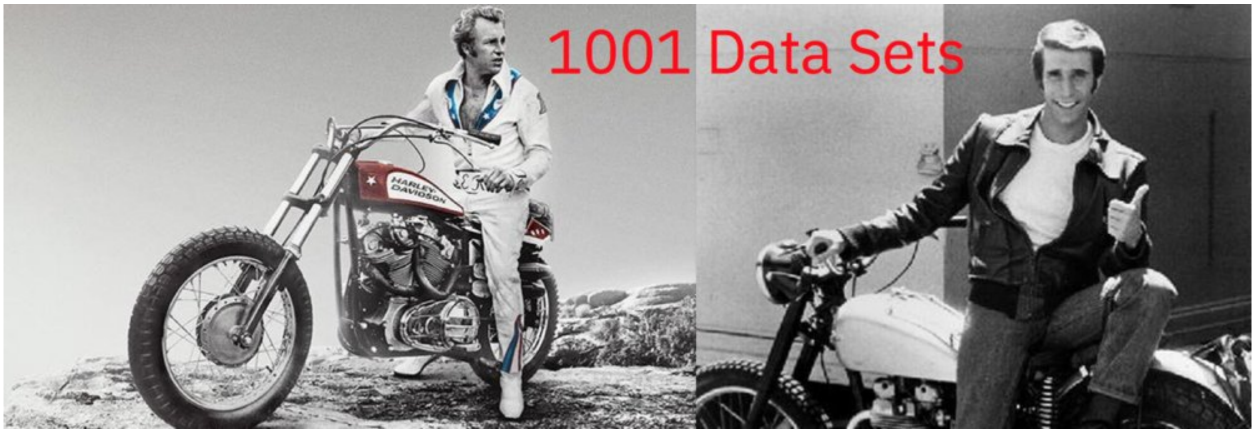
dreamtolearn
becoming, together.

**Dream to Learn is shutting down...** We are very sorry to say that Dream to Learn will be shutting down as of December 28th, 2019. If you have content that you wish to keep, you should make a copy of it before that date.

**1001 Datasets and Data repositories**
**( List of lists of lists )**



**This is a LIST of.... "lists of lists".** ***Messy presentation to pull together Raw Datasets for my hacks. Suggestions to add?*** **Message me or post comment..**

**CTRL-F to "FIND" is your best bet - e.g. CTRL-F "food" or "population"**

100% of the links below are from external sources (not mine)

**Source: IBM Data Asset eXchange Explore useful data sets for enterprise data science**

https://developer.ibm.com/exchanges/data/

**Source: IBM Model Asset eXchange Free, deployable, and trainable code.**

A place for developers to find and use free and open source deep learning models.

https://developer.ibm.com/exchanges/models/

- Audio Classification
- Audio Feature Extraction
- Audio Modeling
- Facial Recognition
- Image Classification

- Image Feature Extraction

- Image-to-Image Translation or Transformation
- Image-to-Text Translation
- Language Modeling
- Named Entity Recognition

- Natural Language Processing
- Object Detection in Images
- Security
- Text Classification
- Text Feature Extraction

- Text-to-Image Translation
- Time Series Prediction
- Video Classification

**Source: Quora**

https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public

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

1. http://datasource.kapsarc.org
2. https://www.kaggle.com/datasets
3. http://www.assetmacro.com
4. http://usgovxml.com
5. http://aws.amazon.com/datasets
6. http://databib.org
7. http://datacite.org
8. http://figshare.com
9. http://linkeddata.org
10. http://reddit.com/r/datasets
11. http://thewebminer.com/
12. http://thedatahub.org alias http://ckan.net
13. http://quandl.com
14. Social Network Analysis Interactive Dataset Library (Social Network Datasets)
15. Datasets for Data Mining
16. Enigma Public
17. *http://www.ufindthem.com/*
18. http://NetworkRepository.com - The First Interactive Network Data Repository
19. http://MLvis.com
20. Open Data Inception - A Comprehensive List of 2500+ Open Data Portals in the World
21. http://data.opendatasoft.com OpenDataSoft catalog

**Single datasets and data repositories**

1. http://archive.ics.uci.edu/ml/
2. http://crawdad.org/
3. http://data.austintexas.gov
4. http://data.cityofchicago.org
5. http://data.govloop.com
6. http://data.gov.uk/
7. data.gov.in
8. http://data.medicare.gov

9. http://data.seattle.gov
10. http://data.sfgov.org
11. http://data.sunlightlabs.com
12. https://datamarket.azure.com/
13. http://developer.yahoo.com/geo/g...
14. http://econ.worldbank.org/datasets
15. http://en.wikipedia.org/wiki/Wik...
16. http://factfinder.census.gov/ser...
17. http://ftp.ncbi.nih.gov/
18. http://gettingpastgo.socrata.com
19. http://googleresearch.blogspot.c...
20. http://books.google.com/ngrams/
21. http://medihal.archives-ouvertes.fr
22. http://public.resource.org/
23. http://rechercheisidore.fr
24. http://snap.stanford.edu/data/in...
25. http://timetric.com/public-data/
26. https://wist.echo.nasa.gov/~wist...
27. http://www2.jpl.nasa.gov/srtm
28. http://www.archives.gov/research...
29. http://www.bls.gov/
30. http://www.crunchbase.com/
31. http://www.dartmouthatlas.org/
32. http://www.data.gov/
33. http://www.datakc.org
34. http://dbpedia.org
35. http://www.delicious.com/jbaldwi...
36. http://www.faa.gov/data_research/
37. http://www.factual.com/
38. http://research.stlouisfed.org/f...
39. http://www.freebase.com/
40. http://www.google.com/publicdata...
41. http://www.guardian.co.uk/news/d...
42. http://www.infochimps.com
43. http://www.kaggle.com/
44. http://build.kiva.org/
45. http://www.nationalarchives.gov....
46. http://www.nyc.gov/html/datamine...
47. http://www.ordnancesurvey.co.uk/...
48. http://www.philwhln.com/how-to-g...
49. http://www.imdb.com/interfaces
50. http://imat-relpred.yandex.ru/en...
51. http://www.dados.gov.pt/pt/catal...
52. http://knoema.com
53. http://daten.berlin.de/
54. http://www.qunb.com
55. http://databib.org/
56. http://datacite.org/
57. http://data.reegle.info/
58. http://data.wien.gv.at/
59. http://data.gov.bc.ca
60. https://pslcdatashop.web.cmu.edu/ (interaction data in learning environments)
61. http://www.icpsr.umich.edu/icpsrweb/CPES/ - 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)
62. http://www.dati.gov.it
63. http://dati.trentino.it
64. http://www.databagg.com/
65. http://networkrepository.com - Network/ML data repository w/ visual interactive analytics

66. Home (United Nations Environment Programme Grid Genava a lot of GIS datasets

**Source: Quora - Alan Morrison PWC**

https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public

Agriculture

- U.S. Department of Agriculture's PLANTS Database
-
- **Biology**
- 1000 Genomes
- Collaborative Research in Computational Neuroscience (CRCNS)
- Gene Expression Omnibus (GEO)
- Human Microbiome Project (HMP)
- ICOS PSP Benchmark
- MIT Cancer Genomics Data
- NIH Microarray data (FTP)
- Protein Data Bank
- PubChem Project
- PubGene (now Coremine Medical)
- Stanford Microarray Data
- The Personal Genome Project or PGP
- UCSC Public Data
- UniGene
- **Climate/Weather**
- Australian Weather
- Canadian Meteorological Centre
- Climate Data from UEA (updated monthly)
- Global Climate Data Since 1929
- NOAA Bering Sea Climate
- NOAA Climate Datasets
- NOAA Realtime Weather Models
- WU Historical Weather Worldwide
- **Complex Networks**
- CrossRef DOI URLs
- DBLP Citation dataset
- NBER Patent Citations
- NIST complex networks data collection
- Protein-protein interaction network
- PyPI and Maven Dependency Network
- Scopus Citation Database
- Stanford GraphBase (Steven Skiena)
- Stanford Large Network Dataset Collection
- The Koblenz Network Collection
- The Laboratory for Web Algorithmics (UNIMI)
- UCI Network Data Repository
- UFL sparse matrix collection
- WSU Graph Database

- **Computer Networks**
- 3.5B Web Pages from CommonCraw 2012
- 53.5B Web clicks of 100K users in Indiana Univ.
- CAIDA Internet Datasets
- ClueWeb09 - 1B web pages
- ClueWeb12 - 733M web pages
- CommonCrawl Web Data over 7 years
- CRAWDAD Wireless datasets from Dartmouth Univ.
- Open Mobile Data by MobiPerf
- UCSD Network Telescope, IPv4 /8 net
- **Data Challenges**
- Challenges in Machine Learning
- DrivenData Competitions for Social Good
- ICWSM Data Challenge (since 2009)
- Kaggle Competition Data
- KDD Cup by Tencent 2012
- Localytics Data Visualization Challenge
- Netflix Prize
- Yelp Dataset Challenge
- **Economics**
- American Economic Ass (AEA)
- EconData from UMD
- Internet Product Code Database
- **Energy**
- AMPds
- BLUEd
- COMBED
- Dataport
- ECO
- EIA
- HFED
- iAWE
- Plaid
- REDD
- UK-Dale
- **Finance**
- CBOE Futures Exchange
- Google Finance
- Google Trends
- NASDAQ
- OANDA
- OSU Financial data
- Quandl
- St Louis Federal
- Yahoo Finance
- **GeoSpace/GIS**
- BODC - marine data of ~22K vars
- EOSDIS - NASA's earth observing system data
- Factual Global Location Data
- Global Administrative Areas Database (GADM)
- Geo Spatial Data from ASU
- GeoNames Worldwide
- Natural Earth - vectors and rasters of the world

- Open Street Map (OSM)
- TIGER/Line - U.S. boundaries and roads
- TwoFishes - Foursquare's coarse geocoder
- TZ Timezones shapfiles

Government

- Australia (abs.gov.au)
- Australia (data.gov.au)
- Canada
- Chicago
- EuroStat
- FedStats
- Germany
- Glasgow, Scotland, UK
- Guardian world governments
- London Datastore, UK
- MassGIS, Massachusetts, U.S.
- Netherlands
- New Zealand
- NYC betanyc
- NYC Open Data
- OECD
- Open Government Data (OGD) Platform India
- San Francisco Data sets
- South Africa
- The World Bank
- U.K. Government Data
- U.S. American Community Survey
- U.S. CDC Public Health datasets
- U.S. Census Bureau
- U.S. Department of Housing and Urban Development (HUD)
- U.S. Federal Government Agencies
- U.S. Federal Government Data Catalog
- U.S. Food and Drug Administration (FDA)
- U.S. Open Government
- UK 2011 Census Open Atlas Project
- United Nations

Healthcare

- EHDP Large Health Data Sets
- Gapminder World, demographic databases
- Medicare Coverage Database (MCD), U.S.
- Medicare Data Engine of medicare.gov Data
- Medicare Data File

Image Processing

- 2GB of Photos of Cats
- Face Recognition Benchmark
- ImageNet - an image database in WordNet hierarchy

Machine Learning

- Delve Datasets for classification and regression (Univ. of Toronto)
- Discogs Monthly Data
- eBay Online Auctions (2012)
- IMDb Database
- Keel Repository for classification, regression and time series
- Lending Club Loan Data
- Machine Learning Data Set Repository
- Million Song Dataset
- More Song Datasets
- MovieLens Data Sets
- RDataMining - "R and Data Mining" ebook data
- Registered Meteorites on Earth
- Restaurants Health Score Data in San Francisco
- UCI Machine Learning Repository
- Yahoo! Ratings and Classification Data

Museums

- Cooper-Hewitt's Collection Database
- Minneapolis Institute of Arts metadata
- Tate Collection metadata
- The Getty vocabularies

Natural Language

- ClueWeb09 FACC
- ClueWeb12 FACC
- DBpedia - 4.58M things with 583M facts
- Flickr Personal Taxonomies
- Google Books Ngrams (2.2TB)
- Google Web 5gram (1TB, 2006)
- Gutenberg eBooks List
- Hansards text chunks of Canadian Parliament
- Machine Translation of European languages
- SMS Spam Collection in English
- USENET postings corpus of 2005~2011
- Wikidata - Wikipedia databases
- Wikipedia Links data - 40 Million Entities in Context
- WordNet databases and tools

Physics

- CERN Open Data Portal
- NSSDC (NASA) data of 550 space spacecraft

Public Domains

- Amazon
- Archive.org Datasets
- CMU JASA data archive

- CMU StatLab collections
- Data360
- Datamob.org
- Google
- Infochimps
- KDNuggets Data Collections
- Numbray
- Reddit Datasets
- RevolutionAnalytics Collection
- Sample R data sets
- Stats4Stem R data sets
- StatSci.org
- The Washington Post List
- UCLA SOCR data collection
- UFO Reports
- Wikileaks 911 pager intercepts
- Yahoo Webscope

Search Engines

- Academic Torrents of data sharing from UMB
- Archive-it from Internet Archive
- Datahub.io
- DataMarket (Qlik)
- Freebase.com of people, places, and things
- Harvard Dataverse Network of scientific data
- ICPSR (UMICH)
- Statista.com - statistics and Studies

Social Sciences

- Ancestry.com Forum Dataset over 10 years
- CMU Enron Email of 150 users
- Facebook Data Scrape (2005)
- Facebook Social Networks from LAW (since 2007)
- Foursquare Social Network in 2010, 2011
- Foursquare from UMN/Sarwat (2013)
- General Social Survey (GSS) since 1972
- GetGlue - users rating TV shows
- GitHub Collaboration Archive
- Mobile Social Networks from UMASS
- PewResearch Internet Survey Project
- SourceForge.net Research Data
- StackExchange Data Explorer
- Titanic Survival Data Set
- Twitter Graph of entire Twitter site
- UCB's Archive of Social Science Data (D-Lab)
- UCLA Social Sciences Data Archive
- UNIMI/LAW Social Network Datasets
- Universities Worldwide
- UPJOHN for Labor Employment Research
- Yahoo! Graph and Social Data
- Youtube Video Social Graph in 2007,2008

Sports

- Betfair Historical Exchange Data
- Cricsheet Matches (baseball)
- Ergast Formula 1, from 1950 up to date (API)
- Football/Soccer resouces (data and APIs)
- Lahman's Baseball Database
- Retrosheet Baseball Statistics

Time Series

- Time Series Data Library (TSDL) from MU
- UC Riverside Time Series Dataset

Transportation

- Airlines OD Data 1987-2008
- Bike Share Systems (BSS) collection
- Hubway Million Rides in MA
- Marine Traffic - ship tracks, port calls and more
- NYC Taxi Trip Data 2013 (FOIA/FOILed)
- OpenFlights - airport, airline and route data
- RITA Airline On-Time Performance data
- RITA/BTS transport data collection (TranStat)
- Transport for London (TFL)
- Travel Tracker Survey (TTS) for Chicago
- U.S. Bureau of Transportation Statistics (BTS)
- U.S. Domestic Flights 1990 to 2009
- U.S. Freight Analysis Framework since 2007

Complementary Collections

- DataWrangling: Some Datasets Available on the Web
- Inside-r: Finding Data on the Internet
- Quora: Where can I find large datasets open to the public?
- like being punched in the brain!: 100+ Interesting Data Sets for Statistics
- StaTrek: Leveraging open data to understand urban lives"

Source: Xiaming's Github caesar0301/awesome-public-datasets, January 2015. Please go to Github for this and other updated lists.

15.9k views · View 25 Upvoters · Answer requested by Kapil Vinja

**Related QuestionsMore Answers Below**

- Where can I find large datasets open to the public csv easy?
- Where can I find large datasets open to the public in excel?
- Where can I find large medical datasets open to the public?
- Where can I find large datasets open to the public on sleep?
- Where can I find large Indian datasets open to the pub

**SOURCE - ERIK HILLE - SMU**

**International Historical Statistics (by Brian Mitchell)**

- **Data:** Aggregate trade (current value), bilateral trade with main trading partners (current value), and major commodity exports by main exporting countries. No data on trade as share of GDP is readily available.
- **Geographical coverage:** Countries around the world
- **Time span:** Long time series with annual observations – from 19th century up to today (2010)
- **Available at:** The books are published in three volumes covering more than 5000 pages. 11 At some universities you can access the online version of the books where data tables can be downloaded as ePDFs and Excel files. The online access ishere.
- *Data from the 19th century onwards for countries around the world is available in the International Historical Statistics (IHS). These statistics – originally published under the editorial leadership of Brian Mitchell (since 1983) – are a collection of data sets taken from many primary sources, including both official national and international abstracts.*

**Penn World Tables**

- **Data:** Real and PPP-adjusted GDP in US millions of dollars, national accounts (household consumption, investment, government consumption, exports and imports), exchange rates and population figures.
- **Geographical coverage:** Countries around the world
- **Time span:** from 1950-2011 (version 8.1)
- **Available at:** Online here
- *Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), "The Next Generation of the Penn World Table" forthcoming American Economic Review, available for download at www.ggdc.net/pwt*

**Correlates of War Bilateral Trade**

- **Data:** Total national trade and bilateral trade flows between states. Total imports and exports of each country in current US millions of dollars and bilateral flows in current US millions of dollars
- **Geographical coverage:** Single countries around the world
- **Time span:** from 1870-2009
- **Available at:** Online at www.correlatesofwar.org
- *This data set is hosted by Katherine Barbieri, University of South Carolina, and Omar Keshk, Ohio State University.*

**World Bank – World Development Indicators**

- **Data:** Trade (% of GDP) and many more specific series: trade in merchandise, trade in services, trade in high-technology, trade in ICT goods, trade in ICT services – always exports and imports separately. Also export and import value index and volume index.
- **Geographical coverage:** Countries and world regions
- **Time span:** Annual since 1960
- **Available at:** Online at http://data.worldbank.org

### UN Comtrade

- **Data:** Bilateral trade flows by commodity
- **Geographical coverage:** Countries around the world
- **Time span:** 1962-2013
- **Available at:** Online here

### UNCTADstat

- **Data:** Many different measures, including trade by volumes and value
- **Geographical coverage:** Countries around the world
- **Time span:** For some series, data is available since 1948 – mostly annual, sometimes quarterly.
- **Available at:** Online here

### Eurostat – COMEXT

- **Data:** Trade flows (also by commodity)
- **Geographical coverage:** Europe (EU and EFTA)
- **Time span:** Mostly since 1988
- **Available at:** Online here
- *Also, the Eurostat website 'Statistics Explained' publishes up-to-date statistical information on international trade in goodsand services.*

### World Trade Organization – WTO

- **Data:** Many series on tariffs and trade flows
- **Geographical coverage:** Countries around the world
- **Time span:** Since 1948 for some series
- **Available at:** Online here

### CEPII database on the World Economy

- **Data:** Many different data sets related to international trade, including trade flows by commodity geographical variables, and variables to estimate gravity models
- **Geographical coverage:** Countries around the world
- **Time span:** Some series go back to the 1990s.
- **Available at:** Online here

### NBER-United Nations Trade Data, 1962-2000

- **Data:** Export and import values and volumes by commodity
- **Geographical coverage:** Single countries
- **Time span:** 1962-2000
- **Available at:** Online here
- *This data is also available from the Center for International Data.*

### Smaller historical trade data sets

- Data on **UK bilateral trade** for the time 1870-1913 was collected by David S. Jacks. It is downloadable in excel format here.
- For the time **1870-1913** 21,000 bilateral trade observations can be found in Mitchener and Weidenmier (2008) – Trade and empire, available in the Economic Journal here.

- Data on **UK, Germany, France, and US** between mid-19th to 20th Century can be found here.
- Data on **Developing Country Export** – in 1840, 1860, 1880 and 1900 – by John Hanson is available here.
- Data on **trade between England and Africa** during the period 1699-1808 is available on the Dutch Data Archiving and Networked Services. It was compiled by Marion Johnson.

Applying these same sources to Education quality in developing countries:

- **Education Index** multiple sheets of excel **data** is available at **Human Development Reports** or you can use their tool to explore the data Human Development Reportsalso google has access to explore the data Google Public Data Explorer additional indexes in this HD report that you might be interest in are: Human Development Indexand Adult Literacy Index and Gross enrollment ratio.
- The World Bank has Literacy rates Adult literacy rate, population 15+ years, both sexes (%)in addition to lots of other data: World Bank Data. Lots of years. Lots of CountriesCountries | Data. Lots of data variables Topics | Data - Indicators | Data - Catalog | The World Bank.
- Our government also likes to stay informed and is willing to share some of that data: CIA -The World Factbook
- Possibly looking at the Human Capital Report 2015 has Rankings of human capital index has various measures of education and productivity capabilities.
- Unveiling the beauty of statistics for a fact based world view. - (http://www.gapminder.org/)
- Data Plotter - has Average Test Scores
- **Penn World Tables - Data:** Real and PPP-adjusted GDP in US millions of dollars, national accounts (household consumption, investment, government consumption, exports and imports), exchange rates and population figures. *Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), "The Next Generation of the Penn World Table" forthcoming American Economic Review, available for download at www.ggdc.net/pwt*

**Source: Tableu - How to find the best sources for free, public data sets**

https://www.tableau.com/about/blog/2019/2/public-data-sets-102221

- FiveThirtyEight - A goldmine of over 100 data sets on sports and politics. Examples: March Madness predictions, political polling, the Bachelorette show, etc.
- The Pudding - This data journalism website aims to explain hotly-debated cultural events with visual essays, sourced from original data sets and primary research. Their GitHub is a hub for pop culture data. Examples: Women's vs. men's pants pockets, weather conditions on Mars, etc.
- Buzzfeed - If you know Buzzfeed, you know that their news site covers a variety of topics in politics, sports, and current events. They also have a rich list of data sets on GitHub. Examples: Trump's tweets, the text of every State of the Union address, etc.
- Washington Post - The Washington Post is a respected news source and their list of open data sets contains topics like NCAA financials and transportation data. Examples: School shootings, police shootings, NFL arrests, etc.
- Viz for Social Good - A hackathon style project that connects the community with non-profit organizations. Examples: Advocating for fatherless boys in Africa, increasing awareness of child refugees, supporting black male entrepreneurs.
- Makeover Monday - A weekly, social-data project to create a discussion around improving data visualizations. Each Sunday, the team posts a link to a visualization and a data set. Your challenge is to create a better version of the visualization in your own creative way. Their weekly data sets are diverse and stay on the site for reuse, so it is a great place to start in your search for clean data. Examples: Wind energy by state, minimum wage, NHL attendance.

- Sports Viz Sunday - A community-led project to create, share, and promote visualizations from the world of sports. Sports Viz Sunday hosts a monthly challenge based on a topical sports theme, regularly sharing updates from the sports visualization world and providing rich data sets across a wide range of sports. Examples: World Cup, the Masters, Formula 1 racing.
- Iron Quest - A project aimed at preparing people for Iron Viz qualifier competitions, offering opportunities to practice finding your own data sets.
- Twitter data - Twitter has an API that allows you to get data about hashtags, keywords, or accounts. Here's a guide on how to connect to Twitter data directly in Tableau. If you're more comfortable working with APIs, you can query to get JSON data, which is a supported data type in Tableau. Here is the complete API documentation. Visualization example: Pulse of Super Bowl LIII.
- **Netflix data** - Download your viewing data by going to netflix.com/viewingactivity. Visualization example: I have created a dashboard that compares people's binges and visualizes Netflix viewing activity over time.
- **Spotify streaming data** - Did you know that you can request your personal listening data from Spotify? If you are familiar working with APIs, you can use the Spotify Web API to get data about music artists, albums, and tracks, directly from the Spotify Data Catalogue.
- Others
    - Kaggle
    - Data.world
    - Data.gov
    - Google dataset search
    - r/datasets

**QuickDraw**

- https://quickdraw.withgoogle.com/data (source: https://quickdraw.withgoogle.com/# )

**Source: Medium: The 50 Best Public Datasets for Machine Learning**

"What are some open datasets for machine learning? After scrapping the web for hours after hours, we have created a great cheat sheet for high quality and diverse machine learning datasets.

**https://medium.com/datadriveninvestor/the-50-best-public-datasets-for-machine-learning-d80e9f030279**

**Dataset Finders**

Kaggle: A data science site that contains a variety of externally contributed interesting datasets. You can find all kinds of niche datasets in its master list, from ramen ratings to basketball data to and even seattle pet licenses.

UCI Machine Learning Repository: One of the oldest sources of datasets on the web, and a great first stop when looking for interesting datasets. Although the data sets are user-contributed, and thus have varying levels of cleanliness, the vast majority are clean. You can download data directly from the UCI Machine Learning repository, without registration.

VisualData: Discover computer vision datasets by category, it allows searchable queries.

**General Datasets**

**Public Government datasets**

Data.gov: This site makes it possible to download data from multiple US government agencies. Data can range from government budgets to school performance scores. Be warned though: much of the data requires additional research.

Food Environment Atlas: Contains data on how local food choices affect diet in the US.

School system finances: A survey of the finances of school systems in the US.

Chronic disease data: Data on chronic disease indicators in areas across the US.

The US National Center for Education Statistics: Data on educational institutions and education demographics from the US and around the world.

The UK Data Service: The UK's largest collection of social, economic and population data.

Data USA: A comprehensive visualization of US public data.

**Finance & Economics**

Quandl: A good source for economic and financial data — useful for building models to predict economic indicators or stock prices.

World Bank Open Data: Datasets covering population demographics, a huge number of economic, and development indicators from across the world.

IMF Data: The International Monetary Fund publishes data on international finances, debt rates, foreign exchange reserves, commodity prices and investments.

Financial Times Market Data: Up to date information on financial markets from around the world, including stock price indexes, commodities and foreign exchange.

Google Trends: Examine and analyze data on internet search activity and trending news stories around the world.

American Economic Association (AEA): A good source to find US macroeconomic data.

**Machine Learning Datasets:**

**Images**

Labelme: A large dataset of annotated images.

ImageNet: The de-facto image dataset for new algorithms, organized according to the WordNet hierarchy, in which hundreds and thousands of images depict each node of the hierarchy.

LSUN: Scene understanding with many ancillary tasks (room layout estimation, saliency prediction, etc.)

MS COCO: Generic image understanding and captioning.

COIL100 : 100 different objects imaged at every angle in a 360 rotation.

Visual Genome: Very detailed visual knowledge base with captioning of ~100K images.

Google's Open Images: A collection of 9 million URLs to images "that have been annotated with labels spanning over 6,000 categories" under Creative Commons.

Labelled Faces in the Wild: 13,000 labeled images of human faces, for use in developing applications that involve facial recognition.

Stanford Dogs Dataset: Contains 20,580 images and 120 different dog breed categories.

Indoor Scene Recognition: A very specific dataset and very useful, as most scene recognition models are better 'outside'. Contains 67 Indoor categories, and 15620 images.

**Sentiment Analysis**

Multidomain sentiment analysis dataset: A slightly older dataset that features product reviews from Amazon.

IMDB reviews: An older, relatively small dataset for binary sentiment classification features 25,000 movie reviews.

Stanford Sentiment Treebank: Standard sentiment dataset with sentiment annotations.

Sentiment140: A popular dataset, which uses 160,000 tweets with emoticons pre-removed.

Twitter US Airline Sentiment: Twitter data on US airlines from February 2015, classified as positive, negative, and neutral tweets

**Natural Language Processing**

HotspotQA Dataset: Question answering dataset featuring natural, multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems.

Enron Dataset: Email data from the senior management of Enron, organized into folders.

Amazon Reviews: Contains around 35 million reviews from Amazon spanning 18 years. Data include product and user information, ratings, and the plaintext review.

Google Books Ngrams: A collection of words from Google books.

Blogger Corpus: A collection 681,288-blog posts gathered from blogger.com. Each blog contains a minimum of 200 occurrences of commonly used English words.

Wikipedia Links data: The full text of Wikipedia. The dataset contains almost 1.9 billion words from more than 4 million articles. You can search by word, phrase or part of a paragraph itself.

Gutenberg eBooks List: Annotated list of ebooks from Project Gutenberg.

Hansards text chunks of Canadian Parliament: 1.3 million pairs of texts from the records of the 36th Canadian Parliament.

Jeopardy: Archive of more than 200,000 questions from the quiz show Jeopardy.

SMS Spam Collection in English: A dataset that consists of 5,574 English SMS spam messages

Yelp Reviews: An open dataset released by Yelp, contains more than 5 million reviews.

UCI's Spambase: A large spam email dataset, useful for spam filtering.

**Self-driving**

Berkeley DeepDrive BDD100k: Currently the largest dataset for self-driving AI. Contains over 100,000 videos of over 1,100-hour driving experiences across different times of the day and weather conditions. The annotated images come from New York and San Francisco areas.

Baidu Apolloscapes: Large dataset that defines 26 different semantic items such as cars, bicycles, pedestrians, buildings, streetlights, etc.

Comma.ai: More than 7 hours of highway driving. Details include car's speed, acceleration, steering angle, and GPS coordinates.

Oxford's Robotic Car: Over 100 repetitions of the same route through Oxford, UK, captured over a period of a year. The dataset captures different combinations of weather, traffic and pedestrians, along with long-term changes such as construction and roadworks.

Cityscape Dataset: A large dataset that records urban street scenes in 50 different cities.

CSSAD Dataset: This dataset is useful for perception and navigation of autonomous vehicles. The dataset skews heavily on roads found in the developed world.

KUL Belgium Traffic Sign Dataset: More than 10000+ traffic sign annotations from thousands of physically distinct traffic signs in the Flanders region in Belgium.

MIT AGE Lab: A sample of the 1,000+ hours of multi-sensor driving datasets collected at AgeLab.

LISA: Laboratory for Intelligent & Safe Automobiles, UC San Diego Datasets: This dataset includes traffic signs, vehicles detection, traffic lights, and trajectory patterns.

Bosch Small Traffic Light Dataset: Dataset for small traffic lights for deep learning.

LaRa Traffic Light Recognition: Another dataset for traffic lights. This is taken in Paris.

WPI datasets: Datasets for traffic lights, pedestrian and lane detection.

Clinical

MIMIC-III: Openly available dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data associated with ~40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more.

**Source: GeoPlatform Data.gov Search - Geospatial Platform**

The GeoPlatform provides shared and trusted geospatial data, services, and applications for use by the public and by government agencies and partners to meet their mission needs.

- https://data.geoplatform.gov/

- https://www.geoplatform.gov/
- https://www.fgdc.gov/dataandservices - shared and trusted geospatial data, services, and applications.

**NLP Datasets - Source: Niderhoff Github nlp-datasets**

https://github.com/niderhoff/nlp-datasets  Alphabetical list of free/public domain datasets with text data for use in Natural Language Processing (NLP). Most stuff here is just raw unstructured text data, if you are looking for annotated corpora or Treebanks refer to the sources at the bottom.

- Apache Software Foundation Public Mail Archives: all publicly available Apache Software Foundation mail archives as of July 11, 2011 (200 GB)
- Blog Authorship Corpus: consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. 681,288 posts and over 140 million words. (298 MB)
- Amazon Fine Food Reviews [Kaggle]: consists of 568,454 food reviews Amazon users left up to October 2012. Paper. (240 MB)
- Amazon Reviews: Stanford collection of 35 million amazon reviews. (11 GB)
- ArXiv: All the Papers on archive as fulltext (270 GB) + sourcefiles (190 GB).
- ASAP Automated Essay Scoring [Kaggle]: For this competition, there are eight essay sets. Each of the sets of essays was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students ranging in grade levels from Grade 7 to Grade 10. All essays were hand graded and were double-scored. (100 MB)
- ASAP Short Answer Scoring [Kaggle]: Each of the data sets was generated from a single prompt. Selected responses have an average length of 50 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students primarily in Grade 10. All responses were hand graded and were double-scored. (35 MB)
- Classification of political social media: Social media messages from politicians classified by content. (4 MB)
- CLiPS Stylometry Investigation (CSI) Corpus: a yearly expanded corpus of student texts in two genres: essays and reviews. The purpose of this corpus lies primarily in stylometric research, but other applications are possible. (on request)
- ClueWeb09 FACC: ClueWeb09 with Freebase annotations (72 GB)
- ClueWeb11 FACC: ClueWeb11 with Freebase annotations (92 GB)
- Common Crawl Corpus: web crawl data composed of over 5 billion web pages (541 TB)
- Cornell Movie Dialog Corpus: contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts: 220,579 conversational exchanges between 10,292 pairs of movie characters, 617 movies (9.5 MB)
- Corporate messaging: A data categorization job concerning what corporations actually talk about on social media. Contributors were asked to classify statements as information (objective statements about the company or it’s activities), dialog (replies to users, etc.), or action (messages that ask for votes or ask users to click on links, etc.). (600 KB)
- Crosswikis: English-phrase-to-associated-Wikipedia-article database. Paper. (11 GB)
- DBpedia: a community effort to extract structured information from Wikipedia and to make this information available on the Web (17 GB)
- Death Row: last words of every inmate executed since 1984 online (HTML table)
- Del.icio.us: 1.25 million bookmarks on delicious.com
- Disasters on social media: 10,000 tweets with annotations whether the tweet referred to a disaster event (2 MB).
- Economic News Article Tone and Relevance: News articles judged if relevant to the US economy and, if so, what the tone of the article was. Dates range from 1951 to 2014. (12 MB)
- Enron Email Data: consists of 1,227,255 emails with 493,384 attachments covering 151 custodians (210 GB)
- Event Registry: Free tool that gives real time access to news articles by 100.000 news publishers worldwide. Has API. (query tool)

- Examiner.com - Spam Clickbait News Headlines [Kaggle]: 3 Million crowdsourced News headlines published by now defunct clickbait website The Examiner from 2010 to 2015. (200 MB)
- Federal Contracts from the Federal Procurement Data Center (USASpending.gov): data dump of all federal contracts from the Federal Procurement Data Center found at USASpending.gov (180 GB)
- Flickr Personal Taxonomies: Tree dataset of personal tags (40 MB)
- Freebase Data Dump: data dump of all the current facts and assertions in Freebase (26 GB)
- Freebase Simple Topic Dump: data dump of the basic identifying facts about every topic in Freebase (5 GB)
- Freebase Quad Dump: data dump of all the current facts and assertions in Freebase (35 GB)
- GigaOM Wordpress Challenge [Kaggle]: blog posts, meta data, user likes (1.5 GB)
- Google Books Ngrams: available also in hadoop format on amazon s3 (2.2 TB)
- Google Web 5gram: contains English word n-grams and their observed frequency counts (24 GB)
- Gutenberg Ebook List: annotated list of ebooks (2 MB)
- Hansards text chunks of Canadian Parliament: 1.3 million pairs of aligned text chunks (sentences or smaller fragments) from the official records (Hansards) of the 36th Canadian Parliament. (82 MB)
- Harvard Library: over 12 million bibliographic records for materials held by the Harvard Library, including books, journals, electronic resources, manuscripts, archival materials, scores, audio, video and other materials. (4 GB)
- Hate speech identification: Contributors viewed short text and identified if it a) contained hate speech, b) was offensive but without hate speech, or c) was not offensive at all. Contains nearly 15K rows with three contributor judgments per text string. (3 MB)
- Hillary Clinton Emails [Kaggle]: nearly 7,000 pages of Clinton's heavily redacted emails (12 MB)
- Home Depot Product Search Relevance [Kaggle]: contains a number of products and real customer search terms from Home Depot's website. The challenge is to predict a relevance score for the provided combinations of search terms and products. To create the ground truth labels, Home Depot has crowdsourced the search/product pairs to multiple human raters. (65 MB)
- Identifying key phrases in text: Question/Answer pairs + context; context was judged if relevant to question/answer. (8 MB)
- Jeopardy: archive of 216,930 past Jeopardy questions (53 MB)
- 200k English plaintext jokes: archive of 208,000 plaintext jokes from various sources.
- Machine Translation of European Languages: (612 MB)
- Material Safety Datasheets: 230,000 Material Safety Data Sheets. (3 GB)
- Million News Headlines - ABC Australia [Kaggle]: 1.3 Million News headlines published by ABC News Australia from 2003 to 2017. (56 MB)
- MCTest: a freely available set of 660 stories and associated questions intended for research on the machine comprehension of text; for question answering (1 MB)
- NEGRA: A Syntactically Annotated Corpus of German Newspaper Texts. Available for free for all Universities and non-profit organizations. Need to sign and send form to obtain. (on request)
- News Headlines of India - Times of India [Kaggle]: 2.7 Million News Headlines with category published by Times of India from 2001 to 2017. (185 MB)
- News article / Wikipedia page pairings: Contributors read a short article and were asked which of two Wikipedia articles it matched most closely. (6 MB)
- NIPS2015 Papers (version 2) [Kaggle]: full text of all NIPS2015 papers (335 MB)
- NYTimes Facebook Data: all the NYTimes facebook posts (5 MB)
- One Week of Global News Feeds [Kaggle]: News Event Dataset of 1.4 Million Articles published globally in 20 languages over one week of August 2017. (115 MB)
- Objective truths of sentences/concept pairs: Contributors read a sentence with two concepts. For example “a dog is a kind of animal” or “captain can have the same meaning as master.” They were then asked if the sentence could be true and ranked it on a 1-5 scale. (700 KB)
- Open Library Data Dumps: dump of all revisions of all the records in Open Library. (16 GB)
- Personae Corpus: collected for experiments in Authorship Attribution and Personality Prediction. It consists of 145 Dutch-language essays by 145 different students. (on request)
- Reddit Comments: every publicly available reddit comment as of july 2015. 1.7 billion comments (250 GB)
- Reddit Comments (May ‘15) [Kaggle]: subset of above dataset (8 GB)
- Reddit Submission Corpus: all publicly available Reddit submissions from January 2006 - August 31, 2015). (42 GB)

- Reuters Corpus: a large collection of Reuters News stories for use in research and development of natural language processing, information retrieval, and machine learning systems. This corpus, known as "Reuters Corpus, Volume 1" or RCV1, is significantly larger than the older, well-known Reuters-21578 collection heavily used in the text classification community. Need to sign agreement and sent per post to obtain. (2.5 GB)
- SaudiNewsNet: 31,030 Arabic newspaper articles alongwith metadata, extracted from various online Saudi newspapers. (2 MB)
- SMS Spam Collection: 5,574 English, real and non-enconded SMS messages, tagged according being legitimate (ham) or spam. (200 KB)
- SouthparkData: .csv files containing script information including: season, episode, character, & line. (3.6 MB)
- Stackoverflow: 7.3 million stackoverflow questions + other stackexchanges (query tool)
- Twitter Cheng-Caverlee-Lee Scrape: Tweets from September 2009 - January 2010, geolocated. (400 MB)
- Twitter New England Patriots Deflategate sentiment: Before the 2015 Super Bowl, there was a great deal of chatter around deflated footballs and whether the Patriots cheated. This data set looks at Twitter sentiment on important days during the scandal to gauge public sentiment about the whole ordeal. (2 MB)
- Twitter Progressive issues sentiment analysis: tweets regarding a variety of left-leaning issues like legalization of abortion, feminism, Hillary Clinton, etc. classified if the tweets in question were for, against, or neutral on the issue (with an option for none of the above). (600 KB)
- Twitter Sentiment140: Tweets related to brands/keywords. Website includes papers and research ideas. (77 MB)
- Twitter sentiment analysis: Self-driving cars: contributors read tweets and classified them as very positive, slightly positive, neutral, slightly negative, or very negative. They were also prompted asked to mark if the tweet was not relevant to self-driving cars. (1 MB)
- Twitter Tokyo Geolocated Tweets: 200K tweets from Tokyo. (47 MB)
- Twitter UK Geolocated Tweets: 170K tweets from UK. (47 MB)
- Twitter USA Geolocated Tweets: 200k tweets from the US (45MB)
- Twitter US Airline Sentiment [Kaggle]: A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service"). (2.5 MB)
- U.S. economic performance based on news articles: News articles headlines and excerpts ranked as whether relevant to U.S. economy. (5 MB)
- Urban Dictionary Words and Definitions [Kaggle]: Cleaned CSV corpus of 2.6 Million of all Urban Dictionary words, definitions, authors, votes as of May 2016. (238 MB)
- Wesbury Lab Usenet Corpus: anonymized compilation of postings from 47,860 English-language newsgroups from 2005-2010 (40 GB)
- Wesbury Lab Wikipedia Corpus Snapshot of all the articles in the English part of the Wikipedia that was taken in April 2010. It was processed, as described in detail below, to remove all links and irrelevant material (navigation text, etc) The corpus is untagged, raw text. Used by Stanford NLP (1.8 GB).
- Wikipedia Extraction (WEX): a processed dump of english language wikipedia (66 GB)
- Wikipedia XML Data: complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. (500 GB)
- Yahoo! Answers Comprehensive Questions and Answers: Yahoo! Answers corpus as of 10/25/2007. Contains 4,483,032 questions and their answers. (3.6 GB)
- Yahoo! Answers consisting of questions asked in French: Subset of the Yahoo! Answers corpus from 2006 to 2015 consisting of 1.7 million questions posed in French, and their corresponding answers. (3.8 GB)
- Yahoo! Answers Manner Questions: subset of the Yahoo! Answers corpus from a 10/25/2007 dump, selected for their linguistic properties. Contains 142,627 questions and their answers. (104 MB)
- Yahoo! HTML Forms Extracted from Publicly Available Webpages: contains a small sample of pages that contain complex HTML forms, contains 2.67 million complex forms. (50+ GB)
- Yahoo! Metadata Extracted from Publicly Available Web Pages: 100 million triples of RDF data (2 GB)
- Yahoo N-Gram Representations: This dataset contains n-gram representations. The data may serve as a testbed for query rewriting task, a common problem in IR research as well as to word and sentence similarity task, which is common in NLP research. (2.6 GB)

- Yahoo! N-Grams, version 2.0: n-grams (n = 1 to 5), extracted from a corpus of 14.6 million documents (126 million unique sentences, 3.4 billion running words) crawled from over 12000 news-oriented sites (12 GB)
- Yahoo! Search Logs with Relevance Judgments: Annonymized Yahoo! Search Logs with Relevance Judgments (1.3 GB)
- Yahoo! Semantically Annotated Snapshot of the English Wikipedia: English Wikipedia dated from 2006-11-04 processed with a number of publicly-available NLP tools. 1,490,688 entries. (6 GB)
- Yelp: including restaurant rankings and 2.2M reviews (on request)
- Youtube: 1.7 million youtube videos descriptions (torrent)

### Source - AWESOMEDATA GITHUB

**https://github.com/awesomedata**

- Awesome public datasets/NLP (includes more lists)
- AWS Public Datasets
- CrowdFlower: Data for Everyone (lots of little surveys they conducted and data obtained by crowdsourcing for a specific task)
- Kaggle 1, 2 (make sure though that the kaggle competition data can be used outside of the competition!)
- Open Library
- Quora (mainly annotated corpora)
- /r/datasets (endless list of datasets, most is scraped by amateurs though and not properly documented or licensed)
- rs.io (another big list)
- Stackexchange: Opendata
- Stanford NLP group (mainly annotated corpora and TreeBanks or actual NLP tools)
- Yahoo! Webscope (also includes papers that use the data that is provided)

### Free Public Data Sets for Your First Data Science Project

https://www.springboard.com/blog/free-public-data-sets-data-science-project/

1.
    1. **United States Census Data:** The U.S. Census Bureau publishes reams of demographic data at the state, city, and even zip code level. The data set is fantastic for creating geographic data visualizations and can be accessed on the Census Bureau website. Alternatively, the data can be accessed via an API. One convenient way to use that API is through the choroplethr. In general, this data is very clean and very comprehensive.
    2. **FBI Crime Data:** The FBI crime data set is fascinating. If you're interested in analyzing time series data, you can use it to chart changes in crime rates at the national level over a 20-year period. Alternatively, you can look at the data geographically.
    3. **CDC Cause of Death:** The Centers for Disease Control and Prevention maintains a database on cause of death. The data can be segmented in almost every way imaginable: age, race, year, and so on.
    4. **Medicare Hospital Quality:** The Centers for Medicare & Medicaid Services maintains a database on quality of care at more than 4,000 Medicare-certified hospitals across the U.S., providing for interesting comparisons.
    5. **SEER Cancer Incidence:** The U.S. government also has data about cancer incidence, again segmented by age, race, gender, year, and other factors. It comes from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program.

6. **Bureau of Labor Statistics**: Many important economic indicators for the United States (like unemployment and inflation) can be found on the Bureau of Labor Statistics website. Most of the data can be segmented both by time and by geography.
7. **Bureau of Economic Analysis:** The Bureau of Economic Analysis also has national and regional economic data, including gross domestic product and exchange rates.
8. **IMF Economic Data:** For access to global financial statistics and other data, check out the International Monetary Fund's website.
9. **Dow Jones Weekly Returns:** Predicting stock prices is a major application of data analysis and machine learning. One relevant data set to explore is the weekly returns of the Dow Jones Index from the Center for Machine Learning and Intelligent Systems at the University of California, Irvine.
10. **Data.gov.uk:** The British government's official data portal offers access to tens of thousands of data sets on topics such as crime, education, transportation, and health.
11. **Enron Emails:** After the collapse of Enron, a data set of roughly 500,000 emails with message text and metadata were released. The data set is now famous and provides an excellent testing ground for text-related analysis. You also can explore other research uses of this data set through the page.
12. **Google Books Ngrams:** If you're interested in truly massive data, the Ngram viewer data set counts the frequency of words and phrases by year across a huge number of text sources. The resulting file is 2.2 TB.
13. **UNICEF:** If data about the lives of children around the world is of interest, UNICEF is the most credible source. The organization's public data sets touch upon nutrition, immunization, and education, among others.
14. **Reddit Comments:** Reddit released a data set of every comment that has ever been made on the site. That's over a terabyte of data uncompressed, so if you want a smaller data set to work with Kaggle has hosted the comments from May 2015 on their site.
15. **Wikipedia:** Wikipedia provides instructions for downloading the text of English-language articles, in addition to other projects from the Wikimedia Foundation.
16. **Lending Club:** Lending Club provides data about loan applications it has rejected as well as the performance of loans that it issued. The data set lends itself both to categorization techniques (will a given loan default) as well as regressions (how much will be paid back on a given loan).
17. **Walmart:** Walmart has released historical sales data for 45 stores located in different regions across the United States.
18. **Airbnb: Inside Airbnb offers different data sets related to Airbnb listings in dozens of cities around the world.**
19. **Yelp:** Yelp maintains a dataset for use in personal, educational, and academic purposes. It includes 6 million reviews spanning 189,000 businesses in 10 metropolitan areas. Students are welcome to participate in Yelp's dataset challenge.

**Economic Datasets**

Each year since 1978, the Federal Reserve Bank of Kansas City has sponsored a symposium on an important economic issue facing the U.S. and world economies. Symposium participants include prominent central bankers, finance ministers, academics, and financial market participants from around the world. Papers, commentary, and discussion.

https://www.kansascityfed.org/publications/research/escp/symposiums/escp-archive

**Data From Figure 8**

https://www.figure-eight.com/data-for-everyone/

Image URLs, the matched word, whether the pair matched, and a confidence score for each
https://www.figure-eight.com/wp-content/uploads/2016/03/image-descriptions-DFE.csv

Judge emotions about nuclear energy from Twitter
https://www.figure-eight.com/wp-content/uploads/2016/03/1377191648_sentiment_nuclear_power.csv

Decide whether two English sentences are related
https://www.figure-eight.com/wp-content/uploads/2016/03/1377882923_sentence_pairs.csv

Evaluate how similar are two sets of words on a seven point scale
https://www.figure-eight.com/wp-content/uploads/2016/03/1377883875_similar_word_combinations-1.csv

Sentiment Analysis Global Warming/Climate Change
https://www.figure-eight.com/wp-content/uploads/2016/03/1377884570_tweet_global_warming.csv

Judge Emotion About Brands
https://www.figure-eight.com/wp-content/uploads/2016/03/judge-1377884607_tweet_product_company.csv

tweets that mention Claritin for October, 2012
https://crowdflower.com/blog/2013/03/discovering-drug-side-effects-with-crowdsourcing/

Sentence plausibility- ranked them on a scale of implausible to plausible
https://www.figure-eight.com/wp-content/uploads/2016/03/plausible-sentences-DFE.csv

National Park locations
https://www.figure-eight.com/wp-content/uploads/2016/03/National-Park-Database-DFE.csv

Company categorizations
https://www.figure-eight.com/wp-content/uploads/2016/03/Company-Categorization-DFE.csv

How beautiful is this image? (Buildings and Architecture)
https://www.figure-eight.com/wp-content/uploads/2016/03/How-beautiful-buildings-DFE.csv

How beautiful is this image? (Animals)
https://www.figure-eight.com/wp-content/uploads/2016/03/How-beautiful-animals-DFE.csv

Gender breakdown of Time Magazine covers
https://www.figure-eight.com/wp-content/uploads/2016/03/TIME_Gender_Ratio.csv

Judge the relatedness of familiar words and made-up ones
https://www.figure-eight.com/wp-content/uploads/2016/03/judge-nonce-words.csv

**Audio Content Analysis**

**Source**: Alexander Lerch / Audio Content Analysis

https://www.audiocontentanalysis.org/data-sets/

**AWS Public Data Sets**

**Source**: AWS Public Data Sets https://aws.amazon.com/public-datasets/

**Geospatial and Environmental Datasets**

Learn more about working with geospatial data on AWS at Earth on AWS.

- Landsat on AWS: An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.
- Sentinel-2 on AWS: An ongoing collection of satellite imagery of all land on Earth produced by the Sentinel-2 satellite.
- GOES on AWS: GOES provides continuous weather imagery and monitoring of meteorological and space environment data across North America.
- SpaceNet on AWS: A corpus of commercial satellite imagery and labeled training data to foster innovation in the development of computer vision algorithms.
- OpenStreetMap on AWS: OSM is a free, editable map of the world, created and maintained by volunteers. Regular OSM data archives are made available in Amazon S3.
- MODIS on AWS: Select products from the Moderate Resolution Imaging Spectroradiometer (MODIS) managed by the U.S. Geological Survey and NASA.
- Terrain Tiles: A global dataset providing bare-earth terrain heights, tiled for easy usage and provided on S3.
- NAIP: 1 meter aerial imagery captured during the agricultural growing seasons in the continental U.S.
- NEXRAD on AWS: Real-time and archival data from the Next Generation Weather Radar (NEXRAD) network.
- NASA NEX: A collection of Earth science datasets maintained by NASA, including climate change projections and satellite images of the Earth's surface.
- District of Columbia LiDAR: LiDAR point cloud data for Washington, DC.
- EPA Risk-Screening Environmental Indicators: detailed air model results from EPA's Risk-Screening Environmental Indicators (RSEI) model.
- HIRLAM Weather Model: HIRLAM (High Resolution Limited Area Model) is an operational synoptic and mesoscale weather prediction model managed by the Finnish Meteorological Institute.

**Genomics and Life Science Datasets**

Learn more about genomics in the cloud.

- 1000 Genomes Project: A detailed map of human genetic variation.
- TCGA on AWS: Raw and processed genomic, transcriptomic, and epigenomic data from The Cancer Genome Atlas (TCGA) available to qualified researchers via the Cancer Genomics Cloud.
- ICGC on AWS: Whole genome sequence data available to qualified researchers via The International Cancer Genome Consortium (ICGC).
- 3000 Rice Genome on AWS: Genome sequence of 3,024 rice varieties.
- Genome in a Bottle (GIAB): Several reference genomes to enable translation of whole human genome sequencing to clinical practice.

**Datasets for Machine Learning**

Learn more about artificial intelligence and machine learning on AWS.

- Common Crawl: A corpus of web crawl data composed of over 5 billion web pages.
- Amazon Bin Image Dataset: Over 500,000 bin JPEG images and corresponding JSON metadata files describing products in an operating Amazon Fulfillment Center.
- GDELT: Over a quarter-billion records monitoring the world's broadcast, print, and web news from nearly every corner of every country, updated daily.

- Multimedia Commons: A collection of nearly 100M images and videos with audio and visual features and annotations.
- Google Books Ngrams: A dataset containing Google Books n-gram corpuses.
- SpaceNet on AWS: A corpus of commercial satellite imagery and labeled training data to foster innovation in the development of computer vision algorithms.

**Financial Data**

- Deutsche Börse Public Dataset: Real-time data derived from Deutsche Börse's trading market systems available to the public for free.

**Regulatory and Statistical Data**

- IRS 990 Filings on AWS: Machine-readable data from certain electronic 990 forms filed with the IRS from 2011 to present.
- ACS PUMS on AWS: U.S. Census American Community Survey (ACS) Public Use Microdata Sample (PUMS) is available in a linked data format using the Resource Description Framework (RDF) data model.
- USAspending.gov on AWS: USAspending.gov database, which includes data on all spending by the federal government, including contracts, grants, loans, employee salaries, and more.

**Source: CRAN**

Provides functions to download data from UK Parliament (Text/speeches)

https://cran.r-project.org/web/packages/hansard/index.html

**Source: https://deeplearning4j.org/opendata**

- Open Source Biometric Recognition Data
- Google Audioset: An expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos.
- Uber 2B trip data: Slow rollout of access to ride data for 2Bn trips.
- Yelp Open Dataset: The Yelp dataset is a subset of Yelp businesses, reviews, and user data for use in NLP.
- Kaggle Datasets Page
- Data Portals
- Open Data Monitor
- Quandl Data Portal

**Source: Quora:**

**https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public**

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

1. http://datasource.kapsarc.org
2. https://www.kaggle.com/datasets
3. http://www.assetmacro.com

4. http://usgovxml.com
5. http://aws.amazon.com/datasets
6. http://databib.org
7. http://datacite.org
8. http://figshare.com
9. http://linkeddata.org
10. http://reddit.com/r/datasets
11. http://thewebminer.com/
12. http://thedatahub.org alias http://ckan.net
13. http://quandl.com
14. Social Network Analysis Interactive Dataset Library (Social Network Datasets)
15. Datasets for Data Mining
16. Enigma Public
17. *http://www.ufindthem.com/*
18. http://NetworkRepository.com - The First Interactive Network Data Repository
19. http://MLvis.com
20. Open Data Inception - A Comprehensive List of 2500+ Open Data Portals in the World
21. http://data.opendatasoft.com OpenDataSoft catalog

**Single datasets and data repositories**

1. http://archive.ics.uci.edu/ml/
2. http://crawdad.org/
3. http://data.austintexas.gov
4. http://data.cityofchicago.org
5. http://data.govloop.com
6. http://data.gov.uk/
7. data.gov.in
8. http://data.medicare.gov
9. http://data.seattle.gov
10. http://data.sfgov.org
11. http://data.sunlightlabs.com
12. https://datamarket.azure.com/
13. http://developer.yahoo.com/geo/g...
14. http://econ.worldbank.org/datasets
15. http://en.wikipedia.org/wiki/Wik...
16. http://factfinder.census.gov/ser...
17. http://ftp.ncbi.nih.gov/
18. http://gettingpastgo.socrata.com
19. http://googleresearch.blogspot.c...
20. http://books.google.com/ngrams/
21. http://medihal.archives-ouvertes.fr
22. http://public.resource.org/
23. http://rechercheisidore.fr
24. http://snap.stanford.edu/data/in...
25. http://timetric.com/public-data/
26. https://wist.echo.nasa.gov/~wist...
27. http://www2.jpl.nasa.gov/srtm
28. http://www.archives.gov/research...
29. http://www.bls.gov/
30. http://www.crunchbase.com/
31. http://www.dartmouthatlas.org/
32. http://www.data.gov/
33. http://www.datakc.org
34. http://dbpedia.org
35. http://www.delicious.com/jbaldwi...
36. http://www.faa.gov/data_research/

37. http://www.factual.com/
38. http://research.stlouisfed.org/f...
39. http://www.freebase.com/
40. http://www.google.com/publicdata...
41. http://www.guardian.co.uk/news/d...
42. http://www.infochimps.com
43. http://www.kaggle.com/
44. http://build.kiva.org/
45. http://www.nationalarchives.gov....
46. http://www.nyc.gov/html/datamine...
47. http://www.ordnancesurvey.co.uk/...
48. http://www.philwhln.com/how-to-g...
49. http://www.imdb.com/interfaces
50. http://imat-relpred.yandex.ru/en...
51. http://www.dados.gov.pt/pt/catal...
52. http://knoema.com
53. http://daten.berlin.de/
54. http://www.qunb.com
55. http://databib.org/
56. http://datacite.org/
57. http://data.reegle.info/
58. http://data.wien.gv.at/
59. http://data.gov.bc.ca
60. https://pslcdatashop.web.cmu.edu/ (interaction data in learning environments)
61. http://www.icpsr.umich.edu/icpsrweb/CPES/ - 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)
62. http://www.dati.gov.it
63. http://dati.trentino.it
64. http://www.databagg.com/
65. http://networkrepository.com - Network/ML data repository w/ visual interactive analytics
66. Home (United Nations Environment Programme Grid Genava a lot of GIS datasets

**Source: Google Search**

r-directory > Reference Links > Free Data Sets https://r-dir.com/reference/datasets.html
Big Data Made Simple - 70 WebSites - http://bigdata-madesimple.com/70-websites-to-get-large-data-repositories-for-free/
18 places to find data sets for data science projects https://www.dataquest.io/blog/free-datasets-for-projects/

**Source: IBM - https://apsportal.ibm.com/community**

Consumer Prices

Consumption of ozone-depleting CFCs in ODP...

Contraceptive prevalence (% women 15-49) by...

Country Population by Gender 1985-2005

Country populations 15 years of age and...

Country Statistics - Europe - Population and...

Country Statistics: Airports

Country Statistics: Area

Country Statistics: Birth Rate

Country Statistics: Budget Surplus Or Deficit

Country Statistics: Central Bank Discount...

Country Statistics: Children Under the Age...

Country Statistics: Commercial Bank Prime...

Country Statistics: Crude Oil - Exports

Country Statistics: Crude Oil - Imports

Country Statistics: Crude Oil - Production

Country Statistics: Crude Oil - Proved...

Country Statistics: Current Account Balance

Country Statistics: Death Rate

Country Statistics: Debt - External

Country Statistics: Distribution of Family...

Country Statistics: Education Expenditures

Country Statistics: Electricity - Consumption

Country Statistics: Electricity - Exports

Country Statistics: Electricity - From Other...

Country Statistics: Electricity - From...

Country Statistics: Electricity - From...

Country Statistics: Electricity - From...

Country Statistics: Electricity - Imports

Country Statistics: Electricity - Installed...

Country Statistics: Electricity - Production

Country Statistics: Exports

Country Statistics: GDP - Per Capita (PPP)

Country Statistics: GDP - Real Growth Rate

Country Statistics: GDP (Purchasing Power...

Country Statistics: Gross National Saving

Country Statistics: Health Expenditures

Country Statistics: HIV/AIDS - Adult...

Country Statistics: HIV/AIDS - Deaths

Country Statistics: HIV/AIDS - People Living...

Country Statistics: Imports

Country Statistics: Industrial Production...

Country Statistics: Infant Mortality Rate

Country Statistics: Inflation Rate (Consumer...

Country Statistics: Internet Users

Country Statistics: Labor Force

Country Statistics: Life Expectancy at Birth

Country Statistics: Market Value of Publicly...

Country Statistics: Maternal Mortality Rate

Country Statistics: Merchant Marine

Country Statistics: Military Expenditures

Country Statistics: Natural Gas - Consumption

Country Statistics: Natural Gas - Exports

Country Statistics: Natural Gas - Imports

Country Statistics: Natural Gas - Production

Country Statistics: Natural Gas - Proved...

Country Statistics: Net Migration Rate

Country Statistics: Obesity - Adult...

Country Statistics: Population

Country Statistics: Population Growth Rate

Country Statistics: Public Debt

Country Statistics: Railways

Country Statistics: Refined Petroleum...

Country Statistics: Refined Petroleum...

Country Statistics: Refined Petroleum...

Country Statistics: Refined Petroleum...

Country Statistics: Reserves of Foreign...

Country Statistics: Roadways

Country Statistics: Stock of Broad Money

Country Statistics: Stock of Direct Foreign...

Country Statistics: Stock of Direct Foreign...

Country Statistics: Stock of Domestic Credit

Country Statistics: Stock of Narrow Money

Country Statistics: Taxes And Other Revenues

Country Statistics: Telephones - Fixed Lines

Country Statistics: Telephones - Mobile...

Country Statistics: Total Fertility Rate

Country Statistics: Unemployment Rate

Country Statistics: Unemployment, Youth Ages...

Country Statistics: Waterways

Country Surface Area (sq. km)

Days required to start a business by country

Dry Bulb Temperature, by country, station...

Environment Statistics Database - Waste

Environment Statistics Database - Water

Exports of goods and services (% of GDP)

Exports of goods and services as % of GDP by...

External debt stocks, total (DOD, current...

Fertility rate by country in total births...

Foreign direct investment, net inflows (BoP,...

Forest area by country in sq km

Fortune 100 companies

GDP by country in US$

GDP growth annual % by country

Geographic Coordinates of World Locations

GNI, PPP (current international $) by country

GoSales Transactions for Logistic Regression...

GoSales Transactions for Naive Bayes Model

Government Consumption Expenditure

Greenhouse Gas Emissions worldwide

Gross capital formation as % of GDP by...

Gross domestic expenditure on research and...

Gross National Income per capita, Atlas...

Gross National Income per capita, PPP...

Gross National Income, Atlas method (current...

Haven't found what you're looking for?Visit Stack Overflow

High-tech exports as % of manufactured...

Household Consumption Expenditure

Households by type of household, age and sex...

IBM Data Science Experience

IBM Watson Facebook posts for 2015

Imports of goods and services as % of GDP by...

Improved water source by country: %...

Intentional homicide, number and rate per...

Intents & examples for IBM Watson...

Interest Rates

International Liquidity

Life expectancy at birth by country in total...

Malaria death rate per 100,000 population,...

Marital status of men and women

Market capitalization of listed companies as...

Measles immunization % children 12-23 months...

Merchandise trade as % of GDP by country

Military expenditure as % of GDP by country

Minimum Legal Age for Marriage per Country

Mobile cellular subscriptions per 100 people...

Mobile-cellular telephone subscriptions per...

Natural gas production, 1995 - 2012,...

OrNotebookto default project

Part-time employment rate, worldwide, by...

Percentage of Internet users by country

Population below national poverty line,...

Primary school completion rate % of relevant...

Ratio (% of population) at national poverty...

Ratio of girls to boys in primary and...

Refugees

Refugees, worldwide, 2003 - 2013

Renewable internal freshwater resources per...

Roads paved as % of total roads by country

Roads, paved (% of total roads), Worldwide,...

Services value added as % of GDP by country

Share Price Index (SPI), Worldwide, by...

The Nurse Assignment Problem data

Total employment, by economic activity...

Total population by country

Total population, both sexes combined...

UCI ML Repository: Chronic Kidney Disease...

UCI: Abalone

UCI: Adult - Predict income

UCI: Car evaluation

UCI: Forest fires

UCI: Iris

UCI: Poker hand - testing data set

UCI: Poker hand - training data set

UCI: Wine recognition

United States Demographic Measures: Education

United States Demographic Measures: Income

United States Demographic Measures: Race

United States Demographic Measures:...

Unmet need for family planning, spacing,...

World Marriage Data

World Tourism Data by the World Tourism...

Worldwide County and Region - National...

Worldwide Electricity Demand and Production...

Worldwide Fuel Oil Consumption By Household...

**Source: http://www.data.gov/**

- Agriculture
- Climate
- Consumer
- Ecosystems
- Education
- Energy
- Finance

- Health
- Local Government
- Manufacturing
- Maritime
- Ocean
- Public Safety
- Science & Research

Check out Data.gov's new Metrics Pag - July 31, 2017 *By Data.gov*

=====

- **Community Categories**
  - geospatial (80348)
  - temperature (25708)
  - physical (22940)
  - profile (20343)
  - water depth (15116)
  - unknown (14638)
  - Access (63)
  - api (68)
  - application/octet-s... (51)
  - application/vnd.lot... (3263)
  - application/xslt+xml (326)
  - Safety (852)
  - Research (799)
  - Energy (776)
  - Consumer (412)
  - Ocean (406)
  - Federal Government (54725)
  - Multiple Sources (21940)
  - University (7133)
  - State Government (3375)
  - Other (102)
  - National Oceanic an... (32227)
  - NSGIC GIS Inventory... (21940)
  - U.S. Geological Sur... (9924)
  - Earth Data Analysis... (4164)
  - US Census Bureau, D... (3509)
  - Applied Science & T... (600)
  - Fatalities, Casualt... (470)
  - Total Energy (312)
  - Natural Resources a... (275)
  - Markets, Prices, an... (143)
  - Natural Hazards (136)
  - Environment (111)
  - Electricity (107)
  - Physical and Oceano... (82)
  - Housing and Community (80)
  - Petroleum and Other... (76)
  - Biology and Habitats (74)
  - Environment & Envir... (73)
  - Elevation and Bathy... (71)
  - Energy (68)
  - Consumption and Eff... (66)
  - Finance (64)
  - Plants and Plant Sy... (60)
  - Natural Gas (55)

- Exposure Data (55)
- Food/Non-Food Agric... (54)
- Human Health and Nu... (53)
- Compliance, Violati... (48)
- Transportation (42)
- Pre-K to 12 Education (42)
- Corrections Data (41)
- Agriculture Investm... (40)
- Incidents and Crashes (38)
- Agriculture & Food (38)
- Social & Behavioral... (37)
- Rural Development (36)
- Nuclear (34)
- Atmospheric, Earth ... (33)
- Health Care (31)
- Coal (30)
- Local and Regional ... (28)
- Hydropower (23)
- Energy & Energy Con... (22)
- Law Enforcement Data (21)
- Animals and Animal ... (19)
- Education (18)
- Higher Education (18)
- Safety (17)
- Renewable & Alterna... (16)
- International (13)
- Ecosystem Change Dr... (12)
- Science and Technology (11)
- Earth, Atmospheric ... (10)
- Crime Data (10)
- Vocational and Adult (9)

**===**

**SOURCE -**

**http://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#70f9792d6796**

1. Data.gov http://data.gov The US Government pledged last year to make all government data available freely online. This site is the first stage and acts as a portal to all sorts of amazing information on everything from climate to crime.
2. US Census Bureau http://www.census.gov/data.html A wealth of information on the lives of US citizens covering population data, geographic data and education.
3. Socrata is another interesting place to explore government-related data, with some visualisation tools built-in.
4. European Union Open Data Portal http://open-data.europa.eu/en/data/ As the above, but based on data from European Union institutions.
5. Data.gov.uk http://data.gov.uk/ Data from the UK Government, including the British National Bibliography – metadata on all UK books and publications since 1950.
6. Canada Open Data is a pilot project with many government and geospatial datasets.
7. Datacatalogs.org offers open government data from US, EU, Canada, CKAN, and more.
8. The CIA World Factbook https://www.cia.gov/library/publications/the-world-factbook/ Information on history, population, economy, government, infrastructure and military of 267 countries.Healthdata.gov

https://www.healthdata.gov/ 125 years of US healthcare data including claim-level Medicare data, epidemiology and population statistics.

9. NHS Health and Social Care Information Centre http://www.hscic.gov.uk/home Health data sets from the UK National Health Service.
10. UNICEF offers statistics on the situation of women and children worldwide.
11. World Health Organization offers world hunger, health, and disease statistics.
12. Amazon Web Services public datasets http://aws.amazon.com/datasets Huge resource of public data, including the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information and NASA 's database of satellite imagery of Earth.
13. Facebook FB +0.23% Graph https://developers.facebook.com/docs/graph-api Although much of the information on users' Facebook profile is private, a lot isn't – Facebook provide the Graph API as a way of querying the huge amount of information that its users are happy to share with the world (or can't hide because they haven't worked out how the privacy settings work).
14. Face.com: A fascinating tool for facial recognition data.
15. UCLA makes some of the data from its courses public.
16. Data Market is a place to check out  data related to economics, healthcare, food and agriculture, and the automotive industry.
17. Google Public data explorer includes data from world development indicators, OECD, and human development indicators, mostly related to economics data and the world.
18. Junar is a data scraping service that also includes data feeds.
19. Buzzdata is a social data sharing service that allows you to upload your own data and connect with others who are uploading their data.
20. Gapminder http://www.gapminder.org/data/ Compilation of data from sources including the World Health Organization and World Bank covering economic, medical and social statistics from around the world.
21. Google GOOGL +0.25% Trends http://www.google.com/trends/explore Statistics on search volume (as a proportion of total search) for any given term, since 2004.
22. Google Finance https://www.google.com/finance 40 years' worth of stock market data, updated in real time.
23. Google Books Ngrams http://storage.googleapis.com/books/ngrams/books/datasetsv2.html Search and analyze the full text of any of the millions of books digitised as part of the Google Books project.
24. National Climatic Data Center http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim Huge collection of environmental, meteorological and climate data sets from the US National Climatic Data Center. The world's largest archive of weather data.
25. DBPedia http://wiki.dbpedia.org Wikipedia is comprised of millions of pieces of data, structured and unstructured on every subject under the sun. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.
26. New York Times NYT -0.42% http://developer.nytimes.com/docs Searchable, indexed archive of news articles going back to 1851.
27. Freebase http://www.freebase.com/ A community-compiled database of structured data about people, places and things, with over 45 million entries.
28. Million Song Data Set http://aws.amazon.com/datasets/6468931156960467 Metadata on over a million songs and pieces of music. Part of Amazon Web Services.
29. UCI Machine Learning Repository is a dataset specifically pre-processed for machine learning.
30. Financial Data Finder at OSU offers a large catalog of financial data sets.
31. Pew Research Center offers its raw data from its fascinating research into American life.
32. The BROAD Institute offers a number of cancer-related datasets.

====

**Source: Caesar0301 Awesome Data Sets**

**https://github.com/caesar0301/awesome-public-datasets**

**Agriculture**

- U.S. Department of Agriculture's PLANTS Database
- U.S. Department of Agriculture's Nutrient Database

## Biology

- 1000 Genomes
- American Gut (Microbiome Project)
- Broad Bioimage Benchmark Collection (BBBC)
- Broad Cancer Cell Line Encyclopedia (CCLE)
- Cell Image Library
- Complete Genomics Public Data
- EBI ArrayExpress
- EBI Protein Data Bank in Europe
- Electron Microscopy Pilot Image Archive (EMPIAR)
- ENCODE project
- Ensembl Genomes
- Gene Expression Omnibus (GEO)
- Gene Ontology (GO)
- Global Biotic Interactions (GloBI)
- Harvard Medical School (HMS) LINCS Project
- Human Genome Diversity Project
- Human Microbiome Project (HMP)
- ICOS PSP Benchmark
- International HapMap Project
- Journal of Cell Biology DataViewer
- MIT Cancer Genomics Data
- NCBI Proteins
- NCBI Taxonomy
- NCI Genomic Data Commons
- NIH Microarray data or FTP (see FTP link on RAW)
- OpenSNP genotypes data
- Pathguid - Protein-Protein Interactions Catalog
- Protein Data Bank
- Psychiatric Genomics Consortium
- PubChem Project
- PubGene (now Coremine Medical)
- Sanger Catalogue of Somatic Mutations in Cancer (COSMIC)
- Sanger Genomics of Drug Sensitivity in Cancer Project (GDSC)
- Sequence Read Archive(SRA)
- Stanford Microarray Data
- Stowers Institute Original Data Repository
- Systems Science of Biological Dynamics (SSBD) Database
- The Cancer Genome Atlas (TCGA), available via Broad GDAC
- The Catalogue of Life
- The Personal Genome Project or PGP
- UCSC Public Data
- UniGene
- Universal Protein Resource (UnitProt)

## Climate/Weather

- Actuaries Climate Index
- Australian Weather
- Aviation Weather Center - Consistent, timely and accurate weather information for the world airspace system
- Brazilian Weather - Historical data (In Portuguese)
- Canadian Meteorological Centre
- Climate Data from UEA (updated monthly)
- European Climate Assessment & Dataset
- Global Climate Data Since 1929
- NASA Global Imagery Browse Services
- NOAA Bering Sea Climate
- NOAA Climate Datasets
- NOAA Realtime Weather Models
- NOAA SURFRAD Meteorology and Radiation Datasets
- The World Bank Open Data Resources for Climate Change
- UEA Climatic Research Unit
- WorldClim - Global Climate Data
- WU Historical Weather Worldwide

## Complex Networks

- AMiner Citation Network Dataset
- CrossRef DOI URLs
- DBLP Citation dataset
- DIMACS Road Networks Collection
- NBER Patent Citations
- Network Repository with Interactive Exploratory Analysis Tools
- NIST complex networks data collection
- Protein-protein interaction network
- PyPI and Maven Dependency Network
- Scopus Citation Database
- Small Network Data
- Stanford GraphBase (Steven Skiena)
- Stanford Large Network Dataset Collection
- Stanford Longitudinal Network Data Sources
- The Koblenz Network Collection
- The Laboratory for Web Algorithmics (UNIMI)
- The Nexus Network Repository
- UCI Network Data Repository
- UFL sparse matrix collection
- WSU Graph Database

## Computer Networks

- 3.5B Web Pages from CommonCrawl 2012
- 53.5B Web clicks of 100K users in Indiana Univ.
- CAIDA Internet Datasets

- ClueWeb09 - 1B web pages
- ClueWeb12 - 733M web pages
- CommonCrawl Web Data over 7 years
- CRAWDAD Wireless datasets from Dartmouth Univ.
- Criteo click-through data
- OONI: Open Observatory of Network Interference - Internet censorship data
- Open Mobile Data by MobiPerf
- Rapid7 Sonar Internet Scans
- UCSD Network Telescope, IPv4 /8 net

## Data Challenges

- Bruteforce Database
- Challenges in Machine Learning
- CrowdANALYTIX dataX
- D4D Challenge of Orange
- DrivenData Competitions for Social Good
- ICWSM Data Challenge (since 2009)
- Kaggle Competition Data
- KDD Cup by Tencent 2012
- Localytics Data Visualization Challenge
- Netflix Prize
- Space Apps Challenge
- Telecom Italia Big Data Challenge
- TravisTorrent Dataset - MSR'2017 Mining Challenge
- Yelp Dataset Challenge

## Earth Science

- AQUASTAT - Global water resources and uses
- BODC - marine data of ~22K vars
- Earth Models
- EOSDIS - NASA's earth observing system data
- Integrated Marine Observing System (IMOS) - roughly 30TB of ocean measurements or on S3
- Marinexplore - Open Oceanographic Data
- Smithsonian Institution Global Volcano and Eruption Database
- USGS Earthquake Archives

## Economics

- American Economic Association (AEA)
- EconData from UMD
- Economic Freedom of the World Data
- Historical MacroEconomc Statistics
- International Economics Database and various data tools
- International Trade Statistics

- Internet Product Code Database
- Joint External Debt Data Hub
- Jon Haveman International Trade Data Links
- OpenCorporates Database of Companies in the World
- Our World in Data
- SciencesPo World Trade Gravity Datasets
- The Atlas of Economic Complexity
- The Center for International Data
- The Observatory of Economic Complexity
- UN Commodity Trade Statistics
- UN Human Development Reports

## Education

- College Scorecard Data
- Student Data from Free Code Camp

## Energy

- AMPds
- BLUEd
- COMBED
- Dataport
- DRED
- ECO
- EIA
- HES - Household Electricity Study, UK
- HFED
- iAWE
- PLAID - the Plug Load Appliance Identification Dataset
- REDD
- Tracebase
- UK-DALE - UK Domestic Appliance-Level Electricity
- WHITED

## Finance

- CBOE Futures Exchange
- Google Finance
- Google Trends
- NASDAQ
- NYSE Market Data (see FTP link on RAW)
- OANDA
- OSU Financial data
- Quandl
- St Louis Federal

- Yahoo Finance

## GIS

- ArcGIS Open Data portal
- Cambridge, MA, US, GIS data on GitHub
- Factual Global Location Data
- Geo Spatial Data from ASU
- Geo Wiki Project - Citizen-driven Environmental Monitoring
- GeoFabrik - OSM data extracted to a variety of formats and areas
- GeoNames Worldwide
- Global Administrative Areas Database (GADM)
- Homeland Infrastructure Foundation-Level Data
- Landsat 8 on AWS
- List of all countries in all languages
- National Weather Service GIS Data Portal
- Natural Earth - vectors and rasters of the world
- OpenAddresses
- OpenStreetMap (OSM)
- Pleiades - Gazetteer and graph of ancient places
- Reverse Geocoder using OSM data & additional high-resolution data files
- TIGER/Line - U.S. boundaries and roads
- TwoFishes - Foursquare's coarse geocoder
- TZ Timezones shapfiles
- UN Environmental Data
- World boundaries from the U.S. Department of State
- World countries in multiple formats

## Government

- A list of cities and countries contributed by community
- Open Data for Africa
- OpenDataSoft's list of 1,600 open data

## Healthcare

- EHDP Large Health Data Sets
- Gapminder World demographic databases
- GDC supports several cancer genome programs for CCG, TCGA, TARGET etc.
- PhysioBank Databases - a large and growing archive of physiological data
- Medicare Coverage Database (MCD), U.S.
- Medicare Data Engine of medicare.gov Data
- Medicare Data File
- MeSH, the vocabulary thesaurus used for indexing articles for PubMed
- Number of Ebola Cases and Deaths in Affected Countries (2014)
- Open-ODS (structure of the UK NHS)

- OpenPaymentsData, Healthcare financial relationship data
- The Cancer Genome Atlas project (TCGA) (refer to GDC and BigQuery table)
- World Health Organization Global Health Observatory

## Image Processing

- 10k US Adult Faces Database
- 2GB of Photos of Cats or Archive version
- Adience Unfiltered faces for gender and age classification
- Affective Image Classification
- Animals with attributes
- Caltech Pedestrian Detection Benchmark
- Chars74K dataset, Character Recognition in Natural Images (both English and Kannada are available)
- Face Recognition Benchmark
- Flickr: 32 Class Brand Logos
- GDXray: X-ray images for X-ray testing and Computer Vision
- ImageNet (in WordNet hierarchy)
- Indoor Scene Recognition
- International Affective Picture System, UFL
- Massive Visual Memory Stimuli, MIT
- MNIST database of handwritten digits, near 1 million examples
- Several Shape-from-Silhouette Datasets
- Stanford Dogs Dataset
- SUN database, MIT
- The Action Similarity Labeling (ASLAN) Challenge
- The Oxford-IIIT Pet Dataset
- Violent-Flows - Crowd Violence Non-violence Database and benchmark
- Visual genome
- YouTube Faces Database

## Machine Learning

- Context-aware data sets from five domains
- Delve Datasets for classification and regression (Univ. of Toronto)
- Discogs Monthly Data
- eBay Online Auctions (2012)
- IMDb Database
- Keel Repository for classification, regression and time series
- Labeled Faces in the Wild (LFW)
- Lending Club Loan Data
- Machine Learning Data Set Repository
- Free Music Archive
- Million Song Dataset
- More Song Datasets
- MovieLens Data Sets
- New Yorker caption contest ratings
- RDataMining - "R and Data Mining" ebook data
- Registered Meteorites on Earth

- Restaurants Health Score Data in San Francisco
- UCI Machine Learning Repository
- Yahoo! Ratings and Classification Data
- Youtube 8m

## Museums

- Canada Science and Technology Museums Corporation's Open Data
- Cooper-Hewitt's Collection Database
- Minneapolis Institute of Arts metadata
- Natural History Museum (London) Data Portal
- Rijksmuseum Historical Art Collection
- Tate Collection metadata
- The Getty vocabularies

## Natural Language

- POS/NER/Chunk annotated data
- Automatic Keyphrase Extraction
- Blogger Corpus
- CLiPS Stylometry Investigation Corpus
- ClueWeb09 FACC
- ClueWeb12 FACC
- DBpedia - 4.58M things with 583M facts
- Flickr Personal Taxonomies
- Freebase.com of people, places, and things
- Google Books Ngrams (2.2TB)
- Google MC-AFP, generated based on the public available Gigaword dataset using Paragraph Vectors
- Google Web 5gram (1TB, 2006)
- Gutenberg eBooks List
- Hansards text chunks of Canadian Parliament
- Machine Comprehension Test (MCTest) of text from Microsoft Research
- Machine Translation of European languages
- Making Sense of Microposts 2013 - Concept Extraction
- Making Sense of Microposts 2016 - Named Entity rEcognition and Linking
- Microsoft MAchine Reading COmprehension Dataset (or MS MARCO)
- Multi-Domain Sentiment Dataset (version 2.0)
- Open Multilingual Wordnet
- Personae Corpus
- SaudiNewsNet Collection of Saudi Newspaper Articles (Arabic, 30K articles)
- SMS Spam Collection in English
- Universal Dependencies
- USENET postings corpus of 2005~2011
- Webhose - News/Blogs in multiple languages
- Wikidata - Wikipedia databases
- Wikipedia Links data - 40 Million Entities in Context
- WordNet databases and tools

## Neuroscience

- Allen Institute Datasets
- Brain Catalogue
- Brainomics
- CodeNeuro Datasets
- Collaborative Research in Computational Neuroscience (CRCNS)
- FCP-INDI
- Human Connectome Project
- NDAR
- NeuroData
- Neuroelectro
- NIMH Data Archive
- OASIS
- OpenfMRI
- Study Forrest

## Physics

- CERN Open Data Portal
- Crystallography Open Database
- NASA Exoplanet Archive
- NSSDC (NASA) data of 550 space spacecraft
- Sloan Digital Sky Survey (SDSS) - Mapping the Universe

## Psychology/Cognition

- OSU Cognitive Modeling Repository Datasets

## Public Domains

- Amazon
- Archive-it from Internet Archive
- Archive.org Datasets
- CMU JASA data archive
- CMU StatLab collections
- Data.World
- Data360
- Google
- Infochimps
- KDNuggets Data Collections
- Microsoft Azure Data Market Free DataSets
- Microsoft Data Science for Research
- Numbray

- Open Library Data Dumps
- Reddit Datasets
- RevolutionAnalytics Collection
- Sample R data sets
- Stats4Stem R data sets
- StatSci.org
- The Washington Post List
- UCLA SOCR data collection
- UFO Reports
- Wikileaks 911 pager intercepts
- Yahoo Webscope

## Search Engines

- Academic Torrents of data sharing from UMB
- Datahub.io
- DataMarket (Qlik)
- Harvard Dataverse Network of scientific data
- ICPSR (UMICH)
- Institute of Education Sciences
- National Technical Reports Library
- Open Data Certificates (beta)
- OpenDataNetwork - A search engine of all Socrata powered data portals
- Statista.com - statistics and Studies
- Zenodo - An open dependable home for the long-tail of science

## Social Networks

- 72 hours #gamergate Twitter Scrape
- Ancestry.com Forum Dataset over 10 years
- Cheng-Caverlee-Lee September 2009 - January 2010 Twitter Scrape
- CMU Enron Email of 150 users
- EDRM Enron EMail of 151 users, hosted on S3
- Facebook Data Scrape (2005)
- Facebook Social Networks from LAW (since 2007)
- Foursquare from UMN/Sarwat (2013)
- GitHub Collaboration Archive
- Google Scholar citation relations
- High-Resolution Contact Networks from Wearable Sensors
- Indie Map: social graph and crawl of top IndieWeb sites
- Mobile Social Networks from UMASS
- Network Twitter Data
- Reddit Comments
- Skytrax' Air Travel Reviews Dataset
- Social Twitter Data
- SourceForge.net Research Data
- Twitter Data for Online Reputation Management
- Twitter Data for Sentiment Analysis

- Twitter Graph of entire Twitter site
- Twitter Scrape Calufa May 2011
- UNIMI/LAW Social Network Datasets
- Yahoo! Graph and Social Data
- Youtube Video Social Graph in 2007,2008

## Social Sciences

- ACLED (Armed Conflict Location & Event Data Project)
- Canadian Legal Information Institute
- Center for Systemic Peace Datasets - Conflict Trends, Polities, State Fragility, etc
- Correlates of War Project
- Cryptome Conspiracy Theory Items
- Datacards
- European Social Survey
- FBI Hate Crime 2013 - aggregated data
- Fragile States Index
- GDELT Global Events Database
- General Social Survey (GSS) since 1972
- German Social Survey
- Global Religious Futures Project
- Humanitarian Data Exchange
- INFORM Index for Risk Management
- Institute for Demographic Studies
- International Networks Archive
- International Social Survey Program ISSP
- International Studies Compendium Project
- James McGuire Cross National Data
- MacroData Guide by Norsk samfunnsvitenskapelig datatjeneste
- Minnesota Population Center
- MIT Reality Mining Dataset
- Notre Dame Global Adaptation Index (NG-DAIN)
- Open Crime and Policing Data in England, Wales and Northern Ireland
- Paul Hensel General International Data Page
- PewResearch Internet Survey Project
- PewResearch Society Data Collection
- Political Polarity Data
- StackExchange Data Explorer
- Terrorism Research and Analysis Consortium
- Texas Inmates Executed Since 1984
- Titanic Survival Data Set or on Kaggle
- UCB's Archive of Social Science Data (D-Lab)
- UCLA Social Sciences Data Archive
- UN Civil Society Database
- Universities Worldwide
- UPJOHN for Labor Employment Research
- Uppsala Conflict Data Program
- World Bank Open Data
- WorldPop project - Worldwide human population distributions

## Software

- FLOSSmole data about free, libre, and open source software development

## Sports

- Betfair Historical Exchange Data
- Cricsheet Matches (cricket)
- Ergast Formula 1, from 1950 up to date (API)
- Football/Soccer resources (data and APIs)
- Lahman's Baseball Database
- Pinhooker: Thoroughbred Bloodstock Sale Data
- Retrosheet Baseball Statistics
- Tennis database of rankings, results, and stats for ATP, WTA, Grand Slams and Match Charting Project

## Time Series

- Databanks International Cross National Time Series Data Archive
- Hard Drive Failure Rates
- Heart Rate Time Series from MIT
- Time Series Data Library (TSDL) from MU
- UC Riverside Time Series Dataset

## Transportation

- Airlines OD Data 1987-2008
- Bay Area Bike Share Data
- Bike Share Systems (BSS) collection
- GeoLife GPS Trajectory from Microsoft Research
- German train system by Deutsche Bahn
- Hubway Million Rides in MA
- Marine Traffic - ship tracks, port calls and more
- Montreal BIXI Bike Share
- NYC Taxi Trip Data 2009-
- NYC Taxi Trip Data 2013 (FOIA/FOILed)
- NYC Uber trip data April 2014 to September 2014
- Open Traffic collection
- OpenFlights - airport, airline and route data
- Philadelphia Bike Share Stations (JSON)
- Plane Crash Database, since 1920
- RITA Airline On-Time Performance data
- RITA/BTS transport data collection (TranStat)
- Toronto Bike Share Stations (XML file)
- Transport for London (TFL)
- Travel Tracker Survey (TTS) for Chicago
- U.S. Bureau of Transportation Statistics (BTS)

- U.S. Domestic Flights 1990 to 2009
- U.S. Freight Analysis Framework since 2007

**Source: United Nations http://data.un.org/DataMartInfo.aspx**

- Commodity Trade Statistics Database
- Energy Statistics Database
- Environment Statistics Database
- FAO Data
- Gender Info
- Global Indicator Database
- Greenhouse Gas Inventory Data
- Human Development Indices: A statistical update 2012
- Indicators on Women and Men
- INDSTAT
- Industrial Commodity Statistics Database
- International Financial Statistics
- Key Indicators of the Labour Market, 7th Edition
- LABORSTA
- Millennium Development Goals Database
- National Accounts Estimates of Main Aggregates
- National Accounts Official Country Data
- OECD Data
- The State of the World's Children
- UIS Data Centre
- UNAIDS Data
- UNHCR Statistical Database
- UNODC Homicide Statistics 2012
- UNSD Demographic Statistics
- WHO Data
- World Contraceptive Use
- World Development Indicators
- World Fertility Data
- World Marriage Data
- World Meteorological Organization Standard Normals
- World Population Prospects: The 2012 Revision
- World Statistics Pocketbook
- World Telecommunication/ICT Indicators Database
- World Tourism Data
- WTI Data

**Source: http://www.kdnuggets.com/datasets/index.html**

1. 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.
2. BigML big list of public data sources.
3. 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.
4. Bitly 1.usa.gov data, anonymized clicks on gov links.
5. Canada Open Data, pilot project with many government and geospatial datasets.
6. Causality Workbench data repository.

7 Corral Big Data repository at Texas Advanced Computing Center, supporting data-centric science.

8 Data Source Handbook, A Guide to Public Data, by Pete Warden, O'Reilly (Jan 2011).

9 Datacatalogs.org, open government data from US, EU, Canada, CKAN, and more.

10 Data.gov.uk, publicly available data from UK (also London datastore.)

11 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.

12 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.

13 Datamob, public data put to good use.

14 DataSF.org, a clearinghouse of datasets available from the City & County of San Francisco, CA.

15 DataFerrett, a data mining tool that accesses and manipulates TheDataWeb, a collection of many on-line US Goverment datasets.

16 Delve, Data for Evaluating Learning in Valid Experiments

17 EconData, thousands of economic time series, produced by a number of US Government agencies.

18 Enron Email Dataset, data from about 150 users, mostly senior management of Enron.

19 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.

20 FEDSTATS (updated) comprehensive source of US statistics and more https://www.usa.gov/statistics

21 FIMI repository for frequent itemset mining, implementations and datasets.

22 Financial Data Finder at OSU, a large catalog of financial data sets.

23 GDELT: The Global Data on Events, Location and Tone, described by Guardian as "a big data history of life, the universe and everything."

24 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.

25 GeoDa Center, geographical and spatial data.

26 Google ngrams datasets, text from millions of books scanned by Google.

27 Grain Market Research, financial data including stocks, futures, etc.

28 Hilary Mason research-quality Big Data sets collection - many text and image datasets.

29 HitCompanies Datasets, comprehensive data on random 10,000 UK companies sampled from HitCompanies, updated automatically using AI/Machine Learning.

30 ICWSM-2009 dataset contains 44 million blog posts made between August 1st and October 1st, 2008.

31 Infochimps, an open catalog and marketplace for data. You can share, sell, curate, and download data about anything and everything.

32 Investor Links, includes financial data

33 KDD Cup center, with all data, tasks, and results.

34 Kevin Chai list of datasets, for text, SNA, and other fields.

35 KONECT, the Koblenz Network Collection, with large network datasets of all types in order to perform research in the area of network mining.

36 Linking Open Data project, at making data freely available to everyone.

37 Million Song Dataset

38 MIT Cancer Genomics gene expression datasets and publications, from MIT Whitehead Center for Genome Research.

39 ML Data, the data repository of the EU Pascal2 networks.

40 NASDAQ Data Store, provides access to market data.

41 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.

42 National Space Science Data Center (NSSDC), NASA data sets from planetary exploration, space and solar physics, life sciences, astrophysics, and more.

43 Open Data Census, assesses the state of open data around the world.

44 OpenData from Socrata, access to over 10,000 datasets including business, education, government, and fun.

45 Open Source Sports, many sports databases, including Baseball, Football, Basketball, and Hockey.

46 Peter Skomoroch dataset Bookmarks

47 PubGene(TM) Gene Database and Tools, genomic-related publications database

48 Quandl, a collaboratively curated portal to millions of financial and economic time-series datasets.

49 qunb, a platform to find and visualize quantitative data.

50 Robert Schiller data on housing, stock market, and more from his book Irrational Exuberance.

51 SMD: Stanford Microarray Database, stores raw and normalized data from microarray experiments.

52 Jerry Smith dataset collection, with Finance, Government, Machine Learning, Science, and other data.

53 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.

54 StatLib, CMU Datasets Archive.

55 STATOO Datasets part 1 and STATOO Datasets part 2

56 Time Series Data Library

57 Visual Analytics Benchmark Repository.

58 **UCI KDD Database Repository for large datasets used in machine learning and knowledge discovery research.**

59 UCI Machine Learning Repository.

60 UCR Time Series Data Archive, offering datasets, papers, links, and code.

61 United States Census Bureau.

62 Wikiposit, a (virtual) amalgamation of (mostly financial) data from many different sites, allowing users to merge data from different sources

63 Wolfram Alpha disease and patient level dat.

64 Yahoo Sandbox datasets, Language, Graph, Ratings, Advertising and Marketing, Competition

65 Yelp Academic Dataset, all the data and reviews of the 250 closest businesses for 30 universities for students and academics to explore and research.

**Source: http://www.kdnuggets.com/datasets/government-local-public.html**

**Public data catalogs, portals, and services**

- 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.
- Datacatalogs.org, open government data from US, EU, Canada, CKAN, 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.
- Enigma, "Google for public data", provides easy access to government, NGO, and other public domain datasets.
- Freebase, a community-curated database of well-known people, places, and things.
- Google Public Data, with dynamic visualization and exploration tools.
- Knoema World Data Atlas, over 1000 indicators on all countries
- 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.
- Open Data Census, assesses the state of open data around the world.
- Open Data Institute, catalysing the evolution of open data culture to create economic, environmental, and social value.
- Socrata OpenData, provides social data discovery services for opening government, healthcare, energy, education, or environment data.
- Visualing Data big collection of sites and services for accessing data.

**Global, International, UN**

- The World Bank, a comprehensive set of data about development in countries around the globe.
- UN data, a data access system to UN databases
- UNICEF statistics, data analysis and other data about UNICEF work.

**USA: Federal**

- Berkeley-curated US statistical information, by UC Berkeley Graduate School of Journalism, providing a wealth of demographic and other data on a city, county, state and national level.
- Bitly 1.usa.gov data, anonymized clicks on gov links.
- NEW CMS.gov Centers for Medicare and Medicaid Services, Research, Statistics, Data, and Systems.
- 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.
- DataFerrett, a data mining tool that accesses and manipulates TheDataWeb, a collection of many on-line US Goverment datasets.
- EconData, thousands of economic time series, produced by a number of US Government agencies.
- FEDSTATS (updated) a comprehensive source of US statistics and more - new https://www.usa.gov/statistics
- United States Census Bureau.

**USA: State, City, and Local**

- CA: San Francisco Data, a clearinghouse of datasets available from the City and County of San Francisco, CA.
- IL: Chicago data .
- NY: New York NYC Open Dat
- WA: Seattle data .

**Canada**

- Canada Open Data, pilot project with many government and geospatial datasets.

**Europe**

- 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.
- Eurostat, the leading provider of high quality statistics on Europe.
- OECD Data Lab, data visualisations and European data downloads.
- PublicData.eu, access to open, freely reusable datasets from local, regional and national public bodies across Europe.
- Data Publica, l'annuaire des donnees en France, public data about France.
- Paris data.

**Germany**

- Destatis: German Federal Statistics Bureau (Statistisches Bundesamt), all German government statistics.

**Ireland**

- Dublinked, Dublin data.

**Russia**

- Moscow open data (beta), Moscow.
- Perm Open Data Portal (in Russian), Perm.

**UK**

- Data.gov.uk, publicly available data from UK.
- London datastore.
- UK Office for National Statistics (ONS), UK largest independent producer of official statistics and the recognised national statistical institute of the UK.
- UK Data Service, UK largest collection of digital social research data from ESDS, Census Programme, Secure Data Service and others.

**Asia**

**India**

- Census India, data on population, economic activity, literacy, education, housing, urbanisation, fertility, mortality, and more.

**Australia, NZ, and Pacific**

- Data.gov.au provides an easy way to find, access and reuse public datasets from the Australian Government.
- Australian Bureau of Statistics, access to the full range of ABS statistical and reference information.
- Wiki New Zealand, a collaborative website making data about New Zealand accessible for everyone.

**Africa**

- Open Data for Africa, supporting statistical development in Africa as a sound basis for designing and managing effective development policies for reducing poverty on the continent.

**Source:**

List

**Source: http://aws.amazon.com/publicdatasets/**

List

**Available Public Data Sets on AWS**

Click here for the detailed list of available data sets. Here are some examples of popular Public Data Sets:

- NASA NEX: A collection of Earth science data sets maintained by NASA, including climate change projections and satellite images of the Earth's surface
- Common Crawl Corpus: A corpus of web crawl data composed of over 5 billion web pages
- 1000 Genomes Project: A detailed map of human genetic variation
  Google Books Ngrams: A data set containing Google Books n-gram corpuses
- US Census Data: US demographic data from 1980, 1990, and 2000 US Censuses

- Freebase Data Dump: A data dump of all the current facts and assertions in the Freebase system, an open database covering millions of topics

**Source: http://kevinchai.net/datasets**

List

### Blog articles which provide dataset directories

http://conflate.net/inductio/2008/02/a-meta-index-of-data-sets/ – excellent article listing available data sets in the area of machine learning and inference
http://www.datawrangling.com/some-datasets-available-on-the-web.html
http://www.daniel-lemire.com/blog/data-for-data-mining/ – has blog, tag cloud, wiki dataset categories
http://www.kirix.com/blog/category/data-tagssearch/
http://mobblog.cs.ucl.ac.uk/datasets/
http://www.readwriteweb.com/archives/where_to_find_open_data_on_the.php – Article containing a list of available dataset websites

### Dataset directories

http://www.quora.com/Data/Where-can-I-get-large-datasets-open-to-the-public – Public datasets listed on a Quora Q&A thread.
http://caw2.barcelonamedia.org/node/7 – Content Analysis for the Web 2.0 (CAW 2.0) Workshop – part of 18th International Conference of the World Wide Web. Contains training and test datasets from Twitter, MySpace, Slashdot, Ciao and Kongregate.
http://kdd.ics.uci.edu/ – has a machine learning repository
http://archive.ics.uci.edu/ml/datasets.html http://ckan.net/ – listing of links to various datasets
http://www.ldc.upenn.edu/Obtaining/ – Linguistic data consortium catalog
http://www.swivel.com/data_sets
http://datamob.org/datasets
http://infochimps.org/
http://www.freebase.com/
http://numbrary.com/
http://theinfo.org/
http://www.trustlet.org/wiki/Repositories_of_datasets
http://del.icio.us/kirixstrata/publicdata
http://services.alphaworks.ibm.com/manyeyes/browse/data?q=null
http://googleresearch.blogspot.com/ – google research has stated thathttp://research.google.com will soon host open-source scientific datasets –http://blog.wired.com/wiredscience/2008/01/google-to-provi.html – watch this space.
http://data.un.org/
http://www.data360.org/index.aspx
http://tunedit.org/search?q=arff – 800 datasets in ARFF format for different problems and application domains
http://wikiposit.com
http://gsociology.icaap.org/dataupload.html – The Global Social Change Research Project – social, political and economic datasets

### Data sets for a specific field

http://kaggle.com/ – machine learning competitions with data provided by organisations with prize money
http://theinfo.org/get/data – good list here – pay attention to web/news/blogs and Text/Language categories as well as trust network data
http://research.microsoft.com/nlp/ – look under data sets
http://nlp.stanford.edu/links/statnlp.html – look under corpora
http://trec.nist.gov/data/reuters/reuters.html – Reuters Corpora – contains large collection of news stories for use in Natural Language Processing, Information Retrieval and Machine Learning Systems (need to order CDs)

http://trec.nist.gov/data.html – Text retrieval. Has spam, web, question answering, blog and ad hoc (e.g. relevance judgement) tracks
http://plg.uwaterloo.ca/~gvcormac/treccorpus/ (300MB) – Spam Corpus 2005
http://plg.uwaterloo.ca/~gvcormac/treccorpus06/ (75MB – english, 60MB chinese) – Spam Corpus 2006
http://trec.nist.gov/data/reljudge_eng.html – Relevance Judgement
http://ir.dcs.gla.ac.uk/test_collections/blog06info.html (25GB – costs 400 GBP) – Blog 06 data
http://trec.nist.gov/data/qamain.html – Question Answering (many tracks)
http://trec.nist.gov/data/novelty.html – Novelty (some relevance) -

http://infochimps.org/tag/language/datasets – languages
http://infochimps.org/tag/lexicon/datasets – lexicon
http://infochimps.org/tag/lexical/datasets – lexical

http://wordnet.princeton.edu/ – Lexical database that is handy for computational linguistics and natural language processing
http://www.dmoz.org/Computers/Artificial_Intelligence/Machine_Learning/Datasets/ – Machine learning datasets
http://cervisia.org/machine_learning_data.php – Machine learning datasets – benchmark data for comparing different algorithms of your classifier is recommended fromhttp://www.ci.tuwien.ac.at/~meyer/benchdata/
http://mill.ucsd.edu/index.php?page=Datasets&subpage=Overview
http://www.trustlet.org/wiki/Trust_network_datasets#Released_datasets – Trust datasets – includes Epinions
http://stuff.metafilter.com/infodump/ – Metafilter – contains posts, comments, tags, favourites, contact and user data
http://an.kaist.ac.kr/traces/IMC2007.html – YouTube dataset
http://socialnetworks.mpi-sws.mpg.de/ – social network dataset
http://people.csail.mit.edu/jrennie/20Newsgroups/ – newsgroup dataset
http://www.yr-bcn.es/webspam/datasets/ – Webspam datasets

**Link Analysis**

http://www.cs.toronto.edu/~tsap/experiments/datasets/index.html
http://www.cs.toronto.edu/~tsap/experiments/download/download.html

**Recommender systems**

http://www.grouplens.org/ – MovieLens
http://www.ieor.berkeley.edu/~goldberg/jester-data/ – Jester
http://www.netflixprize.com/ – Netflix
http://www.informatik.uni-freiburg.de/~cziegler/BX/ – Book Crossing

**Forums**

http://weimo.de/node/642 – Nabble.com + user ratings of posts

**Blogs**

http://ebiquity.umbc.edu/resource/html/id/212/Splog-Blog-Dataset – Spam blogs (splogs)
http://www.icwsm.org/data.html – 14 million posts, 3 million weblogs – apparently no longer available since Dec 8, 2006
http://ir.dcs.gla.ac.uk/test_collections/blog06info.html – but costs 400 GBP!

**Wikis**

http://labs.systemone.at/wikipedia3 – wikipedia 3 providing wikipedia datasets
http://download.wikipedia.org/ – official wikipedia database dumps (very large)
http://download.freebase.com/wex/ – English wikipedia articles that have been transformed into XML – all files ~ 55GB
http://dbpedia.org/About – structured information from wikipedia – dataset of this is available

**Webpages**

http://www.archive.org/web/web.php – 85 billion webpages archived since 1996

**Misc**

http://opentick.com/ – Stock data
http://lib.stat.cmu.edu/datasets/ – miscellaneous datasets
http://lib.stat.cmu.edu/jasadata/ – datasets from Journal of the American Statistical Association
http://musicbrainz.org/ – music dataset
http://www.jigsaw.com/ – directory of company & business professional dataset
http://www.librarything.com/ – library catalogue
http://www.imeem.com/developers – media library
http://www.scribd.com/doc/9582/integrating-wikipediawordnet – article talking about integrating Wordnet and Wikipedia with YAGO (an extensible and light-weight ontology)
http://wiki.openstreetmap.org/index.php/Potential_Datasources – country maps
http://rdf.dmoz.org/ – open directory project dataset

**Source: http://www.quora.com/Data/Where-can-I-find-large-datasets-open-to-the-public**

List

**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://quandl.com
- Social Network Analysis Interactive Dataset Library (Social Network Datasets)
- 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
https://datamarket.azure.com/
http://developer.yahoo.com/geo/g...
http://econ.worldbank.org/datasets
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://timetric.com/public-data/
https://wist.echo.nasa.gov/~wist...
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.delicious.com/jbaldwi...
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/
http://www.nationalarchives.gov....
http://www.nyc.gov/html/datamine...
http://www.ordnancesurvey.co.uk/...
http://www.philwhln.com/how-to-g...
http://www.imdb.com/interfaces
http://imat-relpred.yandex.ru/en...
http://www.dados.gov.pt/pt/catal...
http://knoema.com
http://daten.berlin.de/
http://www.qunb.com
http://databib.org/
http://datacite.org/
http://data.reegle.info/
http://data.wien.gv.at/
http://data.gov.bc.ca

- https://d396qusza40orc.cloudfron... (large collection from Coursera's *Data Analysis* course)
- NCAR - Climate Data Guide
- | Ecological Data Wiki
- http://www.reddit.com/r/datasets
- http://berkeleyearth.org/dataset... - Berkeley Earth dataset
- http://static.reddit.com/RedditS... - **massive** survey of Redditors and their preferences - see http://blog.reddit.com/2011/09/w... for some analysis

- Welcome to the CRCNS data sharing website - for neuroscience
- http://archiveteam.org/index.php... - Old archives of websites that no longer exist. Includes data on the affinities of 60,000+ Reddit users
- http://www.r-bloggers.com/datase... - Datasets to practice your data mining - discussed at http://www.reddit.com/r/MachineL...
- http://sarahsinbox.com/ - Sarah Palin emails - analyzed by Edwin Chen usingLatent Dirichlet Allocation (LDA) - see http://blog.echen.me/2011/06/27/...
- http://www.ers.usda.gov/Data/ - USDA Economic Research Service datasets
- http://www.mortality.org/ - human mortality datasets
- http://www.fda.gov/Food/FoodSafe... - FDA pesticide datasets
- http://www.ams.usda.gov/AMSv1.0/pdp - USDA pesticide datasets
- http://geosci.uchicago.edu/~rtp1... - Principles of Planetary Climate - also contains links to **a lot** of Earth climate data
- Climatology: What are some historical weather databases?
- http://www.epa.gov/data/ - EPA data
- http://data.giss.nasa.gov/ - NASA GISS data
- http://jimwatsonsequence.cshl.edu/ - James Watson's DNA sequence
- http://evidence.personalgenomes.... - public genomes of people enrolled in the personal genome project - includes genomes of Steven Pinker and Esther Dyson. http://evidence.personalgenomes.... for their genomes
- http://voteview.org/downloads.asp - Congressional Voting datasets (probably contains *everything* about what any politician voted for)
- http://www.norc.uchicago.edu/GSS... - General Social Survey. For tutorial, see http://blogs.discovermagazine.co...
- http://www.cfa.harvard.edu/hitran/ - high-resolution transmission molecular absorption database. HITRAN on the web: http://hitran.iao.ru/molecule

Some others:

- http://www.cdc.gov/nchs/nhanes/n... - National Health and Nutrition Examination Survey
- http://www.nlsinfo.org/ordering/... - NSLY data (sociology) [1]
- http://road.hmdc.harvard.edu/ - election datasets (only 1984-1990 though)

**Source: http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html**

| | |
|---|---|
| datasets-package | The R Datasets Package |

**-- A --**

| | |
|---|---|
| ability.cov | Ability and Intelligence Tests |
| airmiles | Passenger Miles on Commercial US Airlines, 1937-1960 |
| AirPassengers | Monthly Airline Passenger Numbers 1949-1960 |
| airquality | New York Air Quality Measurements |
| anscombe | Anscombe's Quartet of 'Identical' Simple Linear Regressions |
| attenu | The Joyner-Boore Attenuation Data |
| attitude | The Chatterjee-Price Attitude Data |

| | |
|---|---|
| austres | Quarterly Time Series of the Number of Australian Residents |

**-- B --**

| | |
|---|---|
| beaver1 | Body Temperature Series of Two Beavers |
| beaver2 | Body Temperature Series of Two Beavers |
| beavers | Body Temperature Series of Two Beavers |
| BJsales | Sales Data with Leading Indicator |
| BJsales.lead | Sales Data with Leading Indicator |
| BOD | Biochemical Oxygen Demand |

**-- C --**

| | |
|---|---|
| cars | Speed and Stopping Distances of Cars |
| ChickWeight | Weight versus age of chicks on different diets |
| chickwts | Chicken Weights by Feed Type |
| CO2 | Carbon Dioxide Uptake in Grass Plants |
| co2 | Mauna Loa Atmospheric CO2 Concentration |
| crimtab | Student's 3000 Criminals Data |

**-- D --**

| | |
|---|---|
| datasets | The R Datasets Package |
| discoveries | Yearly Numbers of Important Discoveries |
| DNase | Elisa assay of DNase |

**-- E --**

| | |
|---|---|
| esoph | Smoking, Alcohol and (O)esophageal Cancer |
| euro | Conversion Rates of Euro Currencies |
| euro.cross | Conversion Rates of Euro Currencies |
| eurodist | Distances Between European Cities |
| EuStockMarkets | Daily Closing Prices of Major European Stock Indices, 1991-1998 |

**-- F --**

| | |
|---|---|
| faithful | Old Faithful Geyser Data |
| fdeaths | Monthly Deaths from Lung Diseases in the UK |
| Formaldehyde | Determination of Formaldehyde |
| freeny | Freeny's Revenue Data |
| freeny.x | Freeny's Revenue Data |
| freeny.y | Freeny's Revenue Data |

**-- H --**

| | |
|---|---|
| HairEyeColor | Hair and Eye Color of Statistics Students |
| Harman23.cor | Harman Example 2.3 |

| | |
|---|---|
| Harman74.cor | Harman Example 7.4 |

**-- I --**

| | |
|---|---|
| Indometh | Pharmacokinetics of Indomethacin |
| infert | Infertility after Spontaneous and Induced Abortion |
| InsectSprays | Effectiveness of Insect Sprays |
| iris | Edgar Anderson's Iris Data |
| iris3 | Edgar Anderson's Iris Data |
| islands | Areas of the World's Major Landmasses |

**-- J --**

| | |
|---|---|
| JohnsonJohnson | Quarterly Earnings per Johnson & Johnson Share |

**-- L --**

| | |
|---|---|
| LakeHuron | Level of Lake Huron 1875-1972 |
| ldeaths | Monthly Deaths from Lung Diseases in the UK |
| lh | Luteinizing Hormone in Blood Samples |
| LifeCycleSavings | Intercountry Life-Cycle Savings Data |
| Loblolly | Growth of Loblolly pine trees |
| longley | Longley's Economic Regression Data |
| lynx | Annual Canadian Lynx trappings 1821-1934 |

**-- M --**

| | |
|---|---|
| mdeaths | Monthly Deaths from Lung Diseases in the UK |
| morley | Michelson Speed of Light Data |
| mtcars | Motor Trend Car Road Tests |

**-- N --**

| | |
|---|---|
| nhtemp | Average Yearly Temperatures in New Haven |
| Nile | Flow of the River Nile |
| nottem | Average Monthly Temperatures at Nottingham, 1920-1939 |
| npk | Classical N, P, K Factorial Experiment |

**-- O --**

| | |
|---|---|
| occupationalStatus | Occupational Status of Fathers and their Sons |
| Orange | Growth of Orange Trees |
| OrchardSprays | Potency of Orchard Sprays |

**-- P --**

| | |
|---|---|
| PlantGrowth | Results from an Experiment on Plant Growth |

| | |
|---|---|
| precip | Annual Precipitation in US Cities |
| presidents | Quarterly Approval Ratings of US Presidents |
| pressure | Vapor Pressure of Mercury as a Function of Temperature |
| Puromycin | Reaction Velocity of an Enzymatic Reaction |

**-- Q --**

| | |
|---|---|
| quakes | Locations of Earthquakes off Fiji |

**-- R --**

| | |
|---|---|
| randu | Random Numbers from Congruential Generator RANDU |
| rivers | Lengths of Major North American Rivers |
| rock | Measurements on Petroleum Rock Samples |

**-- S --**

| | |
|---|---|
| Seatbelts | Road Casualties in Great Britain 1969-84 |
| sleep | Student's Sleep Data |
| stack.loss | Brownlee's Stack Loss Plant Data |
| stack.x | Brownlee's Stack Loss Plant Data |
| stackloss | Brownlee's Stack Loss Plant Data |
| state | US State Facts and Figures |
| state.abb | US State Facts and Figures |
| state.area | US State Facts and Figures |
| state.center | US State Facts and Figures |
| state.division | US State Facts and Figures |
| state.name | US State Facts and Figures |
| state.region | US State Facts and Figures |
| state.x77 | US State Facts and Figures |
| sunspot.month | Monthly Sunspot Data, from 1749 to "Present" |
| sunspot.year | Yearly Sunspot Data, 1700-1988 |
| sunspots | Monthly Sunspot Numbers, 1749-1983 |
| swiss | Swiss Fertility and Socioeconomic Indicators (1888) Data |

**-- T --**

| | |
|---|---|
| Theoph | Pharmacokinetics of Theophylline |
| Titanic | Survival of passengers on the Titanic |
| ToothGrowth | The Effect of Vitamin C on Tooth Growth in Guinea Pigs |
| treering | Yearly Treering Data, -6000-1979 |
| trees | Girth, Height and Volume for Black Cherry Trees |

**-- U --**

| | |
|---|---|
| UCBAdmissions | Student Admissions at UC Berkeley |
| UKDriverDeaths | Road Casualties in Great Britain 1969-84 |

| | |
|---|---|
| UKgas | UK Quarterly Gas Consumption |
| UKLungDeaths | Monthly Deaths from Lung Diseases in the UK |
| USAccDeaths | Accidental Deaths in the US 1973-1978 |
| USArrests | Violent Crime Rates by US State |
| USJudgeRatings | Lawyers' Ratings of State Judges in the US Superior Court |
| USPersonalExpenditure | Personal Expenditure Data |
| uspop | Populations Recorded by the US Census |

**-- V --**

| | |
|---|---|
| VADeaths | Death Rates in Virginia (1940) |
| volcano | Topographic Information on Auckland's Maunga Whau Volcano |

**-- W --**

| | |
|---|---|
| warpbreaks | The Number of Breaks in Yarn during Weaving |
| women | Average Heights and Weights for American Women |
| WorldPhones | The World's Telephones |
| WWWusage | Internet Usage per Minute |

**Source: http://www.reddit.com/r/datasets/**

List

- Looking for data sets that illustrate poisson distributions. (self.datasets)
- Looking for International Income by Quartile/Quintile (self.datasets)
- Looking for Comic Book Datasets (self.datasets)
- Looking for US National Parks Datasets (self.datasets)
- [ProPublica/CMMS] Average wait times for emergency rooms across the country (projects.propublica.org)
- [REQUEST] Solar power projects logged (self.datasets)
- Request: Looking for a dataset with age, income, crime, etc. by location (self.datasets)
- Dataset from Boston Police's suspended license plate scanning program ALPR (muckrock.com)
- The 911Dataset Project: 3TB across 254,822 files (911datasets.org)
- Snowfall totals? (self.datasets)
- [REQUEST] Looking for a large dataset containing recipes (self.datasets)
- NFL Game Metadata Since 1980 (aragorn.org)
- Where can I find/buy retail price data? (self.datasets)
- Looking for temporal food pattern data (self.datasets)
- NYC Crime Map - Fully interactive pinpoint mapping of all recorded felonies from 2012 to YTD (maps.nyc.gov)
- Please help me test my platform for testing my demographic crowdsourcing platform (self.datasets)
- Advanced Play by Play Stats for Active NBA Players (aragorn.org)
- Meta data merged with play by play data from the two earlier posts (self.datasets)
- Metadata on the 420 Players on NBA Rosters (aragorn.org)
- Medicare Provider Charge Data: Inpatient - Centers for Medicare & Medicaid Services (cms.gov)
- starcraft datset (archive.ics.uci.edu)
- [REQUEST] Network dataset with one large hub (self.datasets)
- Where should I look for environmental data sets? (self.datasets)

**Source: http://www.nber.org/data/**

| | |
|---|---|
| Official Business Cycle Dates | NBER |
| "The American Business Cycle: Continuity and Change"   Historic Data Tables | Gordon |
| Experimental Coincident, Leading and Recession Indexes | Stock, Watson |
| Index of African Governance | Rotberg, Gisselquist |
| Penn-World Tables | Feenstra, Inklaar, Timmer |
| Barro-Lee | Barro, Lee |
| Cross-country Historical Adoption of Technology (CHAT) data | Comin, Hobijn |
| Economic Policy Uncertainty | Baker, Bloom, Davis |
| A History of U.S. Foreign-Exchange-Market Interventions | Bordo, Humpage, Schwartz |
| Occupational Wages around the World | Freeman, Oostendorp |
| Macro History Database | NBER |
| Savings, Investment, and Gold in 13 countries (1850-1945) | Jones, Obstfeld |
| Social Security Pension Reform in Europe | Feldstein, Siebert |
| Historical Cross-Country Technological Adoption: Dataset | Comin, Hobijn |
| Facts and Fantasies about Commodity Futures | Gorton, Rouwenhorst |
| US Industrial Production Index 1790 - 1915 | Davis |

**Industry, Productivity, and Digitization Data**

| | |
|---|---|
| Job Creation and Destruction Data | Haltiwanger et al |
| Management Practices Data | Bloom, Van Reenen |
| Manufacturing Industry Productivity Database | Becker, Gray, Marvakov |
| Internet and Economy Digitization Report | Shiller |
| Public Sector Collective Bargaining Law Data | Valletta, Freeman |
| Form 990 data on tax exempt organizations | IRS |

**International Trade Data**

| | |
|---|---|
| Price Quantity Indexes and Values for U.S. Exports and Imports, 1879-1923 | Lipsey |

| | |
|---|---|
| SITC Rev 2 and NAICS (1997) | Feenstra,Lipsey |
| U.S. Trade by 1972-SIC category, 1958-1994 | Feenstra |
| U.S. Trade by 1987-SIC, 1972-2005; NAICS 1989-2005; HS 1989-2008 | Schott |
| Concordance between HS and SIC/NAICS; Concordance of HS codes over time | Pierce and Schott |
| U.S. Imports by TSUSA, HS, SITC, 1972-2001 | Feenstra |
| U.S. Imports by SAS and Stata, 1972-2001 | Feenstra |
| U.S. Exports by TSUSA, HS, SITC, 1972-2001 | Feenstra |
| U.S. Exports by SAS and Stata, 1972-2001 | Feenstra |
| U.S. Tariffs, 1989-2001 | Romalis |
| U.S. Antidumping Database and Links | Blonigen |
| World Trade Data ( choose World Import and Export Data ) | Feenstra, Lipsey |

### Individual Data

| | |
|---|---|
| Angrist Archive | Joshua Angrist |
| Boston Youth Labor (Market) Survey, 1980, 1989 | Freeman, Katz |
| Collaborative Perinatal (CPP) | NINCDS |
| Consumer Expenditure Survey Extracts | Harris, Sabelhaus (CBO) |
| Current Population Survey | BLS |
| Fatality Analysis Reporting System (FARS) Data | NHTSA |
| Gould Sample | Costa |
| National Health and Nutrition Examination Survey (NHANES) | NCHS |
| Reading National Health Interview Survey (NHIS) Data with SAS, SPSS, or Stata | Roth |
| Survey of Economic Expectations | Dominitz, Manski |
| Survey of Income and Program Participation | Census |
| Survey of Program Dynamics | Census |
| Thorndike-Hagen | Thorndike, Hagen |
| Union Army Data Set | Fogel |
| Worker Representation survey | Freeman, Rogers |

### Hospital/Provider Data

| | |
|---|---|
| CMS' Prospective Payment System (PPS) | CMS |
| Reading CMS'  Healthcare Cost Report Information System (HCRIS) datasets using SAS | CMS |
| CMS's National Plan and Provider Enumeration System (NPPES) Files | CMS |
| CMS' National Provider Identifier (NPI) to Unique Physician Identification Number (UPIN) Crosswalk | CMS |
| CMS' National Provider Identifier (NPI) to State License Crosswalk | CMS |
| CMS' Provider of Service (POS) files | CMS |
| CMS' Medicare Provider Charge Data | CMS |
| CMS' ICD-9-CM to and from ICD-10-CM and ICD-10-PCS Crosswalk or General Equivalence Mappings | CMS |
| CMS's CBSA, MSA, and State Wage Index Files | CMS |
| CMS' SSA to FIPS CBSA and MSA County Crosswalks | CMS |
| CMS' SSA to FIPS State and County Crosswalks | CMS |

**Demographic and Vital Statistics**

| | |
|---|---|
| Vital Statistics Books ( Historical ) | NCHS |
| Vital Statistics Births | NCHS |
| Interactive index to Vital Statistics Births 1931-1968 | NCHS |
| Reading SEER U.S. County Population Data with SAS, SPSS, or Stata 1969-on | Roth |
| Vital Statistics Births and Infant Mortality 1920-1945 | Cutler, Norberg, Norton |
| Vital Statistics Births 1940-1968 | Finkelstein, Heidi Williams |
| Vital Statistics Mortality Data | NCHS |
| Vital Statistics Deaths - Historical 1900 - 1936 | Grant Miller |
| Vital Statistics Marriage and Divorce | NCHS |
| US Decennial Population by County and State 1900-1990 | Roth |
| US Intercensal Population by County and State 1970-2009 | Roth, James Wang |
| US Intercensal Population by State, Age and Sex 1970-1999 | Census |
| Work-Family Policies and Other Data | Waldfogel, Han, Ruhm |

**Patent and Scientific Papers Data**

| | |
|---|---|
| U.S. Patents | Hall, Jaffe, Tratjenberg |
| NBER-Rensselaer Polytechnic Institute Scientific Papers Database | Adams, Clemmons |
| Nobel Laureate Data | Jones, Weinberg |

**Other Data**

| | |
|---|---|
| Data Appendixes from NBER Working Papers and Books | NBER |
| School District Databook (SDDB) | NCES |
| Tax Model File Documentation | Feenberg |
| Segregation Data | Cutler, Glaeser, Vigdor |
| State Constitutions Project | Wallis |
| NIH CRISP (Computer Retrieval of Information on Scientific Projects) | Lichtenberg |
| Business Travel Index | Borenstein |
| Reading PUMS Data (Public Use Microdata Sample) for 1990 and 2000 with SAS, SPSS, or Stata | Roth |
| Reading 1970 Census Summary File 1A with SAS, SPSS, or Stata | Roth |
| Reading large datasets into Access or Excel | Roth |
| Matched set of Indonesian BPS subdistrict (kecamatan) codes for the period 1976-2002 for use with corresponding BPS data | Olken |
| Resume audit experiment computer program | Lahey |
| Older data formerly at anonymous FTP | NBER |
| Links to child health datasets | Norberg |

**Source: http://www.wto.org/english/res_e/statis_e/data_pub_e.htm**

List

- Statistics database
  Online database containing time series on international trade, country trade, tariff and services profiles
- Tariff Download Facility
  Sophisticated detailed and interactive tariff analysis
- Tariff Analysis Online
  Simpler, standardized tariff statistics, mainly for downloading
- RTA database
  Regional trade agreements
- PTA database
  Preferential trade arrangements
- I-TIP database
  A single entry point for WTO information on trade policy measures covering both tariff and non-tariff measures affecting trade in goods as well as information on trade in services

**Source: http://www.imf.org/external/data.htm**

- World Economic Outlook Databases (WEO) updated
- International Financial Statistics (IFS)
- Principal Global Indicators (PGI)
- Balance of Payments Statistics (BOPS)
- Coordinated Direct Investment Survey (CDIS)
- Coordinated Portfolio Investment Survey (CPIS) **updated**
- Currency Composition of Official Foreign Exchange Reserves (COFER)
- Data Template on International Reserves and Foreign Currency Liquidity
- Financial Access Survey (FAS)
- Financial Soundness Indicators (FSIs)
- G-20 Surveillance Notes
- Joint External Debt Hub
- Monitoring of Fund Arrangements Database (MONA)
- Primary Commodity Prices
- Public Sector Debt Statistics Online Centralized Database
- Quarterly External Debt Statistics (QEDS)

**Source: http://blog.visual.ly/data-sources/**

Government and political data

- **Data.gov:** This is the  go-to resource for government-related data. It claims to have up to 400,000 data sets, both raw data and geo spatial, in a variety of formats.
- The only caveat in using the data sets is you have to make sure you clean them, since many have missing values and characters.
- **Socrata** is another good place to explore government-related data. One great thing about Socrata is they have some visualization tools that make exploring the data easier.
- **City-specific government data**: Some cities have their own data portals setup to browse through city-related data. For example, at **San Francisco Data** you can browse through everything from crime statistics to parking spot available in the city.
- The UN and UN-related sites like **UNICEF** and the **World Health Organization** are rich with all kinds of data, from mortality rates to world hunger statistics.
- **The Census Bureau** houses a ton of information about our lives around income, race, education, population and business.

**Data aggregators**

These are the places that house data from all kinds of sources. Sometimes it's easier to find something here related to a specific category.

- **Programmable Web**: A really useful resource to explore API's and also mashups of different API's.
- **Infochimps** have a data marketplace that offers thousands of public and propietary data sets for download and API access, in a wide range of categories, from historical Twitter and OK Cupid data, to geo locations data, in different formats. You can even upload you own data if you like.
- **Data Market** is a good place to explore data related to economics, healthcare, food and agriculture, and the automotive industry.
- **Google Public data explorer** houses a lot of data from world development indicators, OECD and human development indicators, mostly related to economics data and the world.
- **Junar** is a great data scraping service that also houses data feeds.
- **Buzzdata** is a social data sharing service that allows you to upload your own data and connect and follow others who are uploading their own data.

**3. Social data**

Usually, the best place to get social data for an API is the site itself: Instagram, GetGlue, Foursquare, pretty much all social media sites have their own API's. Here are more details on the most popular ones.

- **Twitter**: Access to the Twitter API for historical uses is fairly limited, to 3200 tweets. For more, check out PeopleBrowsr, Gnip (also offers historical access to the WP Automattic data feed),DataSift, Infochimps, Topsy.
- **Foursquare**: They have their own API and you can get it through Infochimps, as well.
- **Facebook**: The Facebook graph API is the best resource for Facebook.
- **Face.com**: A great tool for facial recognition data.

**4. Weather data**

- **Wunderground** has detailed weather information and also let's you search historical data by zip code or city. It gives temperature, wind, precipitation and hourly observations for that day.
- **Weatherbase** has detailed weather stats on temperature, rain and humidity of nearly 27,000 cities.

**5. Sports data**

These three sites have comprehensive information on teams, players coaches and leaders by season.

- **Football**
- **Baseball**
- **Basketball**

**ESPN** recently came up with its own API, too. You have to be a partner to get access to their data.

**6. Universities and research**

Searching the work of academics who specialize in a particular area is always a great place to find some interesting data.

If you come across specific data that you would like to use, say, in a research paper, the best way to go is to contact the professor directly. (That is how we got the data for our What are the Odds piece, which is one of the most-viewed infographics on the web.)

One university that makes some of the datasets used in its courses publicly available is **UCLA**.

**7. News data**

**The New York Times** has a great API and a really good explorer to access any article in the publication. The data is returned in json format.

**The Guardian Data Blog** regularly posts visualizations and makes data available through a Google docs format. The great thing about this is that that the data has already been cleaned.

**CDC Data - Source: http://www.cdc.gov/ncbddd/disabilityandhealth/datasets.html**

Behavioral Risk Factor Surveillance System (BRFSS)
The BRFSS is a telephone survey that tracks national and state-specific health risk behaviors of adults, 18 years of age or older, residing in the United States. The BRFSS is conducted by the 50 states, the District of Columbia, and three territories (Guam, Puerto Rico, and the U.S. Virgin Islands) and is administered and supported by the Division of Adult and Community Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention (CDC).

National Health Interview Survey (NHIS)
The NHIS is a multi-purpose, nationwide household health survey of the U.S. civilian noninstitutionalized population conducted annually by the National Center for Health Statistics (NCHS), CDC, to produce national estimates for a variety of health indicators. In 1994 and 1995, the NHIS included a special supplement on disability.

National Health and Nutrition Examination Survey (NHANES)
NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines information from interviews and physical examinations.

National Survey of Family Growth (NSFG)
The NSFG gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. The survey results are used by the U.S. Department of Health and Human Services and others to plan health services and health education programs, and to do statistical studies of families, fertility, and health.

American Community Survey (ACS)
The ACS is a mail survey that provides demographic, socioeconomic, and housing information about communities in between the 10-year census. The ACS is conducted by the U.S. Census Bureau. The survey is sent to a sample of households in the United States. The ACS identifies serious difficulty in four basic areas of functioning: vision, hearing, ambulation, and cognition. The ACS also includes two questions to identify people with difficulties that might affect their ability to live independently.

Medical Expenditure Panel Survey (MEPS)
The MEPS comprise a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. The MEPS is the most complete source of data on the cost and use of health care and health insurance coverage.

Survey of Income and Program Participation (SIPP)
The SIPP is a multipanel, longitudinal survey conducted by the U.S. Census Bureau. The SIPP covers the civilian, noninstitutionalized population of residents of the United States, and collects data on the sources and amount of individual income, labor force information, program participation and eligibility data, and general demographic characteristics. The SIPP also includes disability supplements that ask questions to determine individual disability status.

Current Population Survey (CPS)
The CPS is a monthly survey of about 50,000 households conducted by the U.S. Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years. In June 2008, questions were

added to the CPS to identify people with a disability among the civilian noninstitutional population 16 years of age or older. Monthly labor force data are released from the CPS for people with a disability. The collection of these data is sponsored by the Department of Labor's Office of Disability Employment.

**Personality Testing - Source: http://personality-testing.info/_rawdata/**

| Updated | Description | Variables | n | Download |
|---|---|---|---|---|
| 5/14/2014 | Answers to Cattell's 16 Personality Factors Test with items from the IPIP. | 163 likert rated items, gender, age, country and accuracy. | 49159 | 16PF |
| 9/6/2012 | Answers to the Narcissistic Personality Inventory, constructed with the version from Raskin and Terry (1988). | 40 multiple choice, gender, age, time elapsed | 11243 | NPI |
| 6/18/2012 | Answers to the Machivallianism Test, a version of the MACH-IV from Christie and Geis (1970). | 20 likert rated items, gender, age, time elapsed | 13156 | MACH2 |
| 5/18/2014 | Answers to the Big Five Personality Test, constructed with items from the International Personality Item Pool. | 50 likert rated statements, gender, age, race, native language, country | 19719 | BIG5 |
| 7/22/2012 | Answers to the Taylor Manifest Anxiety Scale, from Taylor (1953). | 50 true false statements, gender, age | 5410 | TMA |
| 9/6/2012 | Answers to the Humor Styles Questionnaire, from Martin et. al. (2003). | 32 likert rated items, gender, age, self-rated accuracy | 1071 | HSQ |
| 7/16/2012 | Answers to the Empathizing-Systemizing Test, a combined version of Simon Baron-Cohen's empathizing and systemizing quotients. | 120 likert rated items, gender, age, self-rated accuracy | 13256 | EQSQ |
| 8/5/2013 | Answers to the Holland Code (RIASEC) Test, constructed with public domain items from the Interest Item Pool. | 48 likert rated statements, gender, age, country, time elapsed and self-rated accuracy. | 8855 | RIASEC |
| 7/16/2012 | Answers to the Sexual Compulsivity Scale from Kalichman and Rompa (1995). | 10 likert rated statements, gender, age | 3376 | SCS |
| 7/18/2012 | Answers to the IPIP Assertiveness, Social confidence, Adventurousness, and Dominance scales used as part of an experimental personality test. | 40 likert rated items, gender, age | 1005 | AS+SC+AD+DO |
| 2/15/2014 | Answers to the Rosenberg Self-Esteem Scale. | 10 scale rated items, gender, age, country | 47974 | RSE |
| 5/25/2012 | Answers to an experimental IQ Test previously offered on this website. | 25 questions/answers, age, gender. | 400 | IQ1 |
| 5/25/2012 | Answers to a sentence completion survey appended to the Holland Code and big five personality tests; at completion of either test takers were solicited to participate (most did). | 6 incomplete sentence responses, gender, age, and big five or RIASEC traits. | 1425 | SENTANCES1 |
| 8/6/2013 | Answers to the Experinces in Close Relationships Scale. | 36 likert rated items, gender, age, county. | 17386 | ECR |
| 9/26/2012 | Answers to the Consideration of Future Consequences Scale. | 12 likert rated items, gender, age, self-rated accuracy. | 614 | CFCS |
| 8/7/2012 | Answers to the Kentucky Inventory of Mindfulness Skills from Baer, Smith and Allen (2004). | 39 likert rated items, gender, age. | 601 | KIMS |

| | | | | |
|---|---|---|---|---|
| 9/6/2012 | Answers to the Multidimensional Sexual Self-Concept Questionnaire. | 100 likert rated items, gender, age and context. | 289 | MSSCQ |
| 8/8/2013 | Answers to the Woodworth Psychoneurotic Inventory. | 116 yes/no questions, gender, age and country. | 6019 | WPI |
| 12/8/2013 | Answers to the Hypersensitive Narcissism Scale and The Dirty Dozen. | 22 scale rated items, gender, age, accuracy and country. | 53981 | HSNS+DD |
| 3/8/2014 | Answers to the Short Dark Triad by Paulhus and Jones (2011). | 27 scale rated items and country. | 18192 | SD3 |
| 4/21/2014 | Answers to the Feminist Perspectives Scale, from Henley, N.; Meng, K.; O'Brien, D.; McCarthy, W.; Sockloskie, R. (1998). "Developing a Scale to Measure the Diversity of Feminist Attitudes". Psychology of Women Quarterly, 22(2), 317-348. | 60 scale rated items, gender, age, country. | 13477 | FPS |
| 5/21/2014 | Answers to the Wagner Preference Inventory, from Wagner, Rudolph F., and Kelly A. Wells. "A refined neurobehavioral inventory of hemispheric preference." Journal of clinical psychology 41.5 (1985): 671-676. | 12 multiple choice questions, country | 13502 | Wagner |
| 5/23/2014 | A user generated corpus of personality test items from a short survey were users prompted to generate descriptions of what was unqiue about their personality. | 3 free response, age, gender, native language, country | 2722 | itemsgen |
| 6/21/2014 | Answers to the IPIP HEXACO equivalent scales. | 240 scale rated items, country | 22786 | HEXACO |

**Source: Awesome Public Datasets**

https://github.com/caesar0301/awesome-public-datasets

- Agriculture
- Biology
- Climate/Weather
- Complex Networks
- Computer Networks
- Contextual Data
- Data Challenges
- Economics
- Education
- Energy
- Finance
- Geology
- GIS/Environment
- Government
- Healthcare
- Image Processing

- Machine Learning
- Museums
- Natural Language
- Physics
- Psychology/Cognition
- Public Domains
- Search Engines
- Social Networks
- Social Sciences
- Software
- Sports
- Time Series
- Transportation
- Complementary Collections

## Agriculture

- U.S. Department of Agriculture's PLANTS Database

## Biology

- 1000 Genomes
- American Gut (Microbiome Project)
- Broad Cancer Cell Line Encyclopedia (CCLE)
- Broad Bioimage Benchmark Collection (BBBC)
- Cell Image Library
- Collaborative Research in Computational Neuroscience (CRCNS)
- Complete Genomics Public Data
- EBI ArrayExpress
- EBI Protein Data Bank in Europe
- Electron Microscopy Pilot Image Archive (EMPIAR)
- ENCODE project
- Ensembl Genomes
- Gene Expression Omnibus (GEO)
- Gene Ontology (GO)
- Global Biotic Interactions (GloBI)
- Harvard Medical School (HMS) LINCS Project
- Human Genome Diversity Project
- Human Microbiome Project (HMP)
- ICOS PSP Benchmark
- International HapMap Project
- Journal of Cell Biology DataViewer
- MIT Cancer Genomics Data
- NCBI Proteins
- NCBI Taxonomy
- NeuroData
- NIH Microarray data or FTP
- OpenSNP genotypes data

- Pathguid - Protein-Protein Interactions Catalog
- Protein Data Bank
- Psychiatric Genomics Consortium
- PubChem Project
- PubGene (now Coremine Medical)
- Sanger Catalogue of Somatic Mutations in Cancer (COSMIC)
- Sanger Genomics of Drug Sensitivity in Cancer Project (GDSC)
- Sequence Read Archive(SRA)
- Stanford Microarray Data
- Stowers Institute Original Data Repository
- Systems Science of Biological Dynamics (SSBD) Database
- Temple University Hospital EEG Database
- The Cancer Genome Atlas (TCGA), available via Broad GDAC
- The Catalogue of Life
- The Personal Genome Project or PGP
- UCSC Public Data
- Universal Protein Resource (UnitProt)
- UniGene

## Climate/Weather

- Australian Weather
- Brazilian Weather - Historical data (In Portuguese)
- Canadian Meteorological Centre
- Climate Data from UEA (updated monthly)
- European Climate Assessment & Dataset
- Global Climate Data Since 1929
- NASA Global Imagery Browse Services
- NOAA Bering Sea Climate
- NOAA Climate Datasets
- NOAA Realtime Weather Models
- The World Bank Open Data Resources for Climate Change
- UEA Climatic Research Unit
- WorldClim - Global Climate Data
- WU Historical Weather Worldwide

## Complex Networks

- AMiner Citation Network Dataset
- CrossRef DOI URLs
- DBLP Citation dataset
- NBER Patent Citations
- Network Repository with Interactive Exploratory Analysis Tools
- NIST complex networks data collection
- Protein-protein interaction network
- PyPI and Maven Dependency Network
- Scopus Citation Database
- Small Network Data

- Stanford GraphBase (Steven Skiena)
- Stanford Large Network Dataset Collection
- Stanford Longitudinal Network Data Sources
- The Koblenz Network Collection
- The Laboratory for Web Algorithmics (UNIMI)
- The Nexus Network Repository
- UCI Network Data Repository
- UFL sparse matrix collection
- WSU Graph Database
- DIMACS Road Networks Collection

## Computer Networks

- 3.5B Web Pages from CommonCraw 2012
- 53.5B Web clicks of 100K users in Indiana Univ.
- CAIDA Internet Datasets
- ClueWeb09 - 1B web pages
- ClueWeb12 - 733M web pages
- CommonCrawl Web Data over 7 years
- CRAWDAD Wireless datasets from Dartmouth Univ.
- Criteo click-through data
- Open Mobile Data by MobiPerf
- Rapid7 Sonar Internet Scans
- UCSD Network Telescope, IPv4 /8 net

## Contextual Data

- Context-aware data sets from five domains or GitHub

## Data Challenges

- Challenges in Machine Learning
- CrowdANALYTIX dataX
- D4D Challenge of Orange
- DrivenData Competitions for Social Good
- ICWSM Data Challenge (since 2009)
- Kaggle Competition Data
- KDD Cup by Tencent 2012
- Localytics Data Visualization Challenge
- Netflix Prize
- Space Apps Challenge
- Telecom Italia Big Data Challenge
- Yelp Dataset Challenge
- Bruteforce Database

## Economics

- American Economic Ass (AEA)
- EconData from UMD
- Economic Freedom of the World Data
- Historical MacroEconomc Statistics
- International Trade Statistics
- Internet Product Code Database
- Joint External Debt Data Hub
- Jon Haveman International Trade Data Links
- OpenCorporates Database of Companies in the World
- Our World in Data
- SciencesPo World Trade Gravity Datasets
- The Atlas of Economic Complexity
- The Center for International Data
- The Observatory of Economic Complexity
- UN Commodity Trade Statistics
- UN Human Development Reports

## Education

- Student Data from Free Code Camp

## Energy

- AMPds
- BLUEd
- COMBED
- Dataport
- ECO
- EIA
- HFED
- iAWE
- Plaid
- REDD
- UK-Dale

## Finance

- CBOE Futures Exchange
- Google Finance
- Google Trends
- NASDAQ
- OANDA
- OSU Financial data
- Quandl

- St Louis Federal
- Yahoo Finance
- NYSE Market Data

## Geology

- Earth Models
- Smithsonian Institution Global Volcano and Eruption Database
- USGS Earthquake Archives

## GIS/Environment

- BODC - marine data of ~22K vars
- Cambridge, MA, US, GIS data on GitHub
- EOSDIS - NASA's earth observing system data
- Factual Global Location Data
- Geo Spatial Data from ASU
- Geo Wiki Project - Citizen-driven Environmental Monitoring
- GeoFabrik - OSM data extracted to a variety of formats and areas
- GeoNames Worldwide
- Global Administrative Areas Database (GADM)
- Homeland Infrastructure Foundation-Level Data
- Integrated Marine Observing System (IMOS) - roughly 30TB of ocean measurements or on S3
- International Institute for Systems Analysis - GIS Datasets
- Landsat 8 on AWS
- List of all countries in all languages
- Marinexplore - Open Oceanographic Data
- National Weather Service GIS Data Portal
- Natural Earth - vectors and rasters of the world
- OpenAddresses
- OpenStreetMap (OSM)
- Pleiades - Gazetteer and graph of ancient places
- Reverse Geocoder using OSM data & additional high-resolution data files
- TIGER/Line - U.S. boundaries and roads
- TwoFishes - Foursquare's coarse geocoder
- TZ Timezones shapfiles
- UN Environmental Data
- World boundaries from the U.S. Department of State
- World countries in multiple formats

## Government

- OpenDataSoft's list of 1,600 open data portals
- A list of cities and countries contributed by community

## Healthcare

- EHDP Large Health Data Sets
- Gapminder World demographic databases
- Medicare Coverage Database (MCD), U.S.
- Medicare Data Engine of medicare.gov Data
- Medicare Data File
- MeSH, the vocabulary thesaurus used for indexing articles for PubMed
- Number of Ebola Cases and Deaths in Affected Countries (2014)
- Open-ODS (structure of the UK NHS)
- OpenPaymentsData, Healthcare financial relationship data
- The Cancer Genome Atlas project (TCGA) and BigQuery table
- World Health Organization Global Health Observatory

## Image Processing

- 10k US Adult Faces Database
- 2GB of Photos of Cats or Archive version
- Affective Image Classification
- Animals with attributes
- Face Recognition Benchmark
- ImageNet (in WordNet hierarchy)
- Indoor Scene Recognition
- International Affective Picture System, UFL
- Massive Visual Memory Stimuli, MIT
- Several Shape-from-Silhouette Datasets
- Stanford Dogs Dataset
- SUN database, MIT
- The Oxford-IIIT Pet Dataset
- YouTube Faces Database
- Adience Unfiltered faces for gender and age classification
- The Action Similarity Labeling (ASLAN) Challenge
- Violent-Flows - Crowd Violence Non-violence Database and benchmark

## Machine Learning

- Delve Datasets for classification and regression (Univ. of Toronto)
- Discogs Monthly Data
- eBay Online Auctions (2012)
- IMDb Database
- Keel Repository for classification, regression and time series
- Labeled Faces in the Wild (LFW)
- Lending Club Loan Data
- Machine Learning Data Set Repository
- Million Song Dataset
- More Song Datasets
- MovieLens Data Sets
- RDataMining - "R and Data Mining" ebook data

- Registered Meteorites on Earth
- Restaurants Health Score Data in San Francisco
- UCI Machine Learning Repository
- Yahoo! Ratings and Classification Data

## Museums

- Canada Science and Technology Museums Corporation's Open Data
- Cooper-Hewitt's Collection Database
- Minneapolis Institute of Arts metadata
- Natural History Museum (London) Data Portal
- Rijksmuseum Historical Art Collection
- Tate Collection metadata
- The Getty vocabularies

## Natural Language

- Blogger Corpus
- CLiPS Stylometry Investigation Corpus
- ClueWeb09 FACC
- ClueWeb12 FACC
- DBpedia - 4.58M things with 583M facts
- Flickr Personal Taxonomies
- Freebase.com of people, places, and things
- Google Books Ngrams (2.2TB)
- Google Web 5gram (1TB, 2006)
- Gutenberg eBooks List
- Hansards text chunks of Canadian Parliament
- Machine Comprehension Test (MCTest) of text from Microsoft Research
- Machine Translation of European languages
- Personae Corpus
- SaudiNewsNet Collection of Saudi Newspaper Articles (Arabic, 30K articles)
- SMS Spam Collection in English
- USENET postings corpus of 2005~2011
- Wikidata - Wikipedia databases
- Wikipedia Links data - 40 Million Entities in Context
- WordNet databases and tools

## Physics

- CERN Open Data Portal
- Crystallography Open Database
- NASA Exoplanet Archive
- NSSDC (NASA) data of 550 space spacecraft
- Sloan Digital Sky Survey (SDSS) - Mapping the Universe

## Psychology/Cognition

- OSU Cognitive Modeling Repository Datasets

## Public Domains

- Amazon
- Archive-it from Internet Archive
- Archive.org Datasets
- CMU JASA data archive
- CMU StatLab collections
- Data360
- Datamob.org
- Google
- Infochimps
- KDNuggets Data Collections
- Microsoft Azure Data Market Free DataSets
- Numbray
- Open Library Data Dumps
- Reddit Datasets
- RevolutionAnalytics Collection
- Sample R data sets
- Stats4Stem R data sets
- StatSci.org
- The Washington Post List
- UCLA SOCR data collection
- UFO Reports
- Wikileaks 911 pager intercepts
- Yahoo Webscope

## Search Engines

- Academic Torrents of data sharing from UMB
- Datahub.io
- DataMarket (Qlik)
- Harvard Dataverse Network of scientific data
- ICPSR (UMICH)
- Institute of Education Sciences
- National Technical Reports Library
- Open Data Certificates (beta)
- OpenDataNetwork - A search engine of all Socrata powered data portals
- Statista.com - statistics and Studies
- Zenodo - An open dependable home for the long-tail of science

## Social Networks

- 72 hours #gamergate Twitter Scrape
- Ancestry.com Forum Dataset over 10 years
- Cheng-Caverlee-Lee September 2009 - January 2010 Twitter Scrape
- CMU Enron Email of 150 users
- EDRM Enron EMail of 151 users, hosted on S3
- Facebook Data Scrape (2005)
- Facebook Social Networks from LAW (since 2007)
- Foursquare from UMN/Sarwat (2013)
- GetGlue - users rating TV shows
- GitHub Collaboration Archive
- Google Scholar citation relations
- High-Resolution Contact Networks from Wearable Sensors
- Mobile Social Networks from UMASS
- Network Twitter Data
- Reddit Comments
- Skytrax' Air Travel Reviews Dataset
- Social Twitter Data
- SourceForge.net Research Data
- Twitter Data for Sentiment Analysis
- Twitter Data for Online Reputation Management
- Twitter Graph of entire Twitter site
- Twitter Scrape Calufa May 2011
- UNIMI/LAW Social Network Datasets
- Yahoo! Graph and Social Data
- Youtube Video Social Graph in 2007,2008

## Social Sciences

- ACLED (Armed Conflict Location & Event Data Project)
- Canadian Legal Information Institute
- Center for Systemic Peace Datasets - Conflict Trends, Polities, State Fragility, etc
- Correlates of War Project
- Cryptome Conspiracy Theory Items
- Datacards
- European Social Survey
- FBI Hate Crime 2013 - aggregated data
- GDELT Global Events Database
- General Social Survey (GSS) since 1972
- German Social Survey
- Global Religious Futures Project
- Humanitarian Data Exchange
- Institute for Demographic Studies
- International Networks Archive
- International Social Survey Program ISSP
- International Studies Compendium Project
- James McGuire Cross National Data
- MacroData Guide by Norsk samfunnsvitenskapelig datatjeneste
- MIT Reality Mining Dataset
- Open Crime and Policing Data in England, Wales and Northern Ireland

- Paul Hensel General International Data Page
- PewResearch Internet Survey Project
- PewResearch Society Data Collection
- Political Polarity Data
- StackExchange Data Explorer
- Terrorism Research and Analysis Consortium
- Texas Inmates Executed Since 1984
- Titanic Survival Data Set
- UCB's Archive of Social Science Data (D-Lab)
- UCLA Social Sciences Data Archive
- UN Civil Society Database
- Universities Worldwide
- UPJOHN for Labor Employment Research
- World Bank Data
- WorldPop project - Worldwide human population distributions

## Software

- FLOSSmole data about free, libre, and open source software development

## Sports

- Basketball (NBA/NCAA/Euro) Player Database and Statistics
- Betfair Historical Exchange Data
- Cricsheet Matches (cricket)
- Ergast Formula 1, from 1950 up to date (API)
- Football/Soccer resources (data and APIs)
- Lahman's Baseball Database
- Pinhooker: Thoroughbred Bloodstock Sale Data
- Retrosheet Baseball Statistics

## Time Series

- Databanks International Cross National Time Series Data Archive
- Hard Drive Failure Rates
- Heart Rate Time Series from MIT
- Time Series Data Library (TSDL) from MU
- UC Riverside Time Series Dataset

## Transportation

- Airlines OD Data 1987-2008
- Bay Area Bike Share Data
- Bike Share Systems (BSS) collection

- GeoLife GPS Trajectory from Microsoft Research
- German train system by Deutsche Bahn
- Hubway Million Rides in MA
- Marine Traffic - ship tracks, port calls and more
- Montreal BIXI Bike Share
- NYC Taxi Trip Data 2009-
- NYC Taxi Trip Data 2013 (FOIA/FOILed)
- NYC Uber trip data April 2014 to September 2014
- Open Traffic collection
- OpenFlights - airport, airline and route data
- Philadelphia Bike Share Stations (JSON)
- Plane Crash Database, since 1920
- RITA Airline On-Time Performance data
- RITA/BTS transport data collection (TranStat)
- Toronto Bike Share Stations (XML file)
- Transport for London (TFL)
- Travel Tracker Survey (TTS) for Chicago
- U.S. Bureau of Transportation Statistics (BTS)
- U.S. Domestic Flights 1990 to 2009
- U.S. Freight Analysis Framework since 2007

**Complementary Collections**

- Data Packaged Core Datasets
- Database of Scientific Code Contributions
- DataWrangling: Some Datasets Available on the Web
- Inside-r: Finding Data on the Internet
- OpenDataMonitor: An overview of available open data resources in Europe
- Quora: Where can I find large datasets open to the public?
- RS.io: 100+ Interesting Data Sets for Statistics
- StaTrek: Leveraging open data to understand urban lives

**Source: Neo4J**

https://neo4j.com/developer/example-data/

- The Panama Papers
- Northwind Database Import
- Importing Stack Overflow into Neo4j
- The Cosmic Web of Galaxies
- Chicago Crime Dataset
- How I met your Mother Series
- Awesome Public Datasets
- Consumer Complaint Data
- Football(Soccer) Worldcup, Data Model
- Flight & Airline, Music, Train Schedules
- Kaggle Publication Dataset
- GitHub Event Data

**Source: Vincent Arel-Bundock**

**https://vincentarelbundock.github.io/Rdatasets/datasets.html**

List

| PackageItemTitlecsvdoc | | | |
|---|---|---|---|
| datasets | AirPassengers | Monthly Airline Passenger Numbers 1949-1960 | CSV DOC |
| datasets | BJsales | Sales Data with Leading Indicator | CSV DOC |
| datasets | BOD | Biochemical Oxygen Demand | CSV DOC |
| datasets | CO2 | Carbon Dioxide Uptake in Grass Plants | CSV DOC |
| datasets | Formaldehyde | Determination of Formaldehyde | CSV DOC |
| datasets | HairEyeColor | Hair and Eye Color of Statistics Students | CSV DOC |
| datasets | InsectSprays | Effectiveness of Insect Sprays | CSV DOC |
| datasets | JohnsonJohnson | Quarterly Earnings per Johnson & Johnson Share | CSV DOC |
| datasets | LakeHuron | Level of Lake Huron 1875-1972 | CSV DOC |
| datasets | LifeCycleSavings | Intercountry Life-Cycle Savings Data | CSV DOC |
| datasets | Nile | Flow of the River Nile | CSV DOC |
| datasets | OrchardSprays | Potency of Orchard Sprays | CSV DOC |
| datasets | PlantGrowth | Results from an Experiment on Plant Growth | CSV DOC |
| datasets | Puromycin | Reaction Velocity of an Enzymatic Reaction | CSV DOC |
| datasets | Titanic | Survival of passengers on the Titanic | CSV DOC |
| datasets | ToothGrowth | The Effect of Vitamin C on Tooth Growth in Guinea Pigs | CSV DOC |
| datasets | UCBAdmissions | Student Admissions at UC Berkeley | CSV DOC |
| datasets | UKDriverDeaths | Road Casualties in Great Britain 1969-84 | CSV DOC |
| datasets | UKgas | UK Quarterly Gas Consumption | CSV DOC |
| datasets | USAccDeaths | Accidental Deaths in the US 1973-1978 | CSV DOC |
| datasets | USArrests | Violent Crime Rates by US State | CSV DOC |
| datasets | USJudgeRatings | Lawyers' Ratings of State Judges in the US Superior Court | CSV DOC |
| datasets | USPersonalExpenditure | Personal Expenditure Data | CSV DOC |
| datasets | VADeaths | Death Rates in Virginia (1940) | CSV DOC |
| datasets | WWWusage | Internet Usage per Minute | CSV DOC |
| datasets | WorldPhones | The World's Telephones | CSV DOC |
| datasets | airmiles | Passenger Miles on Commercial US Airlines, 1937-1960 | CSV DOC |
| datasets | airquality | New York Air Quality Measurements | CSV DOC |
| datasets | anscombe | Anscombe's Quartet of 'Identical' Simple Linear Regressions | CSV DOC |
| datasets | attenu | The Joyner-Boore Attenuation Data | CSV DOC |
| datasets | attitude | The Chatterjee-Price Attitude Data | CSV DOC |
| datasets | austres | Quarterly Time Series of the Number of Australian Residents | CSV DOC |
| datasets | cars | Speed and Stopping Distances of Cars | CSV DOC |
| datasets | chickwts | Chicken Weights by Feed Type | CSV DOC |
| datasets | co2 | Mauna Loa Atmospheric CO2 Concentration | CSV DOC |
| datasets | crimtab | Student's 3000 Criminals Data | CSV DOC |
| datasets | discoveries | Yearly Numbers of Important Discoveries | CSV DOC |
| datasets | esoph | Smoking, Alcohol and (O)esophageal Cancer | CSV DOC |

| | | | |
|---|---|---|---|
| datasets | euro | Conversion Rates of Euro Currencies | CSV DOC |
| datasets | faithful | Old Faithful Geyser Data | CSV DOC |
| datasets | freeny | Freeny's Revenue Data | CSV DOC |
| datasets | infert | Infertility after Spontaneous and Induced Abortion | CSV DOC |
| datasets | iris | Edgar Anderson's Iris Data | CSV DOC |
| datasets | islands | Areas of the World's Major Landmasses | CSV DOC |
| datasets | lh | Luteinizing Hormone in Blood Samples | CSV DOC |
| datasets | longley | Longley's Economic Regression Data | CSV DOC |
| datasets | lynx | Annual Canadian Lynx trappings 1821-1934 | CSV DOC |
| datasets | morley | Michelson Speed of Light Data | CSV DOC |
| datasets | mtcars | Motor Trend Car Road Tests | CSV DOC |
| datasets | nhtemp | Average Yearly Temperatures in New Haven | CSV DOC |
| datasets | nottem | Average Monthly Temperatures at Nottingham, 1920-1939 | CSV DOC |
| datasets | npk | Classical N, P, K Factorial Experiment | CSV DOC |
| datasets | occupationalStatus | Occupational Status of Fathers and their Sons | CSV DOC |
| datasets | precip | Annual Precipitation in US Cities | CSV DOC |
| datasets | presidents | Quarterly Approval Ratings of US Presidents | CSV DOC |
| datasets | pressure | Vapor Pressure of Mercury as a Function of Temperature | CSV DOC |
| datasets | quakes | Locations of Earthquakes off Fiji | CSV DOC |
| datasets | randu | Random Numbers from Congruential Generator RANDU | CSV DOC |
| datasets | rivers | Lengths of Major North American Rivers | CSV DOC |
| datasets | rock | Measurements on Petroleum Rock Samples | CSV DOC |
| datasets | sleep | Student's Sleep Data | CSV DOC |
| datasets | stackloss | Brownlee's Stack Loss Plant Data | CSV DOC |
| datasets | sunspot.month | Monthly Sunspot Data, from 1749 to "Present" | CSV DOC |
| datasets | sunspot.year | Yearly Sunspot Data, 1700-1988 | CSV DOC |
| datasets | sunspots | Monthly Sunspot Numbers, 1749-1983 | CSV DOC |
| datasets | swiss | Swiss Fertility and Socioeconomic Indicators (1888) Data | CSV DOC |
| datasets | treering | Yearly Treering Data, -6000-1979 | CSV DOC |
| datasets | trees | Girth, Height and Volume for Black Cherry Trees | CSV DOC |
| datasets | uspop | Populations Recorded by the US Census | CSV DOC |
| datasets | volcano | Topographic Information on Auckland's Maunga Whau Volcano | CSV DOC |
| datasets | warpbreaks | The Number of Breaks in Yarn during Weaving | CSV DOC |
| datasets | women | Average Heights and Weights for American Women | CSV DOC |
| boot | acme | Monthly Excess Returns | CSV DOC |
| boot | aids | Delay in AIDS Reporting in England and Wales | CSV DOC |
| boot | aircondit | Failures of Air-conditioning Equipment | CSV DOC |
| boot | aircondit7 | Failures of Air-conditioning Equipment | CSV DOC |
| boot | amis | Car Speeding and Warning Signs | CSV DOC |
| boot | aml | Remission Times for Acute Myelogenous Leukaemia | CSV DOC |
| boot | bigcity | Population of U.S. Cities | CSV DOC |
| boot | brambles | Spatial Location of Bramble Canes | CSV DOC |
| boot | breslow | Smoking Deaths Among Doctors | CSV DOC |
| boot | calcium | Calcium Uptake Data | CSV DOC |
| boot | cane | Sugar-cane Disease Data | CSV DOC |

| | | | |
|---|---|---|---|
| boot | capability | Simulated Manufacturing Process Data | CSV DOC |
| boot | catsM | Weight Data for Domestic Cats | CSV DOC |
| boot | cav | Position of Muscle Caveolae | CSV DOC |
| boot | cd4 | CD4 Counts for HIV-Positive Patients | CSV DOC |
| boot | channing | Channing House Data | CSV DOC |
| boot | city | Population of U.S. Cities | CSV DOC |
| boot | claridge | Genetic Links to Left-handedness | CSV DOC |
| boot | cloth | Number of Flaws in Cloth | CSV DOC |
| boot | co.transfer | Carbon Monoxide Transfer | CSV DOC |
| boot | coal | Dates of Coal Mining Disasters | CSV DOC |
| boot | darwin | Darwin's Plant Height Differences | CSV DOC |
| boot | dogs | Cardiac Data for Domestic Dogs | CSV DOC |
| boot | downs.bc | Incidence of Down's Syndrome in British Columbia | CSV DOC |
| boot | ducks | Behavioral and Plumage Characteristics of Hybrid Ducks | CSV DOC |
| boot | fir | Counts of Balsam-fir Seedlings | CSV DOC |
| boot | frets | Head Dimensions in Brothers | CSV DOC |
| boot | grav | Acceleration Due to Gravity | CSV DOC |
| boot | gravity | Acceleration Due to Gravity | CSV DOC |
| boot | hirose | Failure Time of PET Film | CSV DOC |
| boot | islay | Jura Quartzite Azimuths on Islay | CSV DOC |
| boot | manaus | Average Heights of the Rio Negro river at Manaus | CSV DOC |
| boot | melanoma | Survival from Malignant Melanoma | CSV DOC |
| boot | motor | Data from a Simulated Motorcycle Accident | CSV DOC |
| boot | neuro | Neurophysiological Point Process Data | CSV DOC |
| boot | nitrofen | Toxicity of Nitrofen in Aquatic Systems | CSV DOC |
| boot | nodal | Nodal Involvement in Prostate Cancer | CSV DOC |
| boot | nuclear | Nuclear Power Station Construction Data | CSV DOC |
| boot | paulsen | Neurotransmission in Guinea Pig Brains | CSV DOC |
| boot | poisons | Animal Survival Times | CSV DOC |
| boot | polar | Pole Positions of New Caledonian Laterites | CSV DOC |
| boot | remission | Cancer Remission and Cell Activity | CSV DOC |
| boot | salinity | Water Salinity and River Discharge | CSV DOC |
| boot | survival | Survival of Rats after Radiation Doses | CSV DOC |
| boot | tau | Tau Particle Decay Modes | CSV DOC |
| boot | tuna | Tuna Sighting Data | CSV DOC |
| boot | urine | Urine Analysis Data | CSV DOC |
| boot | wool | Australian Relative Wool Prices | CSV DOC |
| KMsurv | aids | data from Section 1.19 | CSV DOC |
| KMsurv | alloauto | data from Section 1.9 | CSV DOC |
| KMsurv | allograft | data from Exercise 13.1, p418 | CSV DOC |
| KMsurv | azt | data from Exercise 4.7, p122 | CSV DOC |
| KMsurv | baboon | data from Exercise 5.8, p147 | CSV DOC |
| KMsurv | bcdeter | data from Section 1.18 | CSV DOC |
| KMsurv | bfeed | data from Section 1.14 | CSV DOC |
| KMsurv | bmt | data from Section 1.3 | CSV DOC |
| KMsurv | bnct | data from Exercise 7.7, p223 | CSV DOC |

| | | | |
|---|---|---|---|
| KMsurv | btrial | data from Section 1.5 | CSV DOC |
| KMsurv | burn | data from Section 1.6 | CSV DOC |
| KMsurv | channing | data from Section 1.16 | CSV DOC |
| KMsurv | drug6mp | data from Section 1.2 | CSV DOC |
| KMsurv | drughiv | data from Exercise 7.6, p222 | CSV DOC |
| KMsurv | hodg | data from Section 1.10 | CSV DOC |
| KMsurv | kidney | data from Section 1.4 | CSV DOC |
| KMsurv | kidrecurr | Data on 38 individuals using a kidney dialysis machine | CSV DOC |
| KMsurv | kidtran | data from Section 1.7 | CSV DOC |
| KMsurv | larynx | data from Section 1.8 | CSV DOC |
| KMsurv | lung | data from Exercise 4.4, p120 | CSV DOC |
| KMsurv | pneumon | data from Section 1.13 | CSV DOC |
| KMsurv | psych | data from Section 1.15 | CSV DOC |
| KMsurv | rats | data from Exercise 7.13, p225 | CSV DOC |
| KMsurv | std | data from Section 1.12 | CSV DOC |
| KMsurv | stddiag | data from Exercise 5.6, p146 | CSV DOC |
| KMsurv | tongue | data from Section 1.11 | CSV DOC |
| KMsurv | twins | data from Exercise 7.14, p225 | CSV DOC |
| robustbase | Animals2 | Brain and Body Weights for 65 Species of Land Animals | CSV DOC |
| robustbase | CrohnD | Crohn's Disease Adverse Events Data | CSV DOC |
| robustbase | NOxEmissions | NOx Air Pollution Data | CSV DOC |
| robustbase | SiegelsEx | Siegel's Exact Fit Example Data | CSV DOC |
| robustbase | aircraft | Aircraft Data | CSV DOC |
| robustbase | airmay | Air Quality Data | CSV DOC |
| robustbase | alcohol | Alcohol Solubility in Water Data | CSV DOC |
| robustbase | ambientNOxCH | Daily Means of NOx (mono-nitrogen oxides) in air | CSV DOC |
| robustbase | biomassTill | Biomass Tillage Data | CSV DOC |
| robustbase | bushfire | Campbell Bushfire Data | CSV DOC |
| robustbase | carrots | Insect Damages on Carrots | CSV DOC |
| robustbase | cloud | Cloud point of a Liquid | CSV DOC |
| robustbase | coleman | Coleman Data Set | CSV DOC |
| robustbase | condroz | Condroz Data | CSV DOC |
| robustbase | cushny | Cushny and Peebles Prolongation of Sleep Data | CSV DOC |
| robustbase | delivery | Delivery Time Data | CSV DOC |
| robustbase | education | Education Expenditure Data | CSV DOC |
| robustbase | epilepsy | Epilepsy Attacks Data Set | CSV DOC |
| robustbase | exAM | Example Data of Antille and May - for Simple Regression | CSV DOC |
| robustbase | foodstamp | Food Stamp Program Participation | CSV DOC |
| robustbase | hbk | Hawkins, Bradu, Kass's Artificial Data | CSV DOC |
| robustbase | heart | Heart Catherization Data | CSV DOC |
| robustbase | kootenay | Waterflow Measurements of Kootenay River in Libby and Newgate | CSV DOC |
| robustbase | lactic | Lactic Acid Concentration Measurement Data | CSV DOC |
| robustbase | milk | Daudin's Milk Composition Data | CSV DOC |
| robustbase | pension | Pension Funds Data | CSV DOC |
| robustbase | phosphor | Phosphorus Content Data | CSV DOC |

| | | | |
|---|---|---|---|
| robustbase | pilot | Pilot-Plant Data | CSV DOC |
| robustbase | possumDiv | Possum Diversity Data | CSV DOC |
| robustbase | pulpfiber | Pulp Fiber and Paper Data | CSV DOC |
| robustbase | radarImage | Satellite Radar Image Data from near Munich | CSV DOC |
| robustbase | salinity | Salinity Data | CSV DOC |
| robustbase | starsCYG | Hertzsprung-Russell Diagram Data of Star Cluster CYG OB1 | CSV DOC |
| robustbase | telef | Number of International Calls from Belgium | CSV DOC |
| robustbase | toxicity | Toxicity of Carboxylic Acids Data | CSV DOC |
| robustbase | vaso | Vaso Constriction Skin Data Set | CSV DOC |
| robustbase | wagnerGrowth | Wagner's Hannover Employment Growth Data | CSV DOC |
| robustbase | wood | Modified Data on Wood Specific Gravity | CSV DOC |
| car | AMSsurvey | American Math Society Survey Data | CSV DOC |
| car | Adler | Experimenter Expectations | CSV DOC |
| car | Angell | Moral Integration of American Cities | CSV DOC |
| car | Anscombe | U. S. State Public-School Expenditures | CSV DOC |
| car | Baumann | Methods of Teaching Reading Comprehension | CSV DOC |
| car | Bfox | Canadian Women's Labour-Force Participation | CSV DOC |
| car | Blackmore | Exercise Histories of Eating-Disordered and Control Subjects | CSV DOC |
| car | Burt | Fraudulent Data on IQs of Twins Raised Apart | CSV DOC |
| car | CanPop | Canadian Population Data | CSV DOC |
| car | Chile | Voting Intentions in the 1988 Chilean Plebiscite | CSV DOC |
| car | Chirot | The 1907 Romanian Peasant Rebellion | CSV DOC |
| car | Cowles | Cowles and Davis's Data on Volunteering | CSV DOC |
| car | Davis | Self-Reports of Height and Weight | CSV DOC |
| car | DavisThin | Davis's Data on Drive for Thinness | CSV DOC |
| car | Depredations | Minnesota Wolf Depredation Data | CSV DOC |
| car | Duncan | Duncan's Occupational Prestige Data | CSV DOC |
| car | Ericksen | The 1980 U.S. Census Undercount | CSV DOC |
| car | Florida | Florida County Voting | CSV DOC |
| car | Freedman | Crowding and Crime in U. S. Metropolitan Areas | CSV DOC |
| car | Friendly | Format Effects on Recall | CSV DOC |
| car | Ginzberg | Data on Depression | CSV DOC |
| car | Greene | Refugee Appeals | CSV DOC |
| car | Guyer | Anonymity and Cooperation | CSV DOC |
| car | Hartnagel | Canadian Crime-Rates Time Series | CSV DOC |
| car | Highway1 | Highway Accidents | CSV DOC |
| car | KosteckiDillon | Treatment of Migraine Headaches | CSV DOC |
| car | Leinhardt | Data on Infant-Mortality | CSV DOC |
| car | LoBD | Cancer drug data use to provide an example of the use of the skew power distributions. | CSV DOC |
| car | Mandel | Contrived Collinear Data | CSV DOC |
| car | Migration | Canadian Interprovincial Migration Data | CSV DOC |
| car | Moore | Status, Authoritarianism, and Conformity | CSV DOC |
| car | Mroz | U.S. Women's Labor-Force Participation | CSV DOC |
| car | OBrienKaiser | O'Brien and Kaiser's Repeated-Measures Data | CSV DOC |
| car | Ornstein | Interlocking Directorates Among Major Canadian Firms | CSV DOC |

| | | | |
|---|---|---|---|
| car | Pottery | Chemical Composition of Pottery | CSV DOC |
| car | Prestige | Prestige of Canadian Occupations | CSV DOC |
| car | Quartet | Four Regression Datasets | CSV DOC |
| car | Robey | Fertility and Contraception | CSV DOC |
| car | SLID | Survey of Labour and Income Dynamics | CSV DOC |
| car | Sahlins | Agricultural Production in Mazulu Village | CSV DOC |
| car | Salaries | Salaries for Professors | CSV DOC |
| car | Soils | Soil Compositions of Physical and Chemical Characteristics | CSV DOC |
| car | States | Education and Related Statistics for the U.S. States | CSV DOC |
| car | Transact | Transaction data | CSV DOC |
| car | UN | GDP and Infant Mortality | CSV DOC |
| car | USPop | Population of the United States | CSV DOC |
| car | Vocab | Vocabulary and Education | CSV DOC |
| car | WeightLoss | Weight Loss Data | CSV DOC |
| car | Womenlf | Canadian Women's Labour-Force Participation | CSV DOC |
| car | Wong | Post-Coma Recovery of IQ | CSV DOC |
| car | Wool | Wool data | CSV DOC |
| cluster | agriculture | European Union Agricultural Workforces | CSV DOC |
| cluster | animals | Attributes of Animals | CSV DOC |
| cluster | chorSub | Subset of C-horizon of Kola Data | CSV DOC |
| cluster | flower | Flower Characteristics | CSV DOC |
| cluster | plantTraits | Plant Species Traits Data | CSV DOC |
| cluster | pluton | Isotopic Composition Plutonium Batches | CSV DOC |
| cluster | ruspini | Ruspini Data | CSV DOC |
| cluster | votes.repub | Votes for Republican Candidate in Presidential Elections | CSV DOC |
| cluster | xclara | Bivariate Data Set with 3 Clusters | CSV DOC |
| COUNT | affairs | affairs | CSV DOC |
| COUNT | azcabgptca | azcabgptca | CSV DOC |
| COUNT | azdrg112 | azdrg112 | CSV DOC |
| COUNT | azpro | azpro | CSV DOC |
| COUNT | azprocedure | azprocedure | CSV DOC |
| COUNT | badhealth | badhealth | CSV DOC |
| COUNT | fasttrakg | fasttrakg | CSV DOC |
| COUNT | fishing | fishing | CSV DOC |
| COUNT | lbw | lbw | CSV DOC |
| COUNT | lbwgrp | lbwgrp | CSV DOC |
| COUNT | loomis | loomis | CSV DOC |
| COUNT | mdvis | mdvis | CSV DOC |
| COUNT | medpar | medpar | CSV DOC |
| COUNT | nuts | nuts | CSV DOC |
| COUNT | rwm | rwm | CSV DOC |
| COUNT | rwm1984 | rwm1984 | CSV DOC |
| COUNT | rwm5yr | rwm5yr | CSV DOC |
| COUNT | ships | ships | CSV DOC |
| COUNT | smoking | smoking | CSV DOC |
| COUNT | titanic | titanic | CSV DOC |

| | | | |
|---|---|---|---|
| COUNT | titanicgrp | titanicgrp | CSV DOC |
| Ecdat | Accident | Ship Accidents | CSV DOC |
| Ecdat | Airline | Cost for U.S. Airlines | CSV DOC |
| Ecdat | Airq | Air Quality for Californian Metropolitan Areas | CSV DOC |
| Ecdat | Benefits | Unemployement of Blue Collar Workers | CSV DOC |
| Ecdat | Bids | Bids Received By U.S. Firms | CSV DOC |
| Ecdat | BudgetFood | Budget Share of Food for Spanish Households | CSV DOC |
| Ecdat | BudgetItaly | Budget Shares for Italian Households | CSV DOC |
| Ecdat | BudgetUK | Budget Shares of British Households | CSV DOC |
| Ecdat | Bwages | Wages in Belgium | CSV DOC |
| Ecdat | CPSch3 | Earnings from the Current Population Survey | CSV DOC |
| Ecdat | CRANpackages | Growth of CRAN | CSV DOC |
| Ecdat | Capm | Stock Market Data | CSV DOC |
| Ecdat | Car | Stated Preferences for Car Choice | CSV DOC |
| Ecdat | Caschool | The California Test Score Data Set | CSV DOC |
| Ecdat | Catsup | Choice of Brand for Catsup | CSV DOC |
| Ecdat | Cigar | Cigarette Consumption | CSV DOC |
| Ecdat | Cigarette | The Cigarette Consumption Panel Data Set | CSV DOC |
| Ecdat | Clothing | Sales Data of Men's Fashion Stores | CSV DOC |
| Ecdat | Computers | Prices of Personal Computers | CSV DOC |
| Ecdat | Cracker | Choice of Brand for Crakers | CSV DOC |
| Ecdat | Crime | Crime in North Carolina | CSV DOC |
| Ecdat | DM | DM Dollar Exchange Rate | CSV DOC |
| Ecdat | Diamond | Pricing the C's of Diamond Stones | CSV DOC |
| Ecdat | Doctor | Number of Doctor Visits | CSV DOC |
| Ecdat | DoctorAUS | Doctor Visits in Australia | CSV DOC |
| Ecdat | DoctorContacts | Contacts With Medical Doctor | CSV DOC |
| Ecdat | Earnings | Earnings for Three Age Groups | CSV DOC |
| Ecdat | Electricity | Cost Function for Electricity Producers | CSV DOC |
| Ecdat | Fair | Extramarital Affairs Data | CSV DOC |
| Ecdat | Fatality | Drunk Driving Laws and Traffic Deaths | CSV DOC |
| Ecdat | Fishing | Choice of Fishing Mode | CSV DOC |
| Ecdat | Forward | Exchange Rates of US Dollar Against Other Currencies | CSV DOC |
| Ecdat | FriendFoe | Data from the Television Game Show Friend Or Foe ? | CSV DOC |
| Ecdat | Garch | Daily Observations on Exchange Rates of the US Dollar Against Other Currencies | CSV DOC |
| Ecdat | Gasoline | Gasoline Consumption | CSV DOC |
| Ecdat | Griliches | Wage Datas | CSV DOC |
| Ecdat | Grunfeld | Grunfeld Investment Data | CSV DOC |
| Ecdat | HC | Heating and Cooling System Choice in Newly Built Houses in California | CSV DOC |
| Ecdat | HHSCyberSecurityBreaches | Cybersecurity breaches reported to the US Department of Health and Human Services | CSV DOC |
| Ecdat | HI | Health Insurance and Hours Worked By Wives | CSV DOC |
| Ecdat | Hdma | The Boston HDMA Data Set | CSV DOC |
| Ecdat | Heating | Heating System Choice in California Houses | CSV DOC |
| Ecdat | Hedonic | Hedonic Prices of Cencus Tracts in Boston | CSV DOC |

| | | | |
|---|---|---|---|
| Ecdat | Housing | Sales Prices of Houses in the City of Windsor | CSV DOC |
| Ecdat | Icecream | Ice Cream Consumption | CSV DOC |
| Ecdat | Journals | Economic Journals Dat Set | CSV DOC |
| Ecdat | Kakadu | Willingness to Pay for the Preservation of the Kakadu National Park | CSV DOC |
| Ecdat | Ketchup | Choice of Brand for Ketchup | CSV DOC |
| Ecdat | Klein | Klein's Model I | CSV DOC |
| Ecdat | LaborSupply | Wages and Hours Worked | CSV DOC |
| Ecdat | Labour | Belgian Firms | CSV DOC |
| Ecdat | MCAS | The Massashusets Test Score Data Set | CSV DOC |
| Ecdat | Males | Wages and Education of Young Males | CSV DOC |
| Ecdat | Mathlevel | Level of Calculus Attained for Students Taking Advanced Micro-economics | CSV DOC |
| Ecdat | MedExp | Structure of Demand for Medical Care | CSV DOC |
| Ecdat | Metal | Production for SIC 33 | CSV DOC |
| Ecdat | Mode | Mode Choice | CSV DOC |
| Ecdat | ModeChoice | Data to Study Travel Mode Choice | CSV DOC |
| Ecdat | Mofa | International Expansion of U.S. Mofa's (majority-owned Foreign Affiliates in Fire (finance, Insurance and Real Estate) | CSV DOC |
| Ecdat | Mroz | Labor Supply Data | CSV DOC |
| Ecdat | MunExp | Municipal Expenditure Data | CSV DOC |
| Ecdat | NaturalPark | Willingness to Pay for the Preservation of the Alentejo Natural Park | CSV DOC |
| Ecdat | Nerlove | Cost Function for Electricity Producers, 1955 | CSV DOC |
| Ecdat | OFP | Visits to Physician Office | CSV DOC |
| Ecdat | Oil | Oil Investment | CSV DOC |
| Ecdat | PSID | Panel Survey of Income Dynamics | CSV DOC |
| Ecdat | Participation | Labor Force Participation | CSV DOC |
| Ecdat | PatentsHGH | Dynamic Relation Between Patents and R&D | CSV DOC |
| Ecdat | PatentsRD | Patents, R&D and Technological Spillovers for a Panel of Firms | CSV DOC |
| Ecdat | Pound | Pound-dollar Exchange Rate | CSV DOC |
| Ecdat | Produc | Us States Production | CSV DOC |
| Ecdat | RetSchool | Return to Schooling | CSV DOC |
| Ecdat | SP500 | Returns on Standard & Poor's 500 Index | CSV DOC |
| Ecdat | Schooling | Wages and Schooling | CSV DOC |
| Ecdat | Somerville | Visits to Lake Somerville | CSV DOC |
| Ecdat | Star | Effects on Learning of Small Class Sizes | CSV DOC |
| Ecdat | Strike | Strike Duration Data | CSV DOC |
| Ecdat | StrikeDur | Strikes Duration | CSV DOC |
| Ecdat | StrikeNb | Number of Strikes in Us Manufacturing | CSV DOC |
| Ecdat | SumHes | The Penn Table | CSV DOC |
| Ecdat | Tobacco | Households Tobacco Budget Share | CSV DOC |
| Ecdat | Train | Stated Preferences for Train Traveling | CSV DOC |
| Ecdat | TranspEq | Statewide Data on Transportation Equipment Manufacturing | CSV DOC |
| Ecdat | Treatment | Evaluating Treatment Effect of Training on Earnings | CSV DOC |
| Ecdat | Tuna | Choice of Brand for Tuna | CSV DOC |

| | | | |
|---|---|---|---|
| Ecdat | USFinanceIndustry | US Finance Industry Profits | CSV DOC |
| Ecdat | USclassifiedDocuments | Official Secrecy of the United States Government | CSV DOC |
| Ecdat | USstateAbbreviations | Standard abbreviations for states of the United States | CSV DOC |
| Ecdat | UStaxWords | Number of Words in US Tax Law | CSV DOC |
| Ecdat | UnempDur | Unemployment Duration | CSV DOC |
| Ecdat | Unemployment | Unemployment Duration | CSV DOC |
| Ecdat | University | Provision of University Teaching and Research | CSV DOC |
| Ecdat | VietNamH | Medical Expenses in Viet-nam (household Level) | CSV DOC |
| Ecdat | VietNamI | Medical Expenses in Viet-nam (individual Level) | CSV DOC |
| Ecdat | Wages | Panel Datas of Individual Wages | CSV DOC |
| Ecdat | Wages1 | Wages, Experience and Schooling | CSV DOC |
| Ecdat | Workinghours | Wife Working Hours | CSV DOC |
| Ecdat | Yen | Yen-dollar Exchange Rate | CSV DOC |
| Ecdat | Yogurt | Choice of Brand for Yogurts | CSV DOC |
| Ecdat | bankingCrises | Countries in Banking Crises | CSV DOC |
| Ecdat | breaches | Cyber Security Breaches | CSV DOC |
| Ecdat | incomeInequality | Income Inequality in the US | CSV DOC |
| Ecdat | nonEnglishNames | Names with Character Set Problems | CSV DOC |
| Ecdat | politicalKnowledge | Political knowledge in the US and Europe | CSV DOC |
| gap | PD | A study of Parkinson's disease and APOE, LRRK2, SNCA makers | CSV DOC |
| gap | aldh2 | ALDH2 markers and Alcoholism | CSV DOC |
| gap | apoeapoc | APOE/APOC1 markers and Alzheimer's | CSV DOC |
| gap | cf | Cystic fibrosis data | CSV DOC |
| gap | crohn | Crohn's disease data | CSV DOC |
| gap | fa | Friedreich Ataxia data | CSV DOC |
| gap | fsnps | A case-control data involving four SNPs with missing genotype | CSV DOC |
| gap | hla | The HLA data | CSV DOC |
| gap | hr1420 | An example data for Manhattan plot with annotation | CSV DOC |
| gap | l51 | An example pedigree data | CSV DOC |
| gap | lukas | An example pedigree | CSV DOC |
| gap | mao | A study of Parkinson's disease and MAO gene | CSV DOC |
| gap | meyer | A pedigree data on 282 animals deriving from two generations | CSV DOC |
| gap | mfblong | Example data for ACEnucfam | CSV DOC |
| gap | mhtdata | An example data for Manhattan plot | CSV DOC |
| gap | nep499 | A study of Alzheimer's disease with eight SNPs and APOE | CSV DOC |
| ggplot2 | luv_colours | 'colors()' in Luv space. | CSV DOC |
| HistData | Arbuthnot | Arbuthnot's data on male and female birth ratios in London from 1629-1710. | CSV DOC |
| HistData | Armada | La Felicisima Armada | CSV DOC |
| HistData | Bowley | Bowley's data on values of British and Irish trade, 1855-1899 | CSV DOC |
| HistData | Cavendish | Cavendish's Determinations of the Density of the Earth | CSV DOC |
| HistData | ChestSizes | Chest measurements of 5738 Scottish Militiamen | CSV DOC |
| HistData | CushnyPeebles | Cushny-Peebles Data: Soporific Effects of Scopolamine Derivatives | CSV DOC |

| | | | |
|---|---|---|---|
| HistData | CushnyPeeblesN | Cushny-Peebles Data: Soporific Effects of Scopolamine Derivatives | CSV DOC |
| HistData | Dactyl | Edgeworth's counts of dactyls in Virgil's Aeneid | CSV DOC |
| HistData | DrinksWages | Elderton and Pearson's (1910) data on drinking and wages | CSV DOC |
| HistData | Fingerprints | Waite's data on Patterns in Fingerprints | CSV DOC |
| HistData | Galton | Galton's data on the heights of parents and their children | CSV DOC |
| HistData | GaltonFamilies | Galton's data on the heights of parents and their children, by child | CSV DOC |
| HistData | Guerry | Data from A.-M. Guerry, "Essay on the Moral Statistics of France" | CSV DOC |
| HistData | Jevons | W. Stanley Jevons' data on numerical discrimination | CSV DOC |
| HistData | Langren.all | van Langren's Data on Longitude Distance between Toledo and Rome | CSV DOC |
| HistData | Langren1644 | van Langren's Data on Longitude Distance between Toledo and Rome | CSV DOC |
| HistData | Macdonell | Macdonell's Data on Height and Finger Length of Criminals, used by Gosset (1908) | CSV DOC |
| HistData | MacdonellDF | Macdonell's Data on Height and Finger Length of Criminals, used by Gosset (1908) | CSV DOC |
| HistData | Michelson | Michelson's Determinations of the Velocity of Light | CSV DOC |
| HistData | MichelsonSets | Michelson's Determinations of the Velocity of Light | CSV DOC |
| HistData | Minard.cities | Data from Minard's famous graphic map of Napoleon's march on Moscow | CSV DOC |
| HistData | Minard.temp | Data from Minard's famous graphic map of Napoleon's march on Moscow | CSV DOC |
| HistData | Minard.troops | Data from Minard's famous graphic map of Napoleon's march on Moscow | CSV DOC |
| HistData | Nightingale | Florence Nightingale's data on deaths from various causes in the Crimean War | CSV DOC |
| HistData | OldMaps | Latitudes and Longitudes of 39 Points in 11 Old Maps | CSV DOC |
| HistData | PearsonLee | Pearson and Lee's data on the heights of parents and children classified by gender | CSV DOC |
| HistData | PolioTrials | Polio Field Trials Data | CSV DOC |
| HistData | Prostitutes | Parent-Duchatelet's time-series data on the number of prostitutes in Paris | CSV DOC |
| HistData | Pyx | Trial of the Pyx | CSV DOC |
| HistData | Quarrels | Statistics of Deadly Quarrels | CSV DOC |
| HistData | Snow.deaths | John Snow's map and data on the 1854 London Cholera outbreak | CSV DOC |
| HistData | Snow.deaths2 | John Snow's map and data on the 1854 London Cholera outbreak | CSV DOC |
| HistData | Snow.polygons | John Snow's map and data on the 1854 London Cholera outbreak | CSV DOC |
| HistData | Snow.pumps | John Snow's map and data on the 1854 London Cholera outbreak | CSV DOC |
| HistData | Snow.streets | John Snow's map and data on the 1854 London Cholera outbreak | CSV DOC |
| HistData | Wheat | Playfair's Data on Wages and the Price of Wheat | CSV DOC |
| HistData | Wheat.monarchs | Playfair's Data on Wages and the Price of Wheat | CSV DOC |
| HistData | Yeast | Student's (1906) Yeast Cell Counts | CSV DOC |

| | | | |
|---|---|---|---|
| HistData | YeastD.mat | Student's (1906) Yeast Cell Counts | CSV DOC |
| HistData | ZeaMays | Darwin's Heights of Cross- and Self-fertilized Zea May Pairs | CSV DOC |
| lattice | barley | Yield data from a Minnesota barley trial | CSV DOC |
| lattice | environmental | Atmospheric environmental conditions in New York City | CSV DOC |
| lattice | ethanol | Engine exhaust fumes from burning ethanol | CSV DOC |
| lattice | melanoma | Melanoma skin cancer incidence | CSV DOC |
| lattice | singer | Heights of New York Choral Society singers | CSV DOC |
| MASS | Aids2 | Australian AIDS Survival Data | CSV DOC |
| MASS | Animals | Brain and Body Weights for 28 Species | CSV DOC |
| MASS | Boston | Housing Values in Suburbs of Boston | CSV DOC |
| MASS | Cars93 | Data from 93 Cars on Sale in the USA in 1993 | CSV DOC |
| MASS | Cushings | Diagnostic Tests on Patients with Cushing's Syndrome | CSV DOC |
| MASS | DDT | DDT in Kale | CSV DOC |
| MASS | GAGurine | Level of GAG in Urine of Children | CSV DOC |
| MASS | Insurance | Numbers of Car Insurance claims | CSV DOC |
| MASS | Melanoma | Survival from Malignant Melanoma | CSV DOC |
| MASS | OME | Tests of Auditory Perception in Children with OME | CSV DOC |
| MASS | Pima.te | Diabetes in Pima Indian Women | CSV DOC |
| MASS | Pima.tr | Diabetes in Pima Indian Women | CSV DOC |
| MASS | Pima.tr2 | Diabetes in Pima Indian Women | CSV DOC |
| MASS | Rabbit | Blood Pressure in Rabbits | CSV DOC |
| MASS | Rubber | Accelerated Testing of Tyre Rubber | CSV DOC |
| MASS | SP500 | Returns of the Standard and Poors 500 | CSV DOC |
| MASS | Sitka | Growth Curves for Sitka Spruce Trees in 1988 | CSV DOC |
| MASS | Sitka89 | Growth Curves for Sitka Spruce Trees in 1989 | CSV DOC |
| MASS | Skye | AFM Compositions of Aphyric Skye Lavas | CSV DOC |
| MASS | Traffic | Effect of Swedish Speed Limits on Accidents | CSV DOC |
| MASS | UScereal | Nutritional and Marketing Information on US Cereals | CSV DOC |
| MASS | UScrime | The Effect of Punishment Regimes on Crime Rates | CSV DOC |
| MASS | VA | Veteran's Administration Lung Cancer Trial | CSV DOC |
| MASS | abbey | Determinations of Nickel Content | CSV DOC |
| MASS | accdeaths | Accidental Deaths in the US 1973-1978 | CSV DOC |
| MASS | anorexia | Anorexia Data on Weight Change | CSV DOC |
| MASS | bacteria | Presence of Bacteria after Drug Treatments | CSV DOC |
| MASS | beav1 | Body Temperature Series of Beaver 1 | CSV DOC |
| MASS | beav2 | Body Temperature Series of Beaver 2 | CSV DOC |
| MASS | biopsy | Biopsy Data on Breast Cancer Patients | CSV DOC |
| MASS | birthwt | Risk Factors Associated with Low Infant Birth Weight | CSV DOC |
| MASS | cabbages | Data from a cabbage field trial | CSV DOC |
| MASS | caith | Colours of Eyes and Hair of People in Caithness | CSV DOC |
| MASS | cats | Anatomical Data from Domestic Cats | CSV DOC |
| MASS | cement | Heat Evolved by Setting Cements | CSV DOC |
| MASS | chem | Copper in Wholemeal Flour | CSV DOC |
| MASS | coop | Co-operative Trial in Analytical Chemistry | CSV DOC |
| MASS | cpus | Performance of Computer CPUs | CSV DOC |
| MASS | crabs | Morphological Measurements on Leptograpsus Crabs | CSV DOC |

| | | | |
|---|---|---|---|
| MASS | deaths | Monthly Deaths from Lung Diseases in the UK | CSV DOC |
| MASS | drivers | Deaths of Car Drivers in Great Britain 1969-84 | CSV DOC |
| MASS | eagles | Foraging Ecology of Bald Eagles | CSV DOC |
| MASS | epil | Seizure Counts for Epileptics | CSV DOC |
| MASS | farms | Ecological Factors in Farm Management | CSV DOC |
| MASS | fgl | Measurements of Forensic Glass Fragments | CSV DOC |
| MASS | forbes | Forbes' Data on Boiling Points in the Alps | CSV DOC |
| MASS | galaxies | Velocities for 82 Galaxies | CSV DOC |
| MASS | gehan | Remission Times of Leukaemia Patients | CSV DOC |
| MASS | genotype | Rat Genotype Data | CSV DOC |
| MASS | geyser | Old Faithful Geyser Data | CSV DOC |
| MASS | gilgais | Line Transect of Soil in Gilgai Territory | CSV DOC |
| MASS | hills | Record Times in Scottish Hill Races | CSV DOC |
| MASS | housing | Frequency Table from a Copenhagen Housing Conditions Survey | CSV DOC |
| MASS | immer | Yields from a Barley Field Trial | CSV DOC |
| MASS | leuk | Survival Times and White Blood Counts for Leukaemia Patients | CSV DOC |
| MASS | mammals | Brain and Body Weights for 62 Species of Land Mammals | CSV DOC |
| MASS | mcycle | Data from a Simulated Motorcycle Accident | CSV DOC |
| MASS | menarche | Age of Menarche in Warsaw | CSV DOC |
| MASS | michelson | Michelson's Speed of Light Data | CSV DOC |
| MASS | minn38 | Minnesota High School Graduates of 1938 | CSV DOC |
| MASS | motors | Accelerated Life Testing of Motorettes | CSV DOC |
| MASS | muscle | Effect of Calcium Chloride on Muscle Contraction in Rat Hearts | CSV DOC |
| MASS | newcomb | Newcomb's Measurements of the Passage Time of Light | CSV DOC |
| MASS | nlschools | Eighth-Grade Pupils in the Netherlands | CSV DOC |
| MASS | npk | Classical N, P, K Factorial Experiment | CSV DOC |
| MASS | npr1 | US Naval Petroleum Reserve No. 1 data | CSV DOC |
| MASS | oats | Data from an Oats Field Trial | CSV DOC |
| MASS | painters | The Painter's Data of de Piles | CSV DOC |
| MASS | petrol | N. L. Prater's Petrol Refinery Data | CSV DOC |
| MASS | quine | Absenteeism from School in Rural New South Wales | CSV DOC |
| MASS | road | Road Accident Deaths in US States | CSV DOC |
| MASS | rotifer | Numbers of Rotifers by Fluid Density | CSV DOC |
| MASS | ships | Ships Damage Data | CSV DOC |
| MASS | shrimp | Percentage of Shrimp in Shrimp Cocktail | CSV DOC |
| MASS | shuttle | Space Shuttle Autolander Problem | CSV DOC |
| MASS | snails | Snail Mortality Data | CSV DOC |
| MASS | steam | The Saturated Steam Pressure Data | CSV DOC |
| MASS | stormer | The Stormer Viscometer Data | CSV DOC |
| MASS | survey | Student Survey Data | CSV DOC |
| MASS | synth.te | Synthetic Classification Problem | CSV DOC |
| MASS | synth.tr | Synthetic Classification Problem | CSV DOC |
| MASS | topo | Spatial Topographic Data | CSV DOC |
| MASS | waders | Counts of Waders at 15 Sites in South Africa | CSV DOC |

| | | | | |
|---|---|---|---|---|
| MASS | whiteside | House Insulation: Whiteside's Data | CSV | DOC |
| MASS | wtloss | Weight Loss Data from an Obese Patient | CSV | DOC |
| plm | Cigar | Cigarette Consumption | CSV | DOC |
| plm | Crime | Crime in North Carolina | CSV | DOC |
| plm | EmplUK | Employment and Wages in the United Kingdom | CSV | DOC |
| plm | Gasoline | Gasoline Consumption | CSV | DOC |
| plm | Grunfeld | Grunfeld's Investment Data | CSV | DOC |
| plm | Hedonic | Hedonic Prices of Census Tracts in the Boston Area | CSV | DOC |
| plm | LaborSupply | Wages and Hours Worked | CSV | DOC |
| plm | Males | Wages and Education of Young Males | CSV | DOC |
| plm | Parity | Purchasing Power Parity and other parity relationships | CSV | DOC |
| plm | Produc | US States Production | CSV | DOC |
| plm | RiceFarms | Production of Rice in India | CSV | DOC |
| plm | Snmesp | Employment and Wages in Spain | CSV | DOC |
| plm | SumHes | The Penn World Table, v. 5 | CSV | DOC |
| plm | Wages | Panel Data of Individual Wages | CSV | DOC |
| plyr | baseball | Yearly batting records for all major league baseball players | CSV | DOC |
| pscl | AustralianElectionPolling | Political opinion polls in Australia, 2004-07 | CSV | DOC |
| pscl | AustralianElections | elections to Australian House of Representatives, 1949-2007 | CSV | DOC |
| pscl | EfronMorris | Batting Averages for 18 major league baseball players, 1970 | CSV | DOC |
| pscl | RockTheVote | Voter turnout experiment, using Rock The Vote ads | CSV | DOC |
| pscl | UKHouseOfCommons | 1992 United Kingdom electoral returns | CSV | DOC |
| pscl | absentee | Absentee and Machine Ballots in Pennsylvania State Senate Races | CSV | DOC |
| pscl | admit | Applications to a Political Science PhD Program | CSV | DOC |
| pscl | bioChemists | article production by graduate students in biochemistry Ph.D. programs | CSV | DOC |
| pscl | ca2006 | California Congressional Districts in 2006 | CSV | DOC |
| pscl | iraqVote | U.S. Senate vote on the use of force against Iraq, 2002. | CSV | DOC |
| pscl | politicalInformation | Interviewer ratings of respondent levels of political information | CSV | DOC |
| pscl | presidentialElections | elections for U.S. President, 1932-2012, by state | CSV | DOC |
| pscl | prussian | Prussian army horse kick data | CSV | DOC |
| pscl | unionDensity | cross national rates of trade union density | CSV | DOC |
| pscl | vote92 | Reports of voting in the 1992 U.S. Presidential election. | CSV | DOC |
| reshape2 | french_fries | Sensory data from a french fries experiment. | CSV | DOC |
| reshape2 | smiths | Demo data describing the Smiths. | CSV | DOC |
| reshape2 | tips | Tipping data | CSV | DOC |
| rpart | car.test.frame | Automobile Data from 'Consumer Reports' 1990 | CSV | DOC |
| rpart | car90 | Automobile Data from 'Consumer Reports' 1990 | CSV | DOC |
| rpart | cu.summary | Automobile Data from 'Consumer Reports' 1990 | CSV | DOC |
| rpart | kyphosis | Data on Children who have had Corrective Spinal Surgery | CSV | DOC |
| rpart | solder | Soldering of Components on Printed-Circuit Boards | CSV | DOC |
| rpart | stagec | Stage C Prostate Cancer | CSV | DOC |
| sandwich | PublicSchools | US Expenditures for Public Schools | CSV | DOC |
| sem | Bollen | Bollen's Data on Industrialization and Political Democracy | CSV | DOC |
| sem | CNES | Variables from the 1997 Canadian National Election Study | CSV | DOC |
| sem | Klein | Klein's Data on the U. S. Economy | CSV | DOC |

| | | | |
|---|---|---|---|
| sem | Kmenta | Partly Artificial Data on the U. S. Economy | CSV DOC |
| sem | Tests | Six Mental Tests | CSV DOC |
| survival | bladder | Bladder Cancer Recurrences | CSV DOC |
| survival | cancer | NCCTG Lung Cancer Data | CSV DOC |
| survival | cgd | Chronic Granulotomous Disease data | CSV DOC |
| survival | colon | Chemotherapy for Stage B/C colon cancer | CSV DOC |
| survival | flchain | Assay of serum free light chain for 7874 subjects. | CSV DOC |
| survival | genfan | Generator fans | CSV DOC |
| survival | heart | Stanford Heart Transplant data | CSV DOC |
| survival | kidney | Kidney catheter data | CSV DOC |
| survival | leukemia | Acute Myelogenous Leukemia survival data | CSV DOC |
| survival | logan | Data from the 1972-78 GSS data used by Logan | CSV DOC |
| survival | lung | NCCTG Lung Cancer Data | CSV DOC |
| survival | mgus | Monoclonal gammapothy data | CSV DOC |
| survival | mgus2 | Monoclonal gammapothy data | CSV DOC |
| survival | nwtco | Data from the National Wilm's Tumor Study | CSV DOC |
| survival | ovarian | Ovarian Cancer Survival Data | CSV DOC |
| survival | pbc | Mayo Clinic Primary Biliary Cirrhosis Data | CSV DOC |
| survival | rats | Rat treatment data from Mantel et al | CSV DOC |
| survival | retinopathy | Diabetic Retinopathy | CSV DOC |
| survival | stanford2 | More Stanford Heart Transplant data | CSV DOC |
| survival | tobin | Tobin's Tobit data | CSV DOC |
| survival | transplant | Liver transplant waiting list | CSV DOC |
| survival | veteran | Veterans' Administration Lung Cancer study | CSV DOC |
| vcd | Arthritis | Arthritis Treatment Data | CSV DOC |
| vcd | Baseball | Baseball Data | CSV DOC |
| vcd | BrokenMarriage | Broken Marriage Data | CSV DOC |
| vcd | Bundesliga | Ergebnisse der Fussball-Bundesliga | CSV DOC |
| vcd | Bundestag2005 | Votes in German Bundestag Election 2005 | CSV DOC |
| vcd | Butterfly | Butterfly Species in Malaya | CSV DOC |
| vcd | CoalMiners | Breathlessness and Wheeze in Coal Miners | CSV DOC |
| vcd | DanishWelfare | Danish Welfare Study Data | CSV DOC |
| vcd | Employment | Employment Status | CSV DOC |
| vcd | Federalist | 'May' in Federalist Papers | CSV DOC |
| vcd | Hitters | Hitters Data | CSV DOC |
| vcd | HorseKicks | Death by Horse Kicks | CSV DOC |
| vcd | Hospital | Hospital data | CSV DOC |
| vcd | JobSatisfaction | Job Satisfaction Data | CSV DOC |
| vcd | JointSports | Opinions About Joint Sports | CSV DOC |
| vcd | Lifeboats | Lifeboats on the Titanic | CSV DOC |
| vcd | NonResponse | Non-Response Survey Data | CSV DOC |
| vcd | OvaryCancer | Ovary Cancer Data | CSV DOC |
| vcd | PreSex | Pre-marital Sex and Divorce | CSV DOC |
| vcd | Punishment | Corporal Punishment Data | CSV DOC |
| vcd | RepVict | Repeat Victimization Data | CSV DOC |
| vcd | Saxony | Families in Saxony | CSV DOC |

| | | | |
|---|---|---|---|
| vcd | SexualFun | Sex is Fun | CSV DOC |
| vcd | SpaceShuttle | Space Shuttle O-ring Failures | CSV DOC |
| vcd | Suicide | Suicide Rates in Germany | CSV DOC |
| vcd | Trucks | Truck Accidents Data | CSV DOC |
| vcd | UKSoccer | UK Soccer Scores | CSV DOC |
| vcd | VisualAcuity | Visual Acuity in Left and Right Eyes | CSV DOC |
| vcd | VonBort | Von Bortkiewicz Horse Kicks Data | CSV DOC |
| vcd | WeldonDice | Weldon's Dice Data | CSV DOC |
| vcd | WomenQueue | Women in Queues | CSV DOC |
| Zelig | MatchIt.url | Table of links for Zelig | CSV DOC |
| Zelig | PErisk | Political Economic Risk Data from 62 Countries in 1987 | CSV DOC |
| Zelig | SupremeCourt | U.S. Supreme Court Vote Matrix | CSV DOC |
| Zelig | Weimar | 1932 Weimar election data | CSV DOC |
| Zelig | Zelig.url | Table of links for Zelig | CSV DOC |
| Zelig | approval | U.S. Presidential Approval Data | CSV DOC |
| Zelig | bivariate | Sample data for bivariate probit regression | CSV DOC |
| Zelig | coalition | Coalition Dissolution in Parliamentary Democracies | CSV DOC |
| Zelig | coalition2 | Coalition Dissolution in Parliamentary Democracies, Modified Version | CSV DOC |
| Zelig | eidat | Simulation Data for Ecological Inference | CSV DOC |
| Zelig | free1 | Freedom of Speech Data | CSV DOC |
| Zelig | free2 | Freedom of Speech Data | CSV DOC |
| Zelig | friendship | Simulated Example of Schoolchildren Friendship Network | CSV DOC |
| Zelig | grunfeld | Simulation Data for model Seemingly Unrelated Regression (sur) that corresponds to method SUR of systemfit | CSV DOC |
| Zelig | hoff | Social Security Expenditure Data | CSV DOC |
| Zelig | homerun | Sample Data on Home Runs Hit By Mark McGwire and Sammy Sosa in 1998. | CSV DOC |
| Zelig | immi1 | Individual Preferences Over Immigration Policy | CSV DOC |
| Zelig | immi2 | Individual Preferences Over Immigration Policy | CSV DOC |
| Zelig | immi3 | Individual Preferences Over Immigration Policy | CSV DOC |
| Zelig | immi4 | Individual Preferences Over Immigration Policy | CSV DOC |
| Zelig | immi5 | Individual Preferences Over Immigration Policy | CSV DOC |
| Zelig | immigration | Individual Preferences Over Immigration Policy | CSV DOC |
| Zelig | klein | Simulation Data for model Two-Stage Least Square (twosls) that corresponds to method 2SLS of systemfit | CSV DOC |
| Zelig | kmenta | Simulation Data for model Three-Stage Least Square (threesls) that corresponds to method 3SLS of systemfit | CSV DOC |
| Zelig | macro | Macroeconomic Data | CSV DOC |
| Zelig | mexico | Voting Data from the 1988 Mexican Presidental Election | CSV DOC |
| Zelig | mid | Militarized Interstate Disputes | CSV DOC |
| Zelig | newpainters | The Discretized Painter's Data of de Piles | CSV DOC |
| Zelig | sanction | Multilateral Economic Sanctions | CSV DOC |
| Zelig | seatshare | Left Party Seat Share in 11 OECD Countries | CSV DOC |
| Zelig | sna.ex | Simulated Example of Social Network Data | CSV DOC |
| Zelig | swiss | Swiss Fertility and Socioeconomic Indicators (1888) Data | CSV DOC |
| Zelig | tobin | Tobin's Tobit Data | CSV DOC |

| | | | |
|---|---|---|---|
| Zelig | turnout | Turnout Data Set from the National Election Survey | CSV DOC |
| Zelig | voteincome | Sample Turnout and Demographic Data from the 2000 Current Population Survey | CSV DOC |
| HSAUR | BCG | BCG Vaccine Data | CSV DOC |
| HSAUR | BtheB | Beat the Blues Data | CSV DOC |
| HSAUR | CYGOB1 | CYG OB1 Star Cluster Data | CSV DOC |
| HSAUR | Forbes2000 | The Forbes 2000 Ranking of the World's Biggest Companies (Year 2004) | CSV DOC |
| HSAUR | GHQ | General Health Questionnaire | CSV DOC |
| HSAUR | Lanza | Prevention of Gastointestinal Damages | CSV DOC |
| HSAUR | agefat | Total Body Composision Data | CSV DOC |
| HSAUR | aspirin | Aspirin Data | CSV DOC |
| HSAUR | birthdeathrates | Birth and Death Rates Data | CSV DOC |
| HSAUR | bladdercancer | Bladder Cancer Data | CSV DOC |
| HSAUR | clouds | Cloud Seeding Data | CSV DOC |
| HSAUR | epilepsy | Epilepsy Data | CSV DOC |
| HSAUR | foster | Foster Feeding Experiment | CSV DOC |
| HSAUR | heptathlon | Olympic Heptathlon Seoul 1988 | CSV DOC |
| HSAUR | mastectomy | Survival Times after Mastectomy of Breast Cancer Patients | CSV DOC |
| HSAUR | meteo | Meteorological Measurements for 11 Years | CSV DOC |
| HSAUR | orallesions | Oral Lesions in Rural India | CSV DOC |
| HSAUR | phosphate | Phosphate Level Data | CSV DOC |
| HSAUR | pistonrings | Piston Rings Failures | CSV DOC |
| HSAUR | planets | Exoplanets Data | CSV DOC |
| HSAUR | plasma | Blood Screening Data | CSV DOC |
| HSAUR | polyps | Familial Andenomatous Polyposis | CSV DOC |
| HSAUR | polyps3 | Familial Andenomatous Polyposis | CSV DOC |
| HSAUR | pottery | Romano-British Pottery Data | CSV DOC |
| HSAUR | rearrests | Rearrests of Juvenile Felons | CSV DOC |
| HSAUR | respiratory | Respiratory Illness Data | CSV DOC |
| HSAUR | roomwidth | Students Estimates of Lecture Room Width | CSV DOC |
| HSAUR | schizophrenia | Age of Onset of Schizophrenia Data | CSV DOC |
| HSAUR | schizophrenia2 | Schizophrenia Data | CSV DOC |
| HSAUR | schooldays | Days not Spent at School | CSV DOC |
| HSAUR | skulls | Egyptian Skulls | CSV DOC |
| HSAUR | smoking | Nicotine Gum and Smoking Cessation | CSV DOC |
| HSAUR | students | Student Risk Taking | CSV DOC |
| HSAUR | suicides | Crowd Baiting Behaviour and Suicides | CSV DOC |
| HSAUR | toothpaste | Toothpaste Data | CSV DOC |
| HSAUR | voting | House of Representatives Voting Data | CSV DOC |
| HSAUR | water | Mortality and Water Hardness | CSV DOC |
| HSAUR | watervoles | Water Voles Data | CSV DOC |
| HSAUR | waves | Electricity from Wave Power at Sea | CSV DOC |
| HSAUR | weightgain | Gain in Weight of Rats | CSV DOC |
| HSAUR | womensrole | Womens Role in Society | CSV DOC |
| psych | Bechtoldt | Seven data sets showing a bifactor solution. | CSV DOC |
| psych | Bechtoldt.1 | Seven data sets showing a bifactor solution. | CSV DOC |

| | | | |
|---|---|---|---|
| psych | Bechtoldt.2 | Seven data sets showing a bifactor solution. | CSV DOC |
| psych | Dwyer | 8 cognitive variables used by Dwyer for an example. | CSV DOC |
| psych | Gleser | Example data from Gleser, Cronbach and Rajaratnam (1965) to show basic principles of generalizability theory. | CSV DOC |
| psych | Gorsuch | Example data set from Gorsuch (1997) for an example factor extension. | CSV DOC |
| psych | Harman.5 | 5 socio-economic variables from Harman (1967) | CSV DOC |
| psych | Harman.8 | Correlations of eight physical variables (from Harman, 1966) | CSV DOC |
| psych | Harman.political | Eight political variables used by Harman (1967) as example 8.17 | CSV DOC |
| psych | Holzinger | Seven data sets showing a bifactor solution. | CSV DOC |
| psych | Holzinger.9 | Seven data sets showing a bifactor solution. | CSV DOC |
| psych | Reise | Seven data sets showing a bifactor solution. | CSV DOC |
| psych | Schmid | 12 variables created by Schmid and Leiman to show the Schmid-Leiman Transformation | CSV DOC |
| psych | Thurstone | Seven data sets showing a bifactor solution. | CSV DOC |
| psych | Thurstone.33 | Seven data sets showing a bifactor solution. | CSV DOC |
| psych | Tucker | 9 Cognitive variables discussed by Tucker and Lewis (1973) | CSV DOC |
| psych | ability | 16 ability items scored as correct or incorrect. | CSV DOC |
| psych | affect | Two data sets of affect and arousal scores as a function of personality and movie conditions | CSV DOC |
| psych | bfi | 25 Personality items representing 5 factors | CSV DOC |
| psych | bfi.dictionary | 25 Personality items representing 5 factors | CSV DOC |
| psych | blot | Bond's Logical Operations Test - BLOT | CSV DOC |
| psych | burt | 11 emotional variables from Burt (1915) | CSV DOC |
| psych | cities | Distances between 11 US cities | CSV DOC |
| psych | cubits | Galton's example of the relationship between height and 'cubit' or forearm length | CSV DOC |
| psych | cushny | A data set from Cushny and Peebles (1905) on the effect of three drugs on hours of sleep, used by Student (1908) | CSV DOC |
| psych | epi | Eysenck Personality Inventory (EPI) data for 3570 participants | CSV DOC |
| psych | epi.bfi | 13 personality scales from the Eysenck Personality Inventory and Big 5 inventory | CSV DOC |
| psych | epi.dictionary | Eysenck Personality Inventory (EPI) data for 3570 participants | CSV DOC |
| psych | galton | Galton's Mid parent child height data | CSV DOC |
| psych | heights | A data.frame of the Galton (1888) height and cubit data set. | CSV DOC |
| psych | income | US family income from US census 2008 | CSV DOC |
| psych | iqitems | 16 multiple choice IQ items | CSV DOC |
| psych | msq | 75 mood items from the Motivational State Questionnaire for 3896 participants | CSV DOC |
| psych | neo | NEO correlation matrix from the NEO_PI_R manual | CSV DOC |
| psych | peas | Galton's Peas | CSV DOC |
| psych | sat.act | 3 Measures of ability: SATV, SATQ, ACT | CSV DOC |
| psych | withinBetween | An example of the distinction between within group and between group correlations | CSV DOC |
| quantreg | Bosco | Boscovich Data | CSV DOC |
| quantreg | CobarOre | Cobar Ore data | CSV DOC |
| quantreg | Mammals | Garland(1983) Data on Running Speed of Mammals | CSV DOC |

| | | | | |
|---|---|---|---|---|
| quantreg | barro | Barro Data | CSV | DOC |
| quantreg | engel | Engel Data | CSV | DOC |
| quantreg | gasprice | Time Series of US Gasoline Prices | CSV | DOC |
| quantreg | uis | UIS Drug Treatment study data | CSV | DOC |
| geepack | dietox | Growth curves of pigs in a 3x3 factorial experiment | CSV | DOC |
| geepack | koch | Ordinal Data from Koch | CSV | DOC |
| geepack | ohio | Ohio Children Wheeze Status | CSV | DOC |
| geepack | respdis | Clustered Ordinal Respiratory Disorder | CSV | DOC |
| geepack | respiratory | Data from a clinical trial comparing two treatments for a respiratory illness | CSV | DOC |
| geepack | seizure | Epiliptic Seizures | CSV | DOC |
| geepack | sitka89 | Growth of Sitka Spruce Trees | CSV | DOC |
| geepack | spruce | Log-size of 79 Sitka spruce trees | CSV | DOC |
| texmex | liver | Liver related laboratory data | CSV | DOC |
| texmex | portpirie | Rain, wavesurge and portpirie datasets. | CSV | DOC |
| texmex | rain | Rain, wavesurge and portpirie datasets. | CSV | DOC |
| texmex | summer | Air pollution data, separately for summer and winter months | CSV | DOC |
| texmex | wavesurge | Rain, wavesurge and portpirie datasets. | CSV | DOC |
| texmex | winter | Air pollution data, separately for summer and winter months | CSV | DOC |
| multgee | arthritis | Rheumatoid Arthritis Clinical Trial | CSV | DOC |
| multgee | housing | Homeless Data | CSV | DOC |
| evir | bmw | Daily Log Returns on BMW Share Price | CSV | DOC |
| evir | danish | Danish Fire Insurance Claims | CSV | DOC |
| evir | nidd.annual | The River Nidd Data | CSV | DOC |
| evir | nidd.thresh | The River Nidd Data | CSV | DOC |
| evir | siemens | Daily Log Returns on Siemens Share Price | CSV | DOC |
| evir | sp.raw | SP Data to June 1993 | CSV | DOC |
| evir | spto87 | SP Return Data to October 1987 | CSV | DOC |
| lme4 | Arabidopsis | Arabidopsis clipping/fertilization data | CSV | DOC |
| lme4 | Dyestuff | Yield of dyestuff by batch | CSV | DOC |
| lme4 | Dyestuff2 | Yield of dyestuff by batch | CSV | DOC |
| lme4 | InstEval | University Lecture/Instructor Evaluations by Students at ETH | CSV | DOC |
| lme4 | Pastes | Paste strength by batch and cask | CSV | DOC |
| lme4 | Penicillin | Variation in penicillin testing | CSV | DOC |
| lme4 | VerbAgg | Verbal Aggression item responses | CSV | DOC |
| lme4 | cake | Breakage Angle of Chocolate Cakes | CSV | DOC |
| lme4 | cbpp | Contagious bovine pleuropneumonia | CSV | DOC |
| lme4 | grouseticks | Data on red grouse ticks from Elston et al. 2001 | CSV | DOC |
| lme4 | sleepstudy | Reaction times in a sleep deprivation study | CSV | DOC |
| mosaicData | Alcohol | Alcohol Consumption per Capita | CSV | DOC |
| mosaicData | Birthdays | US Births in 1969 - 1988 | CSV | DOC |
| mosaicData | Births | US Births | CSV | DOC |
| mosaicData | Births78 | US Births in 1978 | CSV | DOC |
| mosaicData | CPS85 | Data from the 1985 Current Population Survey (CPS85) | CSV | DOC |
| mosaicData | CoolingWater | CoolingWater | CSV | DOC |
| mosaicData | Countries | Countries | CSV | DOC |

| | | | |
|---|---|---|---|
| mosaicData | Dimes | Weight of dimes | CSV DOC |
| mosaicData | Galton | Galton's dataset of parent and child heights | CSV DOC |
| mosaicData | Gestation | Data from the Child Health and Development Studies | CSV DOC |
| mosaicData | GoosePermits | Goose Permit Study | CSV DOC |
| mosaicData | HELPfull | Health Evaluation and Linkage to Primary Care | CSV DOC |
| mosaicData | HELPmiss | Health Evaluation and Linkage to Primary Care | CSV DOC |
| mosaicData | HELPrct | Health Evaluation and Linkage to Primary Care | CSV DOC |
| mosaicData | HeatX | Data from a heat exchanger laboratory | CSV DOC |
| mosaicData | KidsFeet | Foot measurements in children | CSV DOC |
| mosaicData | Marriage | Marriage records | CSV DOC |
| mosaicData | Mites | Mites and Wilt Disease | CSV DOC |
| mosaicData | RailTrail | Volume of Users of a Rail Trail | CSV DOC |
| mosaicData | Riders | Volume of Users of a Massachusetts Rail Trail | CSV DOC |
| mosaicData | SAT | State by State SAT data | CSV DOC |
| mosaicData | SaratogaHouses | Houses in Saratoga County (2006) | CSV DOC |
| mosaicData | SnowGR | Snowfall data for Grand Rapids, MI | CSV DOC |
| mosaicData | SwimRecords | 100 m Swimming World Records | CSV DOC |
| mosaicData | TenMileRace | Cherry Blossom Race | CSV DOC |
| mosaicData | Utilities | Utility bills | CSV DOC |
| mosaicData | Utilities2 | Utility bills | CSV DOC |
| mosaicData | Whickham | Data from the Whickham survey | CSV DOC |
| ISLR | Auto | Auto Data Set | CSV DOC |
| ISLR | Caravan | The Insurance Company (TIC) Benchmark | CSV DOC |
| ISLR | Carseats | Sales of Child Car Seats | CSV DOC |
| ISLR | College | U.S. News and World Report's College Data | CSV DOC |
| ISLR | Default | Credit Card Default Data | CSV DOC |
| ISLR | Hitters | Baseball Data | CSV DOC |
| ISLR | OJ | Orange Juice Data | CSV DOC |
| ISLR | Portfolio | Portfolio Data | CSV DOC |
| ISLR | Smarket | S&P Stock Market Data | CSV DOC |
| ISLR | Wage | Mid-Atlantic Wage Data | CSV DOC |
| ISLR | Weekly | Weekly S&P Stock Market Data | CSV DOC |
| Stat2Data | Alfalfa | Alfalfa | CSV DOC |
| Stat2Data | ArcheryData | ArcheryData | CSV DOC |
| Stat2Data | AutoPollution | AutoPollution | CSV DOC |
| Stat2Data | Backpack | Backpack | CSV DOC |
| Stat2Data | BaseballTimes | BaseballTimes | CSV DOC |
| Stat2Data | BeeStings | BeeStings | CSV DOC |
| Stat2Data | BirdNest | BirdNest | CSV DOC |
| Stat2Data | Blood1 | Blood1 | CSV DOC |
| Stat2Data | BlueJays | Blue Jays | CSV DOC |
| Stat2Data | BritishUnions | BritishUnions | CSV DOC |
| Stat2Data | CAFE | CAFE | CSV DOC |
| Stat2Data | CO2 | CO2 | CSV DOC |
| Stat2Data | CalciumBP | CalciumBP | CSV DOC |
| Stat2Data | CancerSurvival | CancerSurvival | CSV DOC |

|  |  |  |  |  |
|---|---|---|---|---|
| Stat2Data | Caterpillars | Caterpillars | CSV | DOC |
| Stat2Data | Cereal | Cereal | CSV | DOC |
| Stat2Data | ChemoTHC | ChemoTHC | CSV | DOC |
| Stat2Data | ChildSpeaks | ChildSpeaks | CSV | DOC |
| Stat2Data | Clothing | Clothing | CSV | DOC |
| Stat2Data | CloudSeeding | Cloud Seeding | CSV | DOC |
| Stat2Data | CloudSeeding2 | Cloud Seeding 2 | CSV | DOC |
| Stat2Data | CrackerFiber | Cracker Fiber in Diets | CSV | DOC |
| Stat2Data | Cuckoo | Cuckoo | CSV | DOC |
| Stat2Data | Day1Survey | Day1Survey | CSV | DOC |
| Stat2Data | Diamonds | Diamonds | CSV | DOC |
| Stat2Data | Diamonds2 | Diamonds2 | CSV | DOC |
| Stat2Data | Election08 | Election08 | CSV | DOC |
| Stat2Data | Ethanol | Ethanol | CSV | DOC |
| Stat2Data | FGByDistance | FGByDistance | CSV | DOC |
| Stat2Data | FantasyBaseball | FantasyBaseball | CSV | DOC |
| Stat2Data | Fertility | Fertility | CSV | DOC |
| Stat2Data | Film | Film | CSV | DOC |
| Stat2Data | FinalFourIzzo | FinalFourIzzo | CSV | DOC |
| Stat2Data | FinalFourLong | FinalFourLong | CSV | DOC |
| Stat2Data | FinalFourShort | FinalFourShort | CSV | DOC |
| Stat2Data | Fingers | Fingers | CSV | DOC |
| Stat2Data | FirstYearGPA | FirstYearGPA | CSV | DOC |
| Stat2Data | FishEggs | FishEggs | CSV | DOC |
| Stat2Data | FlightResponse | FlightResponse | CSV | DOC |
| Stat2Data | Fluorescence | Fluorescence | CSV | DOC |
| Stat2Data | FruitFlies | FruitFlies | CSV | DOC |
| Stat2Data | Goldenrod | Goldenrod Galls | CSV | DOC |
| Stat2Data | Grocery | Grocery | CSV | DOC |
| Stat2Data | Gunnels | Gunnels | CSV | DOC |
| Stat2Data | HawkTail | HawkTail | CSV | DOC |
| Stat2Data | HawkTail2 | HawkTail2 | CSV | DOC |
| Stat2Data | Hawks | Hawks | CSV | DOC |
| Stat2Data | HearingTest | HearingTest | CSV | DOC |
| Stat2Data | HighPeaks | HighPeaks | CSV | DOC |
| Stat2Data | Hoops | Hoops | CSV | DOC |
| Stat2Data | HorsePrices | HorsePrices | CSV | DOC |
| Stat2Data | Houses | Houses | CSV | DOC |
| Stat2Data | ICU | ICU | CSV | DOC |
| Stat2Data | InfantMortality | InfantMortality | CSV | DOC |
| Stat2Data | InsuranceVote | InsuranceVote | CSV | DOC |
| Stat2Data | Jurors | Jurors | CSV | DOC |
| Stat2Data | Kids198 | Kids198 | CSV | DOC |
| Stat2Data | LeafHoppers | LeafHoppers | CSV | DOC |
| Stat2Data | Leukemia | Leukemia | CSV | DOC |
| Stat2Data | LongJumpOlympics | LongJumpOlympics | CSV | DOC |

| | | | |
|---|---|---|---|
| Stat2Data | LostLetter | LostLetter | CSV DOC |
| Stat2Data | MLB2007Standings | MLB2007Standings | CSV DOC |
| Stat2Data | Marathon | Marathon | CSV DOC |
| Stat2Data | Markets | Markets | CSV DOC |
| Stat2Data | MathEnrollment | Math Enrollments | CSV DOC |
| Stat2Data | MathPlacement | Math Placement | CSV DOC |
| Stat2Data | MedGPA | MedGPA | CSV DOC |
| Stat2Data | MentalHealth | Mental Health Admissions | CSV DOC |
| Stat2Data | MetabolicRate | Metabolic Rate of Caterpillars | CSV DOC |
| Stat2Data | MetroHealth83 | MetroHealth83 | CSV DOC |
| Stat2Data | Milgram | Milgram | CSV DOC |
| Stat2Data | MothEggs | Moth Eggs | CSV DOC |
| Stat2Data | NCbirths | NCbirths | CSV DOC |
| Stat2Data | NFL2007Standings | NFL2007Standings | CSV DOC |
| Stat2Data | Nursing | Nursing | CSV DOC |
| Stat2Data | Olives | Olives | CSV DOC |
| Stat2Data | Orings | Orings | CSV DOC |
| Stat2Data | Overdrawn | Overdrawn | CSV DOC |
| Stat2Data | PalmBeach | PalmBeach | CSV DOC |
| Stat2Data | Pedometer | Pedometer | CSV DOC |
| Stat2Data | Perch | Perch | CSV DOC |
| Stat2Data | PigFeed | PigFeed | CSV DOC |
| Stat2Data | Pines | Pines | CSV DOC |
| Stat2Data | Political | Political | CSV DOC |
| Stat2Data | Pollster08 | Pollster08 | CSV DOC |
| Stat2Data | Popcorn | Popcorn | CSV DOC |
| Stat2Data | PorscheJaguar | PorscheJaguar | CSV DOC |
| Stat2Data | PorschePrice | PorschePrice | CSV DOC |
| Stat2Data | Pulse | Pulse | CSV DOC |
| Stat2Data | Putts1 | Putts1 | CSV DOC |
| Stat2Data | Putts2 | Putts2 | CSV DOC |
| Stat2Data | ReligionGDP | ReligionGDP | CSV DOC |
| Stat2Data | Retirement | Retirement | CSV DOC |
| Stat2Data | RiverElements | RiverElements | CSV DOC |
| Stat2Data | RiverIron | River Iron | CSV DOC |
| Stat2Data | SATGPA | SAT scores and GPA | CSV DOC |
| Stat2Data | SampleFG | SampleFG | CSV DOC |
| Stat2Data | SandwichAnts | Sandwich Ants | CSV DOC |
| Stat2Data | SeaSlugs | Sea Slugs | CSV DOC |
| Stat2Data | Sparrows | Sparrows | CSV DOC |
| Stat2Data | SpeciesArea | Species Area | CSV DOC |
| Stat2Data | Speed | Speed | CSV DOC |
| Stat2Data | Swahili | Swahili | CSV DOC |
| Stat2Data | TMS | TMS | CSV DOC |
| Stat2Data | TextPrices | Text Prices | CSV DOC |
| Stat2Data | ThreeCars | Three Cars | CSV DOC |

| | | | |
|---|---|---|---|
| Stat2Data | TipJoke | Tip Joke | CSV DOC |
| Stat2Data | Titanic | Titanic | CSV DOC |
| Stat2Data | TomlinsonRush | LaDainian Tomlinson Rushing Yards | CSV DOC |
| Stat2Data | TwinsLungs | TwinsLungs | CSV DOC |
| Stat2Data | USstamps | USstamps | CSV DOC |
| Stat2Data | Volts | Volts | CSV DOC |
| Stat2Data | WalkingBabies | WalkingBabies | CSV DOC |
| Stat2Data | WeightLossIncentive | WeightLossIncentive | CSV DOC |
| Stat2Data | WeightLossIncentive4 | WeightLossIncentive4 | CSV DOC |
| Stat2Data | WeightLossIncentive7 | WeightLossIncentive7 | CSV DOC |
| Stat2Data | WordMemory | WordMemory | CSV DOC |
| Stat2Data | YouthRisk2007 | YouthRisk2007 | CSV DOC |
| Stat2Data | YouthRisk2009 | YouthRisk2009 | CSV DOC |

**Source: r-dir (r-directory)**

https://r-dir.com/reference/datasets.html

List:

**World Bank Data** - Literally hundreds of datasets spanning many decades, sortable by topic or country. Data is downloadable in Excel or XML formats, or you can make API calls. This is an outstanding resource.

**Gapminder** - Hundreds of datasets on world health, economics, population, etc. All of it is viewable online within Google Docs, and downloadable as spreadsheets.

**The Data Hub** - Hosted by CKAN. Most of these datasets come from the government.

**Datamob** - List of public datasets.

**Numbrary** - Lists of datasets.

**Kaggle** - Kaggle is a site that hosts data mining competitions. Each competition provides a data set that's free for download.

**SNAP** - Stanford's Large Network Dataset Collection. This list has several datasets related to social networking. Lots of fun in here!

**KONECT** - The Koblenz Network Collection. Several datasets related to social networking & Wikipedia.

**Million Song Dataset** - This is a collection of audio features and metadata for a million contemporary popular music tracks.

**Energy Information Administration** - This site offers a number of datasets on energy production, consumption, sources, etc.

**GeoDa Center** - This is a collection of geospatial datasets offered by Arizona State Univeristy's Center for Geospatial Analysis & Computation.

**Reddit Datasets** - This last one isn't a dataset itself, but rather a social news site devoted to datasets. It's updated regularly with news about newly available datasets.

**Quandl** - This is a web-based front end to a number of public data sets. What's nice about this website is that it allows for the combination of data from a number of sources, and can export the data in a number of formats.

**1,001 Datasets** - This is a list of lists of datasets. There's not much organization here, but there really are a LOT of datasets. Dive in and have fun.

**Yahoo! Webscope** - A reference library of interesting and scientifically useful datasets for non-commercial use by academics and other scientists.

**Time Series Data Library** - Curated by Professor Rob Hyndman of Monash University in Australia, this is a collection of over 500 datasets containing time-series data, organized by category.

**Awesome Public Datasets** - Curated list of hundreds of public datasets, organized by topic.

**Common Crawl** - Massive dataset of billions of pages scraped from the web. The data itself is on Amazon Public Datasets, so its easy to load it into an EC2 instance there. The dataset is updated with a new scrape about once per month.

**SOURCE: Amazon Public Datasets** - Collection of datasets that are ready to be loaded into an EC2 instance.

**A Multi-wavelength Infrared Atlas of the Galactic Plane** Open Source tools were used to combine images from five major infrared surveys of the Galactic Plane, archived at the NASA/IPAC Infrared Science Archive (IRSA). The result is a 16-wavelength infrared Atlas of the Galactic Plane that coves the wavelength range 1 µm to 24 µm.

**CCAFS-Climate Data** High resolution climate data to help assess the impacts of climate change primarily on agriculture. These open access datasets of climate projections will help researchers make climate change impact assessments.

**NASA NEX** Three NASA NEX datasets are now available, including climate projections and satellite images of Earth.

**Human Microbiome Project** Human Microbiome Project Data Set

**Enron Email Data** Enron email data publicly released as part of FERC's Western Energy Markets investigation converted to industry standard formats by EDRM. The data set consists of 1,227,255 emails with 493,384 attachments covering 151 custodians. The email is provided in Microsoft PST, IETF MIME, and EDRM XML formats.

**Japan Census Data** Multiple data sets including: (1) Population Census of Japan (1995, 2000, 2005, 2010), (2) Establishment and Enterprise Census of Japan (1999, 2001, 2004, 2006), and (3) Economic Census of Japan (2009).

**Apache Software Foundation Public Mail Archives** A collection of all publicly available Apache Software Foundation mail archives as of July 11, 2011

**Freebase Simple Topic Dump** A data dump of the basic identifying facts about every topic in Freebase

**Freebase Quad Dump** A data dump of all the current facts and assertions in Freebase

**Wikipedia Page Traffic Statistic V3** This dataset contains a 150 GB sample of the data used to power trendingtopics.org. It includes a full 3 months of hourly page traffic statistics from Wikipedia (1/1/2011-3/31/2011).

**Material Safety Data Sheets** 230,000 Material Safety Data Sheets.

**Million Song Dataset** The Million Songs Collection is a collection of 28 datasets containing audio features and metadata for a million contemporary popular music tracks.

**Million Song Sample Dataset** This is a 10,000 song subset of audio features and metadata from the Million Songs collection - a collection of 28 datasets containing audio features and metadata for a million contemporary popular music tracks.

**Marvel Universe Social Graph** This dataset is an example of a social collaboration network based on the characters in The Marvel Universe, that is, the artificial world that takes place in the universe of the Marvel comic books.

**Google Books Ngrams** A data set containing Google Books n-gram corpora. This data set is freely available on Amazon S3 in a Hadoop friendly file format and is licensed under a Creative Commons Attribution 3.0 Unported License. The original dataset is available from http://books.google.com/ngrams/.

**The WestburyLab USENET corpus** The WestburyLab USENET corpus is an anonymized compilation of postings from 47,860 English-language newsgroups from 2005-2010.

**1000 Genomes Project** The 1000 Genomes Project, initiated in 2008, is an international public-private consortium that aims to build the most detailed map of human genetic variation available.

**Wikipedia Traffic Statistics V2** Contains 16 months of hourly pageview statistics for all articles in Wikipedia

**M-Lab dataset: Network Diagnostic Tool (NDT)** NDT test results created through Measurement Lab (M-Lab) between February 2009 and September 2009

**M-Lab dataset: Network Path and Application Diagnosis tool (NPAD)** NPAD test results created through Measurement Lab (M-Lab) between February 2009 and September 2009

**Petroleum Public Data Set (working Title)** Public-domain data for the oil & gas industry, assembled from the contributions of participating agencies in the United States, Canada and around the world. This data provides industry stakeholders with an opportunity to focus their efforts on the analysis and interpretation of this data without concern for the trivial and time-consuming tasks of locating, downloading, reformatting and integrating the data prior to value-added work being performed.

**Sloan Digital Sky Survey DR6 Subset** The Sloan Digital Sky Survey is the most ambitious astronomical survey ever undertaken.

**Wikipedia Page Traffic Statistics** Contains 7 months of hourly pageview statistics for all articles in Wikipedia

**Wikipedia XML Data** A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML.

**Federal Reserve Economic Data - Fred** Database of 20,059 U.S. economic time series.

**Twilio/Wigle.net Street Vector Data Set** Twilio/Wigle.net database of mapped US street names and address ranges.

**Federal Contracts from the Federal Procurement Data Center (USASpending.gov)** A data dump of all federal contracts from the Federal Procurement Data Center found at USASpending.gov.

**University of Florida Sparse Matrix Collection** The University of Florida Sparse Matrix Collection is a large, widely available, and actively growing set of sparse matrices that arise in real applications.

**2008 TIGER/Line Shapefiles** Census 2000 and Current United States shapefiles

**Wikipedia Extraction (WEX)** A processed dump of the English language Wikipedia

**Business and Industry Summary Data** US Business and Industry Summary Data

**2003-2006 US Economic Data** US Economic Data for years 2003 to 2006

**Freebase Data Dump** Freebase is an open database of the world's information, covering millions of topics in hundreds of categories

**DBpedia 3.5.1** DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web

**1980 US Census** Data from the 1980 US Census

**1990 US Census** Data from the 1990 US Census

**2000 US Census** Data from the 2000 US Census

**Transportation Databases** Various transportation statistics

**Labor Statistics Databases** Various Labor Statistics

**Source:**

List:

**Source:**

List:

**Source:**

List:

Enjoy! As mentioned above - 100% of this data is reposted - original source is in links - if I've missed any citations, please let me know and will fix

**3 Comments and 1 private comments**

**Partnerspartner help**

**Your partners**

Partners help you to create. They provide encouragement, and comment on what you're working on.

---

You don't currently have any partners, but it's easy to invite people to help you. Just click the button to get started.

(note that your viewers won't see the 'Partners' section until you have at least one partner)

3followers
13recommends

- Document statistics
- about this Document

**STATS SUMMARY**

**Nov 09 - Nov 12 Traffic**

- 160
- 128
- 96
- 64
- 32
- 0

121views
156views
144views

- Nov 11
- Nov 10
- Nov 09

All Time288,821PAGEVIEWSTotal
3followers
13recommends
Search site:
**What is Dream to Learn?**

Dream to Learn was founded in 2013. We are adding new features rapidly, and for those who share our vision, it's a great time to join and help to build Dream to Learn with us.

- About Us
- About
- Founders Blog
- Technologies
- System Status & Uptime

- Information

10.0.0.181: queries 39, query time 27 ms, raw queries 0, deferred counter queries: 2, ws host: ws2.dreamtolearn.com, page rendered in 82 ms
0