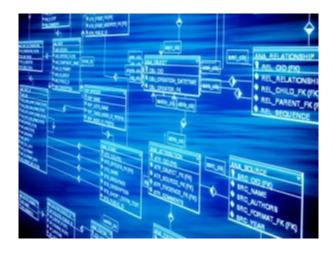
Big Data Sets Available For Free

by Vincent Granville

Dec 30, 2013 (Updated 24 February 2020, Frederic Bien @ GeorgiaTech College of Business)

A few data sets are accessible from our data science apprenticeship web page.



- Source code and data for our Big Data keyword correlation API (see also section in separate chapter, in our book)
- Great statistical analysis: forecasting meteorite hits (see also section in separate chapter, in our book)
- Fast clustering algorithms for massive datasets (see also section in separate chapter, in our book)
- 53.5 billion clicks dataset available for benchmarking and testing
- Over 5,000,000 financial, economic and social datasets
- New pattern to predict stock prices, multiplies return by factor 5 (stock market data, S&P 500; see also section in separate chapter, in our book)
- 3.5 billion web pages: The graph has been extracted from the Common Crawl 2012 web corpus and covers 3.5 billion web pages and 128 billion hyperlinks between these pages
- Another large data set 250 million data points: This is the full resolution GDELT event dataset running January 1, 1979 through March 31, 2013 and containing all data fields for each event record.
- 125 Years of Public Health Data Available for Download

You can find additional data sets at LinkedIn data set. KDNuggets is also a great resource, and for more, check out this link.

Cross-disciplinary data repositories, data collections and data search engines:

- http://usgovxml.com
- http://aws.amazon.com/datasets
- http://databib.org
- http://datacite.org
- http://figshare.com
- http://linkeddata.org
- http://reddit.com/r/datasets
- http://thedatahub.org alias http://ckan.net
- http://guandl.com
- Datasets for Data Mining
- http://enigma.io

Single datasets and data repositories

- http://archive.ics.uci.edu/ml/
- http://crawdad.org/
- http://data.austintexas.gov
- http://data.cityofchicago.org
- http://data.govloop.com
- http://data.gov.uk/
- http://data.medicare.gov
- http://data.seattle.gov
- http://data.sfgov.org
- http://data.sunlightlabs.com
- http://en.wikipedia.org/wiki/Wik...
- http://factfinder.census.gov/ser...
- http://ftp.ncbi.nih.gov/
- http://gettingpastgo.socrata.com
- http://googleresearch.blogspot.c...
- http://books.google.com/ngrams/
- http://medihal.archives-ouvertes.fr
- http://public.resource.org/
- http://rechercheisidore.fr
- http://snap.stanford.edu/data/in...
- http://www2.jpl.nasa.gov/srtm
- http://www.archives.gov/research...
- http://www.bls.gov/
- http://www.crunchbase.com/
- http://www.dartmouthatlas.org/
- http://www.data.gov/
- http://www.datakc.org
- http://dbpedia.org
- http://www.faa.gov/data_research/
- http://www.factual.com/
- http://research.stlouisfed.org/f...
- http://www.freebase.com/
- http://www.google.com/publicdata...

- http://www.guardian.co.uk/news/d...
- http://www.infochimps.com
- http://www.kaggle.com/
- http://build.kiva.org/
- https://opendata.cityofnewyork.us/
- https://moda-nyc.github.io/Project-Library/projects/
- http://www.ordnancesurvey.co.uk/...
- http://www.philwhln.com/how-to-g...
- http://www.imdb.com/interfaces
- https://yandex.ru/q/
- http://www.dados.gov.pt/pt/catal...
- http://knoema.com
- http://daten.berlin.de/
- http://www.gunb.com
- http://databib.org/
- http://datacite.org/
- http://data.reegle.info/
- http://data.wien.gv.at/
- http://data.gov.bc.ca
- https://pslcdatashop.web.cmu.edu/ (interaction data in learning environments)
- https://www.icpsr.umich.edu/icpsrweb/ICPSR/ Inter-university Consortium for Political and Social Research (ICPSR)
- https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/20240 Collaborative Psychiatric Epidemiology Surveys: (A collection of three national surveys focused on each of the major ethnic groups to study psychiatric illnesses and health services use)
- http://www.dati.gov.it
- http://dati.trentino.it