
Programming and frameworks for ML

Where can I find a Dataset?



About Me

Big Data Consultant at Indra / Big Data Lecturer

- More than 20 years of experience in different environments, technologies, customers, countries ...
- Passionate data and technology
- Enthusiastic Big Data world and NoSQL



Daniel Villanueva Jiménez

BigData Developer / Lecturer

INDRA • Universidad Pontificia de Salamanca



Public Datasets



Awesome Public Datasets



This list of [public data sources](#) 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 the [awesome-awesomeness](#) and [sindresorhus's awesome](#) list.

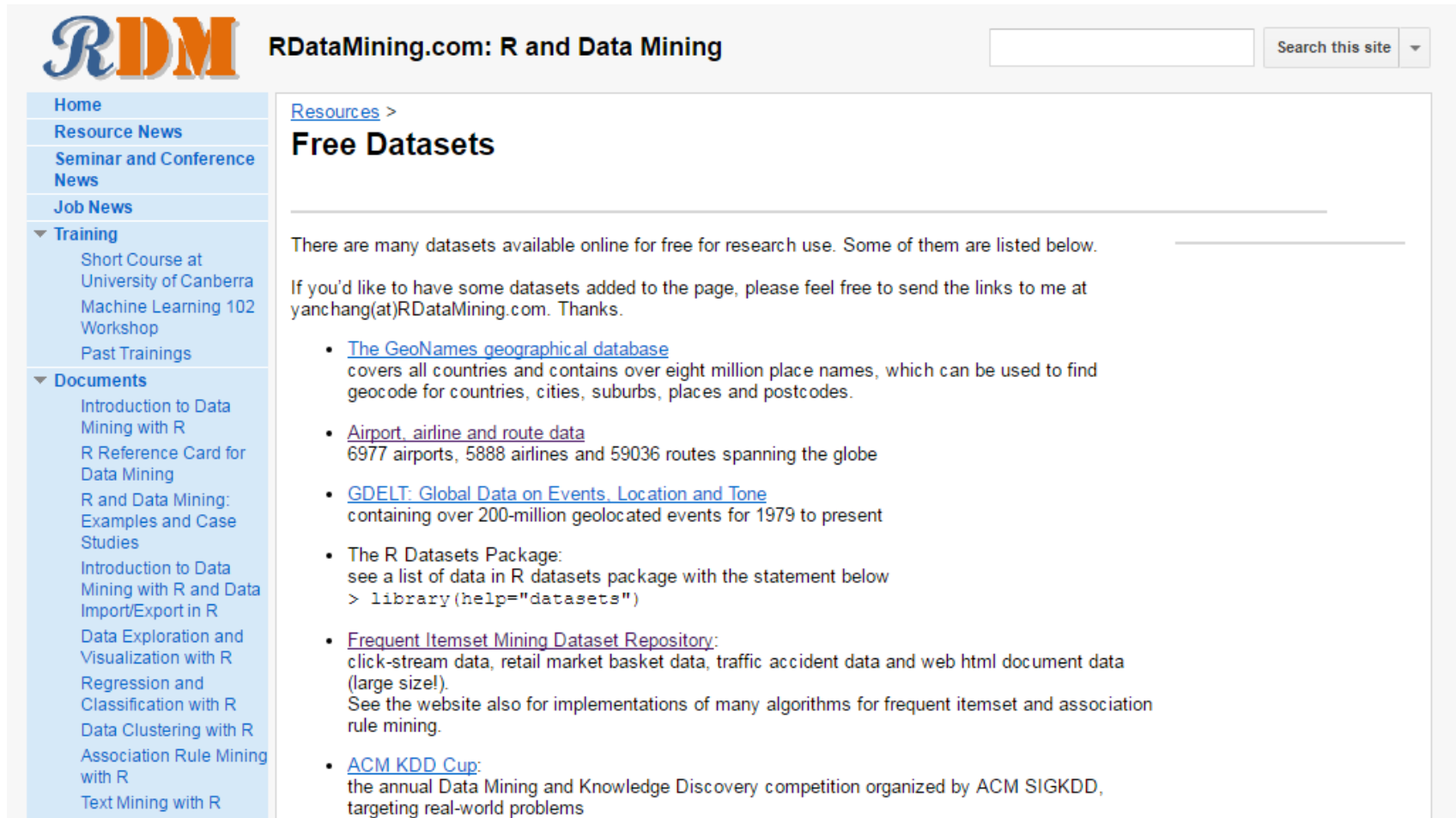
Table of Contents

- [Agriculture](#)
- [Biology](#)
- [Climate/Weather](#)
- [Complex Networks](#)
- [Computer Networks](#)
- [Contextual Data](#)
- [Data Challenges](#)
- [Earth Science](#)
- [Economics](#)
- [Education](#)
- [Energy](#)
- [Finance](#)
- [GIS](#)



<https://github.com/caesar0301/awesome-public-datasets>

Public Datasets



The screenshot shows the RDataMining.com website. The header includes the 'RDM' logo, the site name 'RDataMining.com: R and Data Mining', and a search bar. A left sidebar contains a navigation menu with links to Home, Resource News, Seminar and Conference News, Job News, Training (with sub-links for Short Course at University of Canberra, Machine Learning 102 Workshop, and Past Trainings), and Documents (with sub-links for Introduction to Data Mining with R, R Reference Card for Data Mining, R and Data Mining: Examples and Case Studies, Introduction to Data Mining with R and Data Import/Export in R, Data Exploration and Visualization with R, Regression and Classification with R, Data Clustering with R, Association Rule Mining with R, and Text Mining with R). The main content area is titled 'Resources > Free Datasets'. It contains a paragraph stating that many datasets are available online for free for research use, followed by a request for users to send links to yanchang(at)RDataMining.com. Below this is a list of five resources: The GeoNames geographical database, Airport, airline and route data, GDELT: Global Data on Events, Location and Tone, The R Datasets Package, and Frequent Itemset Mining Dataset Repository. The list ends with the ACM KDD Cup.

RDM RDataMining.com: R and Data Mining

Home
Resource News
Seminar and Conference News
Job News
▼ Training
 Short Course at University of Canberra
 Machine Learning 102 Workshop
 Past Trainings
▼ Documents
 Introduction to Data Mining with R
 R Reference Card for Data Mining
 R and Data Mining: Examples and Case Studies
 Introduction to Data Mining with R and Data Import/Export in R
 Data Exploration and Visualization with R
 Regression and Classification with R
 Data Clustering with R
 Association Rule Mining with R
 Text Mining with R


[Resources](#) >
Free Datasets


There are many datasets available online for free for research use. Some of them are listed below.

If you'd like to have some datasets added to the page, please feel free to send the links to me at [yanchang\(at\)RDataMining.com](mailto:yanchang(at)RDataMining.com). Thanks.

- [The GeoNames geographical database](#)
covers all countries and contains over eight million place names, which can be used to find geocode for countries, cities, suburbs, places and postcodes.
- [Airport, airline and route data](#)
6977 airports, 5888 airlines and 59036 routes spanning the globe
- [GDELT: Global Data on Events, Location and Tone](#)
containing over 200-million geolocated events for 1979 to present
- The R Datasets Package:
see a list of data in R datasets package with the statement below
> `library(help="datasets")`
- [Frequent Itemset Mining Dataset Repository](#):
click-stream data, retail market basket data, traffic accident data and web html document data (large size!).
See the website also for implementations of many algorithms for frequent itemset and association rule mining.
- [ACM KDD Cup](#):
the annual Data Mining and Knowledge Discovery competition organized by ACM SIGKDD, targeting real-world problems

Public Datasets

 Menu



Products ▾

Solutions

Pricing

Software

Support

More ▾

English ▾

My Account ▾

Create an AWS Account

DATASET CATEGORIES

Astronomy >

Biology >

Chemistry >

Climate >

Economics >

Encyclopedic >

Geographic >

Mathematics >

RELATED LINKS

Amazon Machine Images (AMIs)

AWS Public Datasets

Public Datasets on AWS provides a centralized repository of public datasets that can be seamlessly integrated into AWS cloud-based applications. AWS is hosting the public datasets at no charge for the community, and like all AWS services, users pay only for the compute and storage they use for their own applications. Learn more about [Public Datasets on AWS](#).

A Multi-wavelength Infrared Atlas of the Galactic Plane

Open Source tools were used to combine images from five major infrared surveys of the Galactic Plane, archived at the NASA/IPAC Infrared Science Archive (IRSA). The result is a 16-wavelength infrared Atlas of the Galactic Plane that covers the wavelength range 1 μm to 24 μm .

Last Modified: March 9, 2016



<https://aws.amazon.com/es/datasets/>

Public Datasets

DATASETS

[HOT](#)
[NEW](#)
[RISING](#)
[TOP](#)
[GILDED](#)
[WIKI](#)

↑

7

↓

We here at Reddit Gifts would like to make sure that everyone drinks their daily recommended amount of water, stays well hydrated, and remains environmentally conscious! Join us for the **BRAND NEW Water Bottles Exchange!**

Promoted by reddit_exchanges

promoted save hide report pocket

↑

5

↓

March | Monthly discussion thread | 09, 2017 self.datasets

Submitted 6 days ago by **AutoModerator** M - announcement

2 comments share save hide report pocket

↑


18

↓

discussion | **Are there any tools to manage the meta data of my data sets?** self.datasets

Submitted 6 days ago * by **data999** - announcement

28 comments share save hide report pocket




DATASETS

23,748 readers

[Unsubscribe](#)

[Submit Link](#)
[Submit Text](#)





<https://www.reddit.com/r/datasets/>

Public Datasets

Package	Item	Title	csv	doc
datasets	AirPassengers	Monthly Airline Passenger Numbers 1949-1960	CSV	DOC
datasets	BJsales	Sales Data with Leading Indicator	CSV	DOC
datasets	BOD	Biochemical Oxygen Demand	CSV	DOC
datasets	CO2	Carbon Dioxide Uptake in Grass Plants	CSV	DOC
datasets	Formaldehyde	Determination of Formaldehyde	CSV	DOC
datasets	HairEyeColor	Hair and Eye Color of Statistics Students	CSV	DOC
datasets	InsectSprays	Effectiveness of Insect Sprays	CSV	DOC
datasets	JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share	CSV	DOC
datasets	LakeHuron	Level of Lake Huron 1875-1972	CSV	DOC
datasets	LifeCycleSavings	Intercountry Life-Cycle Savings Data	CSV	DOC
datasets	Nile	Flow of the River Nile	CSV	DOC
datasets	OrchardSprays	Potency of Orchard Sprays	CSV	DOC
datasets	PlantGrowth	Results from an Experiment on Plant Growth	CSV	DOC
datasets	Puromycin	Reaction Velocity of an Enzymatic Reaction	CSV	DOC
datasets	Titanic	Survival of passengers on the Titanic	CSV	DOC
datasets	ToothGrowth	The Effect of Vitamin C on Tooth Growth in Guinea Pigs	CSV	DOC
datasets	UCBAdmissions	Student Admissions at UC Berkeley	CSV	DOC
datasets	UKDriverDeaths	Road Casualties in Great Britain 1969-84	CSV	DOC
datasets	UKgas	UK Quarterly Gas Consumption	CSV	DOC
datasets	USAccDeaths	Accidental Deaths in the US 1973-1978	CSV	DOC
datasets	USArrests	Violent Crime Rates by US State	CSV	DOC
datasets	USJudgeRatings	Lawyers' Ratings of State Judges in the US Superior Court	CSV	DOC
datasets	USPersonalExpenditure	Personal Expenditure Data	CSV	DOC
datasets	VADeaths	Death Rates in Virginia (1940)	CSV	DOC
datasets	WWWusage	Internet Usage per Minute	CSV	DOC

Public Datasets

By Jure Leskovec

STANFORD
UNIVERSITY



- SNAP for C++ ▶
- SNAP for Python ▶
- SNAP Datasets ▶
- What's new
- People
- Papers
- Projects ▶
- Citing SNAP
- Links
- About
- Contact us

Open positions

We have filled all the positions for this quarter.

Stanford Large Network Dataset Collection

- **Social networks** : online social networks, edges represent interactions between people
- **Networks with ground-truth communities** : ground-truth network communities in social and information networks
- **Communication networks** : email communication networks with edges representing communication
- **Citation networks** : nodes represent papers, edges represent citations
- **Collaboration networks** : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- **Web graphs** : nodes represent webpages and edges are hyperlinks
- **Amazon networks** : nodes represent products and edges link commonly co-purchased products
- **Internet networks** : nodes represent computers and edges communication
- **Road networks** : nodes represent intersections and edges roads connecting the intersections
- **Autonomous systems** : graphs of the internet
- **Signed networks** : networks with positive and negative edges (friend/foe, trust/distrust)
- **Location-based online social networks** : Social networks with geographic check-ins
- **Wikipedia networks, articles, and metadata** : Talk, editing, voting, and article data from Wikipedia
- **Temporal networks** : networks where edges have timestamps
- **Twitter and Memetracker** : Memetracker phrases, links and 467 million Tweets
- **Online communities** : Data from online communities such as Reddit and Flickr
- **Online reviews** : Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also available from [UF Sparse Matrix collection](#). [Visualizations of SNAP networks](#) by Tim Davis.

• Social networks



<https://snap.stanford.edu/data/>

Public Datasets



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)







☐ Repository ☐ Web



[View ALL Data Sets](#)

Browse Through: 394 Data Sets

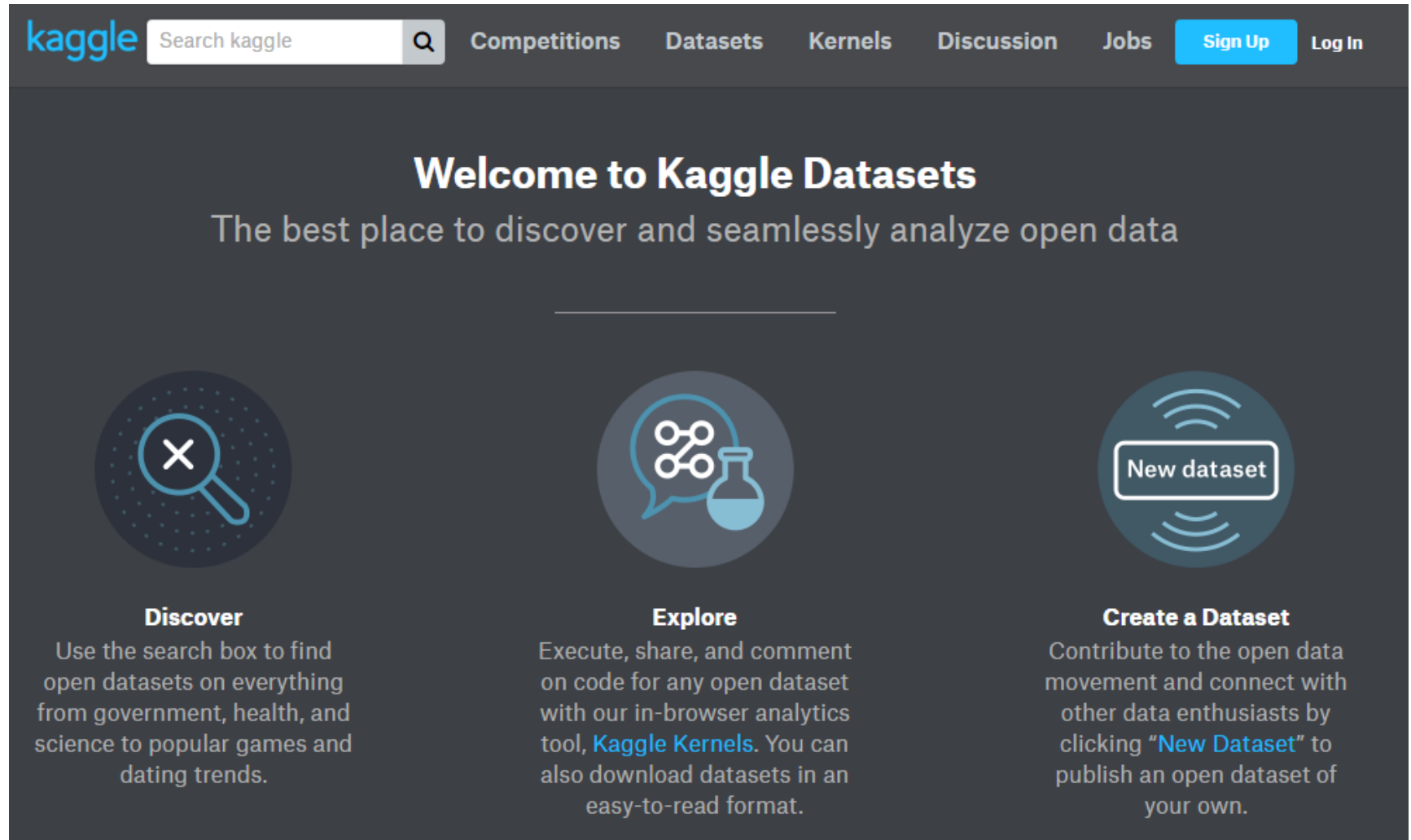
Table View [List View](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (289) Regression (74) Clustering (67) Other (54)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Attribute Type Categorical (37) Numerical (244) Mixed (55)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Data Type Multivariate (306) Univariate (16) Sequential (40) Time-Series (75) Text (37) Domain-Theory (22) Other (21)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Area Life Sciences (89) Physical Sciences (47) CS / Engineering (129) Social Sciences (23) Business (25)	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
	 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
	 Aa			Categorical			



<https://archive.ics.uci.edu/ml/datasets.html>


Public Datasets

The image is a screenshot of the Kaggle Datasets website. At the top, there is a dark navigation bar with the Kaggle logo on the left, a search bar with the text "Search kaggle" and a magnifying glass icon, and several menu items: "Competitions", "Datasets", "Kernels", "Discussion", "Jobs", "Sign Up" (in a blue button), and "Log In". Below the navigation bar, the main content area has a dark background. It features a large heading "Welcome to Kaggle Datasets" in white, followed by the tagline "The best place to discover and seamlessly analyze open data". Below this, there are three circular icons representing different actions: a magnifying glass for "Discover", a network diagram with a flask for "Explore", and a plus sign with a box labeled "New dataset" for "Create a Dataset". Each icon is accompanied by a title and a descriptive paragraph. The "Discover" section says to use the search box to find open datasets from various sources. The "Explore" section describes executing, sharing, and commenting on code using Kaggle Kernels. The "Create a Dataset" section encourages contributing to the open data movement by publishing a new dataset.

kaggle Search kaggle Q Competitions Datasets Kernels Discussion Jobs Sign Up Log In


Welcome to Kaggle Datasets

The best place to discover and seamlessly analyze open data




Discover

Use the search box to find open datasets on everything from government, health, and science to popular games and dating trends.



Explore

Execute, share, and comment on code for any open dataset with our in-browser analytics tool, [Kaggle Kernels](#). You can also download datasets in an easy-to-read format.



Create a Dataset

Contribute to the open data movement and connect with other data enthusiasts by clicking "[New Dataset](#)" to publish an open dataset of your own.



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

Public Datasets

Portal de Transparencia

Participación

Comunidad de Madrid

PLAN DE GOBIERNO

INFORMACIÓN INSTITUCIONAL

INFORMACIÓN JURÍDICA

INFORMACIÓN ECONÓMICA Y ESTADÍSTICA

CANAL DE ACCESO A LA INFORMACIÓN PÚBLICA

Acerca de Transparencia

Datos estadísticos

La Comunidad de Madrid ofrece información detallada sobre las estadísticas de su actividad en distintos ámbitos: Memorias anuales, informes, indicadores, resultados o cartas de servicios.

THANKS FOR YOUR ATTENTION

Daniel Villanueva Jiménez

daniel.villanueva@immune.institute

@dvillaj

