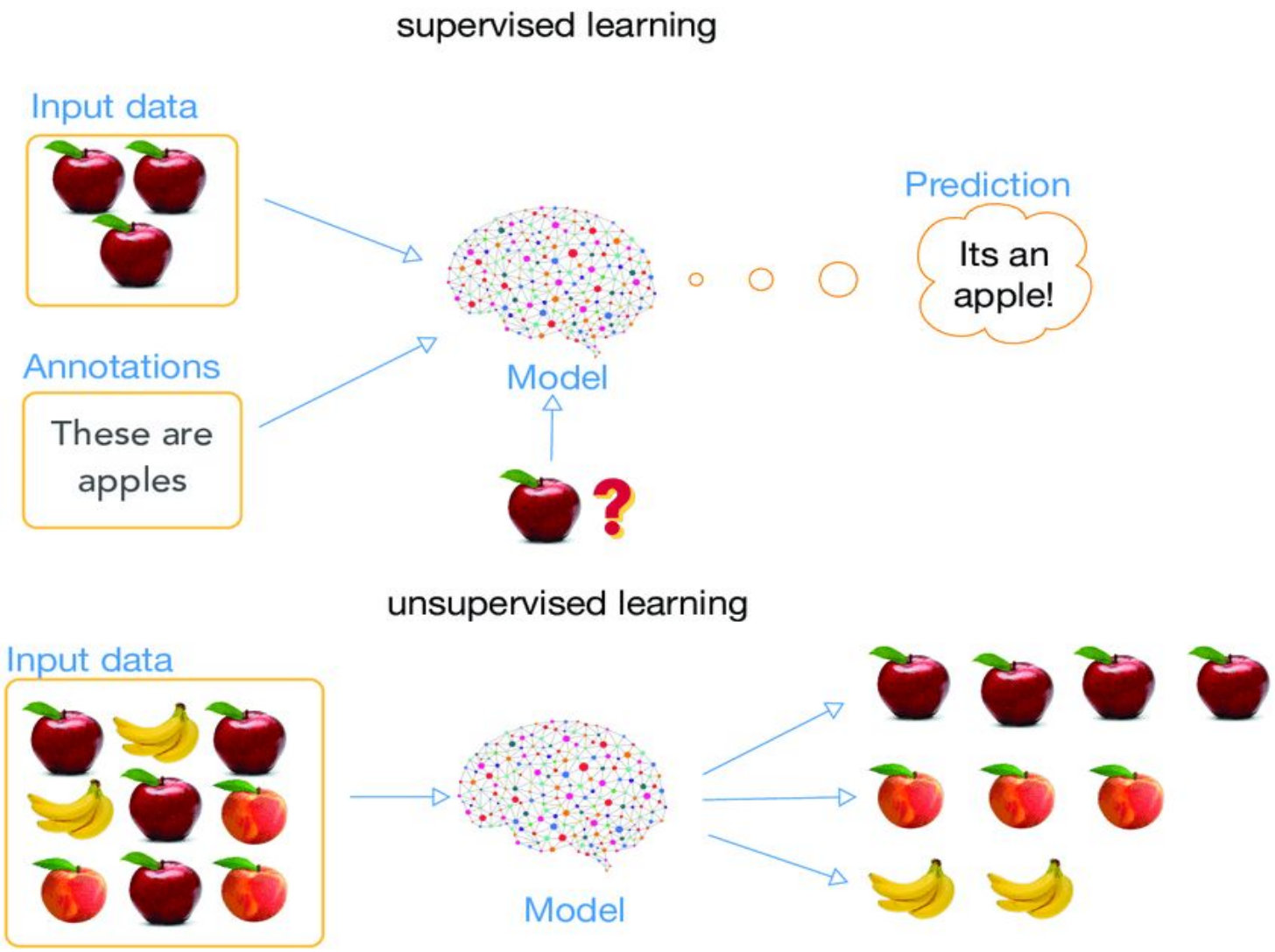
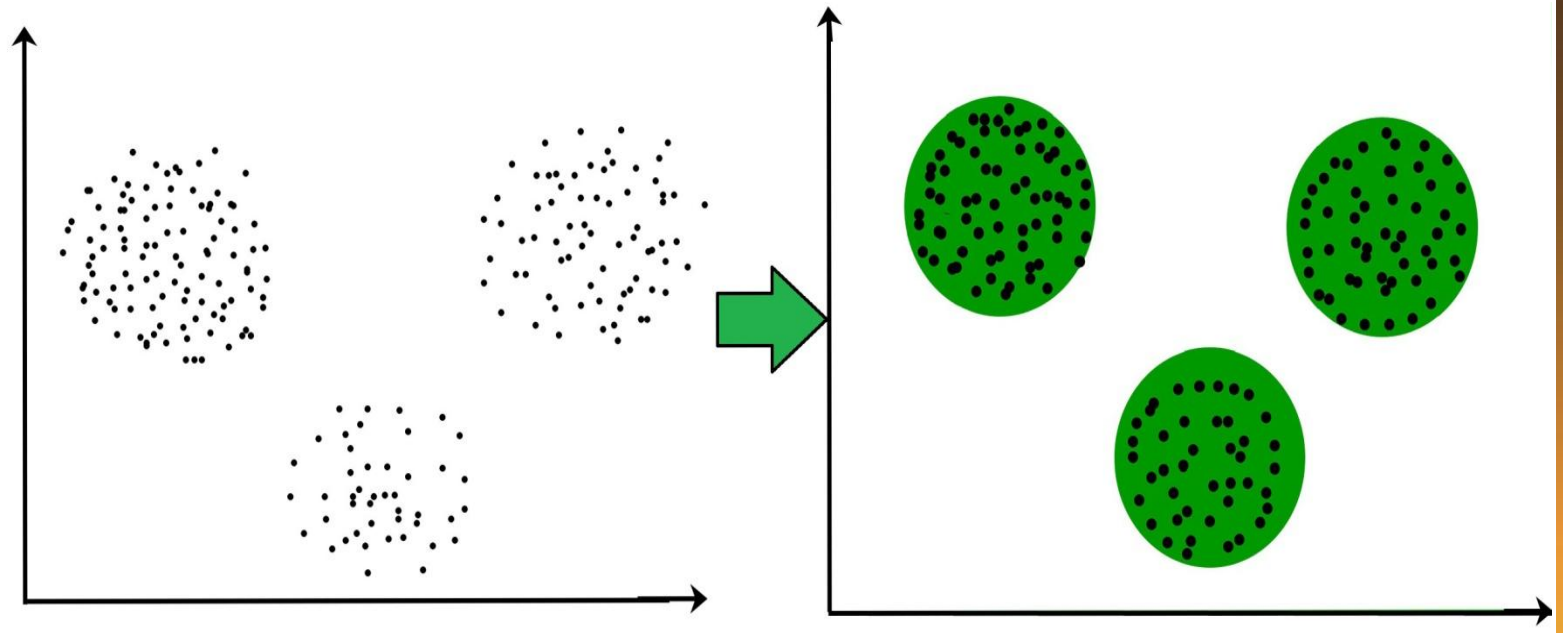
# Examples - Unsupervised learning



# What is Clustering?

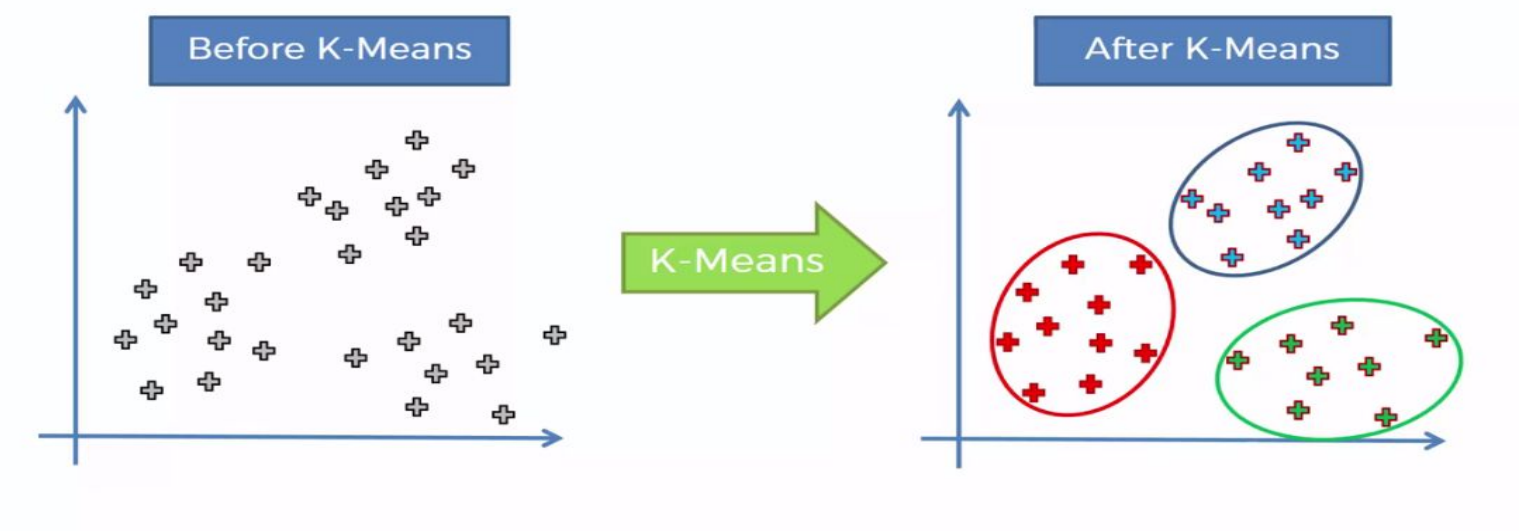
- **Clustering** is the task of dividing the data points into a number of groups such that
  - data points in the same groups are more similar to other data points in the same group
  - It is basically a collection of objects on the basis of similarity and dissimilarity between them.
- **Example-** The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

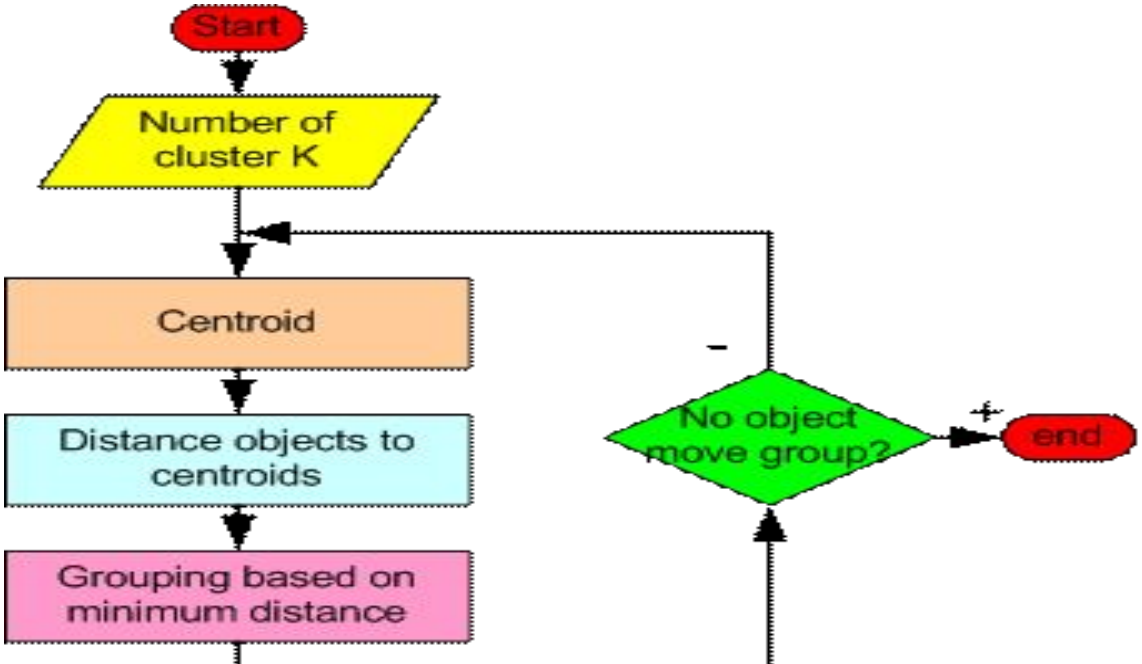
- K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart.

- 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.

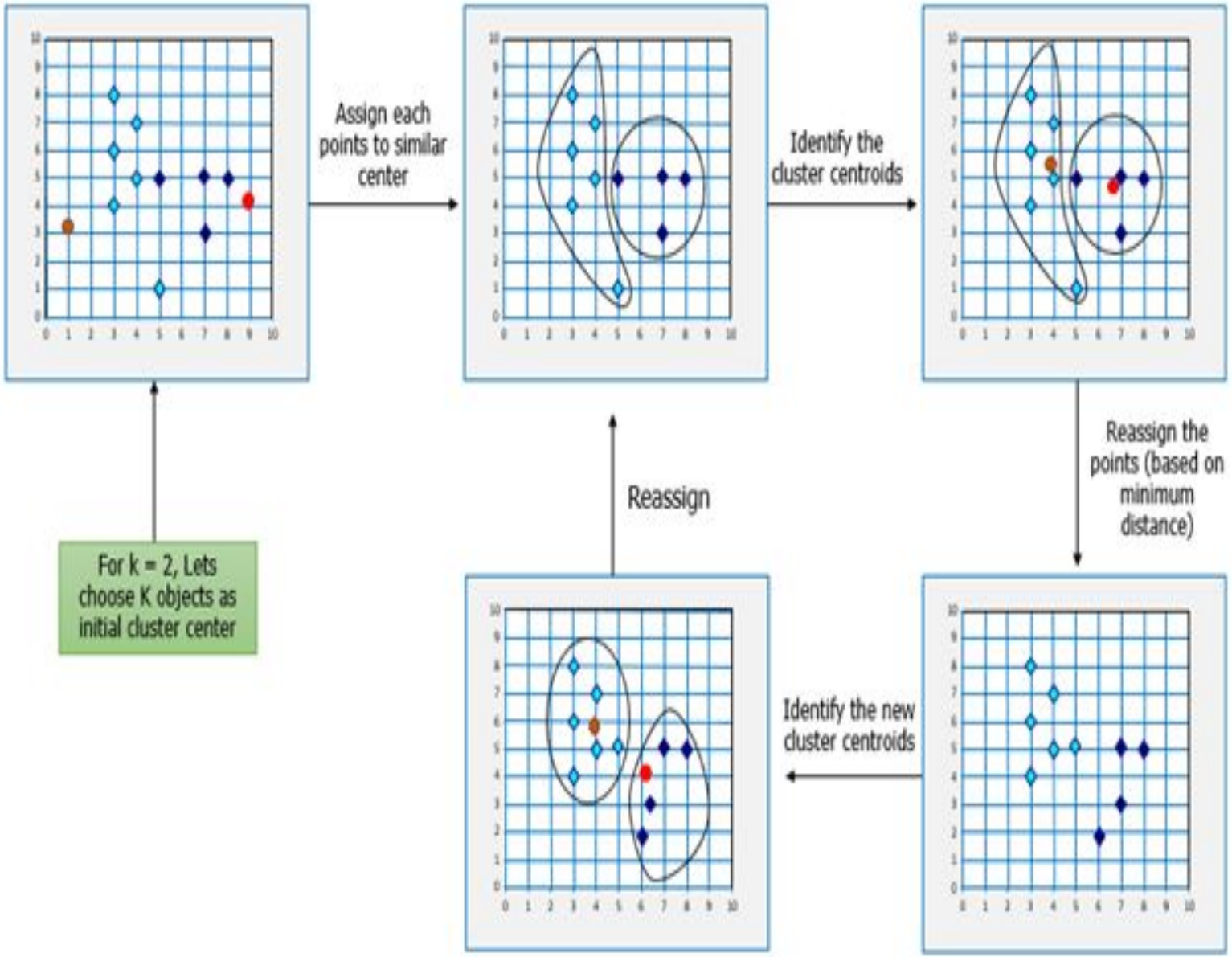
# K-means Working

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.

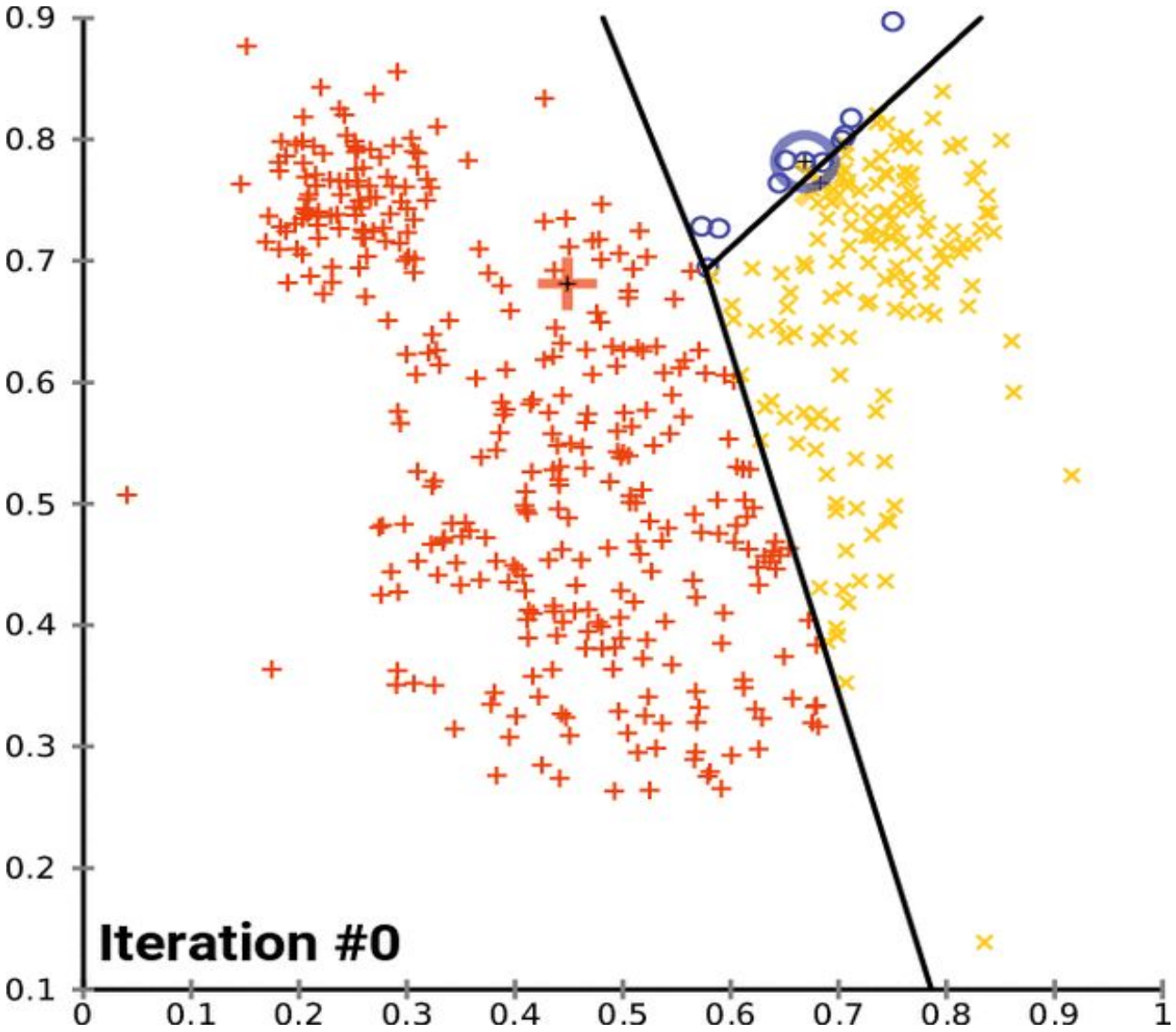


# Step by step process




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

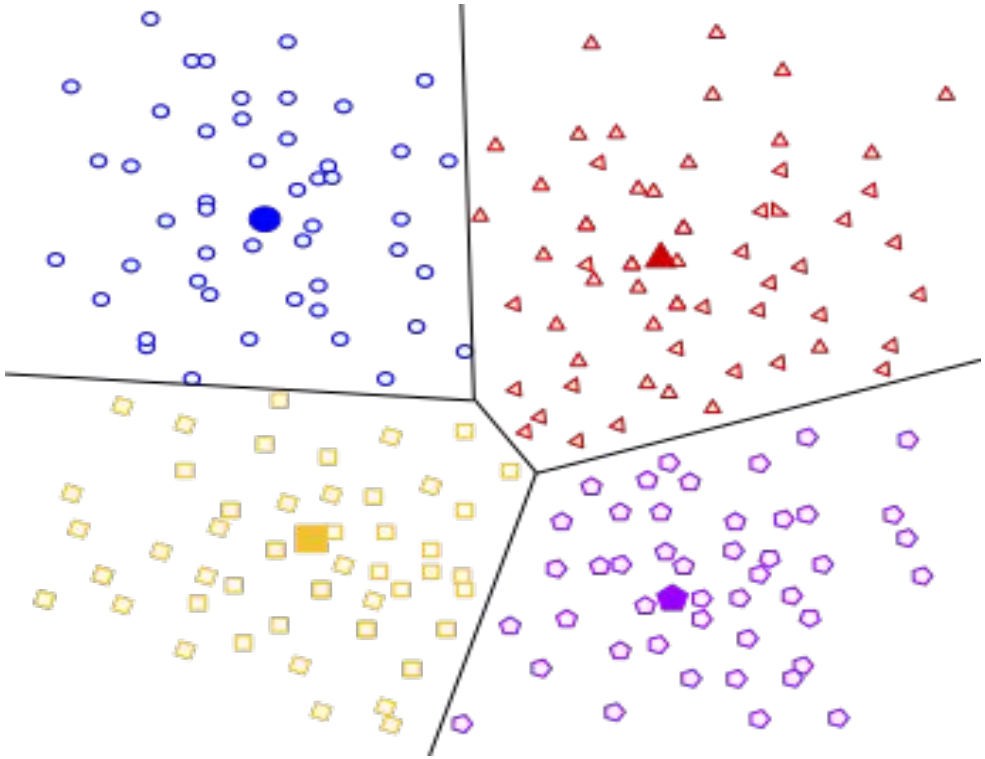
# Applications of Clustering


- **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

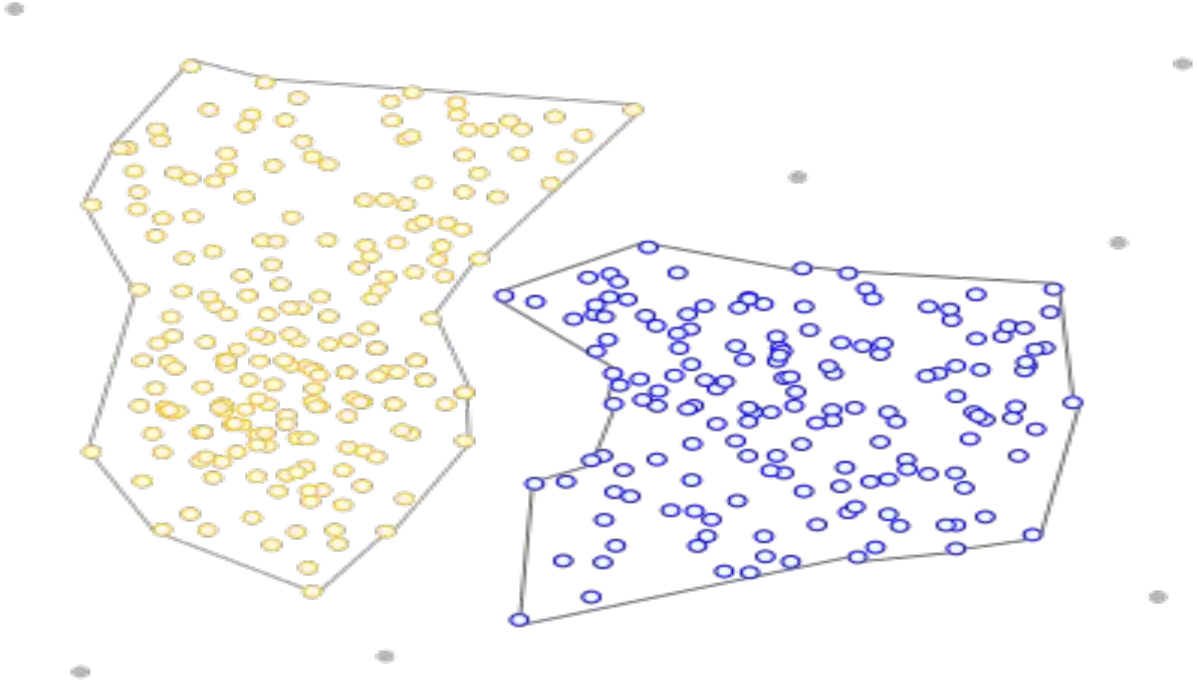
- **Centroid-based clustering** organizes the data into non-hierarchical clusters
- K-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers.



# 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.

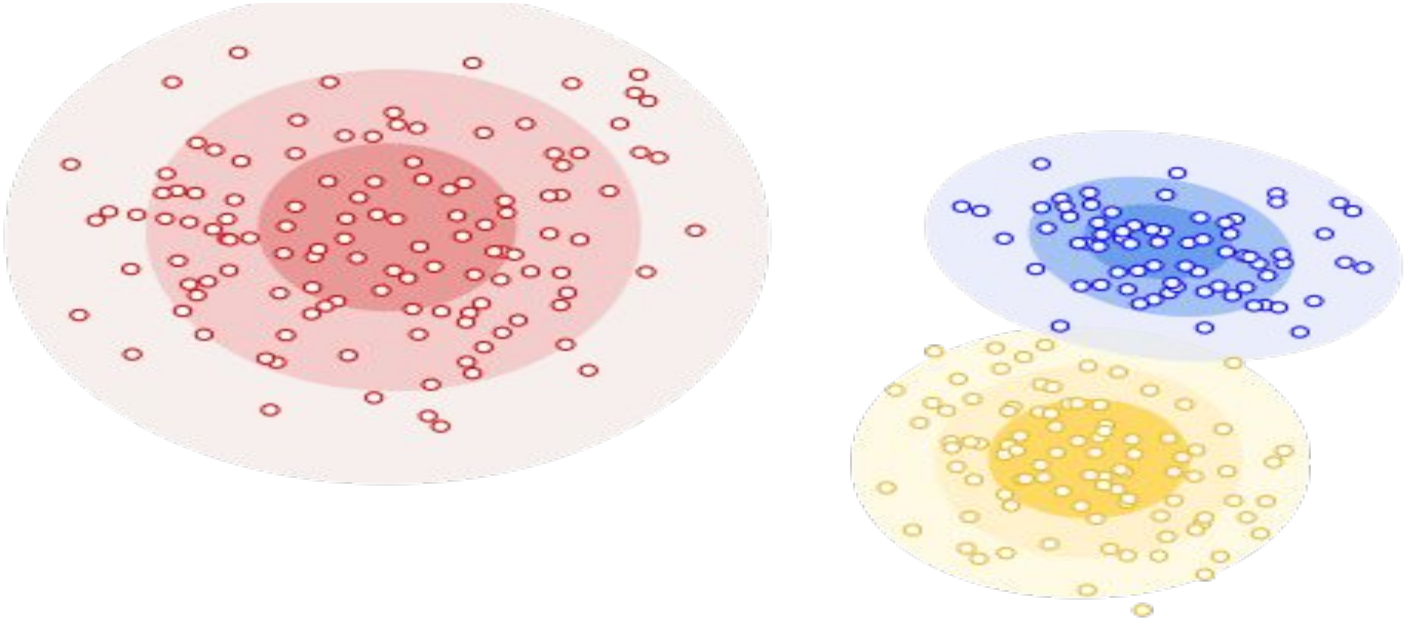


# Types of Clustering

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?