Lecture 4: Clustering

PUSL3123 AI and Machine Learning

Neamah Al-Naffakh

School of Engineering, Computing and Mathematics

Neamah.al-naffakh@plymouth.ac.uk

Today’s Topics

Clustering

Lesson learning outcomes: By the end of today’s lesson, you would be able to:

- Understand the concept of clustering
- **Introduce K-Means algorithm.**
- **Implement K-Means**

UNIVERSITY OF PLYMOUTH
School of Engineering, Computing and Mathematics

What is Clustering?

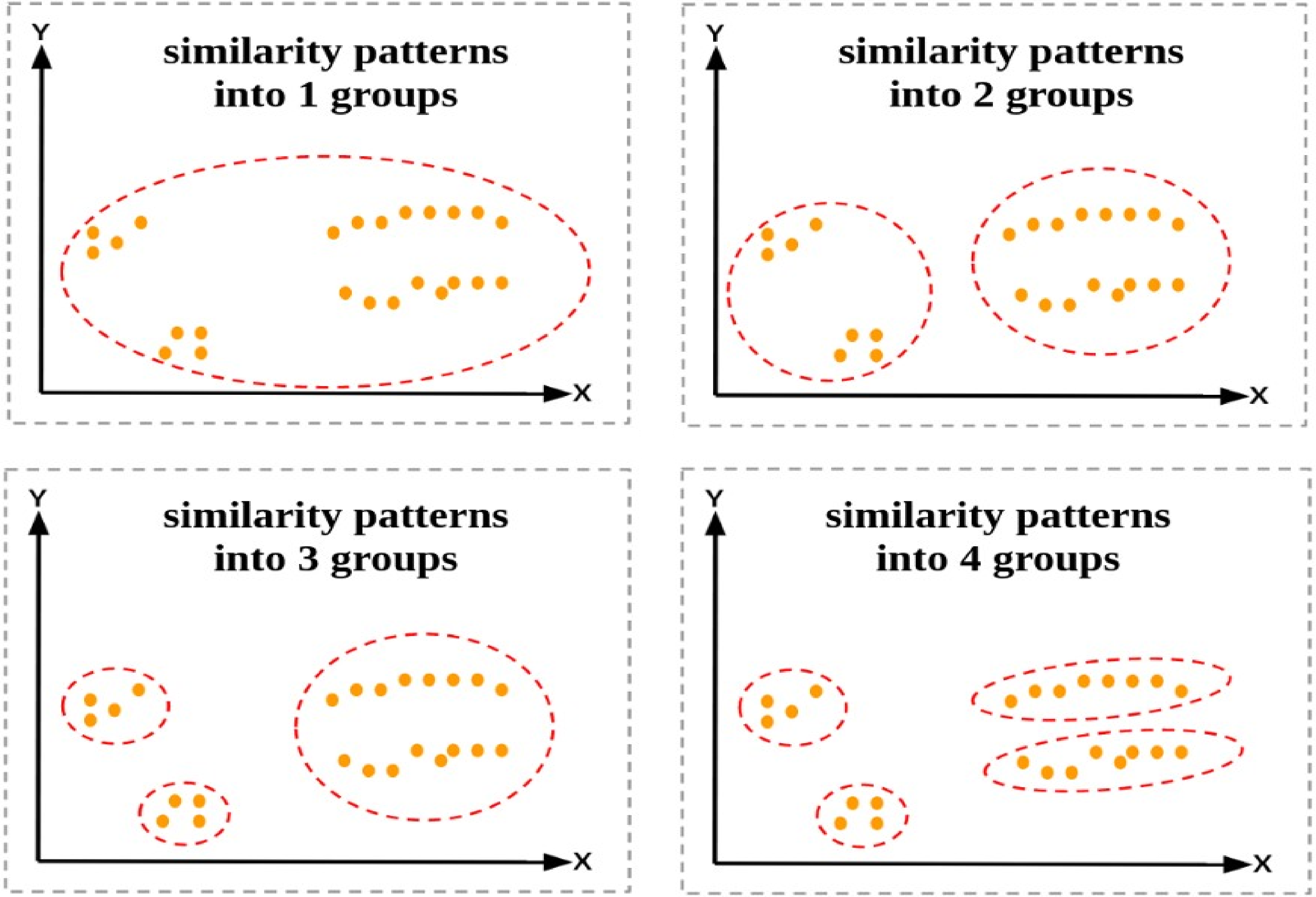
- In everyday terms, **clustering** refers to the grouping together of objects with similar characteristics in other words, the aim is to segregate groups with similar traits and assign them into clusters)

- When it comes to data and data mining, the method of identifying similar groups of data in a data set is called clustering

- It is an unsupervised learning method

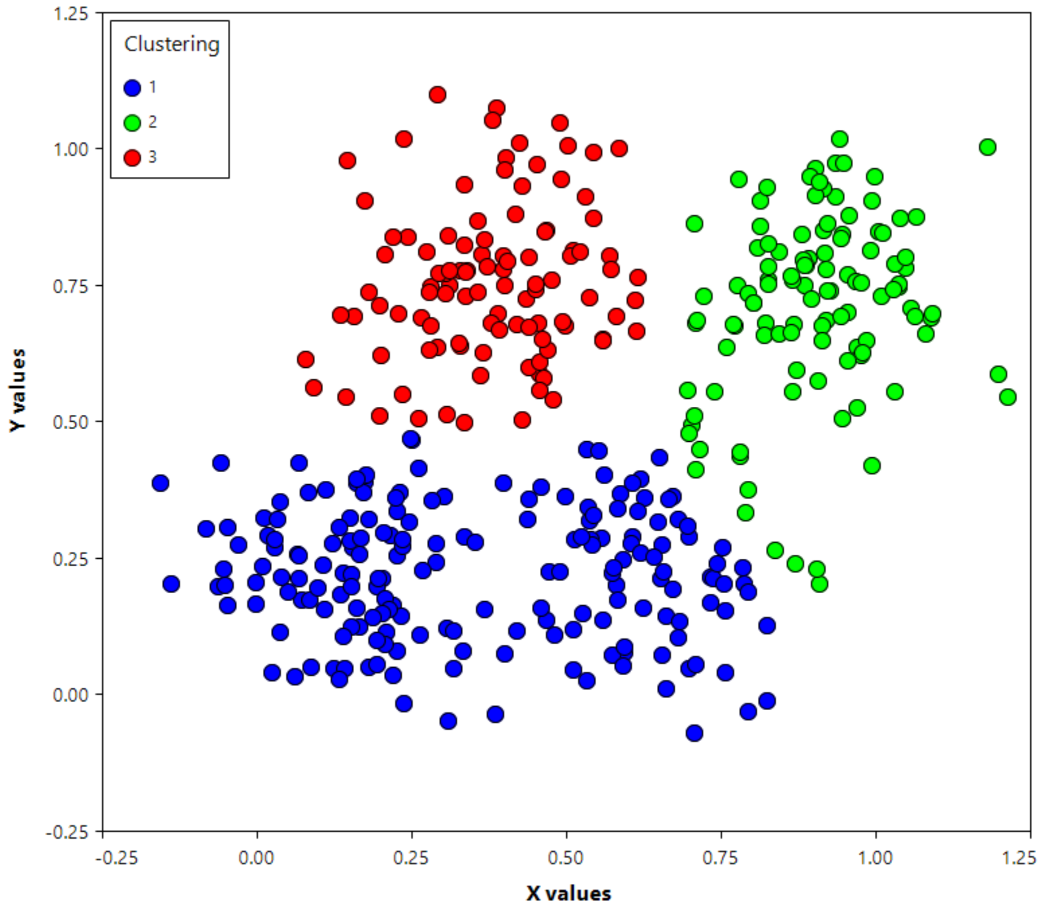
UNIVERSITY OF PLYMOUTH
School of Engineering, Computing and Mathematics

What is Clustering?

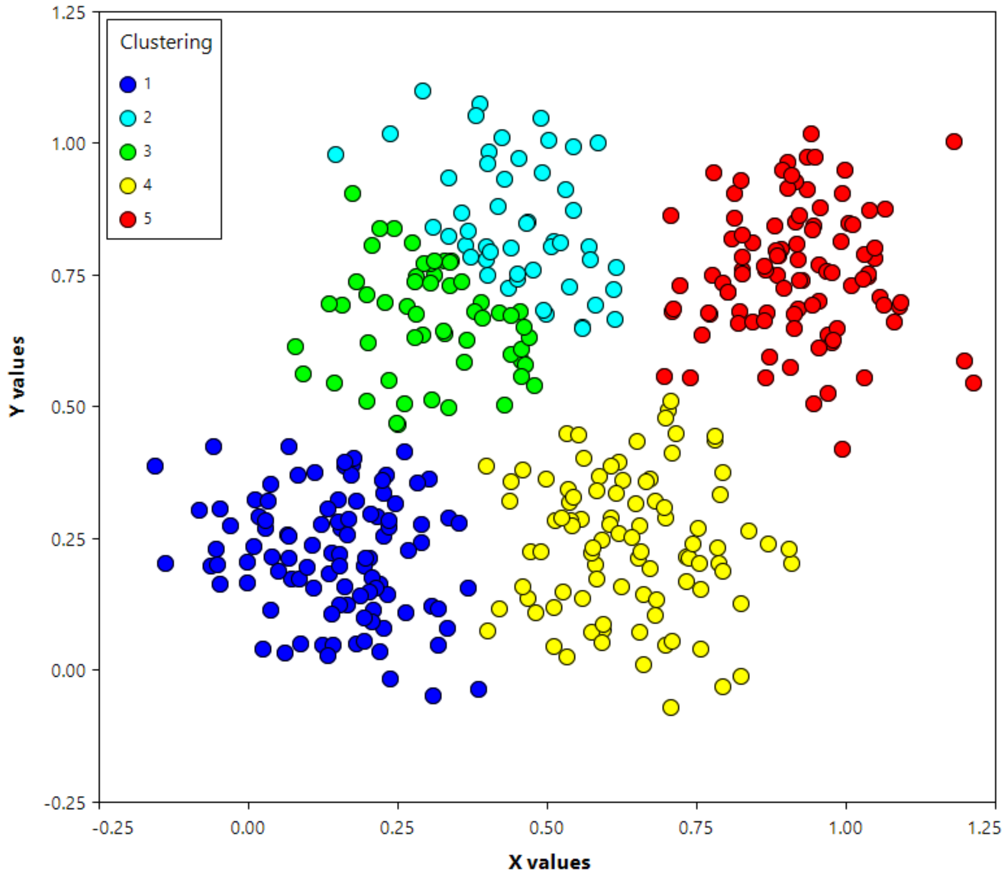


UNIVERSITY OF PLYMOUTH
School of Engineering, Computing and Mathematics

What is Clustering?



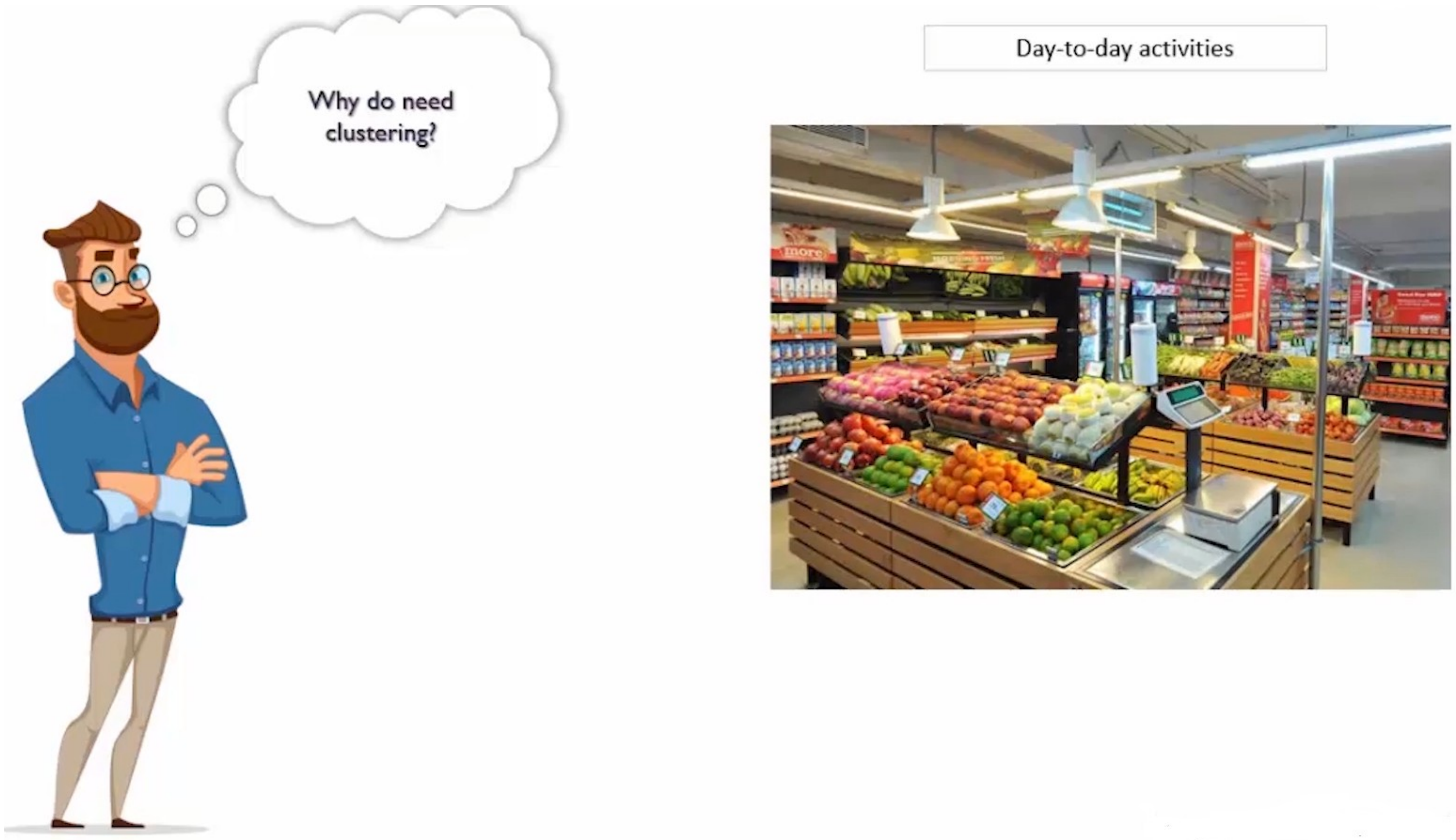
Grouping in 3 clusters



Grouping in 5 clusters

UNIVERSITY OF PLYMOUTH
School of Engineering, Computing and Mathematics

Clustering Example



Common uses of Clustering

- Recommendation engines
- Market Segmentation
- Statistical data analysis
- Social network analysis

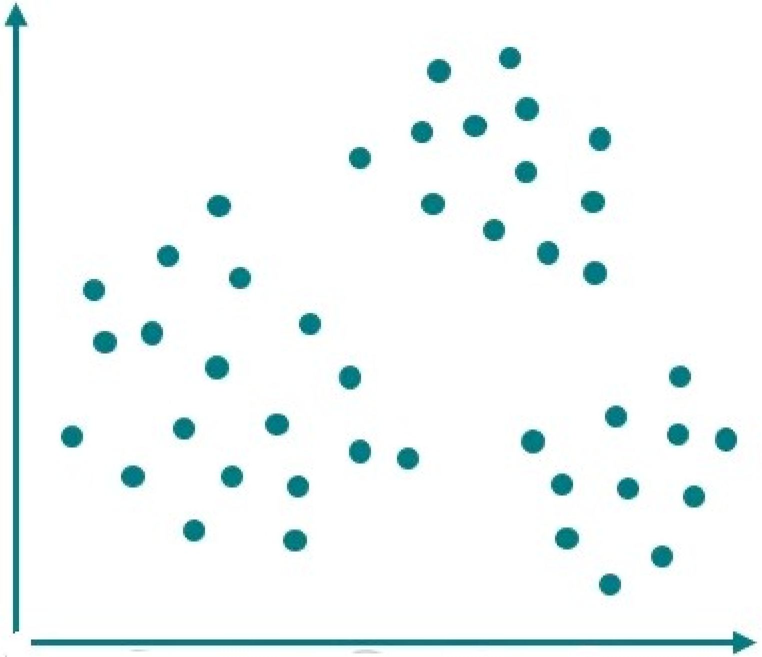
K-Means Clustering

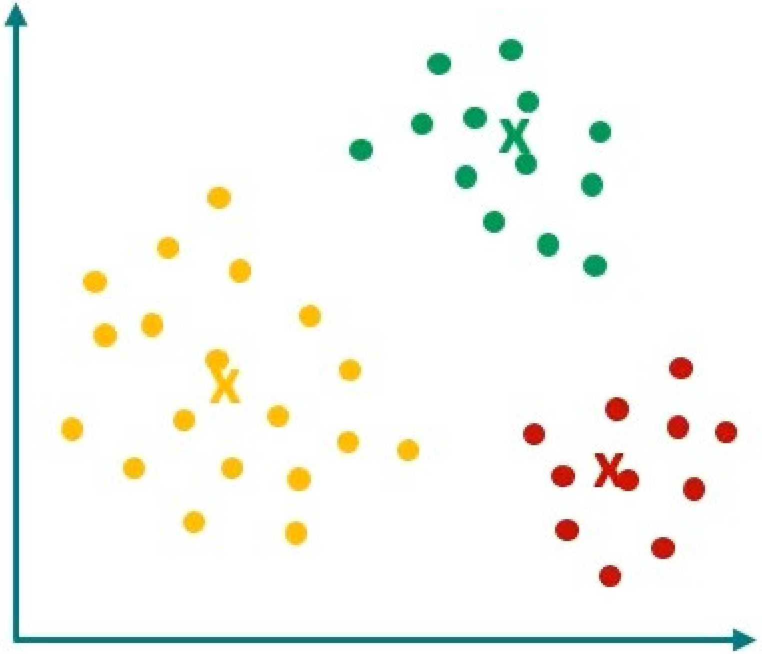
- Old technique, still very popular in use
- Main goal is to group similar data point into one cluster
- Number of clusters are represented by ***K***

- Strengths
  - Simple iterative method
  - It can also scale to large datasets
- Weaknesses
  - Poor performance (local optimum)
  - Difficult to guess the correct "K"

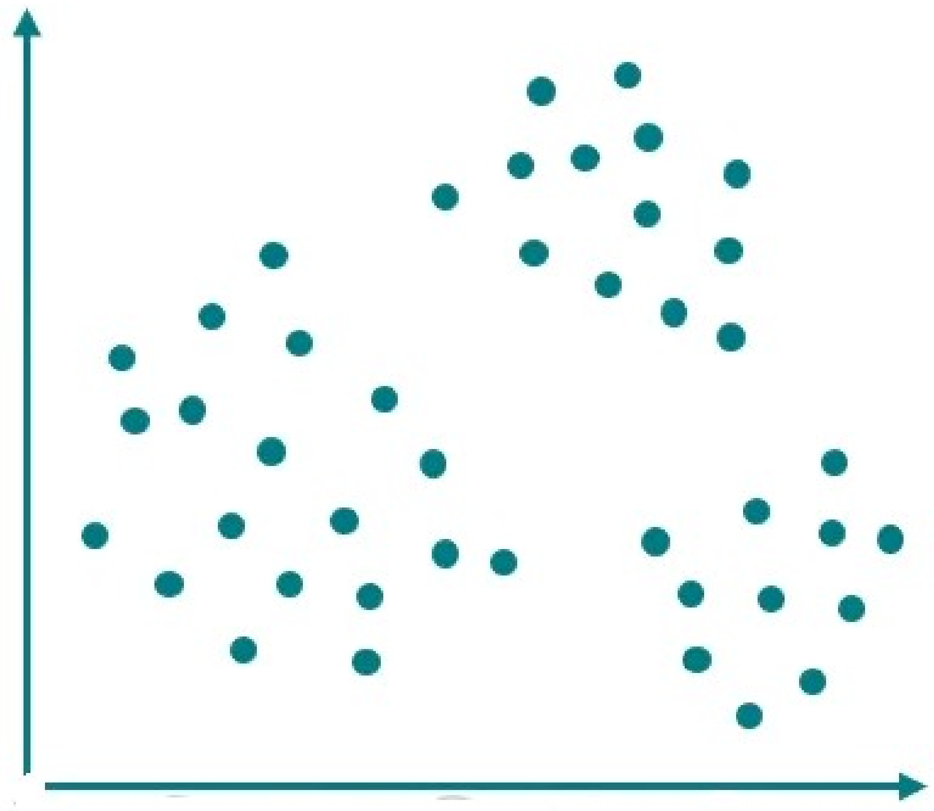
How does K-Means Clustering work?

1. Define number of clusters
2. Randomly set cluster centers (i.e., cluster centroid)
3. **Assign points to clusters**
4. **Re-calculate center of each cluster**
5. **Assign points to the new clusters**



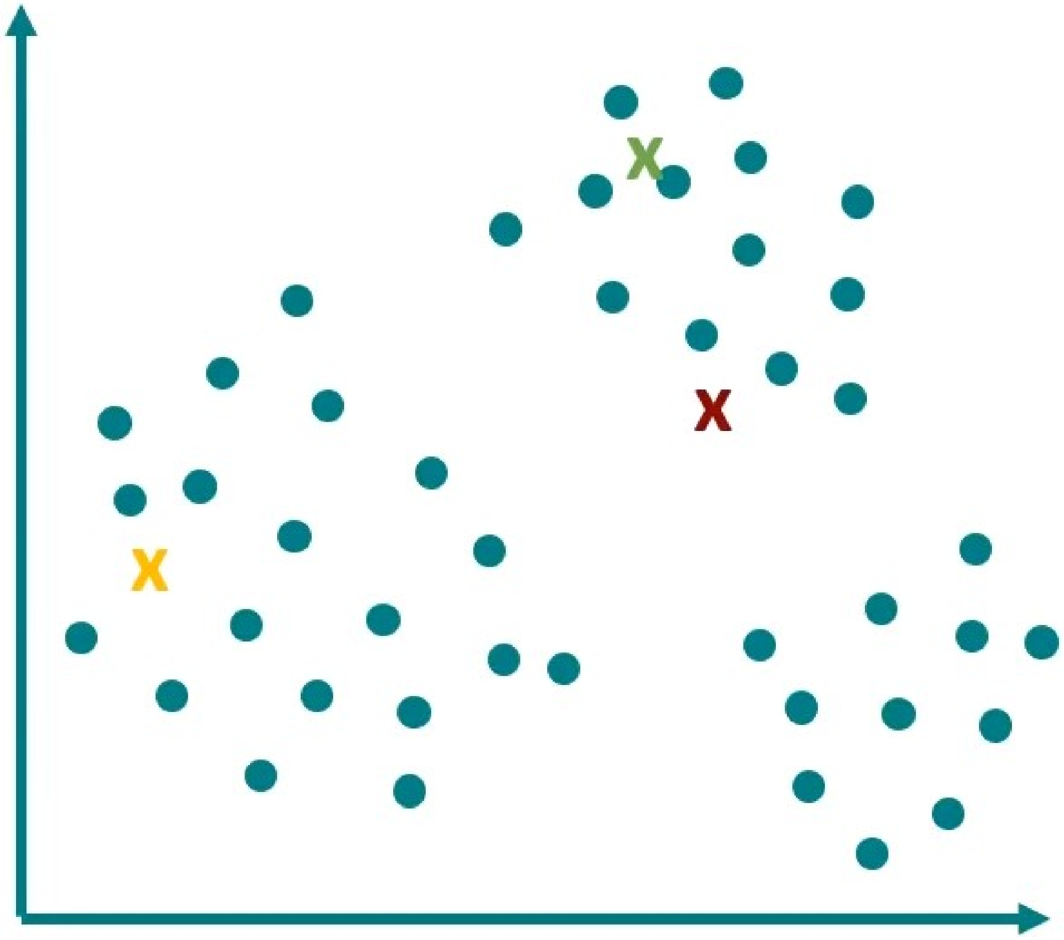


1. Define number of clusters



- Number of clusters in K-Means is ***K*** .
- Let us choose ***K***= 3

2. Randomly set cluster centers

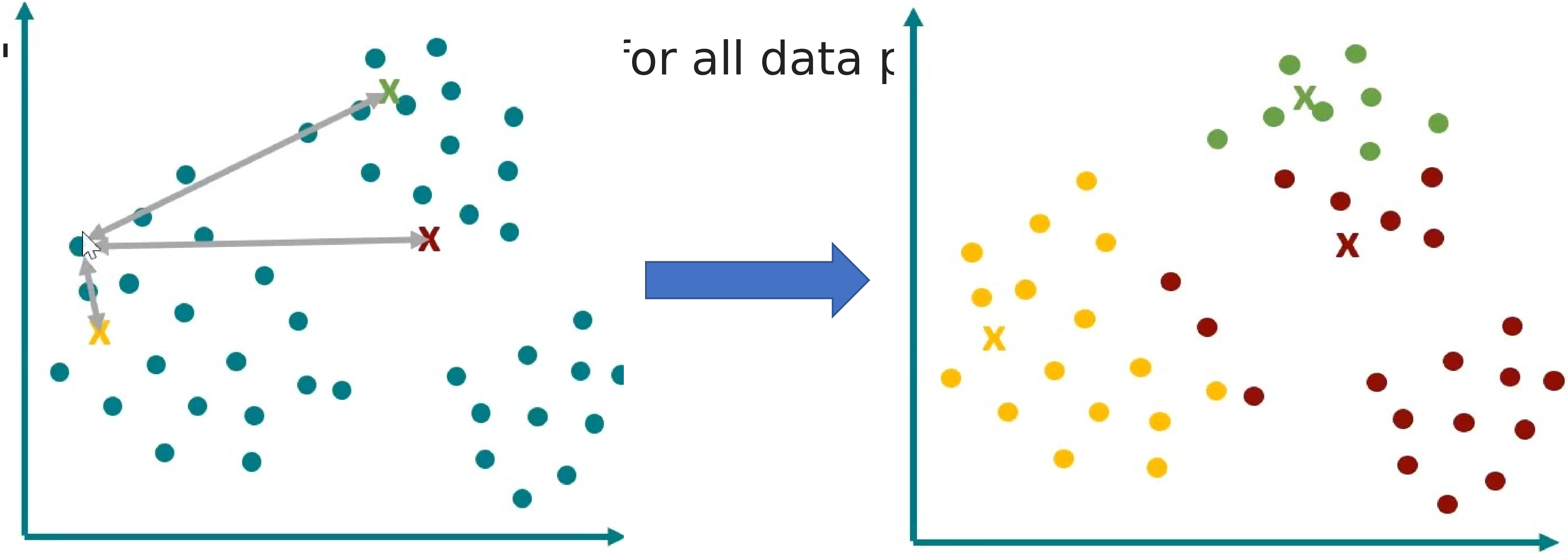


- Define initial Cluster Centroid
- Given that the number of clusters =3 (i.e., ***K***=3), then we have three Centroids
- Each Centroid represent a cluster

3. Assign points to clusters

- Calculate the distance (**Euclidean distance**) between each node to each centroid node, and assign the node to the nearest cluster.

for all data p



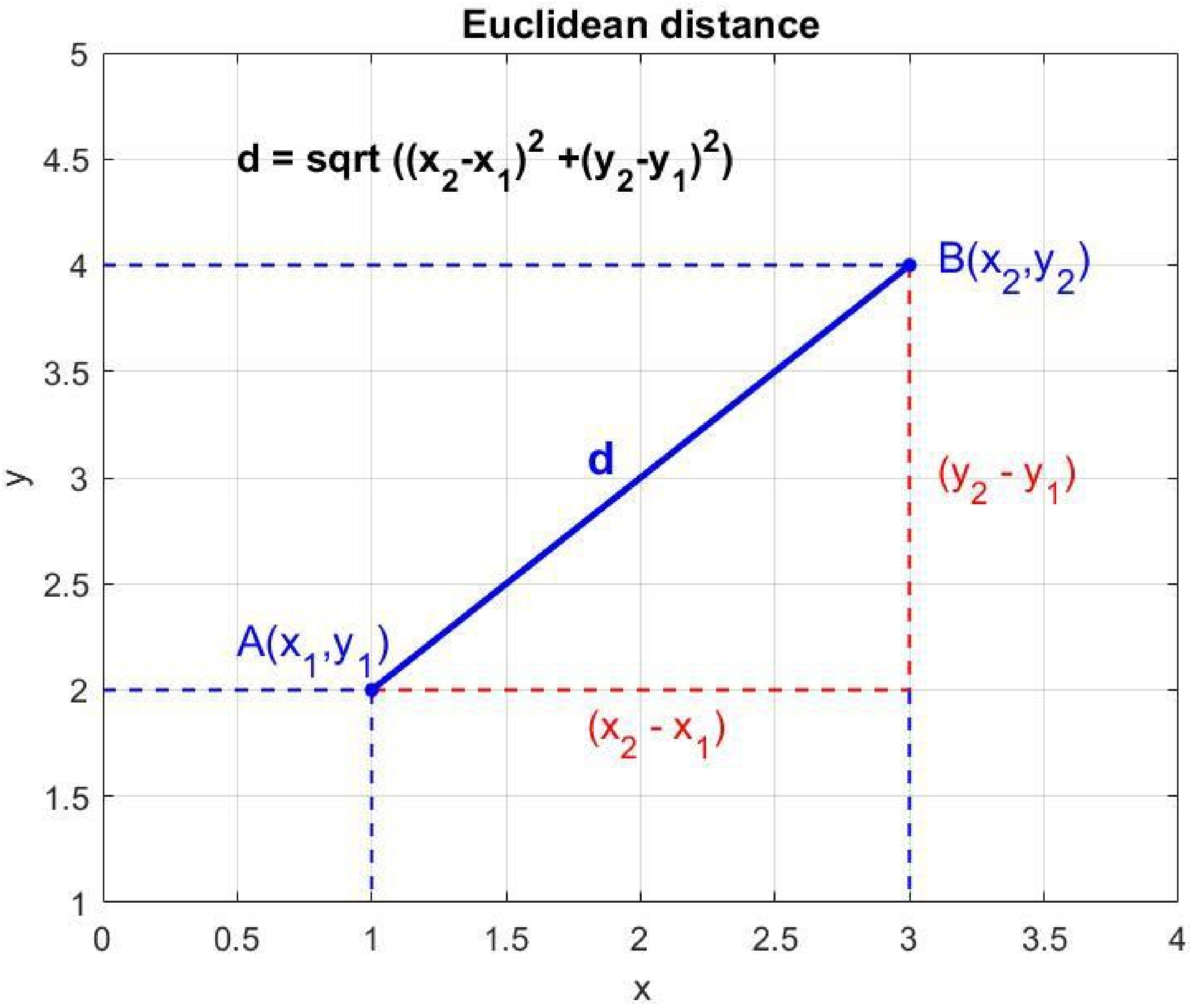
Euclidean Distance

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
$$(\text{distance between two data points } (x_2, y_2)\ and\ (x_1, y_1))$$

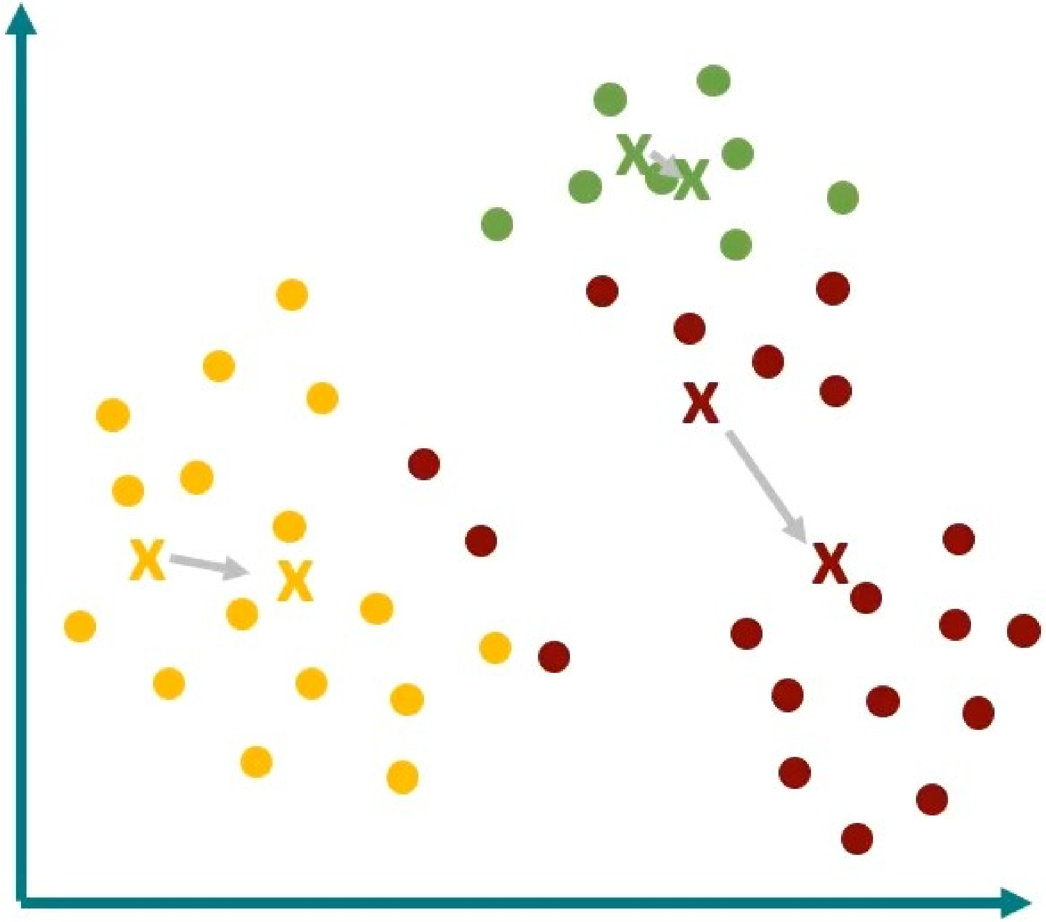
- Assume two centroid nodes with means of $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ select a node $\boldsymbol{x}_i \in (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ where $\boldsymbol{n}$ is the total number of observations.
- (

$$d_{i1} = \sqrt{(x_i - m_1)^T (x_i - m_1)}$$
$$d_{i2} = \sqrt{(x_i - m_2)^T (x_i - m_2)}$$

- $if\ di_1 \leq di_2,\ assign\ x_i\ to\ m_1,$



4. Re-calculate center of each cluster



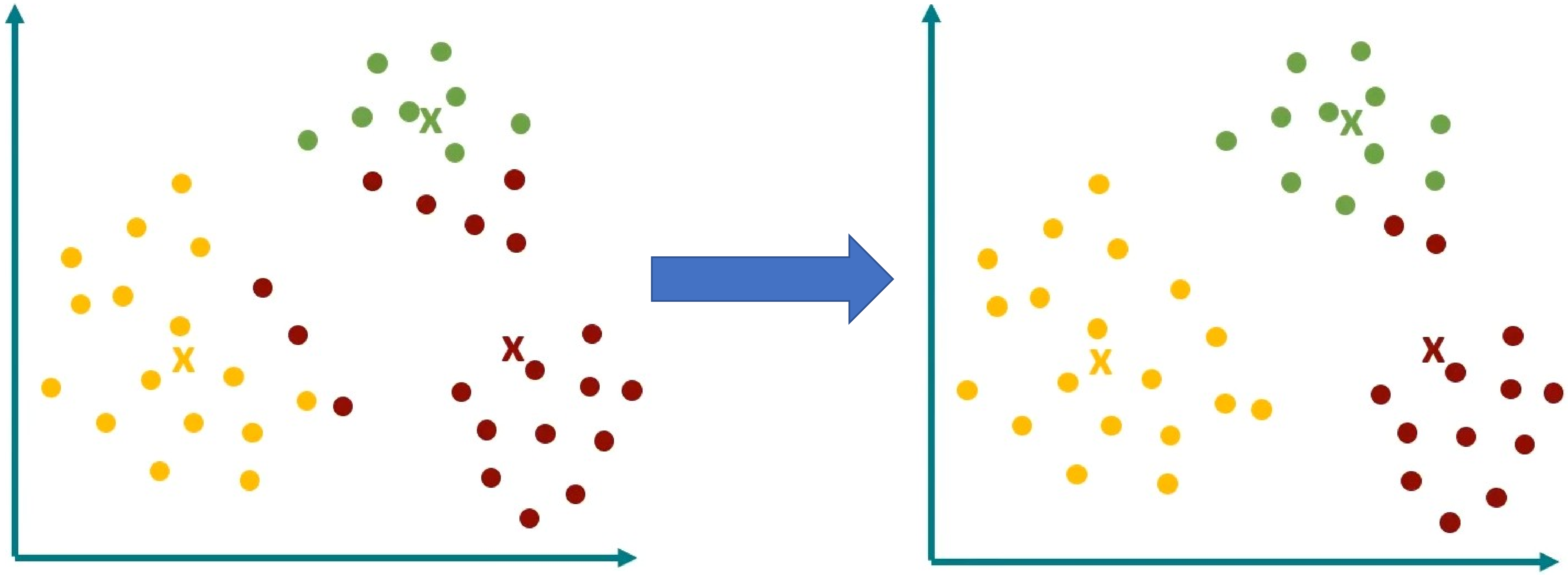
- Update the Centroids for each cluster (i.e., re-assign the Centroids).
- Assume cluster S_i *contains a total of* N_i samples (or observations), for a sample x_i within the cluster, its mean value ca

$$m_i = \frac{1}{N_i} \sum_{i=1}^{N_i} x_i$$

- Move the Cluster Centroids to the cluster

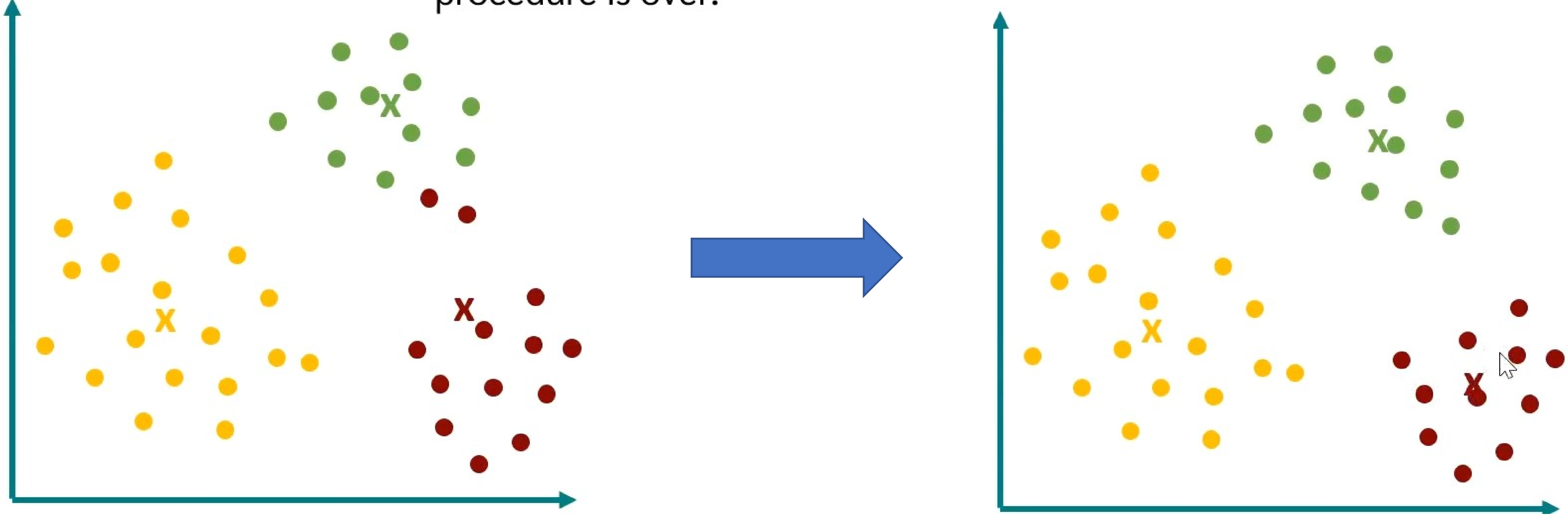
5. Assign points to the new clusters

- Since the Centroids are moved at different point, each data points is again assigned to the cluster that is closed to it (i.e., minimum distance between data point and the Centroid).

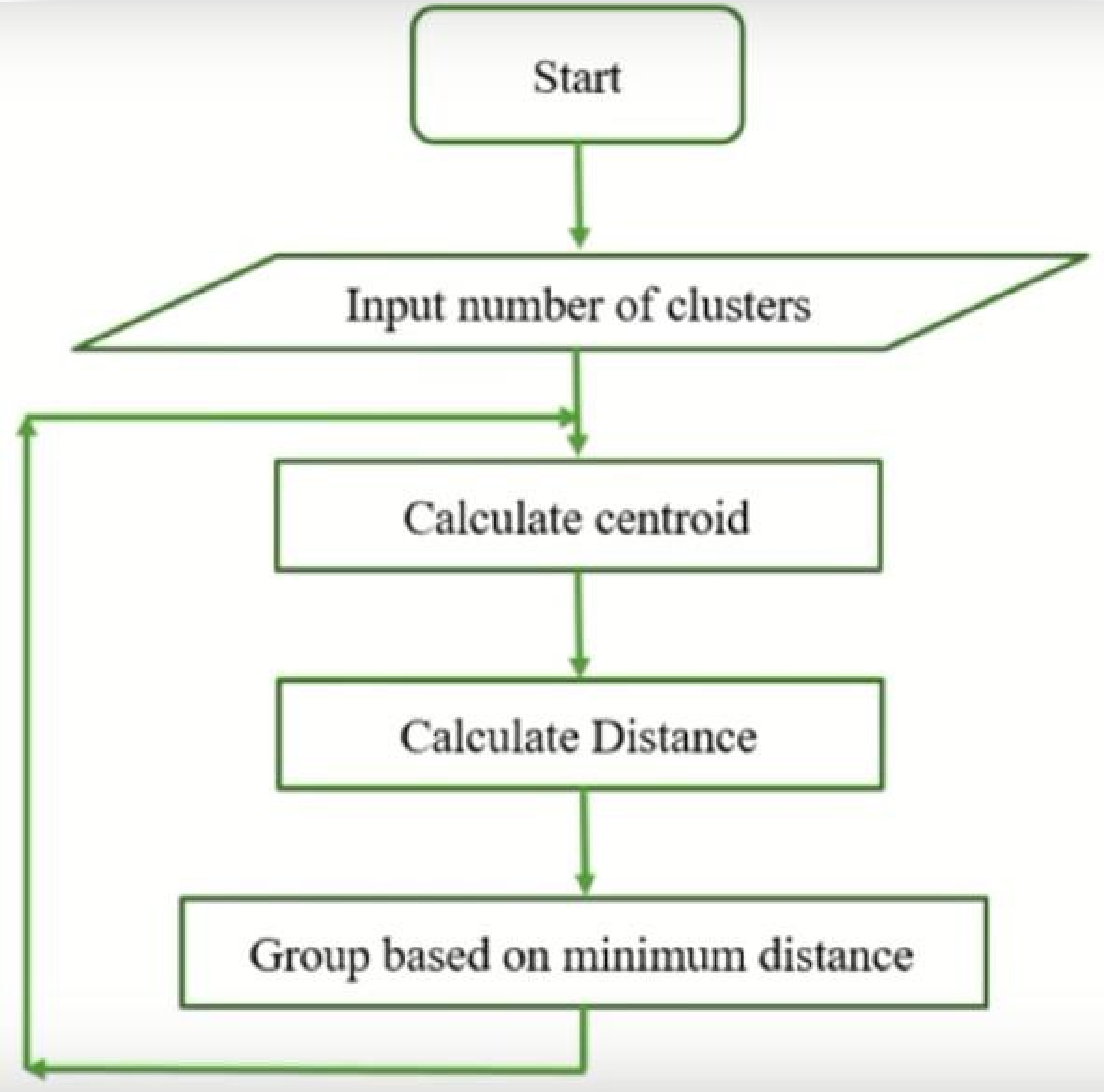


Repeat Step 4 and 5

- Now steps 4) and 5) are repeated until the cluster distribution does not change anymore.
  - Calculate the center of each cluster
  - Assign cluster centroid to the center
  - Assign points to the new clusters
- If the clusters do not change in one iteration, the procedure is over!



K-Means Diagram



15 Minutes Break

UNIVERSITY OF PLYMOUTH
School of Engineering, Computing and Mathematics

Lab 3 Explanation

UNIVERSITY OF PLYMOUTH
School of Engineering, Computing and Mathematics

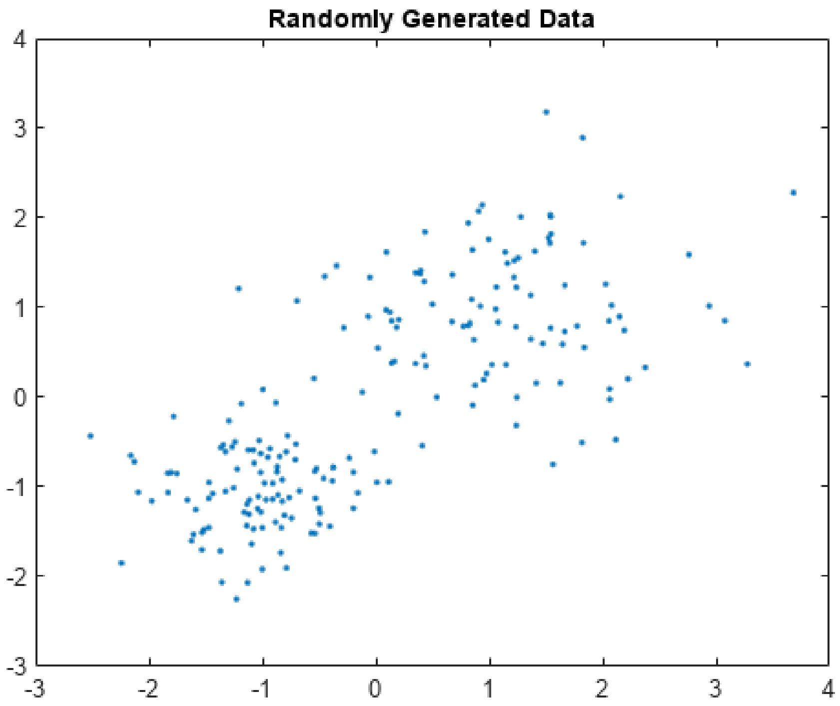
K-Mean MATLAB

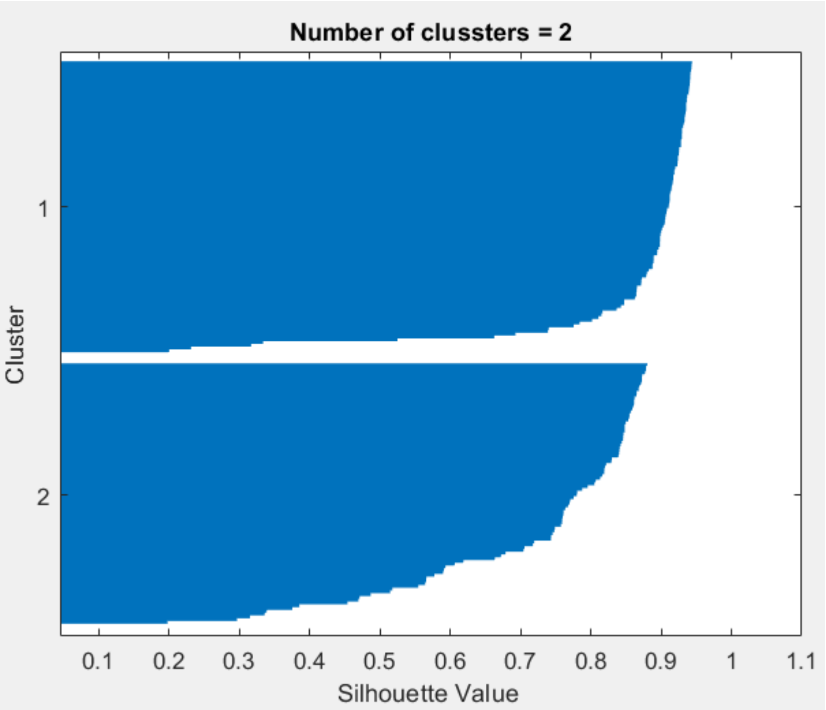
1. Load Data Set or generate random dataset
2. Create Clusters and Examine Separation
3. Use silhouette plot to see if the resulting clusters are well separated.
4. Calculate the mean silhouette (high silhouette value indicating that the clusters are well separated).

- A **Silhouette** plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.
- This measure ranges from **1** to **-1** (**1** indicating that points are very distant from neighboring clusters, while **0** shows that points are not distinctly in one cluster or another and **–1** shows that points are probably assigned to the wrong cluster).

K-Mean Implementation

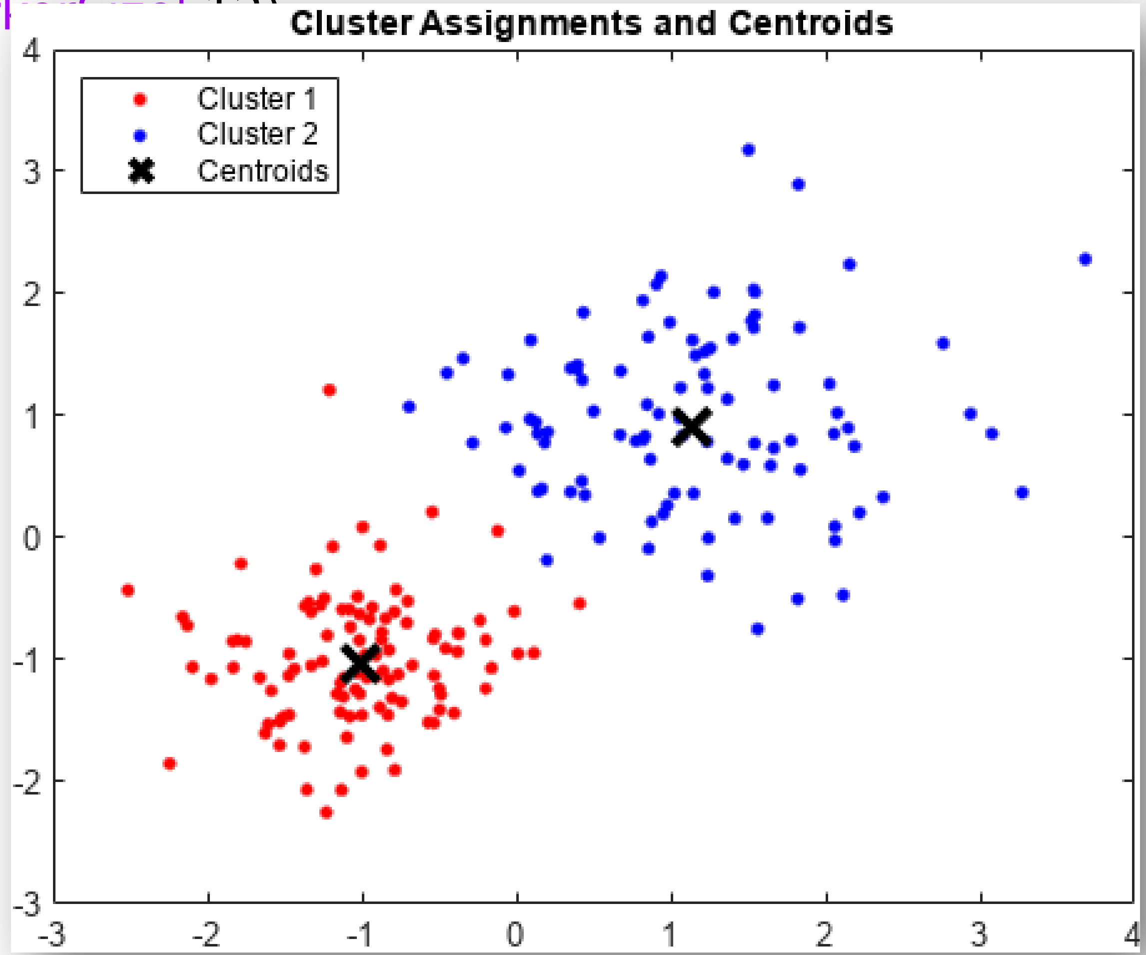
```
clear all,clc,close all
rng default; % For reproducibility
data = [randn(100,2)*0.75+ones(100,2);
randn(100,2)*0.5-ones(100,2)];
figure;
plot(data(:,1),data(:,2),'.');
title 'Randomly Generated Data’;
% idx is a vector of predicted cluster indices
corresponding to the observations in X.
% C is a 3-by-2 matrix containing the final centroid
locations.
k=2;
[idx,C,sumd]=kmeans(data,k);
[silh,h] = silhouette(data,idx,'sqEuclidean');
title(['Number of clussters = ' int2str(k)]);
xlabel 'Silhouette Value '
ylabel 'Cluster'
mean_silh=mean(silh)
```





K-Mean Implementation

```
figure;
plot(data(idx==1,1),data(idx==1,2),'r.','MarkerSize',12)
hold on
plot(data(idx==2,1),data(idx==2,2),'b.','MarkerSize',12)
plot(C(:,1),C(:,2),'kx',...
'MarkerSize',15,'LineWidth',3)
legend('Cluster 1','Cluster 2','Centroids',...
'Location','NW')
title 'Cluster Assignments and Centroids'
hold off
```



K-Mean Implementation

Modify the previous example, to do the following

1- Generate a training data set using three distributions.

2- Partition the training data into three clusters by using kmeans.

3- Plot the clusters and the cluster centroids.

4- Calculate the mean silhouette

5- Assign new data to existing clusters

6- Plot the test data and label the test data usi