[320] Hierarchical Clustering

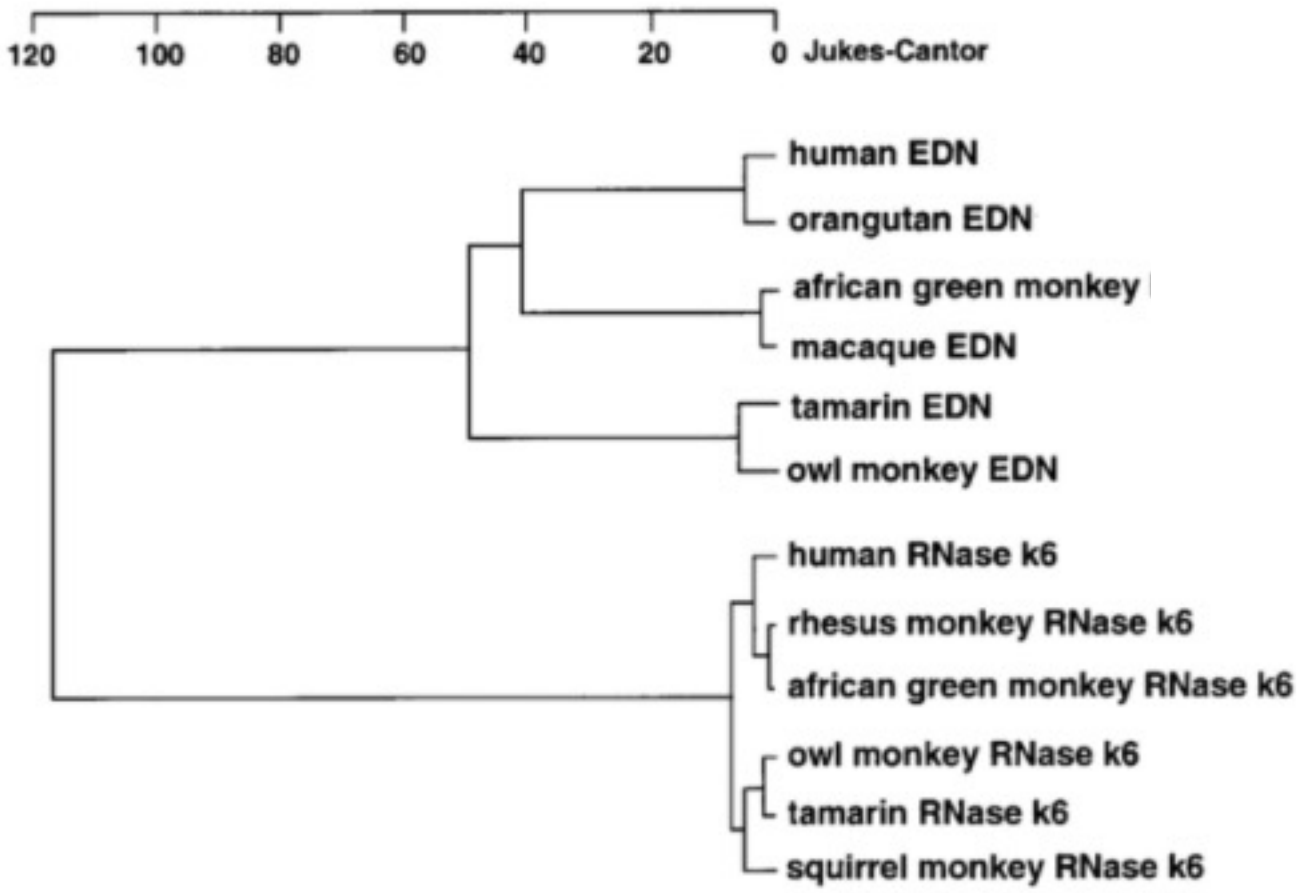
(AgglomerativeClustering and Dendrograms)

Meenakshi Syamkumar

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

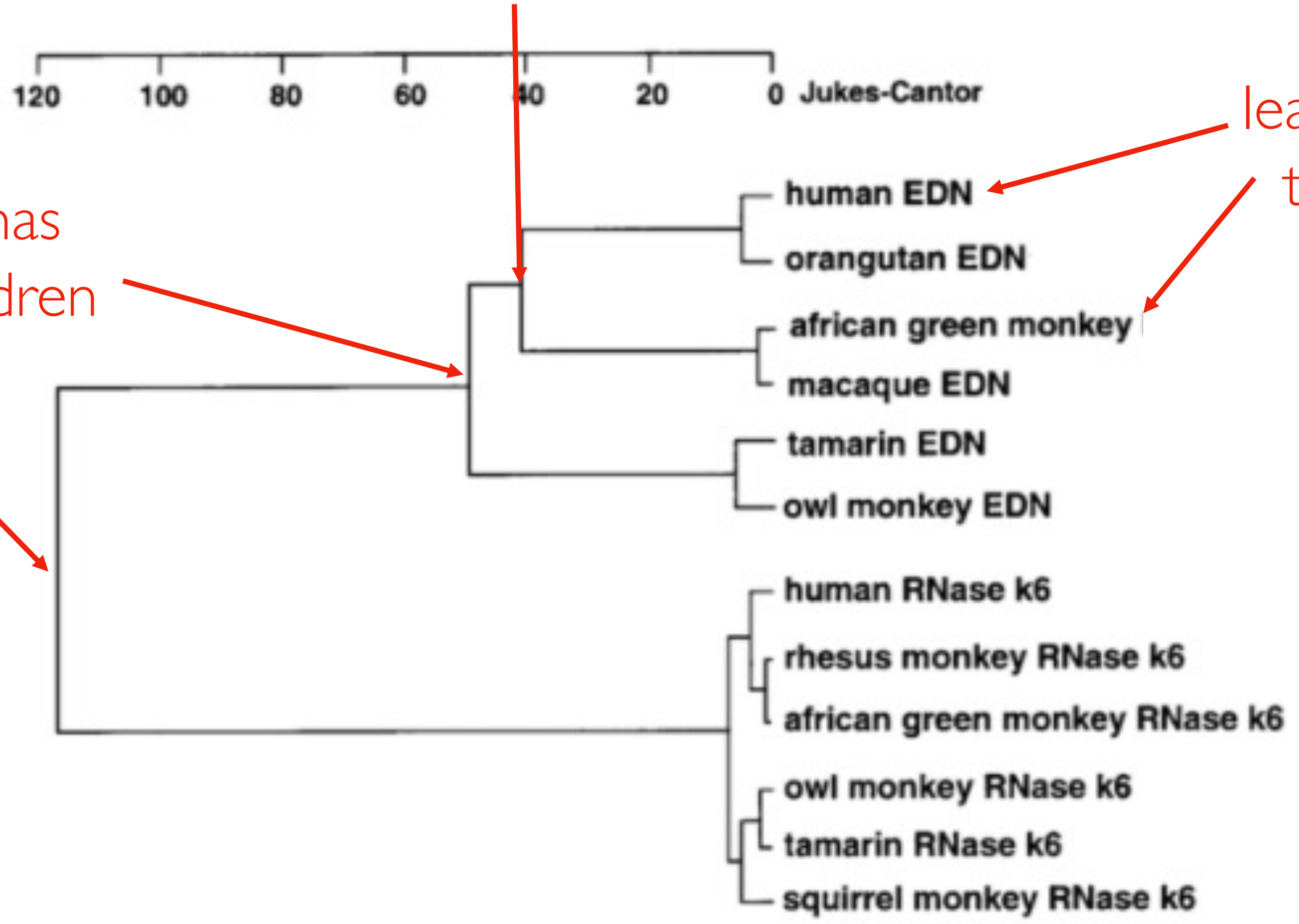
Hierarchical clusters can contain other custers
(example: AgglomerativeClustering)

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



this node's children have ~40 distance

every non-leaf has
EXACTLY 2 children

leaves correspond
to original data

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

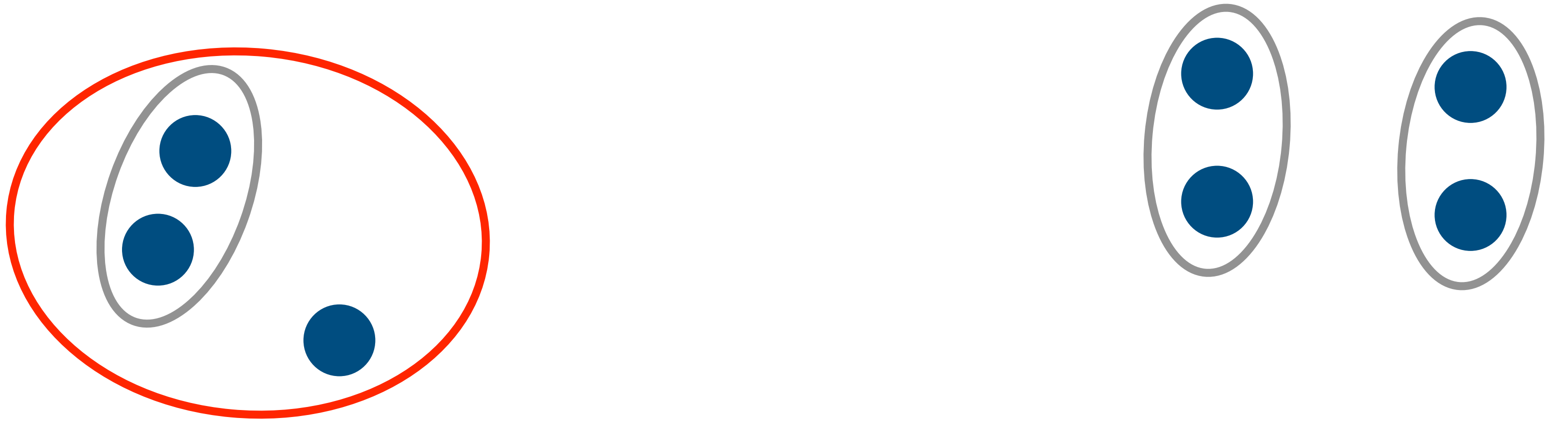
We'll represent hierarchies as special binary trees.

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

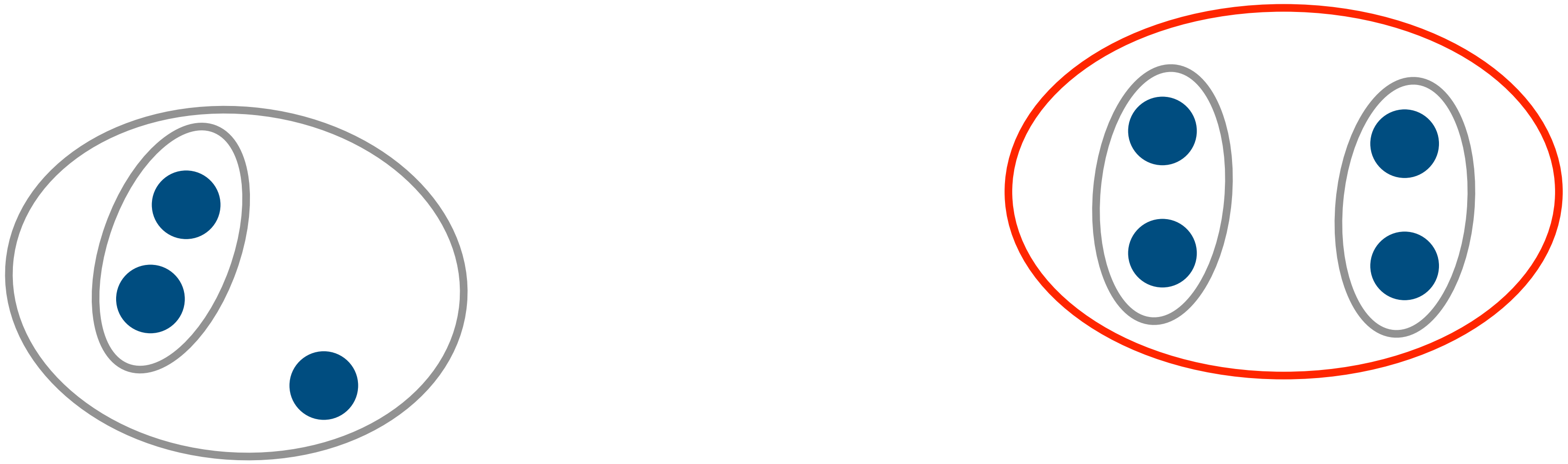
Strategy: Combine Nearby Points/Groups (and repeat!)

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

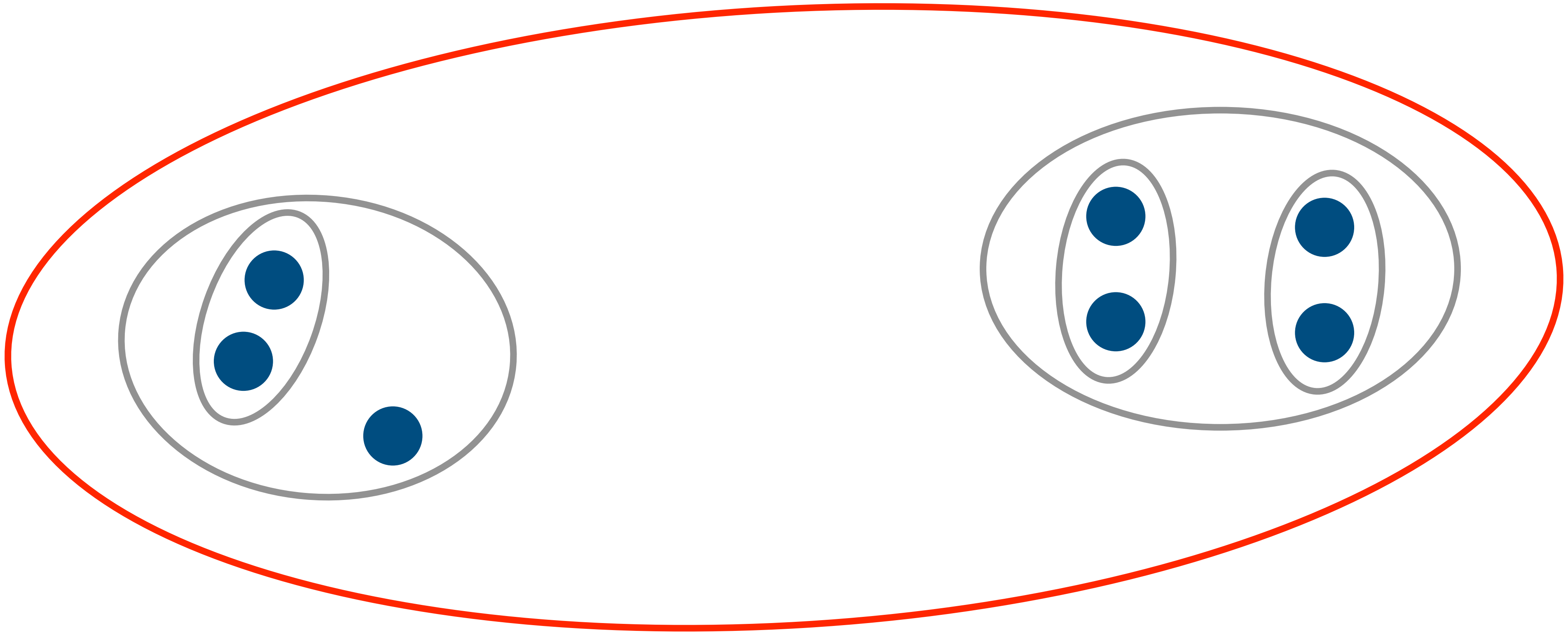
Strategy: Combine Nearby Points/Groups (and repeat!)



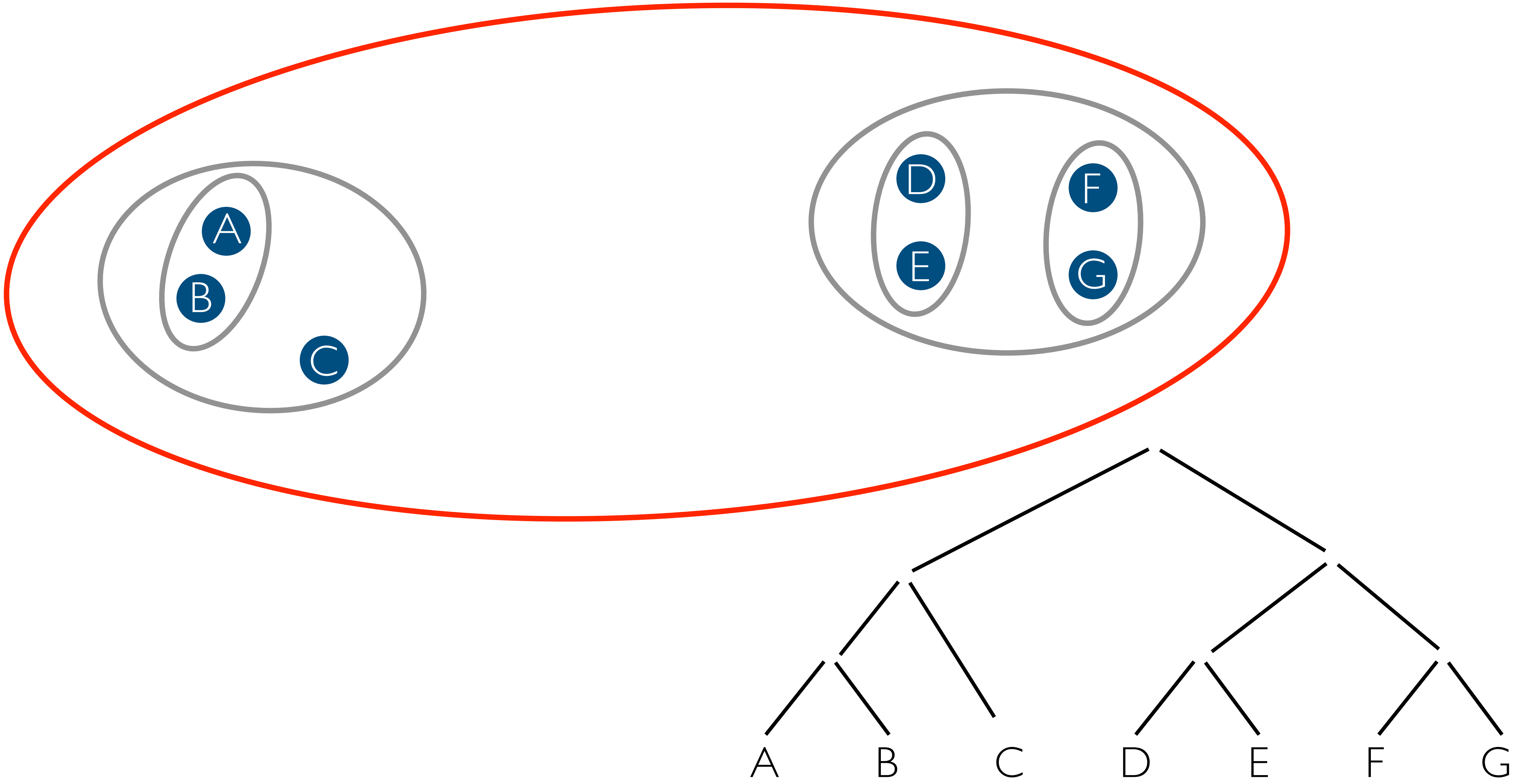
Strategy: Combine Nearby Points/Groups (and repeat!)



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



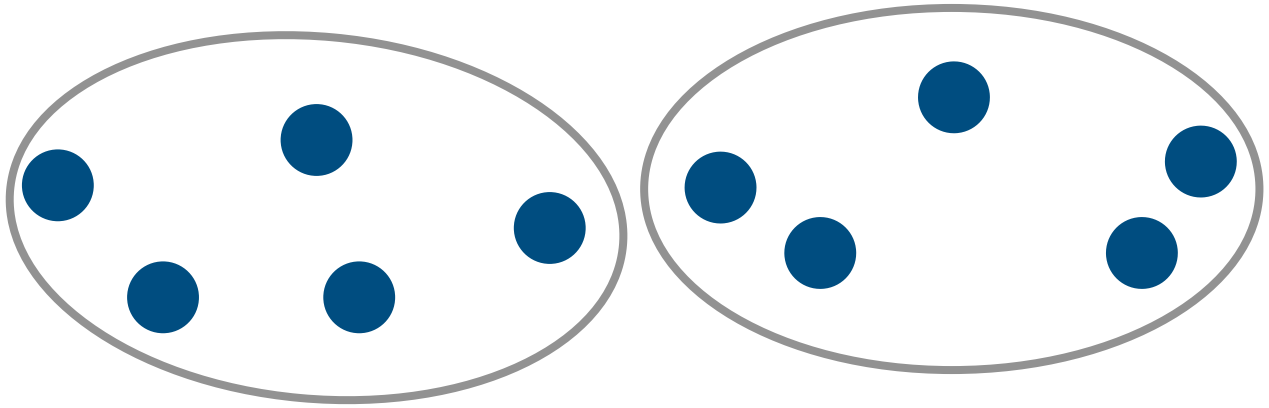
Strategy: Combine Nearby Points/Groups (and repeat!)



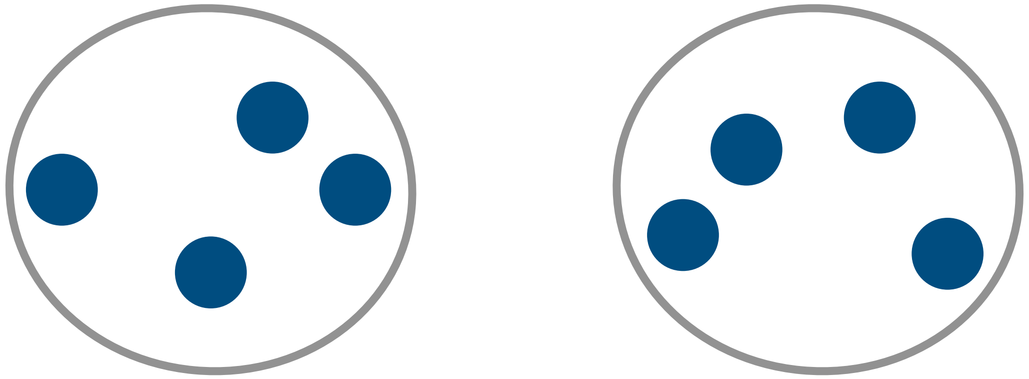
Configuration: what is "nearest"?

option: `linkage`

Configuration: what is "nearest"?

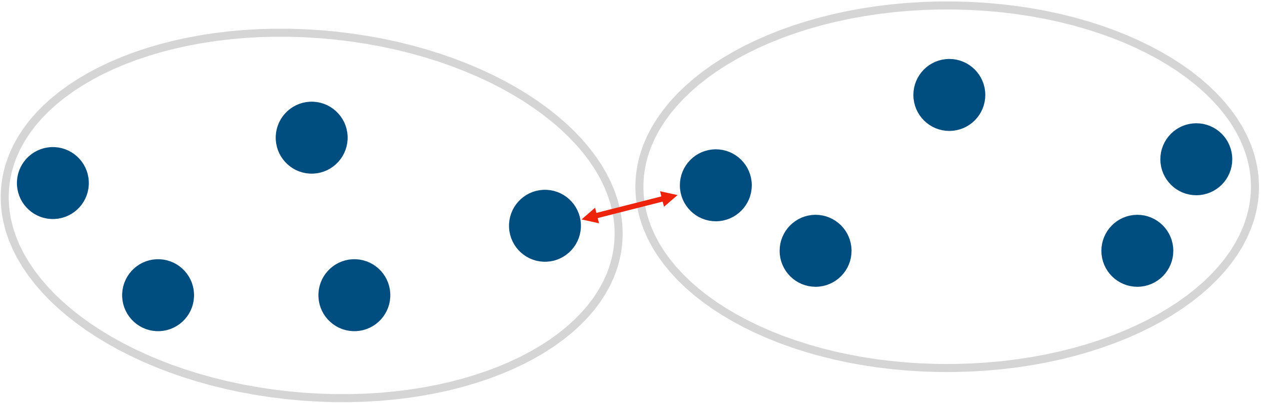


OR...

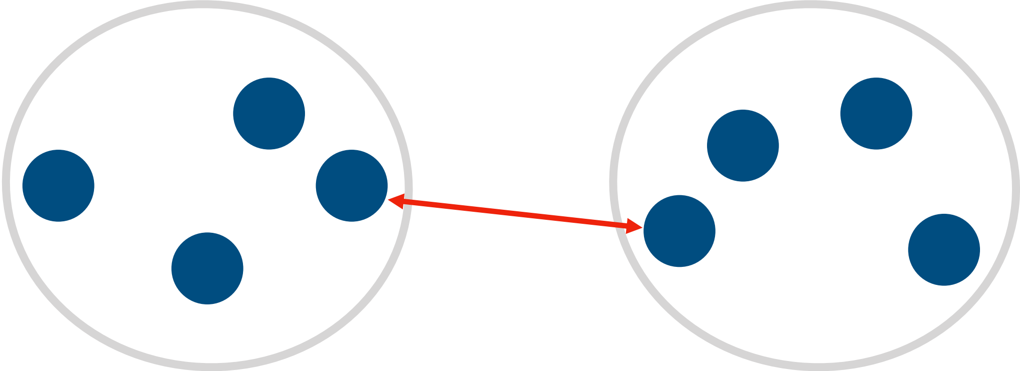


Configuration: what is "nearest"?

linkage="single"

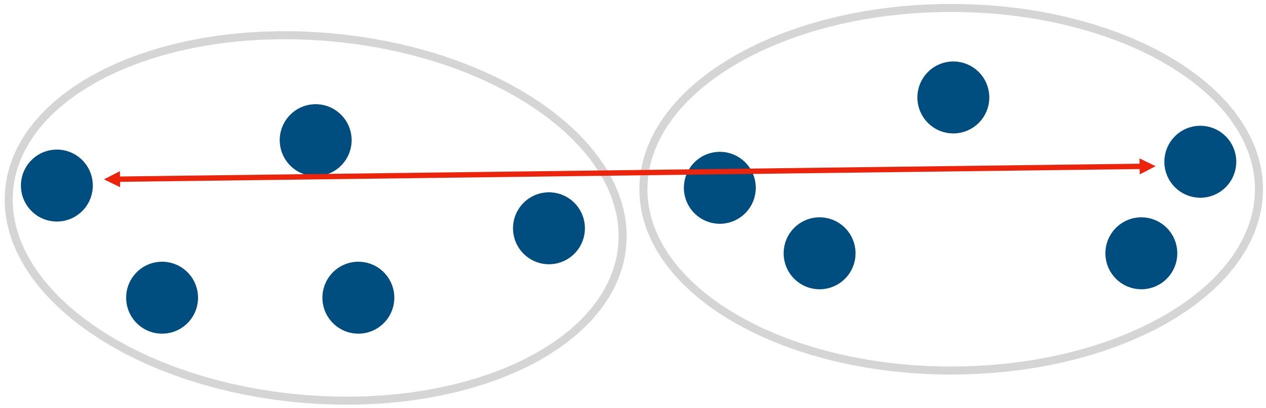


OR...

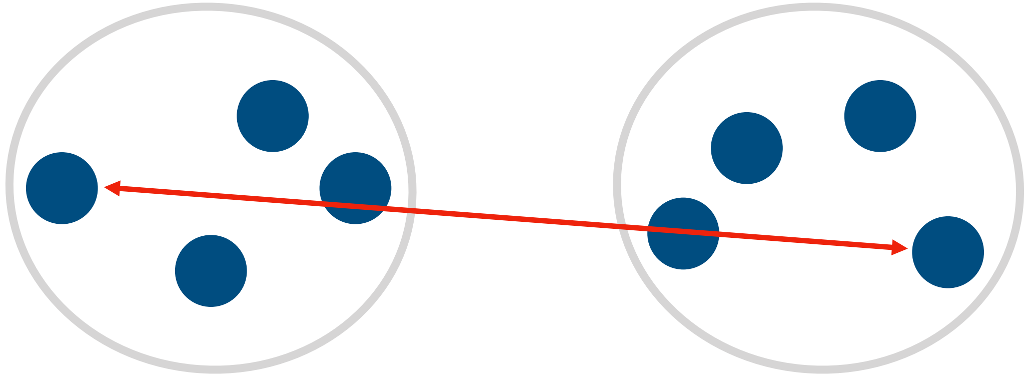


Configuration: what is "nearest"?

linkage="complete"



OR...



Configuration: what is "nearest"?

linkage="????"

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

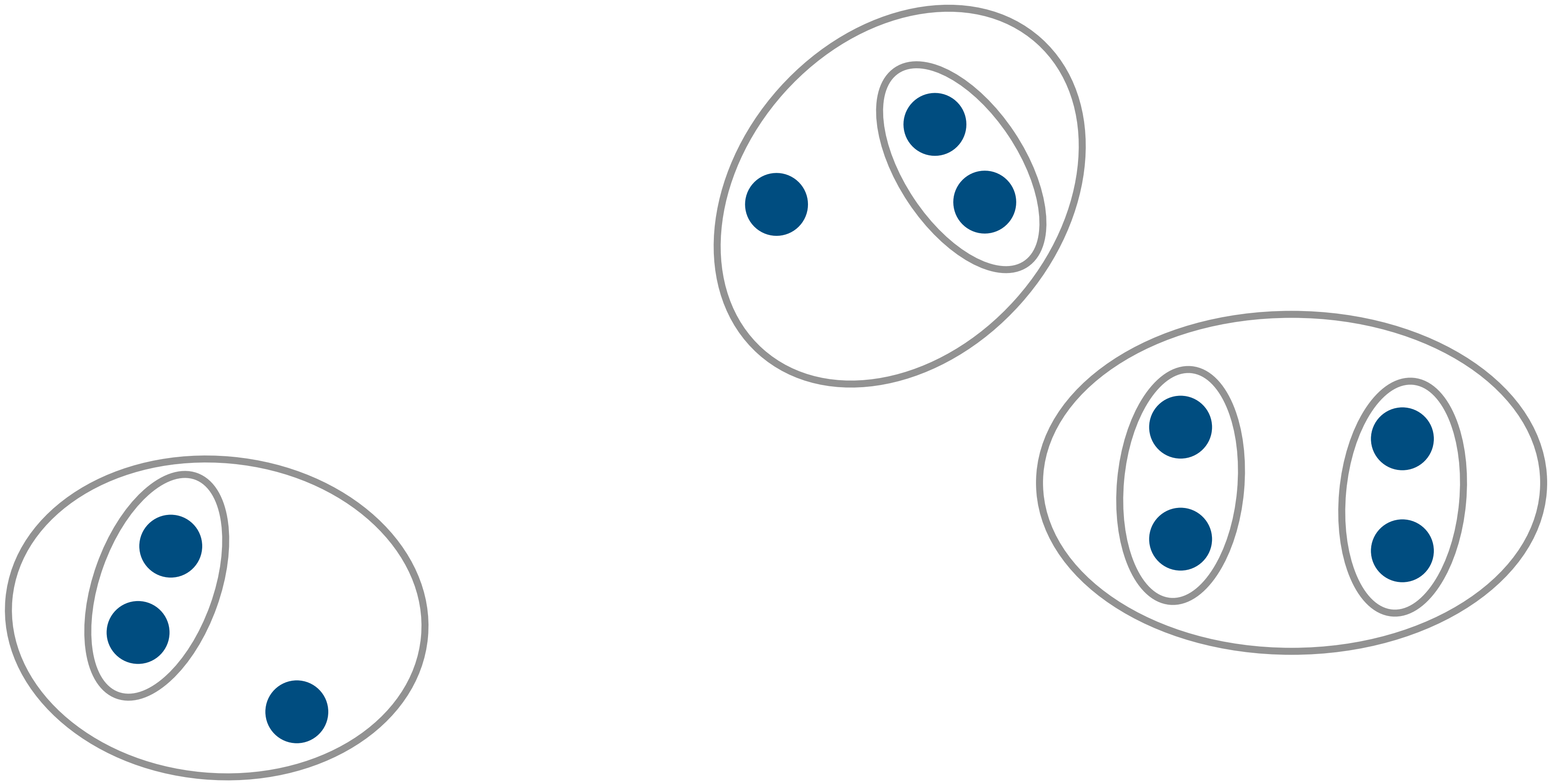
- ward minimizes the variance of the clusters being merged.
- average uses the average of the distances of each observation of the two sets.
- complete or maximum linkage uses the maximum distances between all observations of the two sets.
- single uses the minimum of the distances between all observations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

Configuration: when to stop?

option: `n_clusters` or `distance_threshold`

Configuration: when to stop?

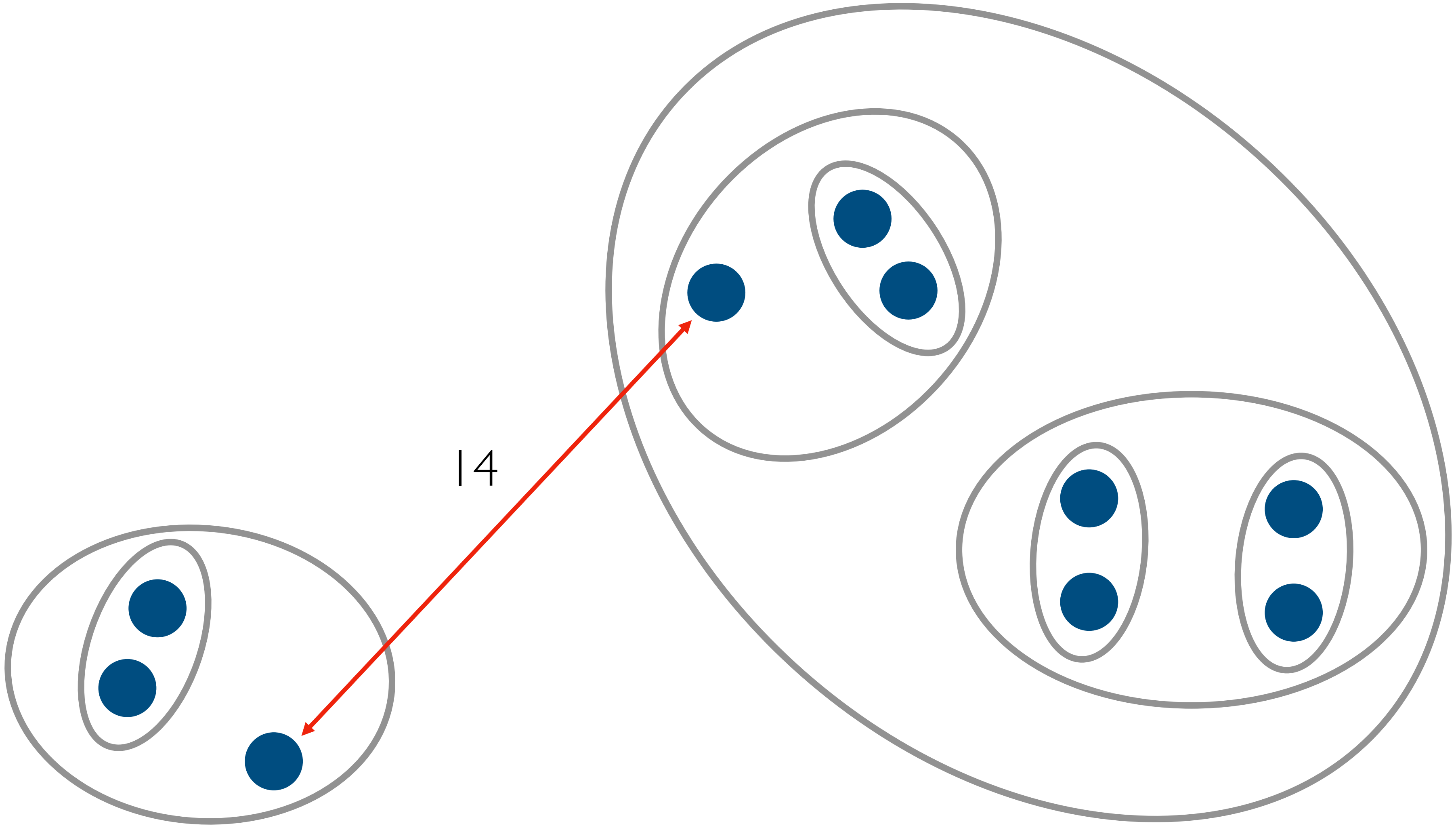
n_clusters=3



each cluster is it's own tree!

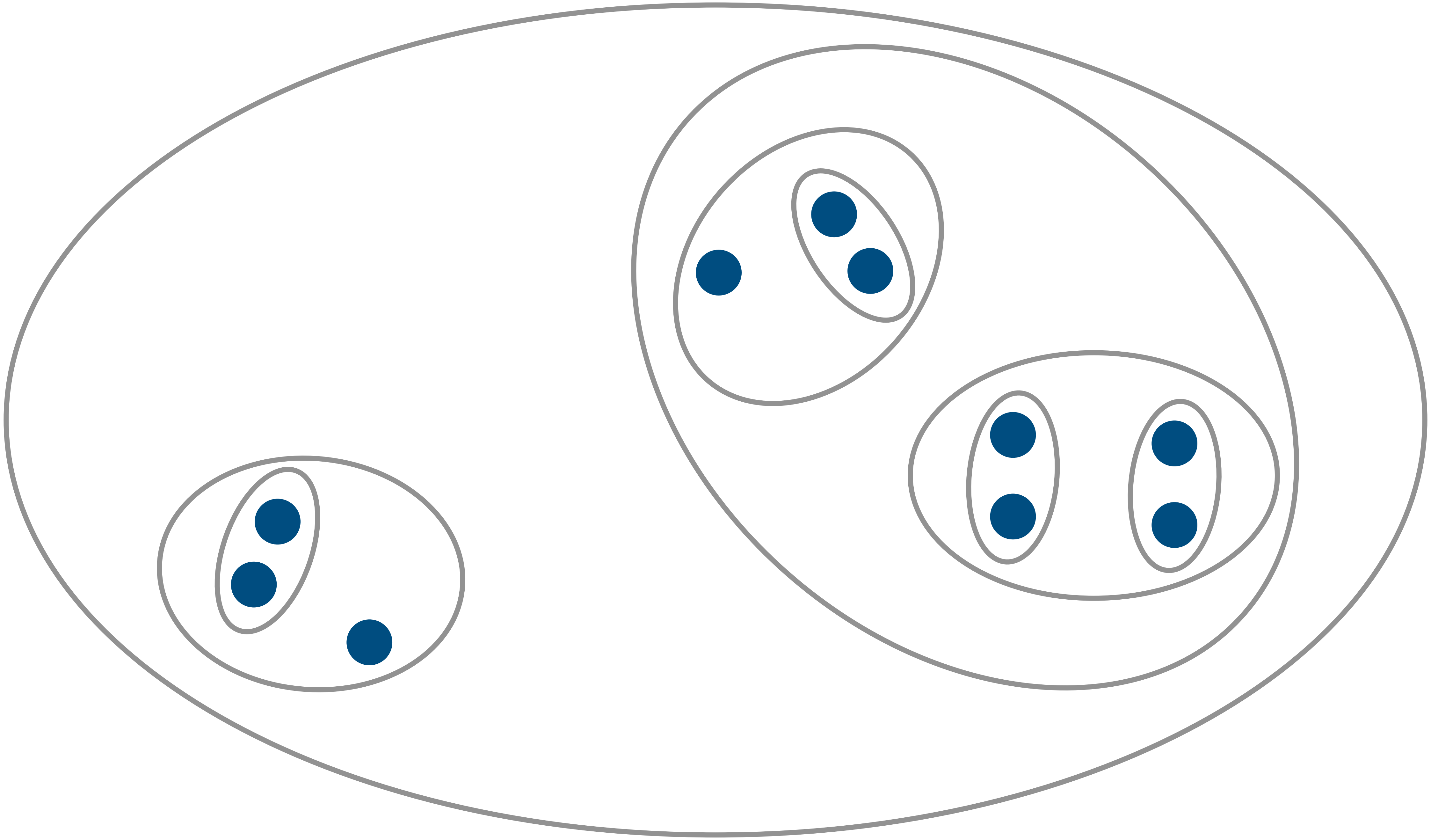
Configuration: when to stop?

distance_threshold=10



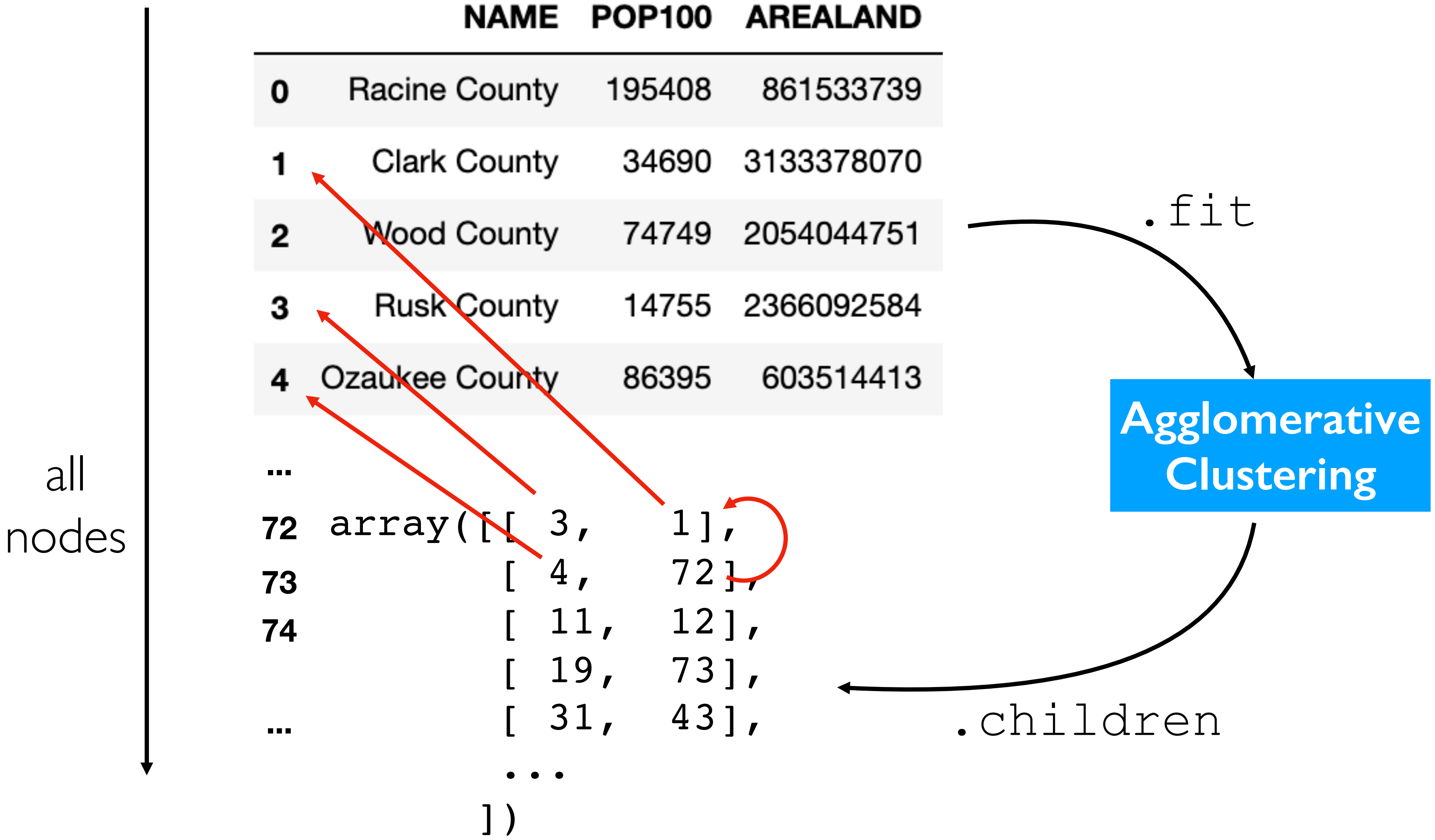
Configuration: when to stop?

distance_threshold=0



Demos…

Node Representation



	NAME	POP100	AREALAND
0	Racine County	195408	861533739
1	Clark County	34690	3133378070
2	Wood County	74749	2054044751
3	Rusk County	14755	2366092584
4	Ozaukee County	86395	603514413
...			

```
72  array([[ 3,    1],
73         [ 4,   72],
74         [ 11,  12],
           [ 19,  73],
...        [ 31,  43],
           ...
          ])
```

Linkage Matrix

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