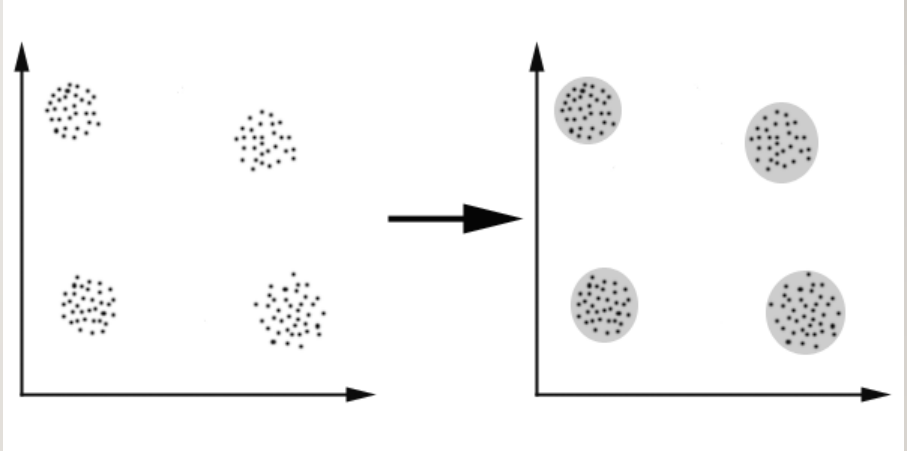
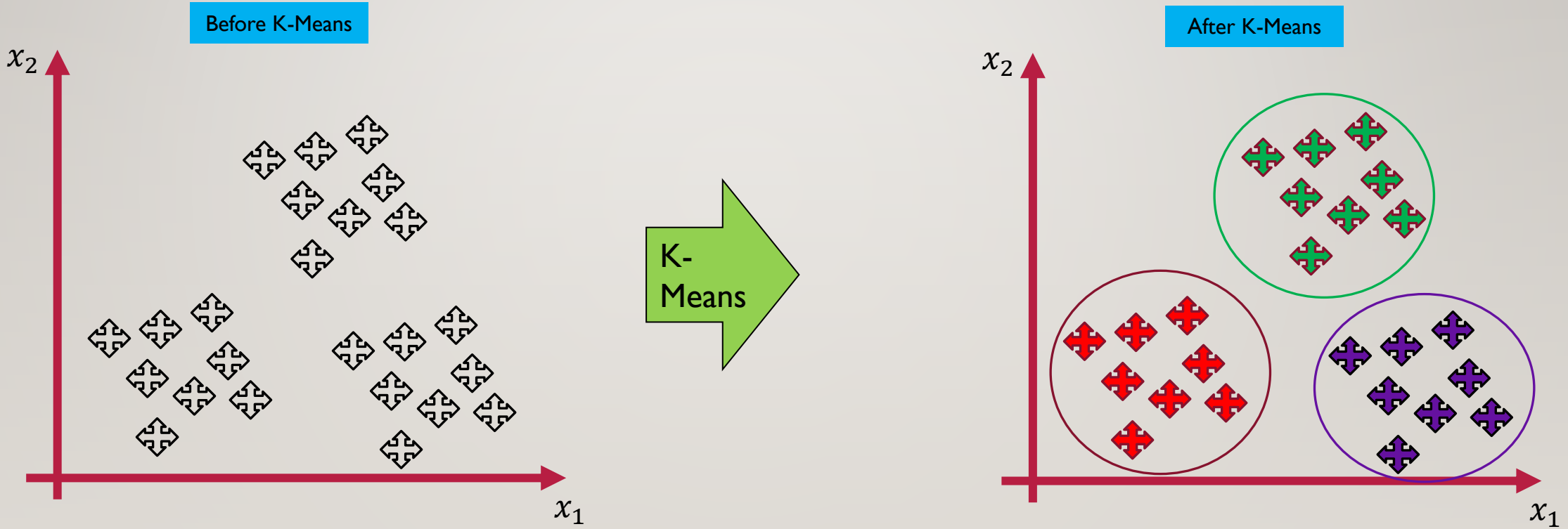
K-MEANS CLUSTERING

CLUSTERING

- Clustering falls into the category of un-supervised learning.
- **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.



WHAT K-MEANS DOES FOR YOU ?



HOW DOES IT WORKS?

STEP 1: Choose the number K clusters.

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

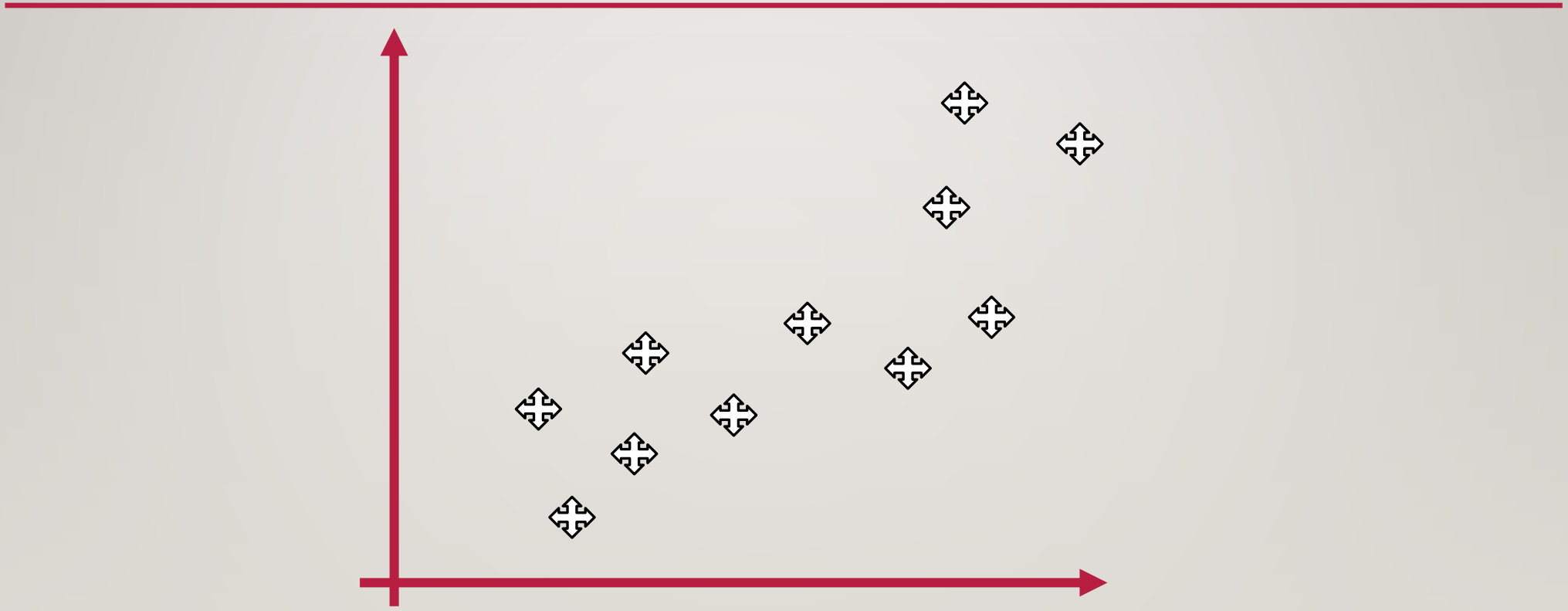
STEP 3: Assign each data point to the closest centroid – That forms K clusters.

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

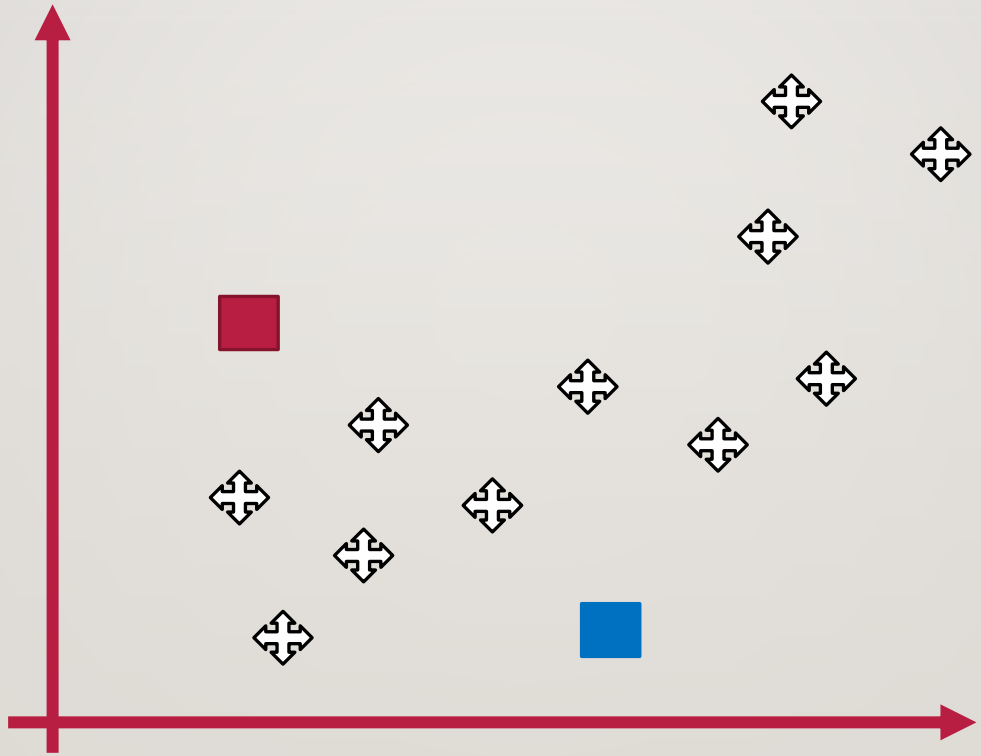
STEP 5: Reassign each data point to the new closet centroid.

If reassignment took place, go to step 4, otherwise go to FIN**.(Model Created)**

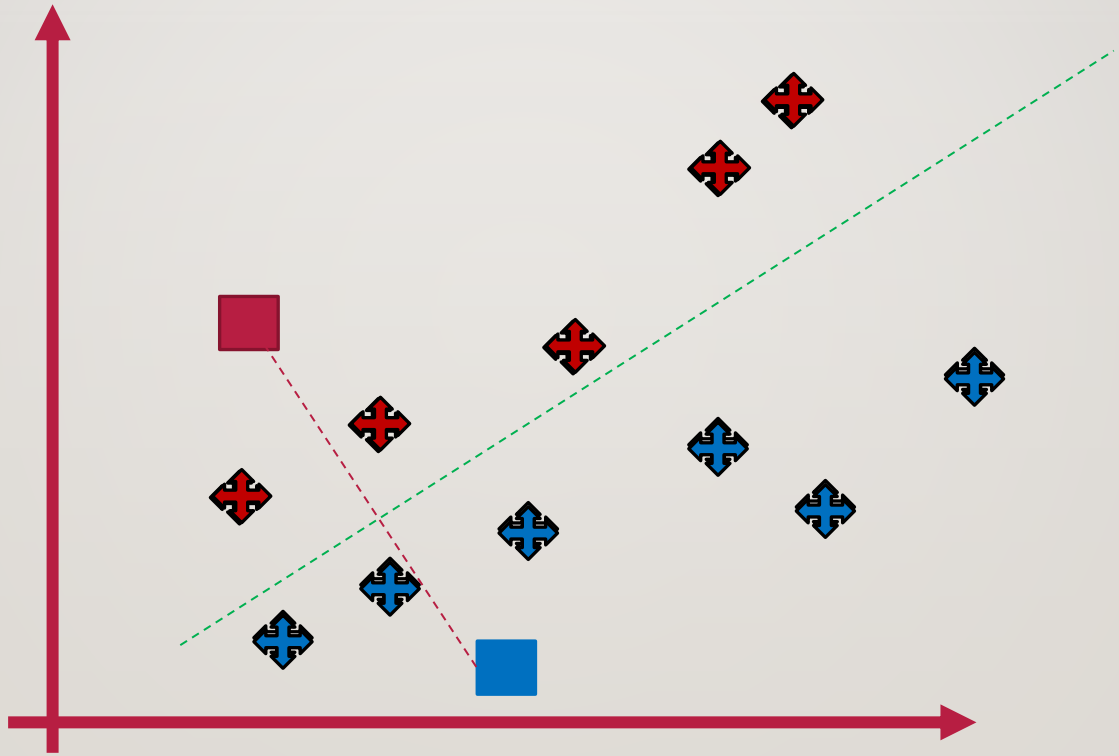
STEP 1: CHOOSE THE NUMBER K OF CLUSTERS: K=2



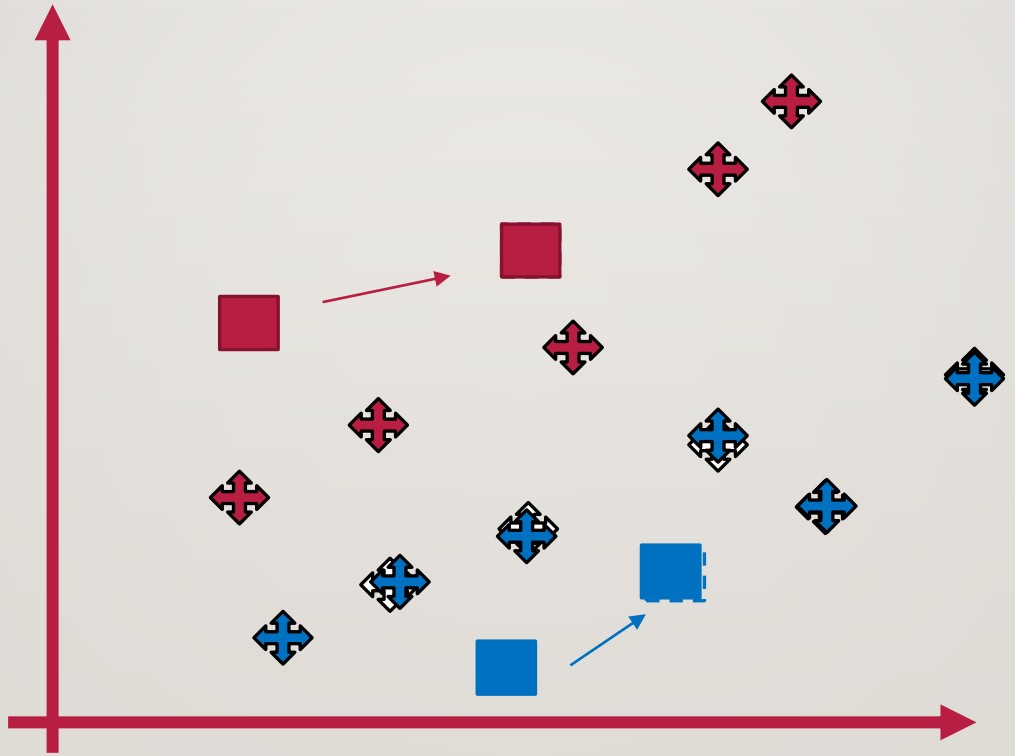
STEP 2: SELECT AT RANDOM K POINTS, THE CENTROID (NOT NECESSARILY FROM YOUR DATASET)



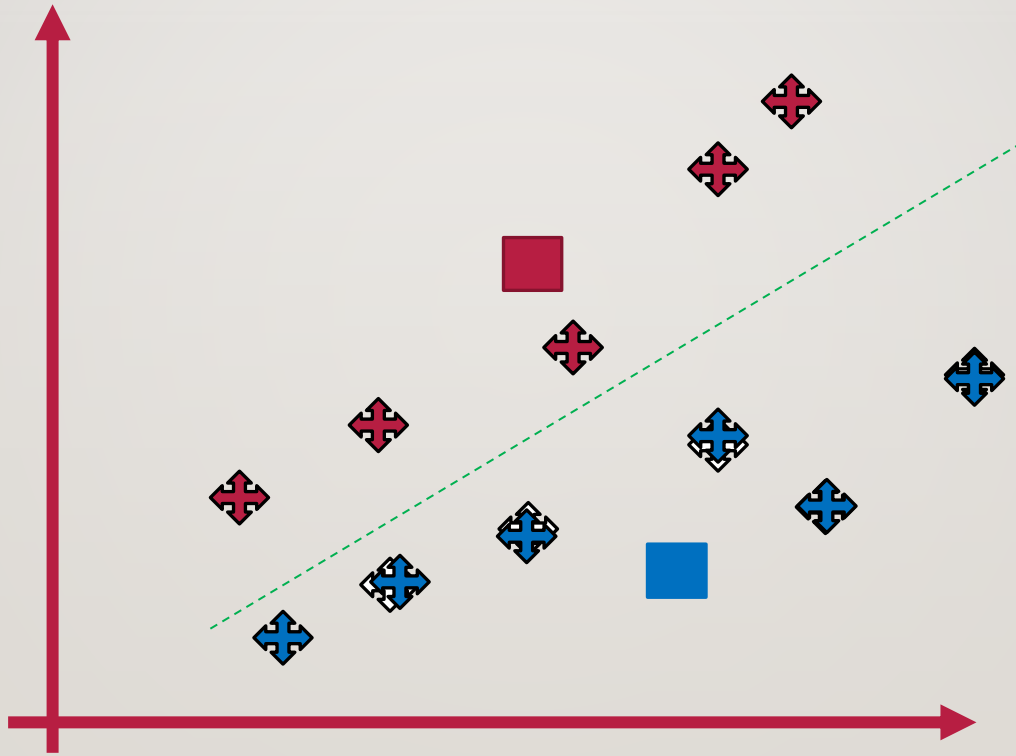
STEP 3: ASSIGN EACH DATA POINT TO THE CLOSEST CENTROID – THAT FORMS K CLUSTERS



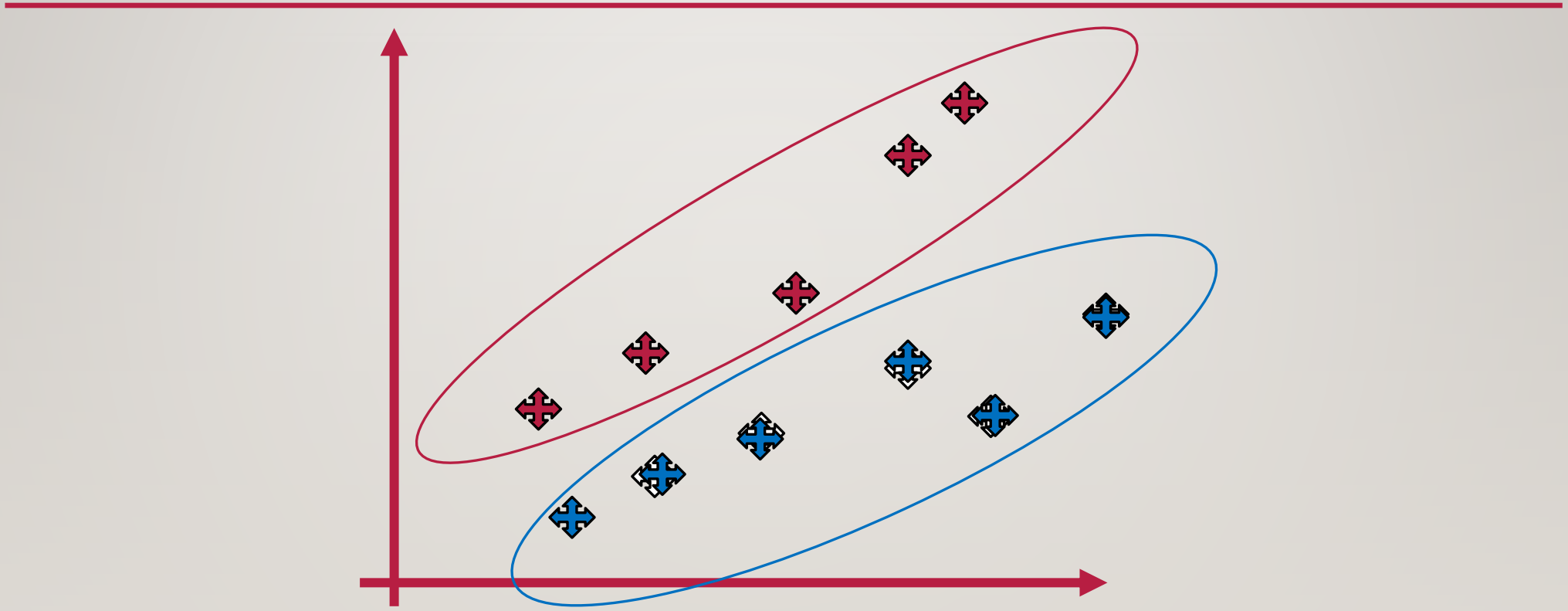
STEP 4: COMPUTE AND PLACE THE NEW CENTROID OF EACH CLUSTER



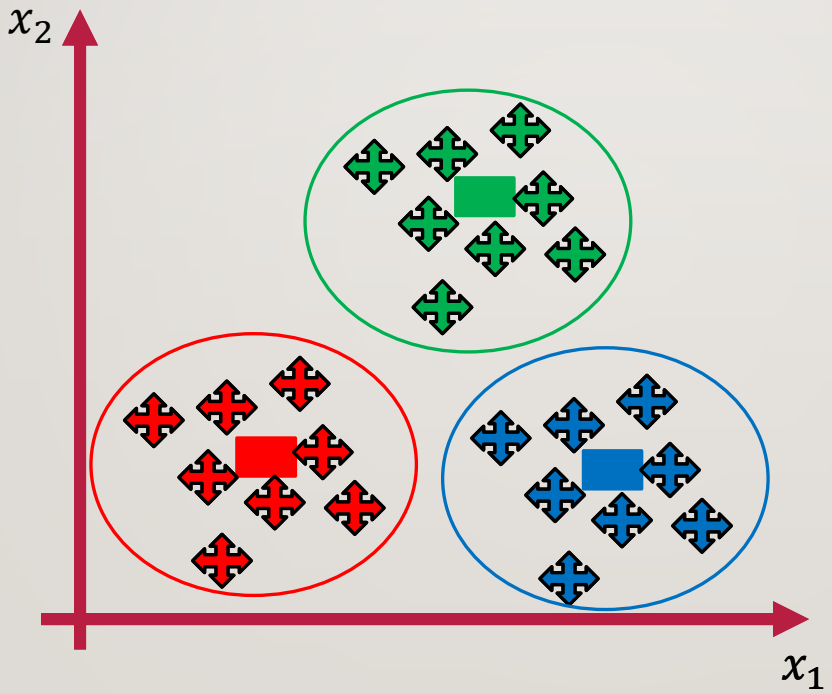
STEP 5: REASSIGN EACH DATA POINT TO THE NEW CLOSEST CENTROID. IF AN REASSIGNMENT TOOK PLACE GO TO STEP 4, OTHERWISE GO TO FIN.



FINISHED MODEL



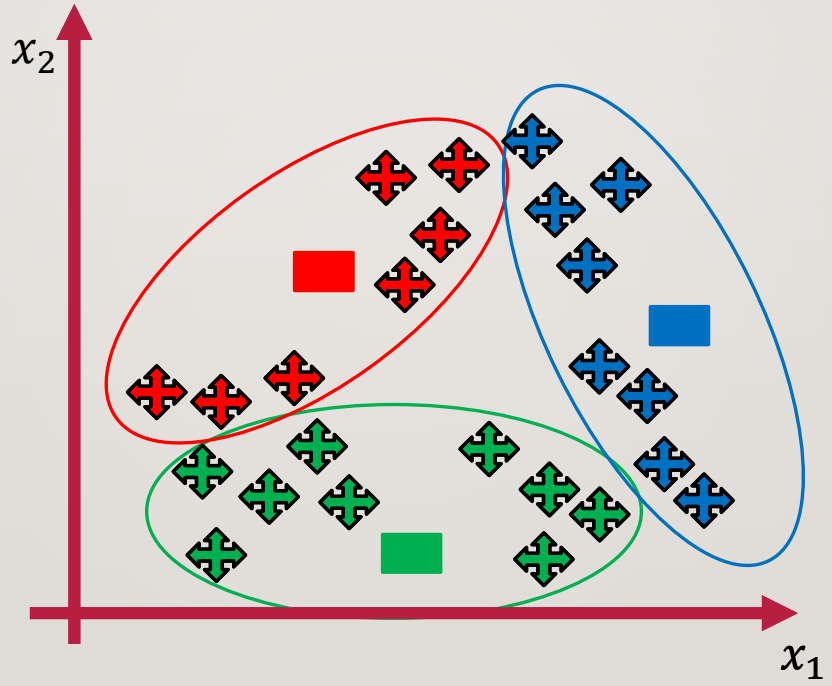
RANDOM INITIALIZATION TRAP



RANDOM INITIALIZATION TRAP

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

RANDOM INITIALIZATION TRAP



AVOIDING THE TRAP

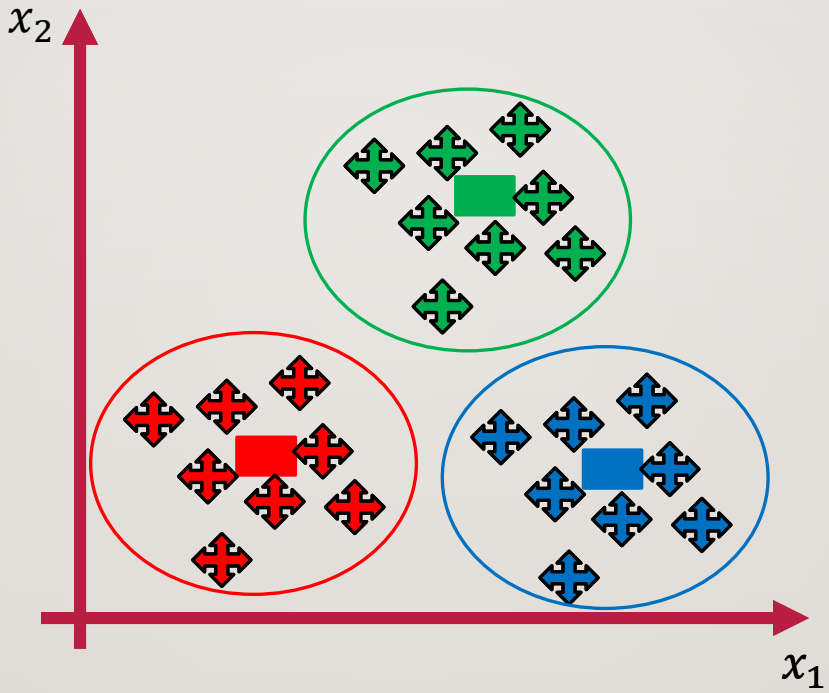
Solution

K-Means++

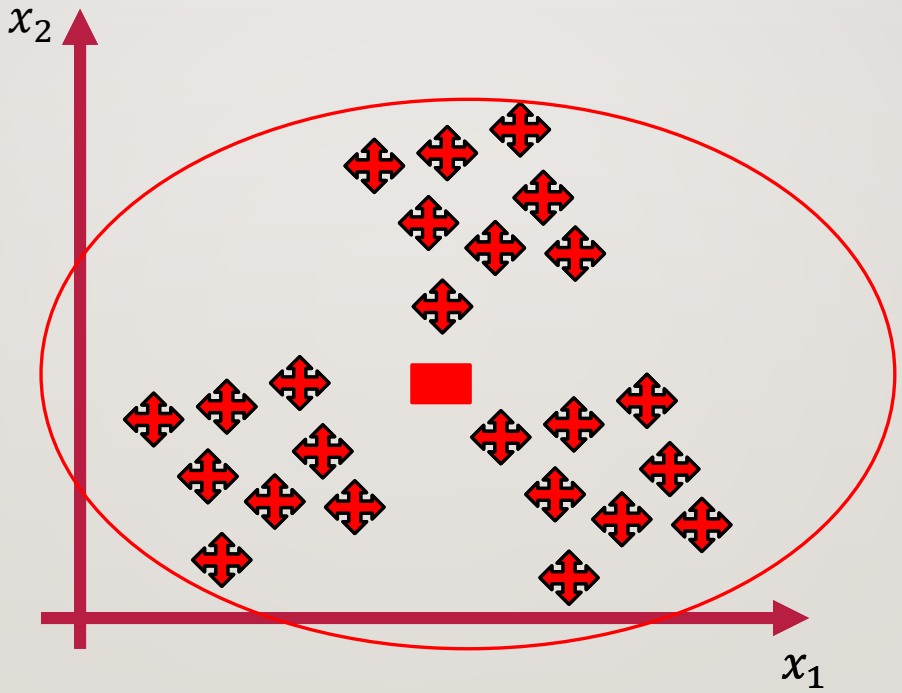
WITHIN CLUSTERS SUM OF SQUARES

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

CONTD..

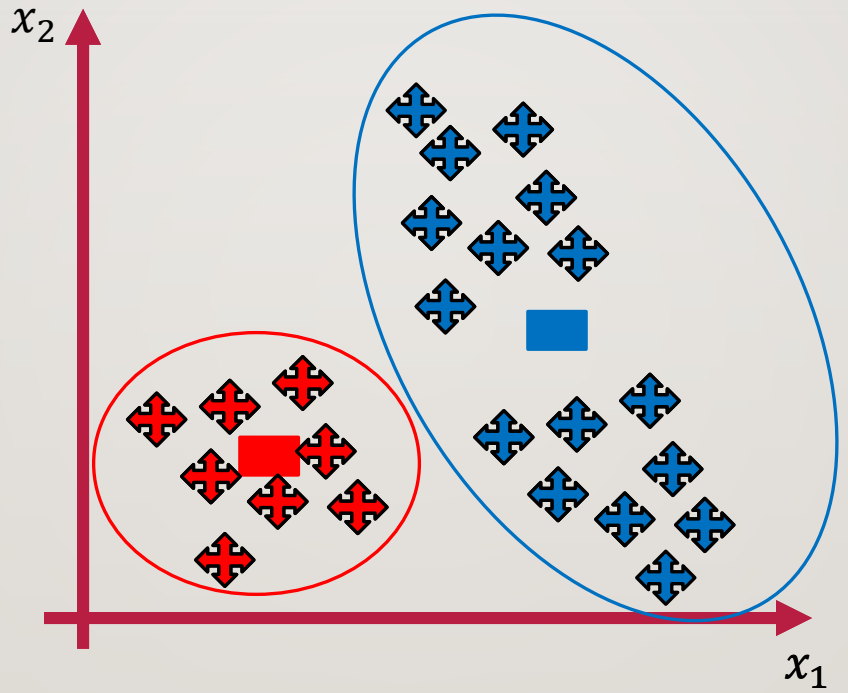


$$\text{WCSS} = \underset{P_i \text{ in cluster } 1}{\Sigma\, distance(P_i, C_1)^2} + \underset{P_i \text{ in cluster } 2}{\Sigma\, distance(P_i, C_2)^2} + \underset{P_i \text{ in cluster } 3}{\Sigma\, distance(P_i, C_3)^2}$$

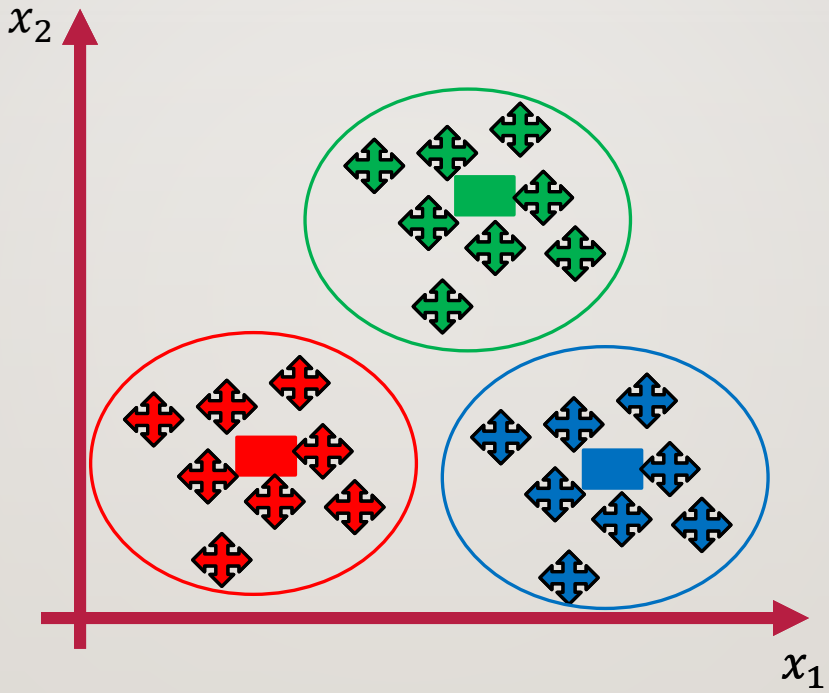


$$\text{WCSS} = \Sigma\, distance(P_i, C_1)^2$$

P_i *in cluster* 1

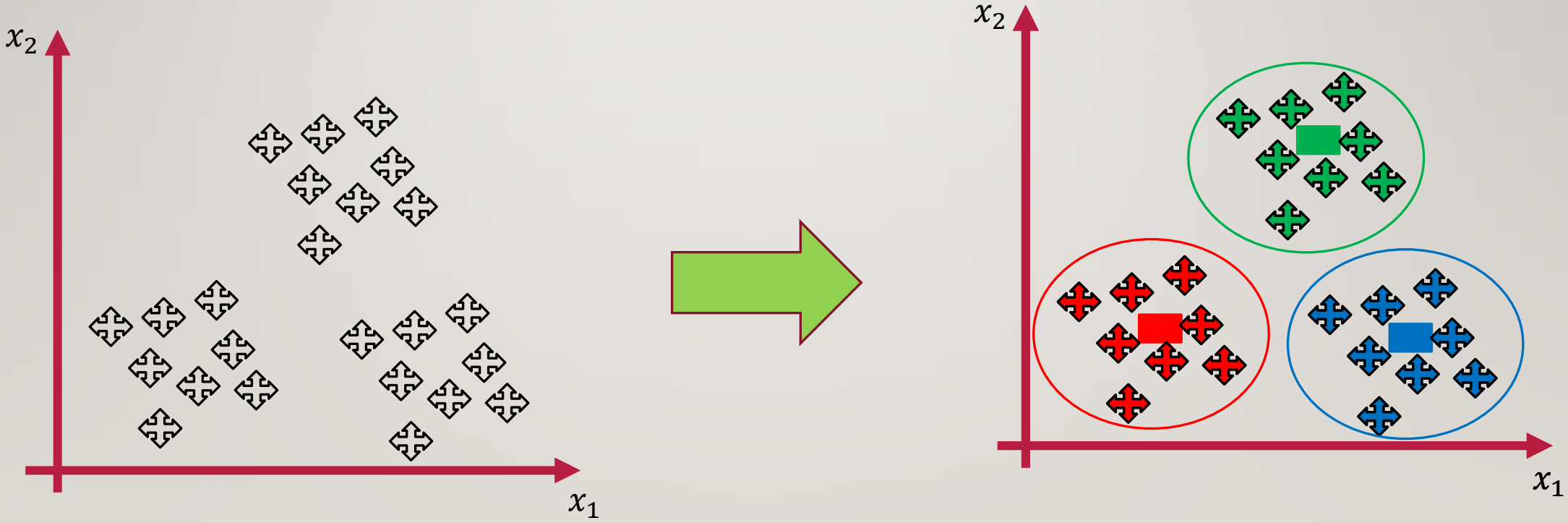


$$\text{WCSS} = \sum_{P_i \text{ in cluster } 1} distance(P_i, C_1)^2 + \sum_{P_i \text{ in cluster } 2} distance(P_i, C_2)^2$$



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

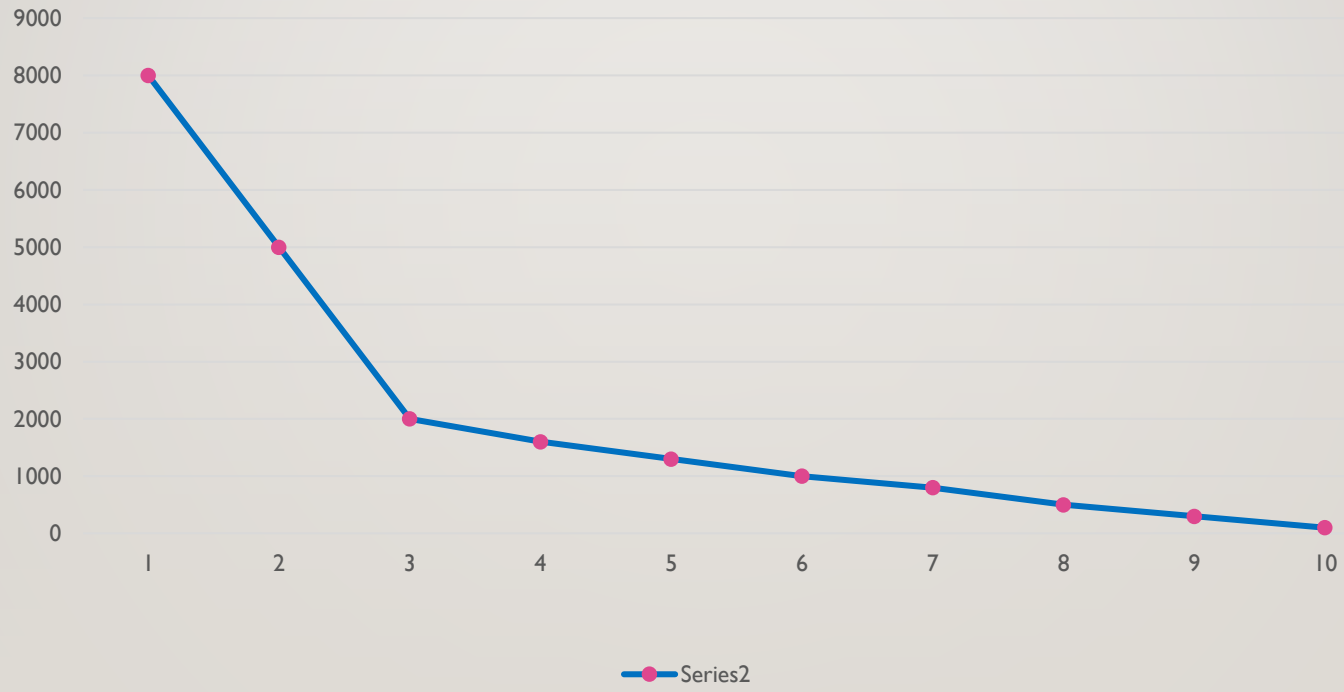
CHOOSING THE RIGHT NUMBER OF CLUSTERS



THE ELBOW METHOD (1/2)

- It is a graph of distortion score (WCSS) vs k
- Distortion can be measured by WCSS as discussed in the previous slides.
- Depending on the distortion appropriate number of clusters k is chosen.

THE ELBOW METHOD (2/2)



ASSUMPTIONS

- If any one of these 3 assumptions is violated, then k-means will fail:
  - k-means assume the variance of the distribution of each attribute (variable) is spherical;
  - All variables have the same variance;
  - the prior probability for all k clusters are the same, i.e. each cluster has roughly equal number of observations
- Clusters in K-means are defined by taking the mean of all the data points in the cluster.

ADVANTAGES

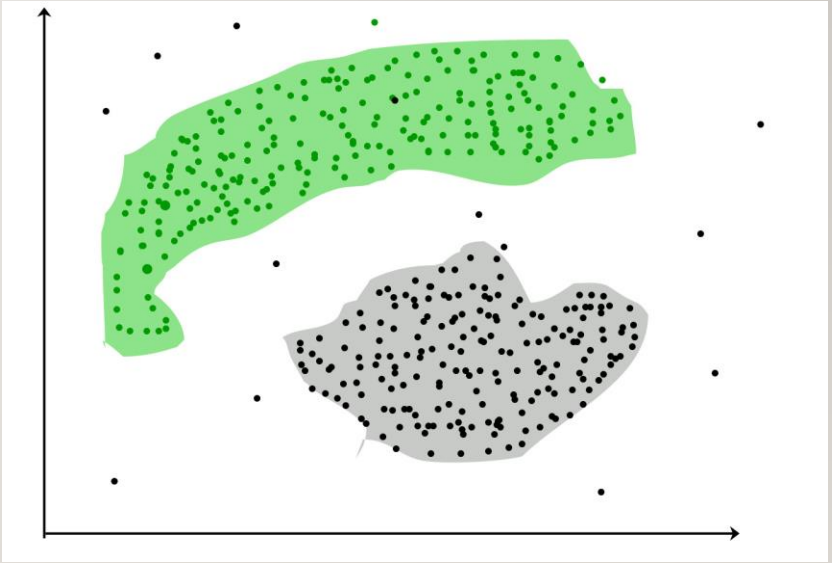
- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

DISADVANTAGES

- The user has to specify k (the number of clusters) in the beginning.
- k-means can only handle numerical data.
- Being dependent on initial values.
- Clustering data of varying sizes and density.
- Clustering outliers.
- Scaling with number of dimensions.

APPLICATIONS

- **Marketing :** It can be used to characterize & discover customer segments for marketing purposes.
- **Biology :** It can be used for classification among different species of plants and animals.
- **Libraries :** It is used in clustering different books on the basis of topics and information.
- **Insurance :** It is used to acknowledge the customers, their policies and identifying the frauds.



THANK YOU