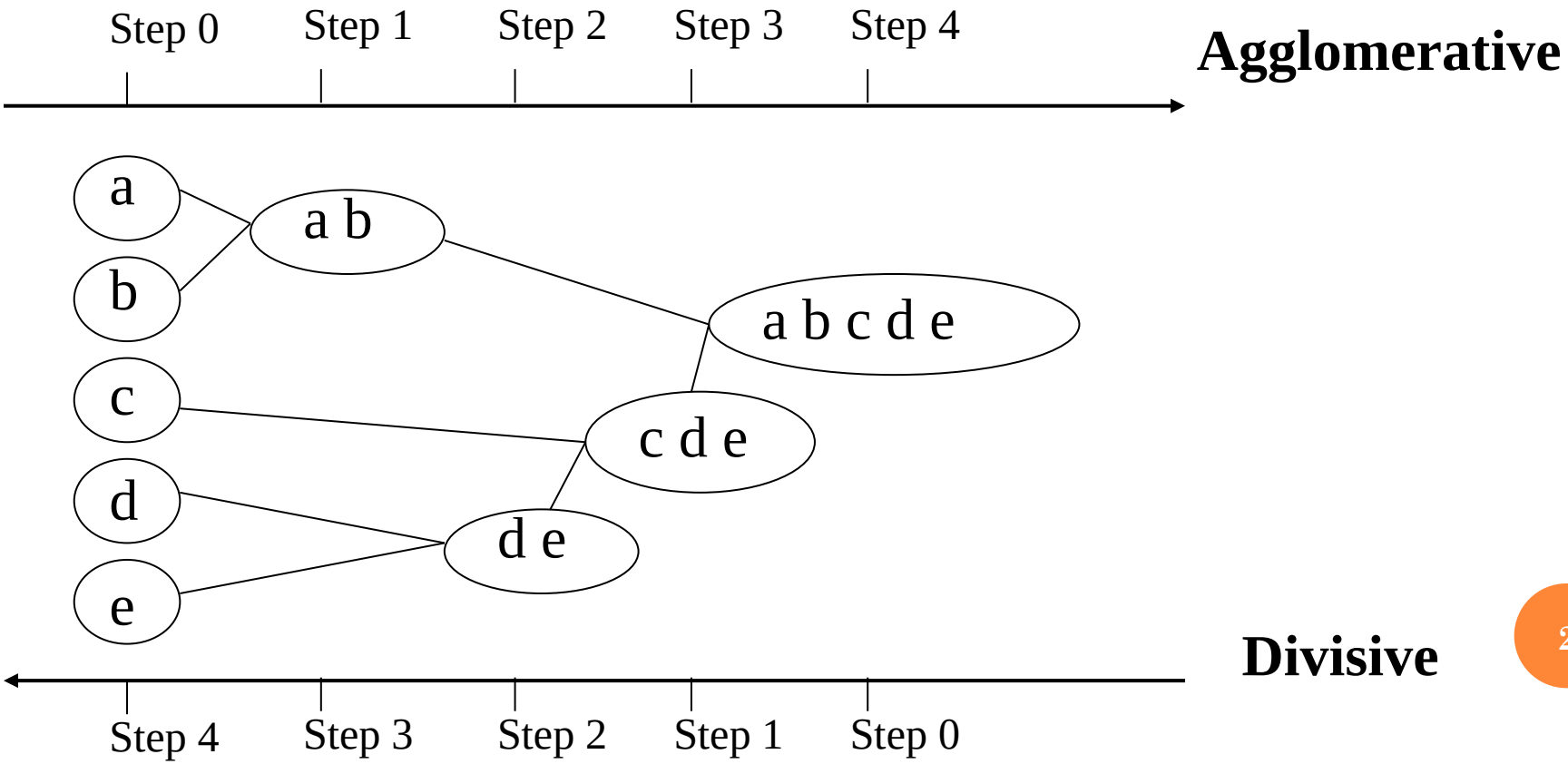
Hierarchical Clustering

Hierachical clustering

- Agglomerative and divisive clustering on the data set {a, b, c, d ,e }

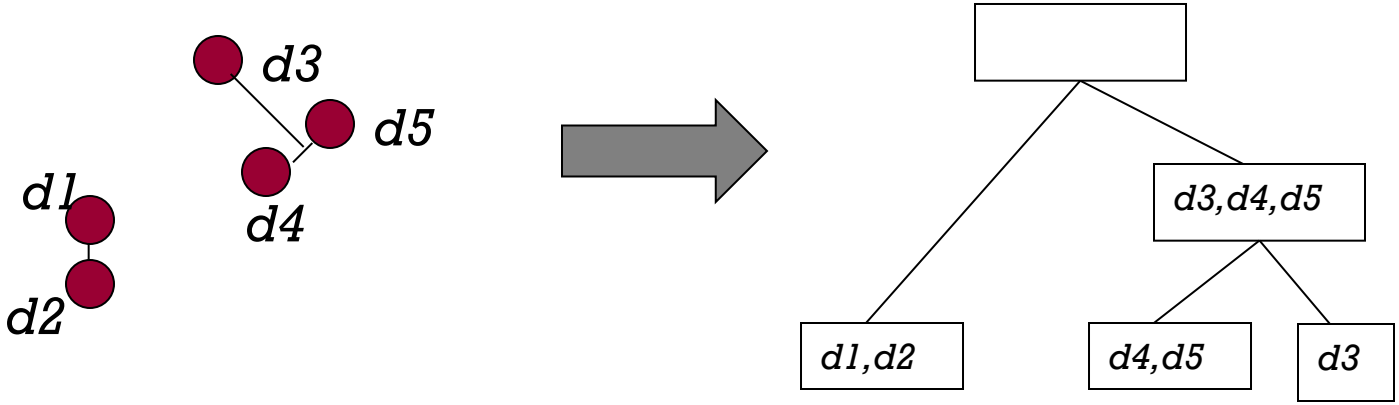


Agglomerative Clustering

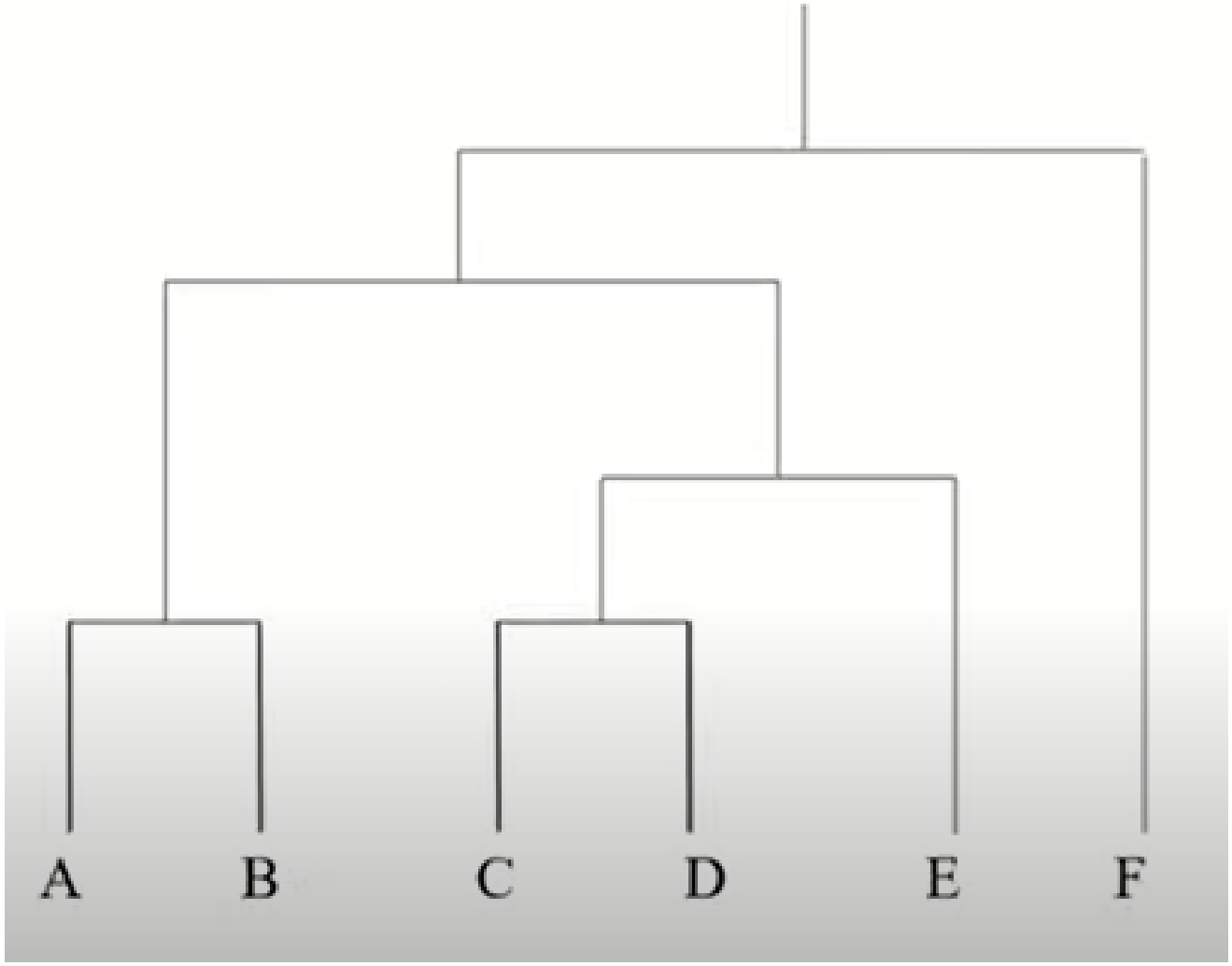
- The **agglomerative clustering** is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.
- The algorithm starts by treating each object as a singleton cluster.
- Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.
- The result is a tree-based representation of the objects, named *dendrogram*

Agglomerative clustering

1. Convert object attributes to distance matrix
2. Set each object as a cluster (thus if we have *N* objects, we will have *N* clusters at the beginning)
3. Repeat until number of cluster is one (or known # of clusters)
    a. Merge two closest clusters
    b. Update distance matrix

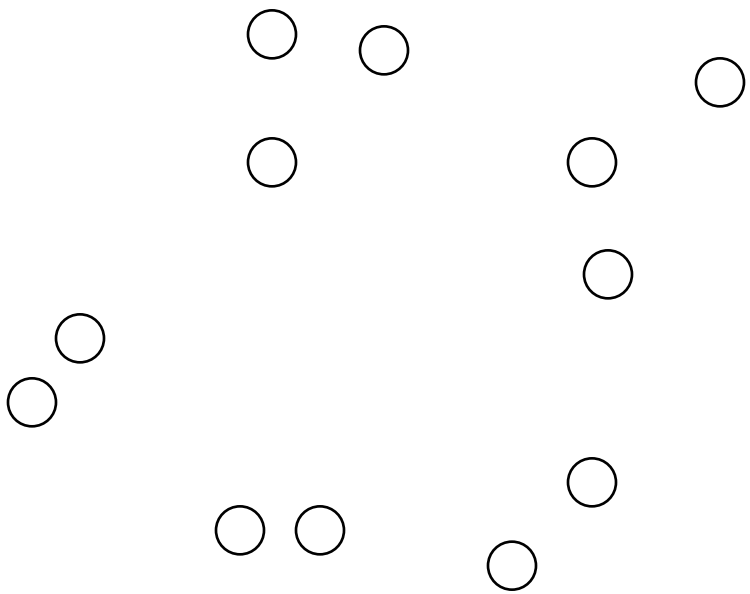


Dendogram



Starting Situation

- Start with clusters of individual points and a distance/proximity matrix



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

Distance Matrix

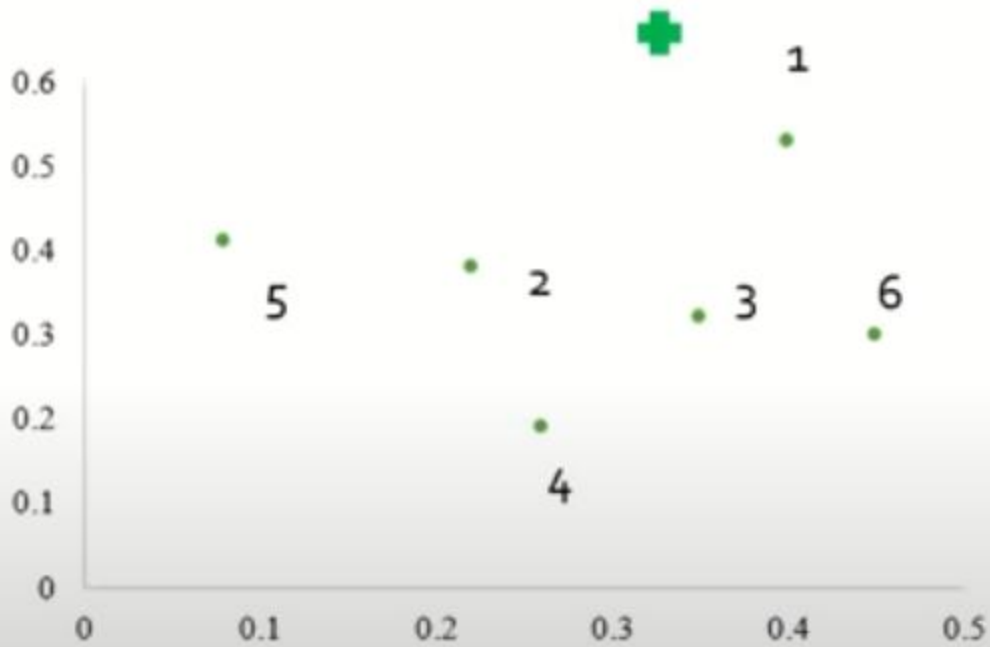
EXAMPLE – SINGLE LINK

- Find the clusters using single link technique. Use Euclidean distance, and draw the dendrogram.

	X	Y
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

EXAMPLE – SINGLE LINK

	X	Y
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



EXAMPLE – SINGLE LINK

- Calculate Euclidean distance, create the distance matrix.

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

Distance (P1,P2) $= \sqrt{(0.40-0.22)^2+(0.53-0.38)^2}$

(0.40,0.53), (0.22,0.38) $= \sqrt{(0.18)^2+(0.15)^2}$

$= \sqrt{0.0324+0.0225}$

$= \sqrt{0.0549}$

$= 0.23$

EXAMPLE – SINGLE LINK

Sample No.	X	Y
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

$$d(p_1, p_2) = \sqrt{(0.22 - 0.40)^2 + (0.38 - 0.53)^2}$$

$$= 0.23$$

$$d(p_1, p_3) = \sqrt{(0.35 - 0.40)^2 + (0.32 - 0.53)^2}$$

$$= 0.22$$

$$d(p_2, p_3) = \sqrt{(0.35 - 0.22)^2 + (0.32 - 0.38)^2}$$

$$= 0.14$$

EXAMPLE – SINGLE LINK

- The distance matrix is

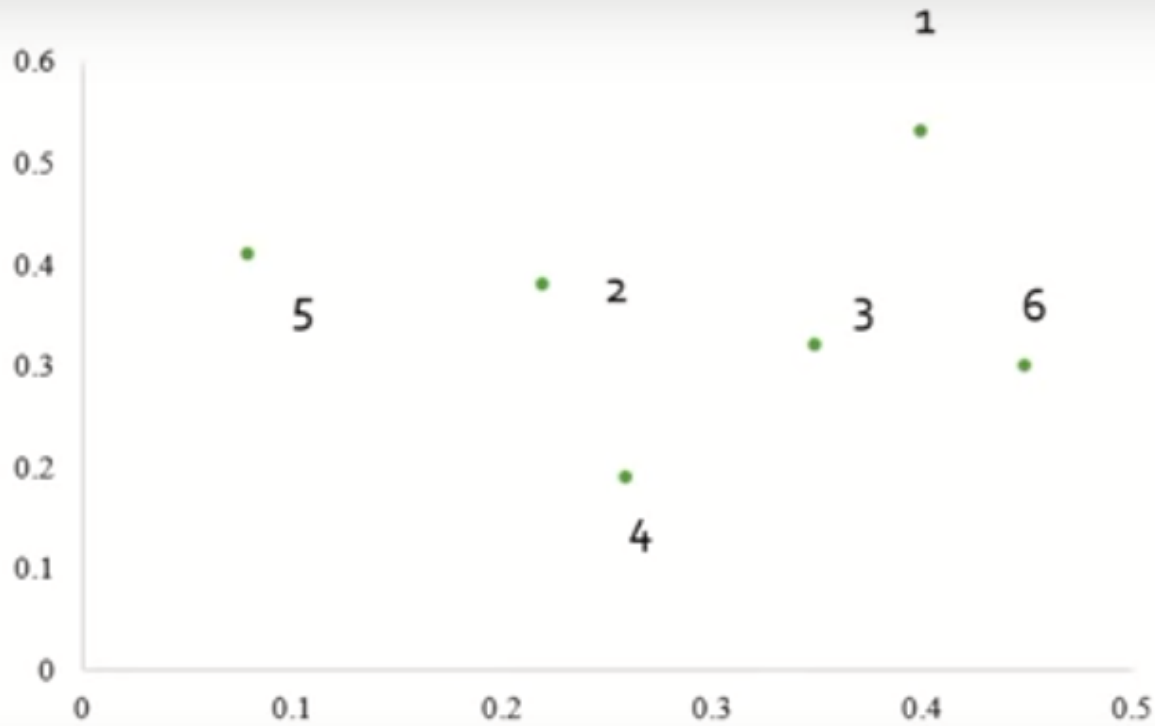
	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

EXAMPLE – SINGLE LINK

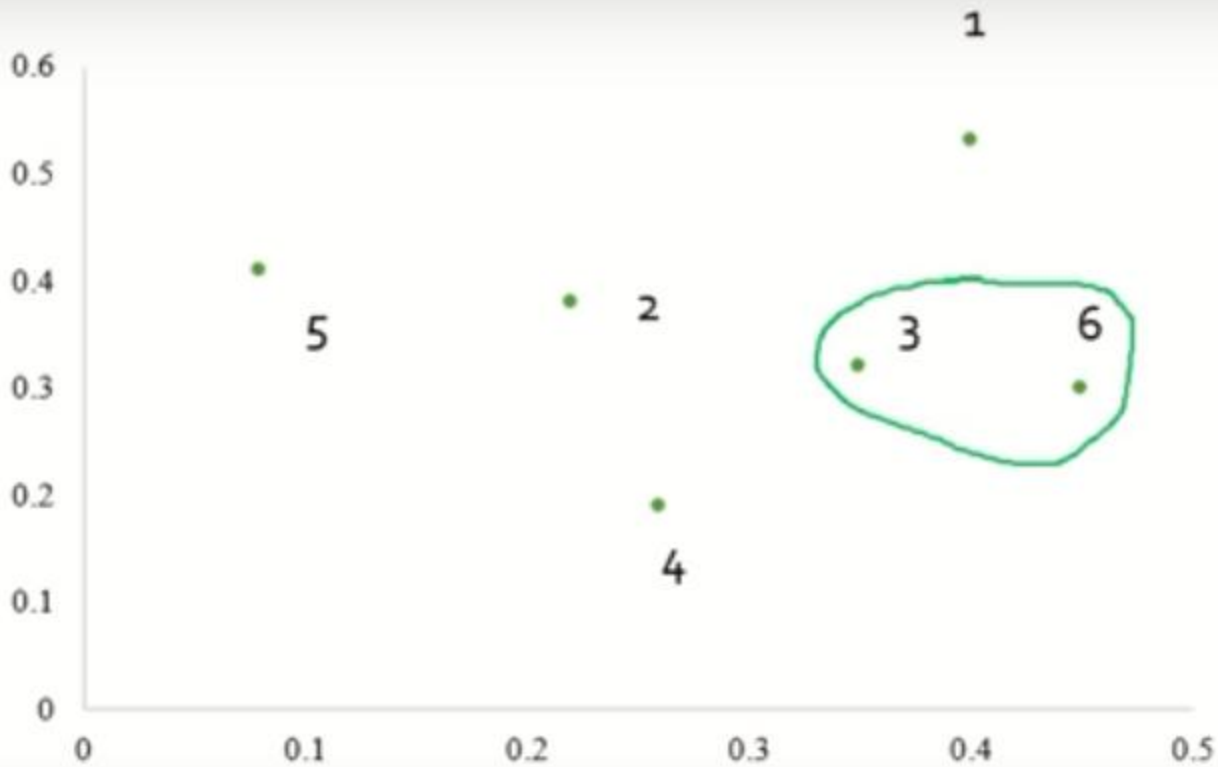
- The distance matrix is

	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

EXAMPLE – SINGLE LINK



EXAMPLE – SINGLE LINK



EXAMPLE – SINGLE LINK

- The distance matrix is

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	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

EXAMPLE – SINGLE LINK

- To update the distance matrix MIN[dist(P3,P6),P1)]
- MIN(dist(P3,P1), (P6,P1))

  = min[(0.22,0.23)]

  = 0.22
- To update the distance matrix MIN[dist(P3,P6),P2)]
- MIN(dist(P3,P2), (P6,P2))

  = min[(0.15,0.25)]

  = 0.15

EXAMPLE – SINGLE LINK

- To update the distance matrix MIN[dist(P3,P6),P4)]
- MIN(dist(P3,P4), (P6,P4))

  = min[(0.15,0.22)]

  = 0.15
- To update the distance matrix MIN[dist(P3,P6),P5)]
- MIN(dist(P3,P5), (P6,P5))

  = min[(0.28,0.39)]

  = 0.28

EXAMPLE – SINGLE LINK

- The updated distance matrix for cluster P3, P6

	P1	P2	P3,P6	P4	P5
P1	0				
P2	0.23	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

EXAMPLE – SINGLE LINK

- The distance matrix is

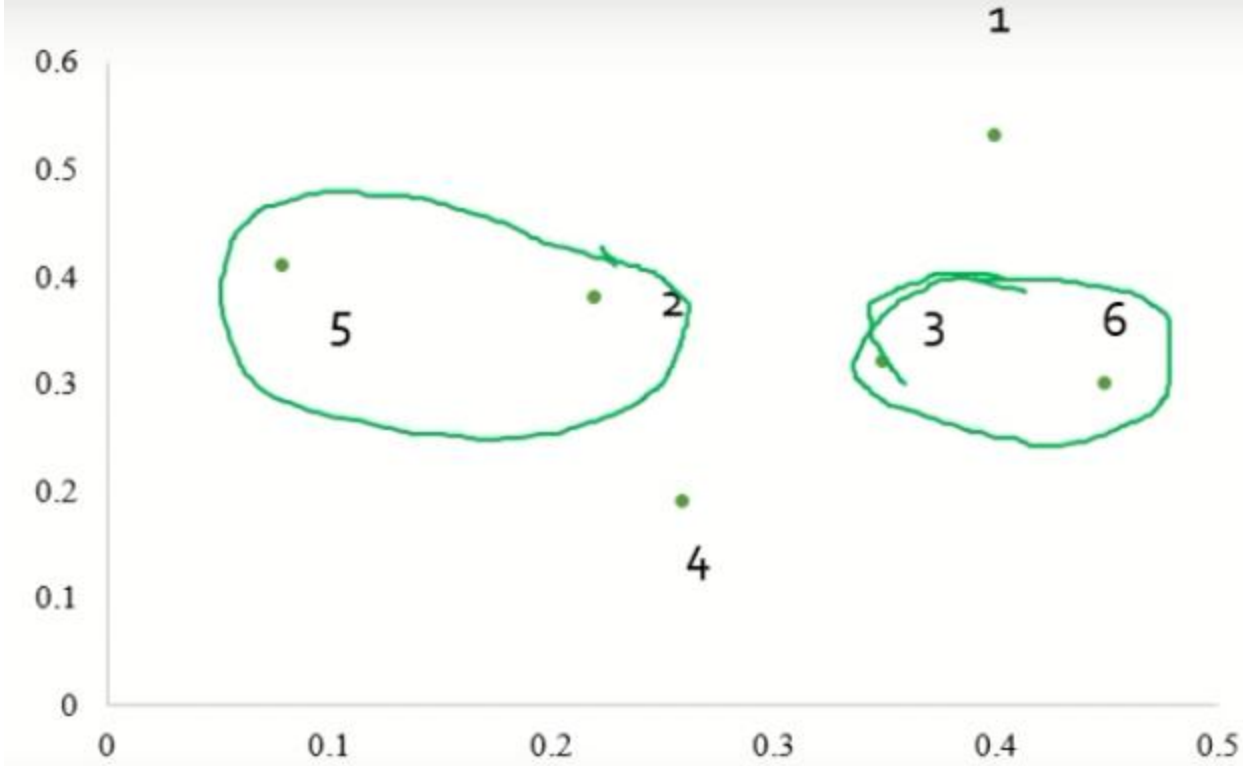
	P1	P2	P3,P6	P4	P5
P1	0				
P2	0.23	0			
P3,P6	0.22	0.15	0		
P4	0.37	0.20	0.15	0	
P5	0.34	0.14	0.28	0.29	0

EXAMPLE – SINGLE LINK

- The distance matrix fro cluster P2, P5

	P1	P2	P3,P6	P4	P5
P1	0				
P2	0.23	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

EXAMPLE – SINGLE LINK



EXAMPLE – SINGLE LINK

- To update the distance matrix MIN[dist(P2,P5),P1)]
- MIN[dist(P2,P1), (P5,P1)]

  = min[(0.23,0.34)]

  = 0.23
- To update the distance matrix MIN[dist(P2,P5),(P3,P6)]
- MIN[dist(P2,(P3,P6)), (P5,(P3,P6))]

  = min[(0.15,0.28)]

  = 0.15

EXAMPLE – SINGLE LINK

- To update the distance matrix MIN[dist(P2,P5),P4)]
- MIN[dist(P2,P4), (P5,P4)]

  = min[(0.20,0.29)]

  = 0.20

EXAMPLE – SINGLE LINK

- The updated distance matrix for cluster P2,P5

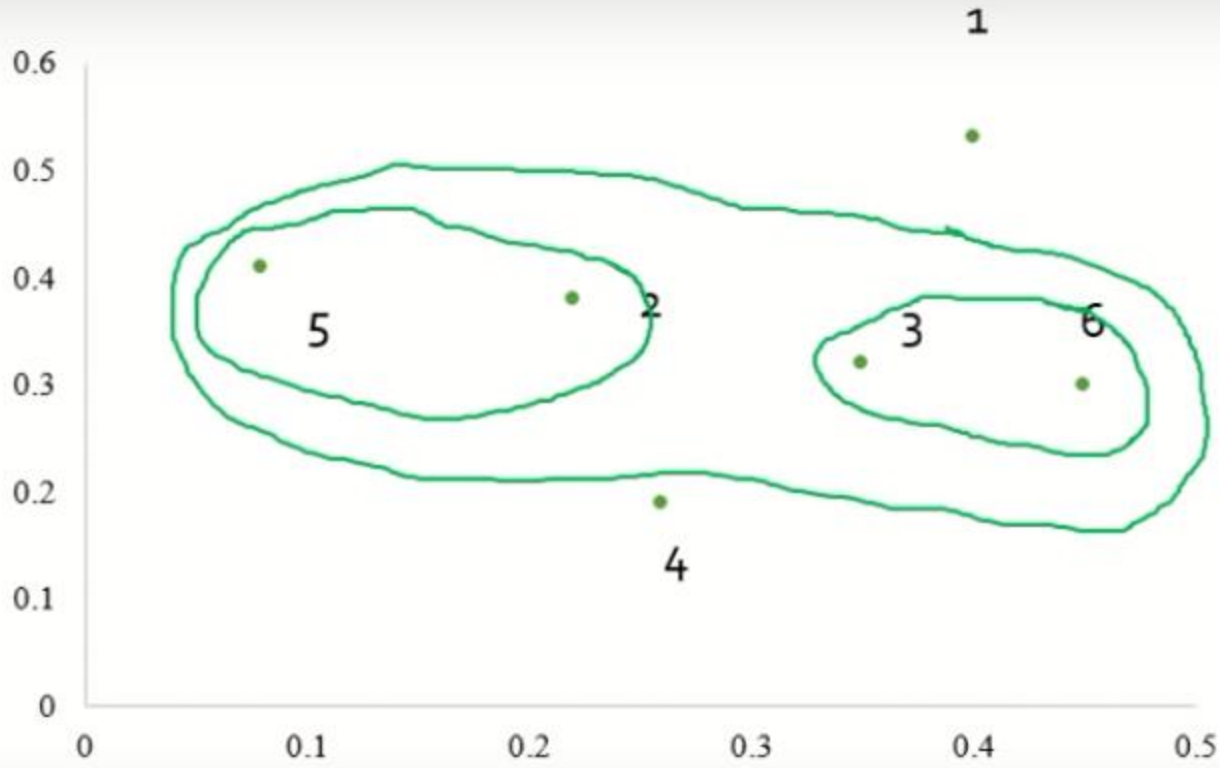
	P1	P2,P5	P3,P6	P4
P1	0			
P2,P5	0.23	0		
P3,P6	0.22	0.15	0	
P4	0.37	0.20	0.15	0

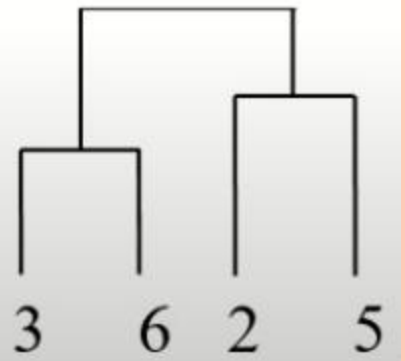
EXAMPLE – SINGLE LINK

- The distance matrix is

	P1	P2,P5	P3,P6	P4
P1	0			
P2,P5	0.23	0		
P3,P6	0.22	0.15	0	
P4	0.37	0.20	0.15	0

EXAMPLE – SINGLE LINK





EXAMPLE – SINGLE LINK

- To update the distance matrix MIN[dist((P2,P5),(P3,P6)),P1]
- MIN[dist((P2,P5),P1), ((P3,P6),P1)]

  = min[(0.23,0.22)]

  = 0.22

EXAMPLE – SINGLE LINK

- To update the distance matrix MIN[dist((P2,P5),(P3,P6)),P4]
- MIN[dist((P2,P5),P4), ((P3,P6),P4)]

  = min[(0.20,0.15)]

  = 0.15

EXAMPLE – SINGLE LINK

- The updated distance matrix for cluster P2,P5,P3,P6

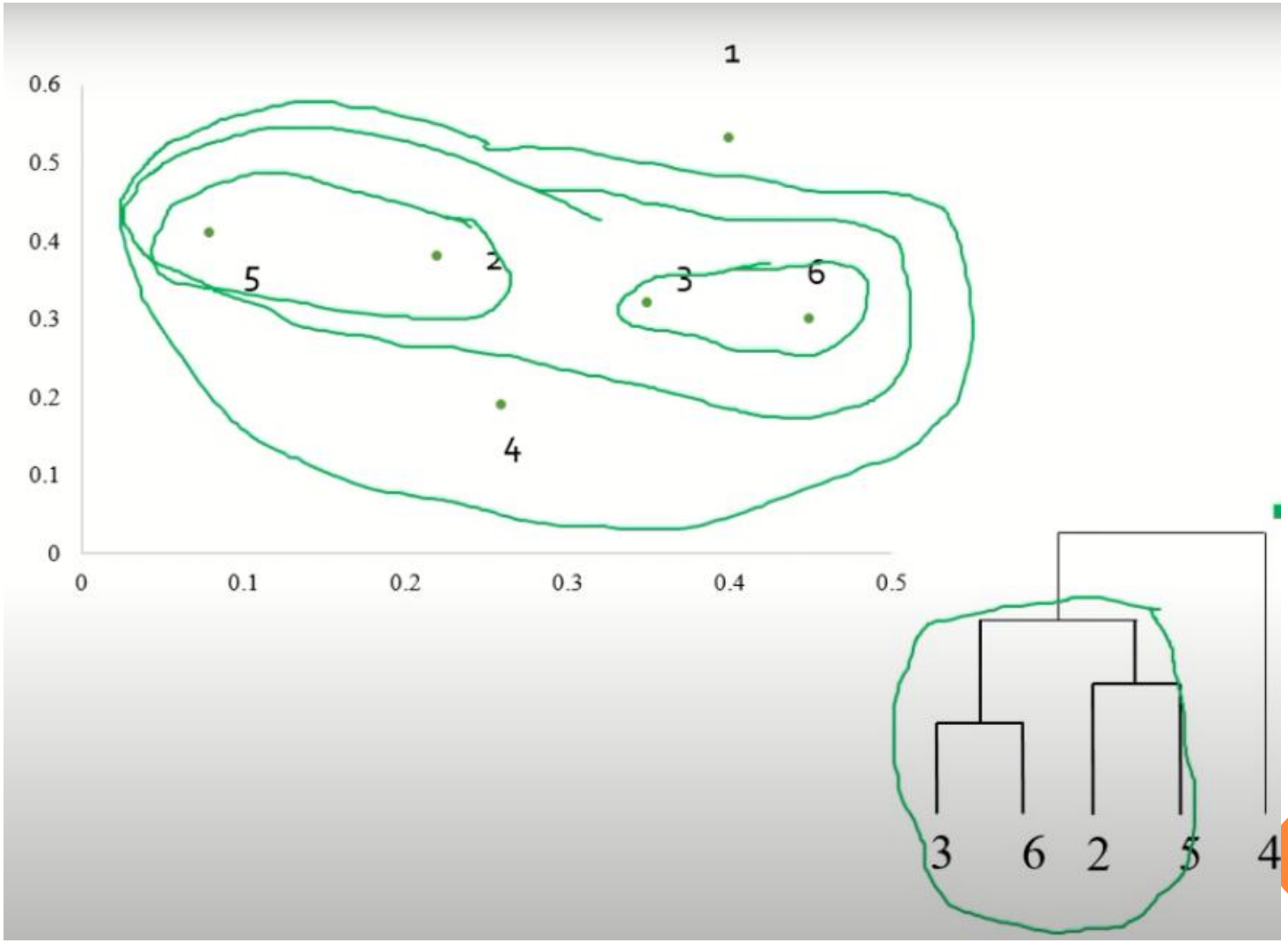
	P1	P2,P5,P3,P6	P4
P1	0		
P2,P5,P3,P6	0.22	0	
P4	0.37	0.15	0

EXAMPLE – SINGLE LINK

- The distance matrix is

	P1	P2,P5,P3,P6	P4
P1	0		
P2,P5,P3,P6	0.22	0	
P4	0.37	0.15	0

EXAMPLE – SINGLE LINK



EXAMPLE – SINGLE LINK

- To update the distance matrix MIN[dist(P2,P5,P3,P6),P4]
- MIN[dist((P2,P5,P3,P6),P1), (P4,P1)]

  = min[(0.22,0.37)]

  = 0.22

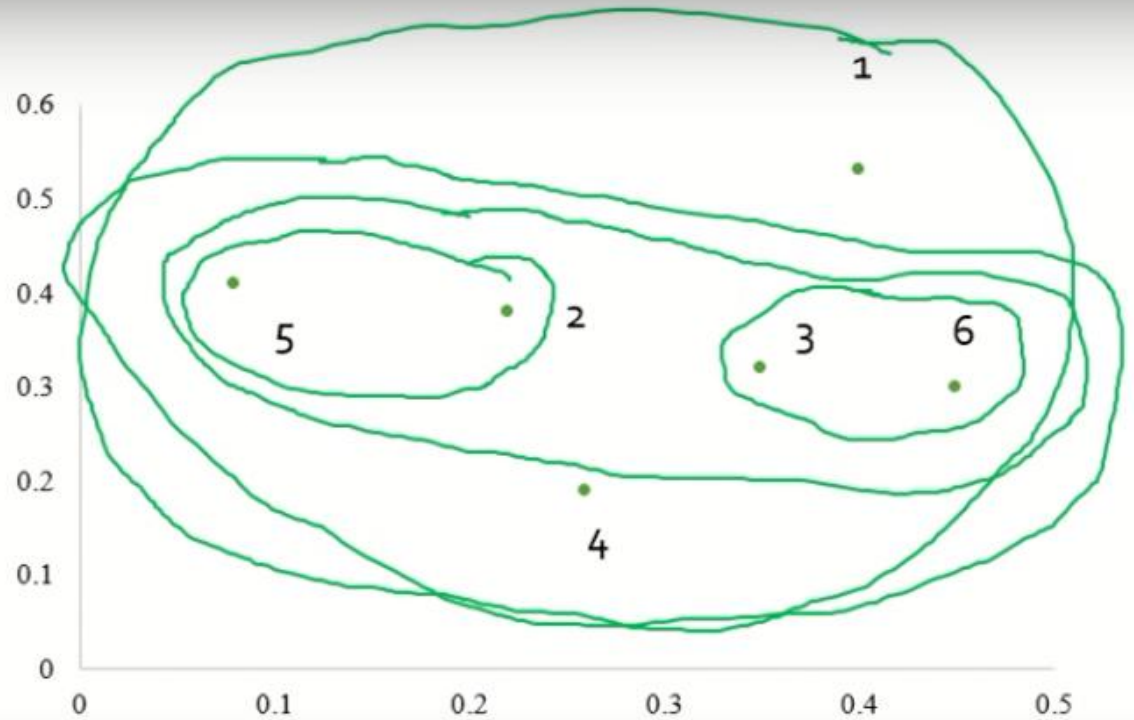
EXAMPLE – SINGLE LINK

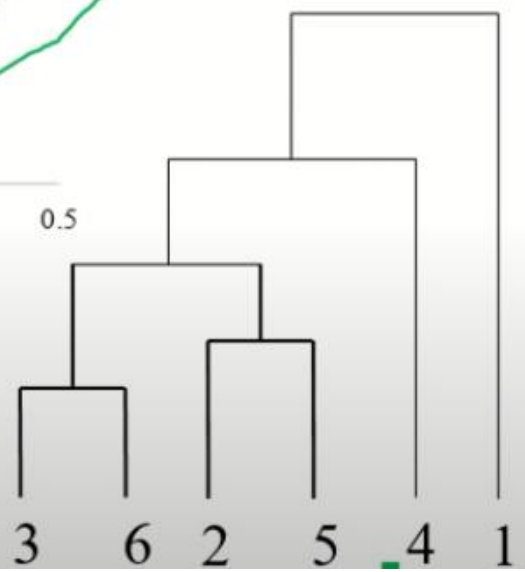
January

- The updated distance matrix for cluster P2,P5,P3,P6,P4

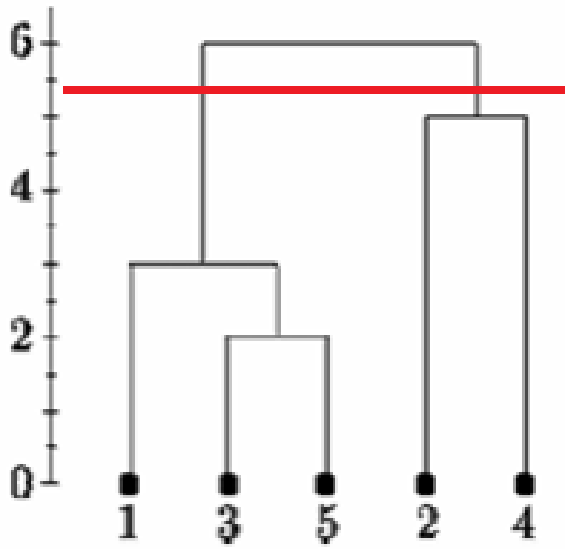
	P1	P2,P5,P3,P6,P4
P1	0	
P2,P5,P3,P6,P4	0.22	0

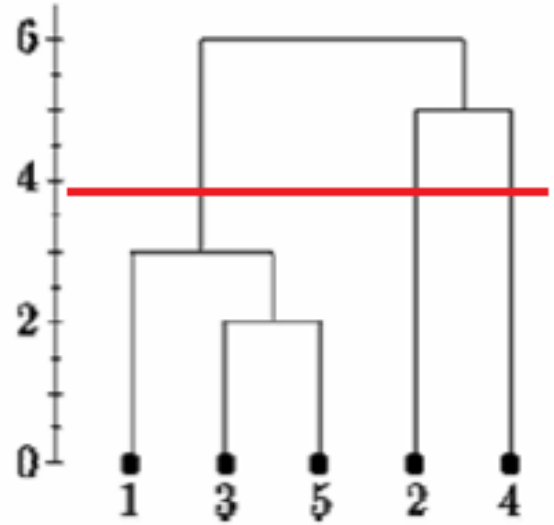
EXAMPLE – SINGLE LINK



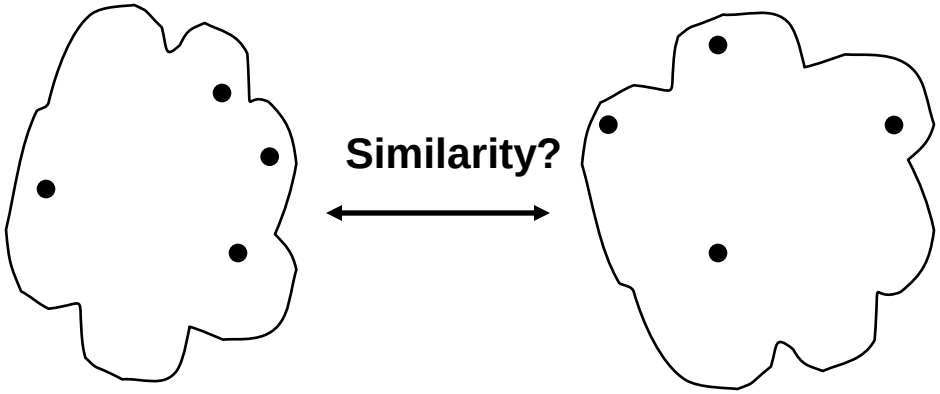


Determining clusters





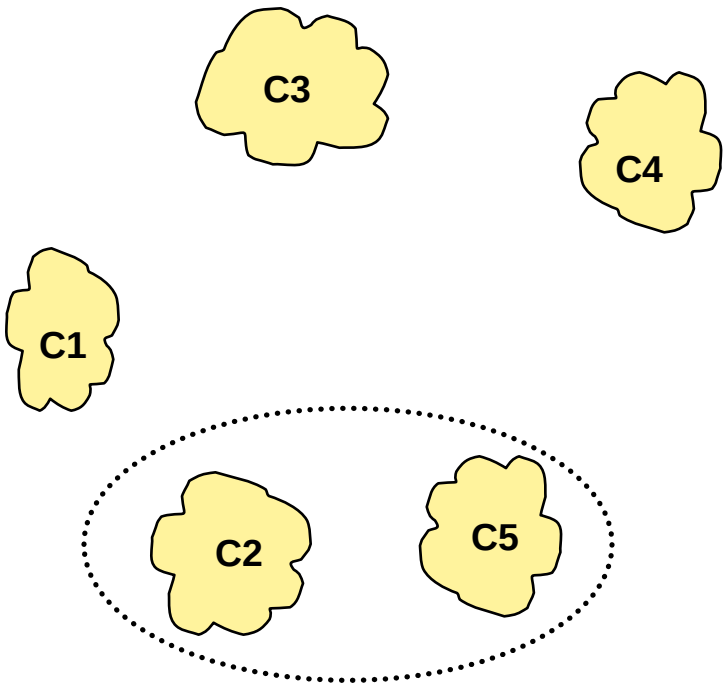
Inter cluster distance measures



- Single Link
- Average Link
- Complete Link
- Distance between centroids

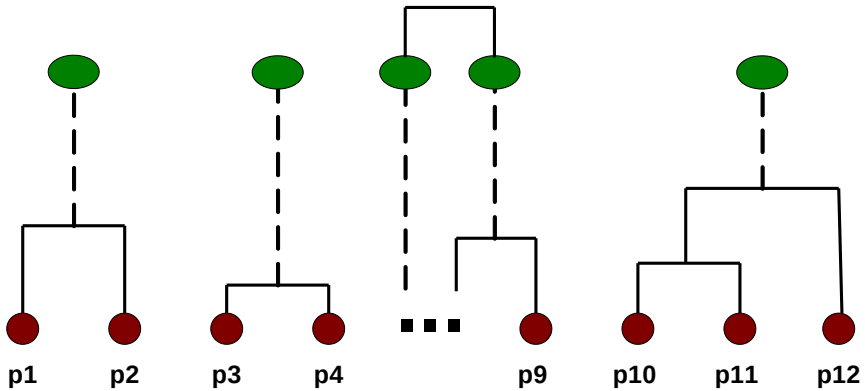
Intermediate situation

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



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

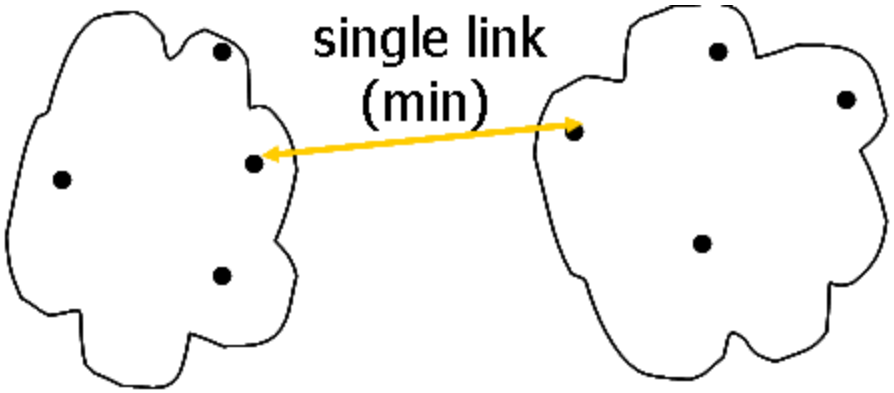
Distance Matrix



Single link

- Smallest distance between an element in one cluster and an element in the other

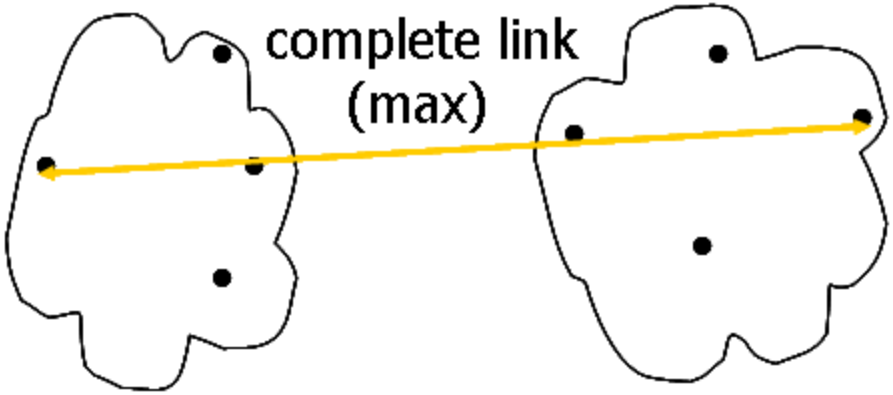
$$D(c_i, c_j) = \min_{x \in c_i, y \in c_j} D(x, y)$$



Complete link

- Largest distance between an element in one cluster and an element in the other

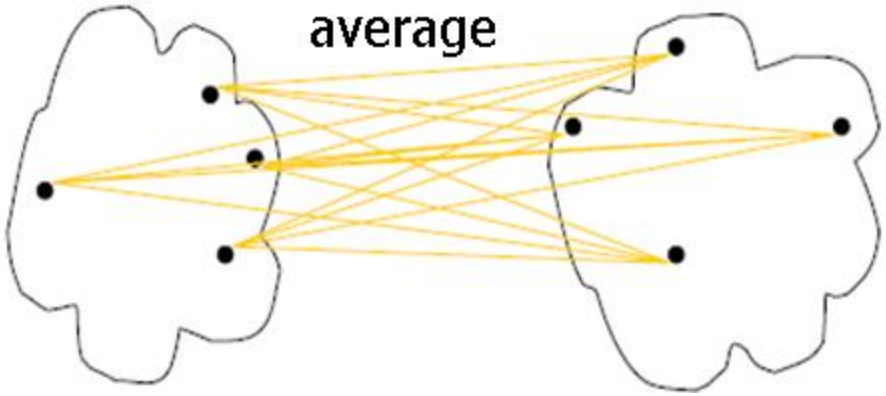
$$D(c_i, c_j) = \max_{x \in c_i, y \in c_j} D(x, y)$$



Average Link

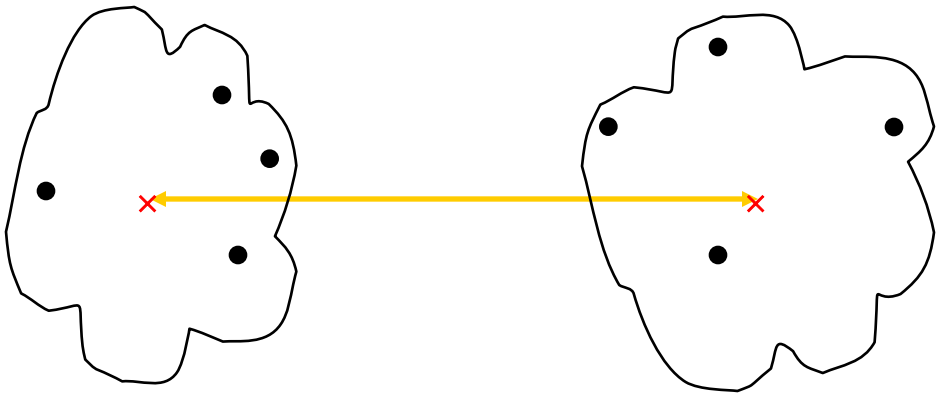
- Avg distance between an element in one cluster and an element in the other

$$D(c_i, c_j) = \underset{x \in c_i, y \in c_j}{avg} D(x, y)$$

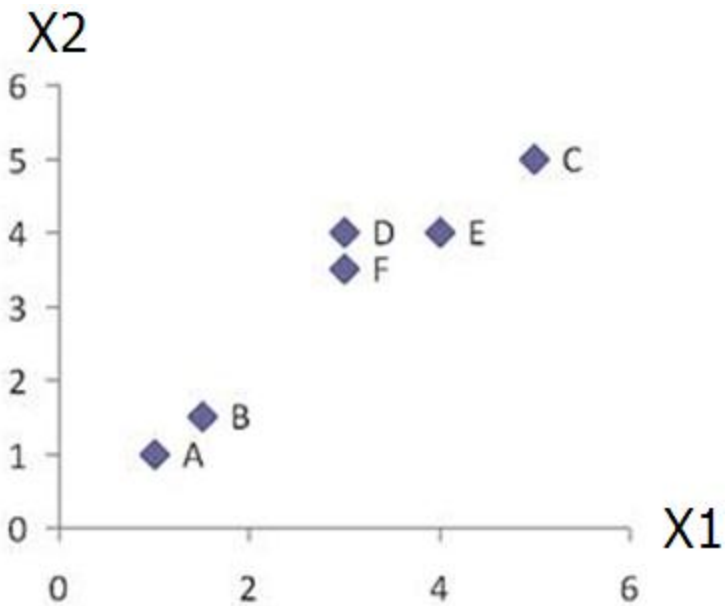


Distance between centroids

- Distance between the centroids of two clusters



Agglomerative Clustering - Example



	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Data matrix

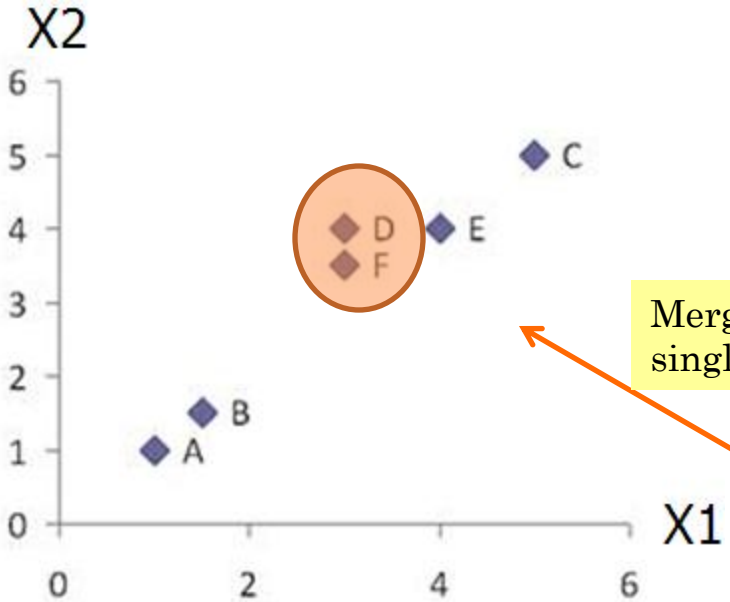
$$d_{AB} = ((1-1.5)^2+(1-1.5)^{2)1/2} = 0.707$$

Euclidean distance

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Merge two closest clusters

Agglomerative Clustering - Example



	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

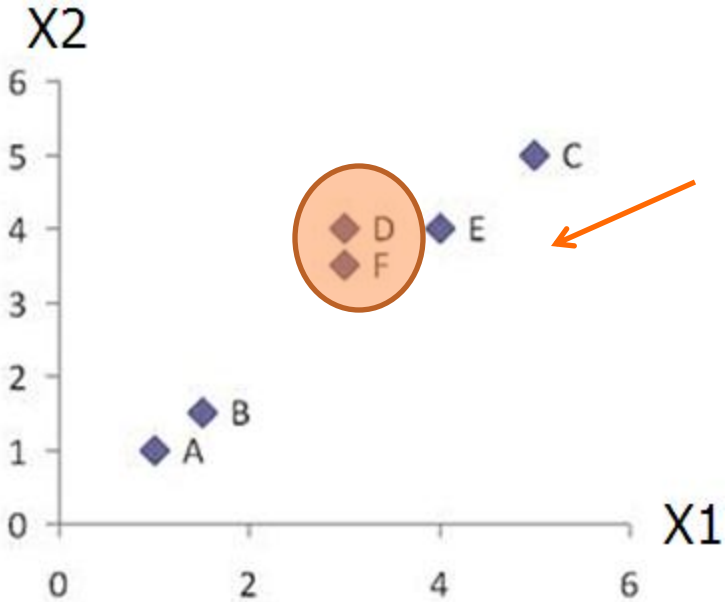
Data matrix

Merge them into single cluster`

Find two closest clusters

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Agglomerative Clustering - Example



Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Dist	A	B	C	D,F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D,F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Agglomerative Clustering - Example

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Min Distance – Single Linkage

$D_{(D,F)\rightarrow A} = \min(d_{DA},d_{FA}) = \min(3.61,3.20) = 3.20$

$D_{(D,F)\rightarrow B} = \min(d_{DB},d_{FB}) = \min(2.92,2.50) = 2.50$

$D_{(D,F)\rightarrow C} = \min(d_{DC},d_{FC}) = \min(2.24,2.50) = 2.24$

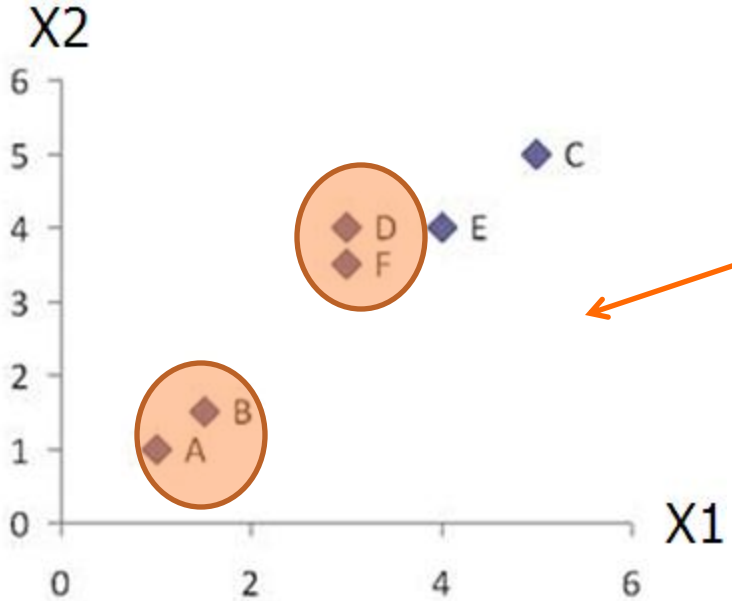
$D_{(D,F)\rightarrow E} = \min(d_{DE},d_{FE}) = \min(1.00,1.12) = 1.00$

Dist	A	B	C	D,F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D,F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Dist	A	B	C	D,F	E
A	0.00	0.71	5.66	**3.20**	4.24
B	0.71	0.00	4.95	**2.50**	3.54
C	5.66	4.95	0.00	**2.24**	1.41
D,F	**3.20**	**2.50**	**2.24**	0.00	**1.00**
E	4.24	3.54	1.41	**1.00**	0.00

Merge two closest clusters

Agglomerative Clustering - Example



Dist	A	B	C	D,F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D,F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00

Dist	A,B	C	D,F	E
A,B	0.00	?	?	?
C	?	0.00	2.24	1.41
D,F	?	2.24	0.00	1.00
E	?	1.41	1.00	0.00

Agglomerative Clustering - Example

Dist	A	B	C	D,F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D,F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00

$D_{(A,B)\rightarrow C} = \min(d_{CA},d_{CB}) = \min(5.66,4.95) = 4.95$

$D_{(A,B)\rightarrow (D,F)} = \min(d_{DA},d_{DB}, d_{FA},d_{FB})$
$= \min(3.61, 2.92, 3.20, 2.50) = 2.50$

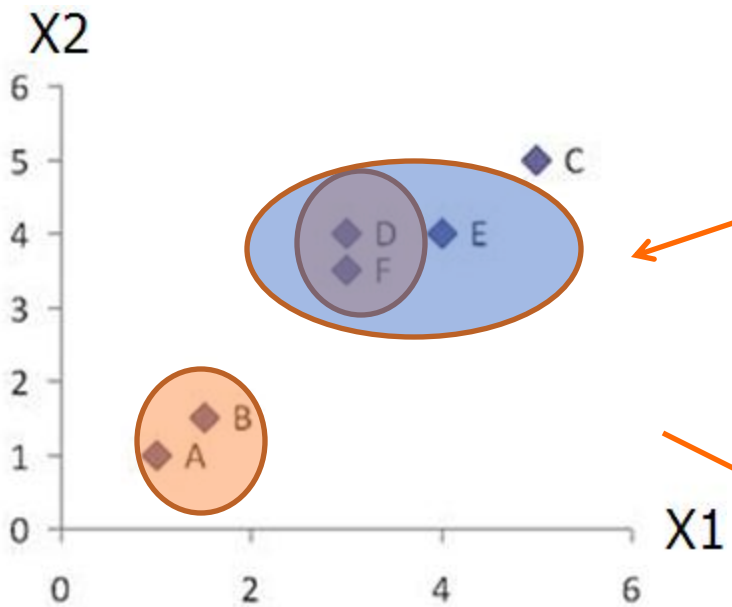
$D_{(A,B)\rightarrow E} = \min(d_{AE},d_{BE}) = \min(4.24,3.54) = 3.54$

Dist	A,B	C	D,F	E
A,B	0.00	?	?	?
C	?	0.00	2.24	1.41
D,F	?	2.24	0.00	1.00
E	?	1.41	1.00	0.00

Dist	A,B	C	D,F	E
A,B	0.00	**4.95**	**2.50**	**3.54**
C	**4.95**	0.00	2.24	1.41
D,F	**2.50**	2.24	0.00	1.00
E	**3.54**	1.41	1.00	0.00

Merge two closest clusters/Update Distance Matrix

Agglomerative Clustering - Example

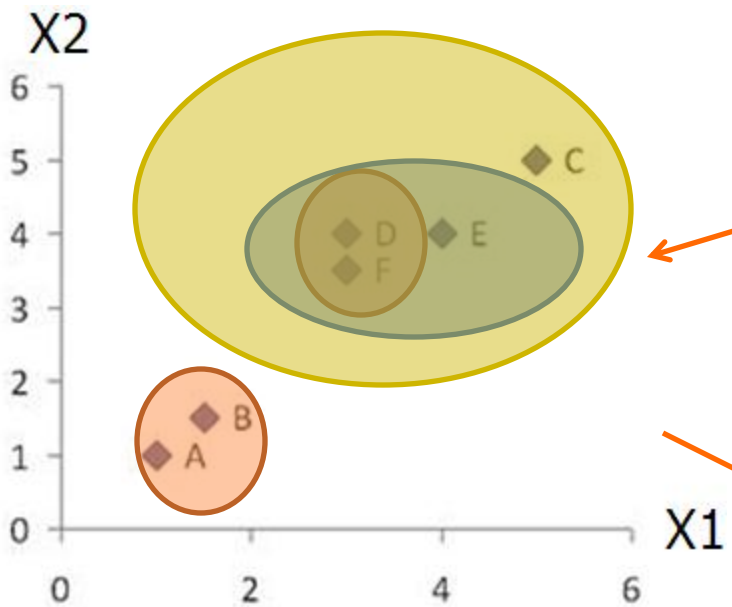


Dist	A,B	C	D,F	E
A,B	0.00	4.95	2.50	3.54
C	4.95	0.00	2.24	1.41
D,F	2.50	2.24	0.00	1.00
E	3.54	1.41	1.00	0.00

Dist	(A,B)	C	(D,F),E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D,F),E	2.50	1.41	0.00

Merge two closest clusters/Update Distance Matrix

Agglomerative Clustering - Example



Dist	(A,B)	C	(D,F),E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D,F),E	2.50	1.41	0.00

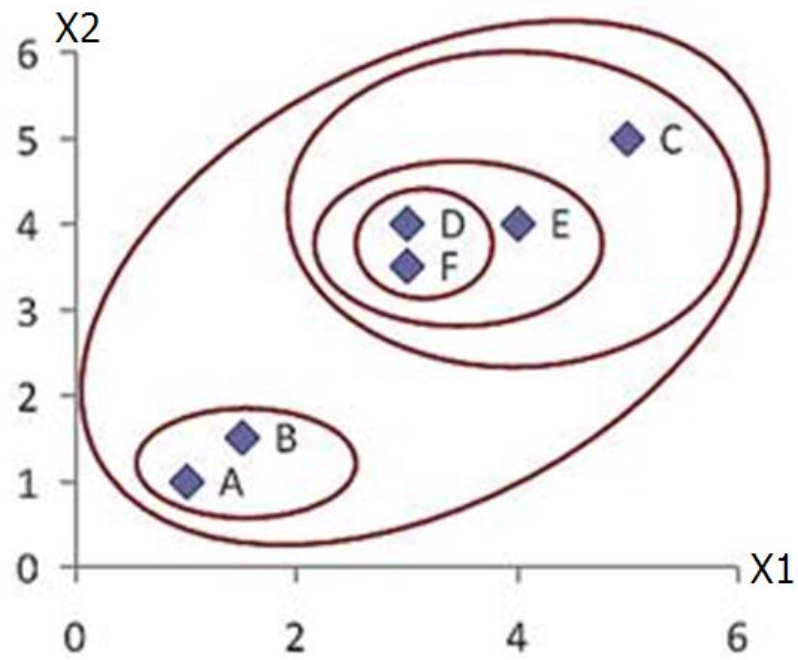
Dist	(A,B)	((D,F),E),C
(A,B)	0.00	2.50
((D,F),E),C	2.50	0.00

Final Result

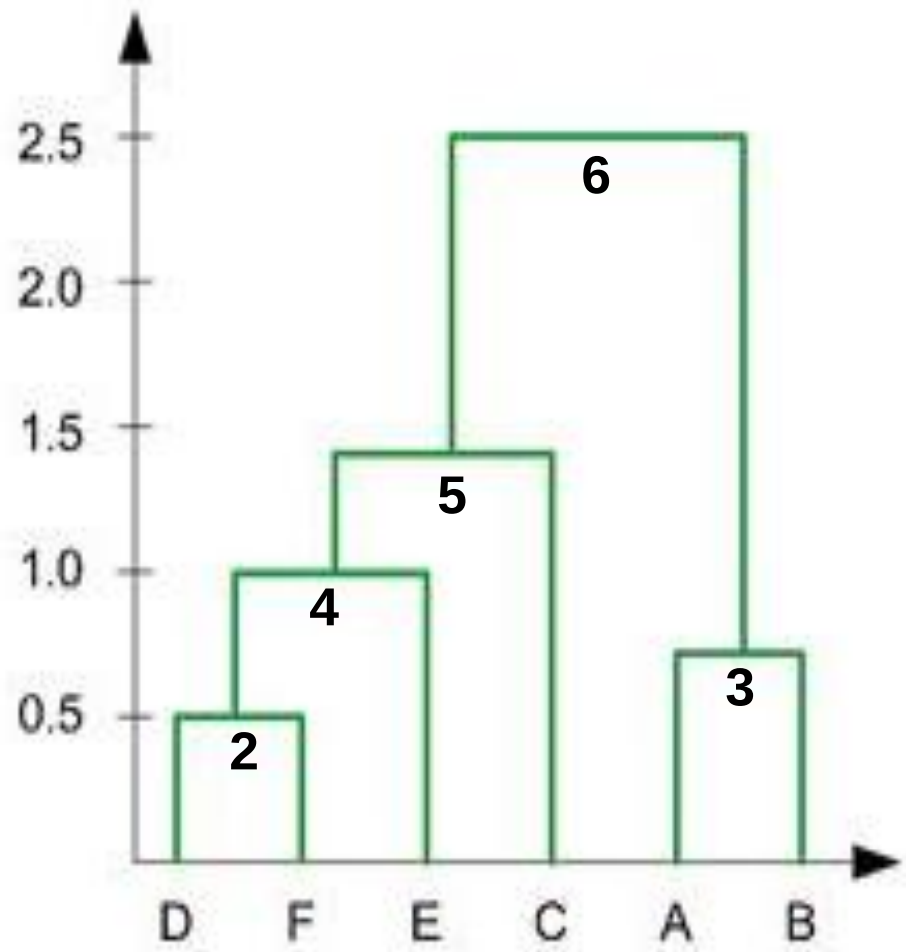
Agglomerative Clustering - Example

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Data matrix



Agglomerative Clustering - Example



1. In the beginning we have 6 clusters: A, B, C, D, E and F
2. We merge cluster D and F into cluster (D, F) at distance 0.50
3. We merge cluster A and cluster B into (A, B) at distance 0.71
4. We merge cluster E and (D, F) into ((D, F), E) at distance 1.00
5. We merge cluster ((D, F), E) and C into (((D, F), E), C) at distance 1.41
6. We merge cluster (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance 2.50
7. The last cluster contain all the objects, thus conclude the computation