



REPOSITORIS DE DADES

#GRBIORetreat23



Outline

1. Link al Drive (per omplir).
2. Exemples
3. Idees?



Mini - resum

49 links, 8 membres, **ongoing**.

1. Dades públiques (obertes) vs. nosaltres recollim (Mortalitat basket,...).

2. Tipus de bases de dades:

- Webs (UCI Machine Learning, Kaggle,...).
- Repositoris (GitHub,...).
- Paquets de R (datasets, cluster.datasets,...).

1. Com organitzar-ho? Link en pàgina del GRBIO?
GitHub GRBIO?
2. Ho considereu útil?
2. Alguna informació més que hauriem de guardar?

Proporcionat per	Link	Descripció	Per què l'utilitzes? Tècnica/Mètode estadístic	Docència/Recerca/els dos/No l'he utilitzat mai	# de bbdd (aprox.)	Altres comentaris
------------------	------	------------	--	--	-----------------------	-------------------

3. Ideas de com tractar-ho?



UC Irvine Machine Learning Repository

The screenshot displays the UC Irvine Machine Learning Repository website. At the top, there is a navigation bar with links for 'Datasets', 'Contribute Dataset', and 'About Us', along with a search bar and a 'Log In' button. The main heading reads 'Welcome to the UC Irvine Machine Learning Repository', followed by a sub-heading: 'We currently maintain 623 datasets as a service to the machine learning community. Here, you can donate and find datasets used by millions of people all around the world!'. Below this are two prominent buttons: 'VIEW DATASETS' and 'CONTRIBUTE A DATASET'. The page is divided into two columns: 'Popular Datasets' and 'New Datasets'. Each column lists several datasets with their names, descriptions, and key statistics like the number of instances and attributes. The footer contains four sections: 'THE PROJECT' (with links to 'About Us', 'CML', and 'National Science Foundation'), 'NAVIGATION' (with links to 'Home', 'View Datasets', and 'Donate a Dataset'), 'LOGISTICS' (with links to 'Contact', 'Privacy Notice', and 'Feature Request or Bug Report'), and the UC Irvine Machine Learning Repository logo.

Popular Datasets

- Iris**
A small classic dataset from Fisher, 1936. One of the earliest datasets us...
Classification | 150 Instances | 4 Attributes
- Heart Disease**
4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach
Classification | 303 Instances | 13 Attributes
- Adult**
Predict whether income exceeds \$50K/yr based on census data. Also kn...
Classification | 48.84K Instances | 14 Attributes
- Dry Bean Dataset**
Images of 13,611 grains of 7 different registered dry beans were taken w...
Classification | 13.61K Instances | 17 Attributes
- Diabetes**
This diabetes dataset is from AIM '94
20 Attributes
- Rice (Cammeo and Osmancik)**
A total of 3810 rice grain's images were taken for the two species, proce...
Classification | 3.81K Instances | 8 Attributes

New Datasets

- Single elder home monitoring: Gas and position**
This dataset contains gas and temperature sensors as well as movement...
Classification | 444.63K Instances | 16 Attributes
- MetroPT-3 Dataset**
From a metro train in an operational context, readings from pressure, te...
Classification | 1.52M Instances | 15 Attributes
- HAR70+**
The Human Activity Recognition 70+ (HAR70+) dataset is a professionall...
Classification | 2.26M Instances | 6 Attributes
- HARTH**
The Human Activity Recognition Trondheim (HARTH) dataset is a profess...
Classification | 6.46M Instances | 8 Attributes
- DeFungi**
DeFungi is a dataset for direct mycological examination of microscopic f...
Classification | 9.11K Instances
- NASA Flood Extent Detection**
This dataset contains synthetic aperture radar (SAR) raster imagery for v...
Other | 50K Instances

THE PROJECT

- About Us
- CML
- National Science Foundation

NAVIGATION

- Home
- View Datasets
- Donate a Dataset

LOGISTICS

- Contact
- Privacy Notice
- Feature Request or Bug Report



Awesome data set - GitHub Repositori

The screenshot shows the GitHub repository page for 'awesome-public-datasets' by the user 'awesomedata'. The repository is public and has 2.3k watchers. The current branch is 'master'. A recent commit by 'caesar0301' is visible, titled 'Update README sha: 842c61e', made 20 hours ago. The repository contains a file named 'Datasets' with a commit message 'Add titanic dataset' from 9 years ago.

awesomedata / awesome-public-datasets

Code Issues 62 Pull requests 60 Actions Security

awesome-public-datasets Public Watch 2.3k

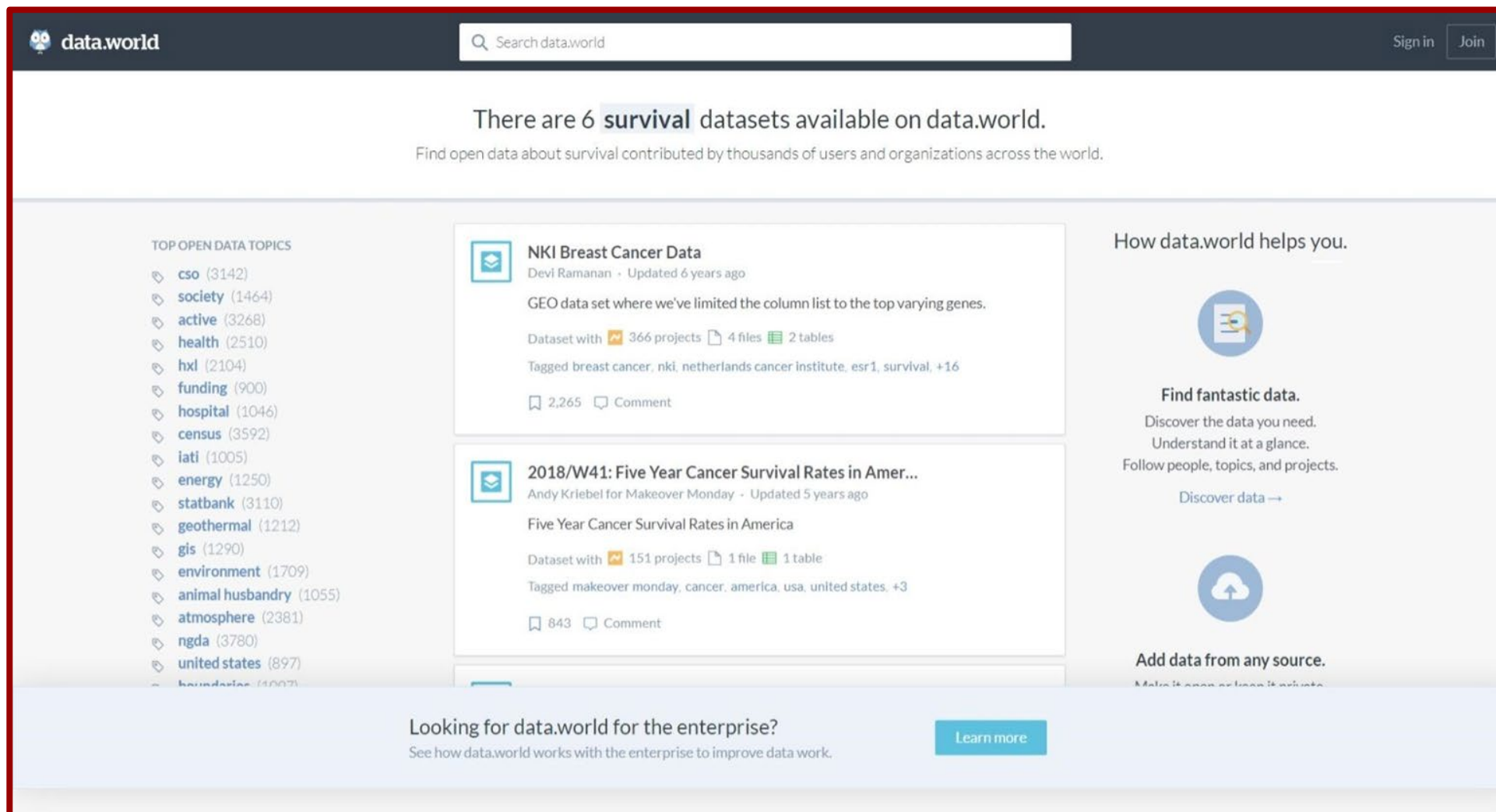
master Go to file Add file <> C

branches Tags

caesar0301 Update README sha: 842c61e 20 hours ago

Datasets Add titanic dataset 9 year

Survival datasets (dintre del dataworld)



data.world Search data.world Sign in Join

There are 6 **survival** datasets available on data.world.
Find open data about survival contributed by thousands of users and organizations across the world.

TOP OPEN DATA TOPICS

- cs0 (3142)
- society (1464)
- active (3268)
- health (2510)
- hxl (2104)
- funding (900)
- hospital (1046)
- census (3592)
- lati (1005)
- energy (1250)
- statbank (3110)
- geothermal (1212)
- gis (1290)
- environment (1709)
- animal husbandry (1055)
- atmosphere (2381)
- ngda (3780)
- united states (897)
- boundaries (1007)

NKI Breast Cancer Data
Devi Ramanan · Updated 6 years ago
GEO data set where we've limited the column list to the top varying genes.
Dataset with 366 projects 4 files 2 tables
Tagged breast cancer, nki, netherlands cancer institute, esr1, survival, +16
2,265 Comment

2018/W41: Five Year Cancer Survival Rates in Amer...
Andy Kriebel for Makeover Monday · Updated 5 years ago
Five Year Cancer Survival Rates in America
Dataset with 151 projects 1 file 1 table
Tagged makeover monday, cancer, america, usa, united states, +3
843 Comment

How data.world helps you.

Find fantastic data.
Discover the data you need.
Understand it at a glance.
Follow people, topics, and projects.
Discover data →

Add data from any source.
Make it open or keep it private.

Looking for data.world for the enterprise?
See how data.world works with the enterprise to improve data work. [Learn more](#)

Competitions

Datasets

Models

Code

Discussions

Courses

...



Search

k

Predict Malicious Websites: XGBoost *Draft saved*

File Edit Insert Run View Help

```
data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)
```

more than
cursor

1. Com organitzar-ho? Link en pàgina del GRBIO?
GitHub GRBIO?
2. Ho considereu útil?
2. Alguna informació més que hauriem de guardar?

Proporcionat per	Link	Descripció	Per què l'utilitzes? Tècnica/Mètode estadístic	Docència/Recerca/els dos/No l'he utilitzat mai	# de bbdd (aprox.)	Altres comentaris
------------------	------	------------	--	--	-----------------------	-------------------

3. Idees de com tractar-ho?