

# Machine Learning (ML) Workshop



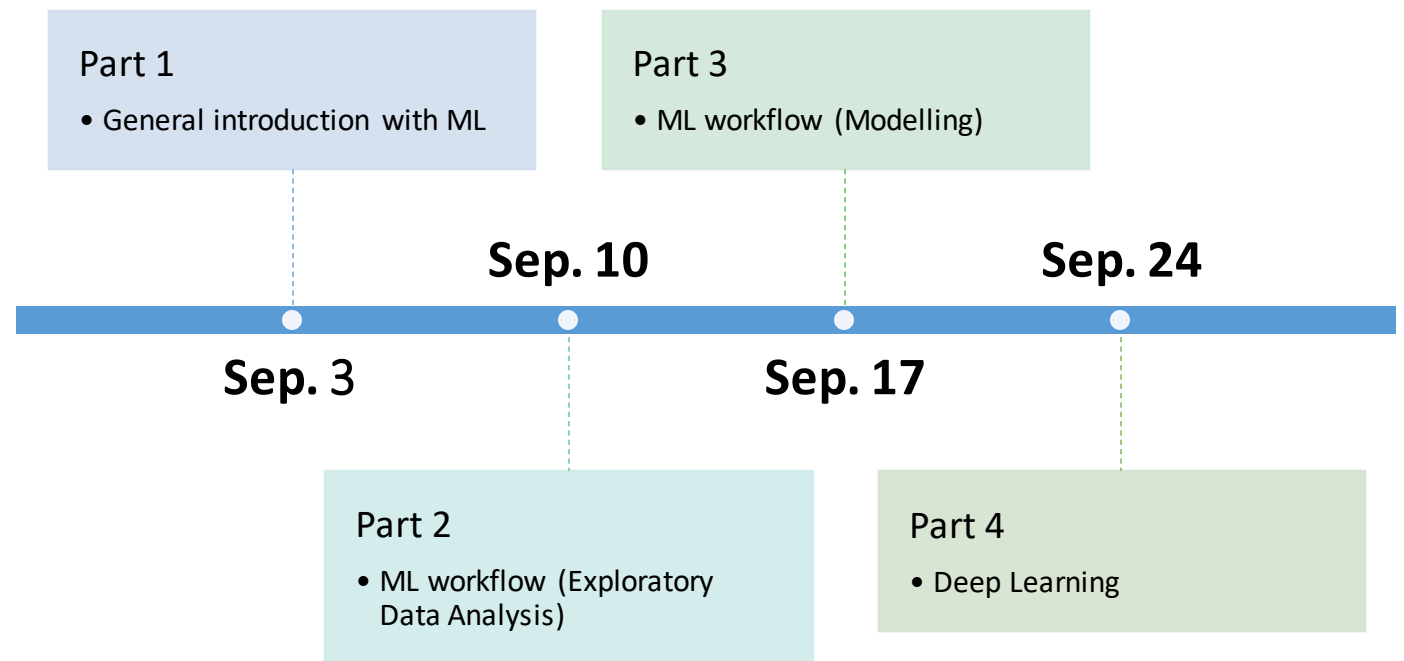
Dr Sara Soltaninejad  
Fall 2021

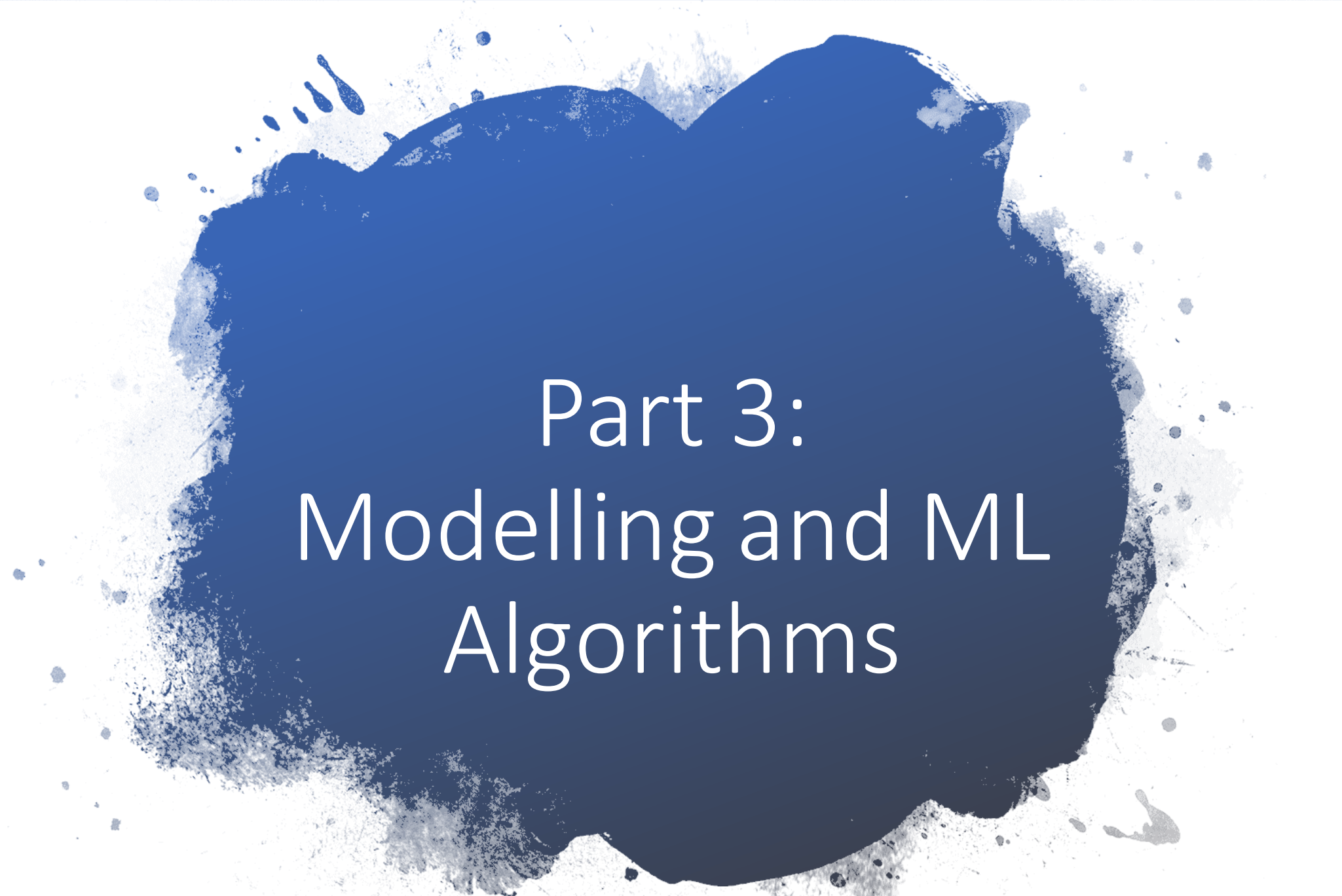
# Who am I?

- Sara Soltaninejad
- PhD in MRC, Computing Science Department, UofA, 2016-2020
  - PhD Thesis: Intelligent Parkinson's Disease Classification and Progress Monitoring Using Non-Invasive Techniques
  - Supervisors:
    - Dr Anup Basu
    - Dr Irene Cheng
- ML Developer AltaML, 2020-Now

# ML Workshop Outline

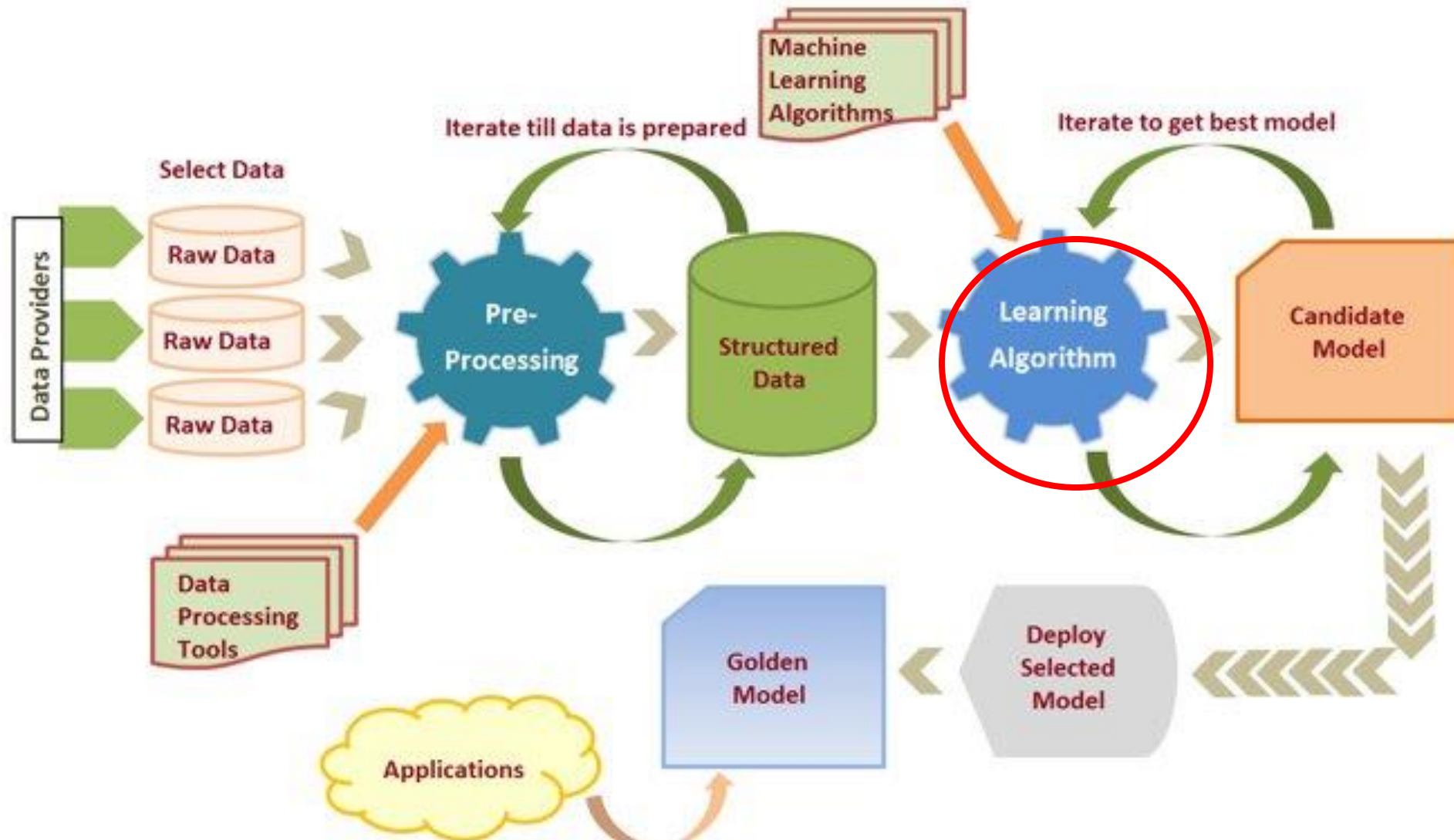
---



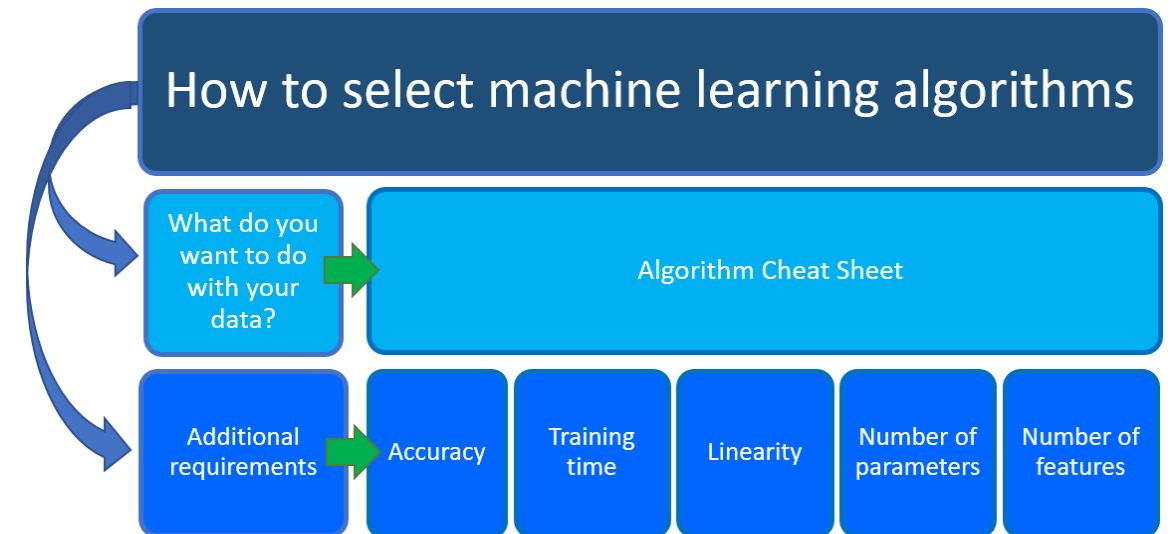
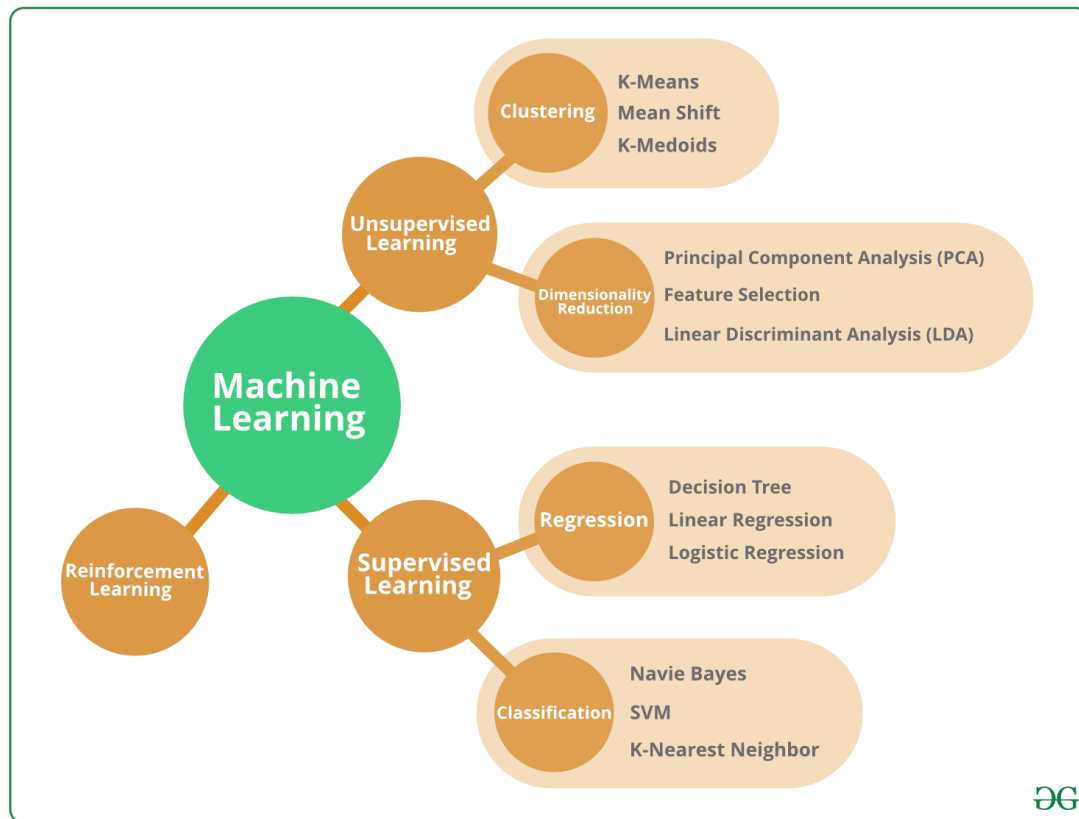


# Part 3: Modelling and ML Algorithms

# Machine Learning (ML) WorkFlow

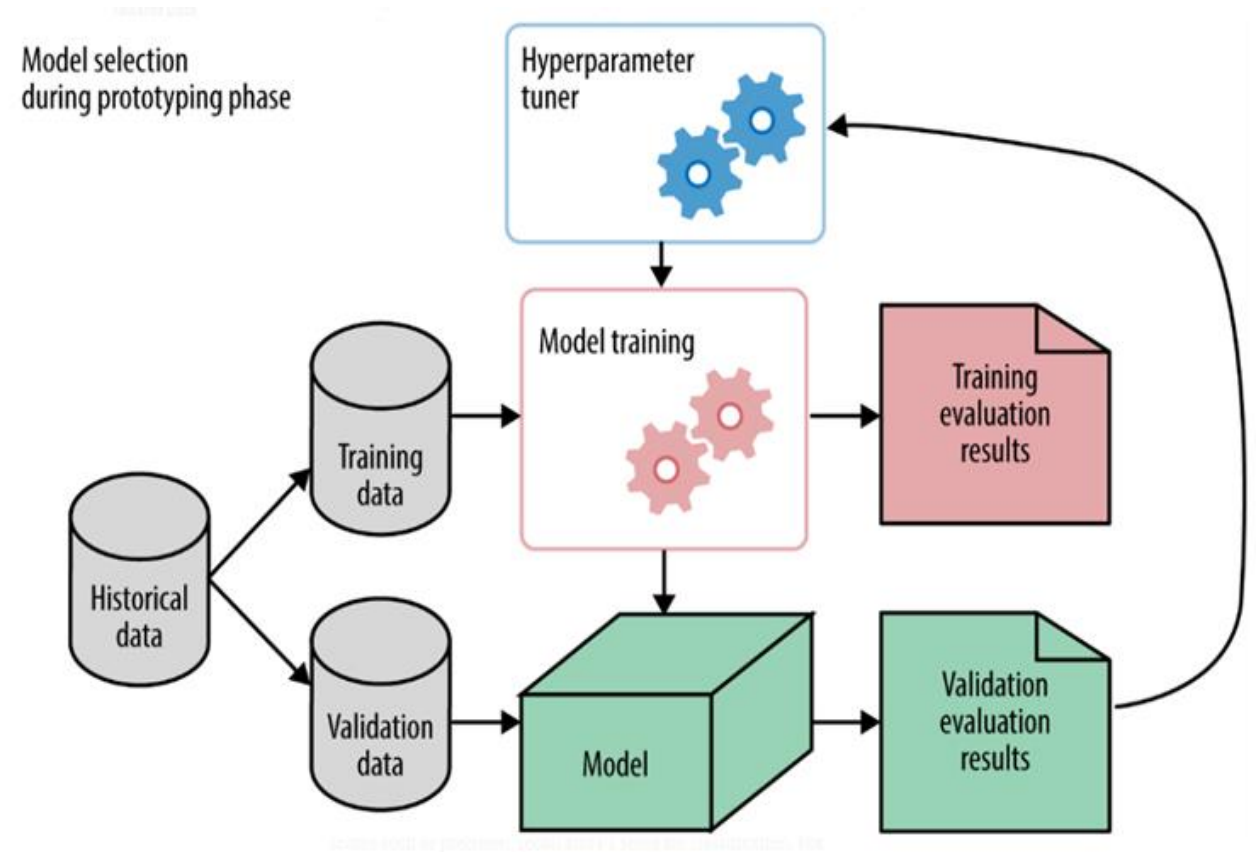


# ML Algorithms



# ML Algorithm hyperparameter Tuning

- Most algorithms have parameters that guide how the algorithms work which have to be tuned to specific problems/data
- Typically a grid search is performed where different sets of parameters are tried
- Model with a given set of parameters is trained and evaluated, using training set
- The set with highest score is selected
- There are also more complex search approaches used. But you have to be careful when you select hyperparameters to make sure the model still works well on unseen data





# ML Algorithm Performance Analysis

## Underfitting



## Overfitting

Your model is underfitting the training data when the model performs poorly on the training data. model unable to capture the underlying pattern of the data.

Reasons:

- Not enough data
- Simple model (Ex: data is nonlinear but our model is linear)

Your model is overfitting your training data when you see that the model performs well on the training data but does not perform well on the evaluation data. The model captures the noise along with the underlying pattern in data.

Reasons:

- Too Complex model
- Noisy data
- Not enough data

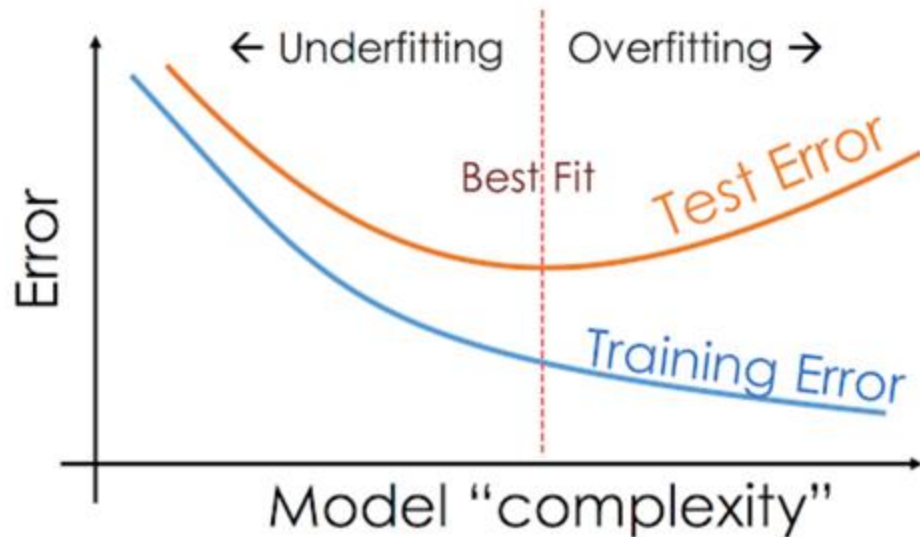


# ML Algorithm Performance Analysis

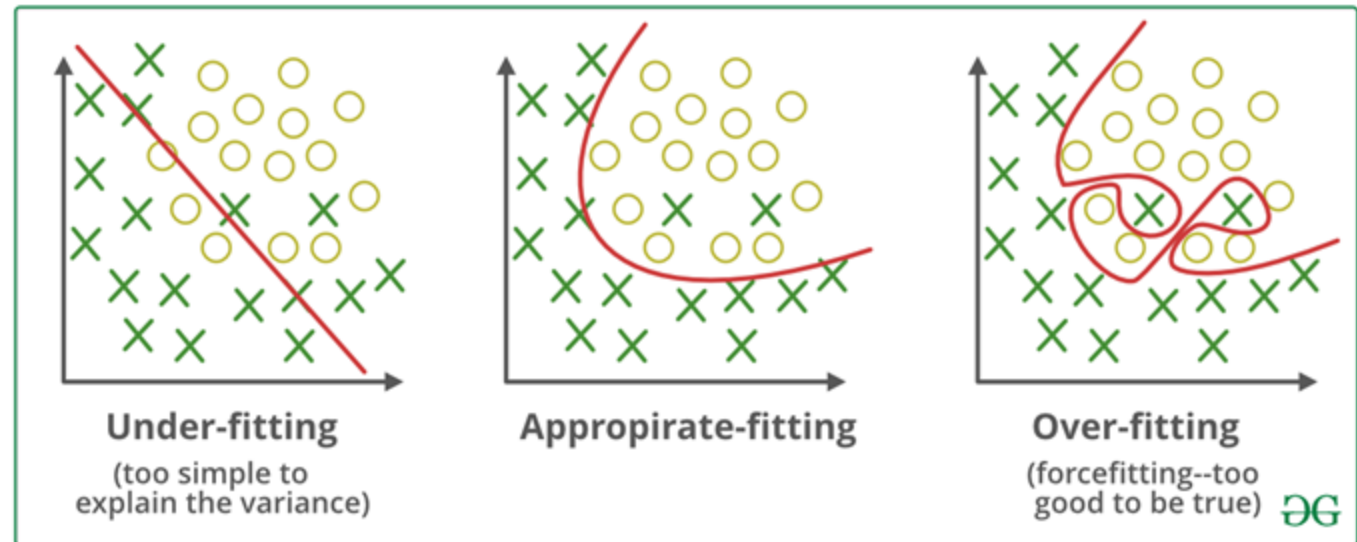
Underfitting



Overfitting

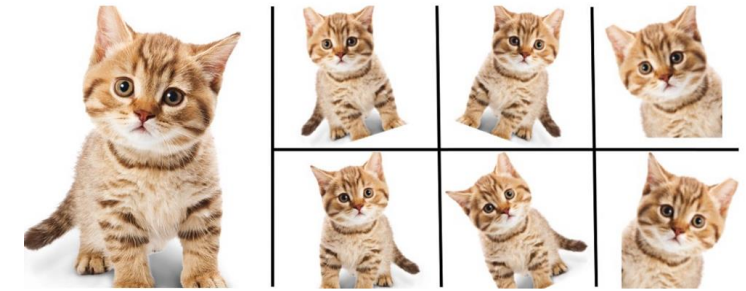


**Bias-Variance  
Trade off**



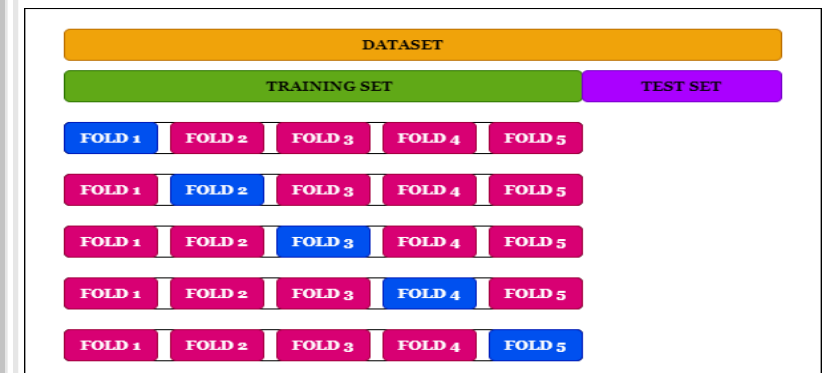
# ML Algorithm Performance Analysis

- Techniques aim at building generalized models
  - Sampling techniques - increase the number of training samples
  - Cross Validation - Split training dataset and use splits to choose best parameters for a model
  - Model Complexity (model type, model hyperparameters numbers,..)
  - Regularization - Allow for error on training data



Enlarge your Dataset

$$\text{Cost} = \underbrace{\sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^M W_j^2}_{\text{Regularization Term}}$$

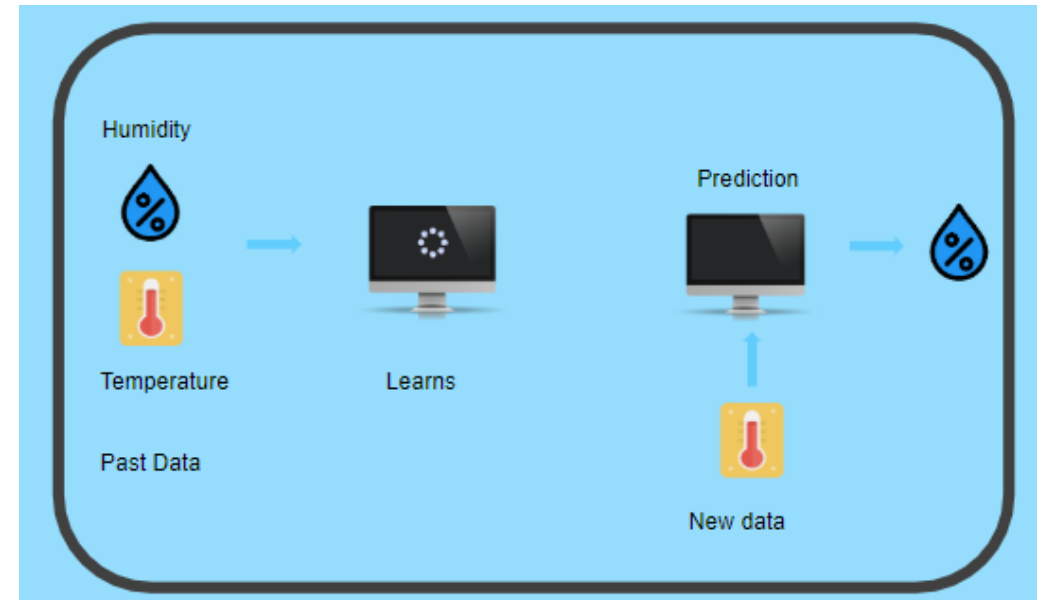
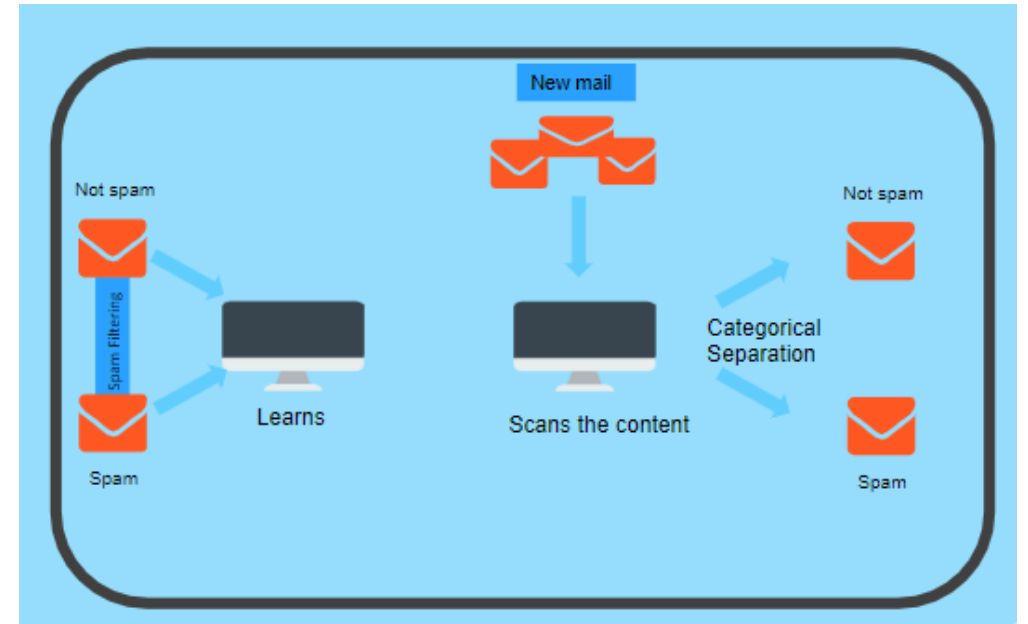


# Supervised ML Algorithms

# Classification vs Regression

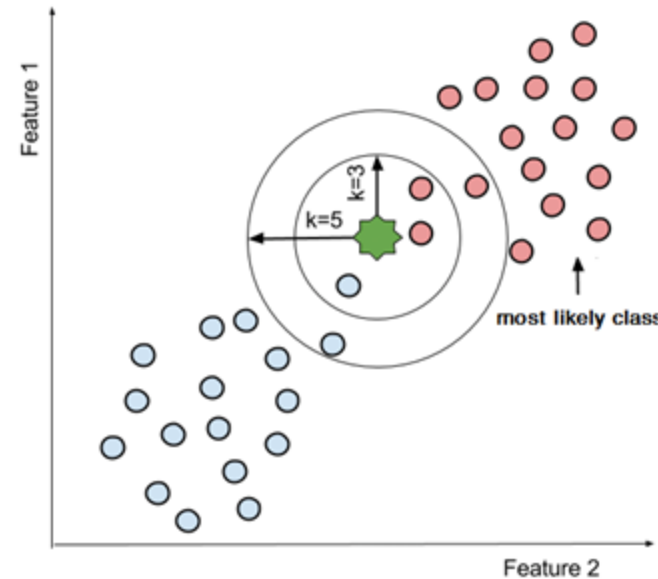
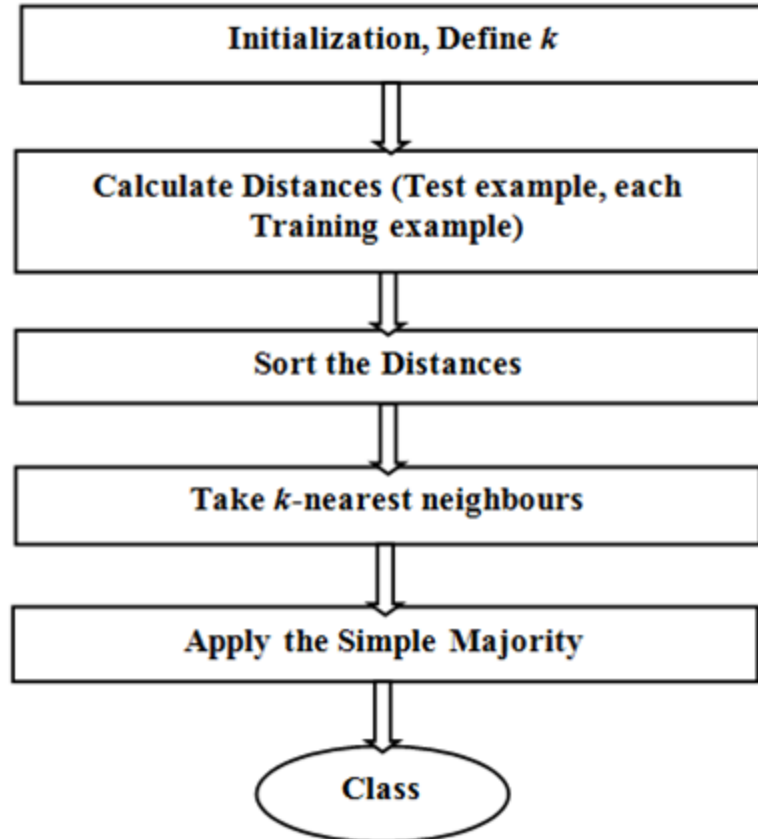
- Classification and Regression come under the same umbrella of Supervised Machine Learning and share the common concept of using past data to make predictions, or take decisions, that's where their similarity ends.

12



# K Nearest Neighbors

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. It's a Lazy and non parametric learning algorithms.



## Distance functions

Euclidean	$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^k  x_i - y_i $
Minkowski	$\left( \sum_{i=1}^k ( x_i - y_i )^q \right)^{1/q}$

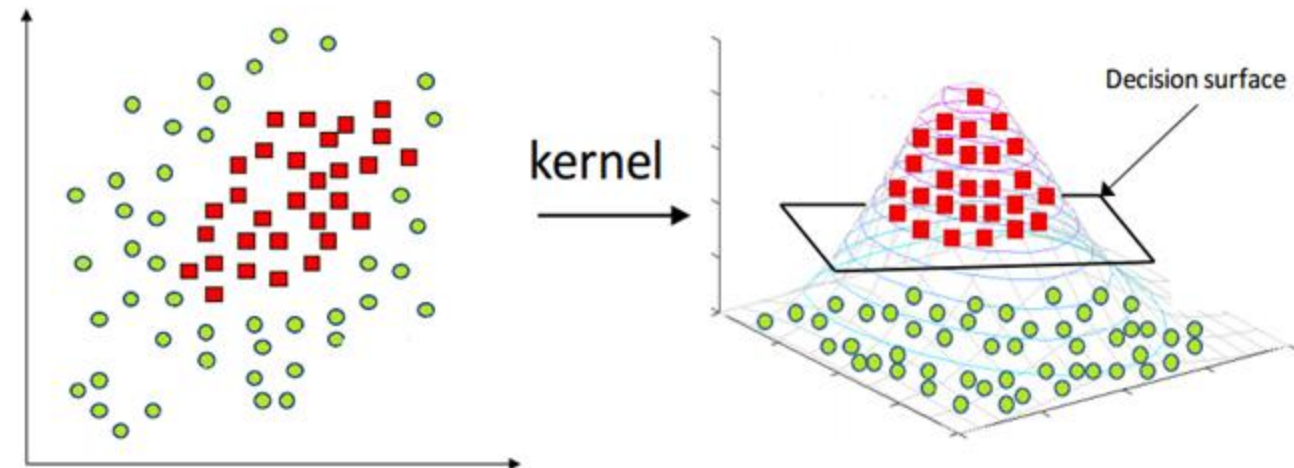
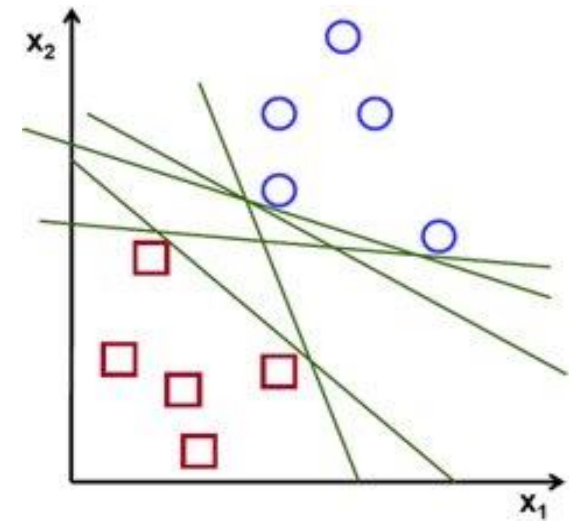
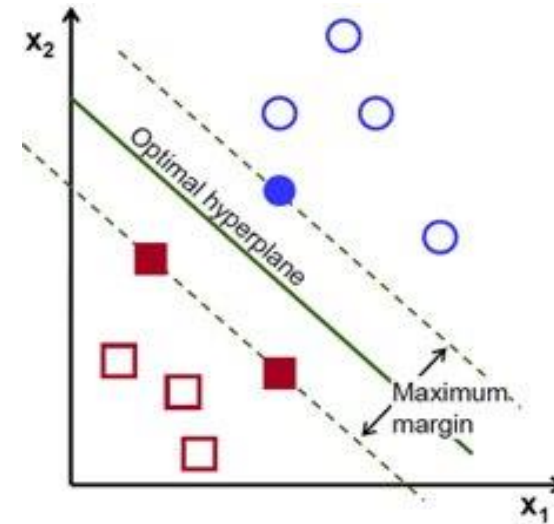
Computationally expensive  
(Time, memory)

# SVM

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes.

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown in the image below

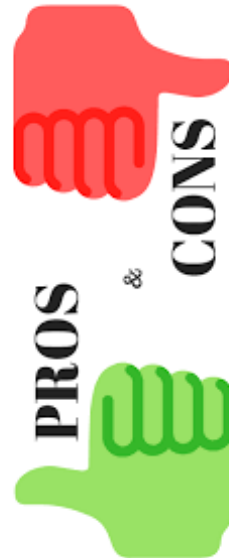
But what happens when there is no clear hyperplane? it's necessary to move away from a 2d view of the data to a 3d view.



# SVM

## Pros:

1. It is really effective in the higher dimension.
2. Effective when the number of features are more than training examples.
3. Best algorithm when classes are separable
4. The hyperplane is affected by only the support vectors thus outliers have less impact.
5. SVM is suited for extreme case binary classification.



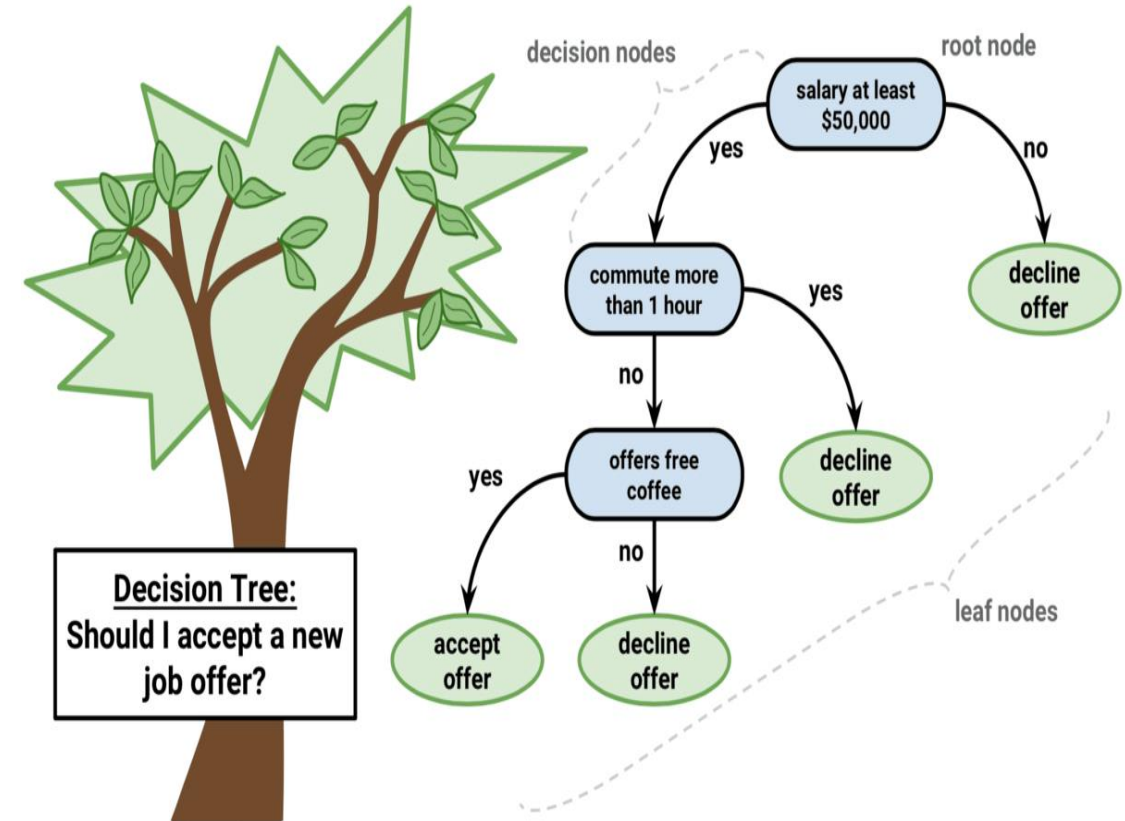
## cons:

1. For larger dataset, it requires a large amount of time to process.
2. Does not perform well in case of overlapped classes.
3. Selecting, appropriately hyperparameters of the SVM that will allow for sufficient generalization performance.
4. Selecting the appropriate kernel function can be tricky.



# Decision Tree

- Decision Tree solves the problem of machine learning by transforming the data into a tree representation.
- A decision tree algorithm can be used to solve both regression and classification problems
- A decision tree is a flowchart-like structure in which:
  - each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails)
  - each leaf node represents a class label (decision taken after computing all features)
  - branches represent conjunctions of features that lead to those class labels.



# Decision Tree

Gini Impurity: What's the probability we classify the datapoint incorrectly

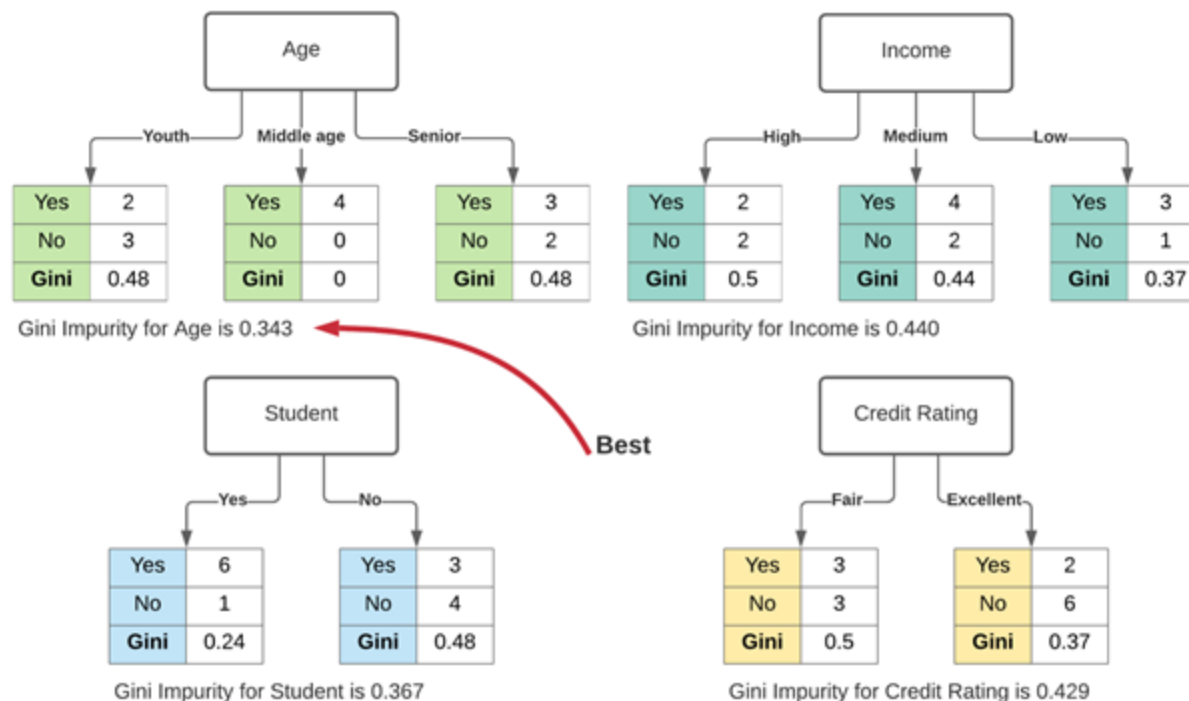
$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

$$= 1 - \sum_{t=0}^t p_t^2$$

Out of 14 instances ,  
yes=9,no=5  
 $1 - (9/14)^2 - (5/14)^2$

$$1 - 0.413 - 0.127 = 0.46$$

Gini = 0.46



## Pros:

- Easy to understand and interpret
- Low effort for data preprocessing (normalization, missing values,..)

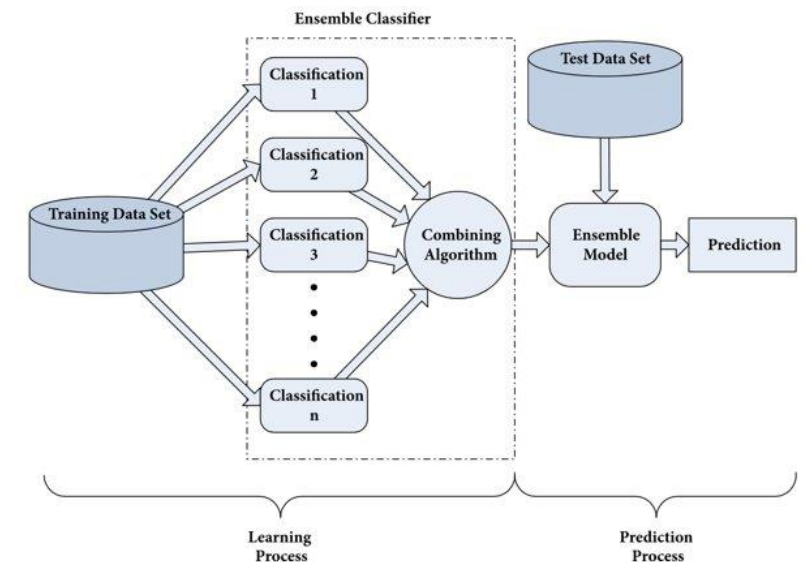


## Cons:

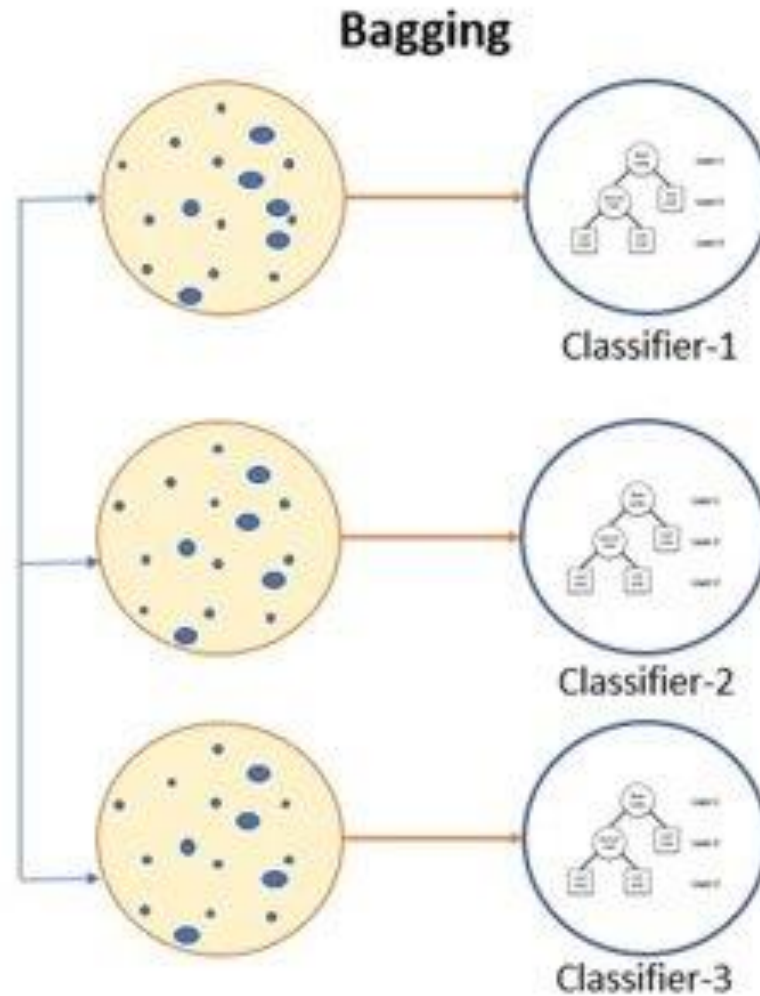
- Computationally expensive (time)
- Performance
- Easy to overfit

# Ensemble Learning Algorithms

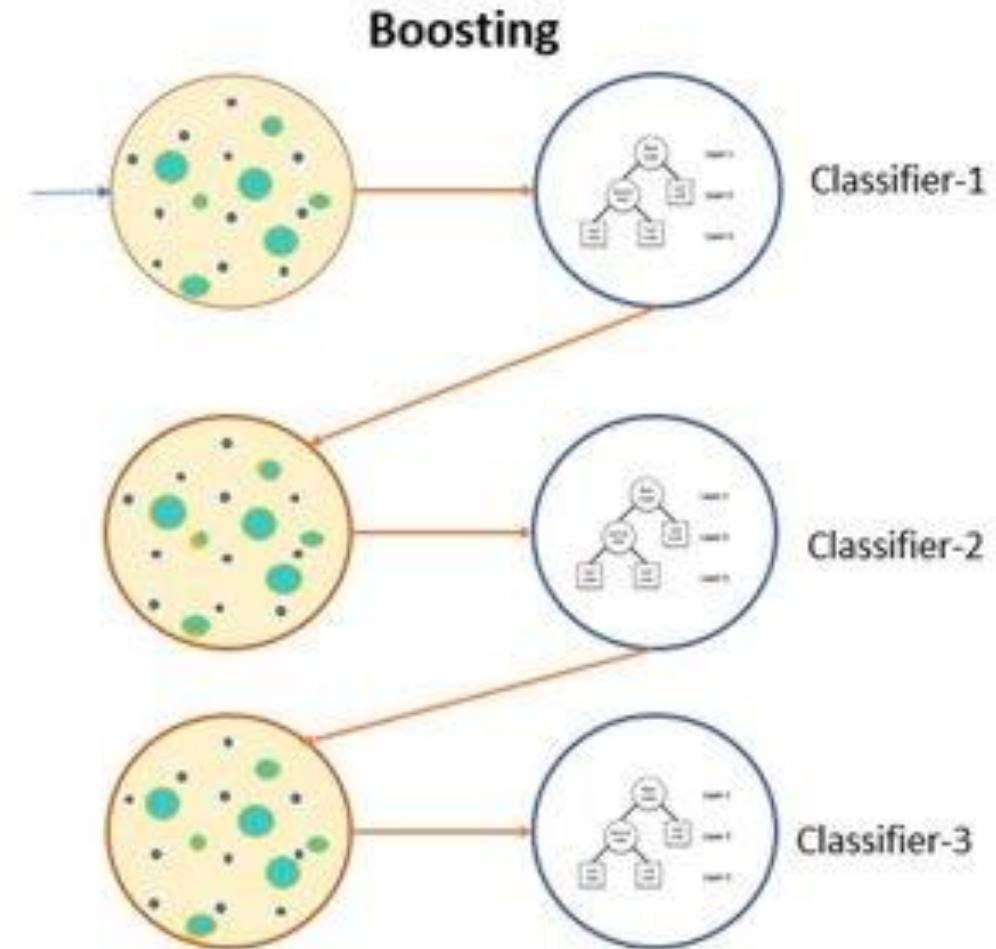
- Ensemble learning is a machine learning paradigm where multiple models (often called “weak learners”) are trained to solve the same problem and combined to get better results.
- The main hypothesis is that when weak models are correctly combined, we can obtain more accurate and/or robust models.



# Ensemble Learning Algorithms

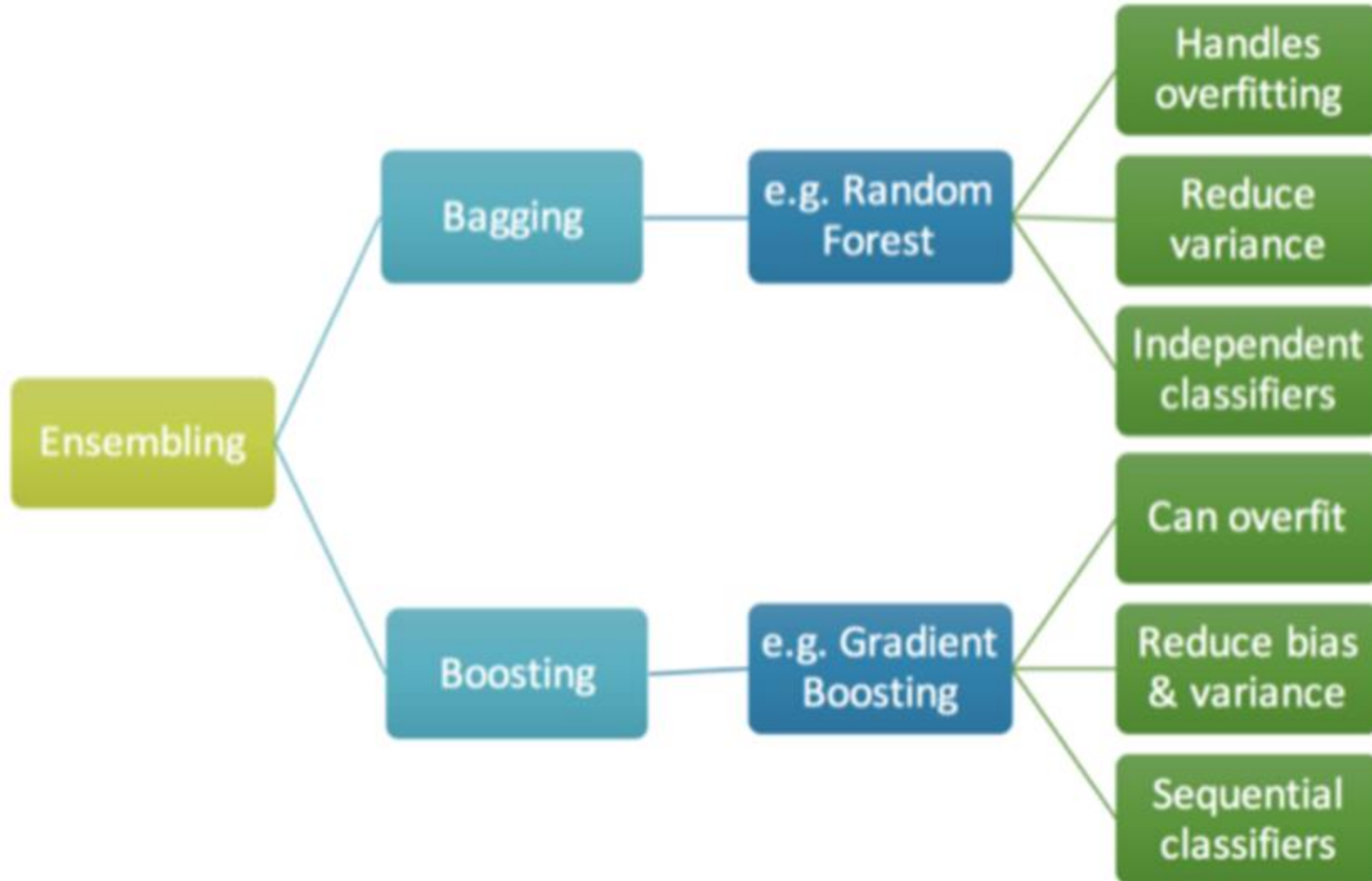


**Parallel**



**Sequential**

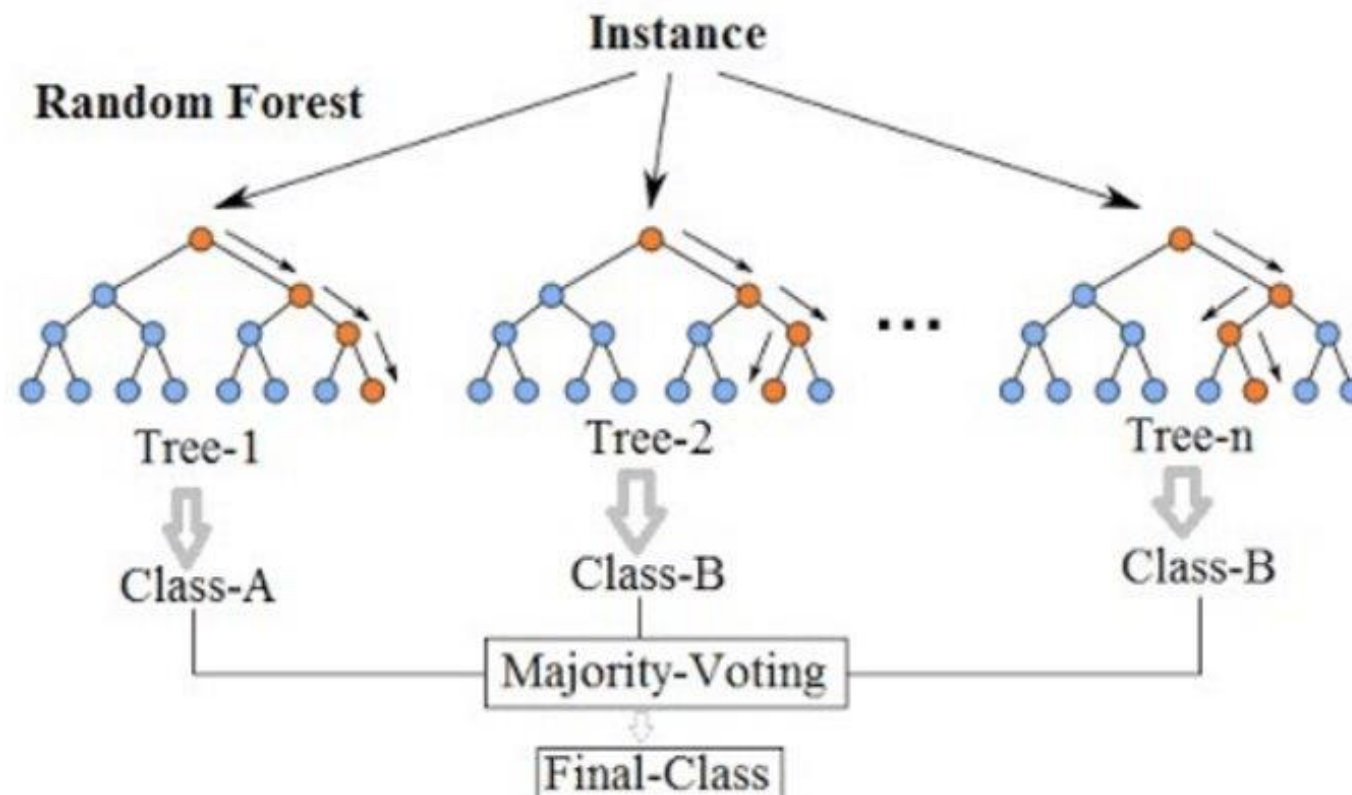
# Ensemble Learning Algorithms





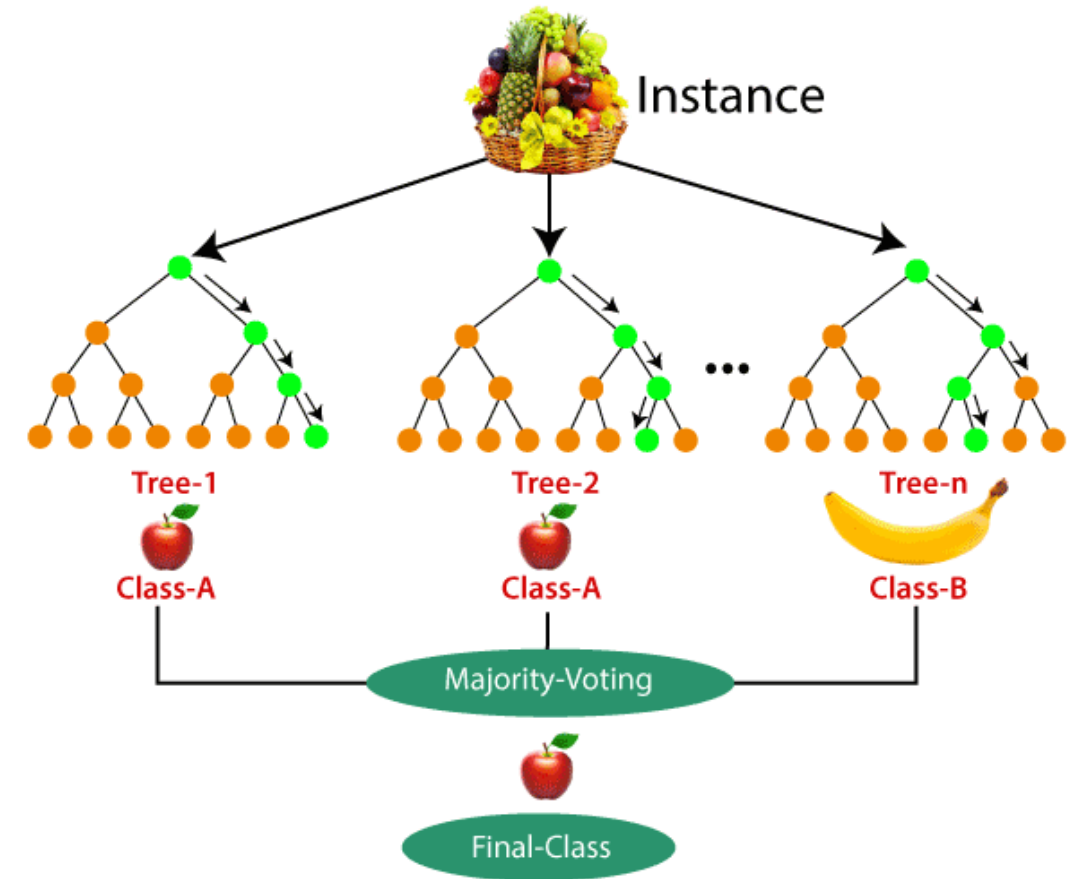
# Random Forest

- Random Forest is an ensemble learning method that operates by constructing multiple decision trees.
- The final decision is made based on the majority of the trees and is chosen by the random forest.
  - Step 1 – First, start with the selection of random samples from a given dataset.
  - Step 2 – Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
  - Step 3 – In this step, voting will be performed for every predicted result.
  - Step 4 – At last, select the most voted prediction result as the final prediction result.



# Random Forest

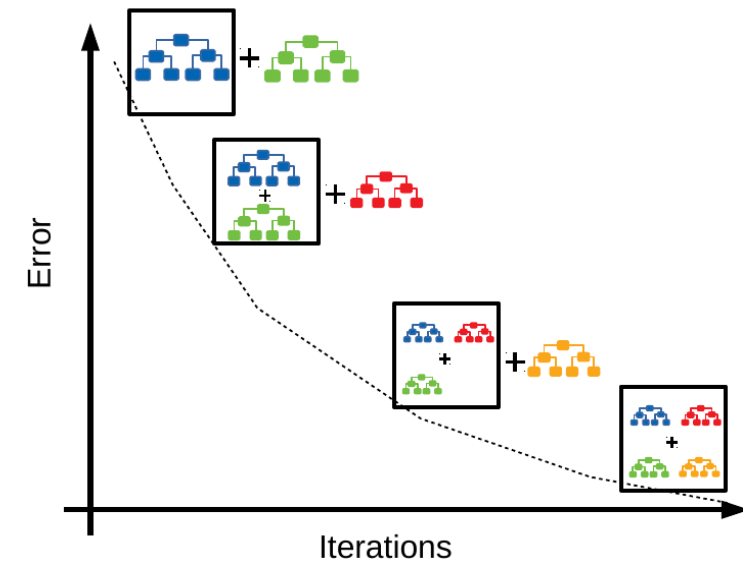
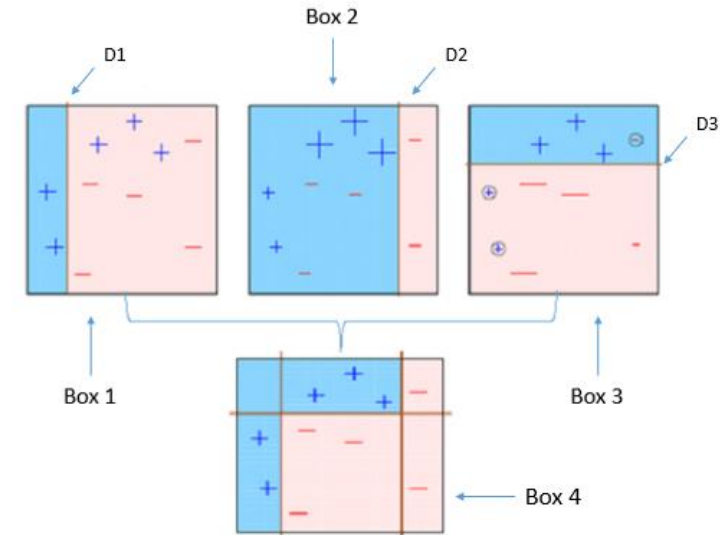
- Pros:
  - Robust to outliers.
  - Works well with non-linear data.
  - Lower risk of overfitting.
  - Runs efficiently on a large dataset.
  - Better accuracy than other classification algorithms.
- Cons:
  - Not descriptive
  - Slow → big number of tree
  - Not suitable for linear methods with a lot of sparse features





# XGBOOST

- XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.
  - Box 1: The first classifier creates a vertical line (split) at D1.
  - Box 2: The next classifier says don't worry I will correct your mistakes. Therefore, it gives more weight to the three + misclassified points (see bigger size of +) and creates a vertical line at D2.
  - Box 3: The next classifier continues to bestow support.
  - Box 4: It is a weighted combination of the weak classifiers. As you can see, it does good job at classifying all the points correctly.
- Should we use it?
  - Pros: High Performance,
  - Cons: Sensitive to outliers, hard to scale up.



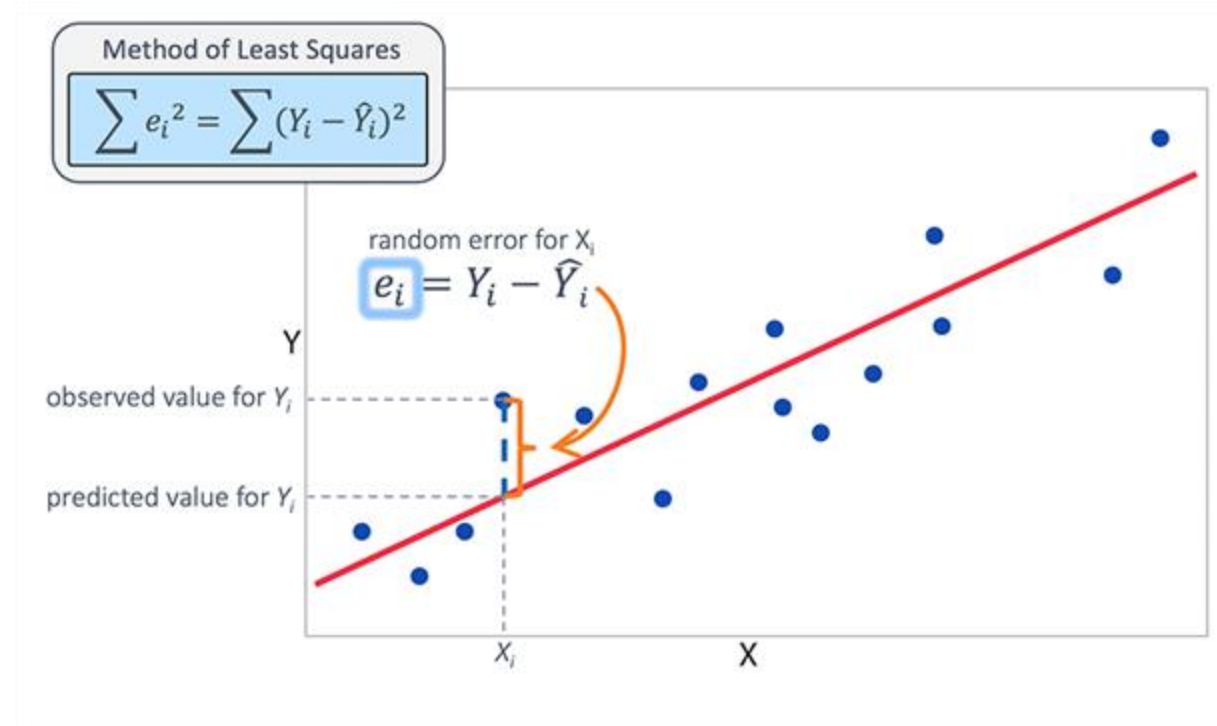
# Regression

ML can also be used to predict continuous target variables

All three classifiers that we have learnt so far can be modified to do that But there are different algorithms just used for this task

Target variable is continuous so there is no confusion matrix nor predicted probabilities

Predict the value of a dependent variable based on the value of at least one independent variable



Predicted value or criterion

Predictor

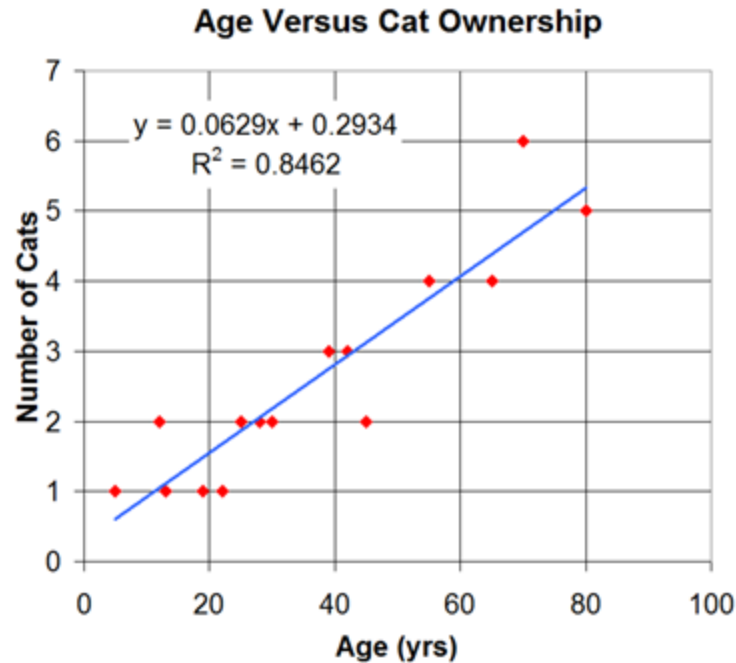
$$Y' = bX + a$$

The slope

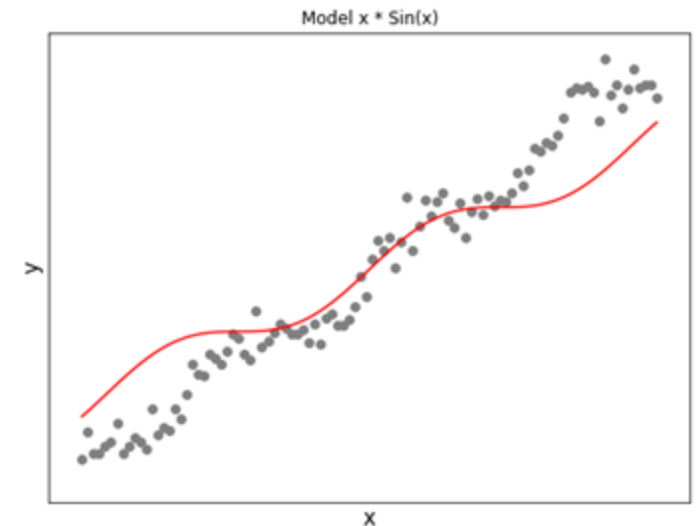
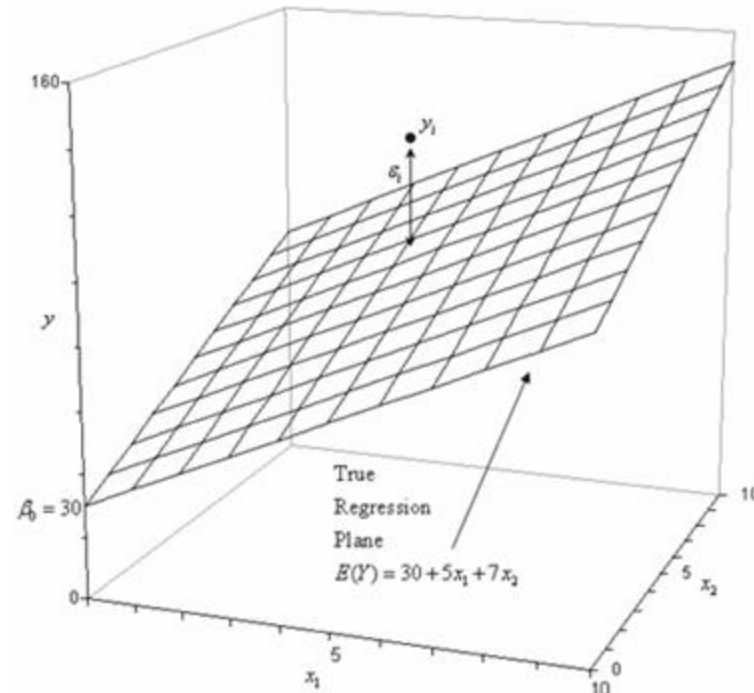
The Y-intercept

# Linear vs Nonlinear Regression vs Multi Regression

## Multi Regression



linear



nonlinear

$$\hat{y} = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon$$

target

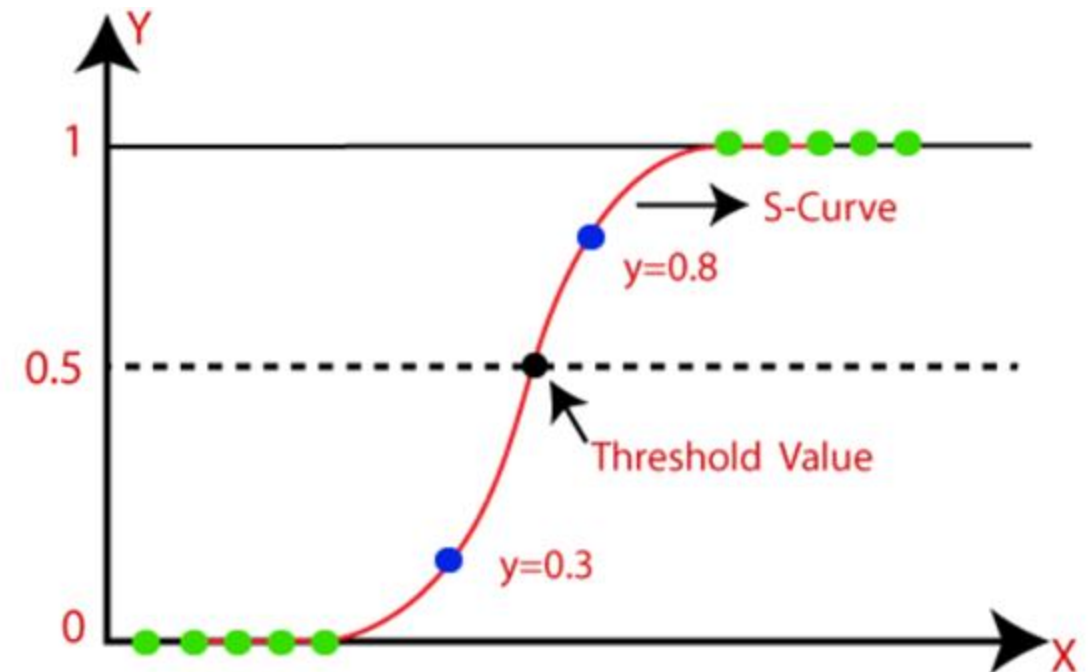
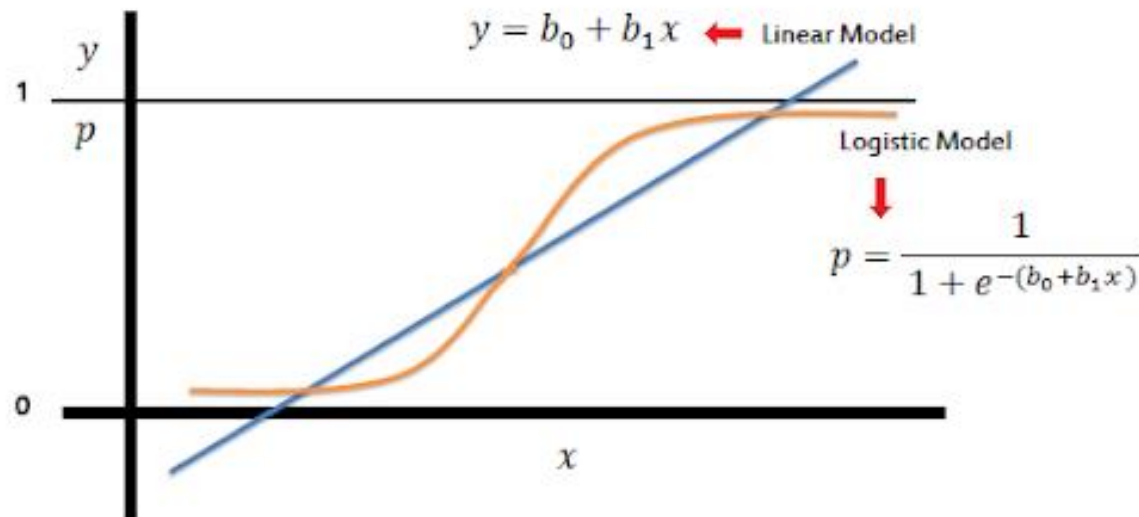
coefficients

inputs

random error

# Logistic Regression

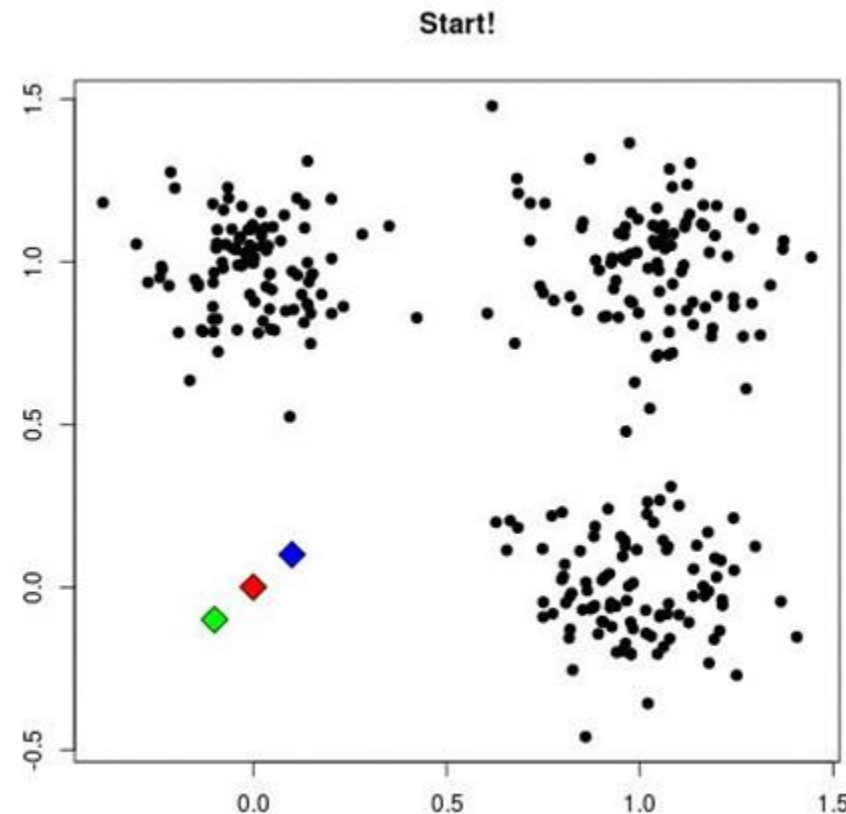
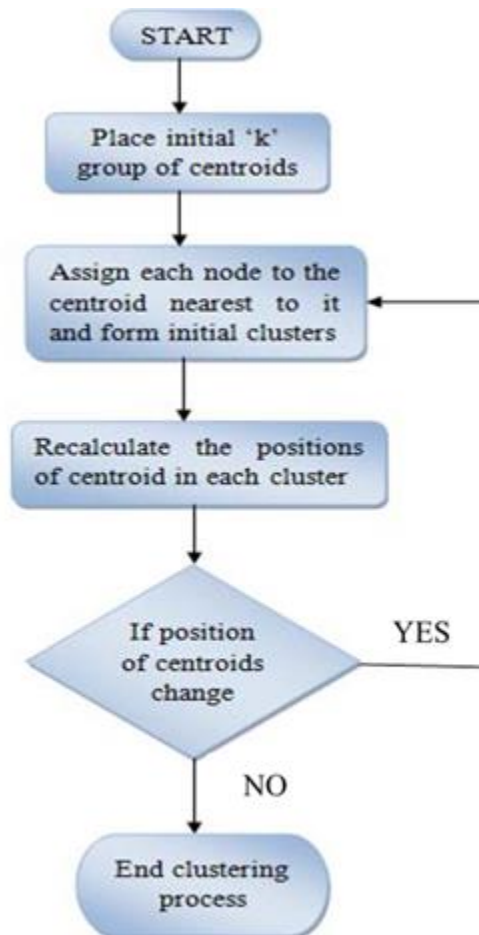
- The other type of regression analysis is logistic regression.
- Logistic regression is essentially used to calculate (or predict) the probability of a binary (yes/no) event occurring.



# Unsupervised ML Algorithms

# K Means

K Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.



**Pros:**  
**Low complexity**

**Cons:**

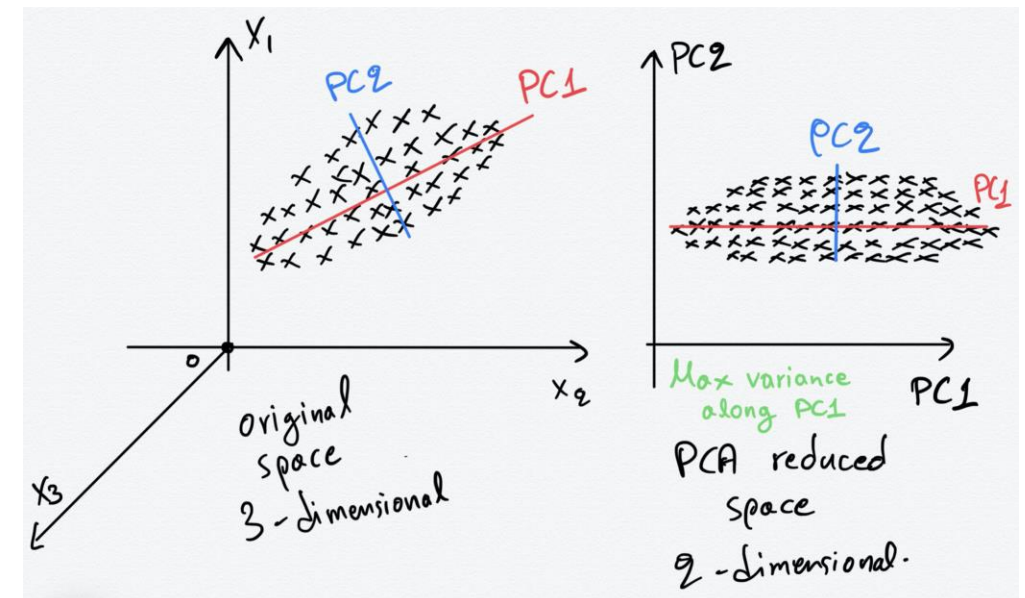
- **Necessity of specifying k**
- **Sensitive to noise and outliers**
- **Sensitive to initial centroids assignments**

# Principal Component Analysis

- Although big data analytics is used in bettering many aspects of human life, it comes with its own problems. One of them is 'Curse of dimensionality'.
- Principal Component Analysis (PCA) is one of the most popular linear dimension reduction.
- it finds best linear combinations of the original variables so that the variance or spread along the new variable is maximum.

## Algorithm:

- Step 1: Get the data from  $m \times n$  matrix  $A$
- Step 2: Calculate the covariance matrix
- Step 3: Calculate the eigenvectors and eigenvalues of the covariance matrix
- Step 4: Choosing principal components and forming a feature vector
- Step 5: Deriving the new data set and forming the clusters







**hands-on**  
TRAINING





# Q&A

contact me



[soltanin@ualberta.ca](mailto:soltanin@ualberta.ca)  
[Sara@altaml.com](mailto:Sara@altaml.com)

*Wants to know more about AltaML*  
<https://www.altaml.com>



*Never  
Stop  
Learning*



Thank  
You