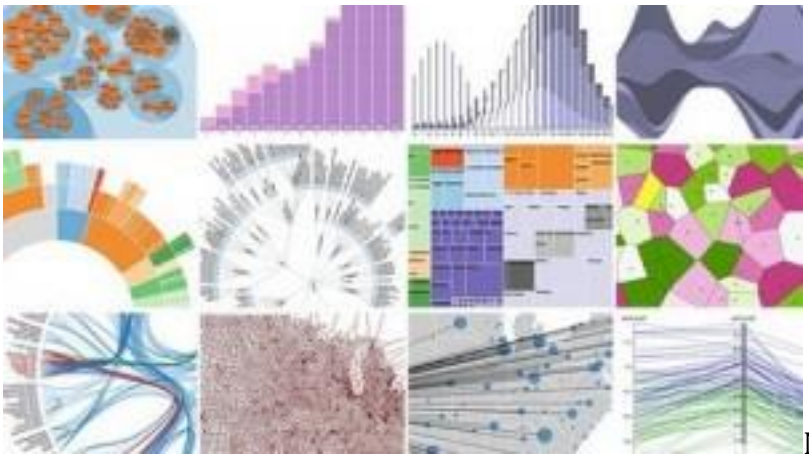
A long, categorized list of large datasets (available for public use) to try your analytics skills on. Which one would you pick?

By Anmol Rajpurohit, @hey_anmol



No matter how many books you read on technology, some knowledge comes only from experience. This is even truer in the field of Big Data. Despite a good number of resources available online (including KDnuggets dataset) for large datasets, many aspirants and practitioners (primarily, the newcomers) are rarely aware of the limitless options when it comes to trying their Data Science skills on real-life large datasets. Thus, we are consistently on the lookout for greater and better datasets available for public use.

In our next endeavor on this journey, we are sharing here an awesome list of public data sources by Xia Ming(bio given at the end) that are collected and organized from blogs, answers, and user responses. Most of the data sets listed below are free, however, some are not.

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
- Small Network Data
- UCI Network Data Repository
- 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.
- Criteo click-through data
- Open Mobile Data by MobiPerf
- UCSD Network Telescope, IPv4 /8 net
- Data Mining Software
- News
- Jobs
- Academic
- Companies
- Courses
- Datasets
- Data Mining Course
- Education
- Meetings
- Polls
- Webcasts

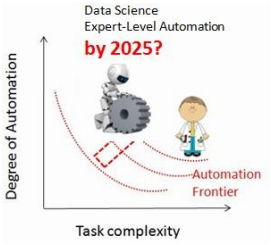
Latest News

- → Salford: 3 Ways to Improve your Targeted Marketing with Analytics, May 28
- → Upcoming Webcasts on Analytics, Big Data, Data Science - May 12 and beyond
- → Interview: Mark Weiner, Temple University Health System on Maturity Assessment of Healthcare Analytics
- → Gaming Analytics Summit 2015, San Francisco – Day 2 Highlights
- → 3 Things About Data Science You Won't Find In Books

New Release: BayesiaLab 5.4, Featuring the New WebSimulator

Top Stories Last week

Most Viewed Last Week



1. **Data Scientists Automated and Unemployed by 2025?**
2. **How To Become a Data Scientist And Get Hired**
3. **The Inconvenient Truth About Data Science**
4. **Poll: What Predictive Analytics, Data Mining, Data Science software/tools you used in the past 12 months?**
5. **Awesome Public Datasets on GitHub**
6. **The Grammar of Data Science: Python vs R**
7. **7 Steps for Learning Data Mining and Data Science**.

Most Shared Last Week

1. **The Inconvenient Truth About Data Science**
2. **Most Viewed Big Data Videos on YouTube**
3. **Data Scientists Automated and Unemployed by 2025?**
4. **Guiding Principles to Build a Demand**

Forecast
5. **Poll: What Predictive Analytics, Data Mining, Data Science software/tools you used in the past 12 months?**
6. **Strata + Hadoop World 2015, London, May 5-7, Watch Live**
7. **Best 5 minutes in Data Science, Season 1**

Awesome Public Datasets on GitHub

Share on facebook Share on linkedin
42
Previous post
Next post

Tags: Datasets, Finance, GitHub, Government, Machine Learning, NLP, Open Data, Time series data

---

A long, categorized list of large datasets (available for public use) to try your analytics skills on. Which one would you pick?

Pages: 1 2

Data Challenges

- Challenges in Machine Learning
- 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
- 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

You can also find various datasets for the following categories:

GeoSpace/GIS

Government

Healthcare

Image Processing

Machine Learning

Museums

Natural Language

Physics

Public Domains

Search Engines

Social Sciences

Sports

Time Series

Transportation

Complementary Collections

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

 Xia Ming is a Ph.D. candidate at Shanghai Jiao Tong Univ. He received B.S. in Optical Information and Science Technology in 2010 at Xidian University, Xi'an, China. His research area is the measurement and analysis of mobile network traffic, especially on the renewed models and characteristics of networks traffic, employing statistical and machine learning techniques on distributed processing platforms such as Apache Spark.