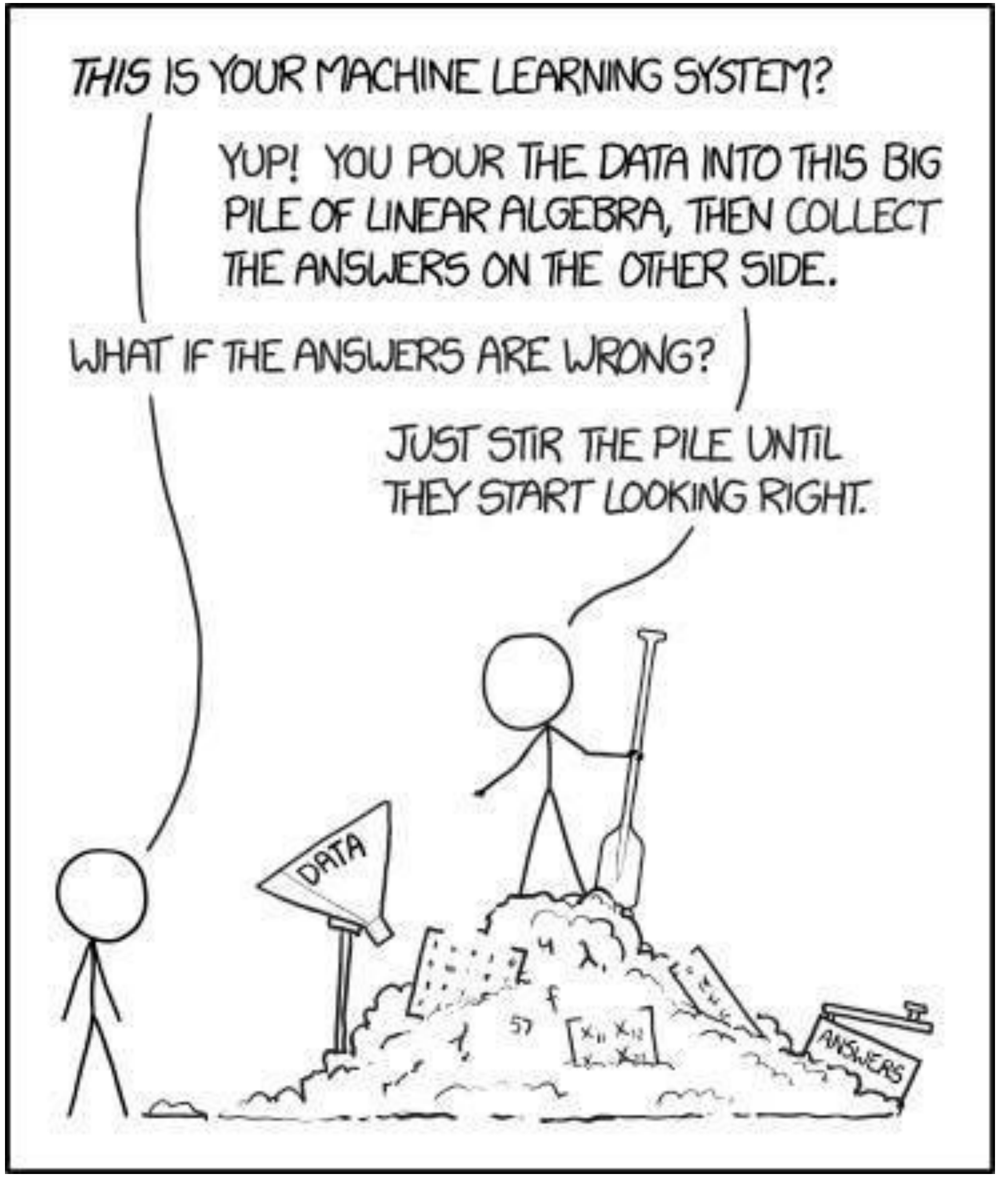
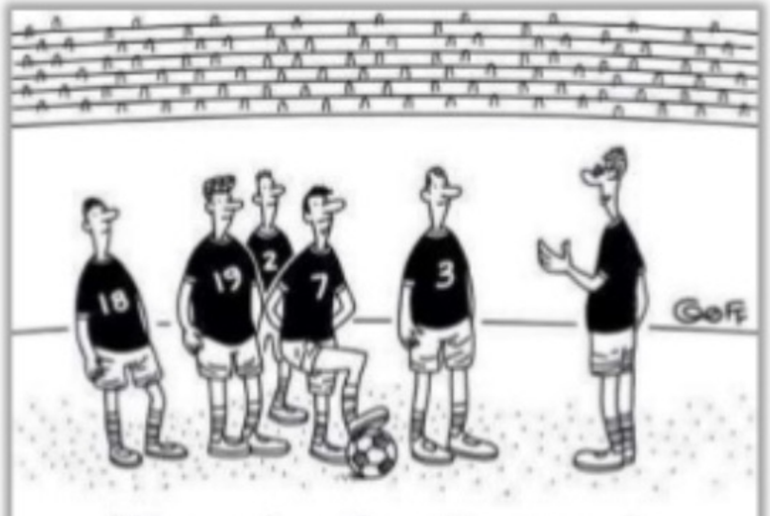
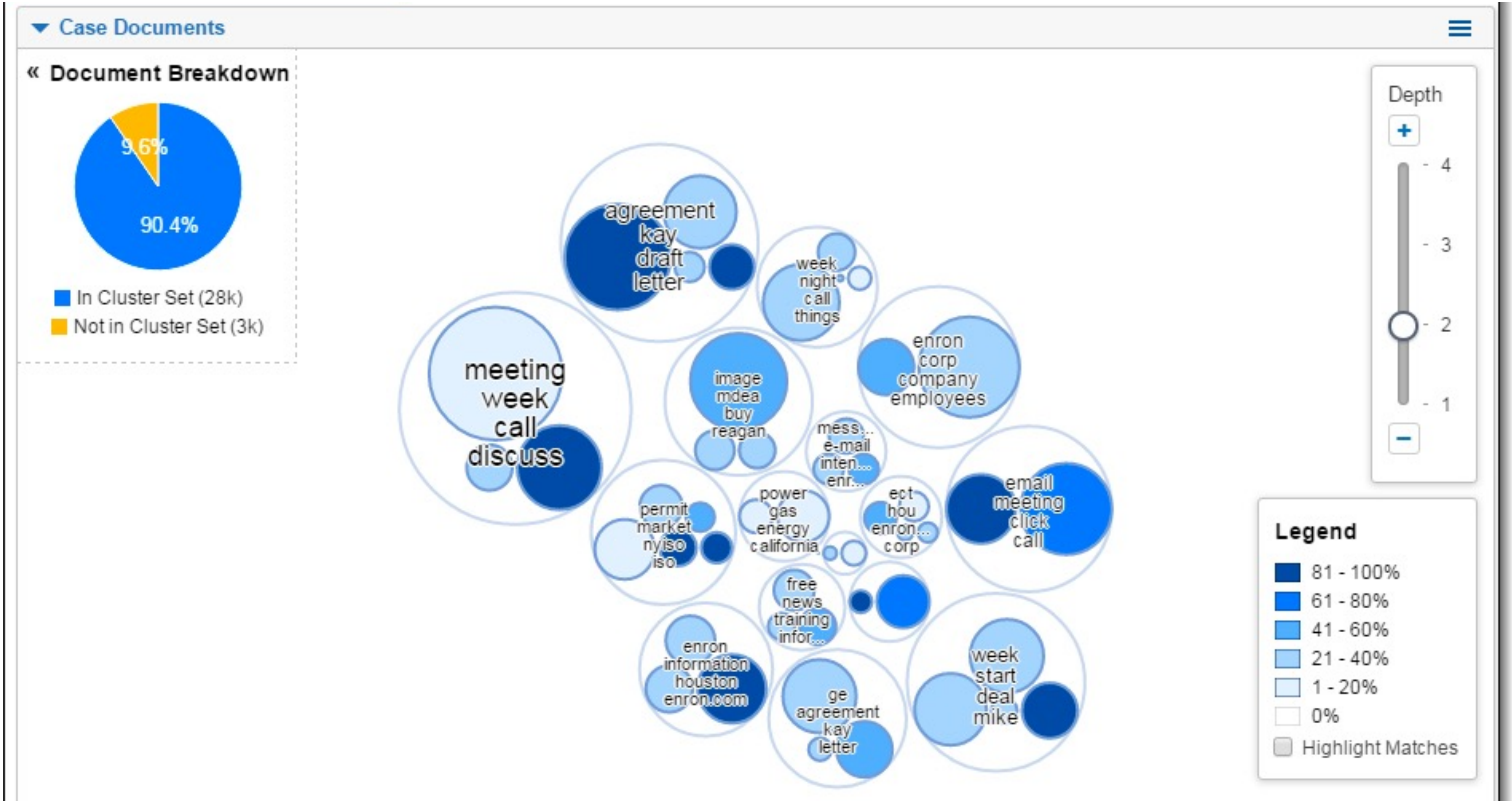




"Remember, the other team is counting on Big Data insights based on previous games. So, kick the ball with your other foot."

Hierarchical Clustering



Case Documents
Document Breakdown
9.6%
90.4%
In Cluster Set (28k)
Not in Cluster Set (3k)
agreement kay draft letter
week night call things
meeting week call discuss
image mdea buy reagan
enron corp company employees
mess... e-mail inten... enr...
permit market nyiso iso
power gas energy california
ect hou enron... corp
email meeting click call
free news training infor...
enron information houston enron.com
ge agreement kay letter
week start deal mike
Depth
4
3
2
1
Legend
81 - 100%
61 - 80%
41 - 60%
21 - 40%
1 - 20%
0%
Highlight Matches

Hierarchical Clustering

- Produces a set of nested clusters, organized as a hierarchical tree
- **Divisive**: Start with one all-inclusive cluster and at each step, split a cluster until only individual points remain
- **Agglomerative**: Start with points as individual clusters and at each step, merge the closest pair of clusters

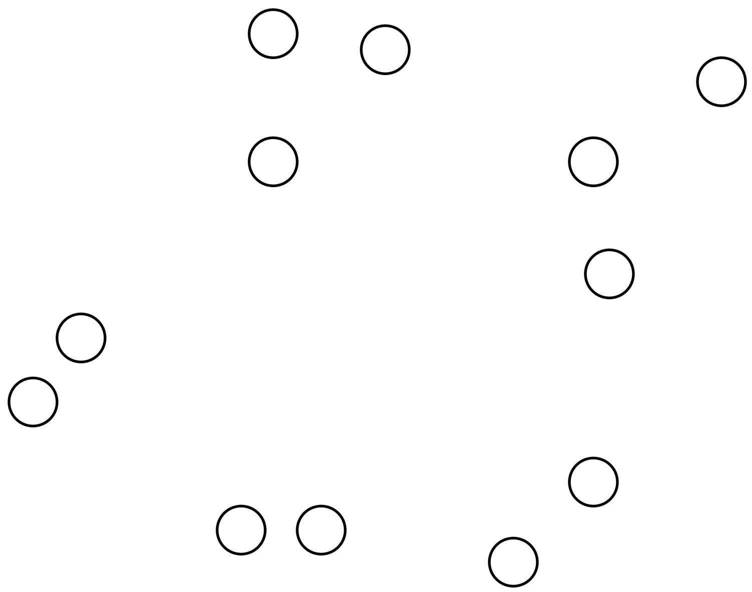
(often abbreviated as HAC: Hierarchical Agglomerative Clustering)

Agglomerative Clustering Algorithm

- Most popular hierarchical clustering technique
- 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

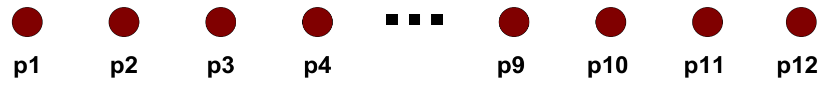
Starting Situation

- Start with clusters of individual points and a proximity matrix



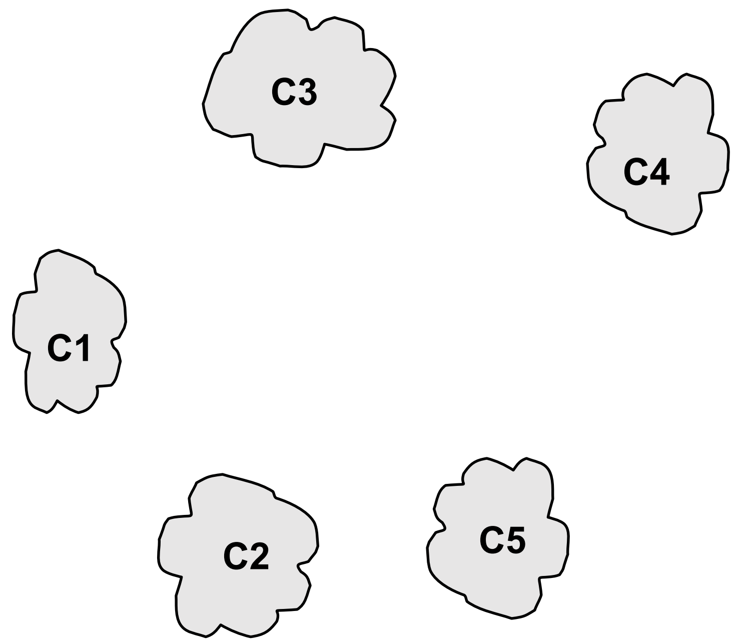
Proximity Matrix

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



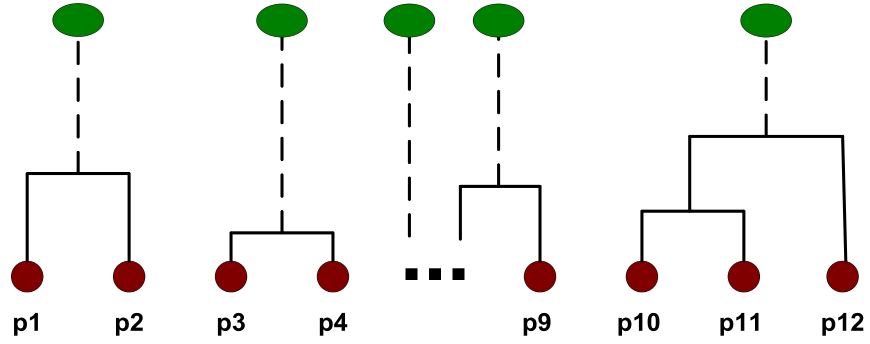
Intermediate Situation

- After some merging steps, we have some clusters



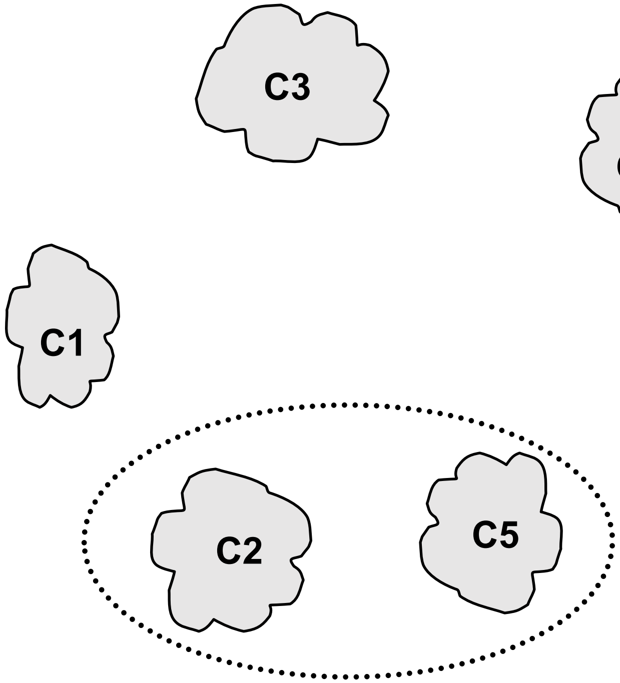
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

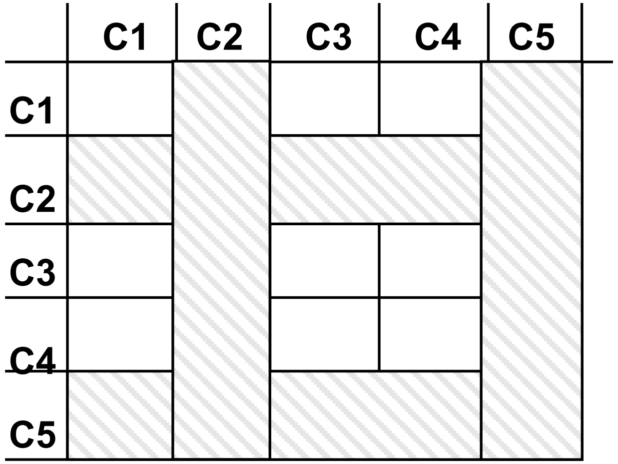
Proximity Matrix



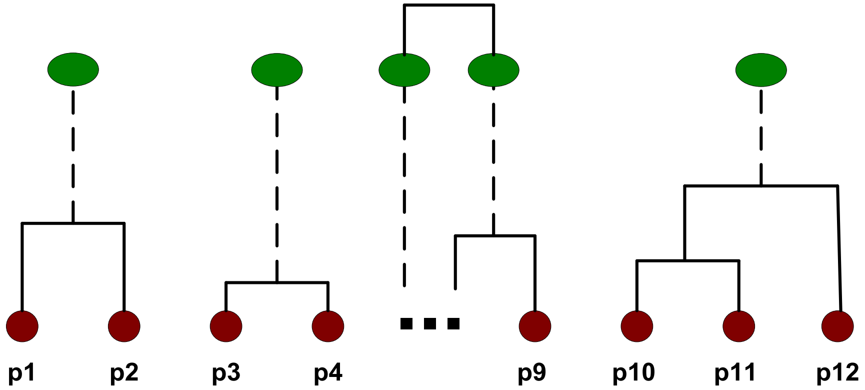
Intermediate Situation

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



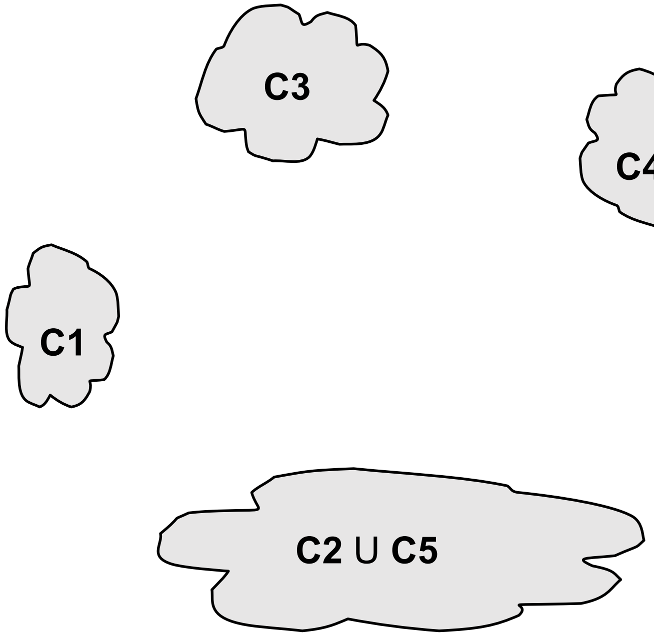


Proximity Matrix



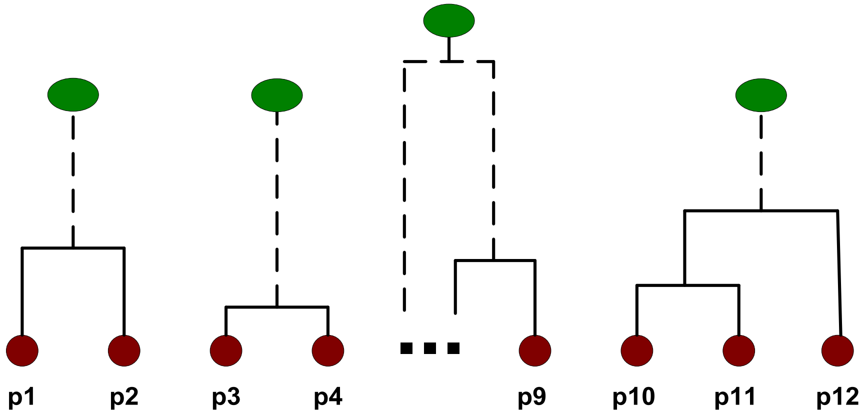
After Merging

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



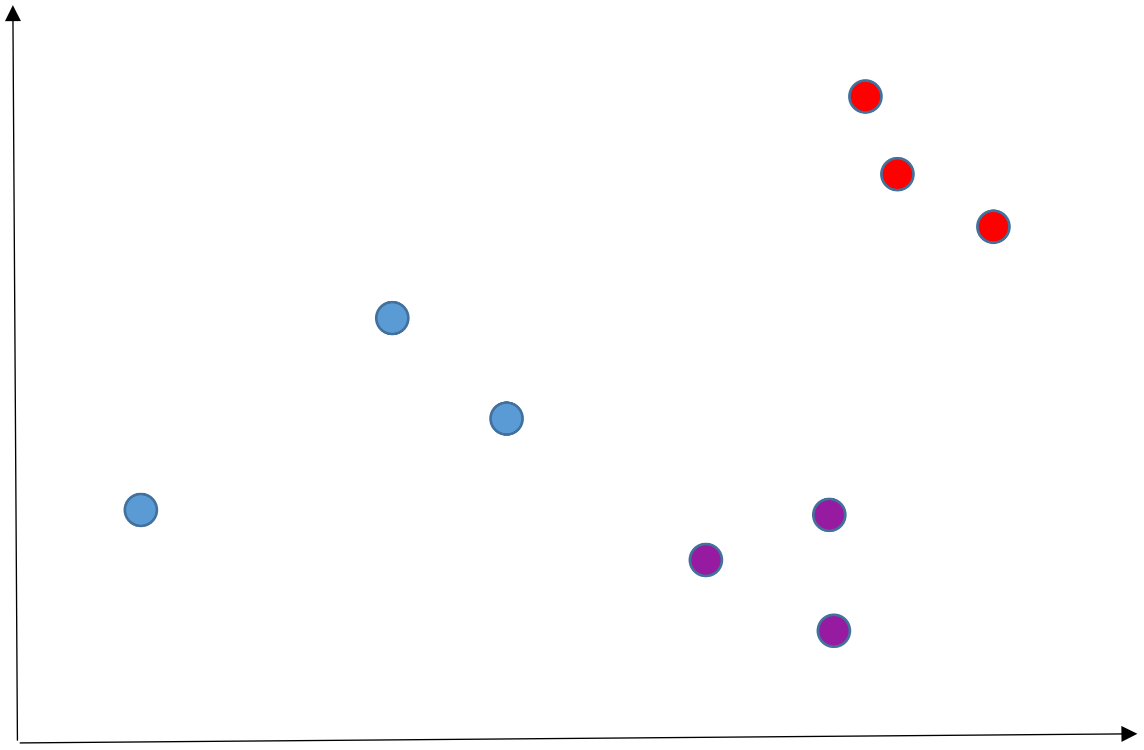
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



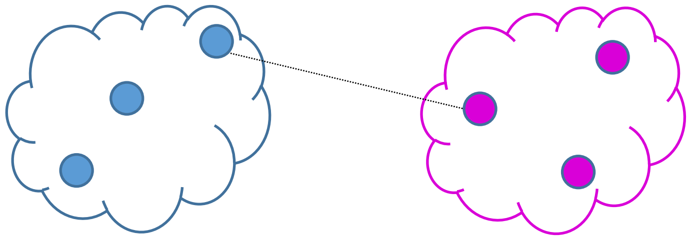
Defining Proximity

- The key operation in hierarchical agglomerative techniques
- Definition of proximity is what differentiates various algorithms

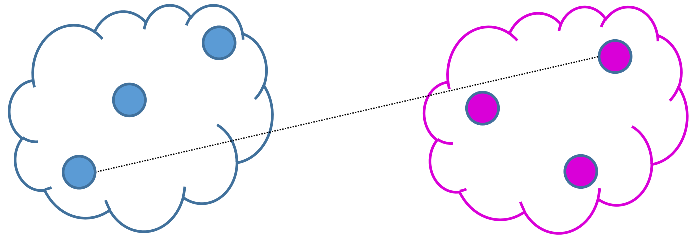


Definitions of proximity

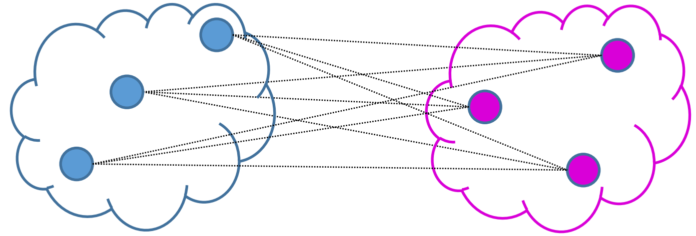
- **MIN/single link**: the proximity between the closest two points in different clusters



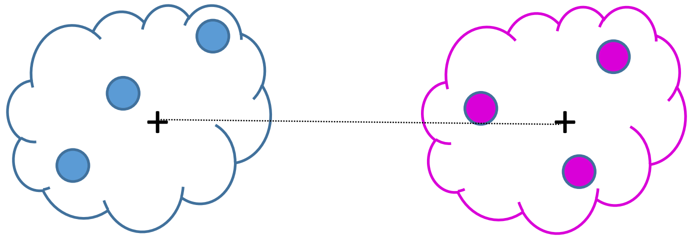
- **MAX/complete link**: the proximity between the farthest two points in different clusters



- **Group Average**: the average pairwise proximities of all pairs of points from different clusters



- **Centroid Method**: the proximity between cluster centroids



- **Ward's Method**: the increase in the SSE that results from merging the two clusters

Proximity Matrix

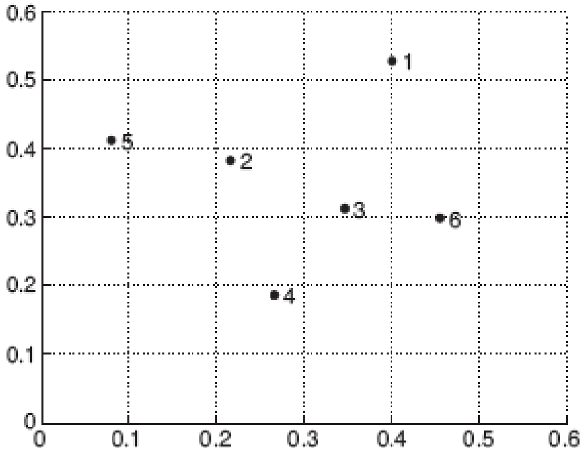


Figure 8.15. Set of 6 two-dimensional points.

Point	x Coordinate	y Coordinate
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

Table 8.3. xy coordinates of 6 points.

	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 8.4. Euclidean distance matrix for 6 points.

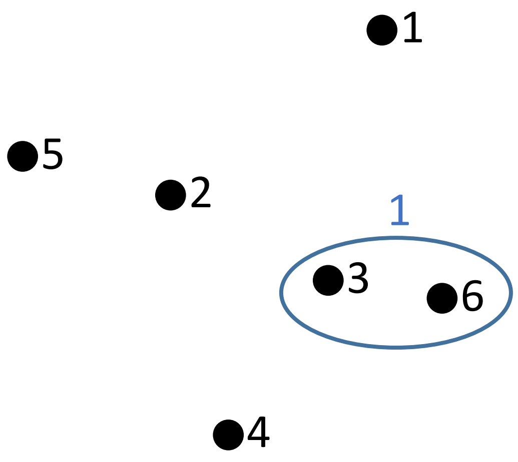
Hierarchical Agglomerative Clustering (HAC) using MIN or Single Link

Proximity matrix

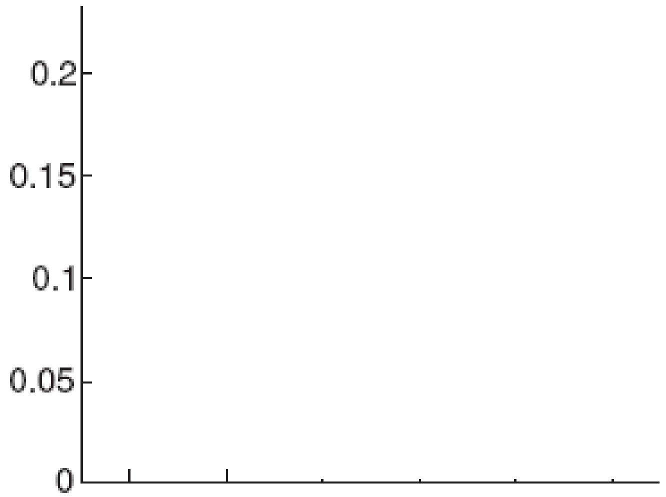
	p1	p2	p3	p4	p5	p6
p1	0	0.24	0.22	0.37	0.34	0.23
p2	0.24	0	0.15	0.20	0.14	0.25
p3	0.22	0.15	0	0.15	0.28	0.11
p4	0.37	0.20	0.15	0	0.29	0.22
p5	0.34	0.14	0.28	0.29	0	0.39
p6	0.23	0.25	0.11	0.22	0.39	0

	p1	p2	{p3,p6}	p4	p5
p1	0	0.24		0.37	0.34
p2	0.24	0		0.20	0.14
{p3,p6}			0		
p4	0.37	0.20		0	0.29
p5	0.34	0.14		0.29	0

Nested cluster diagram



Dendrogram



dist(p1, {p3,p6}) = min(dist(p1,p3), dist(p1,p6))

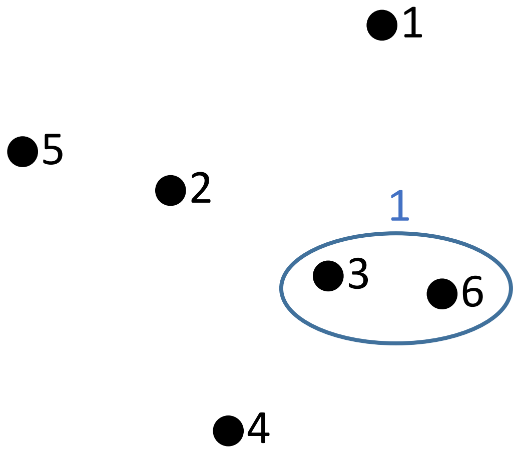
MIN or Single Link

Proximity matrix

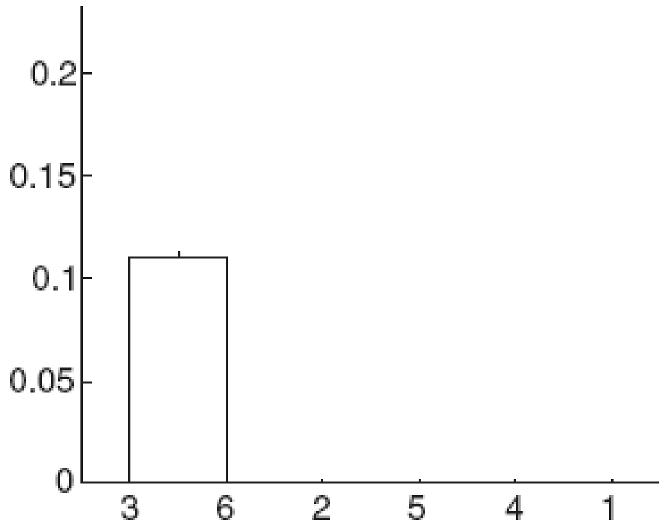
	p1	p2	p3	p4	p5	p6
p1	0	0.24	0.22	0.37	0.34	0.23
p2	0.24	0	0.15	0.20	0.14	0.25
p3	0.22	0.15	0	0.15	0.28	$\boxed{0.11}$
p4	0.37	0.20	0.15	0	0.29	0.22
p5	0.34	0.14	0.28	0.29	0	0.39
p6	0.23	0.25	0.11	0.22	0.39	0

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

Nested cluster diagram



Dendrogram



dist(p1, {p3,p6}) = min(dist(p1,p3), dist(p1,p6))

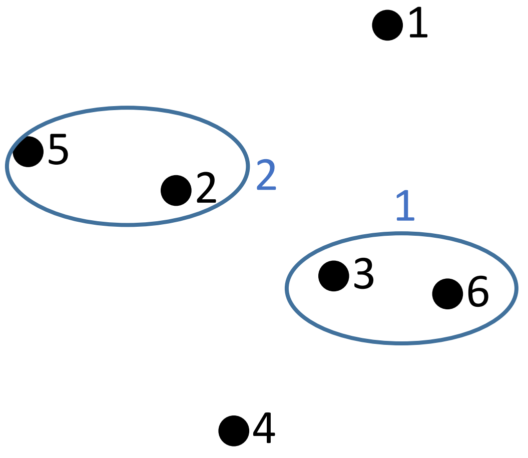
MIN or Single Link

Proximity matrix

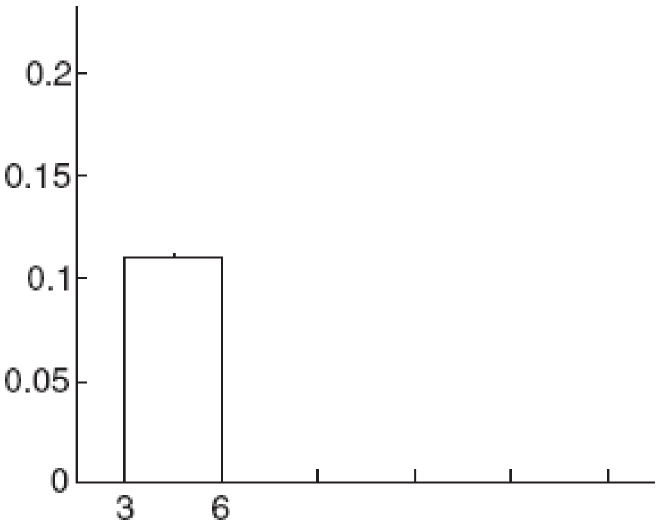
	p1	p2	{p3,p6}	p4	p5
p1	0	0.24	0.22	0.37	0.34
p2	0.24	0	0.15	0.20	$\boxed{0.14}$
{p3,p6}	0.22	0.15	0	0.15	0.28
p4	0.37	0.20	0.15	0	0.29
p5	0.34	0.14	0.28	0.29	0

	p1	**{p2,p5}**	{p3,p6}	p4
p1	0		0.22	0.37
{p2,p5}		0		
{p3,p6}	0.22		0	0.15
p4	0.37		0.15	0

Nested cluster diagram



Dendrogram



dist({p3,p6} , {p2,p5}) = min(dist({p3,p6}, p2), dist({p3,p6}, p5))

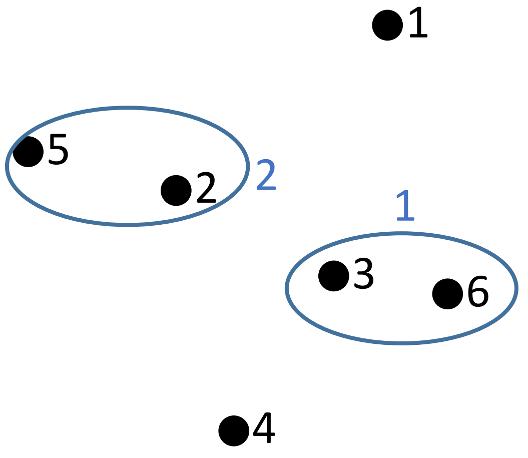
MIN or Single Link

Proximity matrix

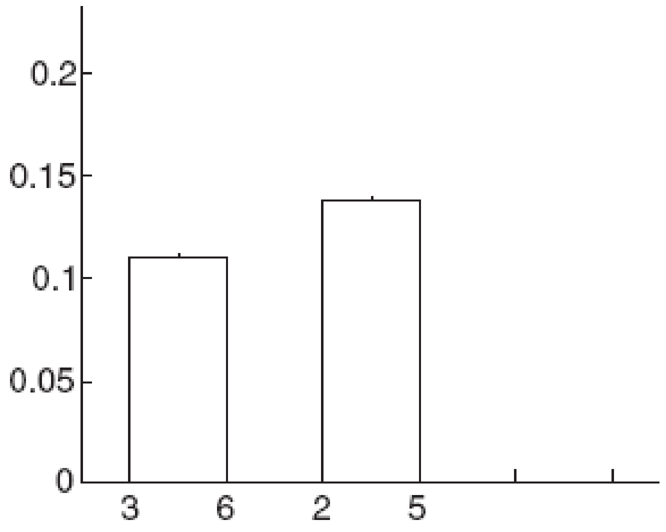
	p1	p2	{p3,p6}	p4	p5
p1	0	0.24	0.22	0.37	0.34
p2	0.24	0	0.15	0.20	$\boxed{0.14}$
{p3,p6}	0.22	0.15	0	0.15	0.28
p4	0.37	0.20	0.15	0	0.29
p5	0.34	0.14	0.28	0.29	0

	p1	**{p2,p5}**	{p3,p6}	p4
p1	0	**0.24**	0.22	0.37
{p2,p5}	**0.24**	**0**	**0.15**	**0.20**
{p3,p6}	0.22	**0.15**	0	0.15
p4	0.37	**0.20**	0.15	0

Nested cluster diagram



Dendrogram



dist({p3,p6} , {p2,p5}) = min(dist({p3,p6}, p2), dist({p3,p6}, p5))

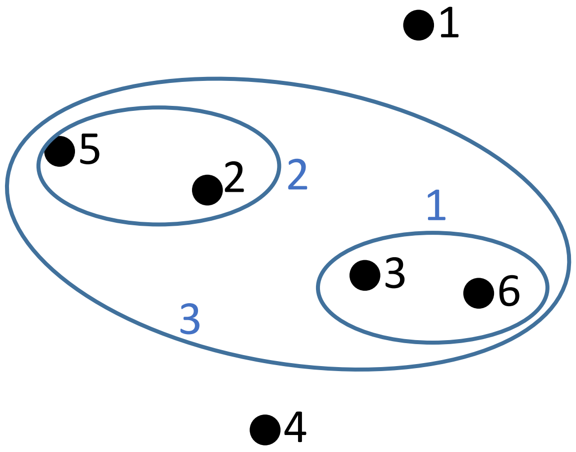
MIN or Single Link

Proximity matrix

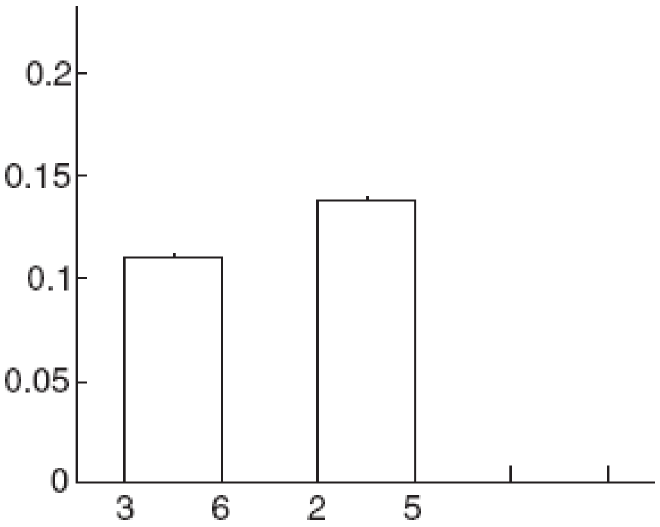
	p1	{p2,p5}	{p3,p6}	p4
p1	0	0.24	0.22	0.37
{p2,p5}	0.24	0	0.15	0.20
{p3,p6}	0.22	0.15	0	0.15
p4	0.37	0.20	0.15	0

	p1	**{p2,p3,p5,p6}**	p4
p1	0		0.37
{p2,p3,p5,p6}		0	
p4	0.37		0

Nested cluster diagram



Dendrogram



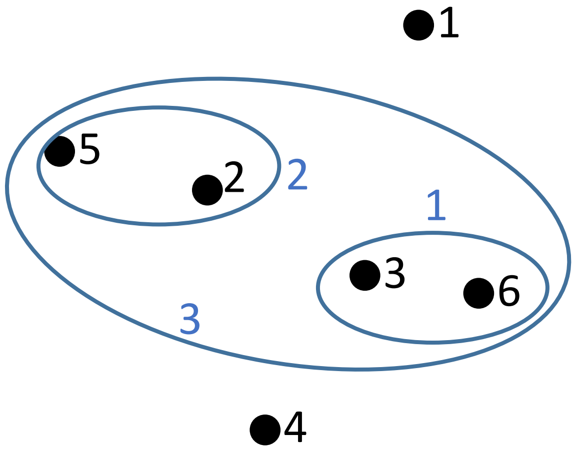
MIN or Single Link

Proximity matrix

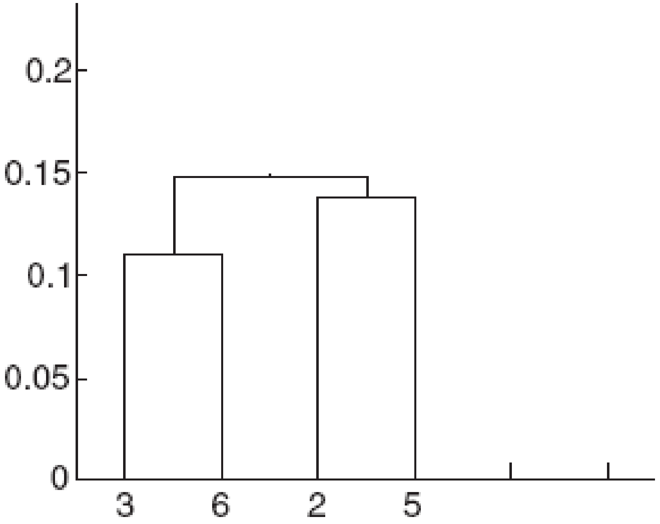
	p1	{p2,p5}	{p3,p6}	p4
p1	~~0~~	0.24	0.22	0.37
{p2,p5}	0.24	~~0~~	0.15	0.20
{p3,p6}	0.22	0.15	~~0~~	0.15
p4	0.37	0.20	0.15	~~0~~

	p1	**{p2,p3,p5,p6}**	p4
p1	0	**0.22**	0.37
{p2,p3,p5,p6}	**0.22**	**0**	**0.15**
p4	0.37	**0.15**	0

Nested cluster diagram



Dendrogram



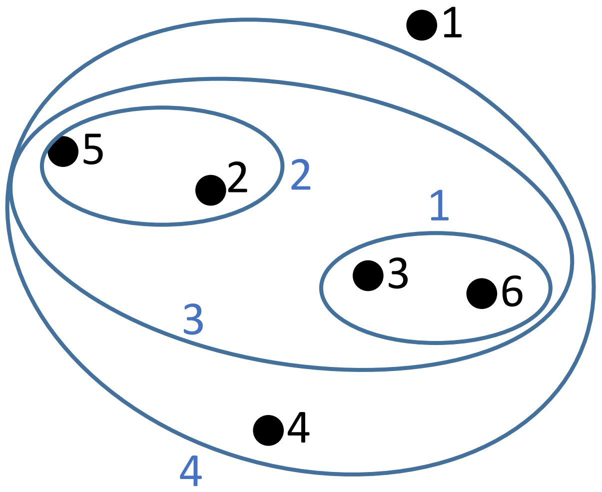
MIN or Single Link

Proximity matrix

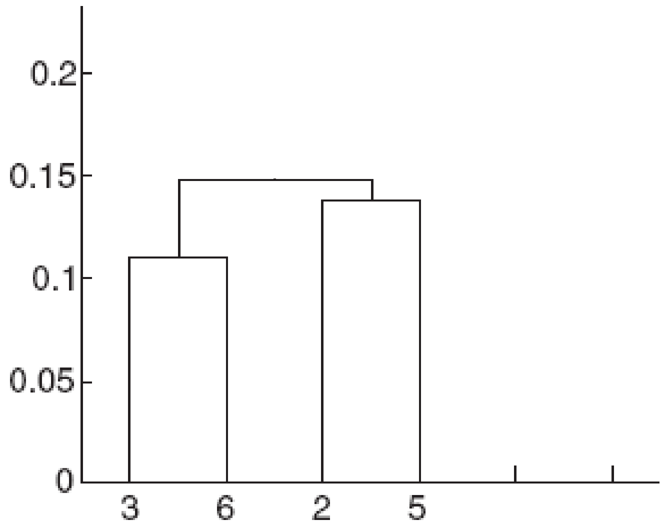
	p1	{p2,p3,p5,p6}	p4
p1	0	0.22	0.37
{p2,p3,p5,p6}	0.22	0	$\boxed{0.15}$
p4	0.37	0.15	0

	p1	**{p2,p3,p4,p5,p6}**
p1	0	**0.22**
{p2,p3,p4,p5,p6}	**0.22**	**0**

Nested cluster diagram



Dendrogram

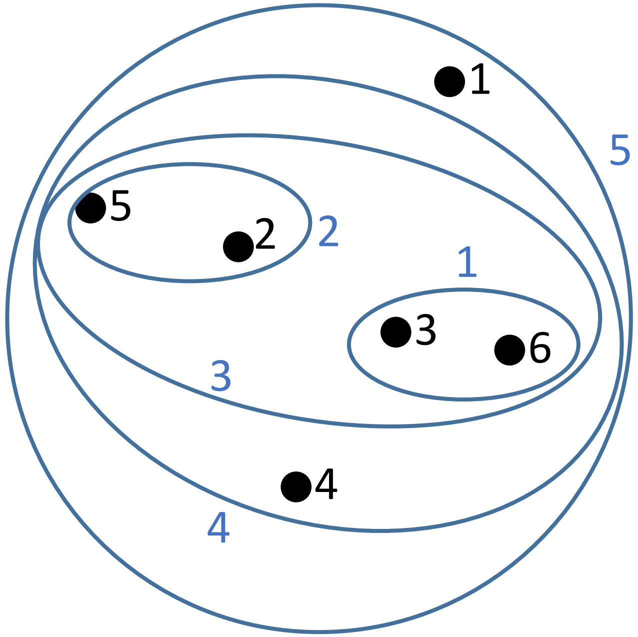


MIN or Single Link

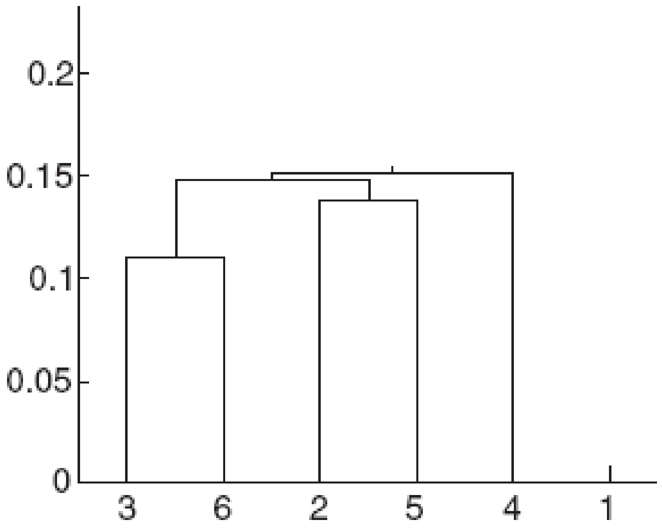
Proximity matrix

	p1	{p2,p3,p4,p5,p6}
p1	0	$\boxed{0.22}$
{p2,p3,p4,p5,p6}	0.22	0

Nested cluster diagram



Dendrogram



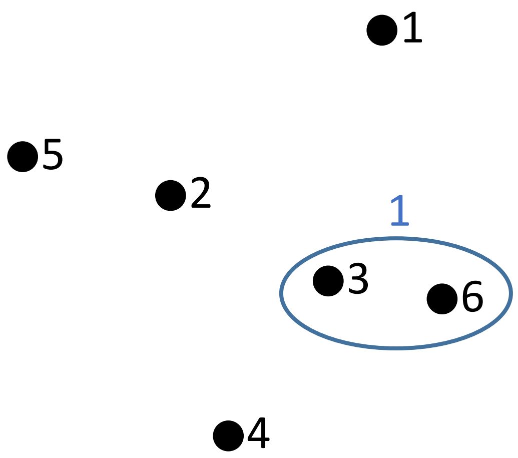
MAX or Complete Link

Proximity matrix

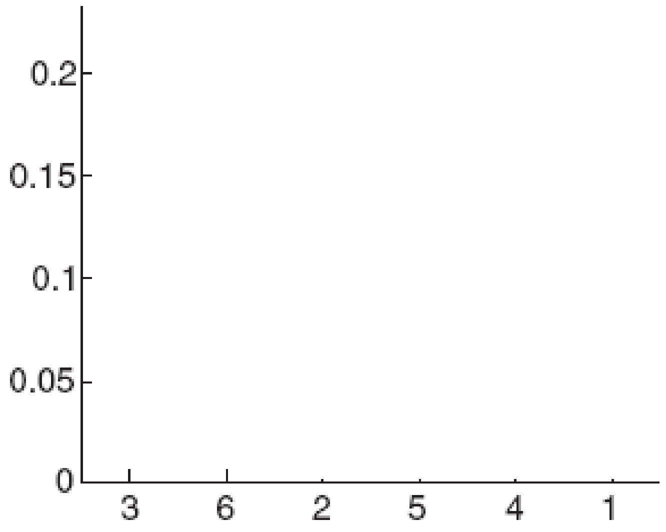
	p1	p2	p3	p4	p5	p6
p1	0	0.24	0.22	0.37	0.34	0.23
p2	0.24	0	0.15	0.20	0.14	0.25
p3	0.22	0.15	0	0.15	0.28	$\boxed{0.11}$
p4	0.37	0.20	0.15	0	0.29	0.22
p5	0.34	0.14	0.28	0.29	0	0.39
p6	0.23	0.25	0.11	0.22	0.39	0

	p1	p2	{p3,p6}	p4	p5
p1	0	0.24	**0.23**	0.37	0.34
p2	0.24	0	**0.25**	0.20	0.14
{p3,p6}	**0.23**	**0.25**	0	**0.22**	**0.39**
p4	0.37	0.20	**0.22**	0	0.29
p5	0.34	0.14	**0.39**	0.29	0

Nested cluster diagram

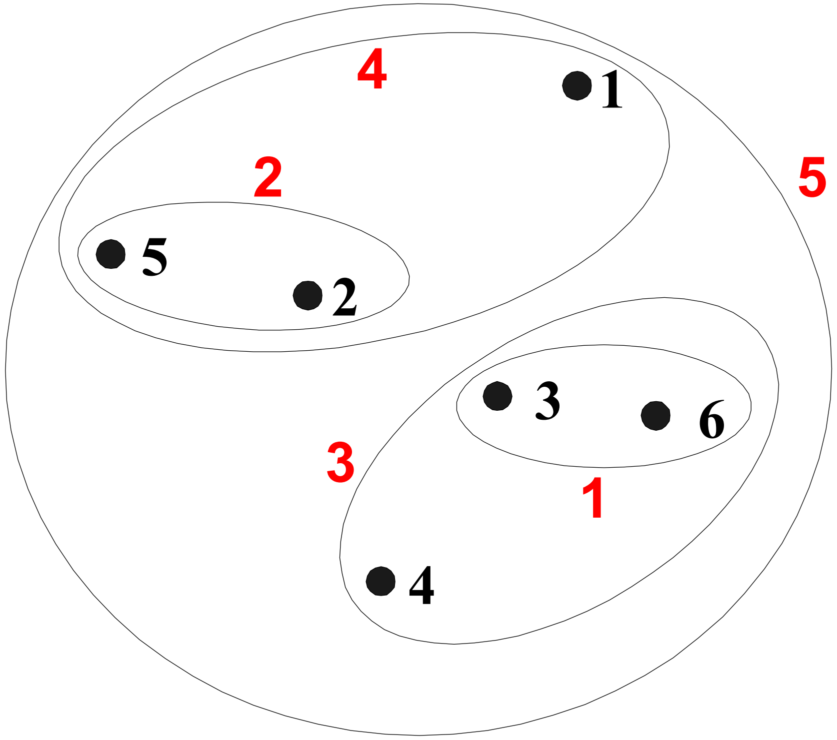


Dendrogram

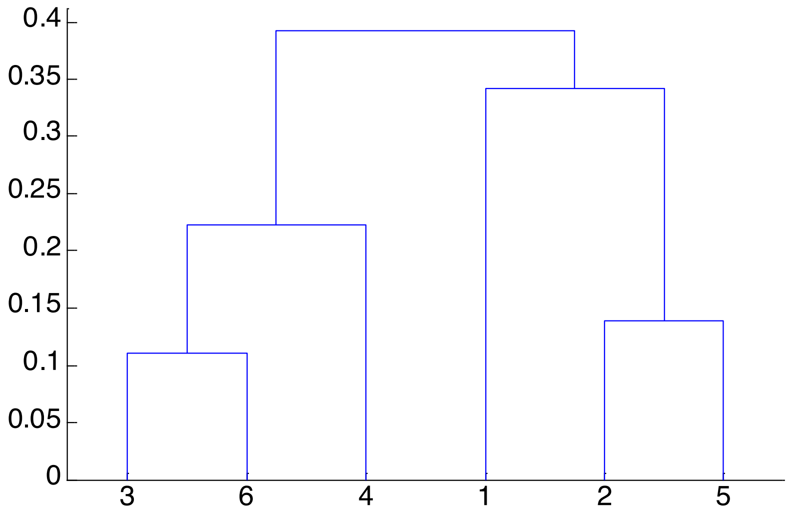


dist(p1, {p3,p6}) = ***max****(dist(p1,p3), dist(p1,p6))*

Max or Complete Link

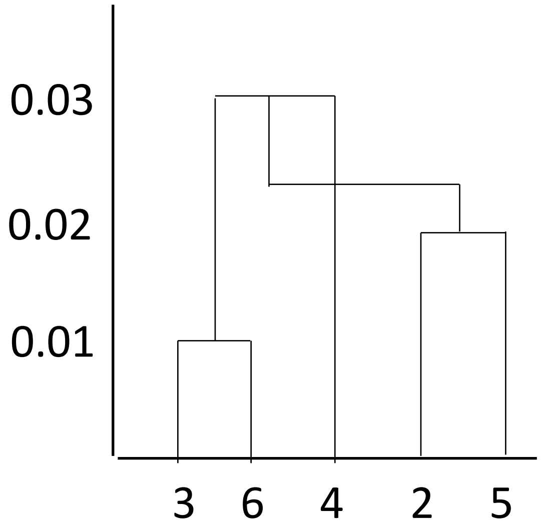


Nested Clusters

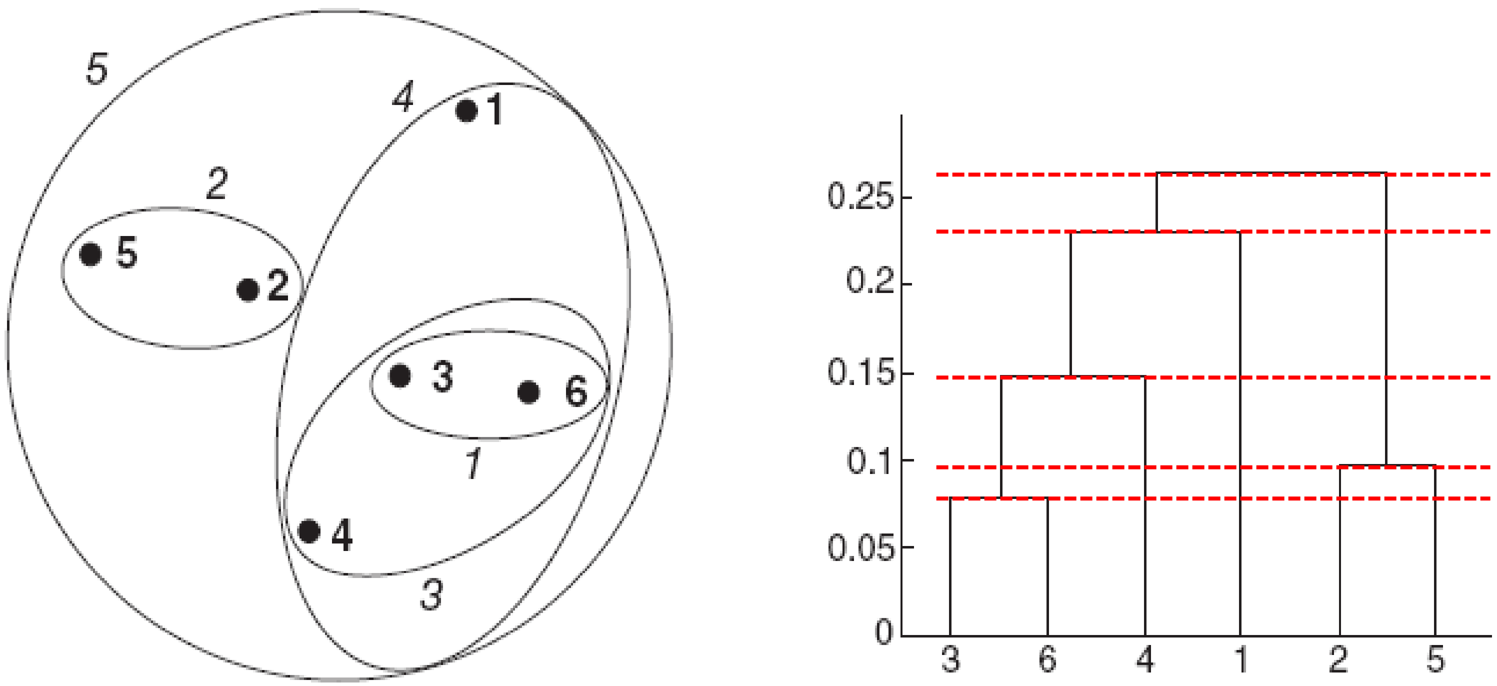


Dendrogram

Note: Centroid methods have a characteristic that other methods don’t have… they can have **inversions.** Cluster merges may happen at a closer distance than previous cluster merges.



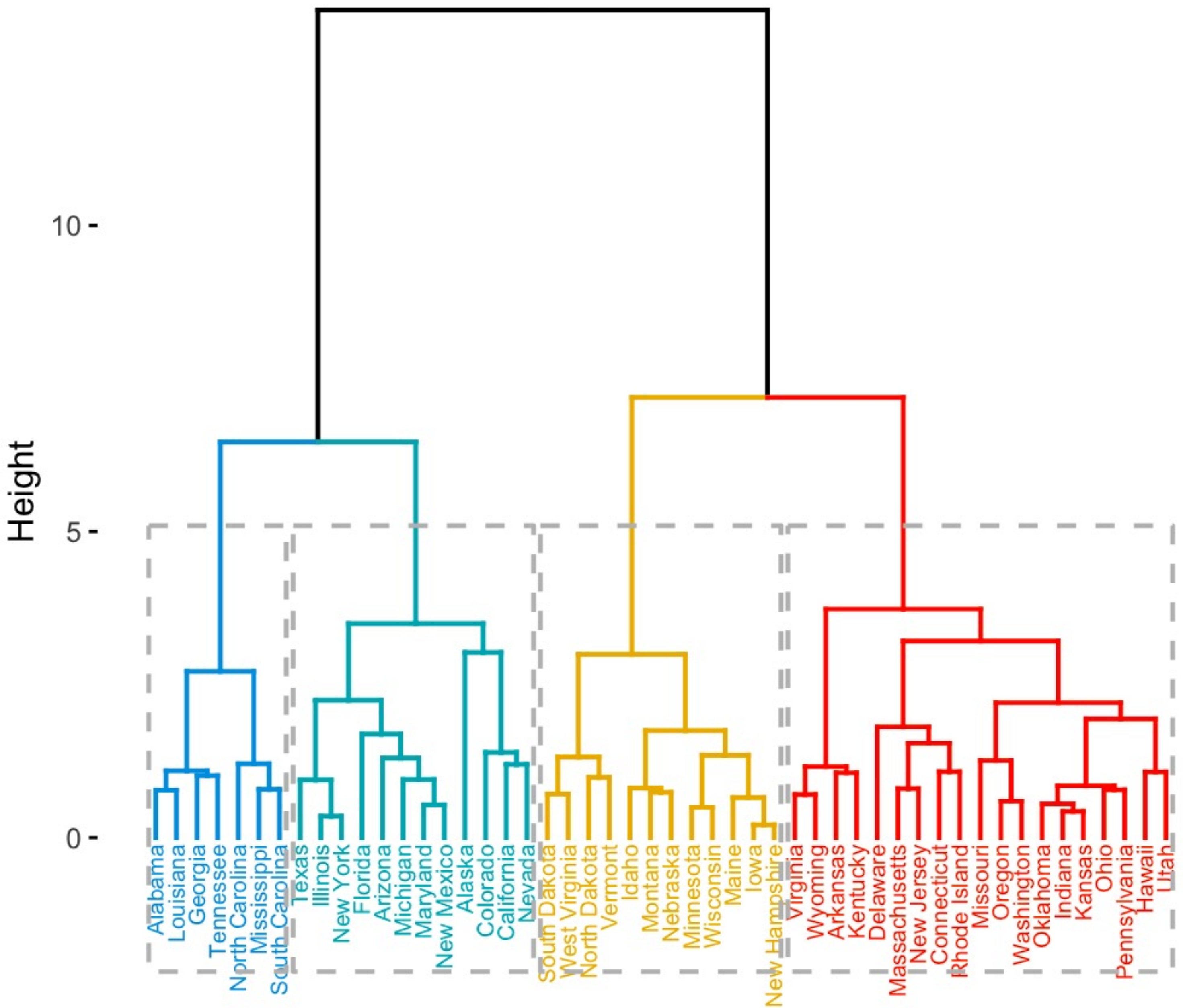
Using Dendrogram to Determine K

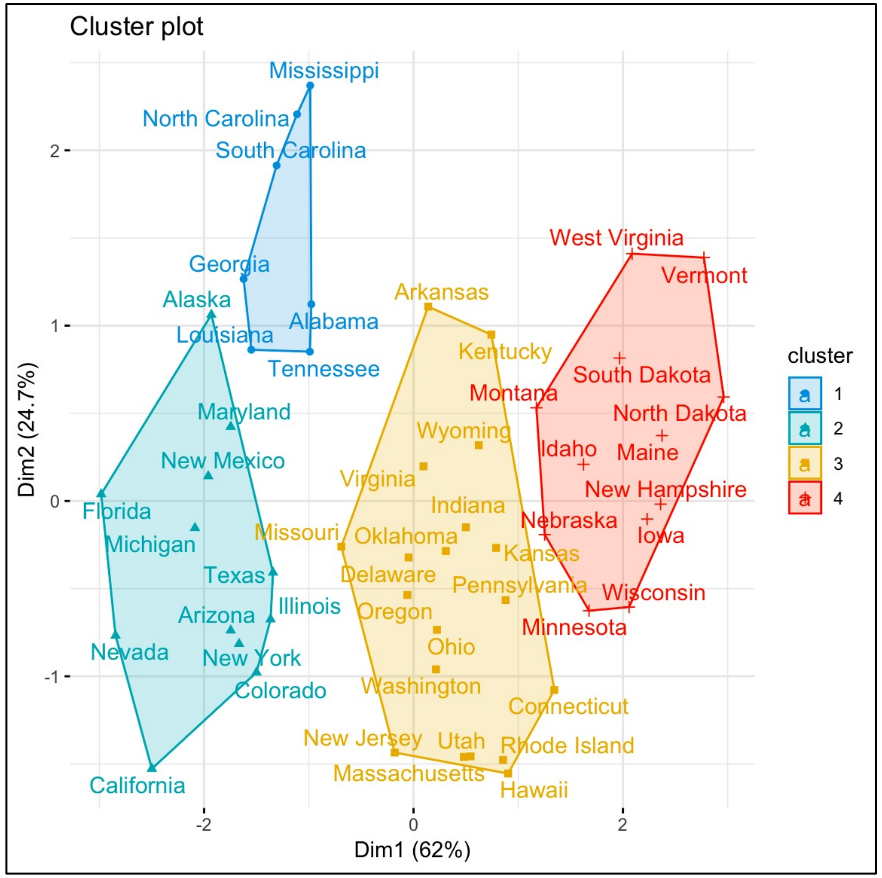


(a) Ward's clustering. (b) Ward's dendrogram.

Figure 8.19. Ward's clustering of the six points shown in Figure 8.15.

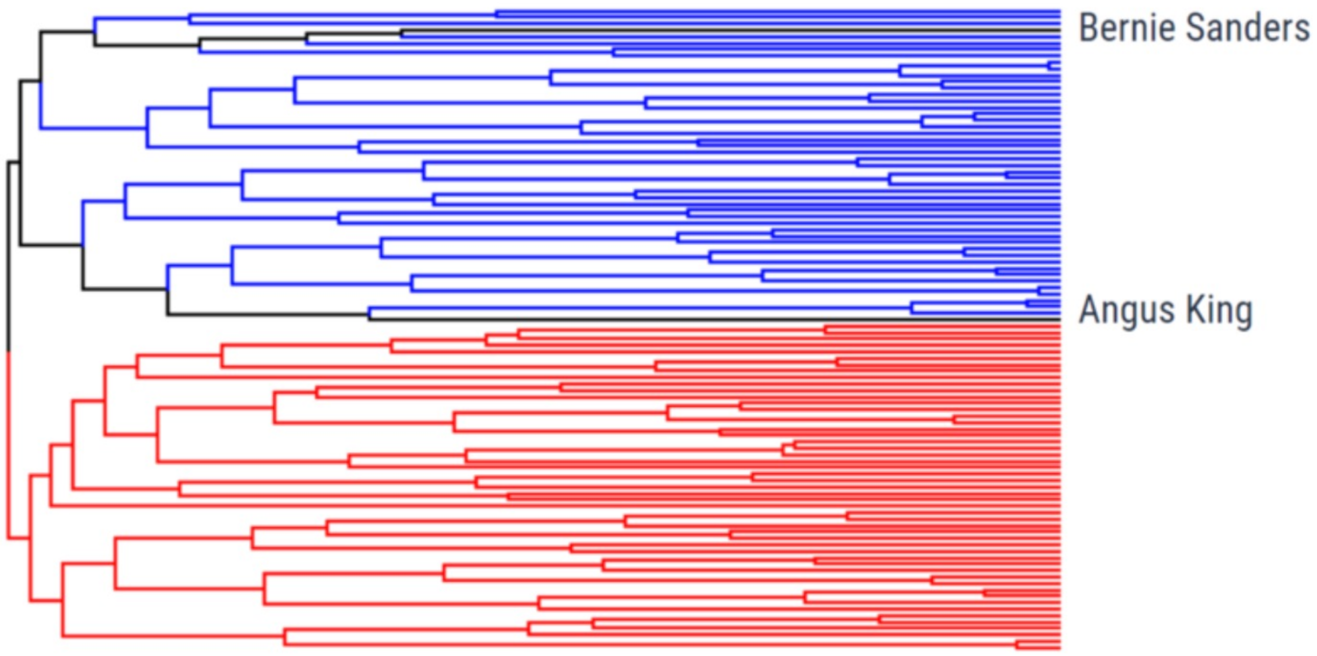
Real World Example





Real World Example

- **US Senator Clustering through Twitter (2018)**
- *Can we find the party lines through Twitter?*
- Our data is simple: we look at which senators follow which senators.
- We use the Walktrap algorithm, which does a random walk through the graph, and estimates the senator similarity by the number of times you end up at a certain senator starting from a different certain senator.
- After getting these similarities, we can use agglomerative clustering to find the dendrogram.



Reds are Republicans, Blues are Democrats, Blacks are independent

Characteristics of Hierarchical Clustering

- Good for data that has an underlying hierarchy
- Expensive in both time and space
- Difficult to choose the “best” proximity measure
- HAC with Ward’s method is often used in conjunction with K-means

“First, perform hierarchical clustering on manageable sample. Visualize & analyze the trees of clusters being formed (the dendrogram) and then use this evaluation to guide how many & what kind of clusters there are in the dataset. In some cases, this can be used directly to initialize k-means on all data in final step.

Or you could do the reverse as well. First create 50–100 micro-clusters using k-means on all data. Then, on the centers of these micro-clusters, perform Hierarchical clustering while interpreting the tree of clusters being formed.”

https://www.quora.com/What-are-the-pros-and-cons-of-k-means-vs-hierarchical-clustering

Practice Problem

You always group the two closest clusters.
The difference is in updating the proximity matrix.