

# 01 Lab: Clustering

# Lab: Clustering

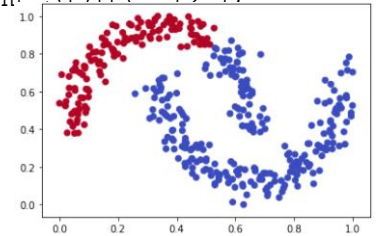
## Data sets for the following exercises

### Option 1) make\_moons

[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make\\_moons.html#sklearn.datasets.make\\_moons](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html#sklearn.datasets.make_moons)

create with: `data, labels = make_moons(n_samples=200, noise=0.1, random_state=123)`

=> IGNORE THE CLASS LABEL (TARGET VARIABLE) for clustering, do not use it



### Option 2) wine data set (see [https://scikit-learn.org/stable/datasets/toy\\_dataset.html](https://scikit-learn.org/stable/datasets/toy_dataset.html))

- ▶ Chemical analysis to determine the origin of wines using the „wine“ data set.
- number of instances: 178
- number of features: 13
- number of „classes“: 3 different origins of Italian wine
- ▶ **features:** Alcohol ; Malic acid ; Ash ; Alcalinity of ash ; Magnesium ; Total phenols ; Flavanoids ; Nonflavanoid phenols ; Proanthocyanins ; Color intensity ; Hue ; OD280/OD315 of diluted wines  
Proline
- ▶ one column „class“: with the types of wine {1, 2, 3}  
=> IGNORE THE CLASS LABEL (TARGET VARIABLE) for clustering, do not use it



## Exercise 1: Python Programming Cluster Analysis (k-means)

**Alternative:** if you prefer not to write programs, experiment here:

<https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/>

[https://www.philippe-fournier-viger.com/tools/kmeans\\_demo.php](https://www.philippe-fournier-viger.com/tools/kmeans_demo.php)

- a) Plot the data set.
- b) Cluster the previously introduced data sets using k-means.
  - ▶ Which method do you know?
  - ▶ Which important parameter do you need to pass to the clustering function?
  - ▶ What could be good value for that parameter?
  - ▶ Try different values and compare the results
- c) Partitioning methods are typically run more than once, why?  
Test different number of runs and compare the results.

## **Exercise 2: Cluster Analysis – Hierarchical methods and further clustering methods**

- a) Cluster the previously introduced data sets using hierarchical and/or further clustering methods.
  
- b) What effects do the hyperparameters have?

# Lab: Clustering

---

## **Exercise 3:** For the very fast ones...

Free experimentation, learning by doing... „own data set“

- ▶ Do you have a favourite structured data set (matrix-like) at hand?  
Otherwise search for a data set on the internet or use some built-in data sets
  
- ▶ Cluster the data and discuss with your neighbours.
  
- ▶ Document with screenshots, so that we can discuss your interesting findings.