Recap

Kmeans – Hard Clustering – each datapoint to 1 cluster

GMM – Soft Clustering – each datapoint has a prob. to be assigned to every cluster

– EM

- Unsupervised methods covered so far:

GMM: $p(x, z)$ — x: data, z: label / hidden variable

$$p(Z_k \mid x) = \frac{p(x, z_k)}{p(x)} = \frac{p(x \mid z_k)\, p(z_k)}{p(x)}$$

$p(x \mid z_k) \sim N(\mu, \Sigma_k)$

$p(z_k) = \pi_k$

$z \sim p(z)$

$x \sim p(x \mid z)$

- Generative vs Discriminative models

$p(z \mid x)$ ← computed directly

Logistic Regression

Neural Network

- Common problem between unsupervised methods covered so far

We need to know k before starting

Georgia Tech

Machine Learning CS 4641

Hierarchical Clustering

Nimisha Roy

Lecturer, SCI, College of Computing, Georgia Tech

Director, Online Undergraduate Initiatives

Georgia Tech

Slides adapted from Max Roozbahani, Rod Borela.

Outline

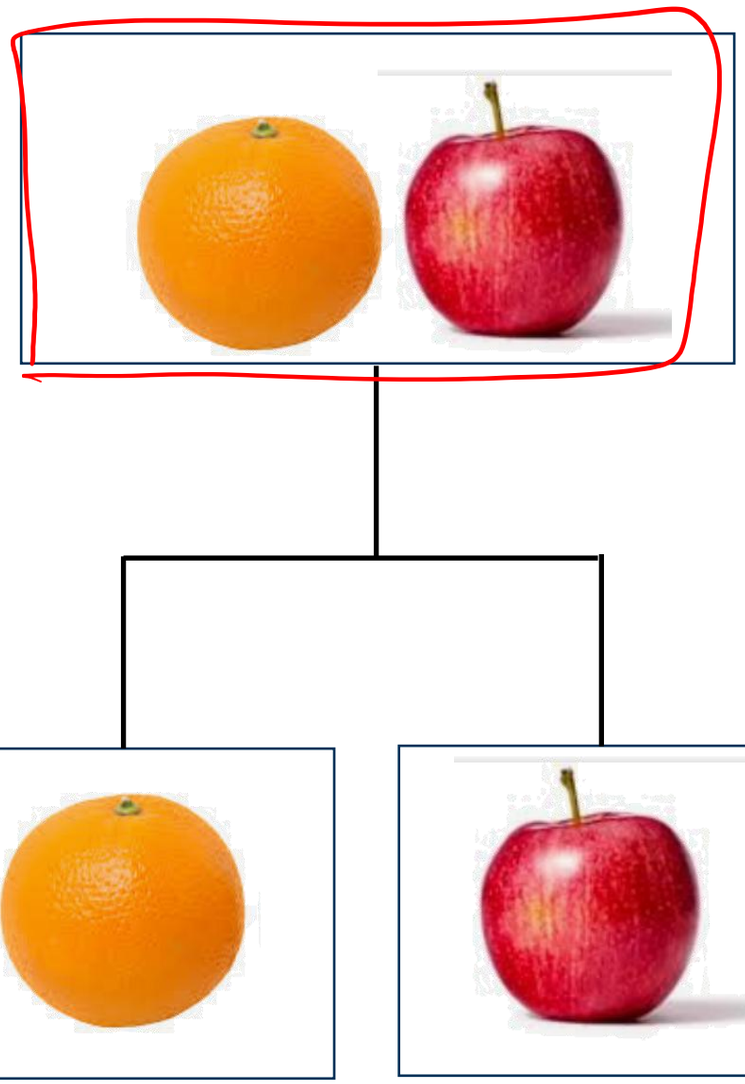
- Overview
- Bottom-Up vs Top-Down Clustering
- Measuring Distance between Clusters

Hierarchical Clustering vs Partitional Clustering

Agglomerative

Divisive

Root Node → 1 cluster

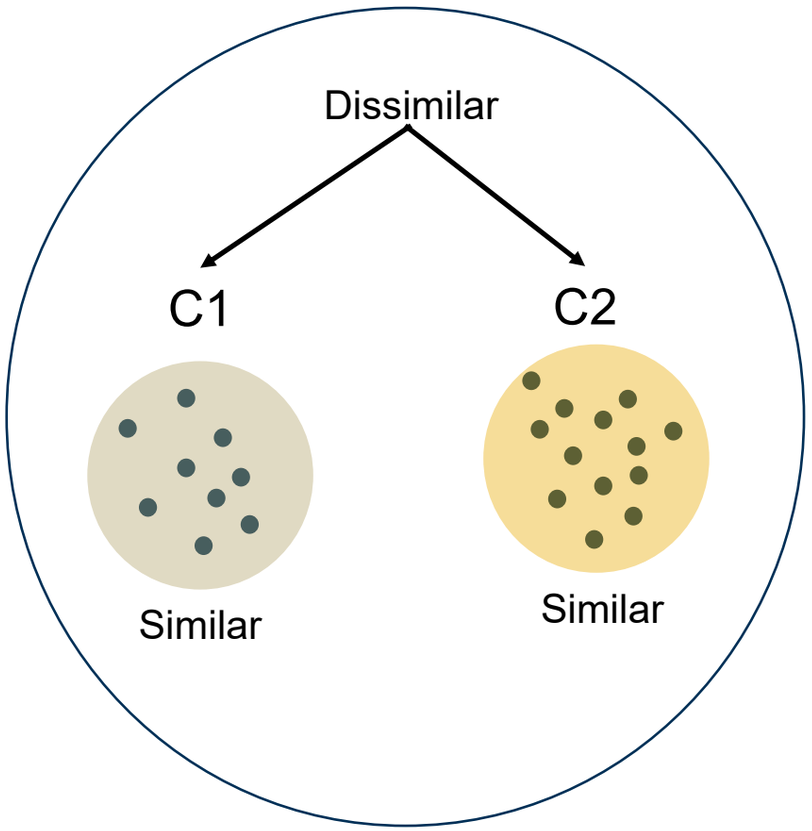


Tree structure (parent-child relationship)

K-Means

decide k=2

tree like structure without knowing k



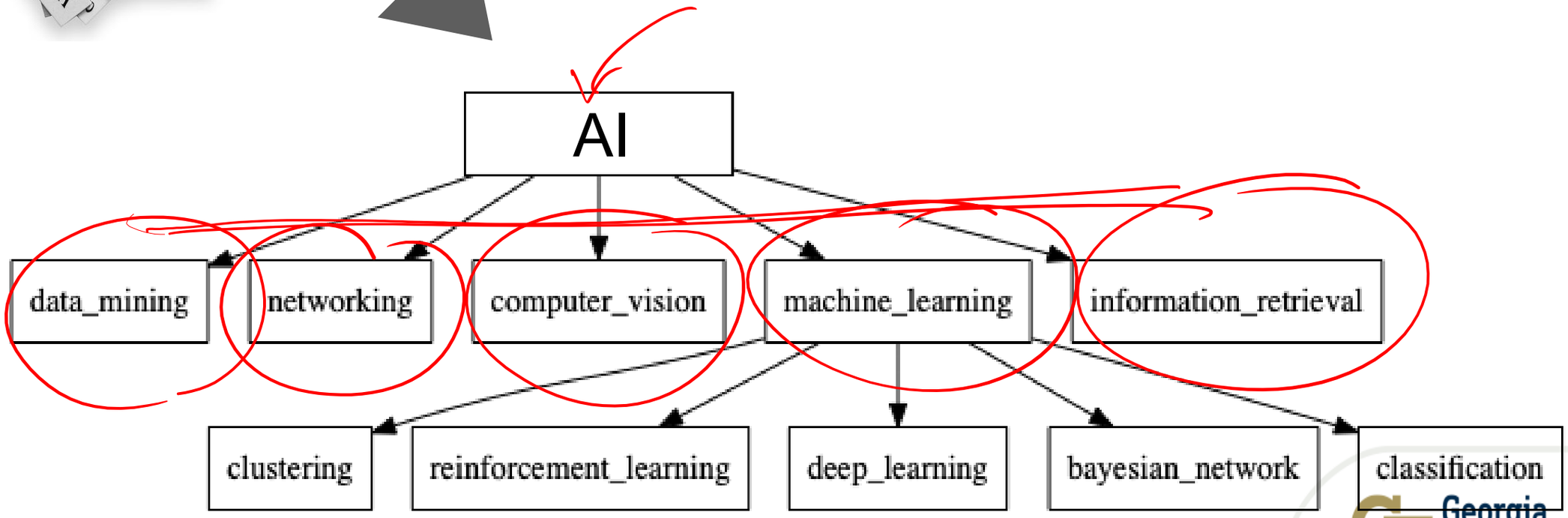
Leaf nodes → each datapoint is its own cluster

Georgia Tech

Hierarchical Clustering

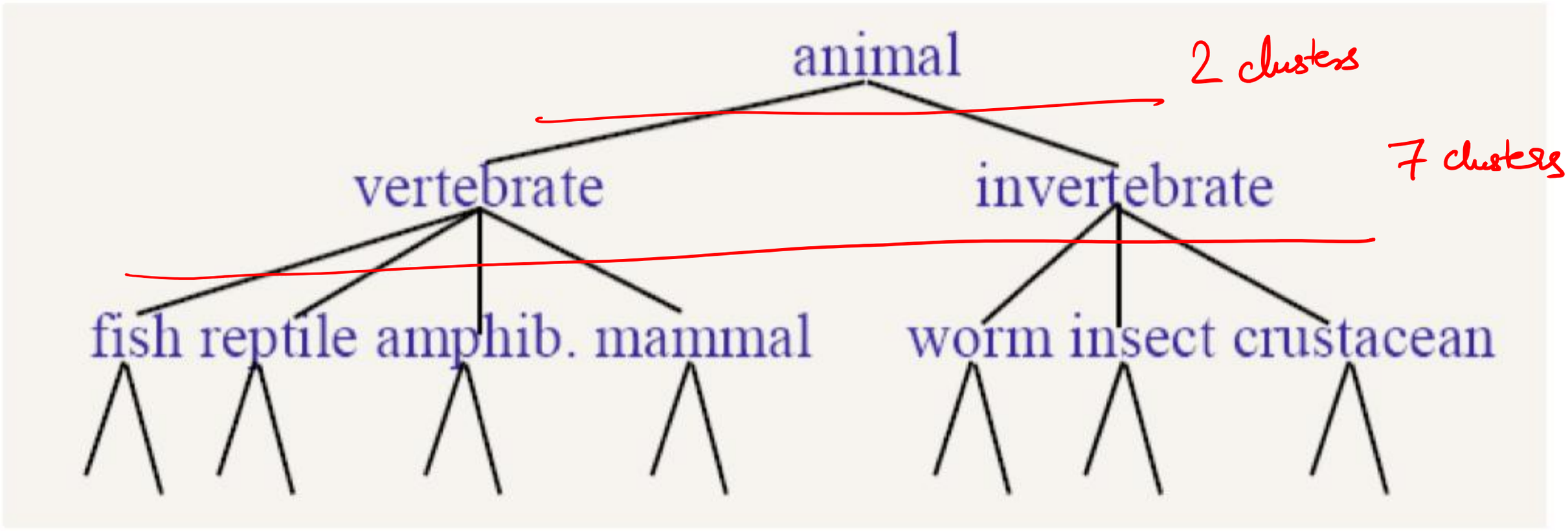
It does not work for very large datasets

- How to organize a set of AI papers into a hierarchy?



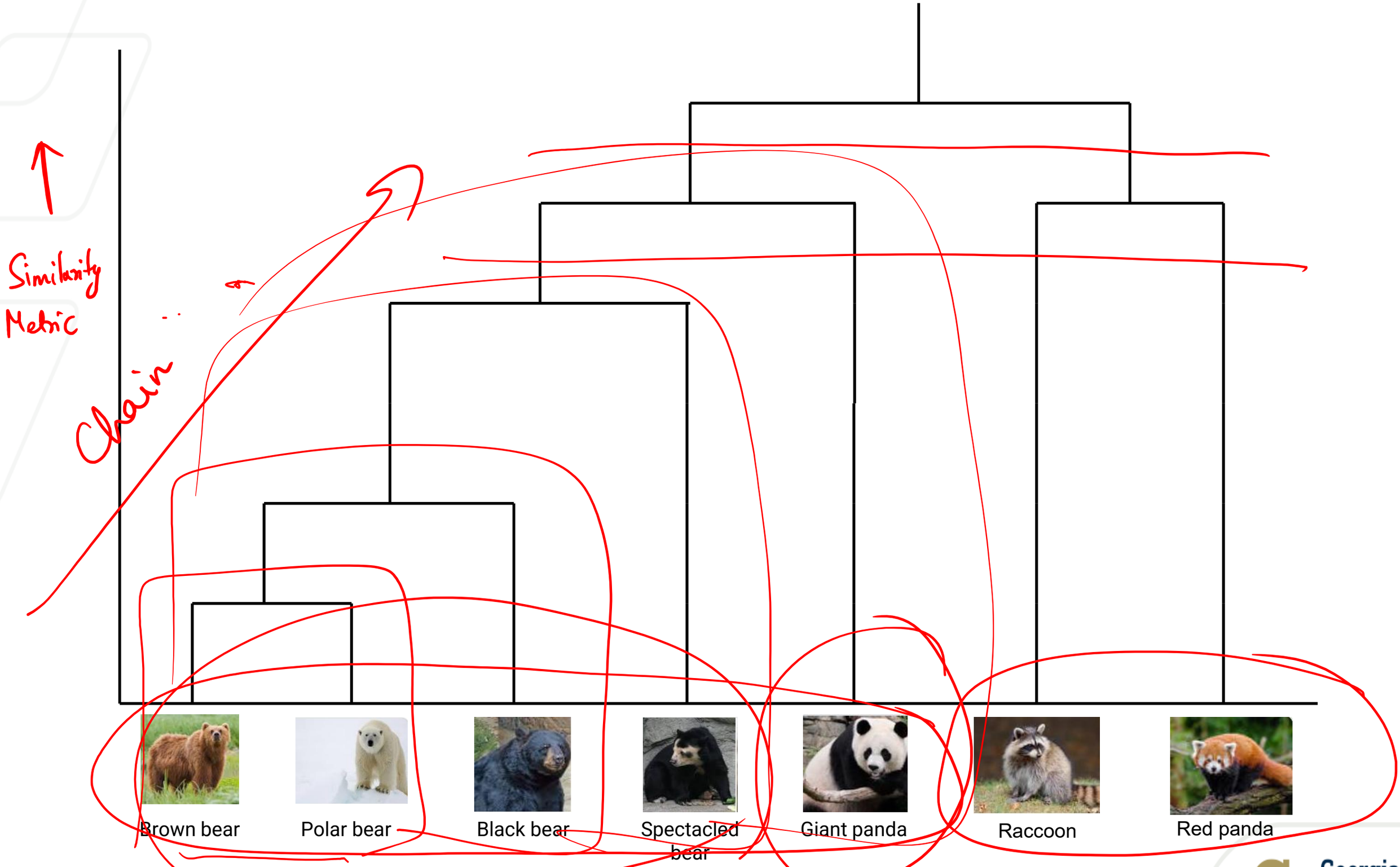
Hierarchical Clustering

- Organize objects into a tree-based hierarchical taxonomy (dendrogram)



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

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

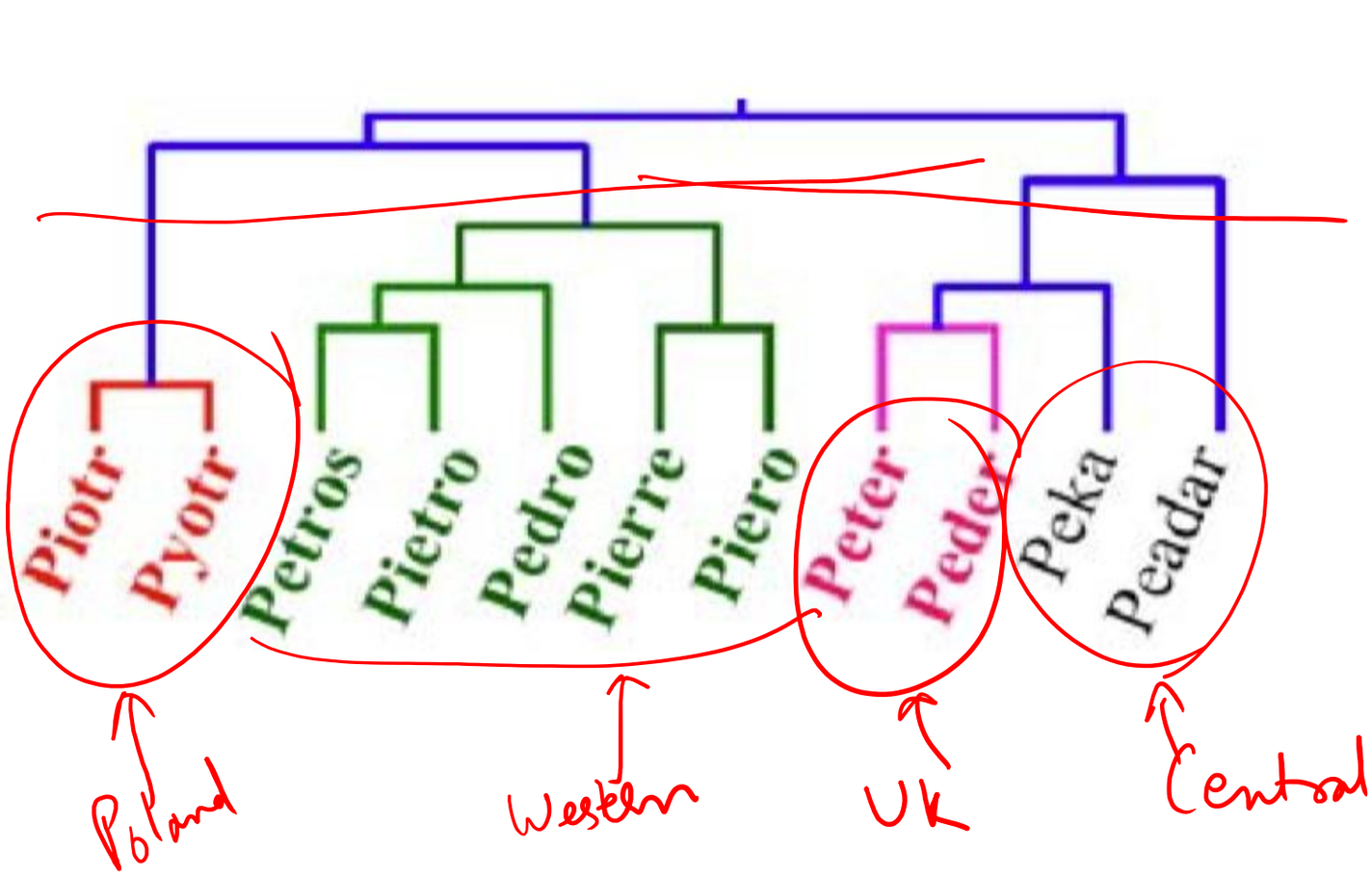


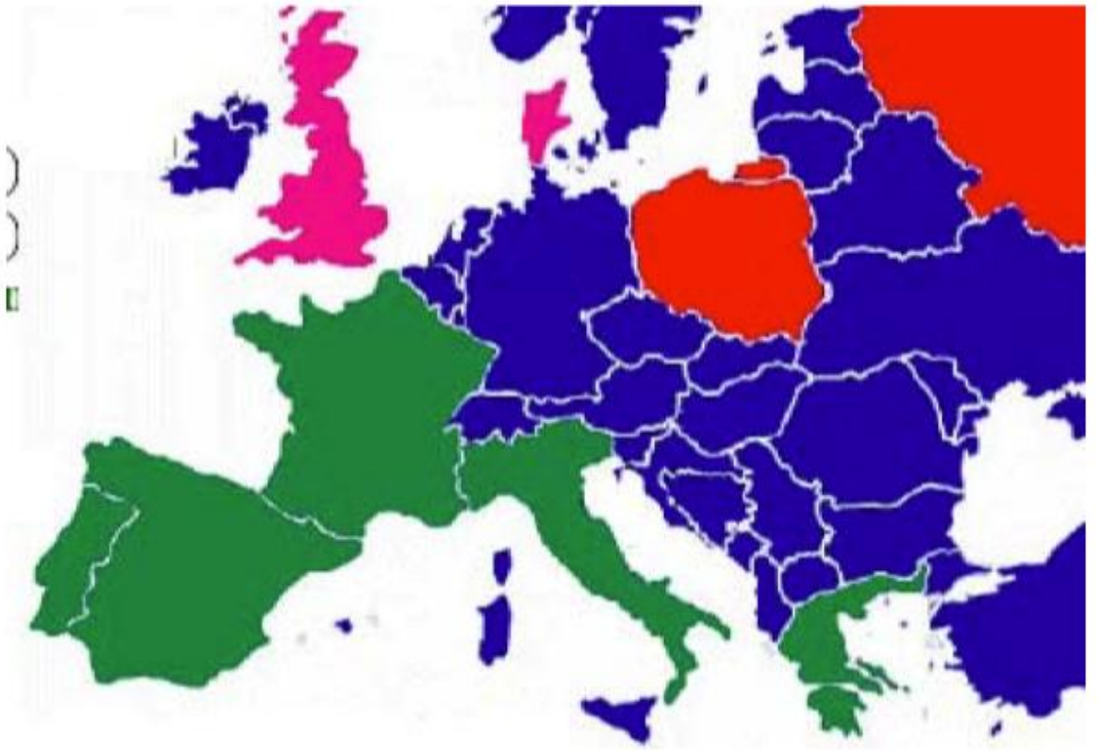
Using Hierarchical clustering, the researchers were able to place the giant pandas closer to bears

Georgia Tech

Hierarchical Clustering

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





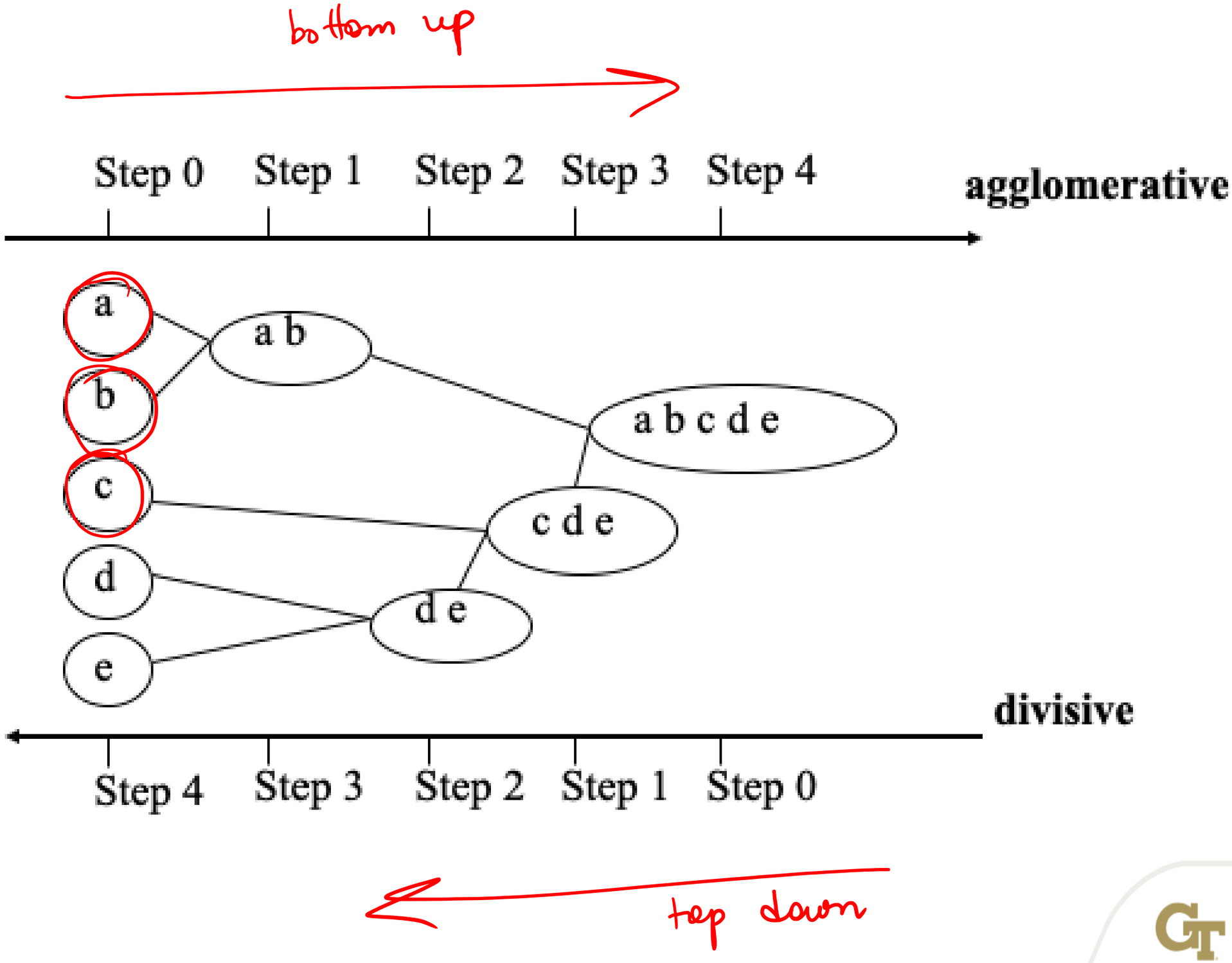
Outline

- Overview
- Bottom-Up vs Top-Down Clustering
- Measuring Distance between Clusters

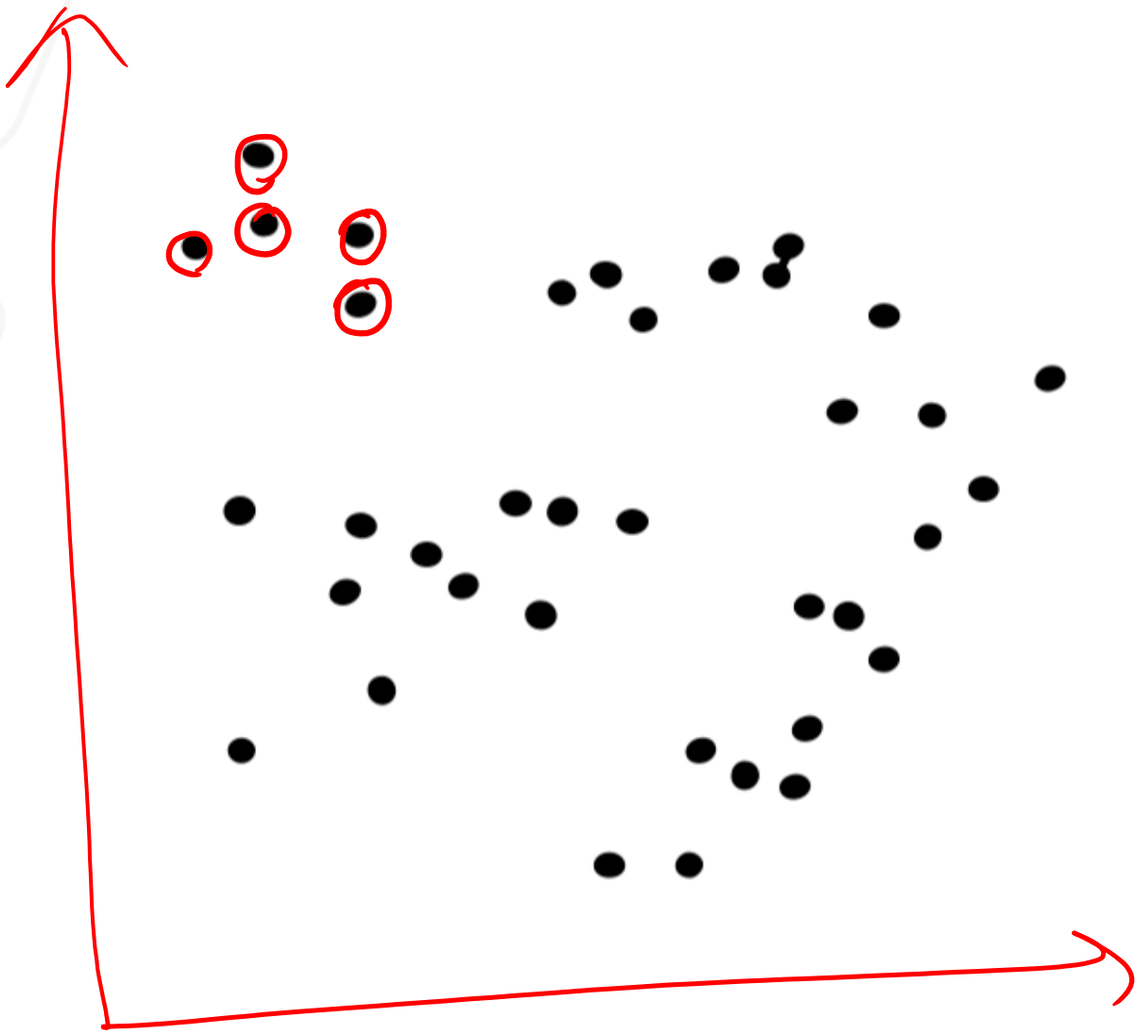
Two Paradigms for Hierarchical Clustering

- 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

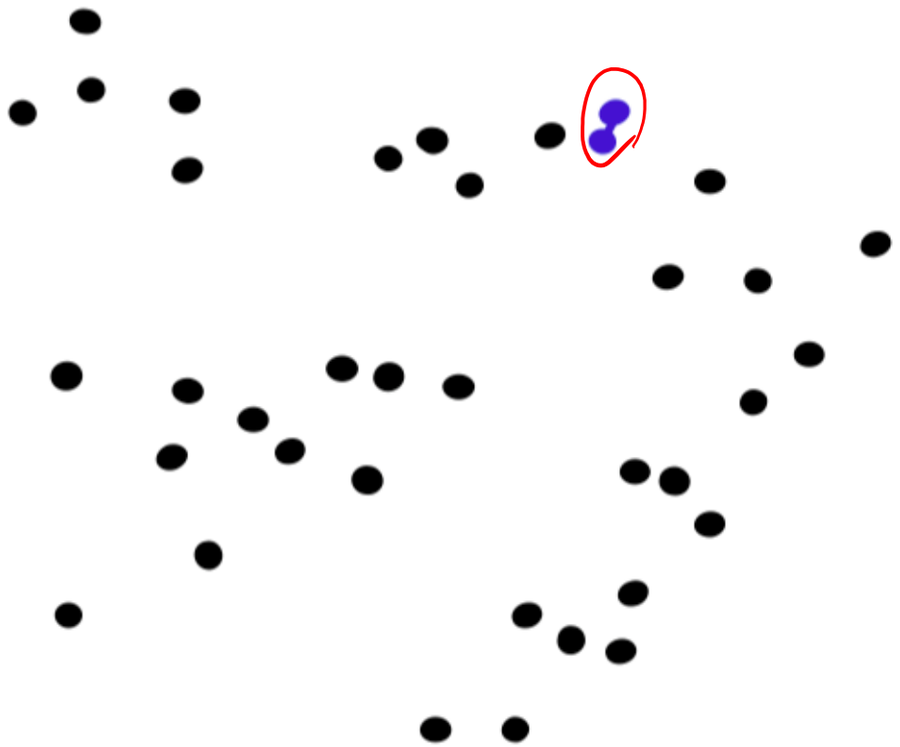


Bottom-Up Agglomerative Clustering



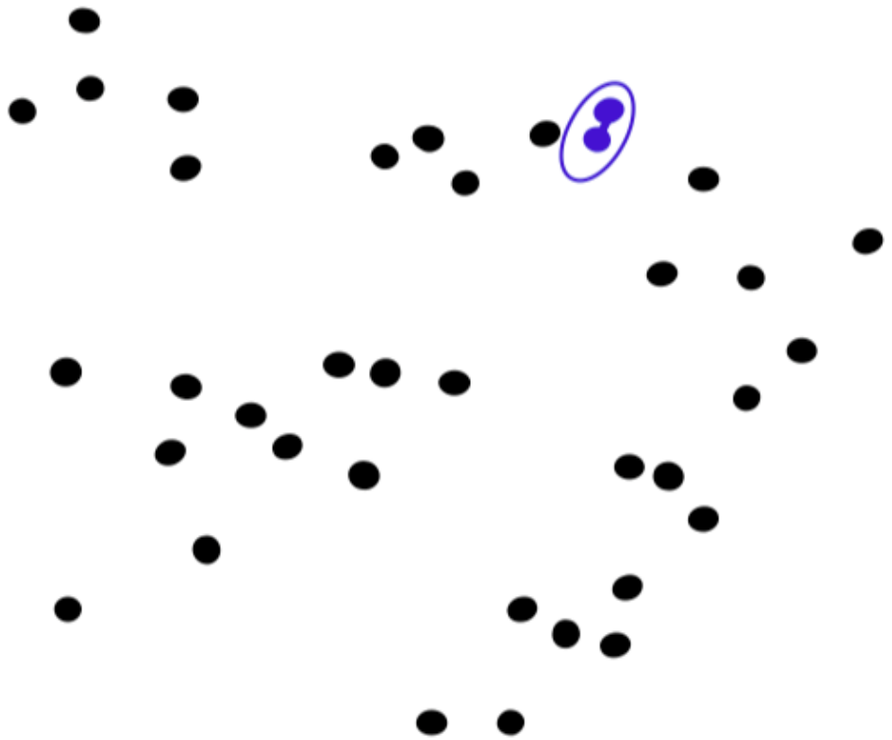
1. Say "Every point is it's own cluster"

Bottom-Up Agglomerative Clustering



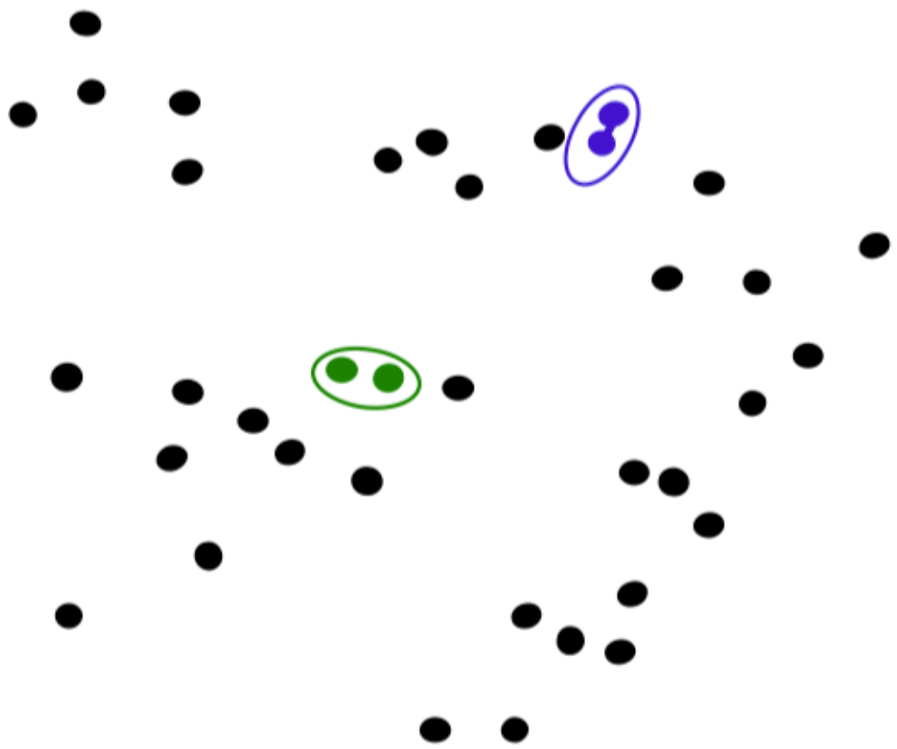
1. Say "Every point is it's own cluster"
2. Find "most similar" pair of clusters

Bottom-Up Agglomerative Clustering

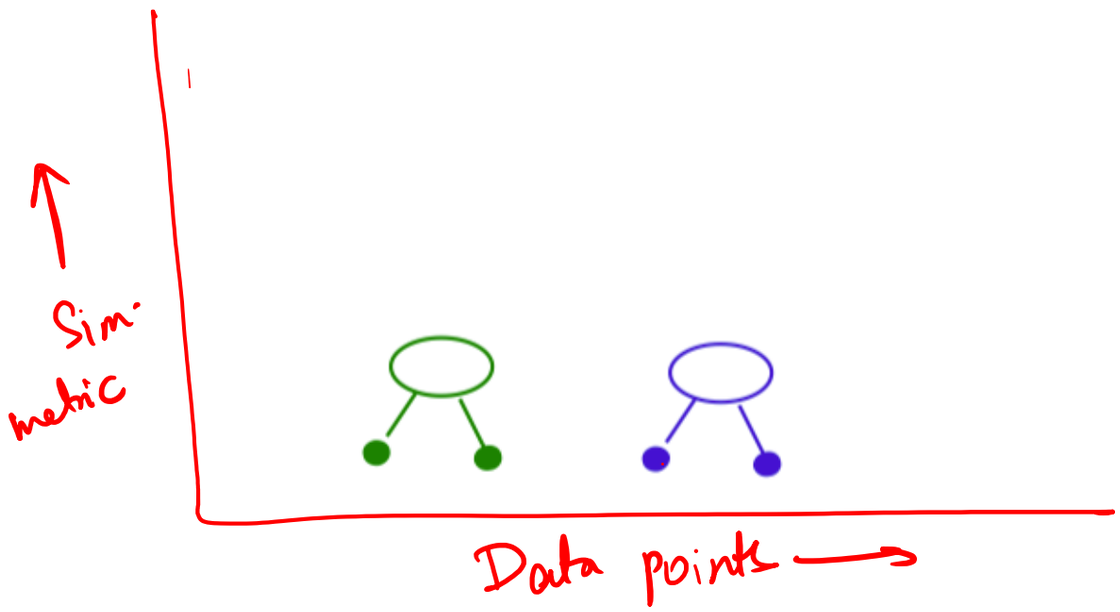


1. Say "Every point is it's own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster

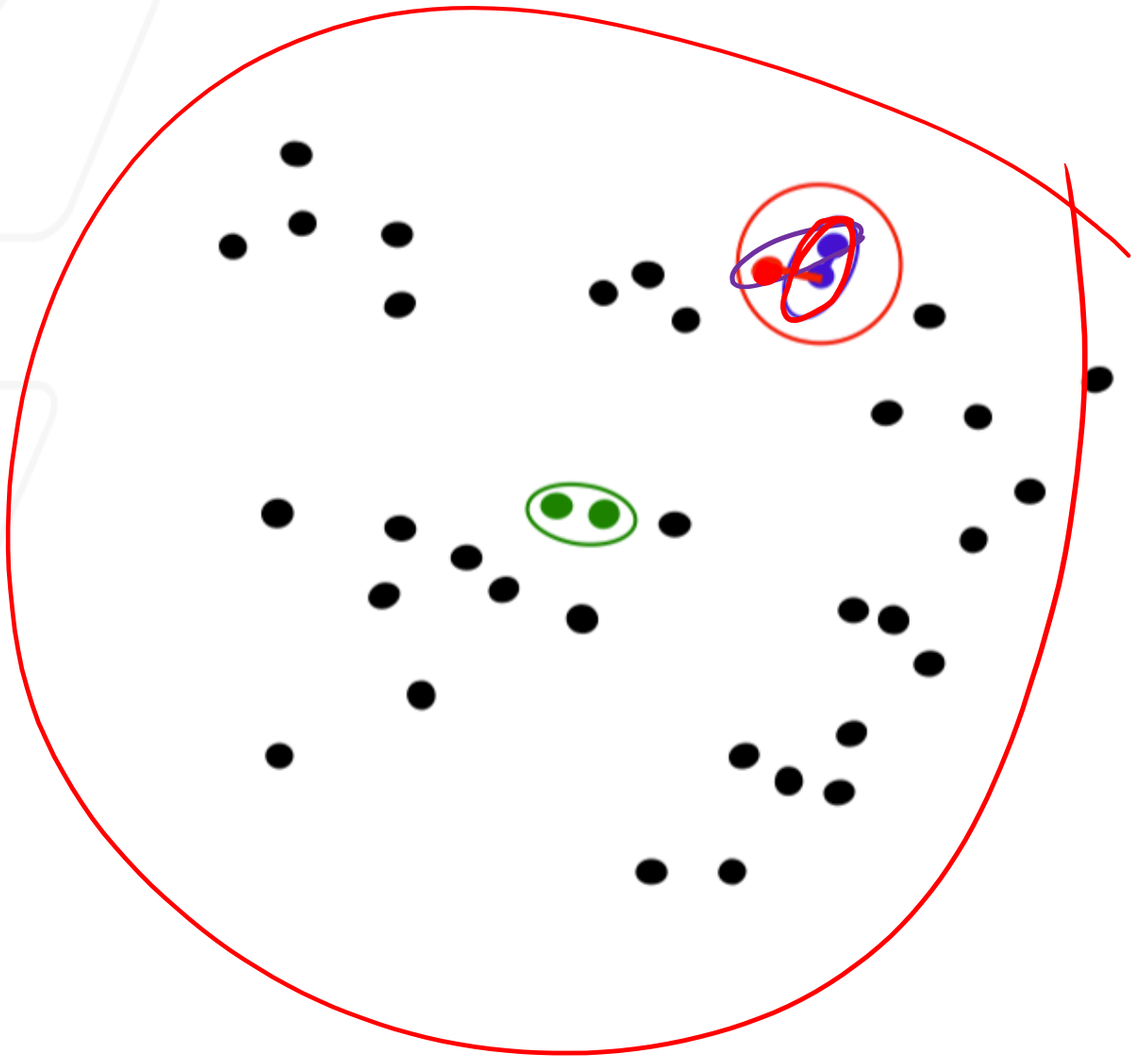
Bottom-Up Agglomerative Clustering



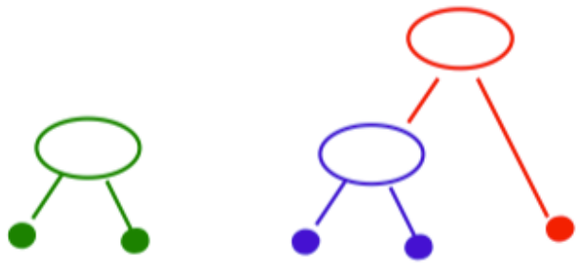
1. Say "Every point is it's own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster
4. Repeat



Bottom-Up Agglomerative Clustering



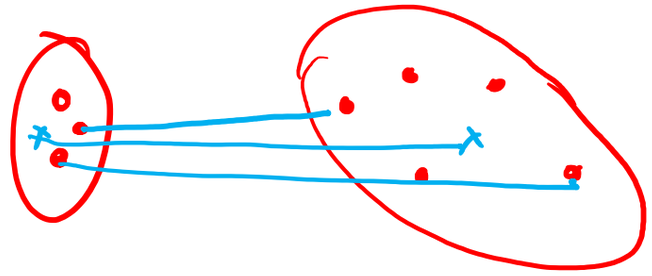
1. Say "Every point is it's own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster
4. Repeat

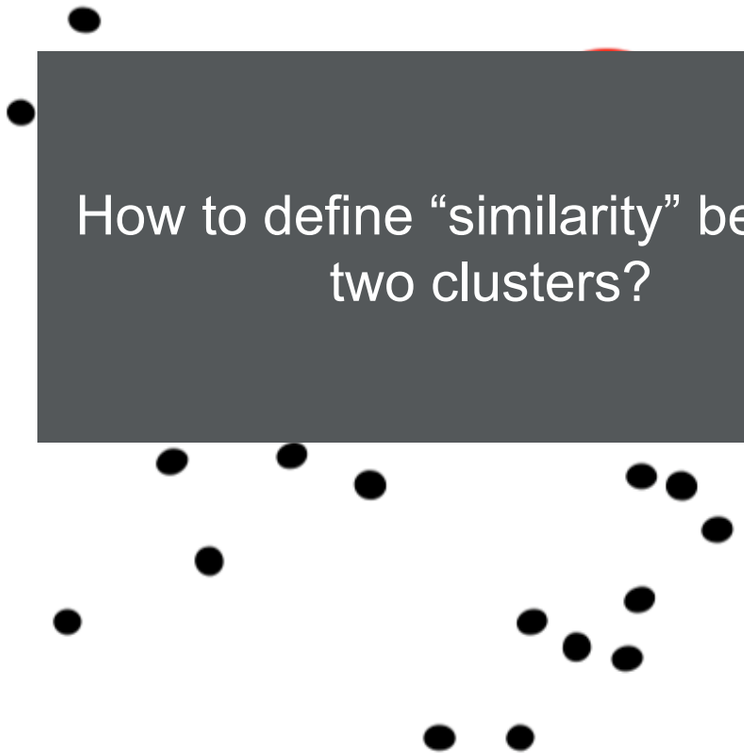


Outline

- Overview
- Bottom-Up vs Top-Down Clustering
- Measuring Distance between Clusters

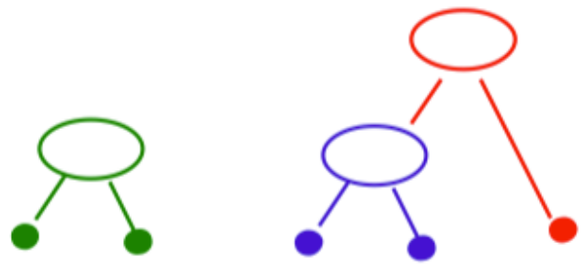
Key Question: Similarity Function





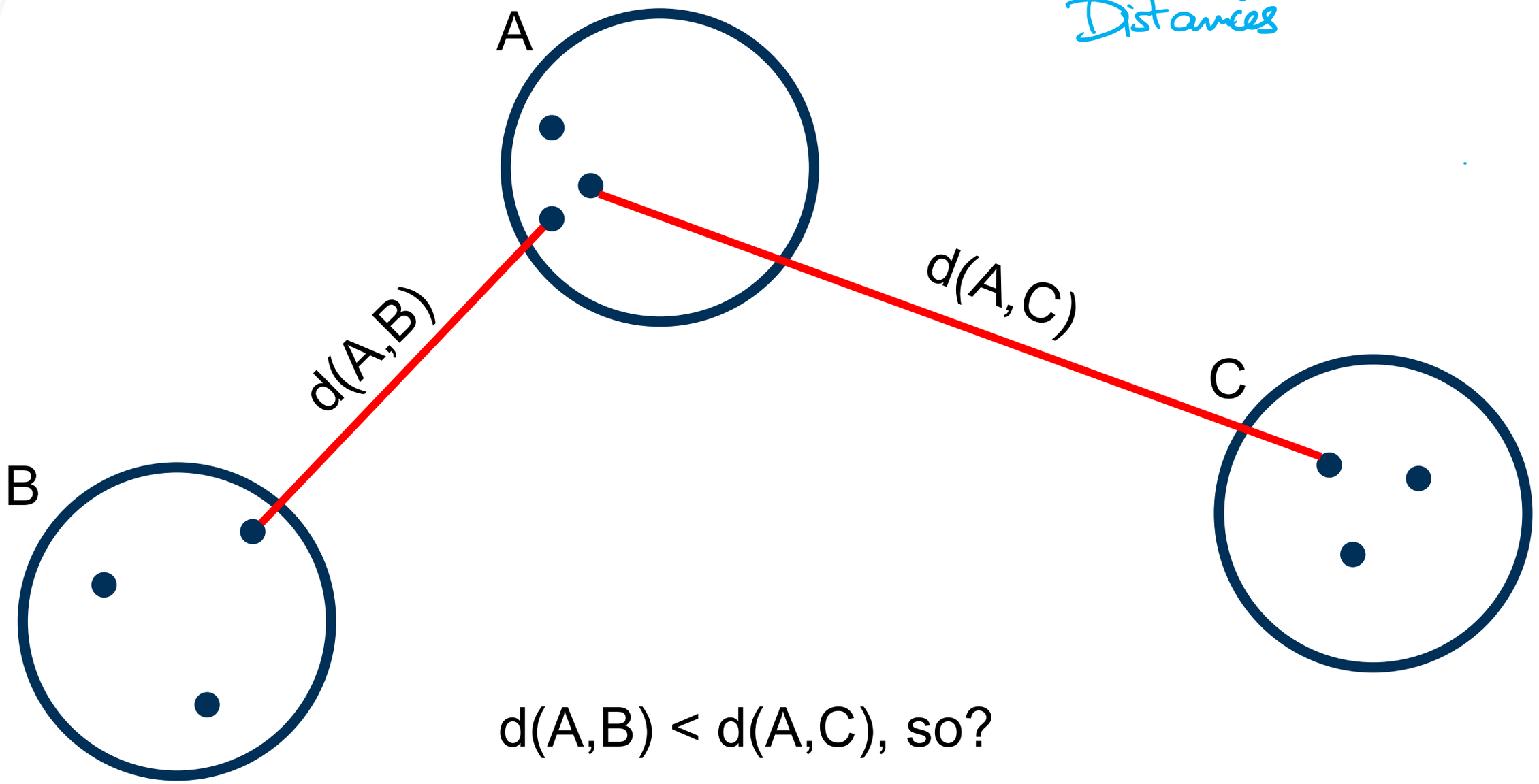
How to define “similarity” between two clusters?

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



I am going to merge A with either B or C. Which one?

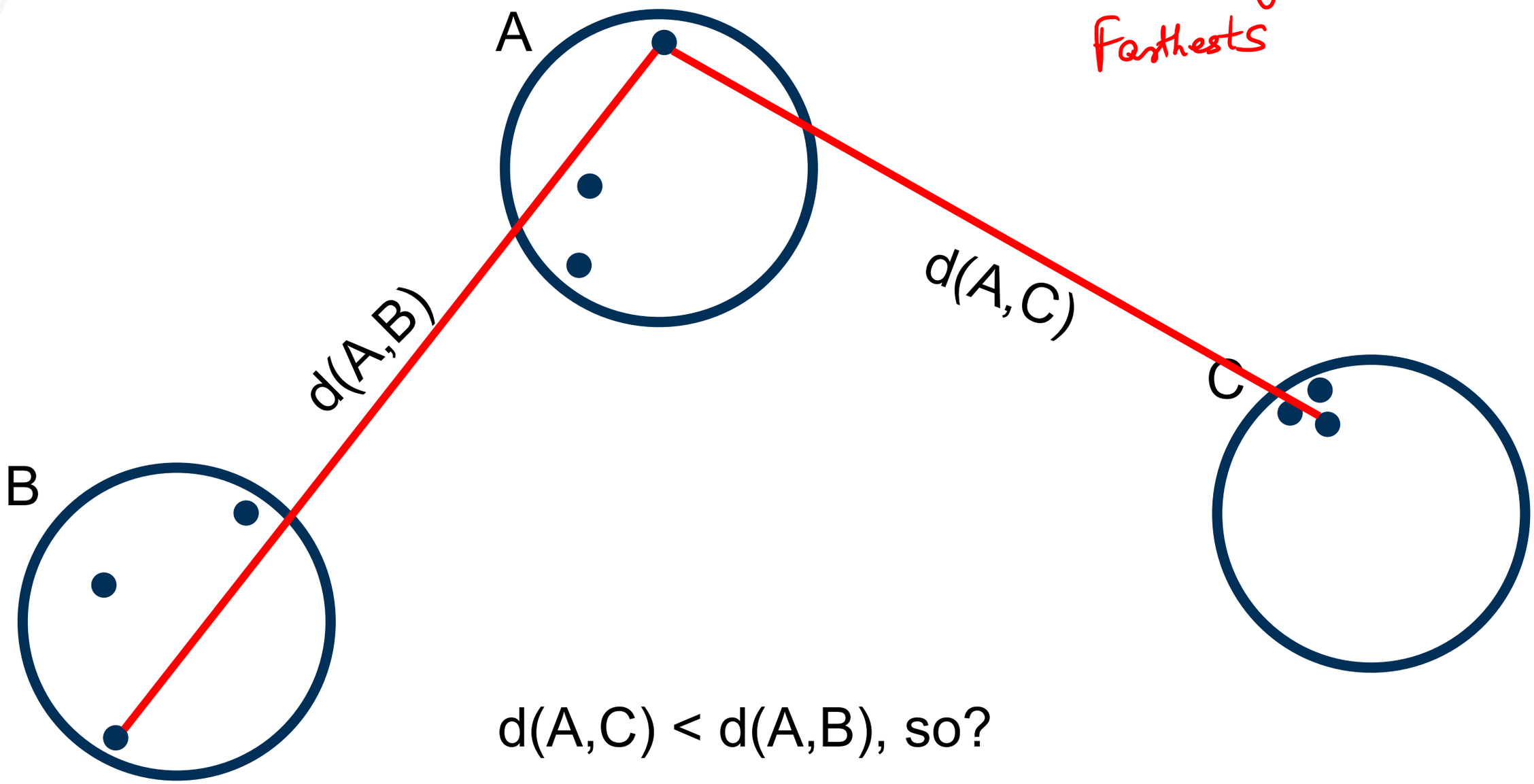
Shortest of Shortest Distances



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

Single Link

I am going to merge A with either B or C. Which one?

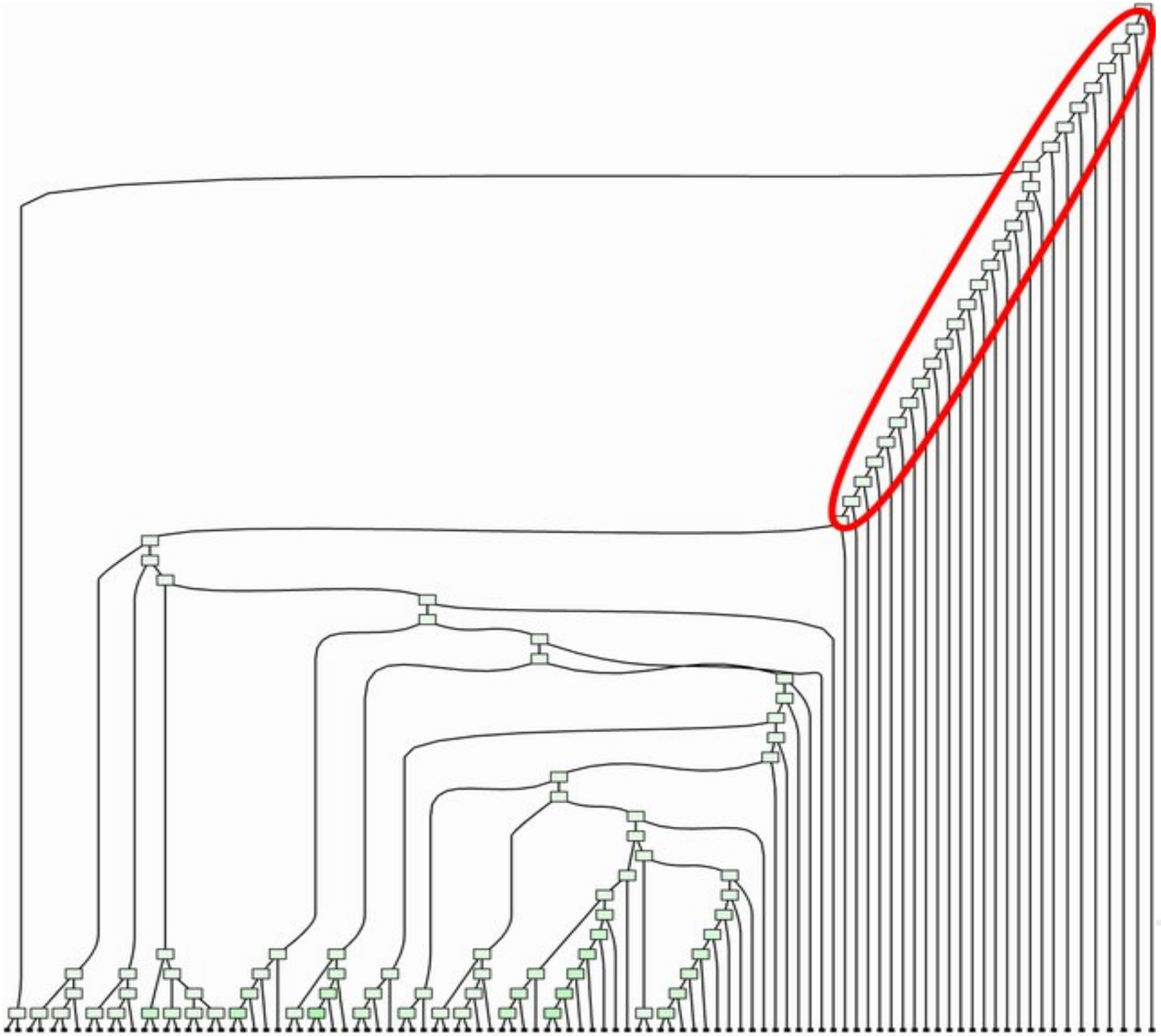


d(A,C) < d(A,B), so?

Complete Link

Single Linkage is prone to chaining

- Clusters get merged because of a sequence of nearby points forming a “chain,” even though the overall clusters are not truly compact or similar.
- So instead of forming tight, well-separated clusters, you get Long, stretched, snake-like clusters.



https://www.researchgate.net/publication/228531785_TNO_Hierarchical_topic_detection_report_at_TDT_2004/figures?lo=1

When is Single Linkage useful?

Single link: A chain of points can be extended for long distances without regard to the overall shape of the emerging cluster. This effect is called *chaining*.

- It is also sensitive to outliers.
- It is faster in general.
- best when d**etecting elongated, irregular shapes.** Example: You're clustering cities along a river. Single-link will connect them in order, capturing the *path-like structure*.
- **When clusters are non-spherical and winding** (like spirals, curves).

When is Complete and Average Linkage useful?

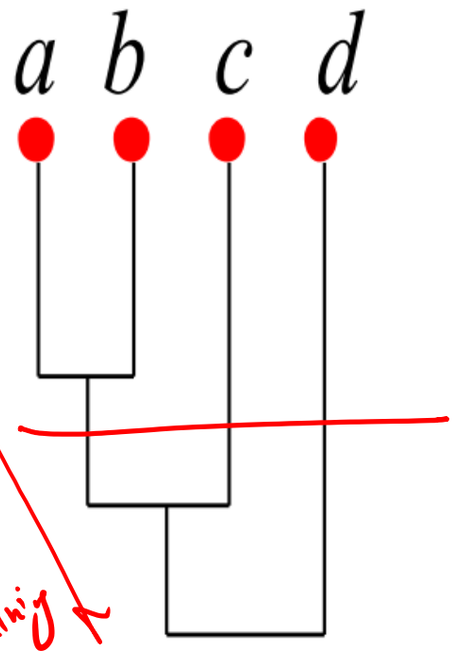
Complete link: Clusters are split into two groups of roughly equal size when we cut the dendrogram at the last merge. In general, this is a more useful organization of the data than a clustering with chains.

- Generally slower.
- Best when **you want compact, well-separated groups.** Example: Customer segmentation in marketing — you want customers in the same group to be similar in *all* features, not just one.
- **Outlier robustness:** Because it considers the farthest points, clusters won't "stretch" easily.

Average link: When you don't know which one may be better for you, start it with the average link method.

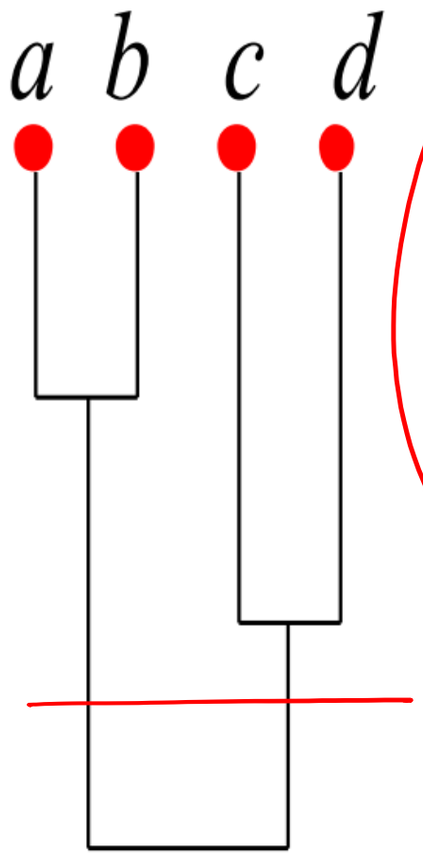
Dendrograms

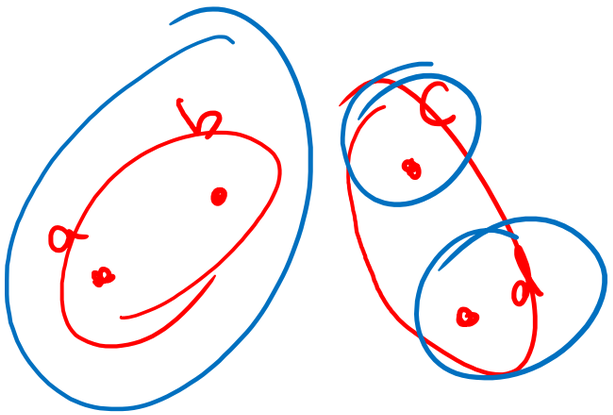
Single-Link



chaining

Complete-Link



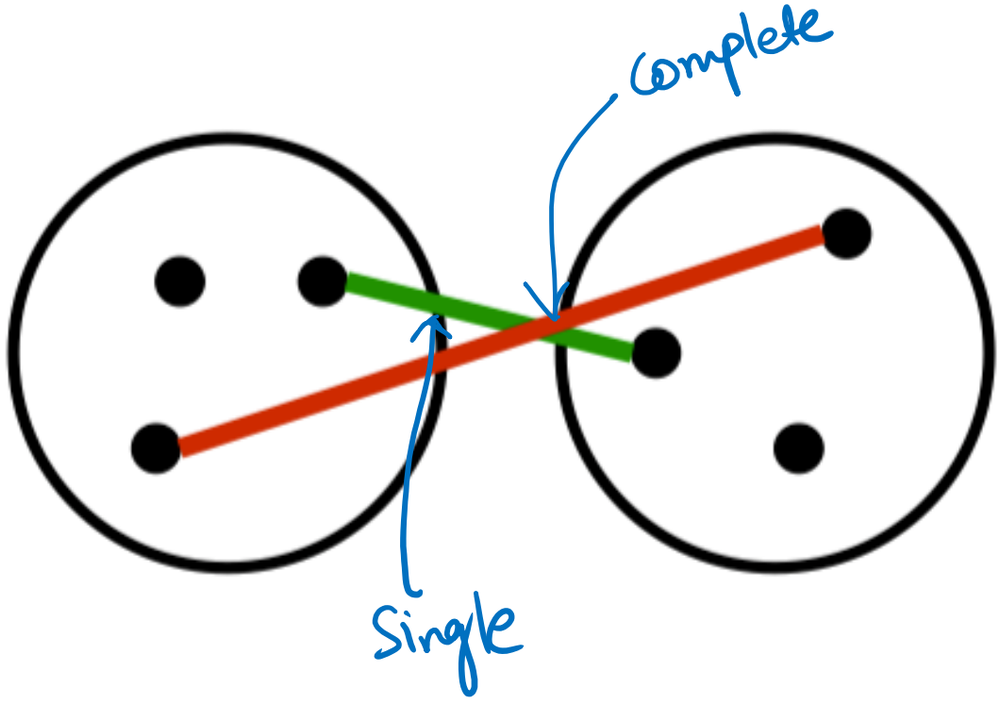


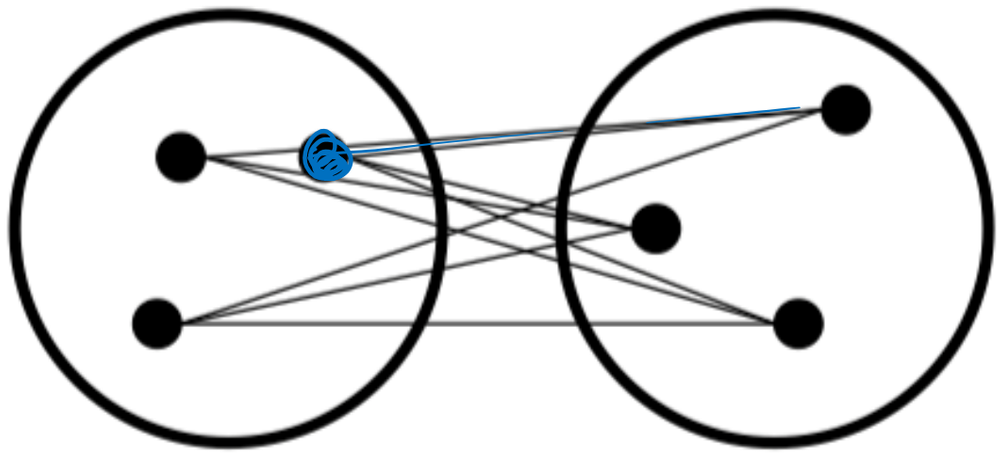
← Silhoutte coefficient

← elbow method

Georgia Tech

How to Define Distance Between Two Clusters?

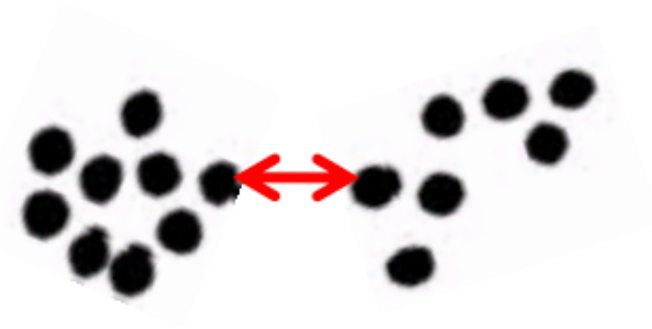




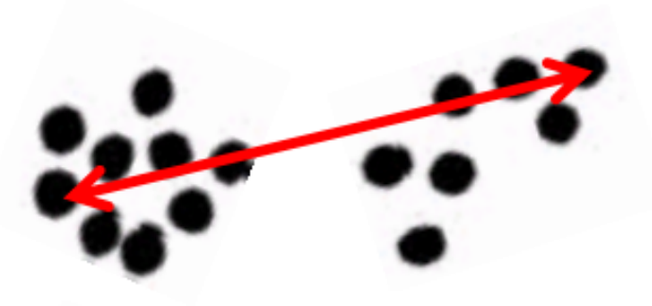
Bottom-up Agglomerative clustering

Different algorithms differ in how the similarities are defined (and hence updated) between two clusters

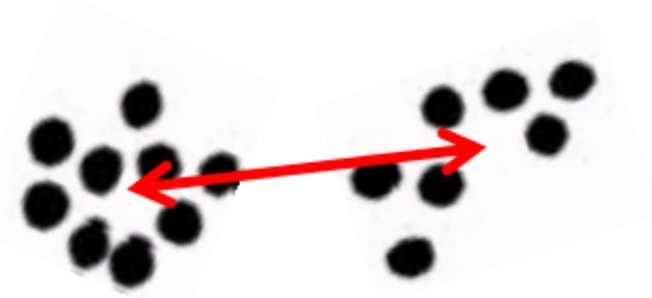
- Single-Link
  - Nearest Neighbor: similarity between their closest members.



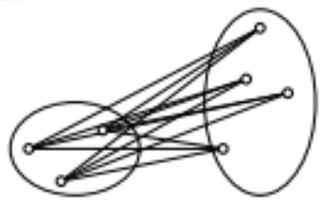
- Complete-Link
  - Furthest Neighbor: similarity between their furthest members.



- Centroid
  - Similarity between the centers of gravity



- Average-Link
  - Average similarity of all cross-cluster pairs.



Distance Between Clusters

Different distance functions can lead to different results!

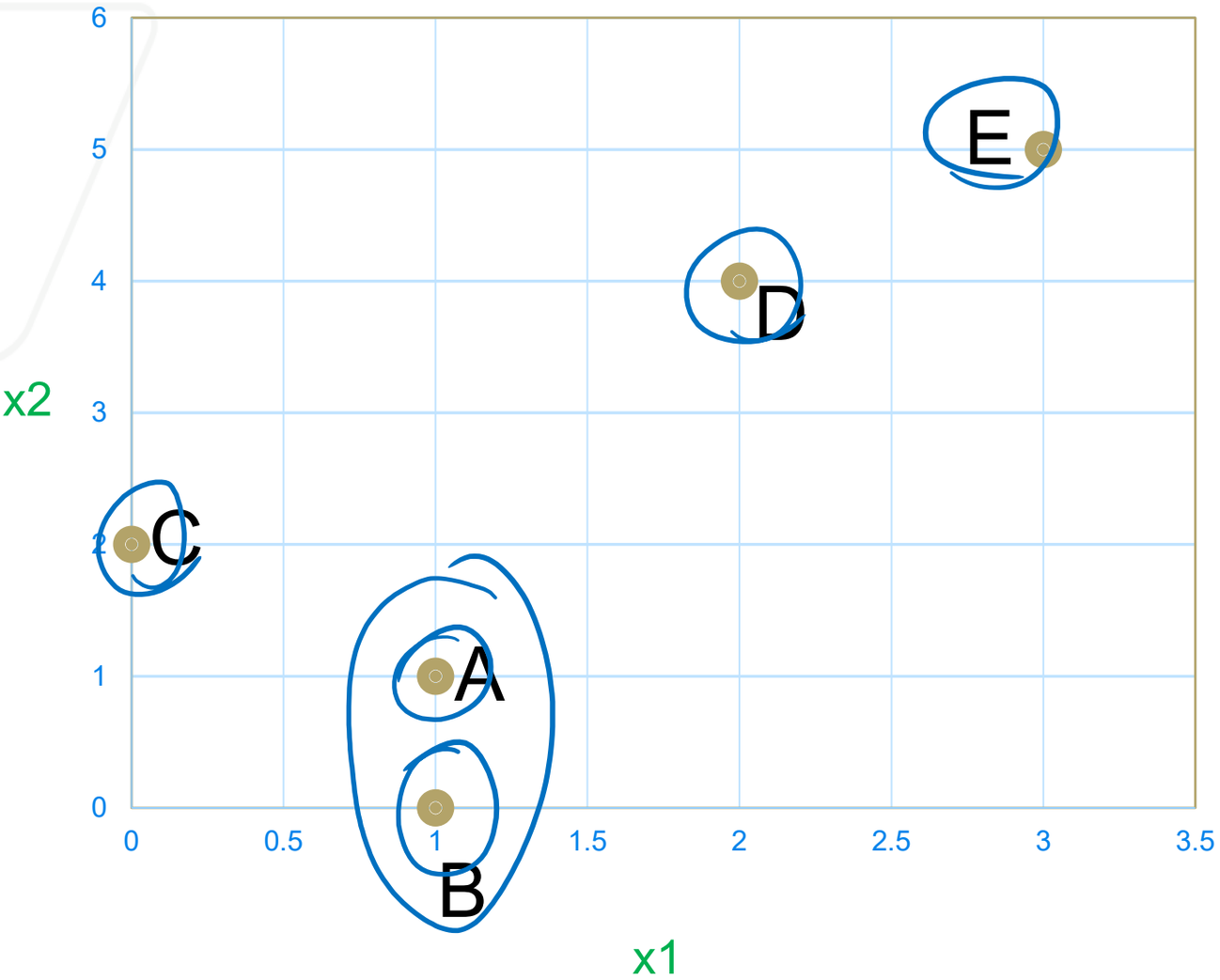
Average Link

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

very expensive table

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

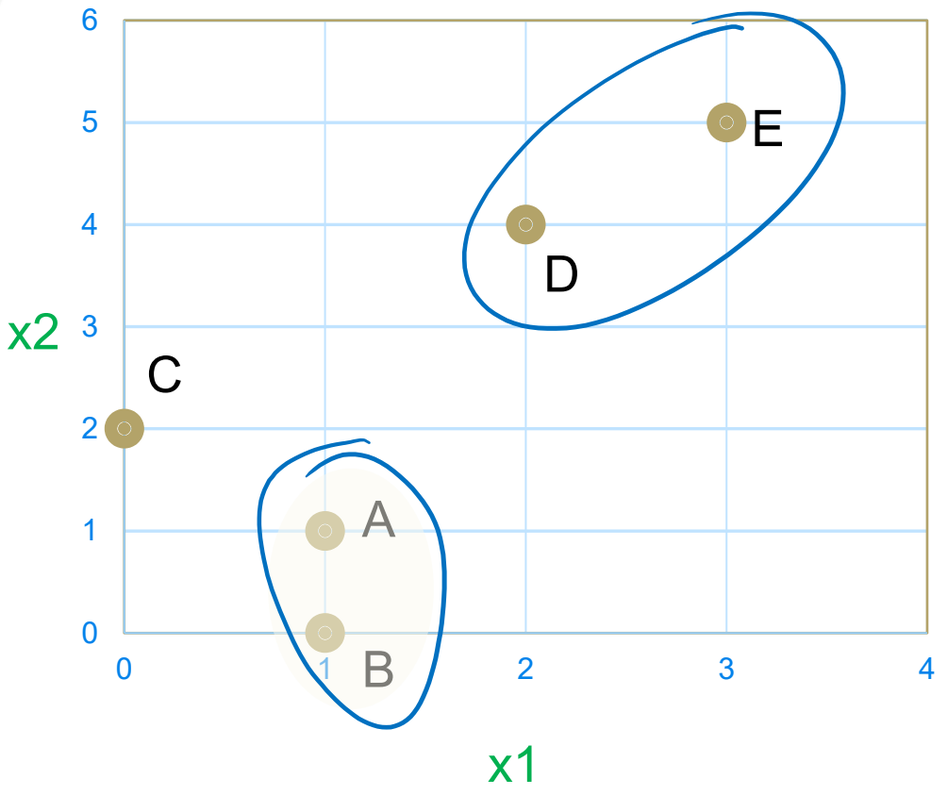


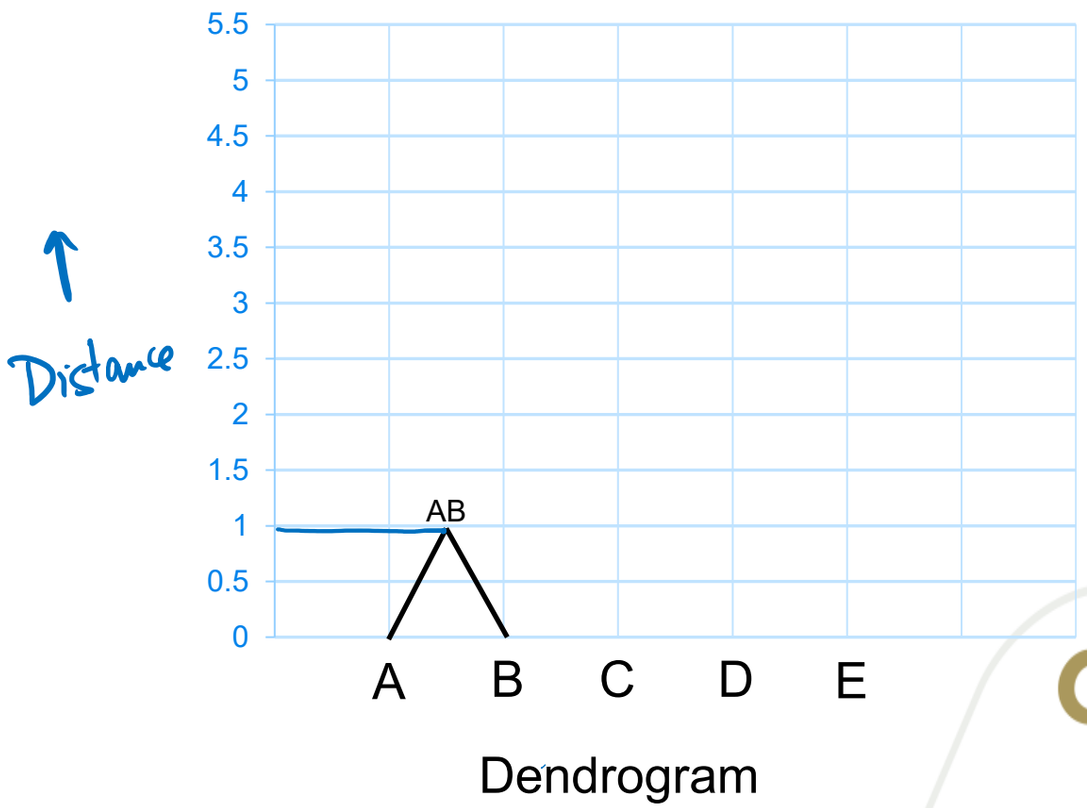
Georgia Tech

Distance based on Average point (Bottom-Up Clustering)

	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



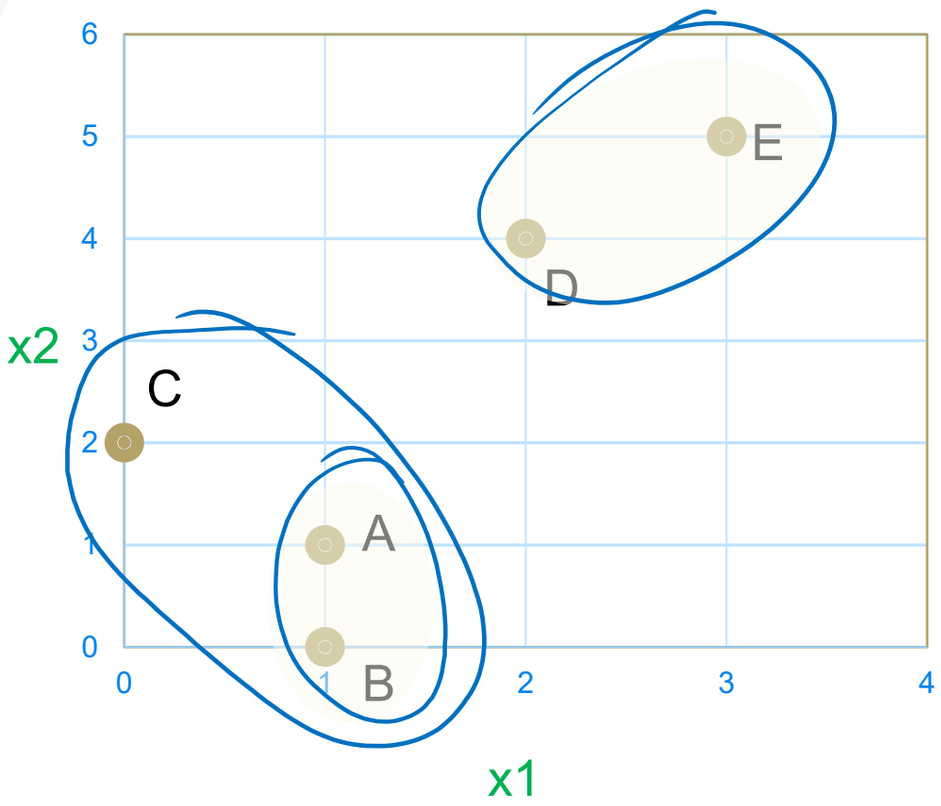


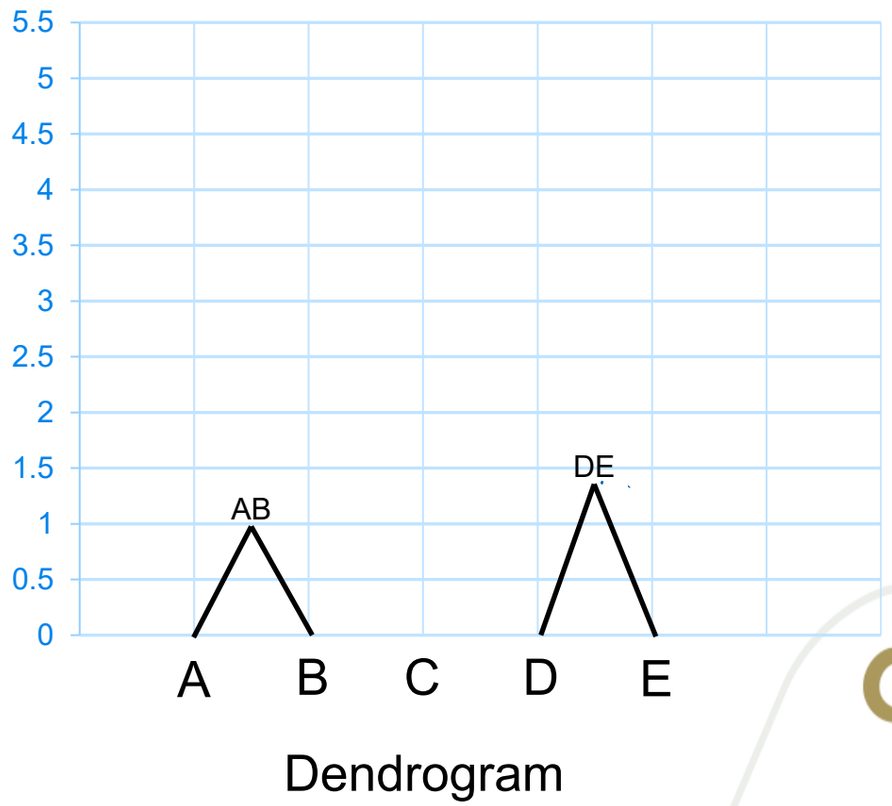
Dendrogram

Distance based on average point (Bottom-Up Clustering)

	(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





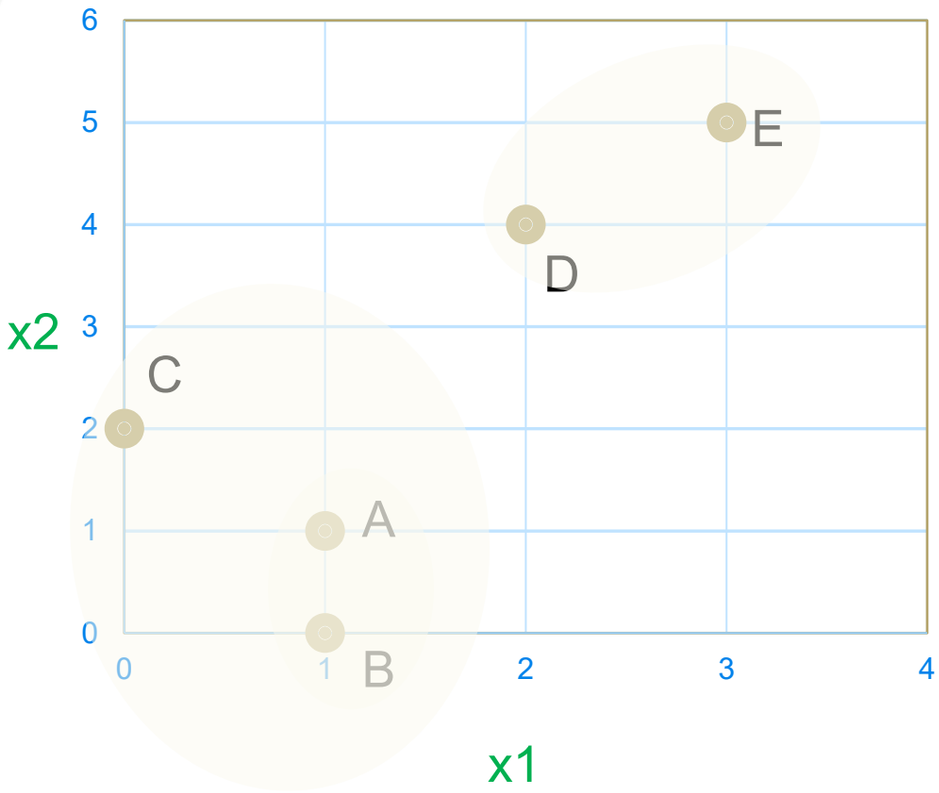
Dendrogram

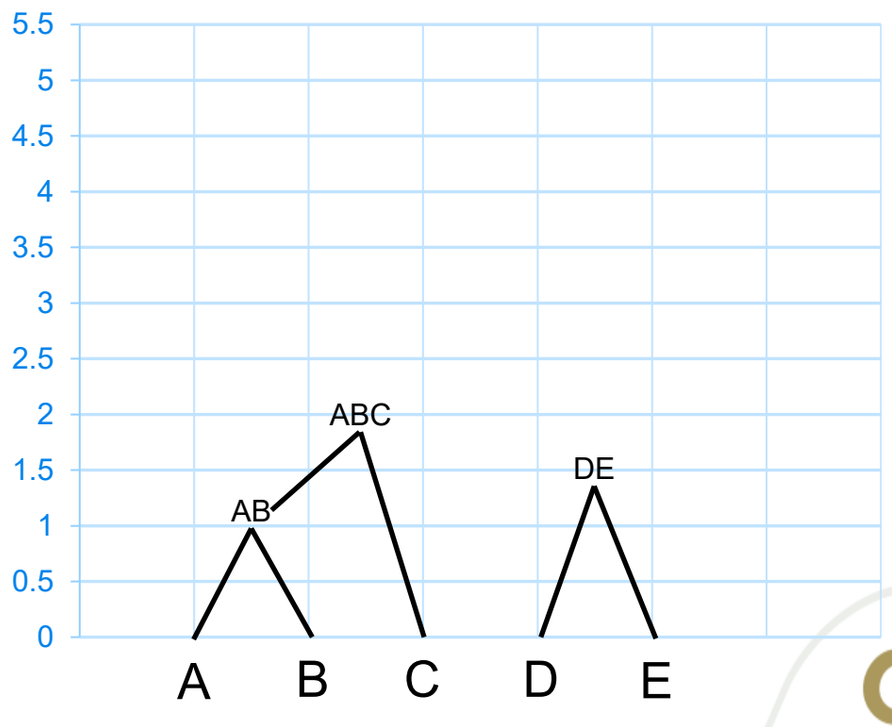
Georgia Tech

Distance based on average point (Bottom-Up Clustering)

	(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





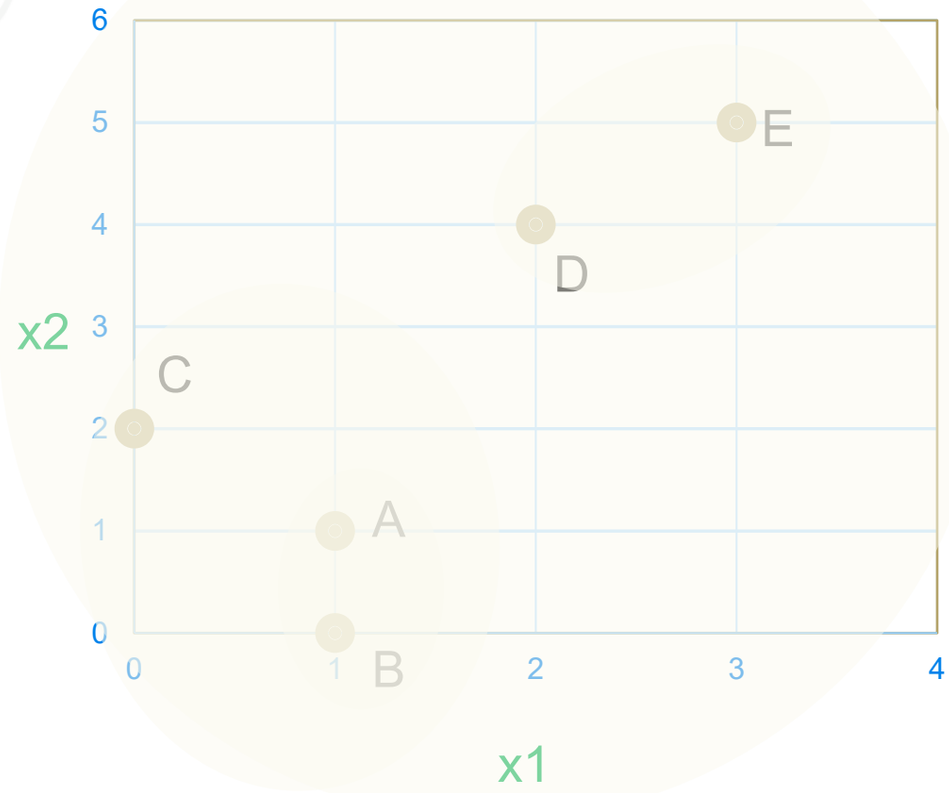
Dendrogram

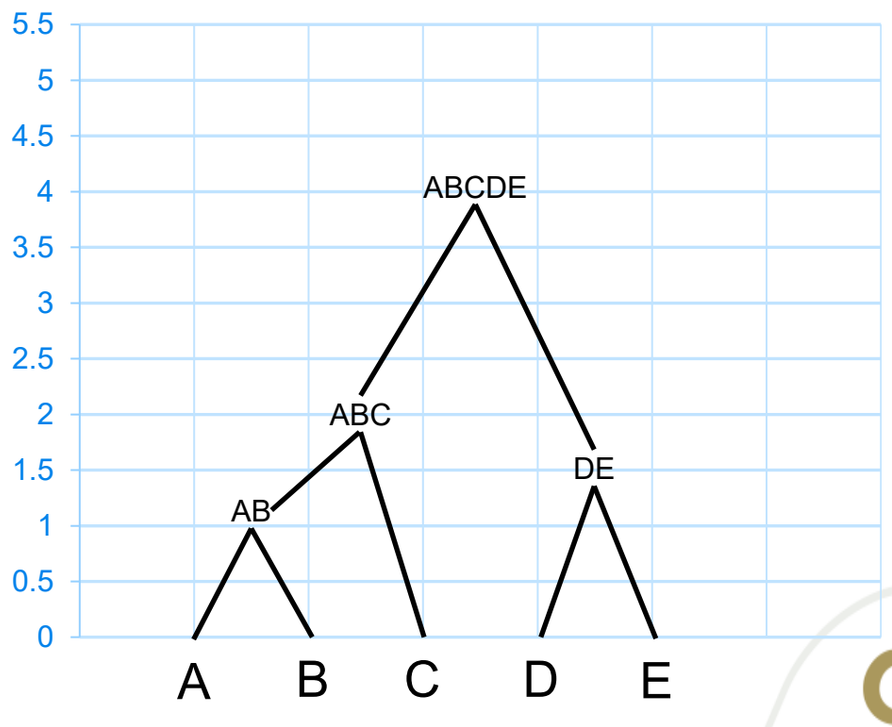
Georgia Tech

Distance based on average point (Bottom-Up Clustering)

	((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





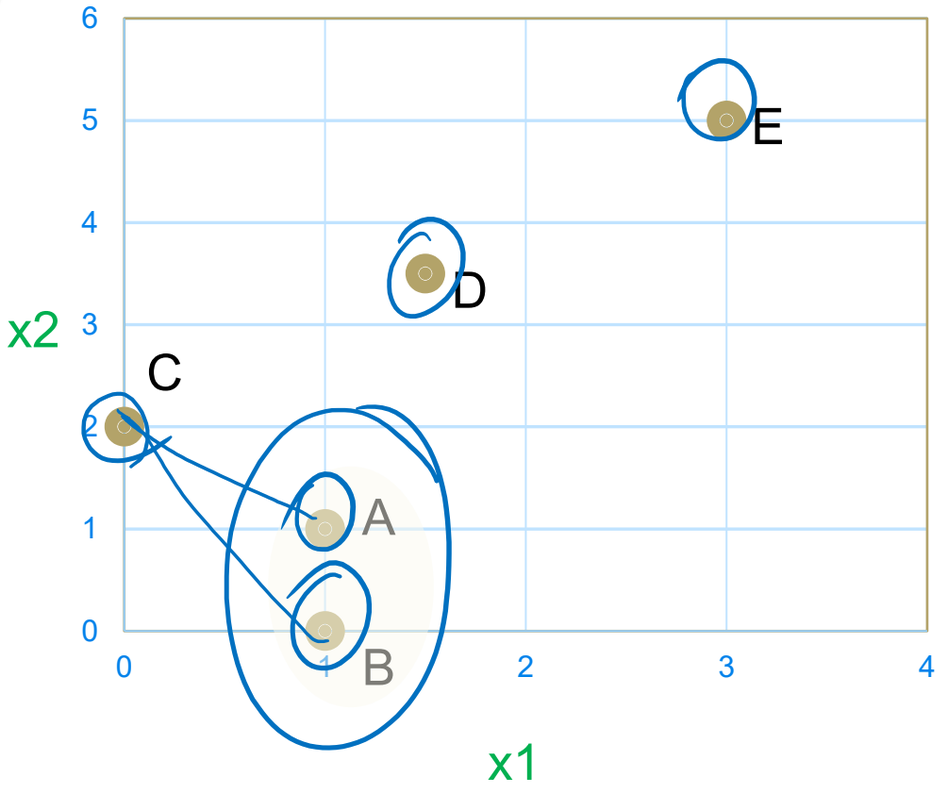
Dendrogram

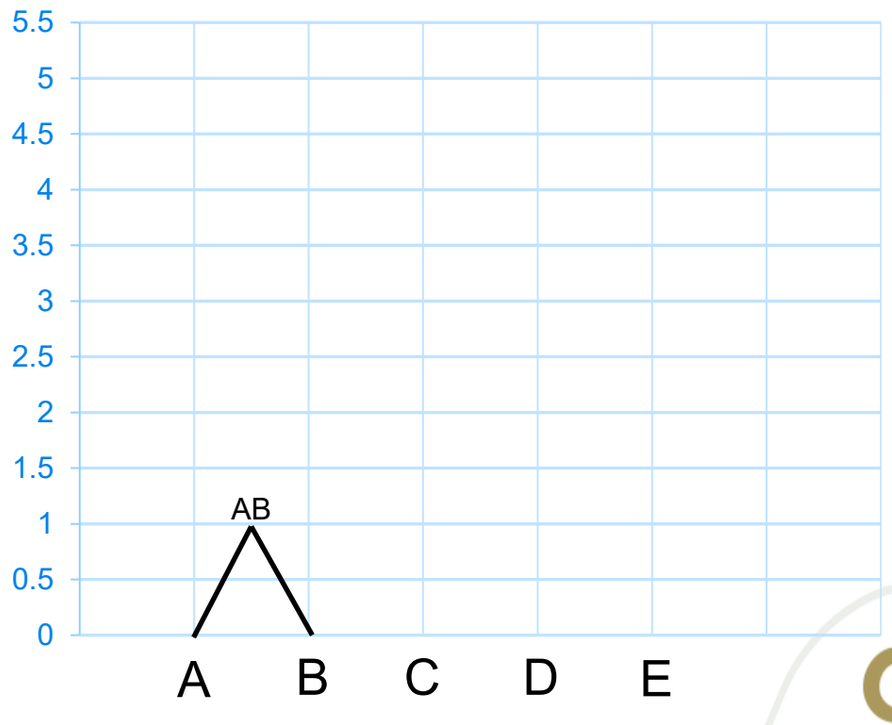
Georgia Tech

Distance based on Single Link (Bottom-Up Clustering)

	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





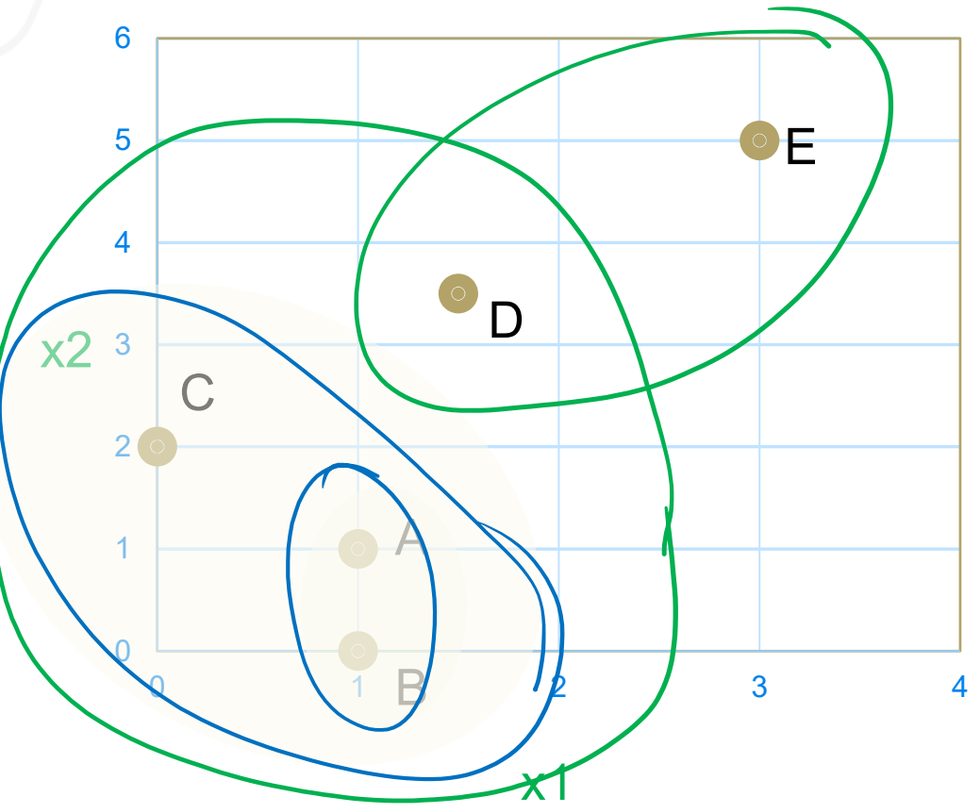
Dendrogram

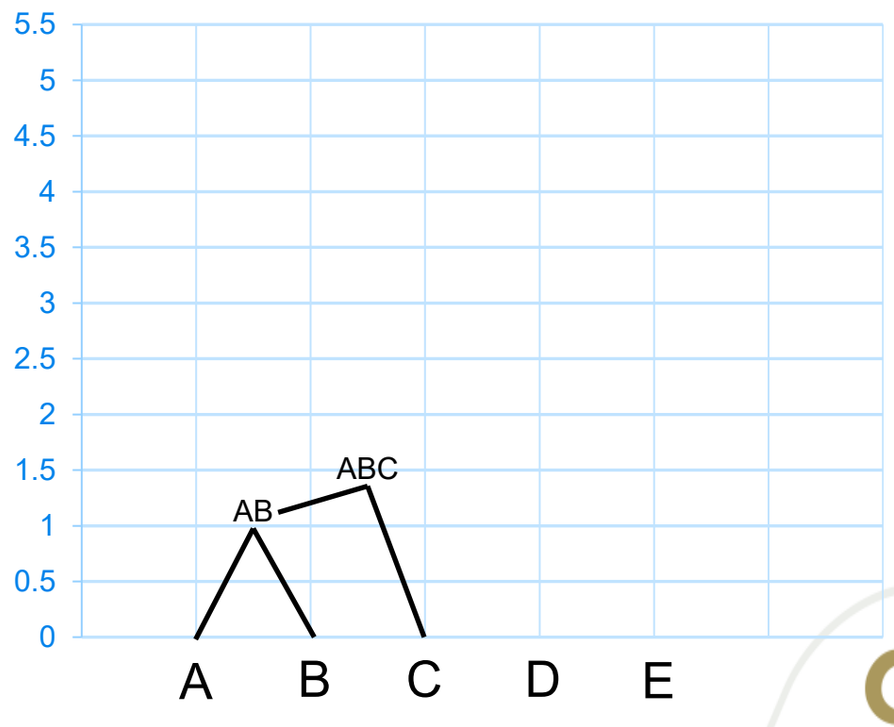
Georgia Tech

Distance based on Single Link (Bottom-Up Clustering)

	(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





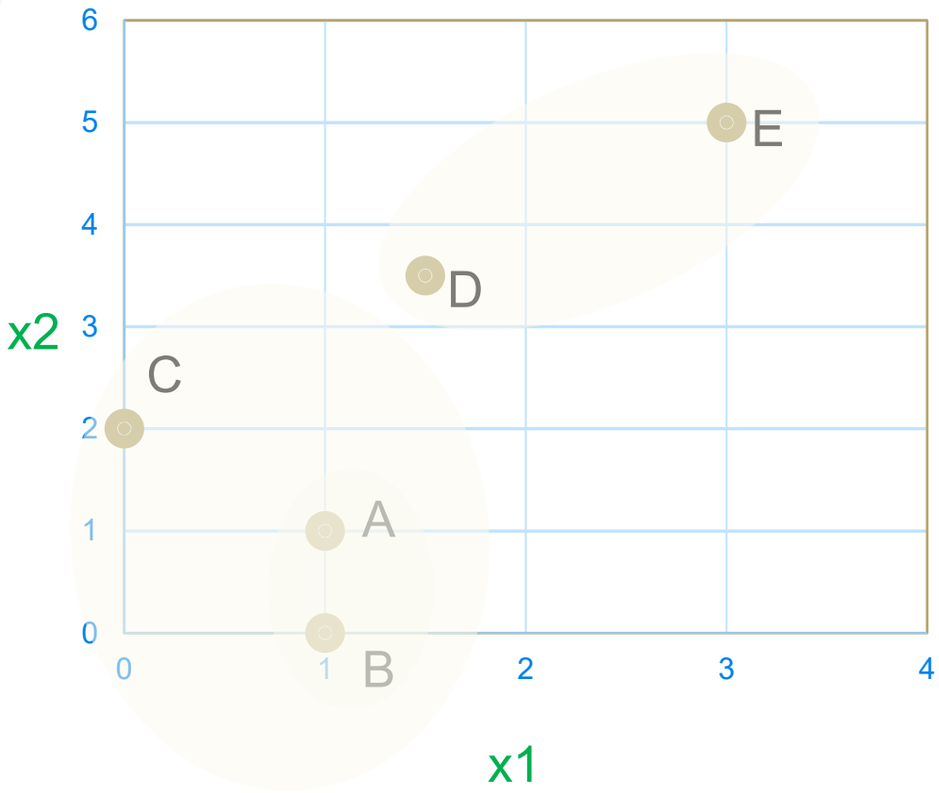
Dendrogram

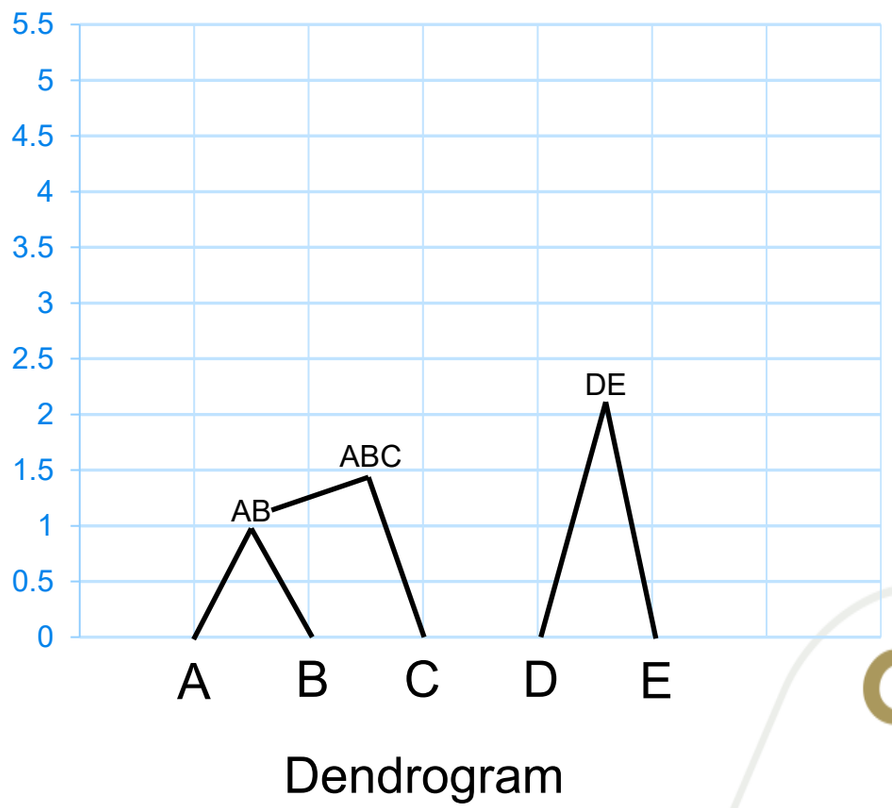
Georgia Tech

Distance based on Single Link (Bottom-Up Clustering)

	(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





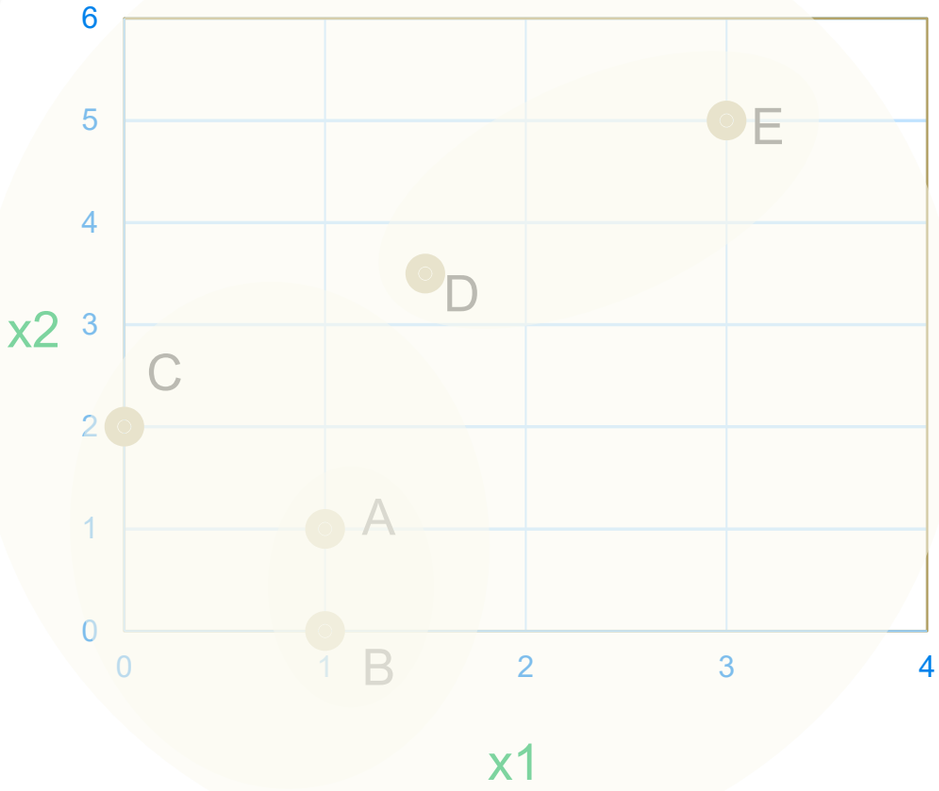
Dendrogram

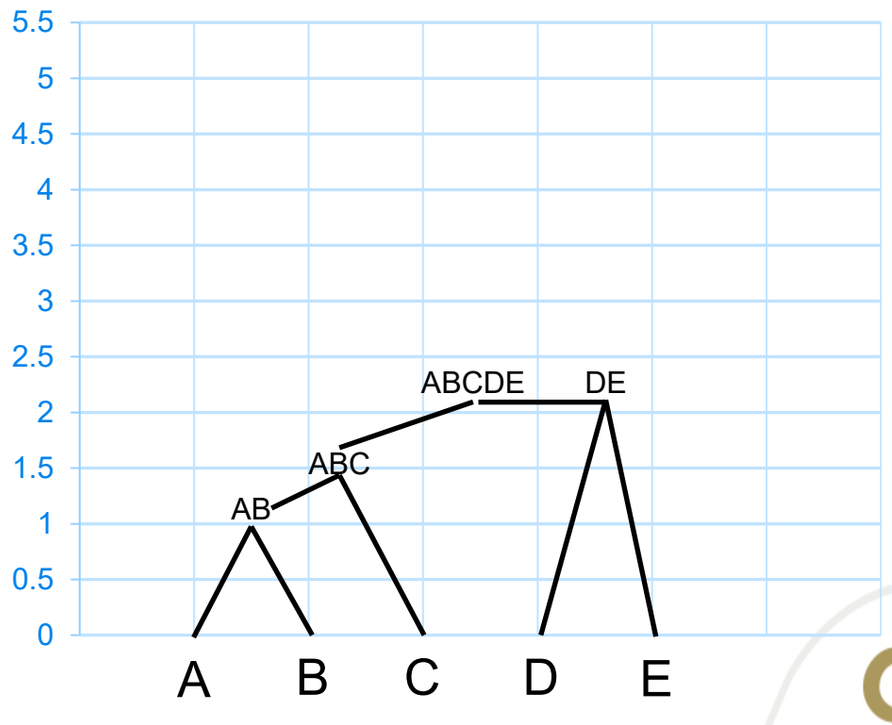
Georgia Tech

Distance based on Single Link (Bottom-Up Clustering)

	((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





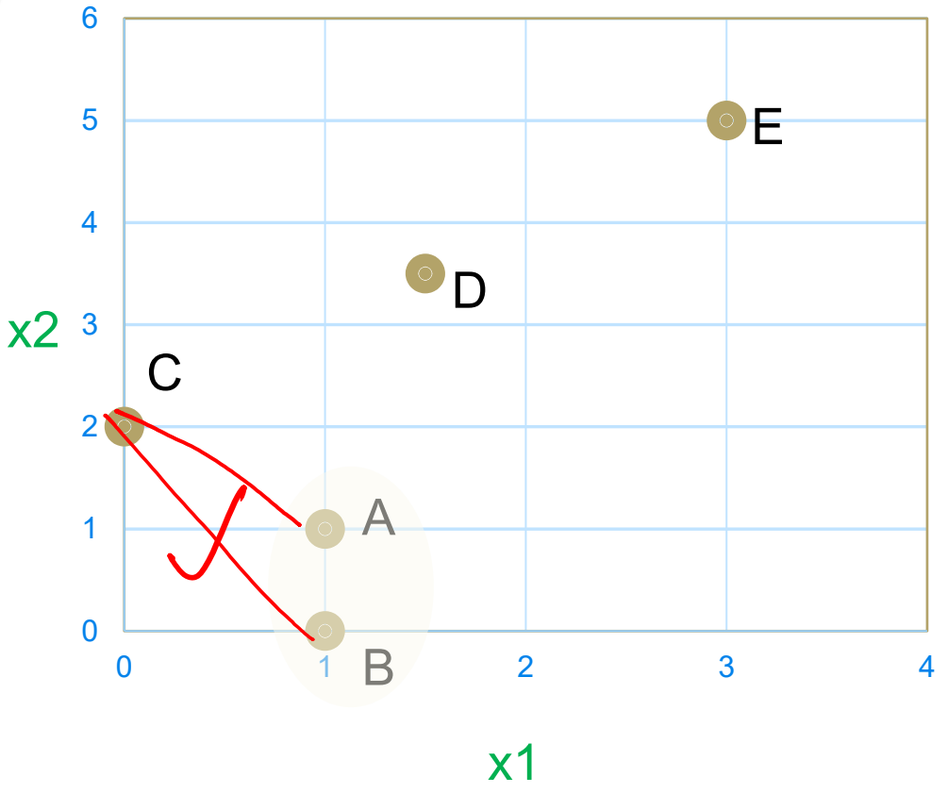
Dendrogram

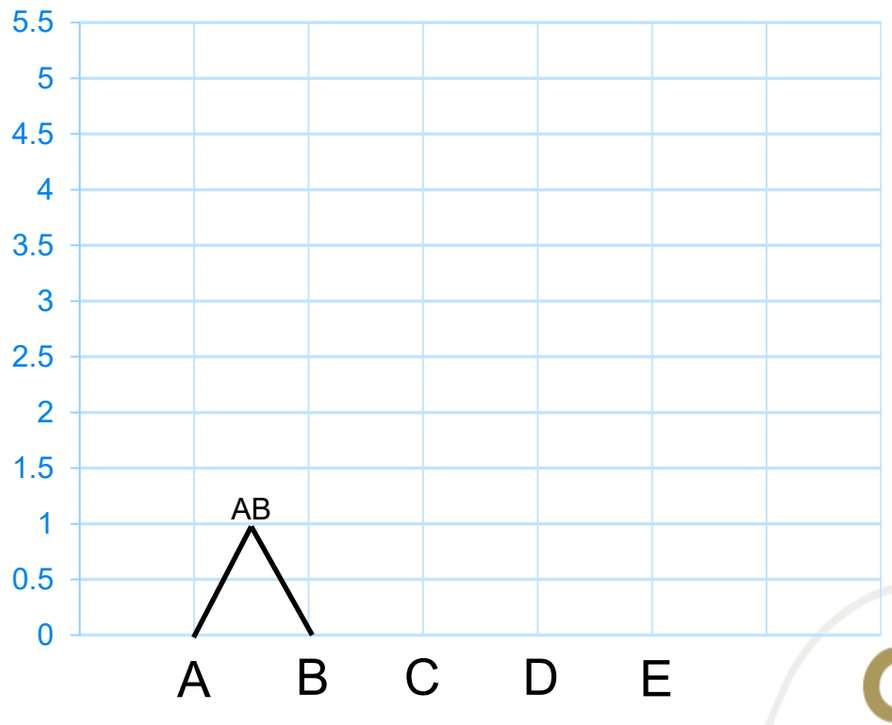
Georgia Tech

Distance based on Complete Link (Bottom-Up Clustering)

	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	3.53	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





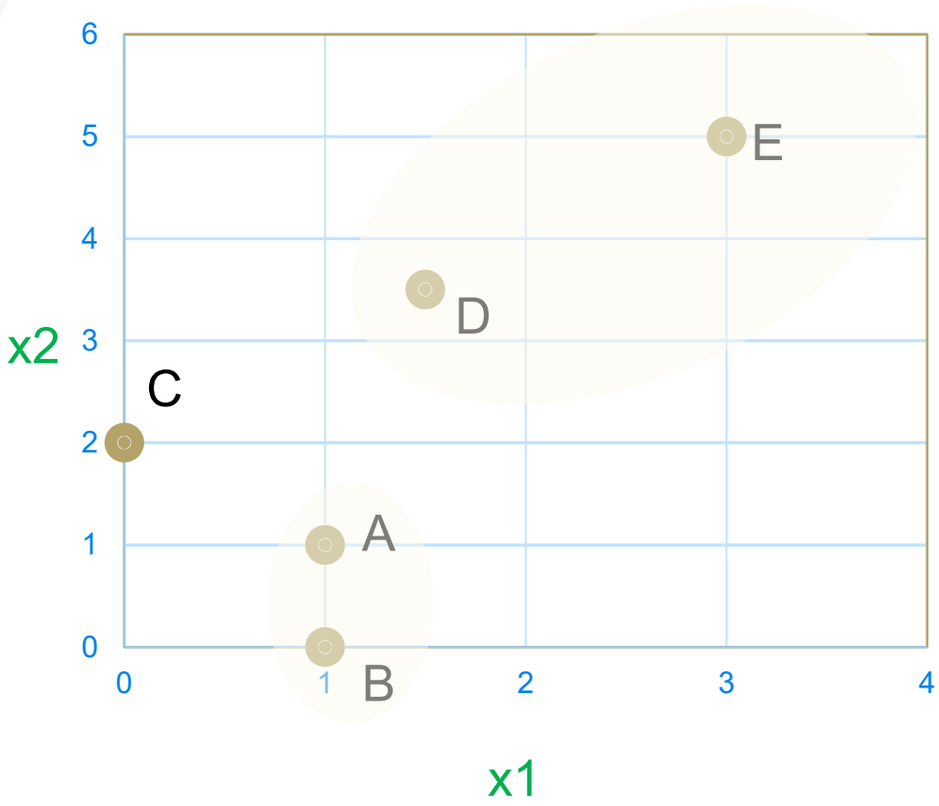
Dendrogram

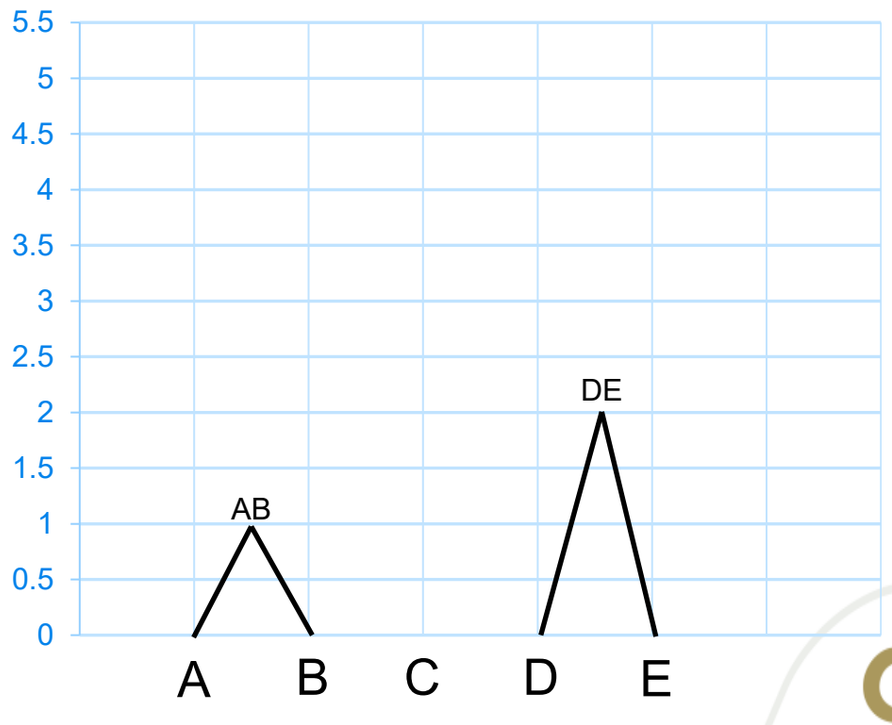
Georgia Tech

Distance based on Complete Link (Bottom-Up Clustering)

	(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





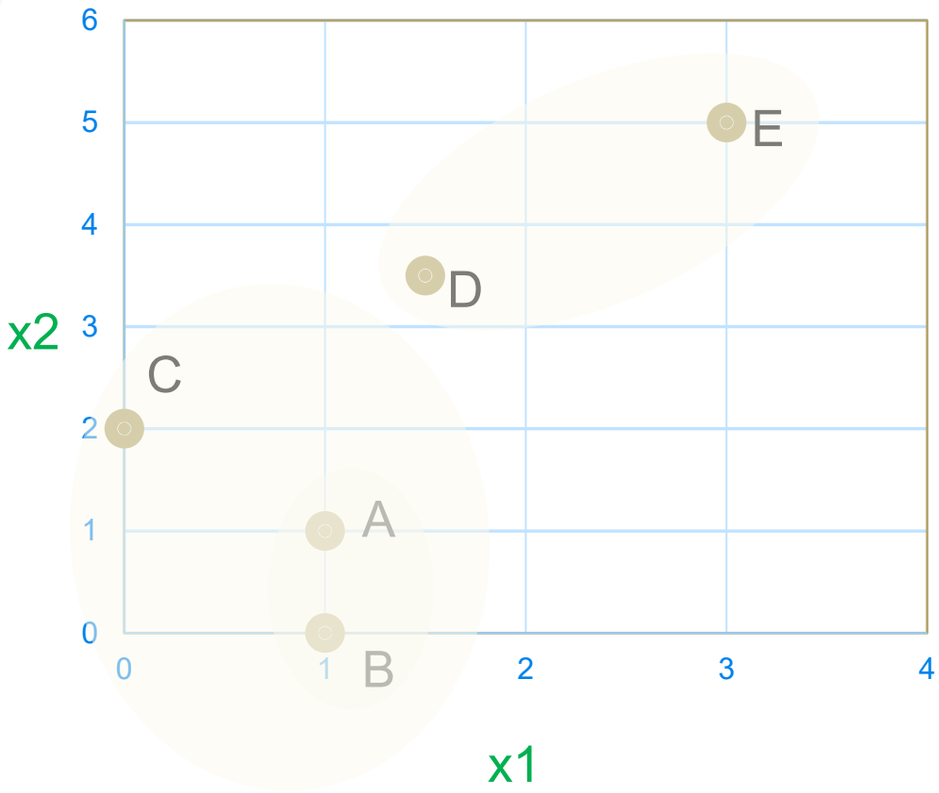
Dendrogram

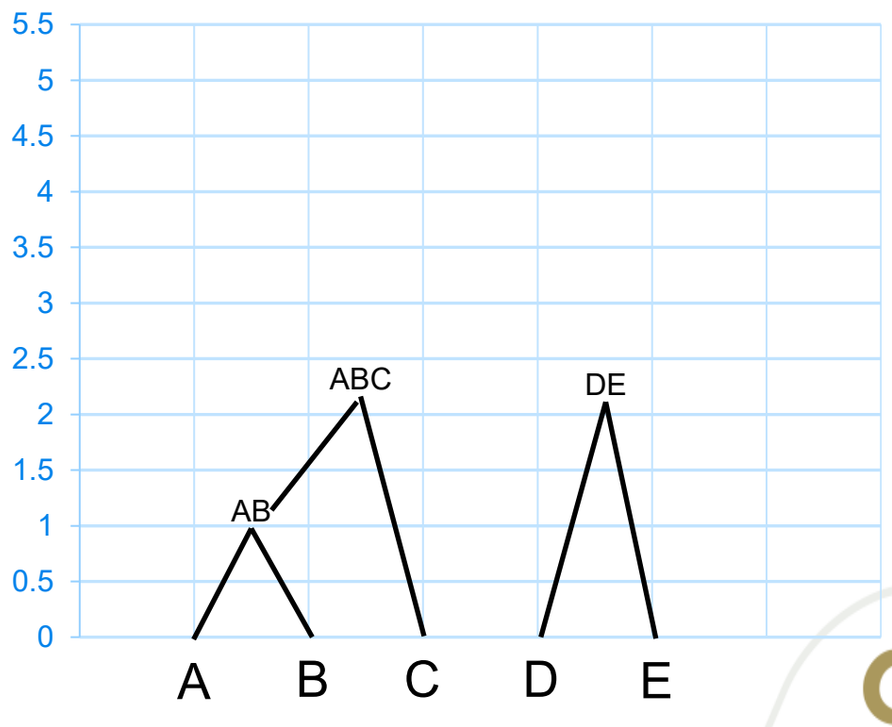
Georgia Tech

Distance based on Single Link (Bottom-Up Clustering)

	(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

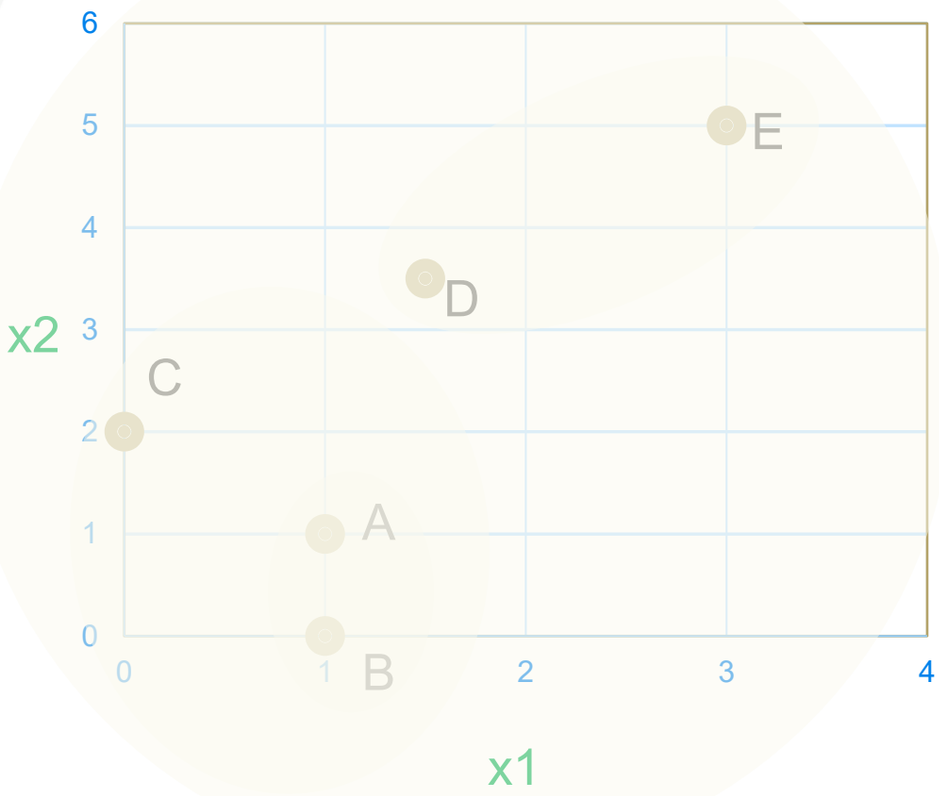
Georgia Tech

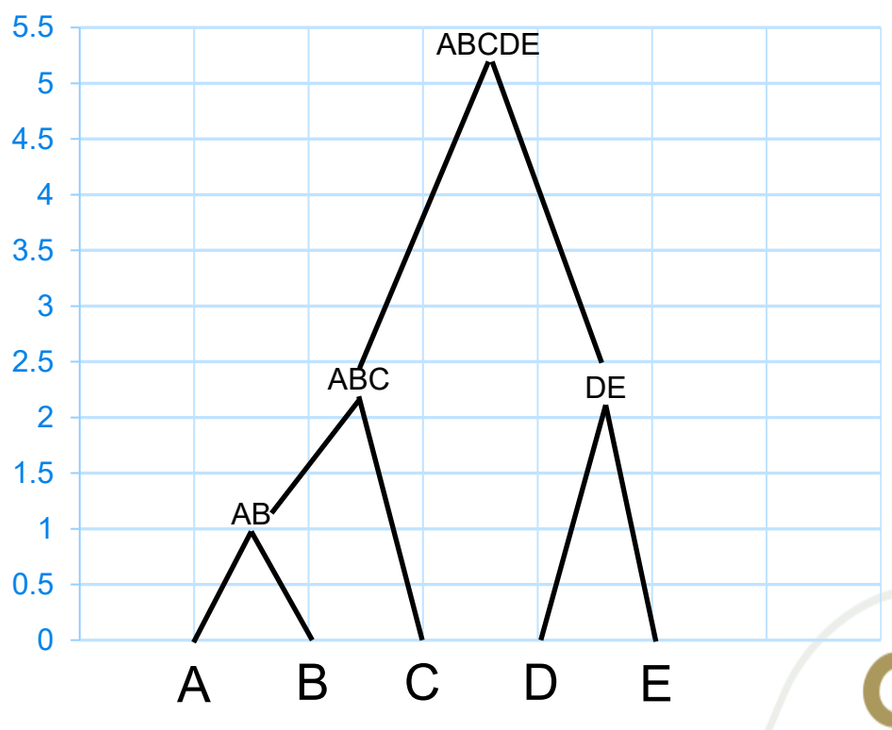
Distance based on Complete Link (Bottom-Up Clustering)

	((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

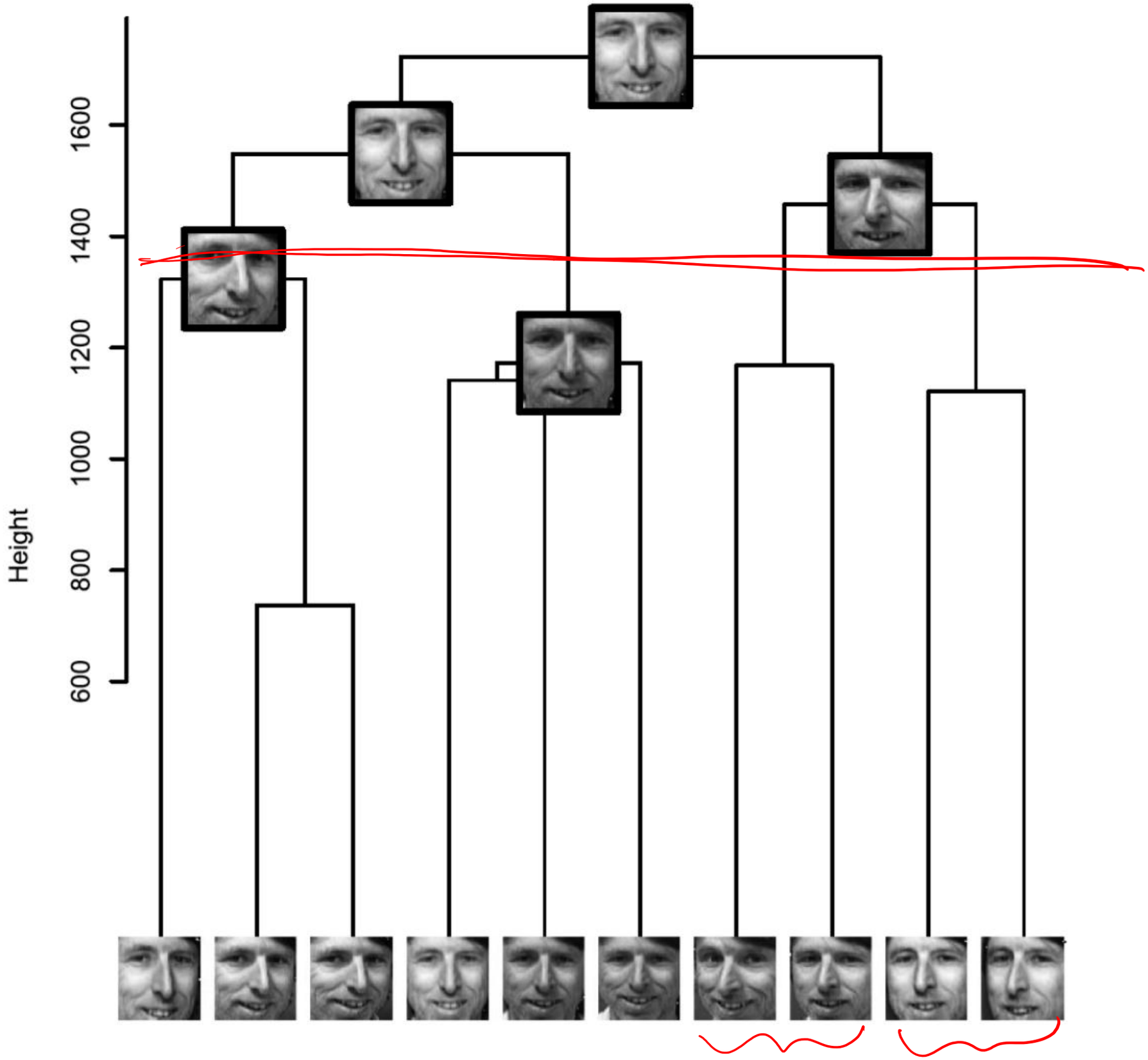
Tall dendogram





Dendrogram

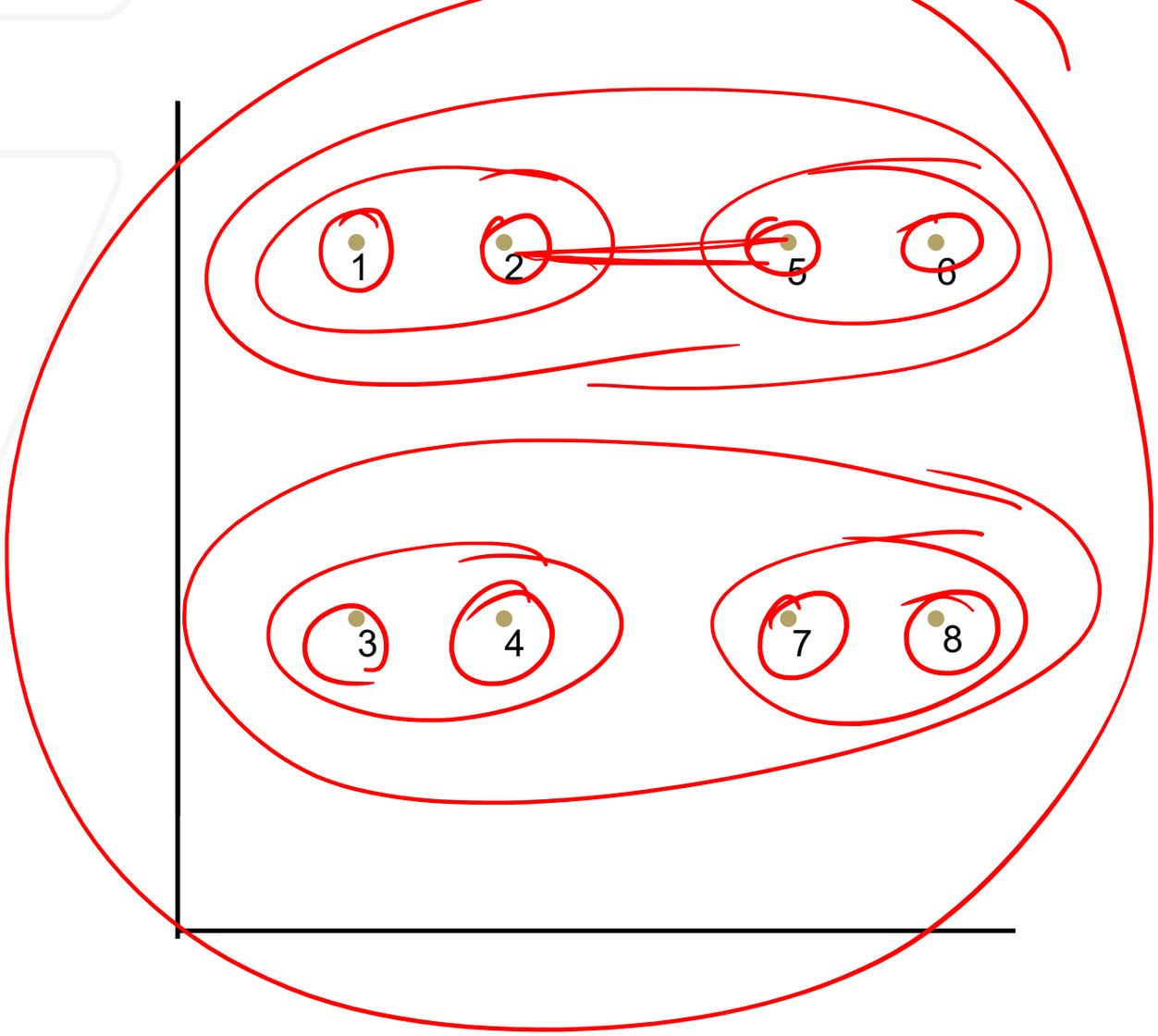
Georgia Tech



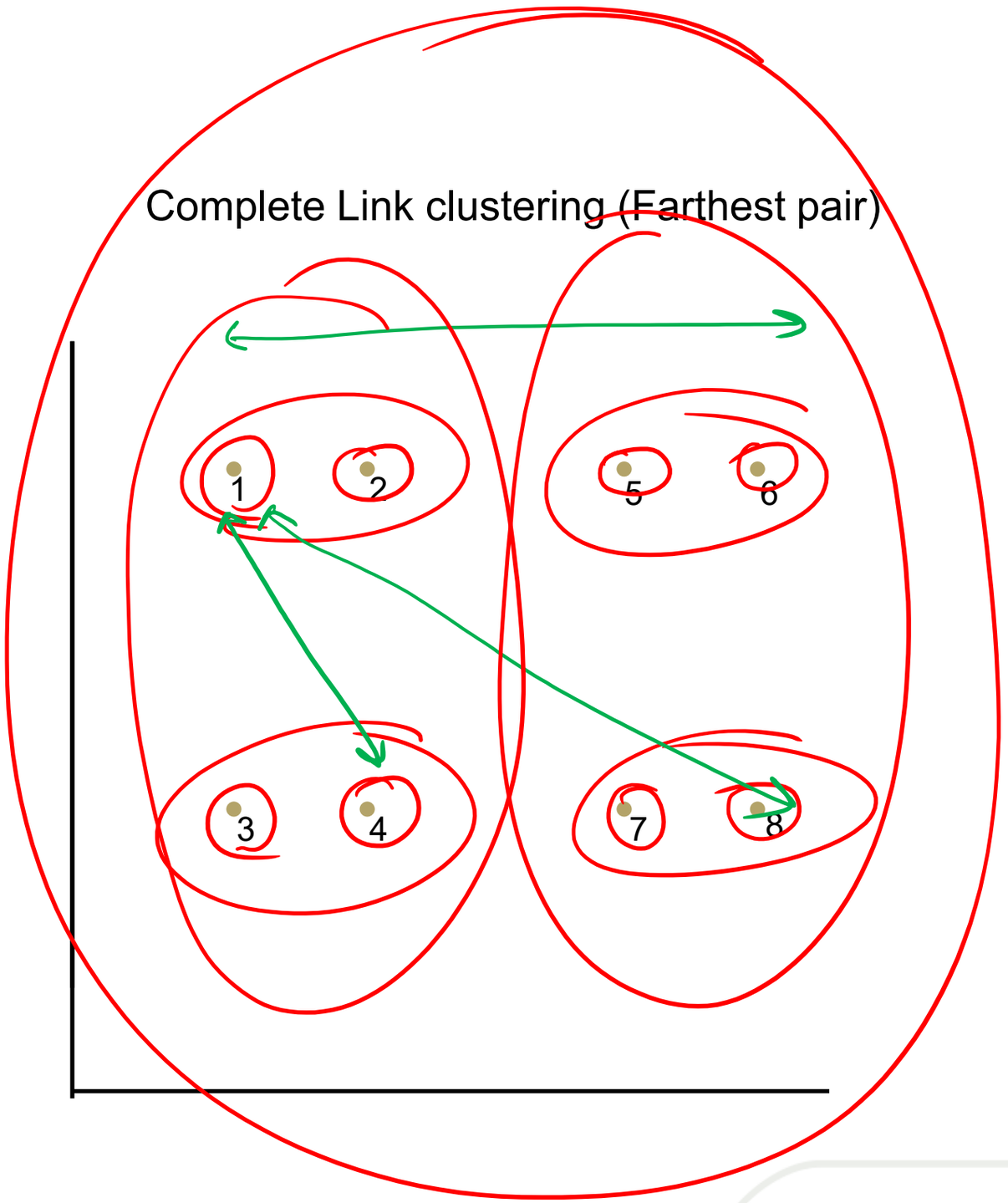
(From Bien et al. (2011))

Another Example

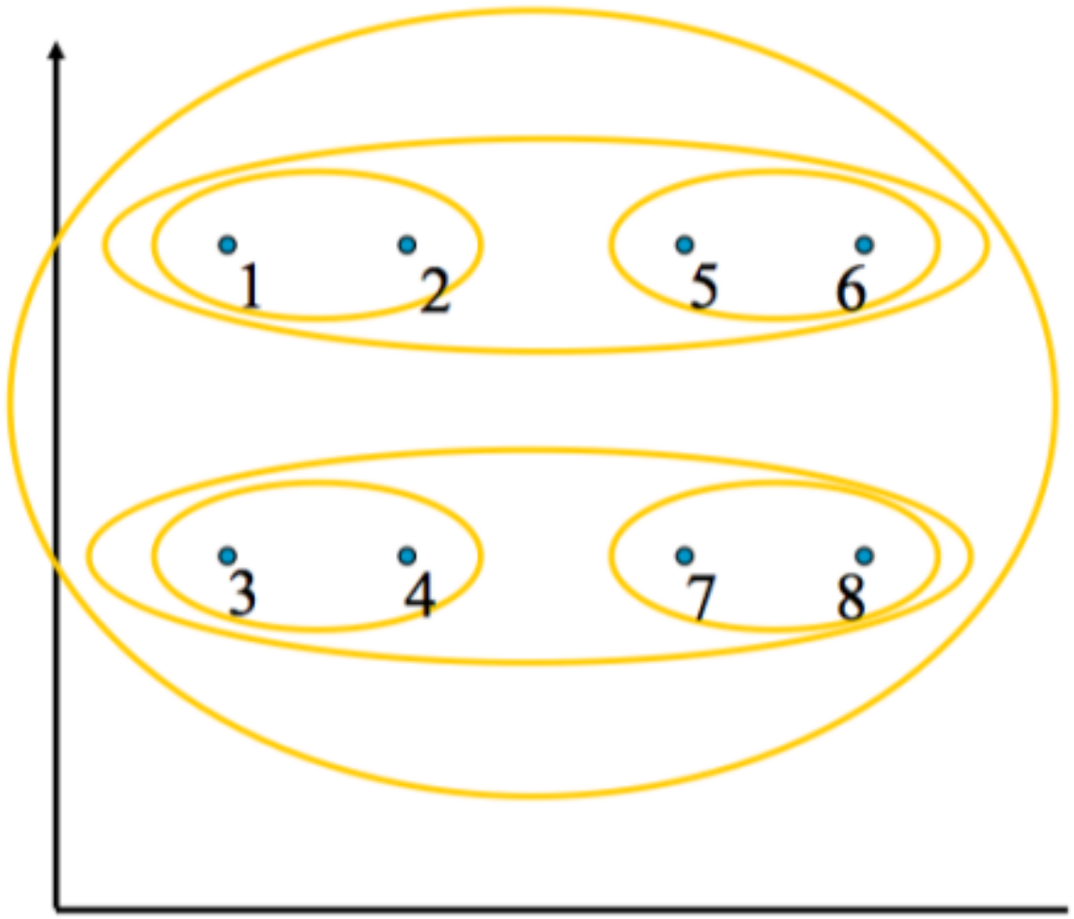
Single Link clustering (Closest pair)



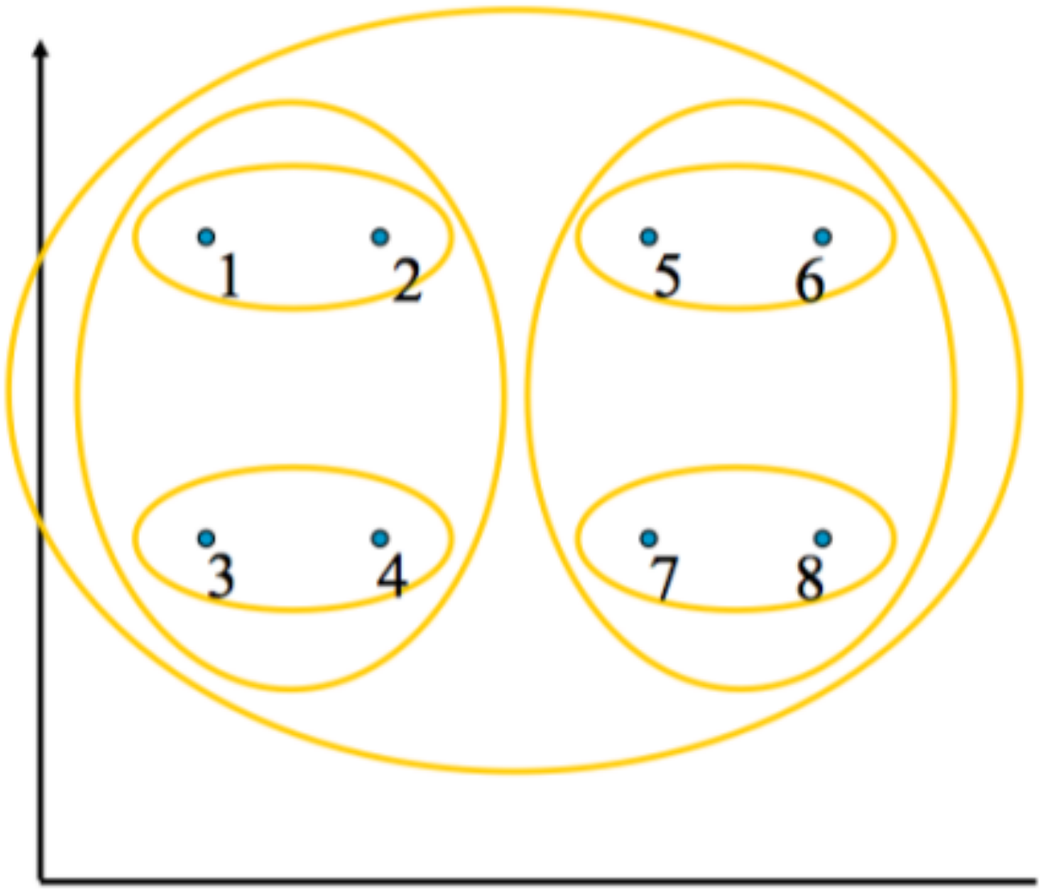
Complete Link clustering (Farthest pair)



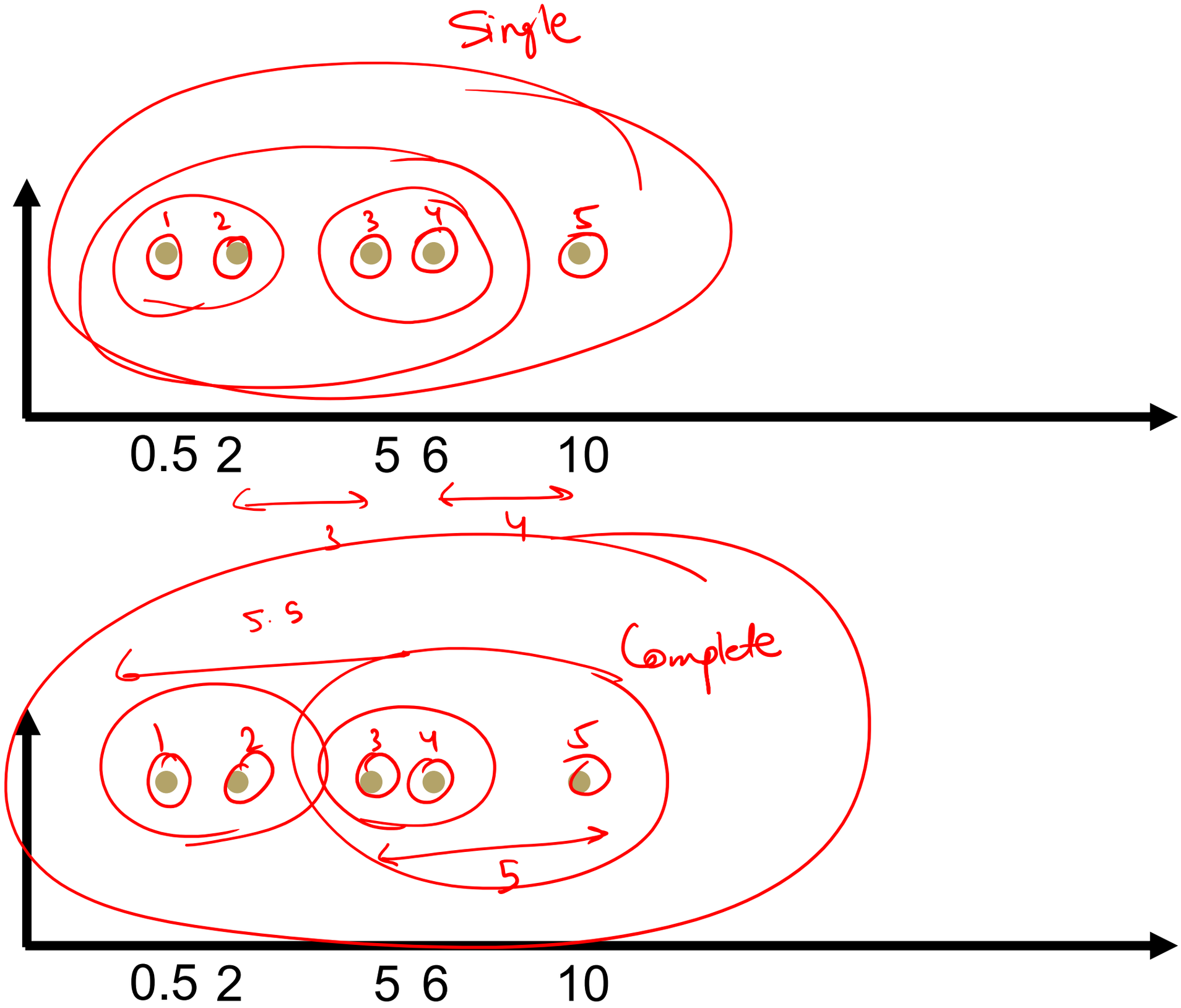
Closest pair
(single-link clustering)

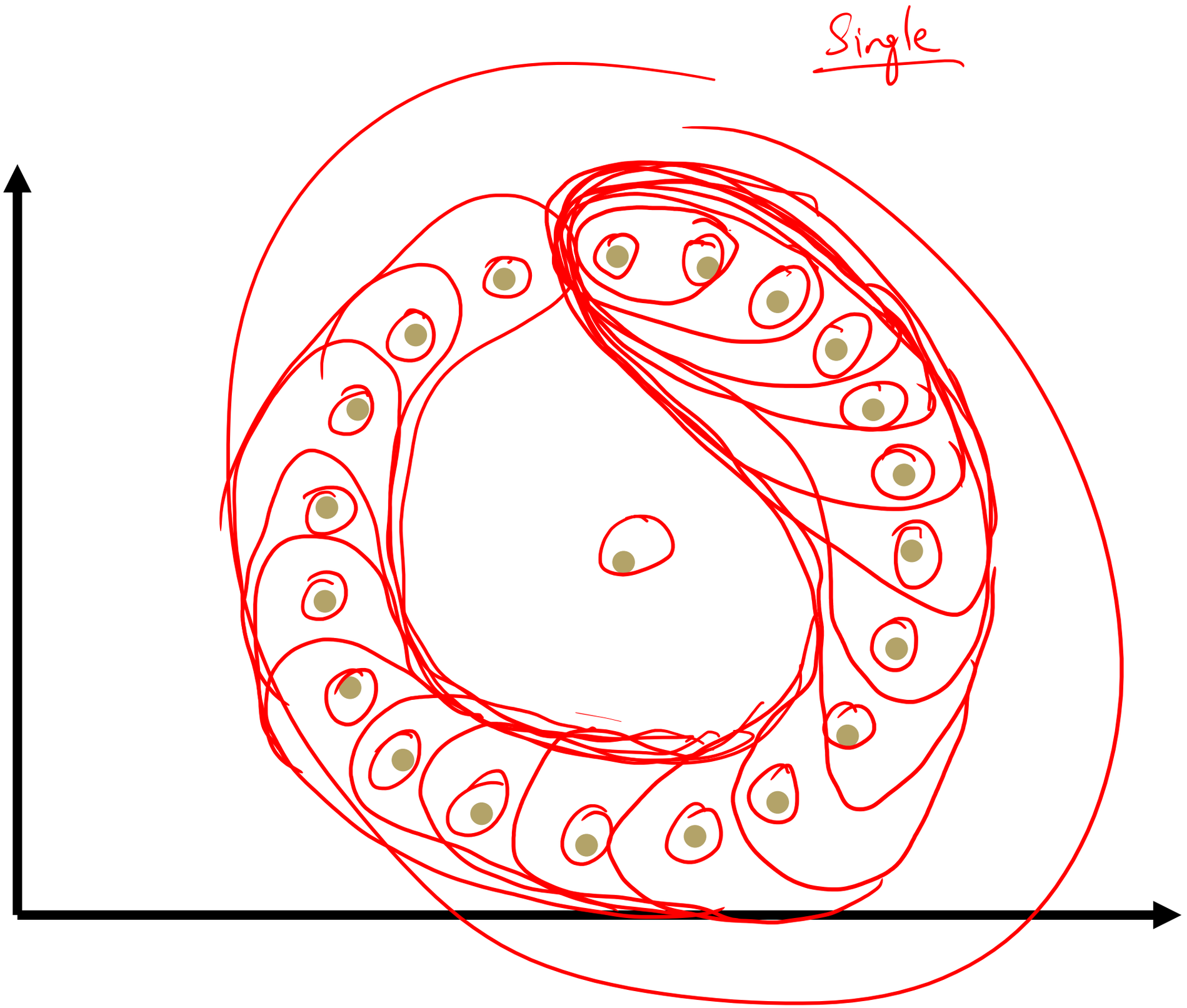


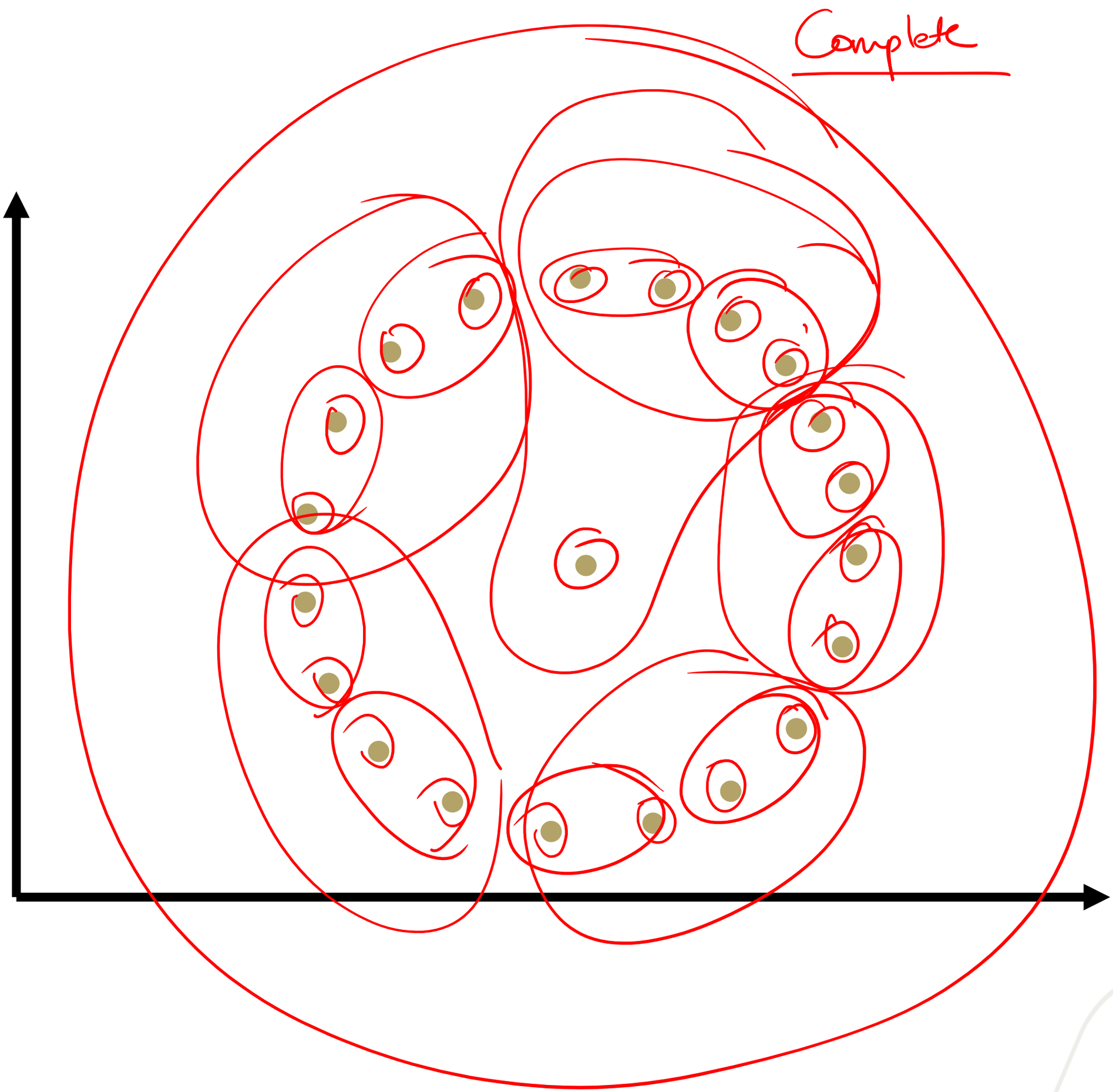
Farthest pair
(complete-link clustering)

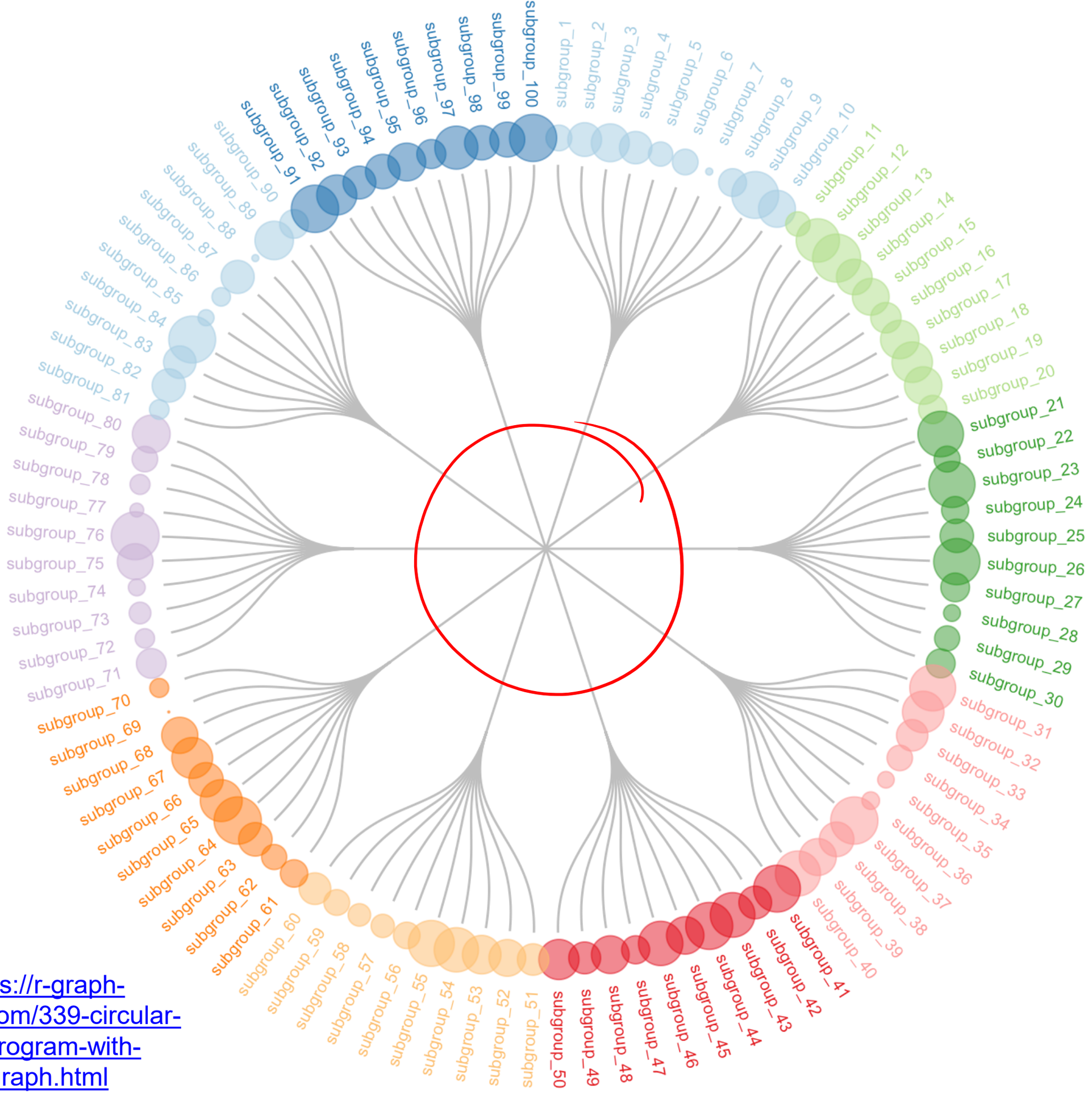


Slide credit: David Sontag.


Single
1
2
3
4
5
0.5
2
5
6
10
3
4
5.5
Complete
1
2
3
4
5
5
0.5
2
5
6
10


Single


Complete



https://r-graph-gallery.com/339-circular-dendrogram-with-ggraph.html

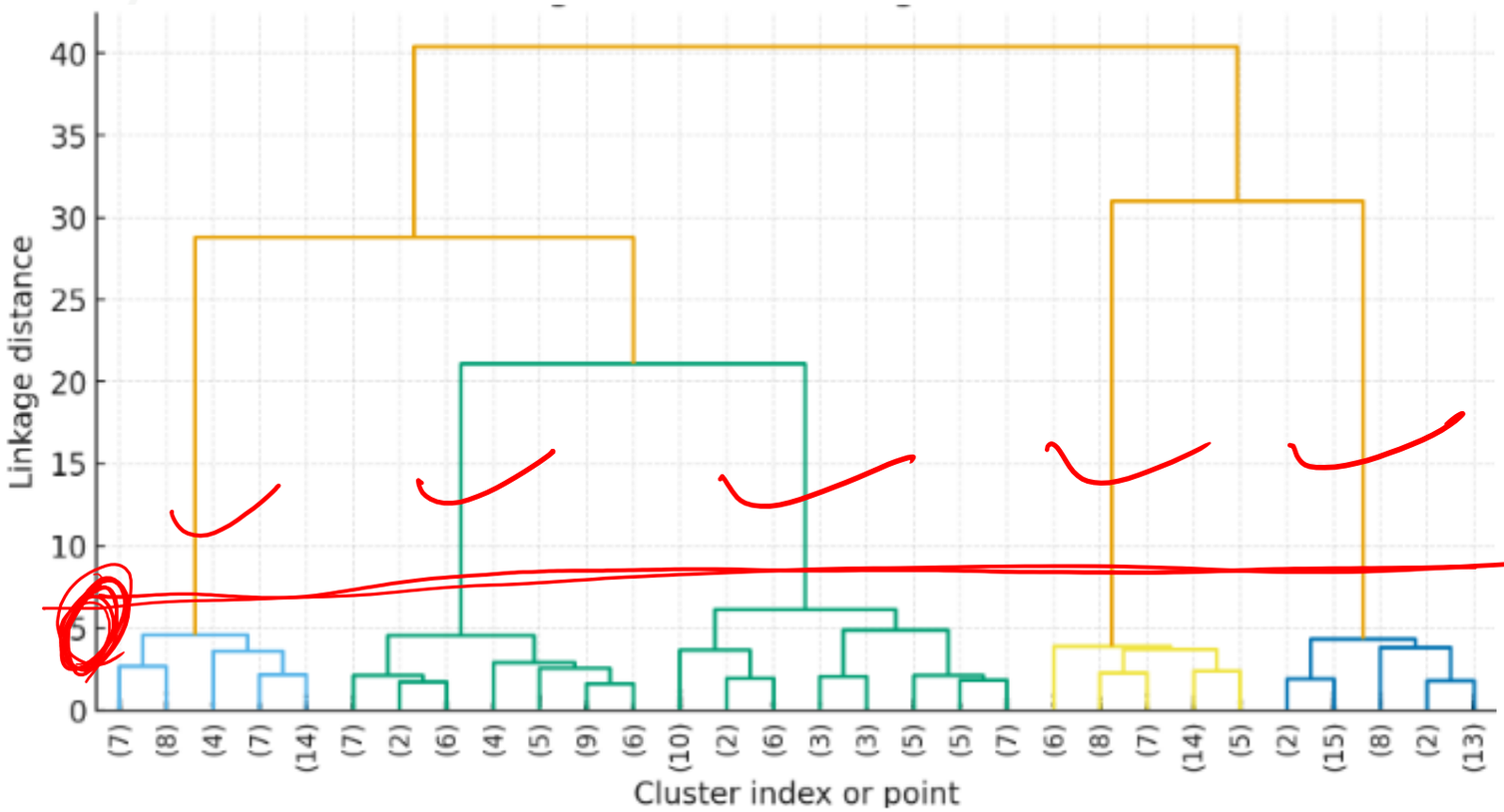
Georgia Tech

Clustering Evaluation

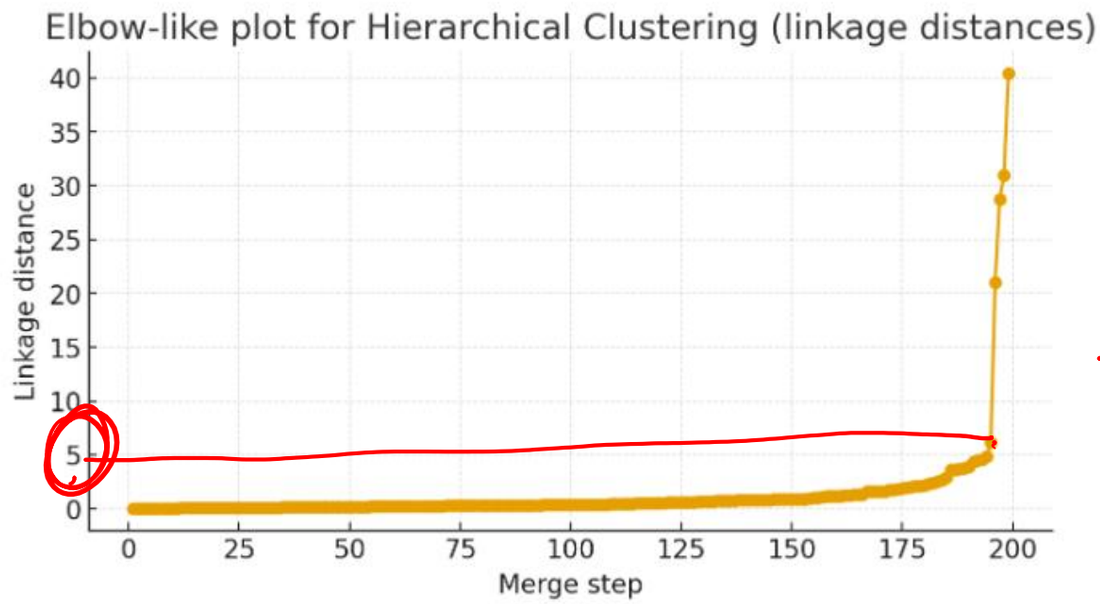
- Internal measures for clustering evaluation
  - Elbow method ← k means
  - Silhouette Coefficient ← $\frac{\text{Within cluster distances}}{\text{Nearest outside cluster distance}}$
  - Graph-based measures (Beta-CV and Normalized cut)

We want intra-cluster datapoints to be as close as possible to each other and inter-clusters to be as far as possible from each other

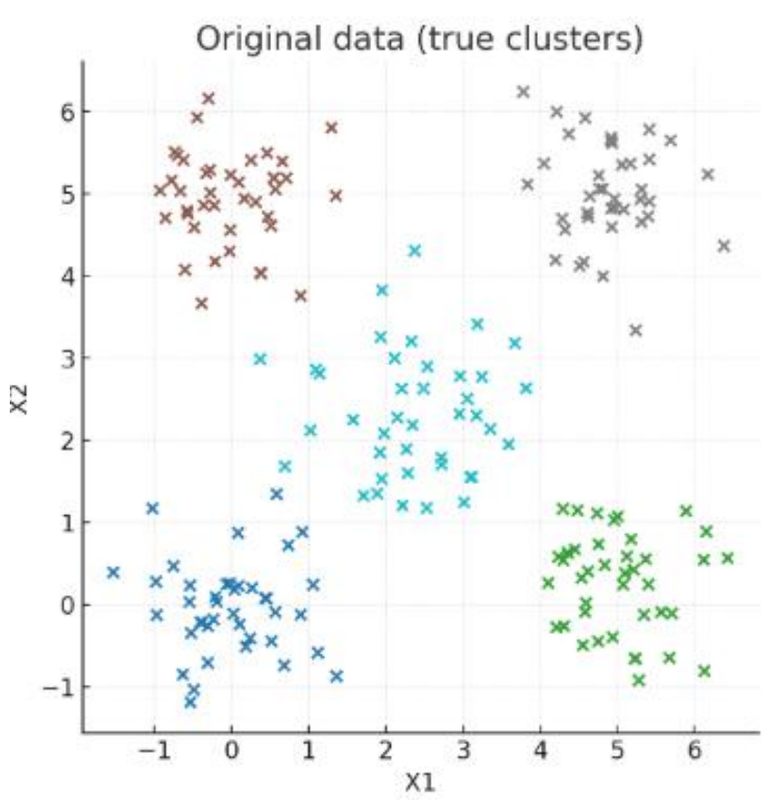
Elbow method in hierarchical clustering

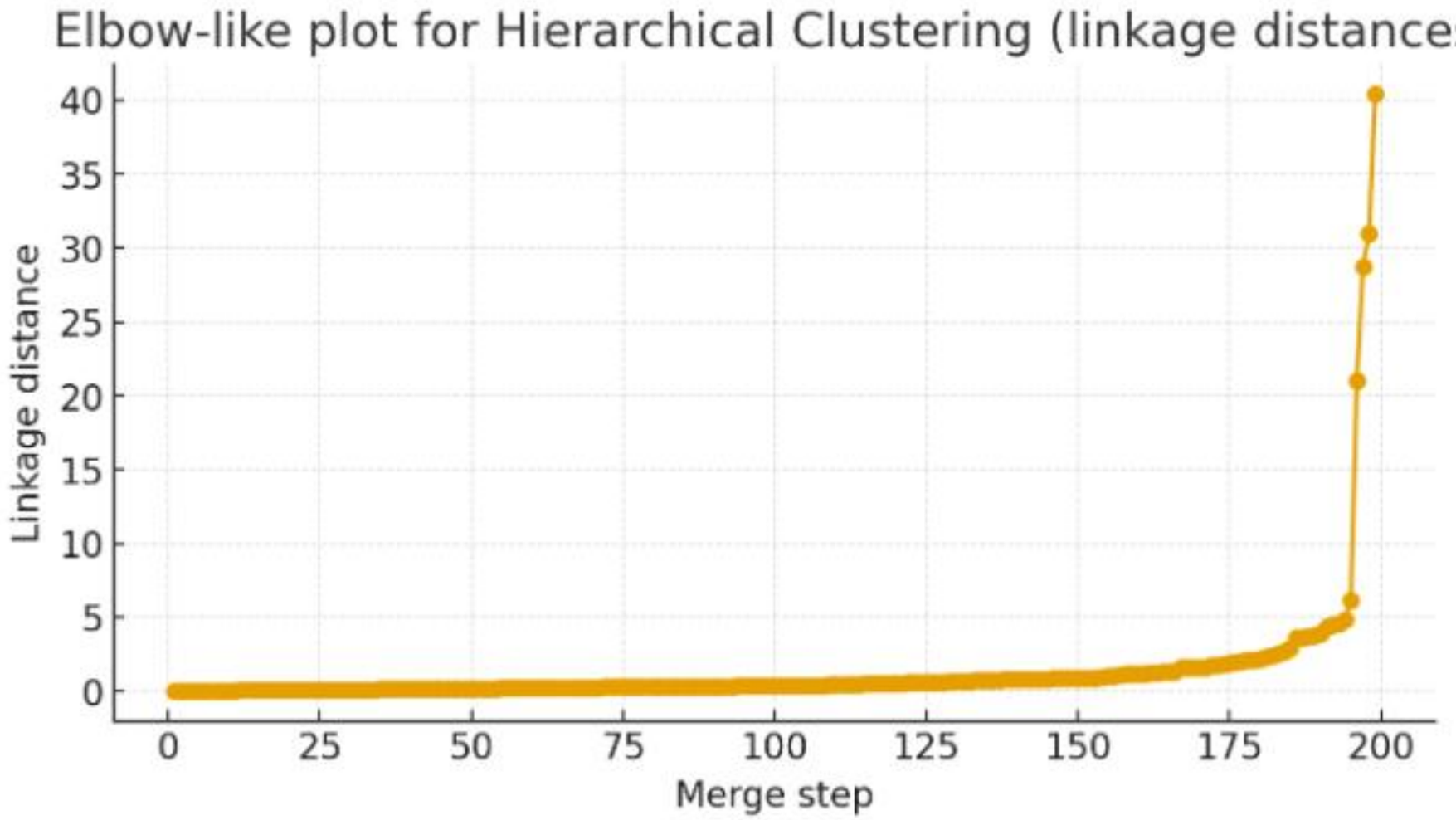


- Based on sudden increases in linkage distance. Early merges = small distances (similar points combined).
- Later merges = big distances (very different clusters forced together).
- The **elbow** is where the jump sharply increases.

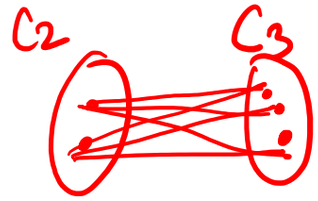


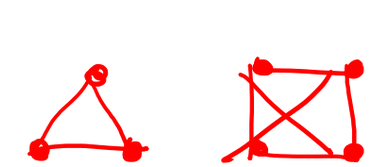
Elbow method in hierarchical clustering



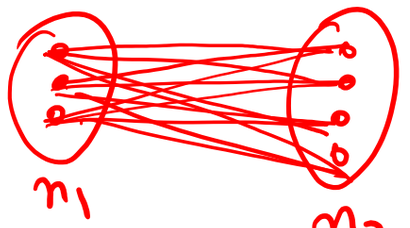


The Beta-CV Measure

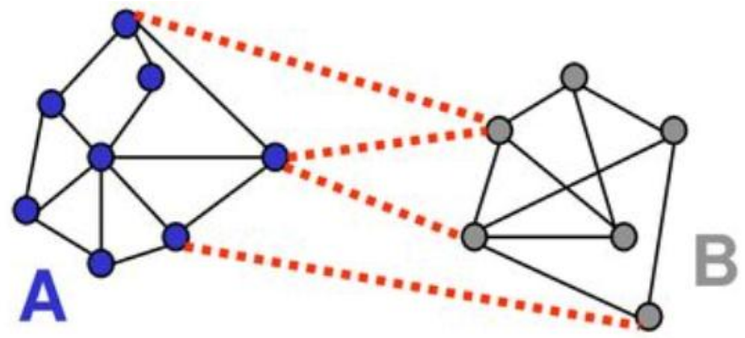




- Let W be the pair-wise distance matrix for all the given points. For any two point sets S and R, we define:



$$W(S,R) = \sum_{\mathbf{x}_i \in S} \sum_{\mathbf{x}_j \in R} w_{ij}$$



The sum of all the intracluster and intercluster weights are given as

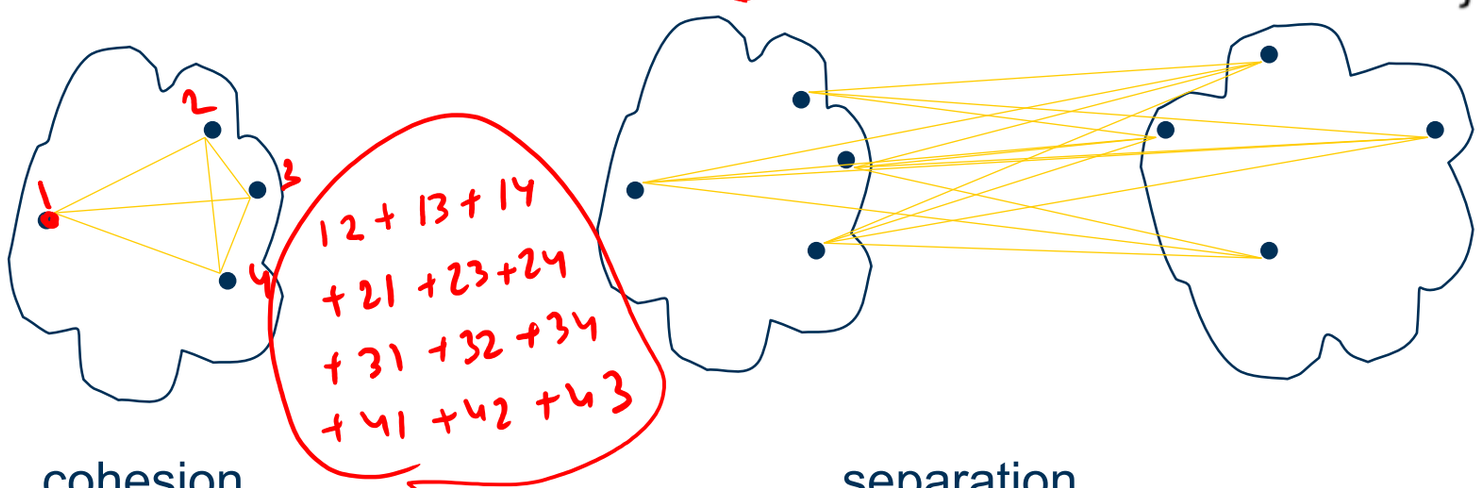
total cohesion

$$W_{in} = \frac{1}{2} \sum_{i=1}^{k} W(C_i, C_i)$$

total separation

$$W_{out} = \frac{1}{2} \sum_{i=1}^{k} W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1} \sum_{j>i} W(C_i, C_j)$$

The distance of each point is measured two times



Georgia Tech

The Beta-CV Measure

The number of distinct intracluster and intercluster edges is given as:

$$N_{in} = \sum_{i=1}^{k} \binom{n_i}{2} \qquad N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i \cdot n_j$$

BetaCV Measure: The BetaCV measure is the ratio of the mean intracluster distance to the mean intercluster distance:

$$BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{N_{out}}{N_{in}} \cdot \frac{W_{in}}{W_{out}} = \frac{N_{out}}{N_{in}} \frac{\sum_{i=1}^{k} W(C_i, C_i)}{\sum_{i=1}^{k} W(C_i, \overline{C_i})}$$

The smaller the BetaCV ratio, the better the clustering.

normalized cohesion / normalized separation

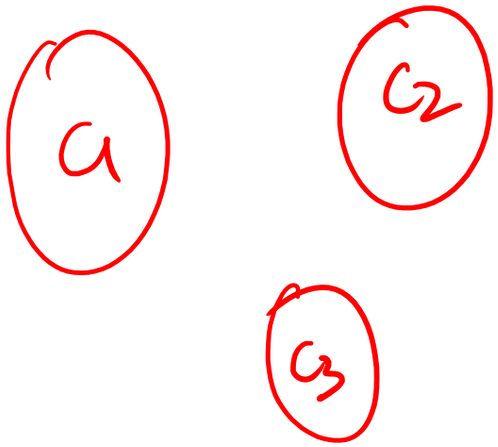
Image credit: Adriana Kovashka

Normalized Cut

Normalized cut: $$NC = \sum_{i=1}^{k} \frac{W(C_i, \overline{C}_i)}{vol(C_i)} = \sum_{i=1}^{k} \frac{W(C_i, \overline{C}_i)}{W(C_i, \mathrm{V})} = \sum_{i=1}^{k} \frac{W(C_i, \overline{C}_i)}{W(C_i, C_i) + W(C_i, \overline{C}_i)} = \sum_{i=1}^{k} \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \overline{C}_i)} + 1}$$

where $vol(C_i) = W(C_i, V)$ is the volume of cluster C_i

The higher normalized cut value, the better the clustering



$$\frac{W(C_1, C_2) + W(C_2, C_3) + W(C_3, C_1)}{W(C_1, C_2) + W(C_2, C_3) + W(C_3, C_1) + W(C_1, C_1) + W(C_2, C_2) + W(C_3, C_3)}$$

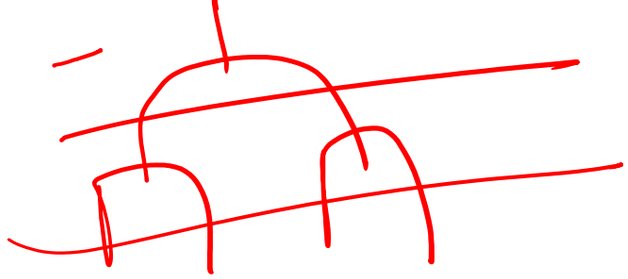
$W(C_i, C_i)$ — Intra-cluster distance

Separation / Separation + cohesion

$W(C_i, \overline{C_i})$ — Inter-cluster distance

high value

Quick Knowledge Check



1. In agglomerative hierarchical clustering, do we start with: a) All data points in one cluster. b) Each data point as its own cluster

2. Which linkage method is most sensitive to chaining and outliers? a) Single link. b) Complete link. c) Average link

3. Cutting a dendrogram at a higher distance threshold produces: a) More clusters. b) Fewer clusters

4. Which of the following is a real-world use case of hierarchical clustering? a) Organizing scientific papers into a taxonomy. b) Predicting stock prices. c) Training a deep reinforcement learning agent

5. The elbow method in hierarchical clustering is based on: a) Within-cluster sum of squared errors. b) Sudden increases in linkage distance. c) Average silhouette score