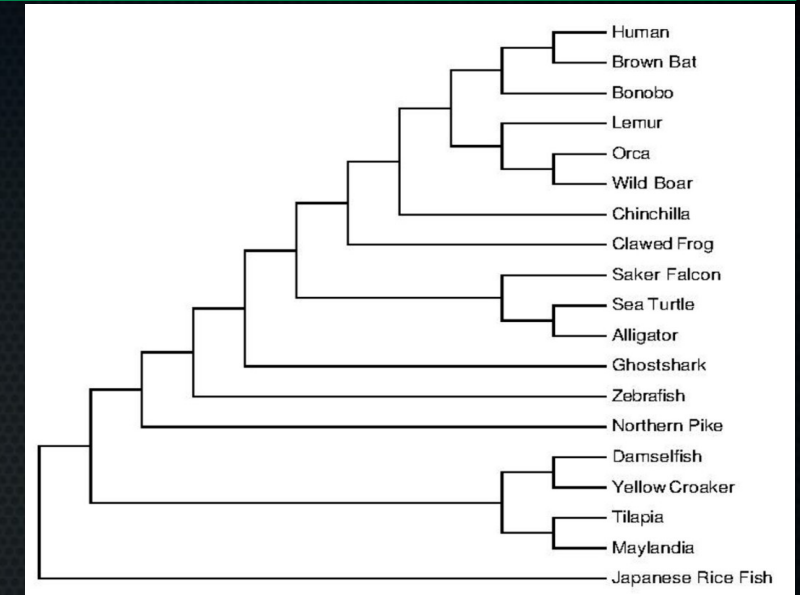


This presentation

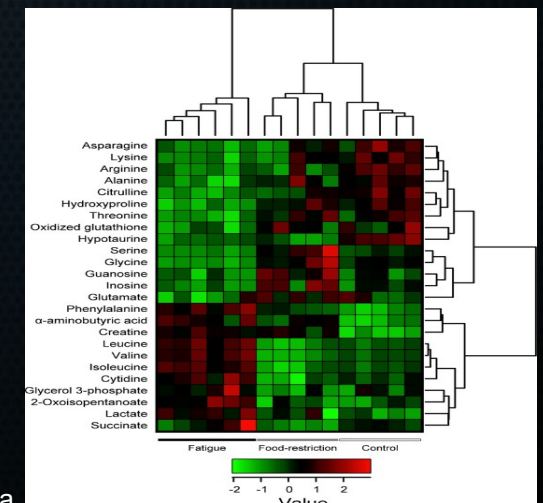
- Hierarchical clustering
- Linkage methods
- ***Note for BiBC Master students: this presentation is exactly equal to the Essentials course one except for 4 slides at the end. If you feel comfortable with hierarchical clustering, feel free to tune out and start on the short practical!***

Agglomerative or hierarchical clustering

- Make a tree connecting all samples, which you can separate into clusters at any level you like
- Start from clusters containing individual samples, stop when you've agglomerated all clusters into one big clustering.



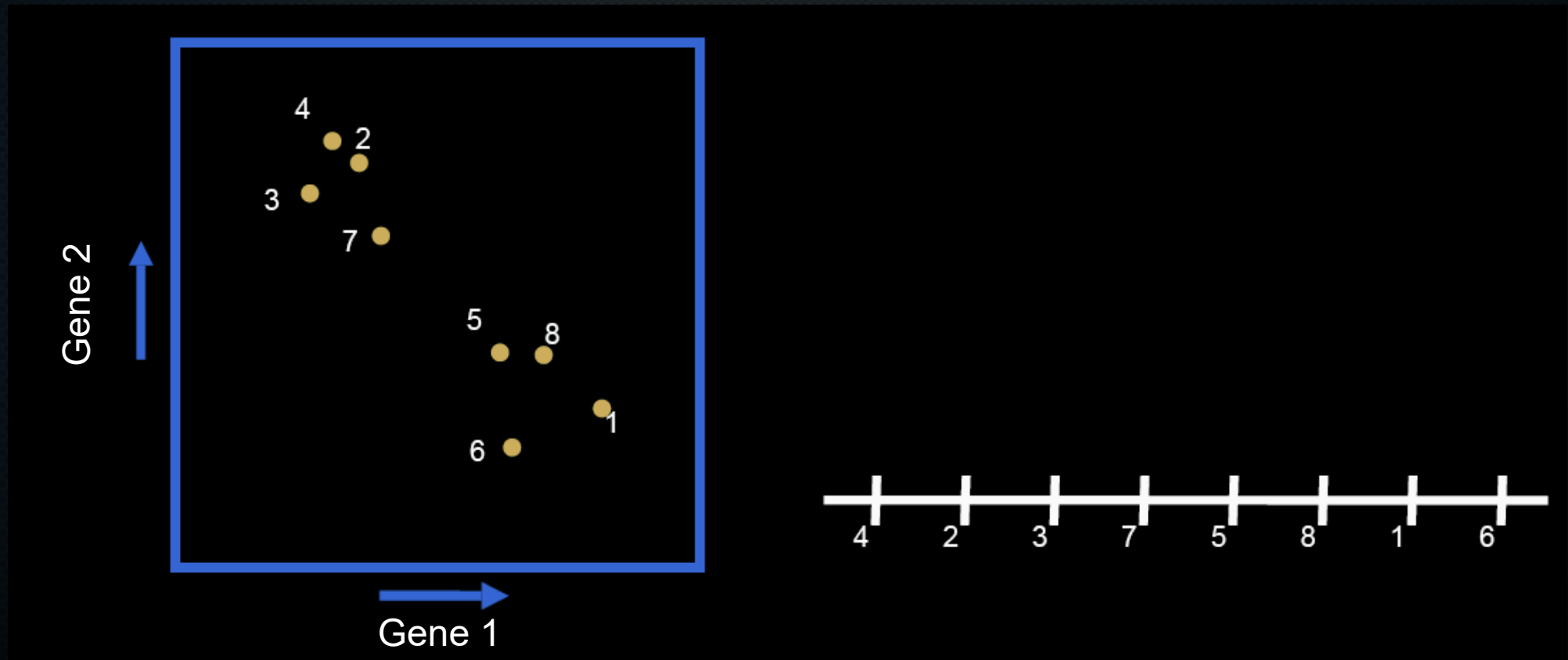
Source: https://commons.wikimedia.org/wiki/File:Phylogenetic_Tree.pdf



Source: Kume, S., Yamato, M., Tamura, Y., Jin, G., Nakano, M., Miyashige, Y., ... & Kataoka, Y. (2015). Potential biomarkers of fatigue identified by plasma metabolome analysis in rats. PloS one, 10(3), e0120106.

Agglomerative or hierarchical clustering

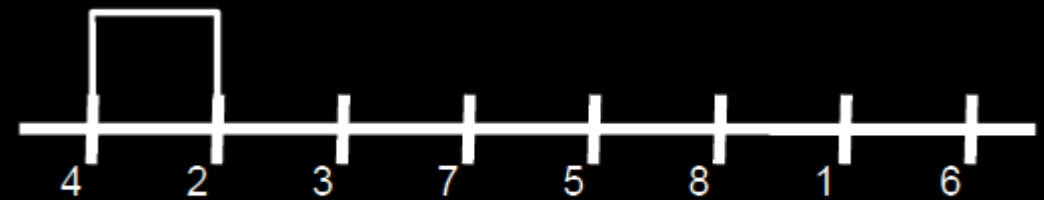
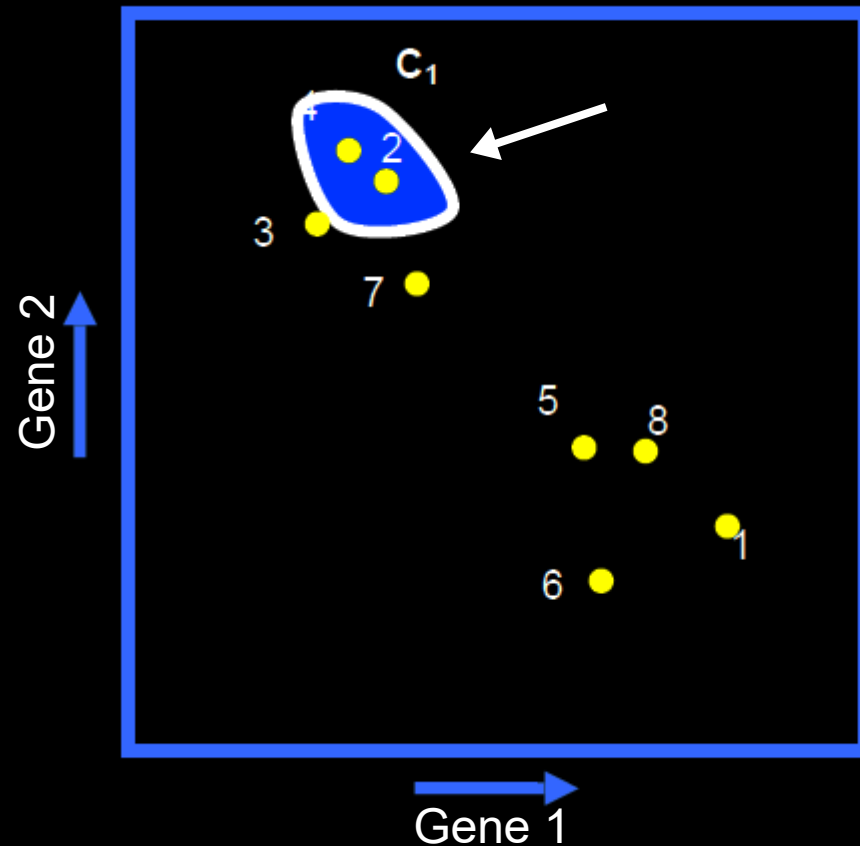
- Let's look at an example: find most similar objects and group them



Source: Jeroen de Ridder

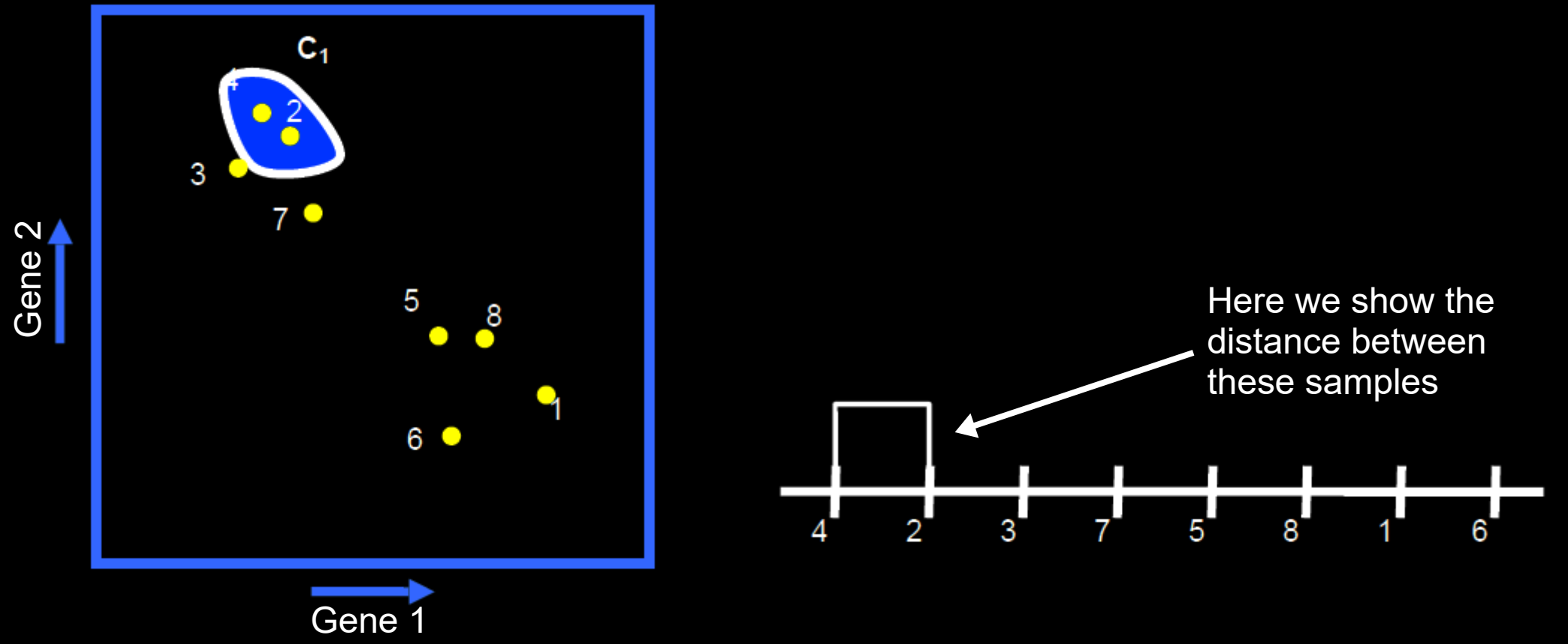
Agglomerative or hierarchical clustering

- Let's look at an example: find most similar objects and group them



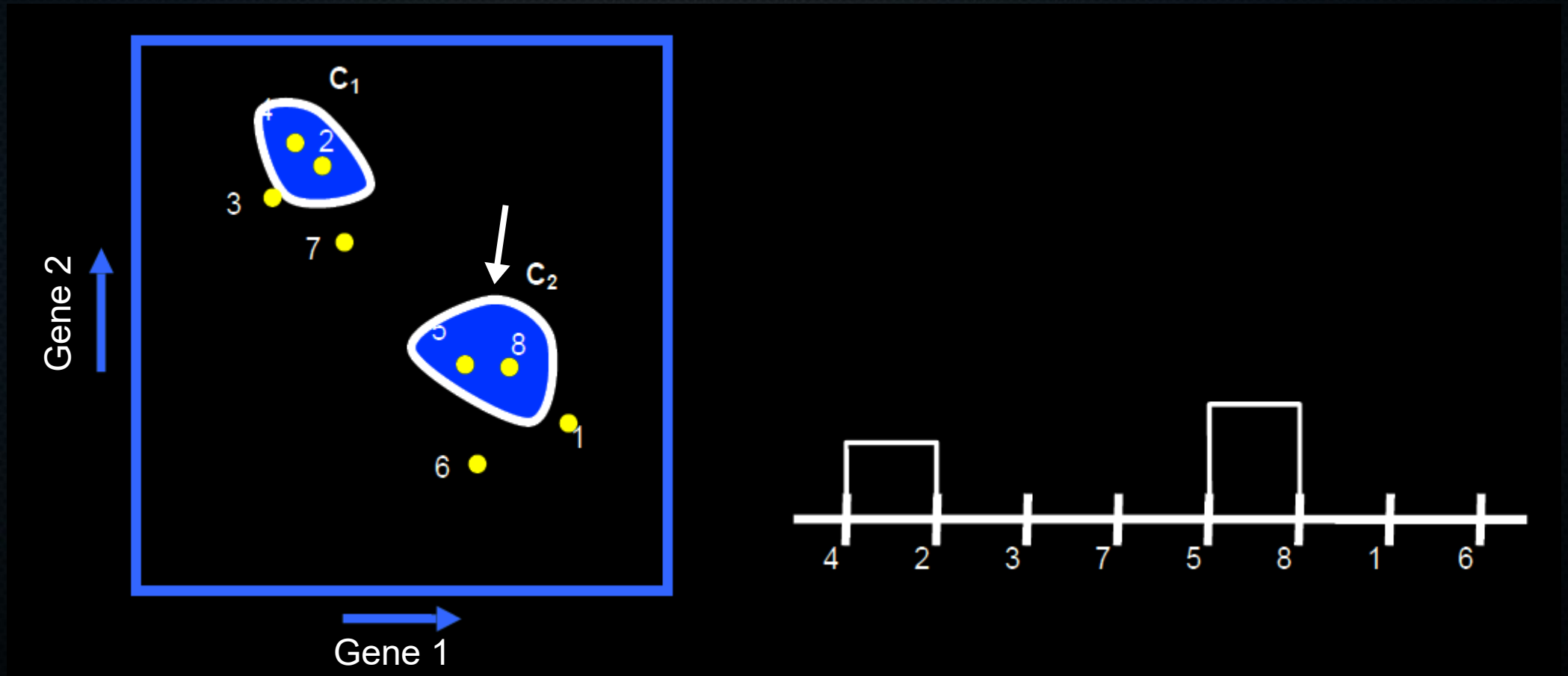
Agglomerative or hierarchical clustering

- Let's look at an example: find most similar objects and group them



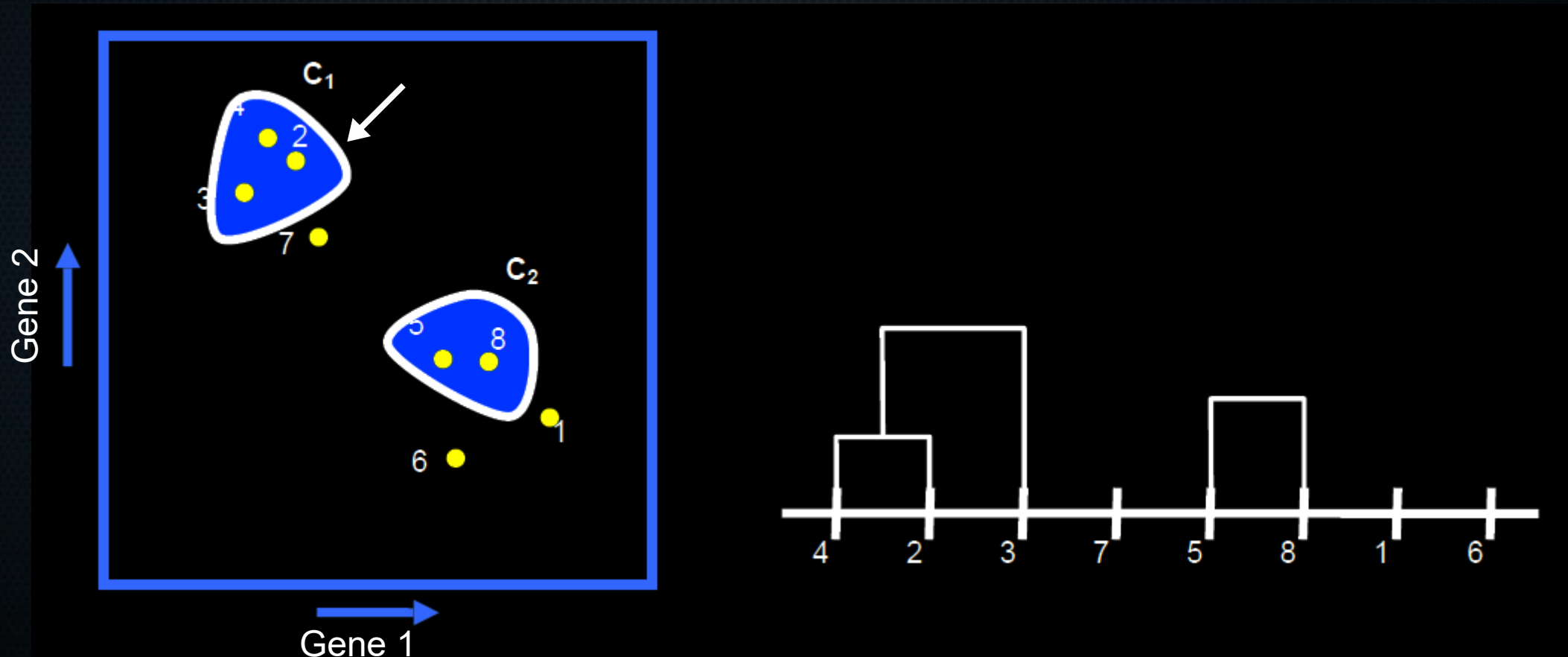
Agglomerative or hierarchical clustering

- Again, find most similar objects and group them



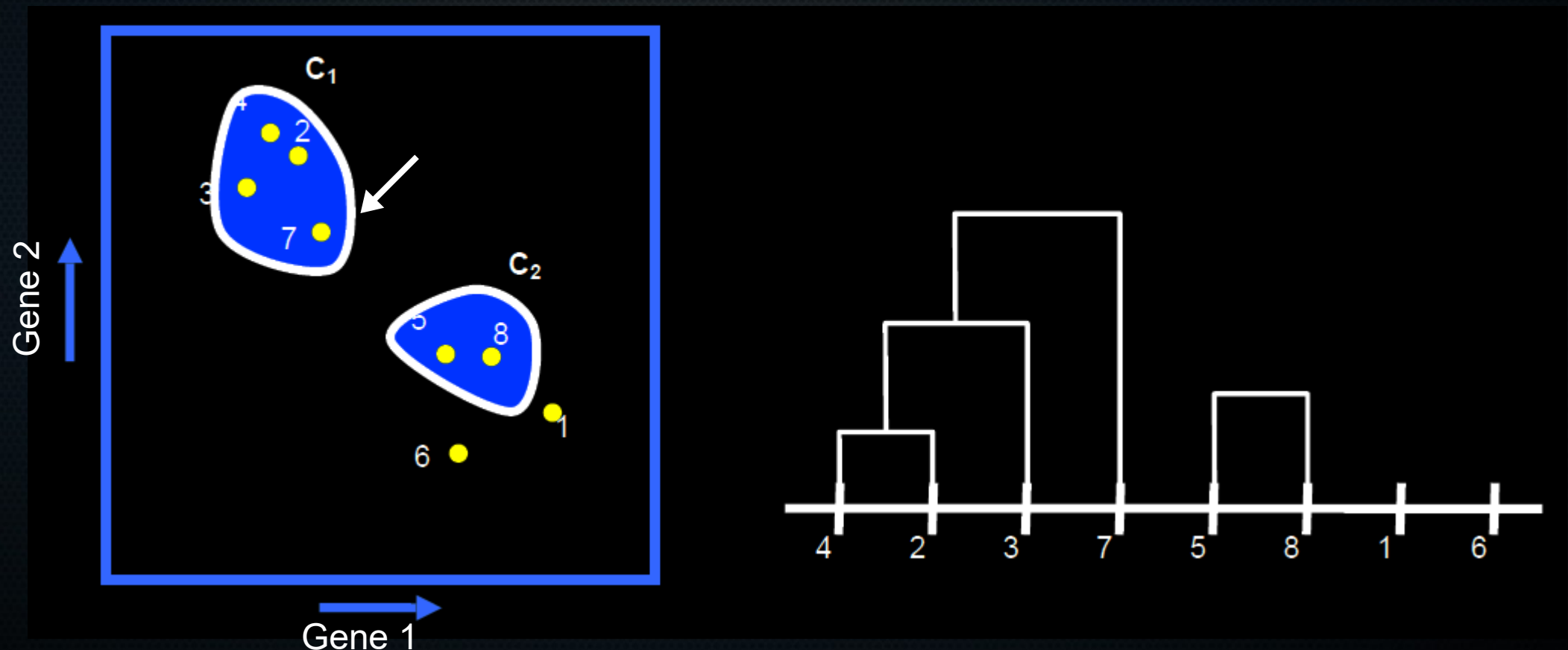
Agglomerative or hierarchical clustering

- Now: closest distance between cluster 1 and point 3!



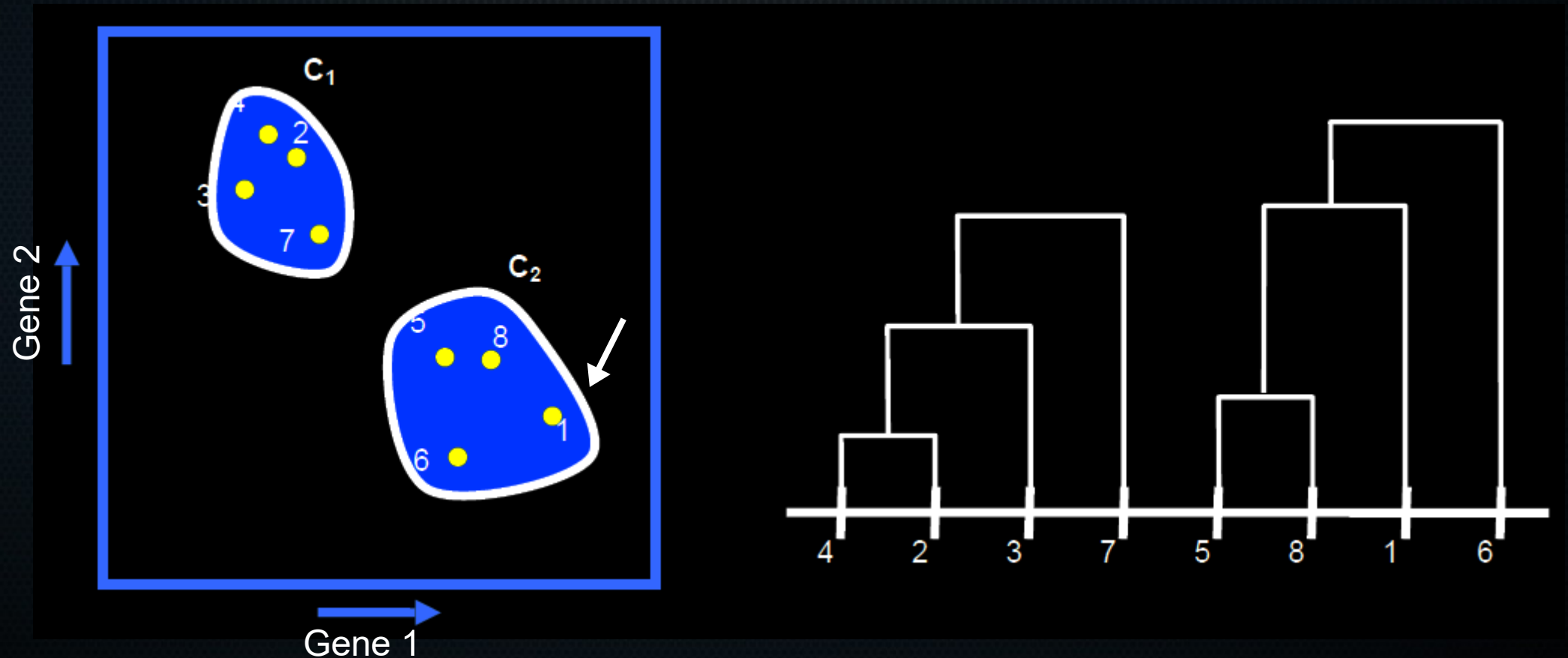
Agglomerative or hierarchical clustering

- Keep iterating until everything is clustered together



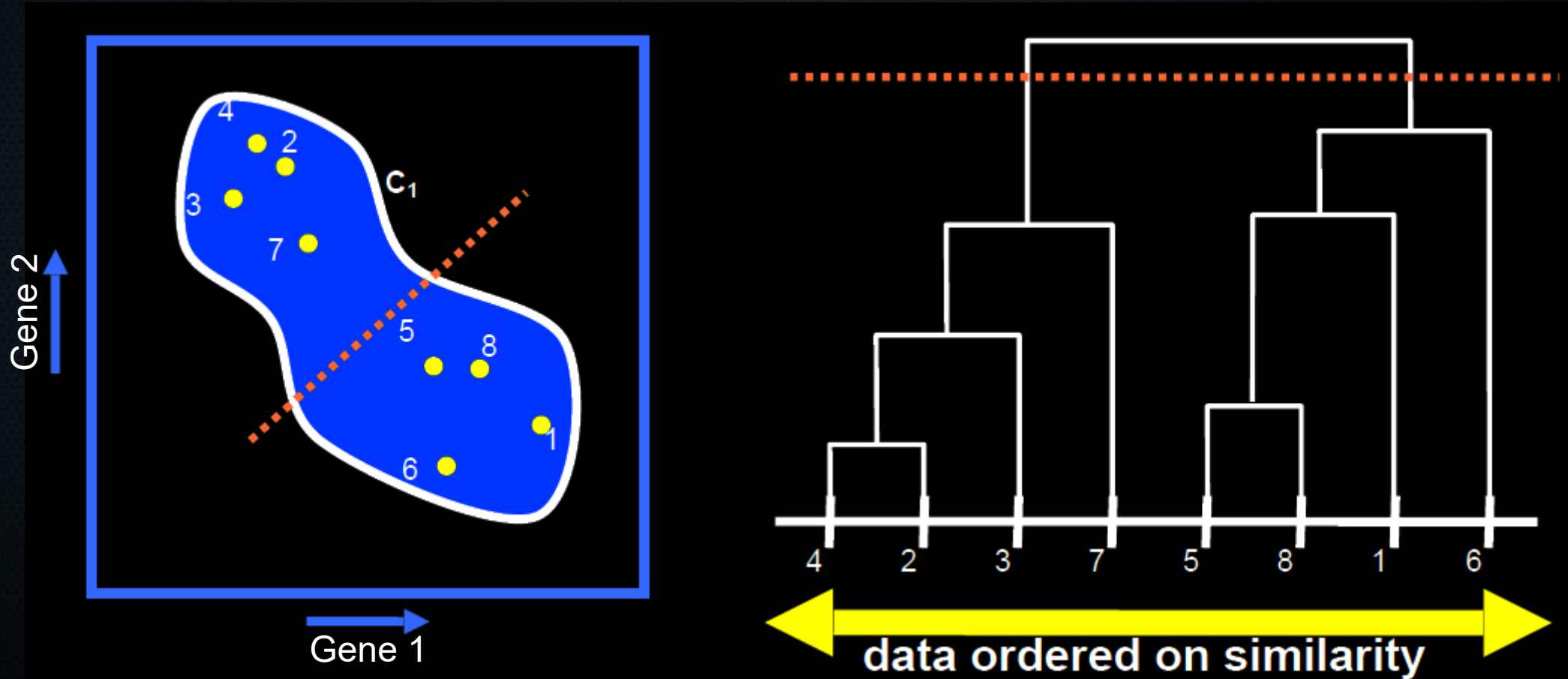
Agglomerative or hierarchical clustering

- Keep iterating until everything is clustered together



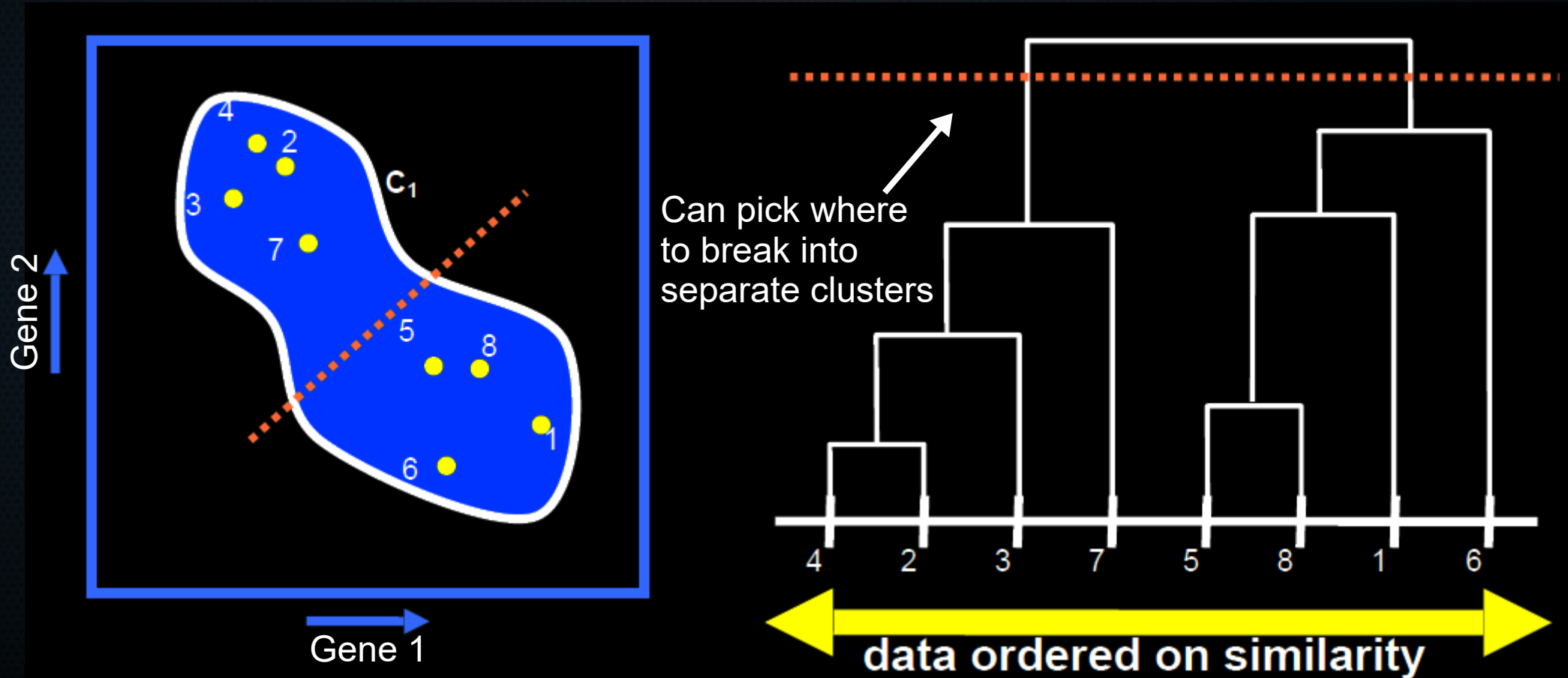
Agglomerative or hierarchical clustering

- Keep iterating until everything is clustered together → Done!



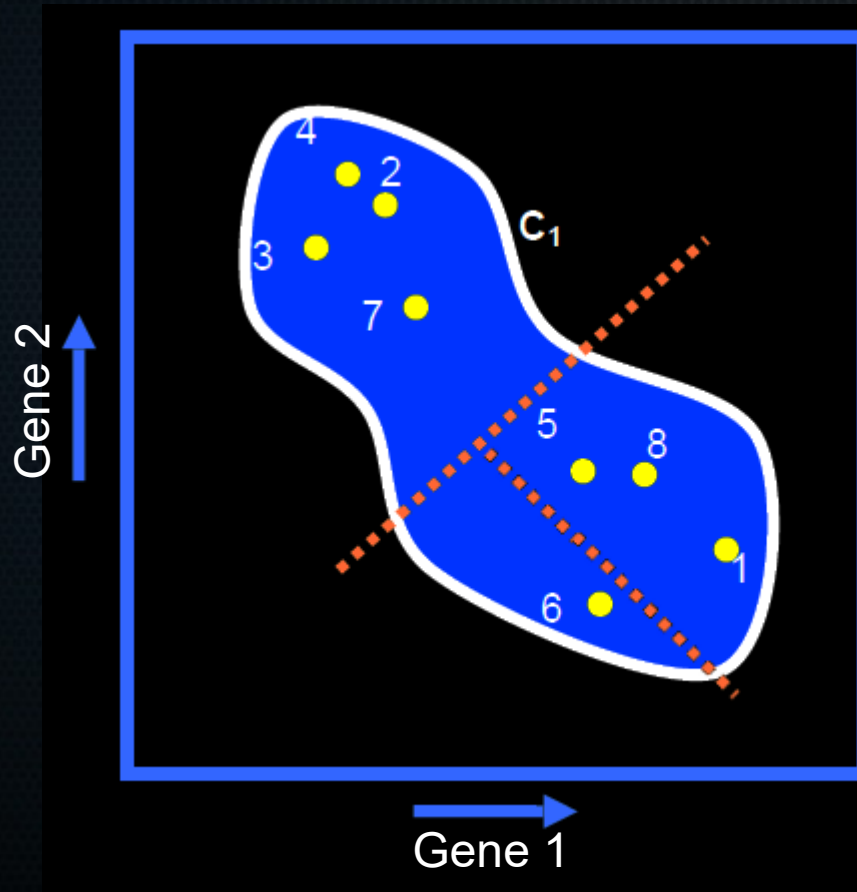
Agglomerative or hierarchical clustering

- Keep iterating until everything is clustered together → Done!



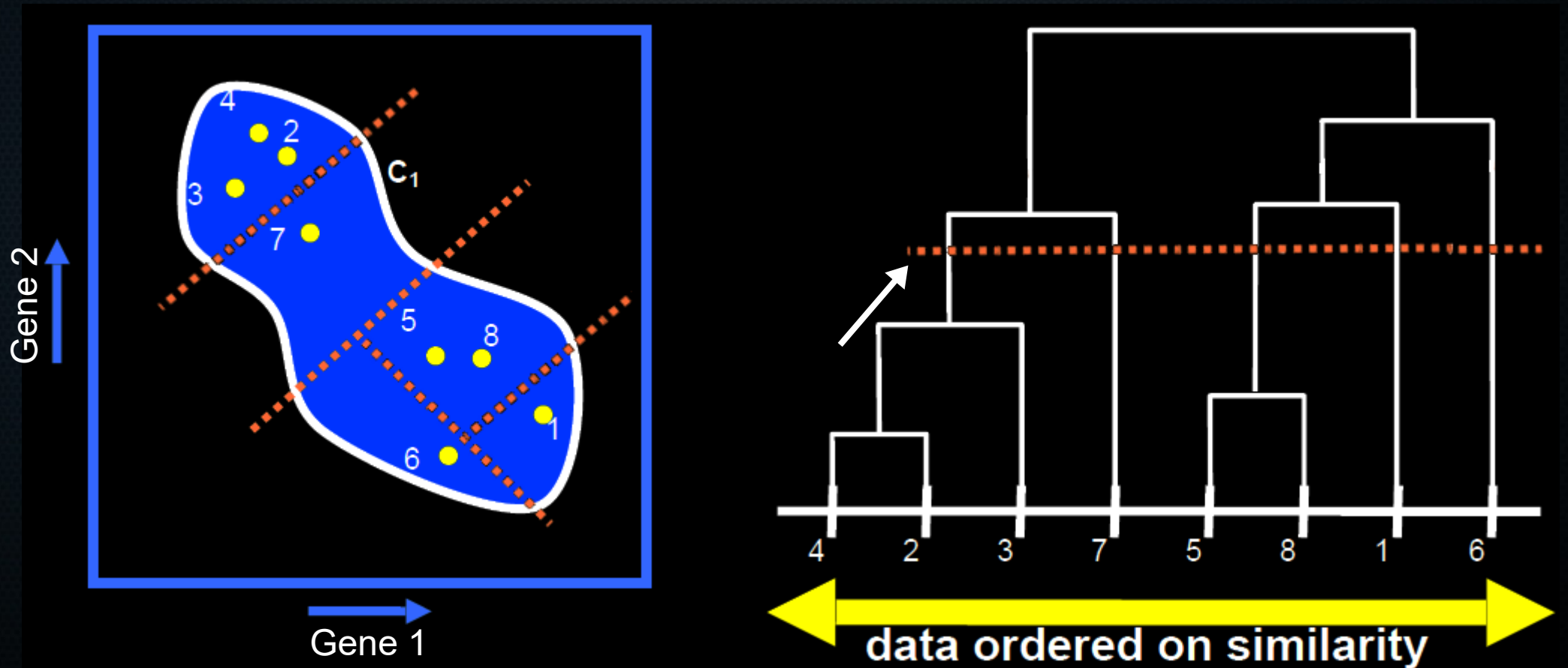
Agglomerative or hierarchical clustering

- Make clusters by cutting the tree at any position



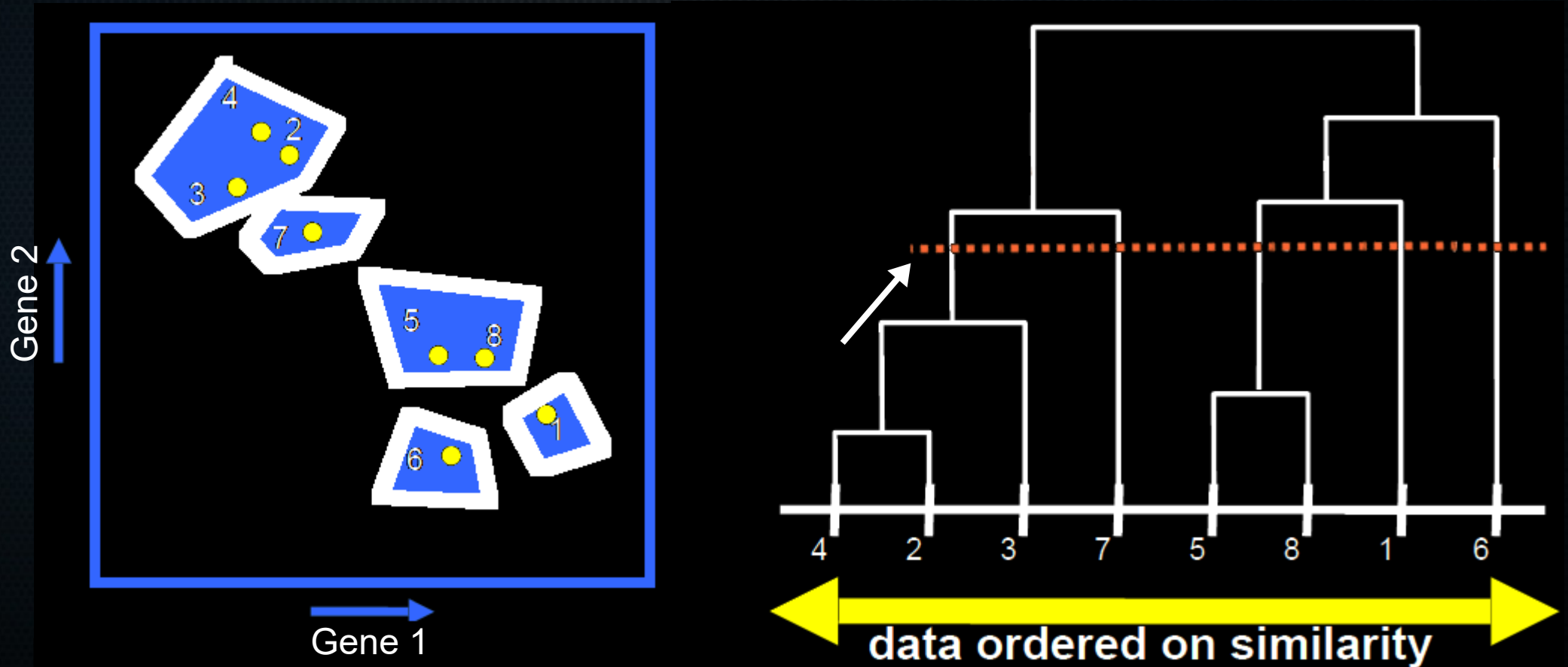
Agglomerative or hierarchical clustering

- Make clusters by cutting the tree at any position



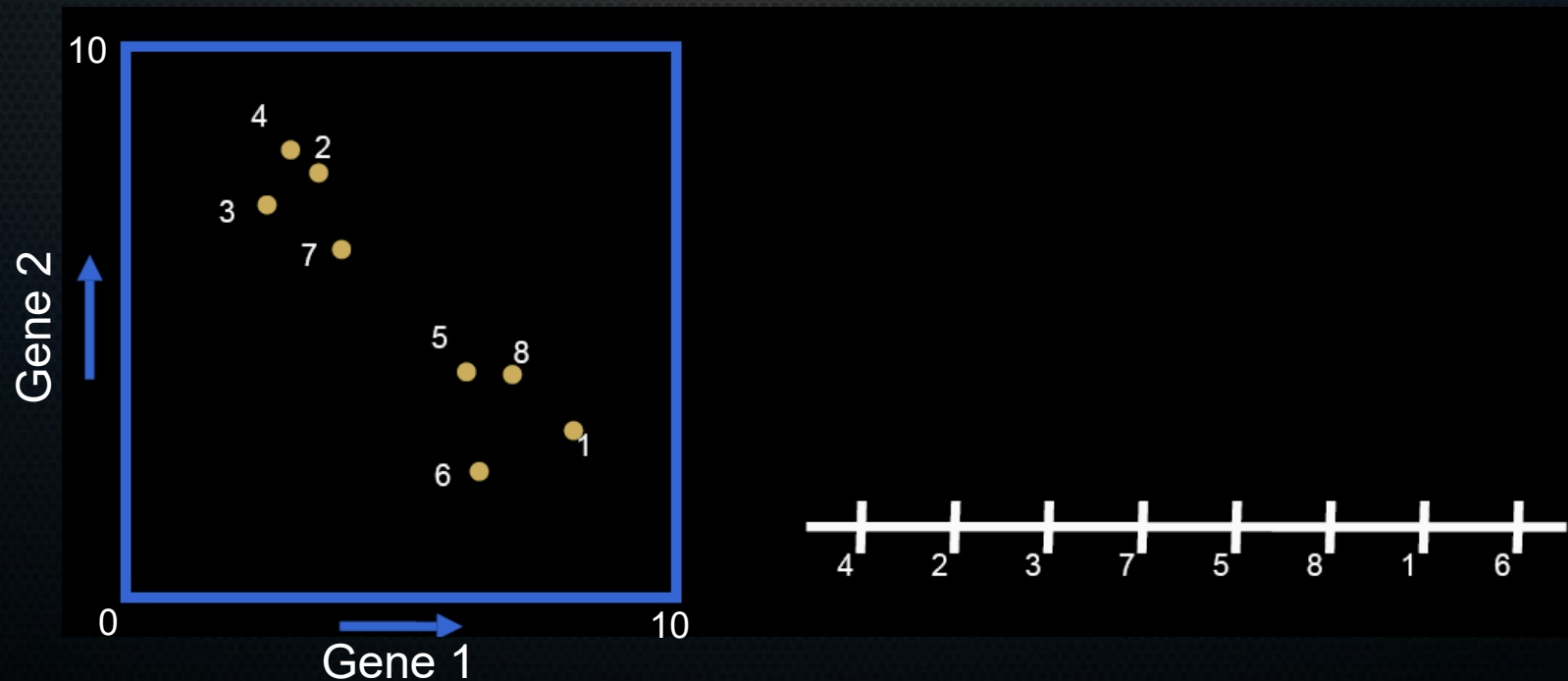
Agglomerative or hierarchical clustering

- Make clusters by cutting the tree at any position



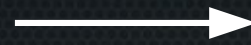
How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8

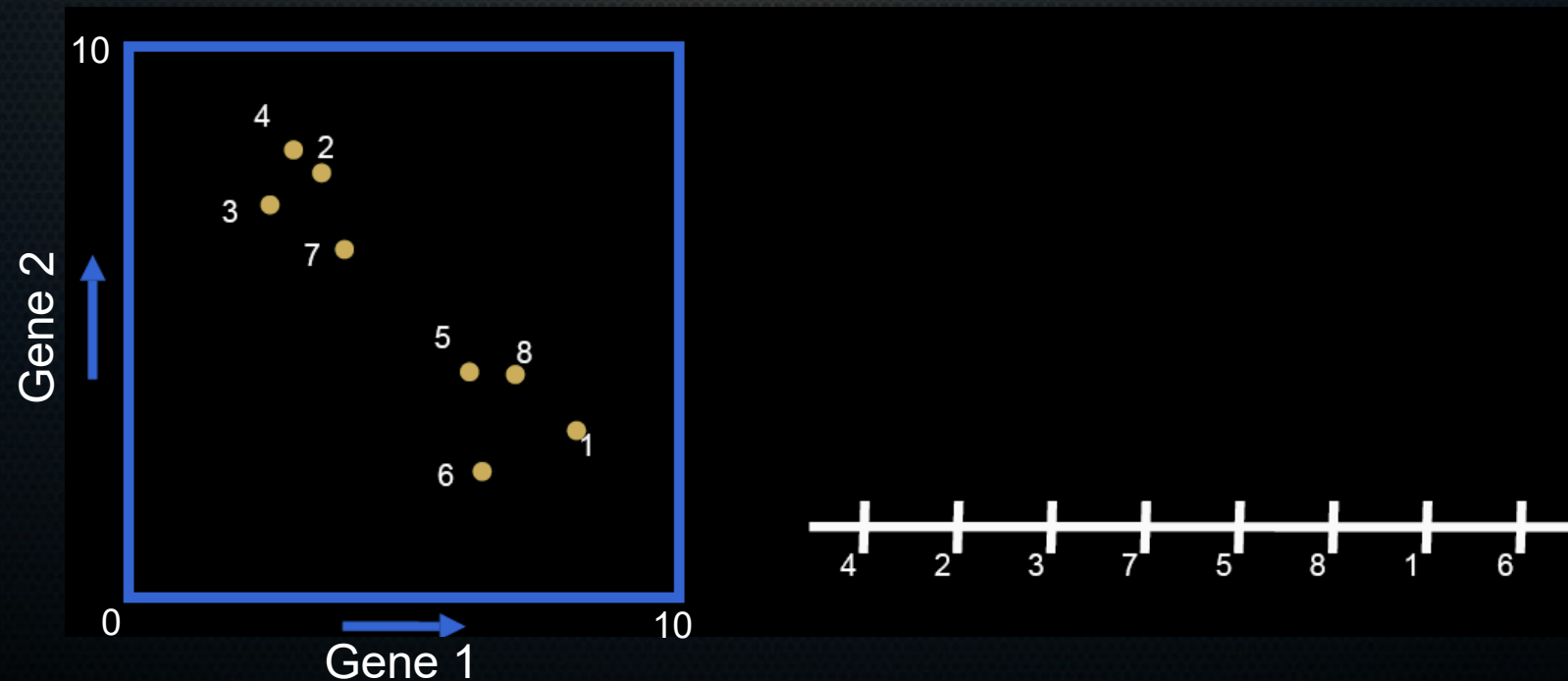


How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



Distance	1	2	3	4
1	0			
2		0		
3			0	
4				0

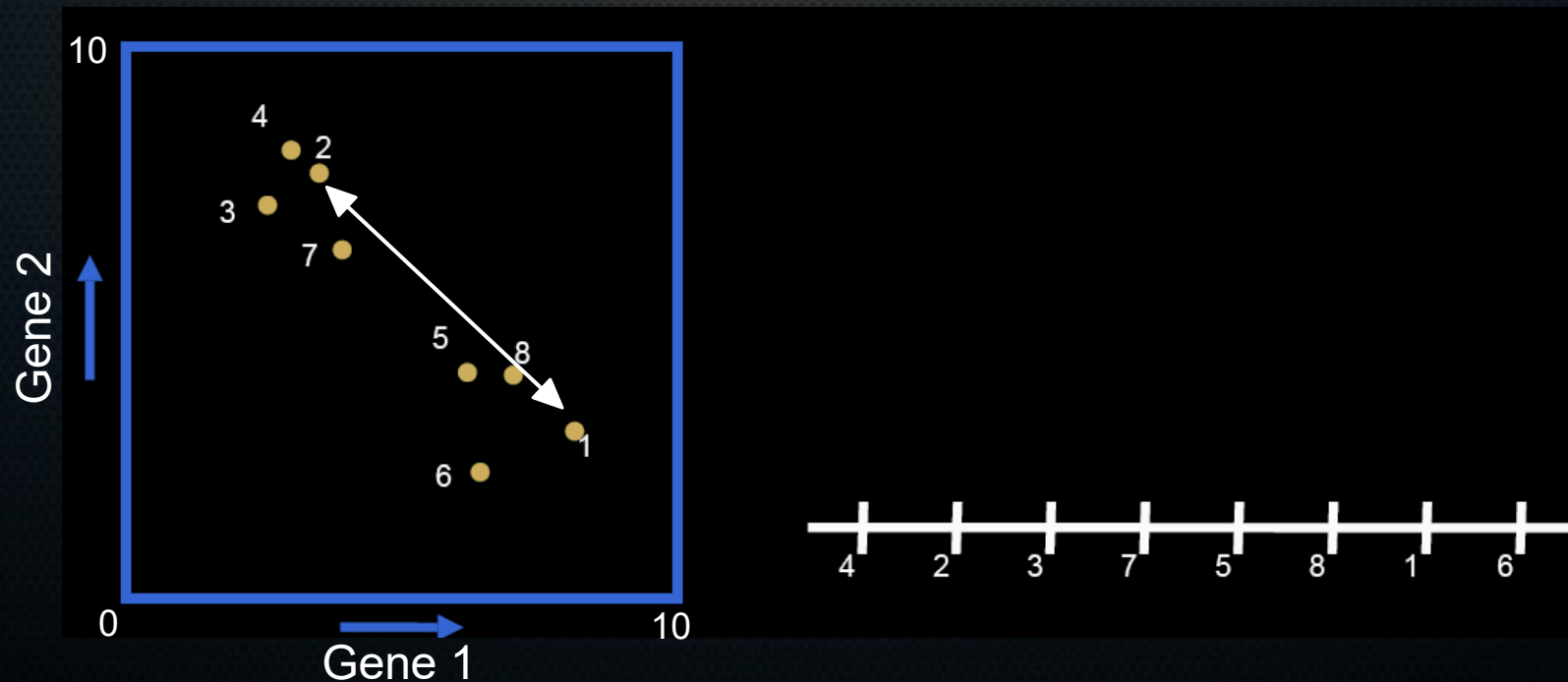


How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



Distance	1	2	3	4
1	0			
2		0		
3			0	
4				0



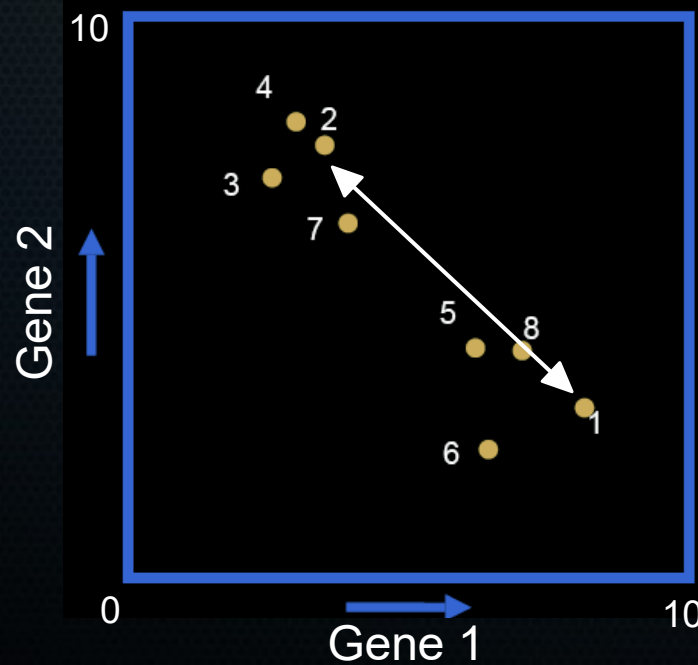
How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



Distance	1	2	3	4
1	0			
2		0		
3			0	
4				0

$$\text{distance} = \sqrt{((\text{Gene1}_1 - \text{Gene1}_2)^2 + (\text{Gene2}_1 - \text{Gene2}_2)^2)}$$



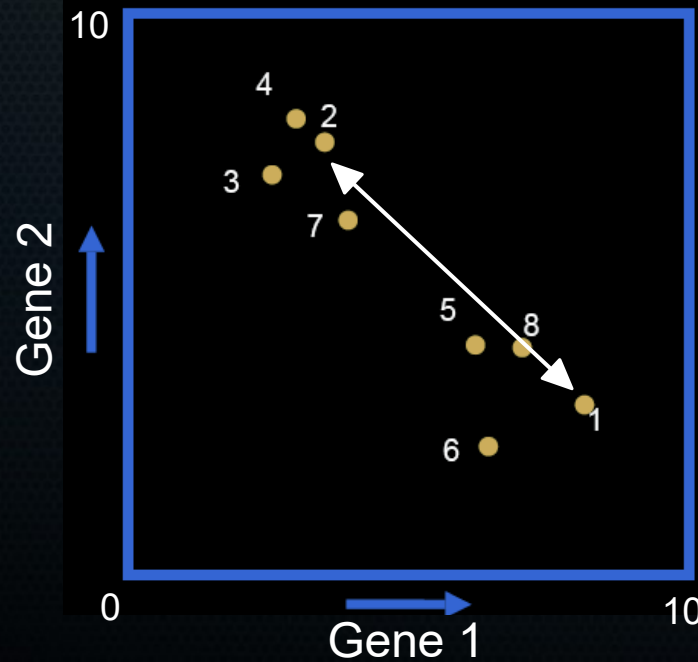
How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



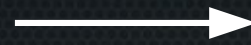
Distance	1	2	3	4
1	0			
2		0		
3			0	
4				0

$$\text{distance} = \sqrt{((8 - 3.2)^2 + (3.5 - 7.6)^2)}$$



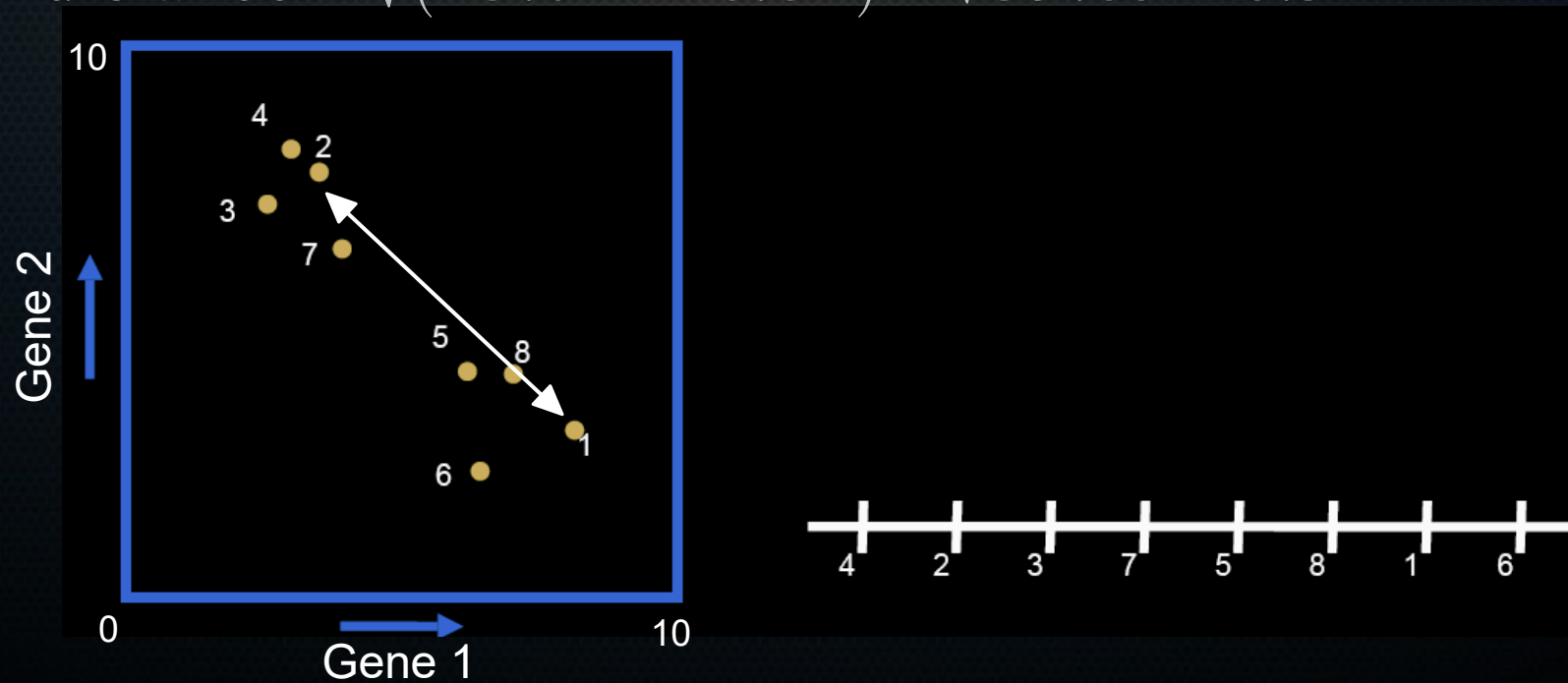
How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



Distance	1	2	3	4
1	0	6.31		
2	6.31	0		
3			0	
4				0

$$\text{distance} = \sqrt{(23.04 + 16.81)} = \sqrt{39.85} \approx 6.31$$

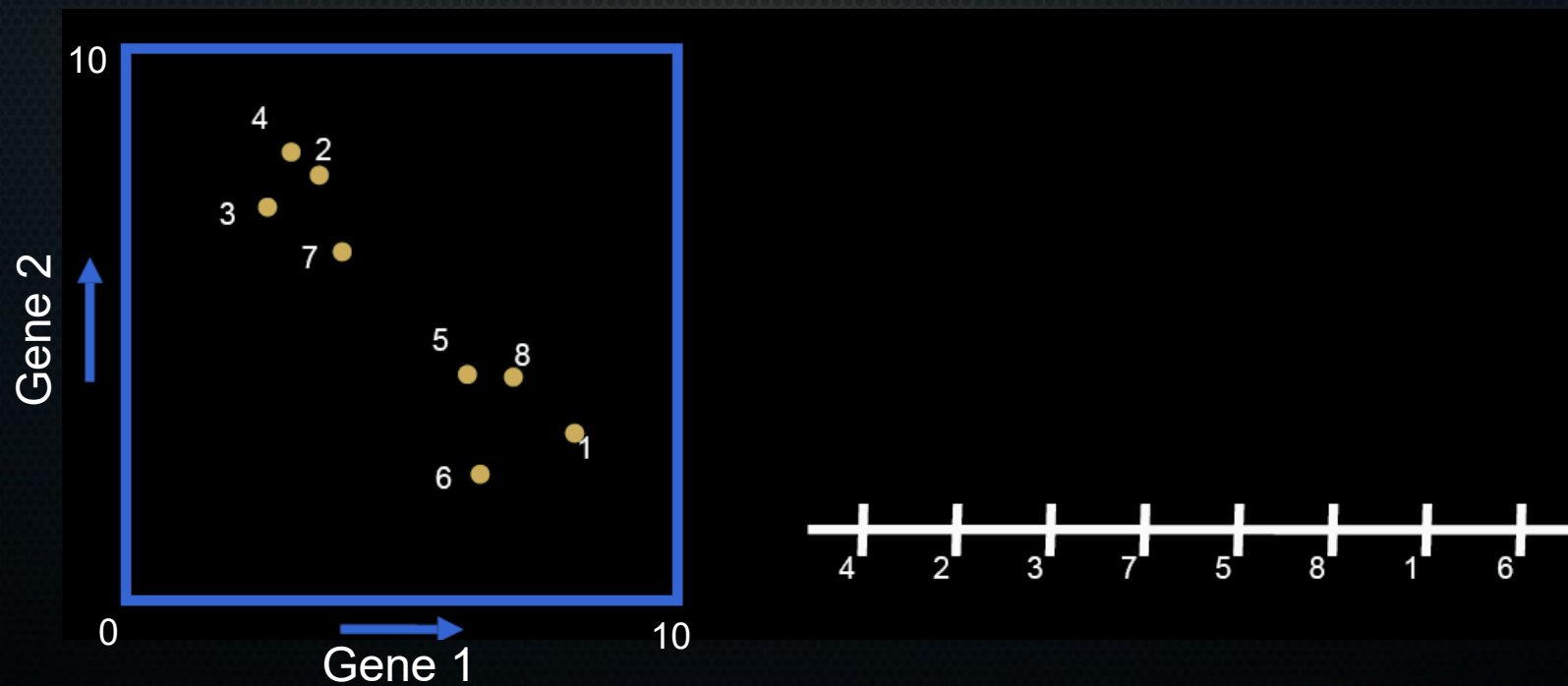


How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



Distance	1	2	3	4
1	0	-	-	-
2	6.31	0	-	-
3	6.52	0.92	0	-
4	6.67	0.36	0.89	0

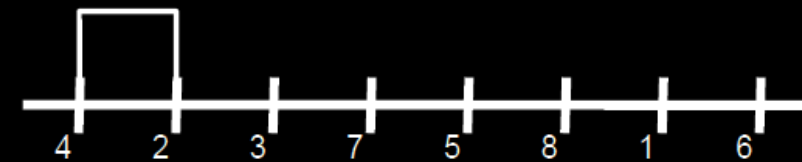
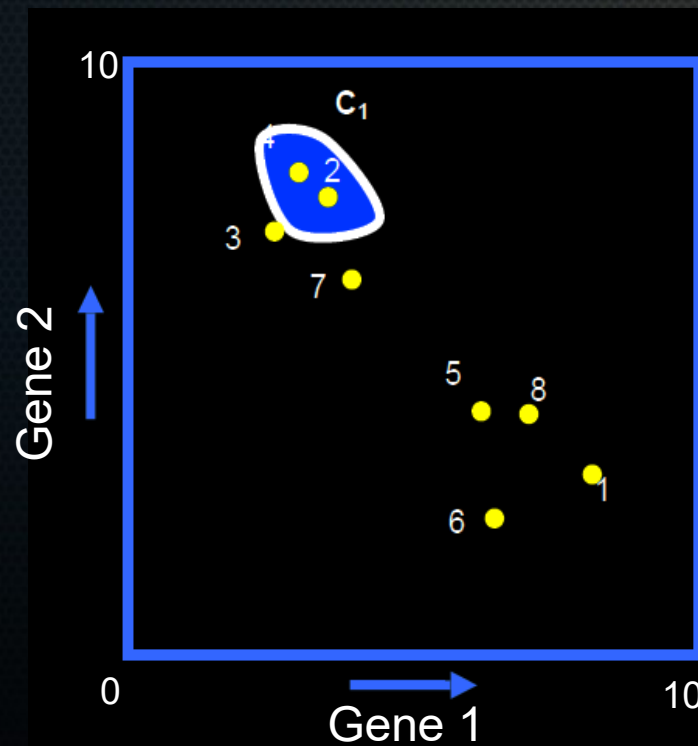


How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



Distance	1	2	3	4
1	0	-	-	-
2	6.31	0	-	-
3	6.52	0.92	0	-
4	6.67	0.36	0.89	0

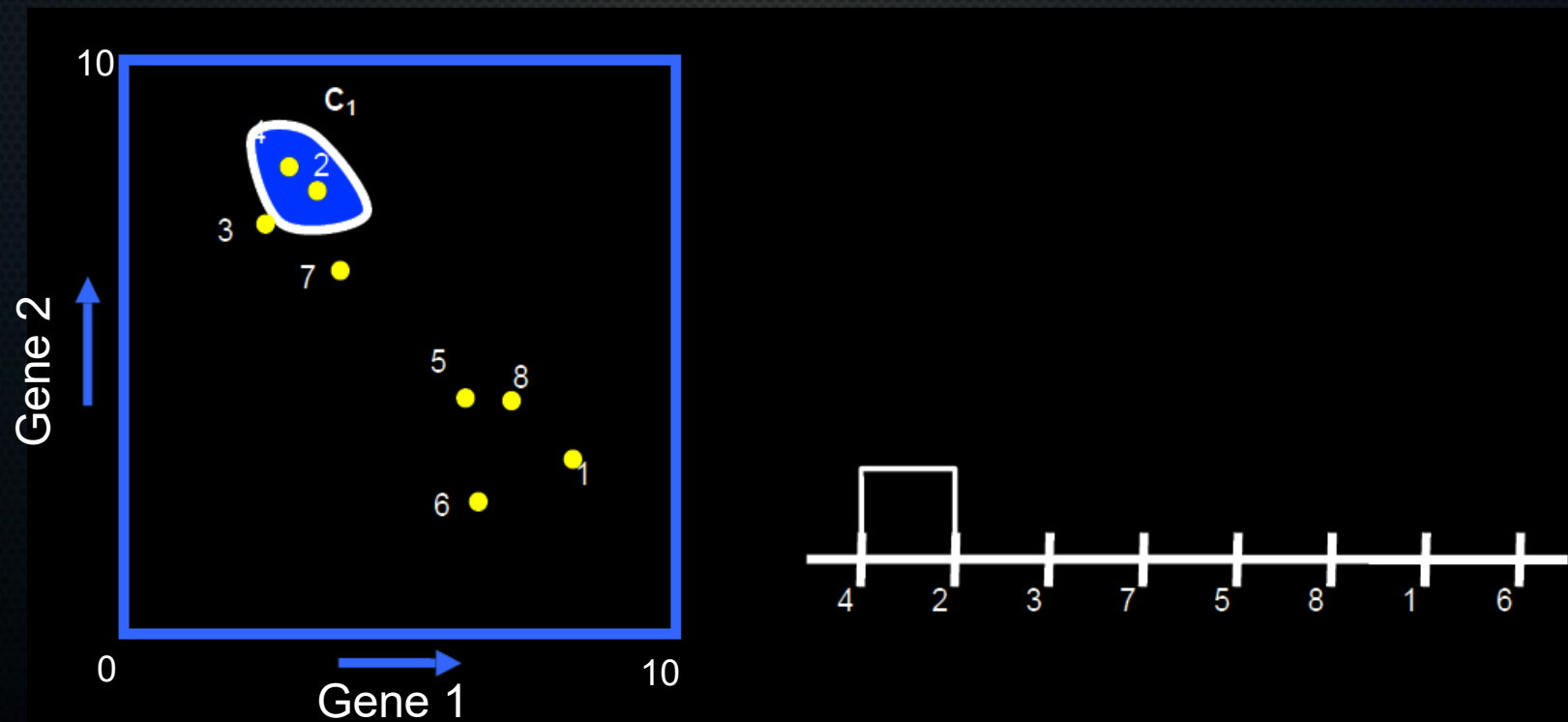


How do we define cluster similarity?

Cluster	Gene 1	Gene 2
1	8	3.5
2+4	?	?
3	2.5	7



Distance	1	2+4	3
1	0	-	-
2+4	?	0	-
3	6.52	?	0



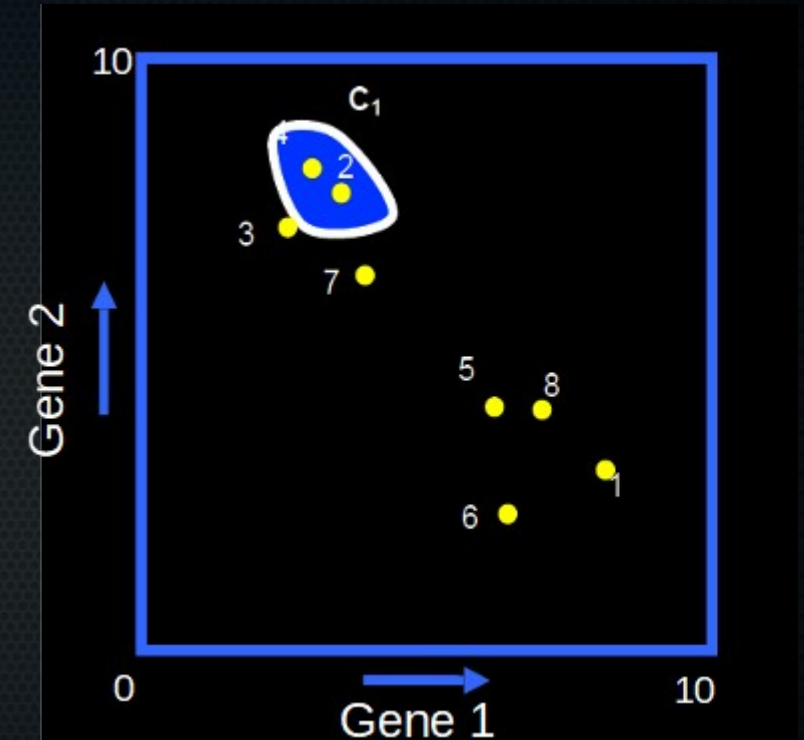
How do we define cluster similarity?

- Three main methods (but there are *MANY*):

How do we define cluster similarity?

- Three main methods:
 - Average linkage (UPGMA)**
Make centroids: distance to a cluster is distance to its mean features

Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8

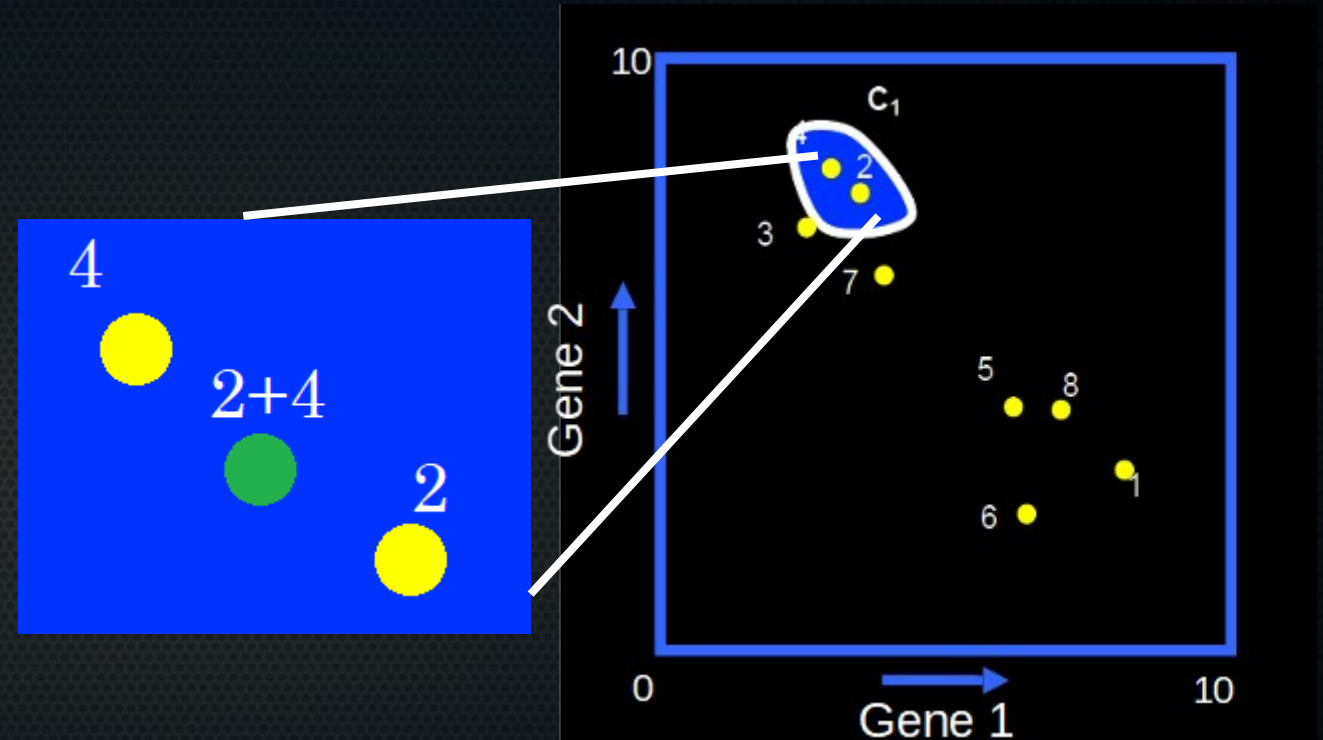


How do we define cluster similarity?

- Three main methods:

- Average linkage (UPGMA)**

Make centroids: distance to a cluster is distance to its mean features



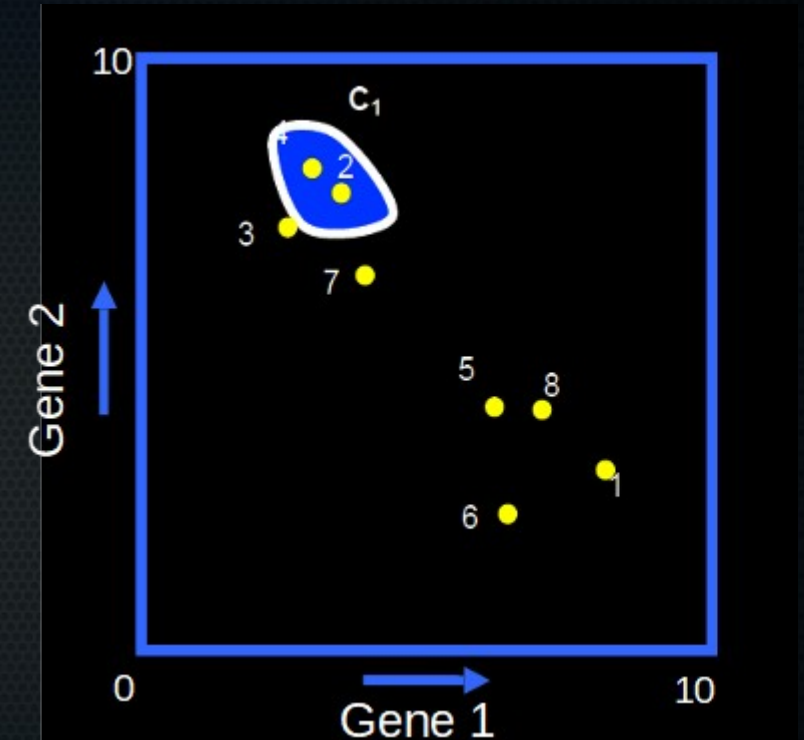
Cluster	Gene 1	Gene 2
1	8	3.5
2	3.2	7.6
3	2.5	7
4	2.9	7.8



Cluster	Gene 1	Gene 2
1	8	3.5
2+4	3.05	7.7
3	2.5	7

How do we define cluster similarity?

- Three main methods:
 - Average linkage (UPGMA)
 - **Single linkage**
Distance between two clusters =
distance between their closest component
points



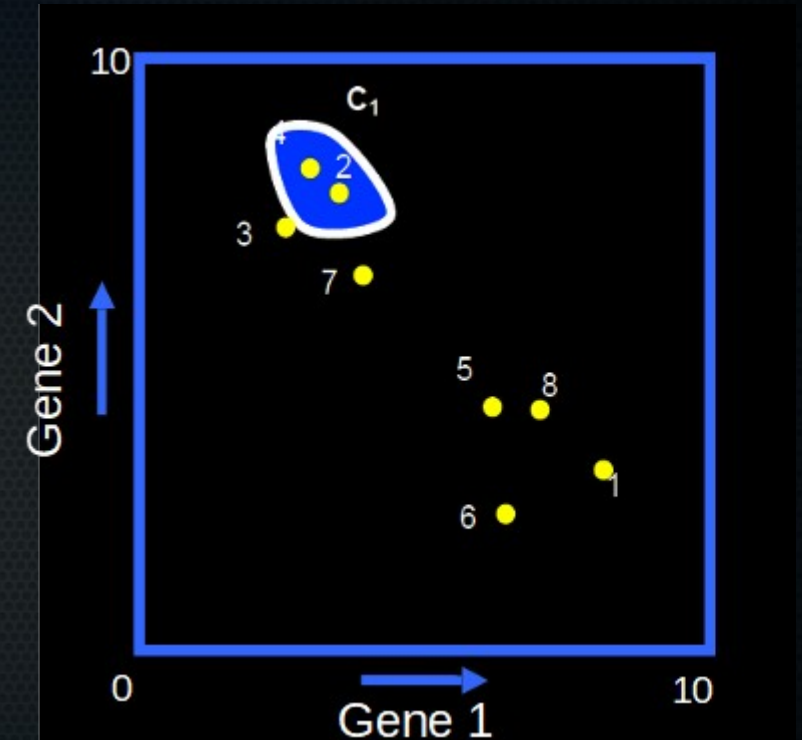
How do we define cluster similarity?

- Three main methods:

- Average linkage (UPGMA)

- **Single linkage**

Distance between two clusters =
distance between their closest component
points



Distance between 3 and 2+4?

Distance	1	2	3	4
1	0	-	-	-
2	6.31	0	-	-
3	6.52	0.92	0	-
4	6.67	0.36	0.89	0

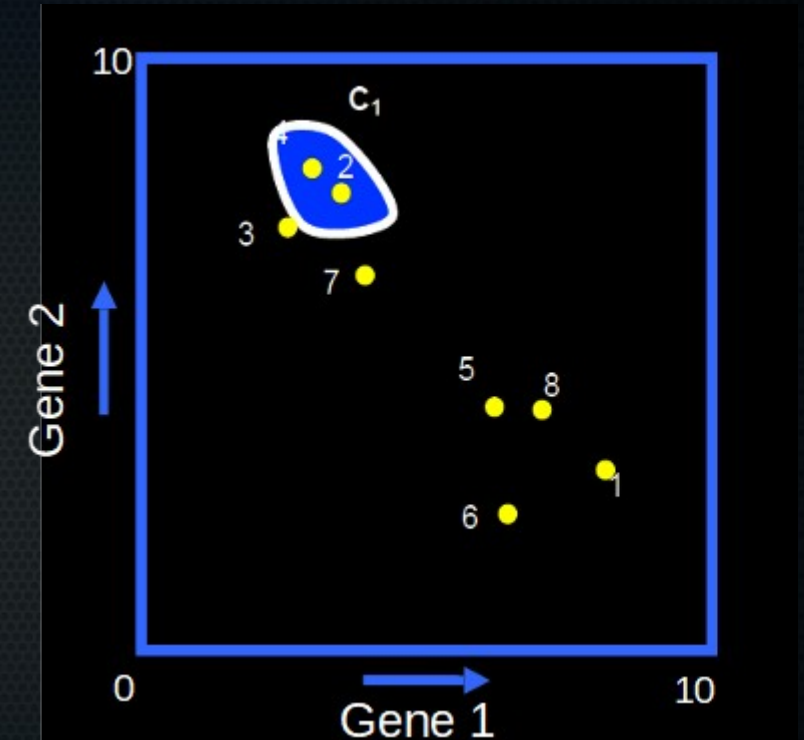
How do we define cluster similarity?

- Three main methods:

- Average linkage (UPGMA)

- **Single linkage**

Distance between two clusters =
distance between their closest component
points



Distance between 3 and 2+4?

Distance	1	2	3	4
1	0	-	-	-
2	6.31	0	-	-
3	6.52	0.92	0	-
4	6.67	0.36	0.89	0

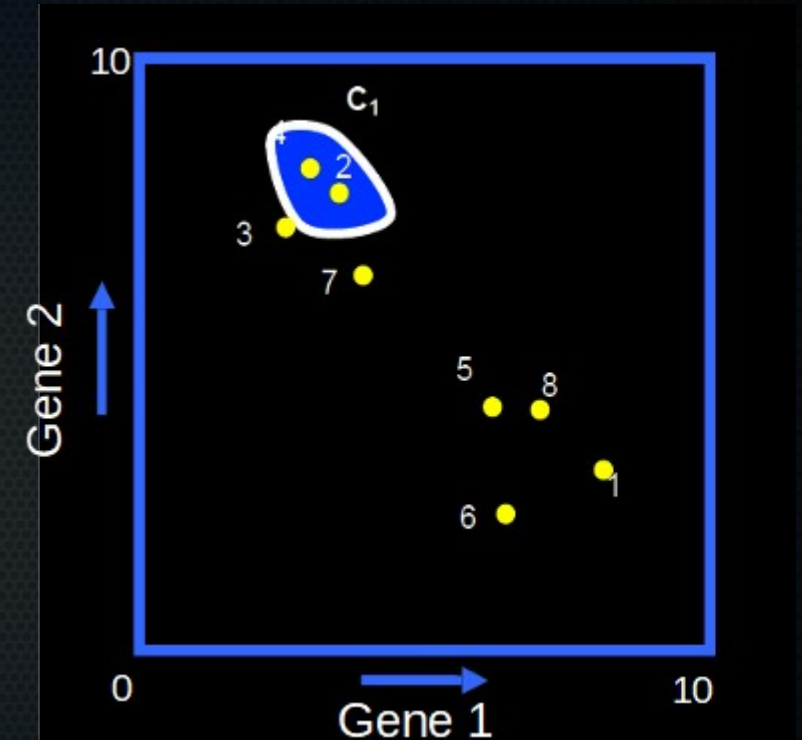
How do we define cluster similarity?

- Three main methods:

- Average linkage (UPGMA)
- Single linkage

- **Complete linkage**

Distance between two clusters =
distance between their furthest component
points

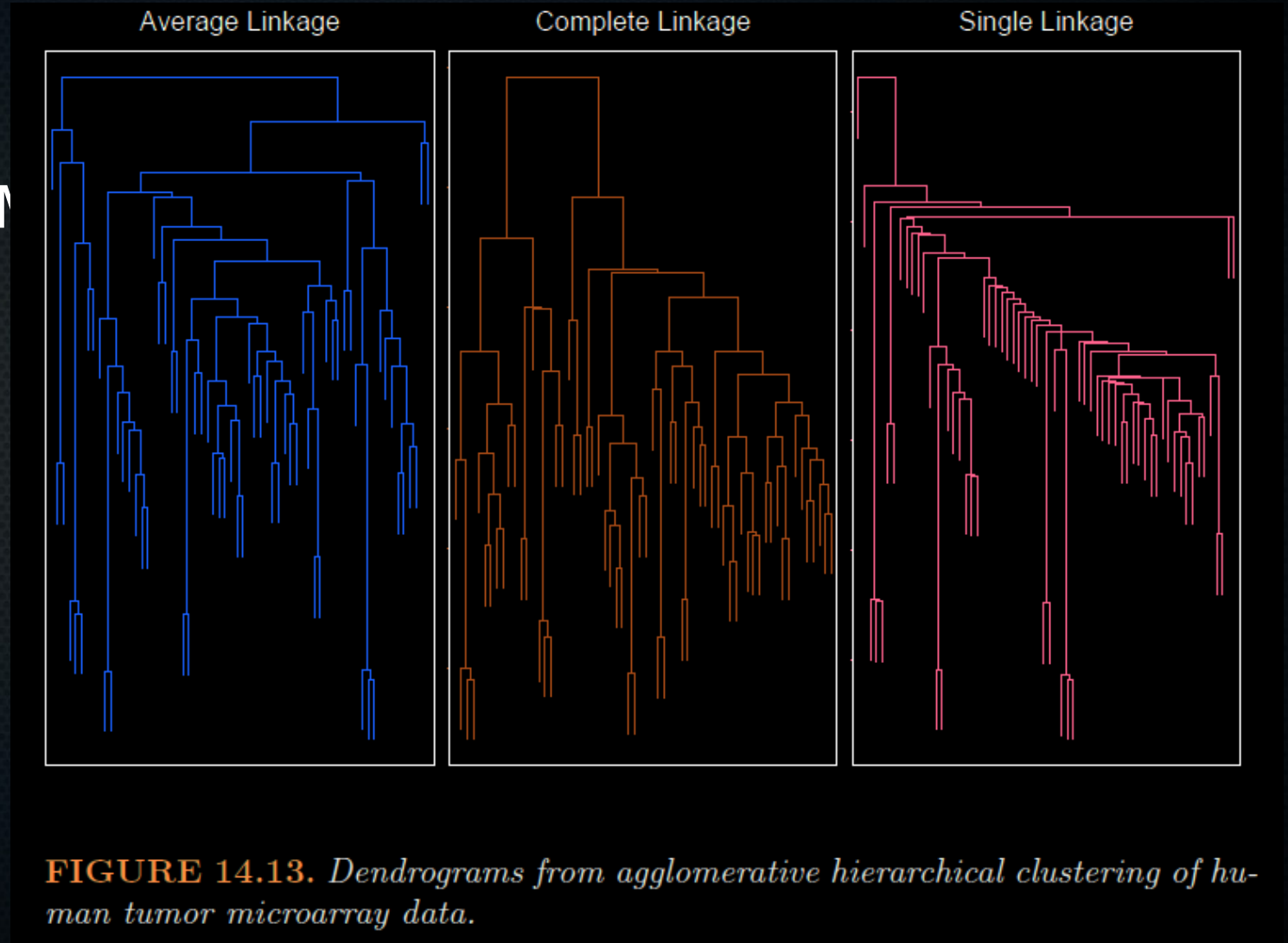


Distance between 3 and 2+4?

Distance	1	2	3	4
1	0	-	-	-
2	6.31	0	-	-
3	6.52	0.92	0	-
4	6.67	0.36	0.89	0

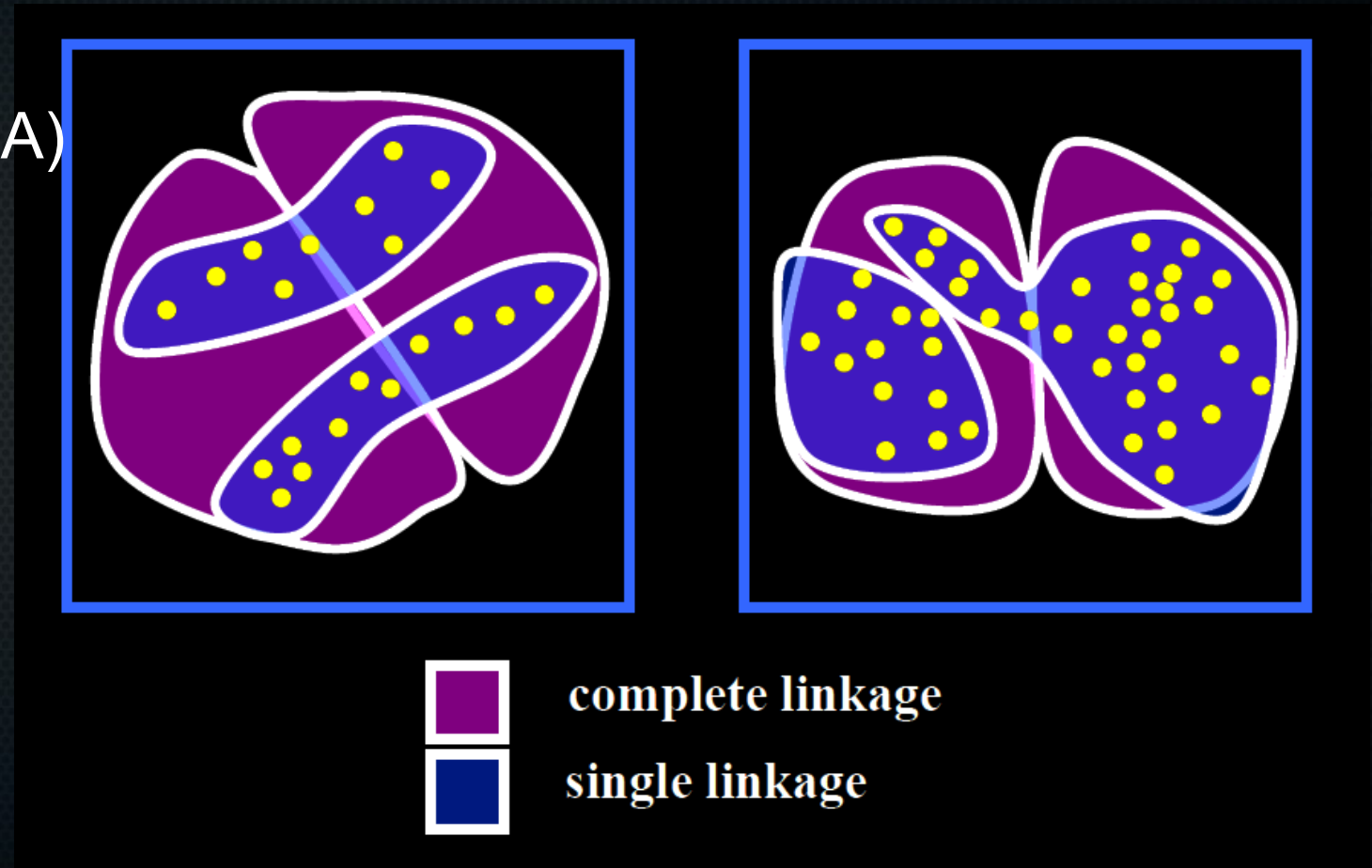
How do we define cluster similarity?

- Three main methods:
 - Average linkage (UPGM)
 - Single linkage
 - Complete linkage
- Quite different results:



How do we define cluster similarity?

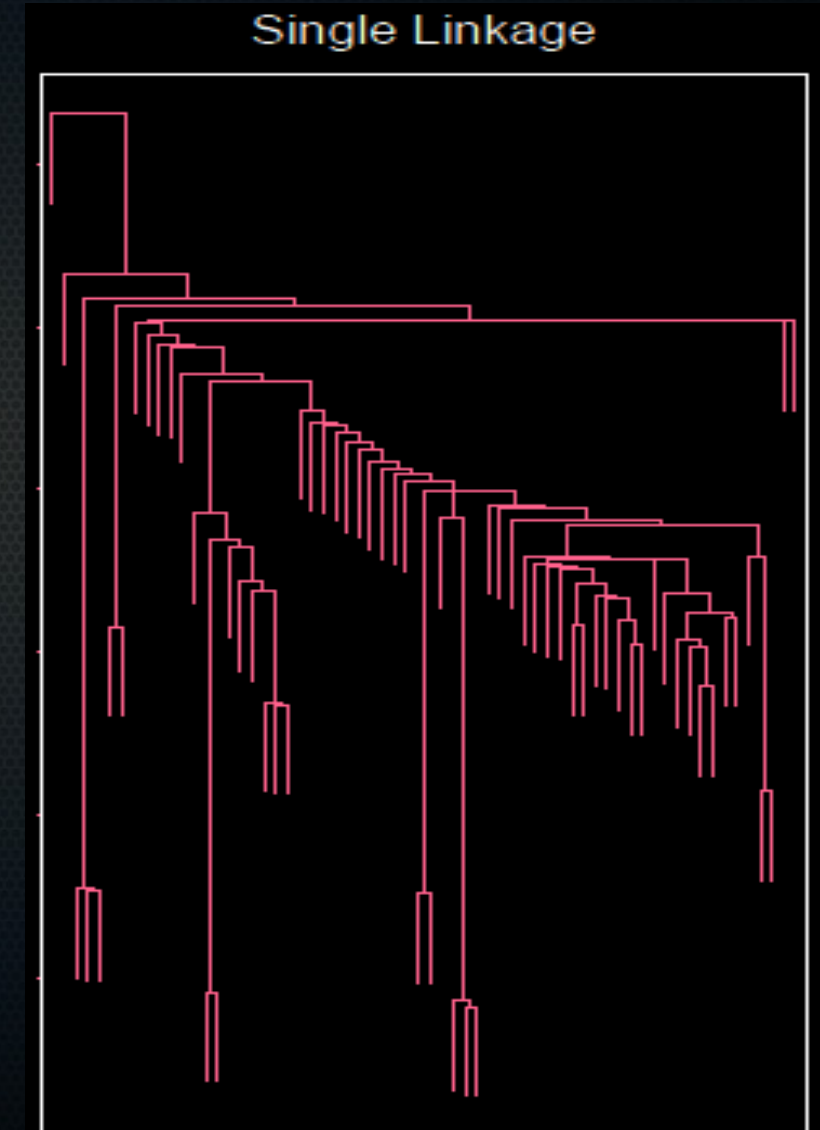
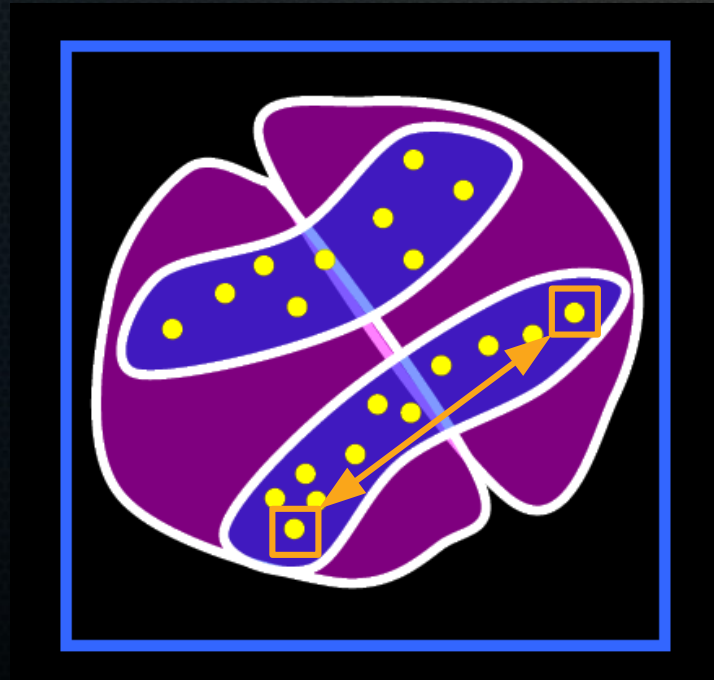
- Three main methods:
 - Average linkage (UPGMA)
 - Single linkage
 - Complete linkage
- Quite different results:



Source: Jeroen de Ridder

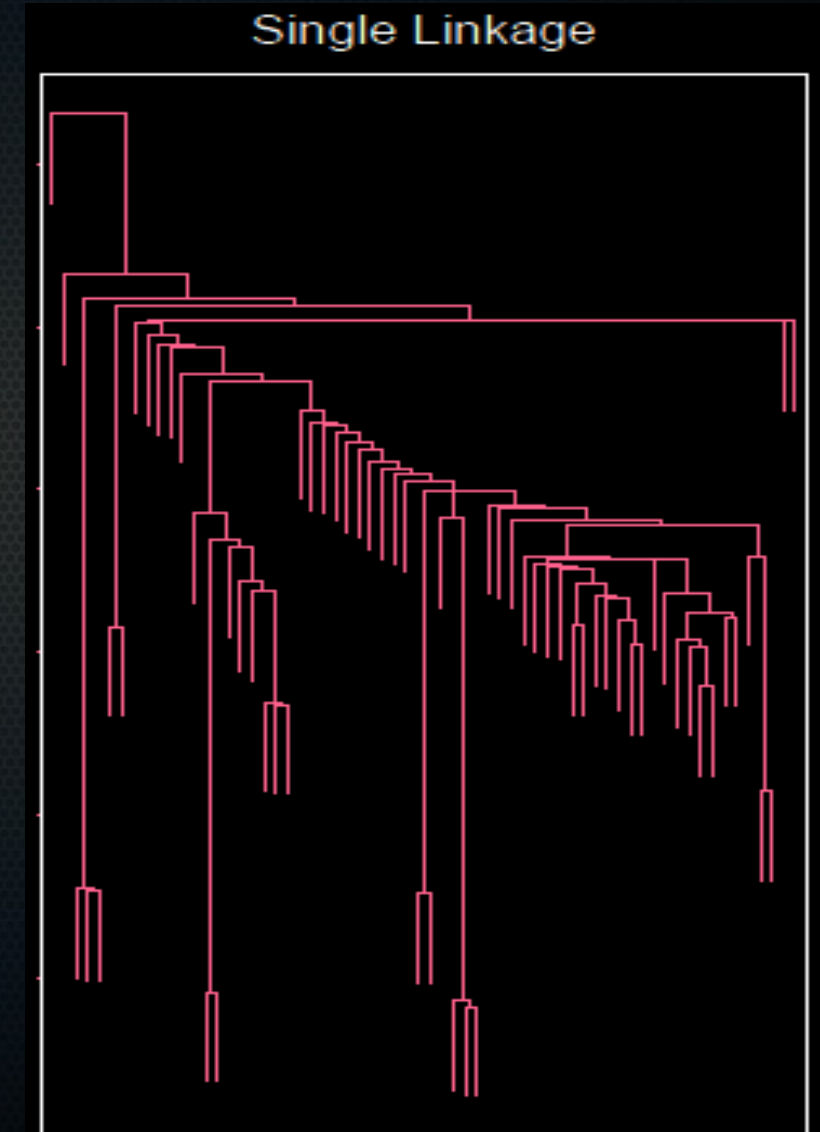
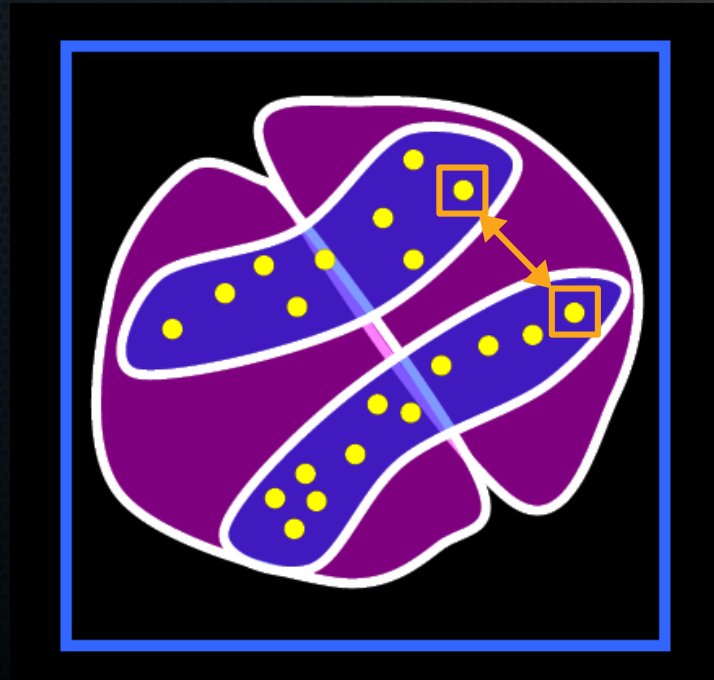
Which linkage method to pick?

- Single linkage drawback:
 - Chaining → combines observations that are very far away by many small intermediate steps



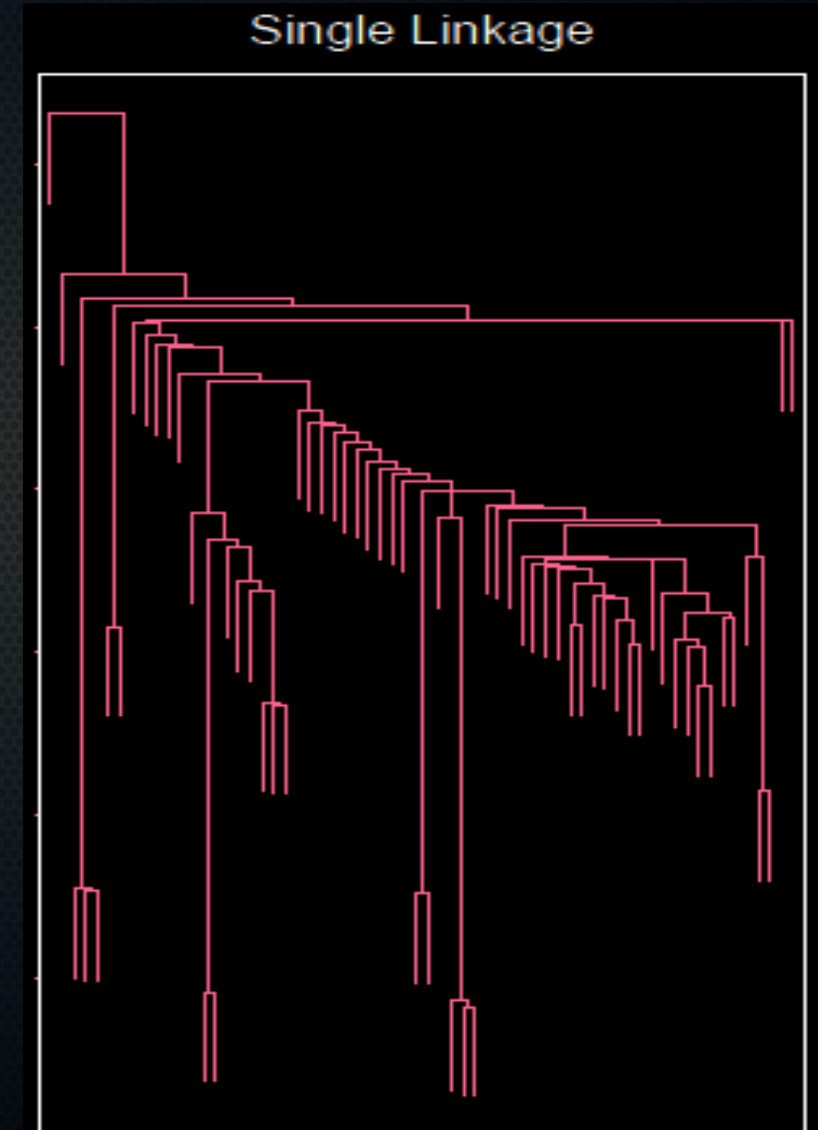
Which linkage method to pick?

- Single linkage drawback:
 - Chaining → combines observations that are very far away by many small intermediate steps



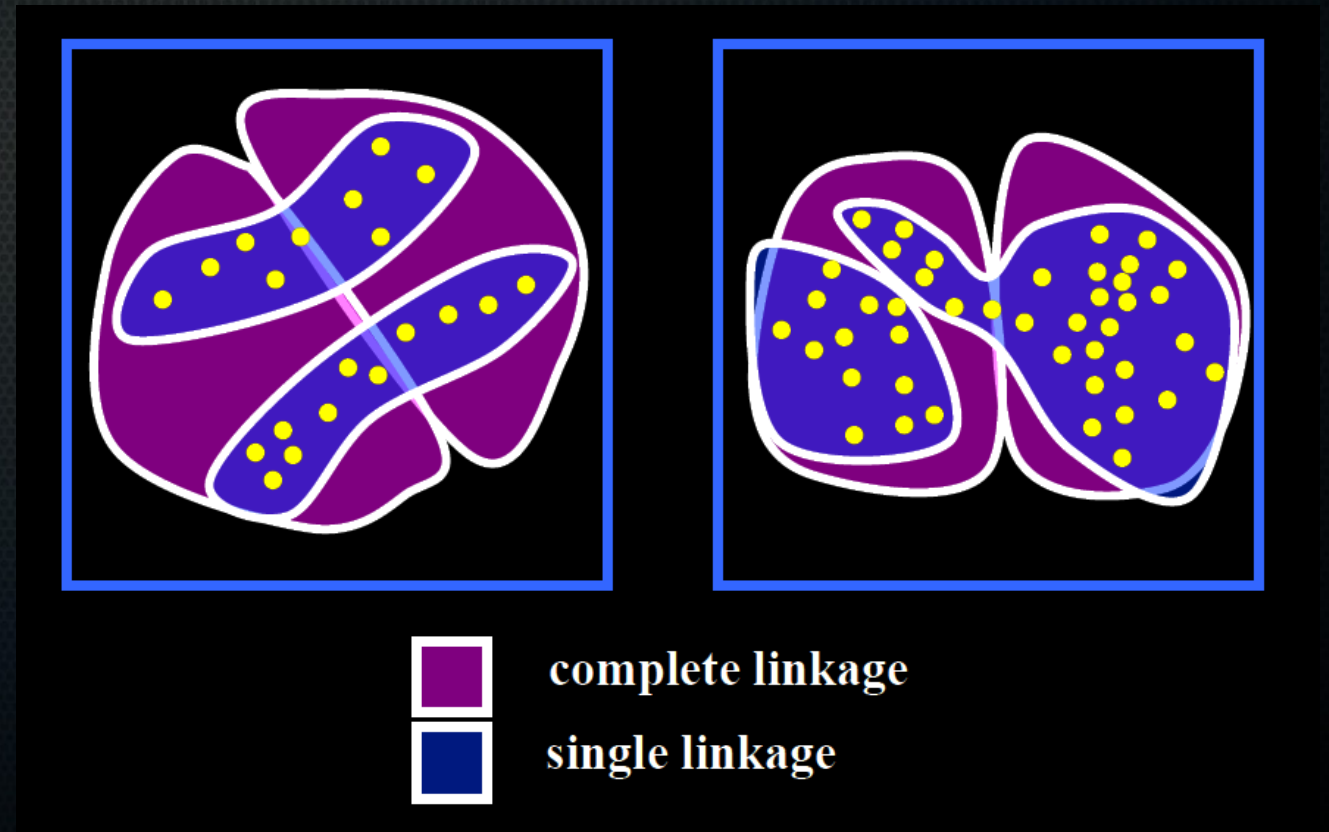
Which linkage method to pick?

- Single linkage drawback:
 - Chaining → combines observations that are very far away by many small intermediate steps
 - Doesn't lead to compact clusters



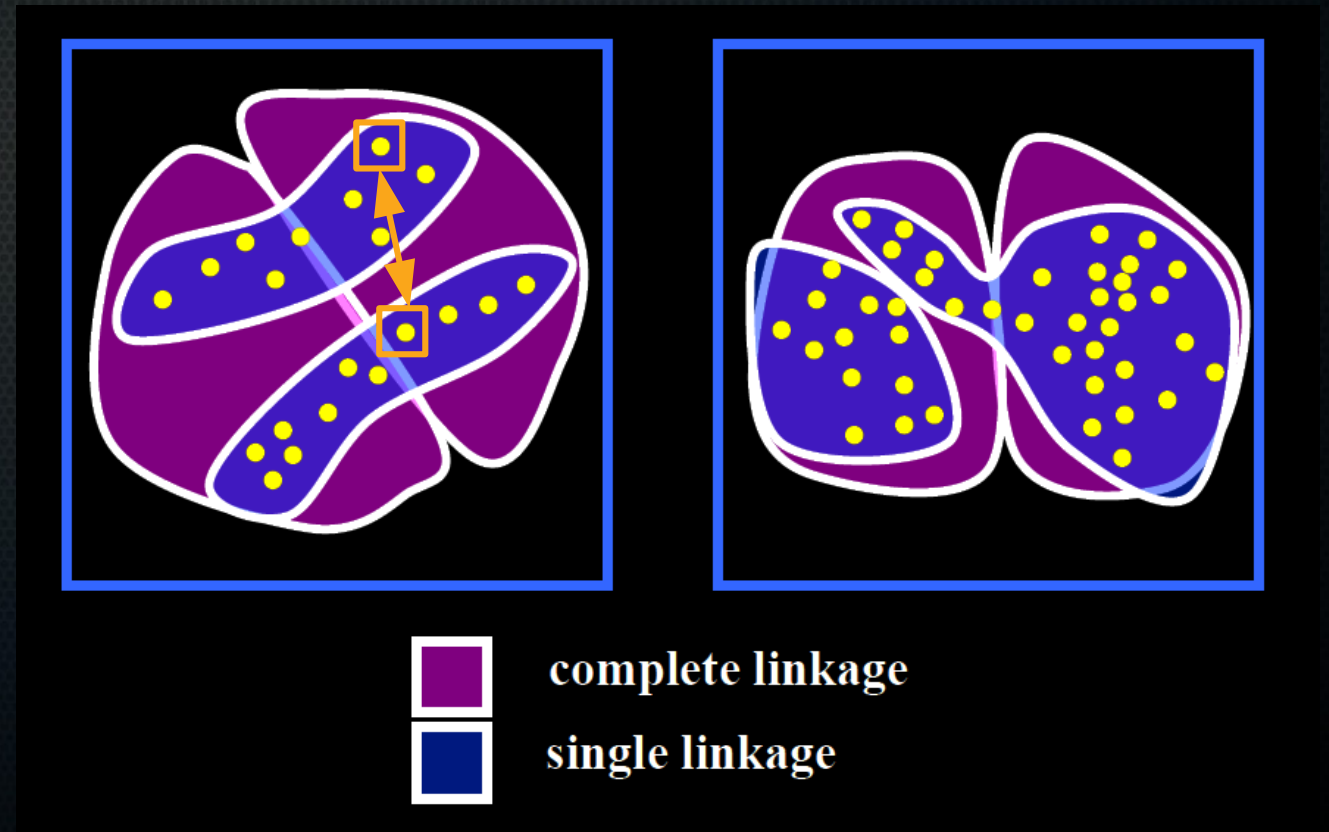
Which linkage method to pick?

- Complete linkage drawback:
 - Opposite problem: close only if close to all members of a group.



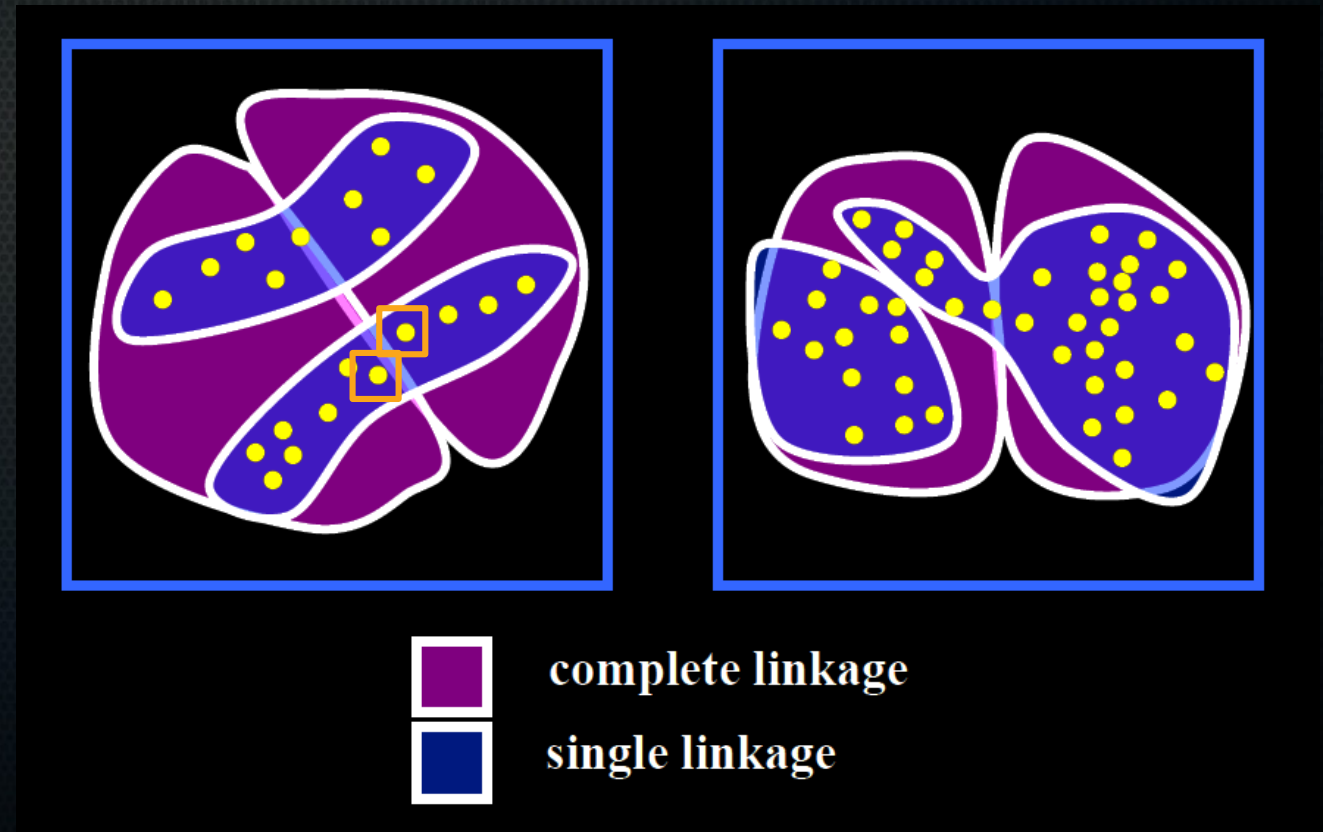
Which linkage method to pick?

- Complete linkage drawback:
 - Opposite problem: close only if close to all members of a group, so *samples* that are close can get assigned to very different clusters



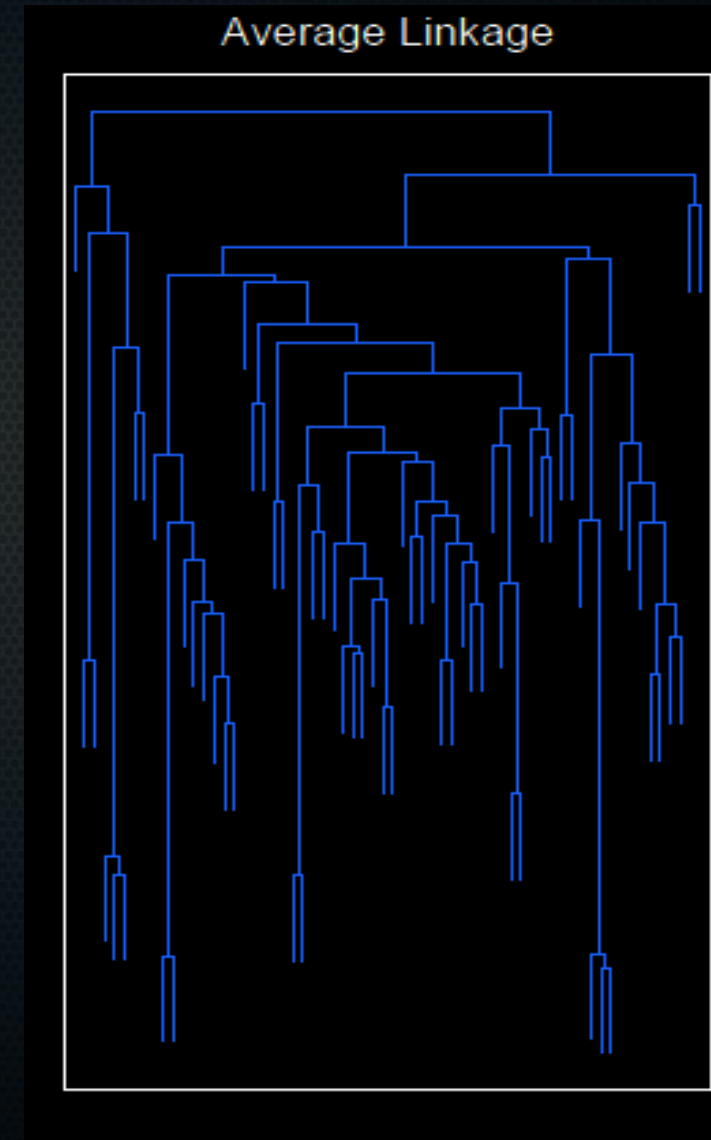
Which linkage method to pick?

- Complete linkage drawback:
 - Opposite problem: close only if close to all members of a group, so *samples* that are close can get assigned to very different clusters



Which linkage method to pick?

- Average linkage (UPGMA):
 - Compromise between the two.
 - Does depend on the numerical scale



Summary agglomerative/hierarchical clustering

- Start with each data point in a cluster by itself
- Calculate distances between all clusters
- Join closest clusters together
- Recalculate distances (depending on linkage method!)
- Iterate until all clusters are connected
- Make clusters by ,cutting through' the tree at any level of clustering.

Divisive clustering

- Similar to bisecting K-means, you can also start with every observation grouped together, and then split off parts.
- Example = DIANA:
 - pick the object with the highest average dissimilarity to all other objects in its cluster (most distant from all others)

Divisive clustering

- Similar to bisecting K-means, you can also start with every observation grouped together, and then split off parts.
- Example = DIANA:
 - pick the object with the highest average dissimilarity to all other objects in its cluster (most distant from all others)
 - put it in a new cluster, cluster it together with everything that is closer to it than to the centroid of the original cluster

Divisive clustering

- Similar to bisecting K-means, you can also start with every observation grouped together, and then split off parts.
- Example = DIANA :
 - pick the object with the highest average dissimilarity to all other objects in its cluster (most distant from all others)
 - put it in a new cluster, cluster it together with everything that is closer to it than to the centroid of the original cluster
- Won't go into further detail here.

Break for short practical
