RECOD
reasoning for complex data

Machine Learning Datasets

Why? Which? For What?

(Largely based on slides from Samuel Fadel)

Prof. Sandra Avila
Institute of Computing (IC/Unicamp)

MC886, September 2, 2019

Deep Neural Network (Pretraining)

SVM

Multi-layered Perceptron (Backpropagation)

ADALINE

XOR Problem

Perceptron

Golden Age

Dark Age ("AI Winter")

Electronic Brain

1943 1957 1960 1969 1986 1995 2006

1940 1950 1960 1970 1980 1990 2000 2010

S. McCulloch - W. Pitts

F. Rosenblatt

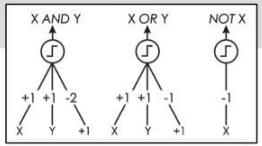
B. Widrow - M. Hoff

M. Minsky - S. Papert

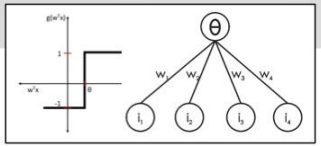
D. Rumelhart - G. Hinton - R. Wiliams

V. Vapnik - C. Cortes

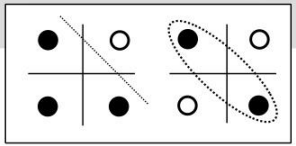
G. Hinton - S. Ruslan



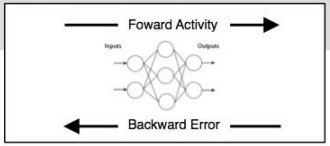
- Adjustable Weights
- Weights are not Learned



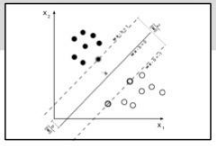
- Learnable Weights and Threshold



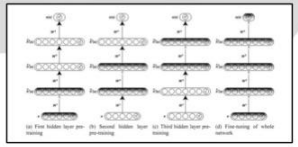
- XOR Problem



- Solution to nonlinearly separable problems
- Big computation, local optima and overfitting



- Limitations of learning prior knowledge
- Kernel function: Human Intervention



- Hierarchical feature Learning

IMAGENET

www.image-net.org

22K categories and 14M images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

Fly agaric, Amanita muscaria

Poisonous (but rarely fatal) woodland fungus having a scarlet cap with white warts and white gills

1444 pictures | 29.59% Popularity Percentile | Wordnet IDs

Numbers in brackets: (the number of synsets in the subtree).

- ImageNet 2011 Fall Release (3232
  - plant, flora, plant life (4486)
  - geological formation, formation
  - natural object (1112)
  - sport, athletics (176)
  - artifact, artefact (10504)
  - fungus (308)
    - false morel (6)
    - pythium (1)
    - truffle, earthnut, earth-ball (
    - candida (1)
    - shiitake, shiitake mushroom,
    - sac fungus (0)
    - lichen (11)
    - brown root rot fungus, Thiela
    - dry rot (0)
    - white fungus, Saprolegnia fer
    - mildew (5)
    - earthball, false truffle, puffba
    - yeast (2)
    - bird's-nest fungus (0)
    - green smut fungus, Ustilaginc
    - hen-of-the-woods, hen of the
    - verticillium (0)
    - scaly lentinus, Lentinus lepide

Treemap Visualization | Images of the Synset | Downloads

Demigod, superman, Ubermensch

A person with great powers and abilities

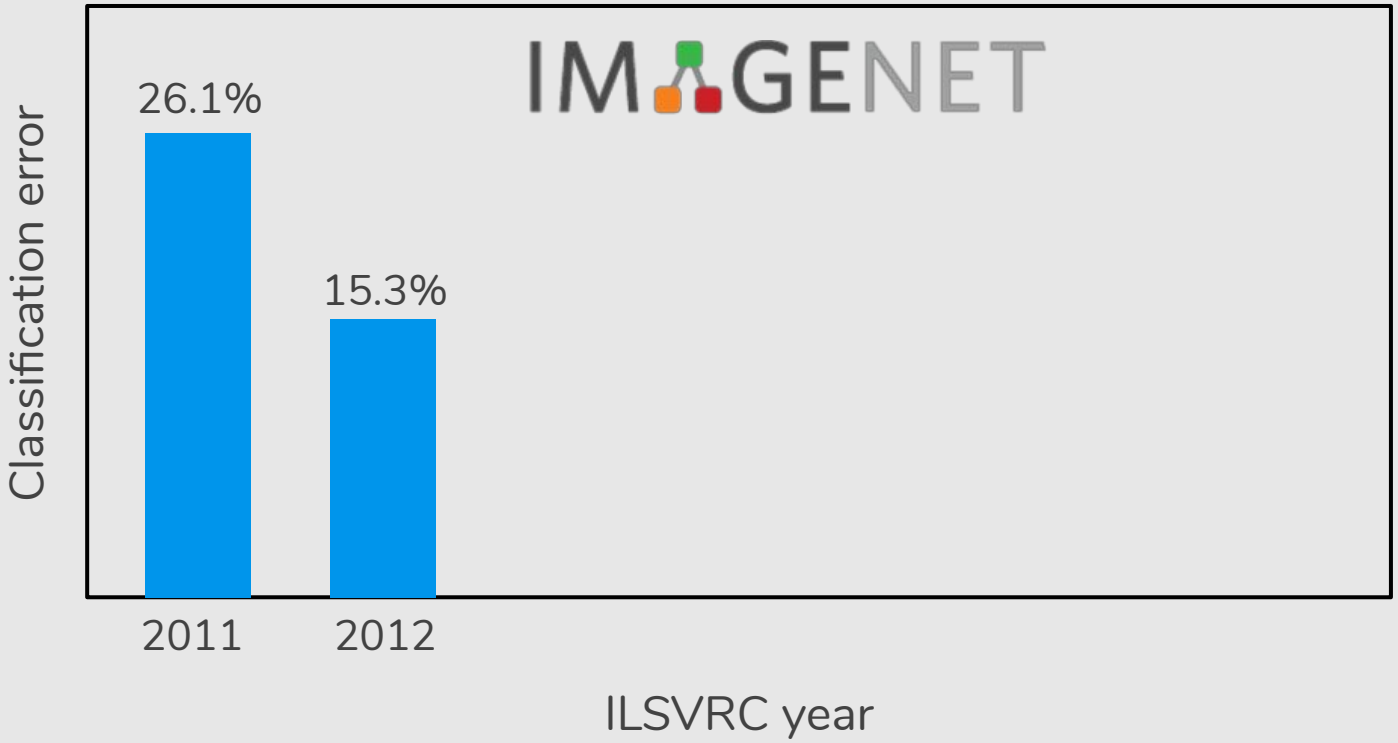
348 pictures

64.87% Popularity Percentile

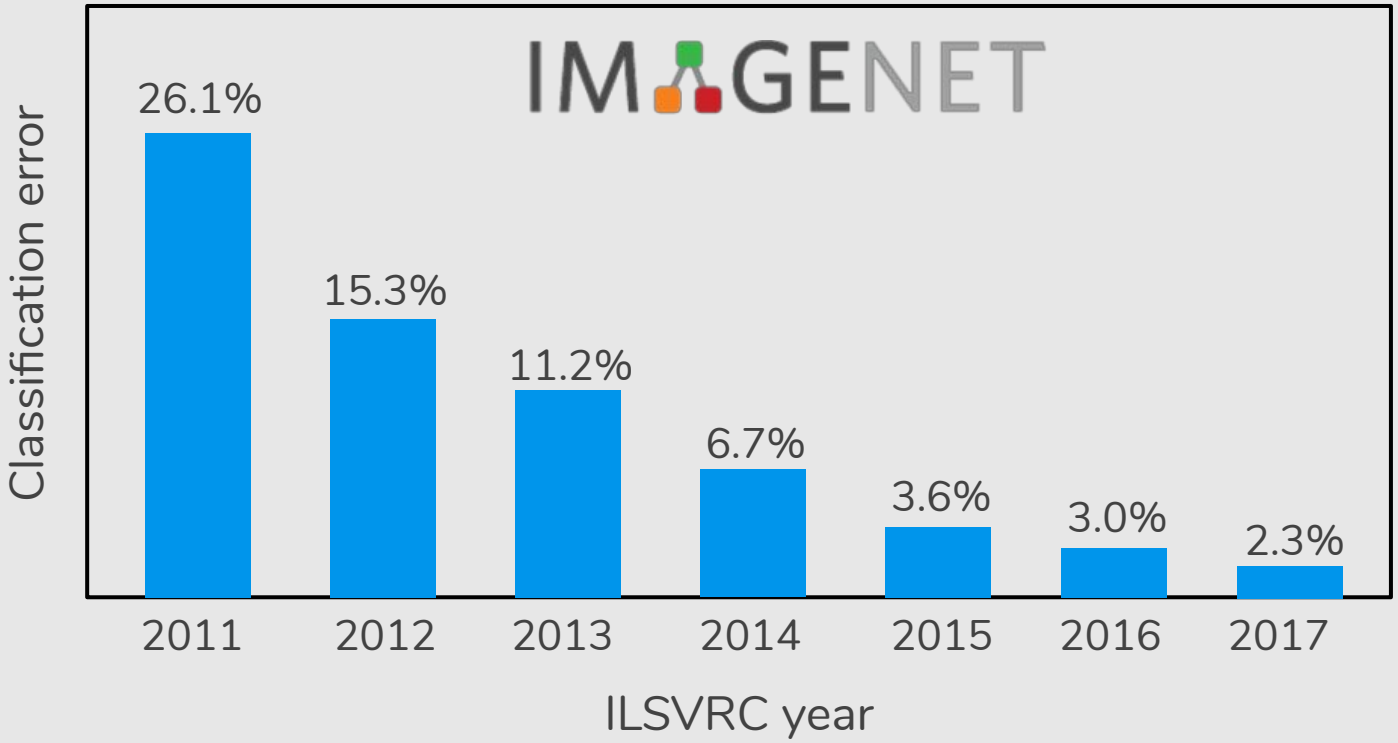
Wordnet IDs

Treemap Visualization | Images of the Synset | Downloads

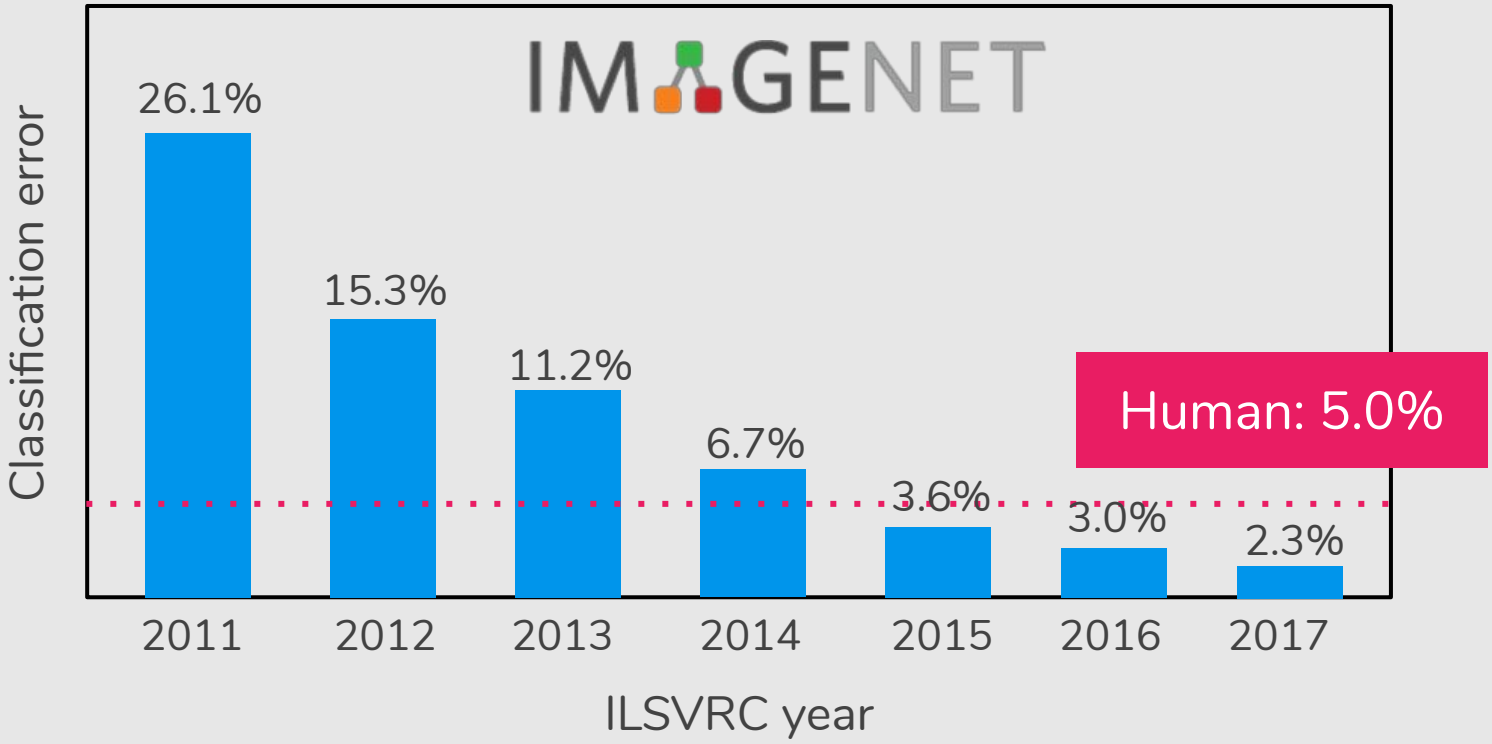
"ImageNet classification with deep convolutional neural networks". NIPS, 2012.



“ImageNet classification with deep convolutional neural networks”. NIPS, 2012.



"ImageNet classification with deep convolutional neural networks". NIPS, 2012.



"ImageNet classification with deep convolutional neural networks". NIPS, 2012.

Why?

Superhuman Pattern Recognition

IJCNN Traffic Sign Recognition Competition (2011)

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

“German Traffic Sign Recognition Benchmark”. IJCNN, 2011.

Superhuman Pattern Recognition

IJCNN Traffic Sign Recognition Competition (2011)

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

  1k classes, 1.2 million training images

"German Traffic Sign Recognition Benchmark". IJCNN, 2011.

Superhuman Pattern Recognition

IJCNN Traffic Sign Recognition Competition (2011)

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

German traffic sign recognition

vs.

Large-scale visual recognition (1k classes)

Why?

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

Which?

Which?


kaggle


UCI
Machine Learning Repository


Google Dataset Search Beta


IMAGENET


WordNet
A lexical database for English


VISUALGENOME


RECOD

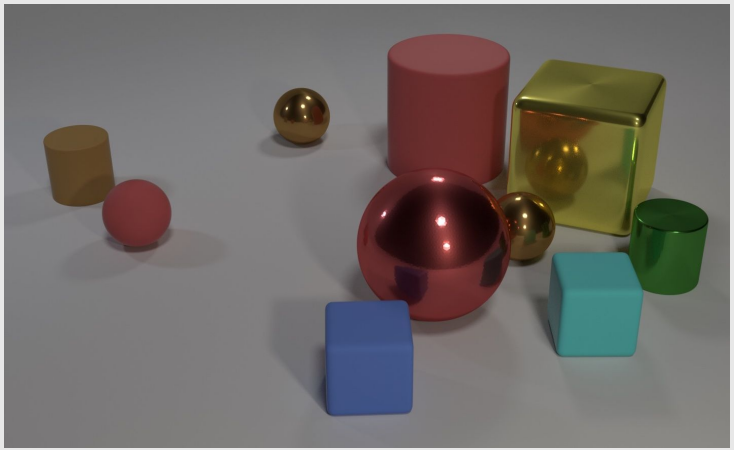
“I find the experimental section of the paper rather weak: it mainly comprises of experiments on **toy** data sets”

“experiments are performed on a set of (rather **artificial**) data sets”

“the experiments should be conducted with more **real** datasets”

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 Search kaggle Competitions Datasets Kernels Discussion Learn

Datasets

Documentation New Dataset

Public Your Datasets Favorites Sort by Hotness

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

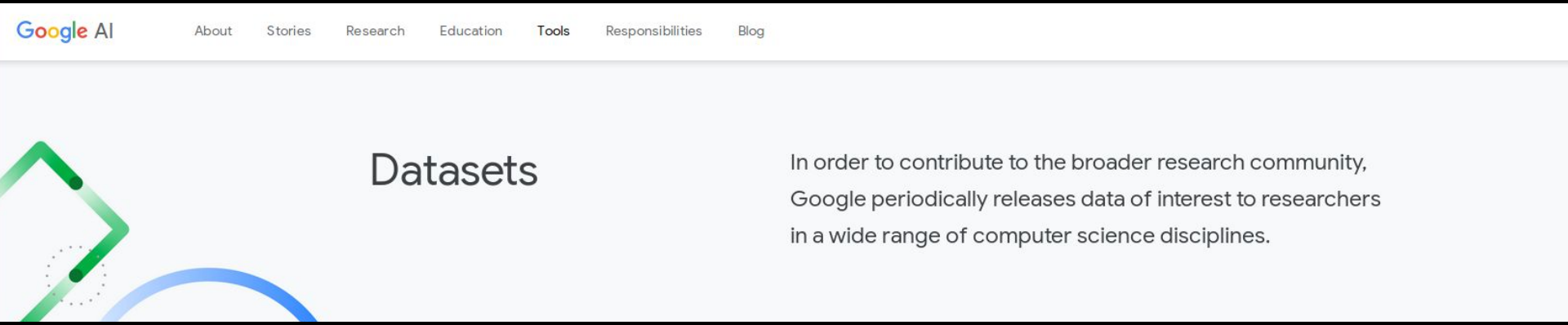
132 

Avocado Prices

Historical data on avocado prices and sales volume in multiple US markets

food and dr... | CSV 628.7 KB | 33 4

Google Dataset https://toolbox.google.com/datasetsearch



Google AI Datasets https://ai.google/tools/datasets

Dataset type

- Image 11
- Video 12
- Audio 19
- Text Annotation 26
- Robotics 7
- Other 6

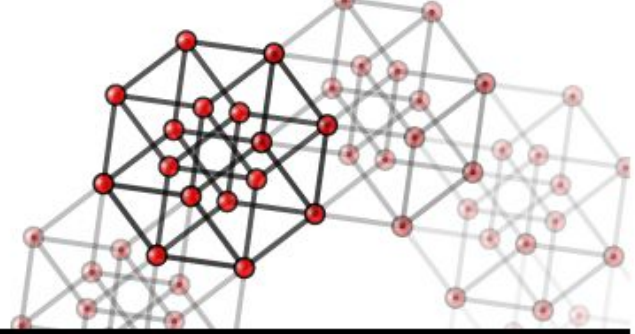
Coached Conversational Preference Elicitation
Wizard-of-Oz preference elicitation conversations between a user and an assistant about movie preferences, with annotated preference statements.

DiscoFuse
A dataset of 60 million examples for training sentence fusion models. The data has been collected from Wikipedia and from Sports articles.

Open Images Extended - Crowdsourced
Additional imagery sets to the main Open Images dataset, to improve its diversity (geographic, cultural, demographic, subject matter, etc). Currently composed of ~478K images contributed by users of the Crowdsource app.

Open Images
A dataset consisting of ~9 million URLs to images that have been annotated with labels spanning over 6000 categories.

MAESTRO
MIDI and Audio Edited for Synchronous TRacks and Organization (MAESTRO) is a dataset composed of over 172 hours of virtuosic piano performances captured with fine alignment (~3 ms) between note labels and audio waveforms.



RECOD

reasoning for complex data

reasoning for complex data

RECOD Code & Data

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

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

4. Flickr-dog Dataset (Vision)

5. VGDB-2016 (Painter Attribution)

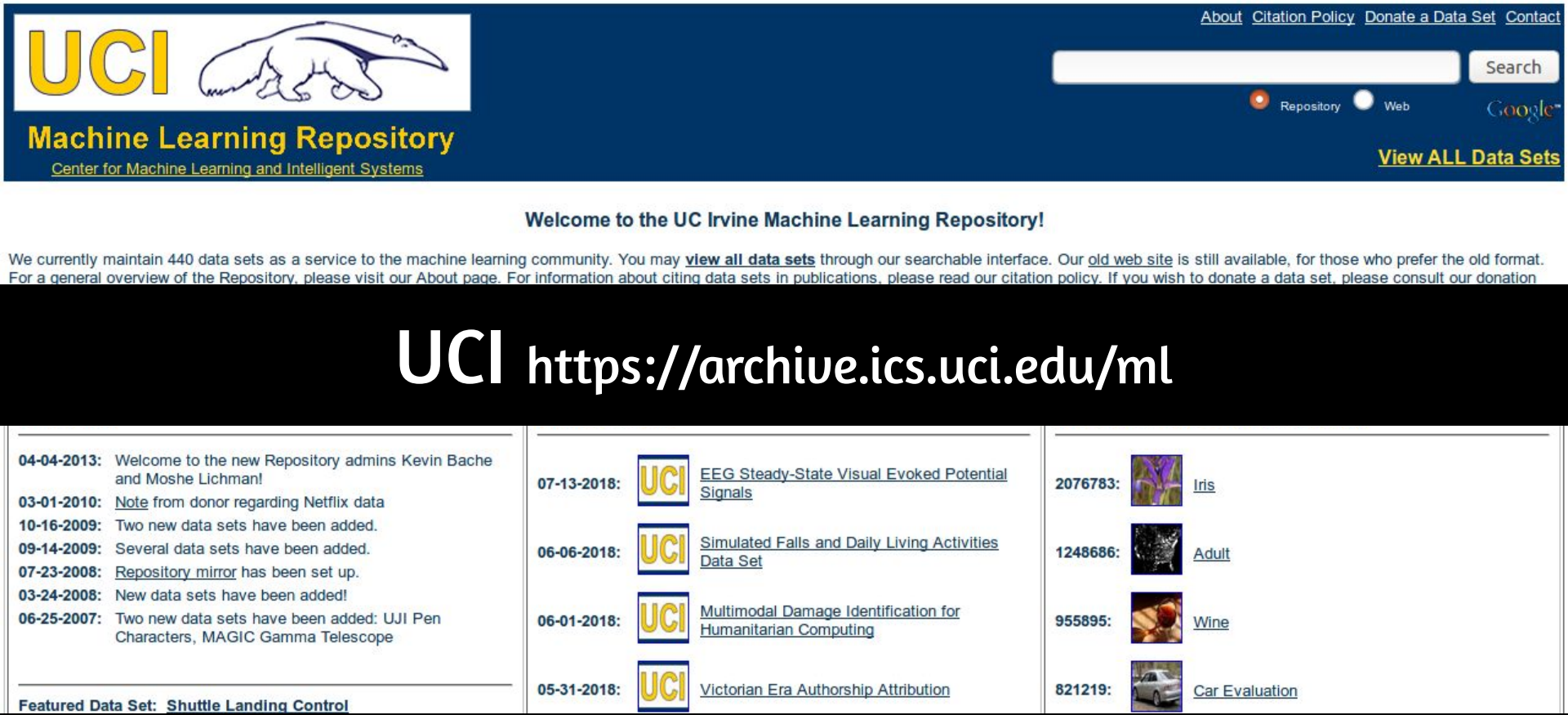
6. UVAD Dataset (Biometric Spoofing Detection)

Blogroll

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

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

Recent Posts

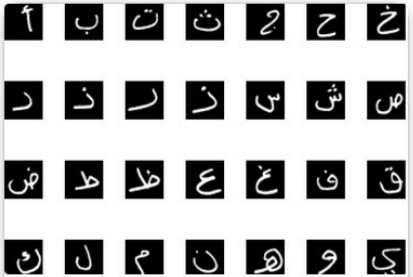

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contact
Search
Repository Web
View ALL Data Sets
Welcome to the UC Irvine Machine Learning Repository!
We currently maintain 440 data sets as a service to the machine learning community. 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

Out the Window

2019.8

Waymo Open Dataset

2019.6



Arabic Handwritten Characters Dataset

2017.8

Arabic Handwritten Digits Dataset

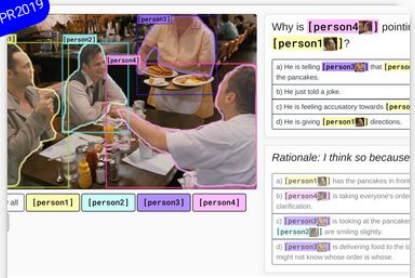
2016.10

VisualData https://www.visualdata.io/

Multiple Light Source Dataset

Contour Drawing Dataset

Pix3D



Visual Commonsense Reasoning (VCR)


BRASIL
Simplifique!
Participe
Acesso à informação
Legislação
Canais
Ir para o conteúdo 1
Ir para o menu 2
Ir para a busca 3
Ir para o rodapé 4
Pesquisar conjuntos de dados...
dados.gov.br
PORTAL BRASILEIRO DE DADOS ABERTOS
Dados | Organizações | Aplicativos | Inventários | Concursos | INDA | Perguntas frequentes | Contato | Sobre o portal
/ Conjuntos de dados
Organizações
Banco Central do Br... (3104)
Instituto Brasileir... (419)
Estado de Alagoas - AL (220)
Agência Nacional de... (210)
Agência Nacional do... (193)
Distrito Federal (168)
Buscar conjunto de dados...
7.064 conjuntos de dados encontrado(s)
Ordenar por:
Relevância
Alunos
Relação de todos os alunos ingressantes e formandos, por ciclo acadêmico.
Portal Brasileiro de Dados Abertos
http://dados.gov.br/dataset
Mostrar mais Organizações
Grupos
Bolsistas
Relação de todos os dados dos alunos bolsistas na instituição (assistência estudantil, PIBIC, BIC-JR, PIBIT, monitoria, PET)

For What?

Skin Cancer Classification

ISIC 2019 Challenge

“The goal is classify dermoscopic images among 9 different diagnostic categories”.

- Classification
  - Melanoma
  - Melanocytic nevus
  - Basal cell carcinoma
  - Actinic keratosis
  - Benign keratosis
  - Dermatofibroma
  - Vascular lesion
  - Squamous cell carcinoma
  - None of the others

Histopathologic Cancer Detection

Challenge 2019 @ Kaggle

“ To identify metastatic cancer in small image patches taken from larger digital pathology scans.”.

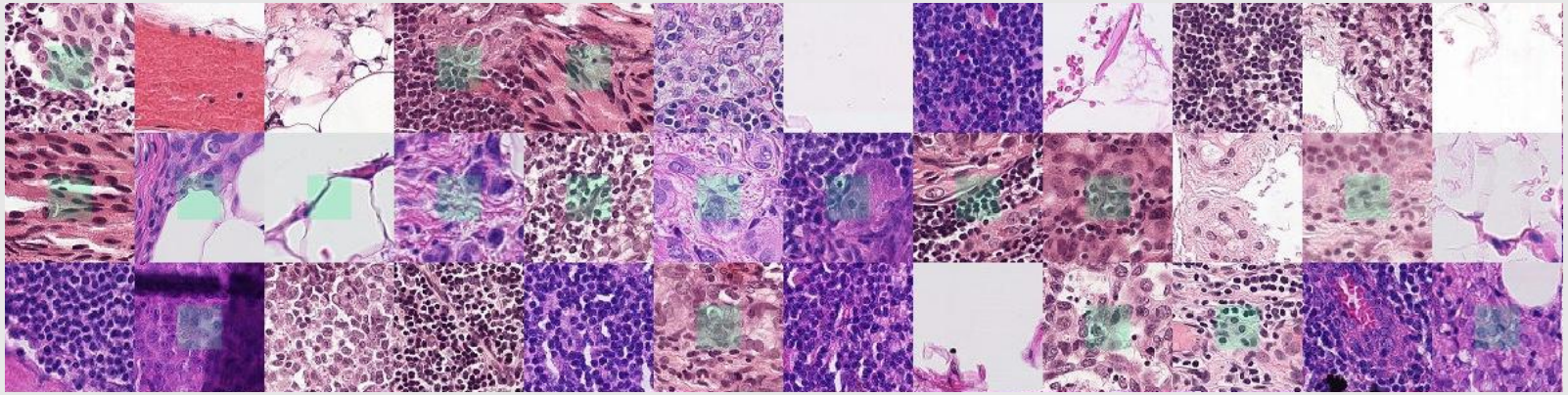


Image Captioning

COCO Captioning Challenge, Conceptual Captions

“*The goal is generate textual description from an image.*”

VQA: Visual Question Answering

[VQA Challenge 2019]

“Given an image and a natural language question about the image, the task is to provide an accurate natural language answer.”

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Large Vocabulary Instance Segmentation

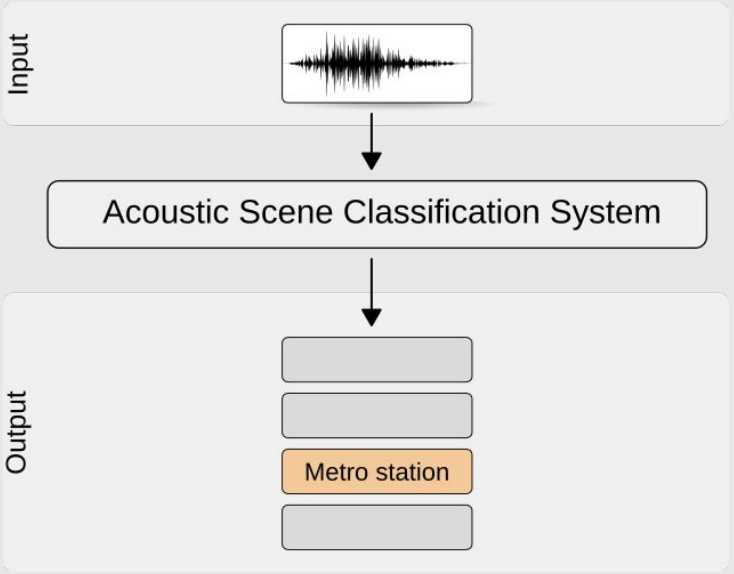
LVIS Challenge 2019

"LVIS presents a novel ***low-shot*** *object detection challenge to encourage new research in object detection."*

Detection and Classification of Acoustic Scenes and Events

DCASE Challenge 2019

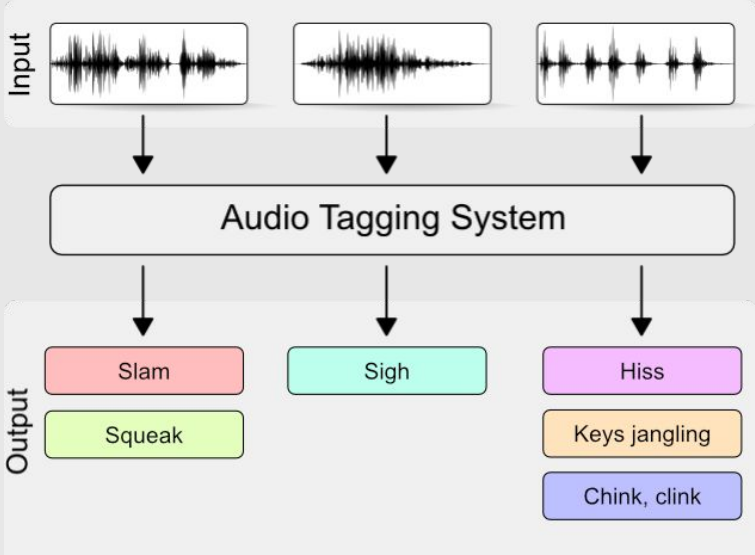
- Acoustic scene classification
- Audio tagging with noisy labels and minimal supervision
- Sound event localization and detection
- Sound event detection in domestic environments
- Urban sound tagging



Detection and Classification of Acoustic Scenes and Events

DCASE Challenge 2019

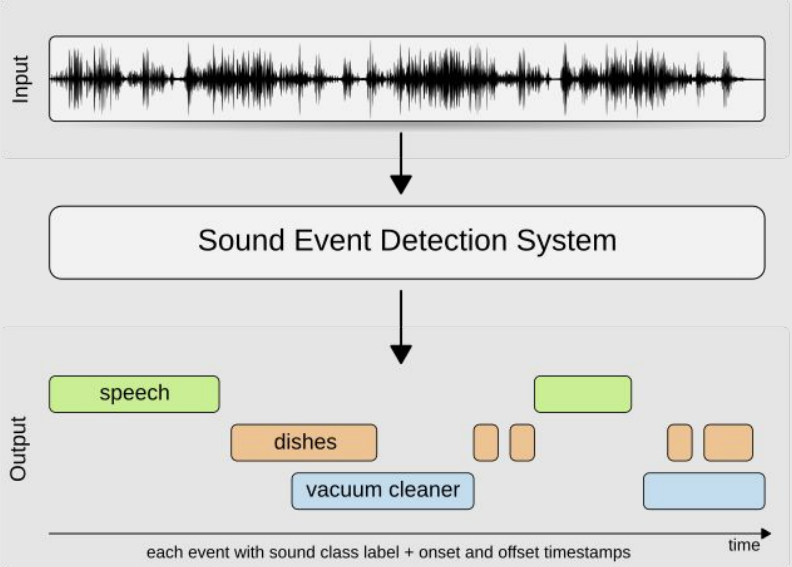
- Acoustic scene classification
- Audio tagging with noisy labels and minimal supervision
- Sound event localization and detection
- Sound event detection in domestic environments
- Urban sound tagging



Detection and Classification of Acoustic Scenes and Events

DCASE Challenge 2019

- Acoustic scene classification
- Audio tagging with noisy labels and minimal supervision
- Sound event localization and detection
- Sound event detection in domestic environments
- Urban sound tagging



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

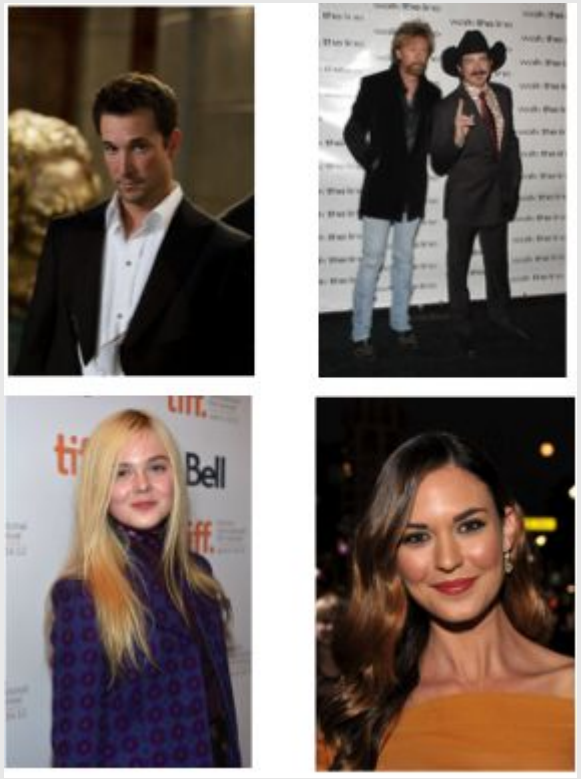
Facebook Comment Volume Dataset

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

Age and Gender Prediction

IMDB-Wiki

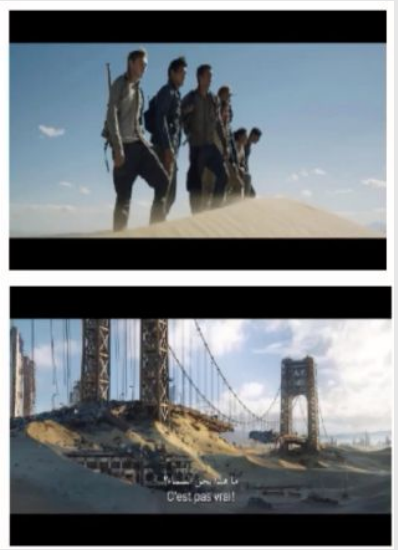
- Classification
  - Predict gender
- Regression
  - Predict age



Predicting Media Interestingness

Media Interestingness Data

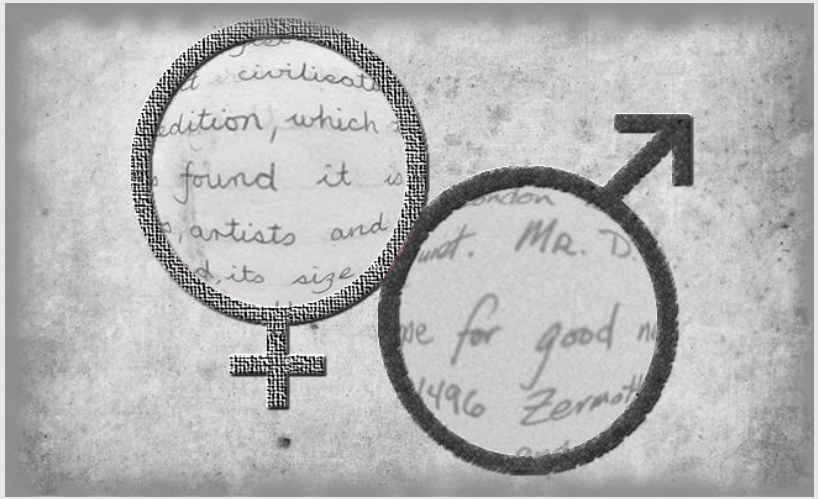
- Image, video, and metadata
- 5,054 samples (train) + 2,342 (test)
- Classification
  - Interesting
  - Not interesting



Gender Prediction from Handwriting

Handwriting Data

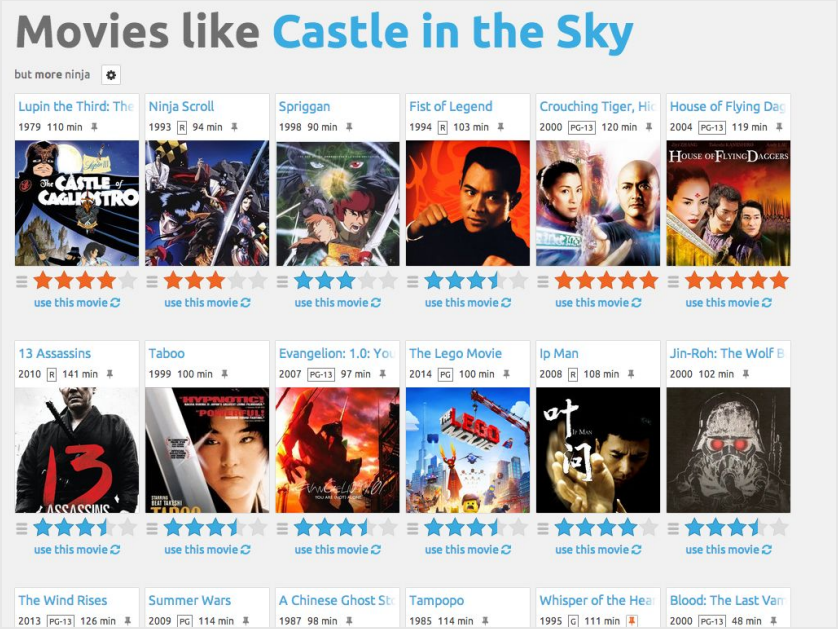
- Images in two languages (English, Arabic)
- Two pages for each language per writer
- Classification
  - Author’s gender from handwriting style



Recommendation System

MovieLens 1M dataset

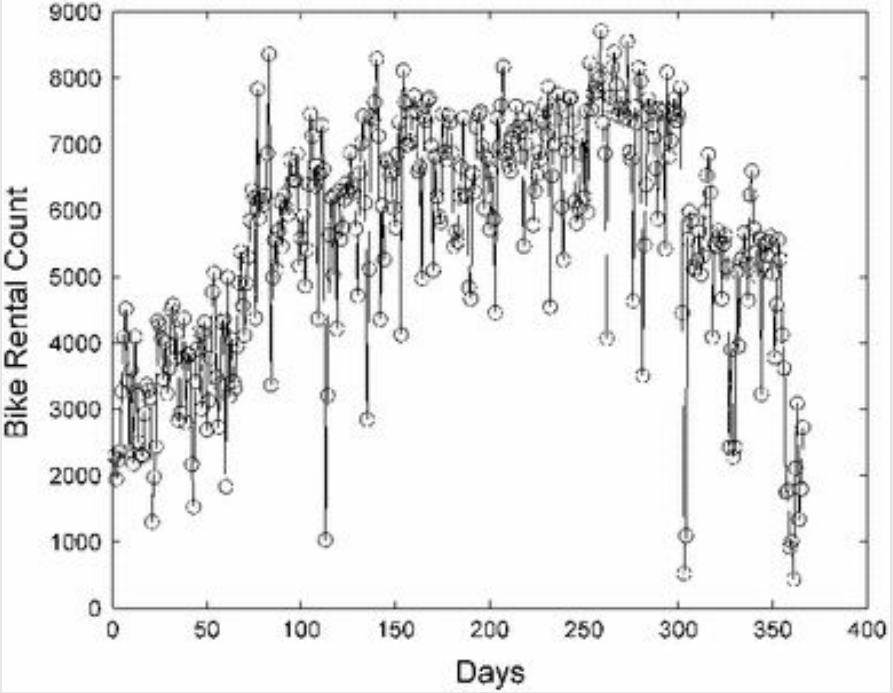
- 1M ratings from 6k users on 4k movies
- Regression
  - Predict ratings (1 to 5)



Bike Sharing

Bike Sharing Dataset

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



AI for Social Good

- Education
- Protecting democracy
- Urban planning
- Assistive technology for people with disabilities
- Health
- Agriculture
- Environmental sustainability
- Social welfare and justice
- Sustainable development

AI for Social Good

https://aiforsocialgood.github.io/neurips2019/

NeurIPS 2018
ICLR 2019
ICML 2019

Joint Workshop on AI for Social Good Workshop

This workshop builds on our AI for Social Good workshop at NeurIPS 2018, ICLR 2019 and ICML 2019.

The accelerating pace of intelligent system research and real world deployment presents three clear challenges for producing "good" intelligent systems: (1) the research community lacks incentives and venues for results centered on social impact, (2) deployed systems often produce unintended negative consequences, and (3) there is little consensus for public policy that maximizes "good" social impacts, while minimizing the likelihood of harm. As a result, researchers often find themselves without a clear path to positive real world impact. The Workshop on AI for Social Good addresses these challenges by bringing together machine learning researchers, social impact leaders, ethicists, and public policy leaders to present their ideas and applications for maximizing the social good. This workshop is a collaboration of three formerly separate lines of research (i.e., this is a "joint" workshop), including researchers in applications-driven AI research, applied ethics, and AI policy. Each of these research areas are unified into a 3-track framework promoting the exchange of ideas between the practitioners of

Deadline for submissions
September 6th
11:59PM PST

Notification of acceptance:
September 24th

Camera-ready submission:
November 17th


Be bold! Be brave!
Credit: https://www.maxtreme.com.br