Public Data Sets

Below are links to publicly available data sets and resources.

Datasets are such an integral part of data science and algorithms that it’s almost impossible to talk about our space without talking about data. This is a small but growing collection of links with public data.

Open City Datasets

Palo Alto Open Data
http://www.cityofpaloalto.org/gov/depts/it/open_data/default.asp

Chicago
https://data.cityofchicago.org/

20 yrs crime data
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

NYC
https://nycopendata.socrata.com/

Rents & Neighborhoods
http://www.huduser.org/portal/datasets/HUD_data_matrix.html

Transportation and Travel

Airlines Dataset
http://stat-computing.org/dataexpo/2009/the-data.html

So far it contains years 1987-2007 (based on http://www.stat.purdue.edu/~sguha/rhipe/doc/html/airline.html)

Data source: http://www.transtats.bts.gov/Fields.asp?Table_ID=236

Open flights database
http://openflights.org/data.html

Capital Bikes Share Data
https://www.capitalbikeshare.com/trip-history-data

Sciences and Engineering

Elements Of Statistics Learning Data
http://www-stat.stanford.edu/~tibs/ElemStatLearn/data.html

NASA Open Data
http://data.nasa.gov/

Seismic Data
http://sioseis.ucsd.edu/segy.header.html

Weather Public Data
http://OpenWeatherMap.org http://OpenMeteoData.org

NIST
http://srdata.nist.gov/gateway/gateway?dblist=0

GitHub Archive
http://www.githubarchive.org

Diverse Data Sets

Many Eyes Community Datasets
http://www-958.ibm.com/software/analytics/manyeyes/

Kaggle Competitions
http://www.kaggle.com/

UCI Machine Learning Library
http://archive.ics.uci.edu/ml/datasets.html

Human Activity Recognition Using Smartphones
http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

MLData repository
http://mldata.org/

GitHub Challenge
https://github.com/blog/1450-the-github-data-challenge-ii

Yelp Dataset Challenge
https://www.yelp.com/dataset_challenge

Netflix Prize
http://stackoverflow.com/questions/1407957/netflix-prize-dataset

Infochimps
http://www.infochimps.com/

Stanford Dataset Library
http://snap.stanford.edu/data/index.html

Million Songs Database
http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset

Caret
http://caret.r-forge.r-project.org/datasets.html

RevolutionR
http://www.revolutionanalytics.com/subscriptions/datasets/

Find your favorite dataset!
http://www.inside-r.org/howto/finding-data-internet

LIBSVM Dataset Compilation
http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

The Data Page NYU
http://people.stern.nyu.edu/adamodar/New_Home_Page/data.html

Public Policy Data

European Open Data (6098 datasets!)
http://open-data.europa.eu/en/

US Open Data
http://www.data.gov/ http://www.data.gov/opendatasites

WorldBank Data
http://data.worldbank.org/data-catalog

Guardian Data
http://www.guardian.co.uk/news/datablog/interactive/2013/jan/14/all-our-datasets-index

Statistics Netherlands
http://www.cbs.nl/en-GB/menu/home/default.htm?Languageswitch=on

Quandl 6M Financial, Economics, and Social Datasets
http://www.quandl.com/

Other

http://grouplens.org/datasets/movielens/

Previous

Next