MLCMS Final Project – A Page Summary
(Group J)

Names: Hu, Wenbin (03779096) ;
Sun, Mei (03755382);
Tang, Yiling (03755346).

Learn and Visualize Representations of Large Data Sets.

Project Summary:

The **curse of dimensionality** long stills in Data Science and Machine Learning caused by the sparseness of data. It will be beneficial to reduce the dimension of the data for data learning efficacy and efficiency. **Manifold Hypothesis** suggests that high-dimensional data often lies on or near low-dimensional manifolds within the larger space. Based on it, there exists some ways to reduce the data dimension, and we will try to explore those methods. The datasets we use are: **Swiss Roll** (Artificial), **Word2vec** (Natural, Language), and **Cifar10** (Natural, Vision). The methods of dimensionality reduction we will use are: **PCA** (Convex, Euclidean Distance), **Diffusion Map** (Convex, Diffusion Distance), and **VAE** (non-Convex, AutoEncoder). We will try to analyze these methods theoretically and conduct some experiments to compare their performance. The metric for the performance of dimensionality reduction is not trivial or unified, and what we try to do is to visualize the result and show what extend the dimension-reduced data will help the ML algorithm.

Task 1/4: Introduce the datasets we use: Swiss Roll, Word2vec, Cifar10.

Show these 3 datasets: do some visualization and describe the datasets. Show how to load these data.

Task 2/4: Introduce the methods we use: PCA, Diffusion Map, and VAE.

Describe, analyse, and categorize these 3 methods. Show why they are used for data dimensionality reduction and compare each method (Pros & Cons). Show what tools do we use / how we use them for implementing these 3 methods. (sklearn, torch)

Task 3/4: Conduct experiments to show the results of the 3 methods on the 3 datasets.

Conduct fair experiments for these methods and datasets and compare their performance.(9 combinations) We will visualize the dimension-reduced data ,compare the performance of related ML algorithms on the original data and the dimension-reduced data ,record not only the run time for each experiment but aslo how large the input datasets can be.

Task 4/4: Conclusion: What really matters when we come to data dimensionality reduction.

From the analysis and the experiments, we draw a conclusion for the key to data dimensionality reduction task. What's the limitations and what's the future direction?