K-Means

K-Means

OPTIMIZATION

Mobile service providers need to setup their network – where to set up the towers so that all its users receive the maximum signal strength

Stringent law enforcement by the government – station the patrol vans- area of high crime rates are in the vicinity of patrol vans

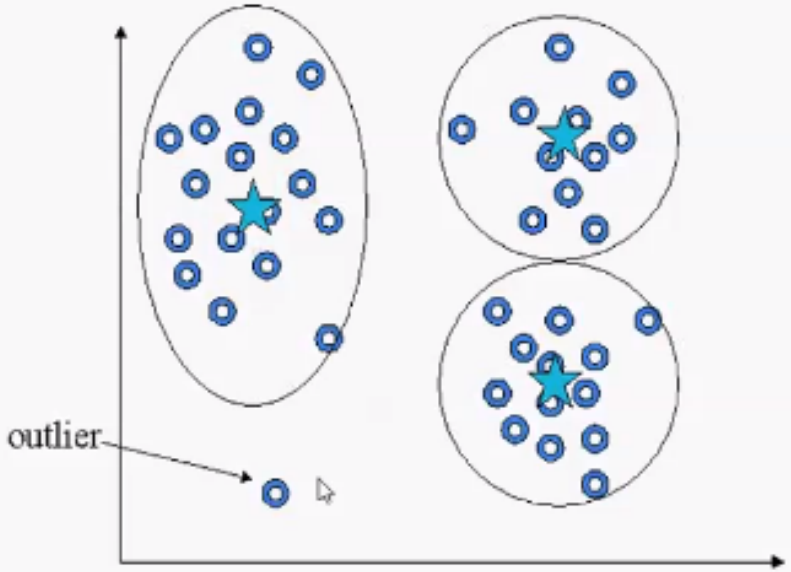
A hospital care chain wants to open series of emergency care wards- max accident prone areas

IDENTIFYING ANOMALY- OUTLIERS

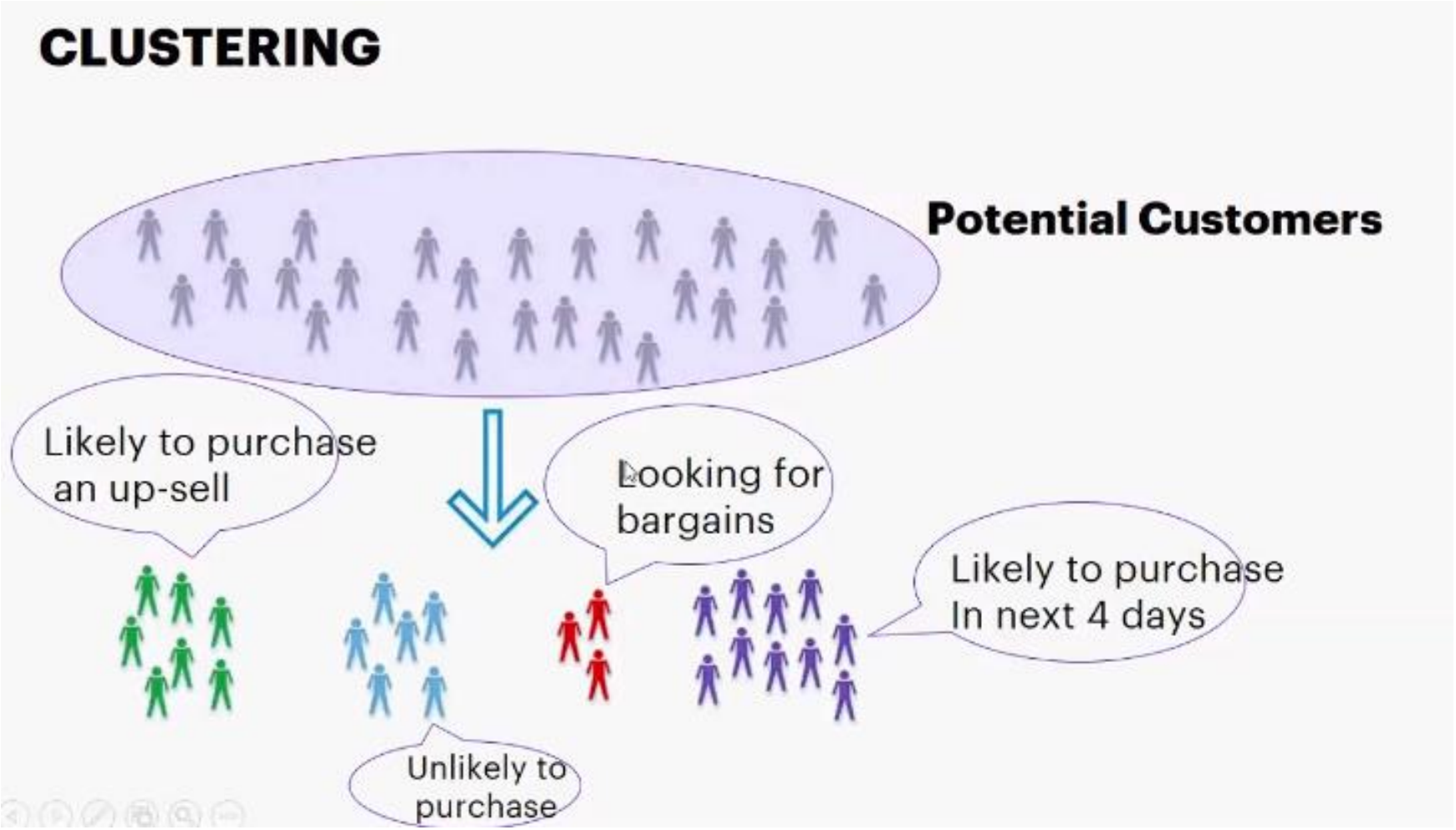
Fraud detection- Communication service providers

Fraud transaction- Banking/Finance

Medicine – abnormal cells – diagnosis



K-Means

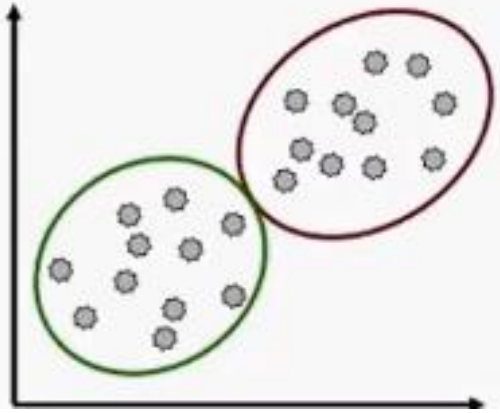

CLUSTERING
Potential Customers
Likely to purchase
an up-sell
Looking for
bargains
Likely to purchase
In next 4 days
Unlikely to
purchase

K-Means

DIFFERENCE BETWEEN CLUSTERING & CLASSIFICATION

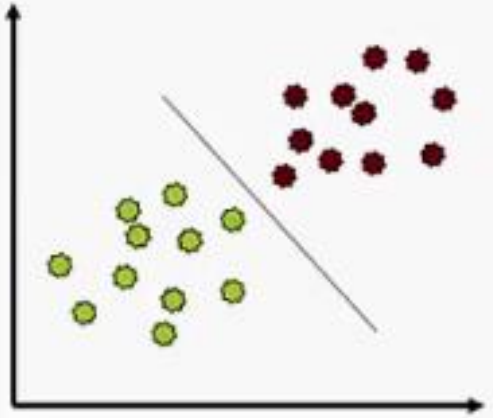
CLUSTERING

- Data is not labeled
- Group points that are "close" to each other
- Identify structure or patterns in data
- Unsupervised learning

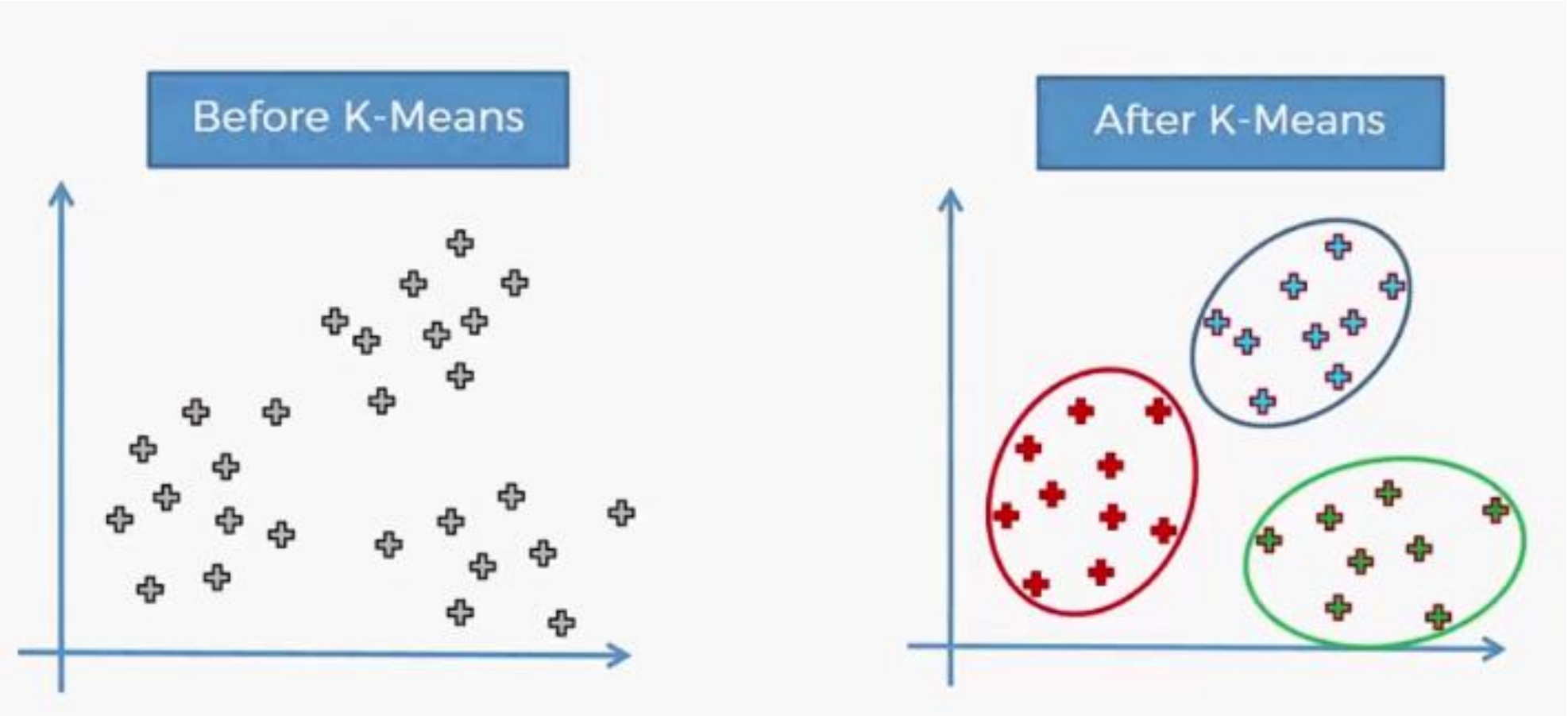


CLASSIFICATION

- Labeled data points
- Want a "rule" that assigns labels to new points
- Supervised learning



K-Means



K-Means

STEPS

Choose the number K of clusters
Select at random K points, the centroids
Assign each data point to the closest centroid
Compute and place the new centroid of each cluster
Reassign each data point to the new closest centroid
  if any reassignment took place, go to previous step
  else finish

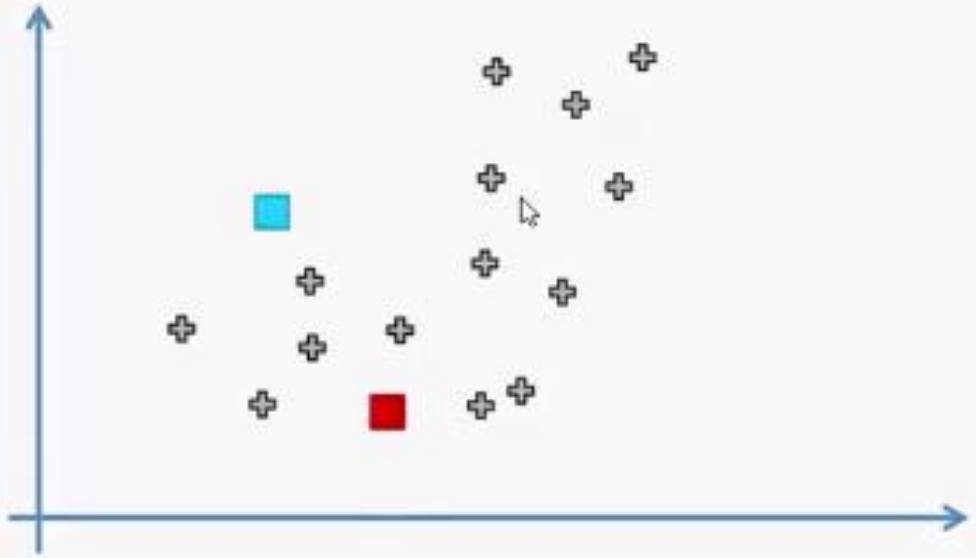
K-Means

K MEANS ALGORITHM

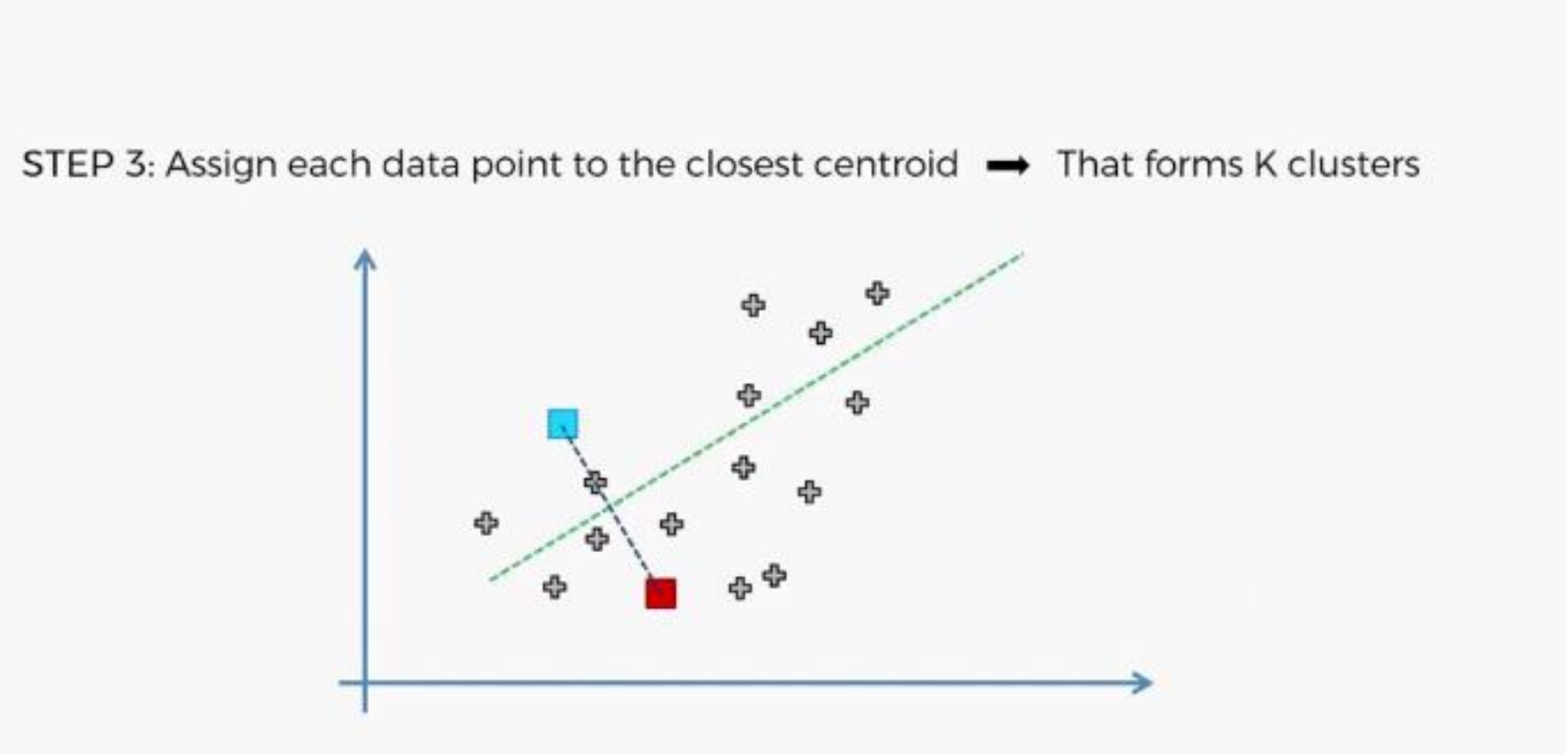
STEP 1: Choose the number K of clusters: K = 2

K-Means

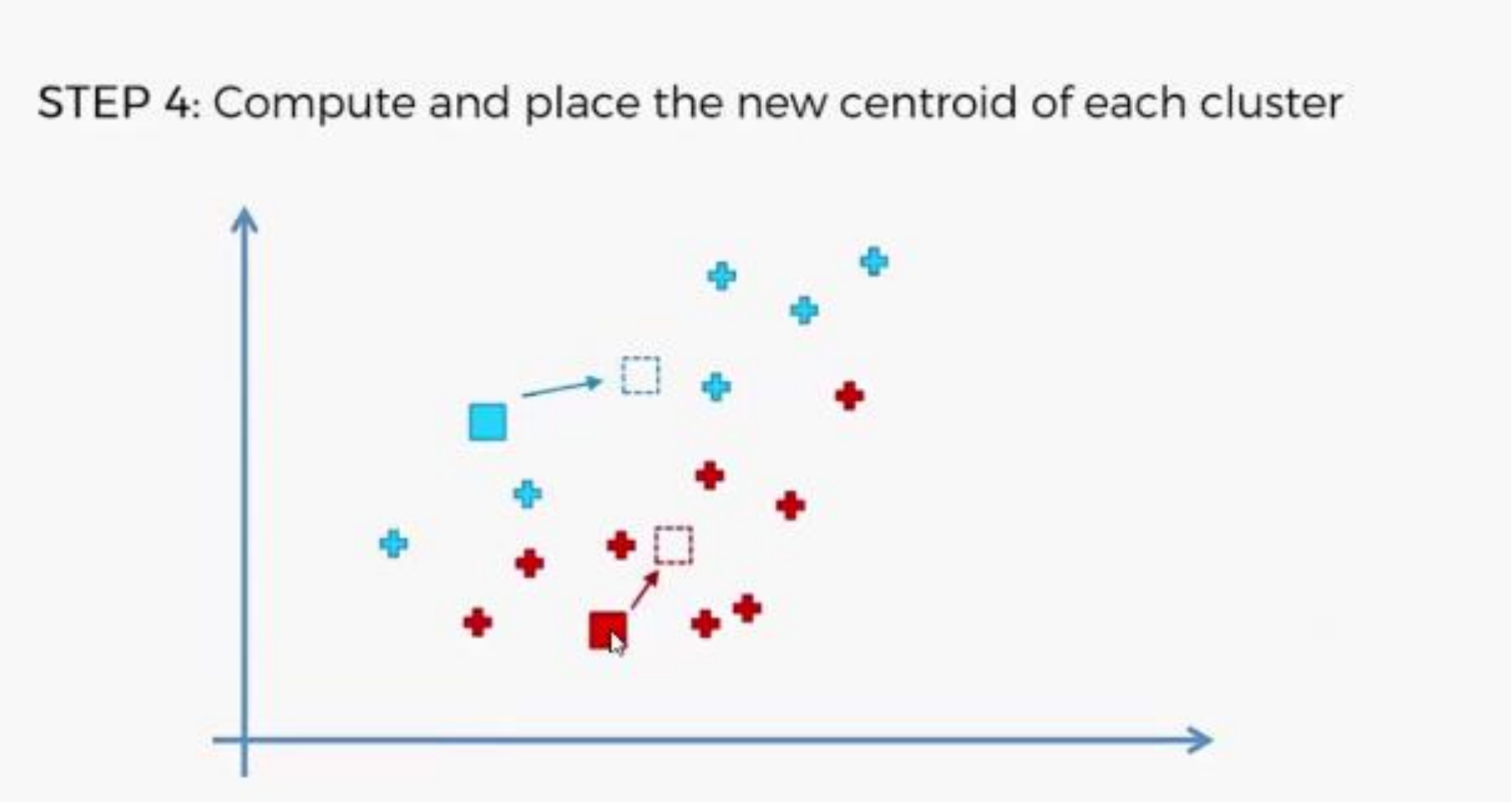
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)



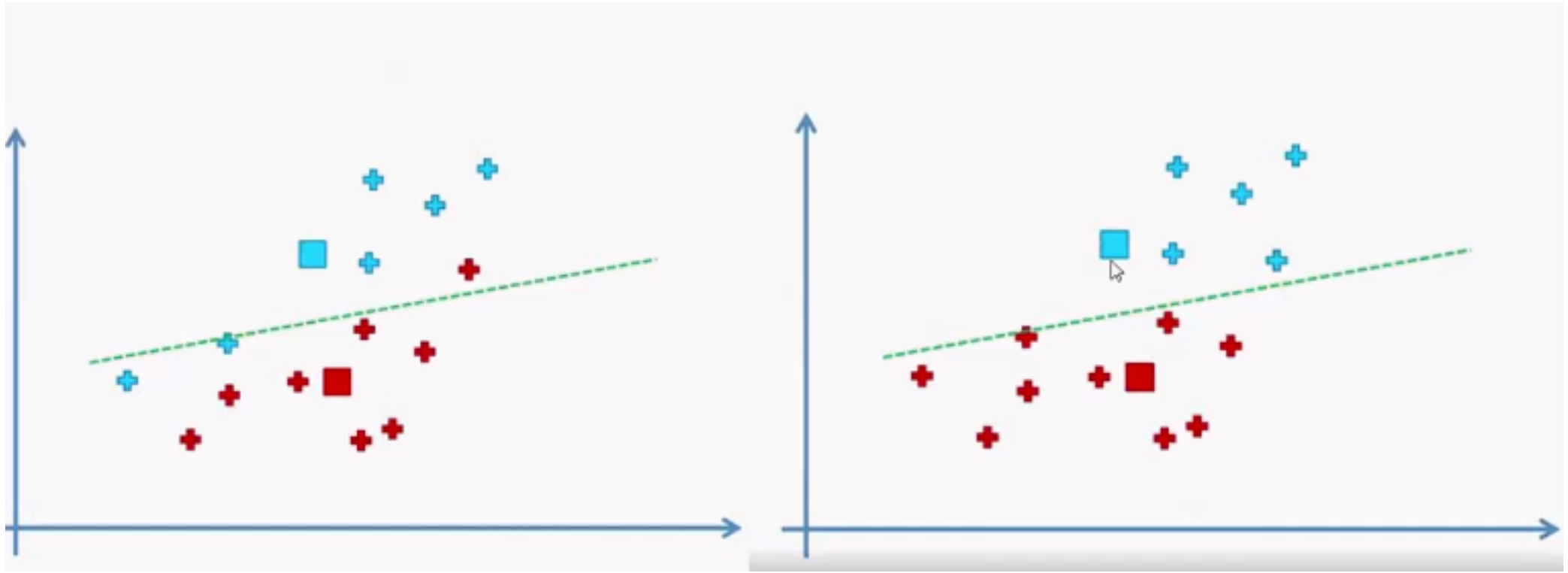
K-Means


STEP 3: Assign each data point to the closest centroid ➡ That forms K clusters

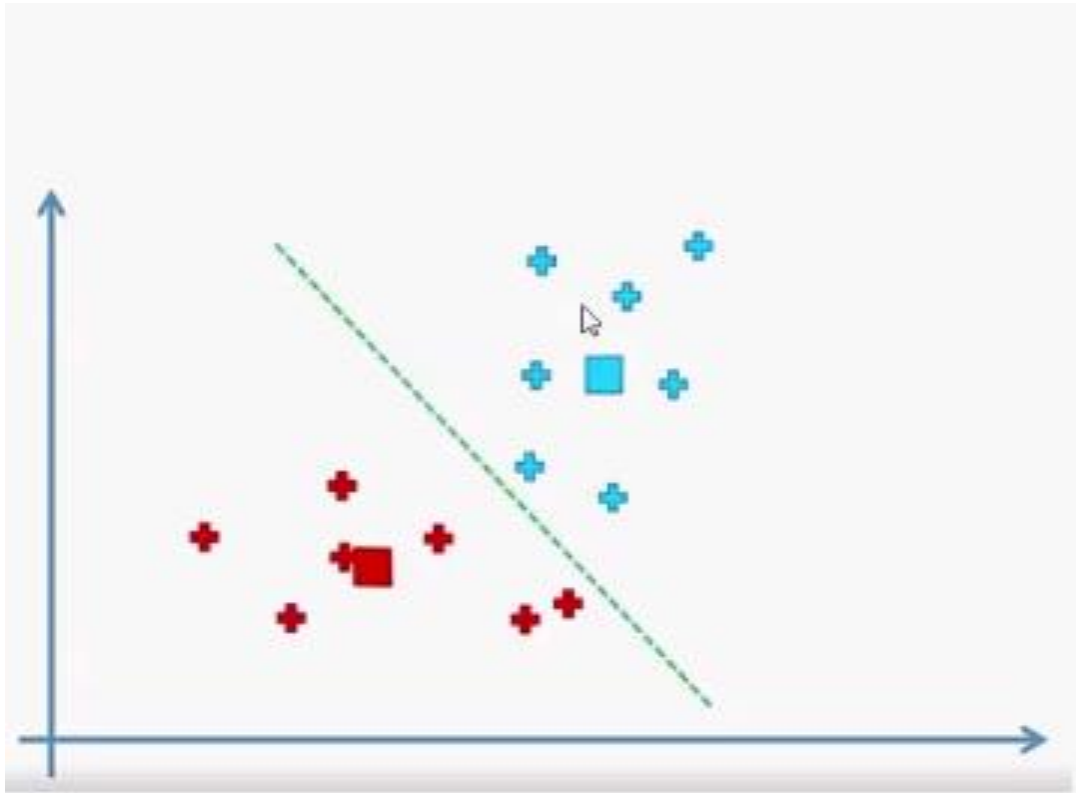
K-Means

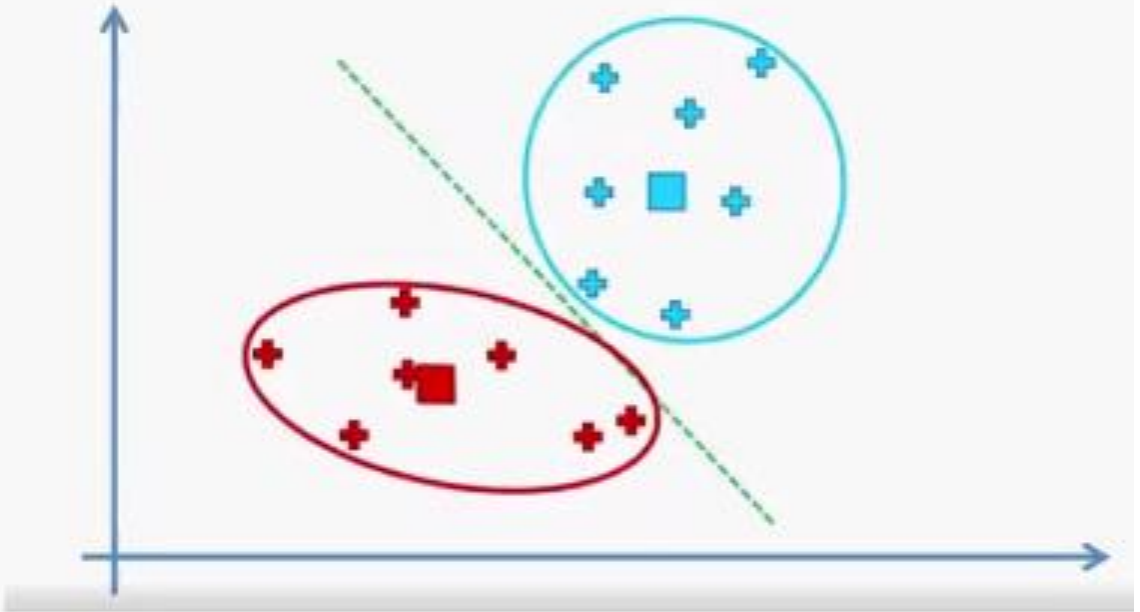


K-Means

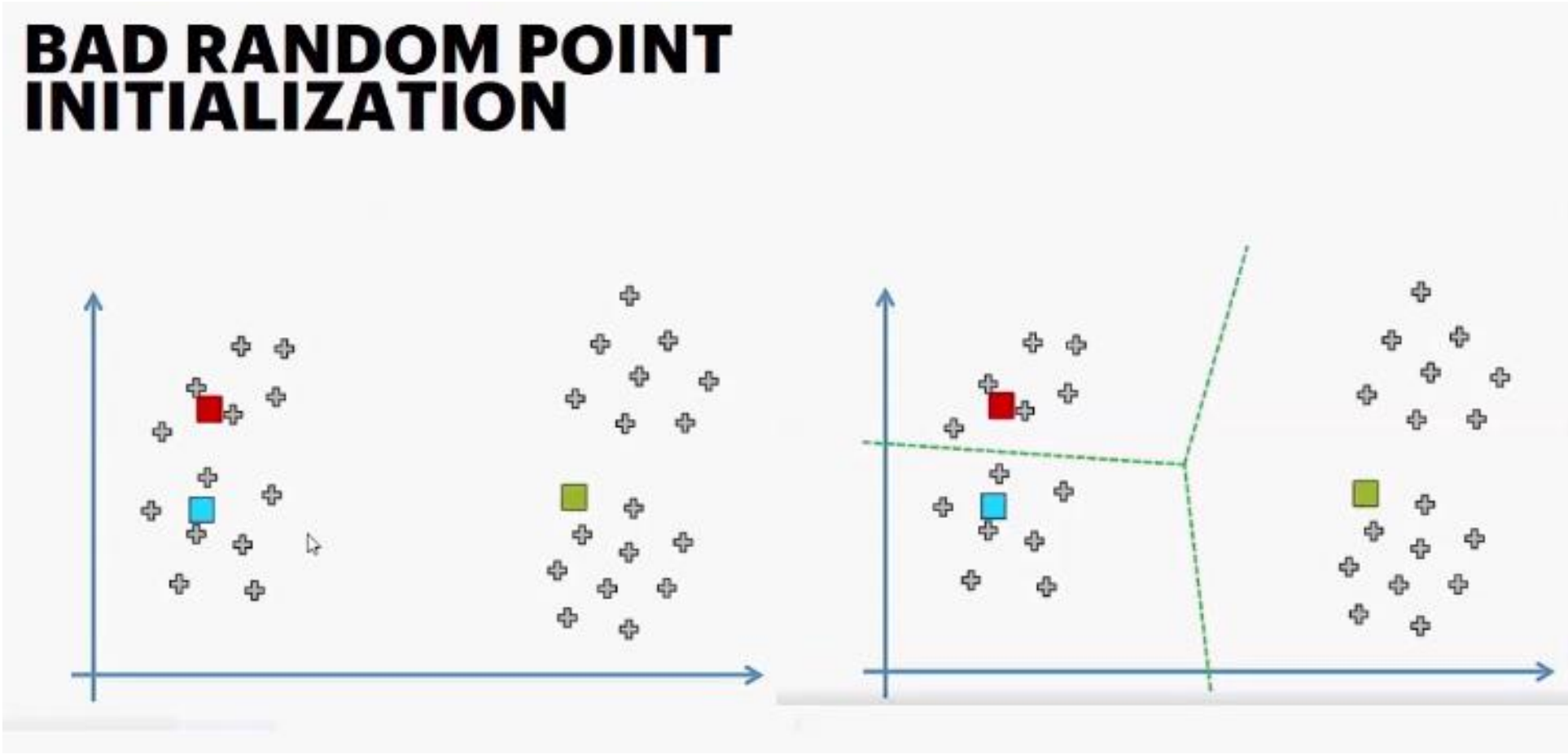


K-Means



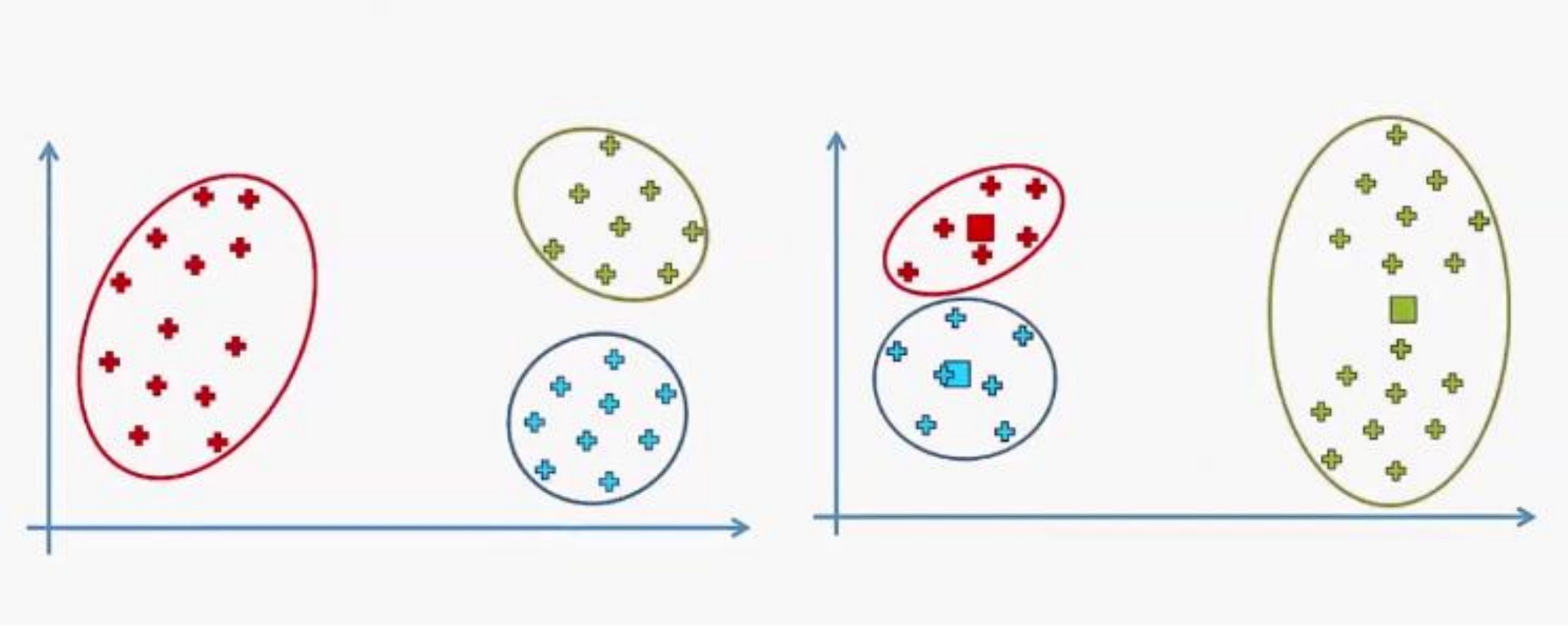


K-Means

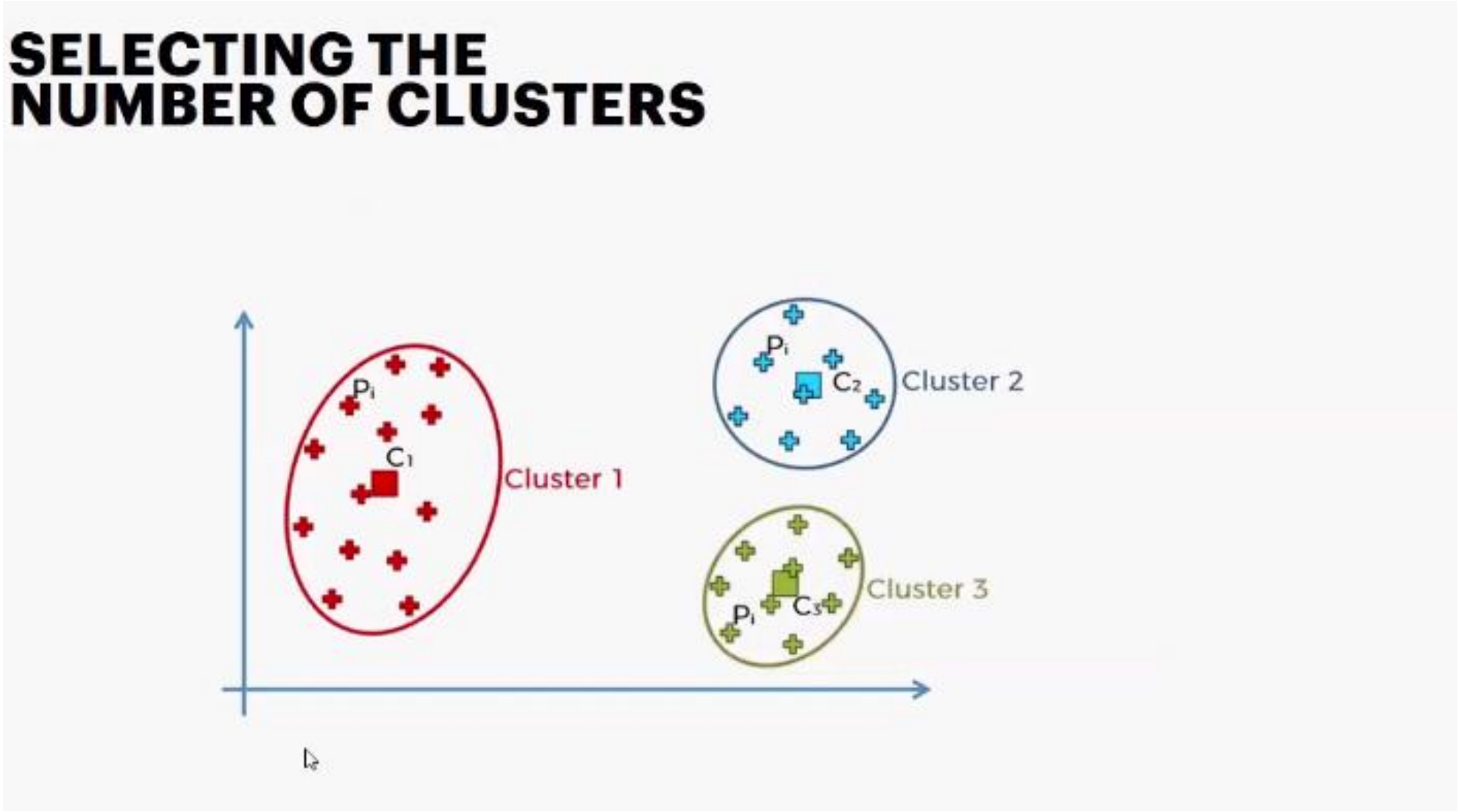


K-Means ++

Using K-means++ we can achieve optimum initial centroids as shown in Fig 1. Where as Fig 2 shows the incorrect initial centroids assignment.



K-Means

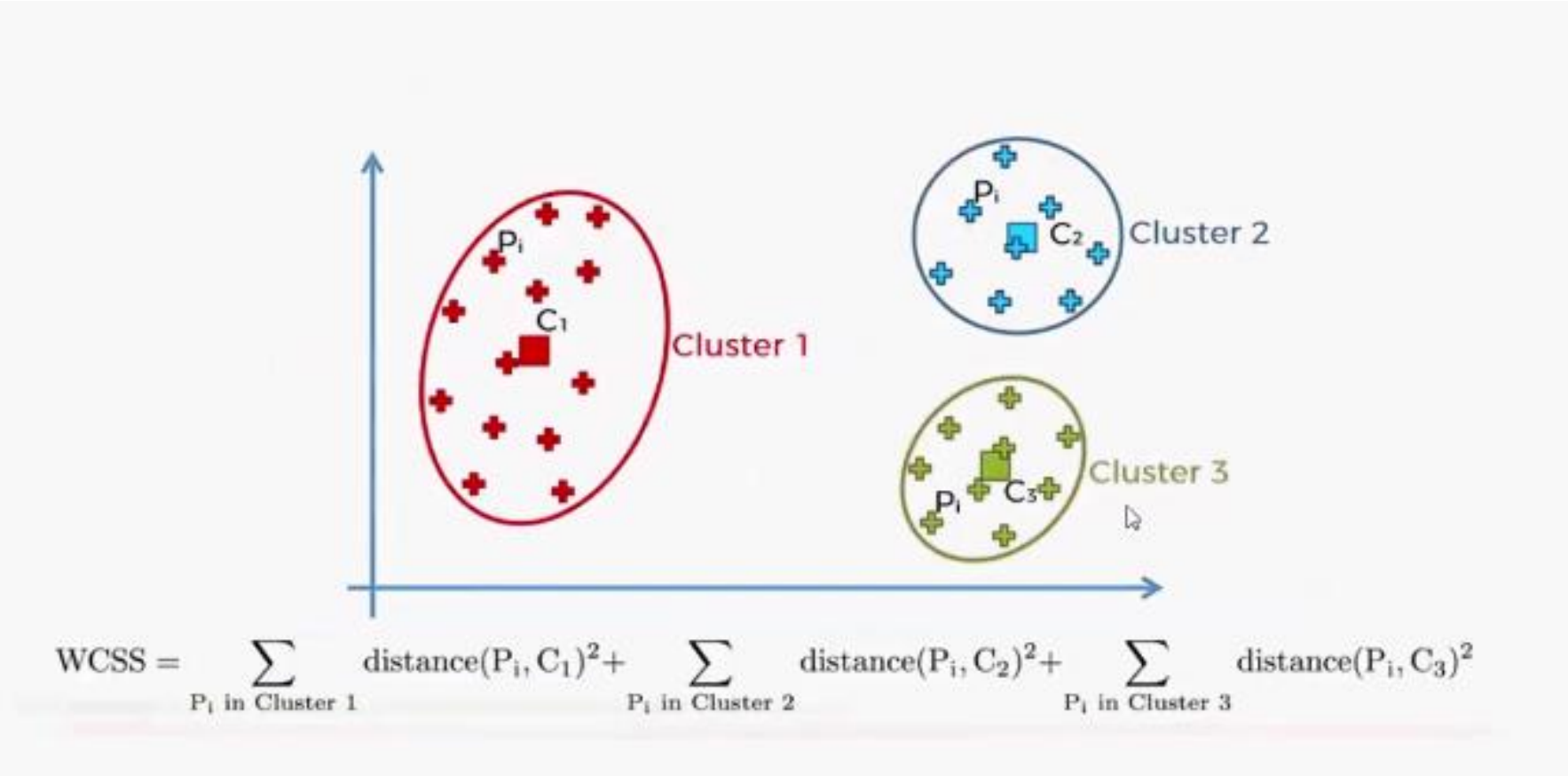

SELECTING THE
NUMBER OF CLUSTERS
Pi
C1
Cluster 1
Pi
C2
Cluster 2
Pi
C3
Cluster 3

K-Means

WITHIN CLUSTER SUM OF SQUARES

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

K-Means



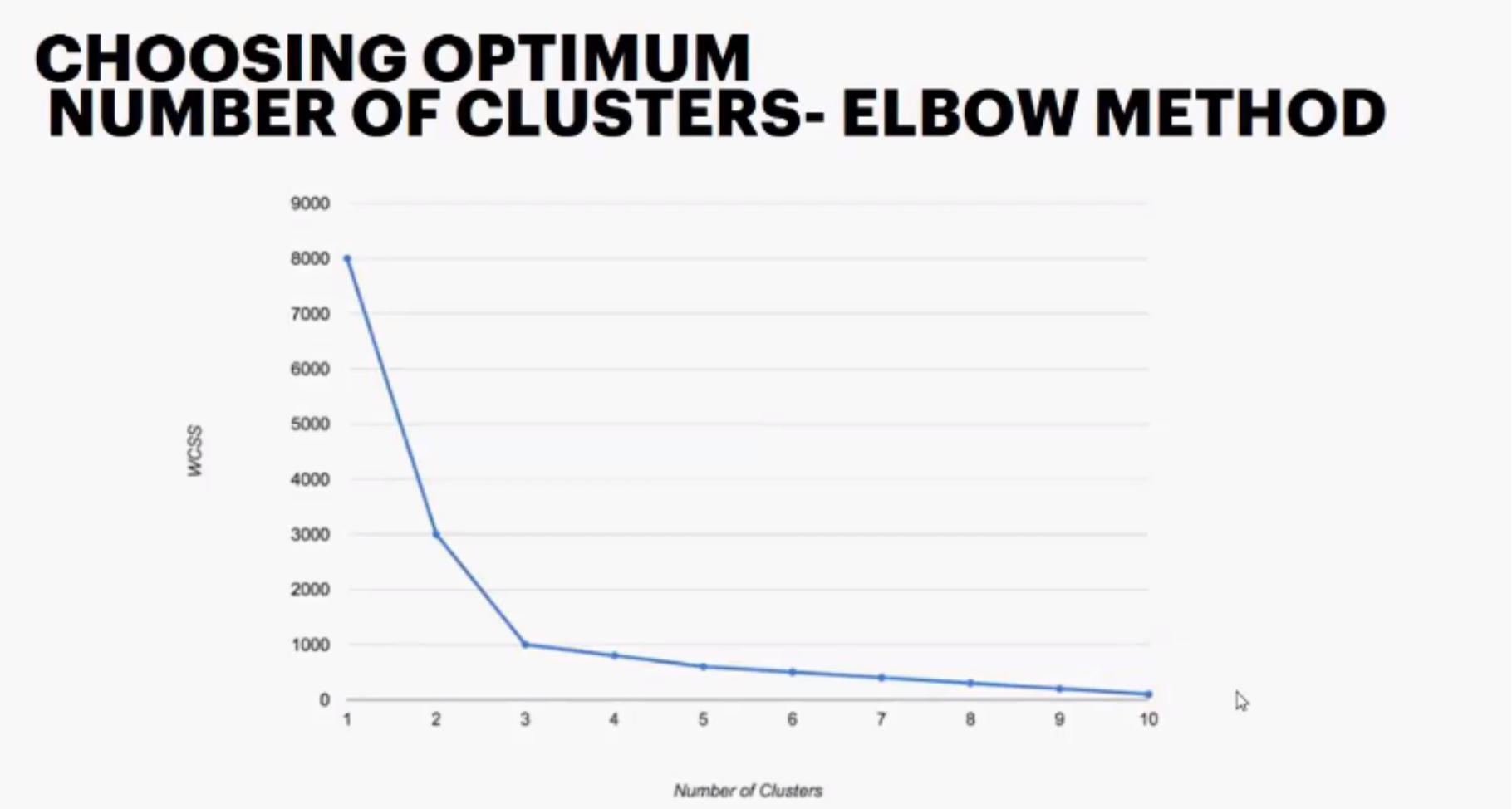
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

K-Means

If the number of clusters = to the number of points

What is WCSS value ?

K-Means


CHOOSING OPTIMUM
NUMBER OF CLUSTERS- ELBOW METHOD
WCSS
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
Number of Clusters

Hierarchical Clustering

Hierarchical Clustering

AGGLOMERATIVE CLUSTERING

Step 1: Make each data point a single point cluster

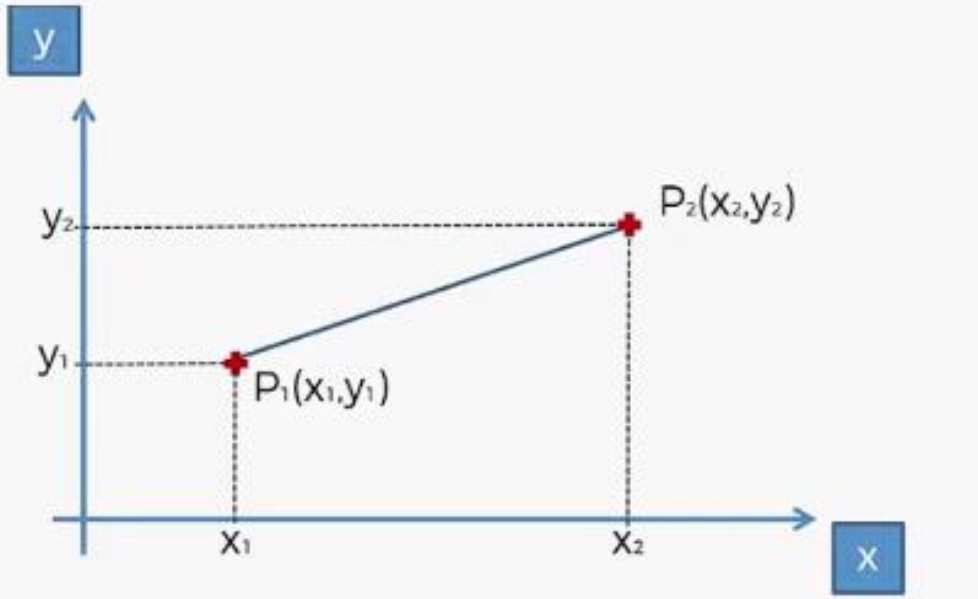
Step 2: Take the two closest data points and make them one cluster

Step3: Take the two closest clusters and make them one cluster

Step4: Repeat Step3 until there is only one cluster

Hierarchical Clustering

EUCLIDEAN DISTANCE



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

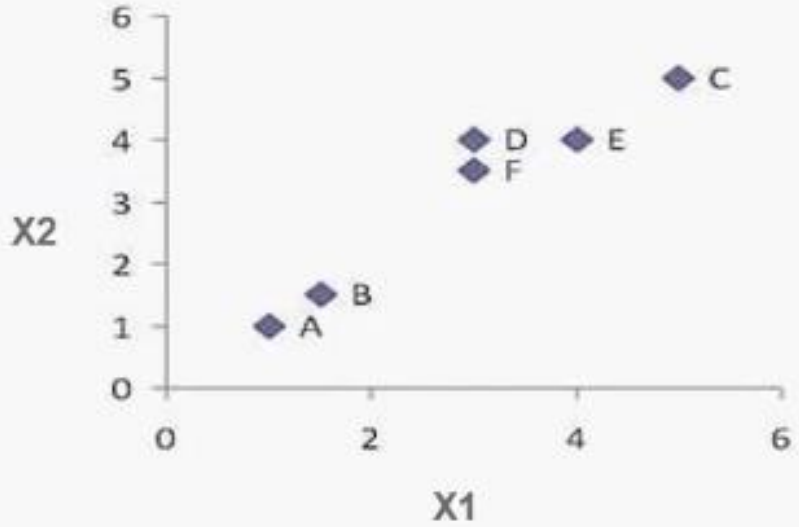
Hierarchical Clustering

HIERARCHICAL CLUSTERING

Simple Example:

Assume that there are have 6 objects namely A, B, C, D, E and F and each object have two measured features X1 and X2.

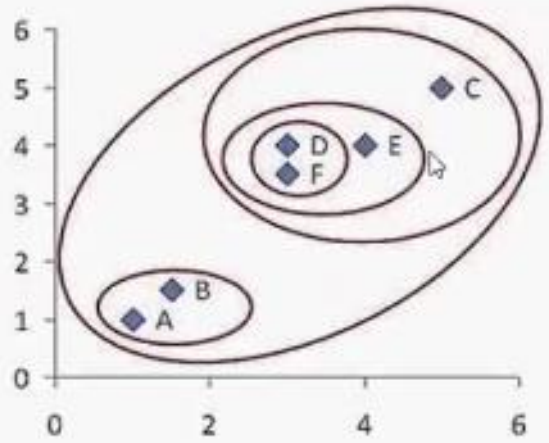
	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



SINGLE LINKAGE HIERARCHICAL CLUSTERING

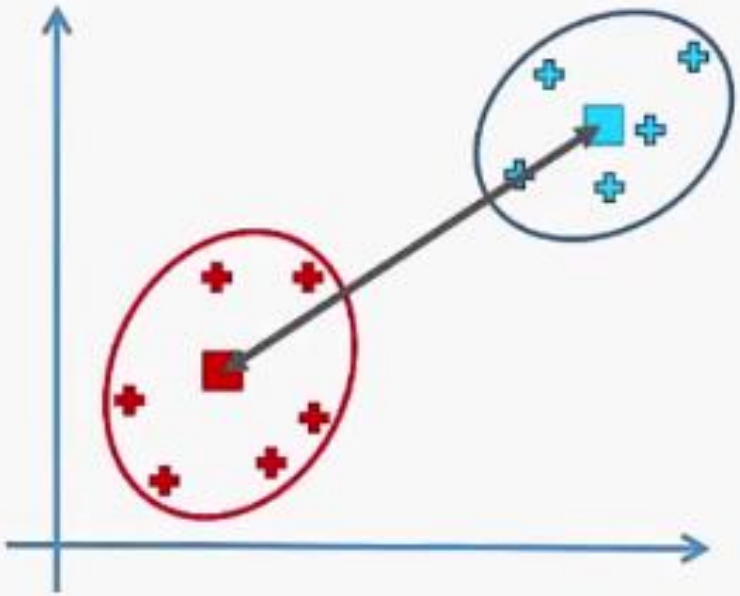
The dendrogram is drawn based on the distances to merge the clusters above

The hierarchy is given as (((D, F), E),C), (A,B). We can also plot the clustering hierarchy into XY space



Hierarchical Clustering

DISTANCE BETWEEN TWO CLUSTERS



1. closest points
2. Furthest points
3. Average distance
4. Distance between centroids

Hierarchical Clustering