Lecture 10. Hierarchical clustering

Xin Chen

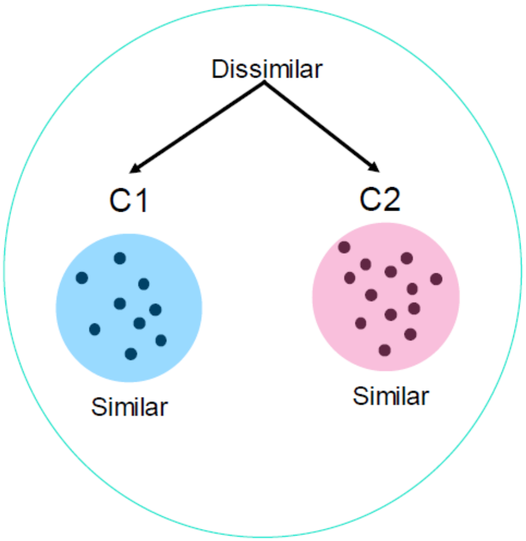
These slides are based on slides from Mahdi Roozbahani

Outline

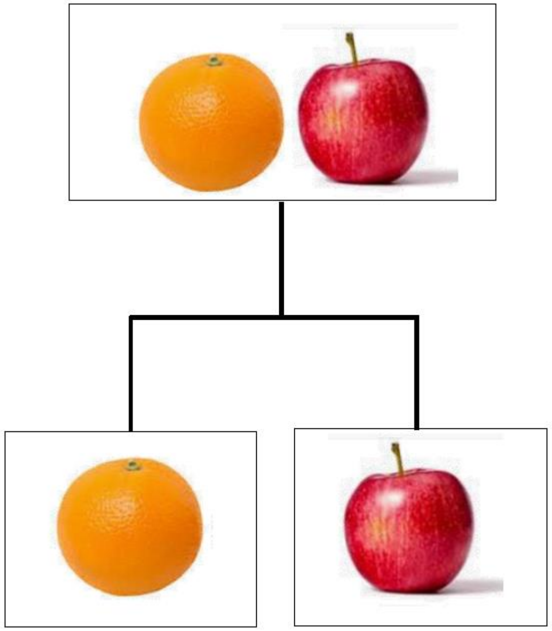
- Overview
- Bottom-up vs. Top-down clustering
- Measure distance between clusters

Partitional clustering vs. Hierarchical clustering

K-Means



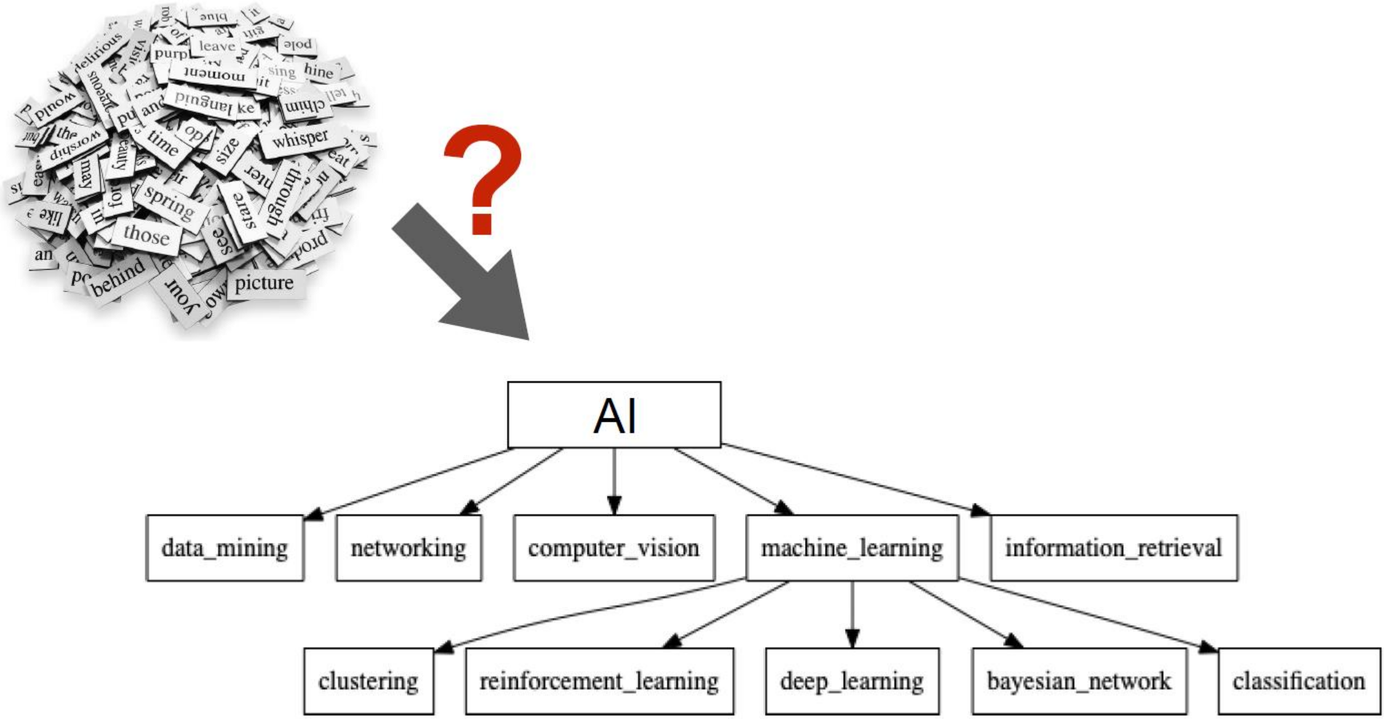
Hierarchical Clustering



Tree structure (parent-child relationship)

Hierarchical clustering

- How to cluster a set of CS papers into a hierarchy?



AI

data_mining

networking

computer_vision

machine_learning

information_retrieval

clustering

reinforcement_learning

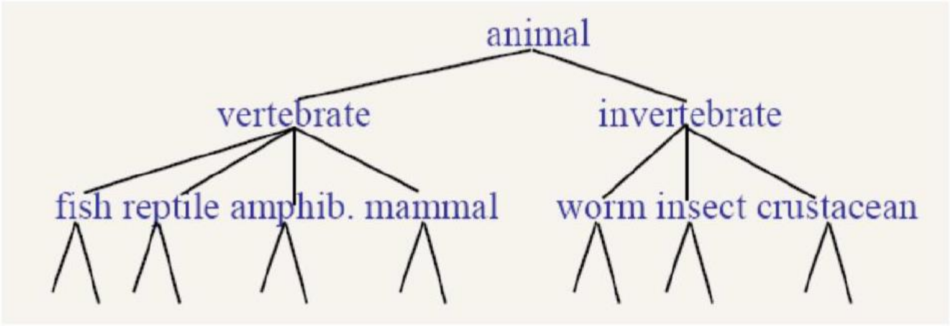
deep_learning

bayesian_network

classification

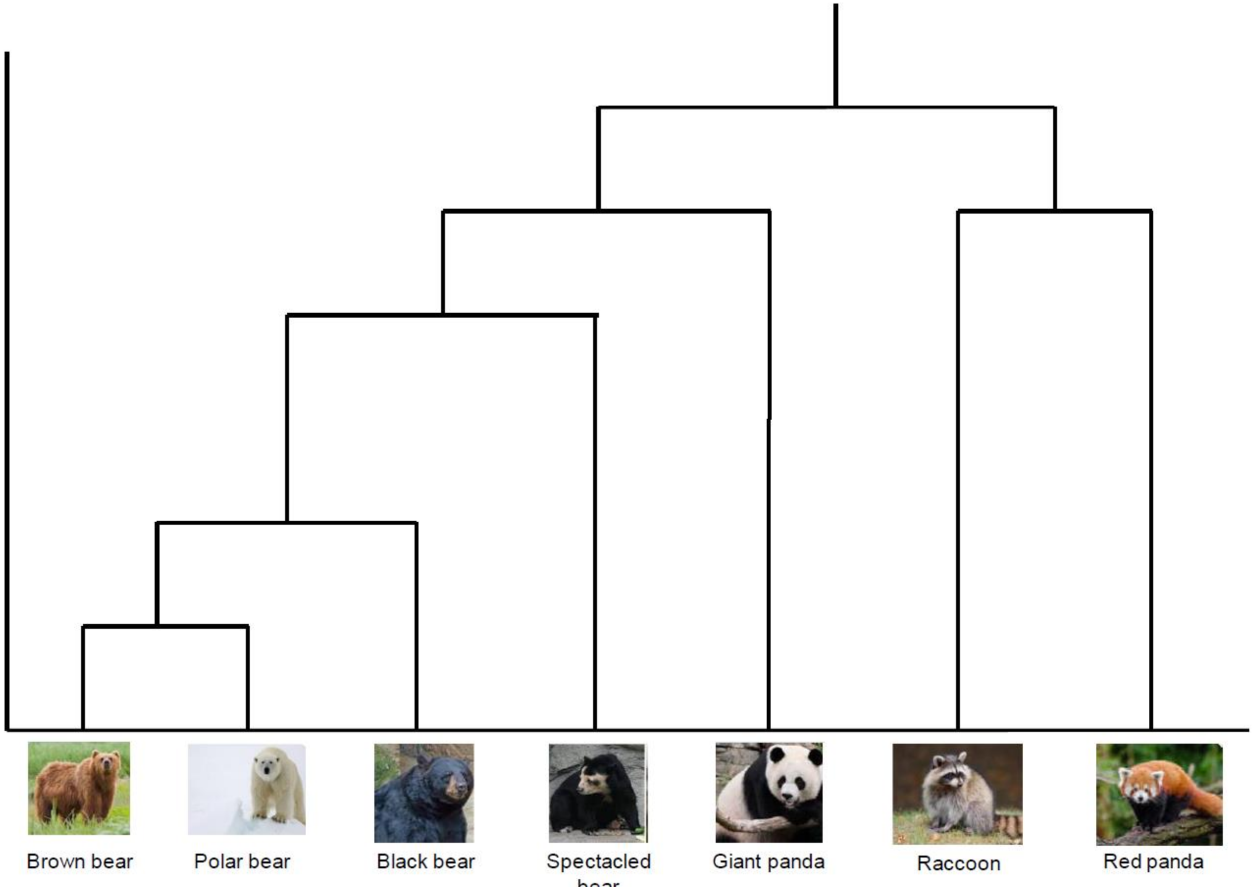
Hierarchical clustering

- Organize objects into a tree-based hierarchical taxonomy



- Many applications in the real world
  – Web pages
  – New articles
  – Scientific papers

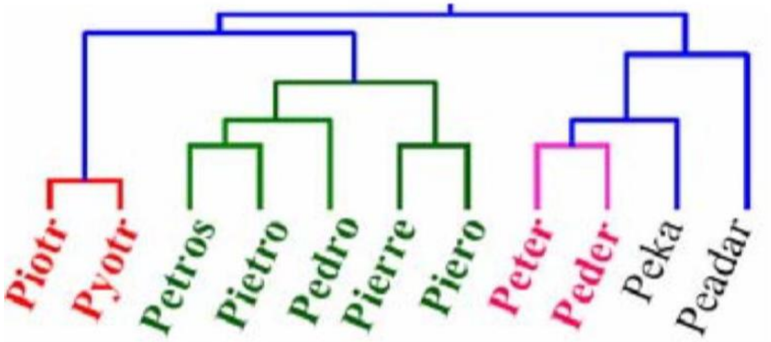
DNA sequence



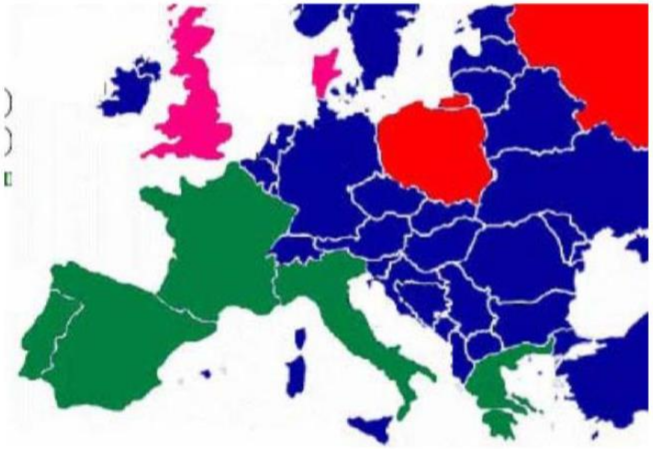
DNA sequencing and hierarchical clustering to find phylogenetic tree of animal evolution.

Using hierarchical clustering, researchers were able to place giant pandas closer to bears

- Organizing data at multiple granularities
- Cutting the dendrogram at a desired level leads to a sub-cluster: each connected component forms a cluster.



Dendrogram



Outline

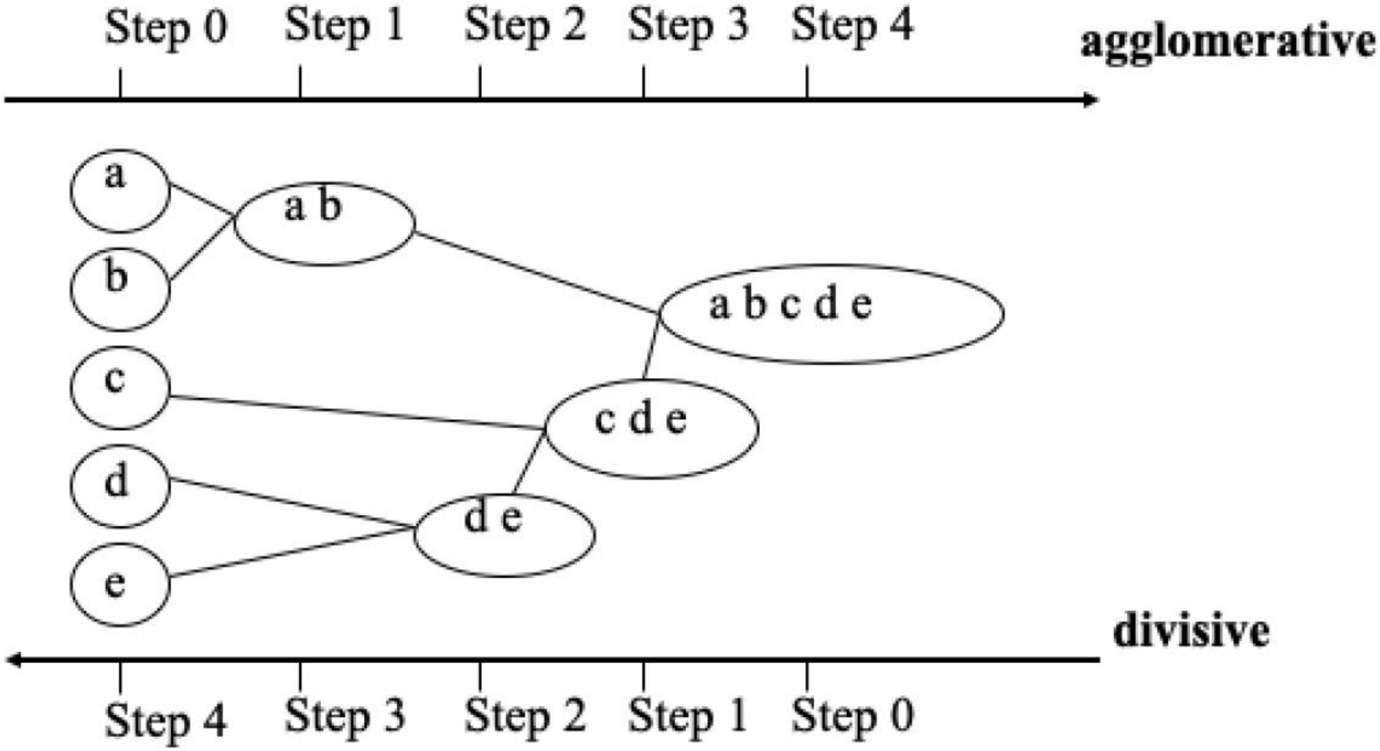
- Overview
- Bottom-up vs. Top-down clustering ←
- Measure distance between clusters

Two paradigms of hierarchical clustering

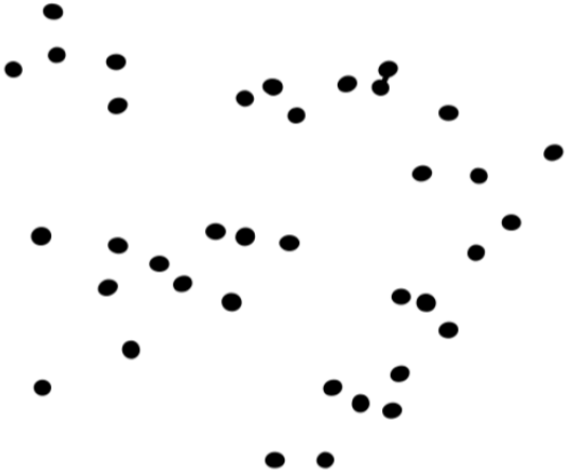
- Bottom-up agglomerative clustering
  - Start by considering each object as a separate cluster
  - Repeatedly join the closest pair of clusters
  - Stop when there is only one cluster left

- Top-down divisive clustering
  - Start by considering all objects as one large cluster
  - Recursively divide each cluster into two sub-clusters
  - Stop when each cluster contains only one object

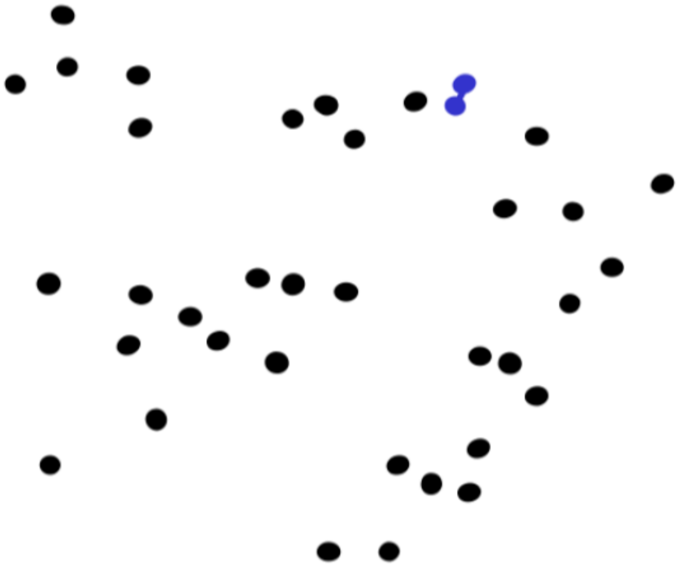
Bottom-up v.s. Top-down



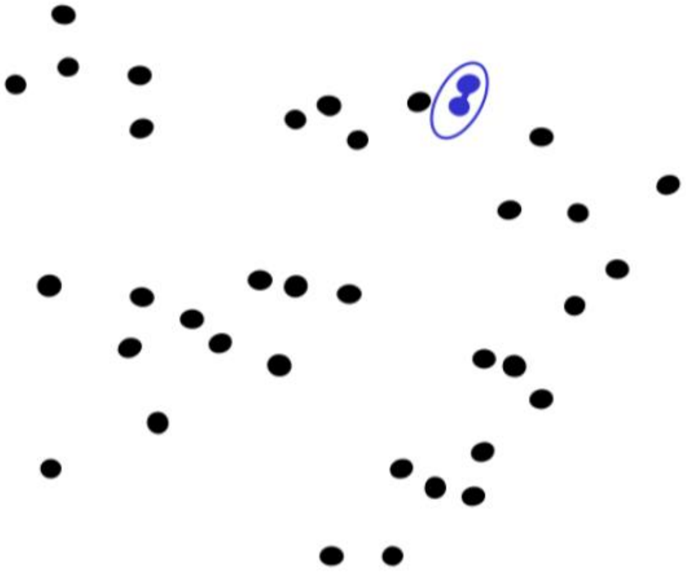
Bottom-up agglomerative clustering



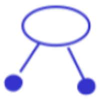
1. Say “every point is its own cluster”

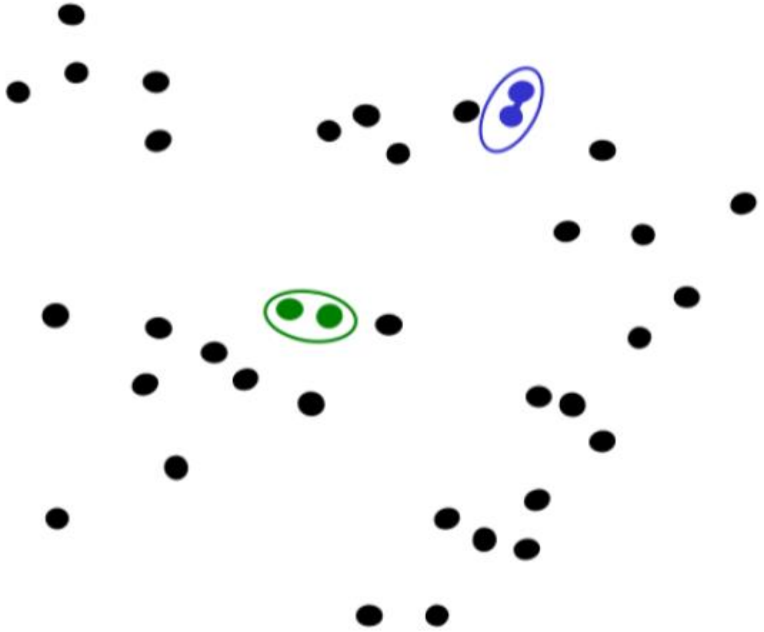


1. Say “every point is its own cluster”
2. Find “most similar” pair of clusters

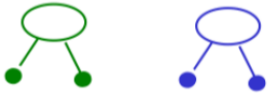


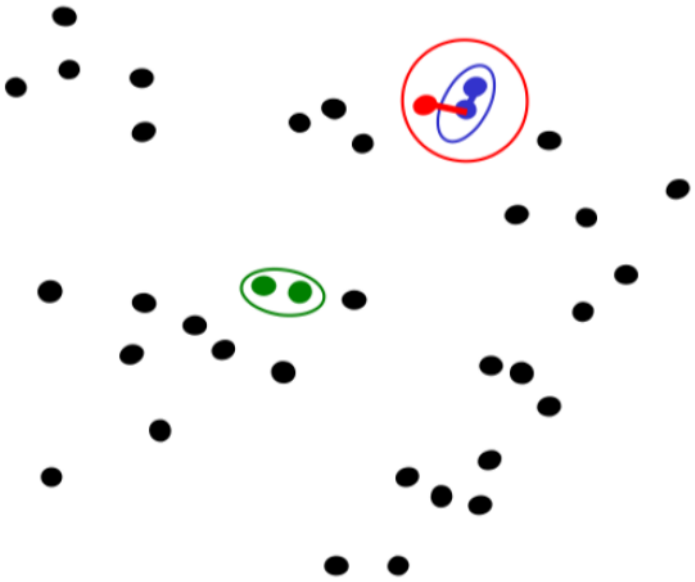
1. Say “every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster



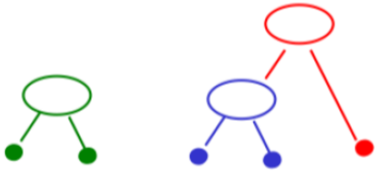


1. Say “every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster
4. Repeat





1. Say “every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster
4. Repeat

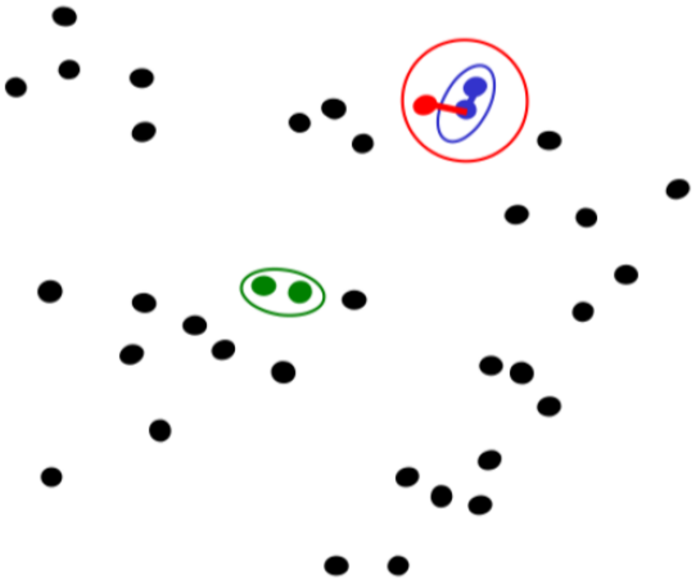


Outline

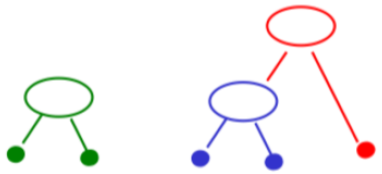
- Overview
- Bottom-up vs. Top-down clustering
- Measure distance between clusters ←

Key question: similarity function

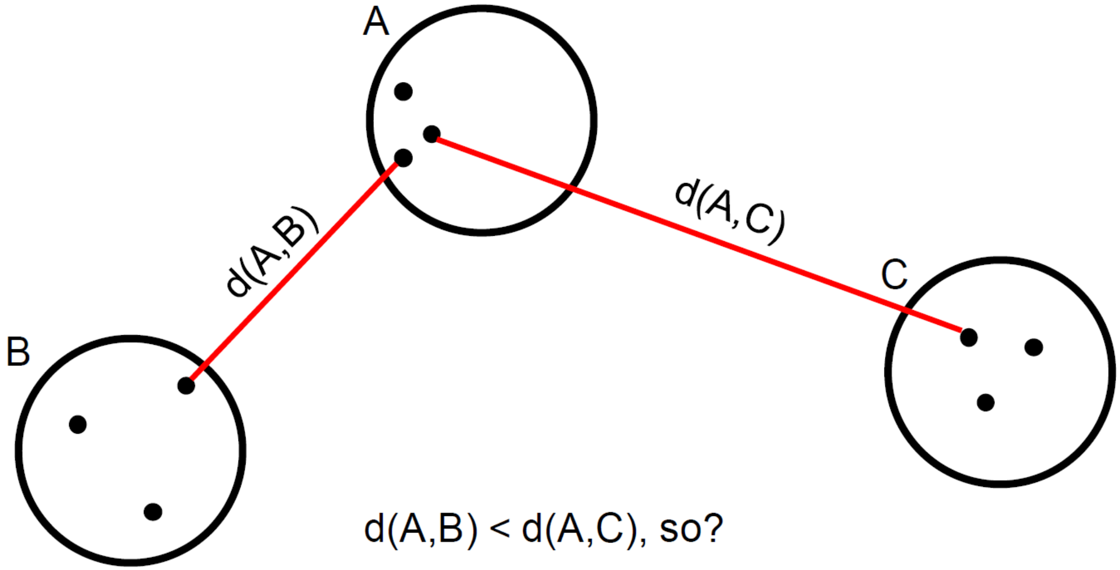
How to define “similarity” between clusters?



1. Say “every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster
4. Repeat

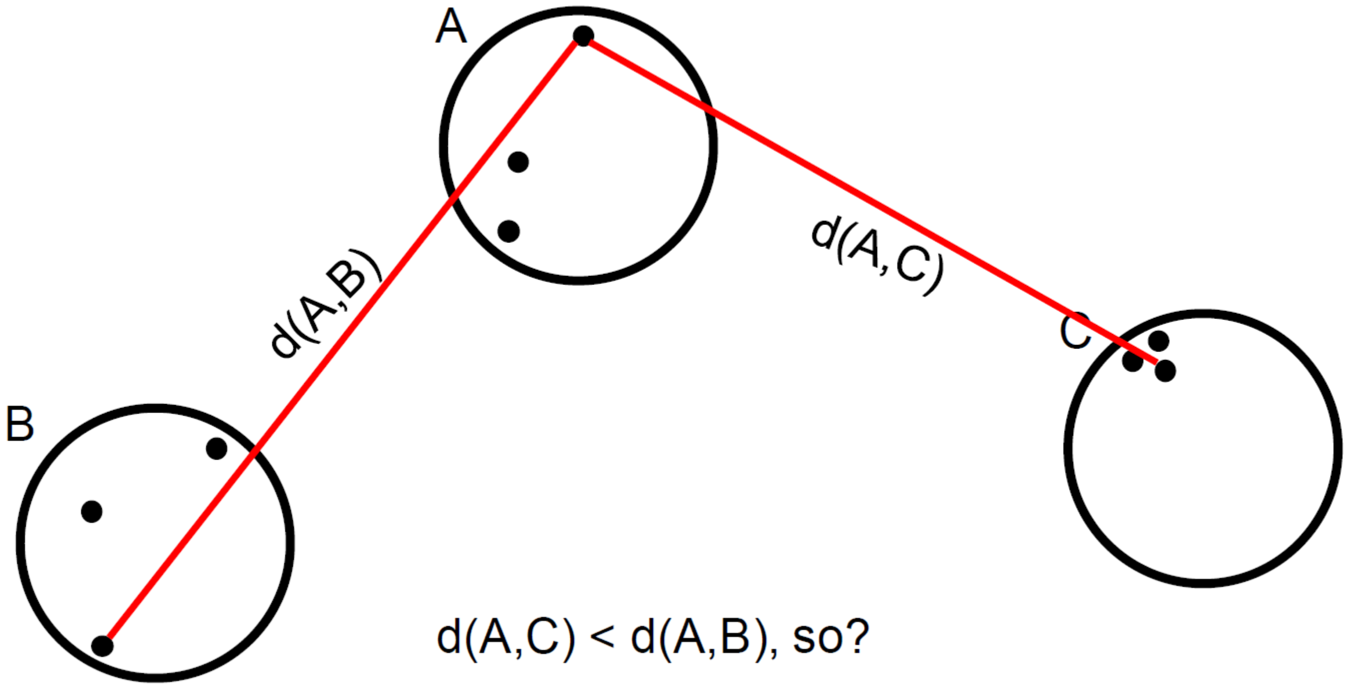


Single link



Merge A with either B or C. Which one?

Complete link



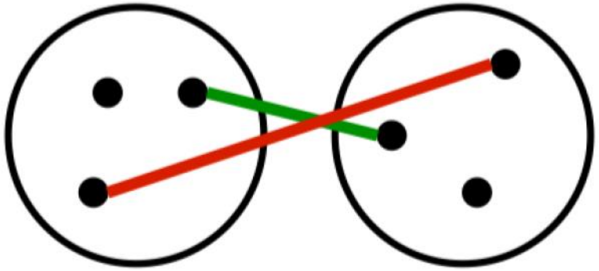
d(A,C) < d(A,B), so?

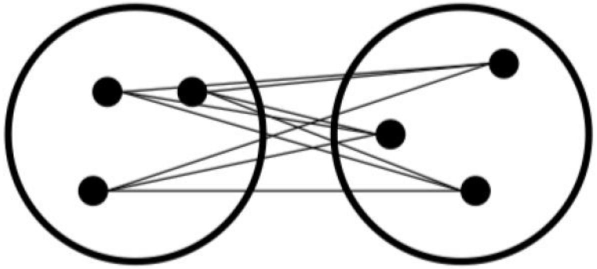
Merge A with either B or C. Which one?

Single link, complete link and average link

- Single link: a chain of points can be extended for long distance without regard to the overall shape of the emerging cluster. This fact is called chaining. It is also sensitive to outliers. It is faster in general.
- Complete link: clusters are split into two groups of roughly equal size when we cut the dendrogram at the last merge. In general, this is a more useful organization of the data than a clustering with chains. It avoids chaining and more robust to outliers. Generally slower.
- Average link: when you don’t know which one may be better for you, start it with the average link method.

Define distance between two cluster

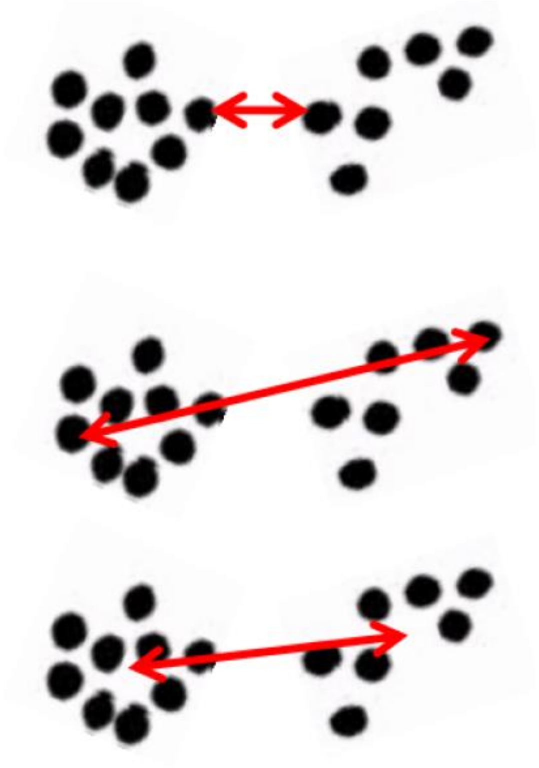




Bottom-up agglomerative clustering

Different algorithms differ in how similarities are defined.

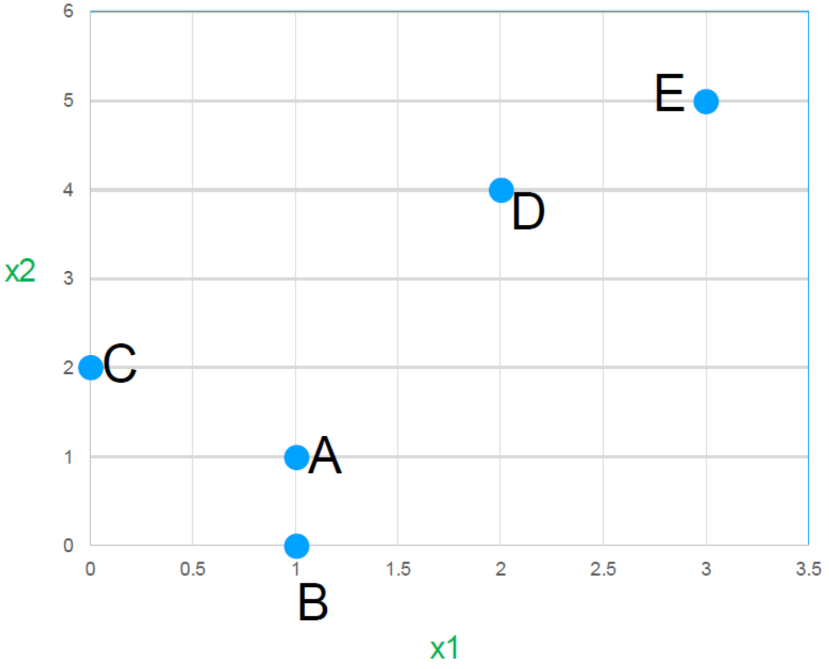
- Single link: nearest neighbor as similarity between their closest members.
- Complete link: furthest neighbor as similarity between their furthest members
- Centroid: similarity between the centers of gravity
- Average link: average similarity of all cross-cluster pairs.



Different distance function can lead to different results.

An example

i	X1	X2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



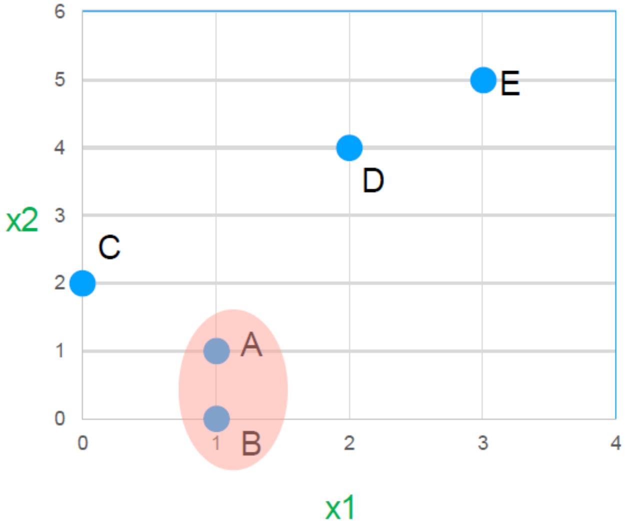
EUCLIDEAN DISTANCE

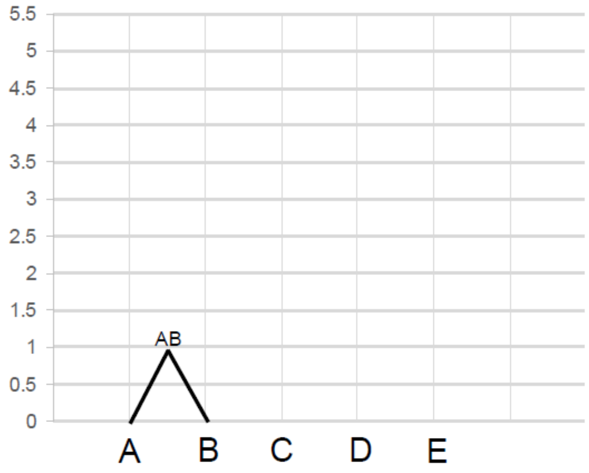
	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0

Distance based on average point

	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0

	(A,B)	C	D	E
(A,B)	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



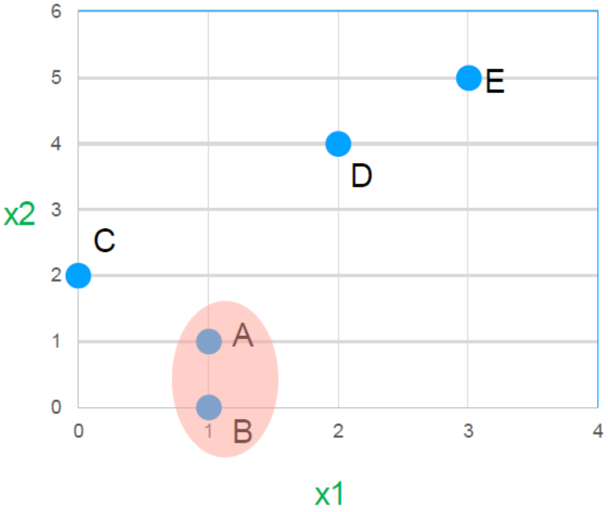


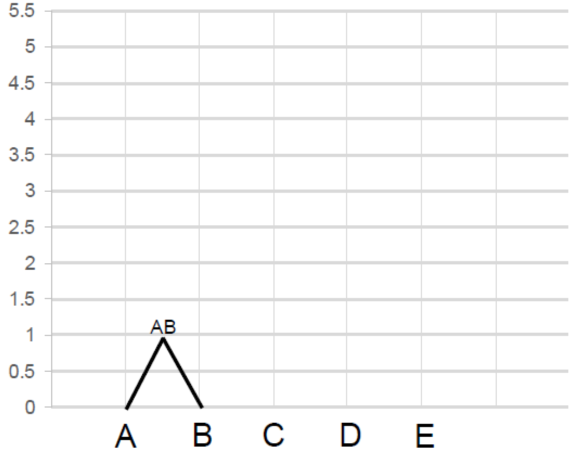
Dendrogram

Distance based on average point

	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0

	(A,B)	C	D	E
(A,B)	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



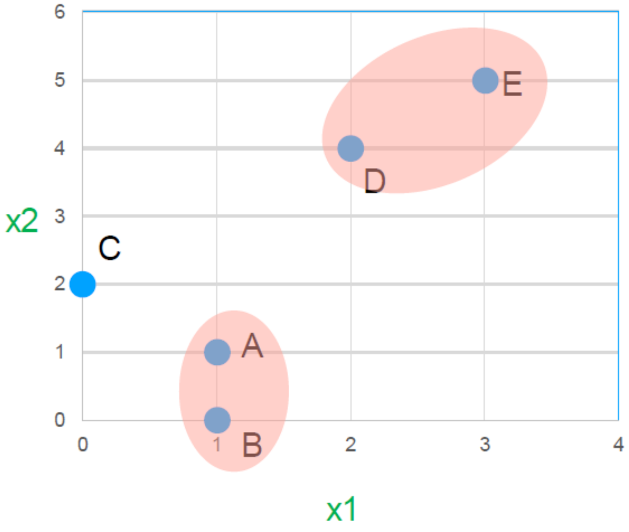


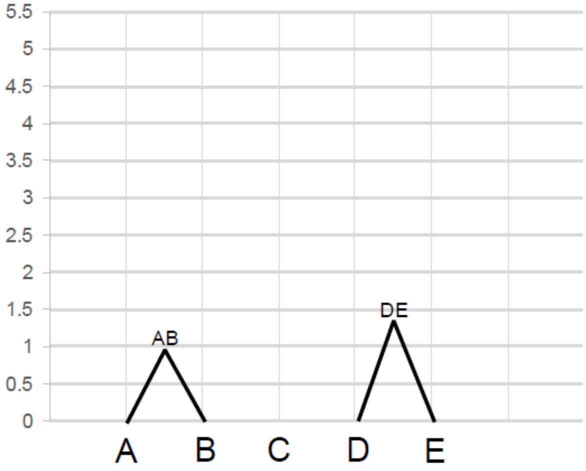
Dendrogram

Distance based on average point

	(A,B)	C	D	E
(A,B)	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0

	(A,B)	C	(D,E)
(A,B)	0	1.8	4.25
C	1.8	0	3.5
(D,E)	4.25	3.5	0



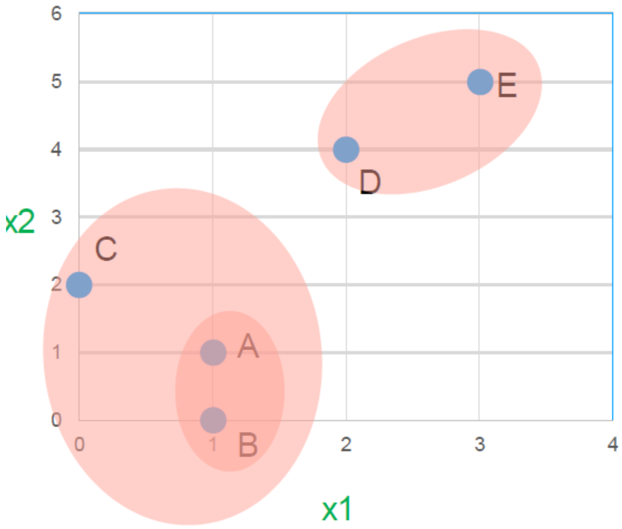


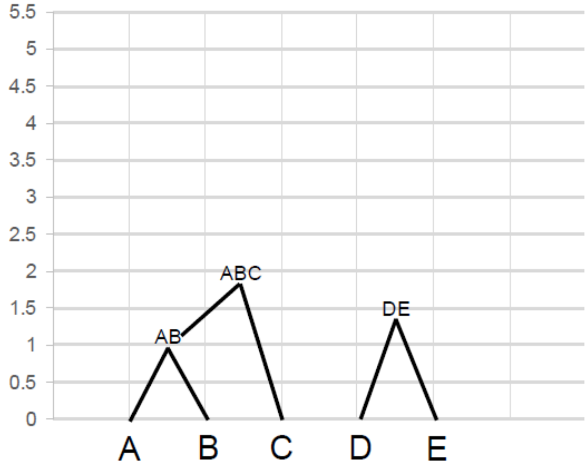
Dendrogram

Distance based on average point

	(A,B)	C	(D,E)
(A,B)	0	1.8	4.25
C	1.8	0	3.5
(D,E)	4.25	3.5	0

	((A,B),C)	(D,E)
((A,B),C)	0	3.875
(D,E)	3.875	0



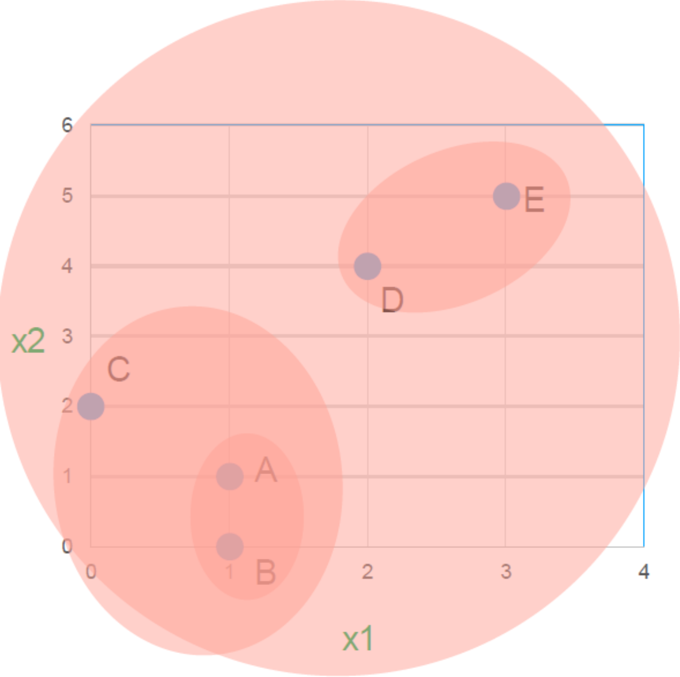


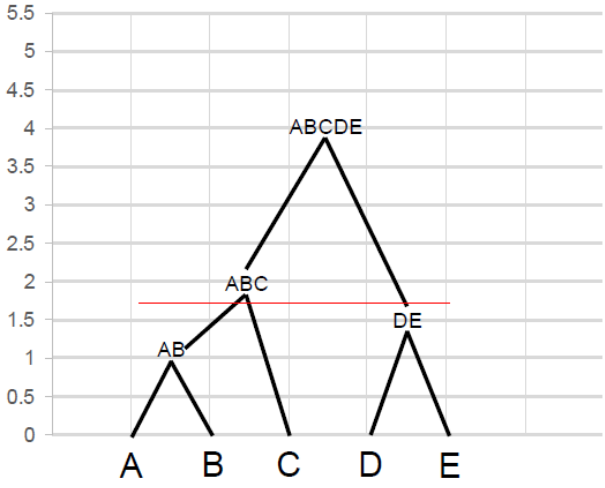
Dendrogram

Distance based on average point

	((A,B),C)	(D,E)
((A,B),C)	0	3.875
(D,E)	3.875	0

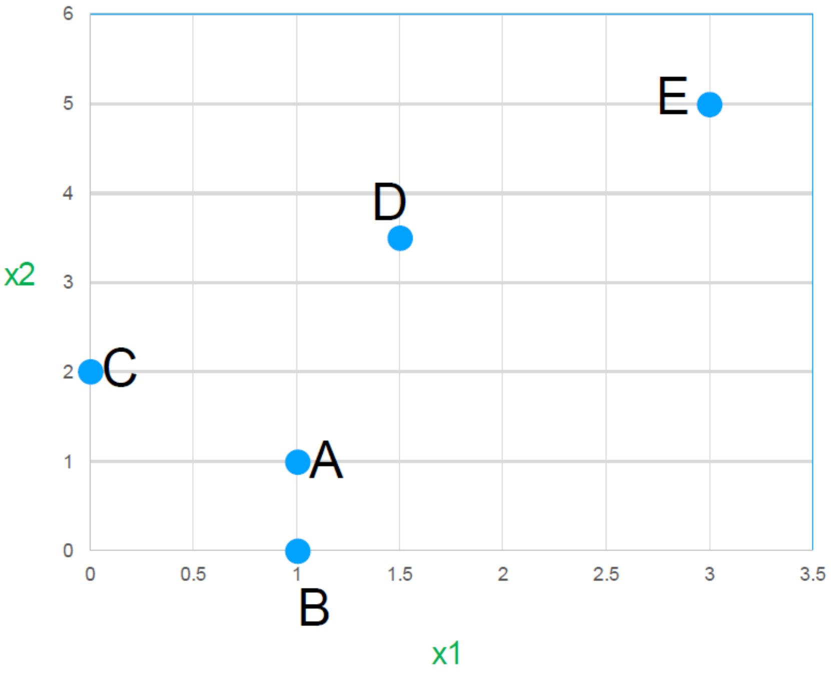
	(((A,B),C),(D,E))
(((A,B),C),(D,E))	0





Dendrogram

i	X1	X2
A	1	1
B	1	0
C	0	2
D	1.5	3.5
E	3	5



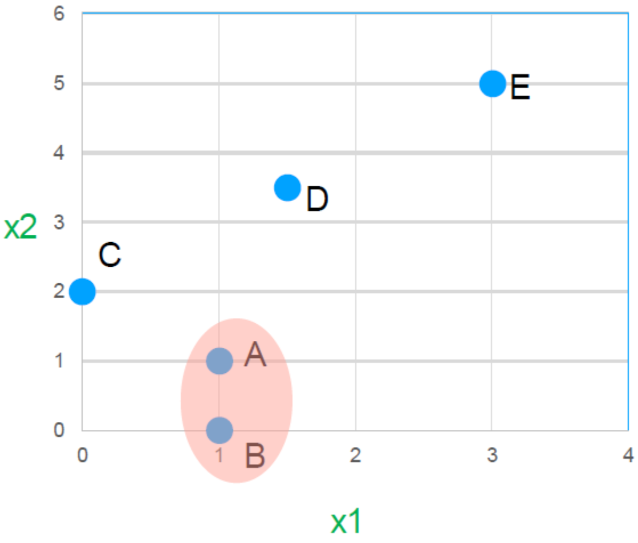
EUCLIDEAN DISTANCE

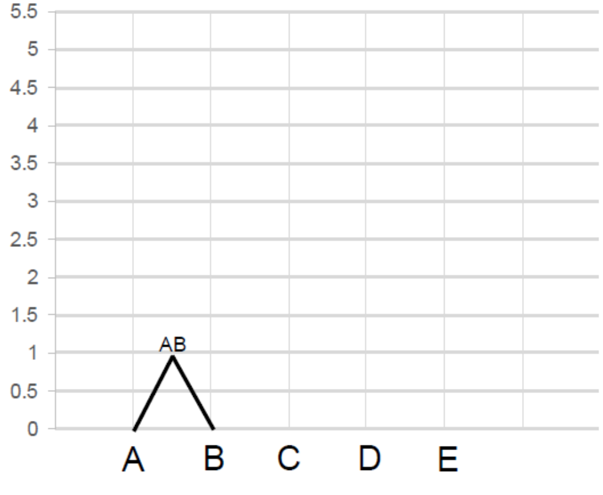
	A	B	C	D	E
A	0	1	1.4	2.55	4.5
B	1	0	2.2	3.53	5.4
C	1.4	2.2	0	2.12	4.2
D	2.55	3.53	2.12	0	2.12
E	4.5	5.4	4.2	2.12	0

Distance based on single link

	A	B	C	D	E
A	0	1	1.4	2.55	4.5
B	1	0	2.2	3.53	5.4
C	1.4	2.2	0	2.12	4.2
D	2.55	3.53	2.12	0	2.12
E	4.5	5.4	4.2	2.12	0

	(A,B)	C	D	E
(A,B)	0	1.4	2.55	4.5
C	1.4	0	2.12	4.2
D	2.55	2.12	0	2.12
E	4.5	4.2	2.12	0



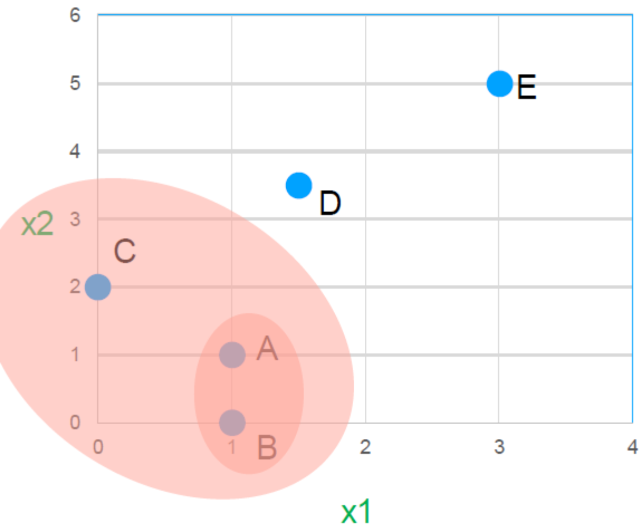


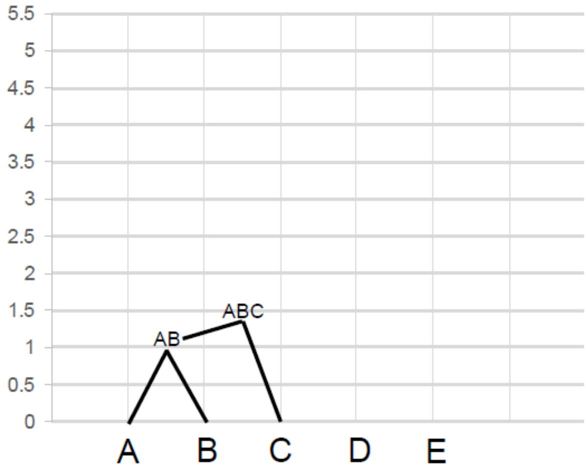
Dendrogram

Distance based on single link

	(A,B)	C	D	E
(A,B)	0	1.4	2.55	4.5
C	1.4	0	2.12	4.2
D	2.55	2.12	0	2.12
E	4.5	4.2	2.12	0

	(A,B),C	D	E
(A,B),C	0	2.12	4.2
D	2.12	0	2.12
E	4.2	2.12	0



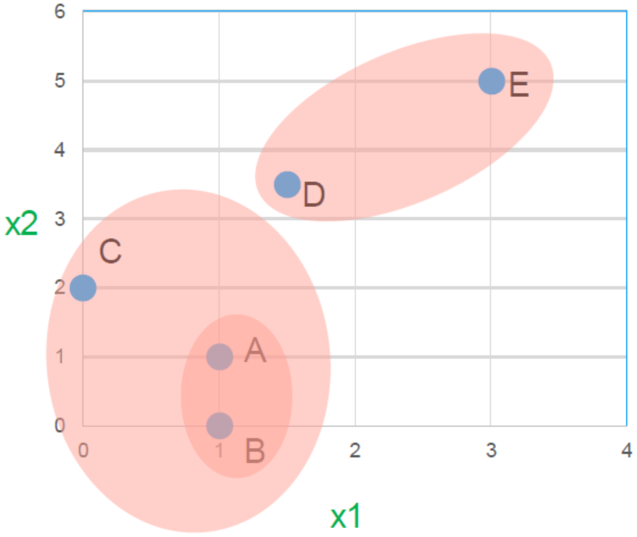


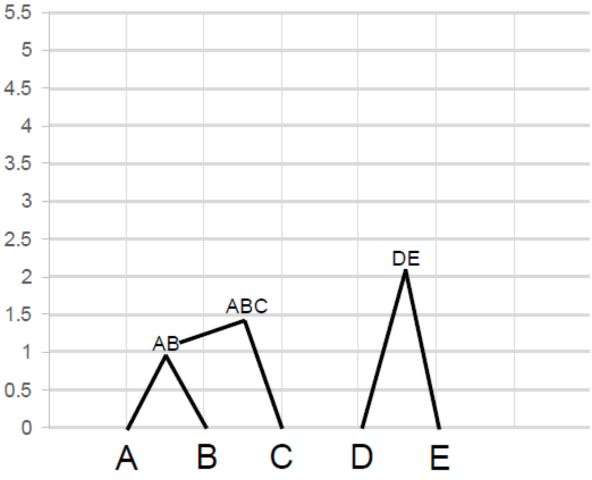
Dendrogram

Distance based on single link

	(A,B), C	D	E
(A,B), C	0	2.12	4.2
D	2.12	0	2.12
E	4.2	2.12	0

	((A,B),C)	(D,E)
((A,B),C)	0	2.12
(D,E)	2.12	0



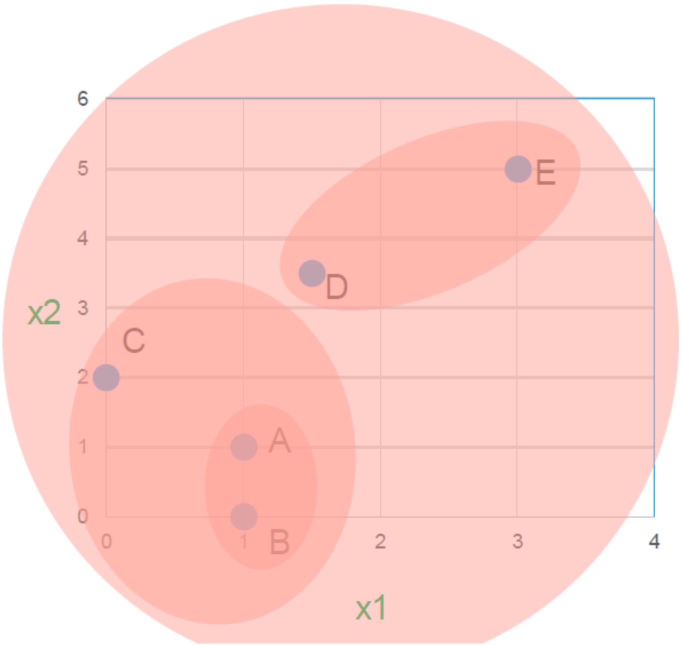


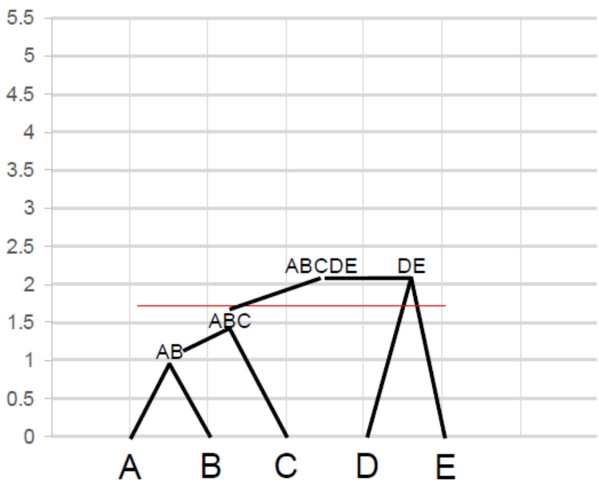
Dendrogram

Distance based on single link

	((A,B),C)	(D,E)
((A,B),C)	0	2.12
(D,E)	2.12	0

	(((A,B),C),(D,E))
(((A,B),C),(D,E))	0



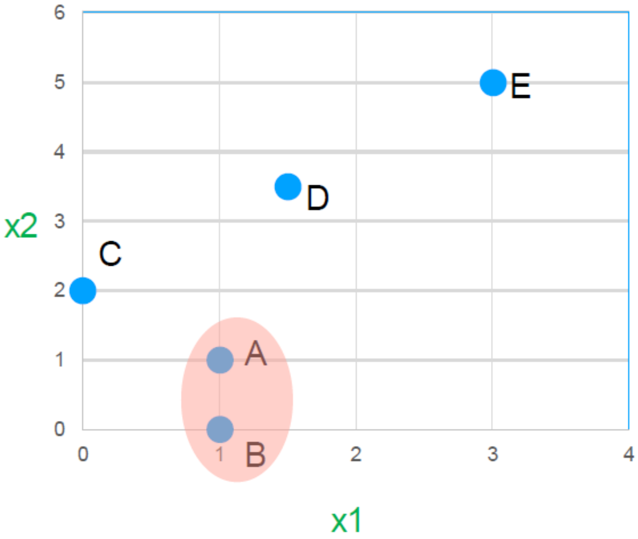


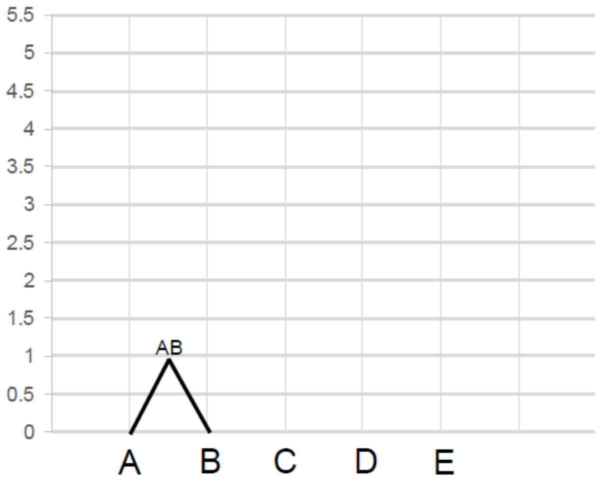
Dendrogram

Distance based on complete link

	A	B	C	D	E
A	0	1	1.4	2.55	4.5
B	1	0	2.2	3.53	5.4
C	1.4	2.2	0	2.12	4.2
D	2.55	3.53	2.12	0	2.12
E	4.5	5.4	4.2	2.12	0

	(A,B)	C	D	E
(A,B)	0	2.2	3.55	5.4
C	2.2	0	2.12	4.2
D	3.55	2.12	0	2.12
E	5.4	4.2	2.12	0



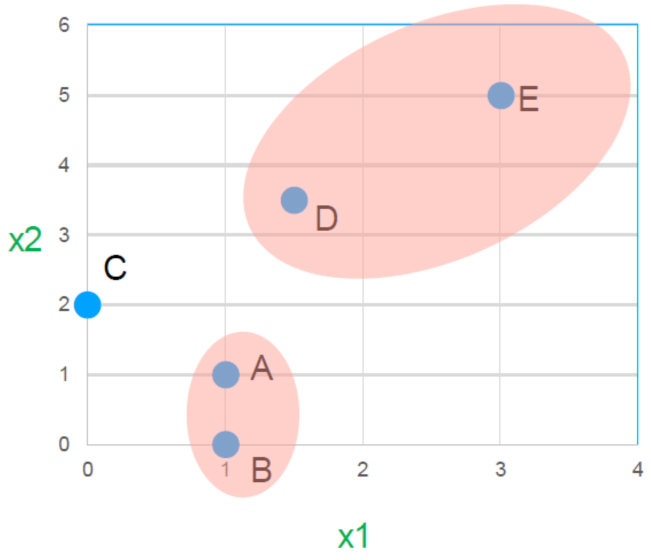


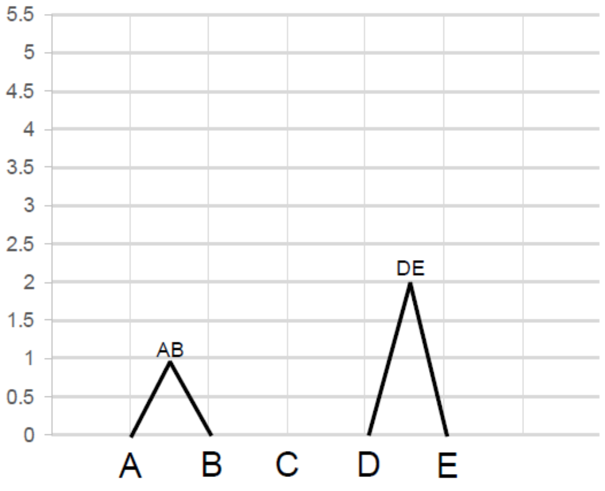
Dendrogram

Distance based on complete link

	(A,B)	C	D	E
(A,B)	0	2.2	3.55	5.4
C	2.2	0	2.12	4.2
D	3.55	2.12	0	2.12
E	5.4	4.2	2.12	0

	(A,B)	C	(D,E)
(A,B)	0	2.2	5.4
C	2.2	0	4.2
(D,E)	5.4	4.2	0



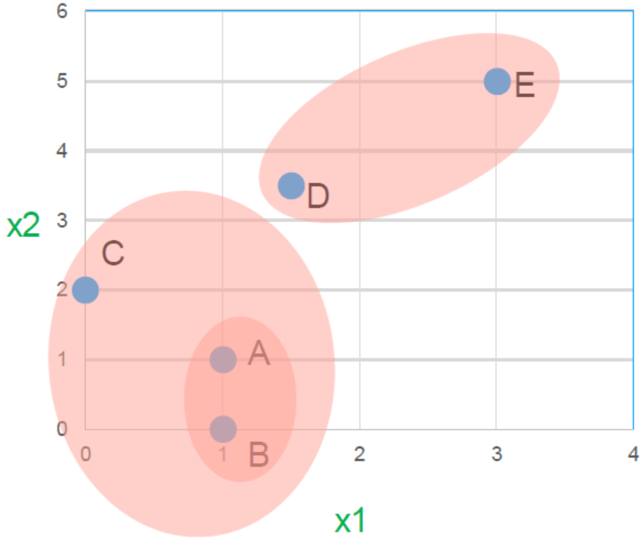


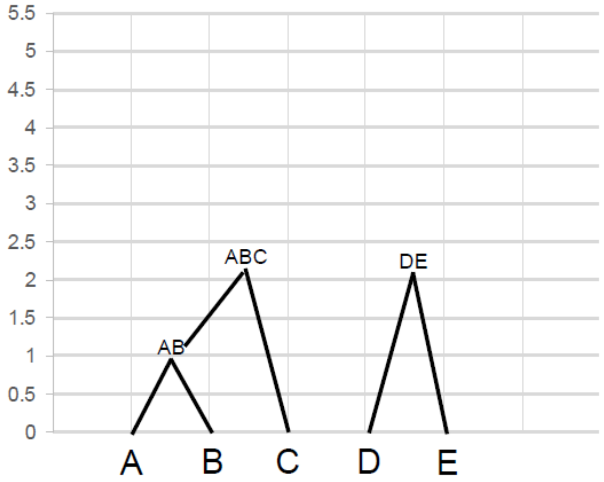
Dendrogram

Distance based on complete link

	(A,B)	C	(D,E)
(A,B)	0	2.2	5.4
C	2.2	0	4.2
(D,E)	5.4	4.2	0

	((A,B),C)	(D,E)
((A,B),C)	0	5.4
(D,E)	5.4	0



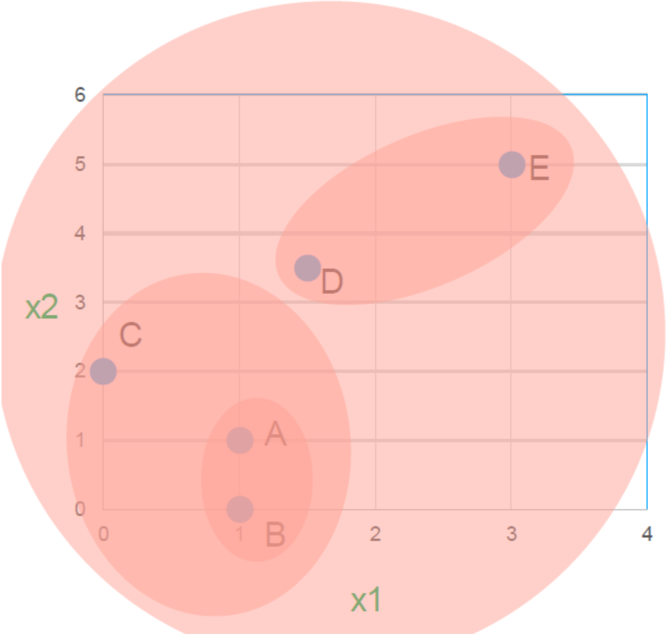


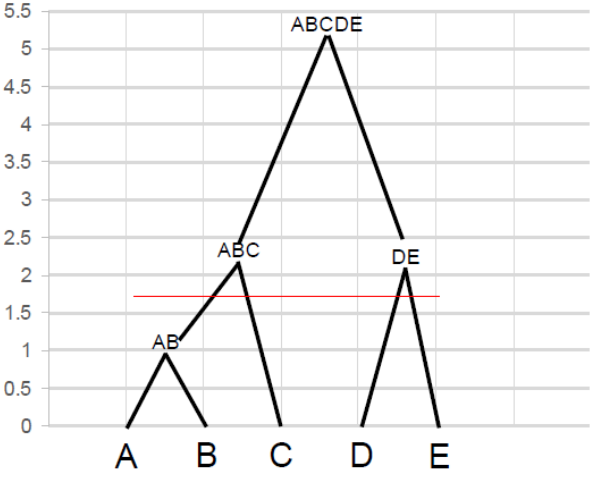
Dendrogram

Distance based on complete link

	((A,B),C)	(D,E)
((A,B),C)	0	5.4
(D,E)	5.4	0

	(((A,B),C),(D,E))
(((A,B),C),(D,E))	0

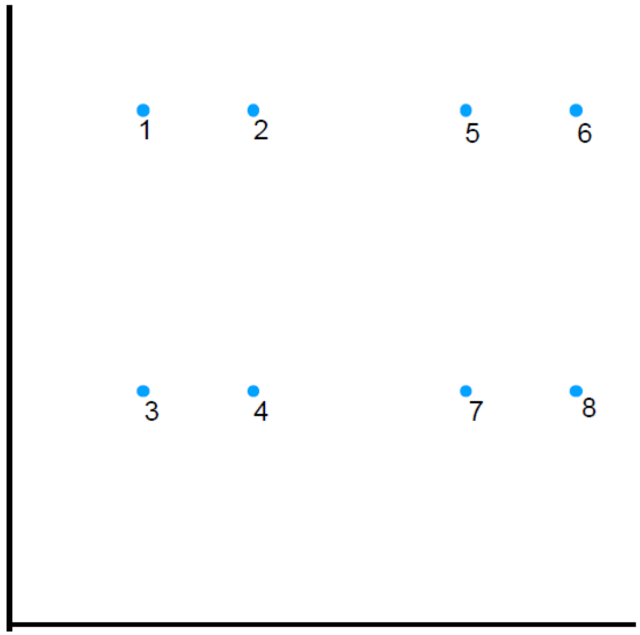




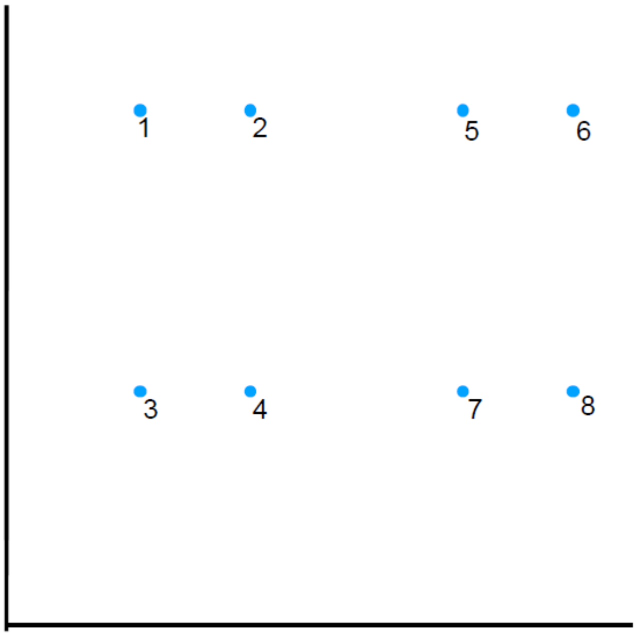
Dendrogram

Another example

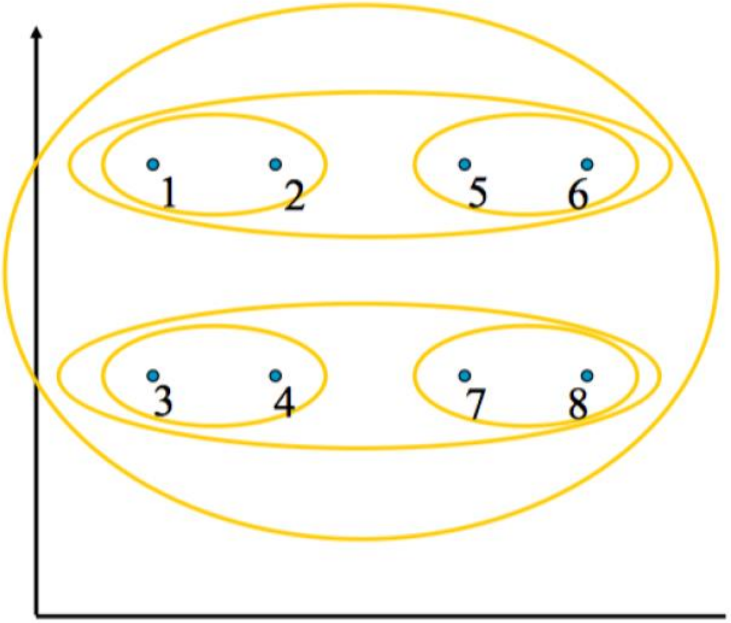
Single Link clustering (closest pair)



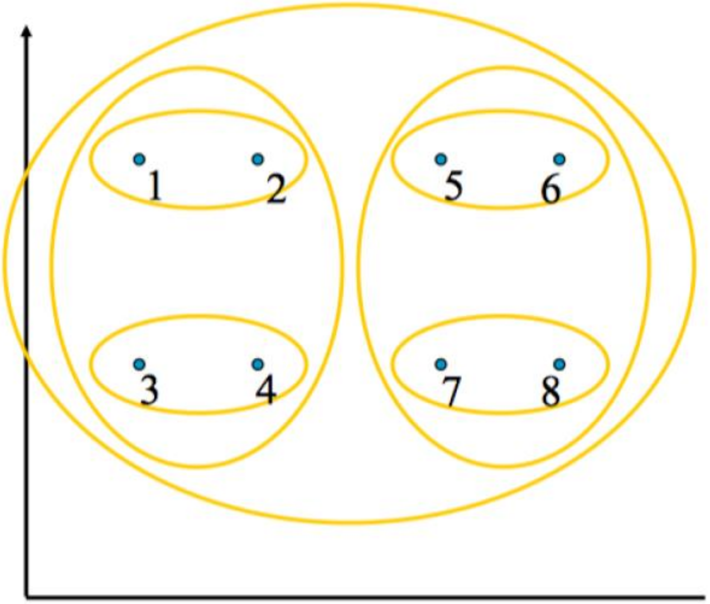
Complete Link clustering (Farthest pair)



Single Link clustering (closest pair)

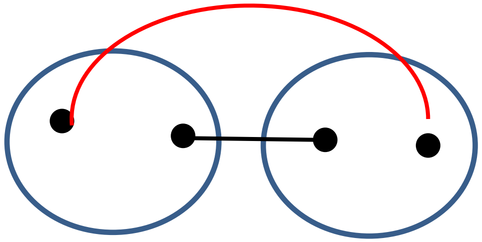


Complete Link clustering (Farthest pair)



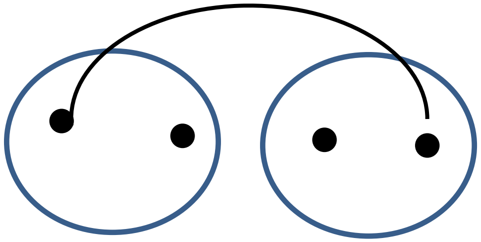
A short summary of the two links

- Single link vs. Full link:
  - Single link suffers from chaining: A->B->C, and it only needs to one pair of points to be close when checking two clusters, while full link avoids chaining.
  - Full link suffers from crowding, as it is based on worst-case similarity.



The two parts are far away from each other.

The clusters are too spread out, not compact enough.



Produces "spherical" clusters. The clusters are compact, but not far enough apart.

Take-home messages

- Hierarchical clustering overview
- Bottom-up and top-down clustering
- Define distance between clusters
- Comparison between single link and complete link