

Как и где искать открытые данные?

Ирина Радченко
кандидат технических наук, доцент
iradche@gmail.com

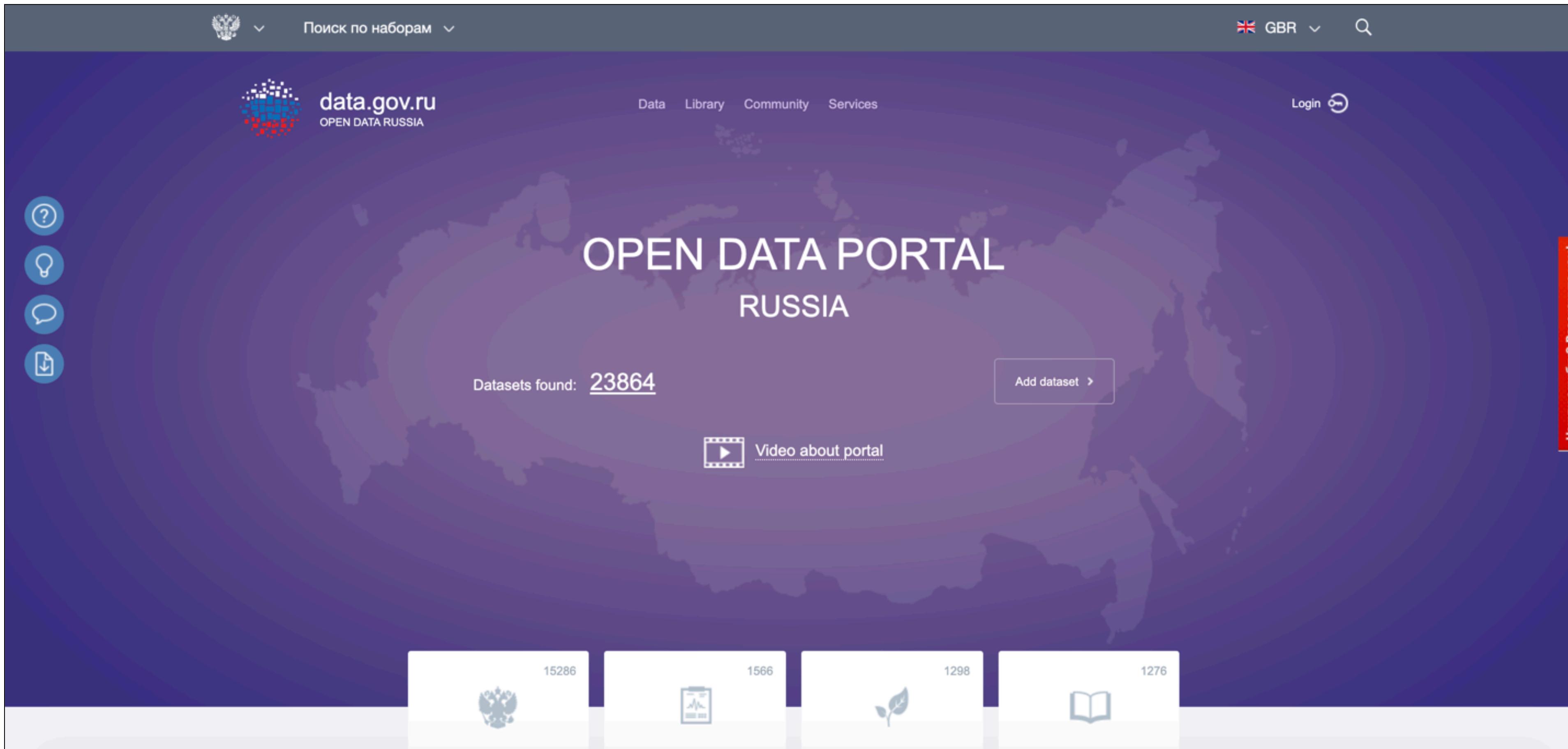
мастер-класс для Skills Lab
24 сентября 2020

Что такое открытые данные?

Открытые данные:

- машиночитаемость
- открытость

Портал российских открытых данных



<https://data.gov.ru/>

Федеральная служба государственной статистики

Федеральная служба
государственной статистики

f t v o Введите свой запрос Eng

О Росстате Статистика Публикации Респондентам Пресс-служба

Главная страница / Статистика В ИЗБРАННОЕ

Статистика

- Официальная статистика
- Переписи и обследования
- Методология и нормативно-справочная информация
- Интерактивные статистические сервисы

Анонсы

- Новости статистики
- Инфографика

Поделиться в соцсетях f t v o

<https://rosstat.gov.ru/statistic>

ЕМИСС государственная статистика

The screenshot shows the homepage of the EMISCS (State Statistics) website. At the top, there is a dark header bar with the logo of the Federal State Statistics Service (Rosstat) on the left, followed by the text "Е М И С С" and "ГОСУДАРСТВЕННАЯ СТАТИСТИКА". To the right of the logo are three navigation links: "Показатели" (Indicators), "Ведомства" (Departments), and a button labeled "Войти" (Log in). Below the header is a large dark grey banner with the text "Официальные статистические показатели" (Official statistical indicators) in white. To the right of this banner are two large white numbers: "7166" above "показателей" (indicators) and "65" above "ведомств" (departments). Below the banner is a search bar with a magnifying glass icon and the placeholder text "Поиск...". Underneath the search bar is a link labeled "Расширенный поиск" (Advanced search). In the bottom left corner of the main content area, there is a box showing "Посетителей в день" (Visitors per day) with the number "1190" and a bar chart. In the bottom right corner, there is a section titled "Популярные показатели" (Popular indicators) with a dropdown menu set to "За неделю" (For the week). This section lists four popular indicators with their respective counts: "Размер и состав денежных доходов и расходов населения" (87), "Экспорт отдельных товаров" (5), "Индексы потребительских цен на товары и услуги" (4), and "Средние потребительские цены (тарифы) на товары и услуги" (4).

Показатели

Ведомства

Войти

Официальные статистические показатели

Поиск...

Расширенный поиск

7166
показателей

65
ведомств

Посетителей в день

1190

Посетителей в неделю

15823

Популярные показатели

За неделю ▾

- 87 Размер и состав денежных доходов и расходов населения
- 5 Экспорт отдельных товаров
- 4 Индексы потребительских цен на товары и услуги
- 4 Средние потребительские цены (тарифы) на товары и услуги

<https://www.fedstat.ru/>

Проекты

Проект ГосЗатрат

Онлайн справочник вопросов и ответов по коронавирусу COVID-19 [ХОЧУ НАЙТИ ОТВЕТЫ >](#)

[RSS](#) [f](#) [vk](#) [Twitter](#) [g+](#) [in](#) [Русский](#) [English](#)

ClearSpending

[Home](#) All news Субсидии Contracts Budgets Customers Suppliers Purchases Аномалии Выгрузка данных NEW Forum About the project

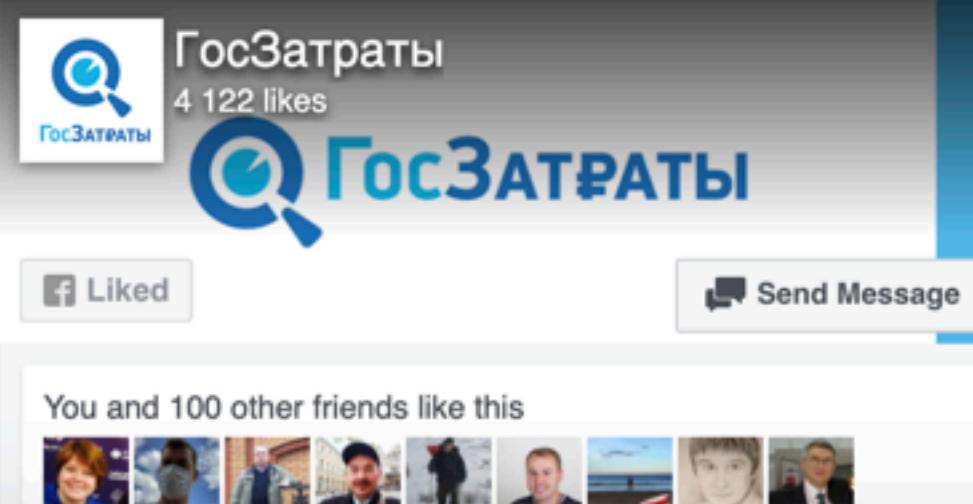
Contract search

Contract subject

Latest news



Мы в facebook



<https://clearspending.ru/>

Проект Хаб открытых данных

The screenshot shows the homepage of the OpenGovData.ru website. The header features a dark blue bar with the project logo (a green stylized 'H'), navigation links for 'Пакеты данных', 'Организации', 'Группы', 'О проекте', and a search bar. On the right, there are links for 'Войти' and 'Зарегистрироваться'. Below the header, a large search box is labeled 'Поиск данных' with placeholder text 'Например, окружающая среда'. A sidebar on the left displays statistics: 'Хаб открытых данных statistics' with values '8,1k' (datasets), '37' (organizations), and '87' (groups). Below these are links to 'наборы данных', 'организации', and 'группы'. The main content area contains two cards: one for 'Администрация муниципального района "Белгородский район" Белгородской области' and another for 'budget.gov.ru'.

Хаб открытых данных - это каталог и хранилище открытых данных для всех русскоязычных пользователей. Хаб создан и поддерживается АНО "Информационная культура" - <http://infoculture.ru>

Предложения, замечания и сообщения об ошибках, пожалуйста, направляйте на infoculture@infoculture.ru

Поддержите проект на <https://yasobe.ru/na/infoculture> или на наш Яндекс.Кошелек 410012648928680

Хаб открытых данных statistics

8,1k **37** **87**

наборы данных организации группы

Администрация муниципального района "Белгородский район" Белгородской области

Администрация муниципального района...

budget.gov.ru

Данные с сайта <http://budget.gov.ru>

Сведения об идентификационных кодах банков, включенных в предусмотренный стат...

<https://opengovdata.ru/>

Академический торрент

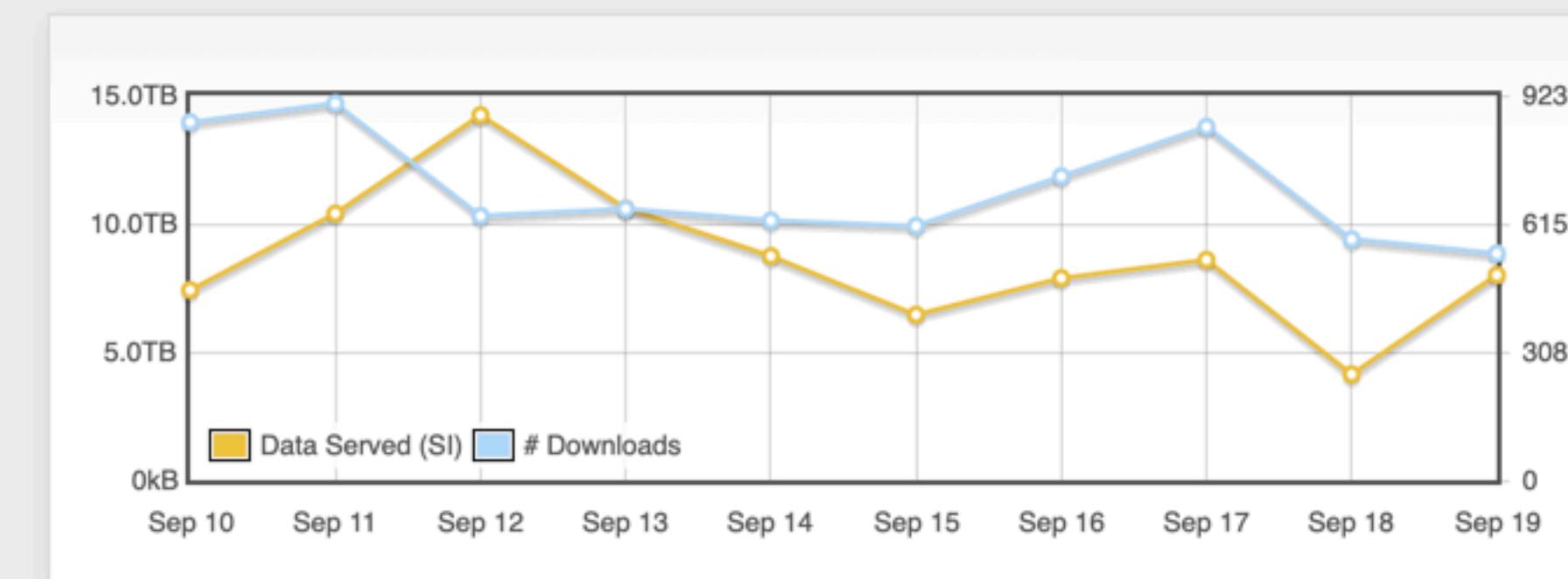
Academic Torrents Q Browse ▾ Upload About Donate Login

paper, author, or dataset Search

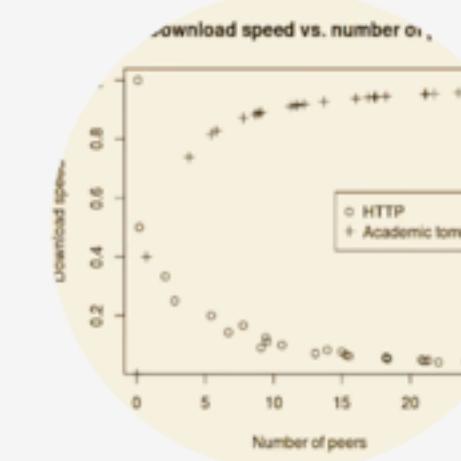
Making over 65TB of research data available!

We've designed a distributed system for sharing enormous datasets - for researchers, by researchers. The result is a scalable, secure, and fault-tolerant repository for data, with blazing fast download speeds. Contact us at contact@academictorrents.com.

View popular! Upload a dataset!



Date	Data Served (SI)	# Downloads
Sep 10	~7.5TB	~923
Sep 11	~10TB	~615
Sep 12	~14TB	~615
Sep 13	~10TB	~615
Sep 14	~9TB	~615
Sep 15	~6TB	~615
Sep 16	~8TB	~615
Sep 17	~9TB	~615
Sep 18	~4TB	~615
Sep 19	~8TB	~615



Download speed vs. number of peers

Number of peers

Download speed

Legend: ○ HTTP + Academic torrents

Accelerate your hosting for free with our academic BitTorrent infrastructure!

Distribute your public data globally for free to ensure it is available forever! ✉ Send Feedback

<https://academictorrents.com/>

Платформы распространения данных

- Открытые государственные данные на платформах
- Международные организации
- Платформы для дата-ученых
- Транснациональные корпорации

US Government's open data

The screenshot shows the official website for US Government's open data. At the top, there is a white header bar with the "DATA.GOV" logo on the left and navigation links for "DATA", "TOPICS", "RESOURCES", "STRATEGY", "DEVELOPERS", and "CONTACT". Below this is a large blue banner with the text "The home of the U.S. Government's open data". It explains that visitors can find data, tools, and resources for research, web and mobile applications, and more. It also provides a link to the Coronavirus.gov site for COVID-19 information. A "GET STARTED" button with a search bar containing "SEARCH OVER 225,075 DATASETS" is visible. A search input field contains the text "Federal Student Loan Program Data". Below this is a blue "HIGHLIGHTS" section featuring the title "Improving Access to Older Adult Health Data for Timely Use Amid COVID-19 and Beyond".

DATA TOPICS ▾ RESOURCES STRATEGY DEVELOPERS CONTACT

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and [more](#).

For information regarding the Coronavirus/COVID-19, please visit [Coronavirus.gov](#).

GET STARTED
SEARCH OVER 225,075 DATASETS

Federal Student Loan Program Data

HIGHLIGHTS

Improving Access to Older Adult Health Data for Timely Use
Amid COVID-19 and Beyond

<https://www.data.gov/>

UK Government's open data

data.gov.uk | Find open data

Publish your data Documentation Support

BETA This is a new service – your [feedback](#) will help us to improve it

Search results

Search data.gov.uk

Q

Filter by

55,214 results found

Sort by

Publisher

[Agricultural Land Classification detailed Post 1988 survey ALCR10794](#)

Best match ▾

Topic

Published by: Natural England
Last updated: 27 September 2016

Format

Survey name: Lower Pennington, Manor Farm (Hants Mins Om. Site 17)
Post 1988 Agricultural Land Classification (ALC) site survey data –
scanned original paper maps and survey reports for individual...

Open Government Licence

[Liverpool Primarily Industrial Sites](#)

<https://data.gov.uk/>

World Bank Open Data

 THE WORLD BANK | Data

This page in: English Español Français العربية 中文

New to this site? [Start Here](#)

Home DataBank Microdata Data Catalog ☰

World Bank Open Data

Free and open access to global development data

Search data e.g. GDP, population, Indonesia

Browse by [Country](#) or [Indicator](#)

MOST RECENT

[Online learning for Open Data in English and Spanish ↗](#)
Tim Herzog, Kenneth Moreno, Sep 16, 2020

[Tracking the socioeconomic impacts of the pandemic in Nigeria: Results from the first three rounds of the Nigeria COVID-19 National Longitudinal Phone Survey ↗](#)
Gbemisola Oseni, Amparo Palacios-Lopez, Kevin McGee, Akuffo Amankwah, Sep 16, 2020

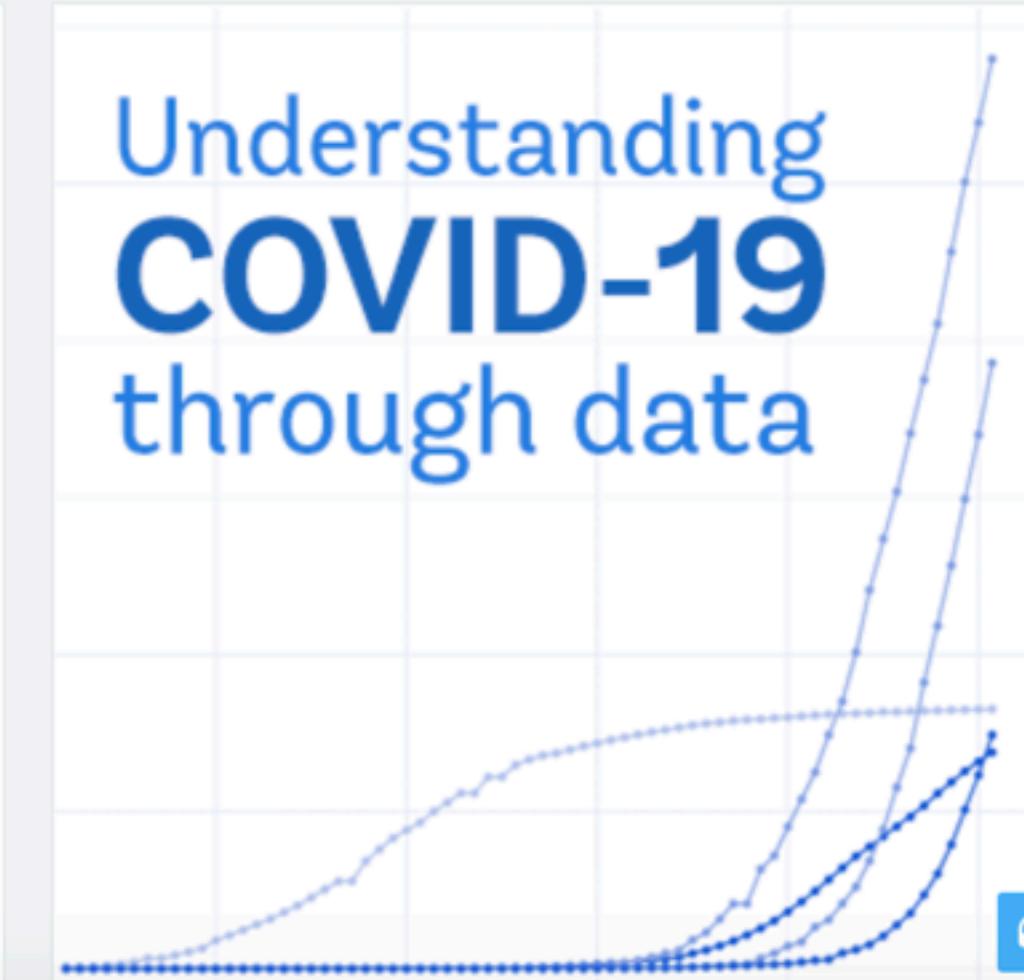
[Exploring links between democracy and](#)

WHAT YOU CAN LEARN WITH OPEN DATA

Extreme Poverty

The proportion of the world's population living in extreme poverty has dropped from 40% to 10%

Understanding COVID-19 through data



Help / Feedback

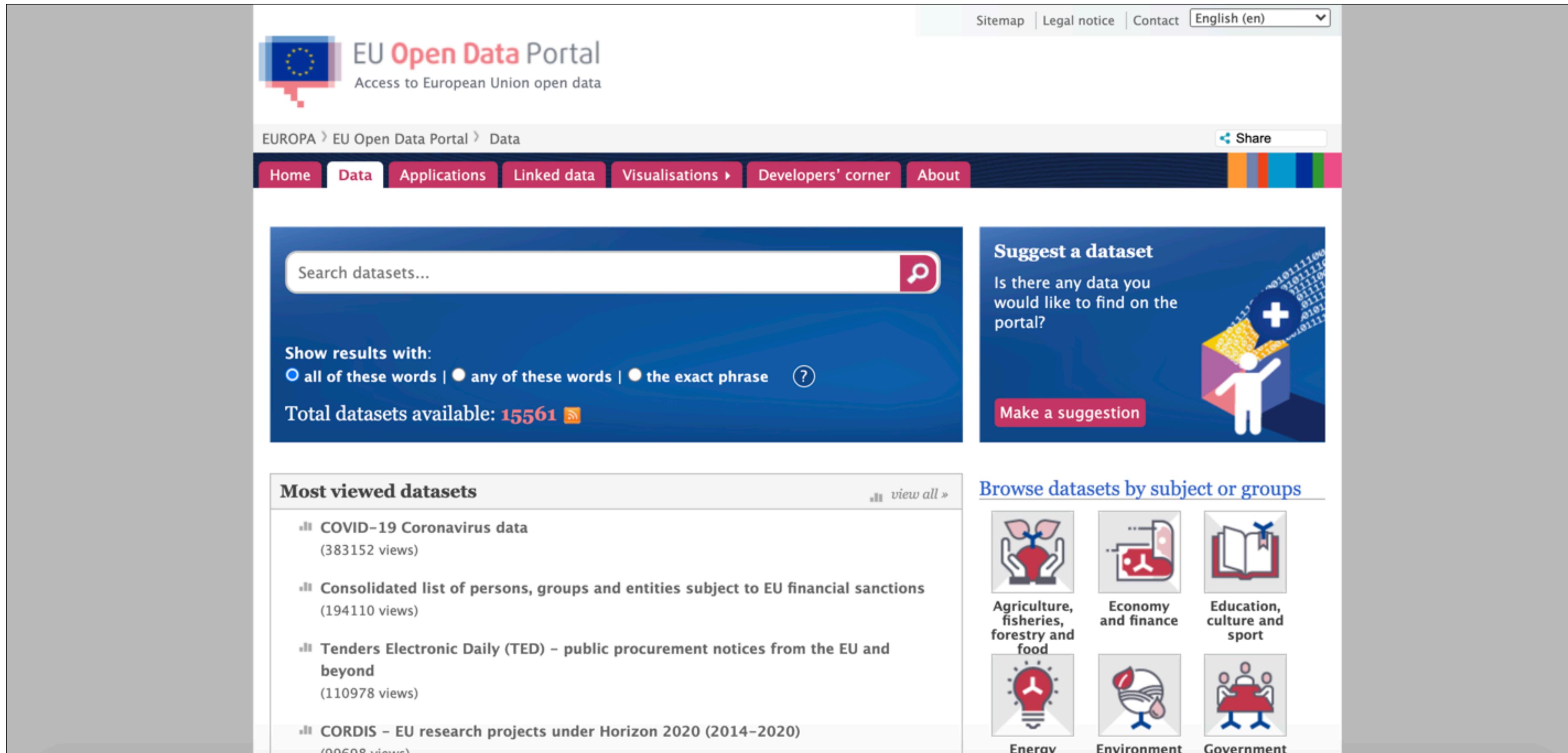
<https://data.worldbank.org/>

European Data portal: сборщик метаданных

The screenshot shows the European Data Portal's user interface. At the top, there is a navigation bar with links for Home, Data, Impact & Studies, Training, News & Events, and About. Below the navigation bar is a secondary menu with tabs for EU and international data, Country data (which is selected), SPARQL Search, Statistics, and Metadata Quality. On the left side, there is a map of Europe with a zoom control (+/-) and a search bar for filtering datasets by location. There is also a 'Settings' section with an 'Operator' dropdown set to 'AND' and a toggle switch for 'OR'. In the center, there is a search bar with placeholder text 'Search datasets...' and a 'Datasets Feed' button with a feed icon. A message indicates '1115343 datasets found'. Below this, a specific dataset is highlighted: 'Journées Européennes du patrimoine 2020 (Liste)'. The description for this dataset is: 'Programme des Journées européennes du patrimoine du 18 au 20 septembre 2020 sur le territoire de Grand Chambéry. Les données sont extraites de la plateforme d'information touristique APIDAE.' It includes download links for json, shp, and csv formats, and creation and update timestamps (20.09.2020 06:00). At the bottom right, there is a French flag and the text 'Plateforme ouverte des données publiques françaises'.

<https://www.europeandataportal.eu/data/datasets?locale=en&minScoring=0>

EU open data portal



The screenshot shows the EU Open Data Portal homepage. At the top, there is a navigation bar with links to Sitemap, Legal notice, Contact, and a language dropdown set to English (en). Below the navigation bar is the EU Open Data Portal logo and the tagline "Access to European Union open data". The main search area features a search bar with placeholder text "Search datasets..." and a magnifying glass icon. Below the search bar are options for search operators: "all of these words" (radio button), "any of these words" (radio button), "the exact phrase" (radio button), and a help link (?). A message indicates "Total datasets available: 15561" with a feed icon. To the right, there is a "Suggest a dataset" section with a "Make a suggestion" button and an illustration of a person pointing at a stack of data cubes. Below the search area, there are two sections: "Most viewed datasets" and "Browse datasets by subject or groups". The "Most viewed datasets" section lists four datasets with their titles and view counts: COVID-19 Coronavirus data (383152 views), Consolidated list of persons, groups and entities subject to EU financial sanctions (194110 views), Tenders Electronic Daily (TED) – public procurement notices from the EU and beyond (110978 views), and CORDIS – EU research projects under Horizon 2020 (2014–2020) (99608 views). The "Browse datasets by subject or groups" section shows six categories with corresponding icons: Agriculture, fisheries, forestry and food (hands holding a plant), Economy and finance (a graph), Education, culture and sport (an open book), Energy (a lightbulb), Environment (a recycling symbol), and Government (two people at a desk).

<https://data.europa.eu/euodp/en/data>

Machine Learning Repository

UCI 

Machine Learning Repository
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact
 Search
 Repository Web Google

[View ALL Data Set](#)

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 557 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:  In Collaboration With: 

Latest News:	Newest Data Sets:	Most Popular Data Sets (hits since 2007):
<p>09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!</p> <p>04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!</p> <p>03-01-2010: Note from donor regarding Netflix data</p> <p>10-16-2009: Two new data sets have been added.</p> <p>09-14-2009: Several data sets have been added.</p> <p>03-24-2008: New data sets have been added!</p> <p>06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope</p>	<p>07-22-2020:  Facebook Large Page-Page Network</p> <p>07-17-2020:  Amphibians</p> <p>07-12-2020:  Early stage diabetes risk prediction dataset.</p> <p>06-28-2020:  Taiwanese Bankruptcy Prediction</p> <p>06-20-2020:  South German Credit (UPDATE)</p> <p>06-17-2020:  BitcoinHeistRansomwareAddressDataset</p>	<p>3533819:  Iris</p> <p>1923463:  Adult</p> <p>1484282:  Wine</p> <p>1326060:  Breast Cancer Wisconsin (Diagnostic)</p> <p>1309344:  Heart Disease</p> <p>1303973:  Wine Quality</p>

Latest News:
09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!
04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010: Note from donor regarding Netflix data
10-16-2009: Two new data sets have been added.
09-14-2009: Several data sets have been added.
03-24-2008: New data sets have been added!
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: [ICU](#)
 Data Type: Multivariate, Time-Series

Newest Data Sets:
07-22-2020:  [Facebook Large Page-Page Network](#)
07-17-2020:  [Amphibians](#)
07-12-2020:  [Early stage diabetes risk prediction dataset.](#)
06-28-2020:  [Taiwanese Bankruptcy Prediction](#)
06-20-2020:  [South German Credit \(UPDATE\)](#)
06-17-2020:  [BitcoinHeistRansomwareAddressDataset](#)

Most Popular Data Sets (hits since 2007):
3533819:  [Iris](#)
1923463:  [Adult](#)
1484282:  [Wine](#)
1326060:  [Breast Cancer Wisconsin \(Diagnostic\)](#)
1309344:  [Heart Disease](#)
1303973:  [Wine Quality](#)

<https://archive.ics.uci.edu/ml/index.php>

Коллекция KDnuggets

The screenshot shows the KDnuggets website with a yellow header bar. The header includes the KDnuggets logo, social media links for Twitter, Facebook, and LinkedIn, and a search bar. Below the header is a navigation menu with links to Blog/News, Opinions, Tutorials, Top stories, Companies, Courses, Datasets, Education, Events (online), Jobs, Software, and Webinars. The main content area has a yellow header "Datasets for Data Mining, Data Science, and Machine Learning". It features social sharing buttons (Like 102, Share 102, Tweet, Share, 8.5K) and a "See also" section with links to Government data sites, Data APIs, and Google Dataset Search. A "Data repositories" section lists various datasets like Anacode Chinese Web Datastore, Appen Open Source Datasets, AssetMacro, Awesome Public Datasets, AWS Public Data Sets, and BigML. To the right, there's a "Latest News" sidebar with links to articles about Python projects, Simpson's Paradox, generative models, Coursera's Machine Learning course, consumer insights, and unpopular opinions. At the bottom, there are "Top Stories Last Week" and "Most Popular" sections.

Datasets for Data Mining, Data Science, and Machine Learning

Like 102 Share 102 Tweet Share 8.5K

See also

- Government, State, City, Local, public data sites and portals
- Data APIs, Hubs, Marketplaces, Platforms, and Search Engines.
- Google Dataset Search

Data repositories

- Anacode Chinese Web Datastore: a collection of crawled Chinese news and blogs in JSON format.
- Appen Open Source Datasets.
- AssetMacro, historical data of Macroeconomic Indicators and Market Data.
- Awesome Public Datasets on github, curated by caesar0301.
- AWS (Amazon Web Services) Public Data Sets, provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications.
- BigML: big list of public data sources

Latest News

- Automating Every Aspect of Your Python Project
- What is Simpson's Paradox and How to Automatically De...
- The Insiders' Guide to Generative and Discriminative ...
- Coursera's Machine Learning for Everyone Fulfils...
- How to Effectively Obtain Consumer Insights in a Data O...
- Unpopular Opinion – Data Scientists Should Be Mor...

Top Stories Last Week

Most Popular

- Free From MIT: Intro to

<https://www.kdnuggets.com/datasets/index.html>

Awesome Public Datasets

[awesomedata / awesome-public-datasets](https://github.com/awesomedata/awesome-public-datasets)

Code Issues Pull requests Actions Security Insights

master 2 branches 1 tag Go to file Add file Code

caesar0301 Update README from APD2: 25bef6fa73932fcffeffe9055f0a83c52a745f... fe7aeae 2 days ago 721 commits

Datasets Add titanic dataset 6 years ago

LICENSE Update license copyright info. 6 years ago

README.rst Update README from APD2: 25bef6fa73932fcffeffe9055f0a83c52a7... 2 days ago

README.rst

Awesome Public Datasets

awesome

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,

About
A topic-centric list of HQ open datasets.

awesomedataworld.slack.com
opendata aaron-swartz
awesome-public-datasets datasets

Readme
MIT License

Releases
1 tags

Packages
No packages published

Contributors 156

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

Time Series Classification

Time Series Classification Home Datasets Algorithms Results Researchers Code Bibliography UEA Papers ▾ About Us

Train Size

[Less than 100 \(49\)](#)
[100 to 500 \(85\)](#)
[Greater than 500 \(44\)](#)

Test Size

[Less than 300 \(84\)](#)
[300 to 1000 \(49\)](#)
[Greater than 1000 \(45\)](#)

Length

[Less than 300 \(91\)](#)
[300 to 700 \(40\)](#)
[Greater than 700 \(47\)](#)

Classes

[Less than 10 \(134\)](#)
[10 to 30 \(34\)](#)
[Greater than 30 \(10\)](#)

Type

[Device \(10\)](#)
[ECG \(9\)](#)
[Image \(34\)](#)
[Motion \(25\)](#)

Dataset listing

The univariate and multivariate classification problems are available in three formats: Weka ARFF, simple text files and sktime ts format. Weka does not allow for unequal length series, so the unequal length problems are all padded with missing values. ts format does allow for this feature.

[Univariate Weka formatted ARFF files and .txt files \(about 500 MB\).](#)

[Univariate sktime formatted ts files \(about 300 MB\).](#)

[Multivariate Weka formatted ARFF files \(and .txt files\) \(about 2 GB\).](#)

[sktime formatted ts files \(about 1.5 GB\).](#)

Lists of the data, including which are unequal length, can be found [here](#). Details on loading sktime data with the Python package are [here](#).

To store multivariate series in ARFF we take advantage of relational attributes. These are fairly unintuitive, so we have provided an overview of this and other basic features of loading data and building classifiers [here](#).

These files provide a simple list of the data characteristics for [univariate](#) and [multivariate](#) problems.

To see how accurate different classifiers are on these data see the [results page](#).

More information on the datasets is given below. Problems with variable length series are listed as length 0.

Search

Dataset	Train Size	Test Size	Length	No. of Classes	Type
AbnormalHeartbeat	303	303	3053	5	AUDIO
ACSF1	100	100	1460	10	DEVICE
Adiac	390	391	176	37	IMAGE

<http://timeseriesclassification.com/aboutus.php>

Платформы для дата-ученых

- Kaggle
- DrivenData

Платформа Kaggle

The screenshot shows the Kaggle datasets homepage. On the left, a sidebar menu titled "kaggle" includes options: Home, Compete, Data (which is selected and highlighted in blue), Notebooks, Discuss, Courses, Jobs, and More. A search bar at the top right contains the placeholder "Search". To the right of the search bar is a notification bell icon and a user profile picture. The main content area features a large title "Datasets" and a sub-section "Create Public Datasets" with a call-to-action button "Create Public Dataset". Below this, there's a search bar for "53,863 datasets" and a "Feedback" link. The dataset list is sorted by "Hottest" and includes a "Filter" button. The first dataset listed is "Health Insurance Cross Sell Prediction" by Anmol Kumar, which is 9 days old, 6 MB in size, and has 10.0 submissions. To the right, there's a section for "Open Tasks" showing a competition for "Health Insurance Cross Sell Prediction" with 14 submissions. At the bottom right, there's a "Competition" section for "Solar Power Generation" with 17 submissions.

datasets

Find and use datasets or complete tasks. [Learn more.](#)

+ New Dataset

Create Public Datasets

Open a dialogue, accept contributions, and get insights: improve your dataset by publishing it on Kaggle.

Create Public Dataset

Search 53,863 datasets

Feedback Filter

Sort by: Hottest

Public Your Datasets Favorites

Health Insurance Cross Sell Prediction

Anmol Kumar

9 days 6 MB 10.0 3 Files (CSV) 1 Task

Open Tasks

Health Insurance Cross Sell Prediction

14 Submissions · In Health Insurance Cross ...

Competition

17 Submissions · In Solar Power Generation

<https://www.kaggle.com/datasets>

Платформа DrivenData

The screenshot shows the homepage of the DrivenData platform. At the top, there is a navigation bar with links for COMPETITIONS, ABOUT, CAREERS, DRIVENDATA^{LABS}, BLOG, MY PROFILE, and LOG OUT. The logo 'DRIVENDATA' is on the left, featuring a colorful bar chart icon. Below the navigation, a large text area reads 'Data science competitions to build a better world'. Two buttons at the bottom left say 'I want to join a competition →' and 'I want to run a competition →'. To the right of the text is a large, colorful sunburst chart with many radial segments in various colors (green, yellow, orange, red, blue, purple) radiating from a central white circle.

<https://www.drivendata.org/>

NLP Database

The Big Bad NLP Database

For database updates follow on [Twitter](#) or [Medium](#)

Want to add a dataset, edit? [Edit](#)

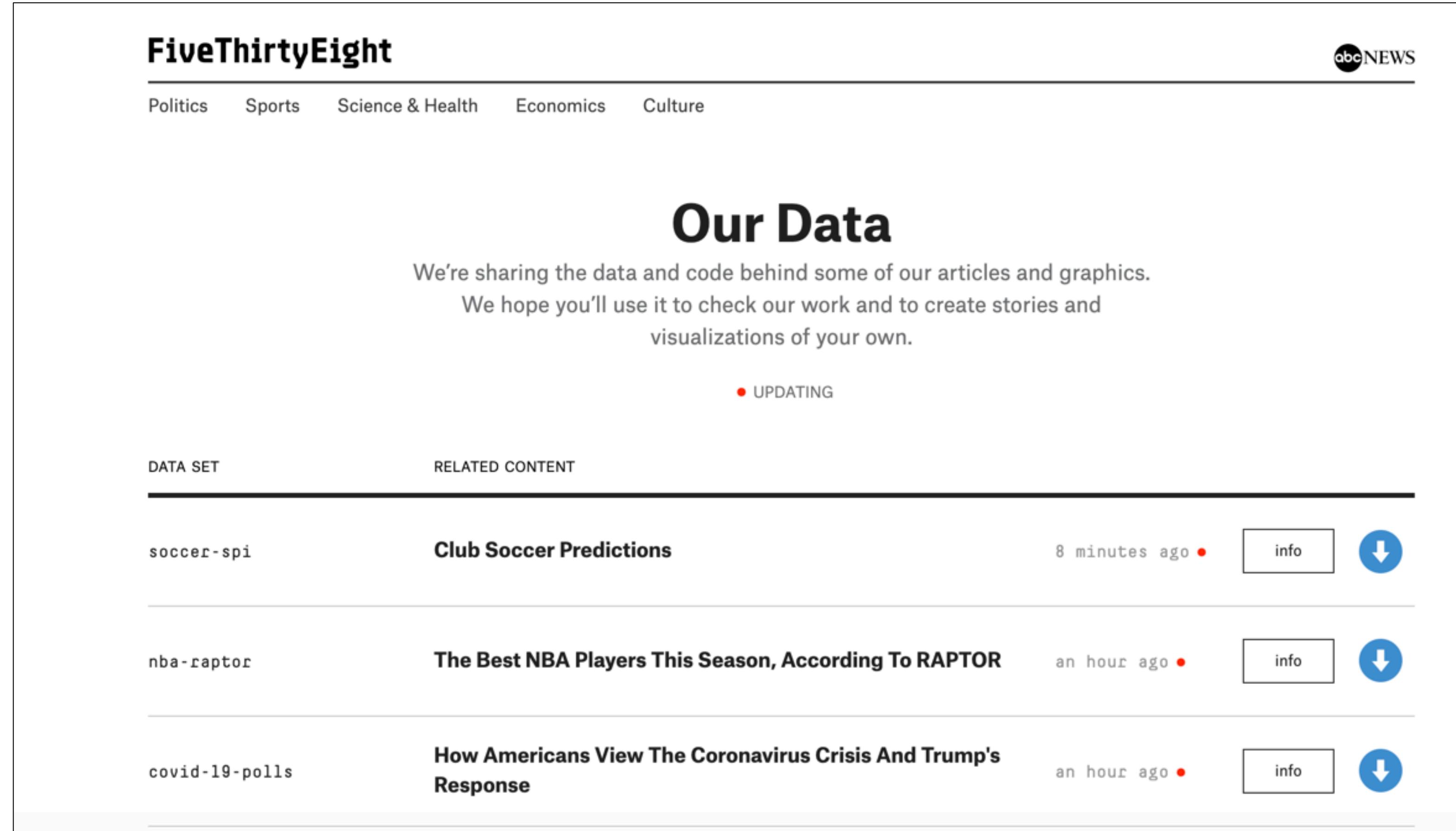
updated 09.19.20 584 datasets

search...

Dataset	Added	Lang	Description	Inst	Format	Task	Year	Creator	Source
Visual Genome	09.19.20	English	Dataset contains over 100K images where each image has an average of 21 objects, 18 attributes, and 18 pairwise relationships between objects.	100,000+	JSON	Visual Question Answering, Knowledge Base	2016	Krishna et al.	LINK PAPER
CraigsListBargain	09.19.20	English	Dataset contains 6,682 human-human dialogues where 2 agents negotiate the sale/purchase of an item.	6,682	JSON	Dialogue	2018	He et al.	LINK PAPER
A Multi-Turn, Multi-Domain Dialogue Dataset (KVRET)	09.19.20	English	Dataset contains 3,031 multi-turn dialogues in three distinct domains appropriate for an in-car assistant: calendar scheduling, weather information retrieval, and point-of-interest navigation.	3,301	JSON	Dialogue	2017	Eric et al.	LINK PAPER
CMU_ARCTIC	09.19.20	English	Dataset contains 1,150 utterances carefully selected from out-of-copyright texts from Project Gutenberg. The databases include US English male (bdl) and female (slt) speakers (both experienced voice talent) as well as other accented speakers.	1,150	WAV	Speech Recognition	2004	CMU	LINK PAPER
			Dataset includes recordings from twenty-four (24) non-native speakers of						

<https://datasets.quantumstat.com/>

Проект FiveThirtyEight

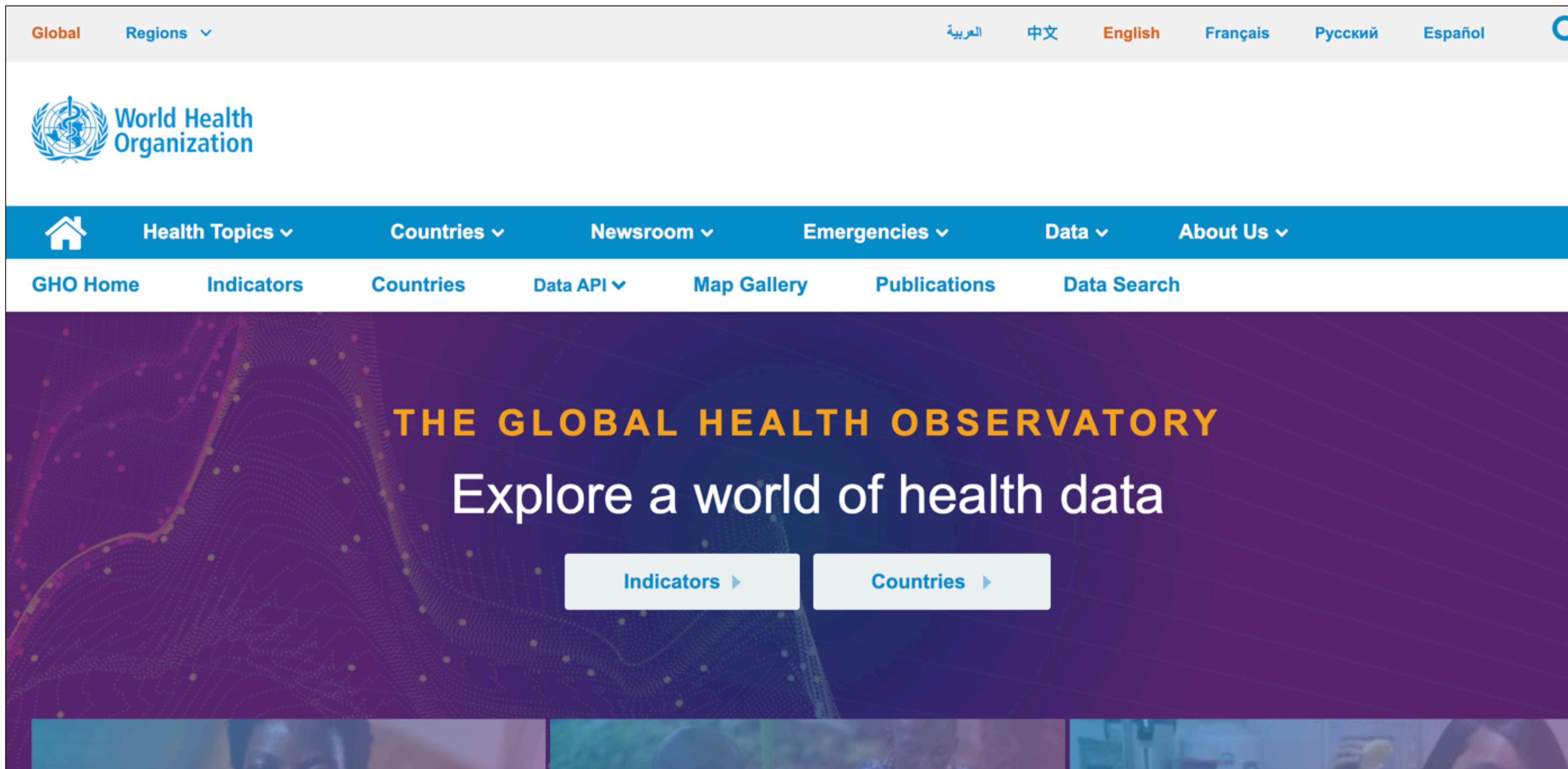


The screenshot shows the 'Our Data' section of the FiveThirtyEight website. At the top, there's a navigation bar with categories: Politics, Sports, Science & Health, Economics, and Culture. The ABC News logo is also present. Below the navigation, the title 'Our Data' is displayed in a large, bold font. A subtext explains: 'We're sharing the data and code behind some of our articles and graphics. We hope you'll use it to check our work and to create stories and visualizations of your own.' A note indicates that the data is 'UPDATING'. There are two tabs at the top of the data list: 'DATA SET' and 'RELATED CONTENT', with 'DATA SET' being the active tab. Three data sets are listed:

DATA SET	CONTENT	LAST UPDATED	INFO	DOWNLOAD
soccer-spi	Club Soccer Predictions	8 minutes ago	info	download
nba-raptor	The Best NBA Players This Season, According To RAPTOR	an hour ago	info	download
covid-19-polls	How Americans View The Coronavirus Crisis And Trump's Response	an hour ago	info	download

<https://data.fivethirtyeight.com/>

Медицинские данные: Global Health Data



<https://www.who.int/data/gho>

Vanderbilt Biostatistics

You are here: Vanderbilt Biostatistics Wiki > Main Web > DataSets (26 Feb 2020, FrankHarrell)

[Edit](#) [Attach](#)

Main

[Department Home Page](#)

[Biostatistics Graduate Program](#)

[Vanderbilt University Medical Center](#)

Main Web

Main Web Home

Search

Recent Changes

Changes

Topic list

Biostatistics Webs

Archive

Main

Sandbox

System

Datasets

Most of the datasets on this page are in the S dumpdata and R compressed save() file formats. Some are available in Excel and ASCII (.csv) formats and Stata (.dta). Methods for retrieving and importing datasets may be found [here](#). If you need one of the datasets we maintain converted to a non-S format please e-mail <mailto:charles.dupont@vanderbilt.edu> to make a request.

For R users of the prostate dataset, put `library(chron)` into effect to handle date variables. A simpler approach is to just convert the one date variable to the built-in R format by running the command `prostate$date <- as.Date(prostate$date)`.

Permission is granted to anyone wishing to use the data sets provided here. Please reference the original paper which, for most data sets, is given in our notes linked below, and note "Data obtained from <http://biostat.mc.vanderbilt.edu/DataSets>".

Description	R	S-Plus (.sdd)	Excel	ASCII	contents()
Stata (.dta)					
Meningitis dataset					
abm.html	abm.sav	abm.dta	abm.xls	NA	Cabm.html
Cardiac catheterization diagnostic data					
acath.html	acath.sav	acath.dta	acath.xls.zip	NA	Cacath.html
CRASH-2					
crash2.html	crash2.rda	NA	NA	NA	Ccrash2.html
WHO ARI Multicentre Study of clinical signs and etiologic agents					
Description	ari.sav ari_other.sav	NA	NA	ari.zip	ari.html
Rosner's estriol data					
NA	birth.estriol.sav	NA	NA	birth_estriol.csv	Cbirth.estriol.html
Boston neighborhood housing prices data					

Centers for Disease Control and Prevention

 Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.TM Data.CDC.gov

Home Data Catalog Developers Video Guides [Sign In](#)

Search

Categories

- Administrative
- Biomonitoring
- Disability & Health
- Environmental Health & Toxicology
- Foodborne, Waterborne, and Related Diseases

Show All...

View Types

- Calendars
- Charts

919 Results

Sort by Most Relevant

Featured Content

CDC COVID-19 Data Tracker

External Content



One-stop source for national and county-level cases, deaths, testing and other COVID-19 information.

COVID-19 Case Surveillance Public Use Data

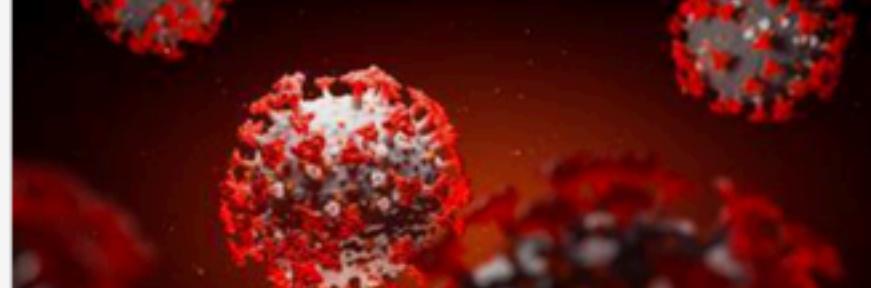
September 1, 2020 208K Views



The COVID-19 case surveillance system database includes patient-level data reported to U.S. states and autonomous reporting entities,...

COVIDView

External Content



A weekly surveillance summary of U.S. COVID-19 Activity.

<https://data.cdc.gov/browse>

Биологические данные

1000 Genomes Project and AWS

The 1000 Genomes Project is an international collaboration which has established the most detailed catalogue of human genetic variation, including SNPs, structural variants, and their haplotype context. The final phase of the project sequenced more than 2500 individuals from 26 different populations around the world and produced an integrated set of phased haplotypes with more than 80 million variants for these individuals.

The Amazon mirror contains the complete data set from the project and the data can be found in the <s3://1000genomes> bucket in the us-east-1 AWS region.

For more information please look at <http://www.1000genomes.org>. If you have any questions about the data, please email info@1000genomes.org.

Accessing 1000 Genomes Data

AWS is making the 1000 Genomes Project data publicly available to the community free of charge. Public Data Sets on AWS provide a centralized repository of public data hosted on Amazon Simple Storage Service (Amazon S3). The data can be seamlessly accessed from AWS services such Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Elastic MapReduce (Amazon EMR), which provide organizations with the highly scalable compute resources needed to take advantage of these large data collections. AWS is storing the public data sets at no charge to the community. Researchers pay only for the additional AWS resources they need for further processing or analysis of the data. Learn more about [Public Data Sets on AWS](#).

The latest 1000 Genomes Project data is publicly available in the [1000genomes Amazon S3 bucket](#).

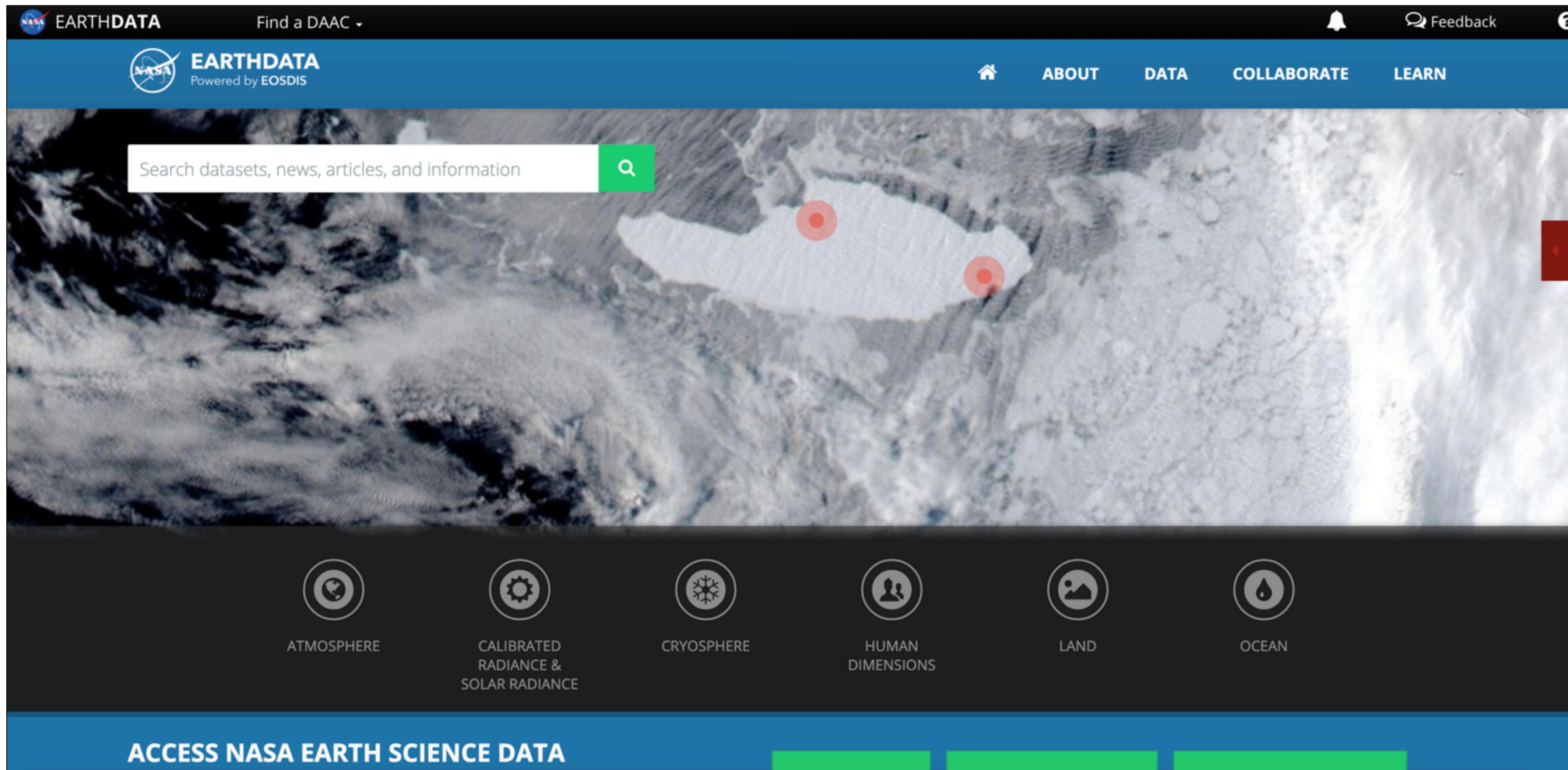
You can access the data via simple HTTP requests, or take advantage of the AWS SDKs in languages such as Ruby, Java, Python, .NET and PHP.

Analyzing 1000 Genomes Data

Researchers can use the Amazon EC2 utility computing service to dive into this data without the usual capital investment required to work with data at this scale. AWS also provides a number of [orchestration](#) and [automation](#) services to help teams make their research available to others to remix and reuse.

Making the data available via a bucket in Amazon S3 also means that customers can crunch the information using

NASA: Earth data



<https://earthdata.nasa.gov/>

NASA: Space data

The screenshot shows the PDS NASA Planetary Data System website. The header includes the PDS logo, a search bar labeled "Find a Node", and a "Data Search" button. Below the header is a navigation menu with links to HOME, DATA SEARCH, TOOLS, and DATA STANDARDS, along with sub-links for Data Search, Keyword Search, Data Set Status, and Data Releases. On the right side of the header is a link to "NASA Port". A sidebar on the left is titled "PDS Nodes" and lists various discipline nodes: Atmospheres (ATM), Geosciences (GEO), Cartography and Imaging Sciences (IMG), Navigational & Ancillary Information (NAIF), Planetary Plasma Interactions (PPI), Ring-Moon Systems (RMS), and Small Bodies (SBN). The main content area is titled "Data Search" and contains text about advanced search tools and contact information for PDS nodes and operators. It also provides links to Keyword Search and Data Set Status. Below this is a search interface with dropdown menus for "Search based on Target" and "Mission". The footer includes the USA.gov logo, links to Privacy / Copyright and Freedom of Information Act, the NASA logo, and contact information for the Webmaster (PDS Operator), NASA Official (Meagan Thompson), and Last updated date (July 2020). A "Need Help?" link is located on the right side of the footer.

<https://pds.nasa.gov/datasearch/data-search/>

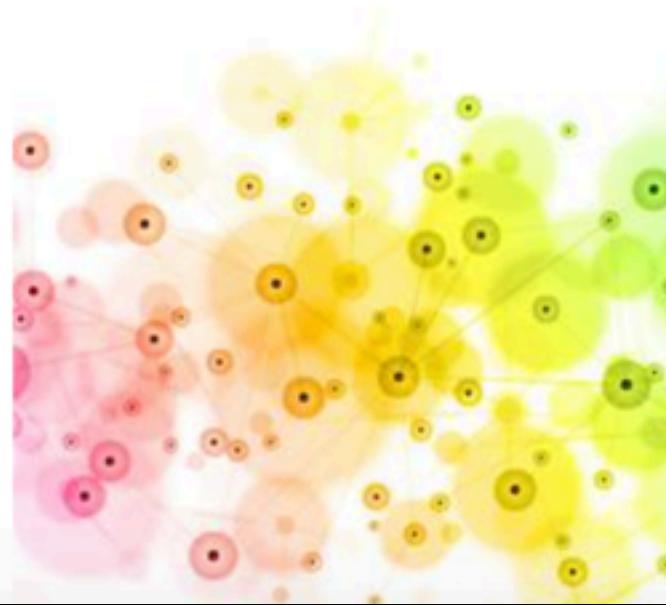
Маркетинговые данные

CHICAGO BOOTH Kilts Center
for Marketing

ABOUT US DATASETS GIVING  

OUR APPROACH SUPPORTING RESEARCH OUR COMMUNITY EVENTS

NIELSEN
DOMINICK'S
ERIM
BAYESM



**FUELING INSIGHTS WITH
DATA**

The provision of marketing data for academic purposes represents one of the key initiatives through which the Kilts Center maintains its ironclad commitment to innovation in marketing science and marketing practice. Academic researchers at Booth and beyond have implemented these data to study such vital topics as the measurement of [the effectiveness of drug](#)

<https://www.chicagobooth.edu/research/kilts/datasets>

Экономические данные



The screenshot shows the homepage of the National Bureau of Economic Research (NBER). The header features the NBER logo and navigation links for 'HOME PAGE', 'Sunday, September 20, 2020', 'log in', 'Search the NBER site', and 'Select one'. A sidebar on the left contains links to 'Working Papers & Publications', 'Activities', 'Meetings', 'NBER Videos', 'Themes in NBER Research', 'Data', 'People', and 'About'. The main content area is titled 'Data' and includes a note about available files and links to 'Resources for Economists', 'new economic releases', 'Google', and 'NBER papers'. It also lists 'SEE ALSO NBER LINKS for other data series.' Below this, a section titled 'Macro Data' lists various datasets with their sources:

Macro Data	Source
Official Business Cycle Dates	NBER
"The American Business Cycle: Continuity and Change" Historic Data Tables	Gordon
Experimental Coincident, Leading and Recession Indexes	Stock, Watson
Index of African Governance	Rotberg, Gisselquist
Jordà-Schularick-Taylor Macrohistory Database	Jordà, Schularick, Taylor
Penn-World Tables	Feenstra, Inklaar, Timmer
Cross-country Historical Adoption of Technology (CHAT) data	Comin, Hobijn
Occupational Wages around the World	Freeman, Oostendorp
Macro History Database	NBER
Savings, Investment, and Gold in 13 countries (1850-1945)	Jones, Obstfeld
Social Security Pension Reform in Europe	Feldstein, Siebert
	<small>Comin, Hobijn</small>

<https://data.nber.org/data/>

Сельскохозяйственные данные

The screenshot shows the 'Data sets' section of the Data Driven Farming Prize website. At the top, there is a navigation bar with links to 'Finalists', 'Partners', 'Judging Criteria', 'Data sets', 'Resources', and 'News & Blogs'. Below the navigation bar, the word 'Data sets' is prominently displayed in a large green font. Underneath it, a sub-section title reads 'Entrants might find below data sets useful in developing their ideas.' A note below this states, 'The data sets have been organised in the following categories:' followed by a list of categories: 'Nepalese Data', 'Global Agriculture', 'Open Data Portals', and 'Food & Nutrition'. The main content area is divided into three columns, each containing a title and a brief description. The first column is titled 'Data Driven Farming GeoNode' and describes it as a platform for sharing, layering, and filtering geospatial data. The second column is titled 'Open Data Nepal' and describes it as a portal containing data on various sectors including geography, demography, economy, water & sanitation, and climate. The third column is titled 'Agriculture and Forest' and describes it as a source of Nepalese Data sets from the National Bureau of Statistics, which generates socio-economic statistics through censuses and surveys.

DATA DRIVEN FARMING PRIZE

Finalists Partners Judging Criteria Data sets Resources News & Blogs

Data sets

Entrants might find below data sets useful in developing their ideas.

The data sets have been organised in the following categories:

Nepalese Data, Global Agriculture, Open Data Portals, Food & Nutrition.

Data Driven Farming GeoNode Data Driven Farming GeoNode is a platform for sharing, layering, and filtering geospatial data. We have loaded and continue to update the platform with key datasets to enable	Open Data Nepal Nepalese Data sets The portal contains data on a variety of sectors, including but not limited to geography, demography, economy, water & sanitation, climate	Agriculture and Forest Nepalese Data sets National bureau of statistics – It generates socio-economic statistics mainly through the operation of censuses and surveys. Agricultural
--	--	--

<https://datadrivenfarming.challenges.org/data-sets-2/>

Pew Research Centre

The screenshot shows the Pew Research Center's website for Internet & Technology. At the top, there's a navigation bar with links for 'ABOUT', 'FOLLOW', 'MY ACCOUNT', and 'DONATE'. Below the header, the main navigation menu includes categories like 'HOME', 'U.S. POLITICS', 'MEDIA & NEWS', 'SOCIAL TRENDS', 'RELIGION', 'INTERNET & TECH' (which is currently selected), 'SCIENCE', 'HISPANICS', 'GLOBAL', and 'METHODS'. A secondary navigation bar below features 'PUBLICATIONS', 'TOPICS', 'PRESENTATIONS', 'DATASETS' (selected), 'INTERACTIVES', 'FACT SHEETS', and 'OUR EXPERTS'. The main content area is titled 'Datasets' and displays 'Displaying 1-10 of 143 results'. To the right, there's a 'REFINE YOUR RESULTS' sidebar with an 'Update' button and a 'DATE' section showing options for 'Past 6 Months (3)', 'Past 12 Months (4)', and 'Past 2 Years (7)'. There's also a dropdown for 'Years' and a 'RANGE OF YEARS' input field with an 'Update' button. On the left, there's a 'Log in' form with fields for 'Email Address *' and 'Password *', both marked with red asterisks. A note says 'By signing in to your account, you agree to our [Terms of Use](#)'. A 'LOGIN' button is at the bottom of the form.

<https://www.pewresearch.org/internet/datasets/>

MNIST database

THE MNIST DATABASE of handwritten digits

[Yann LeCun](#), Courant Institute, NYU

[Corinna Cortes](#), Google Labs, New York

[Christopher J.C. Burges](#), Microsoft Research, Redmond

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Four files are available on this site:

[train-images-idx3-ubyte.gz](#): training set images (9912422 bytes)
[train-labels-idx1-ubyte.gz](#): training set labels (28881 bytes)
[t10k-images-idx3-ubyte.gz](#): test set images (1648877 bytes)
[t10k-labels-idx1-ubyte.gz](#): test set labels (4542 bytes)

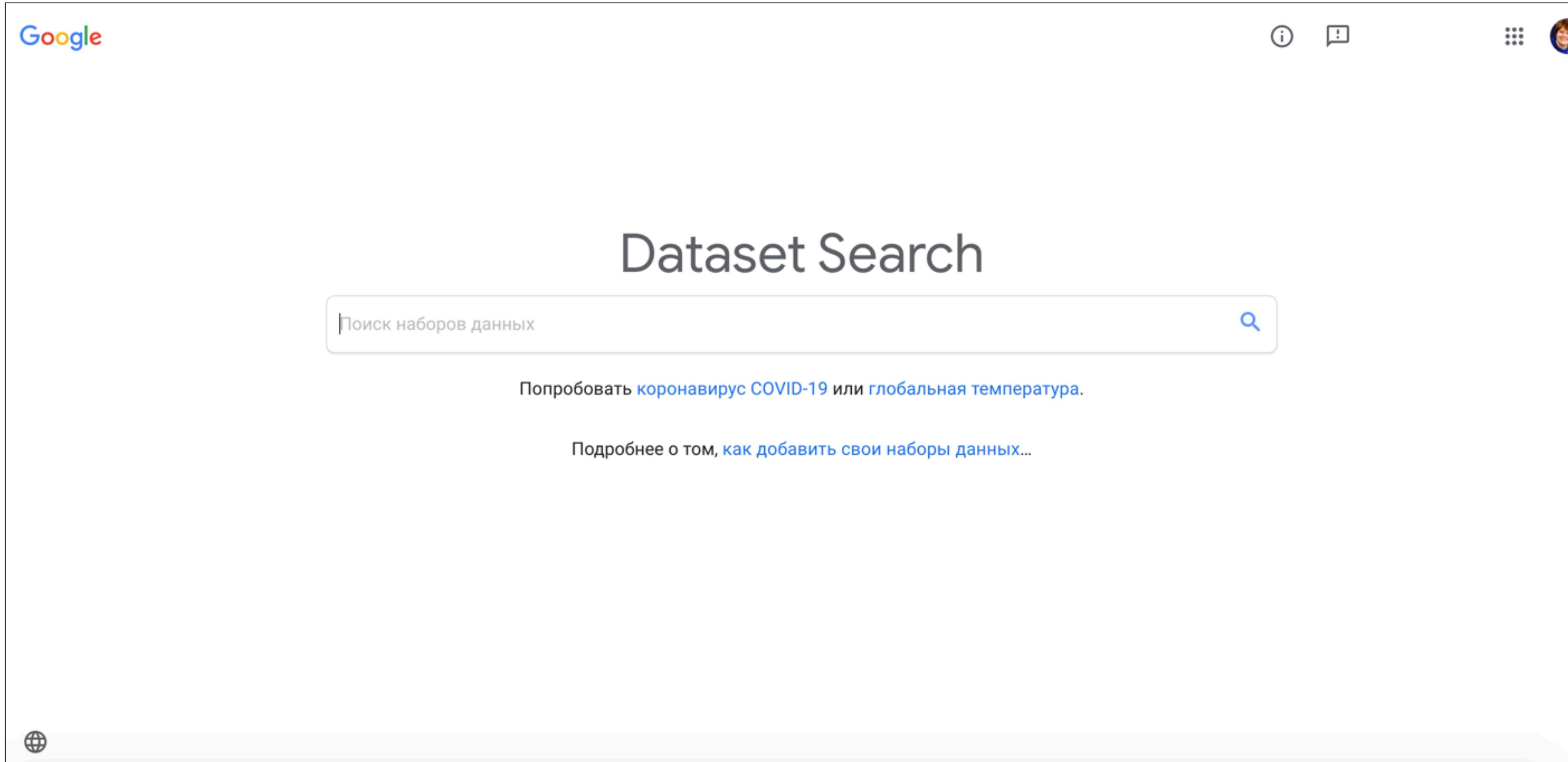
please note that your browser may uncompress these files without telling you. If the files you downloaded have a larger size than the above, they have been uncompressed by your browser. Simply rename them to remove the .gz extension. Some people have asked me "my application can't open your image files". These files are not in any standard image format. You have to write your own (very simple) program to read them. The file format is described at the bottom of this page.

The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. the images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field.

With some classification methods (particularly template-based methods, such as SVM and K-nearest neighbors), the error rate improves when the digits are centered by bounding box rather than center of mass. If you do this kind of pre-processing, you should report it in your publications.

The MNIST database was constructed from NIST's Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and SD-1 as their test set. However, SD-3 is much cleaner and easier to recognize than SD-1. The reason for this can be found on the fact that SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-school students. Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples. Therefore it was necessary to build a new database by mixing NIST's datasets.

Google: Проект Dataset Search



<https://datasetsearch.research.google.com/>

Google: BigQuery public datasets

The screenshot shows the Google Cloud BigQuery public datasets documentation page. The page has a header with navigation links like 'Google Cloud', 'Why Google', 'Solutions', 'Products', 'Pricing', 'Getting Started', a search bar, 'Docs', 'Support', 'Language', 'Console', and a user profile icon. Below the header, there's a secondary navigation bar with 'Data analytics products', 'Guides', 'Reference', 'Support', 'Resources' (which is underlined), 'Contact Sales', and a 'Get started for free' button. On the left, a sidebar for 'BigQuery' includes links for 'All resources', 'Pricing', 'Quotas and limits', 'Release notes', 'Public datasets' (which is highlighted with a blue background), 'Solution providers', and 'Service level agreement'. The main content area displays the title 'BigQuery public datasets' with a star rating of 5 stars and a 'Send feedback' button. It explains that the Cloud Public Datasets Program catalog is in Google Cloud Marketplace. A 'Go to Datasets in the Cloud Marketplace' button is present. The text describes what a public dataset is and how it can be accessed via various methods. To the right, a vertical sidebar titled 'Contents' lists sections such as 'Before you begin', 'Public dataset locations', 'Accessing public datasets in the BigQuery web UI', 'Other public datasets', 'Sharing a dataset with the public', 'Sample tables', 'Contact us', and 'What's next'. The URL at the bottom of the page is <https://cloud.google.com/bigquery/public-data/>.

<https://cloud.google.com/bigquery/public-data/>

Google: YouTube-8M Segments Dataset

The screenshot shows the official website for the YouTube-8M Segments Dataset. The header features a red bar with the "YouTube | 8M" logo on the left and navigation links for "Dataset", "Explore", "Download", "Workshop", and "About" on the right. The main content area has a white background. A large red section title "YouTube-8M Segments Dataset" is centered at the top. Below it is a descriptive paragraph about the dataset's purpose and collection process. Further down, there are three callout boxes with statistics: "237K Human-verified Segment Labels", "1000 Classes", and "5.0 Avg. Segments / Video". A callout box at the bottom provides more detail on the annotation process.

YouTube-8M Segments Dataset

The YouTube-8M Segments dataset is an extension of the YouTube-8M dataset with human-verified segment annotations. In addition to annotating videos, we would like to temporally localize the entities in the videos, i.e., find out when the entities occur.

We collected human-verified labels on about 237K segments on 1000 classes from the validation set of the YouTube-8M dataset. Each video will again come with time-localized frame-level features so classifier predictions can be made at segment-level granularity. We encourage researchers to leverage the large amount of noisy video-level labels in the training set to train models for temporal localization.

We are organizing a [Kaggle Challenge](#) and [The 3rd Workshop on YouTube-8M Large-Scale Video Understanding at ICCV 2019](#).

237K Human-verified Segment Labels	1000 Classes	5.0 Avg. Segments / Video
--	-----------------	---------------------------------

In addition to annotating the topical entity of the full-video, we want to understand when the entity occurs in videos. Given a 5-second segment and a query class, our human raters are asked to verify whether the entity is identified within the segment. To speed up the annotation process, our human raters do not report presence or absence of non-query classes.

<https://research.google.com/youtube8m/>

Google: AudioSet

The screenshot shows the homepage of the Google AudioSet dataset. The background is a dark blue gradient with a faint, stylized waveform pattern. At the top left is the logo '{ ||| } AudioSet'. At the top right are navigation links: HOME, ONTOLOGY, DATASET, DOWNLOAD, and ABOUT. In the center, the text 'A large-scale dataset of manually annotated audio events' is displayed above a white button labeled 'EXPLORE THE DATA'. Below this section, there are three rows of thumbnail images, each row containing three items. The first row shows 'Speech (1,010,480 annotations in dataset)' with two video thumbnails. The second row shows 'Baby cry, infant cry (2,390 annotations in dataset)' with three video thumbnails. The third row shows 'Fowl (6,248 annotations in dataset)' with two video thumbnails.

{ ||| } AudioSet

HOME ONTOLOGY DATASET DOWNLOAD ABOUT

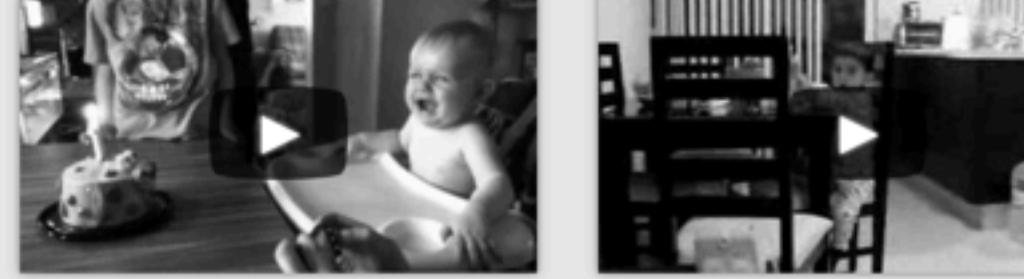
A large-scale dataset of
manually annotated audio events

EXPLORE THE DATA

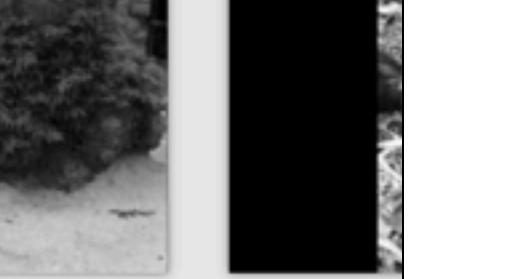
Speech (1,010,480 annotations in dataset)



Baby cry, infant cry (2,390 annotations in dataset)

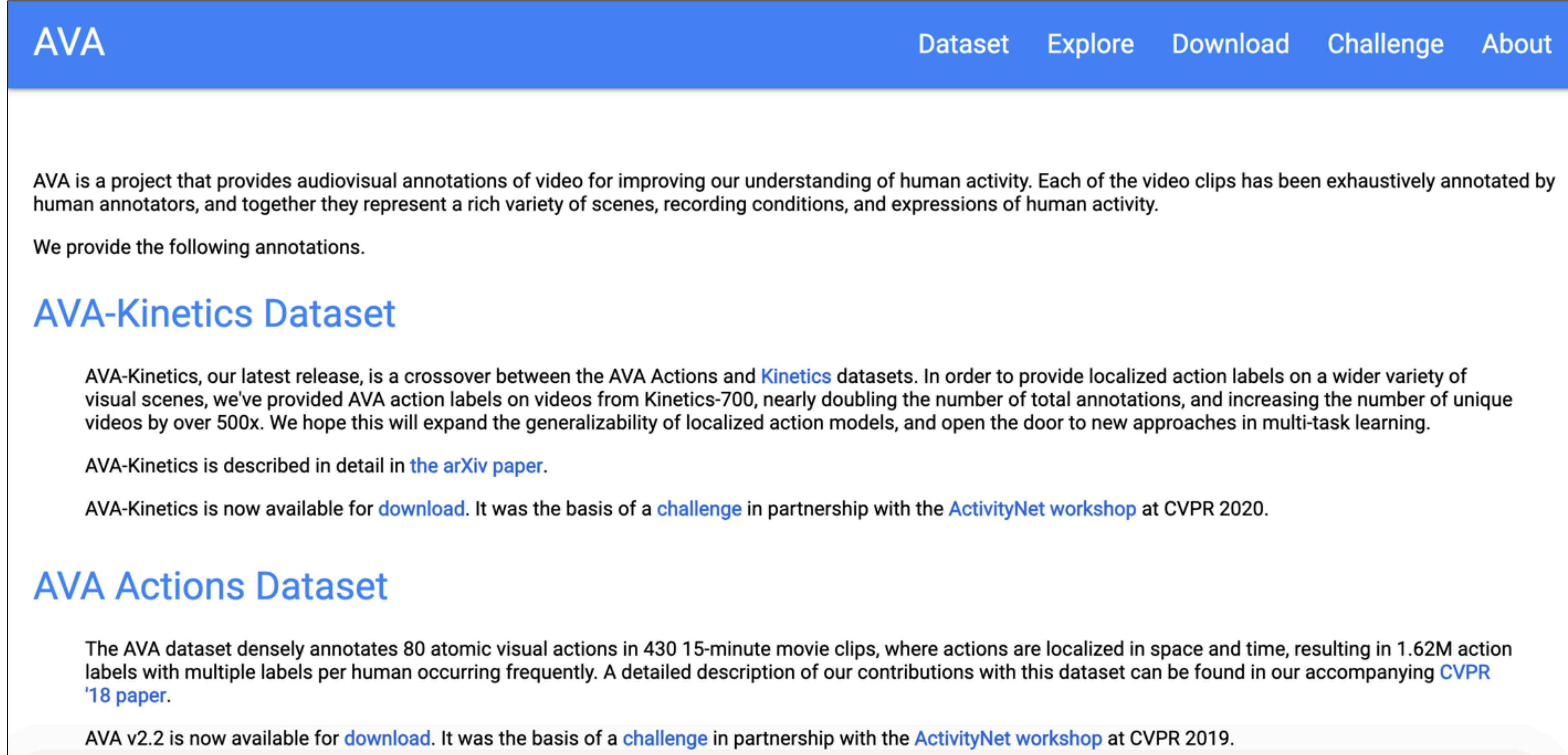


Fowl (6,248 annotations in dataset)



<https://research.google.com/audioset/>

Google: AVA, audiovisual annotations of video

The screenshot shows the homepage of the AVA dataset website. At the top, there is a blue header bar with the word "AVA" on the left and navigation links "Dataset", "Explore", "Download", "Challenge", and "About" on the right. Below the header, there is a large white area containing text and images. On the left side of this area, there is a small thumbnail image of a person. To the right of the image, the text reads: "AVA is a project that provides audiovisual annotations of video for improving our understanding of human activity. Each of the video clips has been exhaustively annotated by human annotators, and together they represent a rich variety of scenes, recording conditions, and expressions of human activity." Below this text, another paragraph states: "We provide the following annotations." Underneath this, there is a section titled "AVA-Kinetics Dataset" in blue text. The text below it describes the AVA-Kinetics dataset as a crossover between the AVA Actions and Kinetics datasets, noting that it provides localized action labels on a wider variety of visual scenes, nearly doubling the number of total annotations, and increasing the number of unique videos by over 500x. It also mentions that this will expand the generalizability of localized action models and open the door to new approaches in multi-task learning. A link to the arXiv paper is provided. Another section titled "AVA Actions Dataset" in blue text follows, describing the AVA dataset as densely annotating 80 atomic visual actions in 430 15-minute movie clips, where actions are localized in space and time, resulting in 1.62M action labels with multiple labels per human occurring frequently. A link to the CVPR '18 paper is provided. A note at the bottom of this section indicates that AVA v2.2 is now available for download and was the basis of a challenge in partnership with the ActivityNet workshop at CVPR 2019.

AVA is a project that provides audiovisual annotations of video for improving our understanding of human activity. Each of the video clips has been exhaustively annotated by human annotators, and together they represent a rich variety of scenes, recording conditions, and expressions of human activity.

We provide the following annotations.

AVA-Kinetics Dataset

AVA-Kinetics, our latest release, is a crossover between the AVA Actions and [Kinetics](#) datasets. In order to provide localized action labels on a wider variety of visual scenes, we've provided AVA action labels on videos from Kinetics-700, nearly doubling the number of total annotations, and increasing the number of unique videos by over 500x. We hope this will expand the generalizability of localized action models, and open the door to new approaches in multi-task learning.

AVA-Kinetics is described in detail in [the arXiv paper](#).

AVA-Kinetics is now available for [download](#). It was the basis of a [challenge](#) in partnership with the [ActivityNet workshop](#) at CVPR 2020.

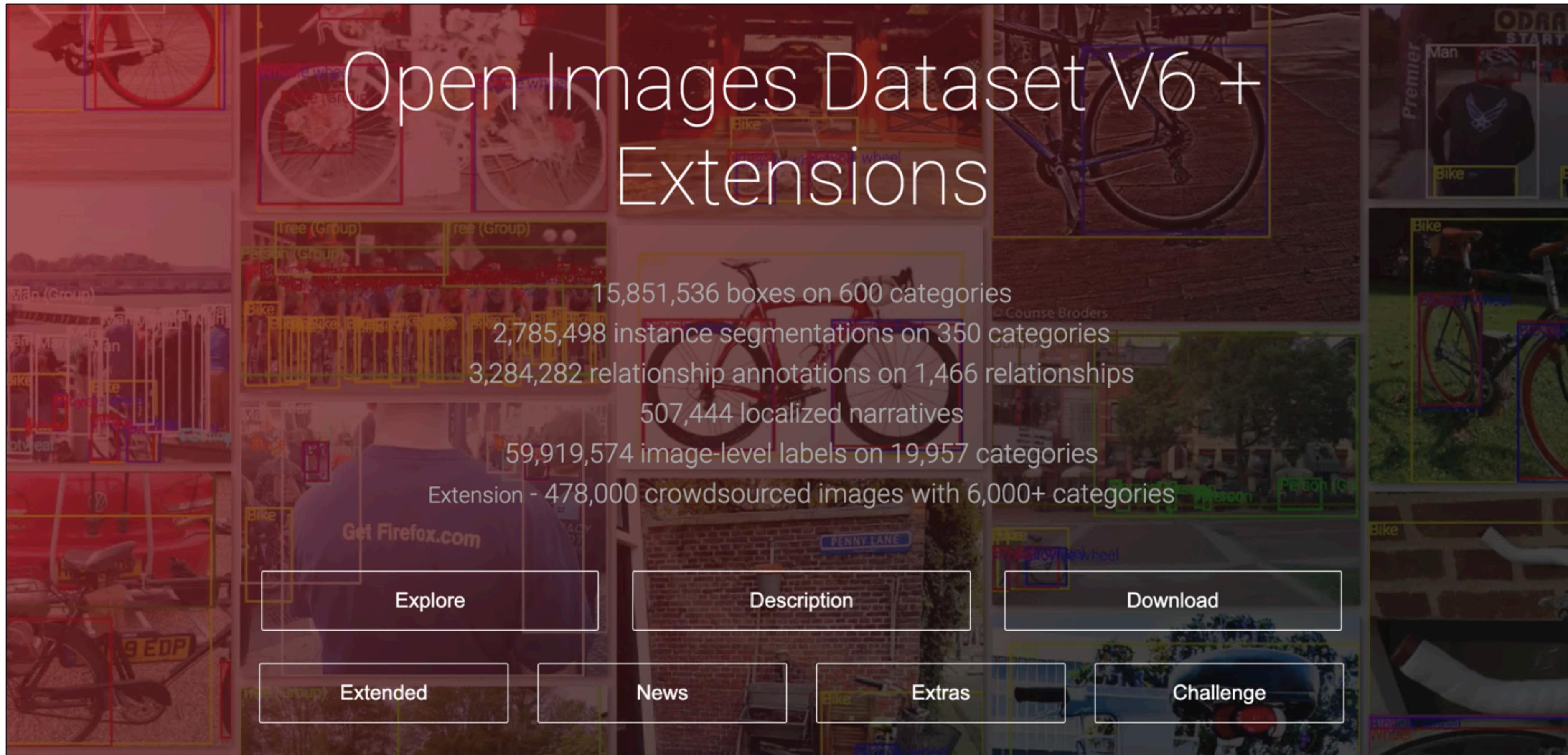
AVA Actions Dataset

The AVA dataset densely annotates 80 atomic visual actions in 430 15-minute movie clips, where actions are localized in space and time, resulting in 1.62M action labels with multiple labels per human occurring frequently. A detailed description of our contributions with this dataset can be found in our accompanying [CVPR '18 paper](#).

AVA v2.2 is now available for [download](#). It was the basis of a [challenge](#) in partnership with the [ActivityNet workshop](#) at CVPR 2019.

<https://research.google.com/ava/>

Google: Open Image



<https://storage.googleapis.com/openimages/web/index.html>

Open Data on Amazon

Registry of Open Data on AWS



About

This registry exists to help people discover and share datasets that are available via AWS resources. [Learn more about sharing data on AWS](#).

See [all usage examples for datasets listed in this registry](#).

See datasets from [Facebook Data for Good](#), [NASA Space Act Agreement](#), [NIH STRIDES](#), [NOAA Big Data Project](#), [Space Telescope Science Institute](#), and [Amazon Sustainability Data Initiative](#).

Search datasets (currently 184 matching datasets)

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the [Registry of Open Data on AWS GitHub repository](#).

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

The Cancer Genome Atlas

[cancer](#) [genomic](#) [life sciences](#) [STRIDES](#) [whole genome sequencing](#)

The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive, multi-dimensional maps of the key genomic changes in major types and subtypes of cancer. TCGA has analyzed matched tumor and normal tissues from 11,000 patients, allowing for the comprehensive characterization of 33 cancer types and subtypes, including 10 rare cancers. The dataset contains open Clinical Supplement, Biospecimen Supplement, RNA-Seq Gene Expression Quantification, miRNA-Seq Isoform Expression Quantificati...

[Details →](#)

Usage examples

- [Spatial Organization And Molecular Correlation Of Tumor-Infiltrating Lymphocytes Using Deep Learning On Pathology Images](#) by Joel Saltz, Rajarsi Gupta, et al.
- [Oncogenic Signaling Pathways in The Cancer Genome Atlas](#) by Francisco Sanchez-Vega, Marco Mina, et al.
- [Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context](#) by Hua-Sheng Chiu, Sonal Somvanshi, et al.
- [Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer](#) by Katherine A. Hoadley, Christina Yau, et al.
- [Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients](#) by André Kahles, Kjønig-Van Lehmann, et al.

[See 29 usage examples →](#)

<https://registry.opendata.aws/>

Microsoft Research Open Data

The screenshot shows the Microsoft Research Open Data homepage. At the top, there is a dark header bar with the Microsoft logo, the text "Microsoft Research Open Data", and links for "Categories", "About", "FAQs", "Feedback", and "Login". Below the header is a large, dark blue banner featuring a network graph pattern. The central text on the banner reads "Microsoft Research Open Data" in white, with a "BETA" badge to its right. Below this, there is a search bar with the placeholder "Search datasets" and a magnifying glass icon. To the left of the search bar, a text block states: "A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and". To the right, another text block says: "domain specific sciences. Download or copy directly to a cloud-based Data Science Virtual Machine for a seamless development experience." At the bottom of the page, there is a section titled "Dataset Categories" with two items: "Computer Science" (represented by a keyboard icon) and "Social Science" (represented by a speech bubble icon). Each category has a "VIEW DATASETS >" link below it.

Microsoft Research Open Data **BETA**

Search datasets

A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and

domain specific sciences. Download or copy directly to a cloud-based Data Science Virtual Machine for a seamless development experience.

Dataset Categories

Computer Science
[VIEW DATASETS >](#)

Social Science
[VIEW DATASETS >](#)

<https://msropendata.com/>

BuzzFeed

The screenshot shows the GitHub profile of the BuzzFeed News organization. The repository count is 108, and there are 3 pinned repositories. A search bar, type filter (All), and language filter (All) are visible. A specific repository, `nics-firearm-background-checks`, is highlighted.

BuzzFeed News
Open-source data, analysis, libraries, tools, and guides from BuzzFeed's newsroom.
[https://github.com/buzzfeednews/...](https://github.com/buzzfeednews/)

Repositories 108 Packages People 3 Projects

Pinned repositories

everything
An index of all our open-source data, analysis, libraries, tools, and guides.
922 stars, 99 forks

Find a repository... Type: All Language: All

nics-firearm-background-checks
Monthly data from the FBI's National Instant Criminal Background Check System, converted from PDF to CSV.
Python, MIT, 106 issues, 107 stars, 0 forks, 0 open issues, Updated 19 days ago

Top languages
Jupyter Notebook, HTML, Python, R, Shell

<https://github.com/BuzzFeedNews>

Data Sharing platforms

CKAN

The CKAN data management platform is in use by numerous governments, organisations and communities around the world. Being [open source](#), we don't know of all the instances, but here is a showcase of the ones we are aware of. We are proud and excited to see all the different uses being made of CKAN – if you would like to see your CKAN instance here [add it to the CKAN Census!](#)

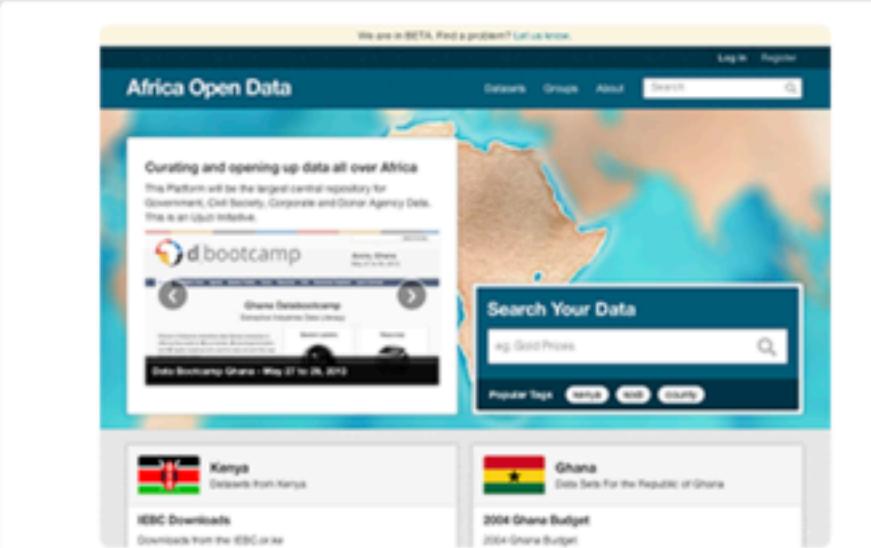
To view instances by region or type, just choose the correct facet below.

Show: [All](#) [Region](#) [Type](#)

199 instances



A.D.E.R.E. - Asociación de Entes Reguladores Eléctricos
A.D.E.R.E. - Asociación de Entes Reguladores



Africa Open Data
The Open Africa Platform initiative aims to be largest repository of Data on the African continent

RECENT POSTS

 [New CKAN 2.9 version released, patch releases for 2.7 and 2.8 available](#)

 [Subscribe to datasets: new CKAN feature explained](#)

 [Getting CKAN 2.9 over the line](#)

TWEETS

We are happy to announce the release of CKAN 2.9! Thanks to all our wonderful contributors. [#opendata #OpenSource](http://t.co/caDv7hRCVA)

<https://ckan.org/about/instances/>

DataHub

DataHub.io by DATAPLAN — We build solutions that unleash the potential of data. Let's start with yours! [Get in touch now >](#)

DATA HUB ABOUT BLOG FIND DATA COLLECTIONS DOCS PRICING TOOLS CHAT • LOGIN JOIN FREE

Search for datasets

There are thousands of datasets from financial market data and population growth to cryptocurrency prices. If you don't find what you are looking for [ask the Data Concierge](#) for a free quote for us to find you the data.

Search

VIX - CBOE Volatility Index updated daily
finance-vix core Files 2 1MB

CBOE Volatility Index (VIX) time-series dataset including daily open, close, high and low. The CBOE Volatility Index (VIX) is a key measure of market expectations of near-term volatility conveyed by S&P 500 stock index option prices introduced in 1993. Data From the VIX FAQ: In 1993, the [Explore Dataset >](#)

Premium Data Offer
Now you can request additional data and/or customized columns!

[Request Quote](#)
or Find out more

<https://datahub.io/search>

Dataverse

The screenshot shows the homepage of the Dataverse Project. At the top, there is a navigation bar with links: 'Dataverse Project' (highlighted in red), 'About ▾', 'Community ▾', 'Best Practices ▾', 'Software ▾', and 'Contact'. Below the navigation is the project's logo, 'The Dataverse® Project', featuring the word 'Dataverse' in large red letters with a registered trademark symbol, and 'Project' in smaller gray letters below it, accompanied by a red circular icon with three nodes connected by lines. The main heading 'Open source research data repository software' is displayed in large gray text. Below this, there are three sections: 'Researchers' (with a green circular icon containing two white circles), 'Journals' (with a blue circular icon containing an open book), and 'Communities' (with a purple circular icon containing four small human figures). Each section contains descriptive text and a blue link for more information.

Dataverse Project About ▾ Community ▾ Best Practices ▾ Software ▾ Contact

The
Dataverse®
Project

Open source research data repository software

 Researchers Enjoy full control over your data. Receive *web visibility, academic credit, and increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. [Want to set up your personal dataverse?](#)

 Journals Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal and associated data*. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. [Want to find out more about journal dataverses?](#)

 Communities Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. [Participate](#)

<https://dataverse.org/>

Thinknum

 Thinknum

Explore Datasets Resources Media Careers More ▾ Login ↗

Explore the breadth of Thinknum Alternative Data

Thinknum Alternative Data provides daily updates on most public and private companies.

Search by company name, ticker, or dataset (i.e. Job Listings)



Tesla Motors
NASDAQ: TSLA

Tesla, Inc. designs, develops, manufactures and sells fully electric vehicles, solar panels, batteries, and energy storage products.

[thinknum.com/datasets/nasdaq:tsla](https://www.thinknum.com/datasets/nasdaq:tsla)



Caterpillar
NYSE: CAT

Caterpillar Inc. designs, develops, engineers, manufactures, markets and sells machinery, engines, and related products.

[thinknum.com/datasets/nysc:cat](https://www.thinknum.com/datasets/nysc:cat)



Chipotle
NYSE: CMG

Chipotle Mexican Grill, Inc. is an American chain of fast casual restaurants specializing in tacos and burritos.

[thinknum.com/datasets/nysc:cmg](https://www.thinknum.com/datasets/nysc:cmg)



Amazon
NASDAQ: AMZN

Amazon.com, Inc., is a multinational technology company focusing in e-commerce, cloud computing, and digital services.

[thinknum.com/datasets/nasdaq:amzn](https://www.thinknum.com/datasets/nasdaq:amzn)



<https://www.thinknum.com/datasets>

Вопросы в ResearchGate

The screenshot shows the 'Ask a question' page on ResearchGate. At the top, there's a navigation bar with 'Home', 'Questions', and 'Jobs'. Below it, a large button says 'Ask a question' with a dropdown arrow. A sub-instruction reads: 'Enter a clear and concise question that others will easily understand. [Learn more](#)'. It also asks for details: 'Please provide any details researchers may need to answer your question.' The main form has two sections: 'Question' (with a text input field labeled 'Enter your question') and 'Description' (with a larger text input field labeled 'Enter an explanation or any details needed to understand your question'). Below these is a section titled 'What is your question about?' with two radio button options: 'I'm looking for a definition of a term or phrase' and 'I have a different research-related question'.

<https://www.researchgate.net/>

LOD cloud

The Linked Open Data Cloud

Browse Submit a dataset Diagram Subclouds About

Datasets

Search dataset... Search

1453 / 1453 datasets

Title	Identifier	View	Edit
Open Data Web	data.odw.tw		
2000 U.S. Census in RDF (rdfabout.com)	2000-us-census-rdf		
2001 Spanish Census to RDF	2001-spanish-census-to-rdf		
AAT-atawil	Altawil		
Amer Nejma	Amer Nejma		
Biblioteca Virtual Miguel de Cervantes	BVMC		
BibSonomy - The blue social bookmark and publication sharing system.	BibSonomy		
Cooperative Patent Classification	CPC		
CaLiGraph	CaLiGraph		

<https://lod-cloud.net/datasets>

Платформа Zenodo

The screenshot shows the Zenodo search interface. The top navigation bar includes the Zenodo logo, a search bar, an upload button, and community links. On the right are 'Log in' and 'Sign up' buttons. Below the header, a sidebar on the left contains filters for 'All versions' (Access Right: Open, Closed, Restricted, Embargoed) and 'File Type' (Pdf, Jpg, Png, Zip, Html, Hdf5). The main content area displays search results for 'Traffic noise data at Qianjiangxincheng' and 'fMRI data: Shape coding in occipito-temporal cortex relies on object silhouette, curvature and medial-axis'.

zenodo

Search Search

Upload Communities

Log in Sign up

All versions

Found 59739 results.

Sort by:
Most recent asc.

< 1 2 3 4 5 6 7 8 9 >

September 15, 2020 (v7) Dataset Open Access

Traffic noise data at Qianjiangxincheng

Mi Binbin; Xia Jianghai;

Traffic noise data recorded at Qianjiangxincheng, Hangzhou, China in 2019. NE161-192

Uploaded on September 20, 2020

6 more version(s) exist for this record

View

September 19, 2020 (v1) Dataset Restricted Access

fMRI data: Shape coding in occipito-temporal cortex relies on object silhouette, curvature and medial-axis

Paolo Papale;

This is the fMRI dataset described in "Shape coding in occipito-temporal cortex relies on object silhouette, curvature and medial-axis", bioRxiv (2019). Note that the data consists of already pre-processed volumes, warped into the MNI152 template (see the paper for details) along with the

Uploaded on September 19, 2020

View

Access Right

- Open (1553090)
- Closed (30029)
- Restricted (3357)
- Embargoed (1073)

File Type

- Pdf (840669)
- Jpg (355929)
- Png (200255)
- Zip (68320)
- Html (39672)
- Hdf5 (15096)

<https://datadrivenfarming.challenges.org/data-sets-2/>

Спасибо за внимание!

Сайт: <http://iRadche.ru>

Телеграм-канал: [@DataPlace](#)



@iRadche



@dadaistka



<https://www.facebook.com/iRadche>



<http://www.slideshare.net/iRadche>