# Syllabus and Setup

Welcome!

This class is an introduction to data cleaning, analysis and visualization. We will walk you through as we analyze real world datasets. Each day, we will spend the first 30 minutes introducing the day's concepts, and the rest of the class will be exercises. We have written a daily walkthrough that you will read and program through in class, and we will be available to help.

This is our first time teaching this course, and we'll be learning as much as you. Don't hesitate to ask us to change something or improve on something. We'll be grateful.

## Prereqs

We assume you have a working knowledge of python (6.01) and are willing to write code. Most of the code you interact with will come with an example that you can modify. Hopefully little specialized code will be generated except for programs you're inspired to write!

## What we will teach

We will teach the basics of data analysis through concrete examples. All of your programming will be written in python. The schedule is as follows:

- Day 0 (today): setup
- Day 1: An end-to-end example getting you from a dataset found online to several plots of campaign contributions.
- Day 2: Lots of visualization examples, and practice going from data to chart.
- Day 3: Statistics basics, including T-Tests, Linear Regression, and statistical significance. We'll use campaign finance and per-county health rankings.
- Day 4: Text processing on a large text corpus (the Enron email dataset) using tf-idf and cosine similarity.
- Day 5: Scaling up to process large datasets using Hadoop/MapReduce on a larger copy of the Enron dataset.
- Day 6: You tell us! Get into groups or work on your own to analyze a dataset of your choosing,

and tell us a story!

# What we will not teach

- *R*. R is a wonderful data analysis, statistics, and plotting framework. We will not be using it because we can achieve all of our objectives in Python, and more MIT undergraduates know Python.
- Visualization using browser technology (canvas, svg, d3, etc) or in non python languages (Processing). These tools are very interesting, and lots of visualizations on the web use these tools (e.g., nytimes visualizations), however they are out of the scope of this class. We'll teach you how to visualize data in static charts. If this is an area of interest for you, the next step will be to build interactive visualizations that the world can explore, and we can point you in the right direction with these.

# Programming Environment (Important!)

Before the class, please set up the environment. You will need to install some software, packages, and download some datasets to get started.

We assume that you are developing in a unix-like environment and are familiar with the common commands (e.g., less, man). If you are a windows user, we assume you are using cygwin but are on your own.

# Tools and Libraries

In this class, you will need to install a number of tools. The major ones are:

- python 2.7
  - Python is usually installed in Mac OSX and major unix distributions. Type `python --version` to make sure it is the right version
- easy_install
  - python package manager.
- pip
  - Makes installing python packages really easy. Requires easy_install.
  - Either install it by typing `sudo easy_install pip` or download the tar.gz file at the

    link above, untar it, go into the newly created directory, and type `sudo python setup.py install`.

- git
    - git is a version control system. Using it, you can check out our code and examples.
    - If everything is working, check the dataiap sourcecode into a directory called `dataiap` using `git clone git://github.com/dataiap/dataiap.git dataiap`
    - We'll be updating the repository periodically. To get the latest copy, go to the `dataiap` directory and type `git pull`.

We will also require a number of python modules:

- numpy 1.6.x: numerical processing module.
    - PIP users can type `sudo pip install numpy`
- scipy 0.10: scientific computing module.
    - Ubuntu users can type `sudo apt-get install python-scipy`
    - PIP users can type `sudo pip install scipy`
    - Even if PIP works, at least on MacOS you might have to install Fortran. We strongly recommend reading and following the installation instructions.
    - Unfortunately, scipy installation might not work from PIP, and you may have to compile it from source (see "Obtaining and Building NumPy and SciPy"). Try something akin to
        - `git clone https://github.com/scipy/scipy.git`
        - `cd scipy`
        - `python setup.py build`
        - `python setup.py install`
- matplotlib 1.1.0
    - PIP users can type `sudo pip install matplotlib`
    - Note: If compiling from source, matplot lib requires a number of other libraries: (libpng, freetype 2)
    - Some MacOS users might run into issues and should just download the binary.
- dateutil
    - PIP users can type `sudo pip install python-dateutil`
- pyparsing
    - PIP users can type `sudo pip install pyparsing`
- mrjob: This is a MapReduce package that we will use it in day 5.

    - PIP users can type `sudo pip install mrjob`
    - If compiling from source, it requires boto (try `sudo pip install boto`).

For convenience, Enthought provides numpy, scipy, matplotlib in a single installable package. Many students that had trouble installing these modules separately were able to install Enthought.

## `dataiap/` Directory Structure

The repository contains the contents of the full course. We will be using

- `dayX/`: files containing the lecture for day X
- `datasets/`: the datasets we will be using should live here
- `resources/`: contains python scripts that you will eventually run
    - `util/`: contains python modules we have written that you will use in this course.
    - `inst/`: instructor python files. Used to setup and test the labs. Please don't view during the course.

## Datasets

We will be working with several datasets in this course. Most of them have been added to the git repository.

The presidential contributions dataset is fairly large. We will use it on the first day, so please download it from ftp://ftp.fec.gov/FEC/Presidential_Map/2008/P00000001/P00000001-ALL.zip.

The datasets we will use are

- 2008 Presidential Campaign Contributions
    - The linked file contains all of the 2008 campaign contributions to each presidential candidate. You can look at the 2012 campaign for various primary candidates as well, but we'll work with 2008 since it's complete.
    - unzip into `dataiap/datasets/pres_campaign/`
- 2011 County Health Rankings
    - The dataset contains per-county health and morbidity statistics.
    - The necessary data should already be uncompressed in `dataiap/datasets/county_health_rankings/additional_measures_cleaned.csv` `dataiap/datasets/county_health_rankings/ypll.csv`

- The Enron email dataset
    - This is the complete set of emails on the enron email server that was released during the scandal. Don't download the dataset as it's huge. We have included subsets of the datasets in the git repository.
        - `dataiap/datasets/emails/kenneth.zip` contains a subset of Kenneth Lay's emails that you will analyze in day 4.
        - `dataiap/datasets/emails/kenneth_json.zip` contains a JSON-encoded subset of Kenneth Lay's emails that you will analyze in day 5.
    - We will upload a JSON encoded version of the full dataset to amazon's S3.

MIT OpenCourseWare
http://ocw.mit.edu

Resource: How to Process, Analyze and Visualize Data
Adam Marcus and Eugene Wu

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.