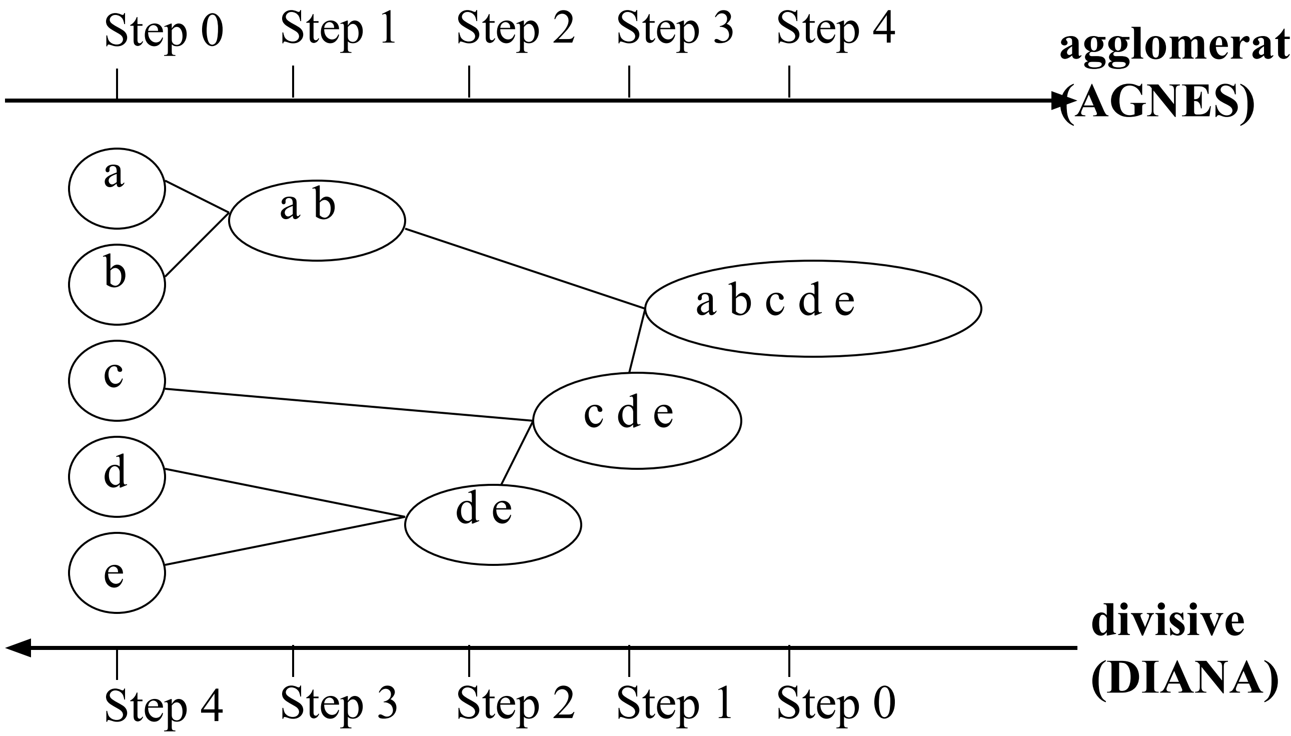
Hierarchical clustering

Hierarchical Clustering algorithms

- **Agglomerative (bottom-up):**
  - Start with each document being a single cluster.
  - Eventually all documents belong to the same cluster.
- **Divisive (top-down):**
  - Start with all documents belong to the same cluster.
  - Eventually each node forms a cluster on its own.
- Does not require the number of clusters ***k*** in advance
- Needs a termination/readout condition
  - The final mode in both Agglomerative and Divisive is of no use.

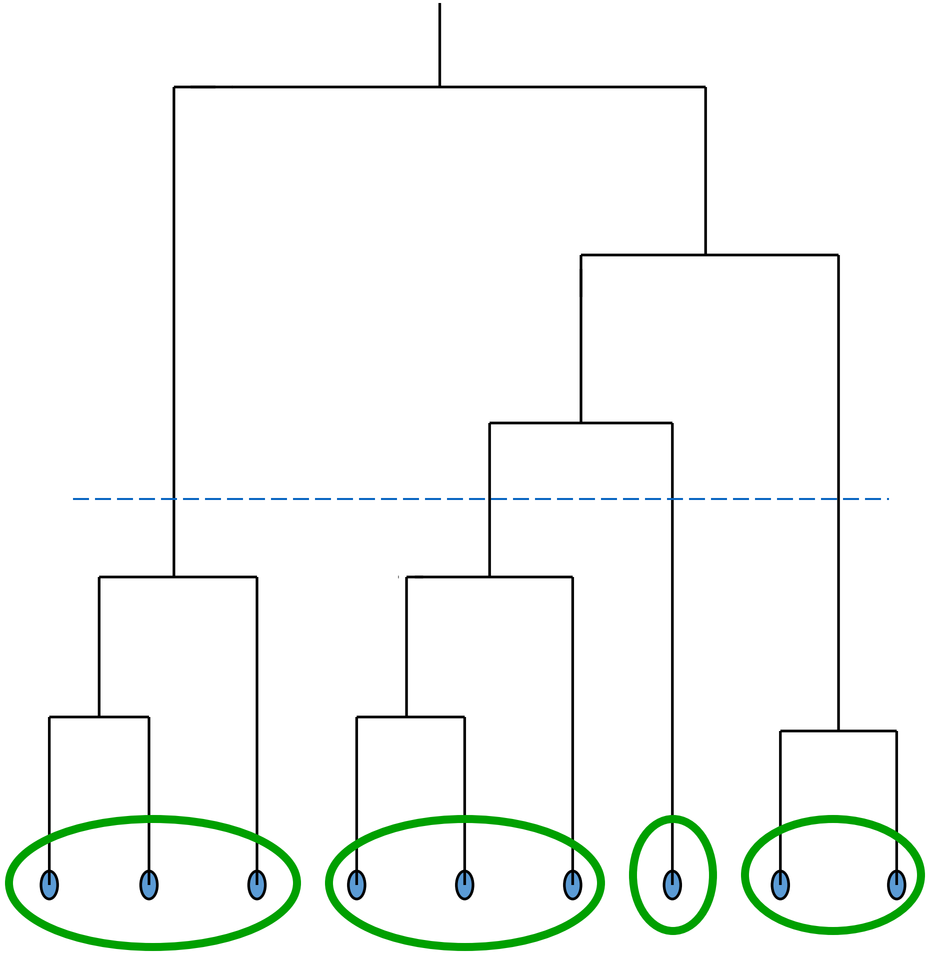
Hierarchical Clustering

- **Agglomerative: Bottom up approach**
- **Divisive :Top down approach**



Dendogram: Hierarchical Clustering

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

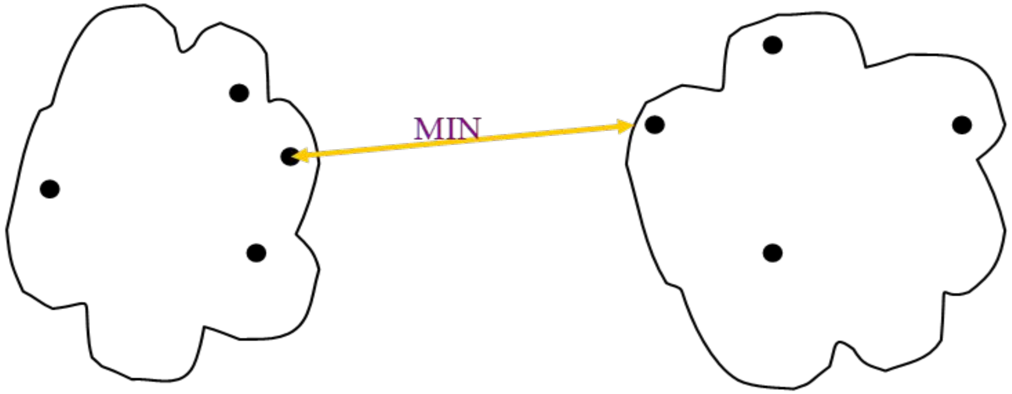


Agglomerative Hierarchical clustering method

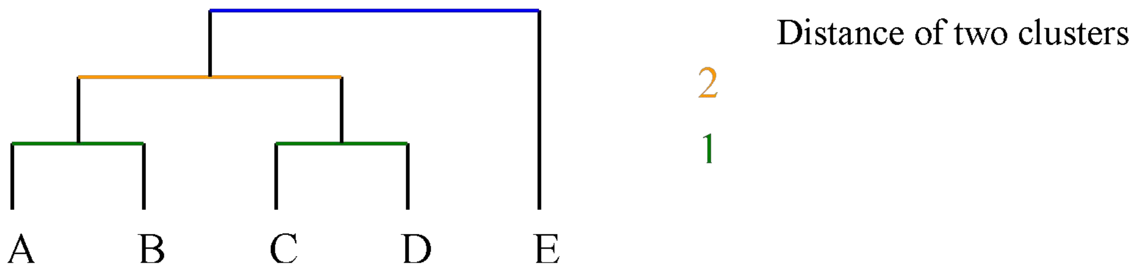
- Single link algorithm
- Complete link algorithm
- Average link algorithm

Single Link Clustering

- Single link algorithm is an example of agglomerative hierarchical clustering method.
- We recall that is a bottom-up strategy: compare each point with each point. Each object is placed in a separate cluster, and at each step we merge the closest pair of clusters, until certain termination conditions are satisfied. This requires defining a notion of cluster proximity.
- For the single link, the proximity of two clusters is defined as the minimum of the distance between any two points in the two clusters.



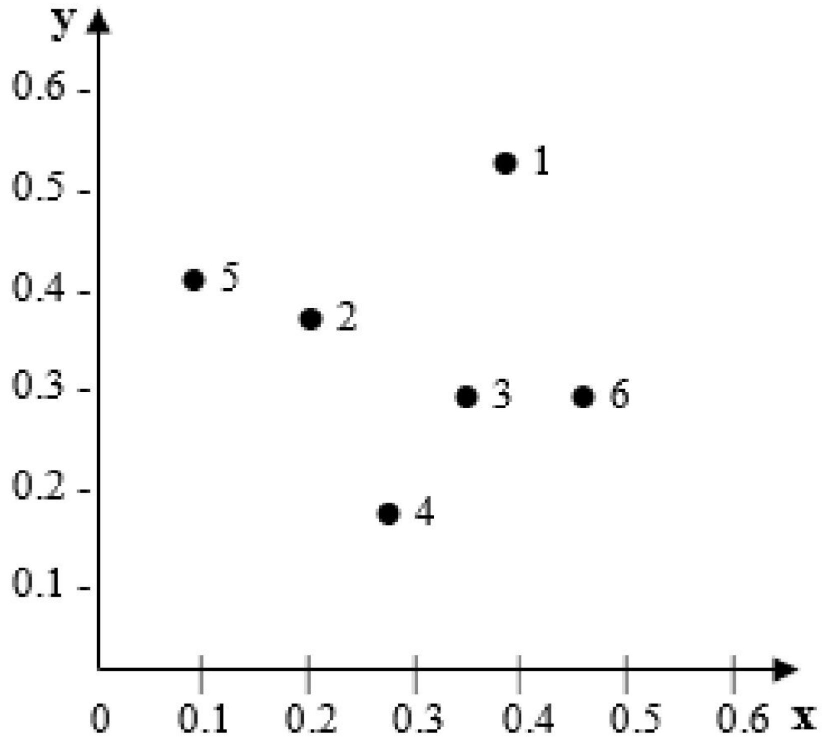
- **Dendrogram** – shows the same information as in the graph above.
- However distance threshold is vertical, and points are at the bottom (horizontal).
- The height at which two clusters are merged in the dendrogram reflects the distance of the two clusters



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

	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

- Solution:
- <u>Step 1.</u> 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:



- 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.

$$D(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j1}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

$$d(p1, p2) = |x_{p1} - x_{p1}|^2 + |y_{p1} - y_{p2}|^2$$

$$= \sqrt{|0.40 - 0.22|^2 + |0.53 - 0.38|^2}$$

$$= \sqrt{|0.18|^2 + |0.15|^2}$$

$$= \sqrt{0.0324 + 0.0225}$$

$$= \sqrt{0.0549}$$

$$= 0.2343$$