Basic Machine Learning: Clustering

Ketentuan Presentasi

Ketentuan Presentasi

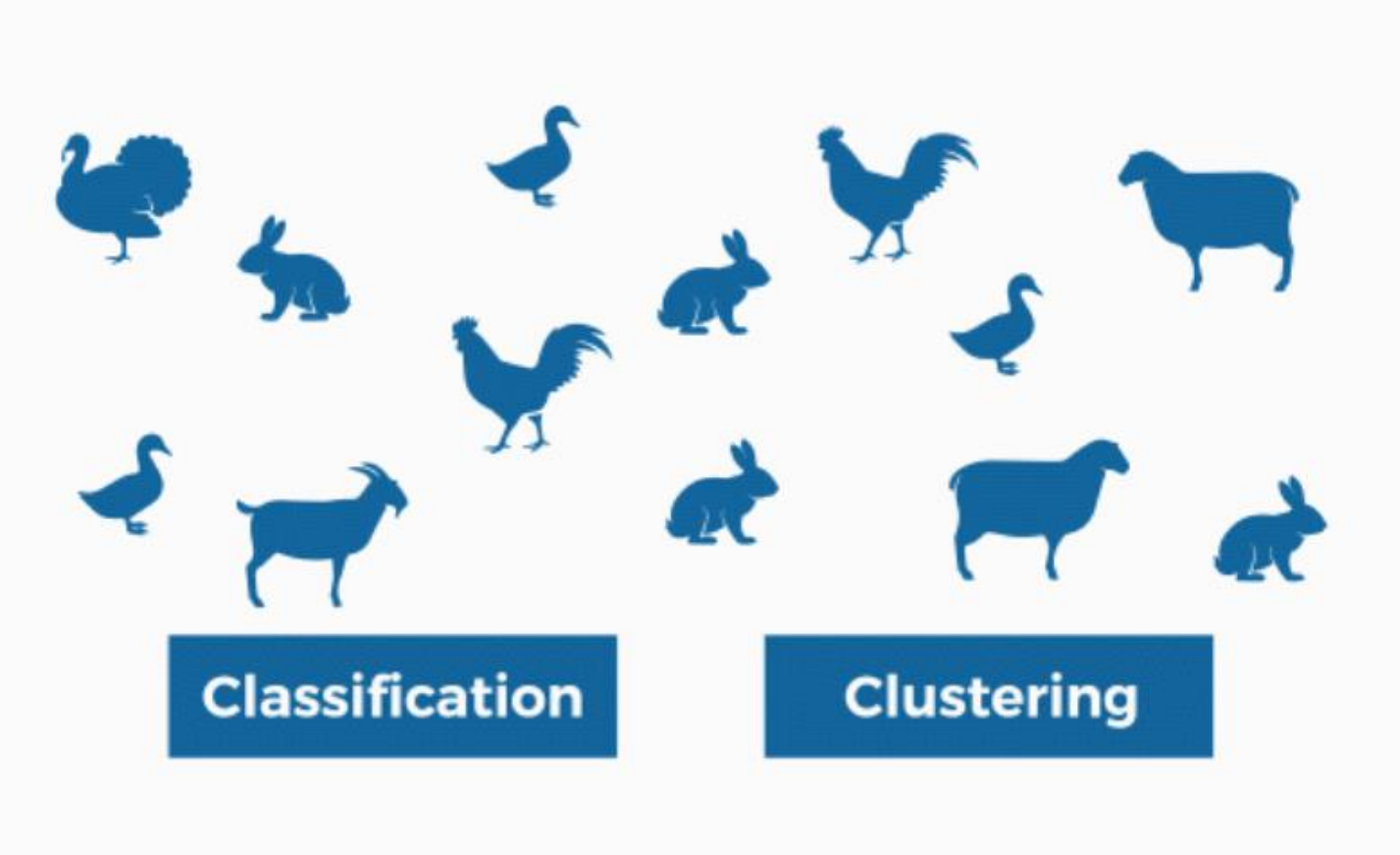
Berikut beberapa ketentuan presentasi mengenai hasil dari pengerjaan *Final Project* yang ada.

Ketentuan

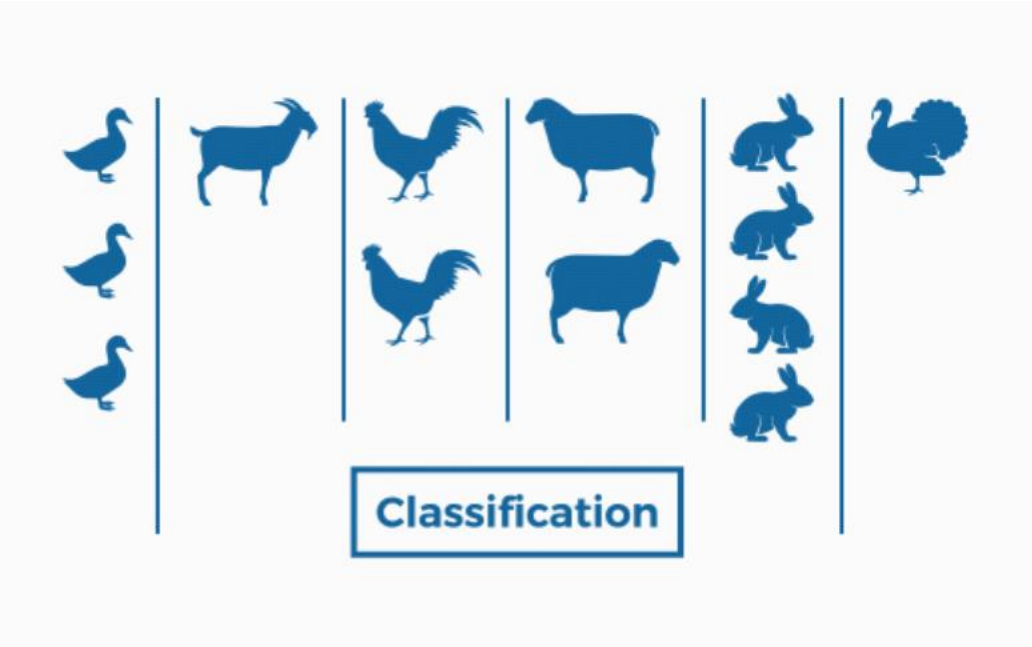
- 30 menit sebelum pertemuan dimulai masing-masing team bisa mengirim slide presentasinya di group Telegram yang ada
- 25 menit waktu yang tersedia untuk masing-masing team mempresentasikan hasil dari *Final Project* yang ada
- 15 menit waktu yang tersedia untuk sesi tanya jawab
- Setiap team bisa secara bebas menentukan berapa banyak presentator yang dibutuhkan
- Diharapkan masing-masing team bisa terlibat secara aktif selama pertemuan berlangsung

Content

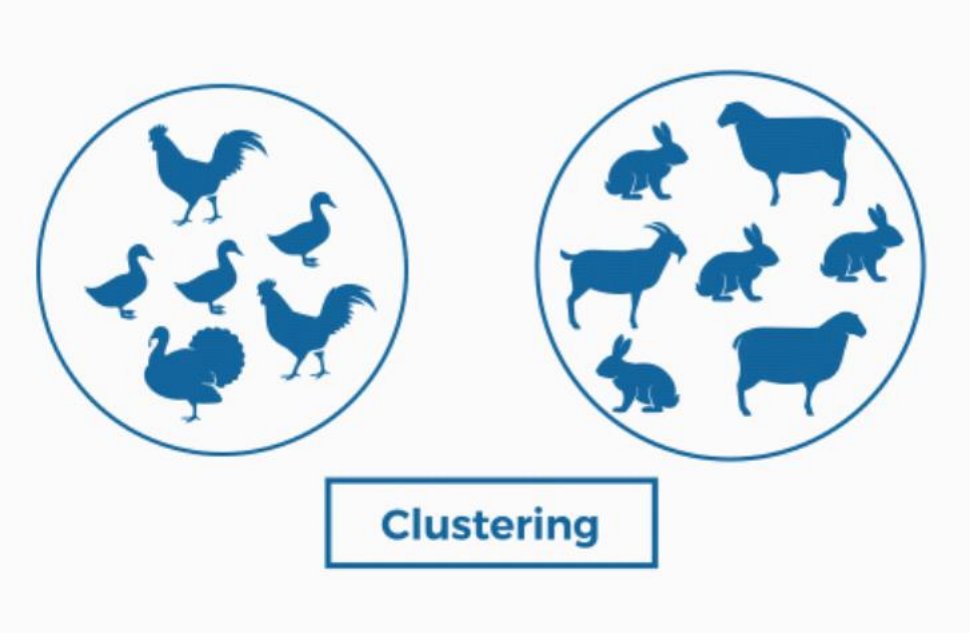
Clustering vs Classification



Clustering vs Classification

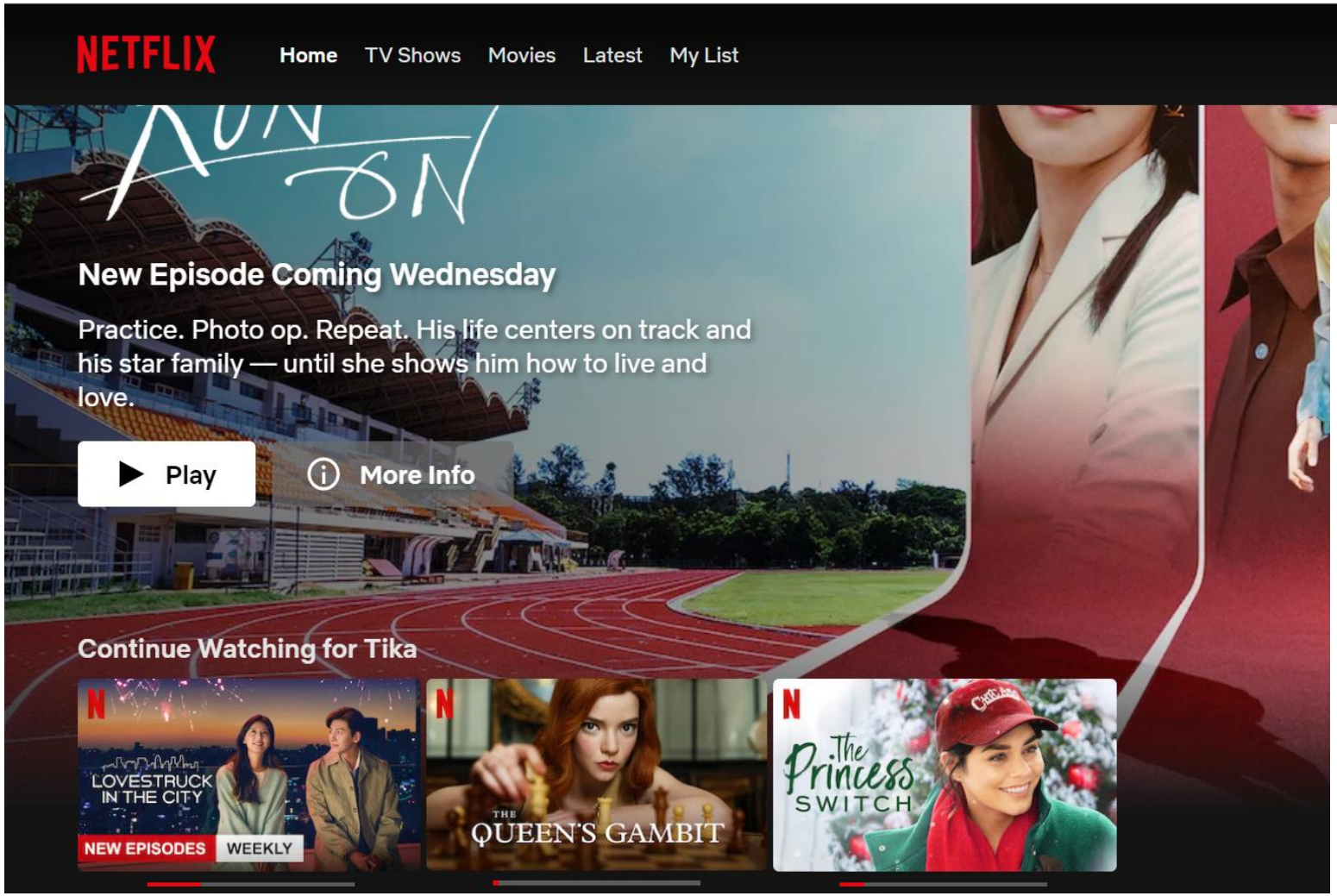


- Supervised Learning
- Pre-defined Label



- Unsupervised Learning
- Not Labelled

Clustering (Concept)



Did You Know?

Netflix began using analytic tools in 2000 to recommend videos for users to rent.

Netflix just has a 90-second window to help viewers find a movie or a TV show before they leave the platform and visit some other service. That's one of the major reasons why Netflix is so obsessed with personalizing recommendations to hook users.

Netflix's personalized recommendation algorithms produce $1 billion a year in value from customer retention.

Majority of Netflix users consider recommendations with 80% of Netflix views coming from the service's recommendations.

Netflix has set up 1300 recommendation clusters based on users viewing preferences.

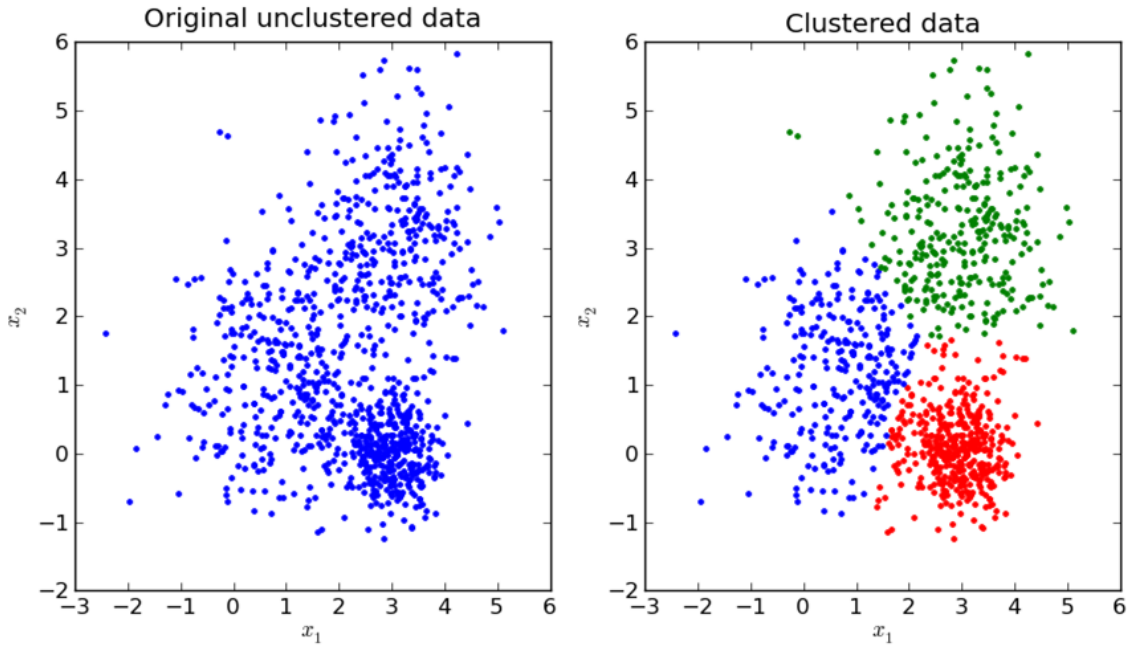
Netflix segments its viewers into over 2K taste groups. Based on the taste group a viewer falls, it dictates the recommendations.

With over 7K TV shows and movies in the catalogue, it is actually impossible for a viewer to find movies they like to watch on their own. Netflix's recommendation engine automates this search process for its users.

Clustering (Concept)

Clustering (or cluster analysis) is a technique that allows us to find groups of similar objects, objects that are more related to each other than to objects in other groups.

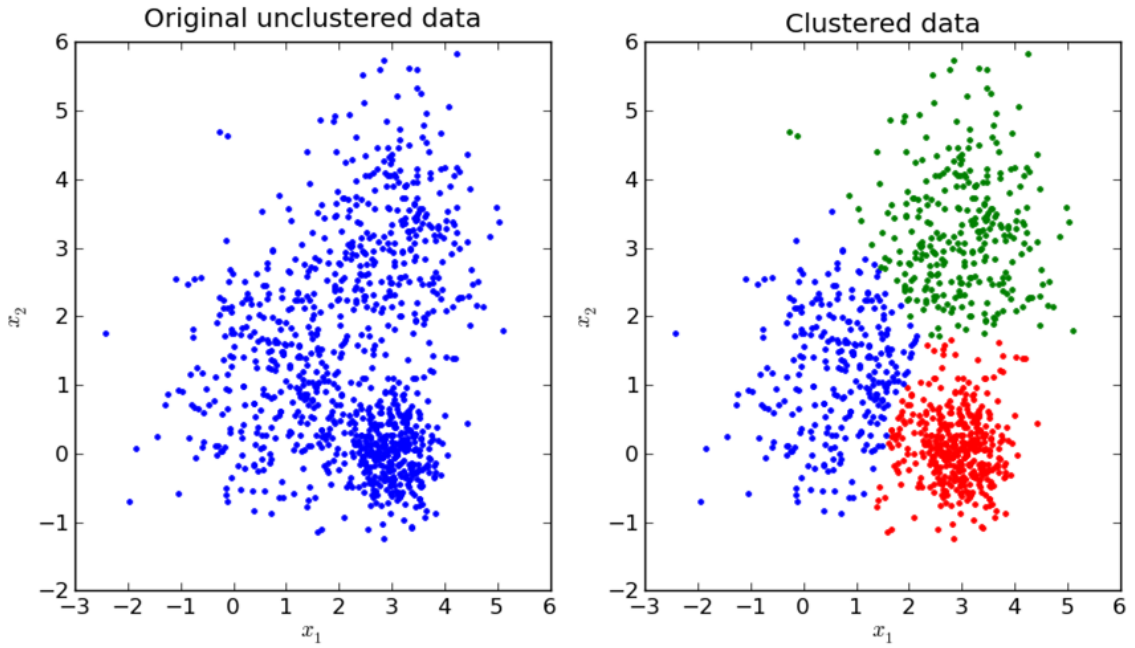
Examples of applications include the grouping of documents, music, and movies by different topics, or finding customers that share similar interests based on common purchase behaviours as a basis for **recommendation engines**.



Clustering (Concept)

The top 3 popular clustering algorithms which is widely used in academia as well as in industry will be discussed as follows:

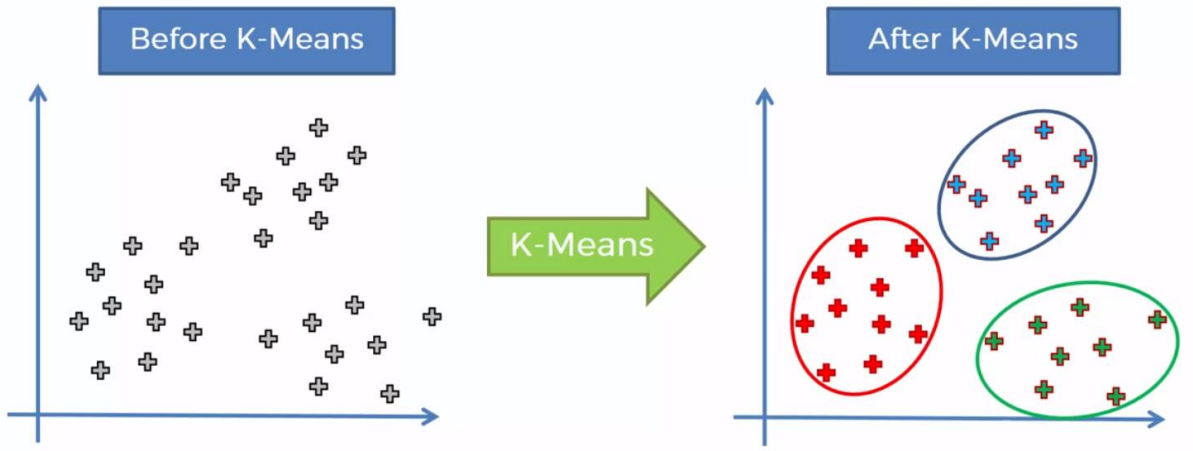
1. K-Means Clustering
2. Hierarchical Clustering
3. DBSCAN



K-Means Clustering

K-Means is probably the most well-known clustering algorithm. It's easy to understand and implement in code.

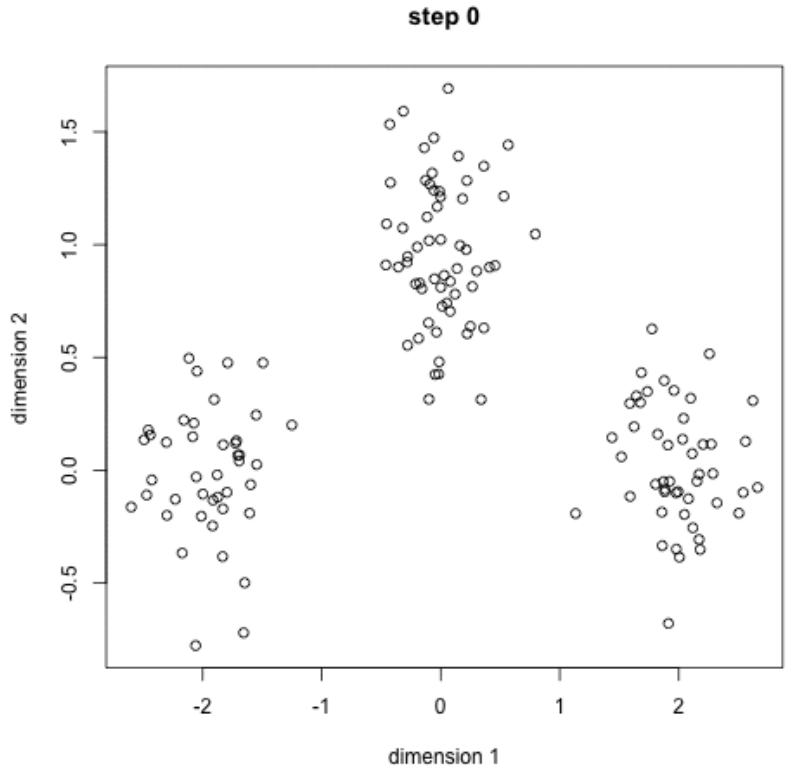
The **K-Means** algorithm clusters data by trying to separate samples in n groups of similar features.



K-Means Clustering

K-Means clustering is an unsupervised machine learning method that works with the following steps:

1. Choose number of clusters (K)
2. Select at random K points centroid
3. Assign each data point to the closest centroid that form K clusters
4. Recompute and place the new centroid of each cluster
5. Reassign each data points to the new closest centroid, if any reassignment took place do step 4 again, otherwise model is finished



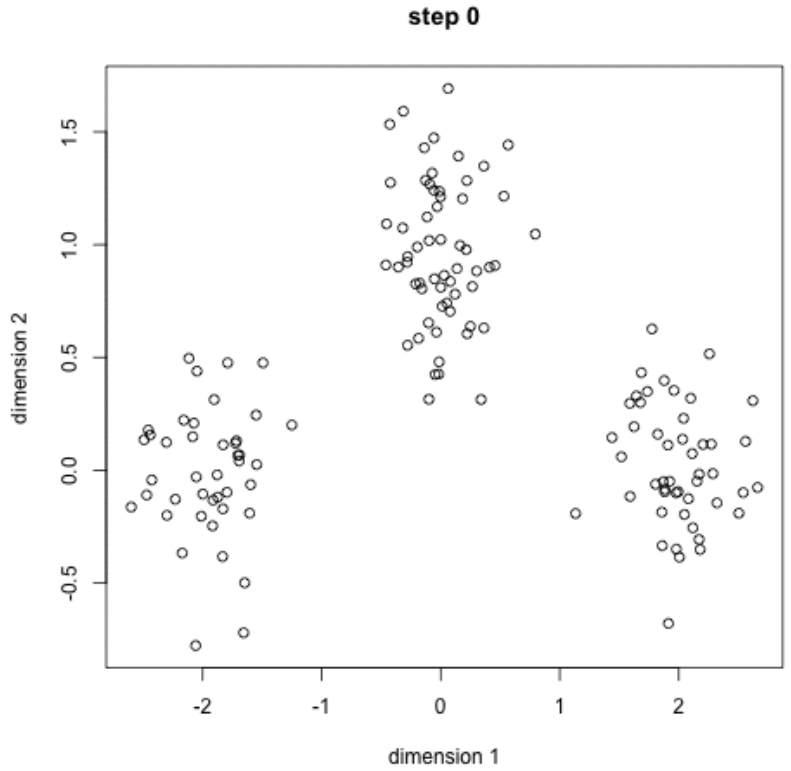
K-Means Clustering

Advantages:

- Faster process since the complexity of process is linear to the amount of data

Disadvantages:

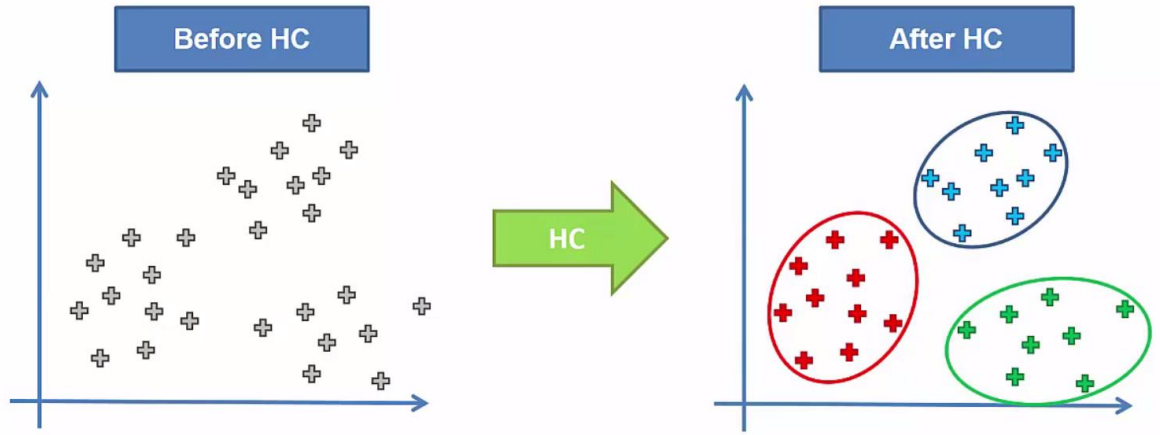
- Decide centroid number is not trivial
- Lack of consistency due to the initial random of centroid



Hierarchical Clustering

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is **represented as a tree** (or dendrogram).

The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

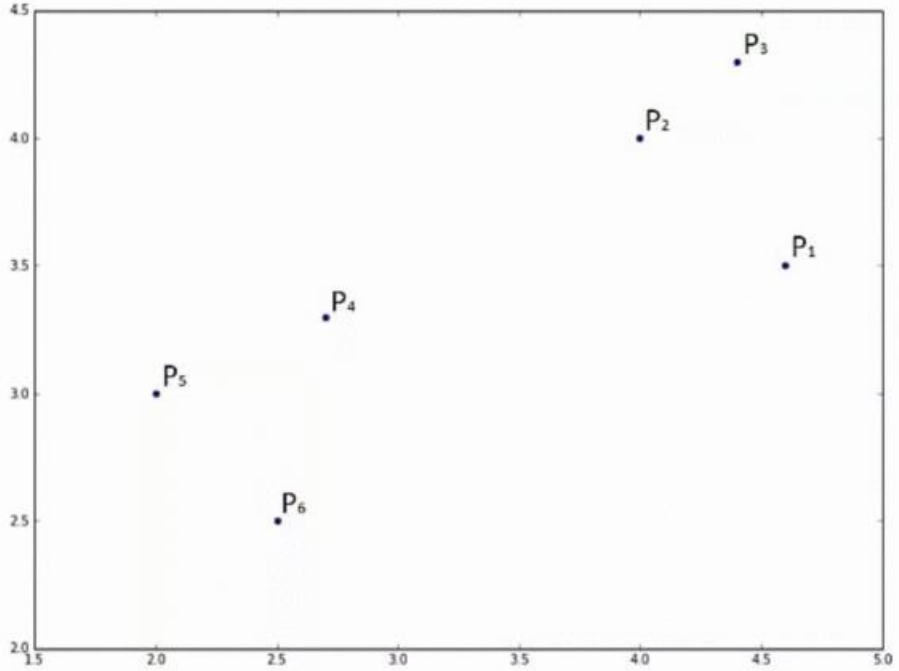


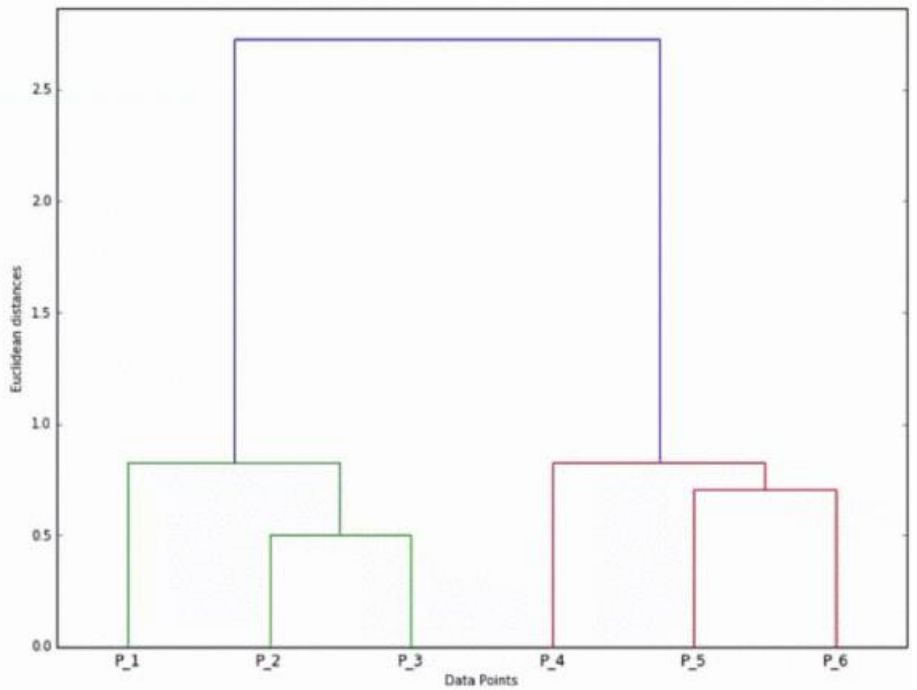
Hierarchical Clustering

Hierarchical Clustering is also an unsupervised machine learning method that works with the following steps:

1. Make each data points a single point cluster (That forms N clusters)
2. Take the two closest data points ad make the one cluster (That forms N-1 cluster)
3. Take the two closest and make them one cluster (That forms N-2 clusters)
4. Repeat step 3 until there is only one cluster

Hierarchical Clustering





Hierarchical Clustering

Advantages:

- Does not require us to specify the number of clusters
- We can even select which number of clusters looks best based on the resulting tree
- Good use case of hierarchical clustering methods is when the underlying data has a hierarchical structure

Disadvantages:

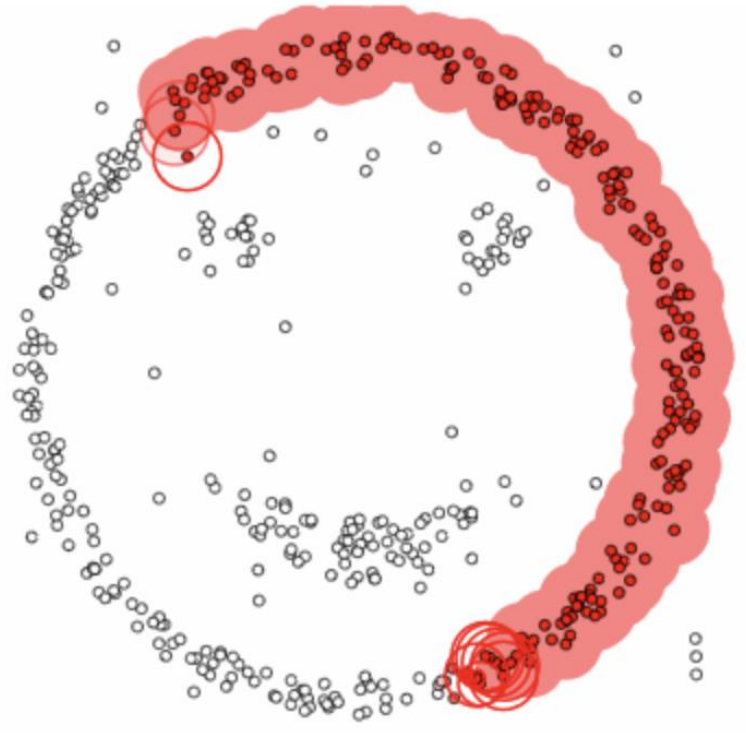
- Lower efficiency as a cost of computation

DBSCAN

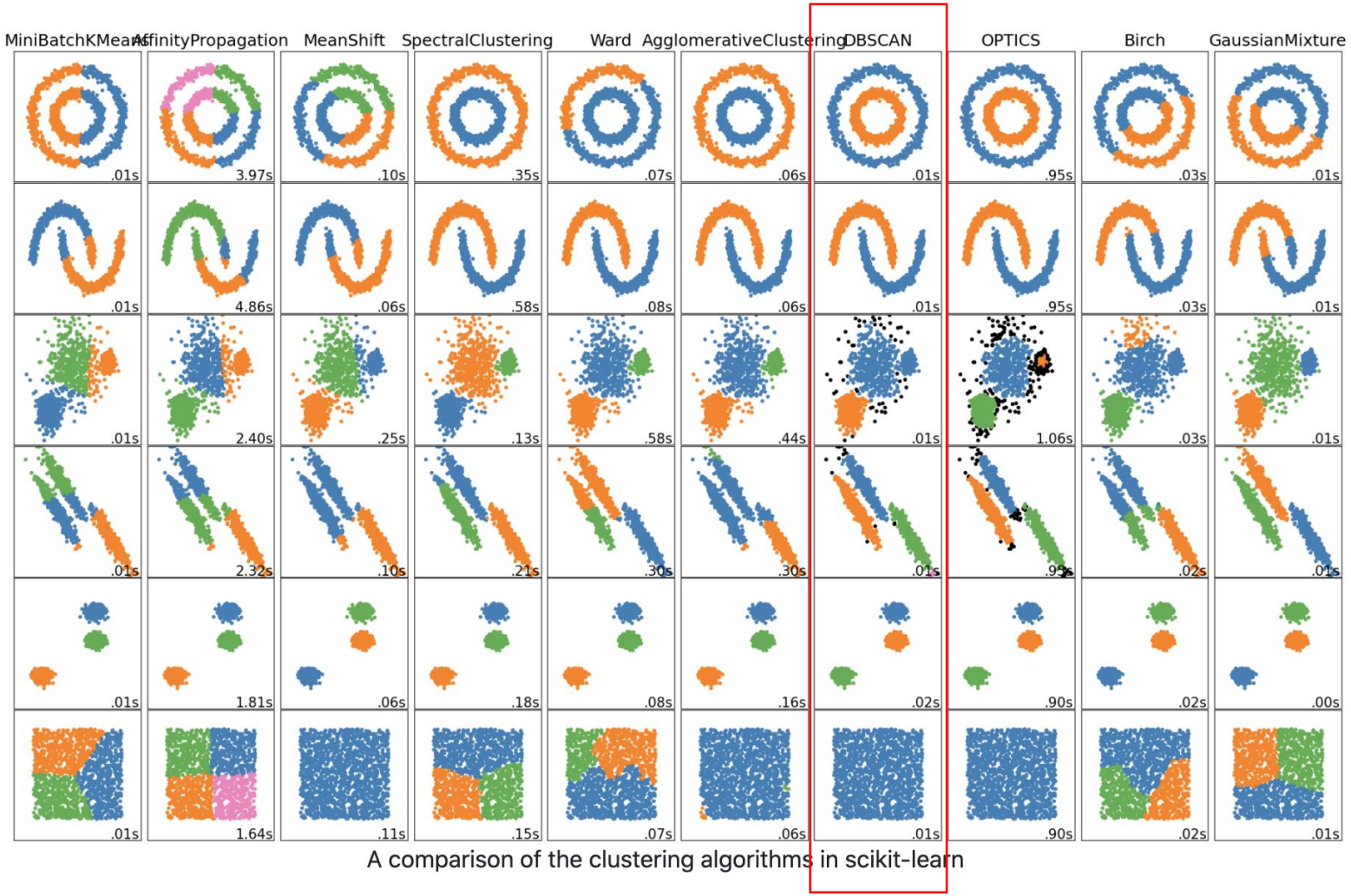
DBSCAN is a density-based clustered algorithm but with a couple of notable advantages. Firstly, it does not require a pe-set number of clusters at all. It also identifies outliers as noises.

Additionally, it can **recognize complexly shaped clusters** very well.

[Click for Visualizing DBSCAN!]



DBSCAN



A comparison of the clustering algorithms in scikit-learn

Thanks!