

Where to find datasets for your upcoming Python project



Ruslan Korniiichuk | 10.12.2024

TABLE OF CONTENTS

01

Python native
datasets

02

General-purpose
repositories

03

Domain-specific
repositories

04

Data APIs

05

Data
marketplaces

06

Nonstandard
ways to find

01

Python native
datasets



Main

Tasks

Libraries

Languages

Licenses

Other

Modalities



3D



Audio



Geospatial



Image



Tabular



Text



Time-series



Video

Size (rows)

< 1K

> 1T

Format



json



csv



parquet



imagefolder



soundfolder



webdataset



text



» arrow

Datasets 257,511

↑↓ Sort: Trending

01-OPEN/Open01-SFT

Preview • Updated 16 days ago • ↓ 955 • ♥ 134

fka/awesome-chatgpt-prompts

Viewer • Updated Sep 3 • ⌵ 170 • ↓ 8.77k • ♥ 6.44k

HuggingFaceTB/smoltalk

Viewer • Updated 11 days ago • ⌵ 2.2M • ↓ 6.09k • ♥ 231

alpindale/two-million-bluesky-posts

Viewer • Updated 10 days ago • ⌵ 2.11M • ↓ 2.05k • ♥ 173

foursquare/fsq-os-places

Viewer • Updated 4 days ago • ⌵ 105M • ↓ 694 • ♥ 39

Xkev/LLaVA-CoT-100k

Viewer • Updated 10 days ago • ⌵ 98.6k • ↓ 1.03k • ♥ 42

5CD-AI/LLaVA-CoT-o1-Instruct

Viewer • Updated 10 days ago • ⌵ 58.5k • ↓ 304 • ♥ 48

CohereForAI/Global-MMLU

Viewer • Updated 1 day ago • ⌵ 602k • ↓ 903 • ♥ 25

opendiffusionai/cc12m-a_woman

Viewer • Updated 1 day ago • ⌵ 24.8k • ↓ 92 • ♥ 22

allenai/tulu-3-sft-mixture

Viewer • Updated 5 days ago • ⌵ 939k • ↓ 4.27k • ♥ 70

TIGER-Lab/OmniEdit-Filtered-1.2M

Viewer • Updated 2 days ago • ⌵ 1.2M • ↓ 2.05k • ♥ 16

CohereForAI/include-base-44

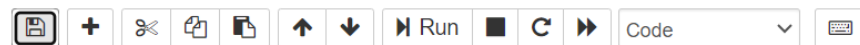
Viewer • Updated 5 days ago • ⌵ 22.6k • ↓ 189 • ♥ 16

argilla/magpie-ultra-v1.0

Viewer • Updated 11 days ago • ⌵ 3.22M • ↓ 2.05k • ♥ 26

wikimedia/wikipedia

Viewer • Updated Jan 9 • ⌵ 61.6M • ↓ 53.7k • ♥ 634



```
In [1]: import datasets
import pandas as pd
```

```
In [2]: %%capture

dataset = datasets.load_dataset("imdb", split="train")
```

```
In [3]: df = pd.DataFrame(dataset)
```

```
In [4]: df.head()
```

Out[4]:

	text	label
0	I rented I AM CURIOUS-YELLOW from my video sto...	0
1	"I Am Curious: Yellow" is a risible and preten...	0
2	If only to avoid making this type of film in t...	0
3	This film was probably inspired by Godard's Ma...	0
4	Oh, brother...after hearing about this ridicul...	0

Section Navigation

- 1. Supervised learning ▾
- 2. Unsupervised learning ▾
- 3. Model selection and evaluation ▾
- 4. Inspection ▾
- 5. Visualizations
- 6. Dataset transformations ▾
- 7. Dataset loading utilities** ▴
- 7.1. Toy datasets
- 7.2. Real world datasets
- 7.3. Generated datasets
- 7.4. Loading other datasets
- 8. Computing with scikit-learn ▾
- 9. Model persistence
- 10. Common pitfalls and recommended practices
- 11. Dispatching ▾
- 12. Choosing the right estimator
- 13. External Resources, Videos and Talks

[Home](#) > [User Guide](#) >[Show Source](#)

7. Dataset loading utilities

The `sklearn.datasets` package embeds some small toy datasets and provides helpers to fetch larger datasets commonly used by the machine learning community to benchmark algorithms on data that comes from the ‘real world’.

To evaluate the impact of the scale of the dataset (`n_samples` and `n_features`) while controlling the statistical properties of the data (typically the correlation and informativeness of the features), it is also possible to generate synthetic data.

General dataset API. There are three main kinds of dataset interfaces that can be used to get datasets depending on the desired type of dataset.

The dataset loaders. They can be used to load small standard datasets, described in the [Toy datasets](#) section.

The dataset fetchers. They can be used to download and load larger datasets, described in the [Real world datasets](#) section.

Both loaders and fetchers functions return a `Bunch` object holding at least two items: an array of shape `n_samples * n_features` with key `data` (except for 20newsgroups) and a numpy array of length `n_samples`, containing the target values, with key `target`.

The Bunch object is a dictionary that exposes its keys as attributes. For more information about Bunch object, see [Bunch](#).

It’s also possible for almost all of these function to constrain the output to be a tuple containing only the data and the target, by setting the `return_X_y` parameter to `True`.

The datasets also contain a full description in their `DESCR` attribute and some contain `feature_names` and `target_names`. See the dataset descriptions below for details.

The dataset generation functions. They can be used to generate controlled synthetic datasets, described



```
In [1]: import pandas as pd
import sklearn.datasets
```

```
In [2]: dataset = sklearn.datasets.fetch_california_housing() # returns a Bunch object
```

```
In [3]: df = pd.DataFrame(dataset.data, columns=dataset.feature_names)
df['target'] = pd.Series(dataset.target)
```

```
In [4]: df.head()
```

Out[4]:

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422



Datasets

[Overview](#)[Catalog](#)[Community Catalog](#)[Guide](#)[API](#)

TFDS now supports the [Croissant](#) 🍷 format! Read the [documentation](#) to know more.

TensorFlow Datasets: a collection of ready-to-use datasets.

TensorFlow Datasets is a collection of datasets ready to use, with TensorFlow or other Python ML frameworks, such as Jax. All datasets are exposed as [tf.data.Datasets](#), enabling easy-to-use and high-performance input pipelines. To get started see the [guide](#) and our [list of datasets](#).

```
import tensorflow as tf
import tensorflow_datasets as tfds

# Construct a tf.data.Dataset
ds = tfds.load('mnist', split='train', shuffle_files=True)

# Build your input pipeline
ds = ds.shuffle(1024).batch(32).prefetch(tf.data.AUTOTUNE)
for example in ds.take(1):
    image, label = example["image"], example["label"]
```

[RUN IN A CO NOTEBOOK](#)

Datasets

Overview

Catalog

Community Catalog

Guide

API

☰ Filter

Overview

- › Dataset Collections
- › 3d
- › Abstractive text summarization
- › Age
- › Anomaly detection
- › Audio
- › Biology
- › Categorical
- › Common sense reasoning
- › Computer science
- › Conditional image generation
- › Coreference resolution
- › D4rl
- › Density estimation
- › Dependency parsing
- › Dialog act labeling
- › Dialogue

TFDS now supports the [Croissant 🥐 format!](#) Read the [documentation](#) to know more.

TensorFlow > Resources > Datasets > Catalog

Was this helpful? 👍 🗨

Datasets



★ **Note:** The datasets documented here are from **HEAD** and so not all are available in the current **tensorflow-datasets** package. They are all accessible in our nightly package **tfds-nightly**.

Usage

See our [getting-started guide](#) for a quick introduction.

```
for ex in tfds.load('cifar10', split='train'):
```



On this page

[Usage](#)[All Datasets](#)[Dataset Collections](#)[3d](#)[Abstractive text summarization](#)[Age](#)[Anomaly detection](#)[Audio](#)[Biology](#)[Categorical](#)[Common sense reasoning](#)[Computer science](#)[Conditional image generation](#)[Coreference resolution](#)[D4rl](#)



```
In [1]: import pandas as pd
import tensorflow as tf
import tensorflow_datasets as tfds
```

```
In [2]: # Construct a tf.data.Dataset
ds, ds_info = tfds.load("imdb_reviews", split="train", with_info=True)
```

```
In [3]: df = tfds.as_dataframe(ds.take(3), ds_info)
```

```
In [4]: df
```

Out[4]:

	label	text
0	0 (neg)	This was an absolutely terrible movie. Don't be lured in by Christopher Walken or Michael Ironside. Both are great actors, but this must simply be their worst role in history. Even their great acting could not redeem this movie's ridiculous storyline. This movie is an early nineties US propaganda piece. The most pathetic scenes were those when the Columbian rebels were making their cases for revolutions. Maria Conchita Alonso appeared phony, and her pseudo-love affair with Walken was nothing but a pathetic emotional plug in a movie that was devoid of any real meaning. I am disappointed that there are movies like this, ruining actor's like Christopher Walken's good name. I could barely sit through it.
1	0 (neg)	I have been known to fall asleep during films, but this is usually due to a combination of things including, really tired, being warm and comfortable on the sette and having just eaten a lot. However on this occasion I fell asleep because the film was rubbish. The plot development was constant. Constantly slow and boring. Things seemed to happen, but with no explanation of what was causing them or why. I admit, I may have missed part of the film, but i watched the majority of it and everything just seemed to happen of its own accord without any real concern for anything else. I cant recommend this film at all.
2	0 (neg)	Mann photographs the Alberta Rocky Mountains in a superb fashion, and Jimmy Stewart and Walter Brennan give enjoyable performances as they always seem to do. But come on Hollywood - a Mountie telling the people of Dawson City, Yukon to elect themselves a marshal (yes a marshal!) and to enforce the law themselves, then gunfighters battling it out on the streets for control of the town? Nothing even remotely resembling that happened on the Canadian side of the border during the Klondike gold rush. Mr. Mann and company appear to have mistaken Dawson City for Deadwood, the Canadian North for the American Wild West. Canadian viewers be prepared for a Reefer Madness type of enjoyable howl with this ludicrous plot, or, to shake your head in disgust.

seaborn-data

Public

Watch 64

Fork 3k

Star 1.6k

master

1 Branch

0 Tags

Go to file

t

Add file

<> Code

About

Data repository for seaborn examples

seaborn.pydata.org/

Readme

Activity

1.6k stars

64 watching

3k forks

Report repository

Releases

No releases published

Packages

No packages published

Contributors 3

mwaskom	Update README.md	71e2436 · now	48 Commits
png	More updates	3 years ago	
process	Remove one-off 2021 datapoint from healthexp dataset	2 years ago	
raw	Add dowjones dataset	2 years ago	
README.md	Update README.md	now	
anagrams.csv	Rename messy anagrams dataset	4 years ago	
anscombe.csv	Add anscombe dataset	10 years ago	
attention.csv	Add attention dataset	11 years ago	
brain_networks.csv	Add brain networks dataset	10 years ago	
car_crashes.csv	Add 538 car crash dataset	9 years ago	
dataset_names.txt	Add a file containing all available dataset names	last year	
diamonds.csv	Add diamonds dataset	6 years ago	



```
In [1]: import pandas as pd
import seaborn
```

```
In [2]: seaborn.get_dataset_names()
```

```
Out[2]: ['anagrams',
'anscombe',
'attention',
'brain_networks',
'car_crashes',
'diamonds',
'dots',
'dowjones',
'exercise',
'flights',
'fmri',
'geyser',
'glue',
'healthexp',
'iris',
'mpg',
'penguins',
'planets',
'seaice',
'taxis',
'tips',
'titanic']
```

```
In [3]: dataset = seaborn.load_dataset("tips")
```

```
In [4]: dataset.head()
```

```
Out[4]:
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

02

General-purpose repositories

General-purpose repositories

- Google Dataset Search
- Kaggle
- UCI Machine Learning Repository
- Opendatasoft
- OpenML
- DataHub
- Gapminder
- FiveThirtyEight
- Awesome Public Datasets
- Open Government Data: EU, Poland, USA, Canada

Dataset Search

Search for Data Sets



Try [coronavirus covid-19](#) or [water quality site:canada.ca](#).

[Learn more](#) about Dataset Search.

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

[+ New Dataset](#)[All datasets](#)[Computer Science](#)[Education](#)[Classification](#)[Computer Vision](#)[NLP](#)[Data Visualization](#)[Pre-Trained Model](#)

Trending Datasets

[See All](#)

Filters

Search datasets...



Keywords



Attributes



Data Type



Subject Area



Task



Instances



Features



Feature Type



Python



Browse Datasets



SORT BY # VIEWS, DESC



EXPAND ALL



Iris

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classific...



Classification



Tabular



150 Instances



4 Features



Heart Disease

4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach



Classification



Multivariate



303 Instances



13 Features



Wine Quality

Two datasets are included, related to red and white vinho verde wine samples, from the north of Portuga...



Classification, Regre...



Multivariate



4.9K Instances



12 Features



Adult

Predict whether annual income of an individual exceeds \$50K/yr based on census data. Also known as "C...



Classification



Multivariate



48.84K Instances



14 Features

34,486 **datasets** already open and ready to use

Search datasets



906M

downloads

34k

open datasets

14.28B

API calls



Search



Datasets

5.8k



Tasks



Flows



Runs



Collections



Benchmarks



Task Types



Measures

Learn

Documentation [↗](#)Blog [↗](#)

API's



Contribute



MINIFY



DARK

5849 datasets found

verified



Datasets

Datasets provide training data for machine learning models. OpenML datasets are uniformly formatted and come with rich meta-data to allow automated processing. You can sort or filter them by a range of different properties. [Learn more.](#)



credit-g

This dataset classifies people described by a set of attributes as good or bad credit risks.



506k



1000 x 21



31



11 years ago

v.1



blood-transfusion-service-center

Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan -- this is a classification problem.



469k



748 x 5



1464



10 years ago

v.1



monks-problems-2

Once upon a time, in July 1991, the monks of Corsendonk Priory were faced with a school held in their priory, namely the 2nd European Summer School on Machine Learning. After listening more than one week to a wide variety of learning algorithms, they felt rather confused: Which algorithm



395k



601 x 7



334



10 years ago

v.1



tic-tac-toe

COLLECTIONS

Curated dataset collections

The awesome section presents collections of high quality datasets organized by topic. Home page for awesome collections is located in the **awesome-data** repository on GitHub and should be modified from there.



Air Pollution

Air pollution datasets and other sources of related data



Bibliographic Data

Existing databases or services providing substantial bibliographic data



Broadband

Data on broadband access and usage around the world



Climate Change

A collection of the most important "general" datasets on climate change

awesome-data

Public

Watch 49

Fork 101

Star 807

main

1 Branch

0 Tags

Go to file

Add file

<> Code

 olawayway	Update economic-data.md	84dd6a4 · 2 months ago	139 Commits
 assets	[add,notes/clio-infra.eu][m]: add a datahub.	2 years ago	
 notes	[add,notes/clio-infra.eu][m]: add a datahub.	2 years ago	
 .gitattributes	[git][s]: git lfs track "**.png" "**.jpg".	2 years ago	
 .gitignore	[git][xs]: gitignore obsidian.	2 years ago	
 README.md	Update README.md with the link to postal codes datasets c...	3 months ago	
 air-pollution.md	[air-pollution][xs]: add keywords.	3 years ago	
 bibliographic-data.md	updated meta property name (#31)	6 years ago	
 broadband.md	[broadband][s]: create broadband file and add microsoft dat...	3 years ago	
 climate-change.md	[fix][s]: broken links to /core datasets and [toc] links	3 months ago	
 climate-data.md	Update climate-data.md	4 years ago	
 config.json	Update config.json	2 months ago	
 coronavirus.md	Update coronavirus.md	2 months ago	
 demographics.md	[fix][s]: broken links to /core datasets	3 months ago	
 economic-data.md	Update economic-data.md	2 months ago	
 education.md	[fix][s]: broken links to /core datasets and [toc] links	3 months ago	

About

Curated list of quality open datasets

- 
- data

opendata

open-data

datasets

open-datasets

datasets-csv

-  Readme
-  Activity
-  Custom properties
-  807 stars
-  49 watching
-  101 forks
- Report repository

Releases

No releases published

Packages

No packages published

Contributors 15



Home > **Download the data**

Download the data

Data documentation

[Read more here](#)

Bulk download

The complete datasets with hundreds of indicators are available in Github repositories:

🔌 [Systema Globalis](#) (indicators inherited from Gapminder World, many are still updated)

🔌 [Fast Track](#) (indicators we compile manually)

🔌 [World Development Indicators](#) (direct copy from World Bank)

The data is organized in loose CSV files which can be consumed by any spreadsheet software. If you are developing something and want to work with the full datasets more efficiently you can benefit from [DDF data model](#).

Choose individual indicators

This menu lists all indicators available in Gapminder Tools. Select one to preview the data, then click download as CSV or XLSX.

Our Data

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.

DATA SET	RELATED CONTENT	LAST UPDATED		
polls	Latest Polls	Sept. 27, 2023	info	↓
congress-generic-ballot	Do Voters Want Democrats Or Republicans In Congress?	Sept. 26, 2023	info	↓
mlb-elo	2023 MLB Predictions	June 21, 2023	info	↓
nba-forecasts	2022-23 NBA Predictions	June 14, 2023	info	↓



awesome-public-datasets

Public

 Watch 2268

 Fork 10k

 Star 61.3k

 master

 2 Branches

 2 Tags

 Go to file

t

 Add file

 Code



caesar0301 Update README sha: [7207b0e](#)

98f51db · 3 weeks ago  781 Commits



Datasets

Add titanic dataset

10 years ago



LICENSE

Update license copyright info.

9 years ago



README.rst

Update README sha: [7207b0e](#)

3 weeks ago



README



MIT license



Awesome Public Datasets



awesome

This is a list of [topic-centric public data sources](#) in high quality. They are collected and tidied from blogs, answers, and user responses. Most of the data sets listed below are free, however, some are not. This project was incubated at [OMNILab](#), Shanghai Jiao Tong University during Xiaming Chen's Ph.D. studies. OMNILab is now part of the [BaiYuLan Open AI community](#). Other amazingly awesome lists can be found in [sindresorhus's awesome](#) list.

About

A topic-centric list of HQ open datasets.

 [awesomedataworld.slack.com](#)

[opendata](#)

[datasets](#)

[aaron-swartz](#)

[awesome-public-datasets](#)

 Readme

 MIT license

 Activity

 Custom properties

 61.3k stars

 2.3k watching

 10k forks

Report repository

Releases 1



v20230903

Latest

on Sep 3, 2023

Open Government Data

European data

data.europa.eu

The official portal for European data

[Home](#)

[Data](#) ▾

[EU Open Data Days](#)

[Academy](#)

[Community](#) ▾

[Publications](#) ▾

[Documentation](#) 

[Home](#) > [Datasets](#)

Datasets

High-Value Dataset

HVDs only

☐ Yes ☒ No

High-Value-Dataset category

- Any -



Data scope

- Any -



Search



[Datasets](#)



Datasets found (1 814 266)

[API request](#) 

Sort by:

Relevance



Register of Auditors and Auditing Entities

The dataset contains information from the “Registry of Auditors and Subjects of Audit Activities”. There are two resources in the dataset: subjects – information about auditing entities and auditors – information about...



Poland's Data Portal

Use public data free of charge, also for commercial purposes



Projekt: "Dane 3.0 - wymiana, wartość"

Fundusze Europejskie

na Rozwój Cyfrowy



Dofinansowane przez
Unię Europejską



MC





CELEBRATING **15 YEARS** OF DATA.GOV

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.

301,899 DATASETS AVAILABLE

Search

[Most Viewed Datasets](#)



[Recently Added Datasets](#)



[Datasets by Organization](#)



[Geospatial](#)



Government
of Canada

Gouvernement
du Canada

[Français](#)

MENU ▾

[Canada.ca](#) > [Open Government](#)

Open Government Portal

Search



[Open Data Portal Catalogue Dataset](#)



[Download Search Results](#)



[Search Tips](#)

Filters

► Organization

► Portal Type

► Collection Type

41868 records

Order By

Updated ▾

[Proactive Disclosure - Travel Expenses](#)

Federal

Travel Expenses: The rules and principles governing travel are outlined in the Policies for Ministers' Offices, the National Joint Council's ...

03

Domain-specific repositories

Domain-specific repositories

- FBI Crime Data Explorer
- NASA Open Data Portal and NASA Earth Data
- Global Health Observatory
- BFI Industry Data
- CERN Open Data Portal



FBI CRIME DATA
EXPLORER

Data Explorer ▼

Data Discovery Tool

Documents & Downloads

About

National Data

Explore

The FBI's Crime Data Explorer (CDE) aims to provide transparency, create easier access, and expand awareness of criminal, and noncriminal, law enforcement data sharing; improve accountability for law enforcement; and provide a foundation to help shape public policy in support of a safer nation. Use the CDE to discover data through visualizations, downloads in .csv format, and other large data files.

State/Agency Data

Explore

U.S. Violent Crime Trends, January 2023 -
December 2023



U.S. Property Crime Trends, January 2023 -
December 2023



Filters

View Types

☐ Calendars

☐ Charts

☐ Datasets

☐ External Datasets

☐ Files and Documents

☐ Filtered Views

☐ Forms

☐ Maps

☐ Measures

☐ Stories

Category

☐ Aerospace

☐ Applied Science

Machine Learning Suitable NASA Datasets

 Search catalog... 

If you're a NASA individual adding more data to data.nasa.gov that has a machine-learning application, please use the topic tag usg-artificial-intelligence in your metadata to make sure it returns in these results: <https://data.nasa.gov/browse?tags=usg-artificial-intelligence>

10000 results

Sort by
Recently added 

 NASA Acronyms in Public Abstracts

Last updated
January 31, 2023

Views
13741

 Mars orbital image (HiRISE) labeled data set version 3

Last updated
January 31, 2023

Views
18203

 [TROPICS07 L2B Tropical Cyclone Intensity Estimate \(TCIE\) Algorithm V1.0](#)

The "Time-Resolved Observations of Precipitation structure and storm Intensity with a Constellation of Smallsats" (TROPICS) mission has a goal of providing nearly all-weather...

Last updated
December 2, 2024

Views



[Home](#) / [Find Data](#)

Find Data

NASA acquires, archives, and distributes data from a variety of sources, including remote sensing instruments on satellite and airborne platforms, field campaigns, in situ measurements, and model outputs. Through interagency and international partnerships, NASA also distributes and promotes Earth observation data from non-NASA missions.

[Data Tools](#) → [Data Catalog](#) →



Explore NASA Earth Data

Our data span the breadth of Earth science disciplines and cover even the most remote areas of the planet, providing measurements of Earth's atmosphere, ocean, land, and cryosphere; assessments about how humans interact with the environment; and assessments of calibrated radiance and solar radiance. These measurements

Get Started

Are you new to NASA Earth science data and need a boost? Check out all of our



Global Health Observatory data repository

[Data analysis](#)[By theme](#)[By category](#)[By indicator](#)[Metadata](#)[About the Observatory](#)[Search](#)

Welcome to the Global Health Observatory, WHO's gateway to health-related statistics for more than 1000 indicators for its 194 Member States. Data are organized to monitor progress towards the Sustainable Development Goals (SDGs), including health status indicators to monitor progress towards the overall health goal, indicators to track equity in health indicators, and the indicators for the specific health and health-related targets of the SDGs.

SDG health and health-related targets

- [Mortality and global health estimates](#)
- [Health Equity Monitor](#)

3.1 Maternal mortality

- [Maternal and reproductive health](#)

3.2 Newborn and child mortality

- [Child health](#)
- [Child mortality](#)

3.3 Communicable diseases



Industry data and insights

Read free research data and market intelligence on the UK film industry and other screen sectors.

We publish the annual BFI Statistical Yearbook, and produce official statistics and bespoke research throughout the year.

Weekend box office figures

Weekly figures for the top 15 films released in the UK, all other British releases and newly-released films.

[Find out more >](#)

Statistical Yearbook

The comprehensive resource for statistics, tables, charts and commentary on the UK screen industries.

[Find out more >](#)

Official statistics release calendar

Official statistics on the UK film industry and other screen sectors.

[Find out more >](#)

Explore more than **five petabytes**
of open data from particle physics!

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

Explore

[datasets](#)
[software](#)
[environments](#)
[documentation](#)

Focus on

[ALICE](#)
[ATLAS](#)
[CMS](#)
[DELPHI](#)
[LHCb](#)
[OPERA](#)
[PHENIX](#)
[TOTEM](#)
[Data Science](#)

04

Data APIs

public-apis

Public

Watch 4199 ▾

Fork 34k ▾

★ Starred 319k ▾

master ▾

2 Branches

0 Tags

Go to file t

Add file ▾

<> Code ▾

 apilayer-admin	Update README.md	274ecf0 · 2 months ago	🕒 4,535 Commits
📁 .github	Add files via upload		2 years ago
📁 scripts	Fix false negative http code 404 in verification		2 years ago
📄 .gitattributes	Ignore .github		7 years ago
📄 .gitignore	Create .gitignore file to Python projects		3 years ago
📄 CONTRIBUTING.md	added postman logos, removed deadlinks, created popular...		8 months ago
📄 LICENSE	Update year to 2022		3 years ago
📄 README.md	Update README.md		2 months ago

📖 README

📄 MIT license

✎

☰

Try Public APIs for free

About

A collective list of free APIs

[APILayer.com/?utm_source=Github&utm...](#)

- api
- lists
- open-source
- list
- development
- public
- resources
- dataset
- free
- software
- apis
- public-api
- public-apis

- 📖 Readme
- 📄 MIT license
- 🔔 Activity
- 📋 Custom properties
- ★ 319k stars
- 👁 4.2k watching
- 🍴 34k forks
- Report repository

Contributors 1,256

05

Data marketplaces

[About AWS](#)[Contact Us](#)[Support](#) ▾[English](#) ▾[My Account](#) ▾[Sign In to the Console](#)[re:Invent](#)[Products](#)[Solutions](#)[Pricing](#)[Documentation](#)[Learn](#)[Partner Network](#)[AWS Marketplace](#)[Customer Enablement](#)[Events](#)[Explore More](#)[AWS Data Exchange](#)[Overview](#) ▾[Data Sets](#) ▾[Solutions](#) ▾[Latest Updates](#)[How To](#) ▾[For Providers](#)[Get Started](#)

AWS Data Exchange

Easily find, subscribe to, and use third-party data in the cloud

Browse 3,500+ third-party data sets



Extensive data set catalog



Better data technology with
AWS integration



Streamlined data procurement
and governance



Easy to use for data files, tables,
and APIs

AWS Data Exchange advantage





Marketplace

Marketplace > Data



Marketplace home



Your products



Your orders

Filter Type to filter

Category



Analytics (50)

Big data (35)

Databases (10)

Machine learning (6)

Generative AI (2)

Type

Data



253 results



Cloud-to-Ground Lightning Strikes

NOAA

Aggregated lightning strike data from 1987 to 2018



ZoomInfo - Companies under 1,000 employees offering...

ZoomInfo

Top 100 companies from ZoomInfo's Company Data Cube with under 1,000 employees...



Circana ProScores

Circana LLC

Scalable Purchase-Based Audiences



Affluence Index™

Equifax Inc.

Household-level continuous scoring system; ranks estimated consumer capacity to spend,save...



Economic Cohorts

Equifax Inc.

An advanced consumer segmentation system based on household economic profiles and...



D&B ID Graph

Dun & Bradstreet

B2B2C Identity for True Omnichannel Marketing

06

Nonstandard ways
to find

Nonstandard ways to find

- Find data in your company
- GitHub
- European funds
- Web scraping
- Data augmentation
- Synthetic data generation
- Amazon Mechanical Turk (MTurk)
- Papers With Code
- Academic Torrents

Find data in your company

- [Google Marketing Platform \(GMP\)](#)
- [Meta Business Suite](#)
- [Alation Data Catalog](#)
- [Collibra Data Intelligence Cloud](#)
- [Informatica Cloud Data Governance and Catalog](#)

Filter by

<> Code

39.4k

📁 Repositories

1

🔔 Issues

56

🔗 Pull requests

1

💬 Discussions

2

👤 Users

0

⌵ More

Advanced

+ Owner

+ Size

+ Number of followers

+ Number of forks

+ Number of stars

+ Date created

+ Date pushed

+ Topic

+ License

+ Archived

1 result (134 ms)

Sort by: Best match ▾

🔖 Save

⋮

adychn/TMall-CTR-CVR-Prediction

☆ Star

Use machine learning to predict CTR/CVR TMall dataset.

● Python · ☆ 0 · Updated on Aug 11, 2019

Sponsor open source projects you depend on

Contributors are working behind the scenes to make open source better for everyone—give them the help and recognition they deserve.

[Explore sponsorable projects →](#)

How can we improve search? [Give feedback](#)

💡 **ProTip!** Press the / key to activate the search input again and adjust your query.



Learn more about European Funds



Discover how the Funds work



Look through the documents



Search through the projects



Explore the data

Press to roll down ▾

Good examples of EU funded projects

See examples of the most interesting projects implemented from the National Programmes for the 2014-2020 perspective.

[Search through the projects](#) ➔

Activities results:

124 329
enterprises supported

13 196
product innovations

14 604
premises adjusted to the needs
of persons with disabilities

Start > Projects > Przeprowadzenie prac badawczo-rozwojowych nad opracowaniem modułu automatyzacji przeprowadzania testów typu A/B/x w oparciu o rozwiązania machine learning (ML) w celu zwiększenia współczynnika konwersji strony internetowej (landing page)

[< Return to search](#)

The title of the project

Przeprowadzenie prac badawczo-rozwojowych nad opracowaniem modułu automatyzacji przeprowadzania testów typu A/B/x w oparciu o rozwiązania machine learning (ML) w celu zwiększenia współczynnika konwersji strony internetowej (landing page)

category



research, development, innovation



Name of the beneficiary

**LANDINGI SPÓŁKA Z OGRANICZONĄ
ODPOWIEDZIALNOŚCIĄ**

Project value

4 002 278,80 zł

Co-financing from the EU

2 792 761,78 złVoivodeship
śląskie

County



program

**Regionalny Program Operacyjny Województwa Śląskiego
na lata 2014-2020**

action

1.2. Badania Rozwój i innowacje w przedsiębiorstwach

fund

Europejski Fundusz Rozwoju Regionalnego

perspective

2014 - 2020

Search for Articles:

Title / Keyword

Author / Affiliation / Email

Entropy

All Article Types

Search

Advanced

Journals / Entropy / Volume 23 / Issue 11 / 10.3390/e23111388



Submit to this Journal

Review for this Journal

Propose a Special Issue

Article Menu

Academic Editors



Mehul Motani



Song-Nam Hong

Subscribe SciFeed

Recommended Articles

Related Info Links



Open Access

Article

Conversion Rate Prediction Based on Text Readability Analysis of Landing Pages

by Ruslan Kornichuk * and Mariusz Boryczka

Institute of Computer Science, Faculty of Science and Technology, University of Silesia in Katowice, Będzińska 39, 41-200 Sosnowiec, Poland

* Author to whom correspondence should be addressed.

Entropy 2021, 23(11), 1388; <https://doi.org/10.3390/e23111388>

Submission received: 24 July 2021 / Revised: 21 September 2021 / Accepted: 20 October 2021 /

Published: 23 October 2021

Download

Browse Figures

Versions Notes

Abstract

Digital marketing has been extensively researched and developed remarkably rapidly over the last decade. Within this field, hundreds of scientific publications and patents have been produced, but the accuracy of prediction

Order Article Reprints



Altmetric



Share



Help



Cite

Discuss in
SciProfiles

Endorse



Comment



Fundusze Europejskie
dla Nowoczesnej Gospodarki



[Aktualności](#) | [O Programie](#) | [Kontakt](#)

Szukaj w serwisie Strony programów

Fundusze Europejskie dla Nowoczesnej Gospodarki 2021-2027

Budżet programu:
ok. 10 mld euro

[Sprawdź aktualne i planowane konkursy](#)

Cele programu

1. wzmocnienie innowacyjności przedsiębiorstw
2. wzrost konkurencyjności MŚP
3. cyfrowa i zielona transformacja przedsiębiorstw



Web scraping

Web scraping is an automated method of extracting data from websites. It involves retrieving and transforming information from the web for various purposes, such as data analysis, research, or to create datasets.

Tools and Libraries:

- Requests, AIOHTTP, HTTPX
- Beautiful Soup
- Selenium
- Requestium
- Scrapy



View PDF

[Download full issue](#)[Outline](#)[Abstract](#)[Keywords](#)[References](#)

ELSEVIER

Procedia Computer Science

Volume 192, 2021, Pages 3677-3685



Averaging and Boosting Methods in Ensemble-Based Classifiers for Text Readability

[Ruslan Korniiichuk](#)✉, [Mariusz Boryczka](#)

[Show more](#) ✓

Add to Mendeley



Share



Cite

<https://doi.org/10.1016/j.procs.2021.09.141> ↗[Get rights and content](#) ↗Under a Creative Commons [license](#) ↗

● open access

Part of special issue



Knowledge-Based and
Intelligent Information &
Engineering Systems:
Proceedings of the 25th
International Conference
[KES2021](#)

Edited by Jaroslaw Watrobski,
Wojciech Salabun, Carlos Toro,
Cecilia Zanni-Merk, Robert J.
Howlett, Lakhmi C. Jain

PDF

[Help](#)[Download full issue](#)

Other articles from this
issue



FEEDBACK

Data augmentation

Data augmentation is a technique used to artificially increase the size of a dataset by applying various transformations to the existing data.

Common Transformations:

- rotation
- flip (horizontal/vertical)
- cropping, scaling and zooming
- image translation (i.e., shifting)
- brightness, contrast, color, and saturation
- adding noise

Synthetic data generation

Synthetic data generation involves creating artificial datasets that mimic real-world data patterns without directly using authentic data instances.

Generation methods:

- generative AI
- rule-based generation
- interpolation and extrapolation

Amazon Mechanical Turk

Access a global, on-demand, 24x7 workforce

[Get started with Amazon Mechanical Turk](#)

Looking for data labeling solutions to power Machine Learning models?

Amazon SageMaker Ground Truth allows you to easily build and manage your own data labeling workflows and workforce. Or, use Ground Truth Plus, a turnkey data labeling service that provides an expert workforce and manages it on your behalf. Amazon Mechanical Turk is accessible through both Ground Truth and Ground Truth Plus.

[Learn More »](#)

Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development.

While technology continues to improve, there are still many things that human beings can do much more effectively than computers, such as moderation,

[Browse State-of-the-Art](#)[Datasets](#)[Methods](#)[More ▾](#)[Sign In](#)

Datasets

11,052 machine learning datasets



Share your dataset with the ML community!



Best match



Filter by Modality

Images

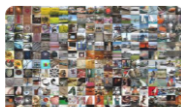
2991



CIFAR-10 (Canadian Institute for Advanced Research, 10 classes)

The CIFAR-10 dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The...

15,229 PAPERS • 108 BENCHMARKS



ImageNet

The ImageNet dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale...

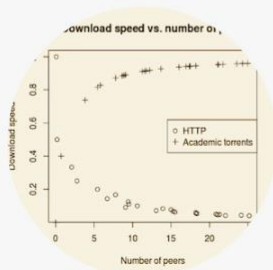
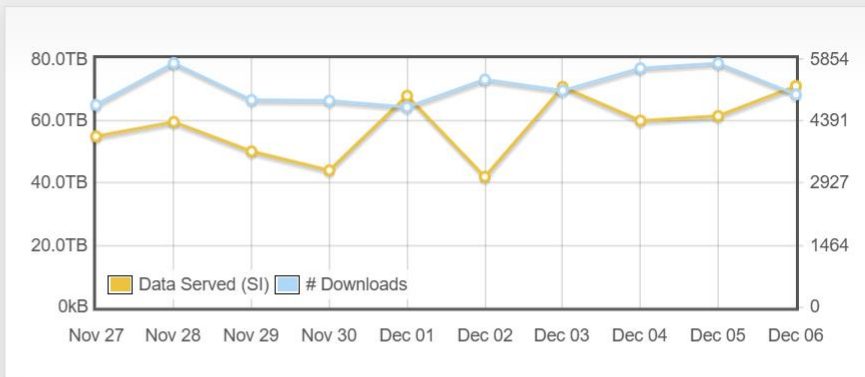
Making over 127.15TB of research data available!

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!



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

tinyurl.com/PYSUMMIT





RUSLAN KORNIICHUK

Software and Platform Engineer in Poland



Hire me

Solid experience of 9+ years in IT. Former Software Engineer and Architect at Fortune 500 companies. Core areas: Cloud Computing, and Software Engineering. Especially interested in Platform Engineering. Click the button above to hire me.

