L09 Hierarchical clustering

Simpsons slides from Tom Mitchell & Ziv Bar-Joseph (CMU 10-601)

Jason Fleischer

Two Types of Clustering

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion (focus of this class)

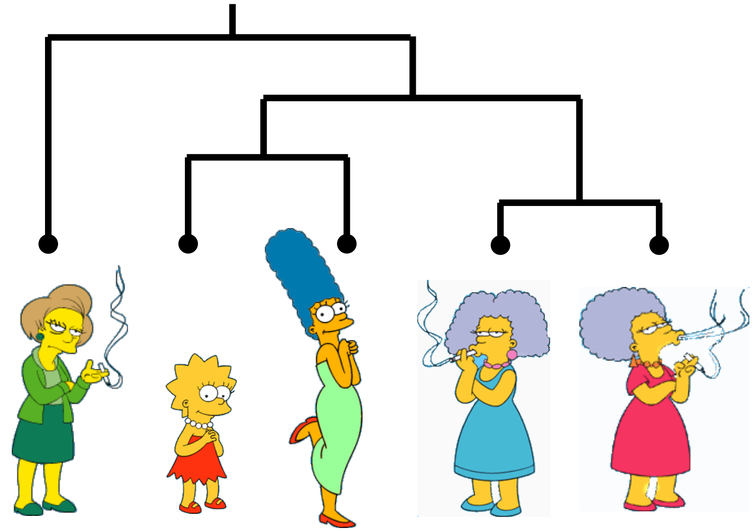
Bottom up or top down

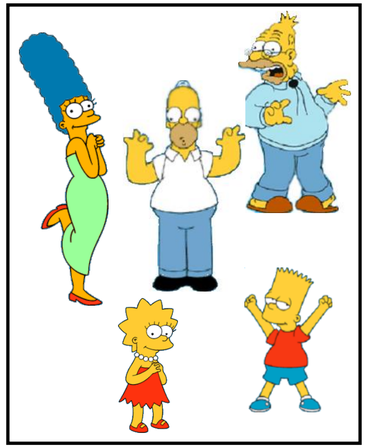
Top down

Hierarchical

Partitional

This is a dendrogram

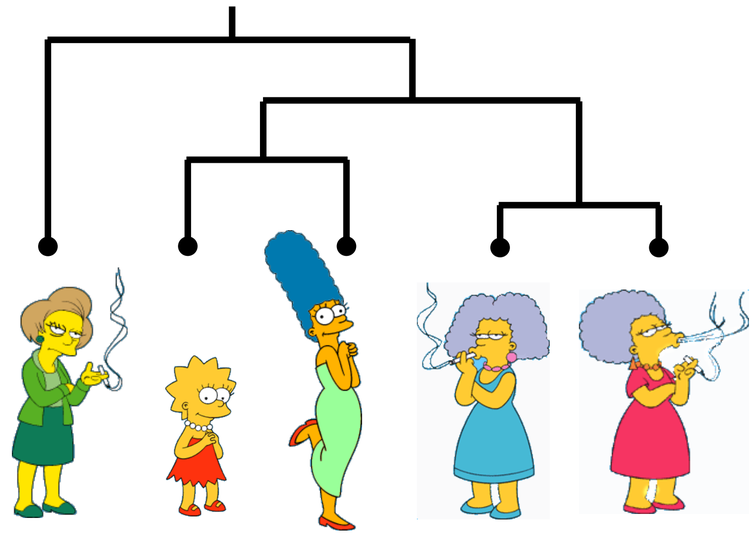




(How-to) Hierarchical Clustering

The number of dendrograms with n leafs $= (2n-3)!/[(2^{(n-2)})(n-2)!]$

Number of Leafs	Number of Possible Dendrograms
2	1
3	3
4	15
5	105
...	…
10	34,459,425

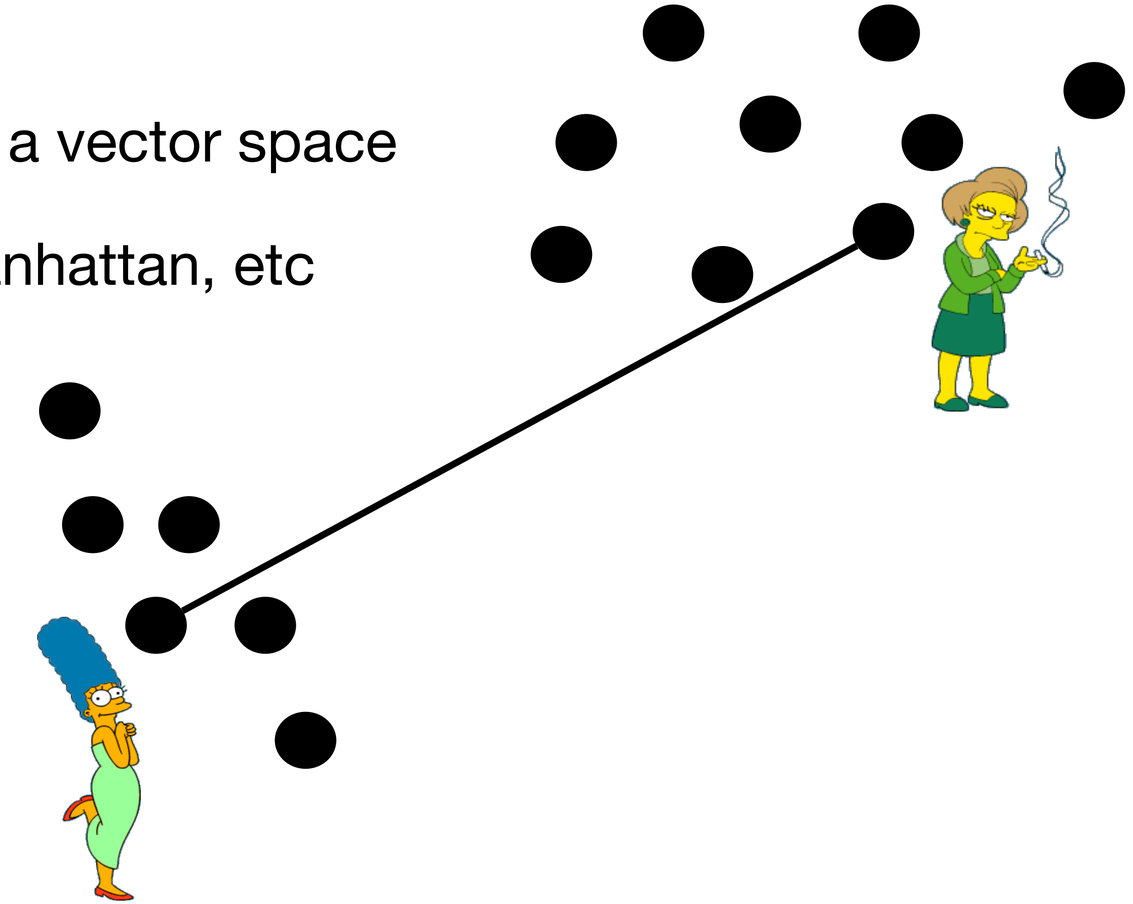


Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

What is the distance between two datapoints?

Start with a distance metric

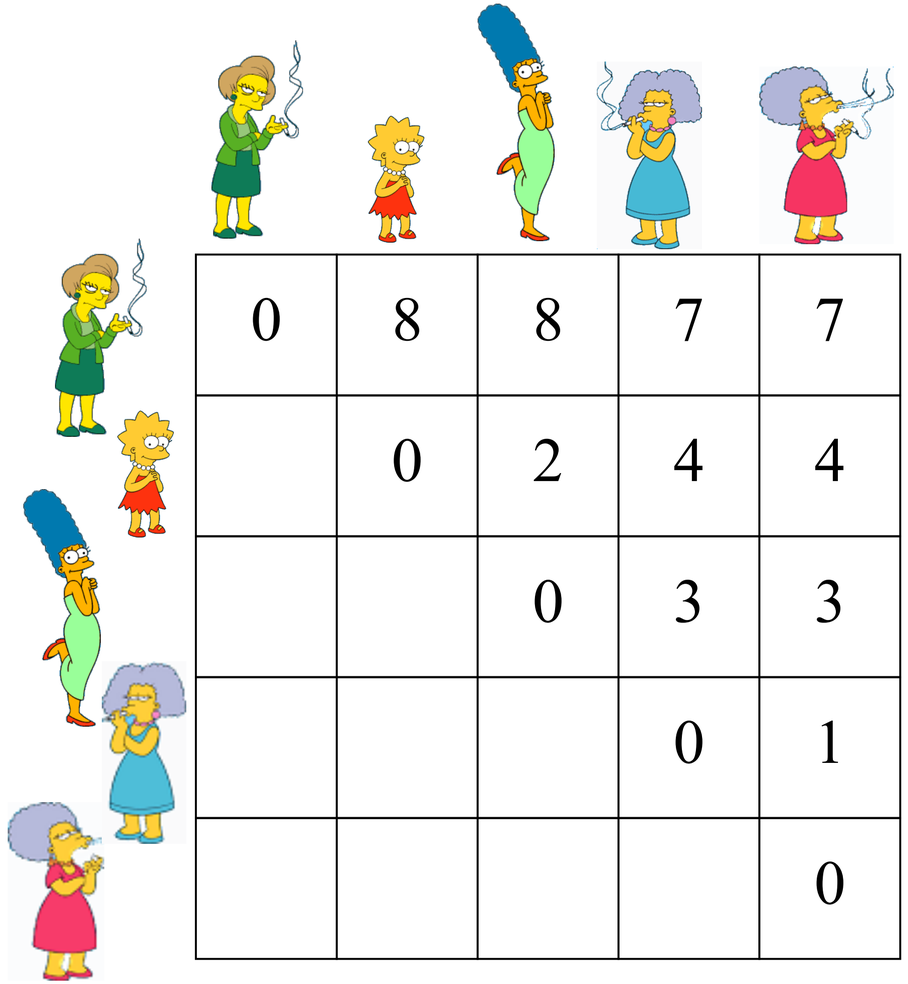
- $d(x, x')$
- Metric operating in a vector space
- e.g., Euclidean, Manhattan, etc



We begin with a distance matrix which contains the distances between every pair of objects in our database.

D() = 8

D() = 1

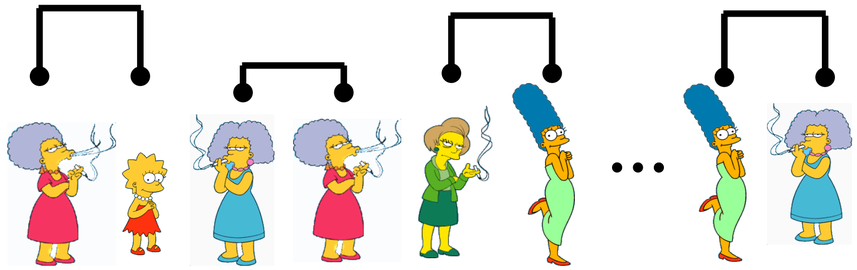


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

Bottom-Up (agglomerative):

Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

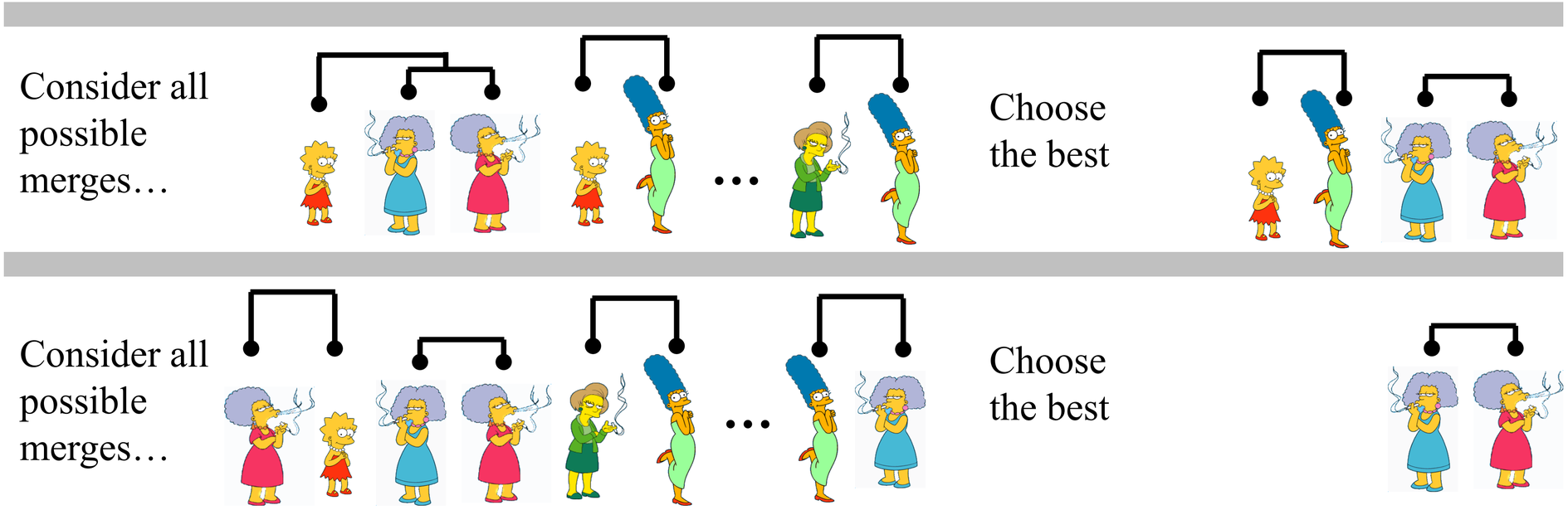
Consider all possible merges…



Choose the best

Bottom-Up (agglomerative):

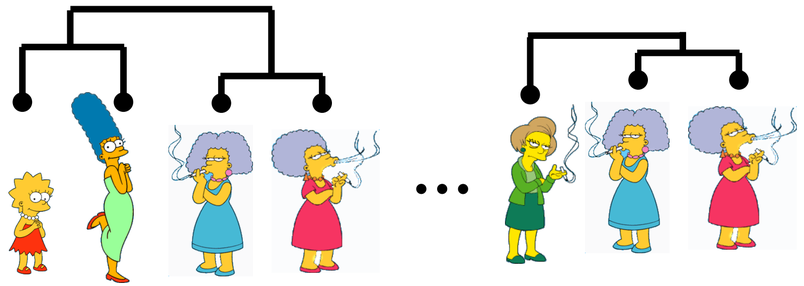
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



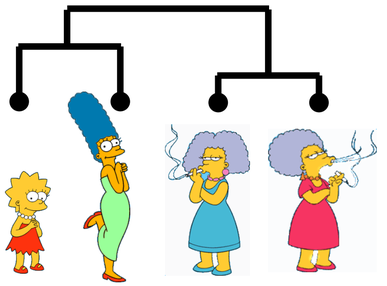
Bottom-Up (agglomerative):

Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

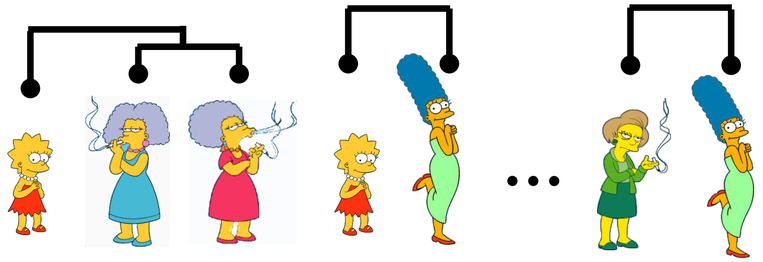
Consider all possible merges…



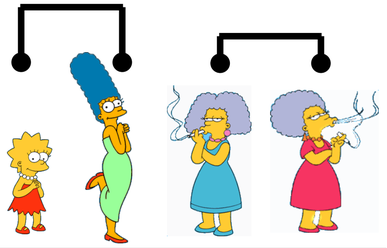
Choose the best



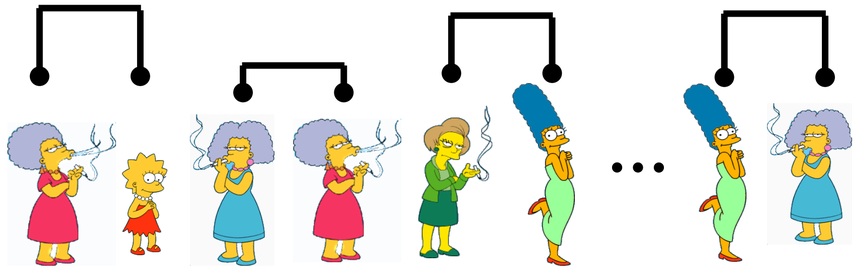
Consider all possible merges…



Choose the best



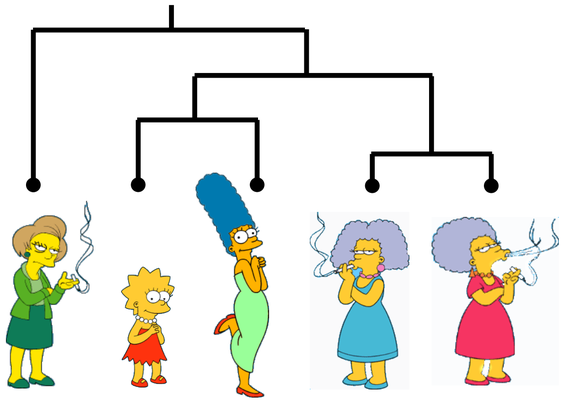
Consider all possible merges…



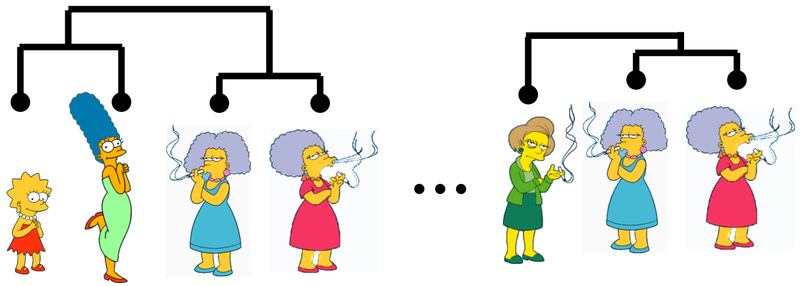
Choose the best

Bottom-Up (agglomerative):

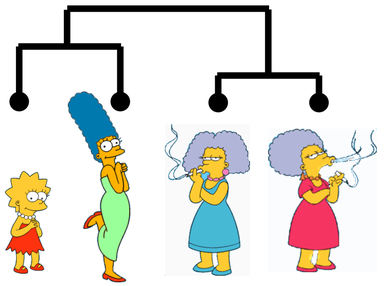
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



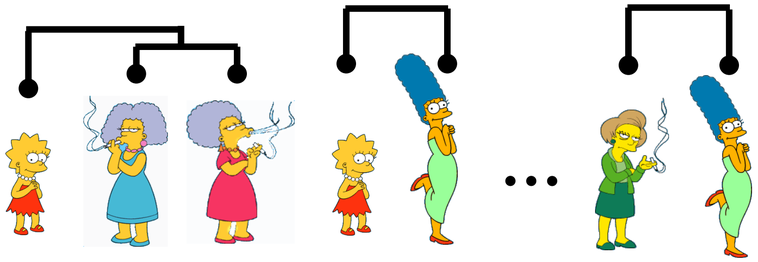
Consider all possible merges…



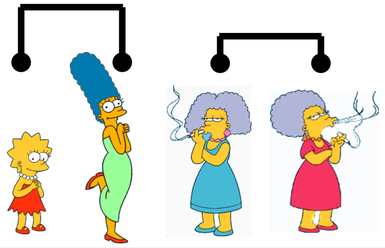
Choose the best



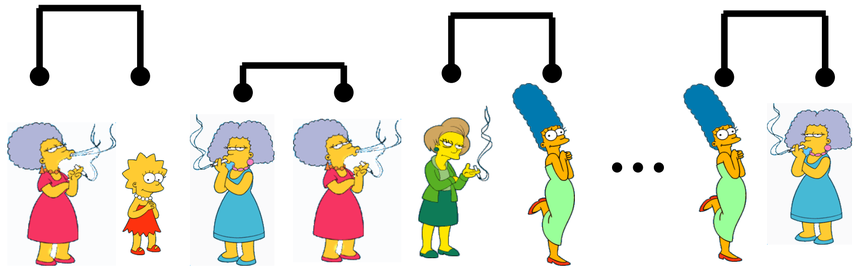
Consider all possible merges…



Choose the best



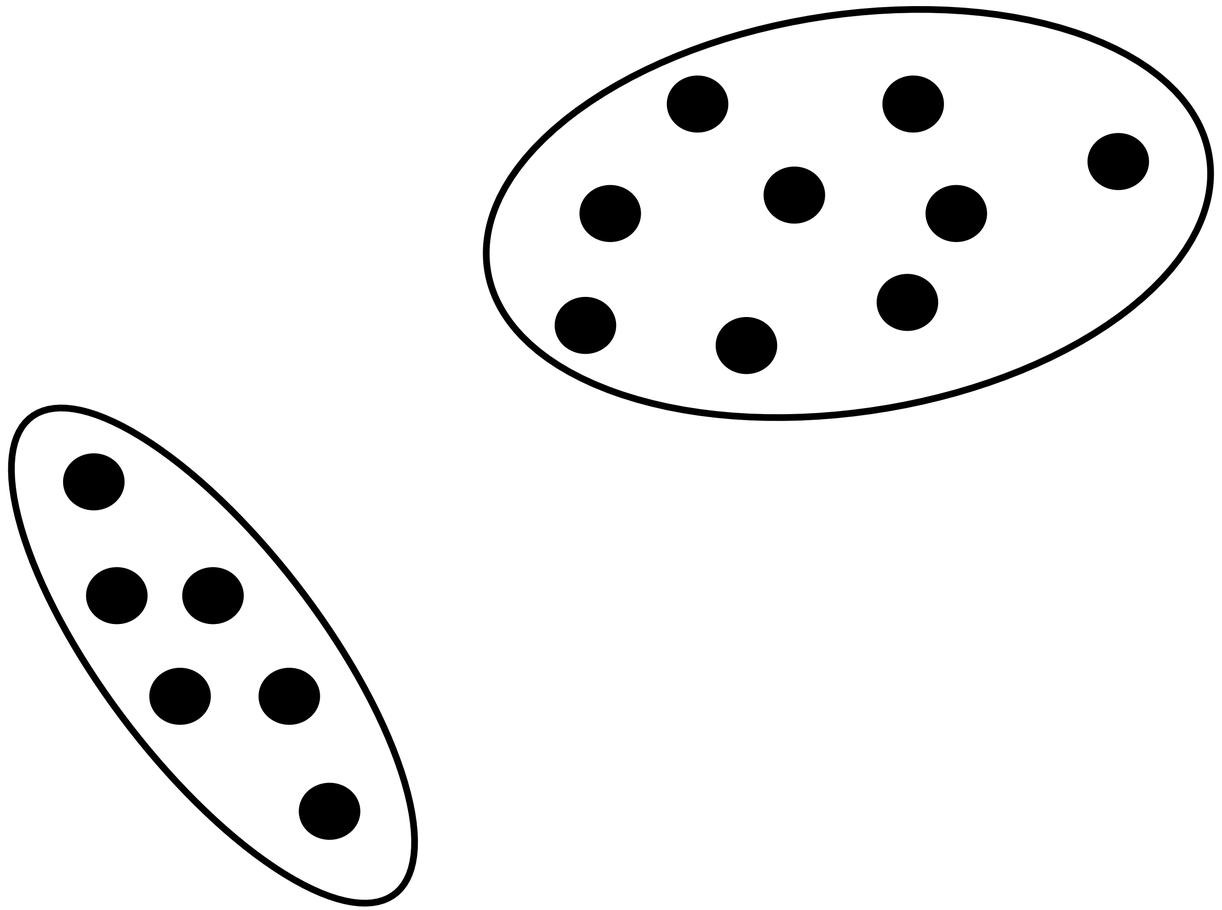
Consider all possible merges…



Choose the best

Distance between two clusters

What does it even mean to measure this?

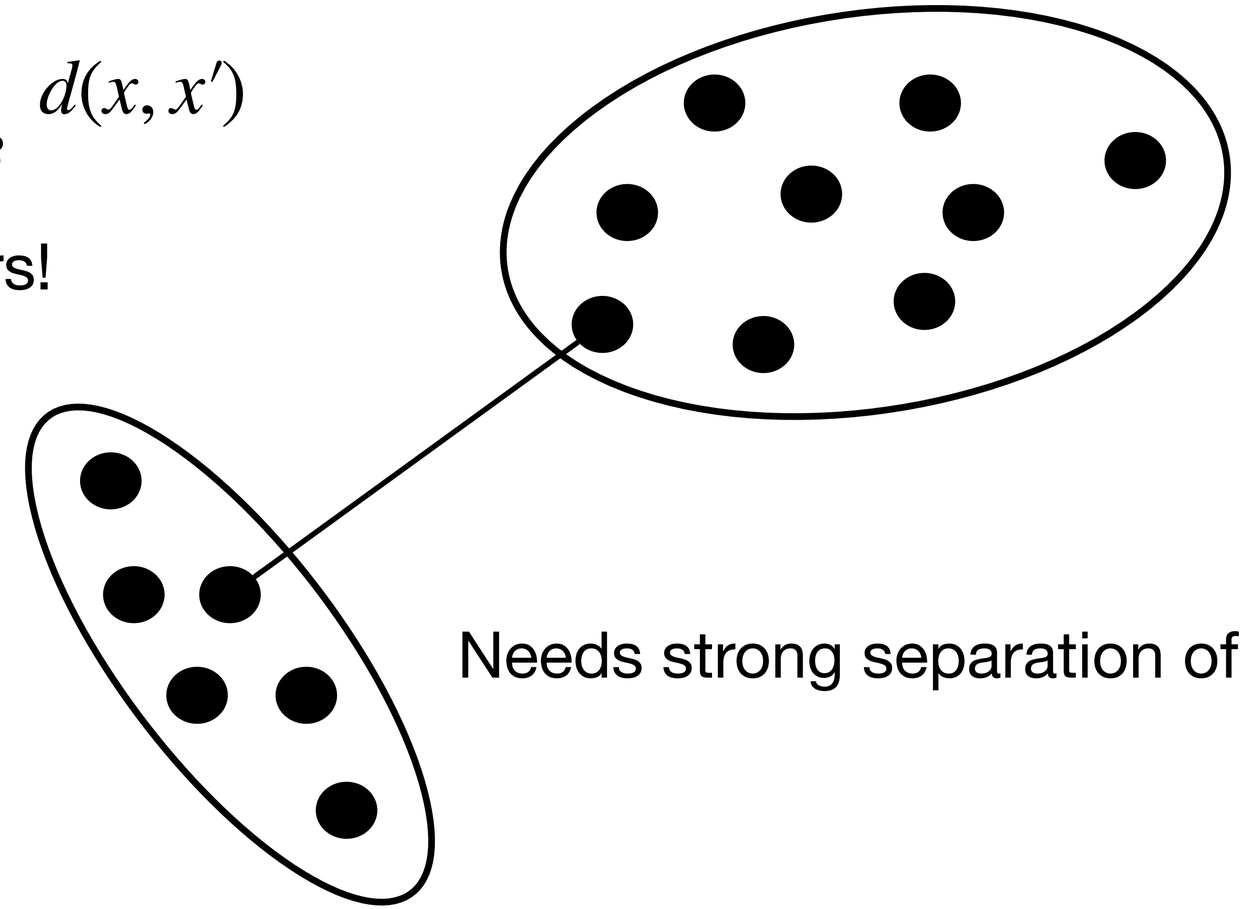


What is the distance between two clusters?

Single linkage

$$d(A, B) = \min_{x \in A, x' \in B} d(x, x')$$

Sensitive to outliers!



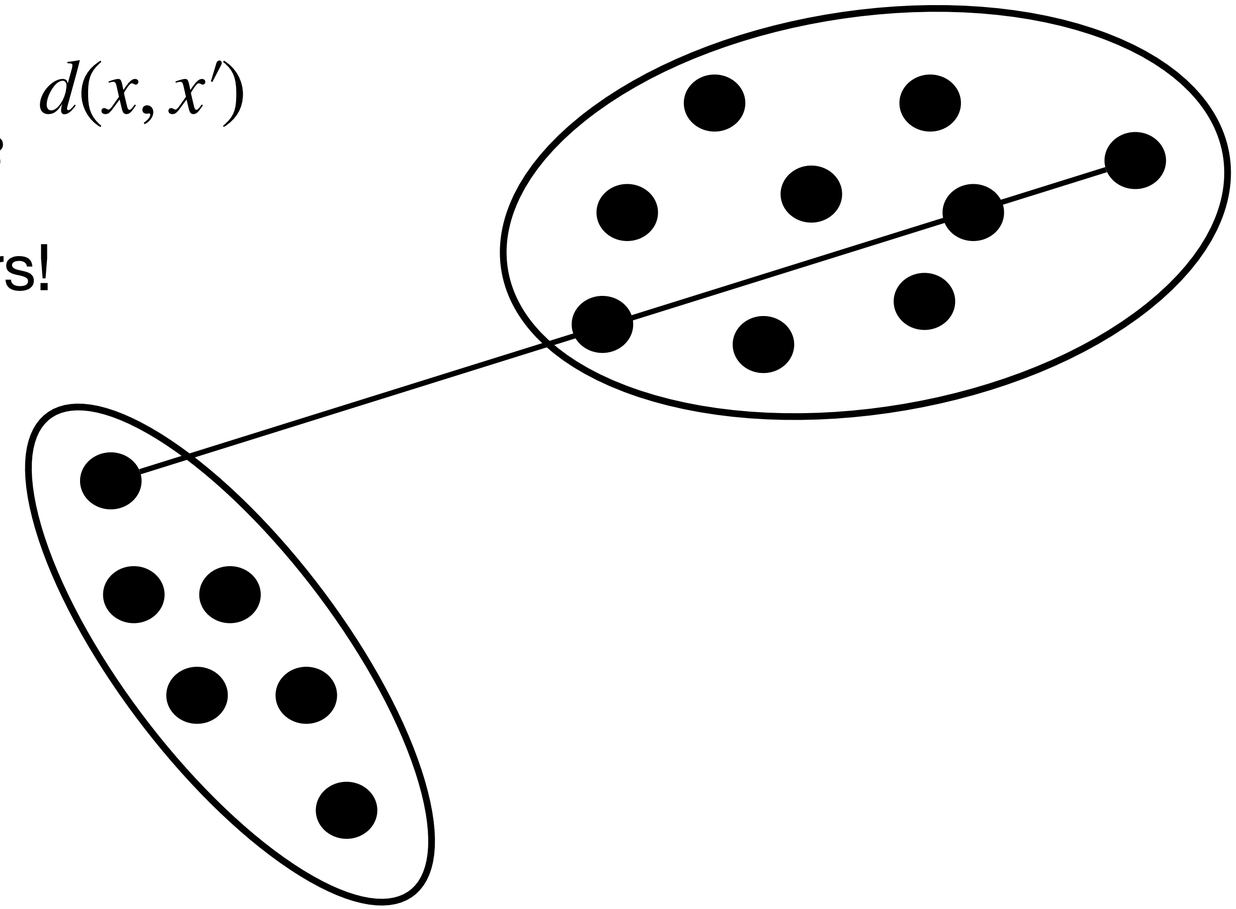
Needs strong separation of compact clusters

What is the distance between two clusters?

Complete linkage

$$d(A, B) = \max_{x \in A, x' \in B} d(x, x')$$

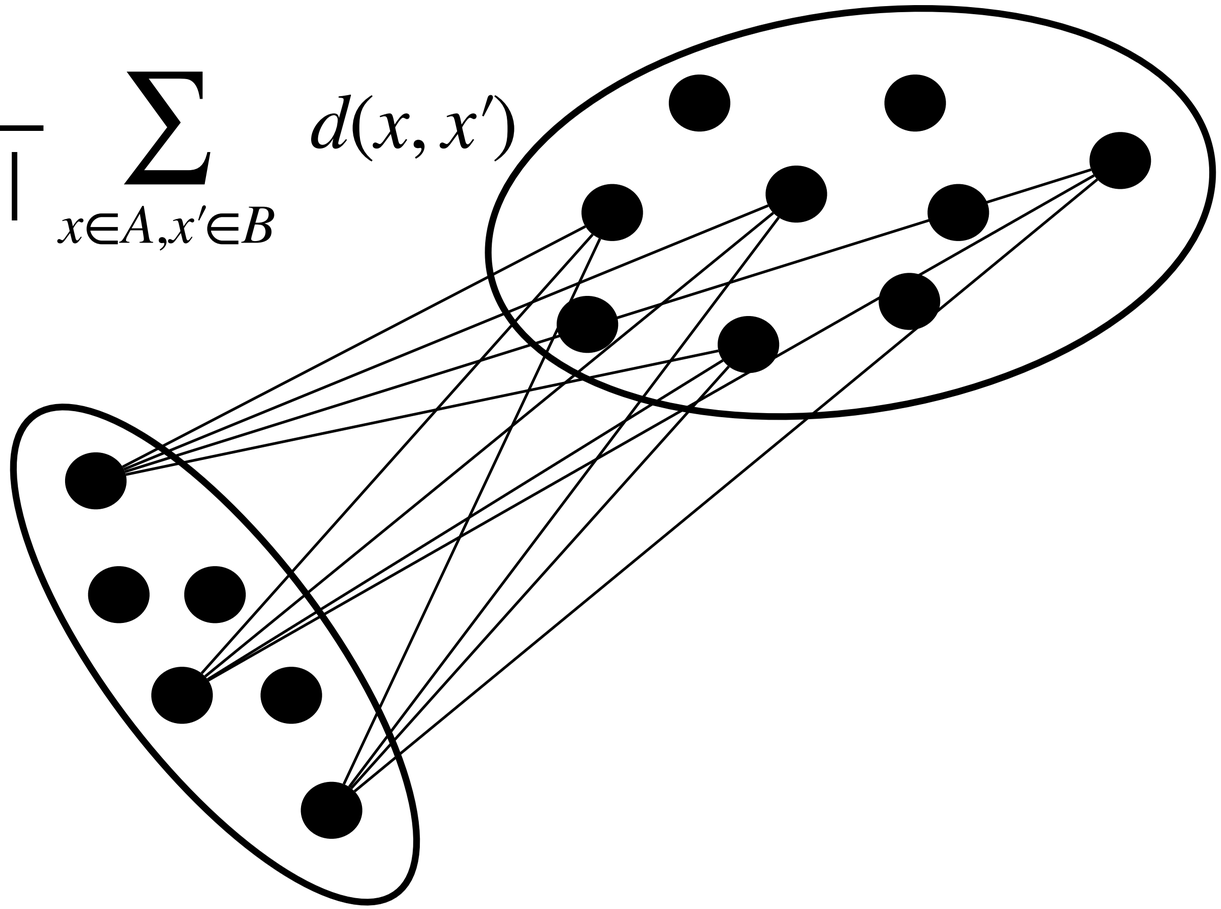
Sensitive to outliers!



What is the distance between two clusters?

Average linkage

$$d(A, B) = \frac{1}{|A||B|} \sum_{x \in A, x' \in B} d(x, x')$$

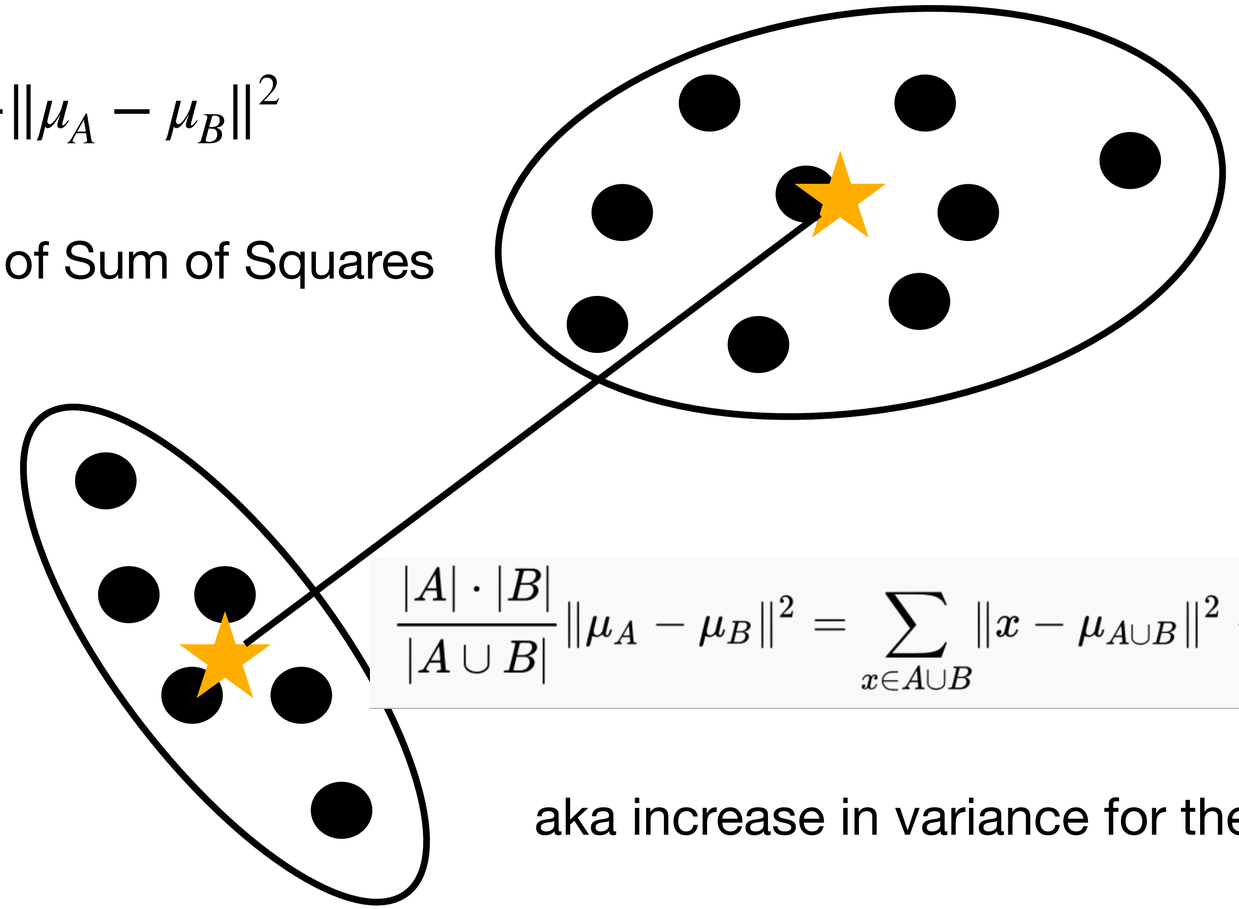


What is the distance between two clusters?

Ward linkage

$$d(A,B) = \frac{|A||B|}{|A|+|B|}\|\mu_A - \mu_B\|^2$$

aka Minimum Increase of Sum of Squares

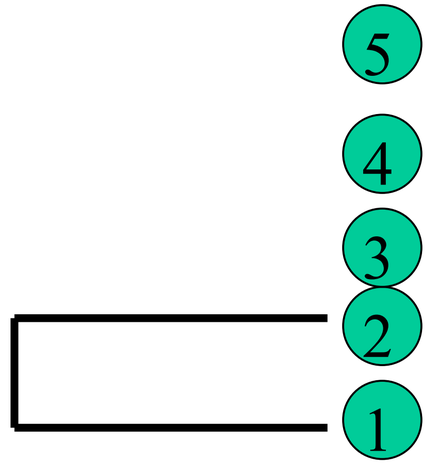


$$\frac{|A|\cdot|B|}{|A\cup B|}\|\mu_A-\mu_B\|^2 = \sum_{x\in A\cup B}\|x-\mu_{A\cup B}\|^2 - \sum_{x\in A}\|x-\mu_A\|^2 - \sum_{x\in B}\|x-\mu_B\|^2$$

aka increase in variance for the cluster being merged

Example: single link

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0



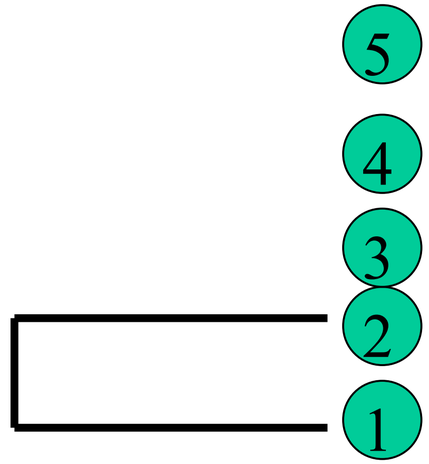
Example: single link

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6,3\} = 3$$

$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10,9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9,8\} = 8$$

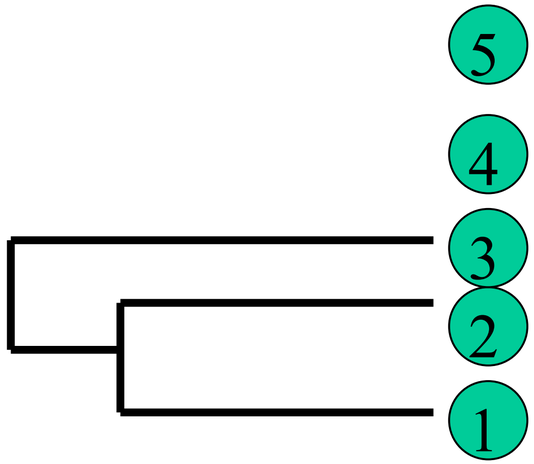


Example: single link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9,7\} = 7$$

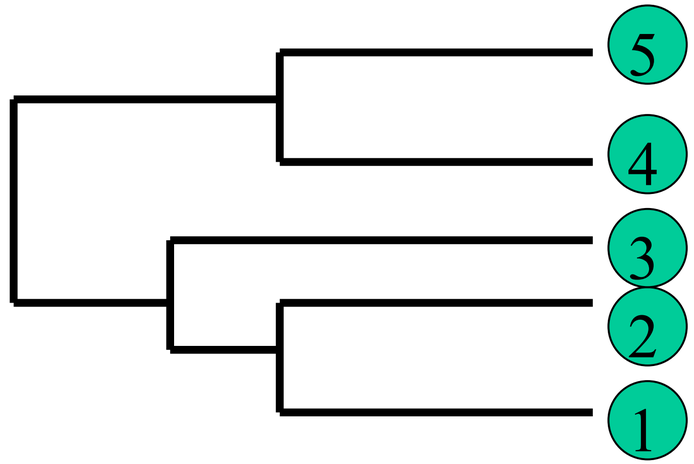
$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8,5\} = 5$$



Example: single link

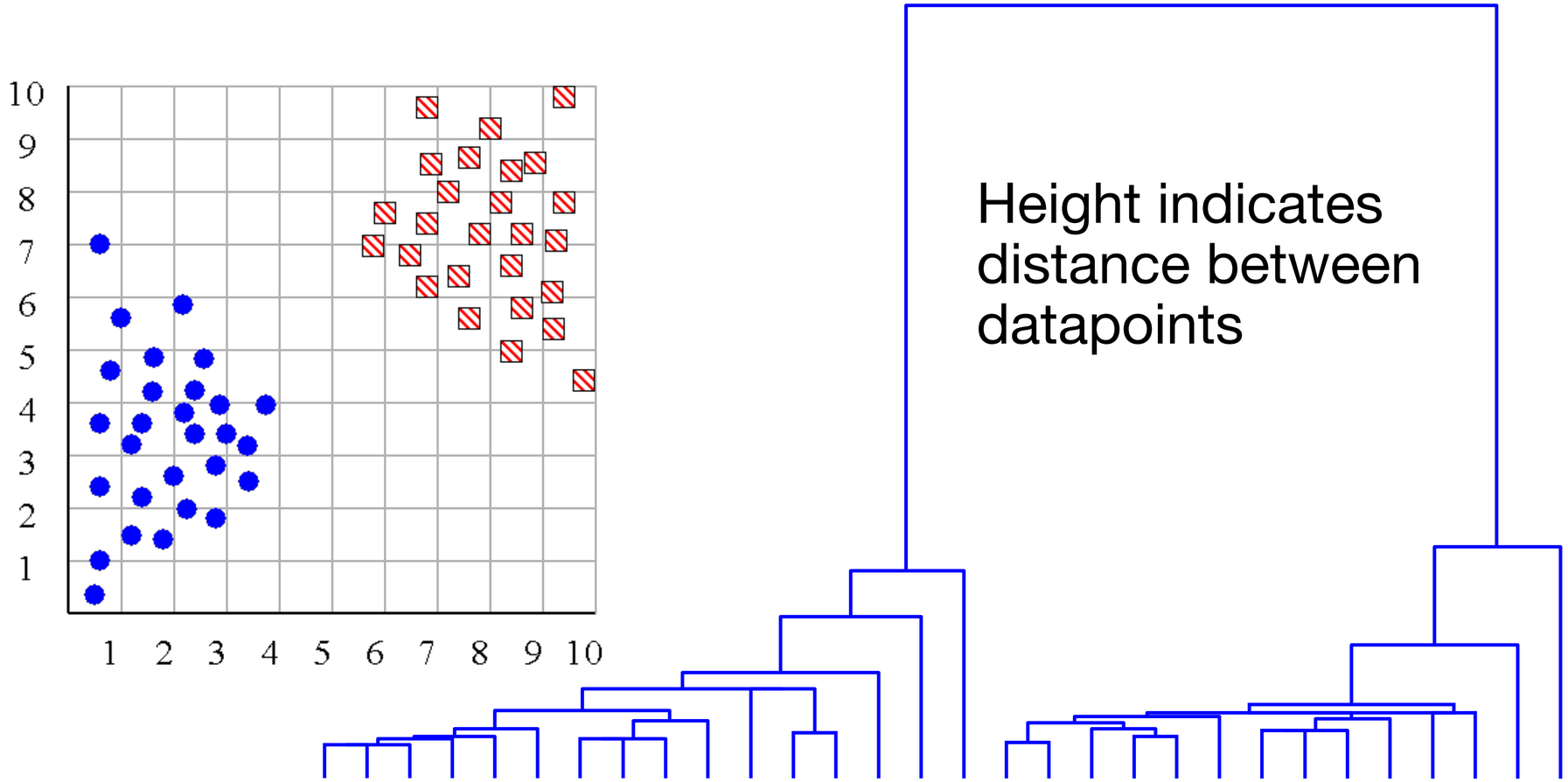
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2,3) \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccc} (1,2,3) & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$



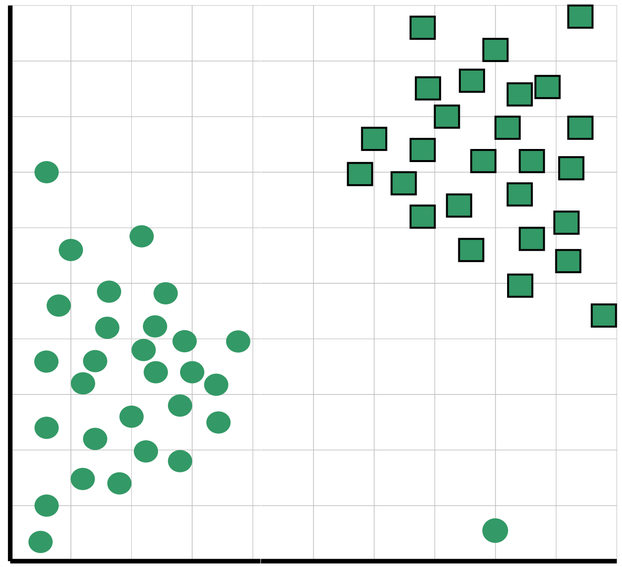
But what are the clusters?

In some cases we can determine the “correct” number of clusters.
However, things are rarely this clear cut, unfortunately.



One potential use of a dendrogram is to detect outliers

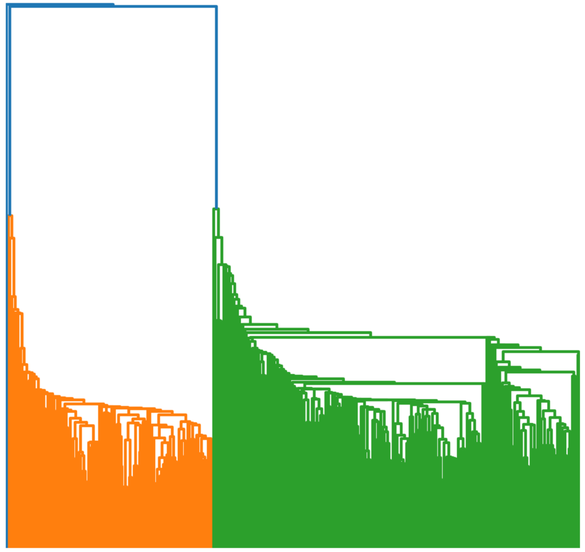
The single isolated branch is suggestive of a data point that is very different to all others



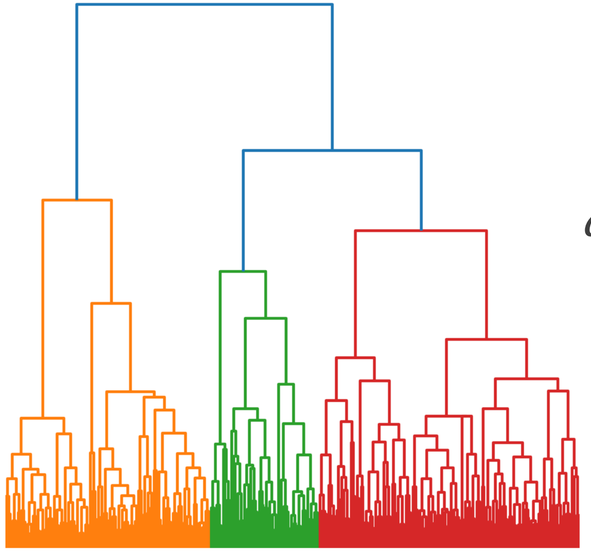
Outlier

$$d(A,B) = \min_{x \in A, x' \in B} d(x,x')$$

Single linkage



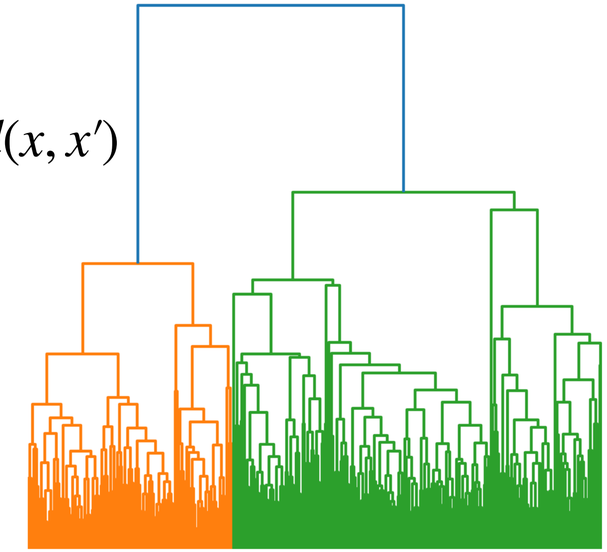
Complete linkage



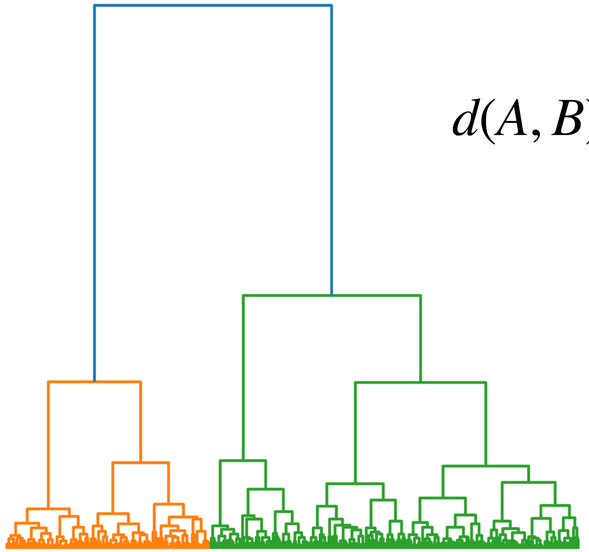
$$d(A,B) = \max_{x \in A, x' \in B} d(x,x')$$

$$d(A,B) = \frac{1}{|A|\,|B|} \sum_{x \in A, x' \in B} d(x,x')$$

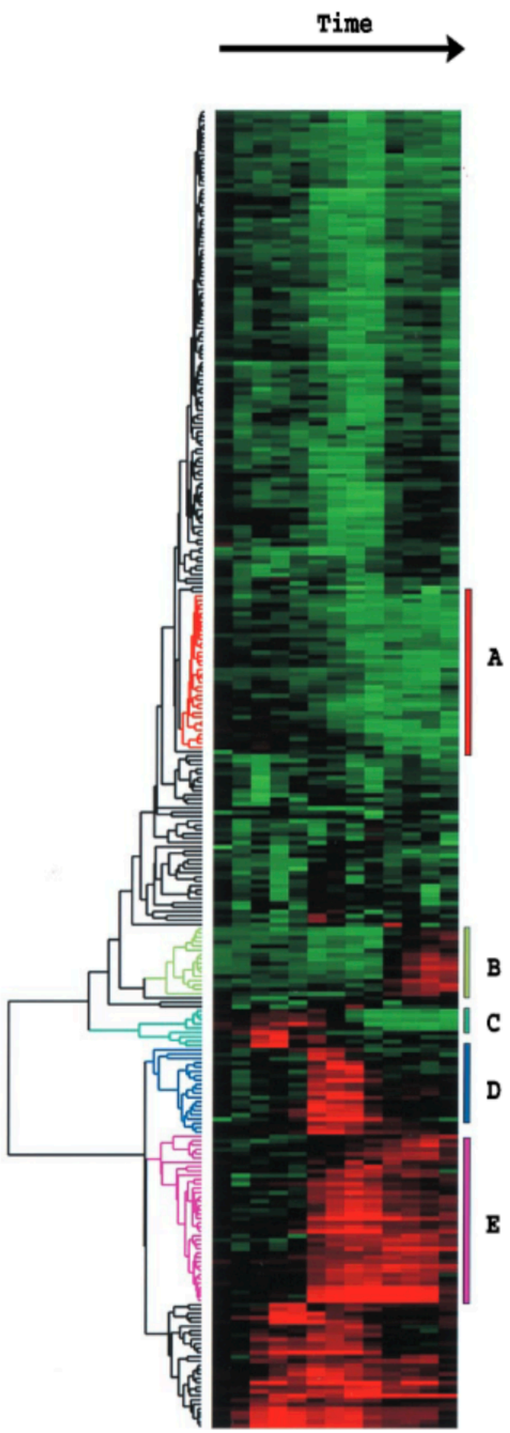
Average linkage



Ward linkage



$$d(A,B) = \frac{|A|\,|B|}{|A| + |B|} \|\mu_A - \mu_B\|^2$$



Hierarchical clustering is frequently used in science, esp genomics

MB Eisen et al, PNAS (1998) >20k citations shows that you can use these techniques to

- Demonstrate gene networks that co-express over time
- Infer function of a gene you didn't know about based on its co-expression partners in the cluster
- (A) cholesterol biosynthesis, (B) the cell cycle, (C) the immediate–early response, (D) angiogenesis, and (E) wound healing and tissue remodeling.