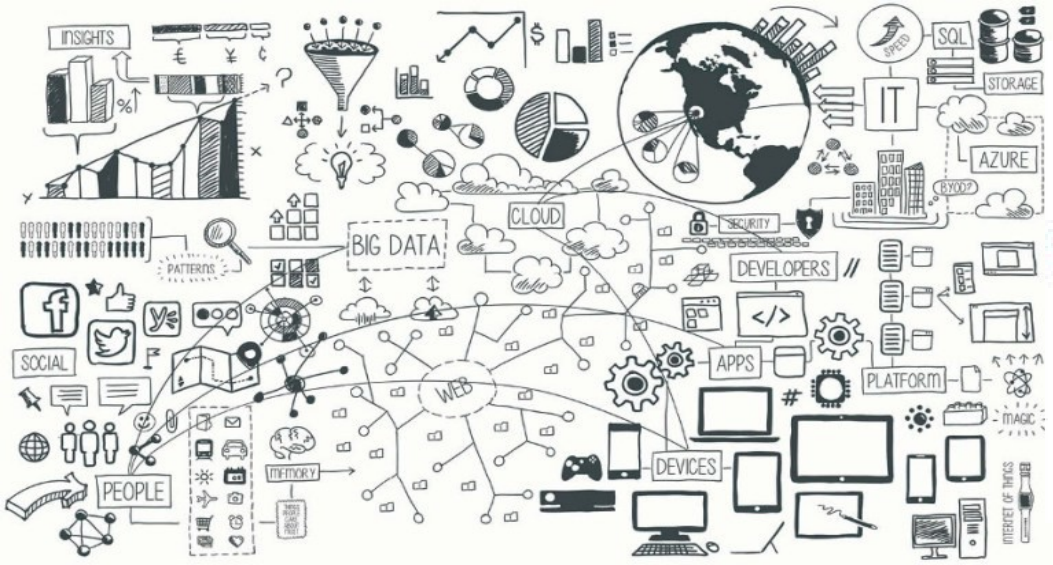
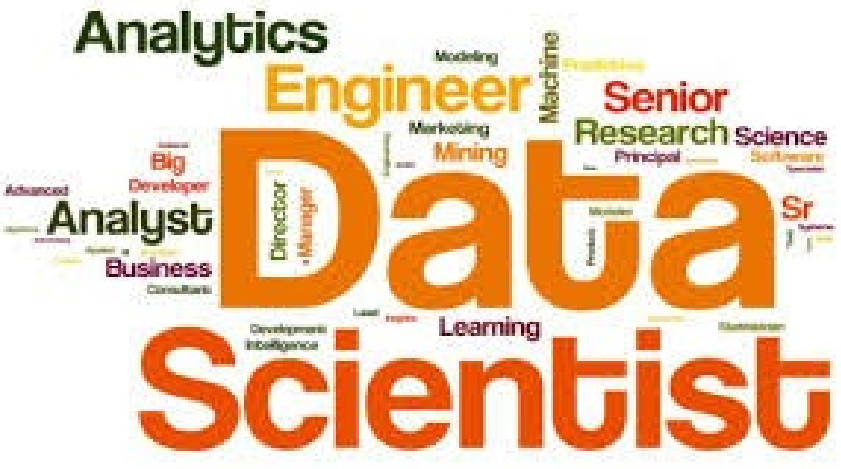
Data Mining (Minería de Datos)

DATASETS & DATA CHALLENGES





Maialen Iturbide

José Manuel Gutiérrez

Grupo de Meteorología

Univ. de Cantabria – CSIC
MACC / IFCA

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Browse Through: **394** Data Sets

Table View List View

Default Task
Classification (289)
Regression (74)
Clustering (67)
Other (54)

Attribute Type
Categorical (37)
Numerical (244)
Mixed (55)

Data Type
Multivariate (306)
Univariate (16)
Sequential (40)
Time-Series (75)
Text (37)
Domain-Theory (22)
Other (21)

Area
Life Sciences (89)
Physical Sciences (47)
CS / Engineering (129)
Social Sciences (23)
Business (25)
Game (10)
Other (67)

 # Attributes
Less than 10 (90)
10 to 100 (182)
Greater than 100 (67)

 # Instances
Less than 100 (19)
100 to 1000 (137)
Greater than 1000 (206)

Format Type
Matrix (275)
Non-Matrix (119)

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
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987
Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992
Auto MPG	Multivariate	Regression	Categorical, Real	398	8	1993
Automobile	Multivariate	Regression	Categorical, Integer, Real	205	26	1987

Most Popular Data Sets (hits since 2007):

1561726: Iris
1019633: Adult
775959: Wine
666464: Car Evaluation
597387: Breast Cancer Wisconsin (Diagnostic)
582860: Forest Fires
559491: Human Activity Recognition Using Smartphones
544173: Heart Disease
538395: Wine Quality

http://archive.ics.uci.edu/ml/datasets.html

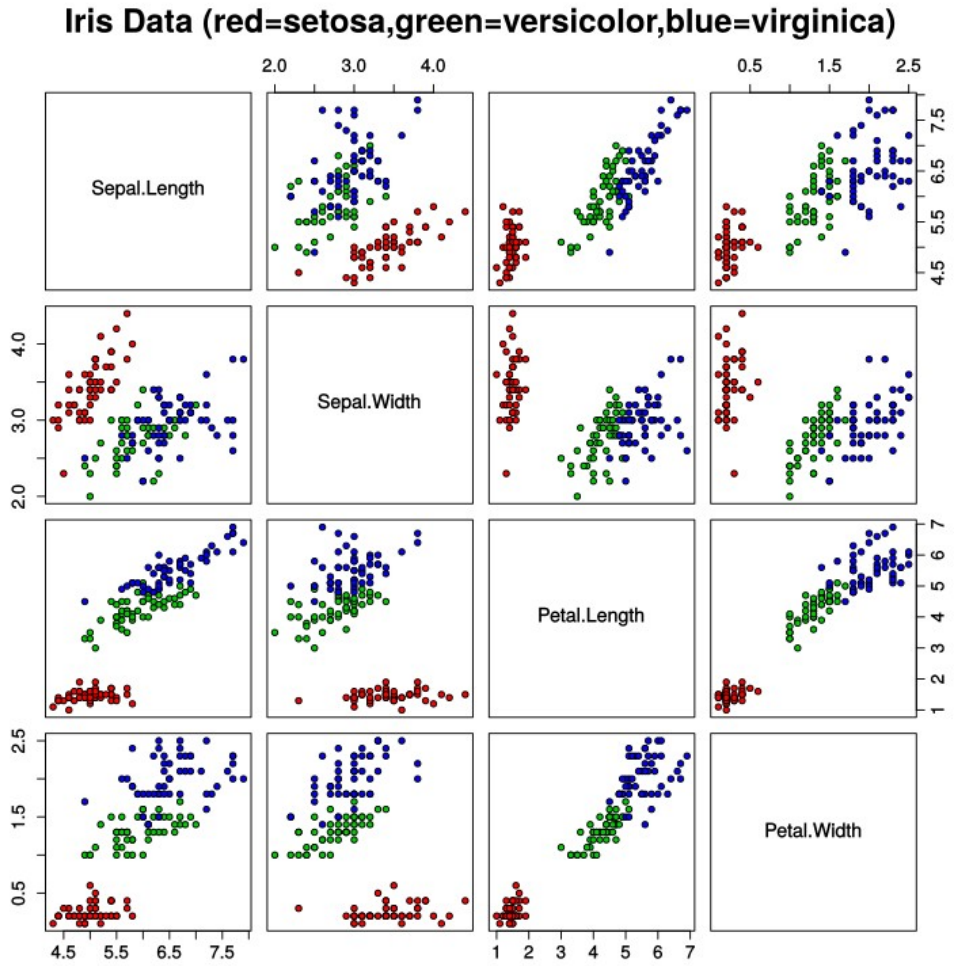
Iris Data Set

Download: Data Folder, Data Set Description

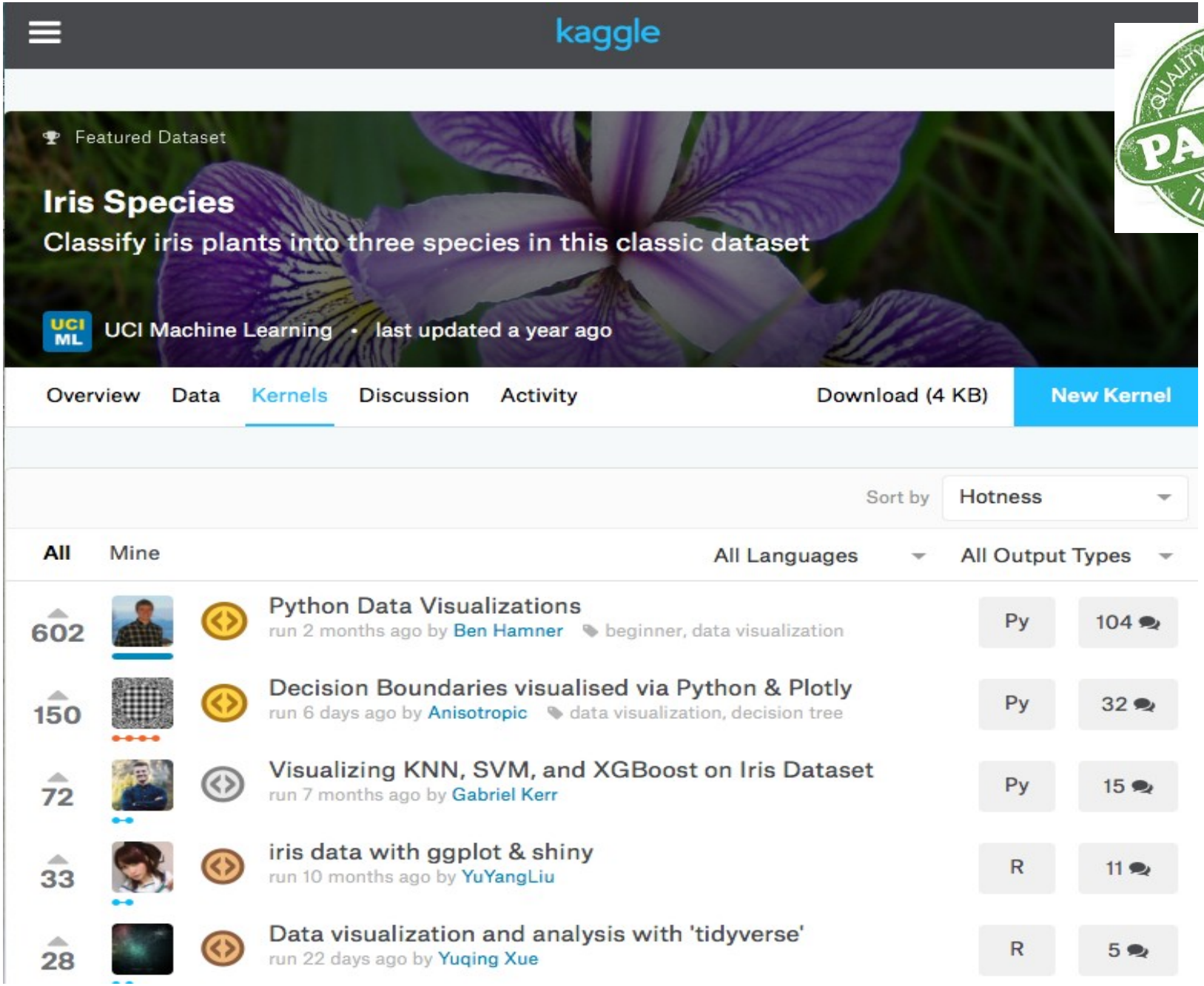
Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1549312

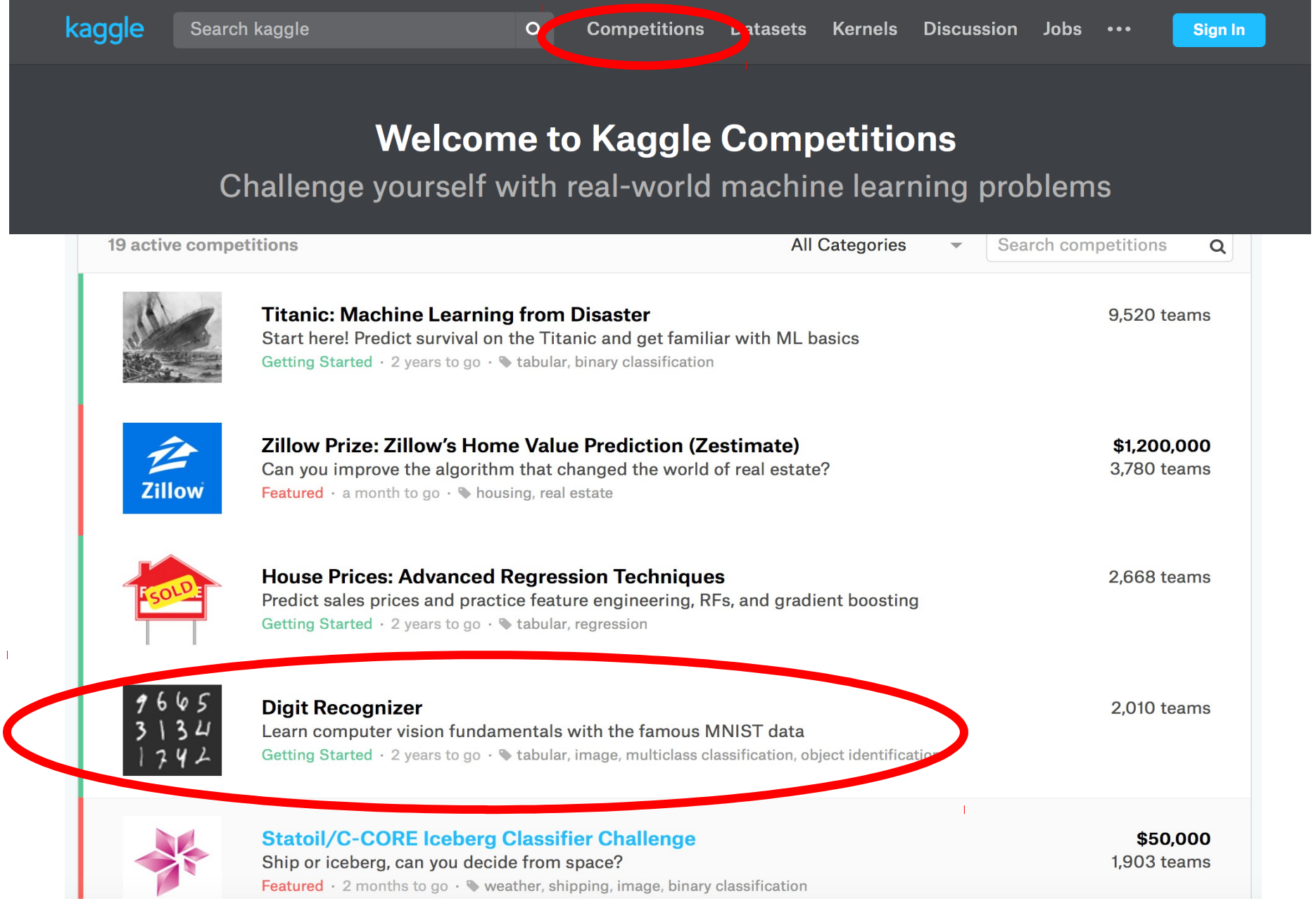
http://archive.ics.uci.edu/ml/datasets/Iris




```
?iris
data(iris)
str(iris)

plot(iris$Petal.Length, iris$Petal.Width,
main="Edgar Anderson's Iris Data")

library(lattice)

xyplot(Petal.Width~Petal.Length, data = iris)

xyplot(Petal.Width~Petal.Length|Species, data = iris)

xyplot(Petal.Width~Petal.Length,
group = Species, data = iris, auto.key = TRUE)

cloud(Sepal.Length ~ Sepal.Length * Petal.Width,
group = Species, data = iris, auto.key = TRUE)

...
```


kaggle
Featured Dataset
Iris Species
Classify iris plants into three species in this classic dataset
UCI ML
UCI Machine Learning • last updated a year ago
Overview
Data
Kernels
Discussion
Activity
Download (4 KB)
New Kernel
Sort by
Hotness
All
Mine
All Languages
All Output Types
602
Python Data Visualizations
run 2 months ago by Ben Hamner
beginner, data visualization
Py
104
150
Decision Boundaries visualised via Python & Plotly
run 6 days ago by Anisotropic
data visualization, decision tree
Py
32
72
Visualizing KNN, SVM, and XGBoost on Iris Dataset
run 7 months ago by Gabriel Kerr
Py
15
33
iris data with ggplot & shiny
run 10 months ago by YuYangLiu
R
11
28
Data visualization and analysis with 'tidyverse'
run 22 days ago by Yuqing Xue
R
5


QUALITY CONTROL
PASSED
INSPECTED



https://www.kaggle.com/c/digit-recognizer/leaderboard

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs ...

Overview Data Kernels Discussion Leaderboard Rules **Late Submission**

Overview

Description
Evaluation
Prizes
Timeline

In this competition, Kaggle is challenging you to build a model that predicts the total ride duration of taxi trips in New York City. Your primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

Longtime Kagglers will recognize that this competition objective is similar to the ECML/PKDD trip time challenge we hosted in 2015. But, this challenge comes with a twist. Instead of awarding prizes to the top finishers on the leaderboard, this playground competition was created to reward collaboration and collective learning.

We are encouraging you (with cash prizes!) to publish additional training data that other participants can use for their predictions. We also have designated bi-weekly and final prizes to reward authors of kernels that are particularly insightful or valuable to the community.

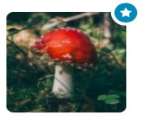
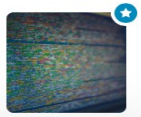
https://www.kaggle.com/headsortails/nyc-taxi-eda-update-the-fast-the-curious/notebook

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs ••• Sign In

Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems

5,361 Datasets | Sizes | File types | Licenses | Tags | Search datasets

490		**Global Terrorism Database** More than 170,000 terrorist attacks worldwide, 1970-2016 **START Consortium** updated 5 months ago	crime terrorism international relati...	CSV 144 MB Other	566 33 135k
364		**World Development Indicators** Explore country development indicators from around the world **World Bank** updated 7 months ago	economics international relati...	CSV 2 GB Other	381 36 118k
209		**Mushroom Classification** Safe to eat or deadly poison? **UCI Machine Learning** updated a year ago	food and drink human medicine plants	CSV 365 KB CC0	365 14 69k
120		**Getting Real about Fake News** Text & metadata from fake & biased news sources around the web **Megan Risdal** updated a year ago	news agencies languages politics	CSV 54 MB CC0	58 10 68k
110		**Wine Reviews** 130k wine reviews with variety, location, winery, price, and description **zackthoutt** updated 7 days ago	critical theory food and drink	CSV 51 MB CC4	24 4 23k
474		**Kaggle ML and Data Science Survey, 2017** A big picture view of the state of data science and machine learning. **Kaggle** updated a month ago	data analysis employment sociology artificial intelligen...	CSV 28 MB ODbL	132 8 73k

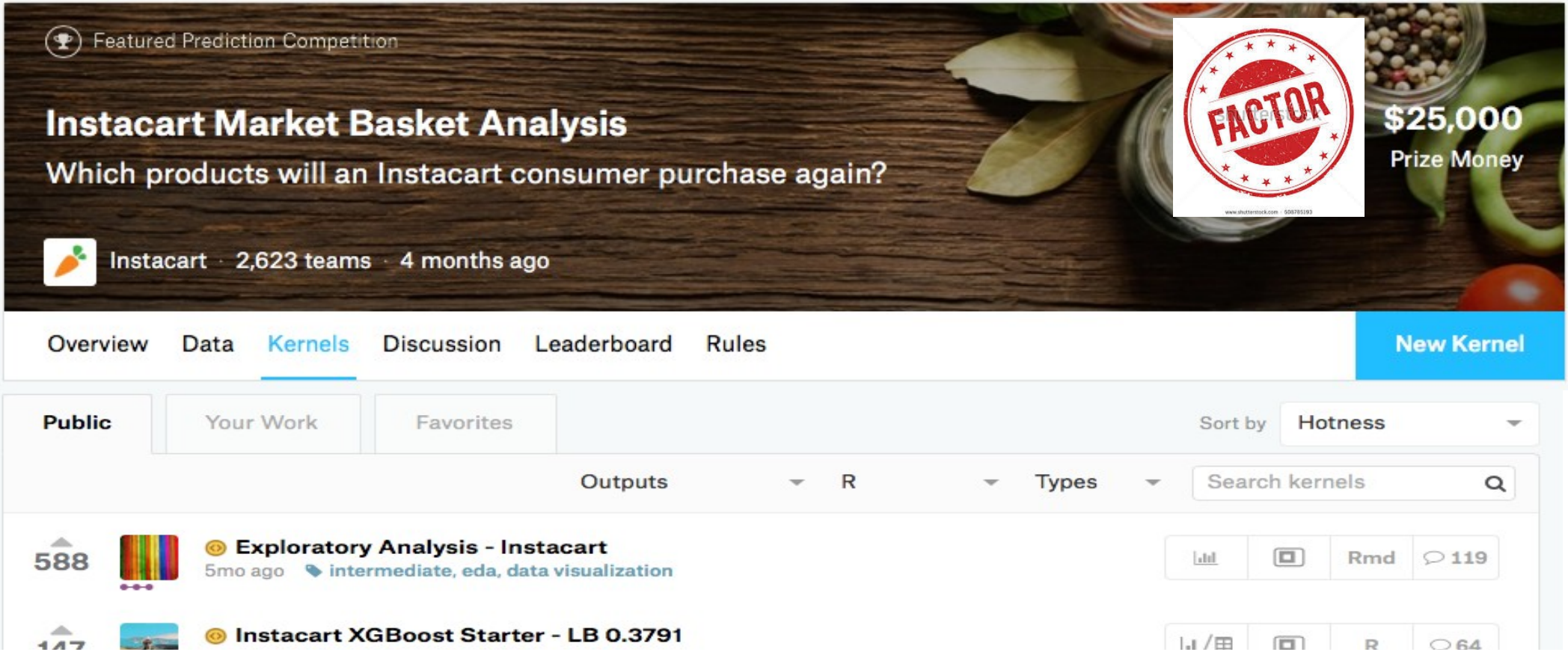
Overview Data Kernels Discussion Activity Download (30 KB) New Kernel

Attribute Information: (classes: edible=e, poisonous=p)

cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
bruises: bruises=t,no=f
odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
gill-attachment: attached=a,descending=d,free=f,notched=n
gill-spacing: close=c,crowded=w,distant=d **gill-size:** broad=b,narrow=n
gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
stalk-shape: enlarging=e,tapering=t
stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
...

Data Set Characteristics:	Multivariate	**Number of Instances:**	8124	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	22	**Date Donated**	1987-04-27
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	298439

https://www.kaggle.com/uciml/mushroom-classification/data



https://www.kaggle.com/philippsp/exploratory-analysis-instacart

A smaller dataset "Groceries" from **arulesViz** package will b- used in the course.


```
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns
            2513             1903             1809
```

Forest Cover Type Prediction

Use cartographic variables to classify forest categories

1,694 teams · 3 years ago

Overview Data Kernels Discussion Leaderboard Rules

13 predictors (d/c), 7 clases

The study area located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m x 30m patch. You are asked to predict an integer classification for the **forest cover type (seven types)**.

- Training set (15120 observations)
- Test set (565892 observations).

Data Fields

Elevation - Elevation in meters
Aspect - Aspect in degrees azimuth
Slope - Slope in degrees
Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water feature
Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway
Hillshade_9am (0 to 255 index) - Hillshade index at 9am, summer solstice
Hillshade_Noon (0 to 255 index) - Hillshade index at noon, summer solstice
Hillshade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition poin

Data Set Characteristics:	Multivariate	**Number of Instances:**	326	**Area:**	Life
Attribute Characteristics:	N/A	**Number of Attributes:**	27	**Date Donated**	2015-05-25
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	30636

https://www.kaggle.com/c/forest-cover-type-prediction

https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping

Reviewed Dataset

Gene expression dataset (Golub et al.)

Molecular Classification of Cancer by Gene Expression Monitoring

Chris Crawford • last updated 4 months ago

Overview | Data | Kernels | Discussion | Activity | Download (1 MB) | New Kernel

Optimization Based Tumor Classification from Microarray Gene Expression Data

Onur Dagliyan[1], Fadime Uney-Yuksektepe[2], I. Halil Kavakli[1], Metin Turkay[3]*

An important use of data obtained from microarray measurements is the classification of tumor types with respect to genes that are either up or down regulated in specific cancer types.

Table 1. Cancer data sets used in this study.

Data set	Samples	Genes	Classes	Reference
Leukemia	72	7129	2	Golub et al. (1999)
Prostate cancer	102	12600	2	Singh et al. (2002)
Prostate outcome	21	12600	2	Singh et al. (2002)
DLBCL	77	7129	2	Shipp et al. (2002)

The highest accuracy is obtained with the optimal gene set consisting of 4 genes:

- Myeloperoxidase (M19507-at),
- adipsin (M84526-at),
- CD33 antigen and
- TCF3 transcription factor 3.

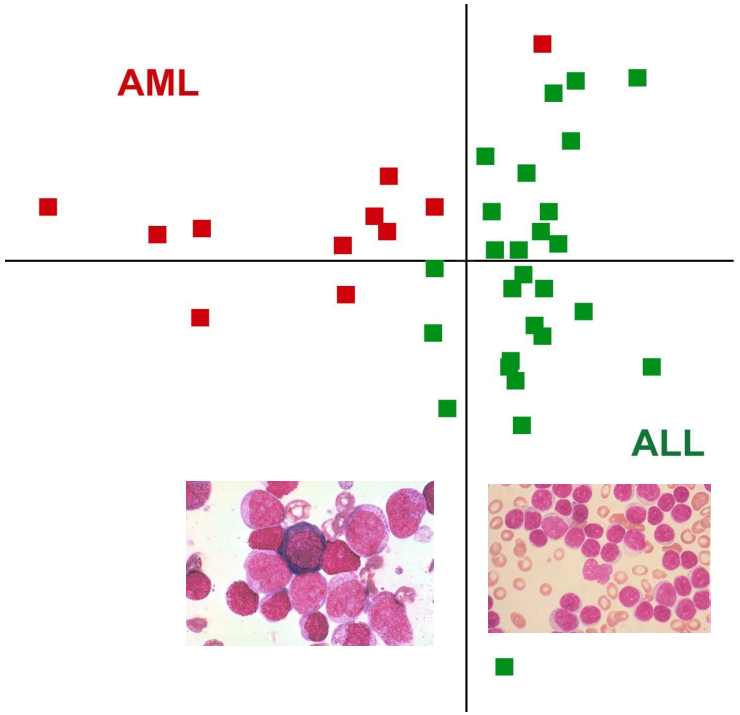


Table 2. Classification results of leukemia (AML-ALL) data set.

Classifier	Test Set	10-CV	LOOCV
HBE	100	97.146 0.903	98.61
BayesNet	94.12	95.71	95.83
LibSVM	58.82	86.576 10.44	91.67
SMO	97.06	93.146 0.571	94.44
Logistic Regression	91.18	96.866 1.67	98.61
RBF Network	97.06	97.43± 1.07	97.22
IBk	97.06	96.006 1.40	95.83
J48	94.12	89.146 1.94	90.28
Random Forest	94.12	93.146 1.07	90.2

Reviewed Dataset

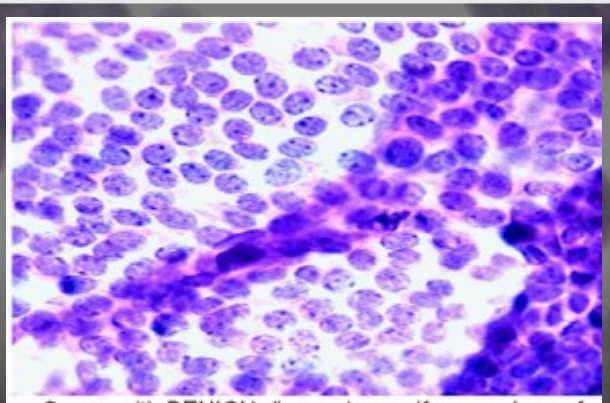
Breast Cancer Wisconsin (Diagnostic) Data Set

Predict whether the cancer is benign or malignant

UCI Machine Learning • last updated a year ago

244

Overview Data Kernels Discussion Activity



Smear with BENIGN diagnosis – uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size

Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	316814

Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	608824

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	w s
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	25.38	17.33	184.60	2019.0	0
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	24.99	23.41	158.80	1956.0	0
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	23.57	25.53	152.50	1709.0	0

Published Dataset

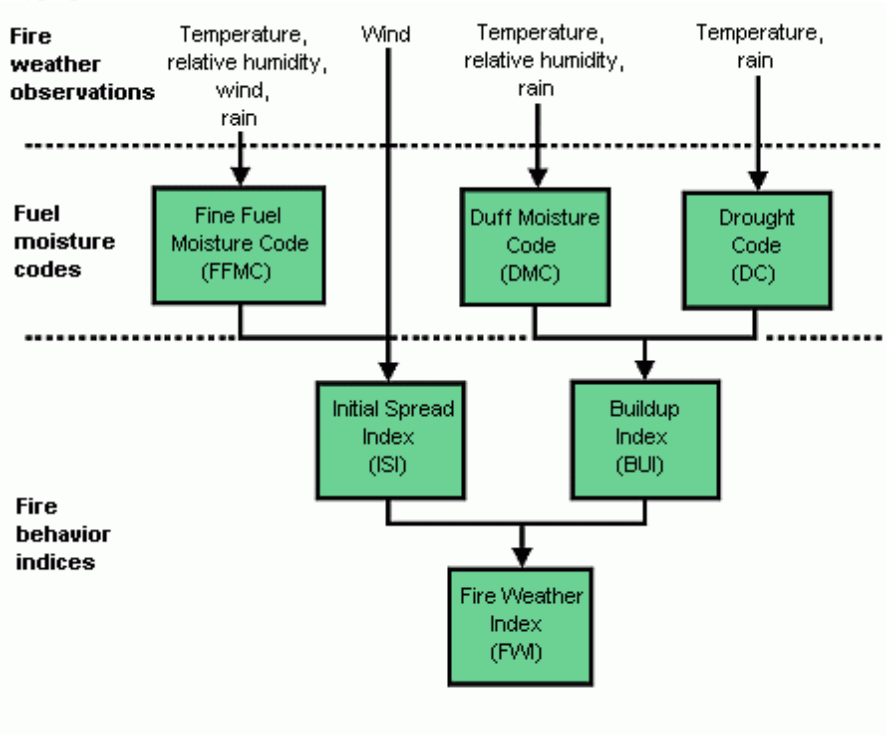
Forest Fires Data Set

predict the burned area of forest fires using meteorological and other data

Ahiale Darlington • last updated 3 months ago

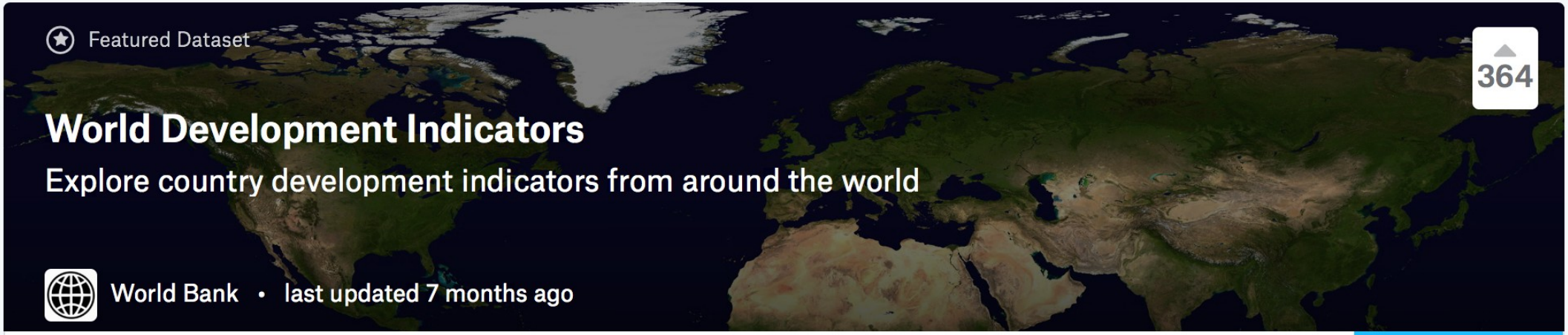
Overview Data Kernels Discussion Activity Download (7 KB) New Kernel

3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84



Data Set Characteristics:	Multivariate	**Number of Instances:**	517
Attribute Characteristics:	Real	**Number of Attributes:**	13
Associated Tasks:	Regression	**Missing Values?**	N/A

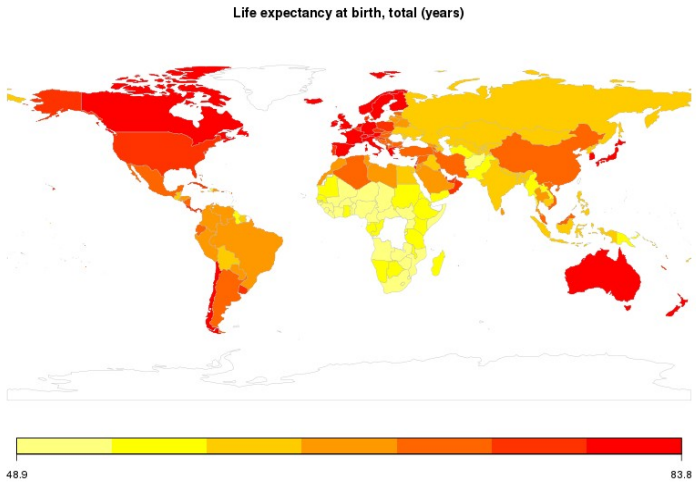
http://archive.ics.uci.edu/ml/datasets/Forest+Fires

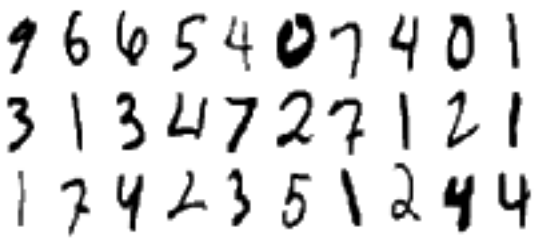


Overview Data Kernels Discussion Activity Download (385 MB) New Kernel

Up to 1300 indices over time for 247 countries

IndicatorCode	IndicatorName	NumCountries	NumYears	FirstYear	LastYear
AG.AGR.TRAC.NO	Agricultural machinery, tractors	219	49	1961	2009
AG.CON.FERT.PT.ZS	Fertilizer consumption (% of fertilizer production)	118	12	2002	2013
AG.CON.FERT.ZS	Fertilizer consumption (kilograms per hectare of arable land)				
AG.LND.AGRI.K2	Agricultural land (sq. km)				
AG.LND.AGRI.ZS	Agricultural land (% of land area)				
AG.LND.ARBL.HA	Arable land (hectares)				
AG.LND.ARBL.HA.PC	Arable land (hectares per person)				
AG.LND.ARBL.ZS	Arable land (% of land area)				
AG.LND.CREL.HA	Land under cereal production (hectares)				





Digit Recognizer

Learn computer vision fundamentals with the famous MNIS

1,996 teams · 2 years to go

http://www.meteo.unican.es/work/train.csv

Overview Data Kernels Discussion Leaderboard Rules

https://www.kaggle.com/c/digit-recognizer#tutorial

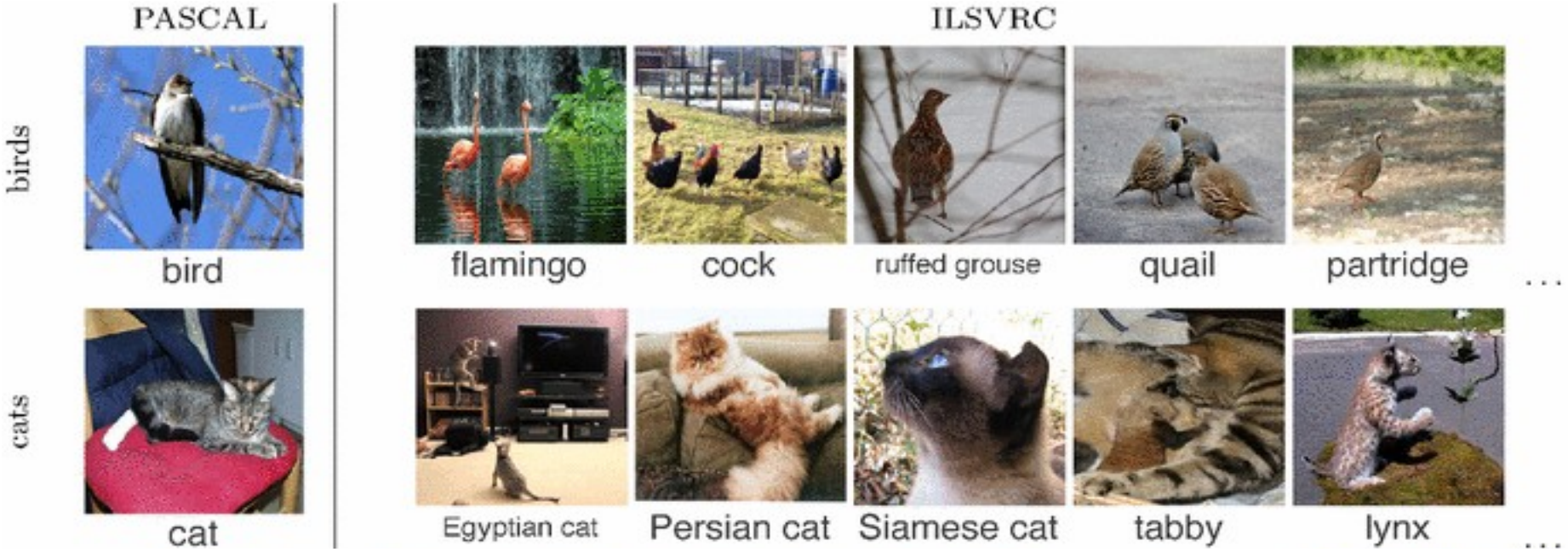
```
library(readr)
train <- data.frame(read_csv("/.../train.csv"))
str(train)

y <- train[,1]; x <- train[,-1]

dim(x)
sqrt(dim(x)[2])

par(mfrow = c(3,2))
image(matrix(as.matrix(x[7,]), nrow = 28, ncol = 28))
for (i in 8:12) {
      image(matrix(as.matrix(x[i,]), nrow = 28, ncol = 28))
}
y[7:12]

y[which(y < 9)] <- 0 ; y[which(y == 9)] <- 1
```

```
data <- data.frame(y,x)

model <- lm(y~., data = data)
out <- model$fitted.values

outbin <- as.double(out > 0.5)

Outbin[7:12]

100*mean(abs(y - outbin))# Error (%)
```

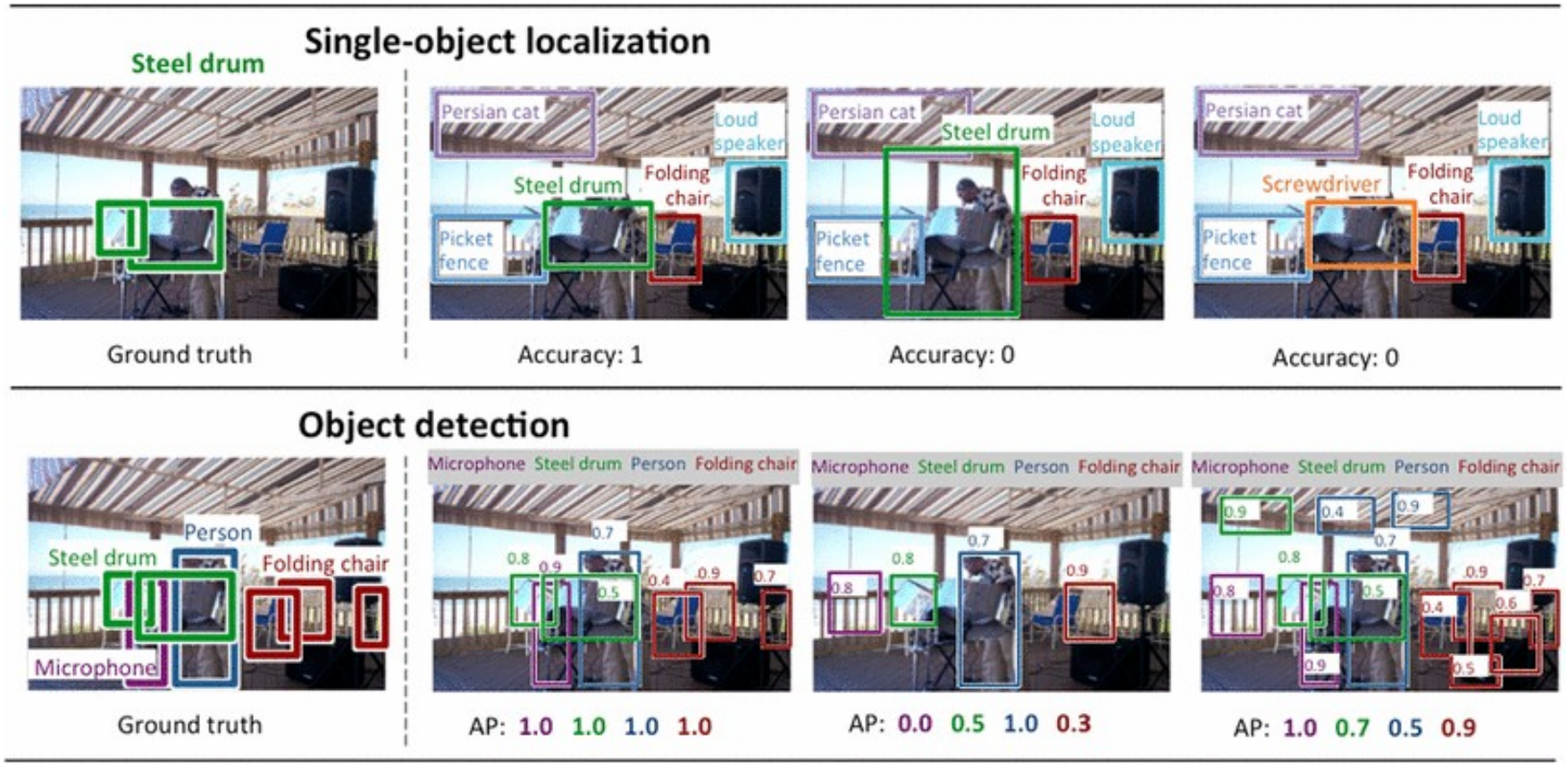
ImageNet is an image database organized according to the (nouns of the) WordNet hierarchy, in which each node of the hierarchy is depicted by an average of over five hundred images.

#synsets: 21841
#images: 14197122

150 GB [kaggle]



David G. Lowe, **Distinctive Image Features from Scale-Invariant Keypoints.** *International Journal of Computer Vision, 2004.*



Validation: top-5 error rate

2017 video included

Inception-v3: 3.46% top-5 and 17.3% top-1 (25 million parameters). [Inception In kaggle]

O. Russakovsk (2015) ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision, 115, 211–252