Recognition:
A machine learning approach

Slides adapted from Fei-Fei Li, Rob Fergus, Antonio Torralba, Kristen Grauman, and Derek Hoiem

The machine learning framework

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

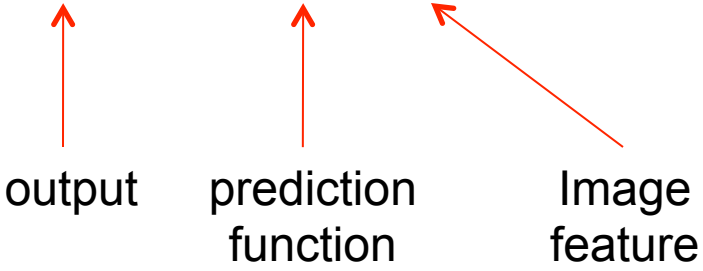
f() = “apple”

f() = “tomato”

f() = “cow”

The machine learning framework

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



- **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

Training

Training Images

Training Labels

Image Features

Training

Learned model

Testing

Test Image

Image Features

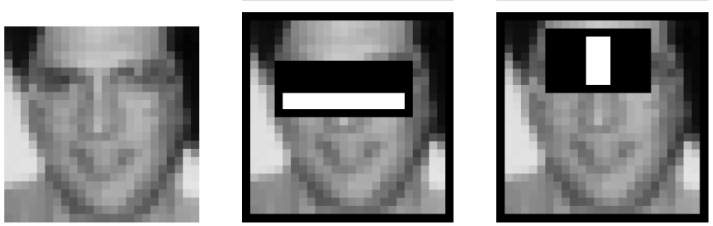
Learned model

Prediction

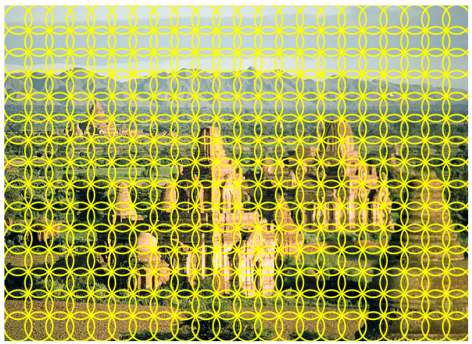
Slide credit: D. Hoiem

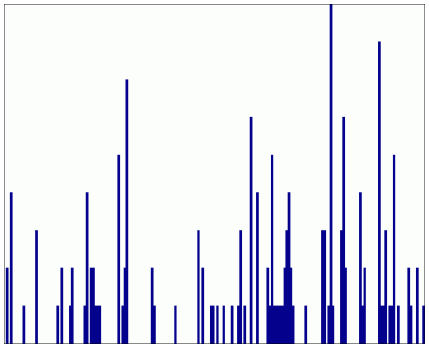
Features (examples)

- Raw pixels (and simple functions of raw pixels)

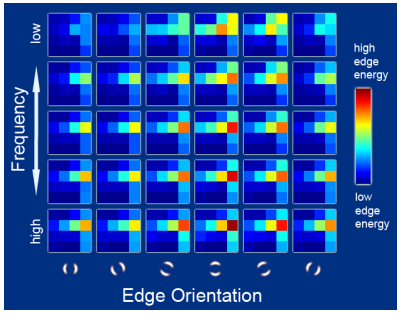


- Histograms, bags of features

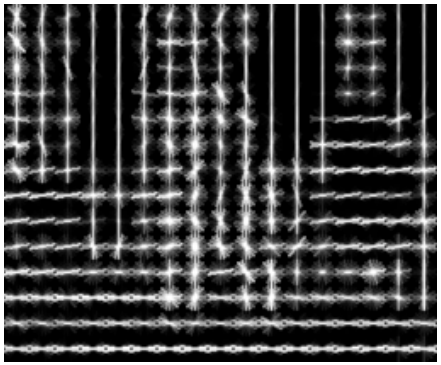




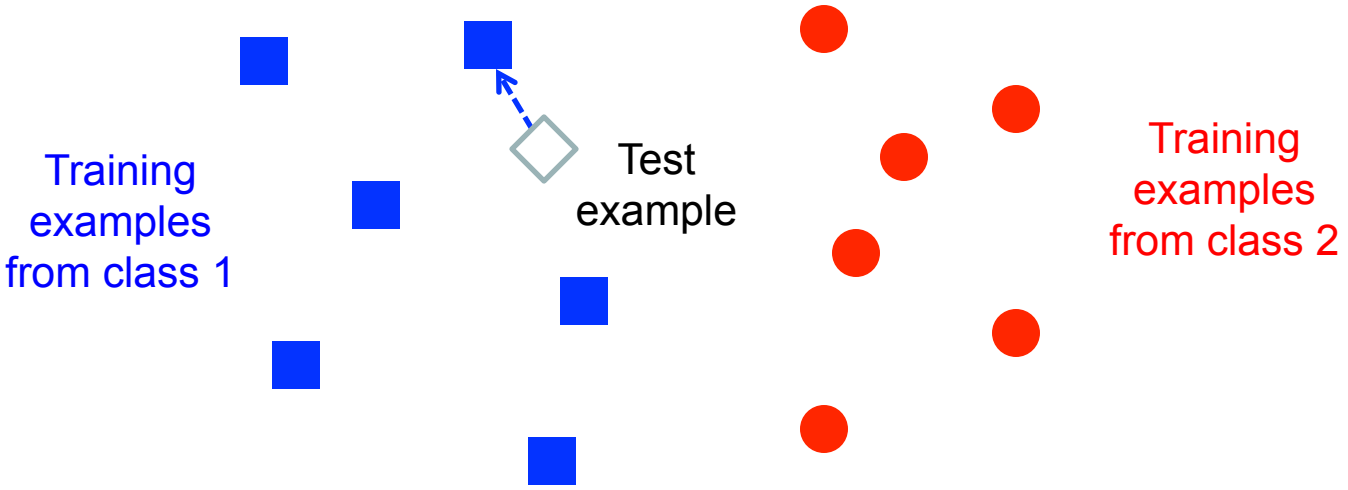
- GIST descriptors



- Histograms of oriented gradients (HOG)



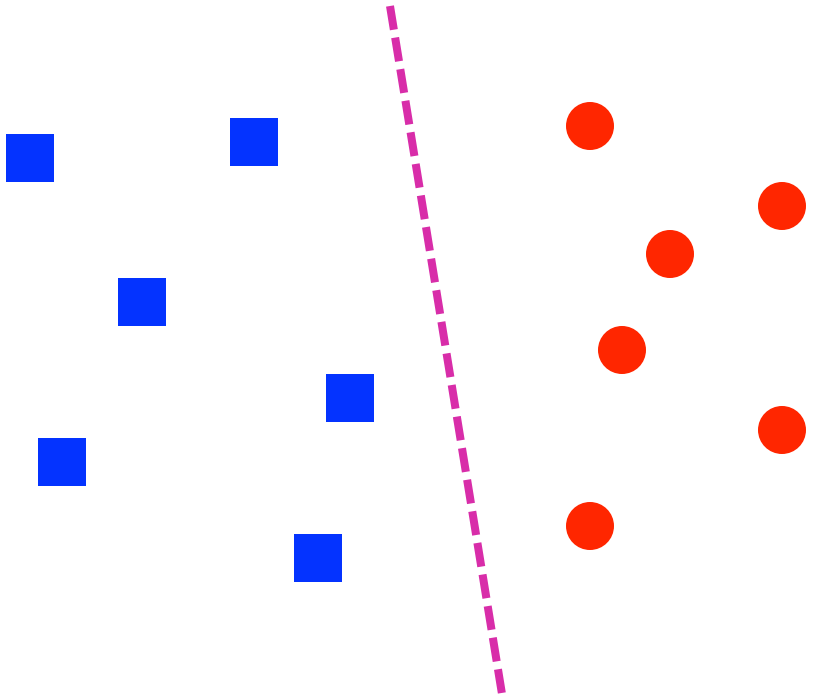
Classifiers: Nearest neighbor



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



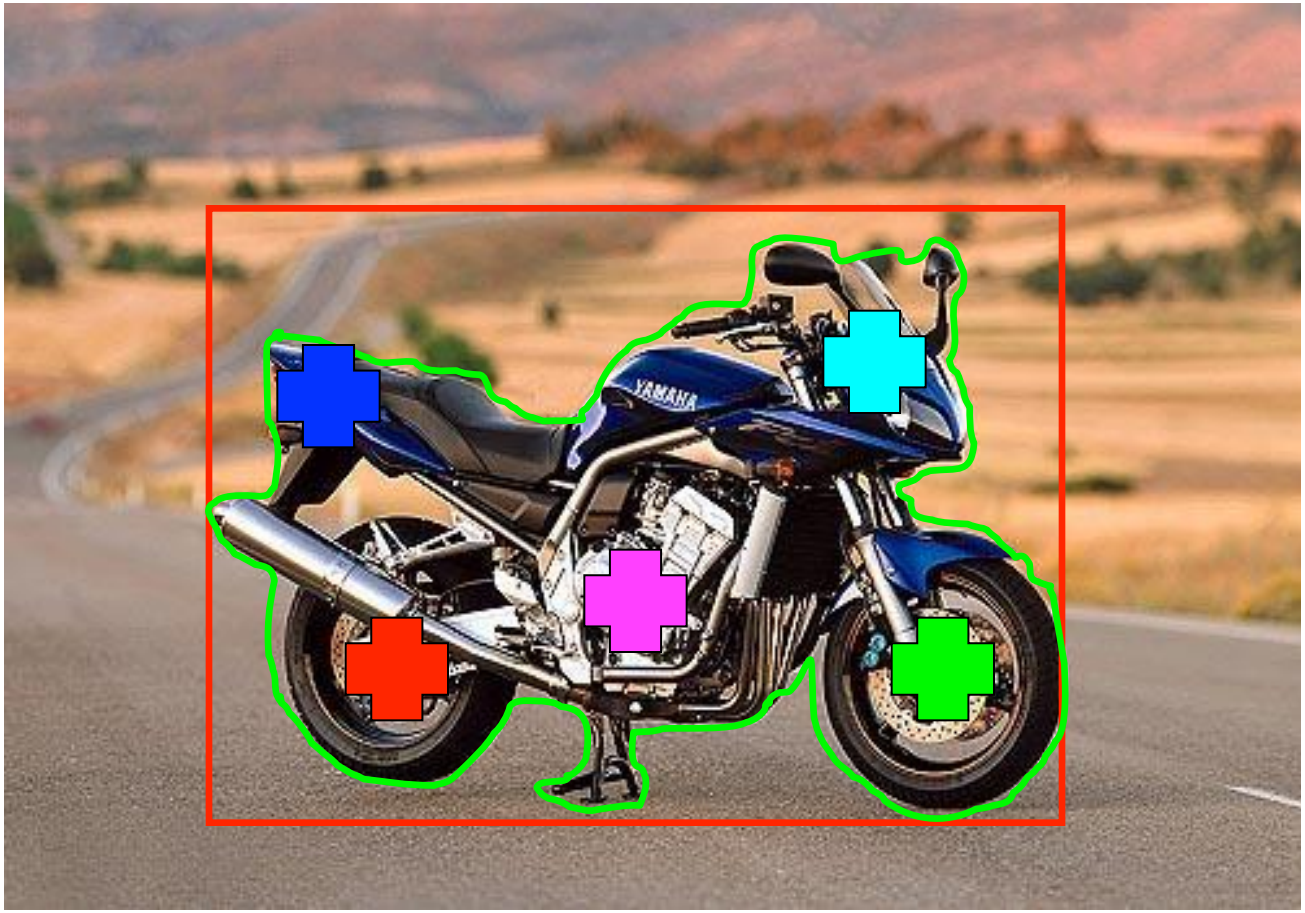
- Find a *linear function* to separate the classes:

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

Recognition task and supervision

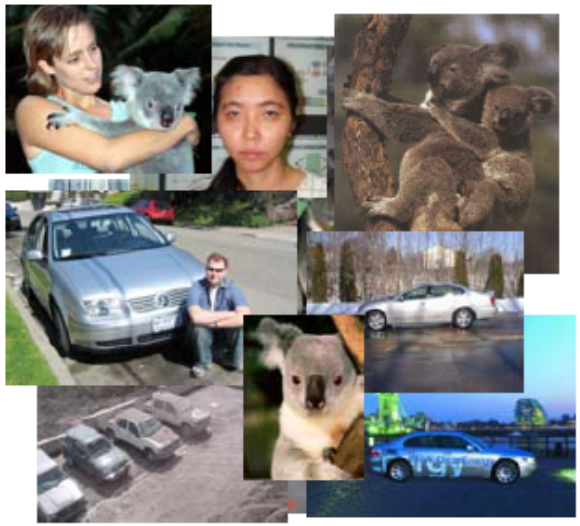
- Images in the training set must be annotated with the "correct answer" that the model is expected to produce

Contains a motorbike

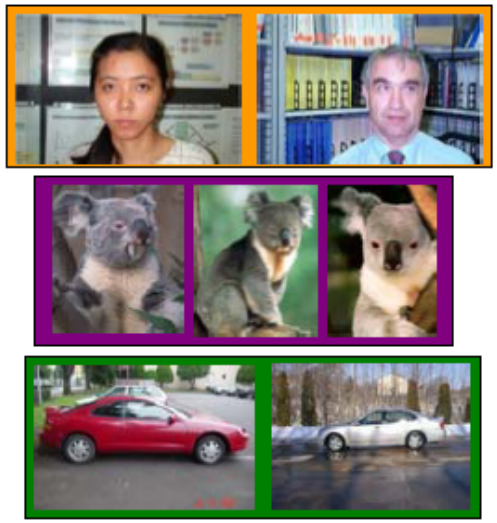


Spectrum of supervision

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

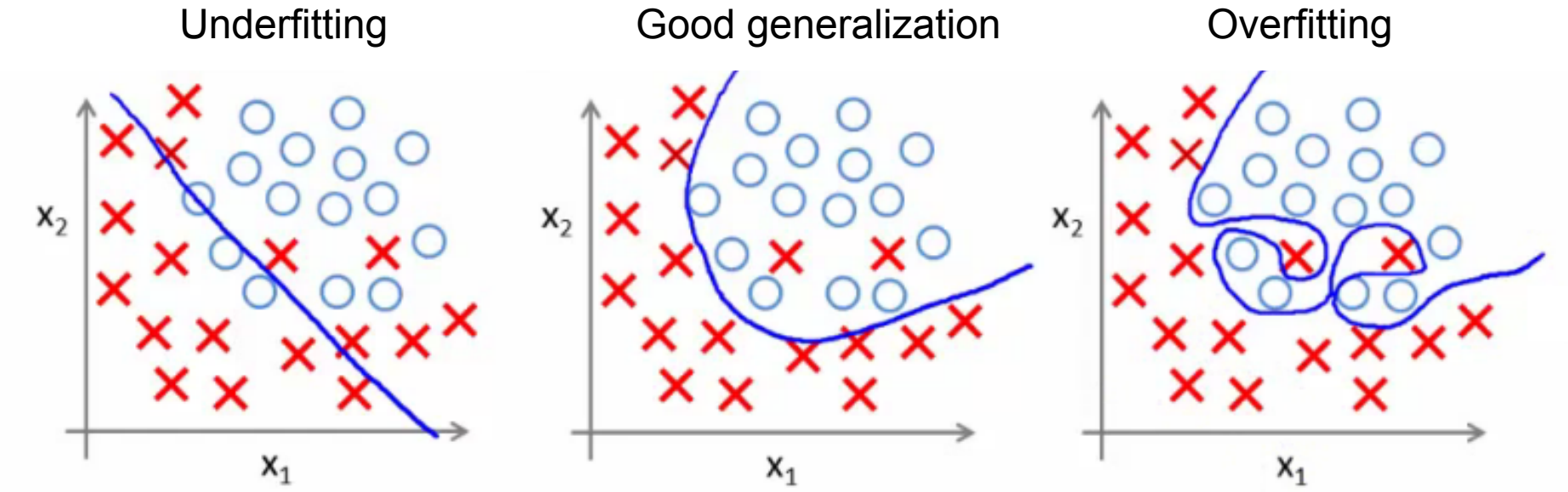
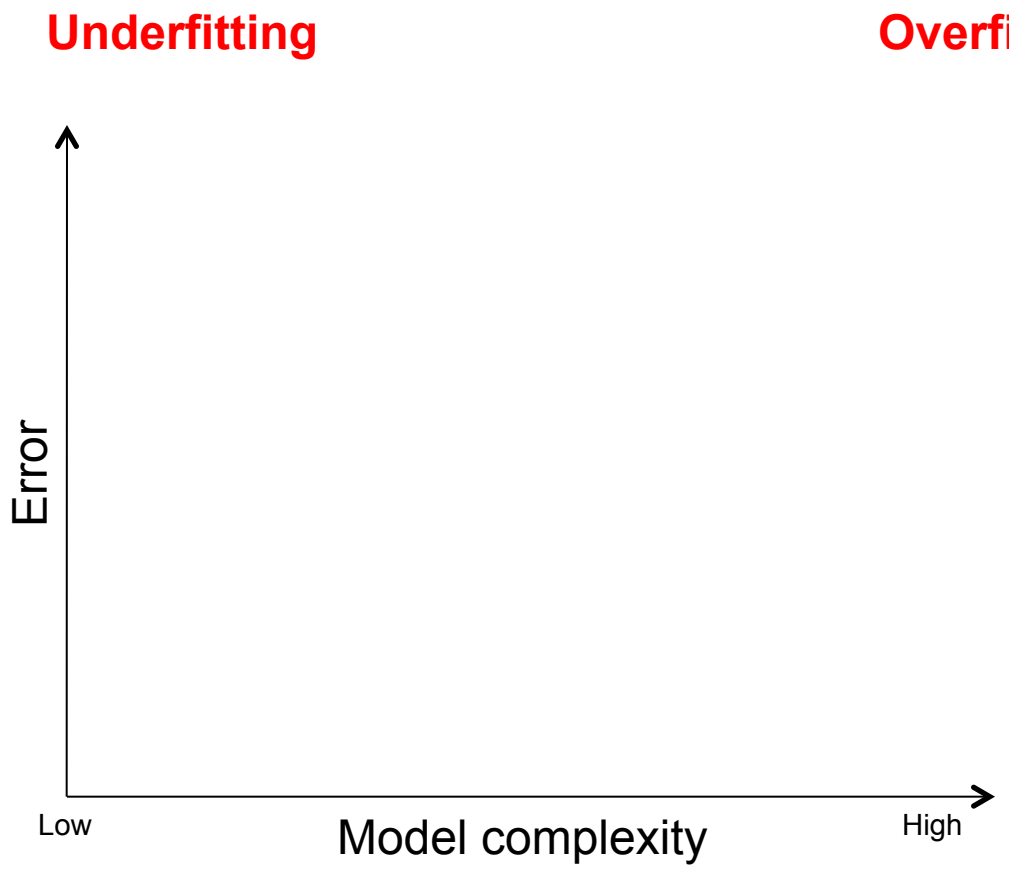


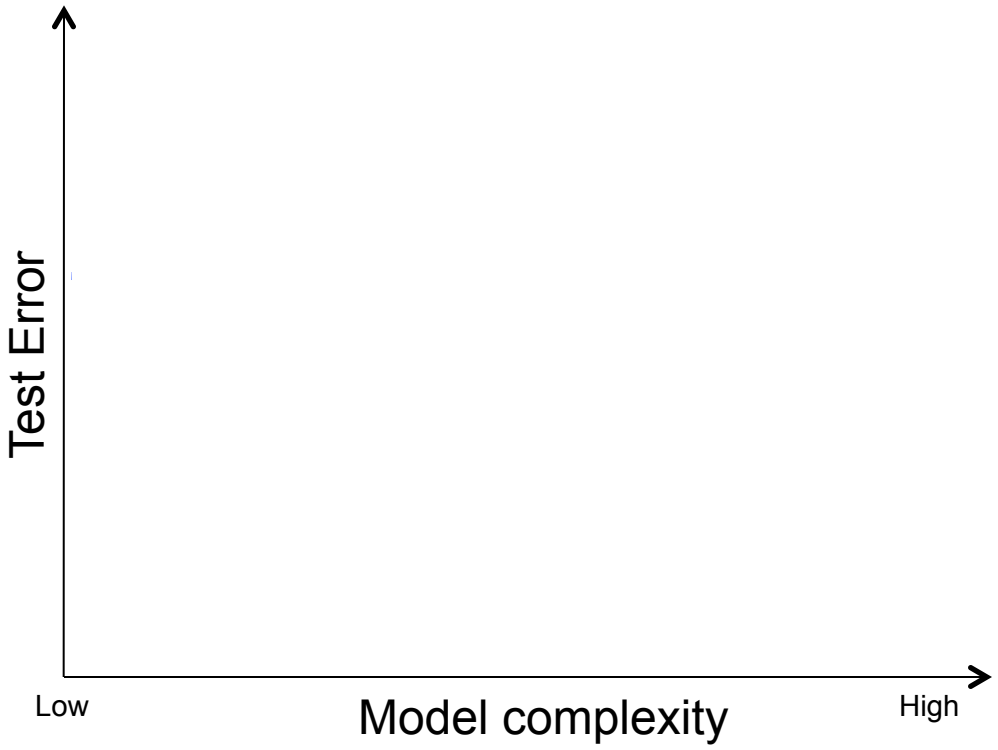
Figure source

Effect of model complexity



Slide credit: D. Hoiem

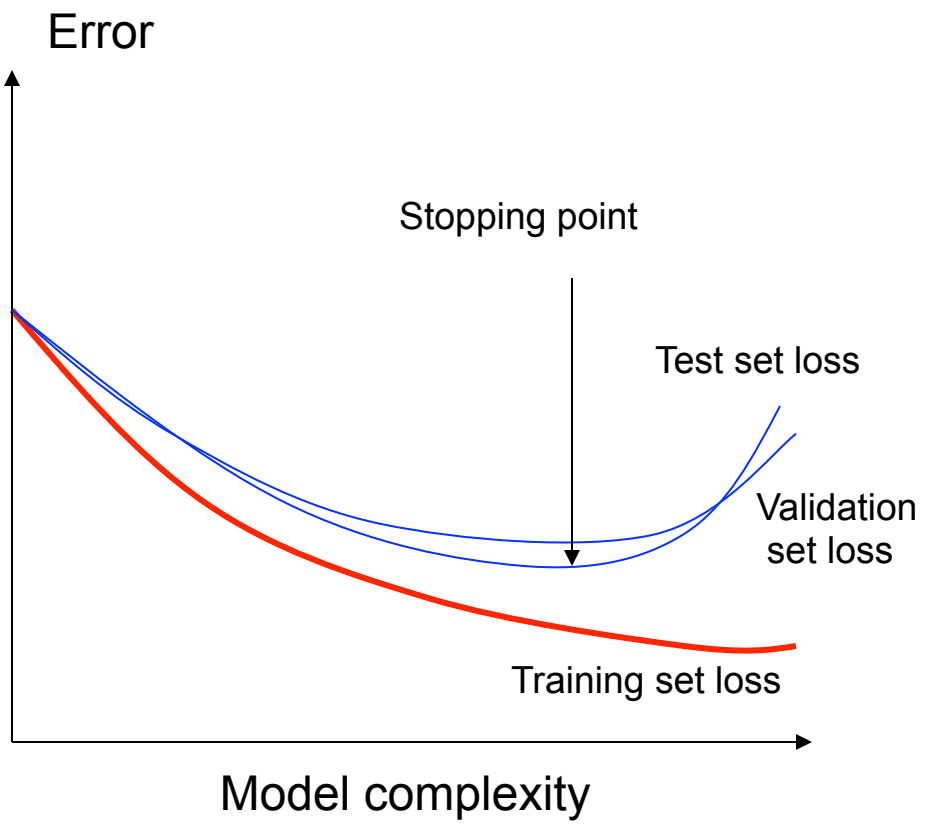
Effect of training set size



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

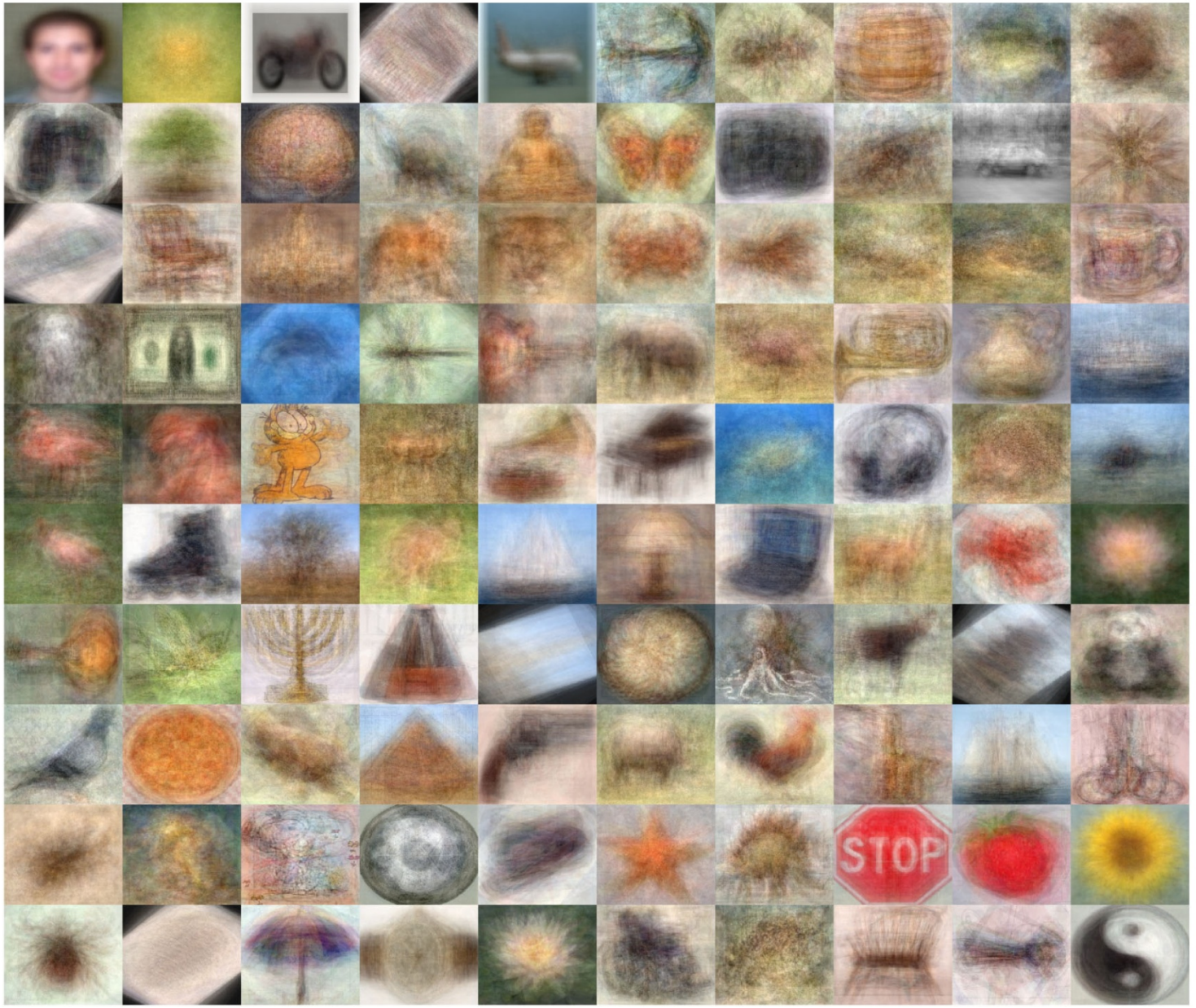
Caltech 101 & 256

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

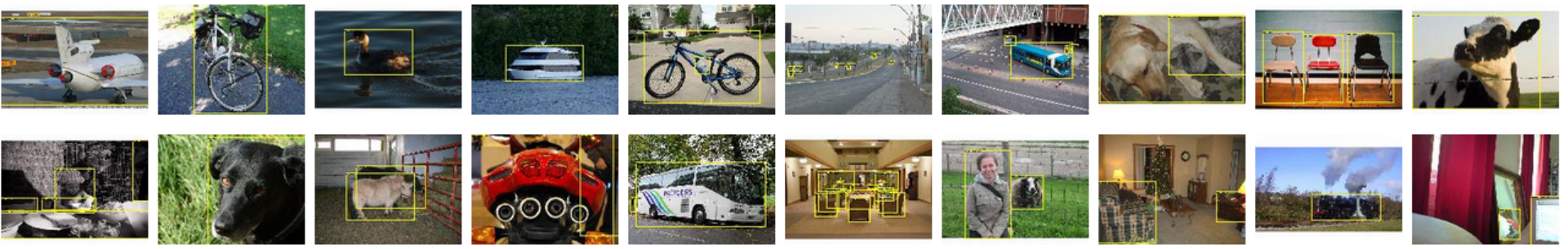


The PASCAL Visual Object Classes Challenge (2005-2012)

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

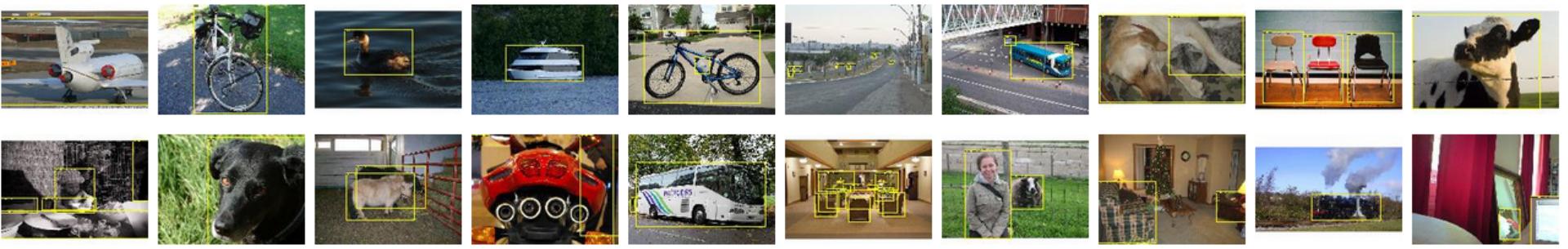
- **Dataset size (by 2012):**
  11.5K training/validation images, 27K bounding boxes, 7K segmentations



PASCAL competitions

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

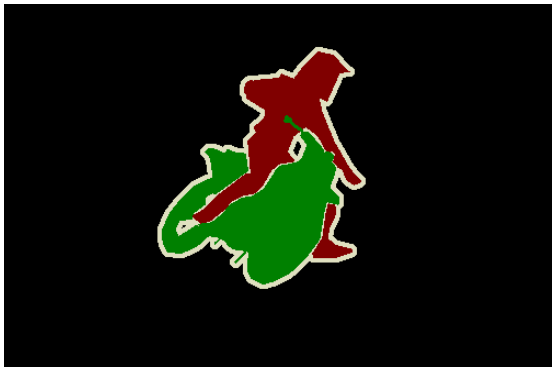
- **Classification:** For each of the twenty classes, predicting presence/absence of an example of that class in the test image
- **Detection:** Predicting the bounding box and label of each object from the twenty target classes in the test image



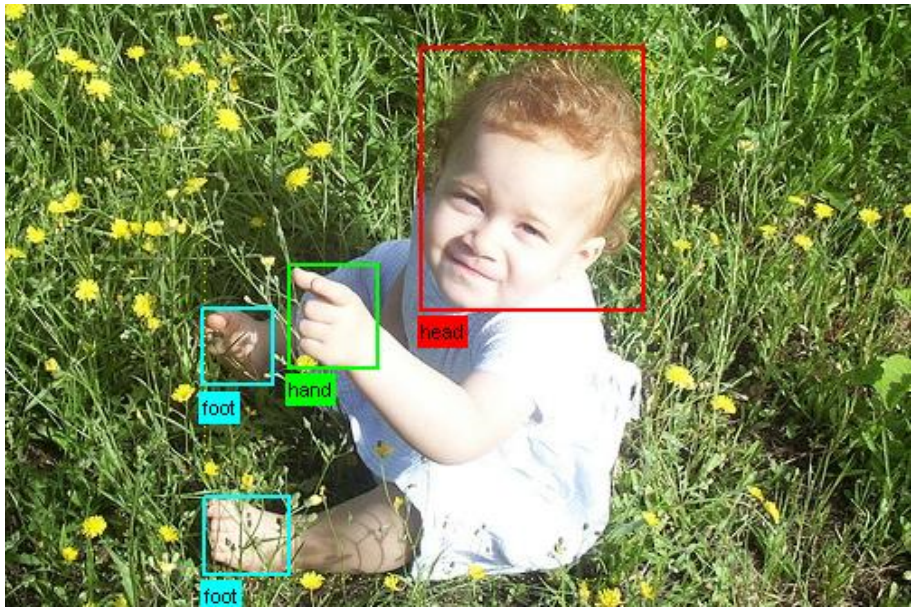
PASCAL competitions

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



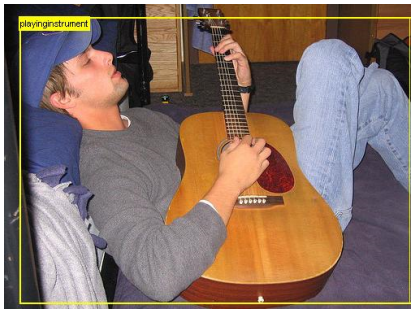
- **Person layout:** Predicting the bounding box and label of each part of a person (head, hands, feet)

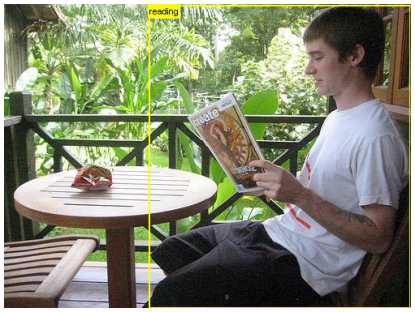


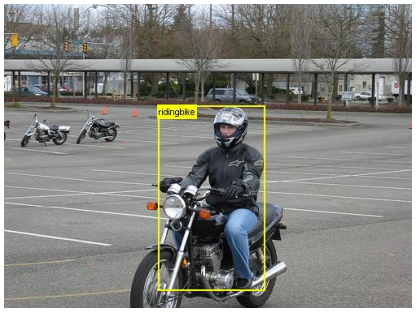
PASCAL competitions

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 Challenge[New] People Publication 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.

SEARCH

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

The ImageNet Challenge 2013 is announced!