Intro to Data Science, Fall 2018
Homework 1: Search and Discover
Questions 1-4 due Aug 27 at start of class; Question 5 due Sep 5

---

1. Find one data repository that we did not discuss in class. Describe it in a few sentences. Tell us the purpose of the repository, the clientele it serves, how big it is, what time frame it spans, how it serves its clients and (if possible) what it might be lacking or how it could be improved. Note that an ideal repository is very large, is comprehensive, possibly spans a large time frame, may include multiple aspects, publicly available, has downloadable data, but can also be queried via the internet.

   Here is an example. The website `imdb.com` is the world's largest movie database. It contains information about every movie made in every country. It also contains comprehensive information about actors and actresses, the directors, producers, and other individuals that go into making a movie. It can be queried and reportedly can be used in software (details are sketchy, however).

2. Generate at least one meaningful analytical question that has not been asked with regard to the data repository you reported above. Why is the answer to your question(s) useful or impactful? Does your question(s) require collection of fresh data? If so, what? How do you think this fresh data can be collected and how should you go about finding the answer to your question.

   For example, it might be interesting to ask if there is a correlation between the number of Academy Award winning actors/actresses born in a certain state and the number of degree programs in theater and drama in that state.

3. Data analysis is best communicated by effective visualization. Find the best example you have seen on the web for visual communication of data analysis. Explain why it is effective.

   For example, we saw the "Temperature Circle", which is an effective way to communicate how warming is a real phenomenon and is "global", not localized to some parts of the world. This is achieved by animating global temperature data over a 100 years.

4. Install Anaconda Navigator on your computer by downloading from `https://www.anaconda.com/download/` and following appropriate instructions. Anaconda is basically a package manager designed for fast experimentation (develope, test, train on a single machine) in Data Science with Python and R. It is also useful for experiments with Machine Learning tools and is available for Linux, Windows, and Mac OS X environment. In particular, it provides a useful Python distribution including NumPy, SciPy, JuPyter, H2O.ai, TensorFlow, matplotlib, and so much more. Enterprise versions are available for the cloud (AWS, GCP, Azure), Spark/Hadoop, and more. The website provides documentation needed to install on your computer. It will help you create "Notebooks" in Python and R for our work.

5. The EPA publishes air quality data on a continuous basis. It is possible to download data on fine particulate matter air pollution, also referred to as PM2.5, which refers to the amount of particulate matter that is smaller than 2.5 micrometers in the air. Higher this number, more is the pollution. The fields are called the following: *RD, Action.Code, State.Code, County.Code, Site.ID, Parameter, POC, Sample.Duration, Unit Method, Date, Start.Time, Sample.Value*. The air quality index is in the column titled *Sample.Value*. The two data files are called:

   `https://users.cs.fiu.edu/~giri/teach/5768/F18/RD_501_88101_1999-0.txt`

   and

   `https://users.cs.fiu.edu/~giri/teach/5768/F18/RD_501_88101_2012-0.txt`

   More details on the descriptions of the data can be found at:

   `https://aqs.epa.gov/aqsweb/airdata/FileFormats.html`.