

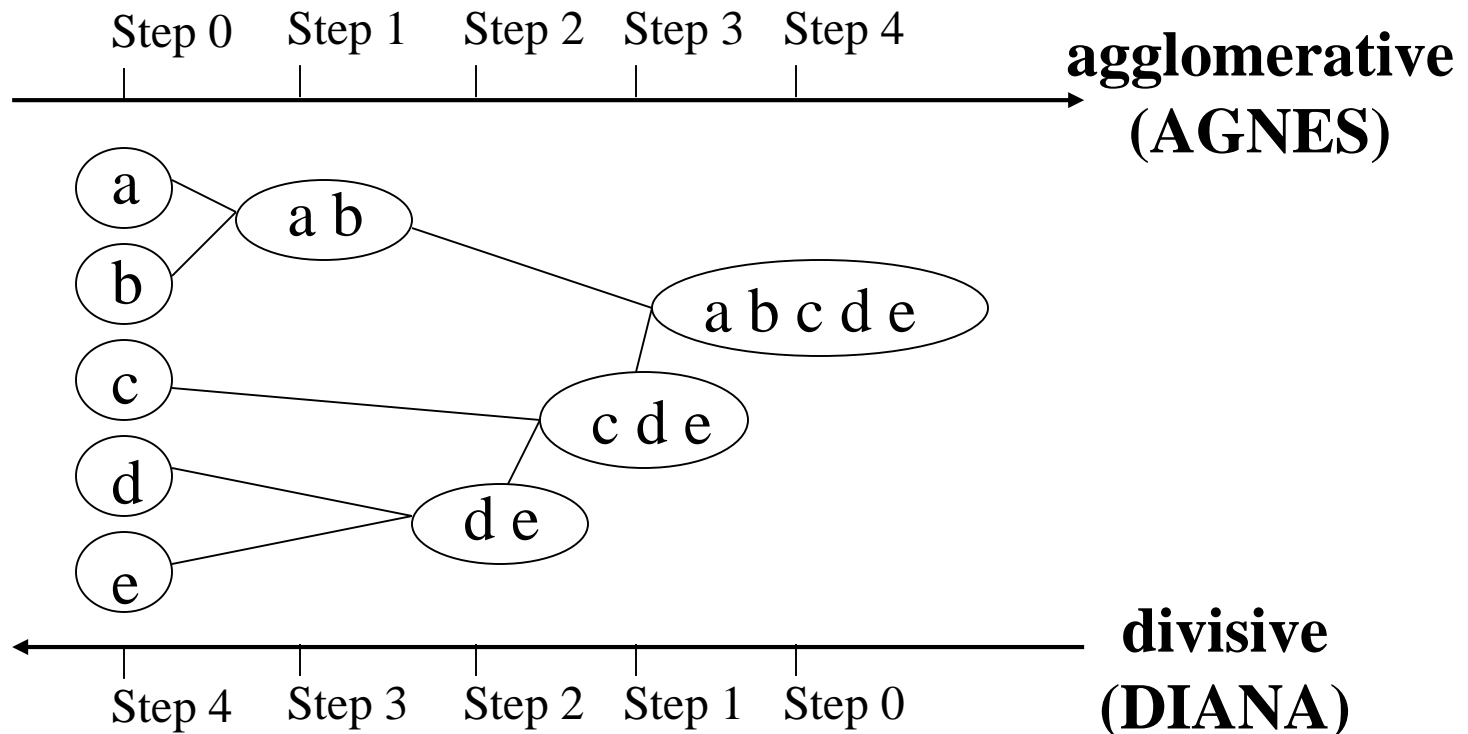
계층 기반 클러스터링 기법

- 계층 단계별로 군집들이 생성됨. 각 레벨별로 다수의 클러스터들이 먼저 생성
- **병합적 방법**
 - 시작시 각 객체를 하나의 클러스터 군집으로 간주함
 - 반복적으로 클러스터링 군집들을 서로 합병함
 - Bottom Up
- **분할적 방법**
 - 시작시 모든 객체들은 한 클러스터에 속함
 - 큰 클러스터 군집들을 반복적으로 분할함
 - Top Down

계층 기반 클러스터링 기법

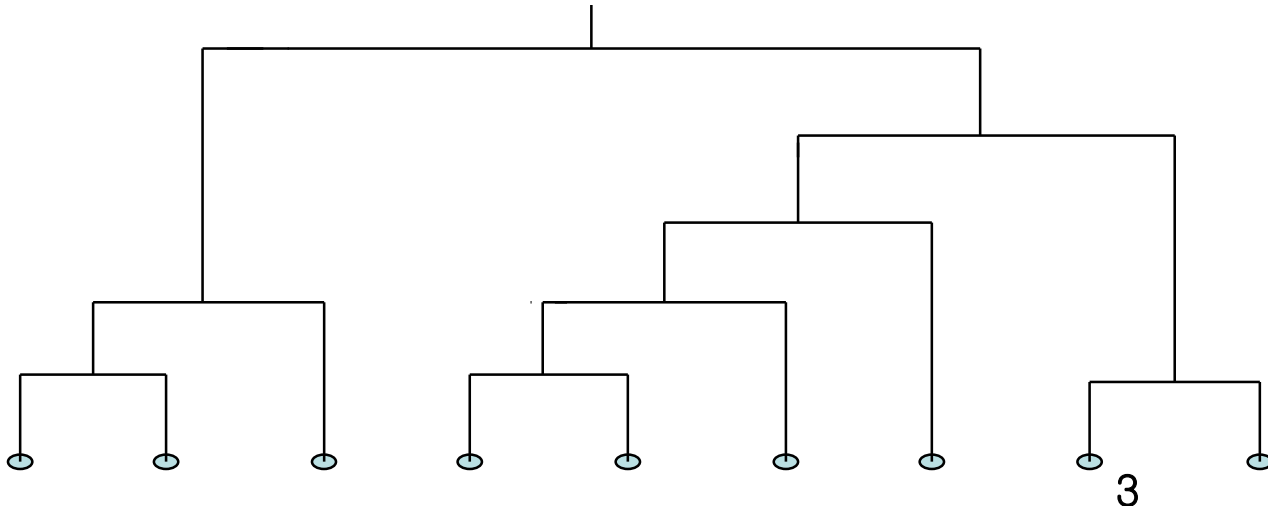
- 군집분석 방법에 따른 거리 매트릭스를 사용함

k 와 같은 값이 필요하지 않으나, 적절한 군집을 나누기 위한 조건이 필요함

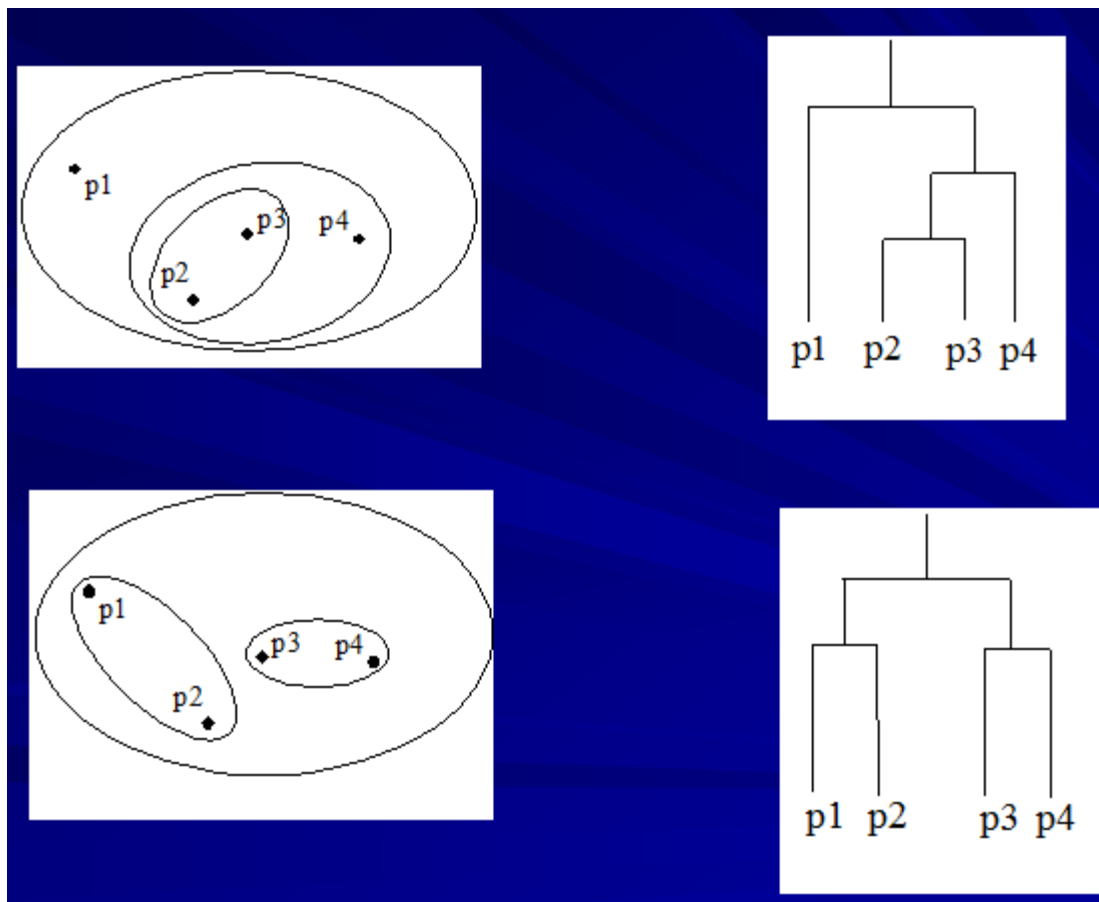


Dendrogram

- 군집들이 계층적으로 합병되는 구조를 보여줌
- 데이터 객체를 계층적으로 포함하는 군집의 트리 형태로 분류함
- 데이터 객체들의 군집은 덴드로그램에서 적정 레벨에서의 분단선으로 찾을 수 있음
이때, 각 연결된 객체들이 동일 클러스터를 구성함



Hierarchical Clustering

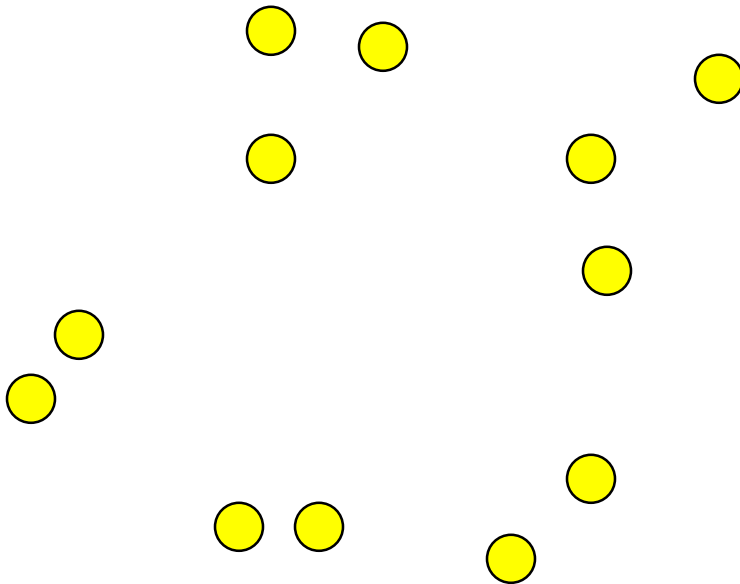


병합적 클러스터링 알고리즘

- 보편적인 계층적 군집 분석 알고리즘
- **Basic algorithm is straightforward**
 1. 객체 간의 거리 기반 매트릭스 구축
 2. 각 객체를 단일 객체의 군집으로 간주
 3. [반복]
 4. 두 인접한 군집을 병합함
 5. 병합에 따라 거리 매트릭스를 갱신함
 6. [병합하여 단 하나의 군집만 남을때까지 반복]
- 두 군집 간의 유사도 계산이 주요 연산
 - 군집 간 유사도 거리 계산 방법에 따라 다양한 방법이 사용됨

Starting Situation

- Start with clusters of individual points and a proximity matrix



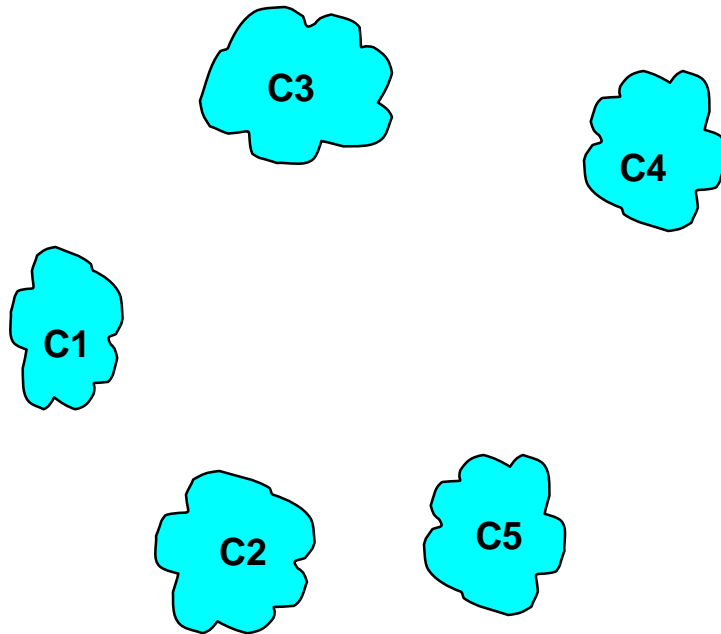
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

p1 p2 p3 p4 ... p9 p10 p11 p12

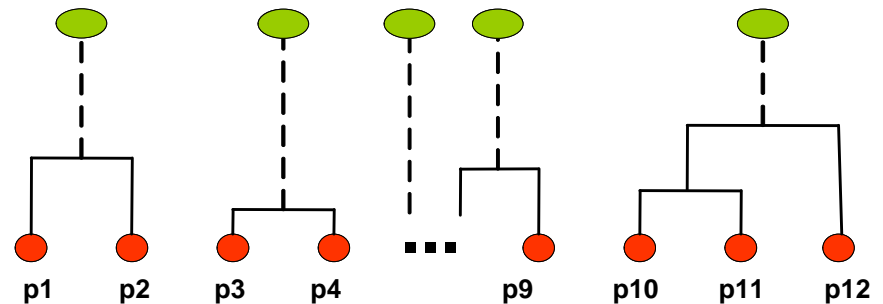
Intermediate Situation

- After some merging steps, we have some clusters



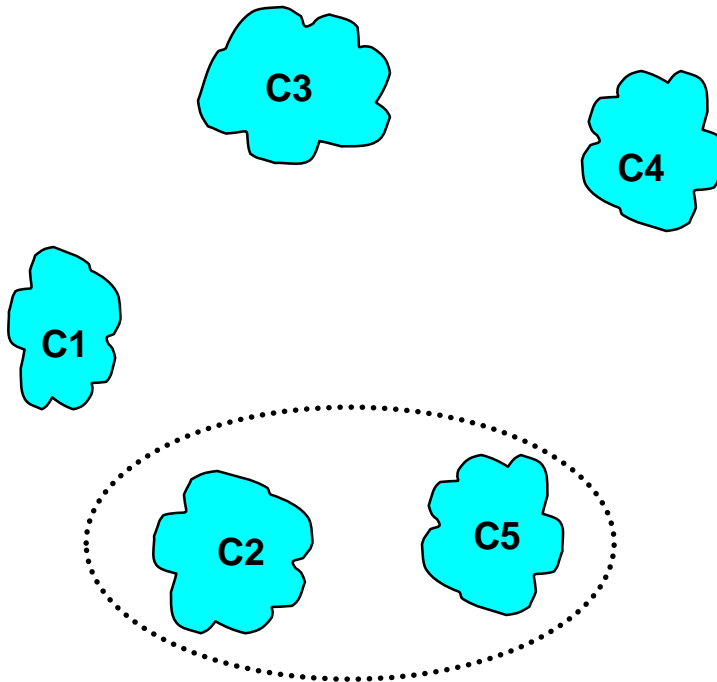
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



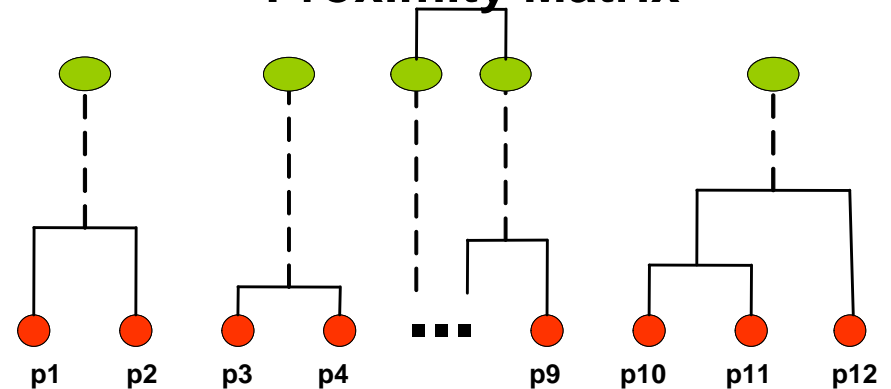
Intermediate Situation

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



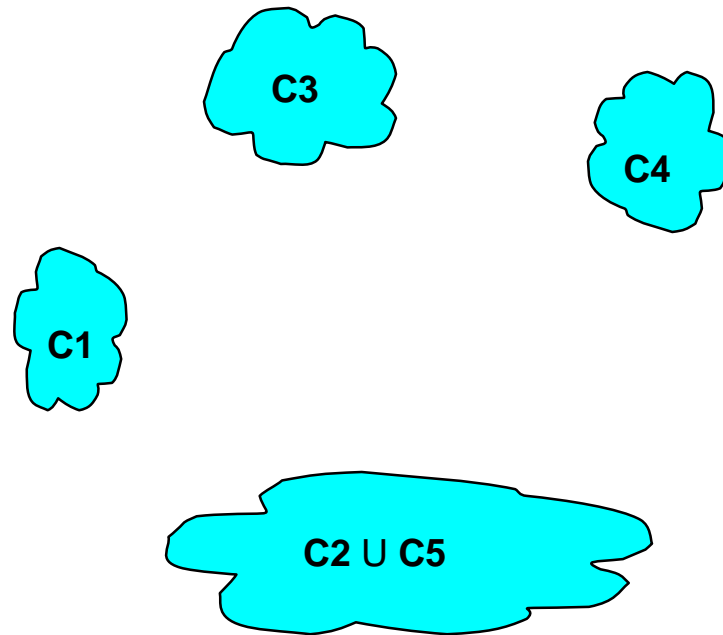
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



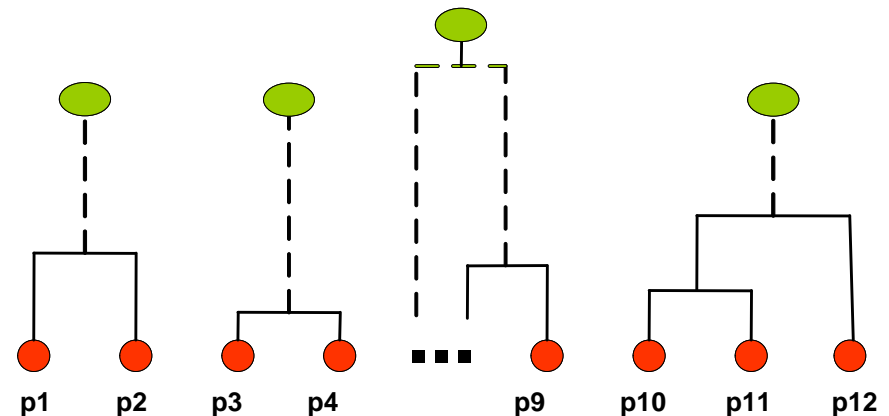
After Merging

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

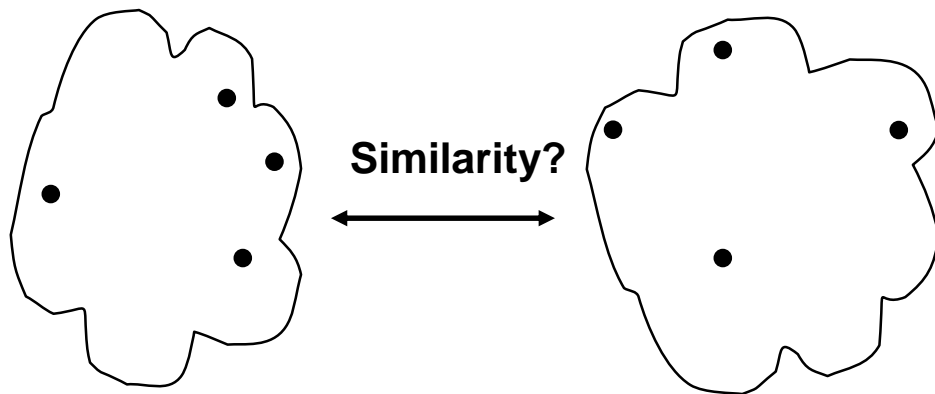


		C2 U C5			
		C1	C5	C3	C4
C1			?		
C2 U C5		?	?	?	?
C3			?		
C4			?		

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

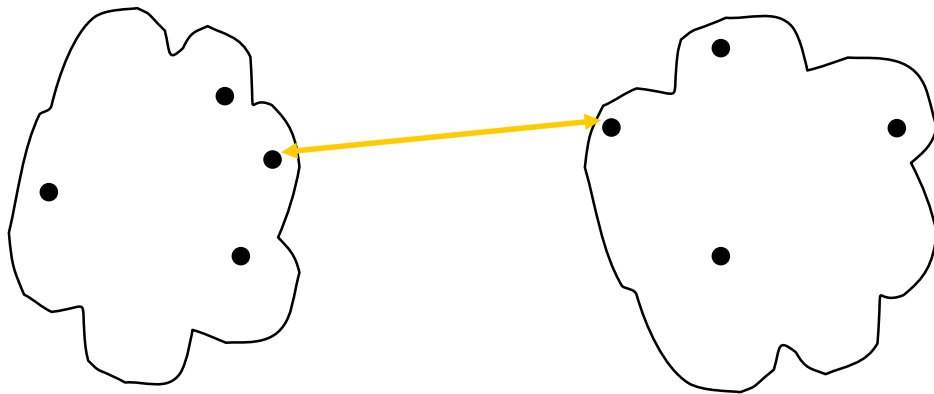
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						

· Proximity Matrix

Distance Between Clusters

- *Single Link(MIN)*
: smallest distance between points
- *Complete Link(MAX)*
: largest distance between points
- *Average Link(GROUP AVERAGE)*
: average distance between points
- *Distance Between Centroids*
: distance between centroids

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

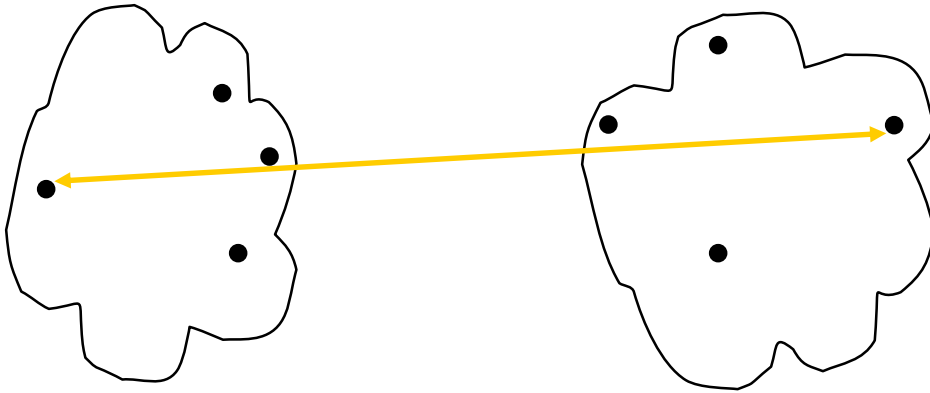
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Single Link

- Uses threshold distances at each level.
- View all items with links (distances) between them.
- Finds maximal connected components in this graph.
- Two clusters are merged if there is at least one edge which connects them.
- Could be agglomerative or divisive.

How to Define Inter-Cluster Similarity

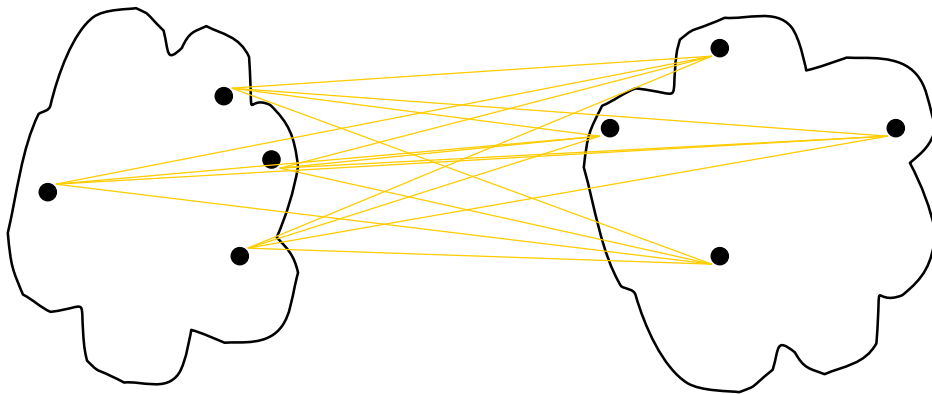


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - Ward's Method uses squared error

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

Proximity Matrix

How to Define Inter-Cluster Similarity

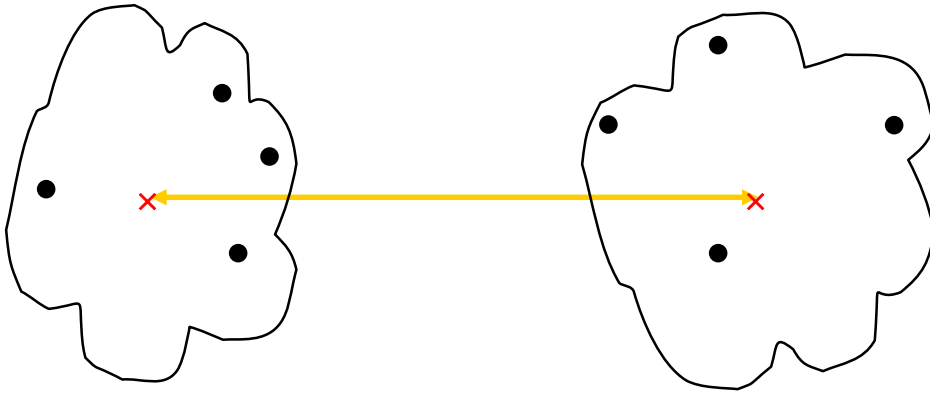


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - Ward's Method uses squared error

	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