Auswärtiges Amt

DAAD Deutscher Akademischer Austauschdienst
German Academic Exchange Service

Geo-IT Online Seminar
Freie Universität Berlin
Institute of Geographical Sciences

Freie Universität Berlin

TRAINING ON GOOGLE EARTH ENGINE

MODULE 6/2 : CROP CLASSIFICATION

Prof. Dr. Eng. Ayman Abdulrahman

Coordinated from

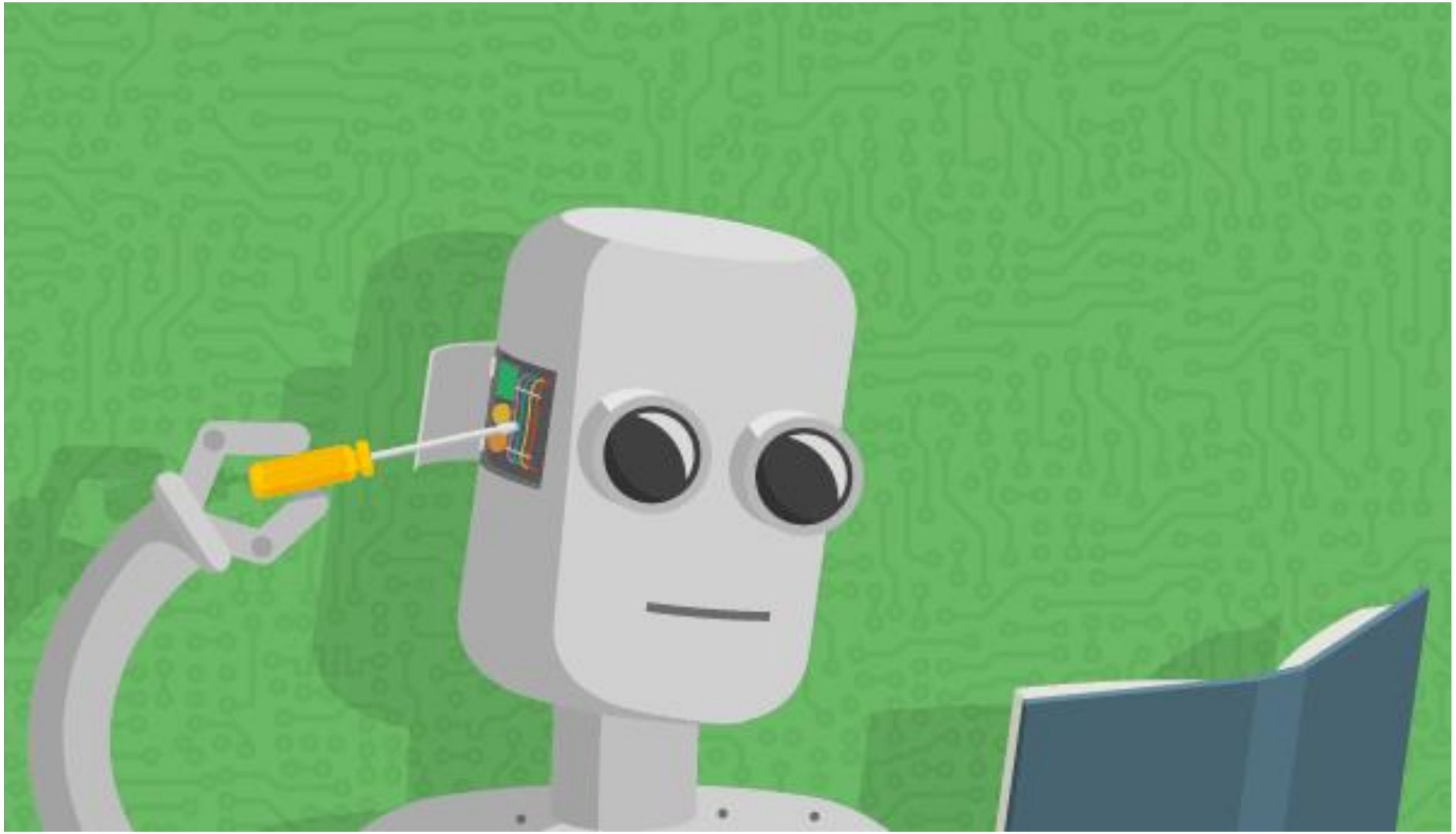
Dr. Wahib Sahwan (Freie Universität Berlin)

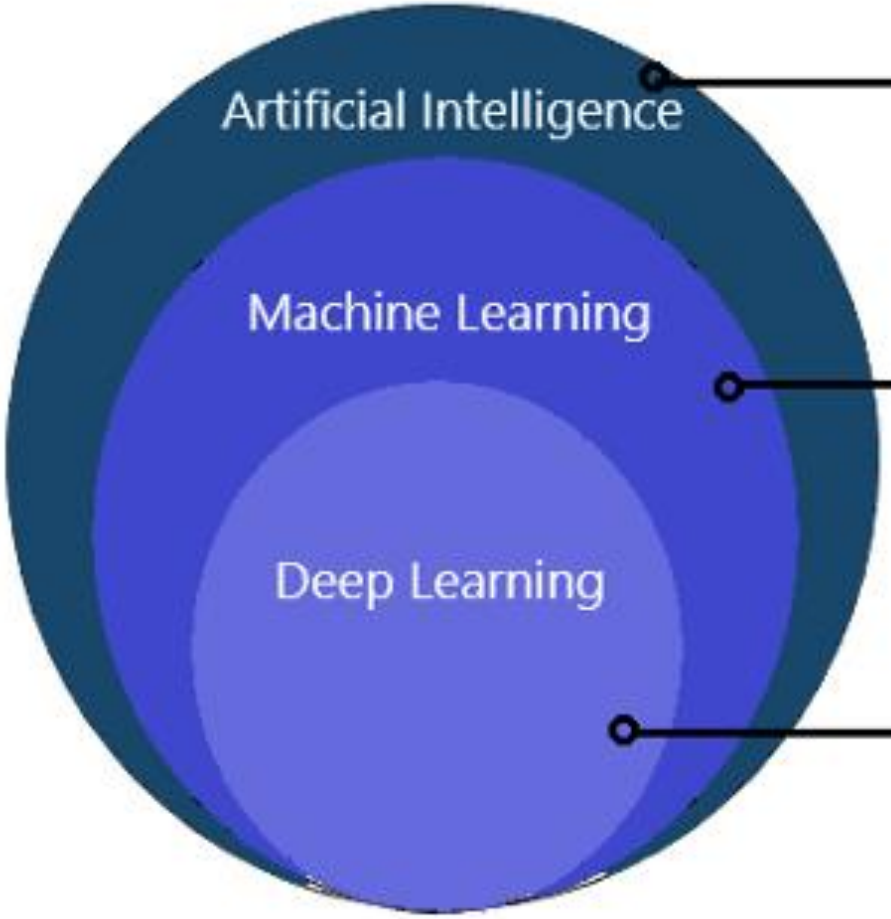
Geo-IT

The Technology of Data Acquisition for Sustainable Development and Crisis Management

(Germany, Jordan, Lebanon and Syria)

Machine learning





Artificial Intelligence

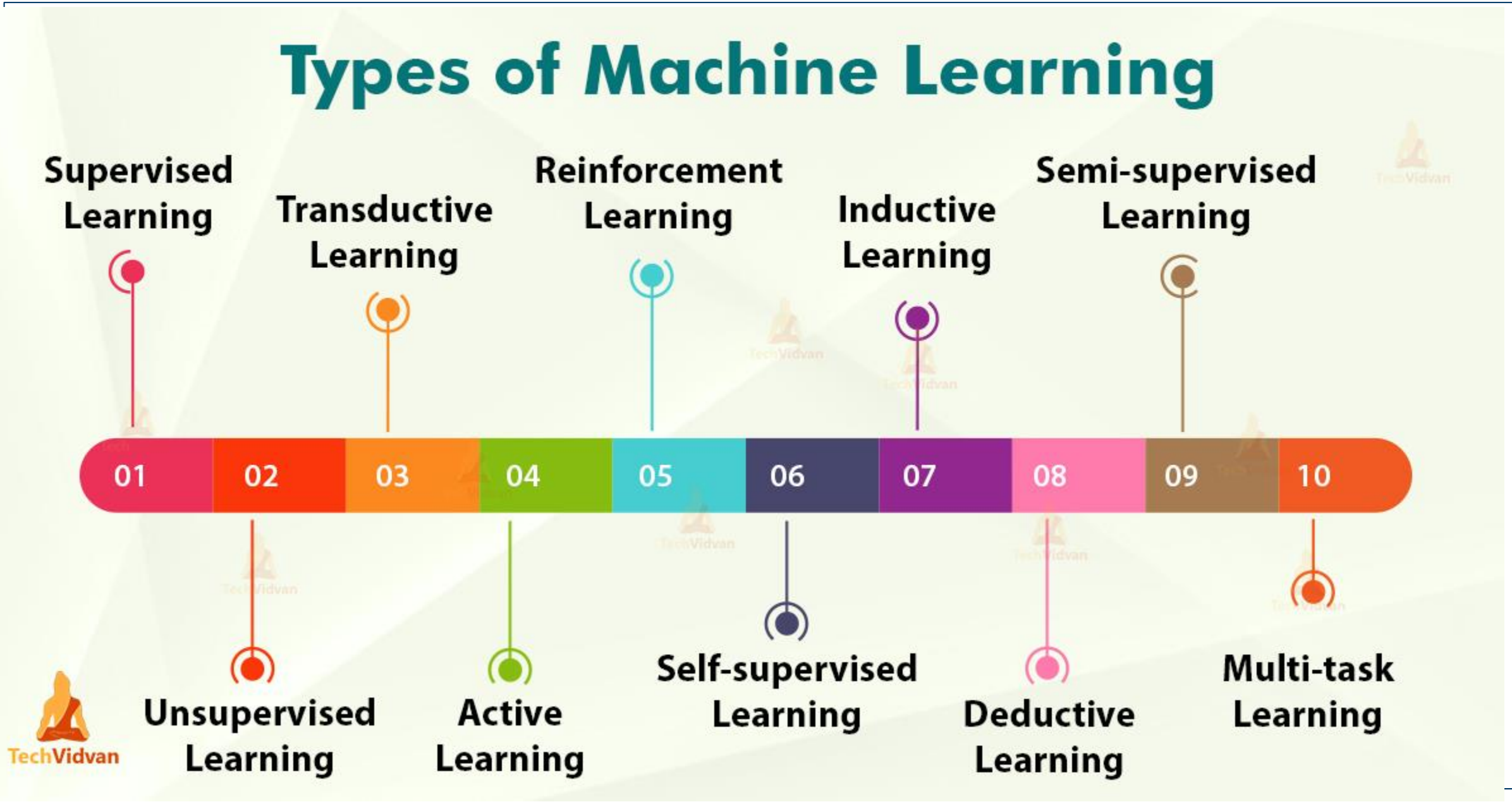
AI enables machines to think without any human intervention.

Machine Learning

Subset of AI that uses statistical learning algorithms that learn pattern in data over time

Deep Learning

Subset of ML that filters the data through multiple layers


Types of Machine Learning
Supervised Learning
Transductive Learning
Reinforcement Learning
Inductive Learning
Semi-supervised Learning
01
02
03
04
05
06
07
08
09
10
Unsupervised Learning
Active Learning
Self-supervised Learning
Deductive Learning
Multi-task Learning
TechVidvan

Machine learning

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

Supervised learning:

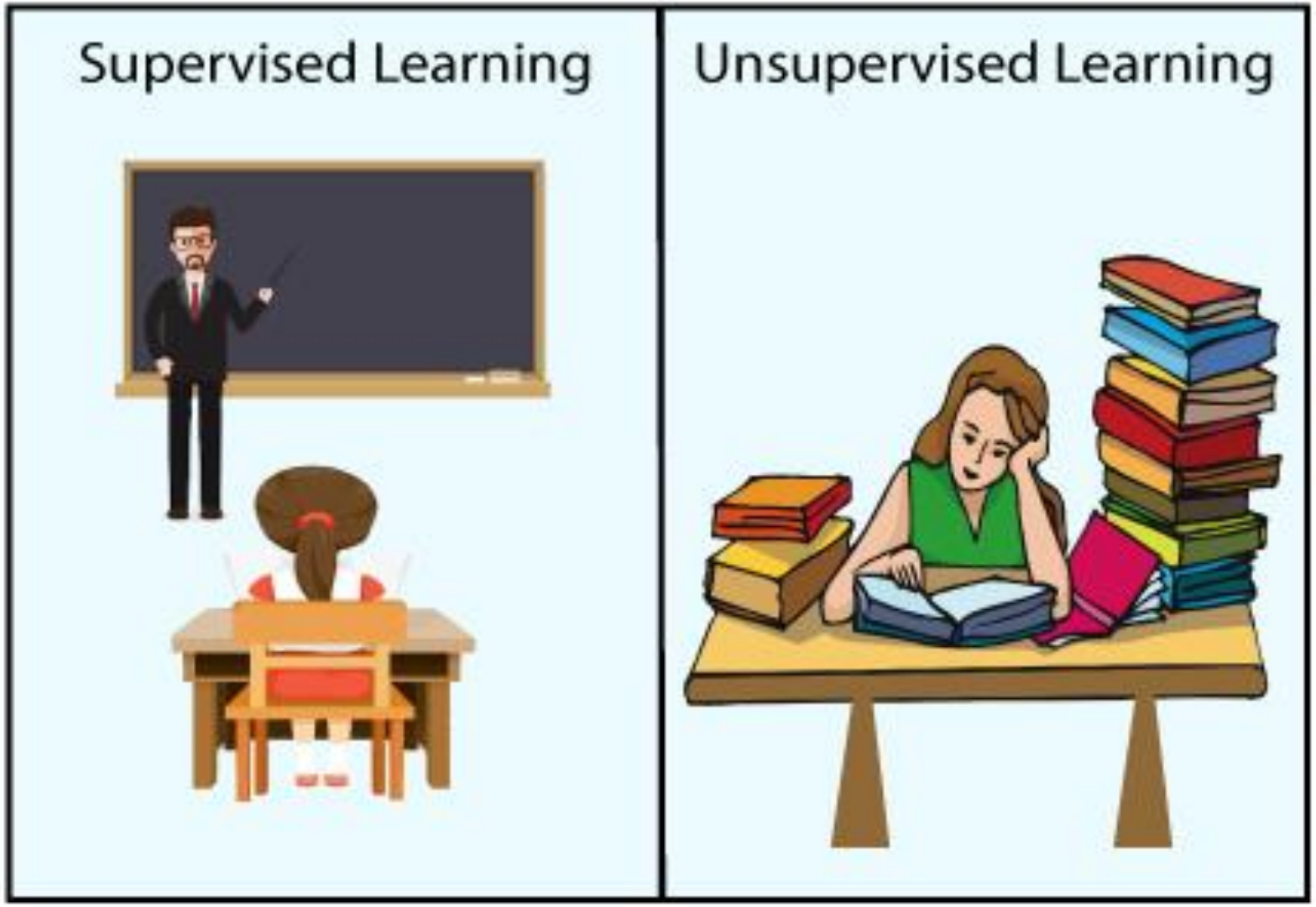
The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

Unsupervised learning:

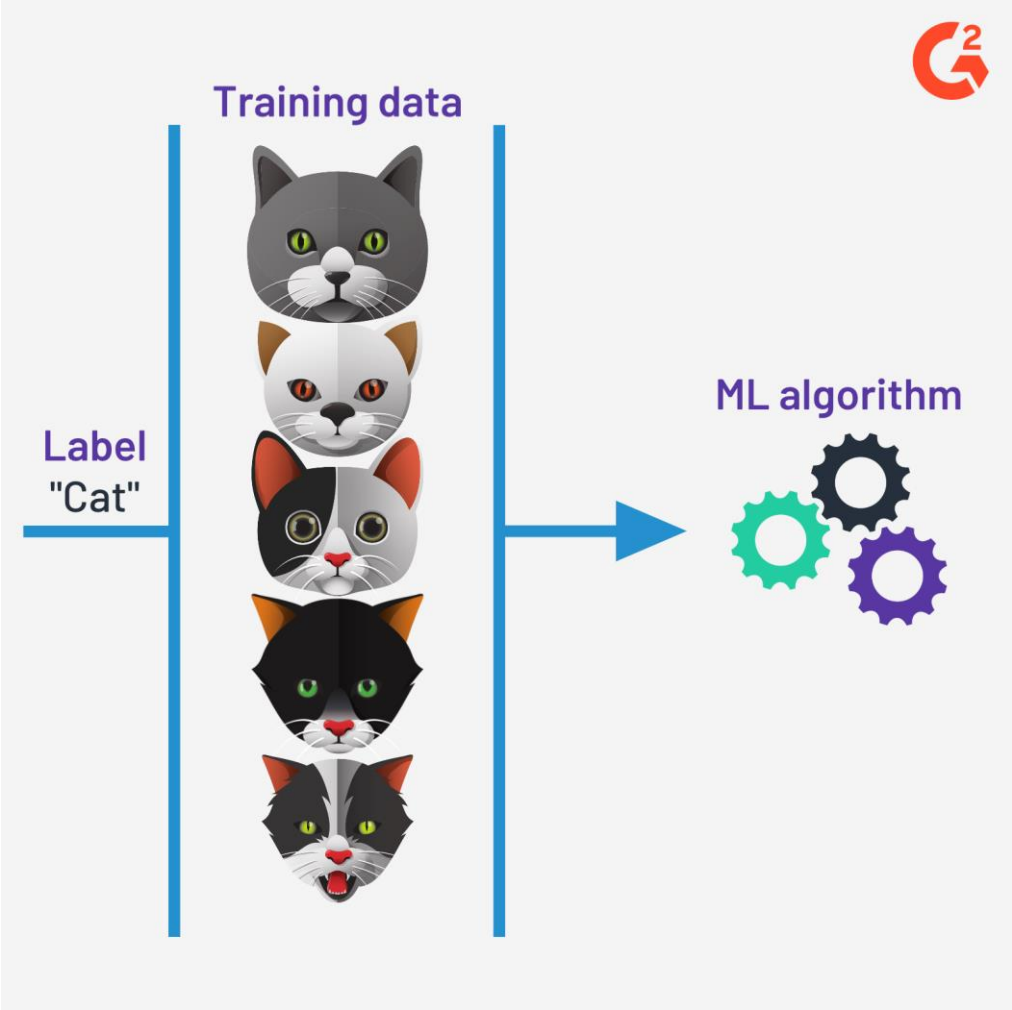
No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

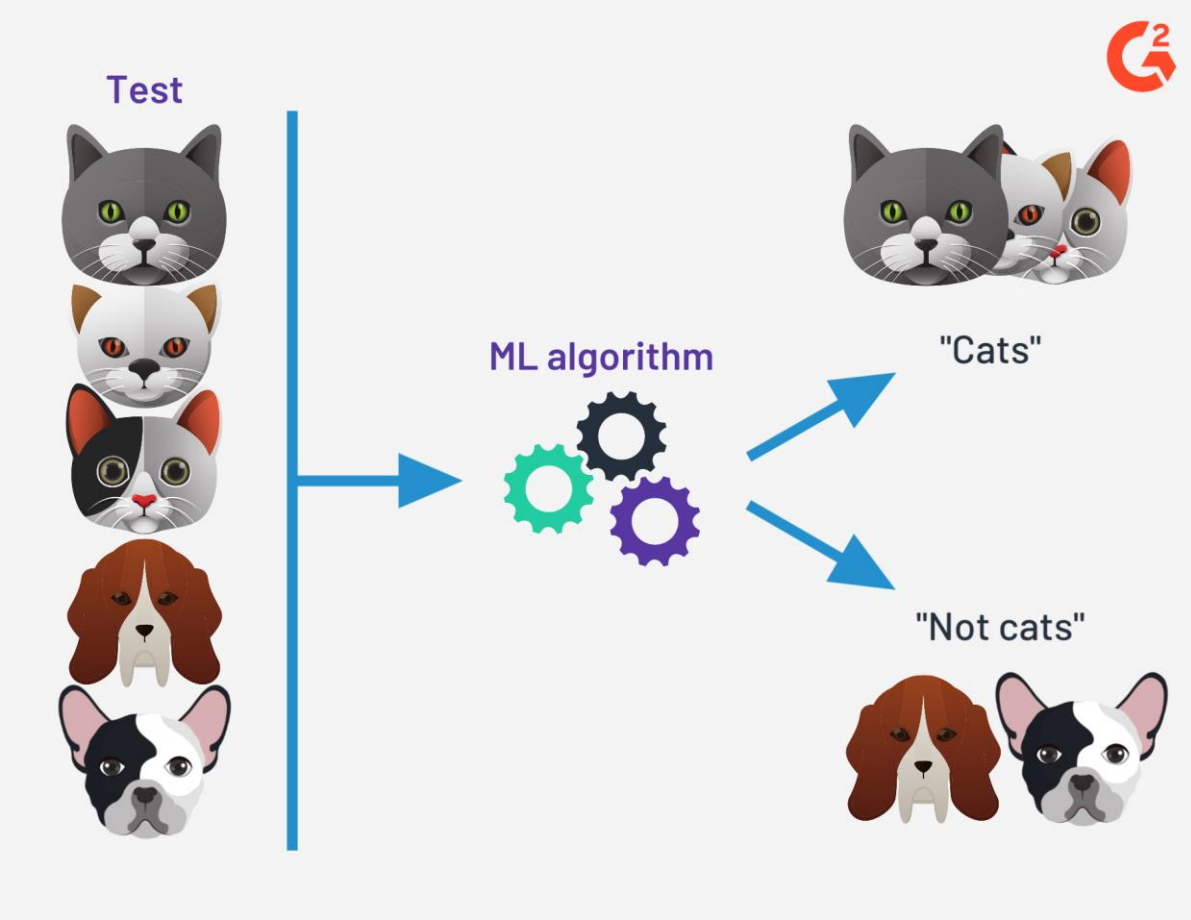
Reinforcement learning:

A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize

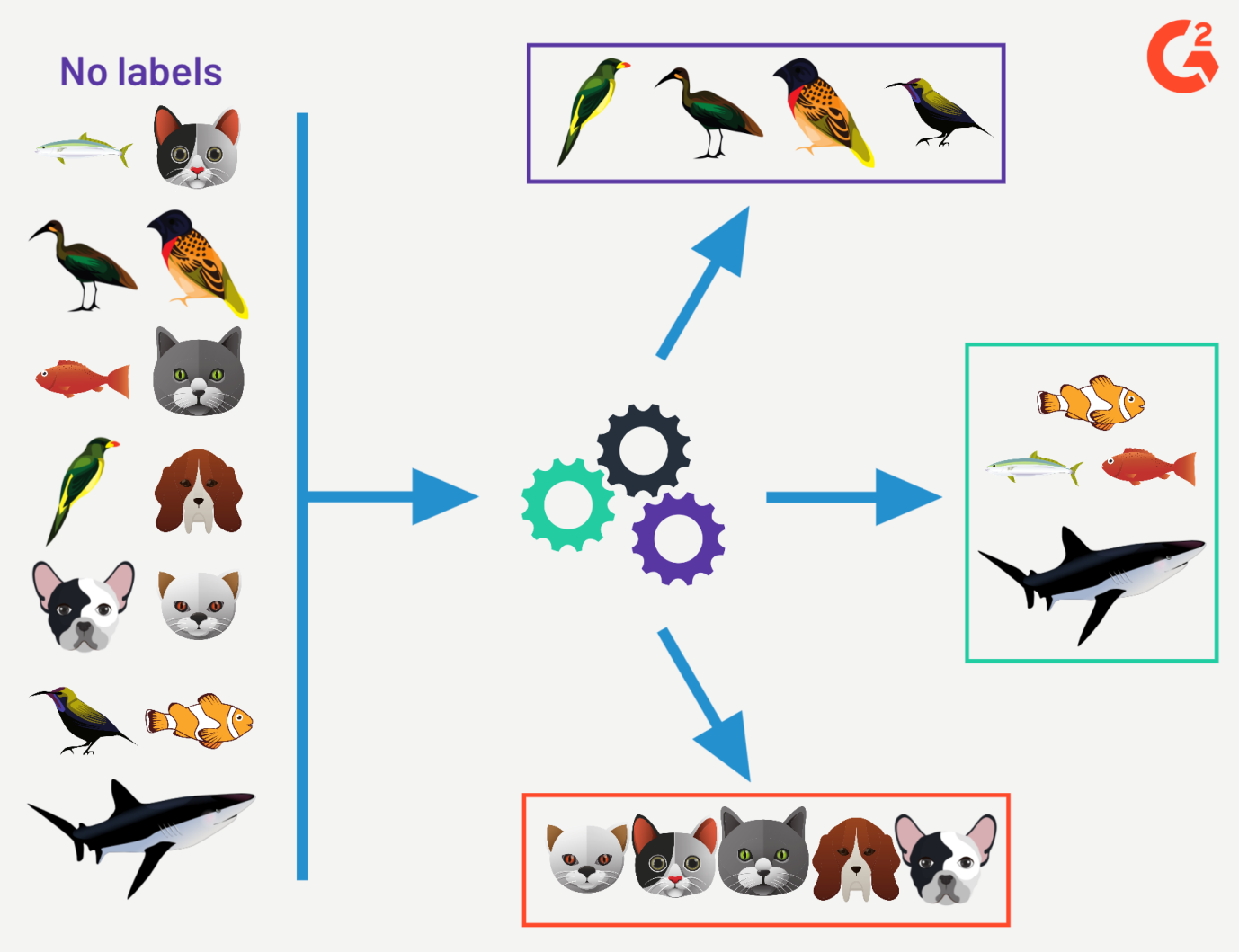

Supervised Learning
Unsupervised Learning

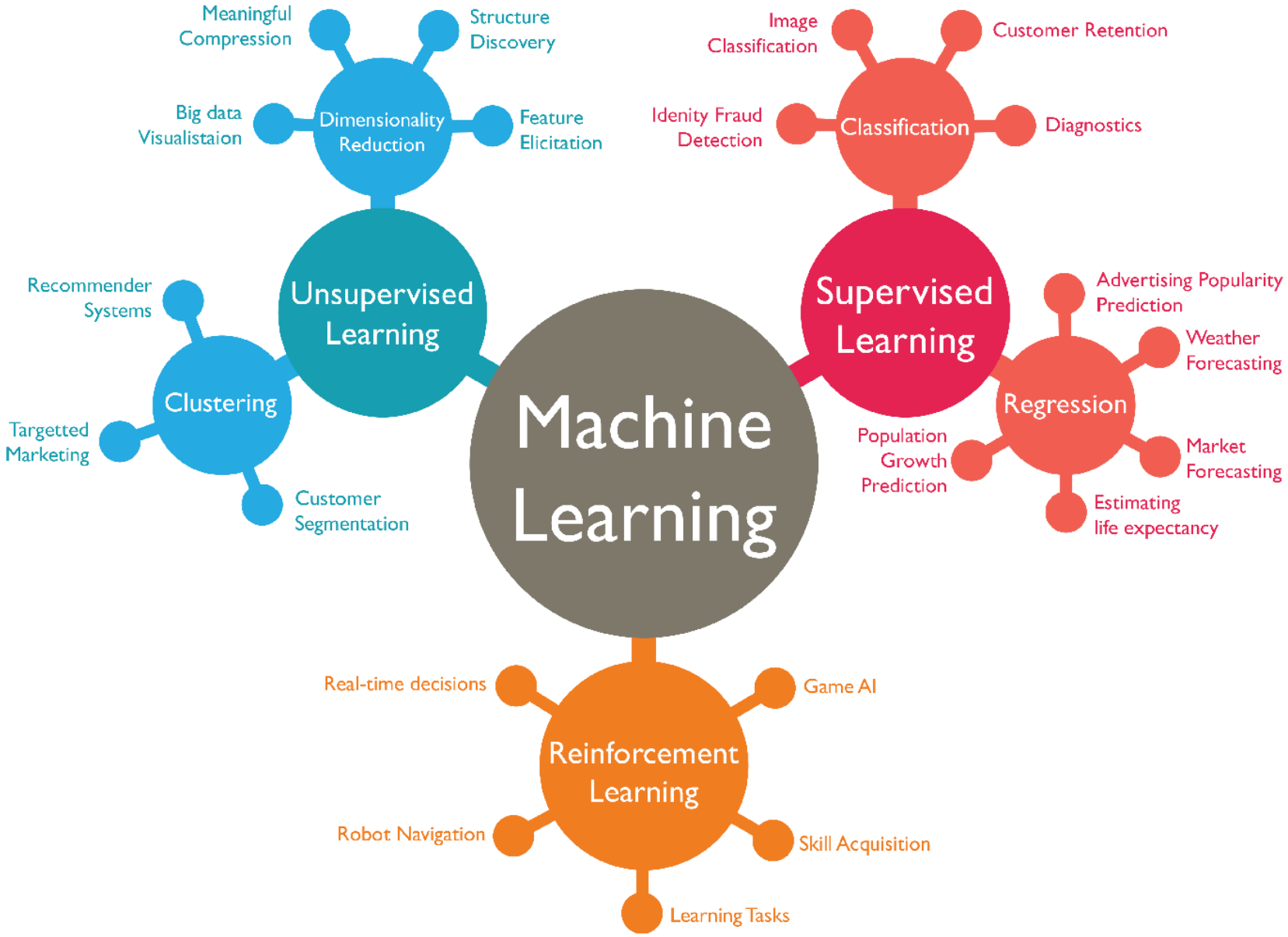
Supervised learning





Unsupervised learning

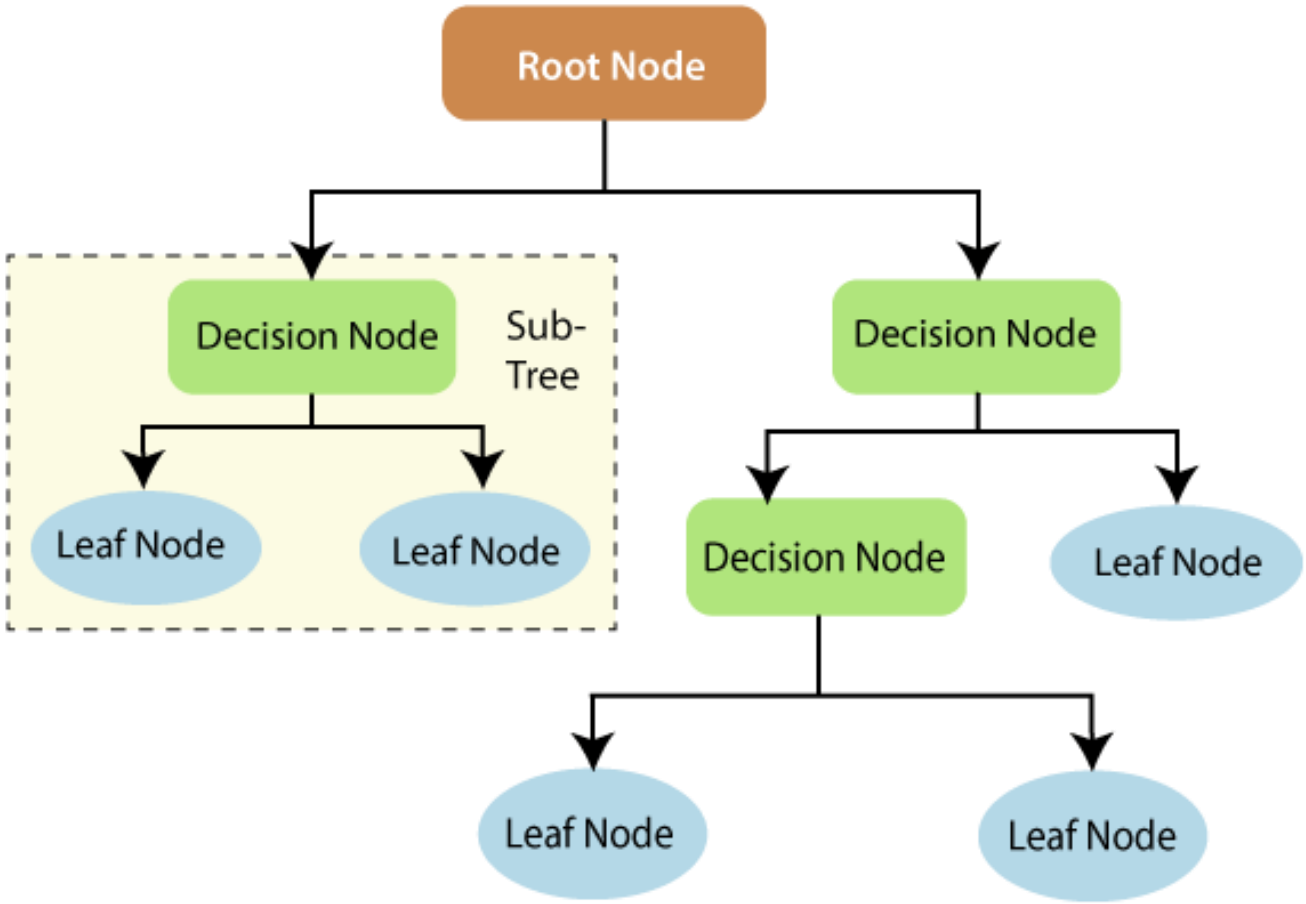


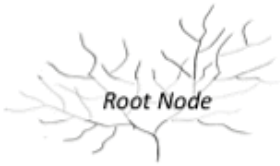

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

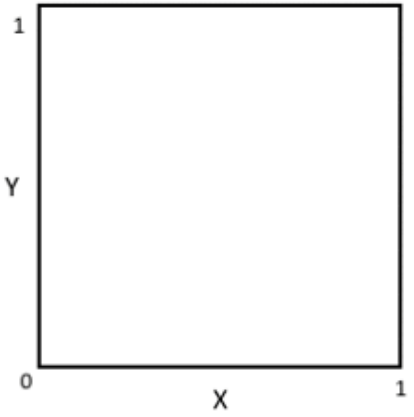
Decision Tree

Decision Tree is a supervised machine learning algorithm used for both classification as well as for regression.
Decision Tree is a tree-structured classifier and can be used to determine a course of action. In this each internal node represents a test case for some attribute, each branch represents the outcome/answers of the test case and each leaf node represents a class label.
Decision tree uses multiple algorithms to decide how to split a node into two or more than two sub-nodes. Homogeneity of the resultant sub-nodes increases with the formation of new sub-nodes.
Decision Tree splits the nodes on all the available variables and then selects the one which results in the most homogeneous sub-node.
Selection of the algorithm to split the node is also based on the types of target variables.

Decision Tree





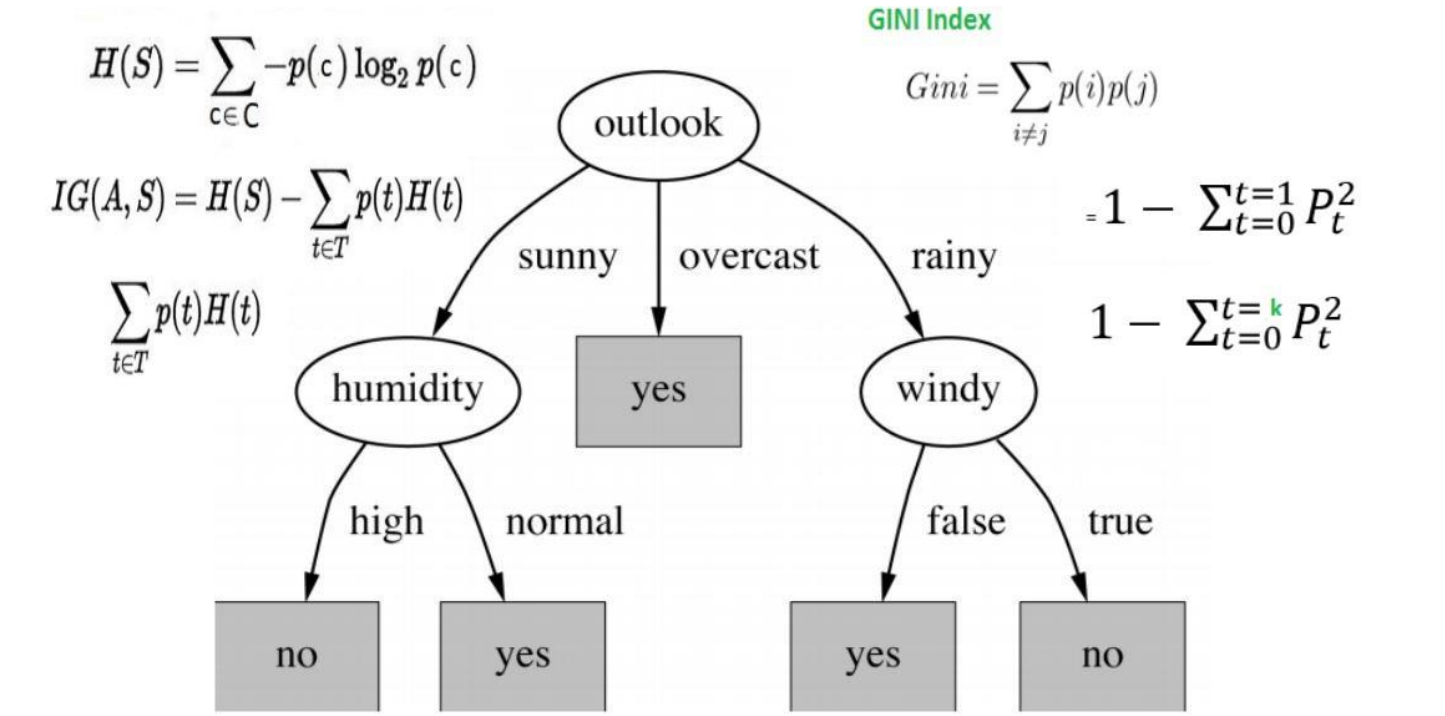


For more tutorials: annalyzin.wordpress.com

Entropy Information gain

Entropy is the measures of impurity, disorder or uncertainty in a bunch of examples.
Entropy controls how a Decision Tree decides to split the data. It actually effects how a Decision Tree draws its boundaries.
Information gain (IG) measures how much “information” a feature gives us about the class.

Final decision tree



Random Forest (RF)

Random Forest is a Supervised Machine Learning algorithm and just like a decision tree it can also be used for the classification as well as for the regression

Random Forest or Random Decision Forest is a method that operates by constructing multiple Decision Trees during the training phase. The decision of the majority of the trees is chosen by the Random Forest as the final decision.

It works on the concept of ensemble learning technique, which operates by combining multiple classifiers to solve a complex problem and to improve the performance of the model

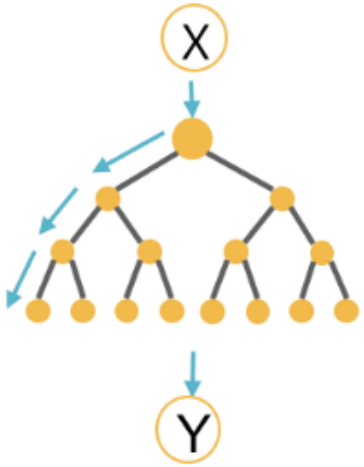
Random Forests have no formal distributional assumptions, they are non-parametric and are able to handle skewed and multi-modal data as well as categorical data that are ordinal or non-ordinal.

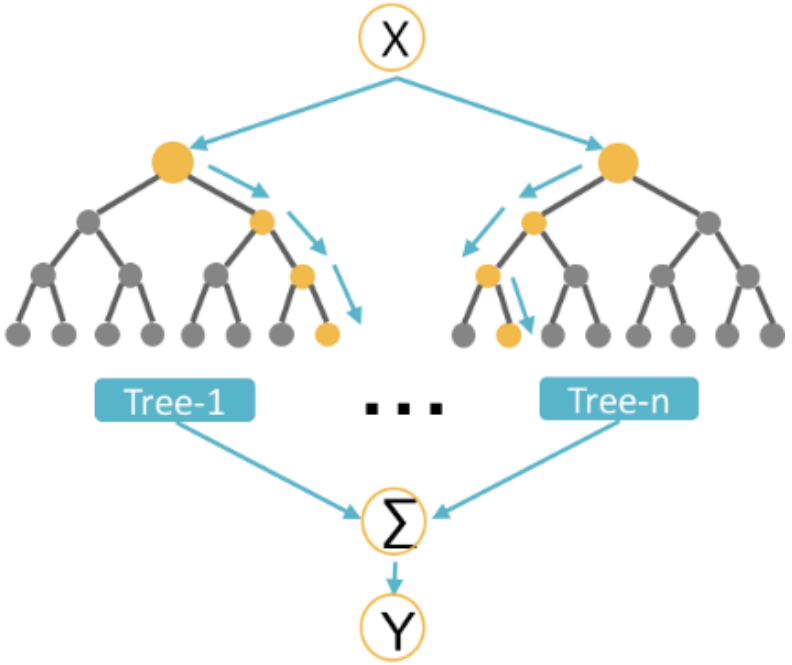
Random Forest Vs Decision Tree

Decision Tree uses the complete dataset while the Random Forest uses only randomly selected rows (or say record or observations) and columns (or say features or variables) to build multiple Decision Tree.

The Decision Tree is simple to understand as compared to Random Forest.

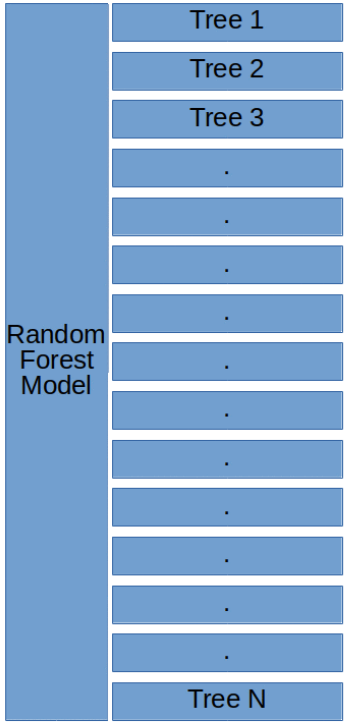
There is a chance of overfitting in the Decision Tree but there is very less chance of overfitting in Random Forest.



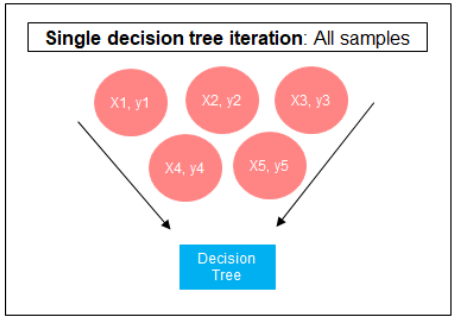


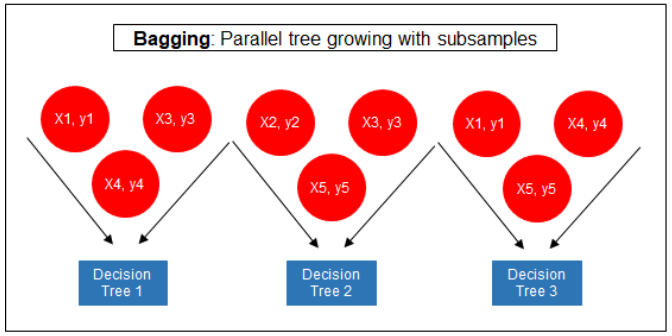
Random Forest (RF)

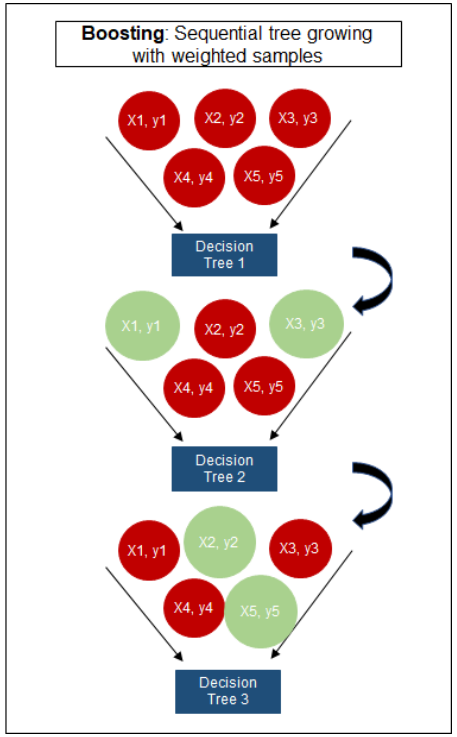
Animation RF



Bagging and Boosting







K-Means Clustering

K-Means Clustering is an Unsupervised Machine Learning Algorithm, Which is used for the Classification Problem.

K-Means separate out the labeled data into different groups (also known as Clusters), on the basis of similar features and common patterns.

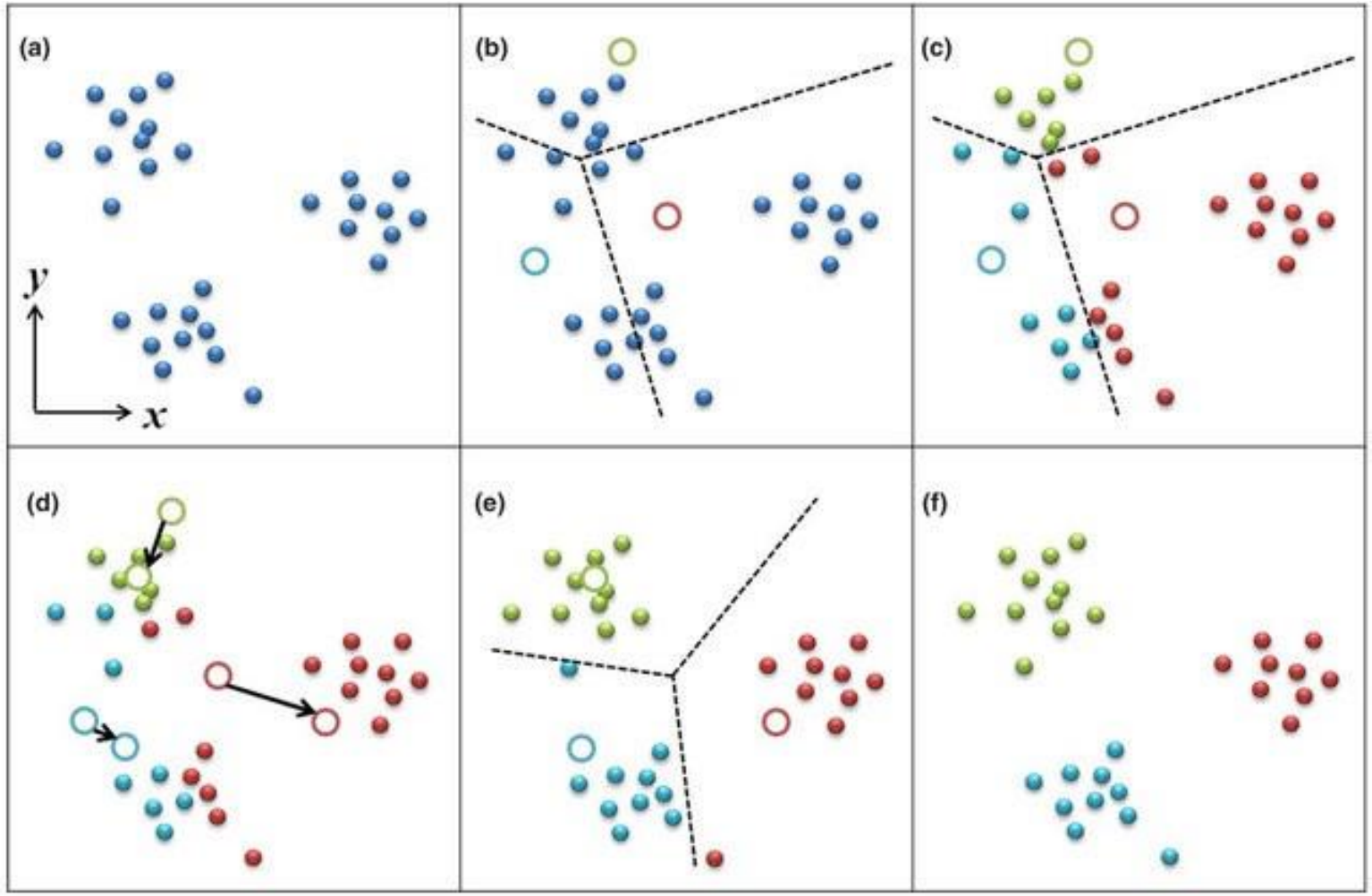
It is an Iterative Algorithm, which divides the whole dataset into K number of Clusters or Subgroup based on similarity and their mean distance from the centroid of that particular cluster formed.

Elbow Method : One of the most important steps in K-Means unsupervised Machine Learning Algorithm is to determine the optimal value of K and we can do so by using the elbow method.

K-Means Clustering

Step 1 : Using the Elbow Method, calculate the optimal value of K to choose the number of clusters.
Step 2 : Randomly initialize the K points (or say Centroids) on the datasets.
Step 3 : All data points should be assigned to their closest centroid.
Step 4 : Calculate mean value and place a new centroid to each cluster.
Step-5 : Repeat Step 3 and Step 4, till no further reassignment occurs.
Step 6 : Following are few criteria based on which we should stop K-Means Algorithm :-

a . If newly formed Centroid does not change
b . If points remain in the same cluster
c . If the Maximum number of iterations are completed



K-Nearest Neighbor(KNN)

K-NN is a supervised Machine Learning algorithm and can be used for both classification as well as for regression problems.

K-NN stands for K-Nearest Neighbors is one of the simplest supervised Machine Learning algorithms and it is mostly used for the classification.

K-NN is also called Lazy Learning Algorithm (instance based learning) because it does not learn from the training set immediately instead it stores the dataset and uses all the data for training while classification.

K-NN is a Non-parametric Learning Algorithm which means it doesn’t assume anything about the underlying data.

K-Nearest Neighbor(KNN)

K-Nearest Neighbors uses the features with similarity to predict the value of new data points. In other words we can say that, it classifies a new data point based on how its neighbors are classified.

In K-NN we are required to choose the value of K which can be any integer but mostly it is chosen as an odd number integer to avoid any confusion between two classes of data.

After choosing the value of K we calculate the K number of neighbors who are nearest to the test data using the Euclidean Distance, Manhattan Distance or Hamming Distance. Most of the time we use Euclidean Distance.

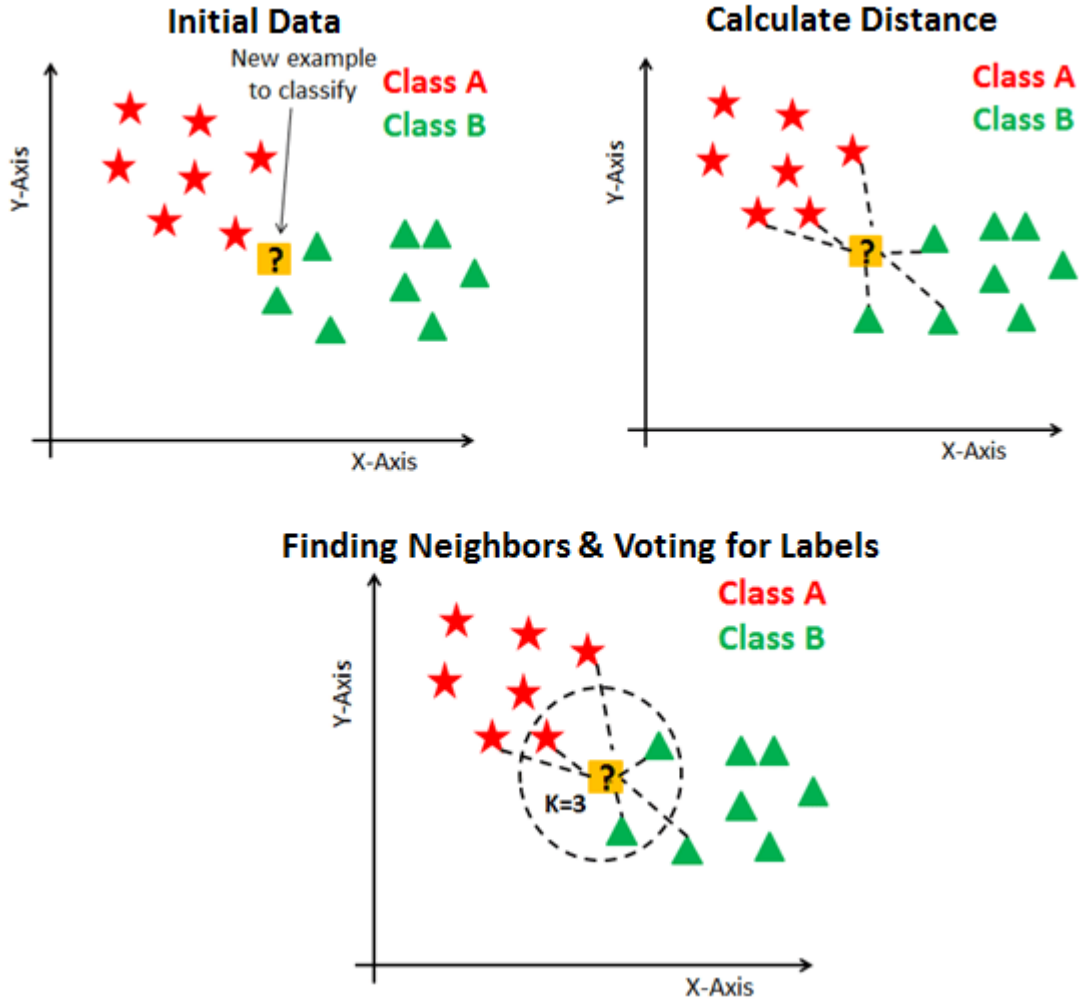
Now test data will be assigned to that category of class which is in majority

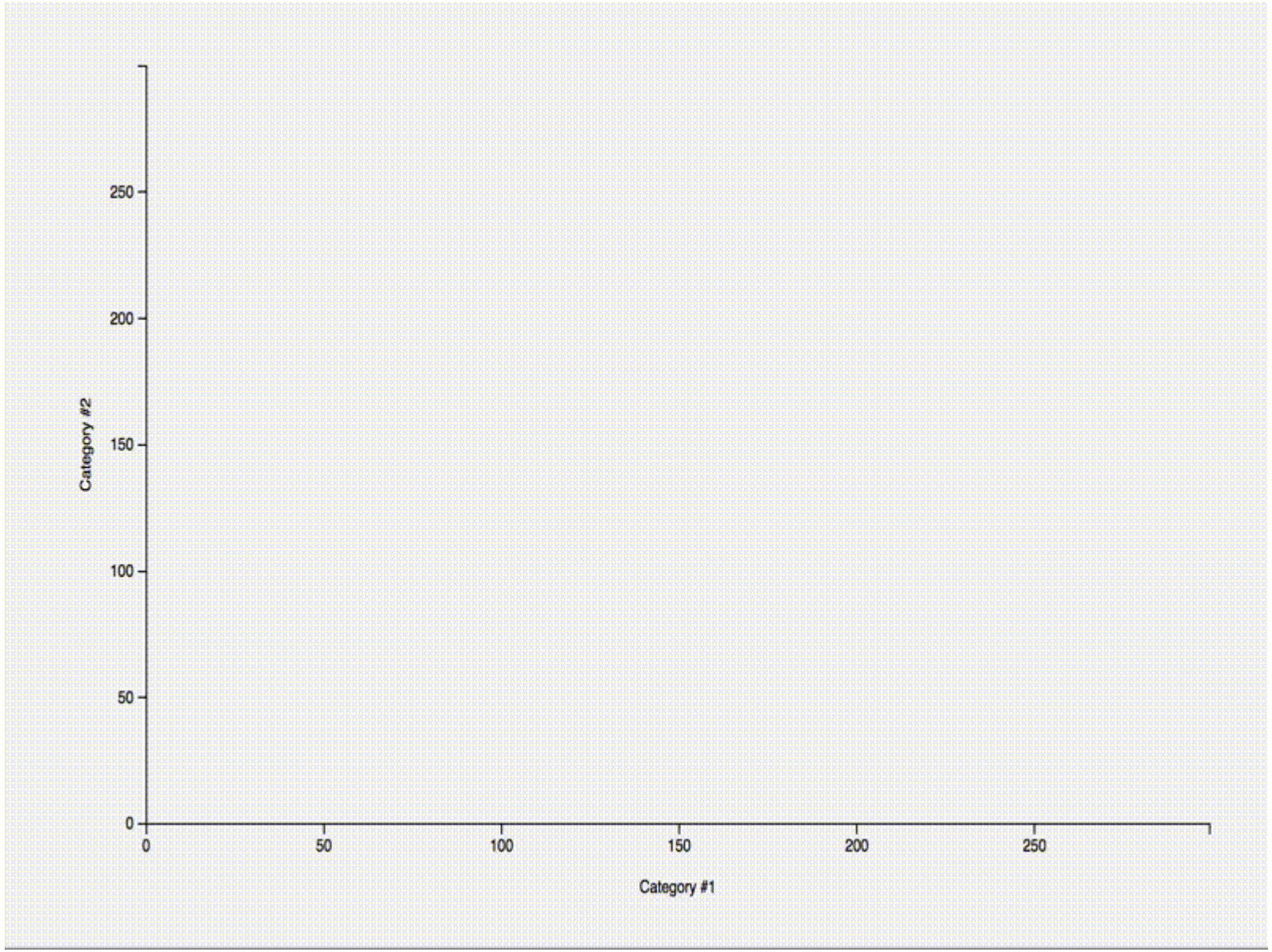
Distance Metrics in K-NN

Following are two major distance metric used in K-NN :-

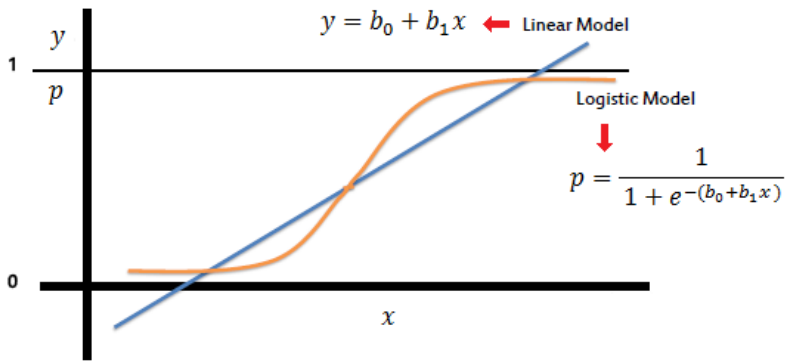
Euclidean Distance :- It is used to represent the shortest distance between the two points.

Manhattan Distance :- It is used to represent the sum of absolute differences between points across all the dimensions.




250
200
150
100
50
0
Category #2
0
50
100
150
200
250
Category #1

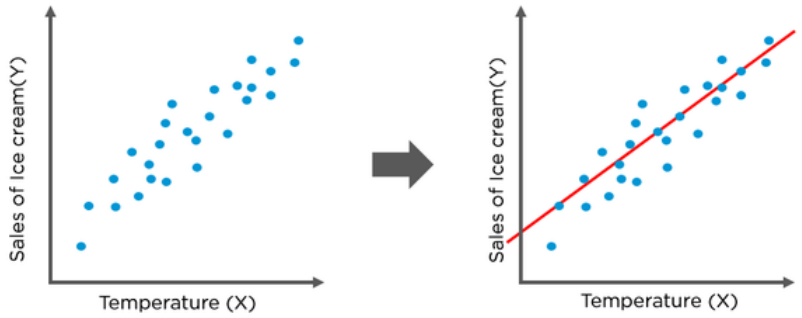
Logistic Regression



Logistic Regression in a Supervised Machine Learning algorithm, which is used where the response variable is categorical. Logistic Regression make the use of logit function, which helps to find the relationship between independent variable and the dependent variables by predicting the probability of their outcome.

Logistic regression uses the sigmoid function which will restrict the value of logistic regression between 0 and 1.

Linear Regression



Linear Regression is basic and most widely used type of predictive analysis.
Linear Regression is one of the simplest supervised machine learning algorithm which helps to find the relationship between one or more independent variables (predictors) denoted as X and the dependent variables (target) denoted as y.
By minimizing the distance (or say error) between all the data points and regression line we can find the best fit line for our dataset.
Our main aim is to minimize the cost function by updating the different values of θ.

Support Vector Machine

SVMs are one of the most robust prediction methods. It is mostly used for Classification problems.

Support Vector Machine is a Supervised Machine Learning Algorithm and it can be used for both Classification as well as for Regression.

The main objective of SVM is to create the best decision boundary (also called Hyperplane) that can separate n-dimension space into classes so that we can easily put the new data point into the correct category in the future.

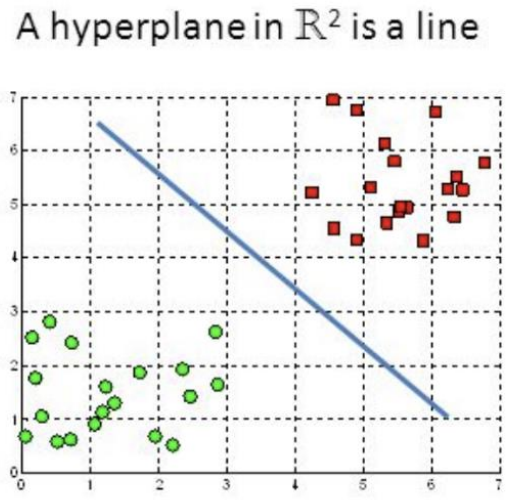
SVM chooses the extreme points (or say vector) that help in creating the hyperplane. These extreme cases are called support vectors and hence the algorithm is termed as Support Vector Machine.

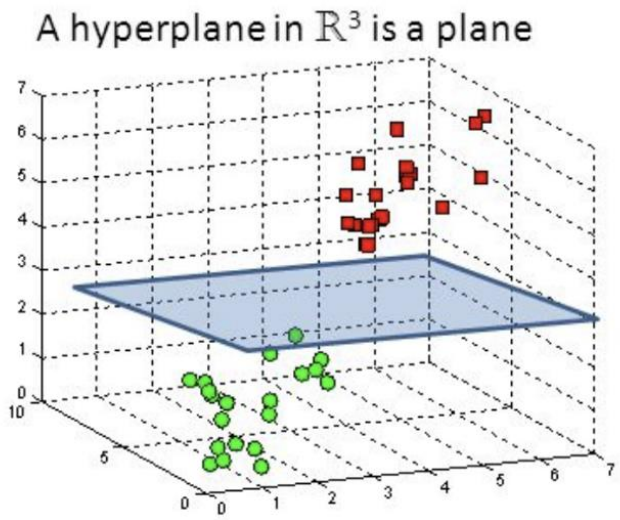
Support Vector Machine

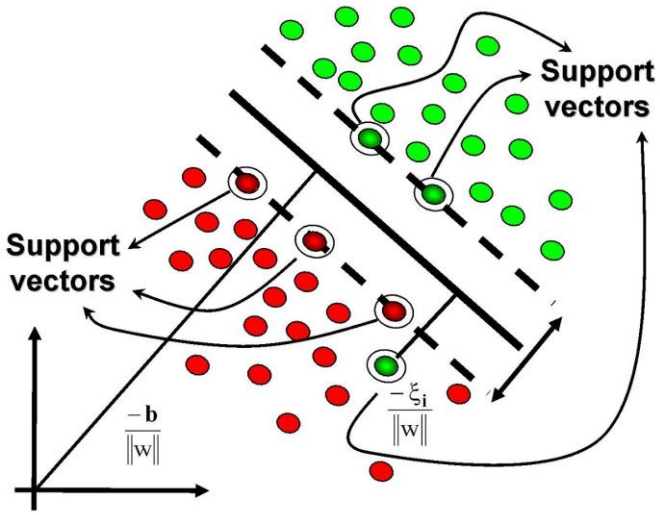
Support Vector :- Support vector is a data point which is located closest to the hyperplane. By using these support vectors, we maximize the margin of the classifier.

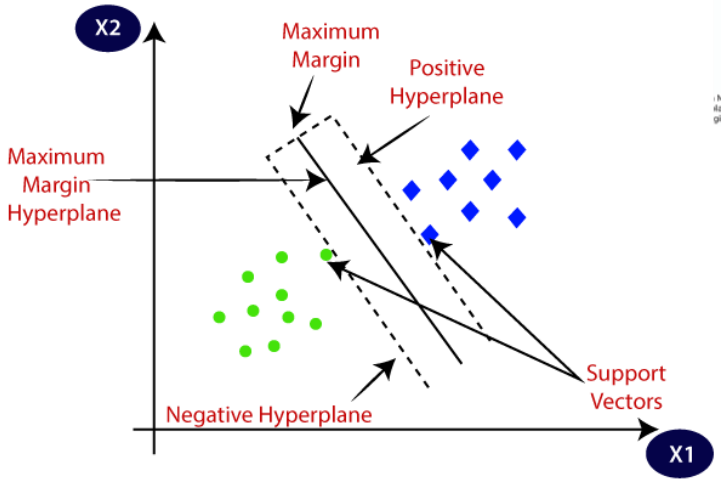
Hyperplane :- It can be defined as the best decision boundary that helps to classify the data points.

Margin :- The distance of the vectors from the hyperplane is called the Margin.









Naive Bayes Algorithm

Naive Bayes is a supervised machine learning algorithm used for classification problems. It is based on Bayes Theorem.

It is a simple mathematical formula and is also known as the Bayes Rule or Bayes Law. It is used for the calculation of conditional probabilities.

Conditional probability :- It can be defined as the probability of an event occurring given that another event has already occurred.

Naive Bayes Theorem assumes that each feature makes an :

equal

independent

contribution to the outcome.

Naive Bayes Algorithm

Bayes Law

Likelihood of B given A is True

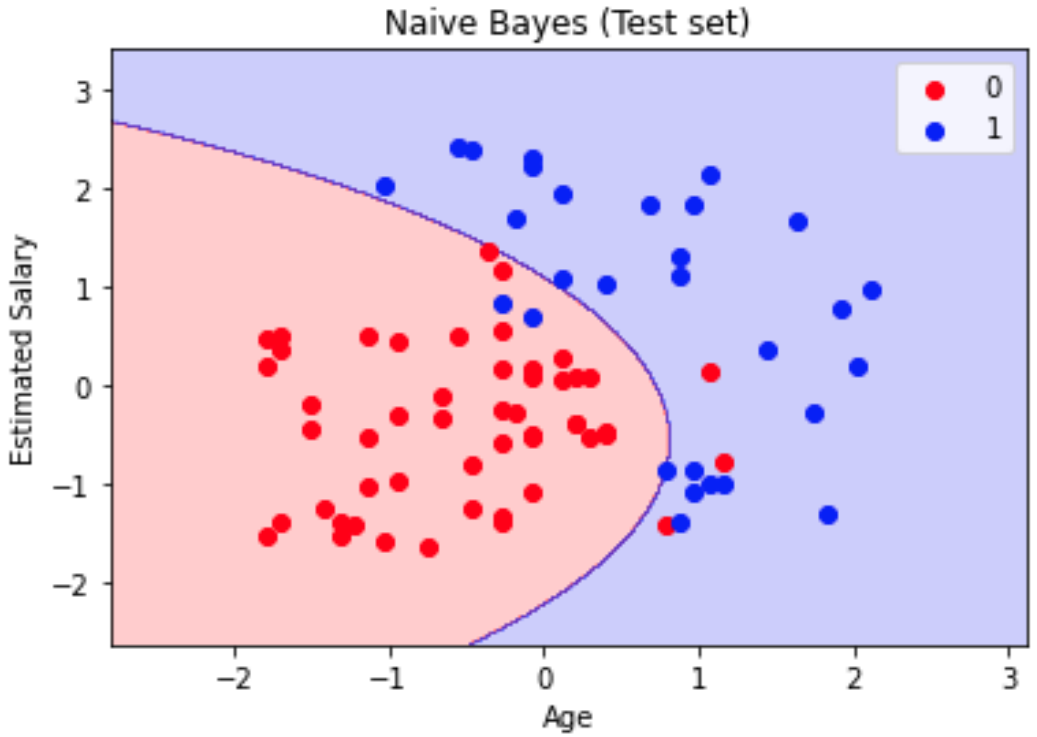
Prior Probability of A

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

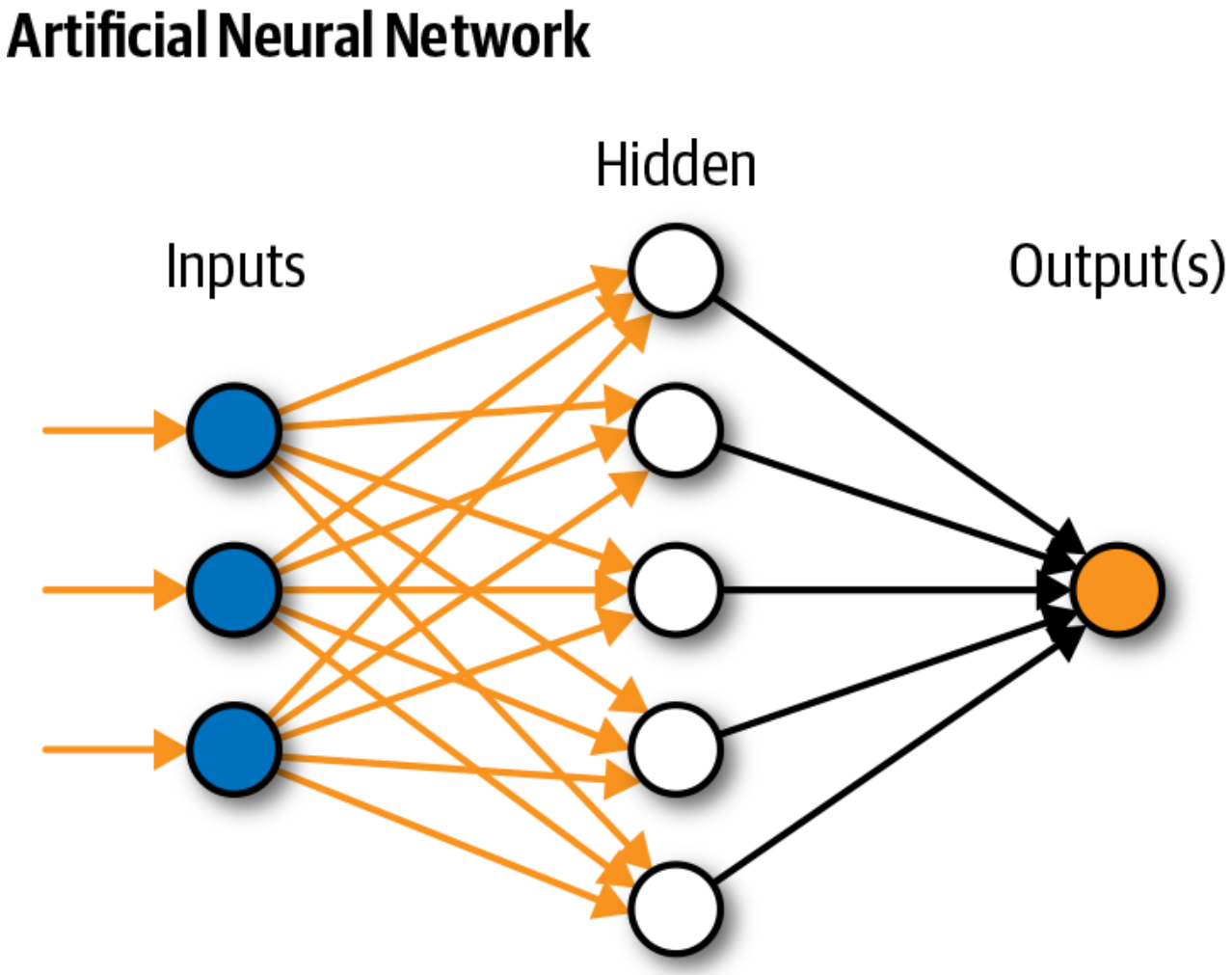
Posterior Probability Of A given B is True

Prior Probability that B is True

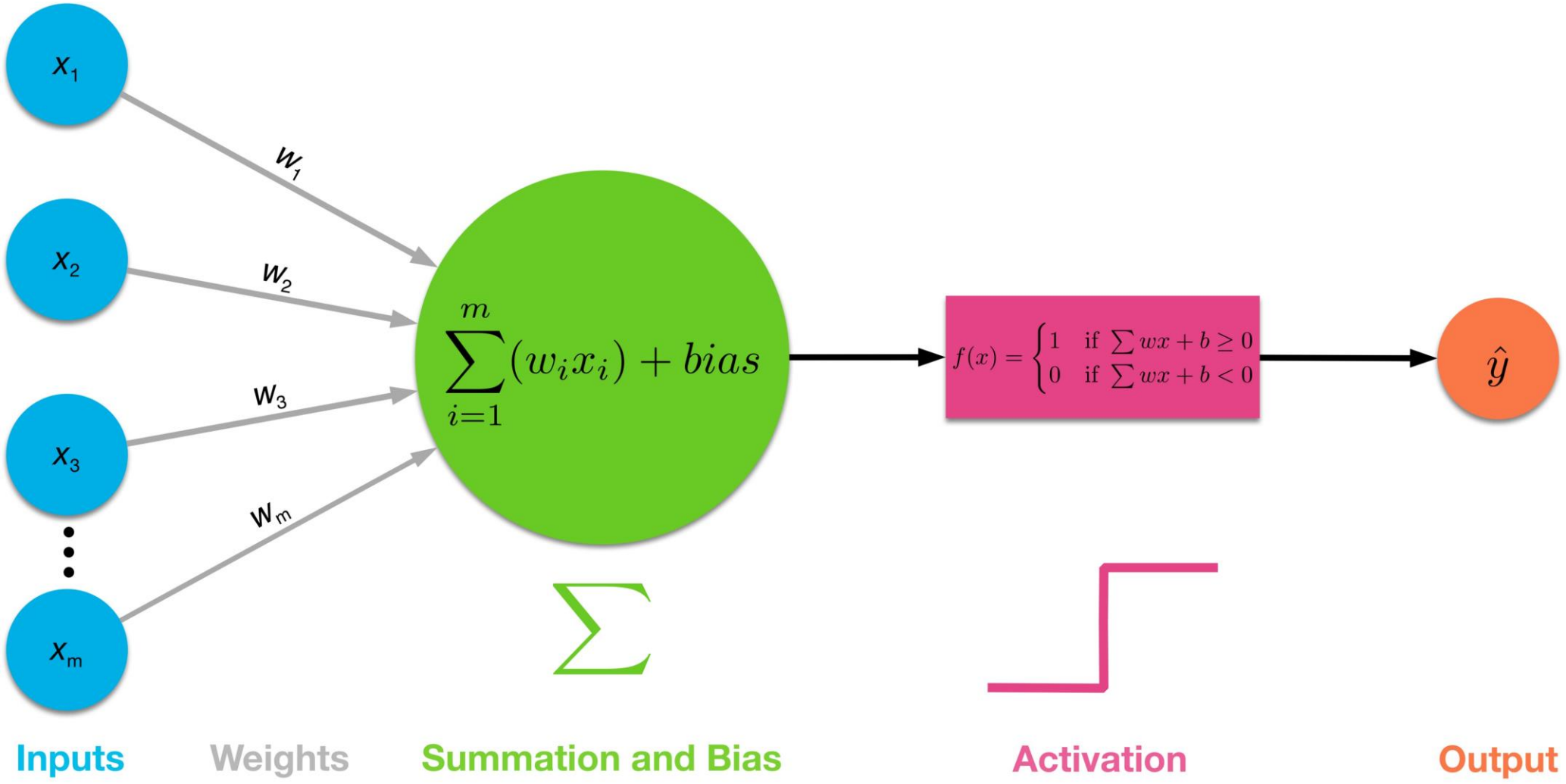
Naive Bayes Classifier



Artificial Neural Network ANN



Artificial Neural Network ANN



Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$