80 Million Tiny Images

IPAM Workshop on Numerical Tools and Fast Algorithms for Massive Data Mining, Search Engines and Applications

October 23rd 2007

Massachusetts Institute of Technology

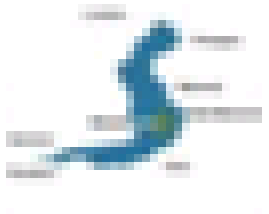
Antonio Torralba

Rob Fergus

William T. Freeman

Overview

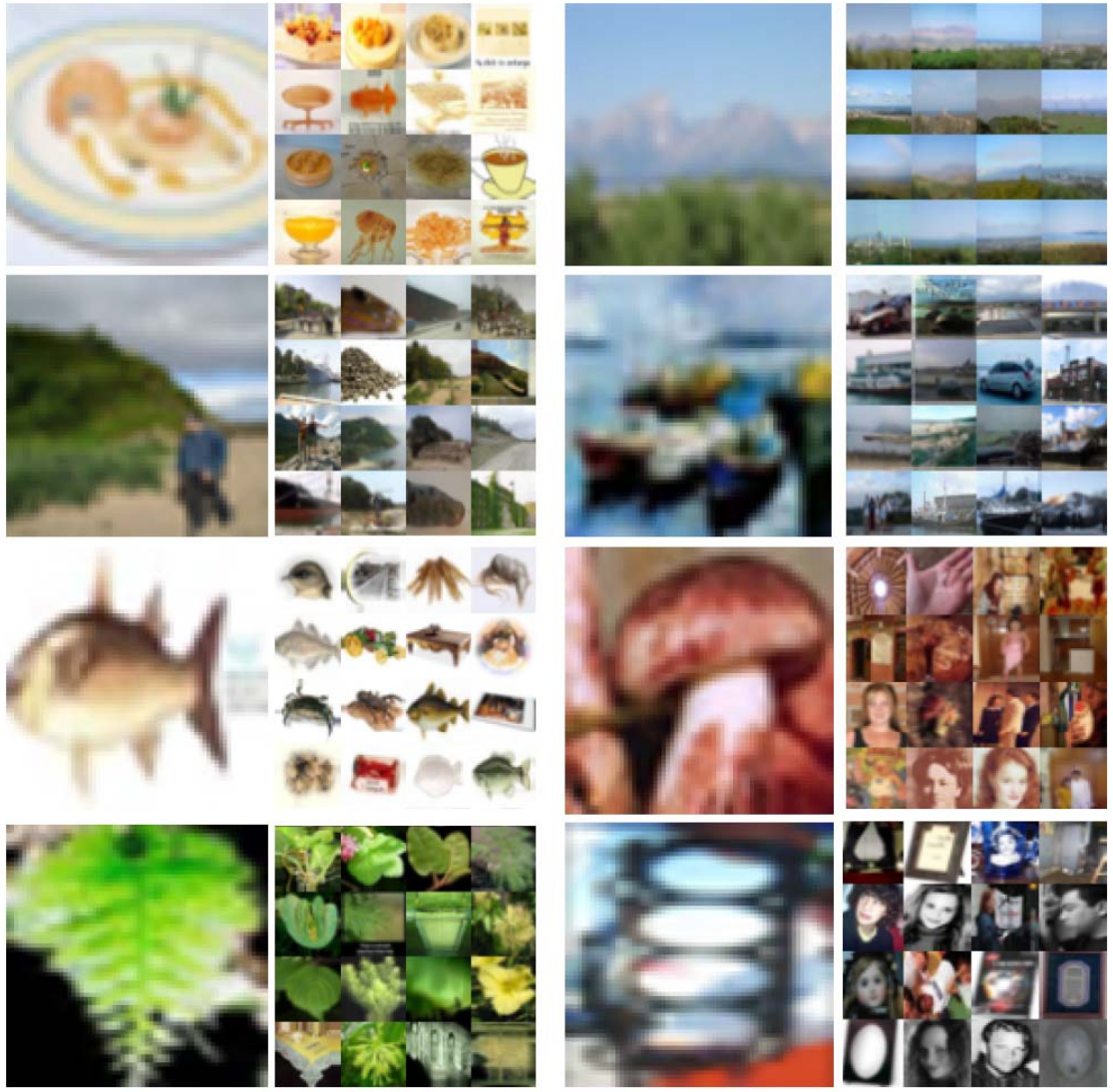
- Non-parametric approach to category-level recognition
- Dataset of 80 million images from Internet



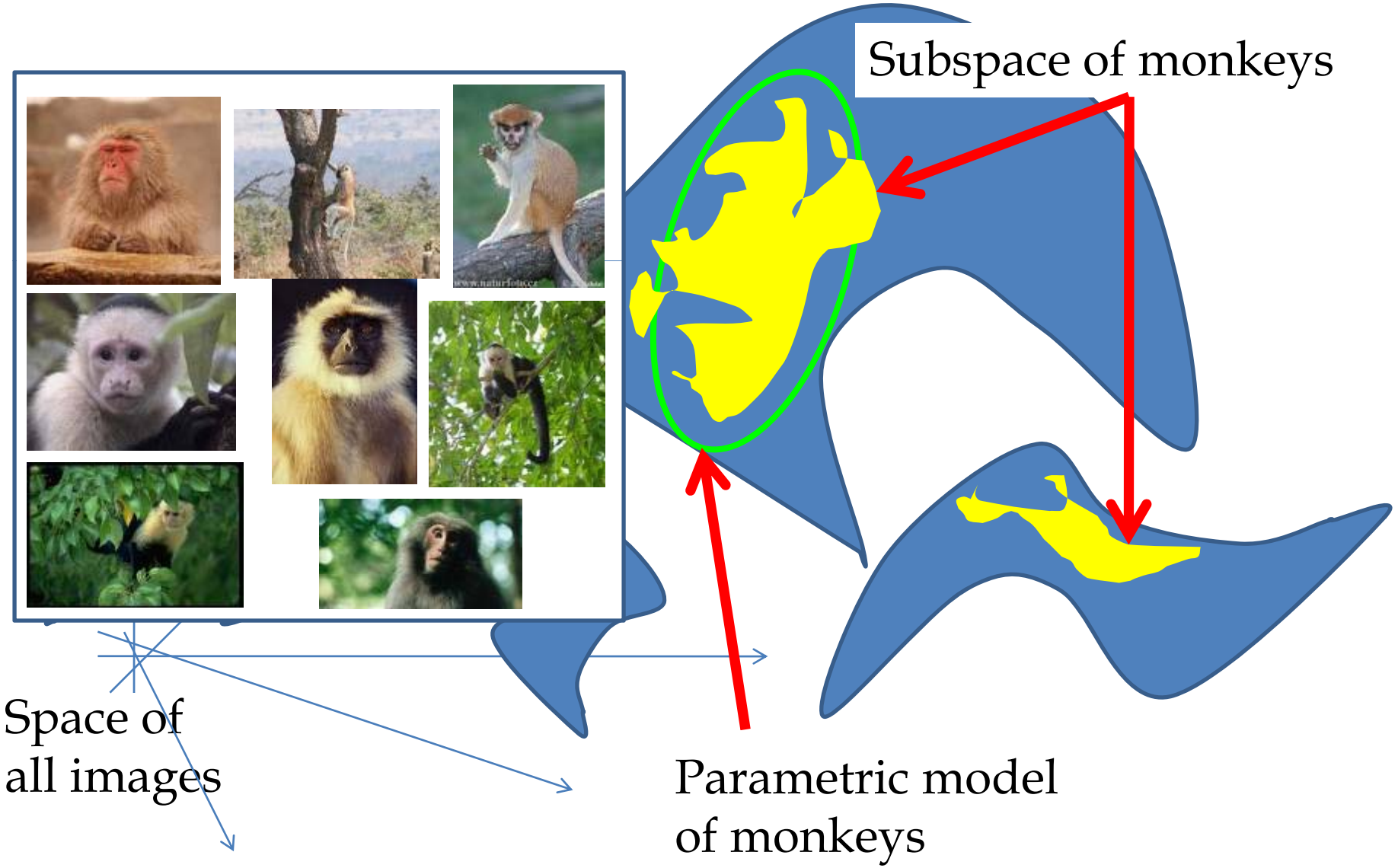
- Use very low resolution images (32x32 color)

Overview

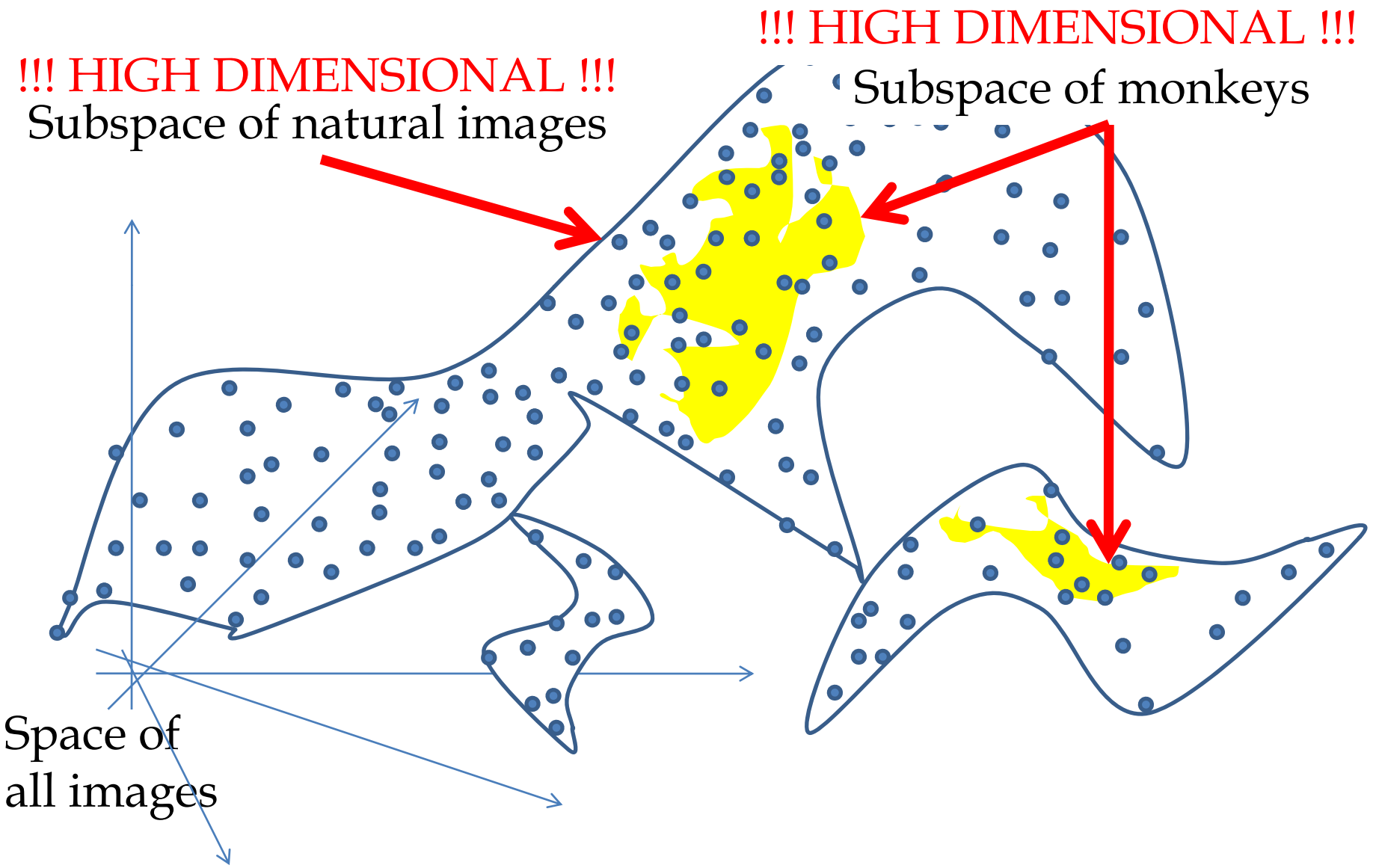
- Use simple algorithms: nearest neighbors



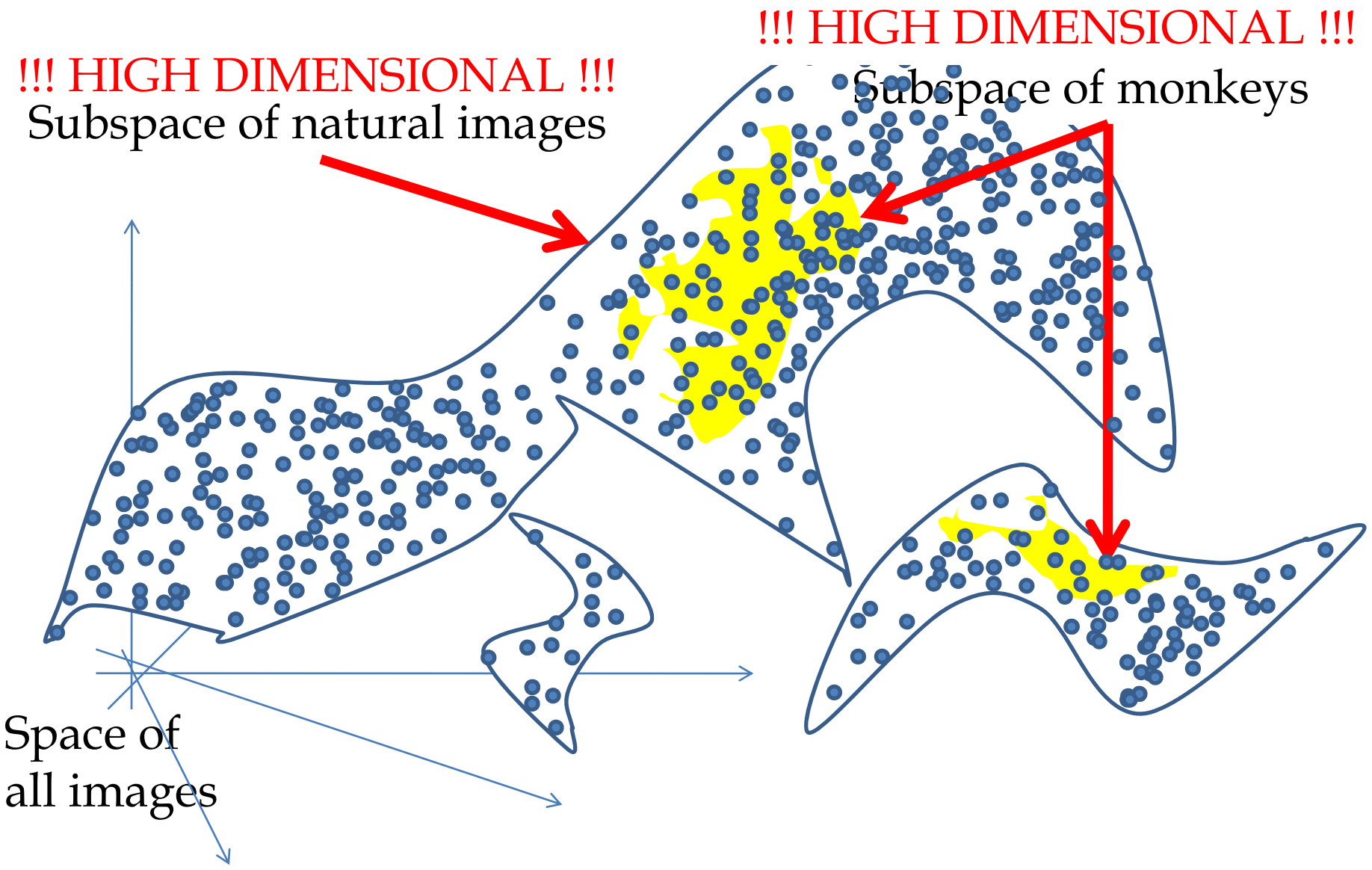
Motivation


Subspace of monkeys
Space of
all images
Parametric model
of monkeys

Non-parametric Approach



Non-parametric Approach



The Data

Thumbnail Collection Project

- Collect images for ALL objects
  - List obtained from WordNet
  - 75,378 non-abstract nouns in English

- Example first 20:

a-bomb
a-horizon
a._conan_doyle
a._e._burnside
a._e._housman
a._e._kennelly
a.e.
a_battery
a_cappella_singing
a_horizon
a_kempis
aalborg
aalii
aalost
aalto
aar
aardvark
aardwolf
aare
aare_river

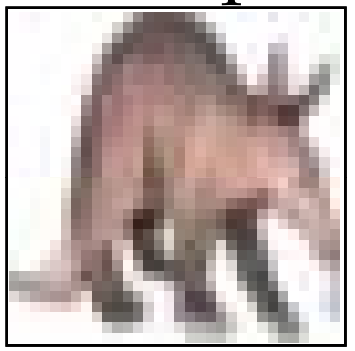
Thumbnail Collection

- 7 different search engines

Dataset Statistics

- Overall stats
  - 79,302,017 images
  - 75,062 different words
- Details
  - Two formats: square & rectangular
  - Gathered at 4.5 images/second
  - Downloaded 97,245,098 images
  - 18% duplicate rate
  - Disk usage: ~ 700Gb
  - Collection time: ~ 9 months

32x32 square



32xN rectangular

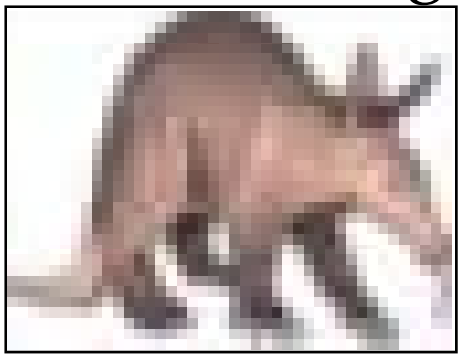
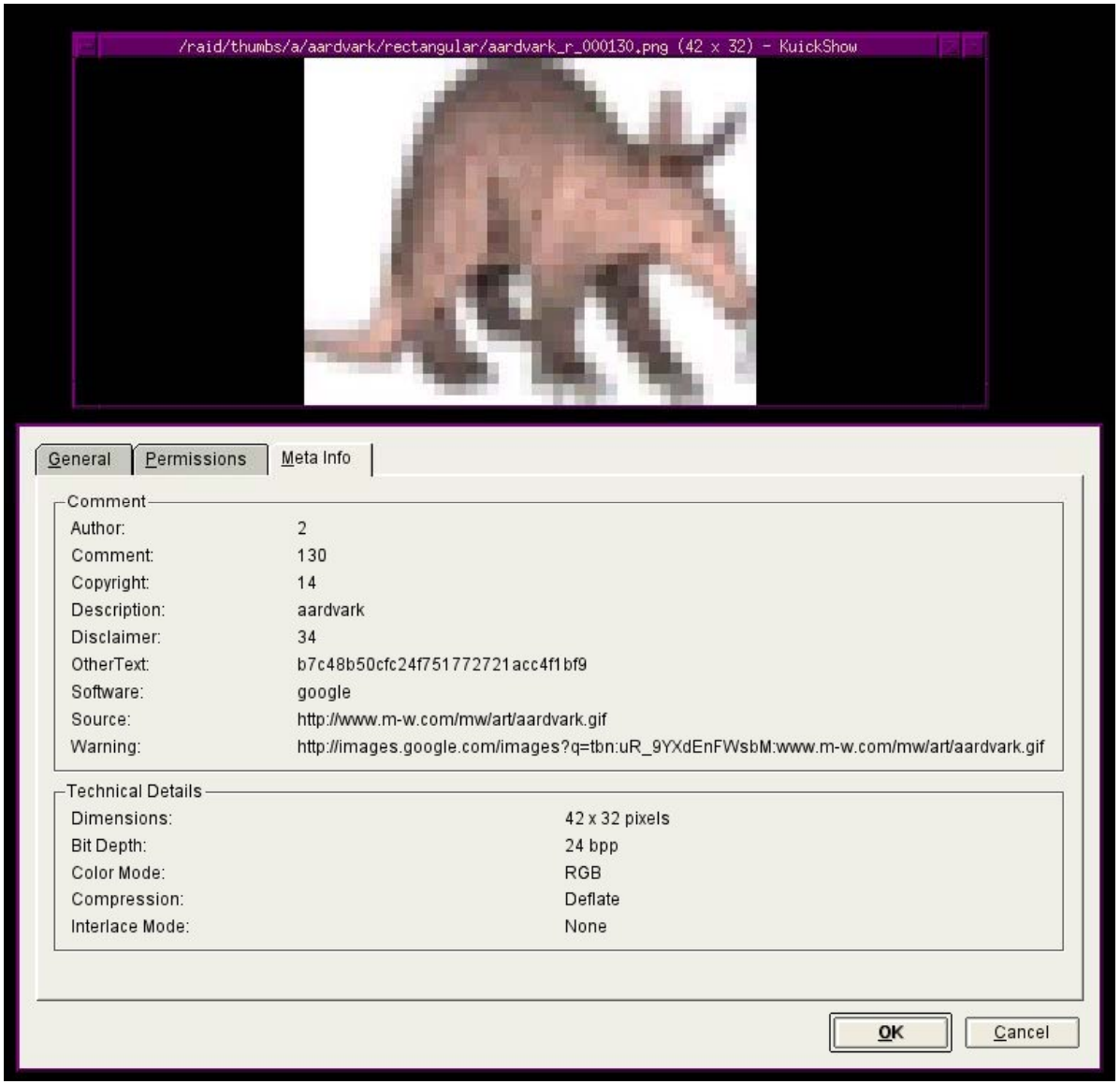
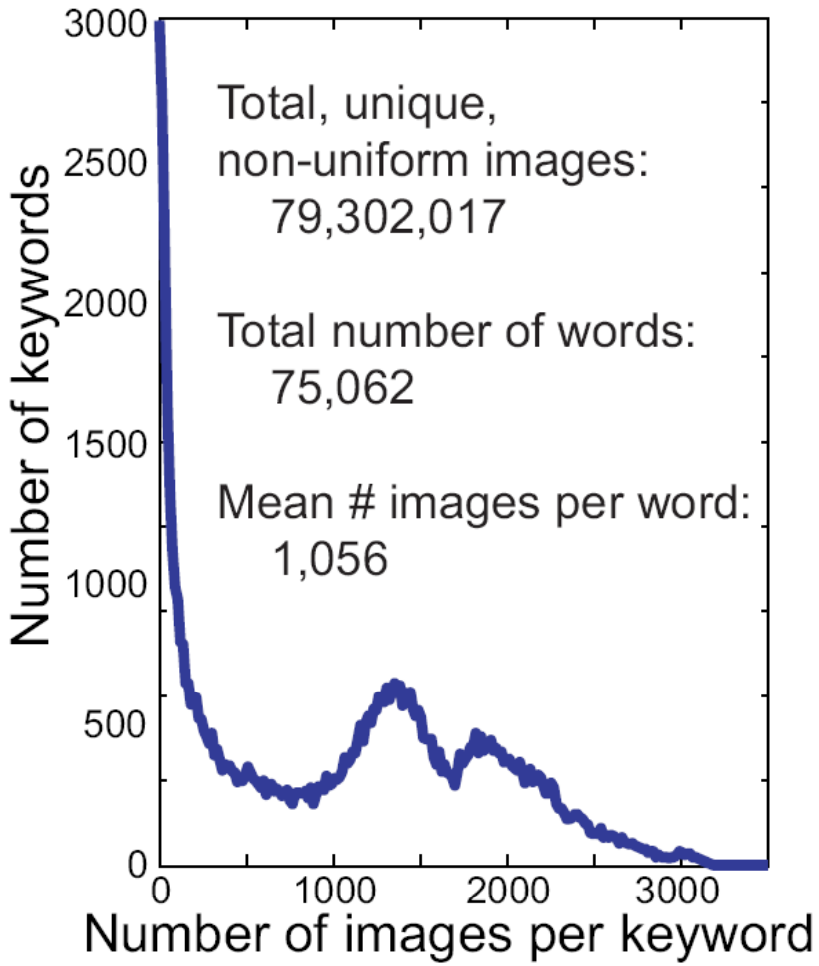


Image Metadata

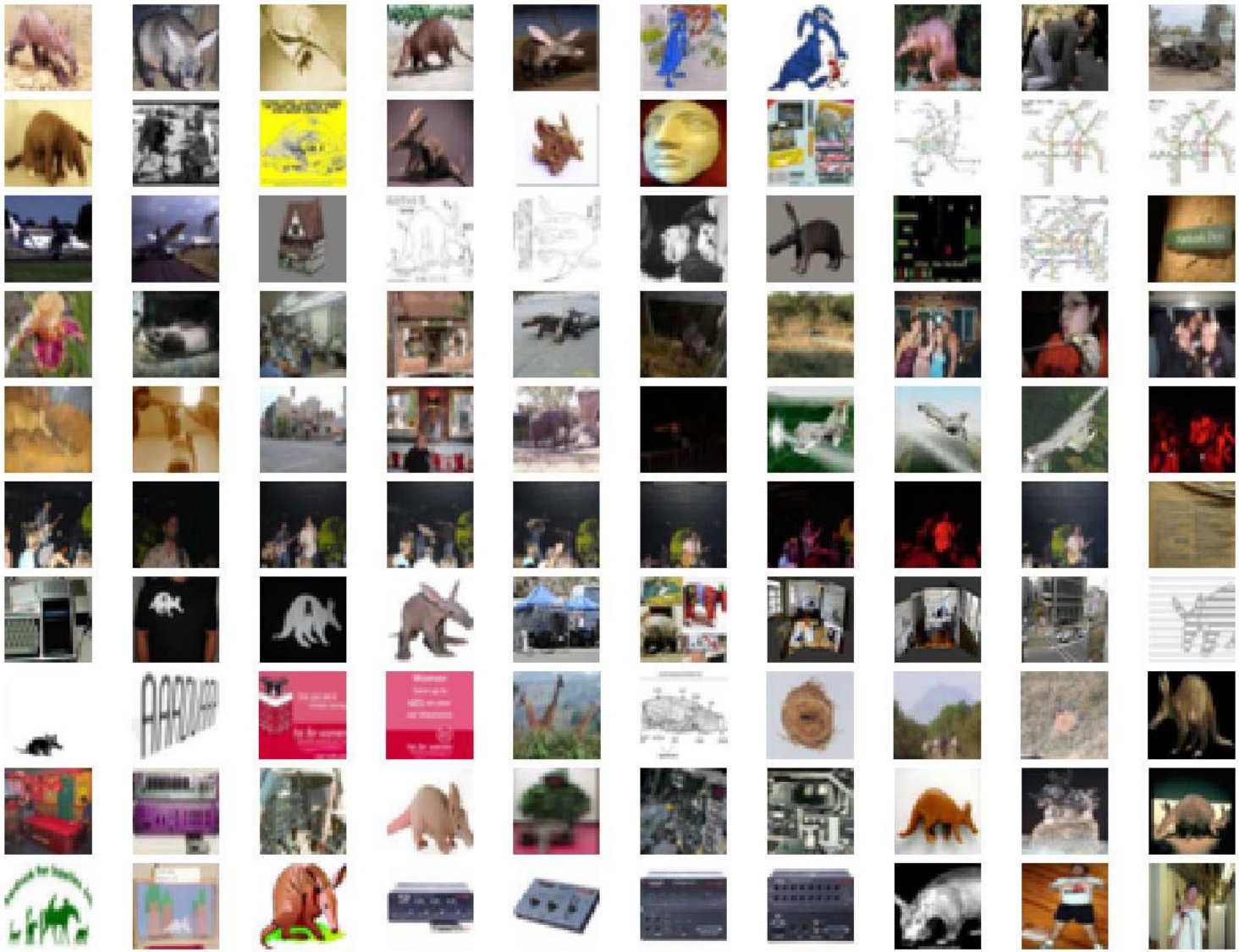
- URL to high-res
- URL of thumbnail
- Engine & Rank



Histogram Images/Word

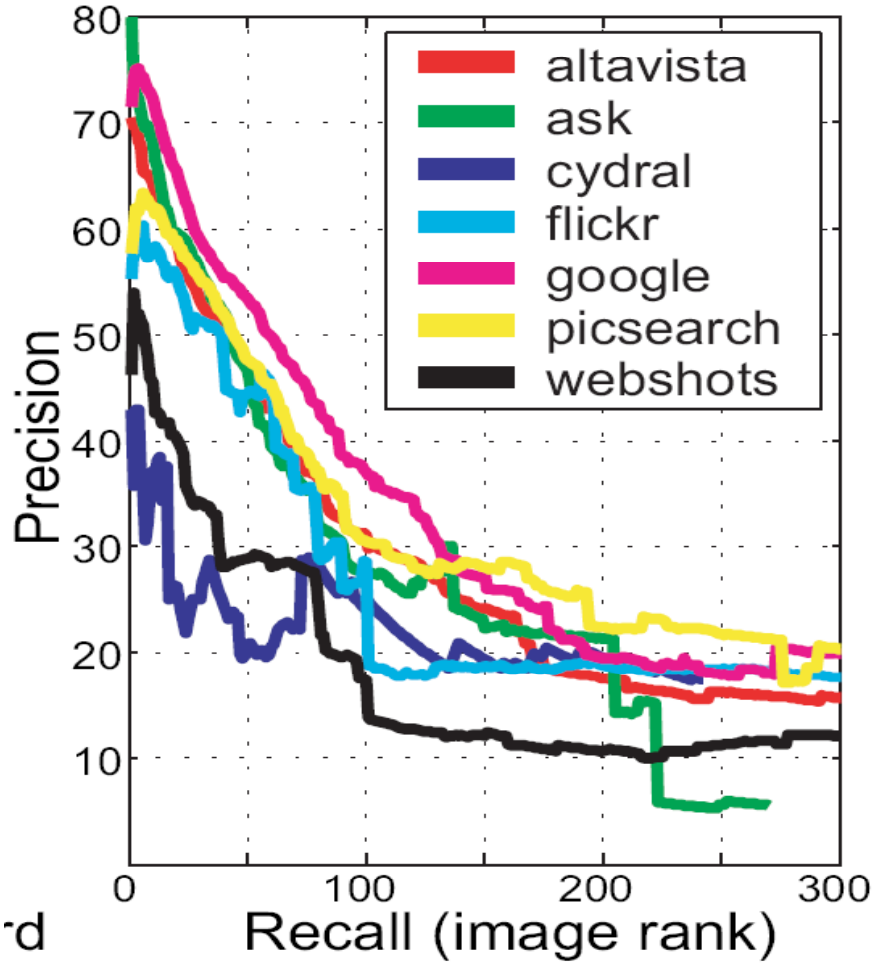


Aardvark Images



Labeling Noise

- Manual labeling of 78 classes
- Best: Google & Altavista
- Worst: Cydral & Webshots



Representation

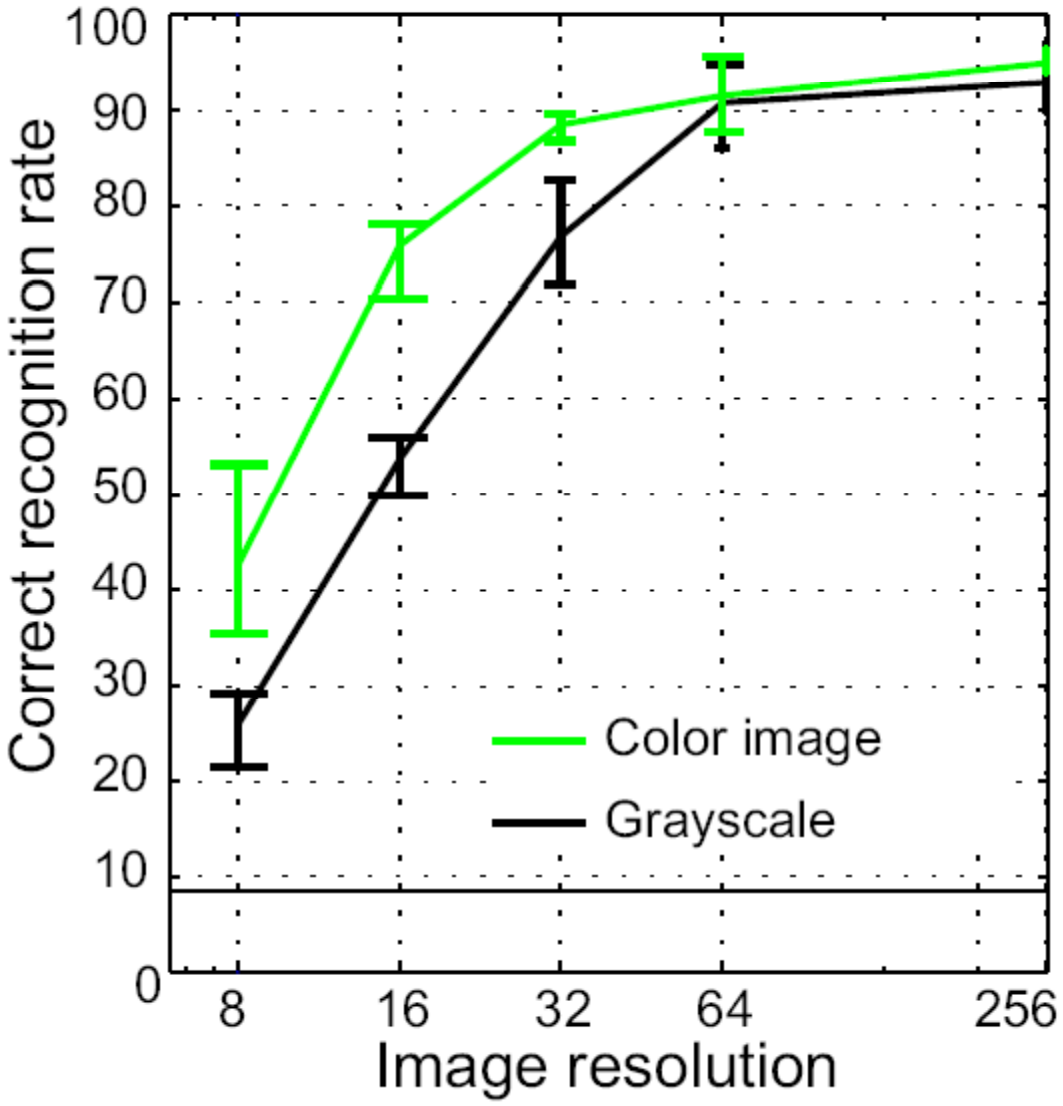
Suitable Image Representation

- Want minimal representation for task:
  - Classifying scene and dominant objects

- Compact representation has low storage requirements

- We blur & subsample to give low-res image (32x32 color)

Human Performance at Scene Recognition



The role of context in object recognition
A. Oliva, A. Torralba
Trends in Cognitive Sciences, in press.
December 2007.

Human Labeling of Tiny Scenes

Image Patches vs Tiny Images

a) Patches

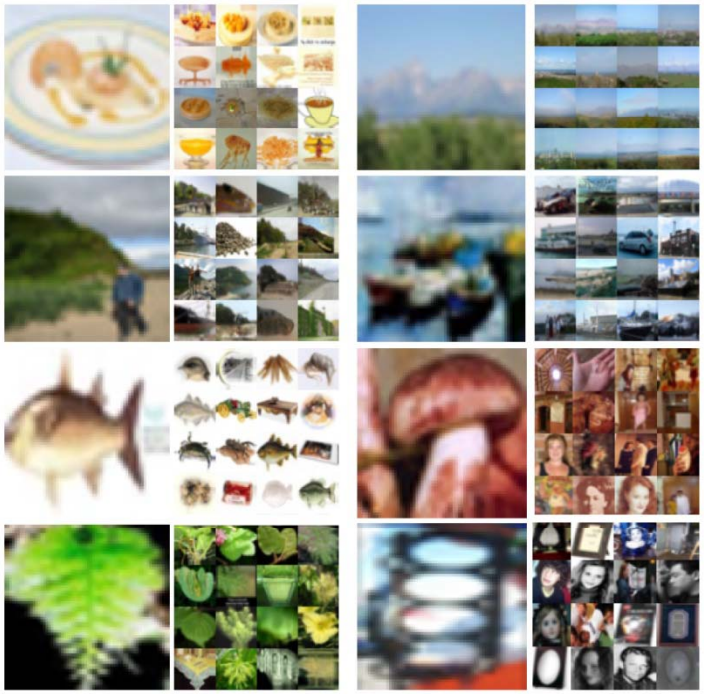
b) Object chips

c) Scenes, small thumbnails

Approach

Non-parametric Classifier

- Nearest-neighbors
- For each query, obtain sibling set (neighbors)
- 3 different types of distance metric
- Hand-designed, use whole image



Metric 1 - D_{ssd}

- Sum of squared differences (SSD)

$$D^2_{ssd} = \sum_{x,y,c} \left[\text{Image 1} - \text{Image 2} \right]^2$$

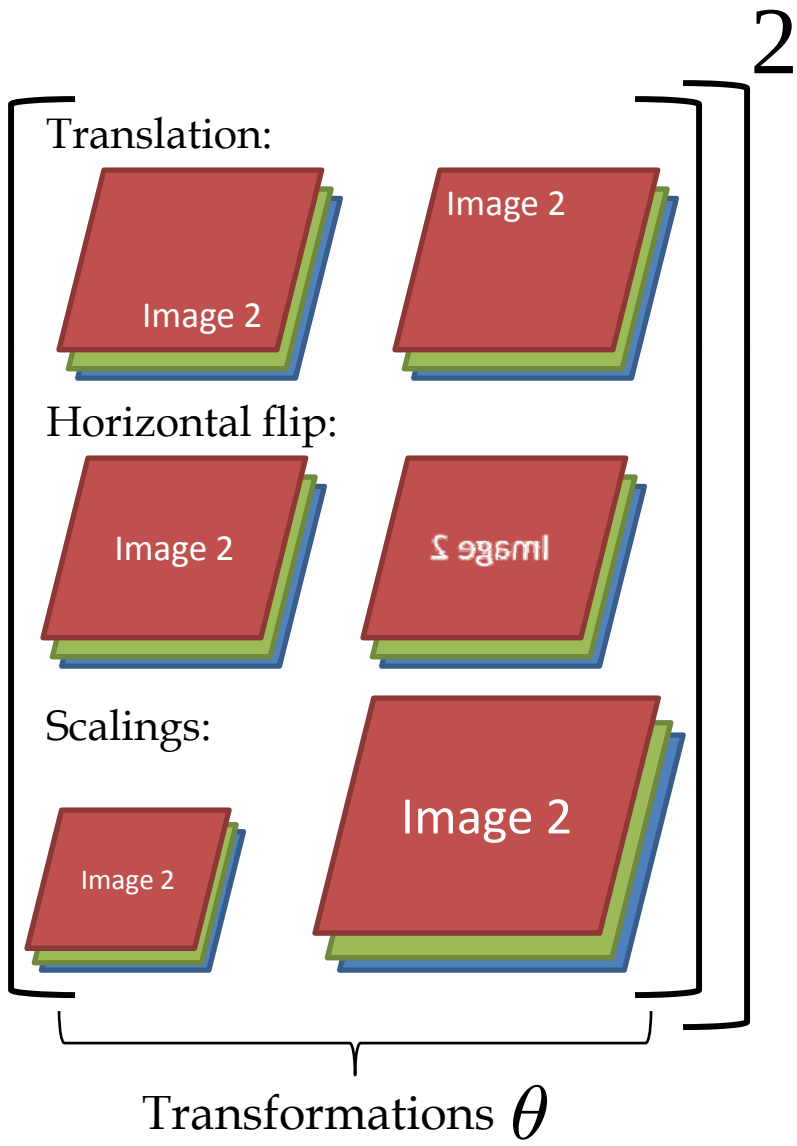
To give invariance to illumination:
Each image normalized to
be zero mean, unit variance

Target Neighbor

Metric 2 - D_{warp}

- SSD but allow small transformations

$$D^2_{warp} = \min_{\theta} \sum_{x,y,c} [\text{Image 1} - \text{Transformations } \theta]^2$$



Find min using gradient descent

Metric 3 - D_{shift}

- As per Warping but also allow sub-window shifts

$$D^2_{shift} = \sum_{x,y,c} \left[\text{Image 1} - \underset{\theta}{\text{Transformed}} \left[\text{Image 2} \right] \right]^2$$

Start with warped version of image 2, as per D_{warp}

Metric 3 - D_{shift}

- As per Warping but also allow sub-window shifts

$$D^2_{shift} = \sum_{x,y,c} \left[\text{  } - \underset{\theta}{\text{Transformed}} \left[\text{  } \right] \right]^2$$

Start with warped version of image 2, as per D_{warp}

Metric 3 - D_{shift}

- As per Warping but also allow sub-window shifts

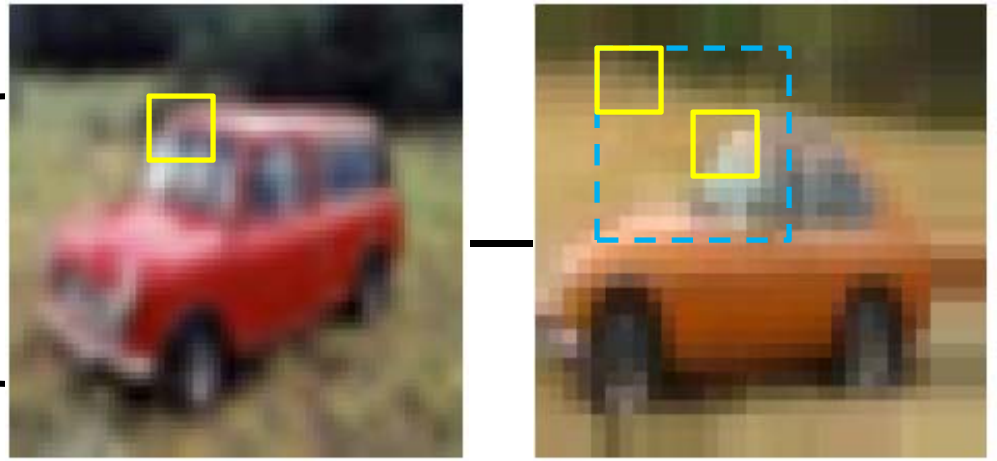
$$D^2_{shift} = \sum_{x,y,c} \left[\text{  } - \text{  } \right]^2$$

Start with warped version of image 2, as per D_{warp}

Metric 3 - D_{shift}

- As per Warping but also allow sub-window shifts

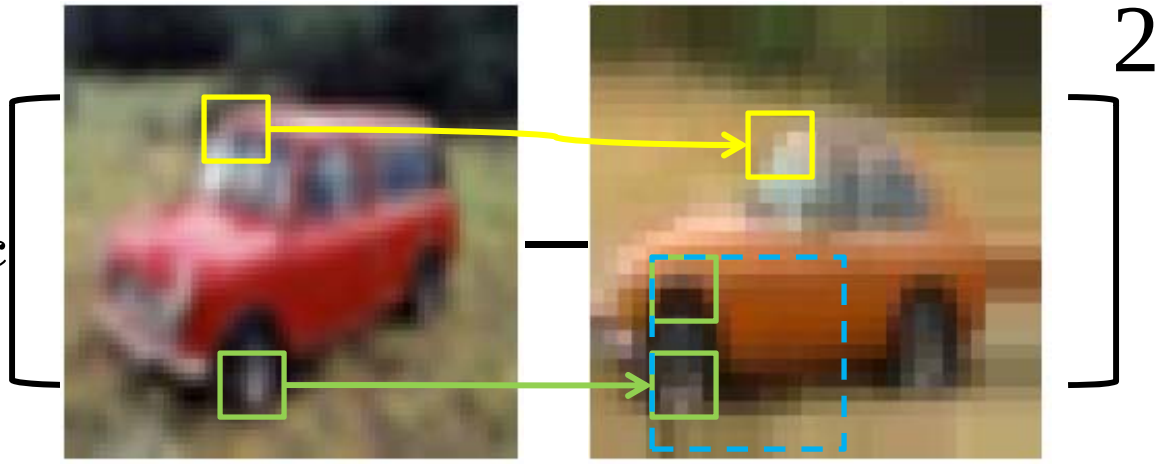
$$D^2_{shift} = \min_{\substack{\text{Local} \\ \text{sub-window}}} \sum_{x,y,c} \left[\quad - \quad \right]^2$$



Metric 3 - D_{shift}

- As per Warping but also allow sub-window shifts

$$D^2_{shift} = \min_{\substack{\text{Local} \\ \text{sub-window}}} \sum_{x,y,c}$$



- Quick since images are so small

Metric 3 - D_{shift}

- As per Warping but also allow sub-window shifts

$$D^2_{shift} = \min_{\substack{\text{Local} \\ \text{sub-window}}} \sum_{x,y,c} \left[\right]^2$$

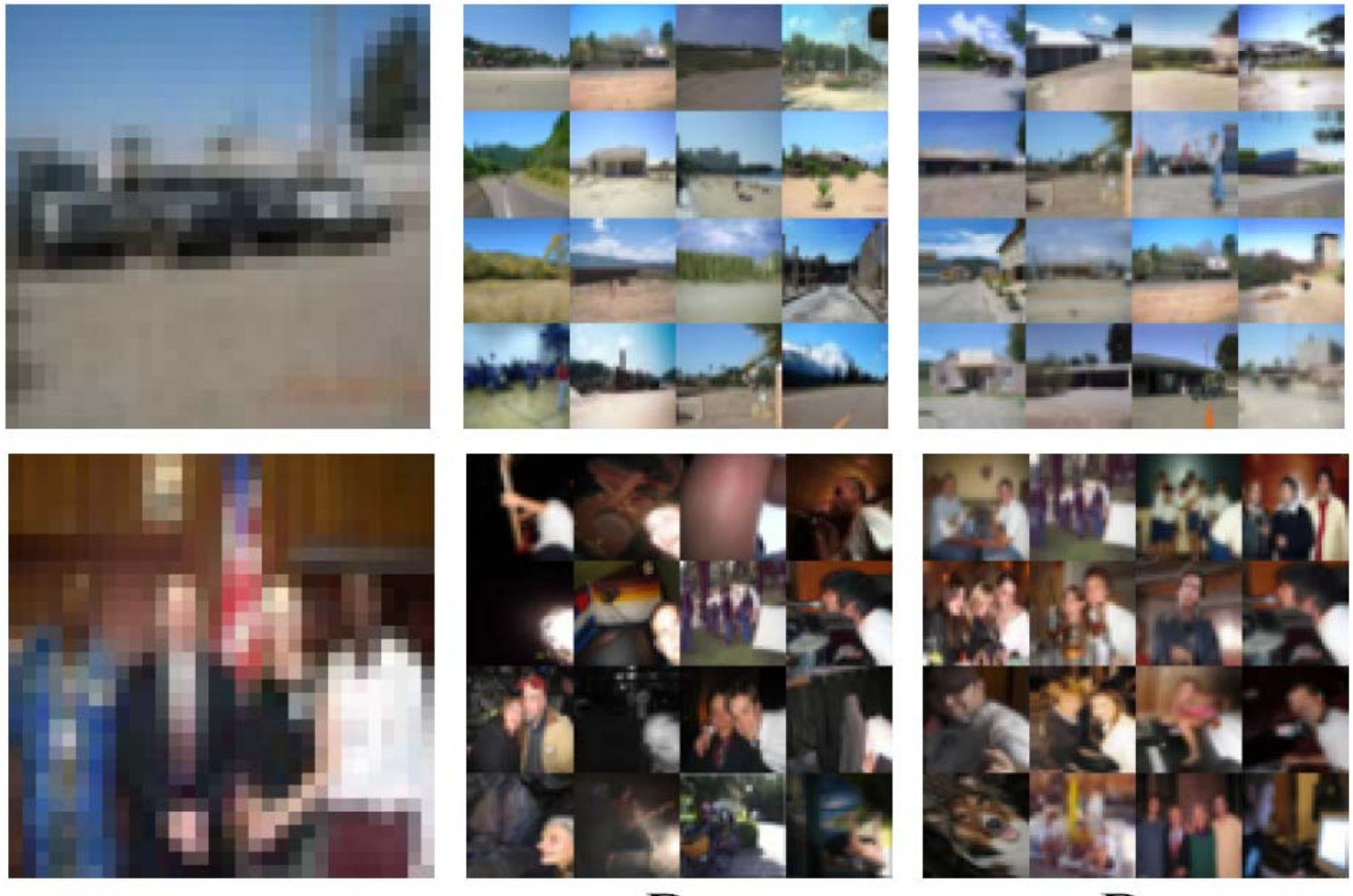
Tried various sizes of sub-window

→ 1x1 (i.e. single pixel) worked best

Comparison of metrics

Sibling Sets with Different Metrics

- Sibling set is 50 images

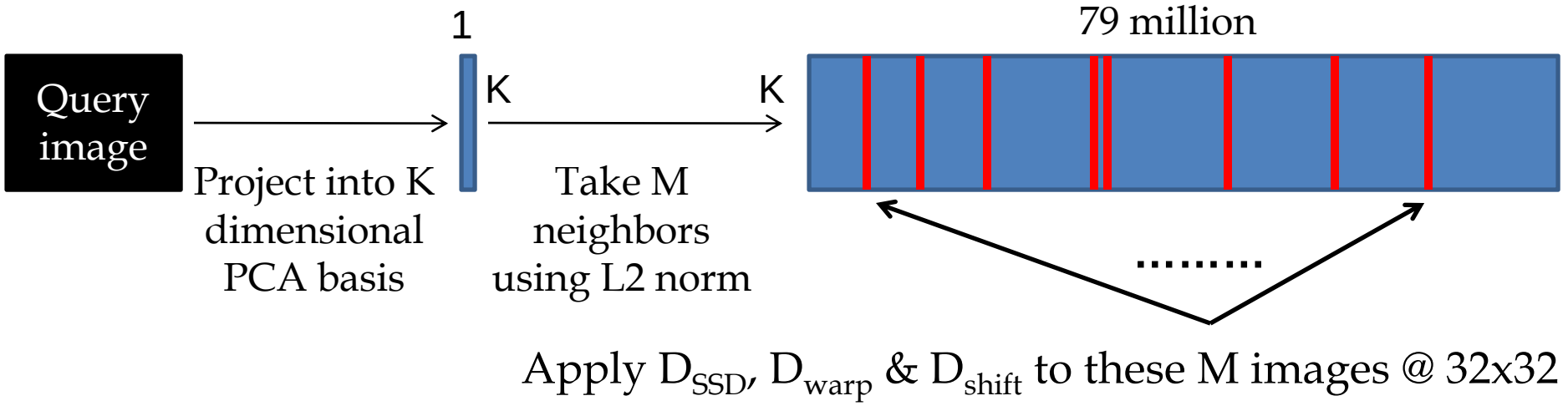


D_{ssd} D_{shift}

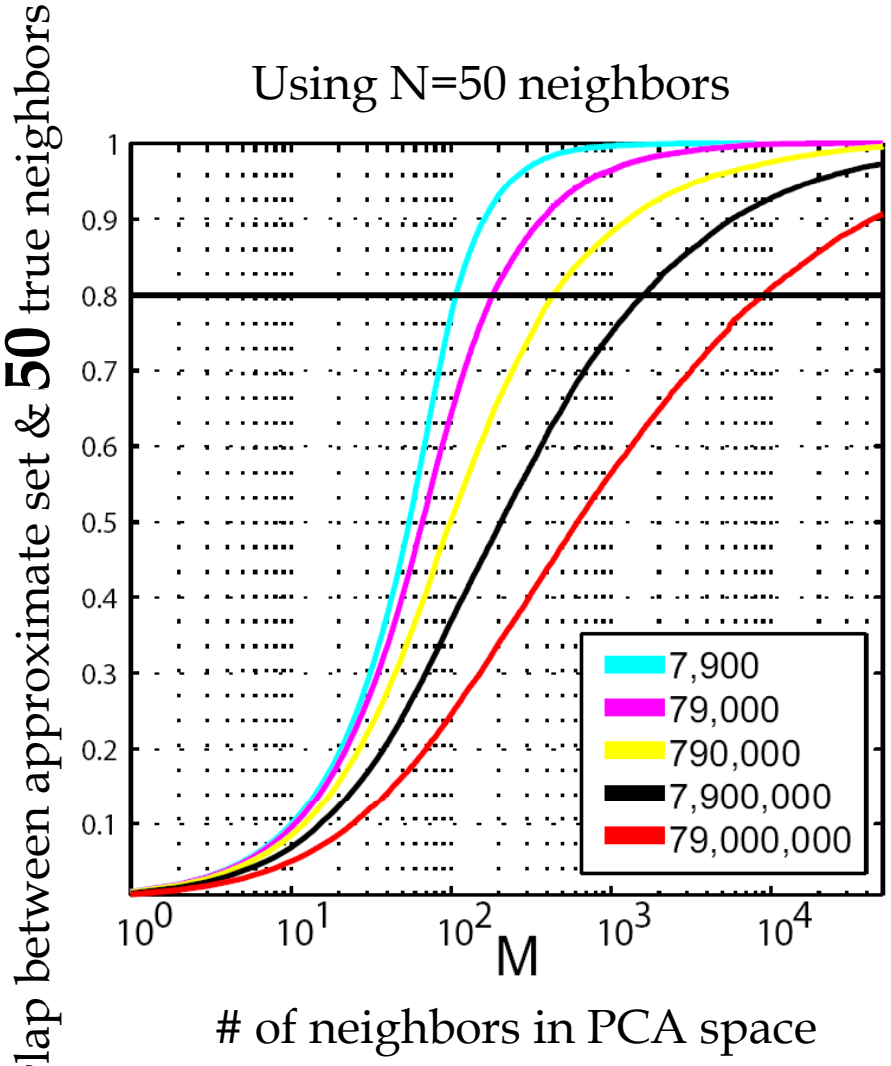
Approximate D_{ssd}

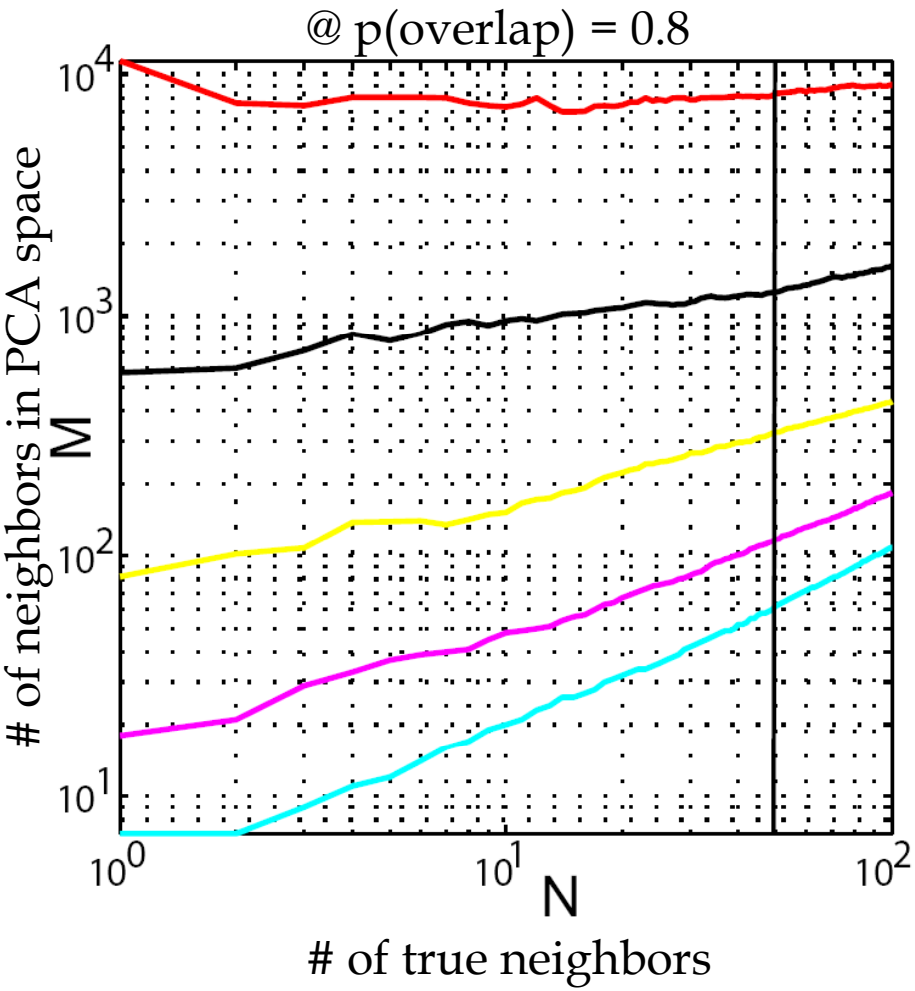
- Exact distance metrics are too expensive to apply to all 79 million images

- Use approximate scheme based on taking first K=19 principal components



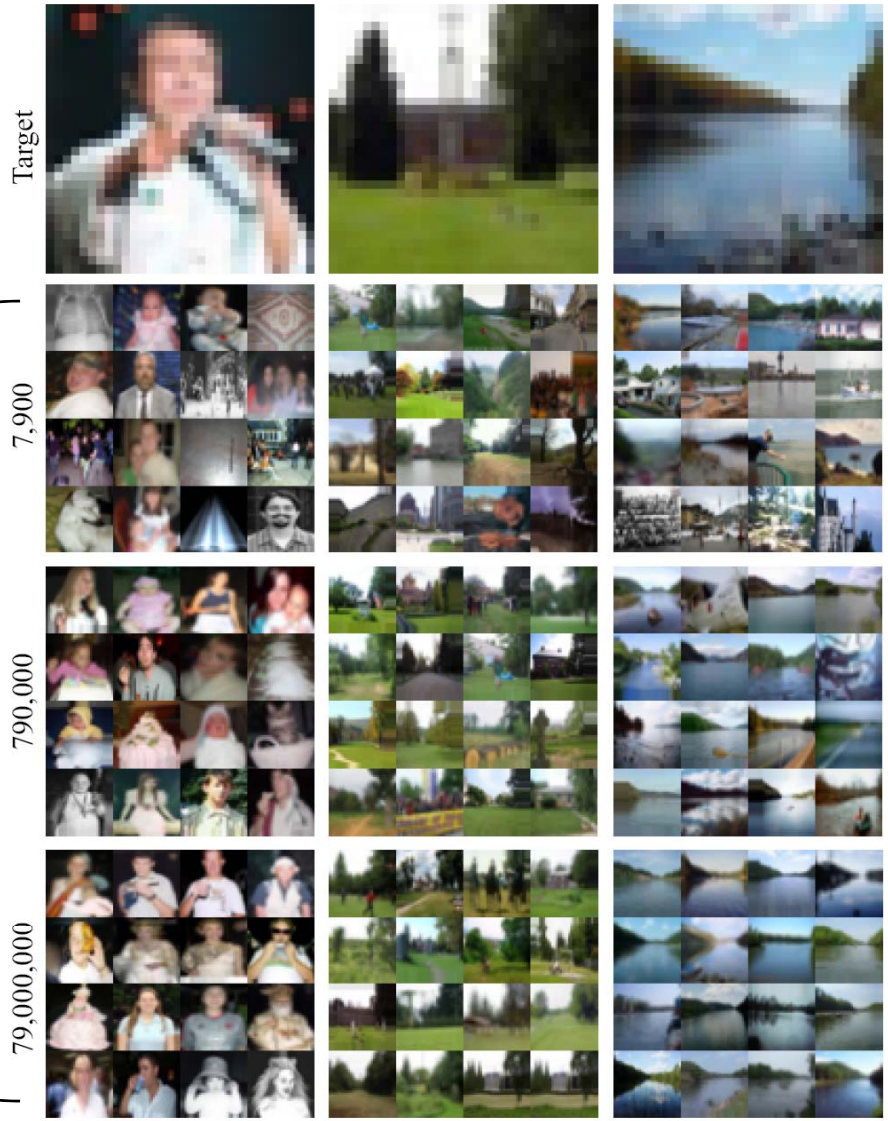
Exact SSD vs Approximate SSD





Quality of Sibling Set using D_{shift}

Size of dataset



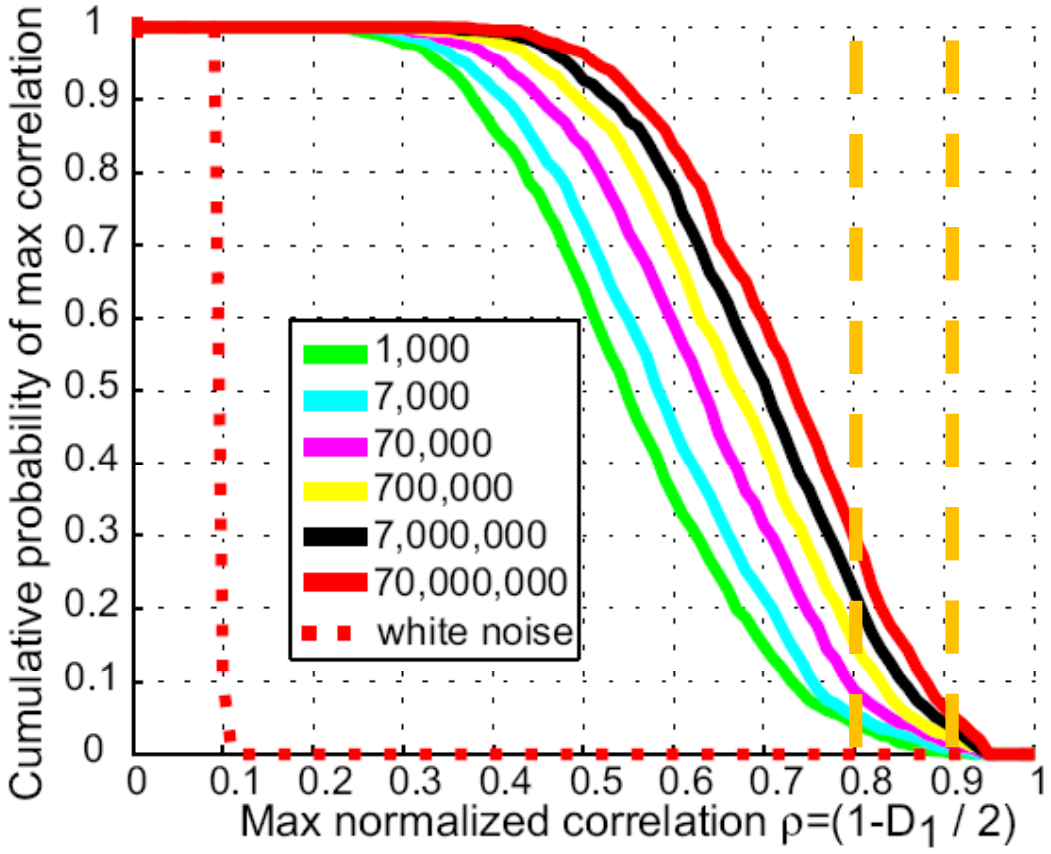
10^5

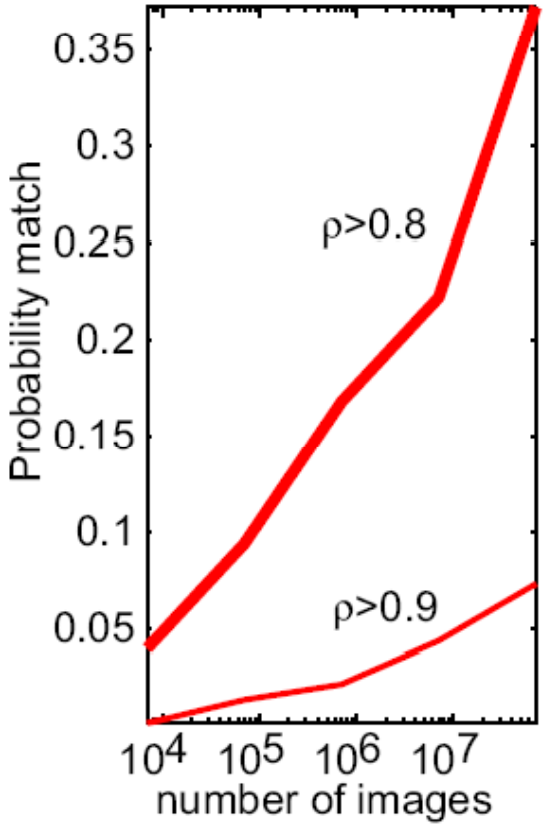
10^6

10^8

Exploring the Manifold of Images

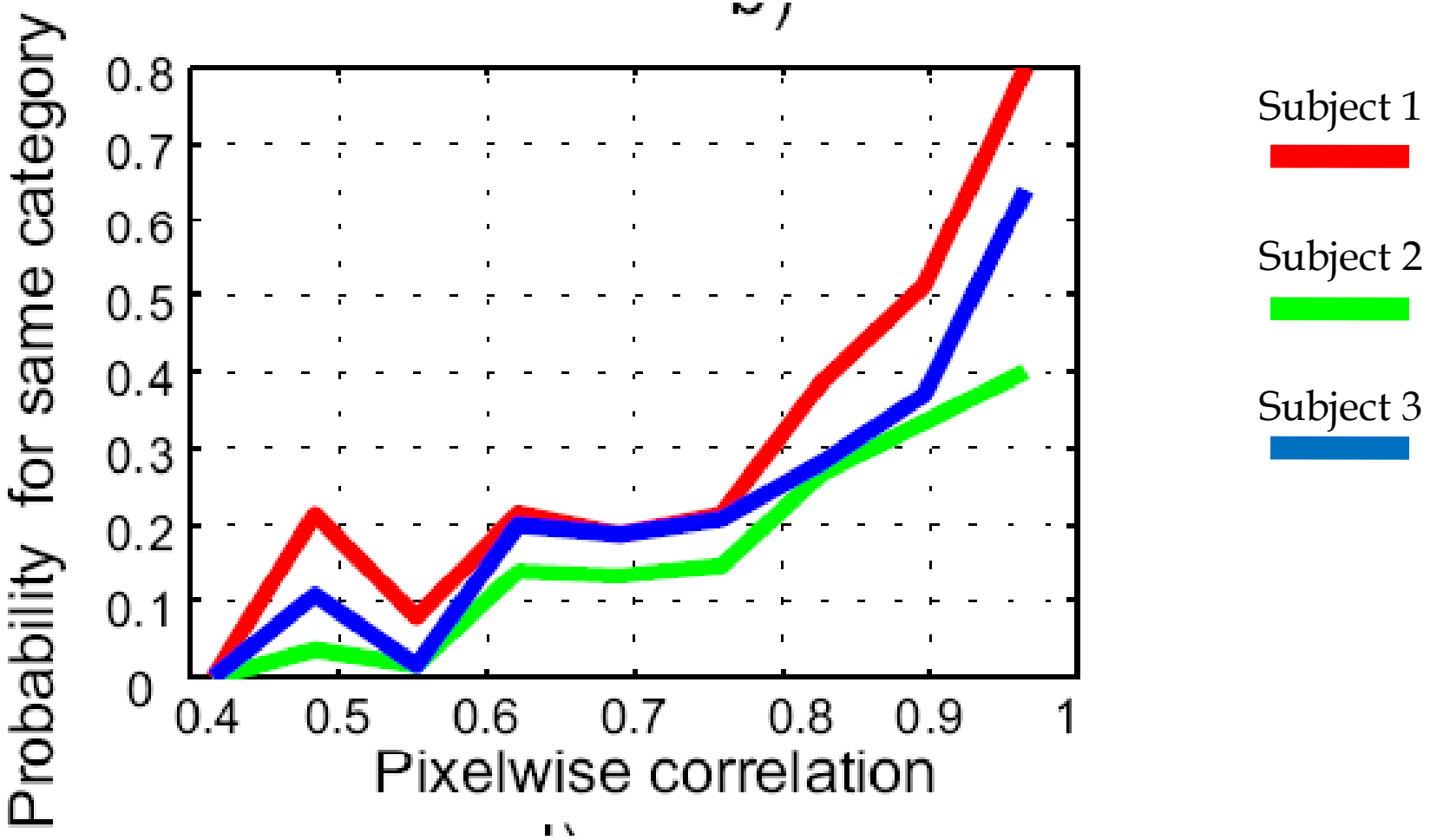
How Many Images Are There?





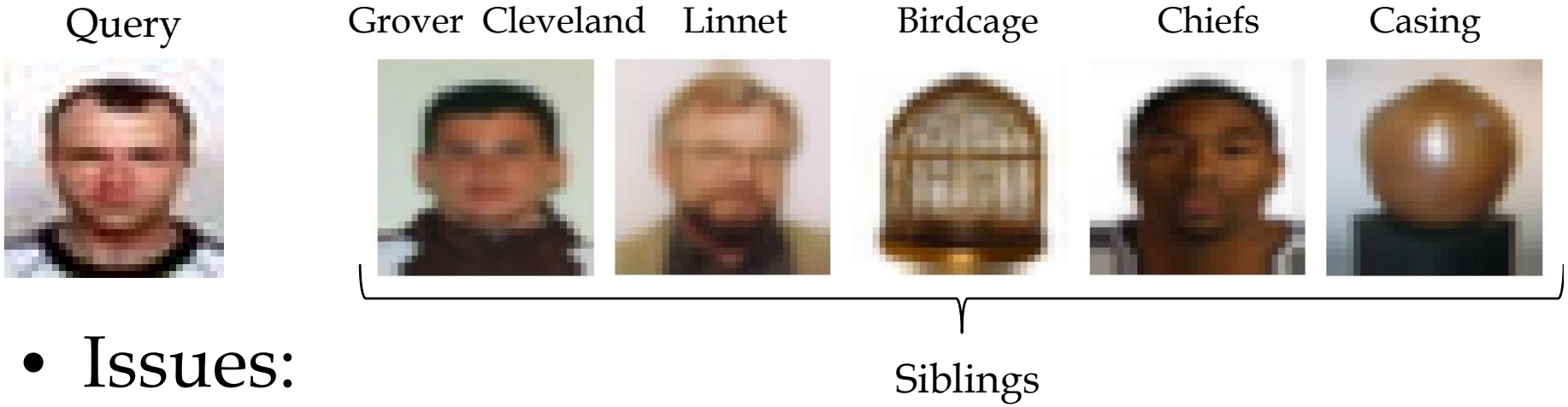
Note: $D_1=D_{SSD}$

How Does D_{ssd} Relate to Semantic Distance?



Label Assignment

- Distance metrics give set of nearby images
- How to compute label?



- Issues:
  - Labeling noise
  - Keywords can be very specific
    - e.g. yellowfin tuna

Wordnet – a Lexical Dictionary

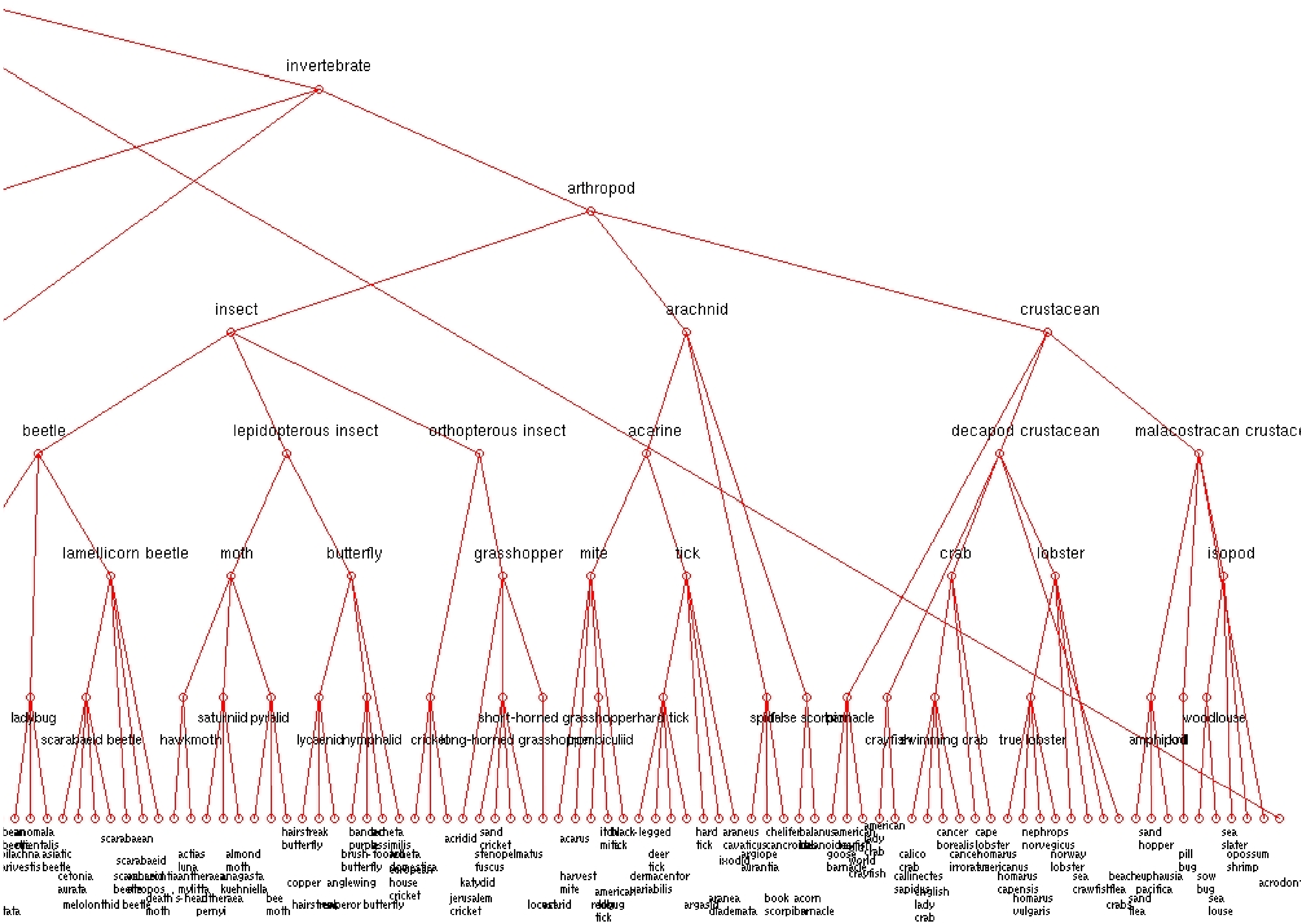
http://wordnet.princeton.edu/

```
Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun aardvark

Sense 1
aardvark, ant bear, anteater, Orycteropus afer
       => placental, placental mammal, eutherian, eutherian mammal
           => mammal
               => vertebrate, craniate
                   => chordate
                       => animal, animate being, beast, brute, creature
                           => organism, being
                               => living thing, animate thing
                                   => object, physical object
                                       => entity
```

Wordnet Hierarchy

```
Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun aardvark

Sense 1
aardvark, ant bear, anteater, Orycteropus afer
       => placental, placental mammal, eutherian, eutherian mammal
           => mammal
               => vertebrate, craniate
                   => chordate
                       => animal, animate being, beast, brute, creature
                           => organism, being
                               => living thing, animate thing
                                   => object, physical object
                                       => entity
```

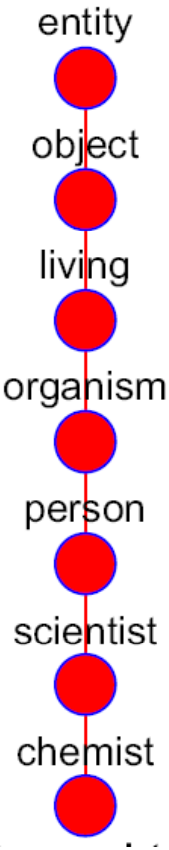
- Convert graph structure into tree by taking most common meaning


invertebrate
arthropod
insect
arachnid
crustacean
beetle
lepidopterous insect
orthopterous insect
acarine
decapod crustacean
malacostracan crustacean
lamellicorn beetle
moth
butterfly
grasshopper
mite
tick
crab
lobster
isopod
ladybug
scarabaeid beetle
hawkmoth
saturniid
pyralid
lycaenid
nymphalid
cricket
long-horned grasshopper
short-horned grasshopper
hard tick
spider
false scorpion
barnacle
crayfish
swimming crab
true lobster
amphipod
woodlouse
scarabaean
asiatic beetle
cetonia aurata
actias luna
almond moth
bee moth
hairstreak butterfly
copper
anglewing
acridid
stenopelmatus fuscus
house cricket
katydid
jerusalem cricket
acarus
harvest mite
deer tick
dermacentor variabilis
hard tick
argiope aurantia
aranea diademata
book scorpion
acorn barnacle
calico crab
english lady crab
cancer borealis
cape lobster
nephrops norvegicus
norway lobster
homarus capensis
homarus vulgaris
sea crawfish
beach flea
sand hopper
sand flea
pill bug
sow bug
sea slater
opossum shrimp
sea louse
acrodont

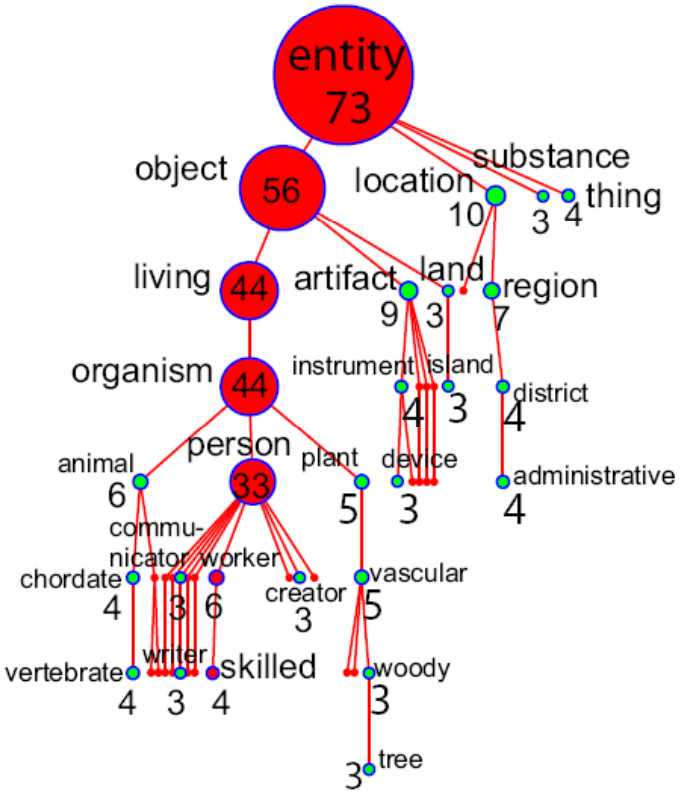
Wordnet Voting Scheme

a) Input image

b) Neighbors



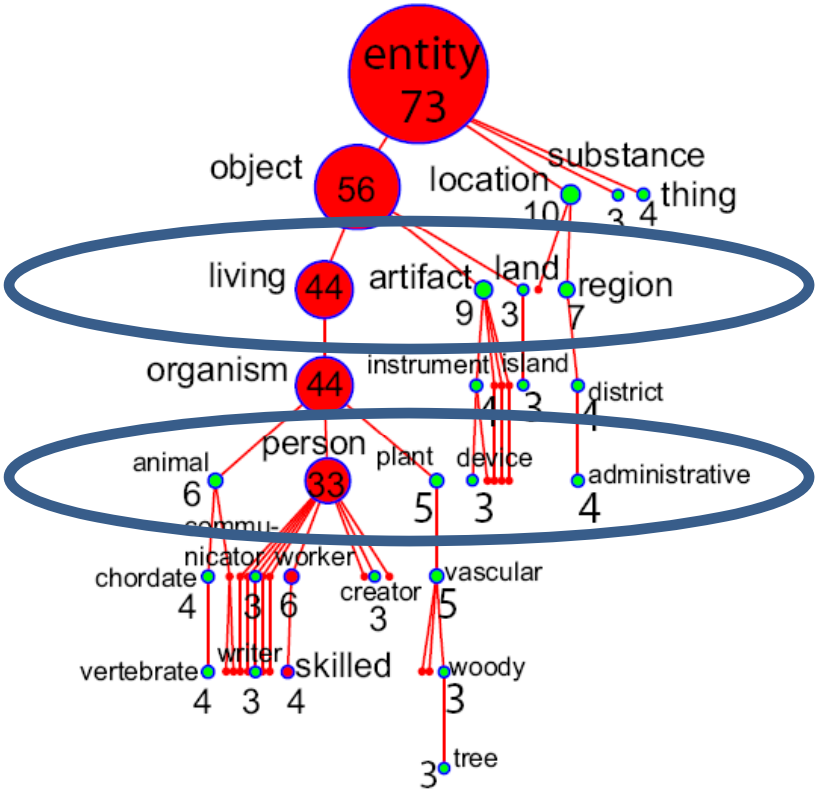
Ground truth



d) Wordnet voted branches

One image – one vote

Classification at Multiple Semantic Levels



d) Wordnet voted branches

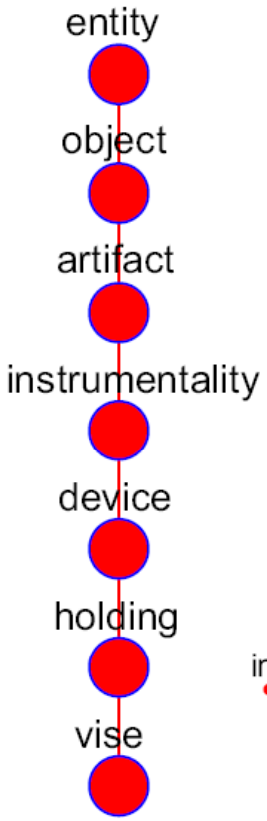
Votes:

Animal	64
Person	93
Plant	5
Region	7
Others	22

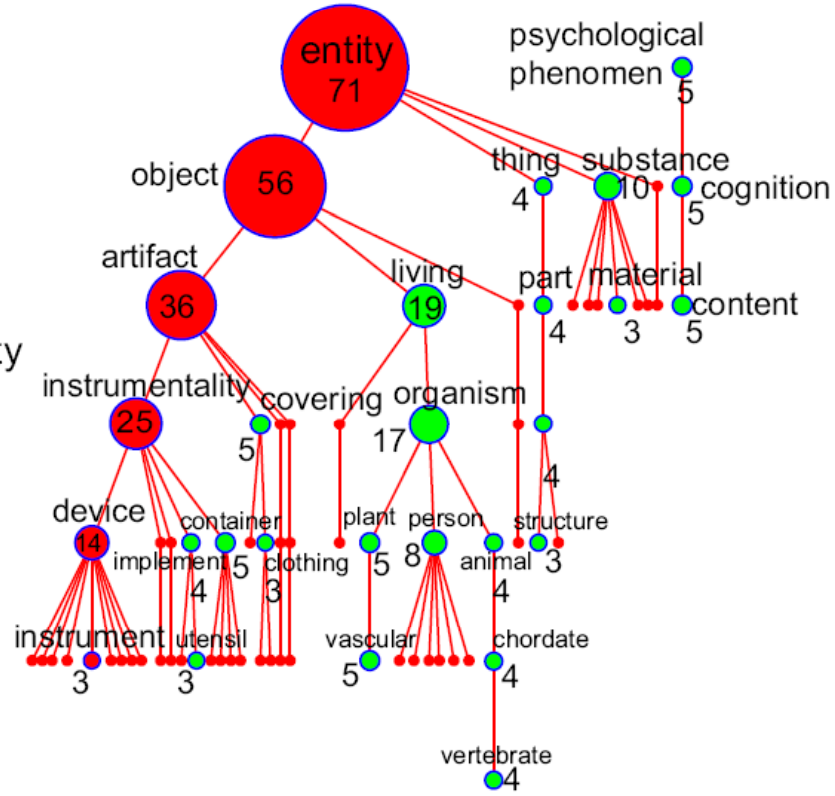
Wordnet Voting Scheme

a) Input image

b) Neighbors



c) Ground truth



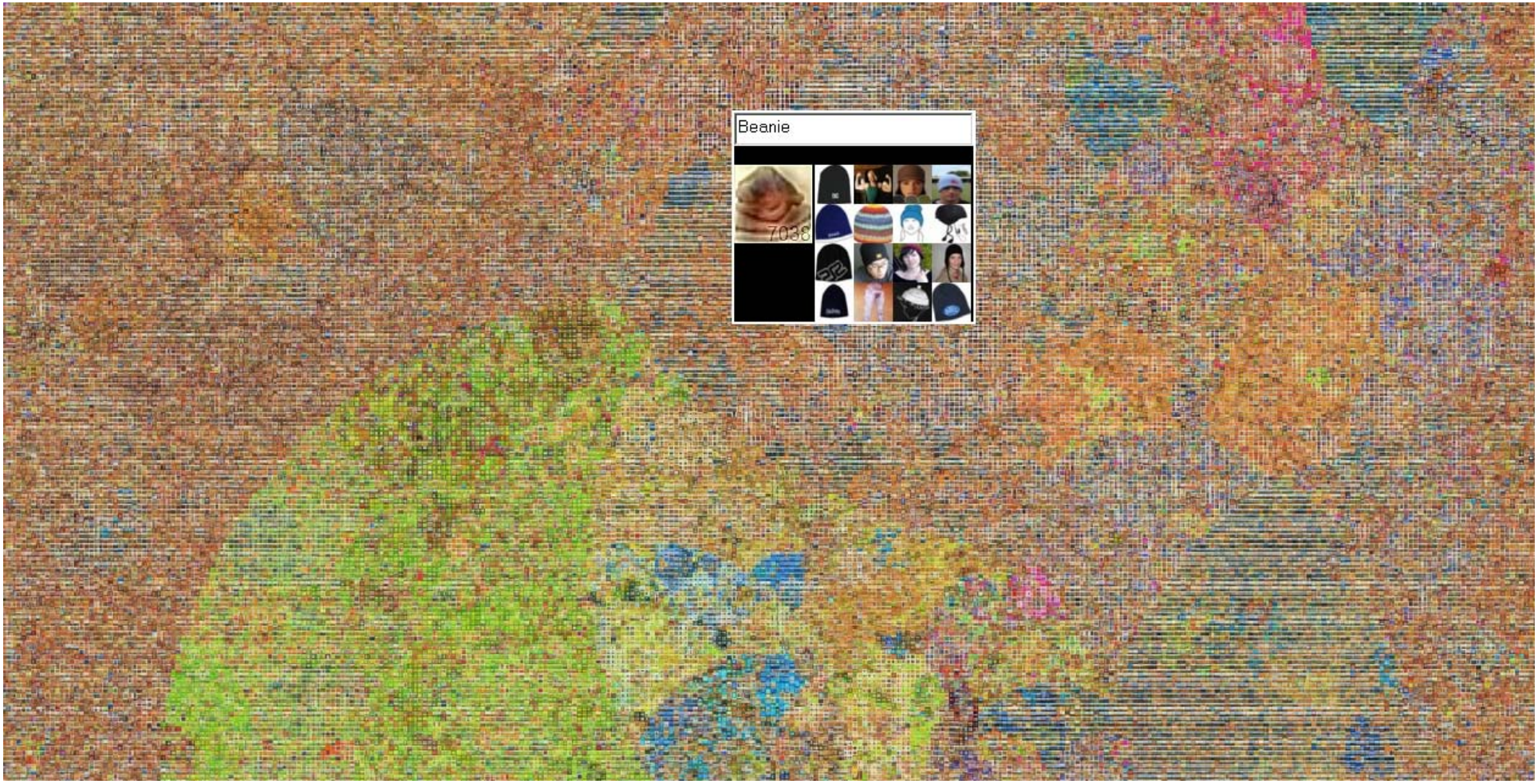
d) Wordnet voted branches

Wordnet Voting

- Overcomes differences in level of semantic labeling:
  - e.g. “person” & “sir arthur conan doyle”
- Totally incorrect labels form hopefully uniform background noise
- Assumes semantic and visual consistency are closely related

Semantic vs Visual Hierarchy

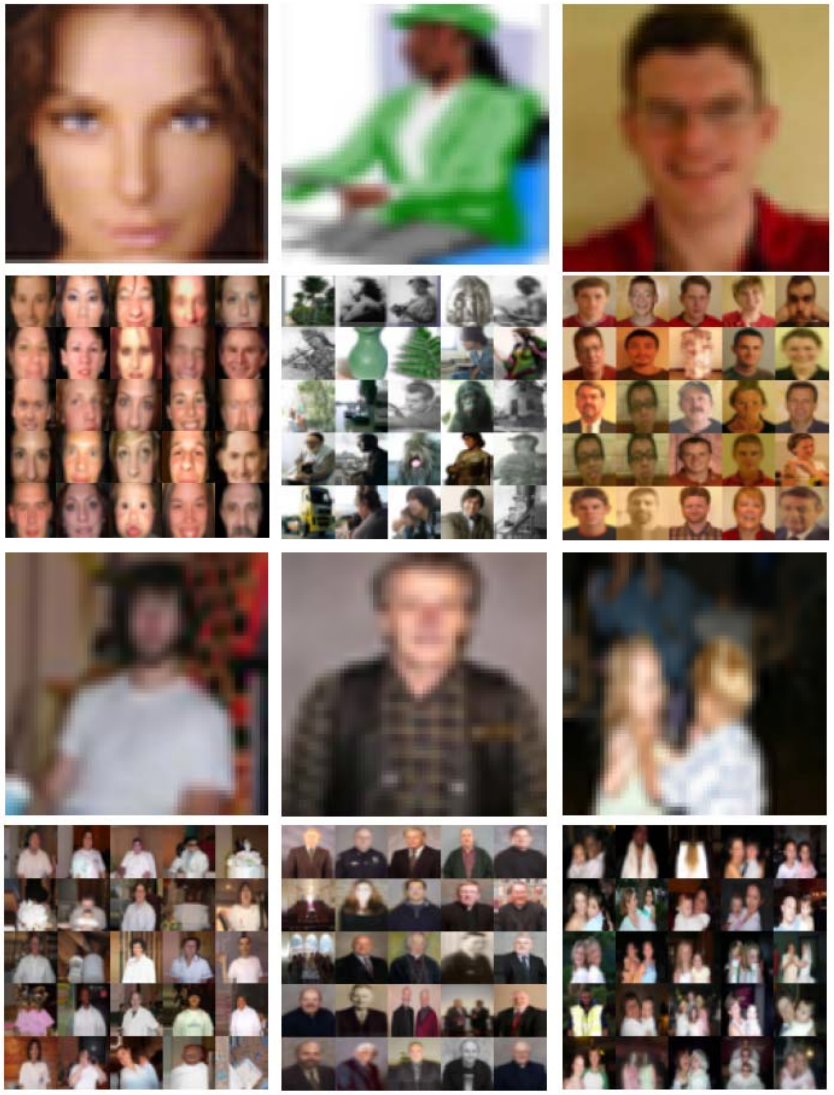
Interactive version at http://people.csail.mit.edu/torralba/tinyimages



Recognition Experiments

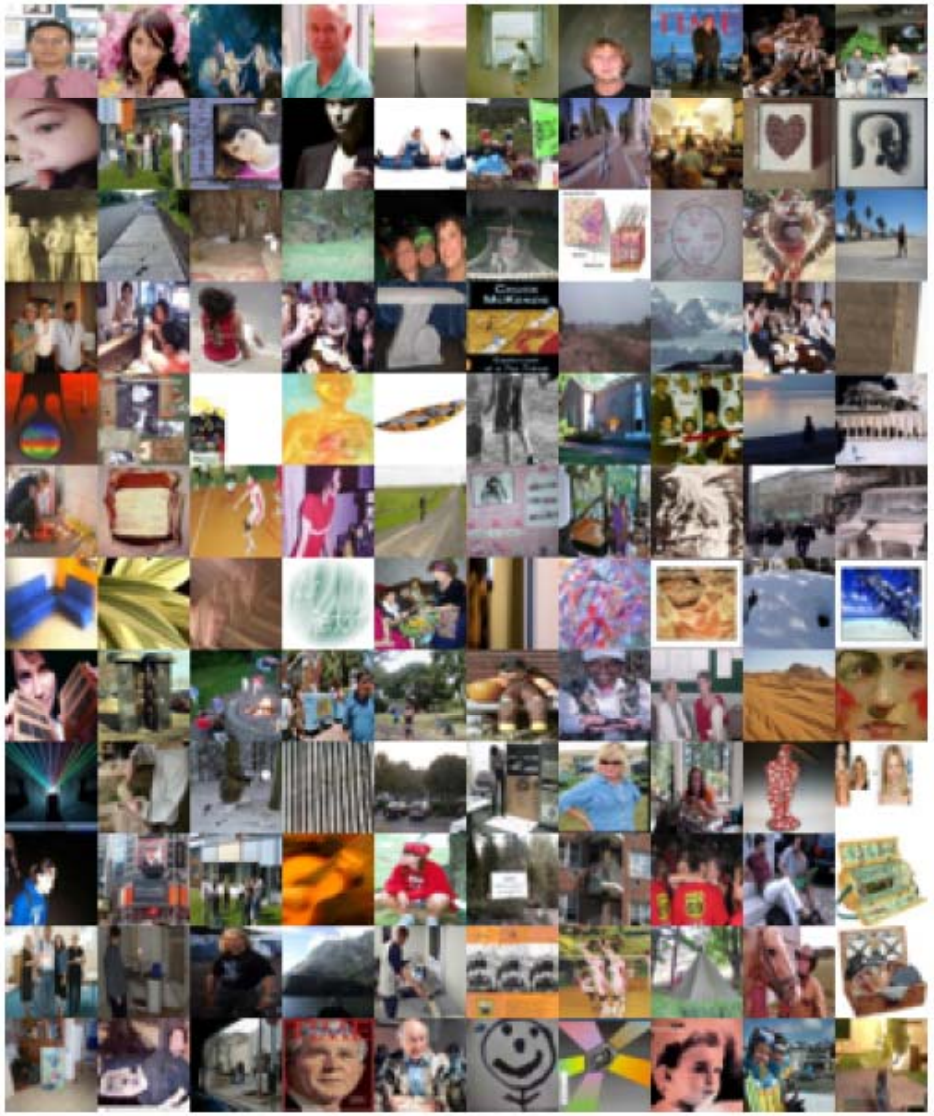
Person Recognition

- 23% of all images in dataset contain people
- Wide range of poses: not just frontal faces



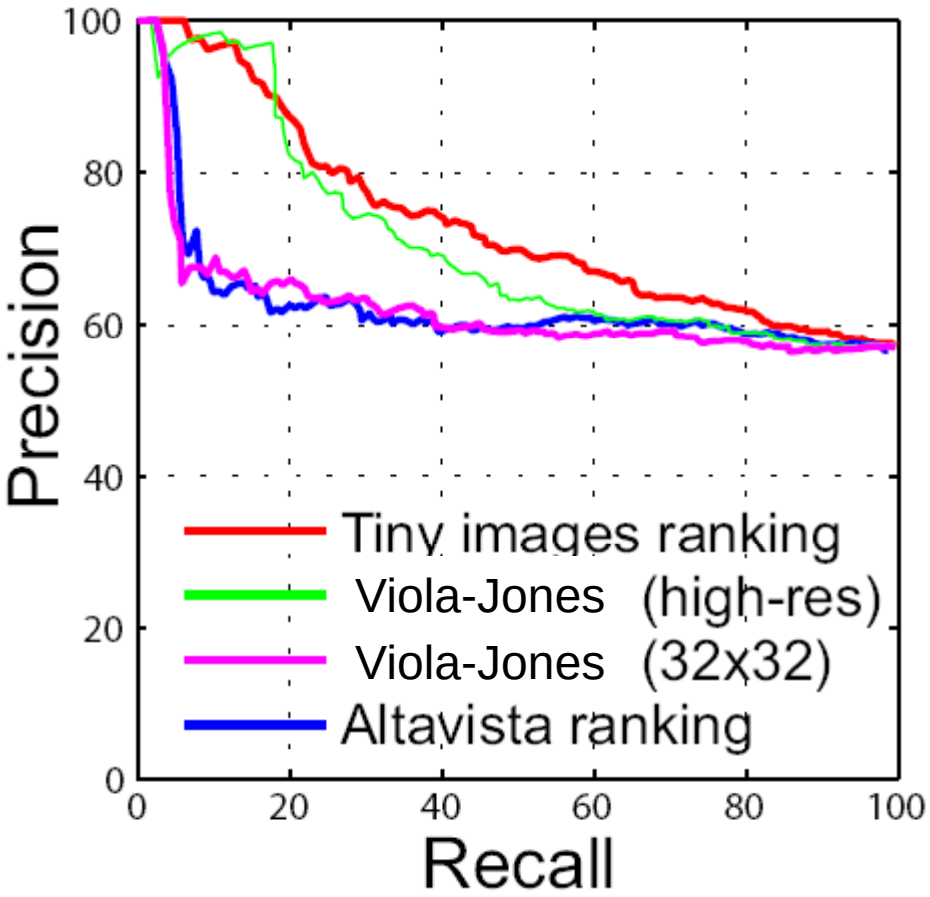
Person Recognition – Test Set

- 1016 images from Altavista using "person" query
- High res and 32x32 available
- Disjoint from 79 million tiny images



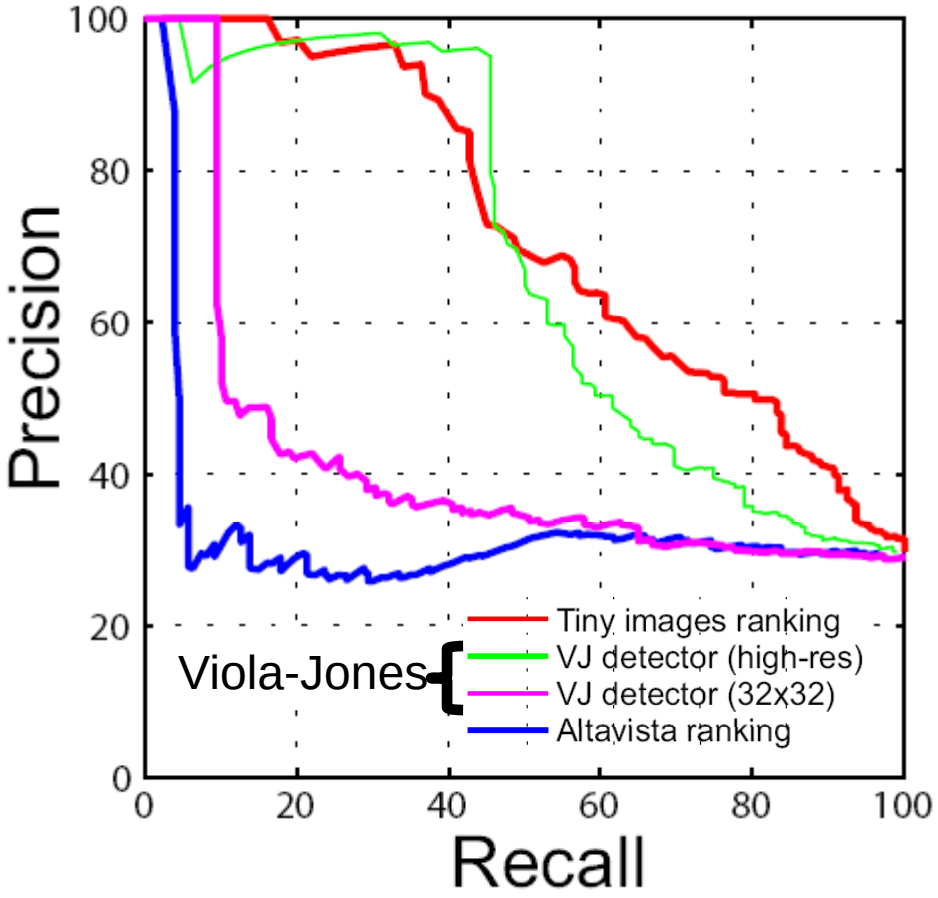
Person Recognition

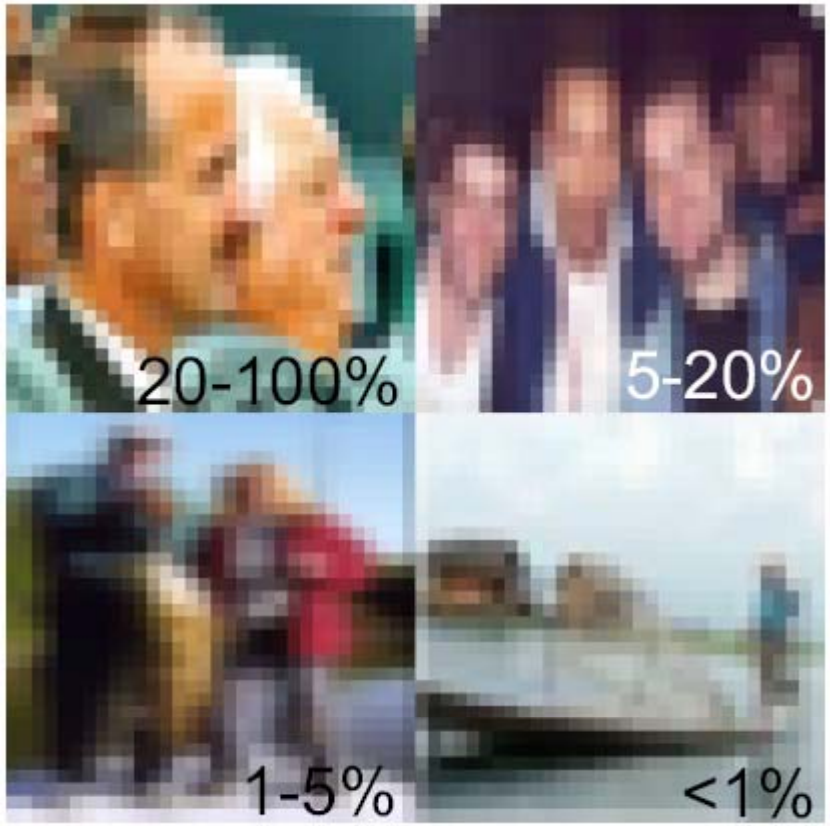
- Task: person in image or not?



Person Recognition

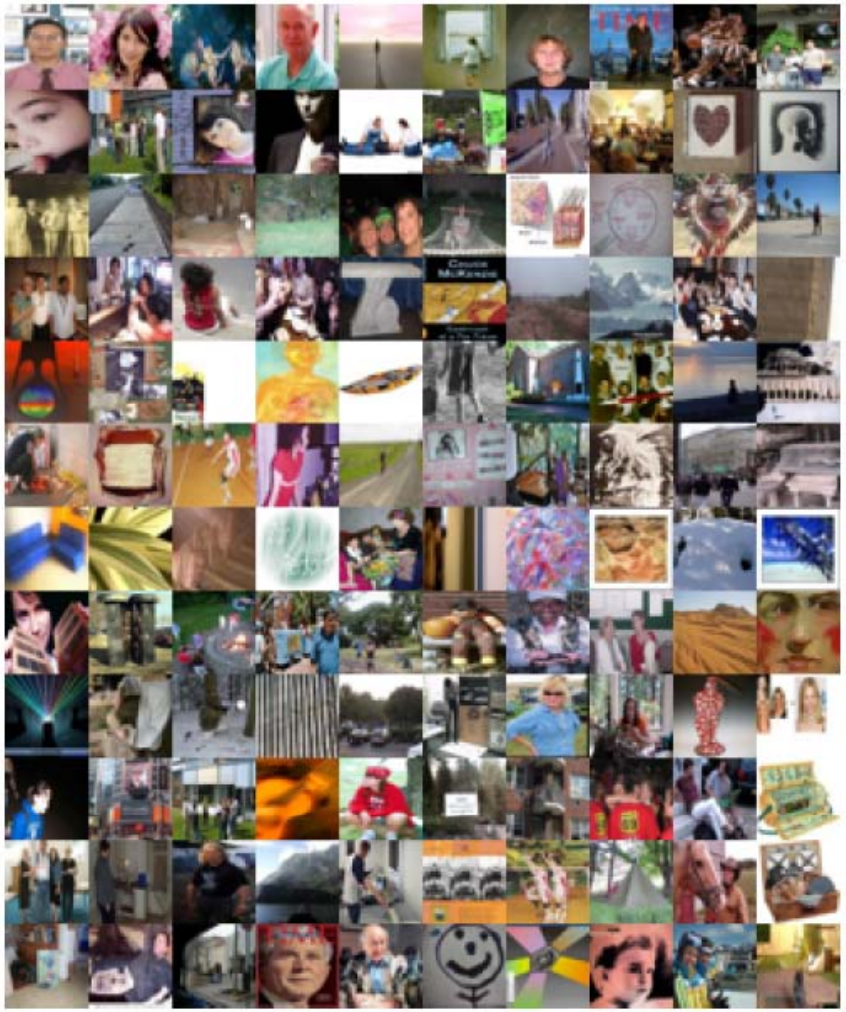
- Subset where face >20% of image



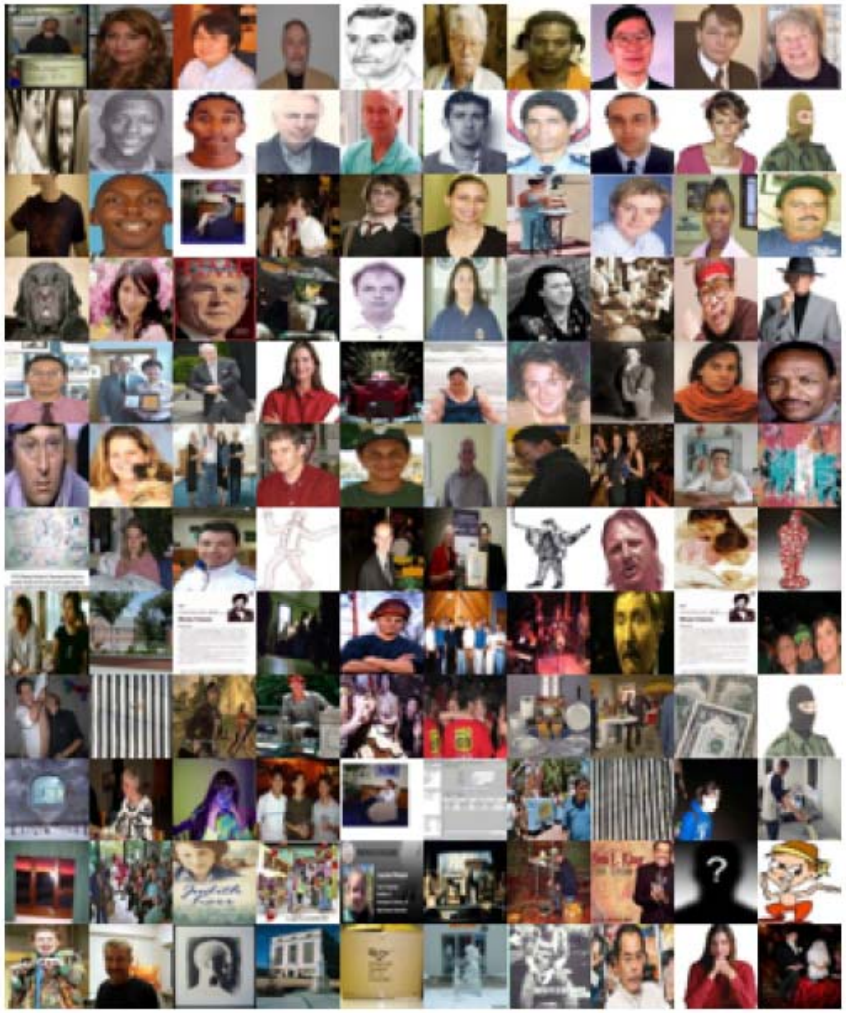


Re-ranked Altavista Images

Original

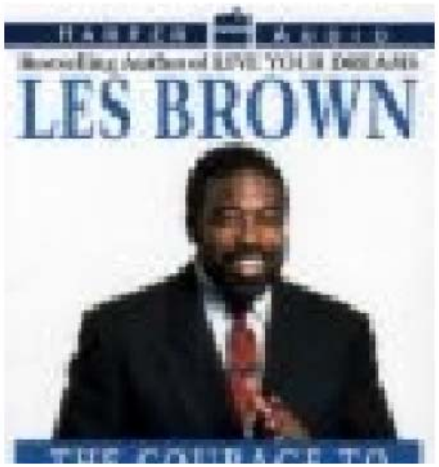


Re-ranked



Person Localization

High-res Query



a)

Ncuts Segmentation

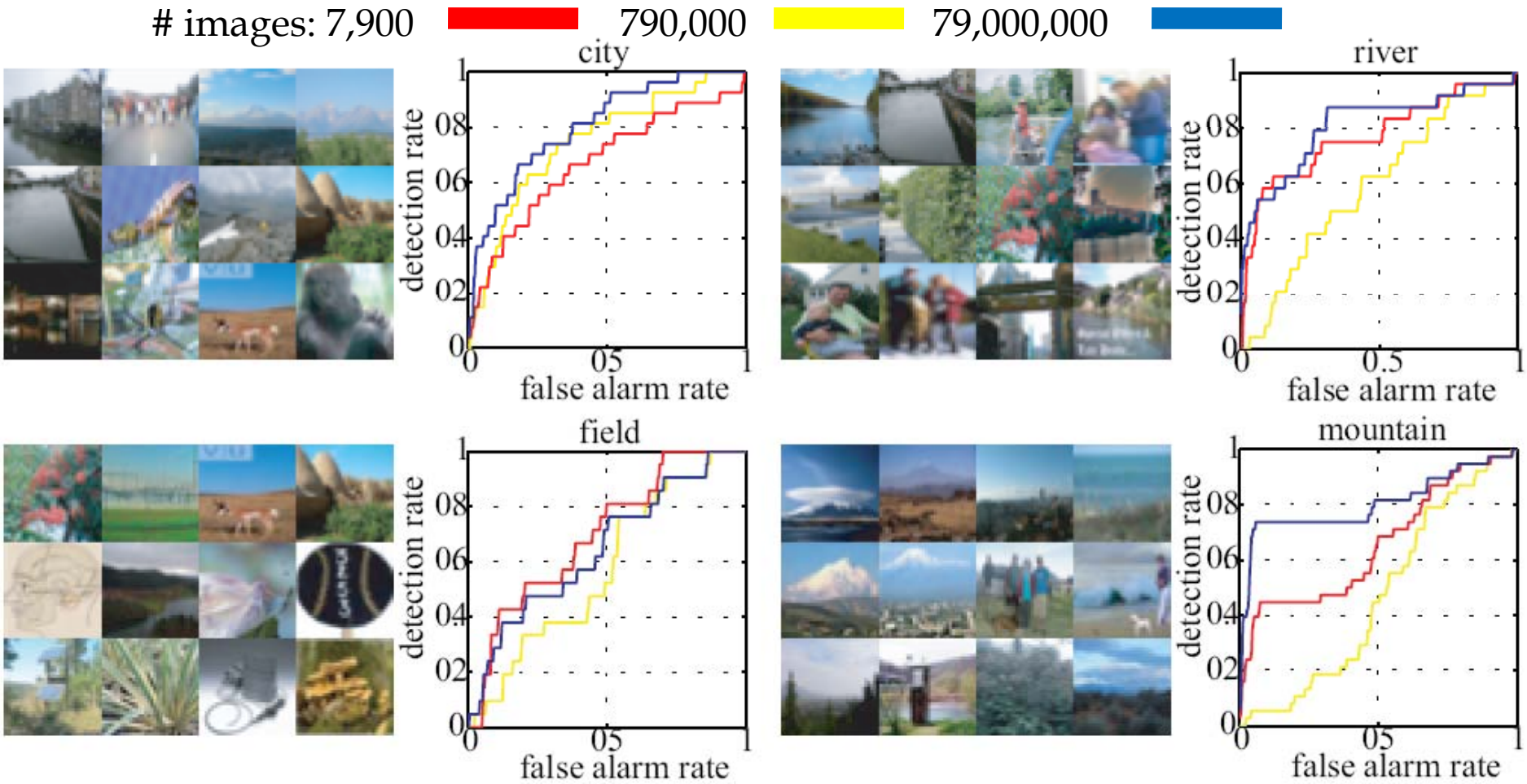
Gives putative crops

b)

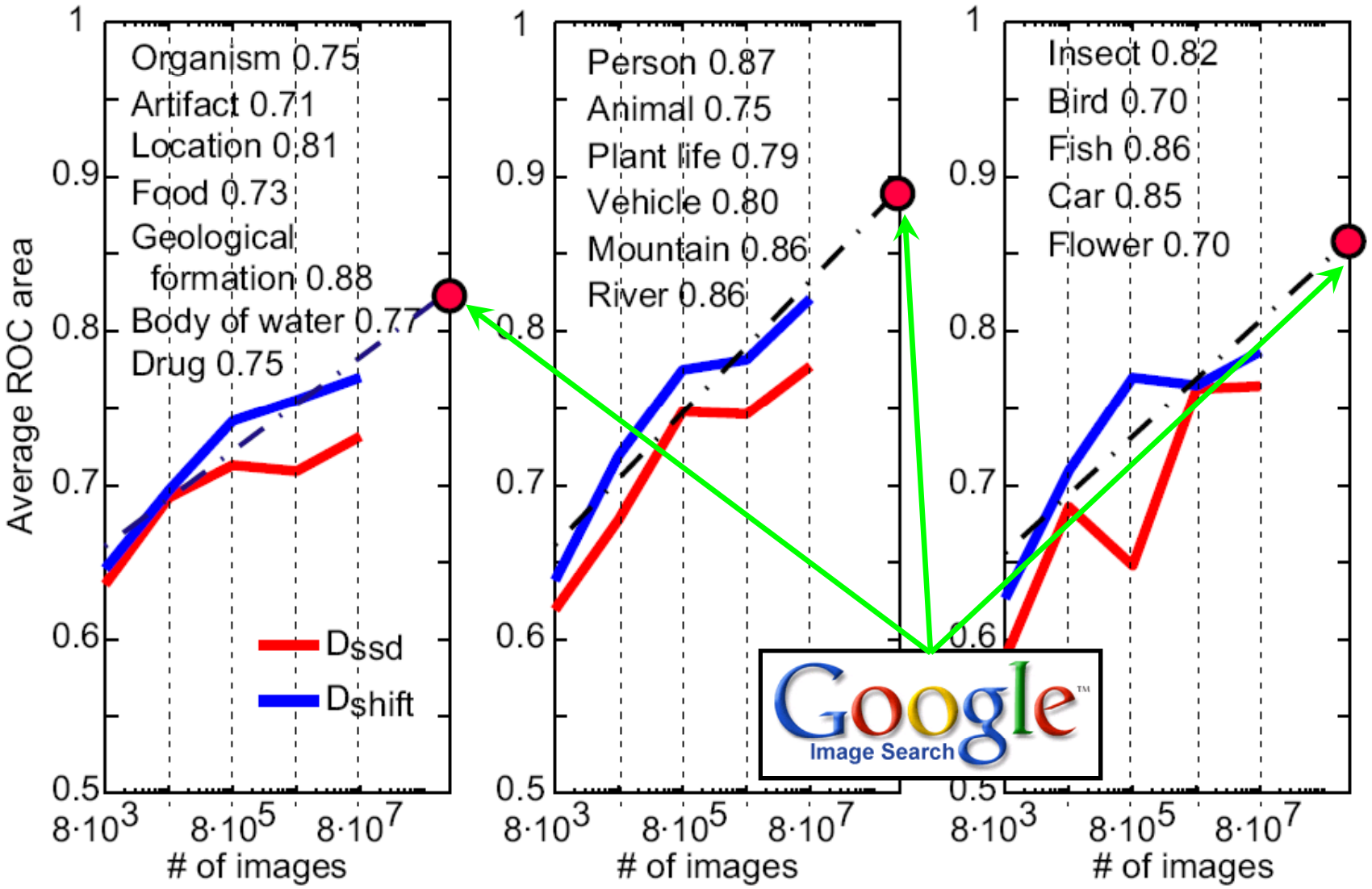
Person Localization Examples

Scene Classification

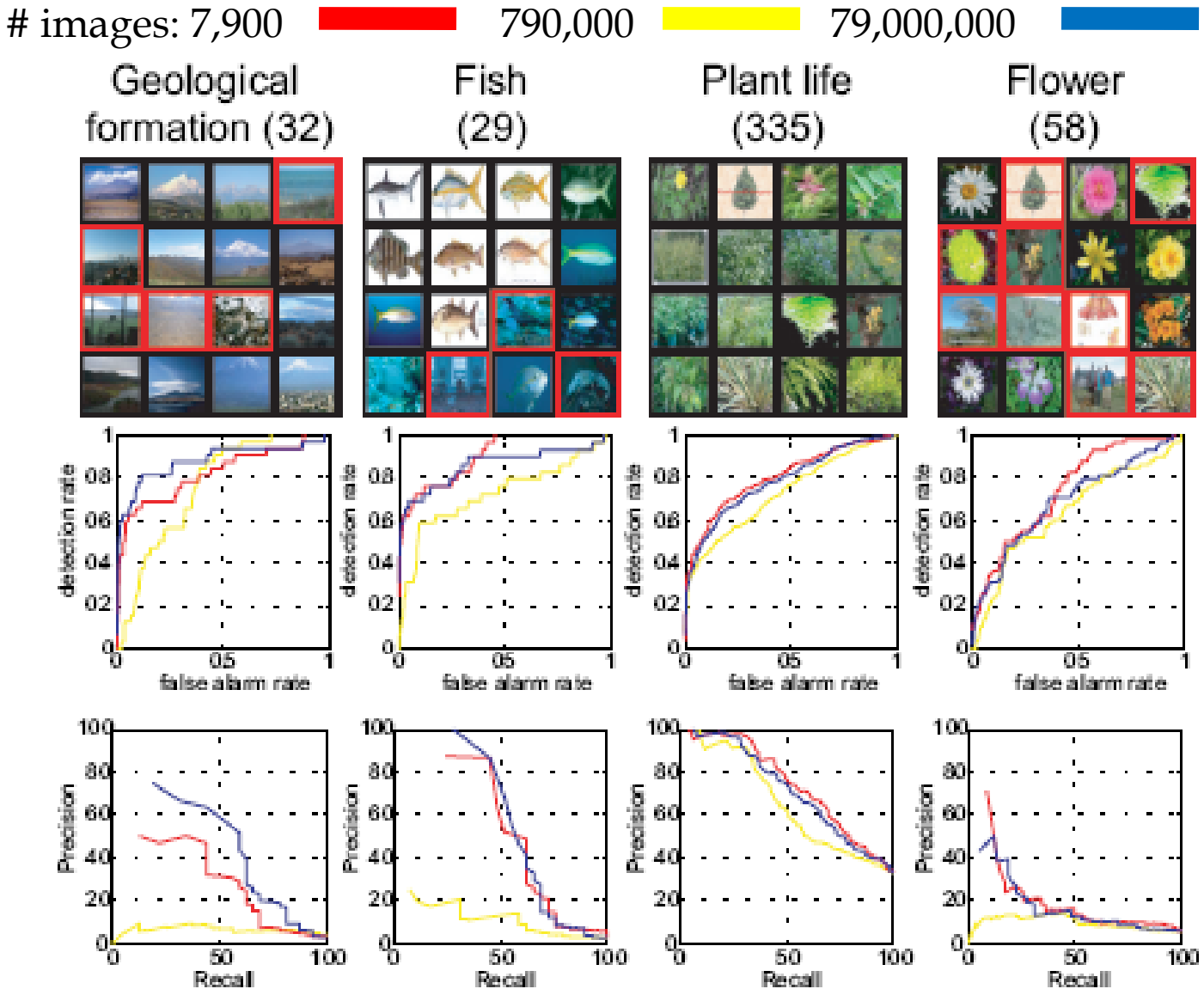
- Test set: 1125 images randomly drawn from 79 million.
- Task: {scene} vs all other classes



Object Classification



Object Classification



Other Applications

Automatic Colorization

Grayscale input
High resolution

Automatic Colorization

Grayscale input
High resolution

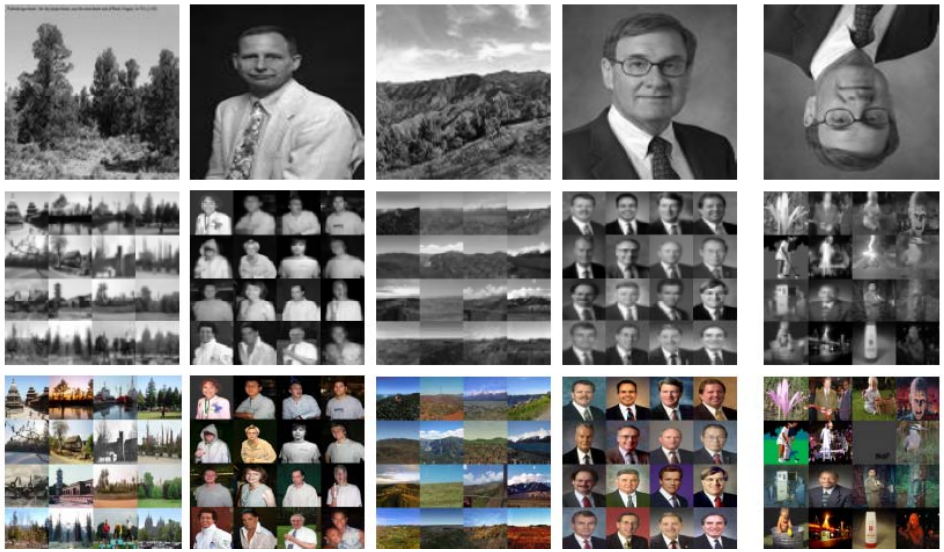
Grayscale
32x32 siblings

Automatic Colorization

Grayscale input
High resolution

Grayscale
32x32 siblings

Color siblings
high resolution



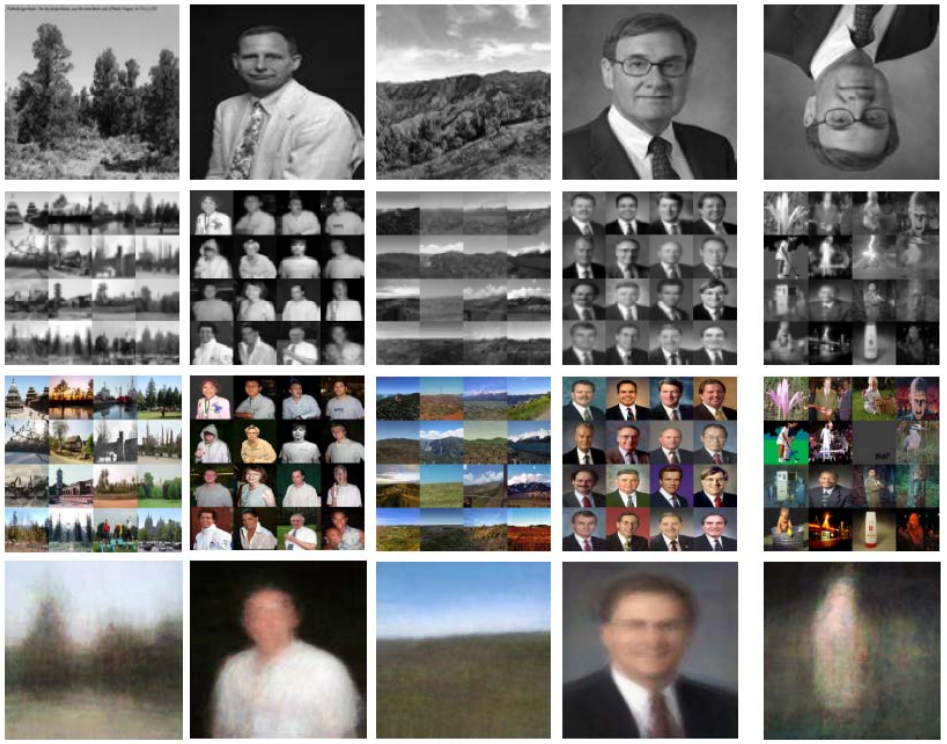
Automatic Colorization

Grayscale input
High resolution

Grayscale
32x32 siblings

Color siblings
high resolution

Average of
color siblings



Automatic Colorization

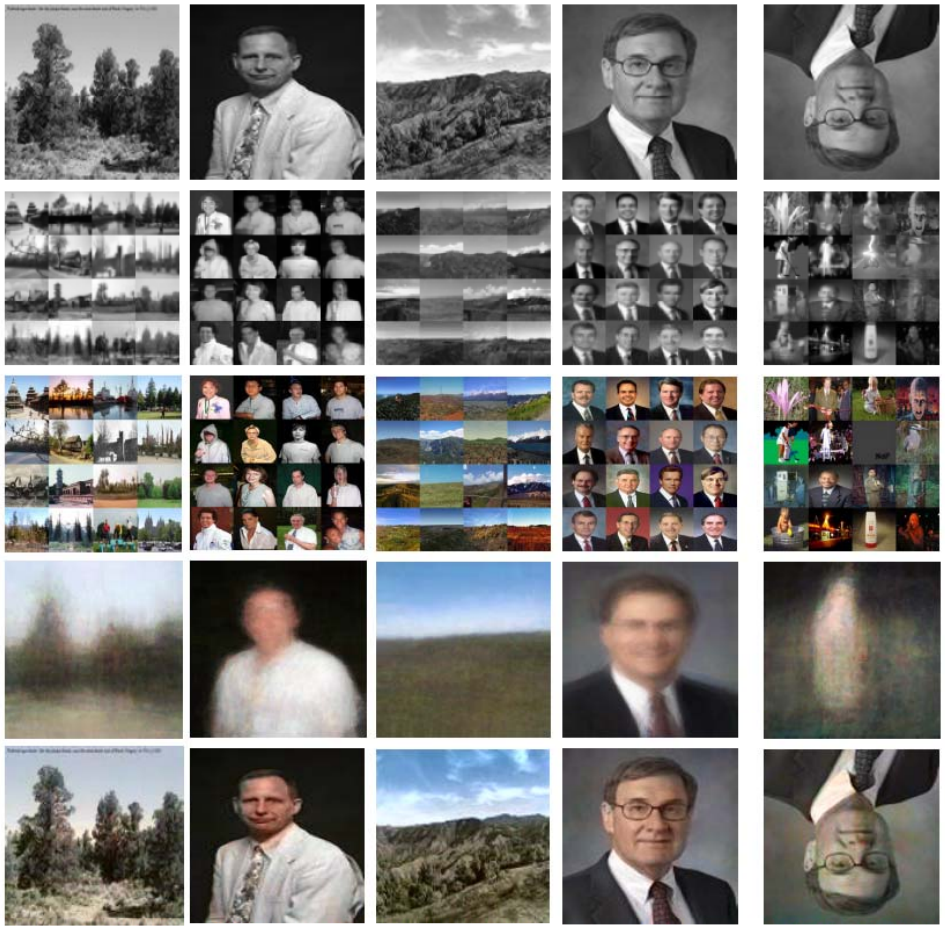
Grayscale input
High resolution

Grayscale
32x32 siblings

Color siblings
high resolution

Average of
color siblings

Colorization of input
using average



Automatic Colorization

Grayscale input
High resolution

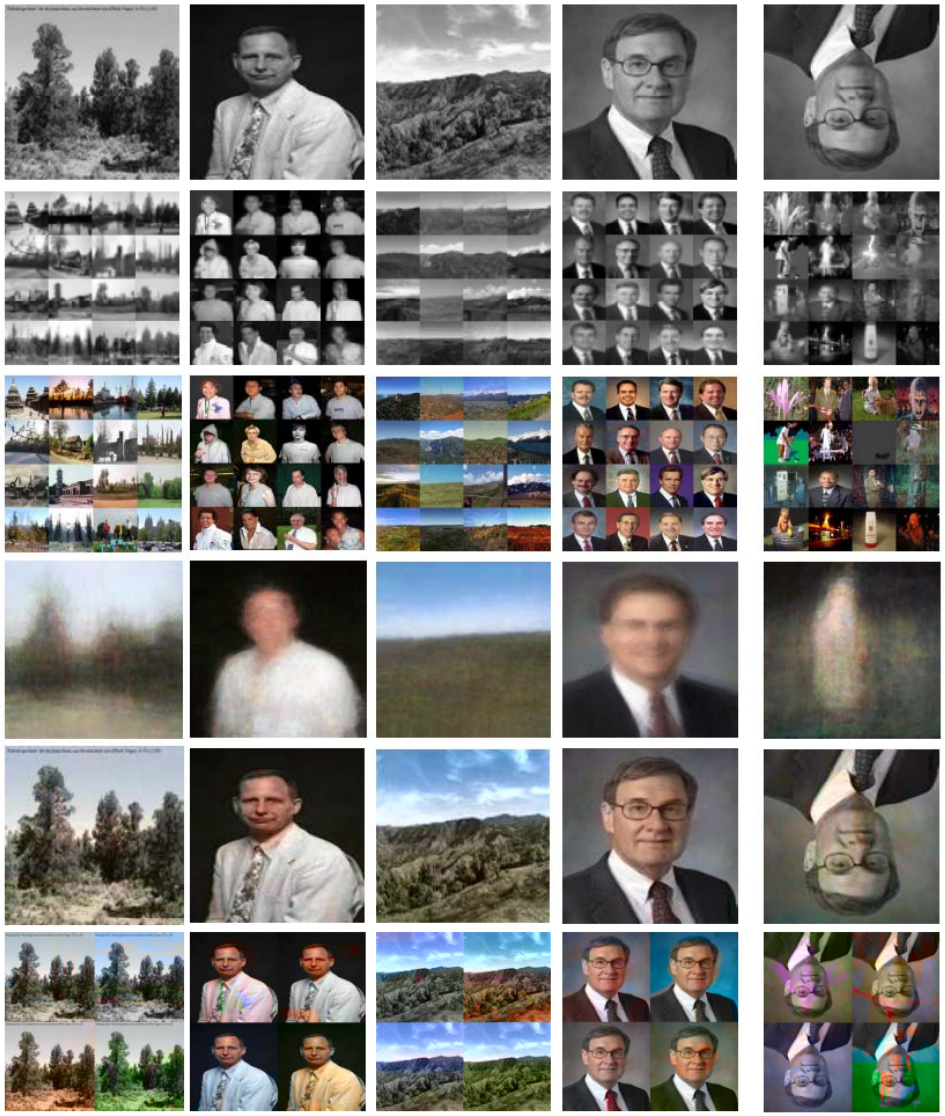
Grayscale
32x32 siblings

Color siblings
high resolution

Average of
color siblings

Colorization of input
using average

Colorization of input
using specific siblings



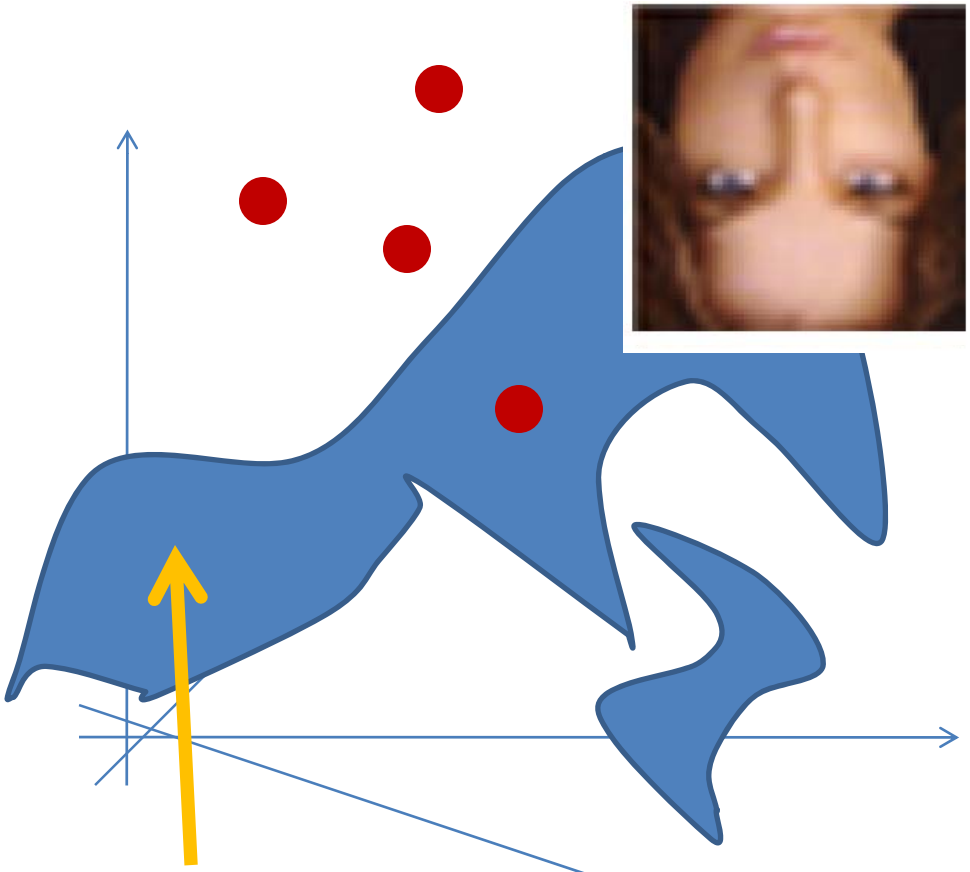
Automatic Colorization Result

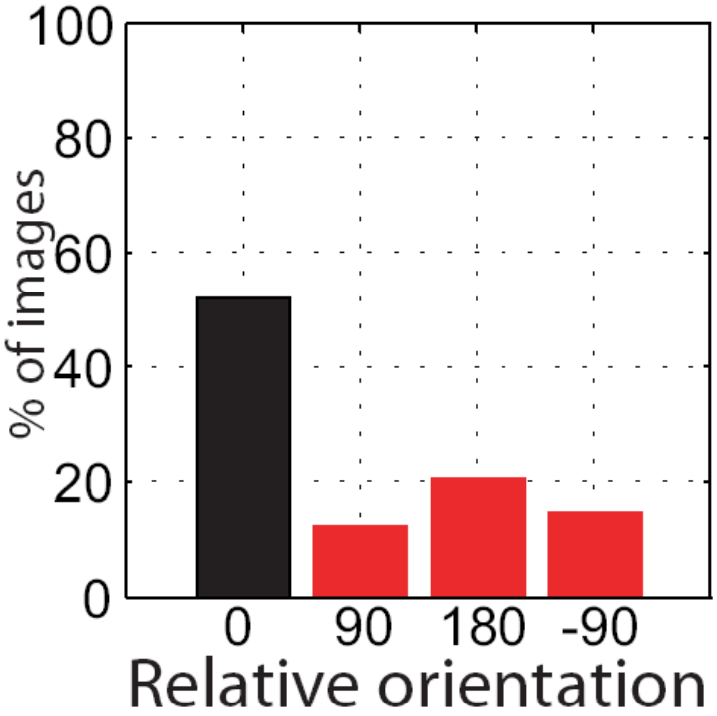
Grayscale input High resolution

Colorization of input using average

Automatic Orientation

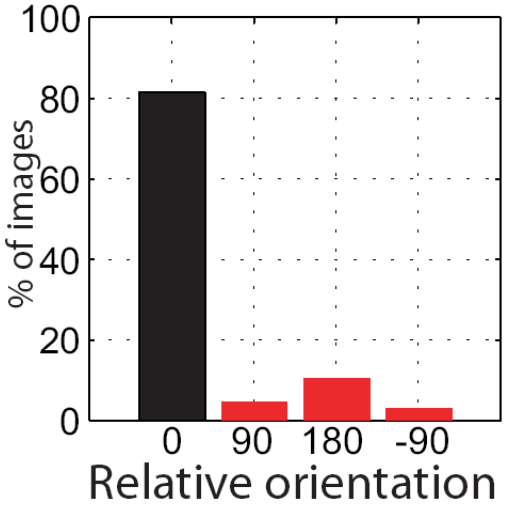
- Look at mean distance to neighbors



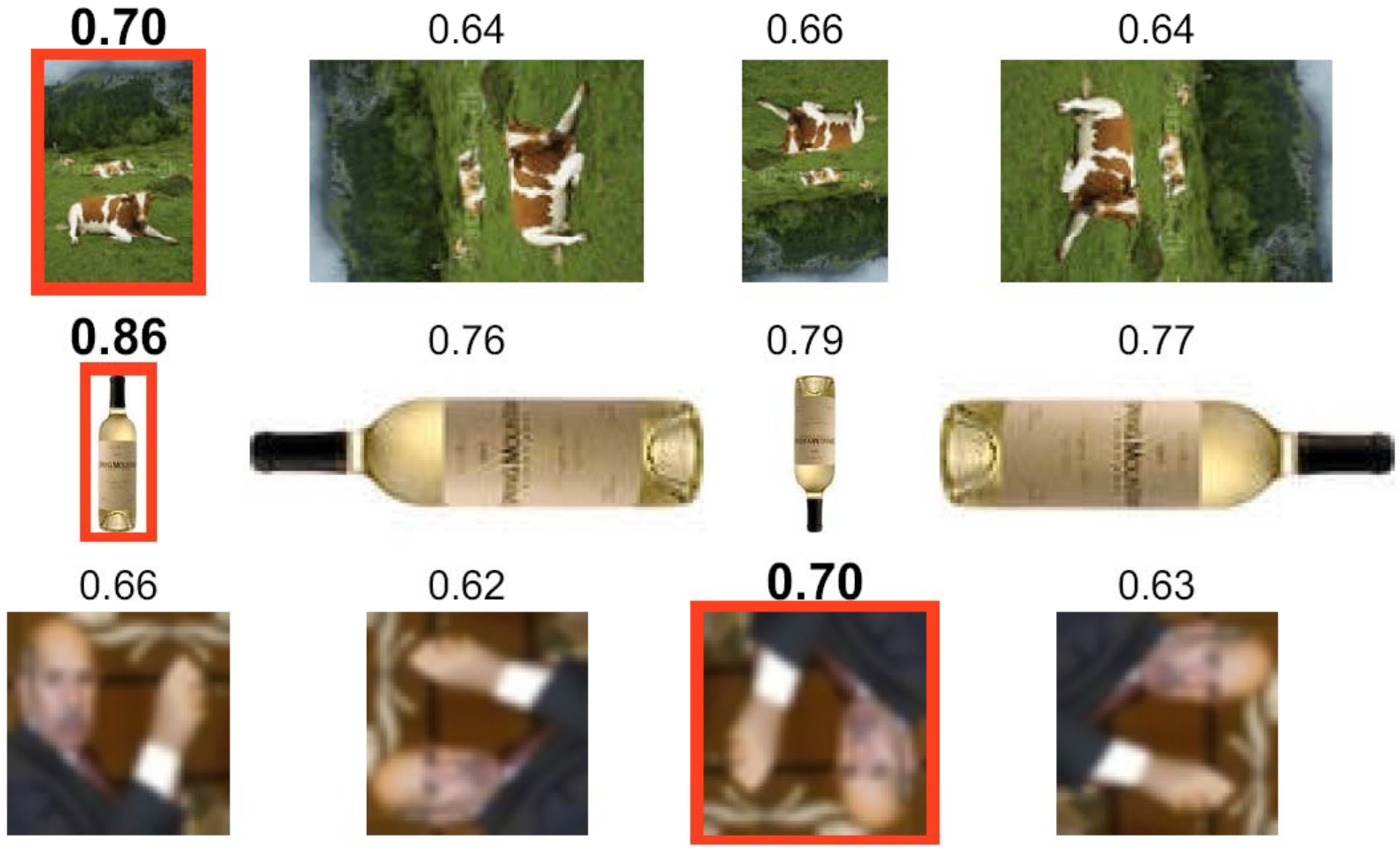


Automatic Orientation

- Many images have ambiguous orientation
- Look at top 25% by confidence:
- Examples of high and low confidence images:



Automatic Orientation Examples

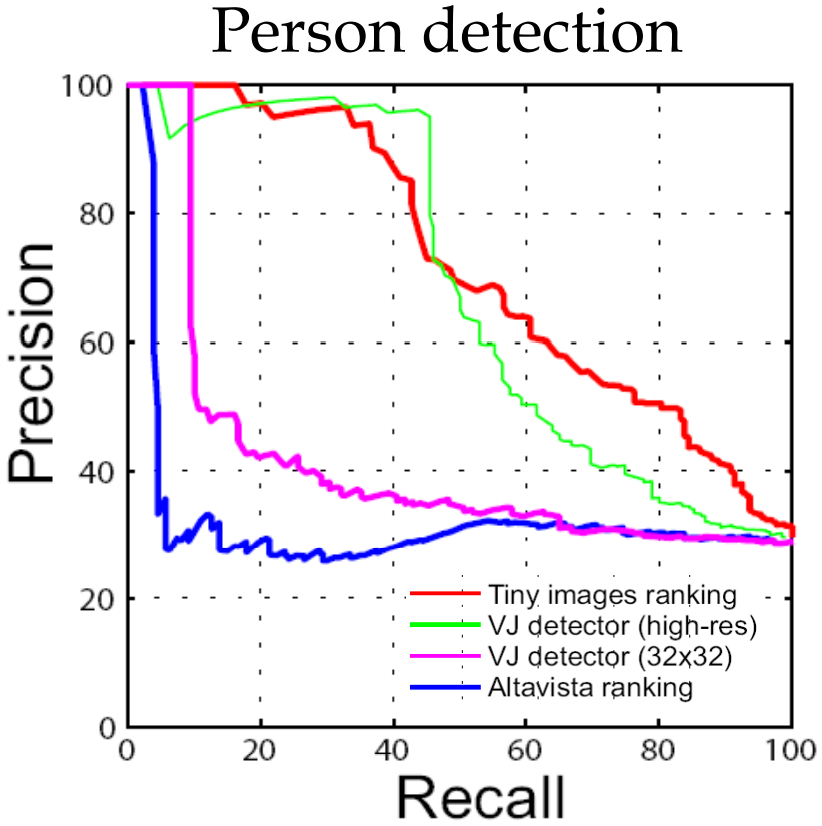


Related Work

- Hayes & Efros, Scene Completion using Millions of photographs, SIGGRAPH 2007.
- Nister & Stewenius. Scalable recognition with a vocabulary tree, CVPR 2006.
- Hoogs & Collins. Object boundary detection in images using a semantic ontology. In *AAAI, 2006*.
- Barnard et al., Matching words and pictures. JMLR, 2003.
- Shakhnarovich et al. Fast pose estimation with parameter sensitive hashing, ICCV 2003

Conclusions

- Can get good results simple algorithms & lots of data
- Issues with Internet images: labeling noise & image baises.
- Bring in learning: Distance metrics, text & images



Webpage: http://people.csail.mit.edu/torralba/tinyimages