**Public Data Source for Data Analysis**

* [Awesome Public Datasets](https://github.com/caesar0301/awesome-public-datasets): A **huge**collection of public datasets. Categorized by field (e.g., biology, economics, machine learning, etc).
* [UCI Machine learning Repository](http://archive.ics.uci.edu/ml/): ”…currently maintain 349 data sets as a service to the machine learning community”
* https://www.kaggle.com/datasets: Also a list of publicly available datasets.
* [Goverment data](https://www.data.gov/): govermental data. Everything from agriculture to science & research. Very interesting.
* [Google Public Data](https://www.google.com/publicdata/directory): Huge collection of different data sources that are public. Seems really nice.
* [Amazon public data sets](https://aws.amazon.com/public-data-sets/): ”AWS hosts a variety of public data sets that anyone can access for free.” Seems interesting.
* [Movielens](https://movielens.org/):  ”Learn more about movies with rich data, images, and trailers. Browse movies by community-applied tags, or apply your own tags. Explore the database with expressive search tools.”  Movie lens is not really a data source in the way that I asked. However, the suggestion was that one could use the movie ratings to learn hadoop/spark/MapReduce. I may give this a try. If I ever get time.
* [Wikipedia ML Datasets](https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research): Wikipedia ML page has various data such as image, text, sound, signal, and so on. For each dataset, you can find its brief description, the data size, format, and machine learning usage. If you are new to deep learning or NLP, you can start exploring the data listed here.