- **Artificial intelligence vs Machine learning**
- **Datasets for Machine Learning**

Jagendra Singh

DATASET FOR MACHINE LEARNING

The key to success in the field of machine learning or to become a great data scientist is to practice with different types of datasets.

But discovering a suitable dataset for each kind of machine learning project is a difficult task. So, in this topic, we will provide the detail of the sources from where you can easily get the dataset according to your project.

Before knowing the sources of the machine learning dataset, let's discuss datasets.

WHAT IS A DATASET?

- **A dataset** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table. This table shows an example of the dataset:

Country	Age	Salary	Purchased
India	38	48000	No
France	43	45000	Yes
Germany	30	54000	No
France	48	65000	No
Germany	40		Yes
India	35	58000	Yes

WHAT IS A DATASET?

- A tabular dataset can be understood as a database table or matrix, where each column corresponds to a **particular variable,** and each row corresponds to the **fields of the dataset.**

- The most supported file type for a tabular dataset is **"Comma Separated File,"** or **CSV.** But to store a "tree-like data," we can use the JSON file more efficiently.

TYPES OF DATA IN DATASETS

- **Numerical data:** Such as house price, temperature, etc.
- **Categorical data:** Such as Yes/No, True/False, Blue/green, etc.
- **Ordinal data:** These data are similar to categorical data but can be measured on the basis of comparison.

NEED OF DATASET

To work with machine learning projects, we need a huge amount of data, because, without the data, one cannot train ML/AI models. Collecting and preparing the dataset is one of the most crucial parts while creating an ML/AI project.

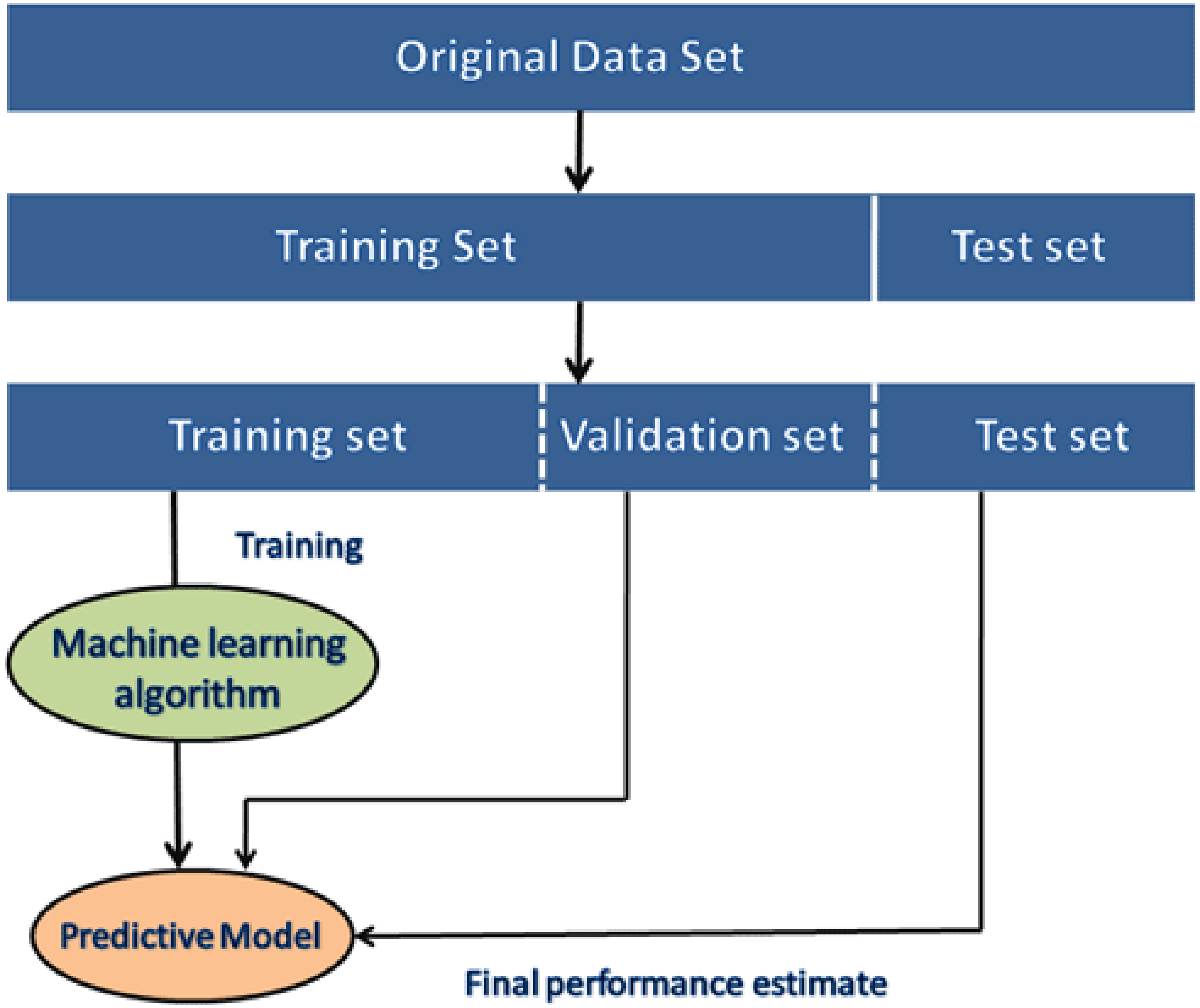
The technology applied behind any ML projects cannot work properly if the dataset is not well prepared and pre-processed.

During the development of the ML project, the developers completely rely on the datasets. In building ML applications, datasets are divided into two parts

Training dataset:

Test Dataset

NEED OF DATASET

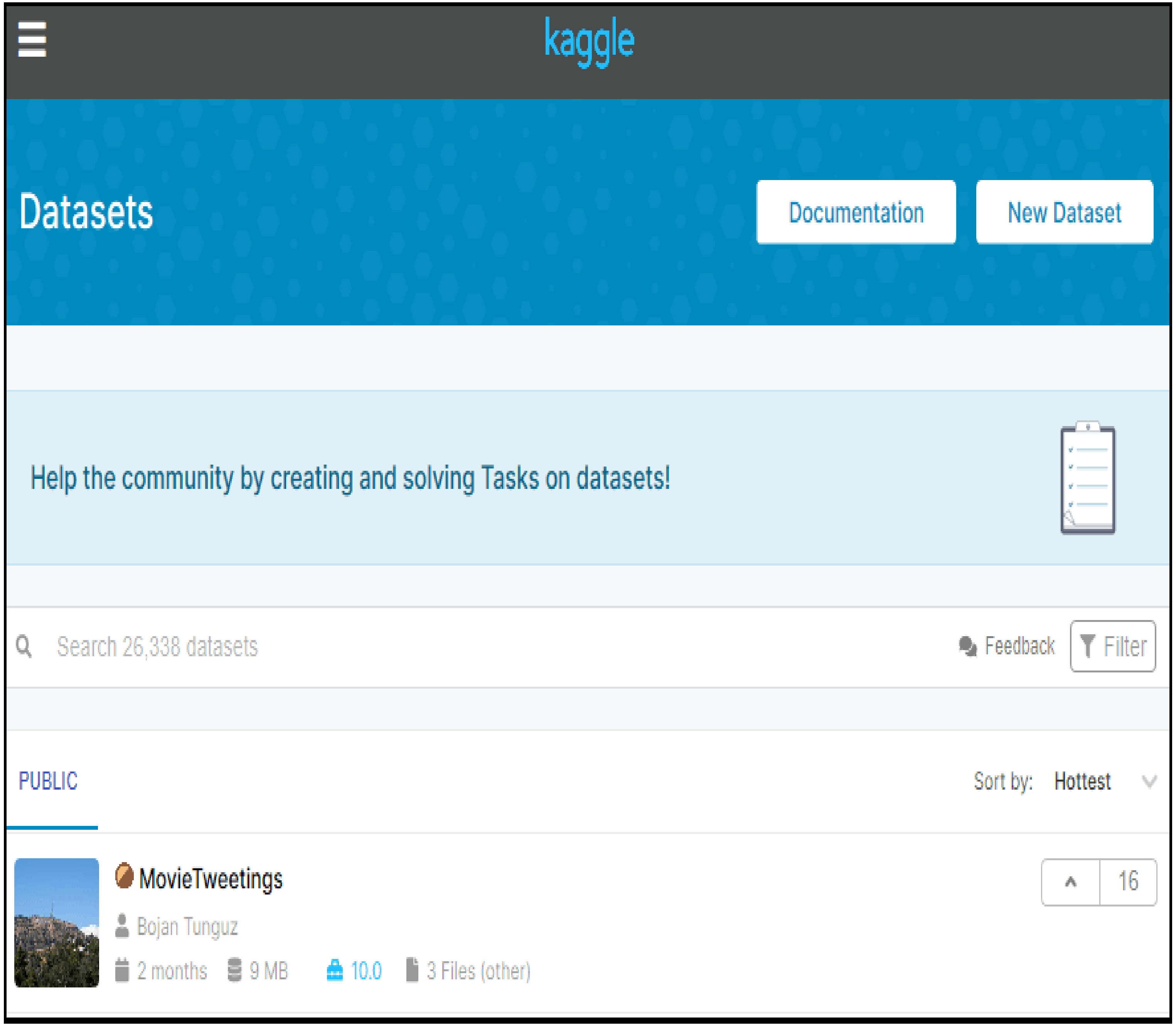


SOURCES FOR MACHINE LEARNING DATASETS

1 – Kaggle Datasets

- Kaggle is one of the best sources for providing datasets for Data Scientists and Machine Learners. It allows users to find, download, and publish datasets in an easy way.
- It also provides the opportunity to work with other machine learning engineers and solve difficult Data Science related tasks.
- Kaggle provides a high-quality dataset in different formats that we can easily find and download.
- The link for the Kaggle dataset is https://www.kaggle.com/datasets

KAGGLE DATASETS



Datasets

Help the community by creating and solving Tasks on datasets!

MovieTweetings
Bojan Tunguz
2 months 9 MB 10.0 3 Files (other)
16

UCI MACHINE LEARNING REPOSITORY

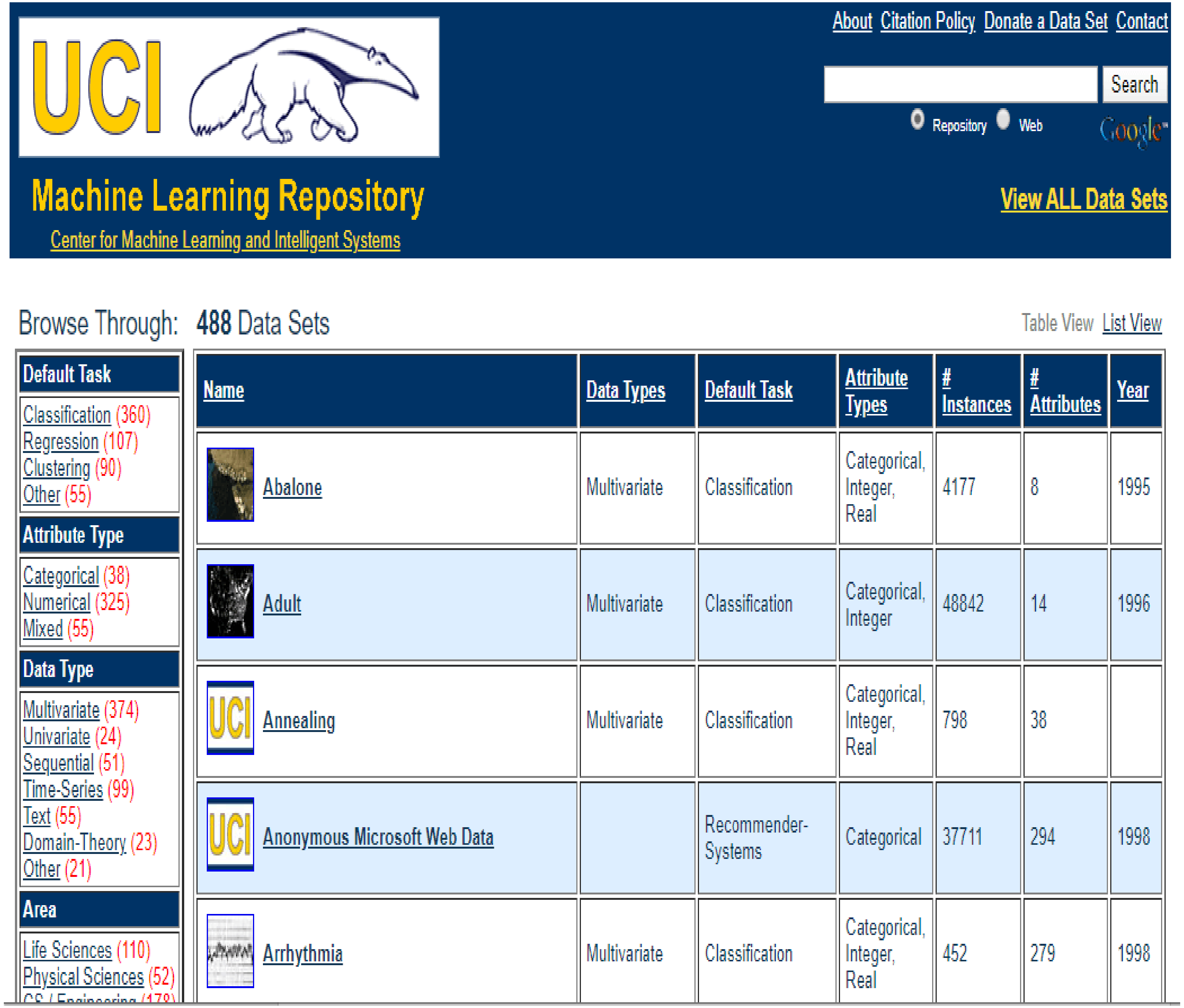
UCI Machine learning repository is one of the great sources of machine learning datasets. This repository contains databases, domain theories, and data generators that are widely used by the machine learning community for the analysis of ML algorithms.

Since the year 1987, it has been widely used by students, professors, researchers as a primary source of machine learning dataset.

It classifies the datasets as per the problems and tasks of machine learning such as Regression, Classification, Clustering, etc. It also contains some of the popular datasets such as the Iris dataset, Car Evaluation dataset, Poker Hand dataset, etc.

The link for the UCI machine learning repository is https://archive.ics.uci.edu/ml/index.php

UCI MACHINE LEARNING REPOSITORY



About Citation Policy Donate a Data Set Contact

Search

○ Repository ○ Web Google

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

View ALL Data Sets

Browse Through: 488 Data Sets

Table View List View

Default Task

- Classification (360)
- Regression (107)
- Clustering (90)
- Other (55)

Attribute Type

- Categorical (38)
- Numerical (325)
- Mixed (55)

Data Type

- Multivariate (374)
- Univariate (24)
- Sequential (51)
- Time-Series (99)
- Text (55)
- Domain-Theory (23)
- Other (21)

Area

- Life Sciences (110)
- Physical Sciences (52)
- CS / Engineering (178)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998

DATASETS VIA AWS

We can search, download, access, and share the datasets that are publicly available via AWS resources. These datasets can be accessed through AWS resources but provided and maintained by different government organizations, researches, businesses, or individuals.

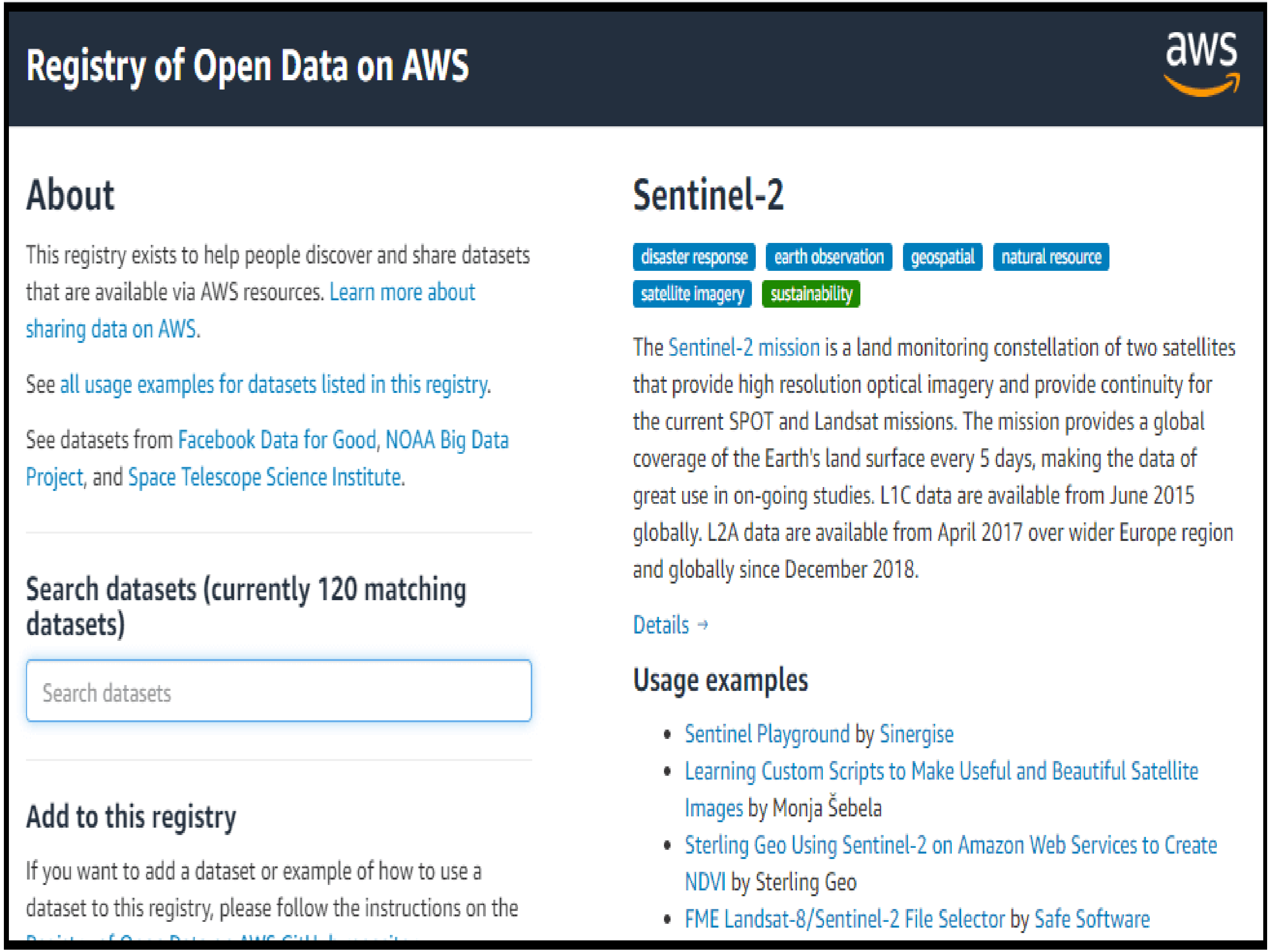
Anyone can analyze and build various services using shared data via AWS resources. The shared dataset on cloud helps users to spend more time on data analysis rather than on acquisitions of data

This source provides the various types of datasets with examples and ways to use the dataset. It also provides the search box using which we can search for the required dataset. Anyone can add any dataset or example to the Registry of Open Data on AWS

https://www.opendatani.gov.uk

The link for the resource is https://registry.opendata.aws/

DATASETS VIA AWS



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, NOAA Big Data Project, and Space Telescope Science Institute.

Search datasets (currently 120 matching datasets)

Search 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.

Sentinel-2

disaster response · earth observation · geospatial · natural resource · satellite imagery · sustainability

The Sentinel-2 mission is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region and globally since December 2018.

Details →

Usage examples

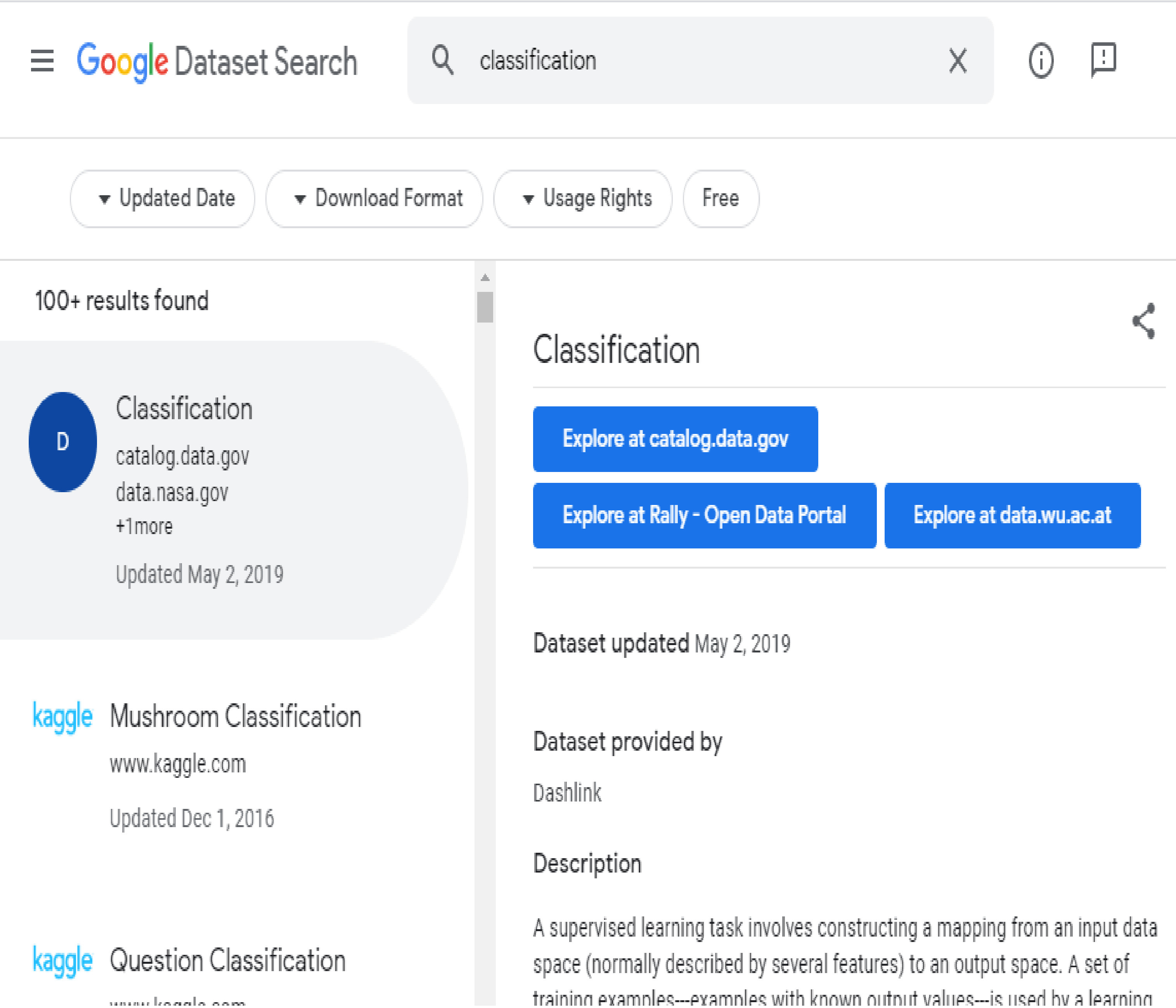
- Sentinel Playground by Sinergise
- Learning Custom Scripts to Make Useful and Beautiful Satellite Images by Monja Šebela
- Sterling Geo Using Sentinel-2 on Amazon Web Services to Create NDVI by Sterling Geo
- FME Landsat-8/Sentinel-2 File Selector by Safe Software

GOOGLE'S DATASET SEARCH ENGINE

Google dataset search engine is a search engine launched by Google on September 5, 2018. This source helps researchers to get online datasets that are freely available for use.

The link for the Google dataset search engine is https://toolbox.google.com/datasetsearch

GOOGLE'S DATASET SEARCH ENGINE



classification

100+ results found

Classification
catalog.data.gov
data.nasa.gov
+1more
Updated May 2, 2019

Mushroom Classification
www.kaggle.com
Updated Dec 1, 2016

Question Classification
www.kaggle.com

Classification

Dataset updated May 2, 2019

Dataset provided by

Dashlink

Description

A supervised learning task involves constructing a mapping from an input data space (normally described by several features) to an output space. A set of training examples---examples with known output values---is used by a learning

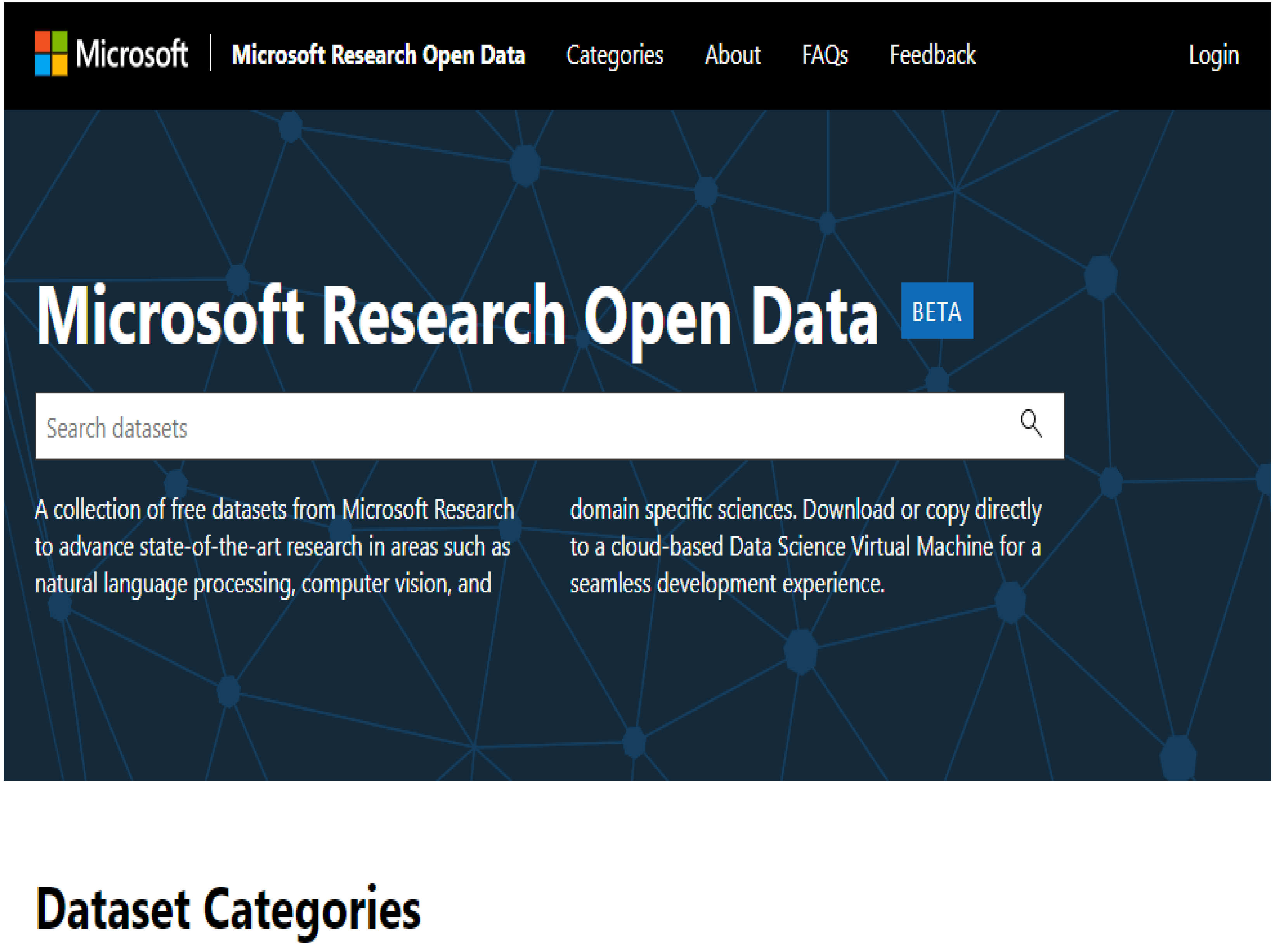
MICROSOFT DATASETS

The Microsoft has launched the **"Microsoft Research Open data"** repository with the collection of free datasets in various areas such as **natural language processing, computer vision, and domain-specific sciences.**

Using this resource, we can download the datasets to use on the current device, or we can also directly use it on the cloud infrastructure.

The link to download or use the dataset from this resource is https://msropendata.com/

MICROSOFT DATASETS



Microsoft | Microsoft Research Open Data

Microsoft Research Open Data BETA

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

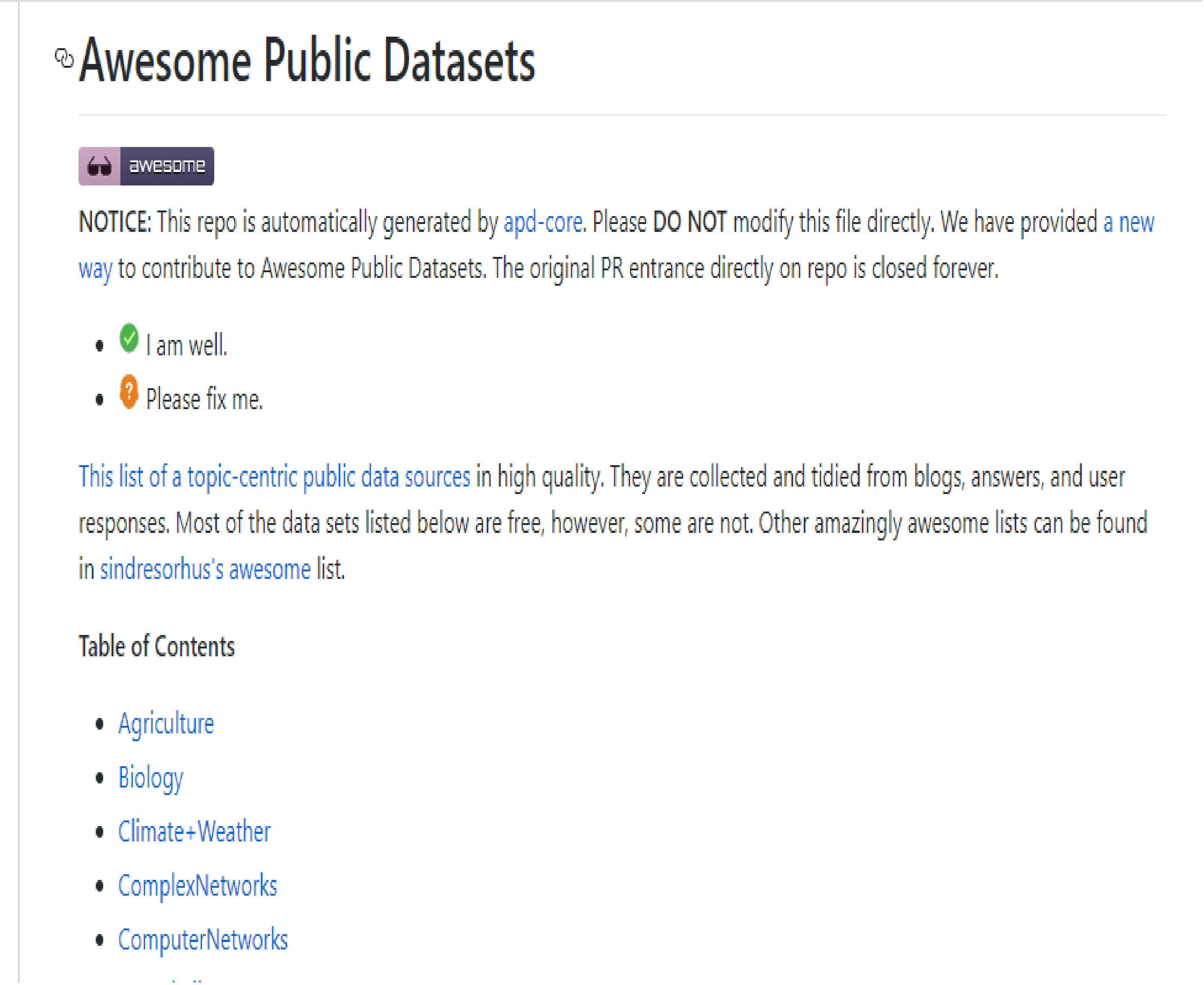
AWESOME PUBLIC DATASET COLLECTION

Awesome public dataset collection provides high-quality datasets that are arranged in a well-organized manner within a list according to topics such as Agriculture, Biology, Climate, Complex networks, etc.

Most of the datasets are available free, but some may not, so it is better to check the license before downloading the dataset.

The link to download the dataset from Awesome public dataset collection is https://github.com/awesomedata/awesome-public-datasets

AWESOME PUBLIC DATASET COLLECTION



Awesome Public Datasets

NOTICE: This repo is automatically generated by apd-core. Please **DO NOT** modify this file directly. We have provided a new way to contribute to Awesome Public Datasets. The original PR entrance directly on repo is closed forever.

- ✓ I am well.
- ? Please fix me.

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

Table of Contents

GOVERNMENT DATASETS

There are different sources to get government-related data. Various countries publish government data for public use collected by them from different departments.

The goal of providing these datasets is to increase transparency of government work among the people and to use the data in an innovative approach. Below are some links of government datasets:

Indian Government dataset

US Government Dataset

Northern Ireland Public Sector Datasets

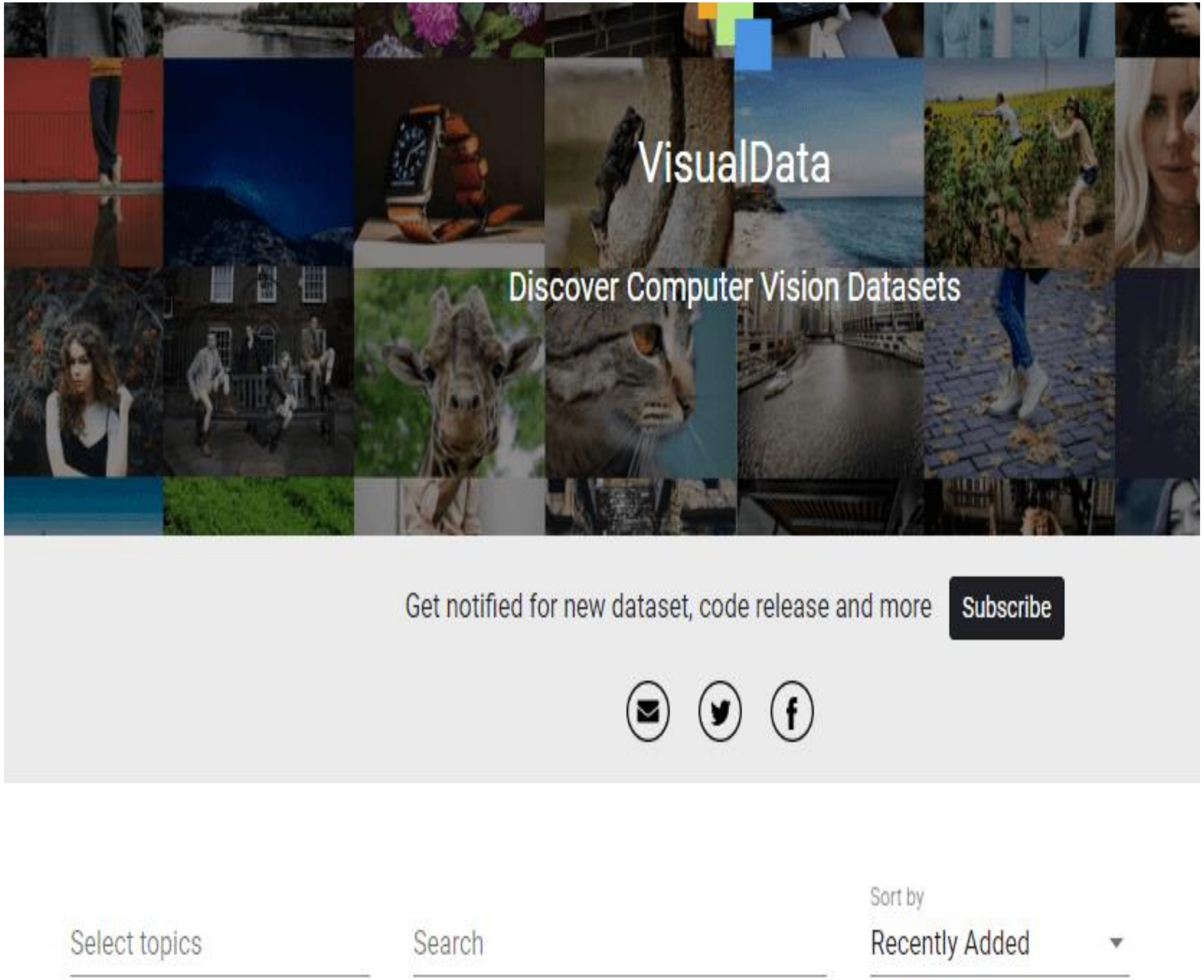
COMPUTER VISION DATASETS

Visual data provides multiple numbers of the great dataset that are specific to computer visions such as Image Classification, Video classification, Image Segmentation, etc.

Therefore, if you want to build a project on deep learning or image processing, then you can refer to this source.

The link for downloading the dataset from this source is https://www.visualdata.io/

COMPUTER VISION DATASETS



VisualData

Discover Computer Vision Datasets

Get notified for new dataset, code release and more

SCIKIT-LEARN DATASET

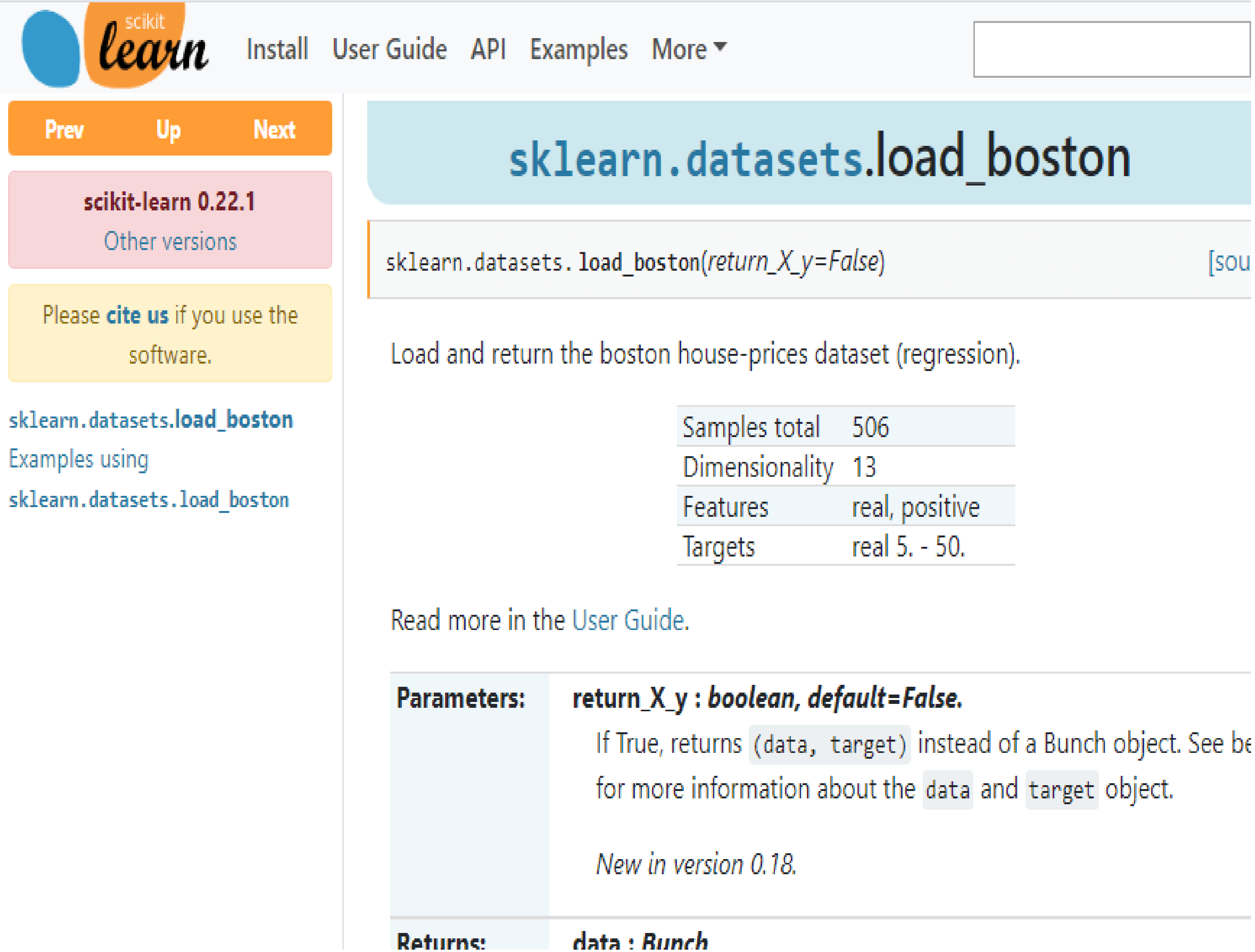
Scikit-learn is a great source for machine learning enthusiasts. This source provides both toy and real-world datasets.

These datasets can be obtained from sklearn.datasets package and using general dataset API.

The toy dataset available on scikit-learn can be loaded using some predefined functions such as, load_boston([return_X_y]), load_iris([return_X_y]), etc, rather than importing any file from external sources. But these datasets are not suitable for real-world projects.

The link to download datasets from this source is https://scikit-learn.org/stable/

SCIKIT-LEARN DATASET



scikit-learn 0.22.1

Other versions

Please **cite us** if you use the software.

`sklearn.datasets`.**load_boston**

Examples using `sklearn.datasets.load_boston`

`sklearn.datasets`.load_boston

`sklearn.datasets.` **load_boston**(*return_X_y=False*) [sou

Load and return the boston house-prices dataset (regression).

Samples total	506
Dimensionality	13
Features	real, positive
Targets	real 5. - 50.

Read more in the User Guide.

Parameters: **return_X_y : *boolean, default=False.***

If True, returns `(data, target)` instead of a Bunch object. See be for more information about the `data` and `target` object.

New in version 0.18.

Returns: **data : *Bunch***

THANK YOU