Distance / Proximity Metrics

Manhattan

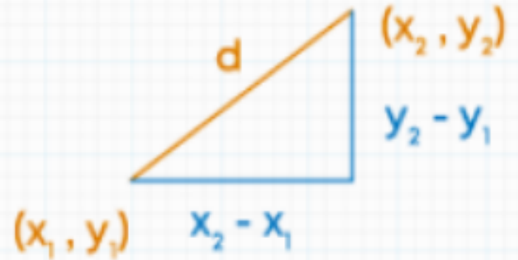
Euclidean

Manhattan distance = $\sum_{i=1}^{n} |x_i - y_i|$

For 2 dimensional space,

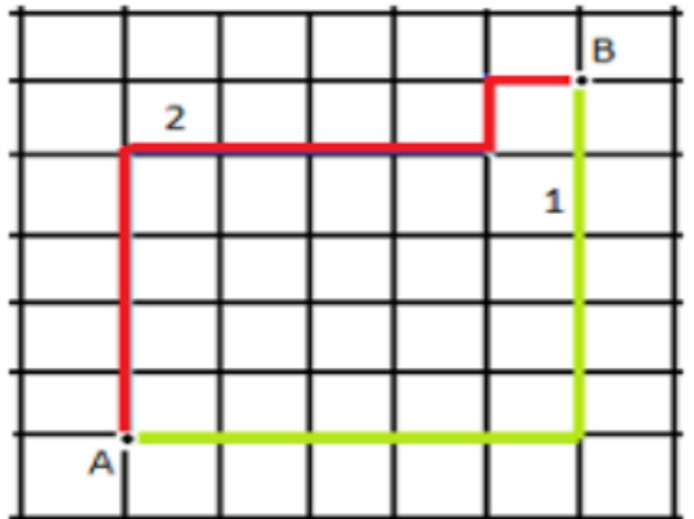
$M_{dist} = |x_2 - x_1| + |y_2 - y_1|$

Euclidean distance (d) = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$



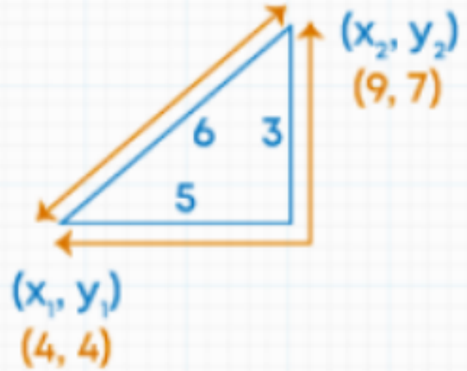
In 3 dimensional space for points (x_1, y_1, z_1) and (x_2, y_2, z_2)

Euclidean distance (d) = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$



→ City Block

→ Can be used to find Hamming distance



Euclidean distance

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

$= \sqrt{(9 - 4)^2 + (7 - 4)^2}$

$= \sqrt{5^2 + 3^2}$

$= \sqrt{25 + 9}$

$= \sqrt{34}$

$= 5.83$

Manhattan distance

$= |x_2 - x_1| + |y_2 - y_1|$

$= |9 - 4| + |7 - 4|$

$= 5 + 3$

$= 8$

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

Distances, such as the Euclidean distance, have some well known properties.

1. **Positivity**
   (a) $d(\mathbf{x}, \mathbf{x}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$,
   (b) $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$.
2. **Symmetry**
   $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$.
3. **Triangle Inequality**
   $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.

where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

A distance that satisfies these properties is a metric

Ques Is the following measure, a metric?

$$d(t_1, t_2) = \begin{cases} t_2 - t_1 & \text{if } t_1 \leq t_2 \\ 24 + (t_2 - t_1) & \text{if } t_1 > t_2 \end{cases}$$

- Distances, such as the Euclidean distance, have some well known properties.

1. **Positivity**
   (a) $d(\mathbf{x}, \mathbf{x}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$,
   (b) $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$.
2. **Symmetry**
   $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$.
3. **Triangle Inequality**
   $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.

where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

- A distance that satisfies these properties is a metric

$d(1Pm, 2Pm) = 15$
$d(2PM, 1PM) = 23$

Ques Is the following measure, a metric?

$$d(t_1, t_2) = \begin{cases} t_2 - t_1 & \text{if } t_1 \leq t_2 \\ 24 + (t_2 - t_1) & \text{if } t_1 > t_2 \end{cases}$$

NOT a metric

Eudidean Measure
↳ Metric

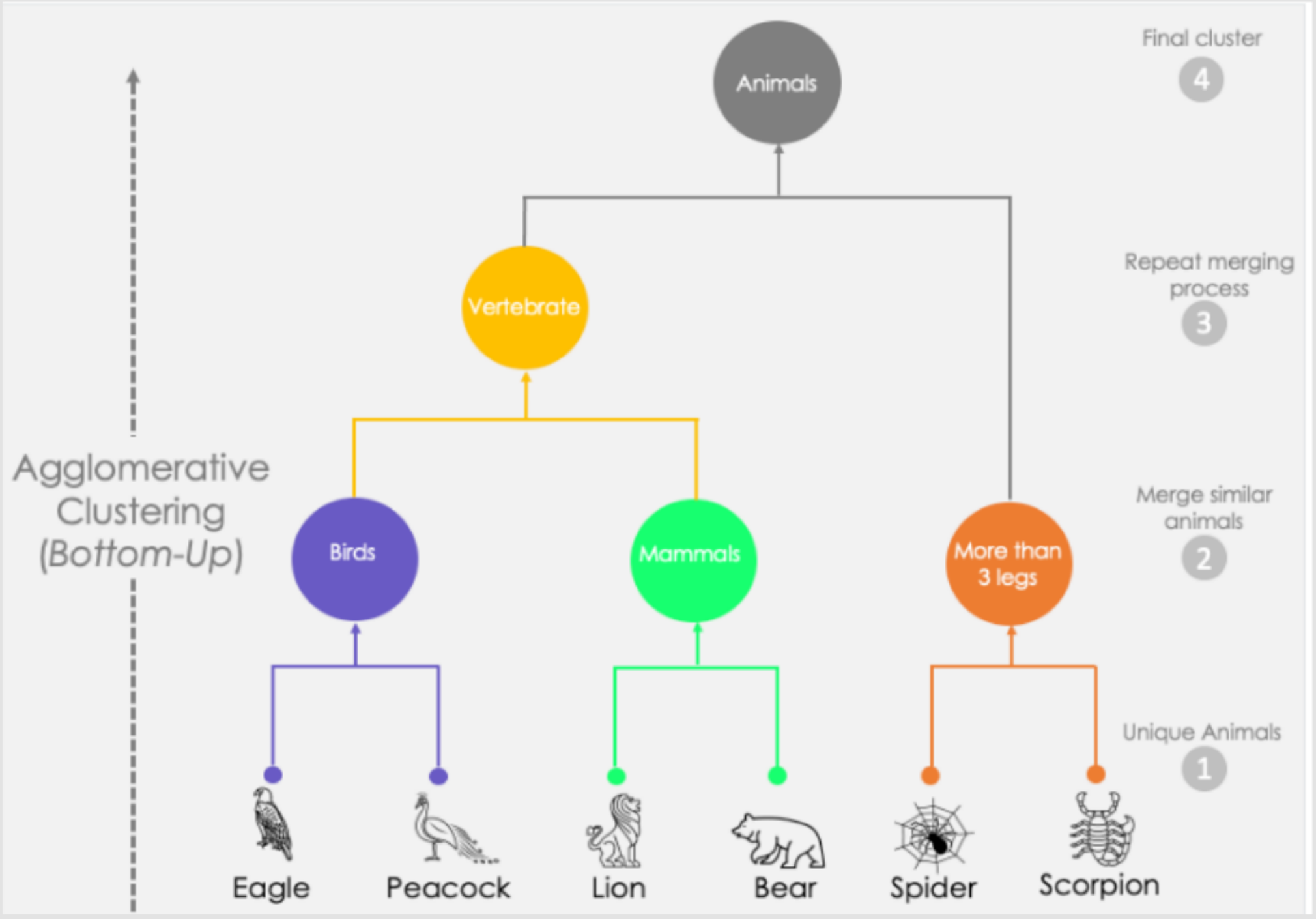
	P_1	P_2	P_3	P_4	P_5	P_6
C_1	1	0	1	1	0	0
C_2	0	1	0	0	1	0
C_3	1	0	0	1	0	0
C_4	0	1	0	0	1	1

$$d(C_1, C_3) = \sqrt{(1-1)^2 + (0-0)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2} = 1$$

$$d(C_1, C_2) = \sqrt{5}$$

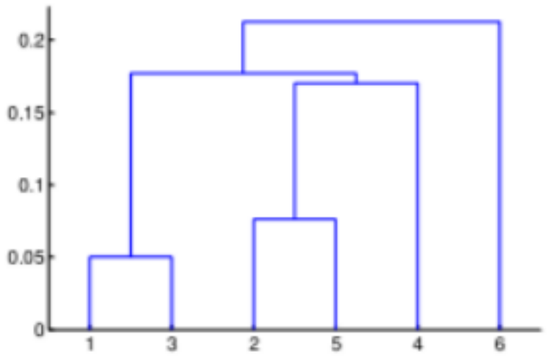
$$d(C_1, C_4) = \sqrt{6}$$

Hierarchical Clustering


Agglomerative Clustering (Bottom-Up)
Animals
Vertebrate
Birds
Mammals
More than 3 legs
Eagle
Peacock
Lion
Bear
Spider
Scorpion
Final cluster
4
Repeat merging process
3
Merge similar animals
2
Unique Animals
1

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that represent cluster-subcluster relationships and records the sequences (order) of merges or splits



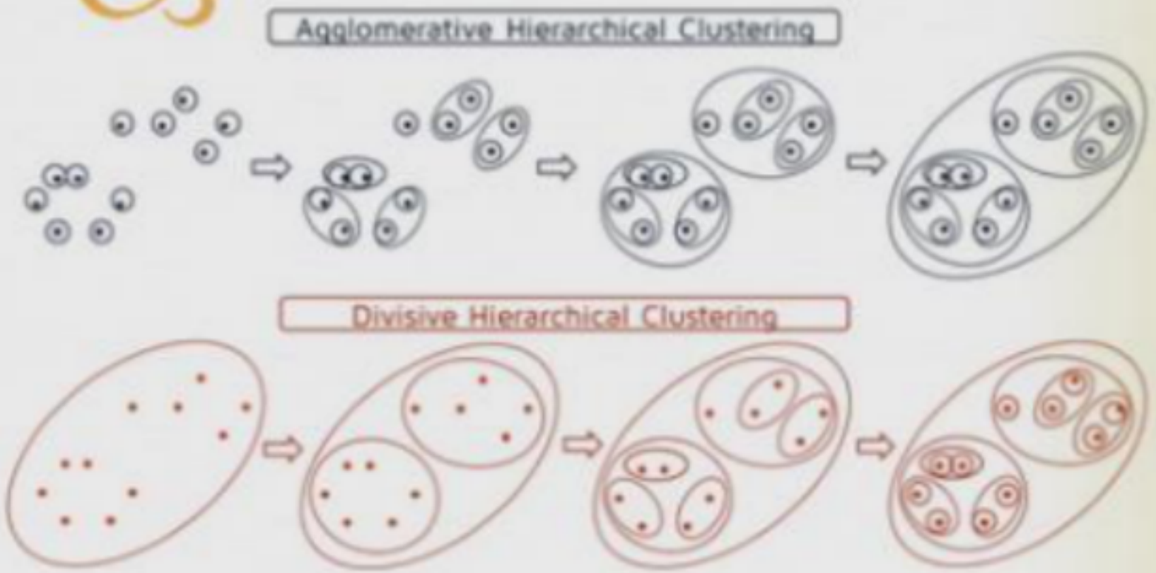
Strengths of Hierarchical Clustering

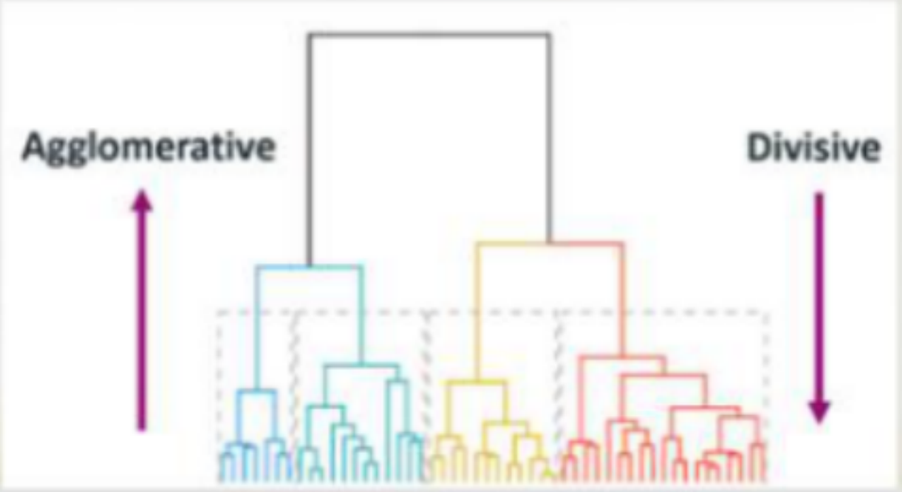
- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level

- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

Hierarchical Clustering

- Clusters are created in levels actually creating sets of clusters at each level.
- *Agglomerative*
  - Initially each item in its own cluster
  - Iteratively clusters are merged together
  - Bottom Up
- *Divisive*
  - Initially all items in one cluster
  - Large clusters are successively divided
  - Top Down





Hierarchical Clustering Approaches

1. Agglomerative: start with data points as *individual clusters (bottom-up)*
   - at each step merge the closest pair of clusters
   - *Definition of "cluster proximity" needed.*
2. Divisive: start with one all-inclusive cluster *(top-down)*
   - at each step split a cluster until only singleton clusters remain
   - Need to decide which cluster to split and how to do splitting

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

Agglomerative Clustering Algorithm

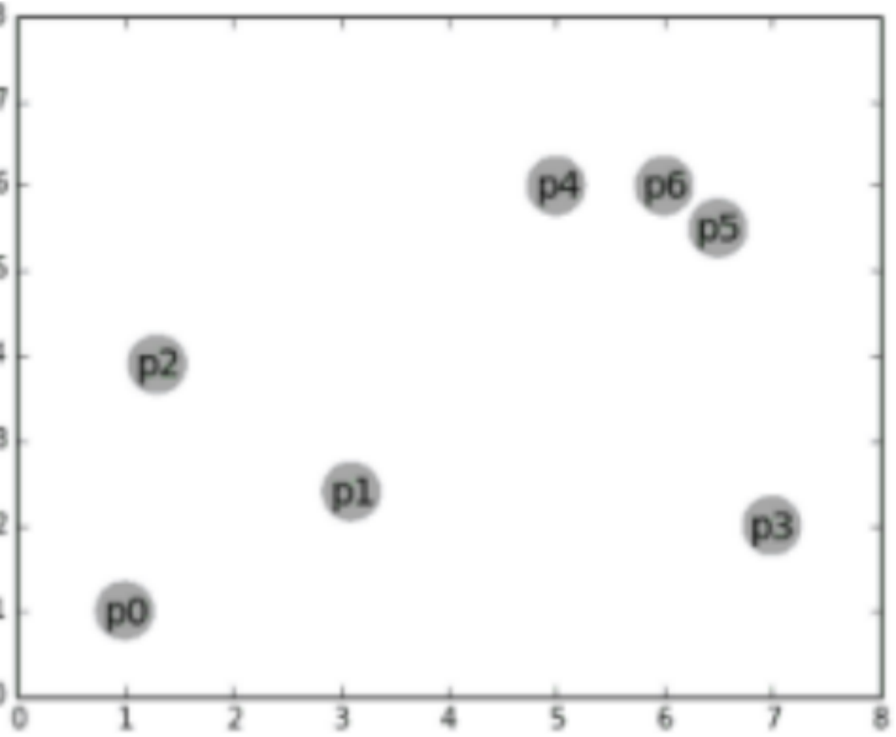
- Most popular hierarchical clustering technique
  - **Key Idea: Successively merge closest clusters**

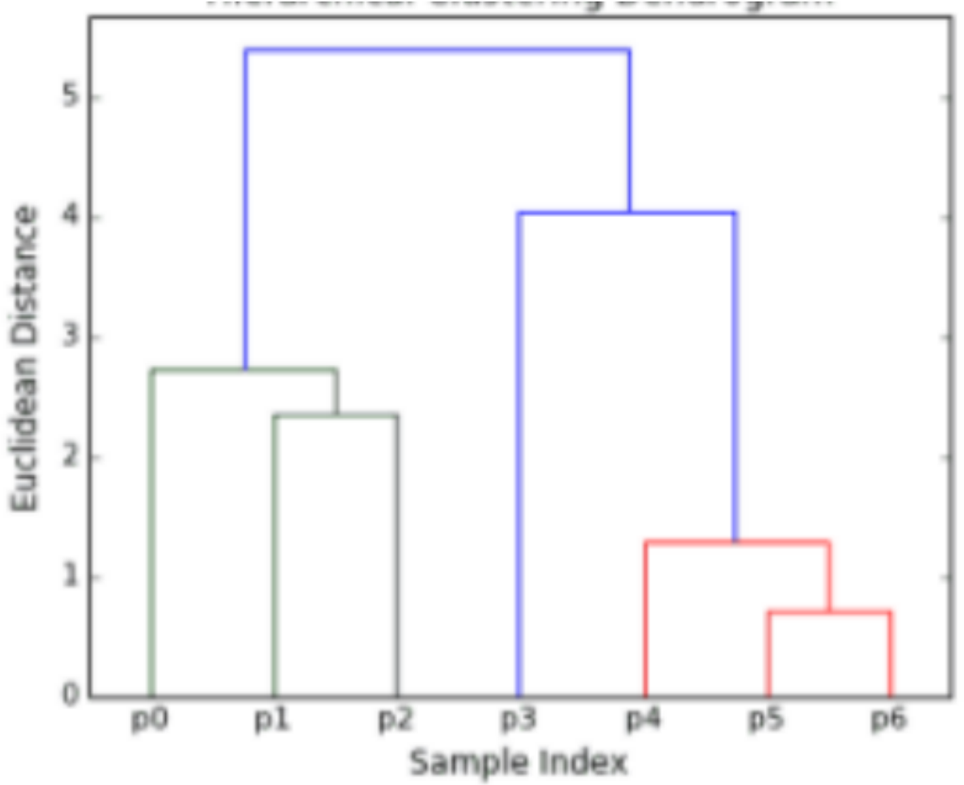
1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

Originally, the distance between two points

Update with distance between two clusters.
- How to define?

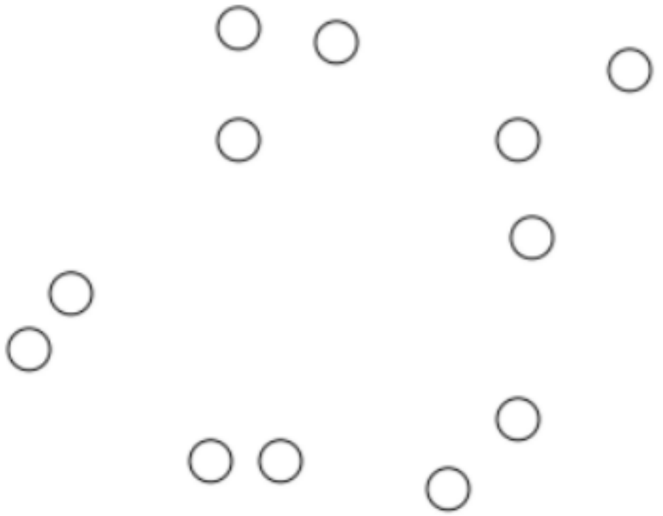
- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms


p4
p6
p5
p2
p1
p3
p0
0
1
2
3
4
5
6
7
8


Euclidean Distance
0
1
2
3
4
5
p0
p1
p2
p3
p4
p5
p6
Sample Index

Starting Situation

- Start with clusters of individual points and a proximity matrix



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

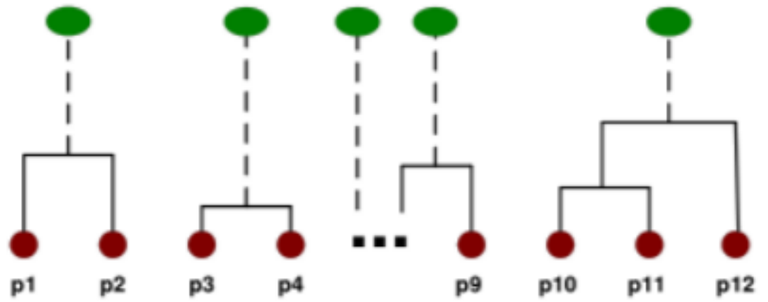
Proximity Matrix

Intermediate Situation

- After some merging steps, we have some clusters

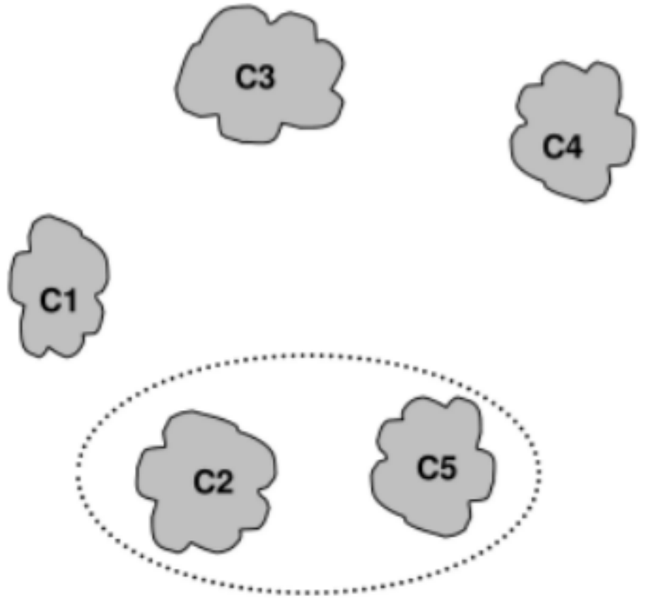
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



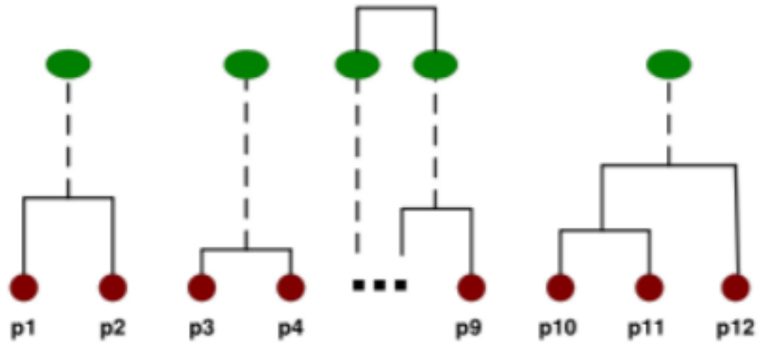
Intermediate Situation

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



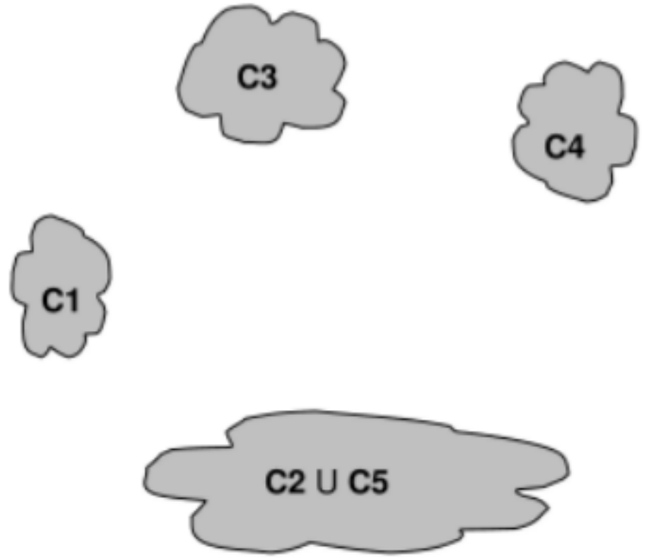
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



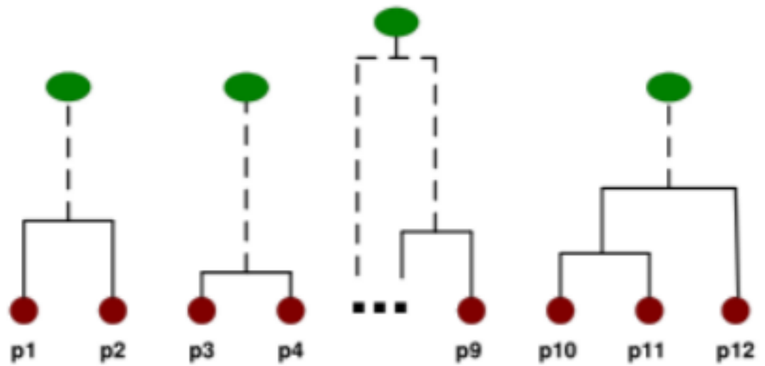
After Merging

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

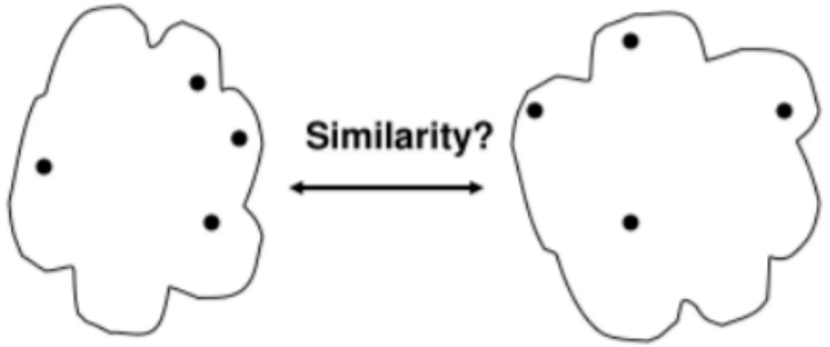


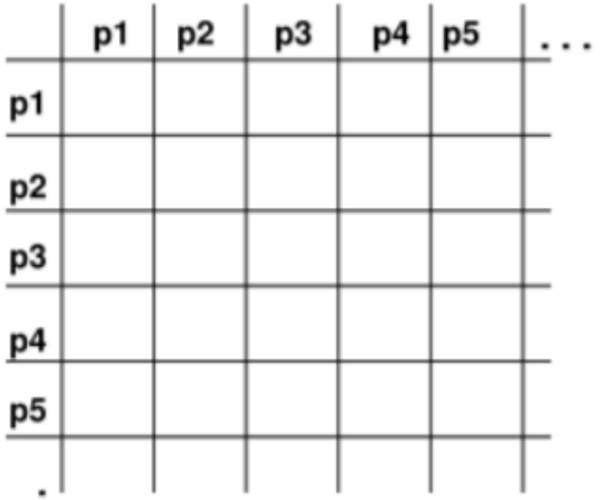
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Distance



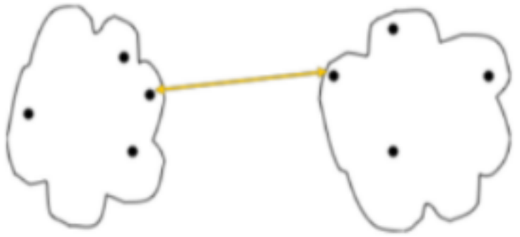


Proximity Matrix

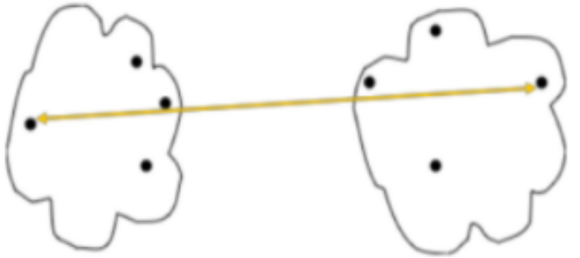
- **MIN** (short/single-link)
- **MAX** (suggestive/complete-link)
- **Group Average**
- **Distance Between Centroids**
- Other methods driven by an objective function
  - Ward's Method uses squared error

efining Proximity between Clusters

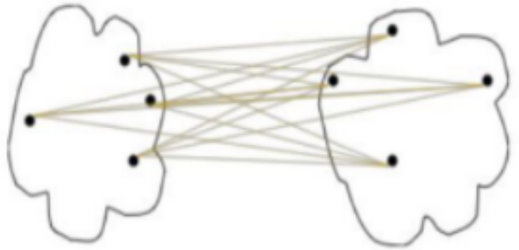
N (single-link)



MAX (complete-link)



Group Average



- **Single Linkage**

  $D(c_1,c_2) = \min D(x_1,x_2)$

  Minimum distance or distance between closest elements in clusters

- **Complete Linkage**

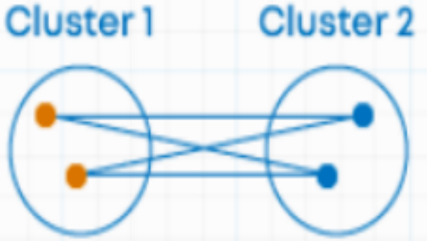
  $D(c_1,c_2) = \max D(x_1,x_2)$

  Maximum distance between elements in clusters

- **Average Linkage**

  $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\Sigma\Sigma D(x_1,x_2)$

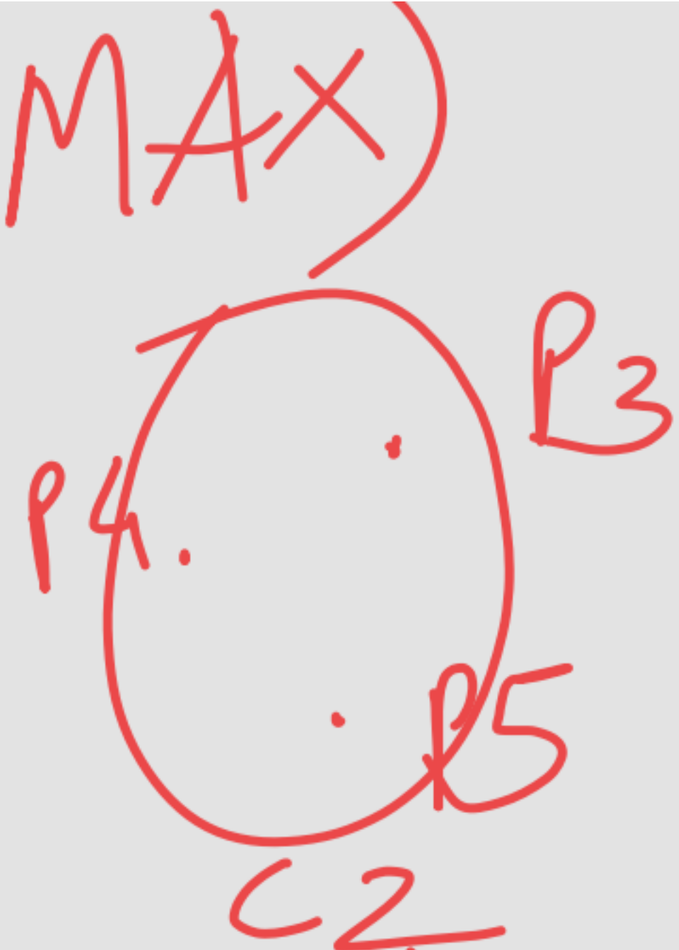
  Average of the distances of all pairs



- **Centroid Method**

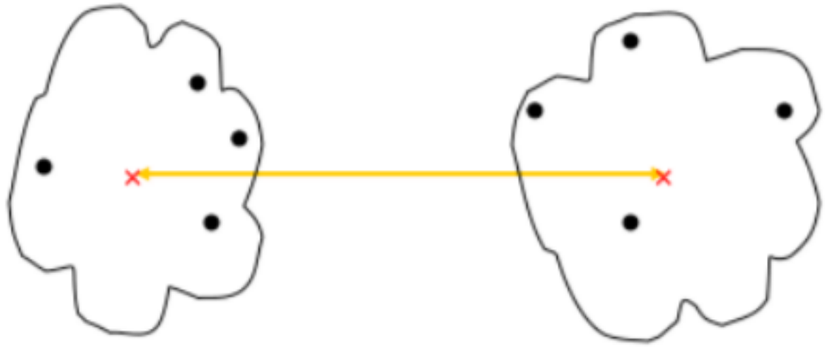
  Combining clusters with minimum distance between the centroids of the two clusters

Complete Linkage (MAX)



$$d(C_1, C_2) = \max\big(d(P1,P3), d(P1,P4), d(P1,P5), d(P2,P3), d(P2,P4), d(P2,P5)\big)$$

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

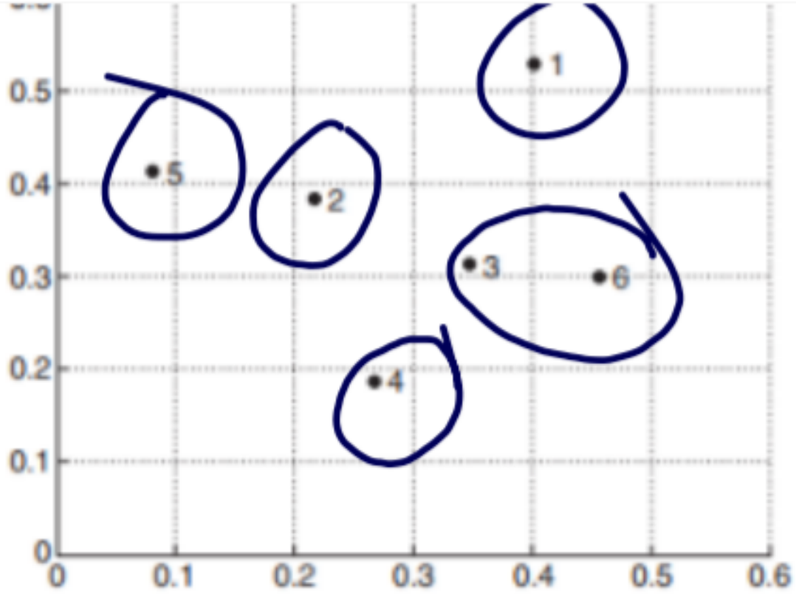


Figure 7.15. Set of six two-dimensional points.

Point	x Coordinate	y Coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table 7.3. xy-coordinates of six points.

Proximity Matrix

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

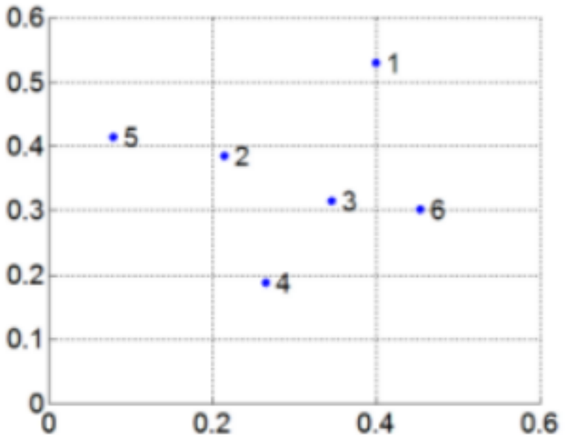
$$= \sqrt{|0.40 - 0.22|^2 + |0.53 - 0.38|^2}$$

	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

Table 7.4. Euclidean distance matrix for six points.

MIN or Single Link

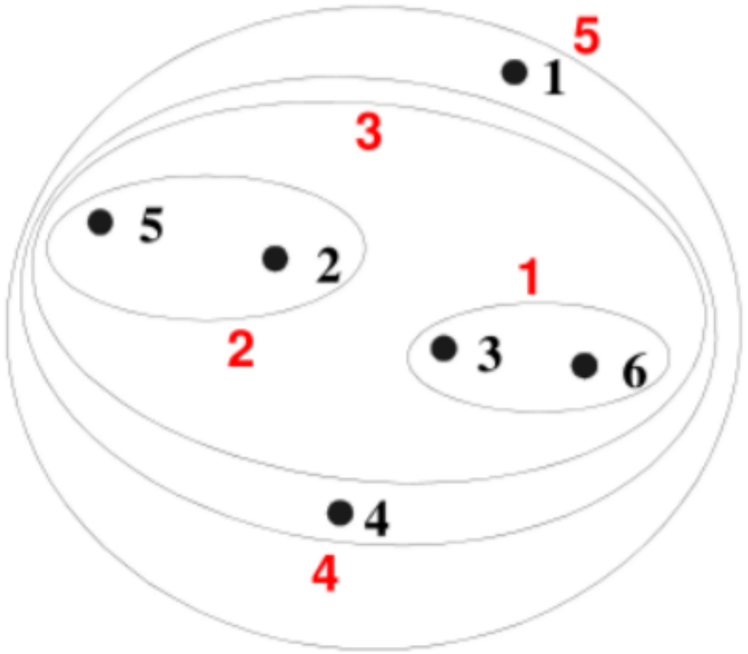
- Proximity of two clusters is based on the two closest points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph
- Example:



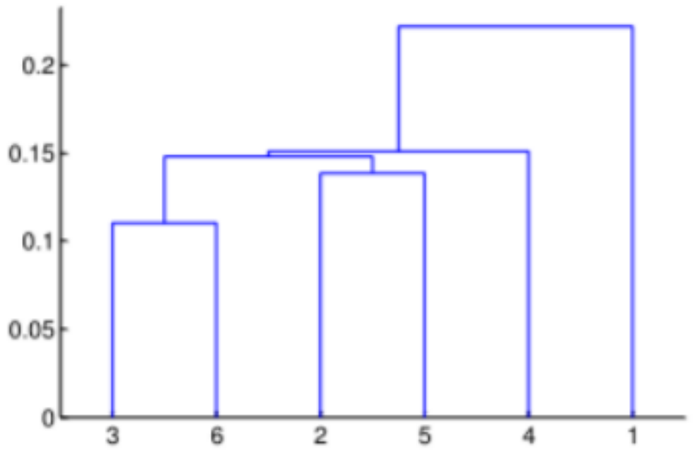
Distance Matrix:

	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

Let's say we now have:

- Cluster {3,6}
- Cluster {2,5}
- Cluster {4}
- Cluster {1}

What's the next merge?

MIN({3,6},{2,5}) = 0.15
Is the next smallest value

Min Linkage

Distance matrix

	p1	p2	p3	p4	p5	p6
p1	0					
p2	0.24	0				
p3	0.22	0.15	0			
p4	0.37	0.20	0.15	0		
p5	0.34	0.14	0.28	0.29	0	
p6	0.23	0.25	0.11	0.22	0.39	0

d(P1, P3UP6) = ? min(d(P1,P3), d(P1,P6))

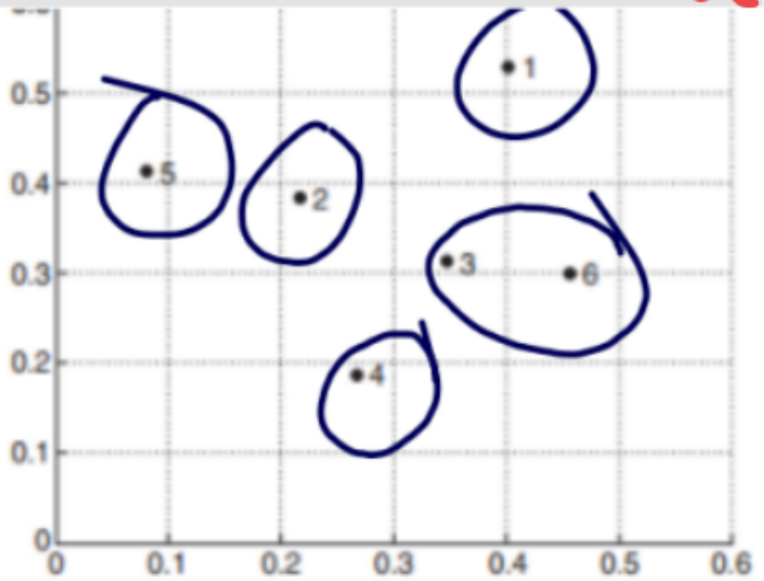


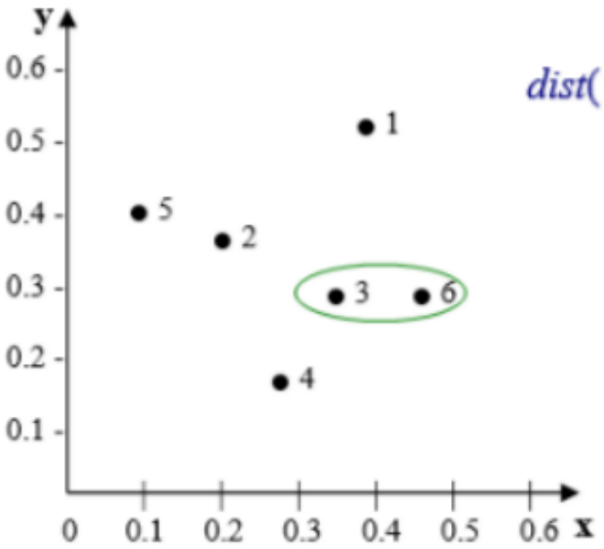
Figure 7.15. Set of six two-dimensional points.

Point	x Coordinate	y Coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table 7.3. xy-coordinates of six points.

space

dendogram



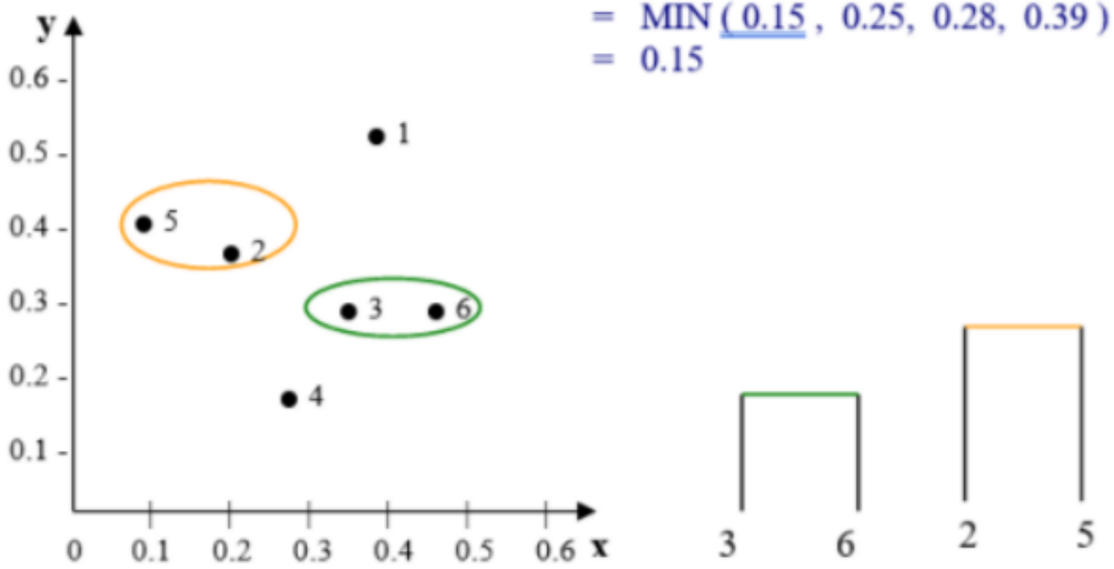
dist((p3, p6), p1) = MIN (*dist*(p3, p1) , *dist*(p6, p1))
= MIN (0.22 , 0.23)
= 0.22

Distance matrix

p1	0				
p2	0.24	0			
(p3, p6)	**0.22**	**0.15**	0		
p4	0.37	0.20	**0.15**	0	
p5	0.34	0.14	**0.28**	0.29	0
	p1	p2	**(p3, p6)**	p4	p5

space

dist((p3, p6), (p2, p5)) = MIN (*dist*(p3, p2) , *dist*(p6, p2), *dist*(p3, p5), *dist*(p6, p5))
= MIN (0.15 , 0.25, 0.28, 0.39) //from original matrix
= 0.15

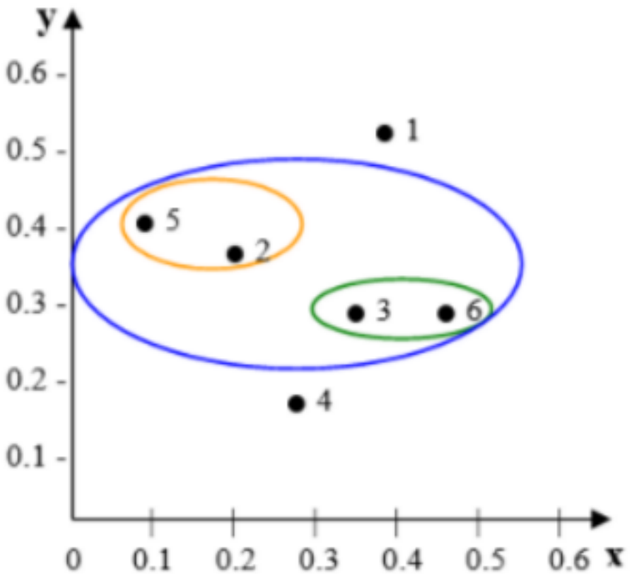


Distance matrix

p1	0			
(p2, p5)	**0.24**	0		
(p3, p6)	0.22	**0.15**	0	
p4	0.37	**0.20**	0.15	0
	p1	**(p2, p5)**	(p3, p6)	p4

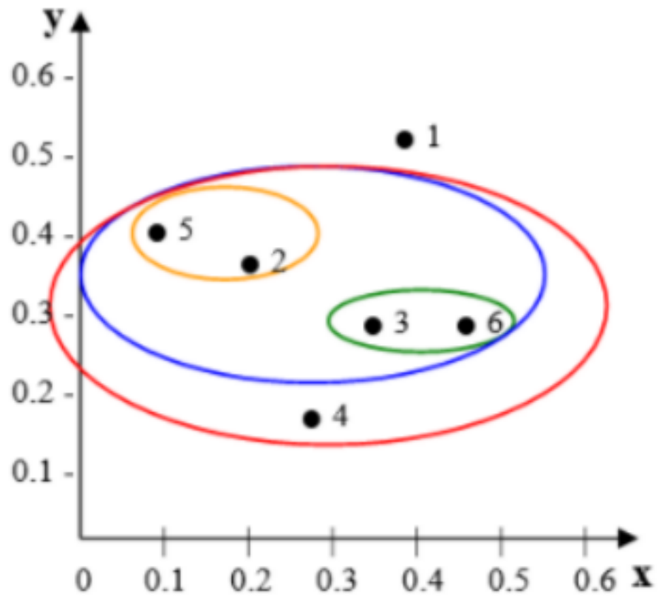
space

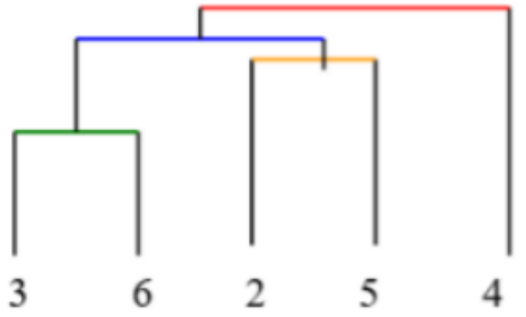
dendogram



Distance matrix

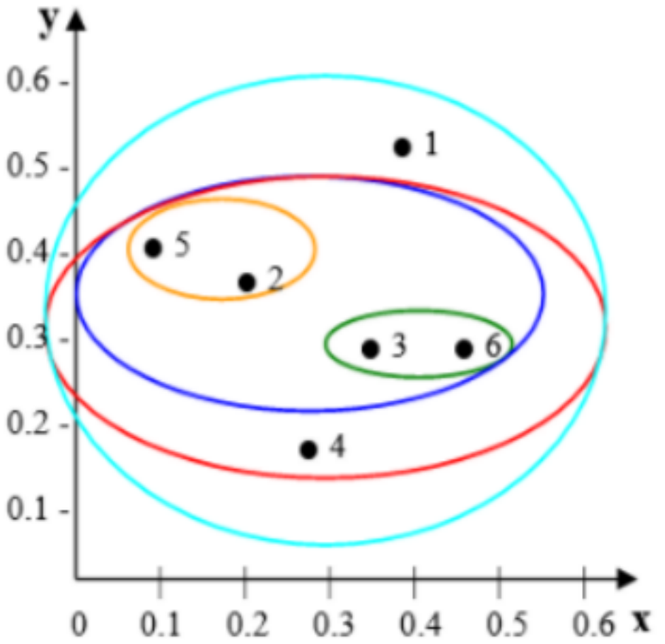
p1	0		
(p2, p5, p3, p6)	**0.22**	0	
p4	0.37	**0.15**	0
	p1	**(p2, p5, p3, p6)**	p4

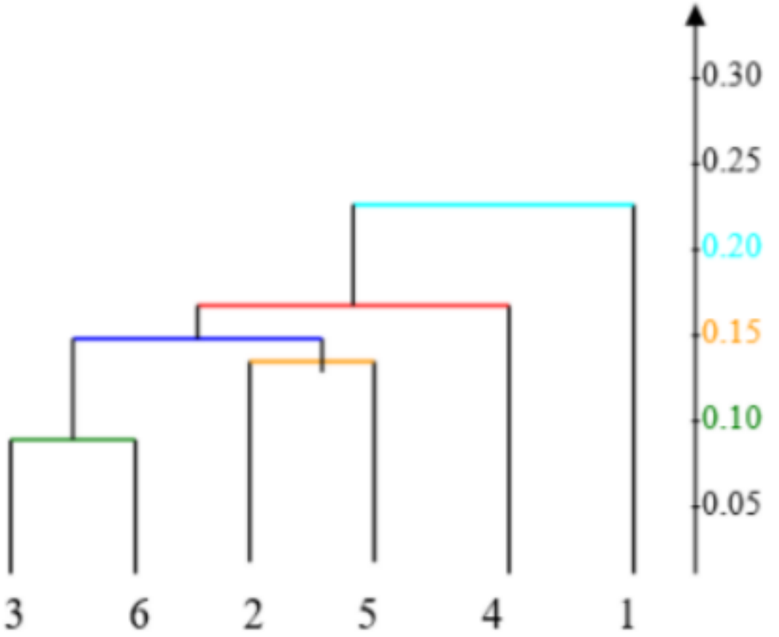




Distance matrix

p1	0	
(p2, p5, p3, p6, p4)	0.22	0
	p1	(p2, p5, p3, p6, p4)





Single Link Clustering - Example 7

Strength of MIN

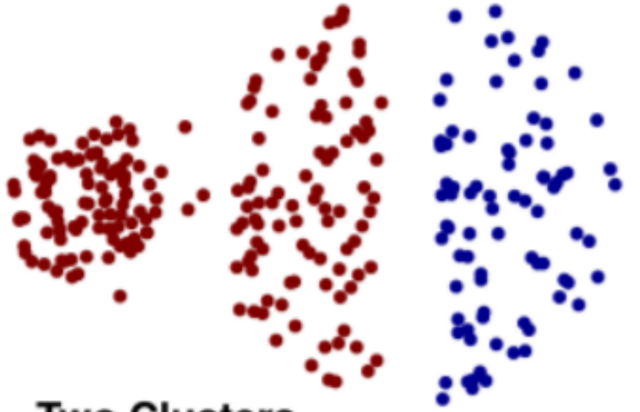
Original Points

Six Clusters

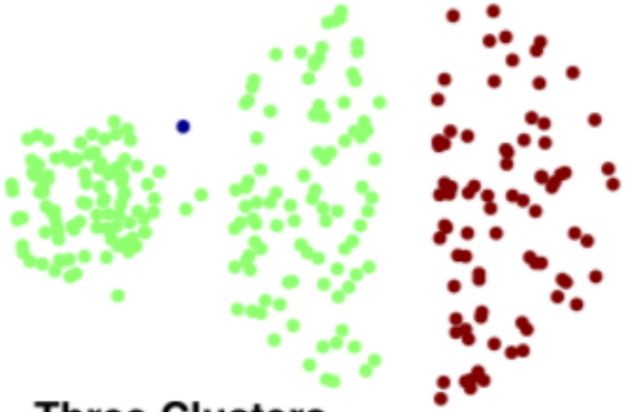
- Can handle non-elliptical shapes

Limitations of MIN

Original Points



Two Clusters

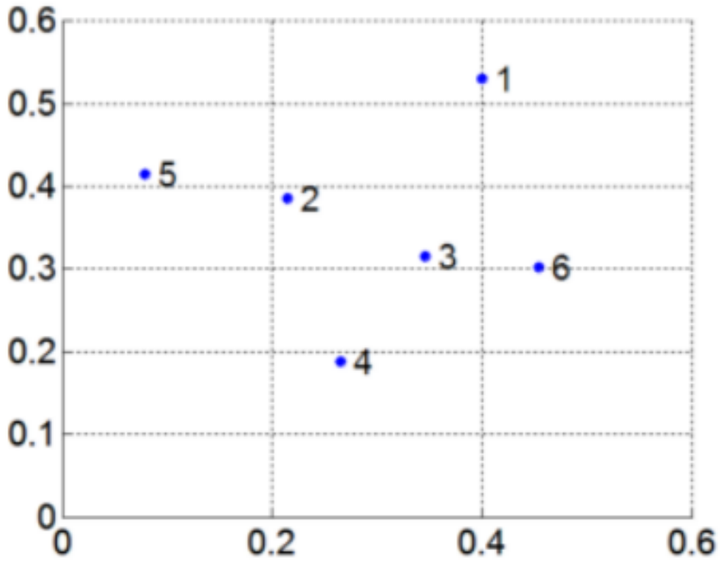


Three Clusters

- Sensitive to noise and outliers

MAX or Complete Linkage or CLIQUE

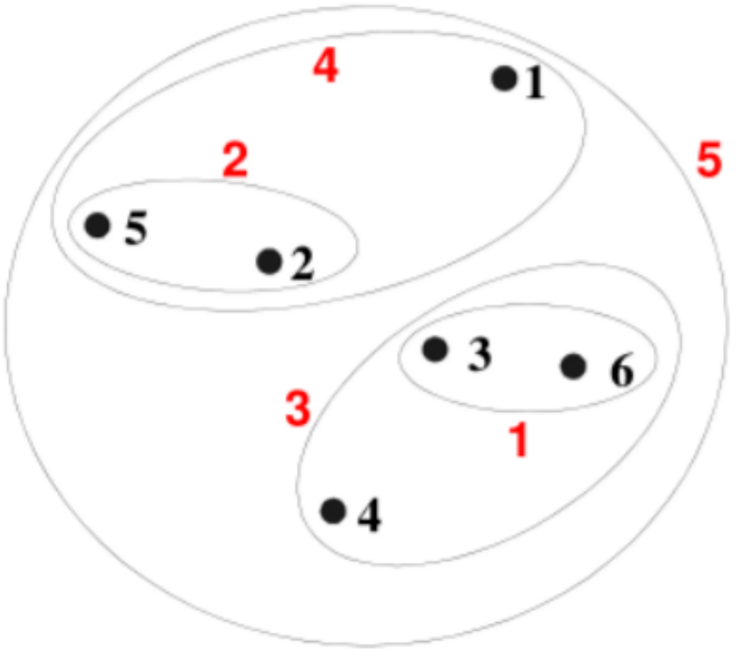
- Proximity of two clusters is based on the two most distant points in the different clusters
  - Determined by all pairs of points in the two clusters



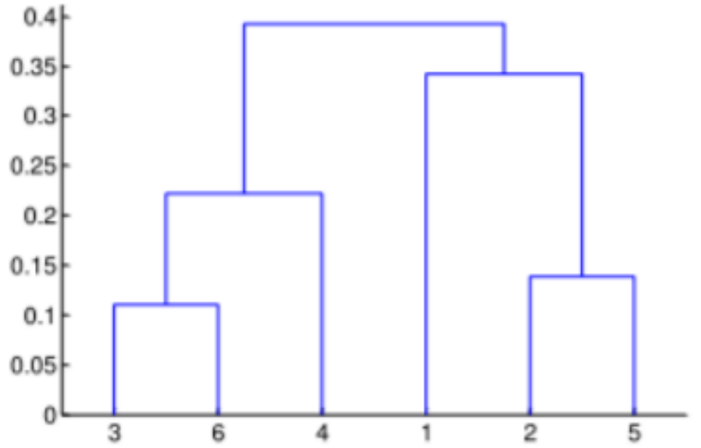
Distance Matrix:

	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



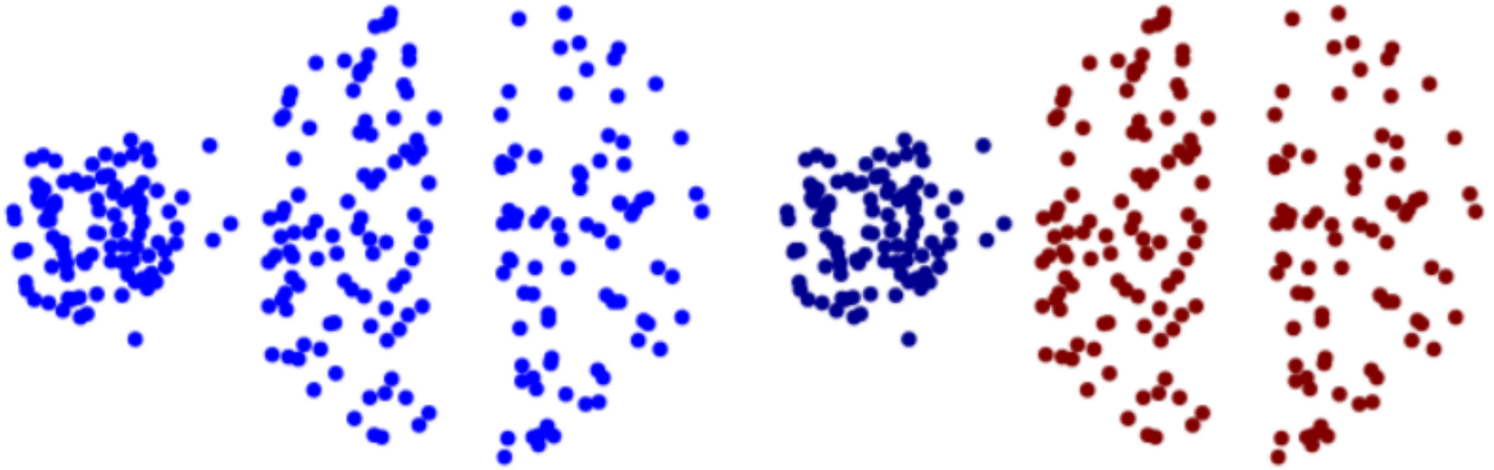
Distance Matrix:

	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

As with single link, points 3 and 6
are merged first. However, $\{3,6\}$ is merged with $\{4\}$, instead of $\{2,5\}$ or $\{1\}$ because

$$
\begin{aligned}
dist(\{3,6\},\{4\}) &= \max(dist(3,4),dist(6,4)) \\
&= \max(0.15,0.22) \\
&= 0.22. \\
dist(\{3,6\},\{2,5\}) &= \max(dist(3,2),dist(6,2),dist(3,5),dist(6,5)) \\
&= \max(0.15,0.25,0.28,0.39) \\
&= 0.39. \\
dist(\{3,6\},\{1\}) &= \max(dist(3,1),dist(6,1)) \\
&= \max(0.22,0.23) \\
&= 0.23.
\end{aligned}
$$

Strength of MAX



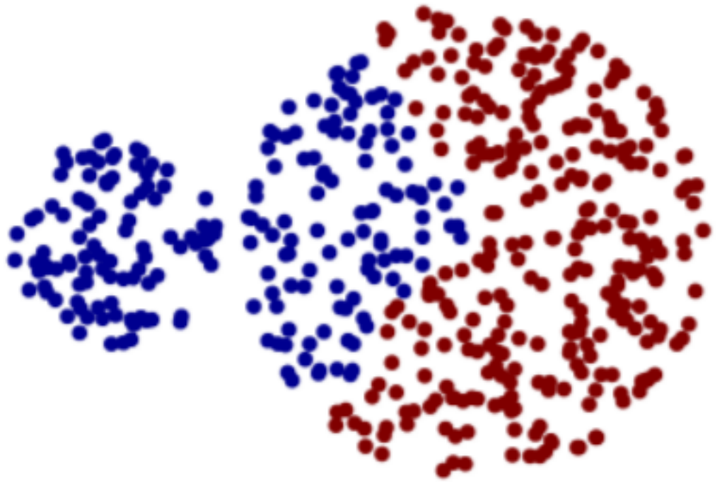
Original Points

Two Clusters

- Less susceptible to noise and outliers

Limitations of MAX

Original Points



Two Clusters

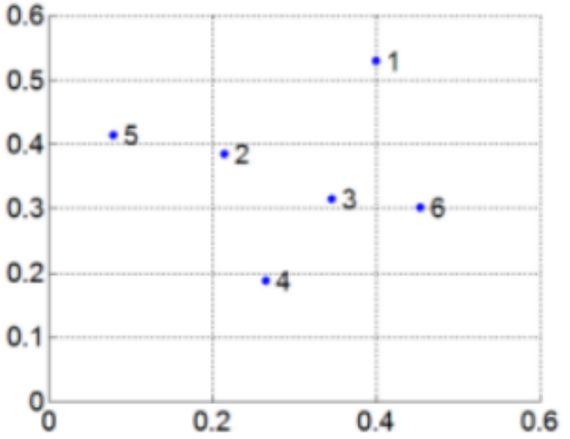
- Tends to break large clusters
- Biased towards globular clusters

Group Average

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

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

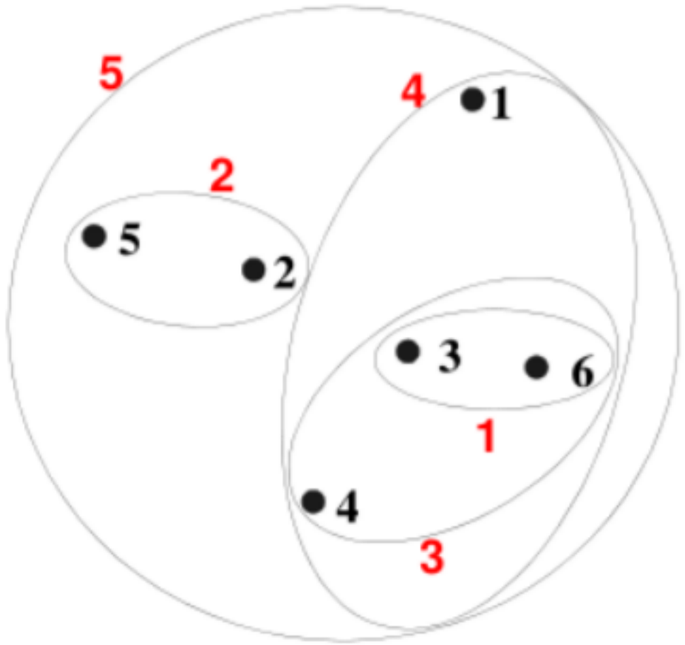
- Need to use average connectivity for scalability since total proximity favors large clusters



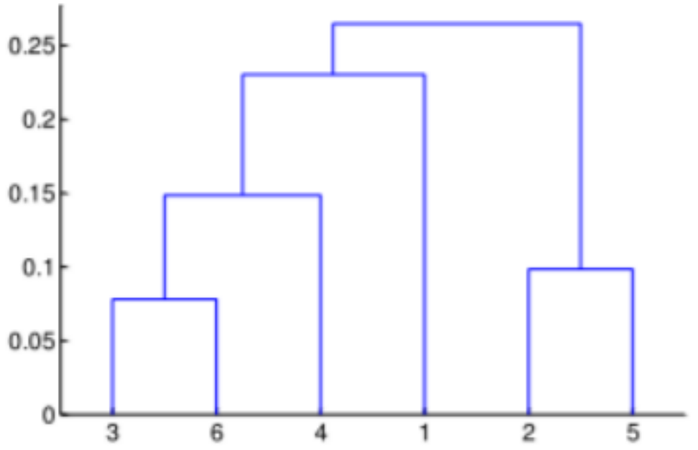
Distance Matrix:

	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: Group Average



Nested Clusters



Dendrogram

Example 8.6 (Group Average). Figure 8.18 shows the results of applying the group average approach to the sample data set of six points. To illustrate how group average works, we calculate the distance between some clusters.

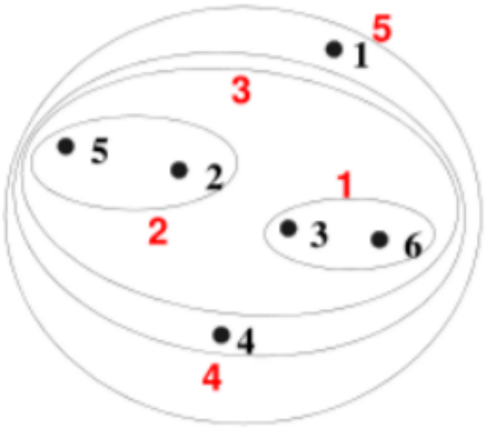
$$
\begin{aligned}
dist(\{3,6,4\},\{1\}) &= (0.22+0.37+0.23)/(3*1) \\
&= 0.28 \\
dist(\{2,5\},\{1\}) &= (0.2357+0.3421)/(2*1) \\
&= 0.2889 \\
dist(\{3,6,4\},\{2,5\}) &= (0.15+0.28+0.25+0.39+0.20+0.29)/(6*2) \\
&= 0.26
\end{aligned}
$$

Because $dist(\{3,6,4\},\{2,5\})$ is smaller than $dist(\{3,6,4\},\{1\})$ and $dist(\{2,5\},\{1$ clusters $\{3,6,4\}$ and $\{2,5\}$ are merged at the fourth stage. ■

Hierarchical Clustering: Group Average

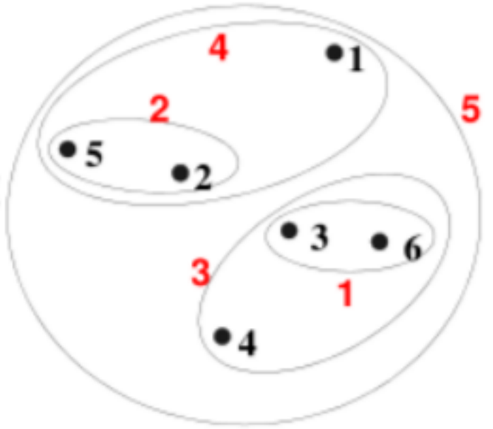
- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise and outliers
- Limitations
  - Biased towards globular clusters

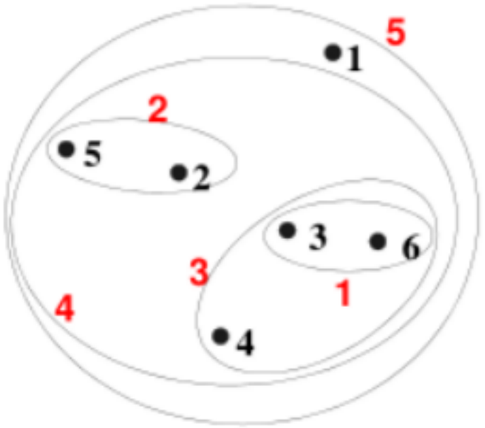
Hierarchical Clustering: Comparison



MIN

MAX





Group Average

Hierarchical Clustering: Time Complexity

- Space complexity: $O(n^2)$
- Time complexity:

	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

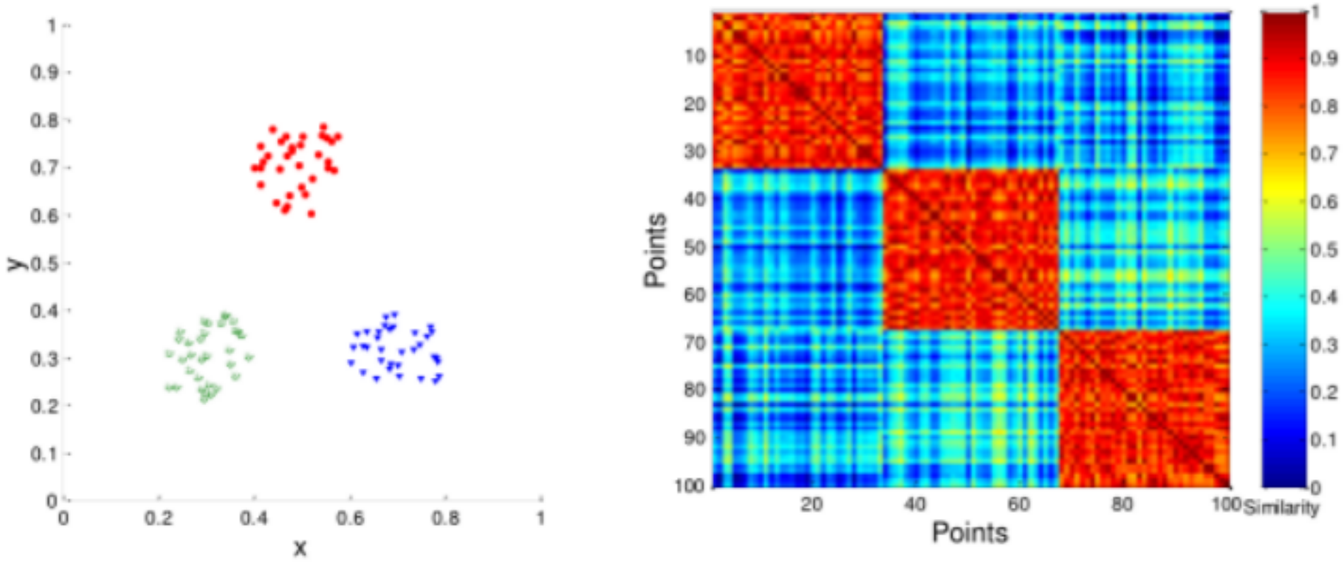
- $O(n^3)$
  - *n* steps (number of merges)
  - At each step: proximity matrix must be searched: n^2

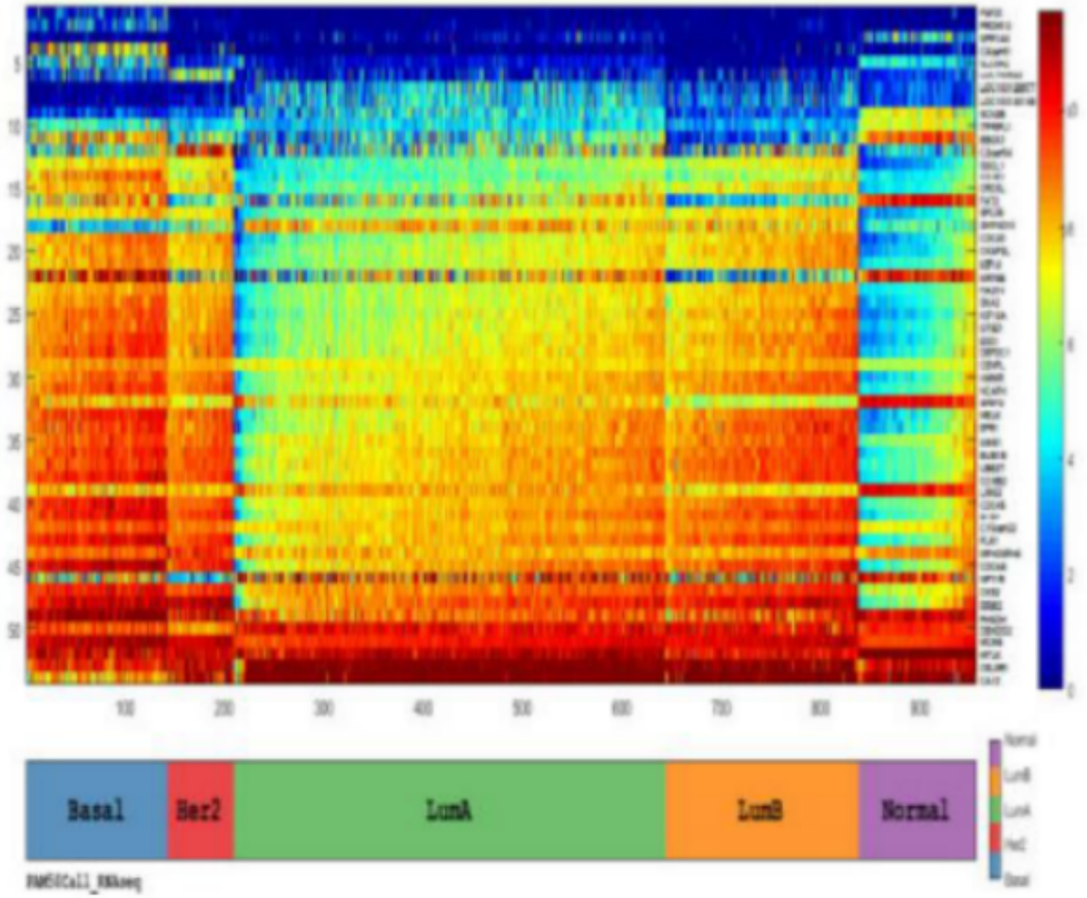
Hierarchical Clustering: Problems and Limitations

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

Using Similarity Matrix for Cluster Validation

- Order the similarity matrix with respect to cluster labels and inspect visually.

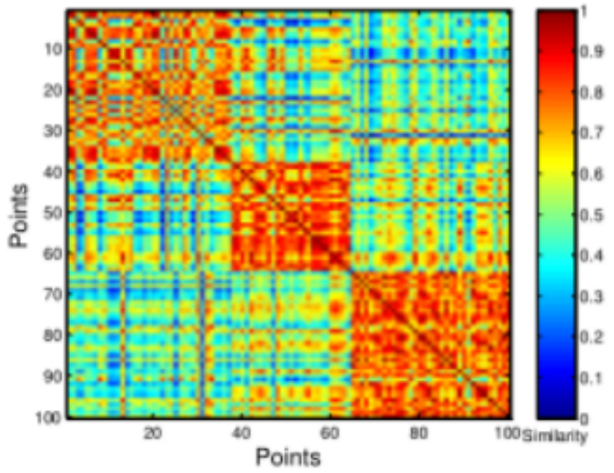


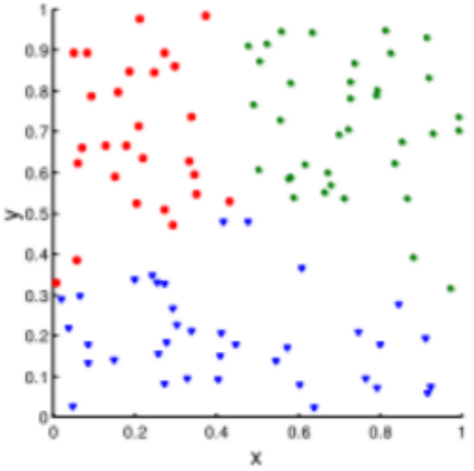


(a) Segregation of five breast cancer subtypes based on differential expression of identified 54 biomarker genes for 956 patients.

Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp





K-means

Internal Measures: SSE

- Clusters in more complicated figures aren't well separated
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
  - SSE
- SSE is good for comparing two clusterings or two clusters
- Can also be used to estimate the number of clusters