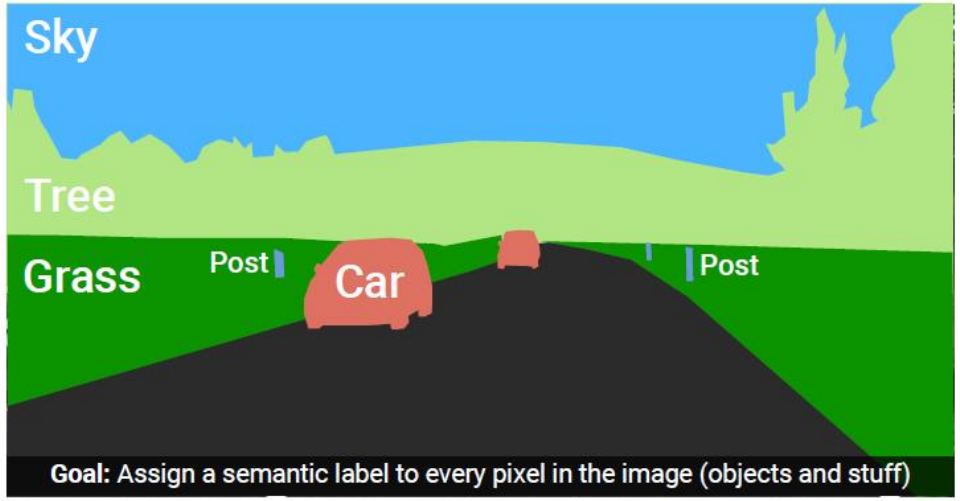
Image Classification

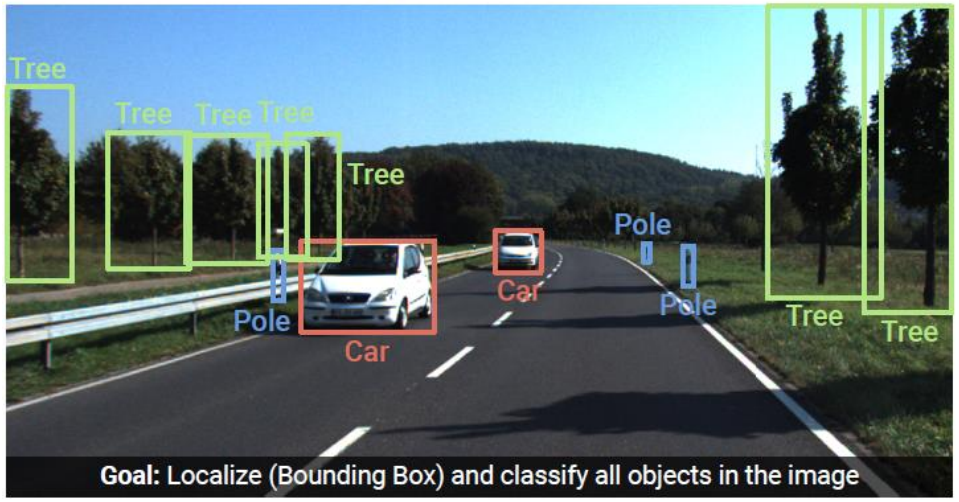
Deep Learning and Image Processing

Image Understanding (Recognition)

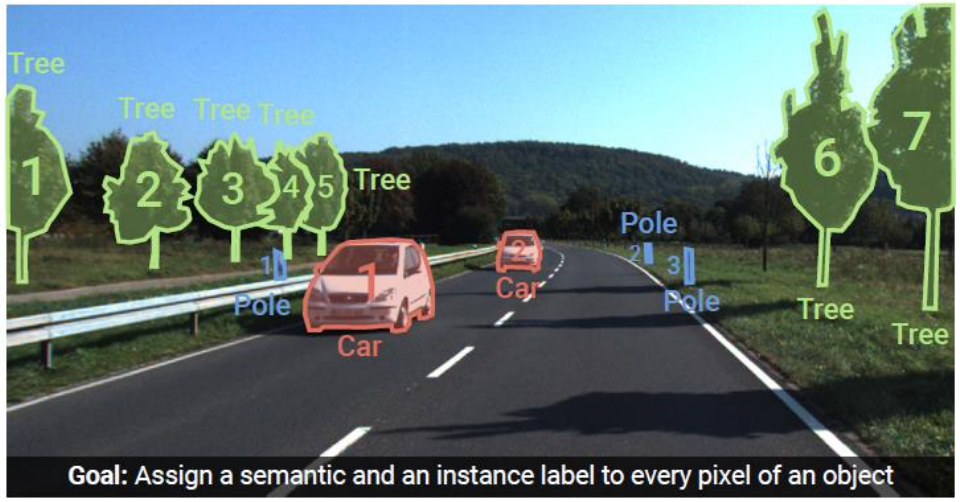
Image Classification



Semantic Segmentation



Object Detection

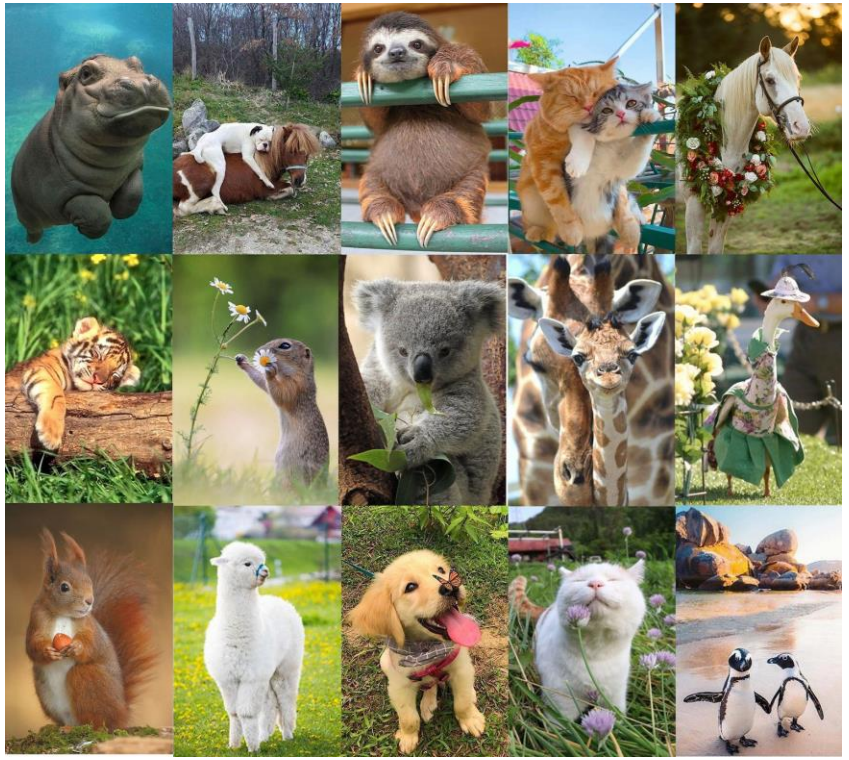


Instance Segmentation

Need for ML

prime example (and also foundation for detection and segmentation):

image classification (whole-image class recognition) according to generic object categories (e.g., cat)



plain keypoint-feature matching only really works for specific instances of a class

→ need to compare with generic objects (e.g., kind of “abstract cat”)

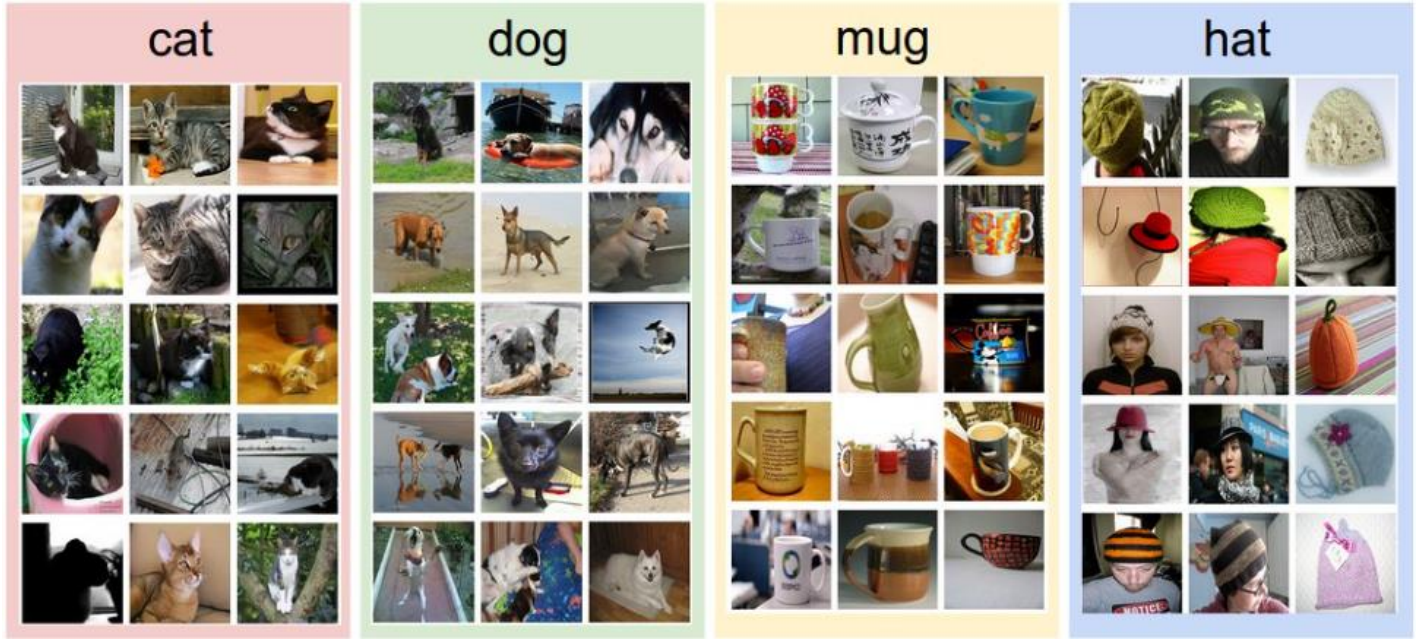
Let’s learn it from data …

What if you haven’t seen this very cat before?

Image Classification

training data set:

labels:



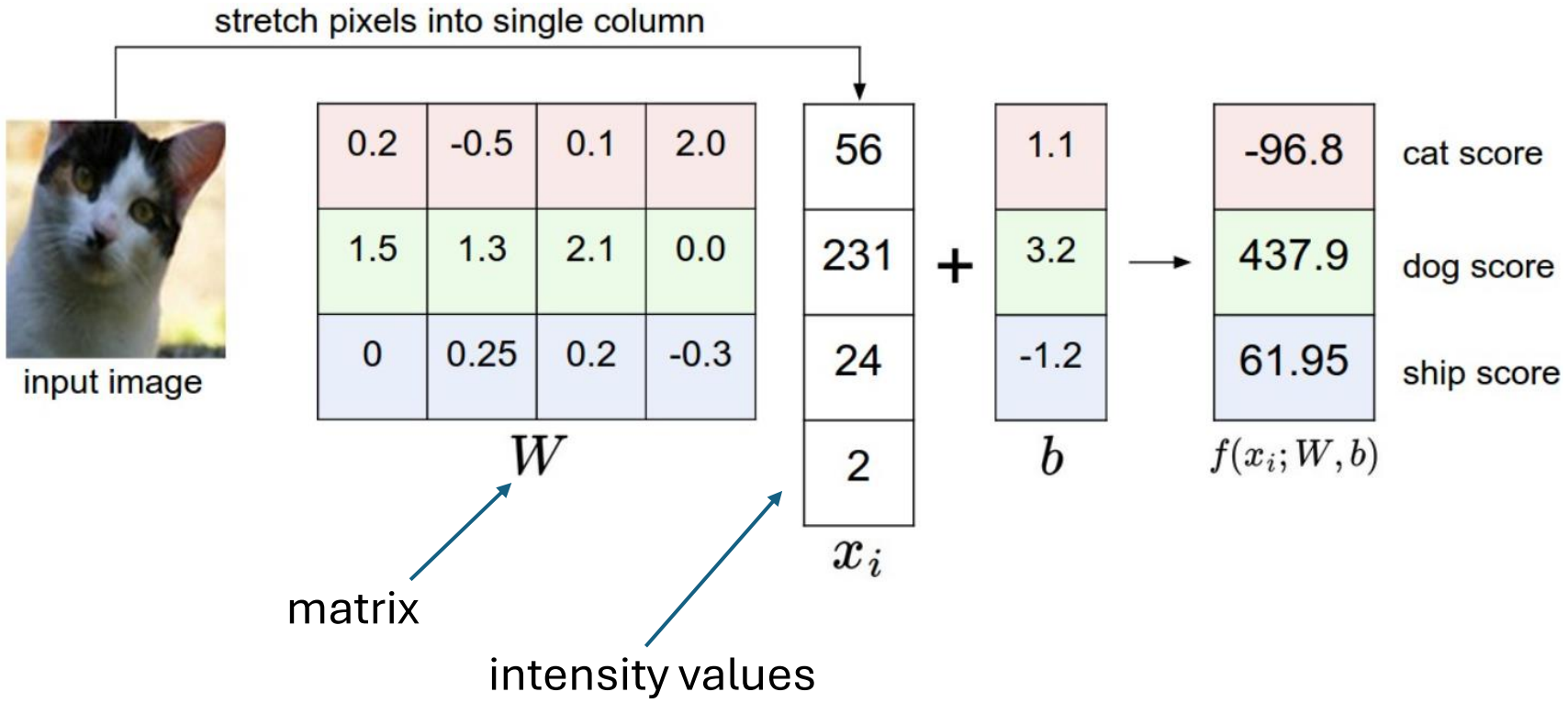
test data with learned classifier:

$f($  $) =$ “Cat”

$f($  $) =$ “Dog”

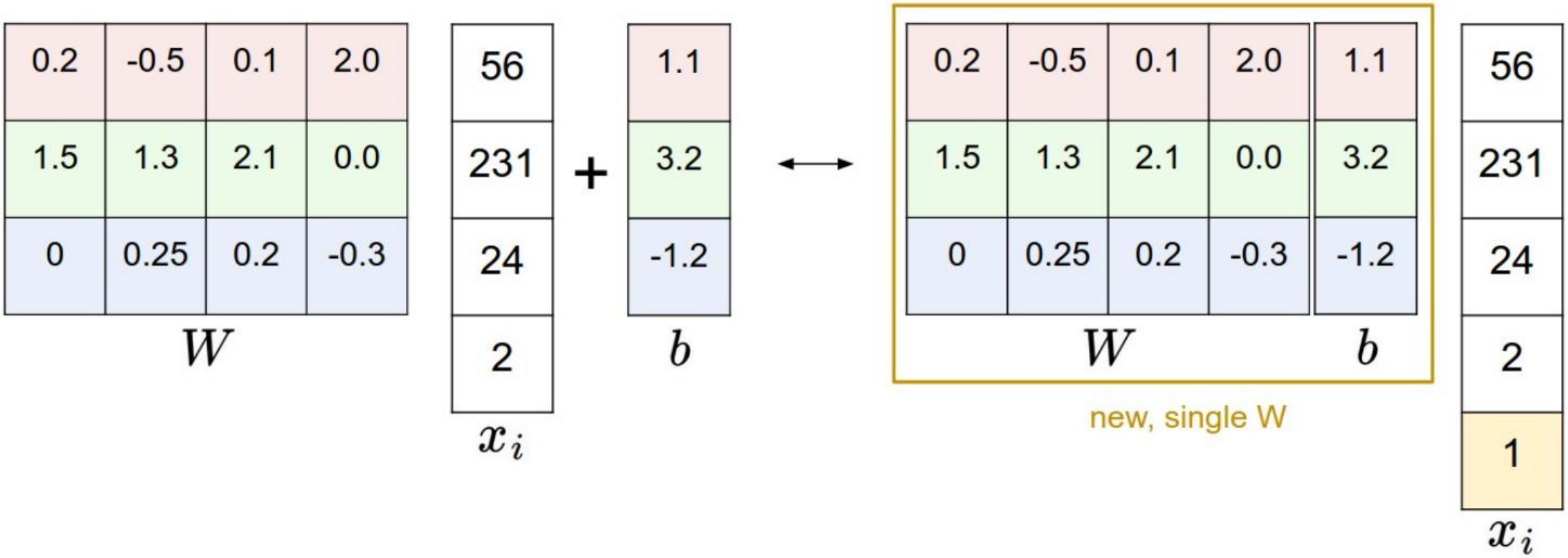
Image Classification with Linear Regression

simplified example: 4-pixel image, 3 classes

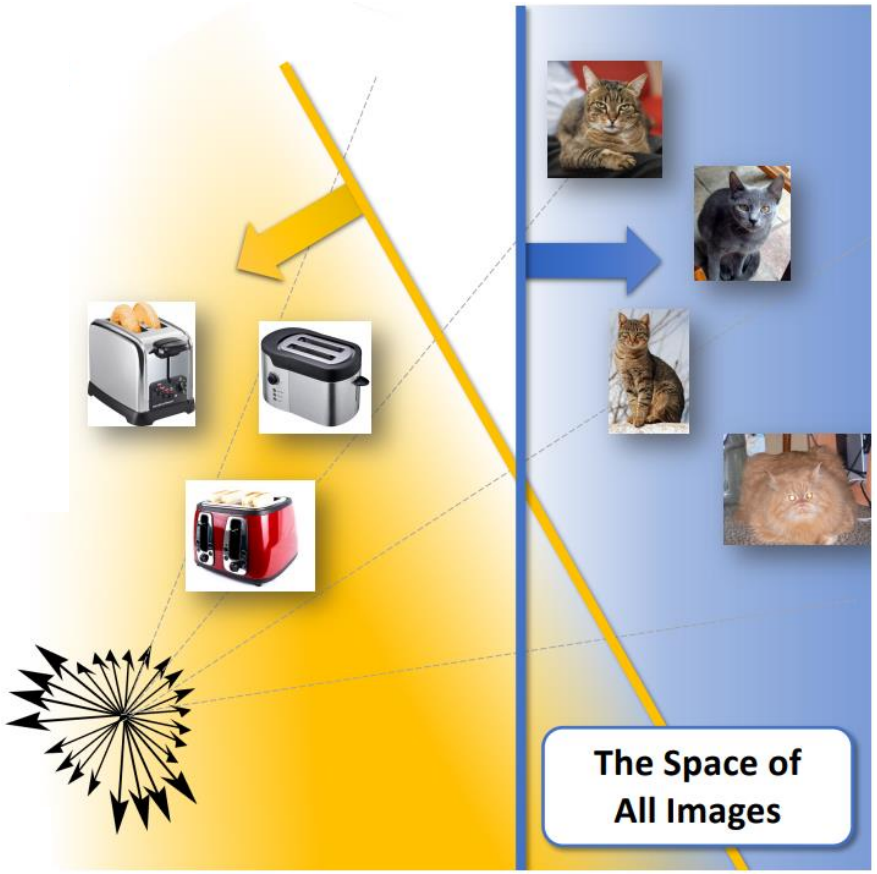


need for one common image size per model

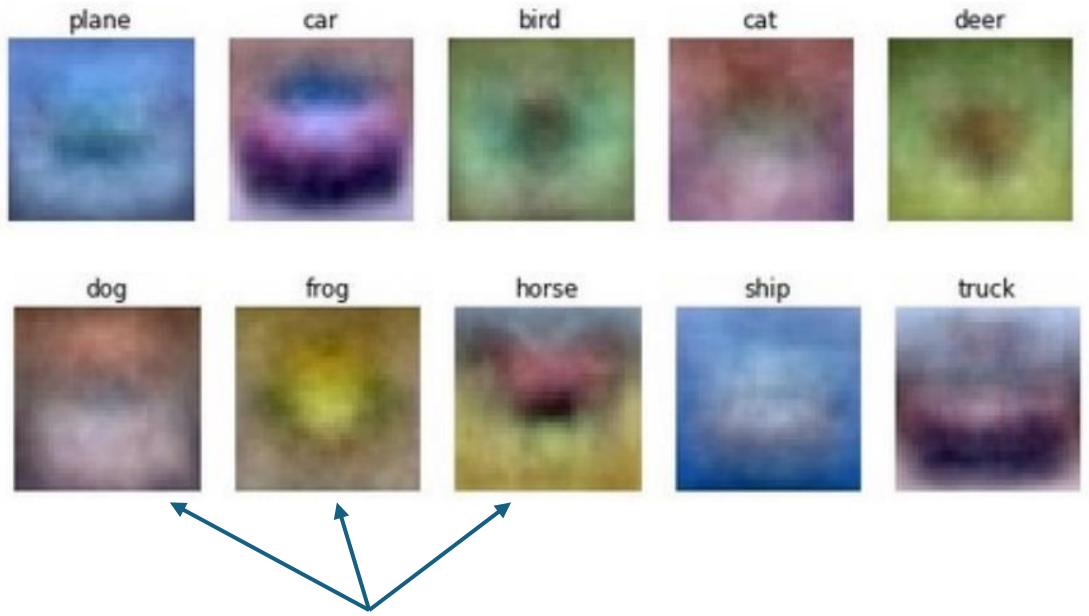
Simplified Matrix Multiplication: Bias Trick



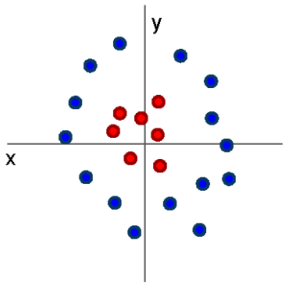
geometric interpretation:
separating hyperplanes



matching interpretation:
generic class templates



different rows in matrix W



Cannot separate red and blue points with linear classifier

raw-pixel images not linearly separable
→ linear model has not enough representational power

Image Classification with kNN

choose an image distance, e.g., L1 distance:

$$d(\boldsymbol{I}_1, \boldsymbol{I}_2) = \sum_p |I_1(p) - I_2(p)|$$

test image

56	32	10	18
90	23	128	133
24	26	178	200
2	0	255	220

-

training image

10	20	24	17
8	10	89	100
12	16	178	170
4	32	233	112

=

pixel-wise absolute value differences

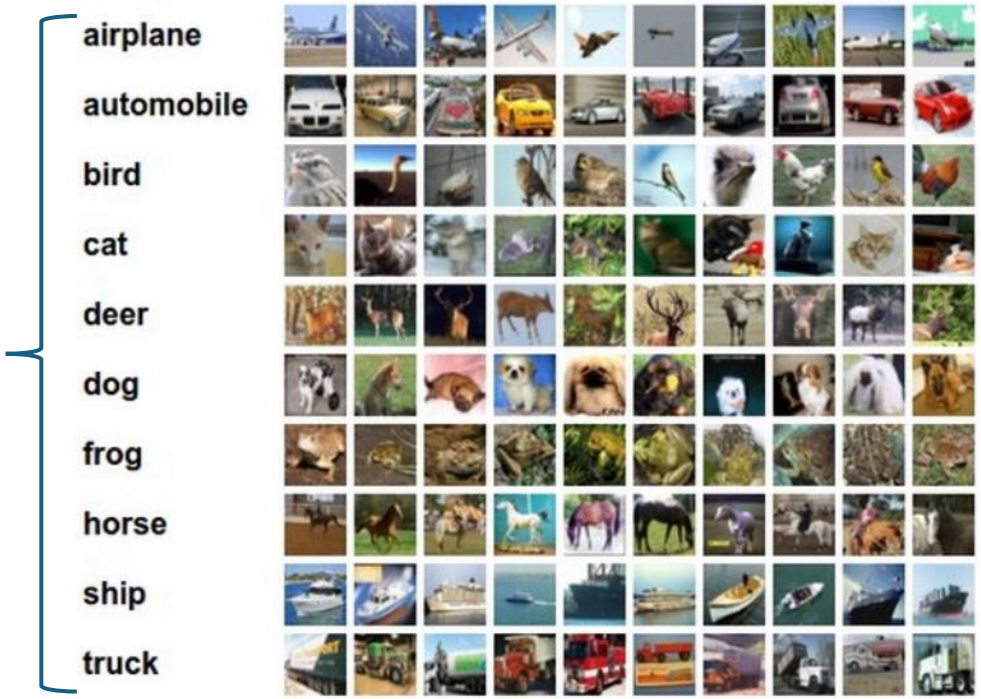
46	12	14	1
82	13	39	33
12	10	0	30
2	32	22	108

→ 456

Image Classification with kNN

training examples CIFAR-10 data set

10 classes



10 nearest neighbors to some test images

better than random guessing, but also not very impressive

Which Features for Image Classification?

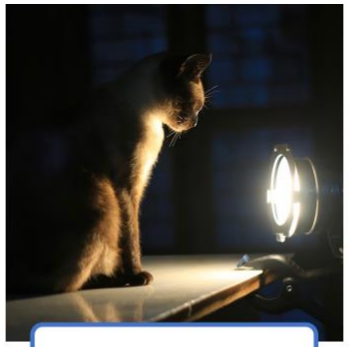
linear regression & kNN (same for tree-based methods):

learning directly from raw pixel intensities does not work great

→ try learning from pre-extracted features, such as HOG or SIFT

challenges:

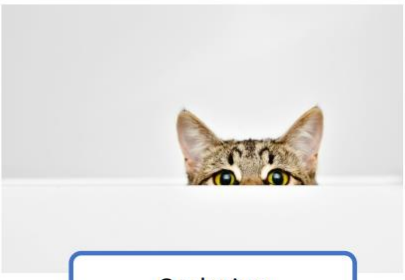
Viewpoint Variation



Lighting Variation

Deformation

Background Clutter

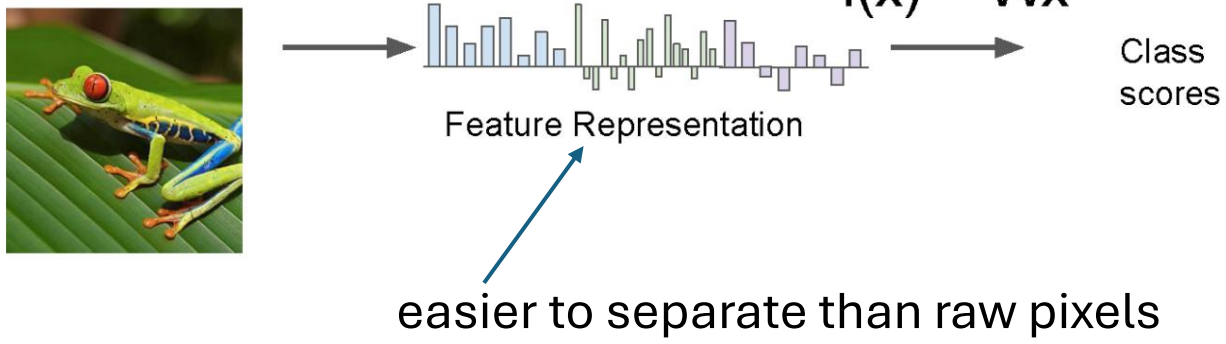


Occlusion

Global Features in Classic ML Method

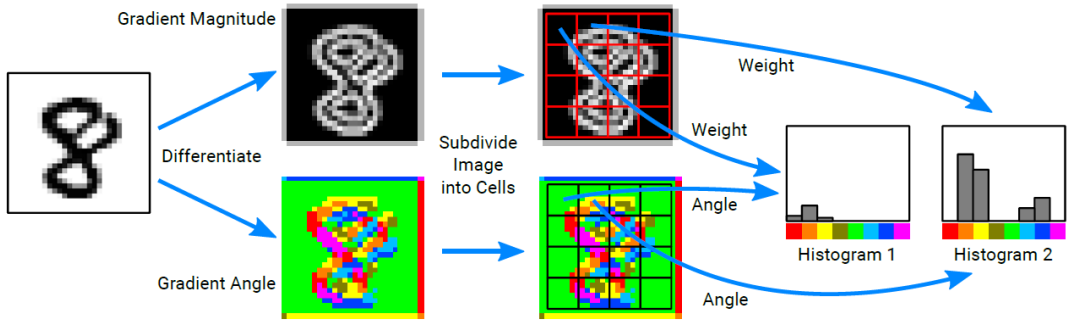
features covering entire image
(instead of only local patches)

or random forest, support-vector machine, ...

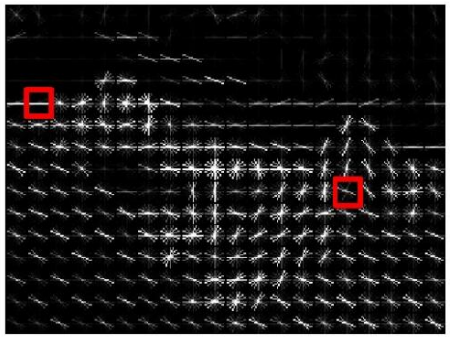


easier to separate than raw pixels

often used: HOG



Divide image into 8x8 pixel regions
Within each region quantize edge direction into 9 bins



Example: 320x240 image gets divided into 40x30 bins; in each bin there are 9 numbers so feature vector has 30*40*9 = 10,800 numbers

Lowe, "Object recognition from local scale-invariant features", ICCV 1999
Dalal and Triggs, "Histograms of oriented gradients for human detection," CVPR 2005

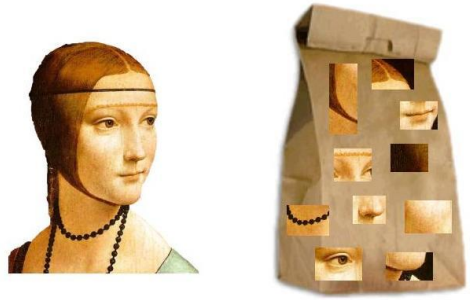
important disadvantage: not translation invariant (due to ordering of patches)

Local Features in Classic ML Method

local features like SIFT do not cover the entire image

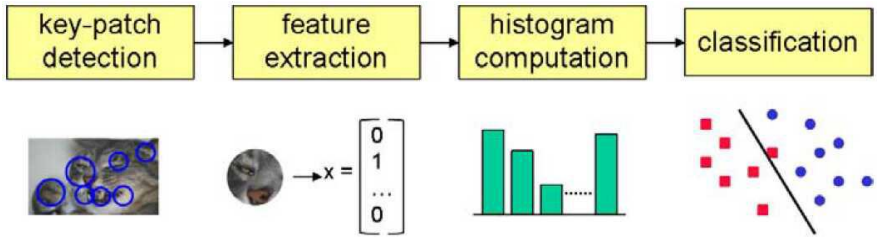
→ need for some processing to identify features in different images (to be able to compare different images with ML)

in HOG this is done by ordering of patches making up the complete image



one popular approach, called bag-of-words model (as it is orderless):

1. learn clustering of SIFT vectors (e.g., with K-means)
2. quantization of different clusters into visual "vocabulary" (see embeddings in language models)
3. create (sparse) histogram of visual "word" occurrences
4. train ML model (e.g., random forest) with histogram bins as features



important disadvantage: ignores spatial relationships among different patches

Need for Feature Learning

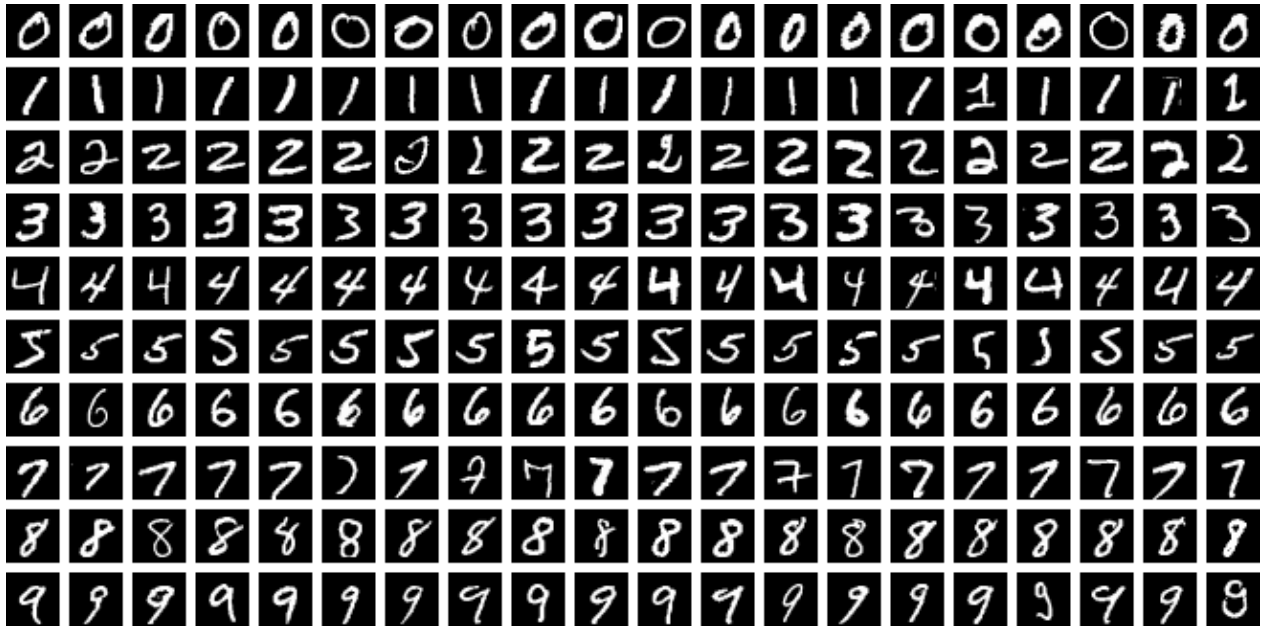
many hand-designed components in approach using pre-extracted features → poor generalization

Is there maybe some way after all to learn features end-to-end from raw pixel intensities?

Image Data Sets

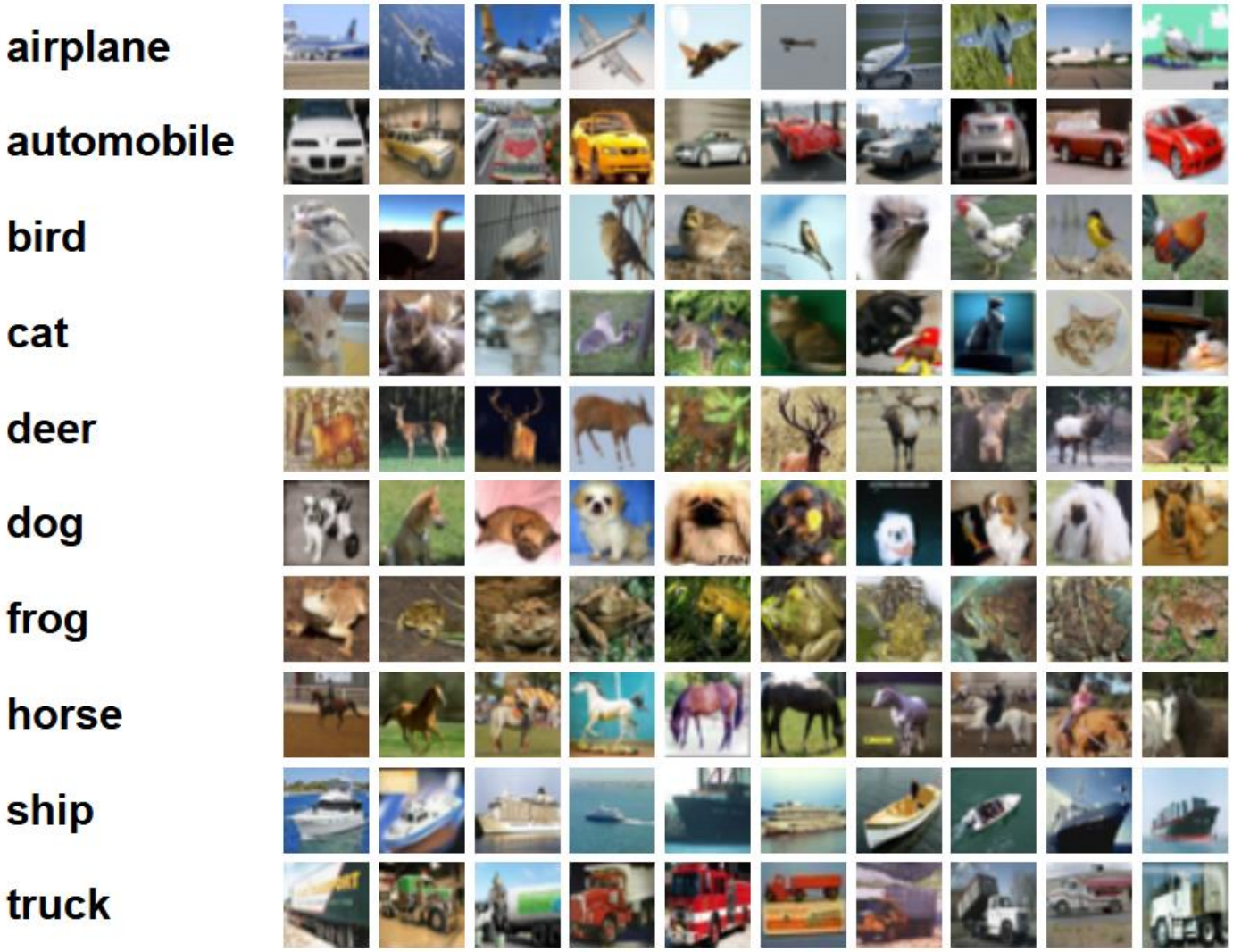
MNIST

- Modified National Institute of Standards and Technology
- handwritten digits (10 classes)
- black and white images
- 28 × 28 pixels
- 60k training and 10k test images



CIFAR-10

- Canadian Institute for Advanced Research
- 10 different labeled object classes
- color images
- 32 × 32 pixels
- 50k training and 10k test images



ImageNet

- more than 14 million color images with varying sizes
- more than 20k labeled categories