Datasets

Important Websites

- UC Irvine Machine Learning Repository
- Kaggle
- https://github.com/caesar0301/awesome-public-datasets
  Feel free to add unlisted data sets!
- https://github.com/trahasch/awesome-public-datasets
  This is a fork of the dataset above with additional ressources e.g. for predictive maintenance or German weather.

A-Z

- AssetMacro, historical data of Macroeconomic Indicators and Market Data.
- Awesome Public Datasets on github, curated by caesar0301.
- AWS (Amazon Web Services) Public Data Sets, provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications.
- BigML big list of public data sources.
- Bioassay data, described in *Virtual screening of bioassay data*, by Amanda Schierz, J. of Cheminformatics, with 21 Bioassay datasets (Active / Inactive compounds) available for download.
- Bitly 1.usa.gov data, anonymized clicks on gov links.
- Canada Open Data, pilot project with many government and geospatial datasets.
- Causality Workbench data repository.
- Corral Big Data repository at Texas Advanced Computing Center, supporting data-centric science.
- CrowdFlower Data for Everyone library.
- Data Source Handbook, A Guide to Public Data, by Pete Warden, O'Reilly (Jan 2011).
- Datacatalogs.org, open government data from US, EU, Canada, CKAN, and more.
- Data.gov.uk, publicly available data from UK (also London datastore.)
- Data.gov/Education, central guide for education data resources including high-value data sets, data visualization tools, resources for the classroom, applications created from open data and more.
- DataMarket, visualize the world's economy, societies, nature, and industries, with 100 million time series from UN, World Bank, Eurostat and other important data providers.
- Datamob, public data put to good use.
- Data Planet, The largest repository of standardized and structured statistical data, with over 25 billion data points, 4.3 billion datasets, 400+ source databases.

- Datasets.co, datasets for data geeks, find and share Machine Learning datasets.
- DataSF.org, a clearinghouse of datasets available from the City & County of San Francisco, CA.
- DataFerrett, a data mining tool that accesses and manipulates TheDataWeb, a collection of many on-line US Government datasets.
- Datasets publicly available on Google BigQuery
- Delve, Data for Evaluating Learning in Valid Experiments
- DWD Climate Data Center (CDC) - Deutscher Wetterdienst ftp://ftp-cdc.dwd.de/pub/CDC/
- EconData, thousands of economic time series, produced by a number of US Government agencies.
- Enron Email Dataset, data from about 150 users, mostly senior management of Enron.
- Europeana Data, contains open metadata on 20 million texts, images, videos and sounds gathered by Europeana - the trusted and comprehensive resource for European cultural heritage content.
- FEDSTATS, a comprehensive source of US statistics and more
- FIMI repository for frequent itemset mining, implementations and datasets.
- Financial Data Finder at OSU, a large catalog of financial data sets.
- GDELT: The Global Data on Events, Location and Tone, described by Guardian as "a big data history of life, the universe and everything."
- GEO (GEO Gene Expression Omnibus), a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.
- GeoDa Center, geographical and spatial data.
- Google ngrams datasets, text from millions of books scanned by Google.
- Grain Market Research, financial data including stocks, futures, etc.
- Hilary Mason research-quality Big Data sets collection - many text and image datasets.
- HitCompanies Datasets, comprehensive data on random 10,000 UK companies sampled from HitCompanies, updated automatically using AI/Machine Learning.
- ICWSM-2009 dataset contains 44 million blog posts made between August 1st and October 1st, 2008.
- Infochimps, an open catalog and marketplace for data. You can share, sell, curate, and download data about anything and everything.
- Investor Links, includes financial data
- Kaggle Datasets.

- KDD Cup center, with all data, tasks, and results.
- Kevin Chai list of datasets, for text, SNA, and other fields.
- KONECT, the Koblenz Network Collection, with large network datasets of all types in order to perform research in the area of network mining.
- Lending Club Loan Statistics
- Linking Open Data project, at making data freely available to everyone.
- Million Song Dataset
- MIT Cancer Genomics gene expression datasets and publications, from MIT Whitehead Center for Genome Research.
- ML Data, the data repository of the EU Pascal2 networks.
- MovieLens
- NASDAQ Data Store, provides access to market data.
- National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including countries from Africa, Europe, Asia, and Latin America.
- National Space Science Data Center (NSSDC), NASA data sets from planetary exploration, space and solar physics, life sciences, astrophysics, and more.
- NetworkRepository: Interactive Data Repository, has many collections of graph and networks from social science, machine learning, scientific computing, and other areas.
- Open Data Census, assesses the state of open data around the world.
- OpenData from Socrata, access to over 10,000 datasets including business, education, government, and fun.
- Open Source Sports, many sports databases, including Baseball, Football, Basketball, and Hockey.
- Peter Skomoroch dataset Bookmarks
- PubGene(TM) Gene Database and Tools, genomic-related publications database
- Quandl, a collaboratively curated portal to millions of financial and economic time-series datasets.
- qunb, a platform to find and visualize quantitative data.
- Robert Schiller data on housing, stock market, and more from his book *Irrational Exuberance*.
- SMD: Stanford Microarray Database, stores raw and normalized data from microarray experiments.
- Jerry Smith dataset collection, with Finance, Government, Machine Learning, Science, and other data.
- SourceForge.net Research Data, includes historic and status statistics on approximately 100,000 projects and over 1 million registered users' activities at the project management web site.
- StatLib, CMU Datasets Archive.

- Stanford Large Network Dataset Collection
- The Data Expo of the Joint Statistical Meetings, the Graphics Section and the Computing Section
- The Humanitarian Data Exchange
- Time Series Data Library
- Visual Analytics Benchmark Repository
- UCI Machine Learning Repository for large datasets used in machine learning and knowledge discovery research.
- UCR Time Series Data Archive, offering datasets, papers, links, and code.
- UK Open Postcode Geo, UK/British postcodes with easting, northing, latitude, and longitude.
- United States Census Bureau.
- Web Data Commons, structured data from the Common Crawl, the largest public web corpus.
- Wikipedia:Database download
- Wikiposit, a (virtual) amalgamation of (mostly financial) data from many different sites, allowing users to merge data from different sources
- Wolfram Alpha disease and patient level data.
- Yahoo Sandbox datasets, Language, Graph, Ratings, Advertising and Marketing, Competition
- Yelp Academic Dataset, all the data and reviews of the 250 closest businesses for 30 universities for students and academics to explore and research.