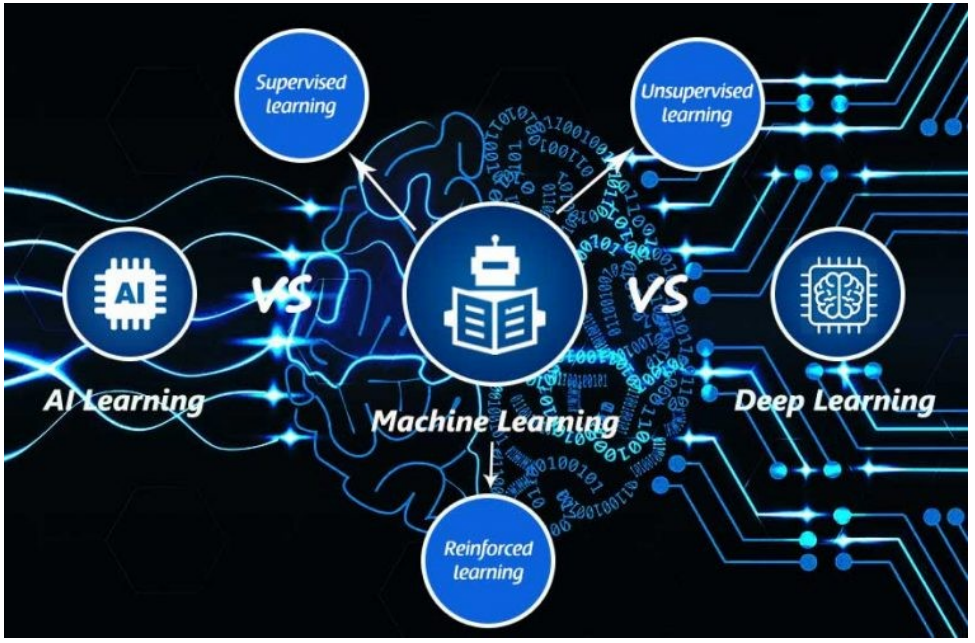
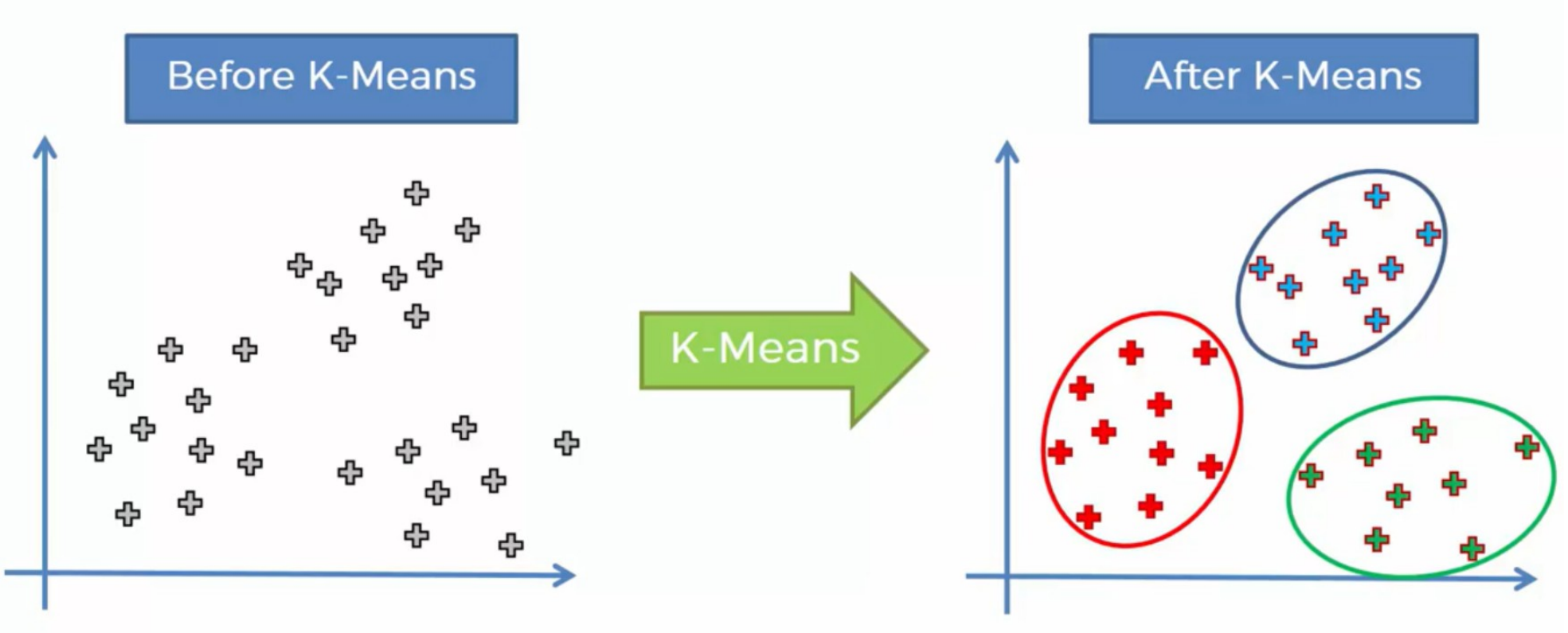


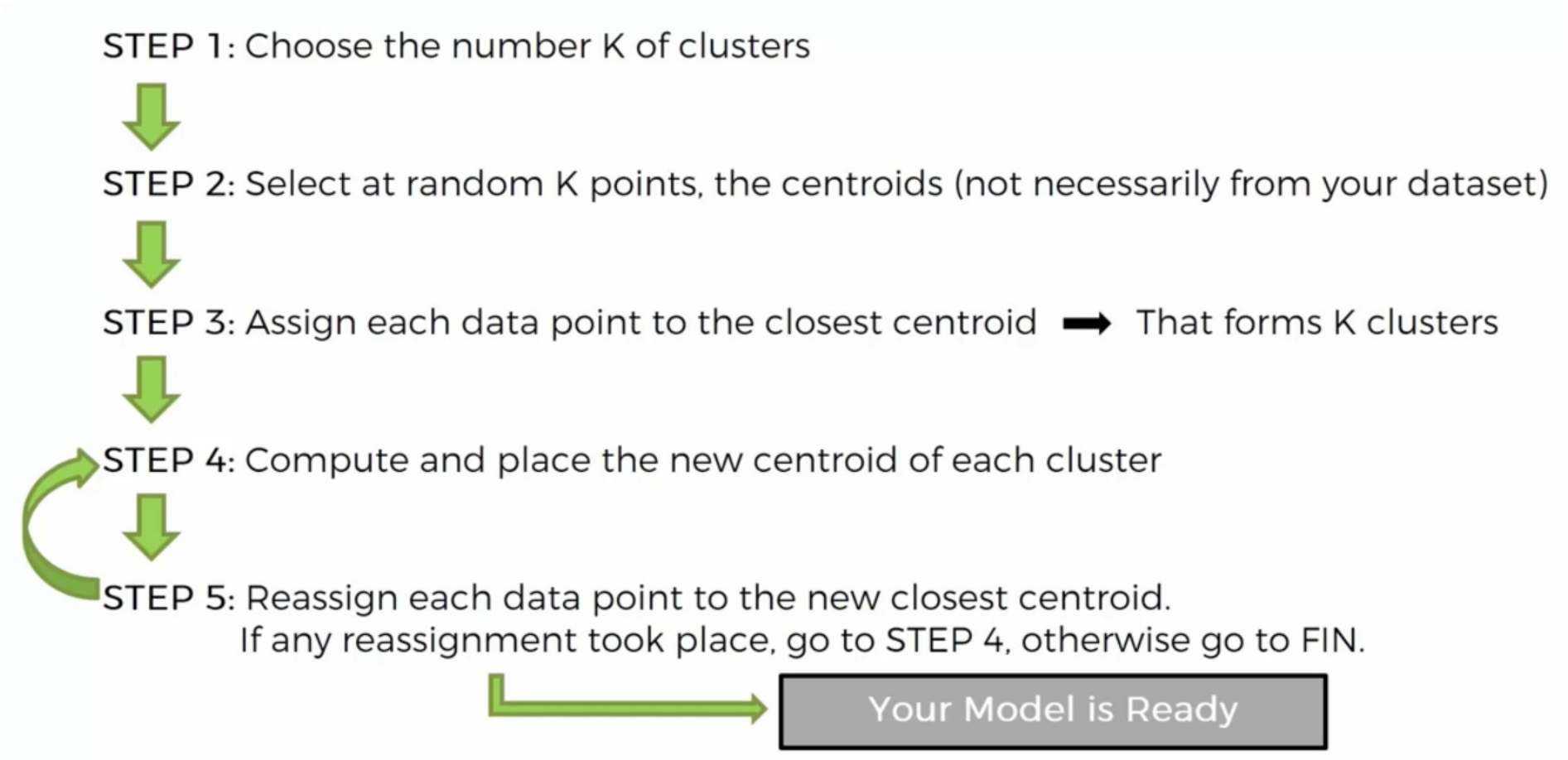
İTÜ
Machine Learning Lessons

Lesson3

K- Means Algorithm



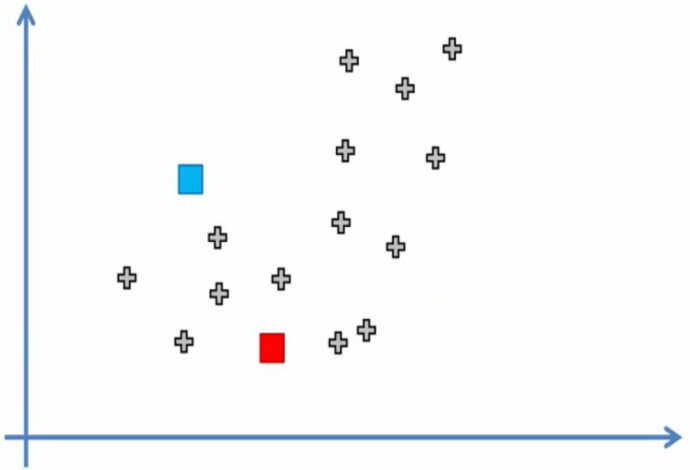
CLUSTERING


STEP 1: Choose the number K of clusters
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)
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 go to FIN.
Your Model is Ready

CLUSTERING

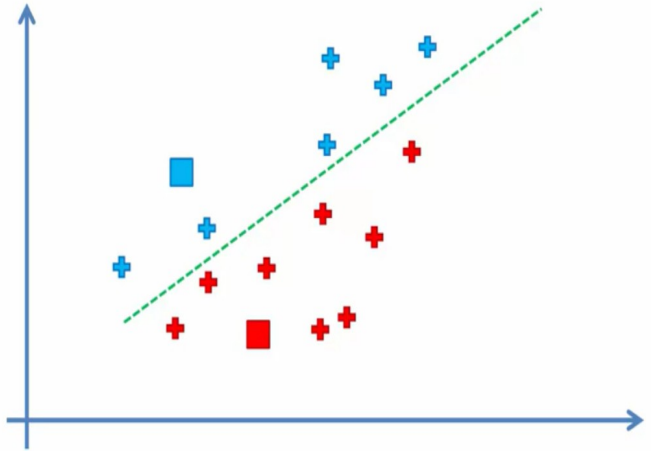
STEP 1: Choose the number K of clusters: K = 2

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

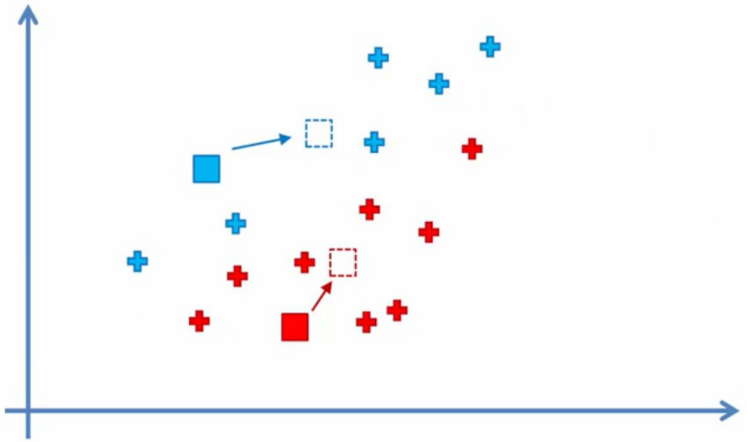


CLUSTERING

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

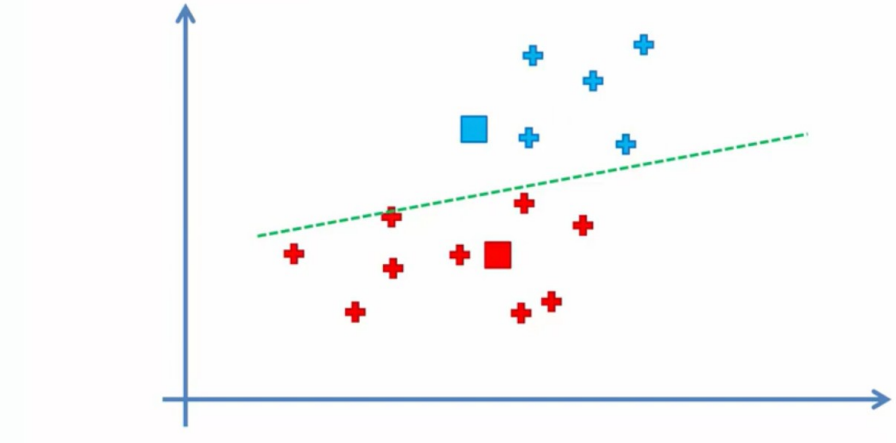


STEP 4: Compute and place the new centroid of each cluster

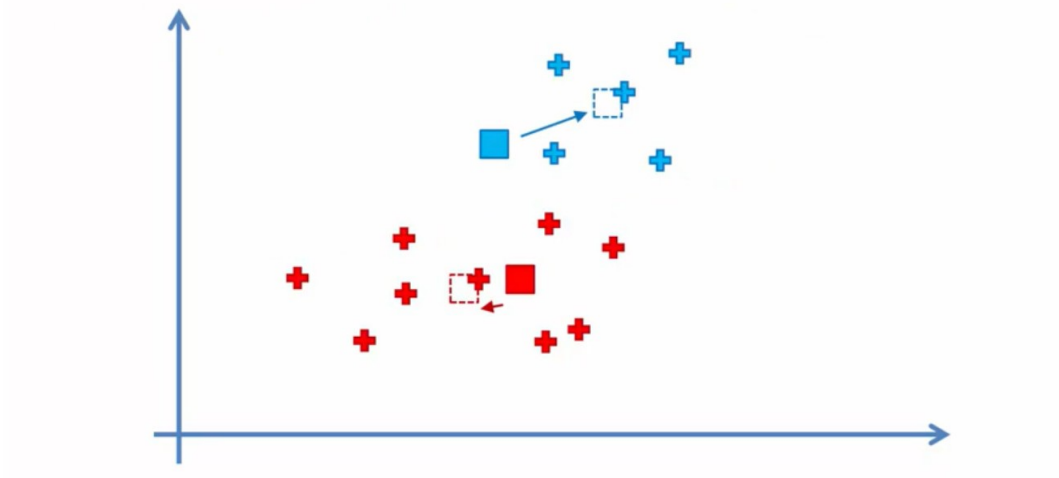


CLUSTERING

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

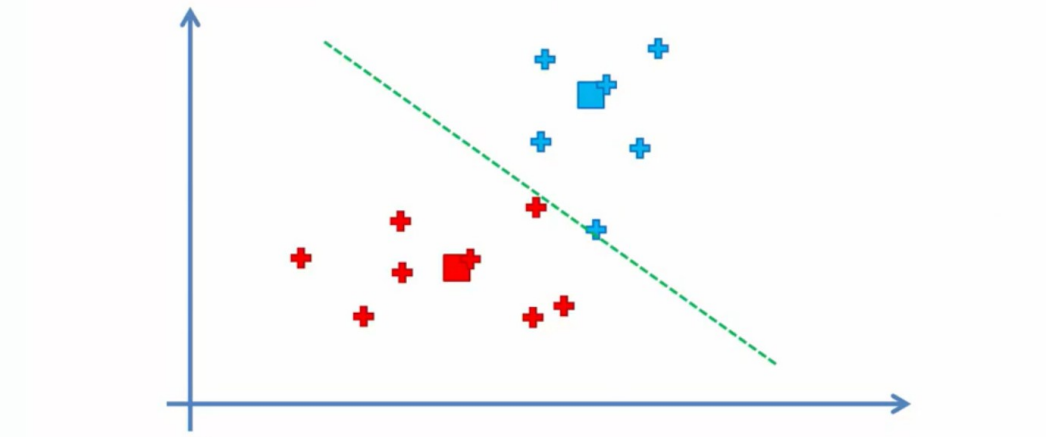


STEP 4: Compute and place the new centroid of each cluster

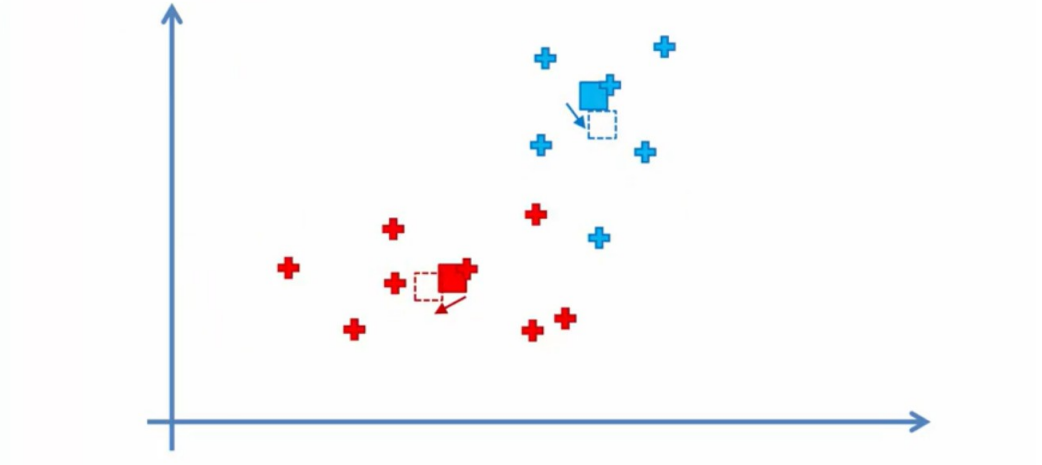


CLUSTERING

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

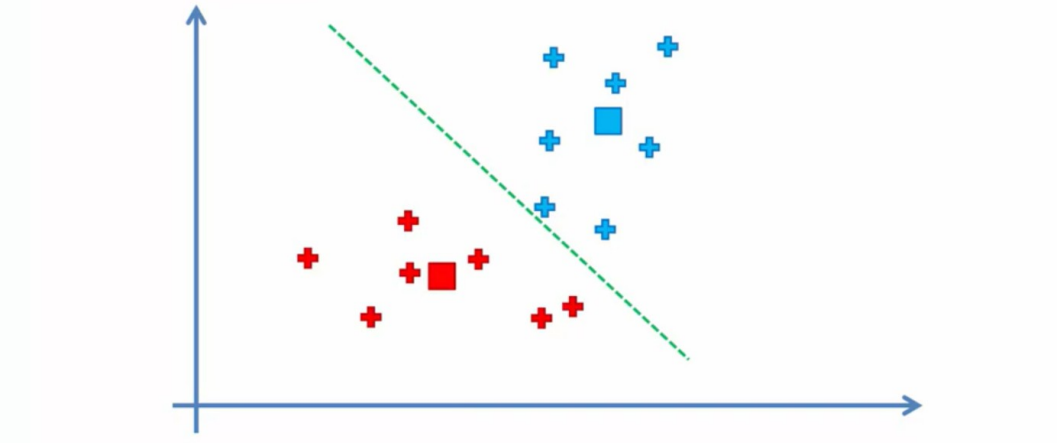


STEP 4: Compute and place the new centroid of each cluster

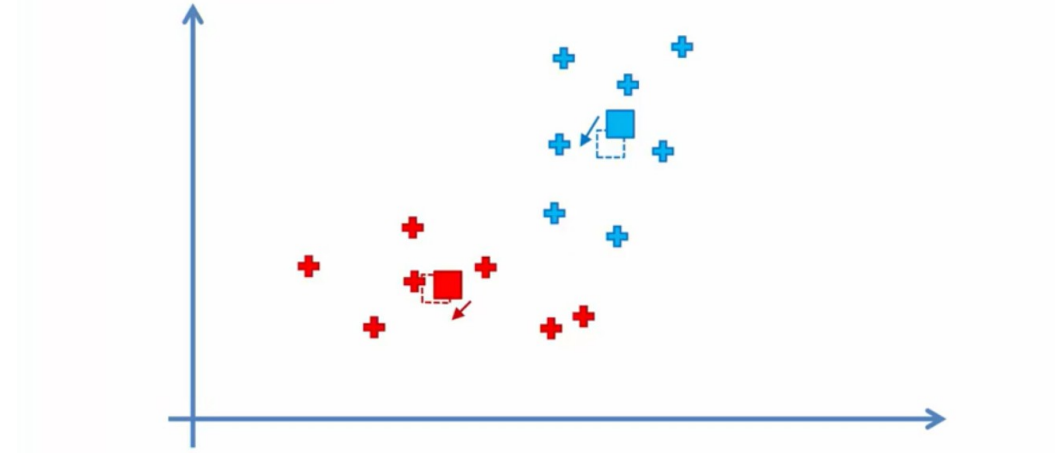


CLUSTERING

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

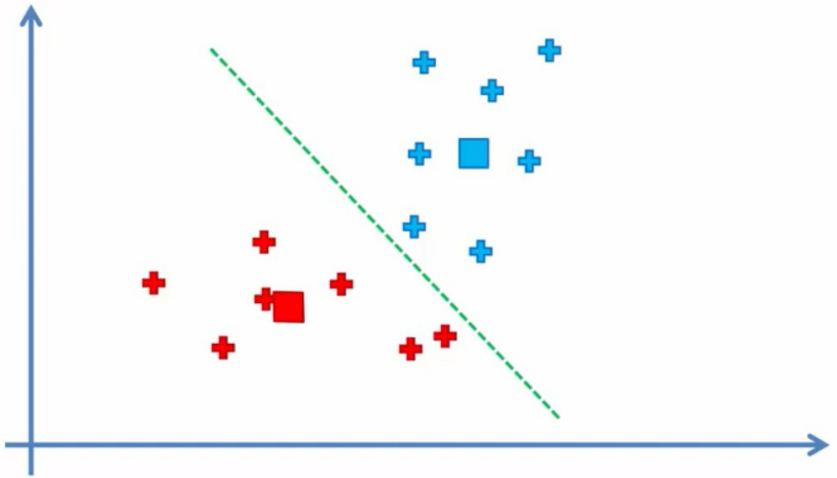


STEP 4: Compute and place the new centroid of each cluster

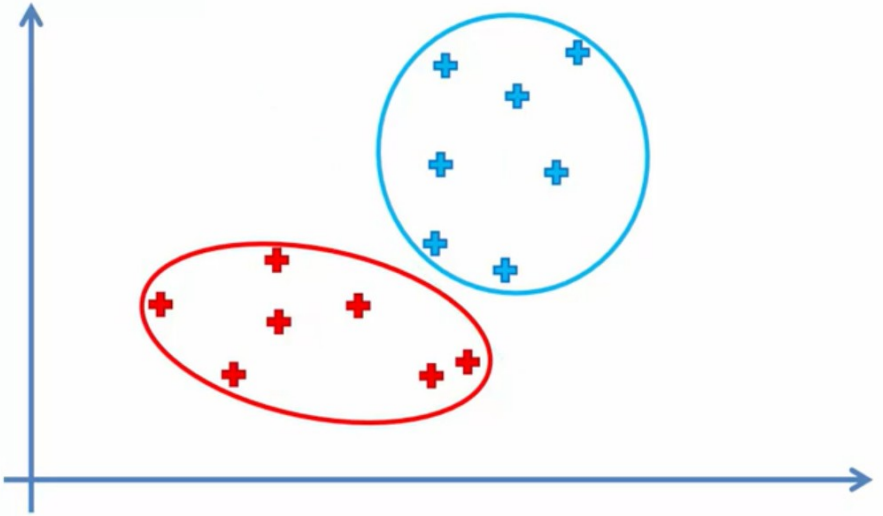


CLUSTERING

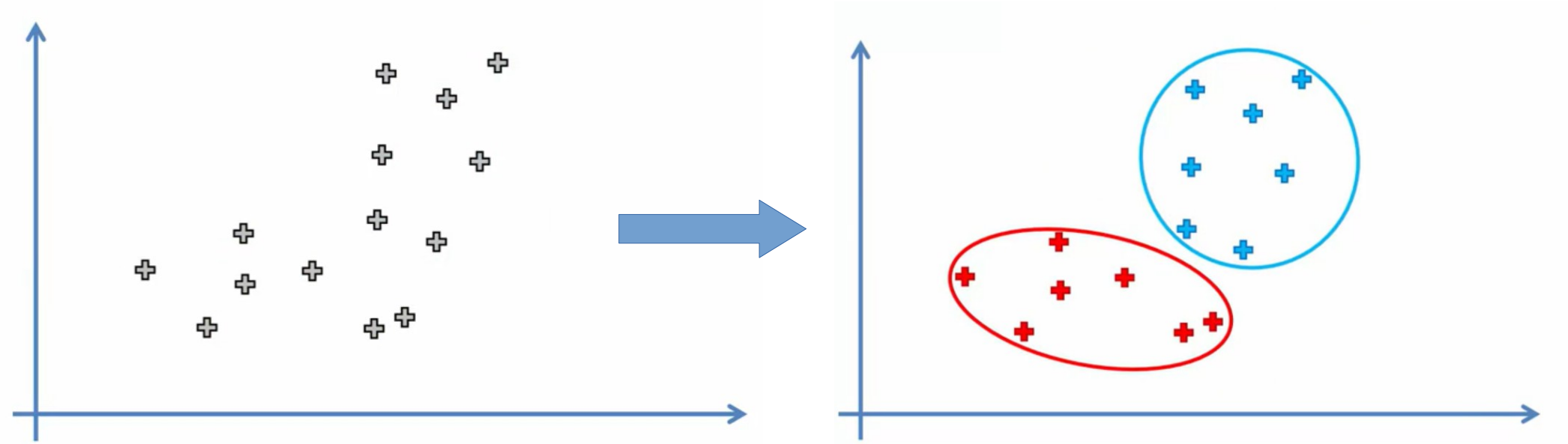
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



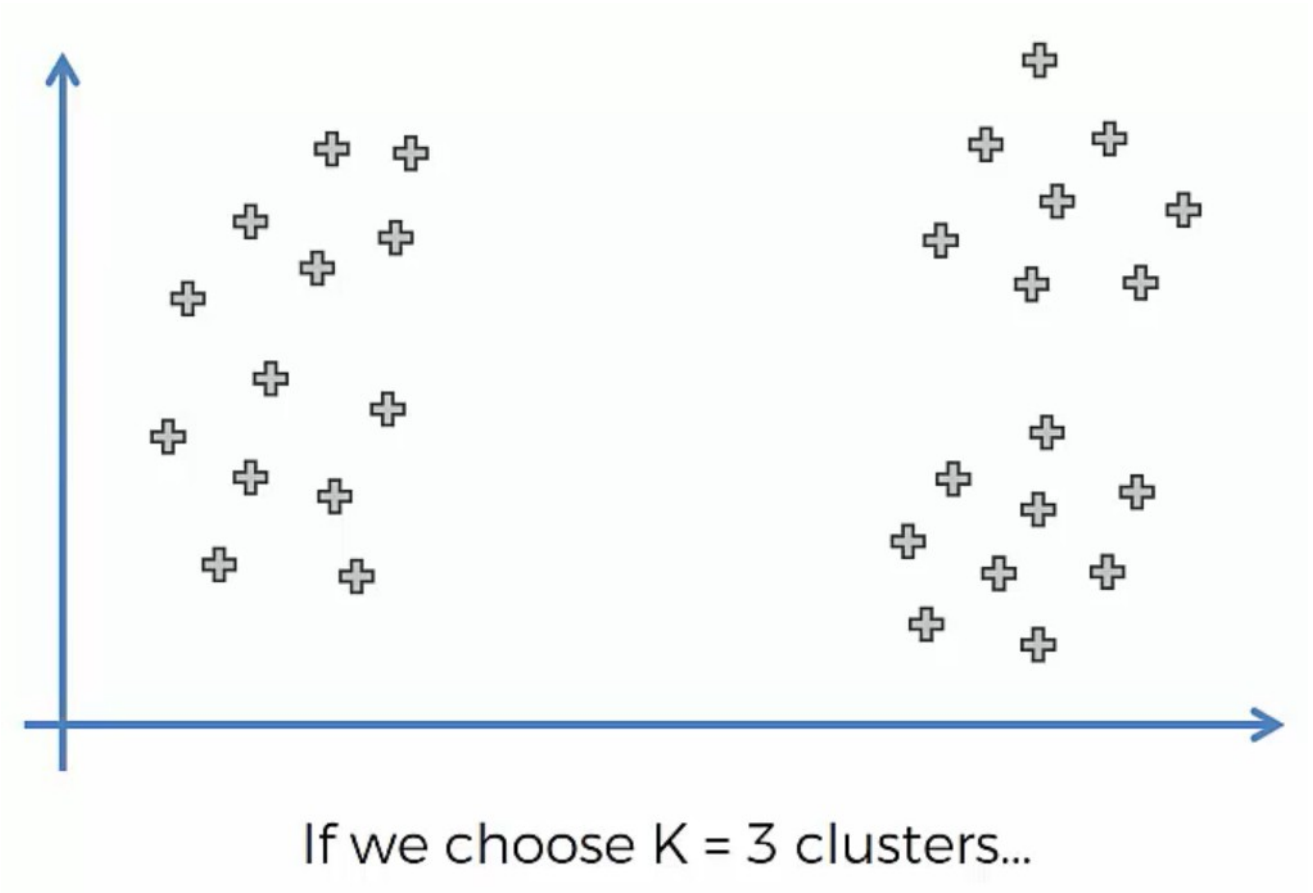
FIN: Your Model Is Ready



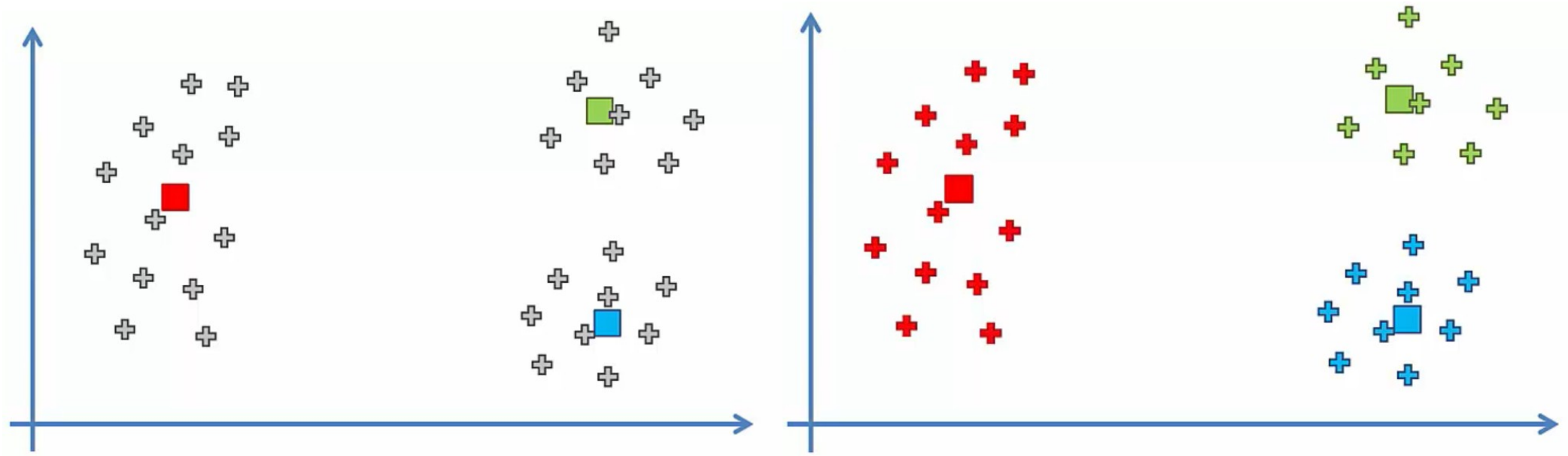
CLUSTERING



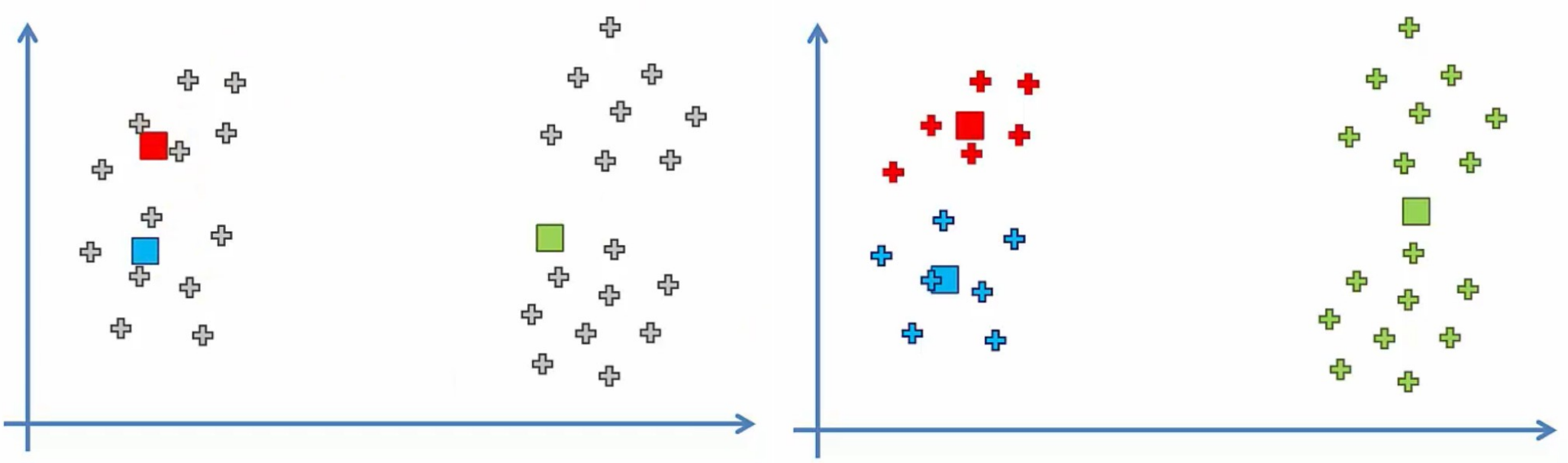
CLUSTERING



CLUSTERING

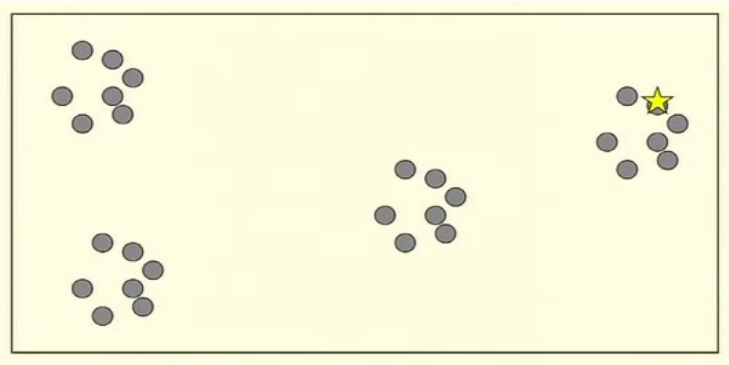


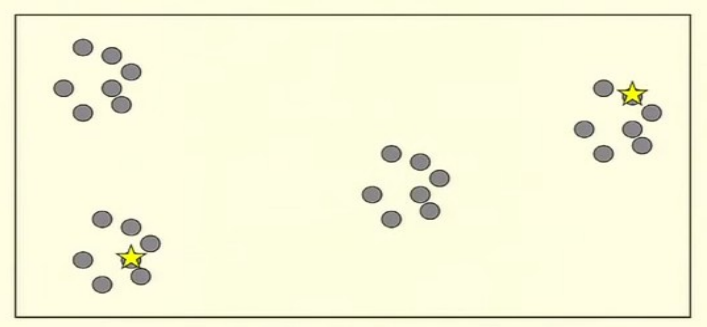
CLUSTERING

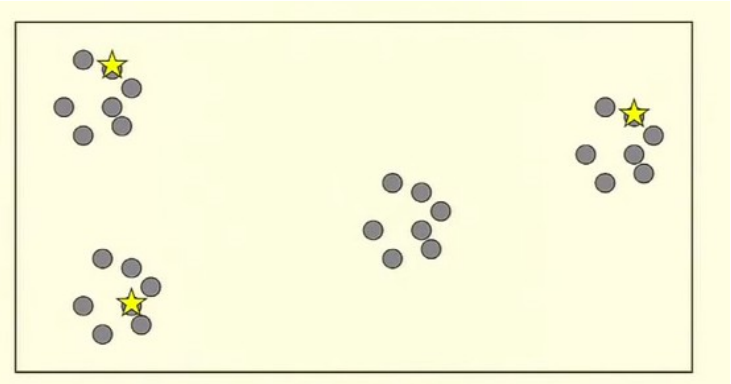


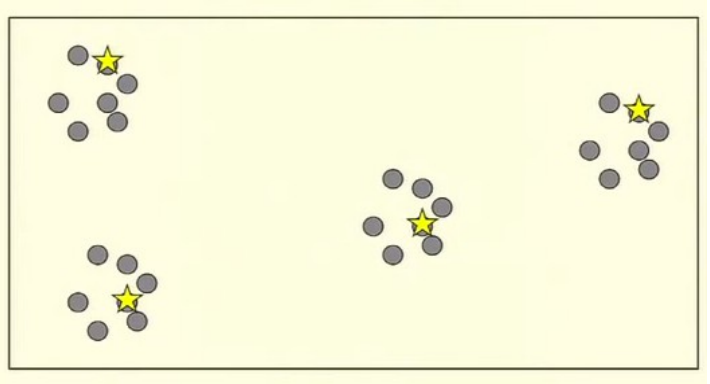
CLUSTERING

SOLUTION : K-MEANS ++ ALGORITHM



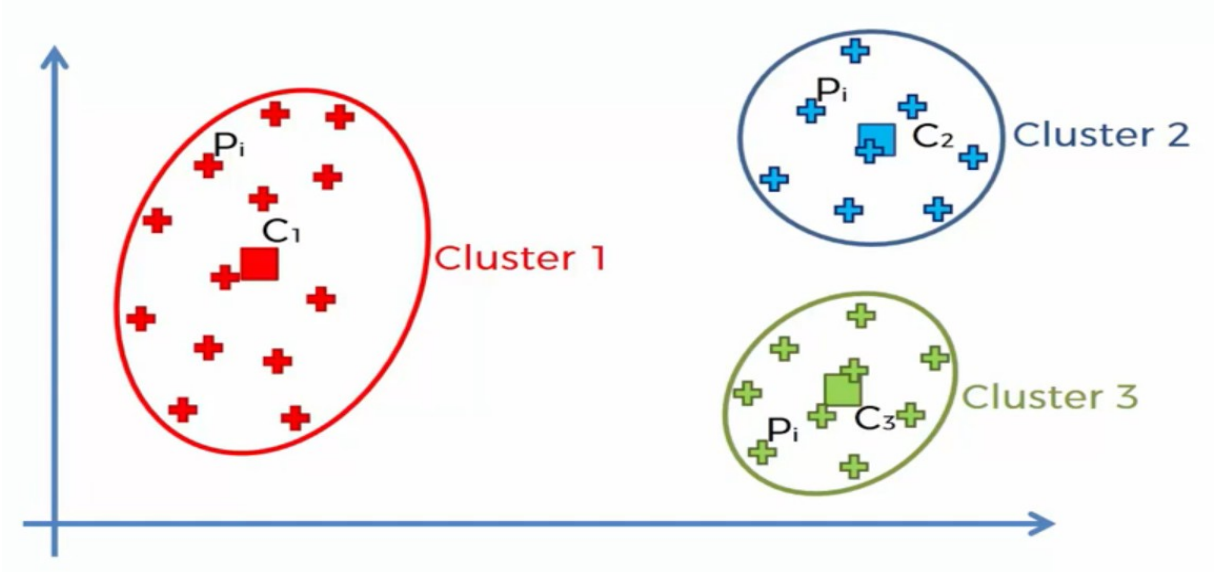






CLUSTERING

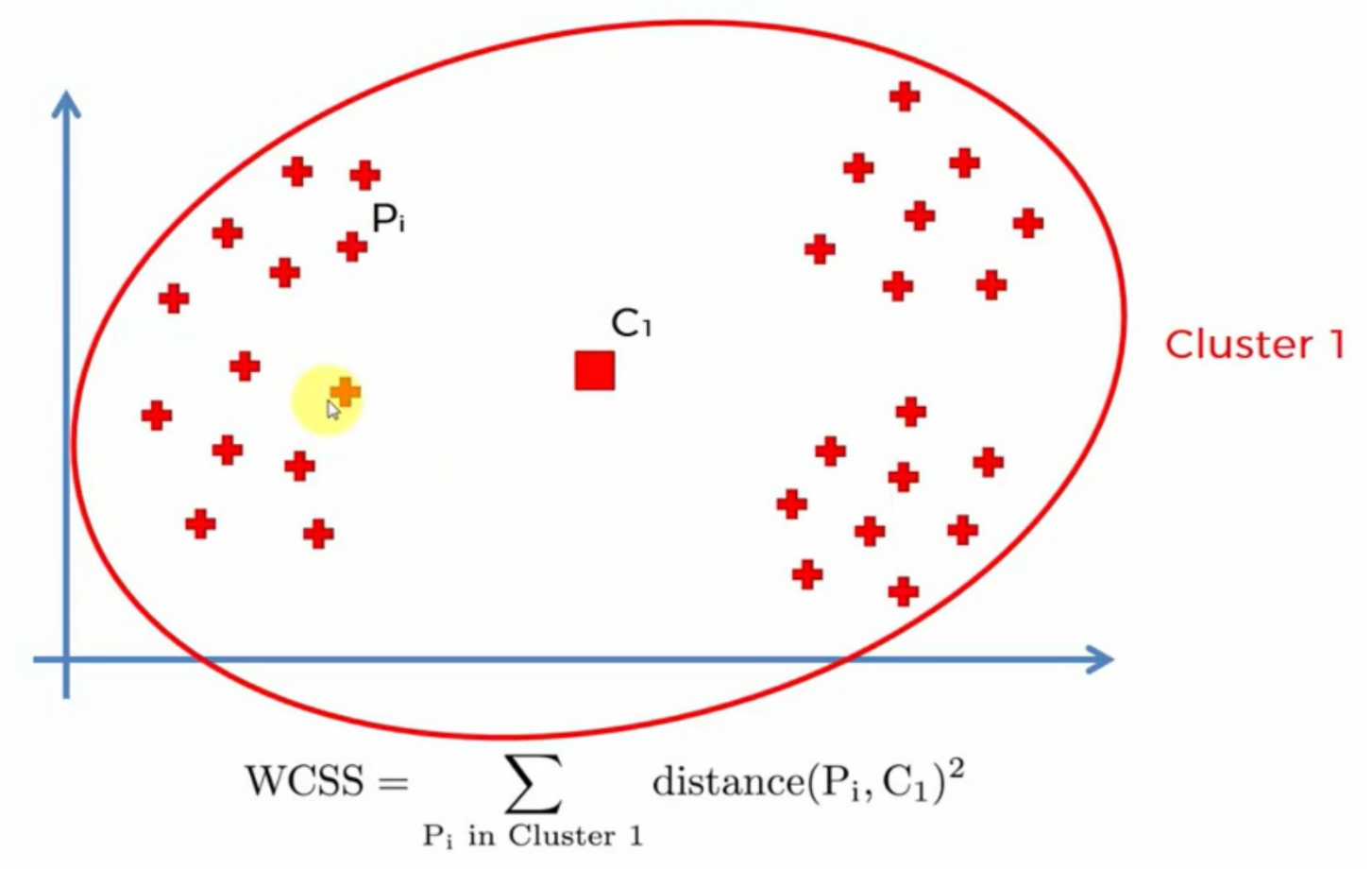
CHOOSING THE K VALUE



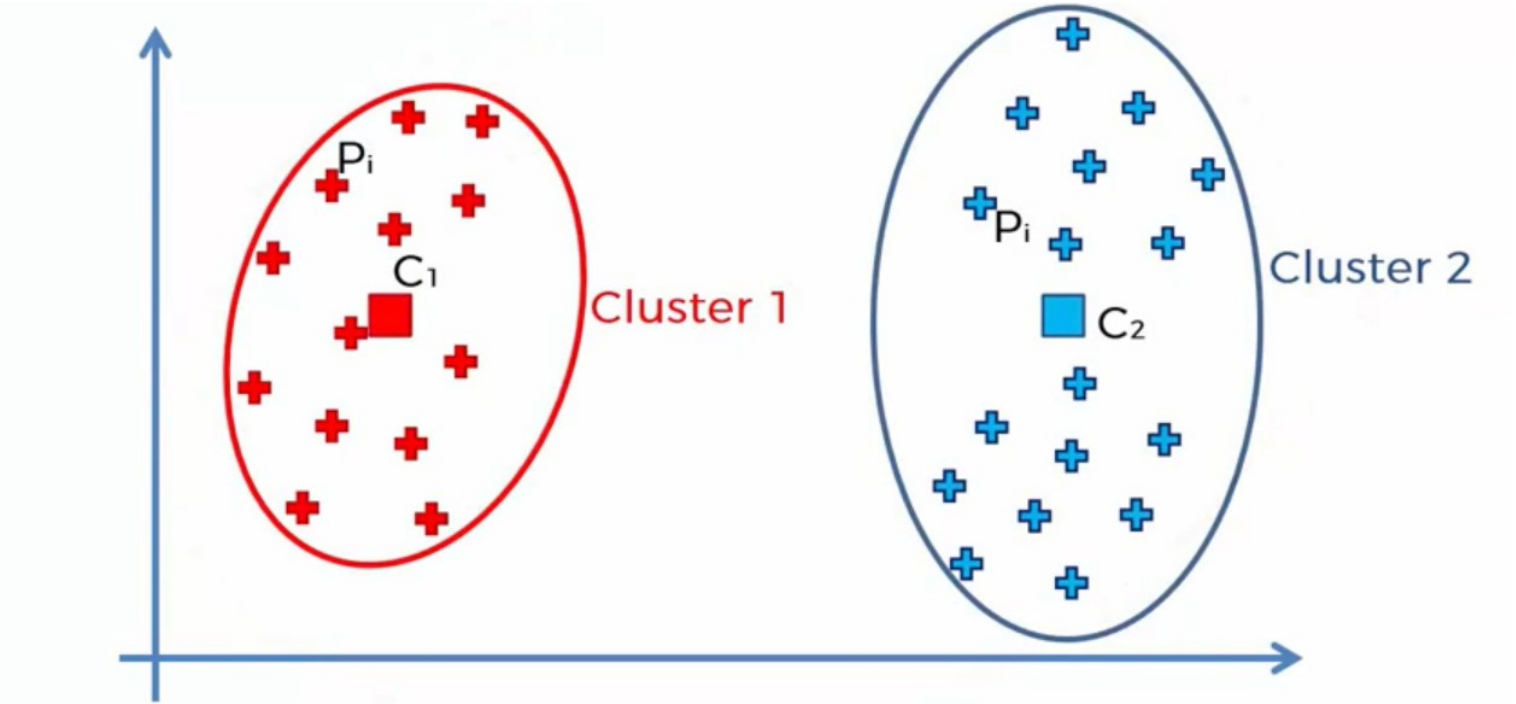
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

(within-cluster sums of squares)

CLUSTERING

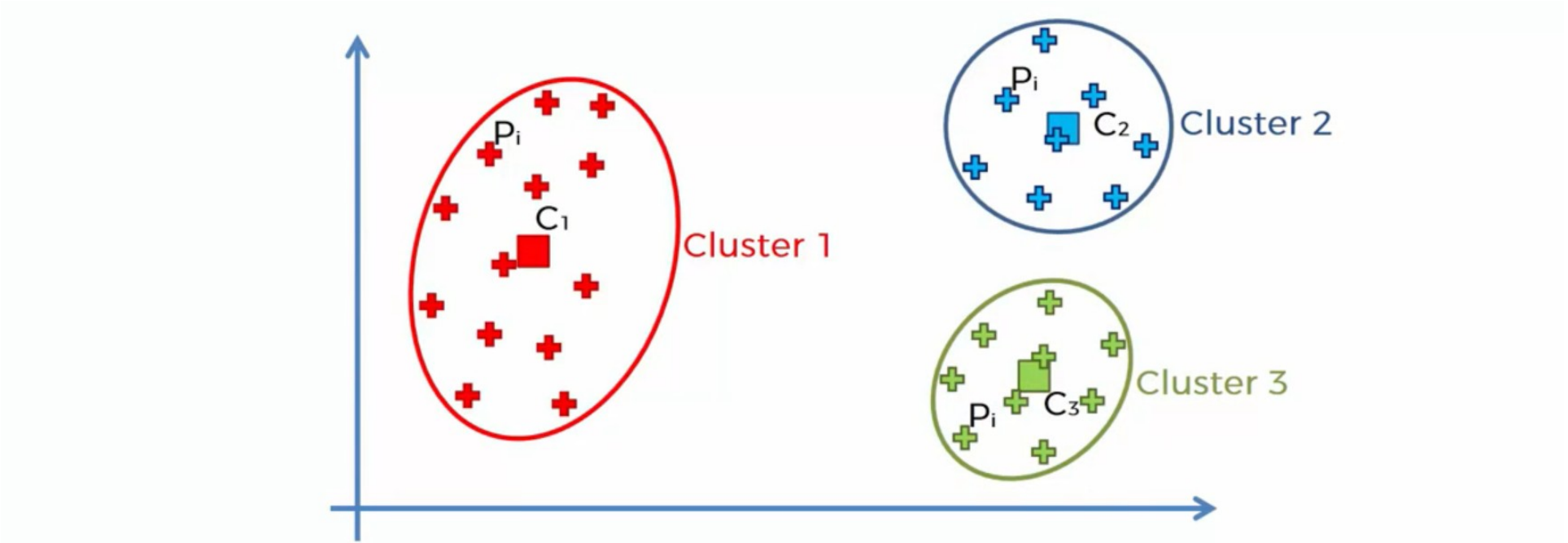

Pi
C1
Cluster 1
WCSS = $\sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2$

CLUSTERING



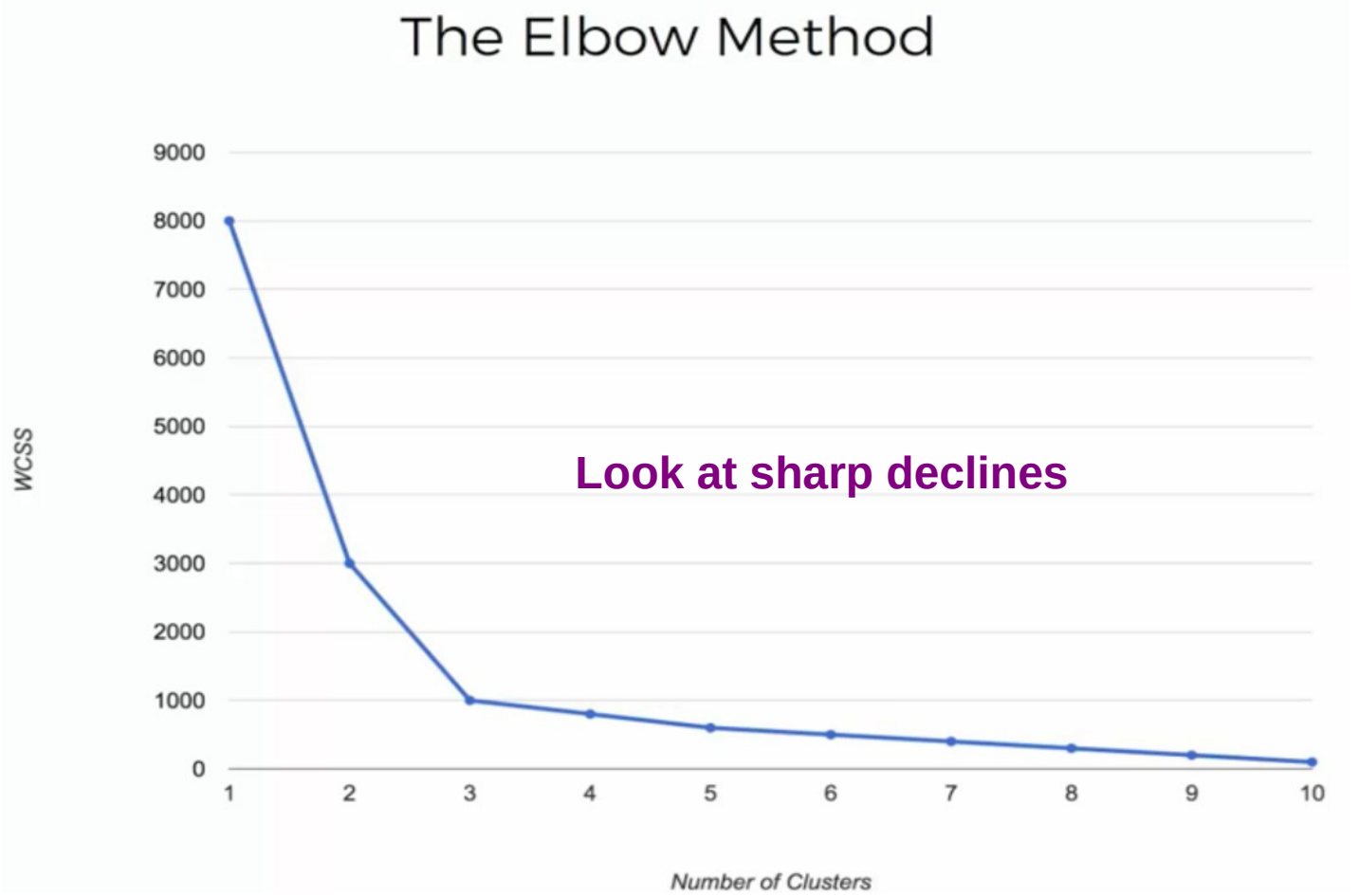
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$

CLUSTERING

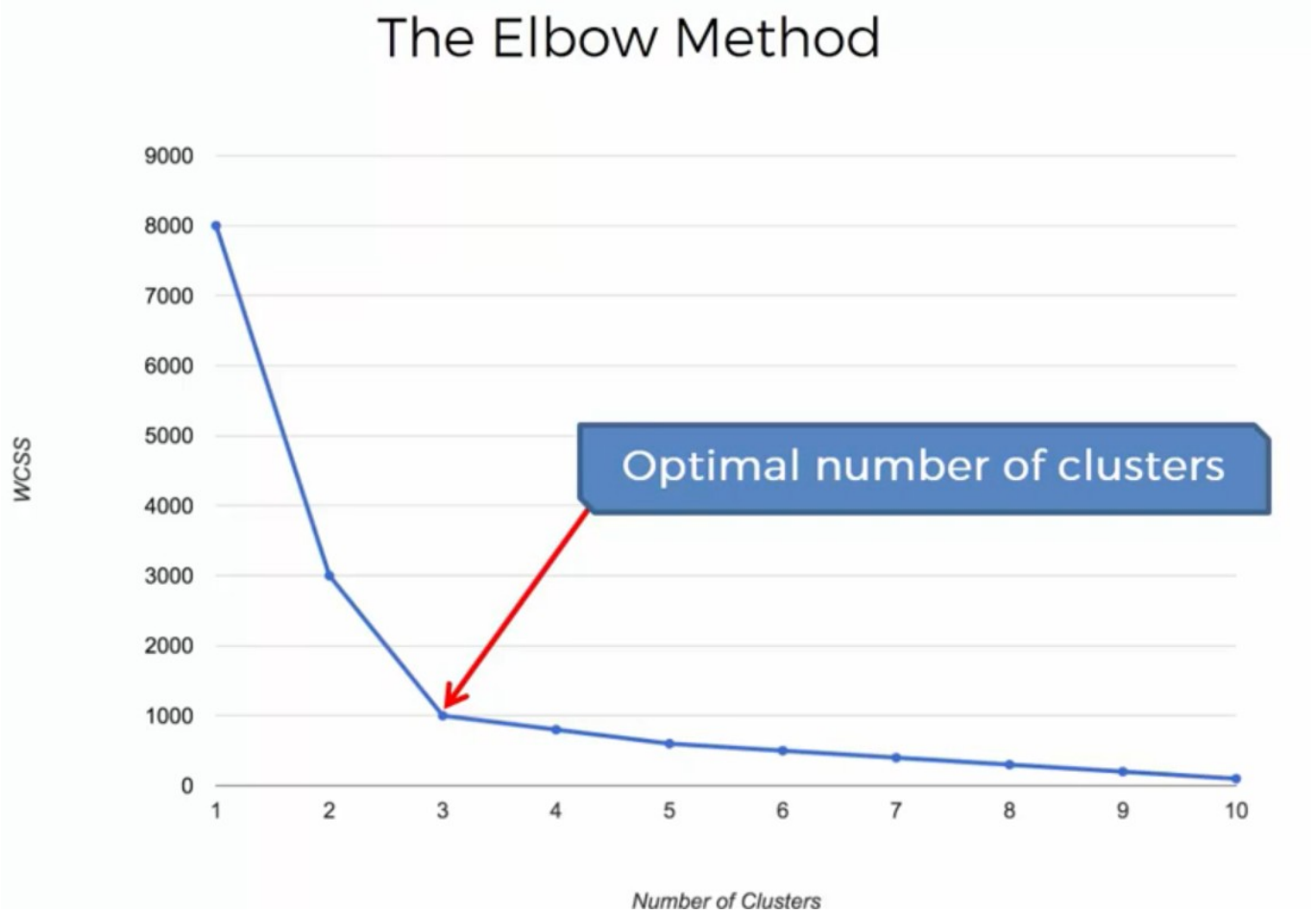


$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

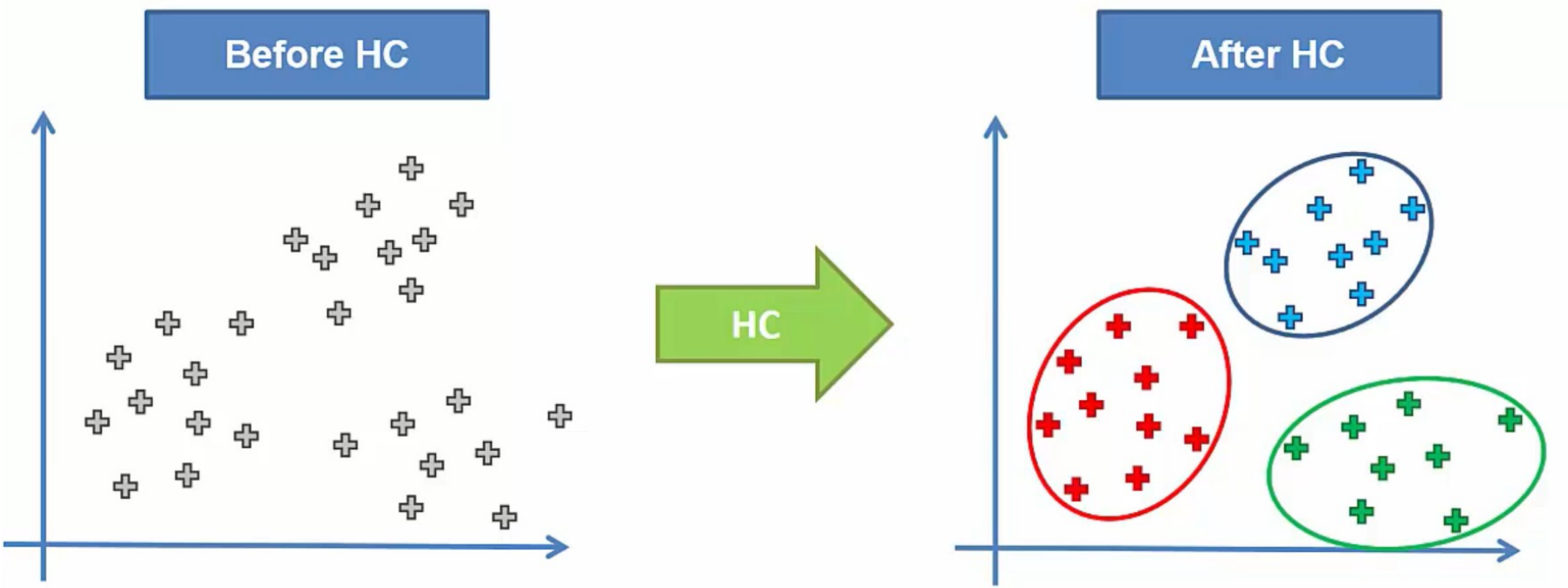
CLUSTERING


The 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
Look at sharp declines

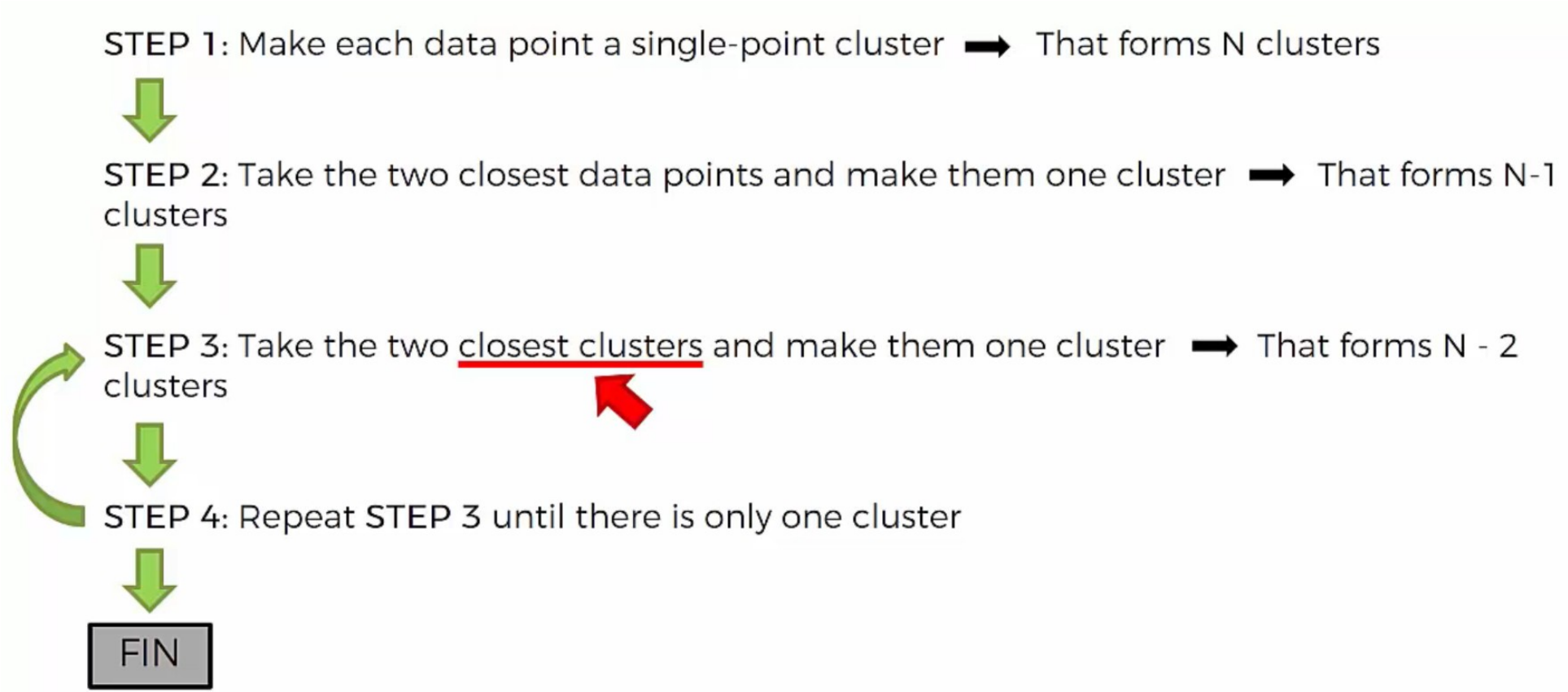
CLUSTERING


The Elbow Method
WCSS
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
Optimal number of clusters
Number of Clusters

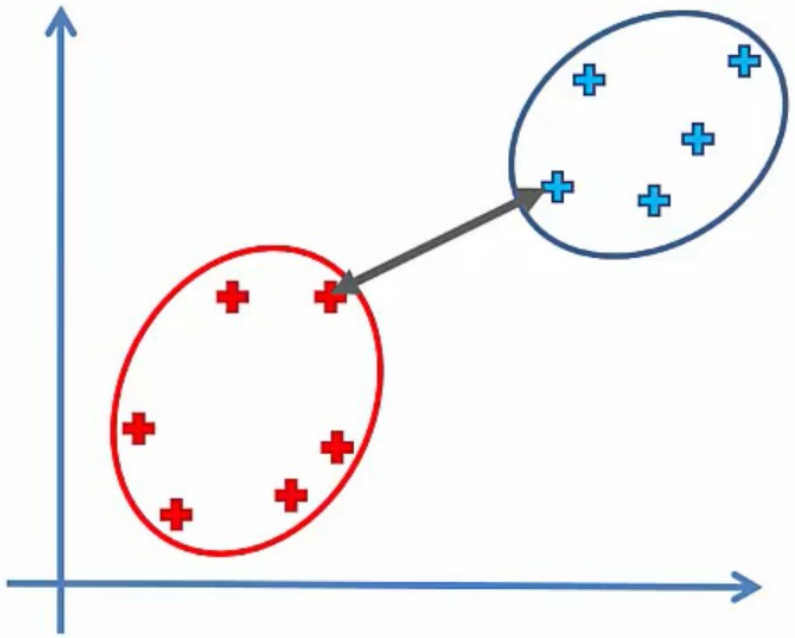
HIERARCHICAL CLUSTERING



HIERARCHICAL CLUSTERING



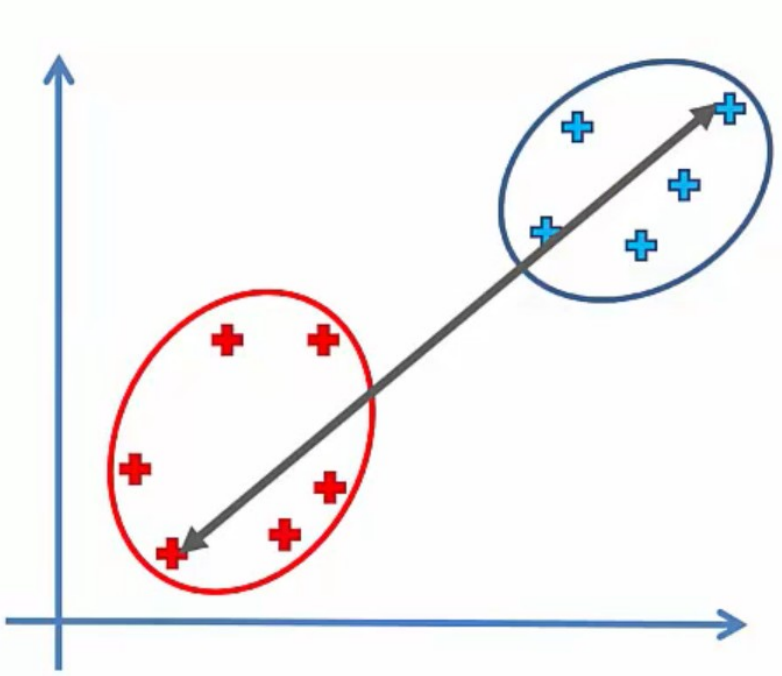
HIERARCHICAL CLUSTERING



Distance Between Two Clusters:

- Option 1: Closest Points

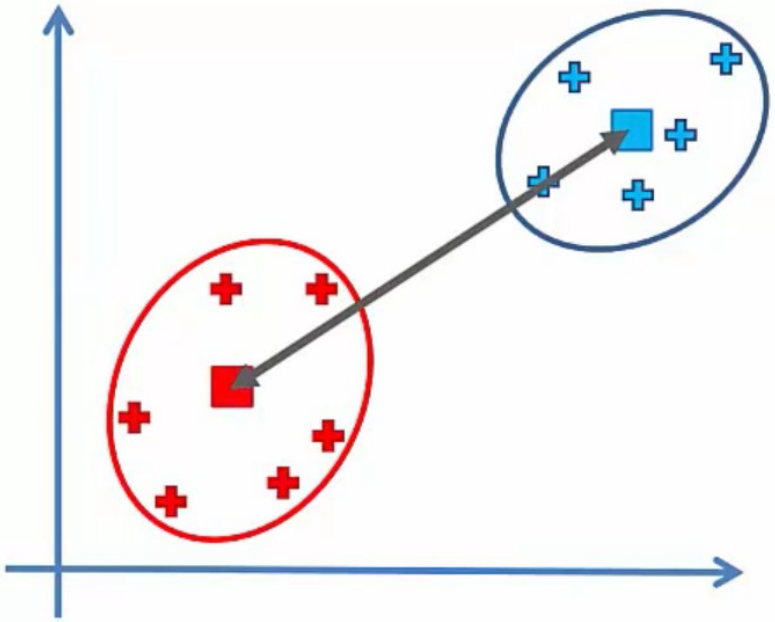
HIERARCHICAL CLUSTERING



Distance Between Two Clusters:

- Option 1: Closest Points
- Option 2: Furthest Points

HIERARCHICAL CLUSTERING



Distance Between Two Clusters:

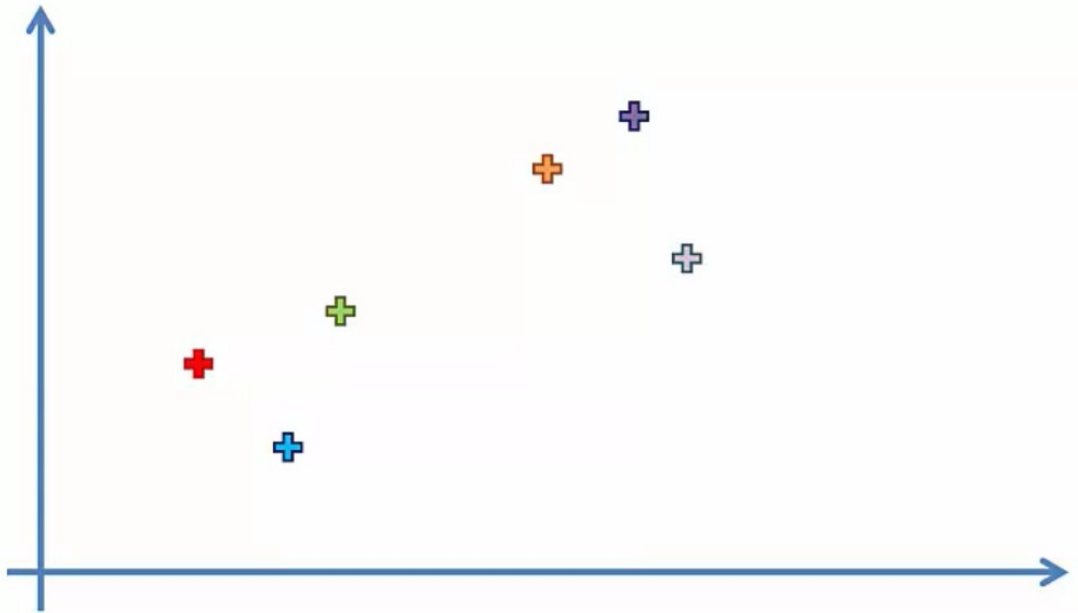
- Option 1: Closest Points
- Option 2: Furthest Points
- Option 3: Average Distance
- Option 4: Distance Between Centroids

HIERARCHICAL CLUSTERING

STEP 1: Make each data point a single-point cluster ➡ That forms 6 clusters

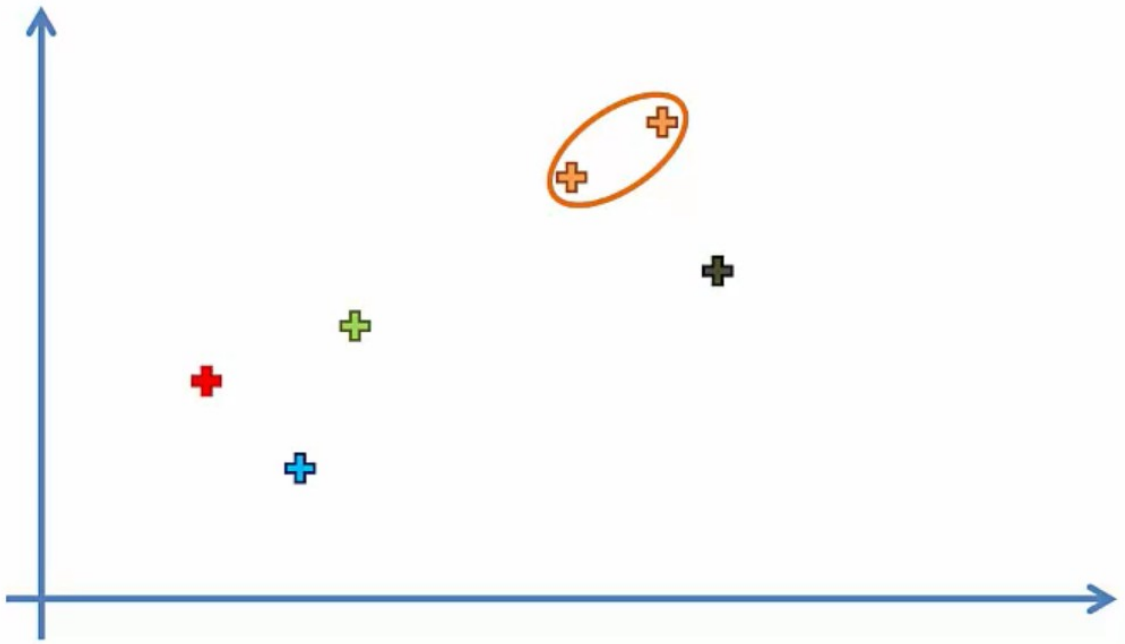
HIERARCHICAL CLUSTERING

STEP 1: Make each data point a single-point cluster ➡ That forms 6 clusters



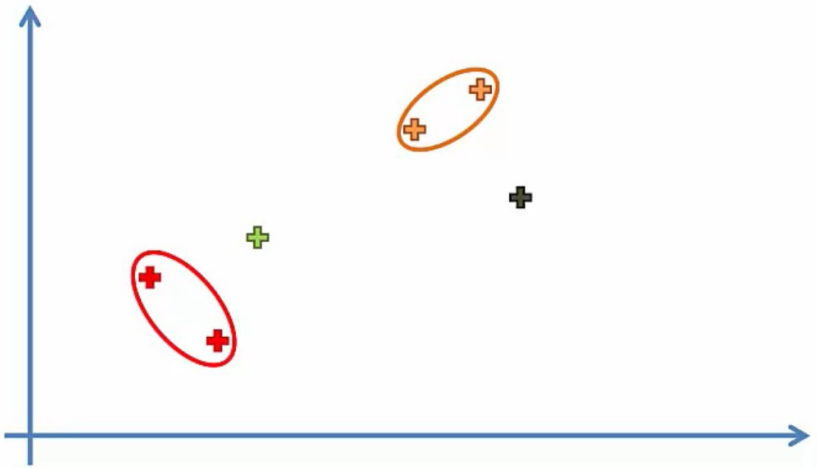
HIERARCHICAL CLUSTERING

STEP 2: Take the two closest data points and make them one cluster
➞ That forms 5 clusters

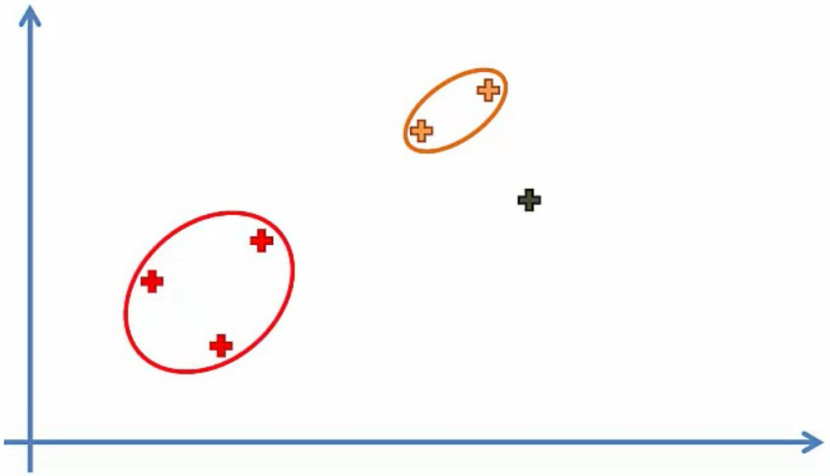


HIERARCHICAL CLUSTERING

STEP 3: Take the two closest clusters and make them one cluster
➡That forms 4 clusters

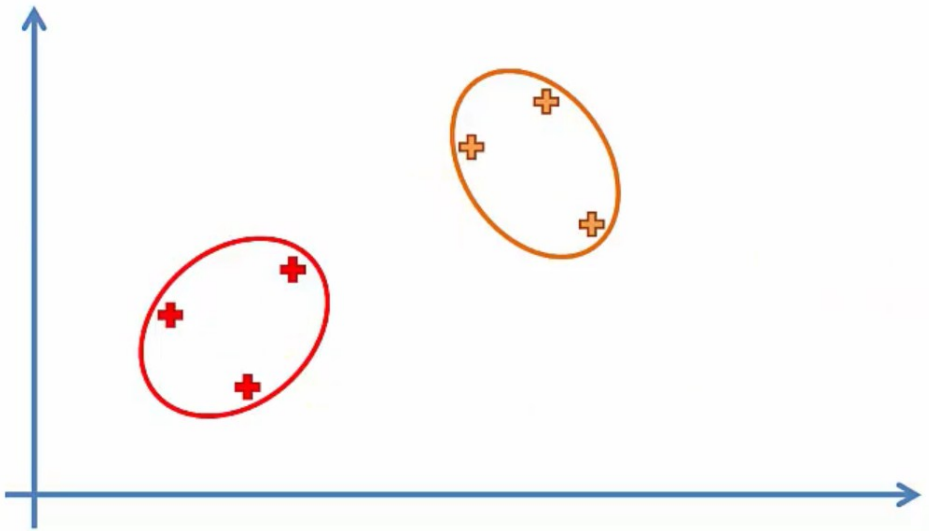


STEP 4: Repeat STEP 3 until there is only one cluster

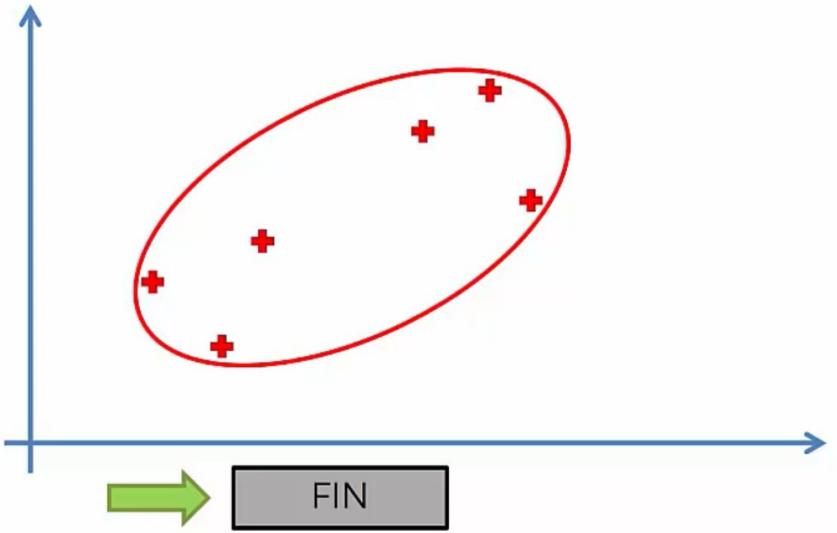


HIERARCHICAL CLUSTERING

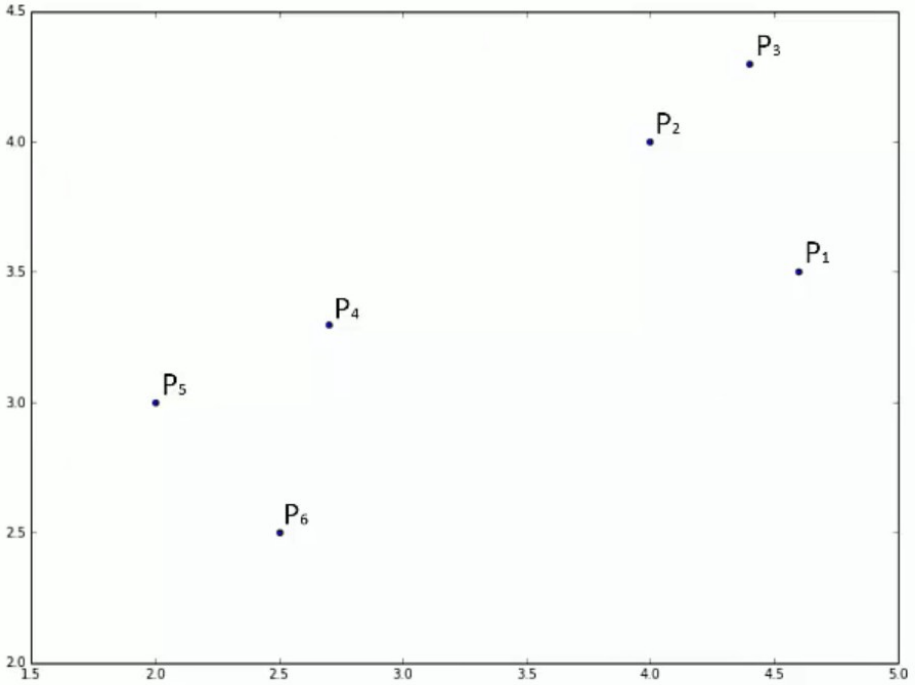
STEP 4: Repeat STEP 3 until there is only one cluster

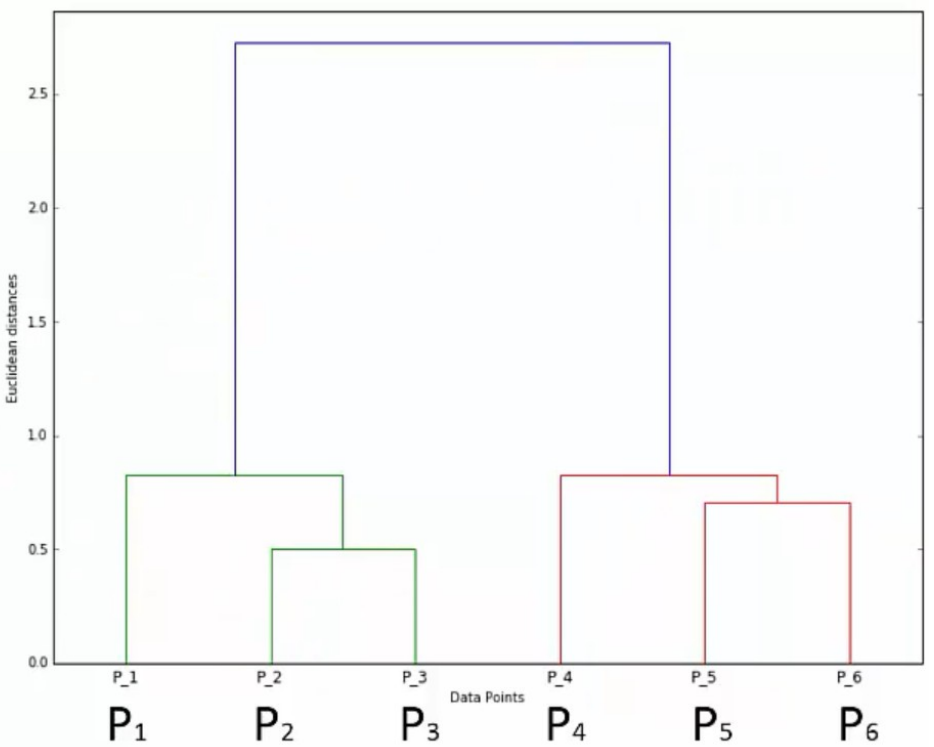


STEP 4: Repeat STEP 3 until there is only one cluster

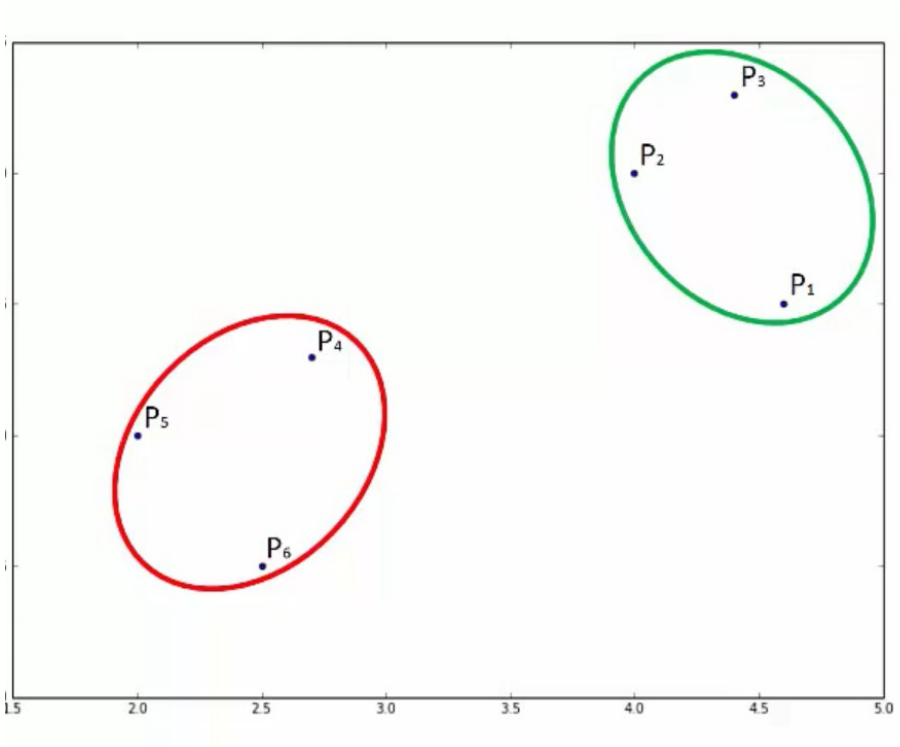


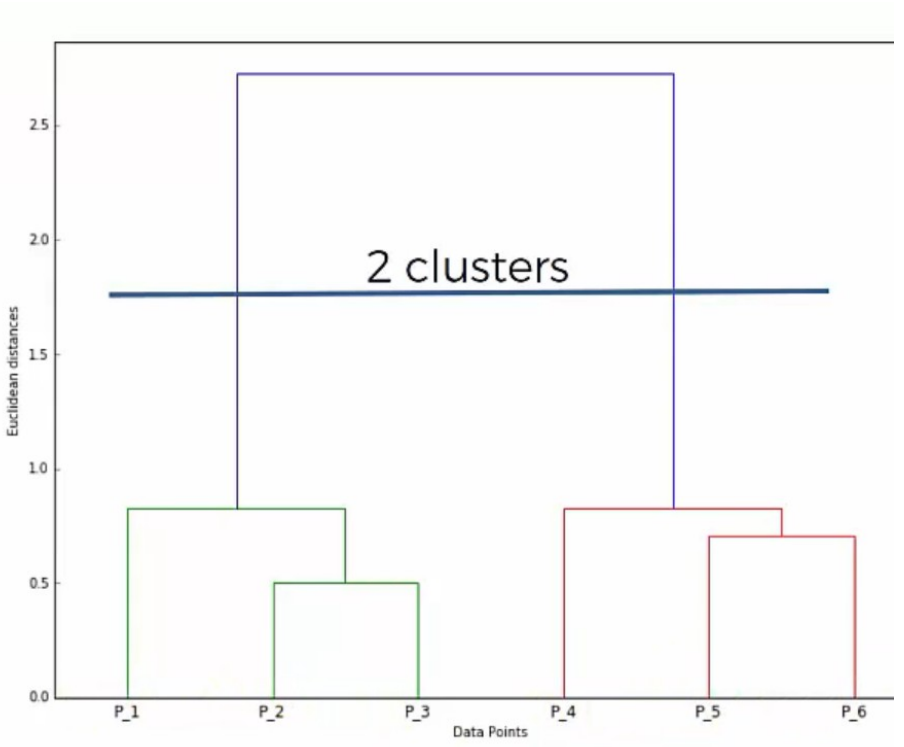
HIERARCHICAL CLUSTERING



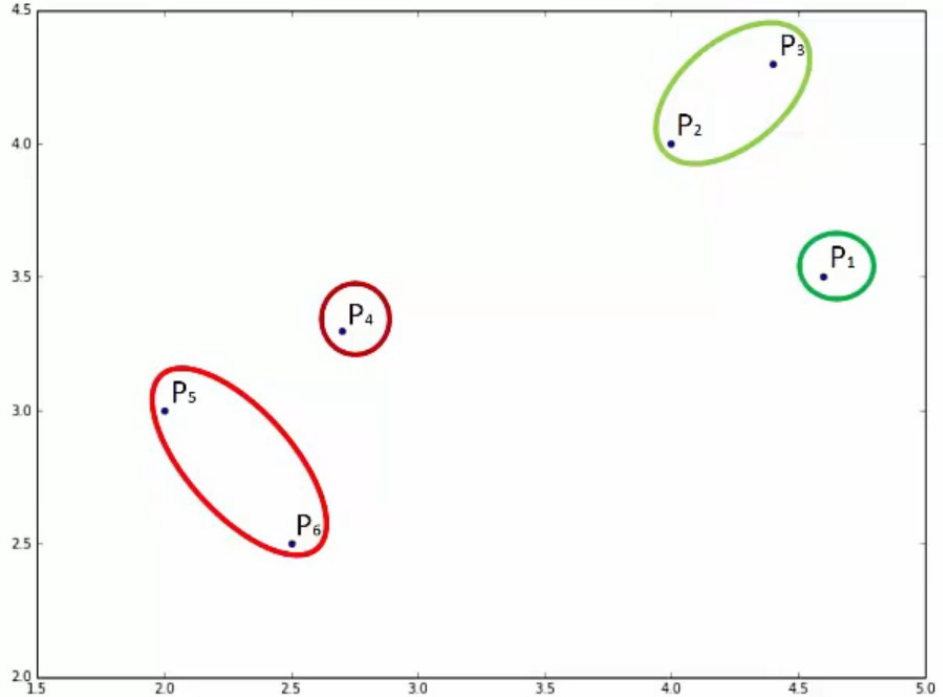


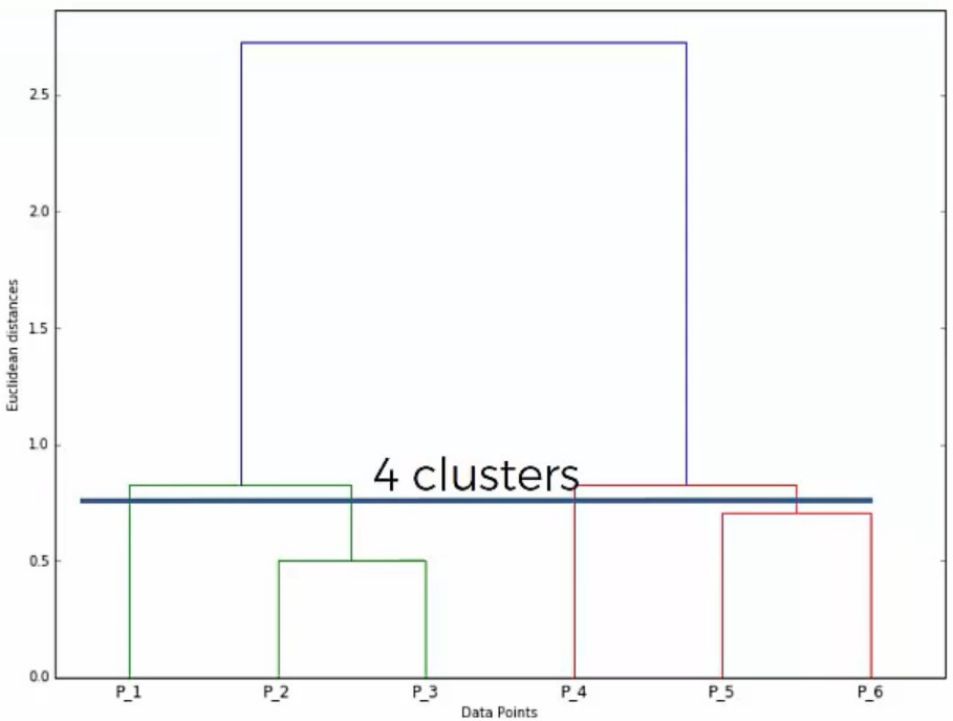
HIERARCHICAL CLUSTERING



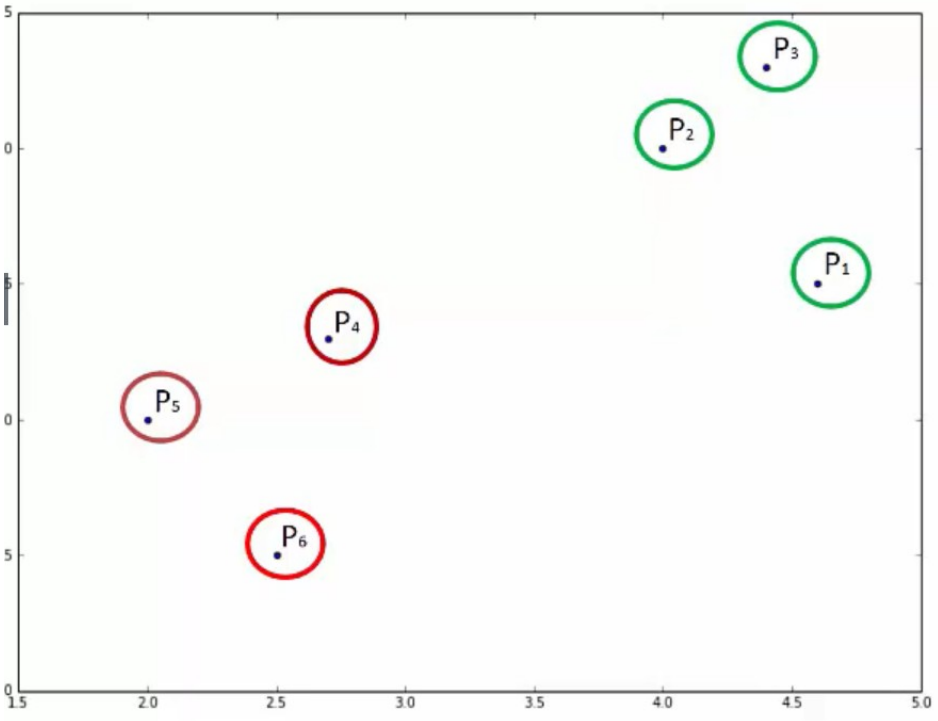


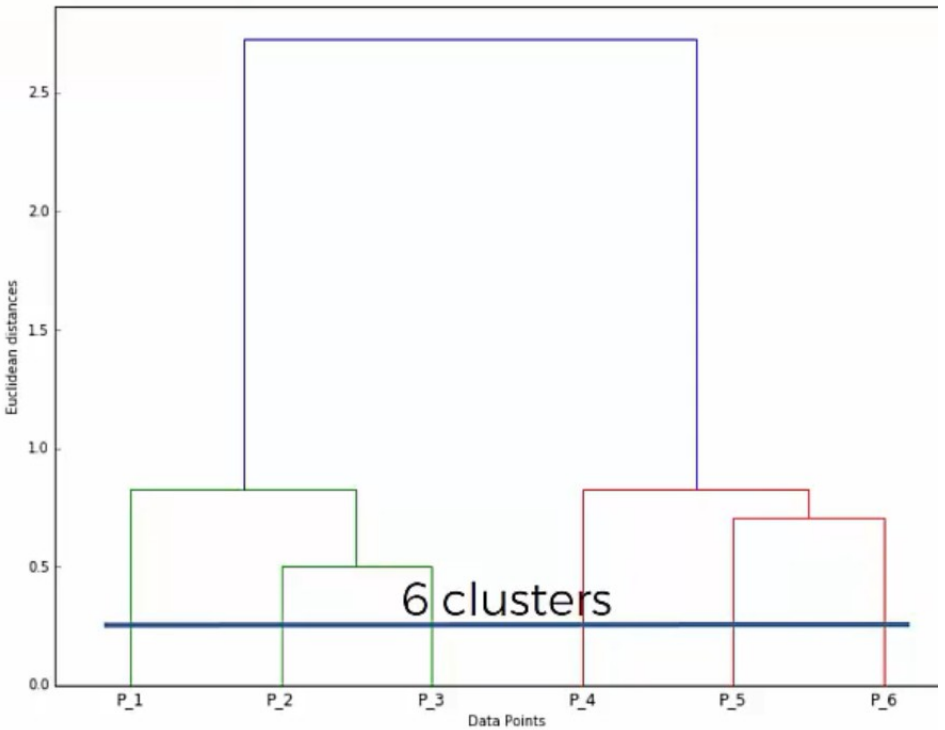
HIERARCHICAL CLUSTERING





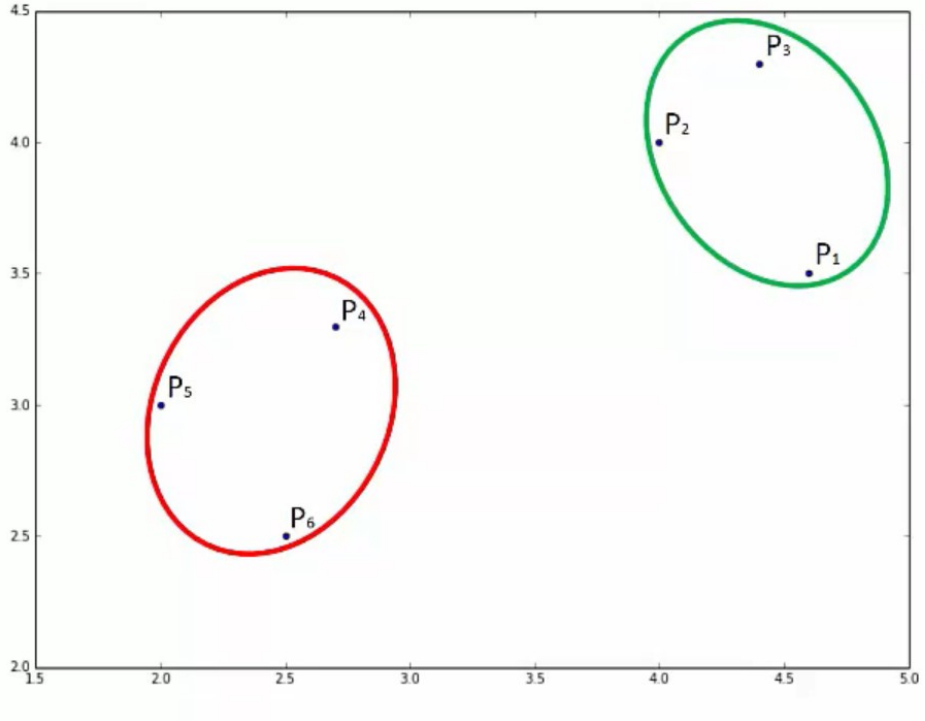
HIERARCHICAL CLUSTERING

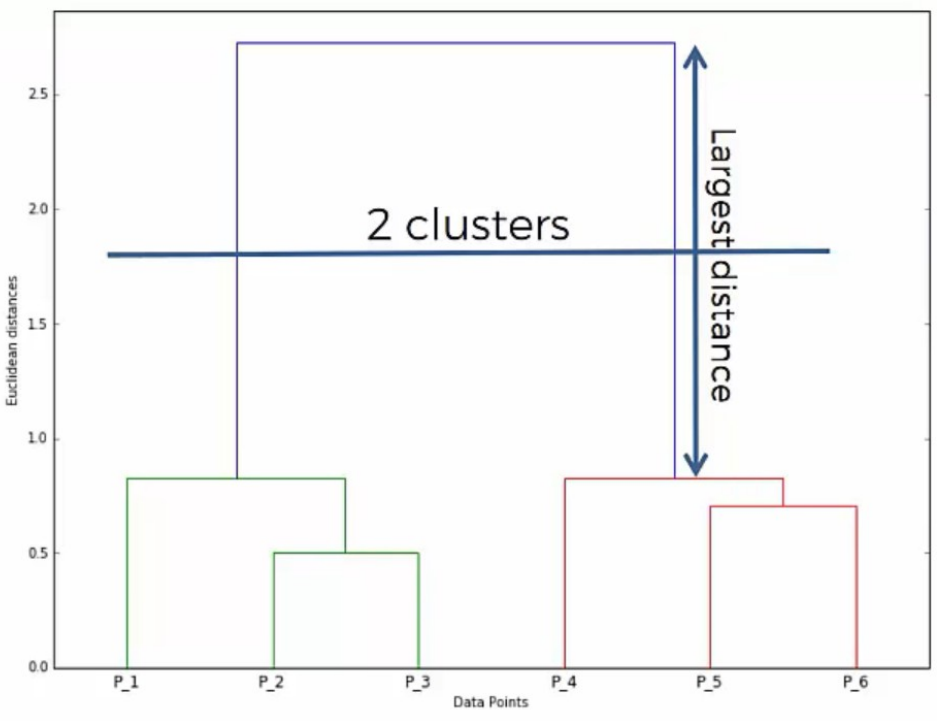




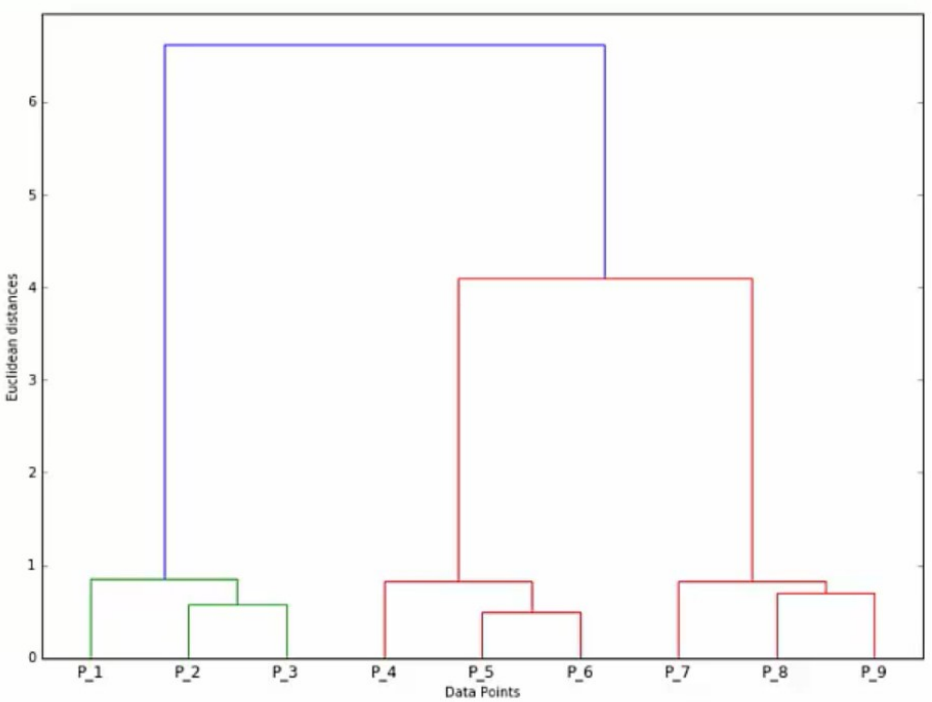
HIERARCHICAL CLUSTERING

OPTIMAL NUMBER OF CLUSTERS

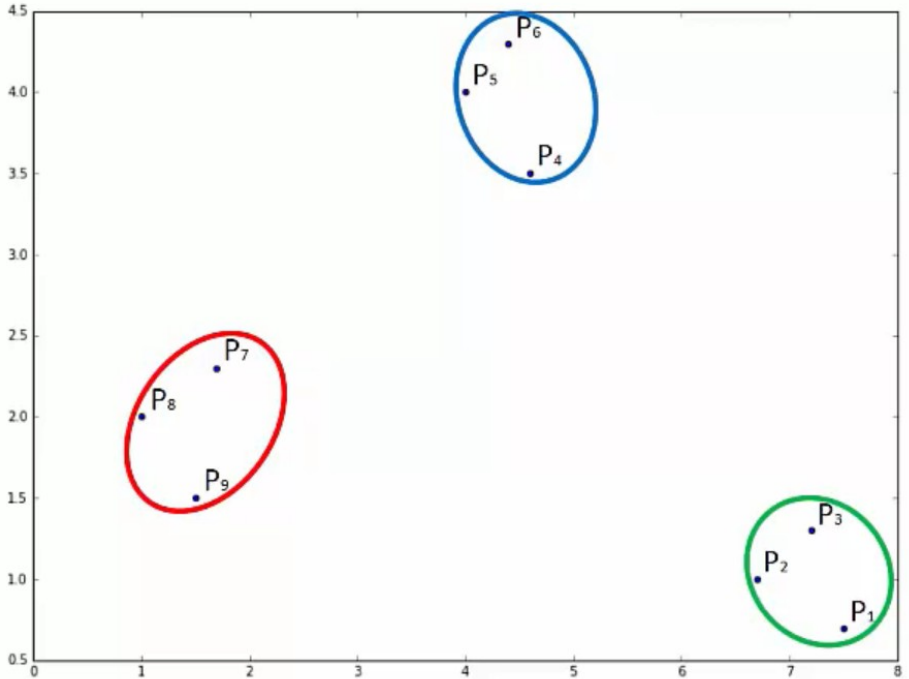


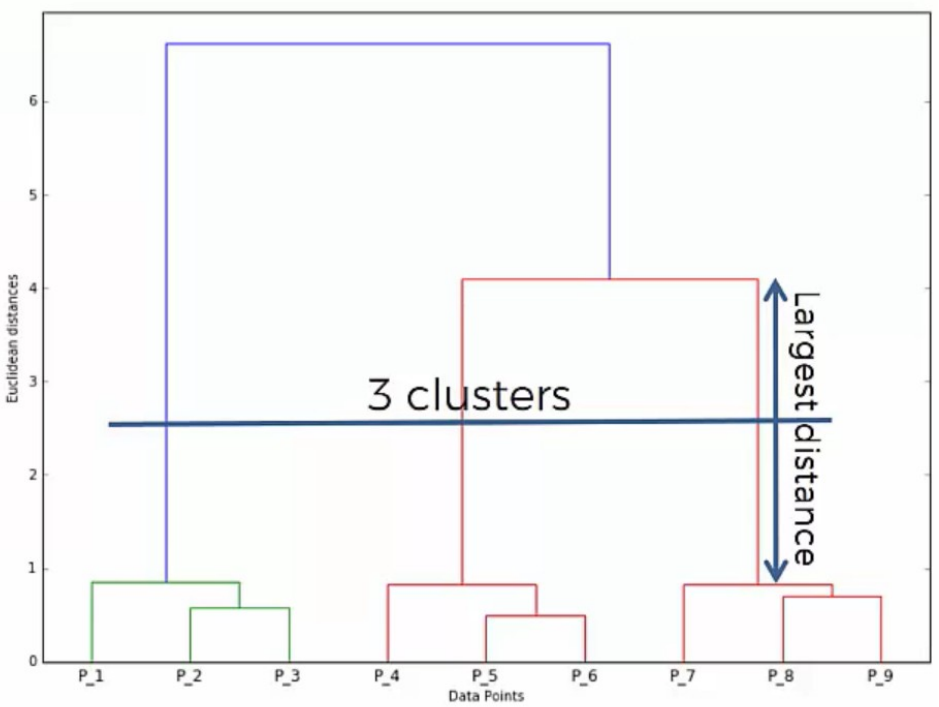


HIERARCHICAL CLUSTERING



HIERARCHICAL CLUSTERING





HIERARCHICAL CLUSTERING

Clustering

Clustering Model	Pros	Cons
K-Means	Simple to understand, easily adaptable, works well on small or large datasets, fast, efficient and performant	Need to choose the number of clusters
Hierarchical Clustering	The optimal number of clusters can be obtained by the model itself, practical visualisation with the dendrogram	Not appropriate for large datasets