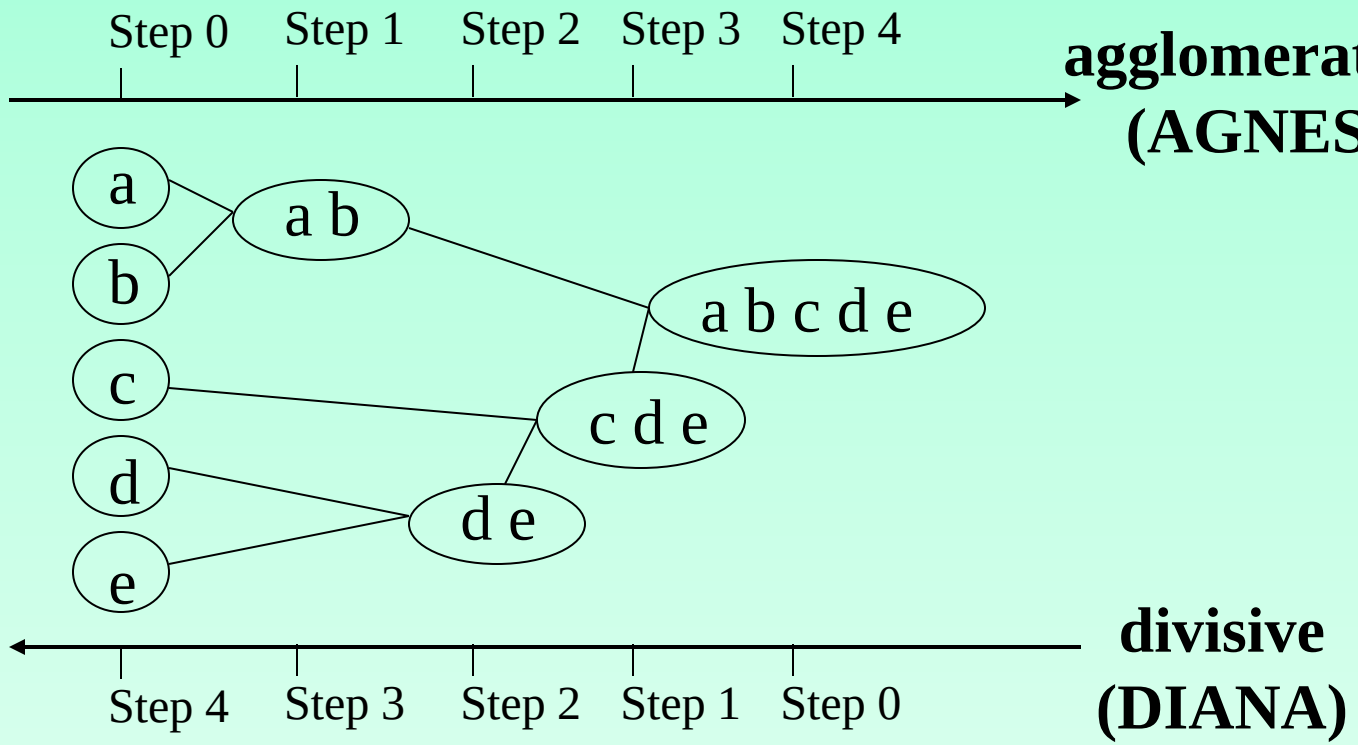
Hierarchical Clustering

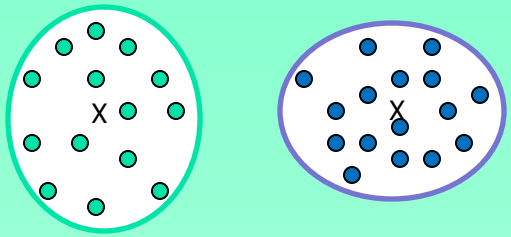
- These methods can be classified as either agglomerative or divisive
- Agglomerative clustering starts on “smaller" clusters, merging them into larger ones
- Divisive clustering works in the opposite direction
- Hierarchical clustering methods are common for several advantages: no need to specify the number of clusters; capability of handling high dimensional data
- However, this class of methods has high computing complexity .

Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



Distance between Clusters



- **Single link:** smallest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = min(t_{ip}, t_{jq})$
- **Complete link:** largest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = max(t_{ip}, t_{jq})$
- **Average:** avg distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = avg(t_{ip}, t_{jq})$

Example for hierarchical clustering

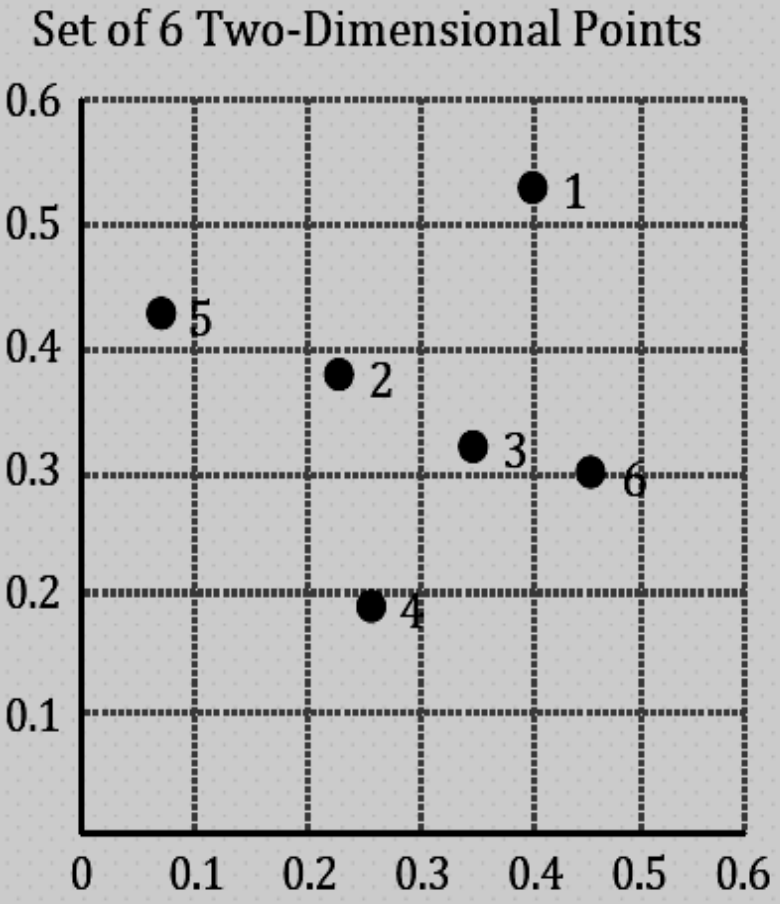
- **Problem:** Assume that the database D is given by the table below. Follow single link technique to find clusters in D. Use Euclidean distance measure.

Given D,

	x	**y**
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Example for hierarchical clustering

Step 1. Plot the objects in *n*-dimensional space (where *n* is the number of attributes). In our case we have 2 attributes – x and y, so we plot the objects p1, p2, … p6 in 2-dimensional space:



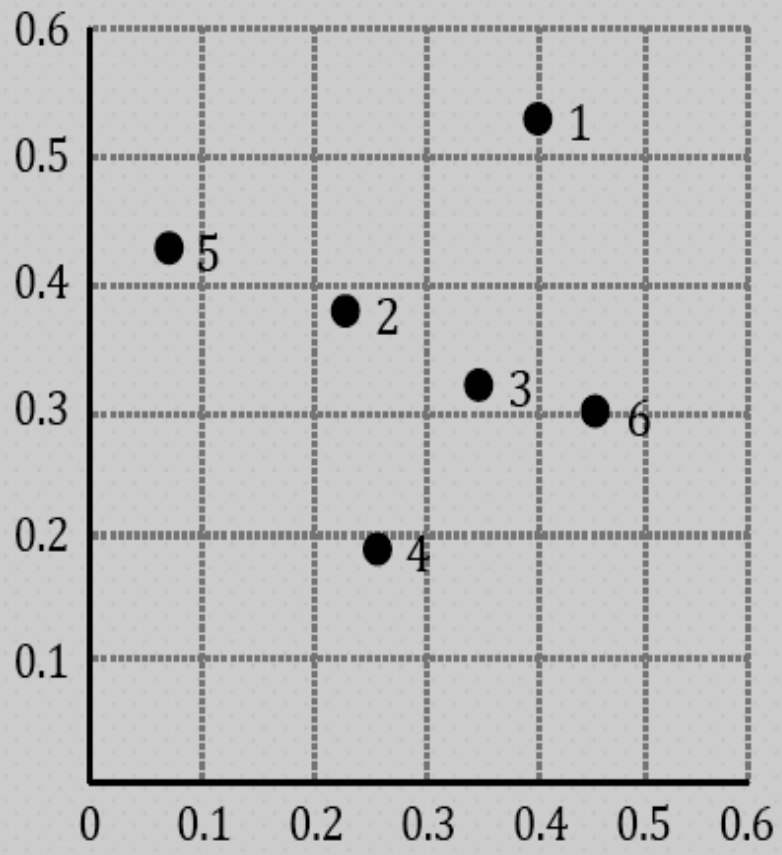
xy Coordinates of 6 Points

Point	*x* Coordinate	*y* Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Example for hierarchical clustering

Step 2. Calculate the distance from each object (point) to all other points, using Euclidean distance measure, and place the numbers in a distance matrix.

Set of 6 Two-Dimensional Points

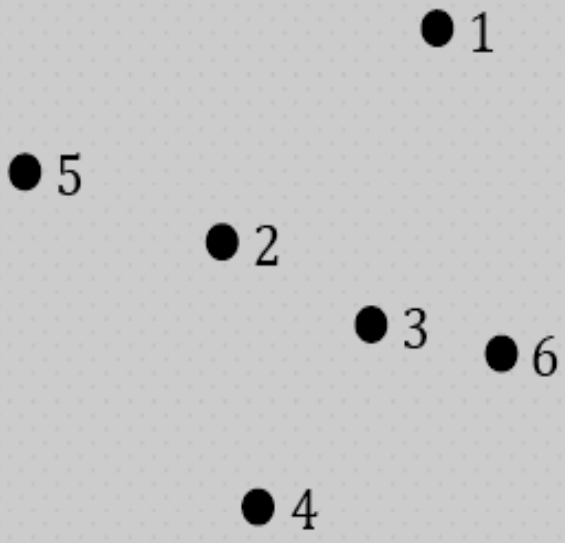


Euclidean Distance Matrix for 6 Points

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Example for hierarchical clustering

Nested Cluster Diagram



Single Link Distance Matrix

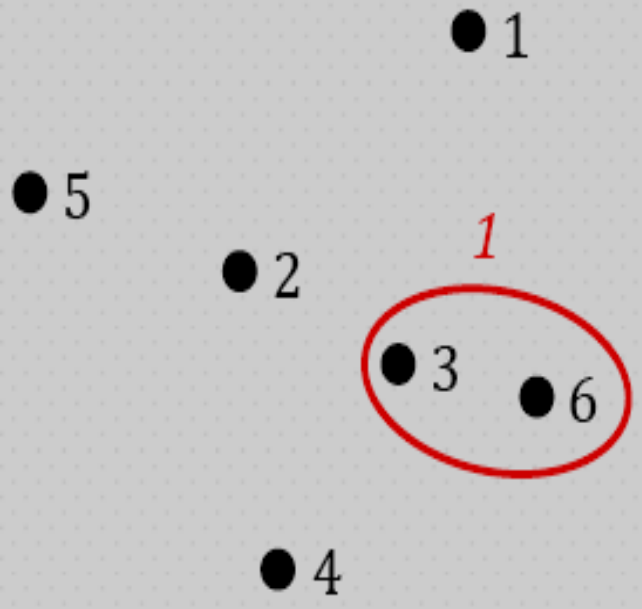
	1	2	3	4	5	6
1	0	0.24	0.22	0.37	0.34	0.23
2		0	0.15	0.20	0.14	0.25
3			0	0.15	0.28	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Step 3 : Identify the two clusters with the shortest distance in the matrix, and merge them together. Re-compute the distance matrix, as those two clusters are now in a single cluster, (no longer exist by themselves).

Example for hierarchical clustering

Points 3 and 6 have the smallest single link proximity distance. Merge these points into one cluster and update the distances to this new cluster. For example, the distance from point 1 to this cluster is 0.22 (the distance to point 3).

Nested Cluster Diagram



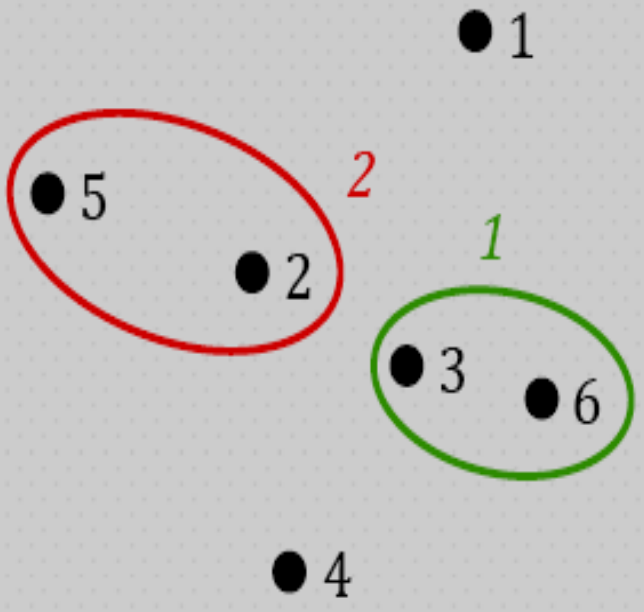
Single Link Distance Matrix

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Example for hierarchical clustering

Points 2 and 5 have the smallest single link proximity distance. Merge these points into one cluster and update the distances to this new cluster.

Nested Cluster Diagram



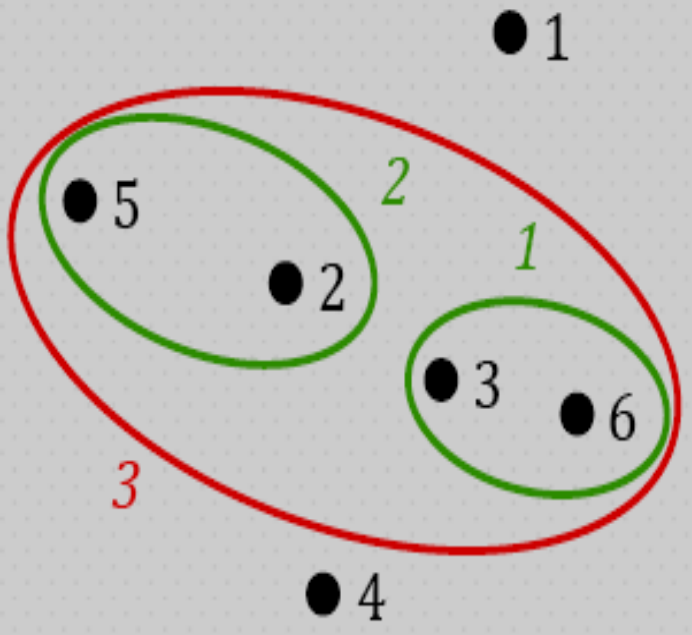
Single Link Distance Matrix

	1	2	4	5	3,6
1	0	*0.24*	0.37	0.34	0.22
2		0	*0.20*	0.14	0.15
4			0	0.29	0.15
5				0	0.28
3,6					0

Example for hierarchical clustering

And iterate…

Nested Cluster Diagram



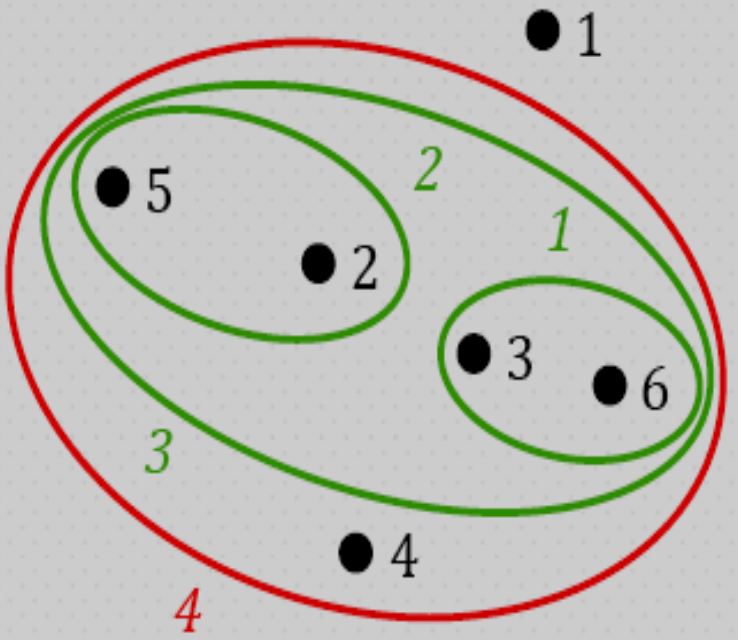
Single Link Distance Matrix

	1	4	2,5	3,6
1	0	0.37	0.24	*0.22*
4		0	0.20	*0.15*
2,5			0	0.15
3,6				0

Example for hierarchical clustering

And iterate…

Nested Cluster Diagram



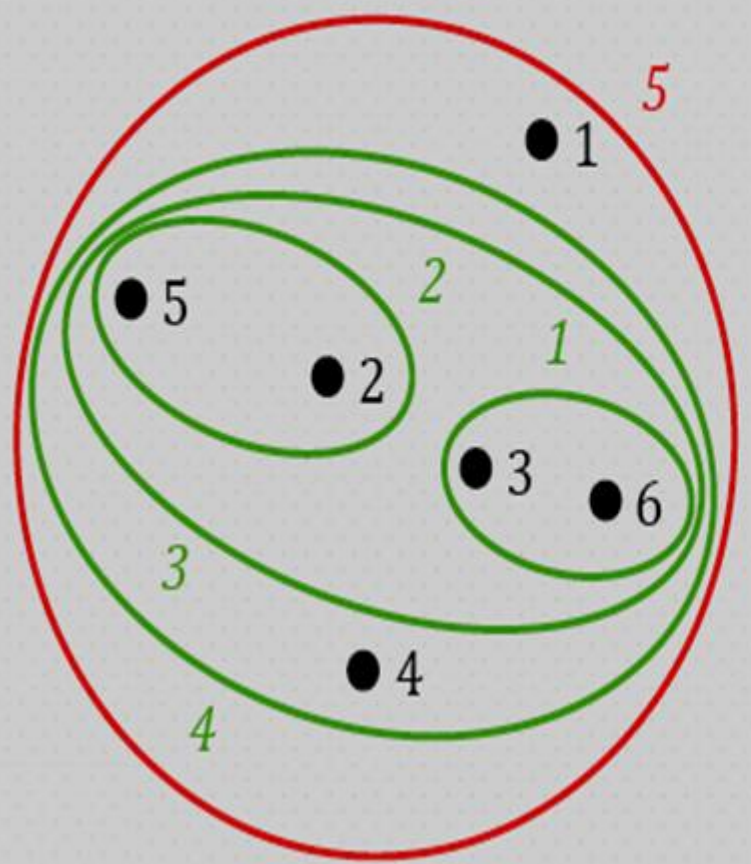
Single Link Distance Matrix

	1	4	2,5,3,6
1	0	0.37	*0.22*
4		0	0.15
2,5,3,6			0

Example for hierarchical clustering

And iterate until there is one all-inclusive cluster. …

Nested Cluster Diagram



Single Link Distance Matrix

	1	4,2,5,3,6
1	0	0.22
2,5,3,6		0

Example for hierarchical clustering