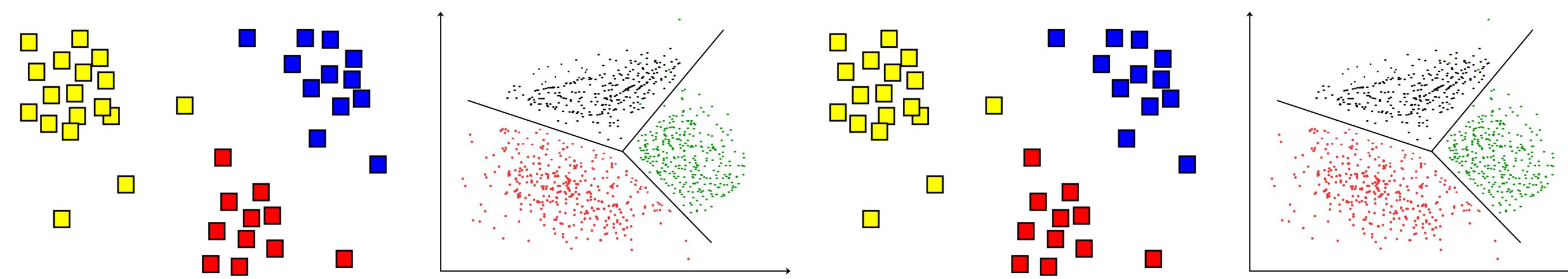


CPSC 425: Computer Vision



Lecture 17: Visual Classification 1, Bag of Words

Menu for Today

Topics:

- **Visual Classification** introduction
- **Instance vs Category** Recognition
- **Bag of Words, K-means**
- **Quiz 5**

Readings:

- **Today's Lecture:** Szeliski 11.4, 12.3-12.4, 9.3, 5.1-5.2

Reminders:

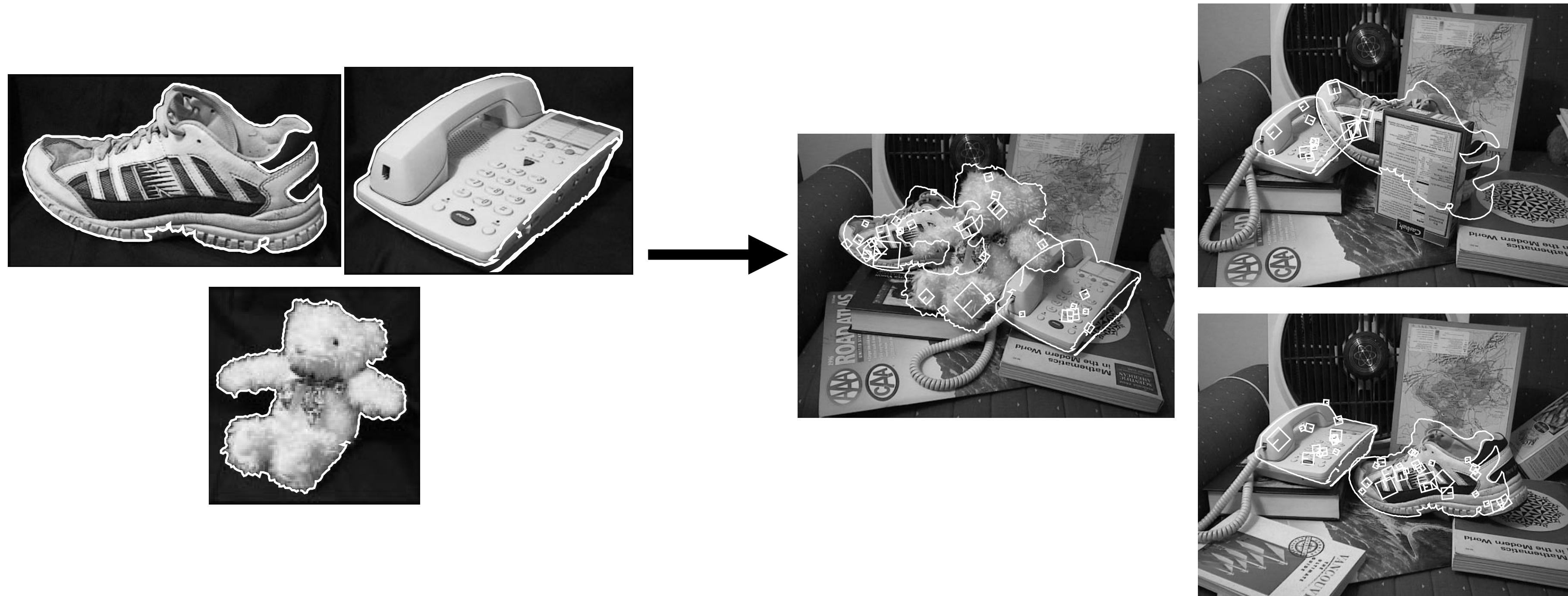
- **Assignment 4:** due today
- **Assignment 5:** Scene Recognition with Bag of Words is now available
- **No lectures next week** (mid term break)

Visual Classification I

- Object recognition introduction: instance, category
- Image classification vs object detection, segmentation
- Visual Words and Bag of Words Classifiers

Object Recognition

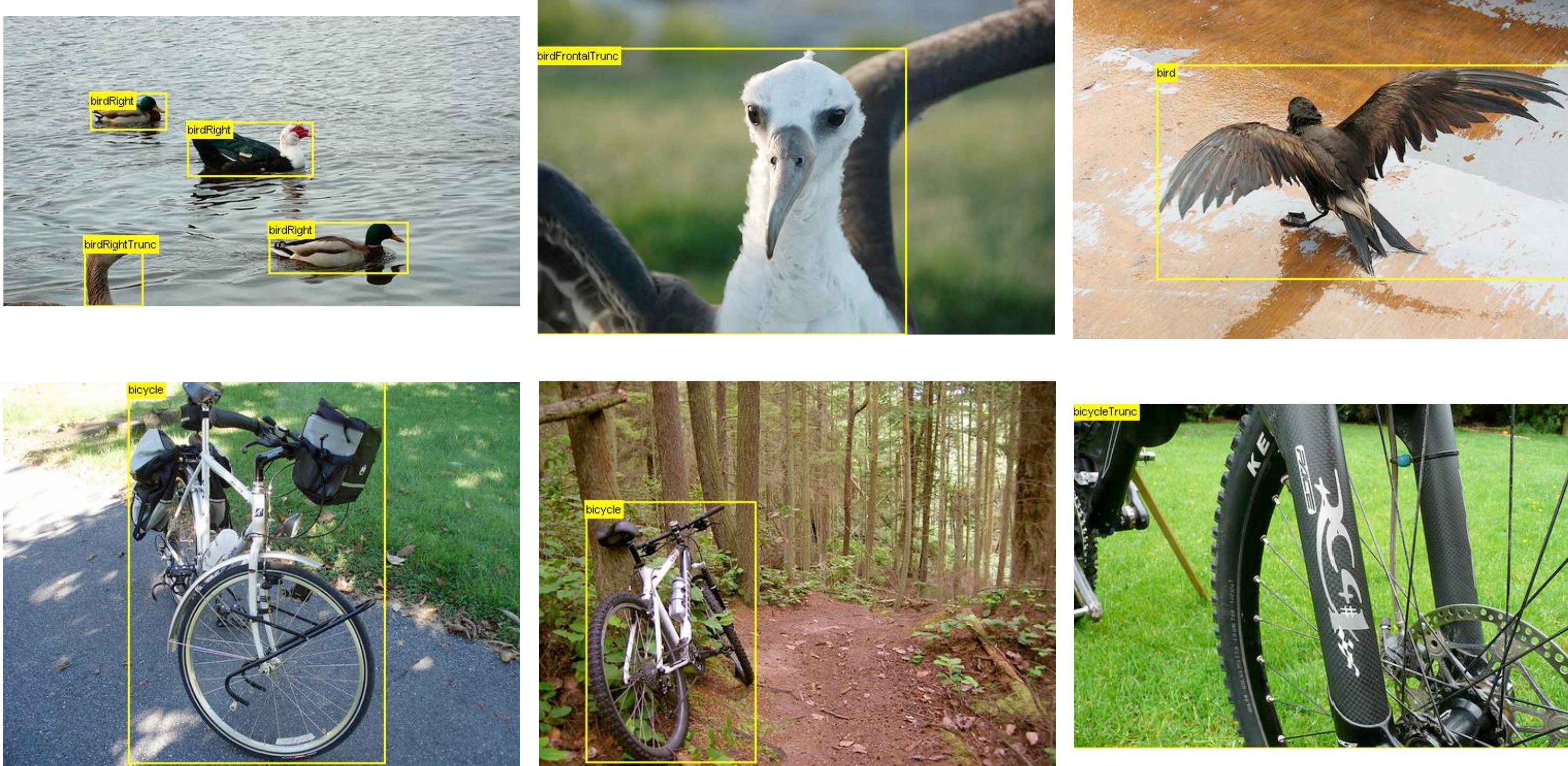
- Object recognition with SIFT features [Lowe 1999]



What is present? Where? What orientation?

Object Recognition

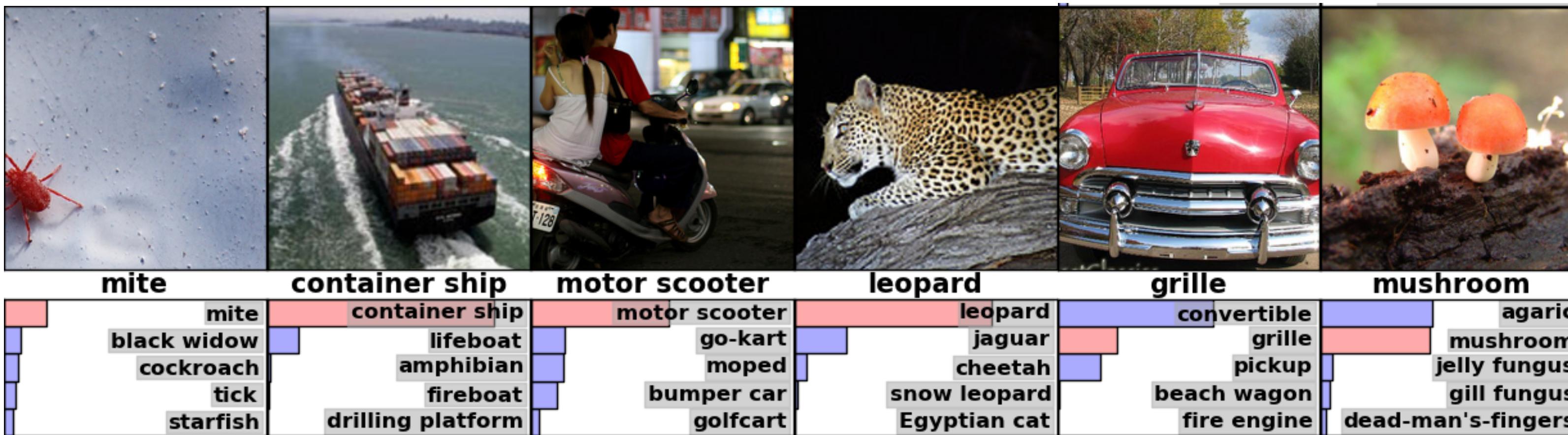
- PASCAL Visual Object Classes Challenges [2005-2012]



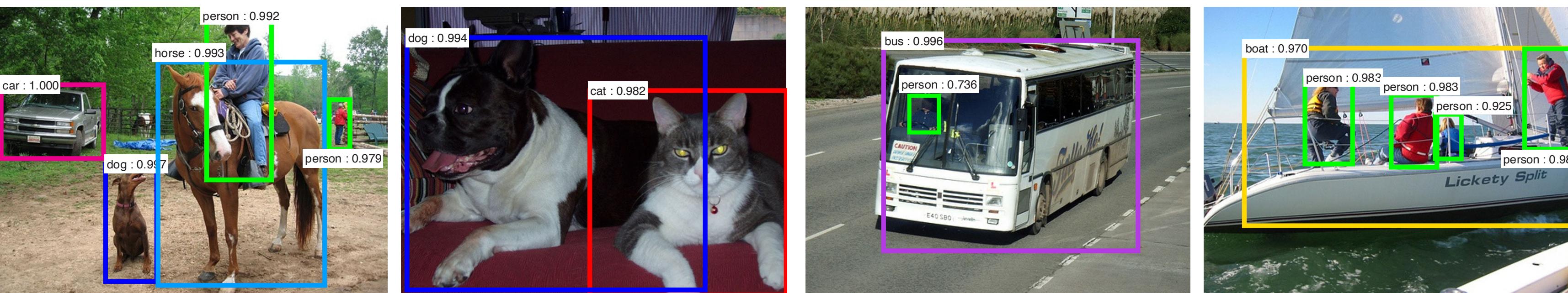
What is present? Where? What orientation?

Classification and Detection

- Classification: Label per image, e.g., ImageNet

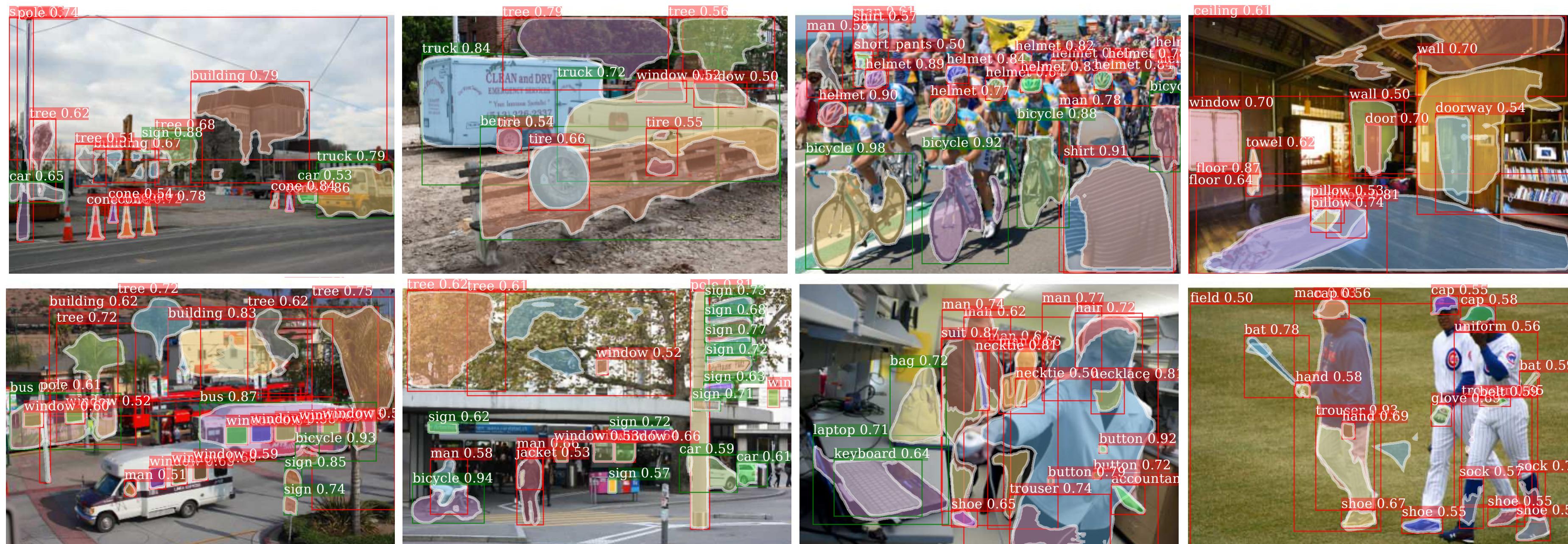


- Detection: Label per region, e.g., PASCAL VOC



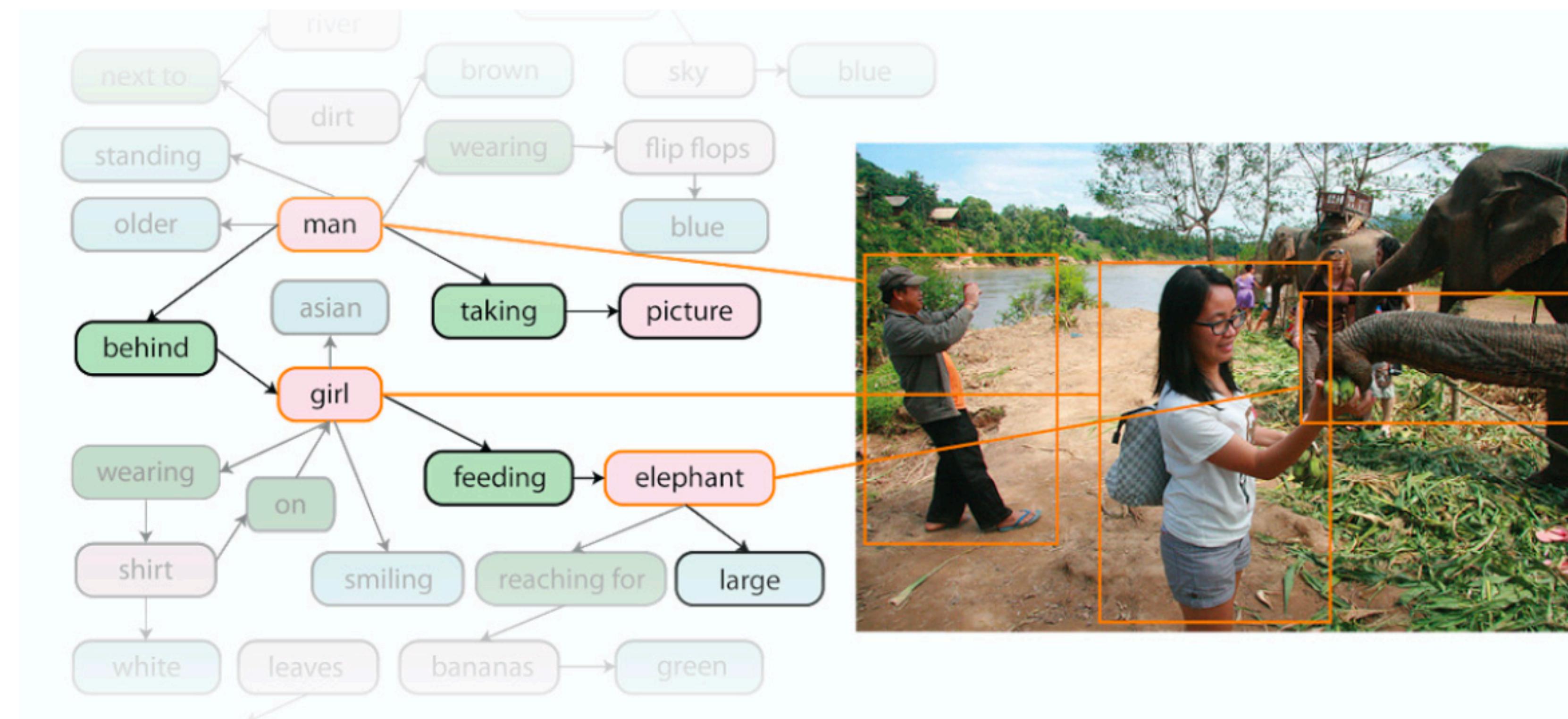
Segmentation

- Segmentation: Label per pixel, e.g., MS COCO



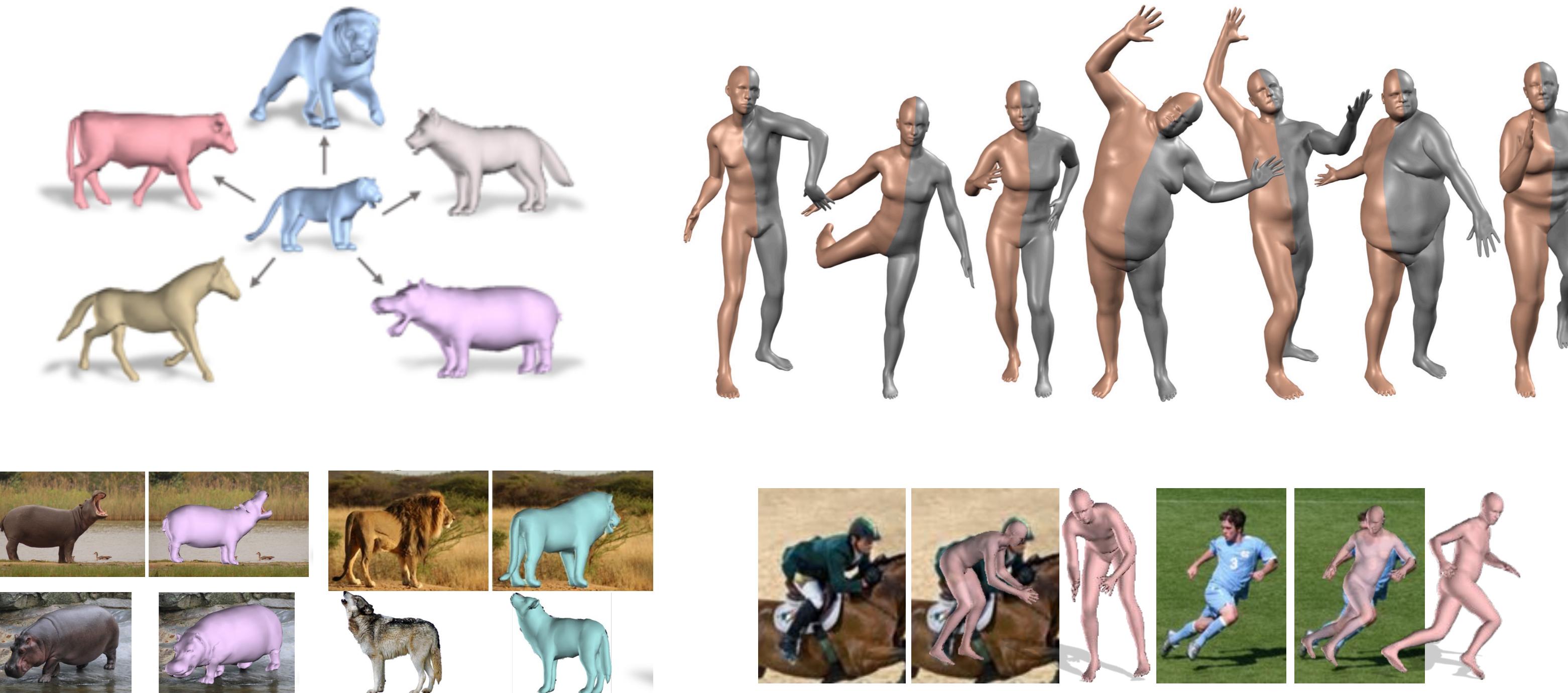
Structured Image Understanding

- “Girl feeding large elephant”
- “A man taking a picture behind girl”



Shape + Tracking

- Other vision applications might need shape modelling (possibly deformable) and/or tracking in video

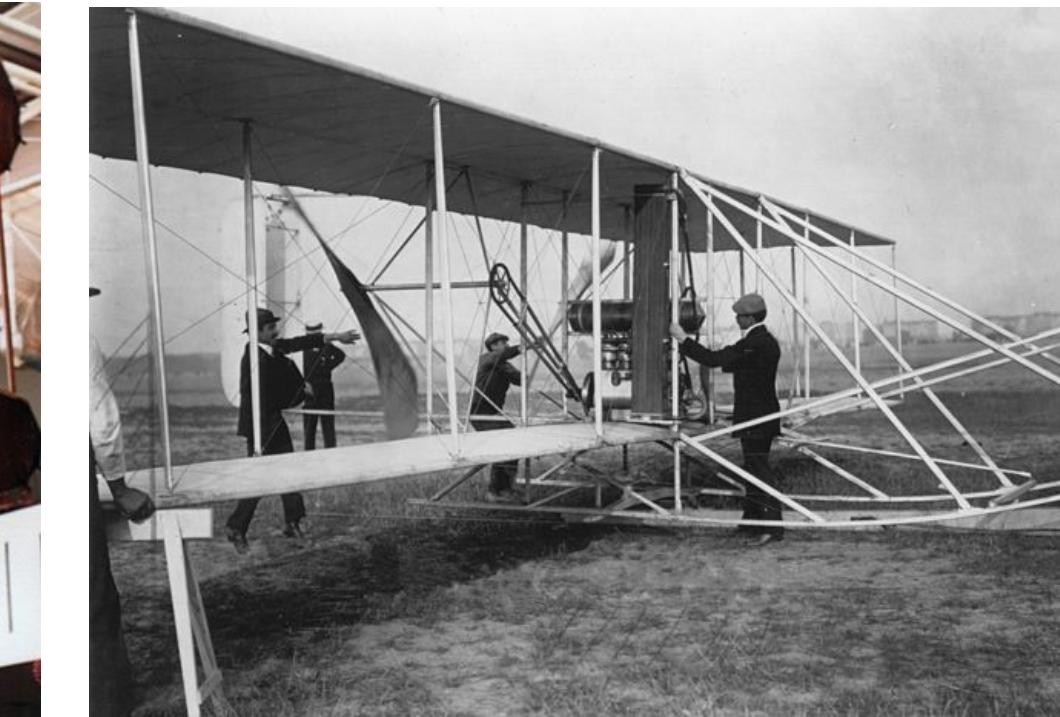


[Zuffi et al 2017]

[SMPL Loper et al 2015]

We'll focus on single image classification today

Classification: Instance vs Category

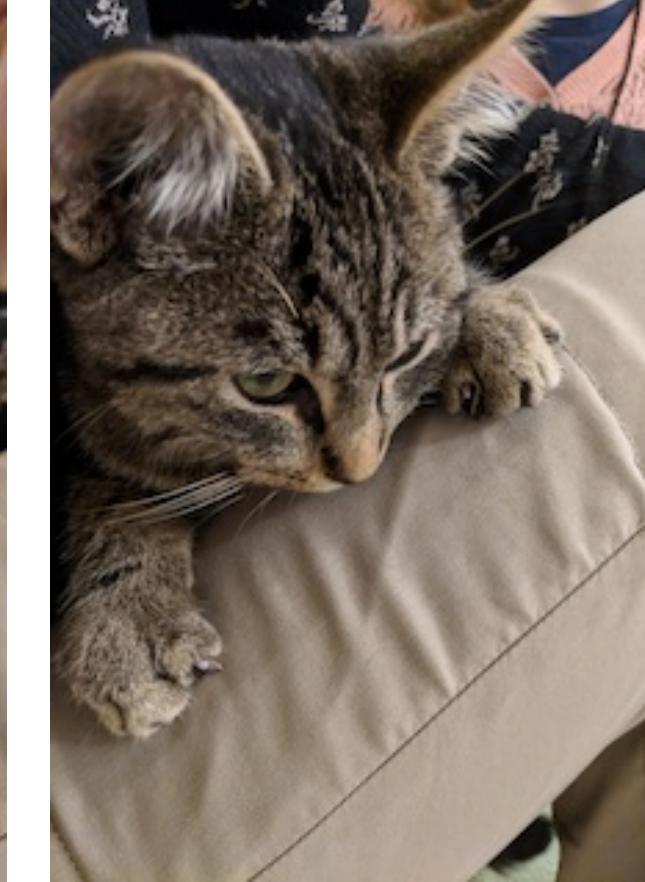
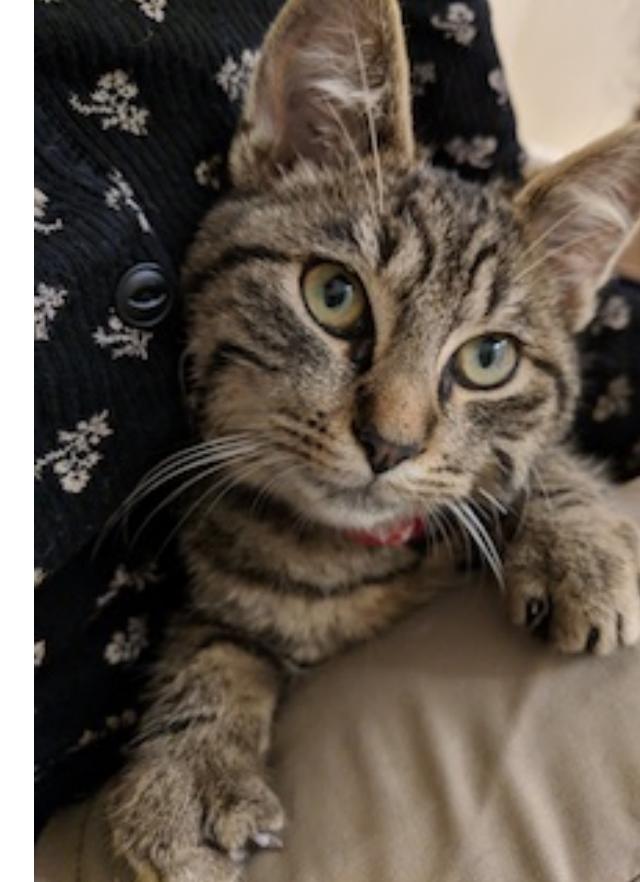


Instance of Aeroplane (Wright Flyer)

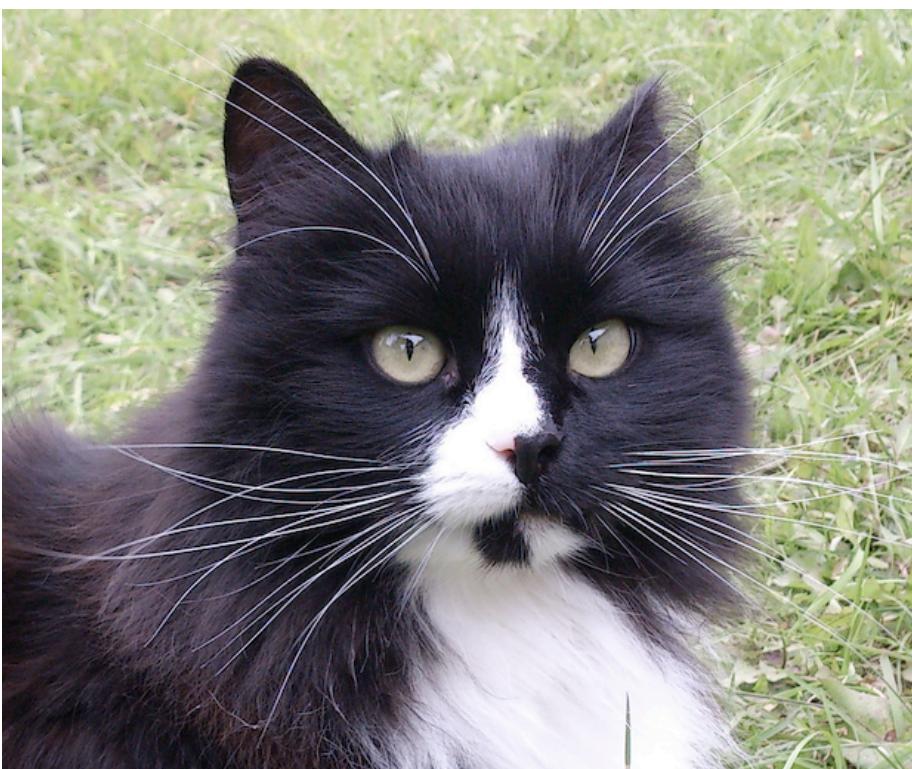


Category of Aeroplanes

Classification: Instance vs Category



Instance of a cat



Category of domestic cats

Taxonomy of Cats

- ↳ Mammals (Class Mammalia)
- ↳ Therians (Subclass Theria)
- ↳ Placental Mammals (Infraclass Placentalia)
- ↳ Ungulates, Carnivorans, and Allies (Superorder Laurasiatheria)
- ↳ Carnivorans (Order Carnivora)
- ↳ Felines (Family Felidae)
- ↳ Small Cats (Subfamily Felinae)
- ↳ Genus *Felis*
 - ↳ Chinese Mountain Cat (*Felis bieti*)
 - ↳ Domestic Cat (*Felis catus*)
 - ↳ Jungle Cat (*Felis chaus*)
 - ↳ African Wildcat (*Felis lybica*)
 - ↳ Sand Cat (*Felis margarita*)
 - ↳ Black-footed Cat (*Felis nigripes*)
 - ↳ European Wildcat (*Felis silvestris*)

Bengal Tiger
[Omveer Choudhary]



Ocelot
[Jitze Couperus]



European Wildcat
[the wasp factory]





vehicle



craft



watercraft



sailing vessel



sailboat



trimaran 13

WordNet

- We can use language to organise visual categories
- This is the approach taken in ImageNet [Deng et al 2009], which uses the WordNet lexical database [wordnet.princeton.edu]
- As in language, visual categories have complex relationships
- e.g., a “sail” is part of a “sailboat” which is a “watercraft”

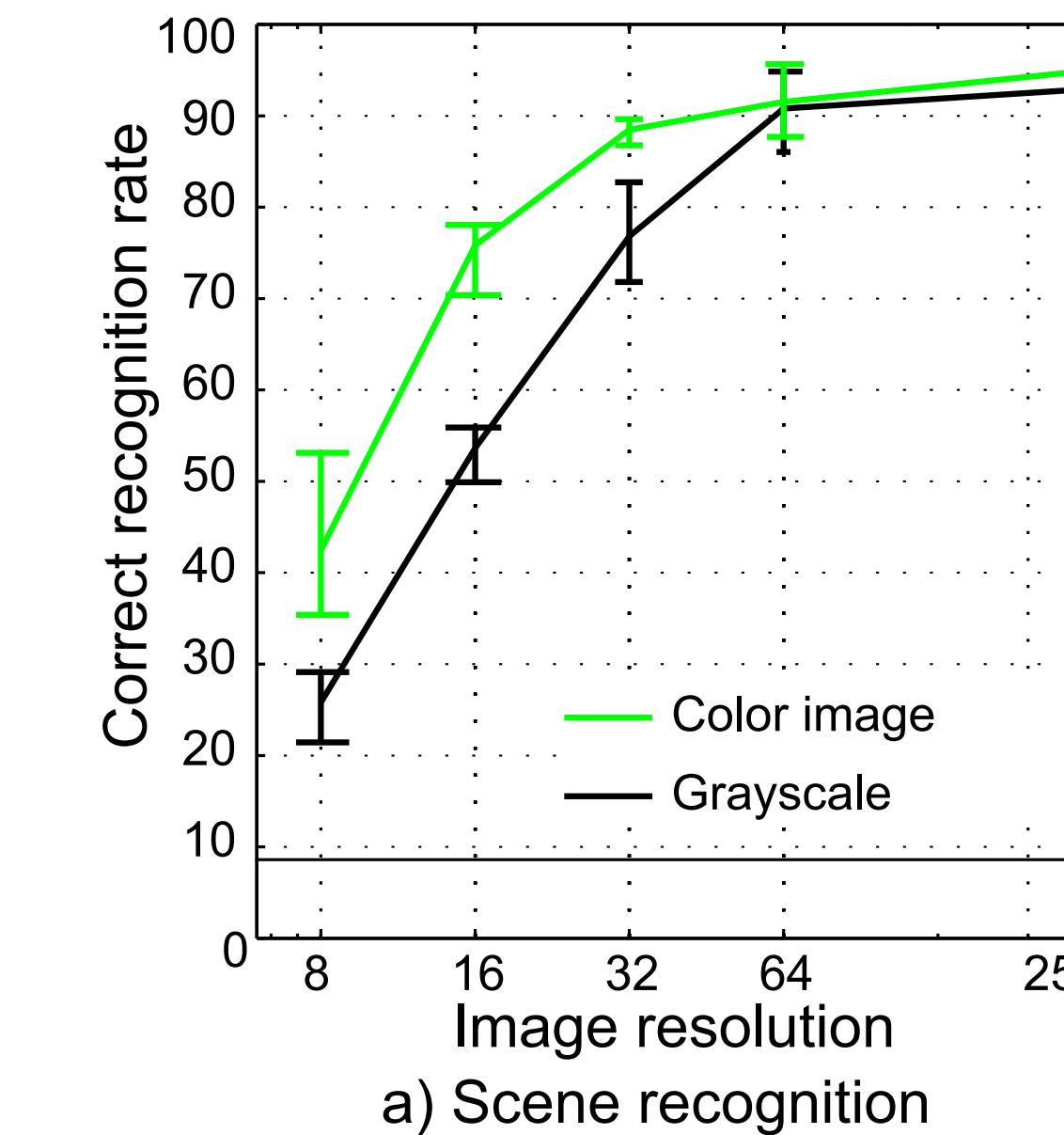
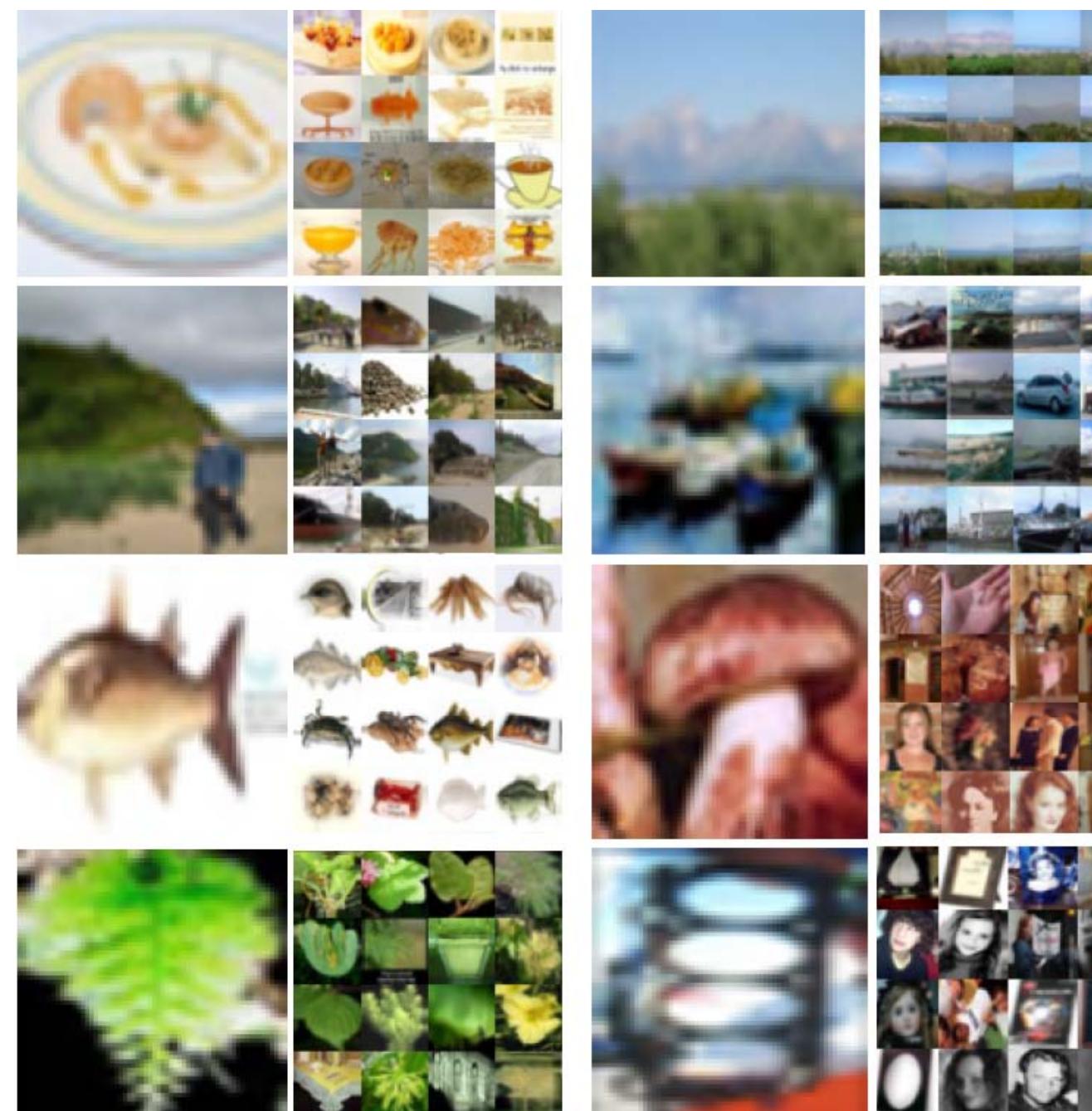
- S: (n) **sailboat**, [sailing boat](#) (a small sailing vessel; usually with a single mast)
 - [direct hyponym](#) / [full hyponym](#)
 - S: (n) [catboat](#) (a sailboat with a single mast set far forward)
 - S: (n) [sharpie](#) (a shallow-draft sailboat with a sharp prow, flat bottom, and triangular sail; formerly used along the northern Atlantic coast of the United States)
 - S: (n) [trimaran](#) (a fast sailboat with 3 parallel hulls)
 - [part meronym](#)
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - S: (n) [sailing vessel](#), [sailing ship](#) (a vessel that is powered by the wind; often having several masts)



If we call a “sailboat” a watercraft, is this wrong? What if we call it a “sail”?

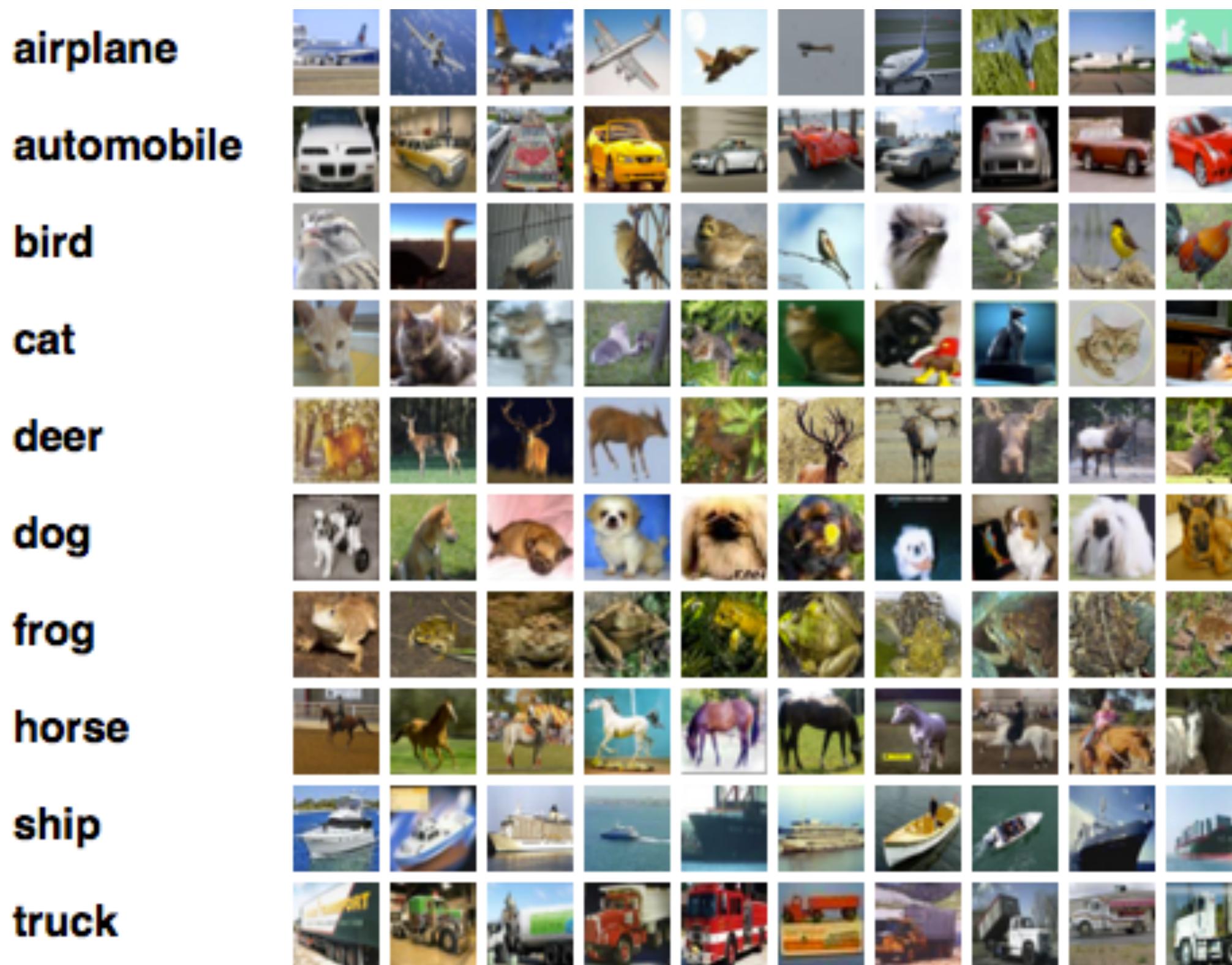
Tiny Image Dataset

- Precursor to ImageNet and CIFAR10/100
- 80 million images collected via image search circa 2008 using 75,062 noun synsets from WordNet (labels are noisy)
- Very small images (32x32xRGB) used to minimise storage
- Note human performance is still quite good at this scale!



CIFAR10 Dataset

- Hand labelled set of 10 categories from Tiny Images dataset
- 60,000 32x32 images in 10 classes (50k train, 10k test)



Good test set for visual recognition problems

Classification

Problem:

Assign new observations into one of a fixed set of categories (classes)

Key Idea(s):

Build a model of data in a given category based on observations of instances in that category

Classification

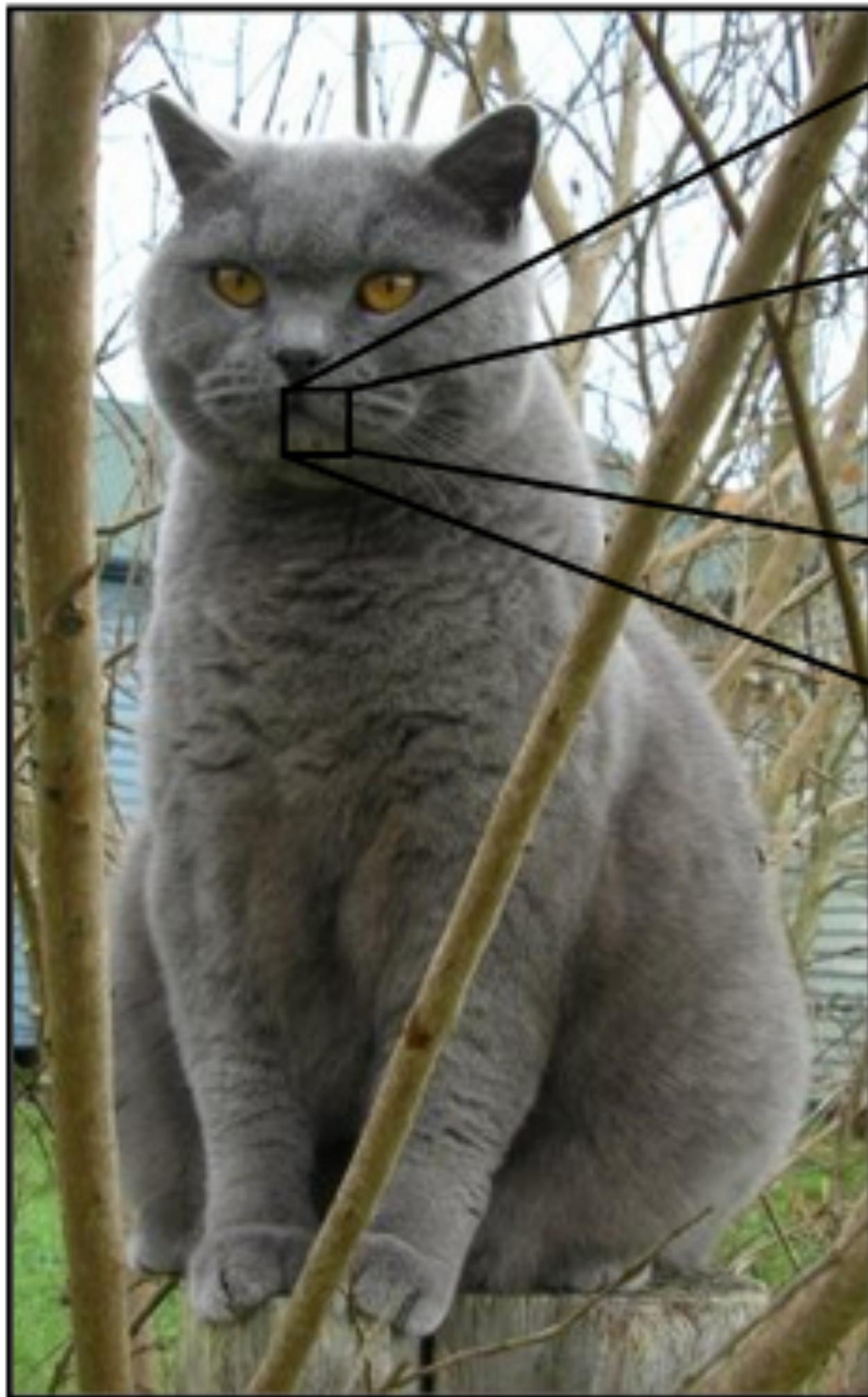


(assume given set of discrete labels)
{dog, cat, truck, plane, ...}



cat

Classification



08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	91	68
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	56	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	63	21	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	65	89	41	92	36	54	22	40	40	28	66	33	13	80
24	47	33	60	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94	21
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
00	44	68	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53	69
04	42	16	73	52	95	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	31	60	99	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	48	66	41	16	23	57	05	54
01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	47	48	00

What the computer sees

image classification

82% cat
15% dog
2% hat
1% mug

Classification

A **classifier** is a procedure that accepts as input a set of features and outputs a class **label**

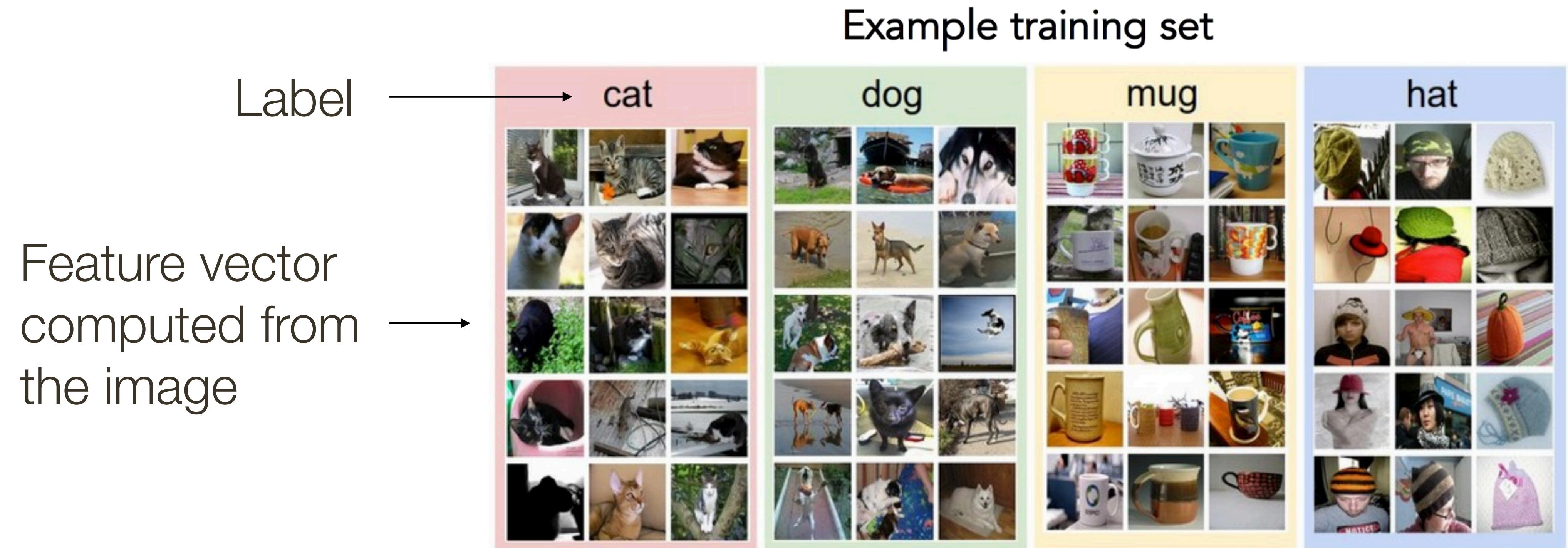
Classifiers can be binary (face vs. not-face) or multi-class (cat, dog, horse, ...).

We build a classifier using a **training set** of labelled examples $\{(\mathbf{x}_i, y_i)\}$, where each \mathbf{x}_i is a feature vector and each y_i is a class label.

Given a previously unseen observation, we use the classifier to predict its class label.

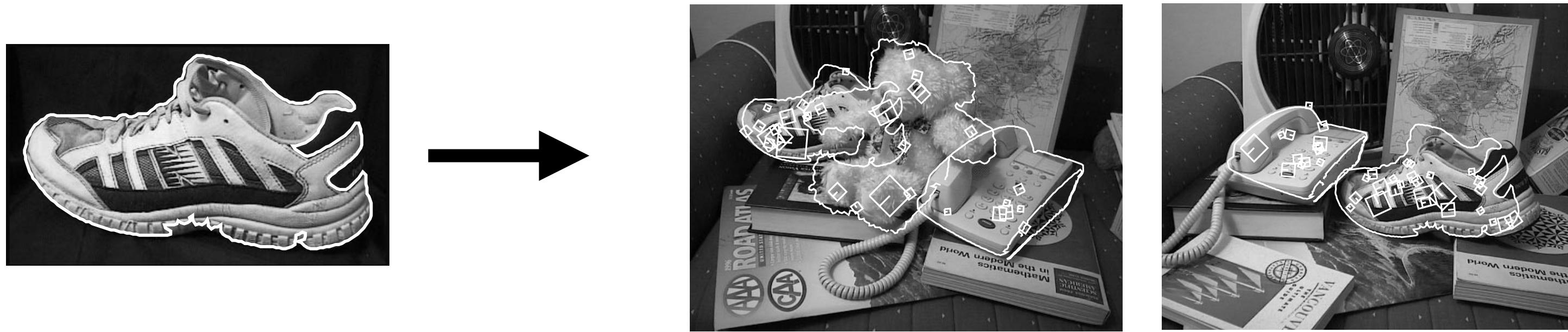
Classification

- Collect a database of images with labels
- Use ML to train an image classifier
- Evaluate the classifier on test images



Instance Recognition using Local Features

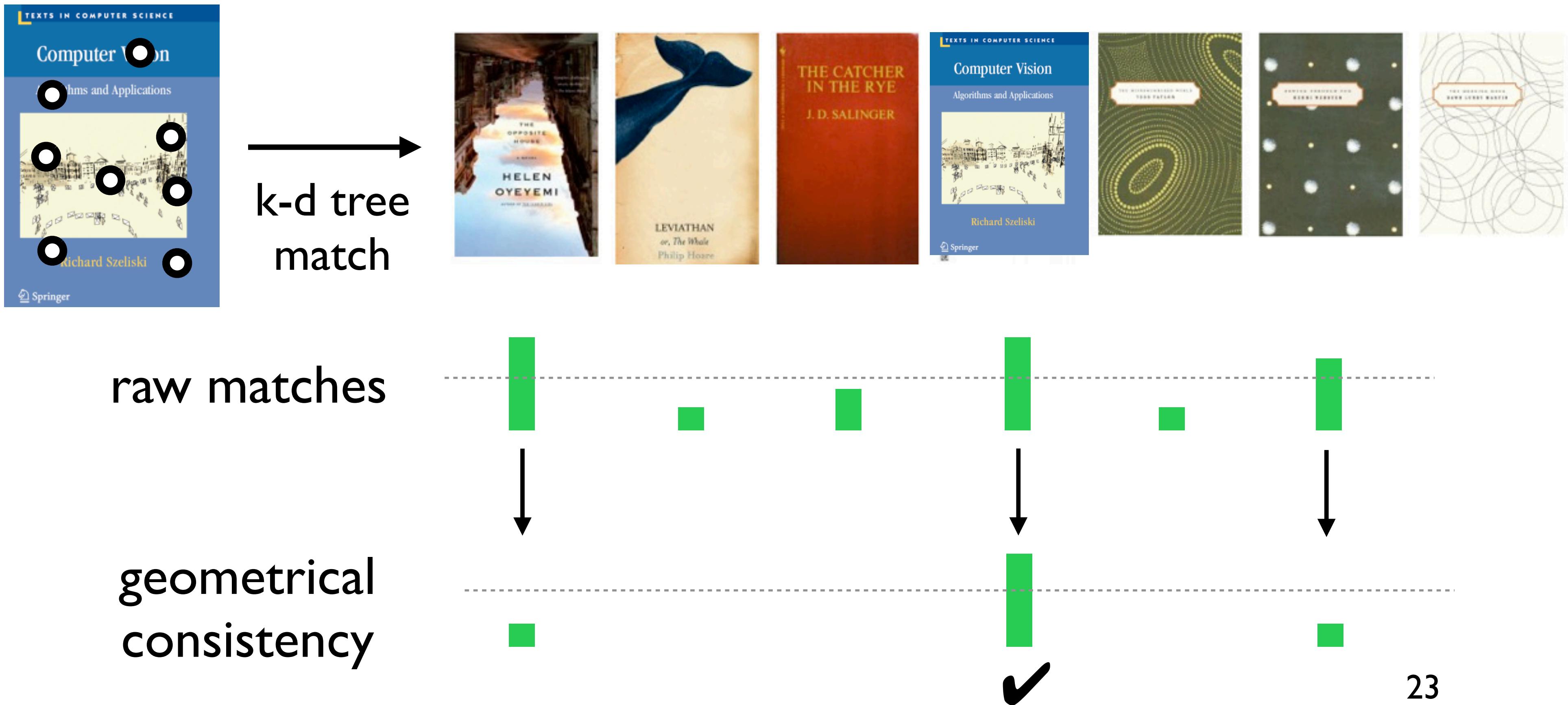
- Feature-based object instance recognition is similar to image registration (2D) or camera pose estimation (3D):



1. Detect Local Features (e.g., SIFT) in all images
 2. Match Features using Nearest Neighbours
 3. Find geometrically consistent matches using RANSAC
(with Affine/Homography or Fundamental matrix)
- The final stage is to verify the match, e.g., require that # consistent matches > threshold

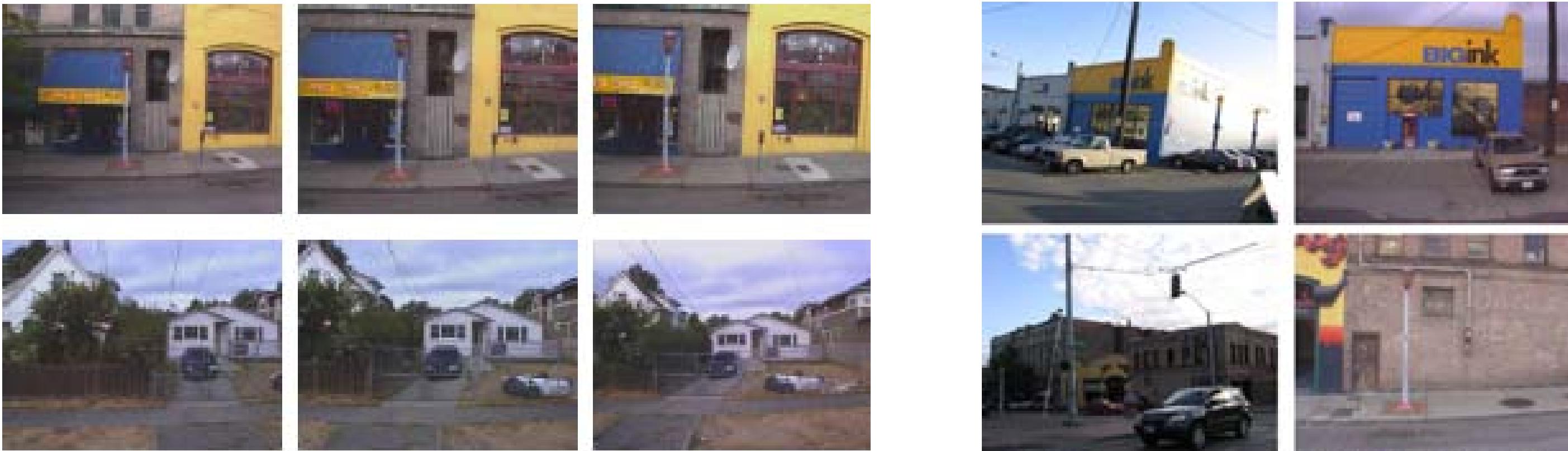
Scaling Local Feature Recognition

- To avoid performing all pairwise comparisons $O(n^2)$:
- Match query descriptors to entire database using k-d tree
- Select subset with max # raw matches and check geometry

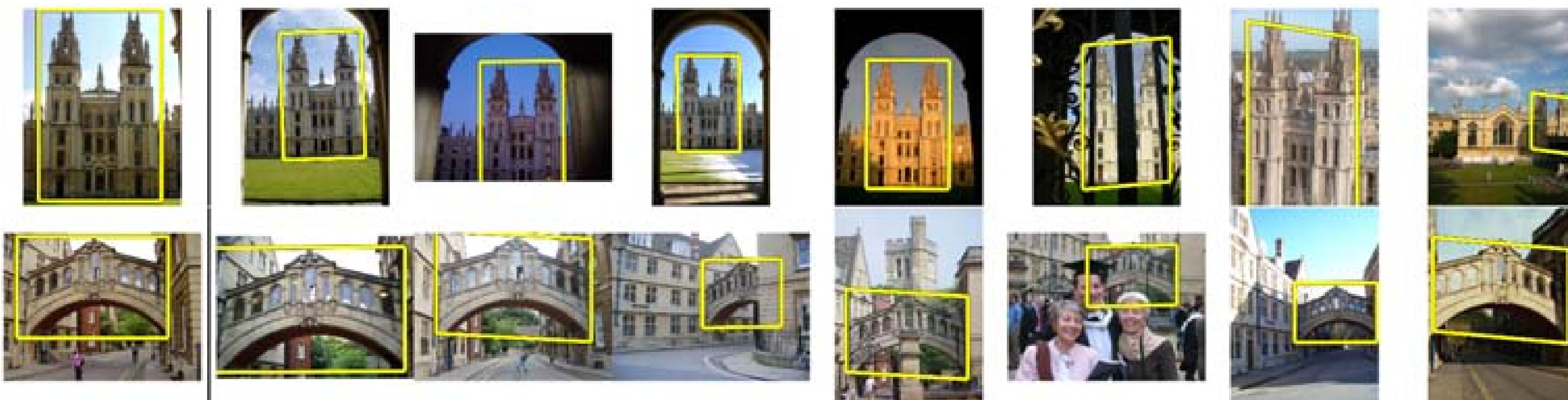


Application: Location Recognition

- Find photo in streetside imagery



[Schindler Brown Szeliski 2007]



[Philbin et al 2007] 24

Local Feature Recognition Failures

- Features + RANSAC fails with large appearance variation, e.g., most object categories and some instance problems



Few correct matches

Local Feature Recognition Failures

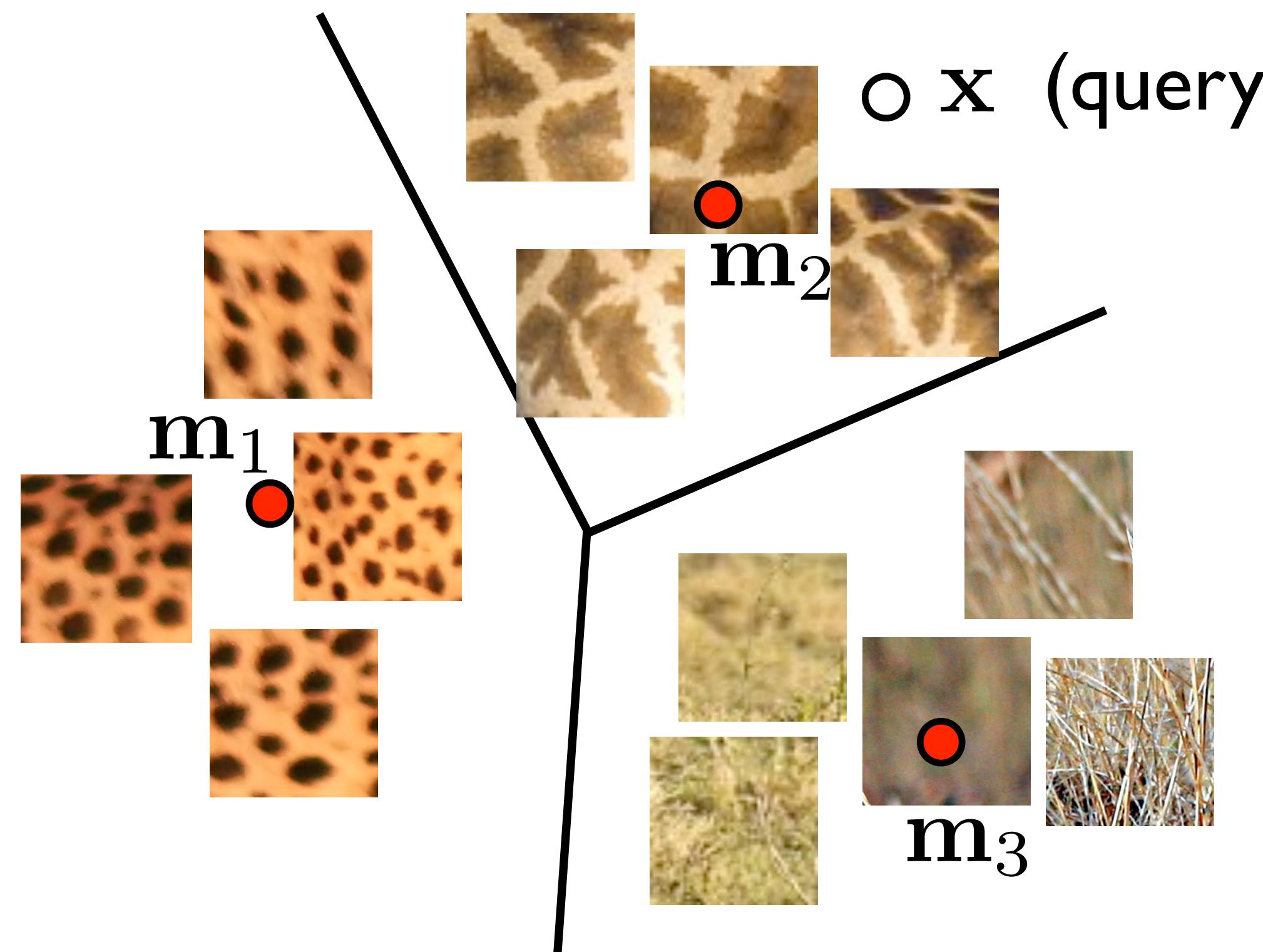
- Features + RANSAC fails with large appearance variation, e.g., most object categories and some instance problems



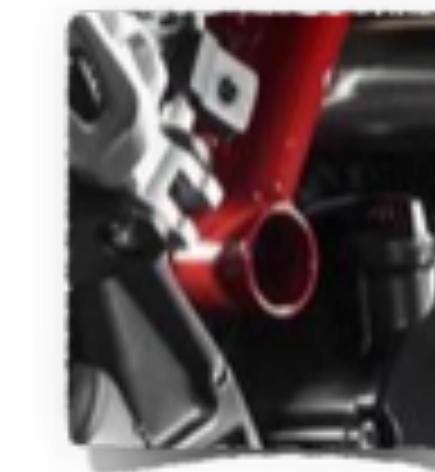
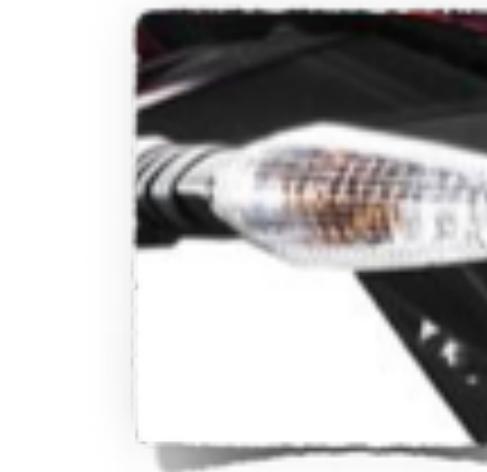
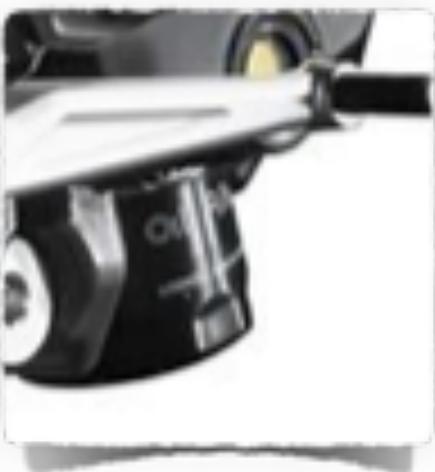
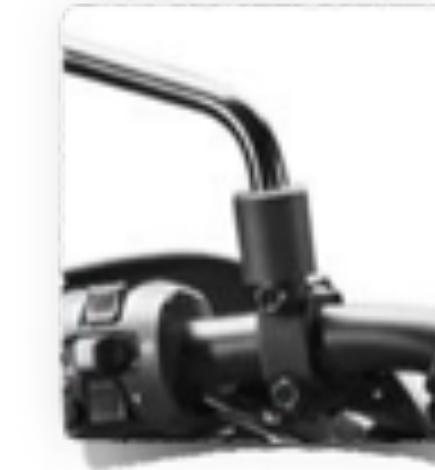
No correct matches

Category Recognition: Visual Words

- The amorphous appearance of visual categories can be modelled using regions of feature space
- A common method is to quantise feature descriptors to a codebook of “visual words” using k-means clustering

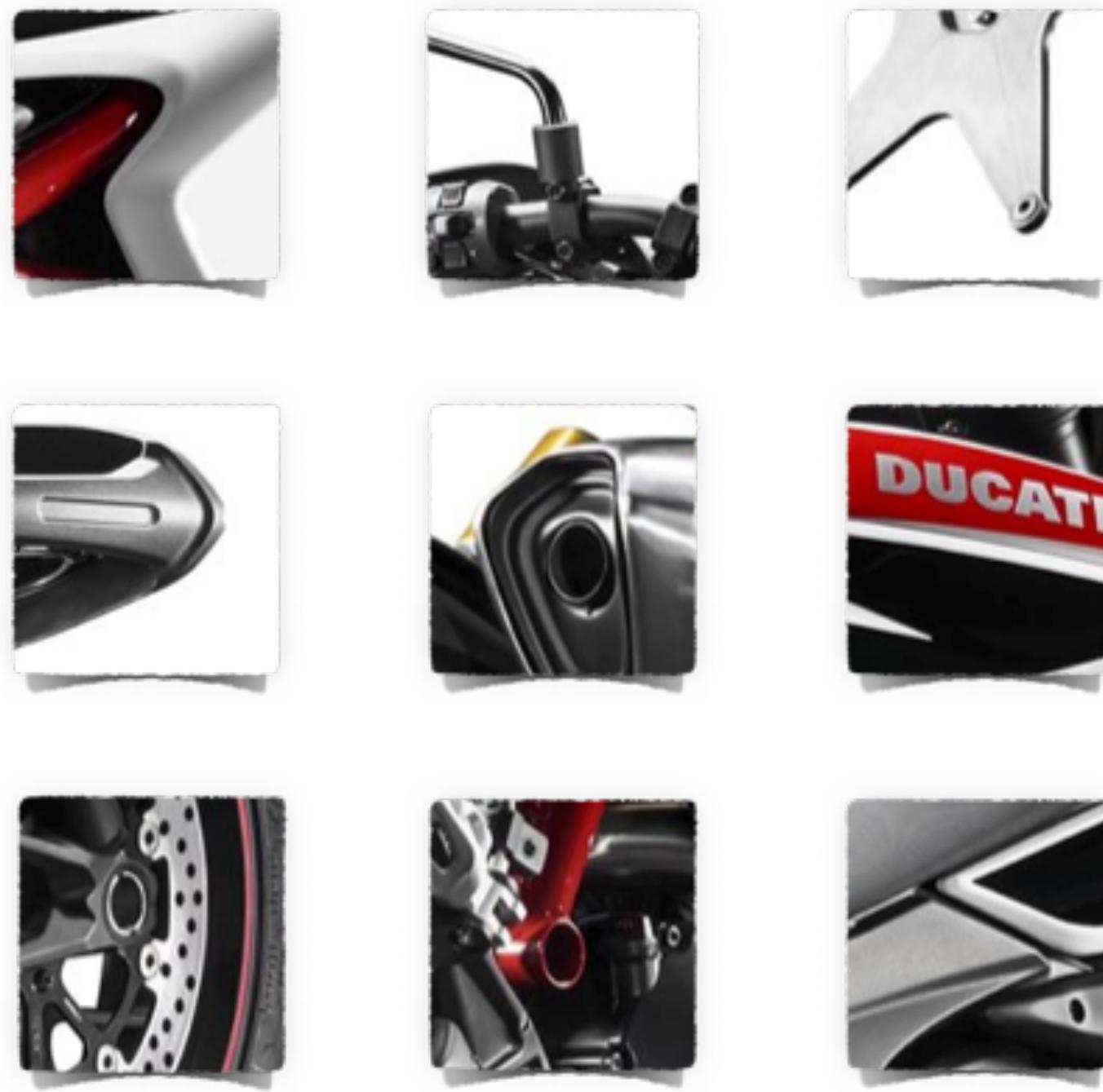
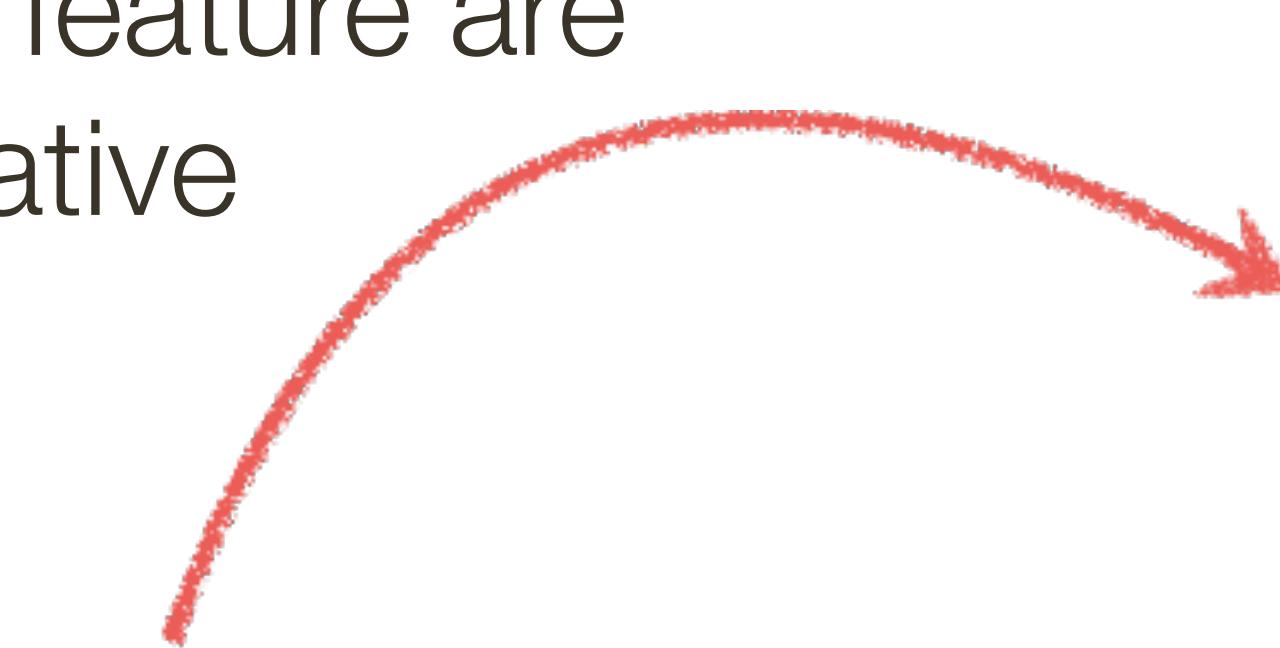


What Objects do These Parts Belong To?



Some local feature are very informative

An object as



a collection of local features
(bag-of-features)

- deals well with occlusion
- scale invariant
- rotation invariant

K-Means Clustering

Assume we know how many clusters there are in the data - denote by K

Each cluster is represented by a cluster center, or mean

Our objective is to minimize the representation error (or quantization error) in letting each data point be represented by some cluster center

Minimize

$$\sum_{i \in \text{clusters}} \left\{ \sum_{j \in i^{\text{th}} \text{ cluster}} \|x_j - \mu_i\|^2 \right\}$$

K-Means Clustering

K-means clustering alternates between two steps:

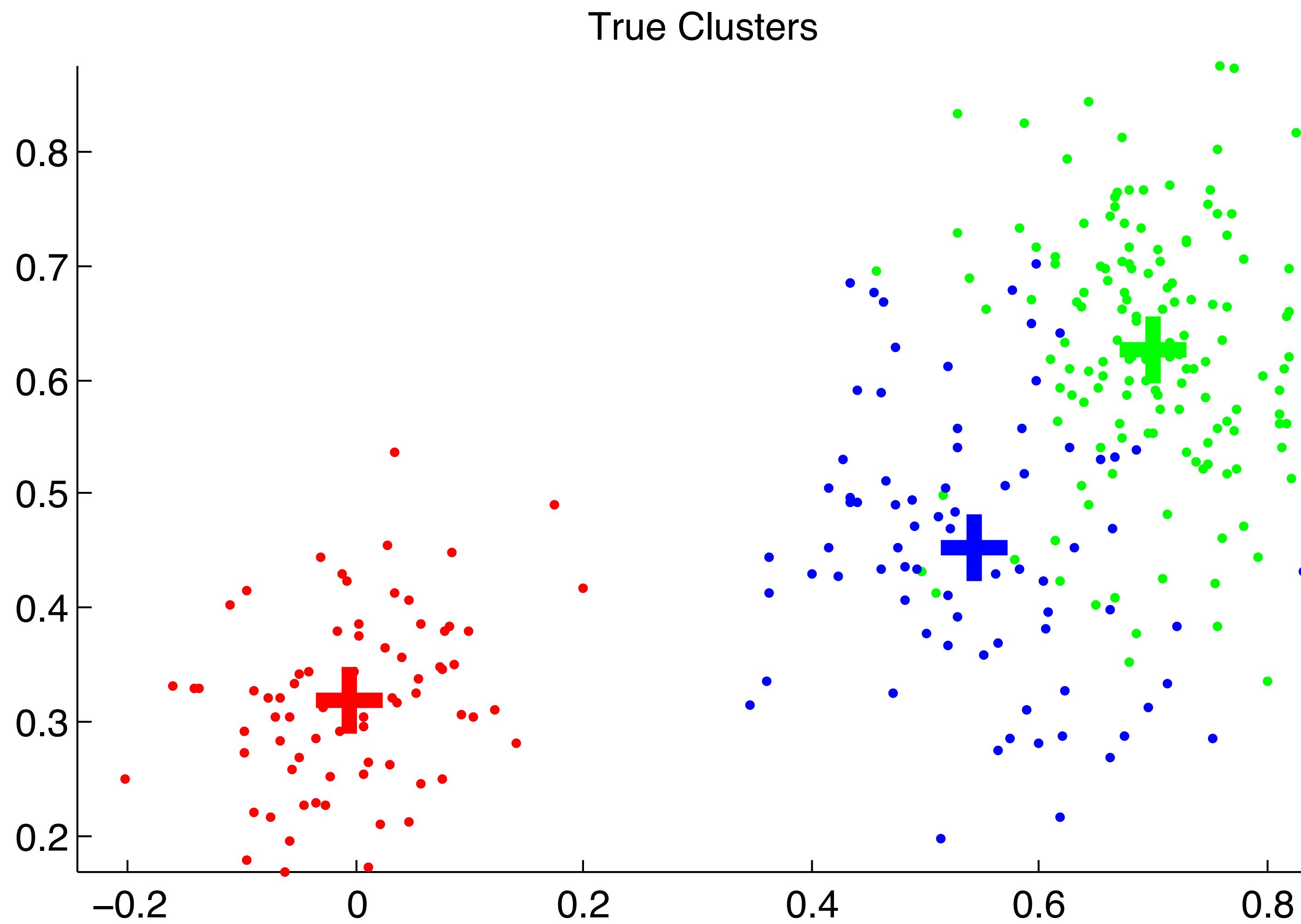
1. Assume the cluster centers are known (fixed). Assign each point to the closest cluster center.
2. Assume the assignment of points to clusters is known (fixed). Compute the best center for each cluster, as the mean of the points assigned to the cluster.

The algorithm is initialized by choosing K random cluster centers

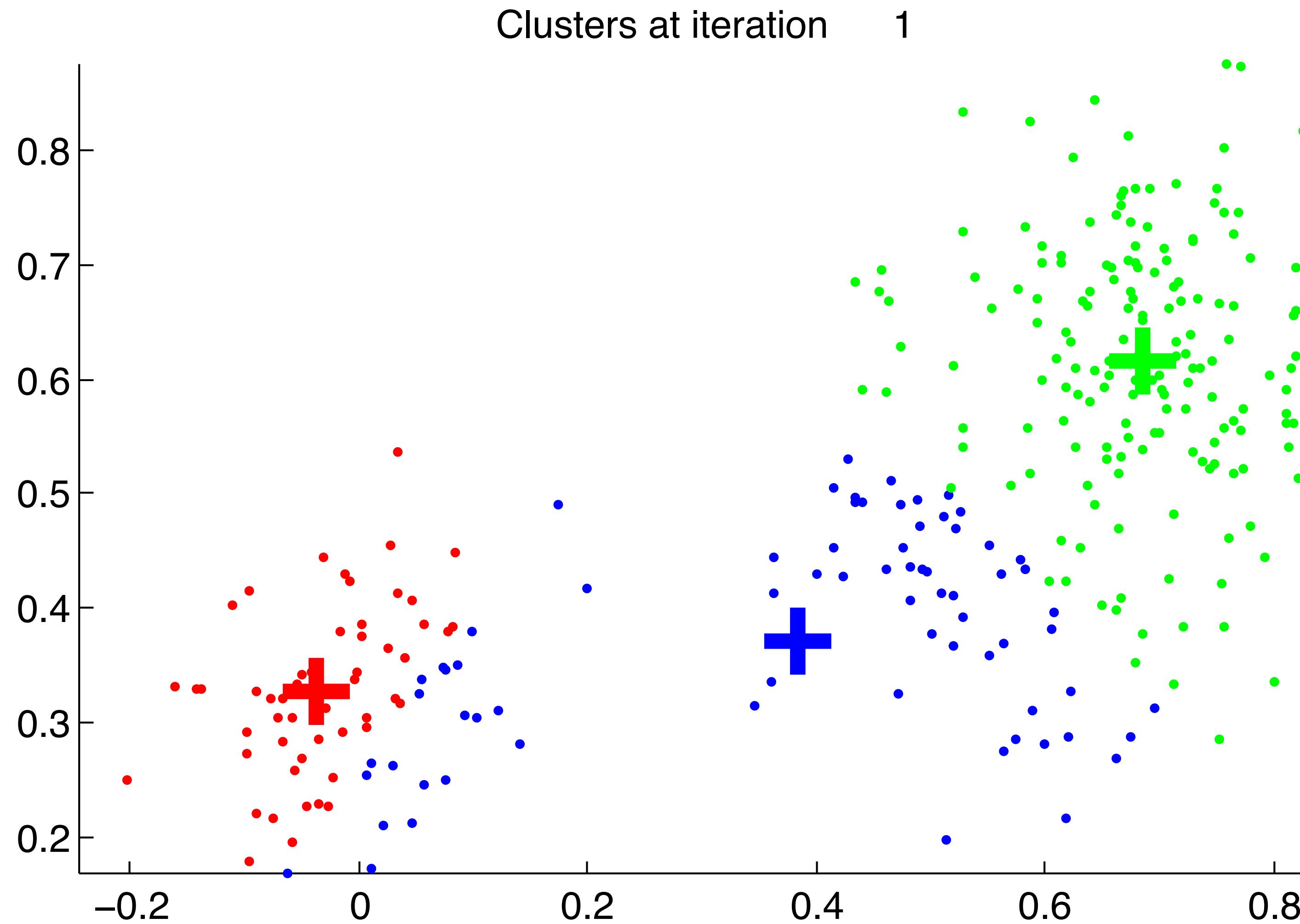
K-means converges to a local minimum of the objective function

- Results are initialization dependent

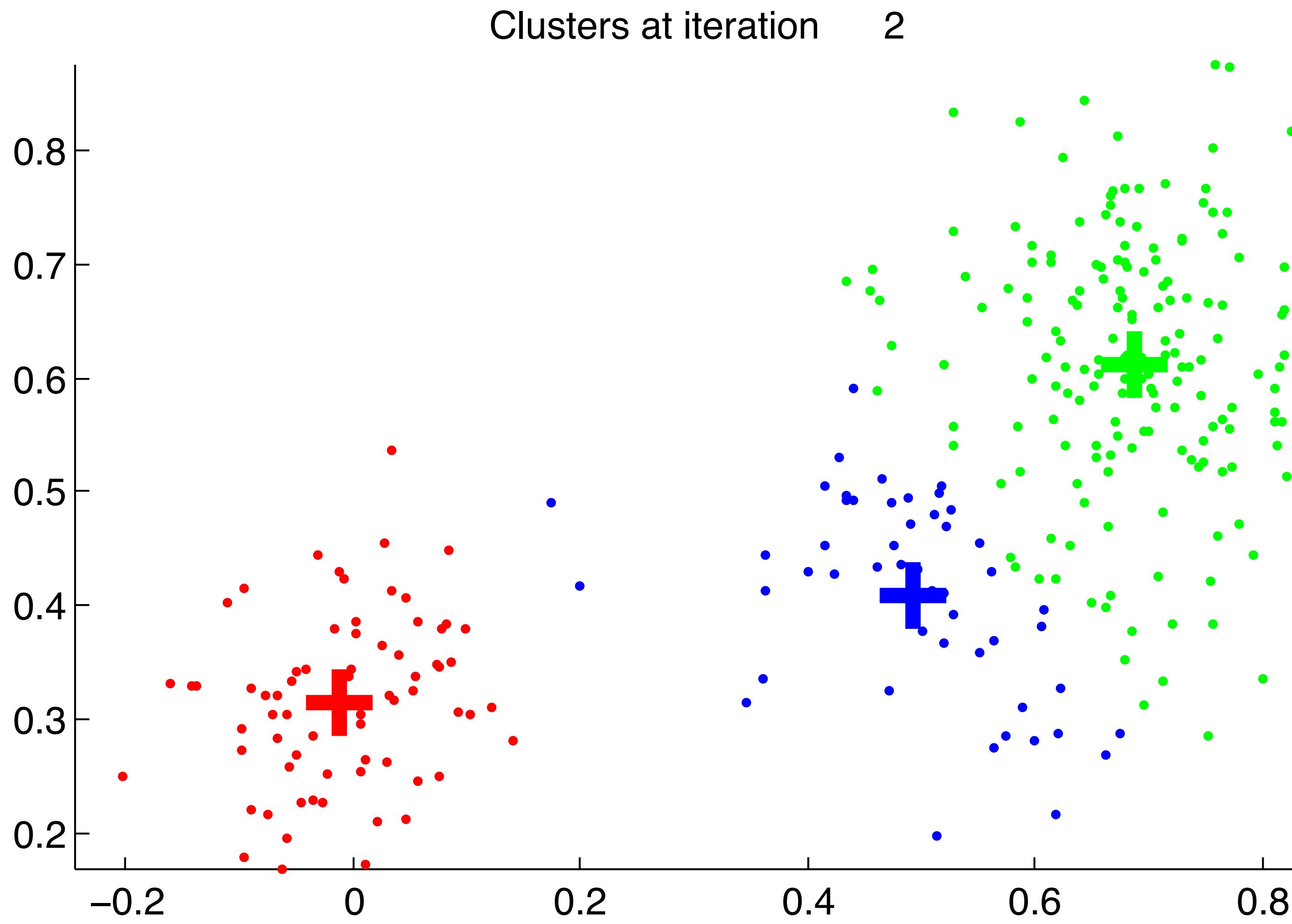
Example 1: K-Means Clustering



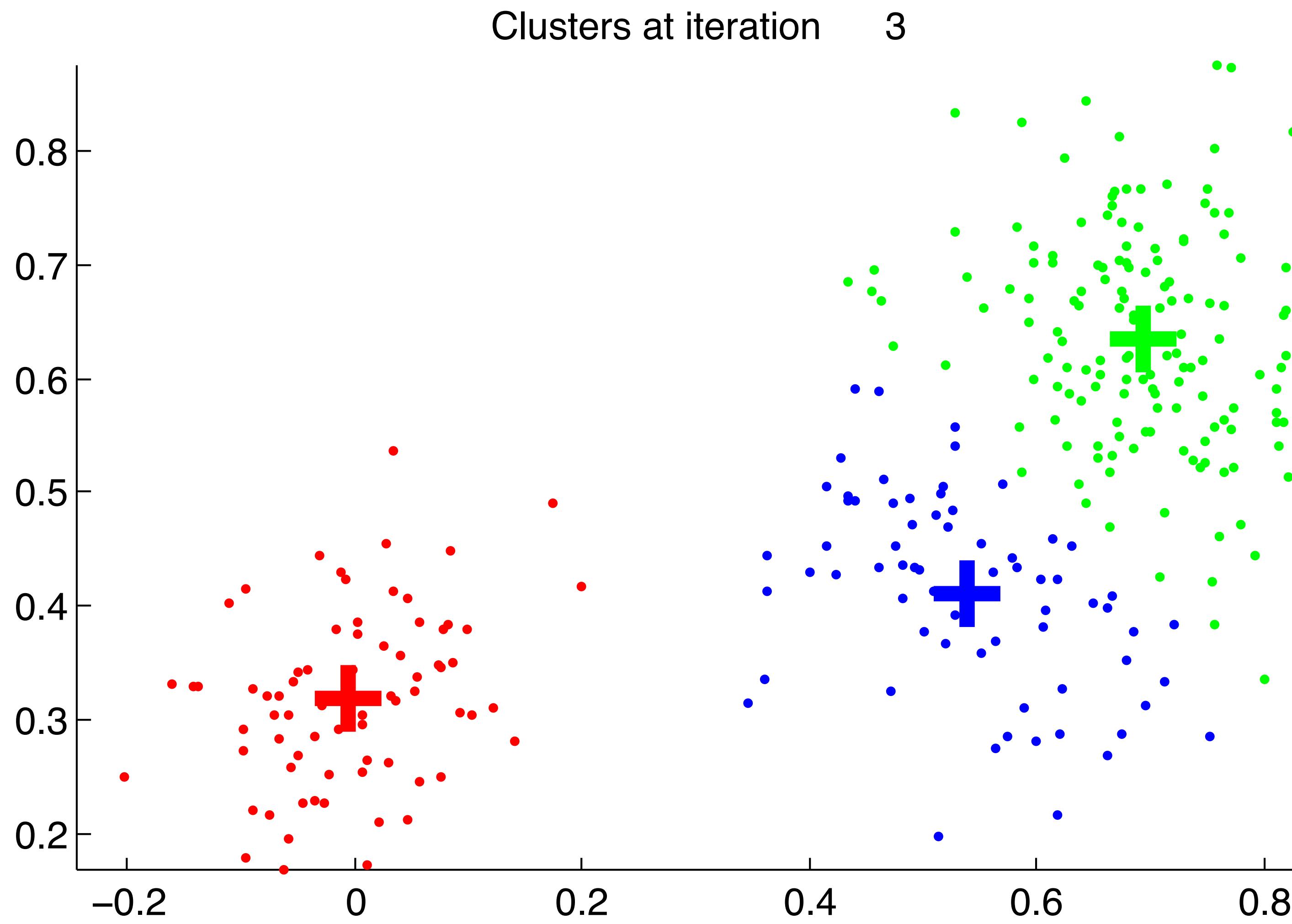
Example 1: K-Means Clustering



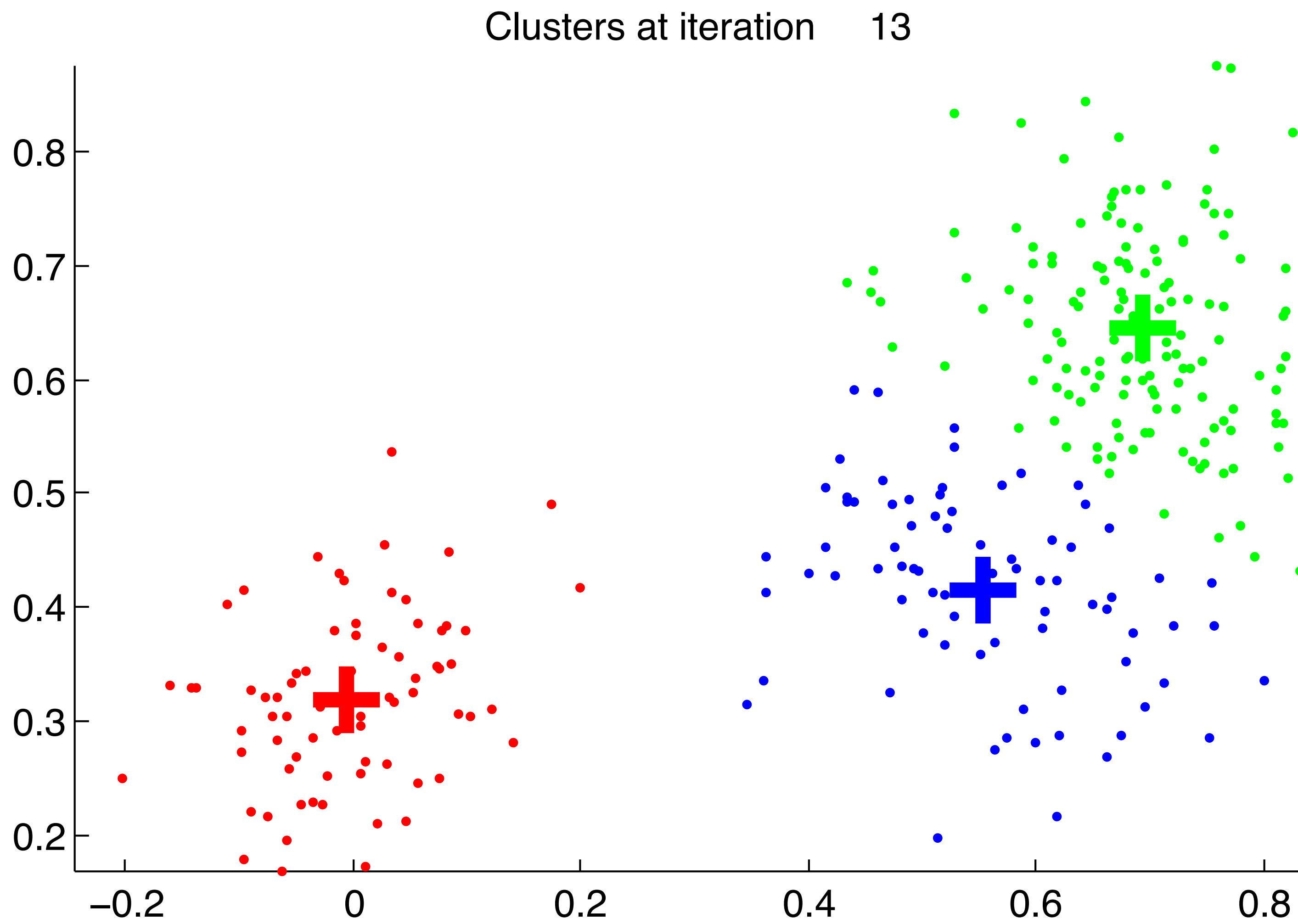
Example 1: K-Means Clustering



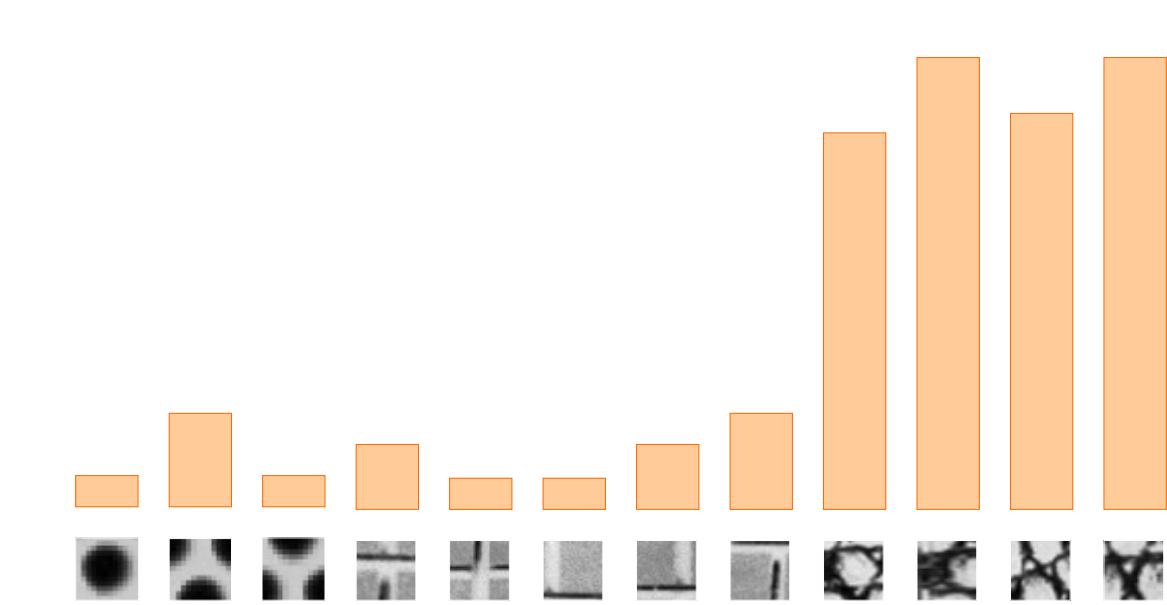
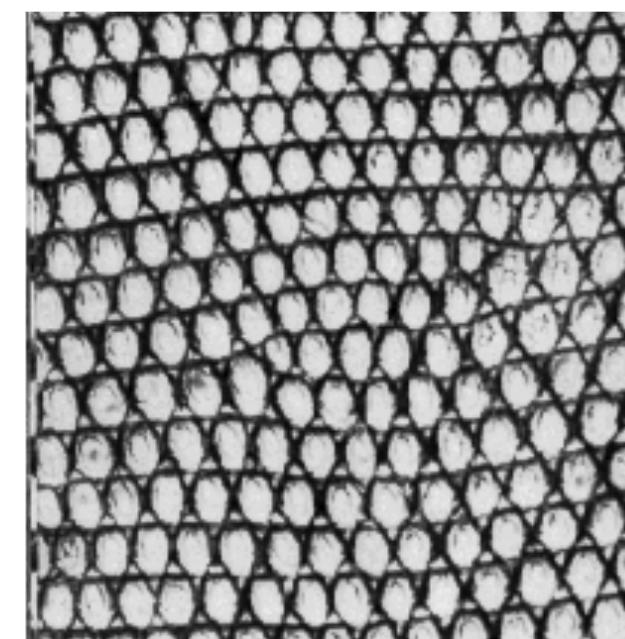
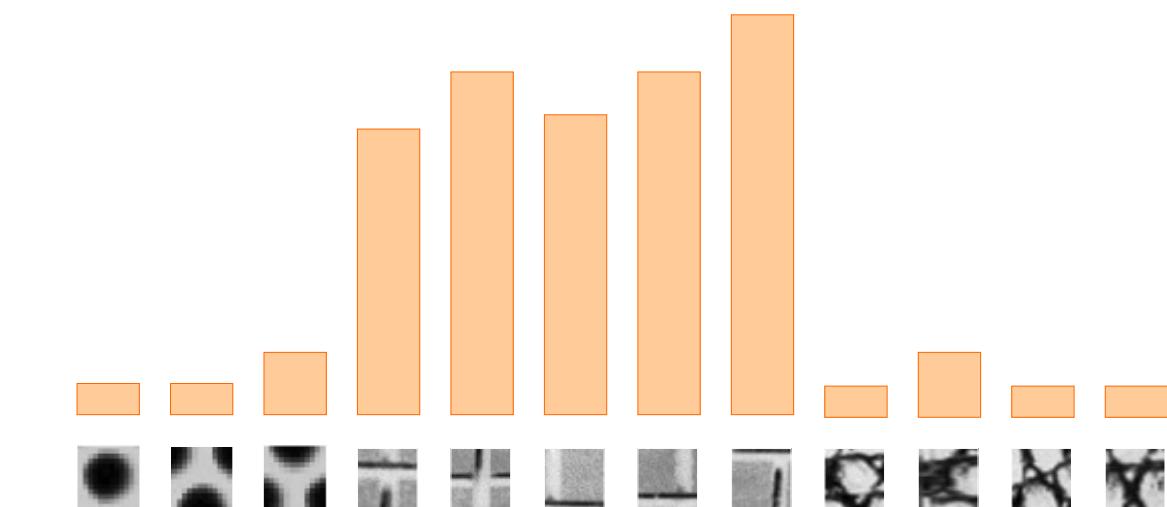
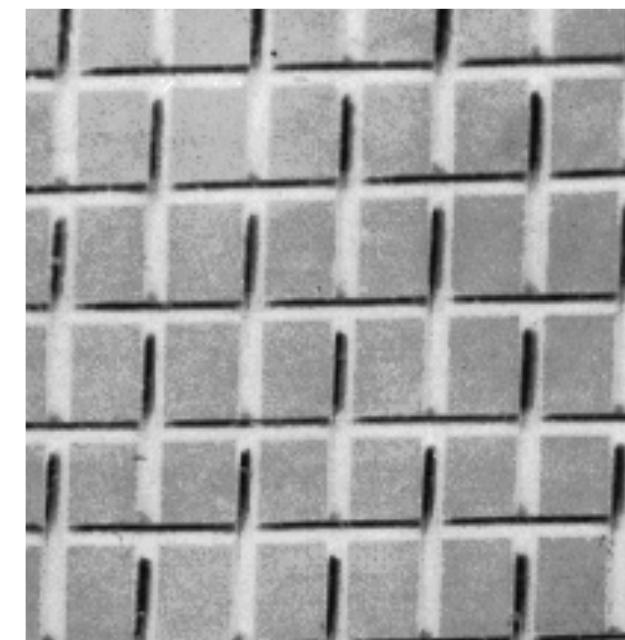
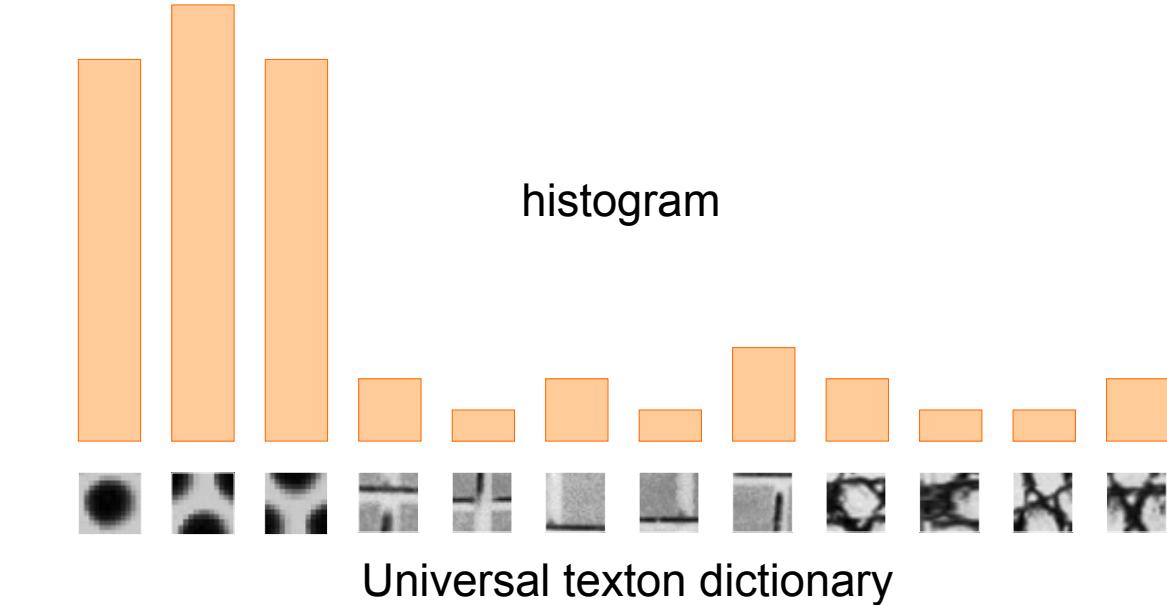
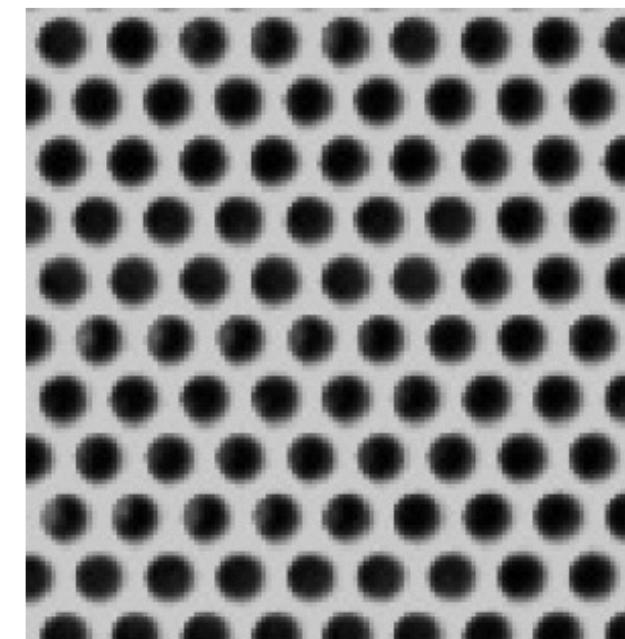
Example 1: K-Means Clustering



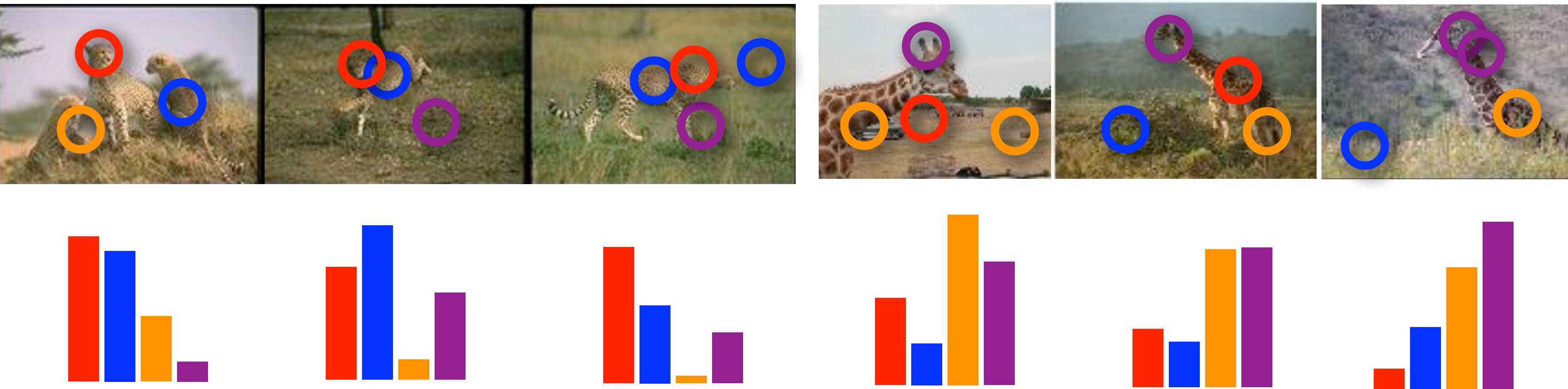
Example 1: K-Means Clustering



Recall: Texture Representation



Visual Word Histogram + SVM



- A popular category recognition method was to use histograms of visual word frequencies to represent each image
- Given a labelled image dataset, a Support Vector Machine (SVM) could be trained to perform image classification, with per-image visual word histograms as input
- Variants on this theme were state-of-the-art for image classification up to around 2011 (deep learning + AlexNet)

Standard **Bag-of-Words** Pipeline (for image classification)

Dictionary Learning:

Learn Visual Words using clustering

Encode:

build Bags-of-Words (BOW) vectors
for each image

Classify:

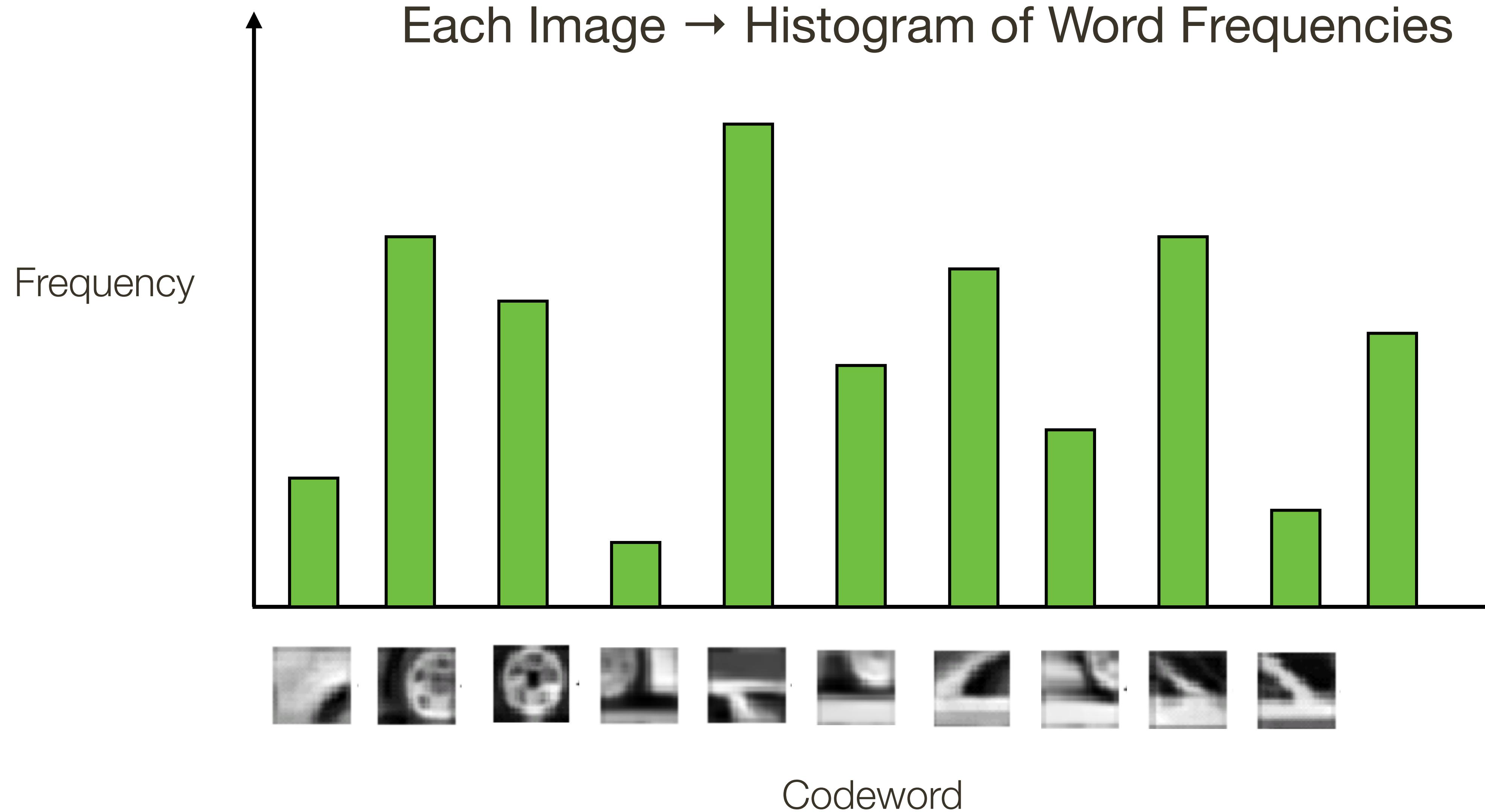
Train and test data using BOWs

1: Construct Visual Dictionary



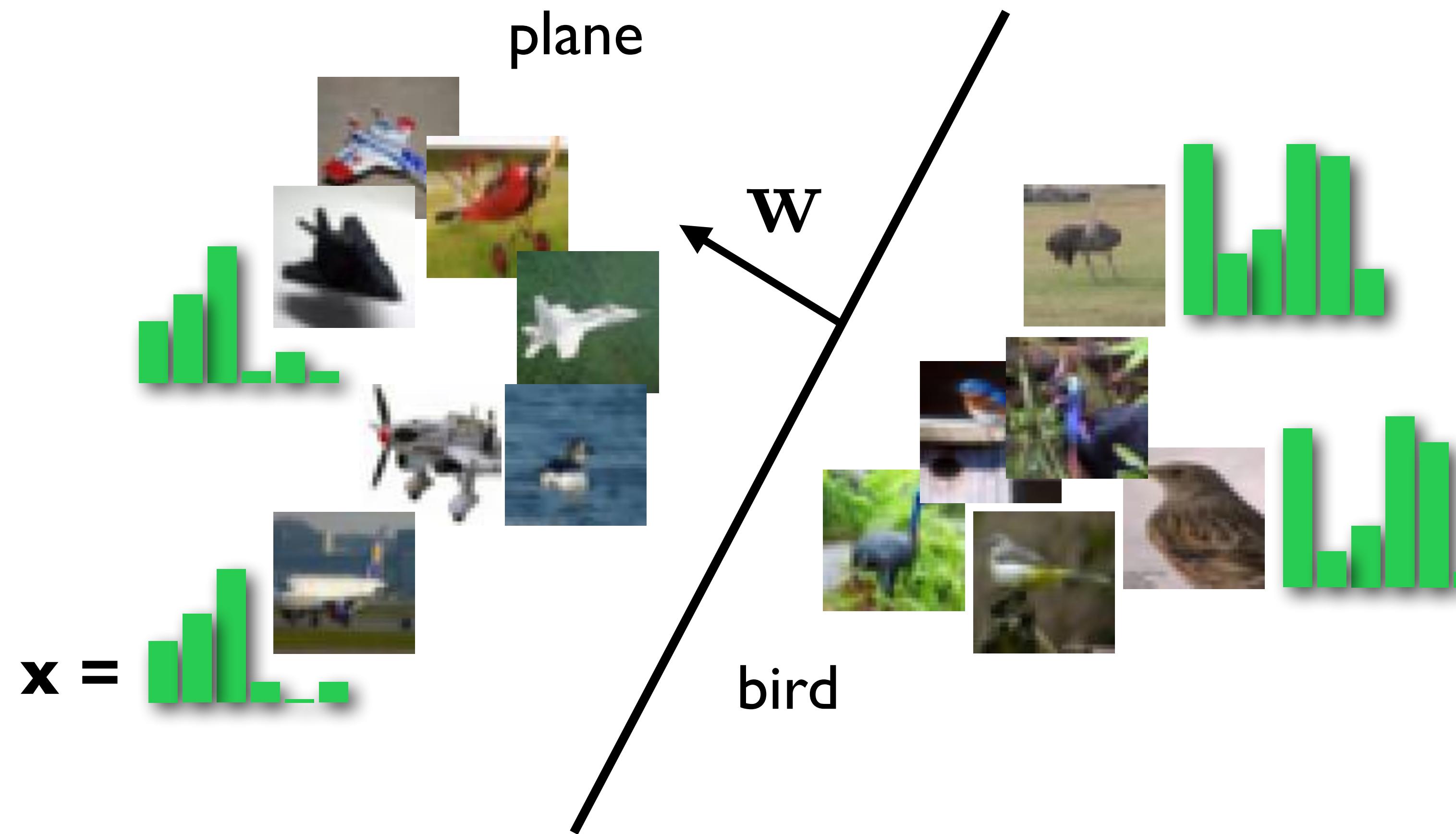
Source: B. Leibe

2. Compute Histogram of Visual Words



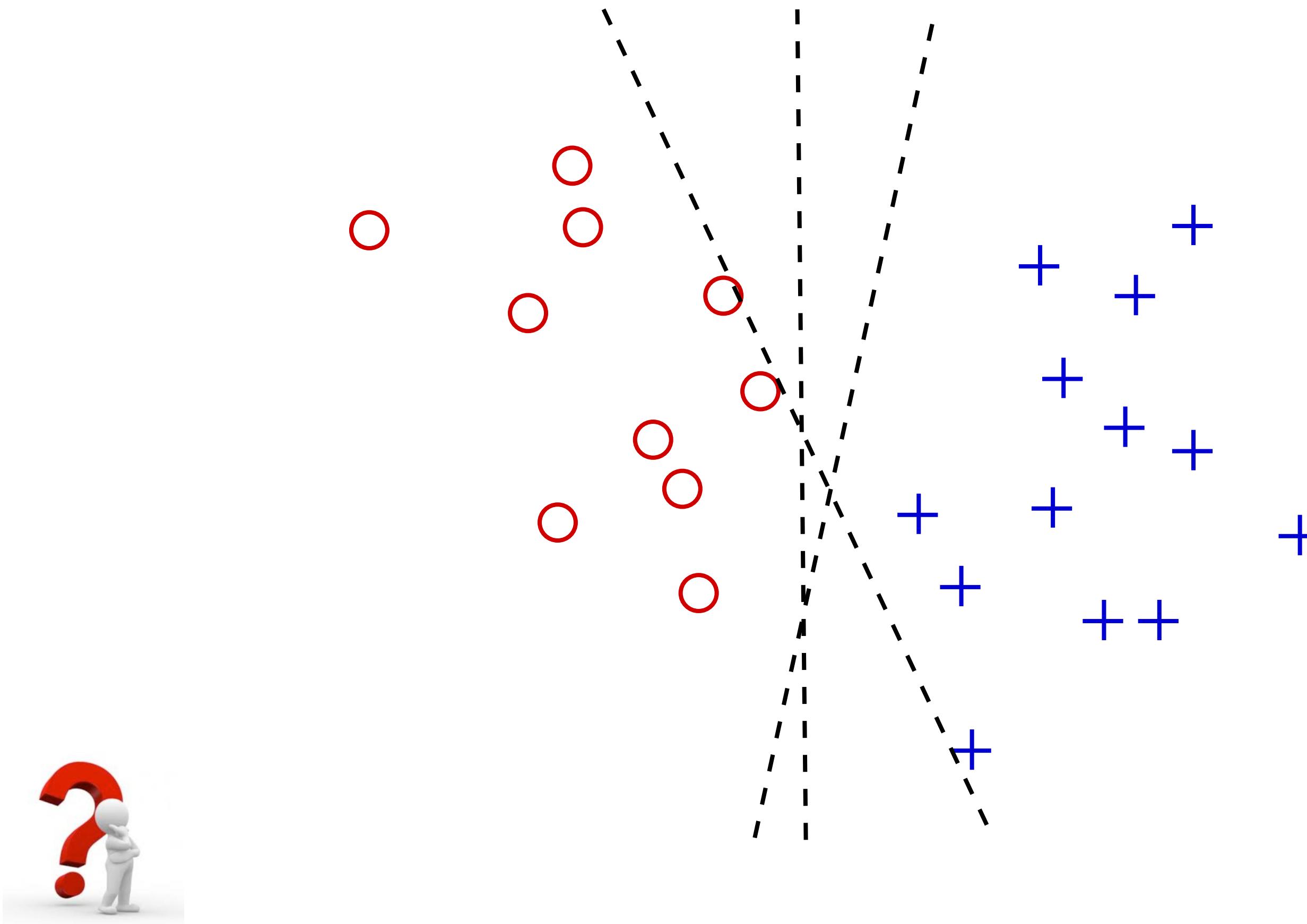
Classify Visual Word Histograms

- e.g., bird vs plane classifier as linear classifier in space of histograms
- Histograms of visual word frequencies = vector \mathbf{x} , linear classifier \mathbf{w}



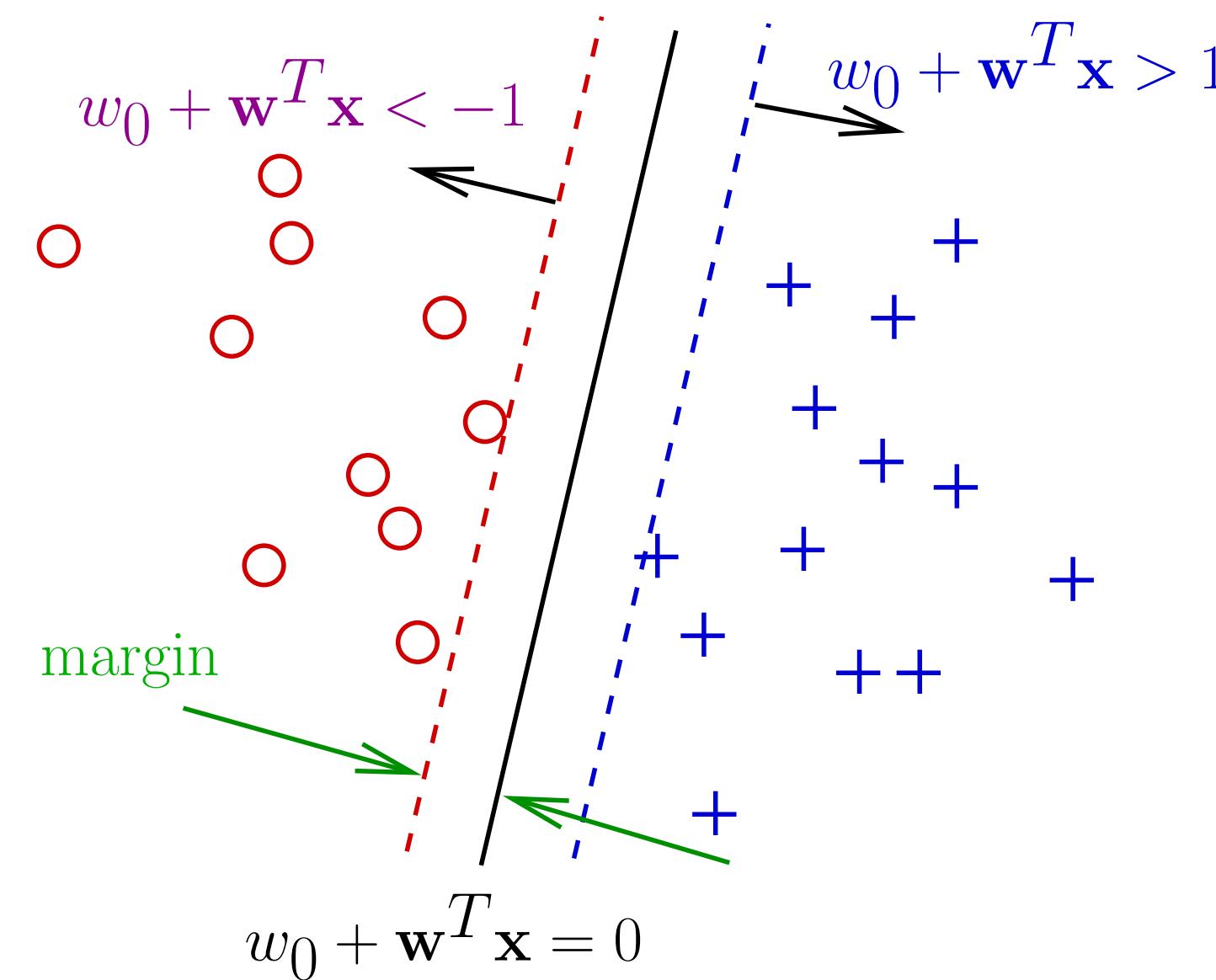
Support Vector Machines

- Which decision boundary is best?



Max-Margin Classifier

- Separation between classes is called the **margin**



- Distance from boundary

$$d_i = y_i(\mathbf{w}^T \mathbf{x}_i + w_0)$$

Note that d_i could be arbitrarily large for large \mathbf{w}

- Maximise the minimum distance for fixed $|\mathbf{w}|$

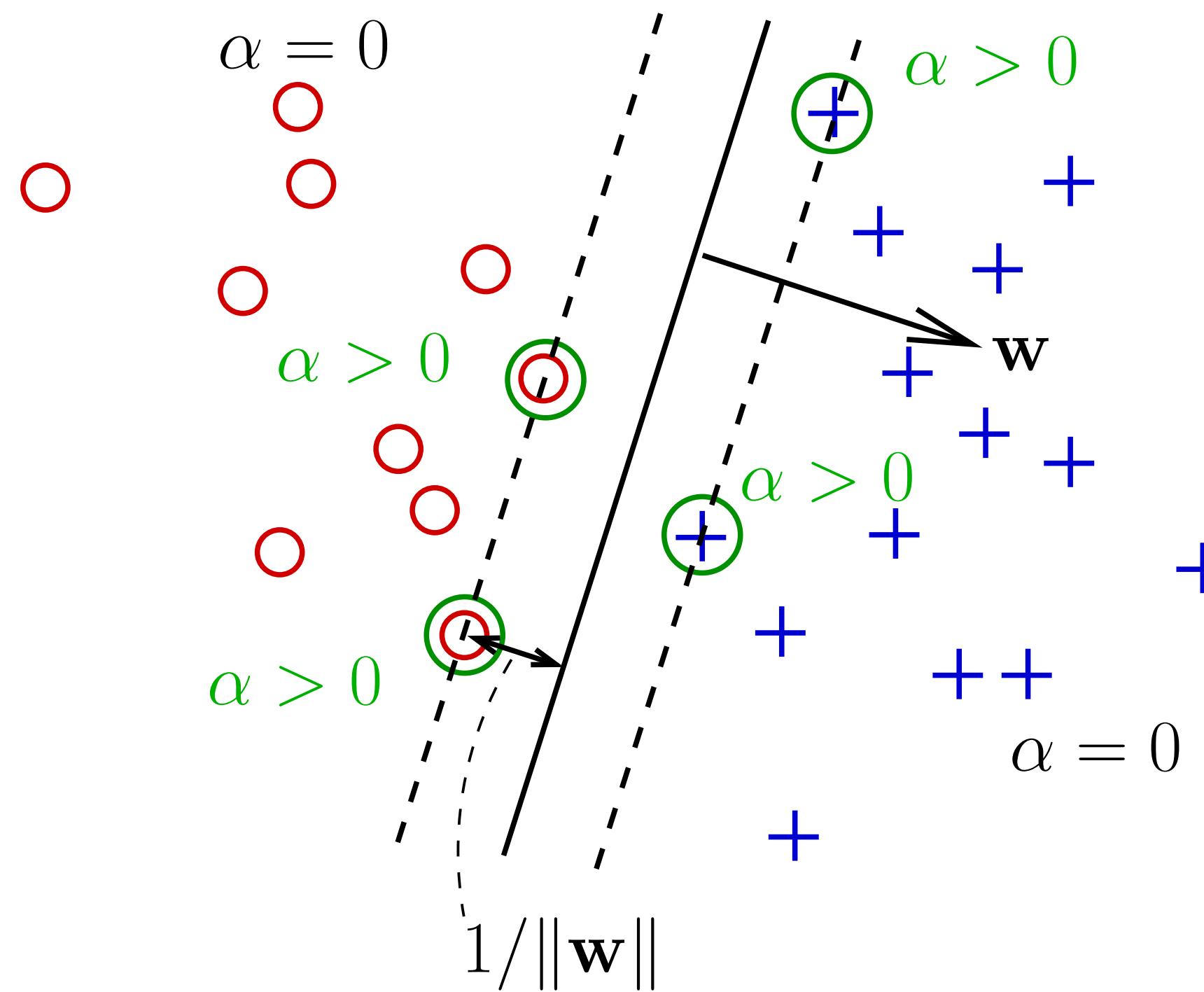
$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \min_i y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \quad s.t. \quad |\mathbf{w}|^2 = 1$$

(quadratic programming)

[Figure credits:
G. Shakhnarovich] 44

Support Vectors

- The active constraints are due to the data that define the classification boundary, these are called **support vectors**



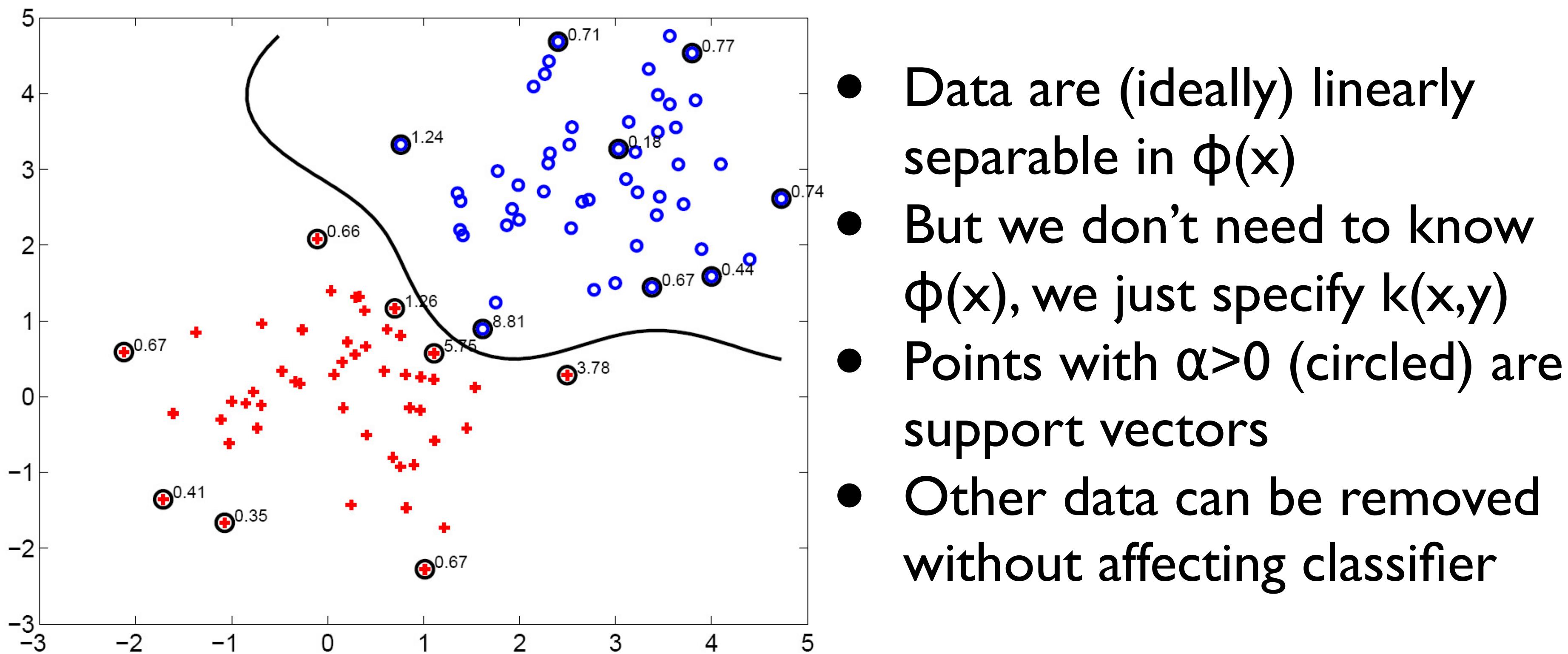
Final classifier can be written in terms of the support vectors:

$$\hat{y} = \text{sign} \left(\hat{w}_0 + \sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} \right)$$

Non-Linear SVM

- Replace inner product with kernel

$$\mathbf{x}_i^T \mathbf{x} \rightarrow \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) \rightarrow k(\mathbf{x}_i, \mathbf{x})$$



Bag-of-Words Representation

Algorithm:

Initialize an empty K -bin histogram, where K is the number of codewords

Extract local descriptors (e.g. SIFT) from the image

For each local descriptor \mathbf{x}

- Map (Quantize) \mathbf{x} to its closest codeword $\rightarrow \mathbf{c}(\mathbf{x})$

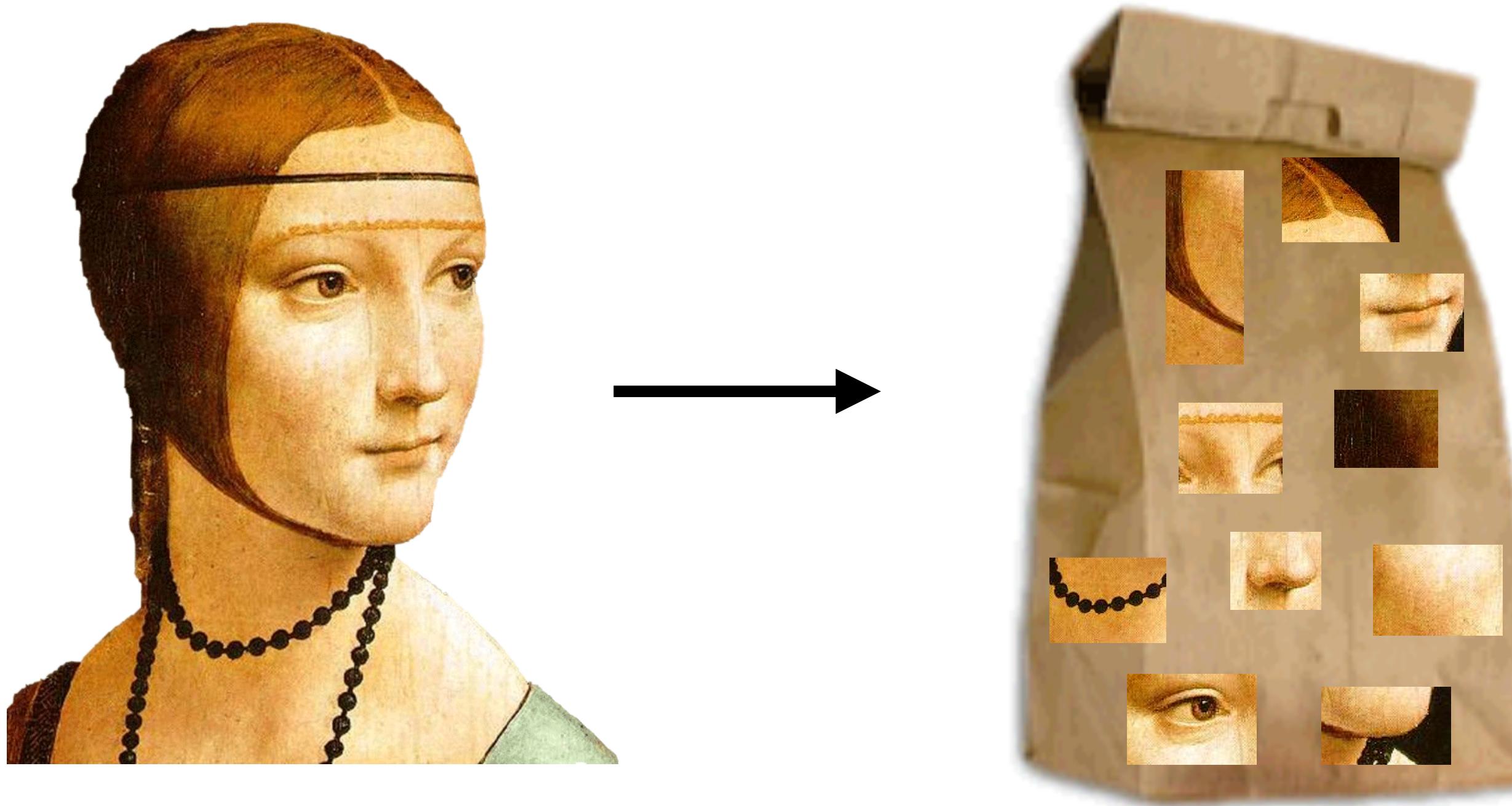
- Increment the histogram bin for $\mathbf{c}(\mathbf{x})$

Return histogram

We can then classify the histogram using a trained classifier, e.g. a support vector machine or k-Nearest Neighbor classifier

Bag of Words

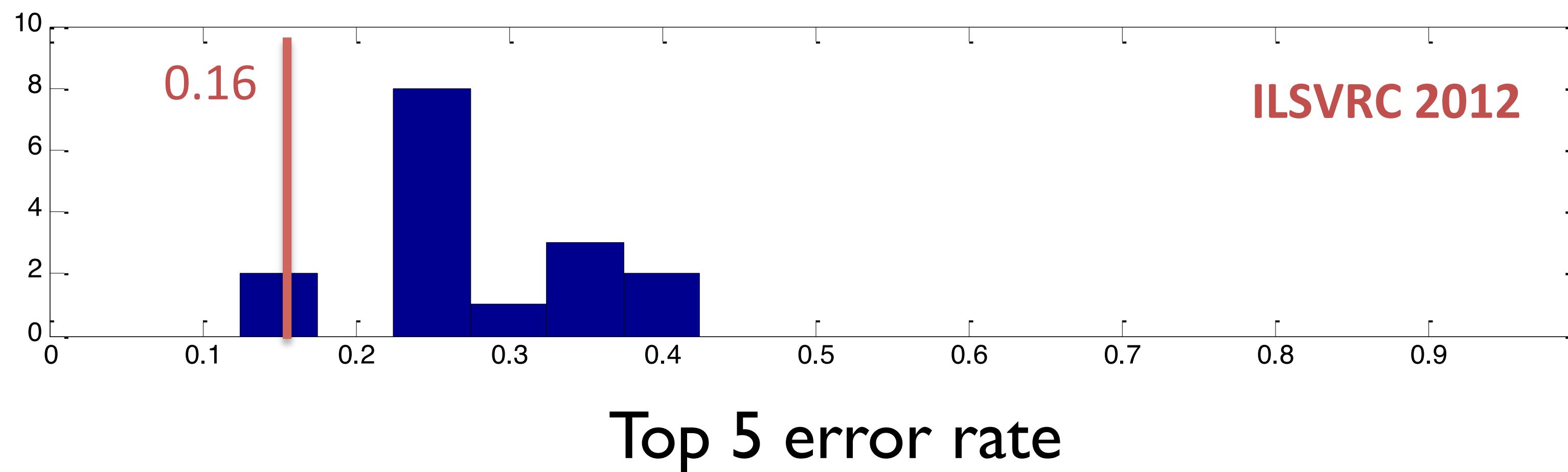
- Images are represented as collections of local visual words (discarding spatial relationships), before SVM classification



- There is some evidence that similar features are effective in CNNs, e.g., pooling of learned small receptive field (17×17) features gives good performance on ImageNet [“BagNets” Brendel Bethge 19]

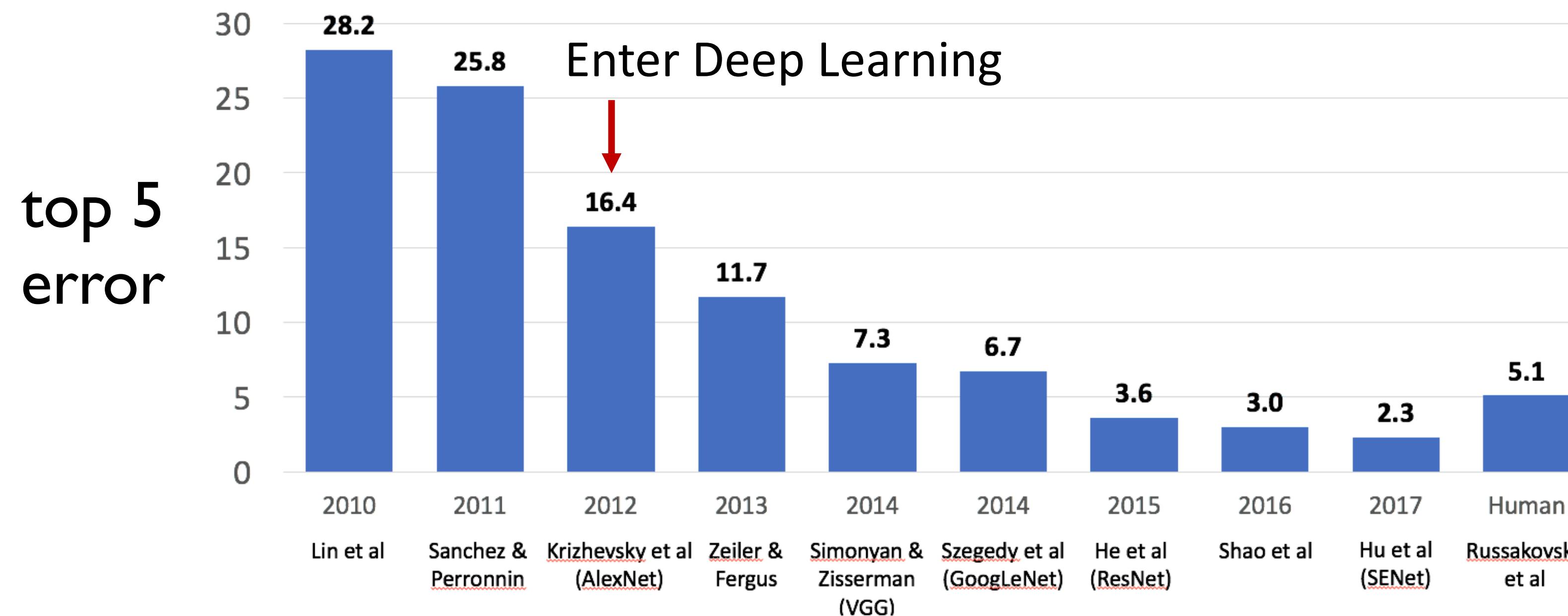
ILSVRC 2012

- For the ImageNet Large Scale Visual Recognition Challenge 2012 competitors had to classify 100K unseen test images into one of 1000 categories
- Alex Krizhevsky and Geoff Hinton used an 8-layer, 60M parameter convolutional neural network trained on two GPUs for 1 week
- This beat all other approaches (visual word/SVM based) by a large margin:



Alexnet

- Won the Imagenet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 by a large margin
- Some ingredients: Deep neural net (Alexnet), Large dataset (Imagenet), Lots of compute (2 GPU weeks), non-saturating activation functions (ReLU)



Summary

Factors that make image classification hard

- intra-class variation, viewpoint, illumination, clutter, occlusion...

A codebook of **visual words** contains representative local patch descriptors

- can be constructed by clustering local descriptors (e.g. SIFT) in training images

The **bag of words** model accumulates a histogram of occurrences of each visual word

An supervised classifier, such as a **Support Vector Machine (SVM)** is then used to classify the word histograms

Menu for Today

Topics:

- **Visual Classification** introduction
- **Instance vs Category** Recognition
- **Bag of Words, K-means**
- **Quiz 5**

Readings:

- **Today's Lecture:** Szeliski 11.4, 12.3-12.4, 9.3, 5.1-5.2

Reminders:

- **Assignment 4:** due today
- **Assignment 5:** Scene Recognition with Bag of Words is now available
- **No lectures next week** (mid term break)