# Machine Learning Course Project

This final project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set.

- It can be conducted in groups of **at most 2 students**.
  Note that each student in the groups of two must submit the project individually.
- You can only use Python or MATLAB for the project programming.
- There are following two deliverables, which must be submitted using Class ssystem.
  (Deadline will be determined later.)

  1. **Report**
  2. **Program**

- Final project report is expected to be in a research paper format (8 pages consistent with point size 12 and single line spacing in MS Word). The final report should roughly have the following format.

  - ✓ Introduction - Motivation
  - ✓ Problem definition
  - ✓ Proposed method
    - Intuition - why should it be the preferred method?
    - Brief description of the algorithm
  - ✓ Experiments
    - Description of your test-bed; list of questions your experiments are designed to answer
    - Details of the experiments; observations
  - ✓ Conclusions

The project may involve applying existing methods (classification/regression, supervised/unsupervised, etc.) to solve an interesting question. Or you may work on developing a new methodology to solve an existing problem on an existing data set.

For possible topics, have a look at Andrew Ng's course projects to get some ideas.
http://cs229.stanford.edu/projects2015.html

Kaggle (https://www.kaggle.com/competitions) has a long list of (machine learning) problems! The problems are cast as open competitions. You can consider picking up a problem from Kaggle (they often have the data available) and maybe even win a prize. But make sure the problem is not too simple (You may consult me via email).

Here are some other sources:

### 1. Anomaly-detection task

The typing anomaly-detection task is to discriminate between the typing of a genuine user trying to gain legitimate access to his or her account and the typing of an impostor trying to gain access illegitimately to that same account. This webpage (http://www.cs.cmu.edu/~keystroke/) is a benchmark data set for keystroke dynamics. The data consist of keystroke-timing information from 51 subjects (typists), each typing a password 400 times. The project would be to use the data on this page to learn a classifier that determines reliably the identity of a given typist.

### 2. Image Segmentation Dataset

The goal is to segment images in a meaningful way. Berkeley collected three hundred images and paid students to hand-segment each one (usually each image has multiple hand segmentations). Two-hundred of these images are training images, and the remaining 100 are test images. The dataset includes code for reading the images and ground-truth labels, computing the benchmark scores, and some other utility functions. It also includes code for a segmentation example. http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

### 3. Email Annotation

The datasets provided below are sets of emails. The goal is to identify which parts of the email refer to a person's name. This task is an example of the general problem area of Information Extraction.
http://www.cs.cmu.edu/~einat/datasets.html

### 4. Object Recognition

The Caltech 256 dataset contains images of 256 object categories taken at varying orientations, varying lighting conditions, and with different backgrounds.
http://www.vision.caltech.edu/Image_Datasets/Caltech256/

**Here is more datasets and challenges:**

- UCI (http://archive.ics.uci.edu/ml/index.php)
- CIFAR(https://www.cs.toronto.edu/~kriz/cifar.html)
- KDD Cup (https://www.kdd.org/kdd-cup)
- Dream-Challenges (http://dreamchallenges.org/)
- Datasets for Data Science (https://www.kdnuggets.com/datasets/index.html)
- Large collection of network datasets (http://networkrepository.com/index.php)
- Awesome Public Datasets (https://github.com/awesomedata/awesome-public-datasets)
- UCI datasets (https://archive.ics.uci.edu/ml/datasets.php)
- NYC Open Data (https://opendata.cityofnewyork.us/)
- NYC Taxi Data (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page)
- Outlier Detection Data Sets (ODDS) (http://odds.cs.stonybrook.edu/)
- Stance identification dataset for fake news detection (http://www.fakenewschallenge.org/)
- Foursquare check-ins (https://sites.google.com/site/yangdingqi/home/foursquare-dataset)

- Product review datasets
    - Amazon product reviews (https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets)
    - Online reviews from SNAP (http://snap.stanford.edu/data/index.html#reviews)
- Amazon product data (http://jmcauley.ucsd.edu/data/amazon/)
- Stack Exchange Data Dump (https://archive.org/details/stackexchange)
- Google public datasets (https://www.google.com/publicdata/directory)
- List of large datasets open to public (https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public)
- Million Song Dataset (http://millionsongdataset.com/)
- Free 'big data' sources'(https://www.datasciencecentral.com/profiles/blogs/the-free-big-data-sources-everyone-should-know)
- AWS Public Datasets (https://registry.opendata.aws/)