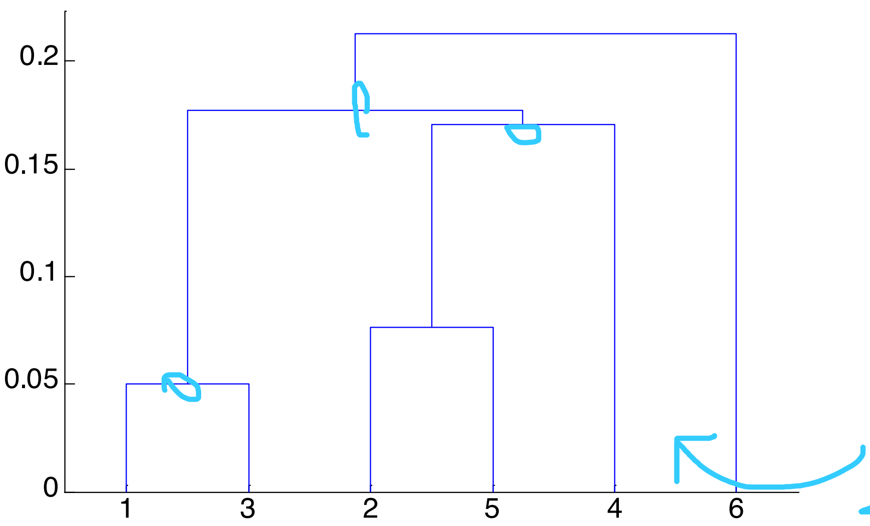
CLUSTERING

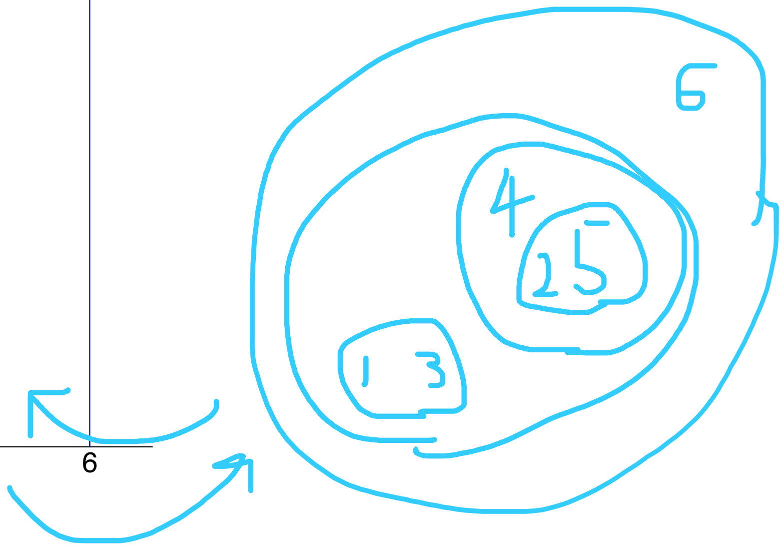
CLUSTERING ALGORITHMS

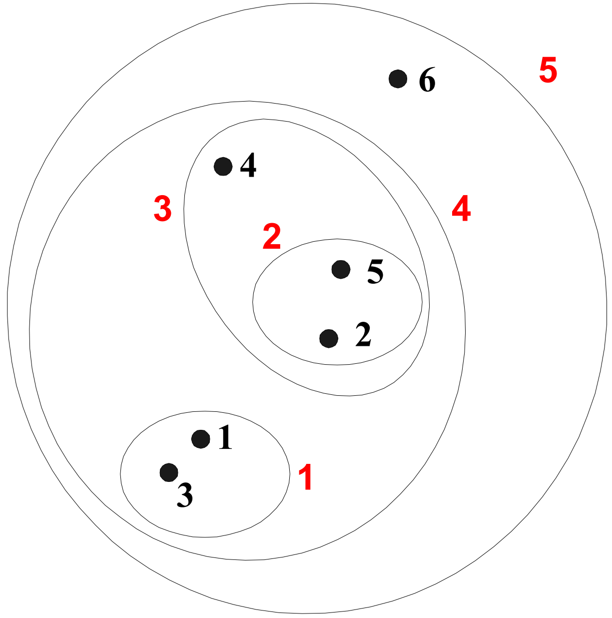
- K-means and its variants
  - What is k-means / algorithm, how to select the initial centroids for k-means (k-means++, bisecting k-means), limitations of k-means (size, density, shape) and the methods to overcome the limitations.
  - partition based clustering (each cluster / group is not overlapped with any other cluster / group.
- Hierarchical clustering
- Density-based clustering

HIERARCHICAL CLUSTERING

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits



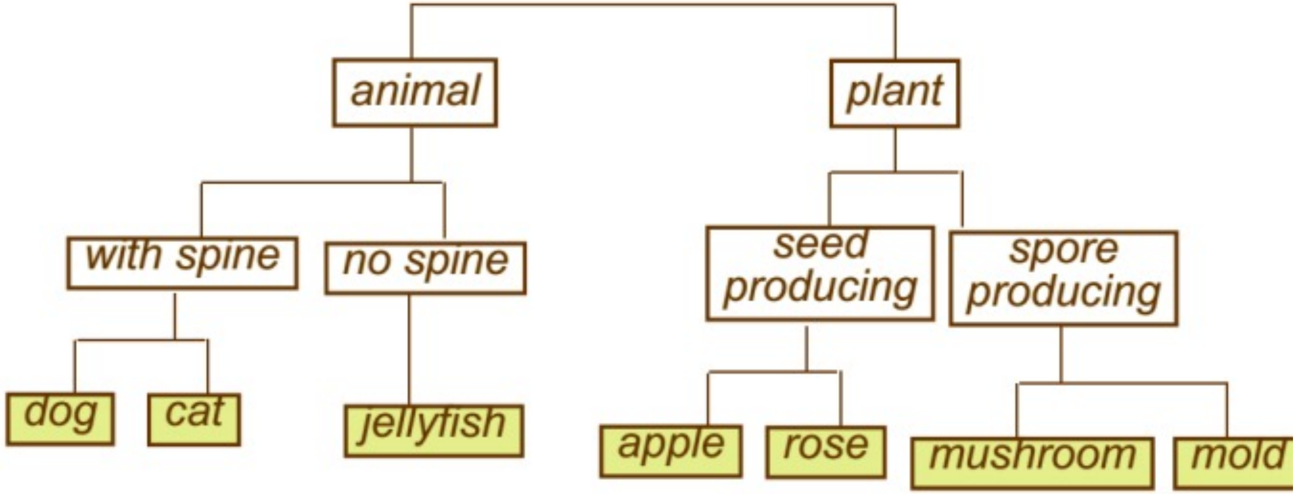




STRENGTHS OF HIERARCHICAL CLUSTERING

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level

- They may correspond to meaningful taxonomies
  - Example: biological science

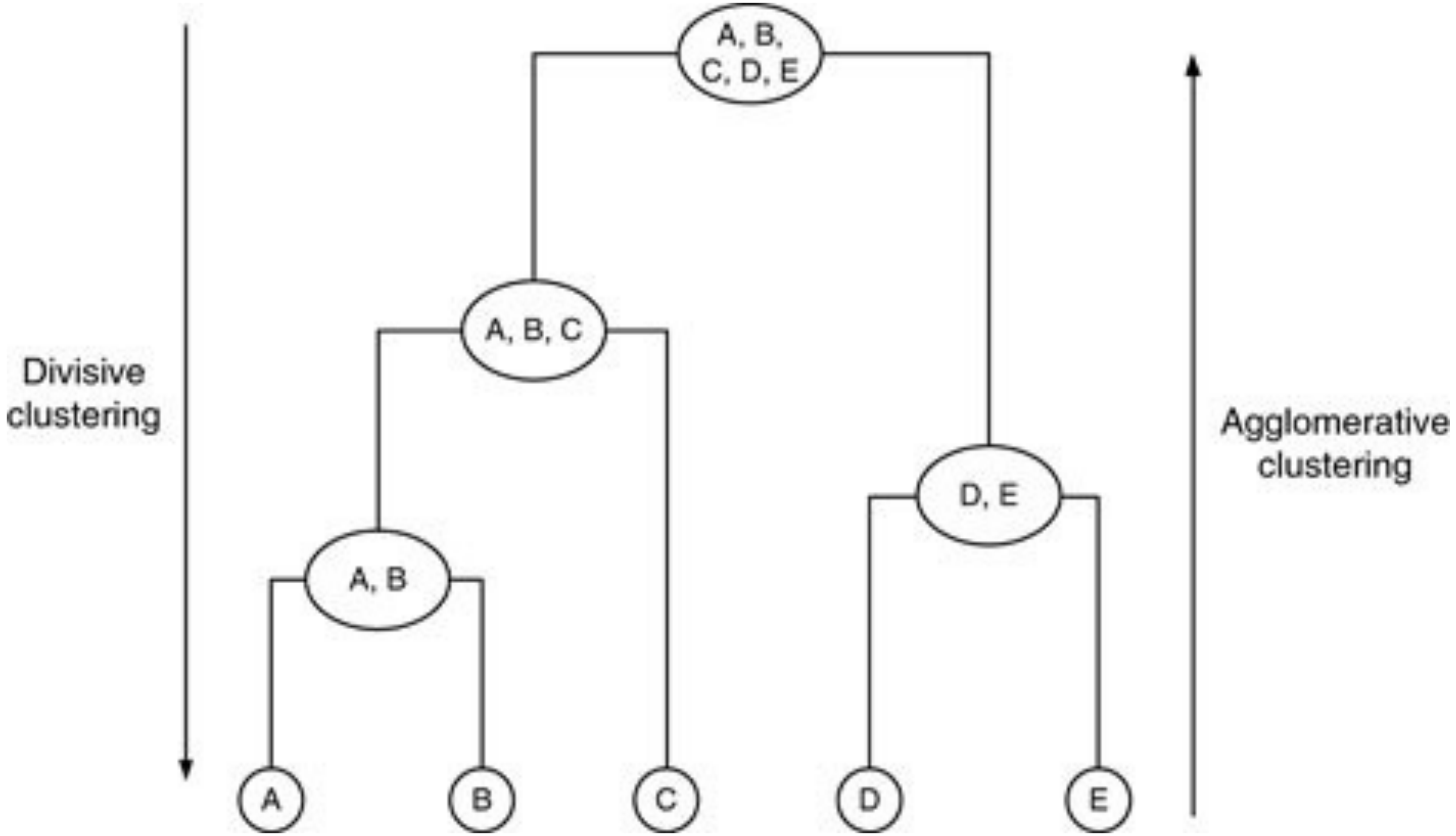


HIERARCHICAL CLUSTERING

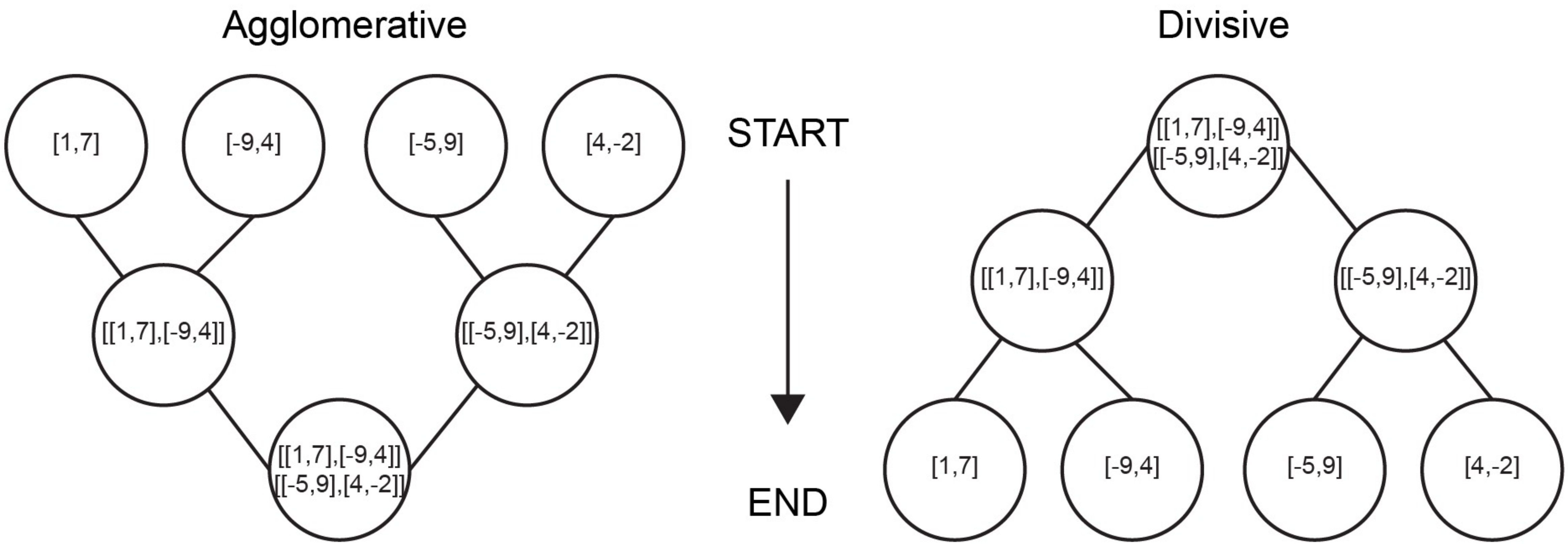
- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains an individual point (or there are k clusters)

- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

HIERARCHICAL CLUSTERING



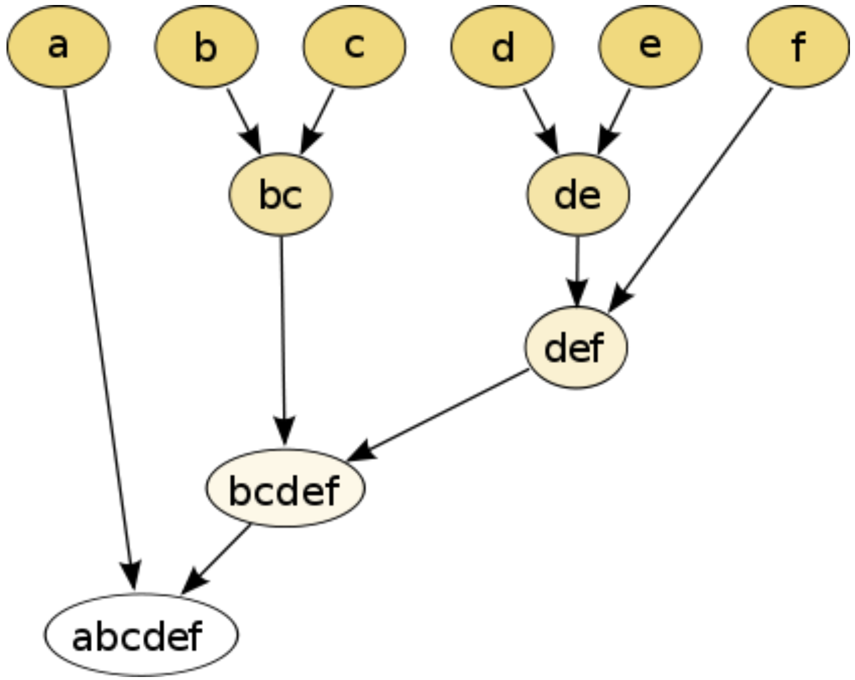
HIERARCHICAL CLUSTERING



AGGLOMERATIVE CLUSTERING ALGORITHM

- **Key Idea: Successively merge closest clusters**

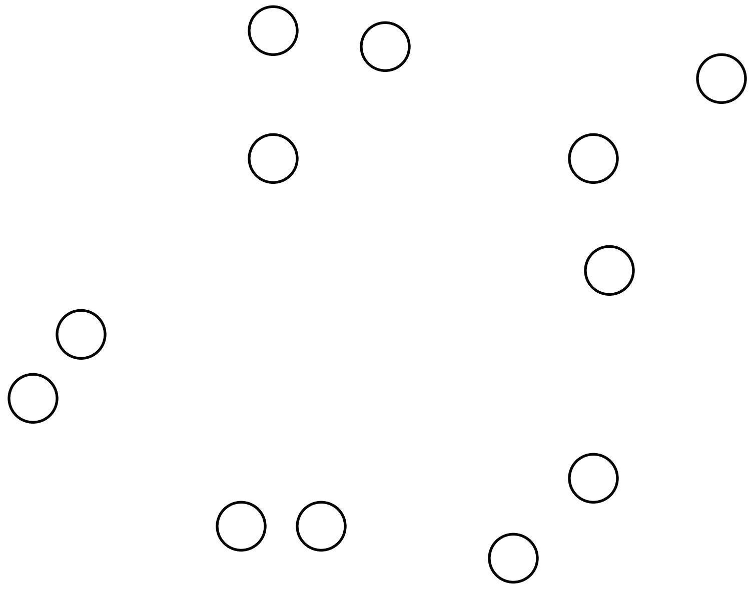
- Basic algorithm
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains



- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

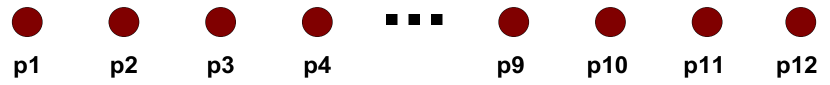
STEPS 1 AND 2

- Start with clusters of individual points and a proximity matrix



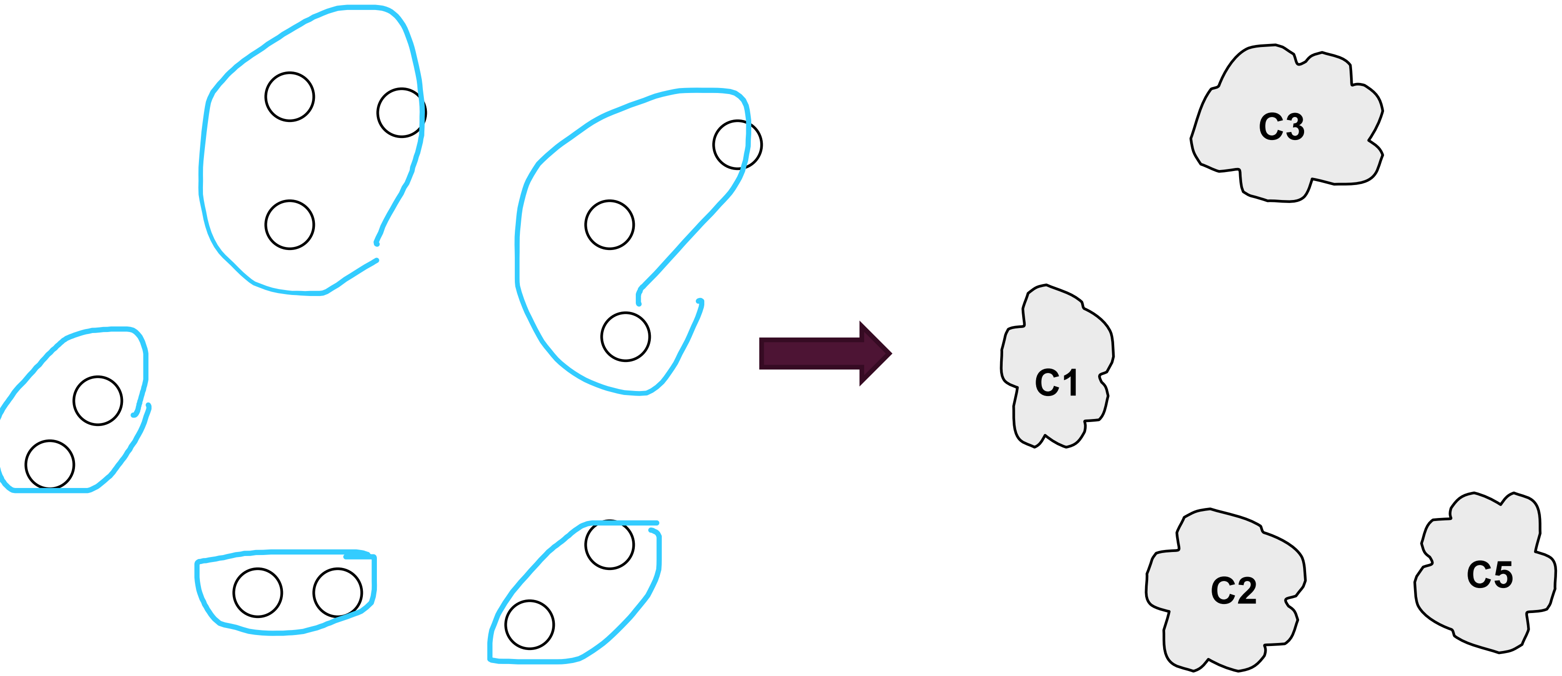
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



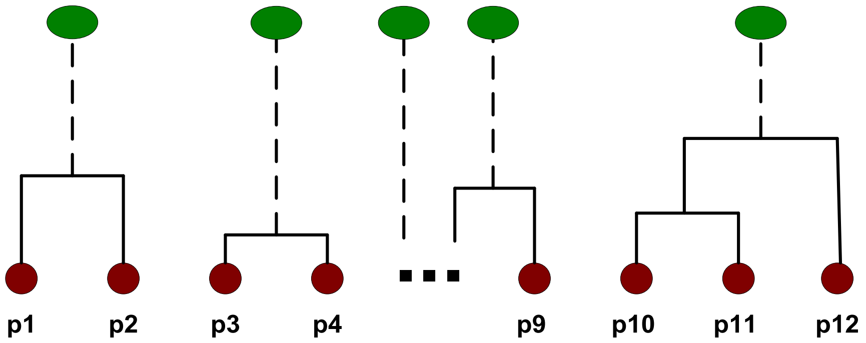
INTERMEDIATE SITUATION

- After some merging steps, we have some clusters



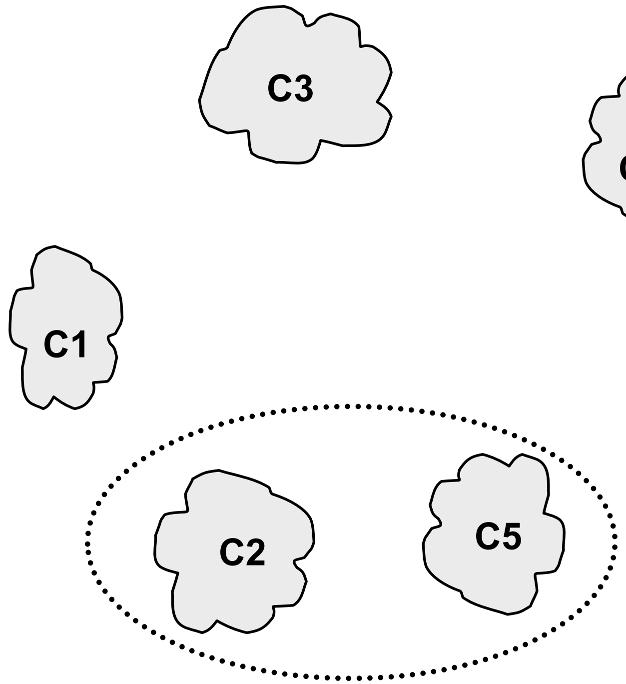
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

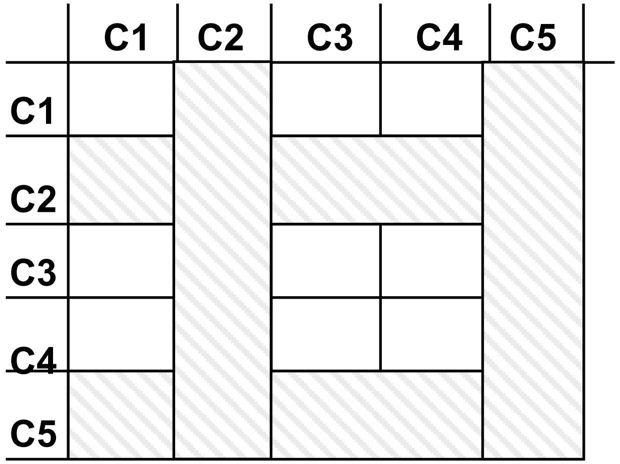
Proximity Matrix



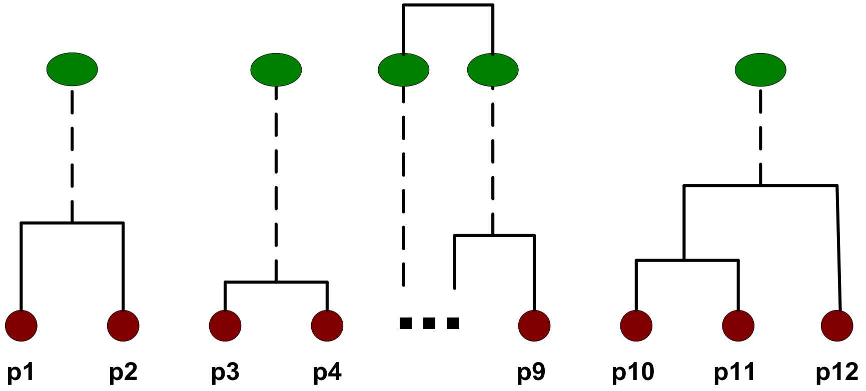
STEP 4

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



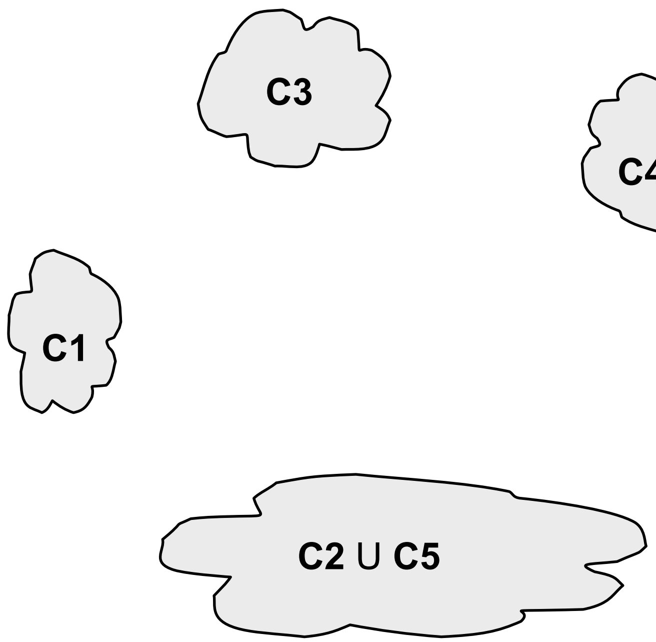


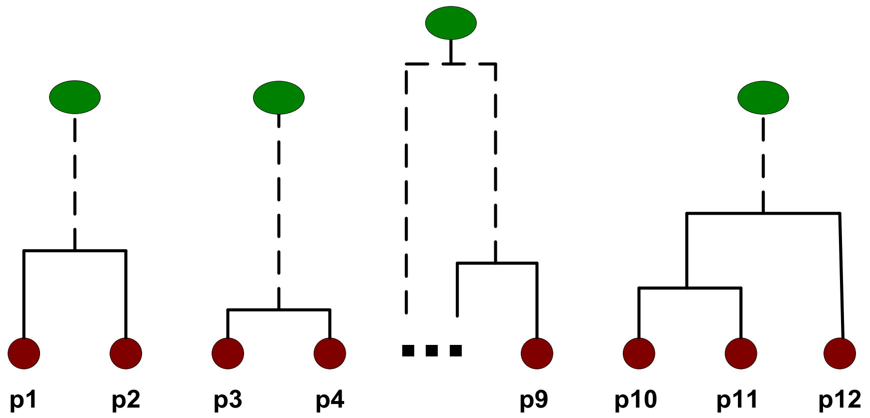
Proximity Matrix



STEP 5

- The question is "How do we update the proximity matrix?"

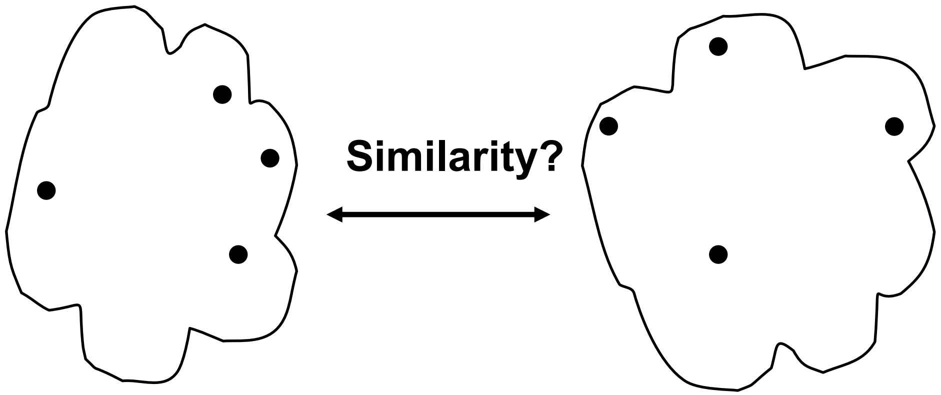




	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix

HOW TO DEFINE INTER-CLUSTER DISTANCE

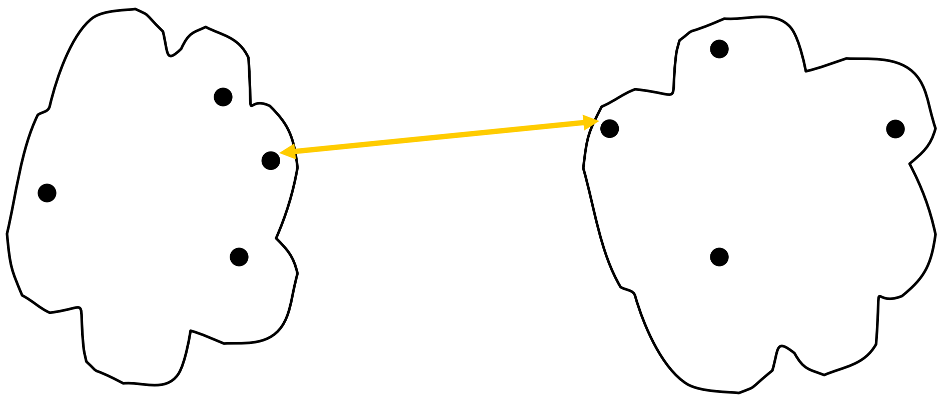


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward’s Method uses squared error

HOW TO DEFINE INTER-CLUSTER SIMILARITY

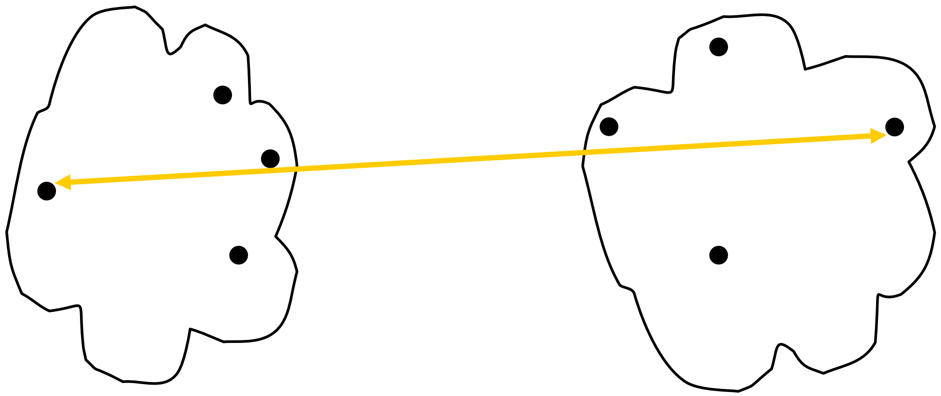


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

HOW TO DEFINE INTER-CLUSTER SIMILARITY

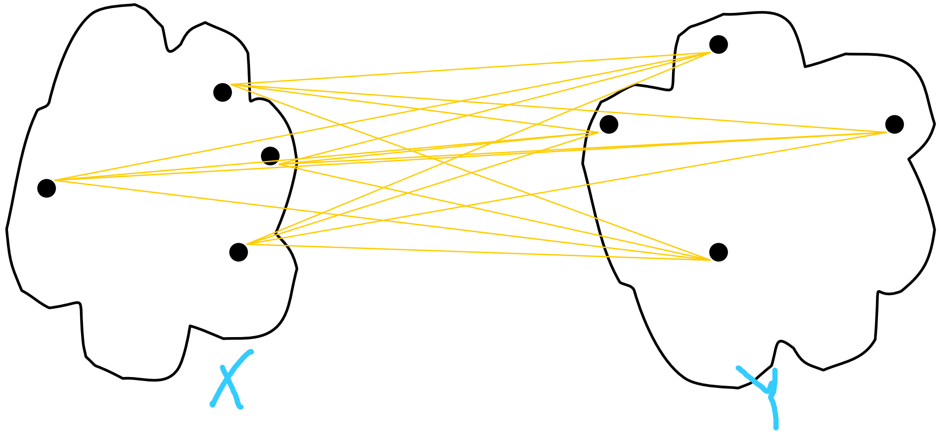


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  – Ward's Method uses squared error

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

Proximity Matrix

HOW TO DEFINE INTER-CLUSTER SIMILARITY



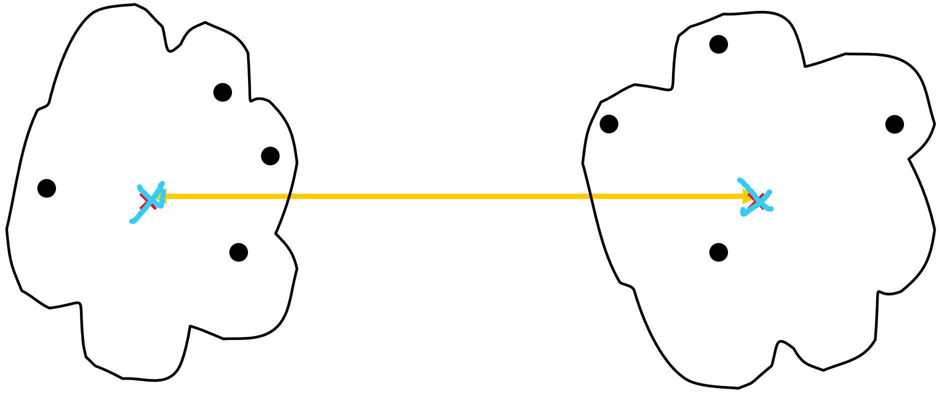
- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

$$\frac{\sum d(x_i, y_i)}{16}$$

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

Proximity Matrix

HOW TO DEFINE INTER-CLUSTER SIMILARITY



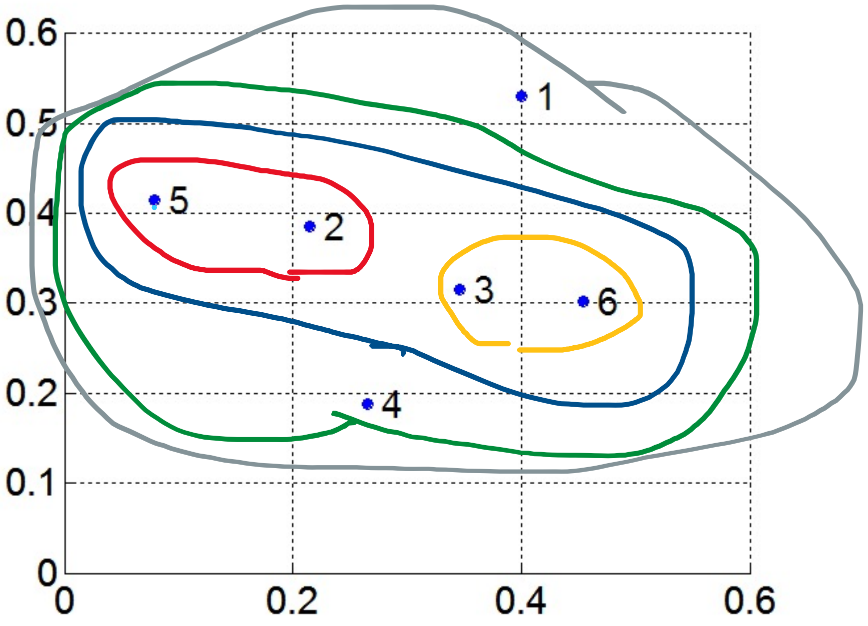
- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

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

Proximity Matrix

MIN OR SINGLE LINK

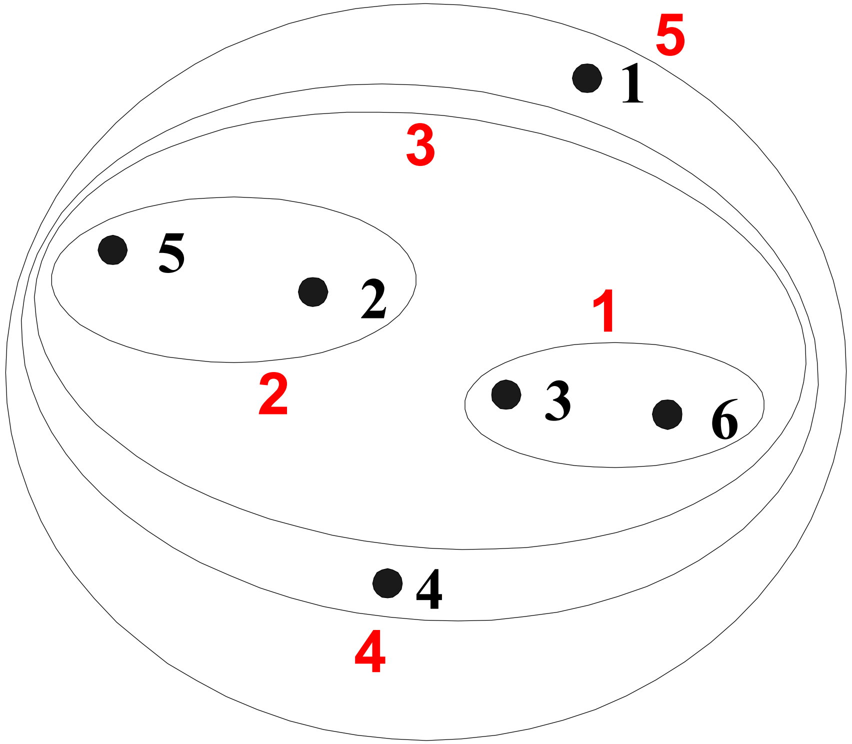
- Proximity of two clusters is based on the two closest points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph
- Example:



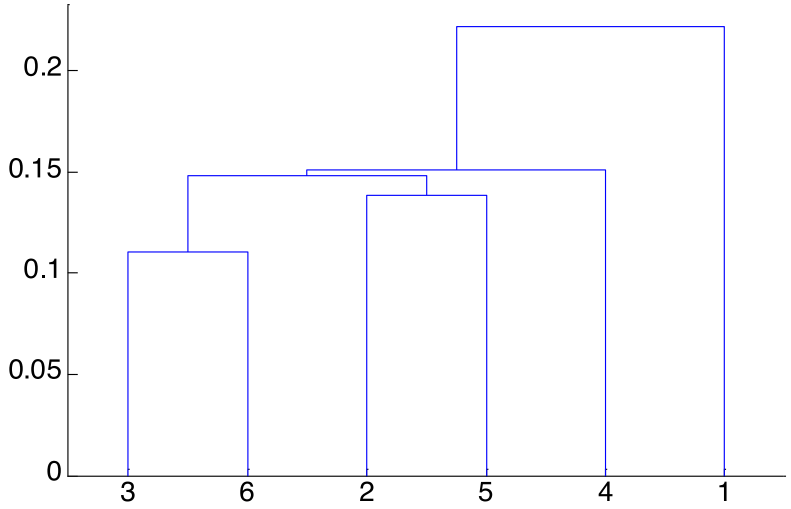
Distance Matrix:

	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

HIERARCHICAL CLUSTERING: MIN

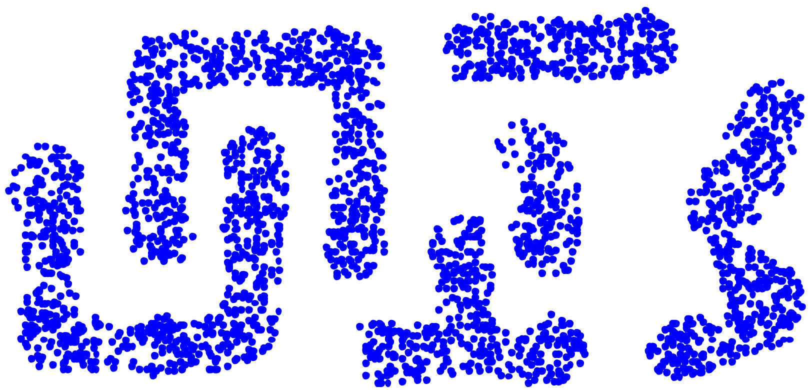


Nested Clusters

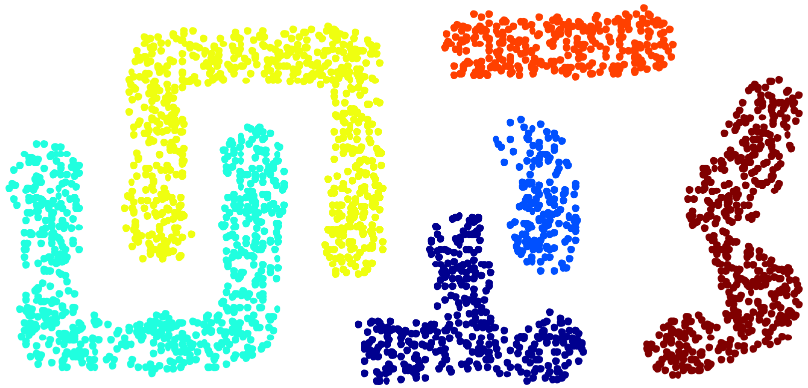


Dendrogram

STRENGTH OF MIN



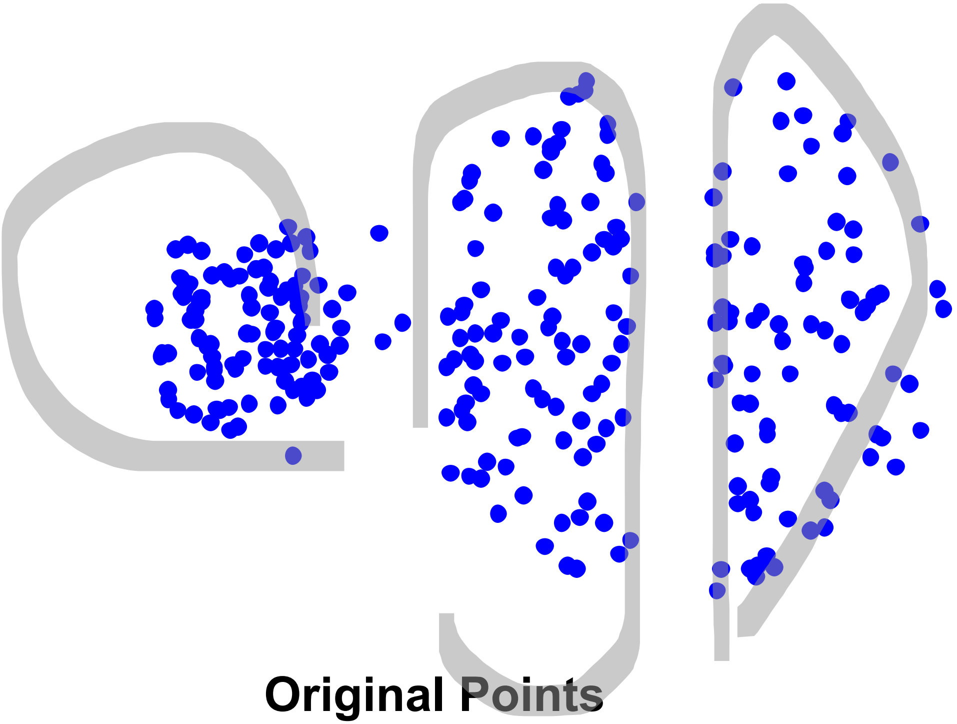
Original Points



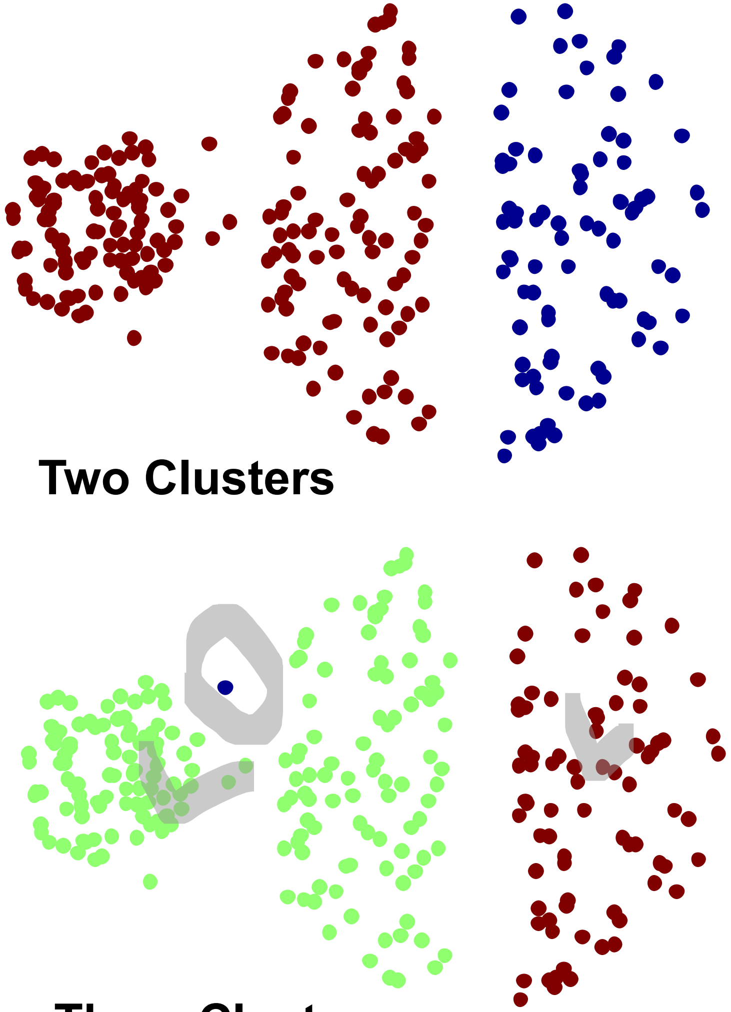
Six Clusters

- Can handle non-elliptical shapes

LIMITATIONS OF MIN



Original Points



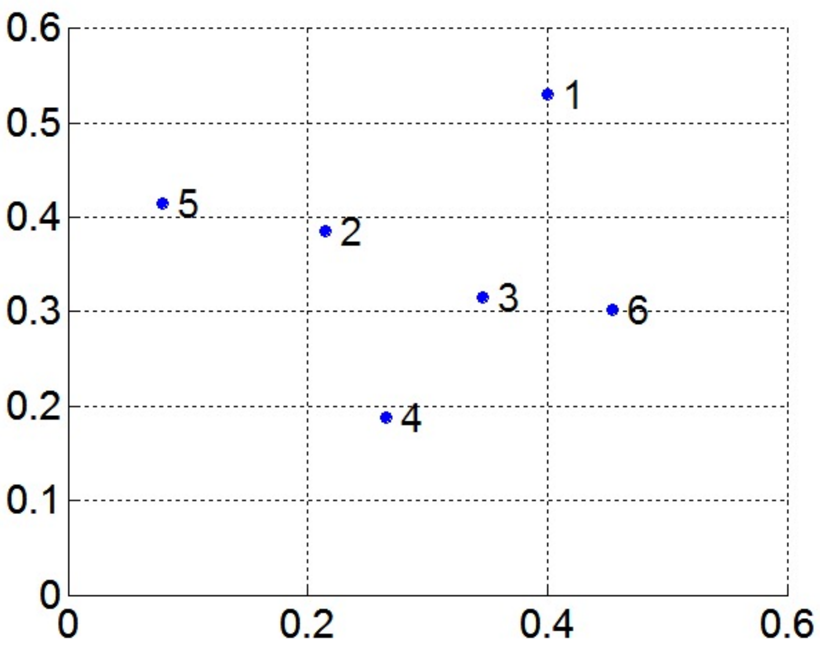
Two Clusters

Three Clusters

• Sensitive to noise

MAX OR COMPLETE LINKAGE

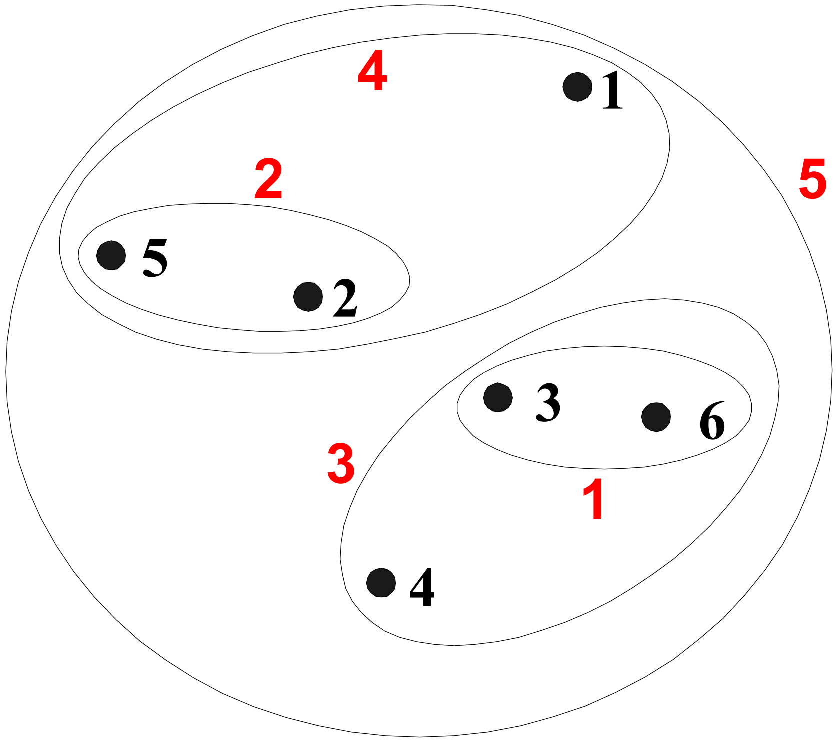
- Proximity of two clusters is based on the two most distant points in the different clusters
  - Determined by all pairs of points in the two clusters



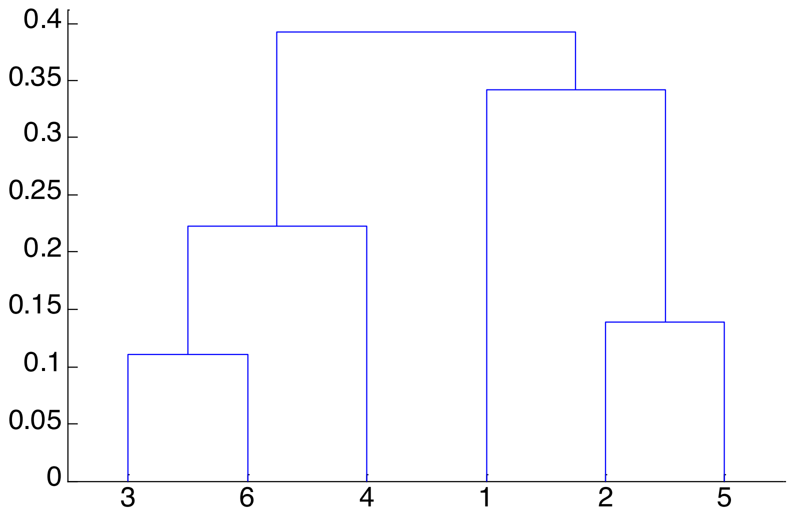
Distance Matrix:

	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

HIERARCHICAL CLUSTERING: MAX

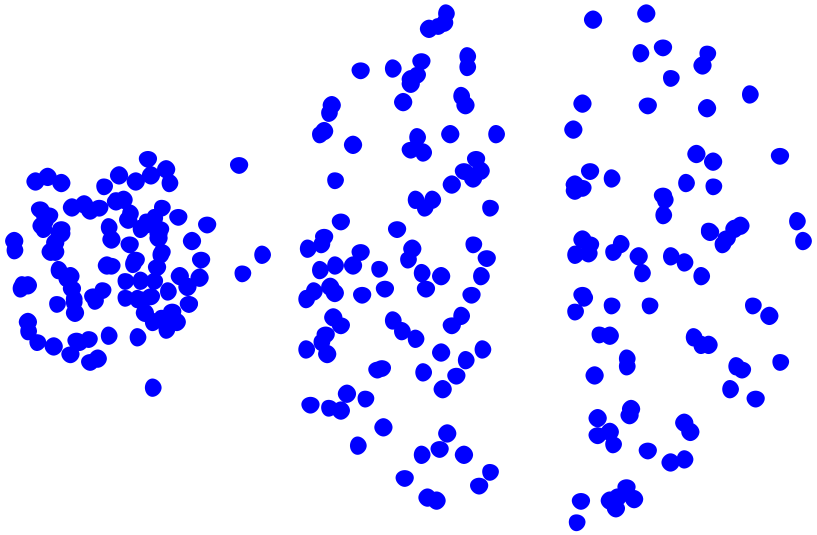


Nested Clusters

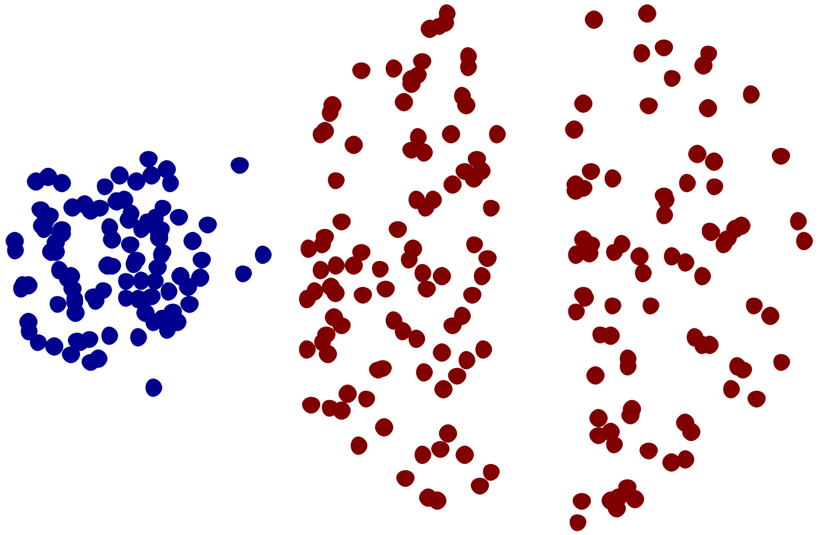


Dendrogram

STRENGTH OF MAX



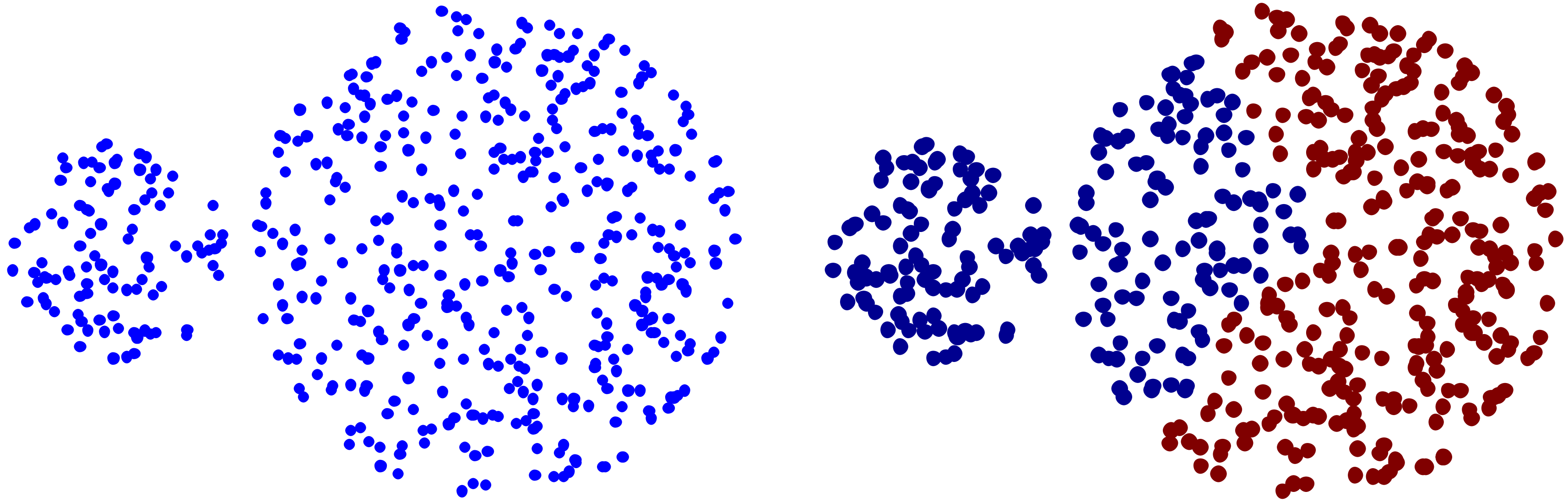
Original Points



Two Clusters

- Less susceptible to noise

LIMITATIONS OF MAX



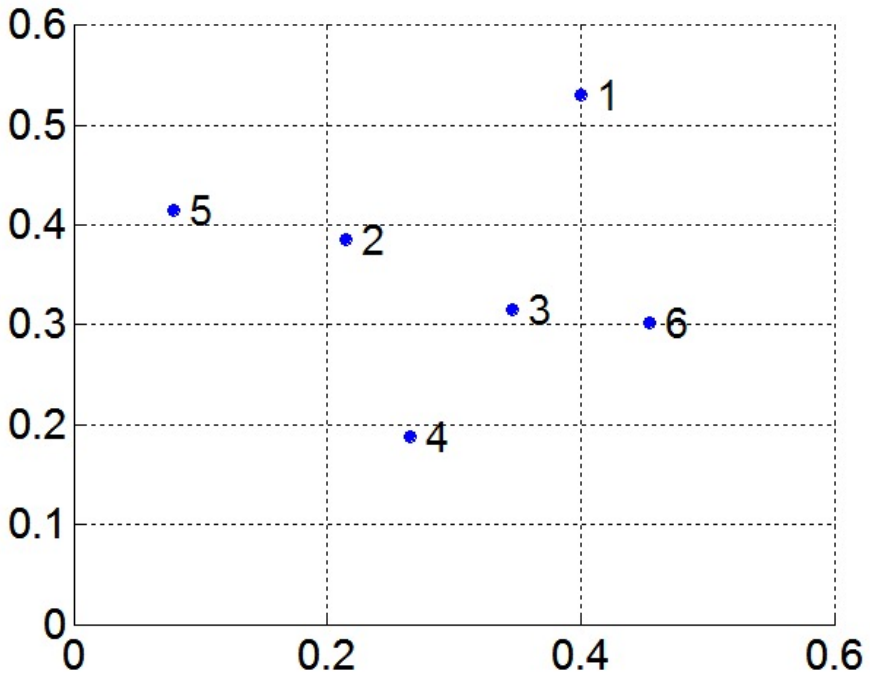
Original Points Two Clusters

- Tends to break large clusters
- Biased towards globular clusters

GROUP AVERAGE

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

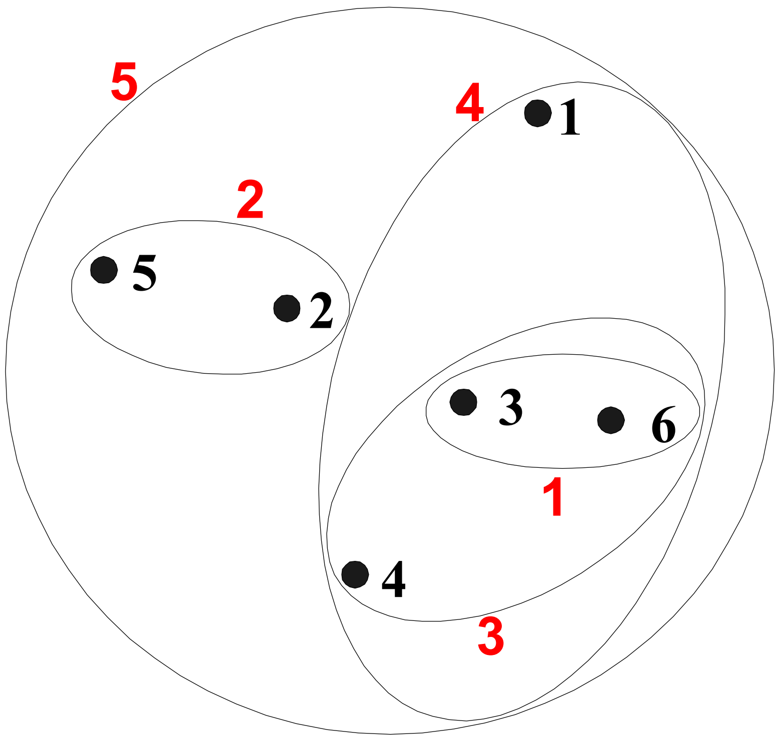
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum\limits_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$



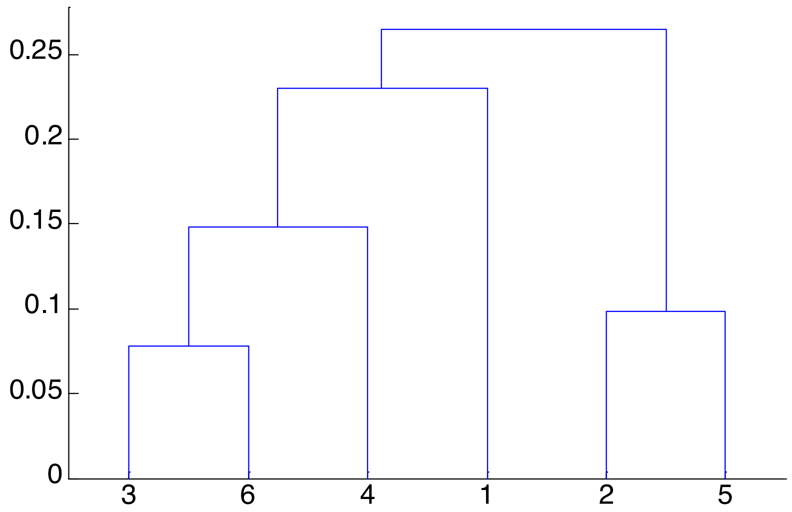
Distance Matrix:

	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

HIERARCHICAL CLUSTERING: GROUP AVERAGE



Nested Clusters



Dendrogram

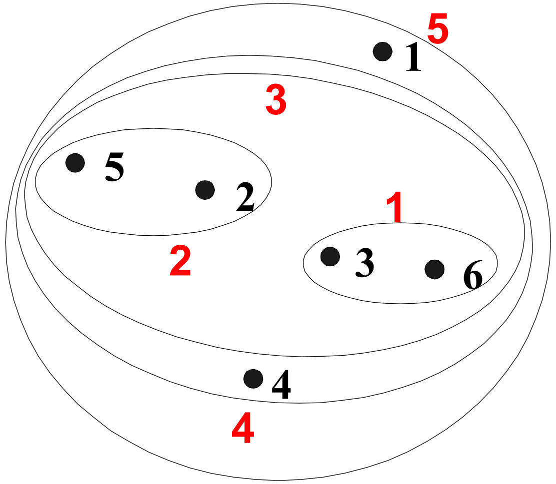
HIERARCHICAL CLUSTERING: GROUP AVERAGE

- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise
- Limitations
  - Biased towards globular clusters

CLUSTER SIMILARITY: WARD'S METHOD

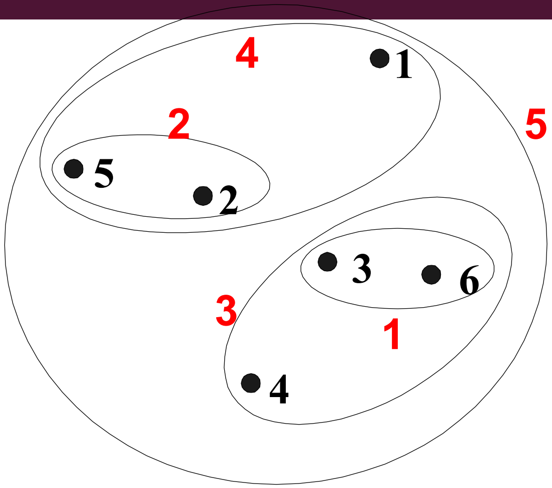
- Similarity of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between points is distance squared
- Less susceptible to noise
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

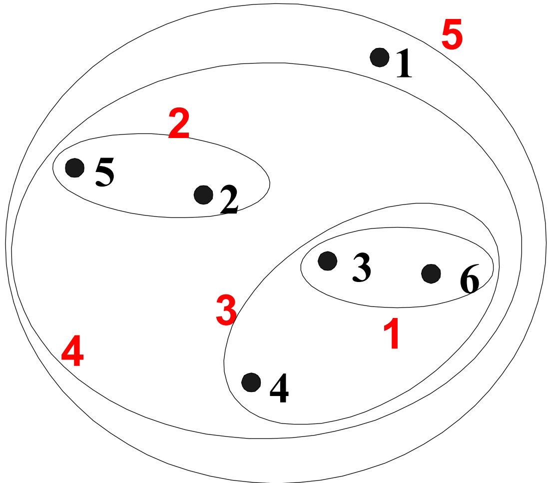
HIERARCHICAL CLUSTERING: COMPARISON



MIN

MAX





Group Average

Ward’s Method