COMP 4220 Machine Learning, Fall 2018
Final Project

The goal of the final project is to show that you have learned something in the class. It is an opportunity for you to explore ideas that you see in the lectures and assignments and extend them. You can think of your project as first steps towards research in machine learning or its applications. You will analyze the problem, design a machine learning solution, implement ML algorithms, and evaluate them on **three** data sets (one for classification and one for regression from "Small Data Sets" and one data set – either classification or regression – from "Large Data Sets").

Milestones and Grading

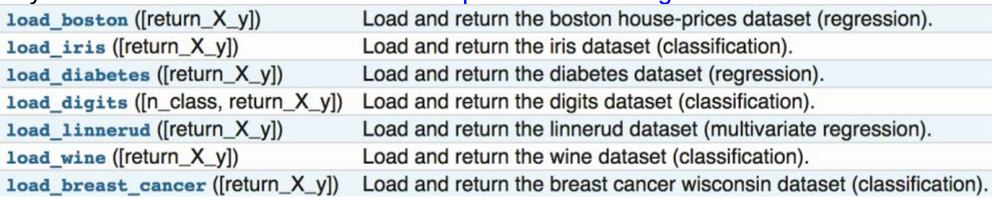
The project is worth 30% of the class grade. This is broken down across the following milestones:

- Form a project team (5 points): All projects must be done in teams of three to four students and inform the instructor who is in your group by Thursday, November 1 (only team leaders).
- Final report (55 points): The final report can be at most five pages (size 11 font, including python codes and figures). The report should be structured like a small research paper. Broadly speaking it should describe:
  - What problem and data sets did you work on?
  - What are the important ideas/methods you explored?
  - What ideas from the class did you use?
  - Reporting the results (cross-validation, easy-to-read figures, etc)
  - Do the results make sense?
  - If you had much more time, how would you continue the project?

  The final report is due by Thursday, December 13 (all team members should submit the final report).
- In-class Presentation (40 points): You should prepare slides for an 8-minute presentation of your project, with 2 minutes for questions. Your slides should contain a summary of methods you implemented, results, and discussions on what these results mean. Each team member should present part of the slides on Monday, December 10.

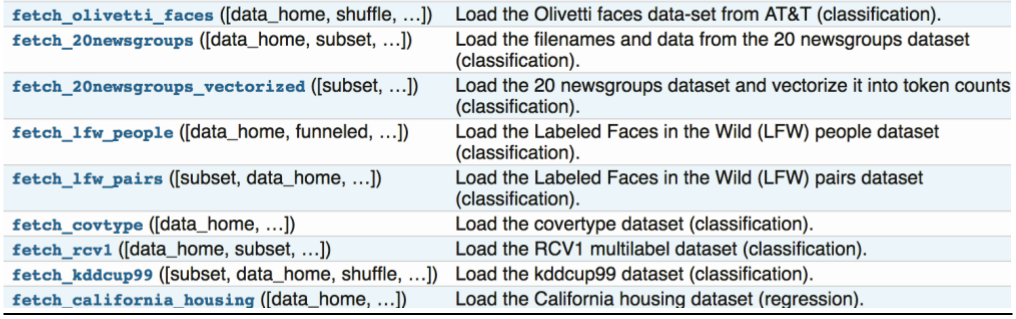
Data Sets

- Small Data Sets: Scikit-learn comes with a few standard data sets that do not require to download any file from some external website: http://scikit-learn.org/stable/datasets/index.html (Section 5.2)



`load_boston` ([return_X_y])	Load and return the boston house-prices dataset (regression).
`load_iris` ([return_X_y])	Load and return the iris dataset (classification).
`load_diabetes` ([return_X_y])	Load and return the diabetes dataset (regression).
`load_digits` ([n_class, return_X_y])	Load and return the digits dataset (classification).
`load_linnerud` ([return_X_y])	Load and return the linnerud dataset (multivariate regression).
`load_wine` ([return_X_y])	Load and return the wine dataset (classification).
`load_breast_cancer` ([return_X_y])	Load and return the breast cancer wisconsin dataset (classification).

- Large Data Sets: Scikit-learn also comes with tools to load larger data sets: http://scikit-learn.org/stable/datasets/index.html (Section 5.3)



`fetch_olivetti_faces` ([data_home, shuffle, …])	Load the Olivetti faces data-set from AT&T (classification).
`fetch_20newsgroups` ([data_home, subset, …])	Load the filenames and data from the 20 newsgroups dataset (classification).
`fetch_20newsgroups_vectorized` ([subset, …])	Load the 20 newsgroups dataset and vectorize it into token counts (classification).
`fetch_lfw_people` ([data_home, funneled, …])	Load the Labeled Faces in the Wild (LFW) people dataset (classification).
`fetch_lfw_pairs` ([subset, data_home, …])	Load the Labeled Faces in the Wild (LFW) pairs dataset (classification).
`fetch_covtype` ([data_home, …])	Load the covertype dataset (classification).
`fetch_rcv1` ([data_home, subset, …])	Load the RCV1 multilabel dataset (classification).
`fetch_kddcup99` ([subset, data_home, shuffle, …])	Load the kddcup99 dataset (classification).
`fetch_california_housing` ([data_home, …])	Load the California housing dataset (regression).