CMPSCI 670: Computer Vision

Introduction to machine learning

University of Massachusetts, Amherst

October 29, 2014

Instructor: Subhransu Maji

Today

- Conclude “introduction to recognition”
- Introduction to machine learning
    - learning to recognize
    - machine learning framework
    - properties of learning algorithms
- Common datasets in computer vision

History of ideas in recognition

1960s – early 1990s: the geometric era

1990s: appearance-based models

1990s – present: sliding window approaches

Late 1990s: local features

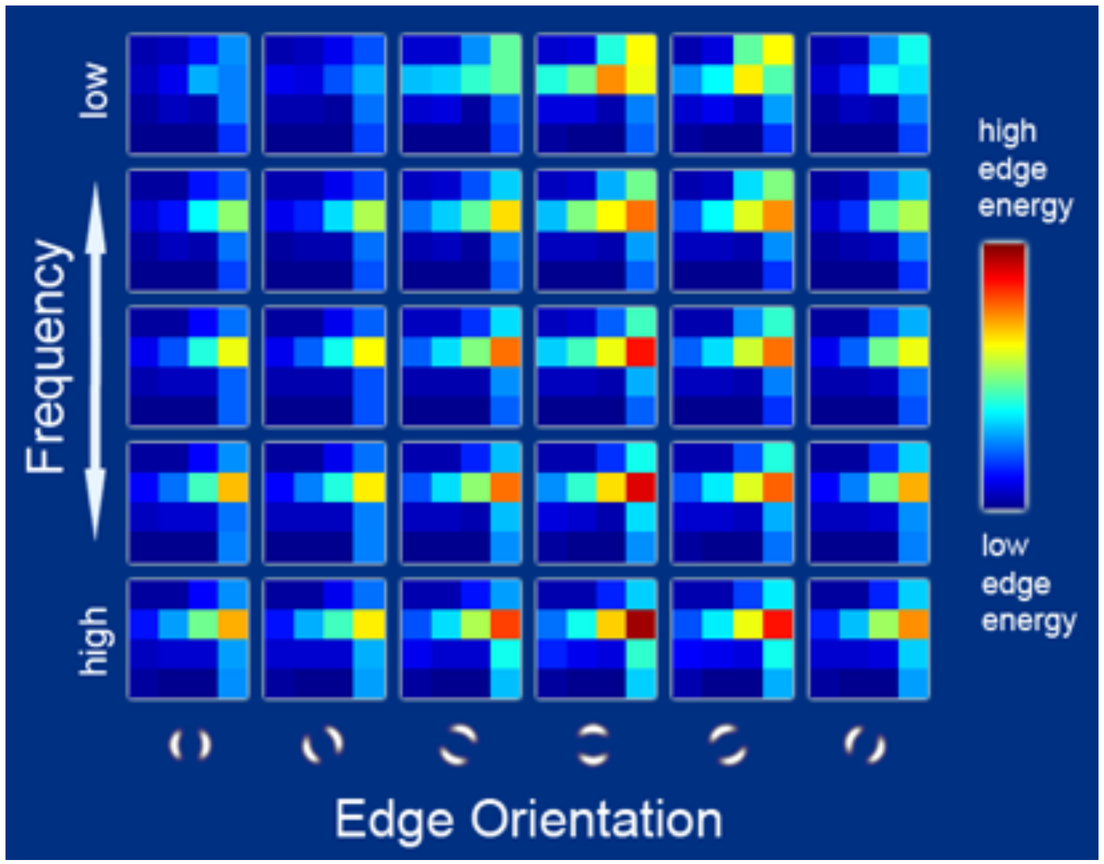
Early 2000s: parts-and-shape models

Mid-2000s: bags of features

Present trends: "big data", context, attributes, combining geometry and recognition, advanced scene understanding tasks, <u>deep learning</u>

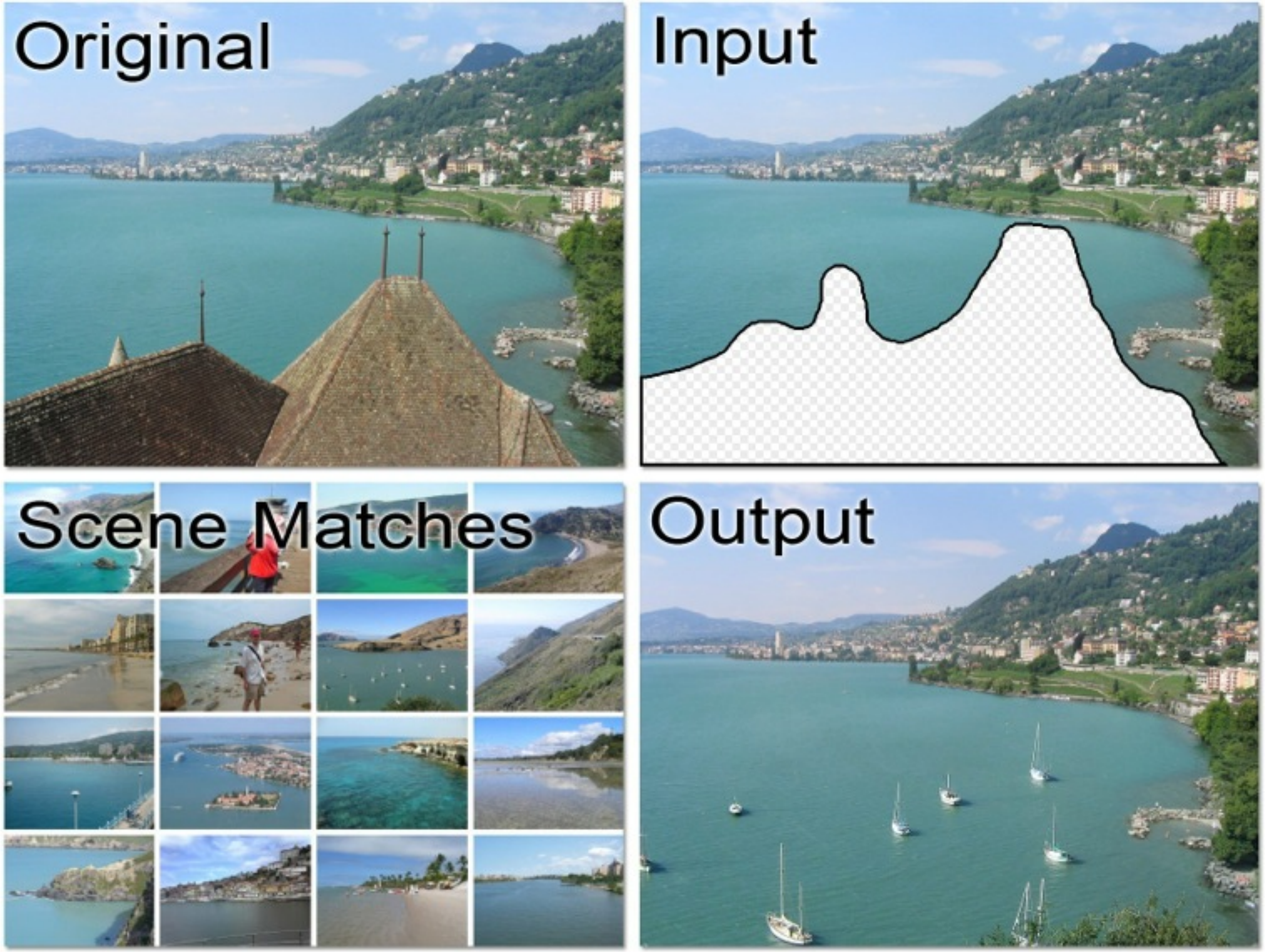
Global appearance models revisited

The “gist” of a scene: Oliva & Torralba (2001)



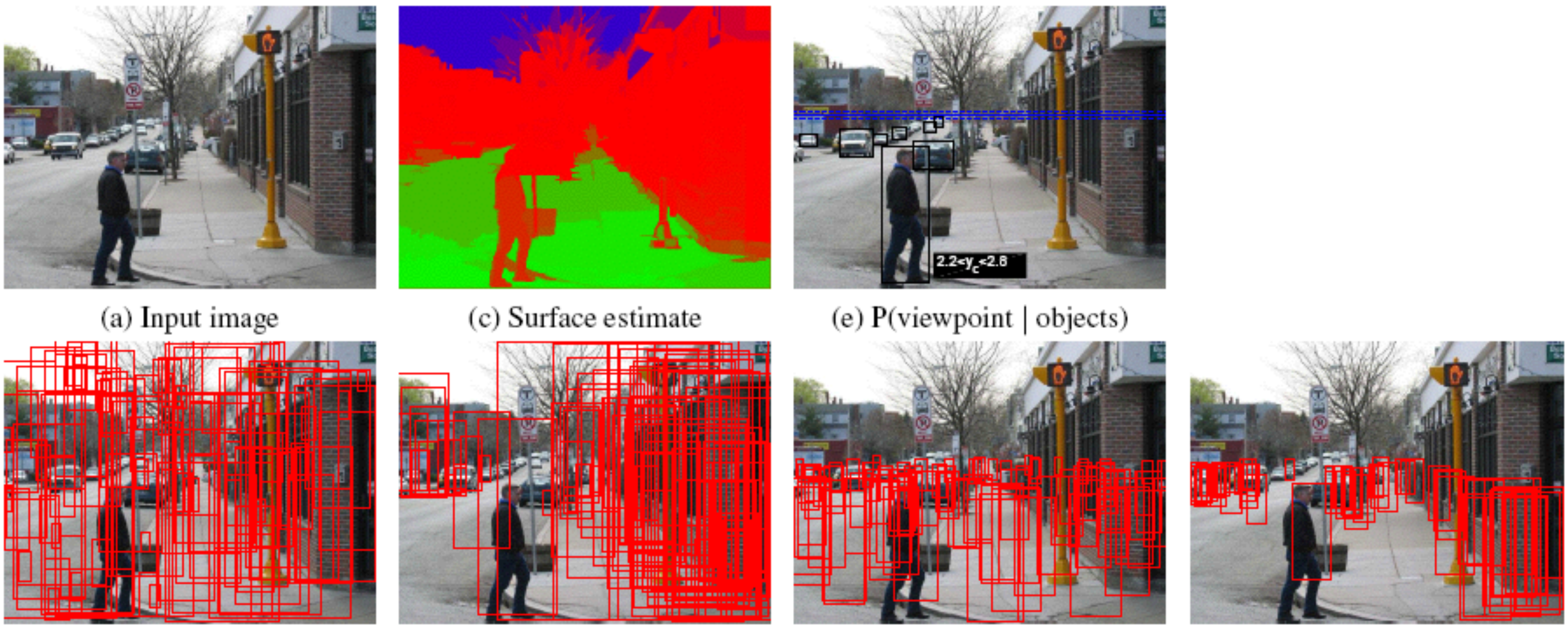
http://people.csail.mit.edu/torralba/code/spatialenvelope/

New applications in graphics



J. Hays and A. Efros, **Scene Completion using Millions of Photographs**, SIGGRAPH 2007

Geometric context



(a) Input image (c) Surface estimate (e) P(viewpoint | objects)

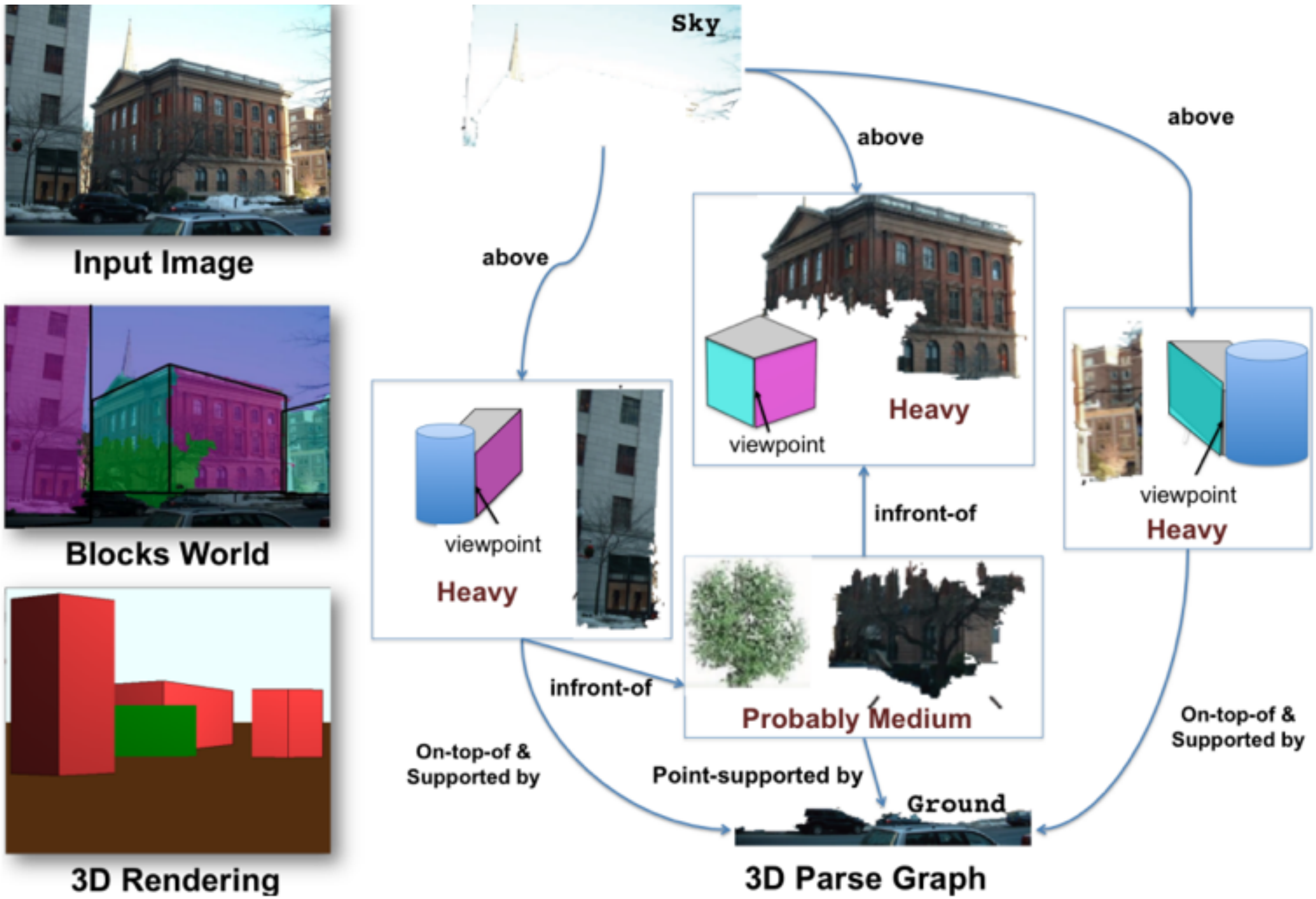
(b) P(person) = uniform (d) P(person | geometry) (f) P(person | viewpoint) (g) P(person|viewpoint,geometry)

D. Hoiem, A. Efros, and M. Herbert, **Putting Objects in Perspective**, CVPR 2006

Geometry and recognition

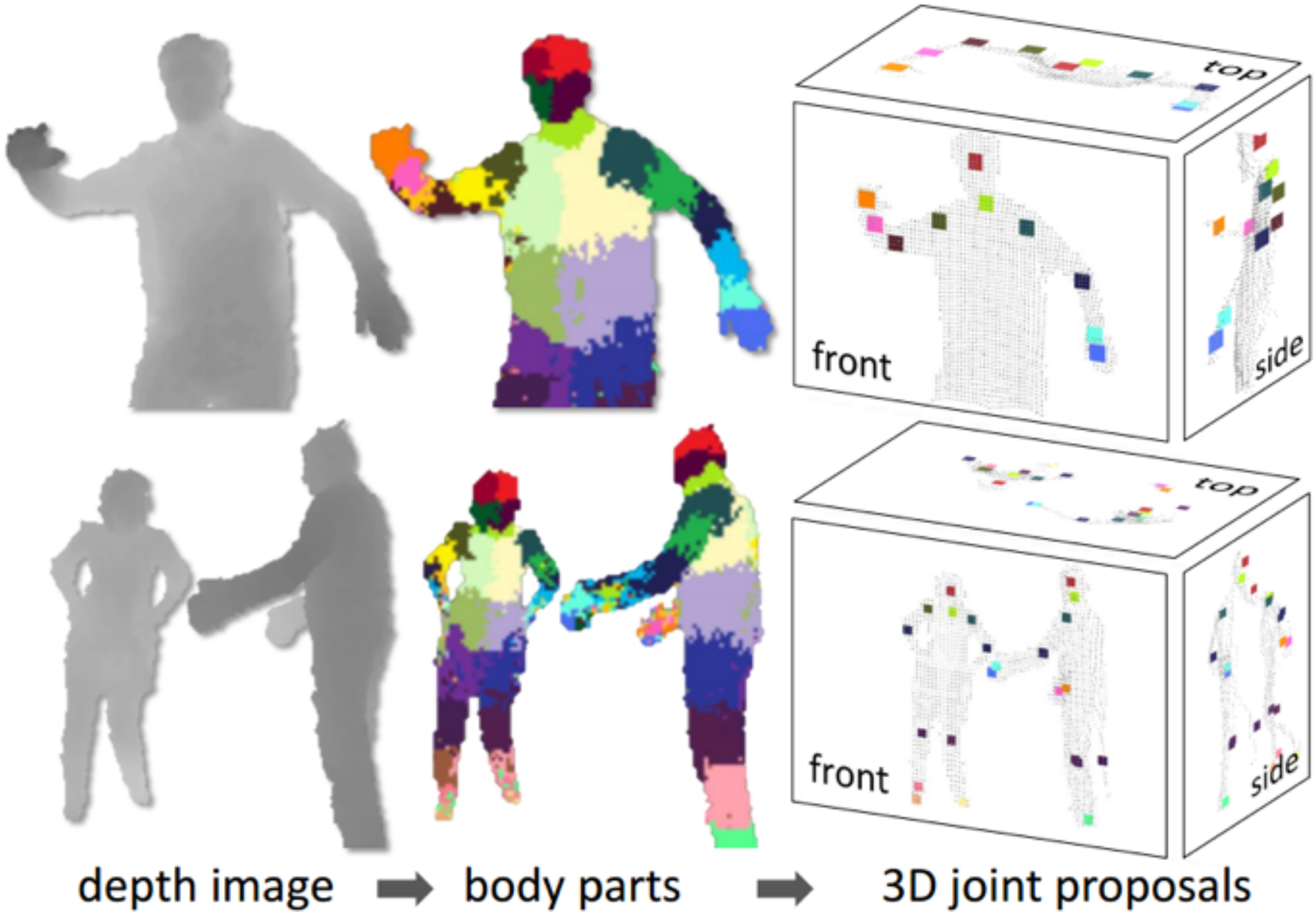
V. Hedau, D. Hoiem, and D. Forsyth, **Recovering the Spatial Layout of Cluttered Rooms**, ICCV 2009.

Geometry and recognition



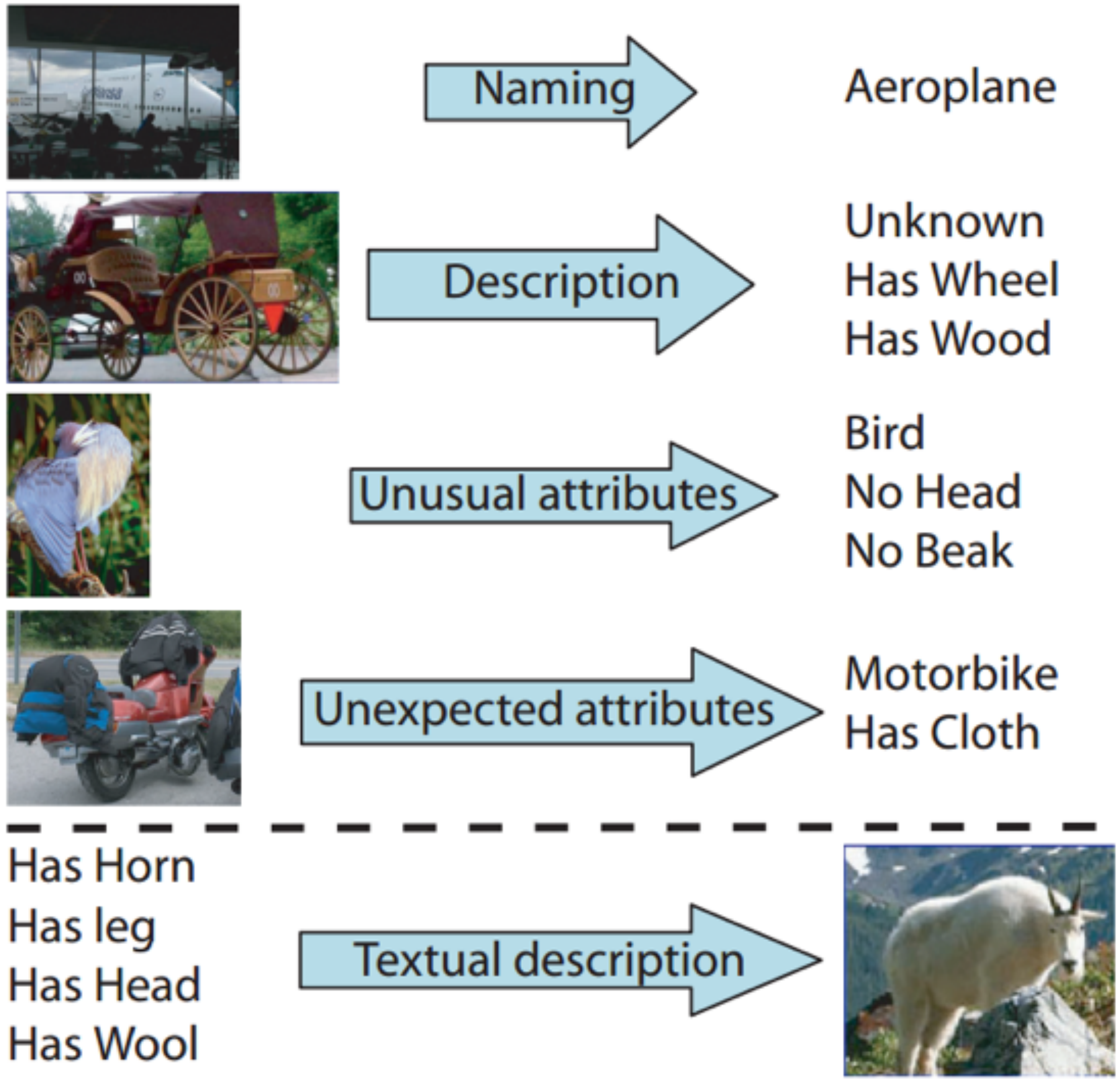
A. Gupta, A. Efros and M. Hebert, [Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics](), ECCV 2010

Recognition from RGBD Images



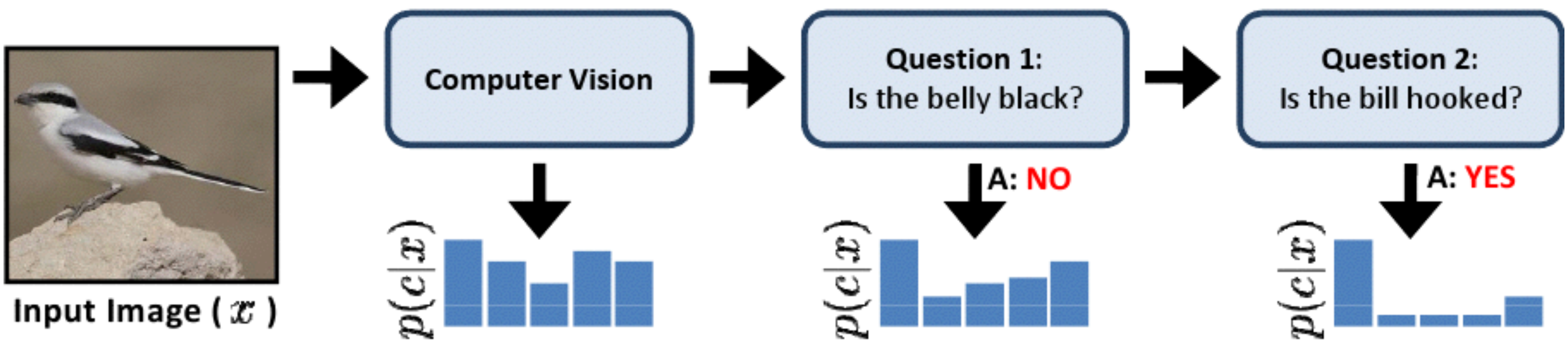
J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, **Real-Time Human Pose Recognition in Parts from a Single Depth Image**, CVPR 2011

Attributes for recognition



A. Farhadi, I. Endres, D. Hoiem, and D Forsyth, **Describing Objects by their Attributes**, CVPR 2009

Human "in the loop" recognition



Branson S., Wah C., Babenko B., Schroff F., Welinder P., Perona P., Belongie S., "Visual Recognition with Humans in the Loop", European Conference on Computer Vision (ECCV), Heraklion, Crete, Sept., 2010.

Crowdsourcing

- Large datasets become the norm (real world settings)
  - LabelMe, PASCAL VOC, ImageNet
  - Enable new machine learning methods (e.g., deep learning)

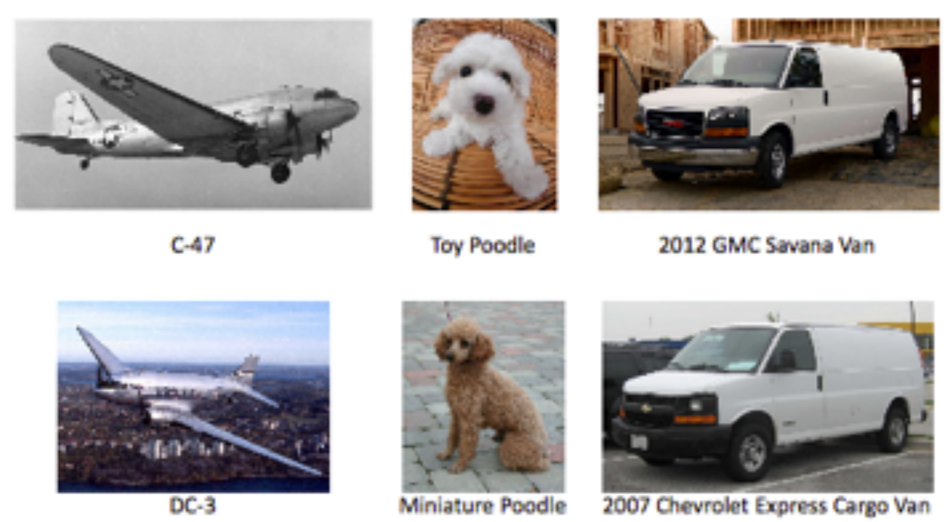
http://www.blogging4jobs.com/hr/solve-your-workplace-issues-by-crowdsourcing/

amazon mechanical turk

Fine-grained recognition

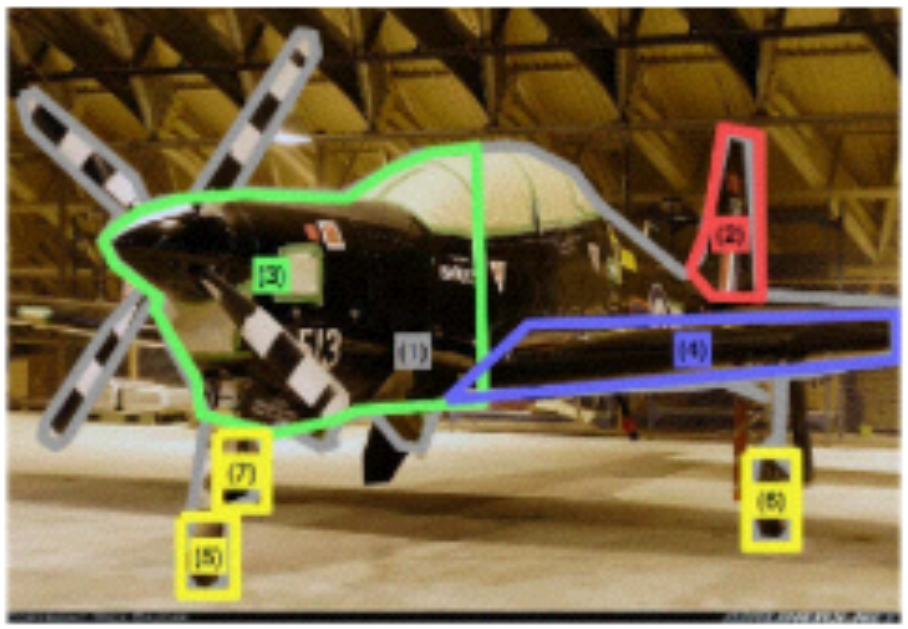
many related classes

often confused



Understanding objects in detail

OID:Aircraft Benchmark



1 **aeroplane** facing-direction: SW; is-airliner: no; is-cargo-plane: no; is-glider: no; is-military-plane: yes; is-propellor-plane: yes; is-seaplane: no; plane-location: on ground/water; plane-size: medium plane; wing-type: single wing plane; undercarriage-arrangement: one-front-two-back; airline: UK–Air Force; model: Short S-312 Tucano T1 2; 2 **vertical stabilizer** tail-has-engine: no-engine 3 **nose** has-engine-or-sensor: has-engine 4 **wing** wing-has-engine: no-engine 5 **undercarriage** cover-type: retractable; group-type: 1-wheel-1-axle; location: front-middle 6 **undercarriage** cover-type: retractable; group-type: 1-wheel-1-axle; location: back-left 7 **undercarriage** cover-type: retractable; group-type: 1-wheel-1-axle; location: back-right.

Vedaldi et al., CVPR 14

Sentence generation


This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.


There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.


There are one cow and one sky. The golden cow is by the blue sky.


There are one dining table, one chair and two windows. The wooden dining table is by the wooden chair, and against the first window, and against the second white window. The wooden chair is by the first window, and by the second white window. The first window is by the second white window.


This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.


Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.


Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.


This is a picture of two dogs. The first dog is near the second furry dog.

G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, T. Berg, [Baby Talk: Understanding and Generating Simple Image Descriptions](), CVPR 2011

Deep learning

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

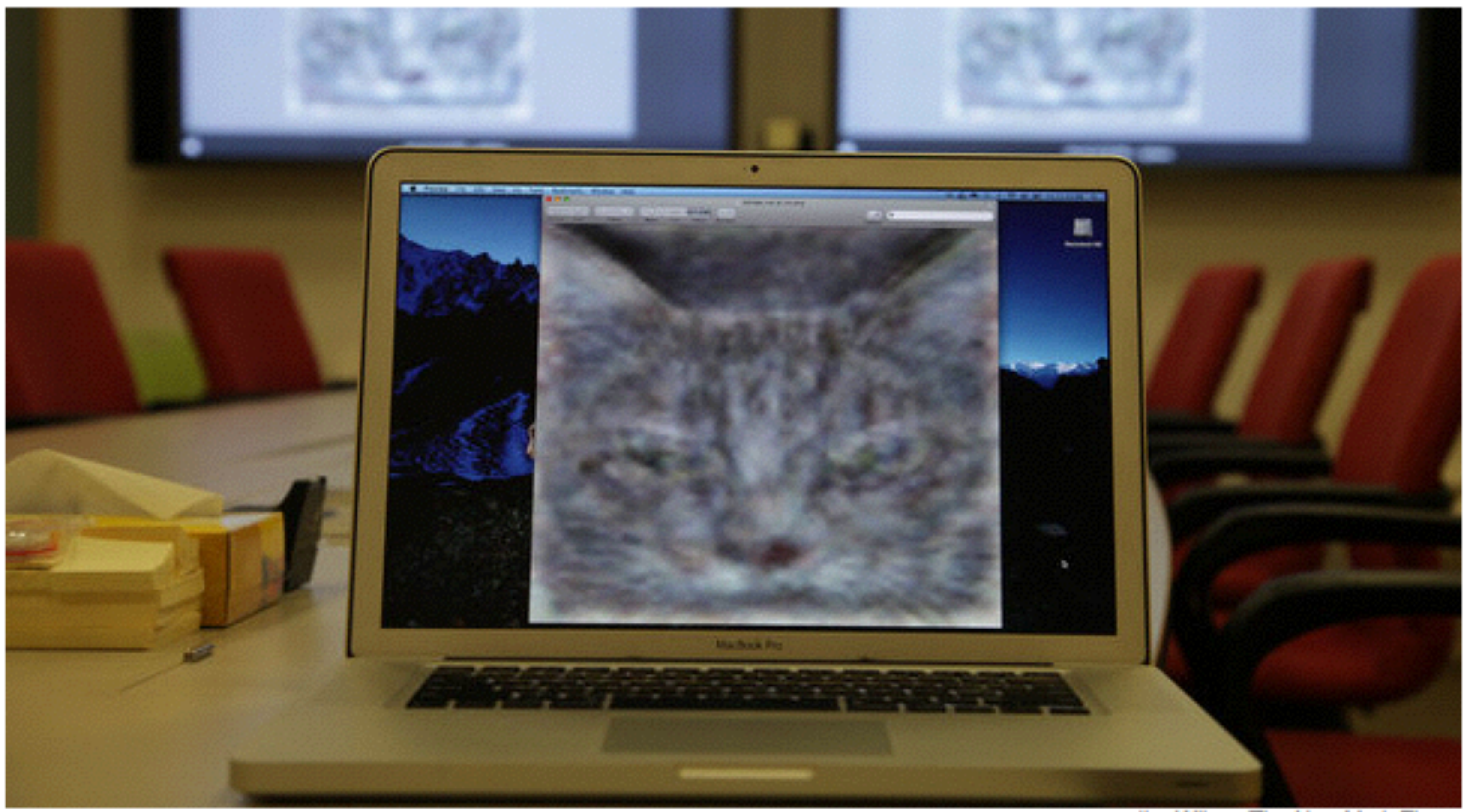
The New York Times

Business Day

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

How Many Computers to Identify a Cat? 16,000



Jim Wilson/The New York Times

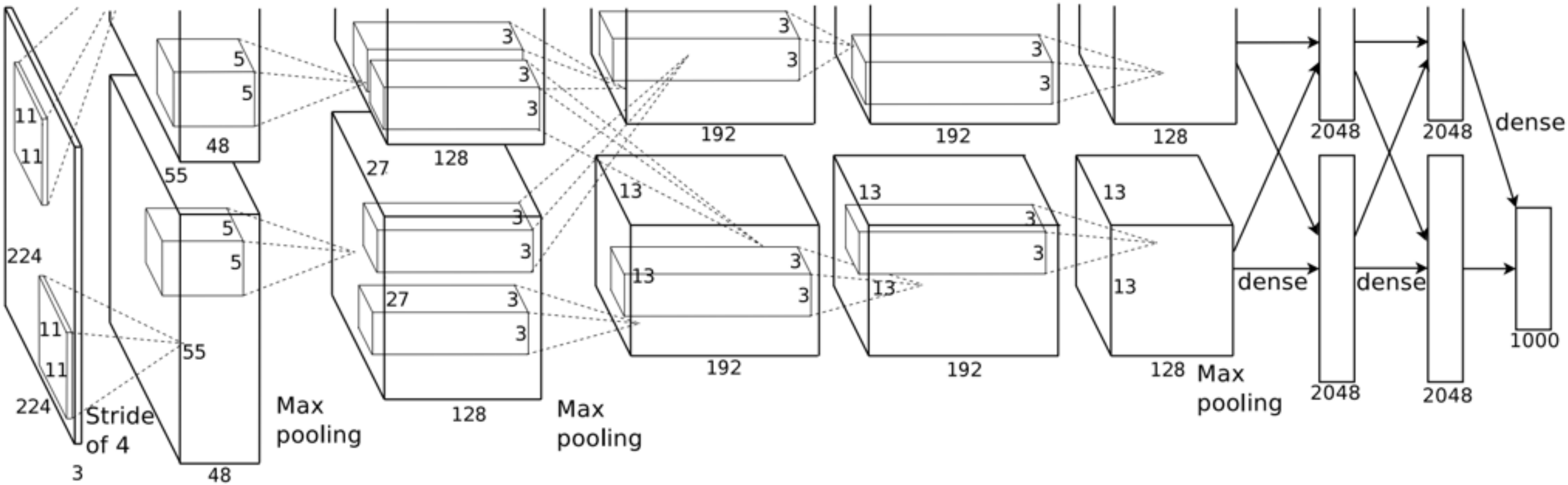
An image of a cat that a neural network taught itself to recognize.

By JOHN MARKOFF

Published: June 25, 2012

[NY Times article]

Recent deep learning breakthroughs…



[ImageNet Classification with Deep Convolutional Neural Networks](#) Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton NIPS 2014

96 filters learned in layer 1

Today

- Conclude “introduction to recognition”
- Introduction to machine learning
  - learning to recognize
  - machine learning framework
  - properties of learning algorithms
- Common datasets in computer vision

Recognition: A machine learning approach

The machine learning framework

Apply a prediction function to a feature representation of the image to get the desired output:

f() = "apple"

f() = "tomato"

f() = "cow"

The machine learning framework

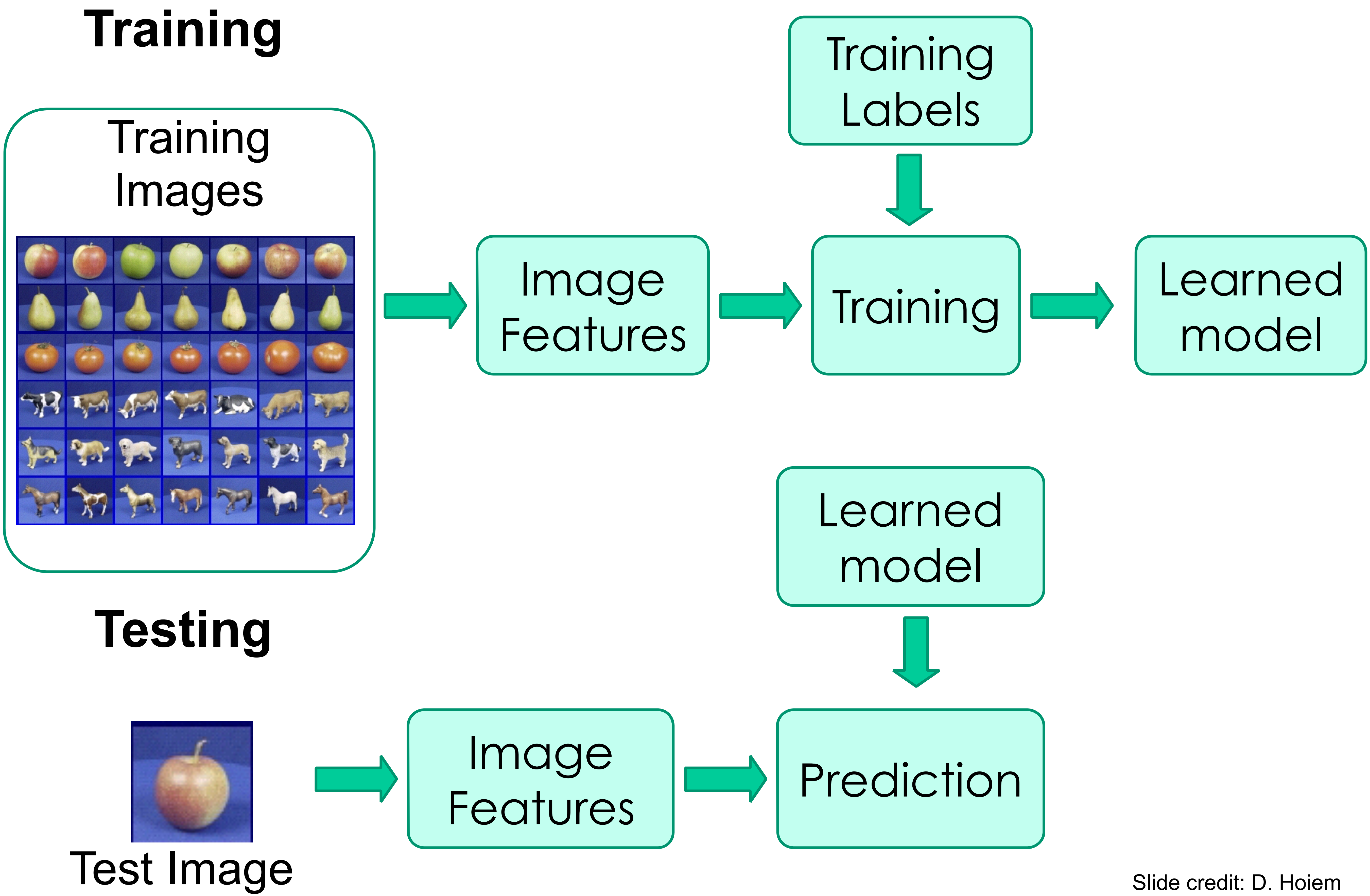
$$y = f(\mathbf{x})$$

output prediction function Image feature

Training: given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

Testing: apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

Steps

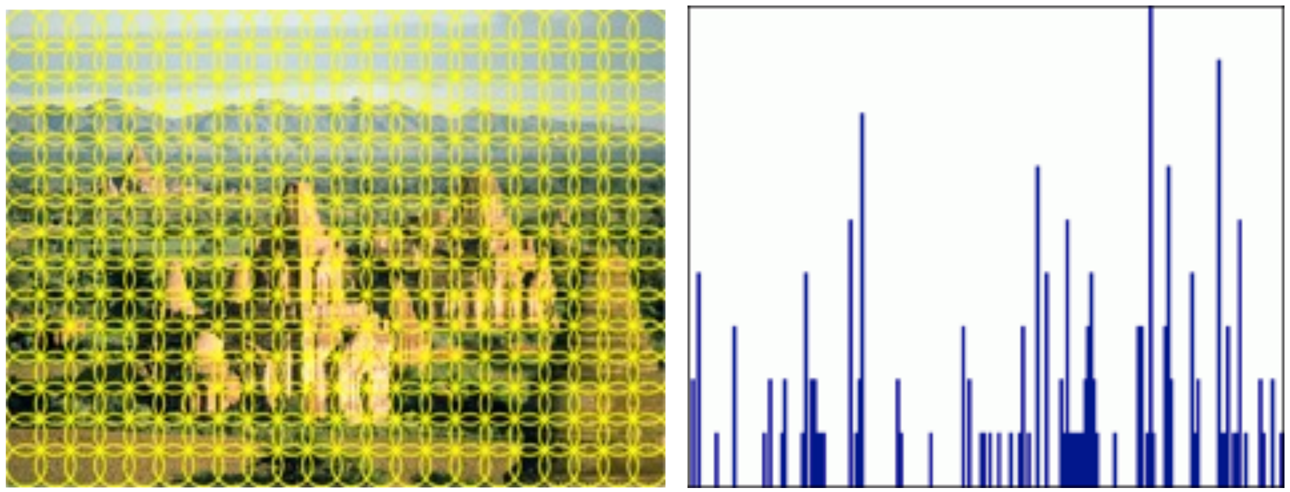


Slide credit: D. Hoiem

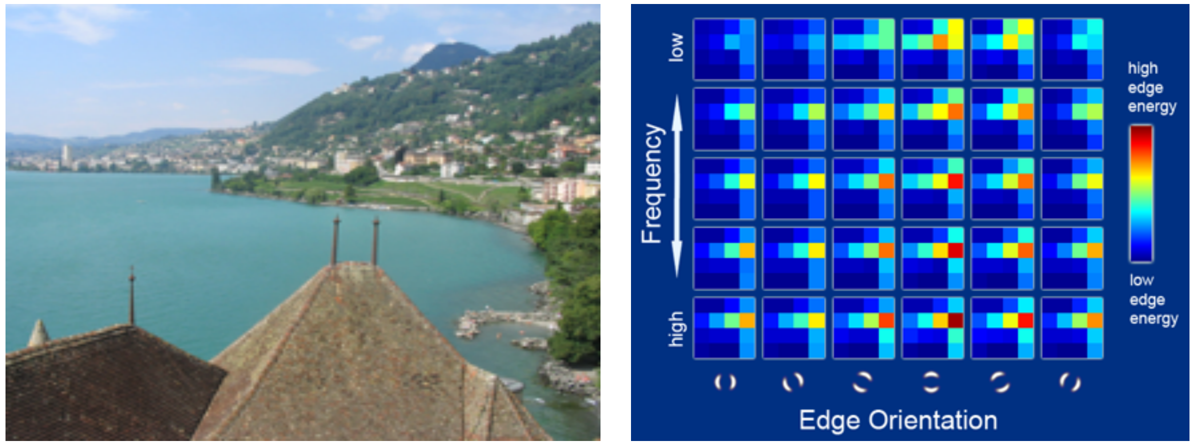
Features (examples)

Raw pixels (and simple functions of raw pixels)

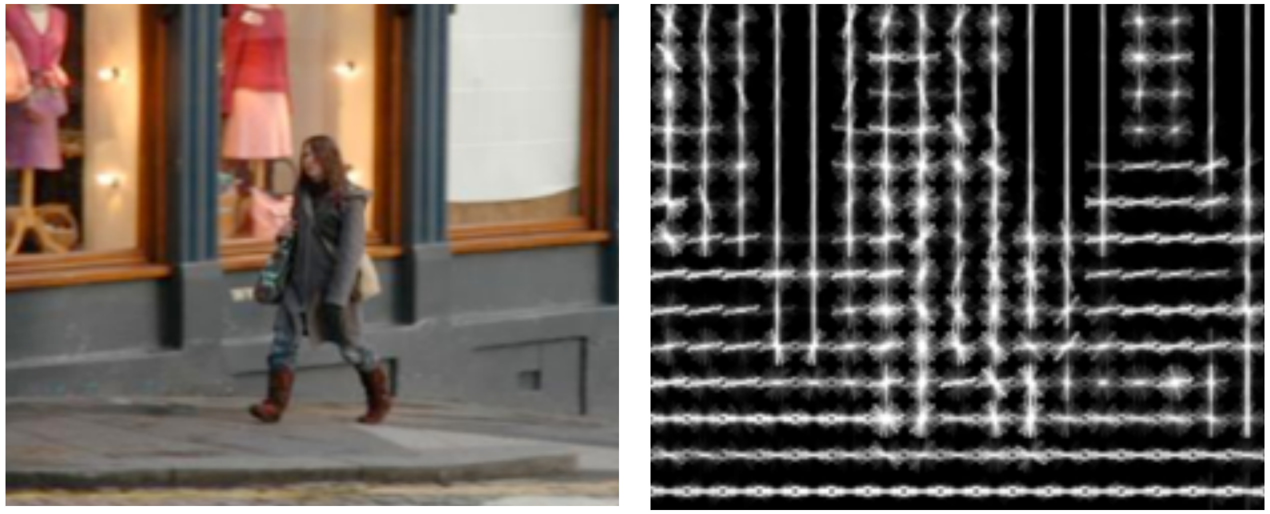
Histograms, bags of features



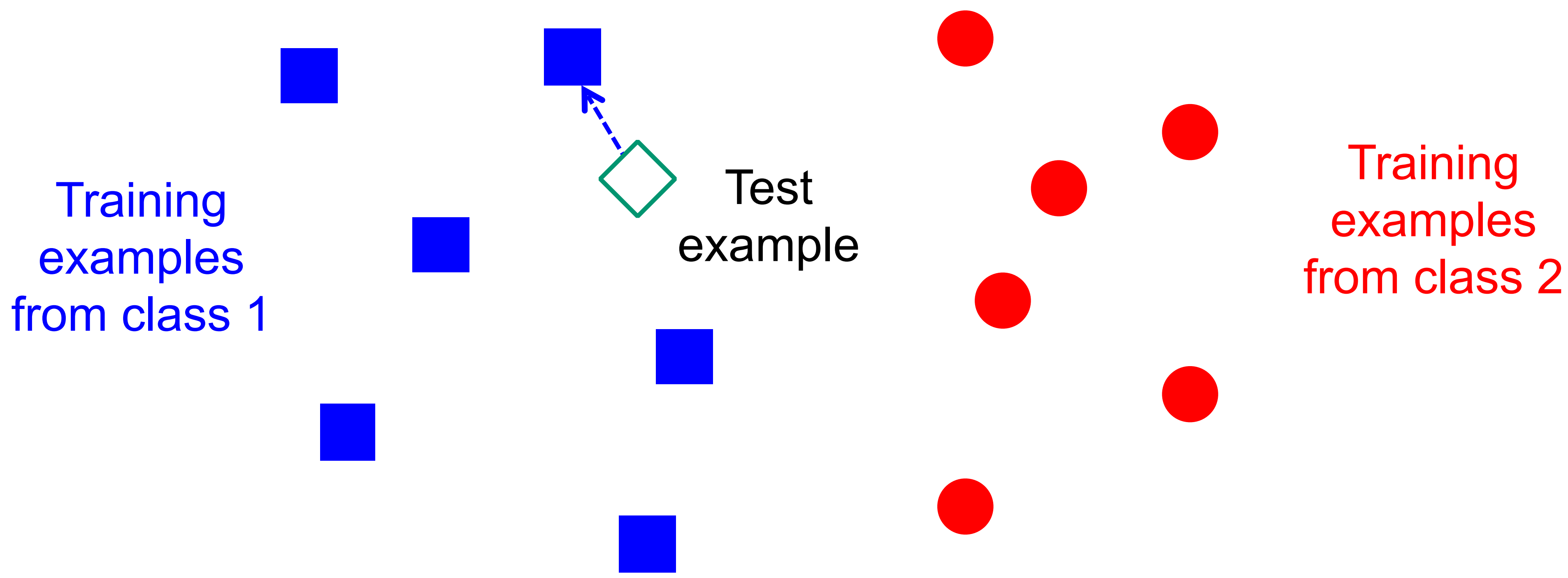
GIST descriptors



Histograms of oriented gradients(HOG)



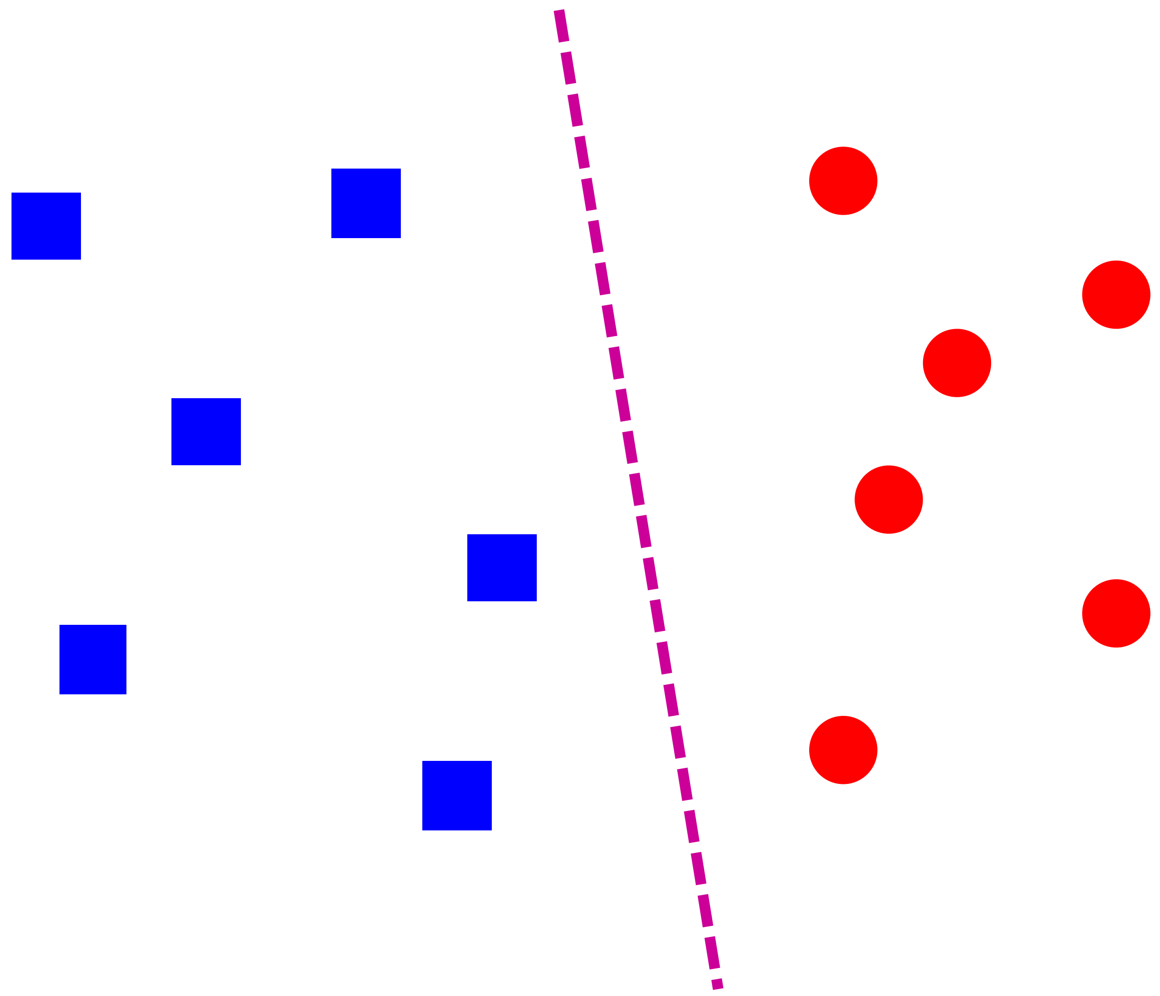
Classifiers: Nearest neighbor



f(**x**) = label of the training example nearest to **x**

All we need is a distance function for our inputs
No training required!

Classifiers: Linear

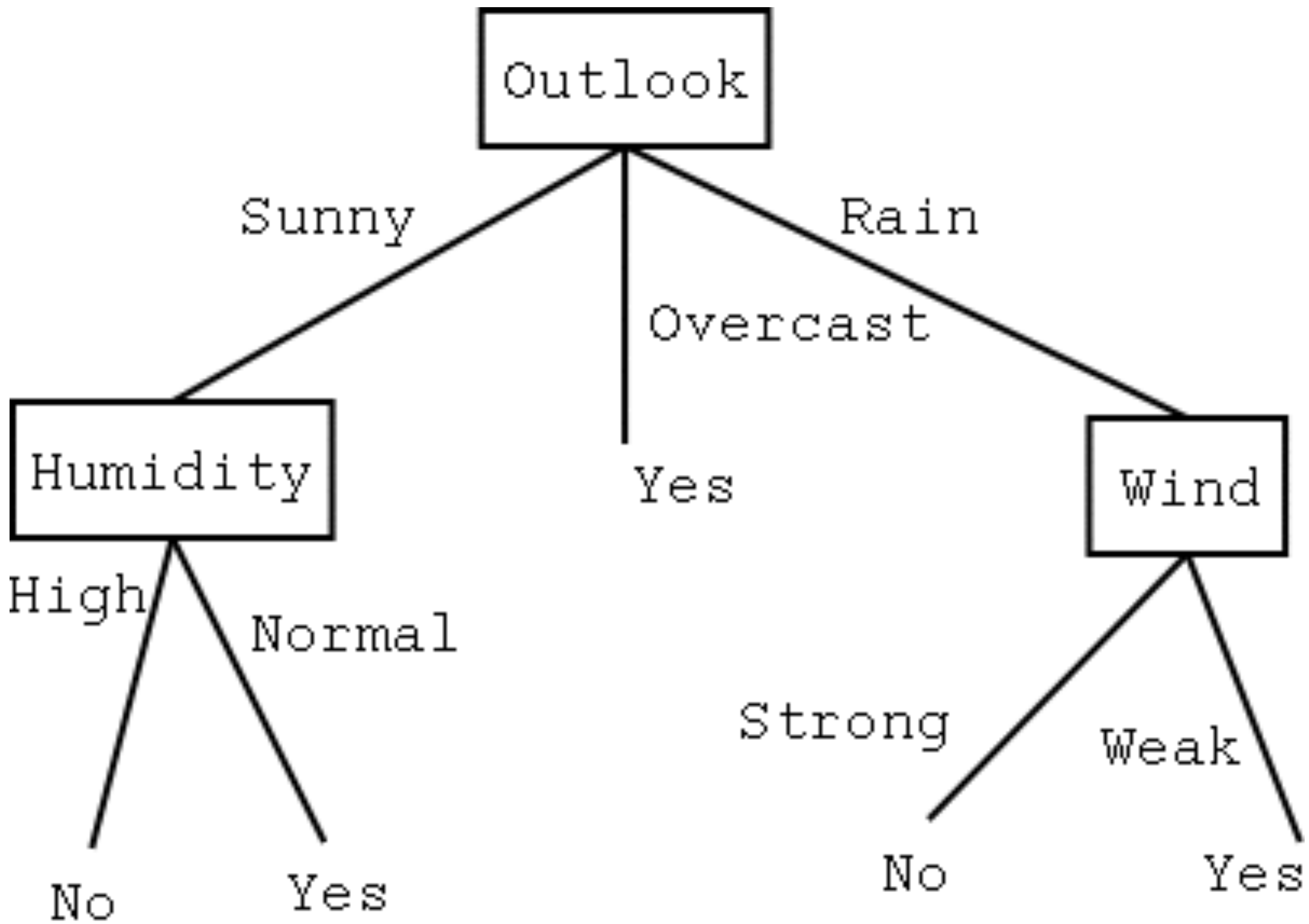


Find a *linear function* to separate the classes:

$$f(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$$

Classifier: Decision trees

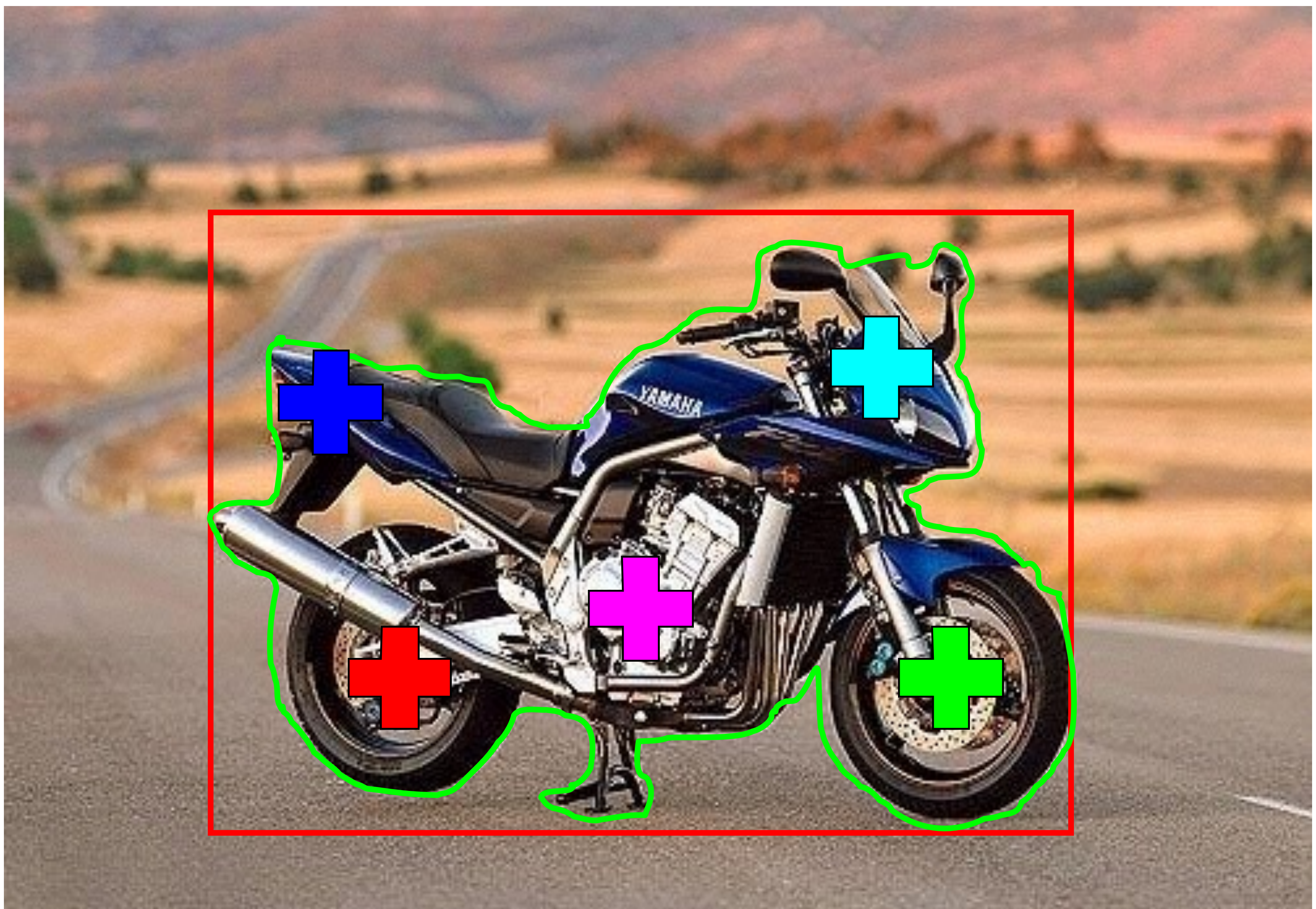
Play tennis?



Recognition task and supervision

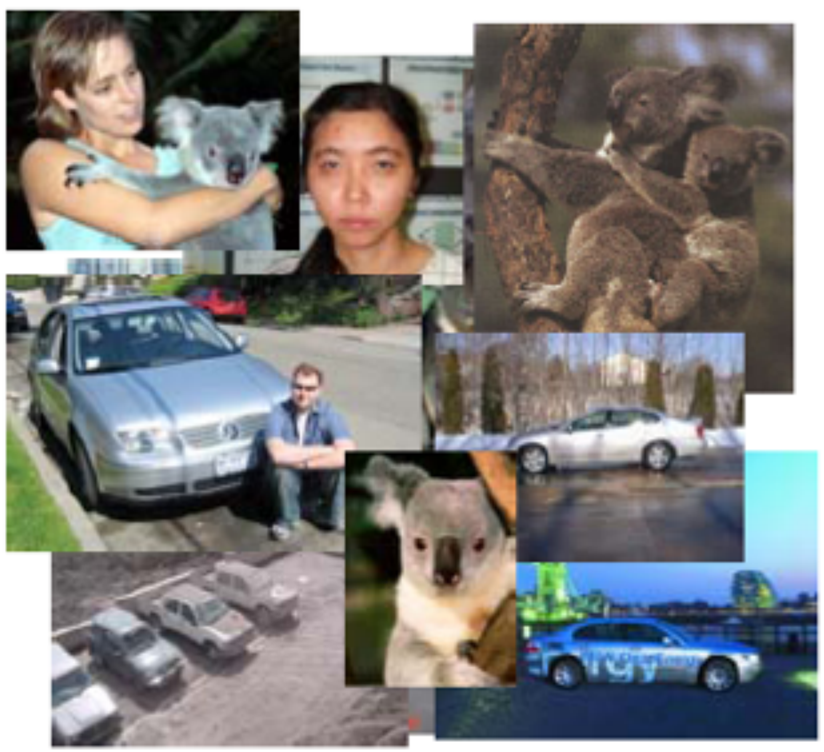
- Images in the training set must be annotated with the "correct answer" that the model is expected to produce

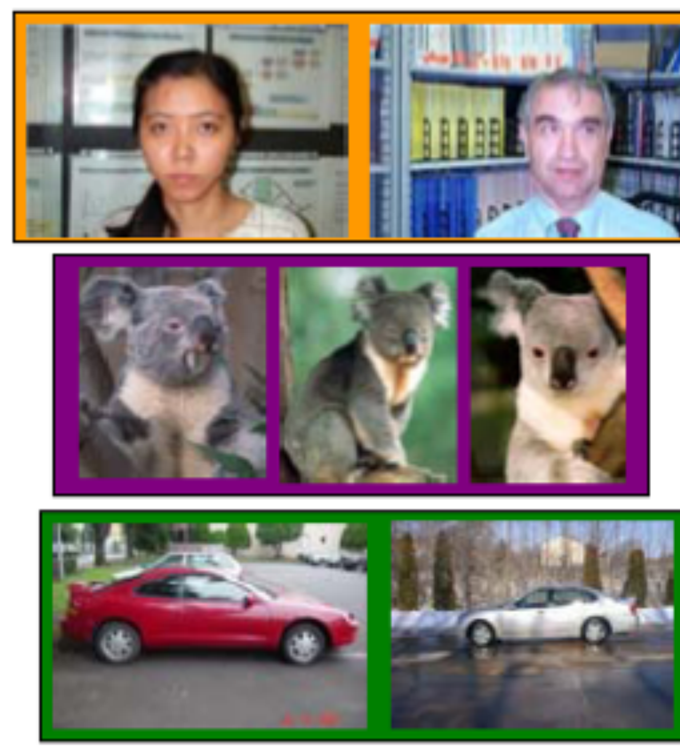
Contains a motorbike



Spectrum of supervision

Less **More**



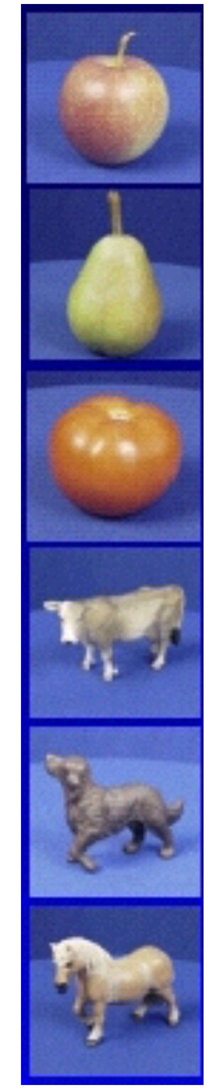


Unsupervised "Weakly" supervised Fully supervised

Definition depends on task

Generalization

Training set (labels known)



Test set (labels unknown)

How well does a learned model *generalize* from the data it was trained on to a new test set?

Diagnosing generalization ability

Training error: how well does the model perform at prediction on the data on which it was trained?

Test error: how well does it perform on a never before seen test set?

Training and test error are both *high*: **underfitting**

- Model does an equally poor job on the training and the test set
- Either the training procedure is ineffective or the model is too "simple" to represent the data

Training error is *low* but test error is *high*: **overfitting**

- Model has fit irrelevant characteristics (noise) in the training data
- Model is too complex or amount of training data is insufficient

Underfitting and overfitting

Underfitting Good generalization Overfitting

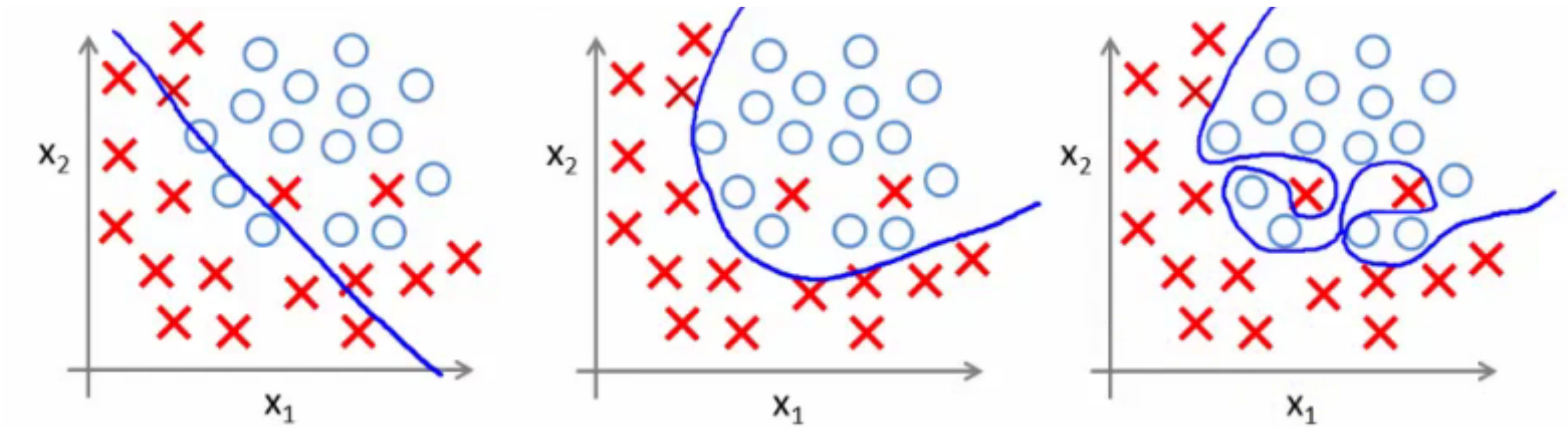
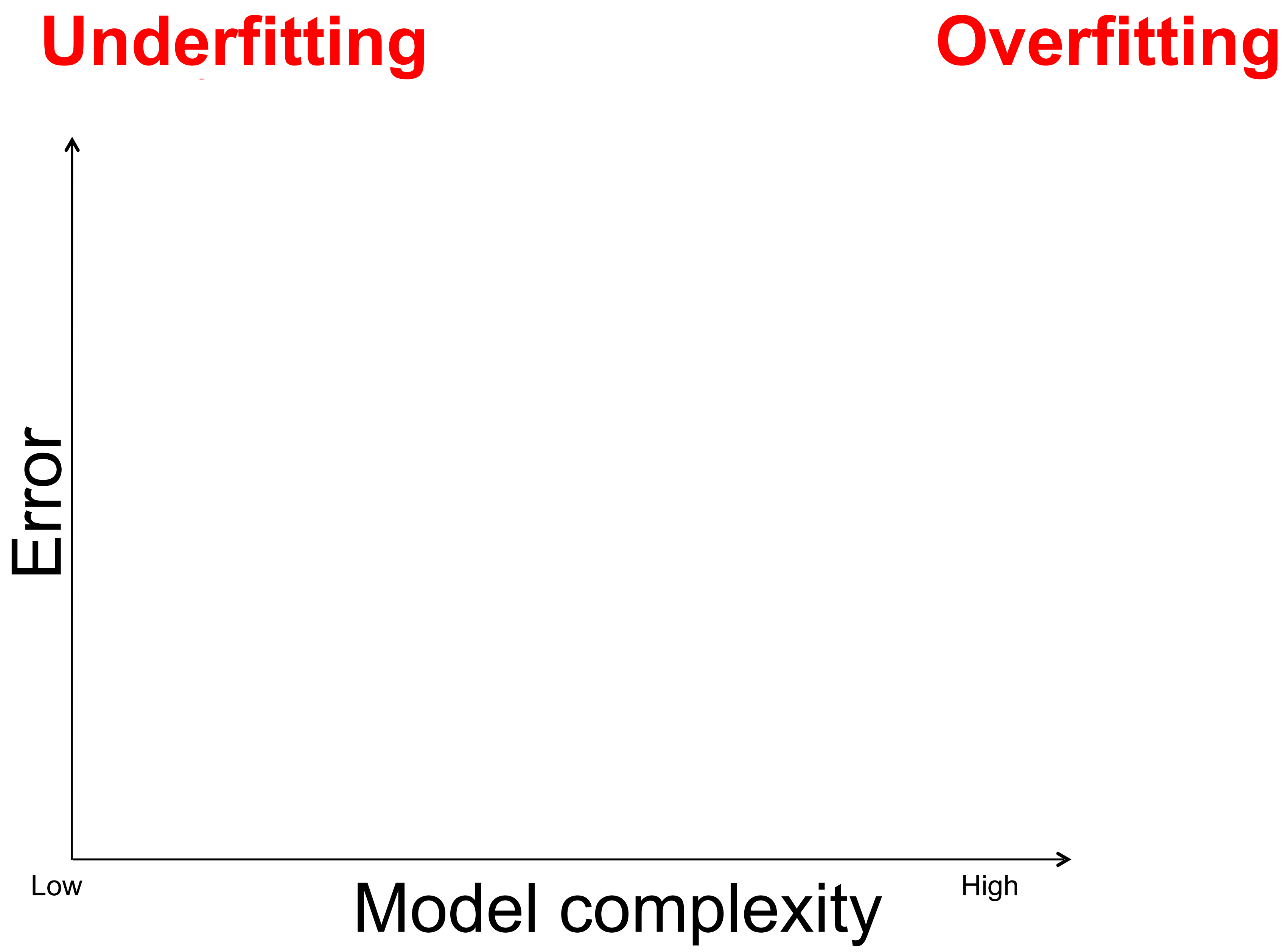


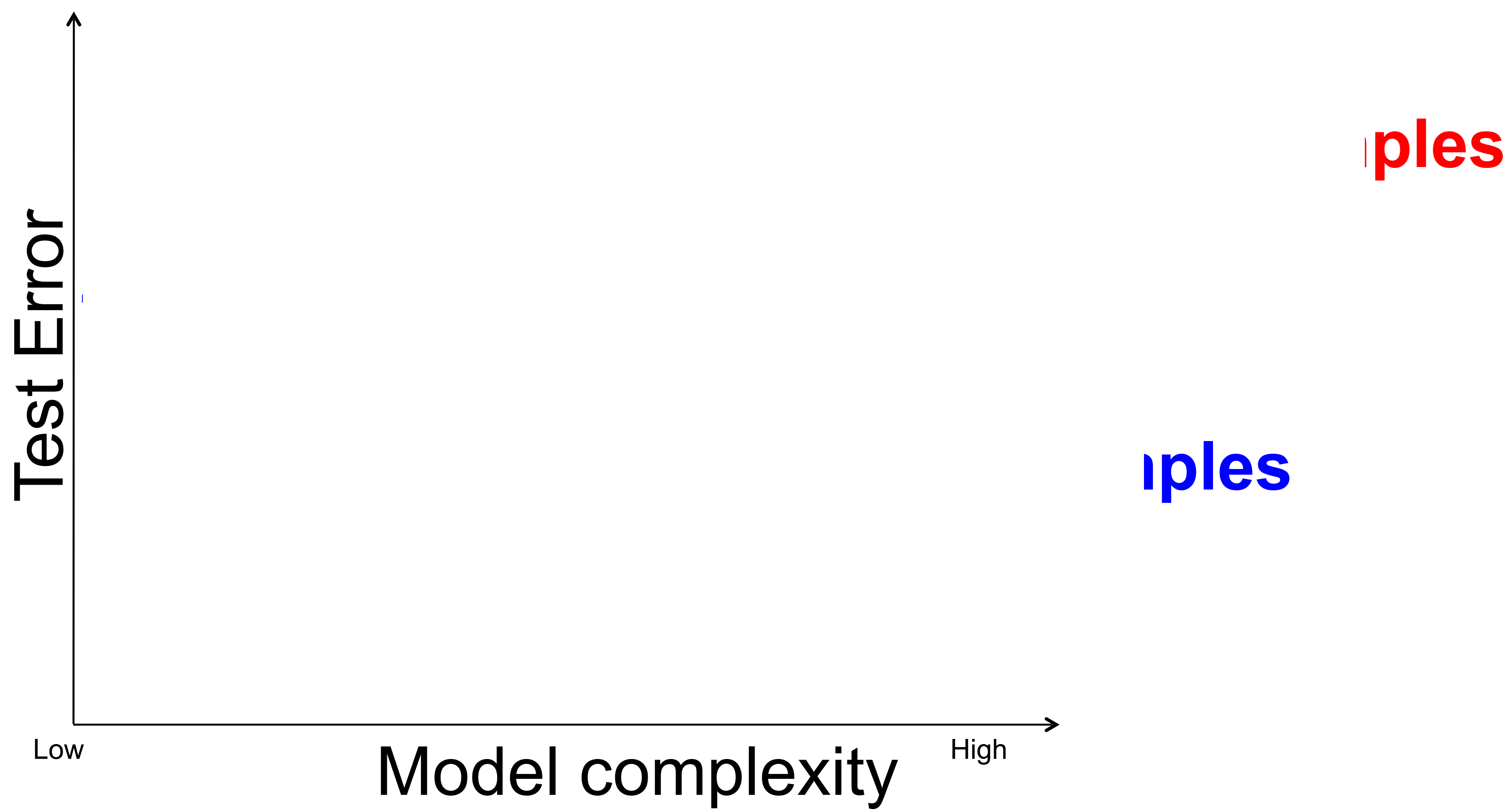
Figure source

Effect of model complexity



Slide credit: D. Hoiem

Effect of training set size



Slide credit: D. Hoiem

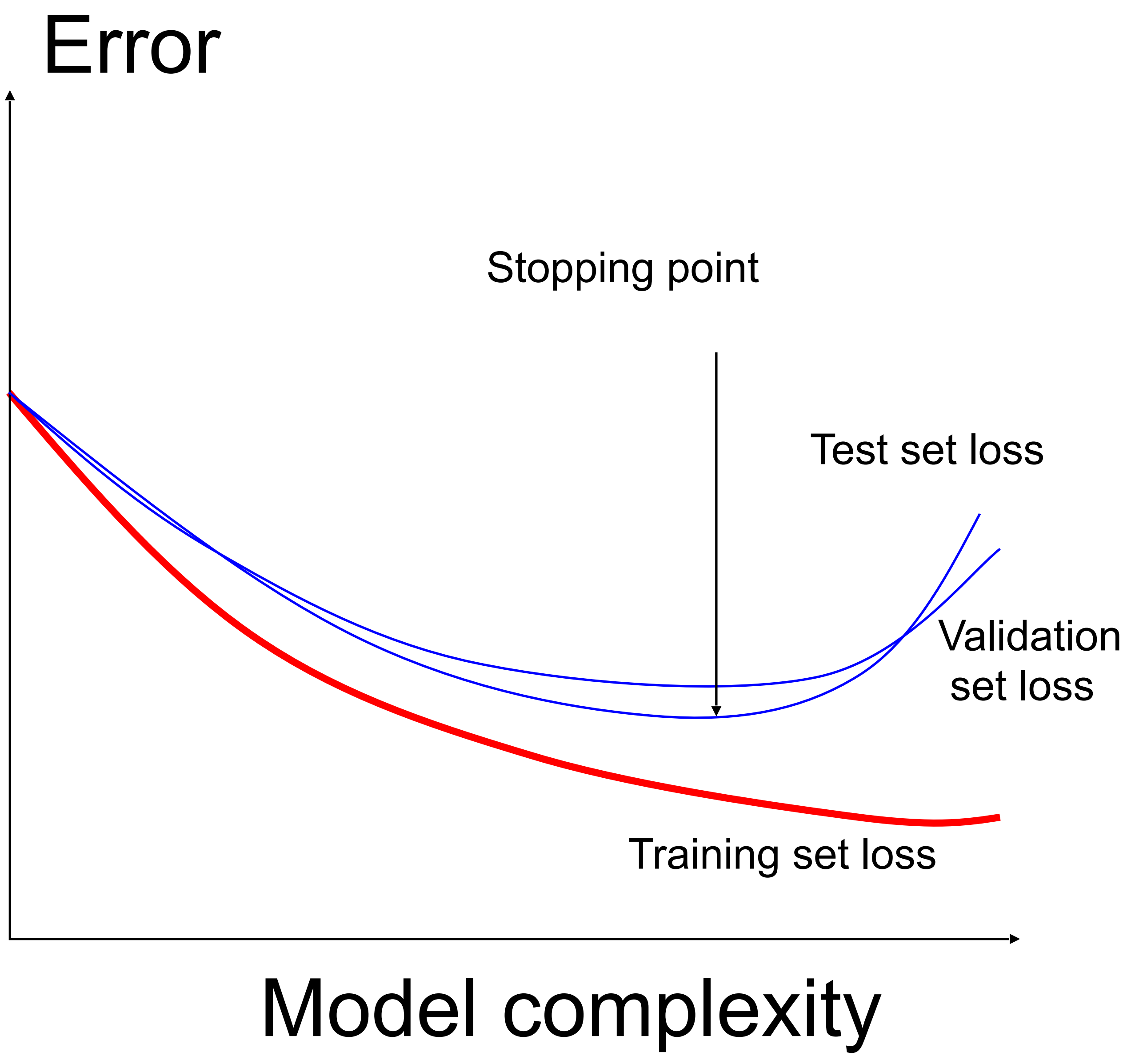
Validation

Split the dataset into **training**, **validation**, and **test** sets

Use training set to **optimize model parameters**

Use validation test to **choose the best model**

Use test set only to **evaluate performance**



Datasets

Circa 2001: five categories, hundreds of images per category

Circa 2004: 101 categories

Today: up to thousands of categories, millions of images

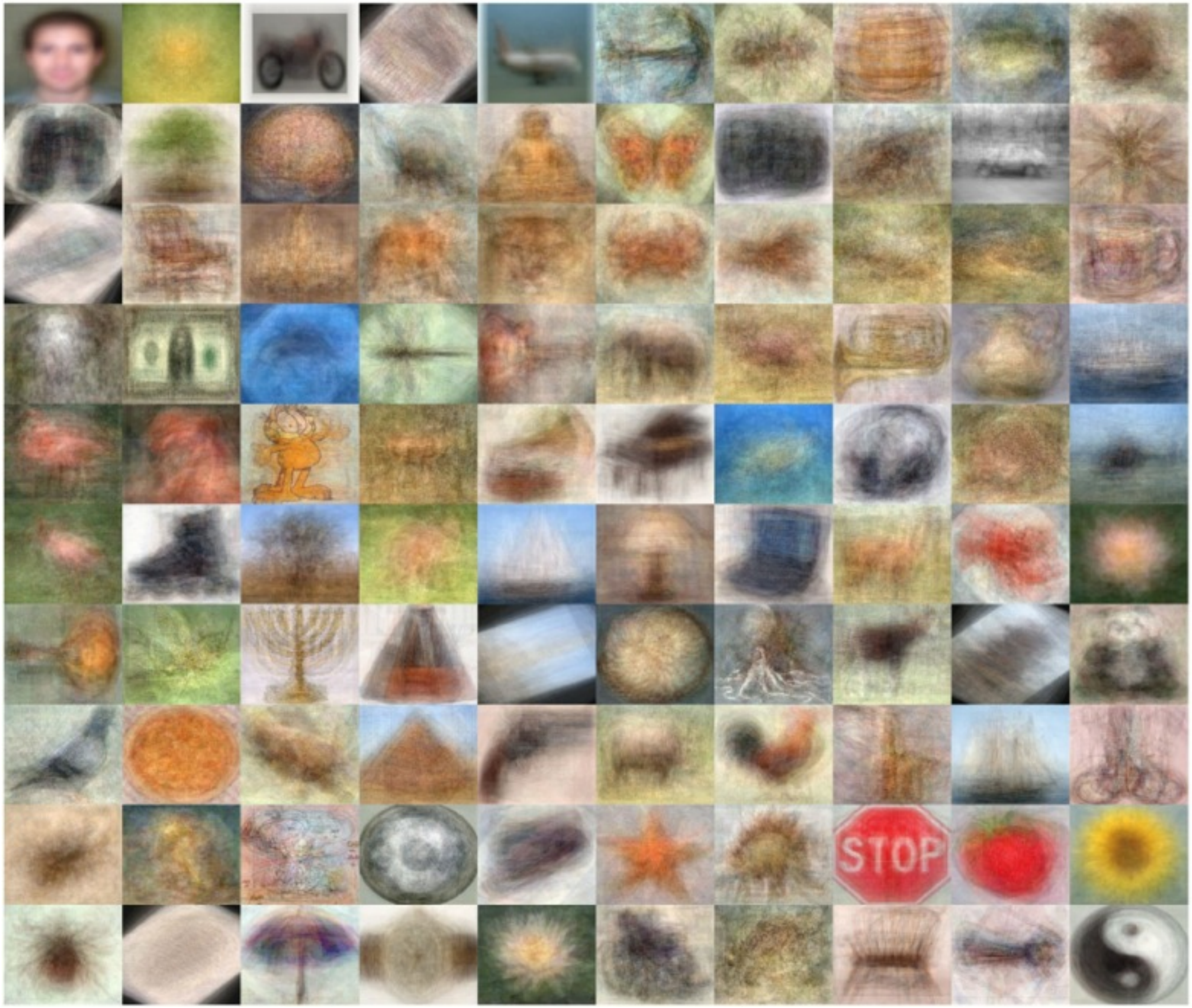
Caltech 101 & 256

http://www.vision.caltech.edu/Image_Datasets/Caltech101/
http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007

Caltech-101: Intra-class variability



The PASCAL Visual Object Classes Challenge (2005-2012)

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

- **Challenge 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

- **Dataset size (by 2012):**
  11.5K training/validation images, 27K bounding boxes, 7K segmentations

PASCAL competitions

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Classification: For each of the twenty classes, predicting presence/absence of an example of that class in the test image

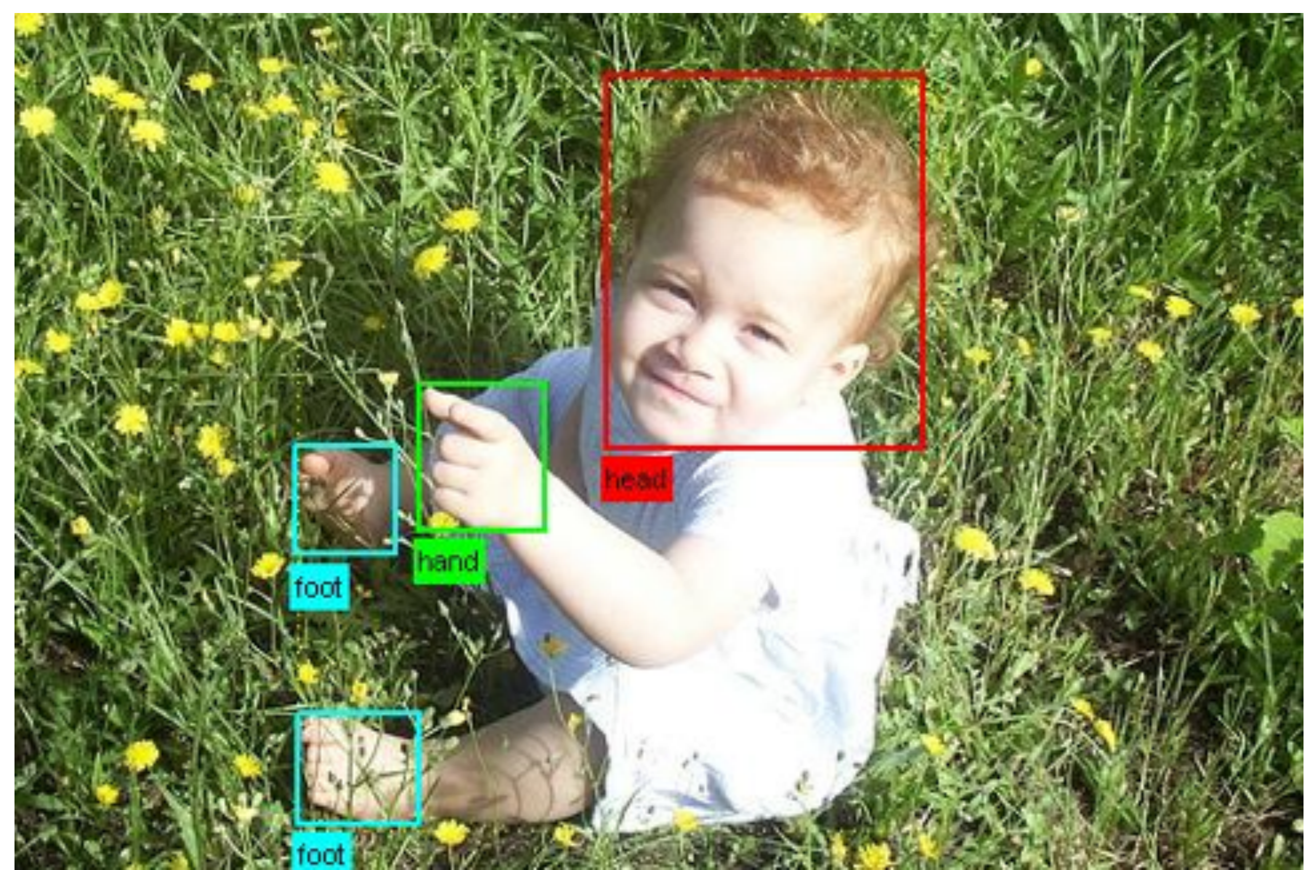
Detection: Predicting the bounding box and label of each object from the twenty target classes in the test image

PASCAL competitions

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Segmentation: Generating pixel-wise segmentations giving the class of the object visible at each pixel, or "background" otherwise

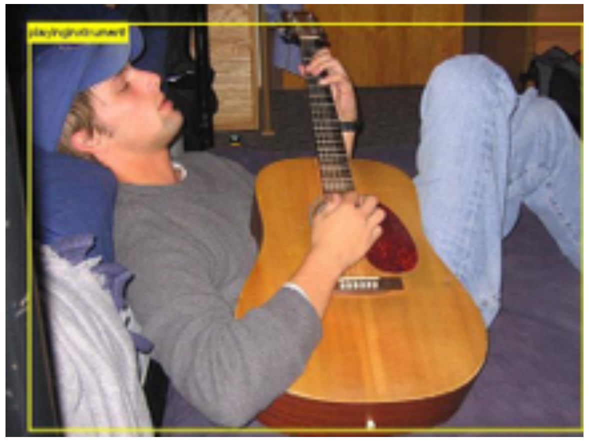
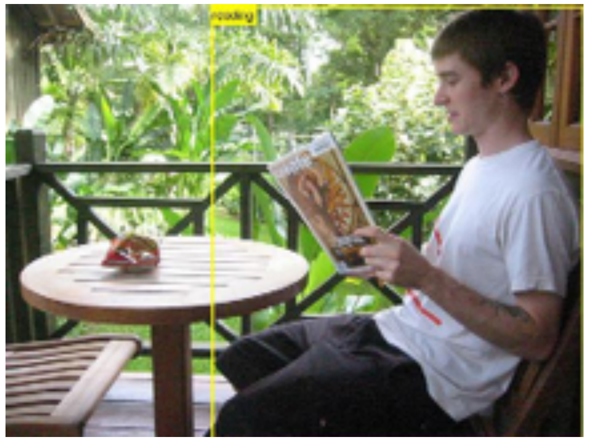
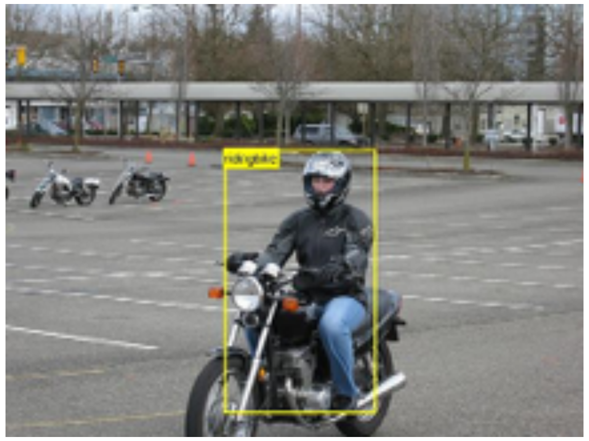
Person layout: Predicting the bounding box and label of each part of a person (head, hands, feet)

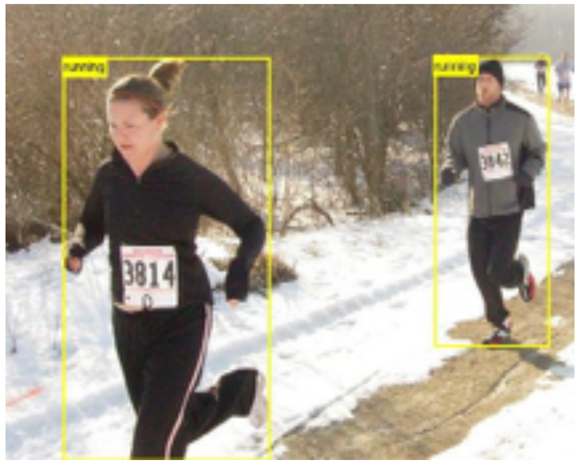
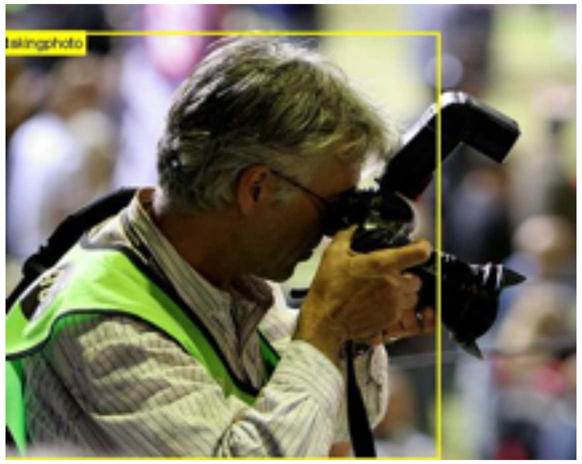


PASCAL competitions

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Action classification (10 action classes)


LabelMe Dataset http://labelme.csail.mit.edu/

Russell, Torralba, Murphy, Freeman, 2008

ImageNet

http://www.image-net.org/

IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications CoolStuff About

Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

What do these images have in common? *Find out!*

Check out the ImageNet Challenge 2014!

Further thoughts and readings

- Chapter 14 of Szeliski's book
- A good reference especially for discriminative learning:

  The Elements of Statistical Learning
  T. Hastie and R Tibshirani (2001,2009)
  http://www.stanford.edu/~hastie/local.ftp/Springer/ESLII_print10.pdf