Machine Learning Datasets

Why? Which? For what?

Samuel G. Fadel
September, 2017

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)

- 40+ classes and ~50k images
- First system to beat humans in visual pattern recognition
  - Error rate: 0.56%
  - Human error rate: 1.16%
  - Closest competitor error rate: 3.86%

Superhuman pattern recognition

IJCNN Traffic Sign Recognition Competition (2011)

- Why was ImageNet 2012 more memorable?

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?


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”*

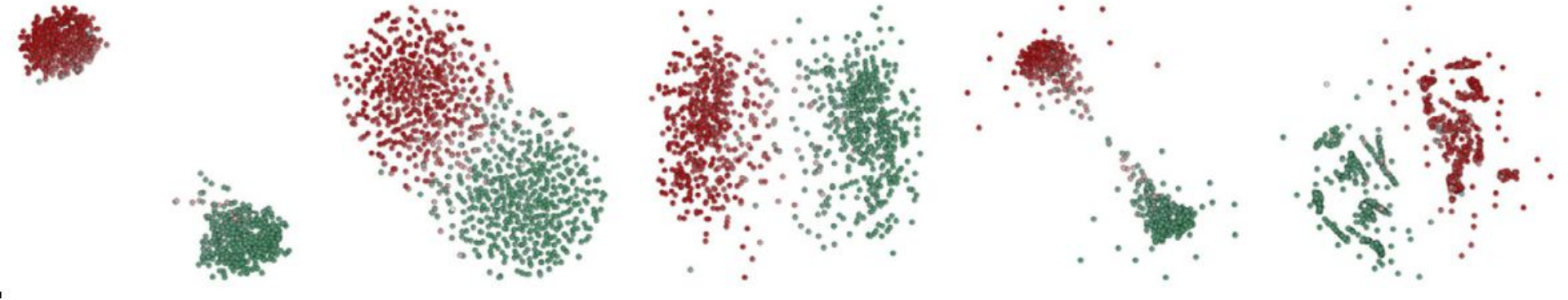
*“experiments are performed on a set of (rather **artificial**) data sets”*

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

Which?

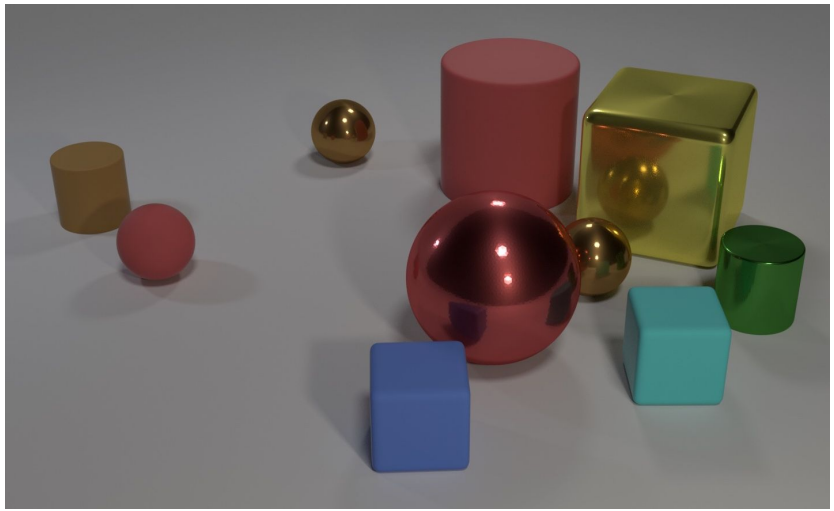
Toy datasets illustrate concepts and are easy to interpret

- 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**?

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

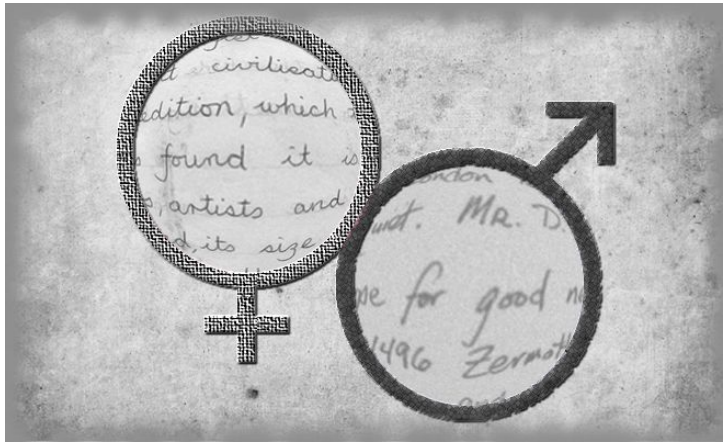
[Media Interestingness Data]

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

Gender prediction from handwriting

Handwriting Data

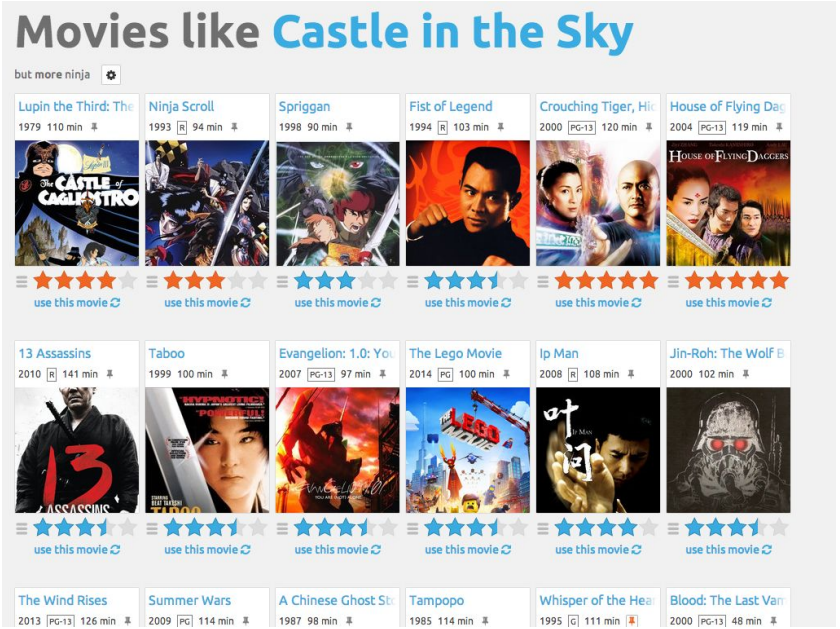
- Images in two languages (English, Arabic)
- Two pages for each language per writer
- Classification
  - Author’s gender from handwriting style



Recommendation system

[MovieLens 1M dataset]

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



Bike Sharing

[Bike Sharing Dataset]

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