

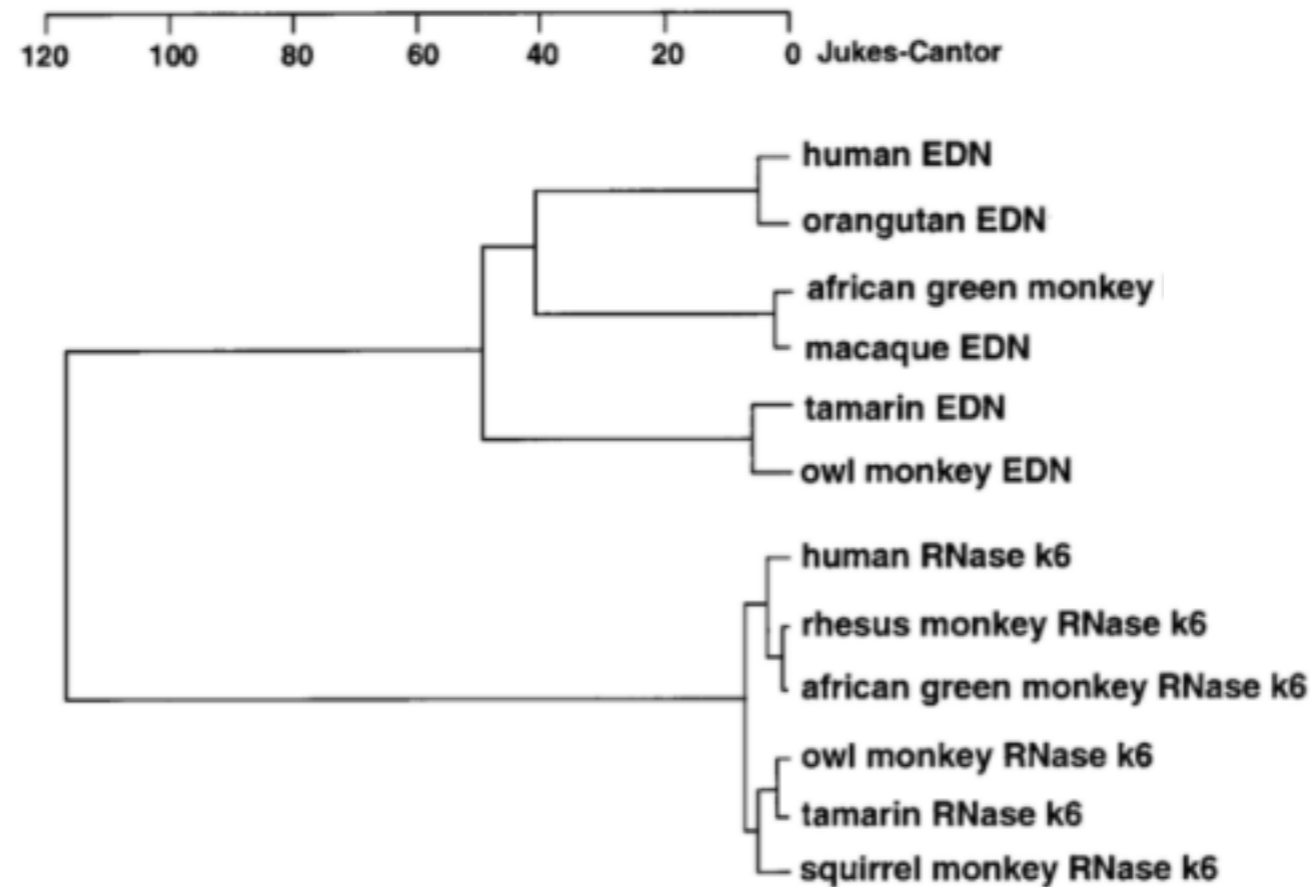
[320] Hierarchical Clustering

(Agglomerative Clustering and Dendrograms)

Non-hierarchical clusters cannot contain other clusters
(example: **KMeans**)

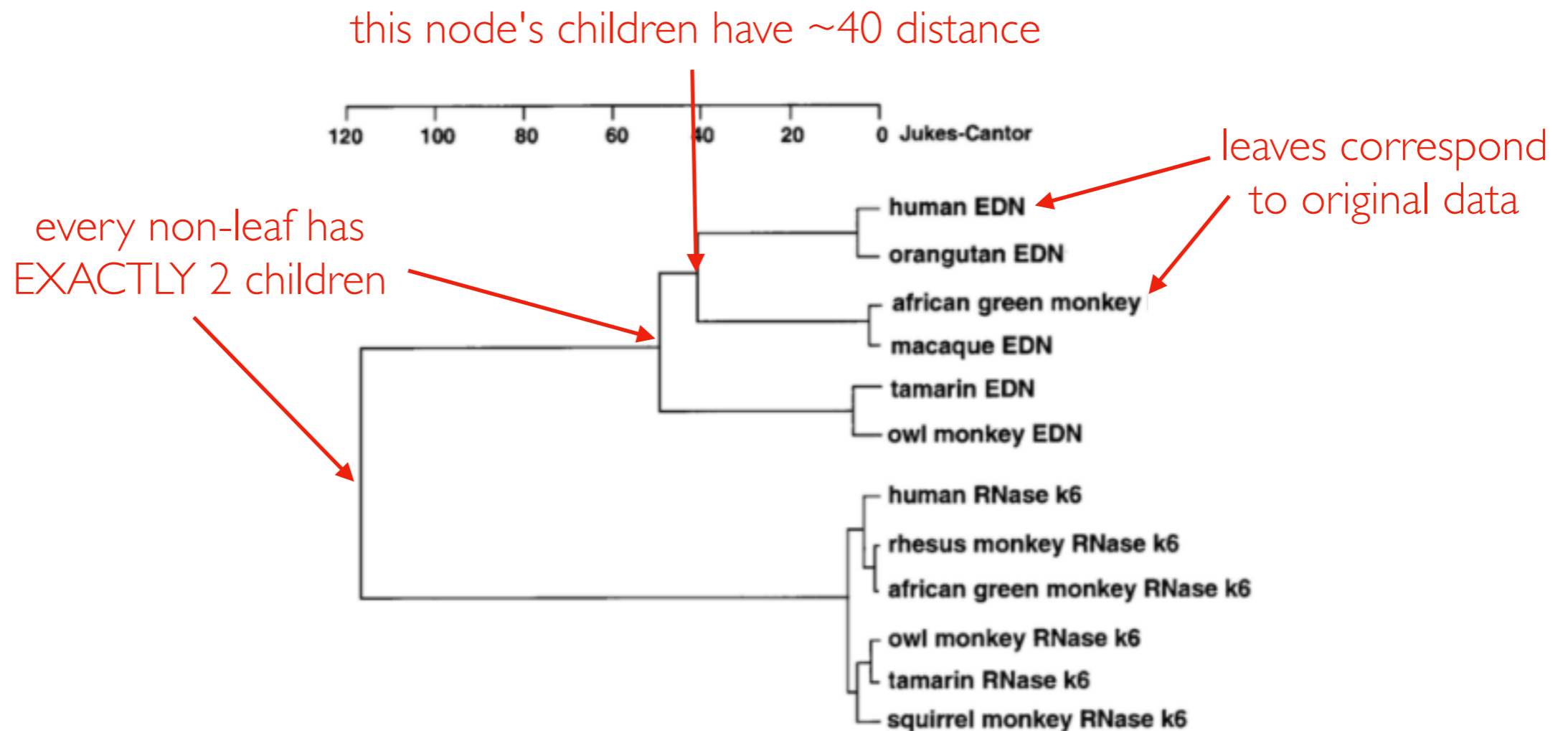
Hierarchical clusters can contain other clusters
(example: **AgglomerativeClustering**)

Hierarchical Clusters with Dendrograms



https://www.researchgate.net/figure/A-Dendrogram-depicting-the-relationships-among-human-and-non-human-primate-EDNs-and_fig1_13459488

Hierarchical Clusters with Dendrograms



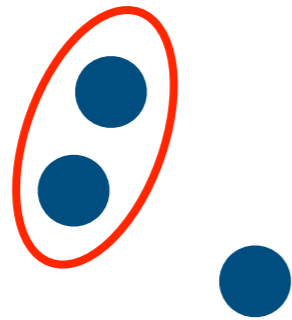
https://www.researchgate.net/figure/A-Dendrogram-depicting-the-relationships-among-human-and-non-human-primate-EDNs-and_fig1_13459488

We'll represent hierarchies as special binary trees.

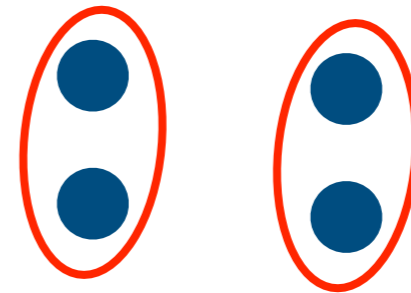
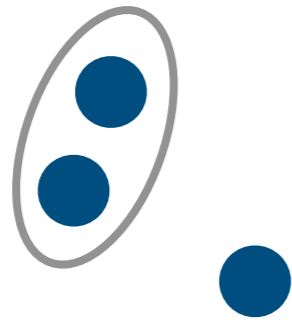
Strategy: Combine Nearby Points/Groups
(and repeat!)



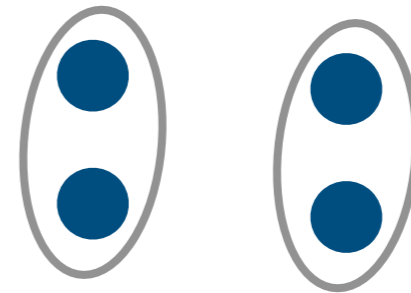
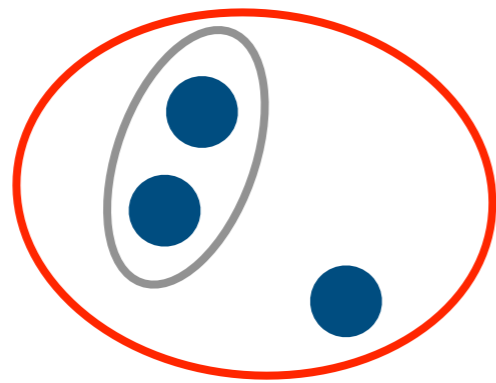
Strategy: Combine Nearby Points/Groups
(and repeat!)



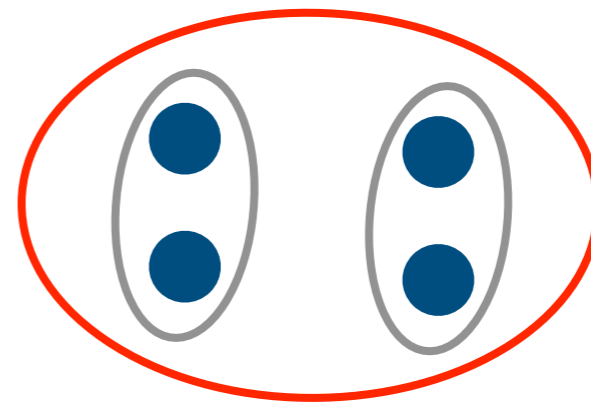
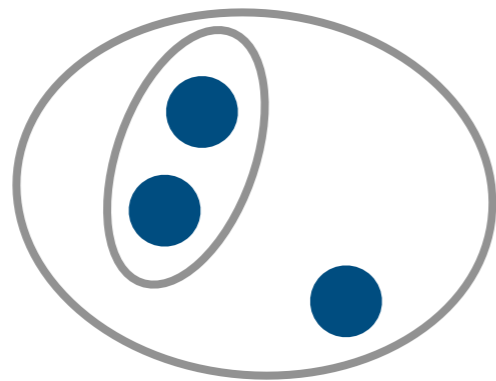
Strategy: Combine Nearby Points/Groups
(and repeat!)



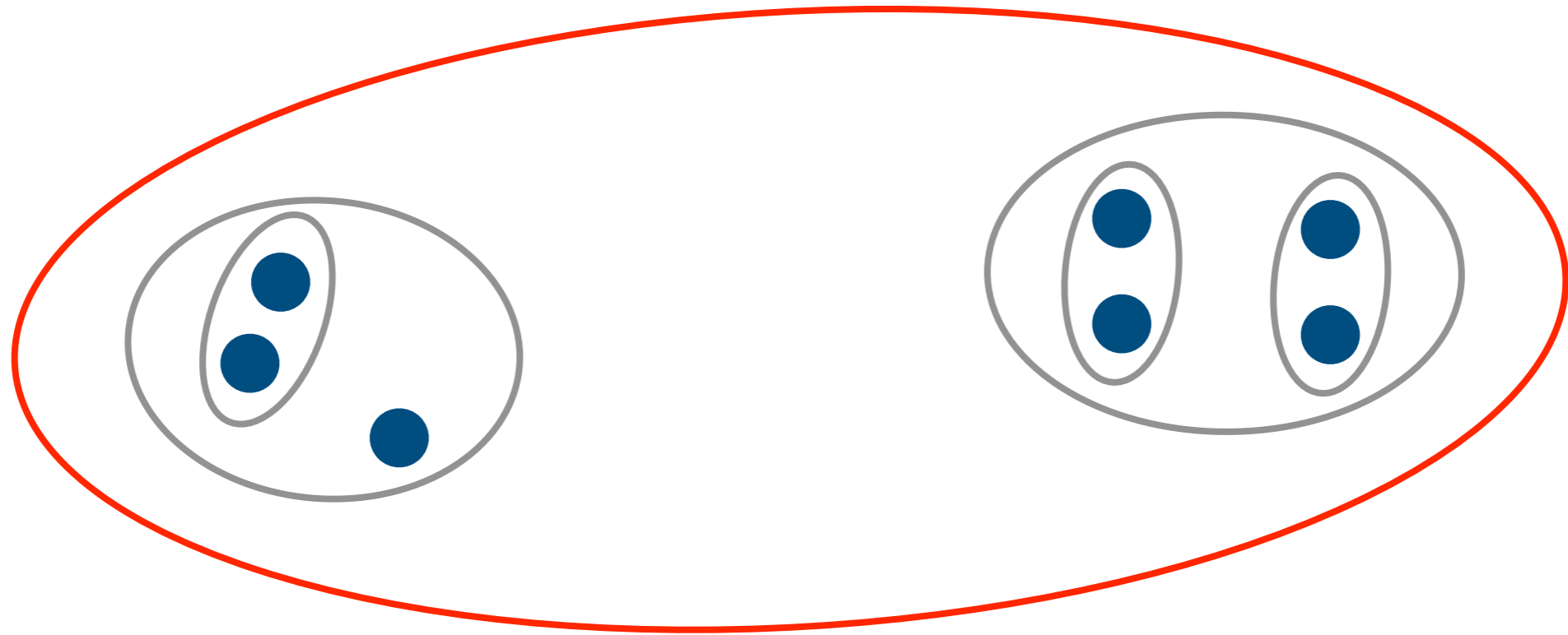
Strategy: Combine Nearby Points/Groups
(and repeat!)



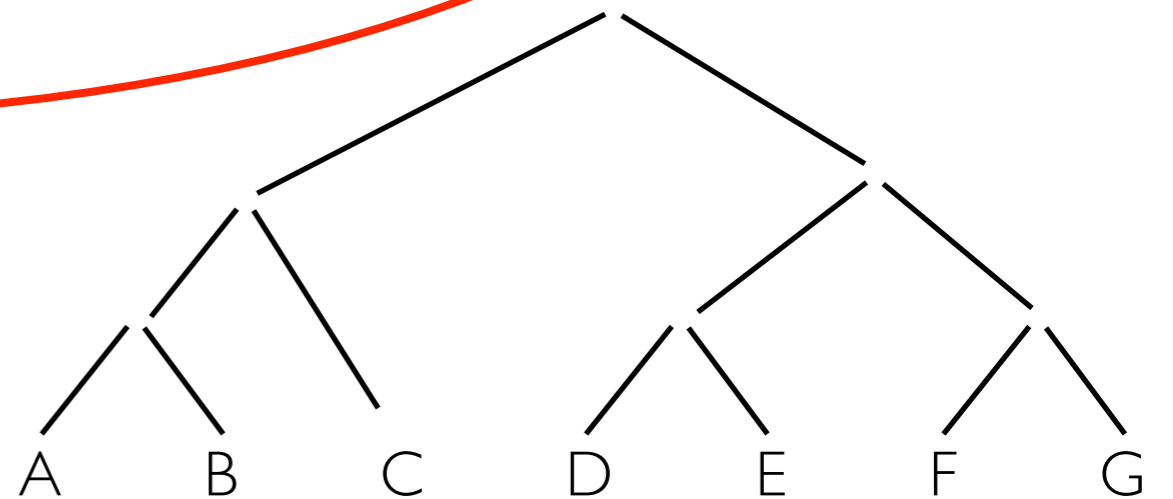
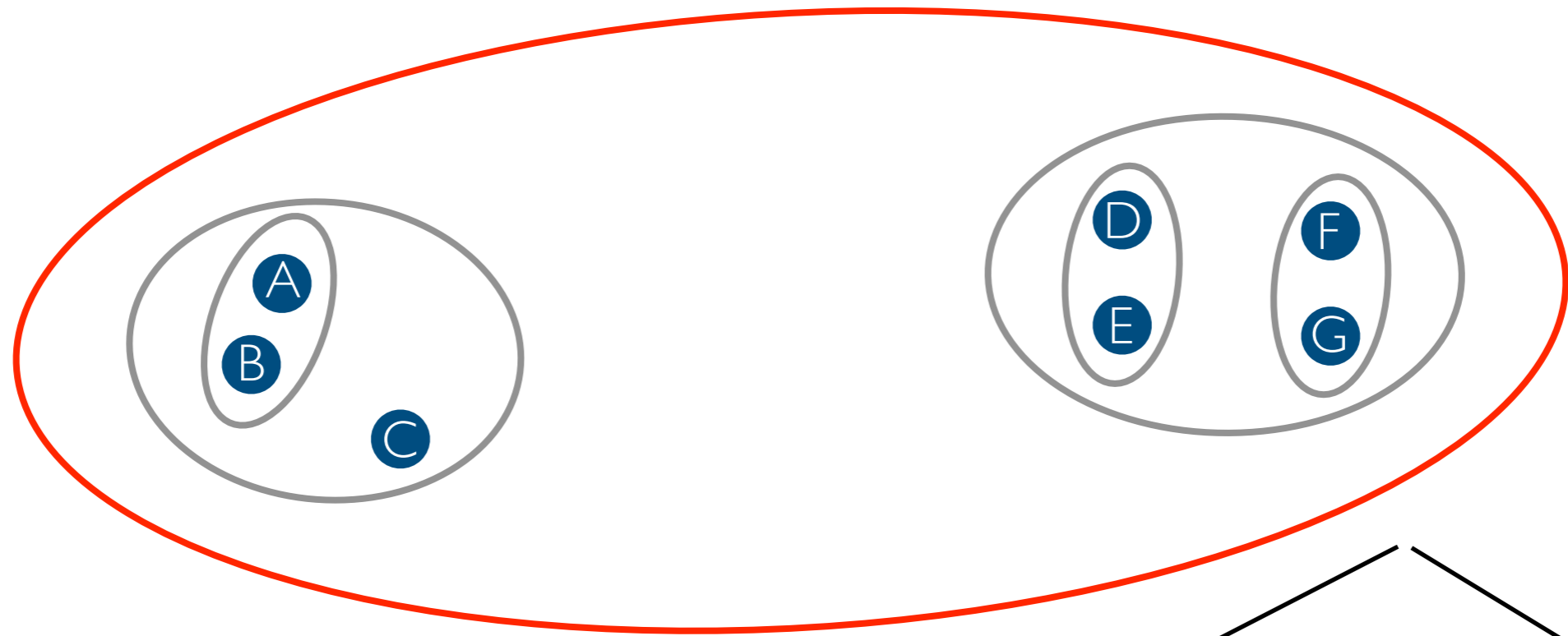
Strategy: Combine Nearby Points/Groups
(and repeat!)



Strategy: Combine Nearby Points/Groups
(and repeat!)



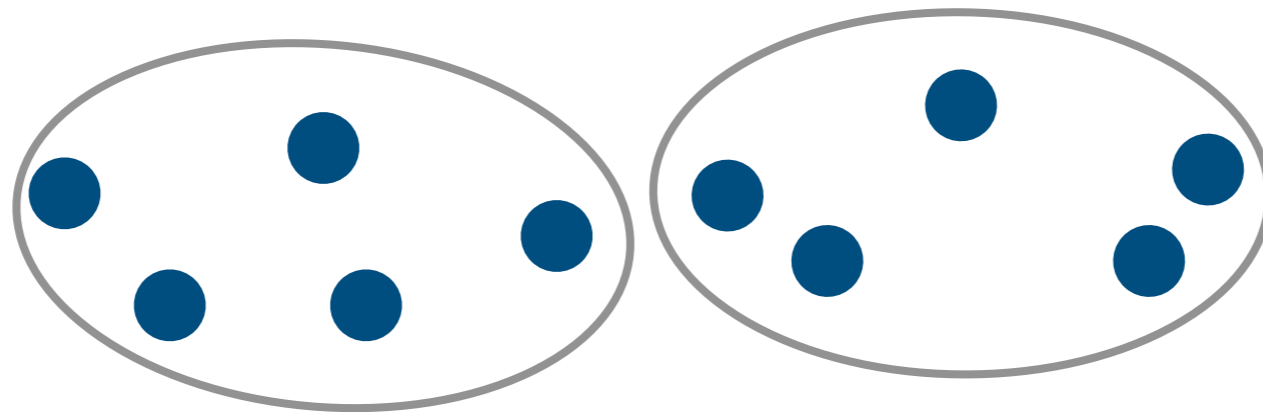
Strategy: Combine Nearby Points/Groups (and repeat!)



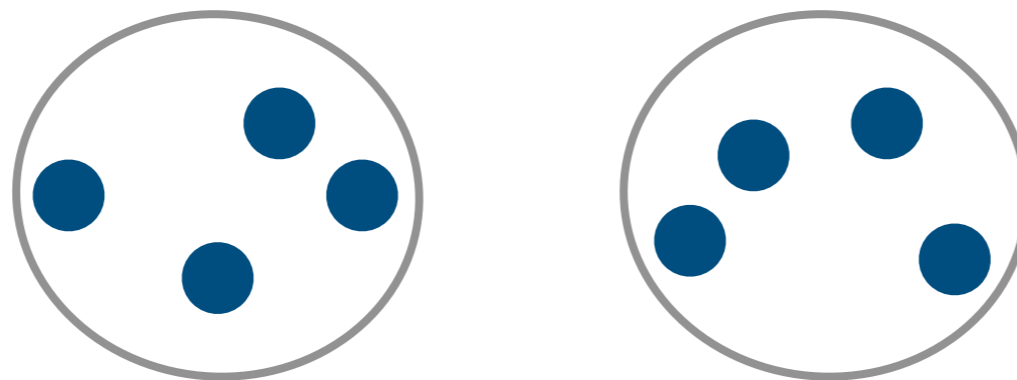
Configuration: what is "nearest"?

option: `linkage`

Configuration: what is "nearest"?

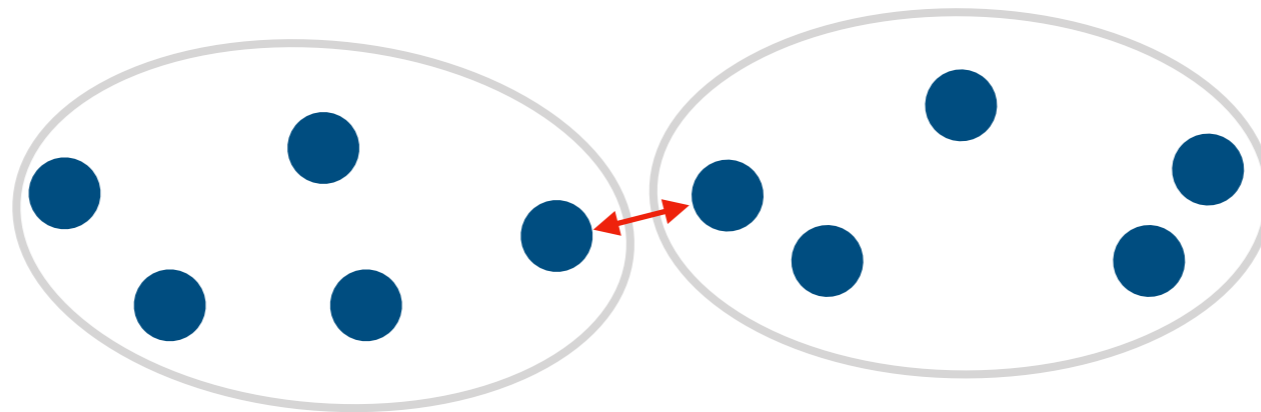


OR...

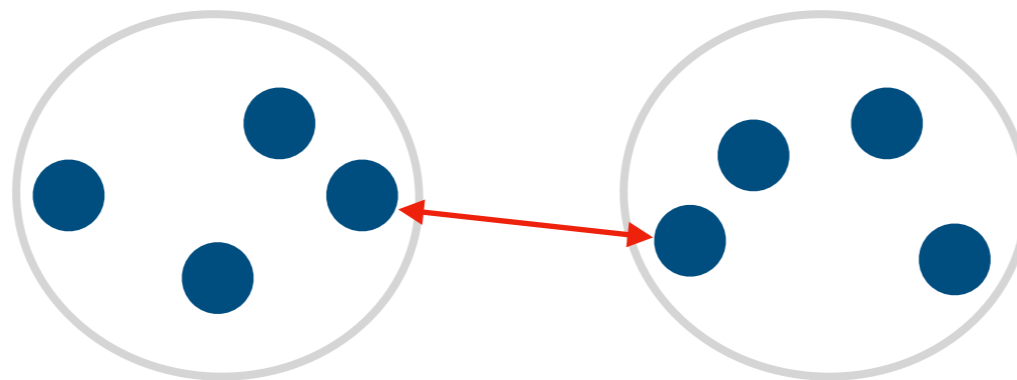


Configuration: what is "nearest"?

linkage="single"

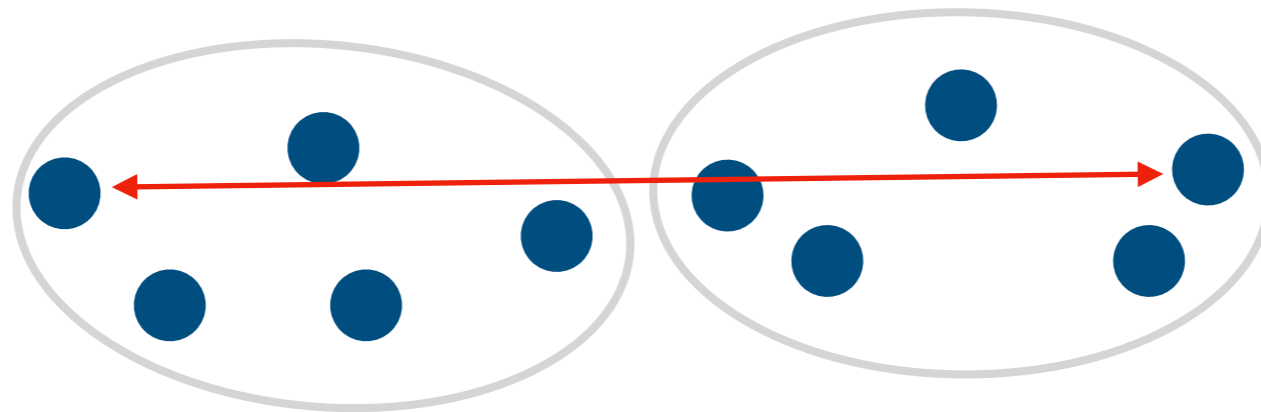


OR...

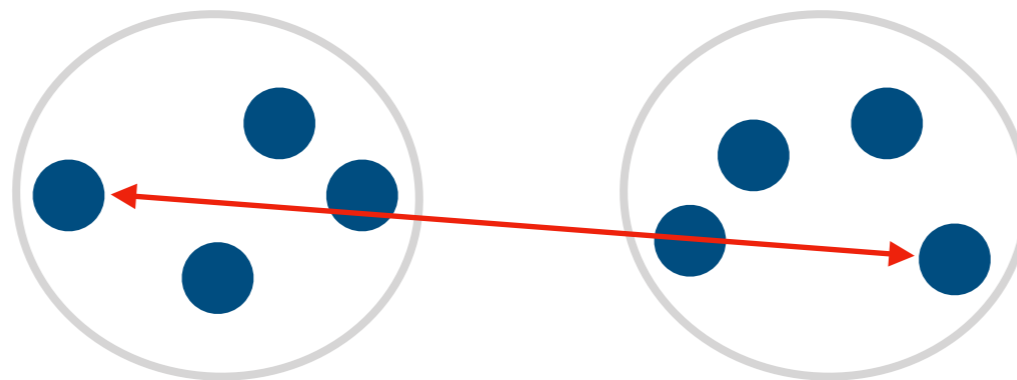


Configuration: what is "nearest"?

linkage="complete"



OR...



Configuration: what is "nearest"?

linkage="????"

From docs: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>

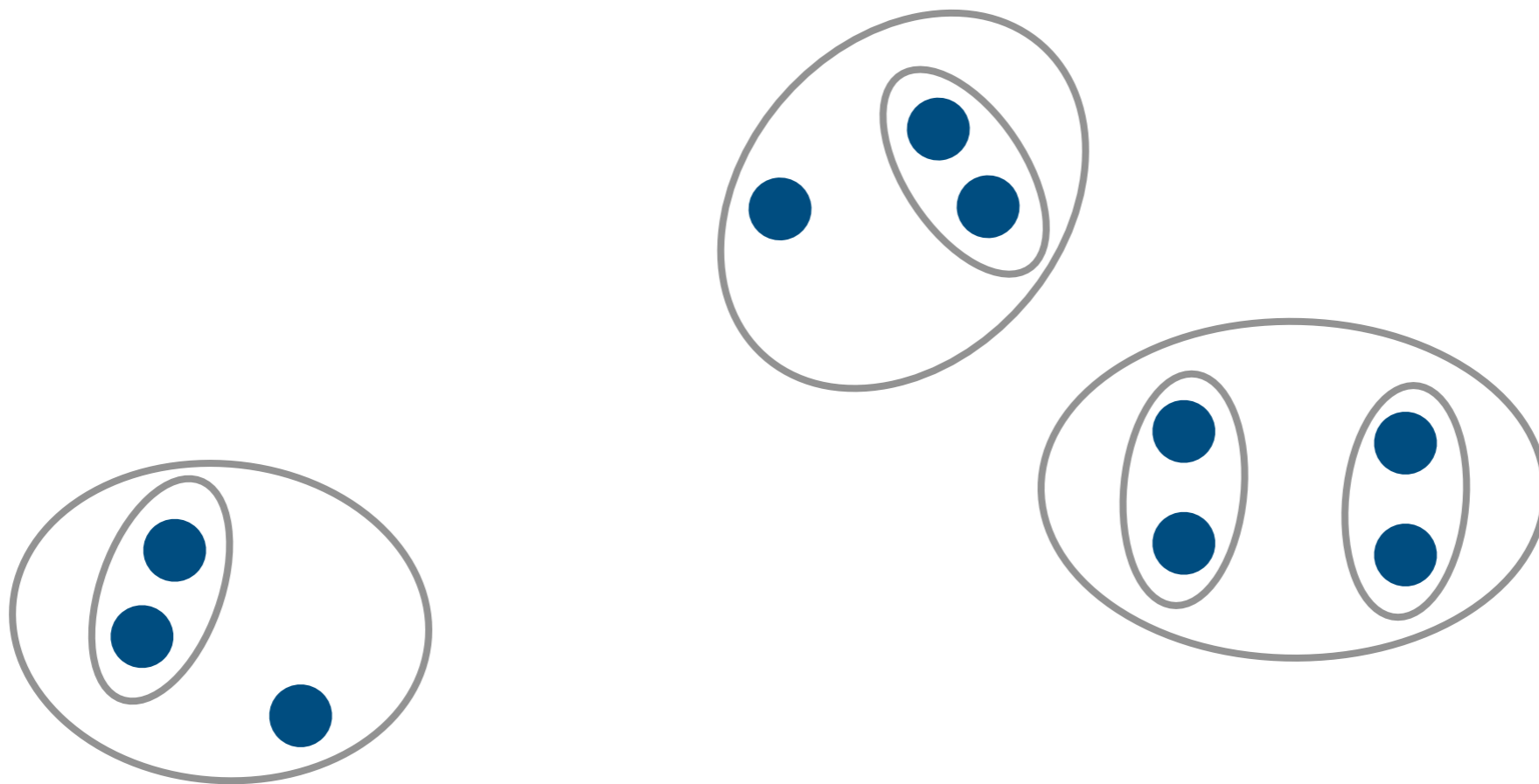
- **ward** minimizes the variance of the clusters being merged.
- **average** uses the average of the distances of each observation of the two sets.
- **complete** or maximum linkage uses the maximum distances between all observations of the two sets.
- **single** uses the minimum of the distances between all observations of the two sets.

Configuration: when to stop?

option: `n_clusters` or `distance_threshold`

Configuration: when to stop?

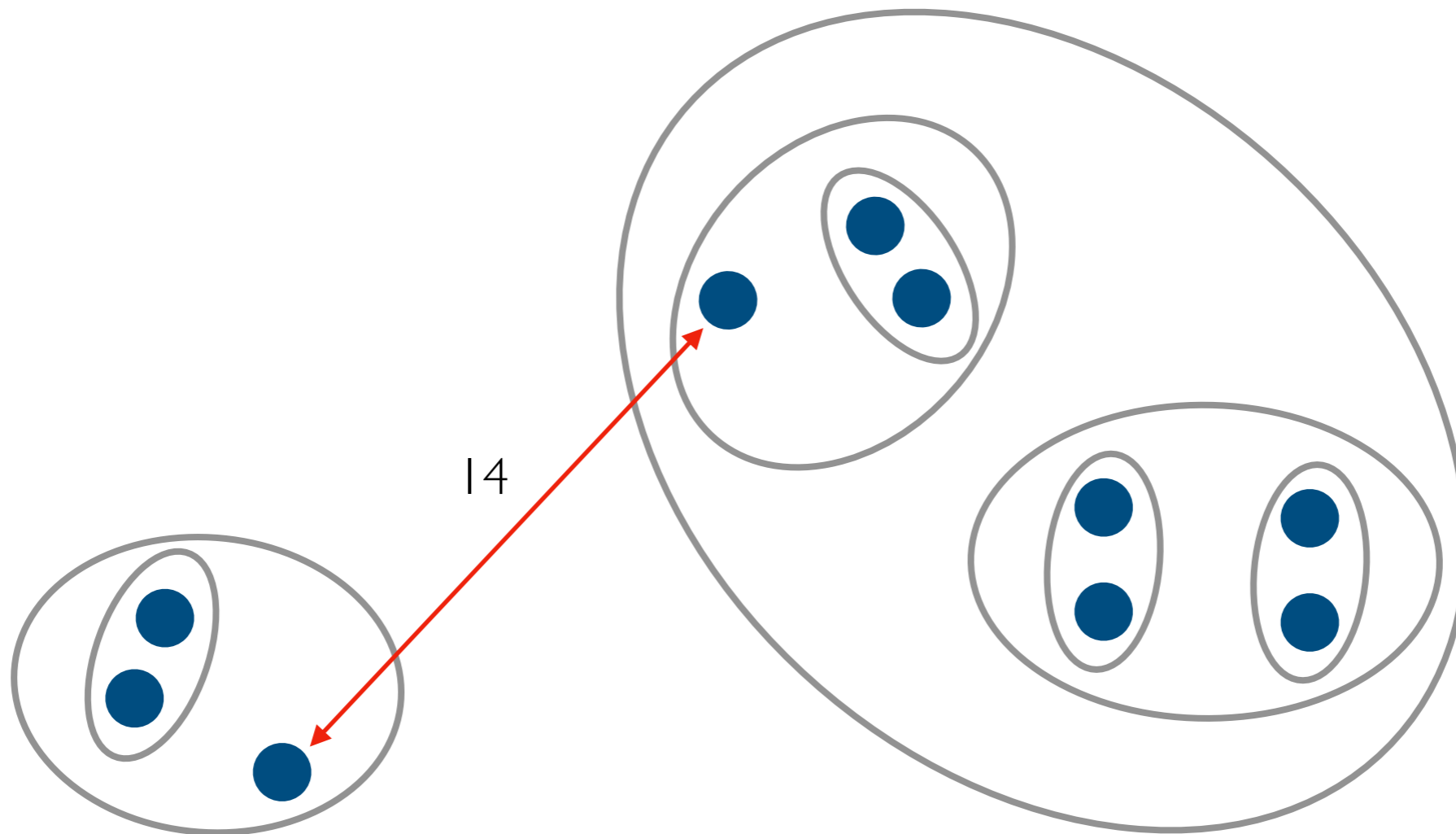
n_clusters=3



each cluster is it's own tree!

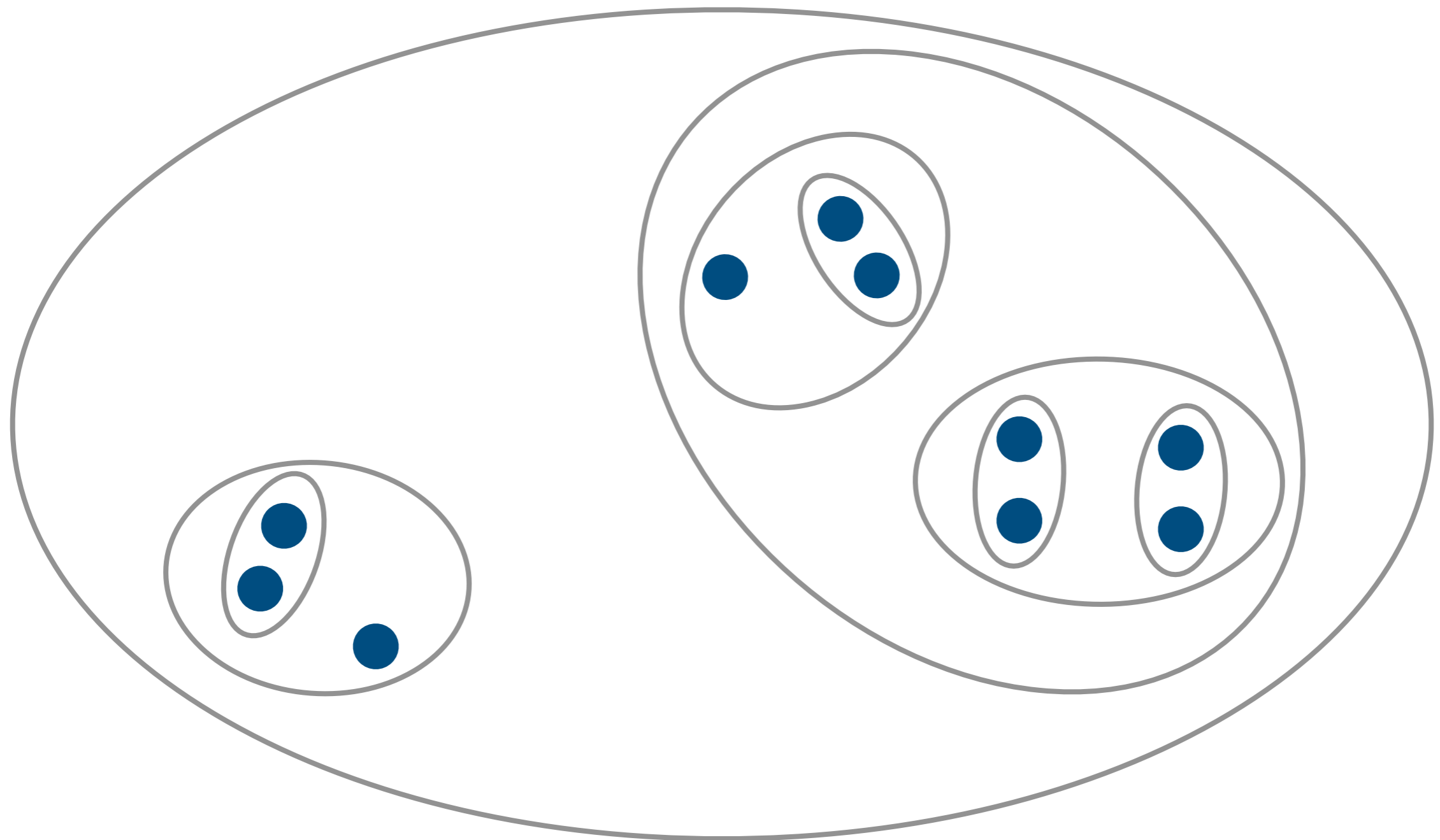
Configuration: when to stop?

distance_threshold=10

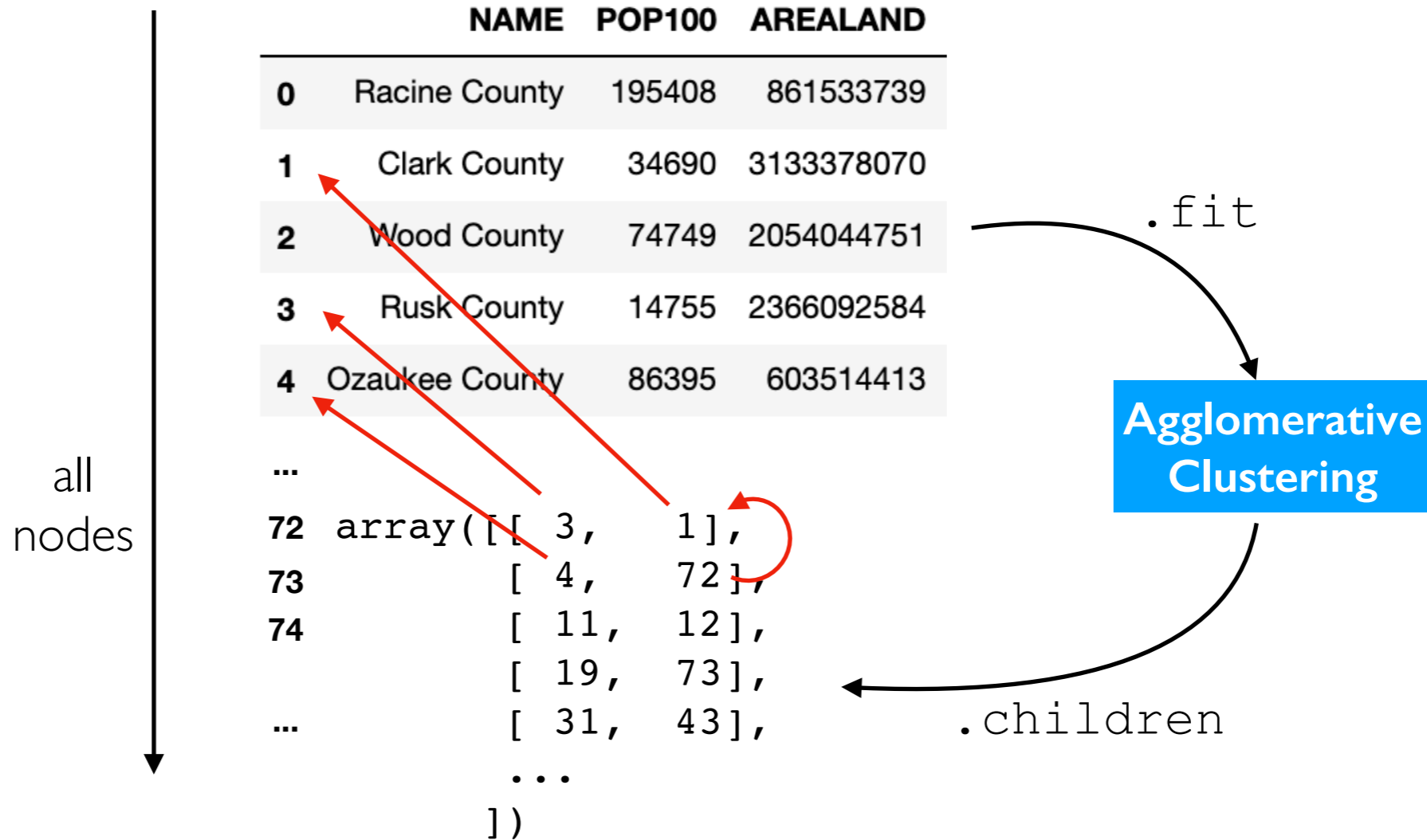


Configuration: when to stop?

distance_threshold=0



Node Representation



Linkage Matrix

	left child	right child	distances	node count
N				
N+1				
N+2				
...				