**University of Tehran**

College of Science

School of Computer Science

Data Mining: HW3 Dr. Sajedi

---

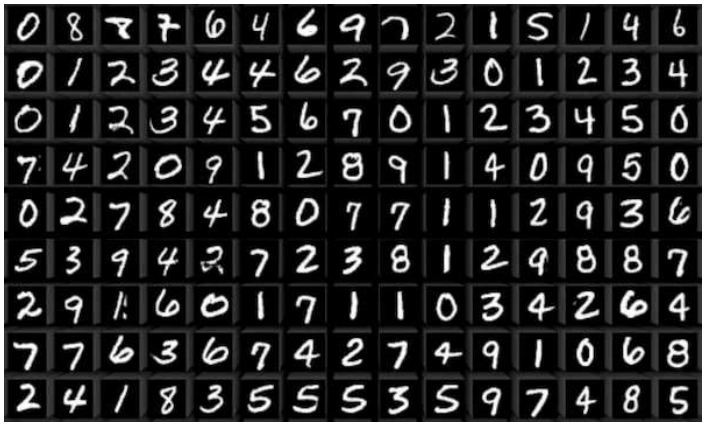
# AutoEncoders & Clustering

## 1. Objective

This project aims to utilize AutoEncoders to reduce the dimensionality of the MNIST dataset, which comprises 70,000 handwritten digit images. After dimensionality reduction, multiple clustering methods, including K-means and DBSCAN, will be applied to segment the data into distinct clusters. Subsequently, the clustered, reduced images will be reconstructed to their original form. Finally, an analysis will be conducted to determine which clustering method performs the best.

## 2. Dataset Description

The MNIST dataset is a large collection of handwritten digits commonly used for training various image processing systems. It contains 70,000 grayscale images of digits (0-9), each 28*28 pixels in size. The dataset is divided into 60,000 training images and 10,000 testing images, providing a standard benchmark for evaluating the performance of machine learning algorithms in digit classification tasks.



# 3. Instructions

### a. Data Preprocessing

Load the dataset and preprocess it by addressing outliers, and any other necessary preprocessing methods that you think is useful. Use libraries such as pandas and NumPy to perform efficient data manipulation.

### b. Feature Reduction

Utilize an AutoEncoder to extract features from the images. Fine-tune hyperparameters such as the number of layers, number of neurons per layer, learning rate, batch size, and activation functions to optimize performance. Build the best possible model based on preferred reconstruction errors, such as Mean Squared Error (MSE) and Mean Absolute Error (MAE), supplemented by visual inspection of the reconstructed images and by tracking model loss over the training period.

### c. Clustering

In this section, you will apply K-Means and a variant from its family, such as Fuzzy C-Means, Weighted K-Means, etc. Additionally, you will use DBSCAN along with a related algorithm like HDBSCAN or VDBSCAN or any preferred algorithm in as its family member. These four clustering methods will be employed to cluster the extracted features from the previous step into the provided clusters.

### d. Image Reconstruction

In this section, you will reconstruct the extracted features in each cluster using the pre-built decoder from your AutoEncoder. Additionally, you will visually inspect the reconstructed images to evaluate the performance of your decoder.

### e. Comparative Analysis of Methods

Now, use the labels from the MNIST dataset to analyze how well each clustering algorithm has grouped the handwritten images according to their actual digits. Compare the performance of the algorithms and include your assessments in your report.

### f. Report Preparation

Write a comprehensive report structured with sections including Introduction, Preprocessing, Model Implementation, Method Comparison, and Conclusion. In the Method Comparison section, explain how each selected clustering algorithm works. Ensure your report includes visual evidence to support the visual assessment

of the clustering methods. Additionally, provide detailed explanations of your work in each section to thoroughly document your process and findings.

# 4. Additional Guidance

Ensure to send a ZIP file containing both your report and your code. Your code must be in ***.ipynb*** (Jupyter Notebook) format; otherwise, your assignment will not be assessed. The file name should follow this pattern: **FULLNAME_STUDENTID_DM_HW3**.