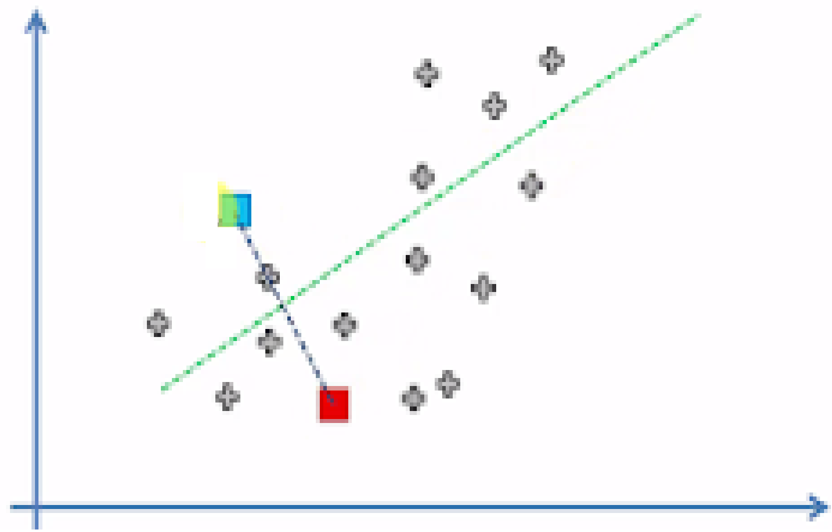
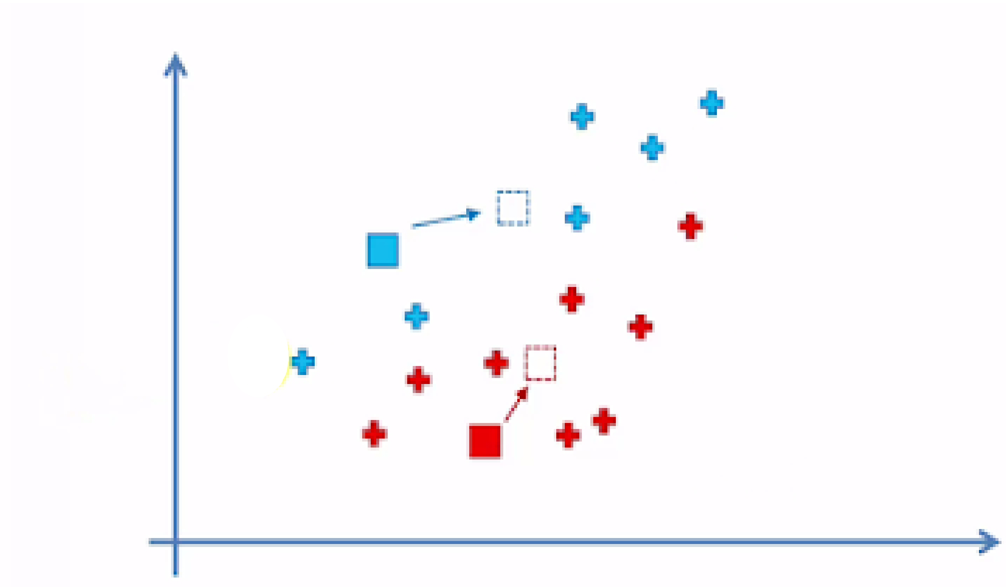
STEP 3

- Assign each data point to the closest centroid ➡ That forms K clusters.



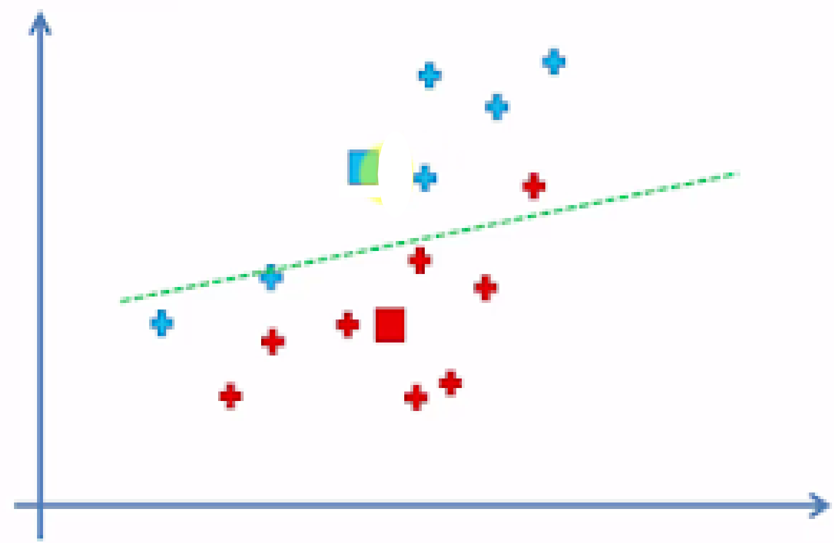
STEP 4

- Compute and place the new centroid of each cluster.

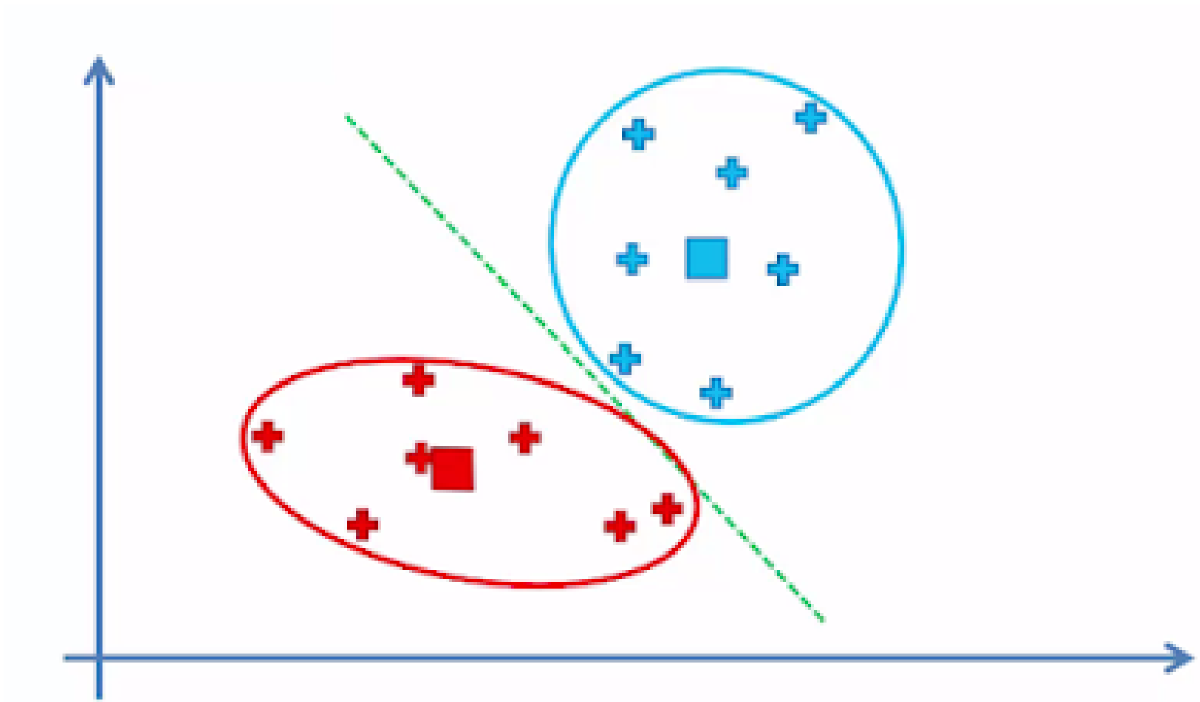


STEP 5

- Reassign each data point to the new closest **centroid**.
- If any **reassignment** took place, go to STEP 4
- Otherwise **FINISH**



Final Model



AGENDA

- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- **Challenges in K Means Clustering**
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

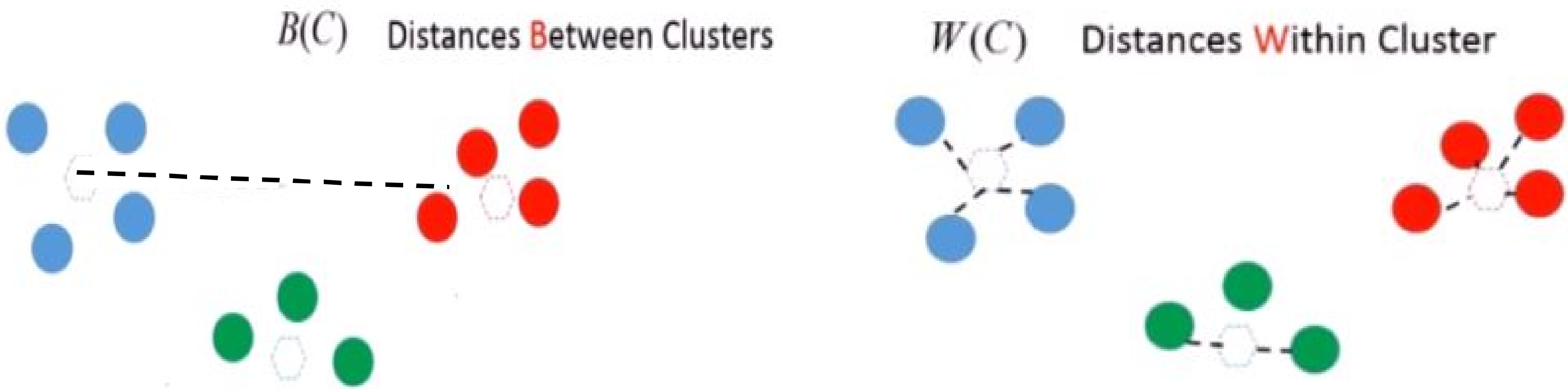
Challenges

How many clusters for this data ?

Two Clusters?

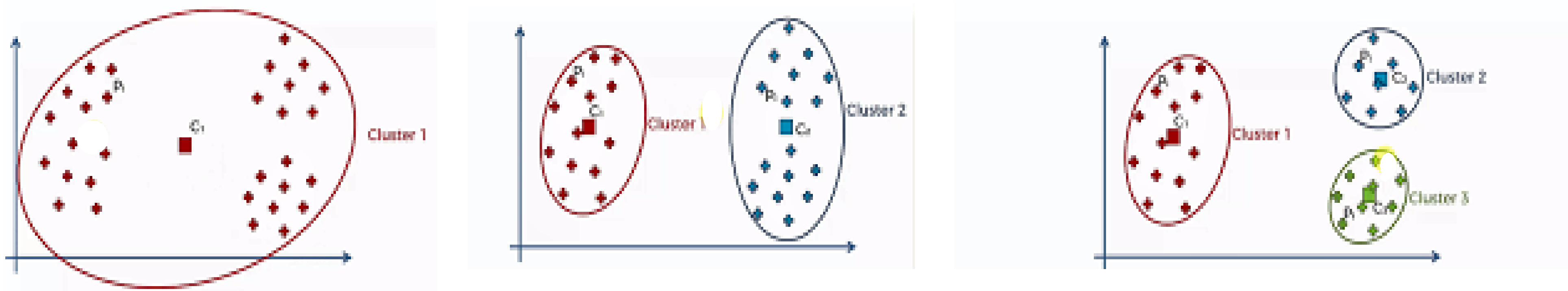
Why not Six Clusters ?

- We use distance between and within the clusters to solve the problem.



The aim of K means is to minimise the distance within the clusters

- The **more number of clusters:** '*within distance*' will be lesser.
- The **less number of clusters:** '*within distance*' will be higher.
- If **number of clusters** are same as data points then within **distance will be Zero**.



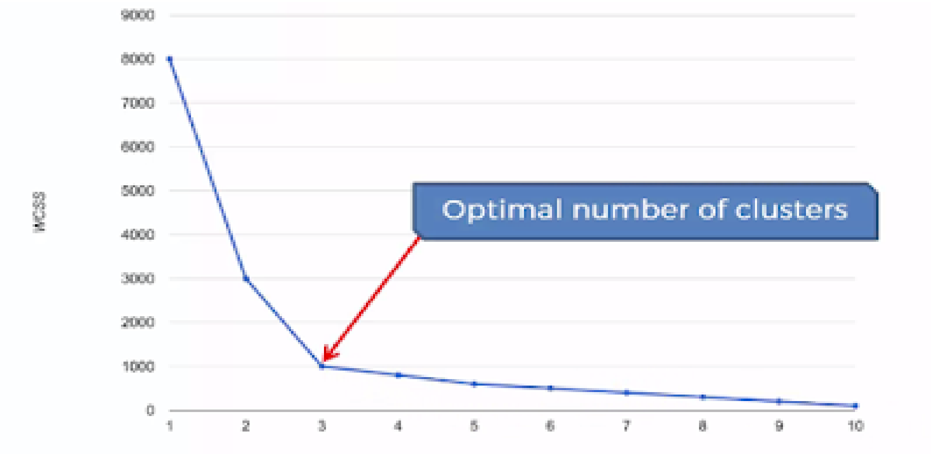
AGENDA

- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- **Elbow Method**
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References
- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

Elbow Method

- Thus, to **choose right number of clusters** use elbow method where you keep increasing number of clusters.
- Wherever you find substantial reduce that will be the **optimum number of clusters** for your problem.



AGENDA

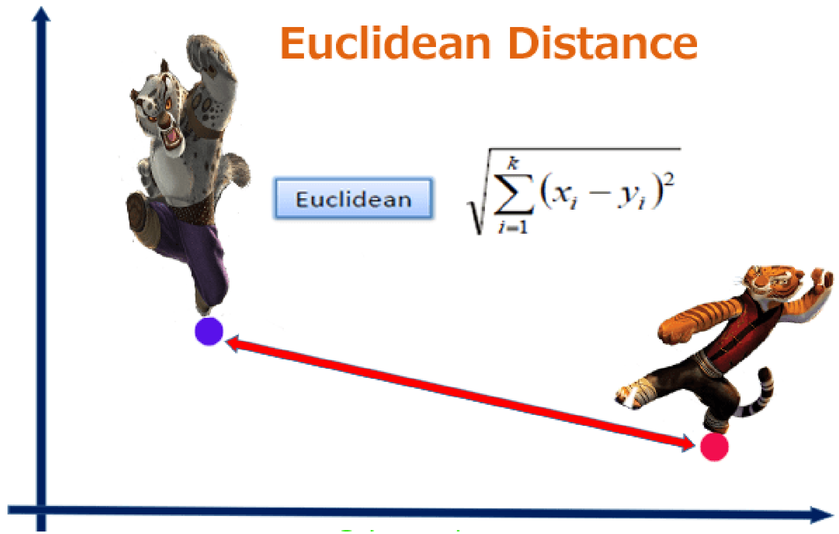
- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
    - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- **Euclidean Distance**
- Illustration of K Means algorithm
- Applications of K Means

References

- Hierarchical Clustering
    - Agglomerative Clustering
    - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
    - Manhattan
    - Minkowski
    - Mahalanobis

Which Distance to use?

- The **Euclidean distance** is the ordinary straight line distance.
- It is the distance between two points in **Euclidean space**.



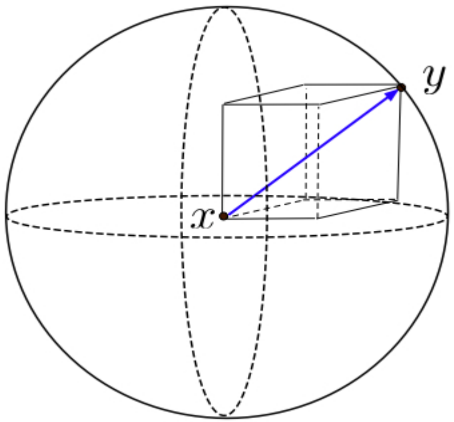
But, why should we use Euclidean distance?

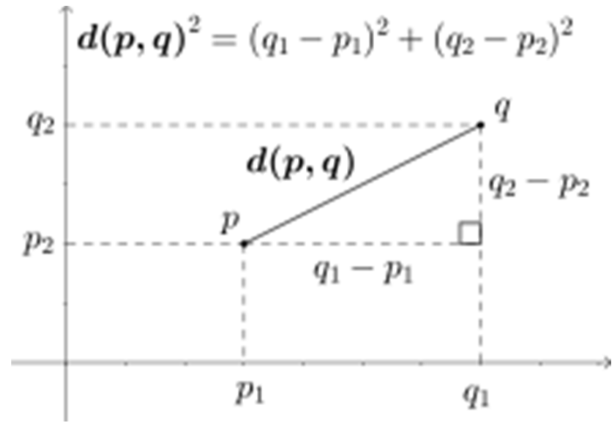
Because…

- K-Means is implicitly based on **pairwise Euclidean distances** b/w data points.
- **Sum of squared deviations from centroid = $\frac{\text{Sum of pairwise squared Euclidean distances}}{\text{Number of points}}$**

$$\Sigma\,(x_i - \bar{x})^2 = \Sigma(x_i^2) - (\Sigma\, x_i)^2 / n$$

- The term "centroid" is itself from **Euclidean geometry**.
- It is **multivariate mean in Euclidean space**.
- Euclidean space is **about Euclidean distances**.





Lets consider a dataset

Sample no.	X	Y
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

STEP 1

- K = **initial cluster size**
- Given, K = 3
- Let’s divide our dataset into 3 clusters and find the **Euclidean Distance.**

- Let **(2, 10), (5, 8)** & **(1, 2)** be our three centroids.