GENERAL CONSIDERATIONS FOR ACQUIRING DATA AND AVAILABLE DATA REPOSITORIES

Before getting the data:

- What legal rights and limitations do you have to the data?
    - liability and terms of use,
    - ability to reproduce or publish results,
    - transfer of data restrictions (e.g. individuals vs. corporate)
- What are the technical limitations?
    - rate limits
    - reliability
    - variety of formats
    - Provenance

Download vs Collect

- Download:
  - a whole data set archive,
  - might be very large,
  - all or nothing proposition,
  - extracts/subsets are up to you.
- APIs for data collection:
  - APIs access data from services,
  - real-time or historical data,
  - subsets are defined by API capabilities,
  - may be focused on exact subsets

Collecting Data

- Things to consider:
  - likely to be writing scripts,
  - that run over long periods of time,
  - need reliable management and monitoring,
  - and estimations of time, space, and bandwidth necessary.
- How will your partition and store your results over time (e.g., AWS S3, files, etc.)?

Downloading data

- Things to consider:
  - similar needs as before,
  - for managing long running processes,
  - first you must download massive amounts of data,
  - process it into a subset.
- What will you keep?

Idempotent Process

- *Idempotent: a process can be applied multiple times and yields the same result.*
- Ideally, you want:
  - a data collection process,
  - that stores just enough data,
  - such that data cleaning, normalization, and subsetting,
  - can be reapplied at will.
- This lets you change your mind later and iterate on the four steps of data analysis.

Some links to Data Sets

- □ http://www.data.gov/open-gov/
- □ http://cnets.indiana.edu/groups/nan/webtraffic/click-dataset/
- □ http://lemurproject.org/clueweb12/
- □ http://www.1000genomes.org/ftpsearch/
- □ https://aws.amazon.com/datasets
- □ http://datahub.io/dataset
- □ http://datacatalogs.org/
- □ http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets
- □ http://blog.archive.org/2012/10/26/80-terabytes-of-archived-web-crawl-data-available-for-research/
- □ http://www.bigdata-madesimple.com/70-websites-to-get-large-data-repositories-for-free/