Ray Chen (rchen63@Buffalo.edu)

Ask:

- What is the purpose of the project?
- What are the inputs?
- What are the outputs?
- Are there any constraints?
- What are PAA and SAX
- What are Euclidean Distance and Hamhattan Distance?

Research/Imagine:

- The purpose of this project is to explore representation and classification of time series.
- The input is a synthetic control data set from the University of California Irvine. The set has 60 columns of data (60-time points) with values between 0 and 100 and 6 different classes.
- The outputs are the PAA set and SAX set.
- PAA stands for Piecewise Aggregate Approximation, the aim of PAA is to treduce the data based on aggregate value in each piecewise region. SAX stands Symbolic Aggregate Approximation. SAX "proves that a distance measure between tow symbolic string lower bounds the true distance between the original time seies in non-trivial." [1]
- "By far the most common distance measure for time series is the Euclidean distance. Given two time series *Q* and *C* of the same length *n*, Equation below defines their Euclidean distance." [1]

$$D(Q,C) \equiv \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2}$$

- Equation of Manhattan Distance, assuming 2 time series *Q* and *C* and length *n:*

$$D(Q,C) \equiv \sum_{i=1}^{n}|q_i - c_i|$$

Plan:

1. Import the data form the website.
   - Download the data file
   - Load the data into matlab
2. Preprocess the data.

- Normalize the data
- Standardize the data.
- Filtering out missing data
- Transform the data.

3. Sampling and representation
    - Create two new representation data sets that reduce the number of samples needed to represent the data. The first set will utilize PAA and the second will use SAX
4. Analysis
    - Classification using Euclidean Distance and Hamhattan Distance, and compare which is better
5. Writing report

REFERENCES

[1] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. 2003. A symbolic representation of time series, with implications for streaming algorithms. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (DMKD '03). Association for Computing Machinery, New York, NY, USA, 2–11. DOI:https://doi.org/10.1145/882082.882086 https://www.cs.ucr.edu/~eamonn/SAX.pdf