# Data Mining II
# Project - Week 1
# Team 2

**Basar Temur, Cagan Yigit Deliktas, Niklas Wichter, Rong Hu, Shiyi Xu, Ting Huang**

**Feature Selection**

Given that there are over 100 features in the dataset, our first step is to filter out "pointless" features. Besides, the dataset encompasses 11 distinct sectors, which are assumed to have their own principal features. Under these two preconditions, we consider applying the following methods for each individual sector.

- PCA
- RFECV

**Missing Values, Outliers, Oversampling (Smote) for class imbalance**

**Models to be Applied**

RandomForest, DecisionTree, SVM, NaiveBayes, KNN, LogisticRegression, GradientBoost, XGBoost, AdaBoost, Voting, Stacking, NeuralNetwork

**Scoring Metrics**

***Weighted F1 Score:** F1 score calculated by taking the average of F1 scores for each class. Average is weighted by support which is the number of true instances for each label.

***AUC One vs One Weighted:** By considering all pairwise combinations of classes, average AUC is calculated. Average is weighted by the support.