Feature Representation and Dimensionality Reduction

Spring 2018

Review

- Last week:
  - K-Nearest Neighbors
  - Support Vector Machine (SVM)
  - Cross-Validation
  - Tuning Hyperparameters

- Assignments (Canvas)
  - Lab assignment due yesterday
  - New problem set out

- Questions?

Today’s Topics

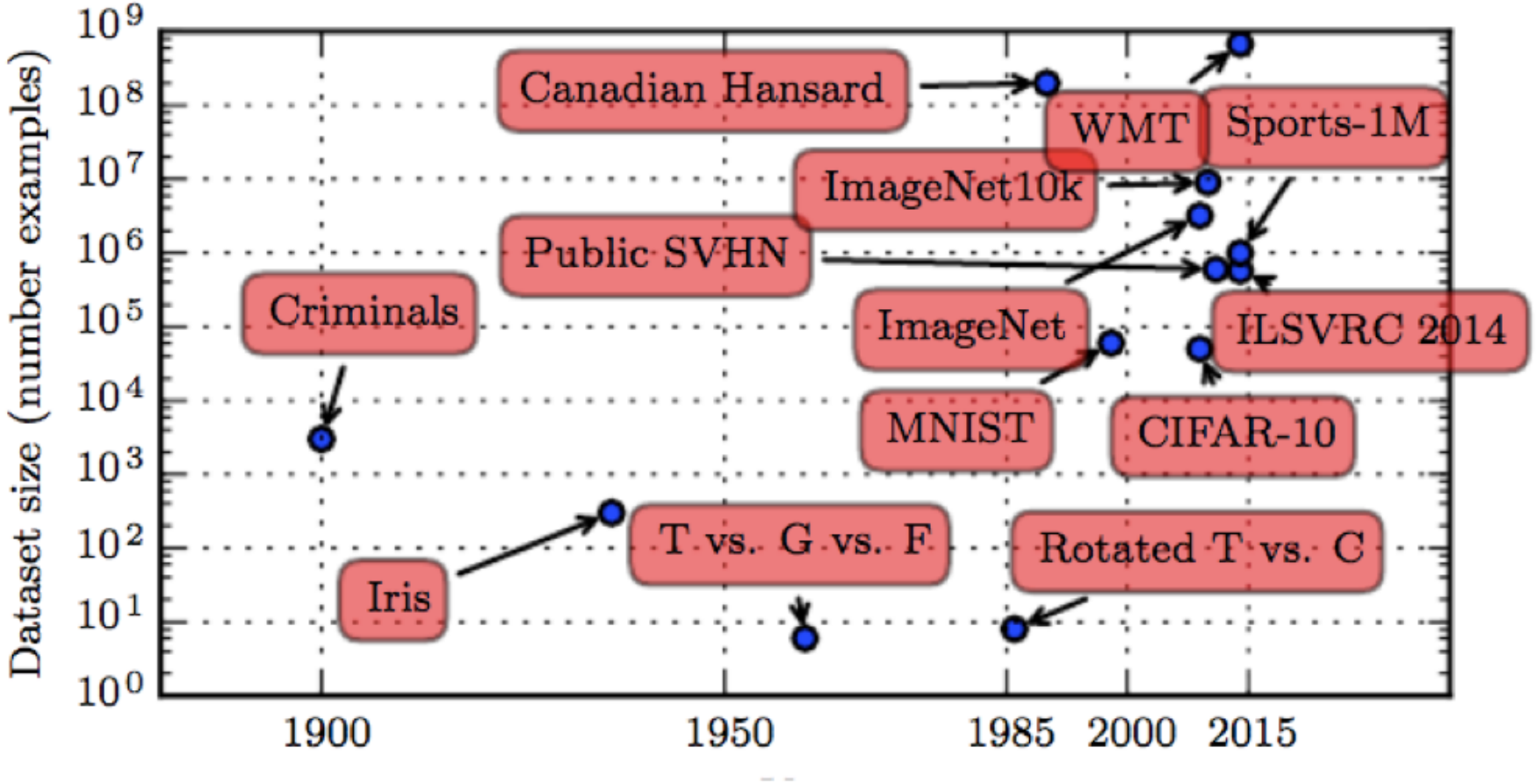
- Real-world community challenges
- Feature Representation
- Dimensionality Reduction
- Classification Evaluation
- Lab

Today’s Topics

- Real-world community challenges
- Feature Representation
- Dimensionality Reduction
- Classification Evaluation
- Lab

Mapping Academic Research Challenges to Modern Technology You Can Use Today

- "Big"-ger Data
  - e.g., internet



Deep Learning, Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Recall Learning Challenge: Sufficient Training

e.g., images

images on basic hard drive: 10^5
(500 GB/2 MB = 250,000)

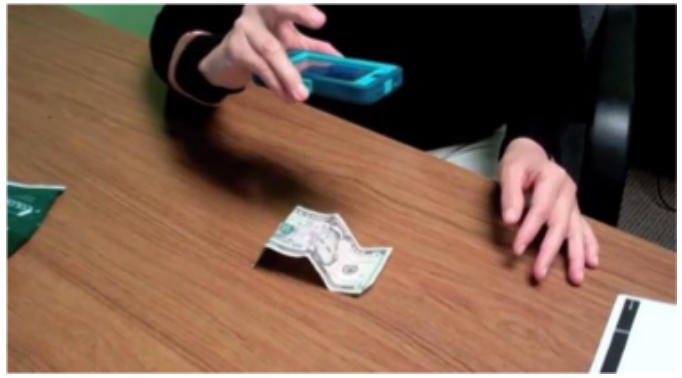
images seen during my first 10 years: 10^9
(24 images/sec * 60 sec * 60 min * 16 hr * 365 days * 10 yrs = 5,045,760,000)

images seen by all humanity: 10^{20}
(7.5 billion humans[1] * 24 images/sec * 60 * 60 * 16 * 365 * 60 yrs = $2.23 * 10^{20}$)

[1] http://www.worldometers.info/world-population/

Adapted from slides by Antonio Torralba

Object Recognition: Industry Products

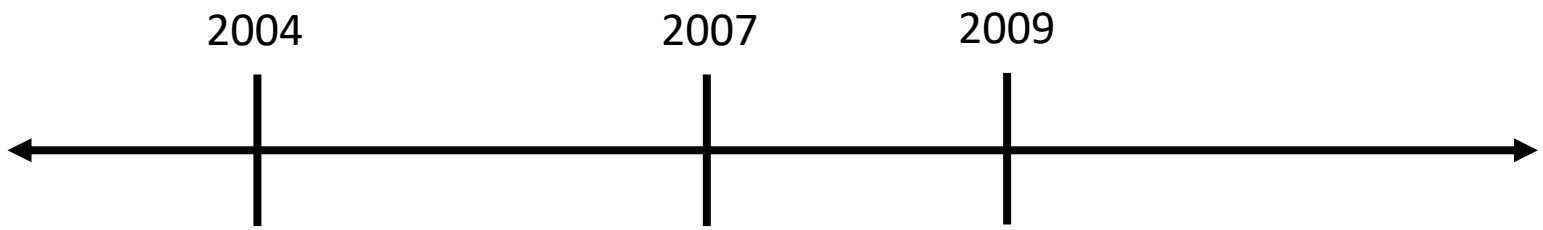


TapTapSee: Assist People with Visual Impairments

e.g., take a picture of an object and find where to buy it

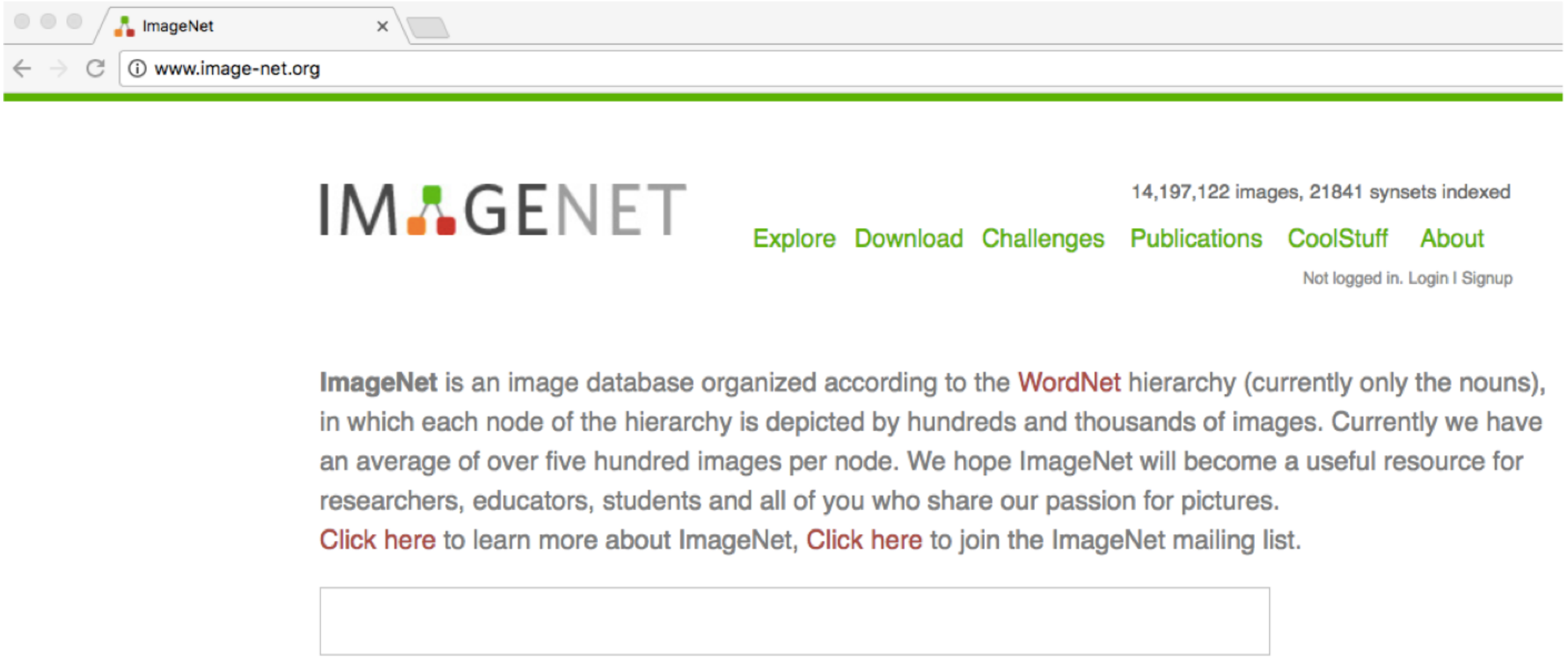
Leafsnap
(Species Identification)

Object Recognition: Academic Research

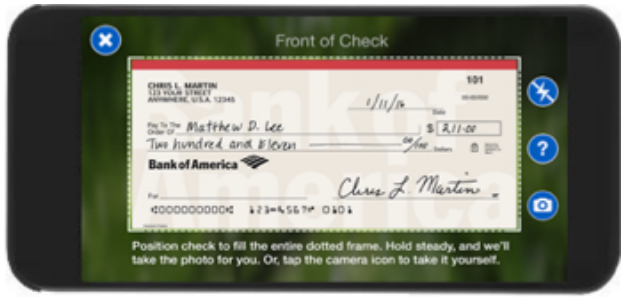


Dataset:	Caltech-101	Caltech-256	CIFAR	ImageNet
# Categories:	101	256	100	1,000
Minimum # Samples/Category:	31+	80+	600	668-3047
# Images:	9,144	30,607	114,000	1,461,406

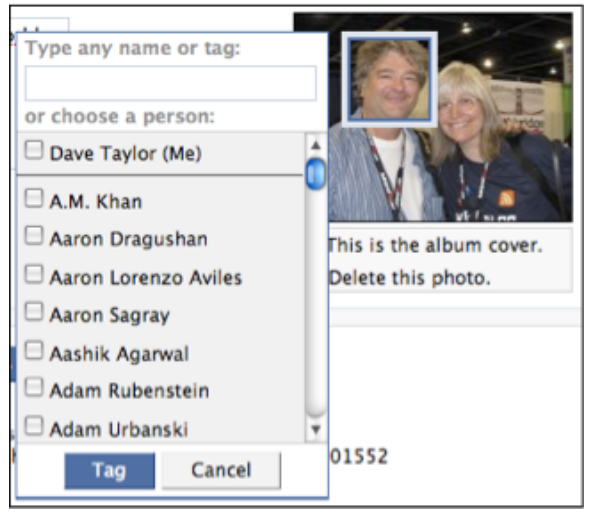
Object Recognition: Premiere World Challenge



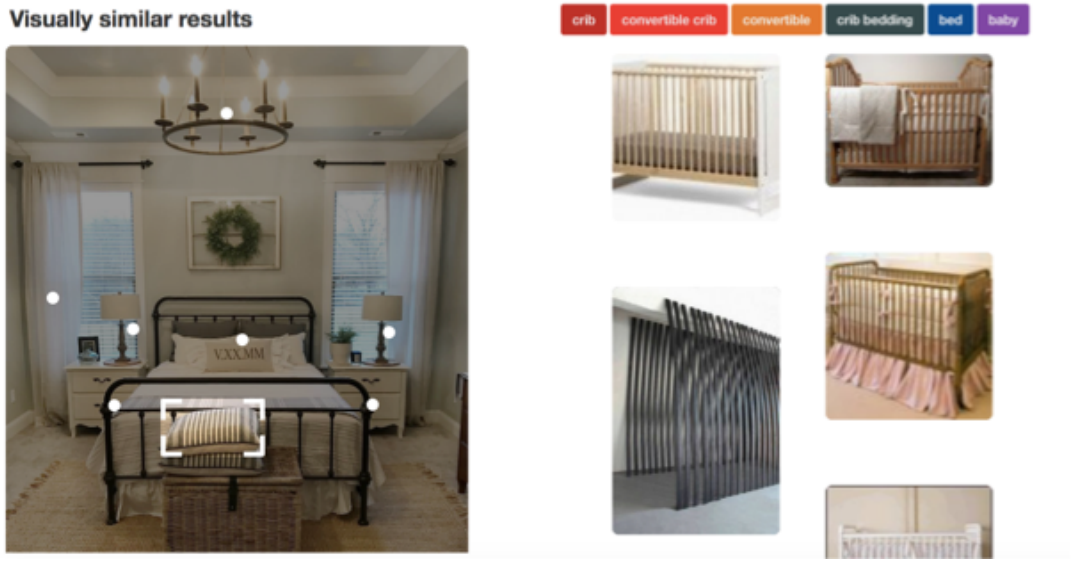
Object Detection: Industry Products



Mobile check deposit
(e.g., Bank of America)



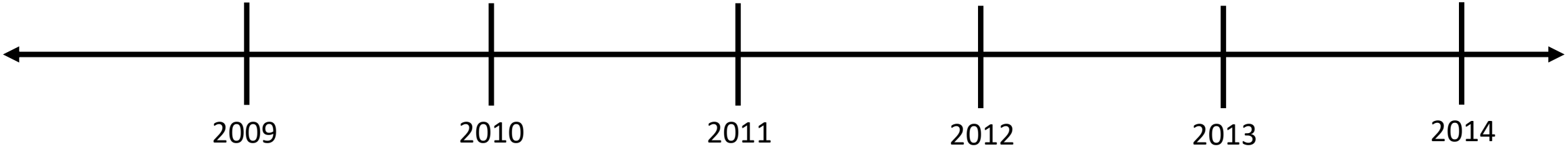
Face detection
(e.g., Facebook)



Search for specific items (e.g., Pinterest)

License Plate Detection (e.g., AllGoVision)

Object Detection: Academic Research



Dataset	PASCAL -VOC	PASCAL -VOC	PASCAL -VOC	PASCAL -VOC	ILSVRC	ILSVRC
# Categories				20	200	200
# Images				21,738	456,182	516,840

Object Detection: A World Challenge

The PASCAL Visual Object Classes Homepage

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

The PASCAL VOC project:

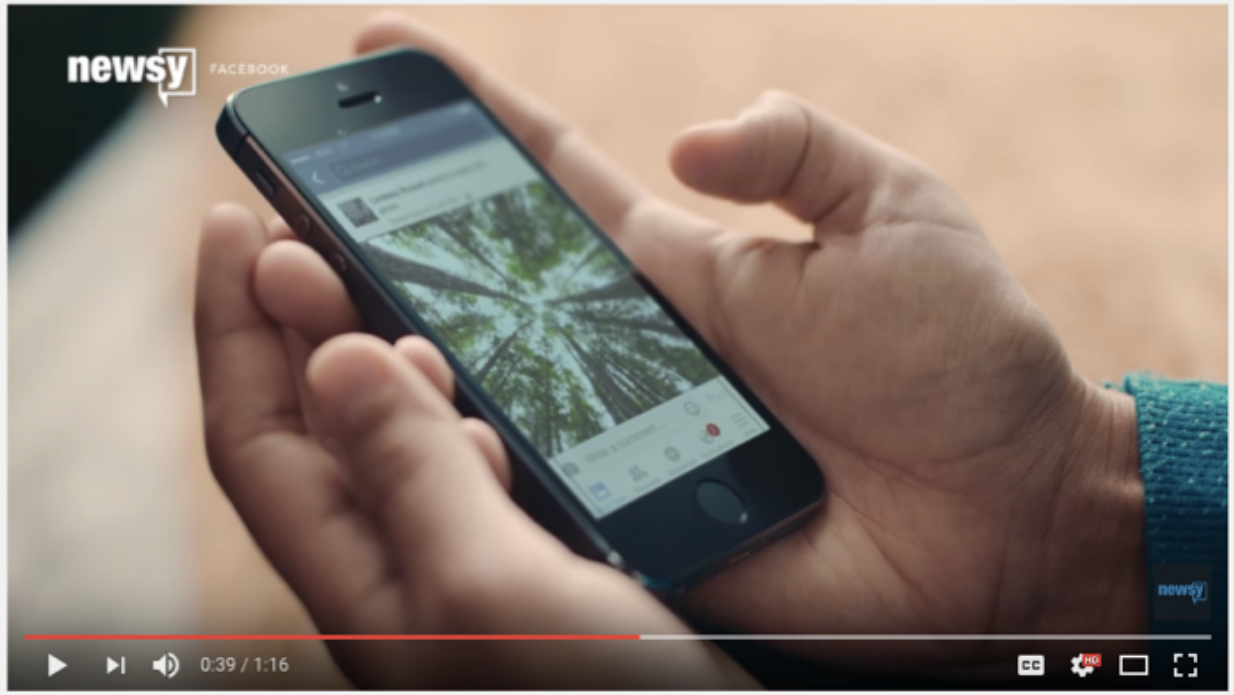
- Provides standardised image data sets for object class recognition
- Provides a common set of tools for accessing the data sets and annotations
- Enables evaluation and comparison of different methods
- Ran challenges evaluating performance on object class recognition (from 2005-2012, now finished)

Pascal VOC data sets

Data sets from the VOC challenges are available through the challenge links below, and evalution of new methods on these data sets can be achieved through the PASCAL VOC Evaluation Server. The evaluation

Image Descriptions: Industry Products

Facebook



Facebook's New AI Tool Is Helping Blind Users 'See' Photos - Newsy

https://www.youtube.com/watch?v=Tjugc8a836Q

Microsoft

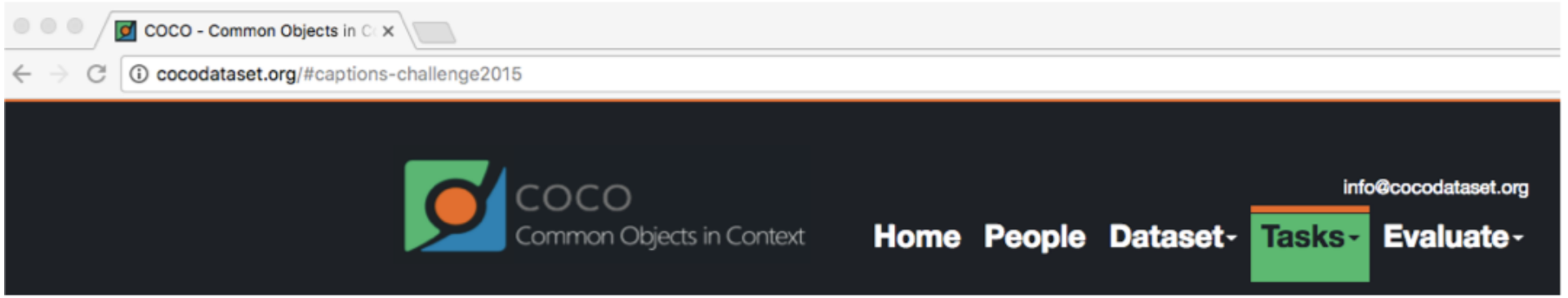
Saqib Shaikh : Microsoft Developer Can 'See' Using Artificial Intelligence Headset

https://www.youtube.com/watch?v=R2mC-NUAmMk

Image Description: Academic Research

	Images	Texts	Judgments	Objects
Pascal1K (Rashtchian et al., 2010)	1,000	5	No	Partial
VLT2K (Elliott & Keller, 2013)	2,424	3	Partial	Partial
Flickr8K (Hodosh & Hockenmaier, 2013)	8,108	5	Yes	No
Flickr30K (Young et al., 2014)	31,783	5	No	No
Abstract Scenes (Zitnick & Parikh, 2013)	10,000	6	No	Complete
IAPR-TC12 (Grubinger et al., 2006)	20,000	1–5	No	Segmented
MS COCO (Lin et al., 2014)	164,062	5	Collected	Partial
BBC News (Feng & Lapata, 2008)	3,361	1	No	No
SBU1M Captions (Ordonez et al., 2011)	1,000,000	1	Collected[7]	No
Déjà-Image Captions (Chen et al., 2015)	4,000,000	Varies	No	No

Image Description: Premiere World Challenge



COCO Captioning Challenge

Winners were announced at CVPR 2015

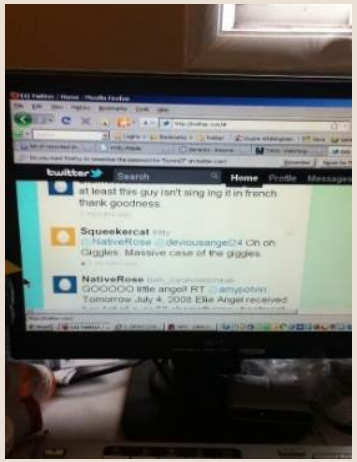
Caption evaluation server remains open!

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Visual Question Answering: Online Demos

Demo: http://visualqa.csail.mit.edu/



Is my monitor on?

Hi there can you please tell me what flavor this is?

Does this picture look scary?

Which side of the room is the toilet on?

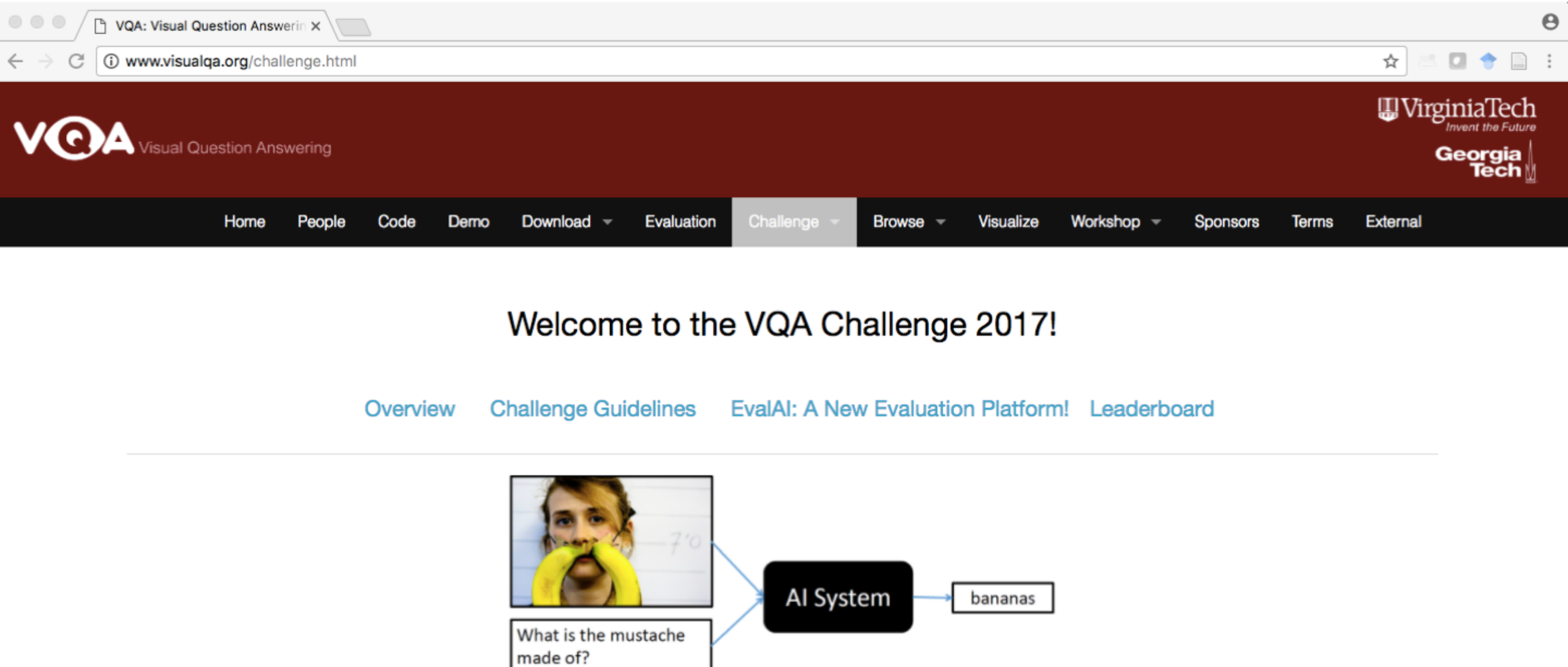
Asked by sighted and blind people

Visual Question Answering: Academic Research

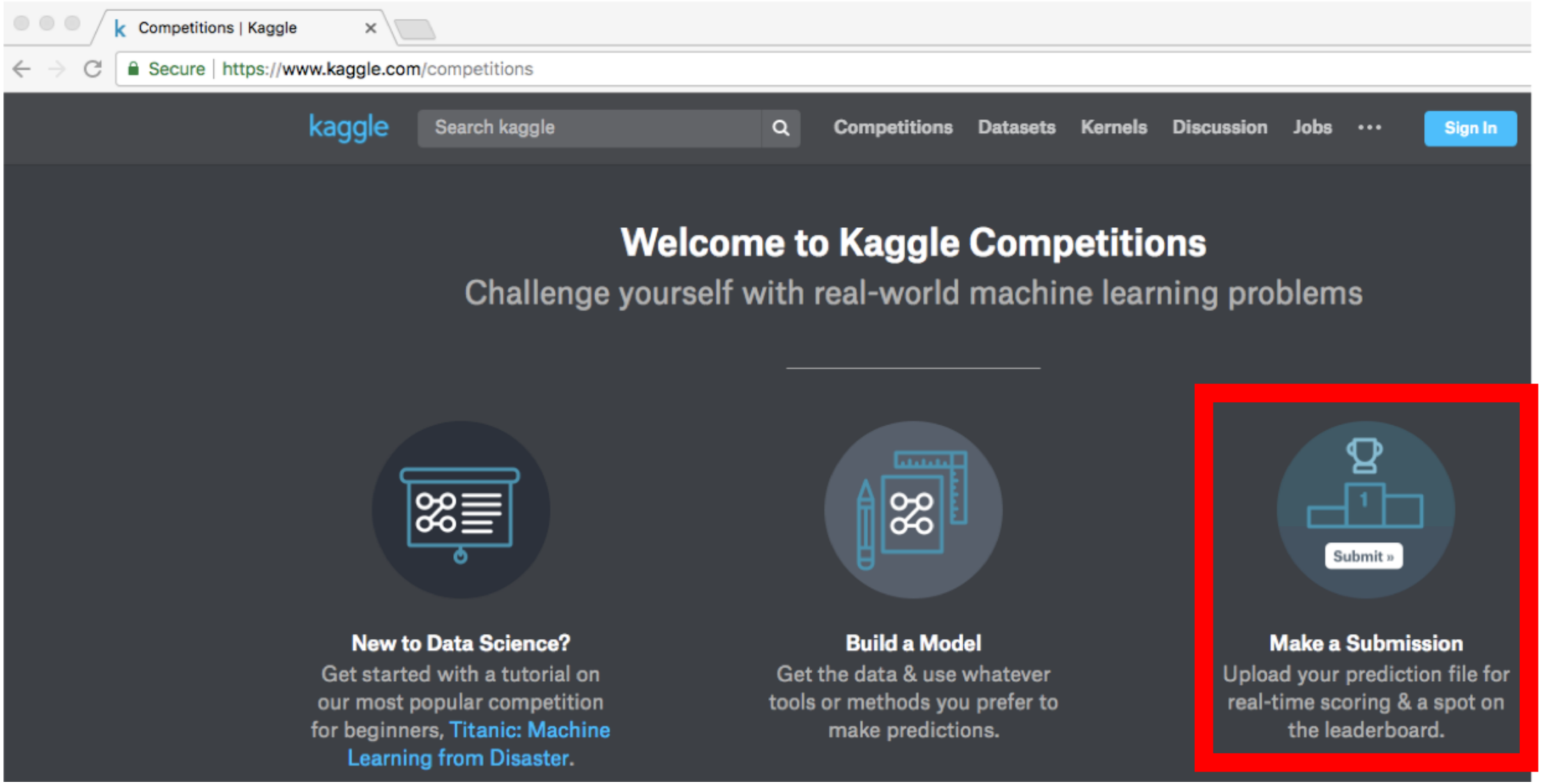
2015-2017

Dataset	Which Images?	Who Asked?	How Asked?
DAQUAR [20]	NYU Depth V2 [23]	Automatically generated (templates)	———
VQA v1.0: Abstract [5]	Abstract Scenes	Crowd workers (AMT)	Typed
VQA v1.0: Real [5]	MSCOCO [18]	Crowd workers (AMT)	Typed
Visual Madlibs [29]	MSCOCO [18]	Automatically generated (templates)	———
FM-IQA [8]	MSCOCO [18]	Crowd workers (Baidu)	Typed
KB-VQA [27]	MSCOCO [18]	In-house participants	Typed
COCO-QA [22]	MSCOCO [18]	Automatically generated (captions)	———
VQA v2.0: Real [9]	MSCOCO [18]	Crowd workers (AMT)	Typed
Visual7W [30]	MSCOCO [18]	Crowd workers (AMT)	Typed
CLEVR [12]	Synthetic Shapes	Automatically generated (templates)	———
SHAPES [4]	Synthetic Shapes	Automatically generated (templates)	———
Visual Genome [17]	MSCOCO [18] & YFCC100M [1]	Crowd workers (AMT)	Typed
FVQA [26]	MSCOCO [18] & ImageNet [7]	In-house participants	Typed
TDIUC [13]	MSCOCO [18] & YFCC100M [1]	Crowd workers (AMT), In-house participants, Automatically generated	Typed

Visual Question Answering: Premiere World Challenge



Kaggle: Challenges Available to All!



What Challenges Typically Have in Common:

1. Publicly-shared train (and validation) dataset with “ground truth” labels
2. Publicly-shared test dataset (“ground truth” labels are hidden)
3. Metrics for evaluating algorithm-generated results on the test set

Why Have Challenges?

- Provide “fair” comparison between algorithms
- Create a community around a shared goal

My Challenge in Fall 2018 in Germany… ☺

VizWiz | VizWiz Dataset | Publications | Contact

VizWiz Dataset

VizWiz dataset was collected using the VizWiz application, which was released in May 2011.

- 11,045 users asked 72,205 visual questions, 48,169 of which were asked by users who agreed to allow their questions to be anonymized and shared. We carefully checked and removed those containing personally identifying information or indecent content, resulting in 31,073 remaining visual questions.
- Each visual question includes an image, a transcription of the question, and 10 answers crowdsourced from Amazon Mechanical Turk workers. Visual questions originated from blind people who each took a picture using a mobile phone and recorded a spoken question about it.

VizWiz v1.0 dataset download:

- 20,000 training images
- 20,000 training questions
- 200,000 training answers
- 3,173 validation images
- 3,173 validation questions
- 31,730 validation answers
- 8,000 test images
- 8,000 test questions

Today’s Topics

- Real-world community challenges
- Feature Representation
- Dimensionality Reduction
- Classification Evaluation
- Lab

Real World Data Challenges

- Different data representations
- Missing data
- Different numerical scales

e.g.,

	Categorical	Numerical	
Dataset	Class Length	Rain (cm)	Attend Class?
Train	Short	1.1	Yes
Train	Medium	2.3	Yes
Train	Medium	0	No
Train	Long	0.7	No
Train	Medium	0.3	Yes
Train	Short	1.5	No
Train	Short	0	Yes
Train	Medium	1.5	Yes
Train	Medium	0.7	Yes
Train		0.6	Yes
Train	Long		Yes
Test	Medium	0.1	Yes
Test	Short		Yes
Test		0.5	No

Categorical Variables

- Categorical
  - Nominal (2 or more categories with no ordering)
    - e.g., gender
  - Ordinal (categories with clear ordering)
    - e.g., t-shirt size, education level
- How to convert categorical to numerical variable?
  - Bad idea to map each category to a number

e.g.,

Categorical

Dataset	Class Length	Rain (cm)	Attend Class?
Train	Short	1.1	Yes
Train	Medium	2.3	Yes
Train	Medium	0	No
Train	Long	0.7	No
Train	Medium	0.3	Yes
Train	Short	1.5	No
Train	Short	0	Yes
Train	Medium	1.5	Yes
Train	Medium	0.7	Yes
Train		0.6	Yes
Train	Long		Yes
Test	Medium	0.1	Yes
Test	Short		Yes
Test		0.5	No

Categorical Variables: One-Hot Encoding

- One-hot encoding: add one new feature per category
- How many features will be made for “Type”?
  - 2
- How many features will be made for “Length”?
  - 3
- How many features would the example dataset have with a one-hot encoding?
  - 6

e.g.,

Type	Length	IMDb_Rating	Liked
Comedy	Short	7.2	Yes
Drama	Medium	9.3	Yes
Comedy	Medium	5.1	No
Drama	Long	6.9	No
Drama	Medium	8.3	Yes
Drama	Short	4.5	No
Comedy	Short	8.0	Yes
Drama	Medium	7.5	Yes

Categorical Variables: One-Hot Encoding

IMDb_Rating	Type_Comedy	Type_Drama	Length_Long	Length_Medium	Length_Short
7.2	1	0	0	0	1
9.3	0	1	0	1	0
5.1	1	0	0	1	0
6.9	0	1	1	0	0
8.3	0	1	0	1	0
4.5	0	1	0	0	1
8.0	1	0	0	0	1
7.5	0	1	0	1	0

- What new challenges arise?
  - Large, sparse matrices
  - Test set may have value not observed in training

Missing Data

- How to replace missing values?
  - Ignore the tuple
  - Manually insert missing values
  - Insert global constant (e.g., 0)
  - Attribute mean
  - Attribute mean for all samples belonging to same class
  - And more…
- Algorithm
  1. Learn on training data
  2. Transform training data
  3. Transform test data

e.g.,

Dataset	Class Length	Rain (cm)	Attend Class?
Train	Short	1.1	Yes
Train	Medium	2.3	Yes
Train	Medium	0	No
Train	Long	0.7	No
Train	Medium	0.3	Yes
Train	Short	1.5	No
Train	Short	0	Yes
Train	Medium	1.5	Yes
Train	Medium	0.7	Yes
Train		0.6	Yes
Train	Long		Yes
Test	Medium	0.1	Yes
Test	Short		Yes
Test		0.5	No

Missing Data: Impute mean values for rain

- Algorithm
  1. Learn on training data
  2. Transform training data
  3. Transform test data
- What is the value to impute?
  - 8.7/10 = 0.87

e.g.,

Dataset	Class Length	Rain (cm)	Attend Class?
Train	Short	1.1	Yes
Train	Medium	2.3	Yes
Train	Medium	0	No
Train	Long	0.7	No
Train	Medium	0.3	Yes
Train	Short	1.5	No
Train	Short	0	Yes
Train	Medium	1.5	Yes
Train	Medium	0.7	Yes
Train		0.6	Yes
Train	Long		Yes
Test	Medium	0.1	Yes
Test	Short		Yes
Test		0.5	No

Missing Data: Impute mean values for rain for all samples belonging to same class

- Algorithm
  1. Learn on training data
  2. Transform training data
  3. Transform test data
- What is the value to impute?
  - "Yes": 6.5/7 = 0.93

e.g.,

Dataset	Class Length	Rain (cm)	Attend Class?
Train	Short	1.1	Yes
Train	Medium	2.3	Yes
Train	Medium	0	No
Train	Long	0.7	No
Train	Medium	0.3	Yes
Train	Short	1.5	No
Train	Short	0	Yes
Train	Medium	1.5	Yes
Train	Medium	0.7	Yes
Train		0.6	Yes
Train	Long		Yes
Test	Medium	0.1	Yes
Test	Short		Yes
Test		0.5	No

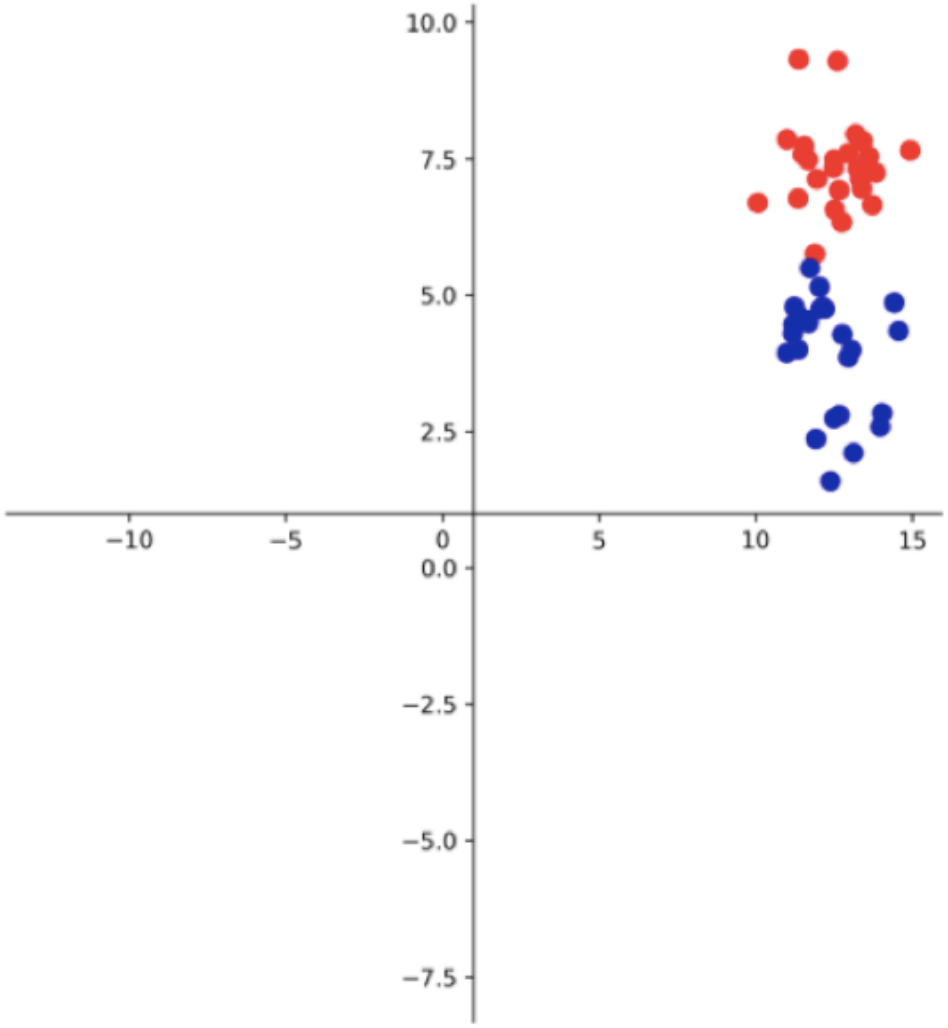
Different numerical scales

e.g.,

Numerical

Dataset	Class Length	Rain (cm)	Attend Class?
Train	Short	1.1	Yes
Train	Medium	2.3	Yes
Train	Medium	0	No
Train	Long	0.7	No
Train	Medium	0.3	Yes
Train	Short	1.5	No
Train	Short	0	Yes
Train	Medium	1.5	Yes
Train	Medium	0.7	Yes
Train		0.6	Yes
Train	Long		Yes
Test	Medium	0.1	Yes
Test	Short		Yes
Test		0.5	No

Different numerical scales



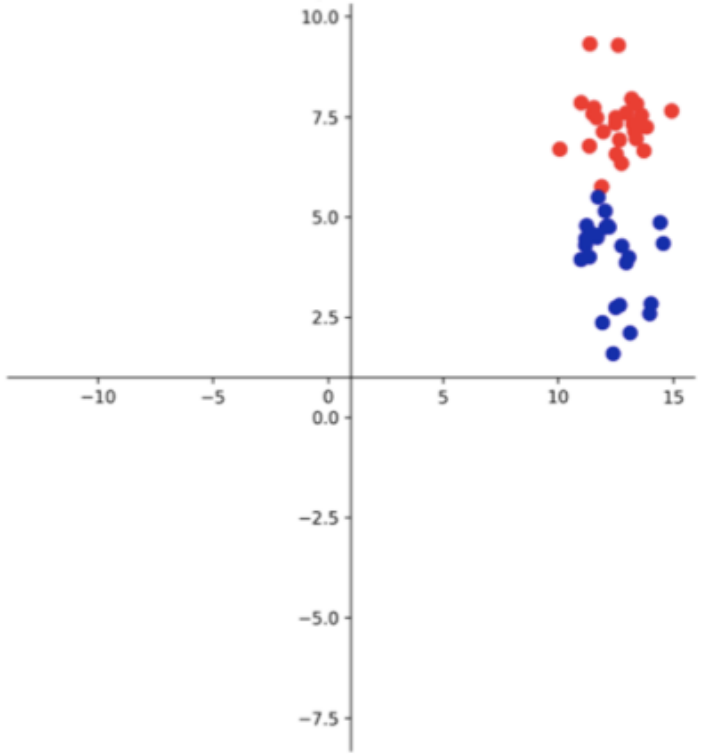
What is range of feature 1 values?

What is range of feature 2 values?

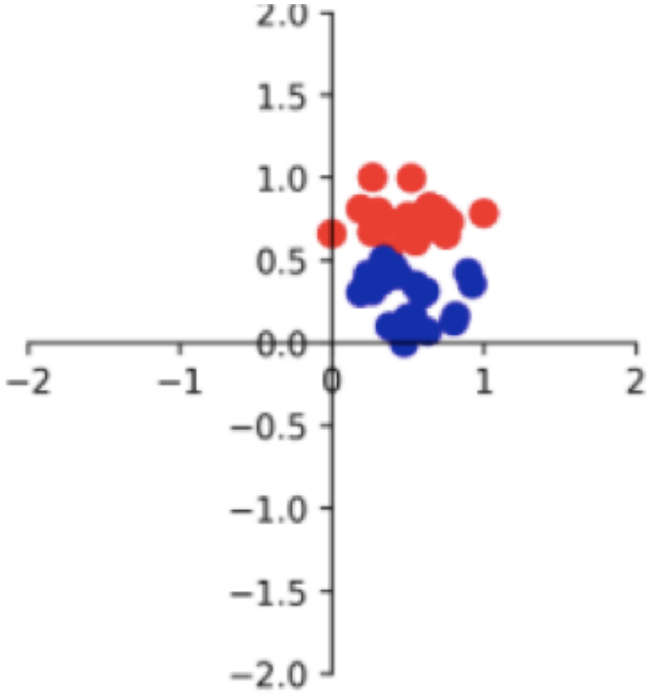
https://github.com/amueller/introduction_to_ml_with_python/blob/master/03-unsupervised-learning.ipynb

Different numerical scales: Solutions

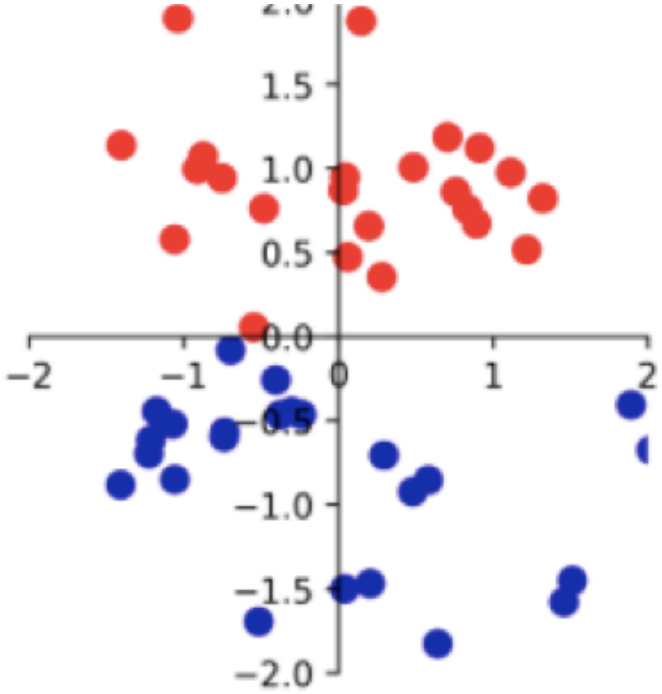
Original Data



Min-Max Scaling



Standardization



https://github.com/amueller/introduction_to_ml_with_python/blob/master/03-unsupervised-learning.ipynb

Different numerical scales: Solutions

- Scaling: puts numerical attributes onto the same scale
  - Min-max scaling: shifts and rescals to range from 0 to1
    - Subtract min value and then divide by the max – min
    - Strength: Bounds values to a specific range
  - Standardization: ensures mean is zero and variance is 1
    - Subtract mean and then divide by the variance
    - Strength: Less affected by outliers
- Algorithm
  1. Learn on training data
  2. Transform training data
  3. Transform test data

Today’s Topics

- Real-world community challenges
- Feature Representation
- Dimensionality Reduction
- Classification Evaluation
- Lab

Problems with High Dimensional Data?

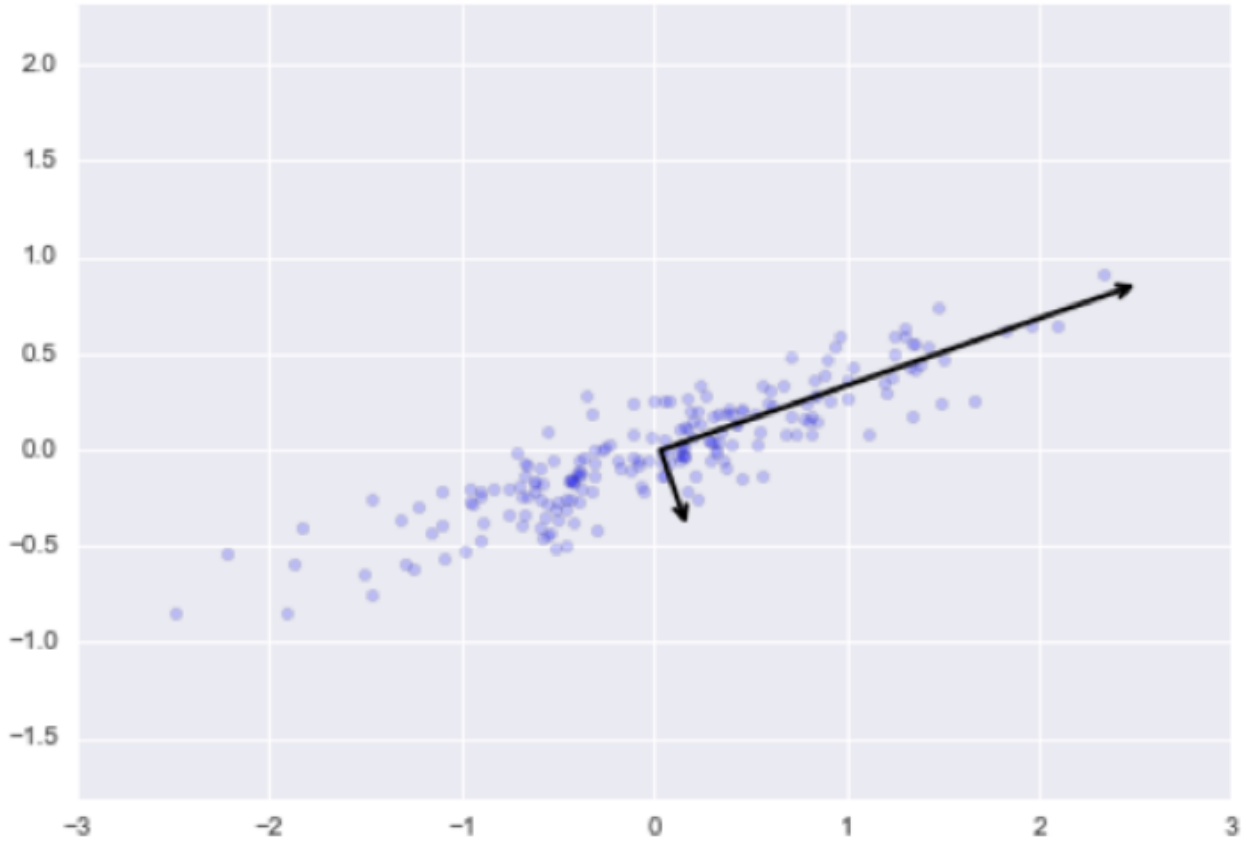
- What are problems of having many features for machine learning?
  - Slower training
  - Slower testing
  - Can be harder to find a good solution, due to greater risk of overfitting
  - Requires lots of memory

Projection Approaches

- Premise:
  - Many features are almost constant
  - Many features are highly correlated; e.g., age and height; degree and job title
- Idea:
  - Training instances actually lie within (or close to) a much lower-dimensional subspace of the high-dimensional space
- Approach:
  - drop dimension(s)

Projection: PCA (Unsupervised)

- Idea: find principle axes and keep most important ones
- Vectors: *principal axes* of data,
- Vector length: variance of the data described when its projected onto that axis.

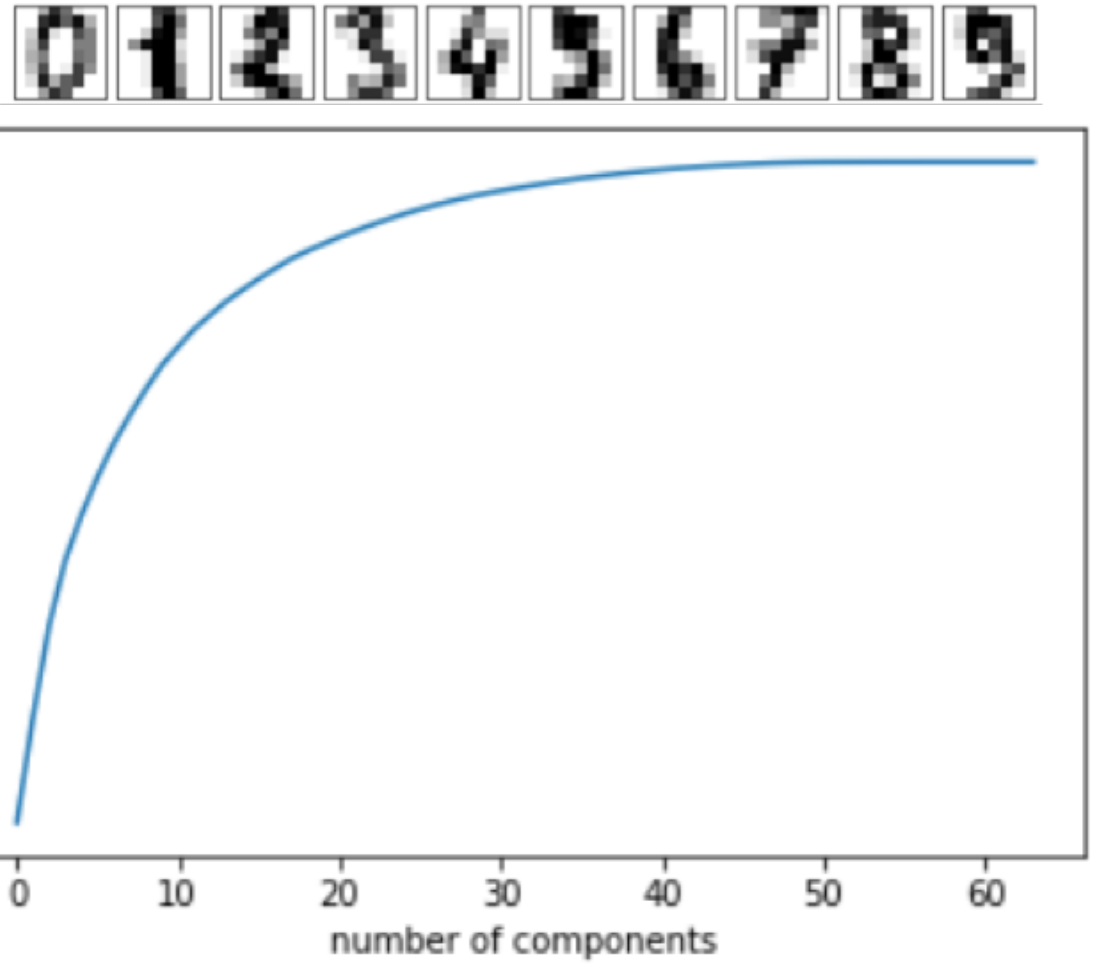


https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

Projection: PCA (Unsupervised)

- Assumption:
  - Data is linearly separable
- Algorithm
  1. Standardize data (recall this centers data around origin)
  2. Construct covariance matrix
  3. Obtain eigenvalues and eigenvectors
  4. Sort eigenvalues by decreasing order to rank eigenvectors
- Key Question: how many principle components to keep?

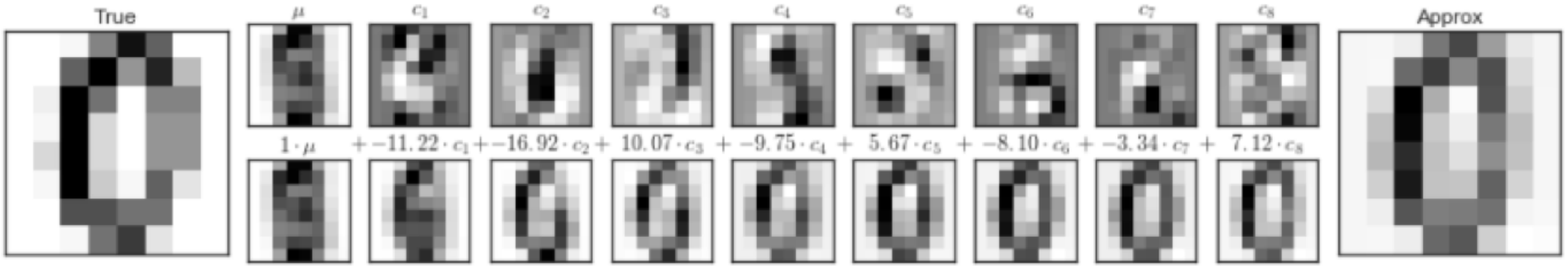
e.g., data with 64 initial values



https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

Projection: PCA (Unsupervised)

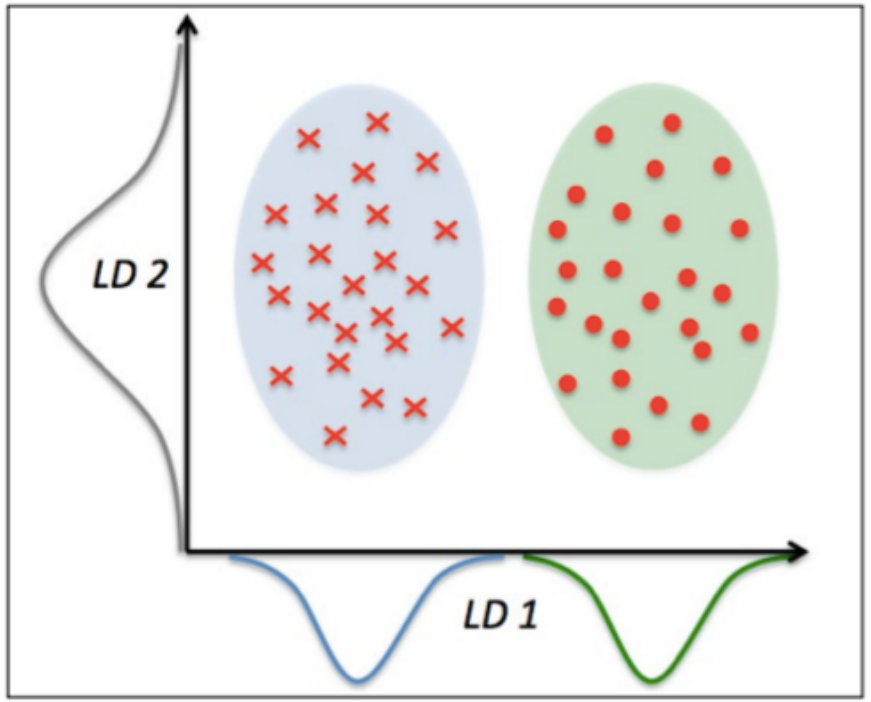
e.g., data with 64 initial values
Reconstruct image using 8 values (principal components) + mean



https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

Projection: Linear Discriminant Analysis (Supervised); established 1936

- Assumptions:
  - Data is normally distributed
  - Data is linearly separable
  - e.g., x-axis would separate the two classes well
  - e.g., y-axis would not separate the two classes well
- Algorithm
  1. Standardize d-dimensional dataset
  2. For each class, compute d-dimensional mean vector
  3. Construct between-class scatter matrix and the within-class scatter matrix
  4. Compute eigenvectors and corresponding eigenvalues
  5. Sort eigenvalues by decreasing the order to rank the corresponding eigenvectors
  6. Choose k eigenvectors that correspond to the k largest eigenvalues
  7. Project samples onto the new feature space



Python Machine Learning; Raschkka & Mirjalili

Manifold Approaches

- Manifold intuition:
  - Imagine a sheet of paper which is a 2-d object/manifold living/embedded in a 3-d world/space
  - Rotating, bending, or crumpling the paper does not change that it is 2d but it does mean that the embedding in 3d space is no longer linear
- Manifold goal:
  - Algorithms seek to learn about the fundamental 2d nature of the paper even as it is contorted to fill the 3d space
- Algorithms:
  - Model the *manifold* on which the training instances lie; i.e., make an assumption or manifold hypothesis that most real-world high-dimensional datasets lie close to a much lower-dimensional manifold
  - e.g., Locally Linear Embedding

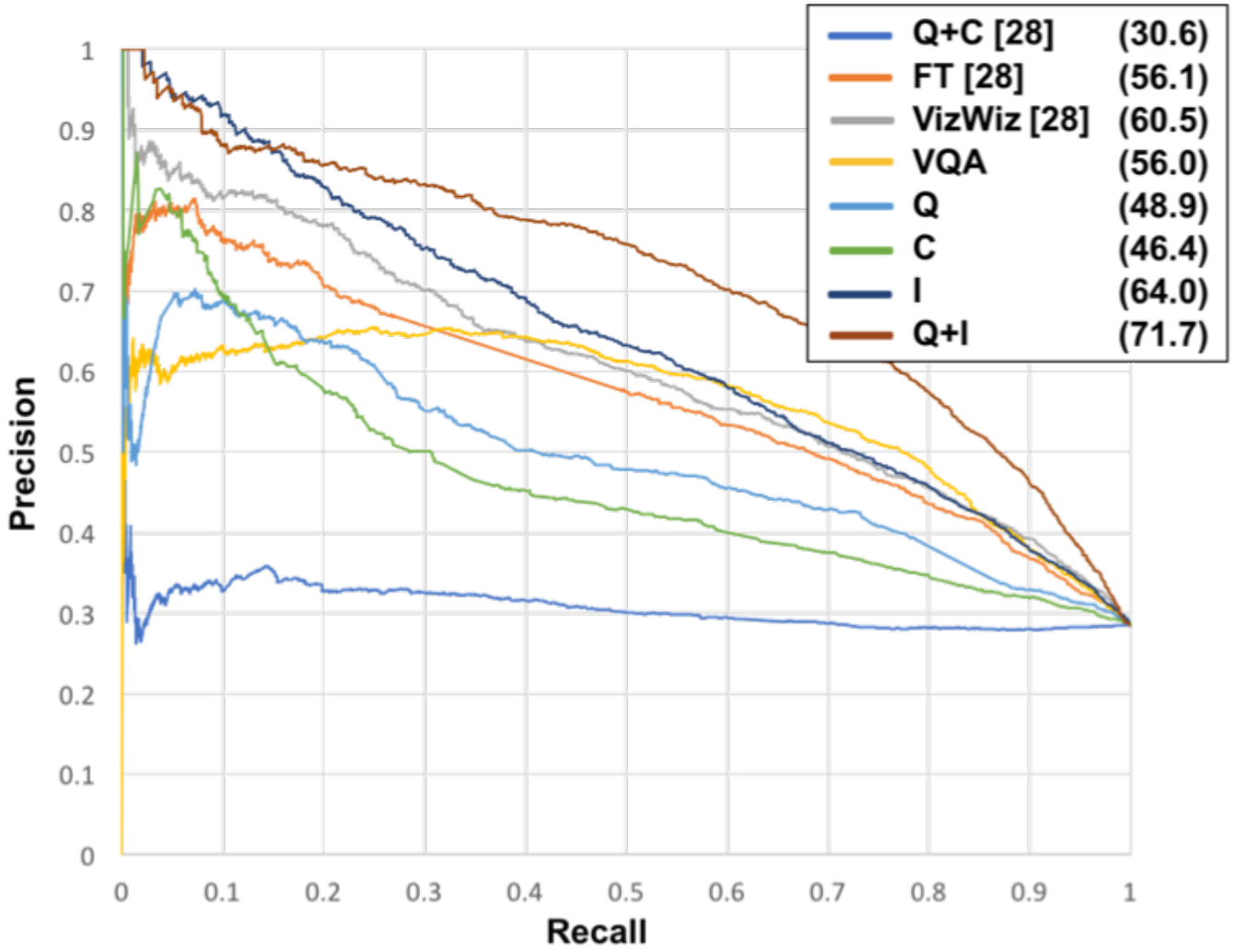
Why Else Use Dimensionality Reduction?

- Visualization
- Data compression
- Noise removal

Today’s Topics

- Real-world community challenges
- Feature Representation
- Dimensionality Reduction
- Classification Evaluation
- Lab

Precision-Recall (PR) Curve

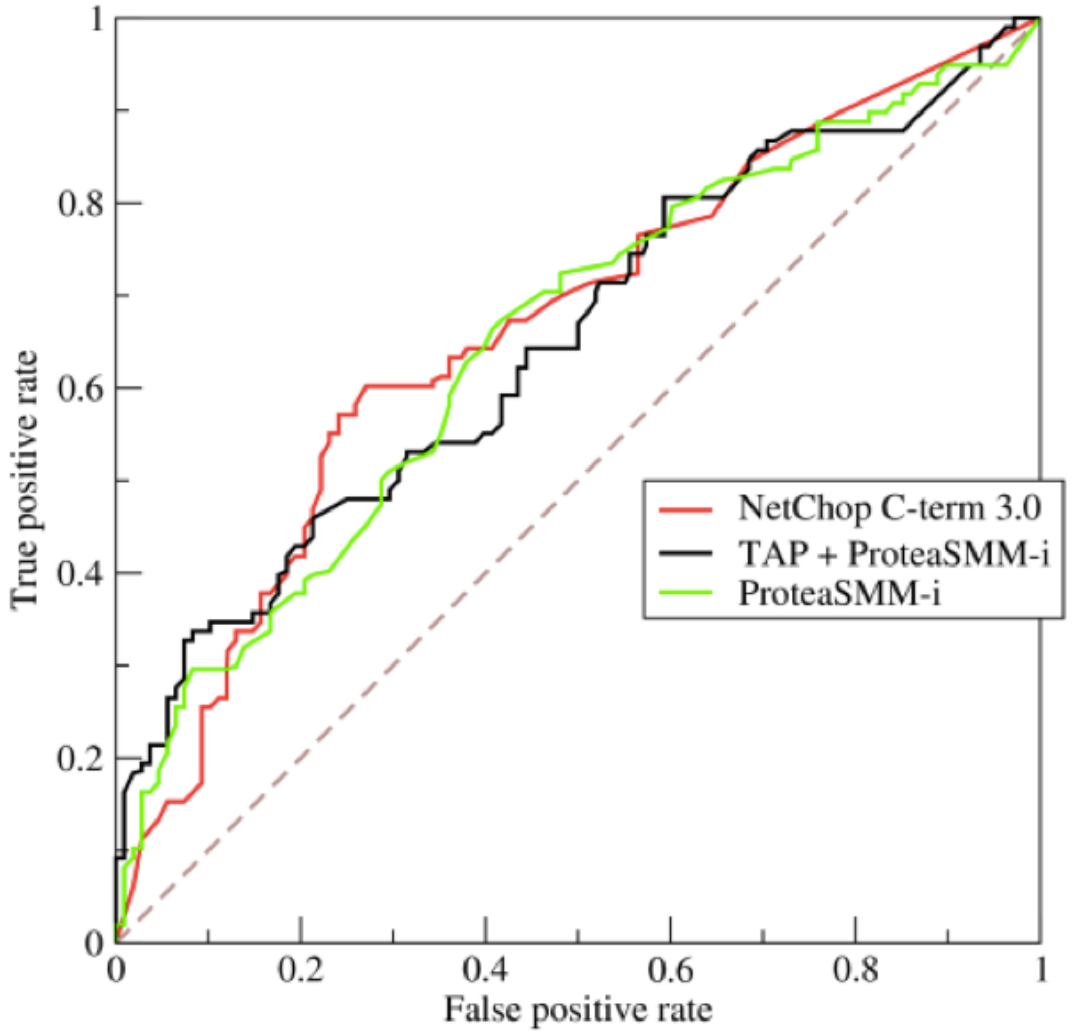


- How would you create a PR curve for 10-nearest neighbor algorithm?
  - Vary classification threshold from 1 to 9 and compute precision and recall for each value
- Which classifier is the best?
- What will be the Average Precision score for a perfect classifier
- Implementation detail: models must return probability or rankings to generate a PR curve

Precision-Recall (PR) Curve

- Precision/Recall trade-off: increasing precision reduces recall, and vice versa
- What are applications where you would prioritize higher precision?
  - Detect offensive content
  - Medical diagnoses
- What are applications where you would prioritize recall?
  - Detect shoplifters
- How do you decide which threshold to use?

Receiver Operating Characteristic (ROC) Curve



TP	FP
FN	TN

$$TPR = TP/P = TP/(TP + FN)$$

$$FPR = FP/N = FP/(FP + TN)$$

Today’s Topics

- Real-world community challenges
- Feature Representation
- Dimensionality Reduction
- Classification Evaluation
- Lab

Introduction to Microsoft Azure