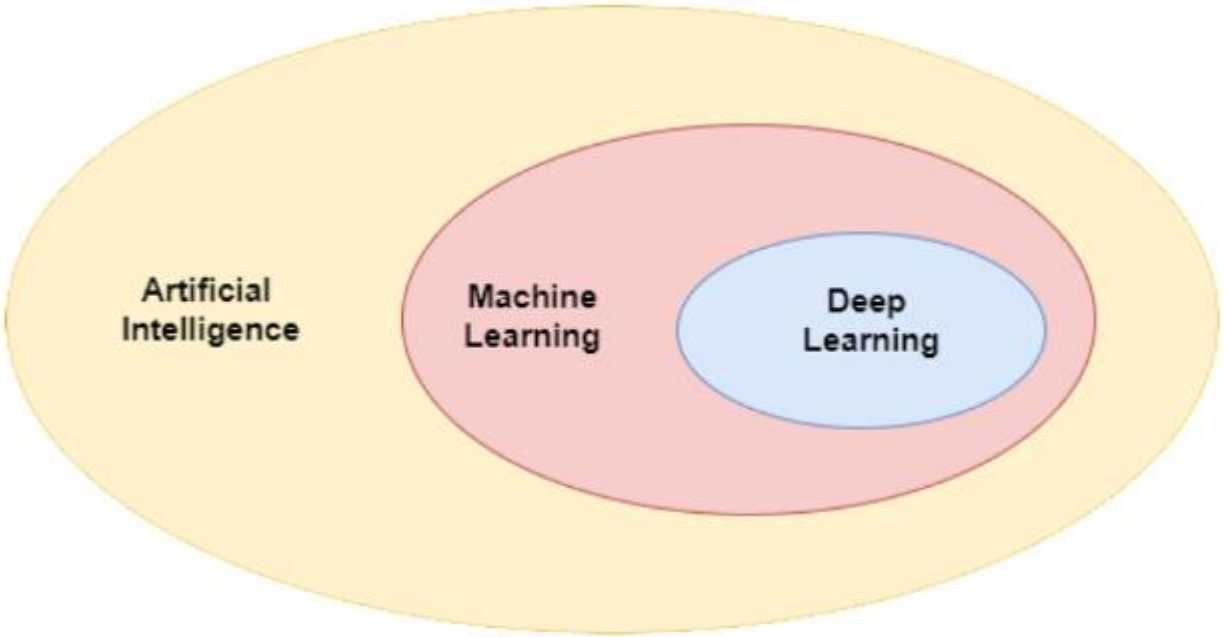
UEM
श्रद्धावान लभते ज्ञानम्
UNIVERSITY OF ENGINEERING & MANAGEMENT
Good Education, Good Jobs

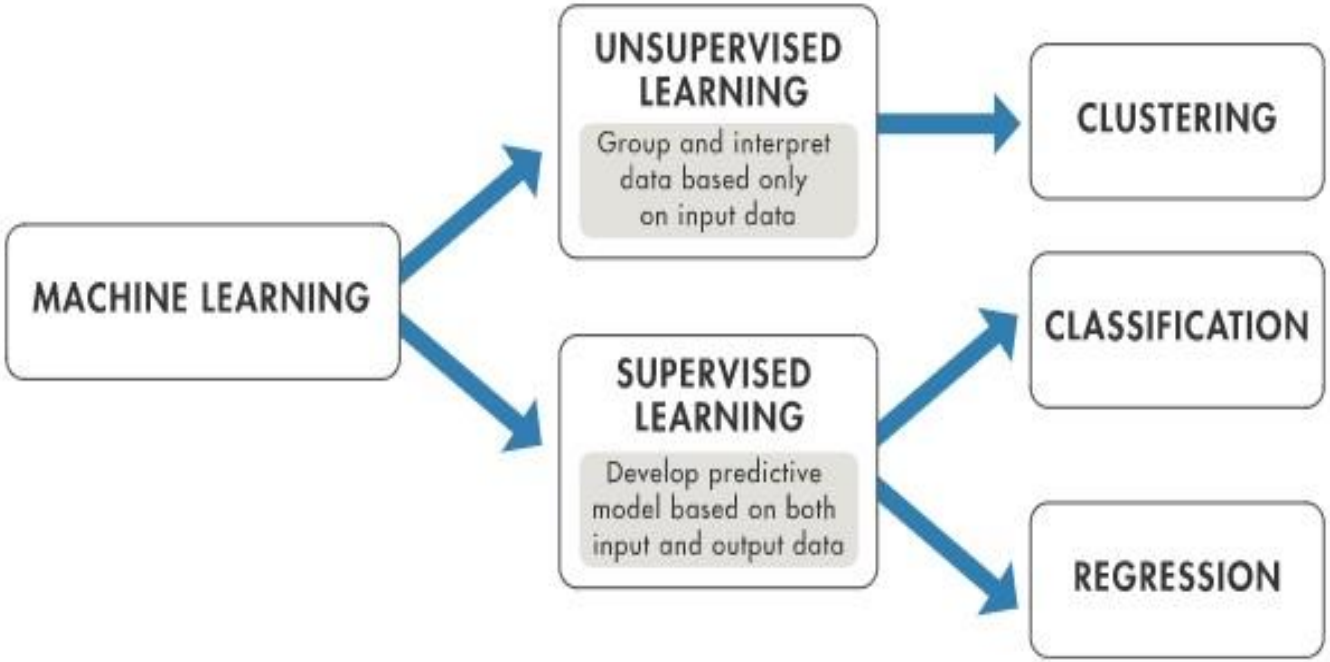
UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

Course Name : AI & ML


Artificial Intelligence
Machine Learning
Deep Learning

Machine Learning?

Tom M. Mitchell: "**Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience .**"


MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

Supervised learning: which trains a model on known input and output data so that it can predict future outputs.

k-Nearest Neighbors, Linear Regression, Logistic Regression, Support Vector Machines(SVMs), Decision Trees & Random forests, Neural Networks

Unsupervised learning: which finds hidden patterns or intrinsic structures in input data.

Clustering: k-means, Hierarchical Cluster Analysis (HCA), Expectation Maximization

Visualization and dimensionality reduction: PCA, LLE, t-SNE

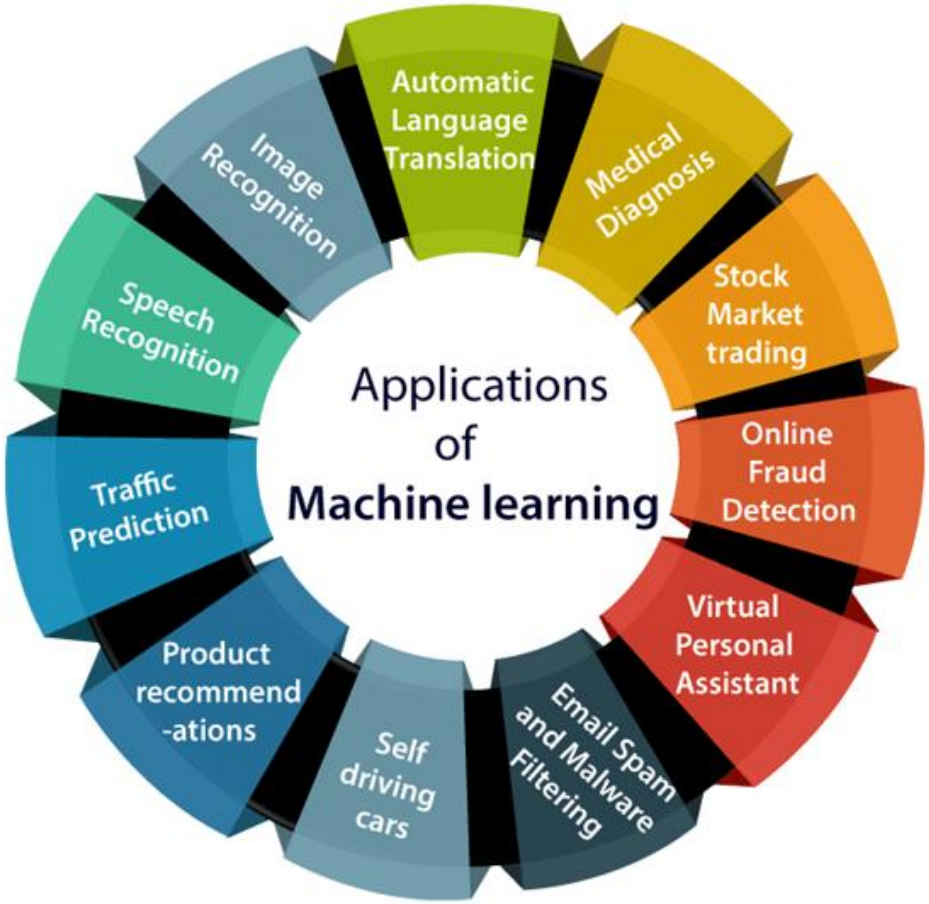
Association rule learning: Apriori, Eclat

Semi-supervised learning: Deals with partially labeled training data, usually a lot unlabeled data and a little bit of labeled data. This type of algorithms are combinations of unsupervised and supervised algorithms.

Reinforcement learning: The learning system called an agent in this context, can observe the environment, select and perform actions and get rewards in return. Then based on this system will find the best policy.

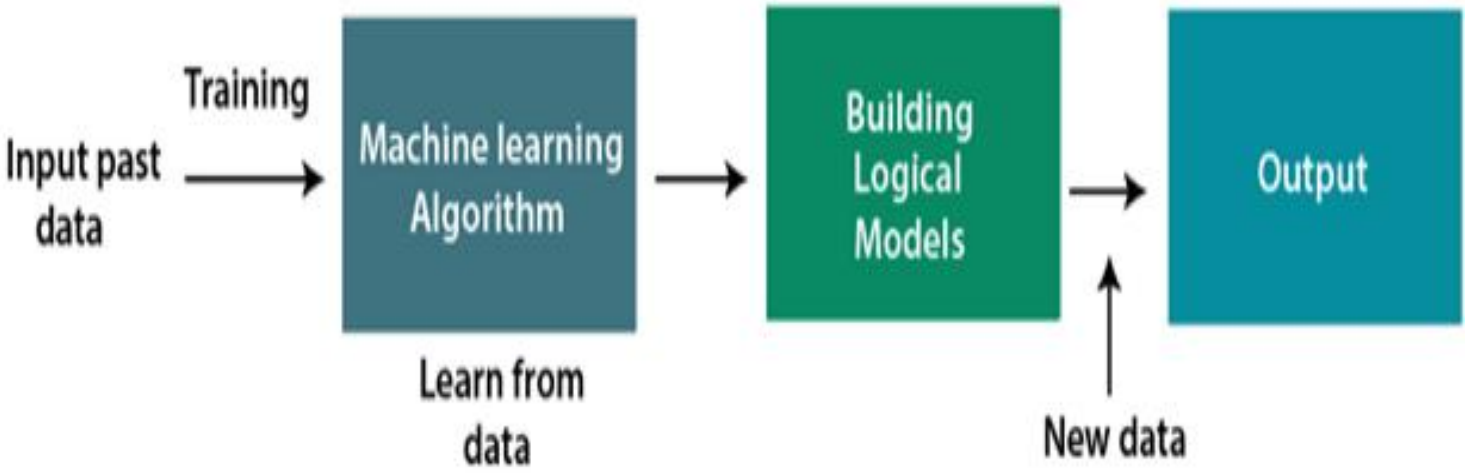
Examples:

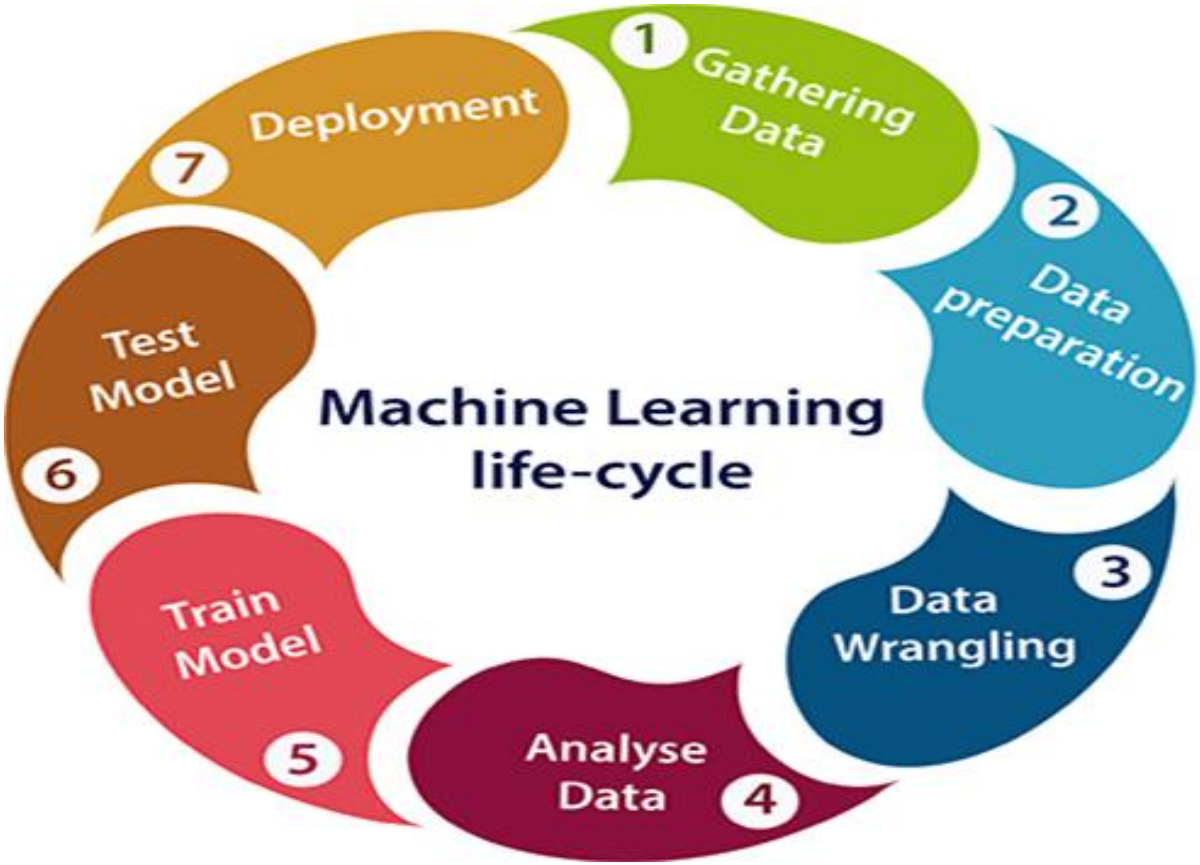
- Housing price prediction
- Mutual fund prediction
- Predicting if an email is spam or not spam
- Whether a tumor is malignant or benign
- Whether a mushroom is poisonous or edible.


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

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.


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


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

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.

Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

Data wrangling is the process of cleaning and converting raw data into a useable format.

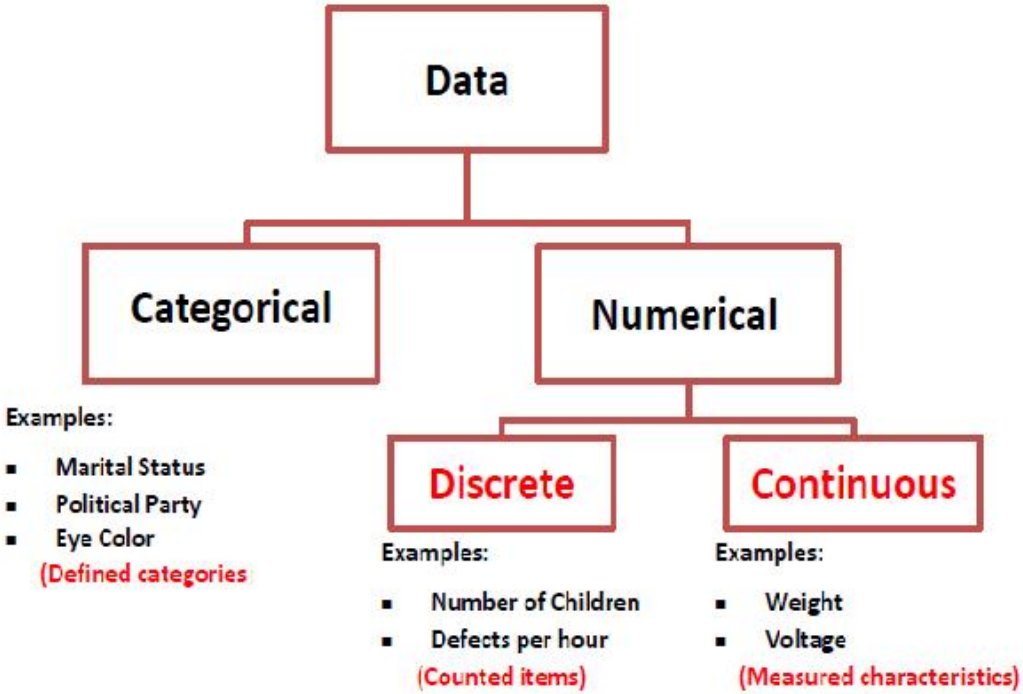
Data analysis is being done using machine learning techniques such as **Classification**, **Regression**, **Cluster analysis**, **Association**, etc.

Training a model is required so that it can understand the various patterns, rules, and, features.

Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

If the above-prepared model is producing an accurate result as per the requirement with acceptable speed, then **deploy** the model in the real system.

Types of Variables



Four different levels of Data

Nominal

Ordinal

Interval

Ratio

- A **nominal** scale classifies data into distinct categories in which no ranking is implied
- Example : Gender, Marital Status
- An **ordinal** scale classifies data into distinct categories in which ranking is implied
- Example : Student Grades: A, B, C, D, F
- An **interval** scale is an ordered scale in which the difference between measurements is a meaningful quantity but the measurements do not have a true zero point.
- Example : Temperature in Fahrenheit and Celsius , Year
- A **ratio** scale is an ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have a true zero point.
- Example : Weight , Age ,Salary

- **Measures of Central Tendency**
- Measures of central tendency yield information about “particular places or locations in a group of numbers.”
- A single number to describe the characteristics of a set of data .

Central Tendency Measures are: Mean, Median, Mode

Arithmetic Mean

- Commonly called ‘the mean’
- It is the average of a group of numbers
- Applicable for interval and ratio data
- Not applicable for nominal or ordinal data
- Affected by each value in the data set, including extreme values
- Computed by summing all values in the data set and dividing the sum by the number of values in the data set

- **Median**
- Middle value in an ordered array of numbers
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data
- Unaffected by extremely large and extremely small values

- **Median: Computational Procedure**
- First Procedure

–Arrange the observations in an ordered array

–If there is an odd number of terms, the median is the middle term of the ordered array

–If there is an even number of terms, the median is the average of the middle two terms

•Second Procedure

- –The median's position in an ordered array is given by (n+1)/2.

- **Median: Example with an Odd Number of Terms**
- Ordered Array
- 3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22
- •There are 17 terms in the ordered array.
- •Position of median = (n+1)/2 = (17+1)/2 = 9
- •The median is the 9th term, 15.
- •If the 22 is replaced by 100, the median is 15.
- •If the 3 is replaced by -103, the median is 15.

- **Median: Example with an Even Number of Terms Ordered Array 3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21**
- •There are 16 terms in the ordered array
- •Position of median = (n+1)/2 = (16+1)/2 = 8.5
- •The median is between the 8th and 9th terms, 14.5
- •If the 21 is replaced by 100, the median is 14.5
- •If the 3 is replaced by -88, the median is 14.5

- **Mode**
- The most frequently occurring value in a data set
- Applicable to all levels of data measurement (nominal, ordinal, interval, and ratio)
- Bimodal -- Data sets that have two modes
- Multimodal -- Data sets that contain more than two modes

Mode -- Example

- The mode is 44
- There are more 44s than any other value

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

- **Dispersion**
- Measures of variability describe the spread or the dispersion of a set of data
- Reliability of measure of central tendency

Dispersion of data

Variance and **Standard Deviation** are essentially a measure of the spread of the data in the data set.

Variance is the average of the squared differences from the mean.

Standard deviation is the square root of the variance.

Standard deviation is an excellent way to identify outliers. Data points that lie more than one standard deviation from the mean can be considered unusual.

Population Variance

- Average of the squared deviations from the arithmetic mean

X	$X-\mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

Population Standard Deviation

- Square root of the variance

X	$X-\mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

$$\sigma = \sqrt{\sigma^2}$$

$$= \sqrt{26.0}$$

$$= 5.1$$

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other.

Where,

xi = data value of x

yi = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values.

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables. It can assume values from -1 to +1. When the correlation coefficient is positive, an increase in one variable also increases the other. When the correlation coefficient is negative, the changes in the two variables are in opposite directions.

$$Correlation = \frac{Cov\,(x, y)}{\sigma x * \sigma y}$$

where:

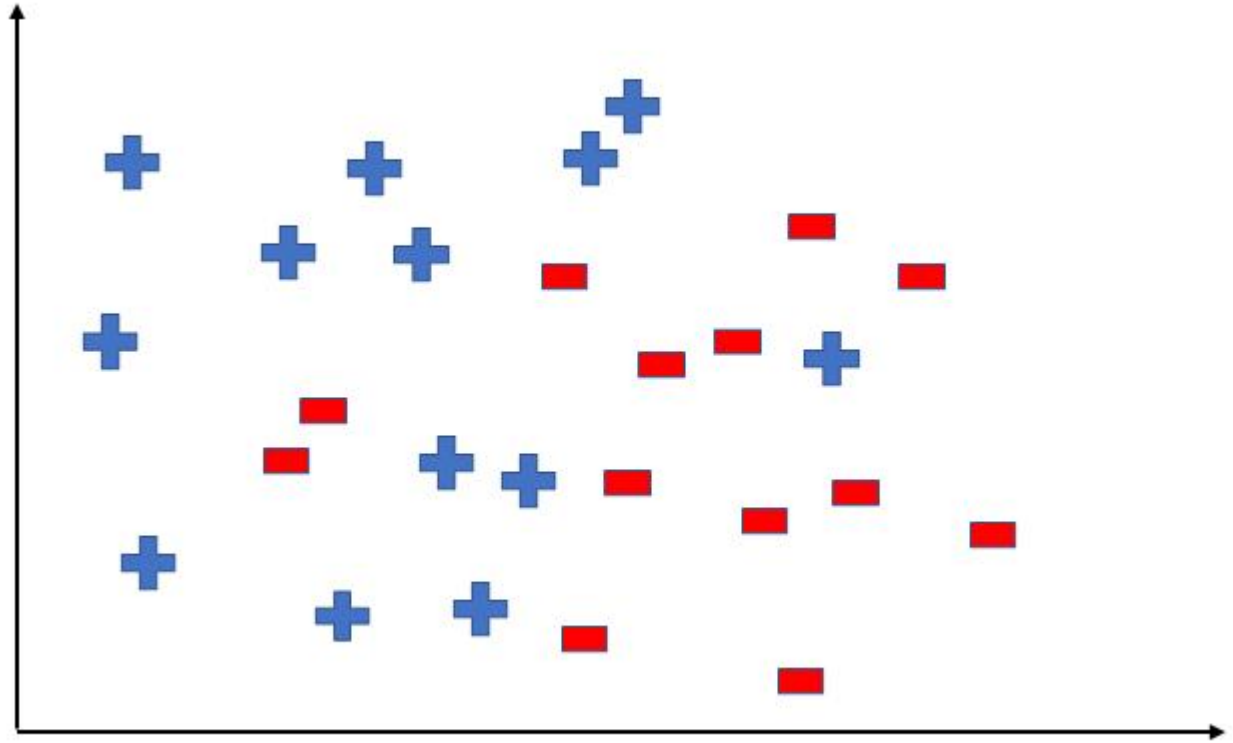
- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

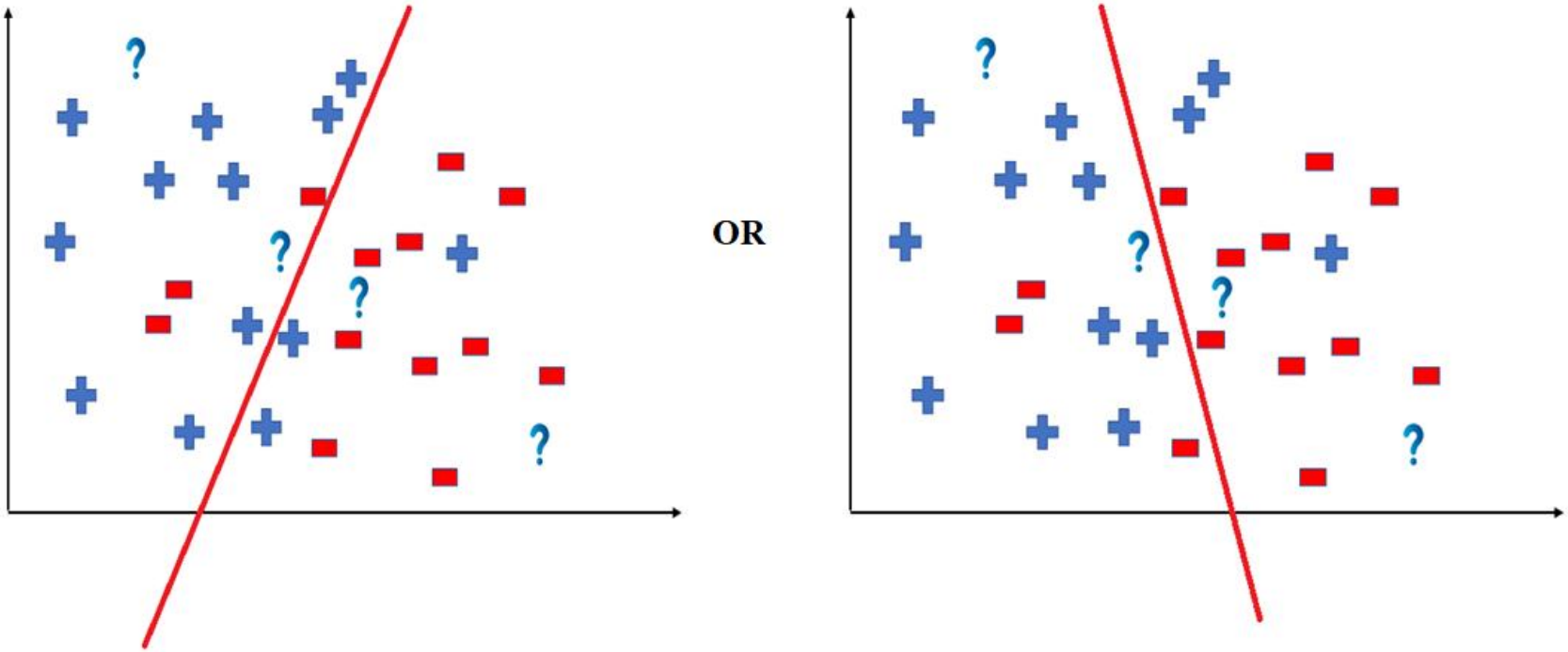
How to ensure Data quality?

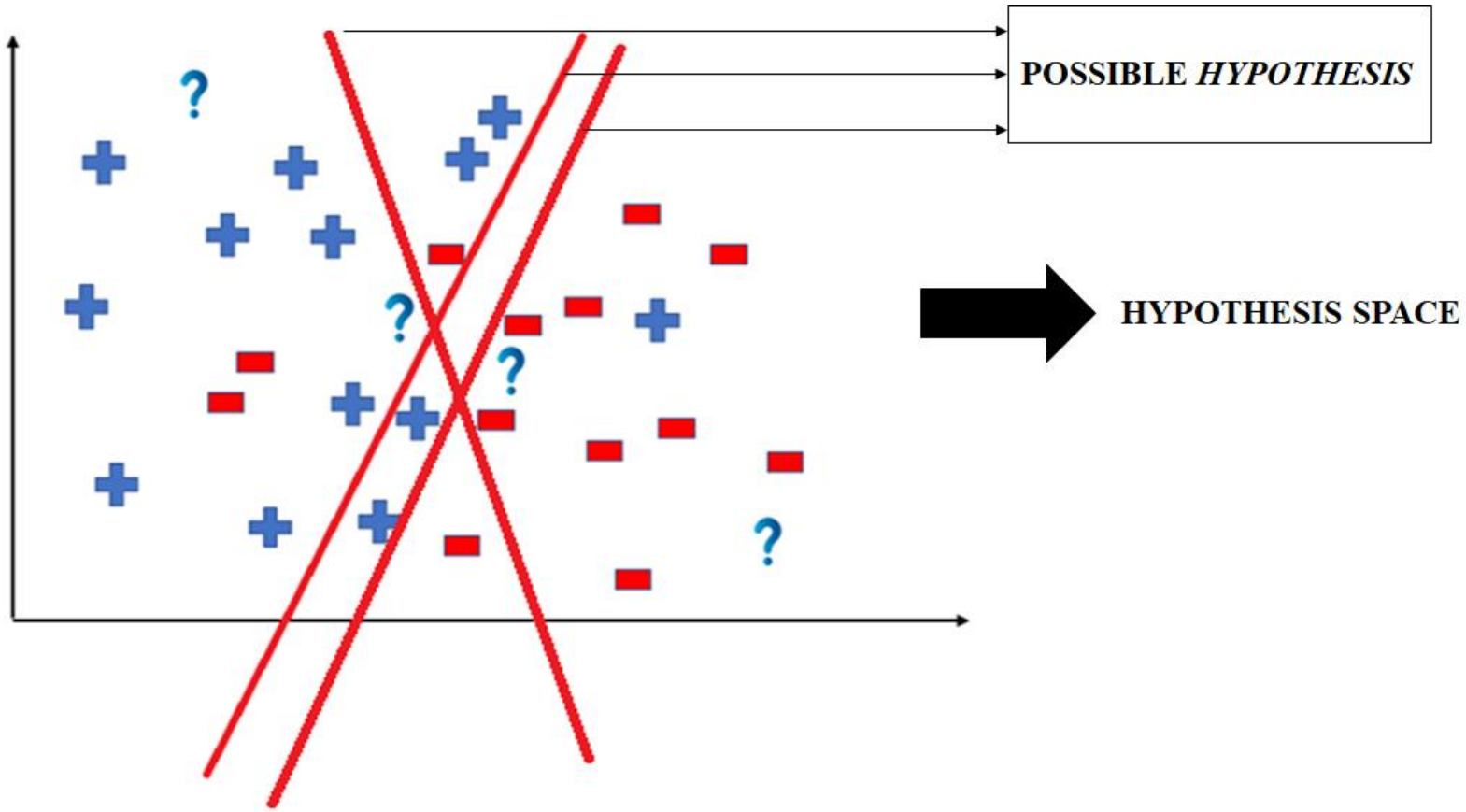
Data quality issues? [Incorrect sample set (ex. Festive data), errors in data collection (missing & outliers value)]

Remedies? [Solving Outliers issue (removing outliers, imputation, capping), Solving Missing value issue (Elimination, Imputing, Estimating)]

- **Hypothesis Space (H):**
  Hypothesis space is the set of all the possible legal hypothesis. This is the set from which the machine learning algorithm would determine the best possible (only one) which would best describe the target function or the outputs.
- **Hypothesis (h):**
  A hypothesis is a function that best describes the target in supervised machine learning. The hypothesis that an algorithm would come up depends upon the data and also depends upon the restrictions and bias that we have imposed on the data.
- In most supervised machine learning algorithm, our main goal is to find out a possible hypothesis from the hypothesis space that could possibly map out the inputs to the proper outputs.
- **NULL** hypotheses are usually the statements which the scientists want to prove wrong, but will start the research towards that goal assuming that the null hypothesis is true.




OR


POSSIBLE *HYPOTHESIS*
HYPOTHESIS SPACE

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Thank You