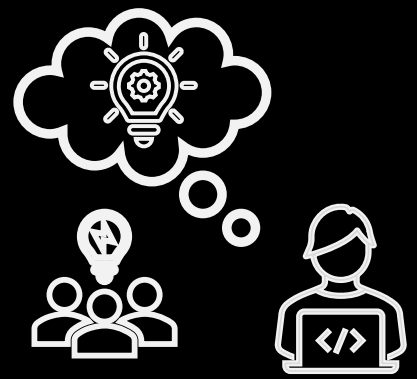
Lab session #7:
DBSCAN Clustering

Giulia Cisotto

Department of Informatics, Systems and Communication

University of Milan-Bicocca

giulia.cisotto@unimib.it



Academic year 2023/2024 – II semester

MOTIVATION

This seventh lab session aims **to design and apply the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm** to an unknown matrix of data. This lab session refers to Prof. Stella's lectures no.7 and 8 "Cluster Analysis: DBSCAN", and "Cluster Analysis: Clustering validation".
You are going to **(re-)use already developed code**: check solutions of previous labs. Also, you are going to add code to properly select the main parameters of DBSCAN (ε, i.e., the radius, and *MinPts*, i.e., the minimum number of points required to define a "dense" core region), and to test this particular algorithm. The use of package **sklearn.cluster.DBSCAN** is the main novelty of this lab session.

Here is the link to **the Python code @Colab for today**

The **data** to work on will be **available on Moodle** at the beginning to the lab session.

The **true labels** will be **available on Moodle** during the lab session.

Useful **packages**: numpy, pandas, scipy, matplotlib, seaborn, sklearn, **scipy.cluster (NEW!)**

You will *not* be given with the parameter Kd, then you need to guess an appropriate no. clusters by analysis.

TASKS

1. Guess the number of clusters (K), if any, in the dataset from visual inspection.

2. Apply DBSCAN in a *naive way* and check results (find the number of *Kd* clusters formed by the algorithm and visualize the clustering solution).

3. After identifying the "knee point“ using the “knee method”, apply DBSCAN again and compare the new results (any change in the number of Kd clusters formed?).

4. Validate the clustering solution obtained at point 3, using one or more measures/strategies to validate a clustering solution (e.g., inter-cluster distanced, silhouette score, ...).

5. Use the true labels to further validate the DBSCAN solution using supervised measures.