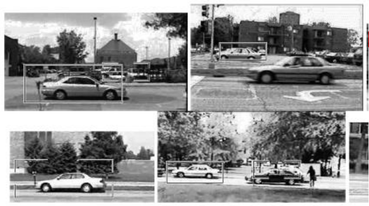
Large-Scale Image Collections

Yimeng Zhang and Henry Shu

10/6/11

Adapted Slides from Feifei's and Rob's slides

Datasets and computer vision

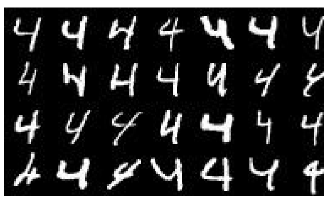


UIUC Cars (2004)
S. Agarwal, A. Awan, D. Roth

CMU/VASC Faces (1998)
H. Rowley, S. Baluja, T. Kanade

FERET Faces (1998)
P. Phillips, H. Wechsler, J. Huang, P. Raus

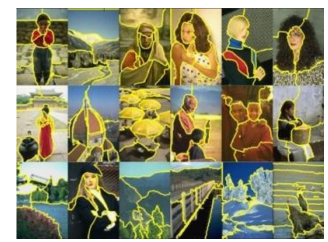
COIL Objects (1996)
S. Nene, S. Nayar, H. Murase



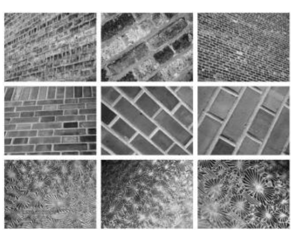
MNIST digits (1998-10)
Y LeCun & C. Cortes

KTH human action (2004)
I. Leptev & B. Caputo

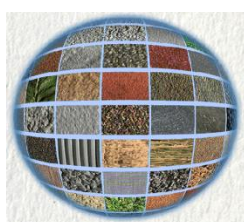
Sign Language (2008)
P. Buehler, M. Everingham, A. Zisserman



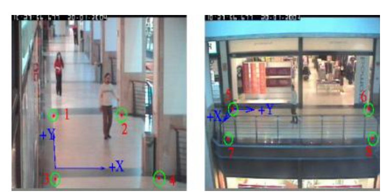
Segmentation (2001)
D. Martin, C. Fowlkes, D. Tal, J. Malik.



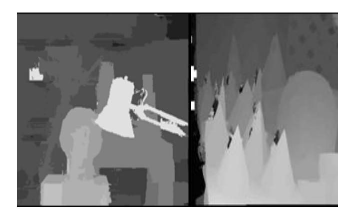
3D Textures (2005)
S. Lazebnik, C. Schmid, J. Ponce



CuRRET Textures (1999)
K. Dana B. Van Ginneken S. Nayar J. Koenderink

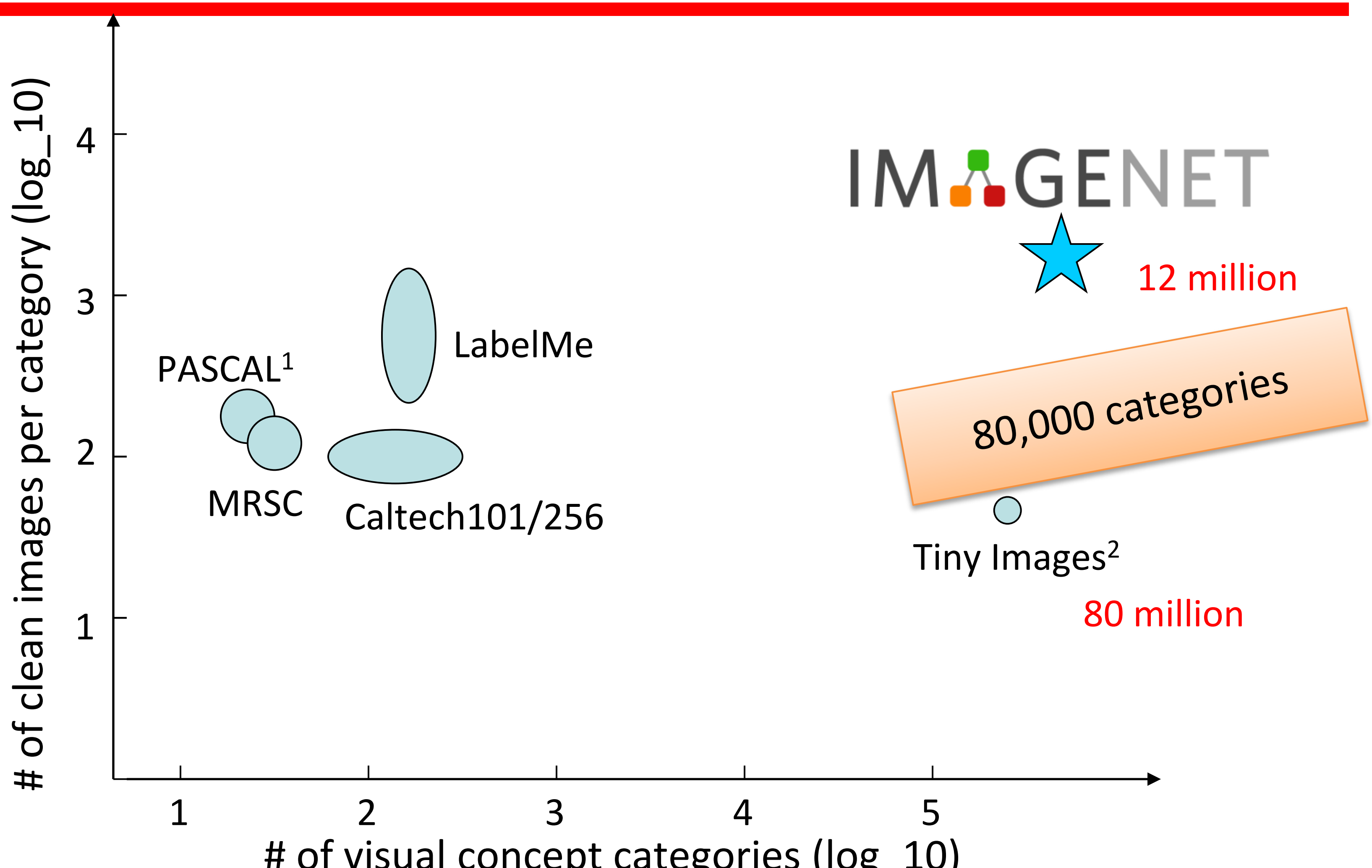


CAVIAR Tracking (2005)
R. Fisher, J. Santos-Victor J. Crowley



Middlebury Stereo (2002)
D. Scharstein R. Szeliski

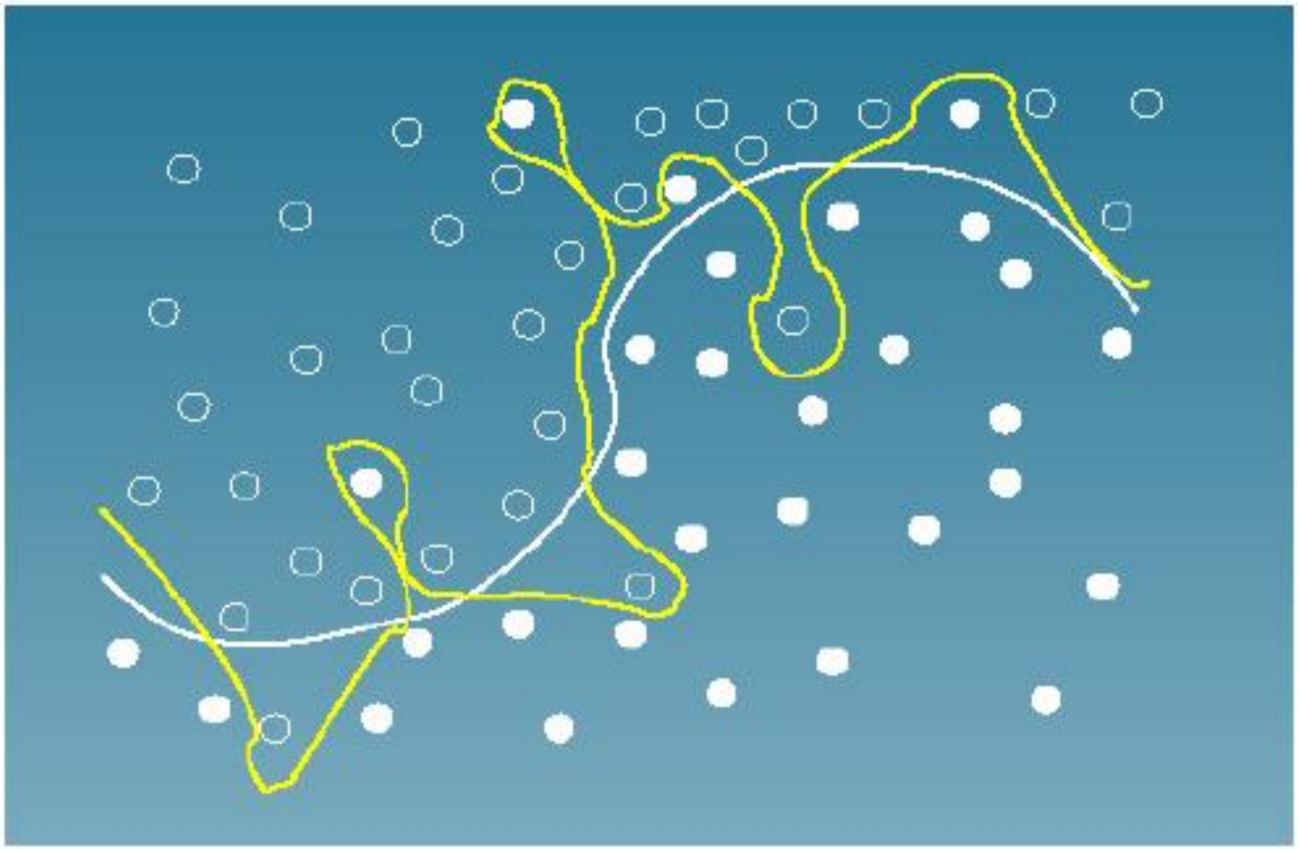
Datasets



1. Excluding the Caltech101 datasets from PASCAL
2. No image in this dataset is human annotated. The # of clean images per category is a rough estimation

Why large dataset?

More training data → Less overfitting



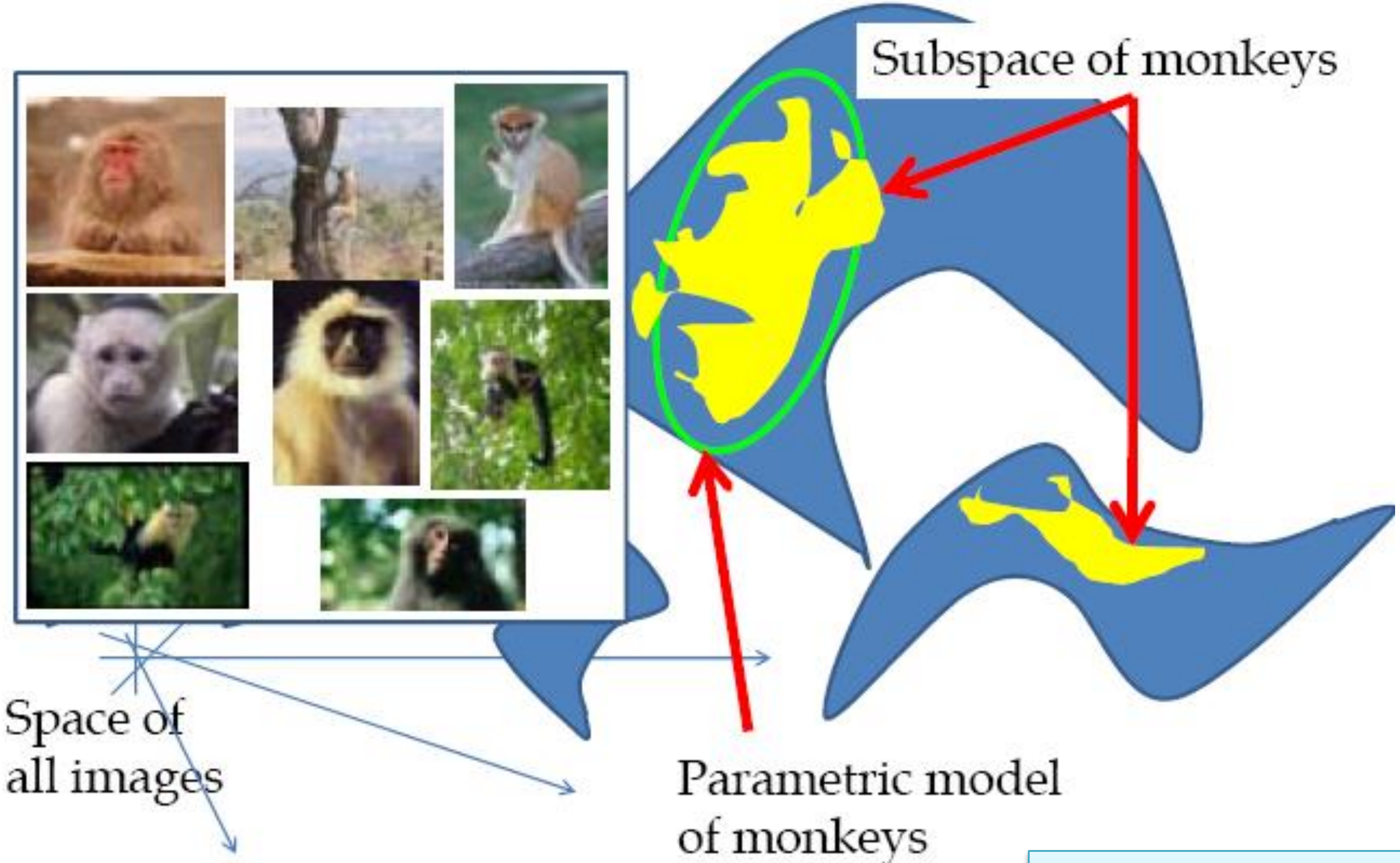
Parametric vs. Non-parametric

Enough amounts of data

→No need for sophisticated learning algorithms
(parametric models)

Nearest neighbor approach is enough
(non-parametric models)

Object Recognition



Parametric Model

Nearest Neighbor

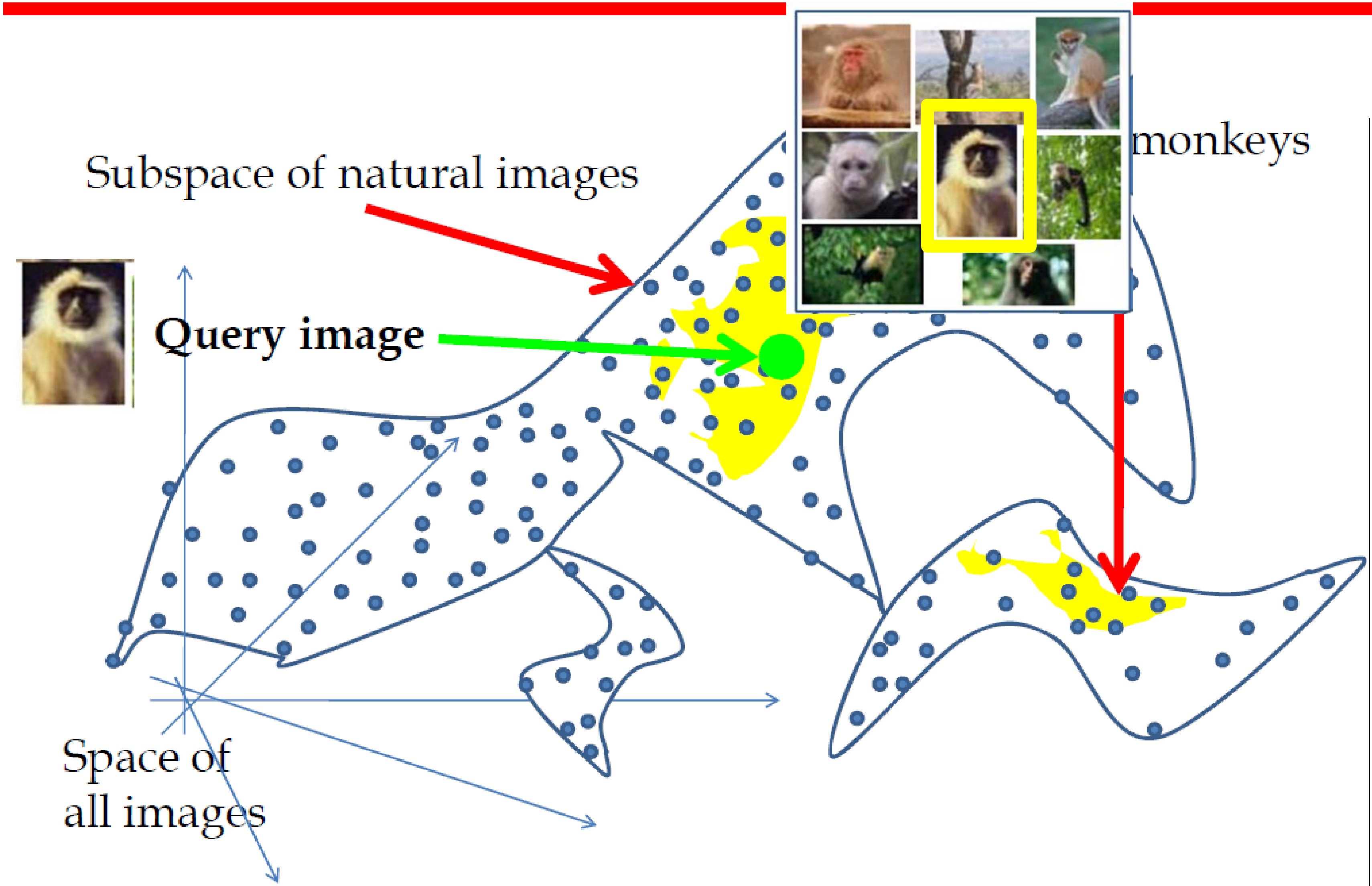
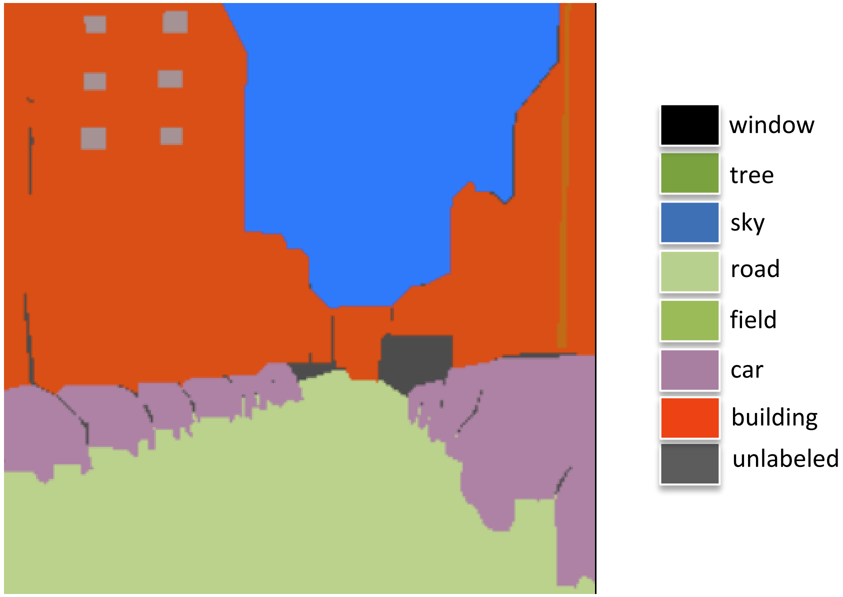


Image Labeling

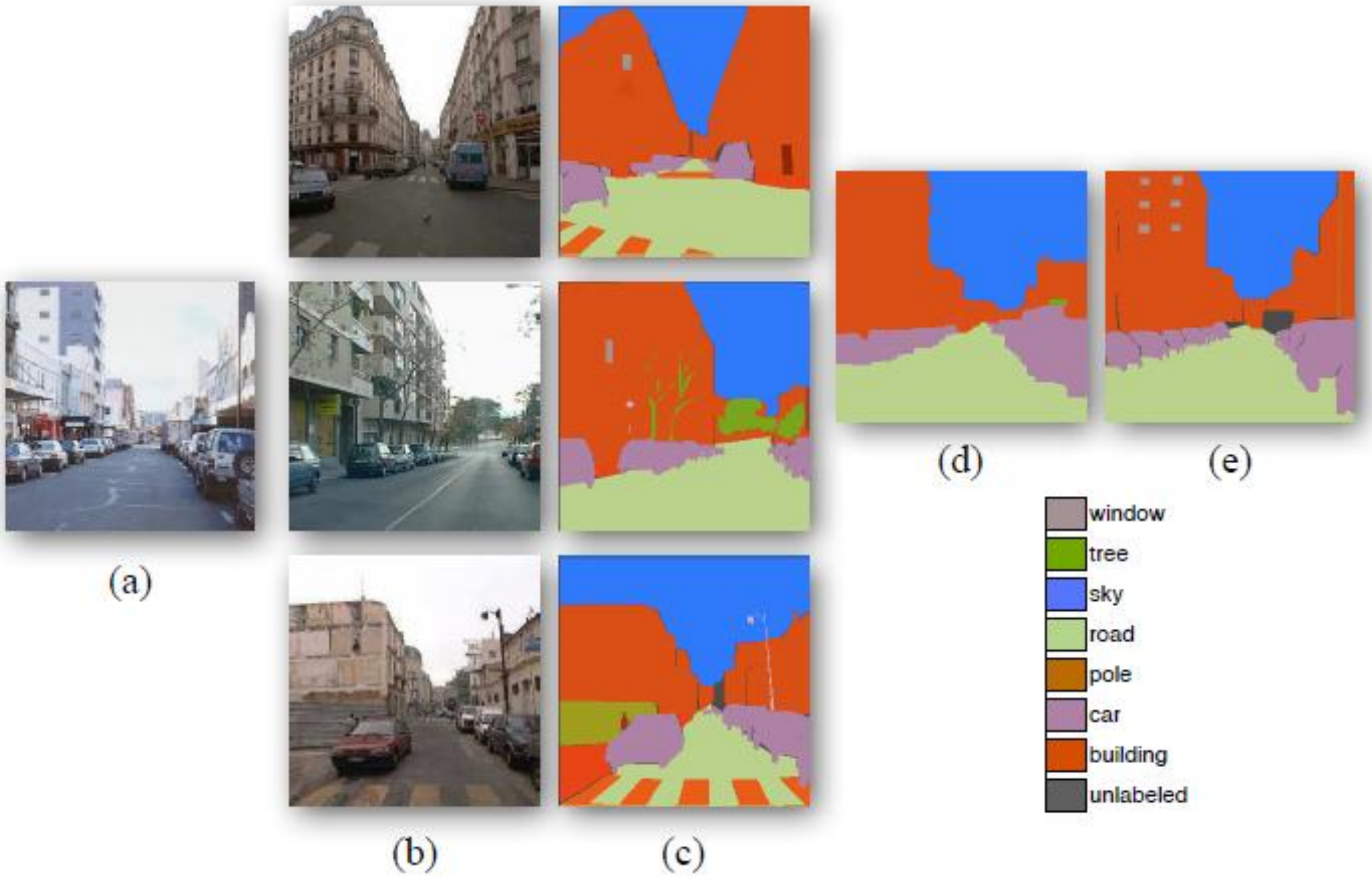
Input



Output

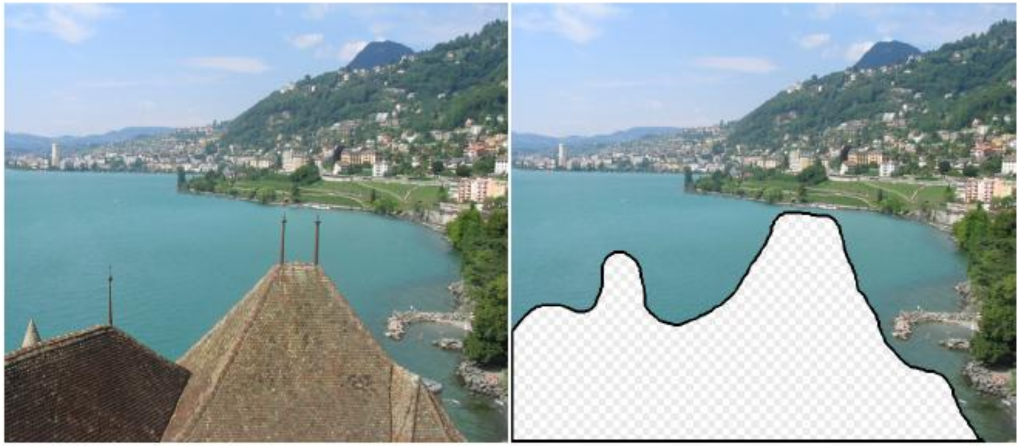
Traditional Method: learn the local appearance for each category, smooth with a MRF/CRF model

Label Transfer



[Liu et al., CVPR09]

Scene Completion



Original Image　　　　Input

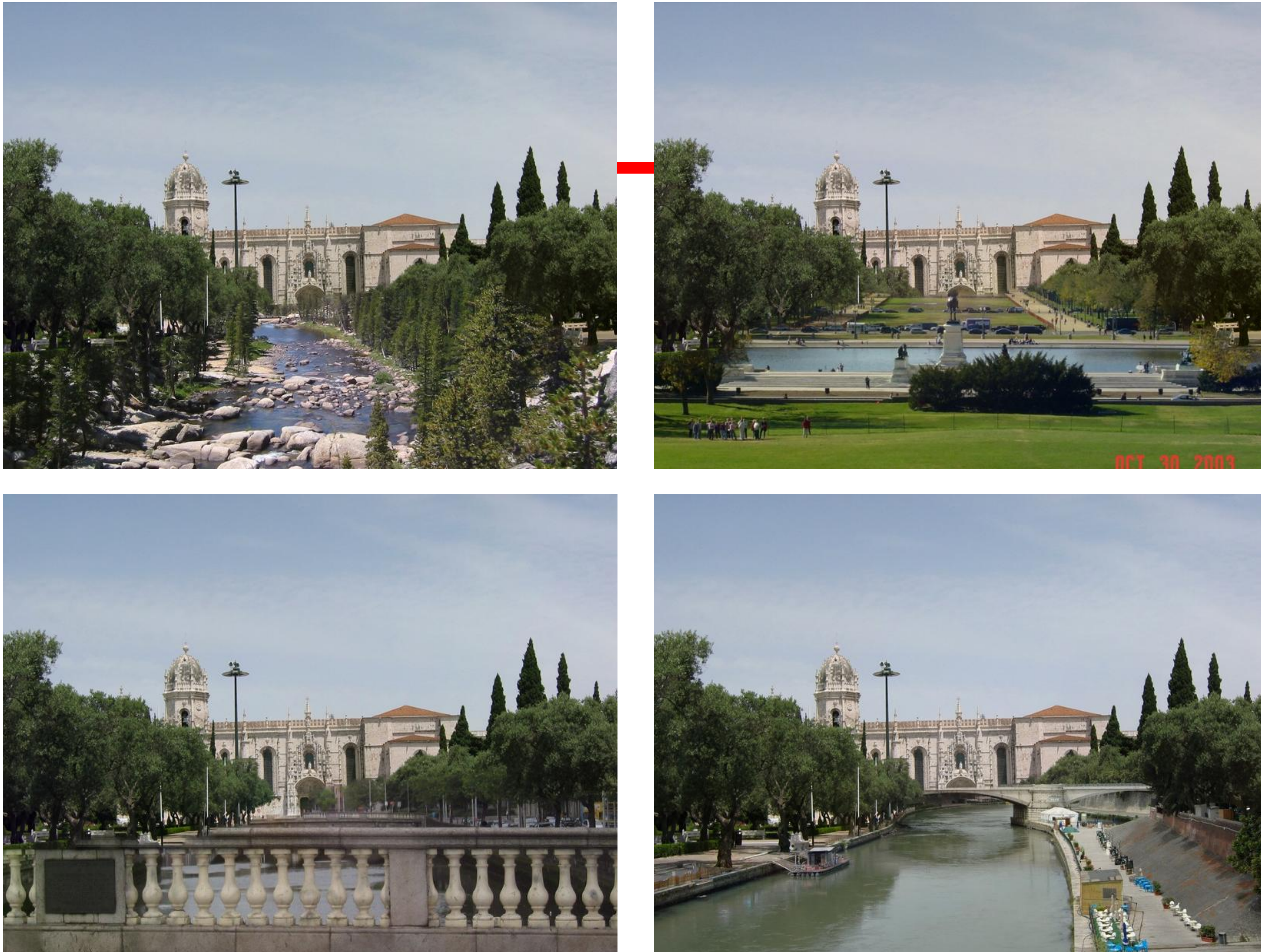
Scene Matches　　　　Output

Context matching + blending

Hays and Efros, SIGGRAPH07

Hays and Efros, SIGGRAPH 2007

Hays and Efros, SIGGRAPH 2007



Hays and Efros, SIGGRAPH 2007

Hays and Efros, SIGGRAPH 2007

Hays and Efros, SIGGRAPH 2007

Image Representation

80 million tiny images: a large dataset for non-parametric object and scene recognition. Torralba et al., PAMI 2008.

Small Codes and Large Image Databases for Recognition, Torralba et al., CVPR 2008

Slides adapted from Rob and Antonio’s slides

How much memory do we need?

- For computation, representation must fit in memory

- Google has few billion images (10^9)

Big PC has ~10 Gbytes

→ Budget of 100 bits/image

1 Megapixel image (10^7 bits)

⇨ Need serious dimension reduction

First Attempt

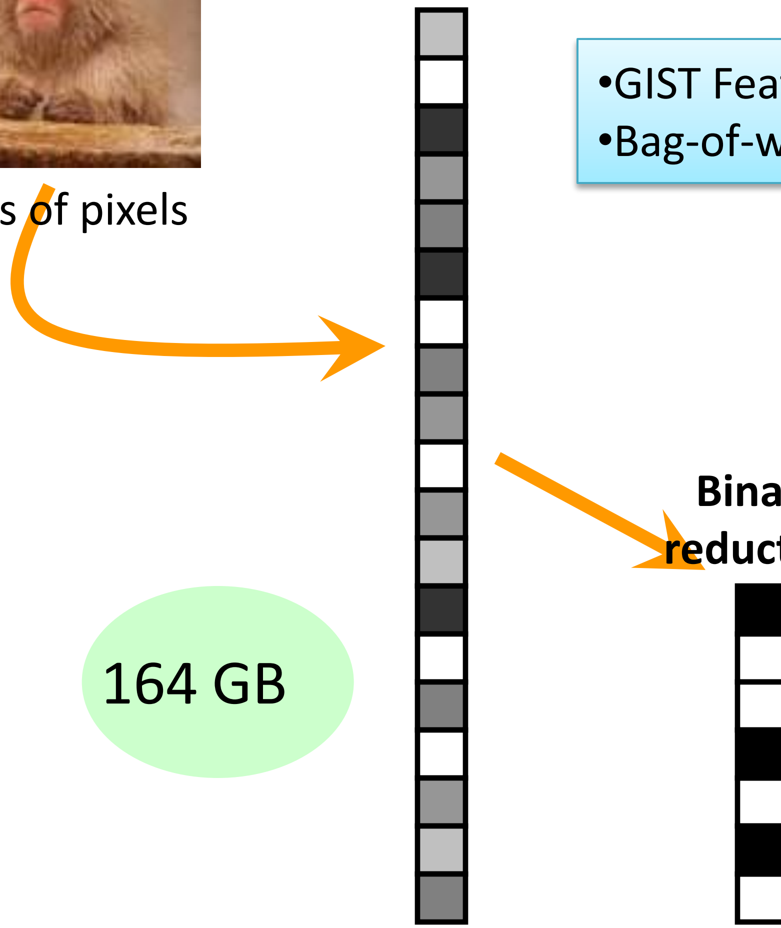
Reduce the resolution

32×32 [Torralba PAMI 08]

Binary Reduction

Lots of pixels

Feature vector



- GIST Features
- Bag-of-words histogram

80 million images?

164 GB

320 MB

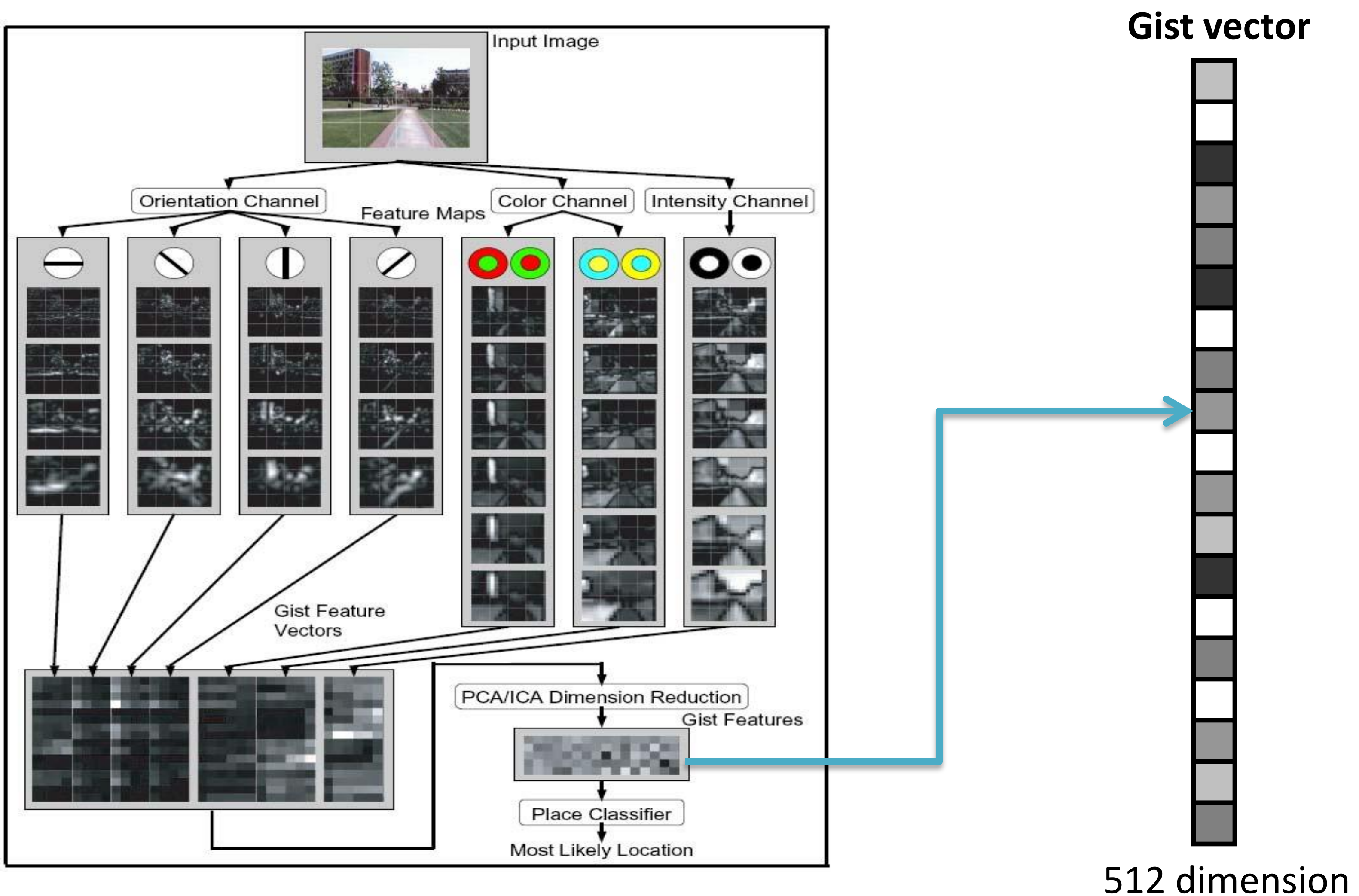
512 values

32 bits

[Torralba, PAMI08]

GIST

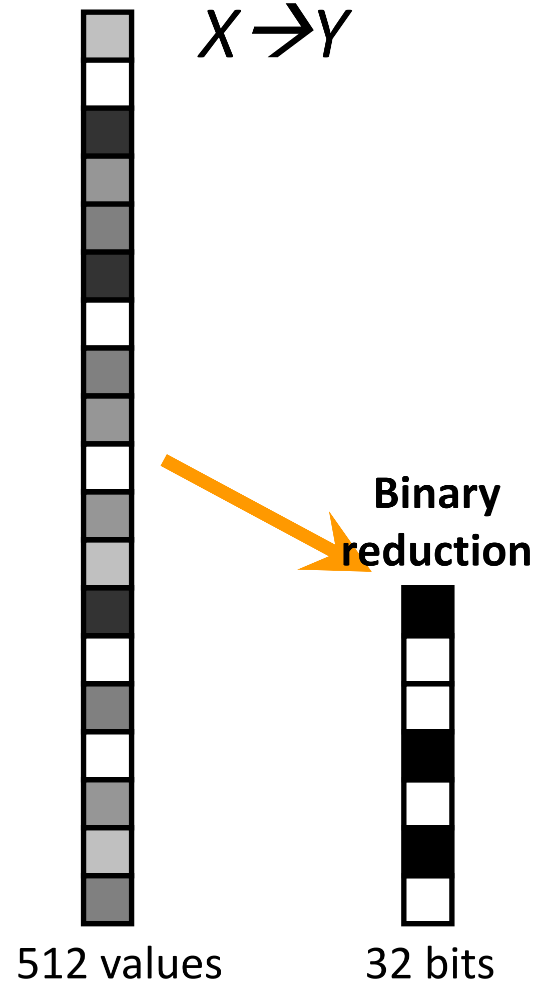
Abstract representation of the scene



Binary Code

Project data to a low dimensional binary space that preserves the *nearest neighbor* relationships

$X \rightarrow Y$



$$y_i = f(x_i) = [h_1(x_i), h_2(x_i), \ldots, h_k(x_i)]$$

$h_j(x_i)$ is binary

Hashing function

Hamming Distance

Definition: the hamming distance between two equal length binary strings is the number of positions for which the bits are different

$$\|1011101, 1001001\|_H = 2$$

$$\|1110101, 1111101\|_H = 1$$

Fast to compute

Binary Code Methods

- Locally Sensitive Hashing

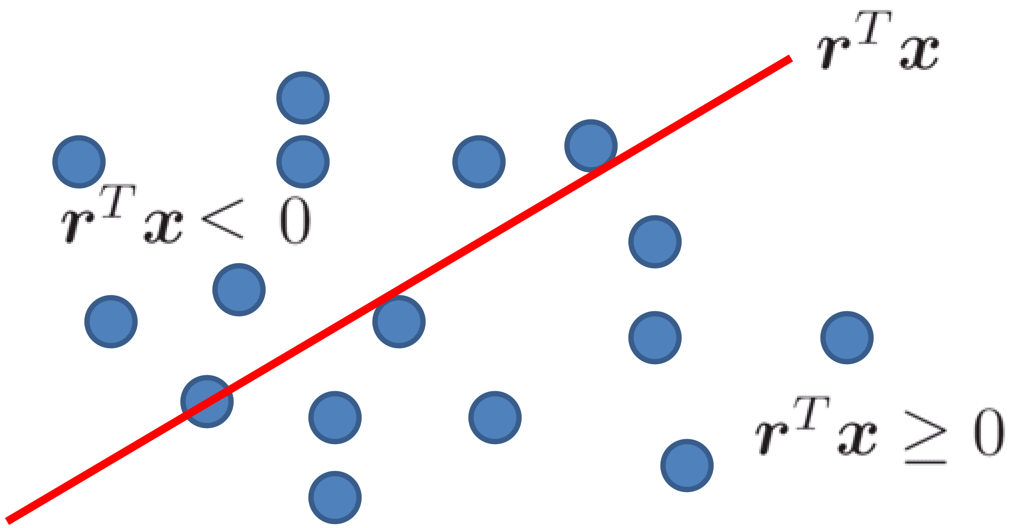
- Learning based method
  - Boost Similarity Sensitive Coding
  - Restricted Boltzmann Machines

Locally Sensitive Hashing

The hashing function of LSH to produce Hash Code

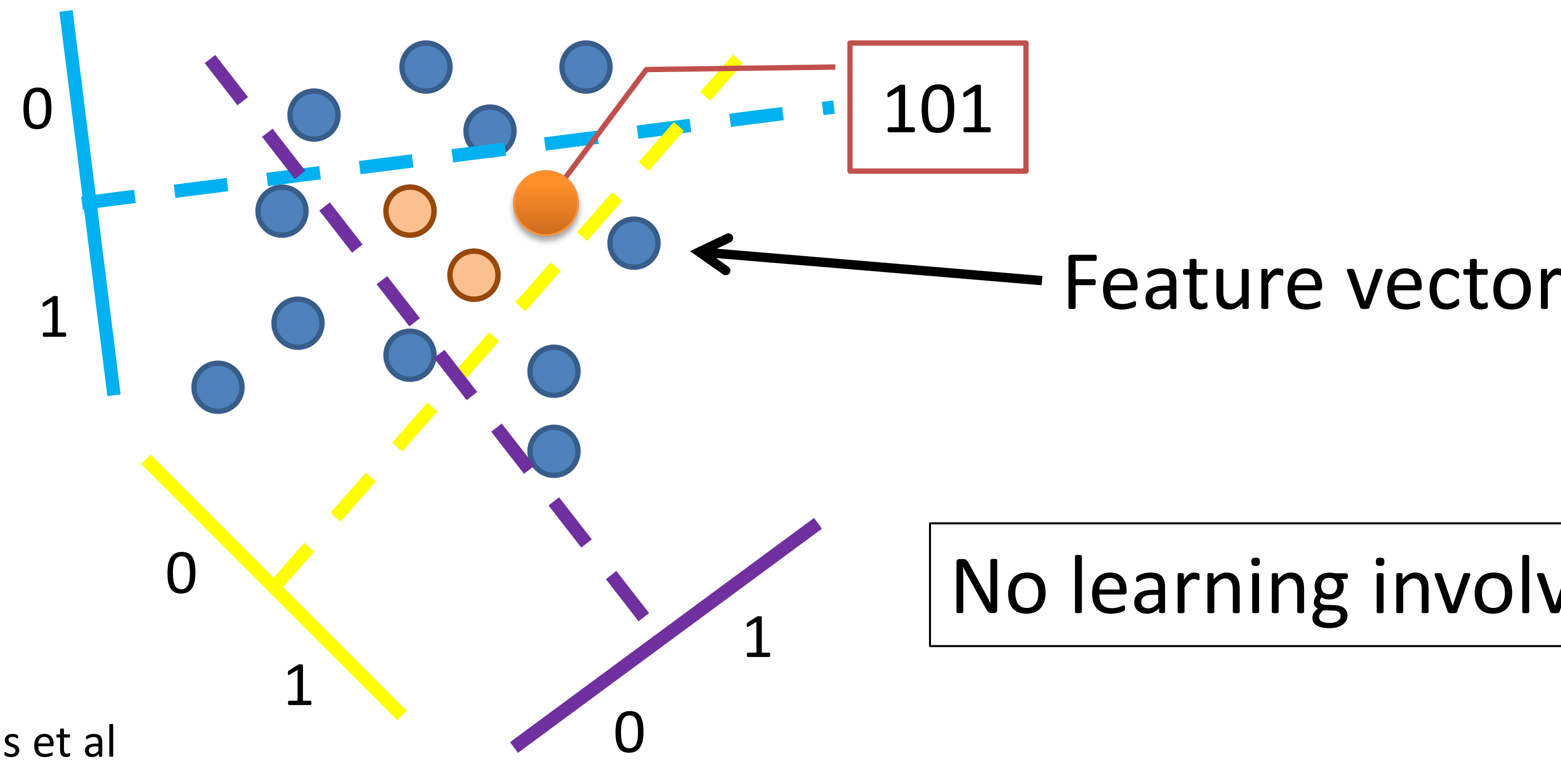
$$h_{\boldsymbol{r}}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{r}^T\boldsymbol{x} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$\boldsymbol{r}^T\boldsymbol{x} \geq 0$ is a hyperplane separating the space (next page for example)



LSH-Hyperplane

- Take **random** projections of data $r^T x$
- Quantize each projection with few bits



No learning involved

Fergus et al

Binary Code Methods

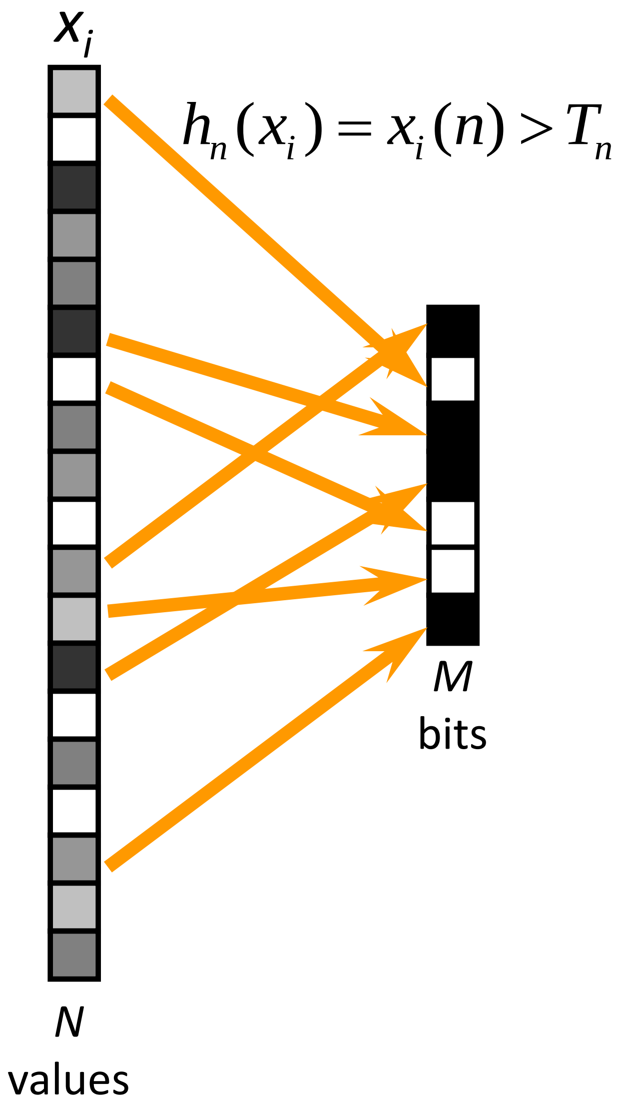
- Locally Sensitive Hashing

- Learning based method
  - Boost Similarity Sensitive Coding
  - Restricted Boltzmann Machines

BoostSSC

Use *boosting* algorithm to select the index and threshold

$$y_i = [h_1(x_i), h_2(x_i), \ldots, h_k(x_i)]$$



Training Examples

Positive example: images pairs that are nearest neighbors

Negative example: image pairs that are not neighbors

&

= 1

&

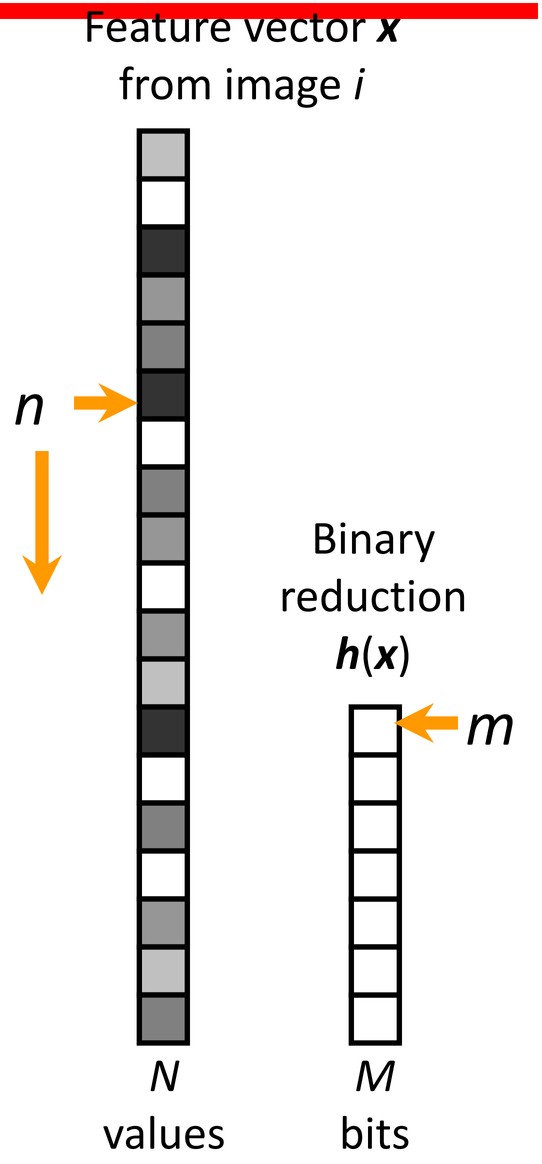
= -1

BoostSSC Training

$$h_n(x_i) = x_i(n) > T_n$$

At each iteration:

Select the index and threshold that minimizes a weighted *error/loss* of the entire training images

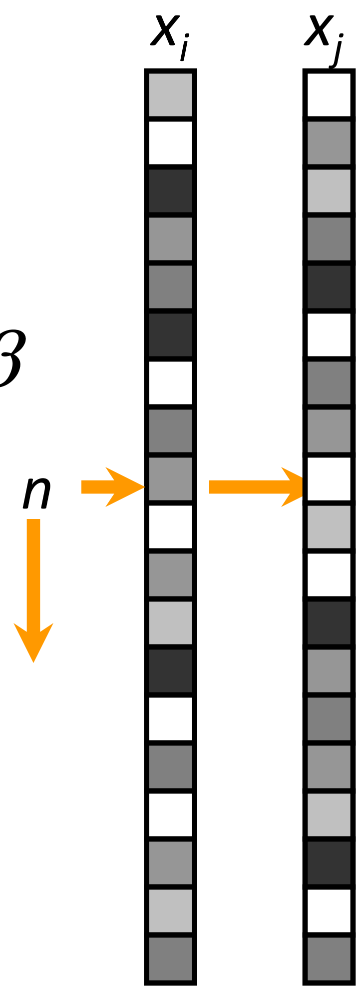


Loss function

- Weak classifier output:

$$h_n(x_i) == h_n(x_j)$$

$$f_n(x_i, x_j) = \alpha[(x_i(n) > T_n) == (x_j(n) > T_n)] + \beta$$



N values

- Loss function

Label: positive/negative

$$\sum_{k=1}^{K} w_n^k (z_k - f_n(x_i^k, x_j^k))^2$$

Weight for a training pair (defined with the loss)

Binary Code Methods

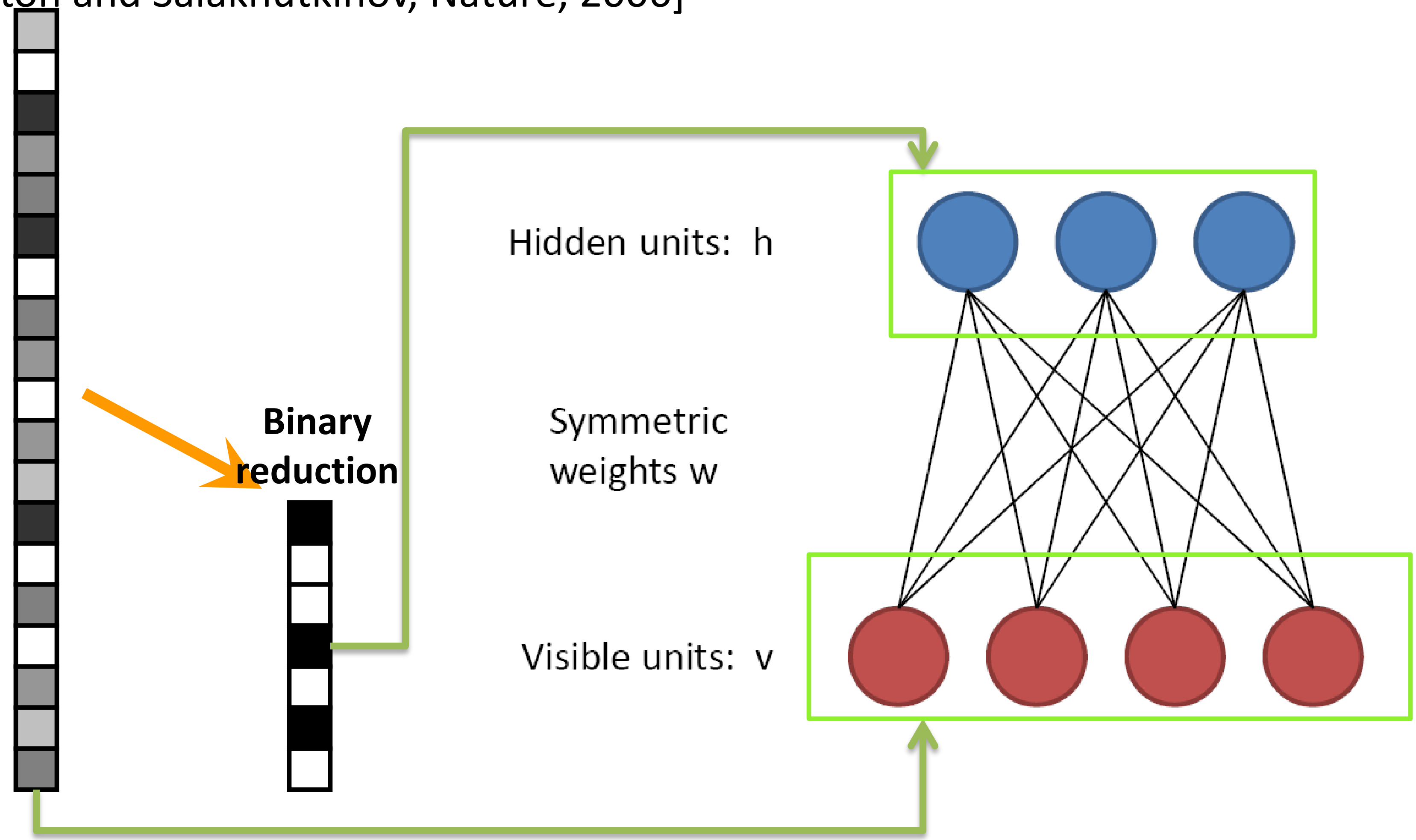
- Locally Sensitive Hashing

- Learning based method
  - Boost Similarity Sensitive Coding
  - Restricted Boltzmann Machines

Restricted Boltzmann Machine (RBM)

Network of binary stochastic units

[Hinton and Salakhutkinov, Nature, 2006]



RBM Training - Unsupervised

Joint Energy function

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i \in \text{visible}} b_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_i w_{ij}$$

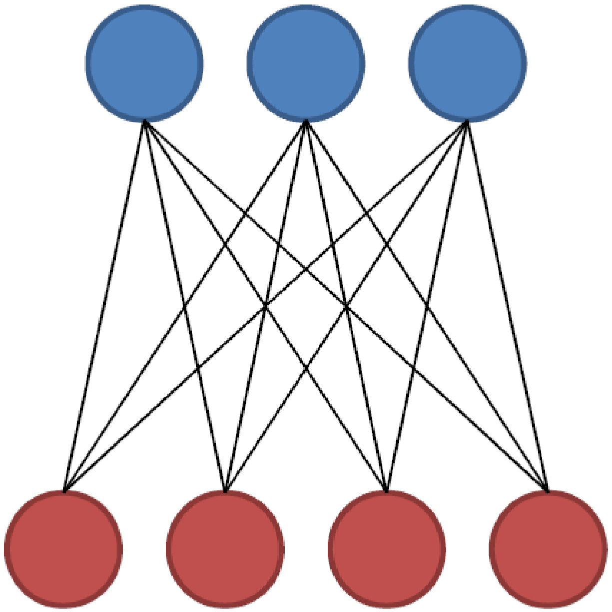
Learn to maximize

$$p(\mathbf{v}) = \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{u},\mathbf{g}} e^{-E(\mathbf{u},\mathbf{g})}}$$

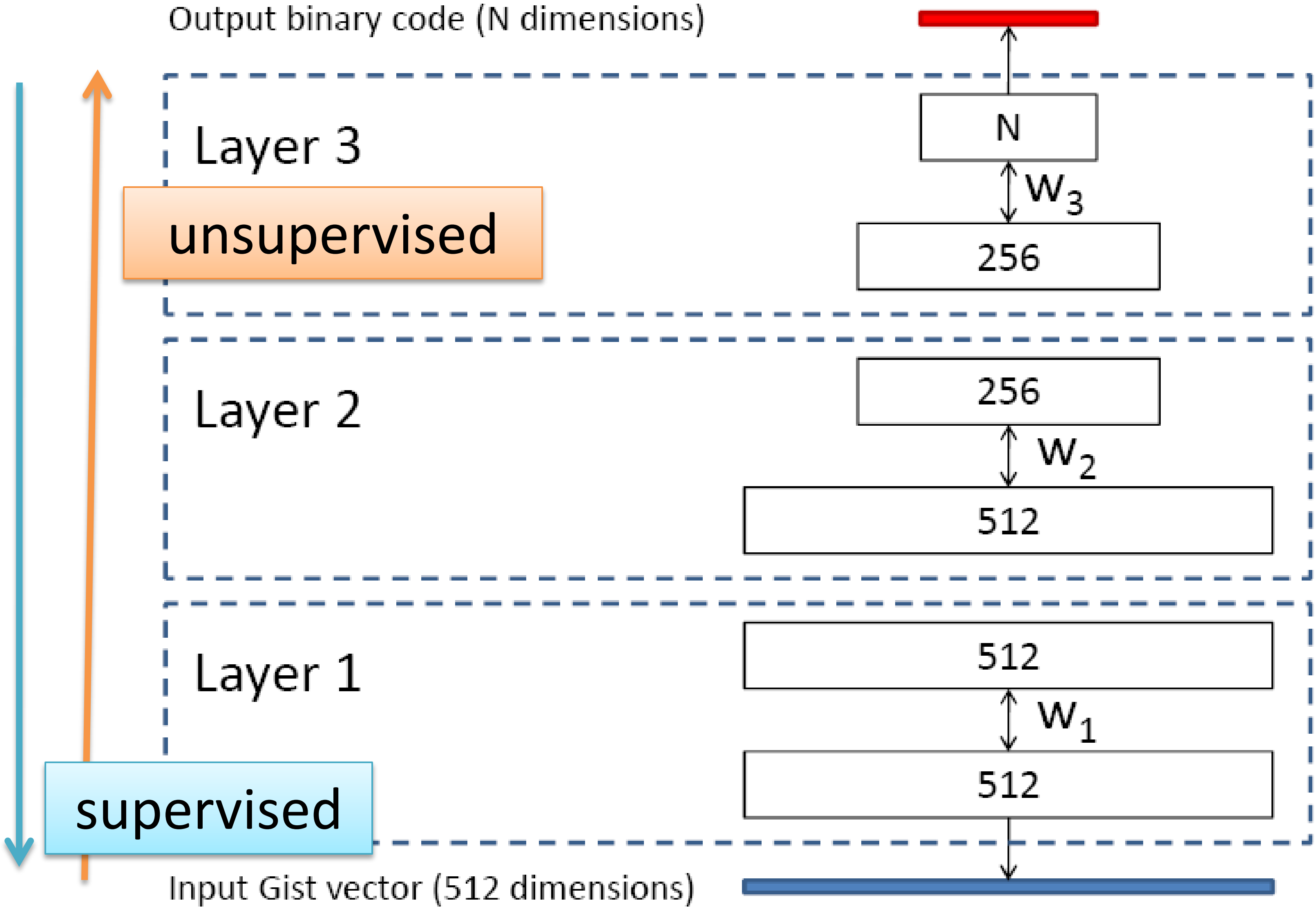
Hidden units: h

Symmetric weights w

Visible units: v



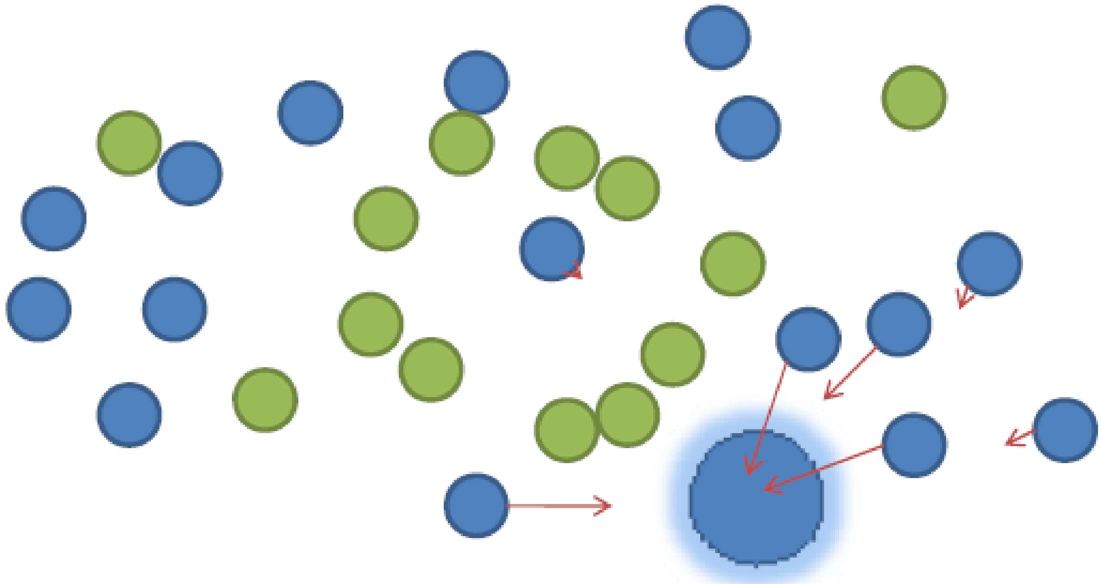
Multi-Layer RBM



Supervised Re-fining

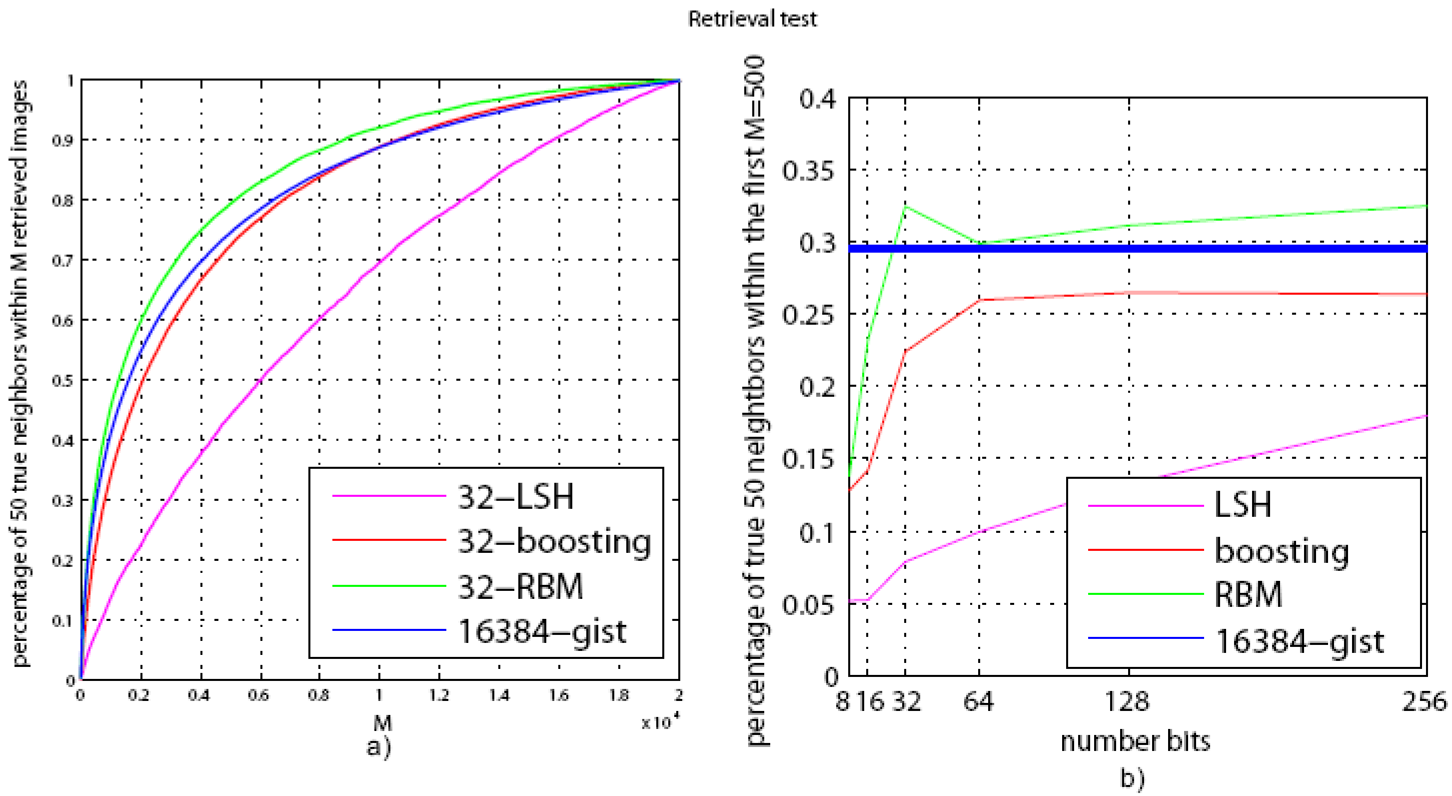
- Goldberger, Roweis,Salakhutdinov & Hinton, NIPS 2004

$$O_{\text{NCA}} = \sum_{k=1}^{K} \sum_{l:c^k=c^l} p_{kl} \qquad p_{kl} = \frac{e^{-||f(\mathbf{x}^k|W)-f(\mathbf{x}^l|W)||^2}}{\sum_{m\neq l} e^{-||f(\mathbf{x}^m|W)-f(\mathbf{x}^l|W)||^2}}$$



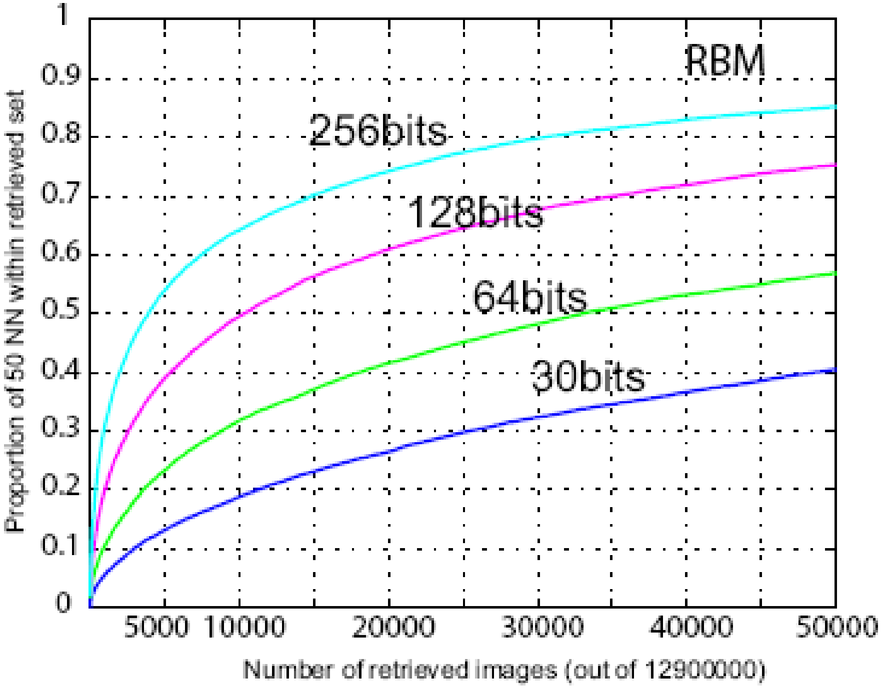
Experiments - LabelMe

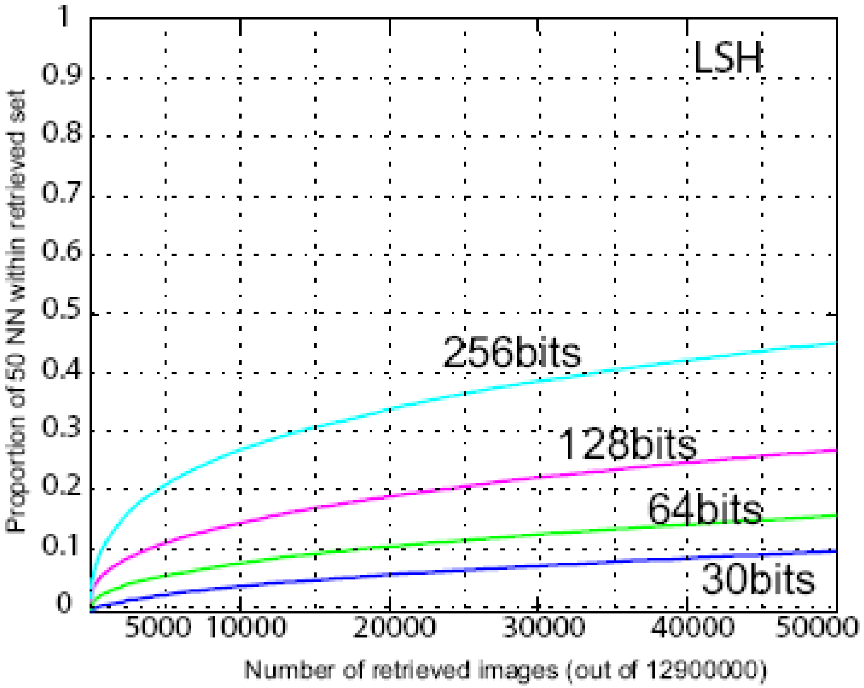
LabelMe (22,000 images)



Experiments - Web

Web (12.9 million)





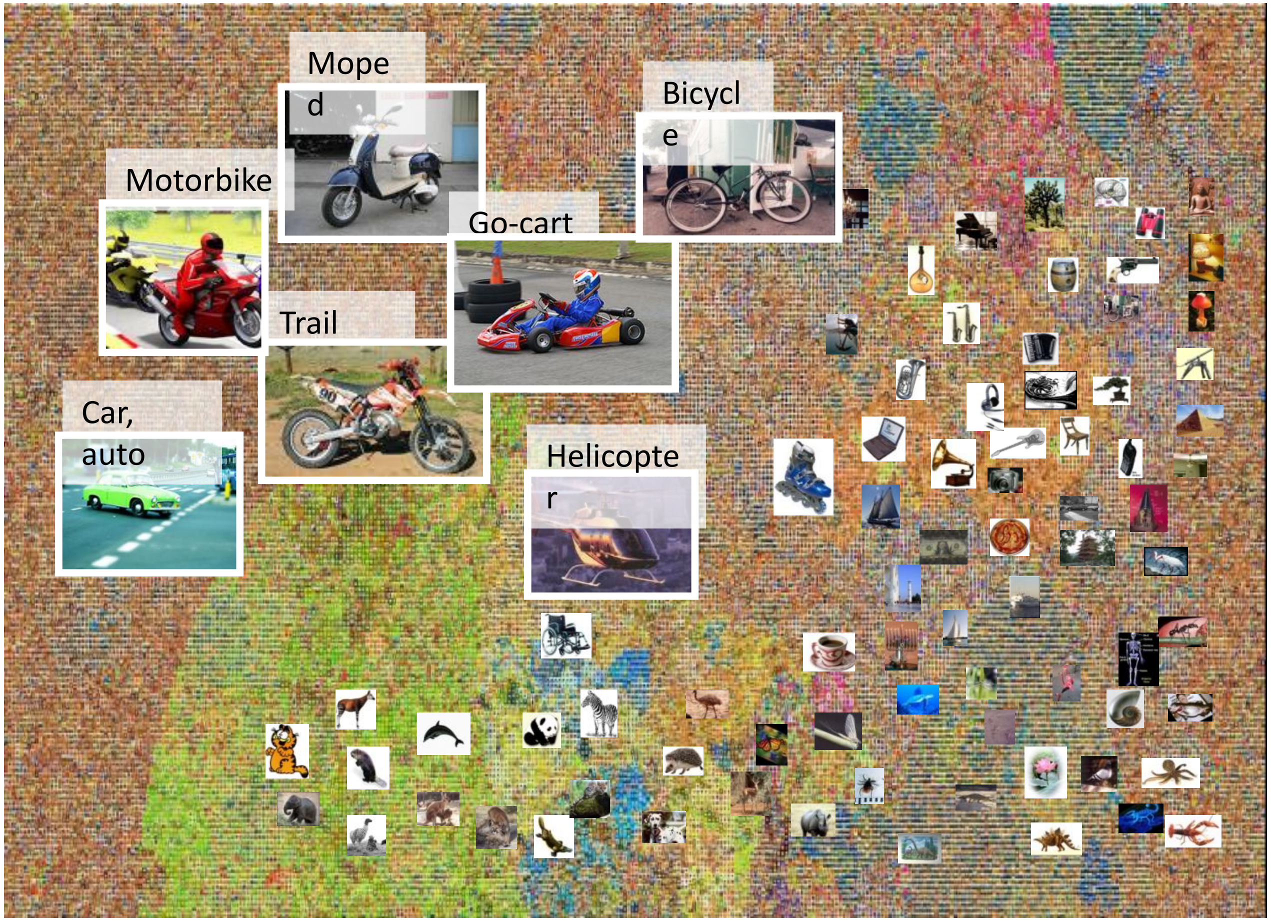
Experiments - Speed

Dataset	LabelMe	Web
# images	2×10^4	1.29×10^7
Gist vector dim.	512	384
Method	**Time (s)**	**Time (s)**
Spill tree - Gist vector	1.05	-
Brute force - Gist vector	0.38	-
Brute force - 30 bit binary	4.3×10^{-4}	0.146
” - 30 bit binary, M/T	2.7×10^{-4}	0.074
Brute force - 256 bit binary	1.4×10^{-3}	0.75
” - 256 bit binary, M/T	4.7×10^{-4}	0.23
Hashing - 30 bit binary	6×10^{-6}	6×10^{-6}

Large Scale Dataset

ImageNet: A Large-Scale Hierarchical Image Database. J. Deng et al., CVPR*, 2009.*
What does classifying more than 10,000 image categories tell us? J. Deng et al., *ECCV. 2010*

IMAGENET



Background image courtesy: Antonio Torralba

WordNet

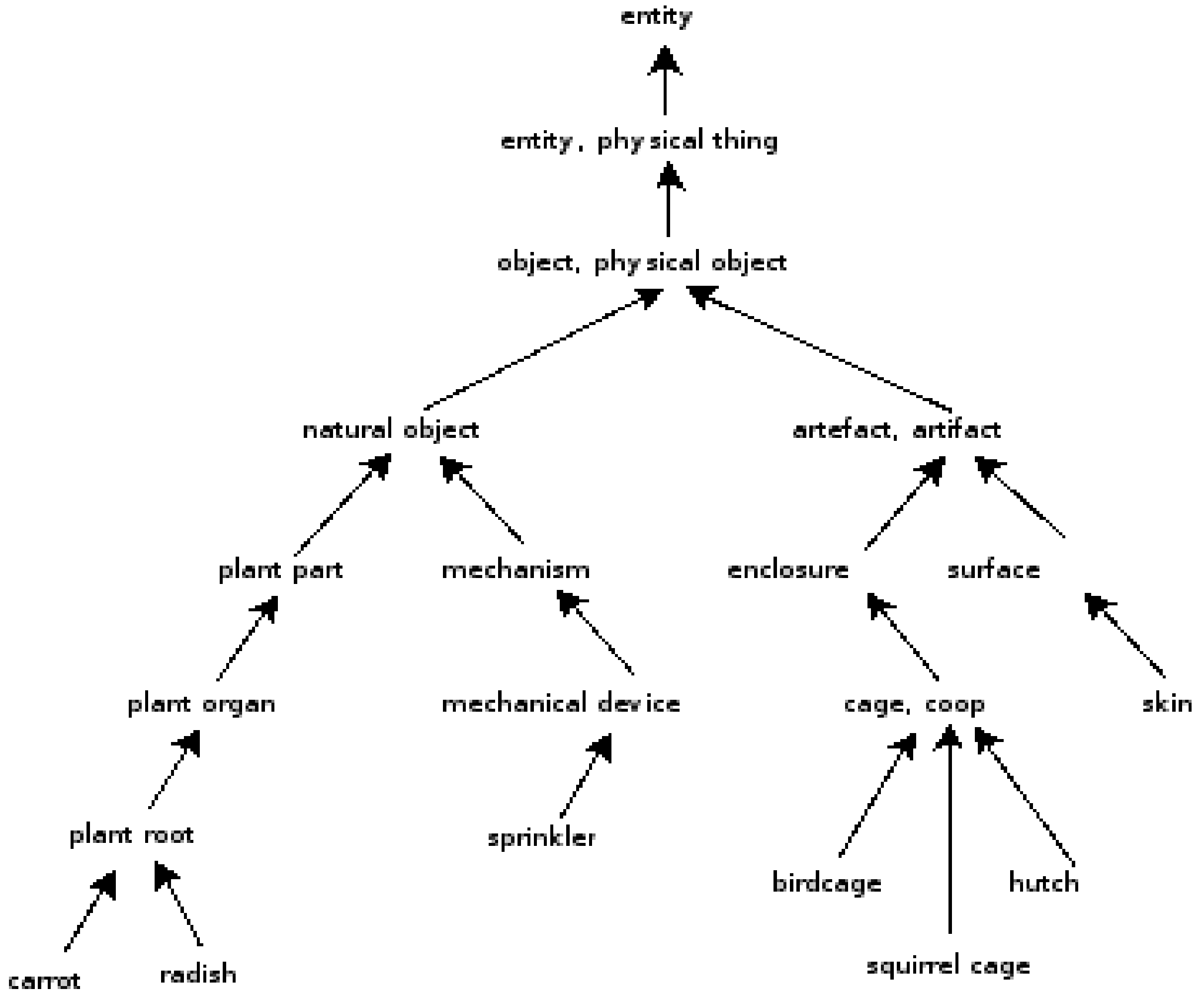
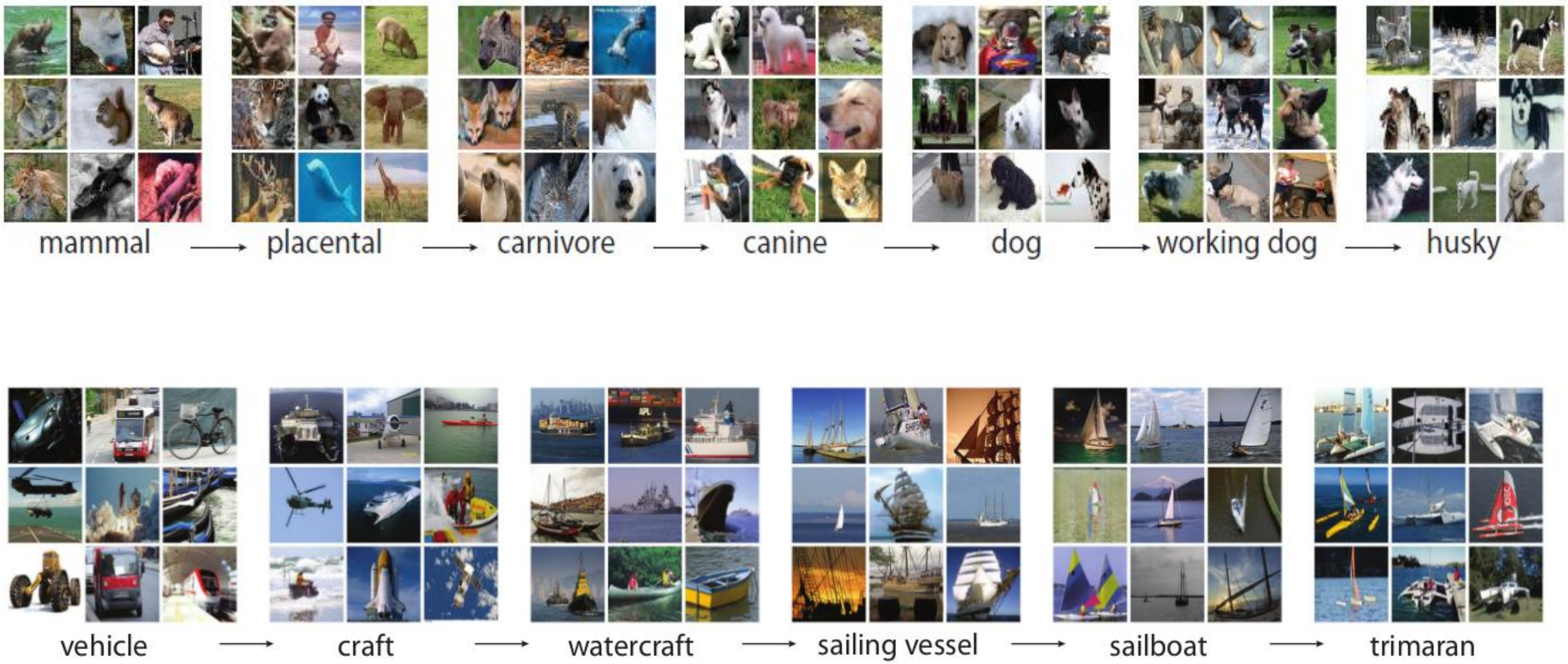


Figure 1. "is a" relation example

ImageNet


mammal
placental
carnivore
canine
dog
working dog
husky
vehicle
craft
watercraft
sailing vessel
sailboat
trimaran

Comparison to Other Data Sets

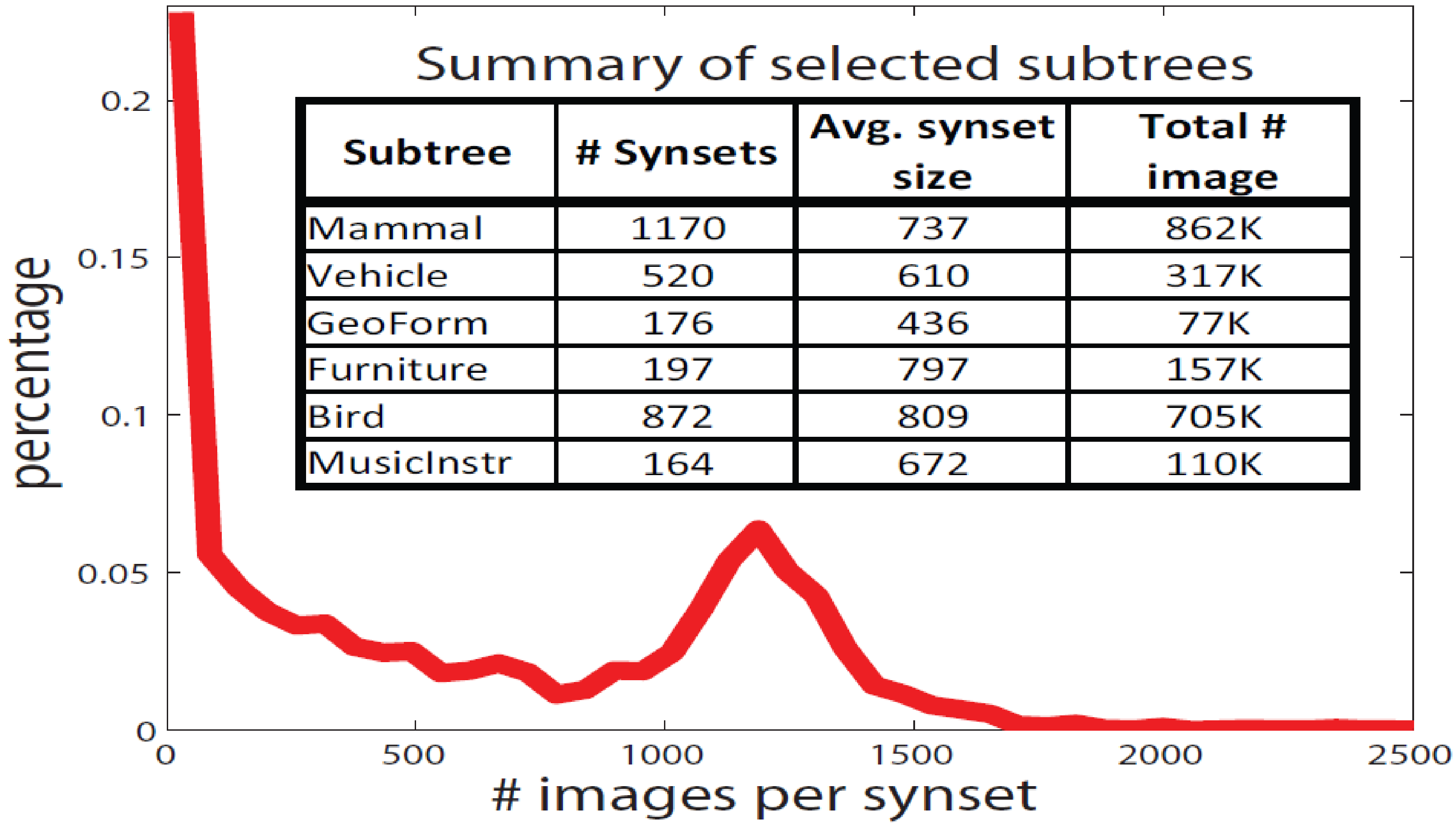
	ImageNet	TinyImage	LabelMe	ESP	LHill
LabelDisam	Y	Y	N	N	Y
Clean	Y	N	Y	Y	Y
DenseHie	Y	Y	N	N	N
FullRes	Y	N	Y	Y	Y
PublicAvail	Y	Y	Y	N	N
Segmented	N	N	Y	N	Y

Scale

- Synsets: 18K
- Images: 12M
- Images w/ bounding box annotations: 658K

High level category	# synset	Avg # images per synset	Total # images
animal	3822	732	2799K
plant	1666	600	999K
mammal	1138	821	934K
vehicle	481	778	374K

Scale



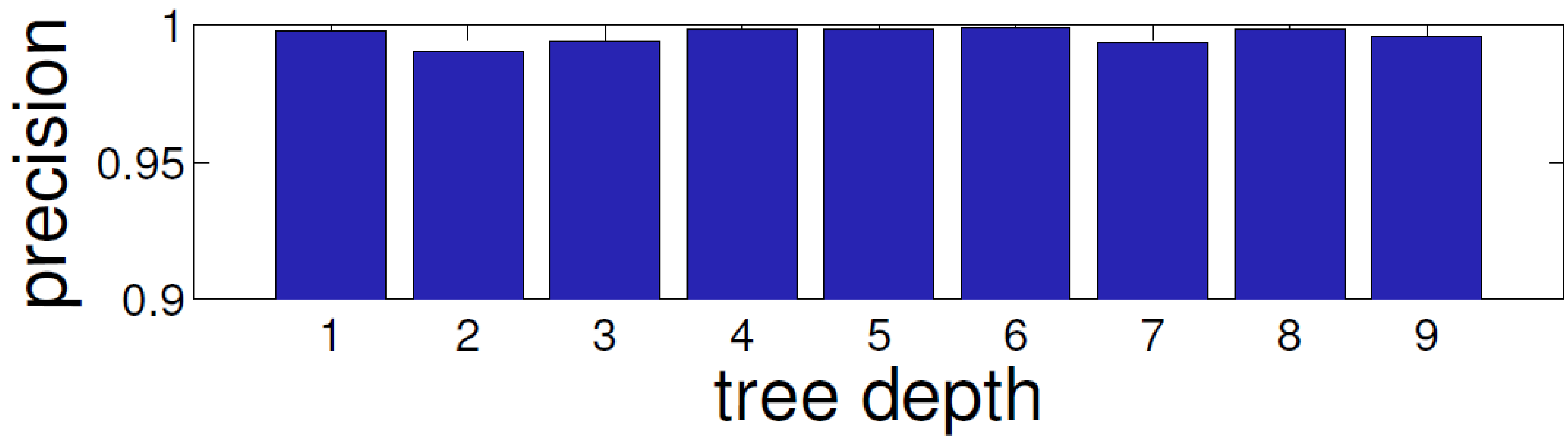
Summary of selected subtrees

Subtree	# Synsets	Avg. synset size	Total # image
Mammal	1170	737	862K
Vehicle	520	610	317K
GeoForm	176	436	77K
Furniture	197	797	157K
Bird	872	809	705K
MusicInstr	164	672	110K

- Over half synsets have > 500 images

Accuracy

- Lower in hierarchy, harder to classify.
  - Dog vs cat
  - Siamese cat vs Burmese cat

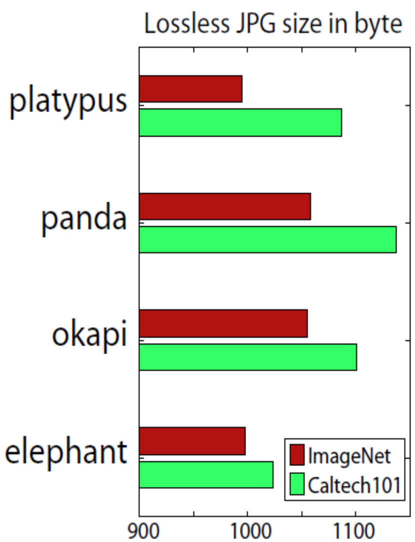


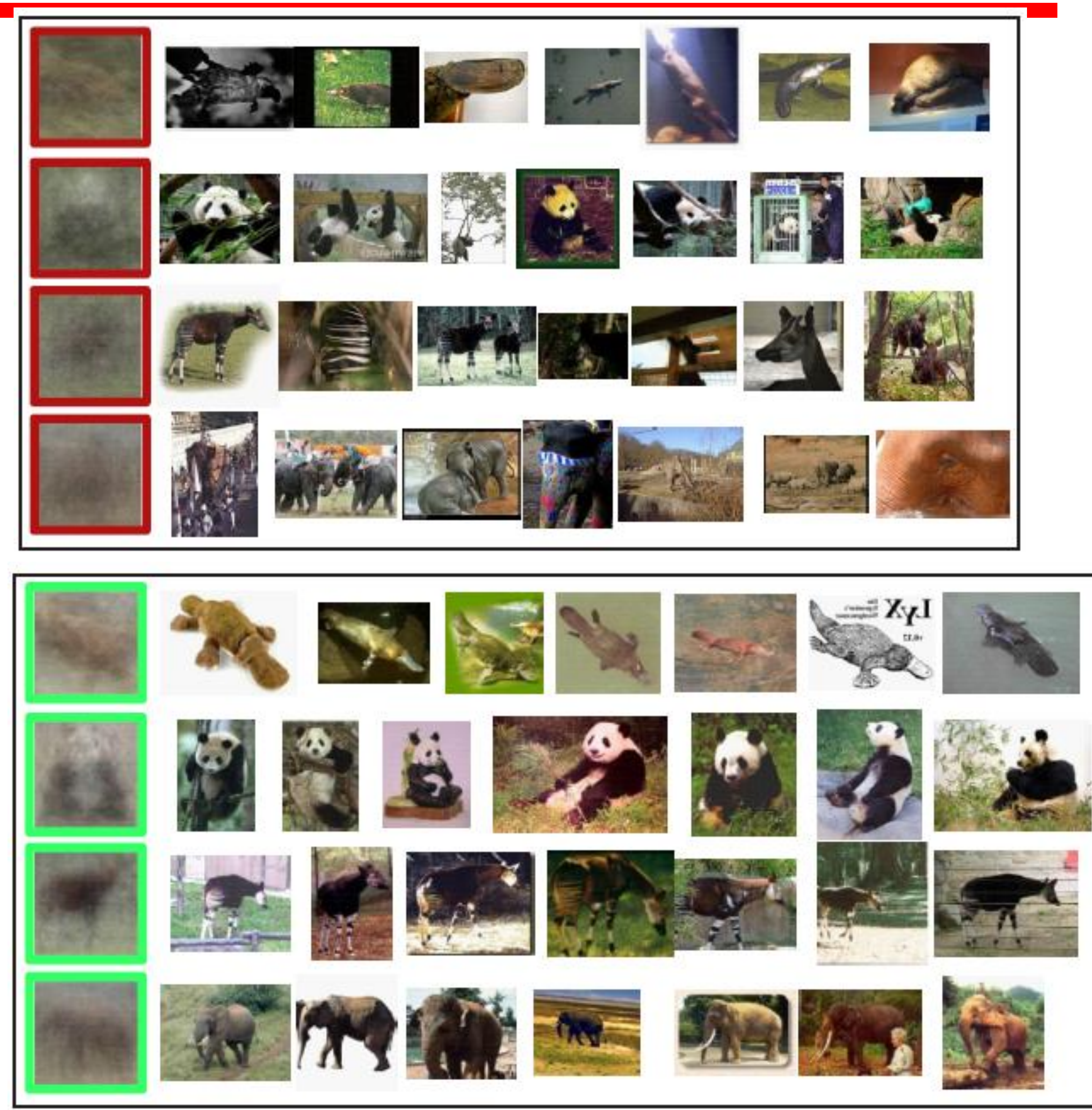
e.g. mammal e.g. dog e.g. German Shepherd

Deng, Dong, Socher, Li, Li, & Fei-Fei, *CVPR*, 2009

Diversity

- Variable appearances, positions, view points, poses, bg clutter, and occlusions.





Constructing ImageNet

Two Steps

Step 1:
Collect candidate images
via the Internet

→

Step 2:
Clean up the candidate
Images by humans

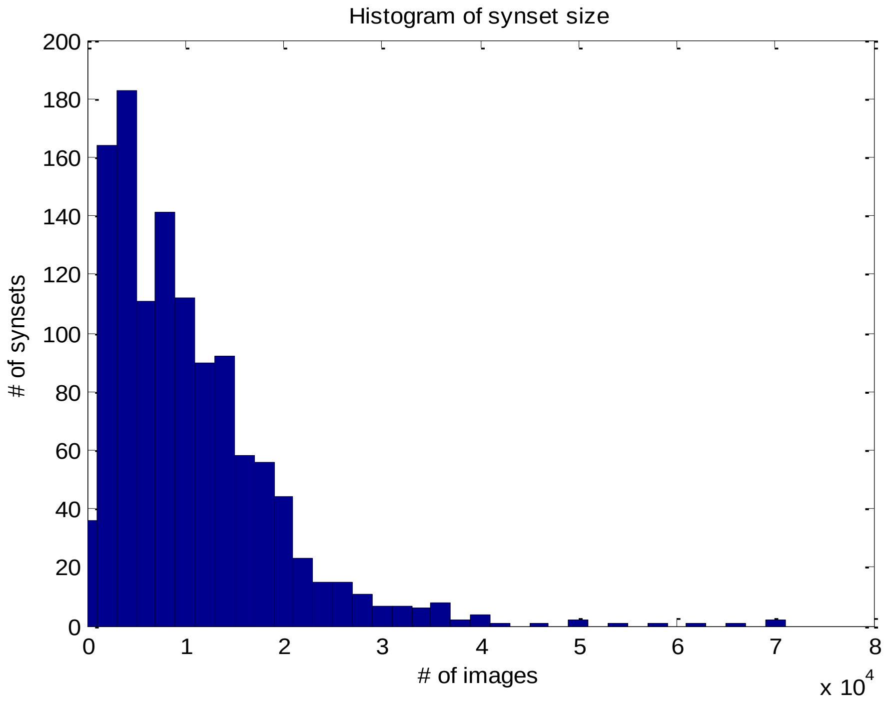
Step 1: Collect Candidate Images from the Internet

- Query expansion
  - Synonyms: *German shepherd, German police dog, German shepherd dog, Alsatian*
  - Appending words from ancestors: *sheepdog, dog*
- Multiple languages
  - Italian, Dutch, Spanish, Chinese

    e.g. ovejero alemán, pastore tedesco,德国牧羊犬
- More engines
- Parallel downloading

Step 1: Collect Candidate Images from the Internet

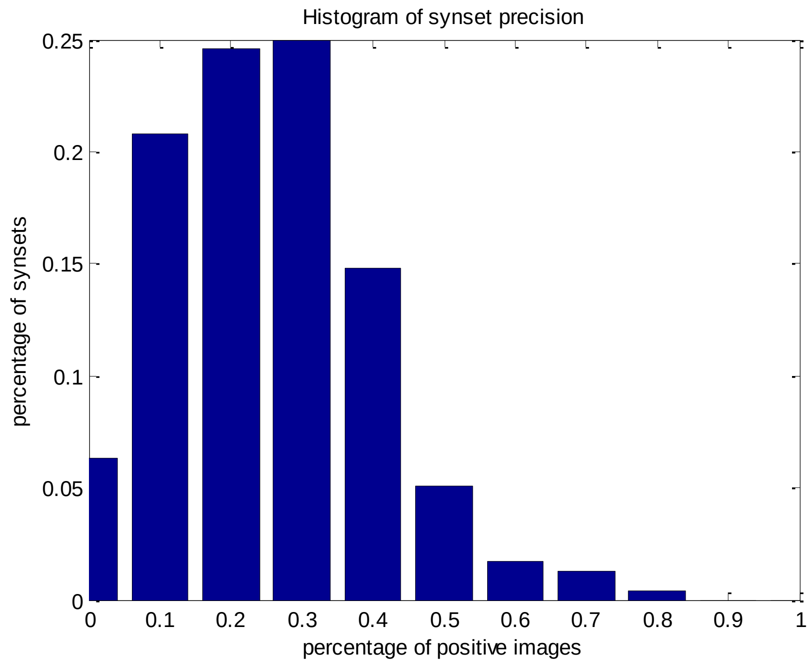
- "Mammal" subtree (1180 synsets)
  – Average # of images per synset: 10.5K



Most populated	Least populated
Humankind (118.5k)	Algeripithecus minutus (90)
Kitty, kitty-cat (69k)	Striped muishond (107)
Cattle, cows (65k)	Mylodonitid (127)
Pooch, doggie (62k)	Greater pichiciego (128)
Cougar, puma (57k)	Damaraland mole rat (188)
Frog, toad (53k)	Western pipistrel (196)
Hack, jade, nag (50k)	Muishond (215)

Step 1: Collect Candidate Images from the Internet

- "Mammal" subtree (1180 synsets)
  - Average accuracy per synset: 26%



Most accurate	Least accurate
Bottlenose dolpin (80%)	Fanaloka (1%)
Meerkat (74%)	Pallid bat (3%)
Burmese cat (74%)	Vaquita (3%)
Humpback whale (69%)	Fisher cat (3%)
African elephant (63%)	Walrus (4%)
Squirrel (60%)	Grison (4%)
Domestic cat (59%)	Pika, Mouse hare (4%)

Step 2: verifying the images by humans

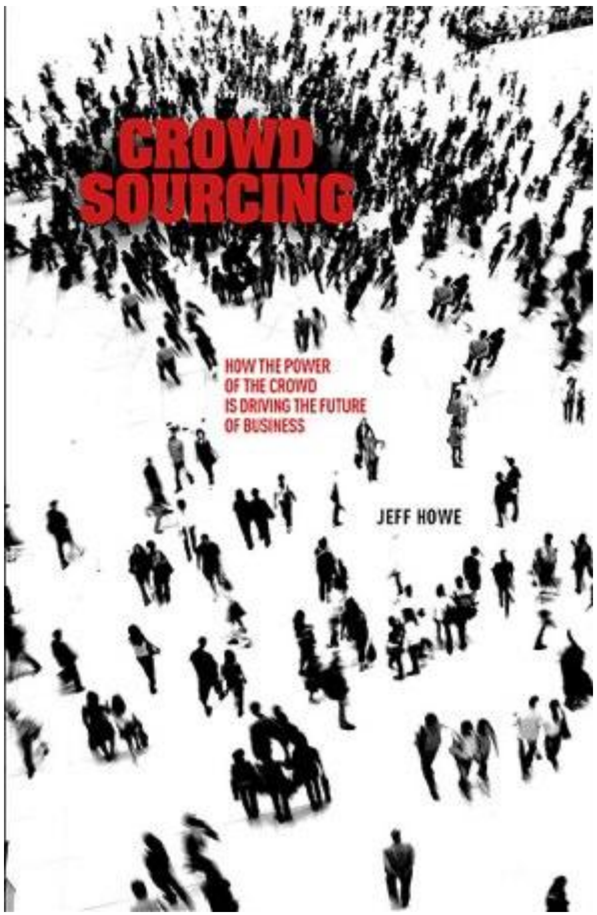
- # of synsets: 40,000 (subject to: imageability analysis)
- # of candidate images to label per synset: 10,000
- # of people needed to verify: 2-5
- Speed of human labeling: 2 images/sec (one fixation: ~200msec)

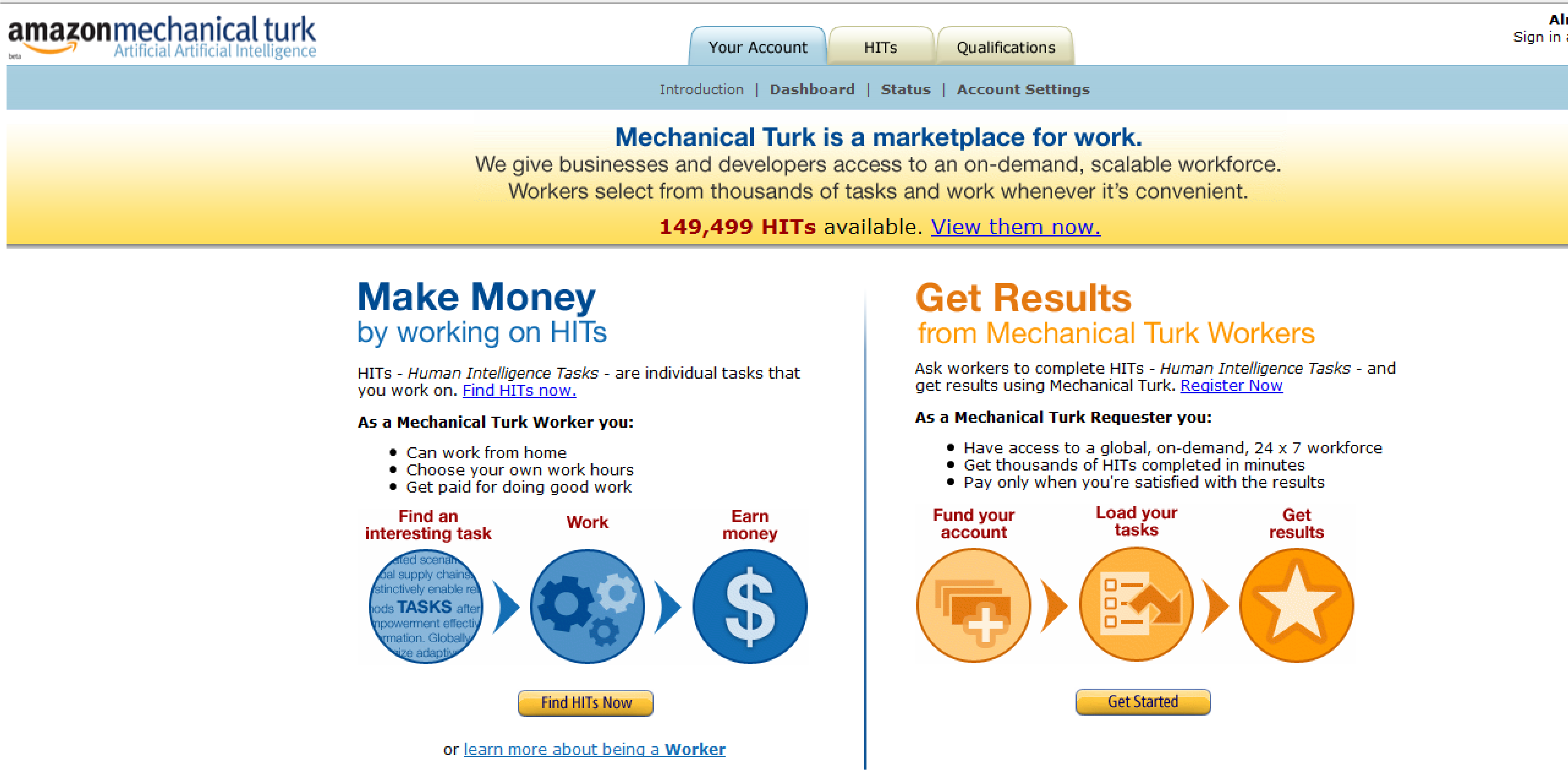
$$40{,}000 \times 10{,}000 \times 3/2 = 600{,}000{,}000 \text{ sec} \approx 19 \text{ year}$$

Moral of the story:

no graduate students would want to do this project!

In summer 2008, we discovered crowdsourcing




amazon mechanical turk
Artificial Artificial Intelligence
beta
Sign in
Your Account
HITs
Qualifications
Introduction | Dashboard | Status | Account Settings
Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.
149,499 HITs available. View them now.
Make Money
by working on HITs
HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.
As a Mechanical Turk Worker you:
Can work from home
Choose your own work hours
Get paid for doing good work
Find an interesting task
Work
Earn money
TASKS
Find HITs Now
or learn more about being a Worker
Get Results
from Mechanical Turk Workers
Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now
As a Mechanical Turk Requester you:
Have access to a global, on-demand, 24 x 7 workforce
Get thousands of HITs completed in minutes
Pay only when you're satisfied with the results
Fund your account
Load your tasks
Get results
Get Started

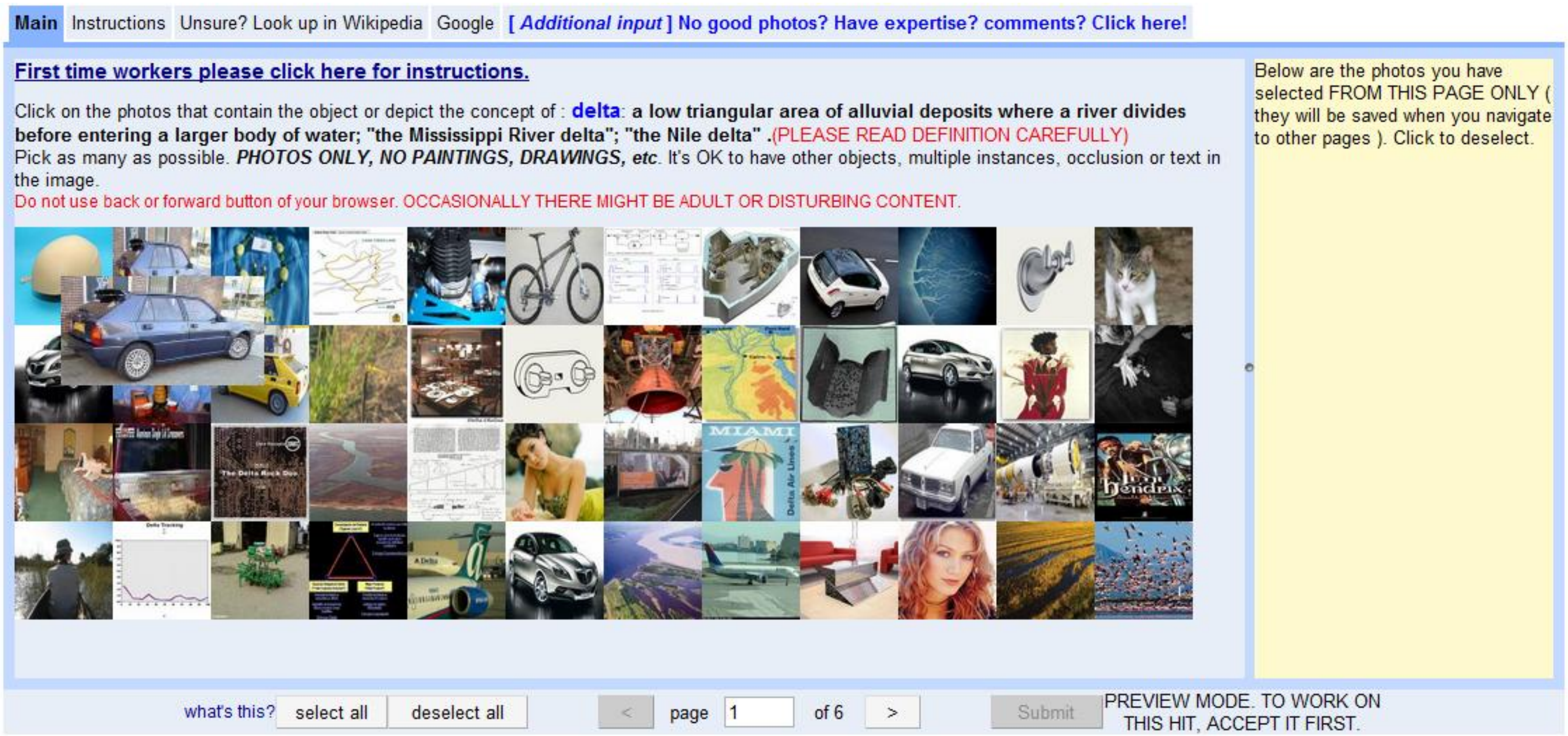
Step 2: verifying the images by humans

- # of synsets: 40,000 (subject to: imageability analysis)
- # of candidate images to label per synset: 10,000
- # of people needed to verify: 2-5
- Speed of human labeling: 2 images/sec (one fixation: ~200msec)
- Massive parallelism (N ~ 10^2-3)

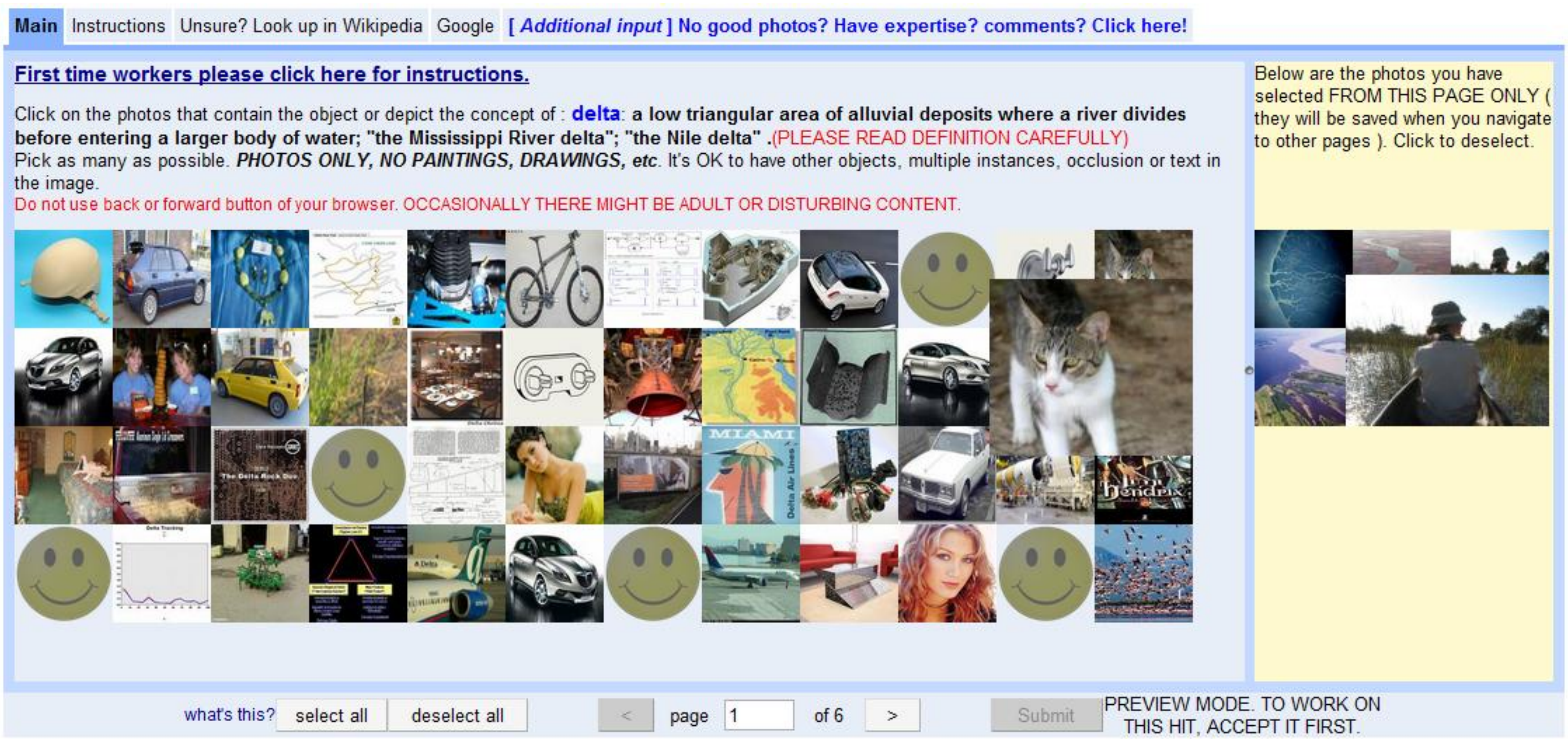
$$40{,}000 \times 10{,}000 \times 3 / 2 = 600{,}000{,}000 \text{ sec} \approx \frac{19 \text{ year}}{N}$$

IMAGENET Basic User Interface

Click on the good images.



IMAGENET Basic User Interface



Enhancement 1

- Provide wiki and google links

Enhancement 2

- Make sure workers read the definition.
  - Words are ambiguous. E.g.
    - Box: *any one of several designated areas on a ball field where the batter or catcher or coaches are positioned*
    - Keyboard: *holder consisting of an arrangement of hooks on which keys or locks can be hung*
  - These synsets are hard to get right
  - Some workers do not read or understand the definition.

Definition quiz

This HIT is about **'delta'**.

Definition: a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta"

Please read the above definition carefully. 'delta' might mean something different from what you think.

I HAVE READ IT

Definition quiz

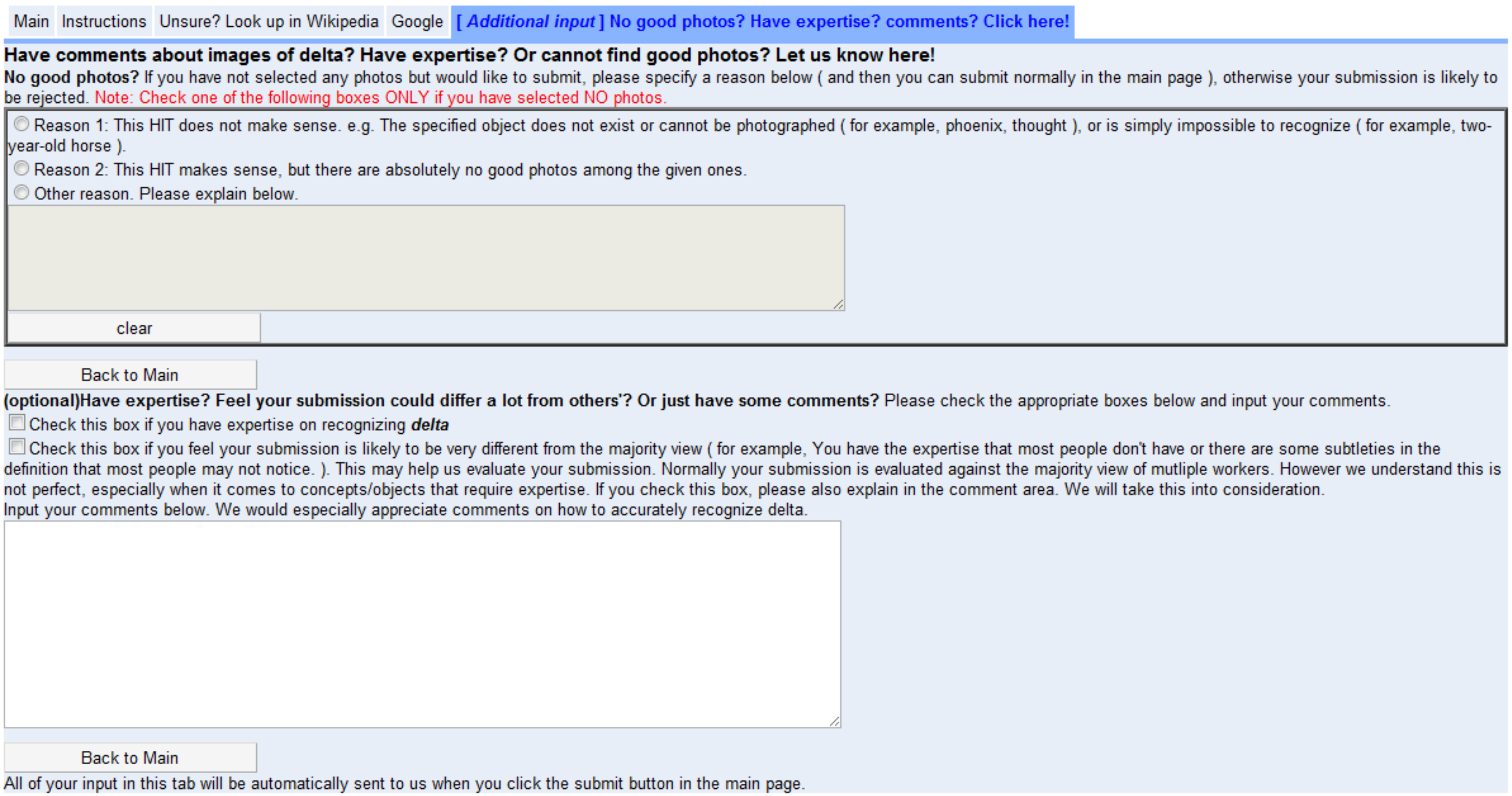
Please answer: what is the meaning of 'delta' in this HIT?

Go back and read the definition again.

- the normal brainwave in the encephalogram of a person in deep dreamless sleep; occurs with high voltage and low frequency (1 to 4 hertz)
- the 4th letter of the Greek alphabet
- a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta"
- an airplane with wings that give it the appearance of an isosceles triangle
- an object shaped like an equilateral triangle

Enhancement 3

- Allow more feedback. E.g. “unimagable synsets” expert opinion



Main | Instructions | Unsure? Look up in Wikipedia | Google | *[Additional input]* No good photos? Have expertise? comments? Click here!

Have comments about images of delta? Have expertise? Or cannot find good photos? Let us know here!

No good photos? If you have not selected any photos but would like to submit, please specify a reason below (and then you can submit normally in the main page), otherwise your submission is likely to be rejected. Note: Check one of the following boxes ONLY if you have selected NO photos.

- Reason 1: This HIT does not make sense. e.g. The specified object does not exist or cannot be photographed (for example, phoenix, thought), or is simply impossible to recognize (for example, two-year-old horse).
- Reason 2: This HIT makes sense, but there are absolutely no good photos among the given ones.
- Other reason. Please explain below.

clear

Back to Main

(optional)Have expertise? Feel your submission could differ a lot from others'? Or just have some comments? Please check the appropriate boxes below and input your comments.

- Check this box if you have expertise on recognizing ***delta***
- Check this box if you feel your submission is likely to be very different from the majority view (for example, You have the expertise that most people don't have or there are some subtleties in the definition that most people may not notice.). This may help us evaluate your submission. Normally your submission is evaluated against the majority view of mutliple workers. However we understand this is not perfect, especially when it comes to concepts/objects that require expertise. If you check this box, please also explain in the comment area. We will take this into consideration.

Input your comments below. We would especially appreciate comments on how to accurately recognize delta.

Back to Main

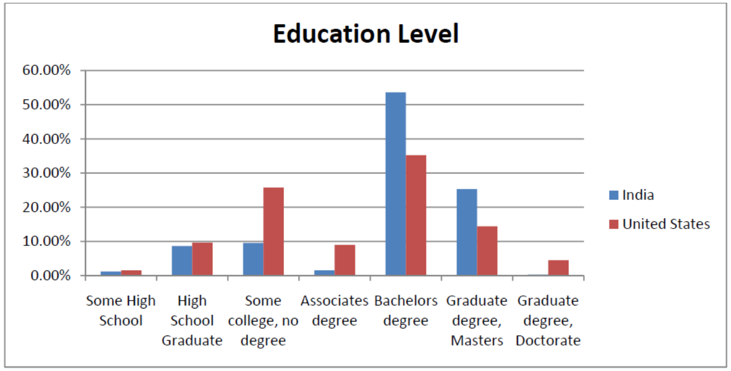
All of your input in this tab will be automatically sent to us when you click the submit button in the main page.

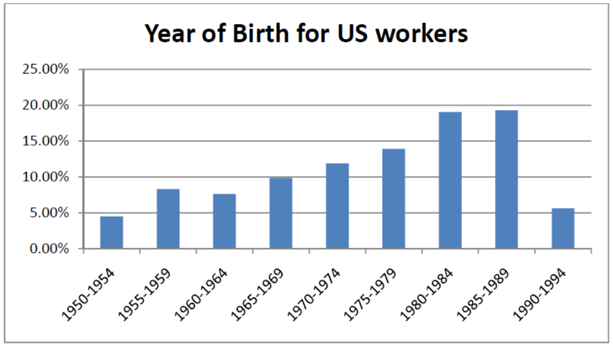
IMAGENET is built by crowdsourcing

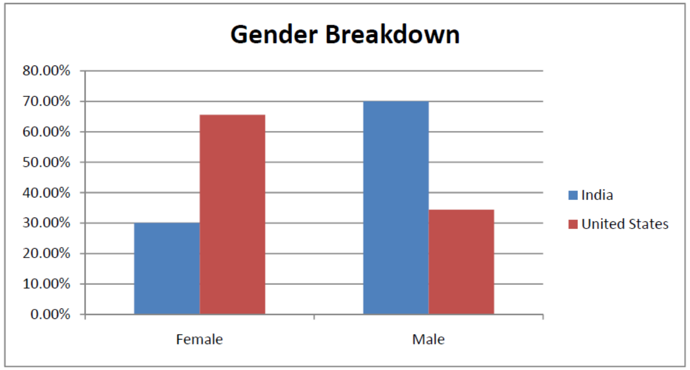
- July 2008: 0 images
- Dec 2008: 3 million images, 6000+ synsets
- April 2010: 11 million images, 15,000+ synsets

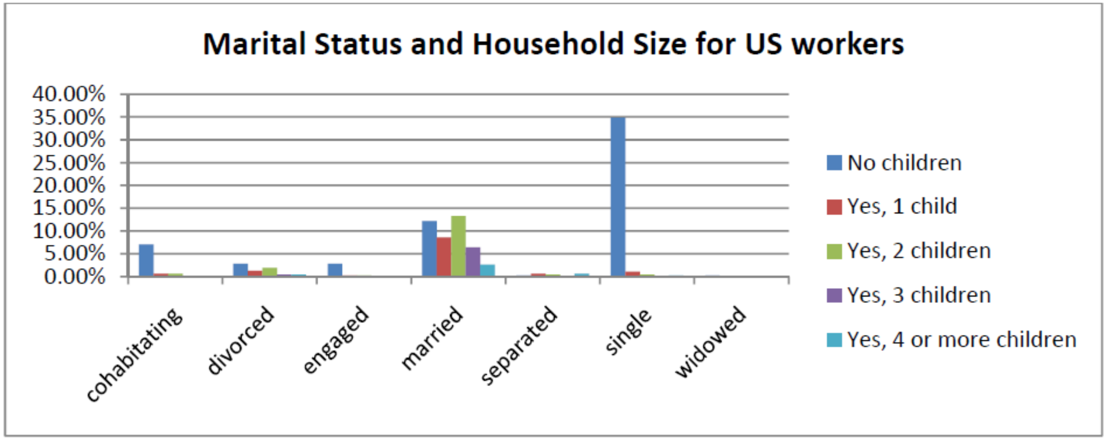
Demography of AMT workers

United States	46.80%
India	**34.00%**
Miscellaneous	19.20%



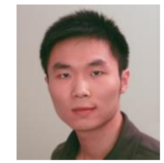




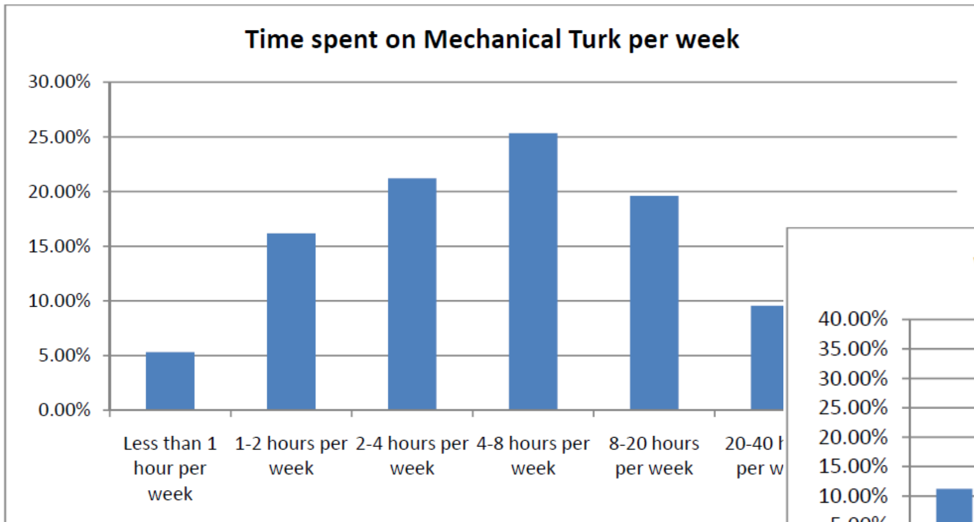


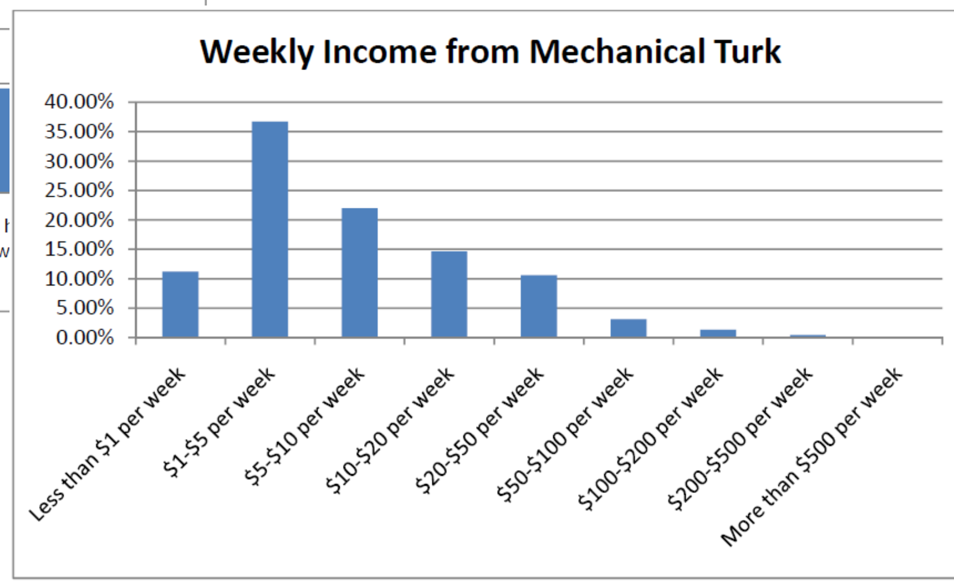
Panos Ipeirotis, NYU, Feb, 2010

Demography of AMT workers



Typical Stanford
Graduate student's income





Panos Ipeirotis, NYU, Feb, 2010

Use Large Dataset Smartly

80 million tiny images: a large dataset for non-parametric object and scene recognition. Torralba, Fergus, Freeman. PAMI 2008.

Nonparametric scene parsing: Label transfer via dense scene alignment , C. Liu, J. Yuen and A. Torralba. *CVPR, 2009*.

What does classifying more than 10,000 image categories tell us?

Background image courtesy: Antonio Torralba

Basic evaluation setup

- IMAGENET
  - 10,000 categories
  - 9 million images
  - 50%-50% train test split
- Multi-class classification in 1-vs-all framework
  - GIST+NN: filter banks; nearest neighbor (Oliva & Torralba, 2001)
  - BOW+NN: SIFT, 1000 codewords, BOW; nearest neighbor
  - BOW+SVM: SIFT, 1000 codewords, BOW; linear SVM
  - SPM+SVM: SIFT, 1000 codewords, Spatial Pyramid; intersection kernel SVM (Lazebnik et al. 2006)

Deng, Berg, Li, & Fei-Fei, *ECCV2010*

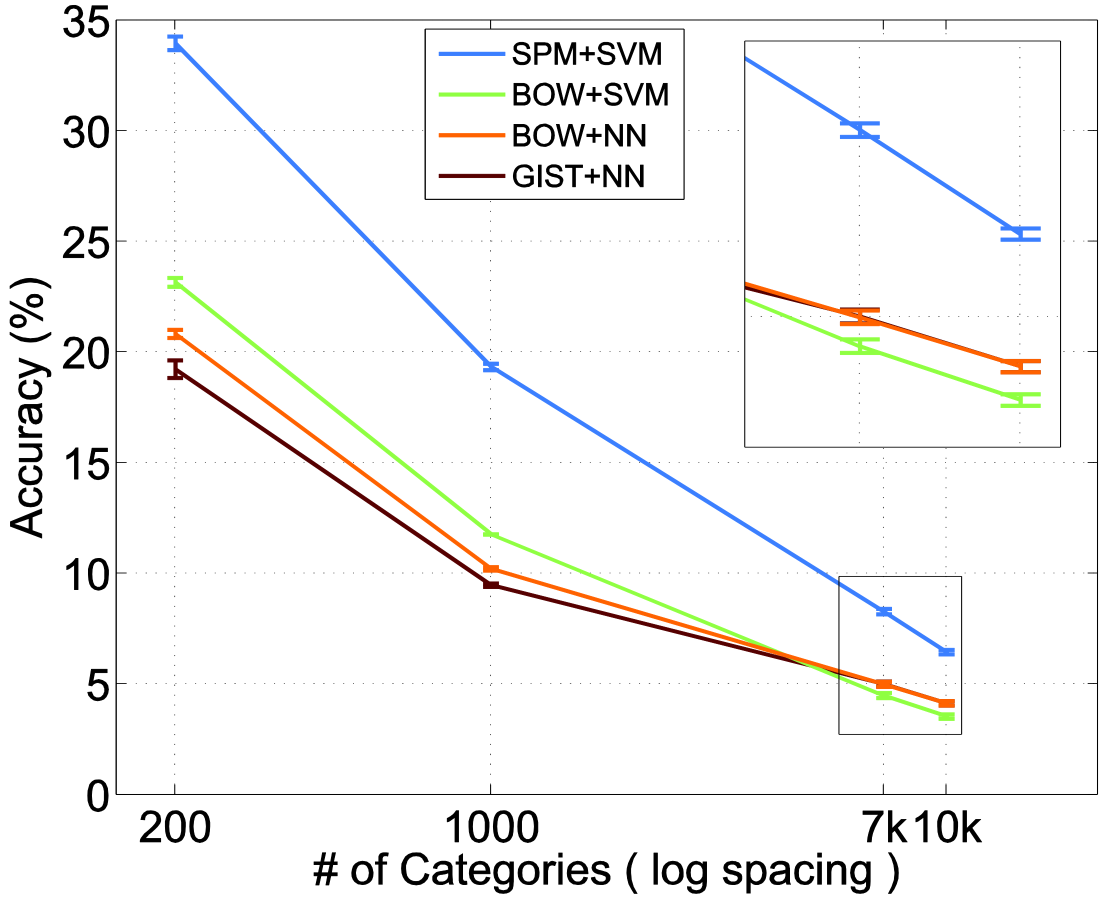
Computation issues first

- BOW+SVM
  - Train one 1-vs-all with LIBLINEAR → 1 CPU hour
  - 10,000 categories → 1 CPU year
- SPM + SVM
  - Maji & Berg 2009, LIBLINEAR with piece-wise linear encoding
  - Memory bottleneck. Modification required.
  - 10,000 categories → 6 CPU year
- Parallelized on a cluster
  - Weeks for a single run of experiments

Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Size matters

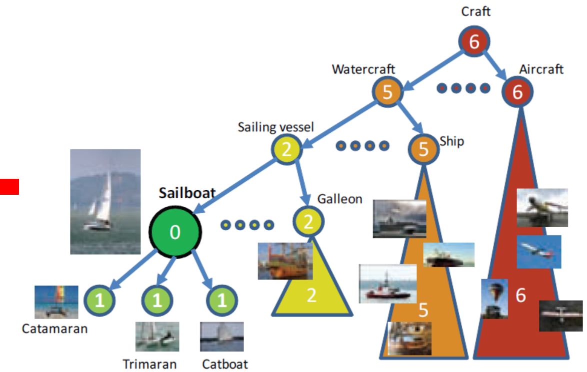
- 6.4% for 10K categories
- Better than we expected (instead of dropping at the rate of 10x; it's roughly at about 2x)
- An ordering switch between SVM and NN methods when the # of categories becomes large

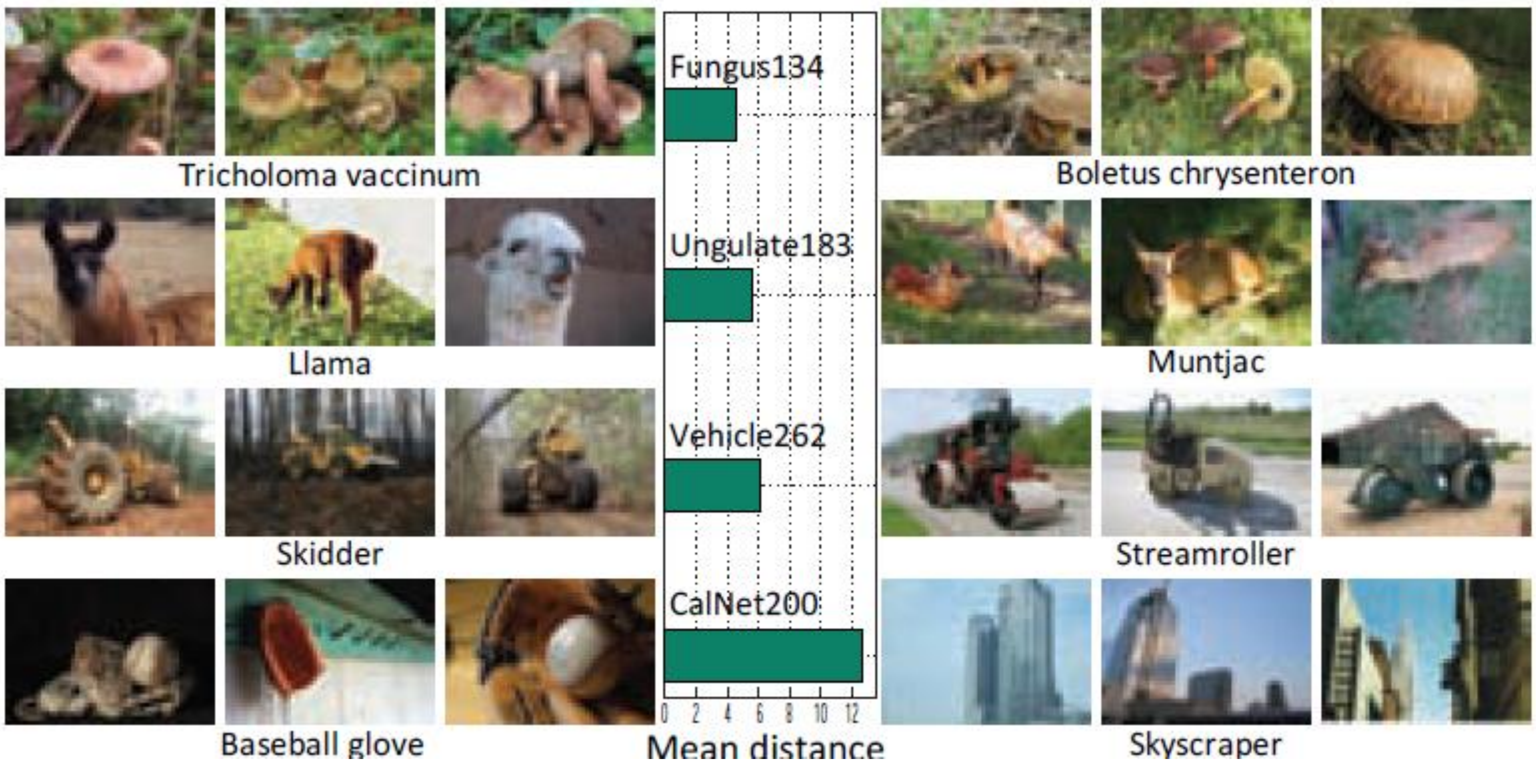


Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Density matters

- Datasets have very different "density" or "sparsity"

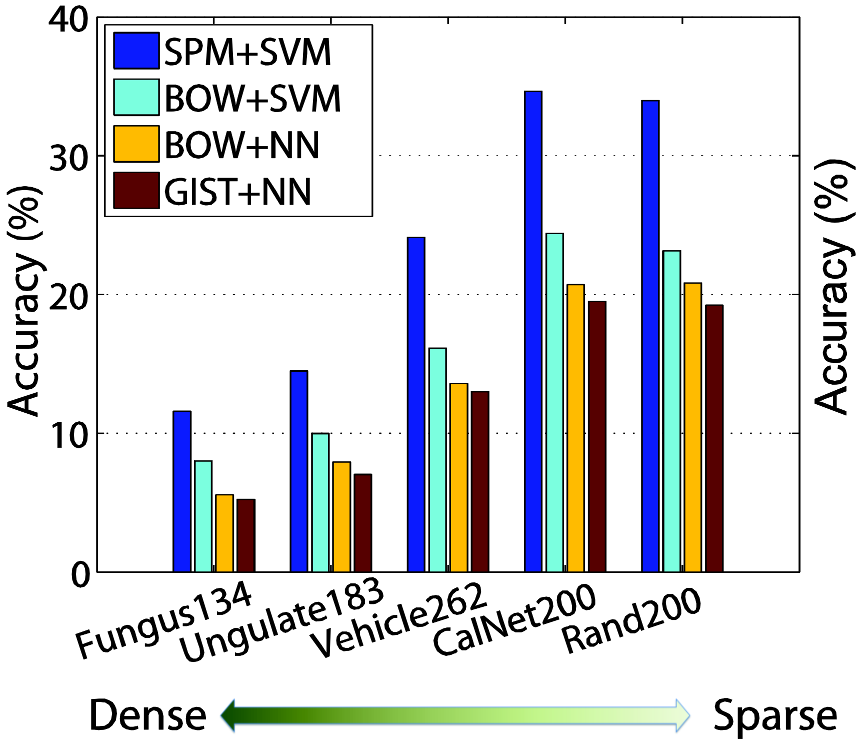


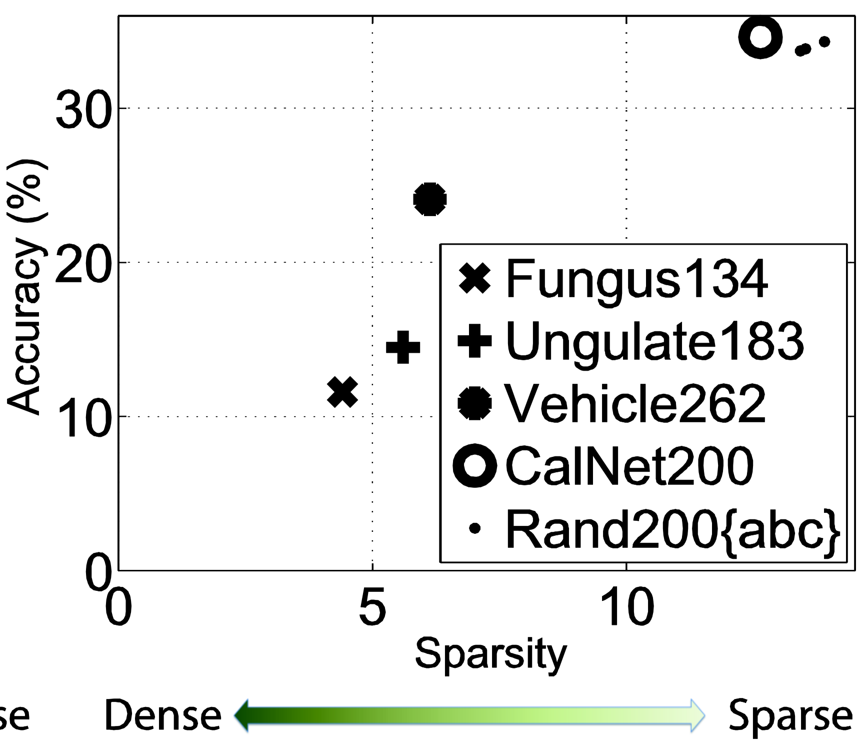


Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Density matters

- Datasets have very different “density” or “sparsity”
- there is a significant difference in difficulty between different datasets, independent of feature and classifier choice.



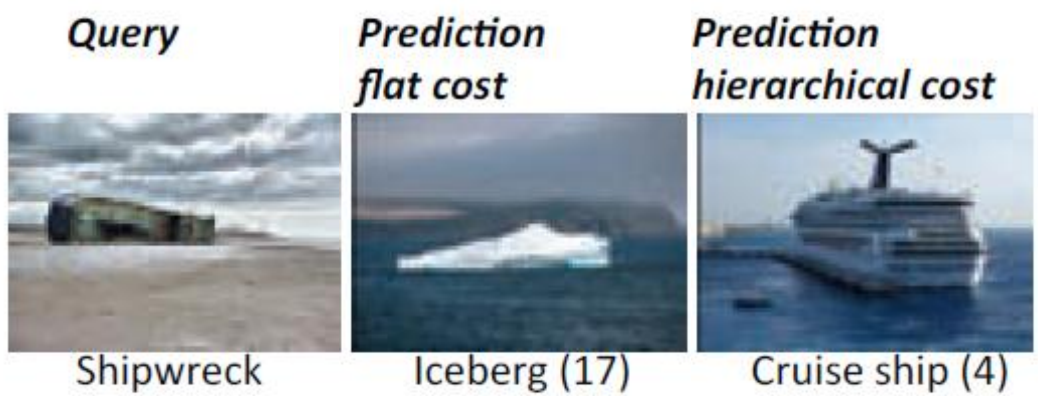


Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Hierarchy matters

- Classifying a "dog" as "cat" is probably not as bad as classifying it as "microwave"
- A simple way to incorporate classification cost

$$C_{ij} = \begin{cases} 0 & i=j \text{, or } i \text{ is a descendent of } j \\ h(i,j) & h \text{ is the height of the lowest common ancestor in WordNet} \end{cases}$$



Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Hierarchy matters

- Classifying a “dog” as “cat” is probably not as bad as classifying it as “microwave”
- A simple way to incorporate hierarchical classification cost

$$C_{ij} = \begin{cases} 0 & i=j, \text{ or } i \text{ is a descendent of } j \\ h(i,j) & h \text{ is the height of the lowest common ancestor in WordNet} \end{cases}$$

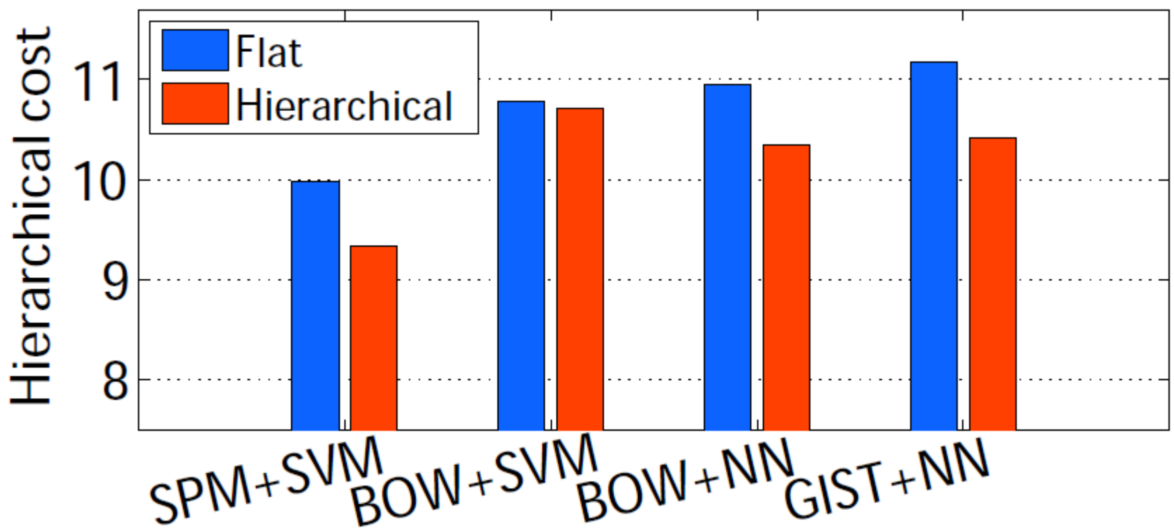
Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Hierarchy matters

- Classifying a "dog" as "cat" is probably not as bad as classifying it as "microwave"
- A simple way to incorporate hierarchical classification cost

i=*j*, or *i* is a descendent of *j*

h is the height of the lowest common ancestor in WordNet

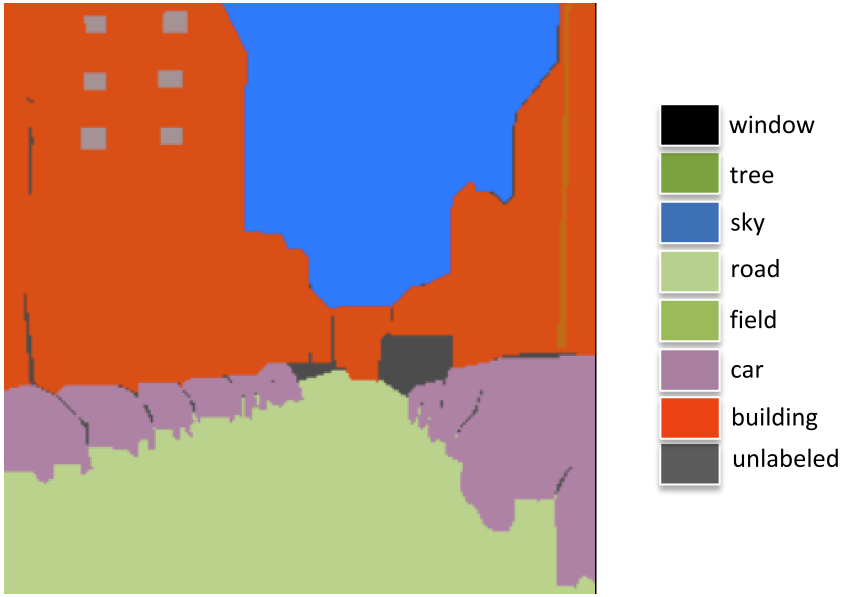


$$\text{cost} \quad f(x) = \arg\min_{i=1,\ldots,K} \sum_{j=1}^{K} C_{i,j}\hat{p}_j(x)$$

Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Image Labeling

Input

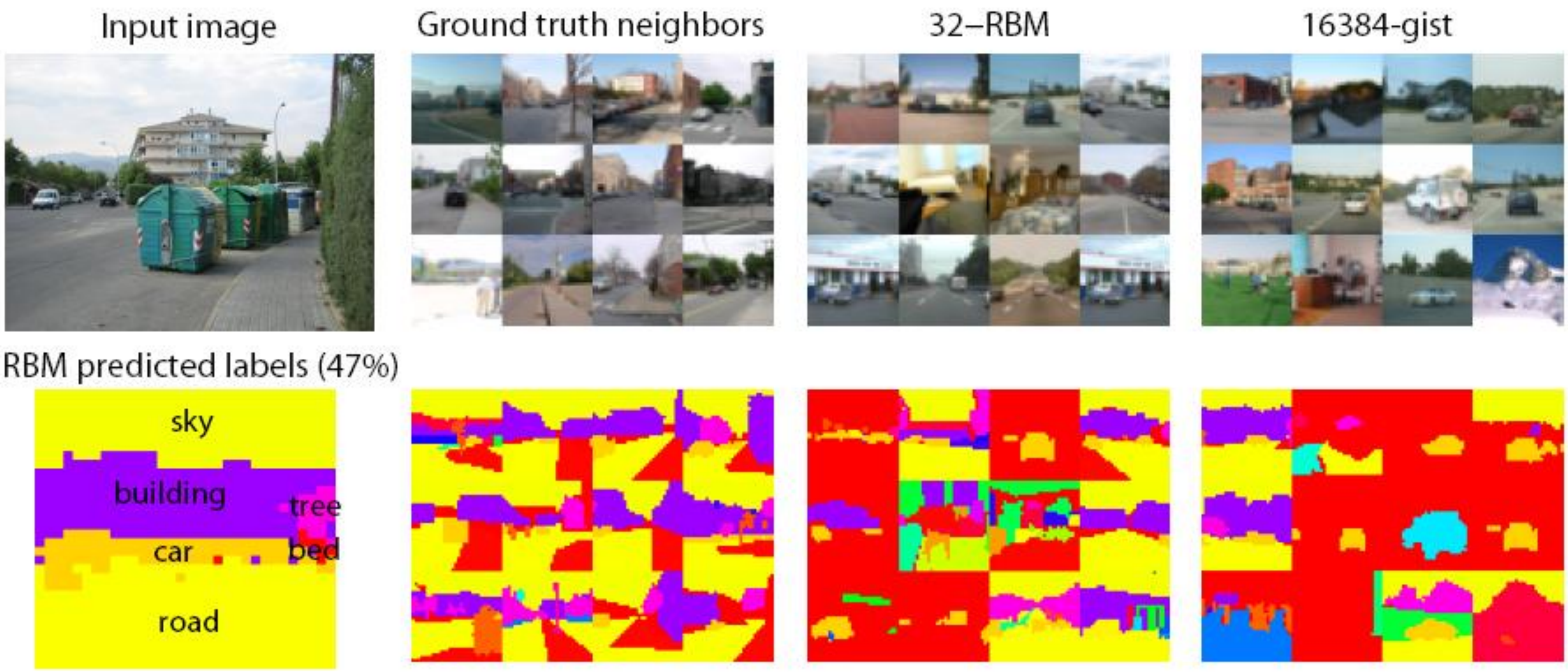


Output

Traditional Method: learn the local appearance for each category, smooth with a MRF/CRF model

Naïve Label Transfer

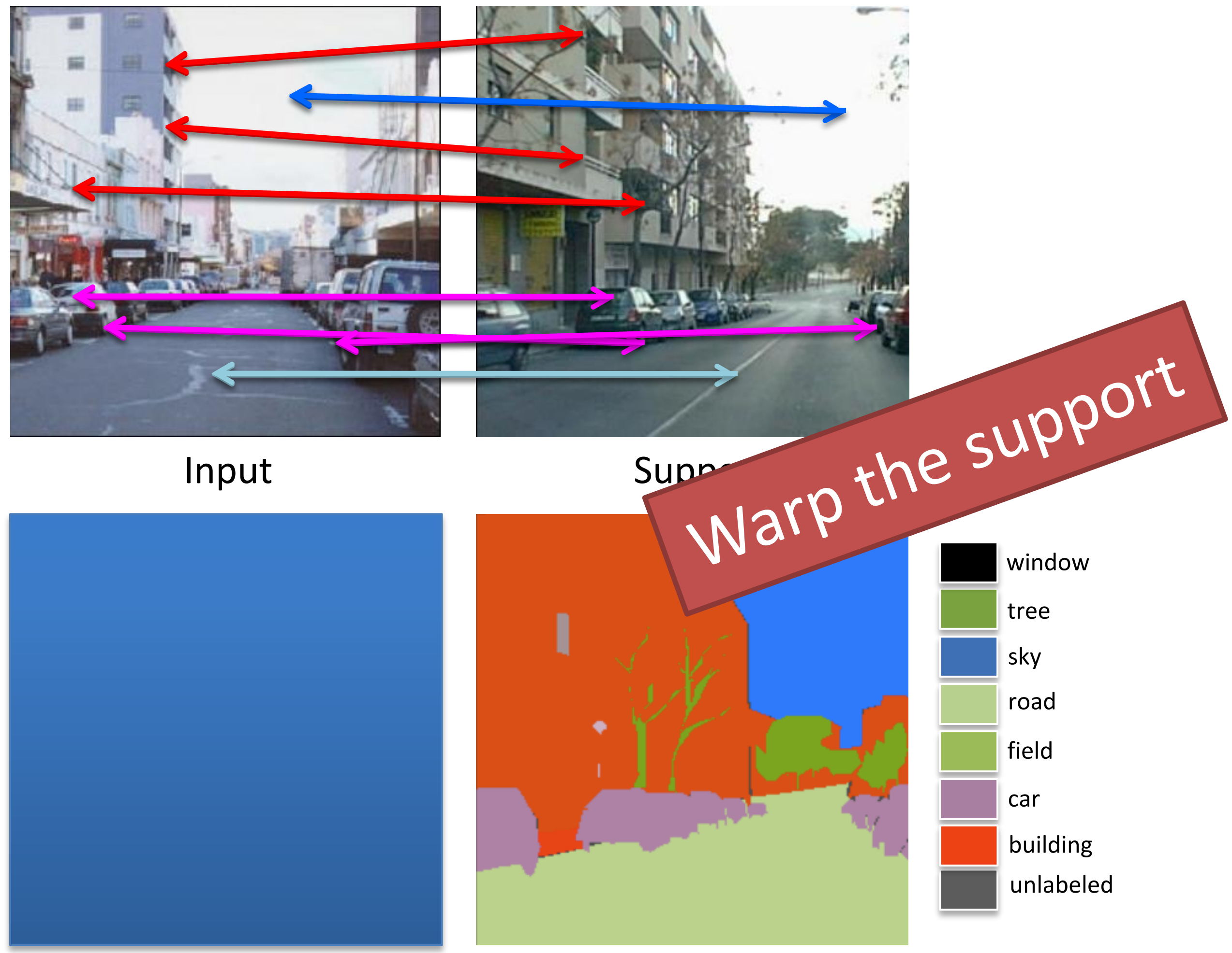
Vote the best label using the labels of the nearest neighbors



[Torralba, CVPR09]

Dense Correspondence

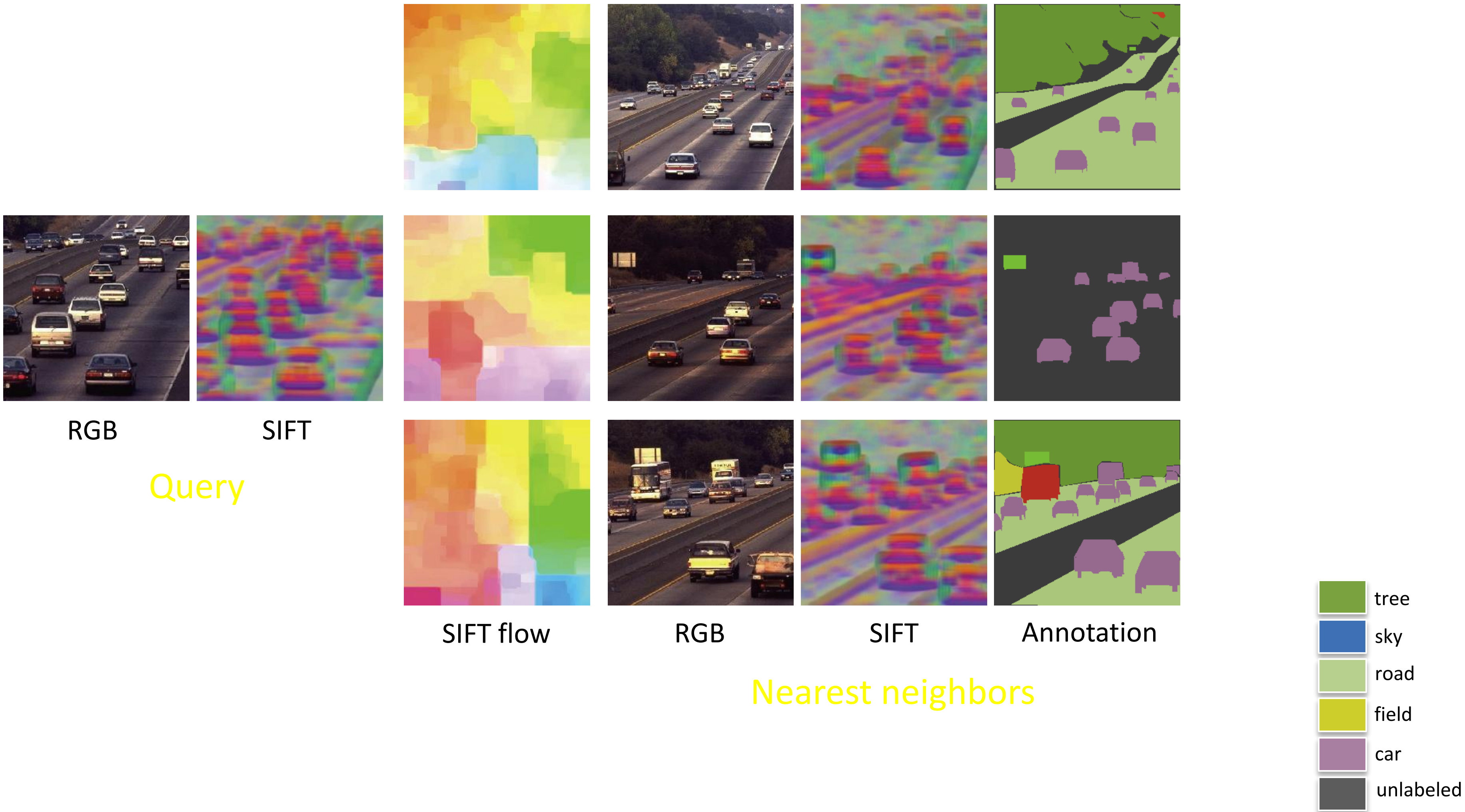
Find better correspondence



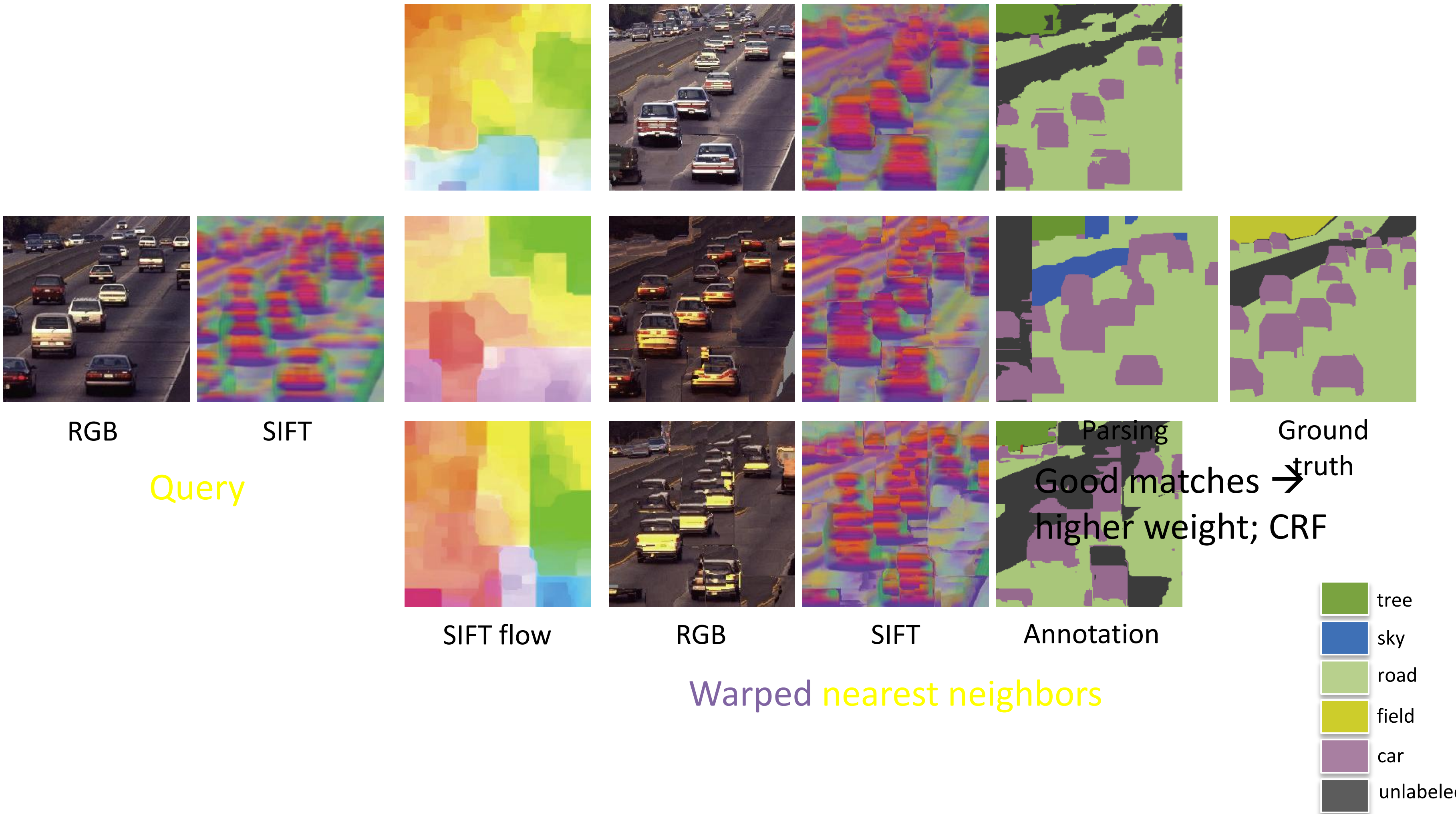
Warped annotation User annotation

Adapted slides of Liu

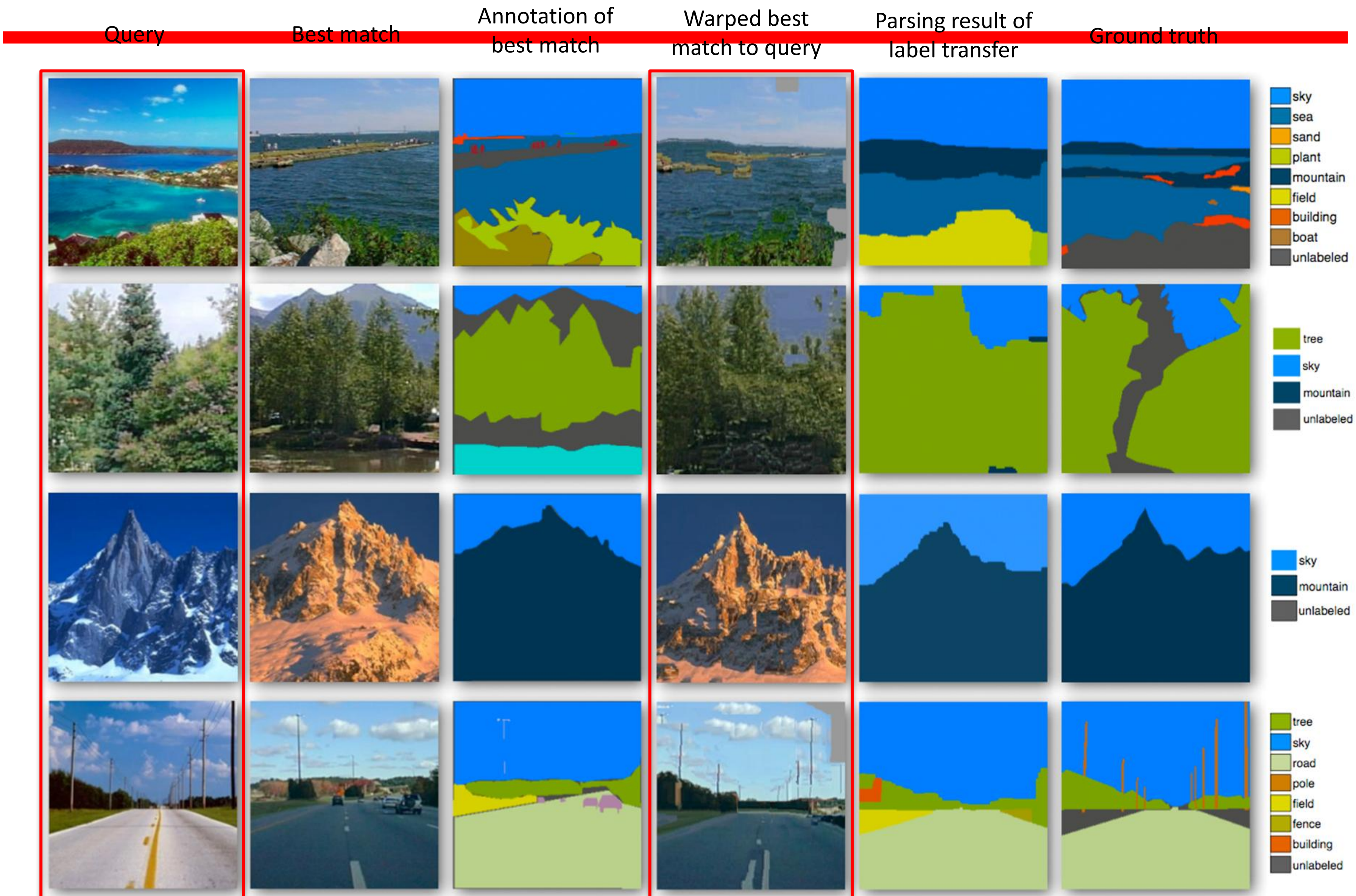
Label transfer system overview



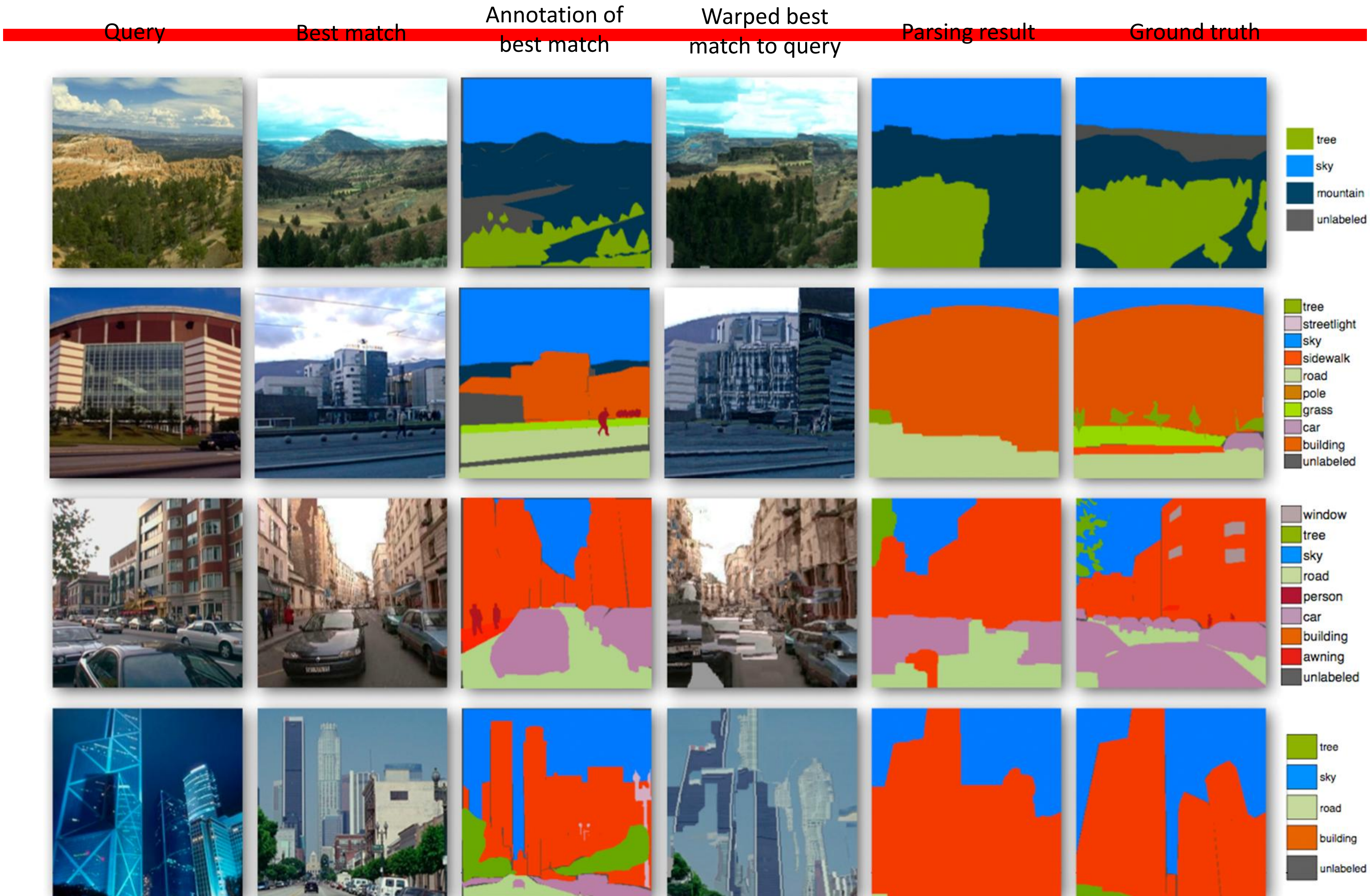
Label transfer system overview



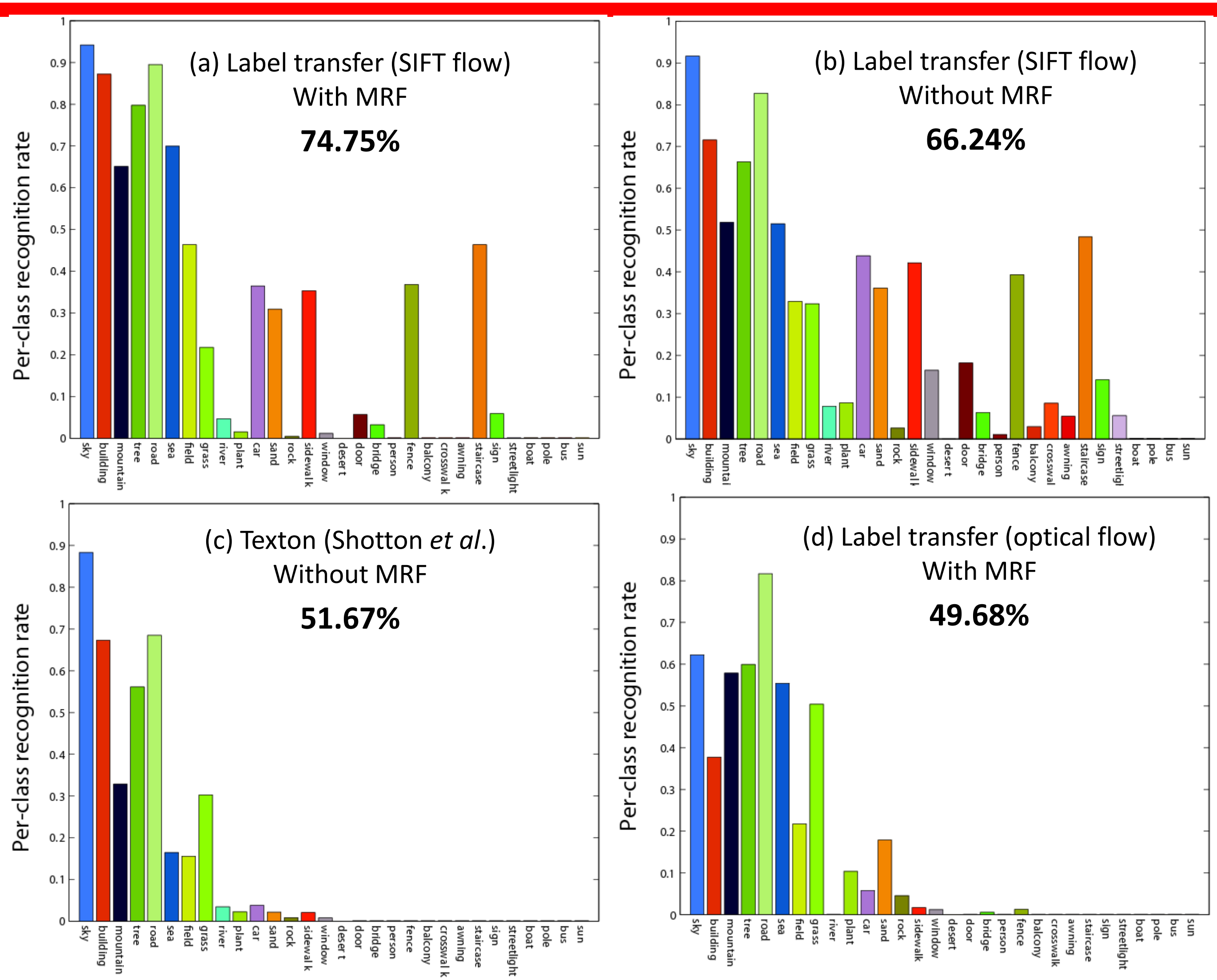
Scene parsing results (1)



Scene parsing results (2)



Comparison with the parametric model



J. Shotton *et al*. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV,* 2006

Conclusion

- Image Dataset is getting larger (tens of millions)
- Memory usage is essential for storing large scale dataset
- Non-parametric models are effective and popular for large dataset
- Hierarchical structure of the object categories can be effectively utilized

Discussion

- How many images do we need?

- What about the quality of the data
  - [Torralba and Efros, CVPR11]

- Is nearest neighbor really the best?

- New research problem
  - How to do learning with the large scale dataset?
  - *Fine-grained* object categorization

- Other interesting things with large dataset?

Thank you