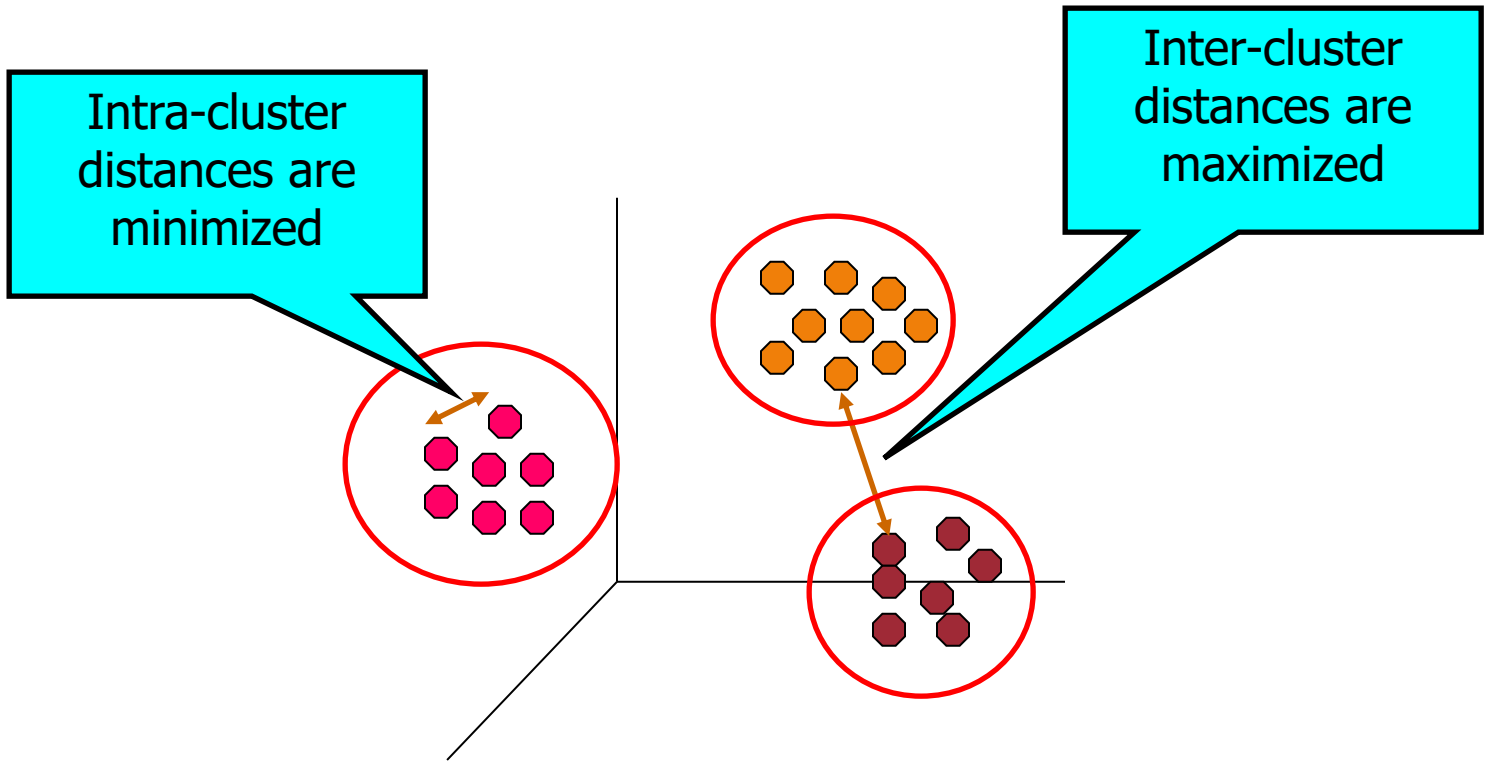
Unsupervised Learning:
k-means clustering, Hierarchical clustering

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



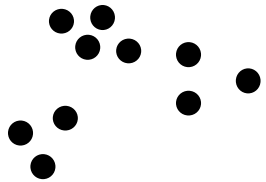
What is Cluster Analysis?

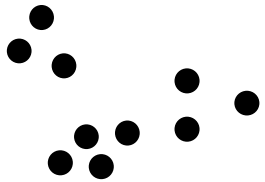
- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised learning: no predefined classes

Some Applications of Clustering

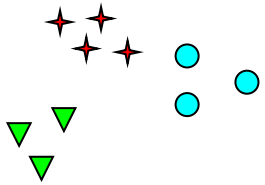
- **Marketing**
  - In the area of marketing, clustering is used to explore and select customers that are potential buyers of the product
  - After the clusters have been developed, *businesses can keep a track of their customers and make necessary decisions to retain them in that cluster*
- **Retail**
  - Retail industries make use of clustering to group customers based on their preferences, style, choice of wear as well as store preferences
  - This allows them to manage their stores in a much more efficient manner.
- **Medical Science**
  - Medicine and health industries make use of clustering algorithms to *facilitate efficient diagnosis and treatment of their patients as well as the discovery of new medicines*
  - Based on the age, group, genetic coding of the patients, these organisations are better capable to understand diagnosis through robust clustering

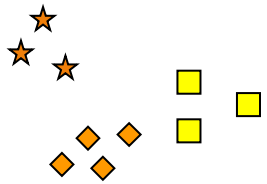
Notion of a Cluster can be Ambiguous



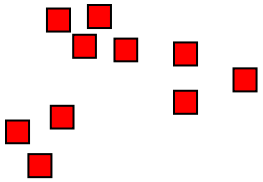


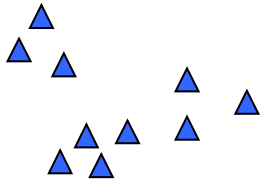
How many clusters?



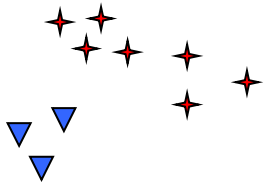


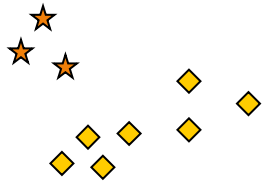
Six Clusters





Two Clusters





Four Clusters

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Types of Clusterings

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters

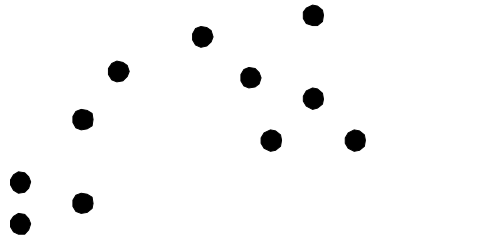
Partitional Clustering

- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

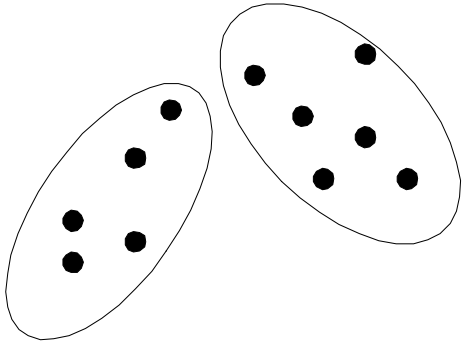
Hierarchical clustering

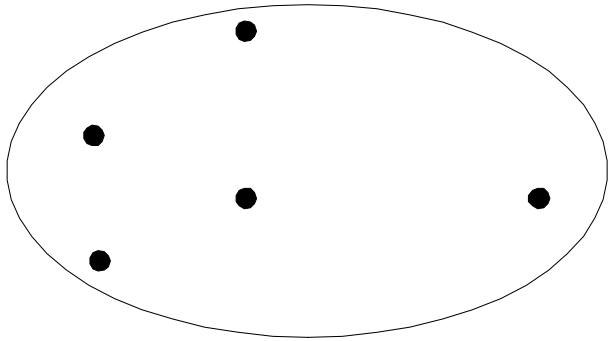
- A set of nested clusters organized as a hierarchical tree

Partitional Clustering



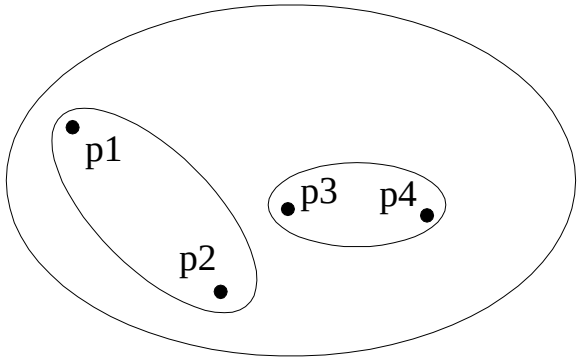
Original Points



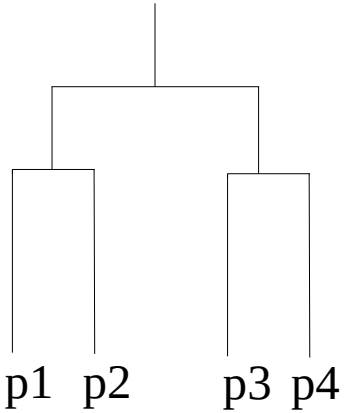


A Partitional Clustering

Hierarchical Clustering



Hierarchical Clustering



Dendrogram

Clustering Algorithms

- K-means and its variants
- Hierarchical clustering

K-means Clustering

- k-Means clustering algorithm proposed by J. Hartigan and M. A. Wong [1979].
- Given a set of *n* distinct objects, the k-Means clustering algorithm partitions the objects into *k* number of clusters such that intracluster similarity is high but the intercluster similarity is low.
- In this algorithm, user has to specify *k*, the number of clusters and consider the objects are defined with numeric attributes and thus using any one of the distance metric to demarcate the clusters.

k-Means Algorithm

The algorithm can be stated as follows.

- First it selects *k* number of objects at random from the set of n objects. These *k* objects are treated as the centroids or center of gravities of *k* clusters.
- For each of the remaining objects, it is assigned to one of the closest centroid. Thus, it forms a collection of objects assigned to each centroid and is called a cluster.
- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).
- The assignment and update procedure is until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no assignment, etc.)

k-Means Algorithm

Input: D is a dataset containing *n* objects, *k* is the number of cluster

Output: A set of *k* clusters

Steps:

1. Randomly choose *k* objects from D as the initial cluster centroids.
2. **For** each of the objects in D **do**
   - Compute distance between the current objects and *k* cluster centroids
   - Assign the current object to that cluster to which it is closest.
3. Compute the "cluster centers" of each cluster. These become the new cluster centroids.
4. Repeat step 2-3 until the convergence criterion is satisfied
5. Stop

Illustration of k-Means clustering algorithms

Table 1. 16 objects with two attributes A_1 and A_2.

A_1	A_2
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Figure 1. Plotting data of Table

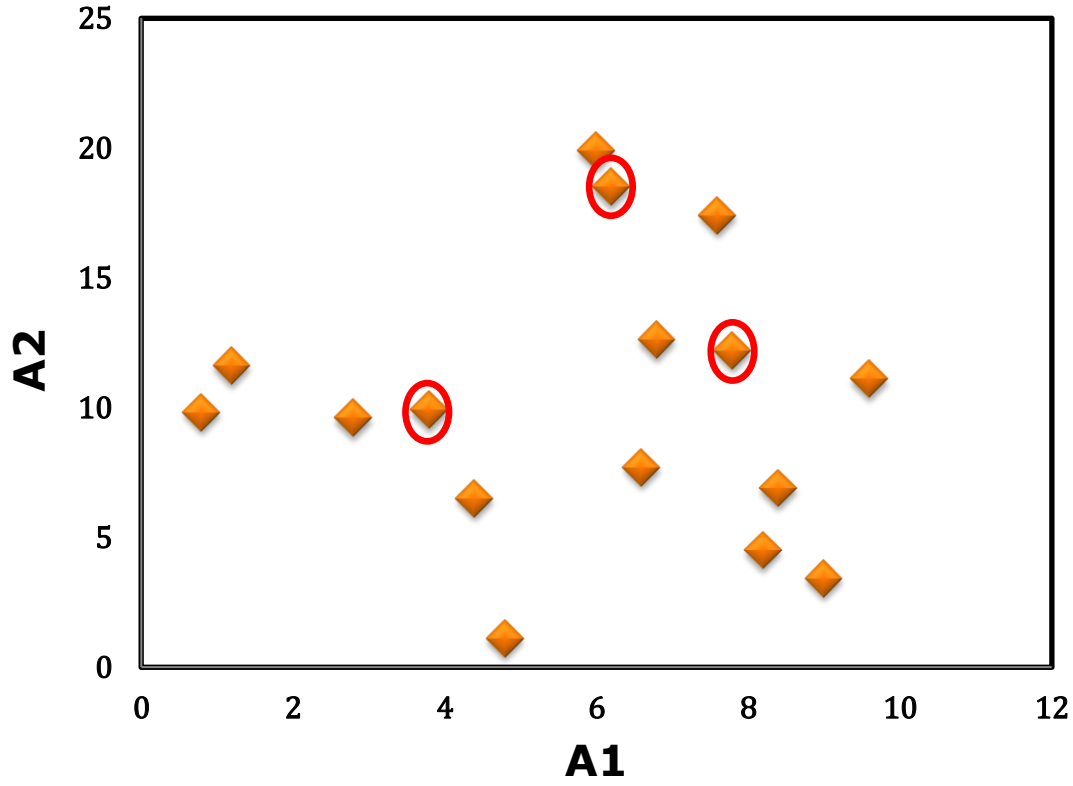


Illustration of k-Means clustering algorithms

- Suppose, k=3. Three objects are chosen at random shown as circled (see Fig 1). These three centroids are shown below.

Initial Centroids chosen randomly

Centroid	Objects	
	A1	A2
c_1	3.8	9.9
c_2	7.8	12.2
c_3	6.2	18.5

- Let us consider the Euclidean distance measure as the distance measurement in our illustration.
- Let d_1, d_2 and d_3 denote the distance from an object to c_1, c_2 and c_3 respectively. The distance calculations are shown in Table 2.
- Assignment of each object to the respective centroid is shown in the right-most column and the clustering so obtained is shown in Fig 2.

Illustration of k-Means clustering algorithms

Table 2: Distance calculation

A_1	A_2	d_1	d_2	d_3	cluster
6.8	12.6	4.0	1.1	5.9	2
0.8	9.8	3.0	7.4	10.2	1
1.2	11.6	3.1	6.6	8.5	1
2.8	9.6	1.0	5.6	9.5	1
3.8	9.9	0.0	4.6	8.9	1
4.4	6.5	3.5	6.6	12.1	1
4.8	1.1	8.9	11.5	17.5	1
6.0	19.9	10.2	7.9	1.4	3
6.2	18.5	8.9	6.5	0.0	3
7.6	17.4	8.4	5.2	1.8	3
7.8	12.2	4.6	0.0	6.5	2
6.6	7.7	3.6	4.7	10.8	1
8.2	4.5	7.0	7.7	14.1	1
8.4	6.9	5.5	5.3	11.8	2
9.0	3.4	8.3	8.9	15.4	1
9.6	11.1	5.9	2.1	8.1	2

Fig 2: Initial cluster with respect to Table 16.2

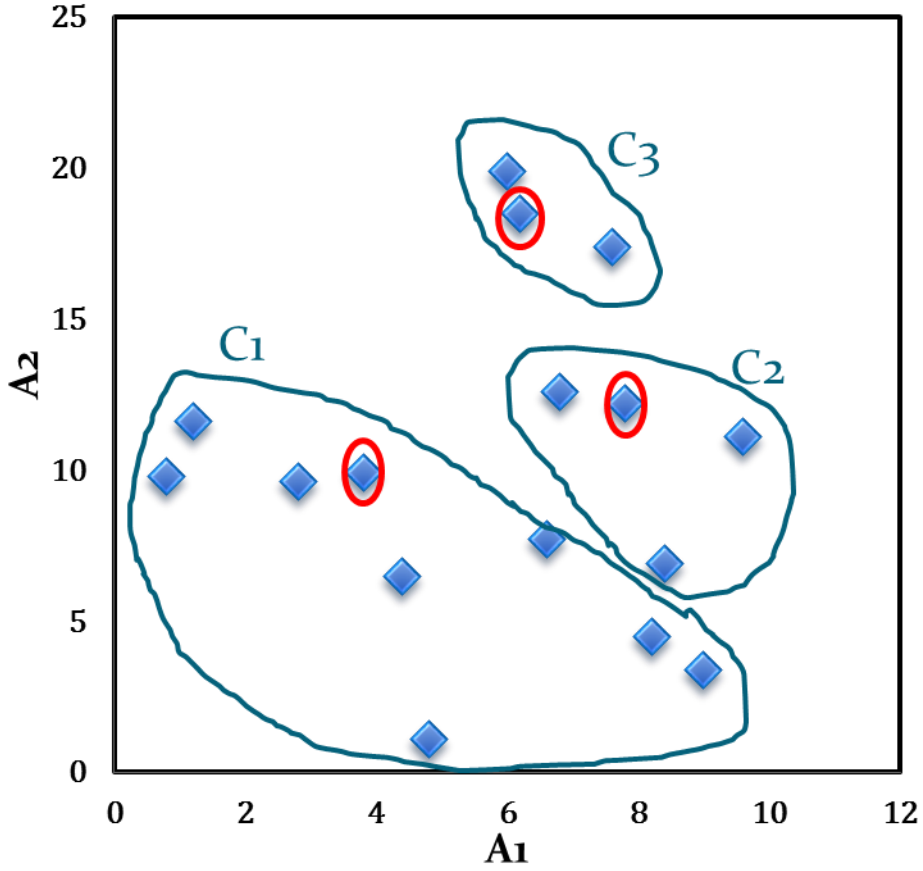


Illustration of k-Means clustering algorithms

The calculation new centroids of the three cluster using the mean of attribute values of A_1 and A_2 is shown in the Table below. The cluster with new centroids are shown in Fig 3.

Calculation of new centroids

New Centroid	Objects	
	A1	A2
c_1	4.6	7.1
c_2	8.2	10.7
c_3	6.6	18.6

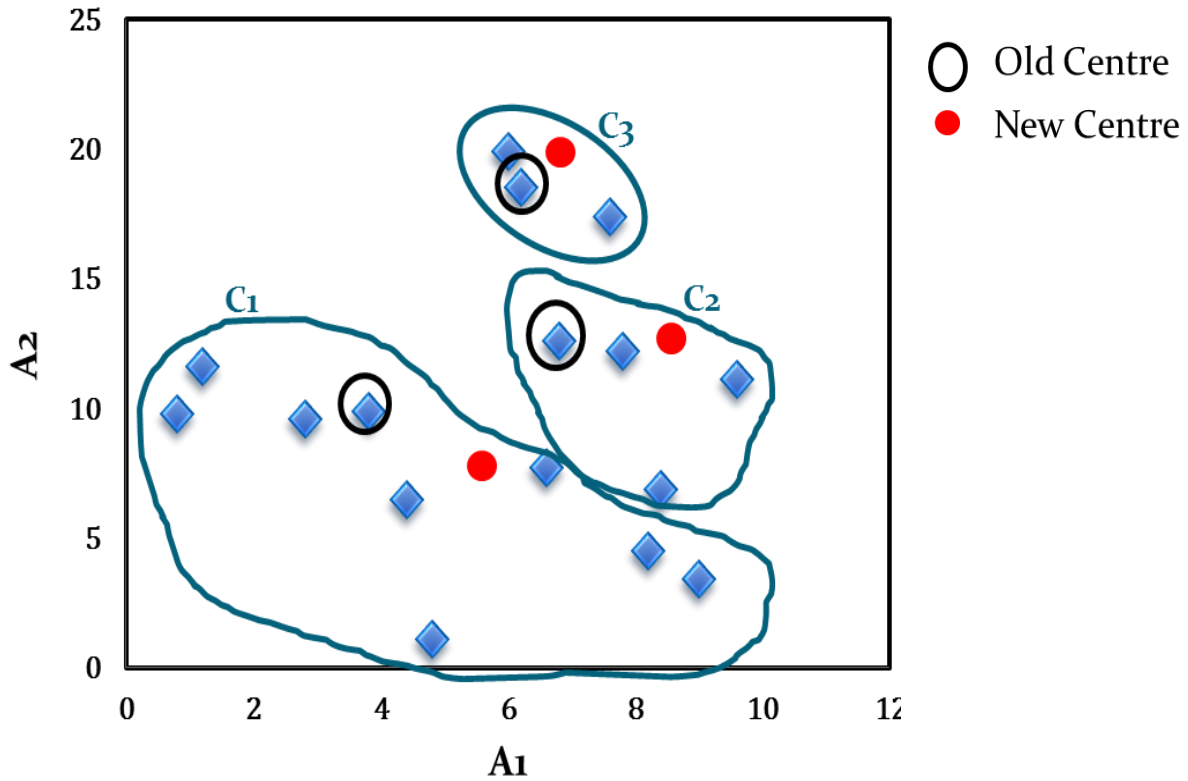


Fig 3: Initial cluster with new centroids

Illustration of k-Means clustering algorithms

We next reassign the 16 objects to three clusters by determining which centroid is closest to each one. This gives the revised set of clusters shown in Fig 4.

Note that point p moves from cluster C_2 to cluster C_1.

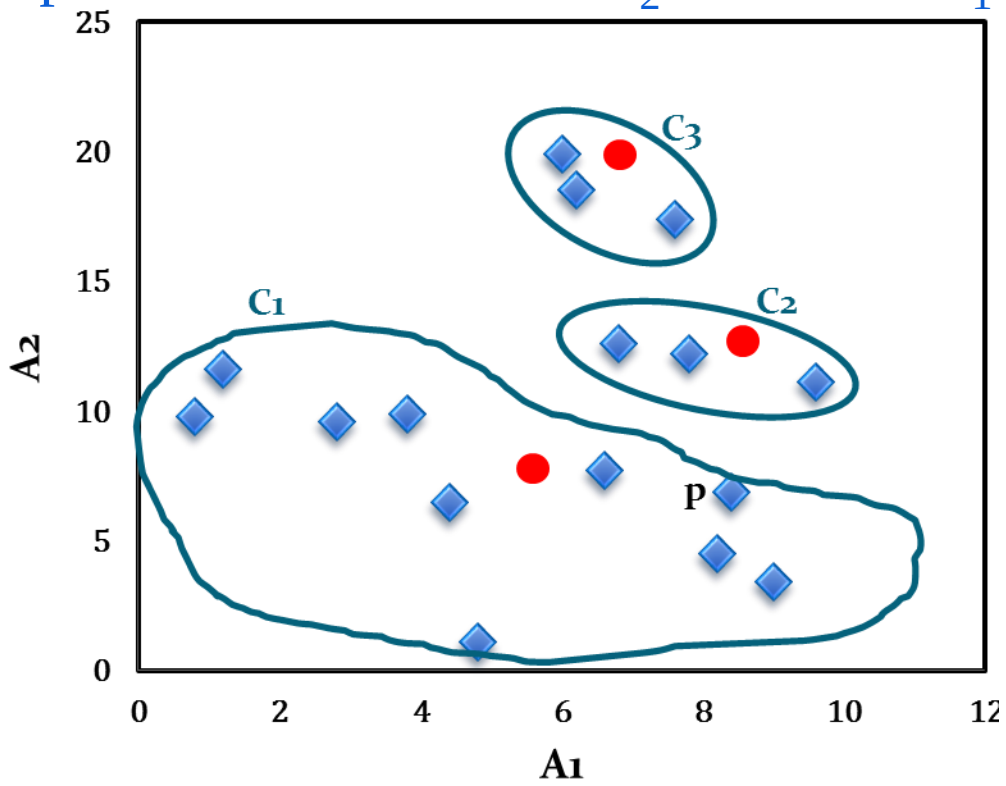


Fig 4: Cluster after first iteration

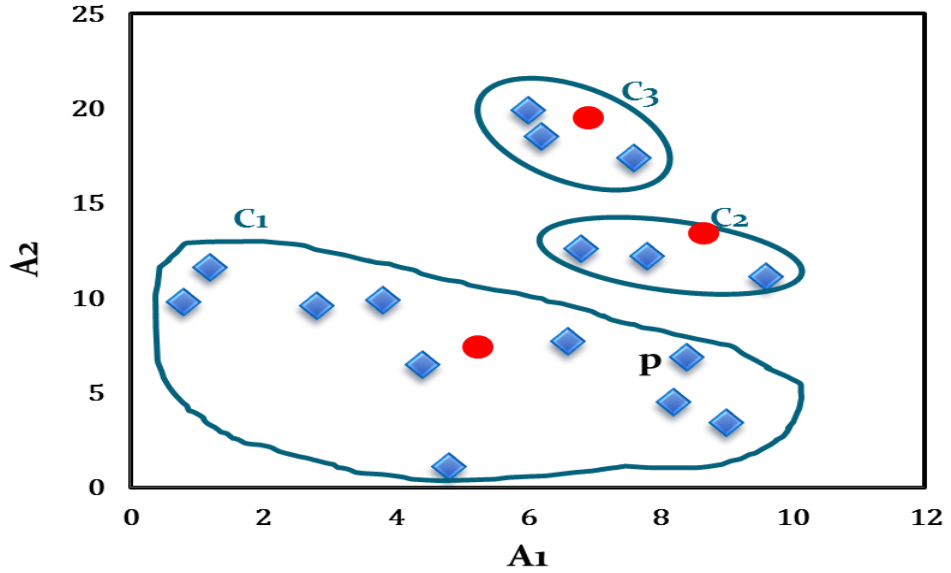
Illustration of k-Means clustering algorithms

- The newly obtained centroids after second iteration are given in the table below. Note that the centroid c_3 remains unchanged, where c_2 and c_1 changed a little.
- With respect to newly obtained cluster centres, 16 points are reassigned again. These are the same clusters as before. Hence, their centroids also remain unchanged.
- Considering this as the termination criteria, the k-means algorithm stops here. Hence, the final cluster in Fig 5 is same as Fig 4.

Cluster centres after second iteration

Centroid	Revised Centroids	
	A1	A2
c_1	5.0	7.1
c_2	8.1	12.0
c_3	6.6	18.6

Fig 5: Cluster after Second iteration



Comments on k-Means algorithm

Let us analyse the k-Means algorithm and discuss the pros and cons of the algorithm.

We shall refer to the following notations in our discussion.

- Notations:
  - x : an object under clustering
  - n : number of objects under clustering
  - $\boldsymbol{C}_i$: the *i-th* cluster
  - c_i : the centroid of cluster $\boldsymbol{C}_i$
  - n_i : number of objects in the cluster $\boldsymbol{C}_i$
  - c : denotes the centroid of all objects
  - k : number of clusters

Comments on k-Means algorithm

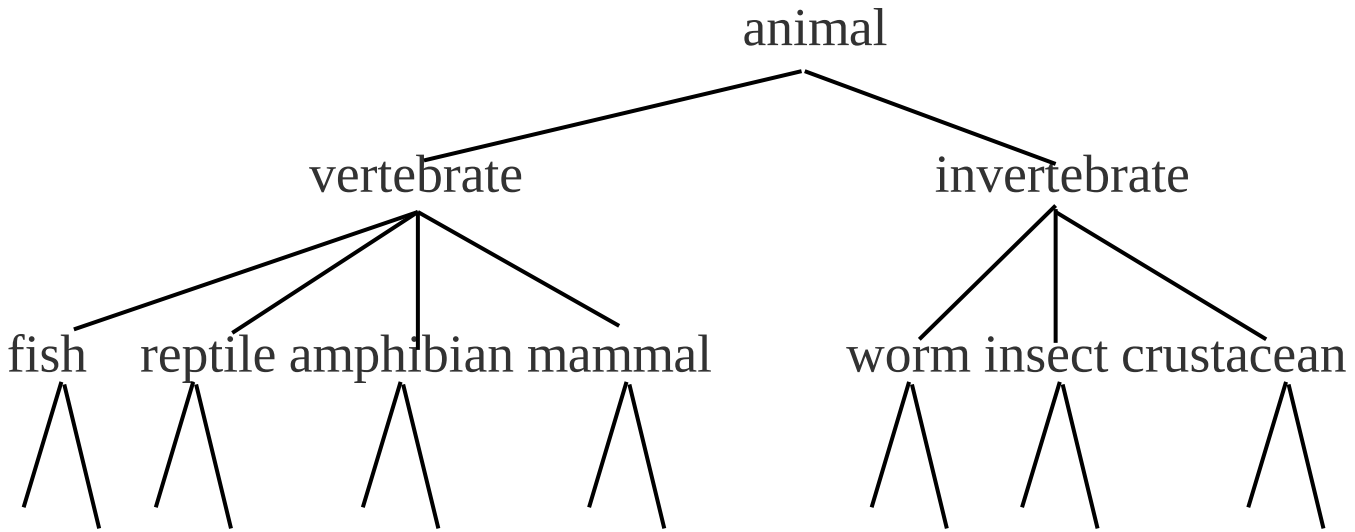
1. **Value of k:**
   - Normally $k \ll n$ and there is heuristic to follow $k \approx \sqrt{n}$.
2. **k versus cluster quality**
   - Usually, there is some objective function to be met as a goal of clustering. One such objective function is sum-square-error denoted by SSE and defined as

$$SSE = \sum_{i=1}^{k} \sum_{x \in \boldsymbol{C}_i} (x - c_i)^2$$

   - Here, $x - c_i$ denotes the error, if x is in cluster $\boldsymbol{C}_i$ with cluster centroid c_i.
   - Usually, this error is measured as distance norms like L_1, L_2 or Cosine similarity, etc.

Hierarchical (Agglomerative) clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*)



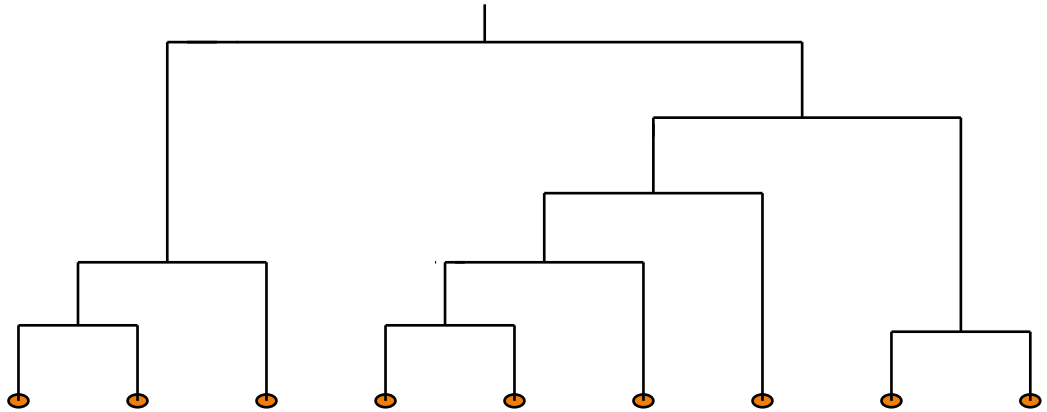
- One approach: recursive application of a partitional clustering algorithm.

Dendrogram: Hierarchical Clustering

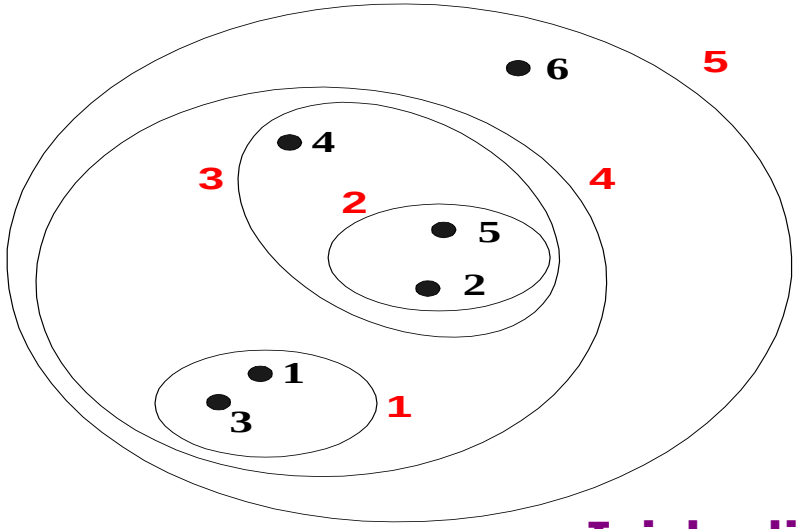
- **Dendrogram (**A *tree graph).* A graphical device for displaying clustering results.

  -Vertical lines represent clusters that are joined together.

  - Each node on the tree is a cluster; each leaf node is a singleton cluster



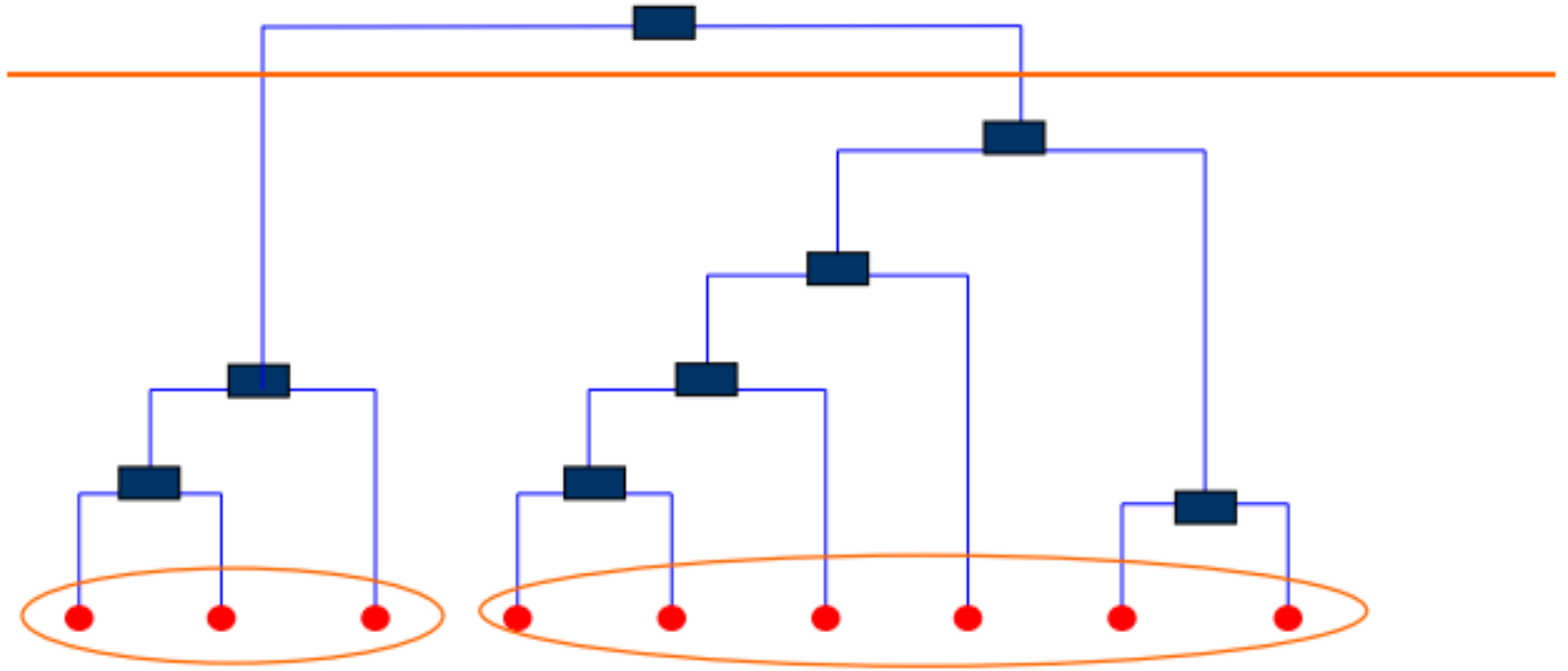
- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.



Icicle diagram.

Dendrogram: Hierarchical Clustering

- A clustering of the data objects is obtained by cutting the *dendrogram* at the desired level, then each connected component forms a cluster



Hierarchical Agglomerative Clustering- Algorithm

- Most popular hierarchical clustering technique
- Basic algorithm
  1. Compute the distance matrix between the input data points
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the distance matrix
  6. **Until** only a single cluster remains
- Key operation is the computation of the distance between two clusters
  - Different definitions of the distance between clusters lead to different algorithms

Select a Similarity Measure

- Similarity measure can be correlations or distances
- The most commonly used measure of similarity is the **Euclidean distance**. The *city-block* distance is also used.
- If variables measured in vastly different units, we must standardize data. Also eliminate outliers
- Use of different similarity/distance measures may lead to different clustering results.
- Hence, it is advisable to use different measures and compare the results.

Hierarchical Agglomerative Clustering-Linkage Method

- **Single-link distance** between clusters C_i and C_j is the ***minimum distance*** between any object in C_i and any object in C_j
  - The distance is **defined by the two most similar objects**

$$D_{sl}(C_i, C_j) = \min_{x,y} \{ d(x, y) | x \in C_i, y \in C_j \}$$

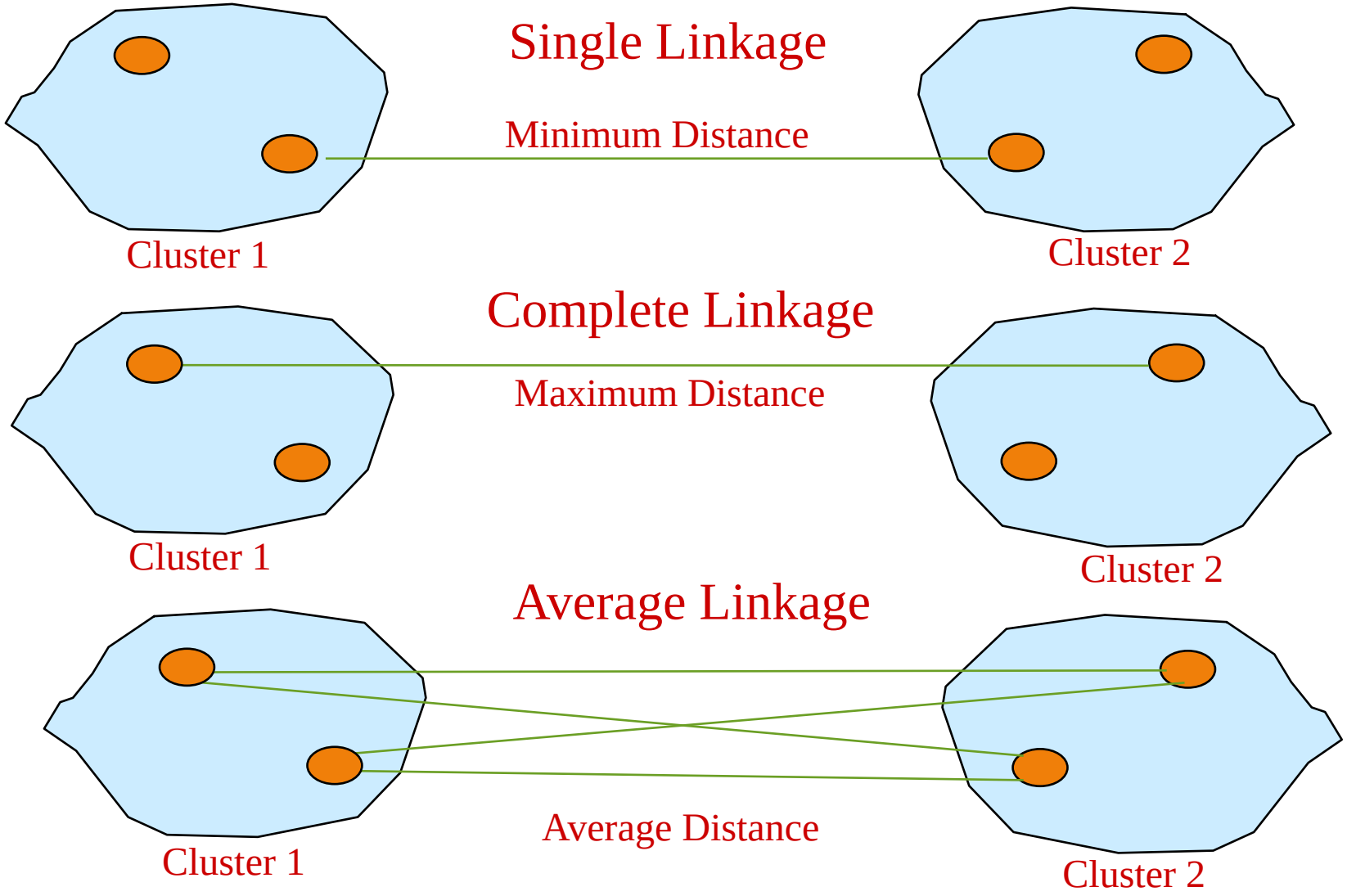
- **Complete-link distance** between clusters C_i and C_j is the ***maximum distance*** between any object in C_i and any object in C_j
  - The distance is **defined by the two most dissimilar objects**

$$D_{cl}(C_i, C_j) = \max_{x,y} \{ d(x, y) | x \in C_i, y \in C_j \}$$

- **Group average distance** between clusters C_i and C_j is the ***average distance*** between any object in C_i and any object in C_j

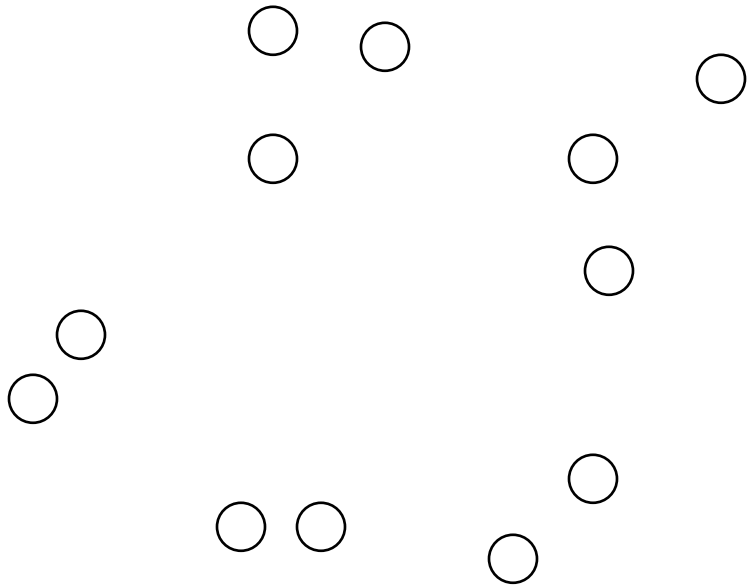
$$D_{avg}(C_i, C_j) = \frac{1}{|C_i| \times |C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

Linkage Methods of Clustering



Input/ Initial setting

- Start with clusters of individual points and a distance/proximity matrix

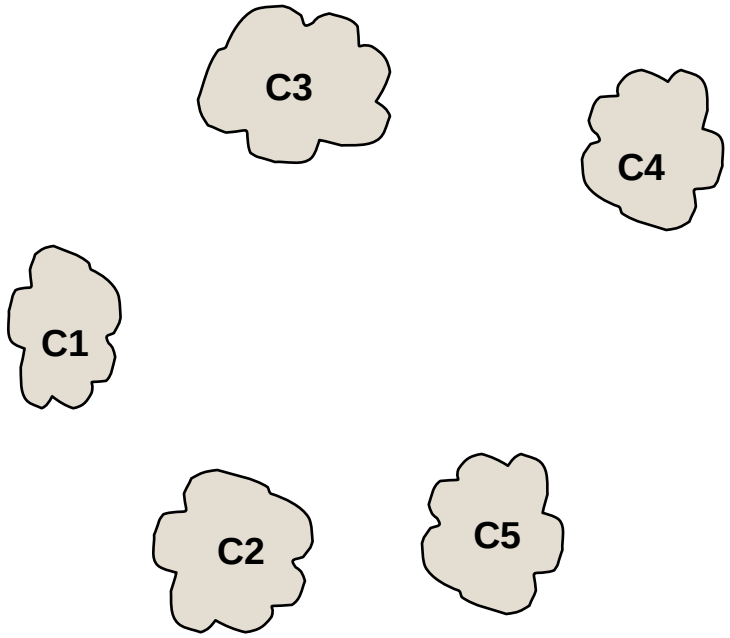


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						

Distance/Proximity Matrix

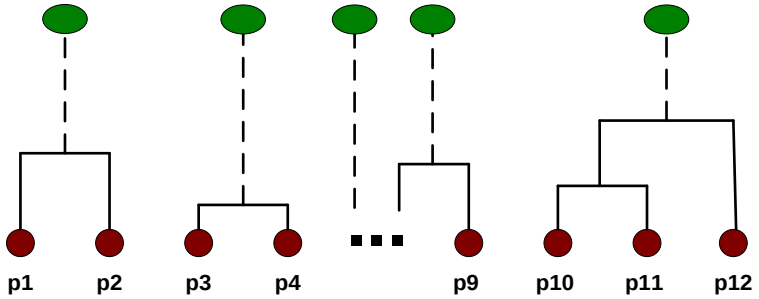
Intermediate State

- After some merging steps, we have some clusters



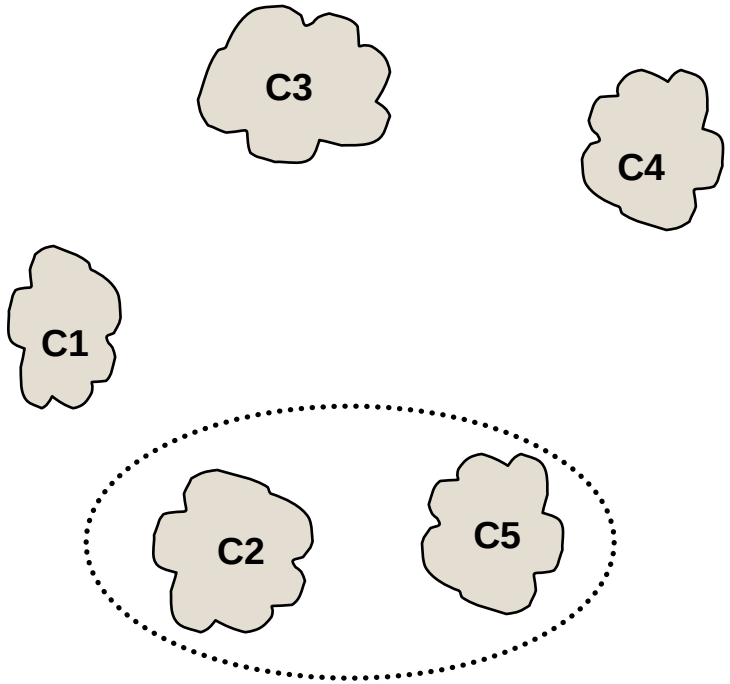
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance/Proximity Matrix



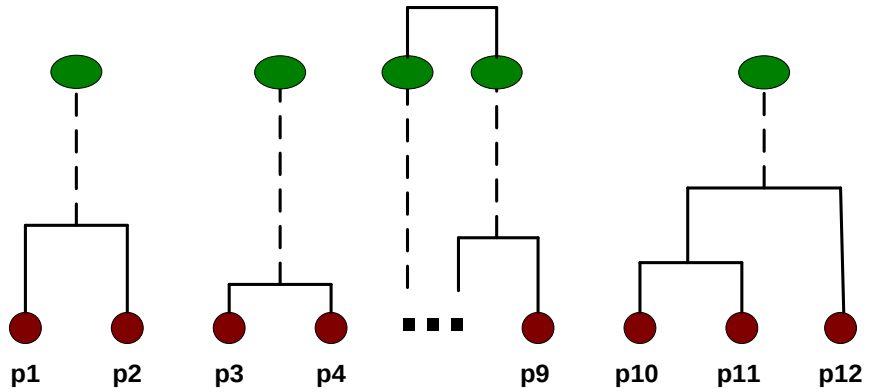
Intermediate State

- Merge the two closest clusters (C2 and C5) and update the distance matrix.



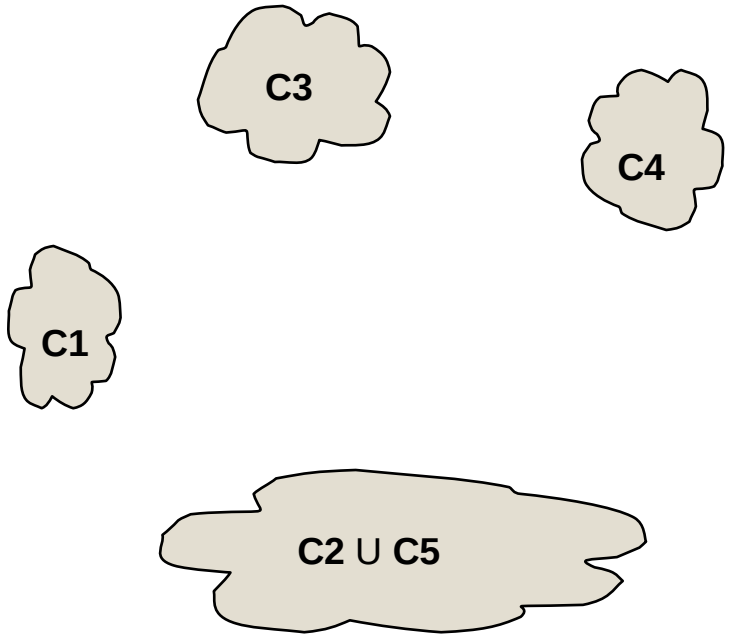
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance/Proximity Matrix

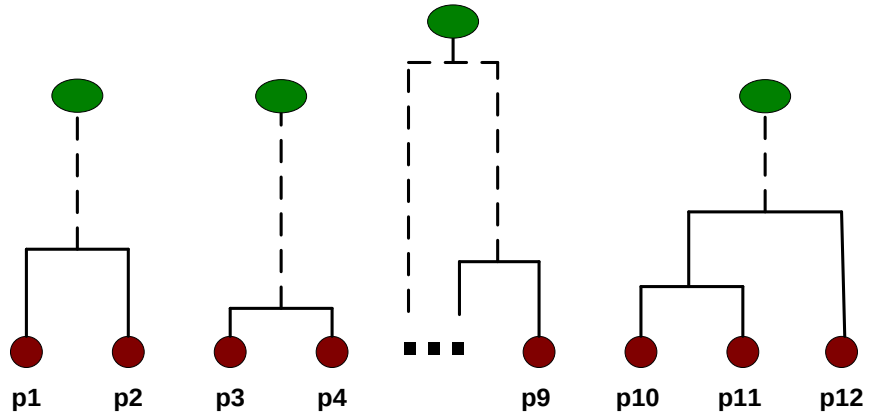


After Merging

- “How do we update the distance matrix?”



	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		



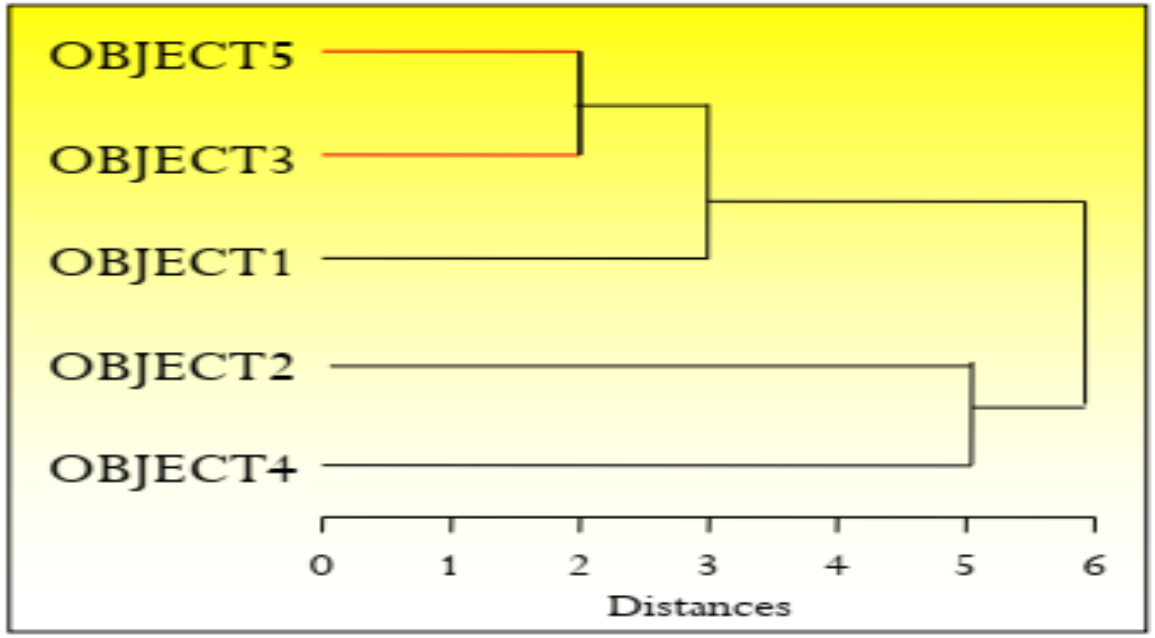
Example-Single linkage (nearest neighbour)

Object	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Distance matrix

Object	5,3	1	2	4
(5,3)	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

Distance	Cluster
0	1,2,3,4,5
2	(5, 3),1, 2, 4
3	(1, 3, 5), 2, 4
5	(1, 3, 5), (2, 4)
6	(1, 3, 5, 2, 4)



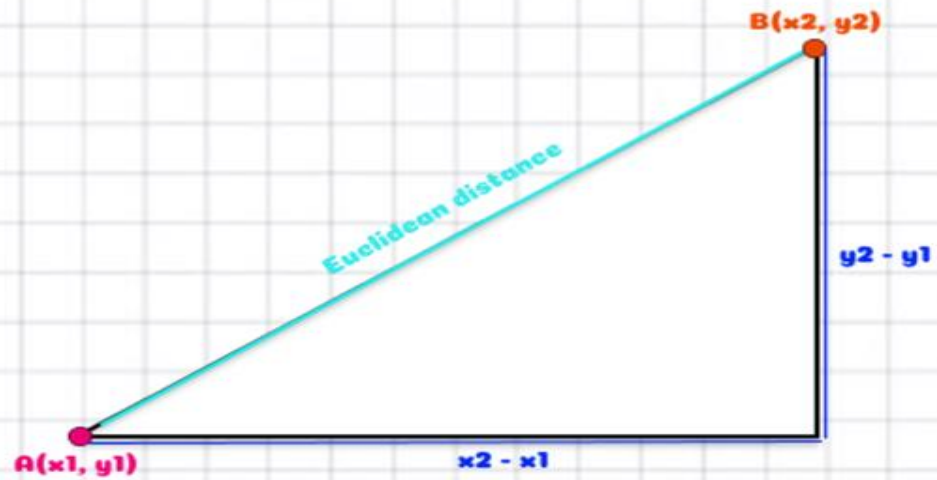
Cluster Tree

Popular Distance Metric

- Euclidean Distance
- Manhattan Distance
- Chebyshev distance

Euclidean Distance represents the shortest distance between two data points.

$$Euclidean\,(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Reference

- http://cse.iitkgp.ac.in/~dsamanta/
- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.
- Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.