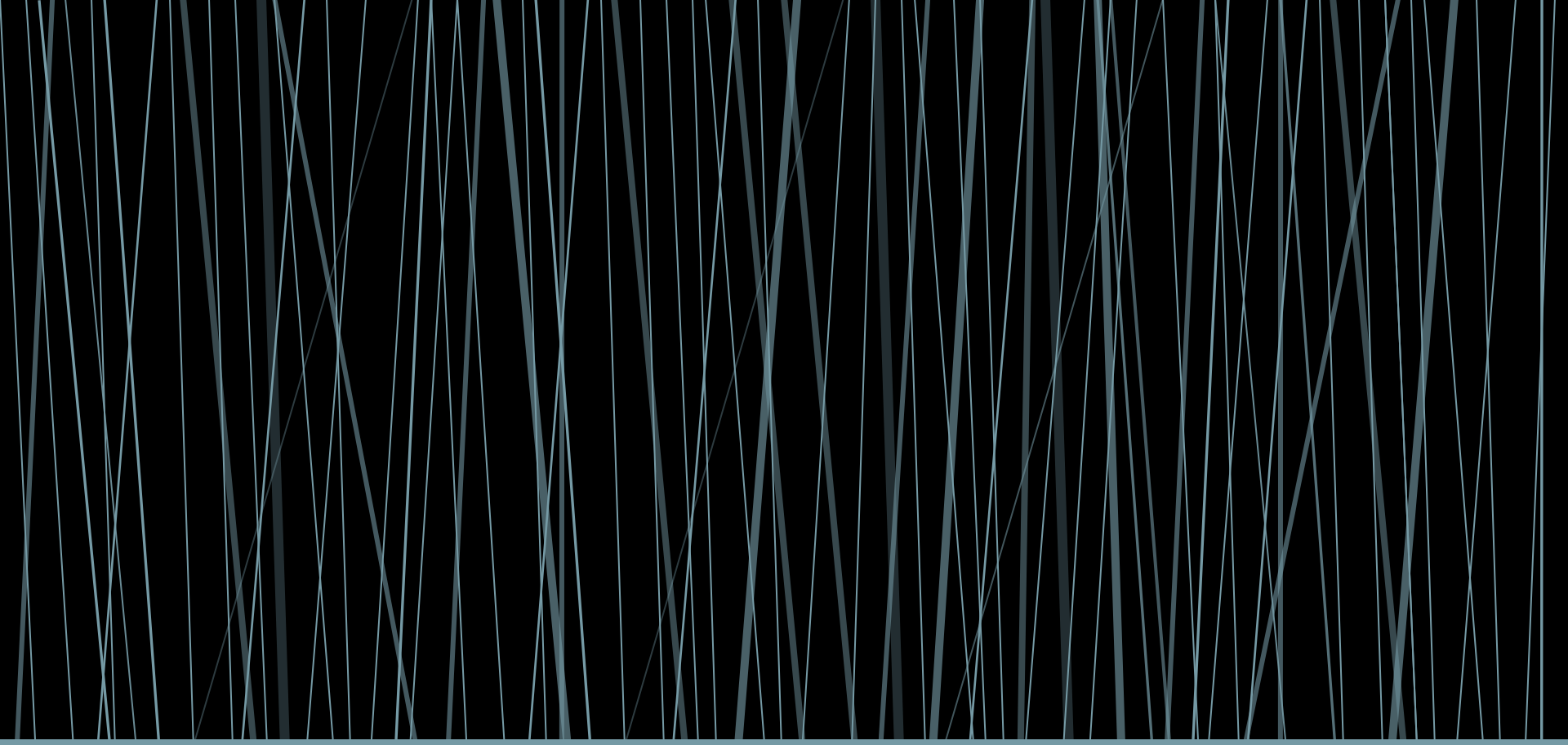
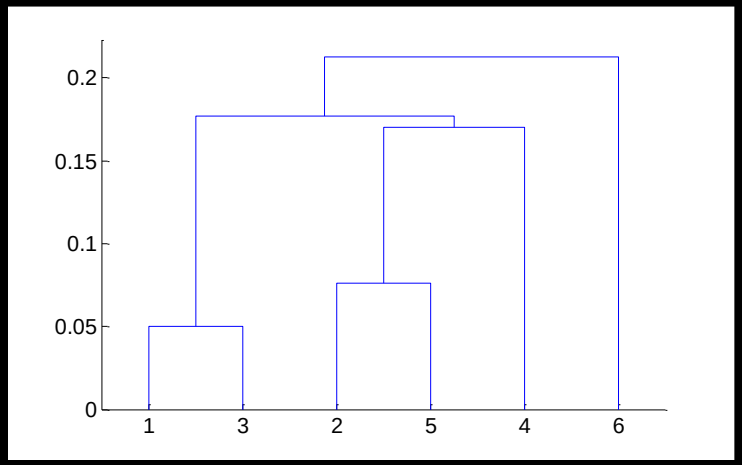
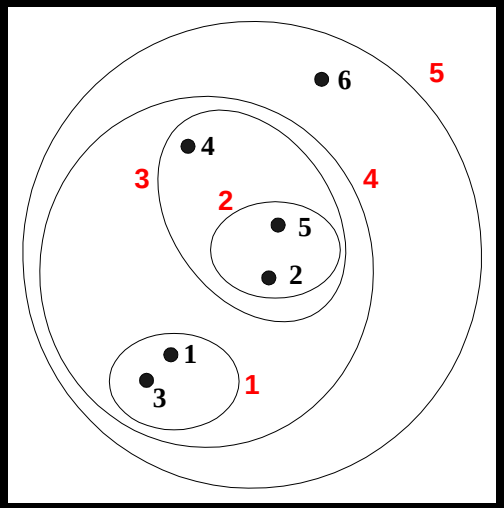


Chapter 5 Hierarchical Clustering

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





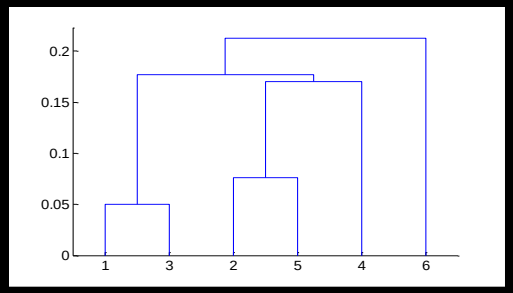
Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level
- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

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

Agglomerative



Divisive

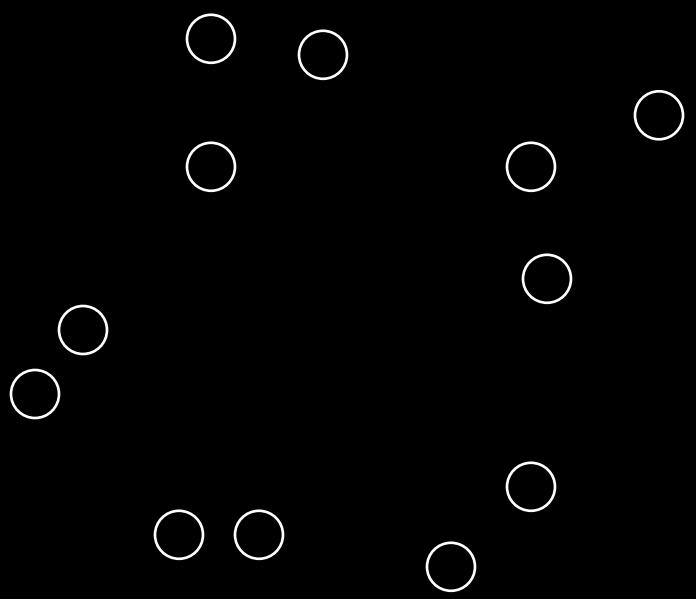
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)

Agglomerative Clustering Algorithm

- Traditional hierarchical algorithms use a similarity or distance matrix (Merge or split one cluster at a time)
- Basic algorithm is straightforward
  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

Starting Situation

- Start with clusters of individual points and a proximity matrix



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

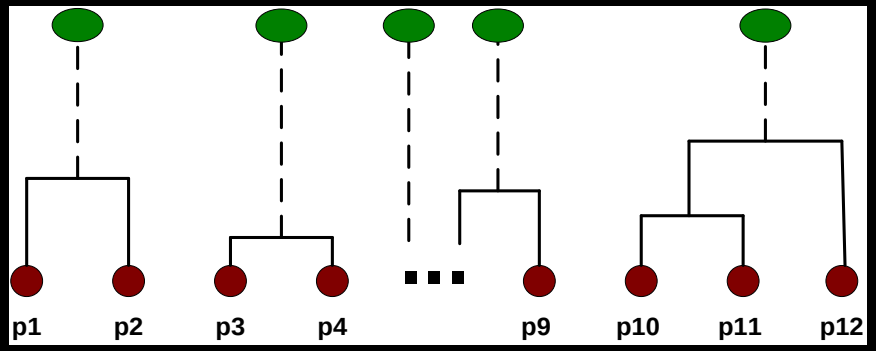
Proximity Matrix

Intermediate Situation

- After some merging steps, we have some clusters

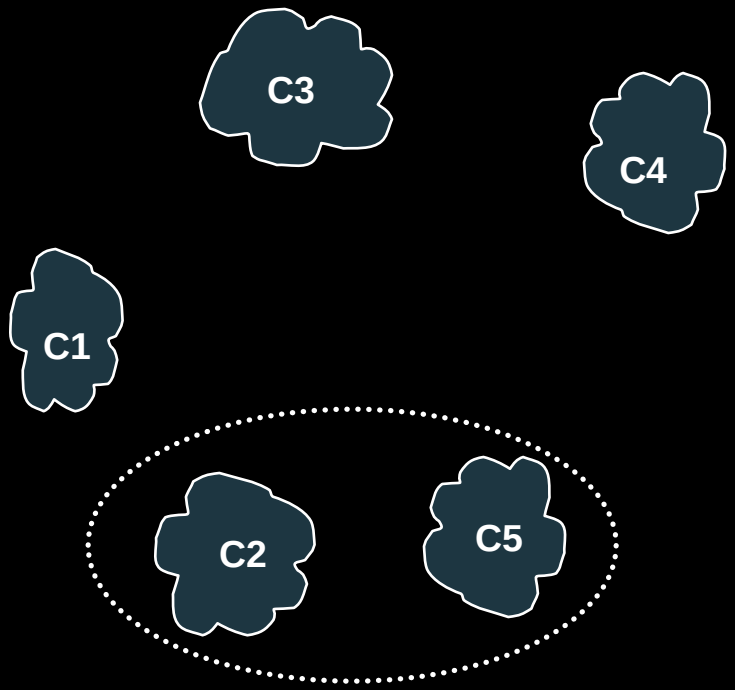
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



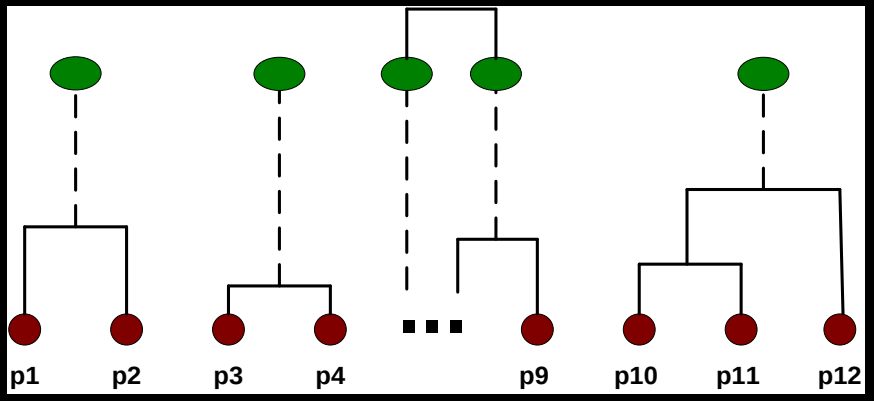
Intermediate Situation

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



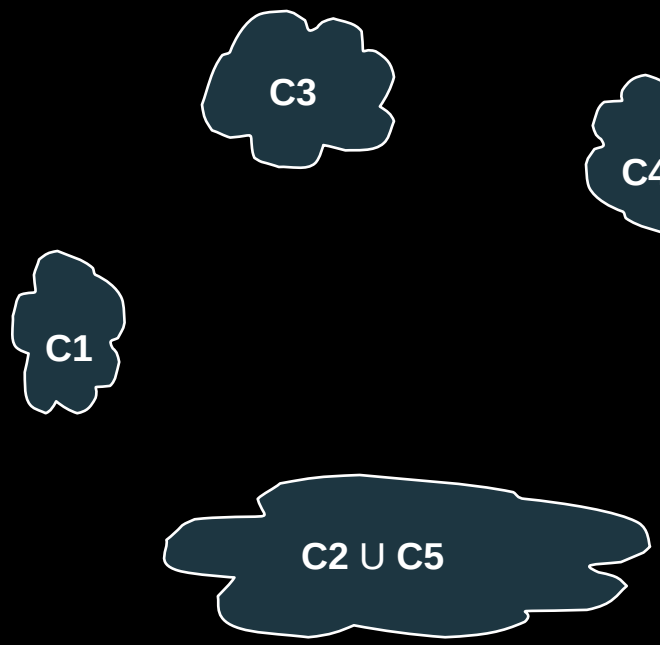
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



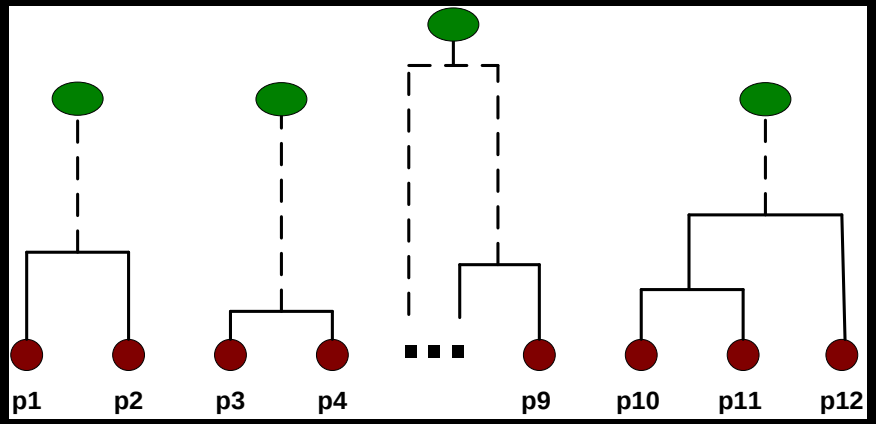
After Merging

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

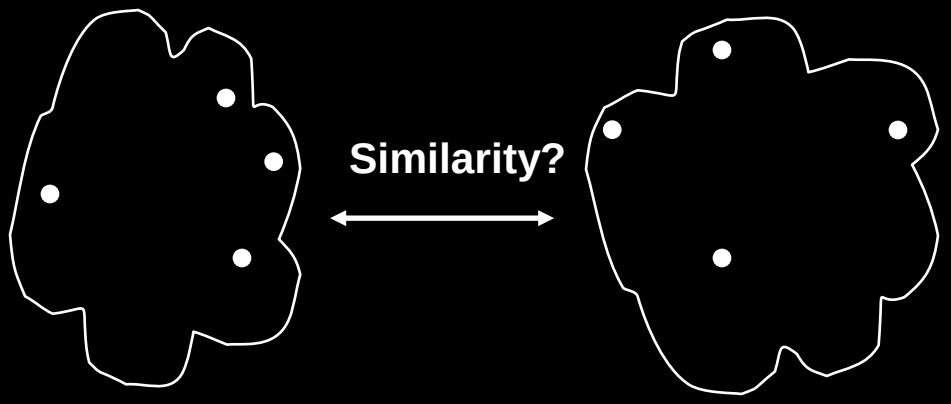


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Similarity

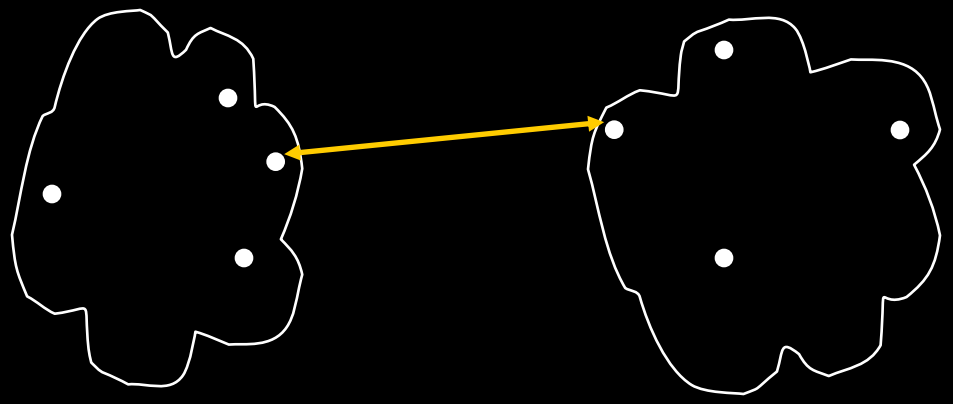


- MIN
- MAX
- Group Average
- Distance Between Centroids

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

Proximity Matrix

How to Define Inter-Cluster Similarity

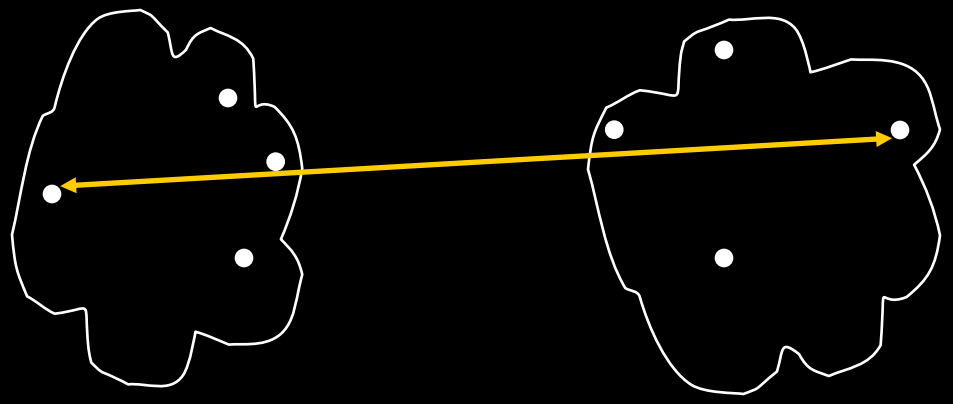


- MIN
- MAX
- Group Average
- Distance Between Centroids

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

Proximity Matrix

How to Define Inter-Cluster Similarity

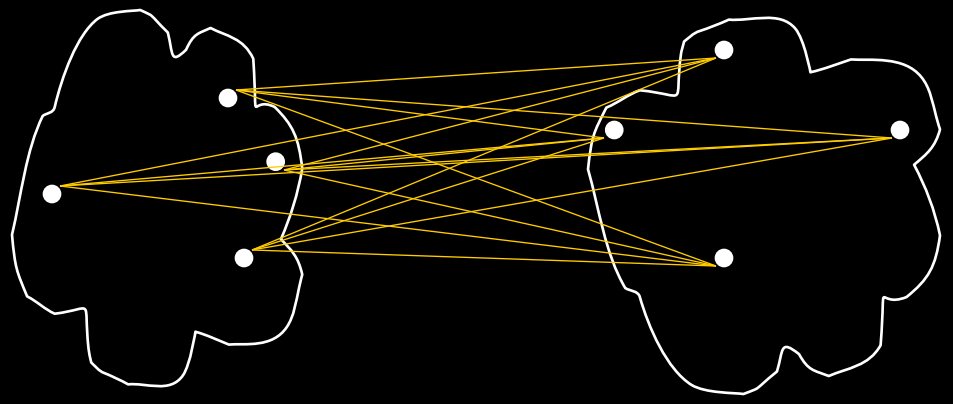


- MIN
- MAX
- Group Average
- Distance Between Centroids

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

Proximity Matrix

How to Define Inter-Cluster Similarity

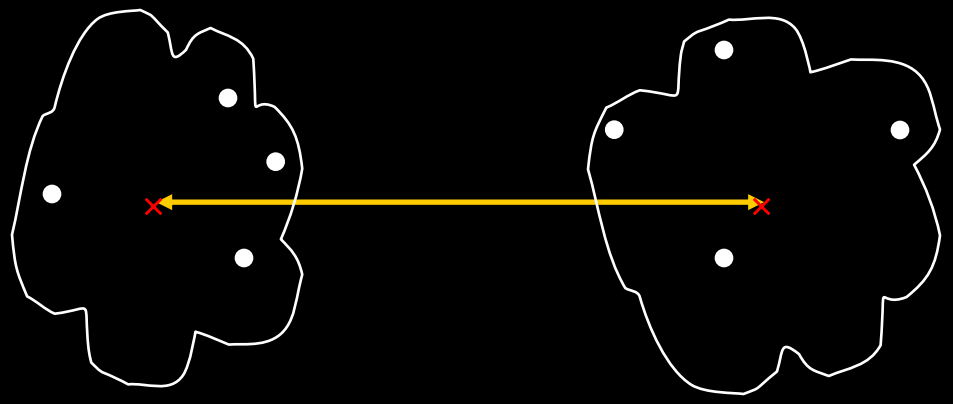


- MIN
- MAX
- Group Average
- Distance Between Centroids

	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

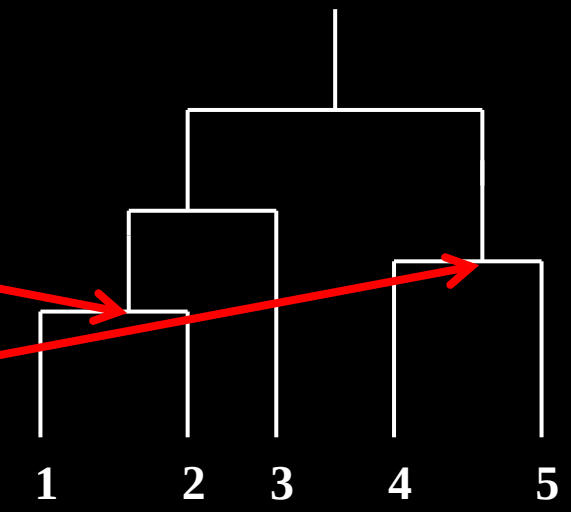
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph.
  - Hierarchical clustering algorithms typically have local objectives

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Cluster Similarity: MIN or Single Link

Example:

https://people.revoledu.com/kardi/tutorial/Clustering/Numerical%20Example.htm

https://www.datavedas.com/hierarchical-clustering/

	D1	D2	D3	D4
D1	0	1.78	3.46	4.97
D2	1.78	0	2.04	3.64
D3	3.46	2.04	0	4.48
D4	4.97	3.64	4.48	0

Cluster Similarity: MIN or Single Link

	D1	D2	D3	D4
D1	0	1.78	3.46	4.97
D2	1.78	0	2.04	3.64
D3	3.46	2.04	0	4.48
D4	4.97	3.64	4.48	0

- Merge D1 and D2 because has the minimum distance value that is 1.78. Then, find the new distance value for ((D1,D2),D3) and ((D1,D2),D4) based on minimum as following:
- New distance value for ((D1,D2),D3) is Min((D1,D3),(D2,D3)) = min (3.46,2.04)=2.04
- New distance value for ((D1,D2),D4) is Min((D1,D4),(D2,D4)) = min (4.97,3.64) = 3.64

Cluster Similarity: MIN or Single Link

	D1,D2	D3	D4
D1,D2	0	2.04	3.64
D3	2.04	0	4.48
D4	3.64	4.48	0

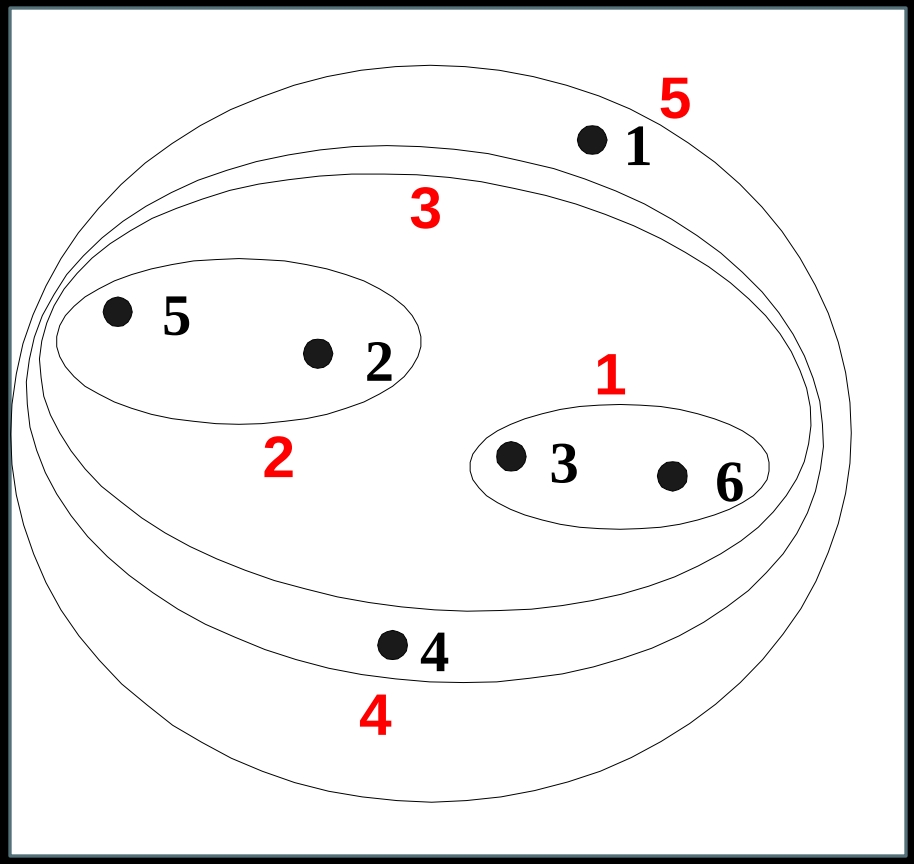
- Merge D1,D2 and D3 because has the minimum distance value that is 2.04. Then, find the new distance value for ((D1,D2,D3),D4) based on minimum as following:
- New distance value for ((D1,D2,D3),D4) is Min((D1,D4),(D2,D4), (D3,D4)) = min(4.97,3.64,4.48) = 3.64

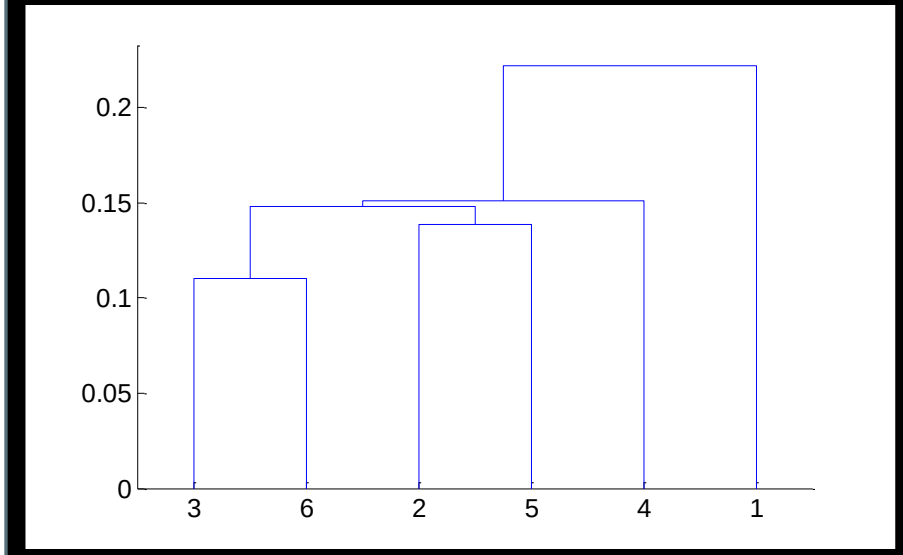
Cluster Similarity: MIN or Single Link

	D1,D2	D3	D4
D1,D2	0	2.04	3.64
D3	2.04	0	4.48
D4	3.64	4.48	0

	D1,D2,D3,D4
D1,D2,D3,D4	0

Hierarchical Clustering: MIN

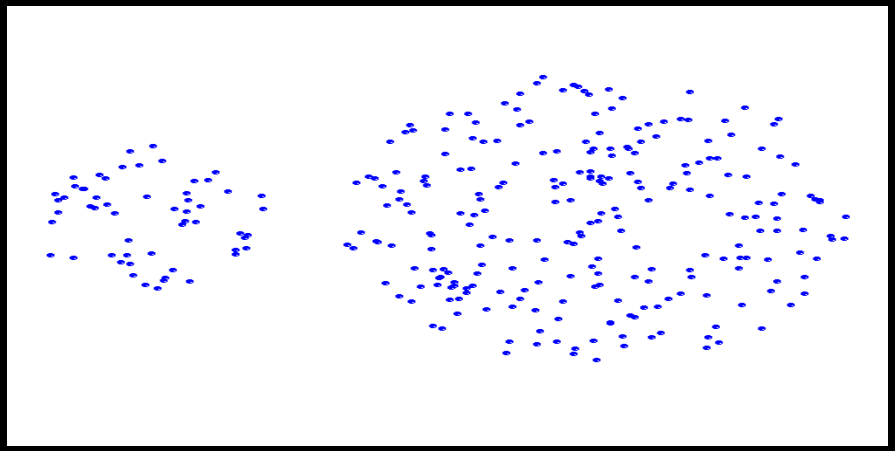




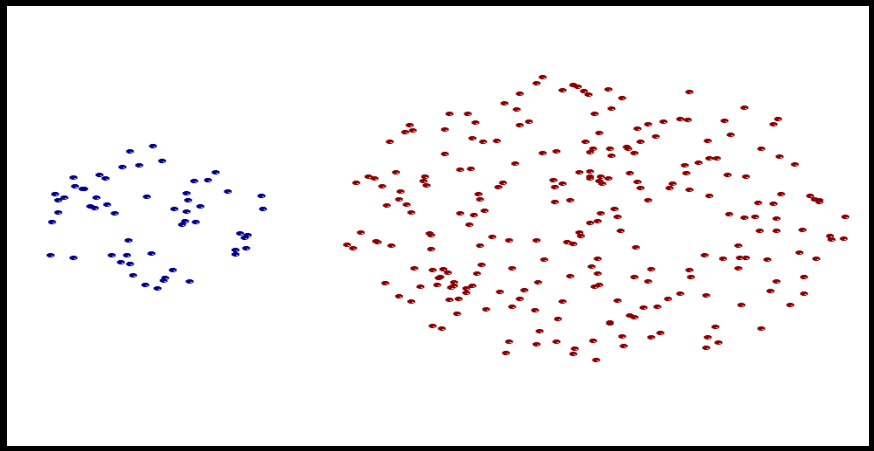
Nested Clusters

Dendrogram

Strength of MIN



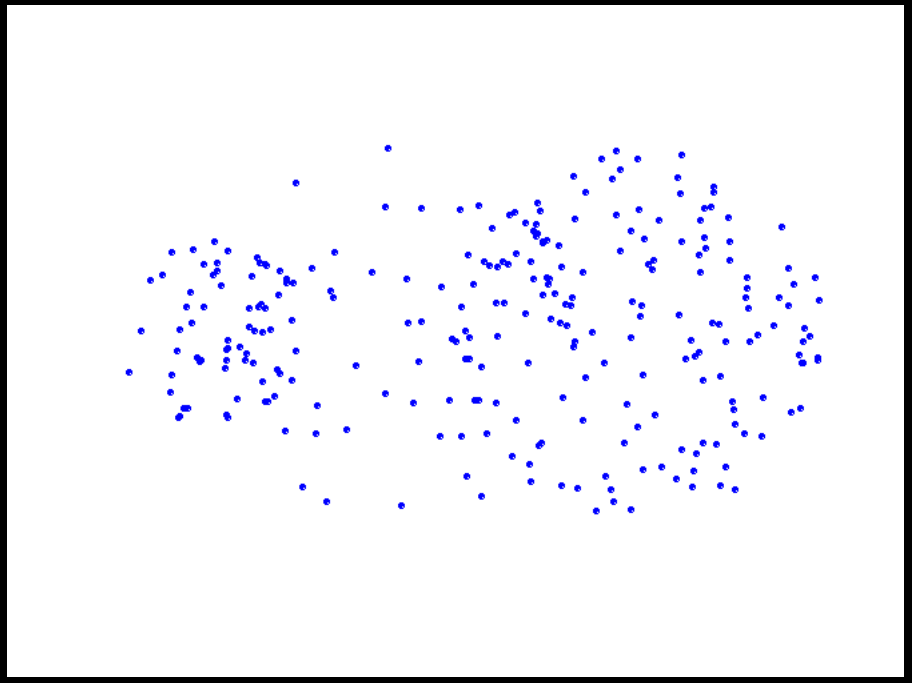
Original Points



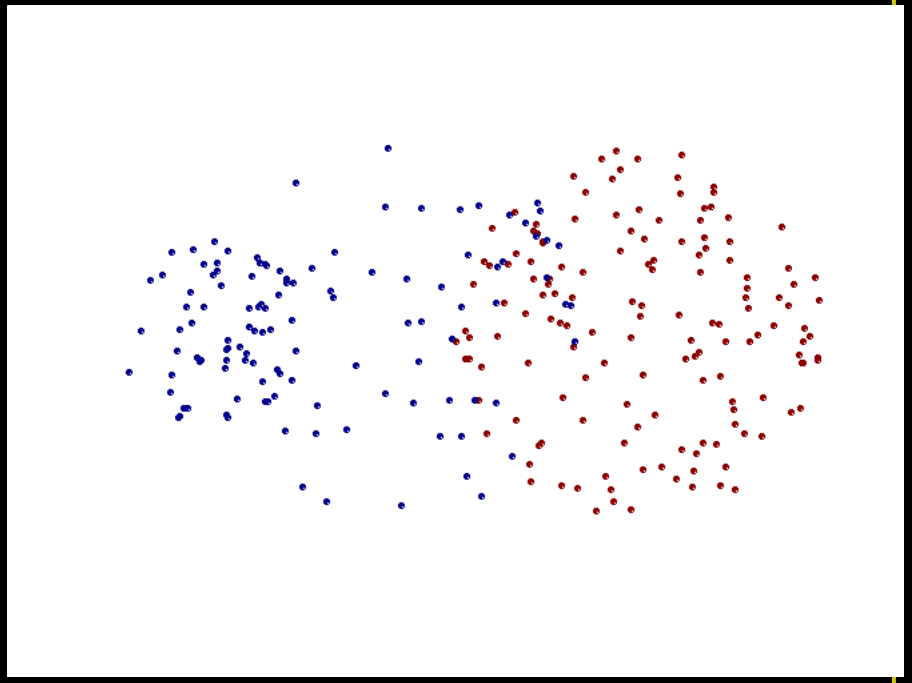
Two Clusters

- Can handle non-elliptical shapes

Limitation of MIN



Original Points



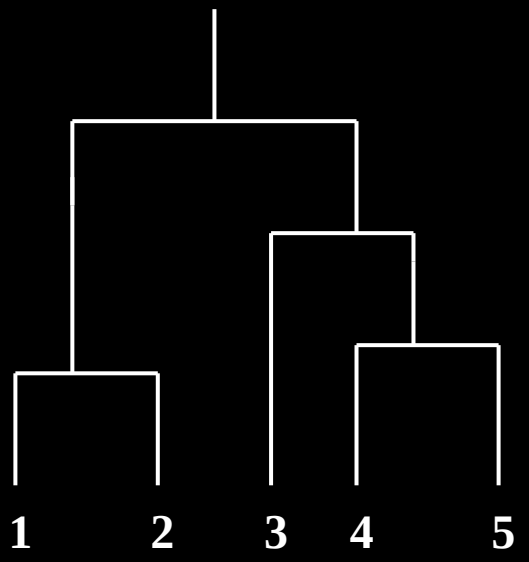
Two Clusters

- Sensitive to noise and outliers

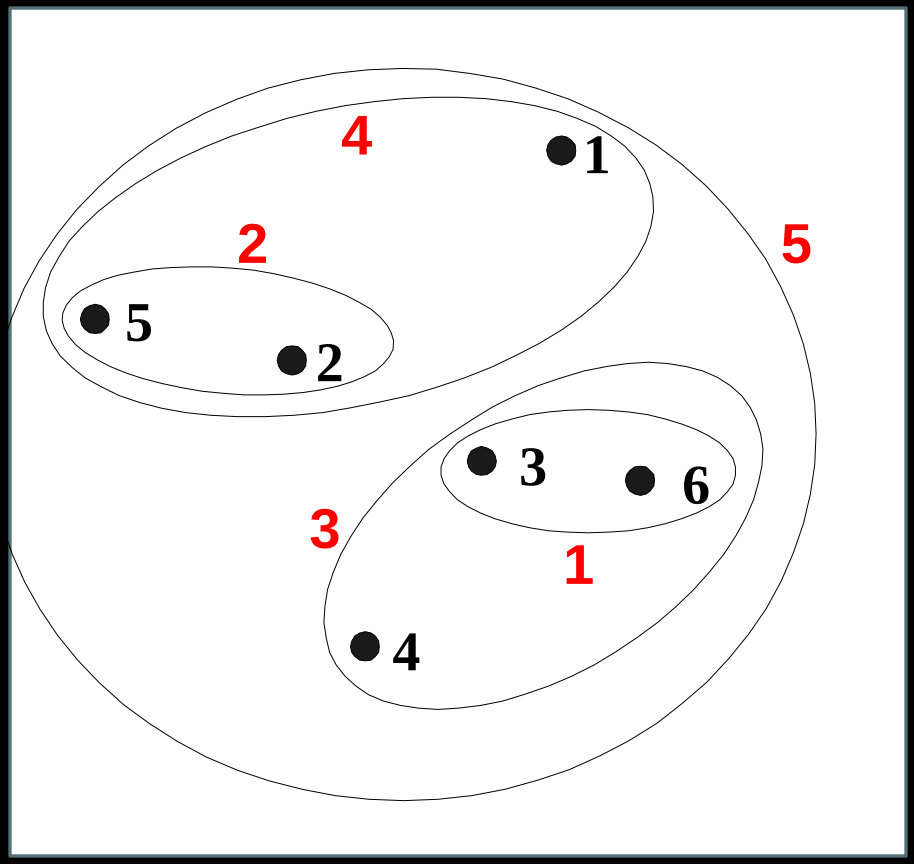
Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters

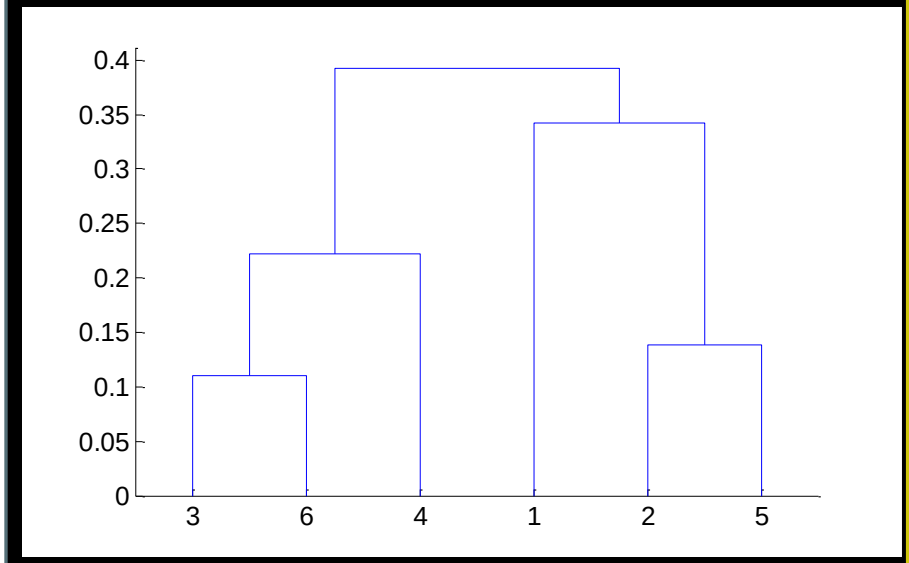
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hierarchical Clustering: MAX

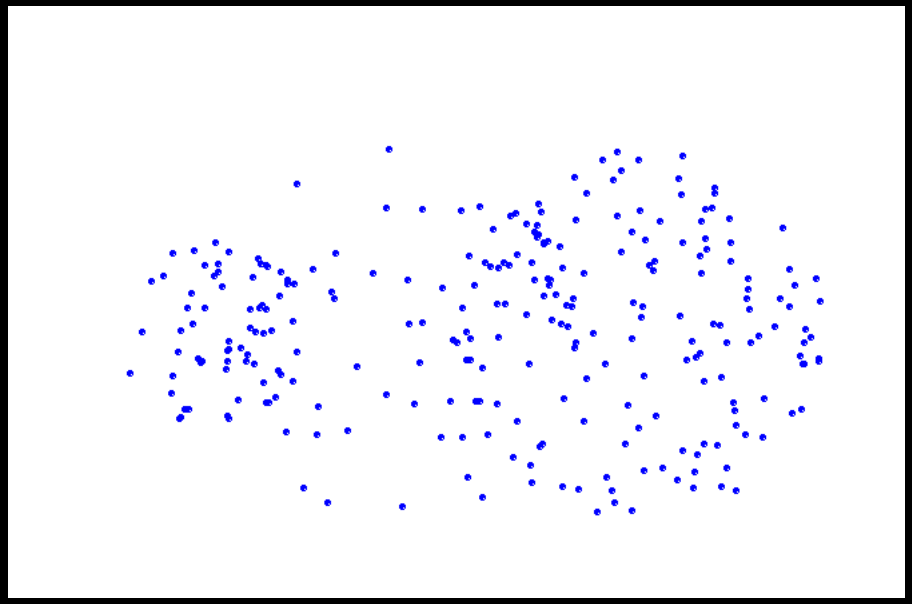


Nested Clusters

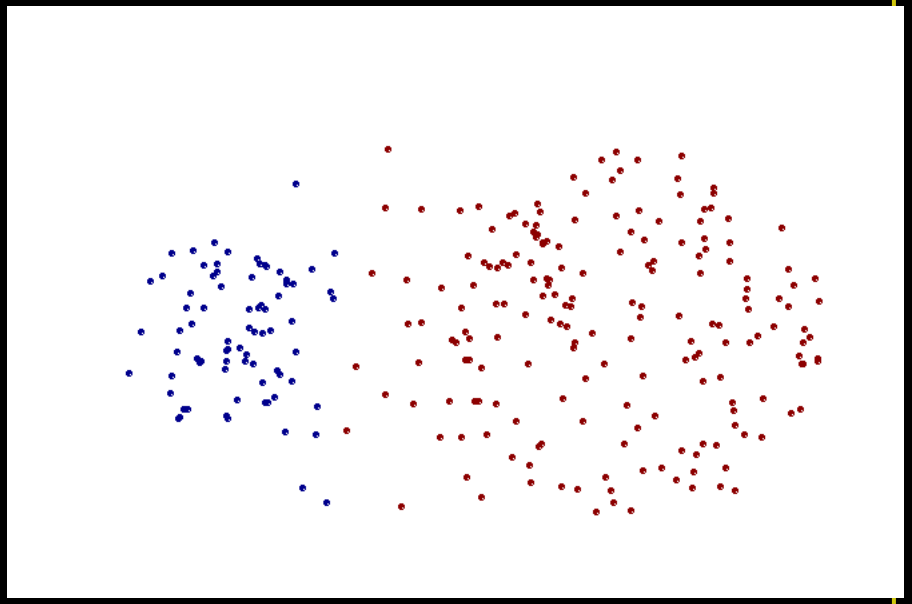


Dendrogram

Strength of MAX



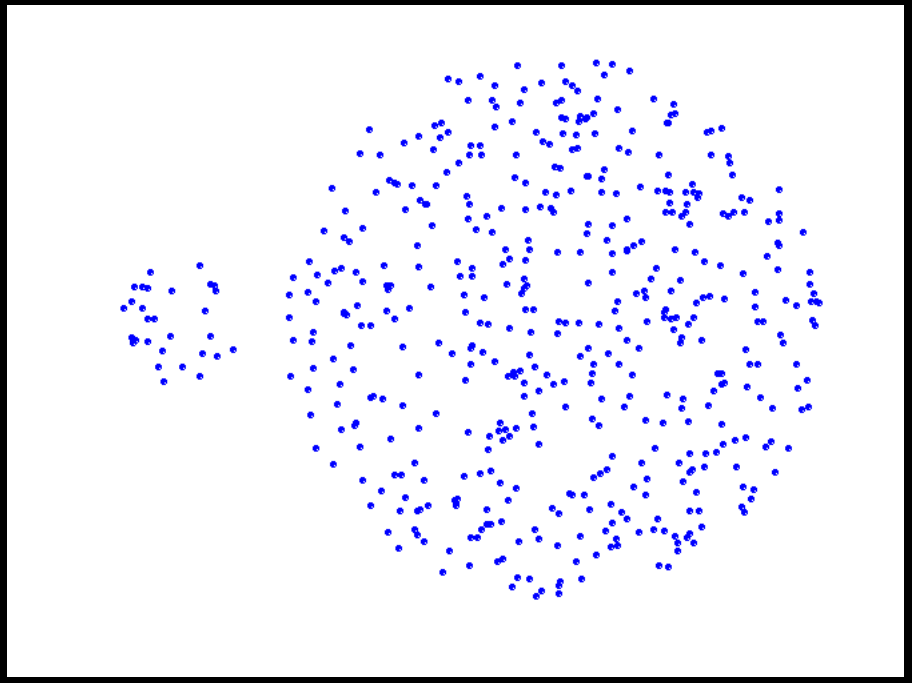
Original Points



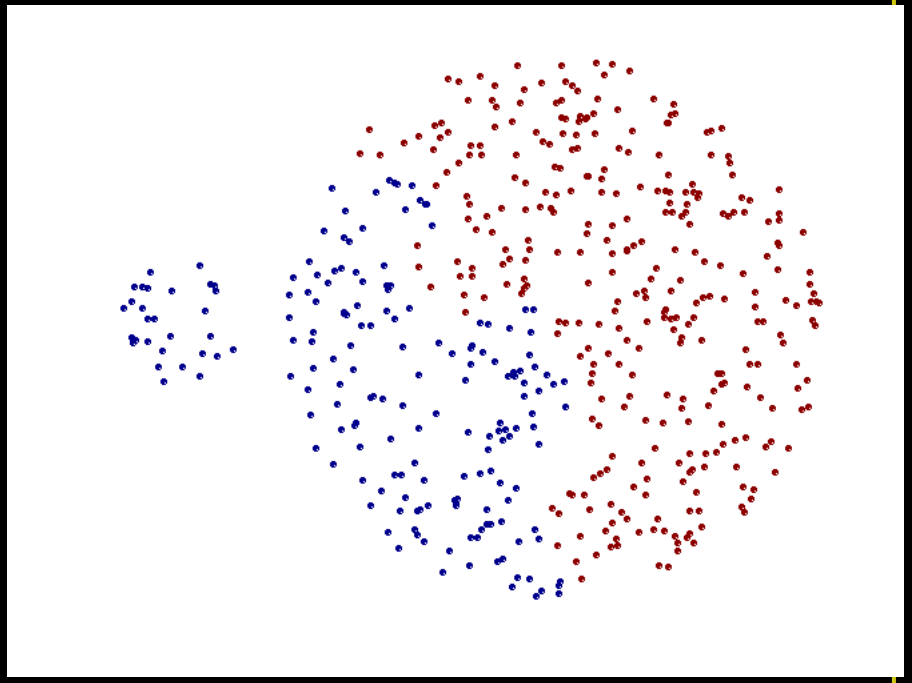
Two Clusters

- Less susceptible to noise and outliers

Limitations of MAX



Original Points



Two Clusters

- Tends to break large clusters
- Biased towards globular clusters

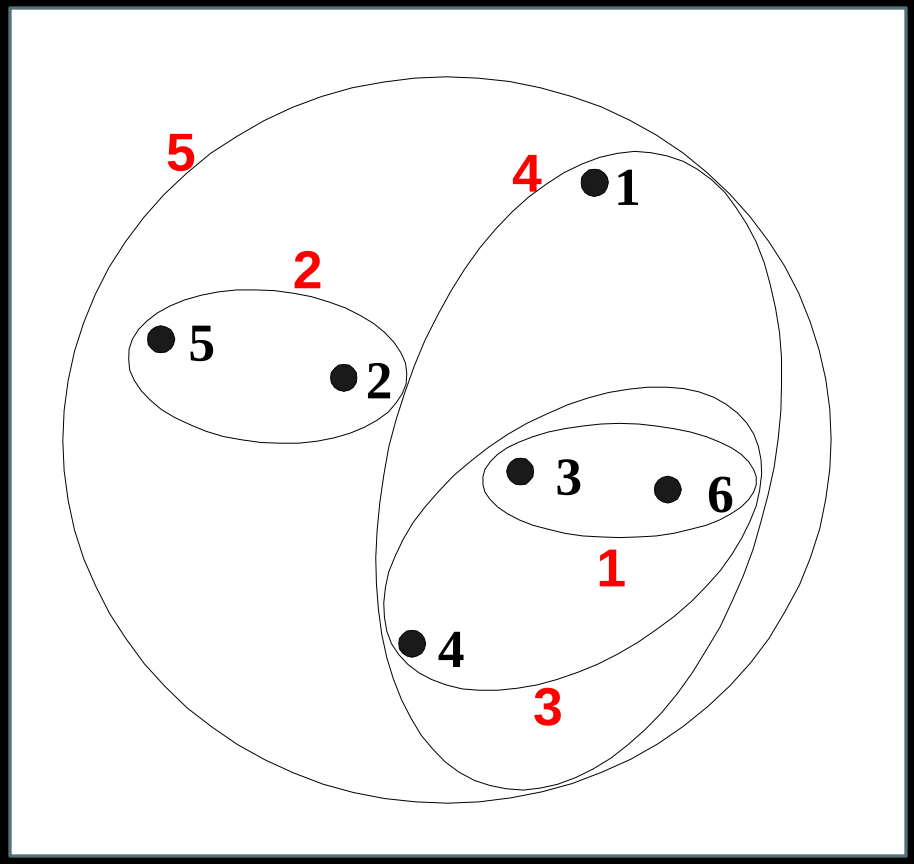
Cluster Similarity: Group Average

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

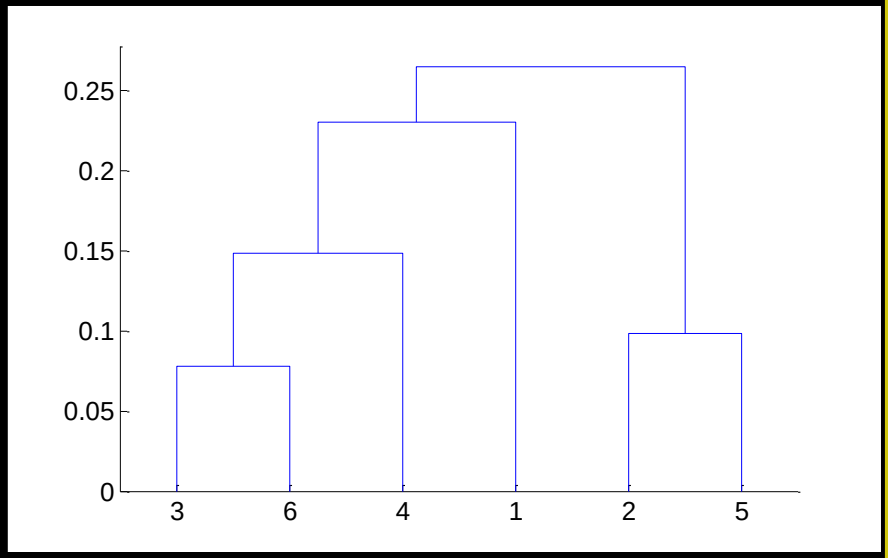
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- Need to use average connectivity for scalability since total proximity favors large clusters

Hierarchical Clustering: Group Ave



Nested Clusters

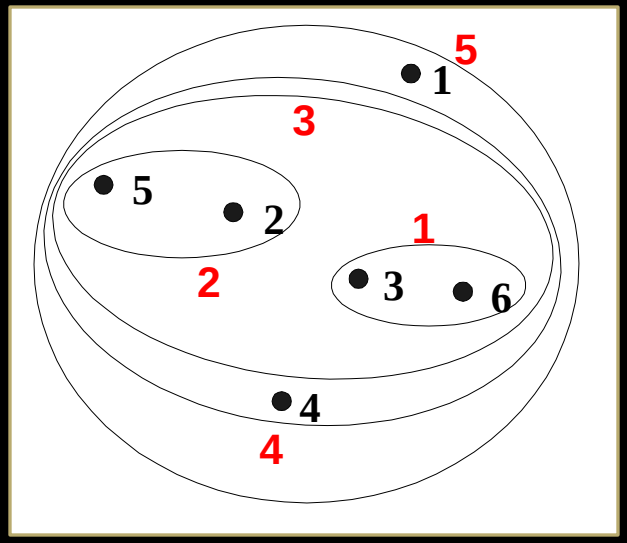


Dendrogram

Hierarchical Clustering: Group Ave

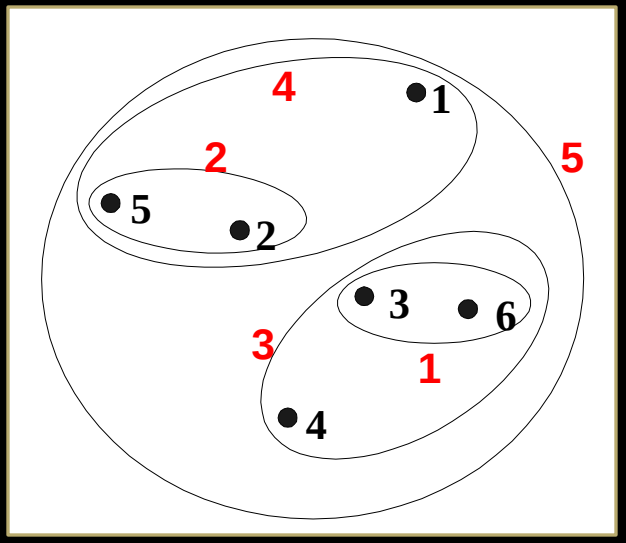
- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise and outliers
- Limitations
  - Biased towards globular clusters

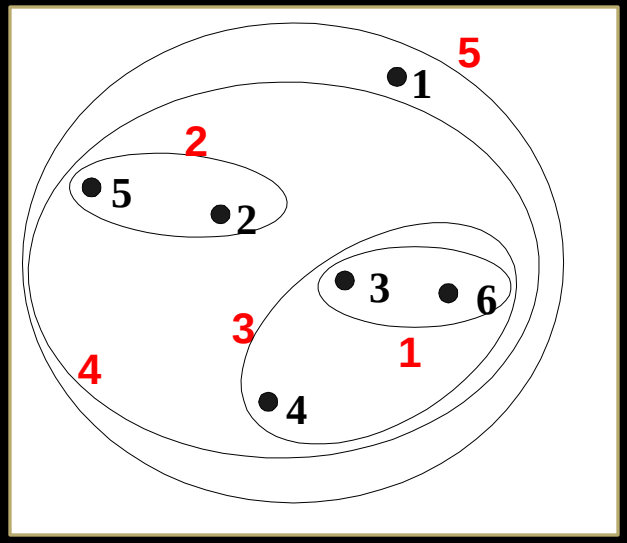
Hierarchical Clustering: Comparison



MIN

MAX





Group Average

Hierarchical Clustering: Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

Cluster Validity

- For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters?
- Purposes of evaluating them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Clusters Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE), Davies-Boulding Index
  - Relative Index: Used to compare two different clustering or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy

External Measures: Clusters Entropy

- Total clusters entropy, ***H(K)*** , is defined as

$$H(K) = \frac{\sum_{k=1}^{K} n_k \cdot H_k}{N}$$

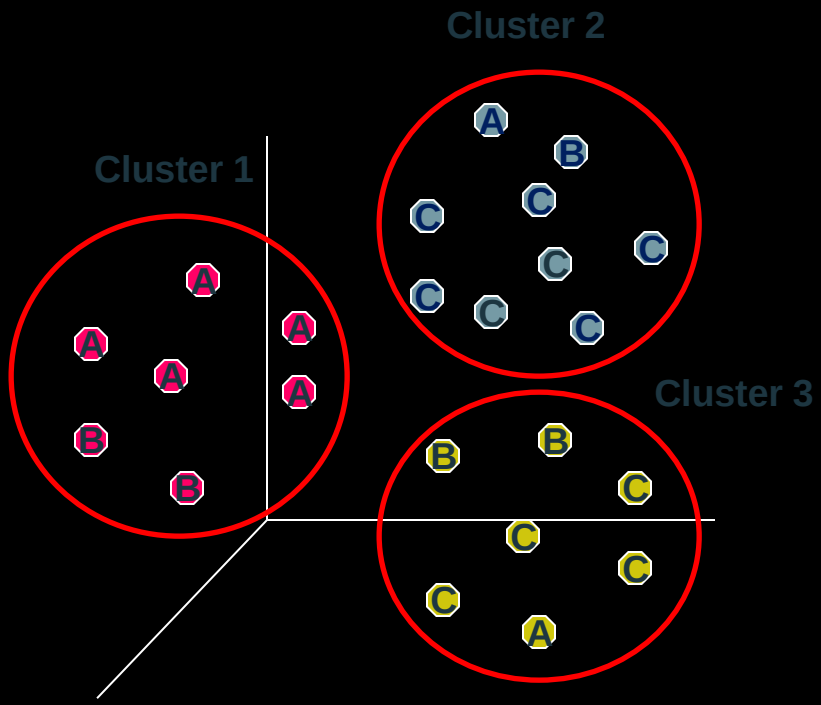
where ***K*** is the number of clusters and n_k is the number of objects in ***k***th cluster, ***N*** is the total number of objects and H_k is the entropy of ***k***th cluster and it is defined as

$$H_k = -\sum_{s=1}^{S} P_{sk} \cdot \log_2(P_{sk})$$

where ***S*** is the number of classes, P_{sk}, is the probability that an object randomly chosen from the ***k***th cluster belongs to the ***s***th class.

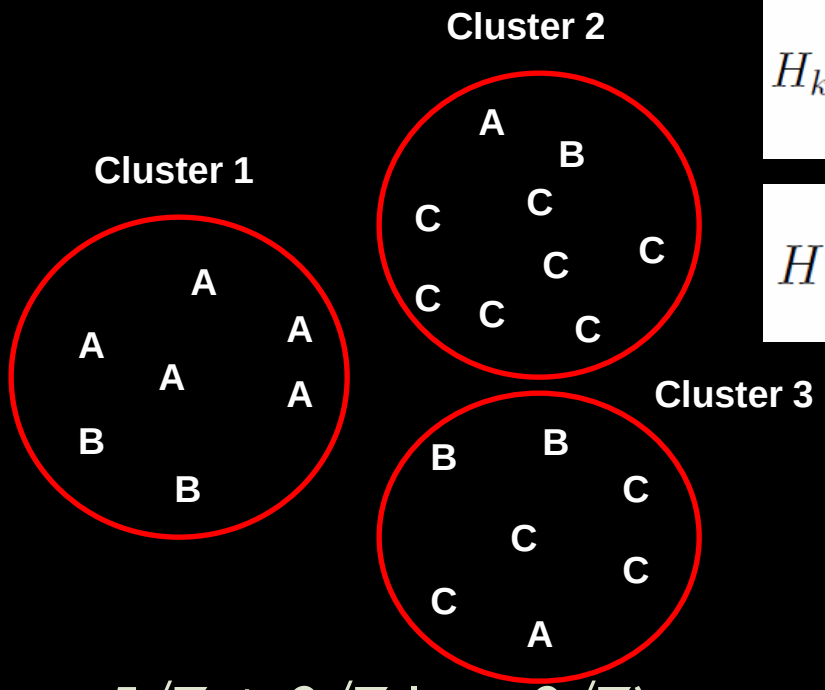
External Measures: Clusters Entropy

- Example:



External Measures: Clusters Entropy

- Example:



$$H_k = -\sum_{s=1}^{S} P_{sk} \cdot \log_2(P_{sk})$$

$$H(K) = \frac{\sum_{k=1}^{K} n_k \cdot H_k}{N}$$

- H1 = - (5/7 $\log_2$ 5/7 + 2/7 $\log_2$ 2/7)
- H2 = - (1/9 $\log_2$ 1/9 + 1/9 $\log_2$ 1/9 + 7/9 $\log_2$ 7/9)
- H3 = -(1/7 $\log_2$ 1/7 + 2/7 $\log_2$ 2/7 + 4/7 $\log_2$ 4/7)
- H(3) = 7/23 (H1) + 9/23 (H2) + 7/23 (H3)

Internal Measures: Davies-Bouldin

- According to Davies-Bouldin validity index (DBI), the best clustering minimizes equation

$$DBI = \frac{1}{C}\sum_{k=1}^{C} max_{l \neq k}\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

where **C** is the number of clusters, **S_c (Q_k)** is the average intra-distance for cluster k

$$CentroidDistance, S_c(Q_k) = \frac{\sum_i \|x_i - c_k\|}{N_k}$$

and **d_{ce}(Q_k, Q_l)** is the inter-distance of cluster k and l

$$CentroidLinkage, d_{ce}(Q_k, Q_l) = \|c_k - c_l\|$$

and $c_k = \frac{1}{N_k}\sum_{x_i \in Q_k} x_i$ is the centroid of cluster k

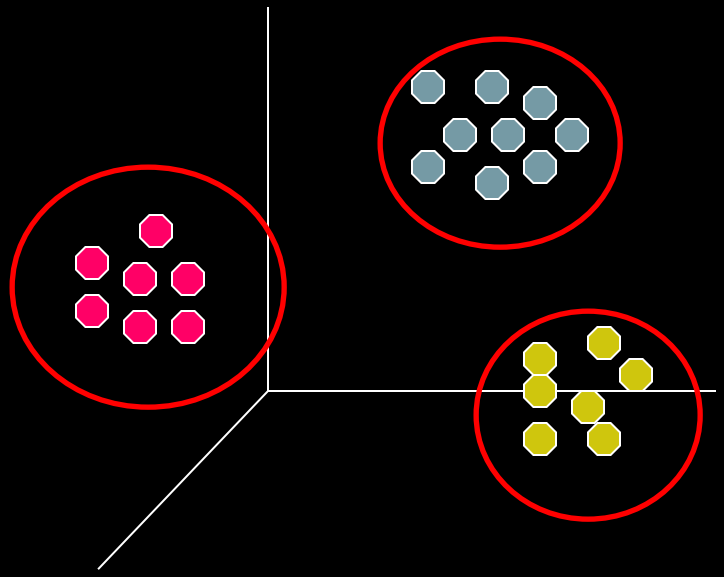
Internal Measures: Davies-Bouldin

- According to Davies-Bouldin validity index (DBI), the best clustering minimizes equation

$$DBI = \frac{1}{C}\sum_{k=1}^{C} max_{l \neq k}\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

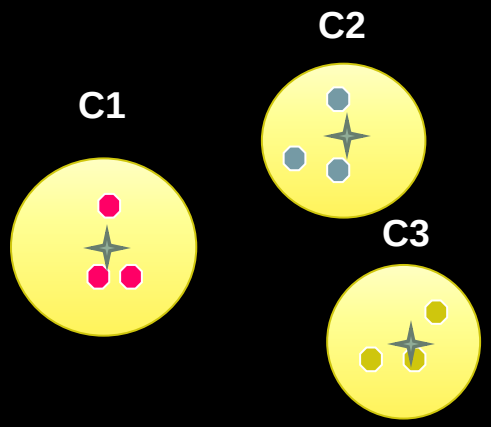
← minimize INTRA-cluster

← maximize INTER-cluster



Centroid of Cluster: $c_k = \frac{1}{N_k} \sum_{x_i \in Q_k} x_i$

- Example:

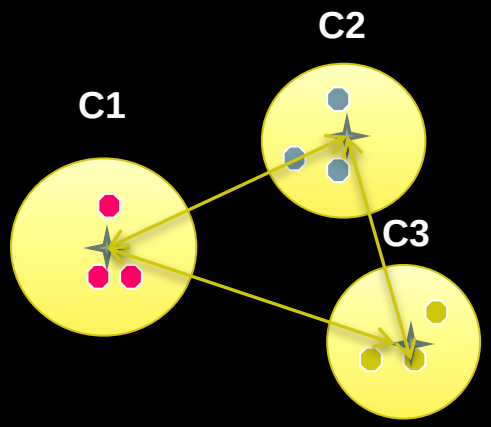


	x	y
Cluster 1	2.0	3.0
	2.5	2.5
	1.0	3.0
Cluster 2	3.0	1.0
	4.0	0.5
	3.5	0.5
Cluster 3	5.0	6.0
	4.0	6.5
	5.5	5.5

C1 = { (2.0 + 2.5 + 1.0)/3, (3.0 + 2.5 + 3.0)/3} = { 1.83, 2.83 }
C2 = { (3.0 + 4.0 + 3.5)/3, (1.0 + 0.5 + 0.5)/3} = { 3.50, 0.67 }
C3 = { (5.0 + 4.0 + 5.5)/3, (6.0 + 6.5 + 5.5)/3} = { 4.83, 6.00 }

Inter distance between Centroids

- Example:



	x	y
Cluster 1	2.0	3.0
Center c1 = { 1.83, 2.83 }	2.5	2.5
	1.0	3.0
Cluster 2	3.0	1.0
Center c2 = { 3.50, 0.67 }	4.0	0.5
	3.5	0.5
Cluster 3	5.0	6.0
Center c3 = { 4.83, 6.00 }	4.0	6.5
	5.5	5.5

$$CentroidLinkage, d_{ce}(Q_k, Q_l) = \|c_k - c_l\|$$

$d_{ce}(Q1,Q2) = \text{sqrt}((1.83 - 3.50)^2 + (2.83 - 0.67)^2) = 2.73$

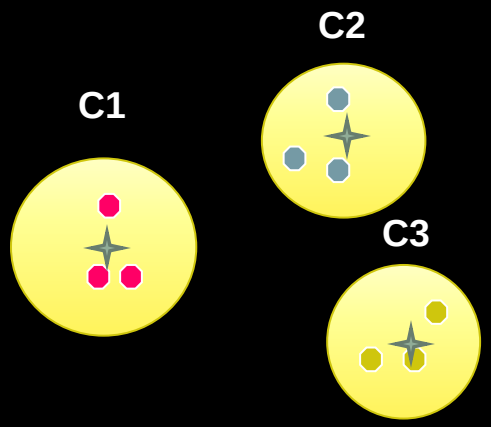
$d_{ce}(Q1,Q3) = \text{sqrt}((1.83 - 4.83)^2 + (2.83 - 6.00)^2) = 4.36$

$d_{ce}(Q2,Q3) = \text{sqrt}((3.50 - 4.83)^2 + (0.67 - 6.00)^2) = 5.49$

Intra distance

$$CentroidDistance, S_c(Q_k) = \frac{\sum_i \|x_i - c_k\|}{N_k}$$

- Example:



	x	y
Cluster 1	2.0	3.0
Center c1 = { 1.83, 2.83 }	2.5	2.5
	1.0	3.0
Cluster 2	3.0	1.0
Center c2 = { 3.50, 0.67 }	4.0	0.5
	3.5	0.5
Cluster 3	5.0	6.0
Center c3 = { 4.83, 6.00 }	4.0	6.5
	5.5	5.5

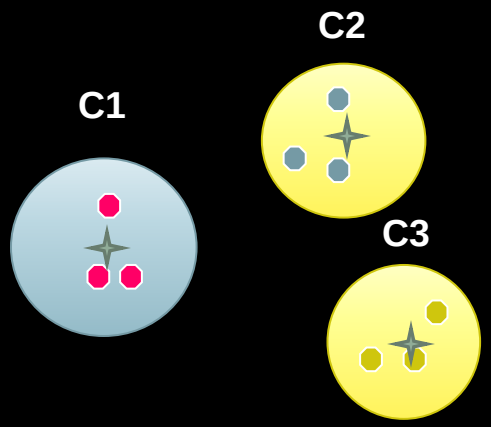
Sc(Q1) = (sqrt((2.0–1.83)^2+(3.0–2.83)^2) + sqrt((2.5–1.83)^2+(2.5–2.83)^2) + sqrt((1.0–1.83)^2+(3.0–2.83)^2))/3
Sc(Q1) = (0.24 + 0.74 + 0.84)/3= 0.61

Sc(Q2) = (sqrt((3.0–3.5)^2+ (1.0–0.67)^2) + sqrt((4.0–3.5)^2+(0.5–0.67)^2) + sqrt ((3.5–3.5)^2+(0.5–0.67)^2))/3
Sc(Q2) = (0.60 + 0.53 + 0.17)/3 = 0.43

Sc(Q3) = (sqrt((5.0–4.83)^2+ (6.0–6.0)^2) + sqrt((4.0–4.83)^2+(6.5–6.0)^2) + sqrt ((5.5–4.83)^2+(5.5–6.0)^2))/3
Sc(Q3) = (0.17 + 0.97 + 0.84)/3 = 0.66

Internal Measures: Davies-Bouldin

- Example:



$$DBI = \frac{1}{C}\sum_{k=1}^{C} max_{l \neq k}\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

For C = 1, for each cluster C ≠ 1, find

$$\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

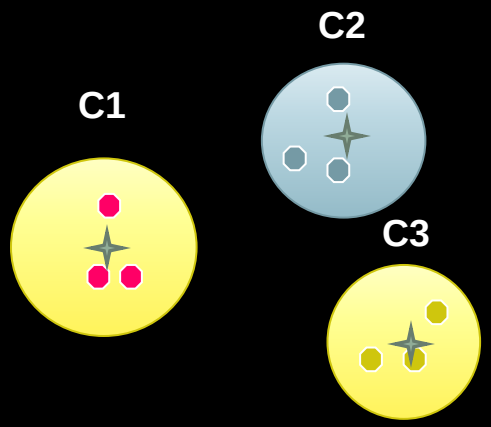
C1 & C2 = $\frac{Sc(Q1) + Sc(Q2)}{d_{ce}(Q1,Q2)} = \frac{(0.61 + 0.43)}{2.73} = 0.38$

C1 & C3 = $\frac{Sc(Q1) + Sc(Q3)}{d_{ce}(Q1,Q3)} = \frac{(0.61 + 0.66)}{4.36} = 0.29$

Take C1DBI =MAX (C1 &C2, C1 &c3) = MAX(0.38, 0.29) = 0.38

Internal Measures: Davies-Bouldin

- Example:



$$DBI = \frac{1}{C}\sum_{k=1}^{C} max_{l \neq k}\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

For C = 2, for each cluster C ≠ 2, find

$$\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

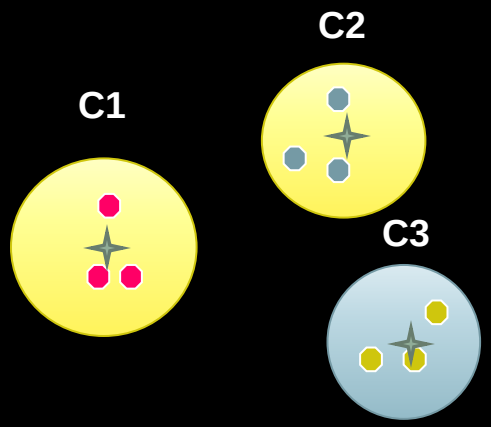
C2 & C1 = $\frac{Sc(Q2) + Sc(Q1)}{d_{ce}(Q2,Q1)} = \frac{(0.43 + 0.61)}{2.73} = 0.38$

C2 & C3 = $\frac{Sc(Q2) + Sc(Q3)}{d_{ce}(Q2,Q3)} = \frac{(0.43 + 0.66)}{5.49} = 0.20$

Take C2DBI =MAX (C2 &C1, C2 &c3) = MAX(0.38, 0.20) = 0.38

Internal Measures: Davies-Bouldin

- Example:



$$DBI = \frac{1}{C}\sum_{k=1}^{C} max_{l \neq k}\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

For C = 3, for each cluster C ≠ 3, find

$$\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

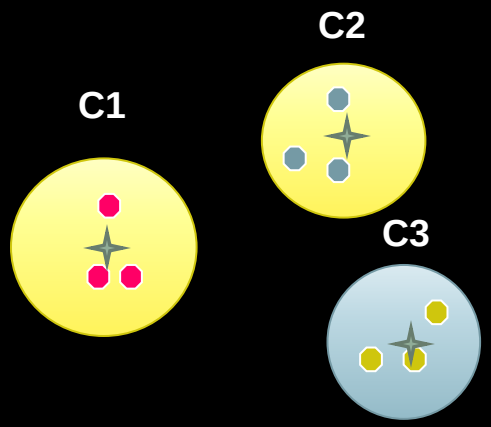
C3 & C1 = $\frac{Sc(Q3) + Sc(Q1)}{d_{ce}(Q3,Q1)} = \frac{(0.66 + 0.61)}{4.36} = 0.29$

C3 & C2 = $\frac{Sc(Q3) + Sc(Q2)}{d_{ce}(Q3,Q2)} = \frac{(0.66 + 0.43)}{5.49} = 0.20$

Take C3DBI =MAX (C3 &C1, C3 &c2) = MAX(0.29, 0.20) = 0.29

Internal Measures: Davies-Bouldin

- Example:



$$DBI = \frac{1}{C}\sum_{k=1}^{C} max_{l \neq k}\left\{\frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)}\right\}$$

$$DBI = \frac{C1DBI + C2DBI + C3DBI}{3}$$

$$DBI = \frac{0.38 + 0.38 + 0.29}{3} = 0.35$$

Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

The End