# Three ideas for capstone project 1

## Project Idea 1:

**Data set: Digit Recognizer**
http:// www.kaggle.com/c/digit-recognizer/data

**Data set introduction:**
The data files train.csv and test.csv contain gray-scale images of hand-drawn digits, from zero through nine. Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total.

The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image. The test data set, (test.csv), is the same as the training set, except that it does not contain the "label" column.

**Story to tell:** To evaluate the proportion of test images that are correctly classified. For example, a categorization accuracy of 0.97 indicates that you have correctly classified all but 3% of the images.

**Potential client:** Digit recognition is widely used in image processing of medical industry and it also plays an important role in hand written recognition.

## Project Idea 2:

**Data set: Default of credit card clients**
http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

**Story to tell:** This research aimed at the case of customer credit card default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. This study is also to investigate if the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

**Potential client:** Insurance Company, bank.

## Project Idea 3:

**Data set:** Twitter data set for Arabic Sentimental Analysis
http://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis

**Data set introduction:** By using a tweet crawler, 2000 labelled tweets (1000 positive tweets and 1000 negative ones) were collected. On various topics such as: politics and arts. These tweets include opinions written in both Modern Standard Arabic (MSA) and the Jordanian dialect.

**Story to tell:** The selected tweets convey some kind of feelings (positive or negative) and the objective of our model is to extract valuable information from such tweets in order to determine the sentiment orientation of the inputted text.

**Potential client:** Social media.