Project 2: Clustering Algorithms

General Introduction:

In this project, you are asked to learn about clustering algorithms. Each team should submit codes and a report via Canvas.

Dataset Description: Two gene datasets (*cho.txt* and *iyer.txt*).
Dataset format: Each row represents a gene:
1) the first column is gene_id.

2) the second column is the ground truth clusters. You can compare it with your results. "-1" means outliers.

3) the rest columns represent gene's expression values (attributes).

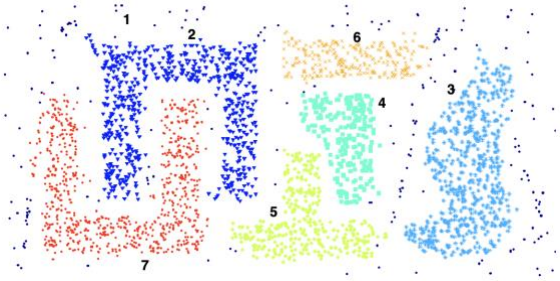
Complete the following tasks:

Select three clustering algorithms to find clusters of genes that exhibit similar expression profiles: choose from 1). K-means, 2). Hierarchical Agglomerative clustering with two different inter-cluster distances, 3). density-based, 4). mixture model, and 5). spectral. Compare their performance, and discuss their pros and cons.

Existing packages of these clustering algorithms can be directly used.

For each of the above tasks, you are required to validate your clustering results using the following methods:

- Choose an external index (e.g., Rand Index or Jaccard Coefficient) and compare the clustering results of different clustering algorithms. The ground truth clusters are provided in the datasets.
- Visualize data sets and clustering results of these algorithms by Principal Component Analysis (PCA). Different clusters are shown in different colors. For example, a clustering result like this:



- You can use existing packages for the external index, PCA, and visualization tools such as Matplotlib.

Project Submission:

• Prepare your submission. **One team only needs to provide one submission**. Make a zipped folder named "**CaseID_CaseID_Proj2.zip**", where "CaseID" refers to your group members' Case IDs. In the folder, you should include:

1. **Report**: A pdf file named **Clustering_report.pdf**. Describe algorithms you selected. Compare the performance of these approaches using visualization and external index on the two given data sets. State the pros and cons of each algorithm and any findings you get from the experiments. The report is suggested to **not exceed** 4 pages.
2. **Code**: A zipped folder named **code.zip**, which contains all codes used in this part. Inside the folder, please also provide a **README** file which describes how to run your code.

Note that copying code/results/report from another group or source is not allowed and may result in an F in the grades of all the team members.

Grades will be given based on the following criterion:

- Report (65 pts):
  - (6x3=18 pts) Algorithm description: brief description of the 3 algorithms
  - (5x3=15 pts) Result visualization
  - (8x3=24 pts) Result evaluation, analysis
  - (8 pts) Overall coherence and clarity
- Code (35 pts)
  - (10x3=30 pts) Correctness and Reproducibility of the results in the report
  - (5) Readme: Readability and clarity