# MATH 829 - Spring 2016
# Introduction to data mining and analysis
# Homework #4

Instructions:

(1) You are allowed to work in groups of 1-4.
(2) Show (and briefly explain) all of your work to receive full credit.
(3) Submit your work before **Saturday, April 16th, 2016**.

---

## Python part

---

**Problem 1.** Consider the `spam` dataset.

a) Use `sklearn` to compute the first 10 principal components of the data.
b) Compute the explained variance ratios for the 10 first principal components.
c) Display the first two principal components with a scatter plot. Use a color code to display the digit corresponding to each point (as in slide 9 of the PCA lecture).

**Problem 2.** The RMS Titanic sank in the North Atlantic Ocean in the early morning of April 15th 1912, after colliding with an iceberg during her maiden voyage from Southampton, UK, to New York City, US. A total of 1502 out of 2224 passengers and crew died during this tragedy. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. The dataset `titanic` provided by Kaggle (available on Sakai) provides information about the Titanic passengers (such as their age, sex, and passenger class), and whether or not they survived the sinking.

a) Use `sklearn` to fit a decision tree to the titanic data. You may want to use only some of the variables, and remove some entries with missing values.
b) Export your decision tree in pdf using graphviz (see `http://scikit-learn.org/stable/modules/tree.html`)
c) Use `sklearn` to fit a random forest model to the titanic data.
d) Measure the prediction accuracy of your decision tree and random forest models with a train/test experiment.