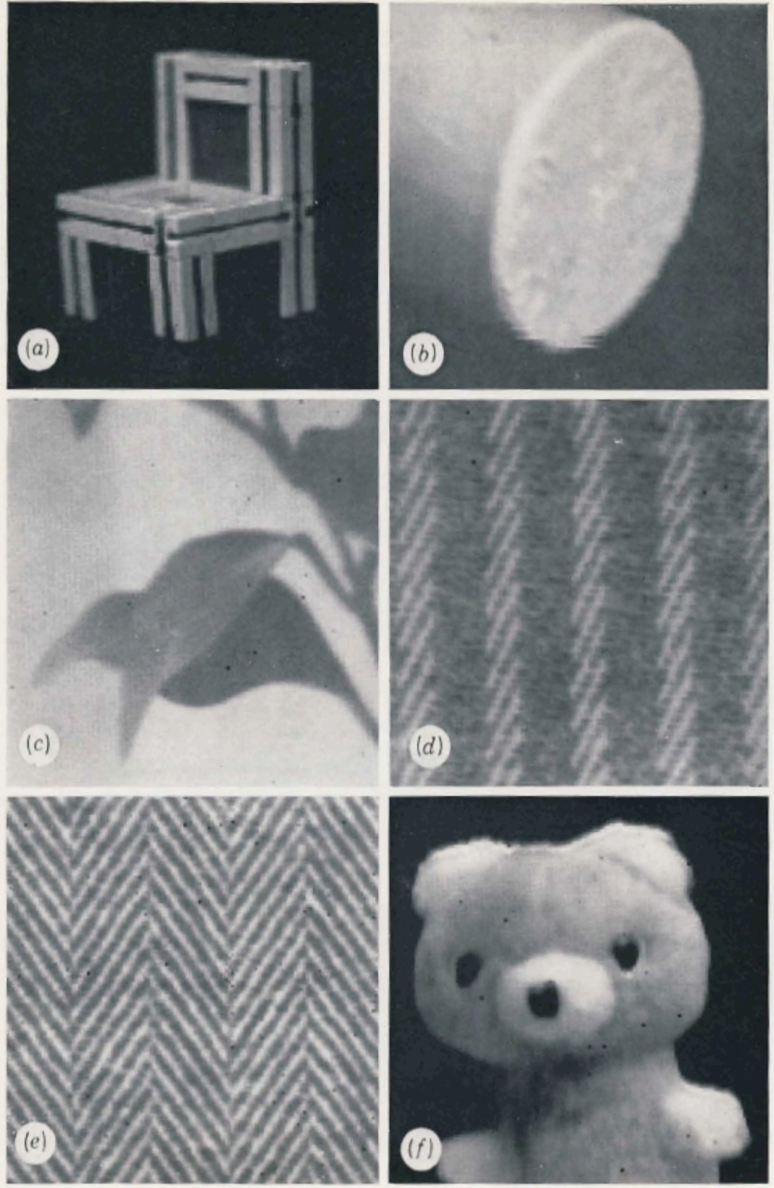
Datasets for object recognition and scene understanding

Slides adapted with gratitude from http://www.cs.washington.edu/education/courses/cse590v/11au/ (Neeraj Kumar and Brian Russell)

1972

Slide credit: A. Torralb

Slide credit: A. Torralba

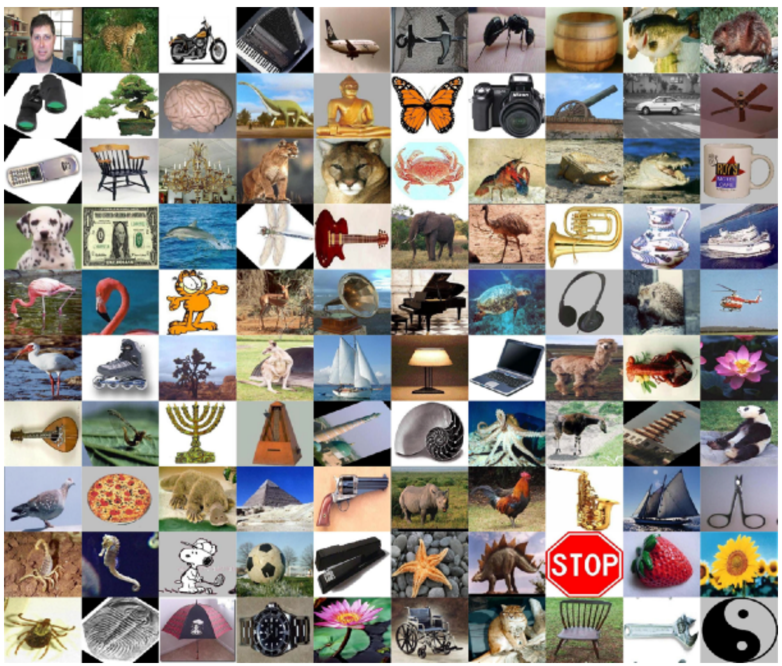


URE 4. This figure provides a high quality reproduction of the six images discussed in the text. (*a*) and (*b*) were taken with a considerably modified Information International Incorporated Vidissector, and the rest were taken with a Telemation TMC-2100 vidicon camera attached to a Spatial Data Systems digitizer (Camera Eye 108). The full dynamic range from black to white is represented by 256 grey-levels. The images reproduced here were created by an Optronics P150ohPhotowriter from intensity arrays that measured 128 elements square. This size of intensity array corresponds to viewing a 1 in square at 5 ft with the human retina. The image of the period at the end of this sentence probably covers more than 40 retinal receptors. The reader should view the images from a distance of about 5 ft when assessing the performance of the programs.

Marr, 1976

Caltech 101 and 256

101 object classes



Fei-Fei, Fergus, Perona, 2004

9,146 images

256 object classes

Griffin, Holub, Perona, 2007

30,607 images

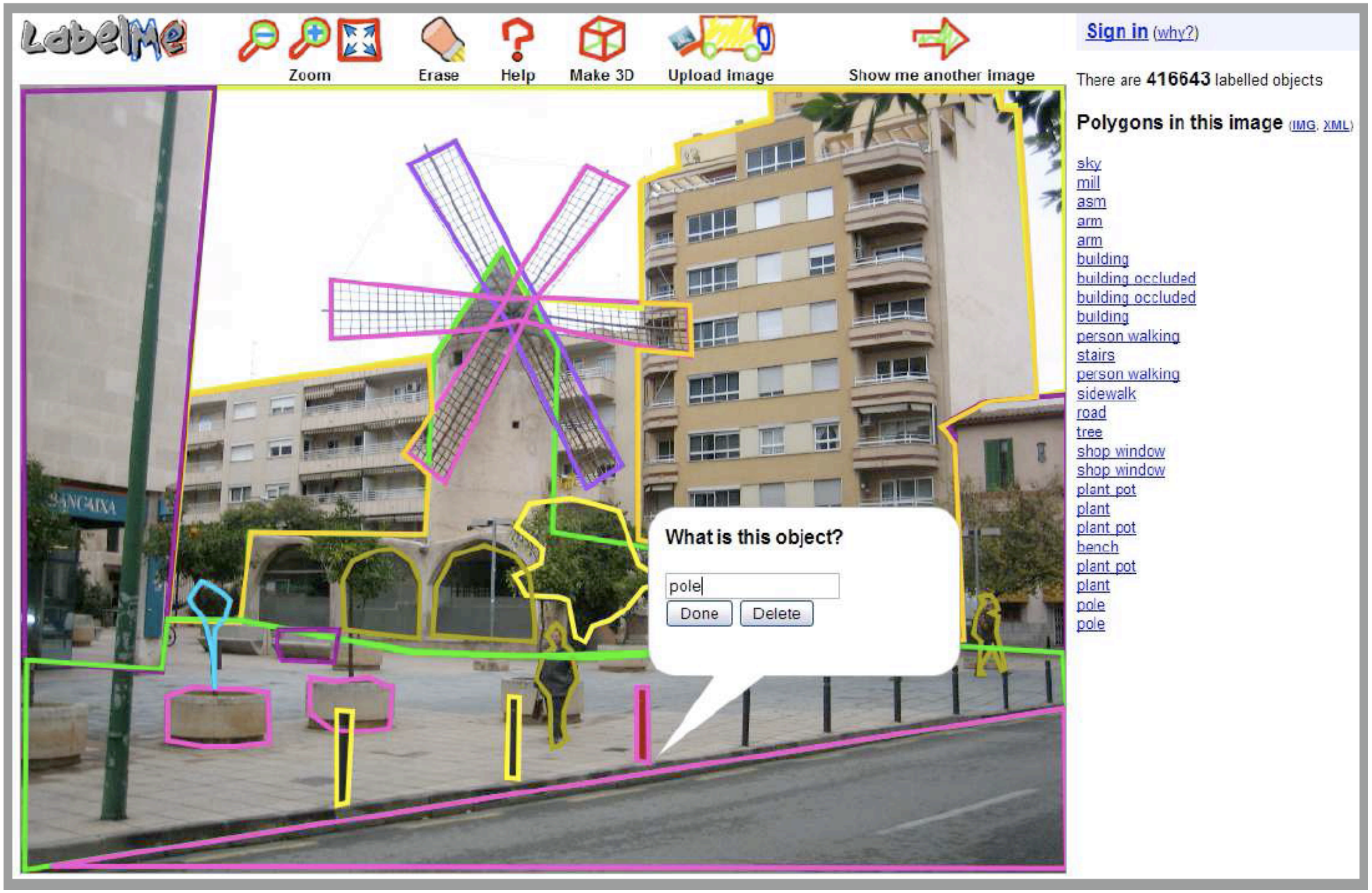
Slide credit: A. Torralba

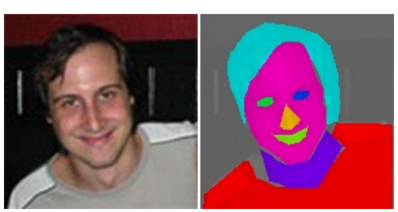
MSRC

591 images, 23 object classes
Pixel-wise segmentation

J. Winn, A. Criminisi, and T. Minka, 2005

LabelMe





Tool went online July 1st, 2005
825,597 object annotations collected
199,250 images available for labeling

labelme.csail.mit.edu

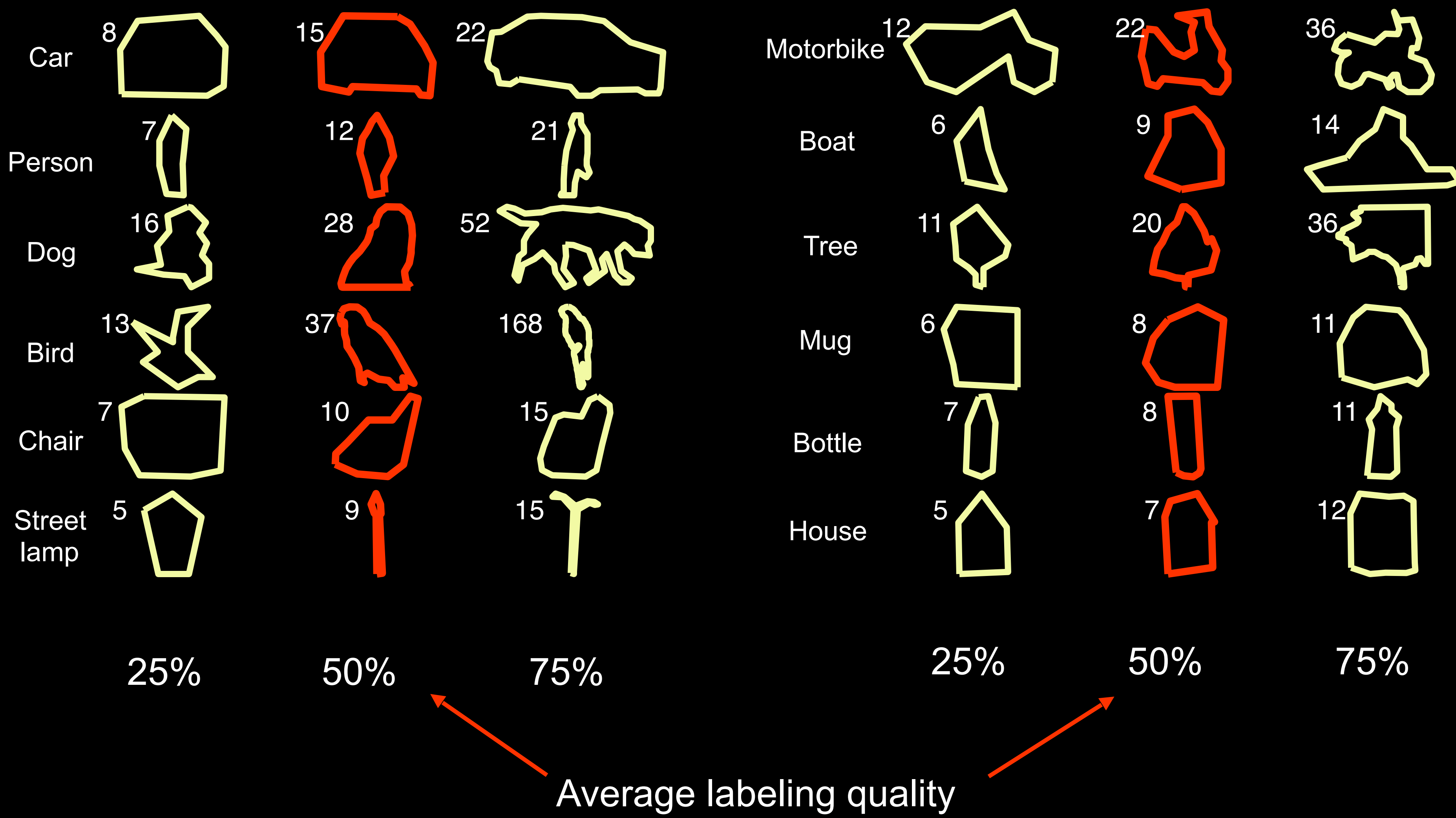
B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, IJCV 2008

Labeling tool Help Downloads New! About Matlab Toolbox Upload Images Stats

Search

Your query (**street**) matches **13238** images

Quality of the labeling

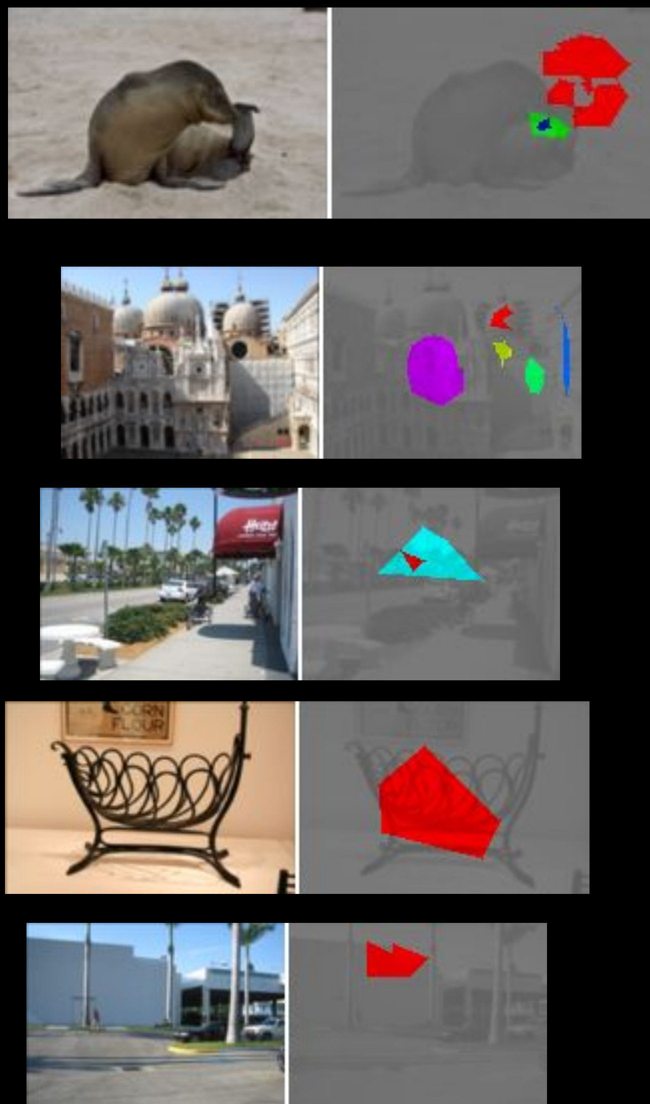


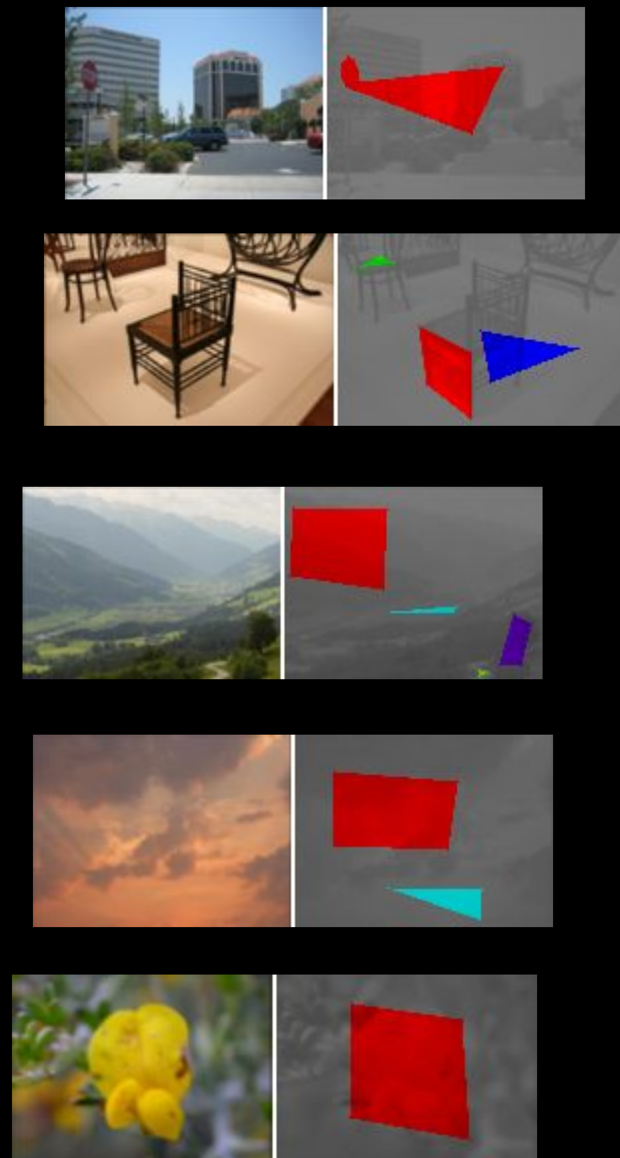
Extreme labeling

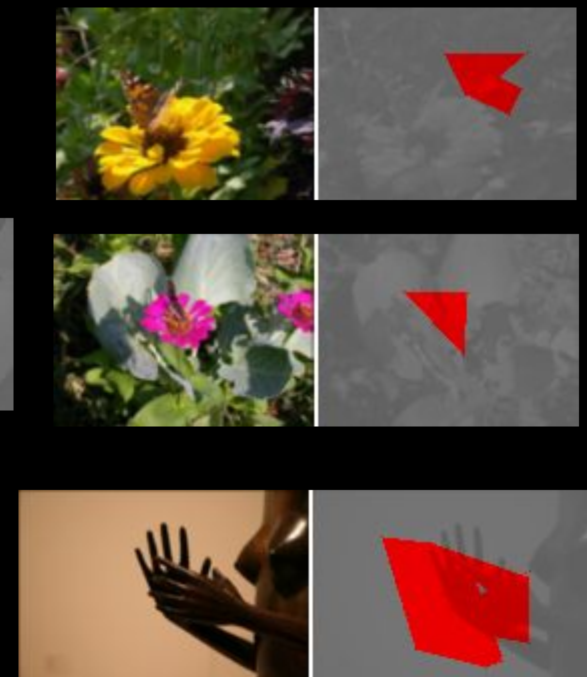
The other extreme of extreme labeling

… things do not always look good…

Testing







Most common labels:

test

adksdsa

woiieiie

…

Sophisticated testing

Most common labels:

Star

Square

Nothing

…

Visual Object Classes Challenge 2011 (VOC2011)

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

[click on an image to see the annotation]

2011 version - 20 object classes:
Person: person
Animal: bird, cat, cow, dog, horse, sheep
Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

The train/val data has 11,530 images containing
27,450 ROI annotated objects and 5,034 segmentations

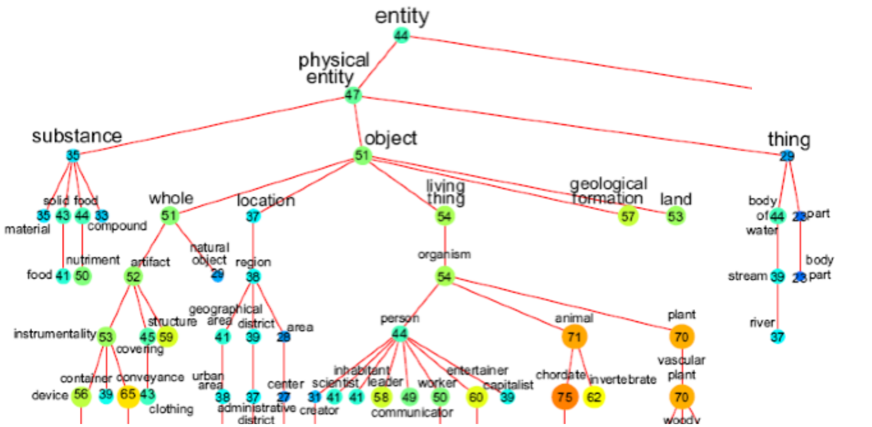
- Three main competitions: classification, detection, and segmentation
- Three "taster" competitions: person layout, action classification, and ImageNet large scale recognition

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman

80.000.000 tiny images

Slide credit: A. Torralb

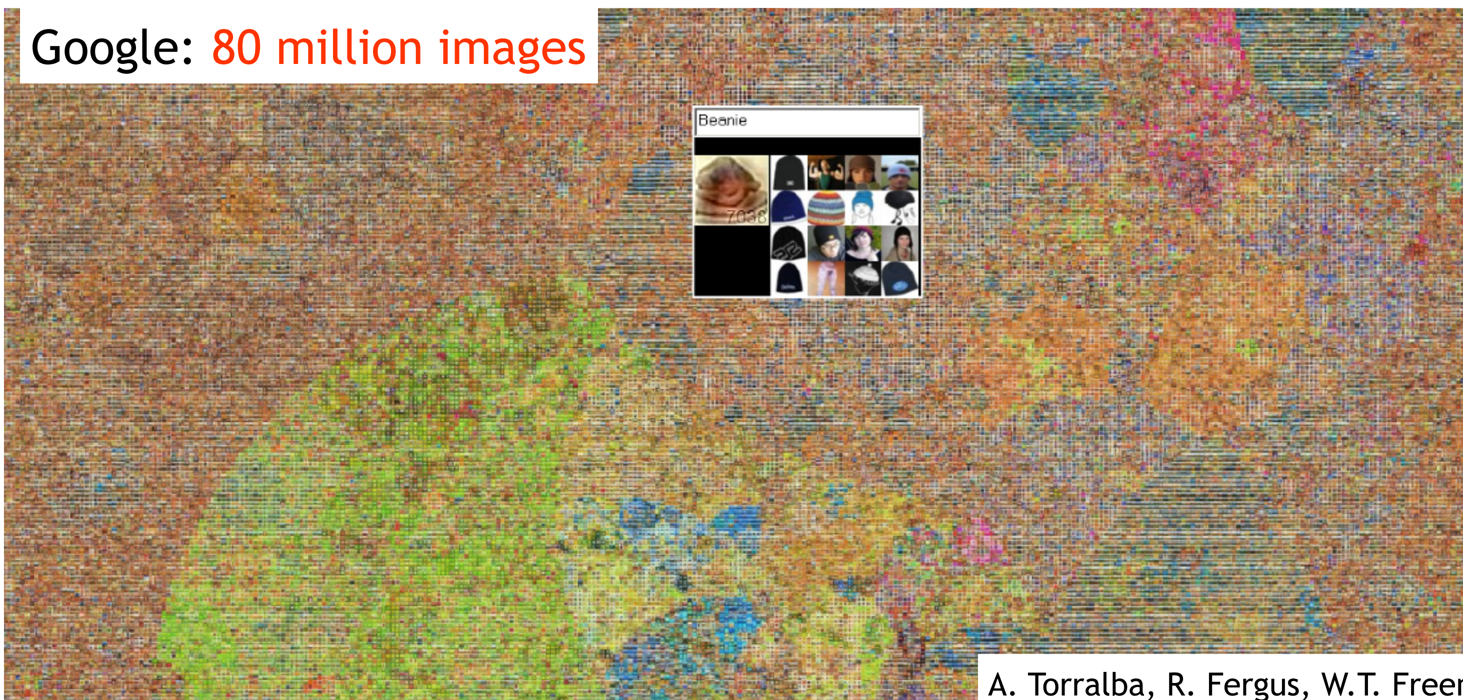
75.000 non-abstract nouns from WordNet



7 Online image search engines

And after 1 year downloading images

Google: 80 million images

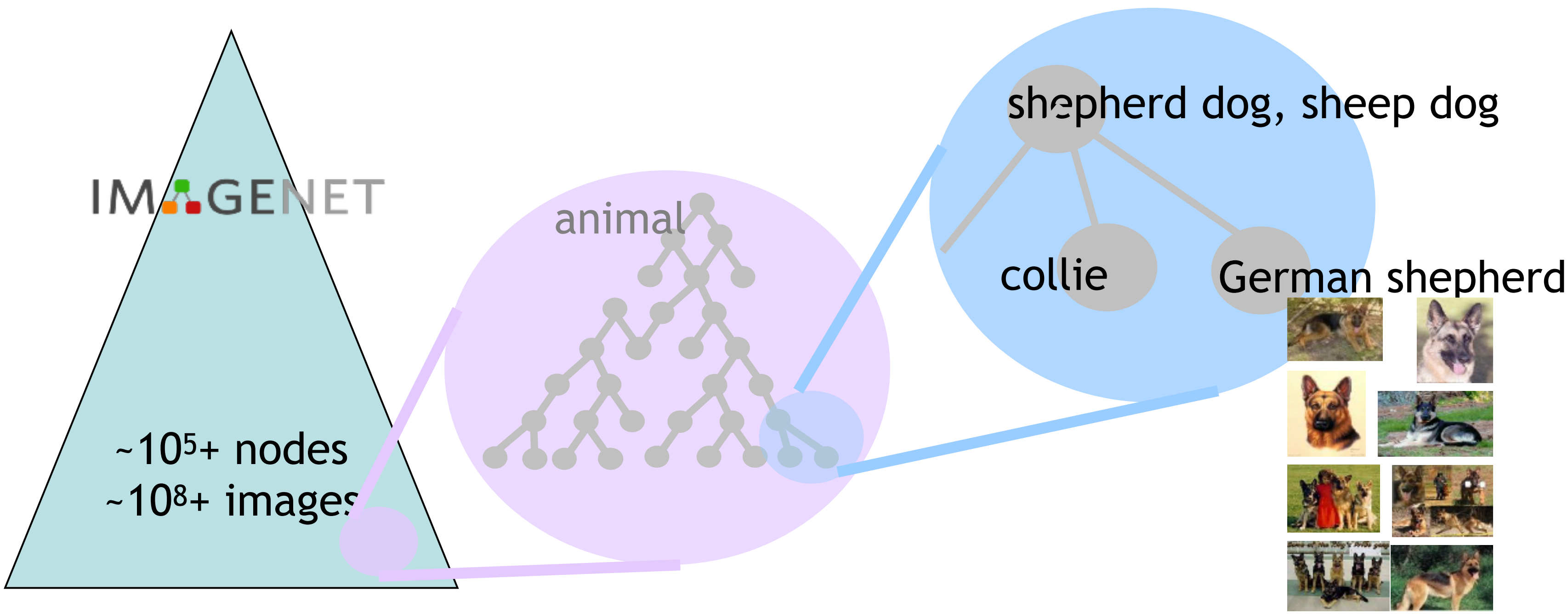


A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008

Slide credit: A. Torralb

IMAGENET

- An ontology of images based on WordNet
  - 22,000+ categories of visual concepts
  - 15 million human-cleaned images
  - **www.image-net.org**



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

- Collected all the terms from WordNet that described scenes, places, and environments
    - Any concrete noun which could reasonably complete the phrase “I am in a place”, or “let’s go to the place”

- 899 scene categories
- 130,519 images
- 397 scene categories with at least 100 images
- 63,726 labeled objects

J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, CVP

All the following slides are from A. Torralba and A. Efros

Unbiased Look at Dataset Bias

Alyosha Efros (CMU)

Antonio Torralba (MIT)

Are datasets measuring the right thing?

- In Machine Learning:

  Dataset is The World

- In Recognition

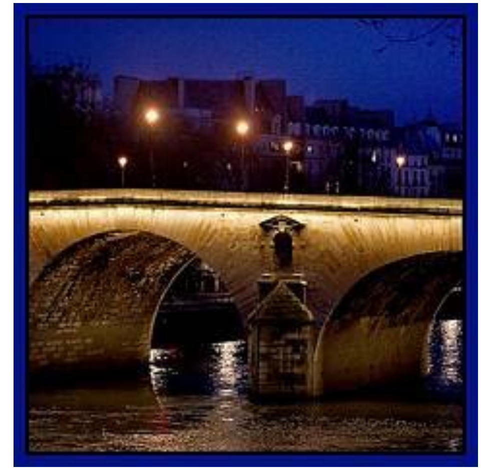
  Dataset is a *representation* of The World

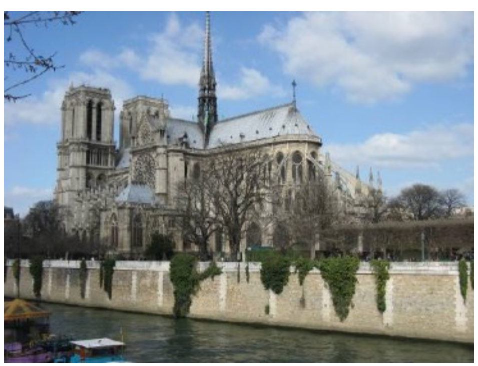
- Do datasets provide a *good* representation?

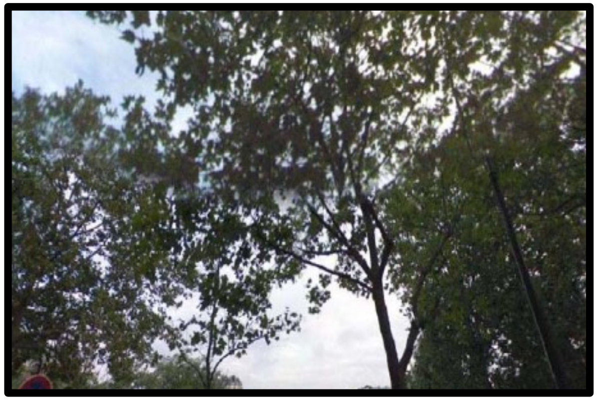
Visual Data is Inherently Biased

- Internet is a tremendous repository of visual data (Flickr, YouTube, Picassa, etc)
- But it's <u>not</u> random samples of visual world

Flickr Paris





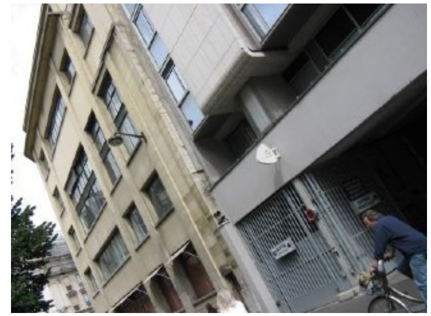


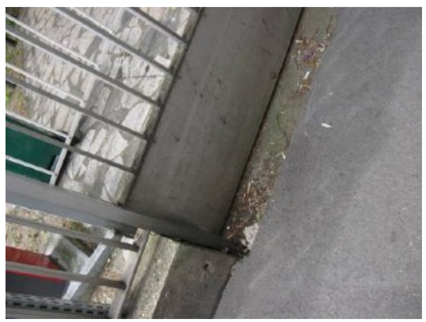
Google StreetView Paris

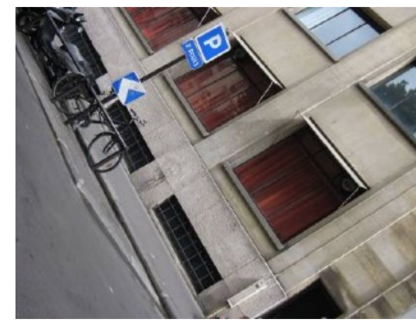
Knopp, Sivic, Pajdla, ECCV 2010

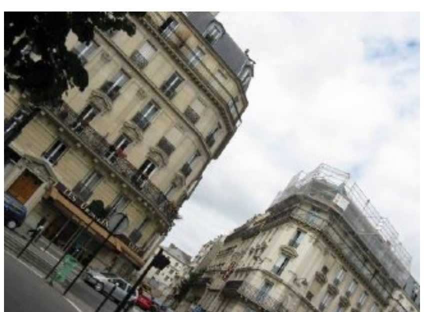
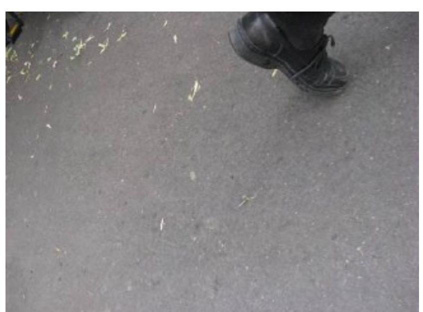
Sampled Alyosha Efros's Paris

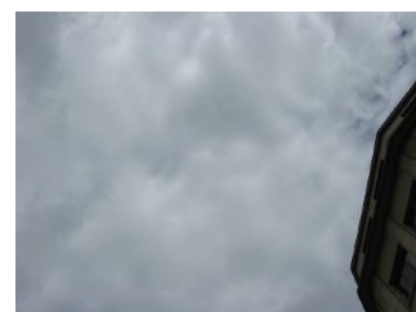






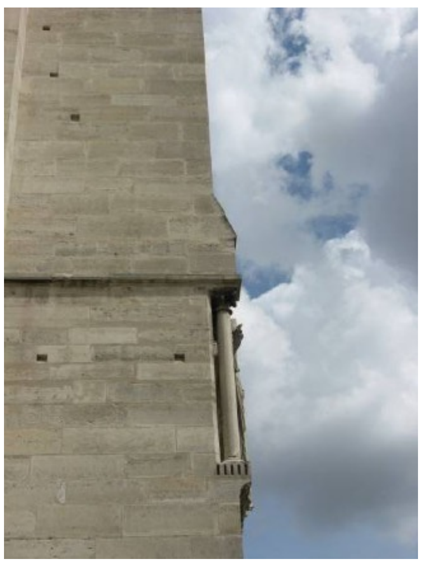


Sampling Bias

- People like to take pictures on vacation

Photographer Bias

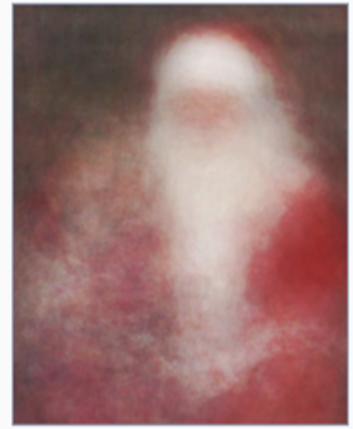
- People want their pictures to be recognizable and/or interesting



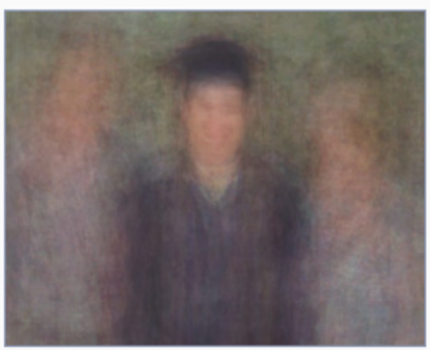
vs.

Social Bias

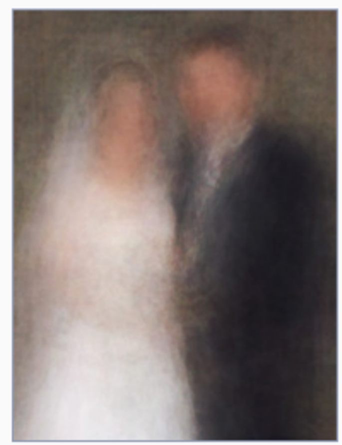
Little Leaguer



Kids with Santa



The Graduate



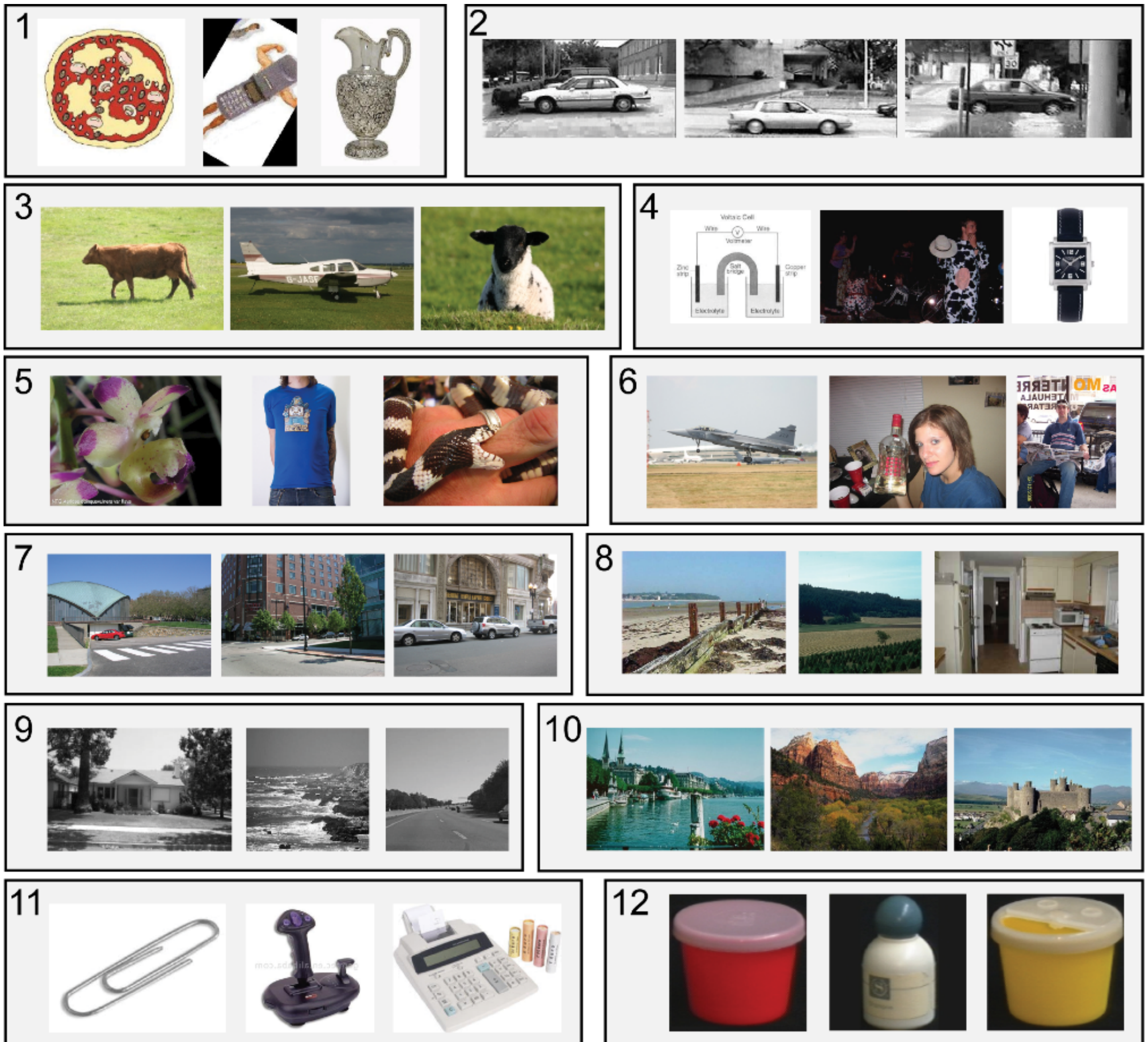
Newlyweds

"100 Special Moments" by Jason Salavon

Our Question

- How much does this bias affect standard datasets used for object recognition?

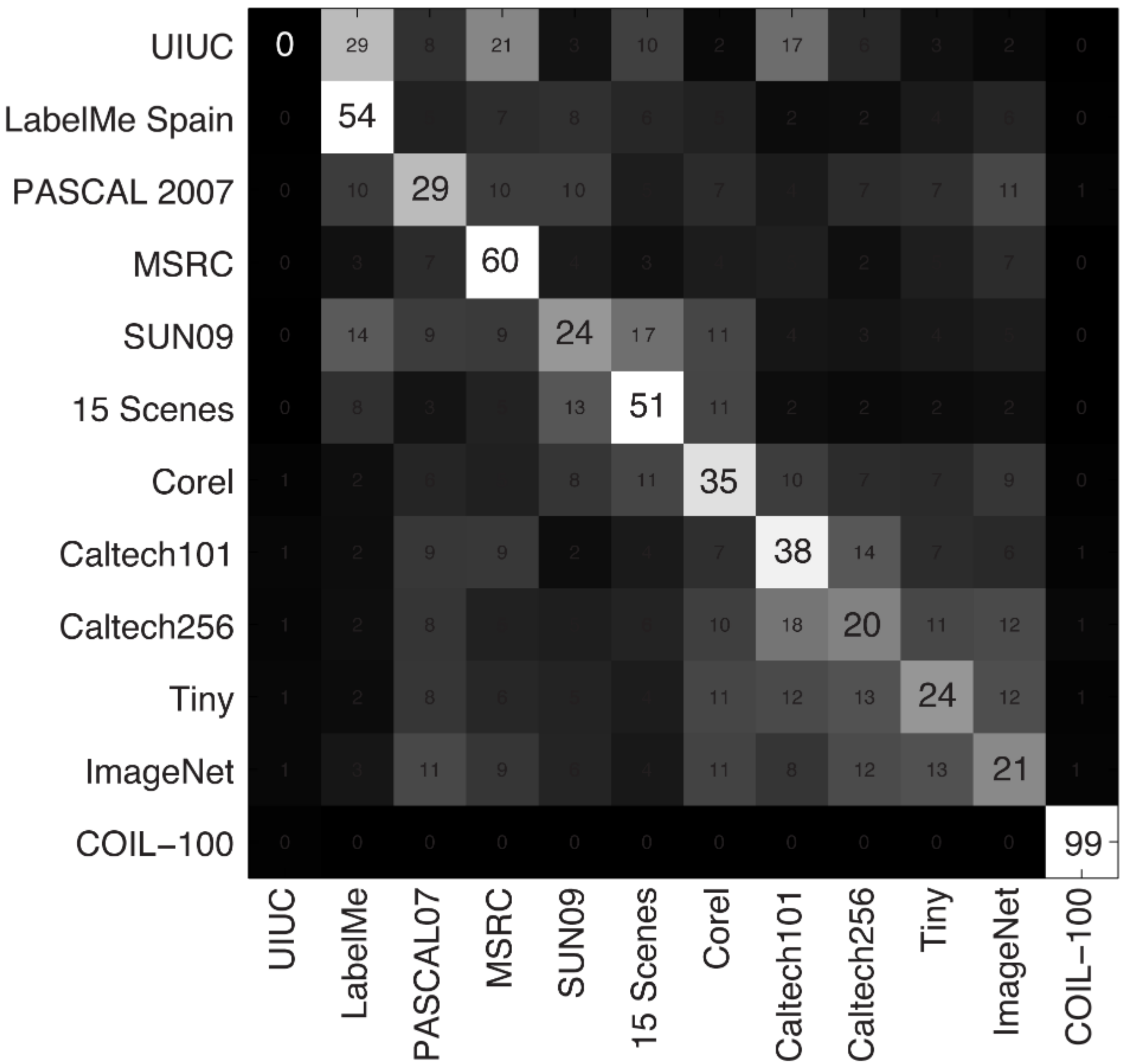
“*Name That Dataset!*” game



- __ Caltech 101
- __ Caltech 256
- __ MSRC
- __ UIUC cars
- __ Tiny Images
- __ Corel
- __ PASCAL 2007
- __ LabelMe
- __ COIL-100
- __ ImageNet
- __ 15 Scenes
- __ SUN’09

SVM plays *"Name that dataset!"*

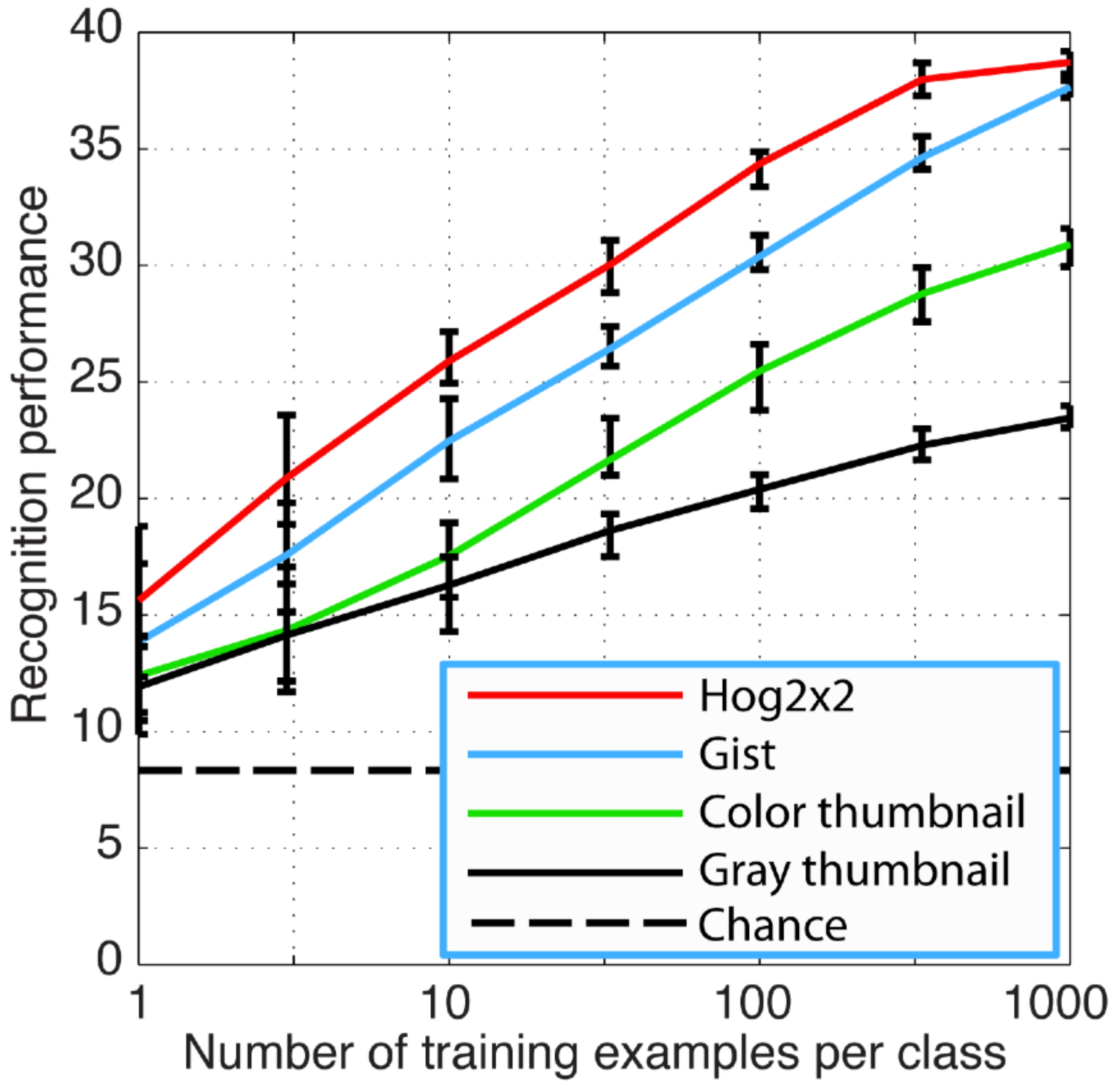
SVM plays *"Name that dataset!"*



- 12 1-vs-all classifiers
- Standard full-image features

- 39% performance (chance is 8%)

SVM plays *"Name that dataset!"*



Datasets have different goals…

- Some are object-centric (e.g. Caltech, ImageNet)
- Otherwise are scene-centric (e.g. LabelMe, SUN’09)

- What about playing *“name that dataset”* on bounding boxes?

Similar results

PASCAL cars

SUN cars

Caltech101 cars

Performance: 61%
(chance: 20%)

ImageNet cars

LabelMe cars

Where does this bias comes from?

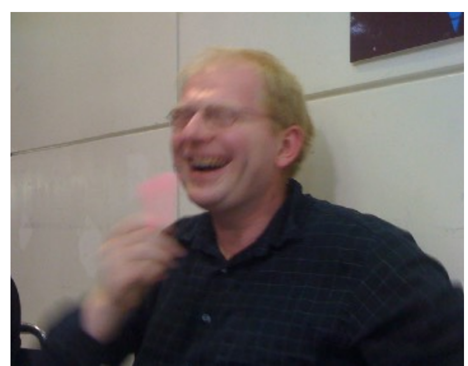
Some bias is in the world

Some bias is in the world

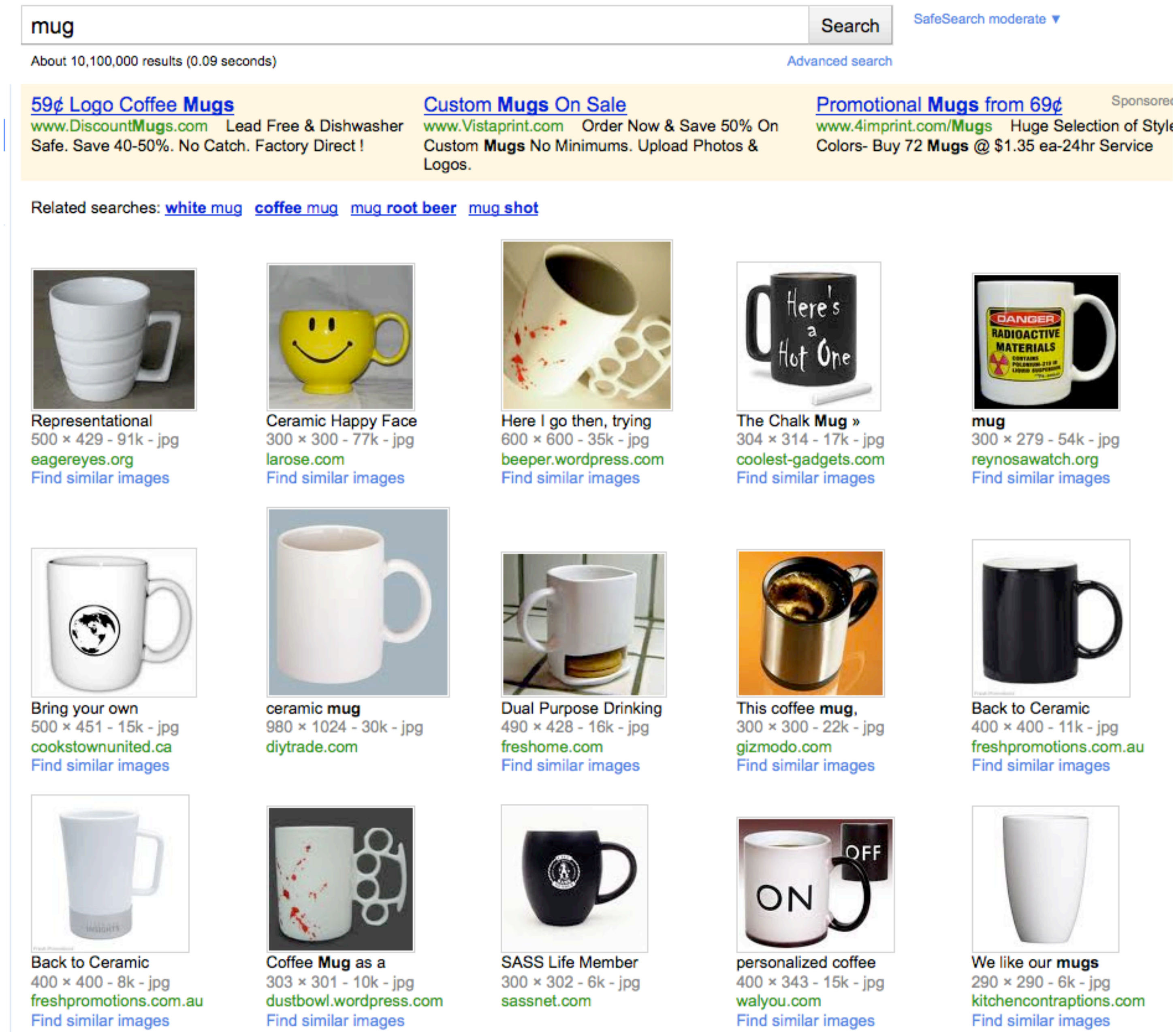


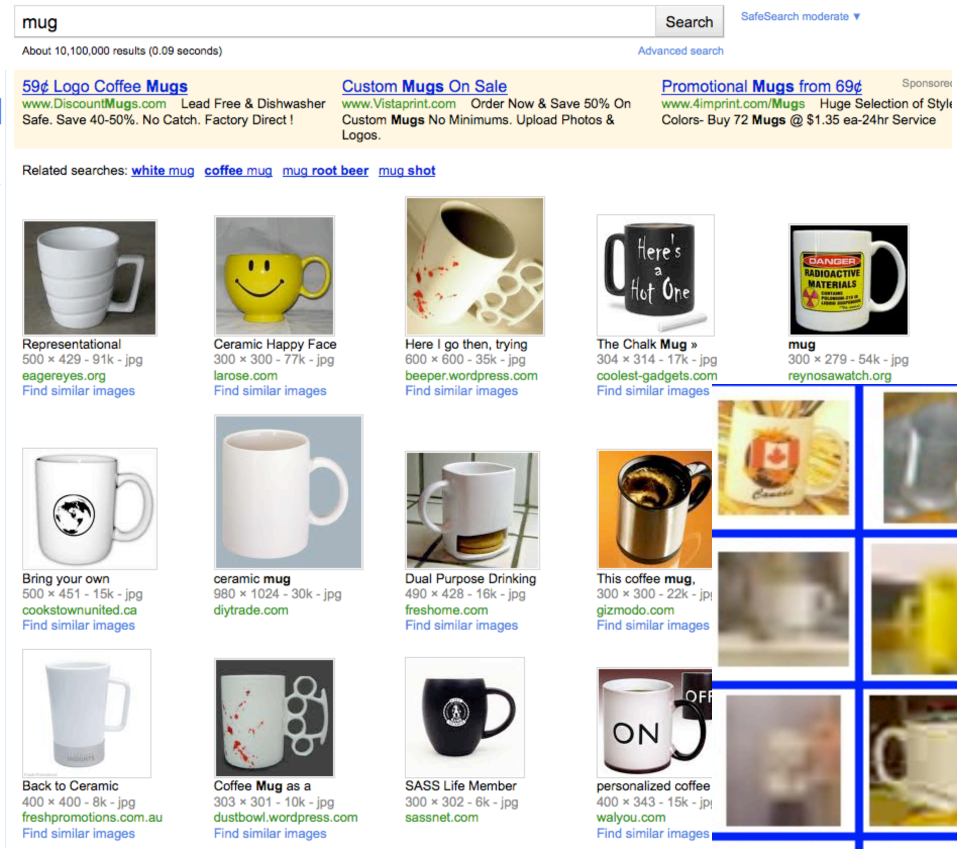







Some bias comes from the way the data is collected





Google mugs

Mugs from LabelMe

Measuring Dataset Bias

Cross-Dataset Generalization

MSRC

Classifier trained on MSRC cars

Cross-dataset Performance

Table 1. Cross-dataset generalization. Object detection and classification performance (AP) for "car" and "person" when training on on dataset (rows) and testing on another (columns), i.e. each row is: training on one dataset and testing on all the others. "Self" refers t training and testing on the same dataset (same as diagonal), and "Mean Others" refers to averaging performance on all except self.

task	Test on: / Train on:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Self	Mean others	Percent drop
"car" classification	SUN09	**28.2**	29.5	16.3	14.6	16.9	21.9	28.2	19.8	**30%**
	LabelMe	14.7	**34.0**	16.7	22.9	43.6	24.5	34.0	24.5	**28%**
	PASCAL	10.1	25.5	**35.2**	43.9	44.2	39.4	35.2	32.6	**7%**
	ImageNet	11.4	29.6	36.0	**57.4**	52.3	42.7	57.4	34.4	**40%**
	Caltech101	7.5	31.1	19.5	33.1	**96.9**	42.1	96.9	26.7	**73%**
	MSRC	9.3	27.0	24.9	32.6	40.3	**68.4**	68.4	26.8	**61%**
	Mean others	10.6	28.5	22.7	29.4	39.4	34.1	53.4	27.5	48%
"car" detection	SUN09	**69.8**	50.7	42.2	42.6	54.7	69.4	69.8	51.9	**26%**
	LabelMe	61.8	**67.6**	40.8	38.5	53.4	67.0	67.6	52.3	**23%**
	PASCAL	55.8	55.2	**62.1**	56.8	54.2	74.8	62.1	59.4	**4%**
	ImageNet	43.9	31.8	46.9	**60.7**	59.3	67.8	60.7	49.9	**18%**
	Caltech101	20.2	18.8	11.0	31.4	**100**	29.3	100	22.2	**78%**
	MSRC	28.6	17.1	32.3	21.5	67.7	**74.3**	74.3	33.4	**55%**
	Mean others	42.0	34.7	34.6	38.2	57.9	61.7	72.4	44.8	**48%**
"person" classification	SUN09	**16.1**	11.8	14.0	7.9	6.8	23.5	16.1	12.8	**20%**
	LabelMe	11.0	**26.6**	7.5	6.3	8.4	24.3	26.6	11.5	**57%**
	PASCAL	11.9	11.1	**20.7**	13.6	48.3	50.5	20.7	27.1	**-31%**
	ImageNet	8.9	11.1	11.8	**20.7**	76.7	61.0	20.7	33.9	**-63%**
	Caltech101	7.6	11.8	17.3	22.5	**99.6**	65.8	99.6	25.0	**75%**
	MSRC	9.4	15.5	15.3	15.3	93.4	**78.4**	78.4	29.8	**62%**
	Mean others	9.8	12.3	13.2	13.1	46.7	45.0	43.7	23.4	**47%**
"person" detection	SUN09	**69.6**	56.8	37.9	45.7	52.1	72.7	69.6	53.0	**24%**
	LabelMe	58.9	**66.6**	38.4	43.1	57.9	68.9	66.6	53.4	**20%**
	PASCAL	56.0	55.6	**56.3**	55.6	56.8	74.8	56.3	59.8	**-6%**
	ImageNet	48.8	39.0	40.1	**59.6**	53.2	70.7	59.6	50.4	**15%**
	Caltech101	24.6	18.1	12.4	26.6	**100**	31.6	100	22.7	**77%**
	MSRC	33.8	18.2	30.9	20.8	69.5	**74.7**	74.7	34.6	**54%**
	Mean others	44.4	37.5	31.9	38.4	57.9	63.7	71.1	45.6	**36%**

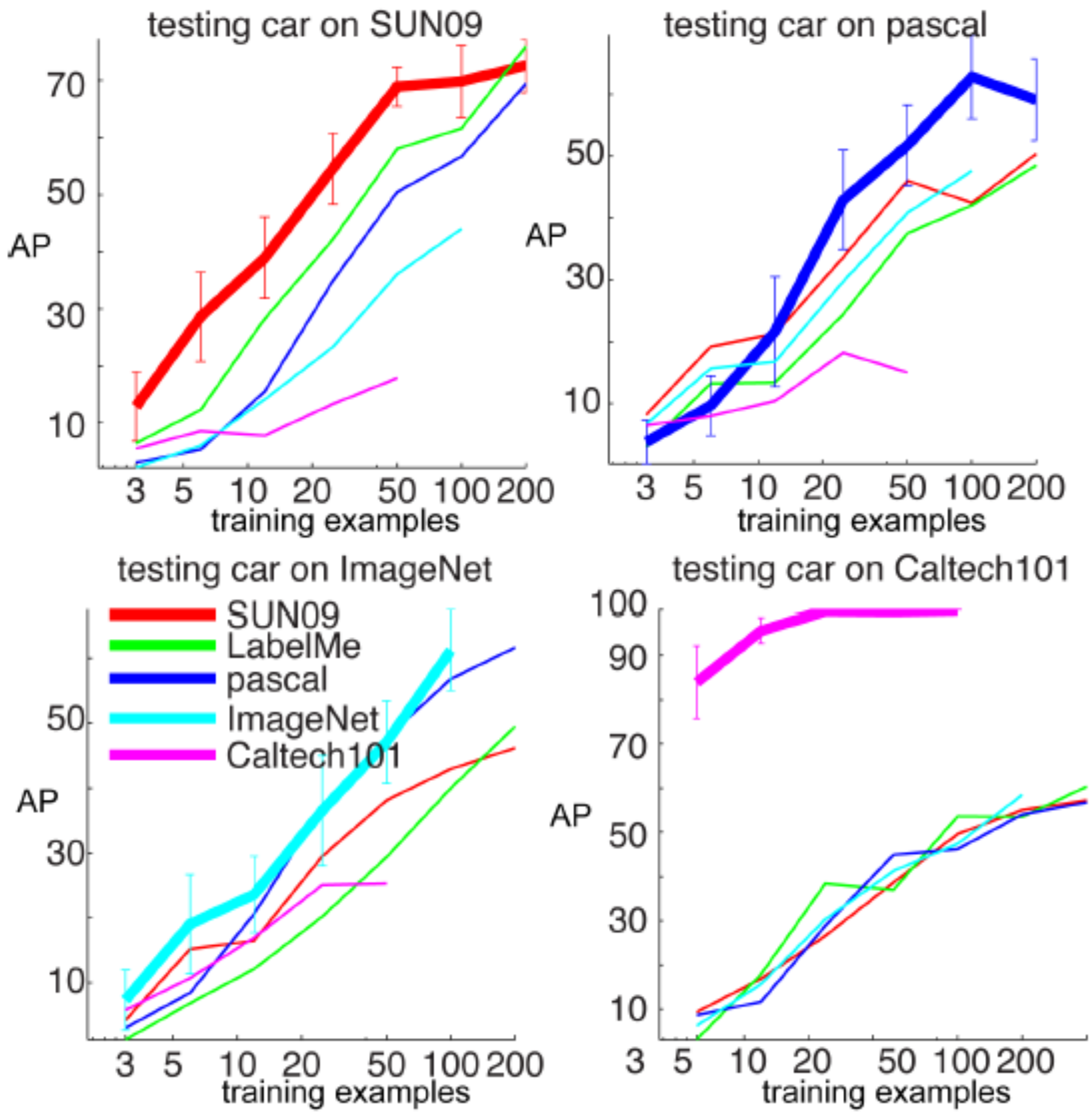


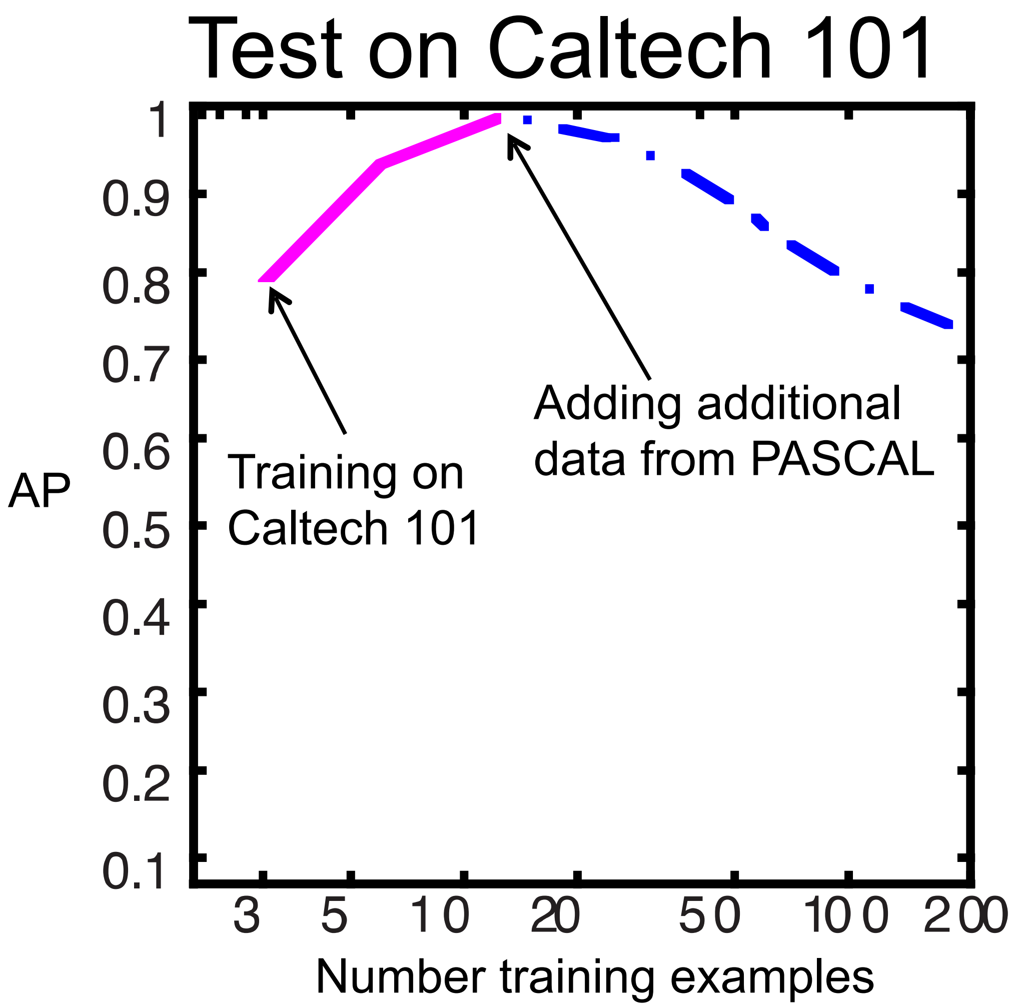
Figure 6. Cross-dataset generalization for "car" detection as function of training data

Dataset Value

Table 3. "Market Value" for a "car" sample across datasets

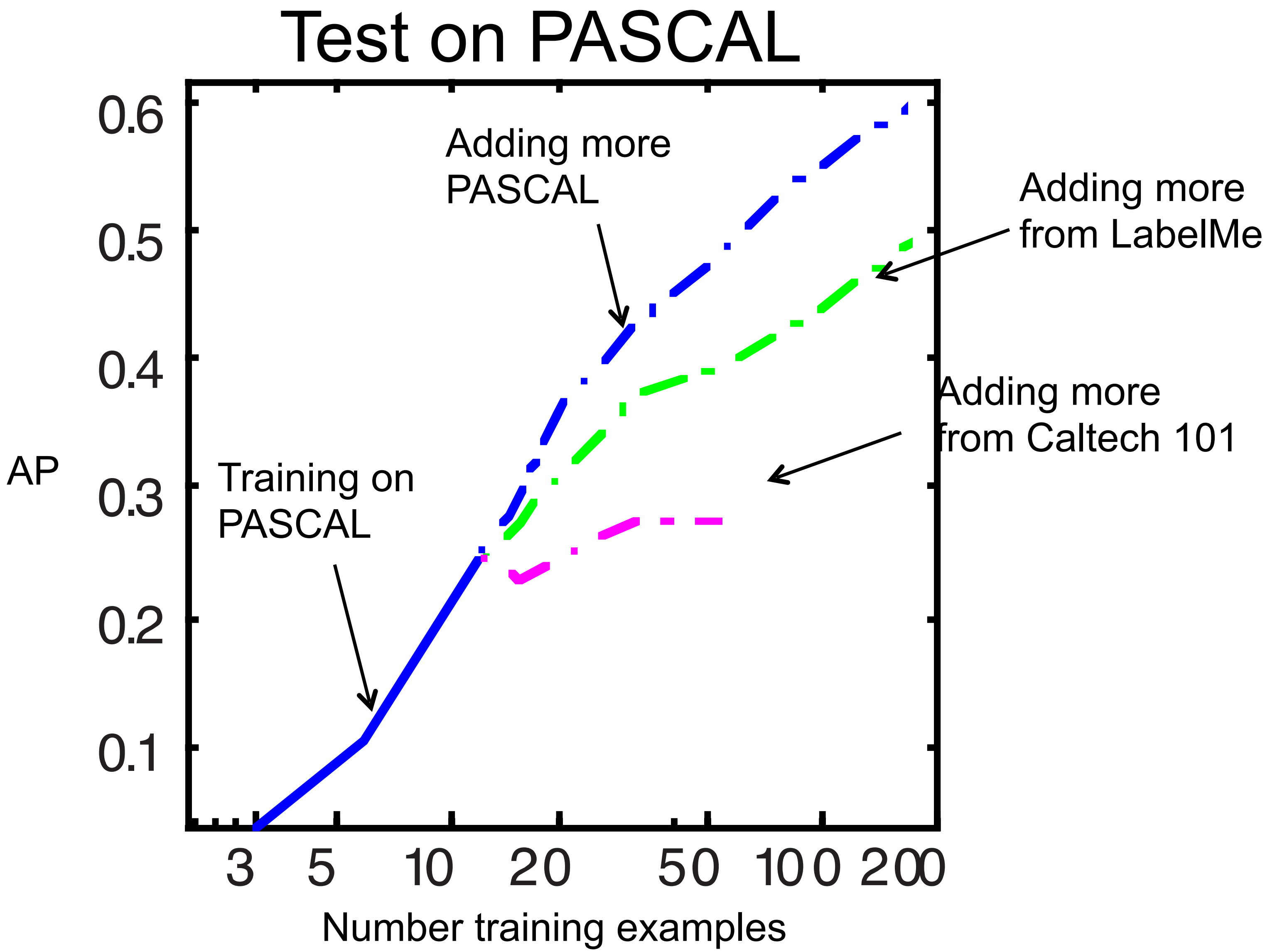
	SUN09 market	LabelMe market	PASCAL market	ImageNet market	Caltech101 market
1 SUN09 is worth	1 SUN09	0.91 LabelMe	0.72 pascal	0.41 ImageNet	0 Caltech
1 LabelMe is worth	0.41 SUN09	1 LabelMe	0.26 pascal	0.31 ImageNet	0 Caltech
1 pascal is worth	0.29 SUN09	0.50 LabelMe	1 pascal	0.88 ImageNet	0 Caltech
1 ImageNet is worth	0.17 SUN09	0.24 LabelMe	0.40 pascal	1 ImageNet	0 Caltech
1 Caltech101 is worth	0.18 SUN09	0.23 LabelMe	0 pascal	0.28 ImageNet	1 Caltech
Basket of Currencies	0.41 SUN09	0.58 LabelMe	0.48 pascal	0.58 ImageNet	0.20 Caltech

Mixing datasets



Task: car detection
Features: HOG

Mixing datasets



Negative Set Bias

Table 2. Measuring Negative Set Bias.

task	Positive Set: / Negative Set:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Mean
"car" detection	self	67.6	62.4	56.3	60.5	97.7	74.5	70.0
	all	53.8	51.3	47.1	65.2	97.7	70.0	64.1
	percent drop	20%	18%	16%	-8%	0%	6%	8%
"person" detection	self	67.4	68.6	53.8	60.4	100	76.7	71.1
	all	52.2	58.0	42.6	63.4	100	71.5	64.6
	percent drop	22%	15%	21%	-5%	0%	7%	9%

Not all the bias comes from the appearance of the objects we care about

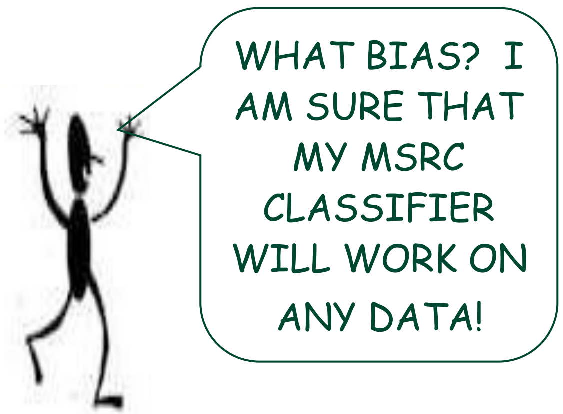
Summary (from 2011)

- Our best-performing techniques just don’t work in the real world
  - e.g., try a person detector on Hollywood film
  - but new datasets (PASCAL, ImageNet) are better than older ones (MSRC, Caltech)

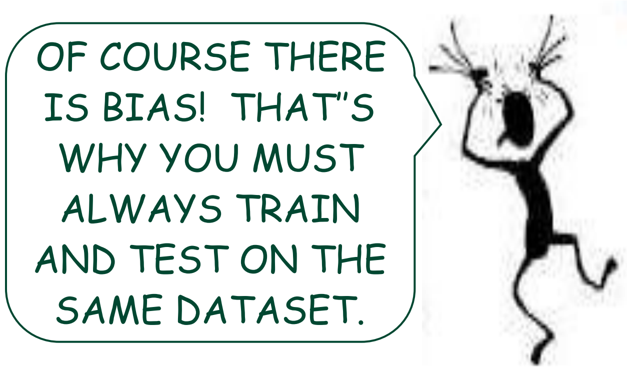
- The classifiers are inherently designed to overfit to type of data it’s trained on.

  - but larger datasets are getting better

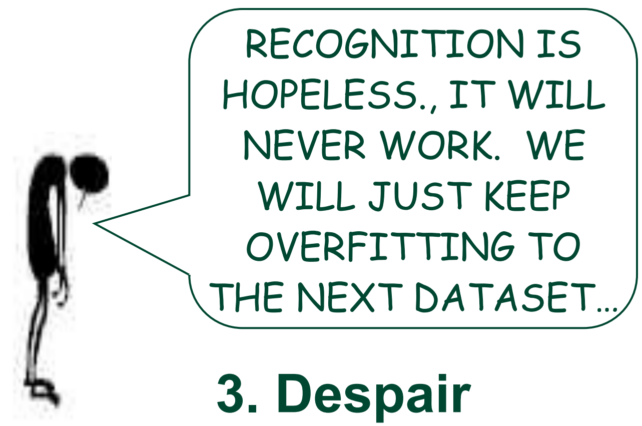
Four Stages of Dataset Grief



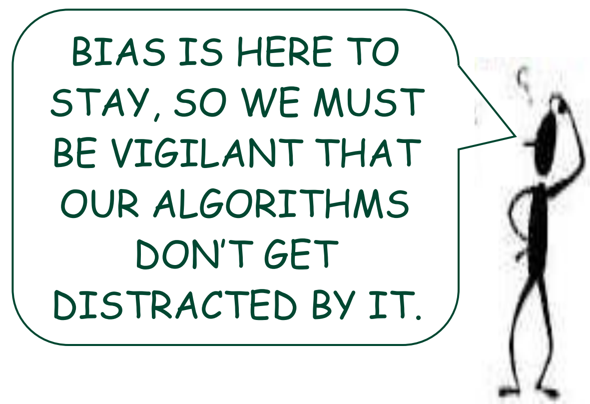
1. Denial



2. Machine Learning



3. Despair



4. Acceptance

Lessons that still apply in 2018

- Datasets are bigger but still very biased
- Specific insights about particular datasets less relevant, but overall message still critical
  - Also, exemplary analysis paper!

- Some work since then
  - Undoing the damage of dataset bias (Khosla et al. https://people.csail.mit.edu/khosla/papers/eccv2012_khosla.pdf)
  - A deeper look at dataset bias (Tommasi et al. https://arxiv.org/pdf/1505.01257.pdf)
  - What makes ImageNet good for transfer learning (Huh et al. https://arxiv.org/pdf/1608.08614.pdf)
  - Work on domain adaptation/transfer learning
  - Work on fairness in machine learning