

Data Science

Public Datasets

Datasets for your next project in

Data Science, Machine Learning, Deep Learning

Himanshu Ramchandani







<https://www.linkedin.com/in/hemansnation/>

Credit: <https://github.com/awesomedata/awesome-public-datasets>


Table of Contents

- Agriculture
- Architecture
- Biology
- Chemistry
- 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 Global Dataset of [...] [Meta]
-  Hyperspectral benchmark dataset on soil moisture - This dataset was measured in a five-day [...] [Meta]
-  Lemons quality control dataset - Lemon dataset has been prepared to investigate the [...] [Meta]
-  Optimized Soil Adjusted Vegetation Index - The IDB is a tool for working with remote sensing [...] [Meta]
-  U.S. Department of Agriculture's Nutrient Database [Meta]
-  U.S. Department of Agriculture's PLANTS Database - The Complete PLANTS Checklist is nearly 7 [...] [Meta]











Architecture

-  Swiss Apartment Models - This dataset contains detailed data on 42,207 apartments (242,257 [...] [Meta]

Biology

-  1000 Genomes - The 1000 Genomes Project ran between 2008 and 2015, creating the largest [...] [Meta]
-  American Gut (Microbiome Project) - The American Gut project is the largest crowdsourced [...] [Meta]
-  BCNB - There are WSIs of 1058 patients, part of tumor regions are annotated in WSIs. Except [...] [Meta]
-  Broad Bioimage Benchmark Collection (BBBC) - The Broad Bioimage Benchmark Collection (BBBC) [...] [Meta]
-  Broad Cancer Cell Line Encyclopedia (CCLE) [Meta]
-  Cell Image Library - This library is a public and easily accessible resource database of [...] [Meta]
-  Complete Genomics Public Data - A diverse data set of whole human genomes are freely [...] [Meta]

-  CytolImageNet - A large-scale dataset of microscopy images. Contains 890,737 total grayscale [...] [Meta]
-  EBI ArrayExpress - ArrayExpress Archive of Functional Genomics Data stores data from high- [...] [Meta]
-  EBI Protein Data Bank in Europe - The Electron Microscopy Data Bank (EMDB) is a public [...] [Meta]
-  ENCODE project - The Encyclopedia of DNA Elements (ENCODE) Consortium is an ongoing [...] [Meta]
-  Electron Microscopy Pilot Image Archive (EMPIAR) - EMPIAR, the Electron Microscopy Public [...] [Meta]
-  Ensembl Genomes [Meta]
-  Gene Expression Omnibus (GEO) - GEO is a public functional genomics data repository [...] [Meta]
-  Gene Ontology (GO) - GO annotation files [Meta]
-  Global Biotic Interactions (GloBI) [Meta]
-  Harvard Medical School (HMS) LINCS Project - The Harvard Medical School (HMS) LINCS Center is [...] [Meta]
-  Human Genome Diversity Project - A group of scientists at Stanford University have [...] [Meta]
-  Human Microbiome Project (HMP) - The HMP sequenced over 2000 reference genomes isolated from [...] [Meta]
-  ICOS PSP Benchmark - The ICOS PSP benchmarks repository contains an adjustable real-world [...] [Meta]
-  International HapMap Project [Meta]
-  Journal of Cell Biology DataViewer [Meta]
-  KEGG - KEGG is a database resource for understanding high-level functions and utilities of [...] [Meta]
-  NCBI Proteins [Meta]
-  NCBI Taxonomy - The NCBI Taxonomy database is a curated set of names and classifications for [...] [Meta]
-  NCI Genomic Data Commons - The GDC Data Portal is a robust data-driven platform that allows [...] [Meta]
-  NIH Microarray data [Meta]
-  OpenSNP genotypes data - openSNP allows customers of direct-to-customer genetic tests to [...] [Meta]

-  Palmer Penguins - The goal of palmerpenguins is to provide a great dataset for data [...] [Meta]
-  Pathguid - Protein-Protein Interactions Catalog [Meta]
-  Protein Data Bank - This resource is powered by the Protein Data Bank archive-information [...] [Meta]
-  Psychiatric Genomics Consortium - The purpose of the Psychiatric Genomics Consortium (PGC) is [...] [Meta]
-  PubChem Project - PubChem is the world's largest collection of freely accessible chemical [...] [Meta]
-  PubGene (now Coremine Medical) - COREMINE™ is a family of tools developed by the Norwegian [...] [Meta]
-  Sanger Catalogue of Somatic Mutations in Cancer (COSMIC) - COSMIC, the Catalogue Of Somatic [...] [Meta]
-  Sanger Genomics of Drug Sensitivity in Cancer Project (GDSC) [Meta]
-  Sequence Read Archive(SRA) - The Sequence Read Archive (SRA) stores raw sequence data from [...] [Meta]
-  Serratus - Analysis of 7.1 million RNA/DNA sequencing datasets to discover the total [...] [Meta]
-  Stanford Microarray Data (Retired NOW) [Meta]
-  Stowers Institute Original Data Repository [Meta]
-  Systems Science of Biological Dynamics (SSBD) Database - Systems Science of Biological [...] [Meta]
-  The Cancer Genome Atlas (TCGA), available via Broad GDAC [Meta]
-  The Catalogue of Life - The Catalogue of Life is a quality-assured checklist of more than 1.8 [...] [Meta]
-  The Personal Genome Project - The Personal Genome Project, initiated in 2005, is a vision and [...] [Meta]
-  UCSC Public Data [Meta]
-  UniGene [Meta]
-  Universal Protein Resource (UniProt) - The Universal Protein Resource (UniProt) is a [...] [Meta]
-  Rfam - The Rfam database is a collection of RNA families, each represented by multiple [...] [Meta]


Chemistry

-  [Ionic Liquids Database - ILThermo \[Meta\]](#)







Climate+Weather










-  [Actuaries Climate Index \[Meta\]](#)
-  [Australian Weather \[Meta\]](#)
-  [Aviation Weather Center - Consistent, timely and accurate weather information for the world \[...\] \[Meta\]](#)
-  [Brazilian Weather - Historical data \(In Portuguese\) - Data related to climate and weather \[...\] \[Meta\]](#)
-  [Canadian Meteorological Centre \[Meta\]](#)
-  [Climate Data from UEA \(updated monthly\) \[Meta\]](#)
-  [Dutch Weather - The KNMI Data Center \(KDC\) portal provides access to KNMI data on weather, \[...\] \[Meta\]](#)
-  [European Climate Assessment & Dataset \[Meta\]](#)
-  [German Climate Data Center \[Meta\]](#)
-  [Global Climate Data Since 1929 \[Meta\]](#)
-  [Charting The Global Climate Change News Narrative 2009-2020 - These four datasets represent \[...\] \[Meta\]](#)
-  [NASA Global Imagery Browse Services \[Meta\]](#)
-  [NOAA Bering Sea Climate \[Meta\]](#)
-  [NOAA Climate Datasets \[Meta\]](#)
-  [NOAA Realtime Weather Models \[Meta\]](#)
-  [NOAA SURFRAD Meteorology and Radiation Datasets \[Meta\]](#)
-  [The World Bank Open Data Resources for Climate Change \[Meta\]](#)
-  [UEA Climatic Research Unit \[Meta\]](#)
-  [WU Historical Weather Worldwide \[Meta\]](#)
-  [Washington Post Climate Change - To analyze warming temperatures in the United States, The \[...\] \[Meta\]](#)
-  [WorldClim - Global Climate Data \[Meta\]](#)

ComplexNetworks



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

ComputerNetworks

-  3.5B Web Pages from CommonCrawl 2012 [Meta]
-  53.5B Web clicks of 100K users in Indiana Univ. [Meta]
-  CAIDA Internet Datasets [Meta]
-  CRAWDAD Wireless datasets from Dartmouth Univ. [Meta]
-  ClueWeb09 - 1B web pages [Meta]
-  ClueWeb12 - 733M web pages [Meta]
-  CommonCrawl Web Data over 7 years [Meta]

-  Shopper Intent Prediction from Clickstream E-Commerce Data with Minimal Browsing Information [Meta]
-  Criteo click-through data [Meta]
-  Internet-Wide Scan Data Repository [Meta]
-  MIRAGE-2019 - MIRAGE-2019 is a human-generated dataset for mobile traffic analysis with [...] [Meta]
-  OONI: Open Observatory of Network Interference - Internet censorship data [Meta]
-  Open Mobile Data by MobiPerf [Meta]
-  The Peer-to-Peer Trace Archive - Real-world measurements play a key role in studying the [...] [Meta]
-  Rapid7 Sonar Internet Scans [Meta]
-  UCSD Network Telescope, IPv4 /8 net [Meta]

CyberSecurity















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

DataChallenges

-  Alcrowd Competitions [Meta]
-  Bruteforce Database [Meta]
-  Challenges in Machine Learning [Meta]
-  CrowdANALYTIX dataX [Meta]
-  D4D Challenge of Orange [Meta]
-  DrivenData Competitions for Social Good [Meta]
-  ICWSM Data Challenge (since 2009) [Meta]
-  KDD Cup by Tencent 2012 [Meta]
-  Kaggle Competition Data [Meta]
-  Localytics Data Visualization Challenge [Meta]
-  Netflix Prize [Meta]














-  Space Apps Challenge [Meta]
-  Telecom Italia Big Data Challenge [Meta]
-  TravisTorrent Dataset - MSR'2017 Mining Challenge [Meta]
-  TunedIT - Data mining & machine learning data sets, algorithms, challenges [Meta]
-  Yelp Dataset Challenge - The Yelp dataset is a subset of our businesses, reviews, and user [...] [Meta]

EarthScience





-  38-Cloud (Cloud Detection) - Contains 38 Landsat 8 scene images and their manually extracted [...] [Meta]
-  AQUASTAT - Global water resources and uses [Meta]
-  BODC - marine data of ~22K vars [Meta]
-  EOSDIS - NASA's earth observing system data [Meta]
-  Earth Models [Meta]
-  Global Wind Atlas - The Global Wind Atlas is a free, web-based application developed to help [...] [Meta]
-  Integrated Marine Observing System (IMOS) - roughly 30TB of ocean measurements [Meta]
-  Marinexplore - Open Oceanographic Data [Meta]
-  Alabama Real-Time Coastal Observing System [Meta]
-  National Estuarine Research Reserves System-Wide Monitoring Program - long-term estuarine [...] [Meta]
-  Oil and Gas Authority Open Data - The dataset covers 12,500 offshore wellbores, 5,000 seismic [...] [Meta]
-  Smithsonian Institution Global Volcano and Eruption Database [Meta]
-  USGS Earthquake Archives [Meta]
-  Wellhead Protection Area (protection zone) prediction using breakthrough curves - This [...] [Meta]

Economics





-  Asian Productivity Organization (APO) - The AEPM provides a graphic dashboard view of [...] [Meta]
-  ASEAN Stats - The ASEANstatsDataPortal was first launched in June 2018. The Portal is [...] [Meta]
-  American Economic Association (AEA) [Meta]
-  Asian KLEMS - Asia KLEMS is an Asian regional research consortium to promote building [...] [Meta]
-  Harvard Atlas of Economic Complexity - A database for people to explore global trade flows [...] [Meta]
-  BIS Financial Database - The files contain the same data as in the BIS Statistics Explorer [...] [Meta]
-  Barro-Lee Education Attainment - Barro-Lee Educational Attainment Data from 1950 to 2010. [...] [Meta]
-  CEPII Database - A database of the world economy, through its country and region profiles, in [...] [Meta]
-  EUKLEMS - EU KLEMS is an industry level, growth and productivity research project. EU KLEMS [...] [Meta]
-  Economic Freedom of the World Data [Meta]
-  Historical National Accounts - The datahub on Comparative Historical National Accounts [...] [Meta]
-  Historical Macroeconomic Statistics [Meta]
-  INFORUM - Interindustry Forecasting at the University of Maryland [Meta]
-  DBnomics – the world's economic database - Aggregates hundreds of millions of time series [...] [Meta]
-  International Trade Statistics [Meta]
-  Internet Product Code Database [Meta]
-  Joint External Debt Data Hub [Meta]
-  Jon Haveman International Trade Data Links [Meta]
-  Latin America KLEMS - LAKLEMS is a technical cooperation project financed by the Inter- [...] [Meta]
-  Long-Term Productivity Database - The Long-Term Productivity database was created as a [...] [Meta]




















-  Maddison Project Database - The Maddison Project Database provides information on comparative [...] [Meta]
-  National Transfer Accounts - The goal of the National Transfer Accounts (NTA) project is to [...] [Meta]
-  OpenCorporates Database of Companies in the World [Meta]
-  Our World in Data [Meta]
-  Penn World Table - PWT version 10.0 is a database with information on relative levels of [...] [Meta]
-  SciencesPo World Trade Gravity Datasets [Meta]
-  The Atlas of Economic Complexity [Meta]
-  The Center for International Data [Meta]
-  The Observatory of Economic Complexity [Meta]
-  UN Commodity Trade Statistics [Meta]
-  UN Human Development Reports [Meta]
-  World Input-Output Database - World Input-Output Tables and underlying data, covering 43 [...] [Meta]
-  World KLEMS - Analytical KLEMS-type data sets for a broad set of countries around the world. [...] [Meta]

Education

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

Energy















-  AMPds - The Almanac of Minutely Power dataset [Meta]
-  BLUEd - Building-Level fully labeled Electricity Disaggregation dataset [Meta]
-  COMBED [Meta]
-  DBFC - Direct Borohydride Fuel Cell (DBFC) Dataset [Meta]

-  DEL - Domestic Electrical Load study datasets for South Africa (1994 - 2014) [Meta]
-  ECO - The ECO data set is a comprehensive data set for non-intrusive load monitoring and [...] [Meta]
-  EIA [Meta]
-  Global Power Plant Database - The Global Power Plant Database is a comprehensive, open source [...] [Meta]
-  HES - Household Electricity Study, UK [Meta]
-  HFED [Meta]
-  MORED: a Moroccan Buildings' Electricity Consumption Dataset - Since spring of 2019, a data [...] [Meta]
-  Marktstammdatenregister - The German Marktstammdatenregister (MaStR) is a database of all [...] [Meta]
-  PEM1 - Proton Exchange Membrane (PEM) Fuel Cell Dataset [Meta]
-  PLAID - The Plug Load Appliance Identification Dataset [Meta]
-  The Public Utility Data Liberation Project (PUDL) - PUDL makes US energy data easier to [...] [Meta]
-  REDD [Meta]
-  SYND - A synthetic energy dataset for non-intrusive load monitoring - With SynD, we present a [...] [Meta]
-  Smart Meter Data Portal - The Smart Meter Data Portal is part of the National Science [...] [Meta]
-  Tracebase [Meta]
-  Ukraine Energy Centre Datasets [Meta]
-  UK-DALE - UK Domestic Appliance-Level Electricity [Meta]
-  WHITED [Meta]
-  iAWE [Meta]










Entertainment

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

Finance

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



























GIS

-  Awesome 3D Semantic City Models - Collection of open 3D semantic city and region models. [Meta]
-  ArcGIS Open Data portal [Meta]
-  Cambridge, MA, US, GIS data on GitHub [Meta]
-  Database of all continents, countries, States/Subdivisions/Provinces and Cities - Database [...] [Meta]
-  Factual Global Location Data [Meta]
-  IEEE Geoscience and Remote Sensing Society DASE Website [Meta]
-  Geo Maps - High Quality GeoJSON maps programmatically generated [Meta]
-  Geo Spatial Data from ASU [Meta]
-  Geo Wiki Project - Citizen-driven Environmental Monitoring [Meta]

-  GeoFabrik - OSM data extracted to a variety of formats and areas [Meta]
-  GeoNames Worldwide [Meta]
-  Global Administrative Areas Database (GADM) - Geospatial data organized by country. Includes [...] [Meta]
-  Homeland Infrastructure Foundation-Level Data [Meta]
-  Landsat 8 on AWS [Meta]
-  List of all countries in all languages [Meta]
-  National Weather Service GIS Data Portal [Meta]
-  Natural Earth - vectors and rasters of the world [Meta]
-  OpenAddresses [Meta]
-  OpenStreetMap (OSM) [Meta]
-  Pleiades - Gazetteer and graph of ancient places [Meta]
-  Reverse Geocoder using OSM data [Meta]
-  Robin Wilson - Free GIS Datasets [Meta]
-  Shadow Accrual Maps - The repository contains the accumulated shadow information for New York [...] [Meta]
-  TIGER/Line - U.S. boundaries and roads [Meta]
-  TZ Timezones shapefile [Meta]
-  TwoFishes - Foursquare's coarse geocoder [Meta]
-  UN Environmental Data [Meta]
-  World boundaries from the U.S. Department of State [Meta]
-  World countries in multiple formats [Meta]

Government

-  Alberta, Province of Canada [Meta]
-  Antwerp, Belgium [Meta]
-  Argentina (non official) [Meta]
-  Datos Argentina - Portal de datos abiertos de la República Argentina. Encontrá datos públicos [...] [Meta]
-  Austin, TX, US [Meta]
-  Australia (abs.gov.au) [Meta]
-  Australia (data.gov.au) [Meta]

-  [Austria \(data.gv.at\) \[Meta\]](#)
-  [Baton Rouge, LA, US \[Meta\]](#)
-  [Beersheba, Israel - Open Data Portal \(Smart7 OpenData\) \[Meta\]](#)
-  [Belgium \[Meta\]](#)
-  [City of Berkeley Open Data \[Meta\]](#)
-  [Brazil \[Meta\]](#)
-  [Buenos Aires, Argentina \[Meta\]](#)
-  [Calgary, AB, Canada \[Meta\]](#)
-  [Cambridge, MA, US \[Meta\]](#)
-  [Canada \[Meta\]](#)
-  [Chicago \[Meta\]](#)
-  [Chile \[Meta\]](#)
-  [China \[Meta\]](#)
-  [Dallas Open Data \[Meta\]](#)
-  [DataBC - data from the Province of British Columbia \[Meta\]](#)
-  [Debt to the Penny - The Debt to the Penny dataset provides information about the total \[...\] \[Meta\]](#)
-  [Denver Open Data \[Meta\]](#)
-  [Durham, NC Open Data \[Meta\]](#)
-  [Edmonton, AB, Canada \[Meta\]](#)
-  [England LGInform \[Meta\]](#)
-  [EuroStat \[Meta\]](#)
-  [EveryPolitician - Ongoing project collating and sharing data on every politician. \[Meta\]](#)
-  [Federal Committee on Statistical Methodology \(FCSM\) \(formerly FedStats\) \[Meta\]](#)
-  [Finland \[Meta\]](#)
-  [France \[Meta\]](#)
-  [Fredericton, NB, Canada \[Meta\]](#)
-  [Gatineau, QC, Canada \[Meta\]](#)
-  [Germany \[Meta\]](#)
-  [Ghent, Belgium \[Meta\]](#)
-  [Glasgow, Scotland, UK \[Meta\]](#)


-  Greece [Meta]
-  Guardian world governments [Meta]
-  Halifax, NS, Canada [Meta]
-  Helsinki Region, Finland [Meta]
-  Hong Kong, China [Meta]
-  Houston, TX, US [Meta]
-  Indian Government Data [Meta]
-  Indonesian Data Portal [Meta]
-  Iowa - Welcome to the State of Iowa's data portal. Please explore data about Iowa and your [...] [Meta]
-  Ireland's Open Data Portal [Meta]
-  Israel's Open Data Portal [Meta]
-  Istanbul Municipality Open Data Portal [Meta]
-  Italy - Il Portale dati.gov.it è il catalogo nazionale dei metadati relativi ai dati [...] [Meta]
-  Jail deaths in America - The U.S. government does not release jail by jail mortality data, [...] [Meta]
-  Japan [Meta]
-  Laval, QC, Canada [Meta]
-  Lexington, KY [Meta]
-  London Datastore, UK [Meta]
-  London, ON, Canada [Meta]
-  Los Angeles Open Data [Meta]
-  Luxembourg - Luxembourgish Open Data Portal [Meta]
-  MassGIS, Massachusetts, U.S. [Meta]
-  Metropolitan Transportation Commission (MTC), California, US [Meta]
-  Mexico [Meta]
-  Mississauga, ON, Canada [Meta]
-  Moldova [Meta]
-  Moncton, NB, Canada [Meta]
-  Montreal, QC, Canada [Meta]
-  Mountain View, California, US (GIS) [Meta]
-  NYC Open Data [Meta]

-  NYC betanyc [Meta]
-  Netherlands [Meta]
-  New York Department of Sanitation Monthly Tonnage - DSNY Monthly Tonnage Data provides [...] [Meta]
-  New Zealand [Meta]
-  OECD [Meta]
-  Oakland, California, US [Meta]
-  Oklahoma [Meta]
-  Open Data for Africa [Meta]
-  Open Government Data (OGD) Platform India [Meta]
-  OpenDataSoft's list of 1,600 open data [Meta]
-  Oregon [Meta]
-  Ottawa, ON, Canada [Meta]
-  Palo Alto, California, US [Meta]
-  OpenDataPhilly - OpenDataPhilly is a catalog of open data in the Philadelphia region. In [...] [Meta]
-  Portland, Oregon [Meta]
-  Portugal - Pordata organization [Meta]
-  Puerto Rico Government [Meta]
-  Quebec City, QC, Canada [Meta]
-  Quebec Province of Canada [Meta]
-  Regina SK, Canada [Meta]
-  Rio de Janeiro, Brazil [Meta]
-  Romania [Meta]
-  Russia [Meta]
-  San Diego, CA [Meta]
-  San Antonio, TX - Community Information Now - CI:Now is a nonprofit serving Bexar (San [...]) [Meta]
-  San Francisco Data sets [Meta]
-  San Jose, California, US [Meta]
-  San Mateo County, California, US [Meta]
-  Saskatchewan, Province of Canada [Meta]
-  Seattle [Meta]

-  Singapore Government Data [Meta]
-  South Africa Trade Statistics [Meta]
-  South Africa [Meta]
-  State of Utah, US [Meta]
-  Switzerland [Meta]
-  Taiwan gov [Meta]
-  Taiwan [Meta]
-  Tel-Aviv Open Data [Meta]
-  Texas Open Data [Meta]
-  The World Bank [Meta]
-  Toronto, ON, Canada [Meta]
-  Tunisia [Meta]
-  U.K. Government Data [Meta]
-  U.S. American Community Survey [Meta]
-  U.S. CDC Public Health datasets [Meta]
-  U.S. Census Bureau [Meta]
-  U.S. Department of Housing and Urban Development (HUD) [Meta]
-  U.S. Federal Government Agencies [Meta]
-  U.S. Federal Government Data Catalog [Meta]
-  U.S. Food and Drug Administration (FDA) [Meta]
-  U.S. National Center for Education Statistics (NCES) [Meta]
-  U.S. Open Government [Meta]
-  UK 2011 Census Open Atlas Project [Meta]
-  US Counties - This is a repository of various data, broken down by US county. While most of [...] [Meta]
-  U.S. Patent and Trademark Office (USPTO) Bulk Data Products [Meta]




















-  Uganda Bureau of Statistics [Meta]
-  Ukraine [Meta]
-  United Nations [Meta]
-  Uruguay [Meta]
-  Valley Transportation Authority (VTA), California, US [Meta]

- 
 Vancouver, BC Open Data Catalog [Meta]
- 
 Victoria, BC, Canada [Meta]
- 
 Vienna, Austria [Meta]
- 
 Statistics from the General Statistics Office of Vietnam - Data in different categories are [...] [Meta]
- 
 U.S. Congressional Research Service (CRS) Reports [Meta]









Healthcare

- 
 AWS COVID-19 Datasets - We're working with organizations who make COVID-19-related data [...] [Meta]
- 
 COVID-19 Case Surveillance Public Use Data - The COVID-19 case surveillance system database [...] [Meta]
- 
 Covid-19 non-processed data of Ecuador - It's a project which provides non-processed datasets [...] [Meta]
- 
 2019 Novel Coronavirus COVID-19 Data Repository by Johns Hopkins CSSE - This is the data [...] [Meta]
- 
 Coronavirus (Covid-19) Data in the United States - The New York Times is releasing a series [...] [Meta]
- 
 COVID-19 Reported Patient Impact and Hospital Capacity by Facility - The following dataset [...] [Meta]
- 
 Composition of Foods Raw, Processed, Prepared USDA National Nutrient Database for Standard [...] [Meta]
- 
 The COVID Tracking Project - The COVID Tracking Project collects and publishes the most [...] [Meta]
- 
 EHDP Large Health Data Sets [Meta]




-  GDC - GDC supports several cancer genome programs for CCG, TCGA, TARGET etc. [Meta]
-  Gapminder World demographic databases [Meta]
-  MeSH, the vocabulary thesaurus used for indexing articles for PubMed [Meta]
-  MeDAL - A large medical text dataset curated for abbreviation disambiguation - Medical [...] [Meta]
-  Medicare Coverage Database (MCD), U.S. [Meta]
-  Medicare Data Engine of medicare.gov Data [Meta]
-  Medicare Data File [Meta]
-  Nightingale Open Science [Meta]
-  Number of Ebola Cases and Deaths in Affected Countries (2014) [Meta]
-  Open-ODS (structure of the UK NHS) [Meta]
-  OpenPaymentsData, Healthcare financial relationship data [Meta]
-  PhysioBank Databases - A large and growing archive of physiological data. [Meta]
-  The Cancer Imaging Archive (TCIA) [Meta]
-  The Cancer Genome Atlas project (TCGA) [Meta]
-  World Health Organization Global Health Observatory [Meta]
-  Yahoo Knowledge Graph COVID-19 Datasets - The Yahoo Knowledge Graph team at Verizon Media is [...] [Meta]
-  Informatics for Integrating Biology and the Bedside [Meta]

ImageProcessing

-  10k US Adult Faces Database [Meta]

-  2GB of Photos of Cats [Meta]
-  Audience Unfiltered faces for gender and age classification [Meta]
-  Affective Image Classification [Meta]
-  Airborne Object Detection and Tracking - The Airborne Object Tracking (AOT) dataset is a [...] [Meta]
-  Animals with attributes [Meta]
-  CADDY Underwater Stereo-Vision Dataset of divers' hand gestures - Contains 10K stereo pair [...] [Meta]
-  Cytology Dataset – CCAgT: Images of Cervical Cells with AgNOR Stain Technique - Contains 9339 [...] [Meta]
-  Caltech Pedestrian Detection Benchmark [Meta]
-  Chars74K dataset - Character Recognition in Natural Images (both English and Kannada are available) [Meta]
-  Cube++ - 4890 raw 18-megapixel images, each containing a SpyderCube color target in their [...] [Meta]
-  Densely Annotated Video Driving Data Set - This data set consists of 28 video sequences of [...] [Meta]
-  Danbooru Tagged Anime Illustration Dataset - A large-scale anime image database with 3.33m+ [...] [Meta]

-  DukeMTMC Data Set - DukeMTMC aims to accelerate advances in multi-target multi-camera [...] [Meta]
-  ETH Entomological Collection (ETHEC) Fine Grained Butterfly (Lepidoptera) Images [Meta]
-  Face Recognition Benchmark [Meta]
-  Flickr: 32 Class Brand Logos [Meta]
-  GDXray - X-ray images for X-ray testing and Computer Vision [Meta]
-  HumanEva Dataset - The HumanEva-I dataset contains 7 calibrated video sequences (4 grayscale [...]) [Meta]
-  ImageNet (in WordNet hierarchy) [Meta]
-  Indoor Scene Recognition [Meta]
-  International Affective Picture System, UFL [Meta]
-  KITTI Vision Benchmark Suite [Meta]
-  Labeled Information Library of Alexandria - Biology and Conservation - Contains over 10 [...] [Meta]
-  MNIST database of handwritten digits, near 1 million examples [Meta]
-  Multi-View Region of Interest Prediction Dataset for Autonomous Driving - Contains 16 driving [...] [Meta]
-  Massive Visual Memory Stimuli, MIT [Meta]
-  Newspaper Navigator - This dataset consists of extracted visual content for 16,358,041 [...] [Meta]
-  Open Images From Google - Pictures with segmentation masks for 2.8 million object instances [...] [Meta]
-  RuFa - Contains images of text written in one of two Arabic fonts (Ruqaa and Nastaliq [...]) [Meta]
-  SUN database, MIT [Meta]
-  SVIRO Synthetic Vehicle Interior Rear Seat Occupancy - 25.000 synthetic scenery's across ten [...] [Meta]
-  Several Shape-from-Silhouette Datasets [Meta]
-  Stanford Dogs Dataset [Meta]
-  The Action Similarity Labeling (ASLAN) Challenge [Meta]
-  The Oxford-IIIT Pet Dataset [Meta]





-  Violent-Flows - Crowd Violence / Non-violence Database and benchmark [Meta]
-  Visual genome [Meta]
-  YouTube Faces Database [Meta]

MachineLearning

-  All-Age-Faces Dataset - Contains 13'322 Asian face images distributed across all ages (from 2 [...]) [Meta]
-  Audi Autonomous Driving Dataset - We have published the Audi Autonomous Driving Dataset [...]) [Meta]
-  B3FD - Facial age (and gender) estimation dataset with 375k images - The B3FD dataset is a [...]) [Meta]
-  Context-aware data sets from five domains [Meta]
-  Delve Datasets for classification and regression [Meta]
-  Discogs Monthly Data [Meta]
-  Fluorescent Neuronal Cells - By releasing this dataset, we aim at providing a new testbed for [...]) [Meta]
-  Free Music Archive [Meta]
-  IMDb Database [Meta]
-  Iranis - A Large-scale Dataset of Farsi/Arabic License Plate Characters [Meta]
-  Keel Repository for classification, regression and time series [Meta]
-  LLVIP - This dataset contains 30976 images, or 15488 pairs, most of which were taken at very [...]) [Meta]
-  Labeled Faces in the Wild (LFW) [Meta]
-  Lending Club Loan Data [Meta]
-  Machine Learning Data Set Repository [Meta]
-  Million Song Dataset [Meta]
-  More Song Datasets [Meta]
-  MovieLens Data Sets [Meta]
-  New Yorker caption contest ratings [Meta]
-  RDataMining - "R and Data Mining" ebook data [Meta]
-  Registered Meteorites on Earth [Meta]














-  Restaurants Health Score Data in San Francisco [Meta]
-  TikTok Dataset - More than 300 dance videos that capture a single person performing dance [...] [Meta]
-  UCI Machine Learning Repository [Meta]
-  Yahoo! Ratings and Classification Data [Meta]
-  YouTube-BoundingBoxes [Meta]
-  Youtube 8m [Meta]
-  eBay Online Auctions (2012) [Meta]








Museums

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












NaturalLanguage

-  Automatic Keyphrase Extraction [Meta]
-  The Big Bad NLP Database [Meta]
-  Blizzard Challenge Speech - The speech + text data comes from professional audiobooks [...] [Meta]
-  Blogger Corpus [Meta]
-  CLiPS Stylometry Investigation Corpus [Meta]
-  ClueWeb09 FACC [Meta]
-  ClueWeb12 FACC [Meta]
-  DBpedia - Structured data from Wikipedia [Meta]
-  Dirty Words - With millions of images in our library and billions of user-submitted keywords, [...] [Meta]









-  Flickr Personal Taxonomies [Meta]
-  Freebase of people, places, and things [Meta]
-  German Political Speeches Corpus - Collection of political speeches from the German [...] [Meta]
-  Google Books Ngrams (2.2TB) [Meta]
-  Google MC-AFP - Generated based on the public available Gigaword dataset using Paragraph Vectors [Meta]
-  Google Web 5gram (1TB, 2006) [Meta]
-  Gutenberg eBooks List [Meta]
-  Hansards text chunks of Canadian Parliament [Meta]
-  LJ Speech - Speech dataset consisting of 13,100 short audio clips of a single speaker reading [...] [Meta]
-  M-AILabs Speech - The M-AILABS Speech Dataset is the first large dataset that we are [...] [Meta]
-  Microsoft MACHine Reading COMprehension Dataset (or MS MARCO) [Meta]
-  Machine Comprehension Test (MCTest) of text from Microsoft Research [Meta]
-  Machine Translation of European languages [Meta]
-  Making Sense of Microposts 2013 - Concept Extraction [Meta]
-  Making Sense of Microposts 2016 - Named Entity rEcognition and Linking [Meta]
-  Multi-Domain Sentiment Dataset (version 2.0) [Meta]
-  No Language Left Behind (NLLB - 200vo) - Dataset based on Meta's metadata for mined bitext. [...] [Meta]
-  Noisy speech database for training speech enhancement algorithms and TTS models - Clean and [...] [Meta]
-  Open Multilingual Wordnet [Meta]
-  POS/NER/Chunk annotated data [Meta]
-  Personae Corpus [Meta]
-  SMS Spam Collection in English [Meta]
-  SaudiNewsNet Collection of Saudi Newspaper Articles (Arabic, 30K articles) [Meta]
-  Stanford Question Answering Dataset (SQuAD) [Meta]
-  USENET postings corpus of 2005~2011 [Meta]

-  Universal Dependencies [Meta]
-  Webhose - News/Blogs in multiple languages [Meta]
-  Wikidata - Wikipedia databases [Meta]
-  Wikipedia Links data - 40 Million Entities in Context [Meta]
-  WordNet databases and tools [Meta]
-  Wordbank - Open, de-identified database of vocabulary development from 84,138 children and [...] [Meta]
-  WorldTree Corpus of Explanation Graphs for Elementary Science Questions - a corpus of [...] [Meta]












Neuroscience

-  Allen Institute Datasets [Meta]
-  Brain Catalogue [Meta]
-  Brainomics [Meta]
-  CodeNeuro Datasets [Meta]
-  Collaborative Research in Computational Neuroscience (CRCNS) [Meta]
-  FCP-INDI [Meta]
-  Human Connectome Project [Meta]
-  NDAR [Meta]
-  NIMH Data Archive [Meta]
-  NeuroData [Meta]
-  NeuroMorpho - NeuroMorpho.Org is a centrally curated inventory of digitally reconstructed [...] [Meta]
-  Neuroelectro [Meta]
-  OASIS [Meta]
-  OpenNEURO [Meta]
-  OpenfMRI [Meta]
-  Study Forrest [Meta]
-  The Nencki-Symfonia EEG/ERP dataset - A high-density electroencephalography (EEG) dataset [...] [Meta]


Physics

-  CERN Open Data Portal [Meta]
-  Crystallography Open Database [Meta]
-  IceCube - South Pole Neutrino Observatory [Meta]
-  Ligo Open Science Center (LOSC) - Gravitational wave data from the LIGO Hanford and [...] [Meta]
-  NASA Exoplanet Archive [Meta]
-  NSSDC (NASA) data of 550 space spacecraft [Meta]
-  Quantum simulations of an electron in a two dimensional potential well - The data was [...] [Meta]
-  Sloan Digital Sky Survey (SDSS) - Mapping the Universe [Meta]



ProstateCancer

-  EOPC-DE-Early-Onset-Prostate-Cancer-Germany - Early Onset Prostate Cancer - Germany. [...] [Meta]
-  GENIE - Data from the Genomics Evidence Neoplasia Information Exchange (GENIE) project of the [...] [Meta]
-  Genomic-Hallmarks-Prostate-Adenocarcinoma-CPC-GENE - Comprehensive genomic profiling of 477 [...] [Meta]
-  MSK-IMPACT-Clinical-Sequencing-Cohort-MSKCC-Prostate-Cancer - Targeted sequencing of clinical [...] [Meta]
-  Metastatic-Prostate-Adenocarcinoma-MCTP - Comprehensive profiling of 61 prostate cancer [...] [Meta]
-  Metastatic-Prostate-Cancer-SU2CPCF-Dream-Team - Comprehensive analysis of 150 metastatic [...] [Meta]
-  NPCR-2001-2015 - Database from CDC's National Program of Cancer Registries (NPCR). The [...] [Meta]
-  NPCR-2005-2015 - Database from CDC's National Program of Cancer Registries (NPCR). The [...] [Meta]
-  NaF-Prostate - NaF Prostate is a collection of F-18 NaF positron emission tomography/computed [...] [Meta]
-  Neuroendocrine-Prostate-Cancer - Whole exome and RNA Seq data of castration resistant [...] [Meta]
-  PLCO-Prostate-Diagnostic-Procedures - The Prostate Diagnostic Procedures dataset (95,837 [...] [Meta]

-  PLCO-Prostate-Medical-Complications - The Prostate Medical Complications dataset (3,350 [...]) [Meta]
-  PLCO-Prostate-Screening-Abnormalities - The Prostate Screening Abnormalities dataset (10,527 [...]) [Meta]
-  PLCO-Prostate-Screening - The Prostate Screening dataset (177,315 records, 35,875 subjects, [...]) [Meta]
-  PLCO-Prostate-Treatments - The Prostate Treatments dataset (13,409 records, 7,614 subjects, [...]) [Meta]
-  PLCO-Prostate - The Prostate dataset is a comprehensive dataset that contains nearly all the [...]) [Meta]
-  PRAD-CA-Prostate-Adenocarcinoma-Canada - Prostate Adenocarcinoma - Canada. Collected by the [...]) [Meta]
-  PRAD-FR-Prostate-Adenocarcinoma-France - Prostate Adenocarcinoma - France. Collected by ten [...]) [Meta]
-  PRAD-UK-Prostate-Adenocarcinoma-United-Kingdom - Prostate Adenocarcinoma - United Kingdom. [...]) [Meta]
-  PROSTATEx-Challenge - Retrospective set of prostate MR studies. All studies included [...]) [Meta]
-  Prostate-3T - The Prostate-3T project provided imaging data to TCIA as part of an ISBI [...]) [Meta]
-  Prostate-Adenocarcinoma-Broad-Cornell-2012 - Comprehensive profiling of 112 prostate cancer [...]) [Meta]
-  Prostate-Adenocarcinoma-Broad-Cornell-2013 - Comprehensive profiling of 57 prostate cancer [...]) [Meta]
-  Prostate-Adenocarcinoma-CNA-study-MSKCC - Copy-number profiling of 103 primary prostate [...]) [Meta]
-  Prostate-Adenocarcinoma-Fred-Hutchinson-CRC - Comprehensive profiling of prostate cancer [...]) [Meta]
-  Prostate Adenocarcinoma (MSKCC/DFCI) - Whole Exome Sequencing of 1013 prostate cancer samples. [Meta]
-  Prostate-Adenocarcinoma-MSKCC - MSKCC Prostate Oncogenome Project. 181 primary, 37 metastatic [...]) [Meta]
-  Prostate-Adenocarcinoma-Organoids-MSKCC - Exome profiling of prostate cancer samples and [...]) [Meta]
-  Prostate-Adenocarcinoma-Sun-Lab - Whole-genome and Transcriptome Sequencing of 65 Prostate [...]) [Meta]

-  Prostate-Adenocarcinoma-TCGA-PanCancer-Atlas - Comprehensive TCGA PanCanAtlas data from 11k [...] [Meta]
-  Prostate-Adenocarcinoma-TCGA - Integrated profiling of 333 primary prostate adenocarcinoma samples. [Meta]
-  Prostate-Diagnosis - PCa T1- and T2-weighted magnetic resonance images (MRIs) were acquired [...] [Meta]
-  Prostate-Fused-MRI-Pathology - The Prostate Fused-MRI-Pathology collection is a combination [...] [Meta]
-  Prostate-MRI - The Prostate-MRI collection of prostate Magnetic Resonance Images (MRIs) was [...] [Meta]
-  Prostate-R - The R package 'ElemStatLearn' contains a prostate cancer dataset from Stamey et [...] [Meta]
-  QIN-PROSTATE-Repeatability - The QIN-PROSTATE-Repeatability dataset is a dataset with [...] [Meta]
-  QIN-PROSTATE - The QIN PROSTATE collection of the Quantitative Imaging Network (QIN) contains [...] [Meta]
-  SEER-YR1973_2015.SEER9 - The SEER November 2017 Research Data files from nine SEER registries [...] [Meta]
-  SEER-YR1992_2015.SJ_LA_RG_AK - The SEER November 2017 Research Data files from the San Jose- [...] [Meta]
-  SEER-YR2000_2015.CA_KY_LO_NJ_GA - The SEER November 2017 Research Data files from the Greater [...] [Meta]
-  SEER-YR2000_2015.CA_KY_LO_NJ_GA - The July - December 2005 diagnoses for Louisiana from their [...] [Meta]
-  TCGA-PRAD-US - TCGA Prostate Adenocarcinoma (499 samples). [Meta]

Psychology+Cognition

-  OSU Cognitive Modeling Repository Datasets [Meta]
-  Open Cognitive Science Data - Publicly available behavioral datasets from across cognitive [...] [Meta]











PublicDomains

-  Aply Open Realtime Data [Meta]

-  Amazon [Meta]
-  Archive.org Datasets [Meta]
-  Archive-it from Internet Archive [Meta]
-  CMU JASA data archive [Meta]
-  CMU StatLab collections [Meta]
-  Data.World [Meta]
-  Data360 [Meta]
-  Enigma Public [Meta]
-  Google [Meta]
-  Grand Comics Database - The Grand Comics Database (GCD) is a nonprofit, internet-based [...] [Meta]
-  Infochimps [Meta]
-  KD Nuggets Data Collections [Meta]
-  Microsoft Azure Data Market Free DataSets [Meta]
-  Microsoft Data Science for Research [Meta]
-  Microsoft Research Open Data [Meta]
-  Open Library Data Dumps [Meta]
-  Reddit Datasets [Meta]
-  RevolutionAnalytics Collection [Meta]
-  Sample R data sets [Meta]
-  Stack Overflow Annual Developer Survey - Annual developer surveys full data sets from 2011 [...] [Meta]
-  StatSci.org [Meta]
-  Stats4Stem R data sets (archived) [Meta]
-  The Washington Post List [Meta]
-  UCLA SOCR data collection [Meta]
-  UFO Reports [Meta]
-  Wikileaks 911 pager intercepts [Meta]
-  Yahoo Webscope [Meta]

SearchEngines

-  Academic Torrents of data sharing from UMB [Meta]



-  Base dos Dados - Data Basis: Open Data Repository for Brazil [Meta]
-  Datahub.io [Meta]
-  Domains Project - Sorted list of Internet domains [Meta]
-  Harvard Dataverse Network of scientific data [Meta]
-  ICPSR (UMICH) [Meta]
-  Institute of Education Sciences [Meta]
-  National Technical Reports Library [Meta]
-  Open Data Certificates (beta) [Meta]
-  OpenDataNetwork - A search engine of all Socrata powered data portals [Meta]
-  Statista.com - statistics and Studies [Meta]
-  Zenodo - An open dependable home for the long-tail of science [Meta]

Social Networks

-  2021 Portuguese Elections Twitter Dataset - 57M+ tweets, 1M+ users - This dataset contains [...] [Meta]
-  72 hours #gamergate Twitter Scrape [Meta]
-  CMU Enron Email of 150 users [Meta]
-  Cheng-Caverlee-Lee September 2009 - January 2010 Twitter Scrape [Meta]
-  China Biographical Database - The China Biographical Database is a freely accessible [...] [Meta]
-  Clubhouse Dataset [Meta]
-  A Twitter Dataset of 40+ million tweets related to COVID-19 - Due to the relevance of the [...] [Meta]
-  43k+ Donald Trump Twitter Screenshots - This archive contains screenshots of 43,475 Donald [...] [Meta]
-  EDRM Enron EMail of 151 users, hosted on S3 [Meta]
-  Facebook Data Scrape (2005) [Meta]
-  Facebook Social Connectedness Index - We use an anonymized snapshot of all active Facebook [...] [Meta]
-  Facebook Social Networks from LAW (since 2007) [Meta]
-  Foursquare from UMN/Sarwat (2013) [Meta]
-  GitHub Collaboration Archive [Meta]

-  Google Scholar citation relations [Meta]
-  High-Resolution Contact Networks from Wearable Sensors [Meta]
-  Indie Map: social graph and crawl of top IndieWeb sites [Meta]
-  Mobile Social Networks from UMASS [Meta]
-  Network Twitter Data [Meta]
-  Reddit Comments [Meta]
-  Skytrax' Air Travel Reviews Dataset [Meta]
-  Social Twitter Data [Meta]
-  SourceForge.net Research Data [Meta]
-  The Reddit COVID dataset - This dataset attempts to capture the full extent of COVID-19 [...] [Meta]
-  Twitch Top Streamer's Data [Meta]
-  Twitter Data for Online Reputation Management [Meta]
-  Twitter Data for Sentiment Analysis [Meta]
-  Twitter Graph of entire Twitter site [Meta]
-  Twitter Scrape Calufa May 2011 [Meta]
-  UNIMI/LAW Social Network Datasets [Meta]
-  United States Congress Twitter Data - Daily datasets with tweets of 1100+ accounts associated [...] [Meta]
-  Yahoo! Graph and Social Data [Meta]
-  Youtube Video Social Graph in 2007,2008 [Meta]









SocialSciences

-  ACLED (Armed Conflict Location & Event Data Project) [Meta]
-  Authoritarian Ruling Elites Database - The Authoritarian Ruling Elites Database (ARED) is a [...] [Meta]
-  Canadian Legal Information Institute [Meta]
-  Center for Systemic Peace Datasets - Conflict Trends, Polities, State Fragility, etc [Meta]
-  Correlates of War Project [Meta]
-  Cryptome Conspiracy Theory Items [Meta]
-  Datacards [Meta]

-  European Social Survey [Meta]
-  FBI Hate Crime 2013 - aggregated data [Meta]
-  Fragile States Index [Meta]
-  GDELT Global Events Database [Meta]
-  General Social Survey (GSS) since 1972 [Meta]
-  German Social Survey [Meta]
-  Global Religious Futures Project [Meta]
-  Gun Violence Data - A comprehensive, accessible database that contains records of over 260k [...] [Meta]
-  Humanitarian Data Exchange [Meta]
-  INFORM Index for Risk Management [Meta]
-  Institute for Demographic Studies [Meta]
-  International Networks Archive [Meta]
-  International Social Survey Program ISSP [Meta]
-  International Studies Compendium Project [Meta]
-  James McGuire Cross National Data [Meta]
-  MIT Reality Mining Dataset [Meta]
-  MacroData Guide by Norsk samfunnsvitenskapelig datatjeneste [Meta]
-  Mass Mobilization Data Project - The Mass Mobilization (MM) data are an effort to understand [...] [Meta]
-  Microsoft Academic Knowledge Graph - The Microsoft Academic Knowledge Graph is a large RDF [...] [Meta]
-  Minnesota Population Center [Meta]
-  Notre Dame Global Adaptation Index (ND-GAIN) [Meta]
-  Open Crime and Policing Data in England, Wales and Northern Ireland [Meta]
-  OpenSanctions - A global database of persons and companies of political, criminal, or [...] [Meta]
-  Paul Hensel General International Data Page [Meta]
-  PewResearch Internet Survey Project [Meta]
-  PewResearch Society Data Collection [Meta]
-  Political Polarity Data [Meta]
-  StackExchange Data Explorer [Meta]
-  Terrorism Research and Analysis Consortium [Meta]

-  Texas Inmates Executed Since 1984 [Meta]
-  Titanic Survival Data Set [Meta]
-  UCB's Archive of Social Science Data (D-Lab) [Meta]
-  UCLA Social Sciences Data Archive [Meta]
-  UN Civil Society Database [Meta]
-  UPJOHN for Labor Employment Research [Meta]
-  Universities Worldwide [Meta]
-  Uppsala Conflict Data Program [Meta]
-  World Bank Open Data [Meta]
-  World Inequality Database - The World Inequality Database (WID.world) aims to provide open [...] [Meta]
-  WorldPop project - Worldwide human population distributions [Meta]








Software

-  FLOSSmole data about free, libre, and open source software development [Meta]
-  GHTorrent - Scalable, queryable, offline mirror of data offered through the GitHub REST API. [Meta]
-  Libraries.io Open Source Repository and Dependency Metadata [Meta]
-  Public Git Archive - a Big Code dataset for all – dataset of 182,014 top-bookmarked Git [...] [Meta]
-  Code duplicates - 2k Java file and 600 Java function pairs labeled as similar or different by [...] [Meta]
-  Commit messages - 1.3 billion GitHub commit messages till March 2019 [Meta]
-  Pull Request review comments - 25.3 million GitHub PR review comments since January 2015 till [...] [Meta]
-  Source Code Identifiers - 41.7 million distinct splittable identifiers collected from 182,014 [...] [Meta]

Sports

-  American Ninja Warrior Obstacles - Contains every obstacle in the history of American Ninja [...] [Meta]
-  Betfair Historical Exchange Data [Meta]
-  Cricsheet Matches (cricket) [Meta]
-  Equity in Athletics - The Equity in Athletics Data Analysis Cutting Tool is brought to you by [...] [Meta]
-  Ergast Formula 1, from 1950 up to date (API) [Meta]
-  Football/Soccer resources (data and APIs) [Meta]
-  Lahman's Baseball Database [Meta]
-  NFL play-by-play data - NFL play-by-play data sourced from: [...] [Meta]
-  Pinhooker: Thoroughbred Bloodstock Sale Data [Meta]
-  Pro Kabadi season 1 to 7 - Pro Kabadi League is a professional-level Kabaddi league in India. [...] [Meta]
-  Retrosheet Baseball Statistics [Meta]
-  Tennis database of rankings, results, and stats for ATP [Meta]
-  Tennis database of rankings, results, and stats for WTA [Meta]
-  Transfermarkt Datasets - Clean, structured and automatically updated football (soccer) data [...] [Meta]
-  USA Soccer Teams and Locations - USA soccer teams and locations. MLS, NWSL, and USL [...] [Meta]




TimeSeries

-  3W dataset - To the best of its authors' knowledge, this is the first realistic and public [...] [Meta]
-  Databanks International Cross National Time Series Data Archive [Meta]
-  Hard Drive Failure Rates [Meta]
-  Heart Rate Time Series from MIT [Meta]
-  Time Series Data Library (TSDL) from MU [Meta]
-  Turing Change Point Dataset - Contains 42 annotated time series collected for the development [...] [Meta]
-  UC Riverside Time Series Dataset [Meta]

Transportation

-  Airlines OD Data 1987-2008 [Meta]
-  Ford GoBike Data (formerly Bay Area Bike Share Data) [Meta]
-  Bike Share Systems (BSS) collection [Meta]
-  Dutch Traffic Information [Meta]
-  GeoLife GPS Trajectory from Microsoft Research [Meta]
-  German train system by Deutsche Bahn [Meta]
-  Hubway Million Rides in MA [Meta]
-  Montreal BIXI Bike Share [Meta]
-  NYC Taxi Trip Data 2009- [Meta]
-  NYC Taxi Trip Data 2013 (FOIA/FOILed) [Meta]
-  NYC Uber trip data April 2014 to September 2014 [Meta]
-  Open Traffic collection [Meta]
-  OpenFlights - airport, airline and route data [Meta]
-  Philadelphia Bike Share Stations (JSON) [Meta]
-  Plane Crash Database, since 1920 [Meta]
-  RITA Airline On-Time Performance data [Meta]
-  RITA/BTS transport data collection (TranStat) [Meta]
-  Renfe (Spanish National Railway Network) dataset [Meta]
-  Toronto Bike Share Stations (JSON and GBFS files) [Meta]
-  Transport for London (TFL) [Meta]
-  Travel Tracker Survey (TTS) for Chicago [Meta]
-  U.S. Bureau of Transportation Statistics (BTS) [Meta]
-  U.S. Domestic Flights 1990 to 2009 [Meta]
-  U.S. Freight Analysis Framework since 2007 [Meta]
-  U.S. National Highway Traffic Safety Administration - Fatalities since 1975 - Contains CSV [...] [Meta]

eSports

-  CS:GO Competitive Matchmaking Data - In this data set we have data about the CSGO matchmaking [...] [Meta]
-  FIFA-2021 Complete Player Dataset [Meta]
-  OpenDota data dump [Meta]

Complementary Collections

- [Data Packaged Core Datasets](#)
- [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](#)
- [CVonline: Image Databases](#)
- [InnoTrek: Leveraging open data to understand urban lives](#)
- [CV Papers: CV Datasets on the web](#)