Clustering

Clustering

What is Clustering?

Clustering – grouping
Unlabelled Data

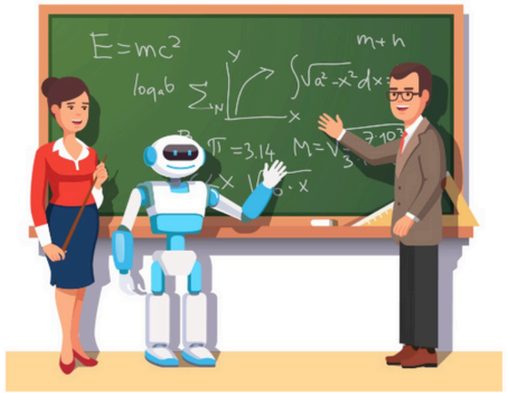
https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

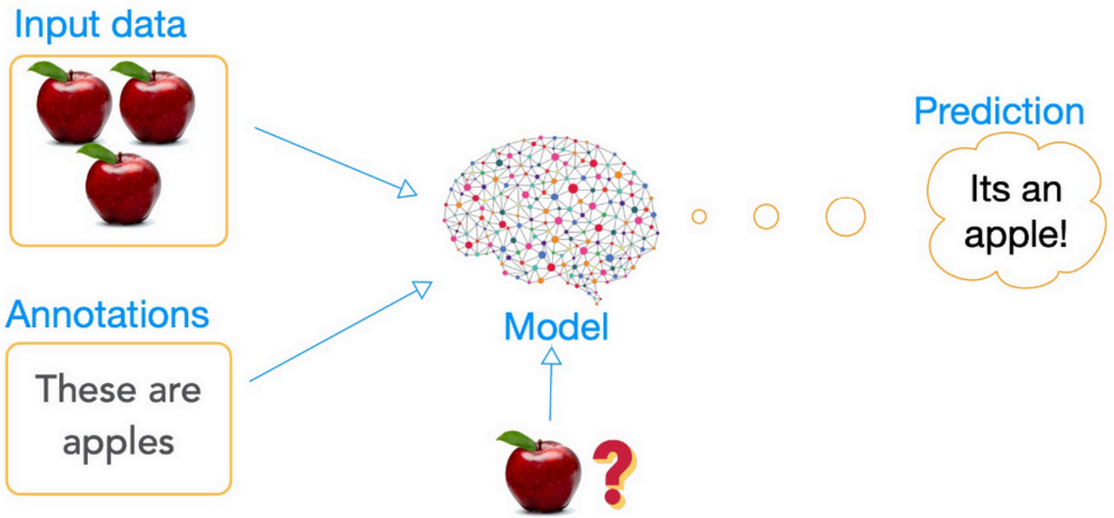
Clustering

أكاديمية طويق
Tuwaiq Academy

What is Clustering?

Supervised Learning
(e.g. Regression, Classification)





Unsupervised Learning
(e.g. Clustering)

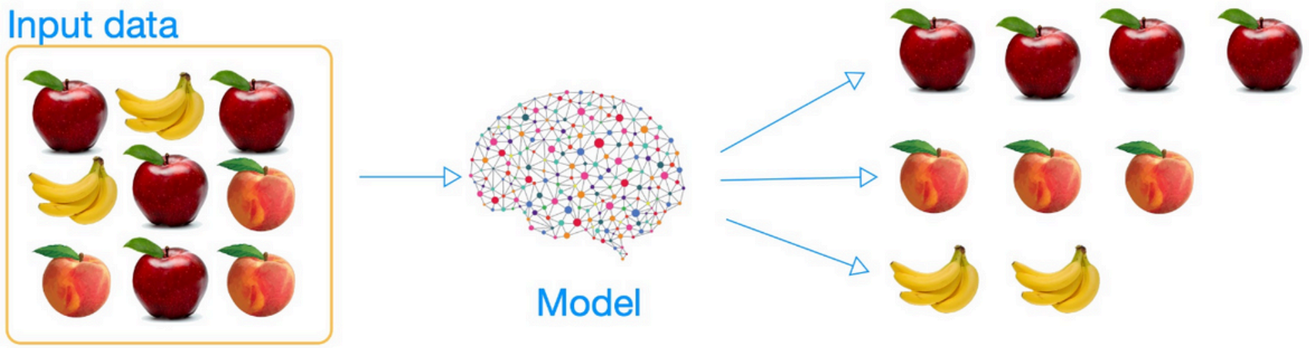


Image source: mdpi.com/2073-8994/10/12/73

Clustering

أكاديمية طويق
Tuwaiq Academy

What is Clustering?

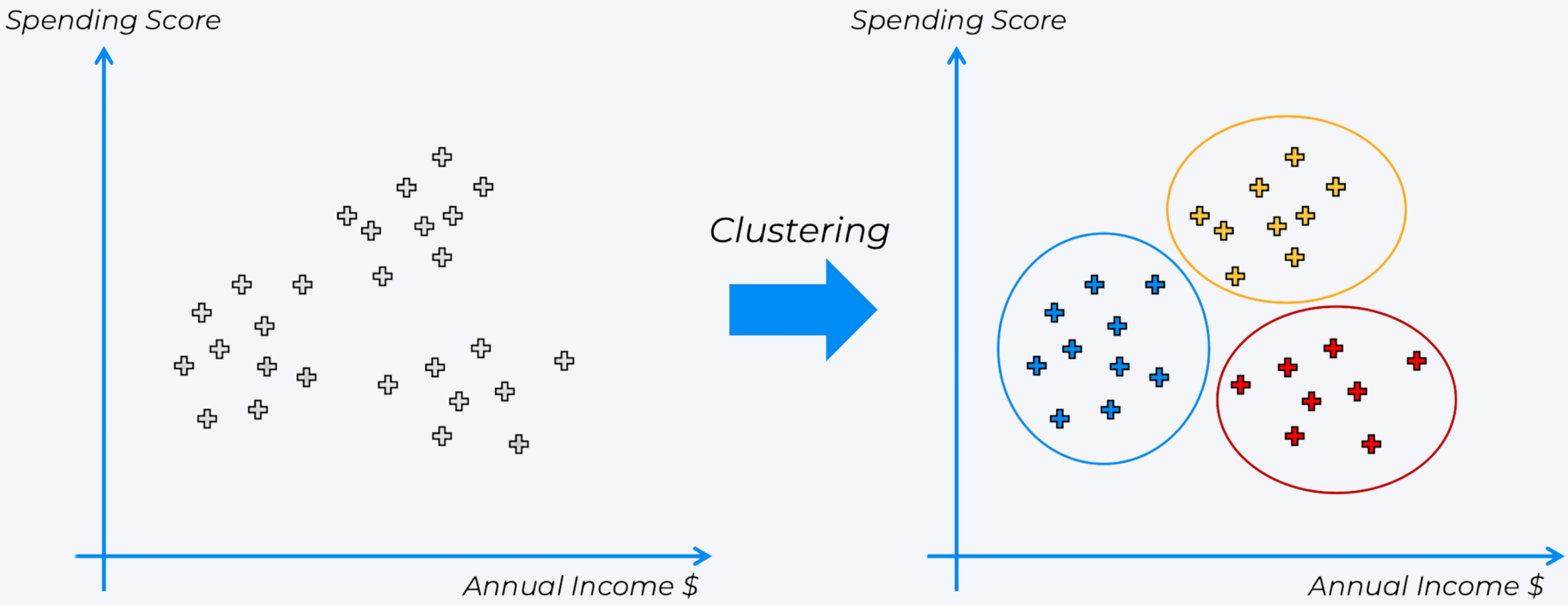


Image source: mdpi.com/2073-8994/10/12/73

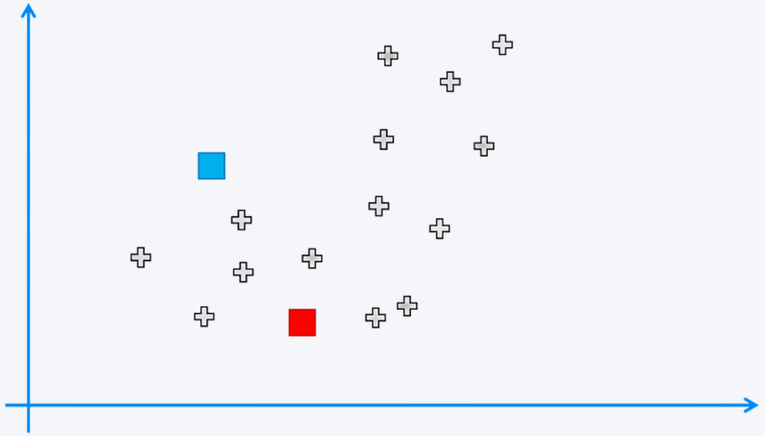
K-Means Clustering

K-Means Clustering

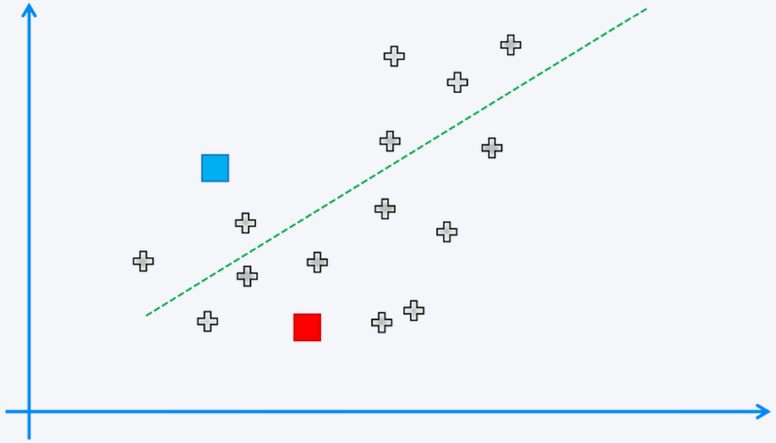
أكاديمية طويق
Tuwaiq Academy

The Intuition Behind K-Means Clustering

No labels data
How many clusters ?
Let consider two



Replace them randomly , any point (centroid)



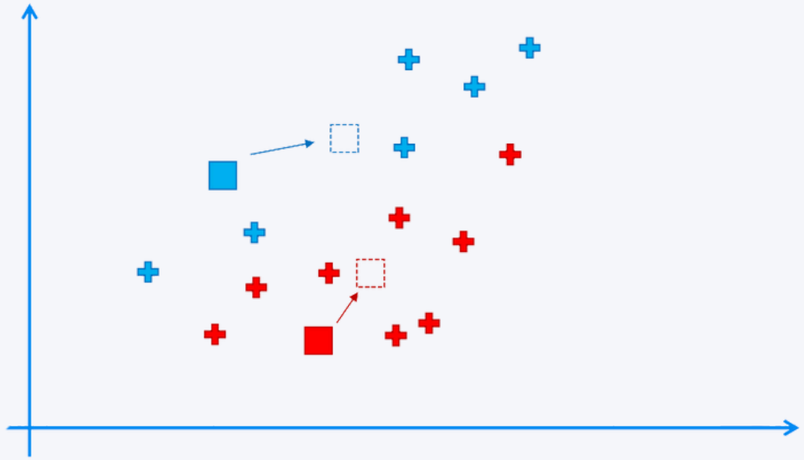
K-means assign distance line to the closest these centroid , easiest way like this

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

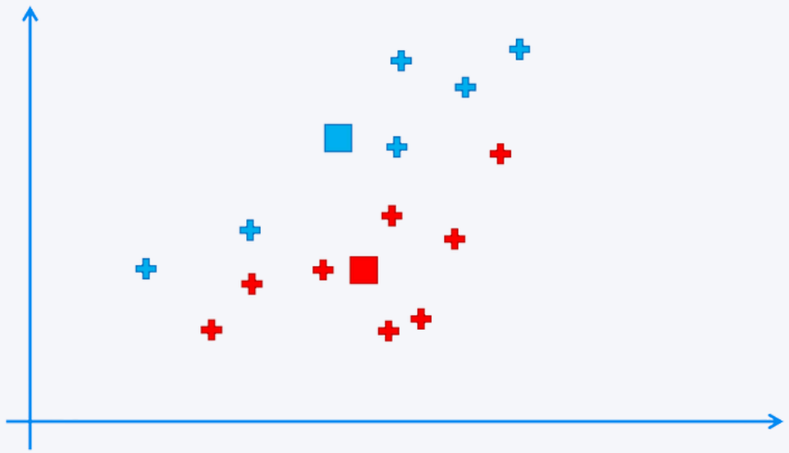
أكاديمية طويق
Tuwaiq Academy

The Intuition Behind K-Means Clustering

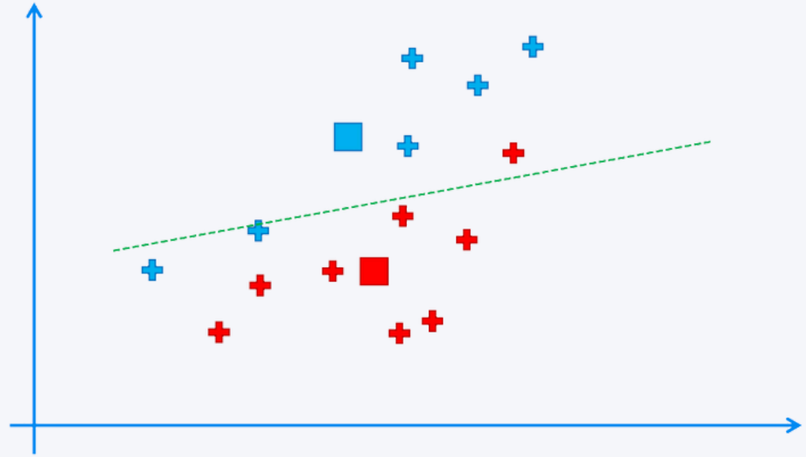


Calculate mass or gravity of each clusters ..

For take all x and get average and for blue take all y and get the average so you will find the cluster mass



move the centriod to new positions and repeat the process



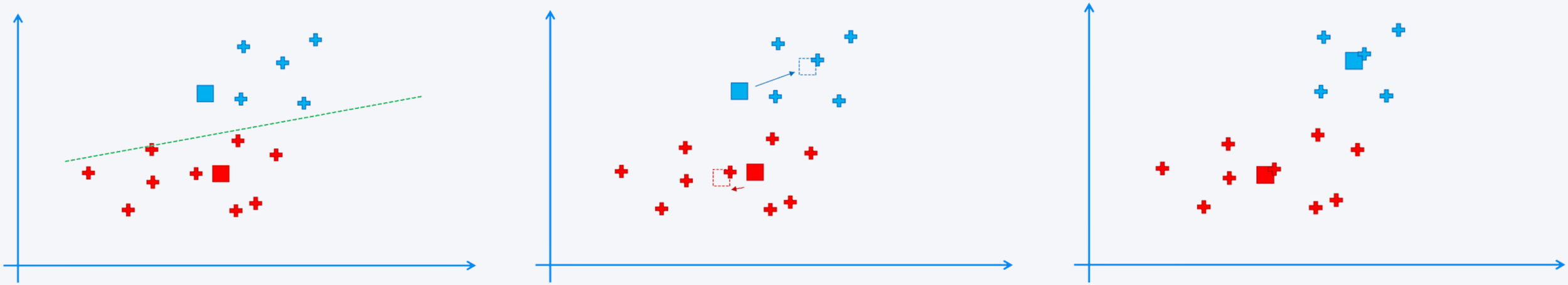
Place the line to the closest centroid

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

أكاديمية طويق
Tuwaiq Academy

The Intuition Behind K-Means Clustering



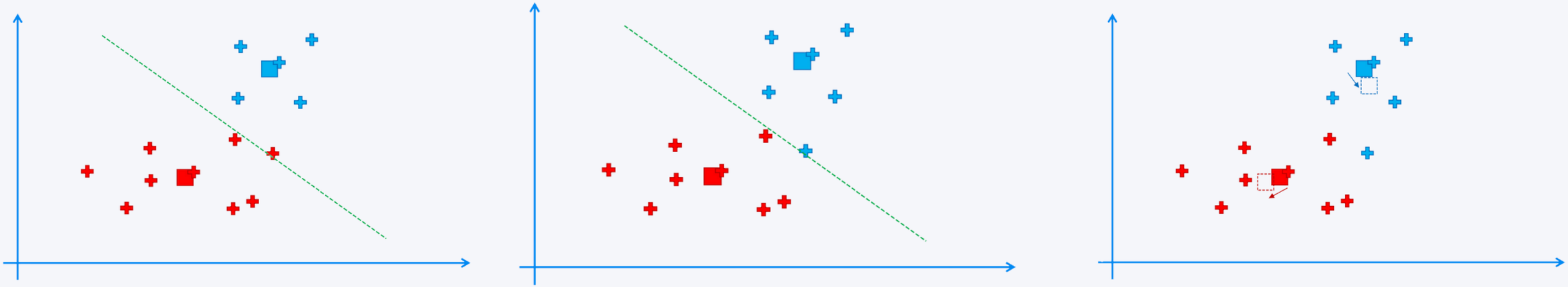
Repeat the process ..

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

أكاديمية طويق
Tuwaiq Academy

The Intuition Behind K-Means Clustering

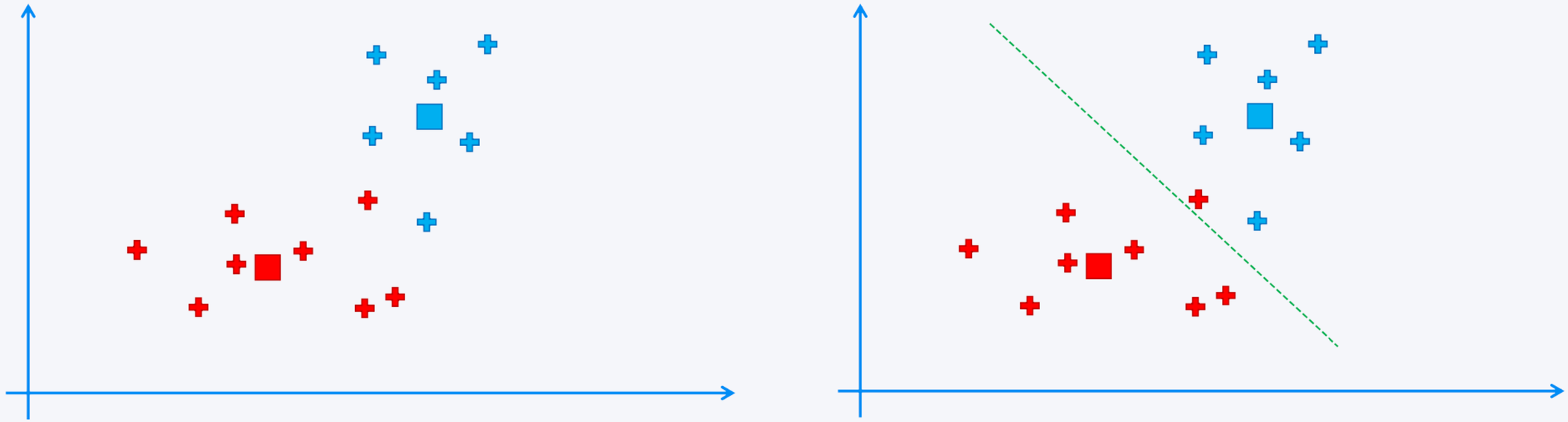


Repeat the process ..

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

The Intuition Behind K-Means Clustering



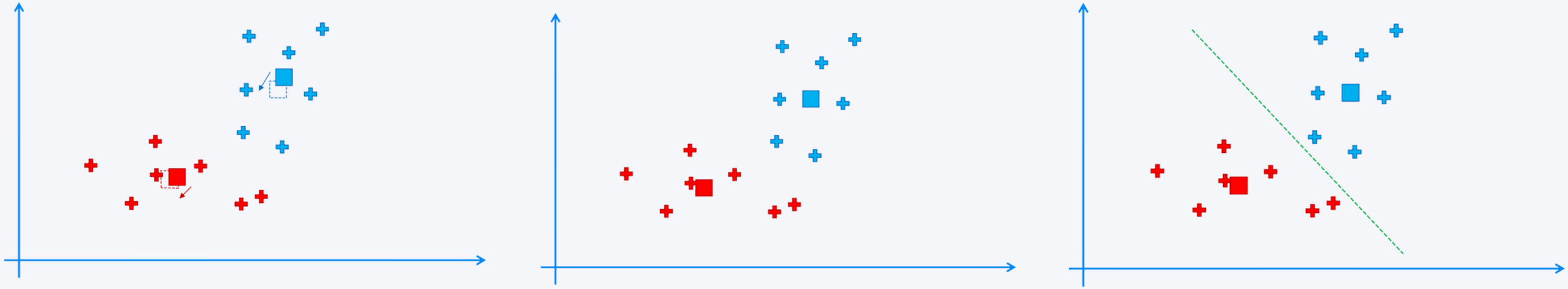
Repeat the process

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

أكاديمية طويق
Tuwaiq Academy

The Intuition Behind K-Means Clustering

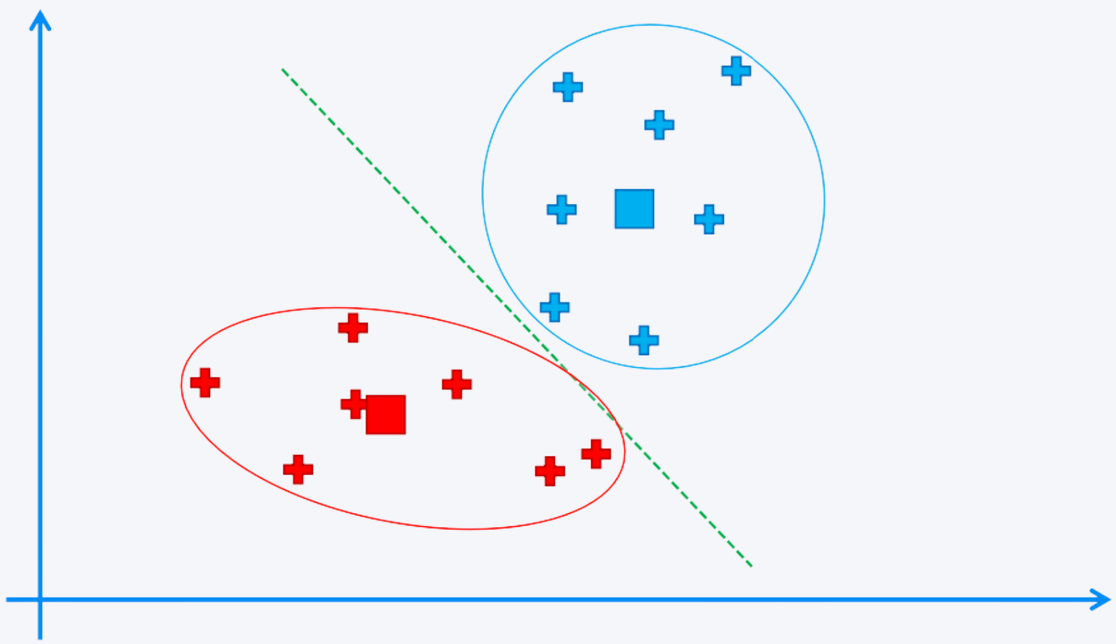


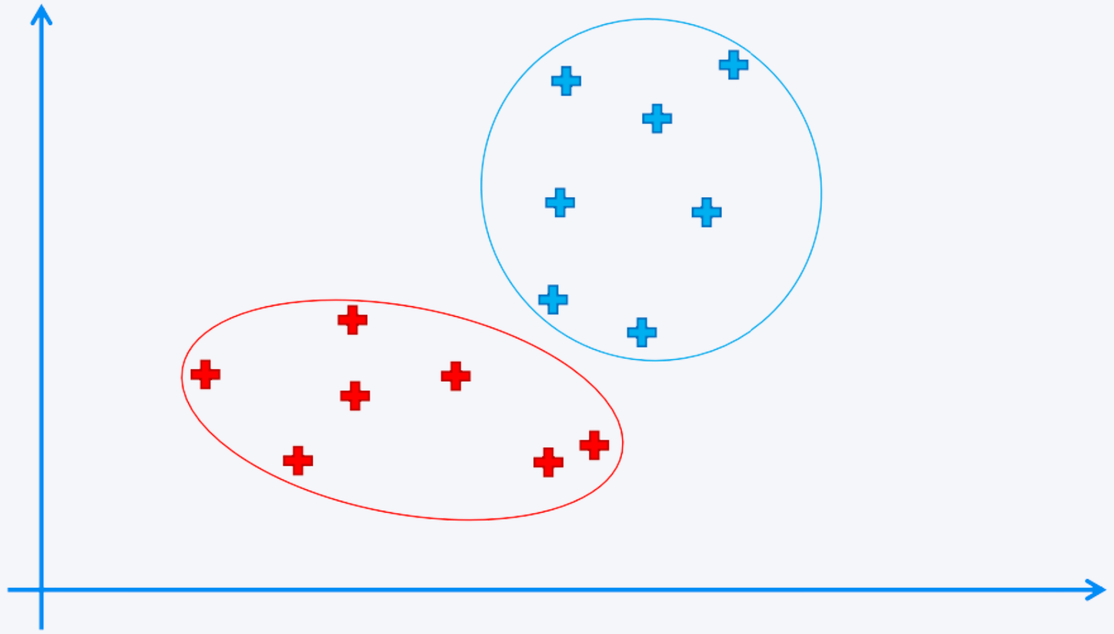
Repeat until the cluster wont change

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

The Intuition Behind K-Means Clustering





So Here is our final clusters

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

أكاديمية طويق
Tuwaiq Academy

The question is how many clusters should be/should we have !

- Sometimes from the knowledge domain
- Sometimes we need to know how many we should have

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

The Elbow Method

Within Cluster Sum of Squares:

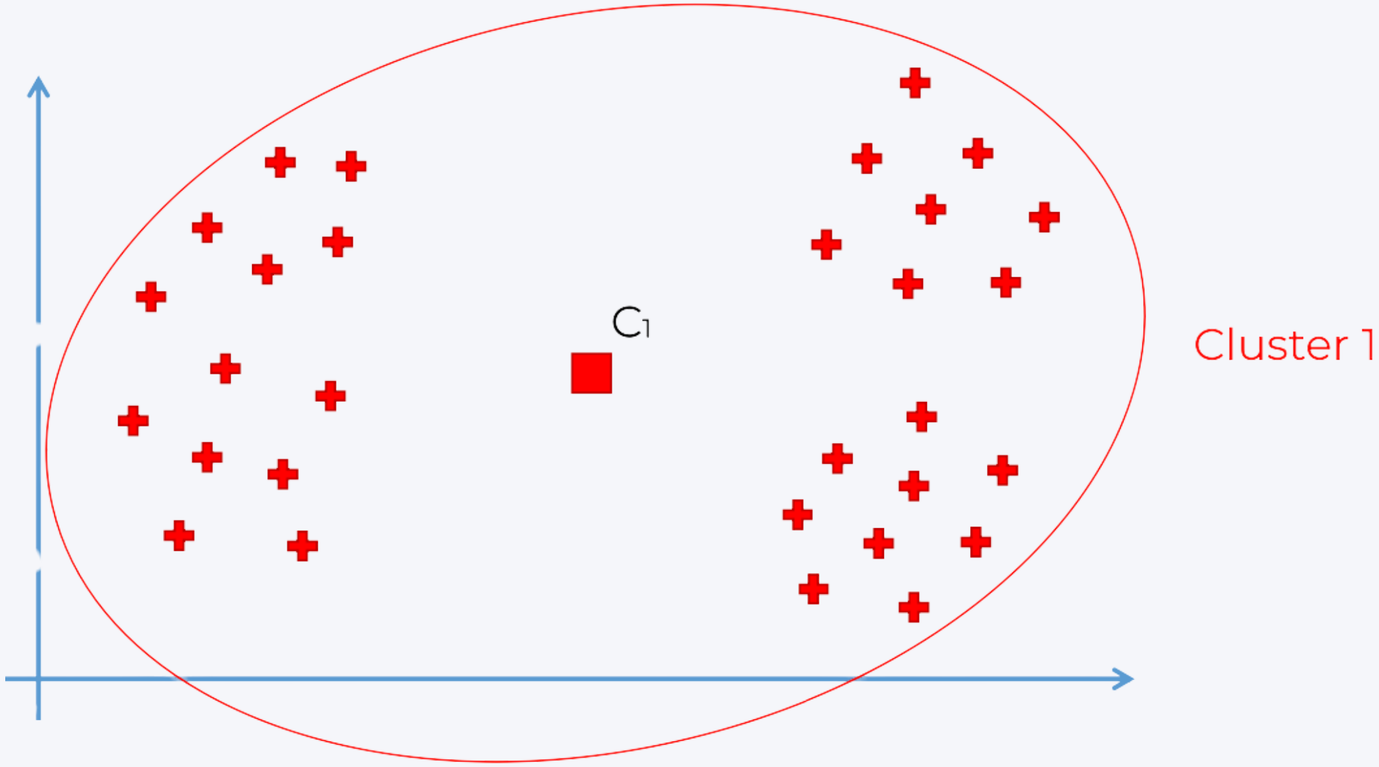
$$\text{WCSS} = \sum_{P_i \text{ in Cluster } 1} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster } 2} \text{distance}(P_i, C_2)^2 + \ldots$$

It basically looks at the distance between each point and the centroid and square it

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

The Elbow Method



Within Cluster Sum of Squares:

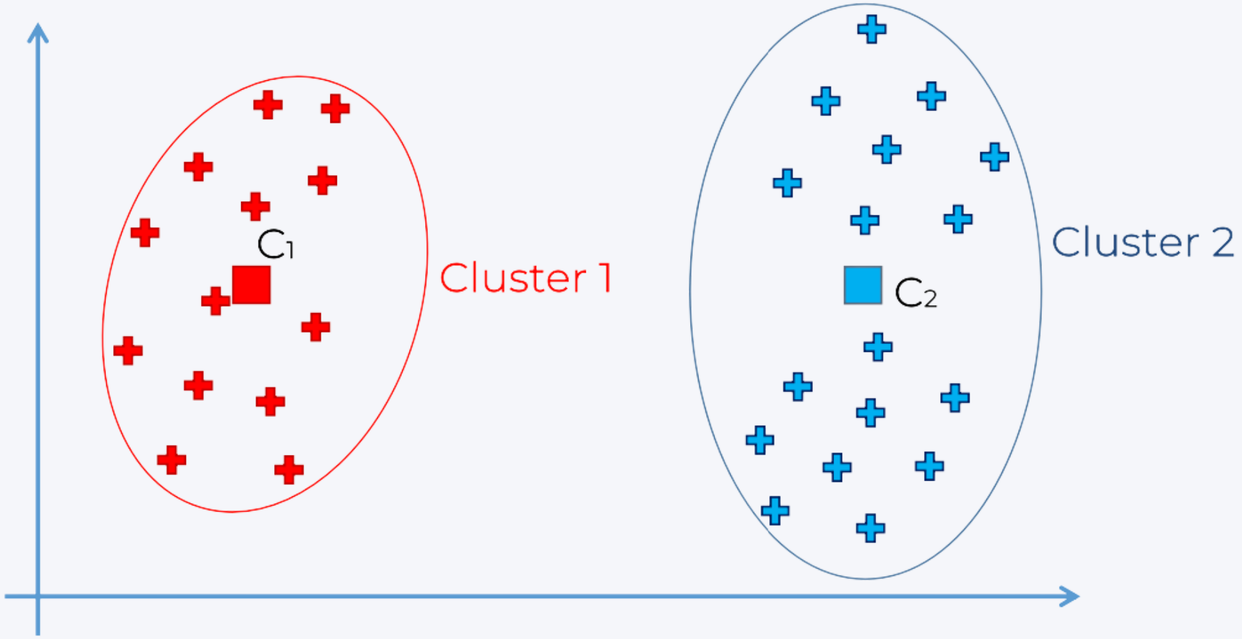
$$\text{WCSS} = \sum_{P_i \text{ in Cluster } 1} \text{distance}(P_i, C_1)^2$$

In this example, measure each points with the centroid and square it to find the WCSS

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

The Elbow Method



Within Cluster Sum of Squares:

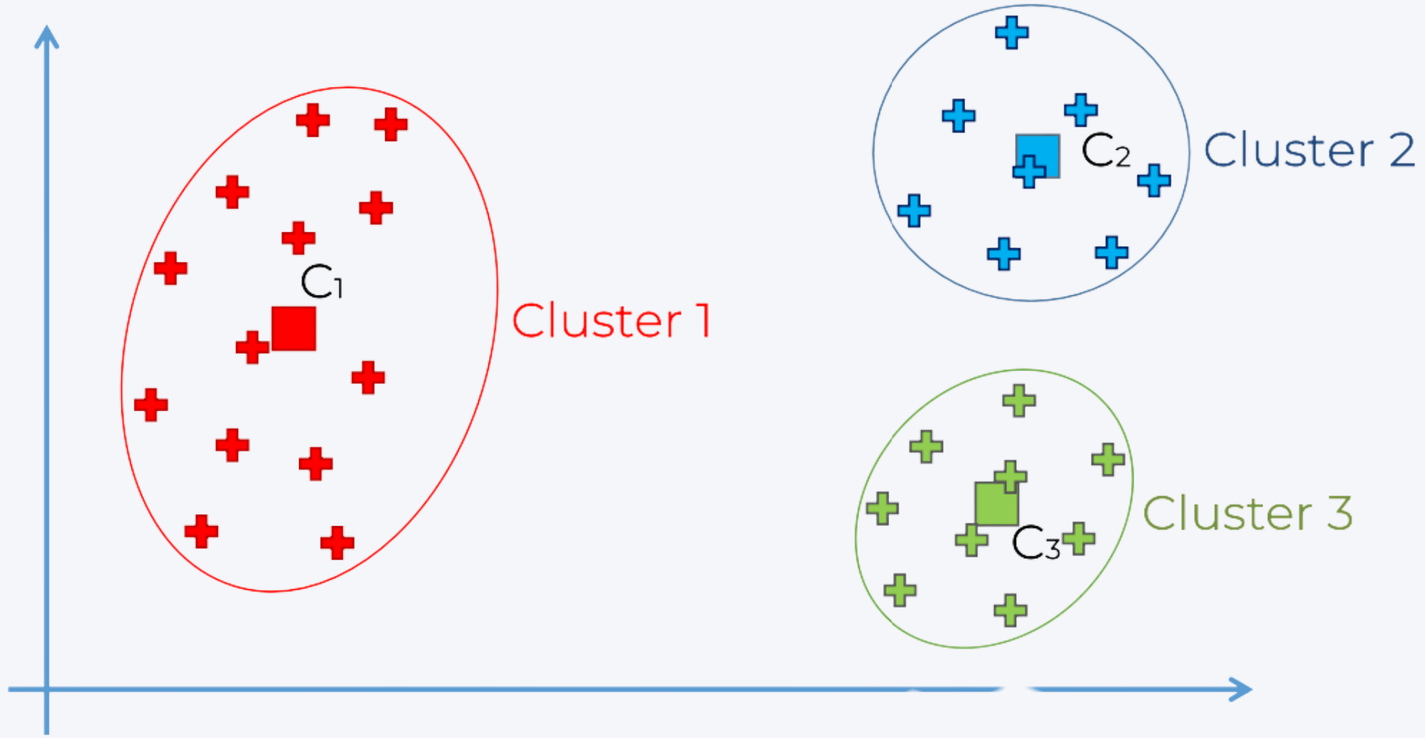
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \ldots$$

Case if we have two clusters

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

The Elbow Method



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

Case if we have three clusters and so on ...

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

The Elbow Method

- If you notice, we need the clusters to find WCSS .. so we need to run it first
- It is **backward** !!
- The more clusters we have, the smaller WCSS we have .. So we can continue increasing numbers of clusters until we get max number of clusters = number of points so WCSS will become 0
- The optimal number of clusters is three because it is the elbow , where WCSS stop dropping as rapidly.

Judgment call !! sometimes not clear, more than candidate .. so

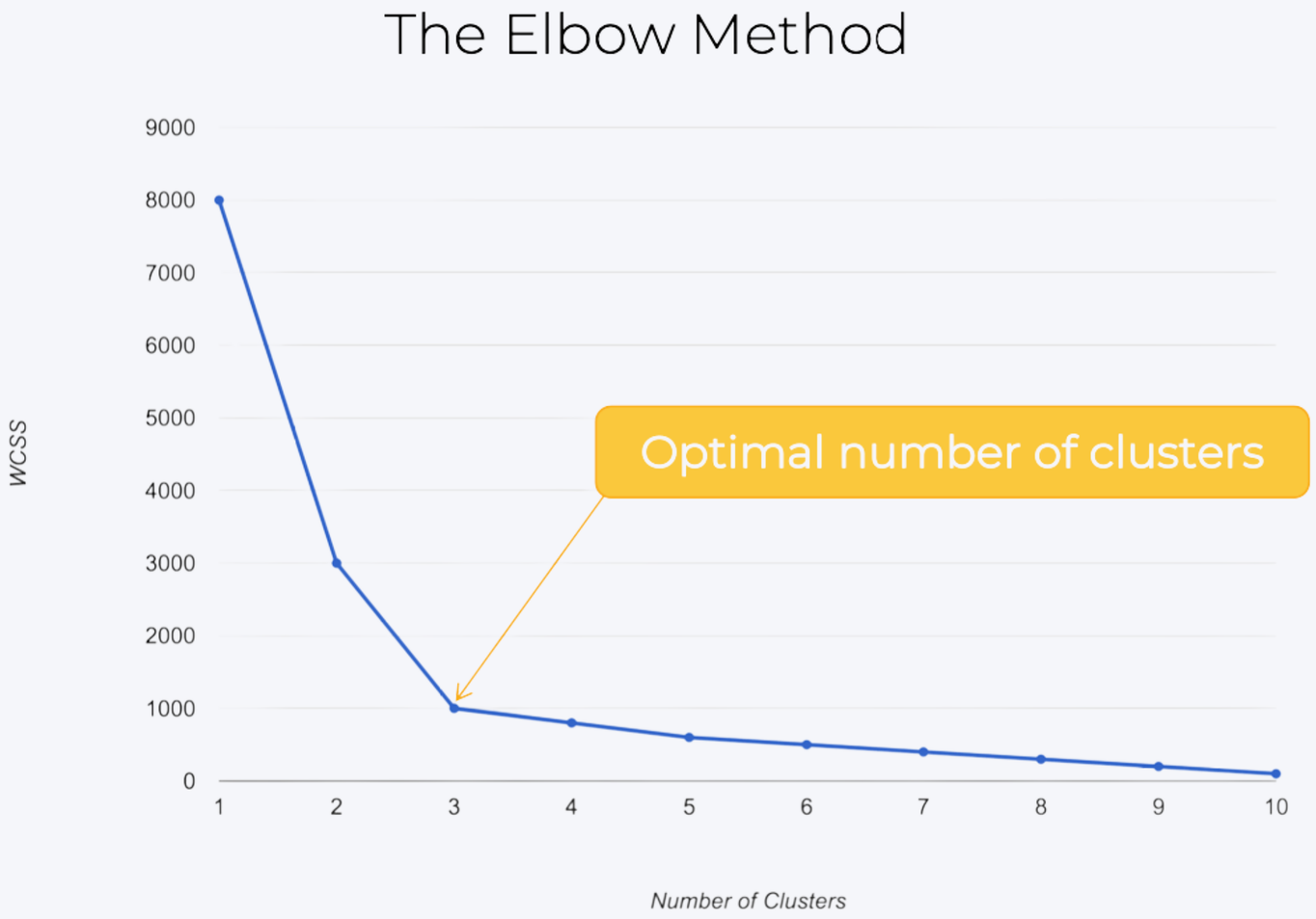


Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

K-Means++

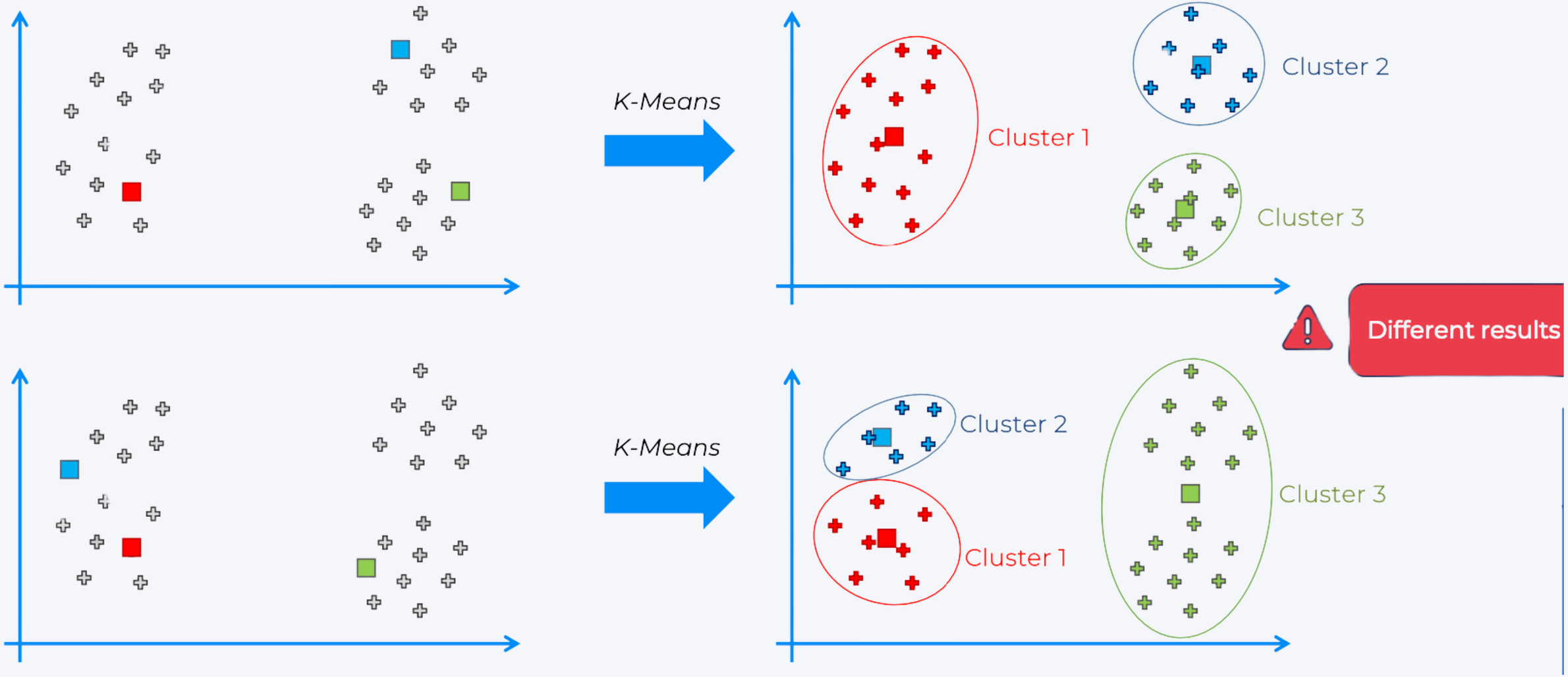


Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

أكاديمية طويق
Tuwaiq Academy

K-Means++

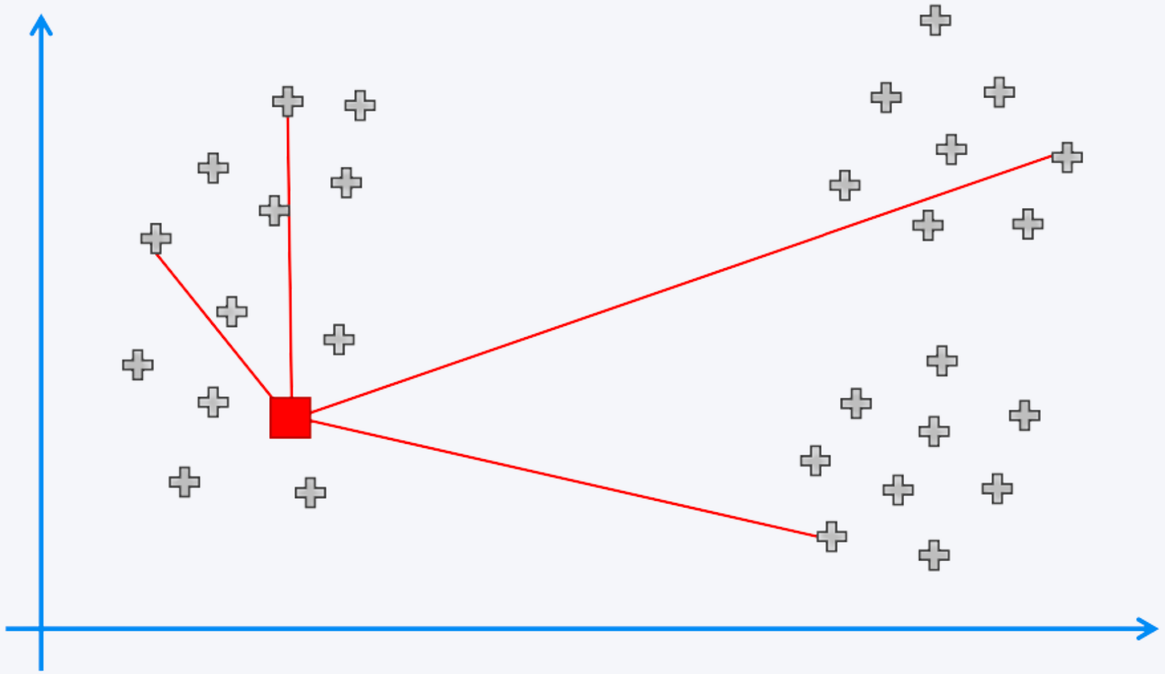
K-Means++ Initialization Algorithm:

- **Step 1:** Choose first centroid at random among data point
- **Step 2:** For each of the remaining data points compute the distance (D) to the nearest out of already selected centroids
- **Step 3:** Choose next centroid among remaining data points using weighted random selection – weighted by D2
- **Step 4:** Repeat Steps 2 and 3 until all k centroids have been selected
- **Step 5**: Proceed with standard k-means clustering

Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

K-Means++



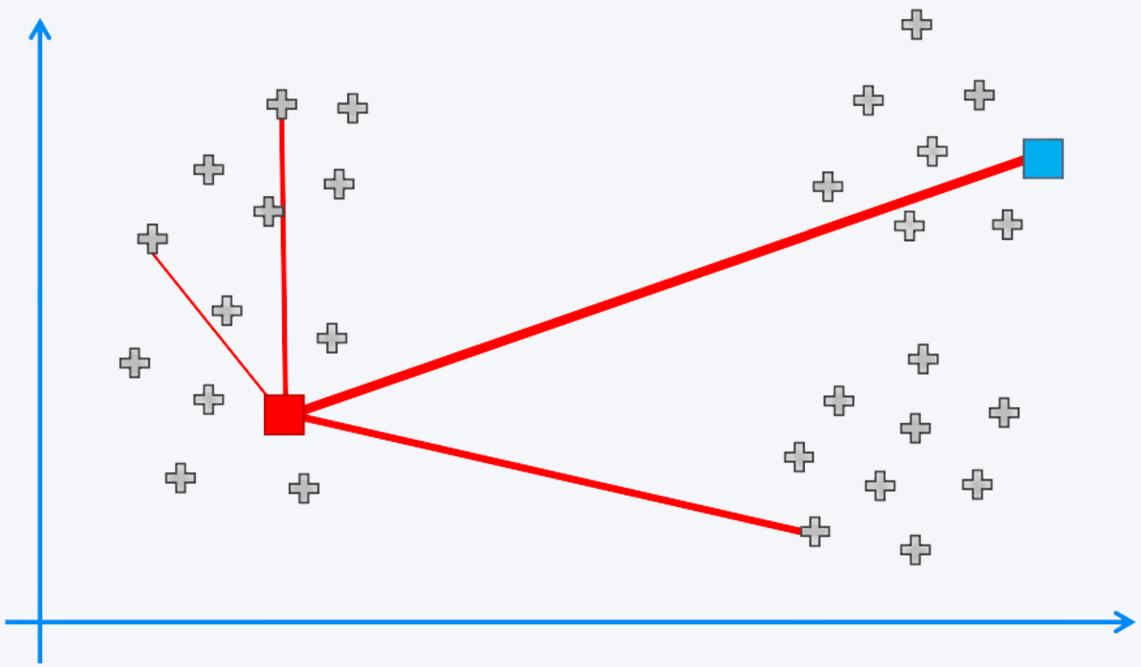


Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

K-Means++

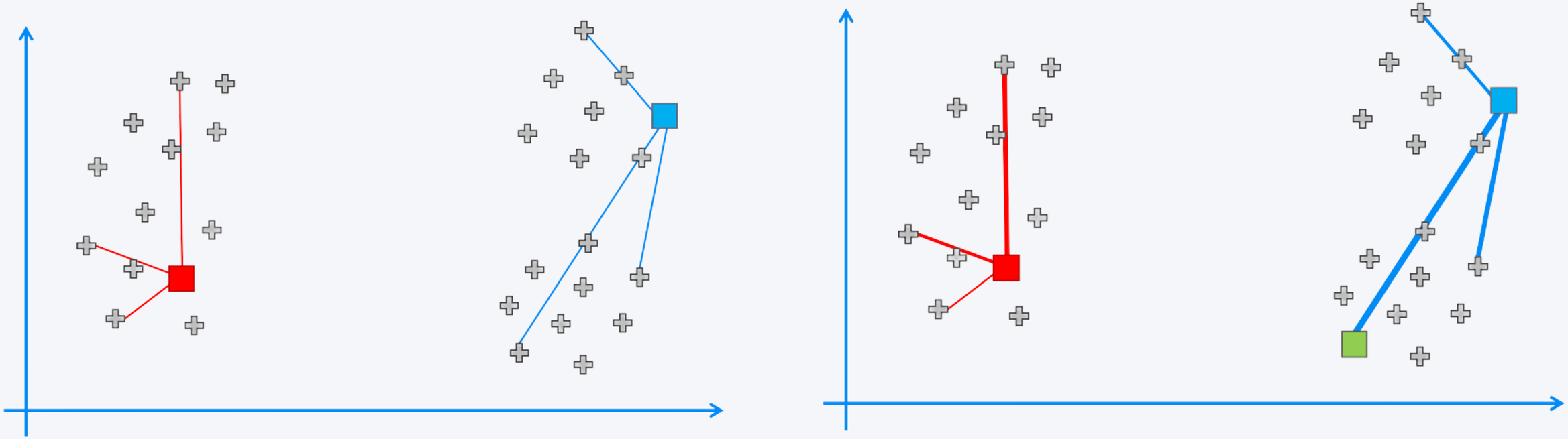
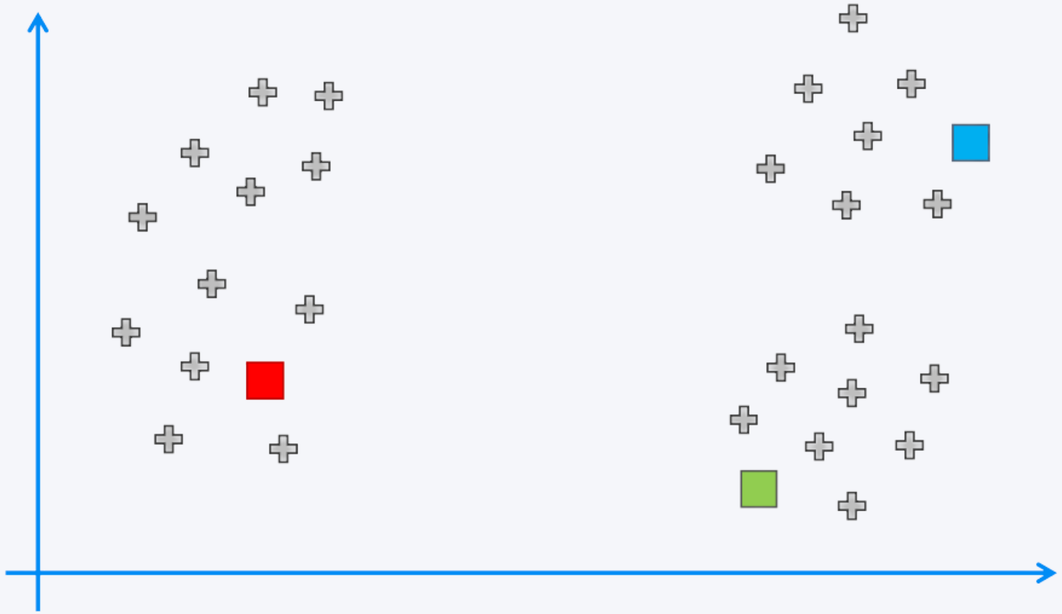


Image source: mdpi.com/2073-8994/10/12/73

K-Means Clustering

أكاديمية طويق
Tuwaiq Academy

K-Means++



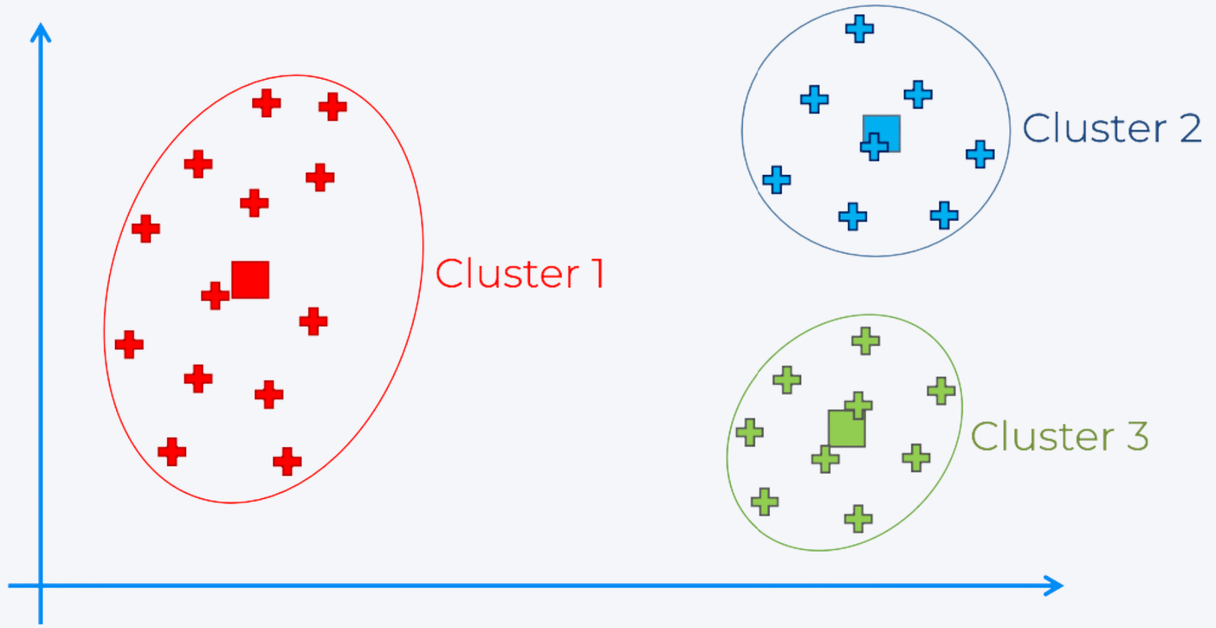


Image source: mdpi.com/2073-8994/10/12/73