

A topic-centric list of HQ open datasets.

 [awesomedataworld.slack.com](#)

 MIT license

☆ 50k stars 🍴 8.8k forks

☆ Star

👁 Watch ▾

Code

Issues

56

Pull requests

61

Actions

Security

Insights

🔗 master ▾

...



caesar0301 ...

on 9 Jun 2021



[View code](#)

☰ README.rst

Awesome Public Datasets



NOTICE: This repo is automatically generated by [apd-core](#). Please **DO NOT** modify this file directly. We have provided [a new way](#) to contribute to Awesome Public Datasets. [Join](#) the [slack community](#) for more communication.

- 🟢 I am well.
- 🟡 Please fix me.





This list of [a topic-centric public data sources](#) in high quality. They are collected and tidied from blogs, answers, and user responses. Most of the data sets listed below are free, however, some are not. Other amazingly awesome lists can be found in [sindresorhus's awesome](#) list.



Table of Contents

- [Agriculture](#)
- [Biology](#)
- [Climate+Weather](#)
- [ComplexNetworks](#)

- [ComputerNetworks](#)
- [CyberSecurity](#)
- [DataChallenges](#)
- [EarthScience](#)
- [Economics](#)
- [Education](#)
- [Energy](#)
- [Entertainment](#)
- [Finance](#)
- [GIS](#)
- [Government](#)
- [Healthcare](#)
- [ImageProcessing](#)
- [MachineLearning](#)
- [Museums](#)
- [NaturalLanguage](#)
- [Neuroscience](#)
- [Physics](#)
- [ProstateCancer](#)
- [Psychology+Cognition](#)
- [PublicDomains](#)
- [SearchEngines](#)
- [SocialNetworks](#)
- [SocialSciences](#)
- [Software](#)
- [Sports](#)
- [TimeSeries](#)
- [Transportation](#)
- [eSports](#)
- [Complementary Collections](#)


















Agriculture

-  [The global dataset of historical yields for major crops 1981–2016 - The \[...\]](#)
-  [Hyperspectral benchmark dataset on soil moisture - This dataset was \[...\]](#)
-  [Lemons quality control dataset - Lemon dataset has been prepared to \[...\]](#)
-  [Optimized Soil Adjusted Vegetation Index - The IDB is a tool for working \[...\]](#)
















-  U.S. Department of Agriculture's Nutrient Database
-  U.S. Department of Agriculture's PLANTS Database - The Complete PLANTS [...] [fixme]






Biology

-  1000 Genomes - The 1000 Genomes Project ran between 2008 and 2015, [...]
-  American Gut (Microbiome Project) - The American Gut project is the [...]
-  Broad Bioimage Benchmark Collection (BBBC) - The Broad Bioimage Benchmark [...]
-  Broad Cancer Cell Line Encyclopedia (CCLE)
-  Cell Image Library - This library is a public and easily accessible [...]
-  Complete Genomics Public Data - A diverse data set of whole human genomes [...]
-  EBI ArrayExpress - ArrayExpress Archive of Functional Genomics Data [...]
-  EBI Protein Data Bank in Europe - The Electron Microscopy Data Bank [...]
-  ENCODE project - The Encyclopedia of DNA Elements (ENCODE) Consortium is [...]
-  Electron Microscopy Pilot Image Archive (EMPIAR) - EMPIAR, the Electron [...]
-  Ensembl Genomes
-  Gene Expression Omnibus (GEO) - GEO is a public functional genomics data [...]
-  Gene Ontology (GO) - GO annotation files
-  Global Biotic Interactions (GloBI)
-  Harvard Medical School (HMS) LINCS Project - The Harvard Medical School [...]
-  Human Genome Diversity Project - A group of scientists at Stanford [...]
-  Human Microbiome Project (HMP) - The HMP sequenced over 2000 reference [...]
-  ICOS PSP Benchmark - The ICOS PSP benchmarks repository contains an [...]
-  International HapMap Project
-  Journal of Cell Biology DataViewer [fixme]
-  KEGG - KEGG is a database resource for understanding high-level functions [...]
-  MIT Cancer Genomics Data
-  NCBI Proteins
-  NCBI Taxonomy - The NCBI Taxonomy database is a curated set of names and [...]
-  NCI Genomic Data Commons - The GDC Data Portal is a robust data-driven [...]
-  NIH Microarray data
-  OpenSNP genotypes data - openSNP allows customers of direct-to-customer [...]
-  Palmer Penguins - The goal of palmerpenguins is to provide a great [...]
-  Pathguid - Protein-Protein Interactions Catalog





















-  Protein Data Bank - This resource is powered by the Protein Data Bank [...]
-  Psychiatric Genomics Consortium - The purpose of the Psychiatric Genomics [...]
-  PubChem Project - PubChem is the world's largest collection of freely [...]
-  PubGene (now Coremine Medical) - COREMINE™ is a family of tools developed [...]
-  Sanger Catalogue of Somatic Mutations in Cancer (COSMIC) - COSMIC, the [...]
-  Sanger Genomics of Drug Sensitivity in Cancer Project (GDSC)
-  Sequence Read Archive(SRA) - The Sequence Read Archive (SRA) stores raw [...]
-  Stanford Microarray Data
-  Stowers Institute Original Data Repository
-  Systems Science of Biological Dynamics (SSBD) Database - Systems Science [...]
-  The Cancer Genome Atlas (TCGA), available via Broad GDAC
-  The Catalogue of Life - The Catalogue of Life is a quality-assured [...]
-  The Personal Genome Project - The Personal Genome Project, initiated in [...]
-  UCSC Public Data
-  UniGene
-  Universal Protein Resource (UnitProt) - The Universal Protein Resource [...]
-  Rfam - The Rfam database is a collection of RNA families, each [...]

Climate+Weather






-  Actuaries Climate Index
-  Australian Weather [fixme]
-  Aviation Weather Center - Consistent, timely and accurate weather [...]
-  Brazilian Weather - Historical data (In Portuguese) - Data related to [...]
-  Canadian Meteorological Centre
-  Climate Data from UEA (updated monthly)
-  Dutch Weather - The KNMI Data Center (KDC) portal provides access to KNMI [...]
-  European Climate Assessment & Dataset
-  German Climate Data Center
-  Global Climate Data Since 1929
-  Charting The Global Climate Change News Narrative 2009-2020 - These four [...]
-  NASA Global Imagery Browse Services
-  NOAA Bering Sea Climate [fixme]
-  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
-  WU Historical Weather Worldwide
-  Wahington Post Climate Change - To analyze warming temperatures in the [...]
-  WorldClim - Global Climate Data

ComplexNetworks



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

ComputerNetworks

















-  3.5B Web Pages from CommonCrawl 2012
-  53.5B Web clicks of 100K users in Indiana Univ.
-  CAIDA Internet Datasets
-  CRAWDAD Wireless datasets from Dartmouth Univ. [fixme]
-  ClueWeb09 - 1B web pages

-  ClueWeb12 - 733M web pages
-  CommonCrawl Web Data over 7 years
-  Criteo click-through data
-  Internet-Wide Scan Data Repository [fixme]
-  MIRAGE-2019 - MIRAGE-2019 is a human-generated dataset for mobile traffic [...] [fixme]
-  OONI: Open Observatory of Network Interference - Internet censorship data
-  Open Mobile Data by MobiPerf
-  The Peer-to-Peer Trace Archive - Real-world measurements play a key role [...]
-  Rapid7 Sonar Internet Scans
-  UCSD Network Telescope, IPv4 /8 net














CyberSecurity

-  CCCS-CIC-AndMal-2020 - The dataset includes 200K benign and 200K malware [...]
-  Traffic and Log Data Captured During a Cyber Defense Exercise - This [...]


















DataChallenges



-  Alcrowd Competitions
-  Bruteforce Database
-  Challenges in Machine Learning
-  CrowdANALYTIX dataX [fixme]
-  D4D Challenge of Orange [fixme]
-  DrivenData Competitions for Social Good
-  ICWSM Data Challenge (since 2009)
-  KDD Cup by Tencent 2012
-  Kaggle Competition Data
-  Localytics Data Visualization Challenge
-  Netflix Prize
-  Space Apps Challenge
-  Telecom Italia Big Data Challenge [fixme]
-  TravisTorrent Dataset - MSR'2017 Mining Challenge
-  TunedIT - Data mining & machine learning data sets, algorithms, challenges [fixme]
-  Yelp Dataset Challenge [fixme]

EarthScience





-  38-Cloud (Cloud Detection) - Contains 38 Landsat 8 scene images and their [...]
-  AQUASTAT - Global water resources and uses
-  BODC - marine data of ~22K vars
-  EOSDIS - NASA's earth observing system data
-  Earth Models [fixme]
-  Global Wind Atlas - The Global Wind Atlas is a free, web-based [...]
-  Integrated Marine Observing System (IMOS) - roughly 30TB of ocean measurements
-  Marinexplore - Open Oceanographic Data
-  Alabama Real-Time Coastal Observing System
-  National Estuarine Research Reserves System-Wide Monitoring Program - [...]
-  Oil and Gas Authority Open Data - The dataset covers 12,500 offshore [...]
-  Smithsonian Institution Global Volcano and Eruption Database
-  USGS Earthquake Archives

Economics



















-  American Economic Association (AEA)
-  EconData from UMD [fixme]
-  Economic Freedom of the World Data
-  Historical Macroeconomic Statistics
-  INFORUM - Interindustry Forecasting at the University of Maryland [fixme]
-  DBnomics – the world's economic database - Aggregates hundreds of [...]
-  International Trade Statistics [fixme]
-  Internet Product Code Database
-  Joint External Debt Data Hub
-  Jon Haveman International Trade Data Links
-  Long-Term Productivity Database - The Long-Term Productivity database was [...]
-  OpenCorporates Database of Companies in the World
-  Our World in Data
-  SciencesPo World Trade Gravity Datasets [fixme]
-  The Atlas of Economic Complexity
-  The Center for International Data
-  The Observatory of Economic Complexity [fixme]

-  UN Commodity Trade Statistics
-  UN Human Development Reports

Education

-  College Scorecard Data
-  New York State Education Department Data - The New York State Education [...]
-  Program for International Student Assessment (PISA) - Contains 15-year- [...]
-  Student Data from Free Code Camp















Energy

-  AMPds - The Almanac of Minutely Power dataset
-  BLUEd - Building-Level fully labeled Electricity Disaggregation dataset
-  COMBED
-  DBFC - Direct Borohydride Fuel Cell (DBFC) Dataset
-  DEL - Domestic Electrical Load study datasets for South Africa (1994 - 2014)
-  ECO - The ECO data set is a comprehensive data set for non-intrusive load [...]
-  EIA
-  Global Power Plant Database - The Global Power Plant Database is a [...]
-  HES - Household Electricity Study, UK
-  HFED
-  PEM1 - Proton Exchange Membrane (PEM) Fuel Cell Dataset
-  PLAID - The Plug Load Appliance Identification Dataset [fixme]
-  The Public Utility Data Liberation Project (PUDL) - PUDL makes US energy [...]
-  REDD
-  SYND - A synthetic energy dataset for non-intrusive load monitoring - [...]
-  Smart Meter Data Portal - The Smart Meter Data Portal is part of the [...]
-  Tracebase
-  Ukraine Energy Centre Datasets
-  UK-DALE - UK Domestic Appliance-Level Electricity
-  WHITED
-  iAWE


















Entertainment












-  Top Streamers on Twitch - This contains data of Top 1000 Streamers from [...]

Finance

-  BIS Statistics - BIS statistics, compiled in cooperation with central [...]
-  Blockmodo Coin Registry - A registry of JSON formatted information files [...]
-  CBOE Futures Exchange [fixme]
-  Complete FAANG Stock data - This data set contains all the stock data of [...]
-  Google Finance
-  Google Trends
-  NASDAQ [fixme]
-  NYSE Market Data
-  OANDA
-  OSU Financial data [fixme]
-  Quandl
-  SEC EDGAR - EDGAR, the Electronic Data Gathering, Analysis, and Retrieval [...]
-  St Louis Federal
-  Yahoo Finance





































GIS

-  Awesome 3D Semantic City Models - Collection of open 3D semantic city and [...]
-  ArcGIS Open Data portal
-  Cambridge, MA, US, GIS data on GitHub
-  Database of all continents, countries, States/Subdivisions/Provinces and [...]
-  Factual Global Location Data
-  IEEE Geoscience and Remote Sensing Society DASE Website
-  Geo Maps - High Quality GeoJSON maps programmatically generated
-  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) - Geospatial data organized [...]
-  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 [fixme]

-  [OpenAddresses](#)
-  [OpenStreetMap \(OSM\)](#)
-  [Pleiades - Gazetteer and graph of ancient places](#)
-  [Reverse Geocoder using OSM data](#)
-  [Robin Wilson - Free GIS Datasets](#)
-  [TIGER/Line - U.S. boundaries and roads](#)
-  [TZ Timezones shapefile](#)
-  [TwoFishes - Foursquare's coarse geocoder](#)
-  [UN Environmental Data](#)
-  [World boundaries from the U.S. Department of State](#)
-  [World countries in multiple formats](#)



Government

-  [Alberta, Province of Canada](#)
-  [Antwerp, Belgium](#)
-  [Argentina \(non official\) \[fixme\]](#)
-  [Datos Argentina - Portal de datos abiertos de la República Argentina. \[...\]](#)
-  [Austin, TX, US](#)
-  [Australia \(abs.gov.au\)](#)
-  [Australia \(data.gov.au\)](#)
-  [Austria \(data.gv.at\)](#)
-  [Baton Rouge, LA, US](#)
-  [Beersheba, Israel - Open Data Portal \(Smart7 OpenData\)](#)
-  [Belgium](#)
-  [City of Berkeley Open Data](#)
-  [Brazil](#)
-  [Buenos Aires, Argentina](#)
-  [Calgary, AB, Canada](#)
-  [Cambridge, MA, US](#)
-  [Canada](#)
-  [Chicago](#)
-  [Chile](#)
-  [China \[fixme\]](#)
-  [Dallas Open Data](#)
-  [DataBC - data from the Province of British Columbia](#)


-  Debt to the Penny - The Debt to the Penny dataset provides information [...]
-  Denver Open Data
-  Durham, NC Open Data
-  Edmonton, AB, Canada
-  England LGInform
-  EuroStat
-  EveryPolitician - Ongoing project collating and sharing data on every [...]
-  Federal Committee on Statistical Methodology (FCSM) (formerly FedStats)
-  Finland
-  France
-  Fredericton, NB, Canada
-  Gatineau, QC, Canada
-  Germany
-  Ghent, Belgium
-  Glasgow, Scotland, UK
-  Greece
-  Guardian world governments
-  Halifax, NS, Canada
-  Helsinki Region, Finland
-  Hong Kong, China
-  Houston, TX, US
-  Indian Government Data
-  Indonesian Data Portal
-  Iowa - Welcome to the State of Iowa's data portal. Please explore data [...]
-  Ireland's Open Data Portal
-  Israel's Open Data Portal [fixme]
-  Istanbul Municipality Open Data Portal
-  Italy - Il Portale dati.gov.it è il catalogo nazionale dei metadati [...]
-  Jail deaths in America - The U.S. government does not release jail by [...]
-  Japan
-  Laval, QC, Canada
-  Lexington, KY
-  London Datastore, UK
-  London, ON, Canada [fixme]
-  Los Angeles Open Data
-  Luxembourg - Luxembourgish Open Data Portal

-  MassGIS, Massachusetts, U.S.
-  Metropolitan Transportation Commission (MTC), California, US
-  Mexico
-  Mississauga, ON, Canada
-  Moldova
-  Moncton, NB, Canada
-  Montreal, QC, Canada
-  Mountain View, California, US (GIS)
-  NYC Open Data [fixme]
-  NYC betanyc
-  Netherlands
-  New York Department of Sanitation Monthly Tonnage - DSNY Monthly Tonnage [...]
-  New Zealand
-  OECD
-  Oakland, California, US [fixme]
-  Oklahoma
-  Open Data for Africa [fixme]
-  Open Government Data (OGD) Platform India
-  OpenDataSoft's list of 1,600 open data
-  Oregon
-  Ottawa, ON, Canada
-  Palo Alto, California, US
-  OpenDataPhilly - OpenDataPhilly is a catalog of open data in the [...]
-  Portland, Oregon
-  Portugal - Pordata organization
-  Puerto Rico Government [fixme]
-  Quebec City, QC, Canada [fixme]
-  Quebec Province of Canada
-  Regina SK, Canada
-  Rio de Janeiro, Brazil
-  Romania
-  Russia
-  San Diego, CA
-  San Antonio, TX - Community Information Now - CI:Now is a nonprofit [...] [fixme]
-  San Francisco Data sets
-  San Jose, California, US

-  San Mateo County, California, US
-  Saskatchewan, Province of Canada
-  Seattle
-  Singapore Government Data
-  South Africa Trade Statistics [fixme]
-  South Africa
-  State of Utah, US
-  Switzerland
-  Taiwan gov
-  Taiwan
-  Tel-Aviv Open Data
-  Texas Open Data
-  The World Bank
-  Toronto, ON, Canada [fixme]
-  Tunisia [fixme]
-  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. National Center for Education Statistics (NCES)
-  U.S. Open Government
-  UK 2011 Census Open Atlas Project
-  US Counties - This is a repository of various data, broken down by US [...]
-  U.S. Patent and Trademark Office (USPTO) Bulk Data Products
-  Uganda Bureau of Statistics [fixme]
-  Ukraine
-  United Nations
-  Uruguay
-  Valley Transportation Authority (VTA), California, US
-  Vancouver, BC Open Data Catalog [fixme]
-  Victoria, BC, Canada
-  Vienna, Austria


































-  Statistics from the General Statistics Office of Vietnam - Data in [...] [fixme]
-  U.S. Congressional Research Service (CRS) Reports





Healthcare

-  AWS COVID-19 Datasets - We're working with organizations who make [...]
-  COVID-19 Case Surveillance Public Use Data - The COVID-19 case [...]
-  2019 Novel Coronavirus COVID-19 Data Repository by Johns Hopkins CSSE - [...]
-  Coronavirus (Covid-19) Data in the United States - The New York Times is [...]
-  COVID-19 Reported Patient Impact and Hospital Capacity by Facility - The [...] [fixme]
-  Composition of Foods Raw, Processed, Prepared USDA National Nutrient Database for Standard [...]
-  The COVID Tracking Project - The COVID Tracking Project collects and [...]
-  EHDP Large Health Data Sets
-  GDC - GDC supports several cancer genome programs for CCG, TCGA, TARGET etc.
-  Gapminder World demographic databases
-  MeSH, the vocabulary thesaurus used for indexing articles for PubMed
-  MeDAL - A large medical text dataset curated for abbreviation [...]
-  Medicare Coverage Database (MCD), U.S.
-  Medicare Data Engine of medicare.gov Data
-  Medicare Data File
-  Number of Ebola Cases and Deaths in Affected Countries (2014)
-  Open-ODS (structure of the UK NHS)
-  OpenPaymentsData, Healthcare financial relationship data
-  PhysioBank Databases - A large and growing archive of physiological data.
-  The Cancer Imaging Archive (TCIA)
-  The Cancer Genome Atlas project (TCGA)
-  World Health Organization Global Health Observatory
-  Yahoo Knowledge Graph COVID-19 Datasets - The Yahoo Knowledge Graph team [...]
-  Informatics for Integrating Biology & the Bedside [fixme]

ImageProcessing

-  10k US Adult Faces Database



-  2GB of Photos of Cats
-  Audience Unfiltered faces for gender and age classification
-  Affective Image Classification
-  Airborne Object Detection and Tracking - The Airborne Object Tracking [...]
-  Animals with attributes
-  CADDY Underwater Stereo-Vision Dataset of divers' hand gestures - [...]
-  Cytology Dataset – CCAgT: Images of Cervical Cells with AgNOR Stain [...]
-  Caltech Pedestrian Detection Benchmark
-  Chars74K dataset - Character Recognition in Natural Images (both English [...]
-  Cube++ - 4890 raw 18-megapixel images, each containing a SpyderCube color [...]
-  Densely Annotated Video Driving Data Set - This data set consists of 28 [...]
-  Danbooru Tagged Anime Illustration Dataset - A large-scale anime image [...]
-  DukeMTMC Data Set - DukeMTMC aims to accelerate advances in multi-target [...]
[fixme]
-  ETH Entomological Collection (ETHEC) Fine Grained Butterfly (Lepidoptra) Images
[fixme]
-  Face Recognition Benchmark
-  Flickr: 32 Class Brand Logos [fixme]
-  GDxray - X-ray images for X-ray testing and Computer Vision
-  HumanEva Dataset - The HumanEva-I dataset contains 7 calibrated video [...]
-  ImageNet (in WordNet hierarchy)
-  Indoor Scene Recognition
-  International Affective Picture System, UFL
-  KITTI Vision Benchmark Suite
-  Labeled Information Library of Alexandria - Biology and Conservation - [...]
-  MNIST database of handwritten digits, near 1 million examples
-  Multi-View Region of Interest Prediction Dataset for Autonomous Driving - [...]
-  Massive Visual Memory Stimuli, MIT
-  Newspaper Navigator - This dataset consists of extracted visual content [...]
-  Open Images From Google - Pictures with segmentation masks for 2.8 [...]
-  RuFa - Contains images of text written in one of two Arabic fonts (Ruqaa [...]
-  SUN database, MIT
-  SVIRO Synthetic Vehicle Interior Rear Seat Occupancy - 25.000 synthetic [...]
-  Several Shape-from-Silhouette Datasets [fixme]
-  Stanford Dogs Dataset
-  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

MachineLearning

-  All-Age-Faces Dataset - Contains 13'322 Asian face images distributed [...]
-  Audi Autonomous Driving Dataset - We have published the Audi Autonomous [...]
-  Context-aware data sets from five domains
-  Delve Datasets for classification and regression
-  Discogs Monthly Data
-  Free Music Archive
-  IMDb Database
-  Iranis - A Large-scale Dataset of Farsi/Arabic License Plate Characters
-  Keel Repository for classification, regression and time series
-  Labeled Faces in the Wild (LFW)
-  Lending Club Loan Data
-  Machine Learning Data Set Repository [fixme]
-  Million Song Dataset [fixme]
-  More Song Datasets [fixme]
-  MovieLens Data Sets
-  New Yorker caption contest ratings
-  RDataMining - "R and Data Mining" ebook data
-  Registered Meteorites on Earth [fixme]
-  Restaurants Health Score Data in San Francisco
-  TikTok Dataset - More than 300 dance videos that capture a single person [...]
-  UCI Machine Learning Repository
-  Yahoo! Ratings and Classification Data
-  YouTube-BoundingBoxes
-  Youtube 8m
-  eBay Online Auctions (2012)















Museums

-  Canada Science and Technology Museums Corporation's Open Data
-  Cooper-Hewitt's Collection Database

















-  Metropolitan Museum of Art Collection API
-  Minneapolis Institute of Arts metadata
-  Natural History Museum (London) Data Portal
-  Rijksmuseum Historical Art Collection
-  Tate Collection metadata
-  The Getty vocabularies

NaturalLanguage

-  Automatic Keyphrase Extraction
-  The Big Bad NLP Database [fixme]
-  Blizzard Challenge Speech - The speech + text data comes from [...]
-  Blogger Corpus
-  CLiPS Stylometry Investigation Corpus [fixme]
-  ClueWeb09 FACC
-  ClueWeb12 FACC
-  DBpedia - Structured data from Wikipedia
-  Dirty Words - With millions of images in our library and billions of [...]
-  Flickr Personal Taxonomies [fixme]
-  Freebase of people, places, and things [fixme]
-  German Political Speeches Corpus - Collection of political speeches from [...]
-  Google Books Ngrams (2.2TB)
-  Google MC-AFP - Generated based on the public available Gigaword dataset [...]
-  Google Web 5gram (1TB, 2006)
-  Gutenberg eBooks List [fixme]
-  Hansards text chunks of Canadian Parliament [fixme]
-  LJ Speech - Speech dataset consisting of 13,100 short audio clips of a [...]
-  M-AILabs Speech - The M-AILABS Speech Dataset is the first large dataset [...]
[fixme]
-  Microsoft MACHine Reading COMprehension Dataset (or MS MARCO)
-  Machine Comprehension Test (MCTest) of text from Microsoft Research
-  Machine Translation of European languages
-  Making Sense of Microposts 2013 - Concept Extraction [fixme]
-  Making Sense of Microposts 2016 - Named Entity rEcognition and Linking
-  Multi-Domain Sentiment Dataset (version 2.0)

-  Noisy speech database for training speech enhancement algorithms and TTS [...] [fixme]
-  Open Multilingual Wordnet
-  POS/NER/Chunk annotated data
-  Personae Corpus [fixme]
-  SMS Spam Collection in English
-  SaudiNewsNet Collection of Saudi Newspaper Articles (Arabic, 30K articles)
-  Stanford Question Answering Dataset (SQuAD)
-  USENET postings corpus of 2005~2011
-  Universal Dependencies
-  Webhose - News/Blogs in multiple languages
-  Wikidata - Wikipedia databases
-  Wikipedia Links data - 40 Million Entities in Context
-  WordNet databases and tools
-  WorldTree Corpus of Explanation Graphs for Elementary Science Questions - [...]






















Neuroscience

-  Allen Institute Datasets
-  Brain Catalogue
-  Brainomics [fixme]
-  CodeNeuro Datasets [fixme]
-  Collaborative Research in Computational Neuroscience (CRCNS)
-  FCP-INDI
-  Human Connectome Project
-  NDAR
-  NIMH Data Archive
-  NeuroData
-  NeuroMorpho - NeuroMorpho.Org is a centrally curated inventory of [...]
-  Neuroelectro
-  OASIS
-  OpenNEURO
-  OpenfMRI
-  Study Forrest

Physics

-  CERN Open Data Portal
-  Crystallography Open Database
-  IceCube - South Pole Neutrino Observatory
-  Ligo Open Science Center (LOSC) - Gravitational wave data from the LIGO [...]
-  NASA Exoplanet Archive
-  NSSDC (NASA) data of 550 space spacecraft
-  Sloan Digital Sky Survey (SDSS) - Mapping the Universe

ProstateCancer

-  EOPC-DE-Early-Onset-Prostate-Cancer-Germany - Early Onset Prostate Cancer [...]
-  GENIE - Data from the Genomics Evidence Neoplasia Information Exchange [...]
-  Genomic-Hallmarks-Prostate-Adenocarcinoma-CPC-GENE - Comprehensive [...]
[fixme]
-  MSK-IMPACT-Clinical-Sequencing-Cohort-MSKCC-Prostate-Cancer - Targeted [...]
[fixme]
-  Metastatic-Prostate-Adenocarcinoma-MCTP - Comprehensive profiling of 61 [...]
[fixme]
-  Metastatic-Prostate-Cancer-SU2CPCF-Dream-Team - Comprehensive analysis of [...]
[fixme]
-  NPCR-2001-2015 - Database from CDC's National Program of Cancer [...]
-  NPCR-2005-2015 - Database from CDC's National Program of Cancer [...]
-  NaF-Prostate - NaF Prostate is a collection of F-18 NaF positron emission [...]
-  Neuroendocrine-Prostate-Cancer - Whole exome and RNA Seq data of [...] [fixme]
-  PLCO-Prostate-Diagnostic-Procedures - The Prostate Diagnostic Procedures [...]
-  PLCO-Prostate-Medical-Complications - The Prostate Medical Complications [...]
-  PLCO-Prostate-Screening-Abnormalities - The Prostate Screening [...]
-  PLCO-Prostate-Screening - The Prostate Screening dataset (177,315 [...])
-  PLCO-Prostate-Treatments - The Prostate Treatments dataset (13,409 [...])
-  PLCO-Prostate - The Prostate dataset is a comprehensive dataset that [...]
-  PRAD-CA-Prostate-Adenocarcinoma-Canada - Prostate Adenocarcinoma - [...]
-  PRAD-FR-Prostate-Adenocarcinoma-France - Prostate Adenocarcinoma - [...]
-  PRAD-UK-Prostate-Adenocarcinoma-United-Kingdom - Prostate Adenocarcinoma [...]
-  PROSTATEx-Challenge - Retrospective set of prostate MR studies. All [...]
-  Prostate-3T - The Prostate-3T project provided imaging data to TCIA as [...]

-  Prostate-Adenocarcinoma-Broad-Cornell-2012 - Comprehensive profiling of [...] [fixme]
-  Prostate-Adenocarcinoma-Broad-Cornell-2013 - Comprehensive profiling of [...] [fixme]
-  Prostate-Adenocarcinoma-CNA-study-MSKCC - Copy-number profiling of 103 [...] [fixme]
-  Prostate-Adenocarcinoma-Fred-Hutchinson-CRC - Comprehensive profiling of [...] [fixme]
-  Prostate Adenocarcinoma (MSKCC/DFCI) - Whole Exome Sequencing of 1013 [...] [fixme]
-  Prostate-Adenocarcinoma-MSKCC - MSKCC Prostate Oncogenome Project. 181 [...] [fixme]
-  Prostate-Adenocarcinoma-Organoids-MSKCC - Exome profiling of prostate [...] [fixme]
-  Prostate-Adenocarcinoma-Sun-Lab - Whole-genome and Transcriptome [...] [fixme]
-  Prostate-Adenocarcinoma-TCGA-PanCancer-Atlas - Comprehensive TCGA [...] [fixme]
-  Prostate-Adenocarcinoma-TCGA - Integrated profiling of 333 primary [...] [fixme]
-  Prostate-Diagnosis - PCa T1- and T2-weighted magnetic resonance images [...]
-  Prostate-Fused-MRI-Pathology - The Prostate Fused-MRI-Pathology [...]
-  Prostate-MRI - The Prostate-MRI collection of prostate Magnetic Resonance [...]
-  Prostate-R - The R package 'ElemStatLearn' contains a prostate cancer [...]
-  QIN-PROSTATE-Repeatability - The QIN-PROSTATE-Repeatability dataset is a [...]
-  QIN-PROSTATE - The QIN PROSTATE collection of the Quantitative Imaging [...]
-  SEER-YR1973_2015.SEER9 - The SEER November 2017 Research Data files from [...]
-  SEER-YR1992_2015.SJ_LA_RG_AK - The SEER November 2017 Research Data files [...]
-  SEER-YR2000_2015.CA_KY_LO_NJ_GA - The SEER November 2017 Research Data [...]
-  SEER-YR2000_2015.CA_KY_LO_NJ_GA - The July - December 2005 diagnoses for [...]
-  TCGA-PRAD-US - TCGA Prostate Adenocarcinoma (499 samples). [fixme]





Psychology+Cognition

-  OSU Cognitive Modeling Repository Datasets [fixme]

PublicDomains






-  [Ably Open Realtime Data](#)
-  [Amazon](#)
-  [Archive.org Datasets](#)
-  [Archive-it from Internet Archive](#)
-  [CMU JASA data archive](#)
-  [CMU StatLab collections](#)
-  [Data.World](#)
-  [Data360 \[fixme\]](#)
-  [Enigma Public](#)
-  [Google](#)
-  [Grand Comics Database - The Grand Comics Database \(GCD\) is a nonprofit, \[...\]](#)
-  [Infochimps \[fixme\]](#)
-  [KDNuggets Data Collections](#)
-  [Microsoft Azure Data Market Free DataSets \[fixme\]](#)
-  [Microsoft Data Science for Research](#)
-  [Microsoft Research Open Data](#)
-  [Open Library Data Dumps](#)
-  [Reddit Datasets \[fixme\]](#)
-  [RevolutionAnalytics Collection \[fixme\]](#)
-  [Sample R data sets](#)
-  [StatSci.org](#)
-  [Stats4Stem R data sets \(archived\)](#)
-  [The Washington Post List](#)
-  [UCLA SOCR data collection](#)
-  [UFO Reports](#)
-  [Wikileaks 911 pager intercepts](#)
-  [Yahoo Webscope](#)







SearchEngines

-  [Academic Torrents of data sharing from UMB](#)
-  [Base dos Dados - Data Basis: Open Data Repository for Brazil](#)
-  [Datahub.io](#)
-  [Domains Project - Sorted list of Internet domains](#)



























-  [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](#)






SocialNetworks

-  [2021 Portuguese Elections Twitter Dataset - 57M+ tweets, 1M+ users - This \[...\]](#)
-  [72 hours #gamergate Twitter Scrape](#)
-  [CMU Enron Email of 150 users](#)
-  [Cheng-Caverlee-Lee September 2009 - January 2010 Twitter Scrape](#)
-  [China Biographical Database - The China Biographical Database is a freely \[...\]](#)
-  [A Twitter Dataset of 40+ million tweets related to COVID-19 - Due to the \[...\]](#)
-  [43k+ Donald Trump Twitter Screenshots - This archive contains screenshots \[...\]](#)
-  [EDRM Enron EMail of 151 users, hosted on S3](#)
-  [Facebook Data Scrape \(2005\)](#)
-  [Facebook Social Connectedness Index - We use an anonymized snapshot of \[...\]](#)
-  [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](#)
-  [Twitch Top Streamer's Data](#)
-  [Twitter Data for Online Reputation Management](#)
-  [Twitter Data for Sentiment Analysis](#)

-  Twitter Graph of entire Twitter site [fixme]
-  Twitter Scrape Calufa May 2011 [fixme]
-  UNIMI/LAW Social Network Datasets
-  United States Congress Twitter Data - Daily datasets with tweets of 1100+ [...]
-  Yahoo! Graph and Social Data
-  Youtube Video Social Graph in 2007,2008

SocialSciences


-  ACLED (Armed Conflict Location & Event Data Project)
-  Authoritarian Ruling Elites Database - The Authoritarian Ruling Elites [...]
-  Canadian Legal Information Institute
-  Center for Systemic Peace Datasets - Conflict Trends, Politics, State Fragility, etc [fixme]
-  Correlates of War Project
-  Cryptome Conspiracy Theory Items
-  Datacards [fixme]
-  European Social Survey
-  FBI Hate Crime 2013 - aggregated data
-  Fragile States Index [fixme]
-  GDELT Global Events Database
-  General Social Survey (GSS) since 1972
-  German Social Survey
-  Global Religious Futures Project
-  Gun Violence Data - A comprehensive, accessible database that contains [...]
-  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
-  MIT Reality Mining Dataset
-  MacroData Guide by Norsk samfunnsvitenskapelig datatjeneste
-  Mass Mobilization Data Project - The Mass Mobilization (MM) data are an [...]
-  Microsoft Academic Knowledge Graph - The Microsoft Academic Knowledge [...]













-  Minnesota Population Center
-  Notre Dame Global Adaptation Index (ND-GAIN)
-  Open Crime and Policing Data in England, Wales and Northern Ireland
-  OpenSanctions - A global database of persons and companies of political, [...]
-  Paul Hensel General International Data Page
-  PewResearch Internet Survey Project
-  PewResearch Society Data Collection
-  Political Polarity Data [fixme]
-  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
-  UPJOHN for Labor Employment Research
-  Universities Worldwide
-  Uppsala Conflict Data Program
-  World Bank Open Data
-  World Inequality Database - The World Inequality Database (WID.world) [...]
-  WorldPop project - Worldwide human population distributions [fixme]

Software








-  FLOSSmole data about free, libre, and open source software development
-  GHTorrent - Scalable, queryable, offline mirror of data offered through [...]
-  Libraries.io Open Source Repository and Dependency Metadata
-  Public Git Archive - a Big Code dataset for all – dataset of 182,014 top- [...]
-  Code duplicates - 2k Java file and 600 Java function pairs labeled as [...]
-  Commit messages - 1.3 billion GitHub commit messages till March 2019
-  Pull Request review comments - 25.3 million GitHub PR review comments [...]
-  Source Code Identifiers - 41.7 million distinct splittable identifiers [...]

Sports

-  American Ninja Warrior Obstacles - Contains every obstacle in the history [...]
-  Betfair Historical Exchange Data














-  Cricsheet Matches (cricket)
-  Equity in Athletics - The Equity in Athletics Data Analysis Cutting Tool [...]
-  Ergast Formula 1, from 1950 up to date (API)
-  Football/Soccer resources (data and APIs)
-  Lahman's Baseball Database
-  NFL play-by-play data - NFL play-by-play data sourced from: [...]
-  Pinhooker: Thoroughbred Bloodstock Sale Data
-  Pro Kabadi season 1 to 7 - Pro Kabadi League is a professional-level [...]
-  Retrosheet Baseball Statistics
-  Tennis database of rankings, results, and stats for ATP
-  Tennis database of rankings, results, and stats for WTA
-  USA Soccer Teams and Locations - USA soccer teams and locations. MLS, [...]

TimeSeries




-  3W dataset - To the best of its authors' knowledge, this is the first [...]
-  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
-  Turing Change Point Dataset - Contains 42 annotated time series collected [...]
-  UC Riverside Time Series Dataset

Transportation

-  Airlines OD Data 1987-2008
-  Ford GoBike Data (formerly Bay Area Bike Share Data) [fixme]
-  Bike Share Systems (BSS) collection
-  Dutch Traffic Information [fixme]
-  GeoLife GPS Trajectory from Microsoft Research
-  German train system by Deutsche Bahn [fixme]
-  Hubway Million Rides in MA [fixme]
-  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](#) [fixme]
-  [RITA/BTS transport data collection \(TranStat\)](#) [fixme]
-  [Renfe](#) (Spanish National Railway Network) dataset
-  [Toronto Bike Share Stations](#) (JSON and GBFS files)
-  [Transport for London](#) (TFL)
-  [Travel Tracker Survey \(TTS\) for Chicago](#) [fixme]
-  [U.S. Bureau of Transportation Statistics \(BTS\)](#)
-  [U.S. Domestic Flights 1990 to 2009](#)
-  [U.S. Freight Analysis Framework since 2007](#)
-  [U.S. National Highway Traffic Safety Administration - Fatalities since \[...\]](#)

eSports

-  [CS:GO Competitive Matchmaking Data](#) - In this data set we have data about [...]
-  [FIFA-2021 Complete Player Dataset](#)
-  [OpenDota data dump](#)

Complementary Collections

- [Data Packaged Core Datasets](#)
- [Database of Scientific Code Contributions](#)
- A growing collection of public datasets: [CoolDatasets](#).
- 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](#)
- CV Papers: [CV Datasets on the web](#)
- CVonline: [Image Databases](#)

Releases

 1 tags

Packages

No packages published

Contributors 157



+ 146 contributors

Environments 1

 **github-pages** Active