Clustering Analysis
Unsupervised learning approach.

Overview.

- Clustering algorithms are unsupervised learning algorithms.
- Clustering is grouping a set of data objects into subsets. Each subset is a cluster.
- The objects(attributes/features) in a cluster are similar to each other and dissimilar to the objects in other clusters.
- The clustering algorithms are useful in discovering the previously unknown groups within the data.
- Clustering algorithms are widely used in finance, biology, web search.

Clustering methods

❑ **Partitioning methods:**

Given a set of ***n*** objects, a partitioning method constructs **k** partitions of the data, where each partition is a cluster and ***k<n.***

- Partitioning methods performs a one level partitioning on the data.
- Most partitioning methods are distance-based.

Clustering methods

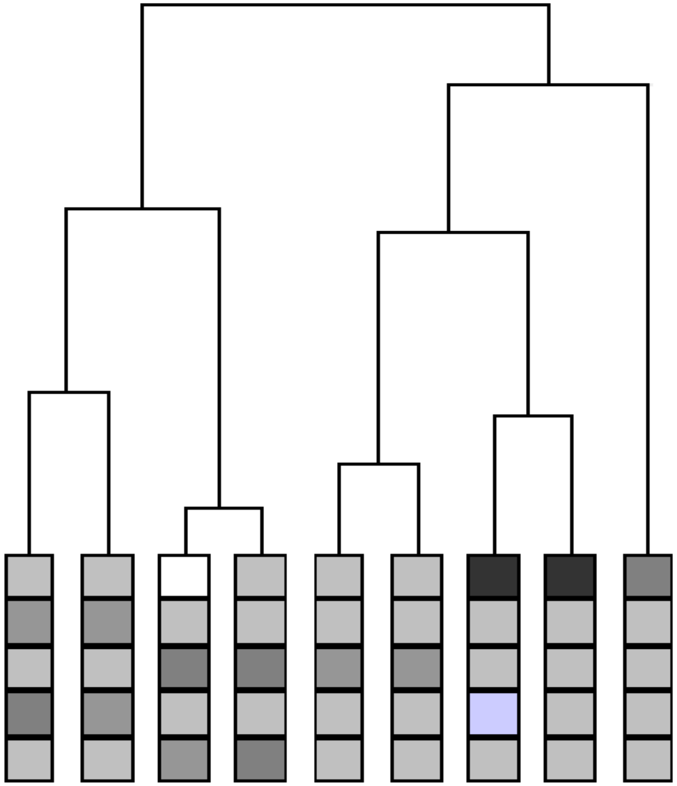
❑ **Hierarchical methods:**

A Hierarchical method creates a hierarchical decomposition of the given dataset.

- The Hierarchical approach can be classified as being either agglomerative(bottom-up approach),or divisive approach(top-down approach).
- *The agglomerative approach* starts with each object forming a separate group. Then , it merges the objects/groups close to one another until all the groups are merged into one.
- *The divisive approach* : starts with all objects in the same cluster. Then, it splits into smaller clusters, until eventually each object is in one cluster, or a termination condition holds.

Hierarchical methods

The divisive approach

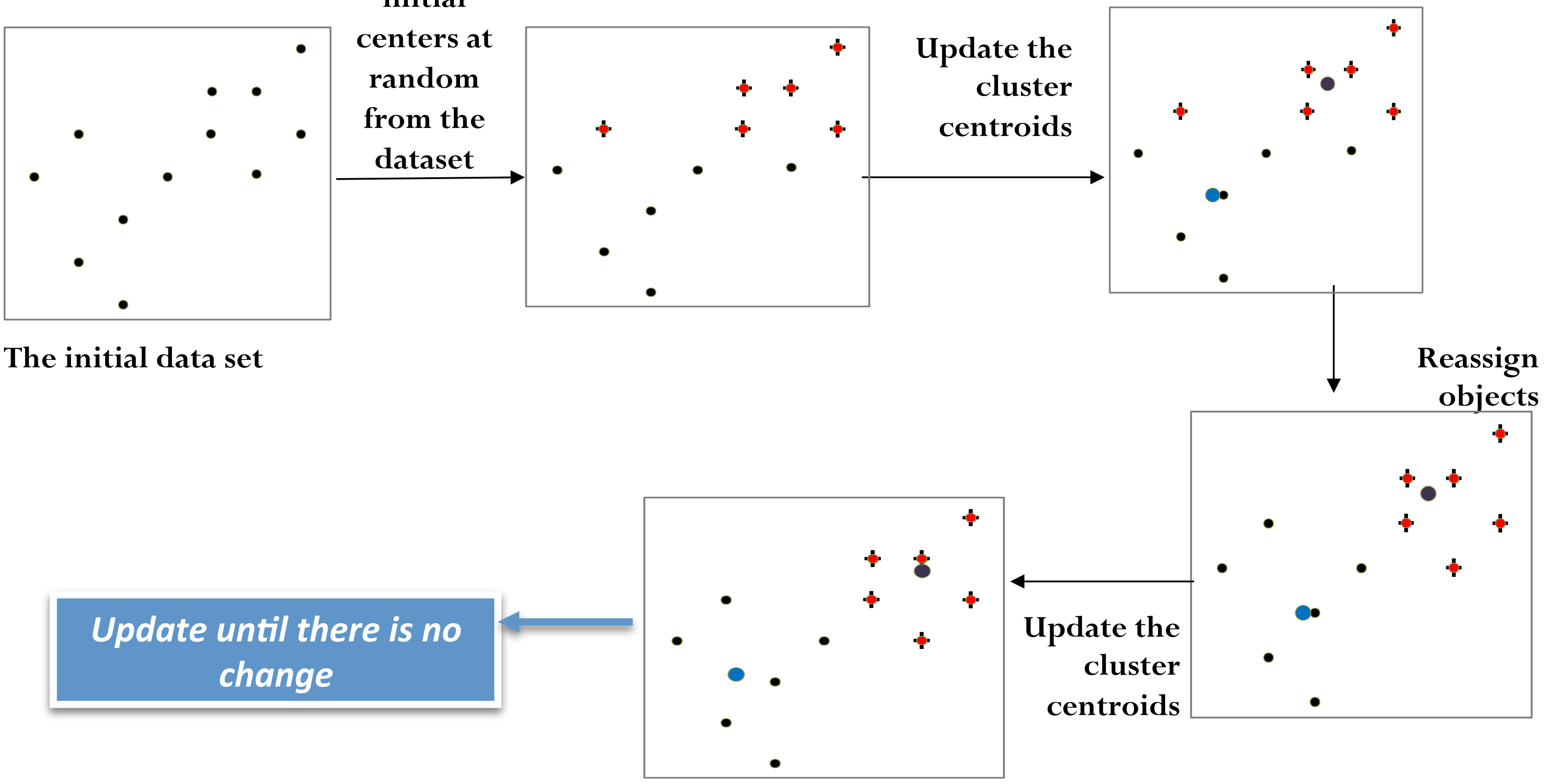


The agglomerative approach

Clustering methods…

- Other clustering methods include :
  - Density-based methods.
  - Grid-based methods.

Partitioning clustering(K-means Algorithm).



K=2

Select initial centers at random from the dataset

The initial data set

Update the cluster centroids

Reassign objects

Update the cluster centroids

Update until there is no change

K-means Algorithm.

1- Number of clustering is given(K).

2- Select initial center at random from the dataset.

3- Use a distant measure such as *Euclidean distance* to compute between the centers and each point in the dataset, and assign each point to the center that it closes to.

4- For each cluster that is formed compute the average to find the appropriate center.

5- Repeat the process until convergence .

➢ The algorithm terminates at a local optimum.

K-means Example algorithm

Type of Treatments	Gene.1 expression	Gene.2 expression
Treatment 1	1	1
Treatment 2	2	1
Treatment 3	4	3
Treatment 4	5	4

The Goal is to find similar treatments ?

K=2

- You can think of each treatment as an (x,y) point in an attribute space.
- At the beginning we choose random centroids for the two clusters , Let us say Treatment.1 (1,1) and Treatment.2 (2,1).

  C1=(1,1)   C2=(2,1)

- We calculate the distance between each cluster centroid and each treatment, using e.g. Euclidian distance .
- (Treatment.1, C1) ,(Treatment.1, C2)= $\sqrt{(1-1)^2+(1-1)^2}=0, \sqrt{(2-1)^2+(1-1)^2}=1$
- (Treatment.2, C1) ,(Treatment.2, C2)=
- (Treatment.3, C1) ,(Treatment.3, C2)= $\sqrt{(1-4)^2+(1-3)^2}=3.6, \sqrt{(2-4)^2+(1-3)^2}=2.8$
- (Treatment.4, C1) ,(Treatment.4, C2)=

- The result can be represented using a distance matrix :

D(T(i),C1)	**0**	**1**	**3.61**	**5**
D(T(i),C2)	1	0	2.83	4.24

Cluster 1	**Treatment.1**
Cluster 2	**Treatment.2** **Treatmenr.3** **Treatment.4**

- Thus, each treatment assigned to the cluster that it closes to,(less distance)

- We re-calculate the centorids again based on the new members in each cluster.
- C1 for cluster.1 remains the same as it is only one member.
- C2 has T2,T3,T4, therefore :
  - » C2= $(\frac{2+4+5}{3},\frac{1+3+4}{3})=(3.7,2,7)$

- After the new centorids for each cluster ,the result is :

D(T(i),C1)	0	1	3.61	5
D(T(i),C2)	3.14	2.36	0.47	1.89

Cluster 1	Treatment.1 Treatment.2
Cluster 2	Treatmenr.3 Treatment.4

- Thus, each treatment assigned to the cluster that it closes to,(less distance)

- We re-calculate the centorids again based on the new members in each cluster.
- C1 for cluster.1(T1,T2)= $(\frac{1+2}{2},\frac{1+1}{2}) = (1.5,1)$
- C2 for cluster.2(T3,T4)= $(\frac{4+5}{2},\frac{3+4}{2}) = (4.5,3.5)$

- After the new centorids for each cluster ,the result is :

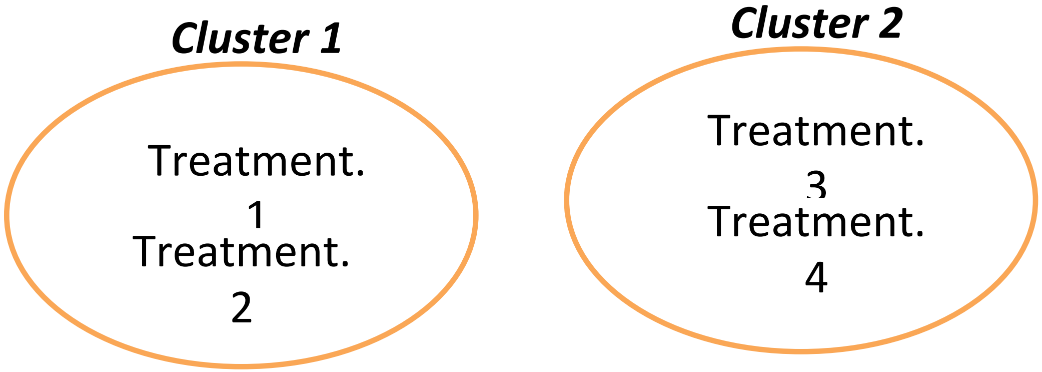
D(T(i),C1)	**0.5**	**0.5**	**3.20**	**4.61**
D(T(i),C2)	4.3	3.54	0.71	0.71

Cluster 1	**Treatment.1** **Treatment.2**
Cluster 2	**Treatmenr.3** **Treatment.4**

- Thus, each treatment assigned to the cluster that it closes to,(less distance)

- Since the clusters in the new grouping remains the same as the previous clusters , then

  » ***K-means terminates at this stage and it is considered to reach its optimal solution***!.



Cluster 1

Treatment. 1

Treatment. 2

Cluster 2

Treatment. 3

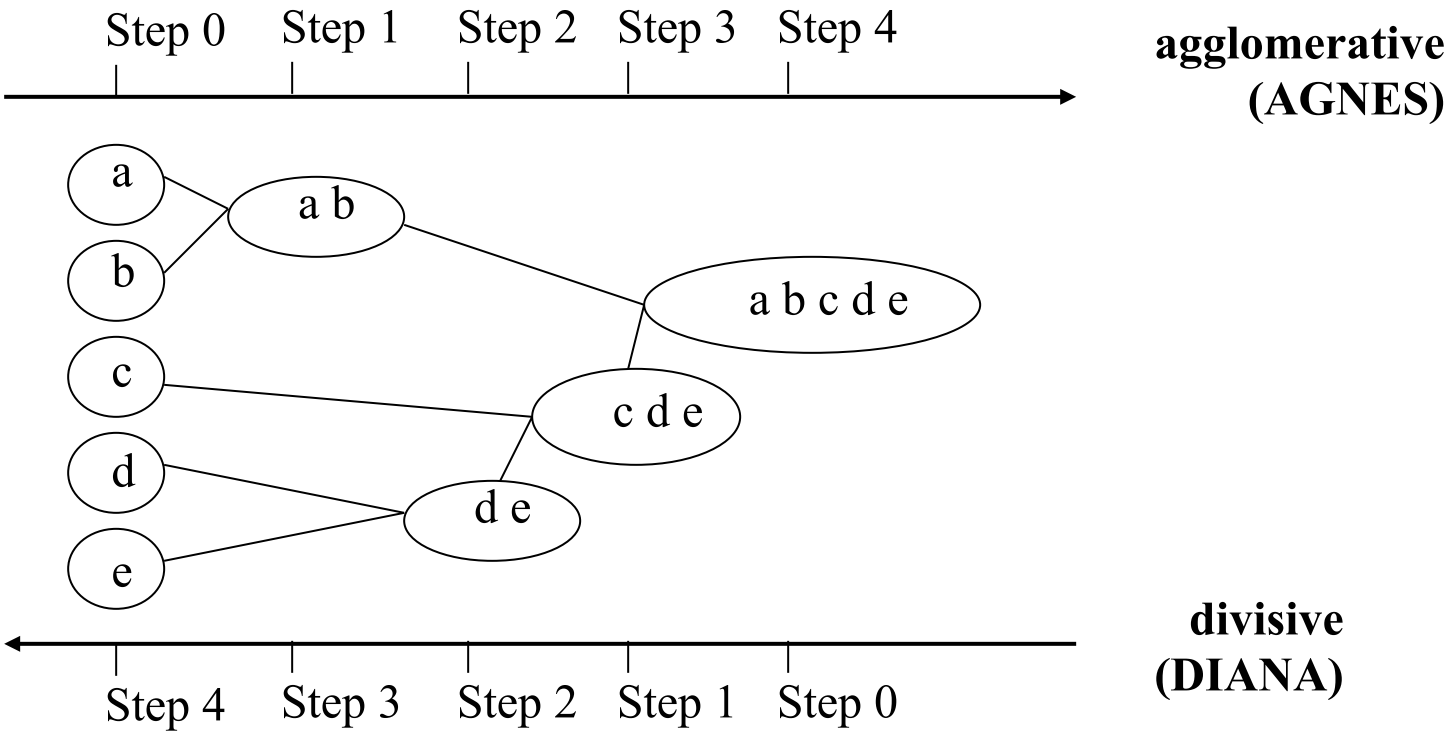
Treatment. 4

Partitioning methods.

- Weaknesses:
  - Applicable when the mean is defined numerically.
  - Needs to specify the k which is not known in advance usually.
  - Difficult to handle noisy and outliers.

Hierarchical Clustering

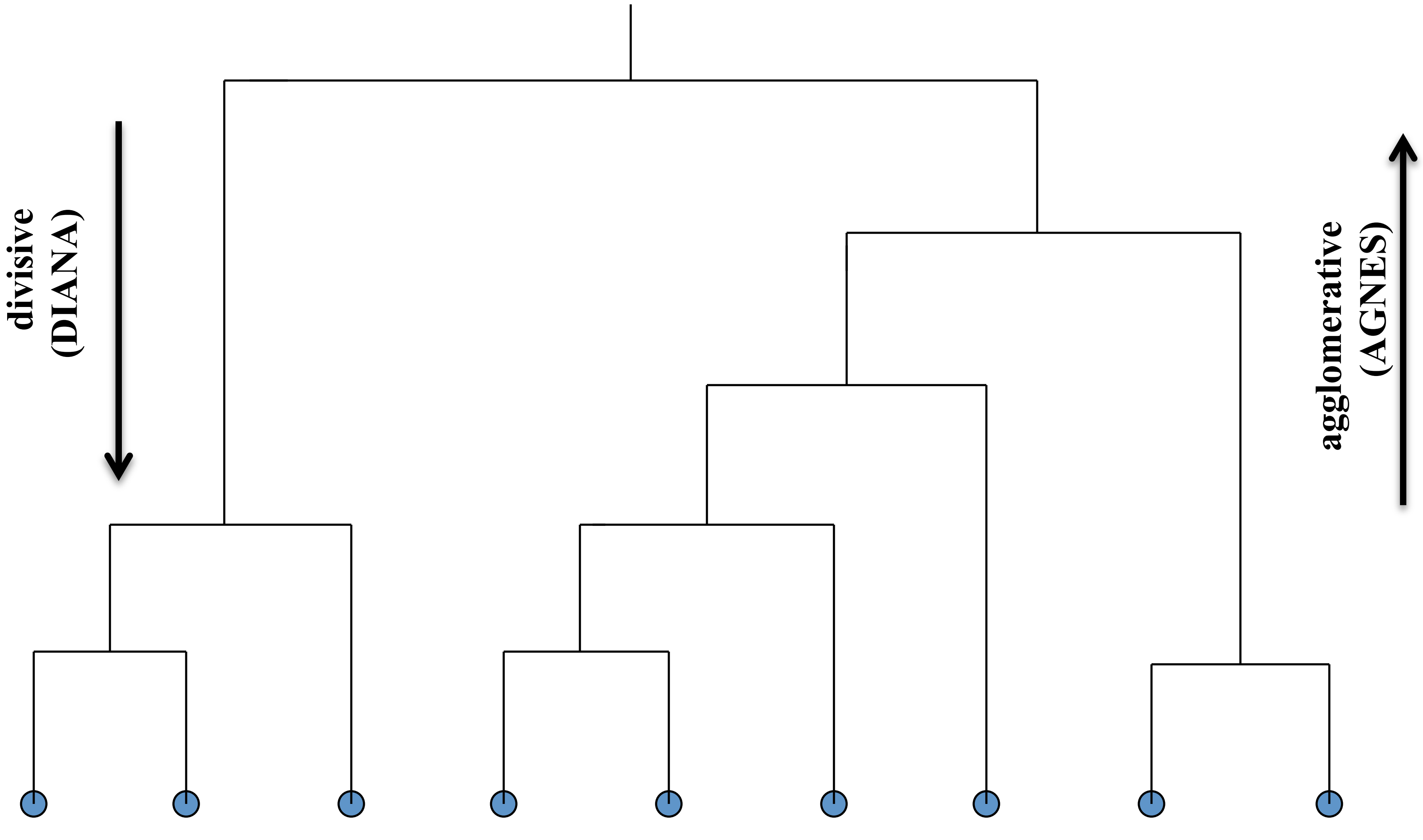
- Agglomerative NESting(AGNES) versus Divisive Analysis(DIANA) hierarchical clustering.



AGNES vs. DIANA .

- Initially, (**AGNES** places each object into a cluster of its own.
- The clusters are then merged step-by-step according to some criterion.
- For example : Cluser.1 and Cluster.2 are merged if an object in C1, and object in C2 form the minimum Euclidean distance between any two objects from different clusters.
- **DIANA**, initially places all objects in one cluster. The cluster is split according to some criterion such as the maximum Euclidean distance between the closest neighbouring objects in the cluster.
- In DIANA , the process of splitting is repeated until each new cluster contains only a single object.

- A tree called a dendrogram is commonly used to represent the process of hierarchical clustering.



Agglomerative NESting(AGNES).

- There are three kinds of agglomerative methods:

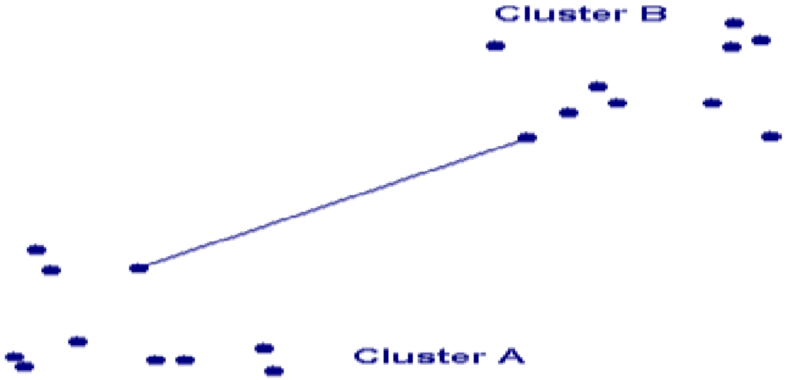
  » *Single linkage.*

  » *Complete linkage.*

  » *Average linkage.*

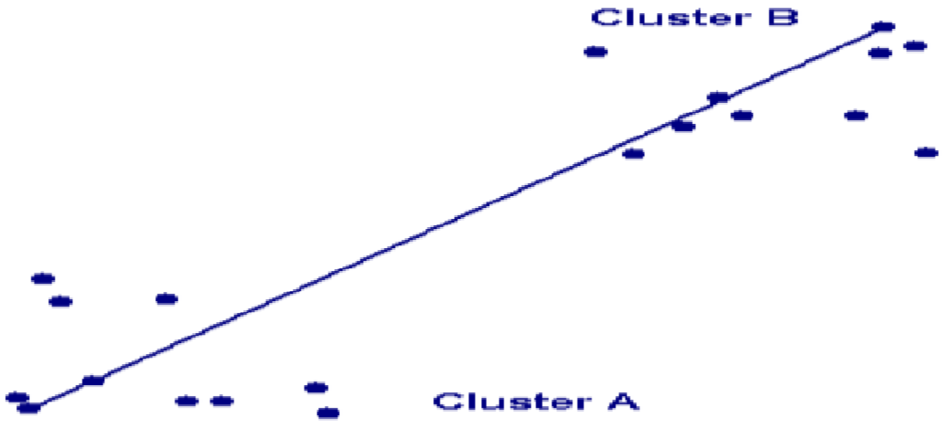
Single linkage.

- Single linkage clustering, also called nearest neighbour technique (**NN**), is one of the simplest agglomerative hierarchal clustering algorithms.
- The distance between each cluster in the single linkage method is defined as the distance between the closest points in two clusters.



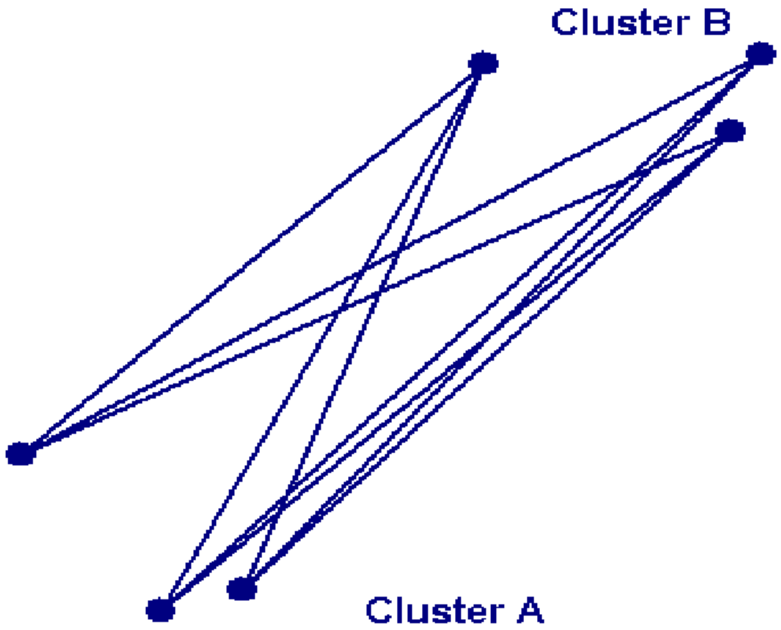
Complete linkage.

- Complete linkage clustering is also known as the farthest neighbour clustering method.
- The distance in complete linkage clustering is defined as the farthest distance between two points in two clusters.



Average linkage.

- The distance in average linkage clustering is the average distance between all points in two clusters.


Cluster B
Cluster A