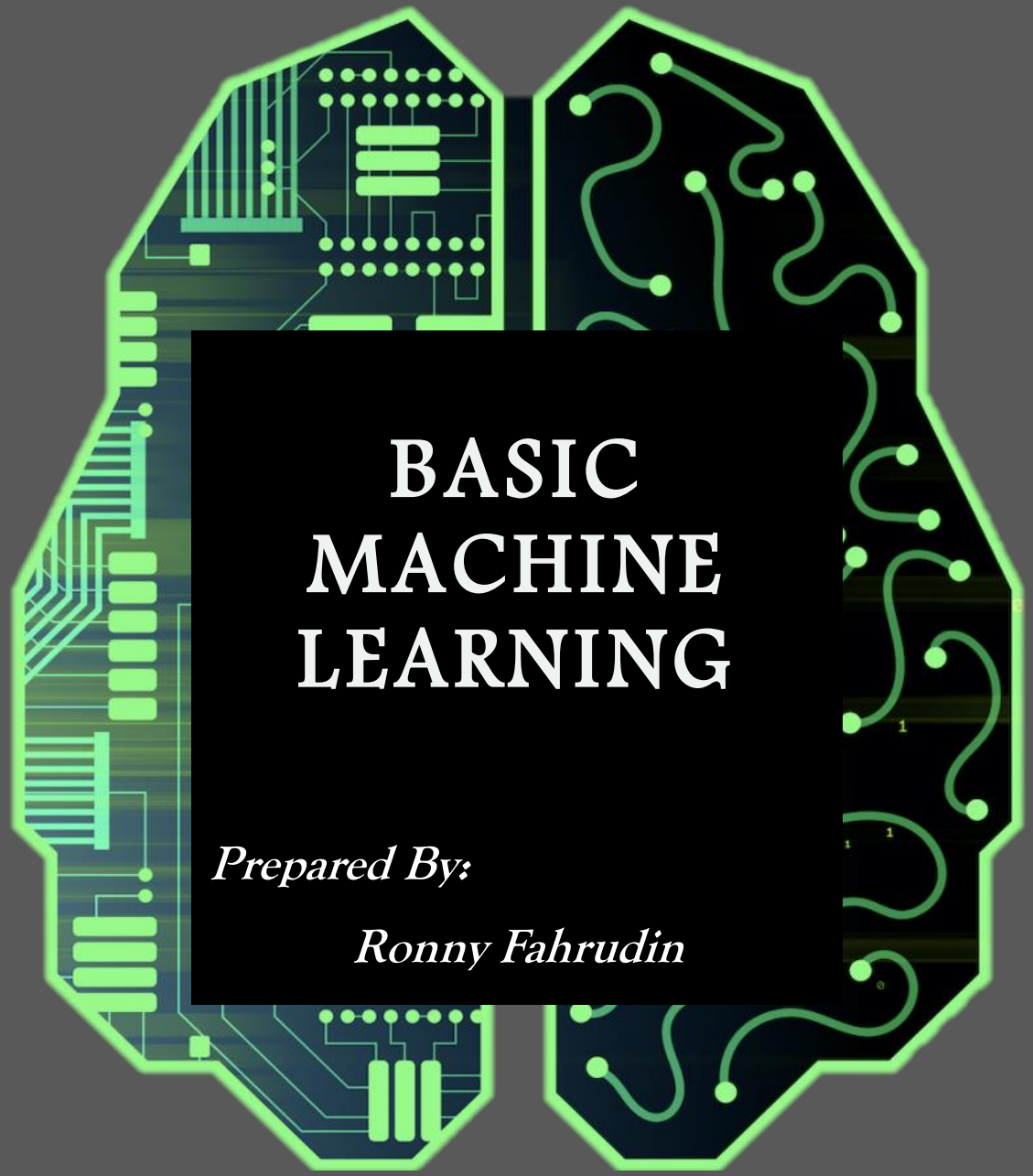


BASIC MACHINE LEARNING

Prepared By:

Ronny Fahrudin

OUTLINE

Machine Learning

What is machine learning?

Why use machine learning?

Machine Learning types Algorithm

Supervised Learning

What is Supervised learning?

When we use Supervised Learning?

Function of Supervised Learning

Types of Supervised Learning Algorithm

Unsupervised Learning

What is Unsupervised learning?

When we use Unsupervised learning?

Function of Unsupervised learning

Types of Unsupervised learning Algorithm

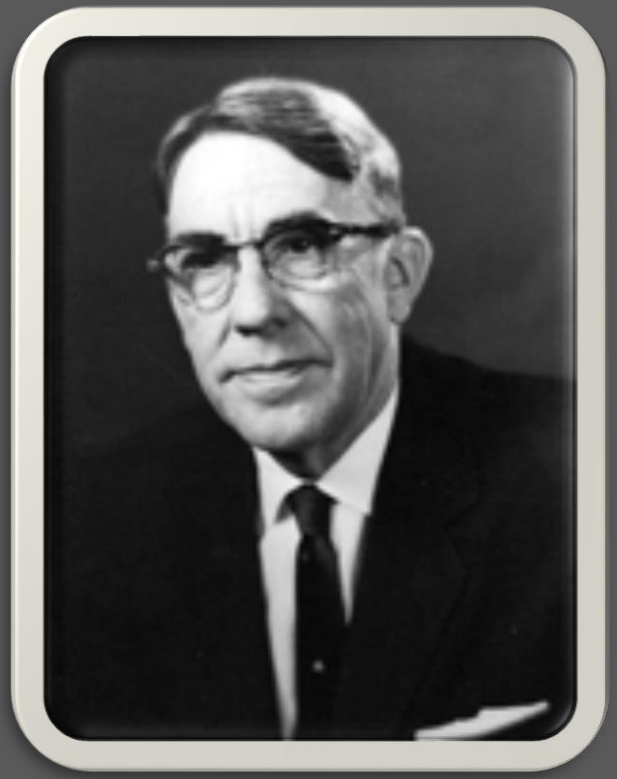
Reinforcement Learning

What is Reinforcement learning?

Function of Reinforcement learning

MACHINE LEARNING

WHAT IS MACHINE LEARNING?



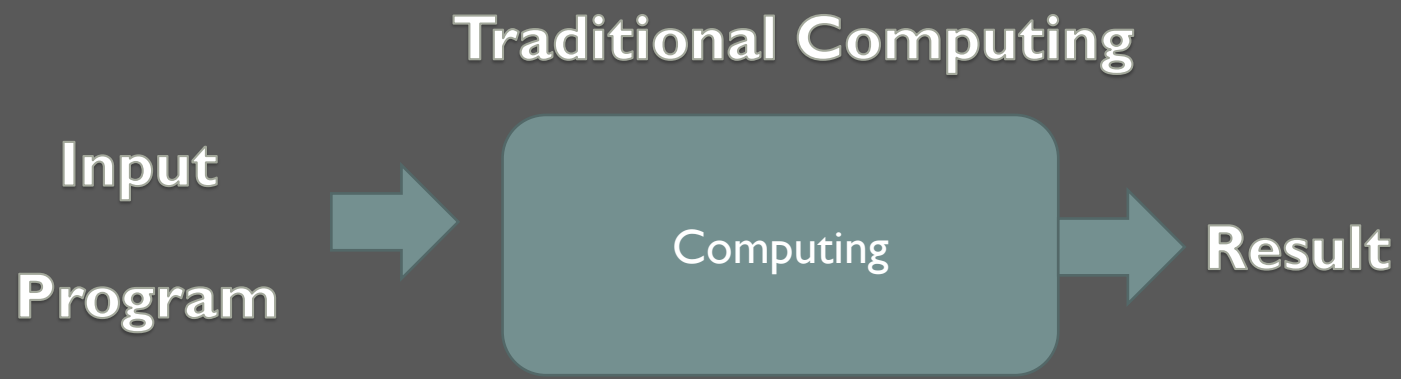
Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

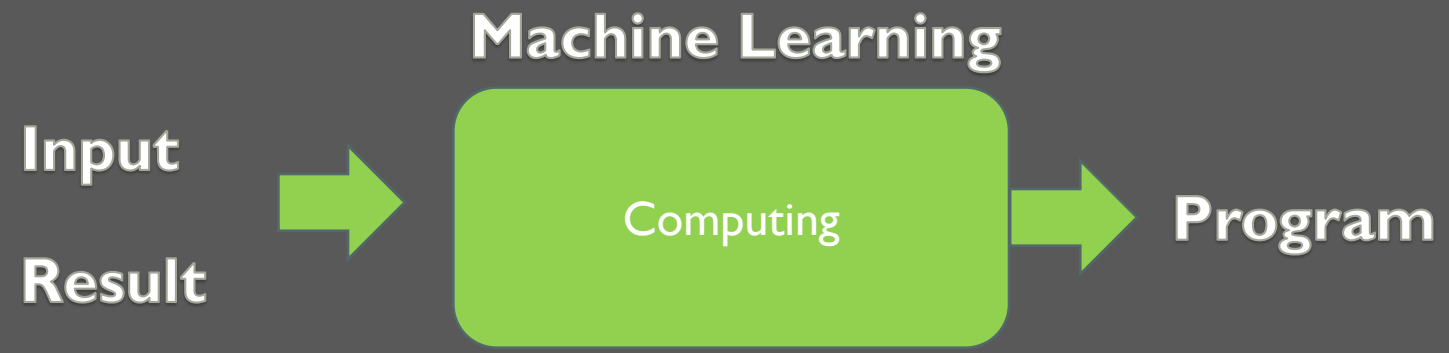
(Arthur Samuel, 1959)

WHY USE MACHINE LEARNING?

1. One Machine learning algorithm can often simplify code and perform better
2. The best machine learning technique can find a solution from complex problem
3. Machine learning system can adapt to new data.
4. Getting insight about complex problems and large amounts of data.

MACHINE LEARNING VS TRADITIONAL COMPUTING





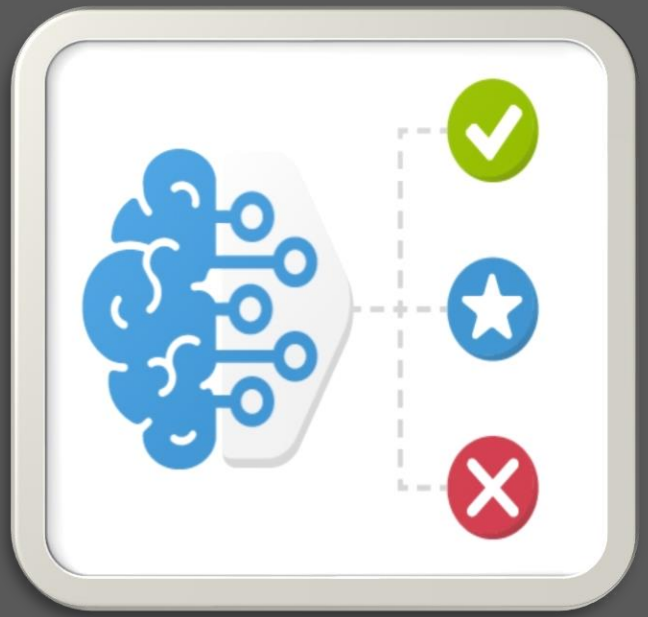
MACHINE LEARNING ALGORITHMS

	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

SUPERVISED LEARNING



WHAT IS SUPERVISED LEARNING?

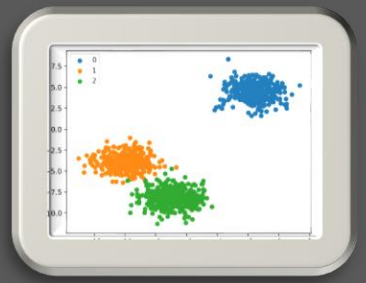
Supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer.

WHEN WE USE SUPERVISED LEARNING?

1. We have labeled datasets
2. We want to make classification modeling
3. We want to make regression modeling

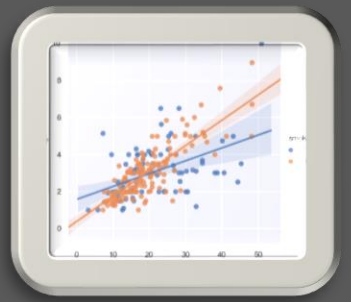
FUNCTION OF SUPERVISED LEARNING

Classifications



- Fraud Detection
- Email Spam Detection
- Diagnostics
- Image Classification

Regression



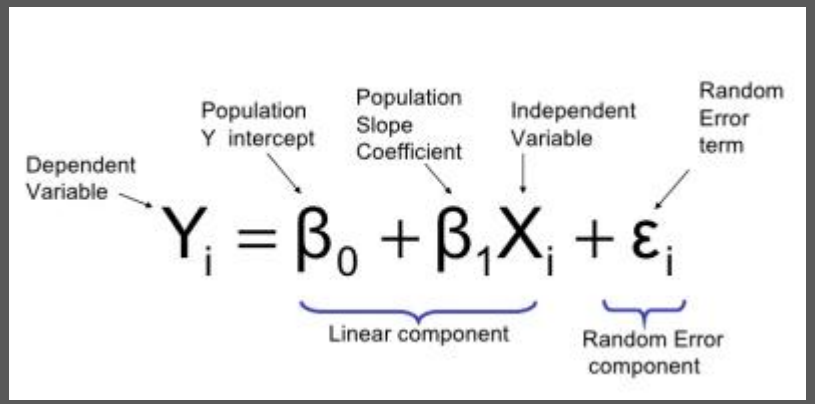
- Risk Assessment
- Score Predictions

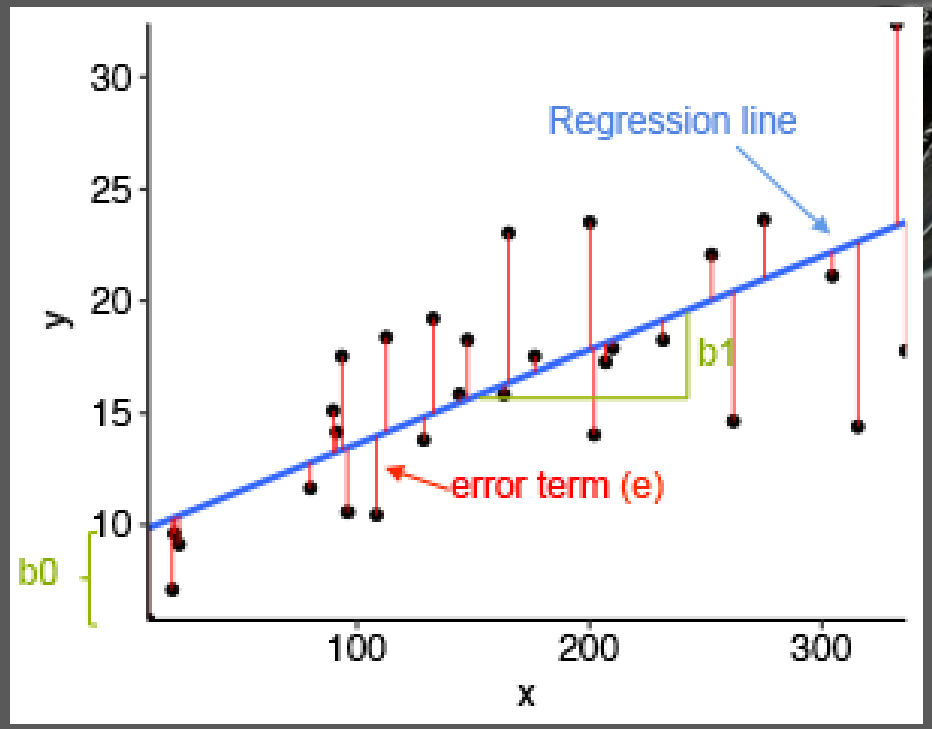
DIFFERENCE BETWEEN CLASSIFICATION & REGRESSION

Classification	Regression
Input variable are discrete	Input variable are continuous
Outcome are class	Outcome are real numbers
To predict of class in future	To predict score in future
Calculation method with measure accuracy	Calculation method with root mean square score

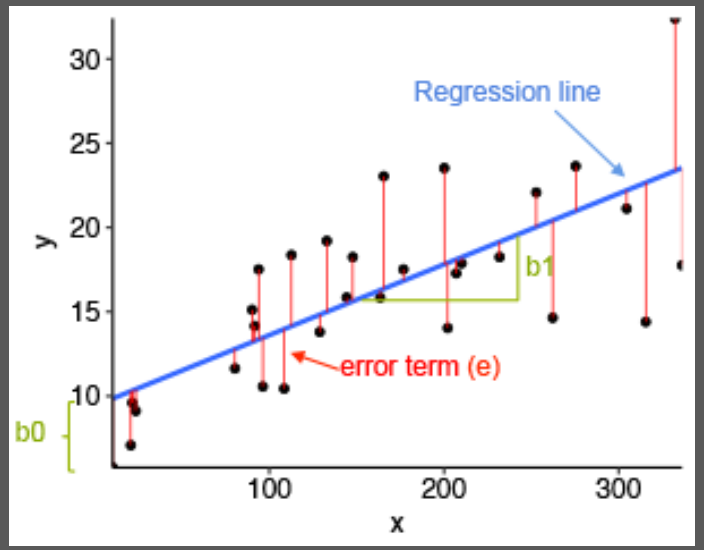
LINEAR REGRESSION

Linear Regression is a linear approach to modelling the relationship between a scalar response(dependent variables) and one or more independent variables.





LINEAR REGRESSION

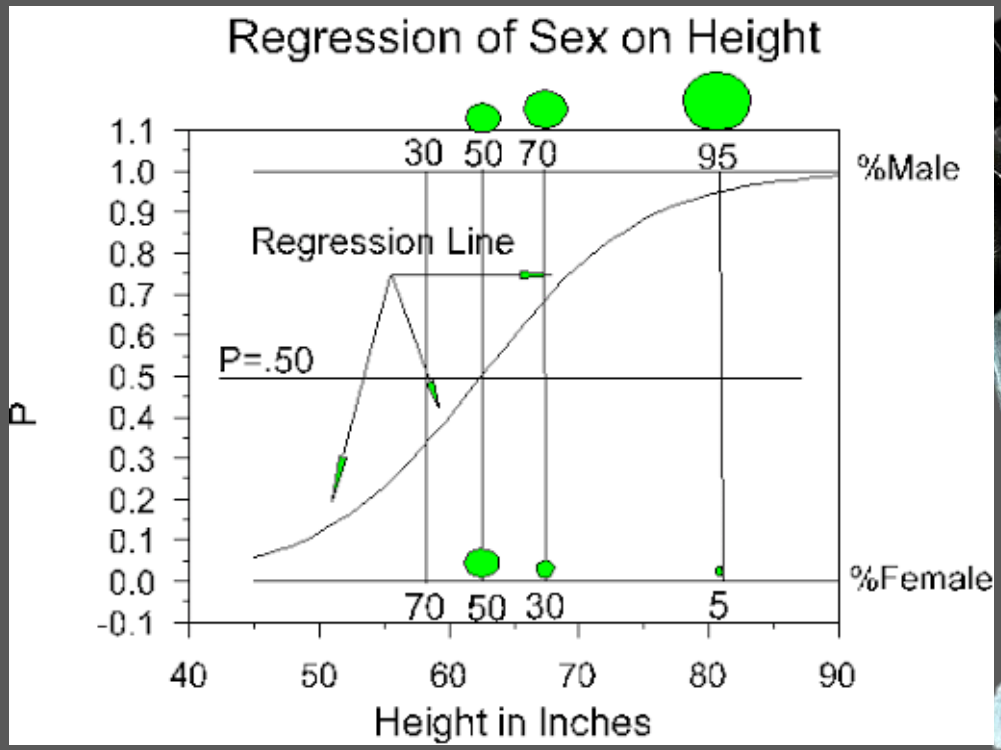


Simple Linear Regression (Bivariate Analysis)

Multiple Linear Regression (Multivariate Analysis)

LOGISTICS REGRESSION

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes.

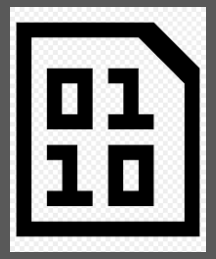


Used to:

1. Classify email spam or not
2. Online transactions fraud or not fraud
3. Tumor Malignant or Benign

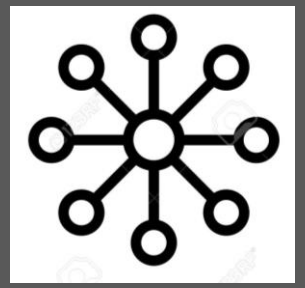
LOGISTICS REGRESSION TYPES

Binomial



- Two possible types of the dependent
- Ex: Pass or Fail

Multinomial



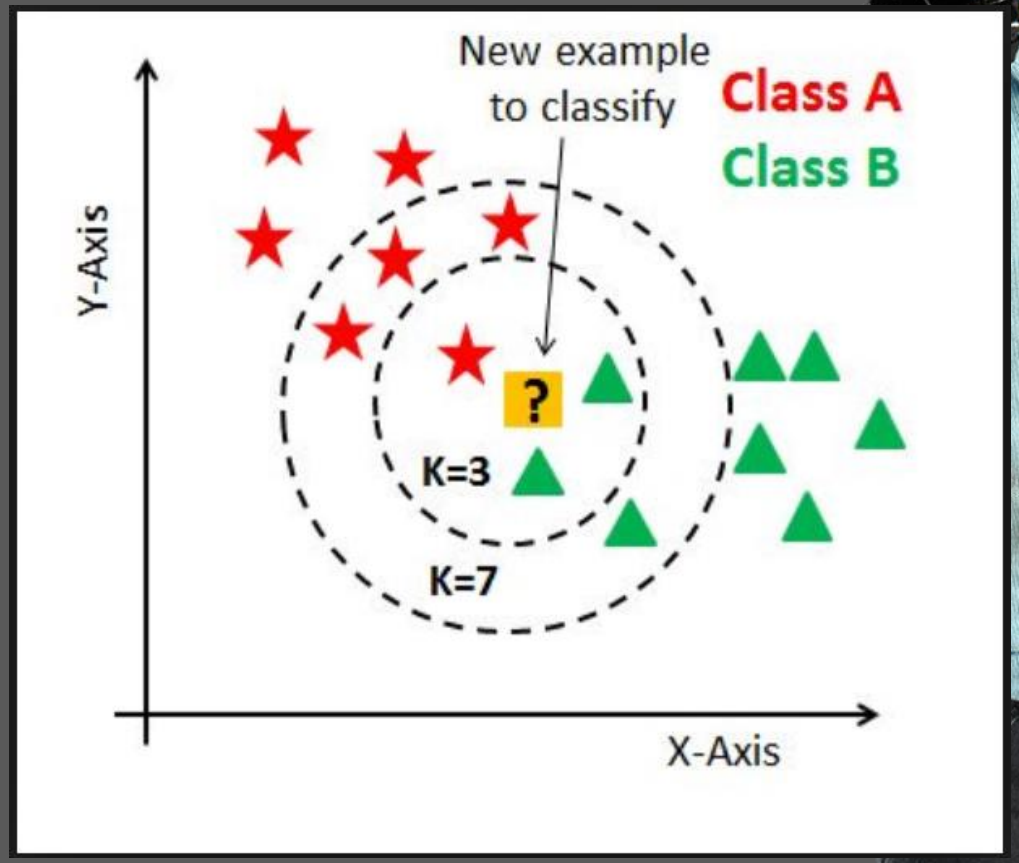
- 3 or more possible unordered types of the dependent
- Ex: Dog, Cat, or Sheep

Ordinal

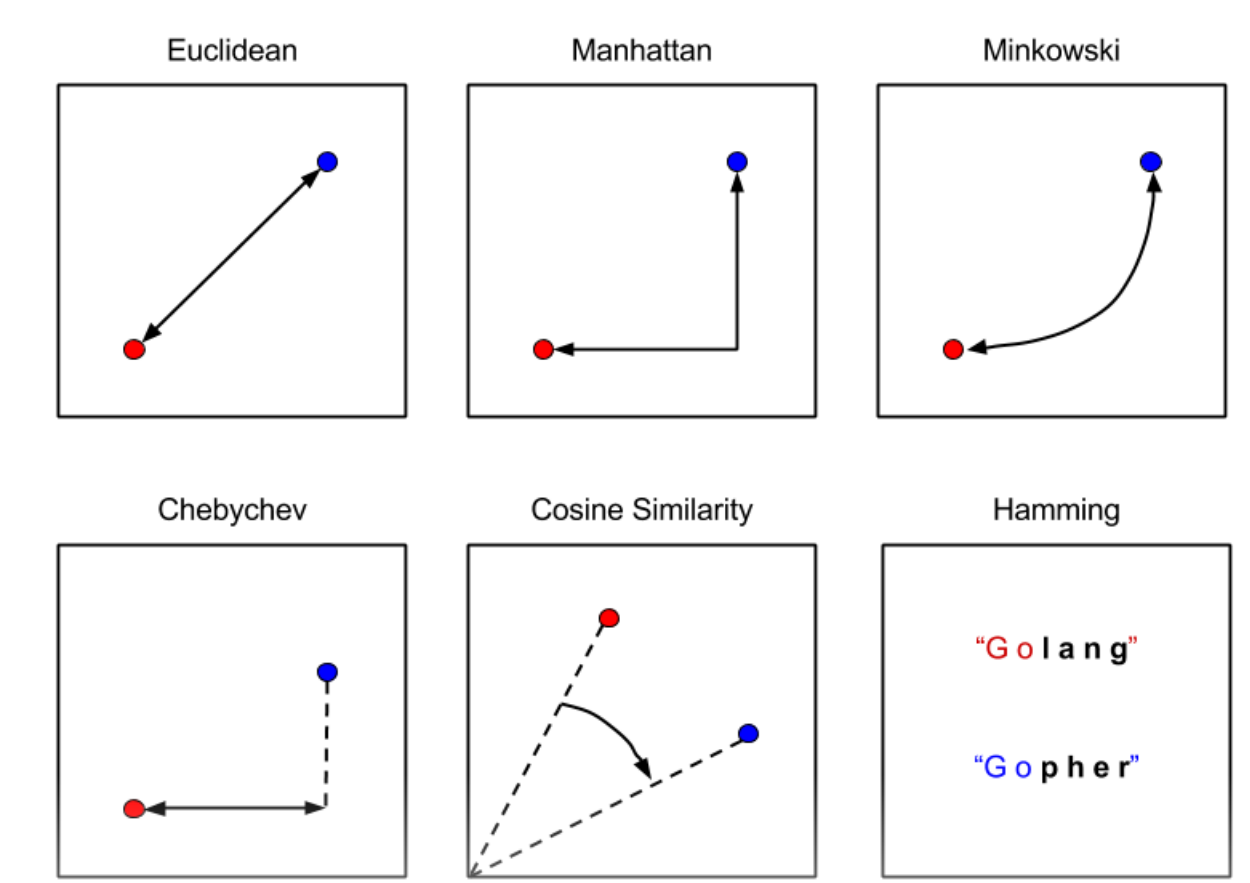
- 3 or more possible ordered types of the dependent
- Ex: rank 1, rank 2, rank 3, rank 4

K-NEAREST NEIGHBOR

- K – Nearest Neighbors is a non parametric method to classification and regression.
- In classification, it can output class membership.
- In regression, it can output the value of average k-nearest neighbors score.

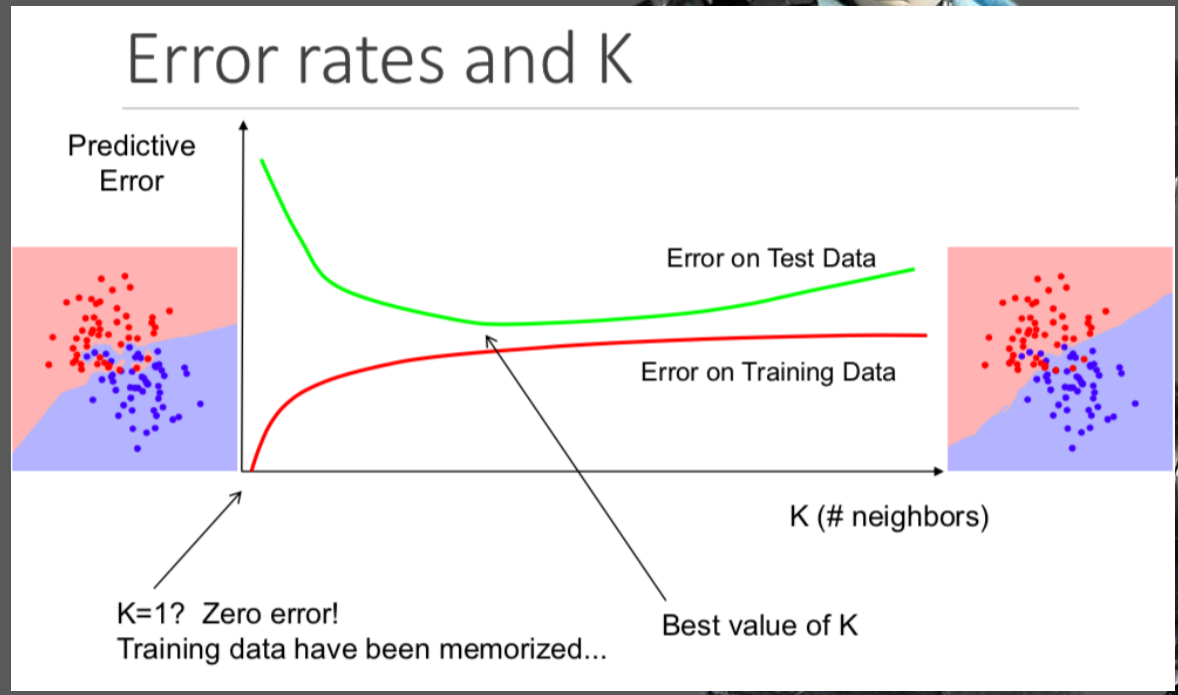


DISTANCE MEASURE K-NN



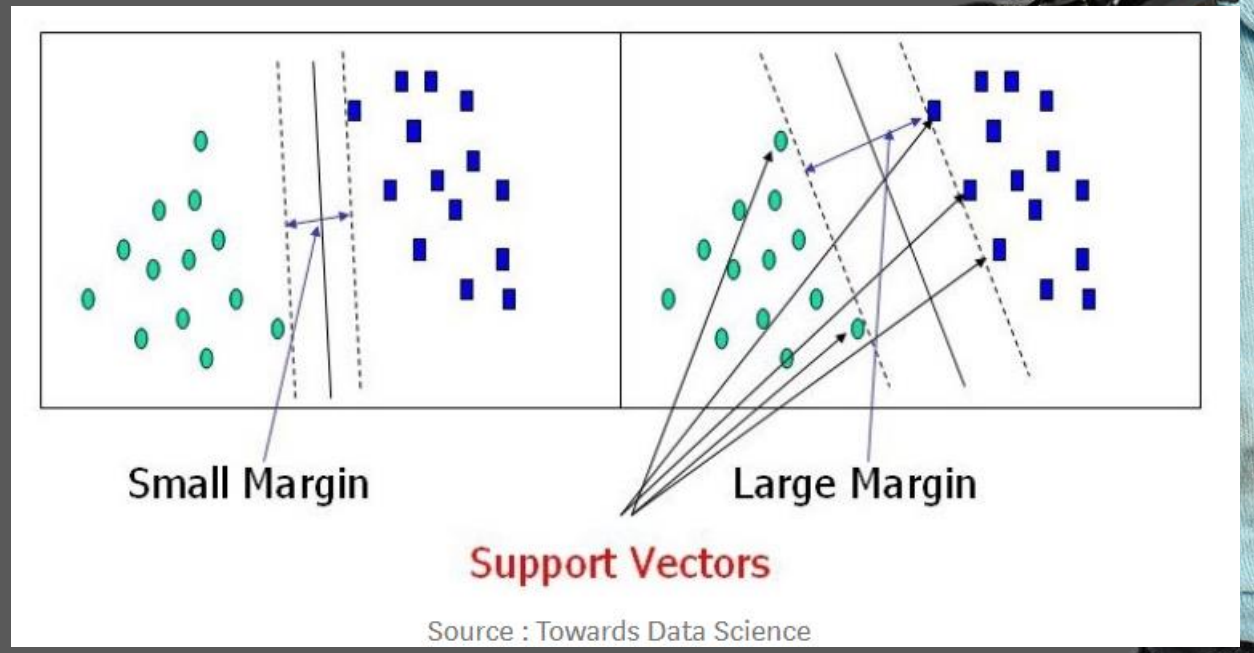
HOW TO CHOOSE VALUE K?

- There are no pre-defined statistical methods to find the most favorable value of K.
- Initialize a random K value and start computing.
- Choosing a small value of K leads to unstable decision boundaries.
- The substantial K value is better for classification as it leads to smoothening the decision boundaries.
- **Derive a plot between error rate and K denoting values in a defined range. Then choose the K value as having a minimum error rate.**



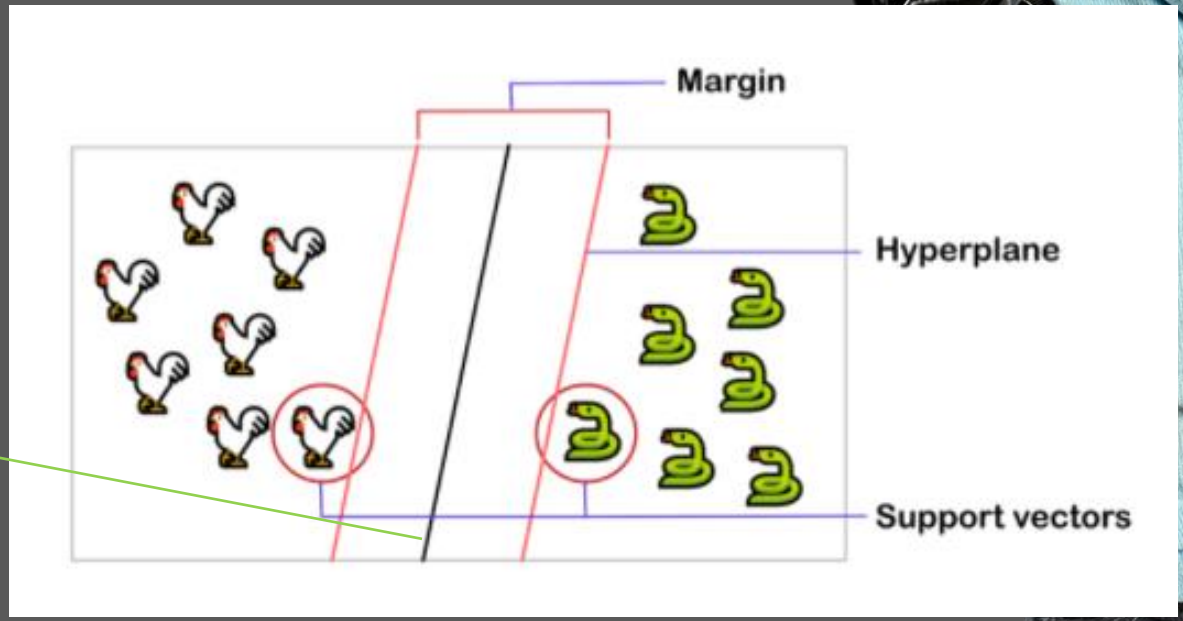
SUPPORT VECTOR MACHINE

- Support vector machines is a relative simple supervised machine learning algorithm used for classification and/or regression.



SUPPORT VECTOR MACHINE

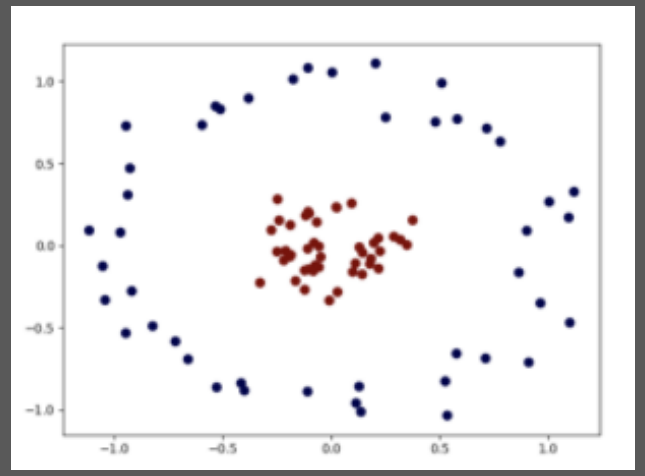
Separated by decision boundary



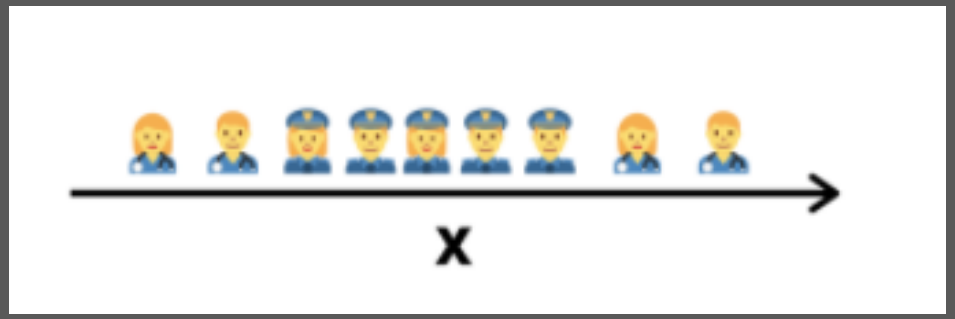
SUPPORT VECTOR MACHINE

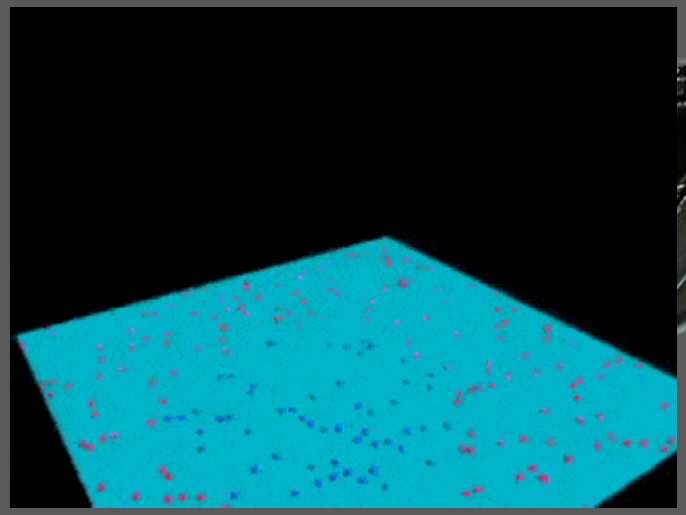
SVM Classification

non linier

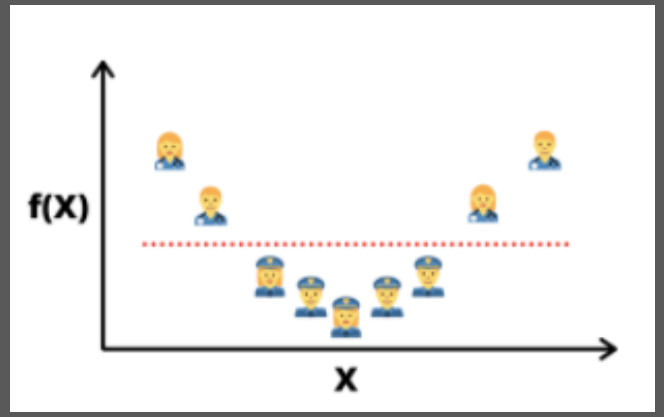


2d non linear

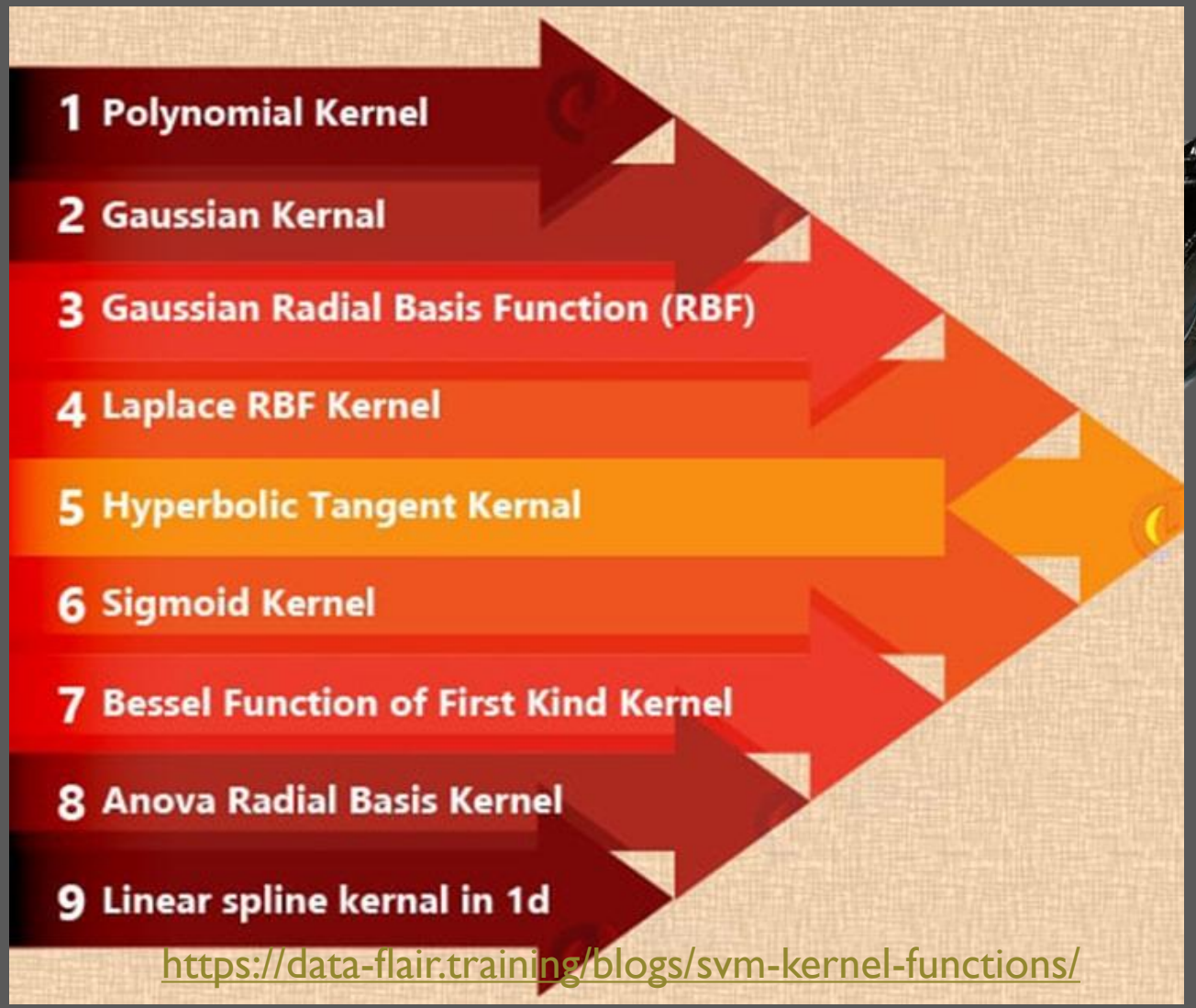




3d kernel trick



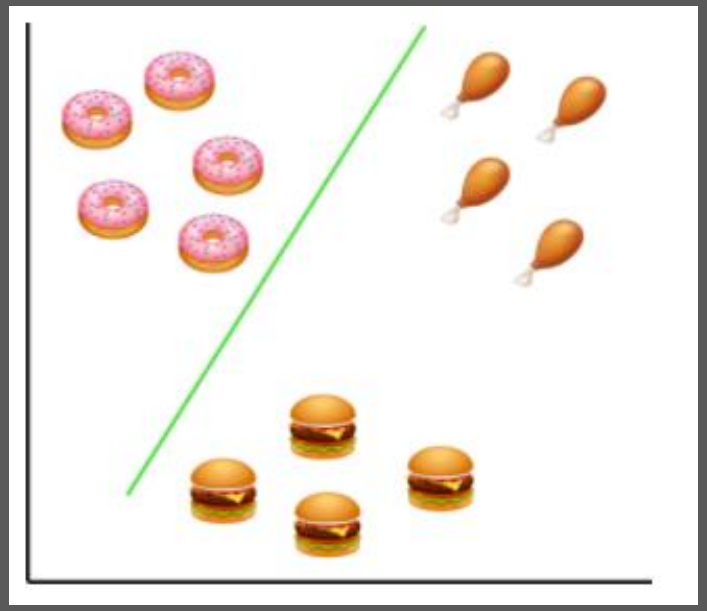
SUPPORT VECTOR MACHINE



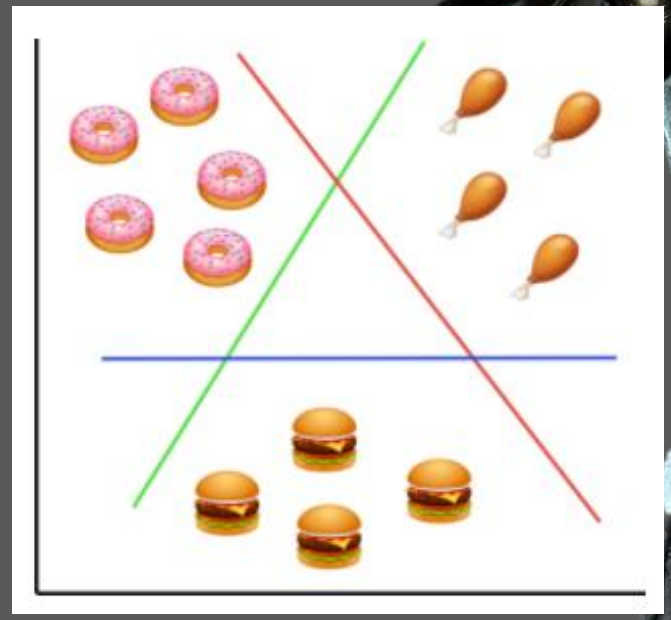
https://data-flair.training/blogs/svm-kernel-functions/

SUPPORT VECTOR MACHINE

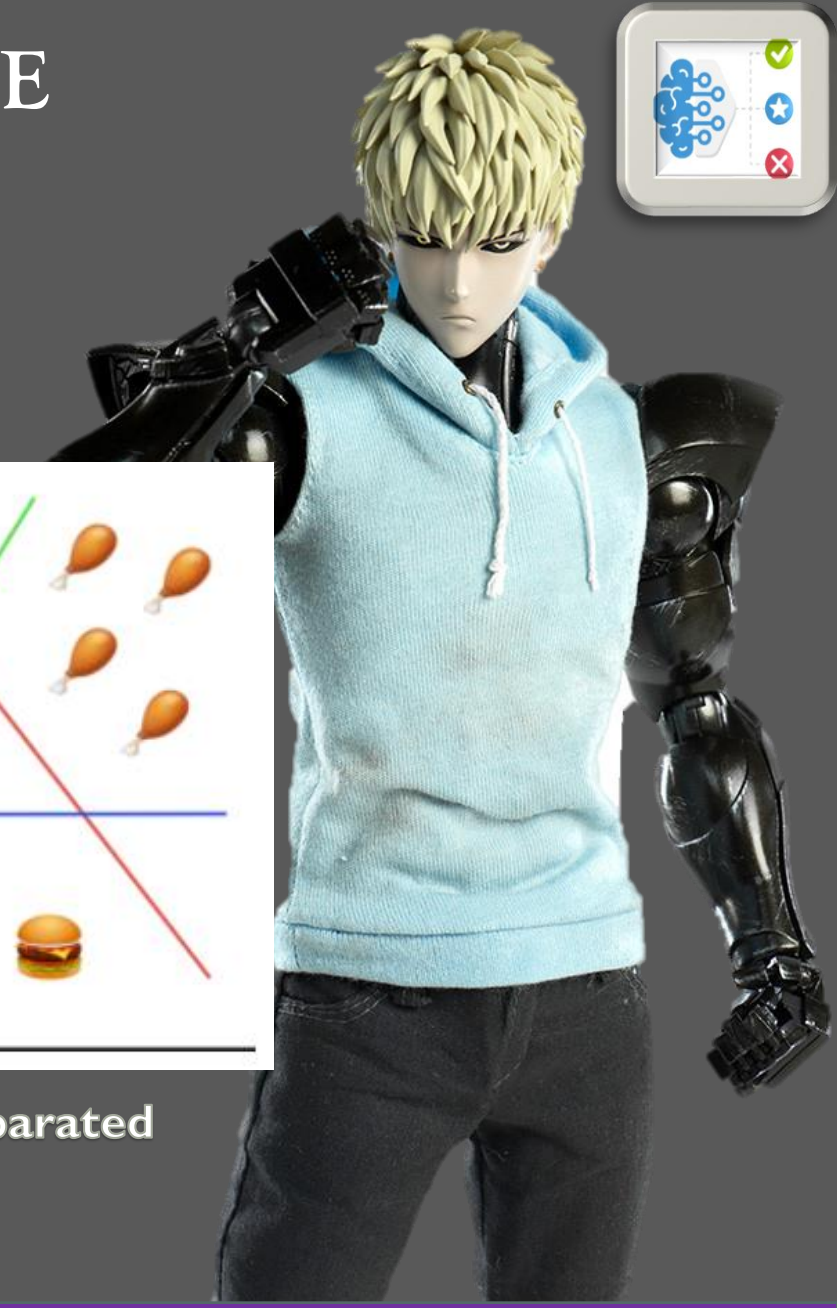
SVM Multi Classification



One separated

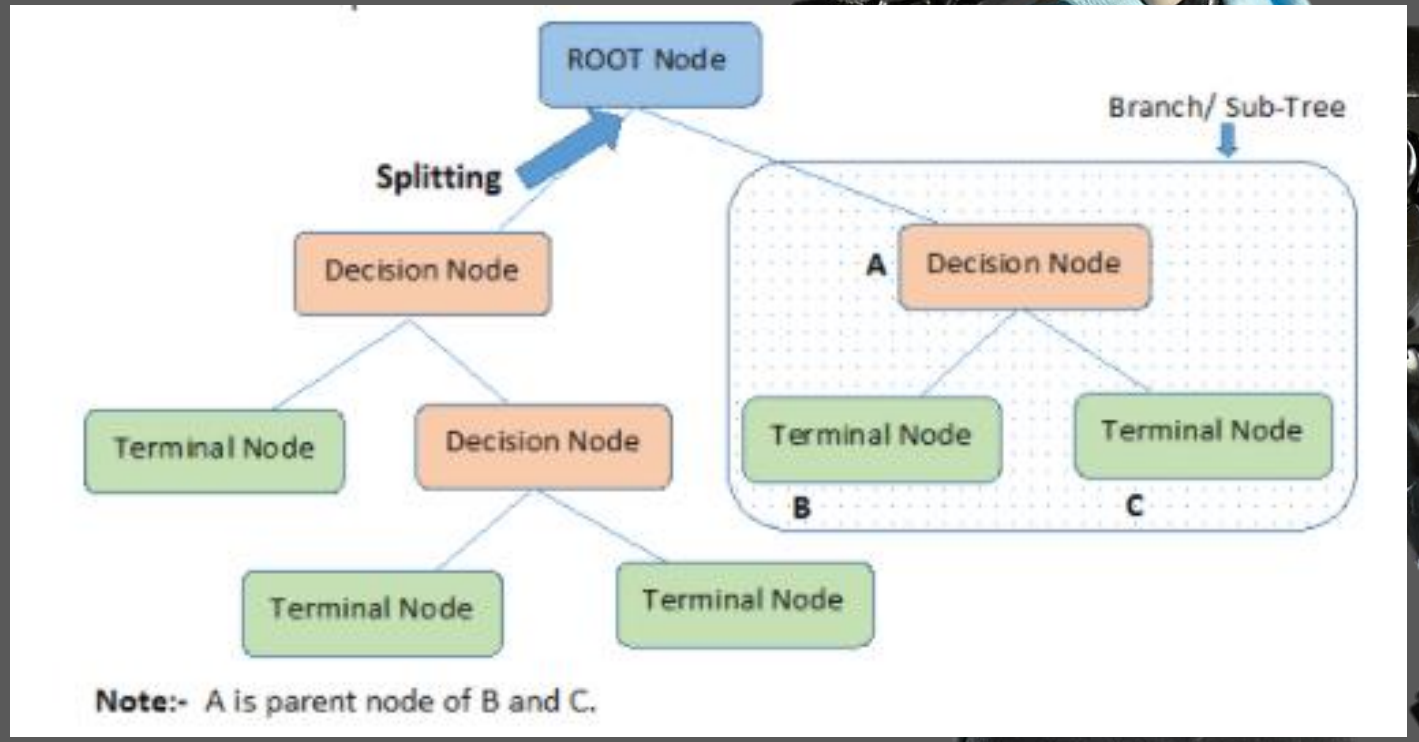


3/Multi separated



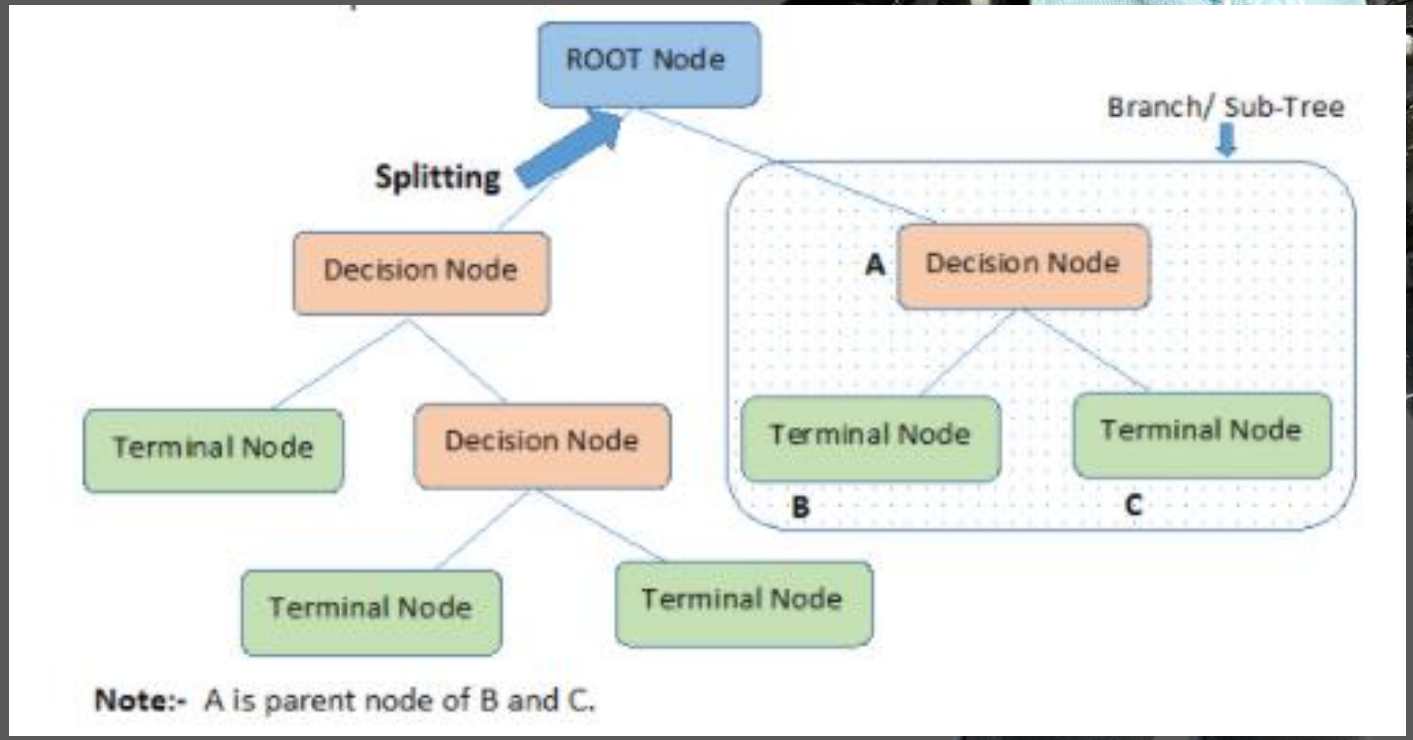
DECISION TREE

Decision Trees are versatile supervised learning algorithms that can perform classification and regression tasks, and even multioutput tasks. It is capable of fitting complex datasets.



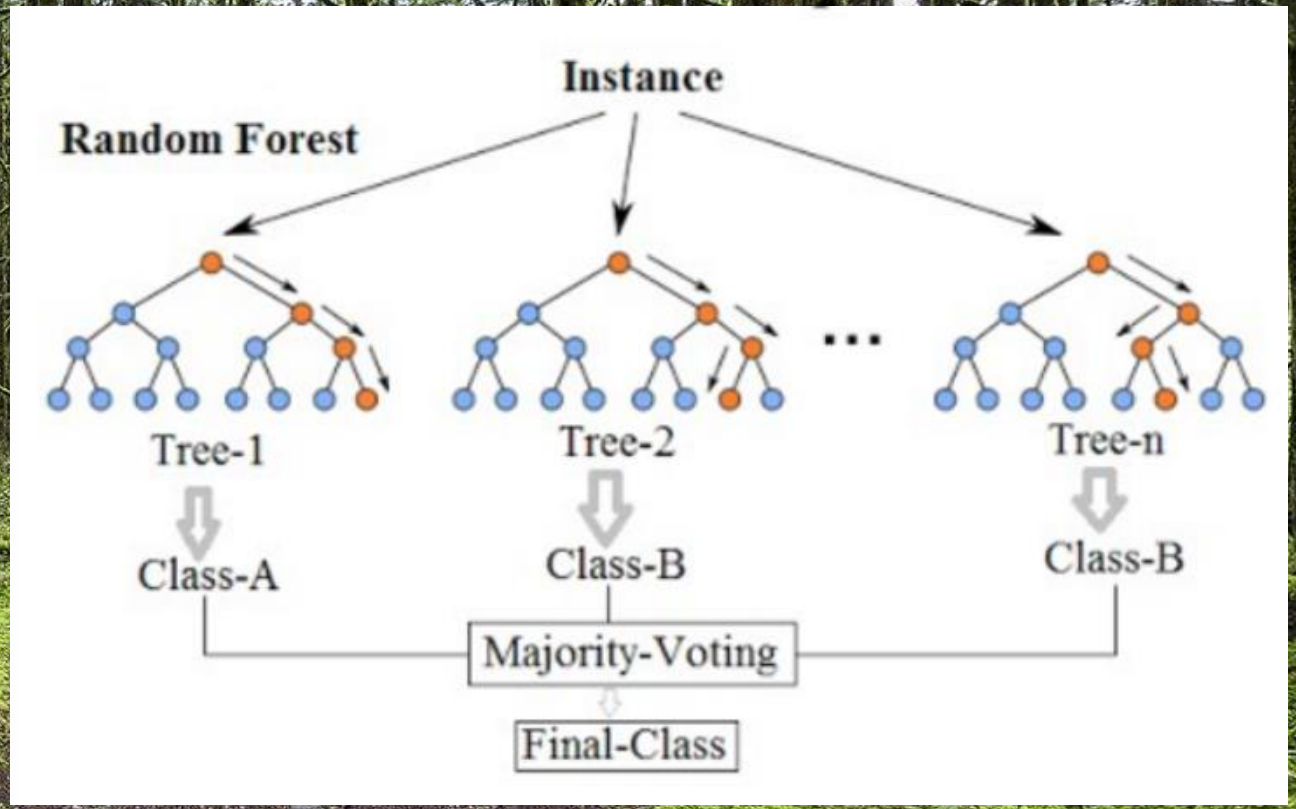
DECISION TREE

Decision Trees are versatile supervised learning algorithms that can perform classification and regression tasks, and even multioutput tasks. It is capable of fitting complex datasets.



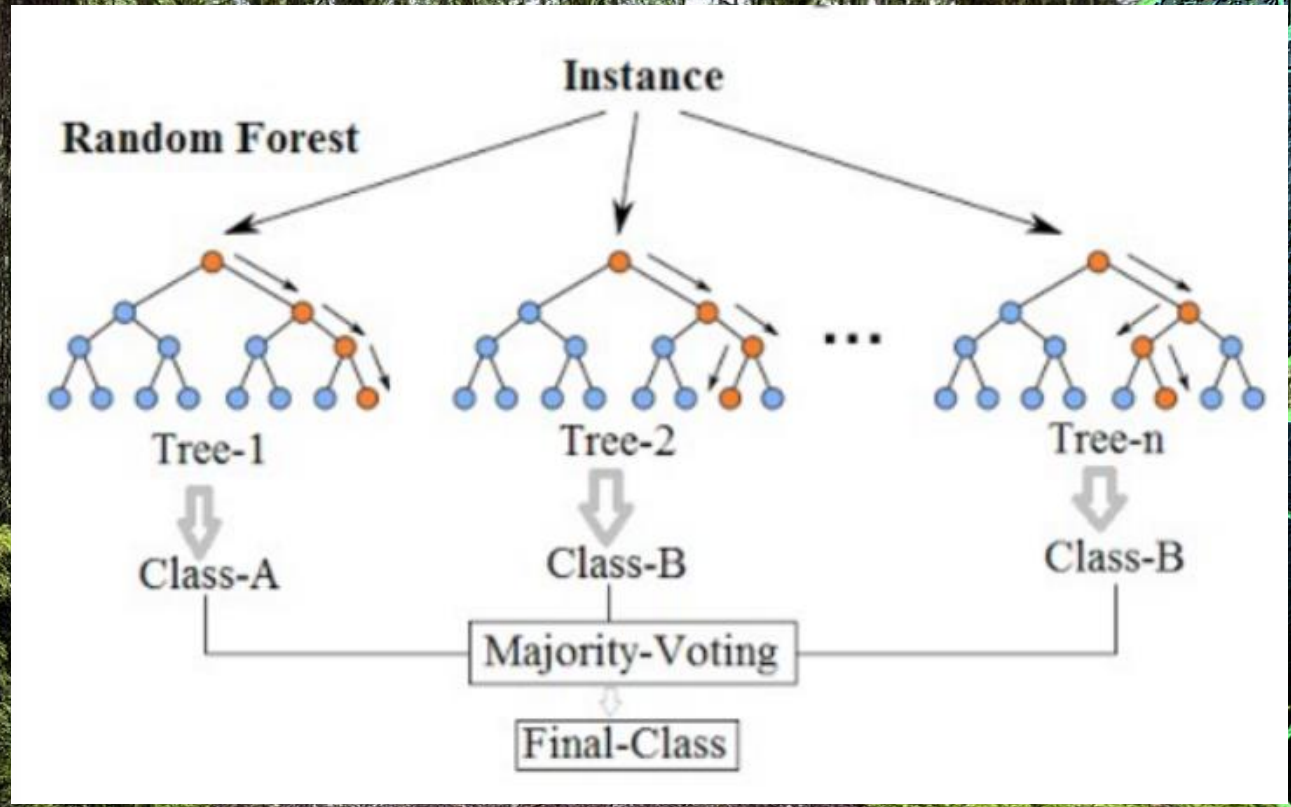
RANDOM FOREST

- Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the bagging method.



WHAT IS BAGGING ?

Bagging or Bootstrap Aggregation is a simple and very powerful ensemble method(combine with multiple algorithm).

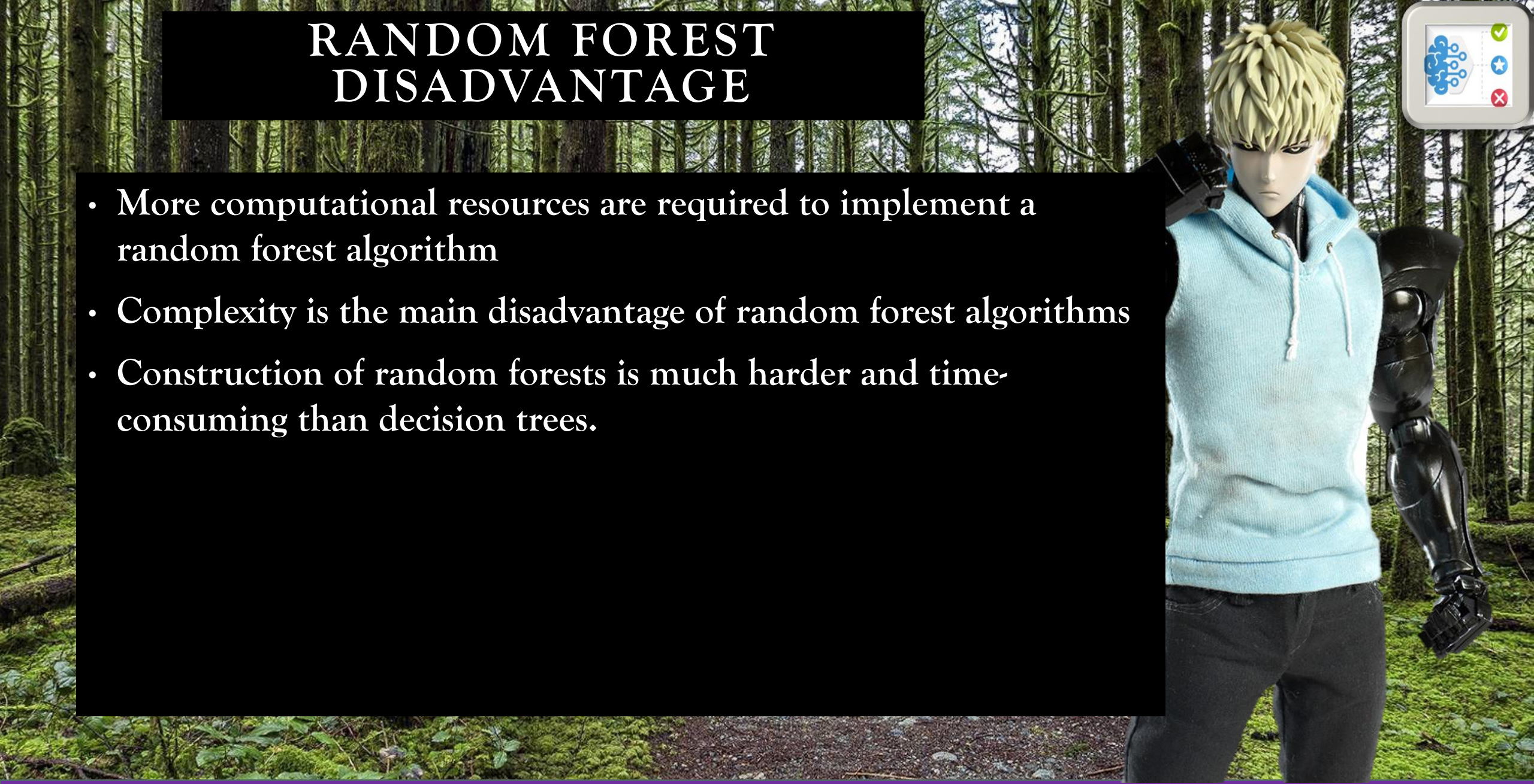


RANDOM FOREST ADVANTAGES

- Solve classification and regression
- Can handle large datasets with higher dimensionality
- It has effective method for estimating missing data and maintains accuracy when large proportion of the data are missing.
- Usually robust to outliers and can handle automatically
- No feature scaling required

RANDOM FOREST DISADVANTAGE

- More computational resources are required to implement a random forest algorithm
- Complexity is the main disadvantage of random forest algorithms
- Construction of random forests is much harder and time-consuming than decision trees.



UNSUPERVISED LEARNING

WHAT IS UNSUPERVISED LEARNING?

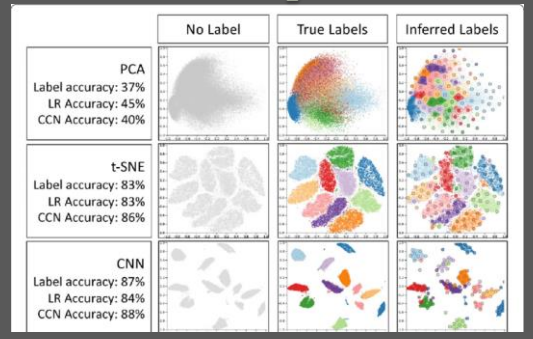
Unsupervised learning as a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.

WHEN WE USE UNSUPERVISED LEARNING?

1. When we have a datasets with no-pre existing label
2. Want to make the dimensional reduction
3. Want to make clustering automatics

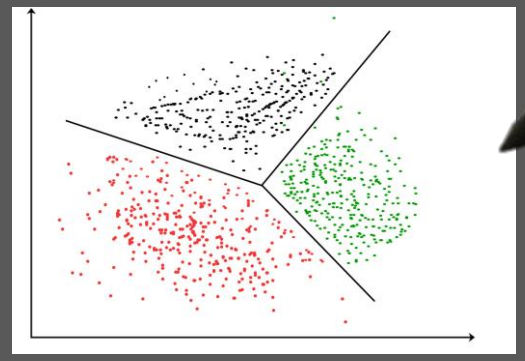
FUNCTION OF UNSUPERVISED LEARNING?

Association Analysis



- Text Mining
- Face Recognition
- Big Data Visualization
- Image Recognition

Clustering

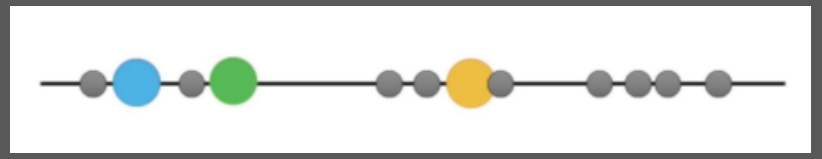


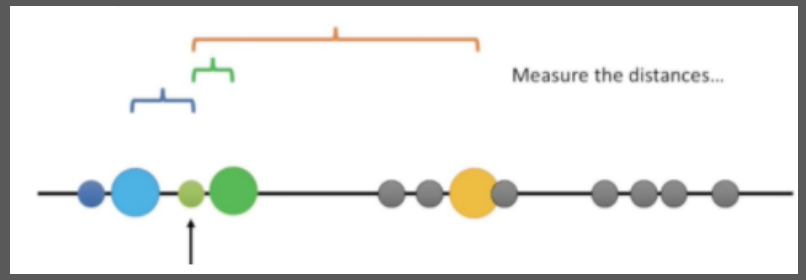
- City Planning
- Targeted Marketing

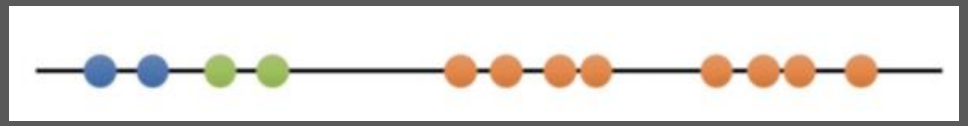
K-MEANS CLUSTERING

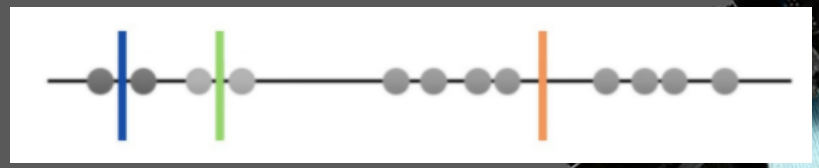
K-Means clustering is a method developed by Stuart Lloyd of Bell Labs in 1957. Lloyd uses this method to convert analog signals to digital signals. This signal conversion process is also known as Pulse Code Modulation.

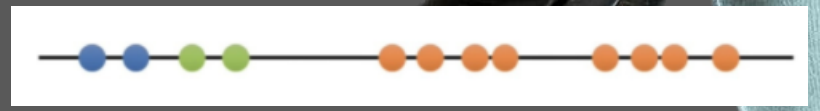
K-MEANS CLUSTERING



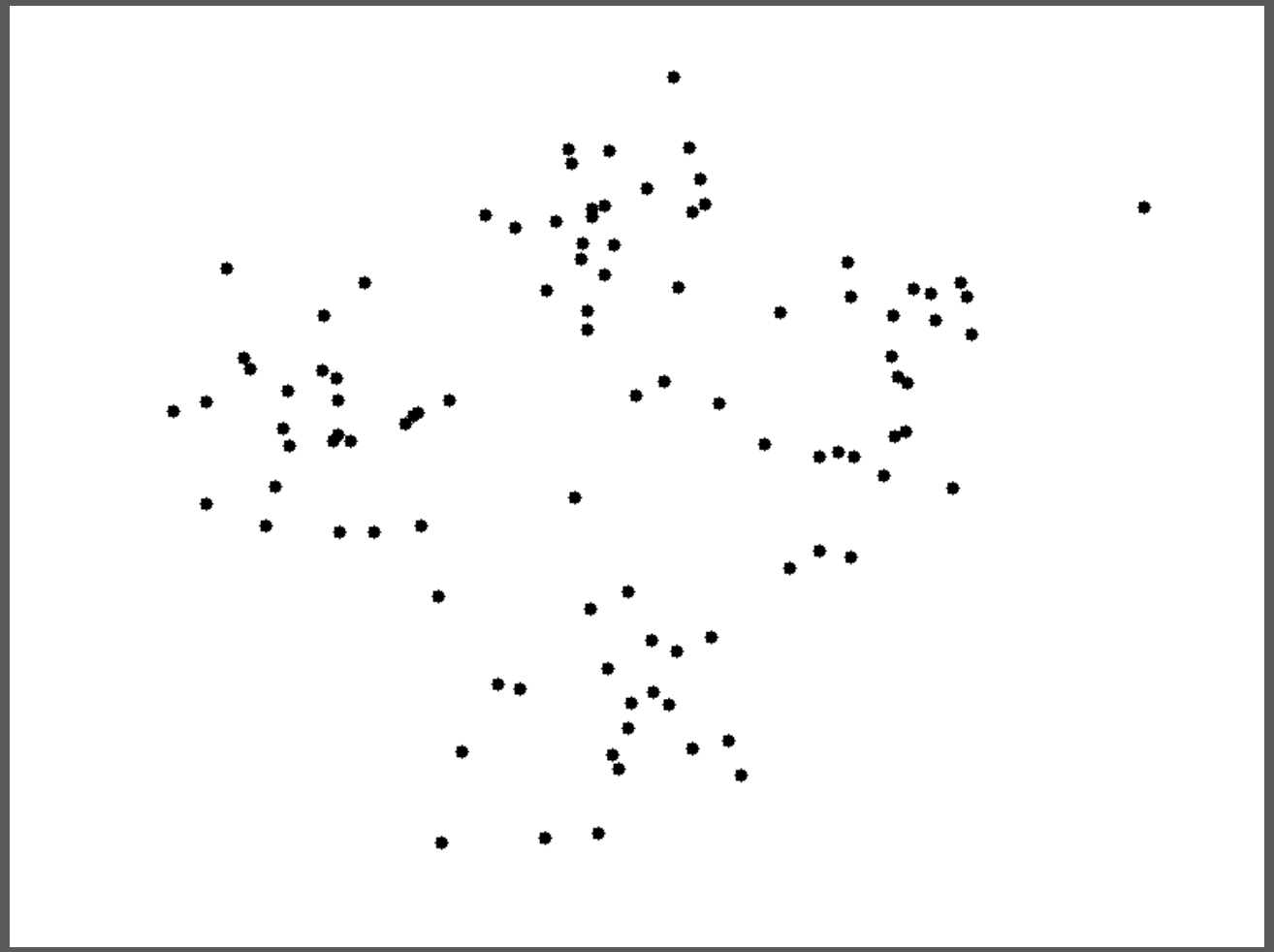




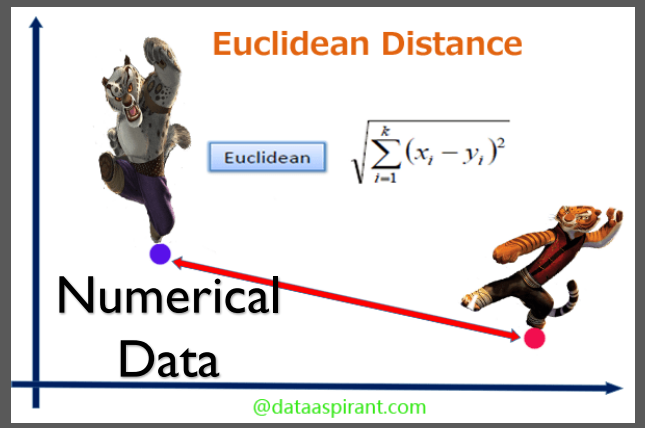


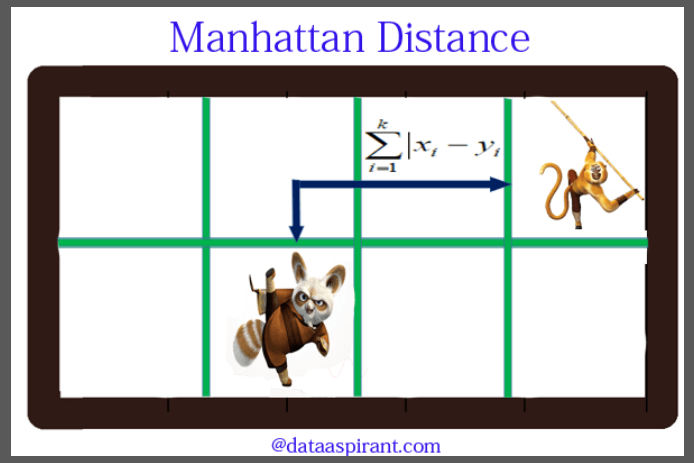


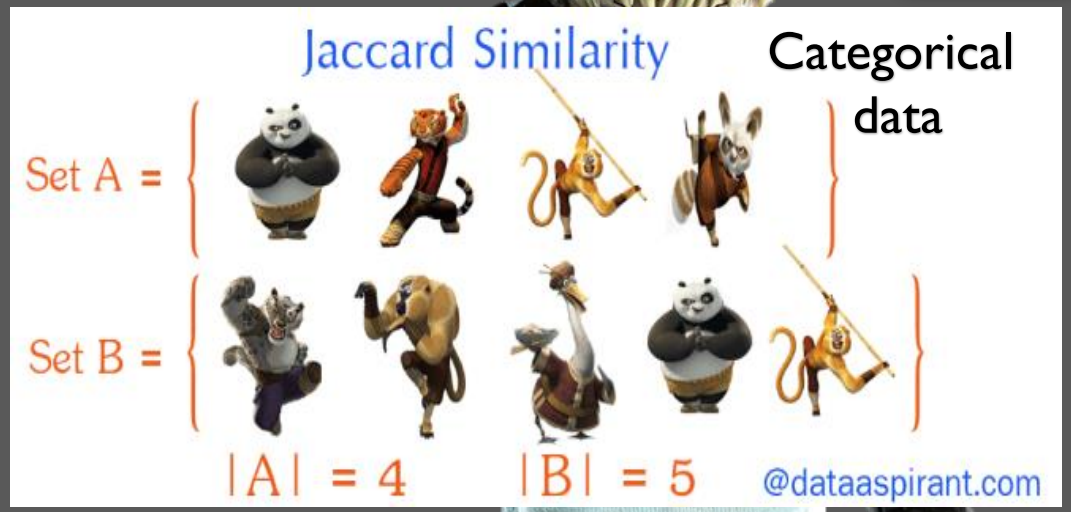
HOW'S K-MEANS WORK?

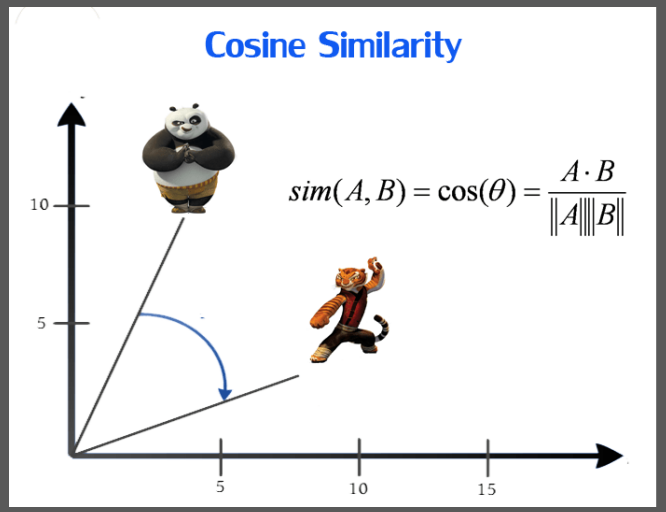


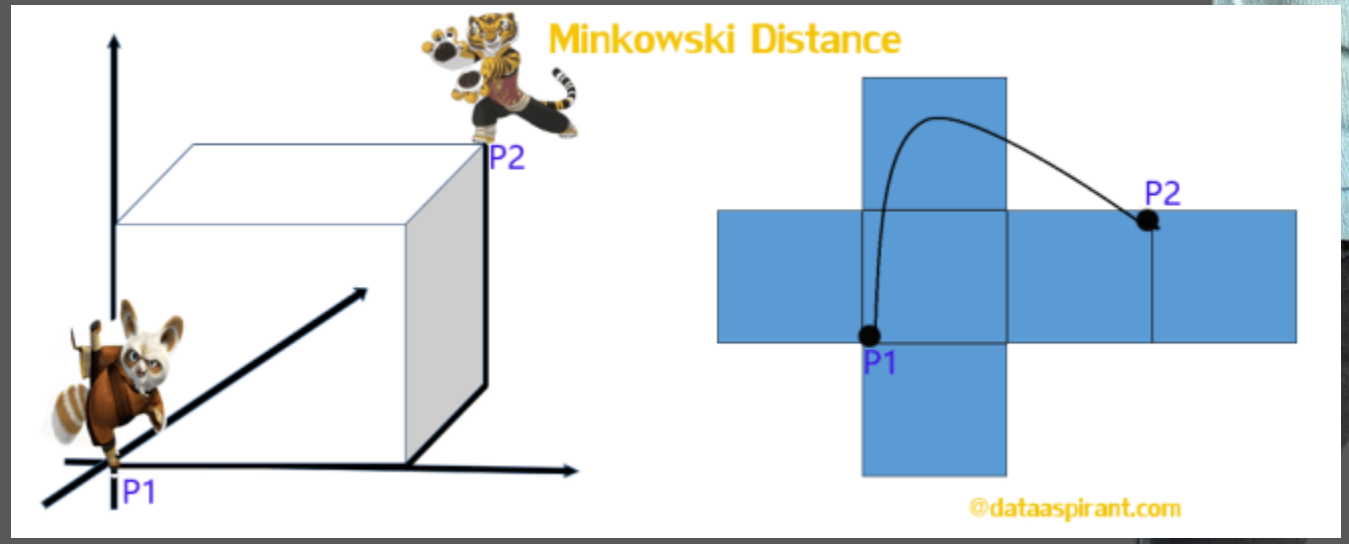
MEASUREMENT OF SIMILARITY











PREPROCESSING BEFORE CLUSTERING

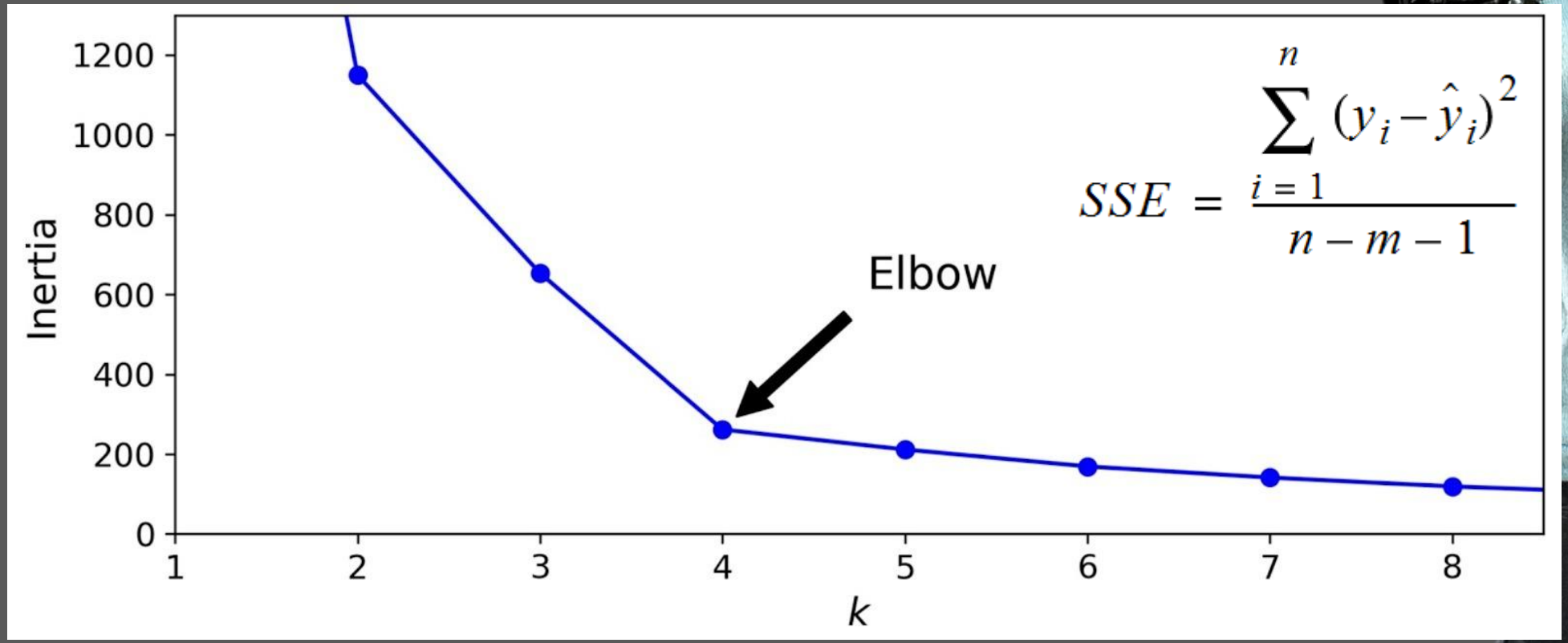
A. Scaling :
   A. Min-Max Normalization,
   B. Z-Score,
   C. Log Transform Check detail
B. Imputation: dealing missing value.

HOW TO CHOSE THE BEST K?

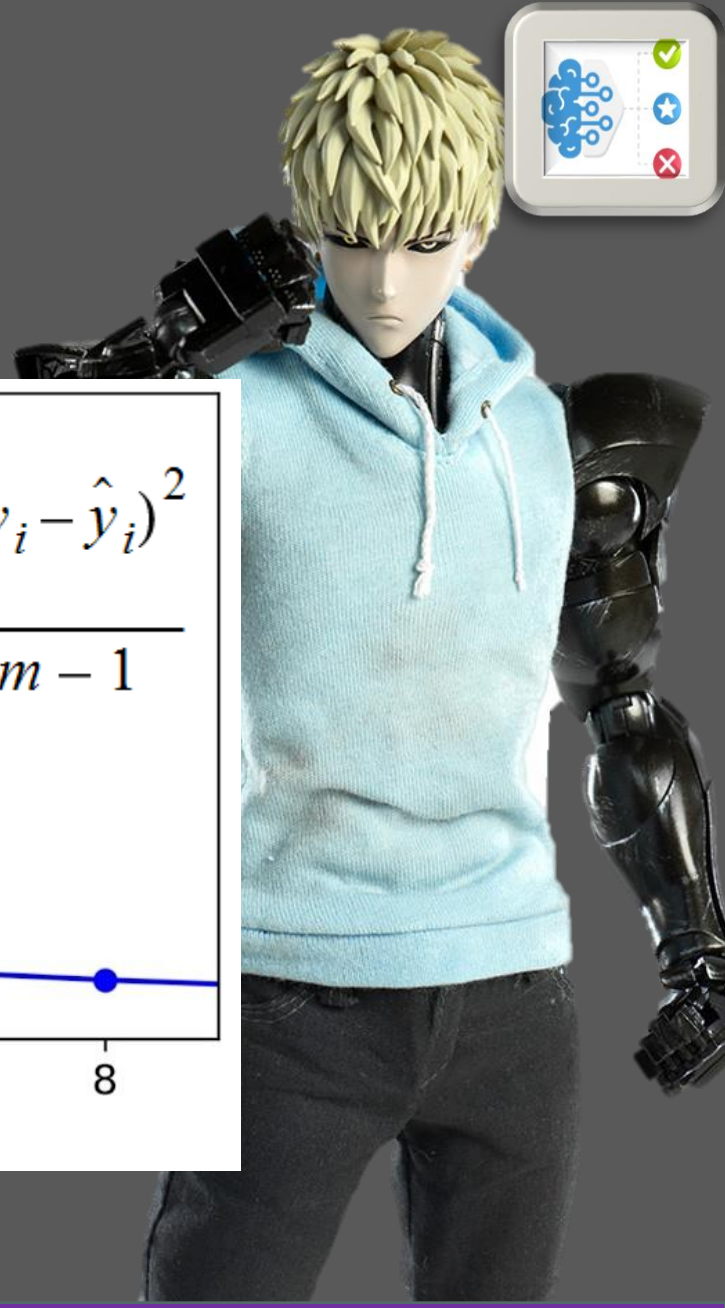
A. There is no “easy” way for choosing the best ‘K’

B. But we can use the **Elbow Method**

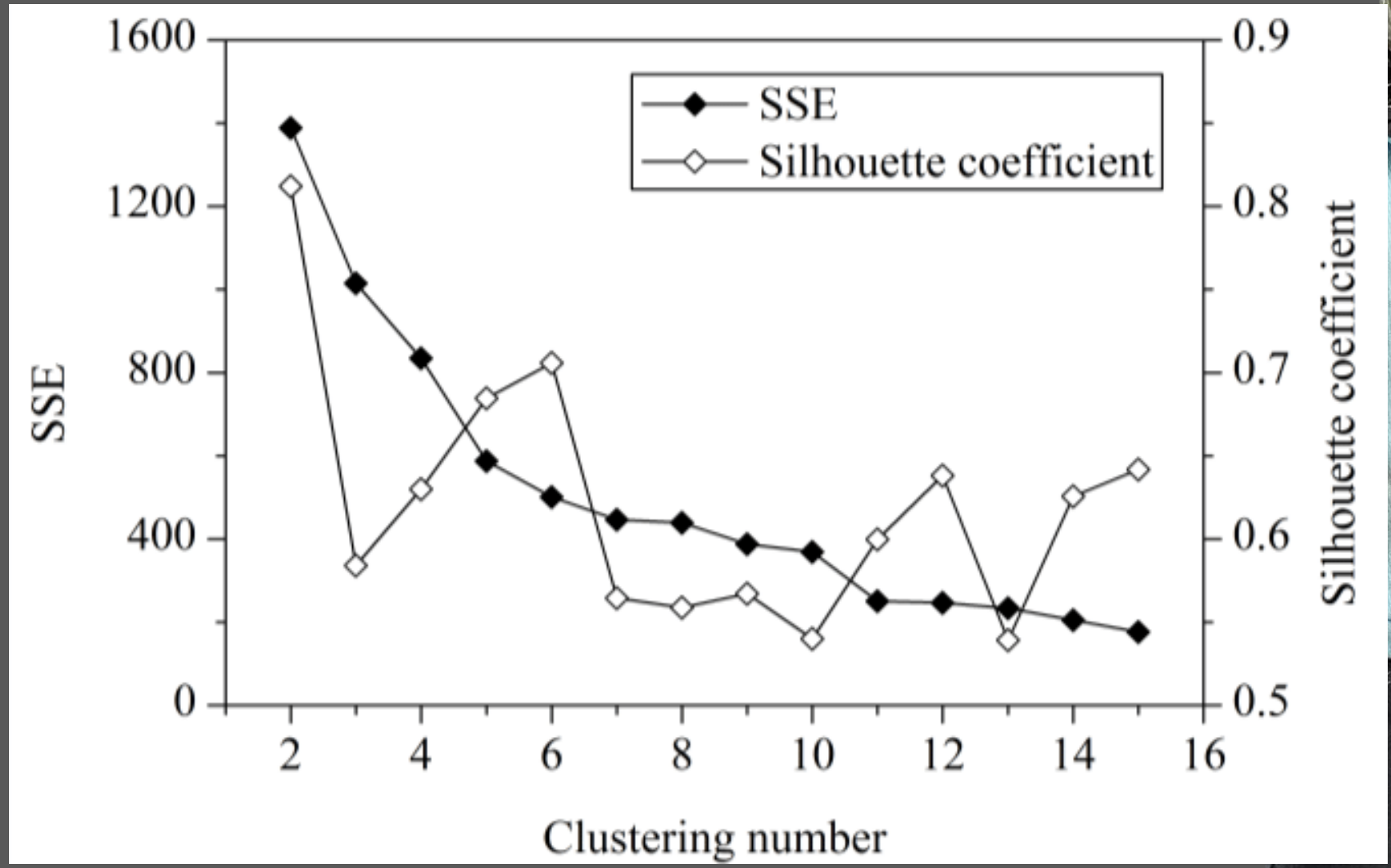
ELBOW METHOD



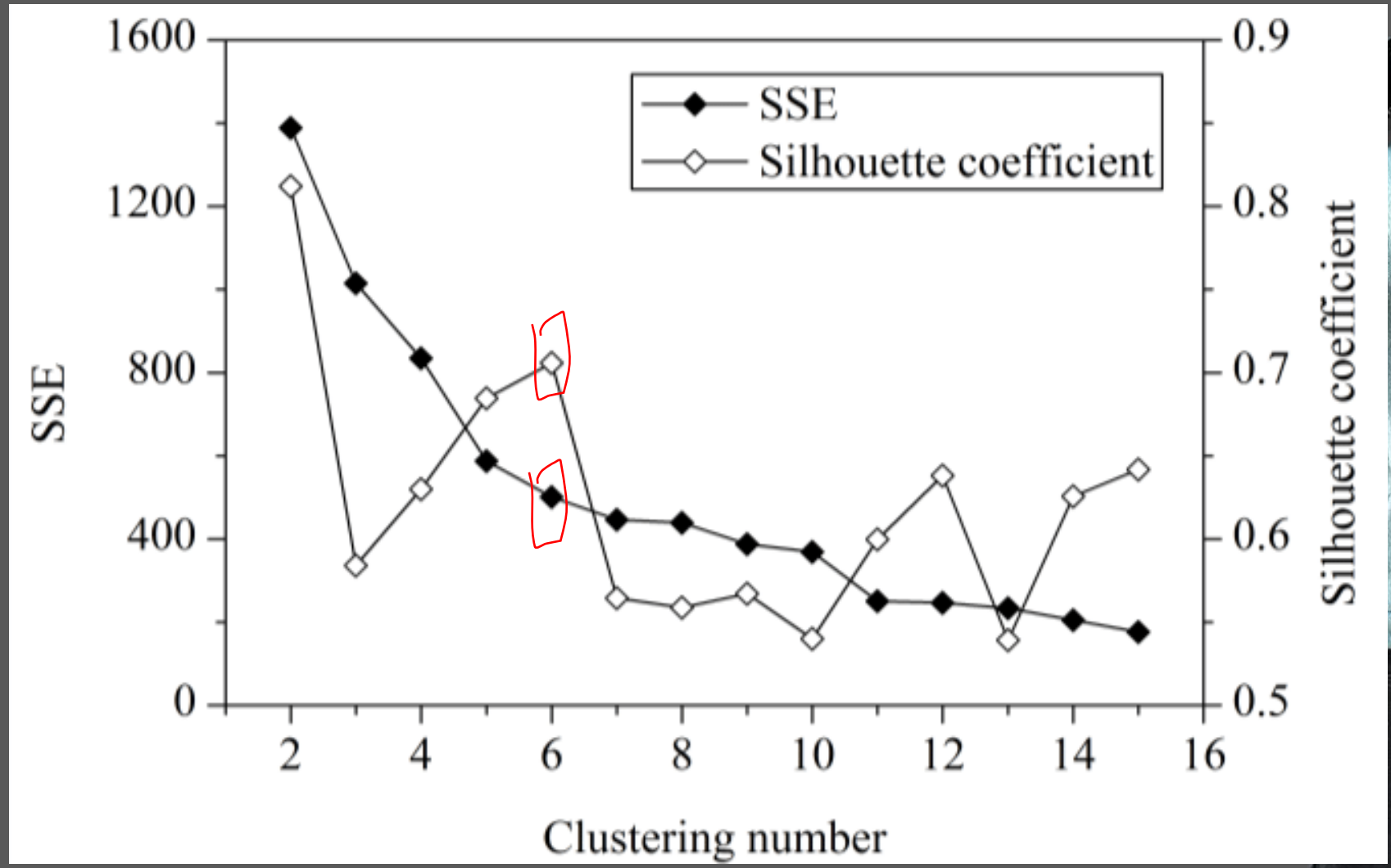
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



ELBOW METHOD



ELBOW METHOD



ADVANTAGES K-MEANS

- Relatively simple to implement
- Scales to large datasets
- Guarantees convergence
- Can warm-start the positions of centroids
- Can warm-start the positions of centroids
- Easily adapts to new examples

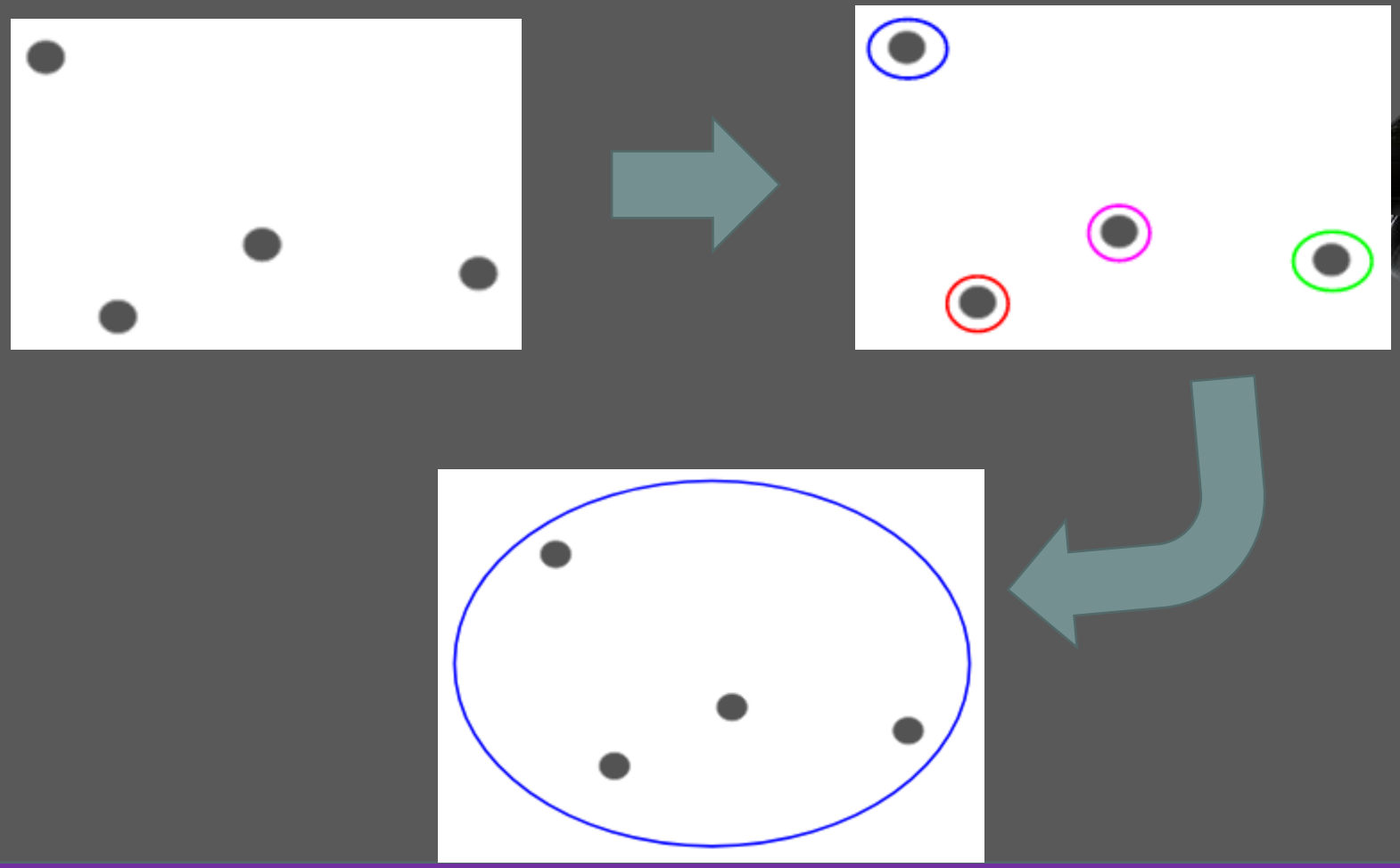
DISADVANTAGES K-MEANS

- Choose k-manually
- Being dependent on initial values
- Clustering data of varying sizes and density
- Clustering outliers
- Scaling with number of dimensions

HIERARCHICAL CLUSTERING

Hierarchical clustering is another unsupervised learning algorithm that is used to group together are unlabeled data points having similar characteristics.

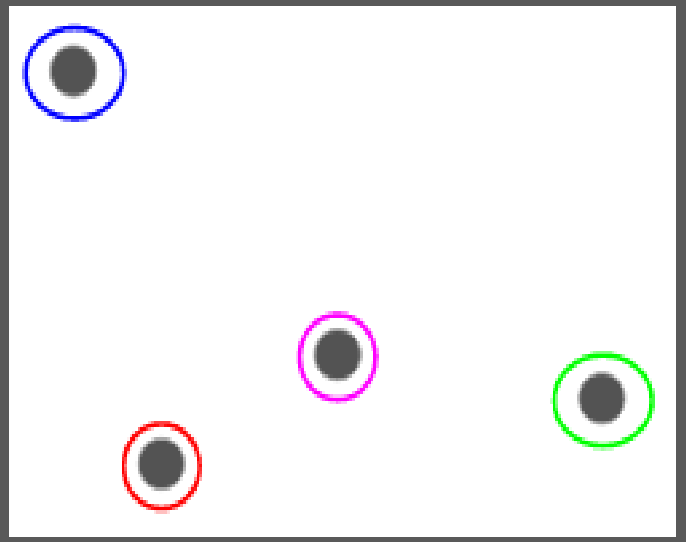
HIERARCHICAL CLUSTERING

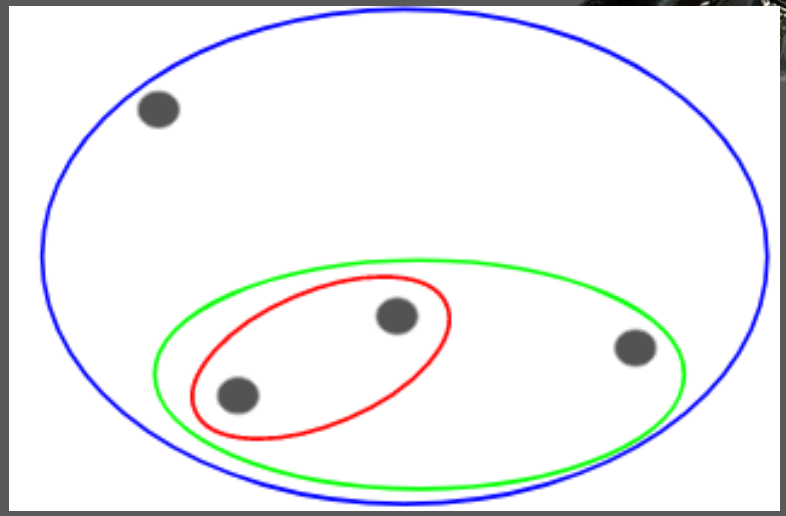


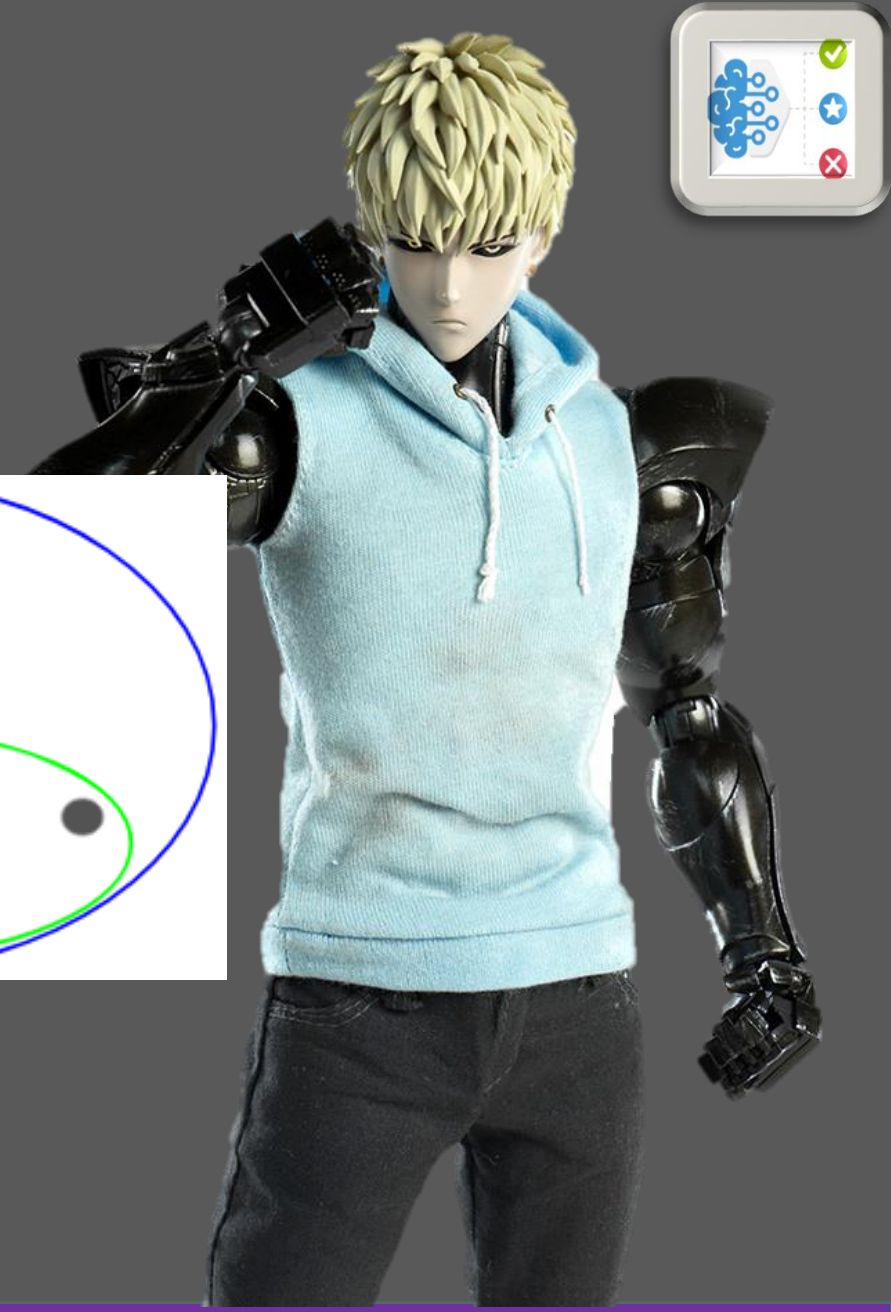
TYPE OF HIERARCHICAL CLUSTERING

1. Agglomerative hierarchical clustering
2. Divisive Hierarchical clustering

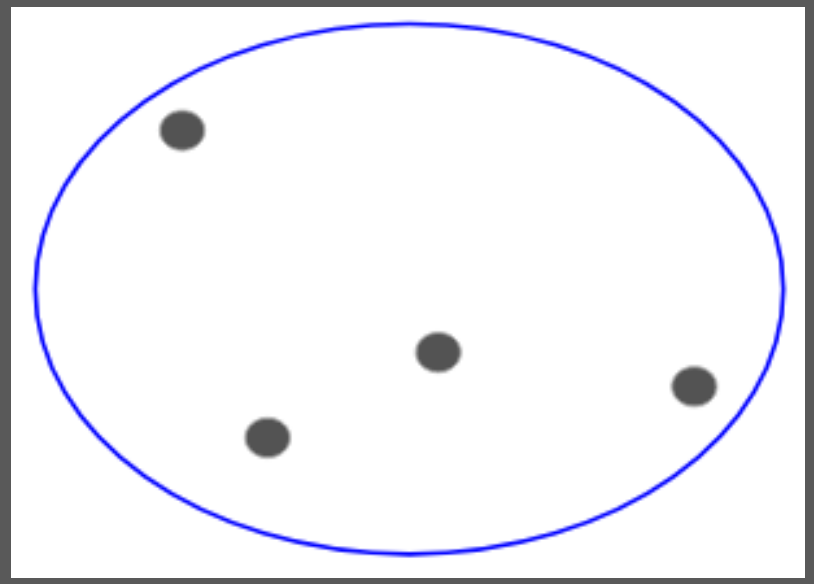
1. Agglomerative hierarchical clustering

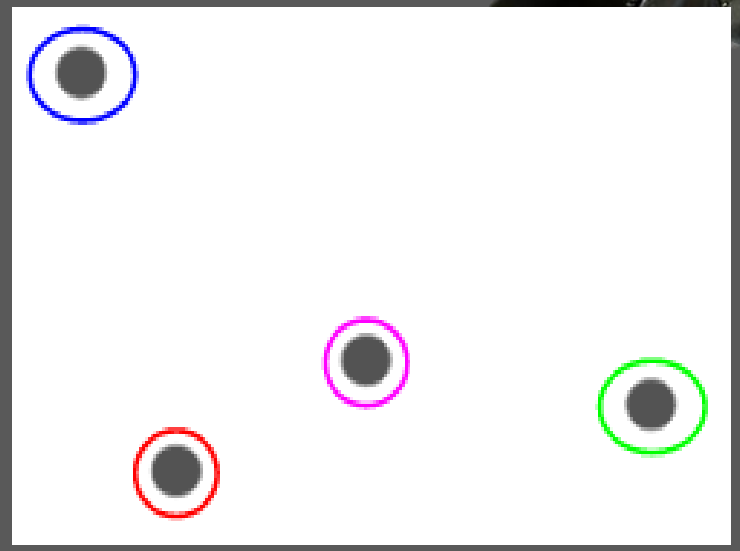






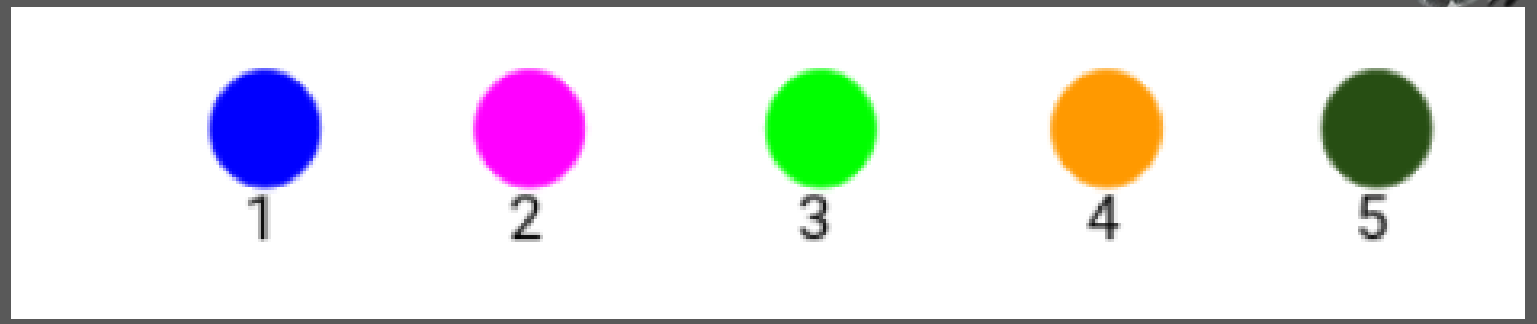
2. Divisive Hierarchical clustering





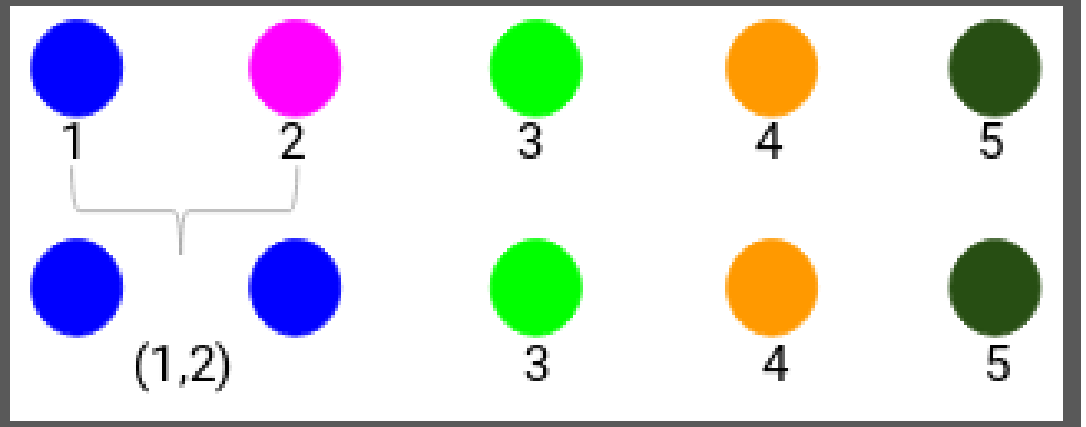
Steps to Perform Hierarchical Clustering

1. Assign all the points to an individual cluster



Steps to Perform Hierarchical Clustering

2. Looks the smallest distance in the proximity matrix and merge the points with smallest distance.

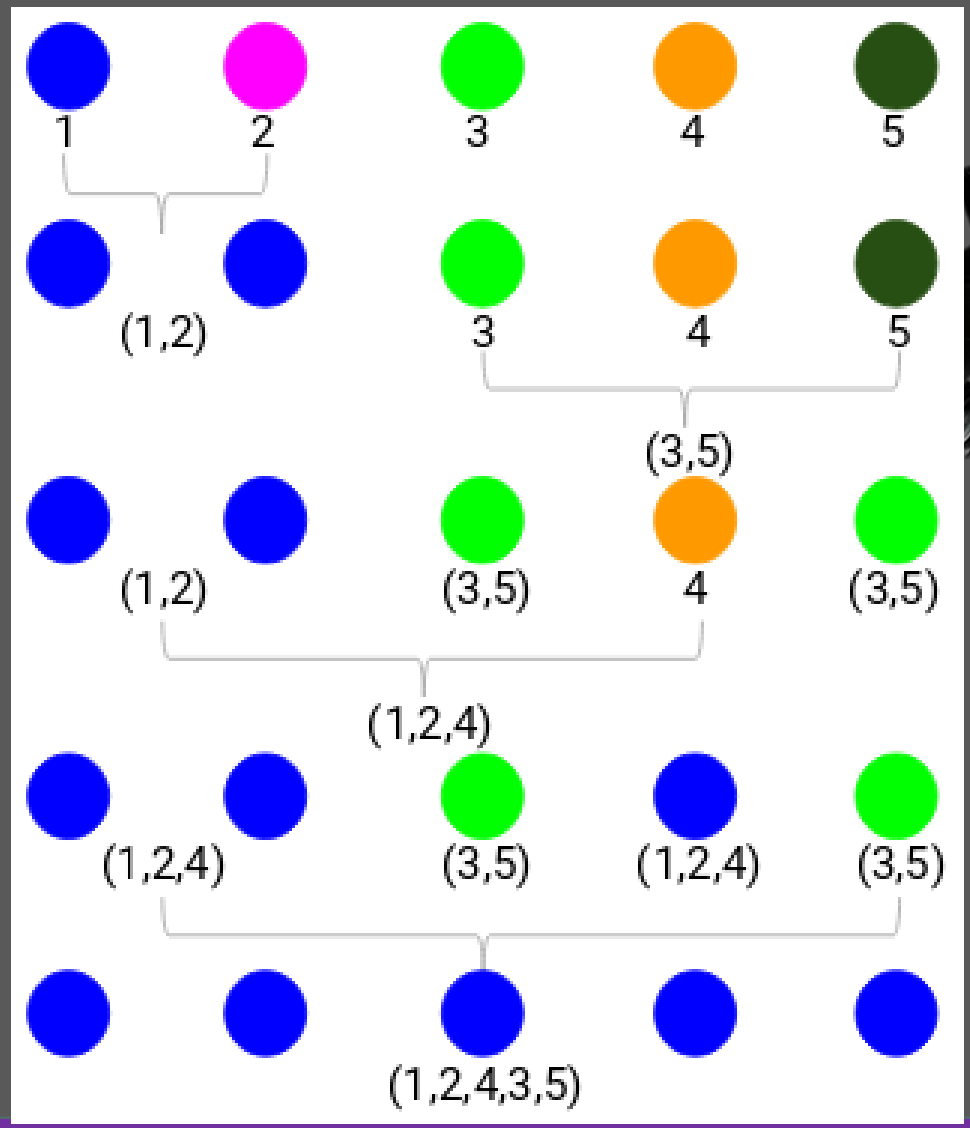


Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0

Steps to Perform Hierarchical Clustering

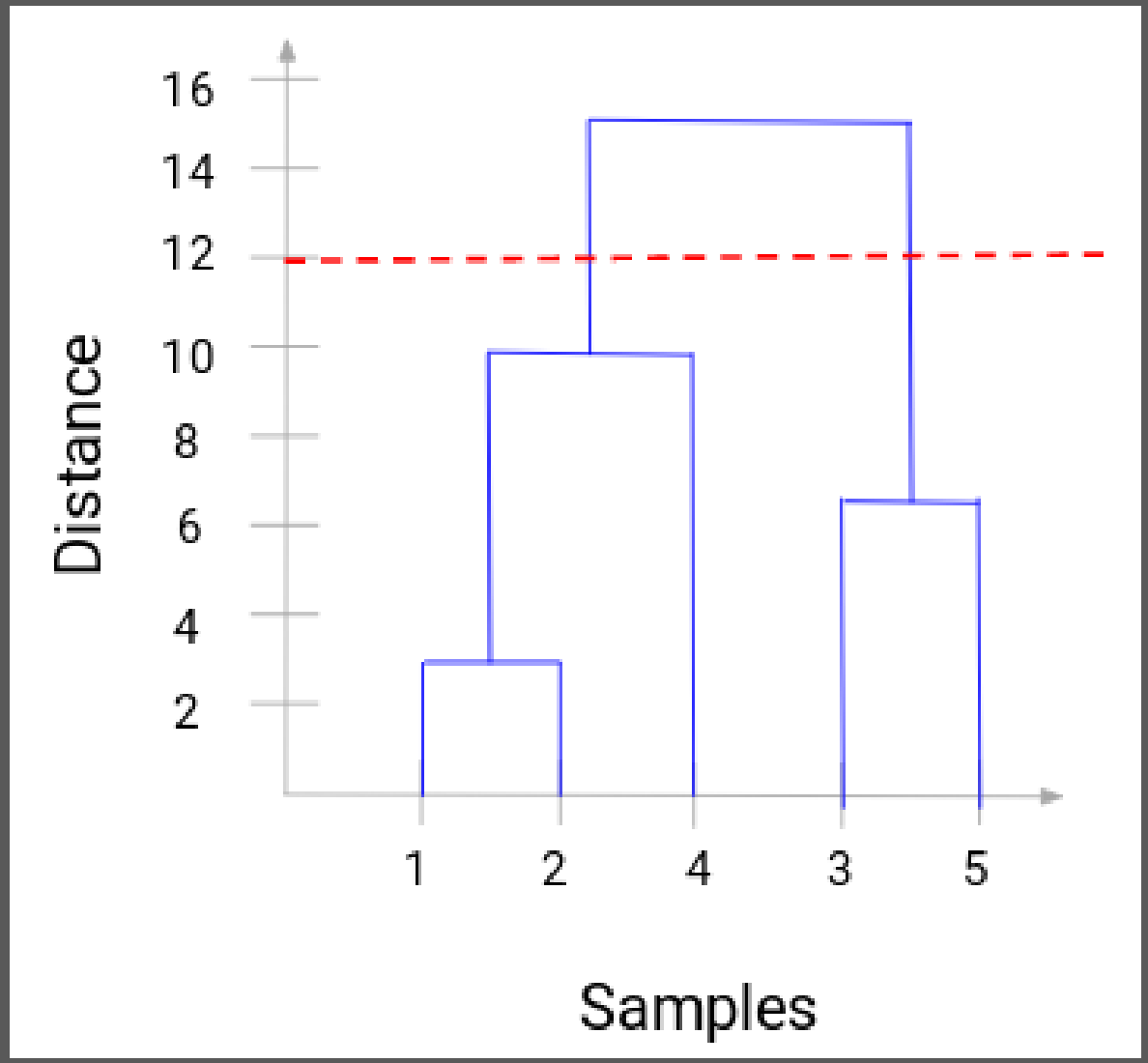
3. Repeat step 2 until only a single cluster is left.



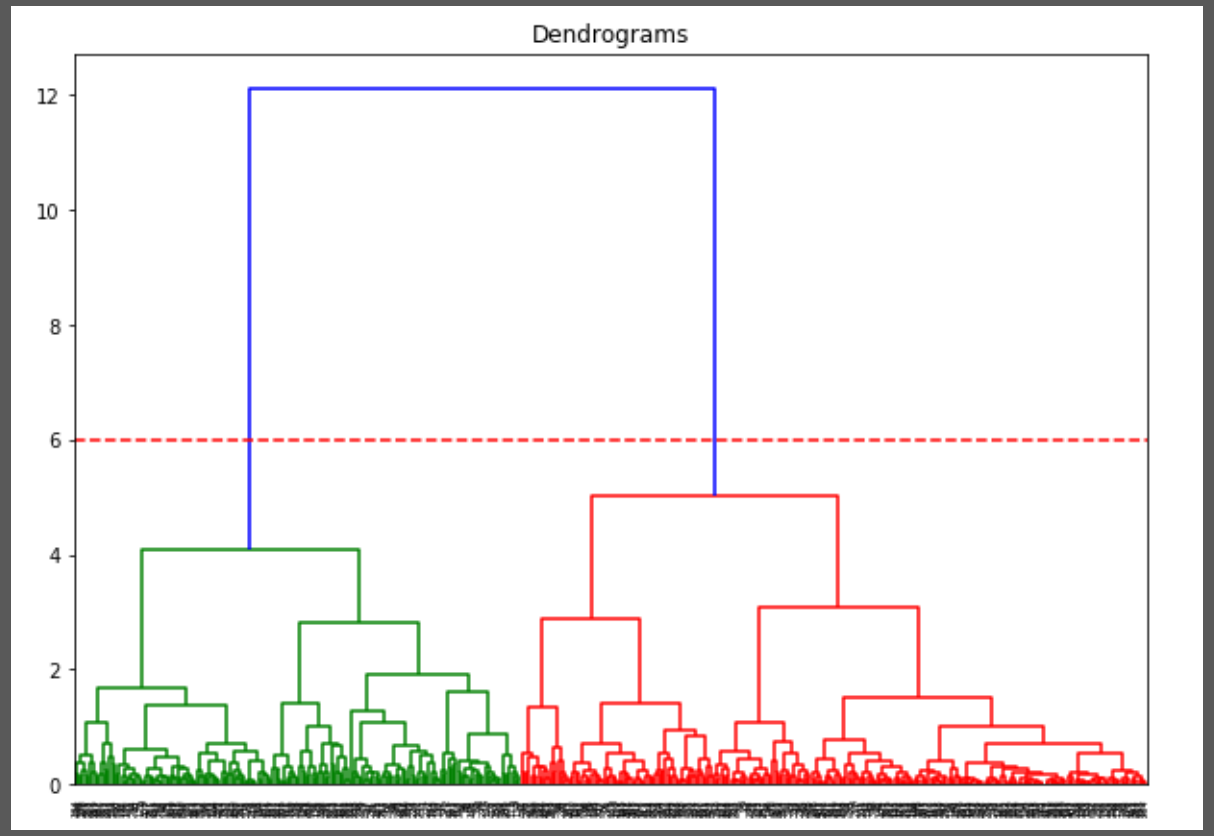
How we choose the number of cluster in hierarchical clustering?

To get the number of clusters for hierarchical clustering, we make use of an awesome concept called a **Dendrogram**.

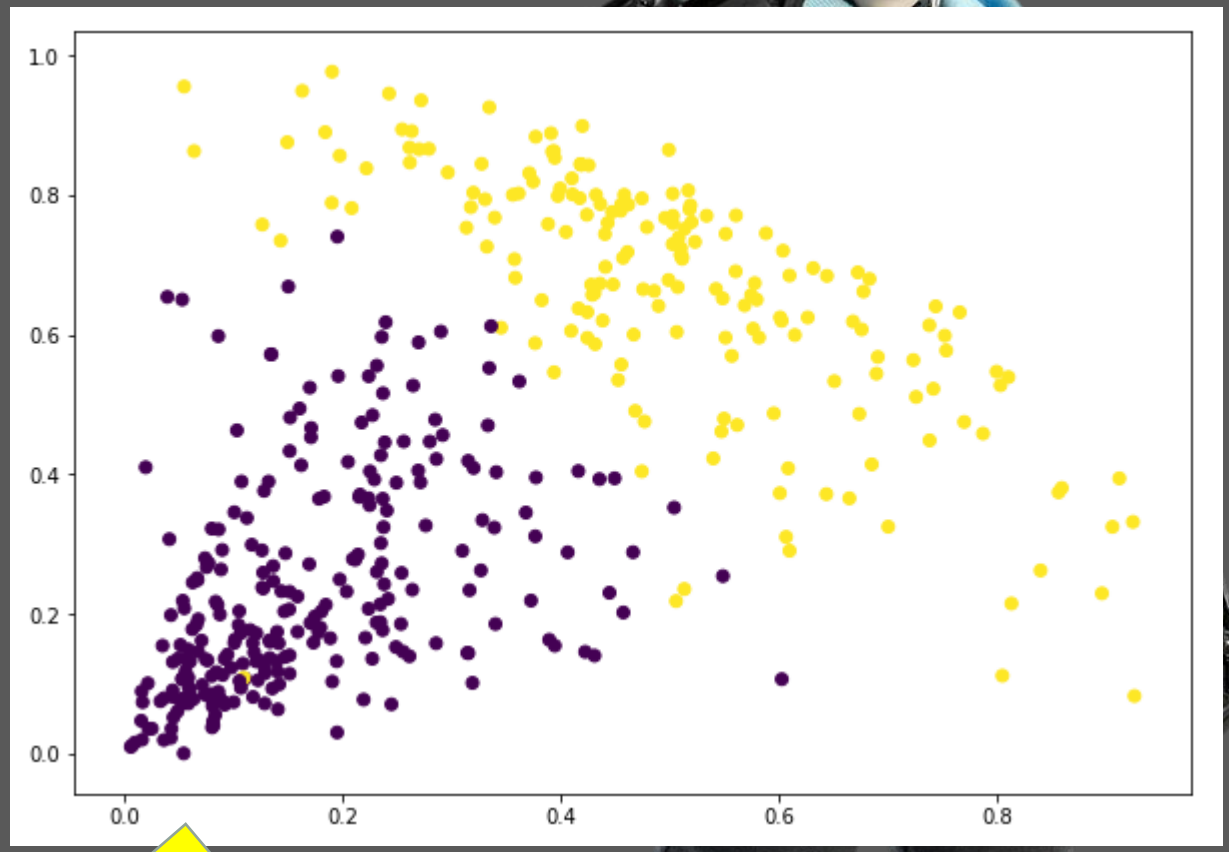
Dendrogram



Dendrogram



Scatterplot clustering



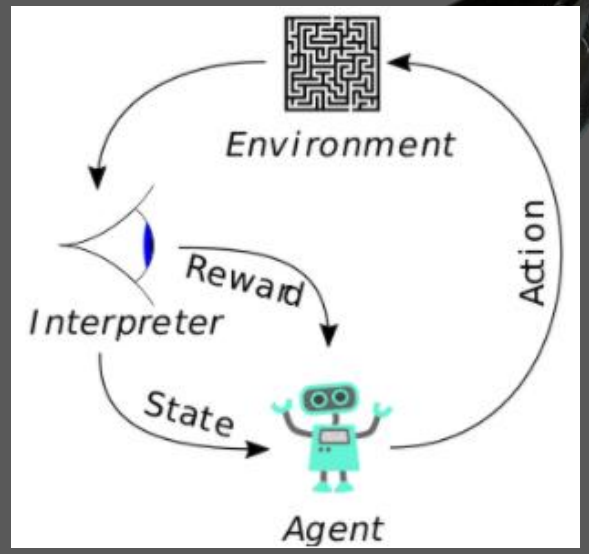
HOW TO CHOOSE ALGORITHM PROPERLY?

- See your data it has label or not?
- What's your goal?
  - Regression?
  - Classification?
  - Clustering?
  - Dimensional Regression ?
- Compare with other algorithms?

REINFORCEMENT LEARNING

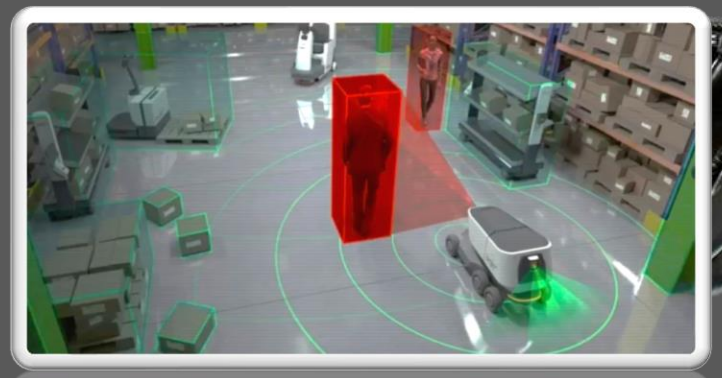
WHAT IS REINFORCEMENT LEARNING?

Reinforcement learning (RL) is an area of **machine learning** concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward.



FUNCTION OF REINFORCEMENT LEARNING?

Inventory Management



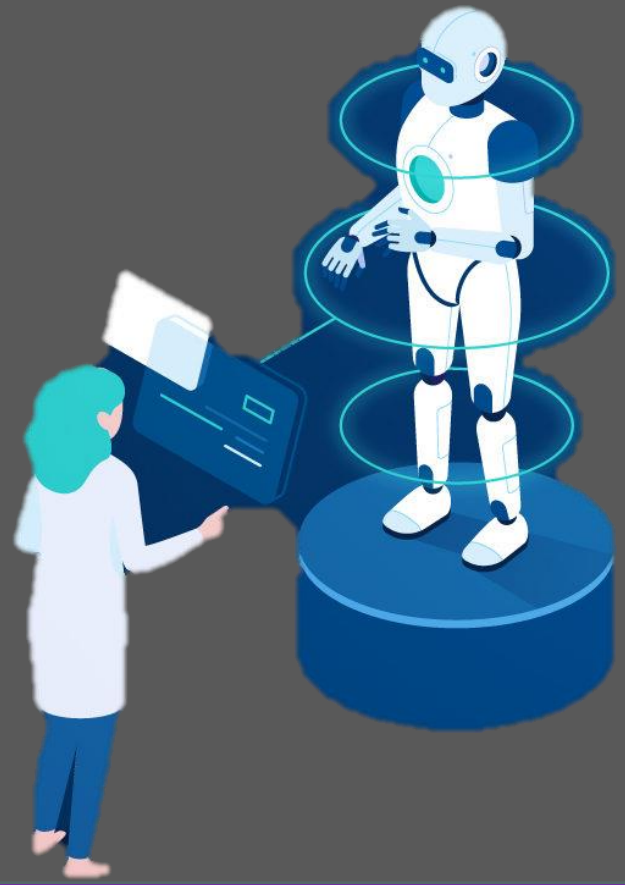
Robot Navigation

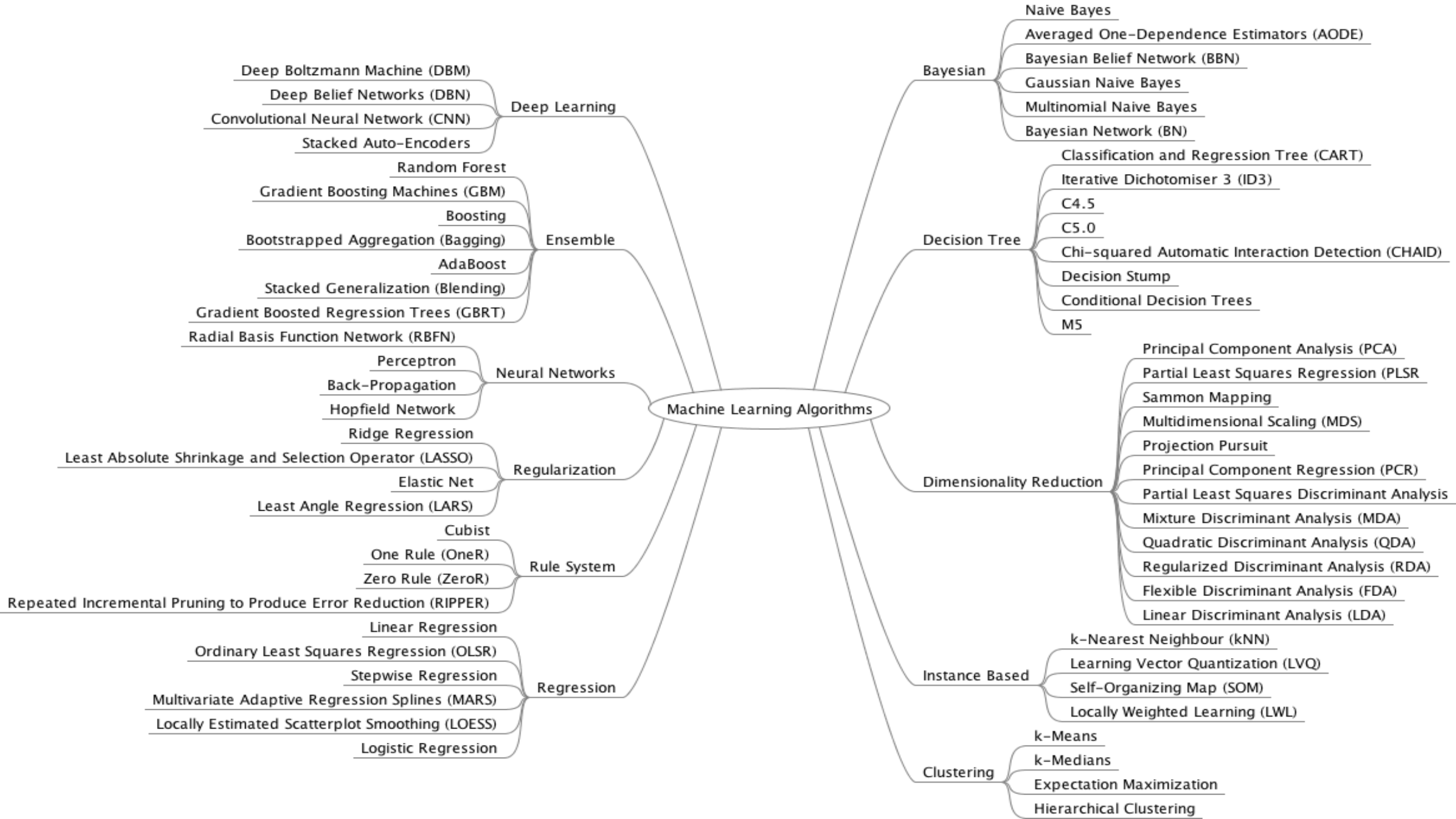
Gaming

Finance

Manufacturing

MIND MAP MACHINE LEARNING




Machine Learning Algorithms
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

SOURCES:

- https://data-flair.training/blogs/types-of-machine-learning-algorithms/
- https://www.lpsm.paris/pageperso/has/source/Hand-on-ML.pdf
- https://searchenterpriseai.techtarget.com/definition/unsupervised-learning
- https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
- https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/#:~:text=Polynomial%20Regression%20is%20a%20form,denoted%20E(y%20%7Cx)

SOURCES:

- https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/
- **Iykra course**
- **Dicoding course machine learning basic**
- https://www.mindmeister.com/690844333/machine-learning?fullscreen=1#
- https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/
- https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_clustering_algorithms_overview.htm

Thank You

Ronny Fahrudin | Data Science Fellowship at IYKRA