Final Project–Big Data

This is Big data final project for two courses. Please do your best to uncover the hidden secrets of this Big Data. While exploring the boundaries of cosmic knowledge, Professor Liao discovered a series of particle accelerator datasets. These datasets may help him identify the so-called “God Particle” — or perhaps the “Devil Particle” — a potential breakthrough that could lead to the next Nobel Prize.

Task:

Your task is to analyze the relationships within this dataset and classify the data into several distinct categories. It is known that if the data has n dimensions, you should be able to clearly observe 4n – 1 clusters. Please attempt to group the data into these 4n – 1 clusters. The actual numerical labels you assign to the clusters are not important — what matters is whether the clustering itself is accurate.

Your results will be evaluated using **Normalized Mutual Information (NMI)** by comparing your clusters with a hidden ground truth. This NMI score will determine your final grade.

There are two types of datasets:

- A **public dataset** with 4 dimensions, for which we will provide the grading script so you can check your performance.
- A **private dataset** with 6 dimensions.

Please write a short report explaining why your algorithm is effective at clustering the data.
Your report should briefly describe:

- The algorithm or method you used
- Why it is suitable for this dataset
- How it handles high-dimensional data
- Any preprocessing, hyperparameters, or assumptions involved

Rules:

1. Individual project – each student must work independently.
2. You may use any clustering methods or algorithms that you find suitable.
3. No plagiarism or cheating. Any violations will result in a zero for the final project and may lead to academic dismissal.

Grading Criteria:

- **60%** Public dataset score
- **30%** Private dataset score
- **20%** Report quality
  = **Total: 110%**