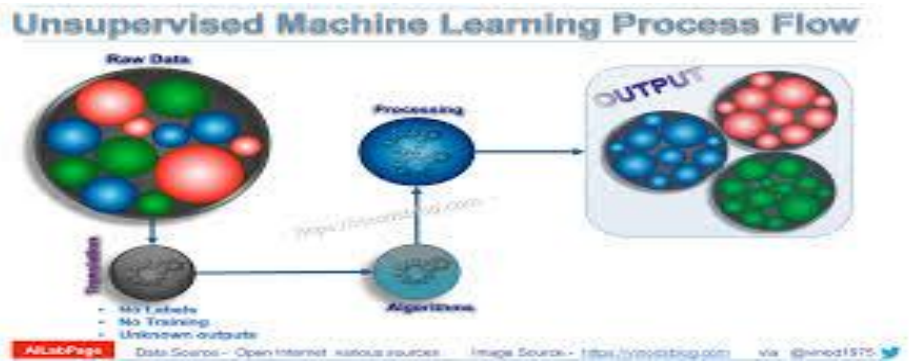
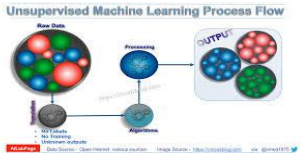


Unsupervised Learning

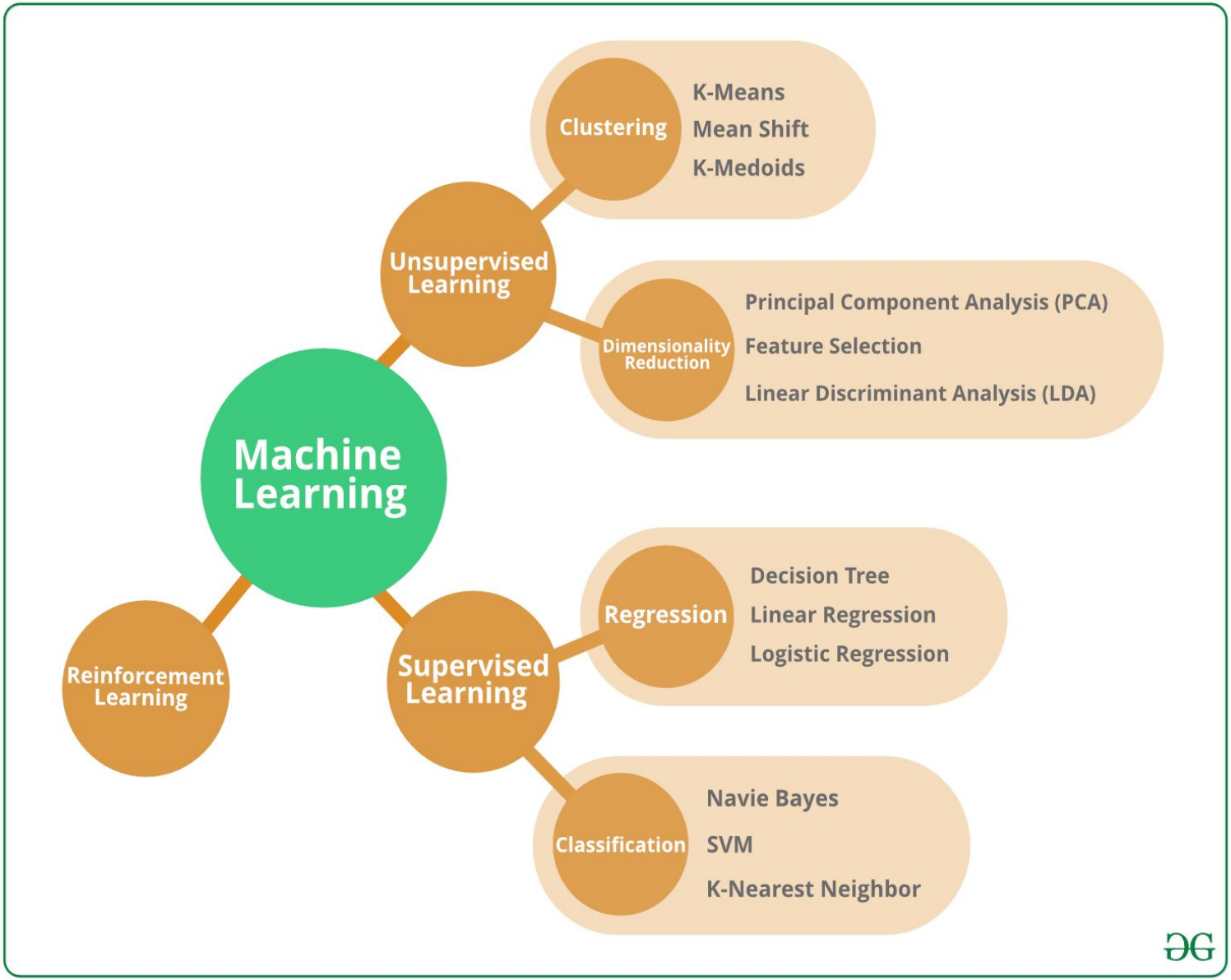
Agenda



INTERNSHIPSTUDIO

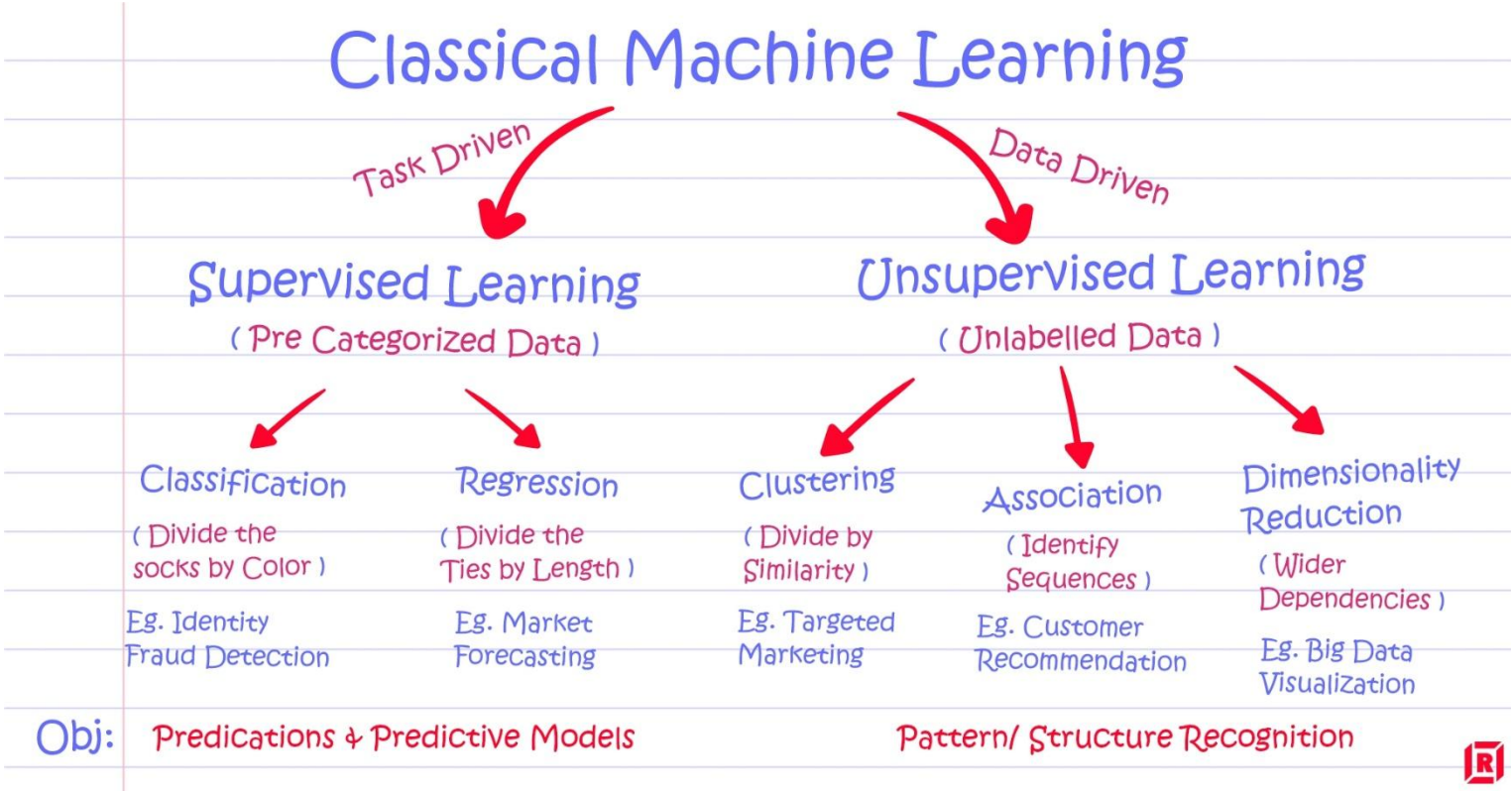
- Difference between Supervised and Unsupervised learning
- K-means Clustering
- Finding optimal number of clusters
- Hierarchical clustering

Recap- ML Algorithms

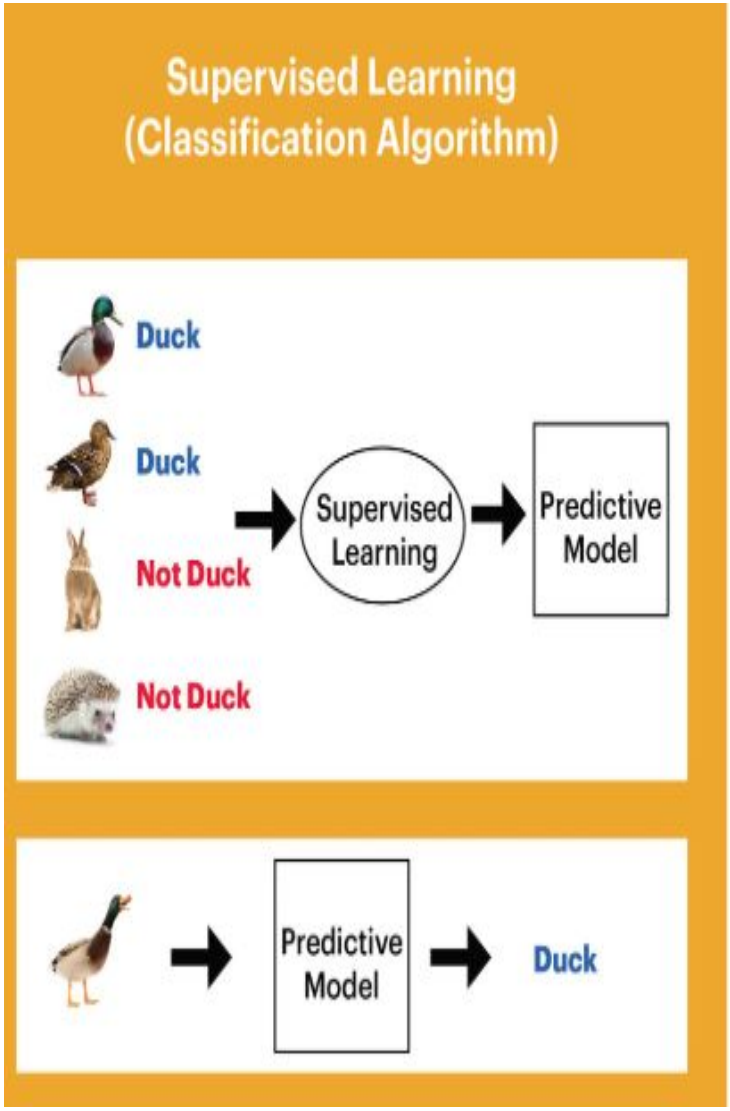


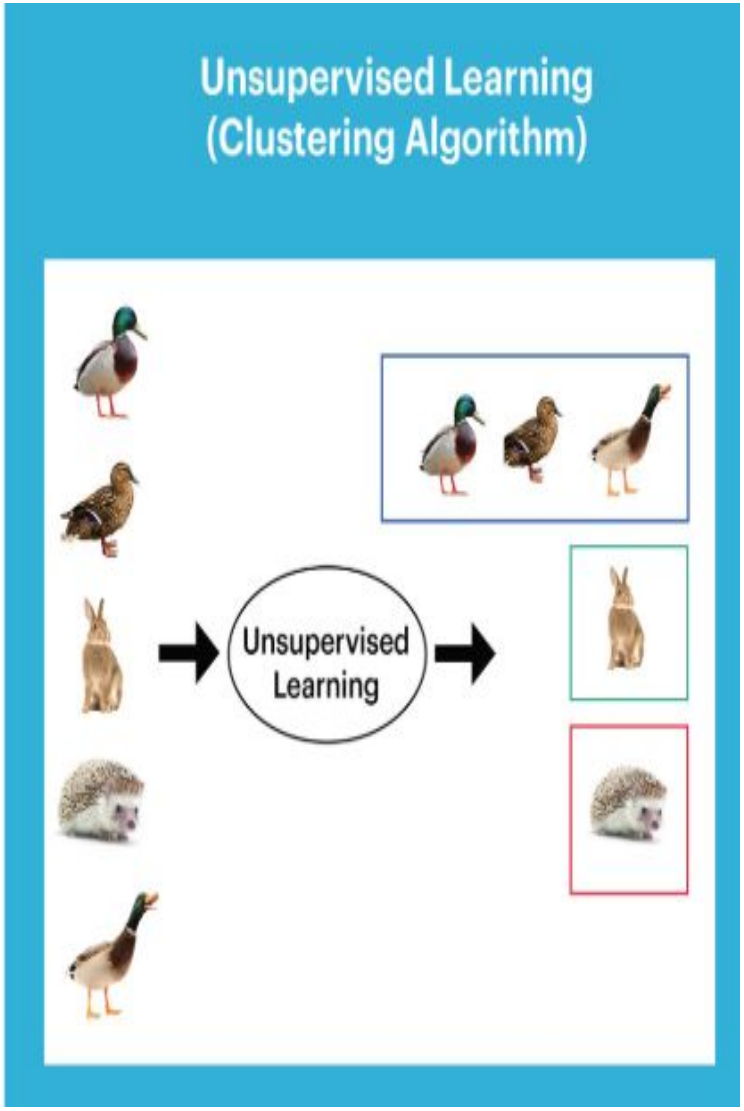
INTERNSHIPSTUDIO

Recap – ML Methods

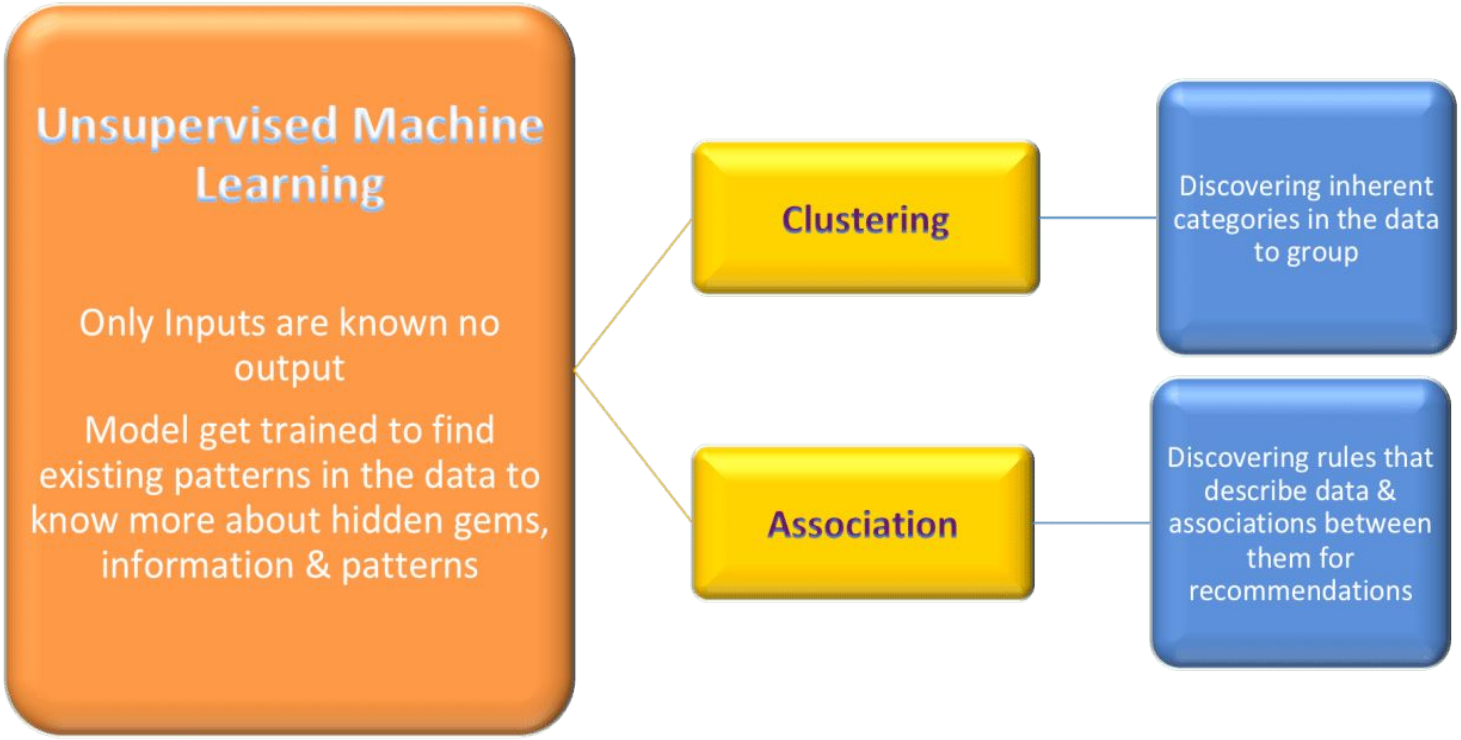


Examples - Unsupervised learning



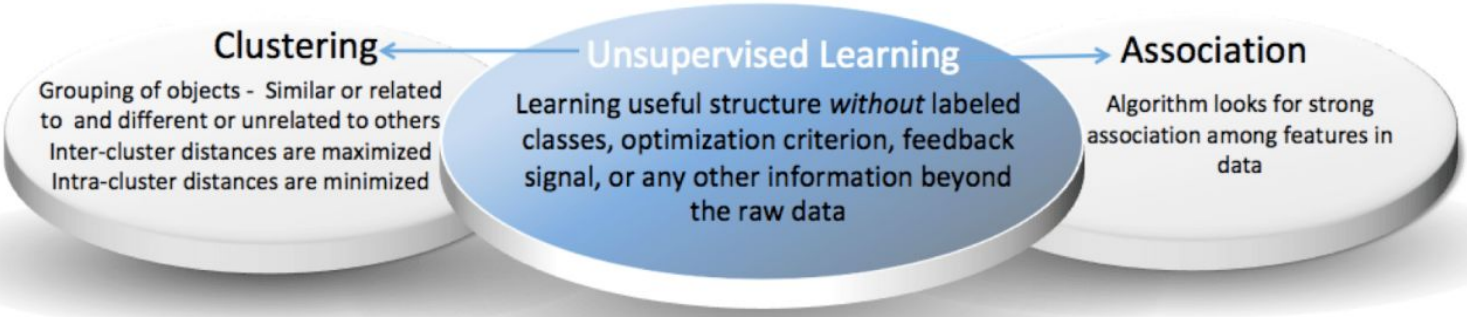


Unsupervised Machine Learning – Screen Shot



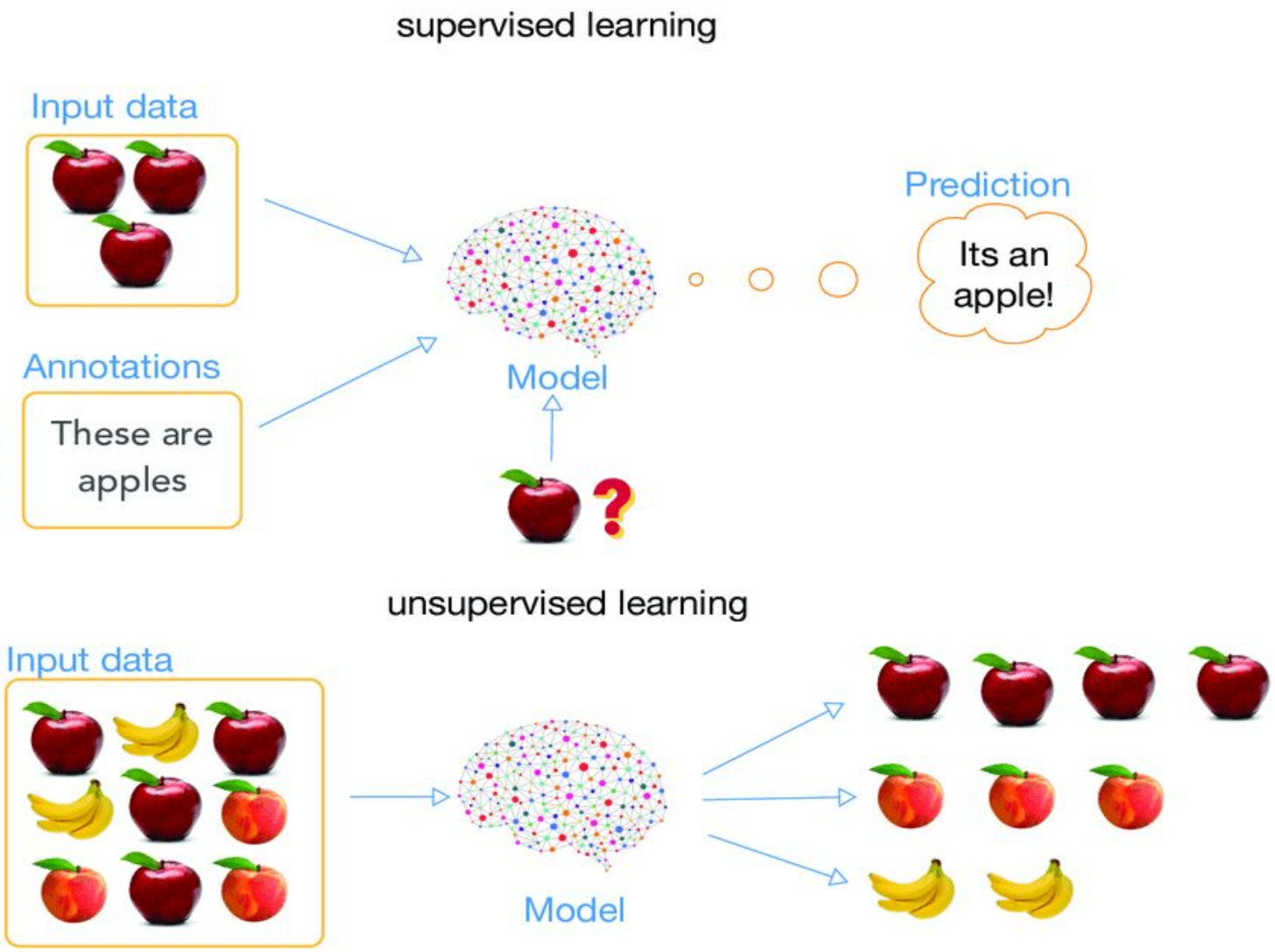
AILabPage Source - Open Internet various sources Image Source - https://vinodsblog.com via @vinod1975

Unsupervised Machine Learning Types



INTERNSHIPSTUDIO

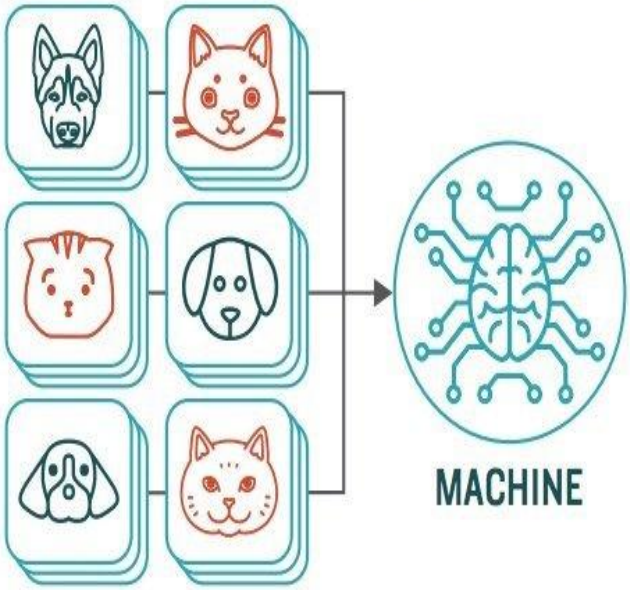
Examples - Unsupervised learning



How **Unsupervised** Machine Learning Works

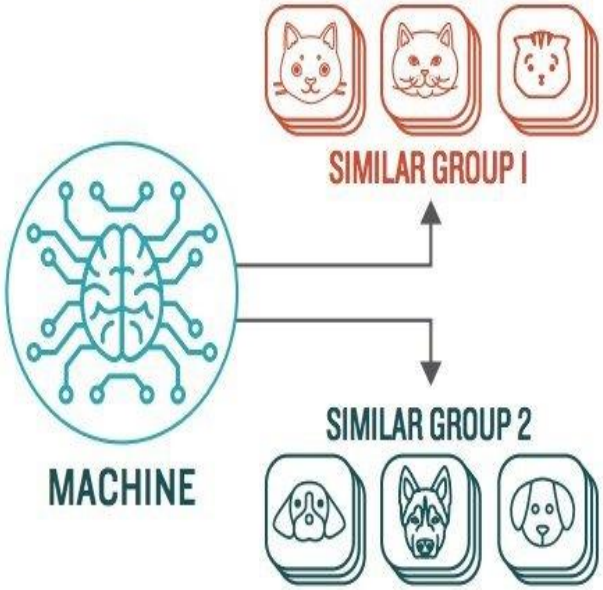
STEP 1

Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

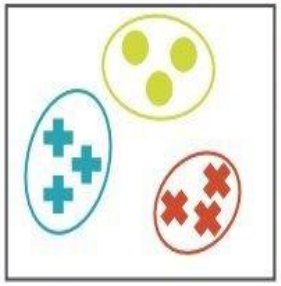


STEP 2

Observe and learn from the patterns the machine identifies



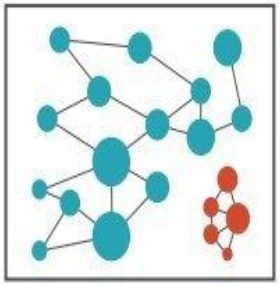
TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?



ANOMALY DETECTION

Identifying abnormalities in data

For Example: Is a hacker intruding in our network?

Difference between Supervised and Unsupervised learning

INTERNSHIPSTUDIO

#1. Method

Supervised Learning

Input
variables and output variables
will be given.

Unsupervised Learning

Only
input data will be given.

#2. Goal

Supervised Learning

Supervised
learning goal is to determine the function so well that when new input data set given, can predict the output.

Unsupervised Learning

Unsupervised
learning goal is to model the hidden patterns or underlying structure in the given input data in order to learn about the data.

INTERNSHIPSTUDIO

Parameters	Supervised machine learning technique	Unsupervised machine learning technique
Process	In a supervised learning model, input and output variables will be given.	In unsupervised learning model, only input data will be given
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labeled
Algorithms Used	Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees.	Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc.
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex
Use of Data	Supervised learning model uses training data to learn a link between the input and the outputs.	Unsupervised learning does not use output data.
Accuracy of Results	Highly accurate and trustworthy method.	Less accurate and trustworthy method.
Real Time Learning	Learning method takes place offline.	Learning method takes place in real time.
Number of Classes	Number of classes is known.	Number of classes is not known.
Main Drawback	Classifying big data can be a real challenge in Supervised	You cannot get precise information regarding data sorting, and the output

INTERNSHIPSTUDIO

Q.1 What is Unsupervised learning?

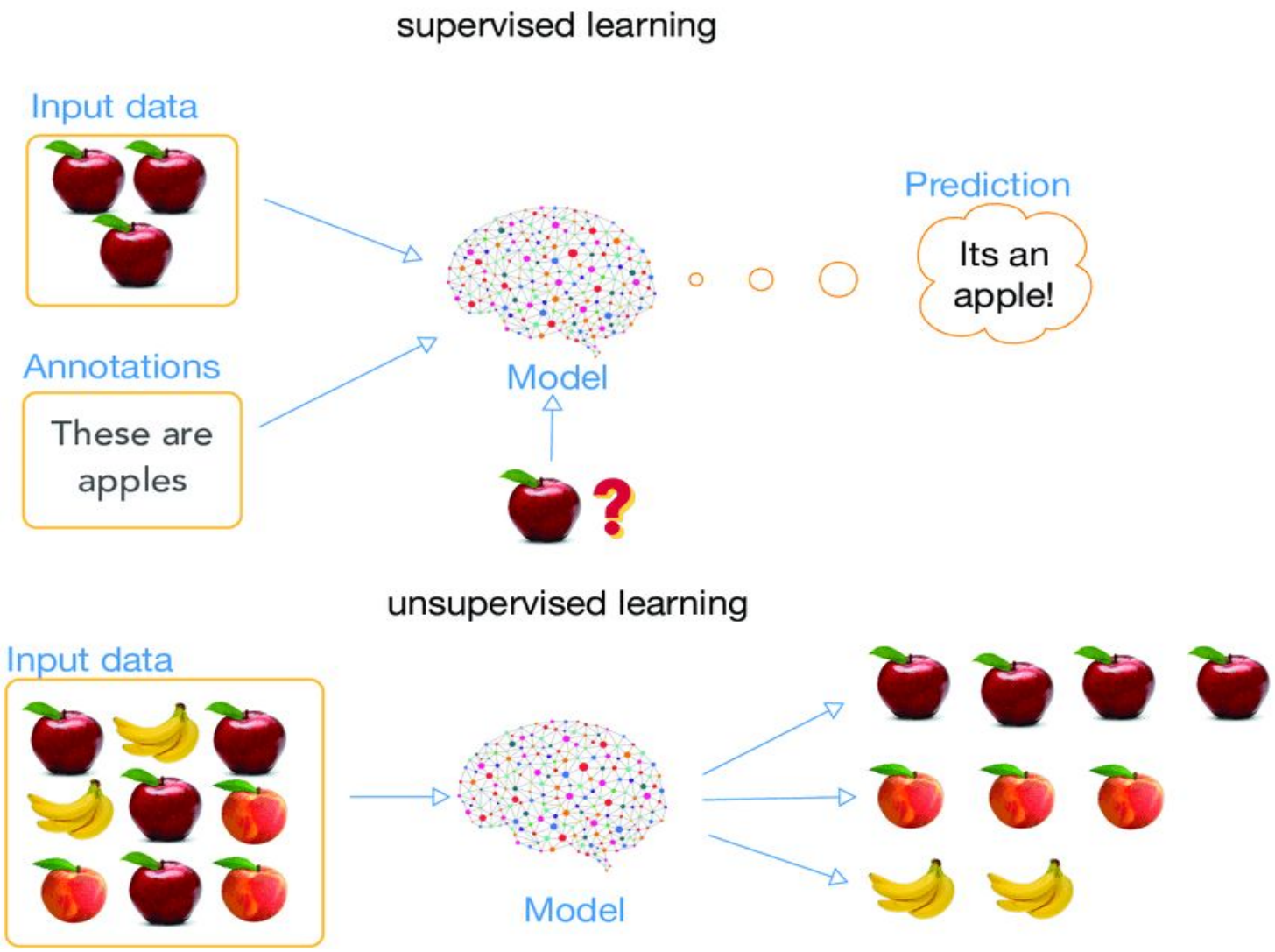
Q.2 What are the types of Supervised and Unsupervised learning?

Q.3 What is difference between the Supervised and Unsupervised learning?

Q.4 What are the advantages and disadvantages of Supervised and Unsupervised learning?

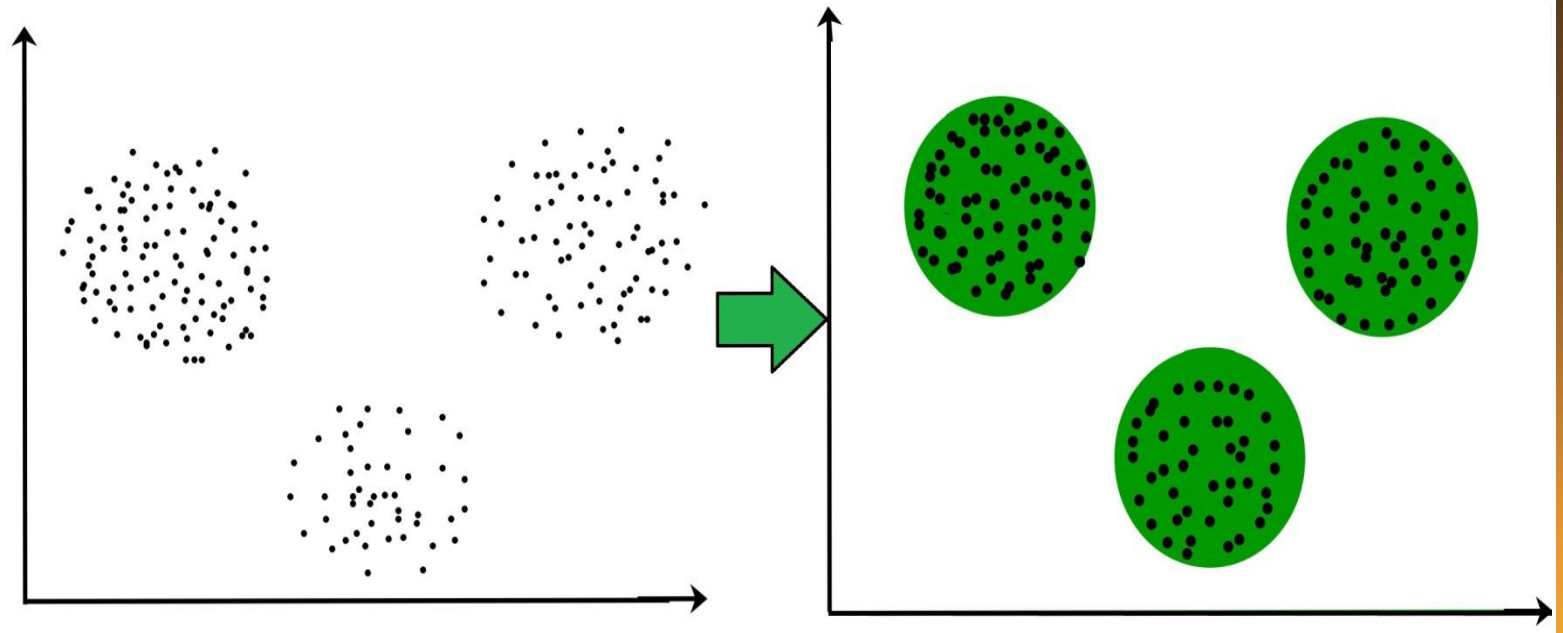
Q.5 Provide few example of Unsupervised learning?

Examples - Unsupervised learning



What is Clustering?

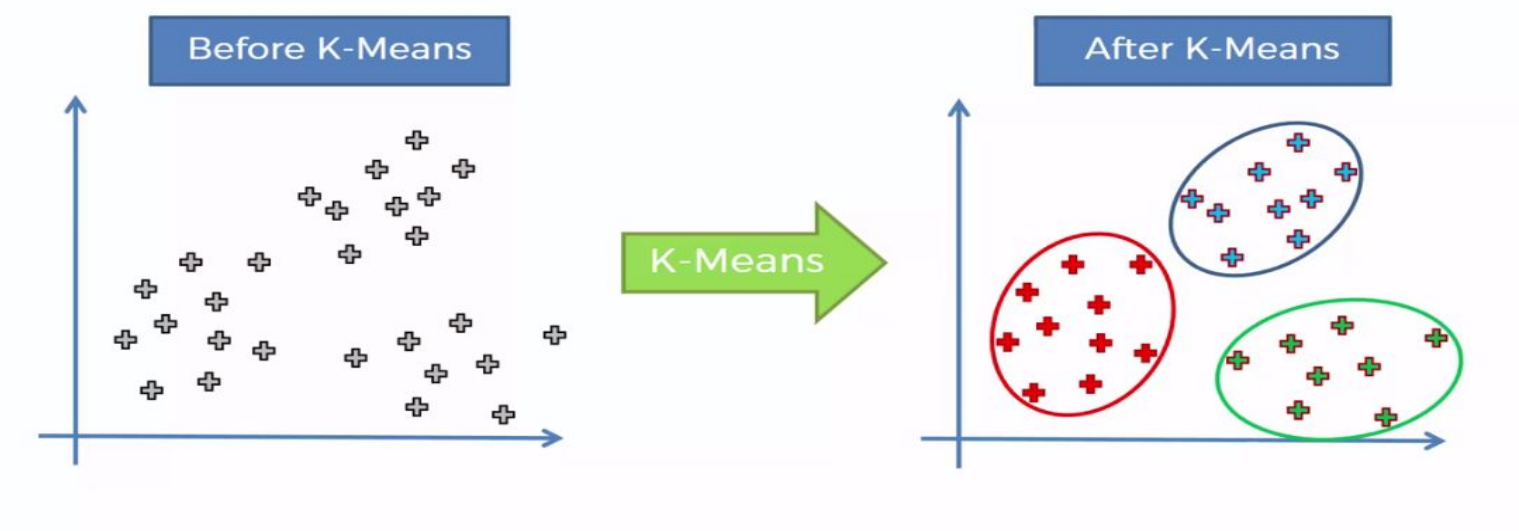
- **Clustering** is the task of dividing the data points into a number of groups such that
  - data points in the same groups are more similar to other data points in the same group
  - It is basically a collection of objects on the basis of similarity and dissimilarity between them.
- **Example-** The data points in the graph below clustered together can be classified into one single group.



K-means Clustering

It partitions the data set such that-

- Each data point belongs to a cluster with the nearest mean.
- Data points belonging to one cluster have high degree of similarity.
- Data points belonging to different clusters have high degree of dissimilarity.



K-means Clustering

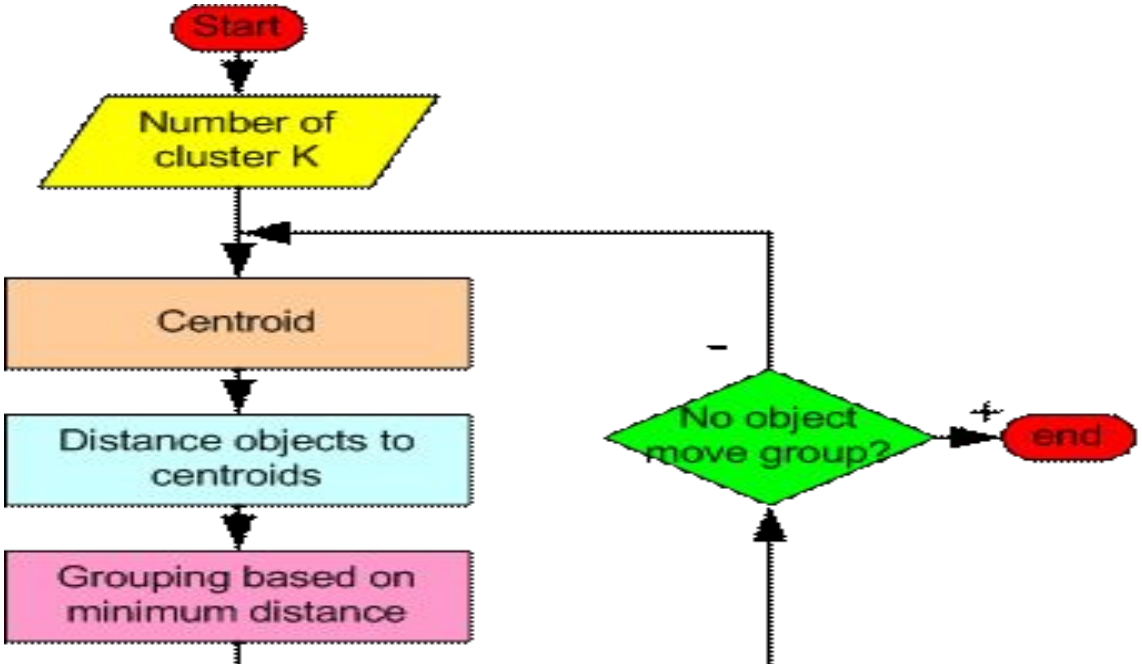
- K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart.

- Creating and optimizing clusters continues till-
  - The centroids have stabilized — there is no change in their values because the clustering has been successful.
  - The defined number of iterations has been achieved.

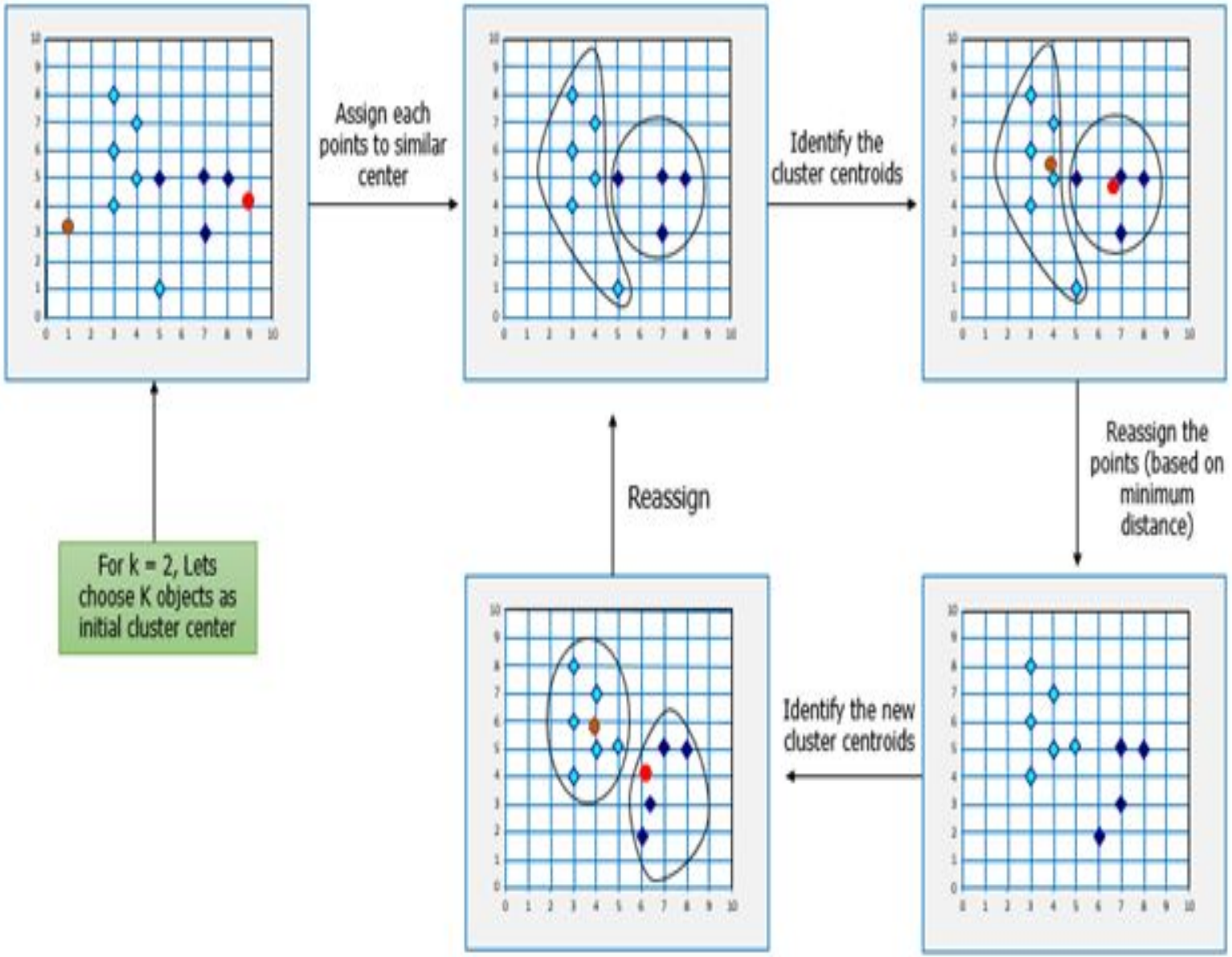
K-means Working

K-means algorithm can be executed in the following steps:

- Partition of objects into k subsets
- Identifying the cluster centroids (mean point) of the current partition.
- Assign each point to a specific cluster
- Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
- After re-allotting the points, find the centroid of the new cluster formed.

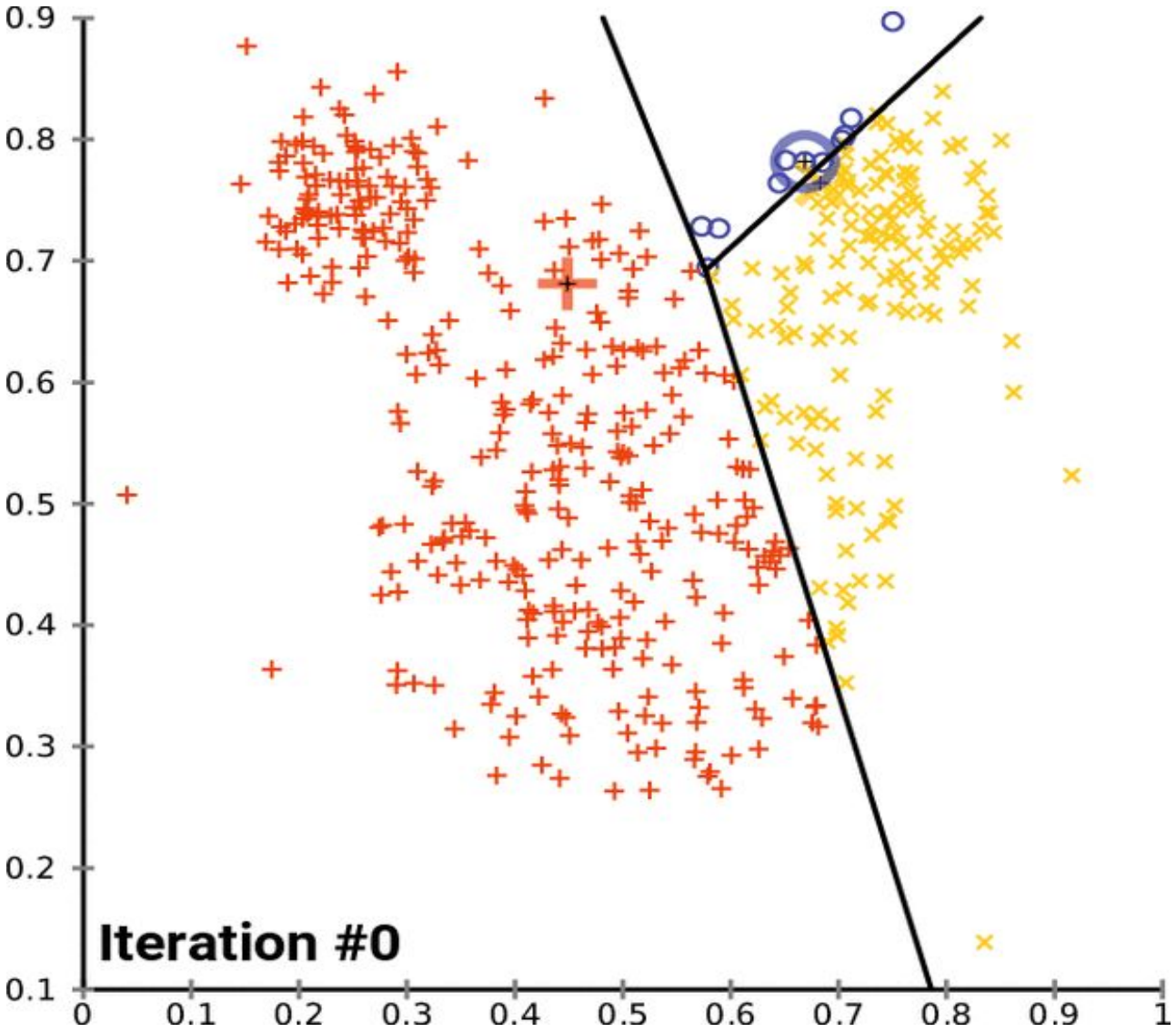


Step by step process



INTERNSHIPSTUDIO

Step by step clustering



INTERNSHIPSTUDIO

Advantages of K-means Clustering

- If variables are huge, then K-Means most of the times computationally faster if we keep k smalls.
- K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Disadvantages

- Difficult to predict K-Value.
- With global cluster, it didn't work well.
- Different initial partitions can result in different final clusters.
- It does not work well with clusters of Different size/density

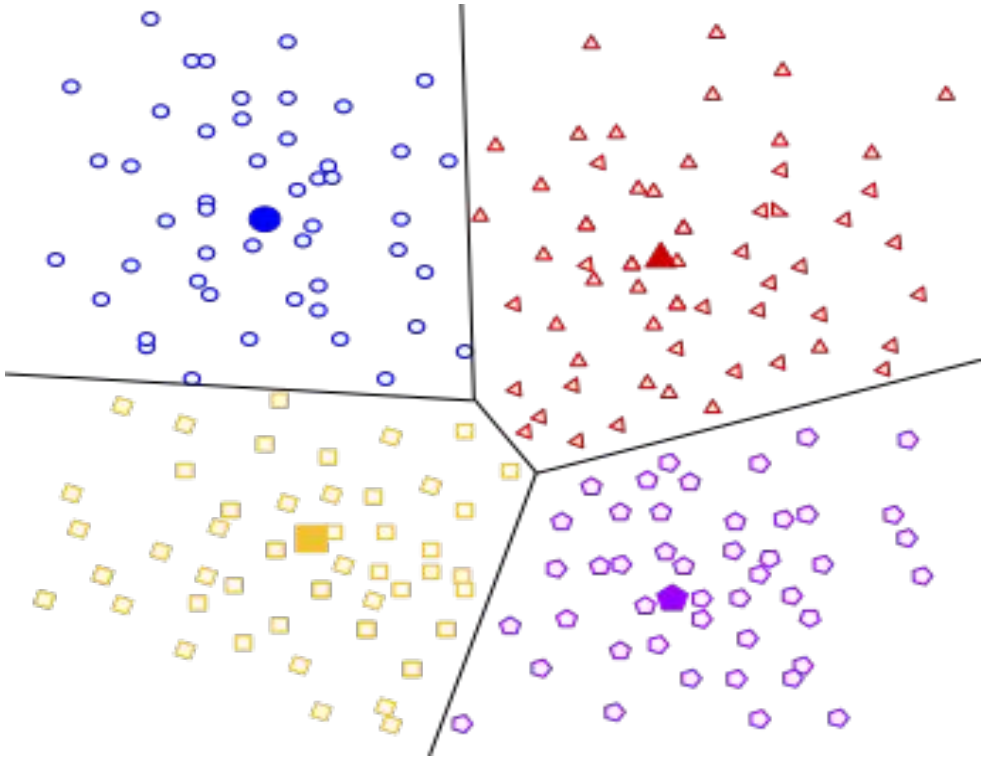
Applications of Clustering

- **Marketing :** Characterize & discover customer segments
- **Biology :** Classification among different species of plants and animals.
- **Libraries :** Clustering books on the basis of topics and information.
- **Insurance :** Acknowledge the customers, their policies and identifying the frauds.
- **City Planning:** To make groups of houses and to study their values based on their geographical locations /other factors
-
- **Earthquake studies:** By learning the earthquake-affected areas we can determine the dangerous zones.

Types of Clustering

INTERNSHIPSTUDIO

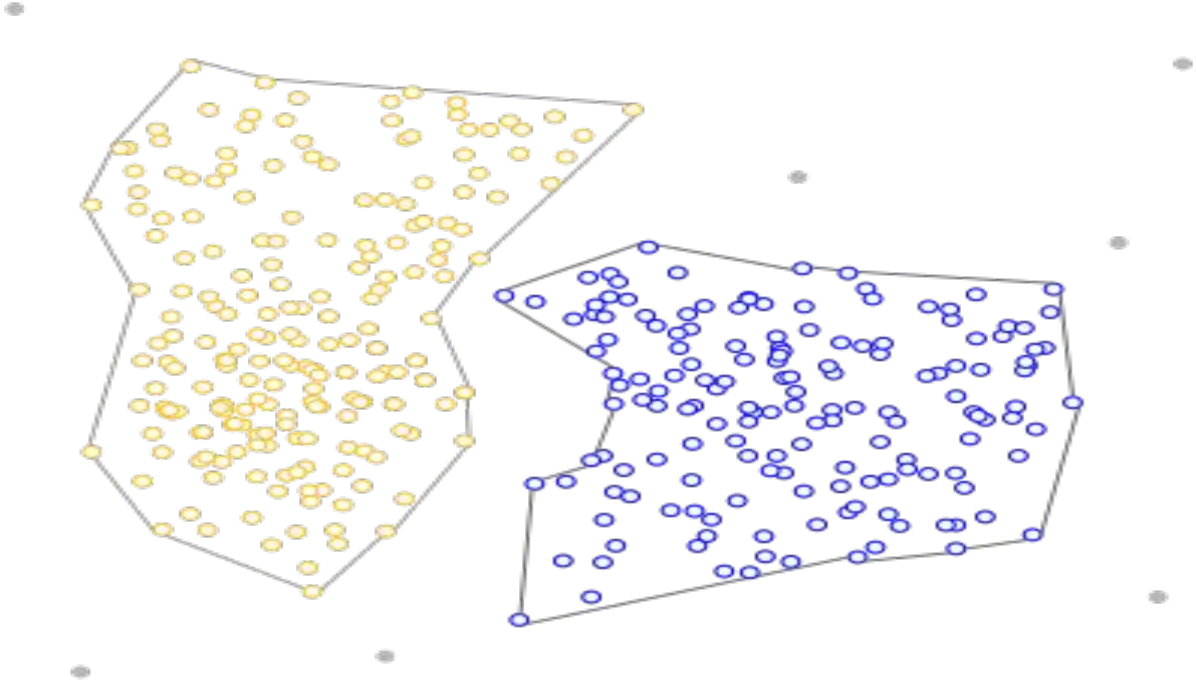
- **Centroid-based clustering** organizes the data into non-hierarchical clusters
- K-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers.



Types of Clustering

Density-based Clustering: Connects areas of high example density into clusters.

- This allows for arbitrary-shaped distributions as long as dense areas can be connected.
- These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.

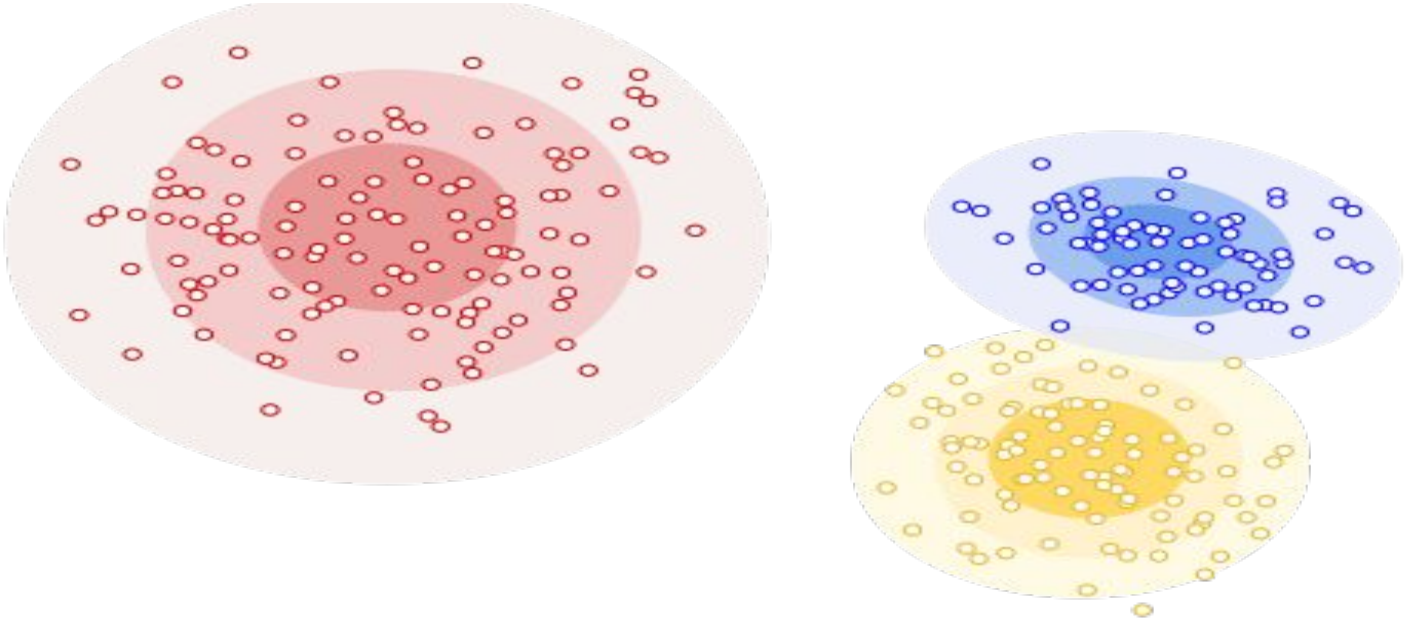


Types of Clustering

INTERNSHIPSTUDIO

Distribution-based Clustering: This clustering approach assumes data is composed of distributions, such as **Gaussian distributions**.

- In Figure , the distribution-based algorithm clusters data into three Gaussian distributions.
- As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases. The bands show that decrease in probability.



INTERNSHIPSTUDIO

Q.1 Explain the concept of clustering?

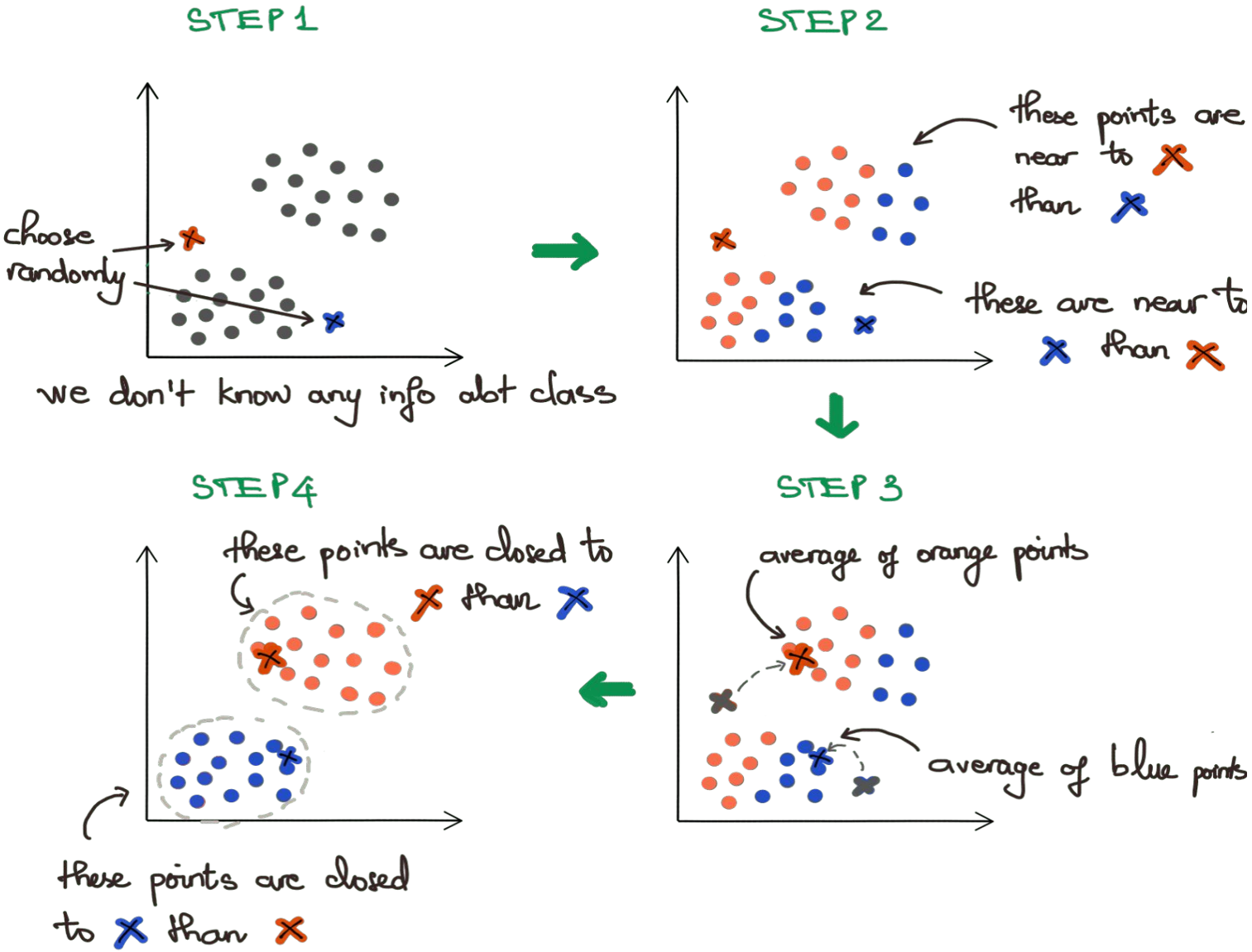
Q.2 What is K-means clustering?

Q.3 Explain the steps of K-means clustering ?

Q.4 What are the advantages applications of K-means clustering ?

Q.5 What are the types of K-means clustering?

Recap- K-means clustering



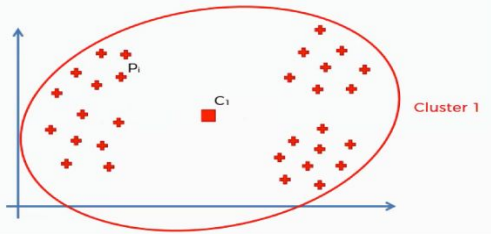
Deciding cluster numbers

WCSS (within-cluster sum of squares) helps us to determine the optimal number of clusters

$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

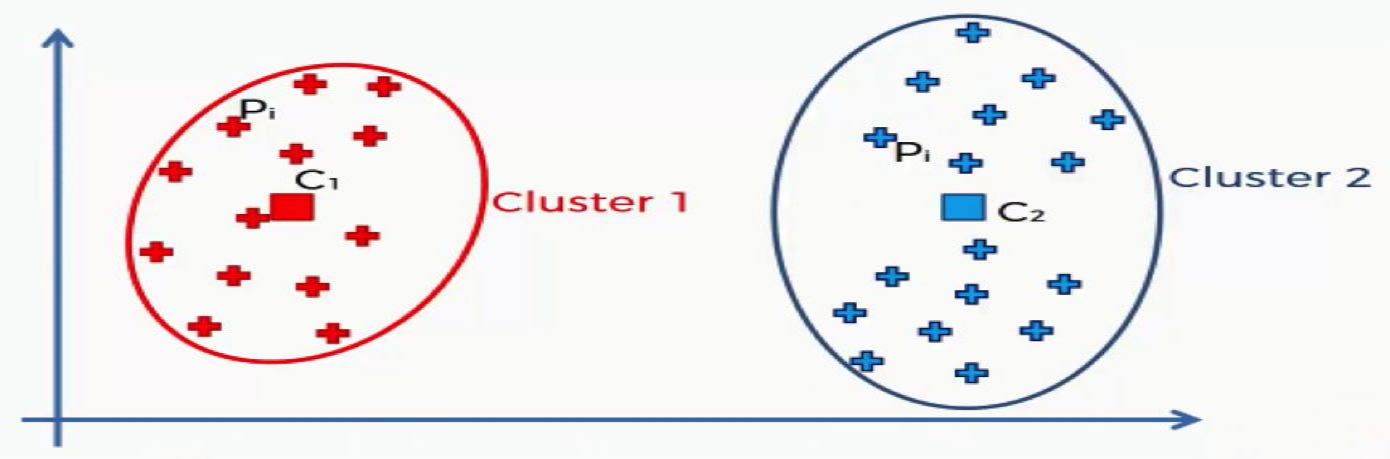
What this equation signifies is this: For Cluster 1, we'll take every point (Pi) that falls within the cluster, and calculate the distance between that point and the centroid (C1) for Cluster 1.

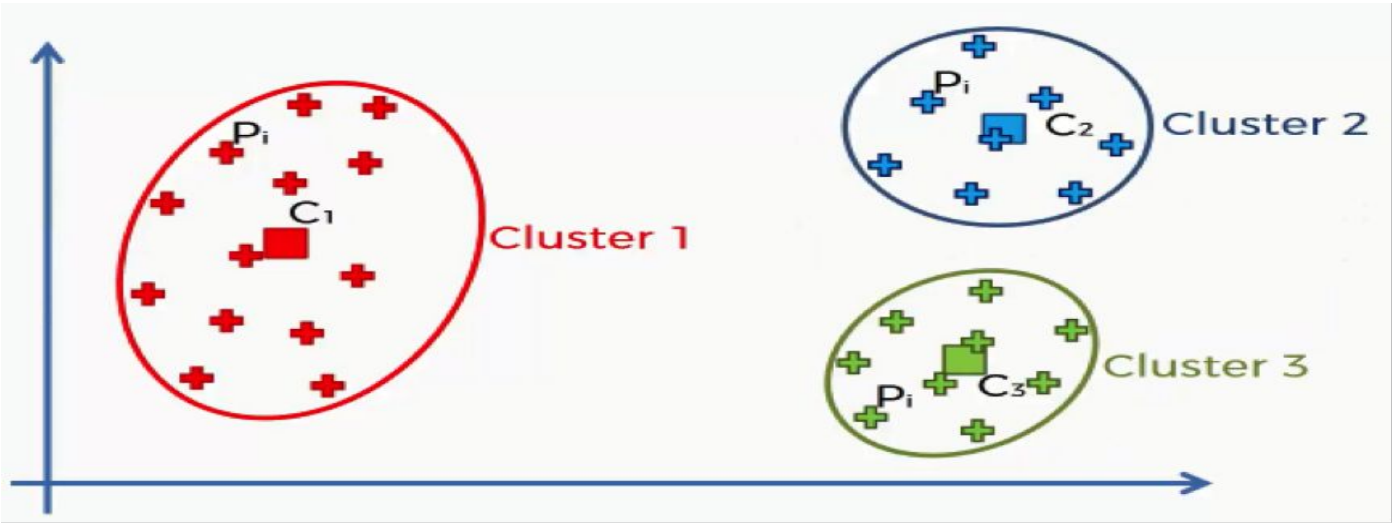
We then square these distances and, finally, calculate the sum of all the squared distances for that cluster. The same is done for all the other clusters. How does this help us in knowing what number of clusters we should use?



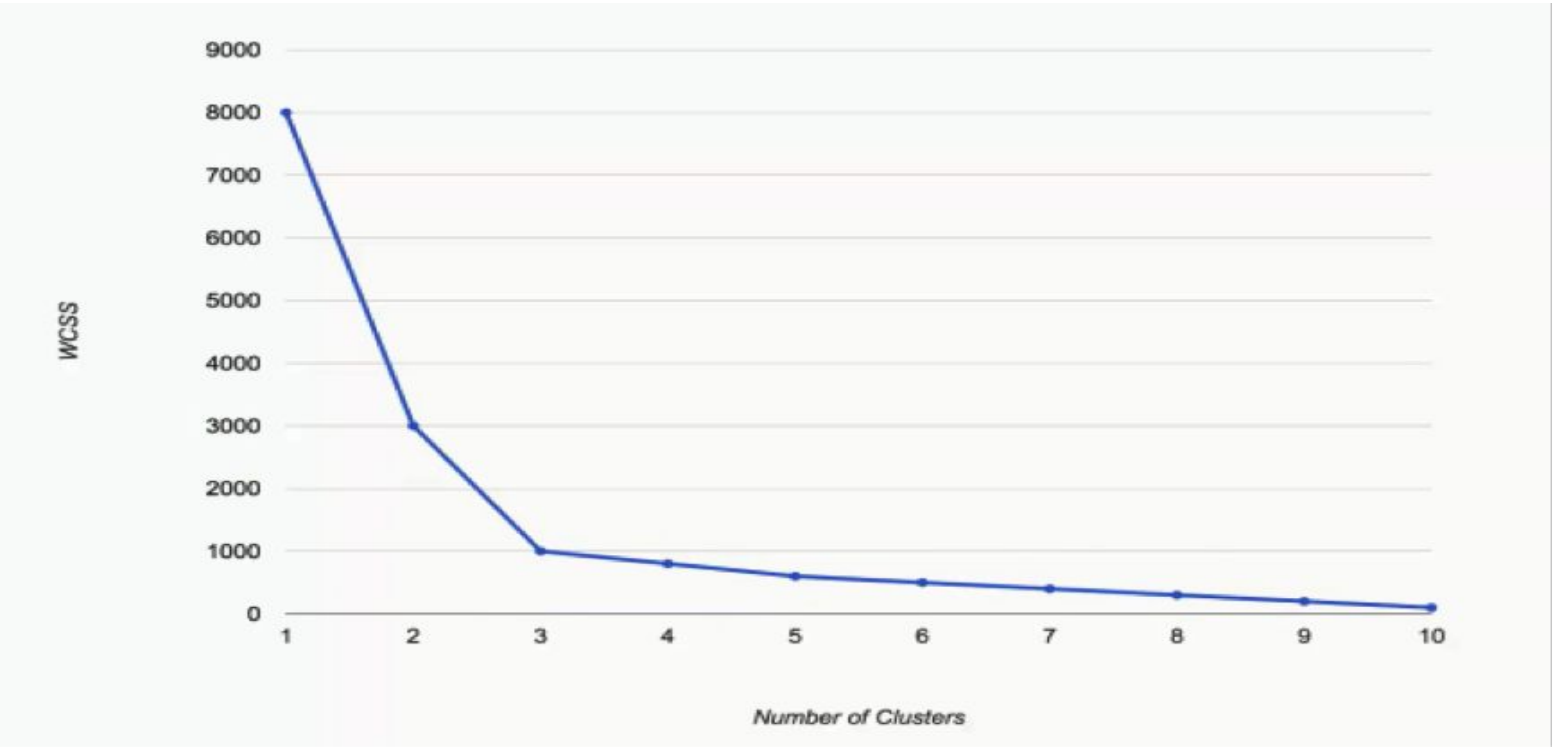
INTERNSHIPSTUDIO

Calculating WCSS

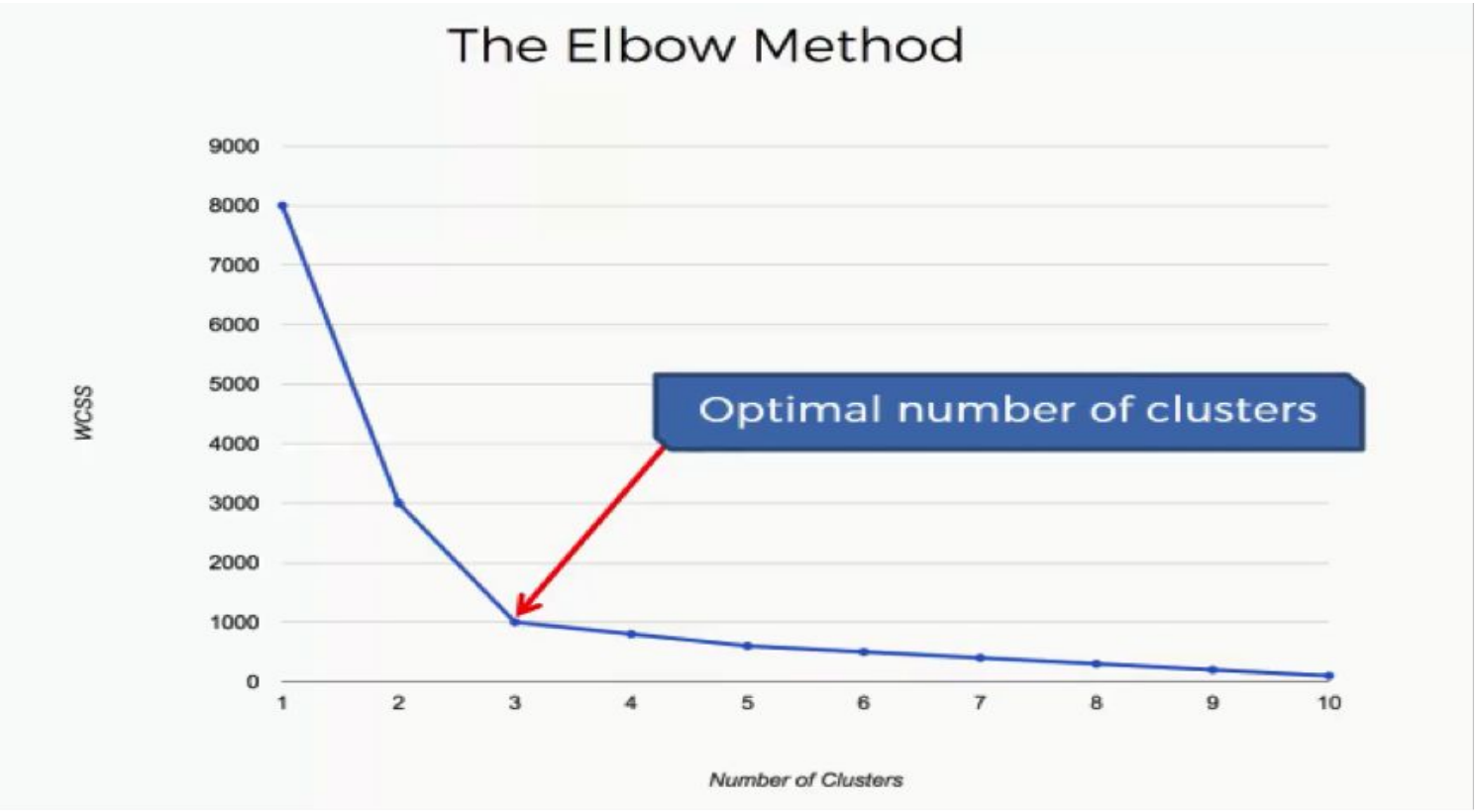




WCSS



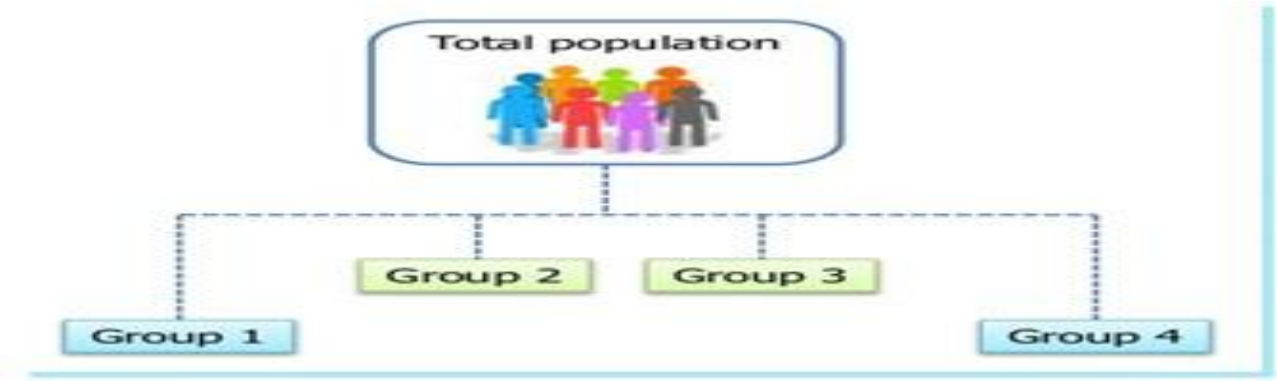
- You can see that as we move from one cluster to two, the WCSS takes a massive fall from 8,000 to 3,000.
- As we move from two to three, the WCSS still decreases substantially, from 3,000 to 1,000. From that point on, however, the changes become very minimal, with each cluster only shaving off 200 WCSS points or less.
- That's our hint when it comes to choosing our optimal number of clusters. The keyword is: Elbow.



The "elbow" in your graph will not always be as obvious as in this example. You're likely to have situations where each person would choose a different point, each thinking that theirs is the optimal point.

That's where you have to make your judgment call as a data scientist. The Elbow Method can only give you a hint at where to look.

Hierarchical Clustering Algorithm

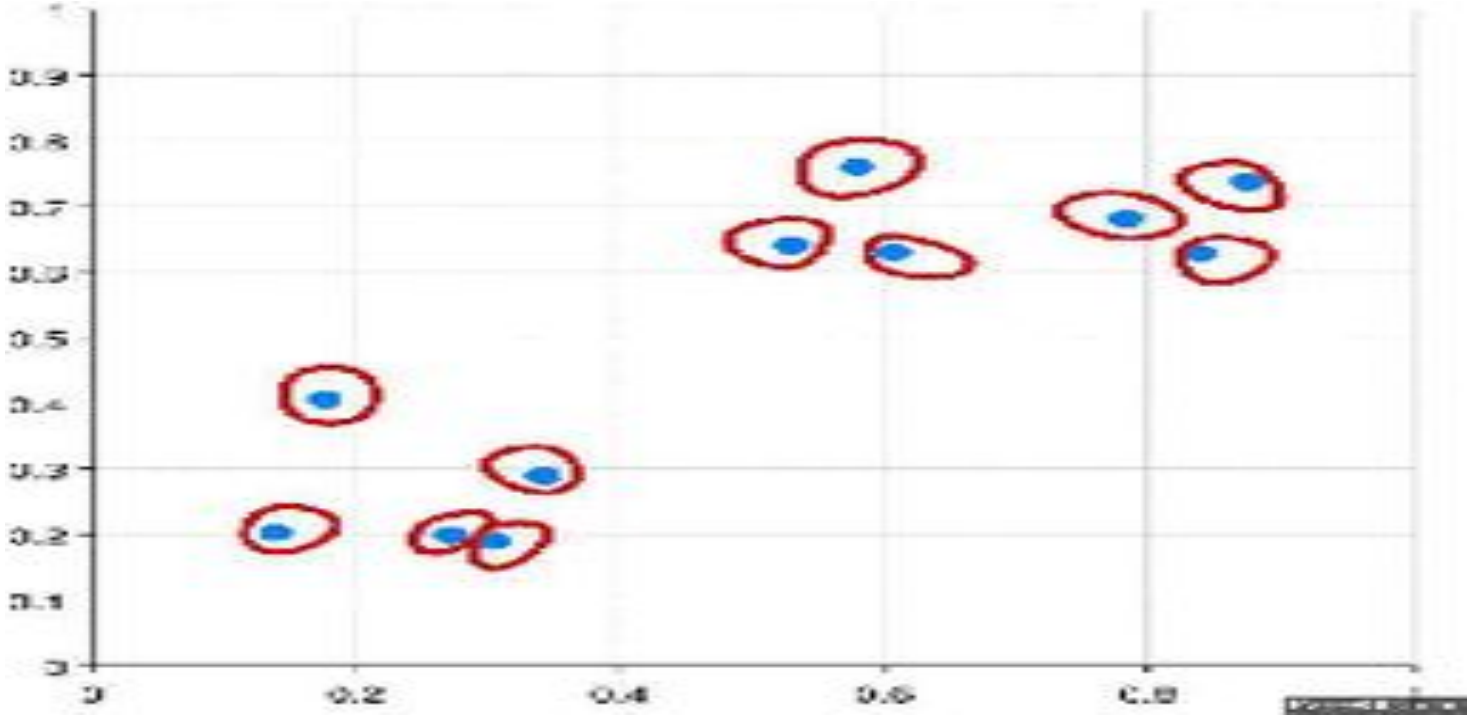


- HCA is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering
- The algorithm groups similar objects into groups called clusters. The endpoint is a set of clusters or groups, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.
- For example, items shown in the image above should be as similar as possible in terms of attributes of the items in each group, and objects in group 1 and group 2 should be as dissimilar as possible.
- Another Example- All files and folders on our hard disk are organized in a hierarchy.

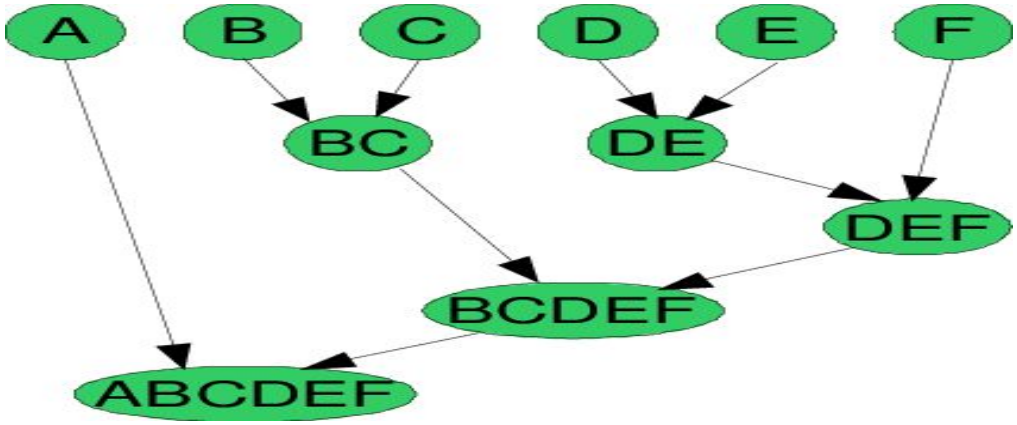
Hierarchical Clustering

Make each data point a single-point cluster → forms N clusters

1. Take the two closest data points and make them one cluster → forms N-1 clusters
2. Take the two closest clusters and make them one cluster → Forms N-2 clusters.
3. Repeat step-3 until you are left with only one cluster.



Hierarchical Clustering Technique



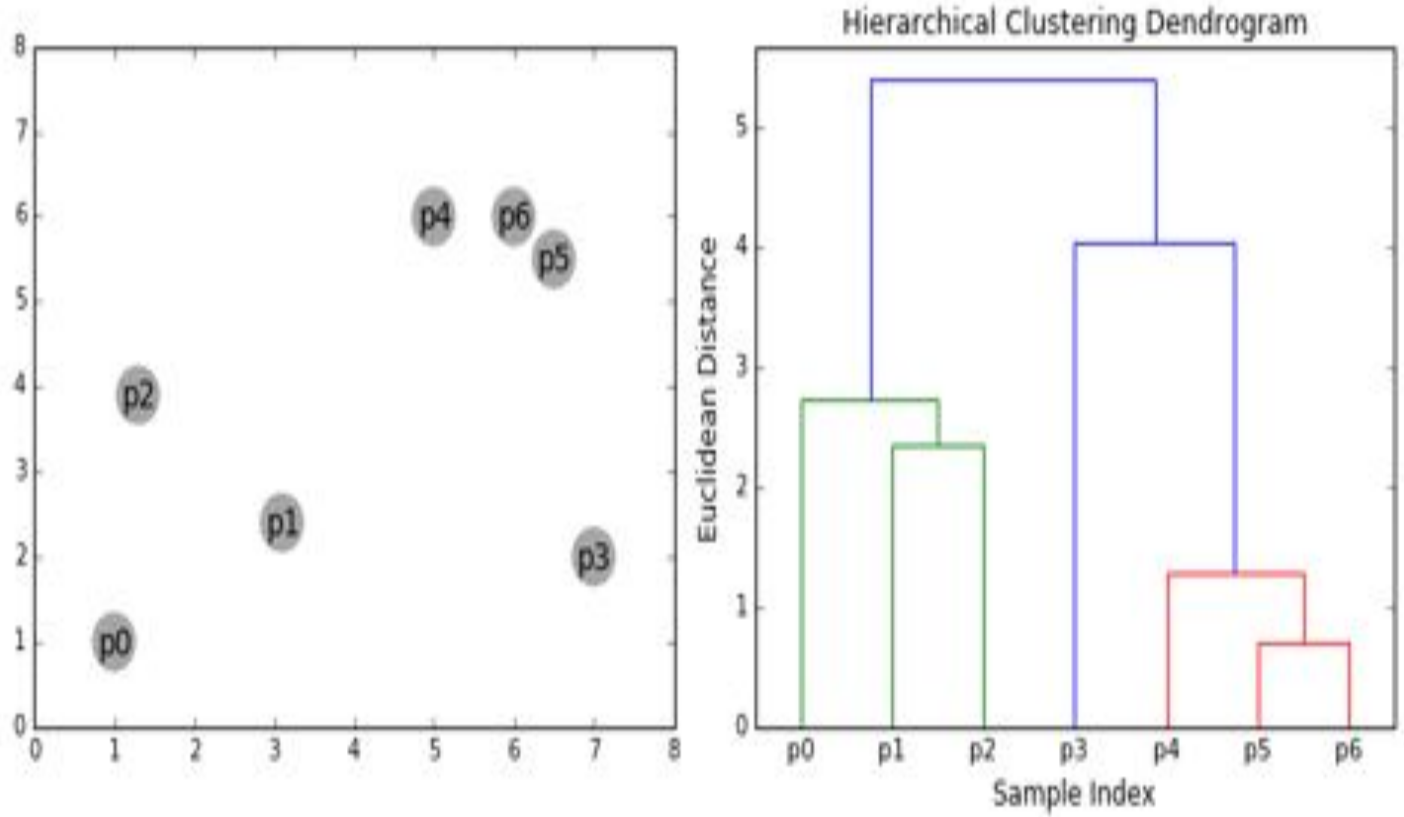
Lets say we have six data points {A,B,C,D,E,F}.

- Step- 1: Calculate the proximity of individual points and consider all the six data points
- Step- 2: similar clusters are merged together and formed as a single cluster. Let's consider B,C, and D,E are similar clusters that are merged in step two. Now, we're left with four clusters which are A, BC, DE, F.
- Step- 3: We again calculate the proximity of new clusters and merge the similar clusters to form new clusters A, BC, DEF.
- Step- 4: Calculate the proximity of the new clusters. The clusters DEF and BC are similar and merged together to form a new cluster. We're now left with two clusters A, BCDEF.
- Step- 5: Finally, all the clusters are merged together

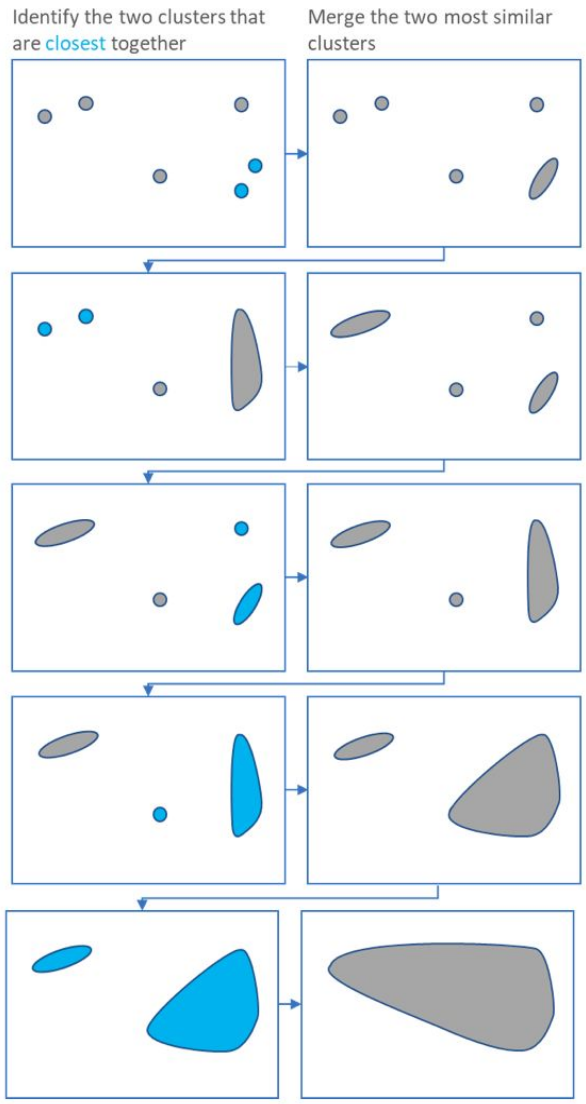
What is a Dendrogram?

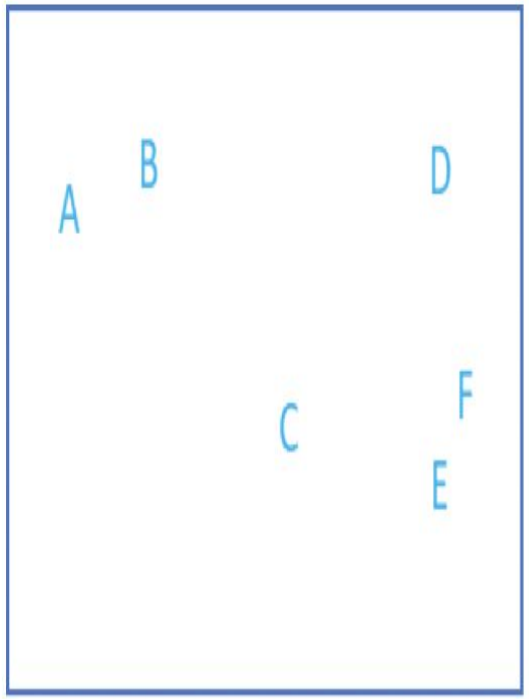
•A Dendrogram is a type of tree diagram showing hierarchical relationships between different sets of data.

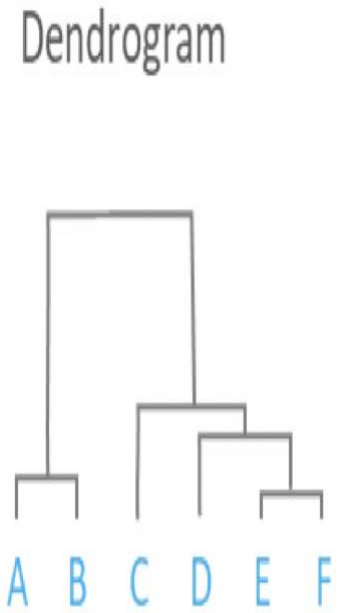
•A Dendrogram contains the memory of hierarchical clustering algorithm, so just by looking at the Dendrogram you can tell how the cluster is formed.



More Example







Industry use Hierarchical Clustering

Business Problem 1: A bank wants to group loan applicants into high/medium/low risk based on attributes such as loan amount, monthly instalments, employment tenure, the number of times the applicant has been delinquent in other payments, annual income, debt to income ratio

- **Business Benefit:** Once the segments are identified, the bank will have a loan applicants' dataset with each applicant labelled as high/medium/low risk.
- Based on these labels, the bank can easily make a decision on
  - Whether to give loan to an applicant
  - how much credit to extend,
  - as well as the interest rate

Business Problem 2: The enterprise wishes to organize customers into groups/segments based on similar traits, product preferences and expectations. Segments are constructed based on customer demographic characteristics, psychographics, past behaviour and product use behaviour.

- **Business Benefit:** Once the segments are identified, marketing messages and products can be customized for each segment.
  - The better the segment(s) chosen for targeting by a particular organization, the more successful the business will be in the market.
  - Hierarchical Clustering can help an enterprise organize data into groups to identify similarities and, equally important, dissimilar groups and characteristics,
  - So that the business can target pricing, products, services, marketing messages and more.

INTERNSHIPSTUDIO

Q.1 Explain Hierarchical Clustering?

Q.2 How the clusters are formed in Hierarchical Clustering?

Q.3 Explain the steps of Hierarchical clustering ?

Q.4 What is a Dendrogram?