



# 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.



**tcs** TATA  
CONSULTANCY  
SERVICES

**Quest**  
global

U  
•  
S  
T

**ibsssoftware**

**Sowparnika**  
Education Infrastructure

# Types of Machine Learning

```
graph TD; A[Types of Machine Learning] --> B[Supervised Learning]; A --> C[Unsupervised Learning]; A --> D[Re-inforcement Learning];
```

## Supervised Learning

- Well defined goals
- Reverse Engineering
- Example – Fraud / Non-Fraud transactions, Inventory management

## Unsupervised Learning

- Outcome is based only on inputs
- Outcome – Typically clustering or segmentation

## Re-inforcement Learning

- Start state and end state are defined
- The agent discovers the path and the relationships on its own

# Supervised vs Unsupervised

## Supervised Learning

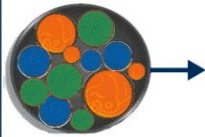
- Known number of classes
- Based on training set
- Used to classify future observations

## Unsupervised Learning

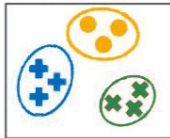
- Unknown number of classes
- No prior knowledge
- Used to understand data

# UNSUPERVISED LEARNING

Raw Data



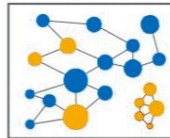
## EXAMPLES



### CLUSTERING

Identifies similarities in groups:

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



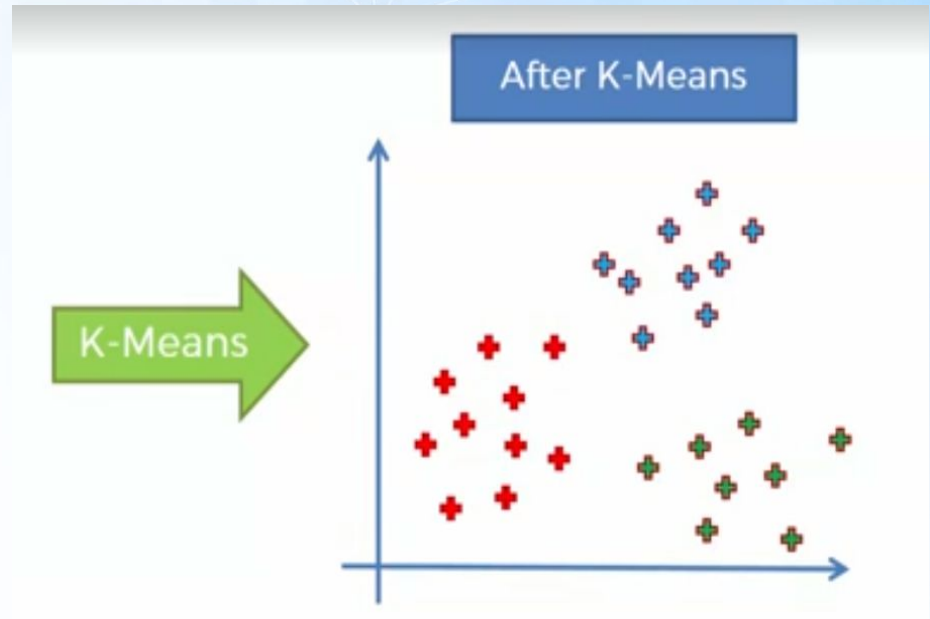
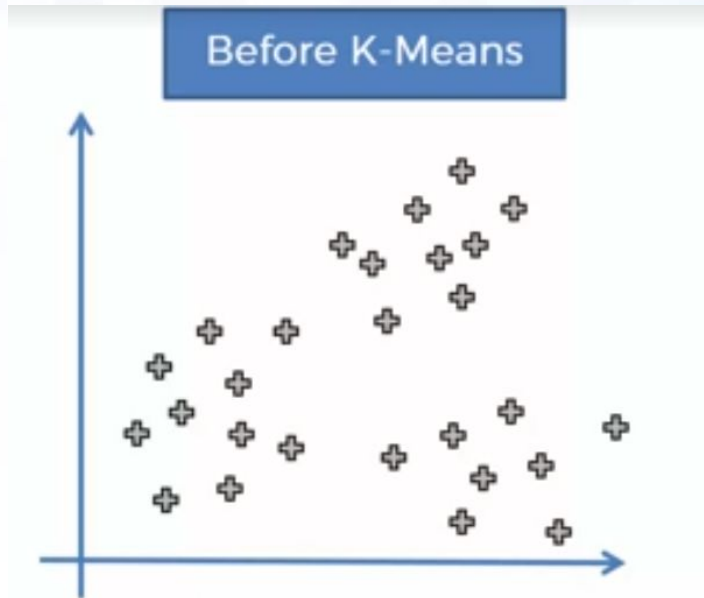
### ANOMALY DETECTION

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?



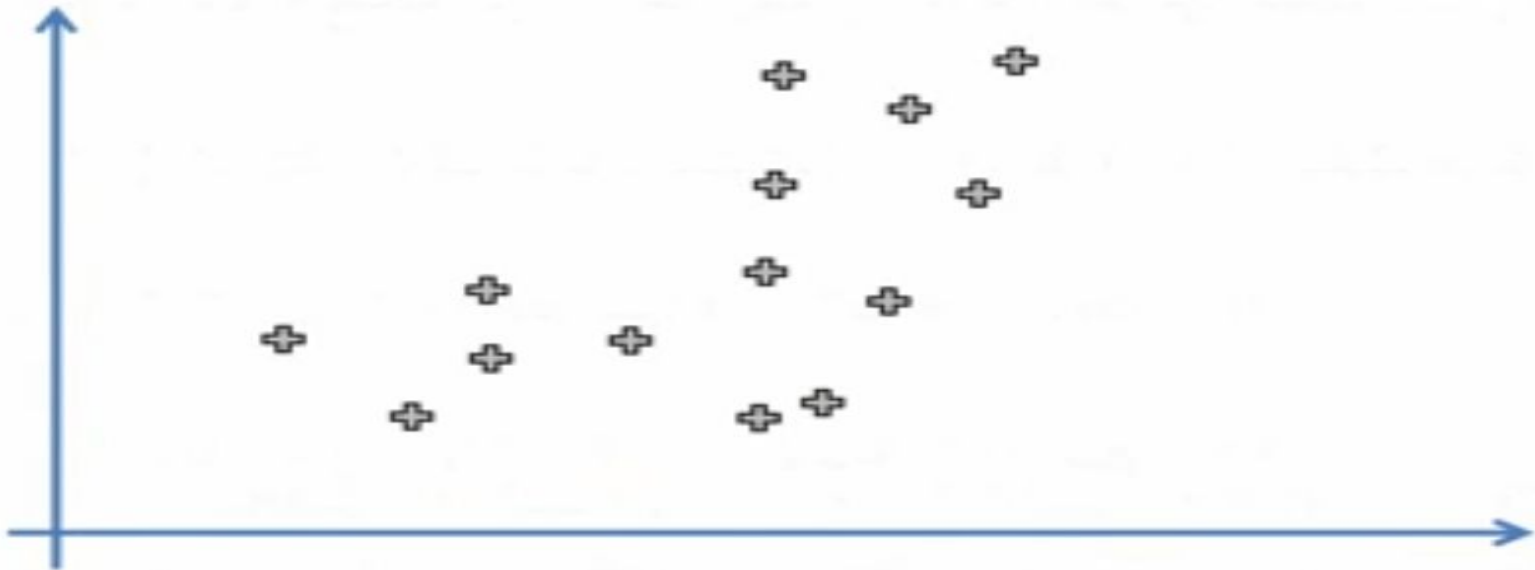
# How K-Means works?

- 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!

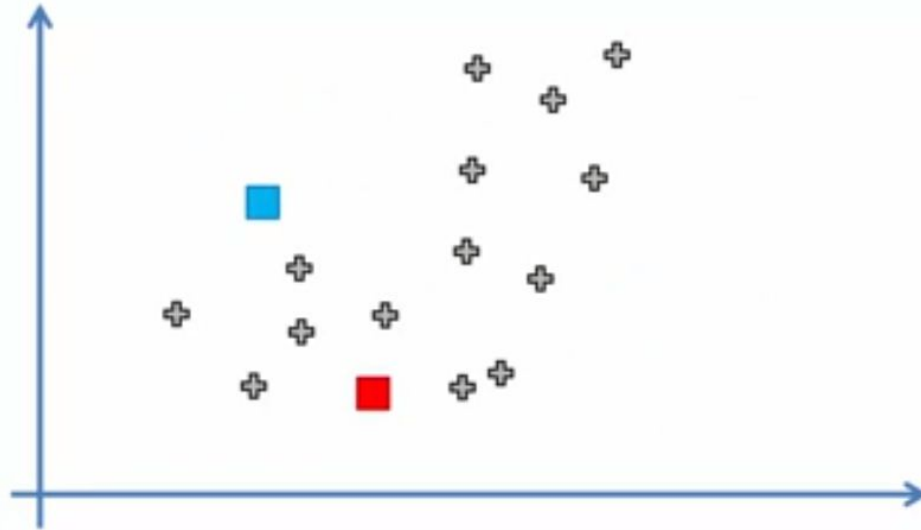
# K-Means Clustering

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



# K-Means Clustering

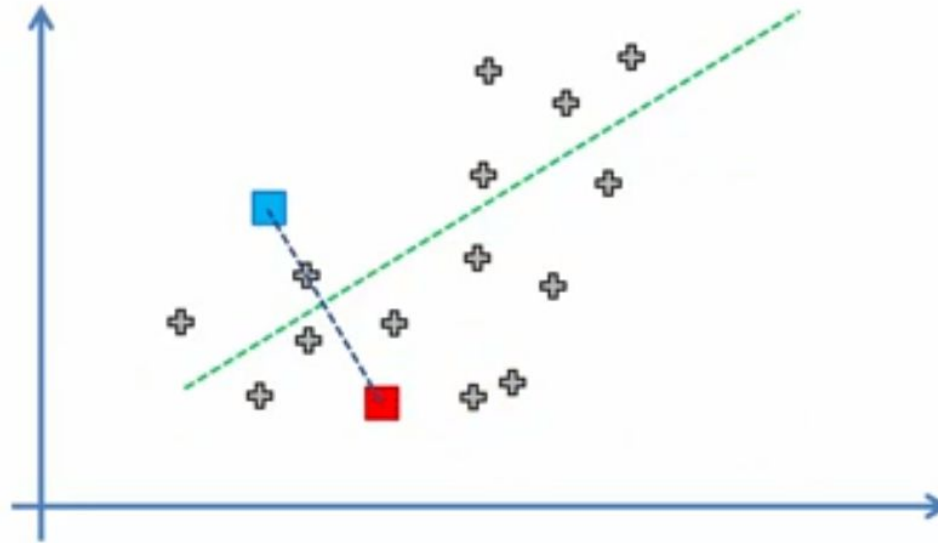
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)





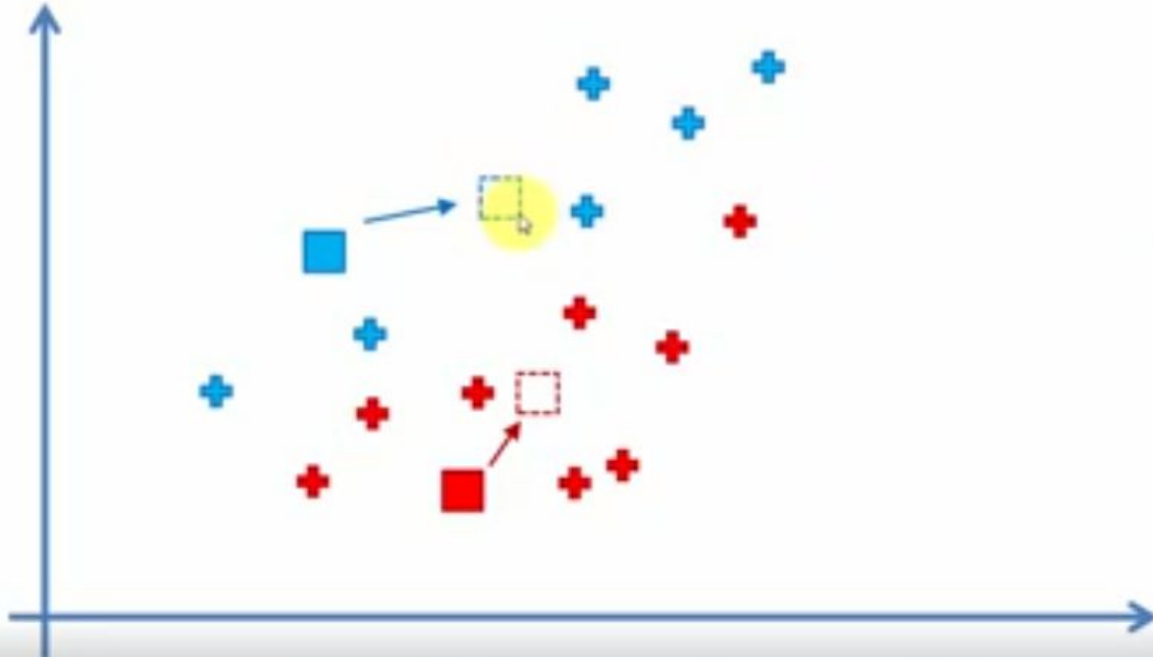
# K-Means Clustering

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



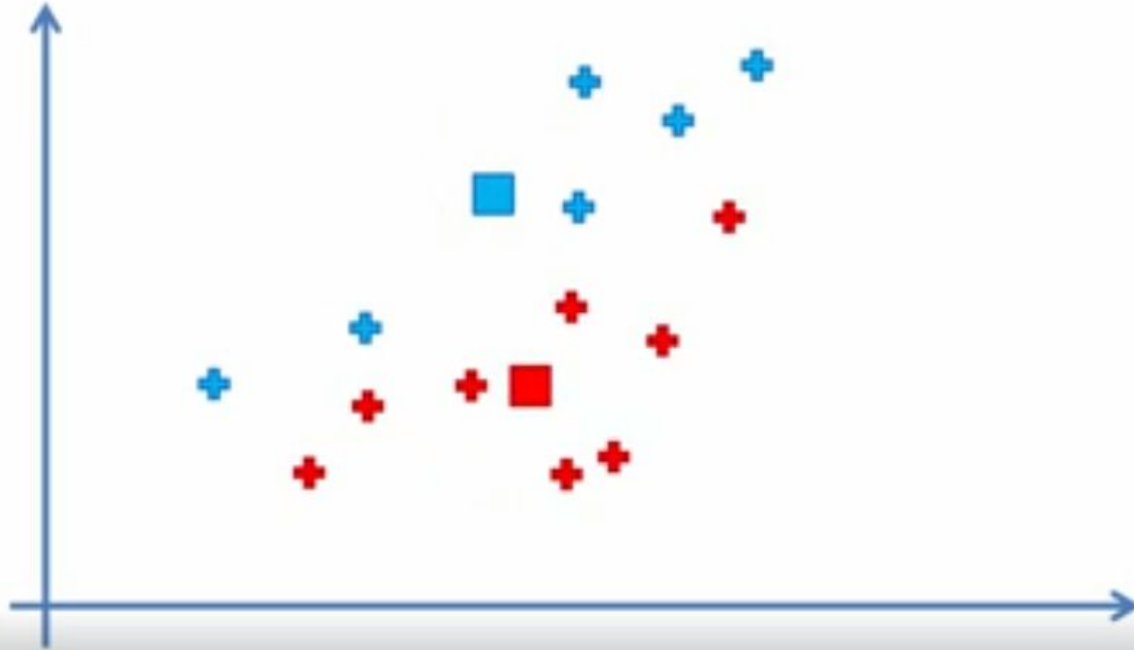
# K-Means Clustering

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



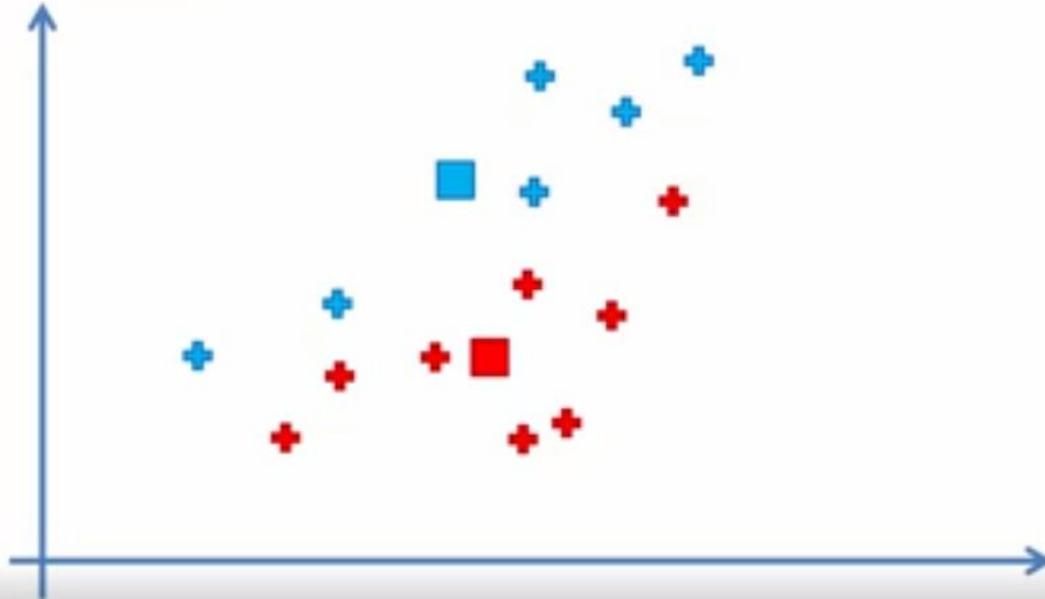
# K-Means Clustering

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



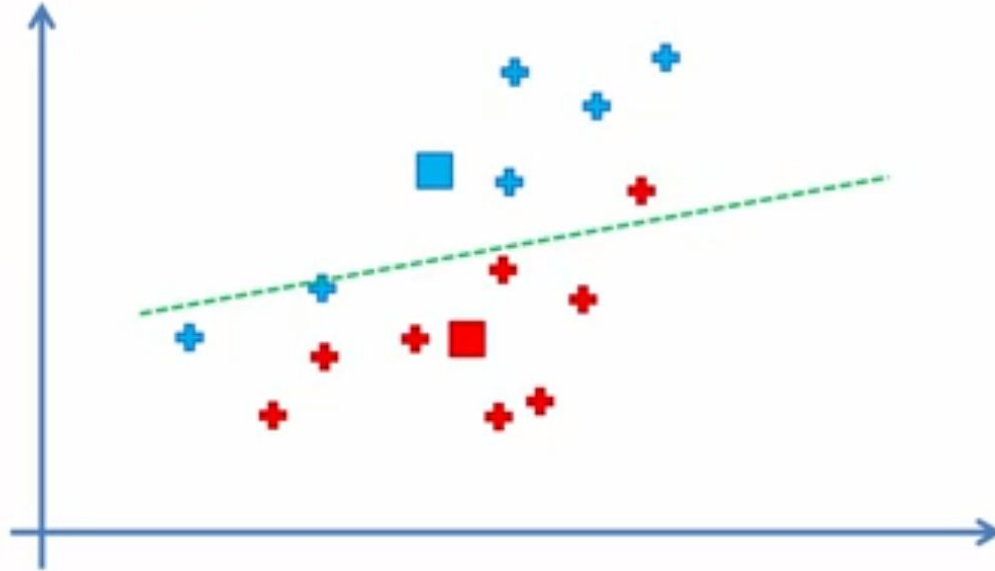
# K-Means Clustering

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



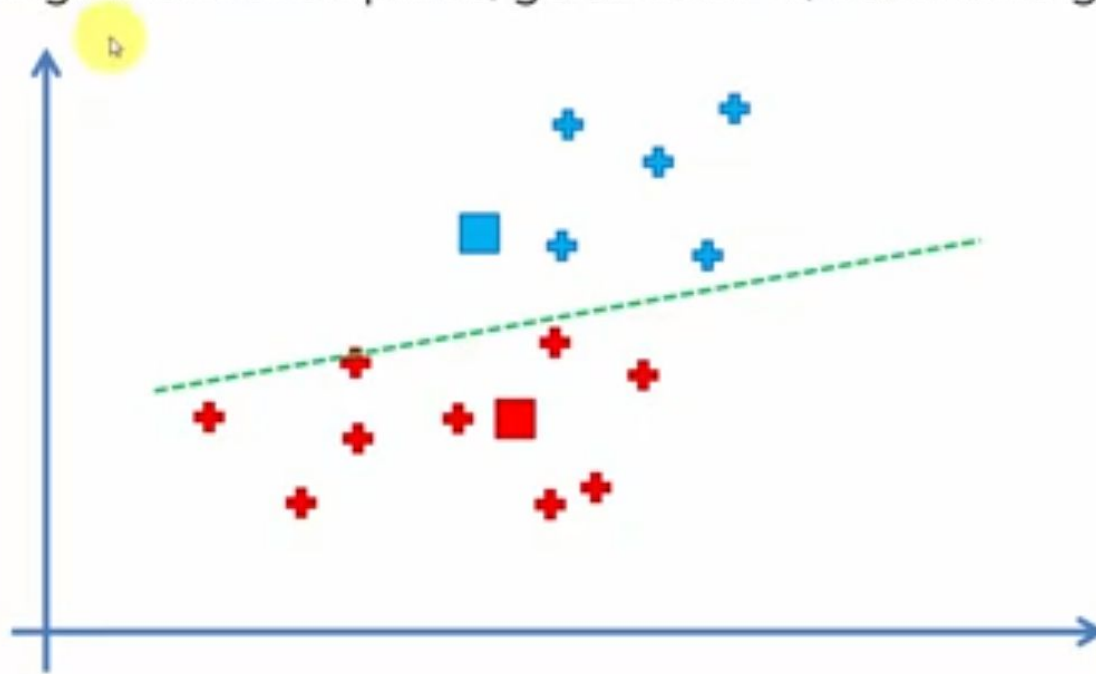
# K-Means Clustering

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



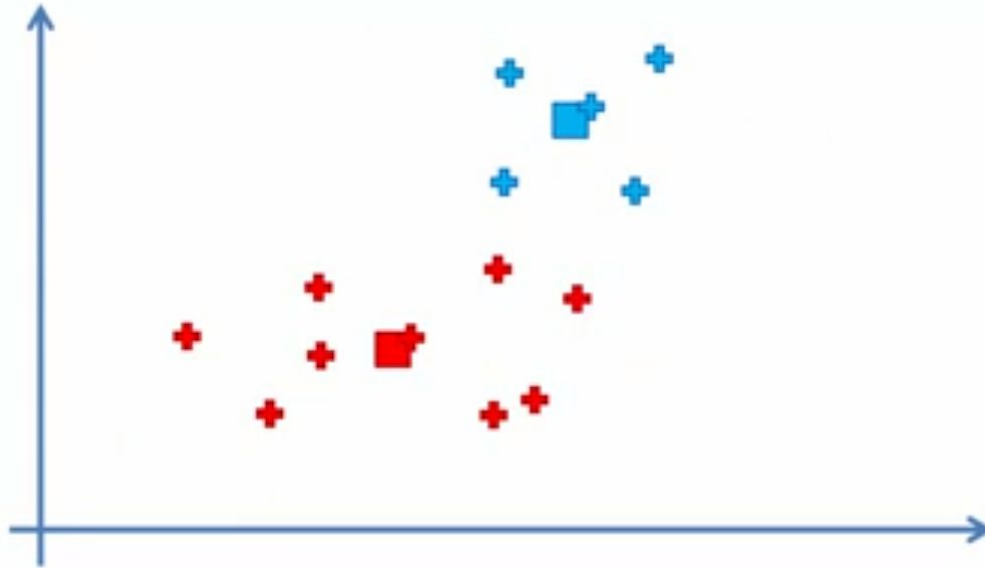
# K-Means Clustering

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



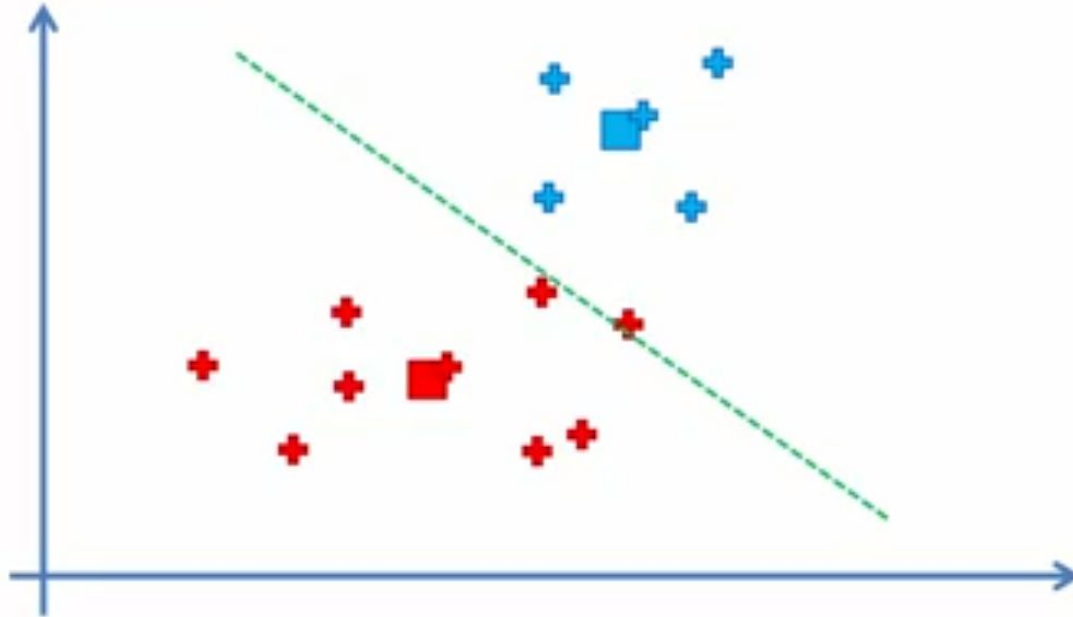
# K-Means Clustering

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



# K-Means Clustering

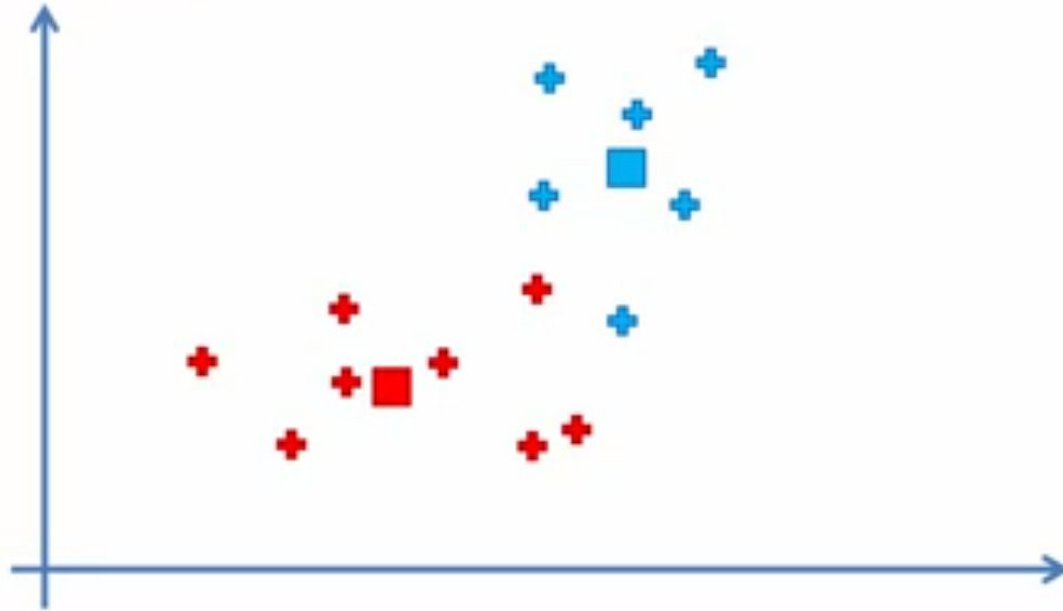
STEP 5: Reassign each data point to the new closest centroid. If any reassignment took place, go to STEP 4, otherwise go to FIN.





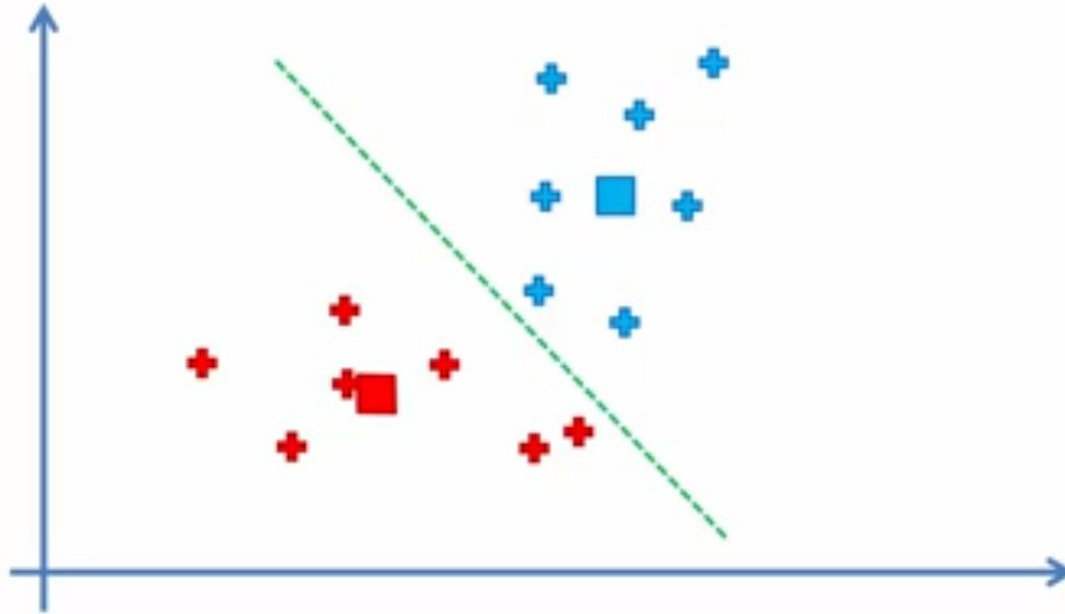
# K-Means Clustering

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



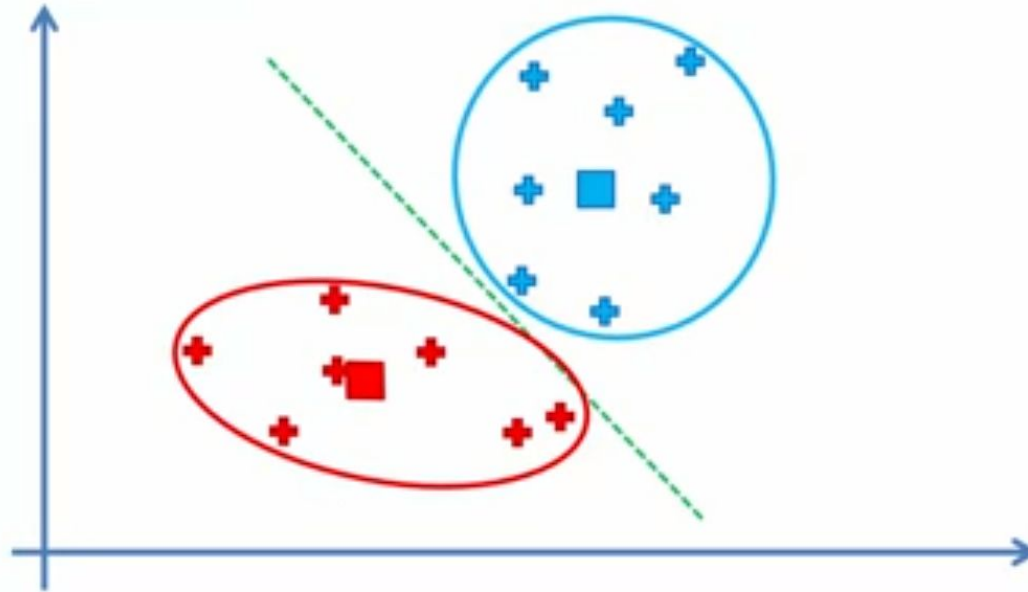
# K-Means Clustering

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



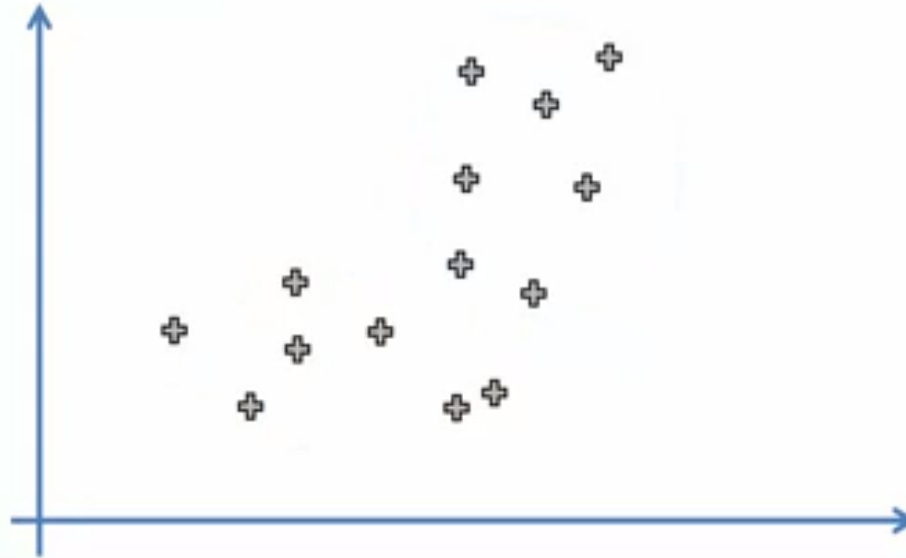
# K-Means Clustering

FIN: Your Model Is Ready



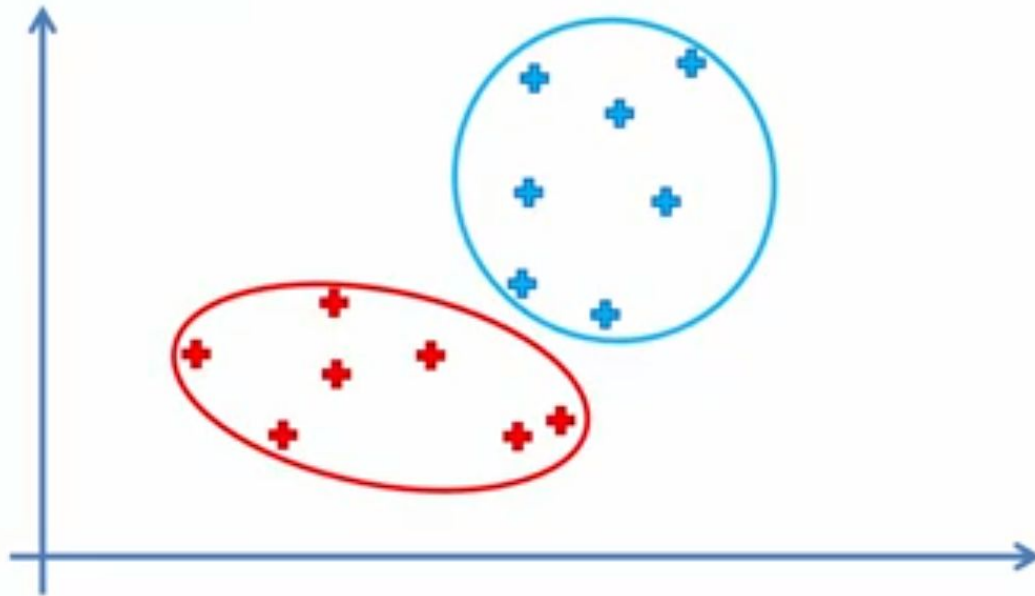
# K-Means Clustering

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

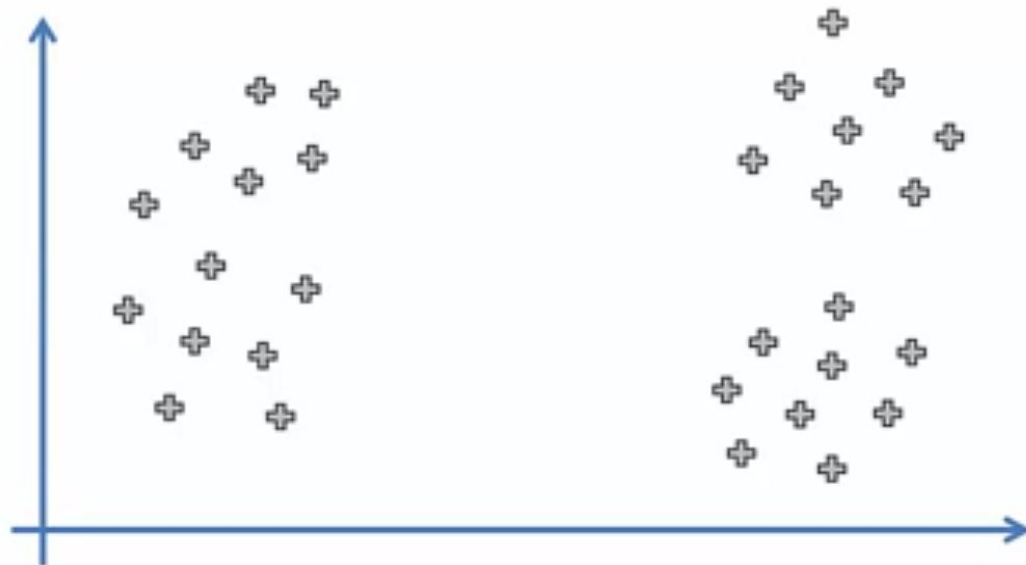


# K-Means Clustering

FIN: Your Model Is Ready

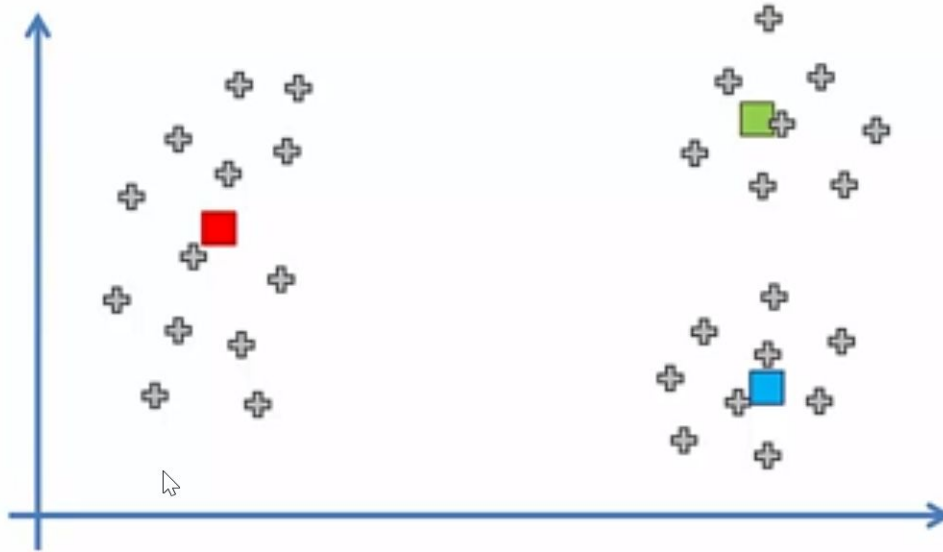


# Random Initialization Trap



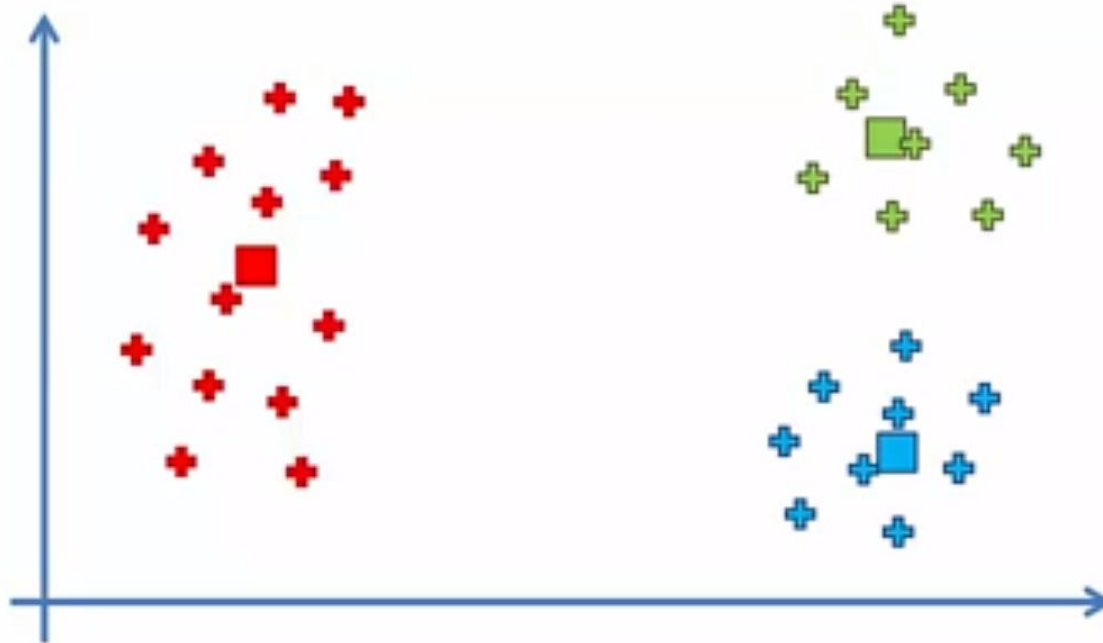
If we choose  $K = 3$  clusters...

# Random Initialization Trap



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

# Random Initialization Trap



...the following three clusters

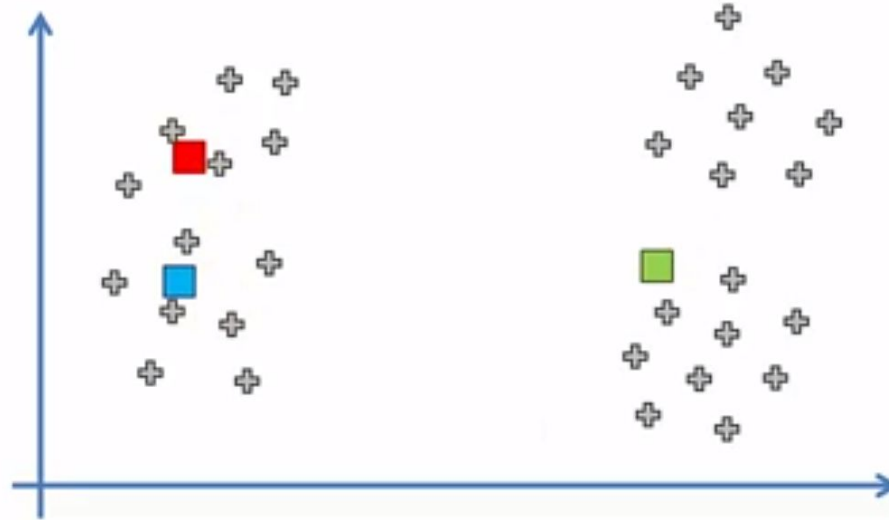


# Random Initialization Trap

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

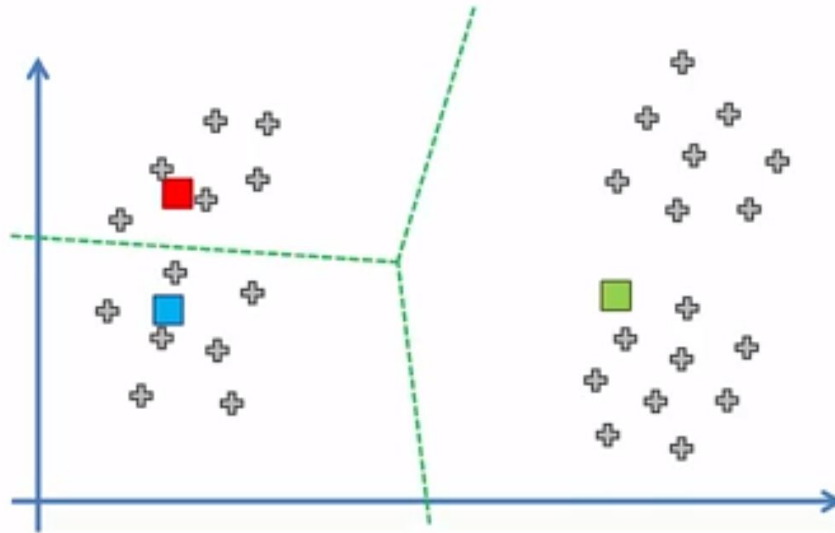
# Random Initialization Trap

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



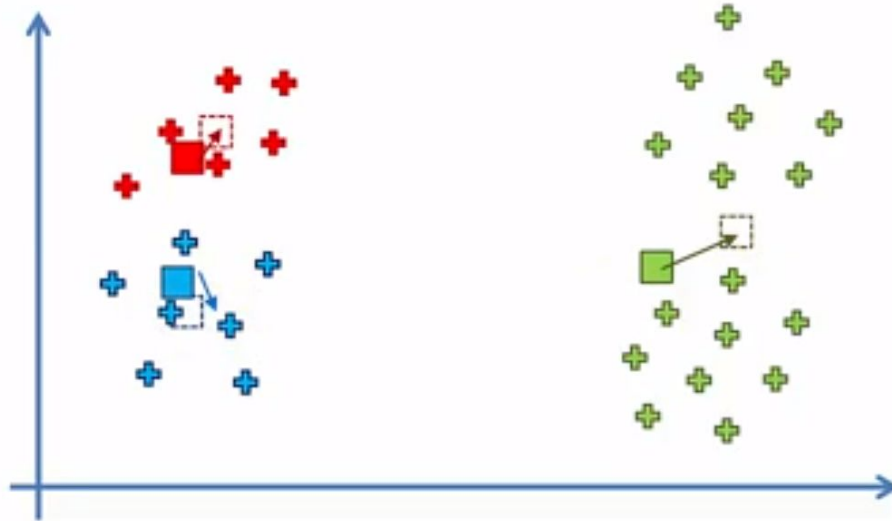
# Random Initialization Trap

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



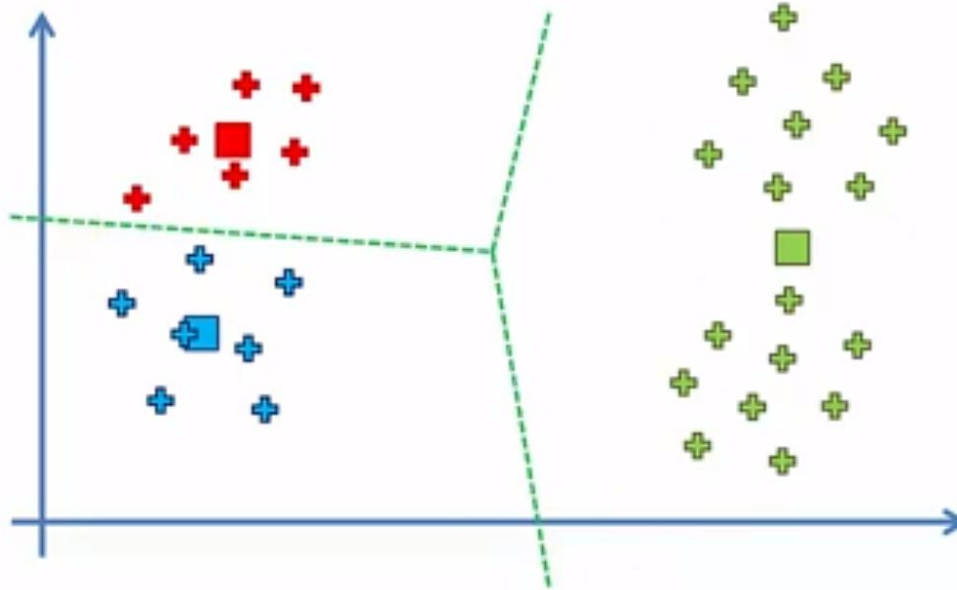
# Random Initialization Trap

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



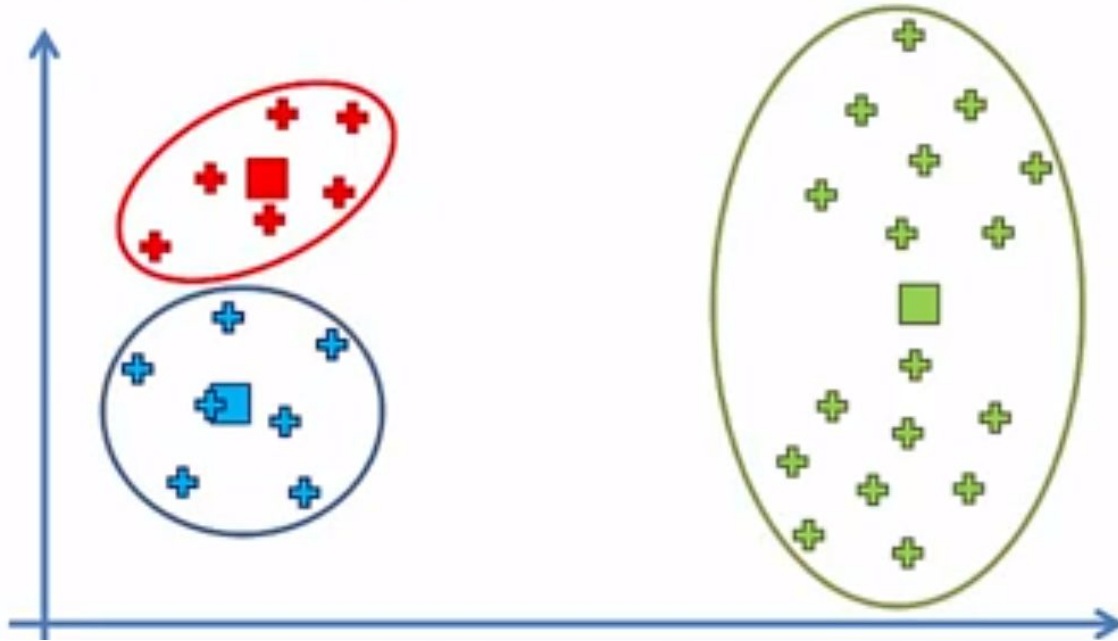
# Random Initialization Trap

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

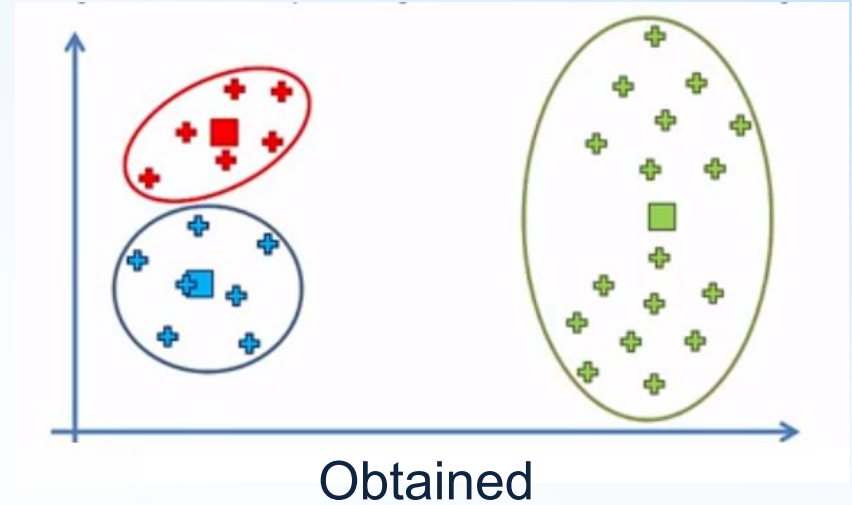
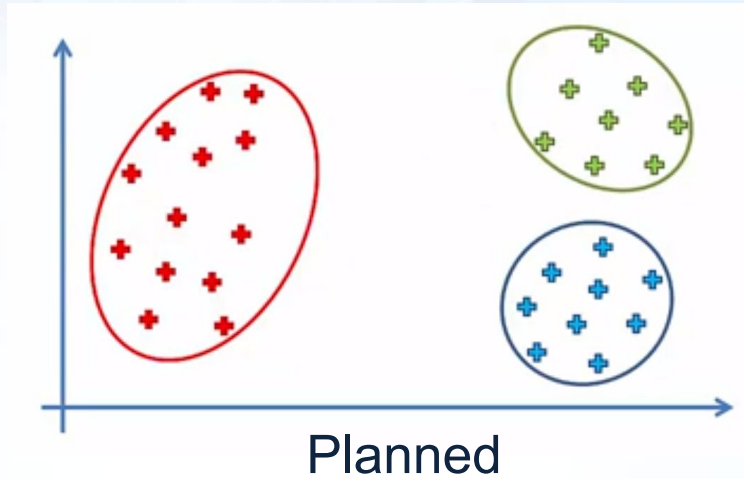


# Random Initialization Trap

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

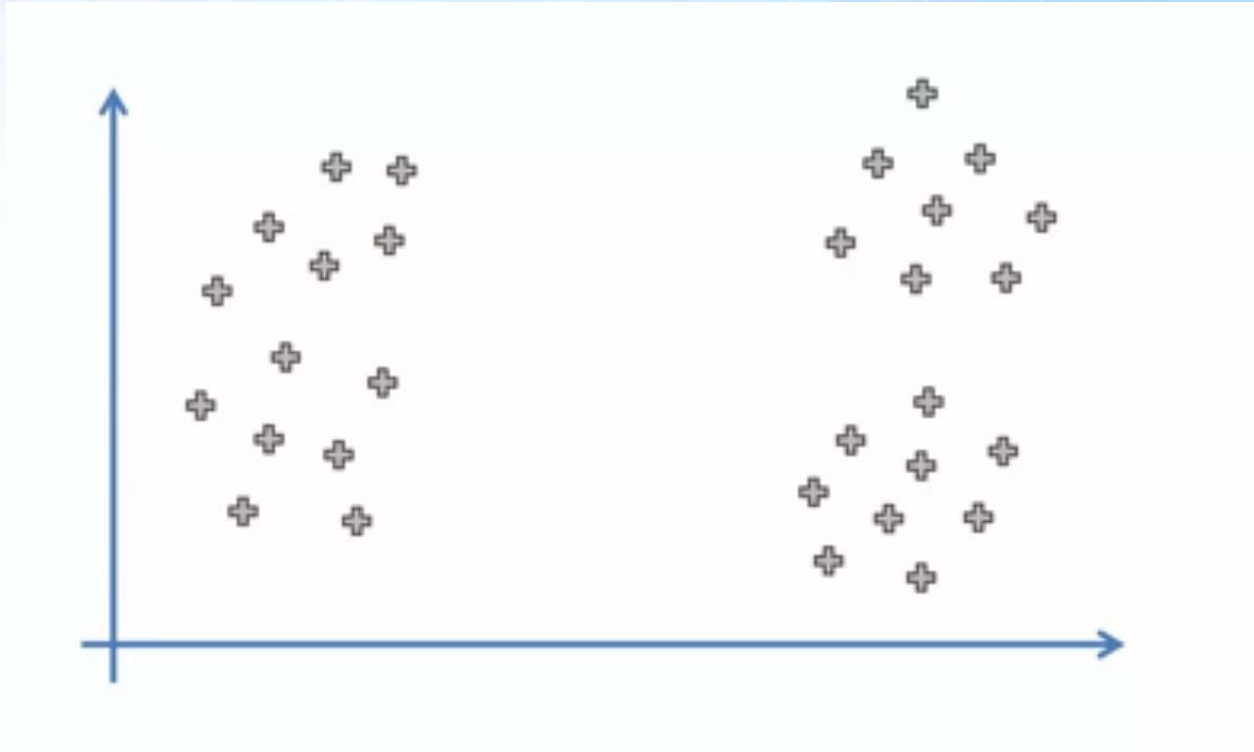


# Random Initialization Trap

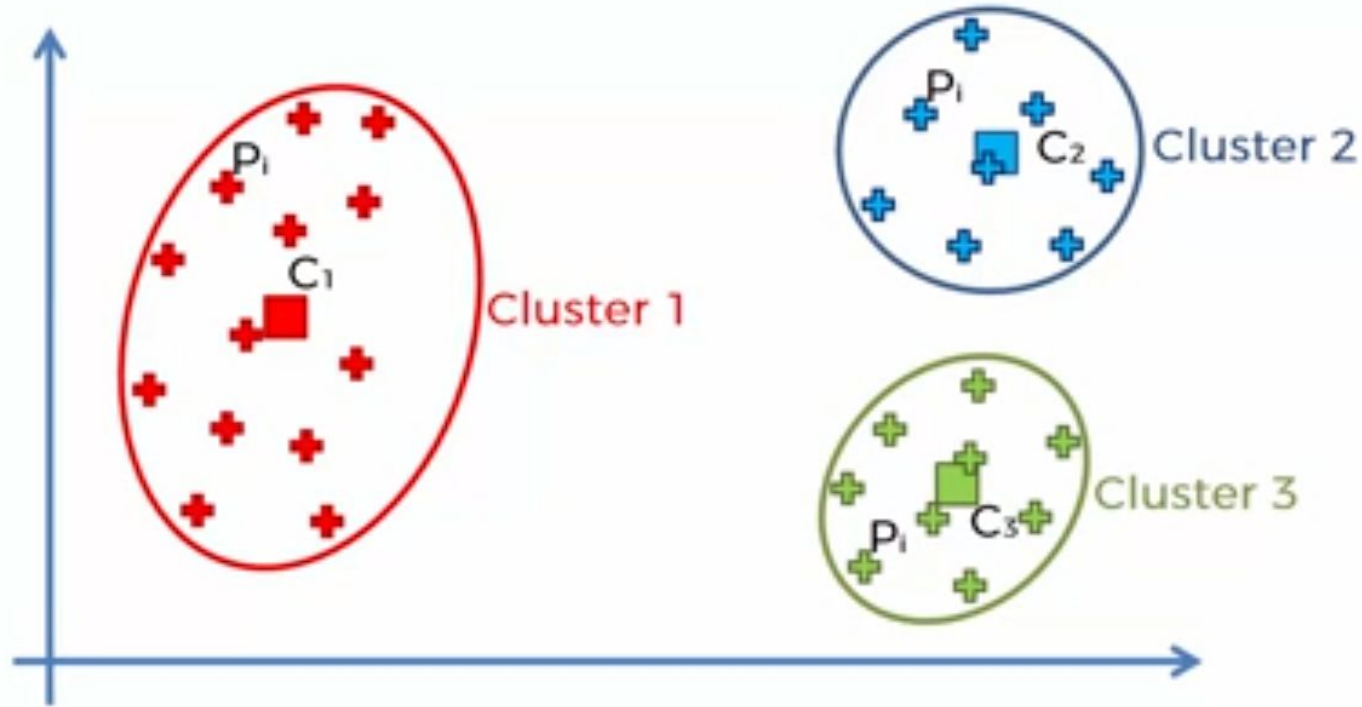




# Choosing Right Number of Clusters



# Choosing Right Number of Clusters

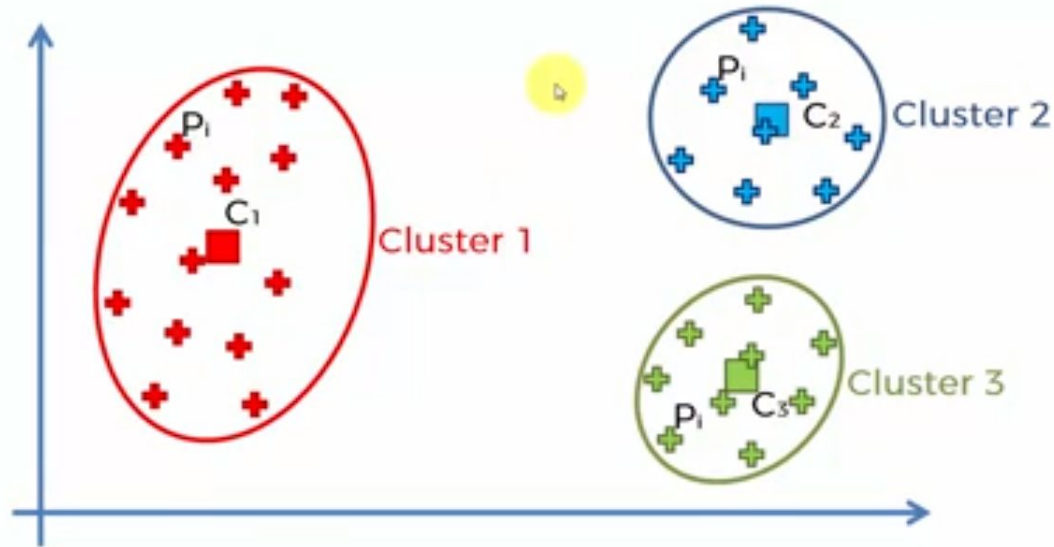


# Choosing Right Number of Clusters

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

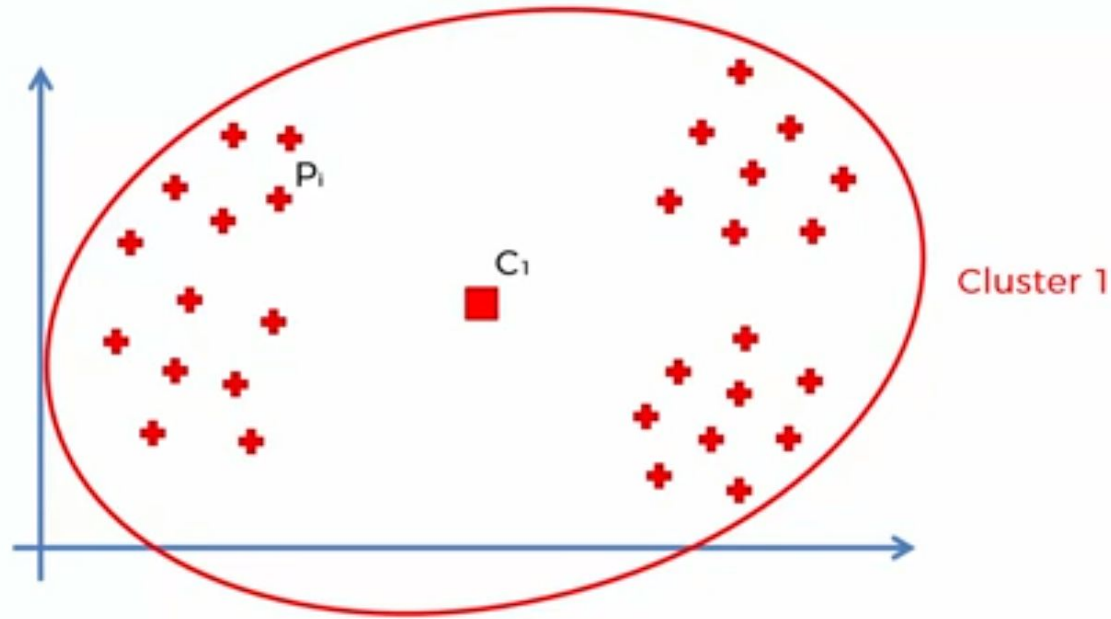
**\* within-cluster sums of squares**

# Choosing Right Number of Clusters



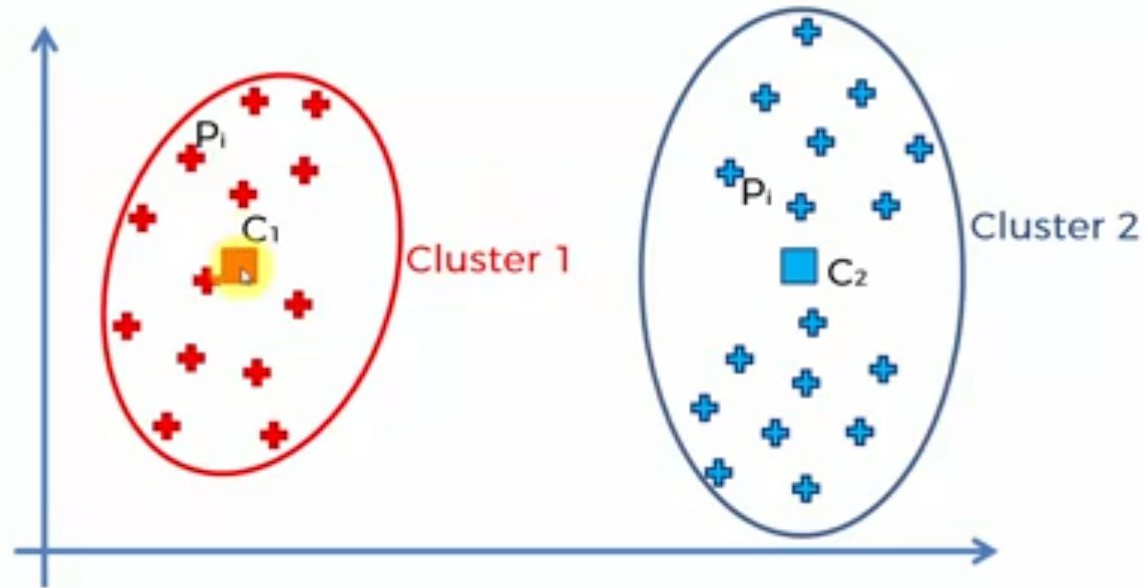
$$WCSS = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

# Choosing Right Number of Clusters



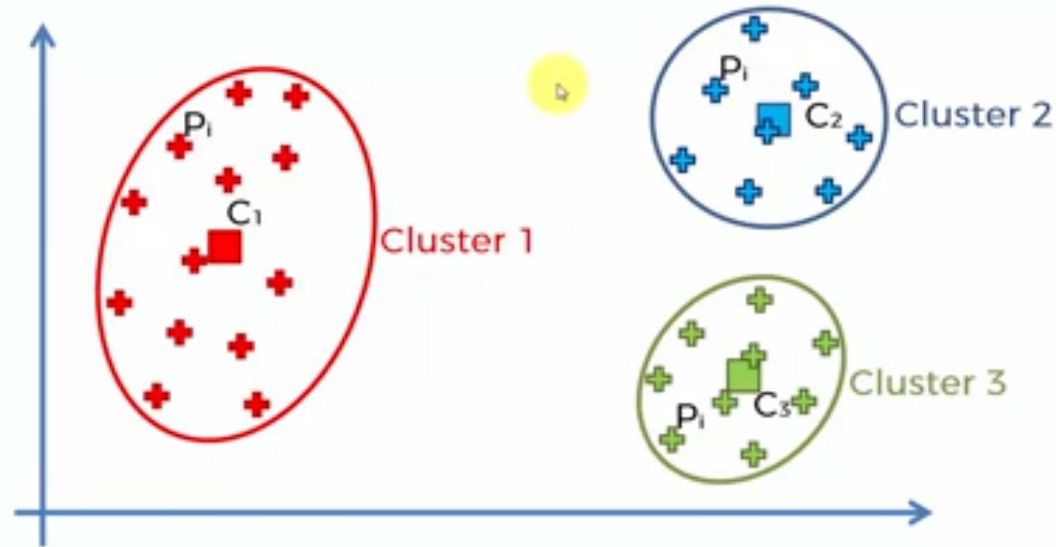
$$WCSS = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2$$

# Choosing Right Number of Clusters



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

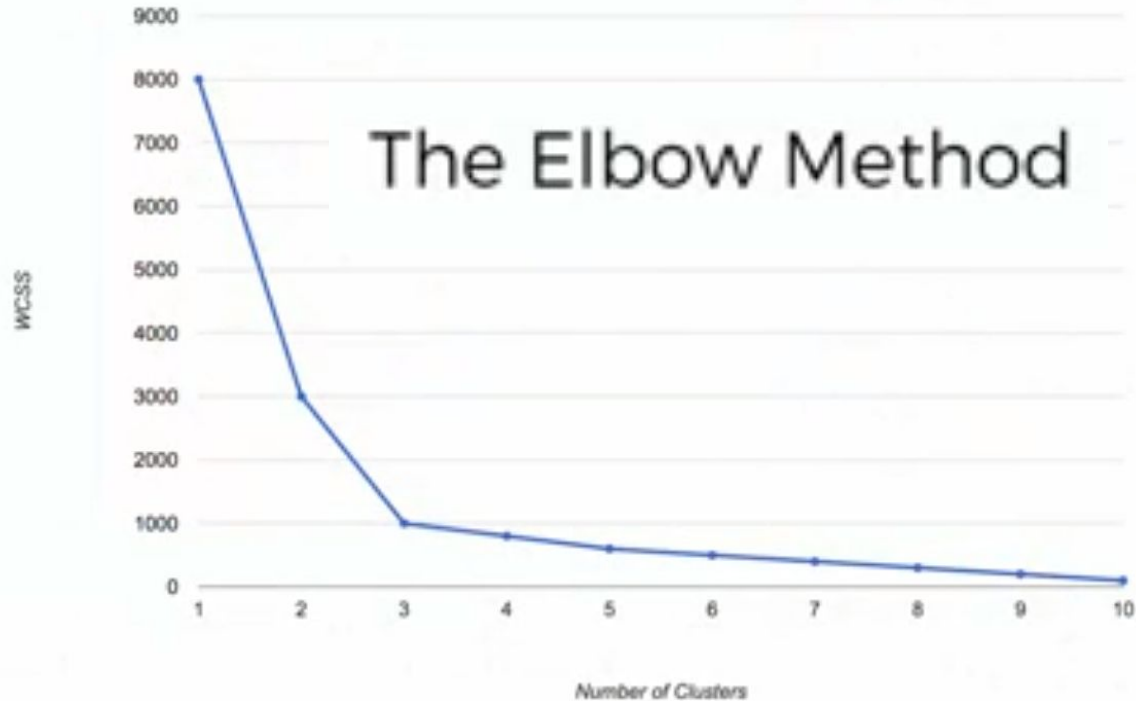
# Choosing Right Number of Clusters



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



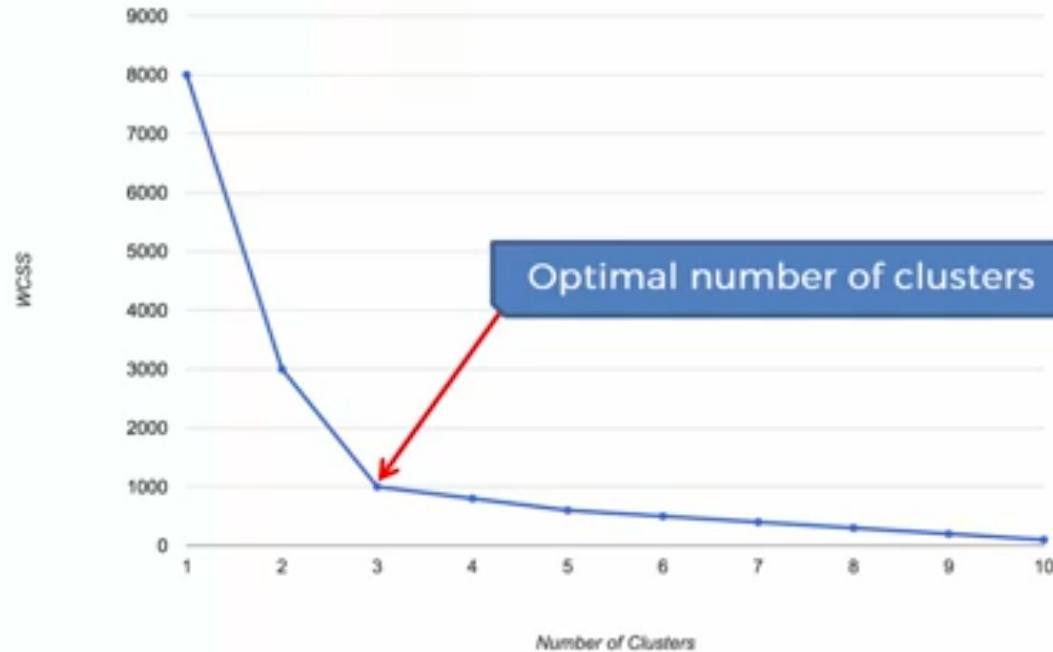
# Choosing Right Number of Clusters





# Choosing Right Number of Clusters

## The Elbow Method





© 2023

## ICT Academy of Kerala

“All rights reserved. Permission granted to reproduce for personal and educational use only. Commercial copying, hiring, lending is prohibited. Any unauthorized broadcasting, public performance, copying or re-recording will constitute an infringement of copyright”