# Clustering (Unsupervised Model)

**Artificial Intelligence**

**School of Cyber Security & Digital Forensics**

**M. Sc. Cyber Security (Semester-I)**

# Clustering

- Clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters
- Applications : **Business Intelligence-** organize customers enhanced customer relationship management

**Project management** – Similar projects to help in auditing and diagnosis

**Web search** – organizing the search results

**Outlier detection** – detection of credit card frauds

# Clustering

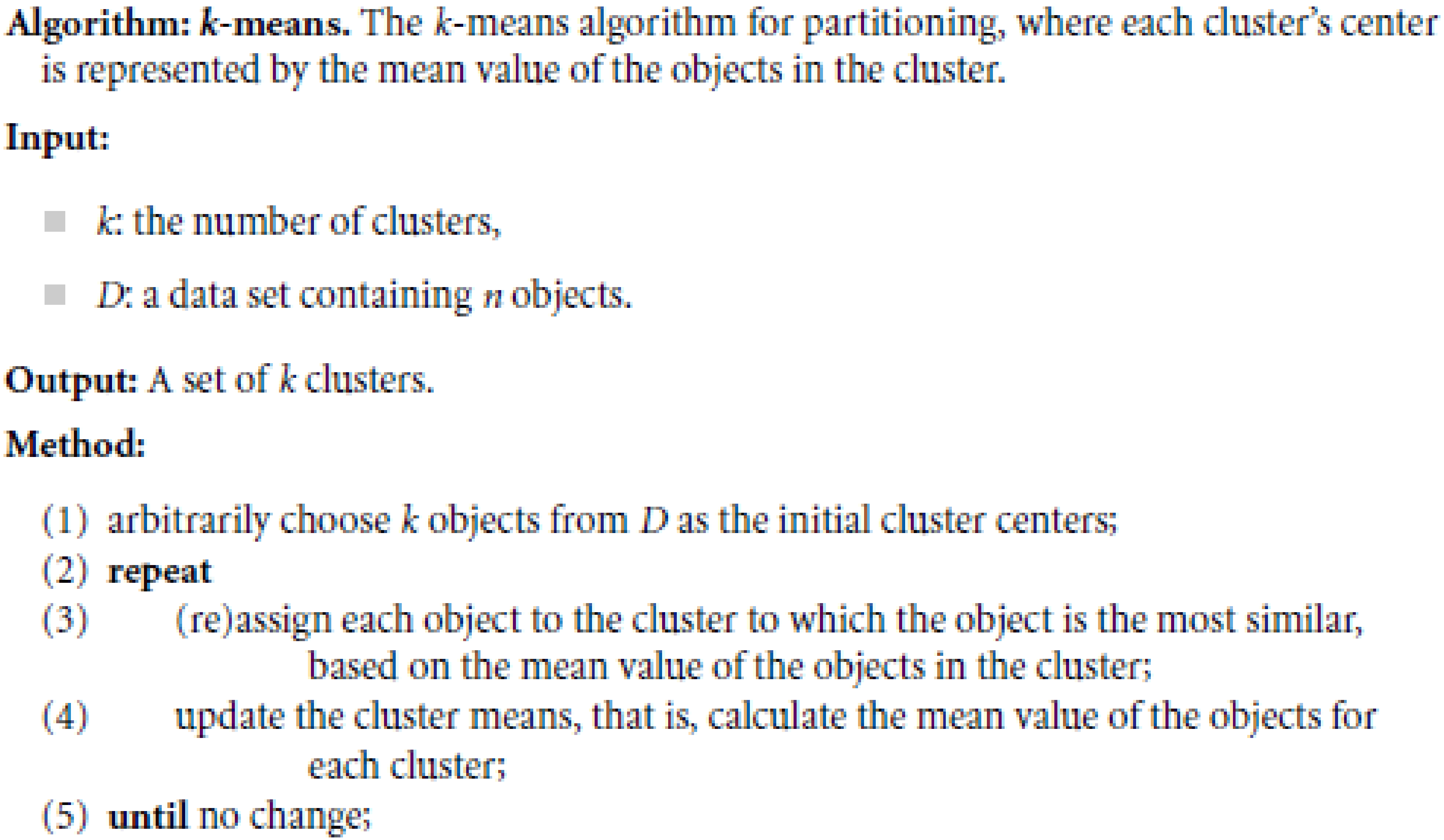
- The major fundamental clustering methods can be classified into the following categories
- **Partitioning methods**
- **Hierarchical methods**
- **Density-based methods**
- **Grid-based methods:**

# K-means clustering: A Centroid-Based Technique

- The objective function aims for high intracluster similarity and low intercluster similarity.
- A centroid-based partitioning technique uses the centroid of a cluster, Ci , to represent that cluster.
- Conceptually, the centroid of a cluster is its center point. The centroid can be defined in various ways such as by the mean or medoid of the objects (or points) assigned to the cluster.
- The quality of cluster Ci can be measured by the within cluster variation, which is the sum of squared error between all objects in p ϵ Ci and the centroid ci, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2,$$

# K-means clustering: Algorithm



**Algorithm: *k*-means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

**Input:**

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

**Output:** A set of *k* clusters.

**Method:**

```
(1) arbitrarily choose $k$ objects from $D$ as the initial cluster centers;
(2) repeat
(3)       (re)assign each object to the cluster to which the object is the most similar,
                    based on the mean value of the objects in the cluster;
(4)       update the cluster means, that is, calculate the mean value of the objects for
                    each cluster;
(5) until no change;
```

# K-means clustering: A Centroid-Based Technique

- Cluster the following data points into three clusters

A1(2,10) A2(2,5) A3(8,4) B1 ( 5,8) B2(7,5) B3(6,4) C1(1,2) C2(4,9)

The distance function is Euclidean distance. Suppose we initially assign A1, B1 and C1 as the centre of each cluster respectively.

i. Show the three cluster centre and clusters after the first round

ii. The final three clusters