Course Project

- Teams of up to 6 people
- Go through entire process of data science
  - Data Cleaning and Preparation
  - Using software tools
  - Analysis
  - Presentation of results
  - Visualization

Data Sources

- Many governments and organizations now warehouse data in a way easily consumable by the public
- `data.gov`: open data covering a large number of topics, US Government
- `open.canada.ca`: Canadian open data
- `data.gov.uk`: United Kingdom open data
- `data.nasa.gov`: Public NASA datasets
- `data.worldbank.org`: Datasets related to development in countries around the world

Additional Data Sources

- In addition to the data warehouses listed above many businesses and popular websites make some of their data available programmatically via a RESTful API
- A good resource to find such data sources is `www.programmableweb.com`, which maintains a list of popular APIs
- Social Media sites in particular are rich sources of data, communities of followers, social network graph
- Facebook, Instagram, Twitter, VK, Twitch all expose APIs

Assignment

- Once you have selected your team members, you should collectively browse through the data sources listed above and choose a topic to study
- Be aware that if you choose to collect your own dataset from an api your team will also have to write a scraper application (be aware of rate limitations etc.)
- Also, depending on the dataset you choose you may be able to simply import a single csv file using (for example) `pandas` or you may need to have a strategy to deal with data in JSON or XML format

Assignment cont.

- As stated in the syllabus the course project will weigh 15 percent of the final grade
- Grade will be assigned collectively based upon team submissions
- The final output of the report will be a written report that communicates the results of the analysis that you perform
- Can create static visualizations using the matplotlib library in python and dynamic/interactive visualizations via the d3 or an equivalent library

In-Class Exercise

- Please use this time to select your team members and peruse the publicly available datasets given here
- In general I would prefer that teams generally take more of a data mining approach versus a machine learning approach
- In other words, focus more on the potential of *unsupervised* learning techniques in your chosen datasets
- Submit the names of your team members and tentative choice of datasets/topic on the moodle