Machine Learning Datasets
Why? Which? For What?

(Largely based on slides from Samuel Fadel)

Edson Bollis (Ph.D. Student)
Institute of Computing (IC/Unicamp)

MC886/MO444, September 6, 2018

Why?

Superhuman Pattern Recognition

IJCNN Traffic Sign Recognition Competition (2011)

- 40+ classes and ~50k images
- **First** system to beat humans in visual pattern recognition

Superhuman Pattern Recognition

IJCNN Traffic Sign Recognition Competition (2011)

- **Why was ImageNet 2012 more memorable?**

  1k classes, 1.2 million training images

Superhuman Pattern Recognition

IJCNN Traffic Sign Recognition Competition (2011)

- **Why was ImageNet 2012 more memorable?**

German traffic sign recognition

versus.

Large-scale visual recognition (1k classes)

ImageNet

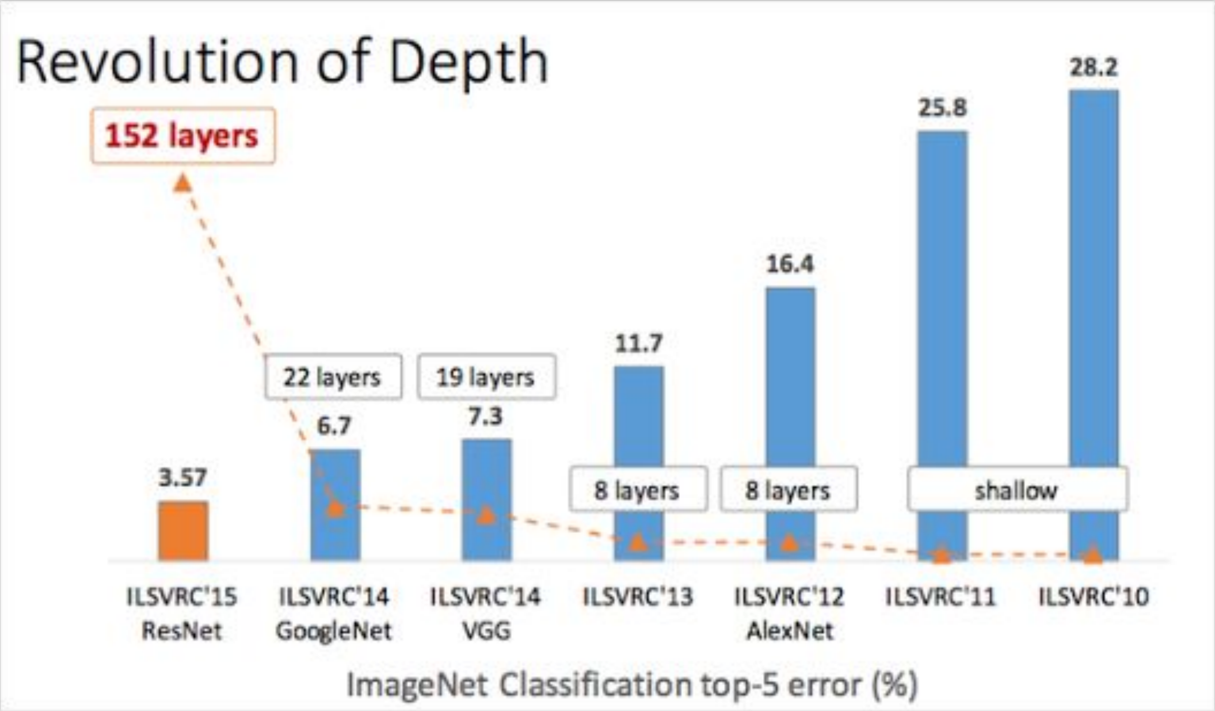


Figure source: Kaiming He

Why?

- Convince audience
- Baseline for comparison with other methods
- Suggest possible applications
- Highlight weaknesses

Which?

Which?

kaggle

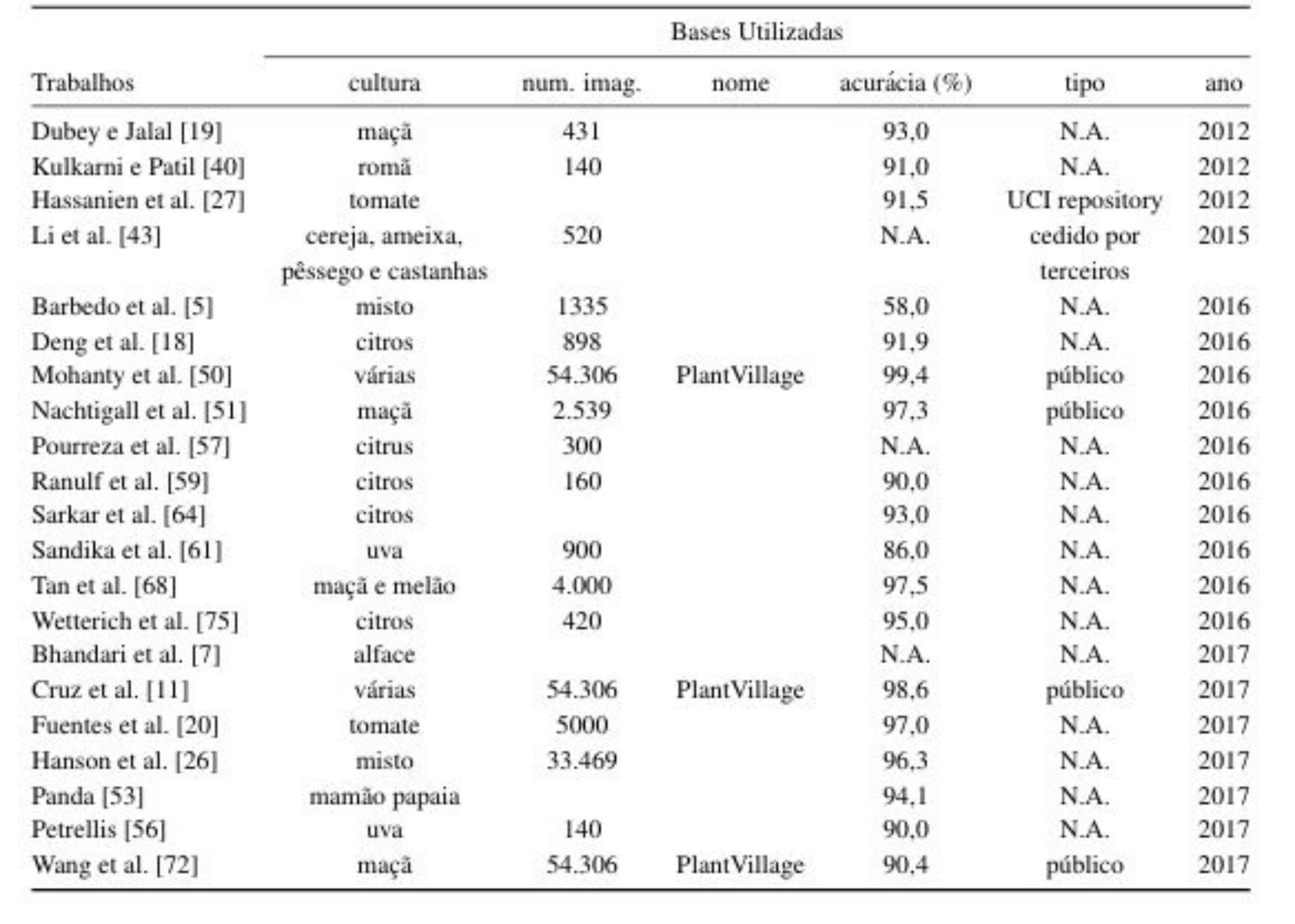
IMAGENET

WordNet
A lexical database for English

VISUALGENOME

UCI
Machine Learning Repository

“I find the experimental section of the paper rather weak: it mainly comprises of experiments on **toy** data sets” “experiments are performed on a set of (rather **artificial**) data sets” “the experiments should be conducted with more **real** datasets”

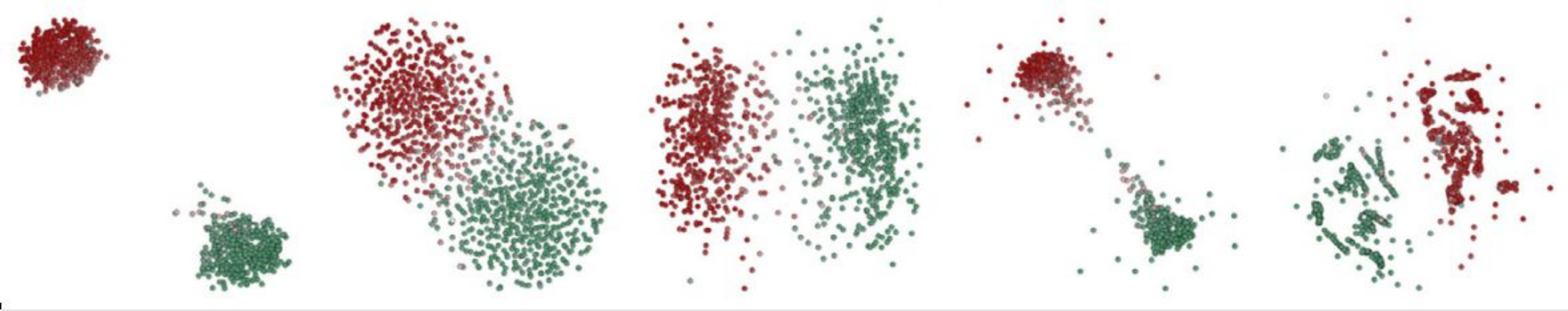


	Bases Utilizadas					
Trabalhos	cultura	num. imag.	nome	acurácia (%)	tipo	ano
Dubey e Jalal [19]	maçã	431		93,0	N.A.	2012
Kulkarni e Patil [40]	romã	140		91,0	N.A.	2012
Hassanien et al. [27]	tomate			91,5	UCI repository	2012
Li et al. [43]	cereja, ameixa, pêssego e castanhas	520		N.A.	cedido por terceiros	2015
Barbedo et al. [5]	misto	1335		58,0	N.A.	2016
Deng et al. [18]	citros	898		91,9	N.A.	2016
Mohanty et al. [50]	várias	54.306	PlantVillage	99,4	público	2016
Nachtigall et al. [51]	maçã	2.539		97,3	público	2016
Pourreza et al. [57]	citrus	300		N.A.	N.A.	2016
Ranulf et al. [59]	citros	160		90,0	N.A.	2016
Sarkar et al. [64]	citros			93,0	N.A.	2016
Sandika et al. [61]	uva	900		86,0	N.A.	2016
Tan et al. [68]	maçã e melão	4.000		97,5	N.A.	2016
Wetterich et al. [75]	citros	420		95,0	N.A.	2016
Bhandari et al. [7]	alface			N.A.	N.A.	2017
Cruz et al. [11]	várias	54.306	PlantVillage	98,6	público	2017
Fuentes et al. [20]	tomate	5000		97,0	N.A.	2017
Hanson et al. [26]	misto	33.469		96,3	N.A.	2017
Panda [53]	mamão papaia			94,1	N.A.	2017
Petrellis [56]	uva	140		90,0	N.A.	2017
Wang et al. [72]	maçã	54.306	PlantVillage	90,4	público	2017

Which?

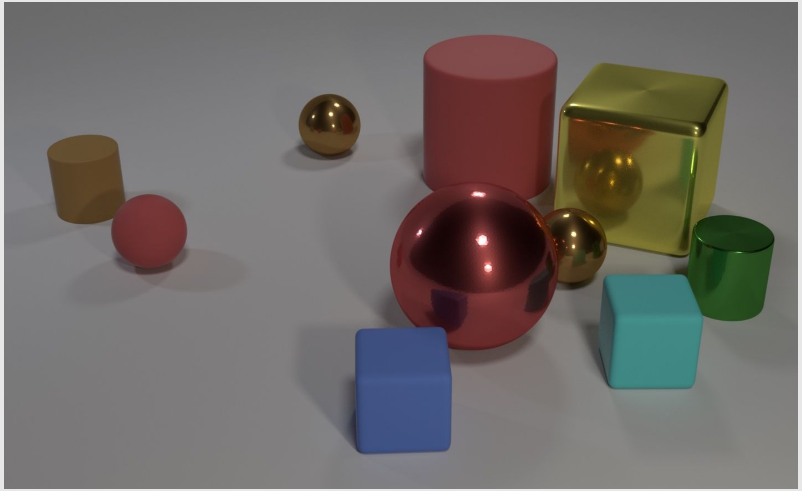
Toy datasets illustrate concepts and are easy to interpret

- Two 20d Gaussians reduced to 2d with 5 methods

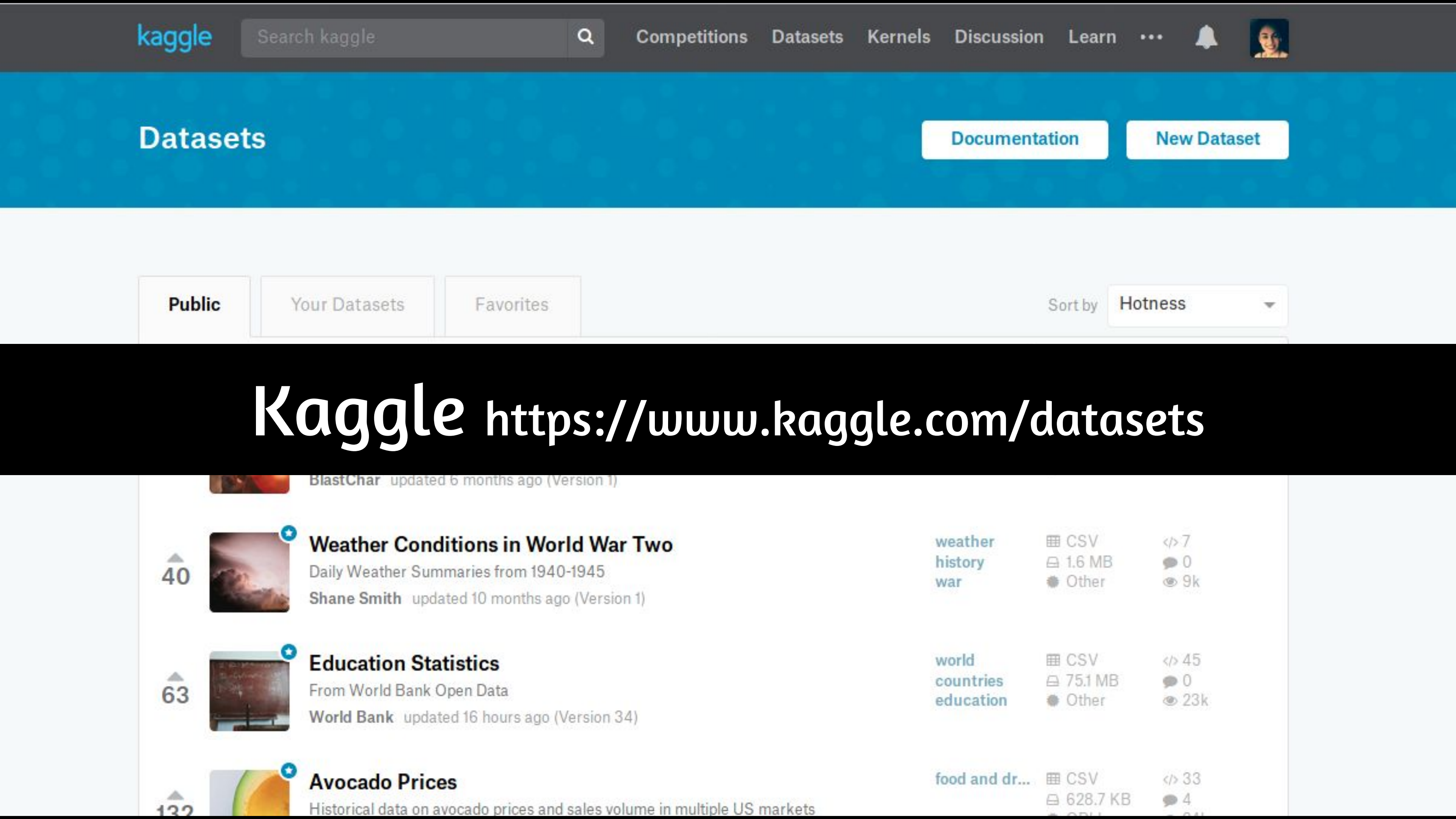


Which?

Challenging datasets are meant **to push** the state of the art.



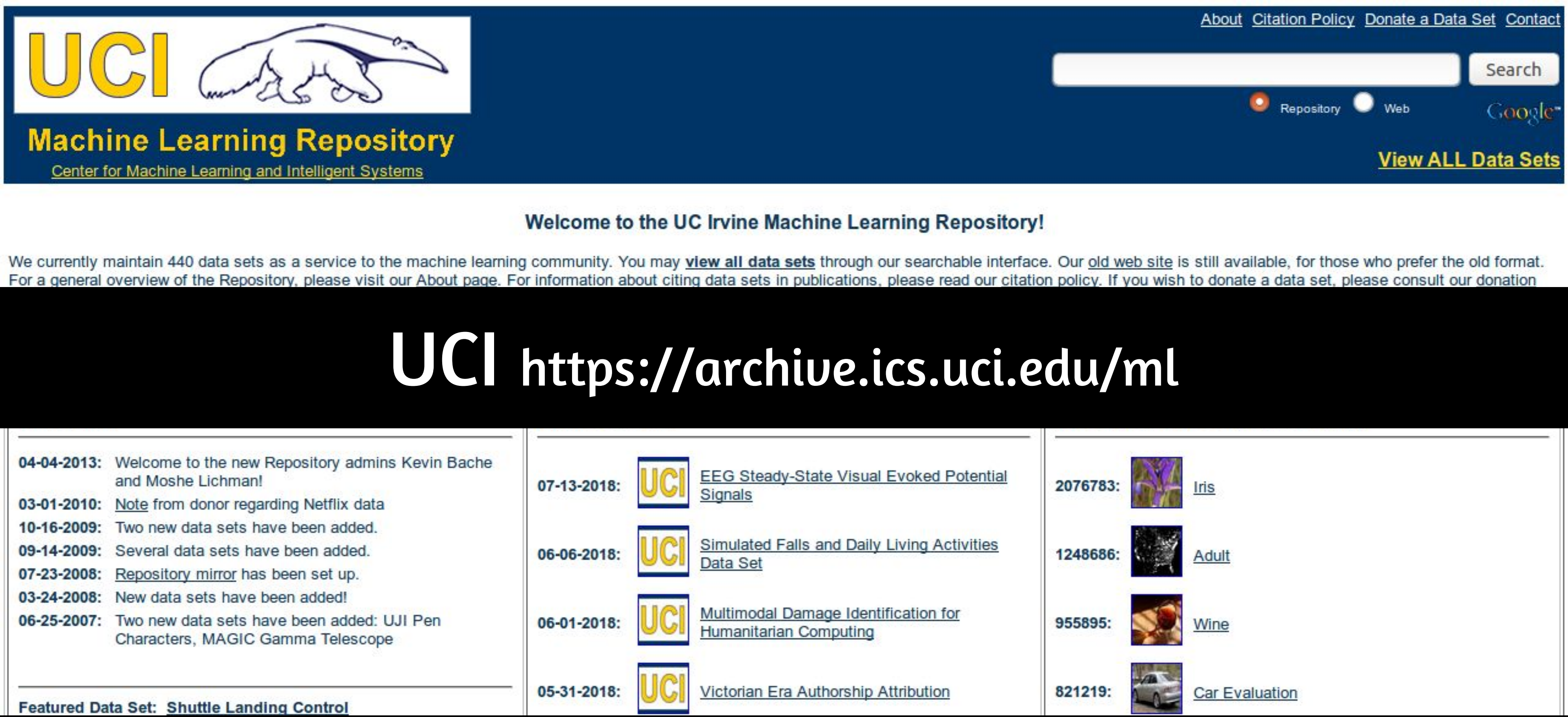
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?



Kaggle https://www.kaggle.com/datasets

BlastChar updated 6 months ago (Version 1)

40	**Weather Conditions in World War Two** Daily Weather Summaries from 1940-1945 Shane Smith updated 10 months ago (Version 1)	weather history war	CSV 1.6 MB Other	7 0 9k
63	**Education Statistics** From World Bank Open Data World Bank updated 16 hours ago (Version 34)	world countries education	CSV 75.1 MB Other	45 0 23k
132	**Avocado Prices** Historical data on avocado prices and sales volume in multiple US markets	food and dr...	CSV 628.7 KB	33 4



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contact
Search
● Repository ○ Web
Google
View ALL Data Sets

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 440 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation

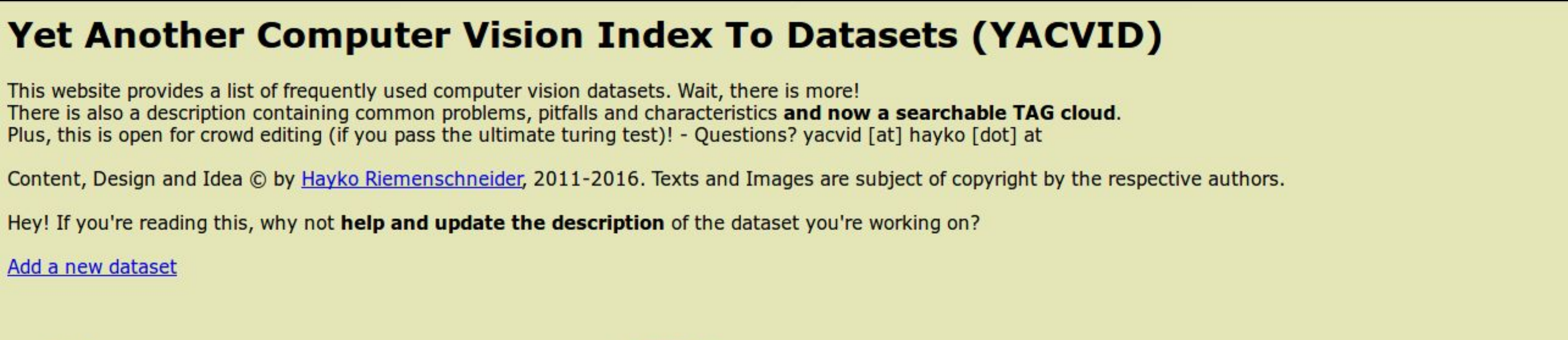
UCI https://archive.ics.uci.edu/ml

04-04-2013:	Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010:	Note from donor regarding Netflix data
10-16-2009:	Two new data sets have been added.
09-14-2009:	Several data sets have been added.
07-23-2008:	Repository mirror has been set up.
03-24-2008:	New data sets have been added!
06-25-2007:	Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: Shuttle Landing Control

07-13-2018:	EEG Steady-State Visual Evoked Potential Signals
06-06-2018:	Simulated Falls and Daily Living Activities Data Set
06-01-2018:	Multimodal Damage Identification for Humanitarian Computing
05-31-2018:	Victorian Era Authorship Attribution

2076783:	Iris
1248686:	Adult
955895:	Wine
821219:	Car Evaluation



Yet Another Computer Vision Index To Datasets (YACVID)

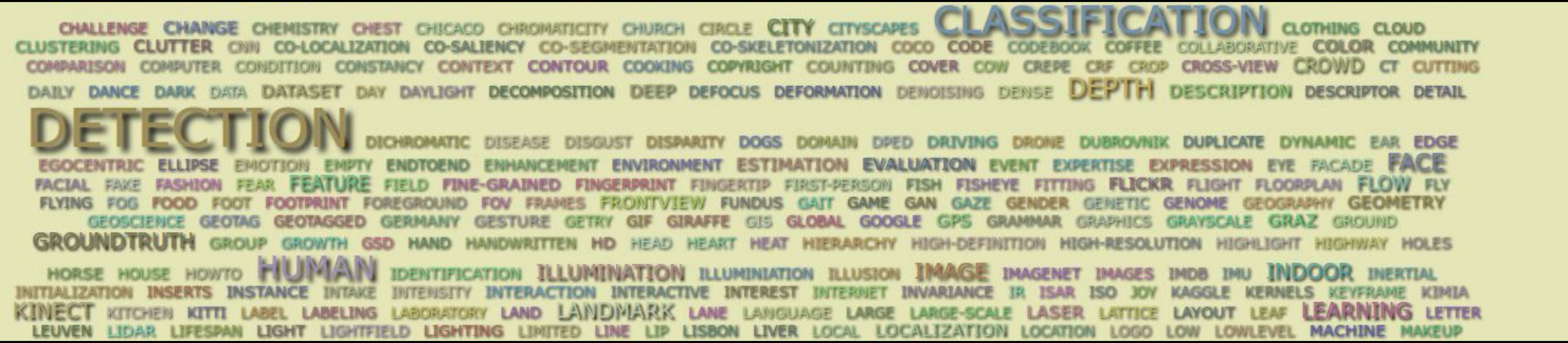
This website provides a list of frequently used computer vision datasets. Wait, there is more!
There is also a description containing common problems, pitfalls and characteristics **and now a searchable TAG cloud**.
Plus, this is open for crowd editing (if you pass the ultimate turing test)! - Questions? yacvid [at] hayko [dot] at

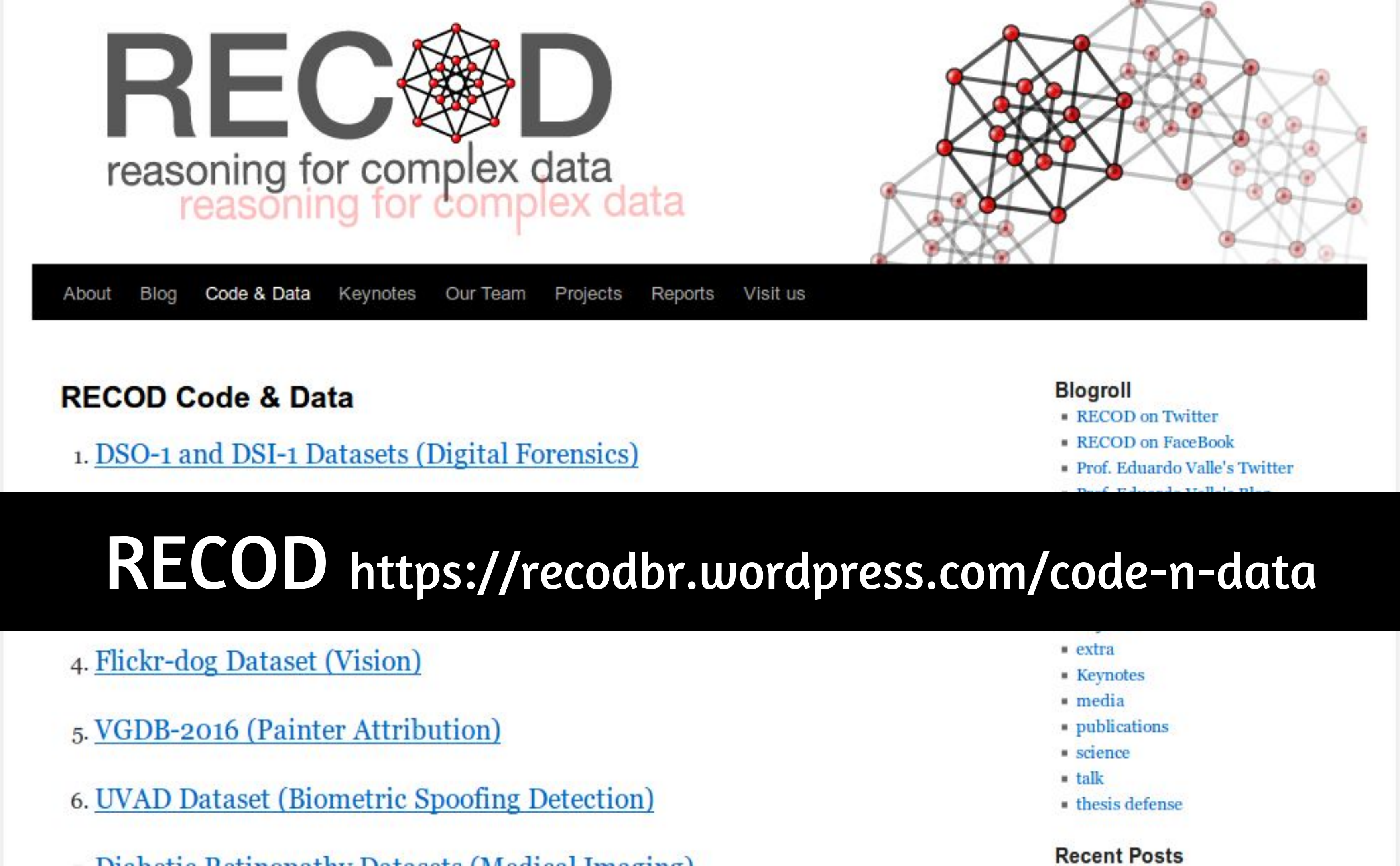
Content, Design and Idea © by Hayko Riemenschneider, 2011-2016. Texts and Images are subject of copyright by the respective authors.

Hey! If you're reading this, why not **help and update the description** of the dataset you're working on?

Add a new dataset

YACVID https://riemenschneider.hayko.at/vision/dataset


CHALLENGE CHANGE CHEMISTRY CHEST CHICAGO CHROMATICITY CHURCH CIRCLE CITY CITYSCAPES CLASSIFICATION CLOTHING CLOUD
CLUSTERING CLUTTER CNN CO-LOCALIZATION CO-SALIENCY CO-SEGMENTATION CO-SKELETONIZATION COCO CODE CODEBOOK COFFEE COLLABORATIVE COLOR COMMUNITY
COMPARISON COMPUTER CONDITION CONSTANCY CONTEXT CONTOUR COOKING COPYRIGHT COUNTING COVER COW CREPE CRF CROP CROSS-VIEW CROWD CT CUTTING
DAILY DANCE DARK DATA DATASET DAY DAYLIGHT DECOMPOSITION DEEP DEFOCUS DEFORMATION DENOISING DENSE DEPTH DESCRIPTION DESCRIPTOR DETAIL
DETECTION DICHROMATIC DISEASE DISGUST DISPARITY DOGS DOMAIN DPED DRIVING DRONE DUBROVNIK DUPLICATE DYNAMIC EAR EDGE
EGOCENTRIC ELLIPSE EMOTION EMPTY ENDTOEND ENHANCEMENT ENVIRONMENT ESTIMATION EVALUATION EVENT EXPERTISE EXPRESSION EYE FACADE FACE
FACIAL FAKE FASHION FEAR FEATURE FIELD FINE-GRAINED FINGERPRINT FINGERTIP FIRST-PERSON FISH FISHEYE FITTING FLICKR FLIGHT FLOORPLAN FLOW FLY
FLYING FOG FOOD FOOT FOOTPRINT FOREGROUND FOV FRAMES FRONTVIEW FUNDUS GAIT GAME GAN GAZE GENDER GENETIC GENOME GEOGRAPHY GEOMETRY
GEOSCIENCE GEOTAG GEOTAGGED GERMANY GESTURE GETRY GIF GIRAFFE GIS GLOBAL GOOGLE GPS GRAMMAR GRAPHICS GRAYSCALE GRAZ GROUND
GROUNDTRUTH GROUP GROWTH GSD HAND HANDWRITTEN HD HEAD HEART HEAT HIERARCHY HIGH-DEFINITION HIGH-RESOLUTION HIGHLIGHT HIGHWAY HOLES
HORSE HOUSE HOWTO HUMAN IDENTIFICATION ILLUMINATION ILLUMINIATION ILLUSION IMAGE IMAGENET IMAGES IMDB IMU INDOOR INERTIAL
INITIALIZATION INSERTS INSTANCE INTAKE INTENSITY INTERACTION INTERACTIVE INTEREST INTERNET INVARIANCE IR ISAR ISO JOY KAGGLE KERNELS KEYFRAME KIMIA
KINECT KITCHEN KITTI LABEL LABELING LABORATORY LAND LANDMARK LANE LANGUAGE LARGE LARGE-SCALE LASER LATTICE LAYOUT LEAF LEARNING LETTER
LEUVEN LIDAR LIFESPAN LIGHT LIGHTFIELD LIGHTING LIMITED LINE LIP LISBON LIVER LOCAL LOCALIZATION LOCATION LOGO LOW LOWLEVEL MACHINE MAKEUP



RECOD

reasoning for complex data

About Blog Code & Data Keynotes Our Team Projects Reports Visit us

RECOD Code & Data

1. DSO-1 and DSI-1 Datasets (Digital Forensics)

4. Flickr-dog Dataset (Vision)
5. VGDB-2016 (Painter Attribution)
6. UVAD Dataset (Biometric Spoofing Detection)
7. Diabetic Retinopathy Datasets (Medical Imaging)

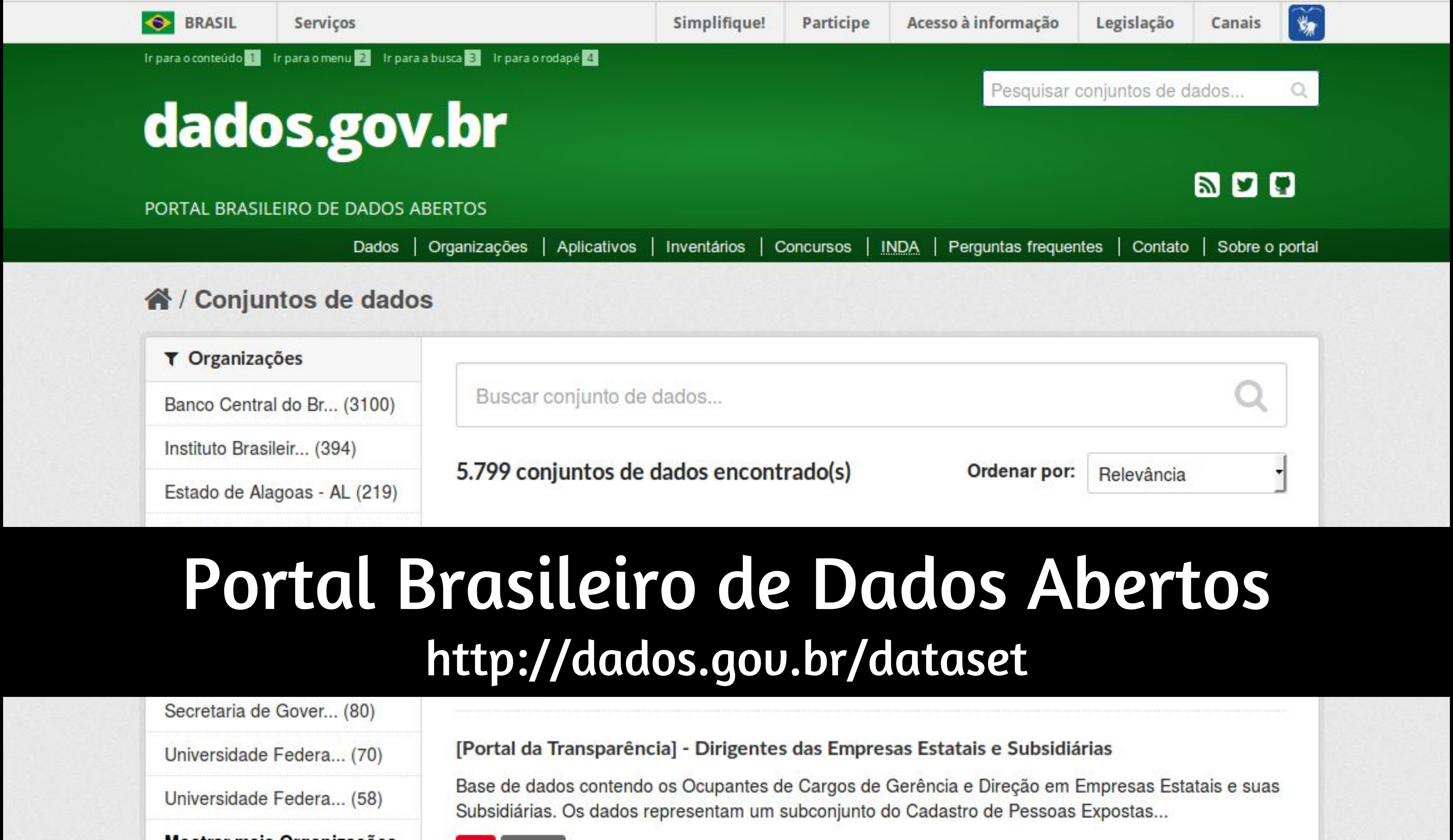
Blogroll

- RECOD on Twitter
- RECOD on FaceBook
- Prof. Eduardo Valle's Twitter
- Prof. Eduardo Valle's Blog

- extra
- Keynotes
- media
- publications
- science
- talk
- thesis defense

Recent Posts

RECOD https://recodbr.wordpress.com/code-n-data



/ Conjuntos de dados

Organizações

Banco Central do Br...	(3100)
Instituto Brasileir...	(394)
Estado de Alagoas - AL	(219)
Secretaria de Gover...	(80)
Universidade Federa...	(70)
Universidade Federa...	(58)

5.799 conjuntos de dados encontrado(s)

Portal Brasileiro de Dados Abertos

http://dados.gov.br/dataset

[Portal da Transparência] - Dirigentes das Empresas Estatais e Subsidiárias

Base de dados contendo os Ocupantes de Cargos de Gerência e Direção em Empresas Estatais e suas Subsidiárias. Os dados representam um subconjunto do Cadastro de Pessoas Expostas...

For What?

Sentiment Analysis

Sentiment Analysis on Movie Reviews

“The movie is surprising with plenty of unsettling plot twists.”

- Classification
  - Negative
  - Somewhat negative
  - Neutral
  - Somewhat positive
  - Positive

Data Compression

ImageNet data, YFCC100M, AudioSet

- Compress audio/image/video

Social Media Engagement Prediction

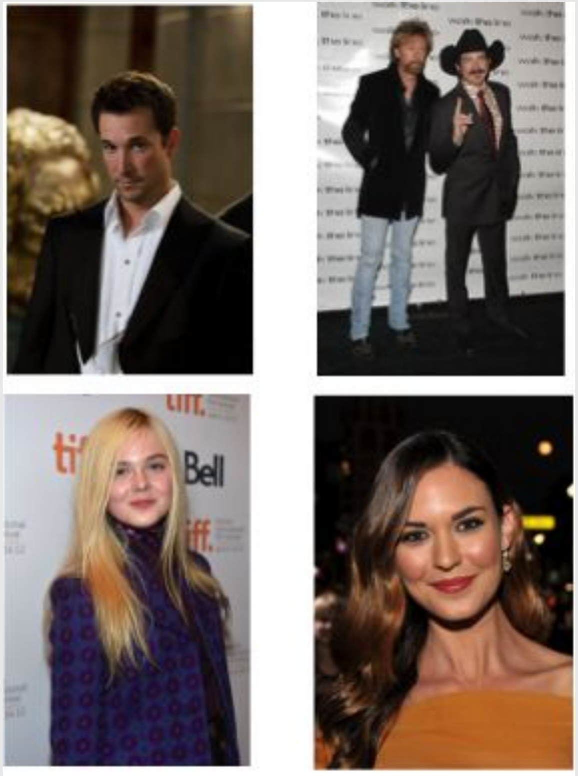
Facebook Comment Volume Dataset

- 480k posts
- Regression
    - Predict number of comments a post will receive

Age and Gender Prediction

IMDB-Wiki

- Classification
    - Predict gender
- Regression
    - Predict age



Predicting Media Interestingness

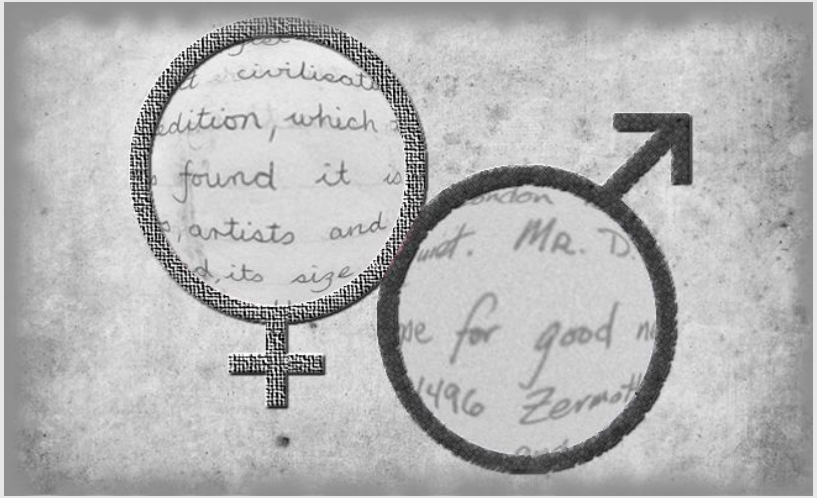
Media Interestingness Data

- Image, video, and metadata
- 5,054 samples (train) + 2,342 (test)
- Classification
  - Interesting
  - Not interesting

Gender Prediction from Handwriting

Handwriting Data

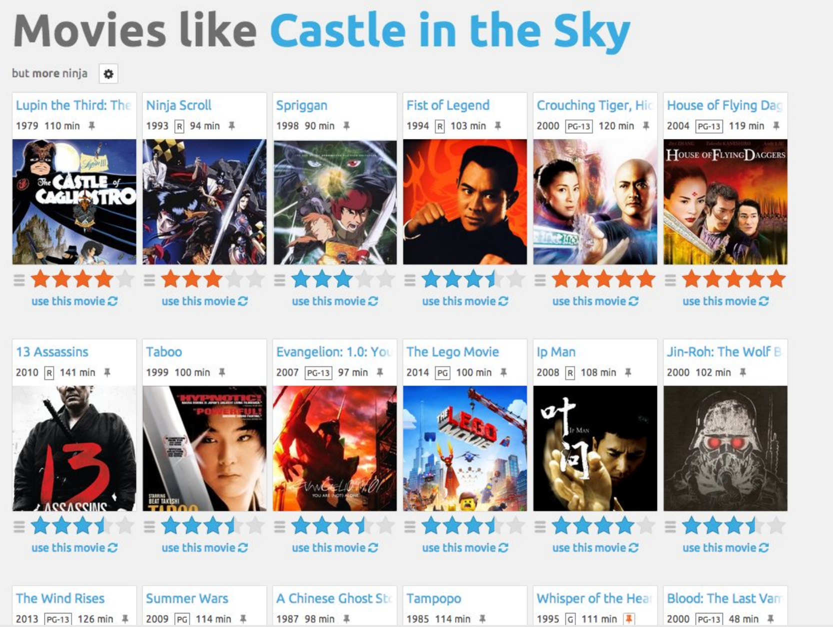
- Images in two languages (English, Arabic)
- Two pages for each language per writer
- Classification
  - Author's gender from handwriting style



Recommendation System

MovieLens 1M dataset

- 1M ratings from 6k users on 4k movies
- Regression
  - Predict ratings (1 to 5)



Movies like Castle in the Sky

but more ninja

Lupin the Third: The
1979 110 min
use this movie

Ninja Scroll
1993 R 94 min
use this movie

Spriggan
1998 90 min
use this movie

Fist of Legend
1994 R 103 min
use this movie

Crouching Tiger, Hic
2000 PG-13 120 min
use this movie

House of Flying Dag
2004 PG-13 119 min
use this movie

13 Assassins
2010 R 141 min
use this movie

Taboo
1999 100 min
use this movie

Evangelion: 1.0: You
2007 PG-13 97 min
use this movie

The Lego Movie
2014 PG 100 min
use this movie

Ip Man
2008 R 108 min
use this movie

Jin-Roh: The Wolf B
2000 102 min
use this movie

The Wind Rises
2013 PG-13 126 min

Summer Wars
2009 PG 114 min

A Chinese Ghost Sto
1987 98 min

Tampopo
1985 114 min

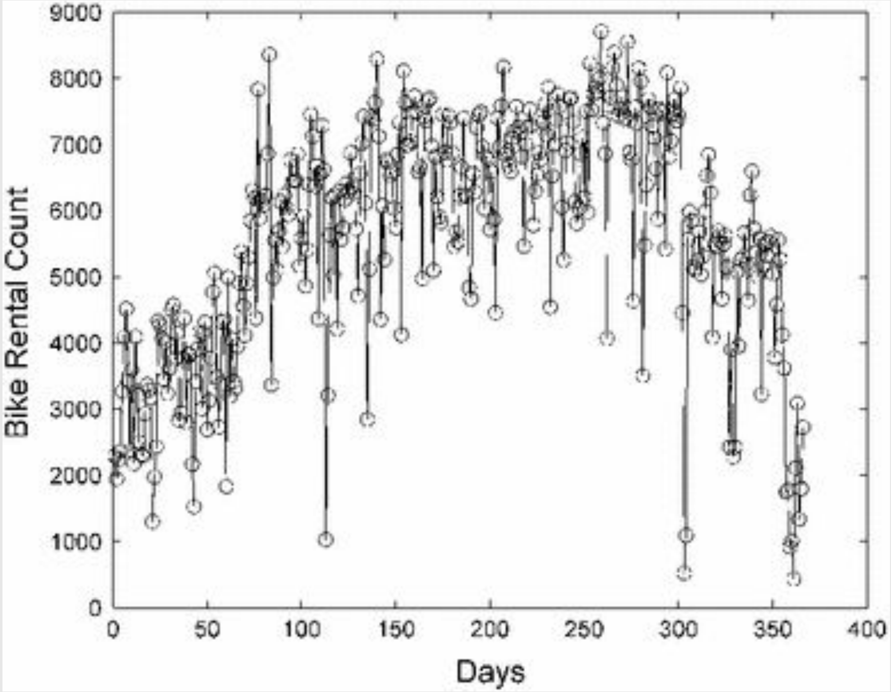
Whisper of the Hea
1995 G 111 min

Blood: The Last Van
2000 PG-13 48 min

Bike Sharing

Bike Sharing Dataset

- Regression
  - Predict bike rental count (hourly or daily)
- Anomaly detection
  - Detect days with spurious rental counts



Be bold!
Be brave!