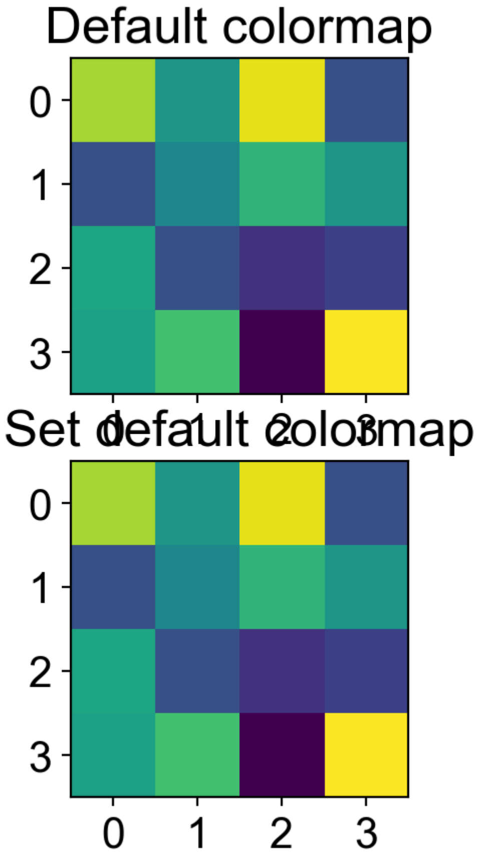
CPSC 330 Lecture 16: DBSCAN, Hierarchical Clustering

Announcements

- HW6 is due next week Wednesday.
  - Computationally intensive (welcome to real-life ML!)
  - heads up: you will need to install many packages (welcome to real-life ML!)


Default colormap
0
1
2
3
0 1 2 3
Set default colormap
0
1
2
3
0 1 2 3

Super cool Demo!

You can download and checkout the Demo here.

All credit to Dr. Varada Kolhatkar for putting this together!

Recap: iClicker Exercise 15.3

Select all of the following statements which are **True**

- a. If you train K-Means with `n_clusters`= the number of examples, the inertia value will be 0.
- b. The elbow plot shows the tradeoff between within cluster distance and the number of clusters.
- c. Unlike the Elbow method, the Silhouette method is not dependent on the notion of cluster centers.
- d. The elbow plot is not a reliable method to obtain the optimal number of clusters in all cases.
- e. The Silhouette scores ranges between -1 and 1 where higher scores indicates better cluster assignments.

iClicker Exercise 16.1

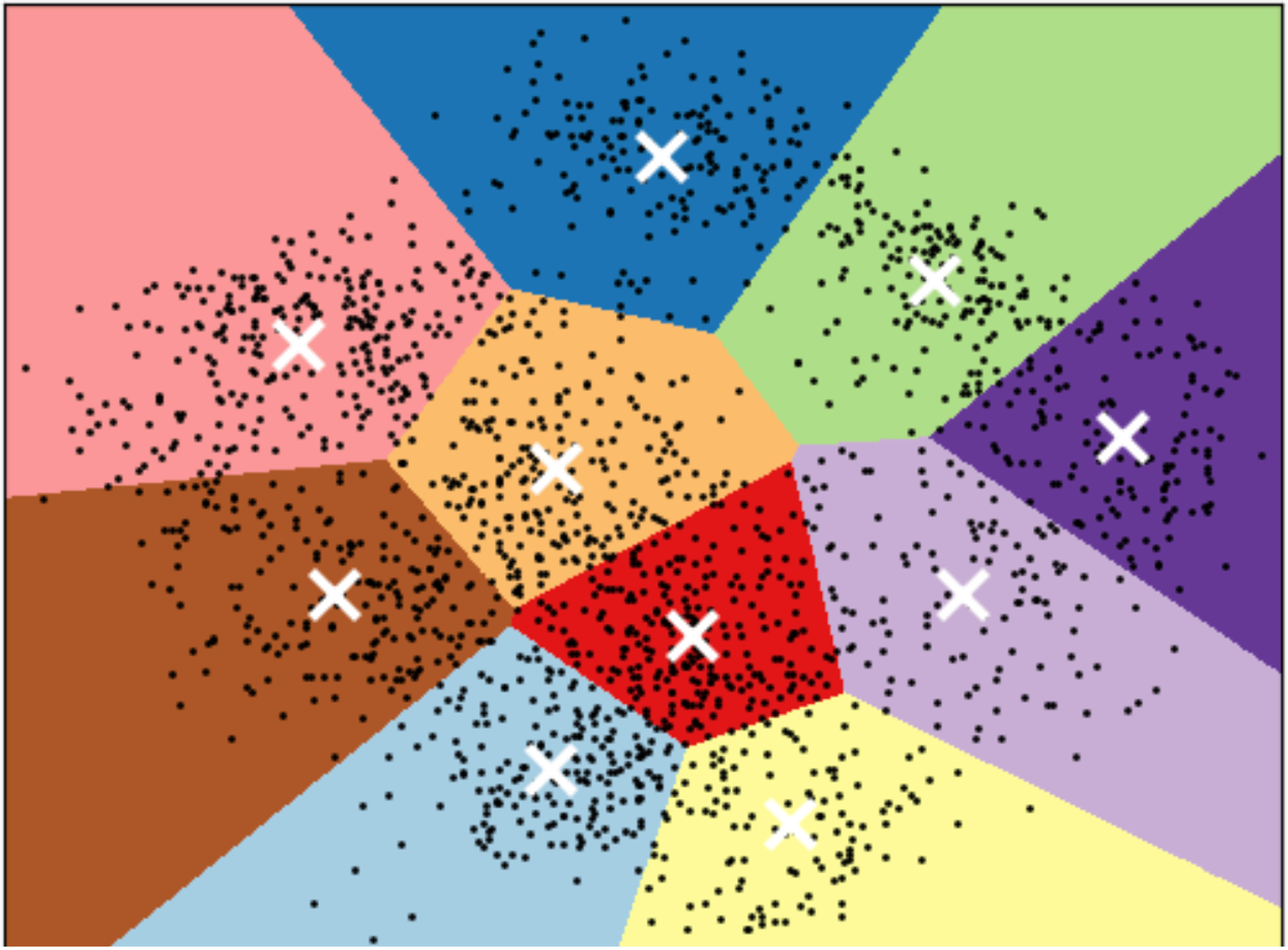
Select all of the following statements which are TRUE.

- a. Similar to K-nearest neighbours, K-Means is a non parametric model.
- b. The meaning of K in K-nearest neighbours and K-Means clustering is very similar.
- c. Scaling of input features is crucial in clustering.
- d. In clustering, it's almost always a good idea to find equal-sized clusters.

Limitations of K-means

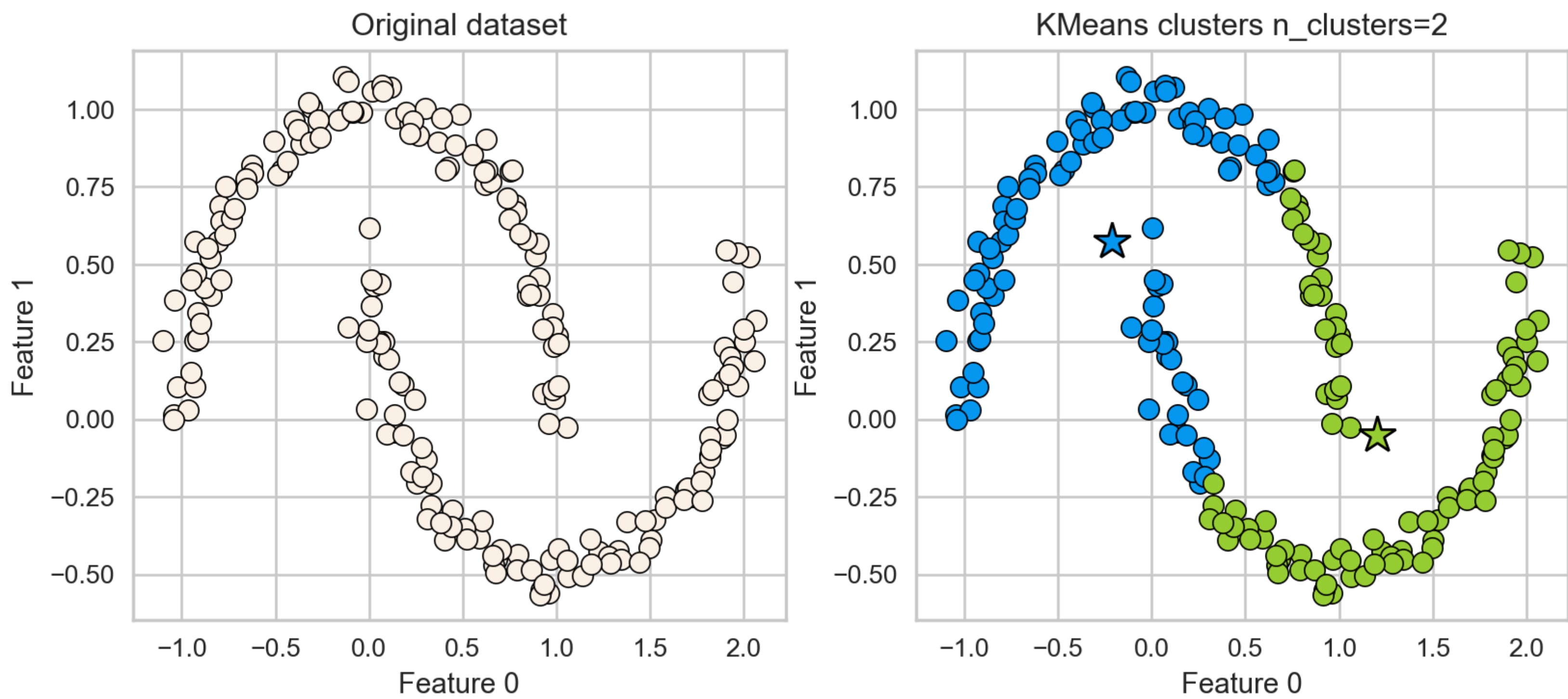
Shape of clusters

- Good for spherical clusters of more or less equal sizes



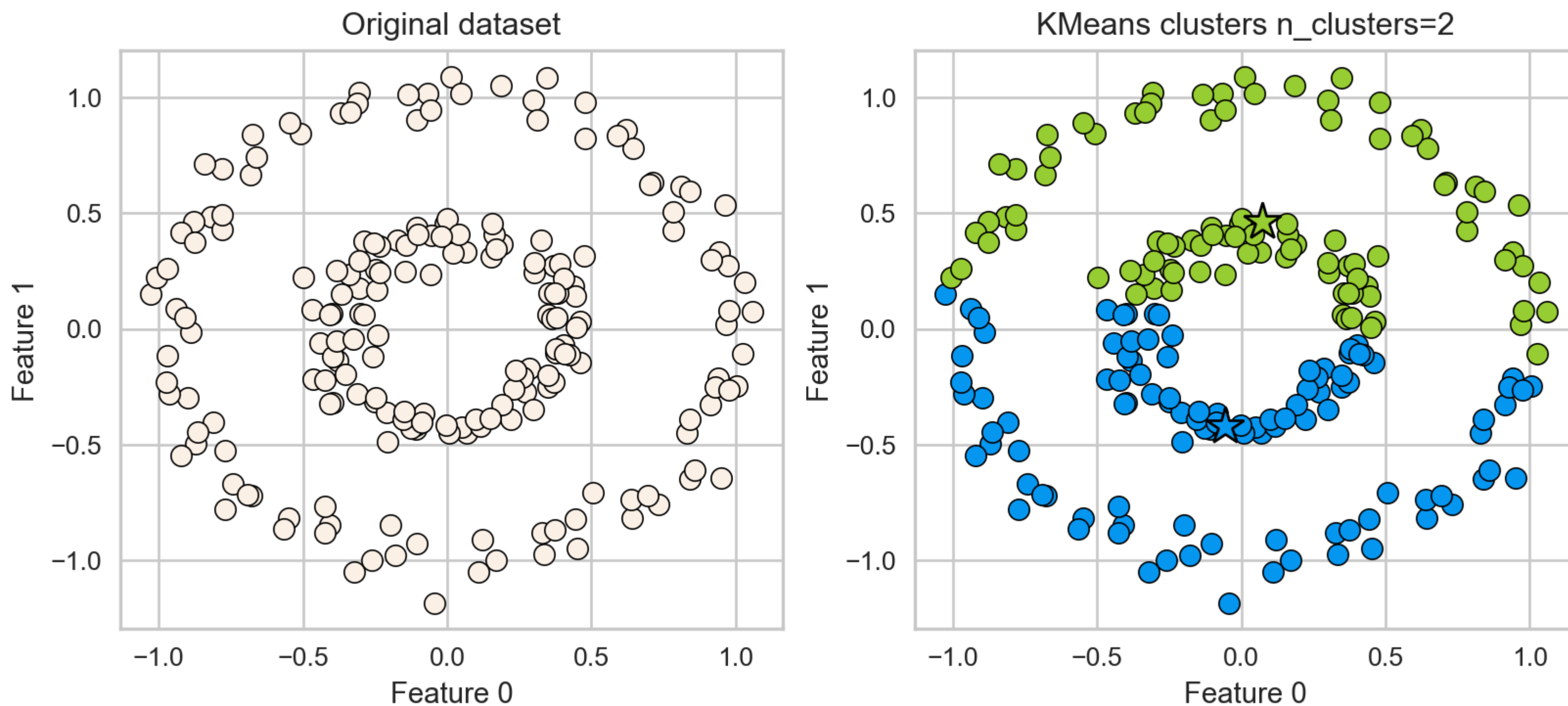
K-Means: failure case 1

- K-Means performs poorly if the clusters have more complex shapes (e.g., two moons data below).



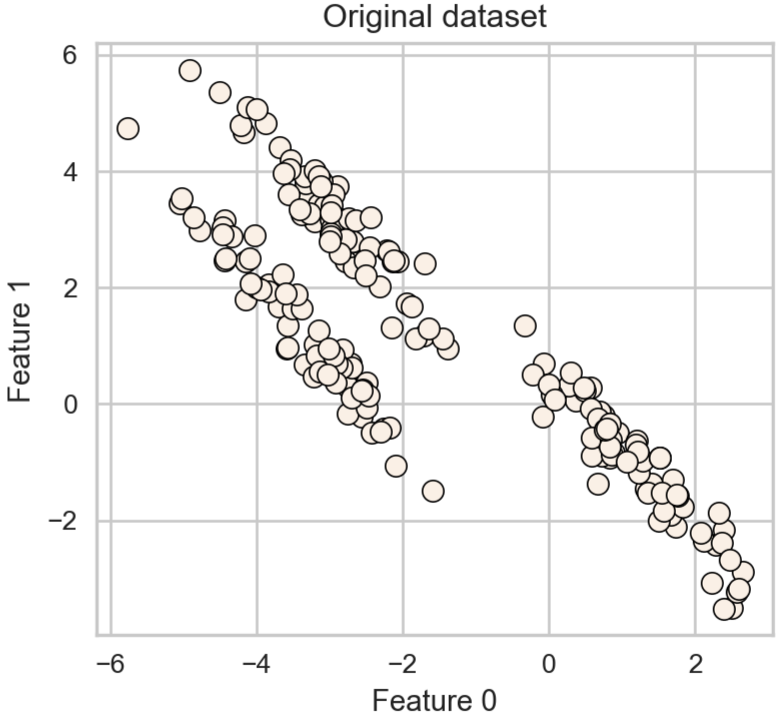
K-Means: failure case 2

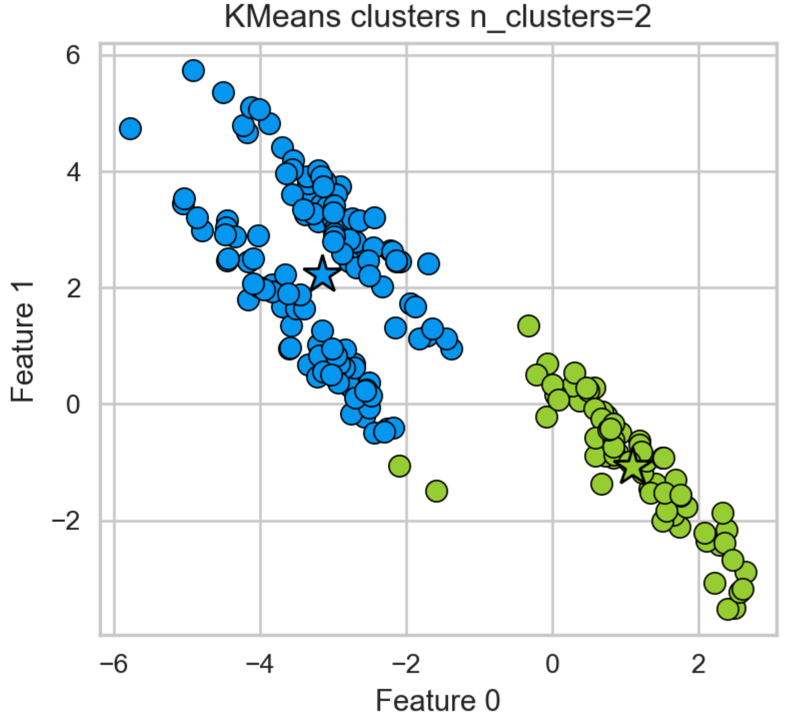
- Again, K-Means is unable to capture complex cluster shapes.



K-Means: failure case 3

- It assumes that all directions are equally important for each cluster and fails to identify non-spherical clusters.





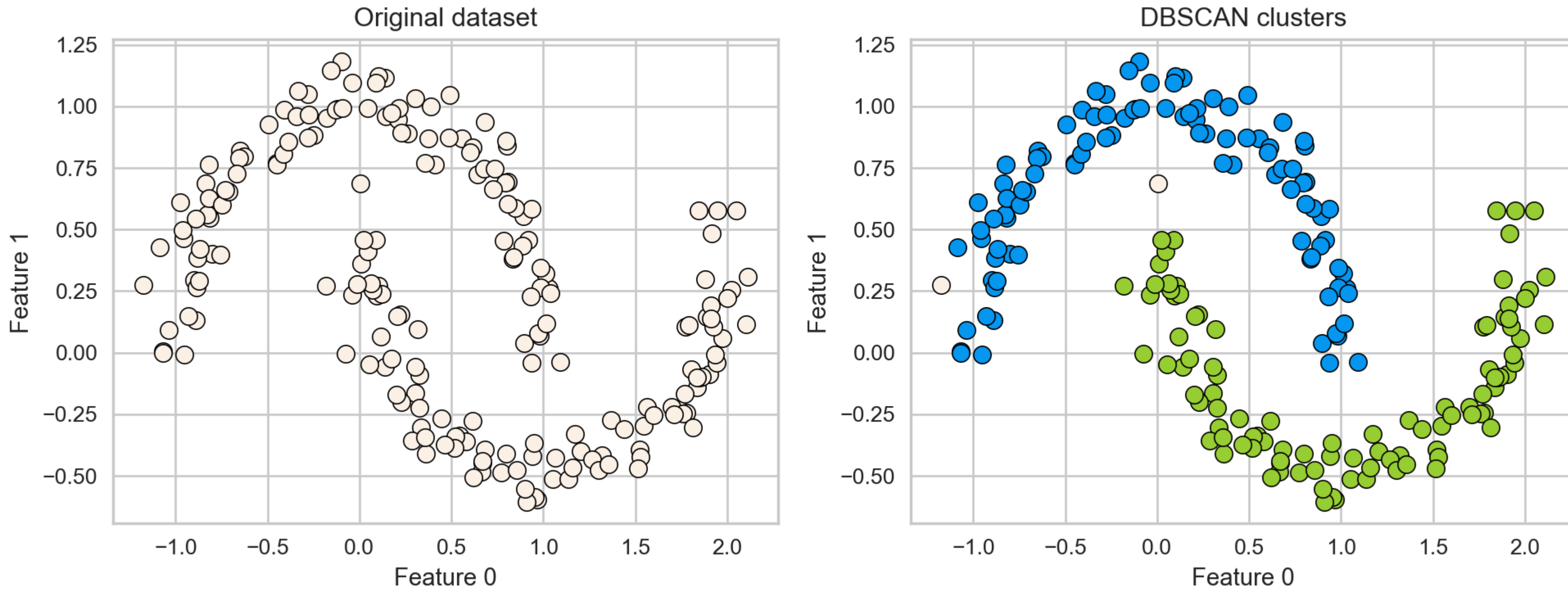
Can we do better?

DBSCAN

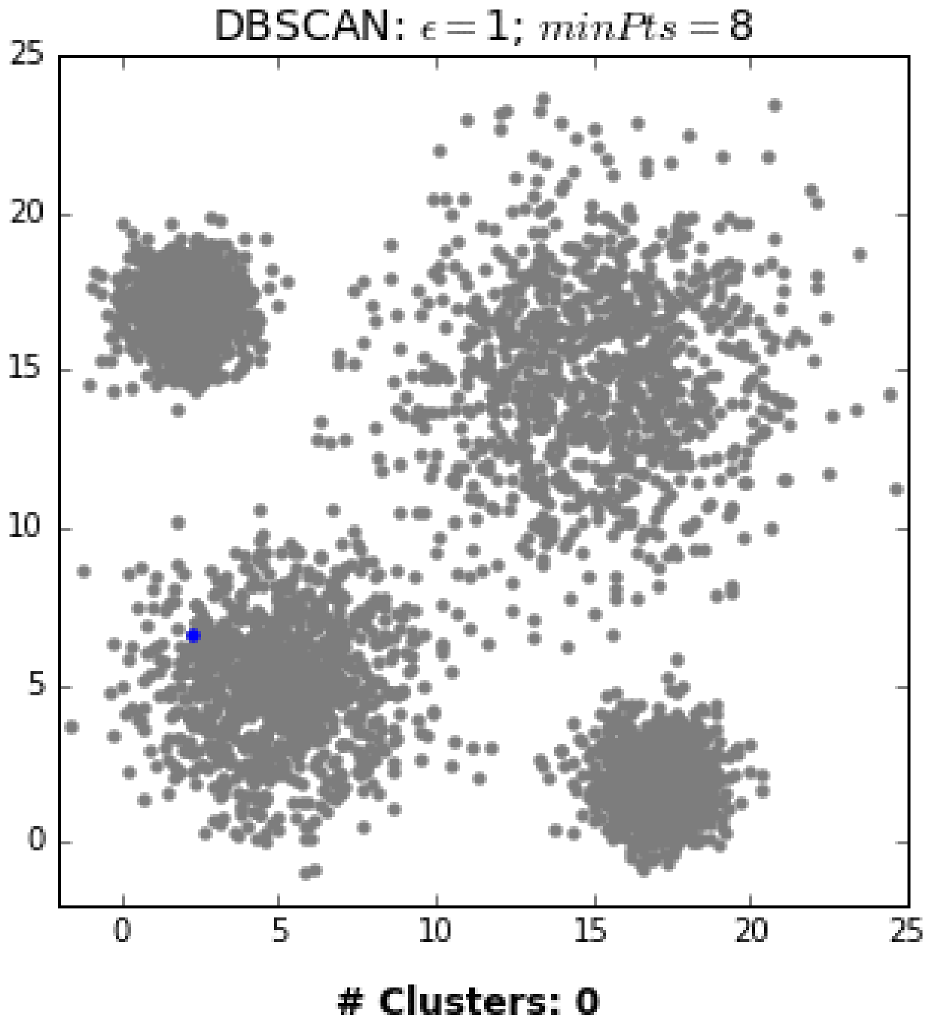
- **D**ensity-Based **S**patial **C**lustering of **A**pplications with **N**oise
- A density-based clustering algorithm

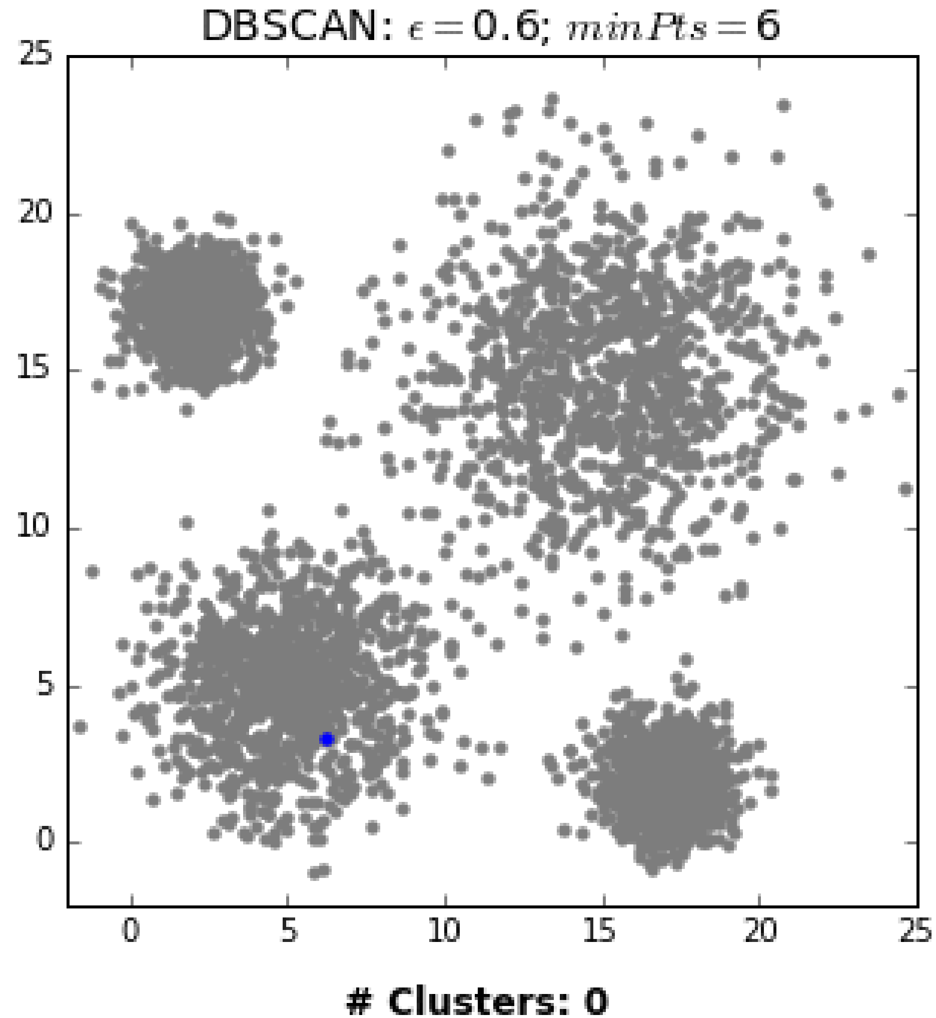
DBSCAN

```
X, y = make_moons(n_samples=200, noise=0.08, random_state=42)
dbscan = DBSCAN(eps=0.2)
dbscan.fit(X)
plot_original_clustered(X, dbscan, dbscan.labels_)
```



How does it work?





DBSCAN Analogy

Consider DBSCAN in a social context:

- Social butterflies (🦋): Core points
- Friends of social butterflies who are not social butterflies: Border points
- Lone wolves (🐺): Noise points

Two main hyperparameters

- `eps`: determines what it means for points to be “close”
- `min_samples`: determines the number of **neighboring points** we require to consider in order for a point to be part of a cluster

DBSCAN: failure cases

- Let's consider this dataset with three clusters of varying densities.
- K-Means performs better compared to DBSCAN. But it has the benefit of knowing the value of K in advance.

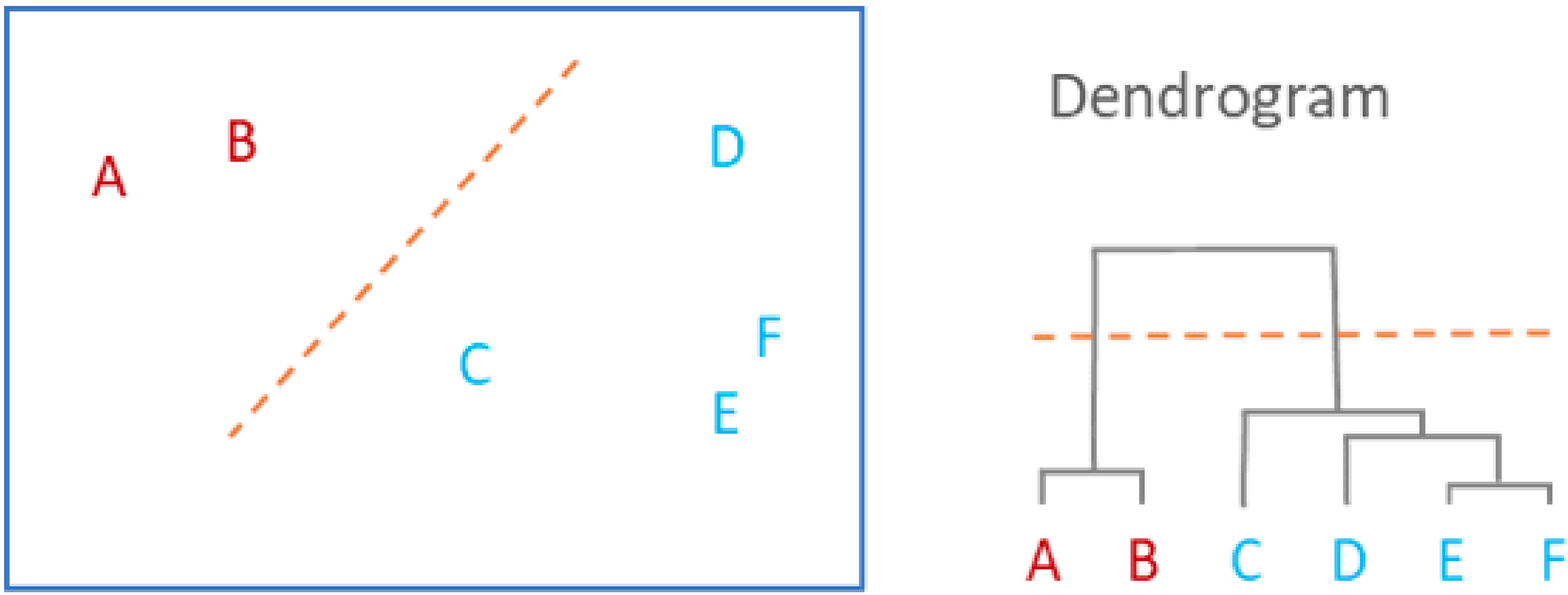
```
[ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15]
```

Break

Let’s take a 10-min break!

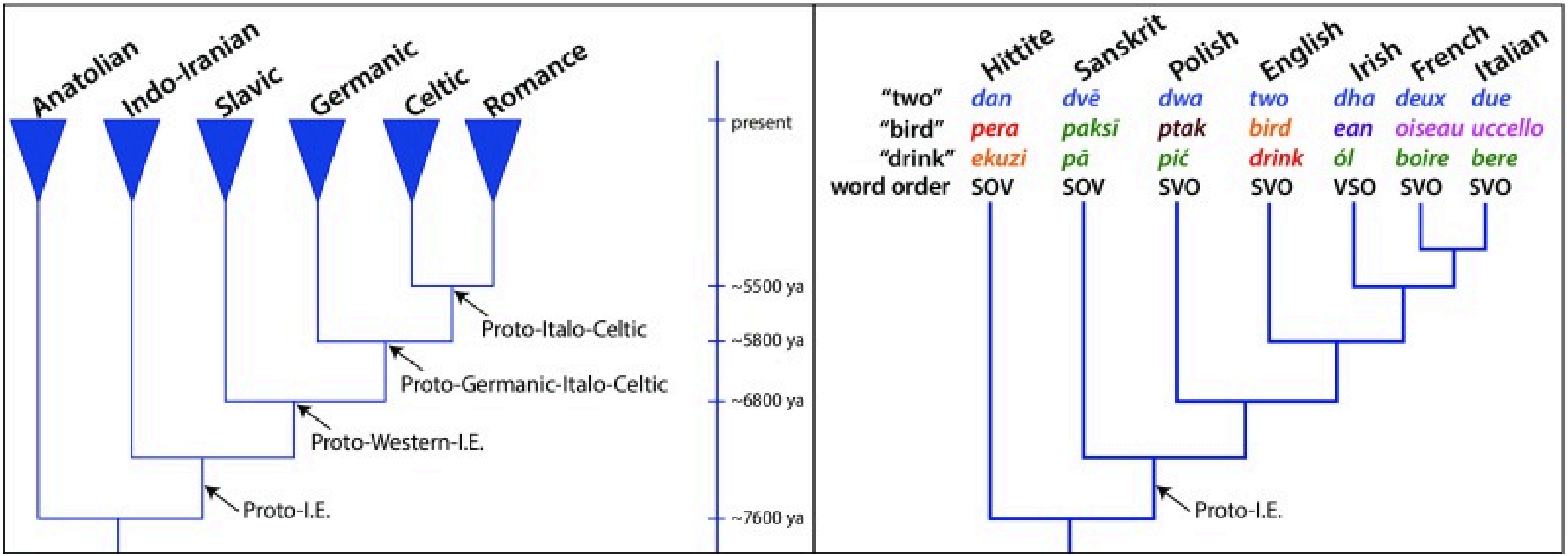
Dendrogram

Definition: visual representation of a tree, in particular, the hierarchical representation of data…



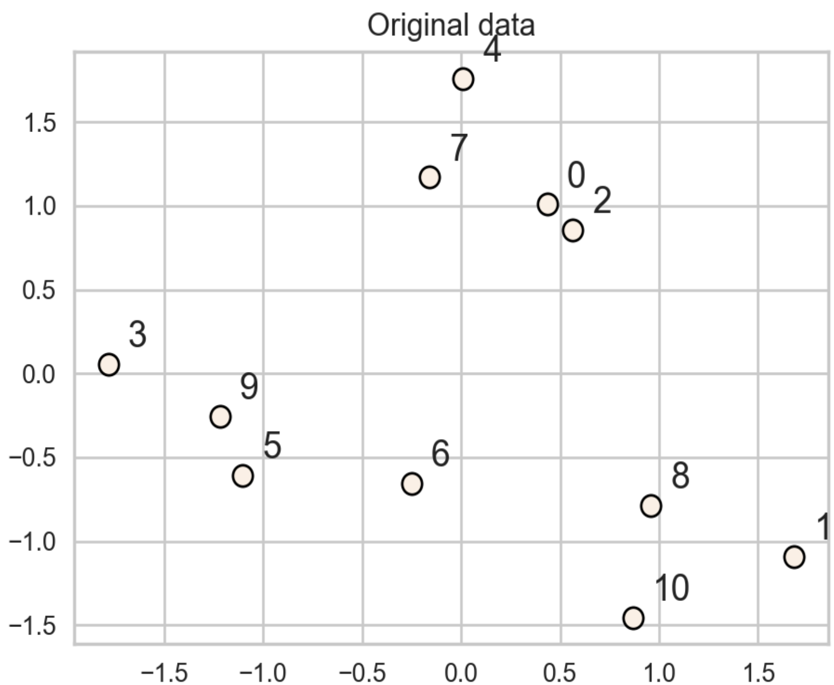
Source

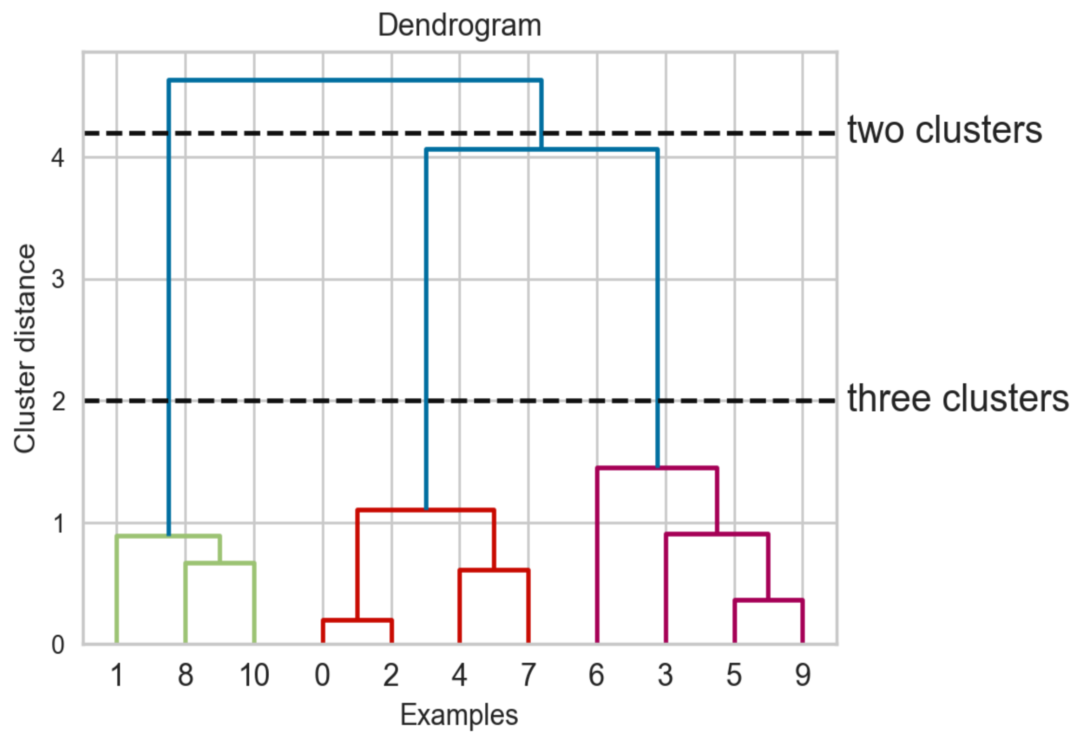
Example: Languages



Source

Hierarchical clustering





Flat clusters

- This is good but how can we get cluster labels from a dendrogram?
- We can bring the clustering to a “flat” format use `fcluster`

Linkage criteria

- When we create a dendrogram, we need to calculate distance between clusters. How do we measure distances between clusters?
- The **linkage criteria** determines how to find similarity between clusters:
- Some example linkage criteria are:
  - Single linkage → smallest minimal distance, leads to loose clusters
  - Complete linkage → smallest maximum distance, leads to tight clusters
  - Average linkage → smallest average distance between all pairs of points in the clusters
  - Ward linkage → smallest increase in within-cluster variance, leads to equally sized clusters

Activity

- Fill in the table below

Clustering Method	KMeans	DBSCAN	Hierarchical Clustering
Approach			
Hyperparameters			
Shape of clusters			
Handling noise			
Examples			