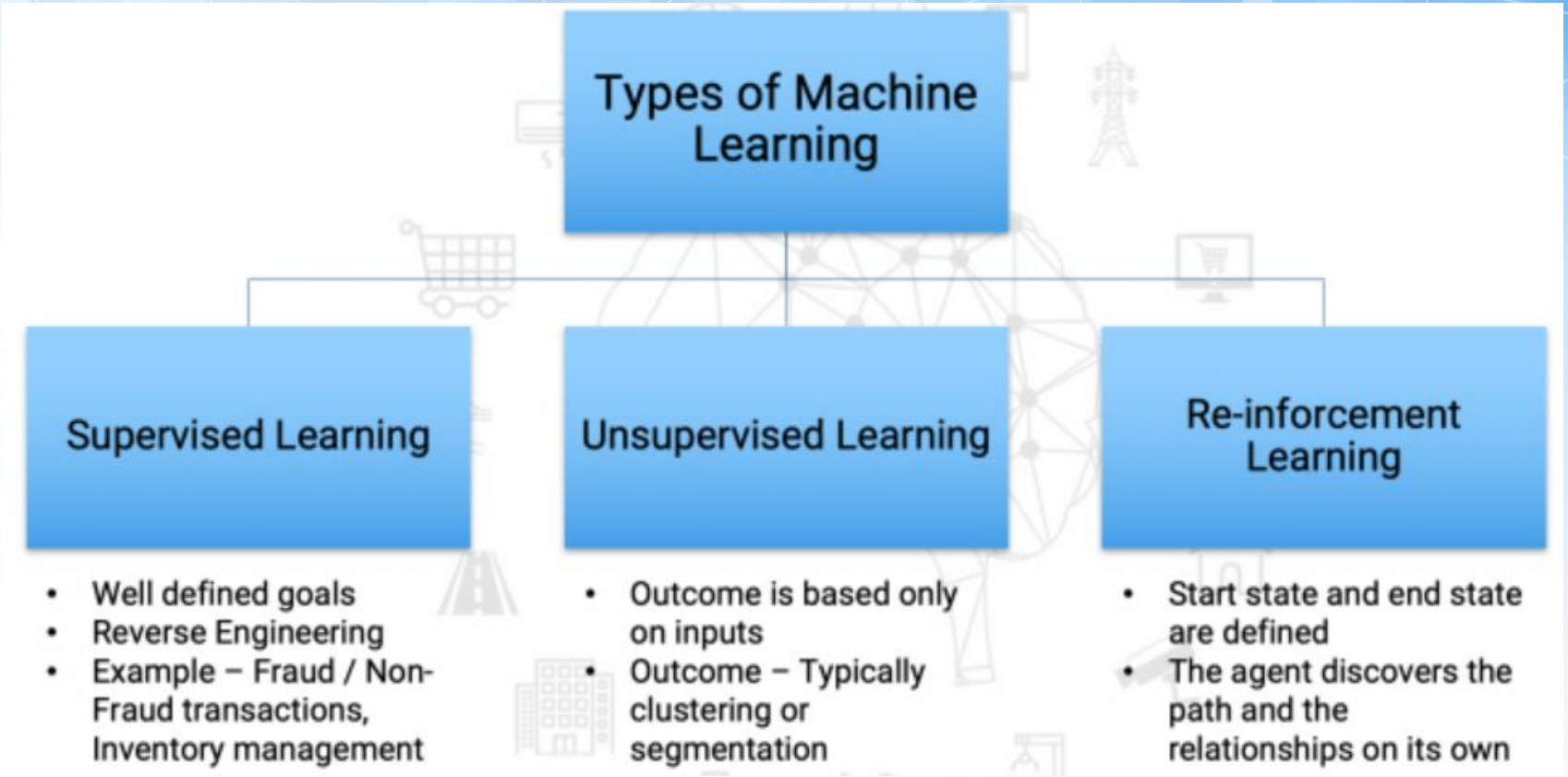
ICT Academy of Kerala

Building the Nation's Future

Unsupervised learning & K-Means Clustering

A **GOVT. OF INDIA** SUPPORTED, **GOVT. OF KERALA** PARTNERED SOCIAL ENTERPRISE.



Supervised vs Unsupervised

Supervised Learning

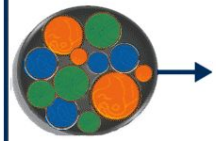
- Known number of classes
- Based on training set
- Used to classify future observations

Unsupervised Learning

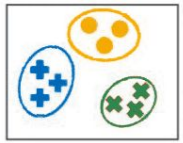
- Unknown number of classes
- No prior knowledge
- Used to understand data

UNSUPERVISED LEARNING

Raw Data



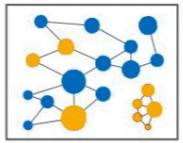
EXAMPLES



CLUSTERING

Identifies similarities in groups:

Are there patterns in the data that indicate which groups to target?



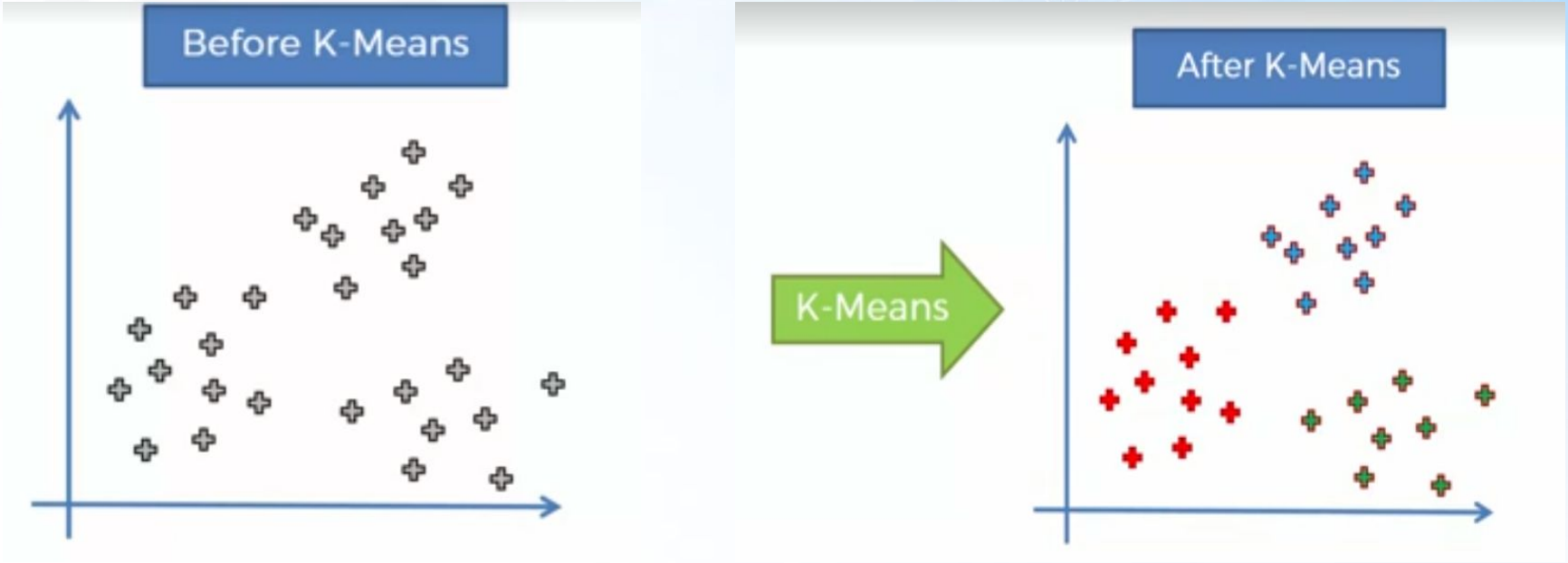
ANOMALY DETECTION

Identifies abnormalities in dataset:

Is the user behaving as it should? Is a hacker intruding the network?

K-Means Clustering

What K-Means does for you?



How K-Means works?

Step 1: Choose the Number of K-Clusters

Step 2: Select at Random K-Points, the centroids (Not necessarily from your dataset.)

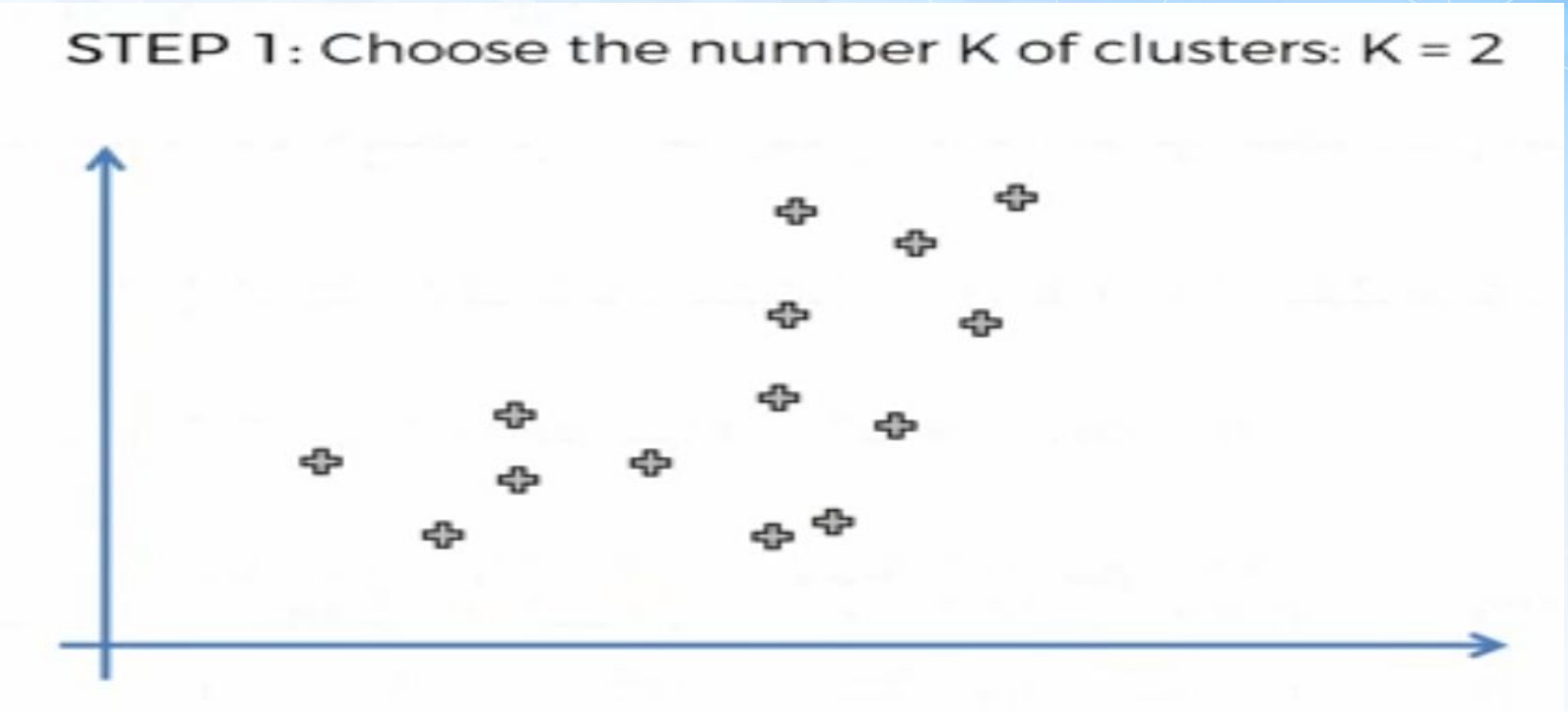
Step 3: Assign each data point to the closest centroid (That forms K-Clusters.)

Step 4: Compute and place the new Centroid of each other

Step 5: Reassign each data point to the new closest Centroid.
(If any reassignment took place, go to Step 4, otherwise FINISH)

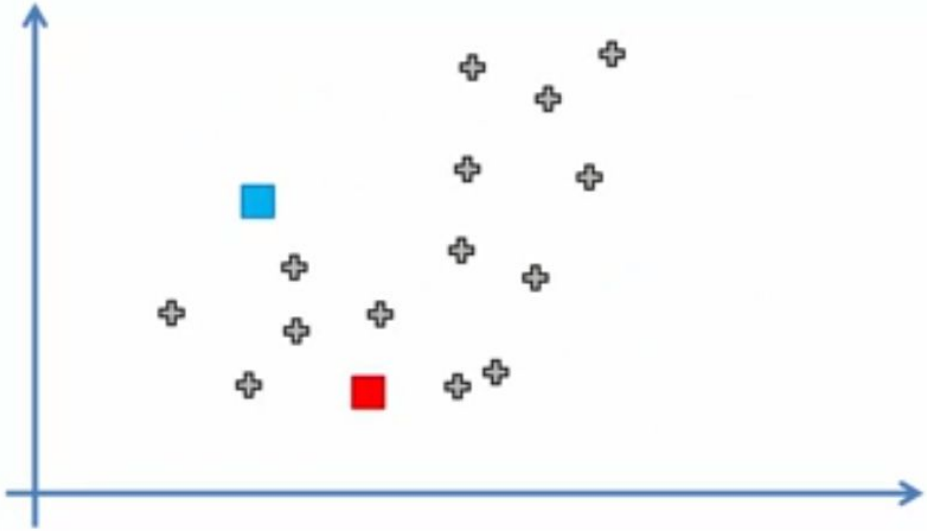
Model is READY!

K-Means Clustering



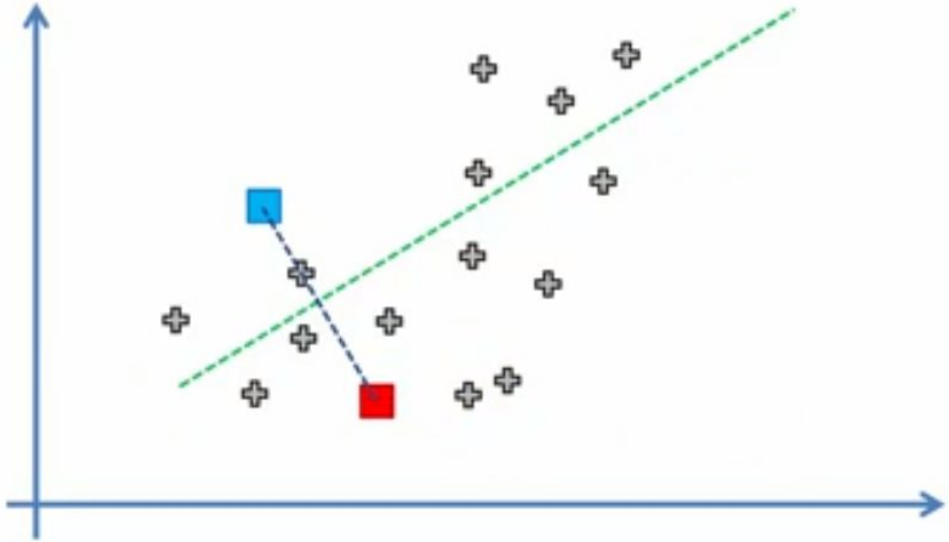
K-Means Clustering

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

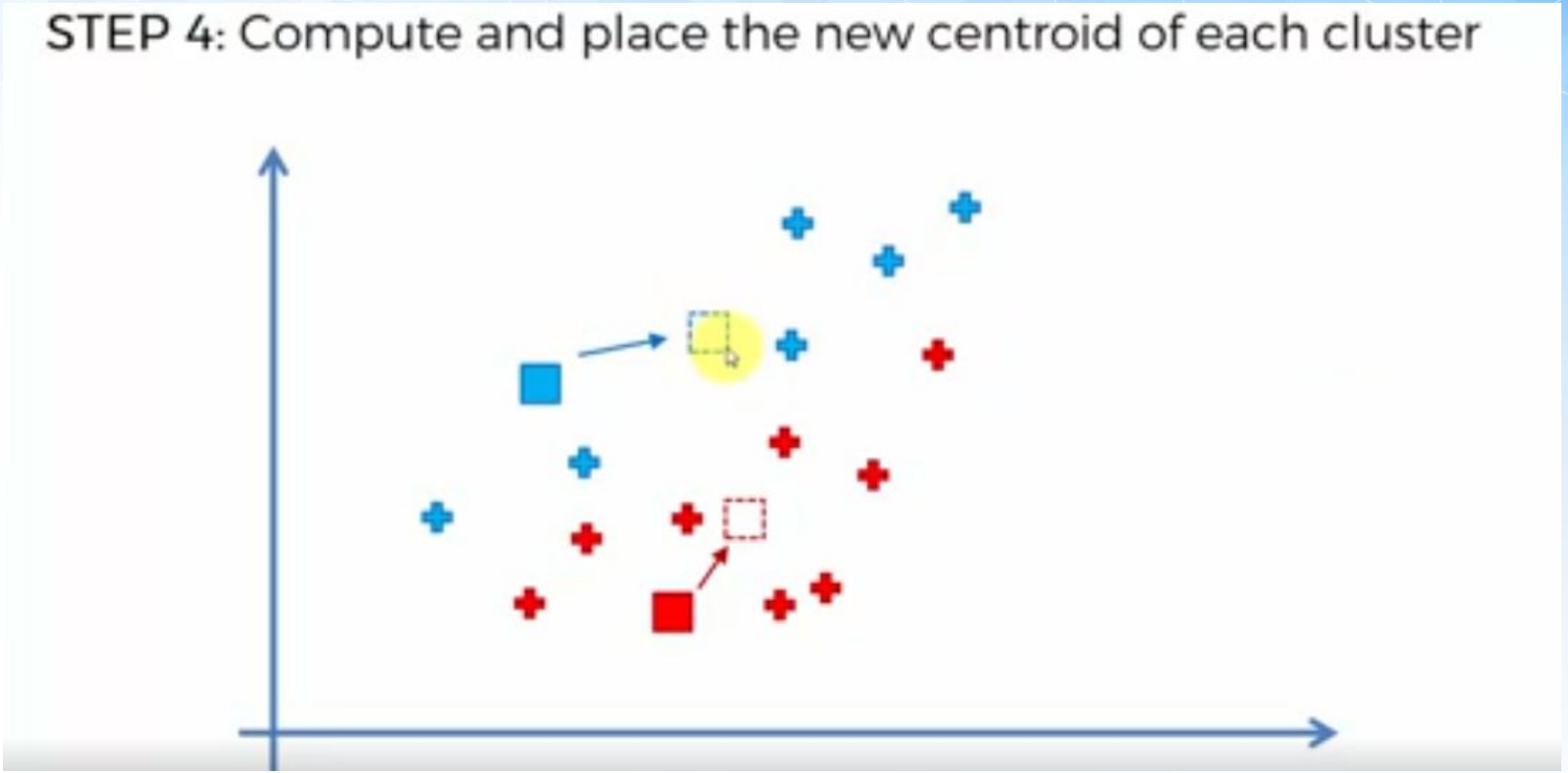


K-Means Clustering

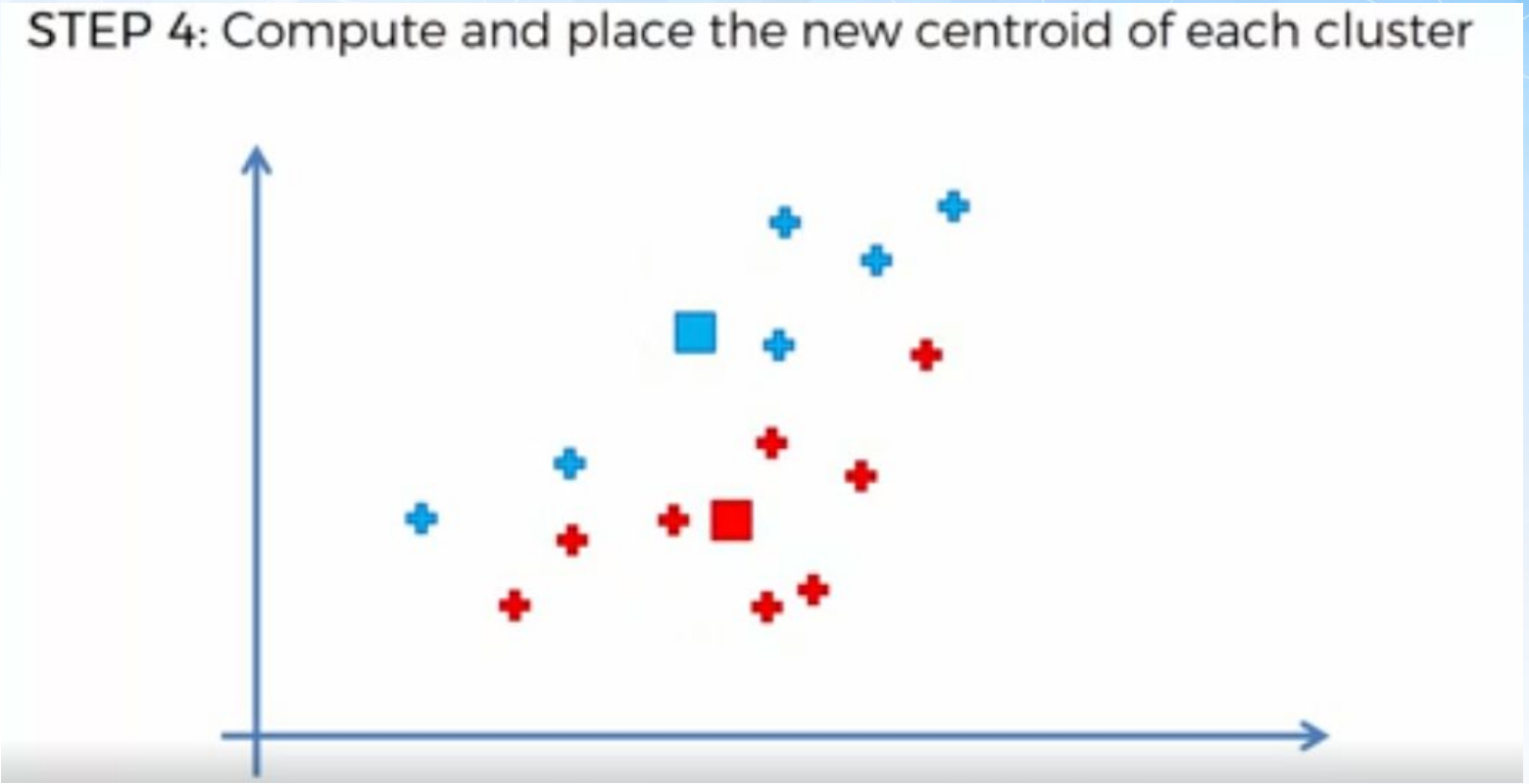
STEP 3: Assign each data point to the closest centroid ➡ That forms K clusters



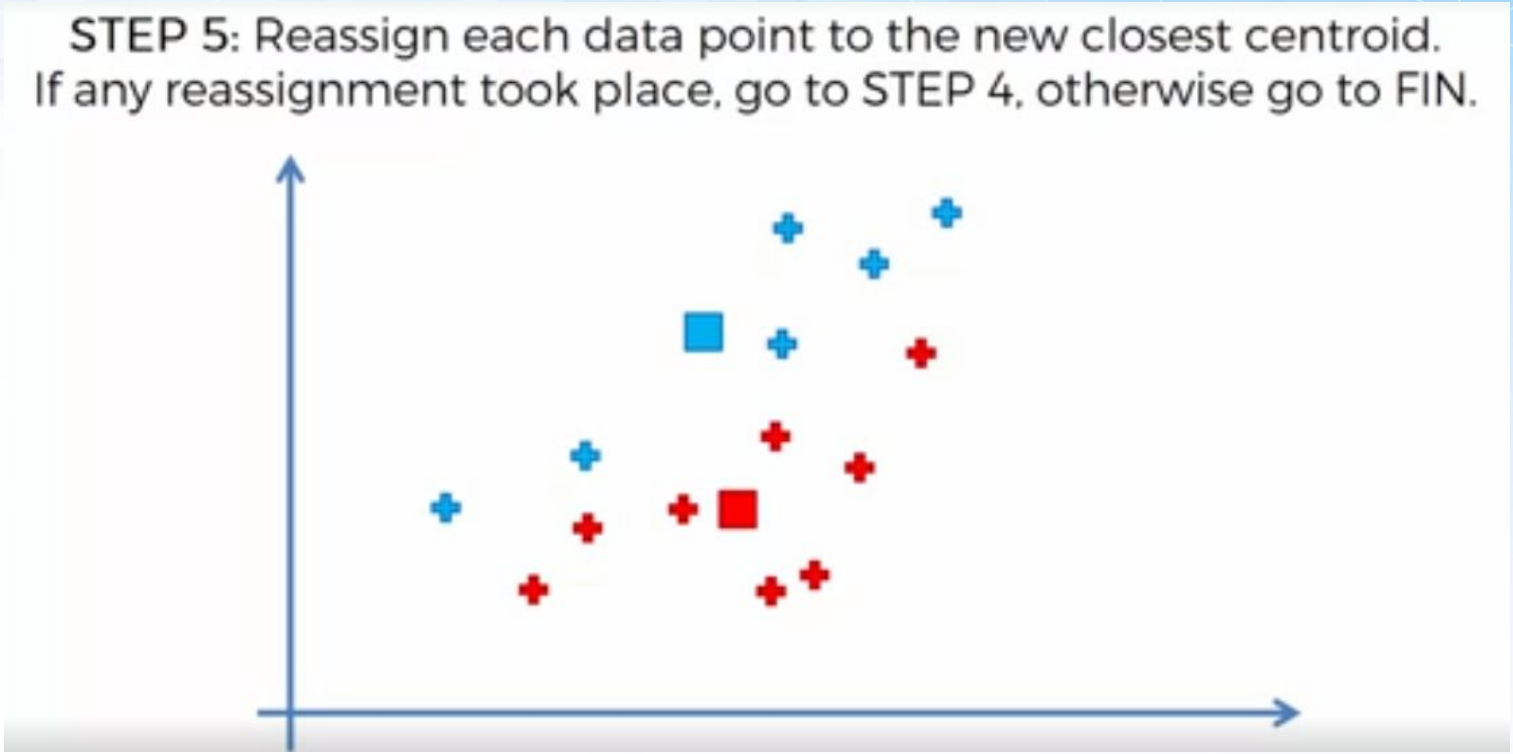
K-Means Clustering



K-Means Clustering

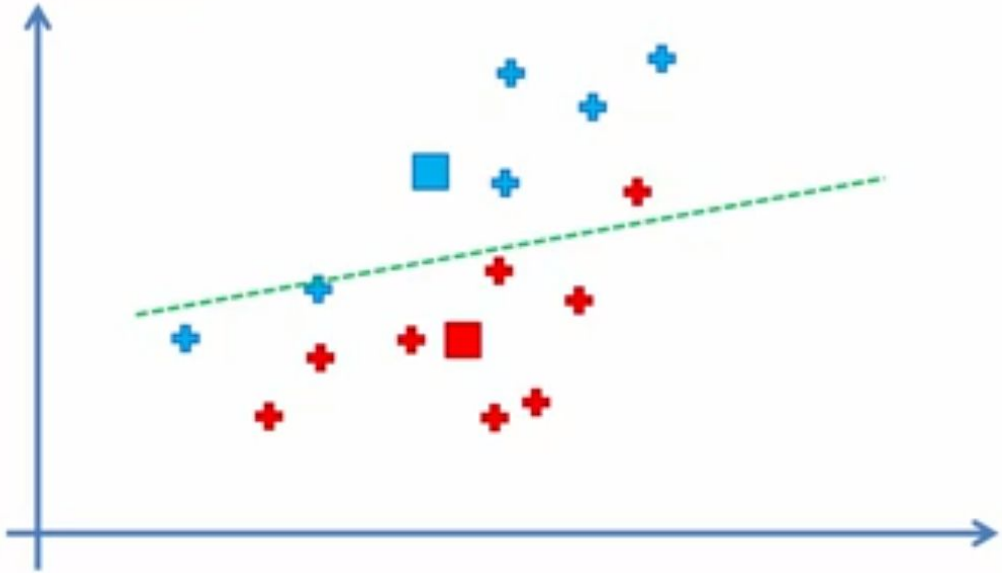


K-Means Clustering



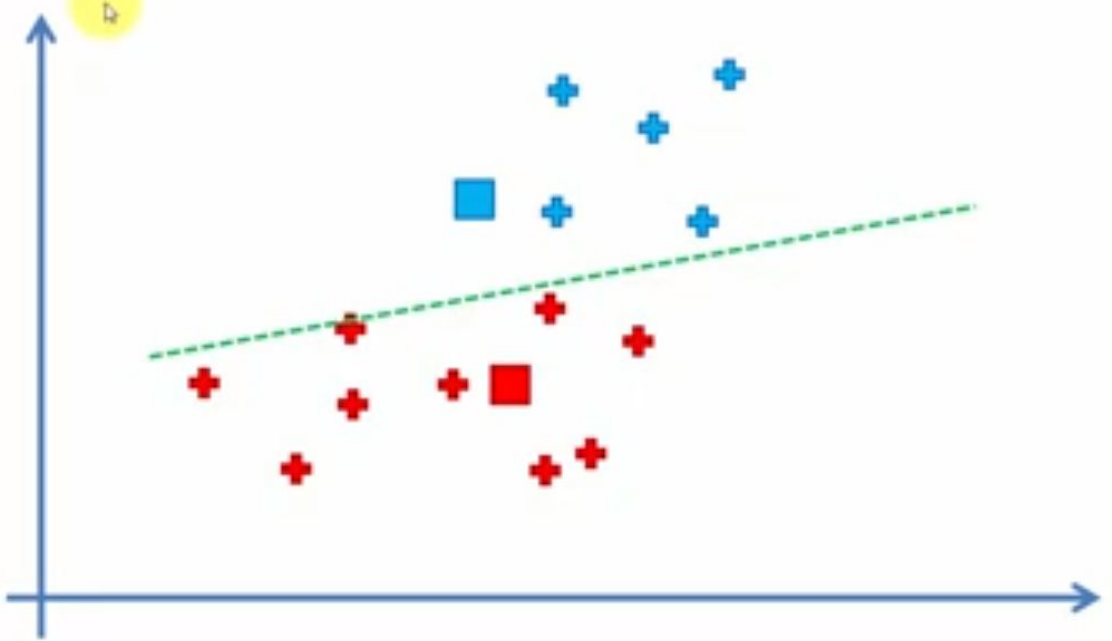
K-Means Clustering

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



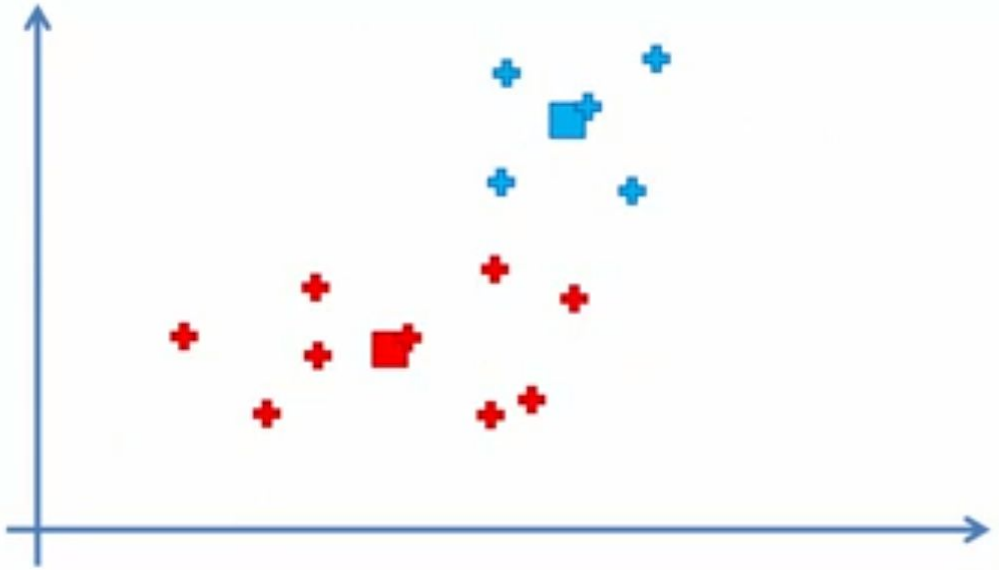
K-Means Clustering

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



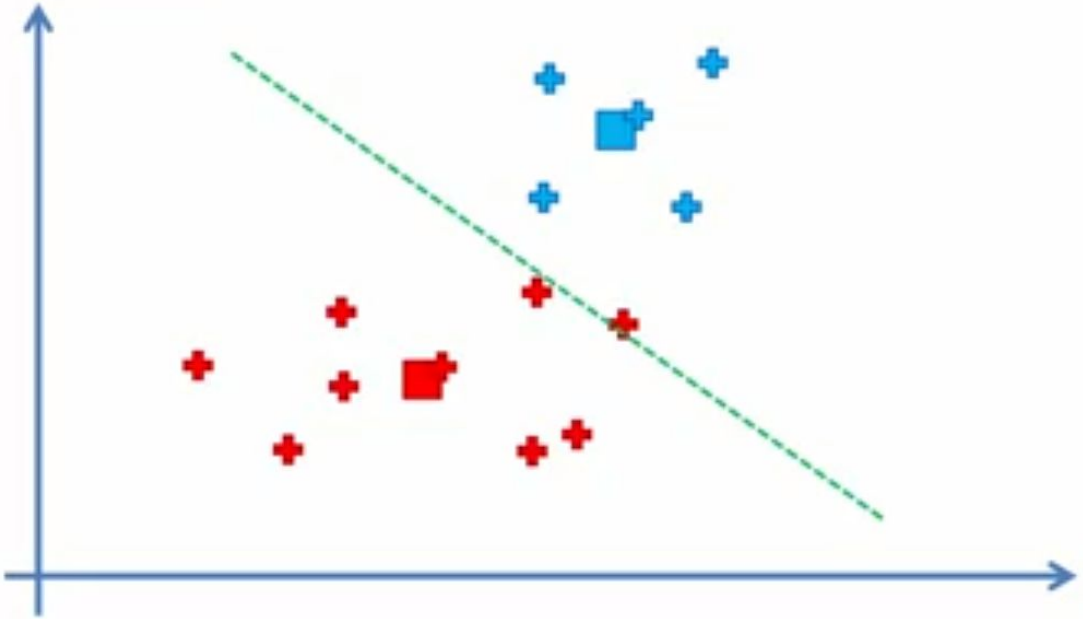
K-Means Clustering

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



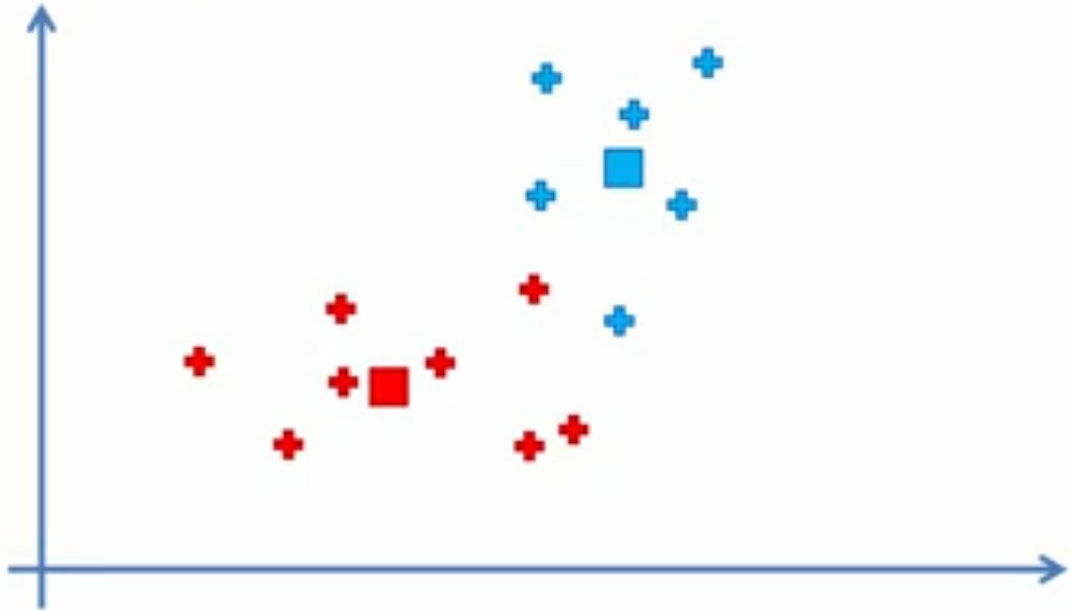
K-Means Clustering

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



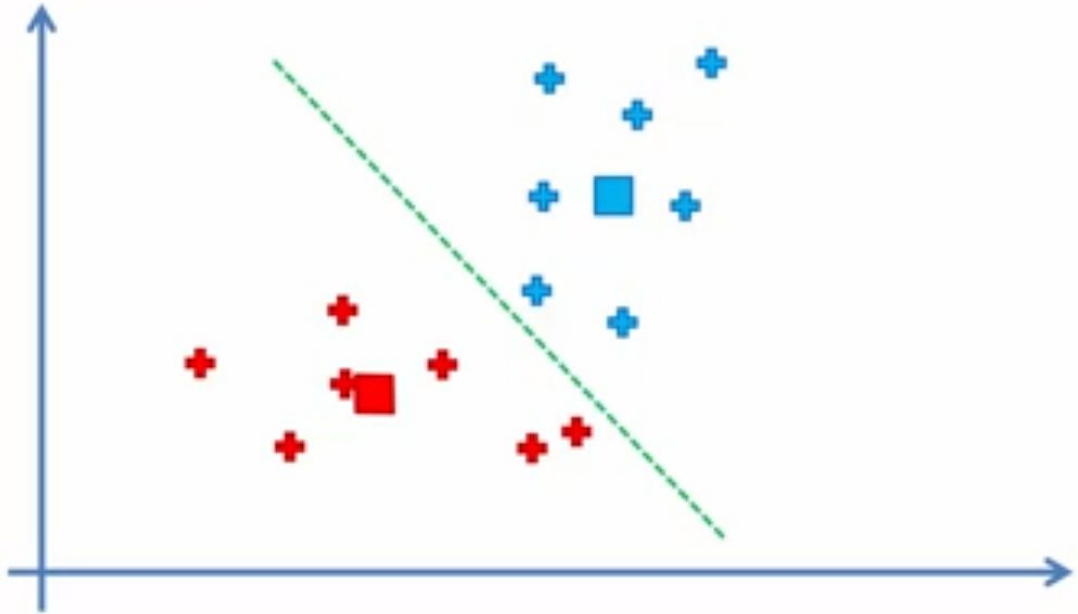
K-Means Clustering

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

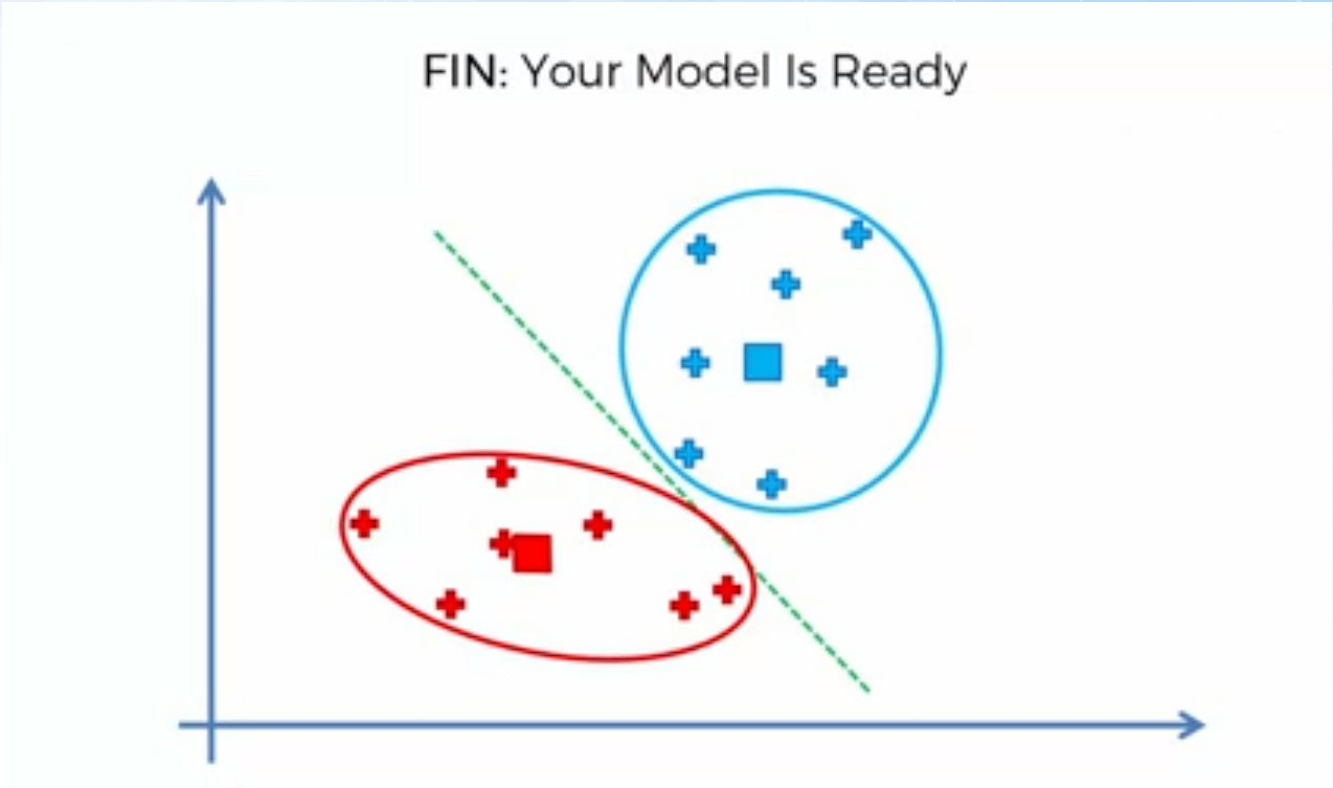


K-Means Clustering

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

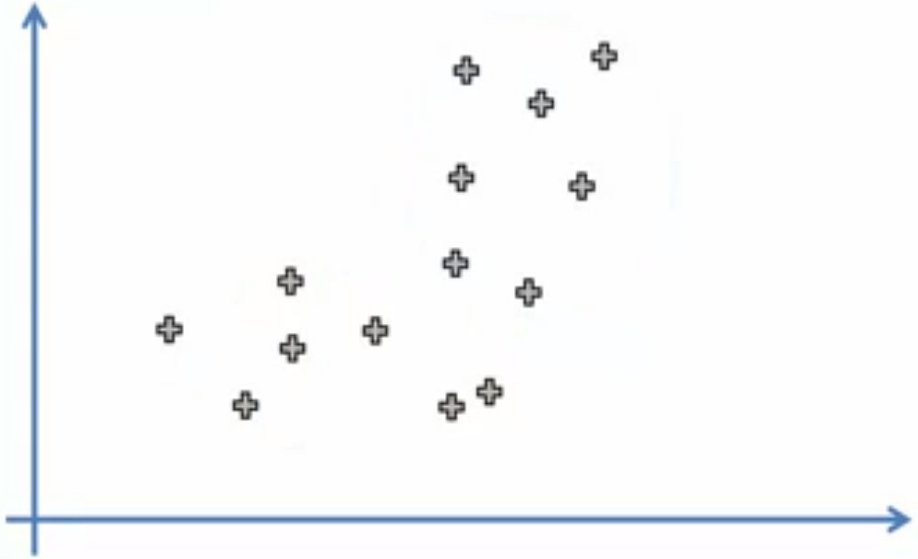


K-Means Clustering

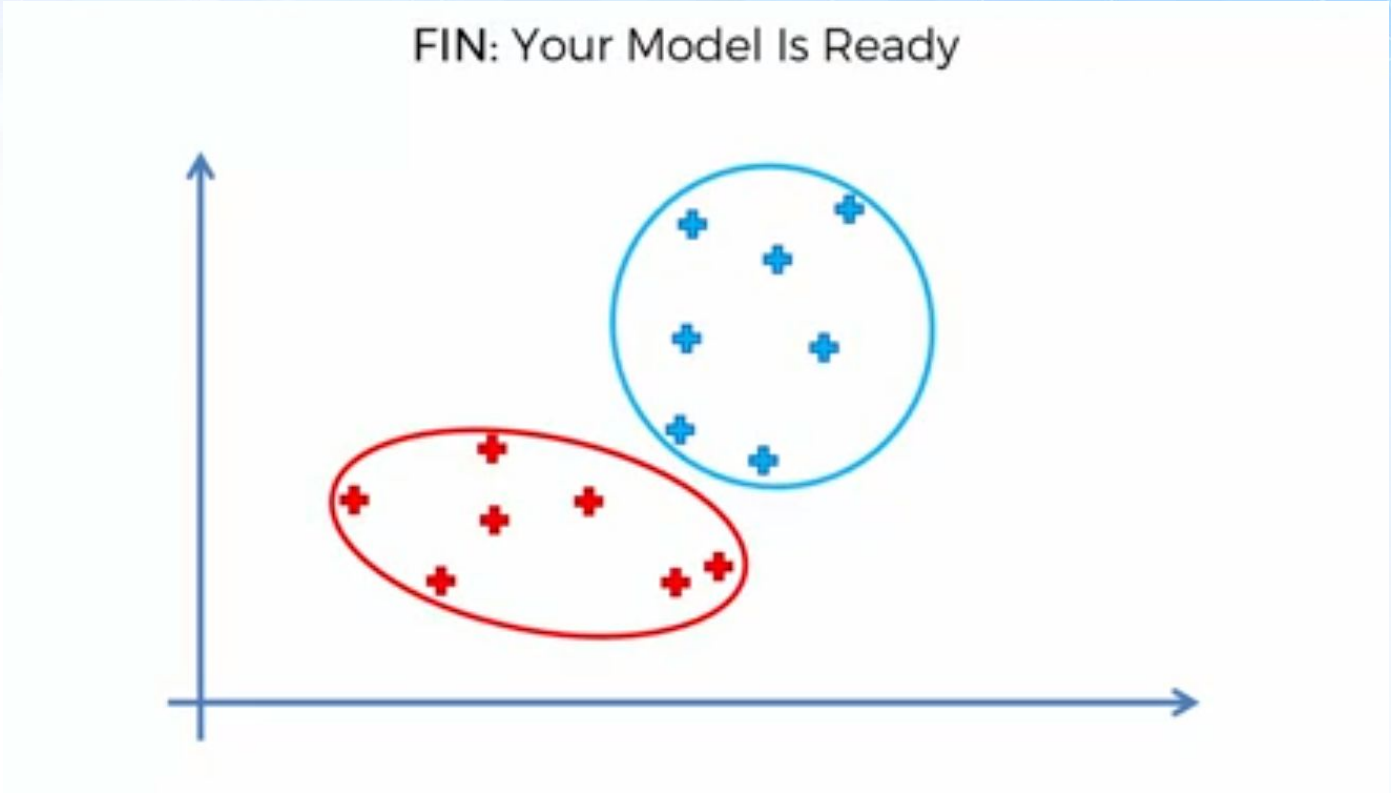


K-Means Clustering

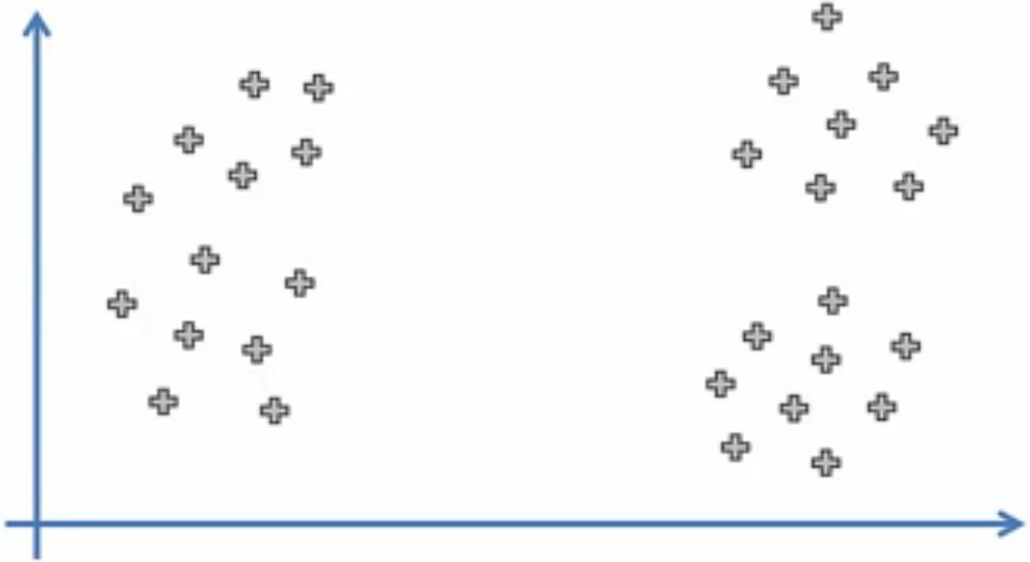
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)



K-Means Clustering

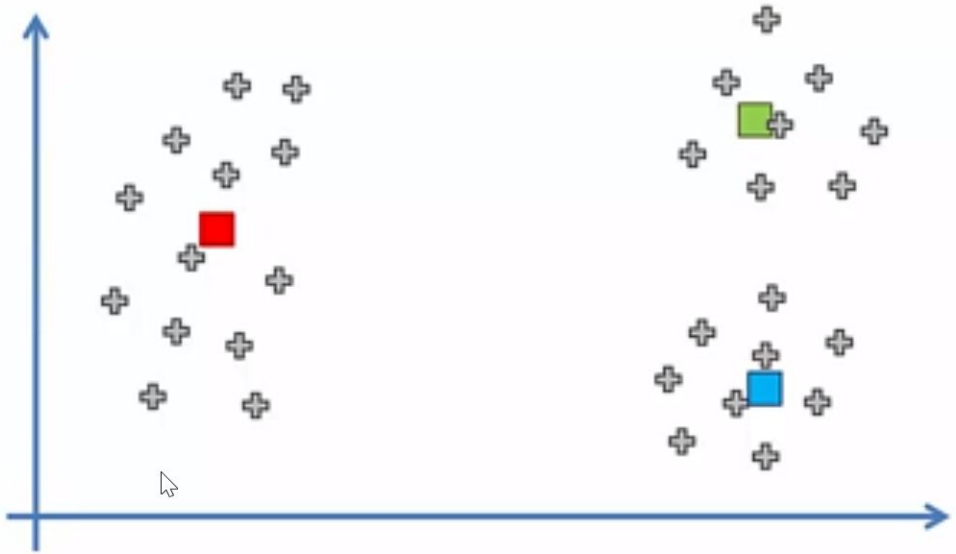


Random Initialization Trap



If we choose K = 3 clusters...

Random Initialization Trap



...this correct random initialisation would lead us to...

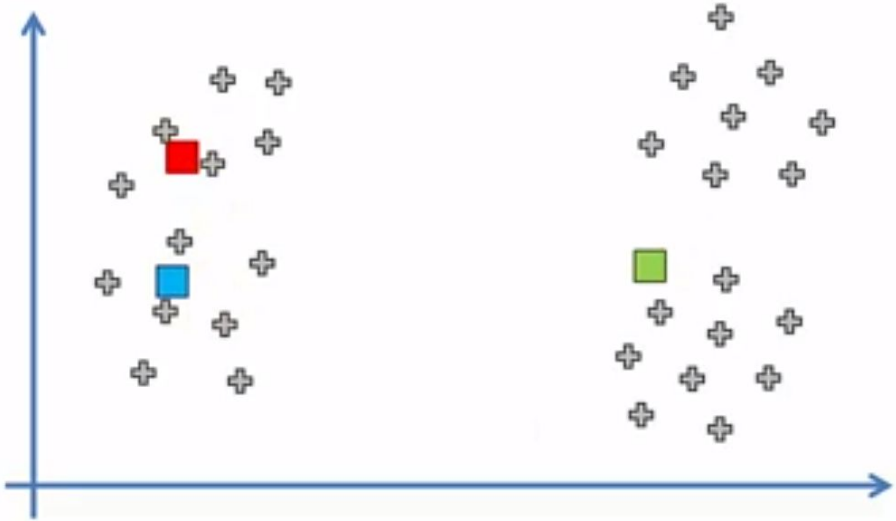
Random Initialization Trap

Random Initialization Trap

But what would happen if we had a bad random initialization?

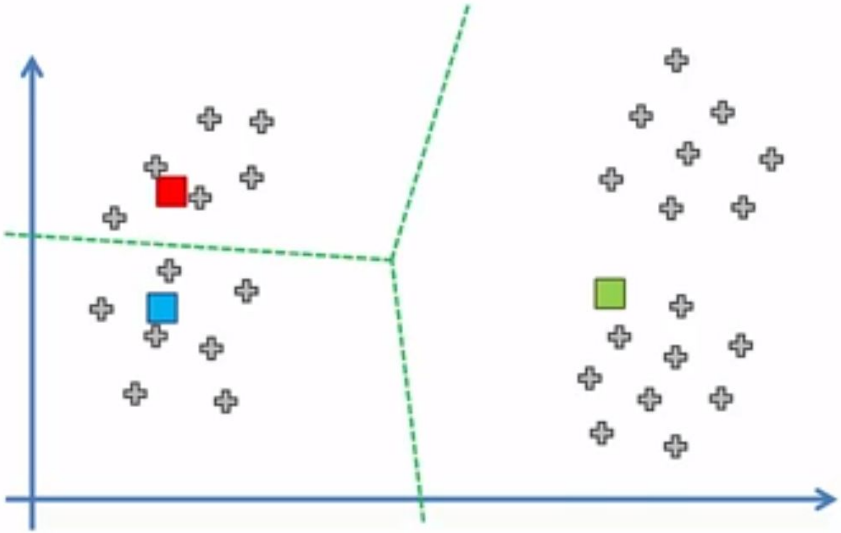
Random Initialization Trap

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



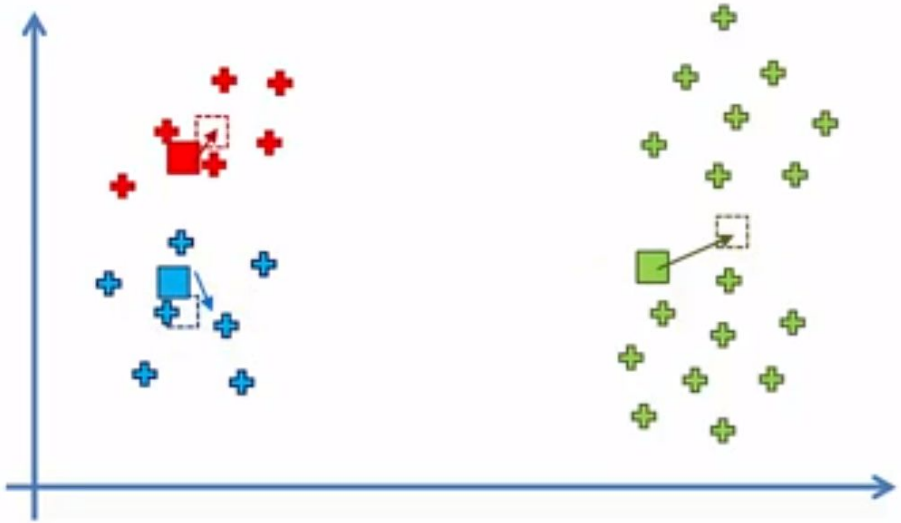
Random Initialization Trap

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



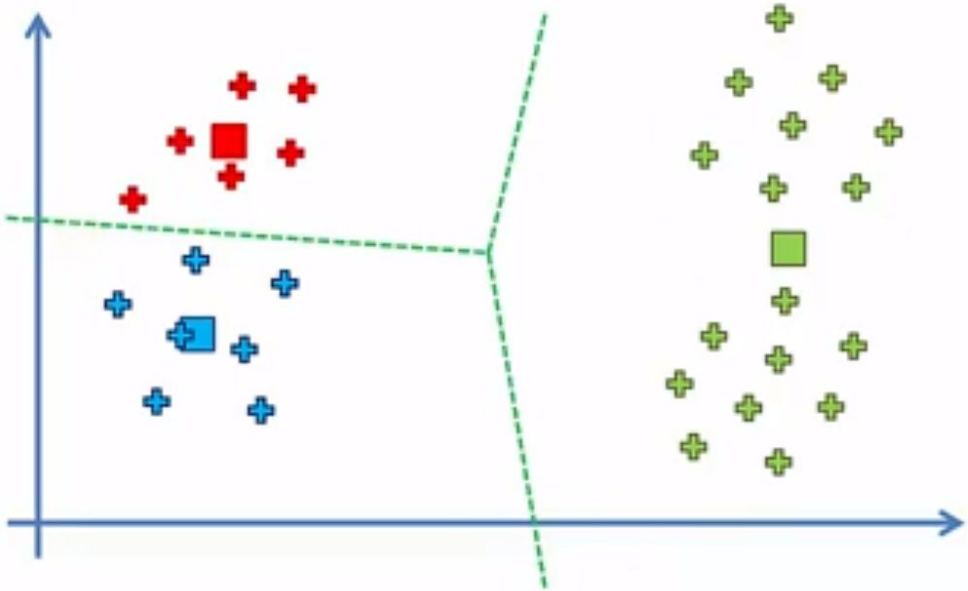
Random Initialization Trap

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



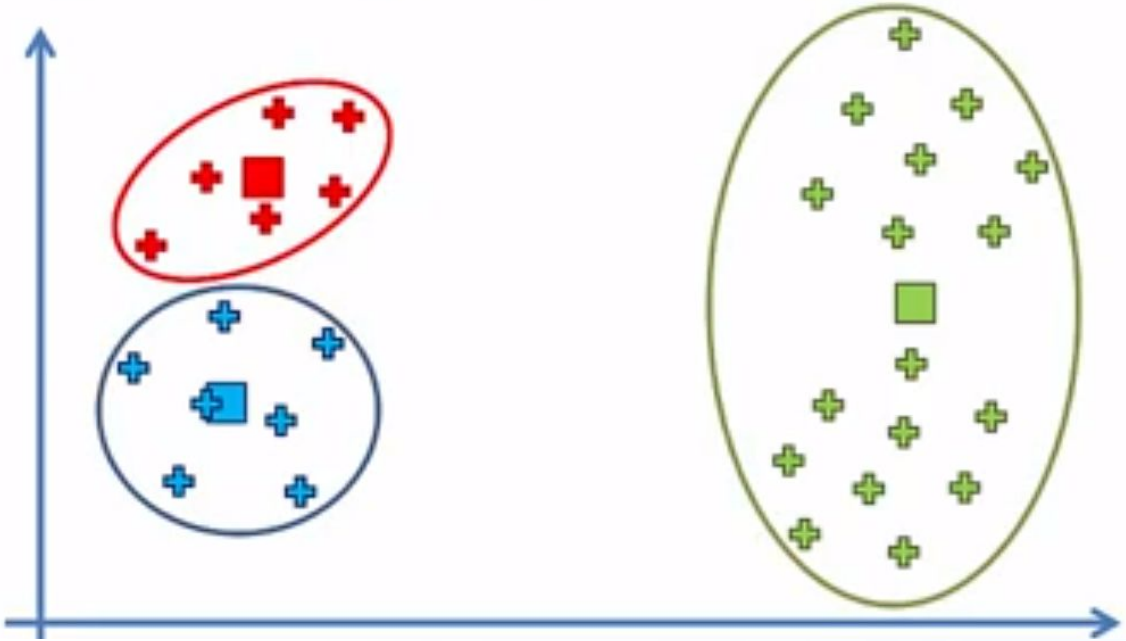
Random Initialization Trap

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

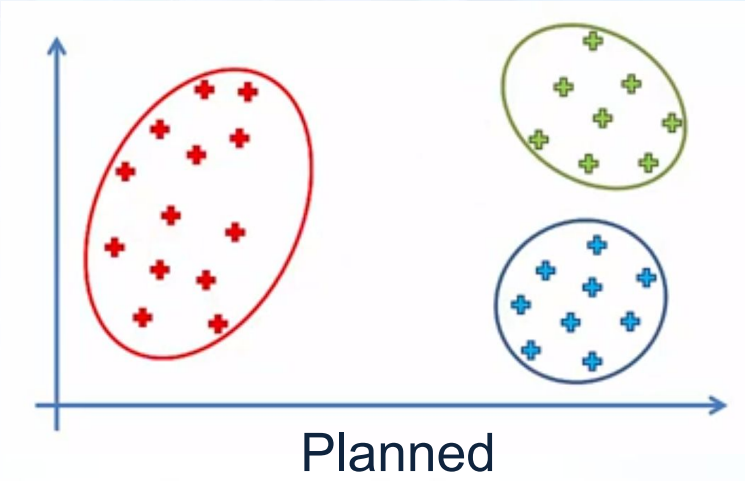


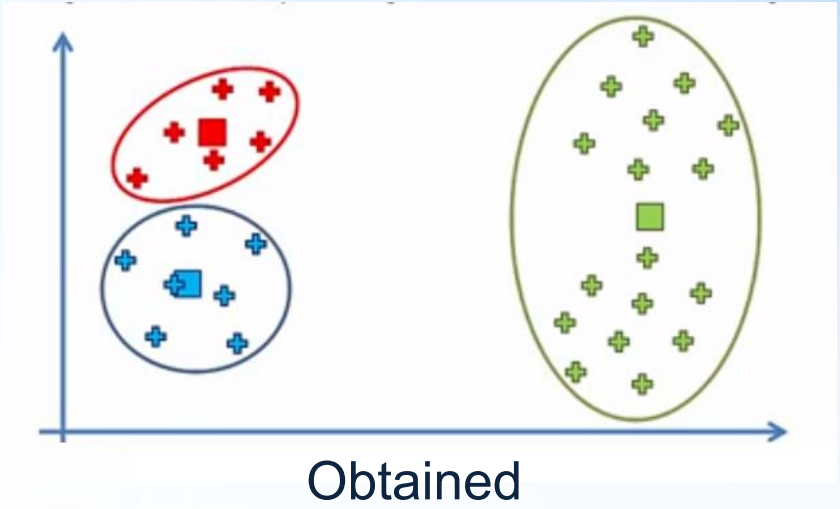
Random Initialization Trap

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



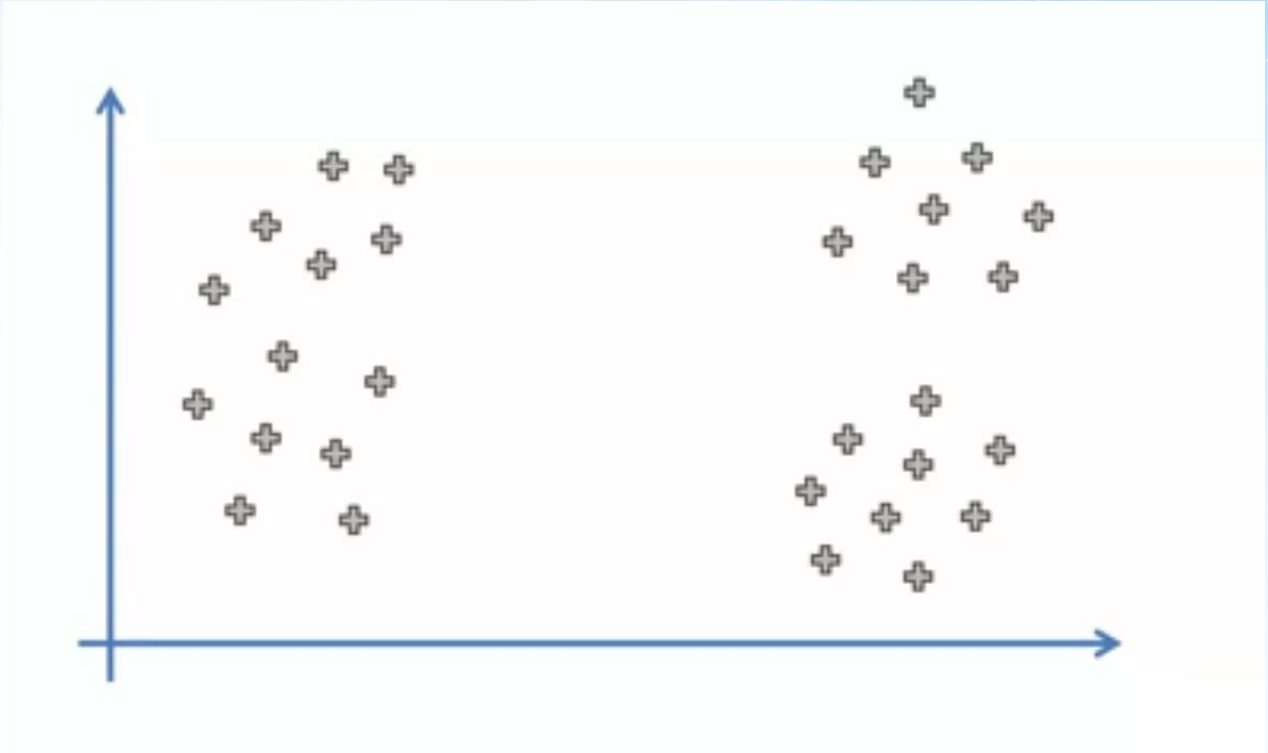
Random Initialization Trap



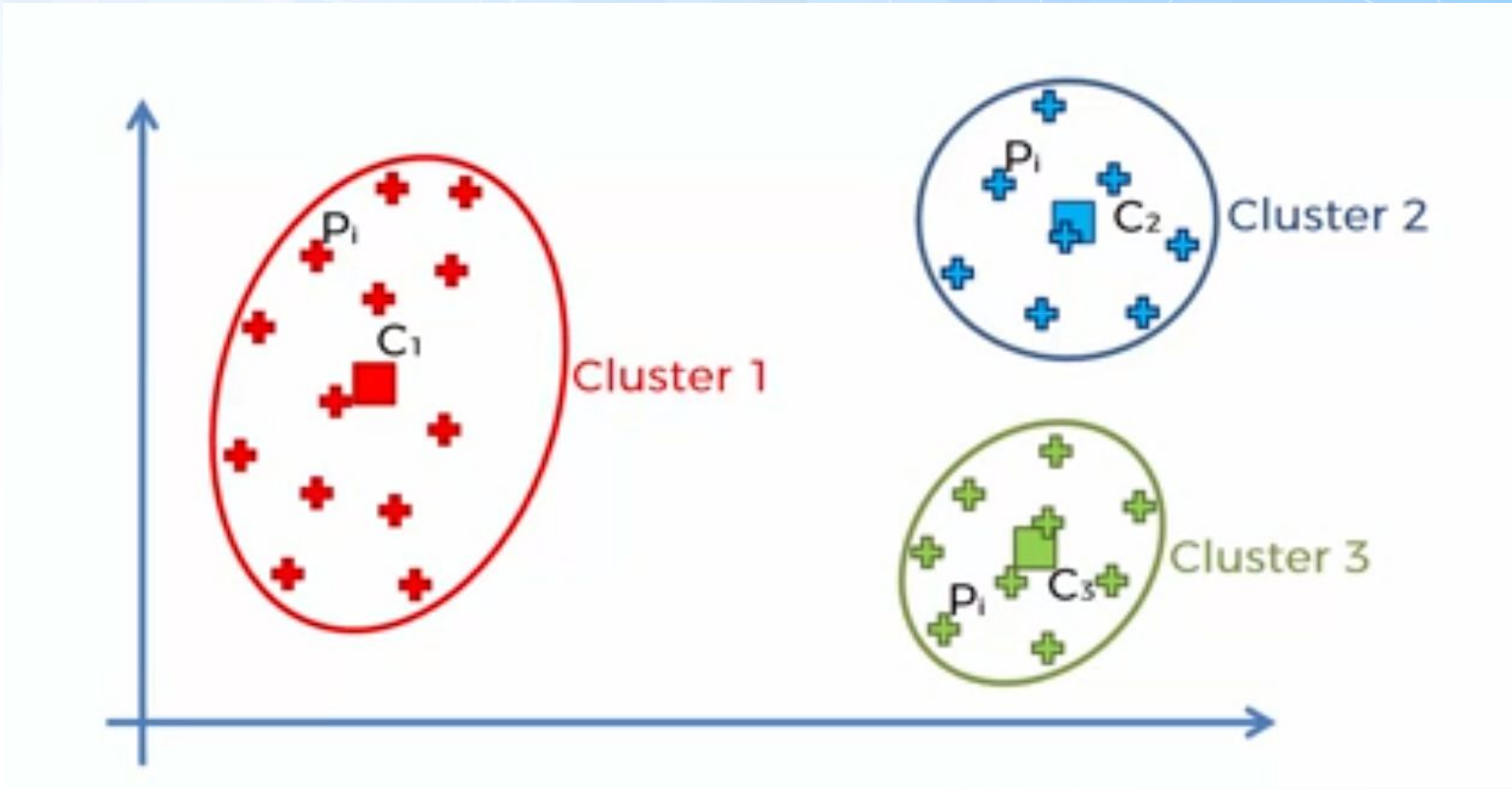


Random Initialization Trap

Choosing Right Number of Clusters



Choosing Right Number of Clusters

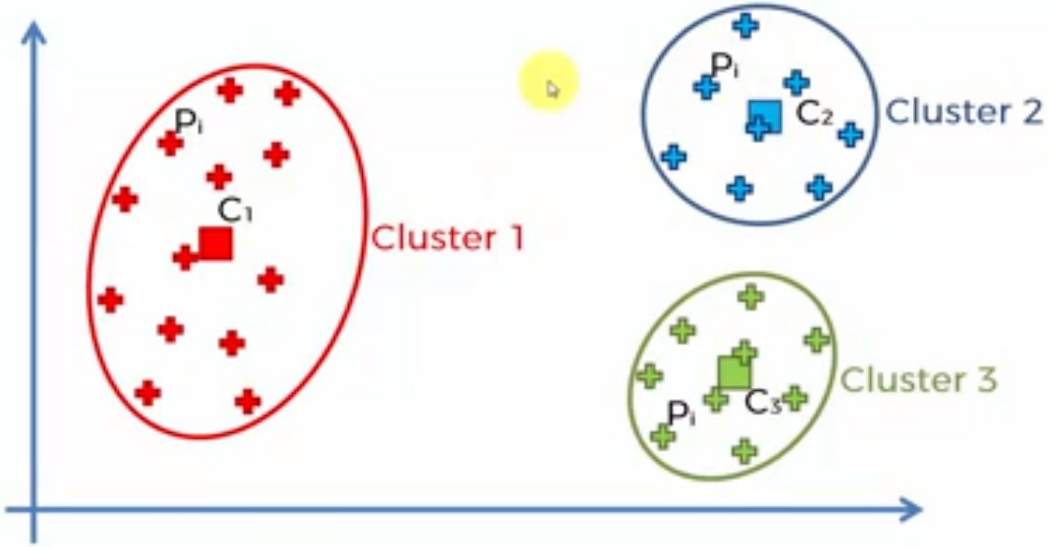


Choosing Right Number of Clusters

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

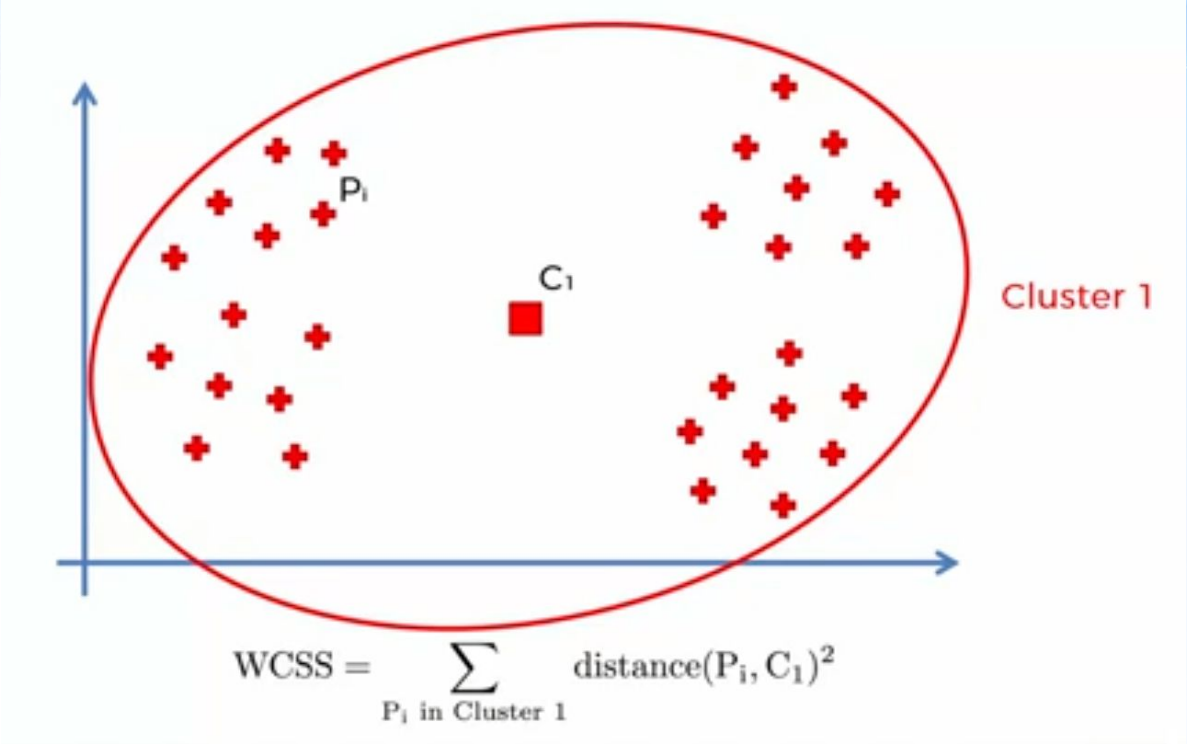
*** within-cluster sums of squares**

Choosing Right Number of Clusters

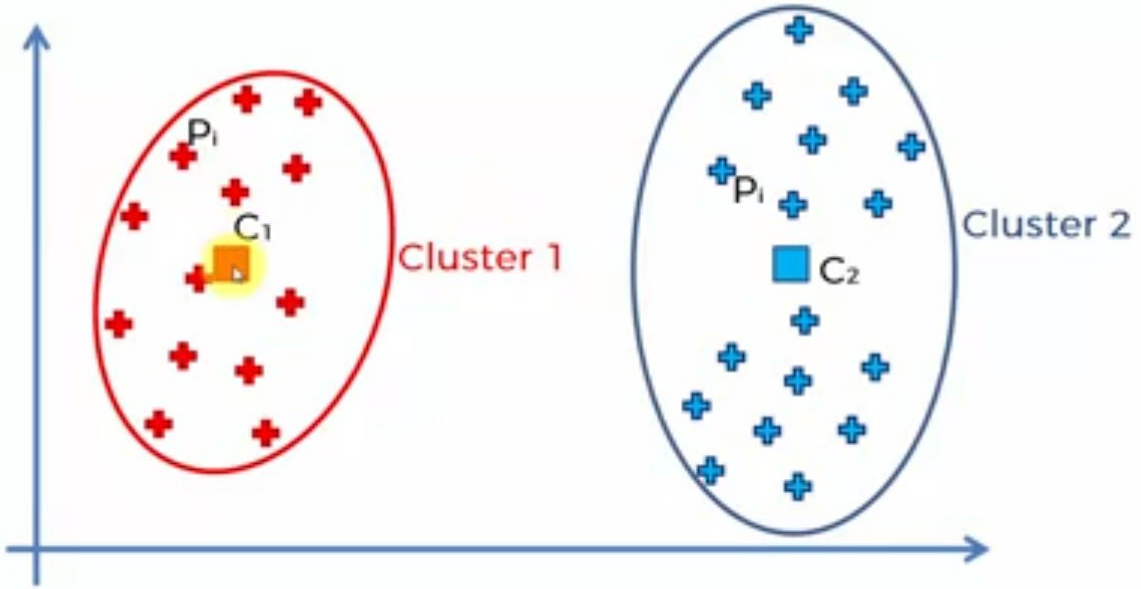


$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Choosing Right Number of Clusters

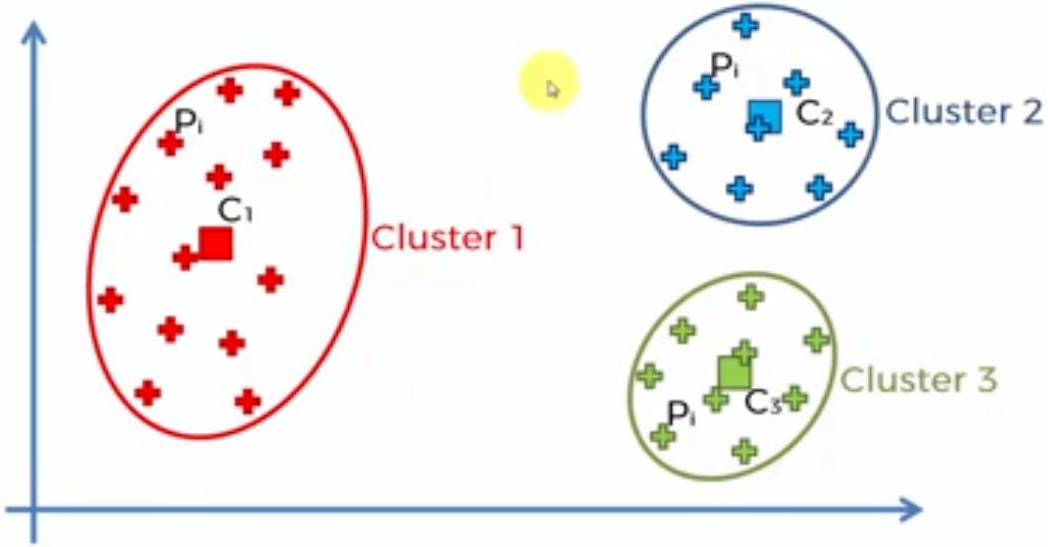


Choosing Right Number of Clusters



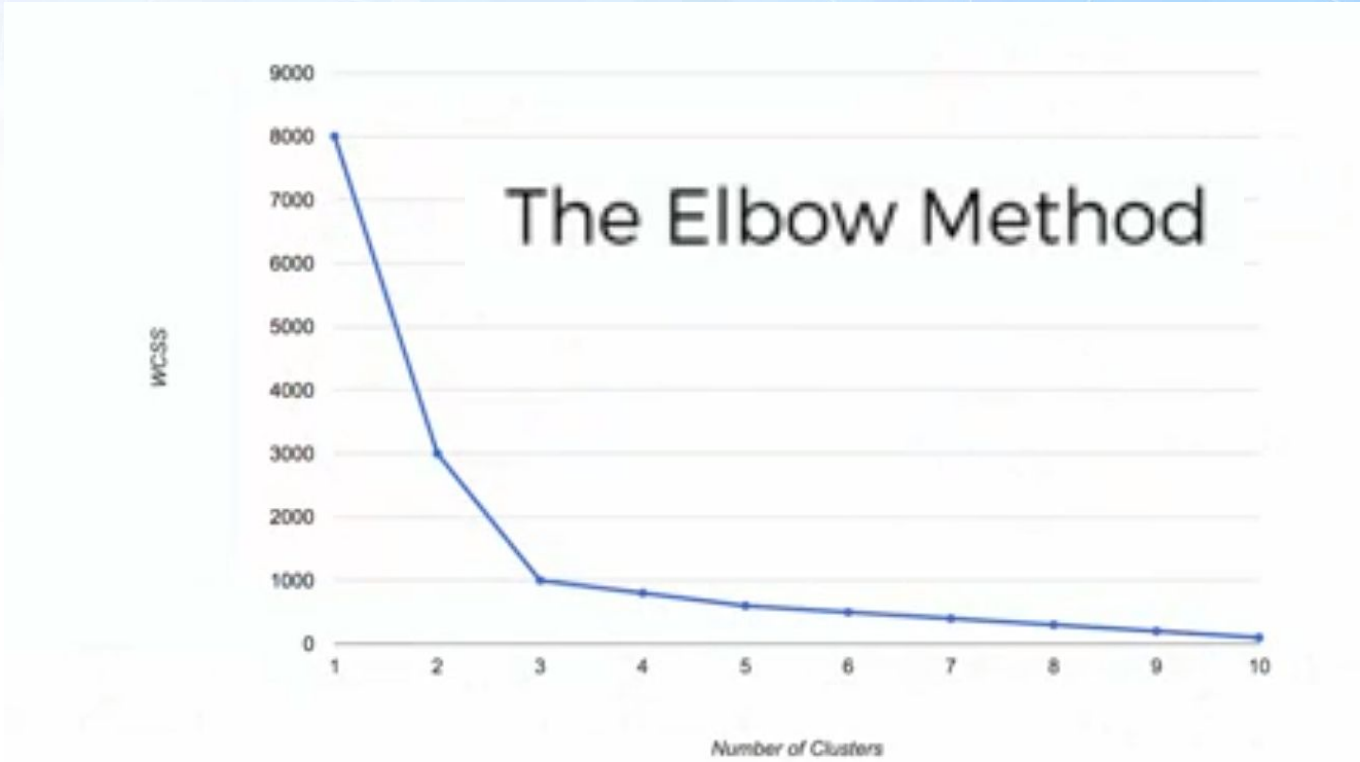
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$

Choosing Right Number of Clusters



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

Choosing Right Number of Clusters



Choosing Right Number of Clusters