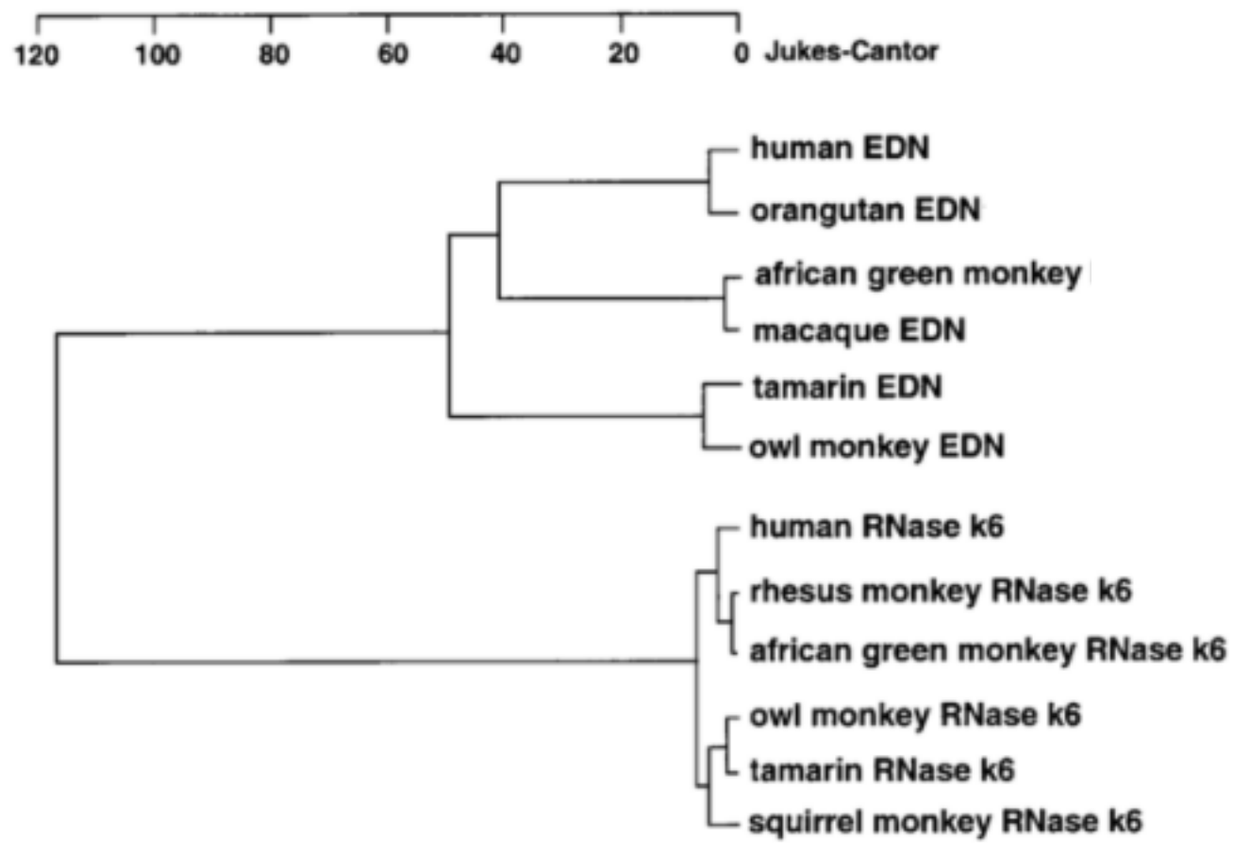
[320] Hierarchical Clustering

(AgglomerativeClustering and Dendrograms)

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

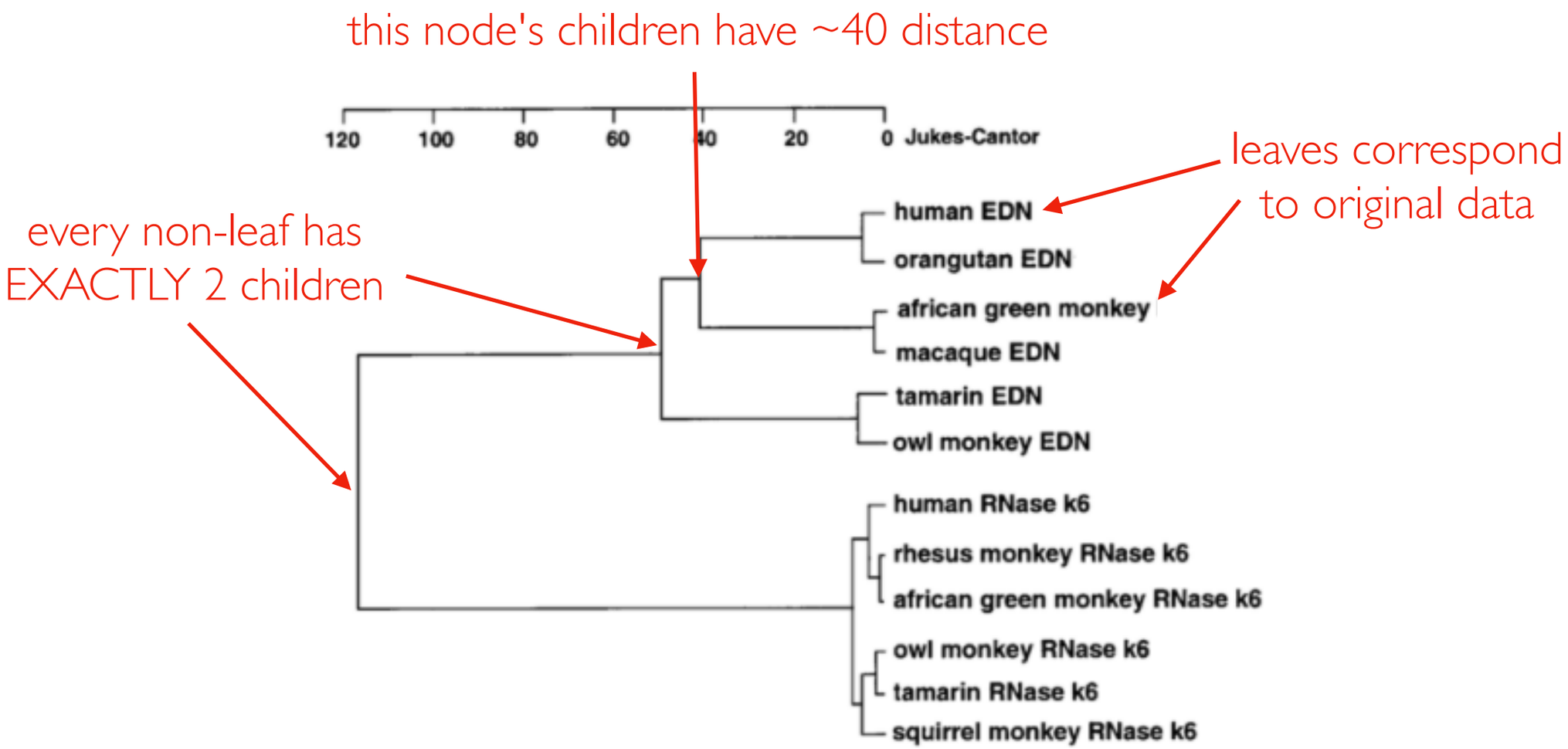
Hierarchical clusters can contain other custers
(example: AgglomerativeClustering)

Hierarchical Clusters with Dendrograms



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

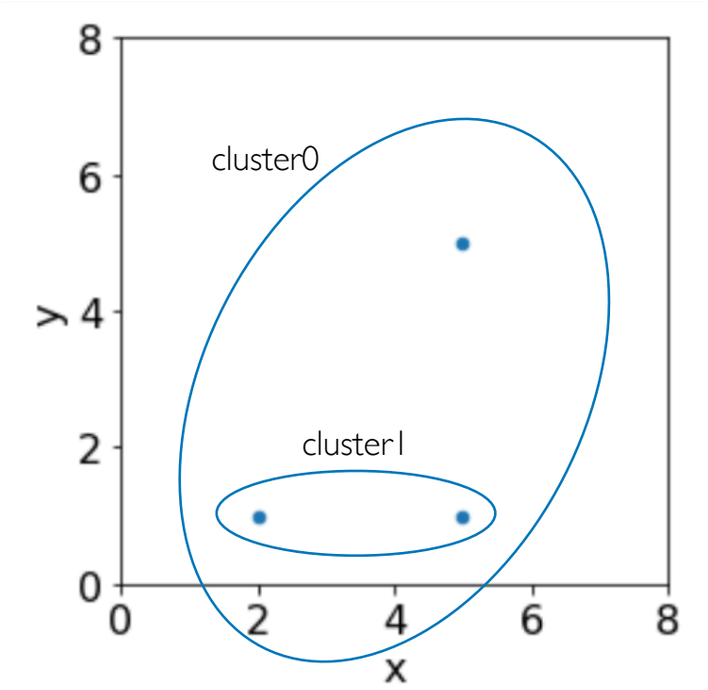
Hierarchical Clusters with Dendrograms



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

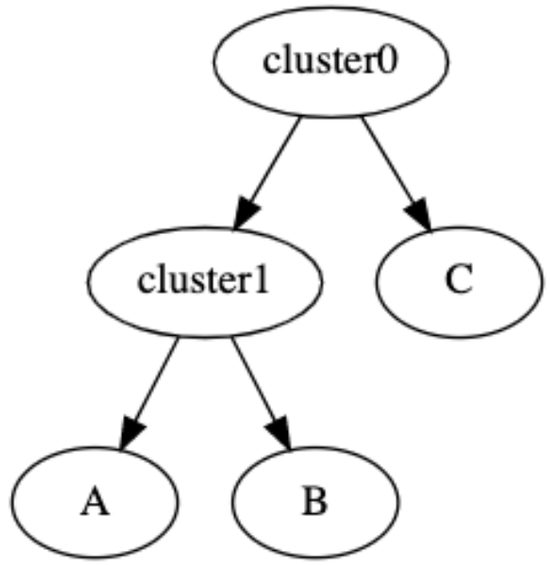
We'll represent hierarchies as special binary trees.

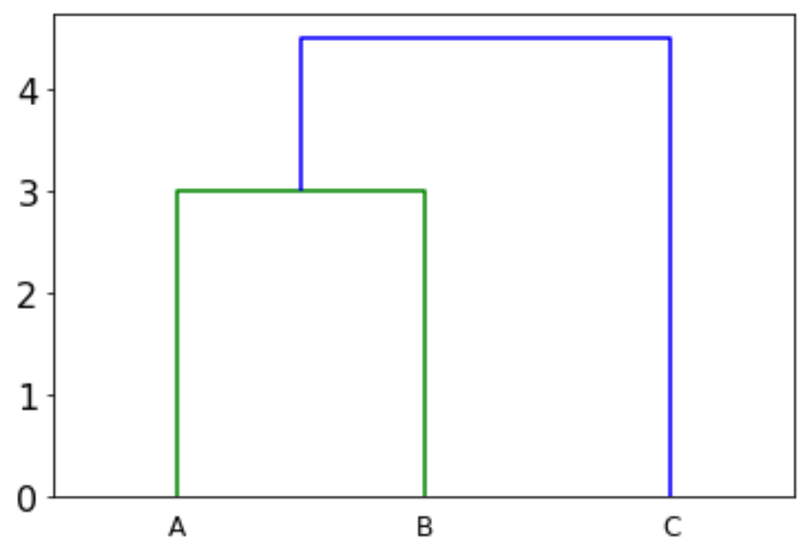
5 views of 3 data points



	x	y
A	2	1
B	5	1
C	5	5

original data



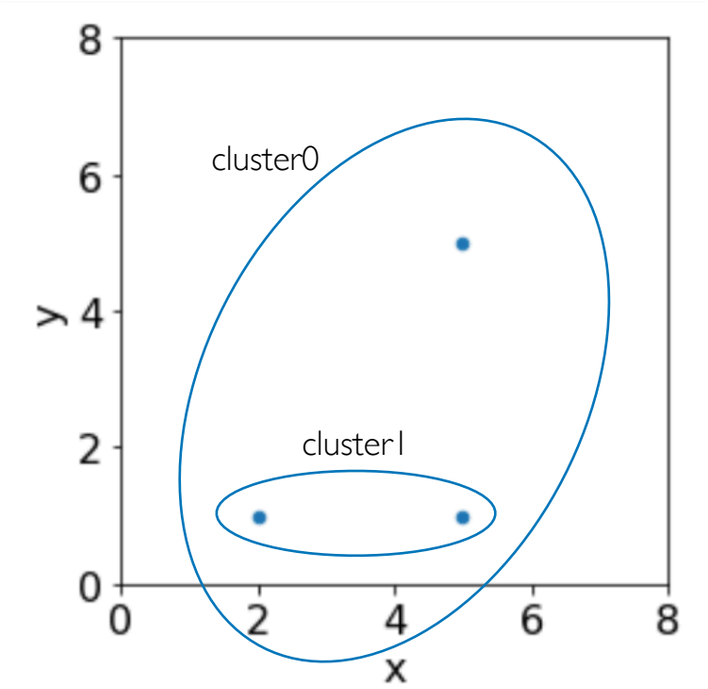


dendrogram

	name	left	right	distance	nodes
0	A			0	1
1	B			0	1
2	C			0	1
3	cluster1	0	1	3	3
4	cluster0	3	2	4.5	2

linkage matrix

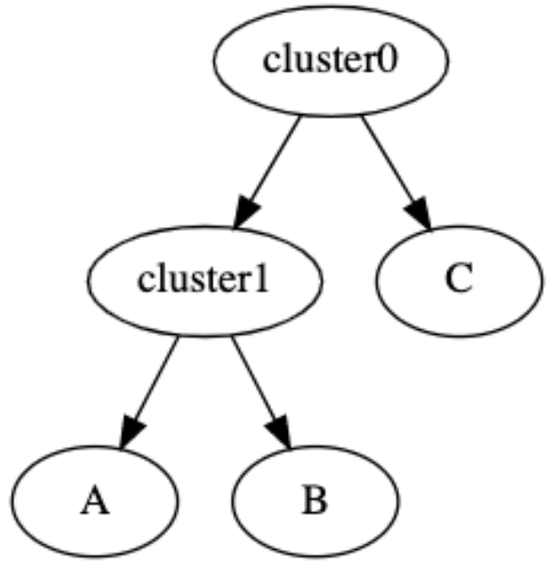
5 views of 3 data points



	x	y
A	2	1
B	5	1
C	5	5

original data

demo 3

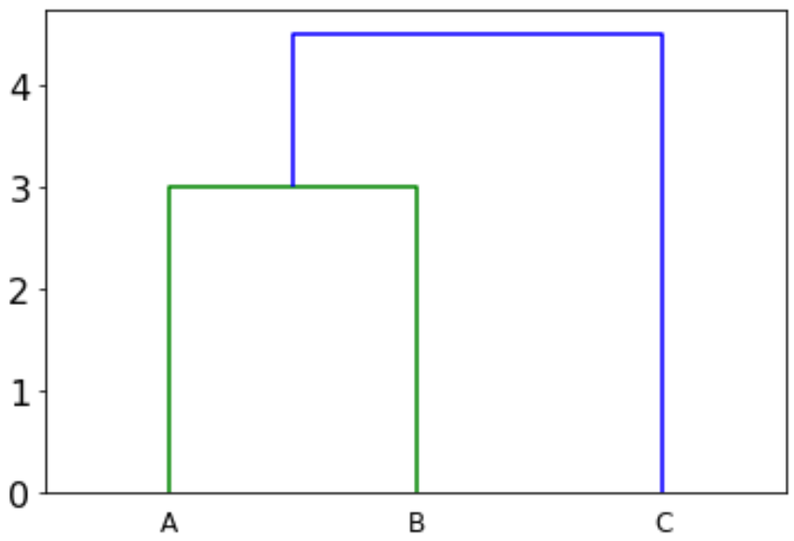


demo 2

	name	left	right	distance	nodes
0	A			0	1
1	B			0	1
2	C			0	1
3	cluster1	0	1	3	3
4	cluster0	3	2	4.5	2

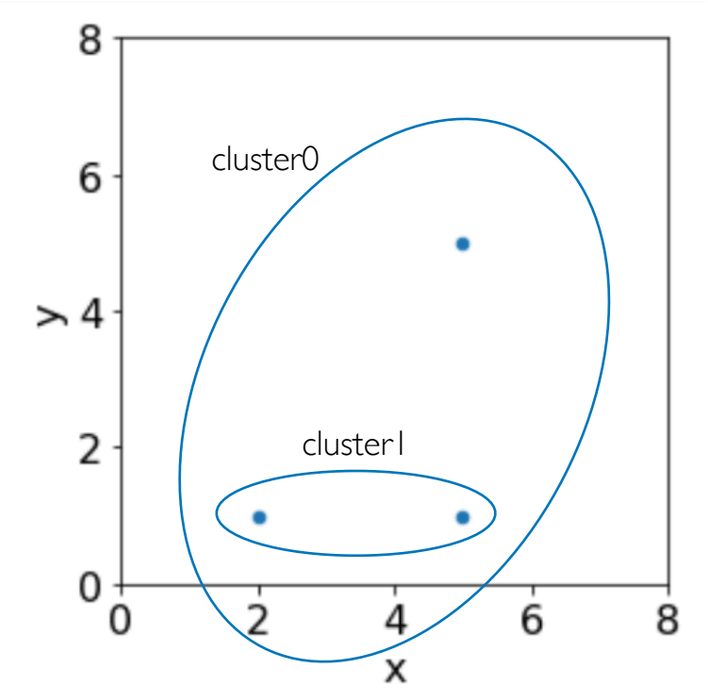
linkage matrix

demo 1



dendrogram

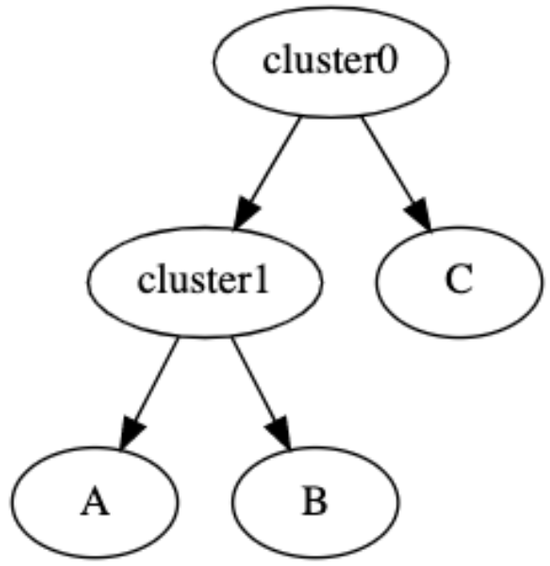
5 views of 3 data points



original data

	x	y
A	2	1
B	5	1
C	5	5

demo 3

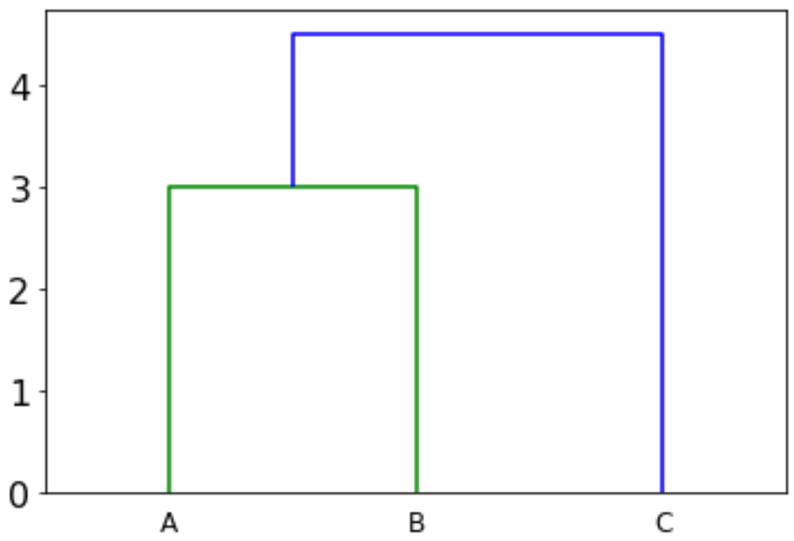


demo 2

linkage matrix

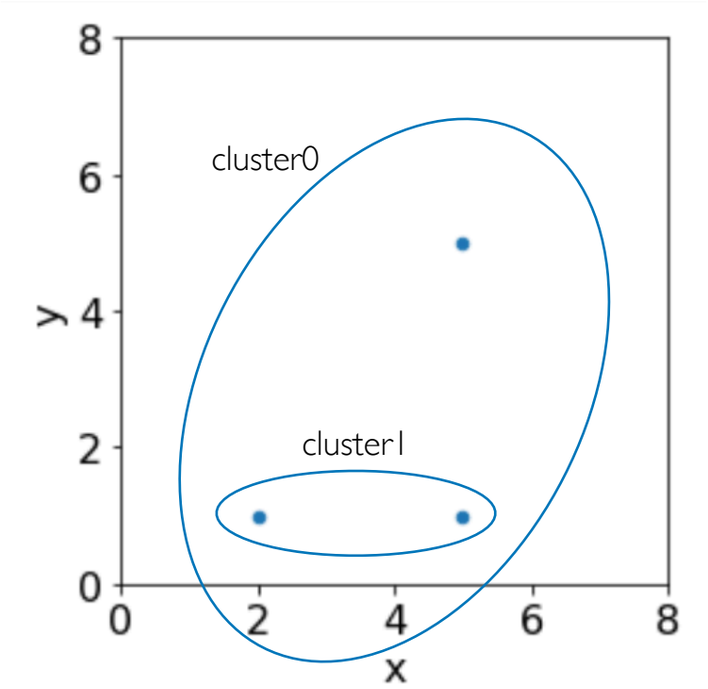
	name	left	right	distance	nodes
0	A			0	1
1	B			0	1
2	C			0	1
3	cluster1	0	1	3	3
4	cluster0	3	2	4.5	2

demo 1



dendrogram

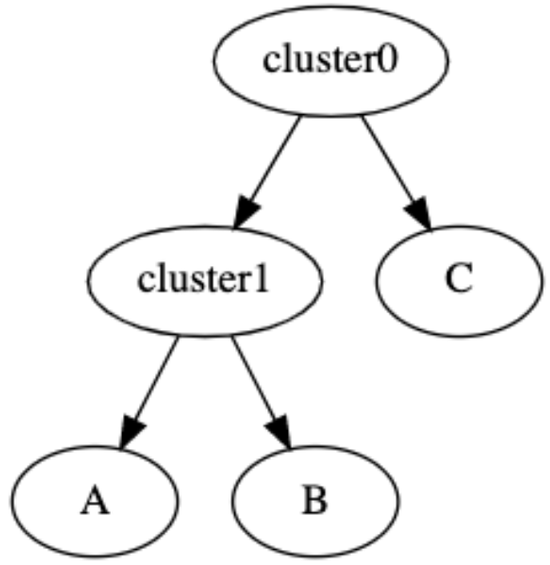
5 views of 3 data points



original data

	x	y
A	2	1
B	5	1
C	5	5

demo 3

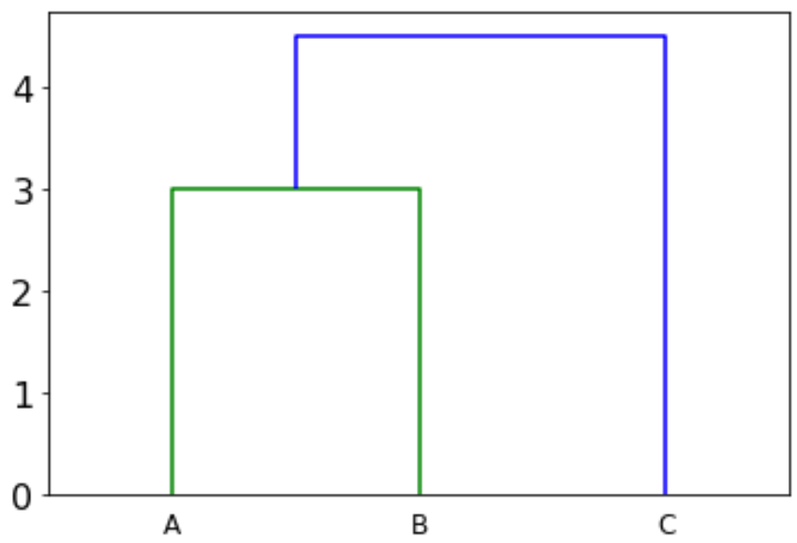


demo 2

linkage matrix

	name	left	right	distance	nodes
0	A			0	1
1	B			0	1
2	C			0	1
3	cluster1	0	1	3	3
4	cluster0	3	2	4.5	2

demo 1



dendrogram

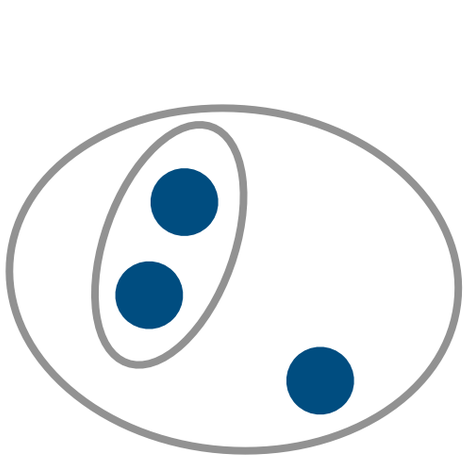
Strategy: Combine Nearby Points/Groups (and repeat!)

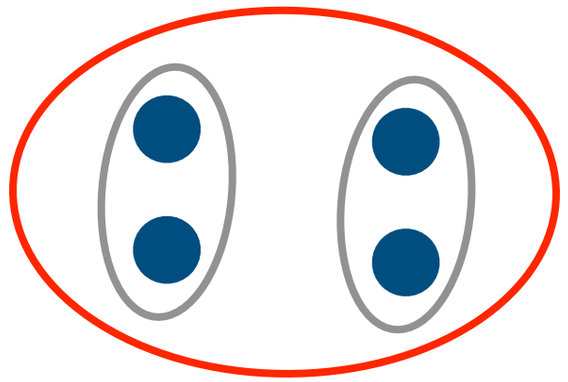
Strategy: Combine Nearby Points/Groups (and repeat!)

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

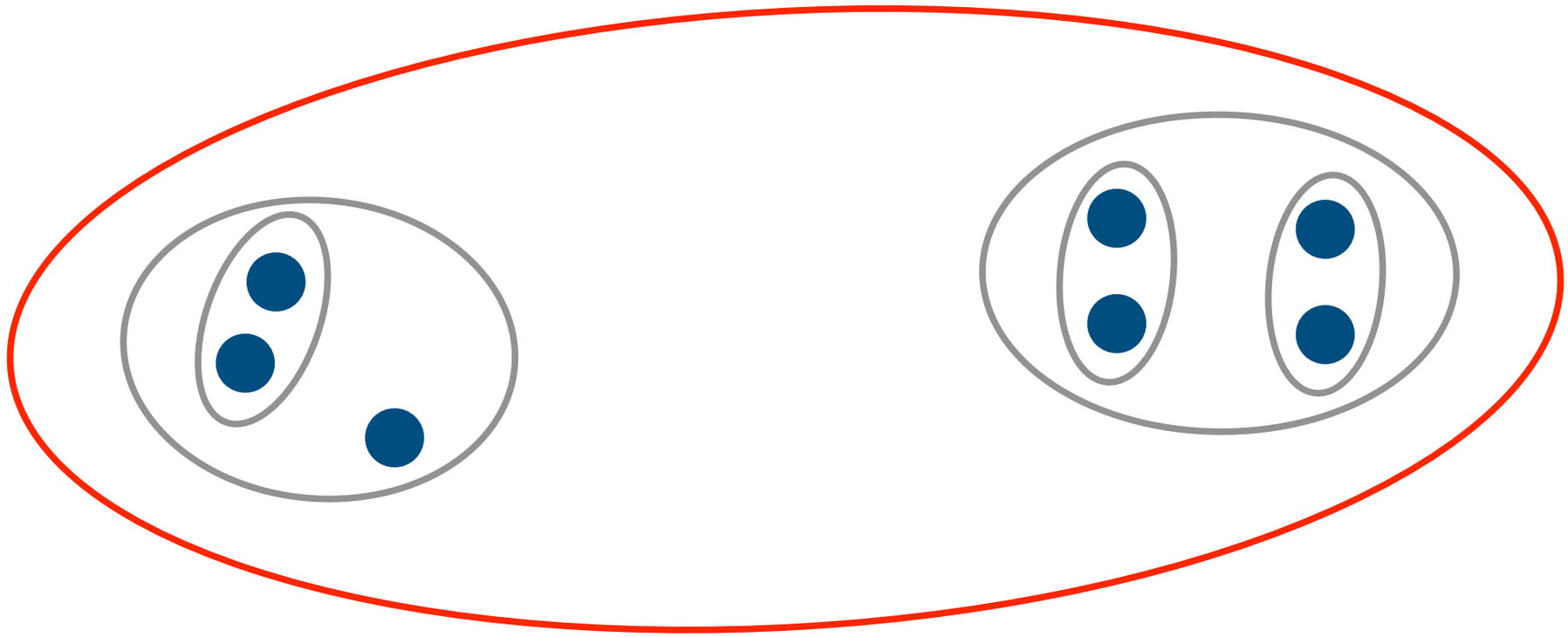
Strategy: Combine Nearby Points/Groups (and repeat!)

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





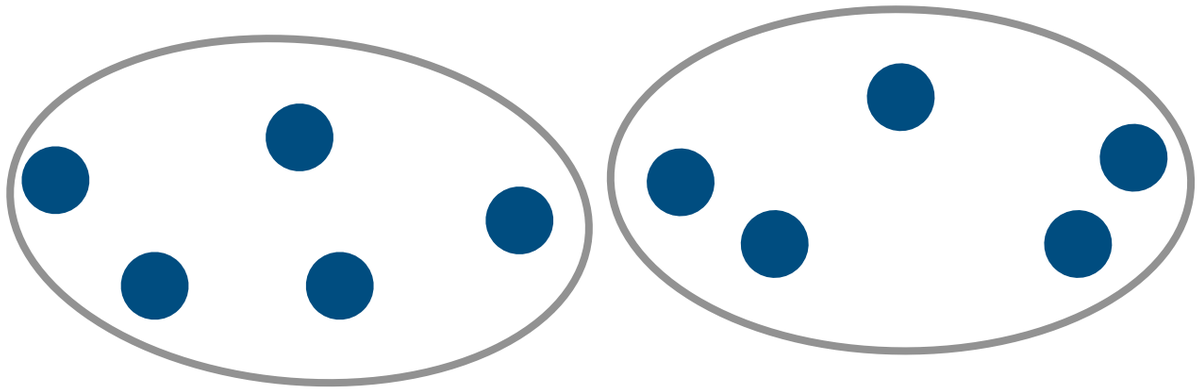
Strategy: Combine Nearby Points/Groups (and repeat!)



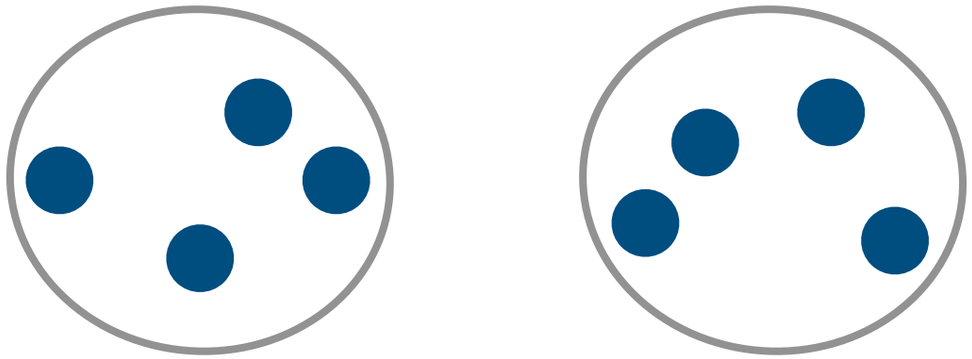
Configuration: what is "nearest"?

option: `linkage`

Configuration: what is "nearest"?

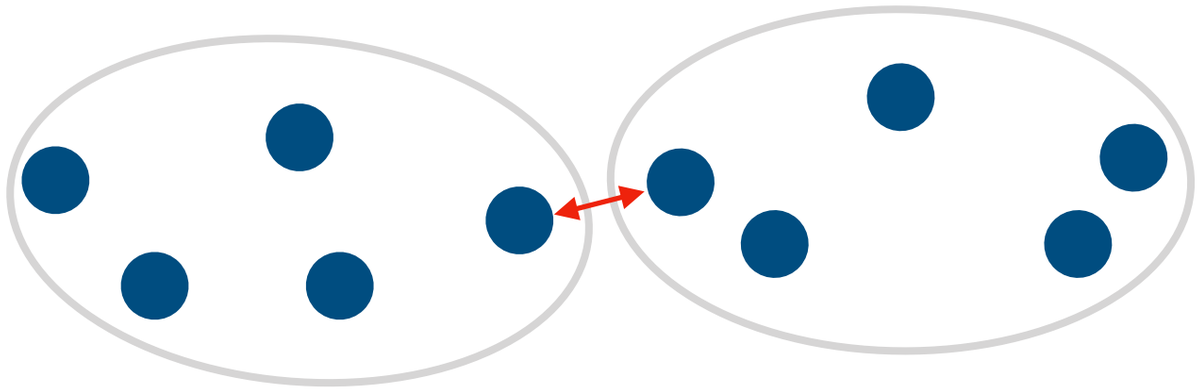


OR...

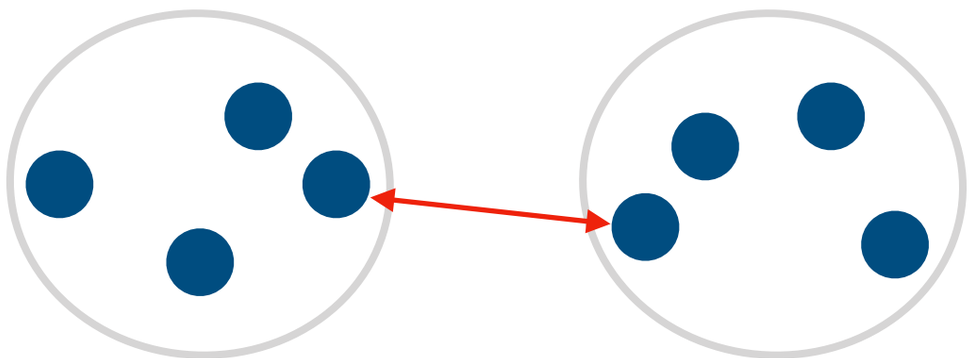


Configuration: what is "nearest"?

linkage="single"

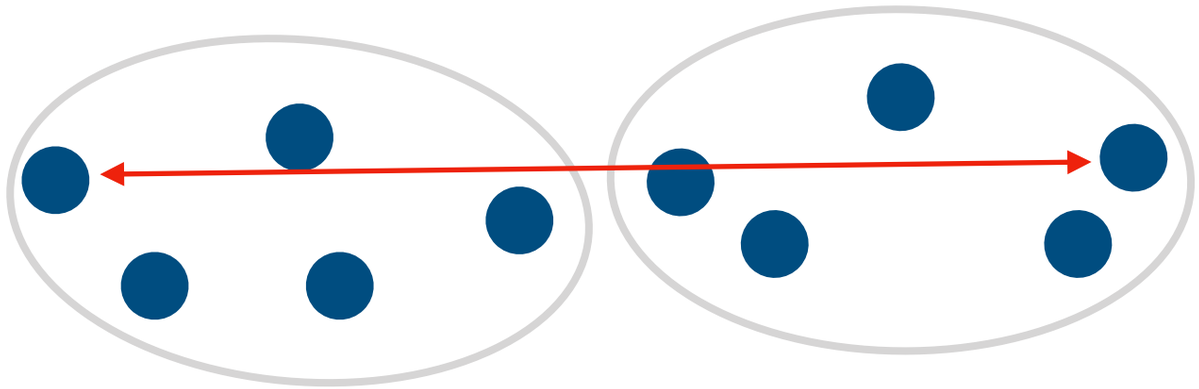


OR...

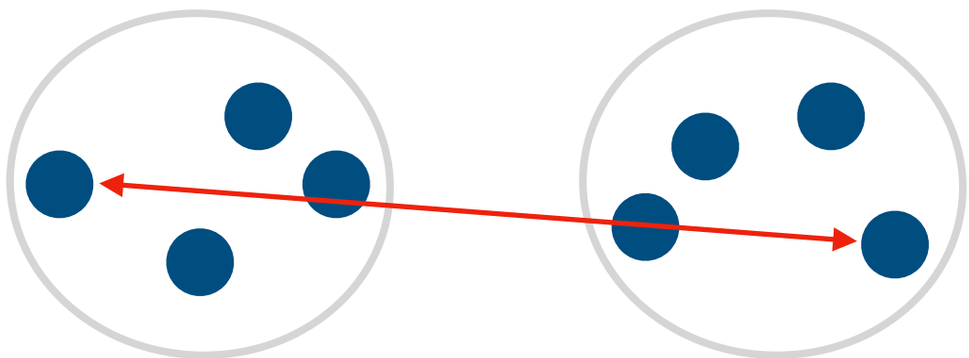


Configuration: what is "nearest"?

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.

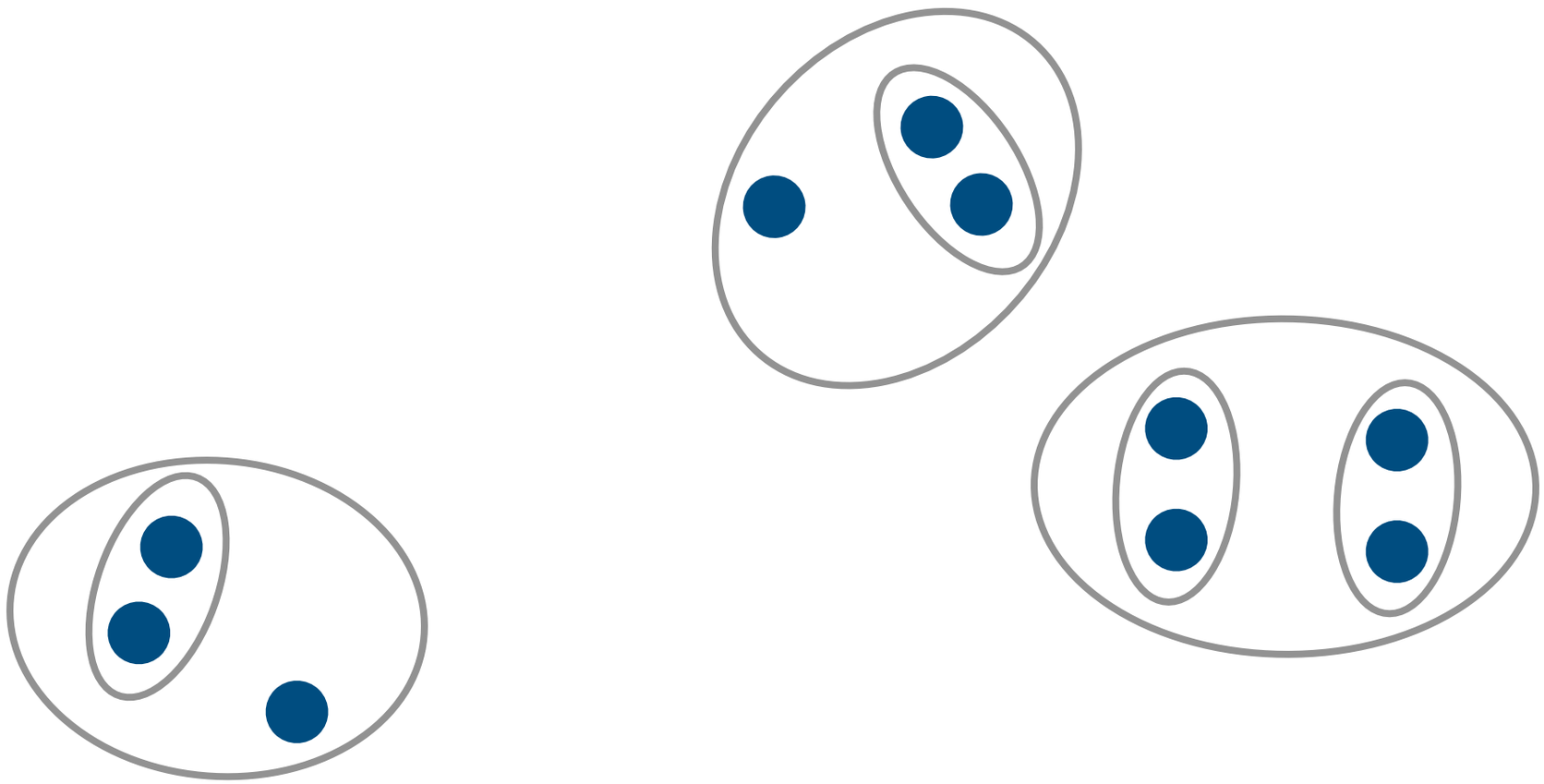
sklearn: AgglomerativeClustering

Configuration: when to stop?

option: `n_clusters` or `distance_threshold`

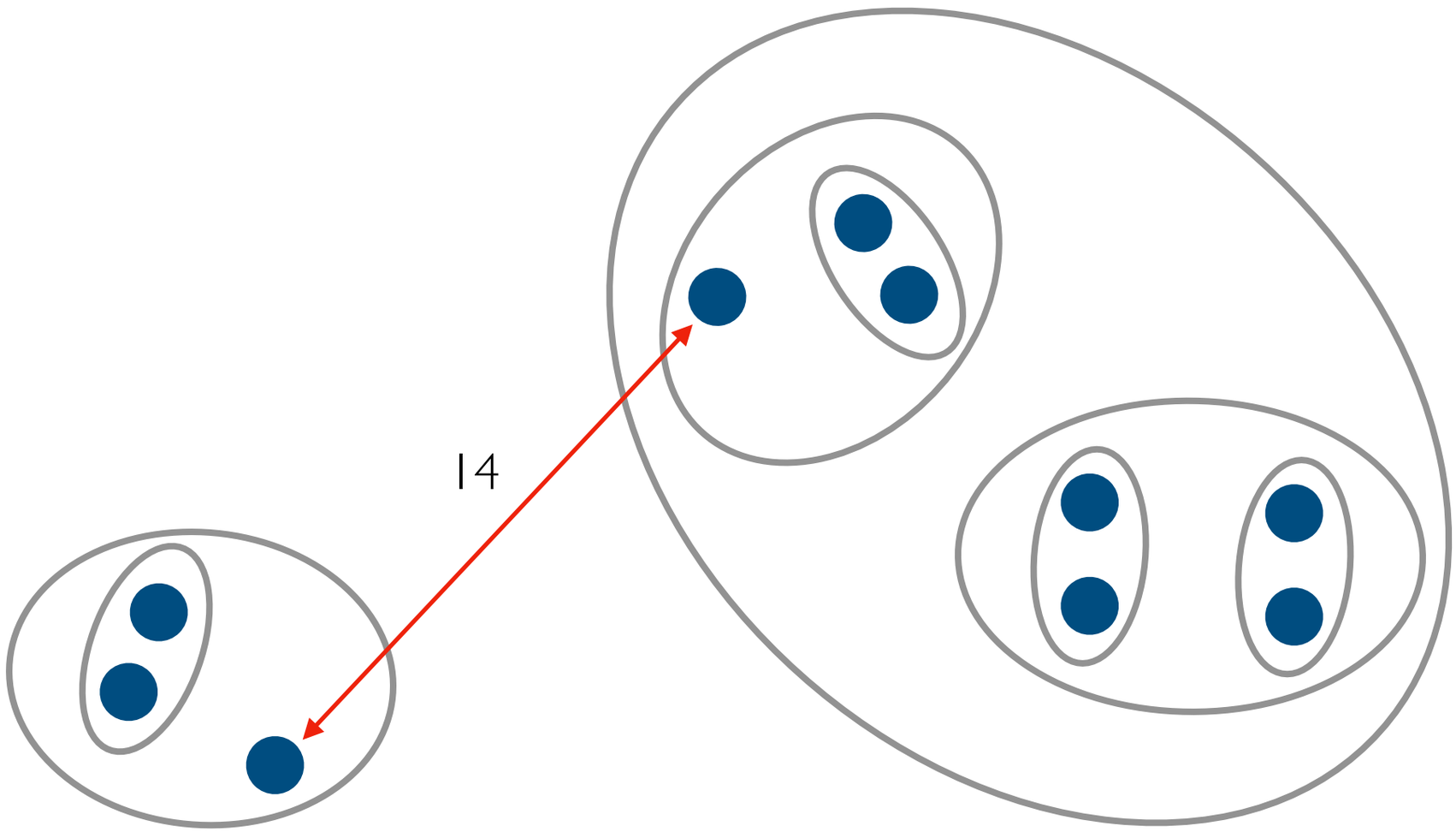
Configuration: when to stop?

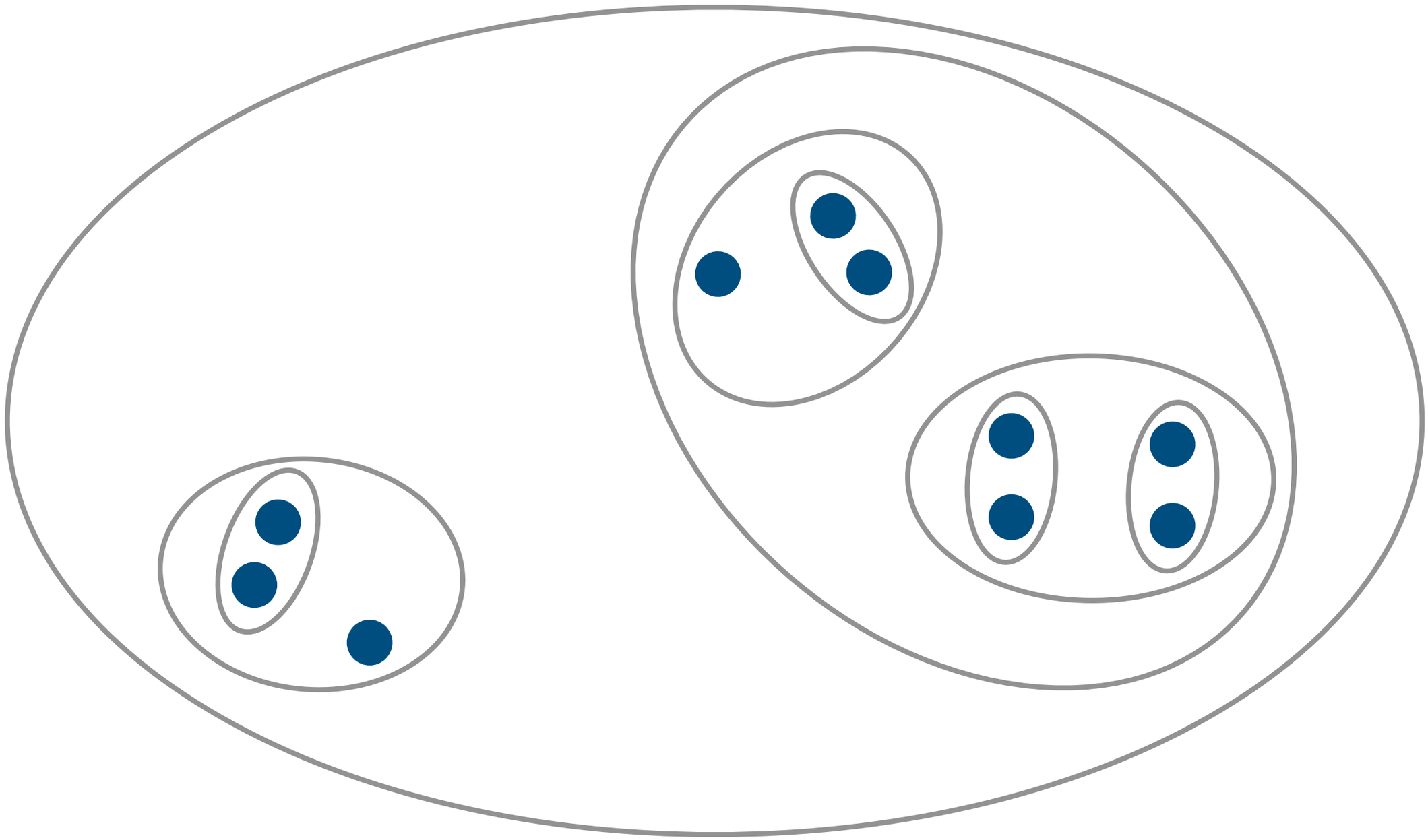
n_clusters=3



Configuration: when to stop?

distance_threshold=10



Configuration: when to stop?

distance_threshold=0