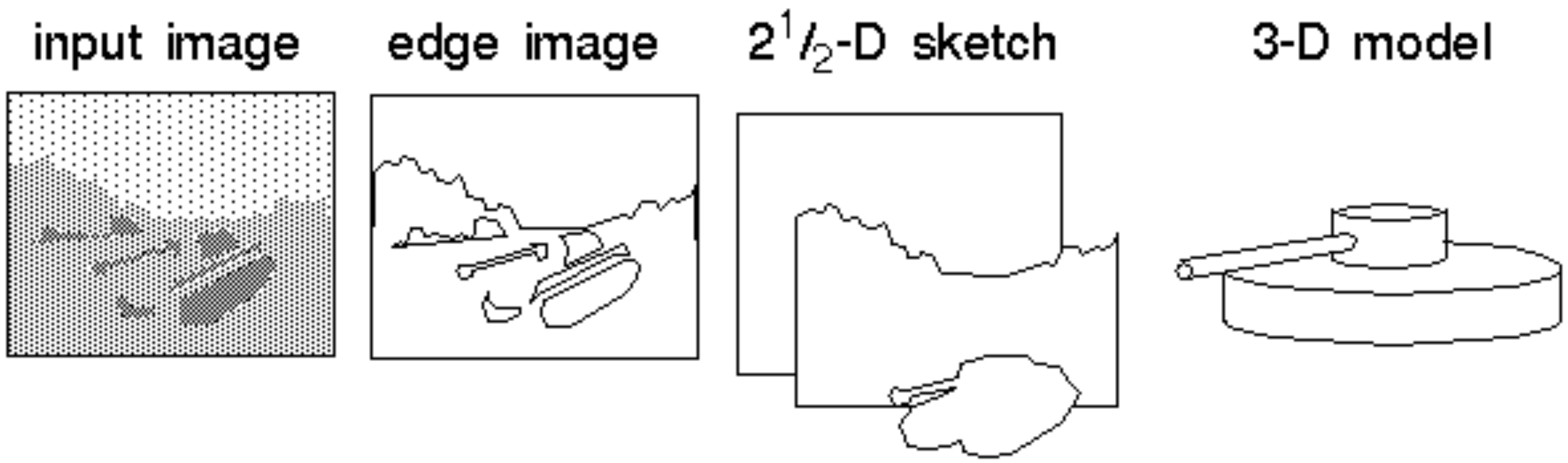
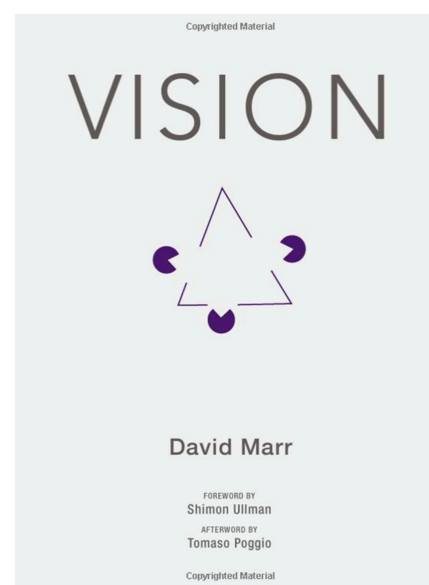
Object Recognition

Computer Vision
Fall 2018
Columbia University

The Big Picture



Low-level Mid-level High-level



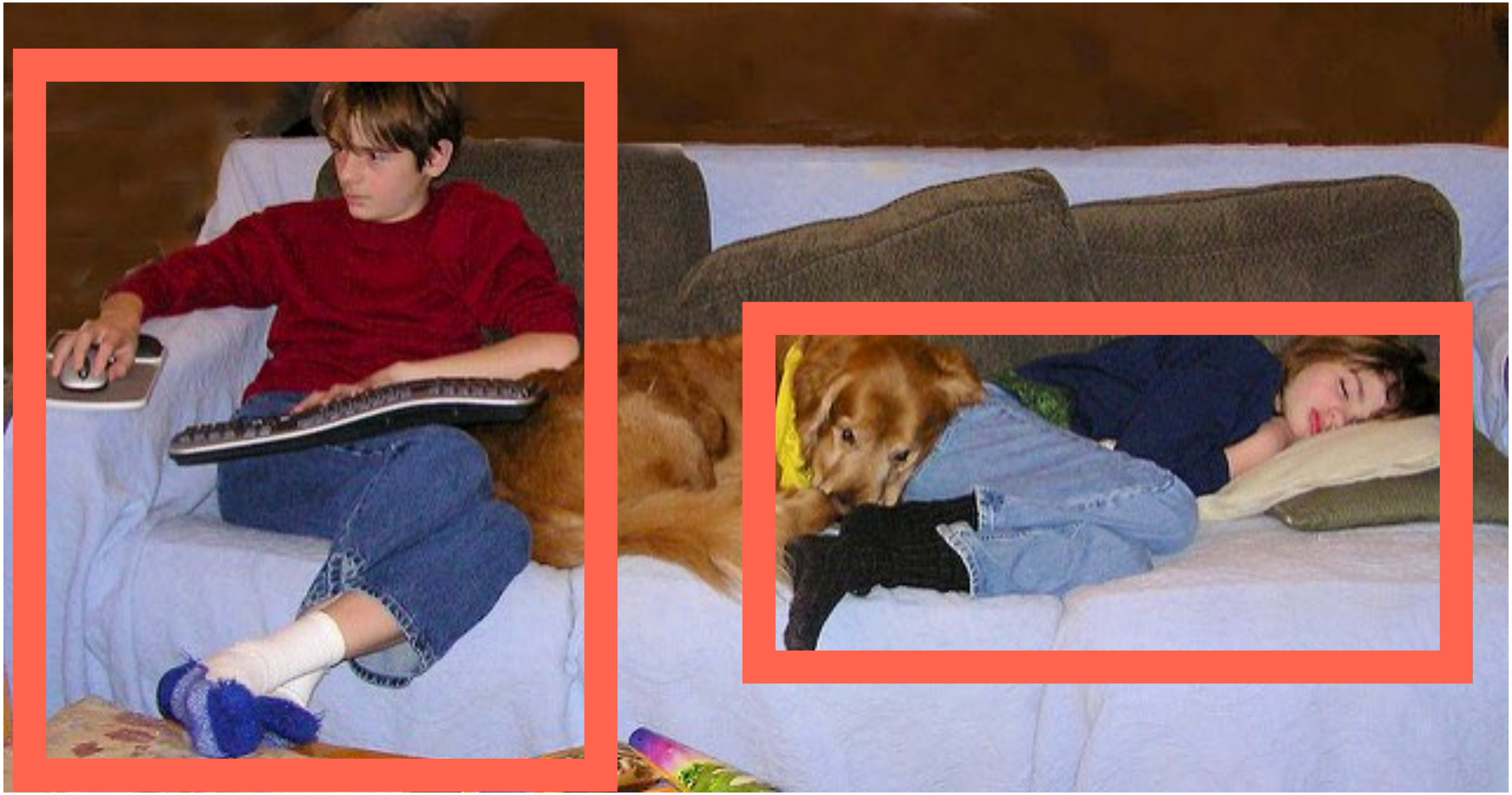
Discussion

1) What does it mean to understand this picture?

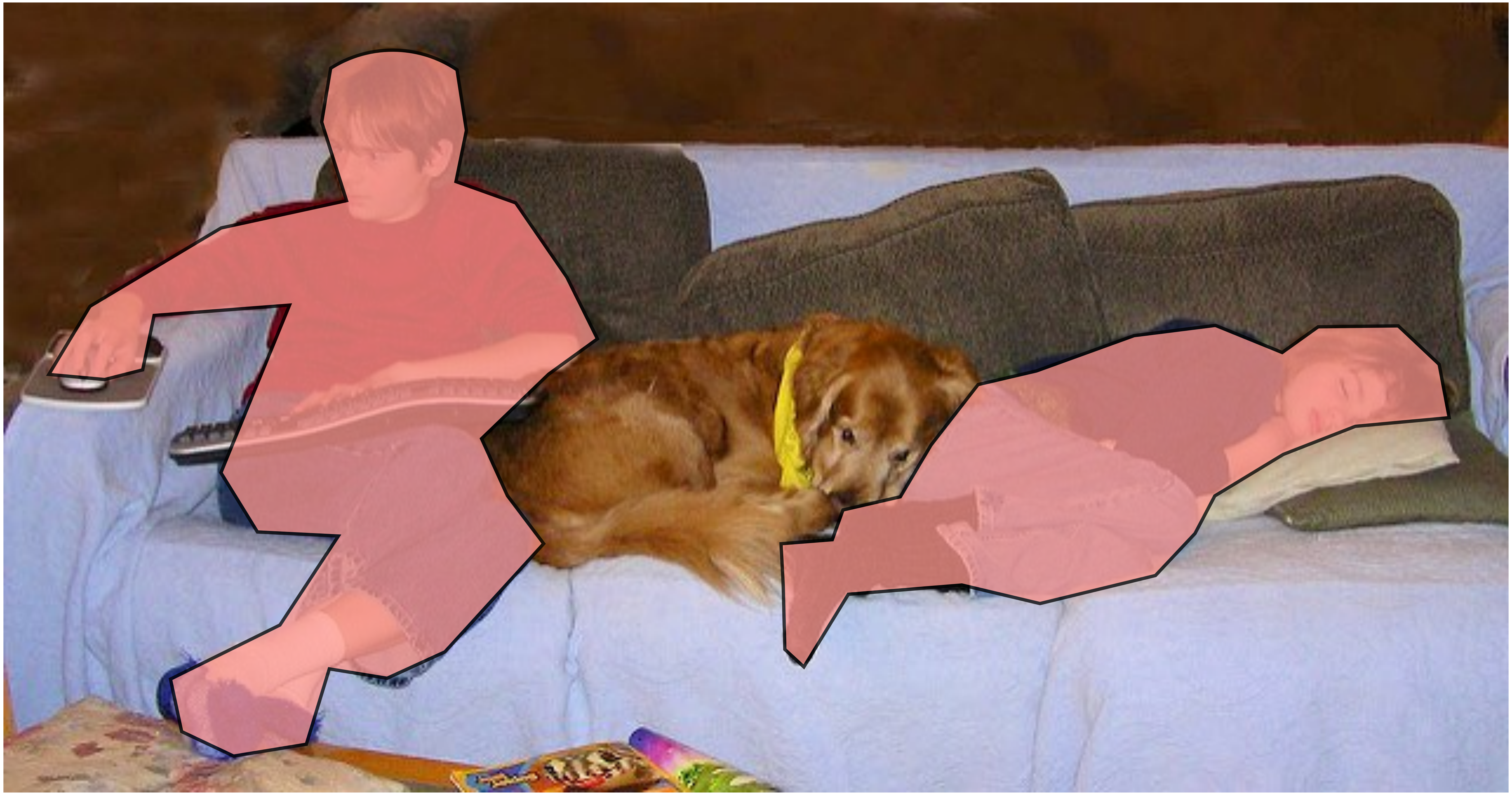
2) How to make software understand this picture?

Classification:
Is there a dog in this image?

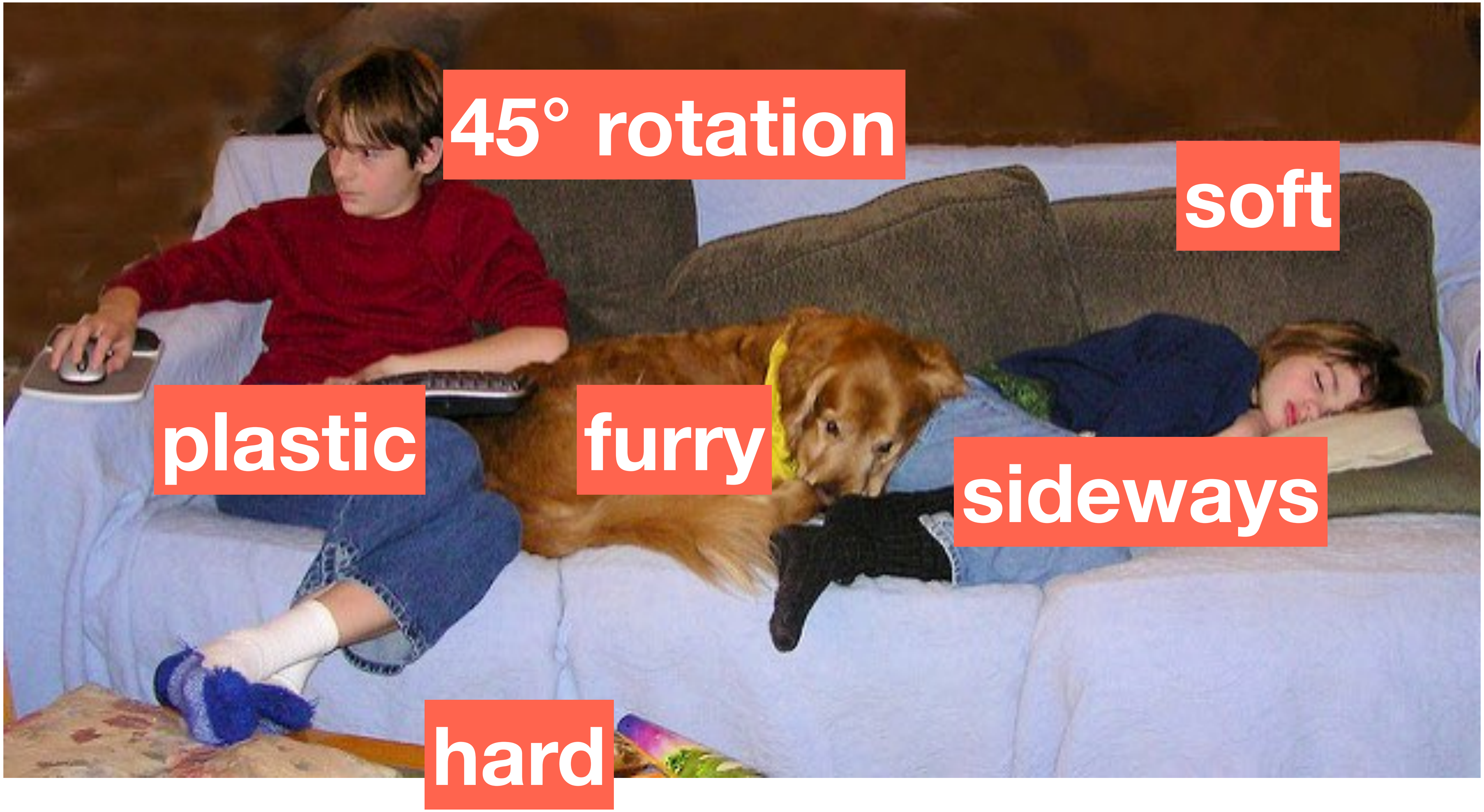
Detection:
Where are the people?



Segmentation:
Where ***really*** are the people?

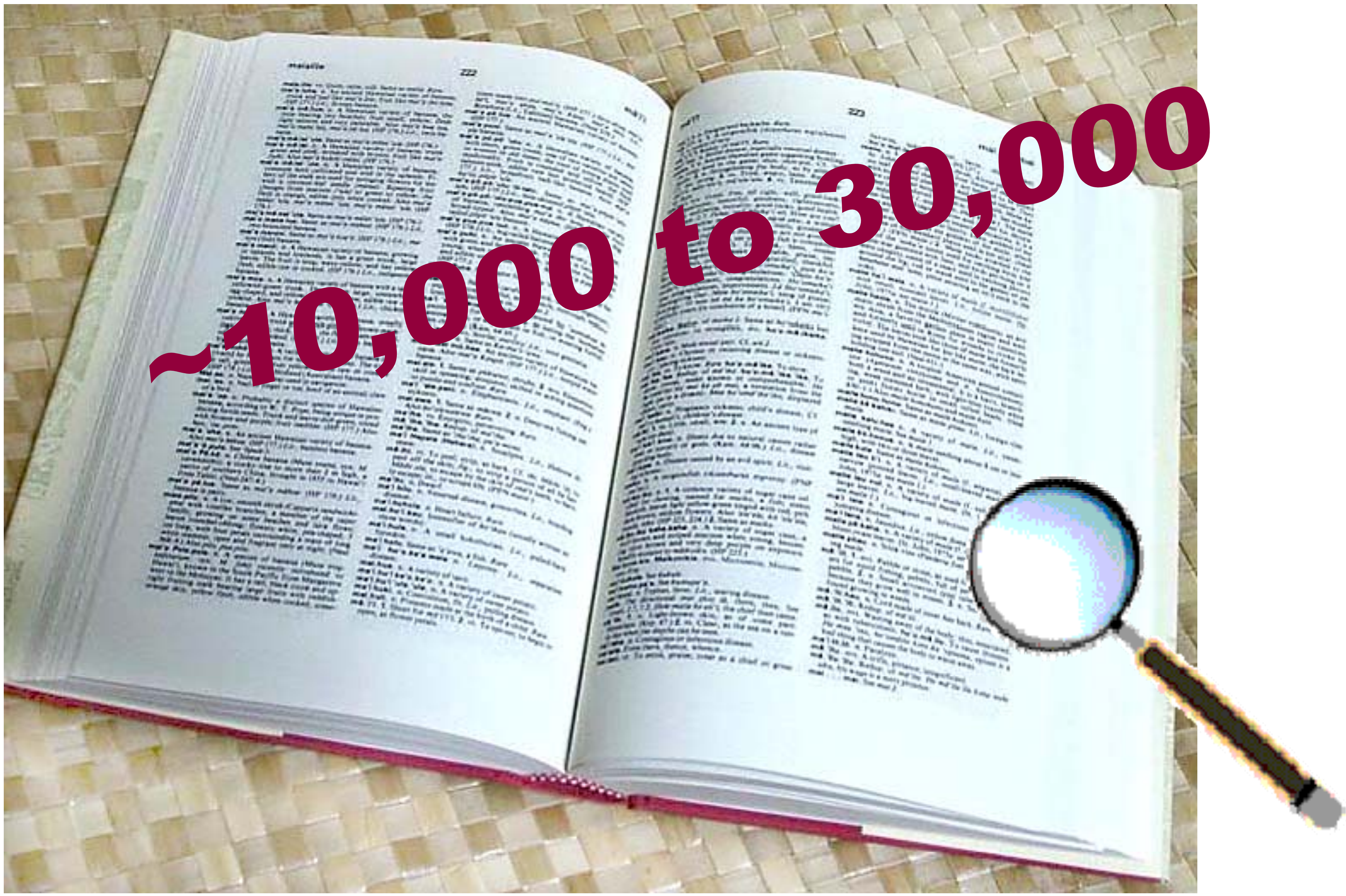


Attributes:
What features do objects have?



Actions:
What are they doing?

How many visual object categories are there?

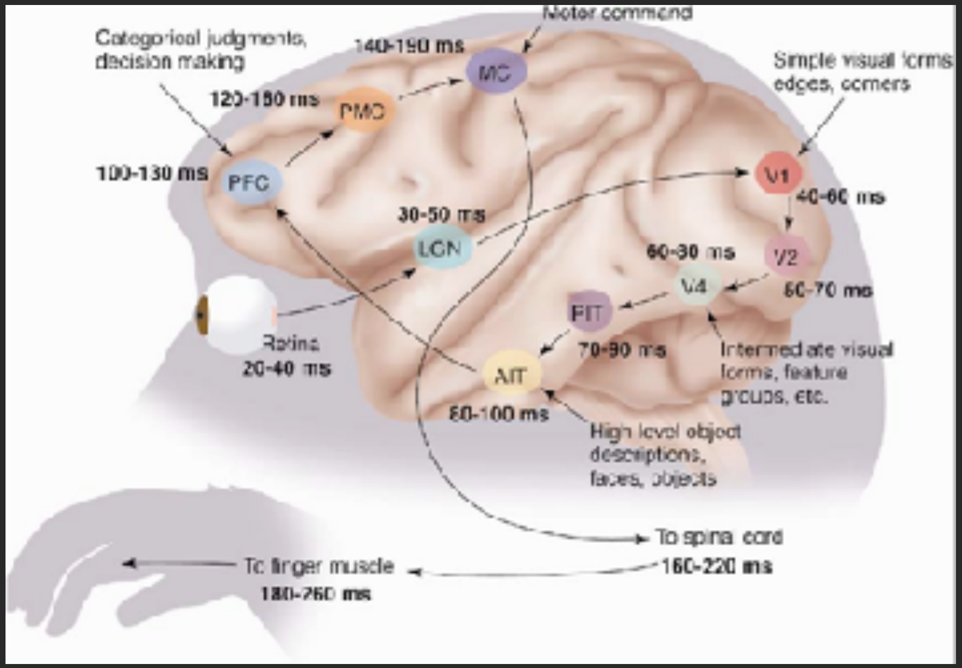


Biederman 1987


~10,000 to 30,000

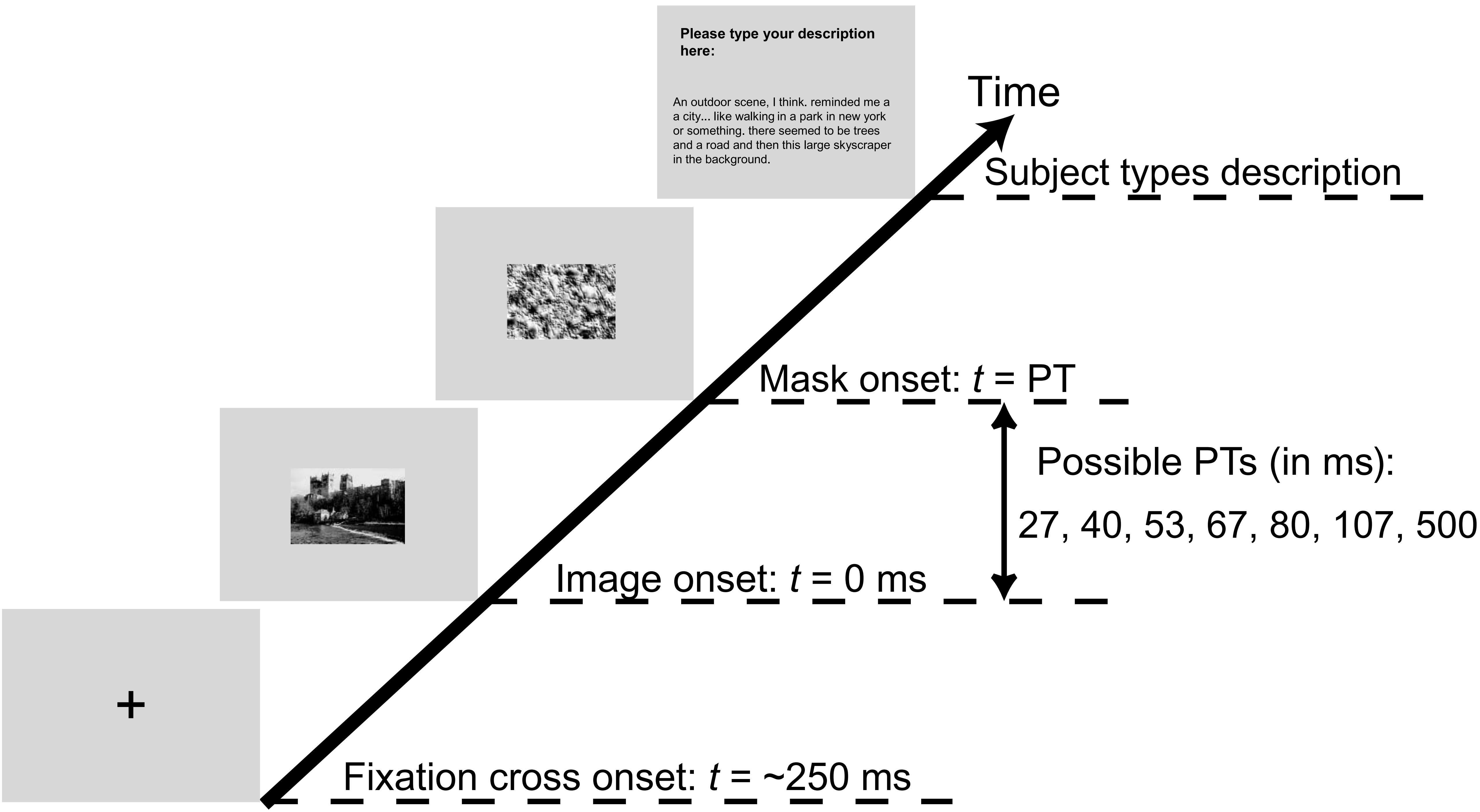
Rapid scene catgorization

People can distinguish high-level concepts (animal/transport) in under 150ms (Thorpe)



Appears to suggest feed-forward computations suffice (or at least dominate)

What do we perceive in a glance of a real-world scene?



PT = 107 ms

This is outdoors. A black, furry dog is running/walking towards the right of the picture. His tail is in the air and his mouth is open. Either he had a ball in his mouth or he was chasing after a ball. (Subject EC)

Inside a house, like a living room, with chairs and sofas and tables, no ppl. (Subject HS)

PT = 500 ms

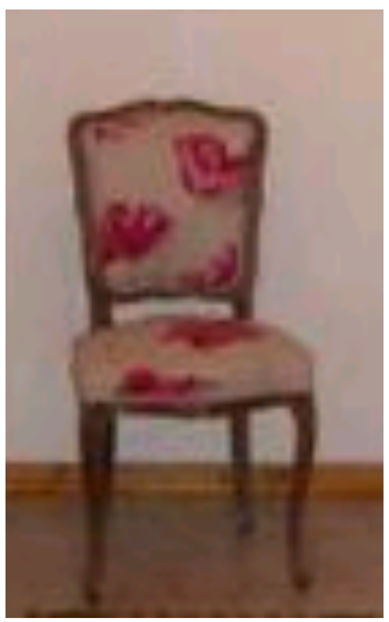
I saw a black dog carrying a gray frisbee in the center of the photograph. The dog was walking near the ocean, with waves lapping up on the shore. It seemed to be a gray day out. (Subject JB)

A room full of musical instruments. A piano in the foreground, a harp behind that, a guitar hanging on the wall (to the right). It looked like there was also a window behind the harp, and perhaps a bookcase on the left. (Subject RW)

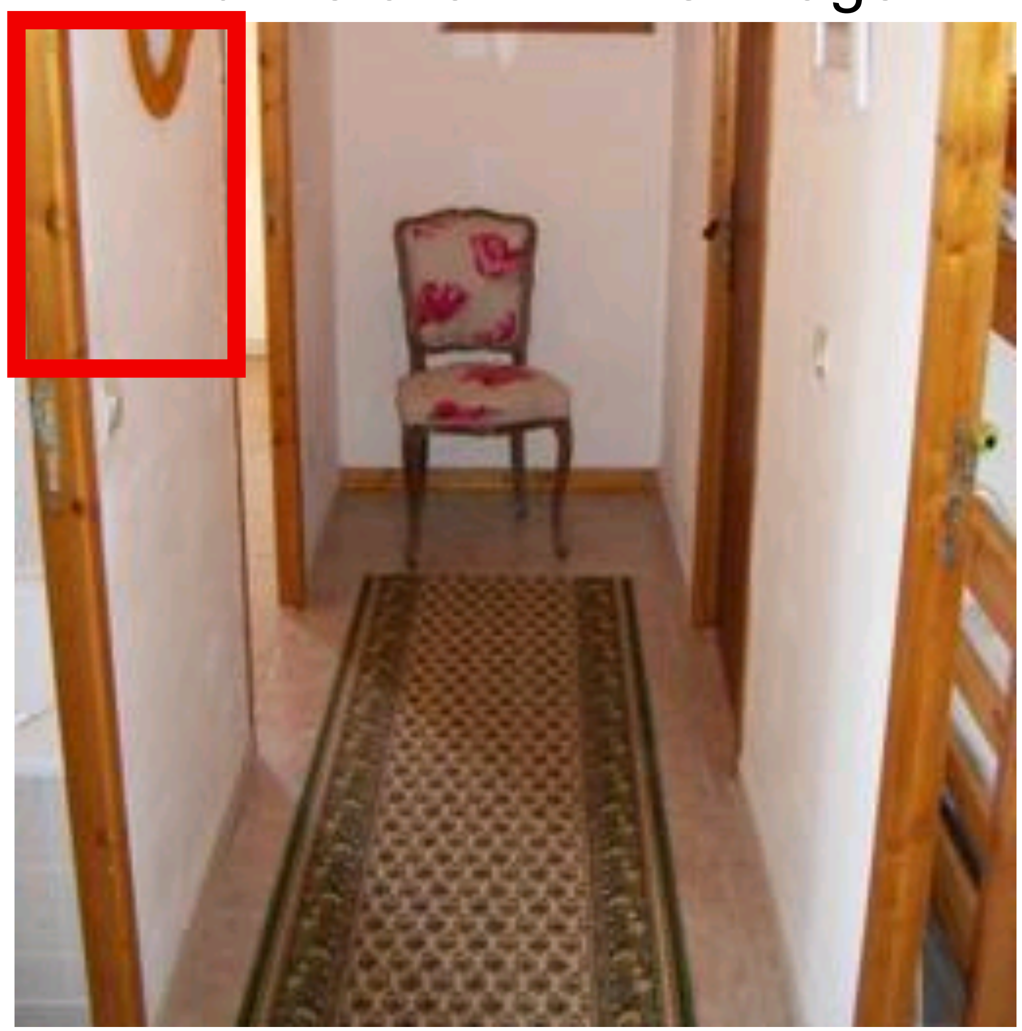
Should language be the right output?

Object recognition
Is it really so hard?

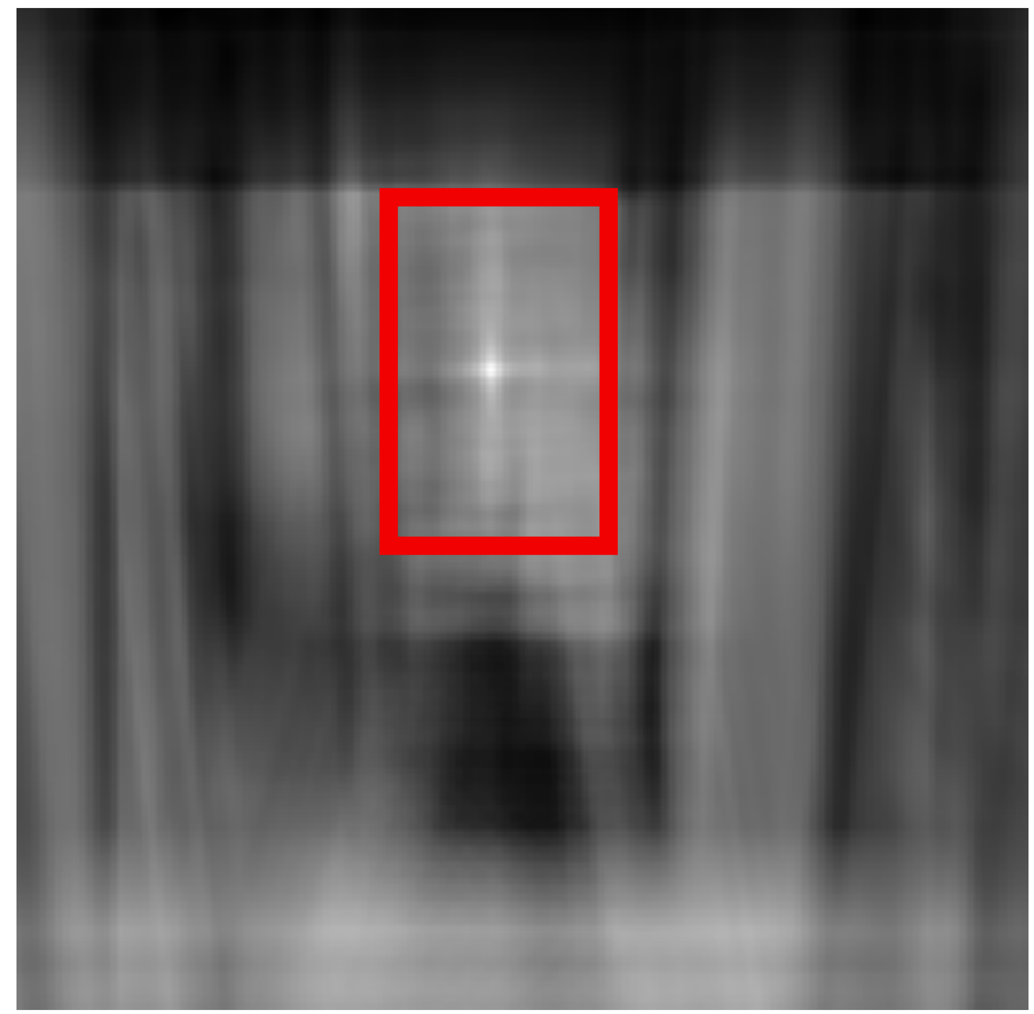
This is a chair

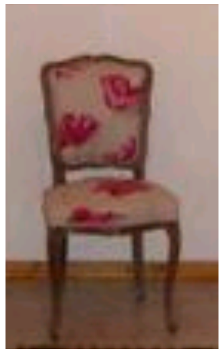


Find the chair in this image



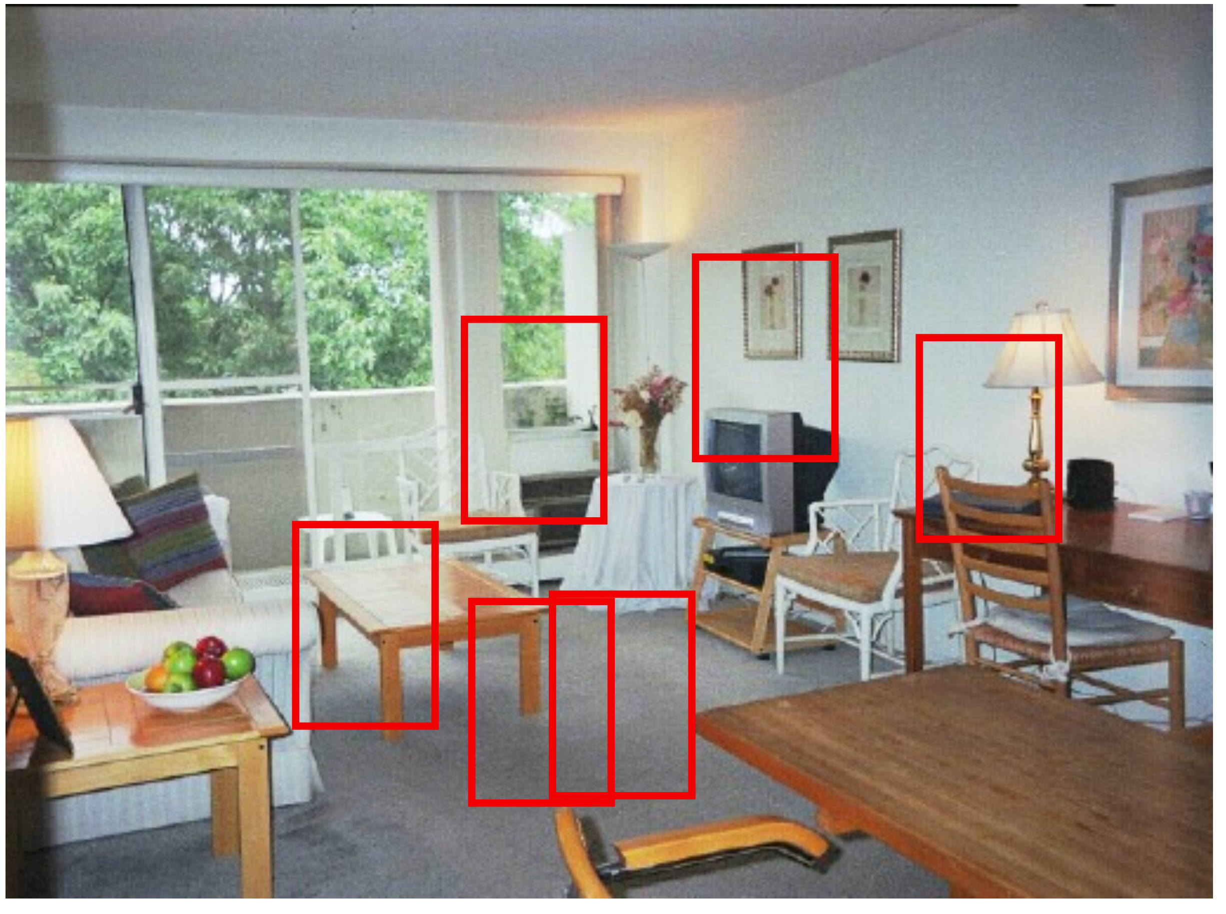
Output of normalized correlation

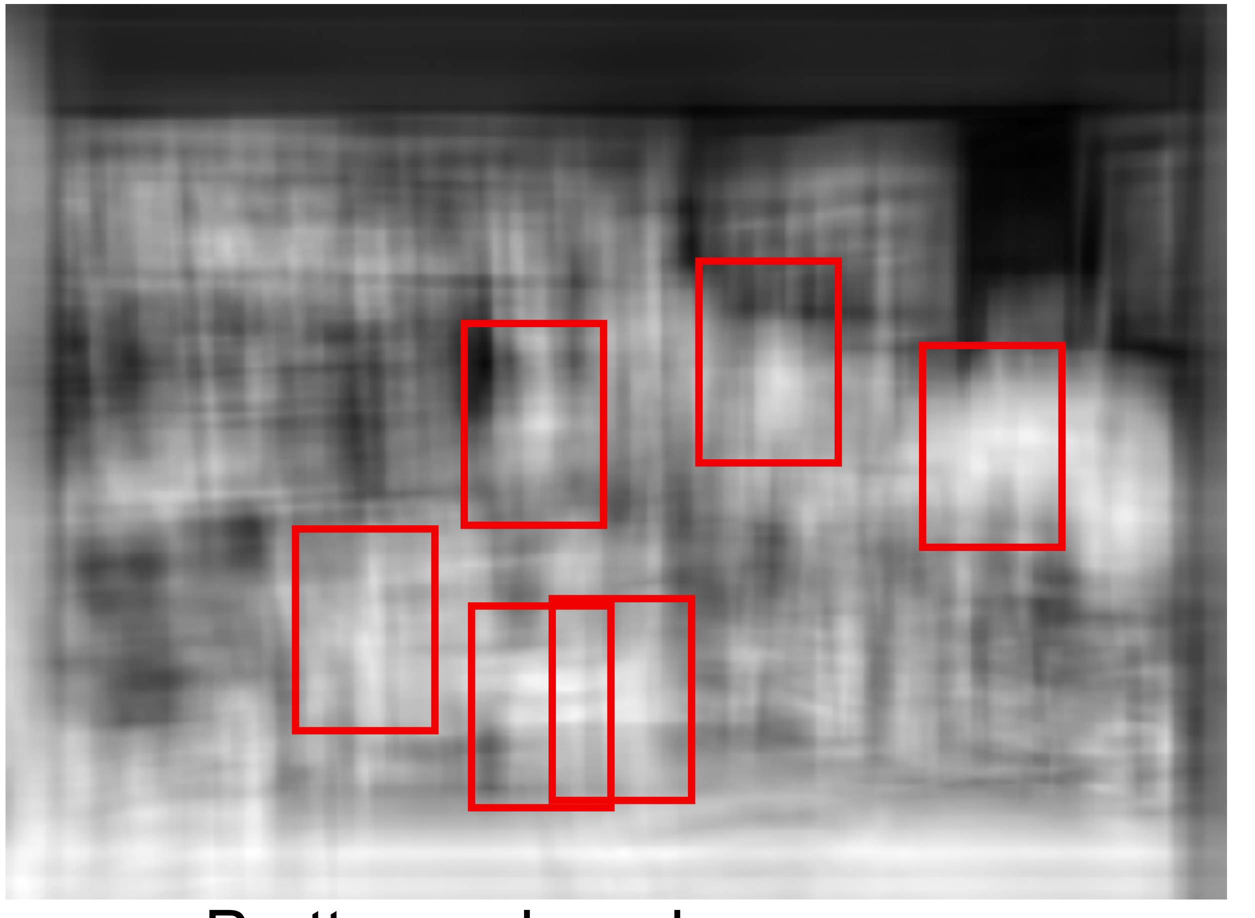




Object recognition
Is it really so hard?

Find the chair in this image



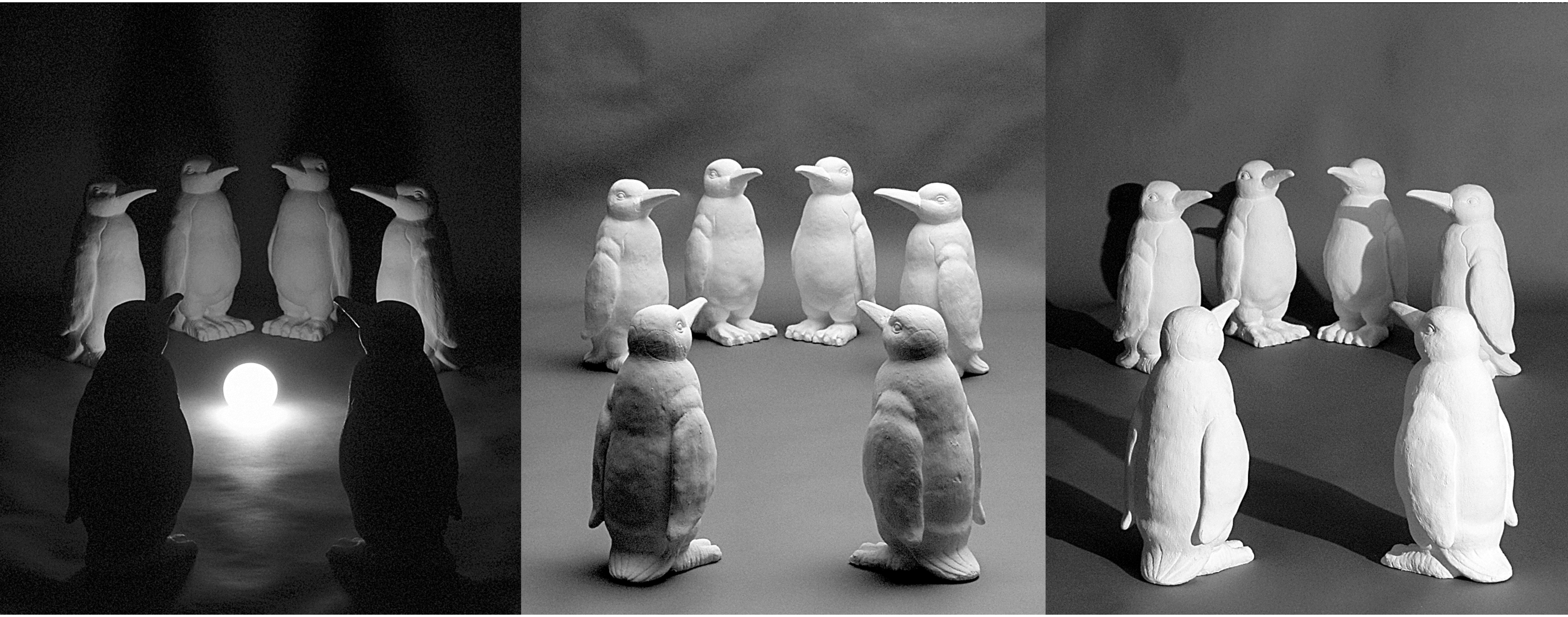


Pretty much garbage
Simple template matching is not going to make it

Challenges: viewpoint variation

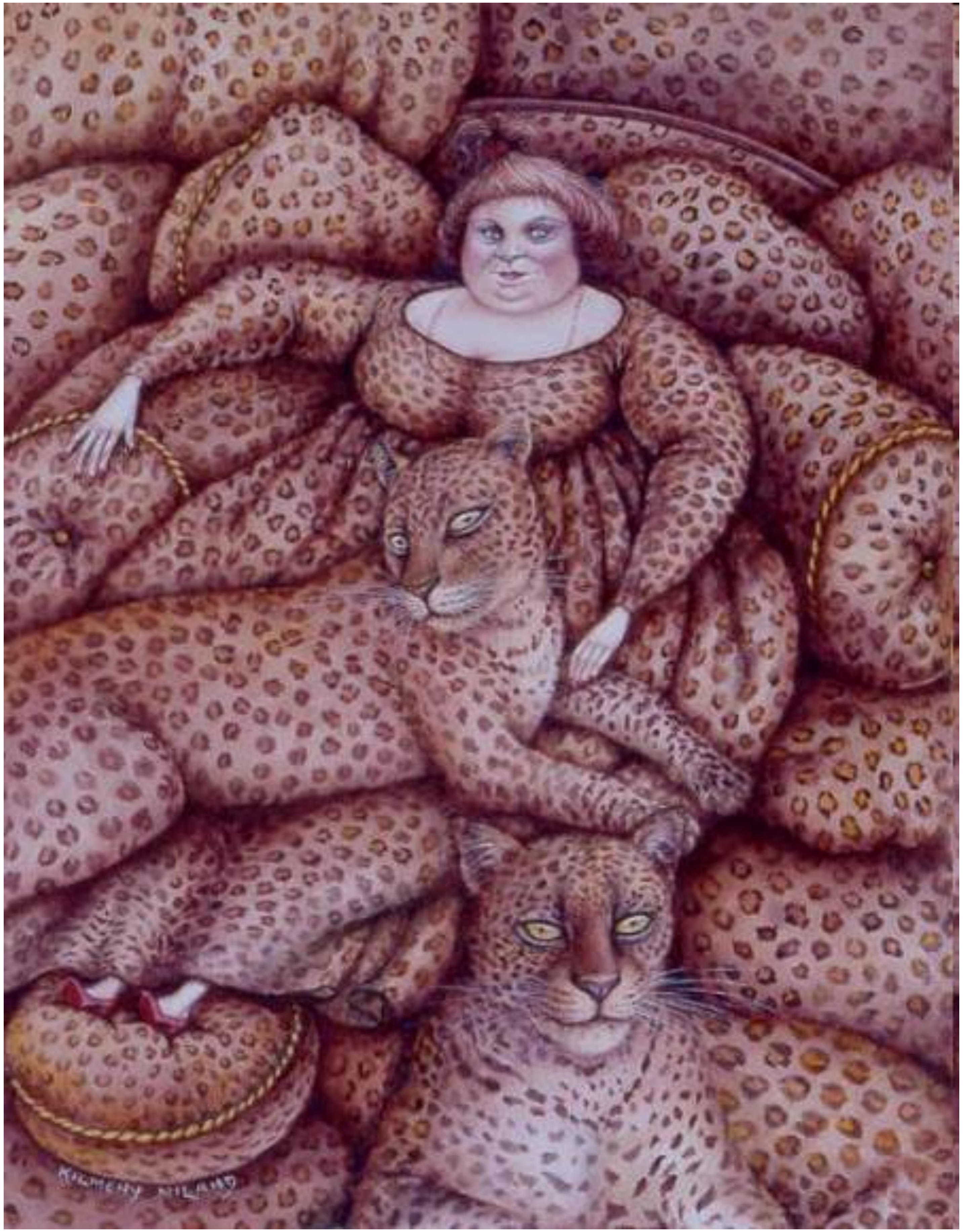
Michelangelo 1475-1564

Challenges: illumination



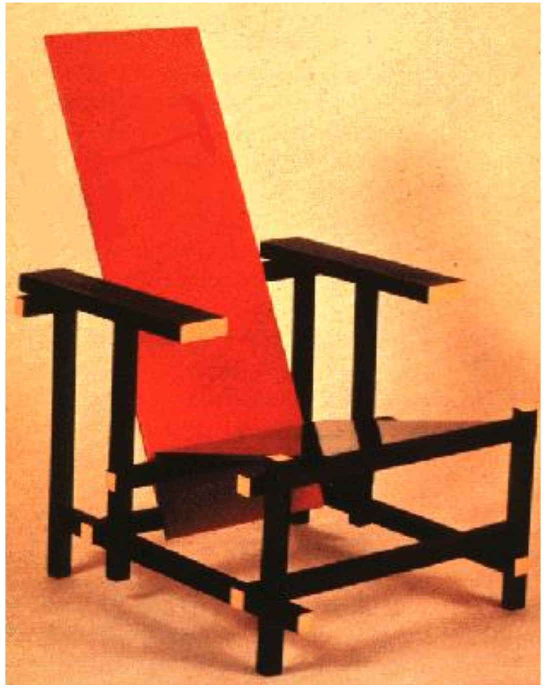
Challenges: scale

Challenges: background clutter



Kilmeny Niland. 1995

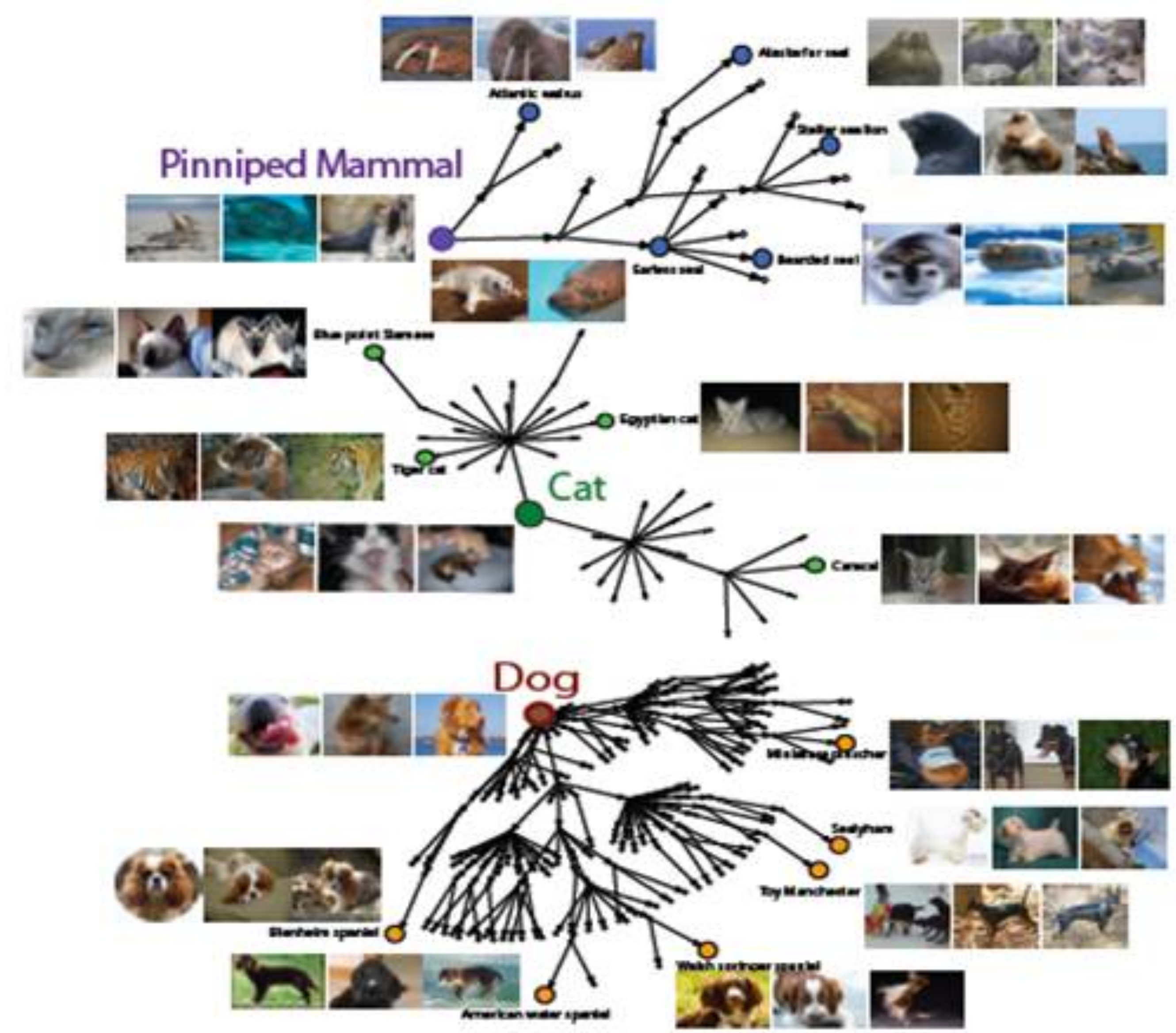
Within-class variations



Svetlana Lazebnik

Supervised Visual Recognition

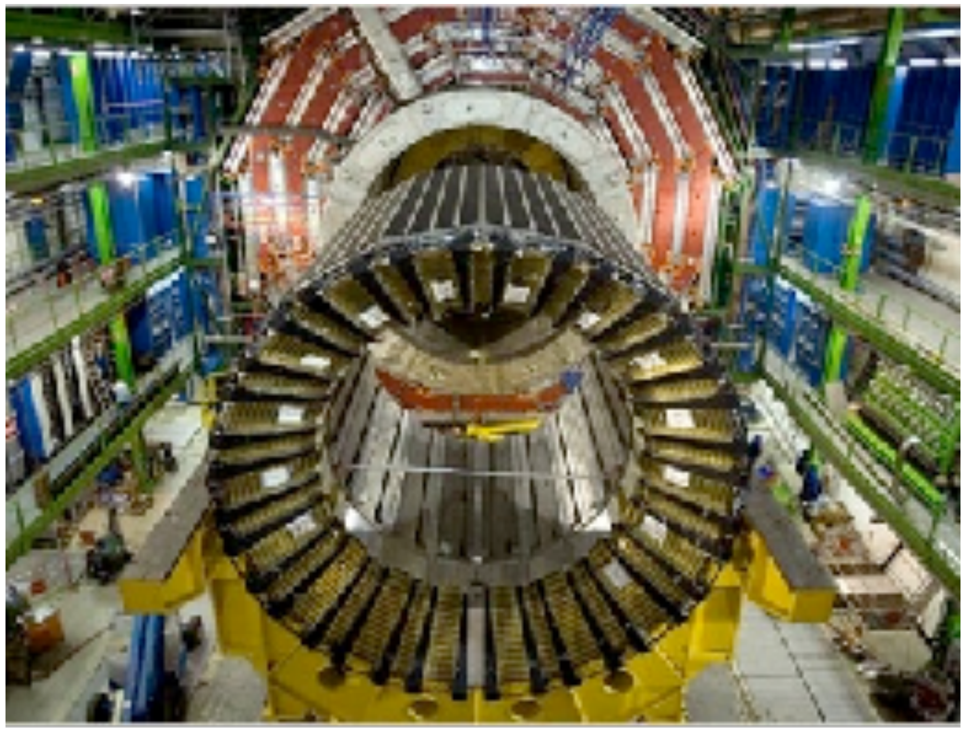
Can we define a canonical list of objects, attributes, actions, materials….?



ImageNet (cf. WordNet, VerbNet, FrameNet,..)

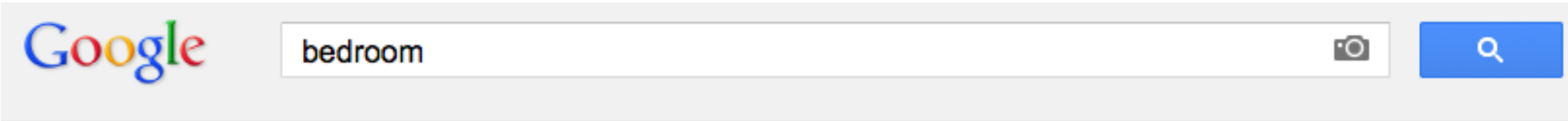
Crowdsourcing

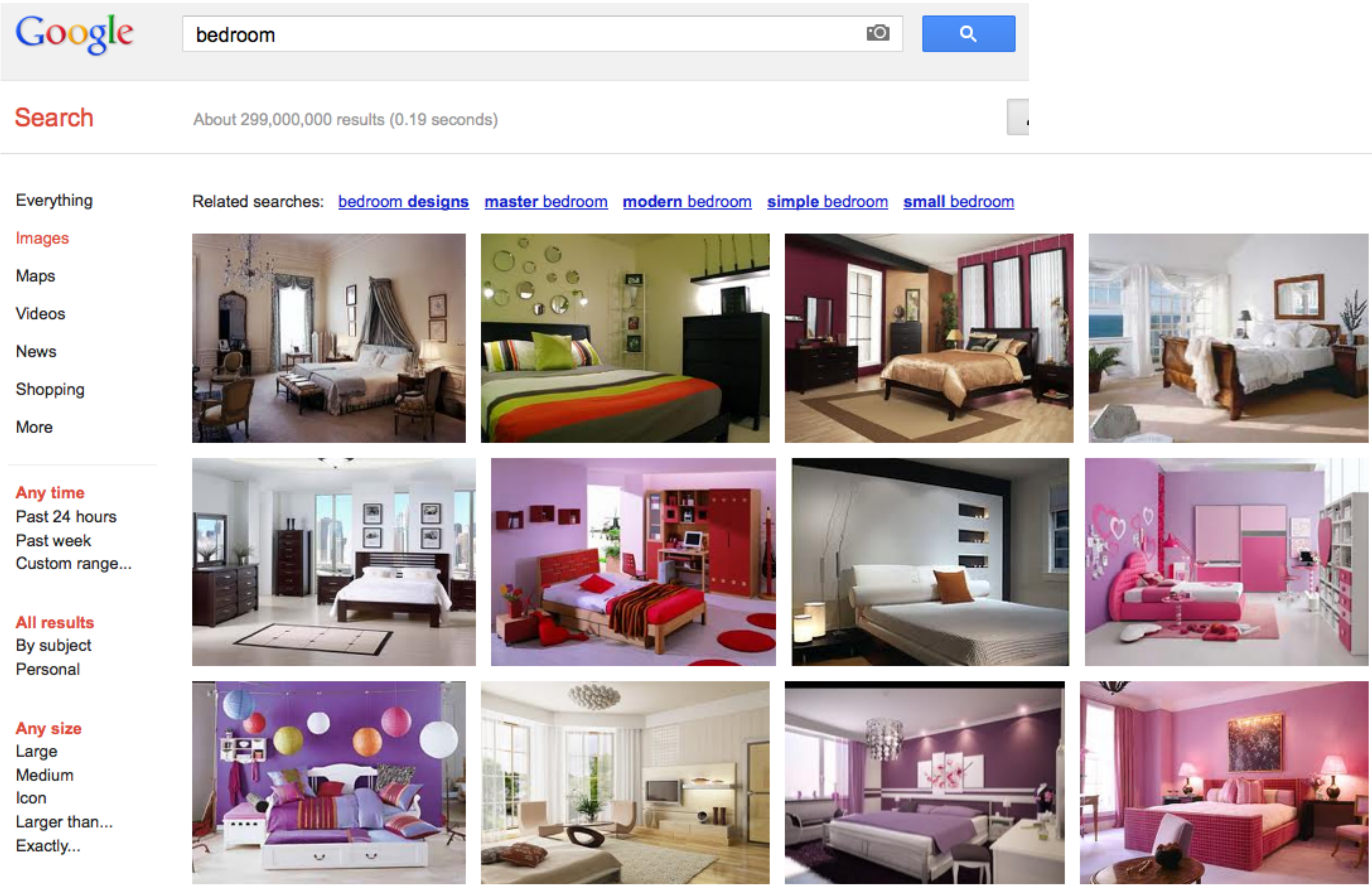
The value of data

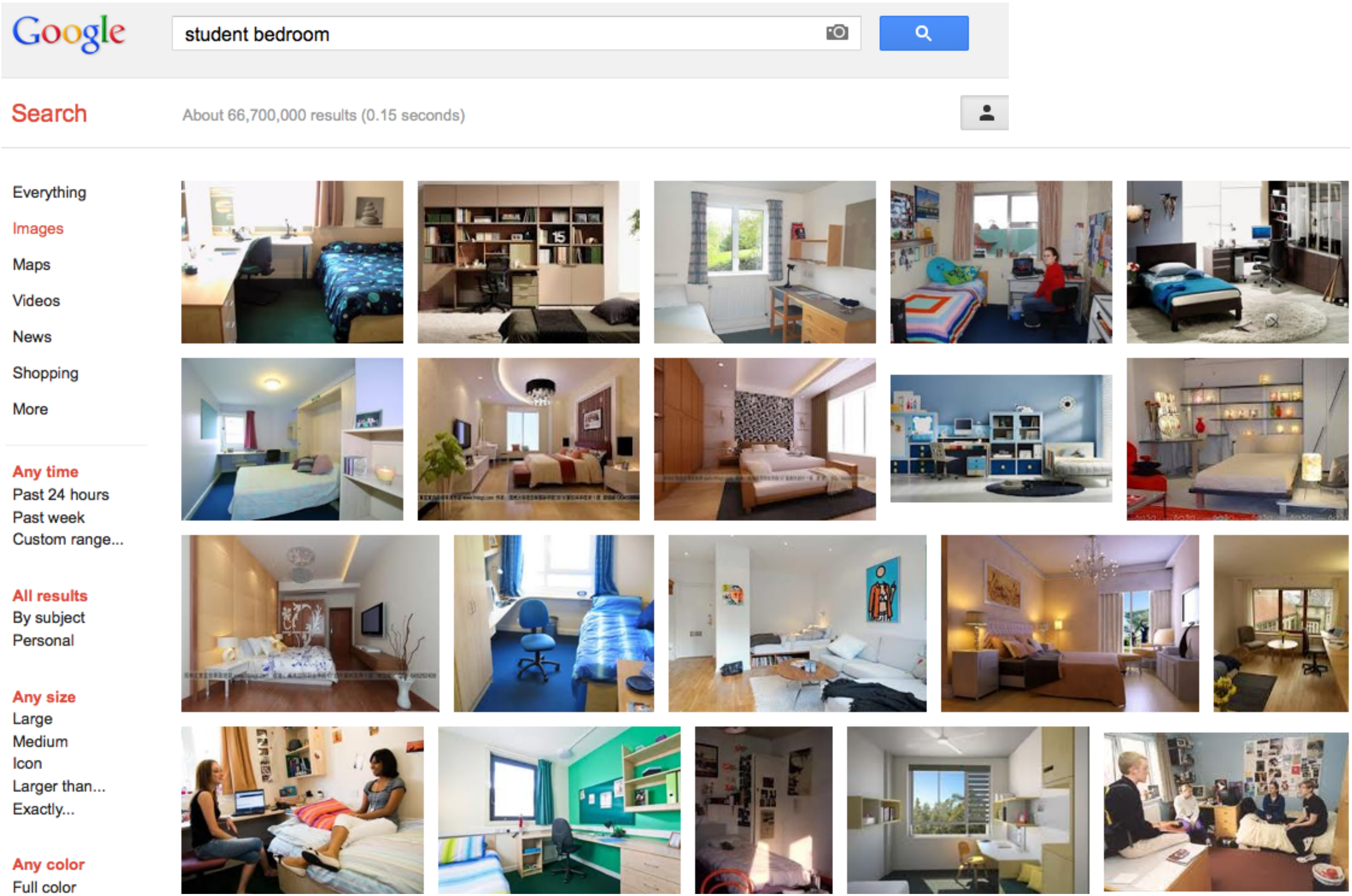


The Large Hadron Collider
$\$ 10^{10}$

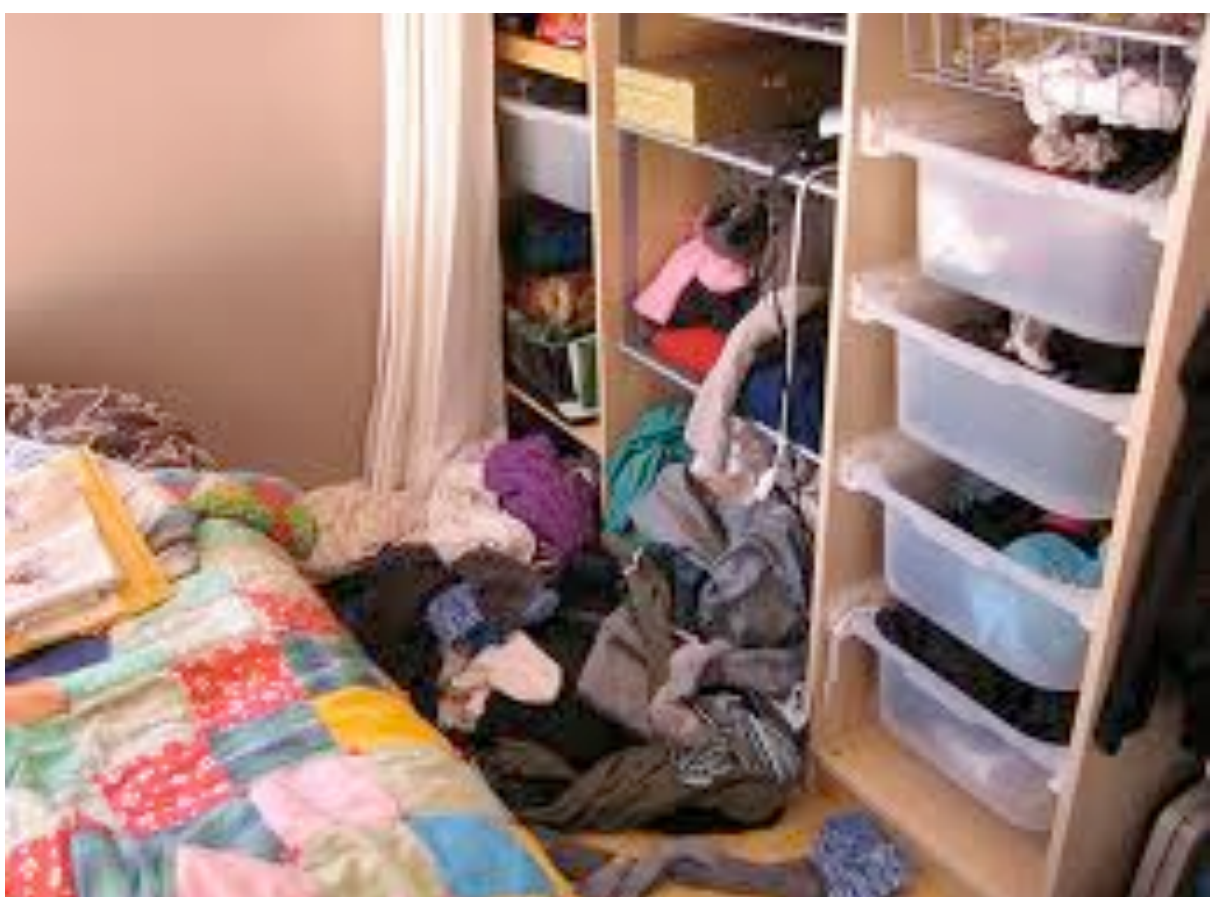
Amazon Mechanical Turk
$\$ 10^{2} - 10^{4}$


bedroom


Google
bedroom
Search
About 299,000,000 results (0.19 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Any time
Past 24 hours
Past week
Custom range...
All results
By subject
Personal
Any size
Large
Medium
Icon
Larger than...
Exactly...
Related searches: bedroom designs master bedroom modern bedroom simple bedroom small bedroom


Google
student bedroom
Search
About 66,700,000 results (0.15 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Any time
Past 24 hours
Past week
Custom range...
All results
By subject
Personal
Any size
Large
Medium
Icon
Larger than...
Exactly...
Any color
Full color


www.bigstock.com · 7067629




Google
mug
Mom

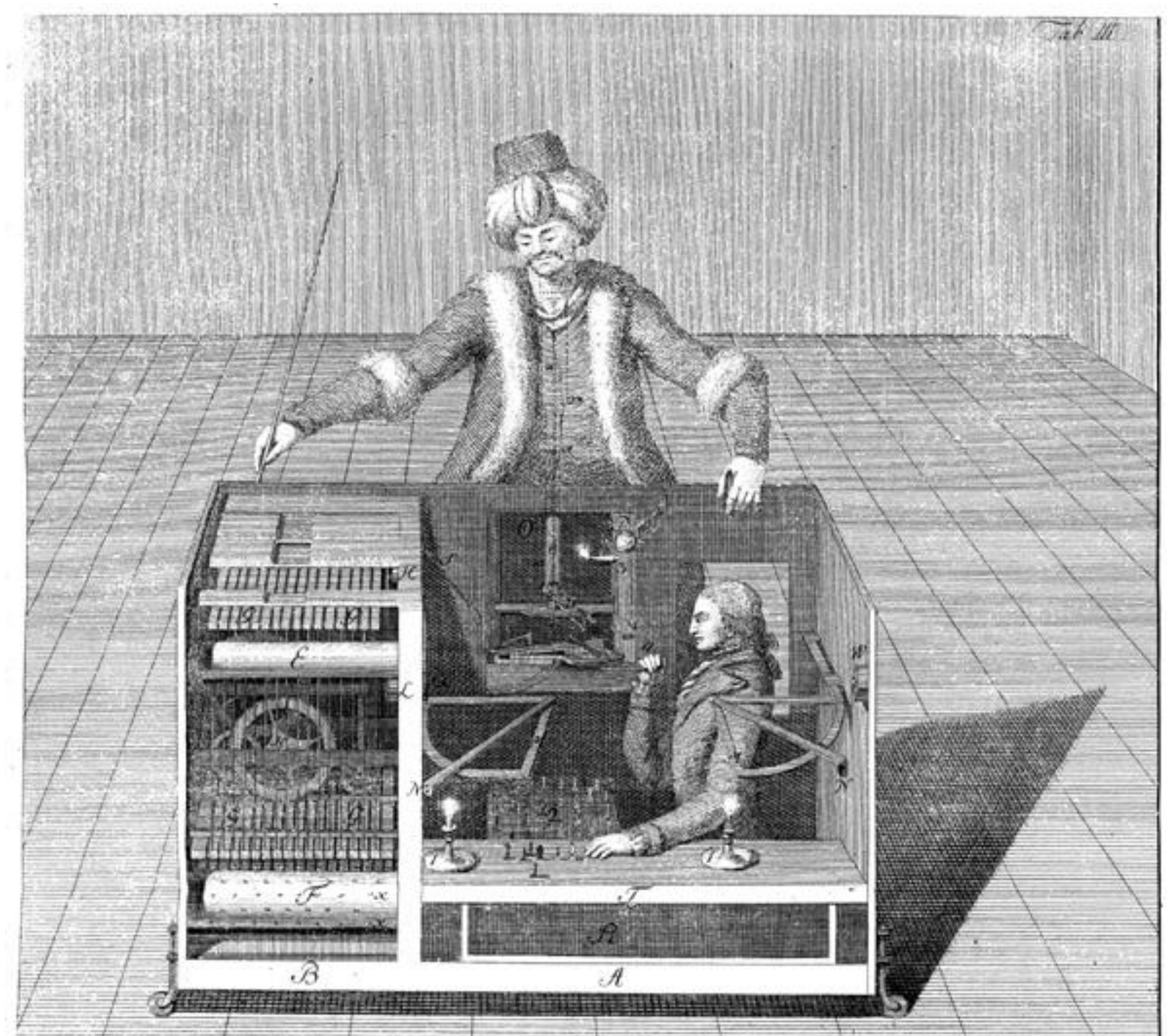
Mechanical Turk

- von Kempelen, 1770.
- Robotic chess player.
- Clockwork routines.
- Magnetic induction (not vision)

- Toured the world; played Napoleon Bonaparte and Benjamin Franklin.

Mechanical Turk

- It was all a ruse!
- Ho ho ho.



Amazon Mechanical Turk

Artificial artificial intelligence.

Launched 2005.
Small tasks, small pay.
Used extensively in data collection.

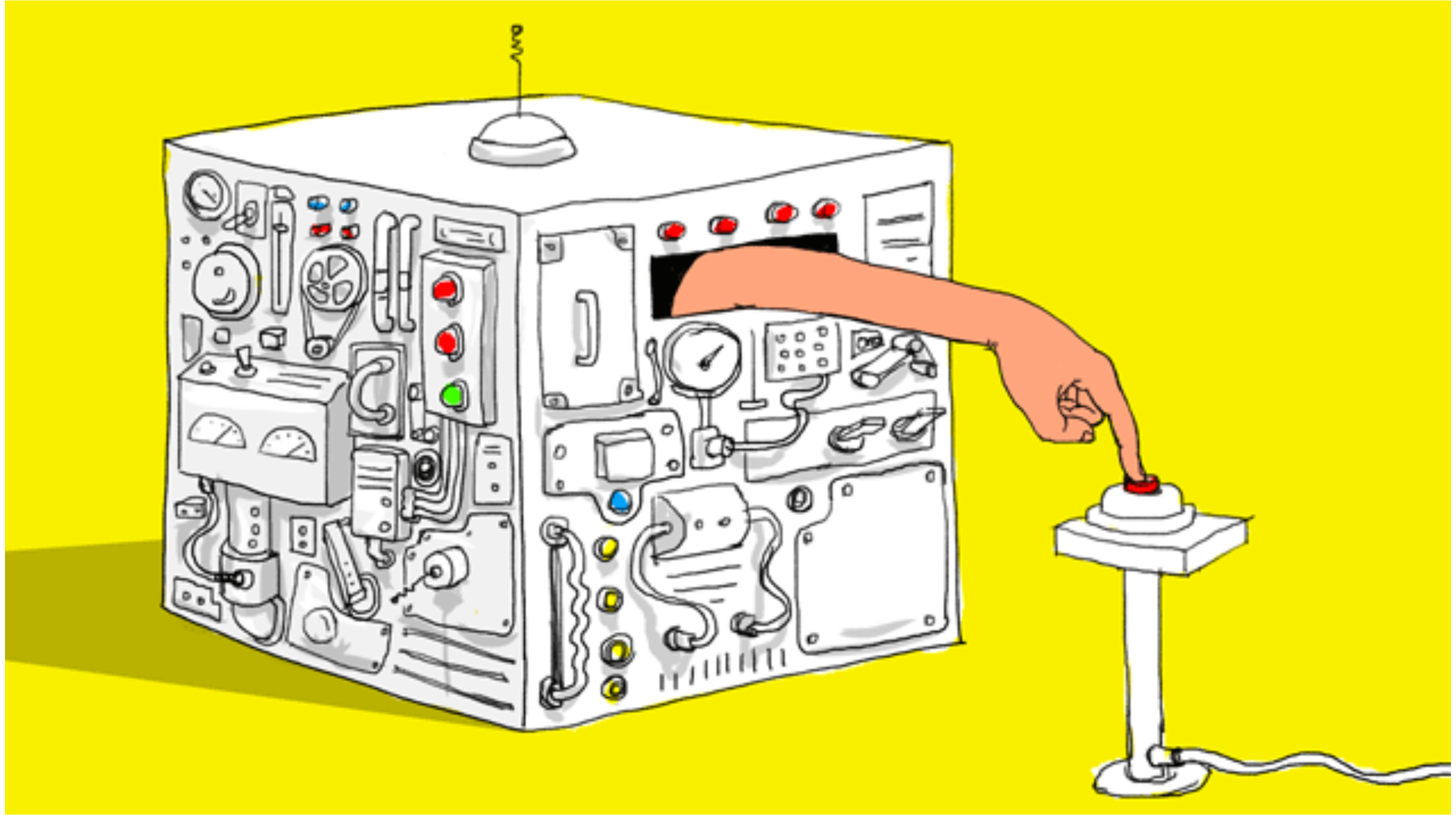


Image: Gizmodo

Beware of the human in your loop

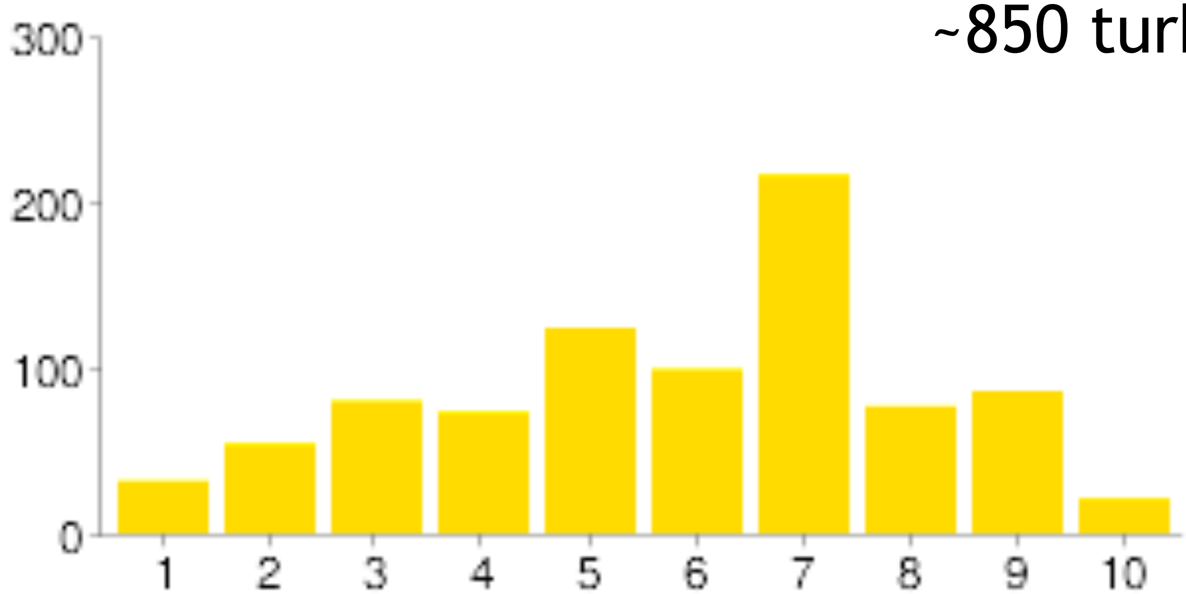
- What do you know about them?
- Will they do the work you pay for?

Let’s check a few simple experiments

Workers are given 1 cent to
randomly pick number between 1 and 10

Workers are given 1 cent to
randomly pick number between 1 and 10

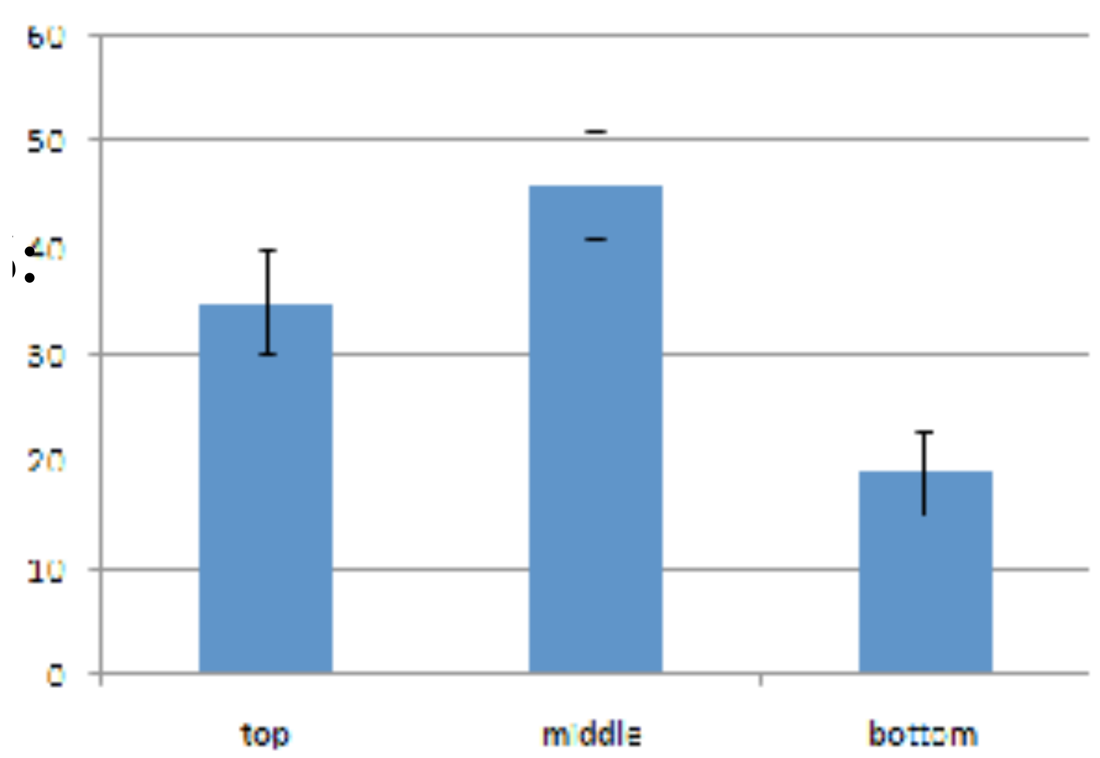
~850 turkers



Experiment by Greg Little
From http://groups.csail.mit.edu/uid/deneme/

Please choose one of the following:

Please choose one of the following:

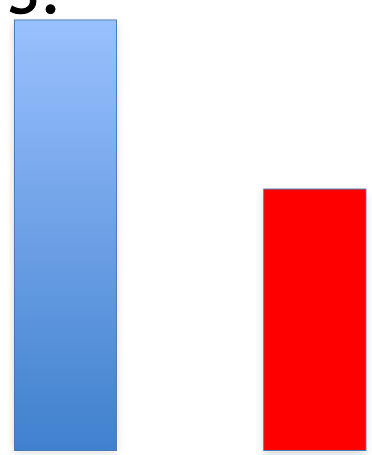


Experiment by Greg Little
From http://groups.csail.mit.edu/uid/deneme/

Please flip an actual coin and report the result

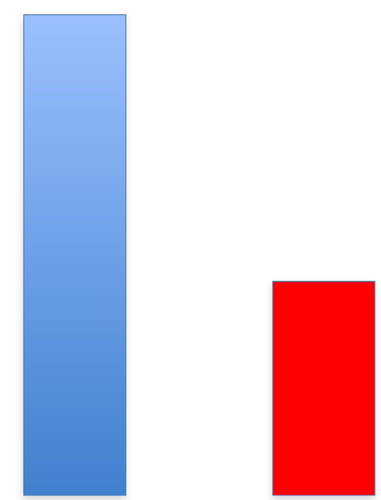
Please flip an actual coin and report the result

After 50 HITS:



31 heads, 19 tails

And 50 more:



34 heads, 16 tails

Experiment by Rob Miller
From http://groups.csail.mit.edu/uid/deneme/

Please click option B:

A

B

C

Please click option B:

A

B

C

Results of 100 HITS

A: 2
B: 96
C: 2

Experiment by Greg Little
From http://groups.csail.mit.edu/uid/deneme/

How do we annotate this?

Notes on image annotation

Adela Barriuso, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology

" I can see the ceiling, a wall and a ladder, but I do not know how to annotate what is on the right side of the picture. Maybe I just need to admit that I can not solve this picture in an easy and fast way. But if I was forced "

Semantic blindspots

Jia Deng, Fei-Fei Li, and many collaborators

What is WordNet?

Original paper by **[George Miller, et al 1990]** cited over 5,000 times

Organizes over 150,000 words into 117,000 categories called *synsets*.

Establishes ontological and lexical relationships in NLP and related tasks.

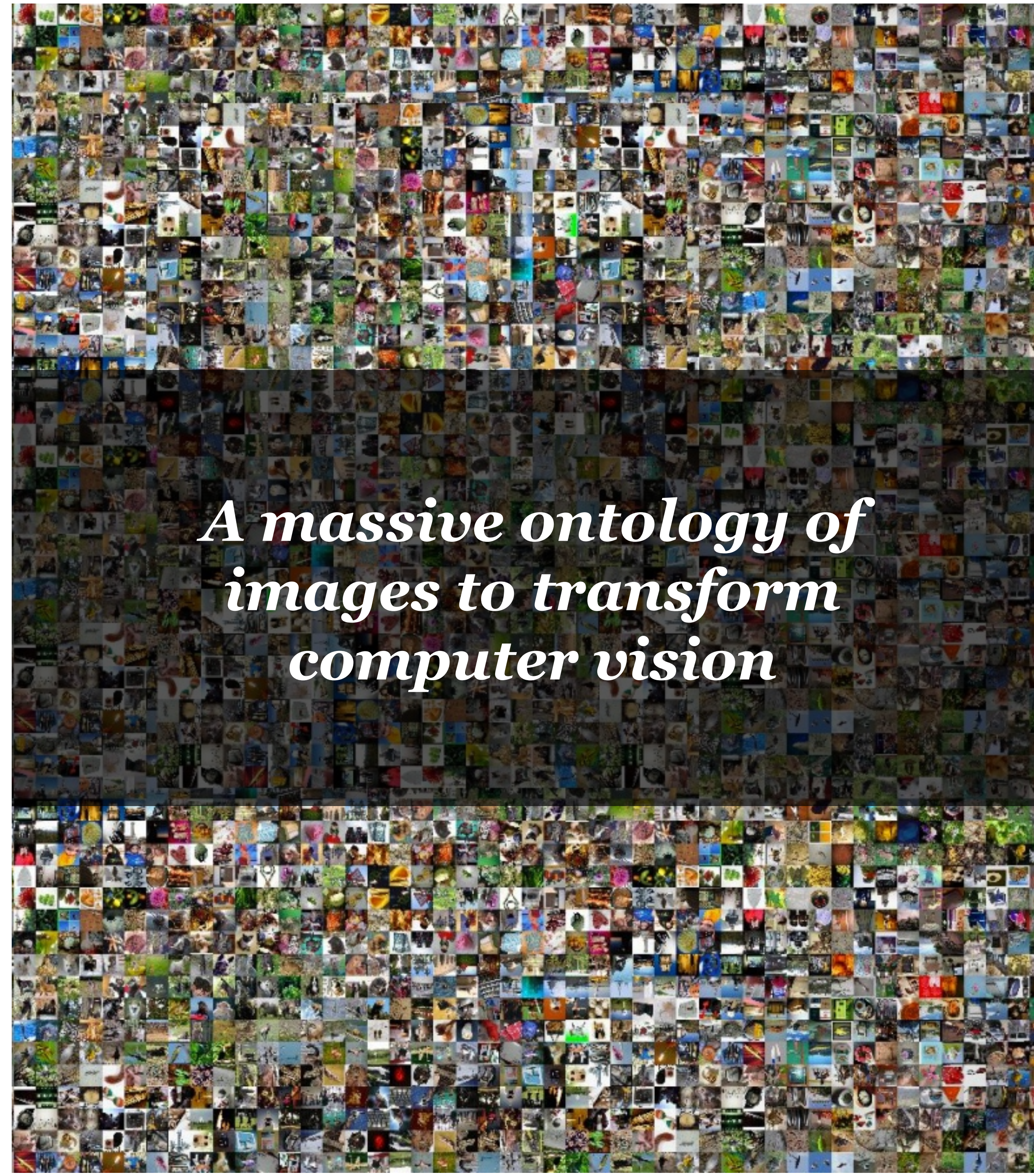
Individually Illustrated WordNet Nodes

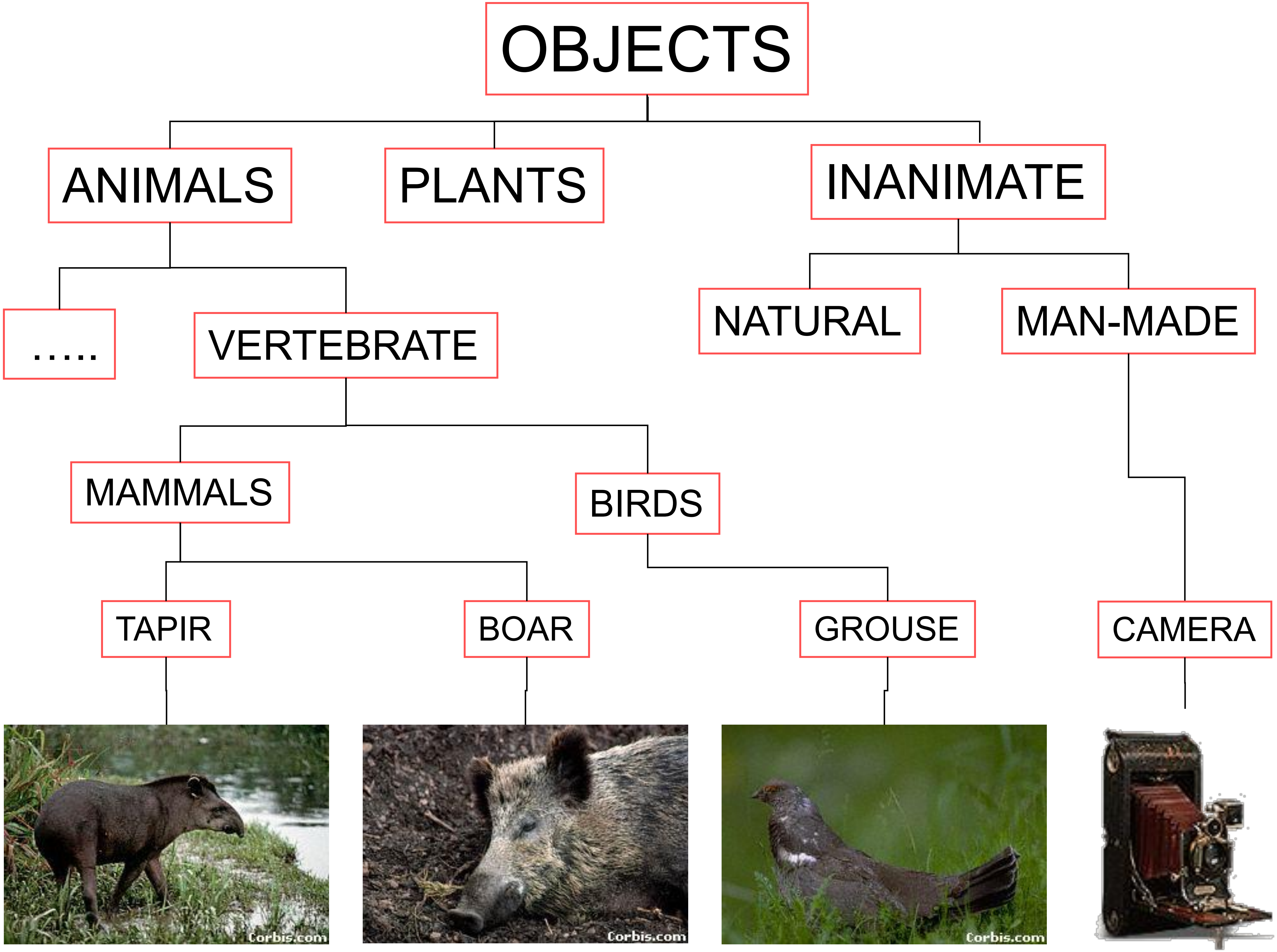
jacket: a short coat

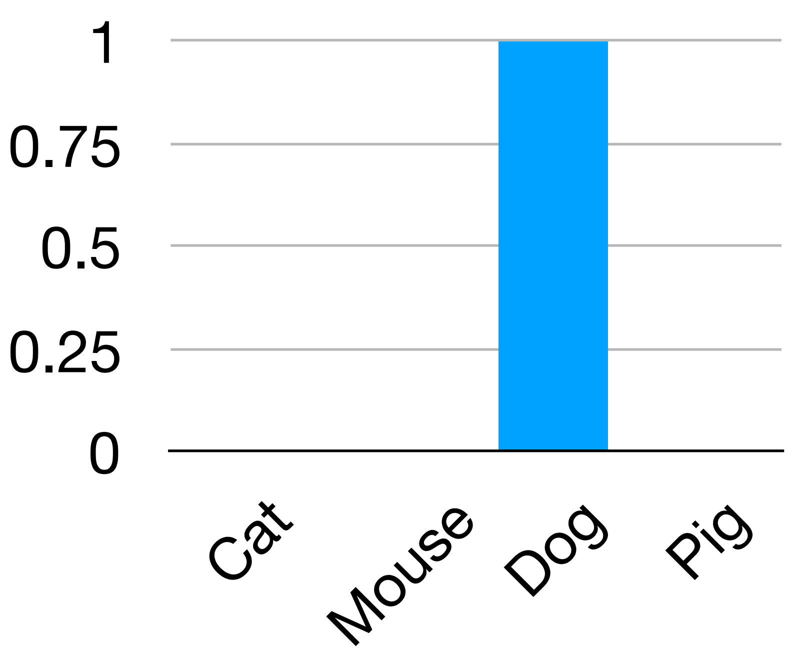
German shepherd: breed of large shepherd dogs used in police work and as a guide for the blind.

microwave: kitchen appliance that cooks food by passing an electromagnetic wave through it.

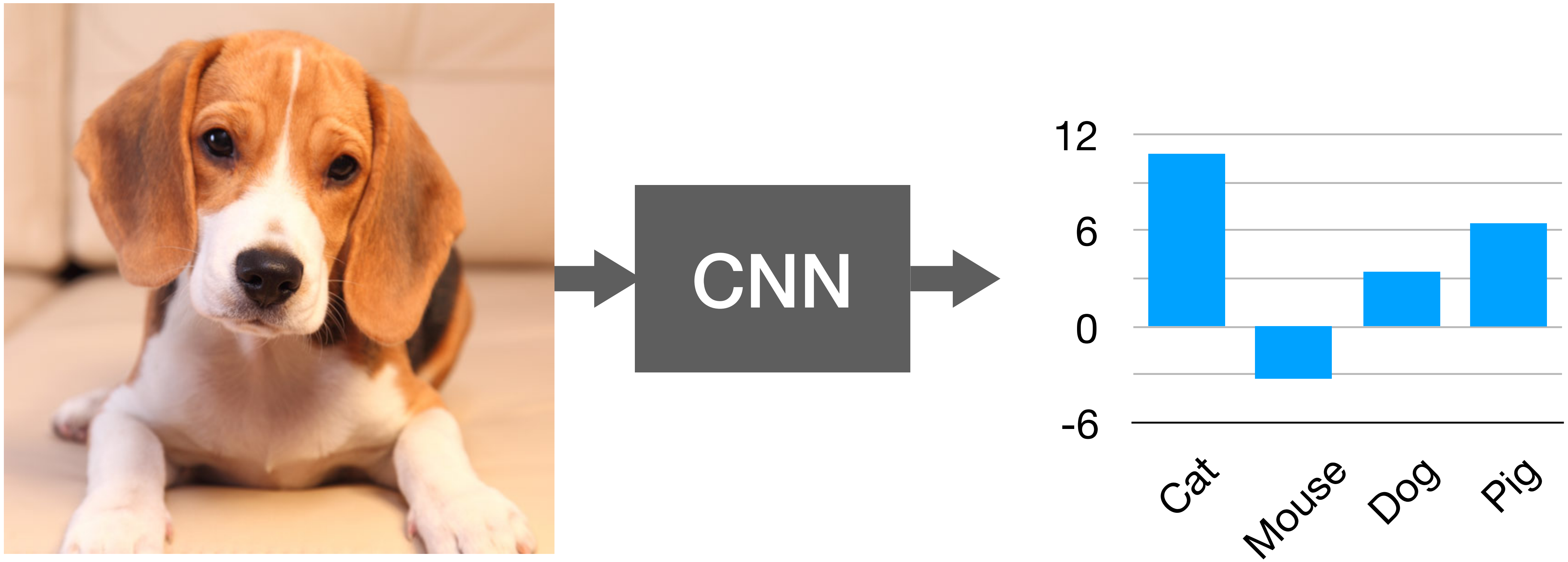
mountain: a land mass that projects well above its surroundings; higher than a hill.



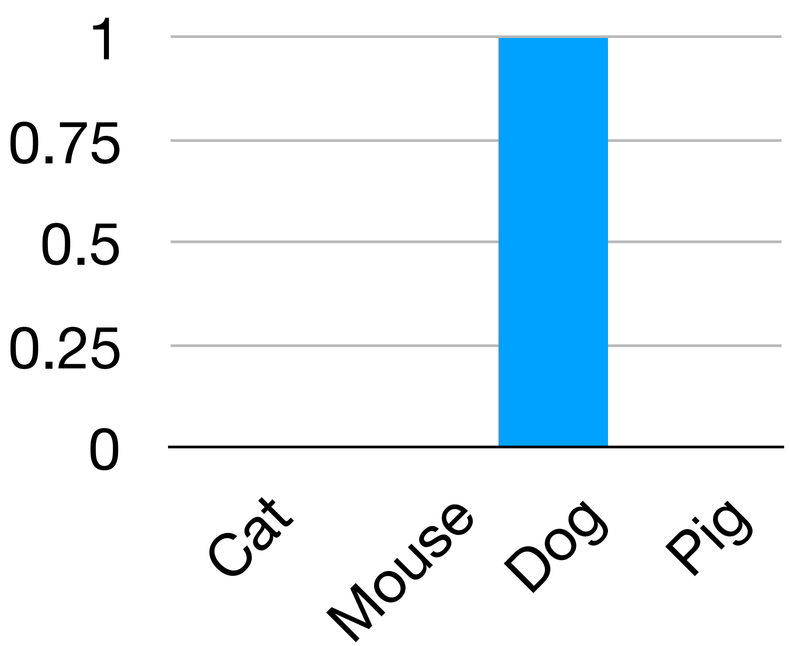

OBJECTS
ANIMALS
PLANTS
INANIMATE
.....
VERTEBRATE
NATURAL
MAN-MADE
MAMMALS
BIRDS
TAPIR
BOAR
GROUSE
CAMERA
Corbis.com
Corbis.com
Corbis.com



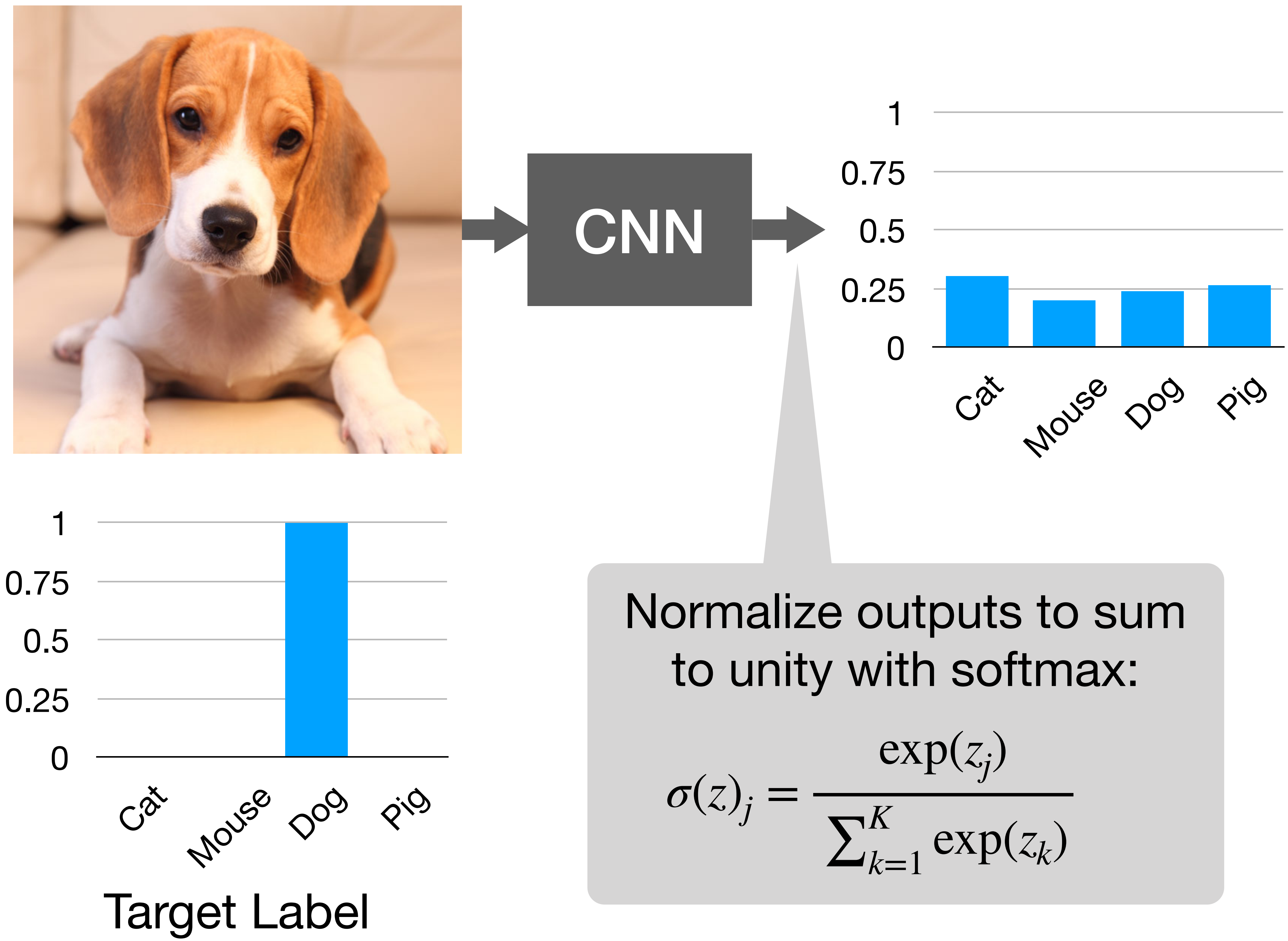
Target Label

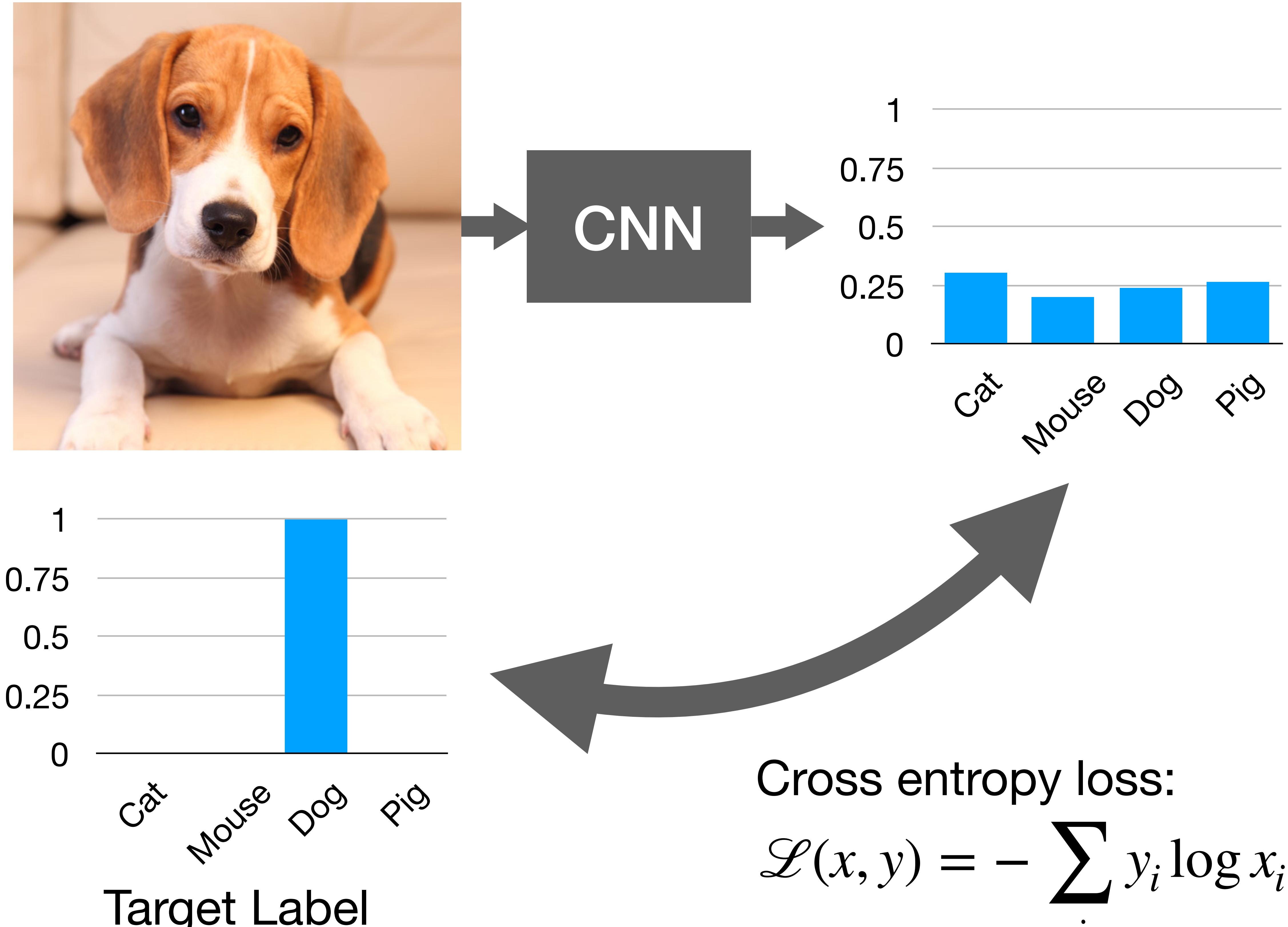


What’s wrong here?



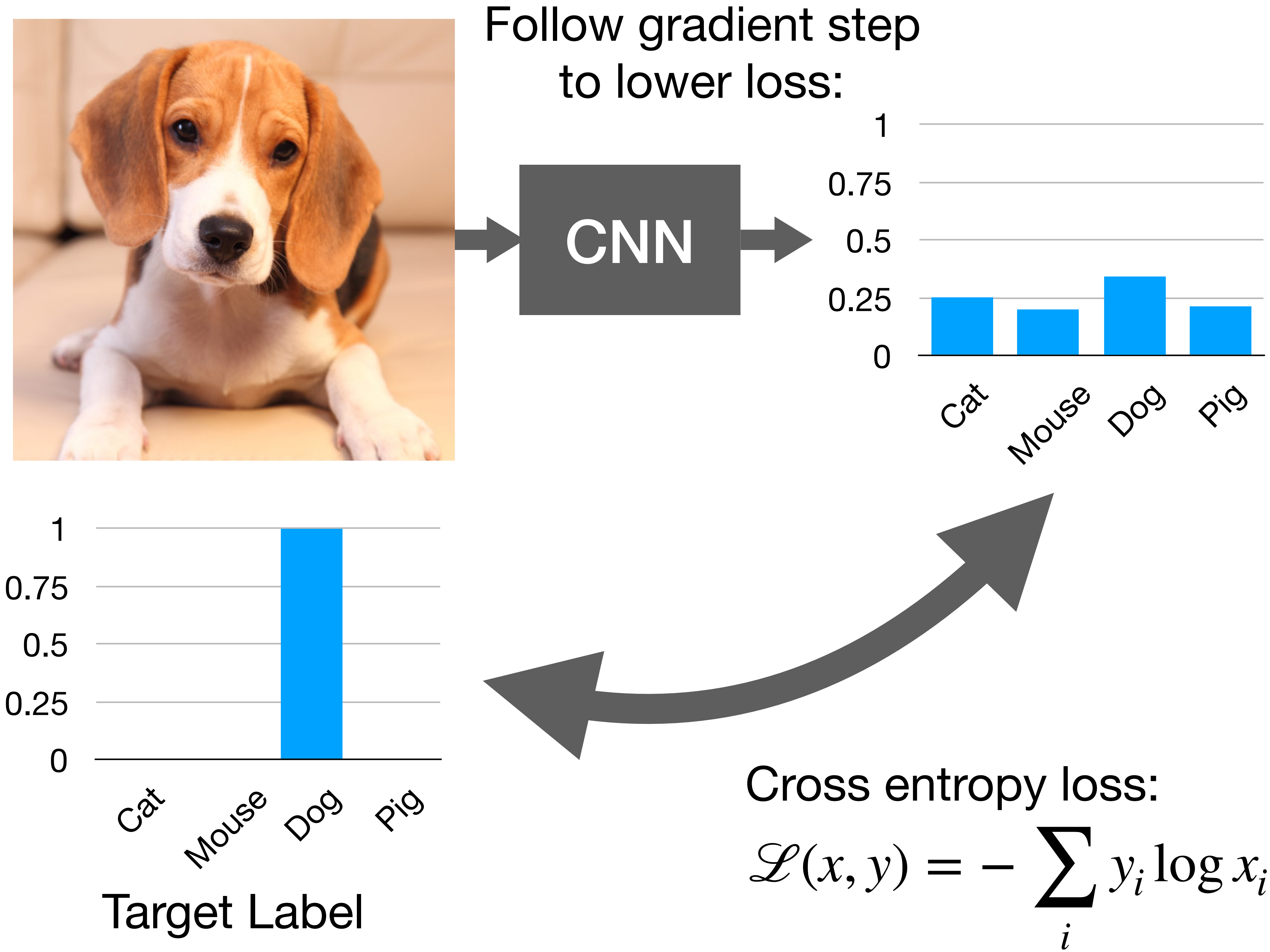
Target Label

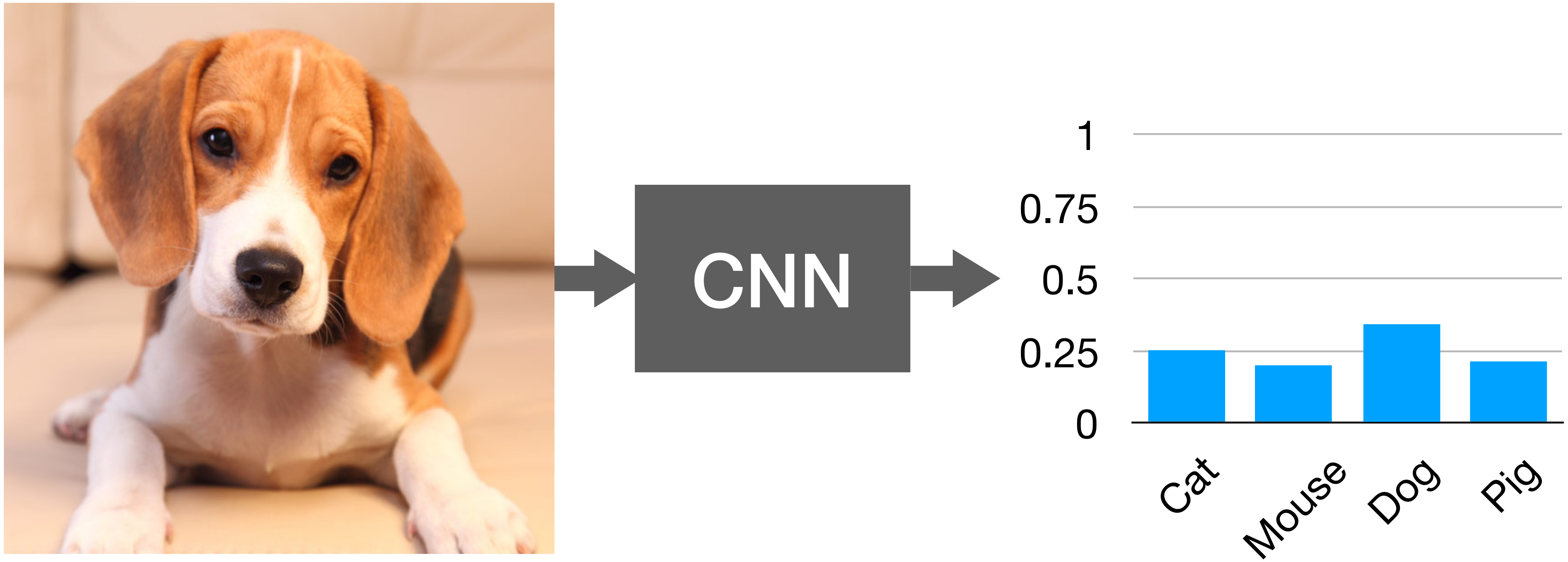

CNN
1
0.75
0.5
0.25
0
Cat
Mouse
Dog
Pig
Target Label
Normalize outputs to sum
to unity with softmax:
$$\sigma(z)_j = \frac{\exp(z_j)}{\sum_{k=1}^{K} \exp(z_k)}$$



Cross entropy loss:

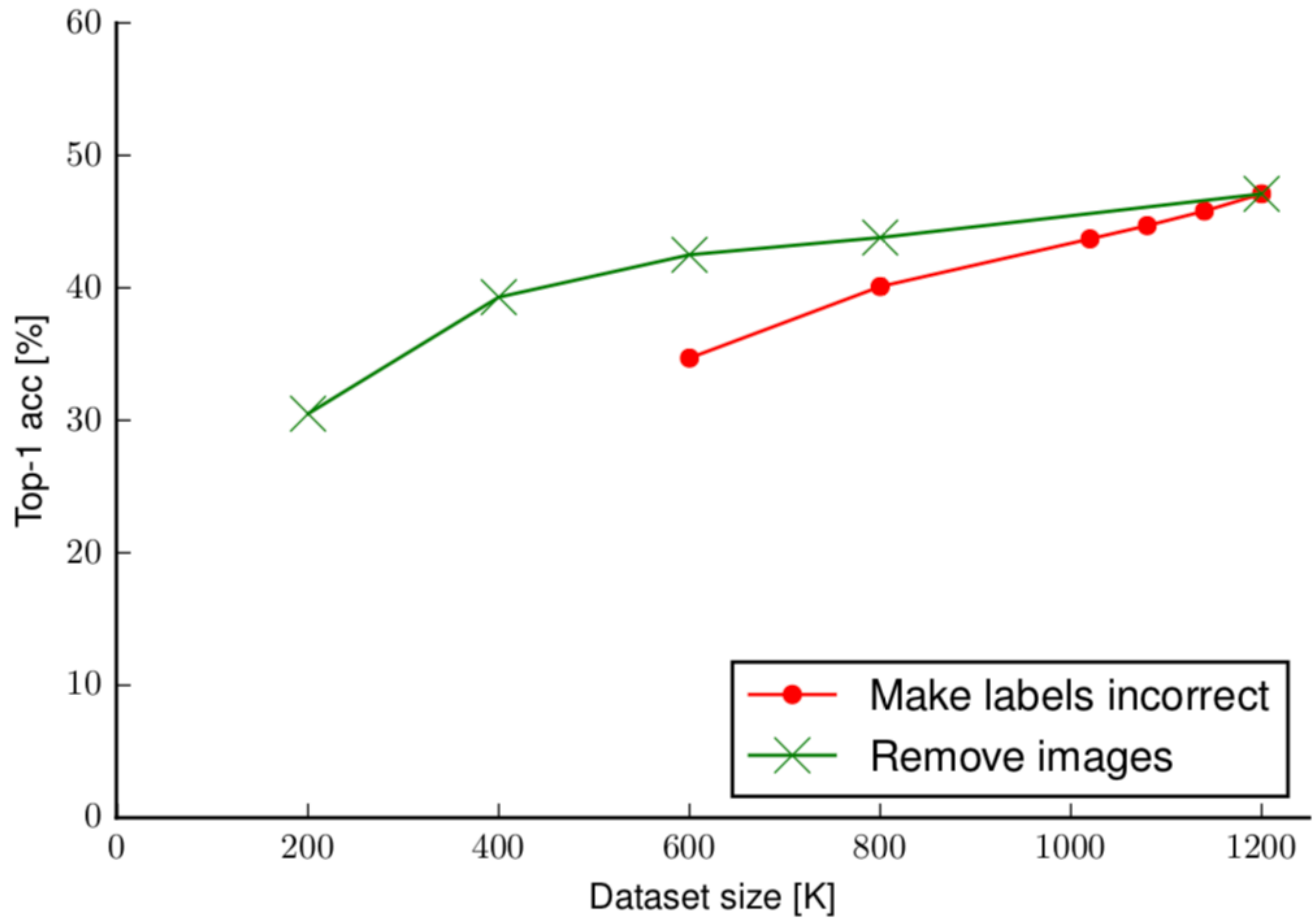
$$\mathscr{L}(x, y) = -\sum_i y_i \log x_i$$


Follow gradient step
to lower loss:
CNN
1
0.75
0.5
0.25
0
Cat
Mouse
Dog
Pig
1
0.75
0.5
0.25
0
Cat
Mouse
Dog
Pig
Target Label
Cross entropy loss:
$\mathscr{L}(x, y) = -\sum_i y_i \log x_i$



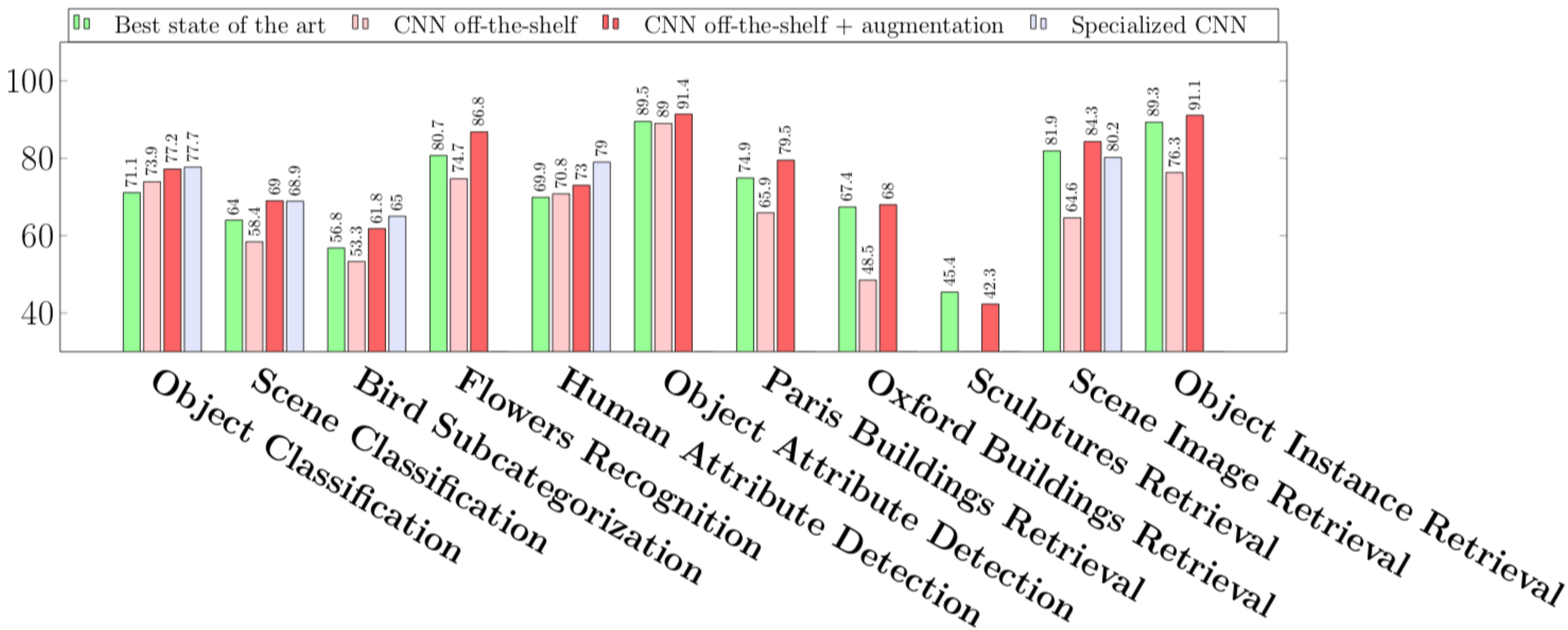
Question: How to localize where objects are?

How much data do you need?



Systematic evaluation of CNN advances on the ImageNet

How much data do you need?



CNN Features off-the-shelf: an Astounding Baseline for Recognition

Short cuts to AI

With billions of images on the web, it's often possible to find a close nearest neighbor.

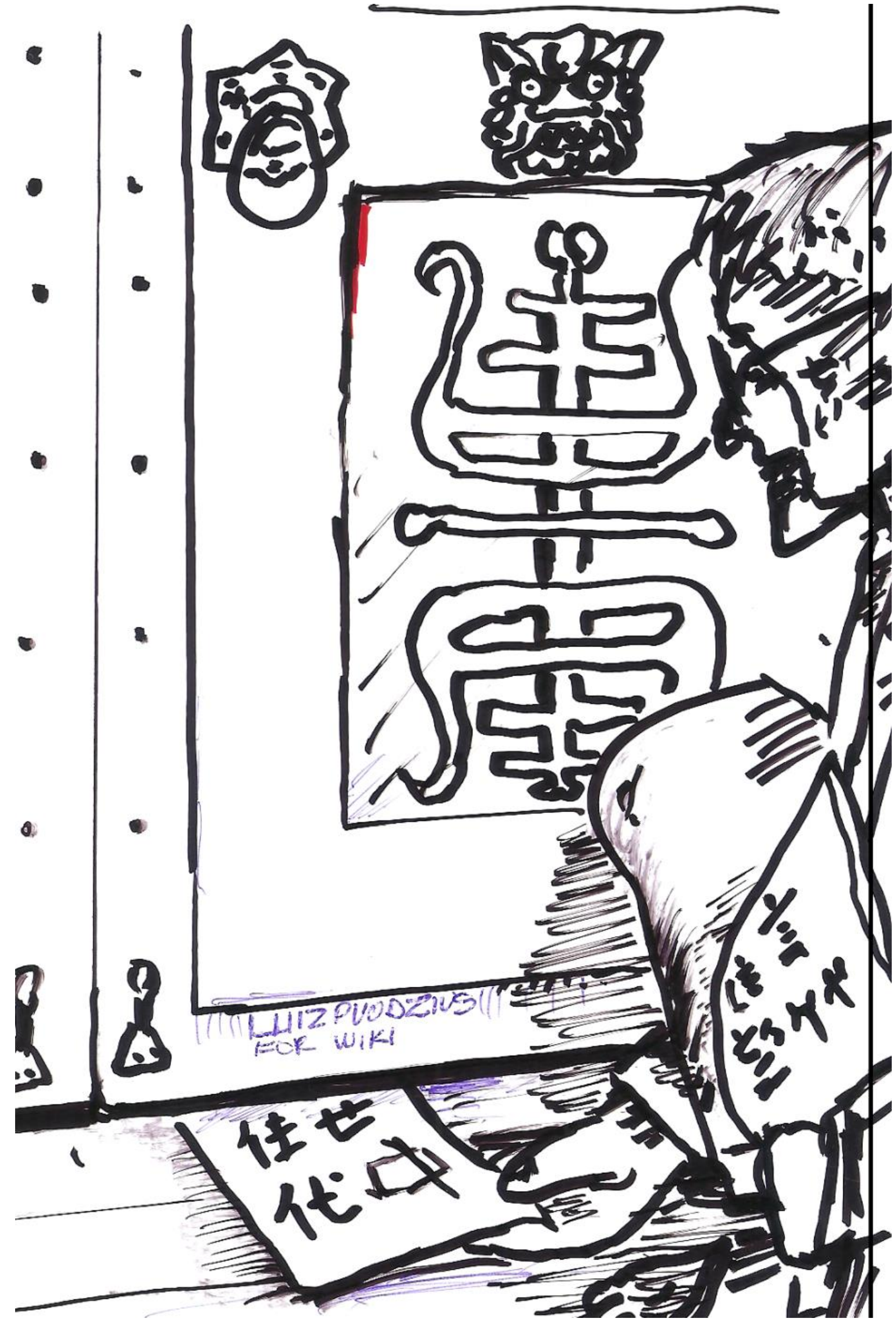
We can shortcut hard problems by "looking up" the answer, stealing the labels from our nearest neighbor.

Chinese Room experiment, John Searle (1980)

Input to program is Chinese, and output is also Chinese. It passes the Turing test.

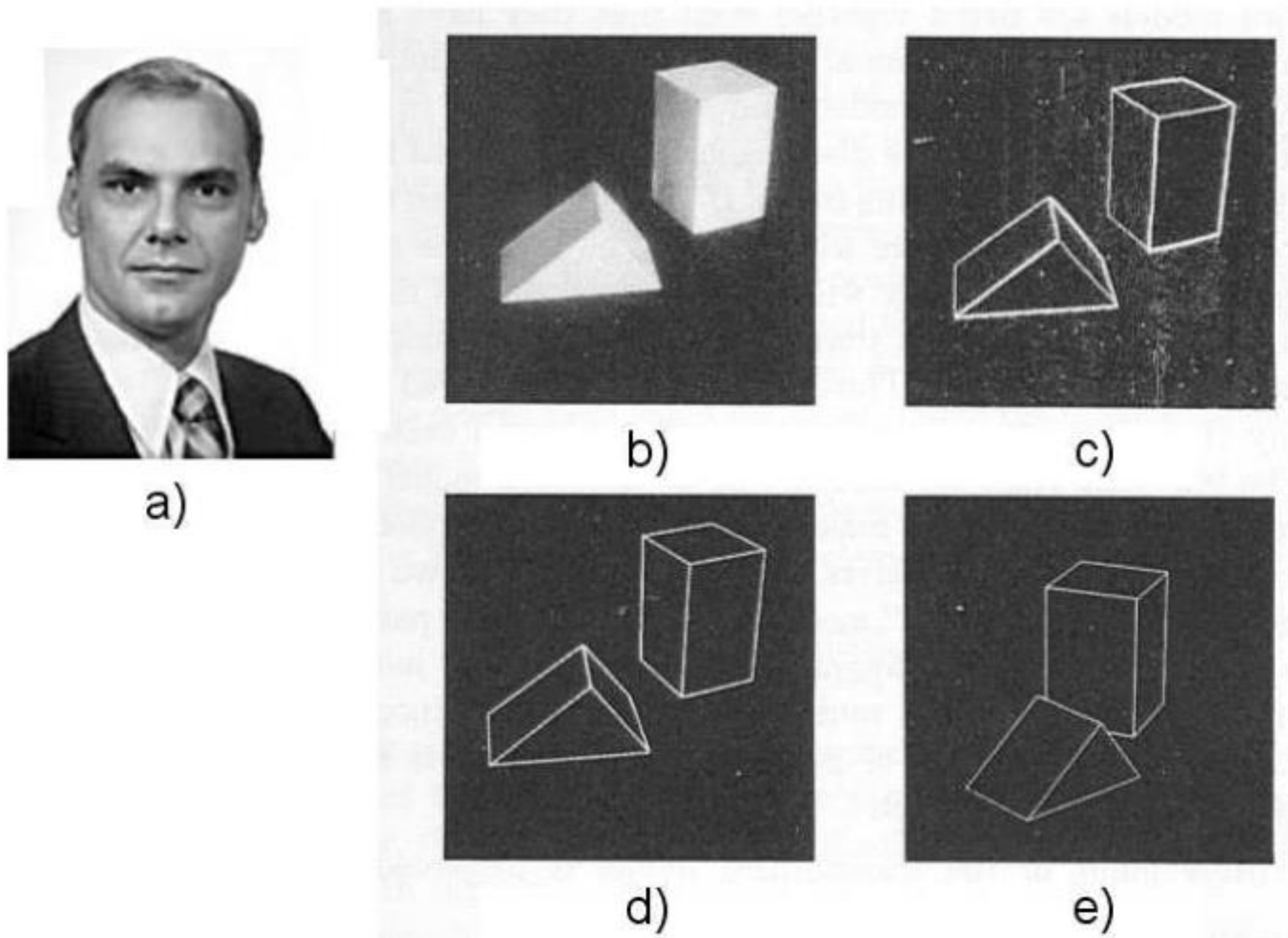
Does the computer “understand” Chinese or just “simulate” it?

What if the software is just a lookup table?



History

Recognition as an alignment problem: Block world

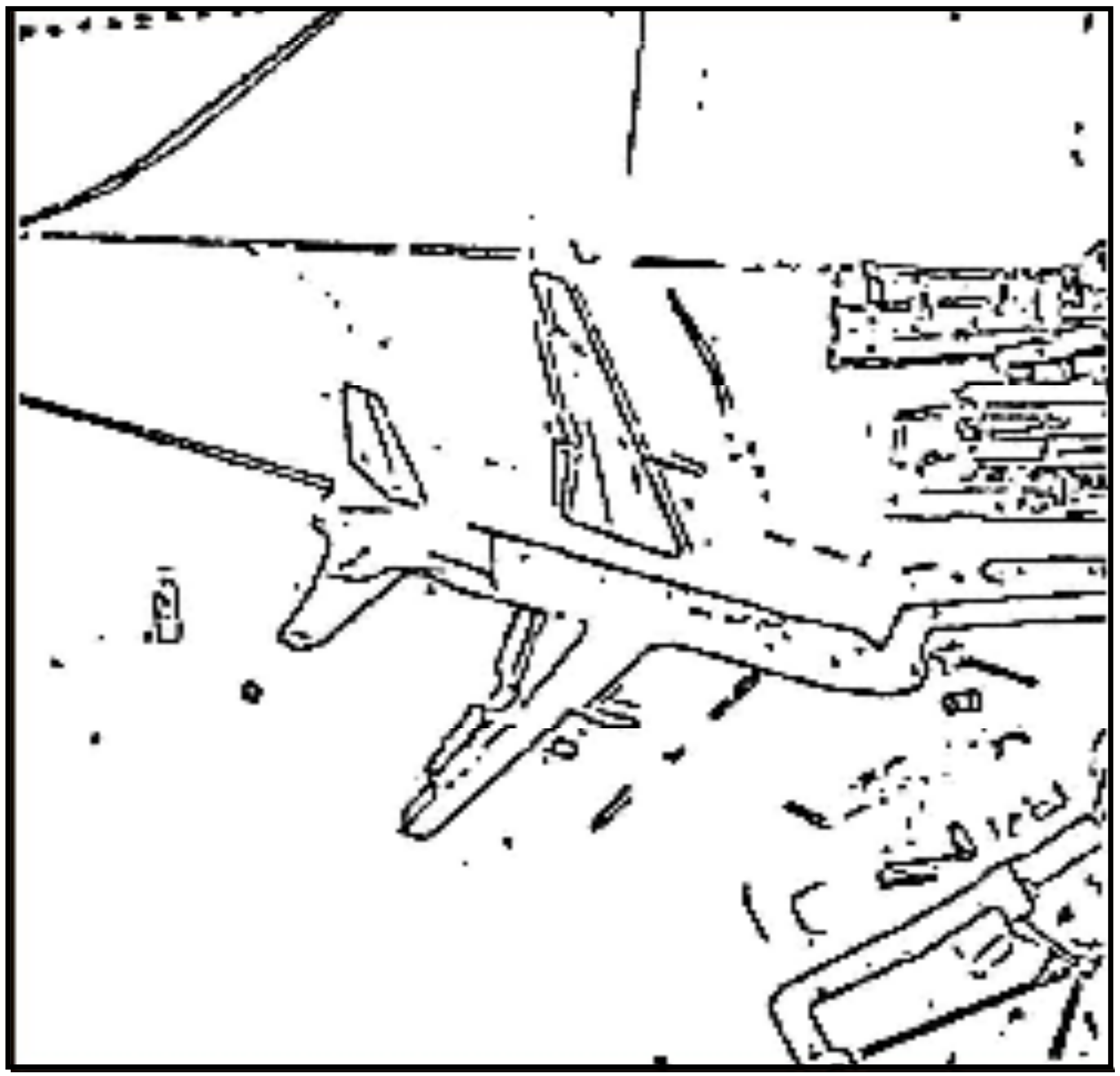


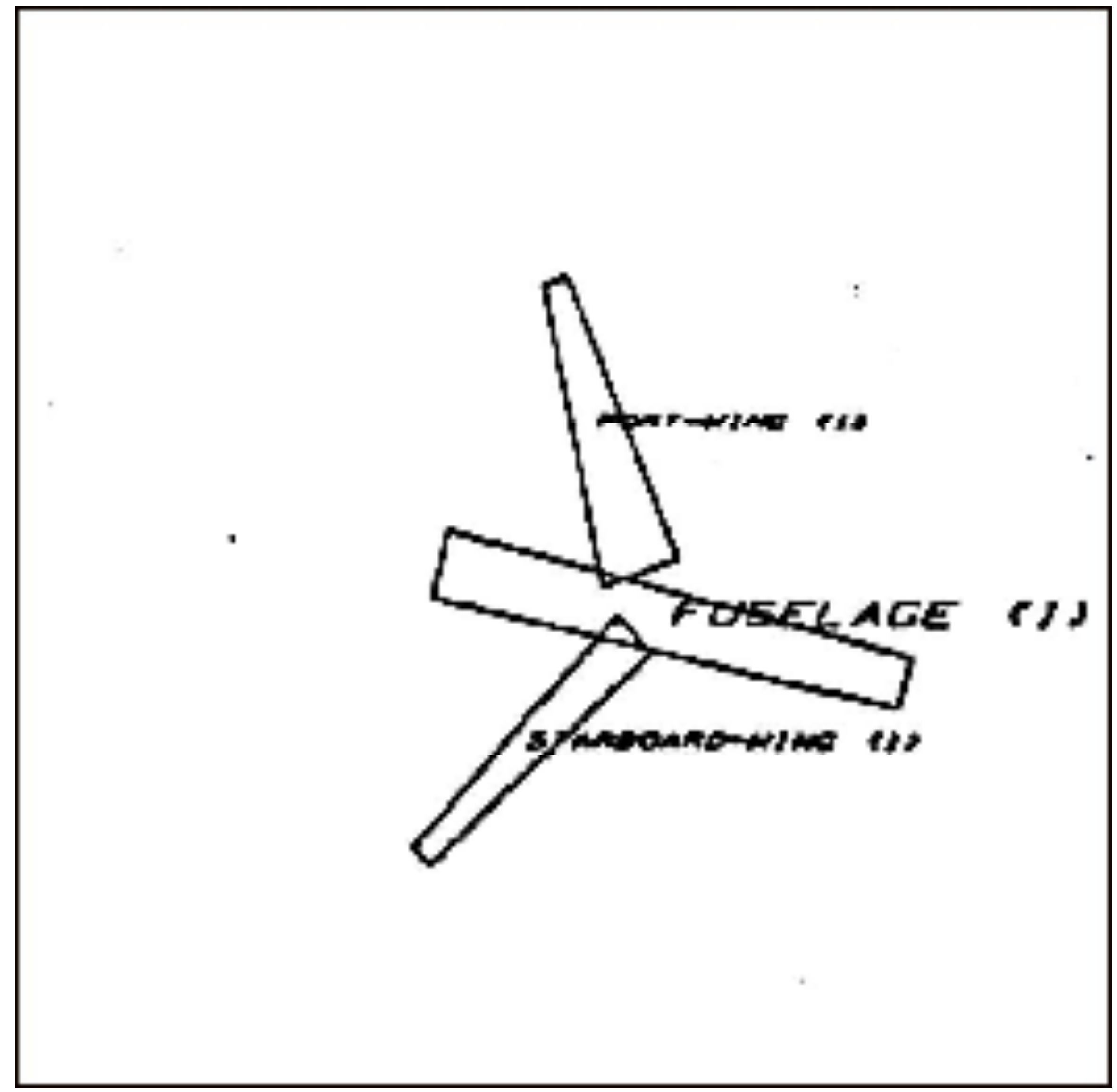
L. G. Roberts
Machine Perception of Three Dimensional Solids,
Ph.D. thesis, MIT
Department of Electrical
Engineering, 1963.

Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b)A blocks world scene. c)Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - e) are taken from [64] with permission MIT Press.)

J. Mundy, Object Recognition in the Geometric Era: a Retrospective, 2006

Representing and recognizing object categories is harder...





ACRONYM (Brooks and Binford, 1981)

Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)

Binford and generalized cylinders

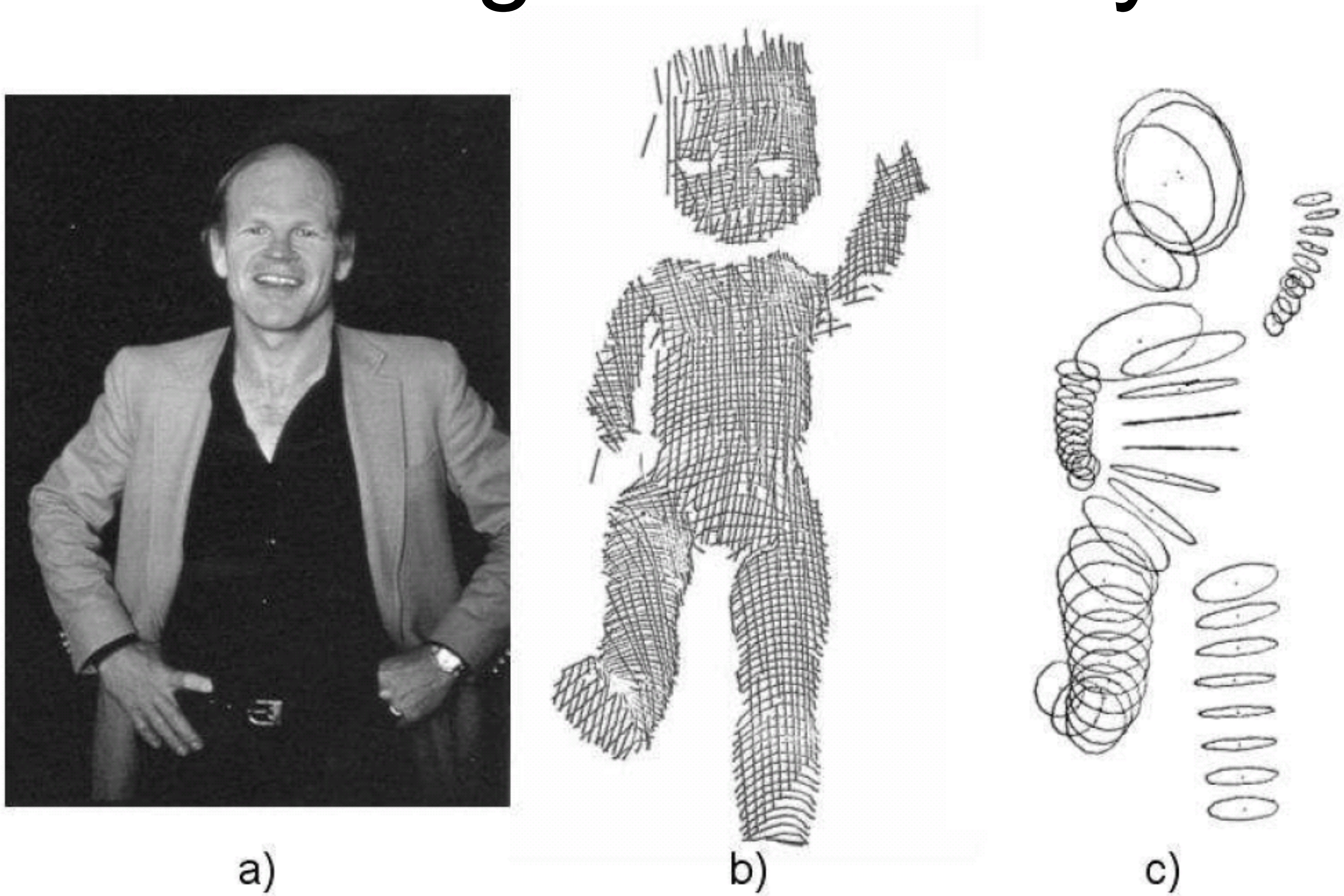
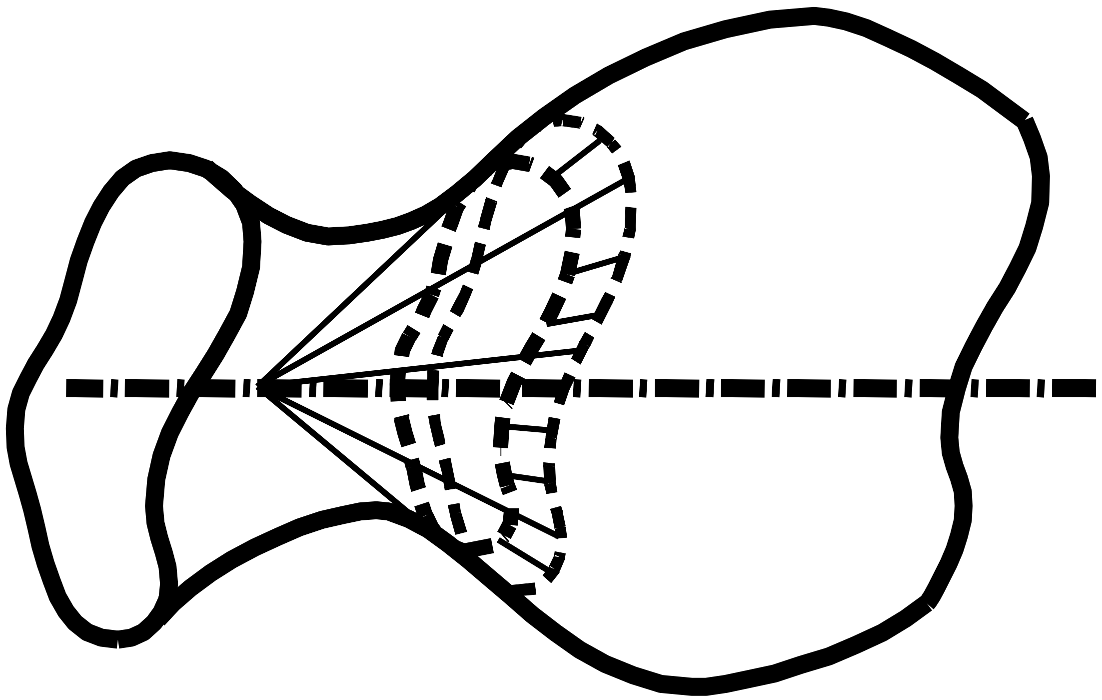


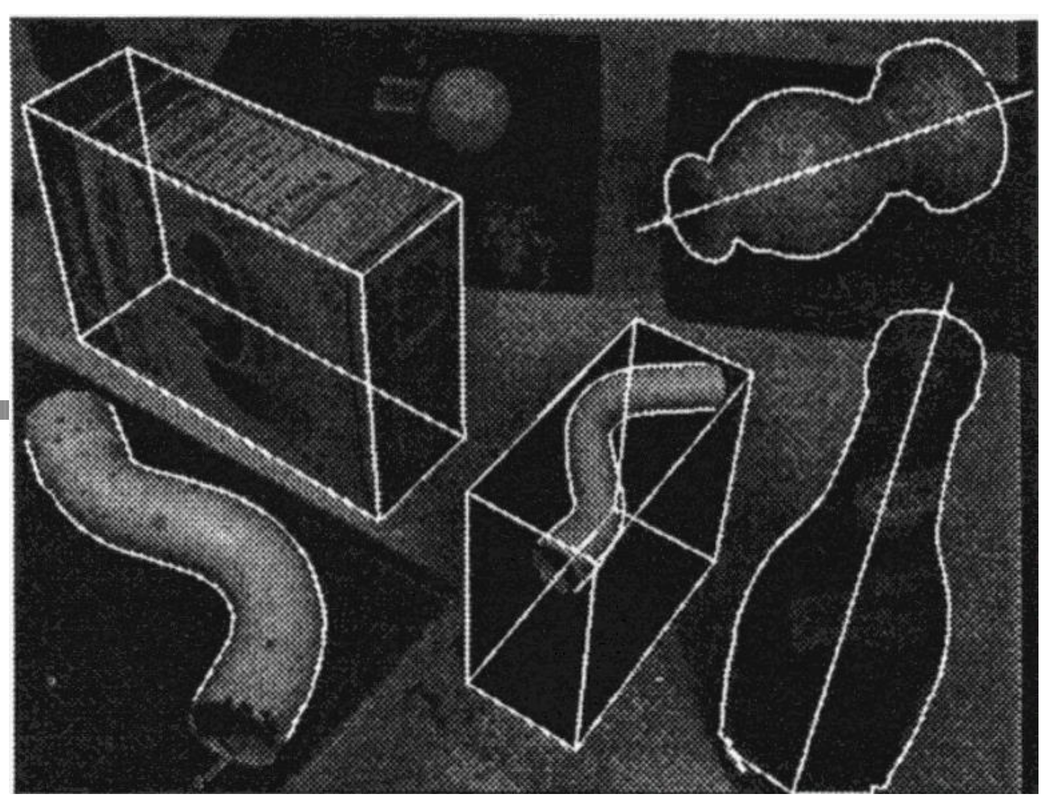
Fig. 3. The representation of objects by assemblies of generalized cylinders. a) Thomas Binford. b) A range image of a doll. c) The resulting set of generalized cylinders. (b) and c) are taken from Agin [1] with permission.)

Object Recognition in the Geometric Era: a Retrospective. Joseph L. Mundy. 2006

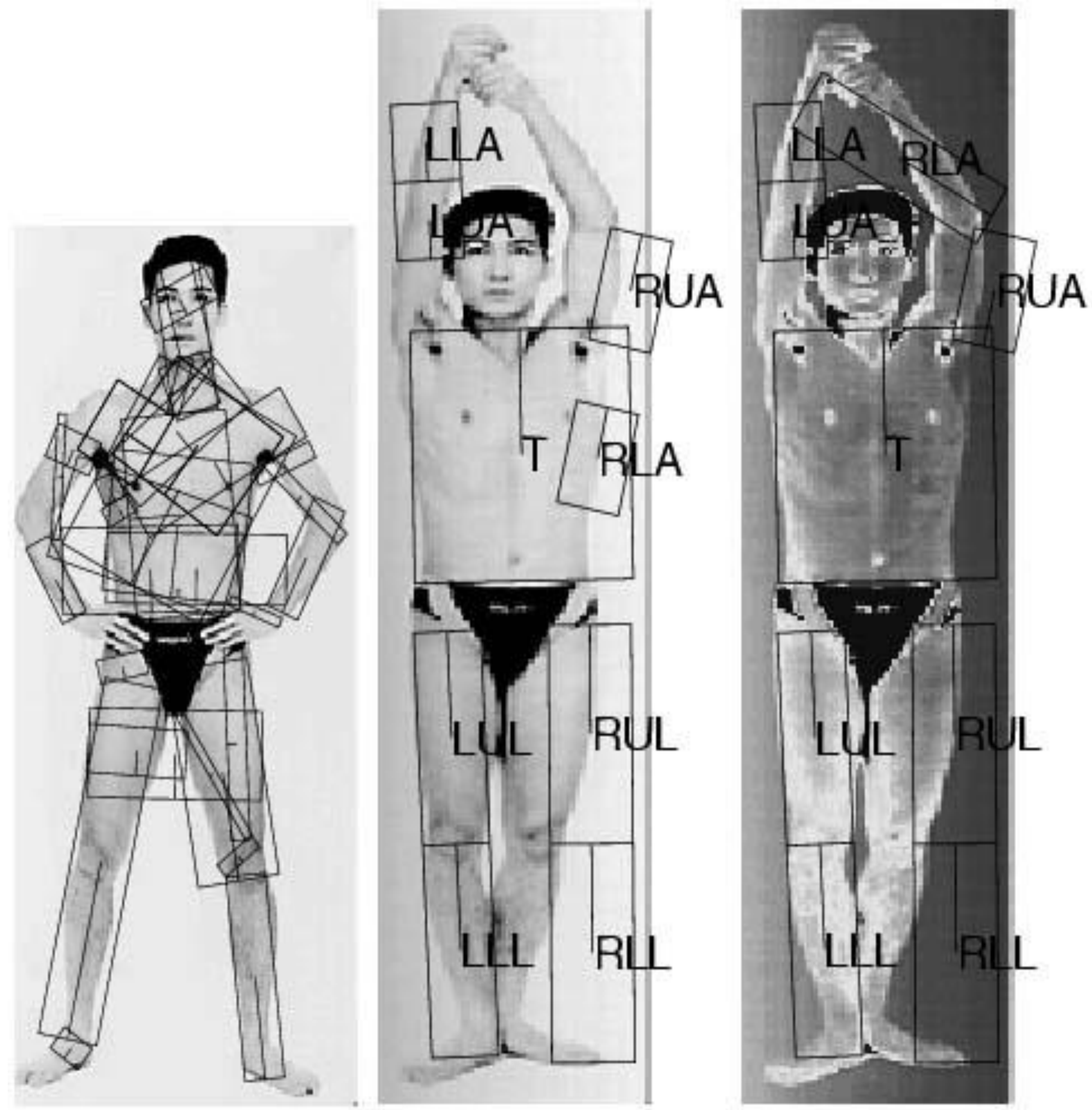
General shape primitives?



Generalized cylinders
Ponce et al. (1989)



Zisserman et al. (1995)



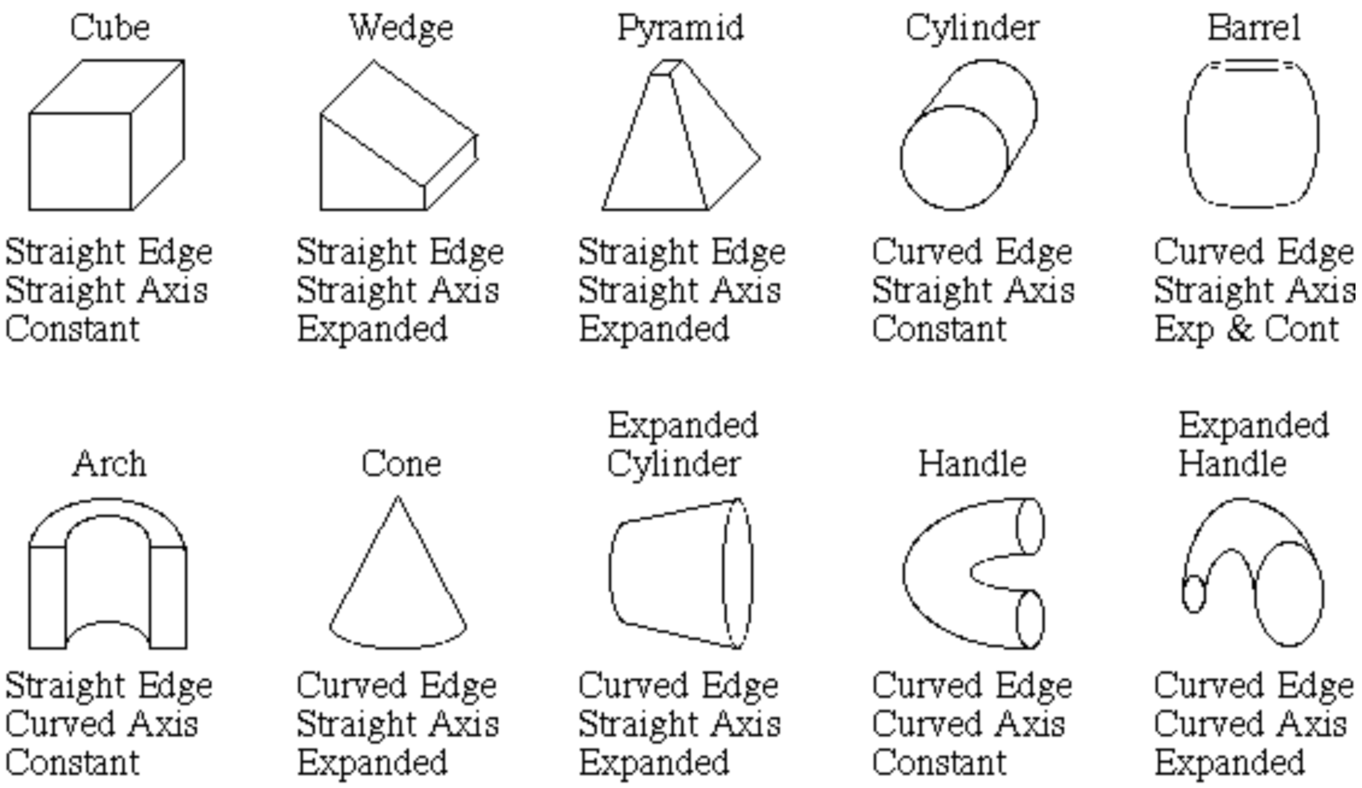
Forsyth (2000)

Svetlana Lazebnik

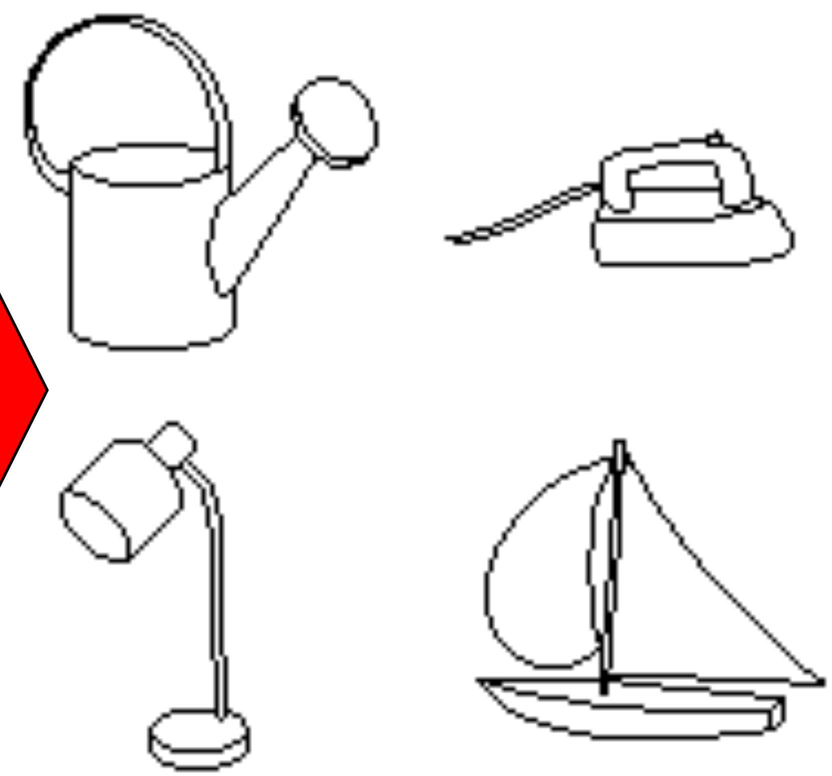
Recognition by components

Biederman (1987)

Primitives (geons)



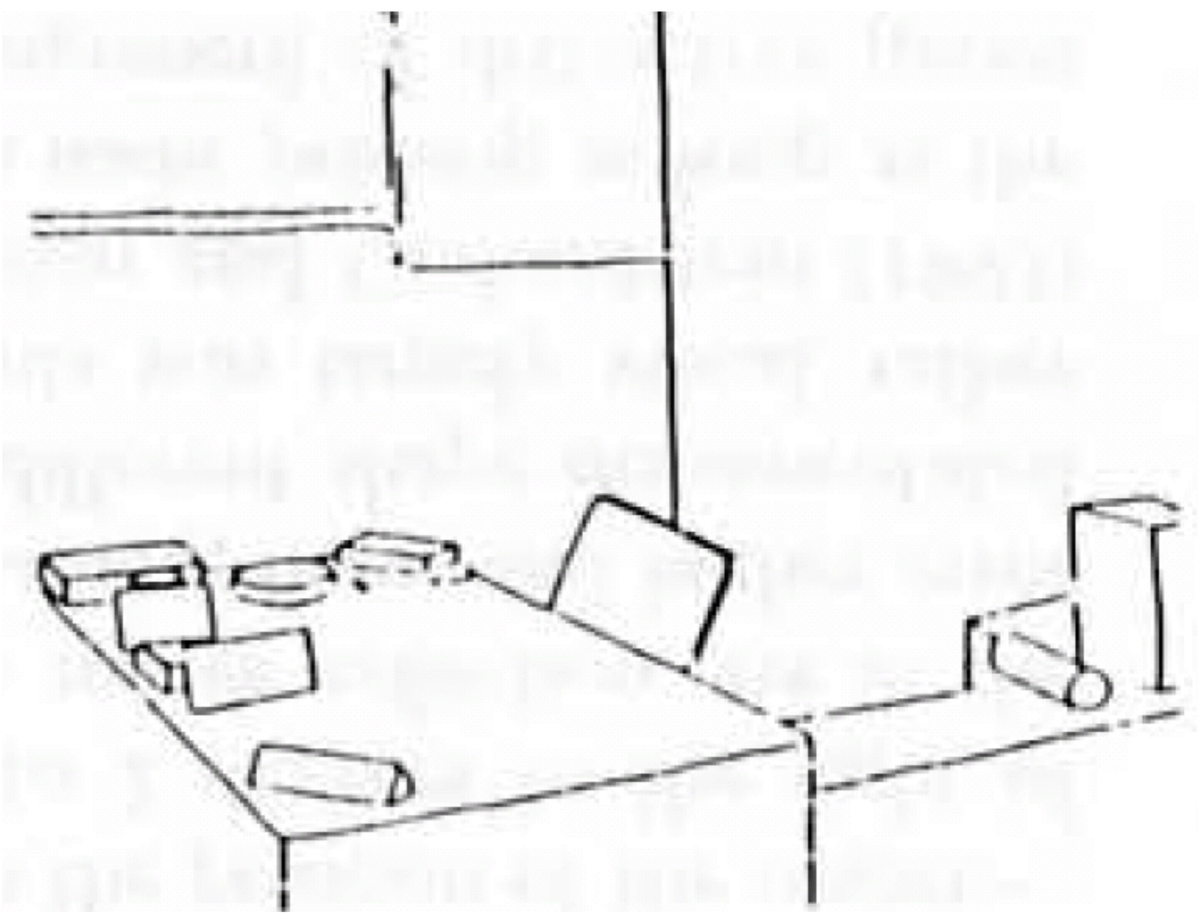
Objects

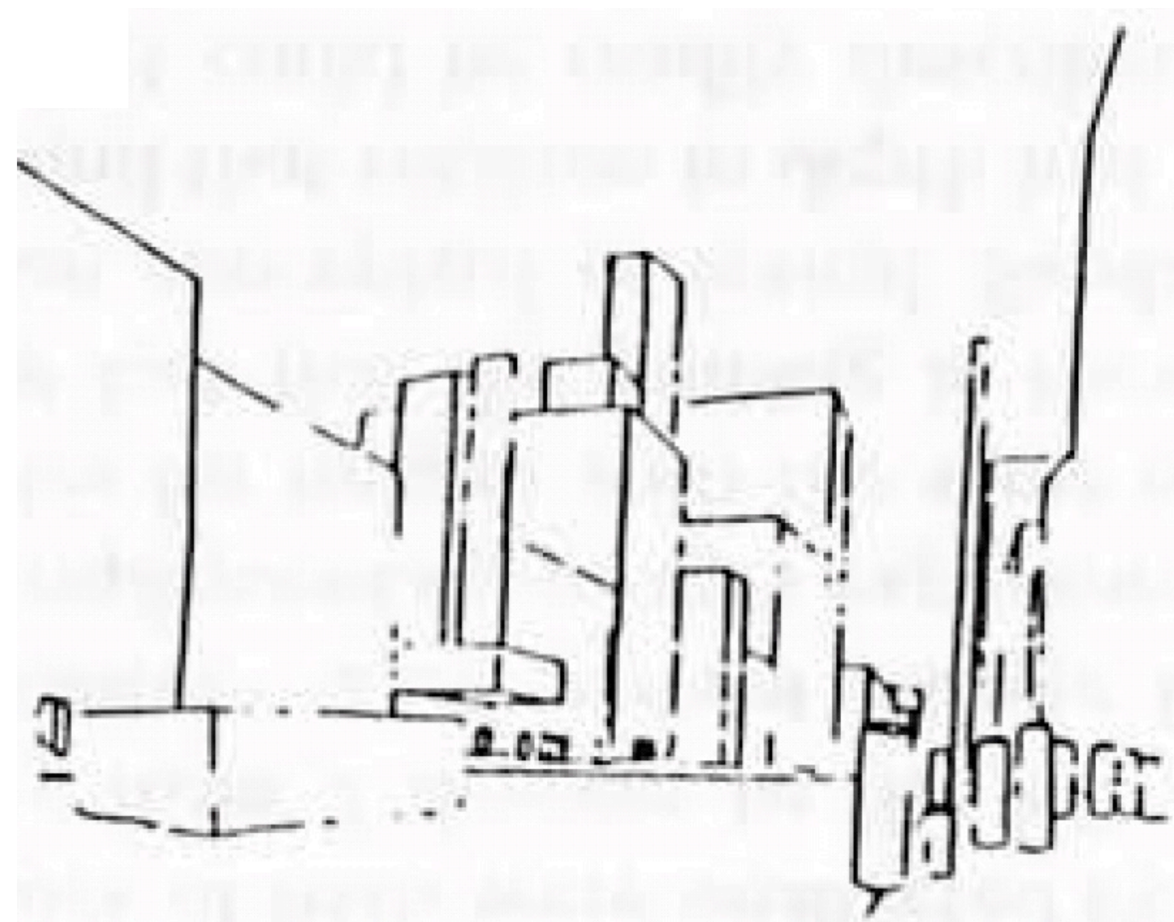


http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

Svetlana Lazebnik

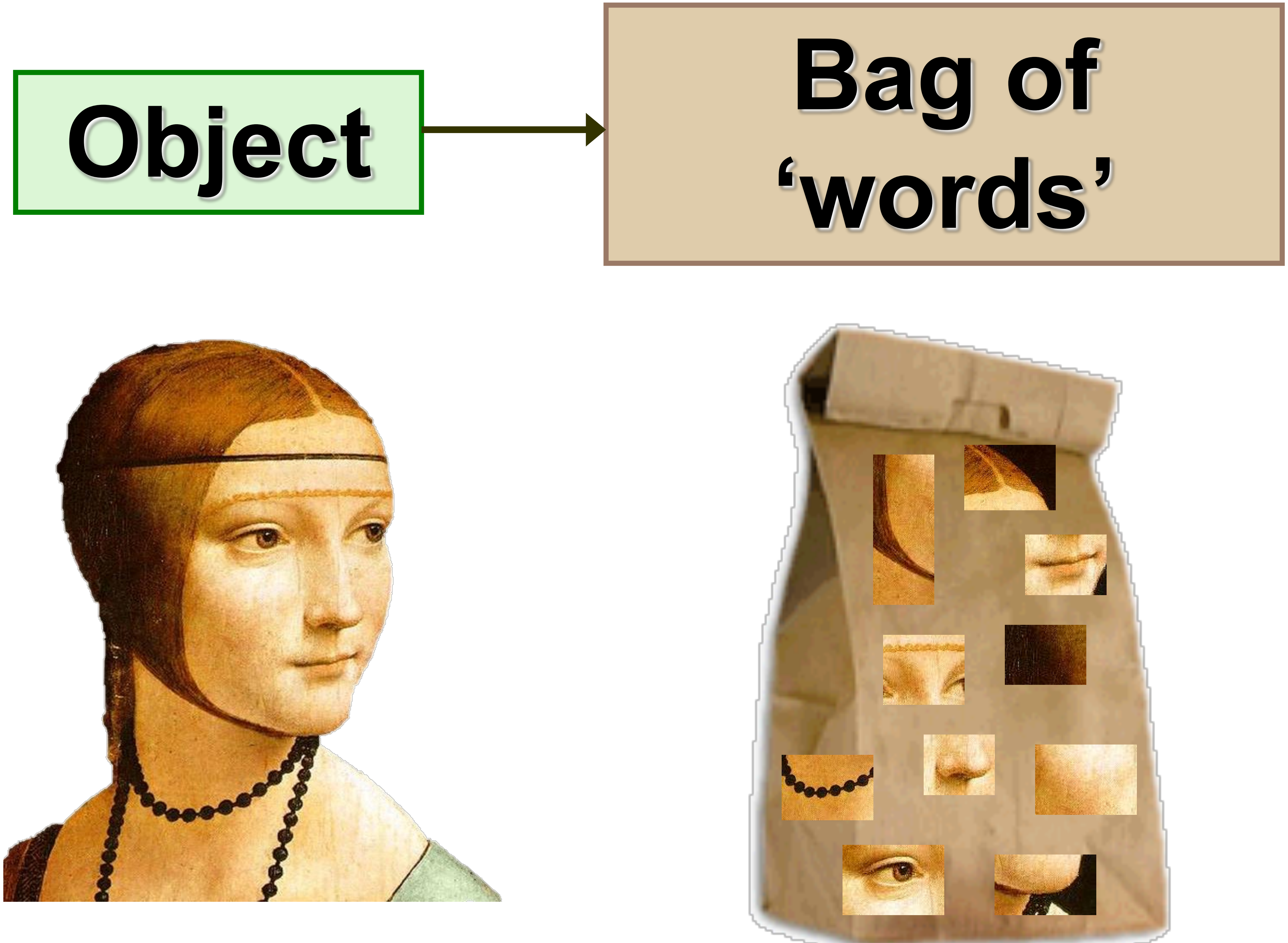
Scenes and geons





Mezzanotte & Biederman

Bag-of-features models



Svetlana Lazebnik

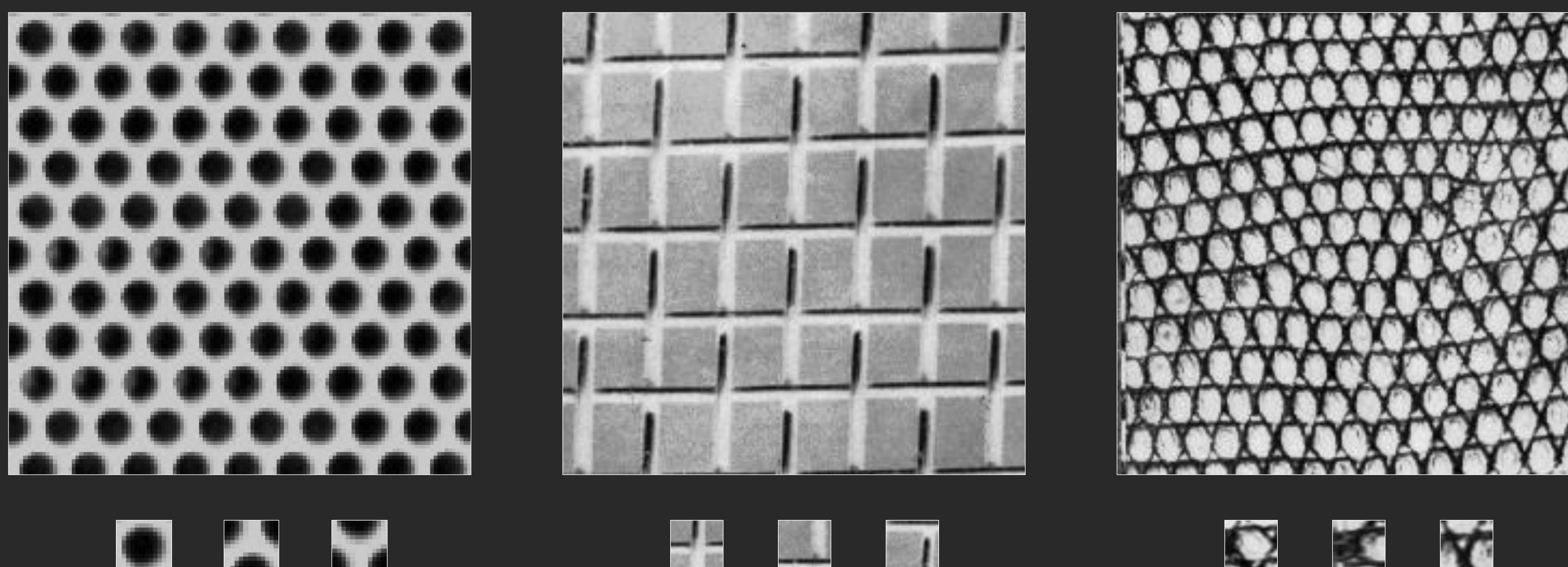
Origin 1: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary   Salton & McGill (1983)

US Presidential Speeches Tag Cloud
http://chir.ag/phernalia/preztags/

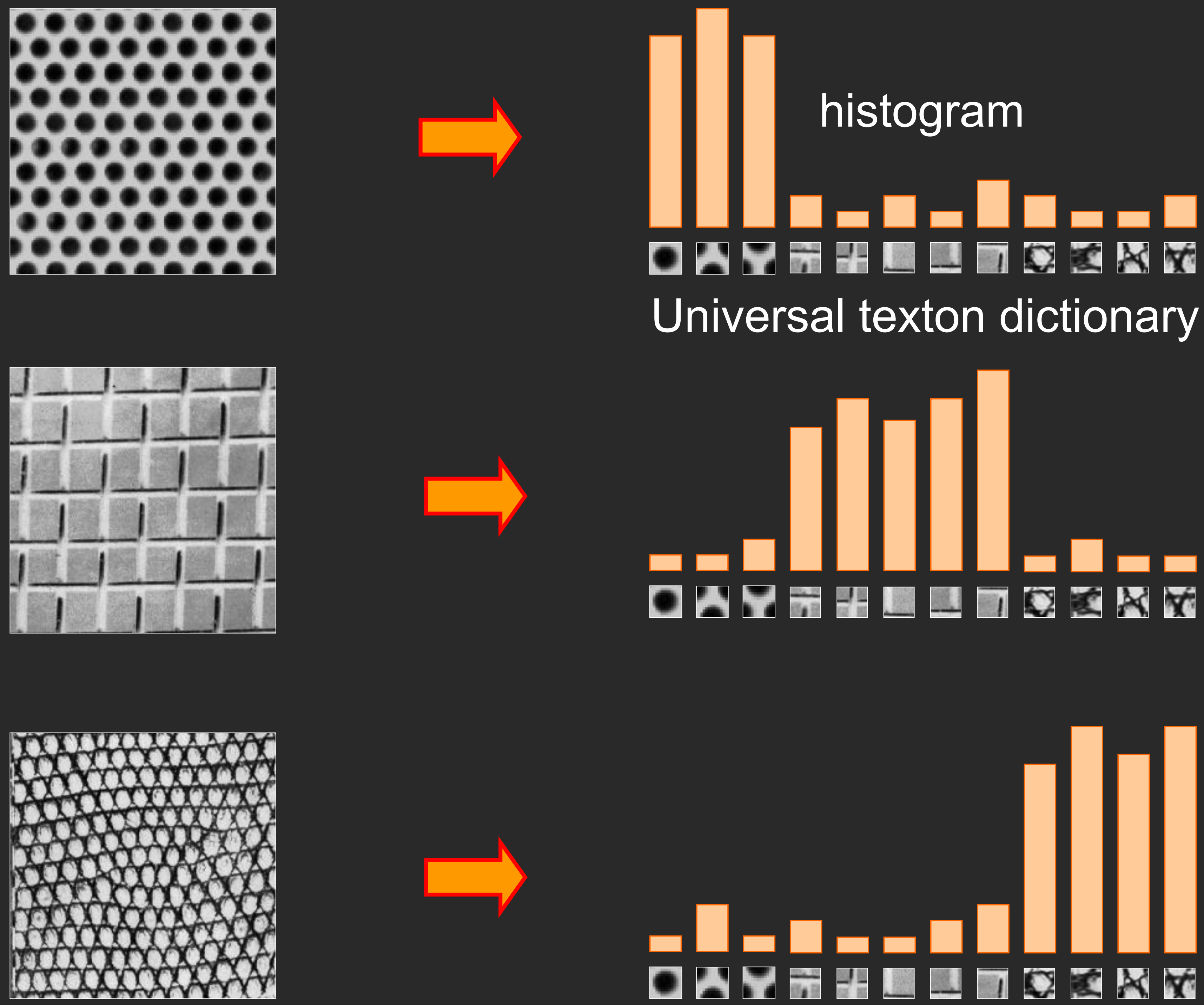
Origin 2: Texture recognition

- Characterized by repetition of basic elements or *textons*
- For stochastic textures, the identity of textons matters, not their spatial arrangement



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Origin 2: Texture recognition



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Bag-of-features models

Svetlana Lazebnik

Objects as texture

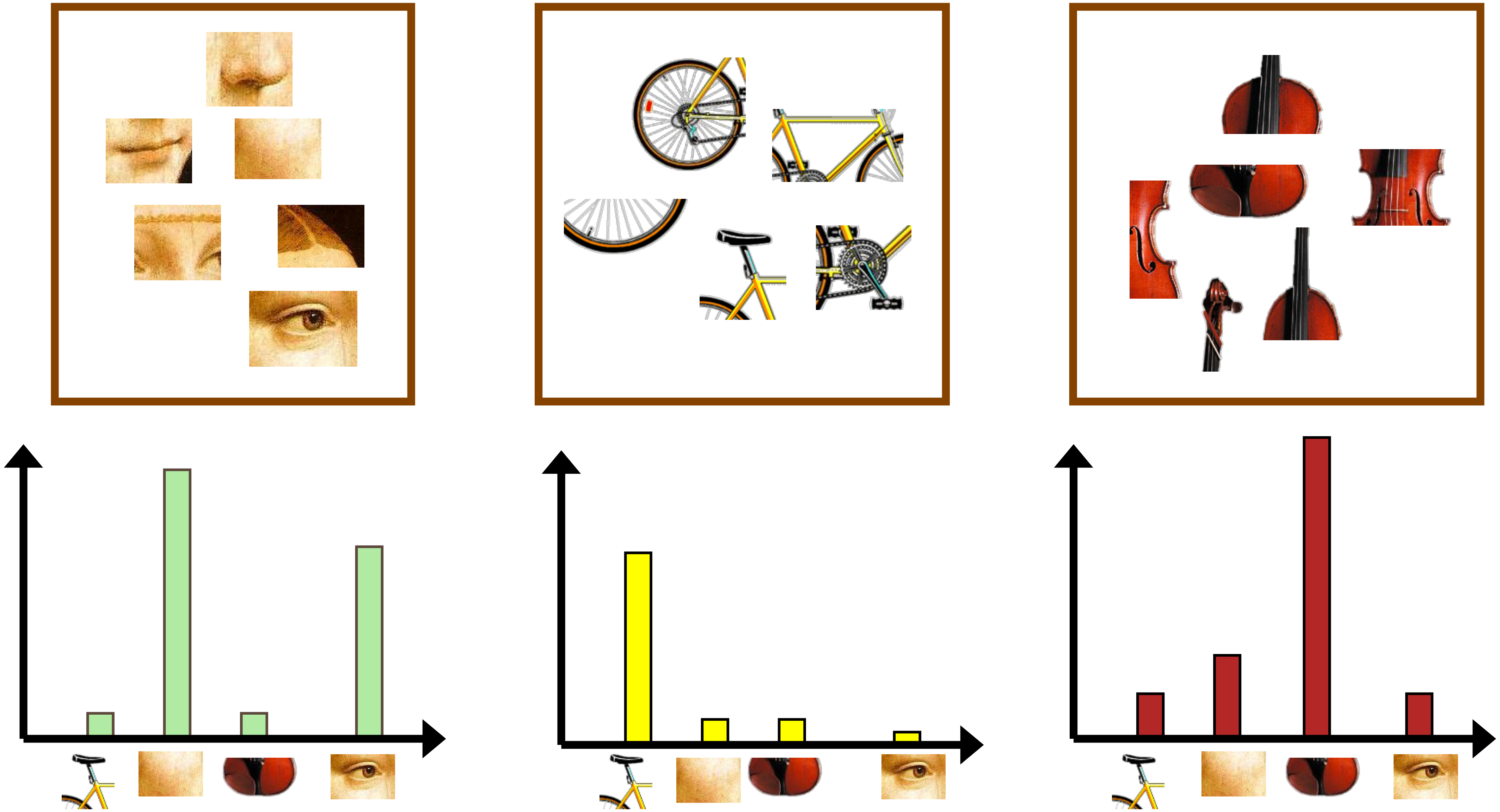
- All of these are treated as being the same

- No distinction between foreground and background: scene recognition?

Svetlana Lazebnik

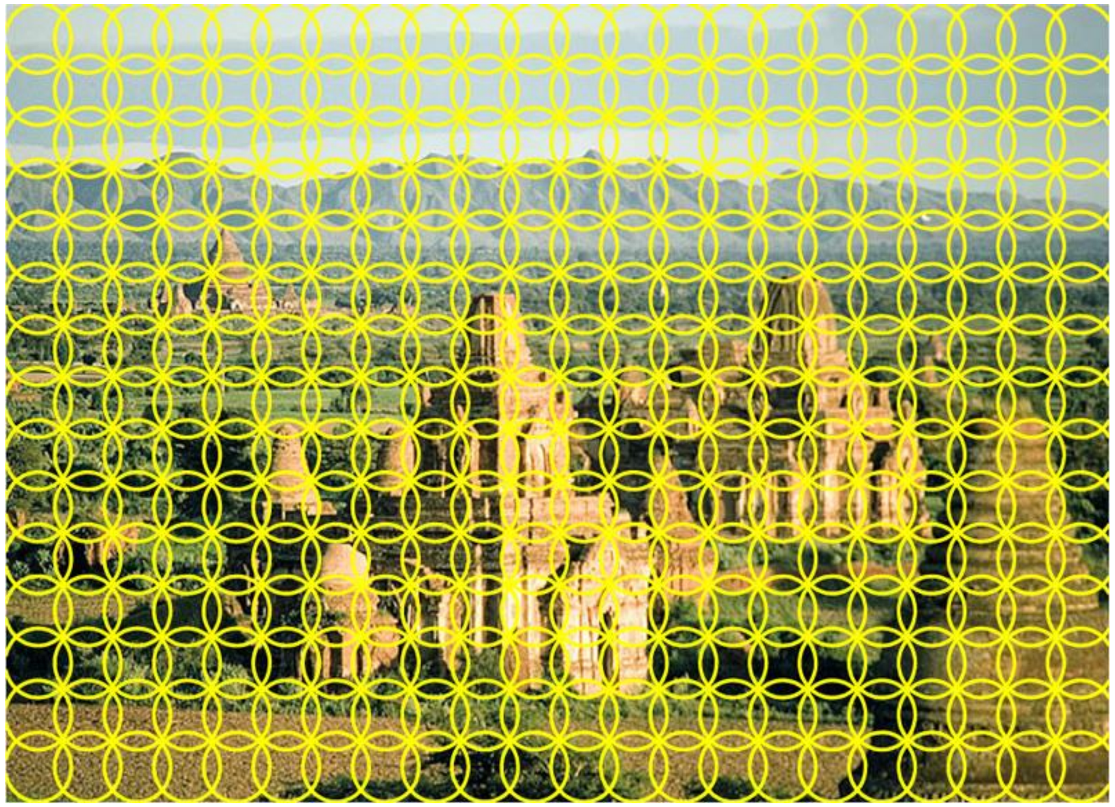
Bag-of-features steps

1. Feature extraction
2. Learn "visual vocabulary"
3. Quantize features using visual vocabulary
4. Represent images by frequencies of "visual words"

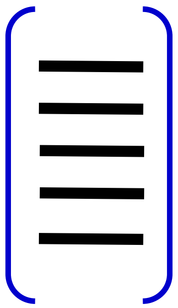


1. Feature extraction

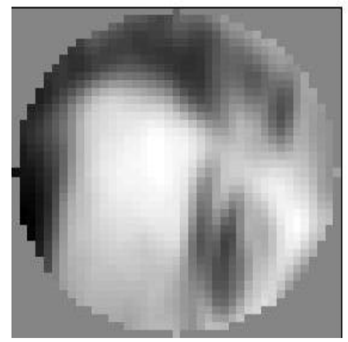
- Regular grid or interest regions



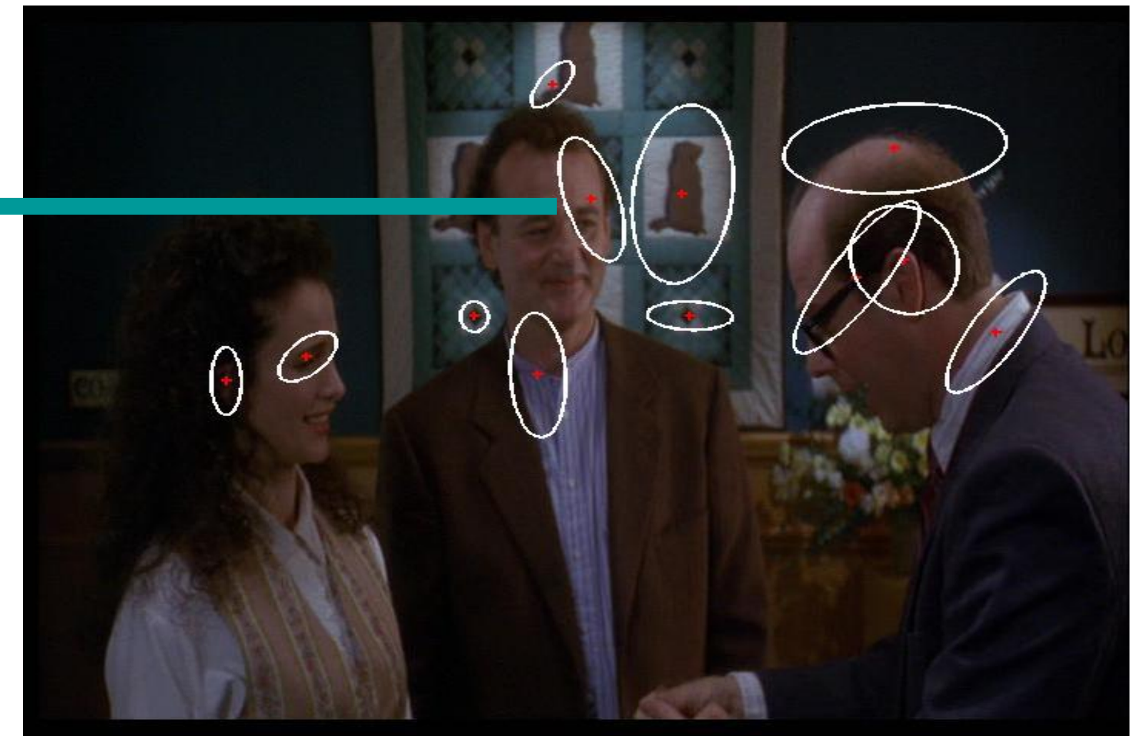
1. Feature extraction



Compute descriptor



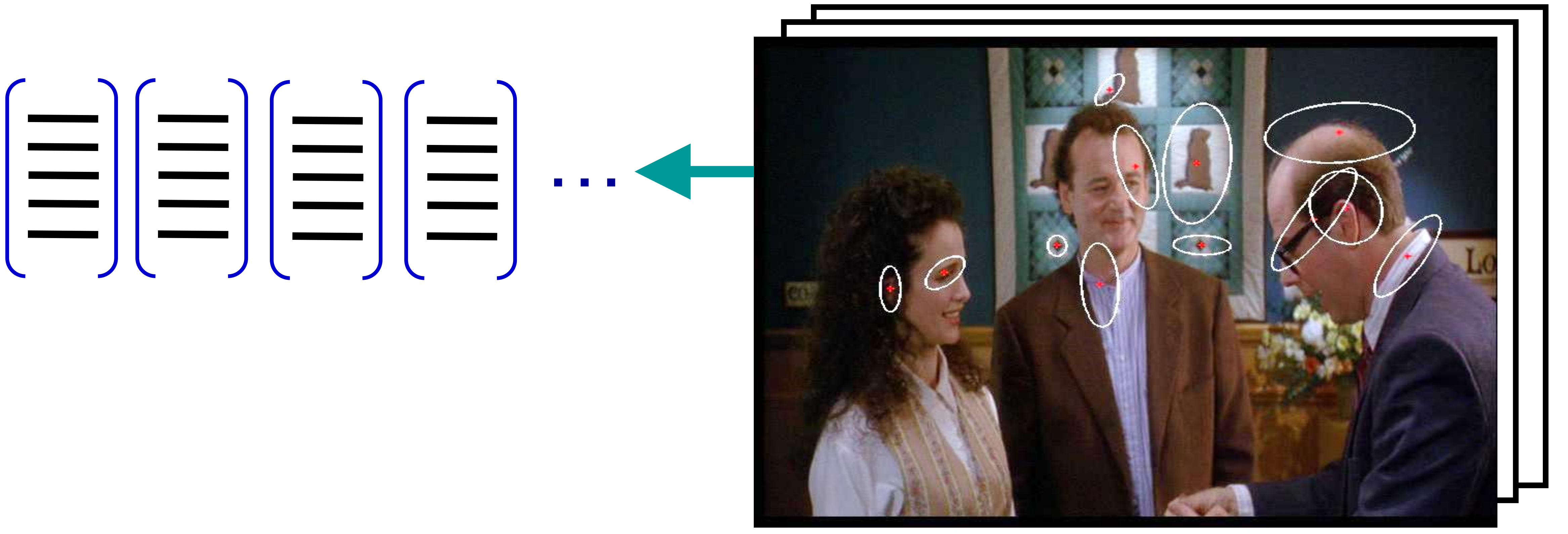
Extract patch



Detect patches

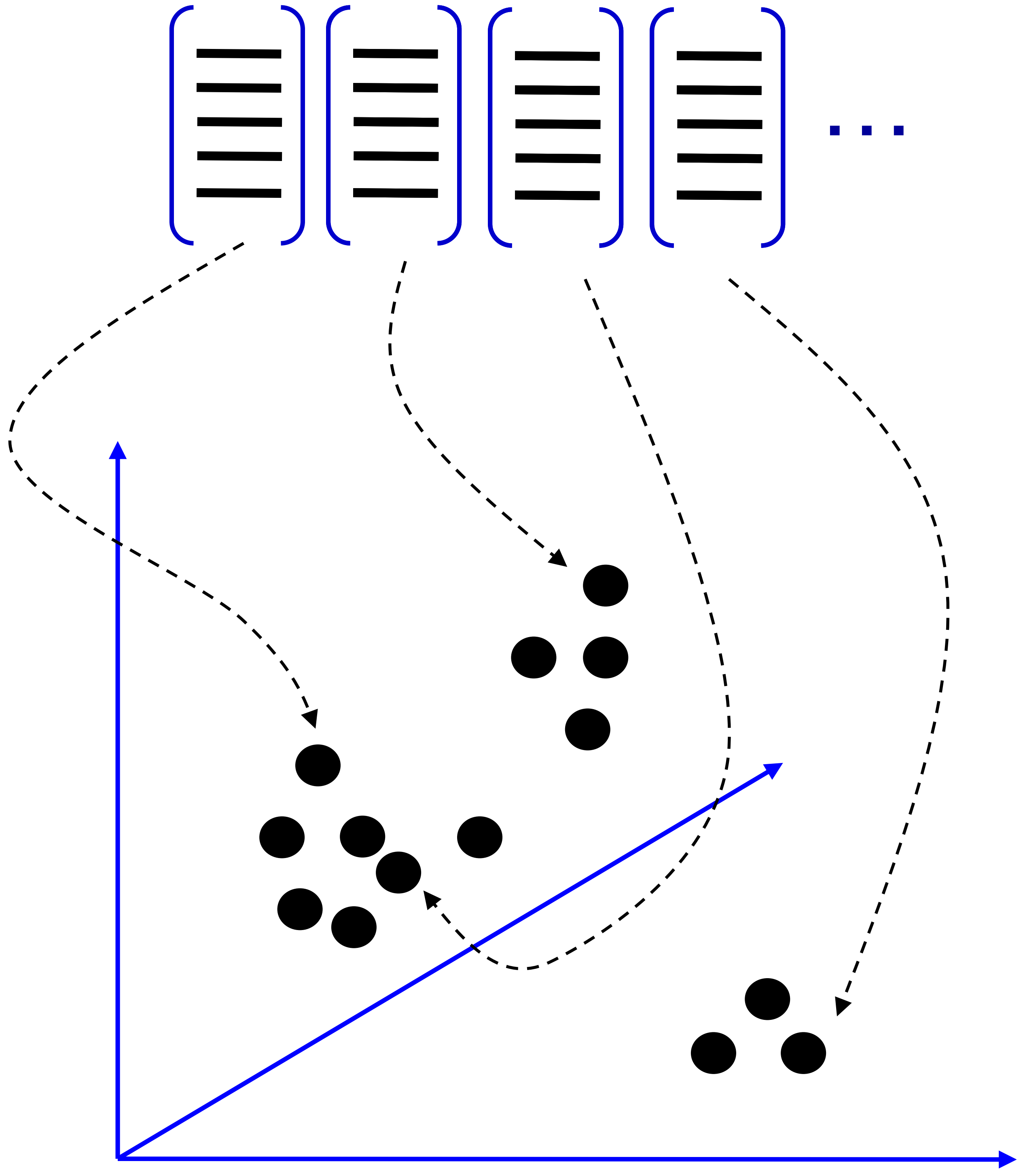
Slide credit: Josef Sivic

1. Feature extraction



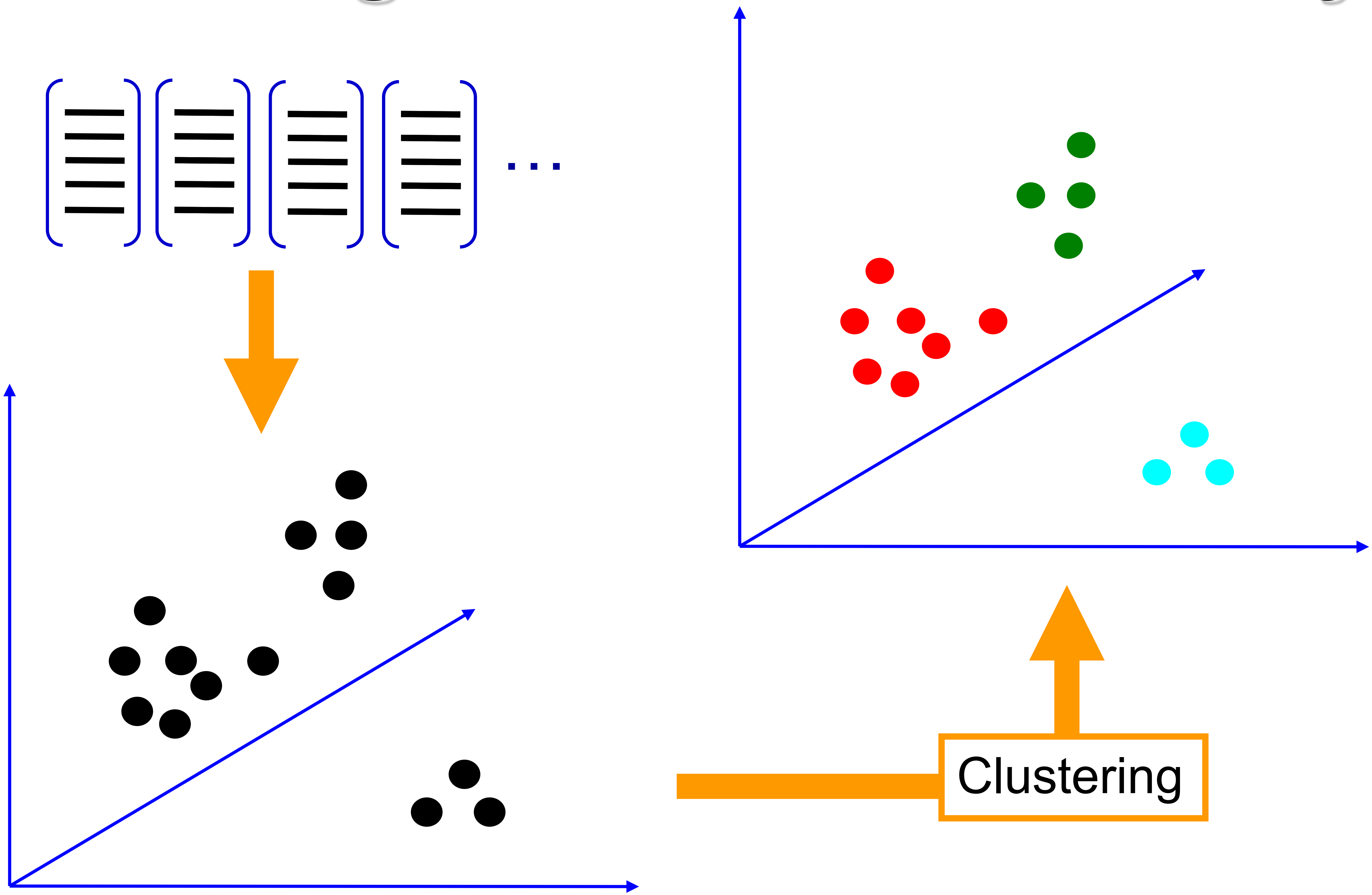
Slide credit: Josef Sivic

2. Learning the visual vocabulary



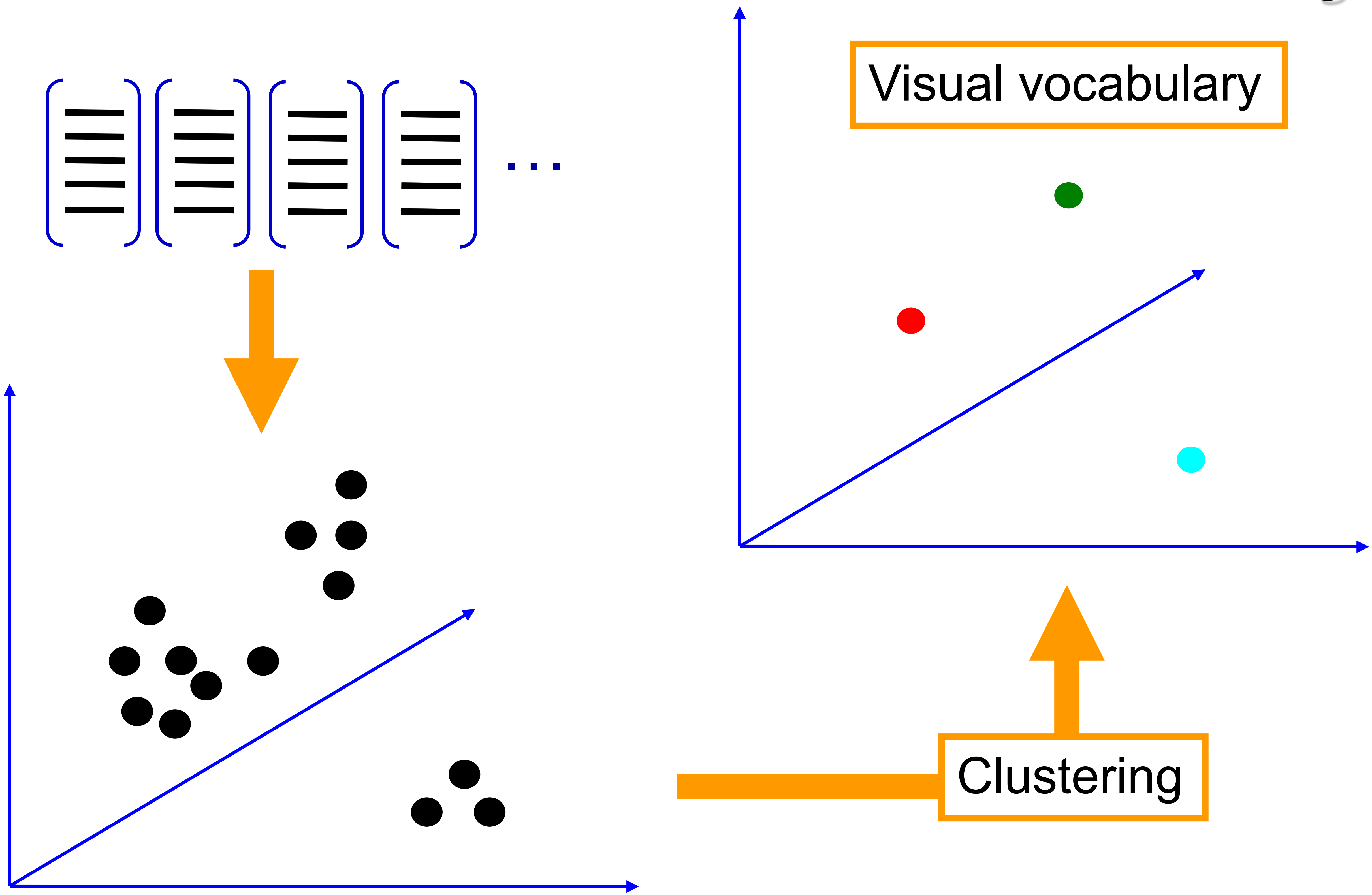
Slide credit: Josef Sivic

2. Learning the visual vocabulary



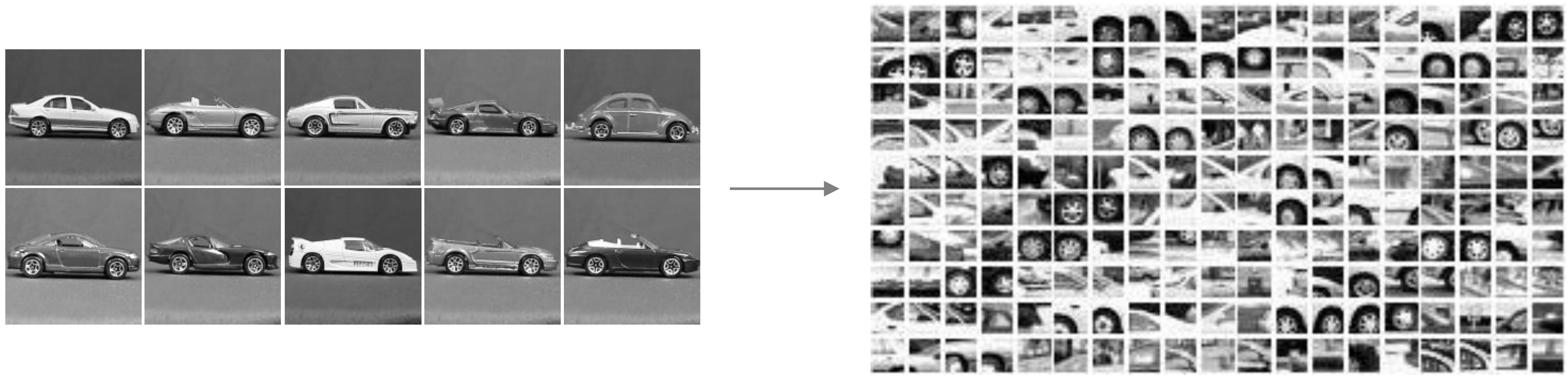
Slide credit: Josef Sivic

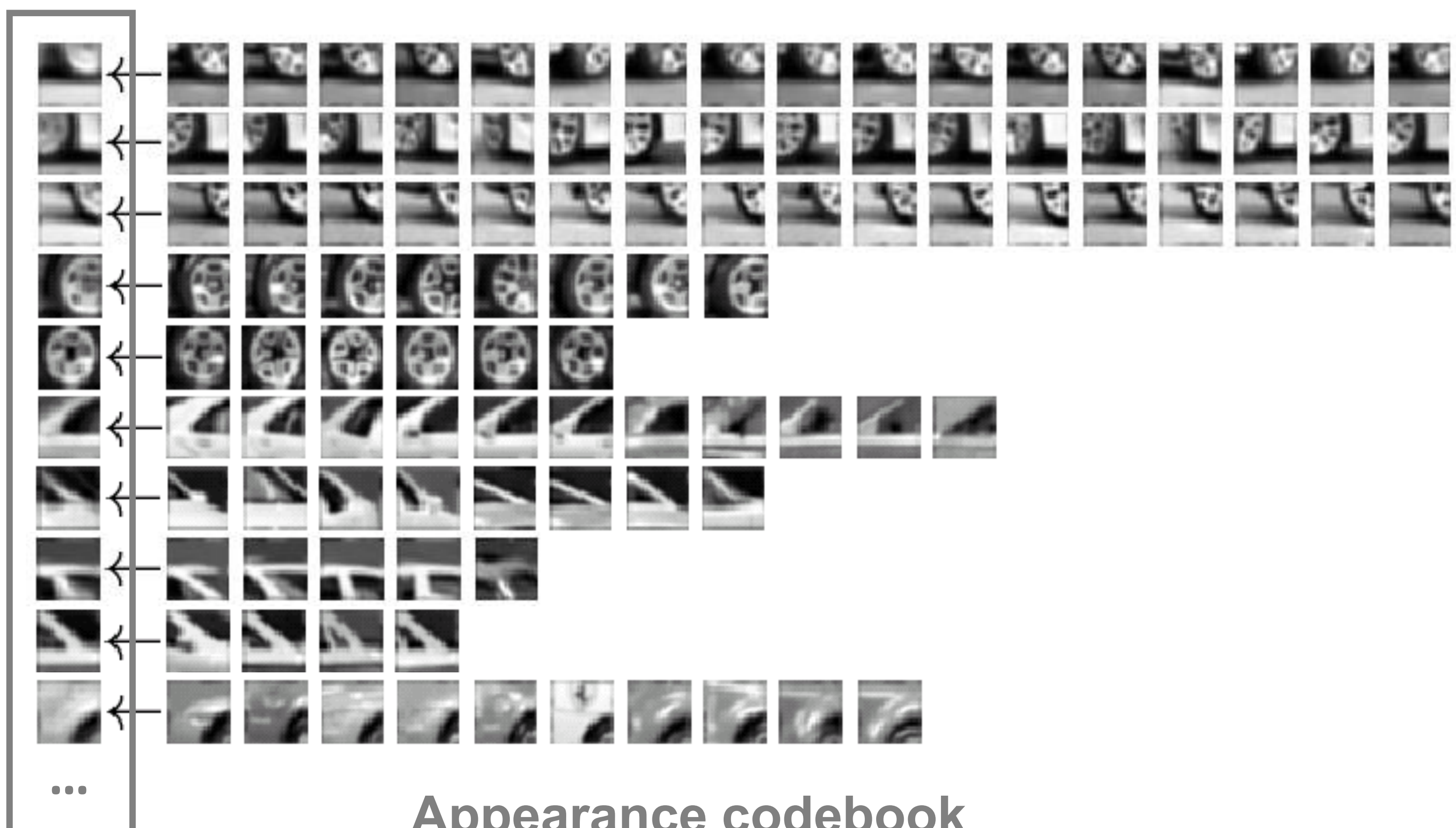
3. Quantize the visual vocabulary



Slide credit: Josef Sivic

Example codebook





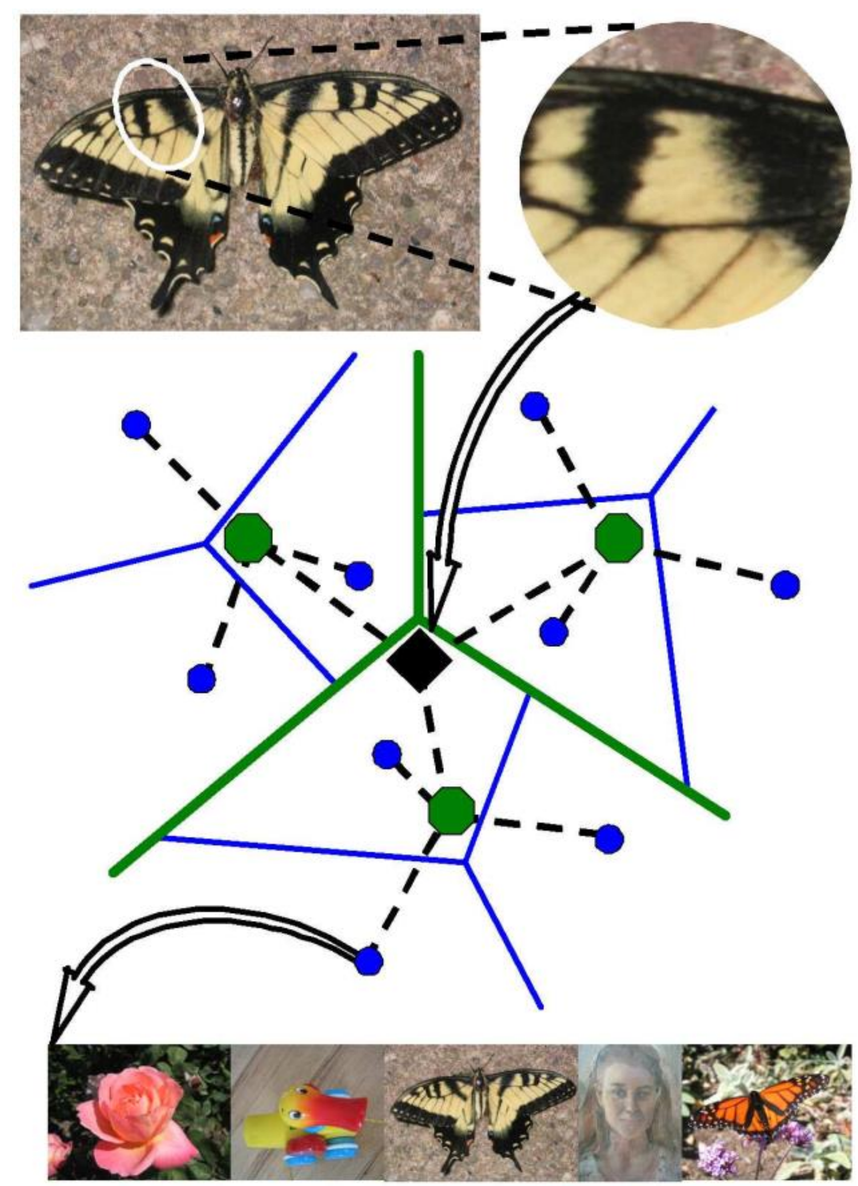
Appearance codebook

Source: B. Leibe

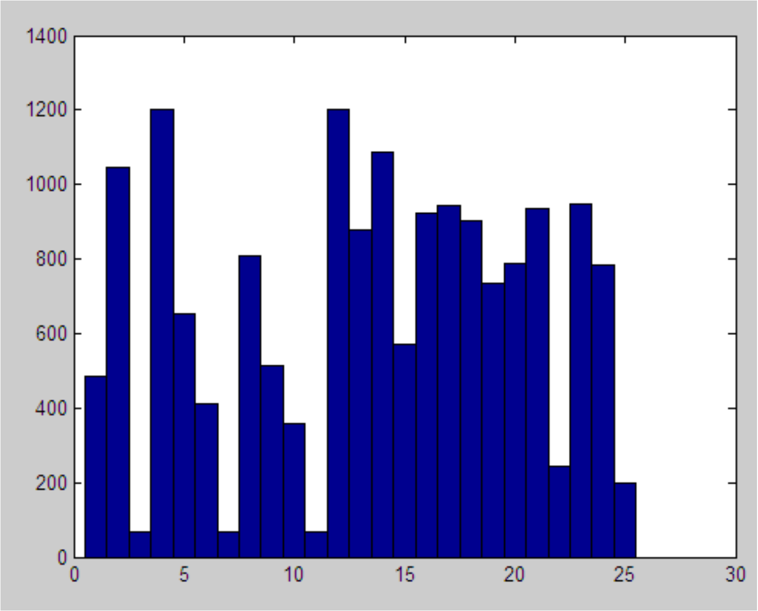
Visual vocabularies: Issues

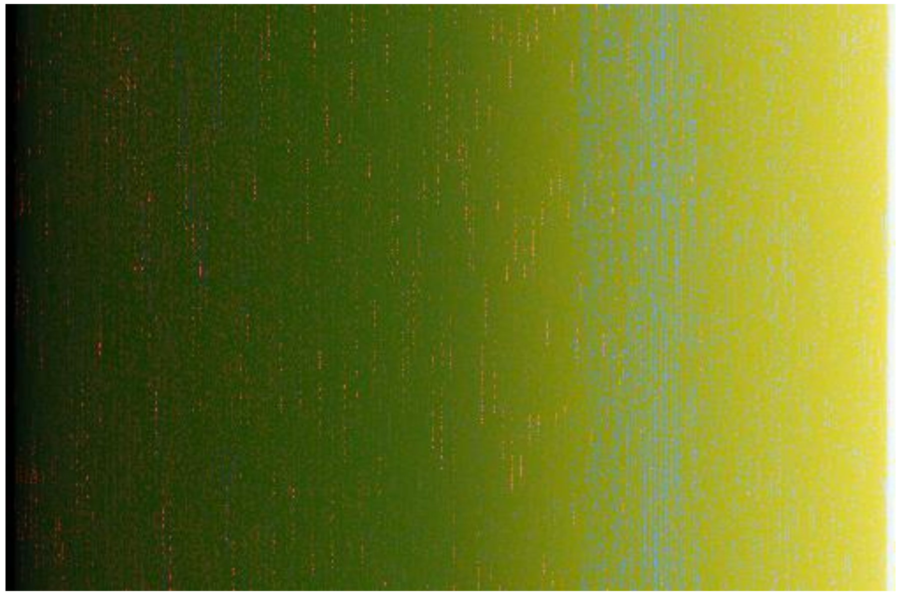
- How to choose vocabulary size?
  - Too small: visual words not representative of all patches
  - Too large: quantization artifacts, overfitting

- Computational efficiency
  - Vocabulary trees
    (Nister & Stewenius, 2006)



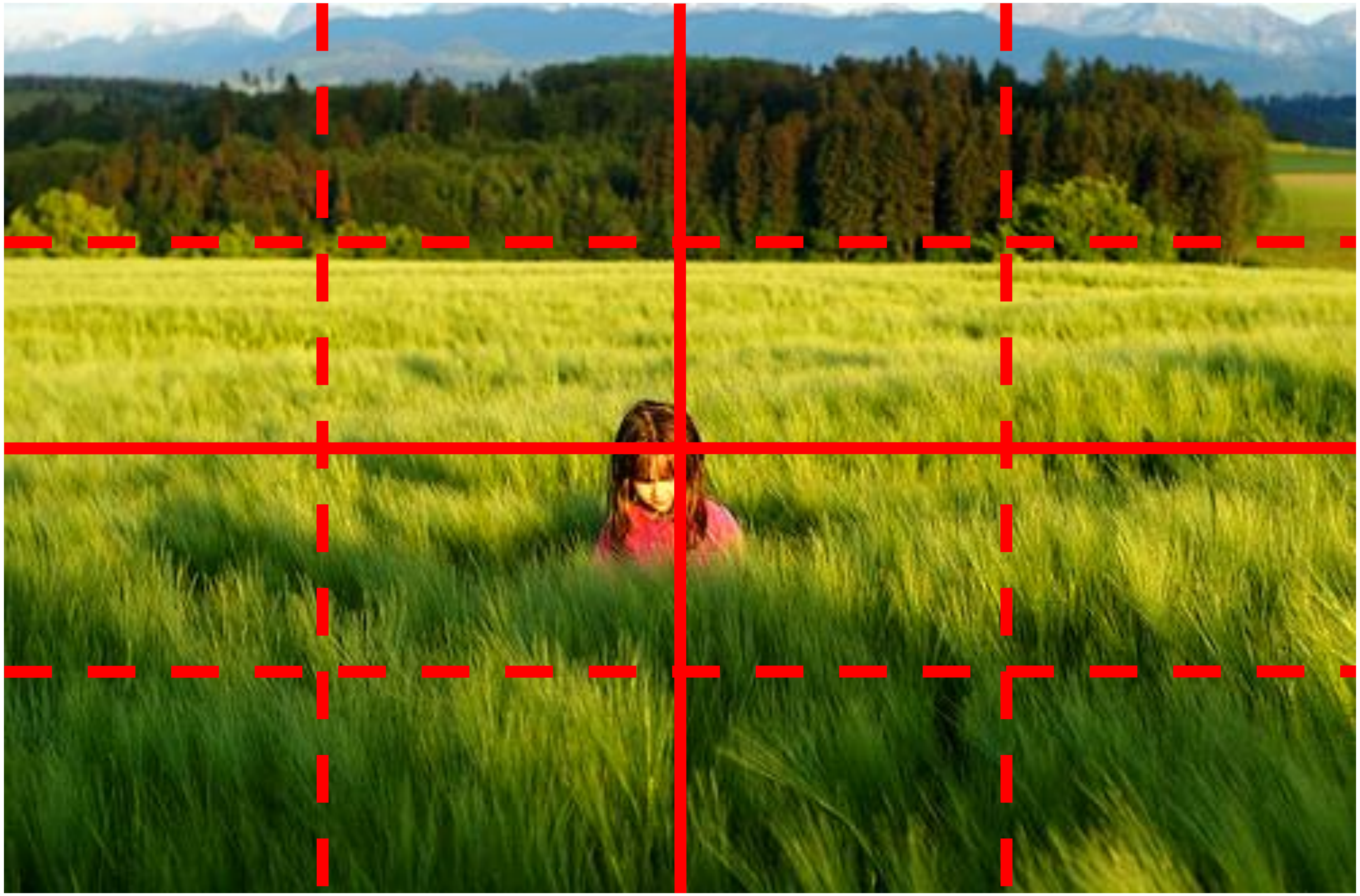
But what about layout?





All of these images have the same color histogram

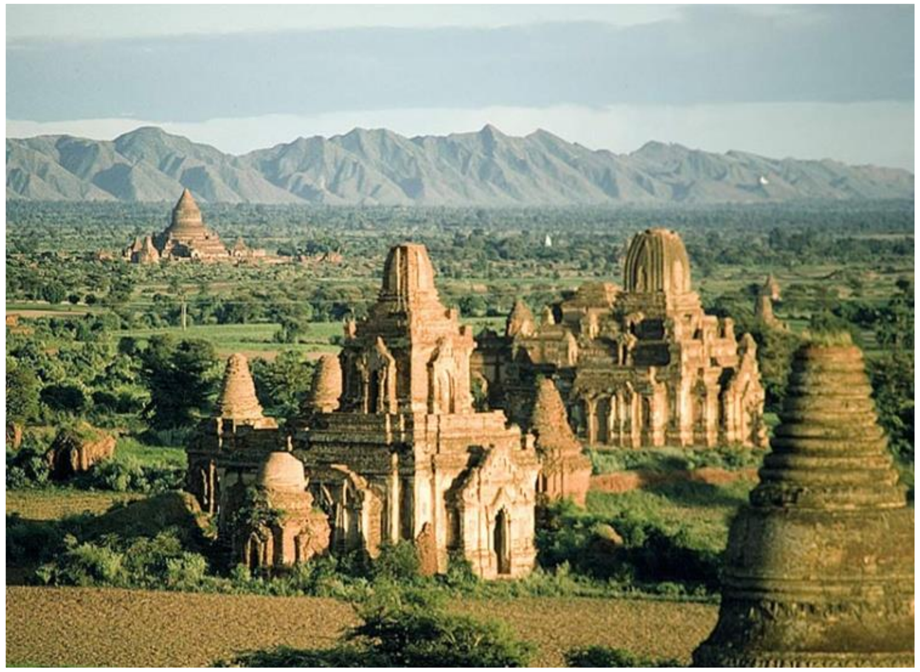
Spatial pyramid

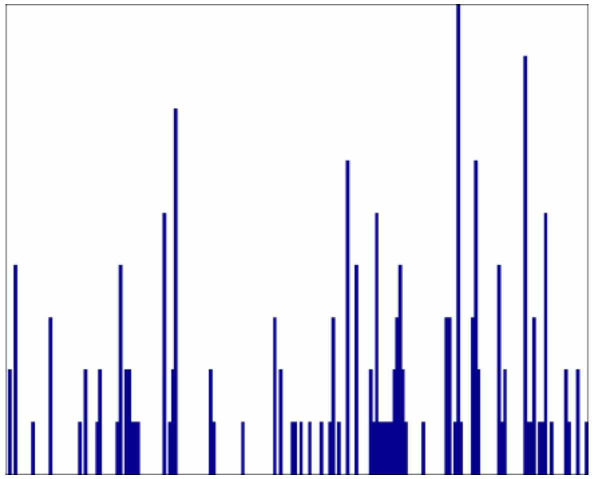


Compute histogram in each spatial bin

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution



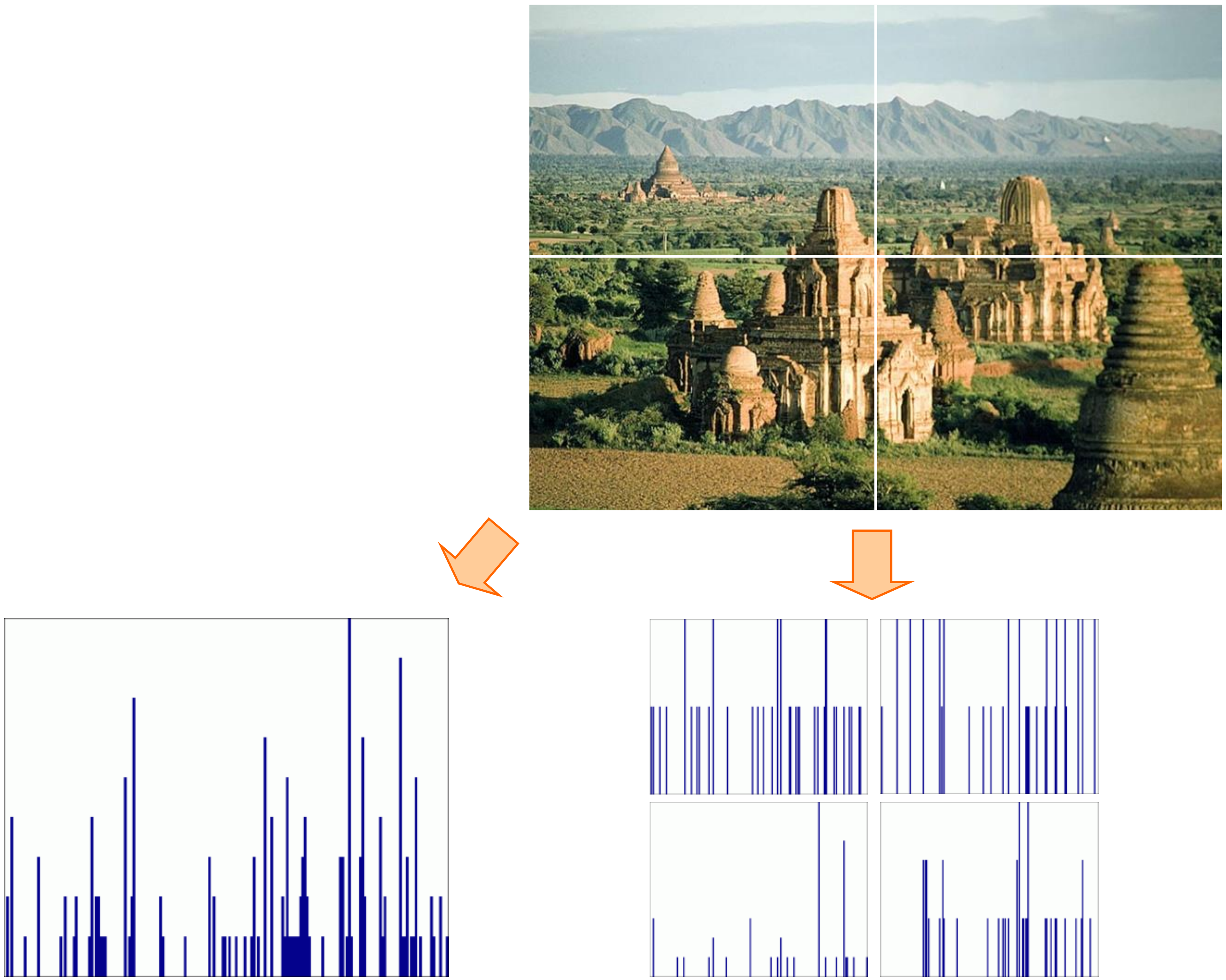


level 0

Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution



level 0 level 1

Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution

Lazebnik, Schmid & Ponce (CVPR 2006)

Representation

- Object as set of parts
  - Generative representation
- Model:
  - Relative locations between parts
  - Appearance of part
- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter

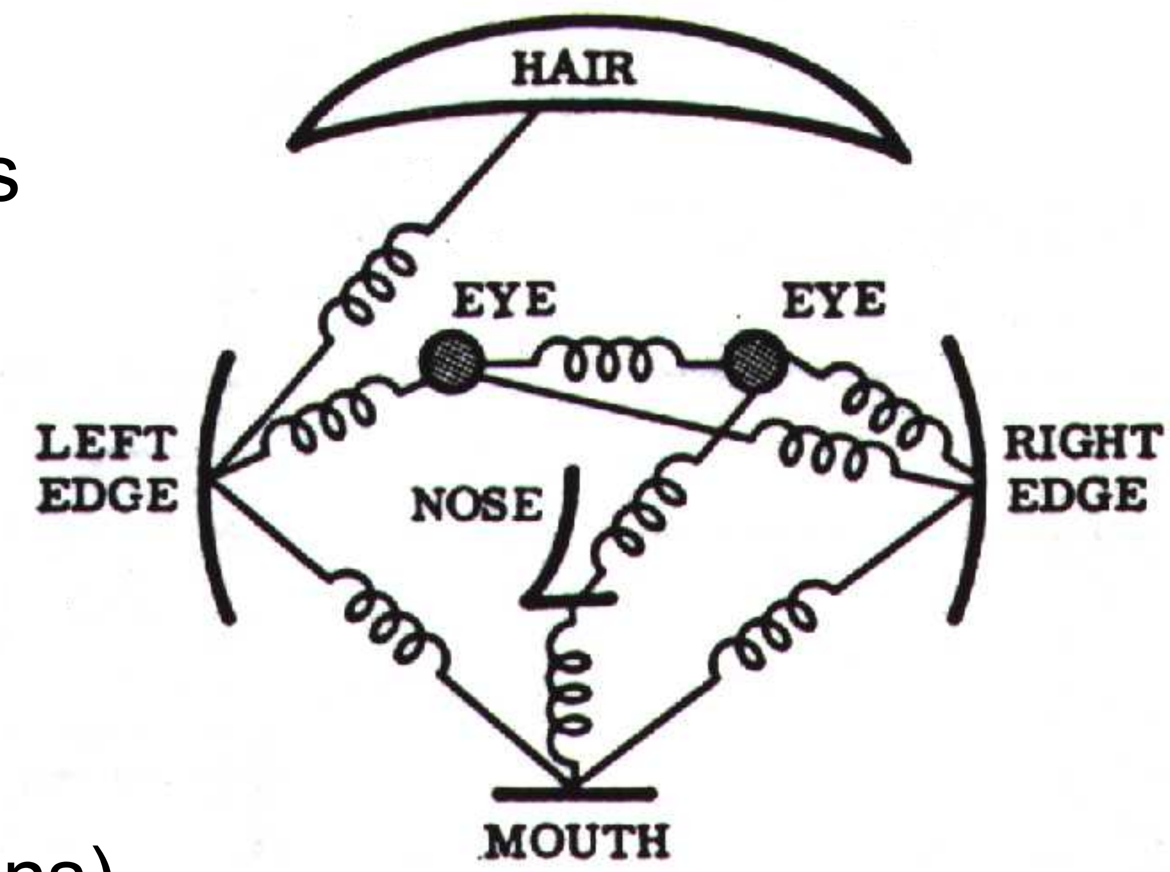


Figure from [Fischler & Elschlager 73]

The Representation and Matching of Pictorial Structures

MARTIN A. FISCHLER AND ROBERT A. ELSCHLAGER

Abstract—The primary problem dealt with in this paper is the following. Given some description of a visual object, find that object in an actual photograph. Part of the solution to this problem is the specification of a descriptive scheme, and a metric on which to base the decision of "goodness" of matching or detection.

We offer a combined descriptive scheme and decision metric which is general, intuitively satisfying, and which has led to promising experimental results. We also present an algorithm which takes the above descriptions, together with a matrix representing the intensities of the actual photograph, and then finds the described object in the matrix. The algorithm uses a procedure similar to dynamic programming in order to cut down on the vast amount of computation otherwise necessary.

One desirable feature of the approach is its generality. A new programming system does not need to be written for every new description; instead, one just specifies descriptions in terms of a certain set of primitives and parameters.

There are many areas of application: scene analysis and description, map matching for navigation and guidance, optical tracking, stereo compilation, and image change detection. In fact, the ability to describe, match, and register scenes is basic for almost any image processing task.

Index Terms—Dynamic programming, heuristic optimization, picture description, picture r
tation.

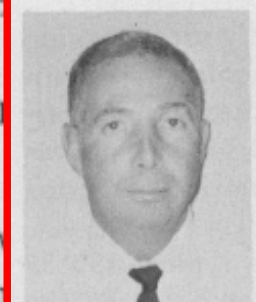
Manuscript received November 30, 1971; revised May 22, 1972, and August 21, 1972.

The authors are with the Lockheed Palo Alto Research Laboratory, Lockheed Missiles & Space Company, Inc., Palo Alto, Calif. 94304.

INT

THE PRIMARY
paper is the follo
a visual object, fi
graph. The object mig
complicated, such as an
can be linguistic, picto
photograph will be ca
dimensional array of gr
being sought is called t

This ability to find a
equivalently, to match
scenes, is basic for alm
Application to such are
tion, map matching for



Martin A. Fischler (S'57–M'58) was born in New York, N. Y., on February 15, 1932. He received the B.E.E. degree from the City College of New York, New York, in 1954 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, Calif., in 1958 and 1962, respectively.

He served in the U. S. Army for two years and held positions at the National Bureau of Standards and at Hughes Aircraft Corporation during the period 1954 to 1958. In 1958 he joined the technical staff of the Lockheed Missiles & Space Company, Inc., at the Lockheed Palo Alto Research Laboratory, Palo Alto, Calif., and currently holds the title of Staff Scientist. He has conducted research and published in the areas of artificial intelligence, picture processing, switching theory, computer organization, and information theory.

Dr. Fischler is a member of the Association for Computing Machinery, the Pattern Recognition Society, the Mathematical Association of America, Tau Beta Pi, and Eta Kappa Nu. He is currently an Associate Editor of the journal *Pattern Recognition* and is a past Chairman of the San Francisco Chapter of the IEEE Society on Systems, Man, and Cybernetics.

Robert A. Elschlager was born in Chicago, Ill., on May 25, 1943. He received the B.S. degree in mathematics from the University of Illinois, Urbana, in 1964, and the M.S. degree in mathematics from the University of California, Berkeley, in 1969.

Since then he has been an Associate Scientist with the Lockheed Missiles & Space Company, Inc., at the Lockheed Palo Alto Research Center, Palo Alto, Calif. His current interests are picture processing, operating systems, computer languages, and computer understanding.

Mr. Elschlager is a member of the American Mathematical Society, the Mathematical Association of America, and the Association for Symbolic Logic.

Object Detection with Discriminatively Trained Part Based Models

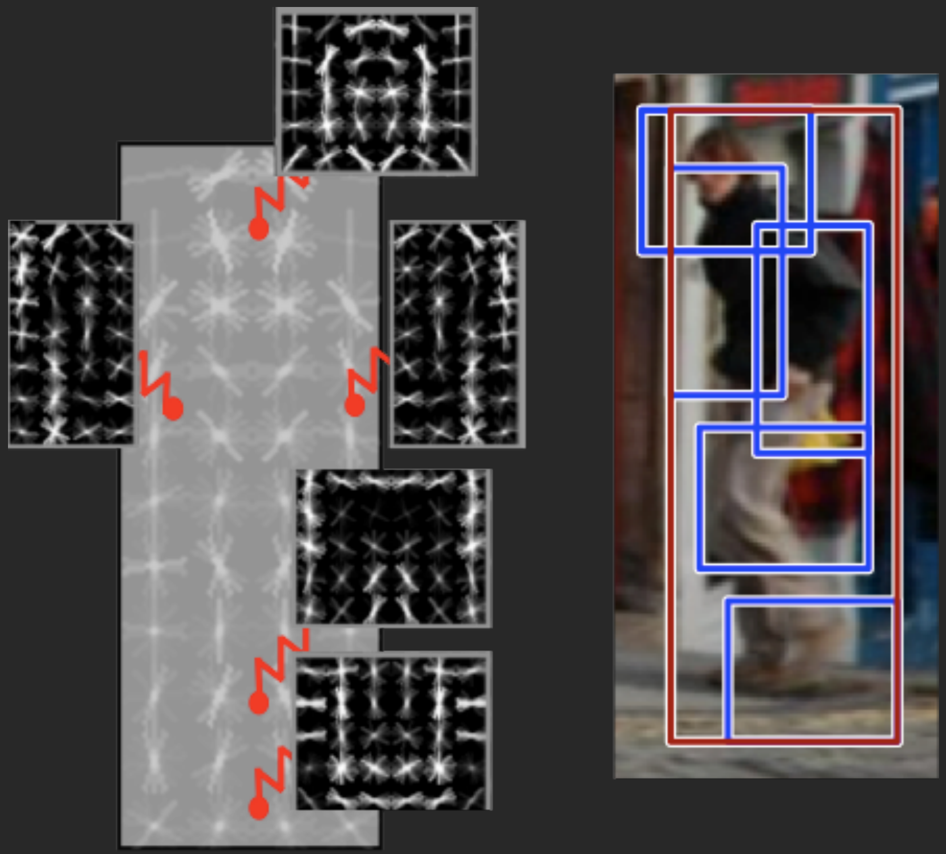
Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan

Abstract—We describe an object detection system based on mixtures of multiscale deformable part models. Our system is able to represent highly variable object classes and achieves state-of-the-art results in the PASCAL object detection challenges. While deformable part models have become quite popular, their value had not been demonstrated on difficult benchmarks such as the PASCAL datasets. Our system relies on new methods for discriminative training with partially labeled data. We combine a margin-sensitive approach for data-mining hard negative examples with a formalism we call *latent SVM*. A latent SVM is a reformulation of MI-SVM in terms of latent variables. A latent SVM is semi-convex and the training problem becomes convex once latent information is specified for the positive examples. This leads to an iterative training algorithm that alternates between fixing latent values for positive examples and optimizing the latent SVM objective function.

Index Terms—Object Recognition, Deformable Models, Pictorial Structures, Discriminative Training, Latent SVM

Combines pictorial structures with machine learning

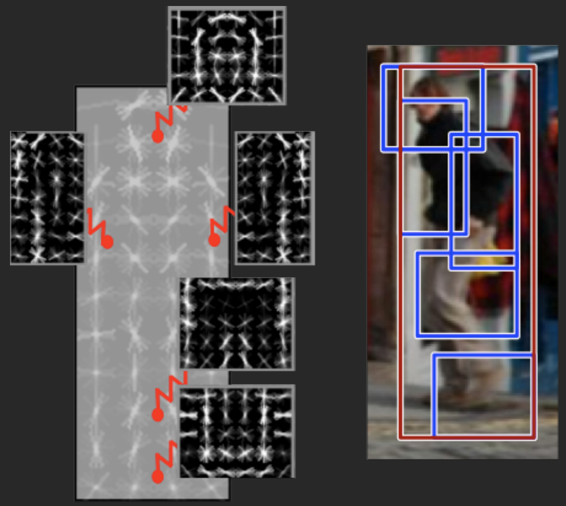
Deformable part models

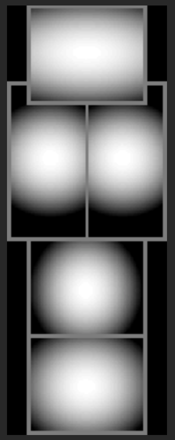


Model encodes local appearance + pairwise geometry

Source: Deva Ramanan

Scoring function





score(x,z)	$= \sum_i w_i \, \phi(x, z_i) +$	$\sum_{i,j} w_{ij} \, \Psi(z_i, z_j)$
x = image $z_i = (x_i, y_i)$ $z = \{z_1, z_2 \ldots\}$	part template scores	spring deformation model

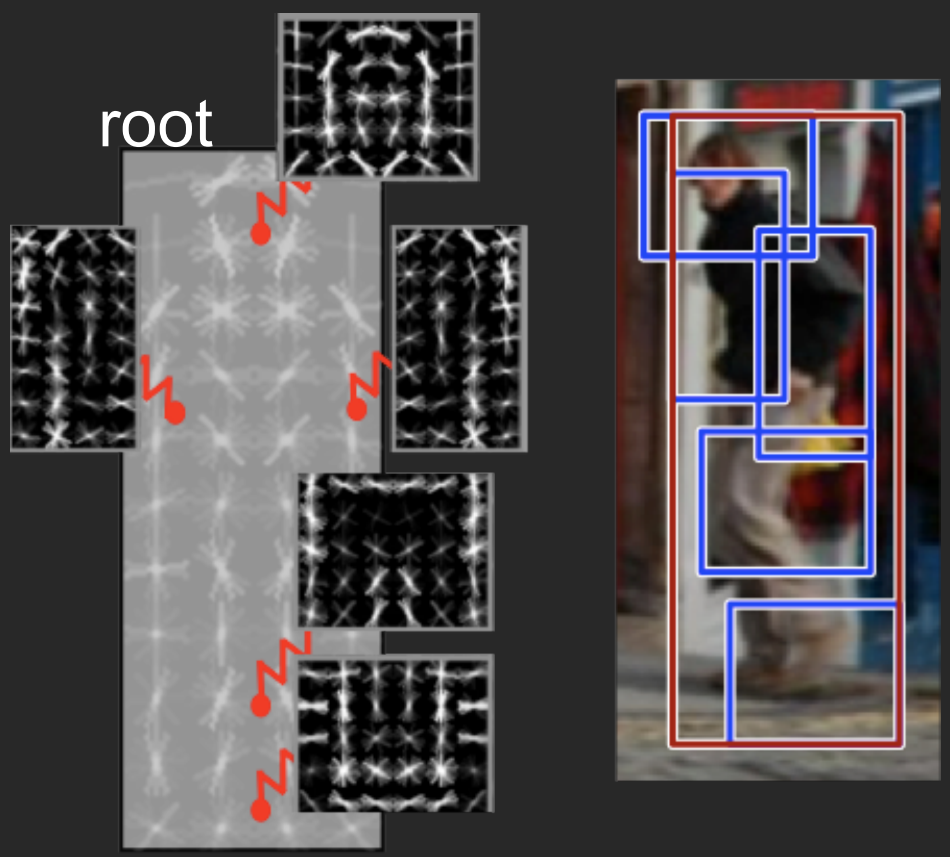
Score is linear in local templates w_i and spring parameters w_{ij}

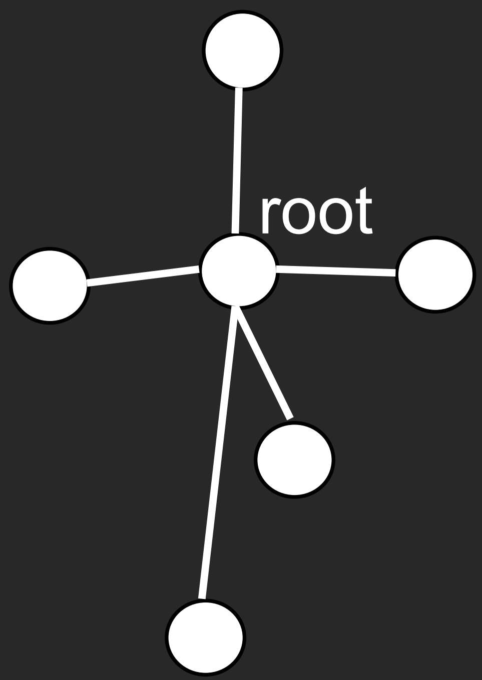
$$\text{score}(x,z) = w \cdot \Phi(x, z)$$

Source: Deva Ramanan

Inference: $\max_z$ score(x,z)

Felzenszwalb & Huttenlocher 05

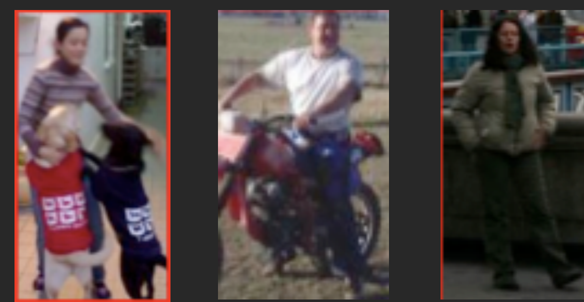




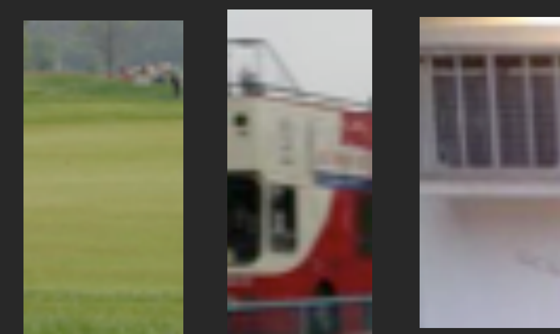
Star model: the location of the root filter is the anchor point
Given the root location, all part locations are independent

Source: Deva Ramanan

Latent SVMs



pos



neg

Given positive and negative training windows {xn}

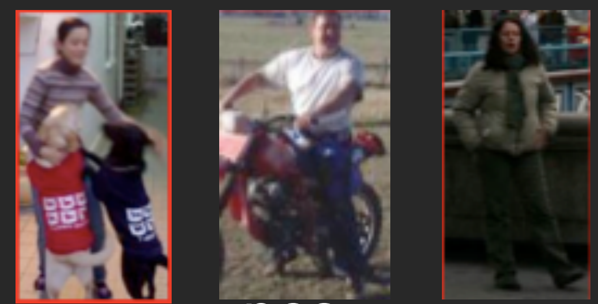
$$L(w) = ||w||^2 + \sum_{n \in \text{pos}} \max(0, 1 - f_w(x_n)) + \sum_{n \in \text{neg}} \max(0, 1 + f_w(x_n))$$

$$f_w(x) = \max_z w \cdot \Phi(x, z)$$

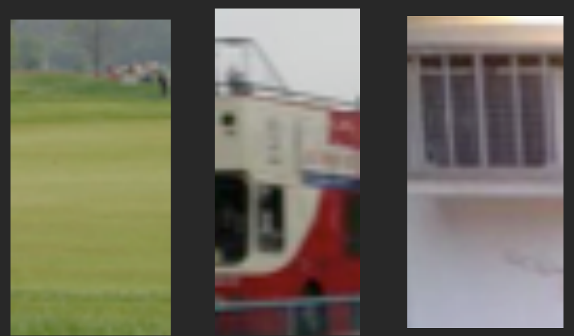
L(w) is “almost” convex

Source: Deva Ramanan

Latent SVMs



pos



neg

Given positive and negative training windows {xn}

$$L(w) = ||w||^2 + \sum_{n \in \text{pos}} \max(0, 1 - \cancel{f_w(x_n)}) + \sum_{n \in \text{neg}} \max(0, 1 + f_w(x_n))$$

$$w \cdot \Phi(x_n, z_n)$$

$$f_w(x) = \max_z w \cdot \Phi(x, z)$$

L(w) is convex if we fix latent values for positives

Source: Deva Ramanan

Coordinate descent

1) Given positive part locations, learn w with a convex program

$$w = \underset{w}{\operatorname{argmin}}\, L(w) \quad \text{with fixed} \quad \{z_n : n \in \text{pos}\}$$

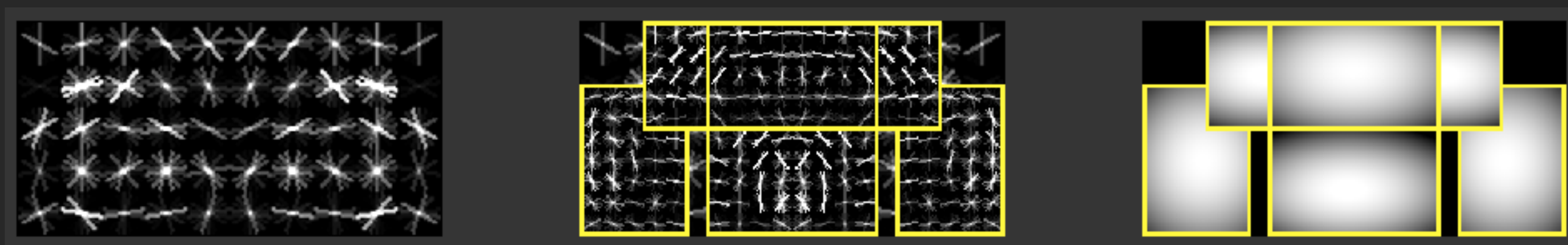
2) Given w, estimate part locations on positives

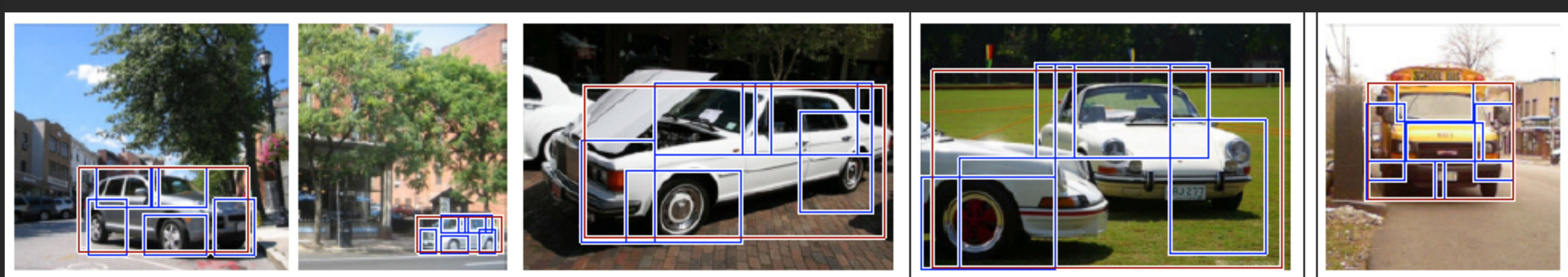
$$z_n = \underset{z}{\operatorname{argmax}}\, w \cdot \Phi(x_n, z) \quad \forall n \in \text{pos}$$

The above steps perform coordinate descent on a joint loss

Source: Deva Ramanan

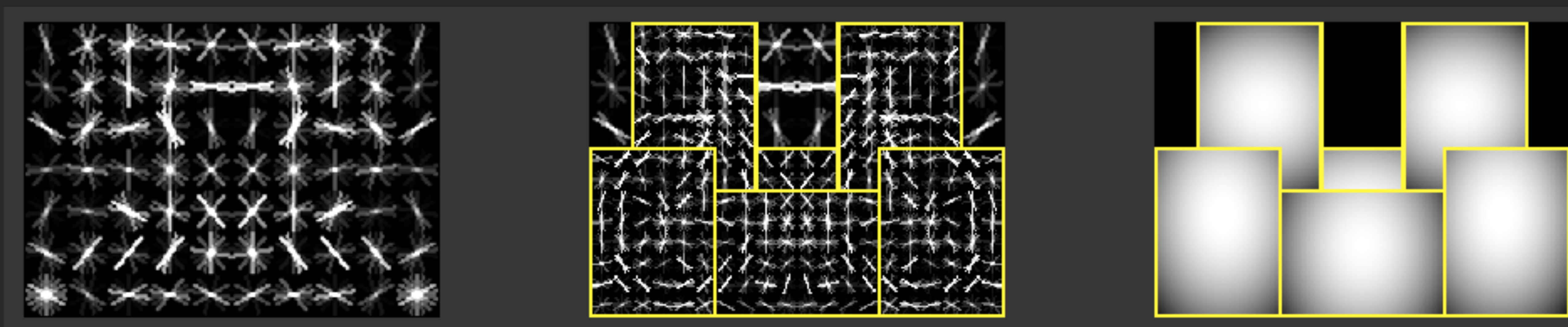
Example models





Source: Deva Ramanan

Example models



Source: Deva Ramanan