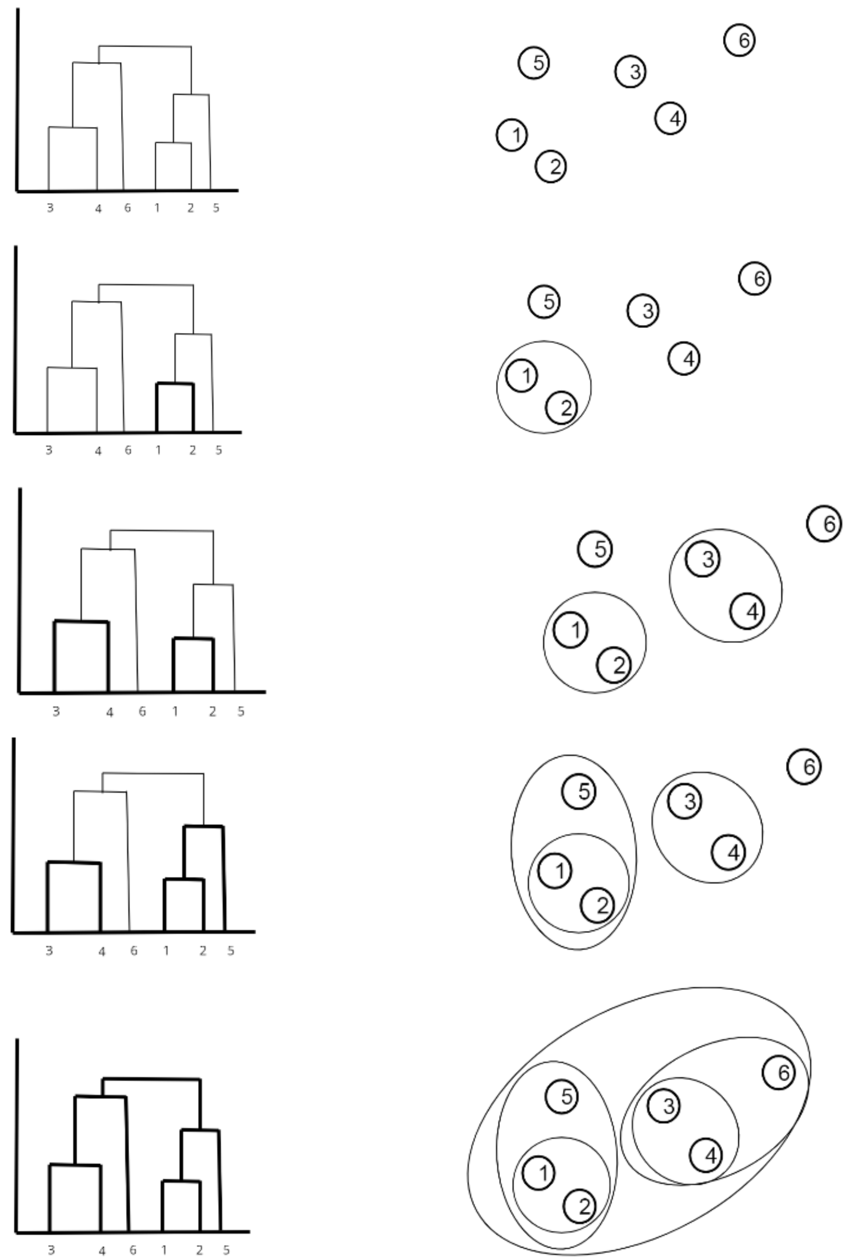
Lecture 5: Hierarchical Clustering

1) Hierarchical Clustering

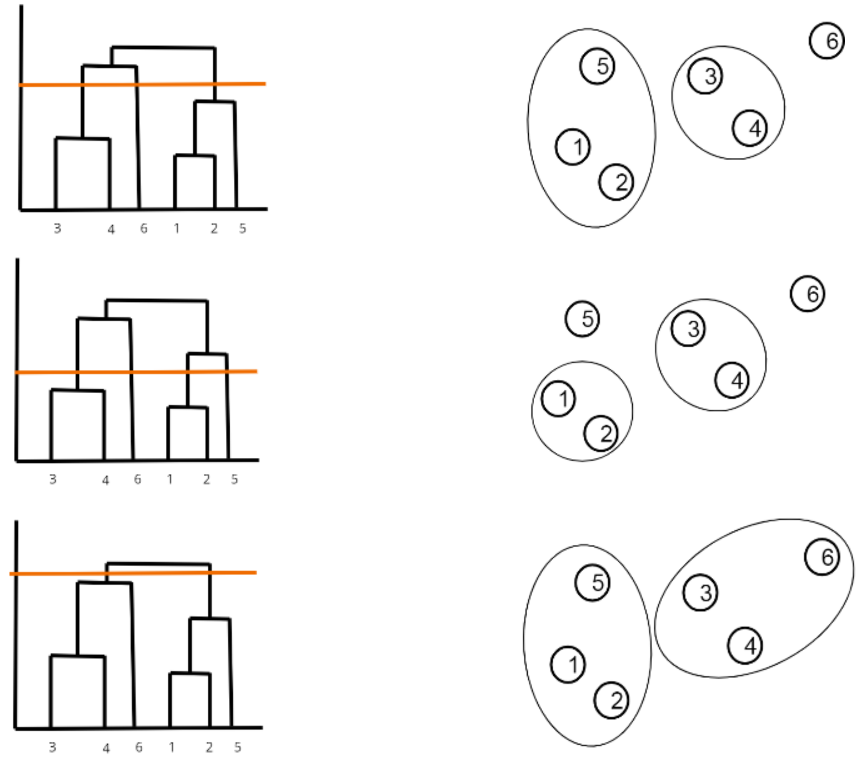
a) Two types:
   i) Agglomerative: our main focus
      1) Start with every point in it own cluster
      2) At each step, merge the two closest clusters
      3) Stop when every point is in the same cluster
   ii) Divisive
      1) Start with every point in the same cluster
      2) At each step, split until every point is in its own cluster

b) Agglomerative Clustering Algorithm
   i) Each point in dataset is its own cluster
   ii) Compute distance between all pairs of clusters
   iii) Merge 2 closest clusters
   iv) Repeat 3 & 4 until all points are in the same cluster

c) At every step, record which clusters were merged in order to produce dendrogram



d) Can cut dendrogram at any threshold to produce any number of clusters



2) Distance Functions

a) Define:

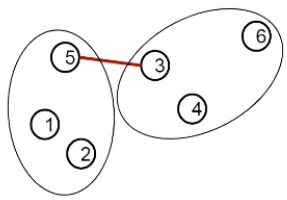
i) Distance between points: d(p_1, p_2)

ii) Distance between clusters: D(C_1, C_2)

b) Single-Link Distance

i) Minimum of all pairwise distances between a point from one cluster and a point from the other cluster

$$D_{SL}(C_1, C_2) = \min \{d(p_1, p_2) \mid p_1 \in C_1, p_2 \in C_2\}$$



Depends on choice of **d**

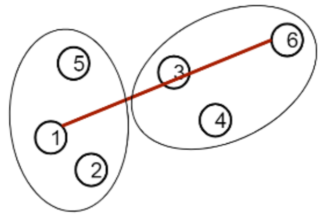
Can handle clusters of different sizes

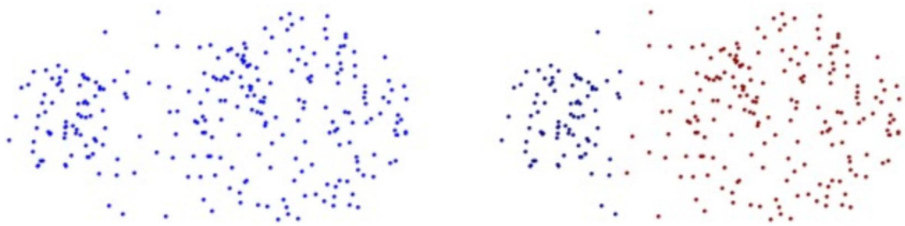
But... Sensitive to noise points
Tends to create elongated clusters

c) Complete Link Distance

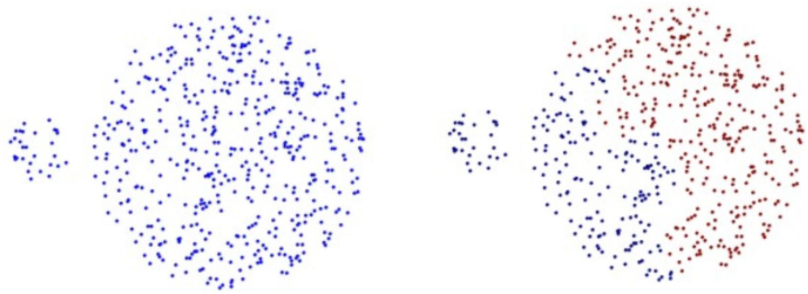
i) Maximum of all pairwise distances between a point from one cluster and a point from the other cluster

$$D_{CL}(C_1, C_2) = \max \{ d(p_1, p_2) \mid p_1 \in C_1, p_2 \in C_2 \}$$





Less susceptible to noise
Creates more balanced (equal diameter) clusters

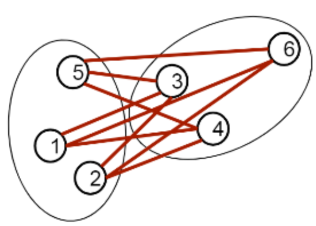


But... Tends to split up large clusters.
All clusters tend to have the same diameter

d) Average Link Distance

i) Average of all pairwise distances between a point from one cluster and a point from the other cluster

$$D_{AL}(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{p_1 \in C_1, p_2 \in C_2} d(p_1, p_2)$$

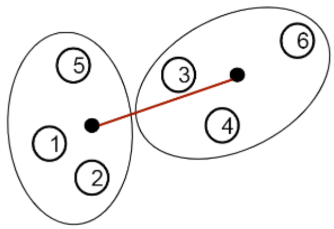


- Less susceptible to noise and outlier, but tend to be biased toward globular clusters

e) Centroid Distance

i) Distance between centroids of clusters

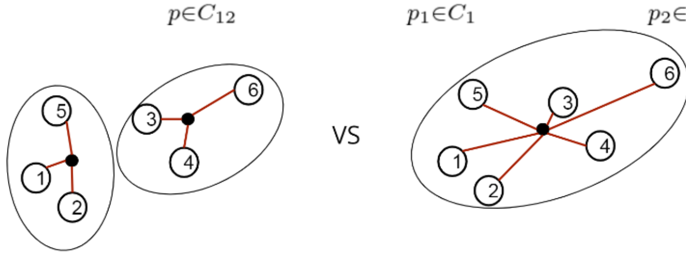
$$D_C(C_1, C_2) = d(\mu_1, \mu_2)$$



f) Ward's Distance

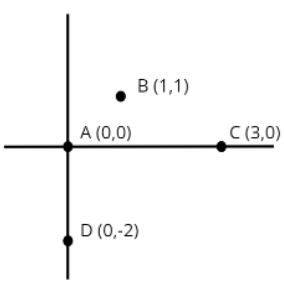
i) Difference between spread/variance of points in merged cluster and unmerged clusters

$$D_{WD}(C_1, C_2) = \sum_{p \in C_{12}} d(p, \mu_{12}) - \sum_{p_1 \in C_1} d(p_1, \mu_1) - \sum_{p_2 \in C_2} d(p_2, \mu_2)$$



3) Example

d = Euclidean
D = Single-Link

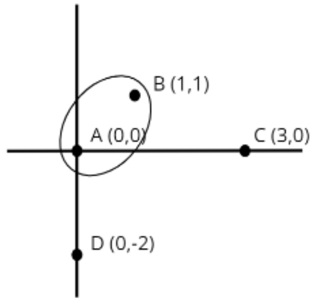


Distance Matrix

	A	B	C	D
A				
B				
C				
D				

Distance Matrix

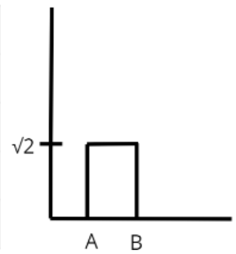
	A	B	C	D
A	0	√2	3	2
B	√2	0	√5	√10
C	5	√5	0	√13
D	2	√10	√13	0



Distance Matrix

	A	B	C	D
A	0	√2	3	2
B	√2	0	√5	√10
C	5	√5	0	√13
D	2	√10	√13	0

Dendrogram

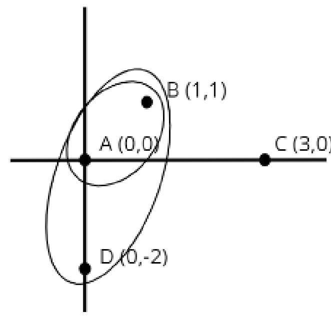


Distance Matrix

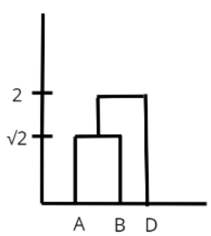
	A & B	C	D
A & B	0		
C		0	√13
D		√13	0

Distance Matrix

	A & B	C	D
A & B	0	√5	2
C	√5	0	√13
D	2	√13	0



Dendrogram

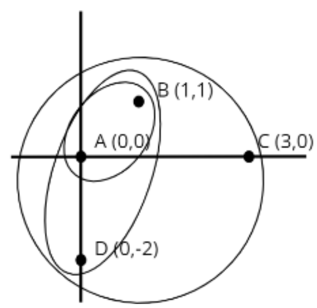


Distance Matrix

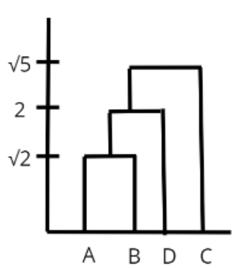
	A & B & D	C
A & B & D	0	
C		0

Distance Matrix

	A & B & D	C
A & B & D	0	√5
C	√5	0



Dendrogram



- Finding threshold to cut dendrogram requires exploration and tuning