How to stop worrying and learn to love Nearest Neighbors

Alexei (Alyosha) Efros

UC Berkeley

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

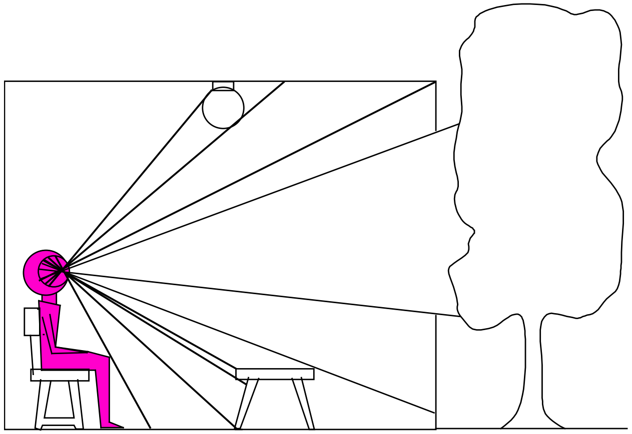
The Tyranny of "Elegant" ideas

"For every complex problem there is an answer that is clear, simple, and wrong."

-- H. L. Mencken

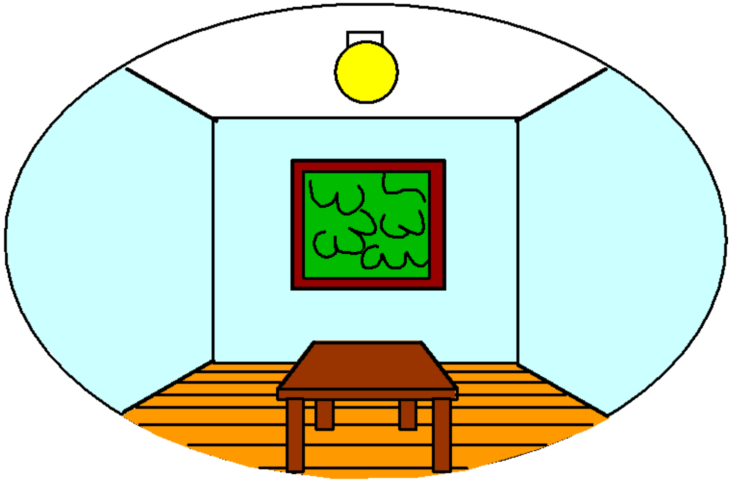
How do humans see 3D?

3D world

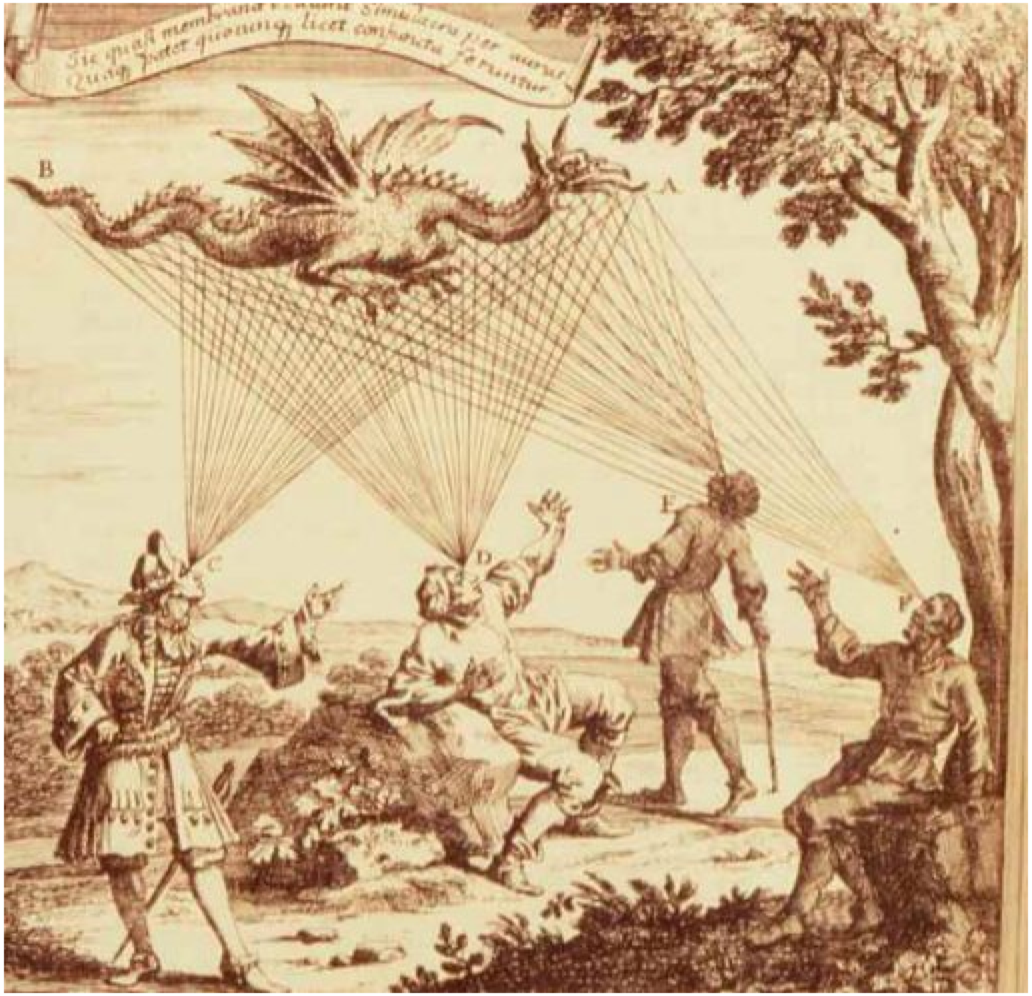


Point of observation

2D image



Emission Theory of Vision



Eyes send out “feeling rays” into the world

Supported by:

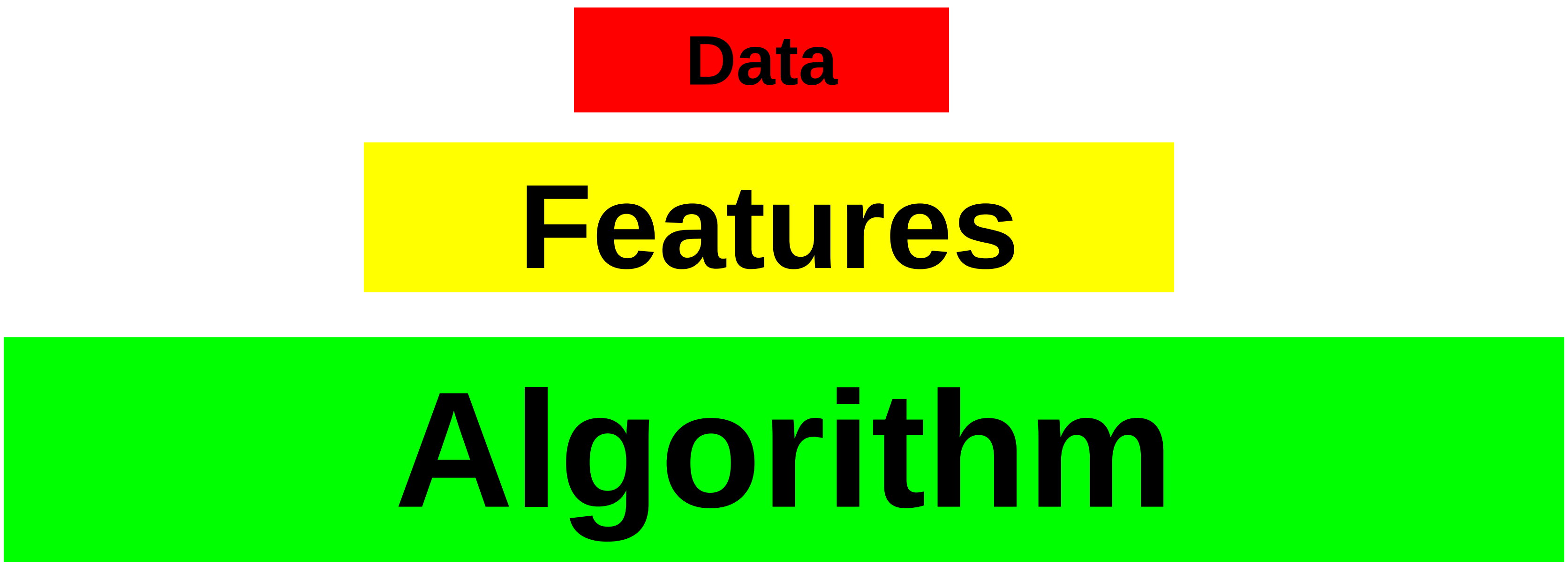
- Empedocles
- Plato
- Euclid (kinda)
- Ptolemy
- …
- 50% of US college students*

*http://www.ncbi.nlm.nih.gov/pubmed/12094435?dopt=Abstract

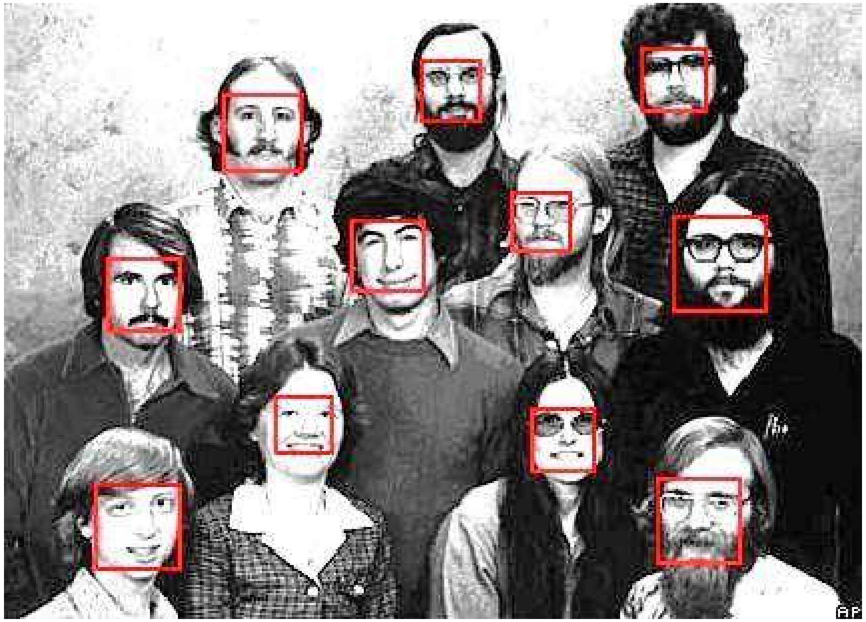
Our Scientific Narcissism

All things being equal, we prefer to credit our own cleverness

We prefer algorithms to data



Face Detection: Big Success Story



- Rowley, Baluja, and Kanade, 1998
    - features: **pixels**, classifier**: neural network**
- Schniderman & Kanade, 1999
    - features: **pairs of wavelet coeff**., classifier**: naïve Bayes**
- Viola & Jones, 2001
    - features: **haar**, classifier: **boosted cascade**

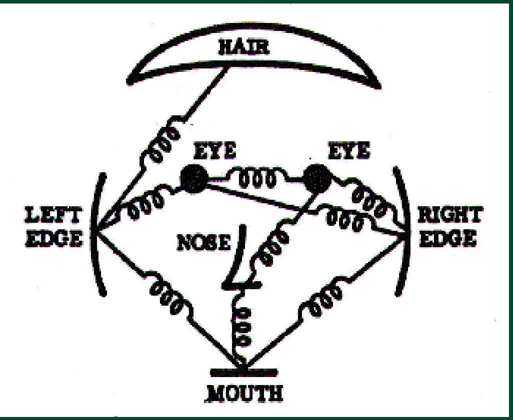
Learning Spectrum

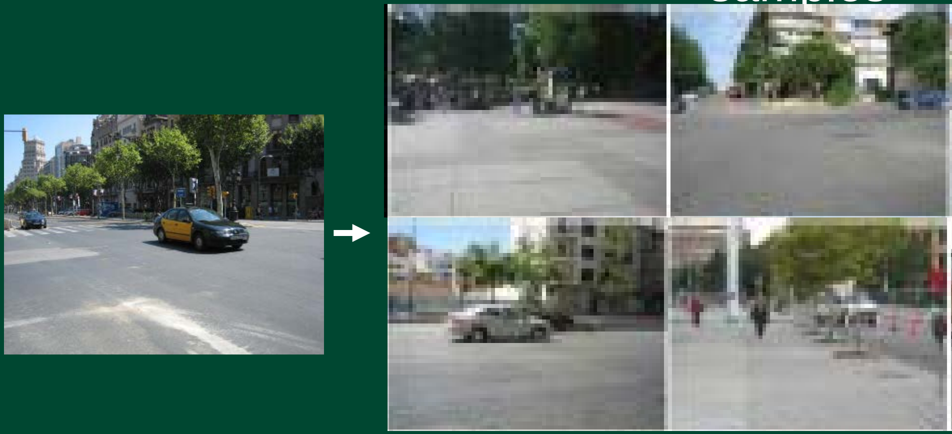
Extrapolation problem
Generalization

Interpolation problem
Correspondence

1 | 10 | 10^2 | 10^3 | 10^4 | 10^5 | 10^6 | ∞

Number of training samples





Slide by Antonio Torralba

“Unreasonable Effectiveness of Data”

[Halevy, Norvig, Pereira 2009]

- Parts of our world can be explained by elegant mathematics:
  – physics, chemistry, astronomy, etc.
- But much cannot:
  – psychology, genetics, economics, etc.

- Enter: The **<u>The Data</u>**
  – Great advances in several fields:
    - e.g. speech recognition, machine translation, vision

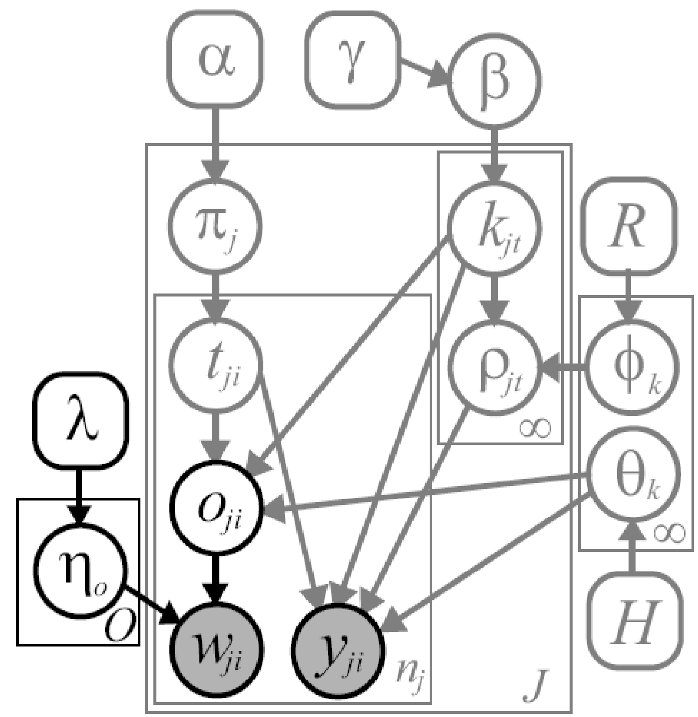
Overfitting to the world

- MNIST Digits
  - 10 digits *
  - ~1,000 variations = 10,000
- English words
  - ~100,000 words *
  - ~5 variations = 500,000
- Natural world
  - ~100,000 objects *
  - ~10,000 variations (pose, scale, lighting, intra-category)
  - **= 1,000,000,000 (1 billion)**
  - Not counting compositionality (will discuss later)

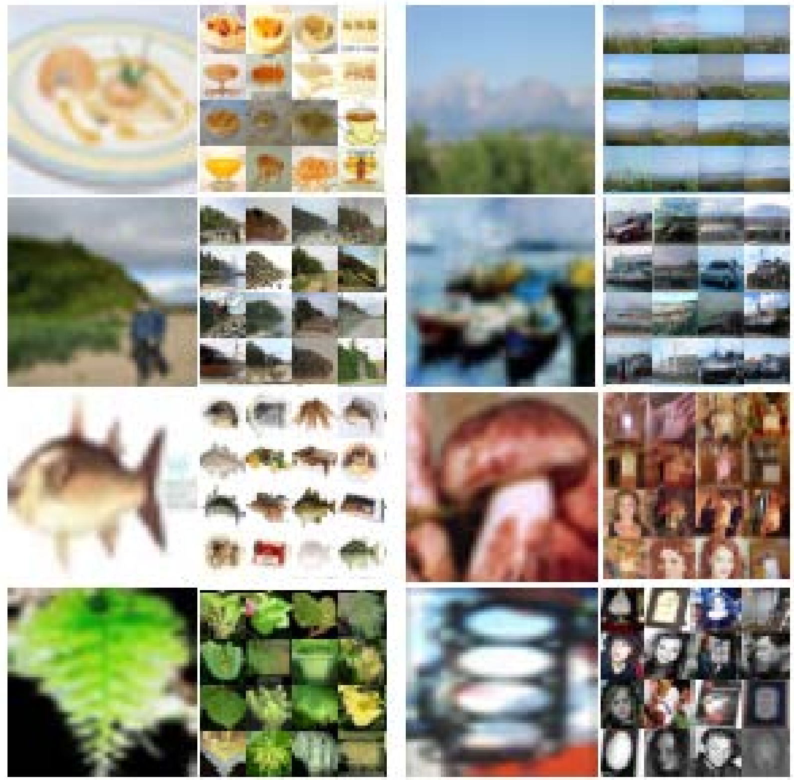
MNIST:
60,000 examples

Topic Models (Blei):
11M examples

Part 1: Nearest Neighbors aren't that bad!



Lots of Tiny Images

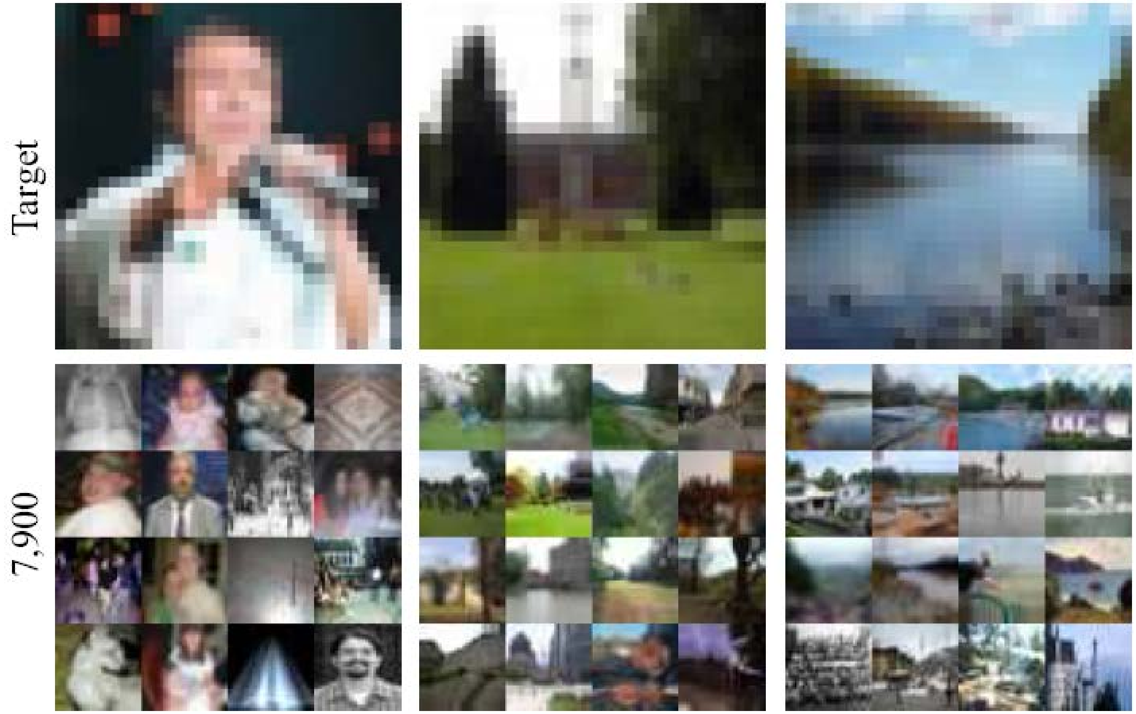


- 80 million tiny images: a large dataset for non-parametric object and scene recognition Antonio Torralba, Rob Fergus and William T. Freeman. PAMI 2008.

Lots

Of

Images

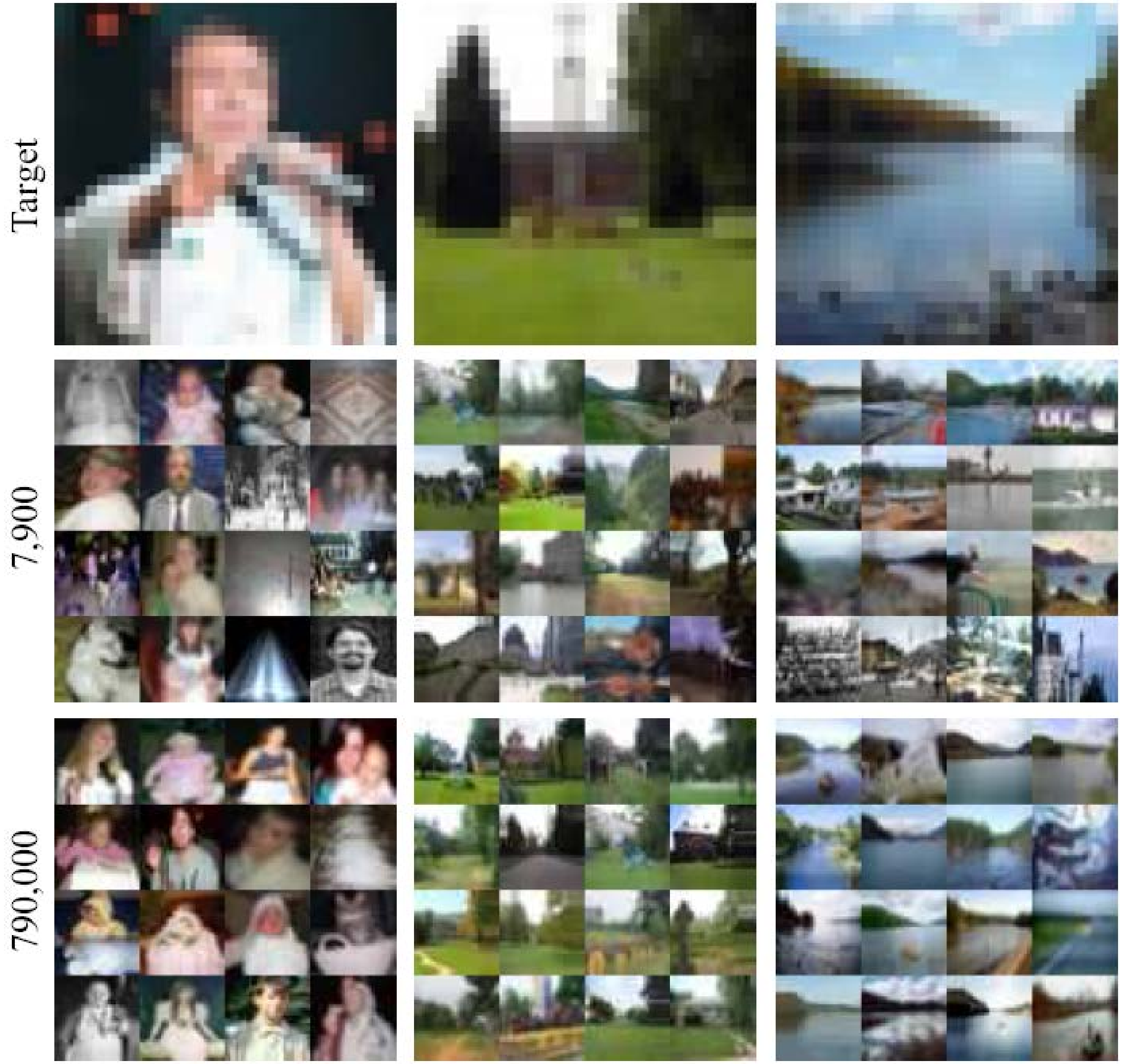


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images

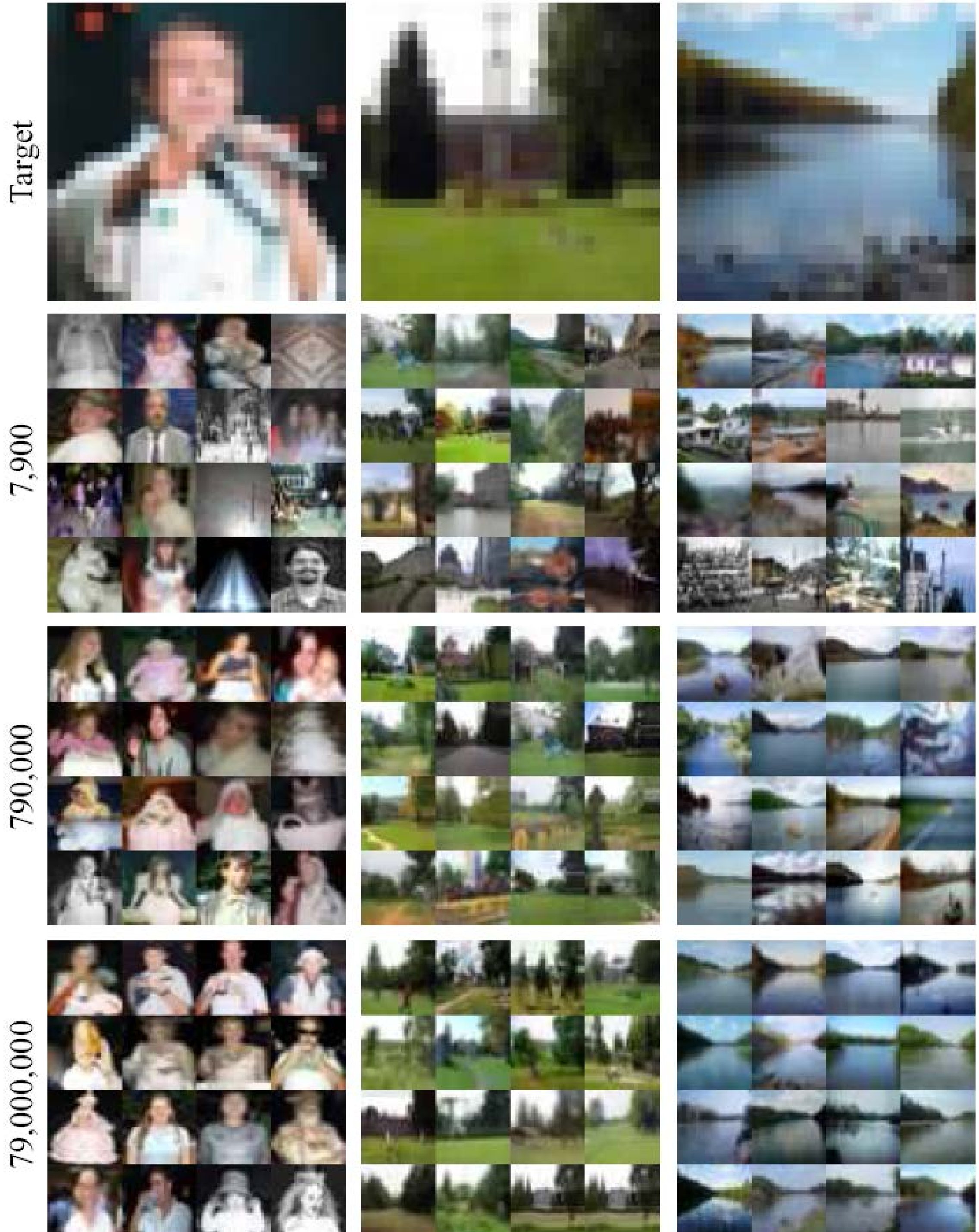


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

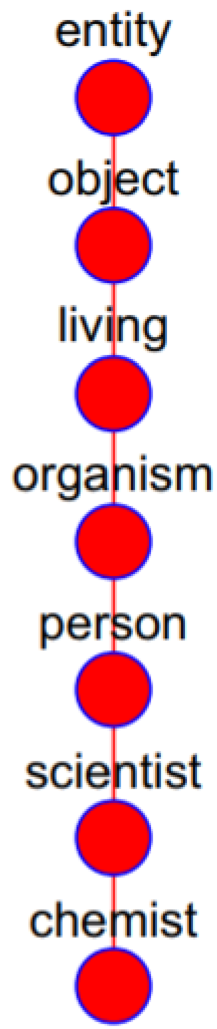
Of

Images

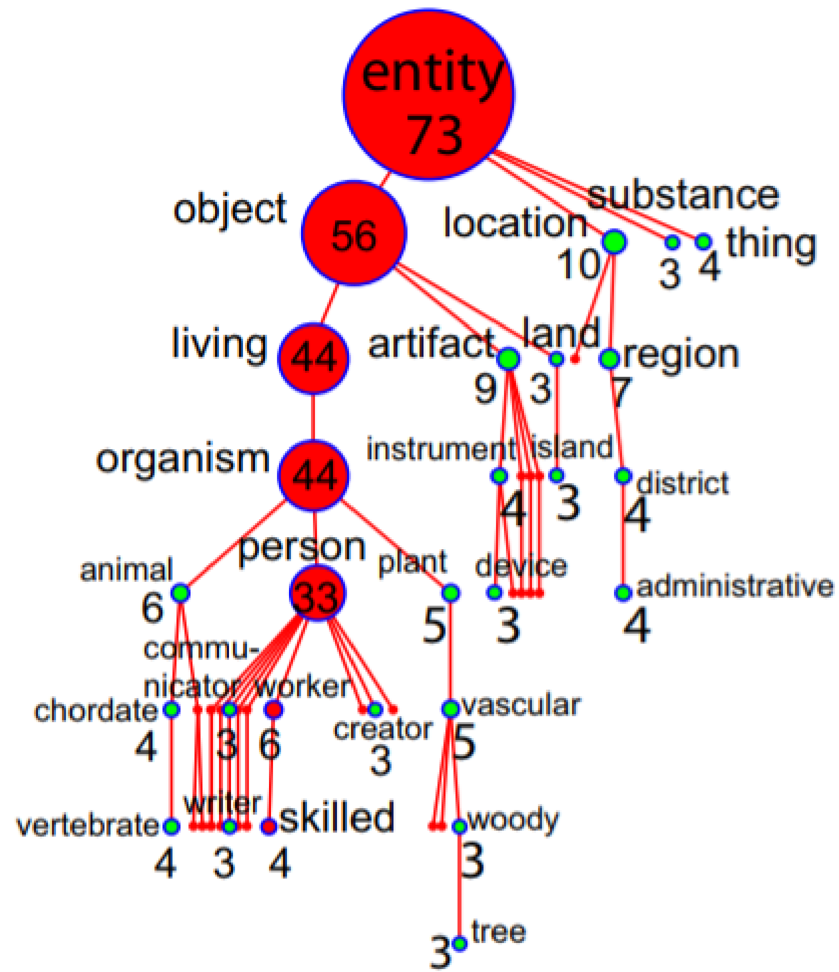


a) Input image

b) Neighbors



c) Ground truth

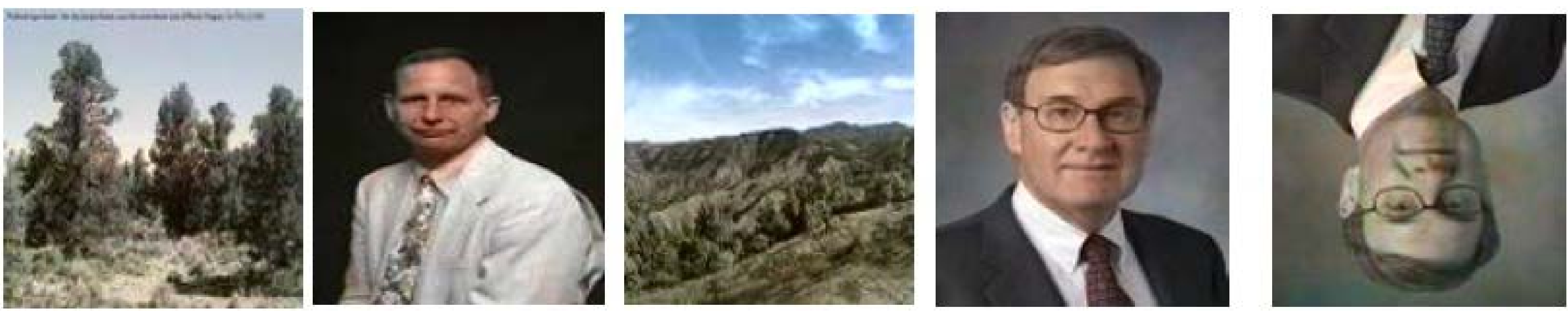


d) Wordnet voted branches

Automatic Colorization

Grayscale input High resolution

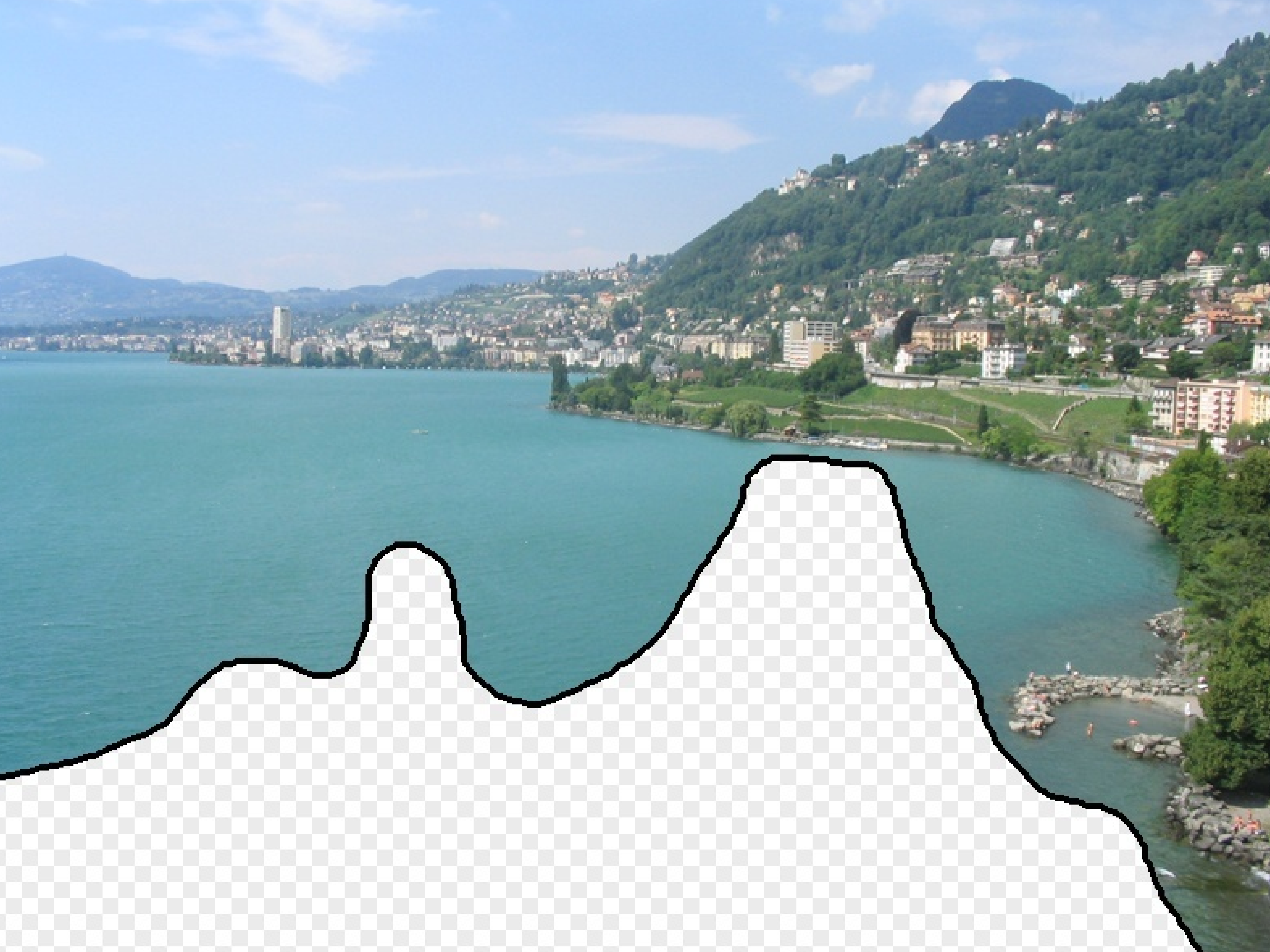
Colorization of input using average



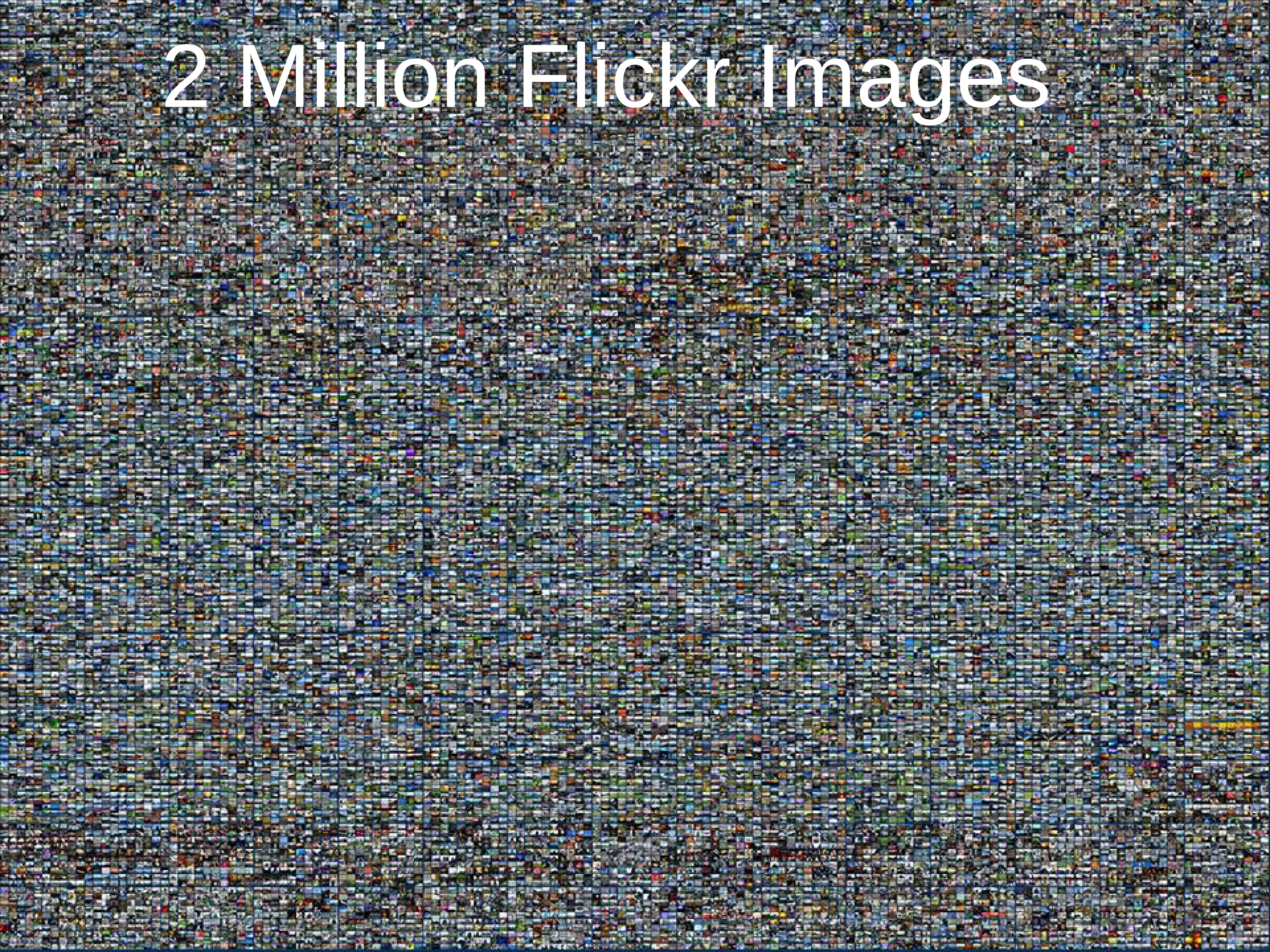
A. Torralba, R. Fergus, W.T.Freeman. 2008

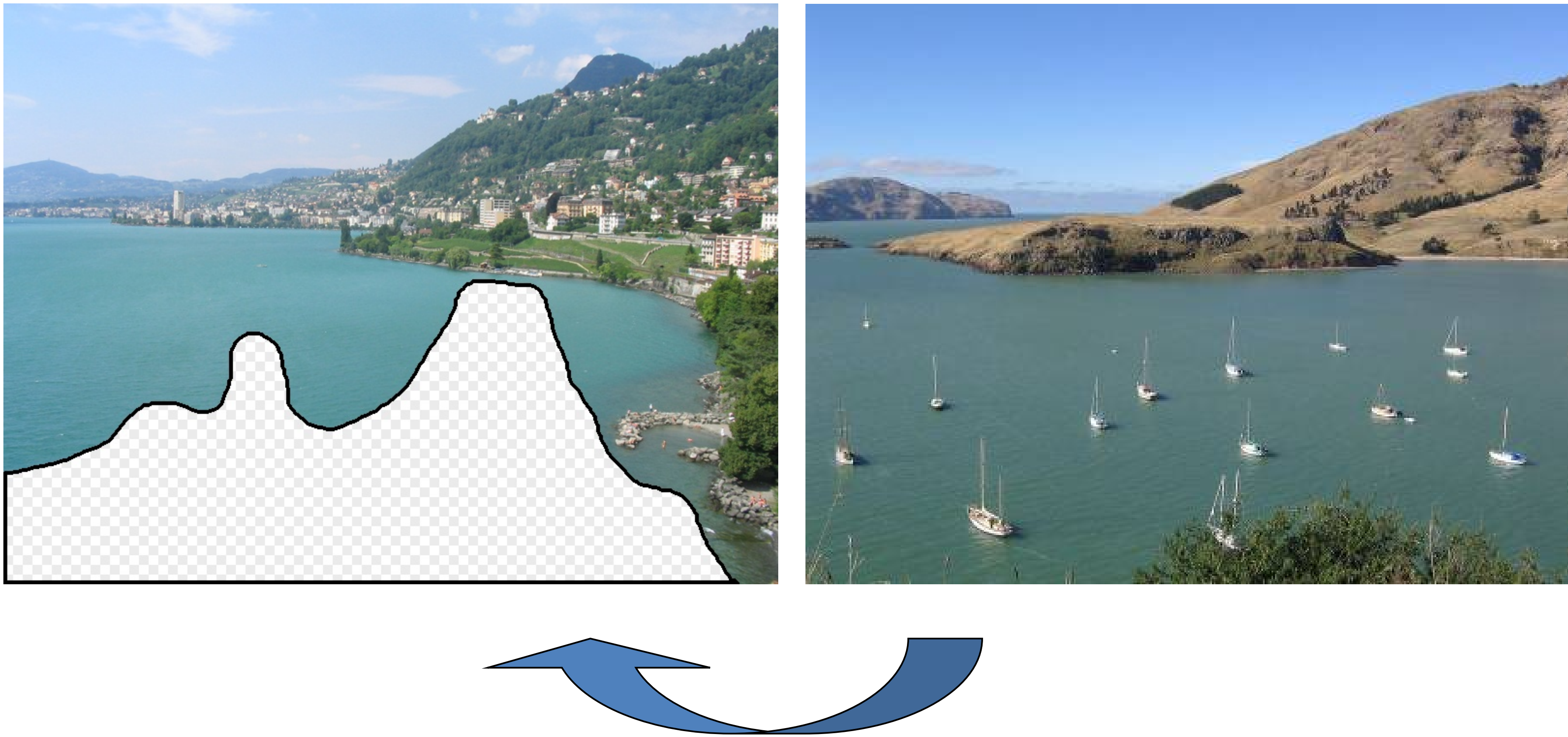
"Size Does Matter"

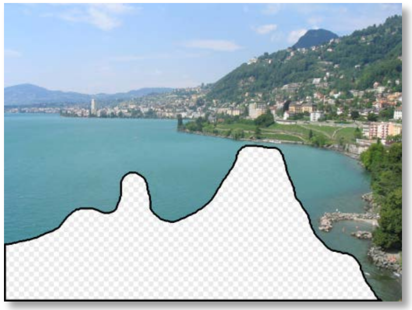
Given enough data, most things will be close-by even with the dumb distance metrics!

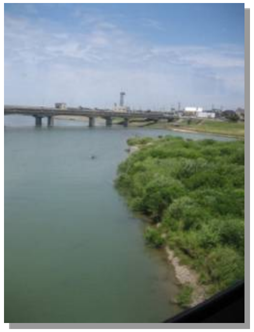


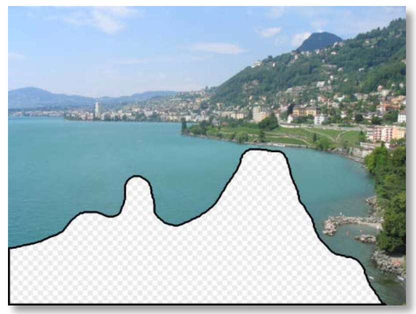
2 Million Flickr Images



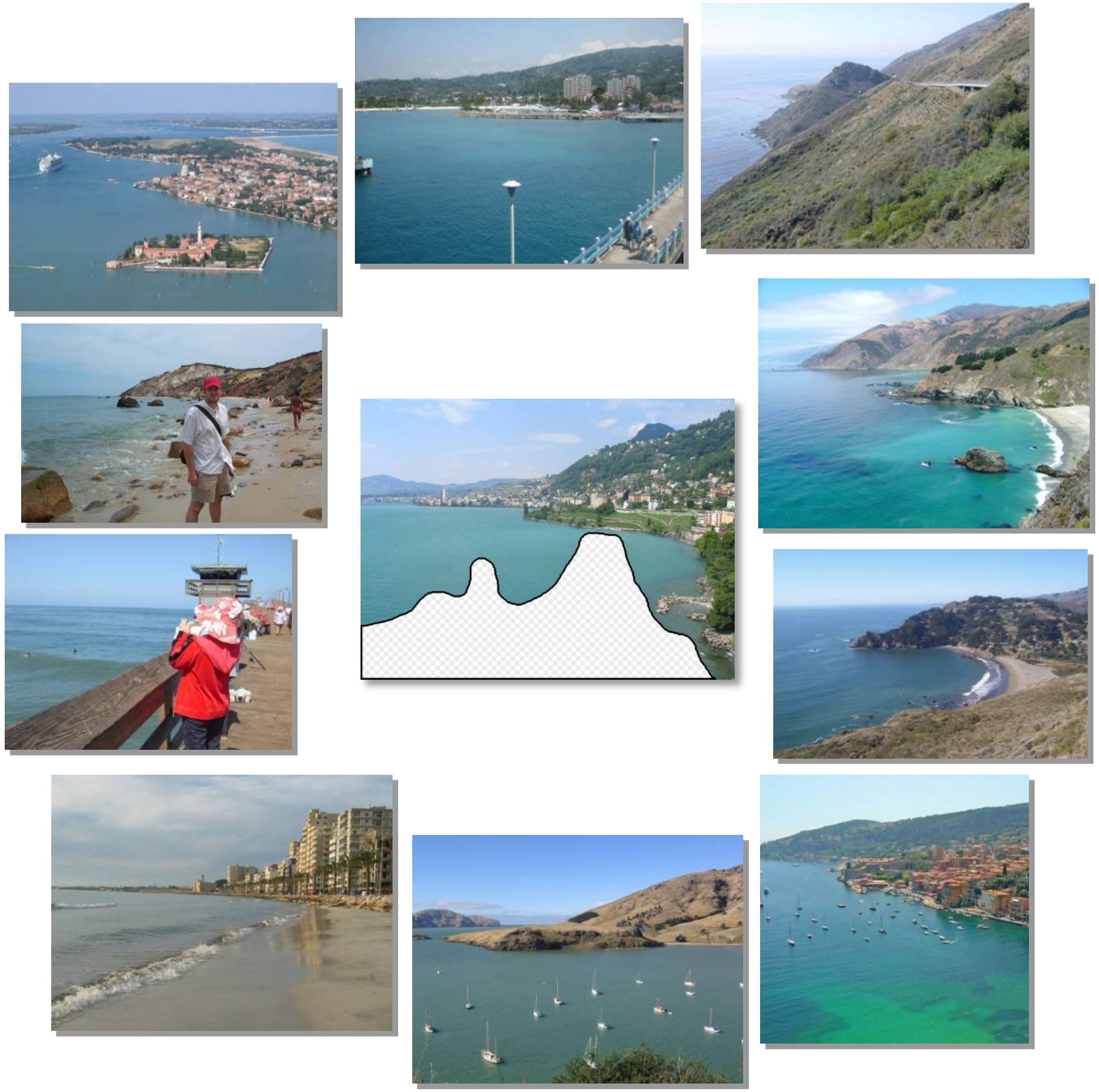




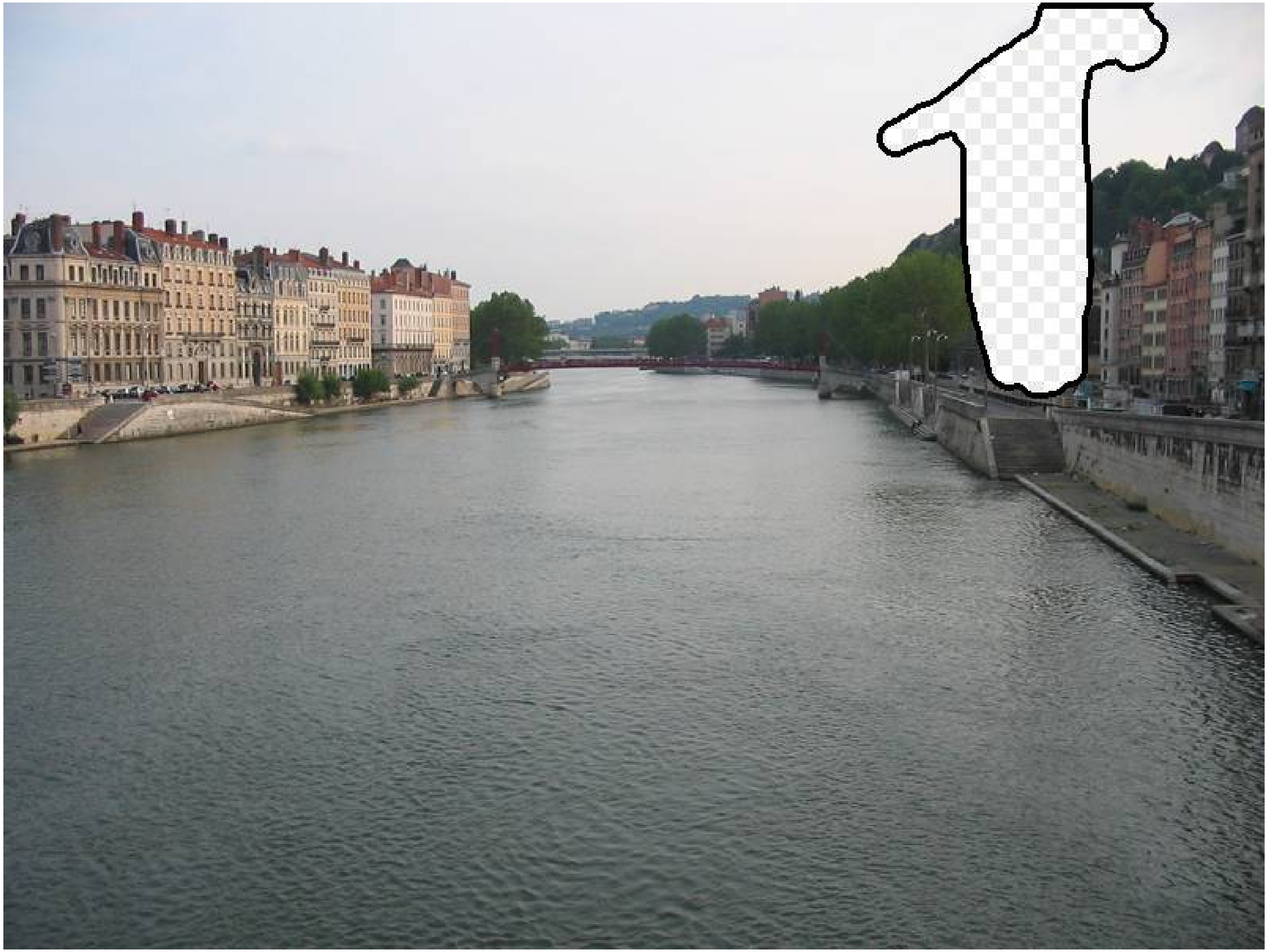


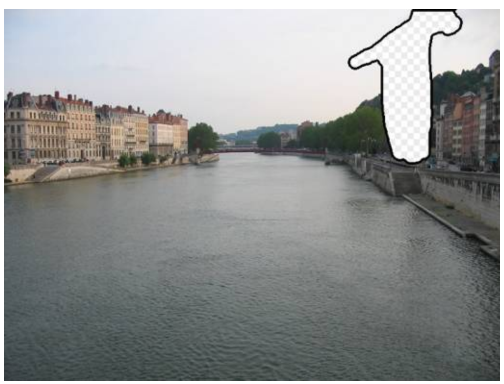


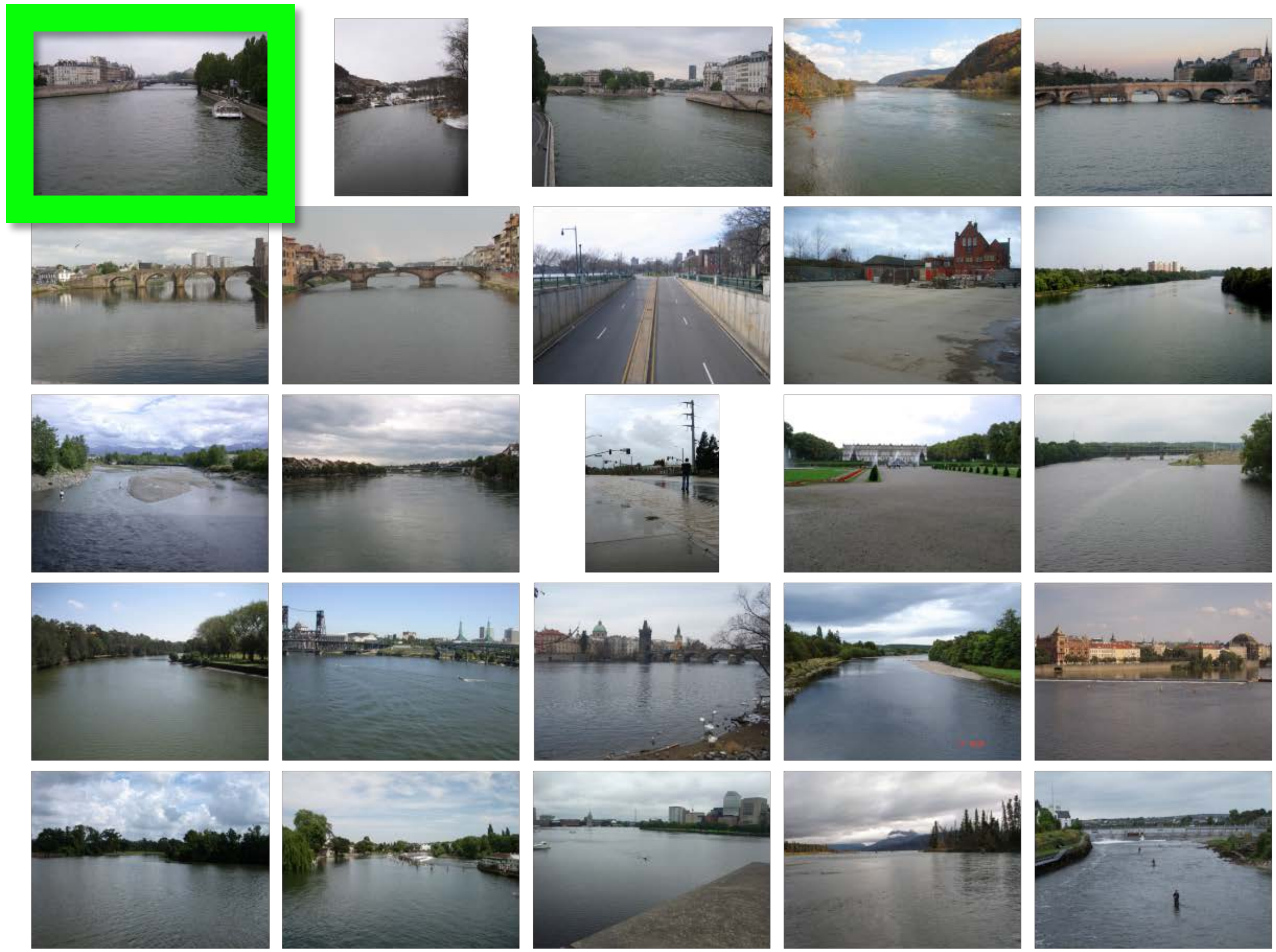
Nearest neighbors from a collection of 20 thousand images



Nearest neighbors from a
collection of 2 million images







… 200 scene matches

e.g. kNN for image understanding

Tags: Sky, Water, Beach, Sunny, …
Time: 1pm, August, 2006, …
Location: Italy, Greece, Hawaii …
Photographer: Flickrbug21, Traveller2

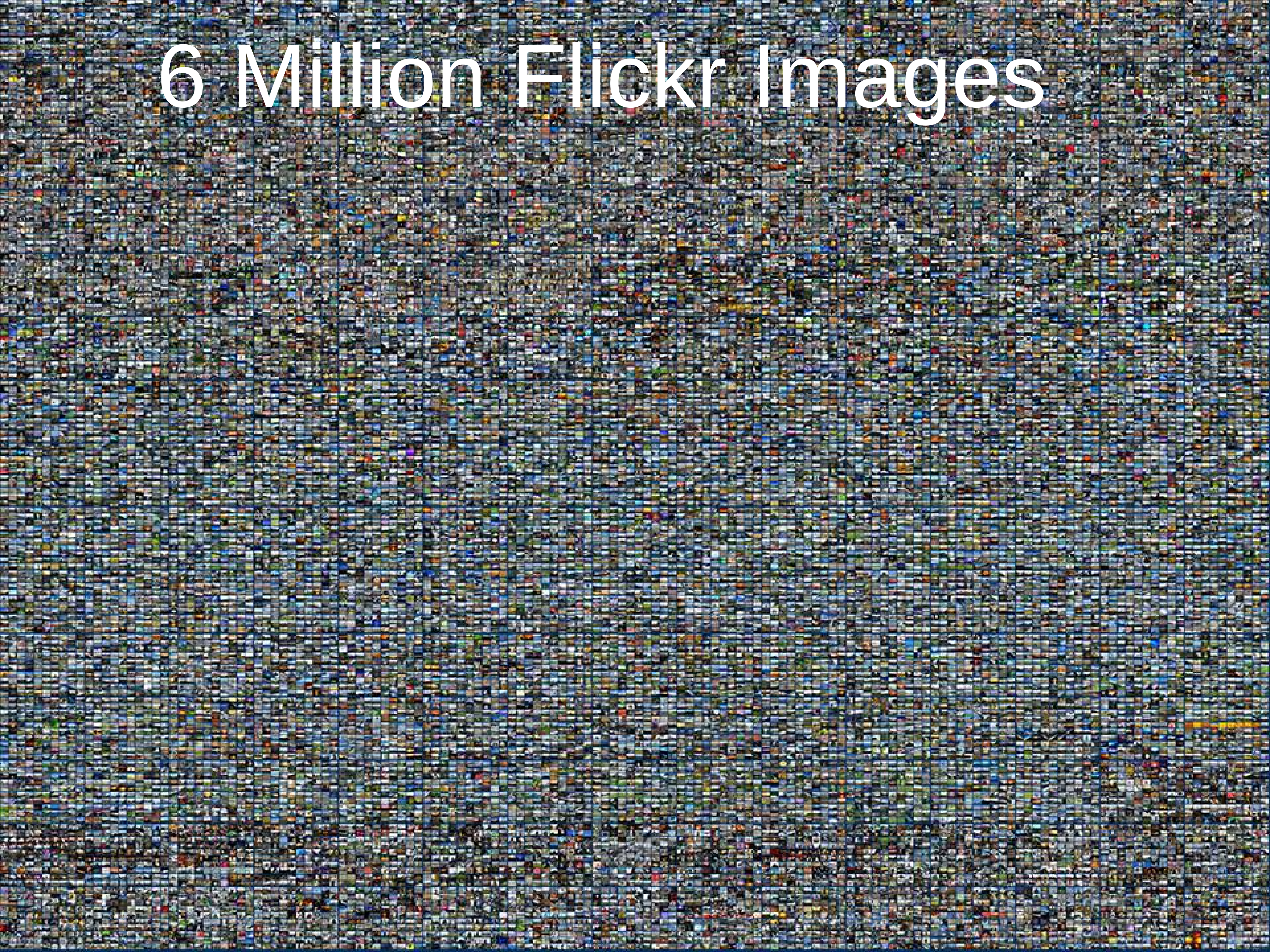
Label Transfer

im2GPS
(using 6 million GPS-tagged Flickr images)

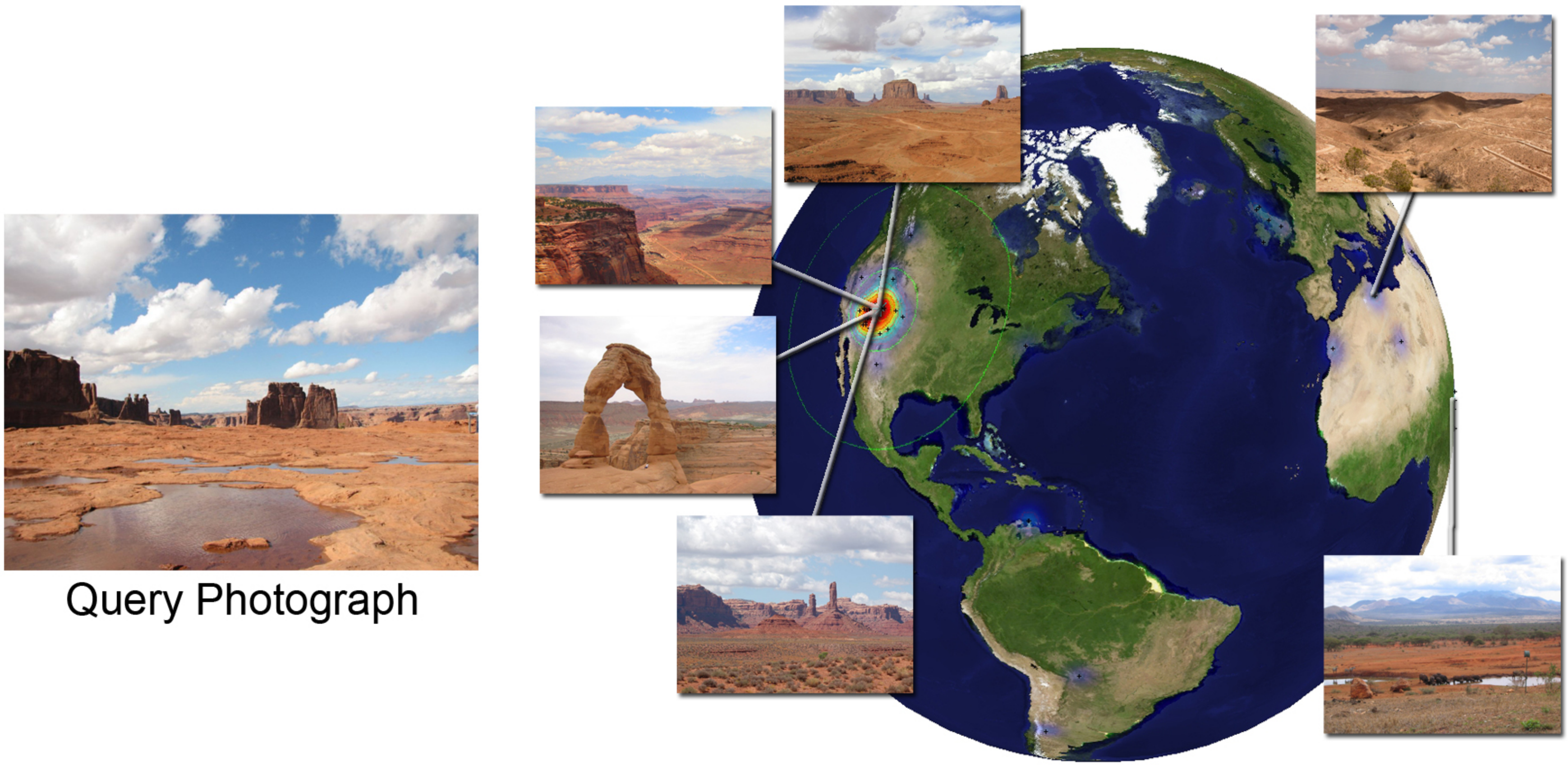
Query Photograph

Im2gps [Hays & Efros, CVPR'08]

6 Million Flickr Images



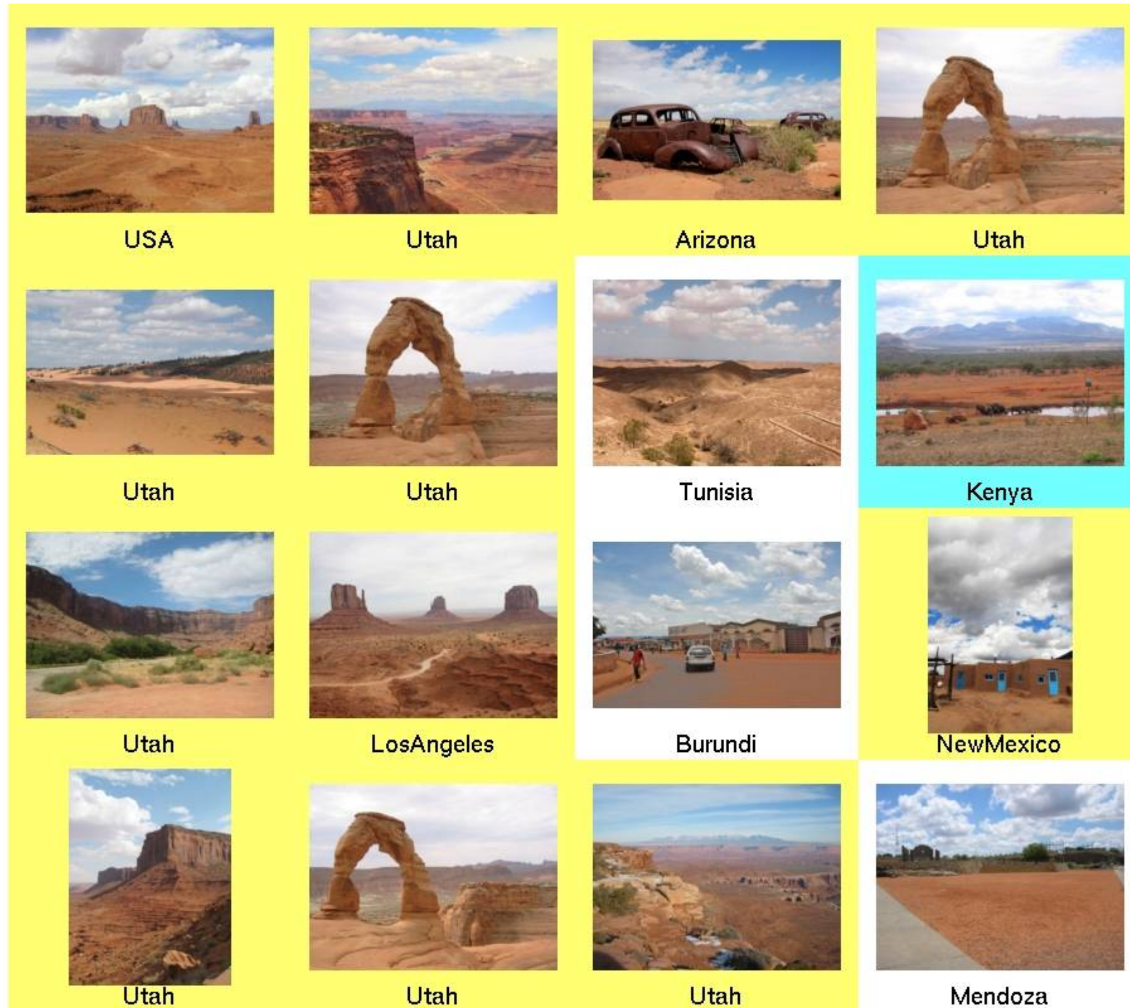
im2GPS
(using 6 million GPS-tagged Flickr images)

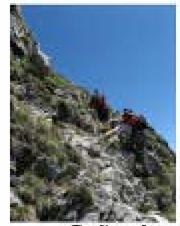


Query Photograph

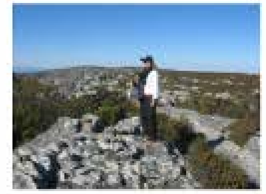
Visually Similar Scenes

Im2gps [Hays & Efros, CVPR'08]


USA
Utah
Arizona
Utah
Utah
Utah
Tunisia
Kenya
Utah
LosAngeles
Burundi
NewMexico
Utah
Utah
Utah
Mendoza



Switzerland



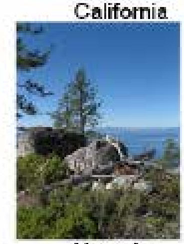
SouthAfrica

California

Barcelona

Italy

Italy



Nevada

Washington

Paris

Madrid

California

Oregon

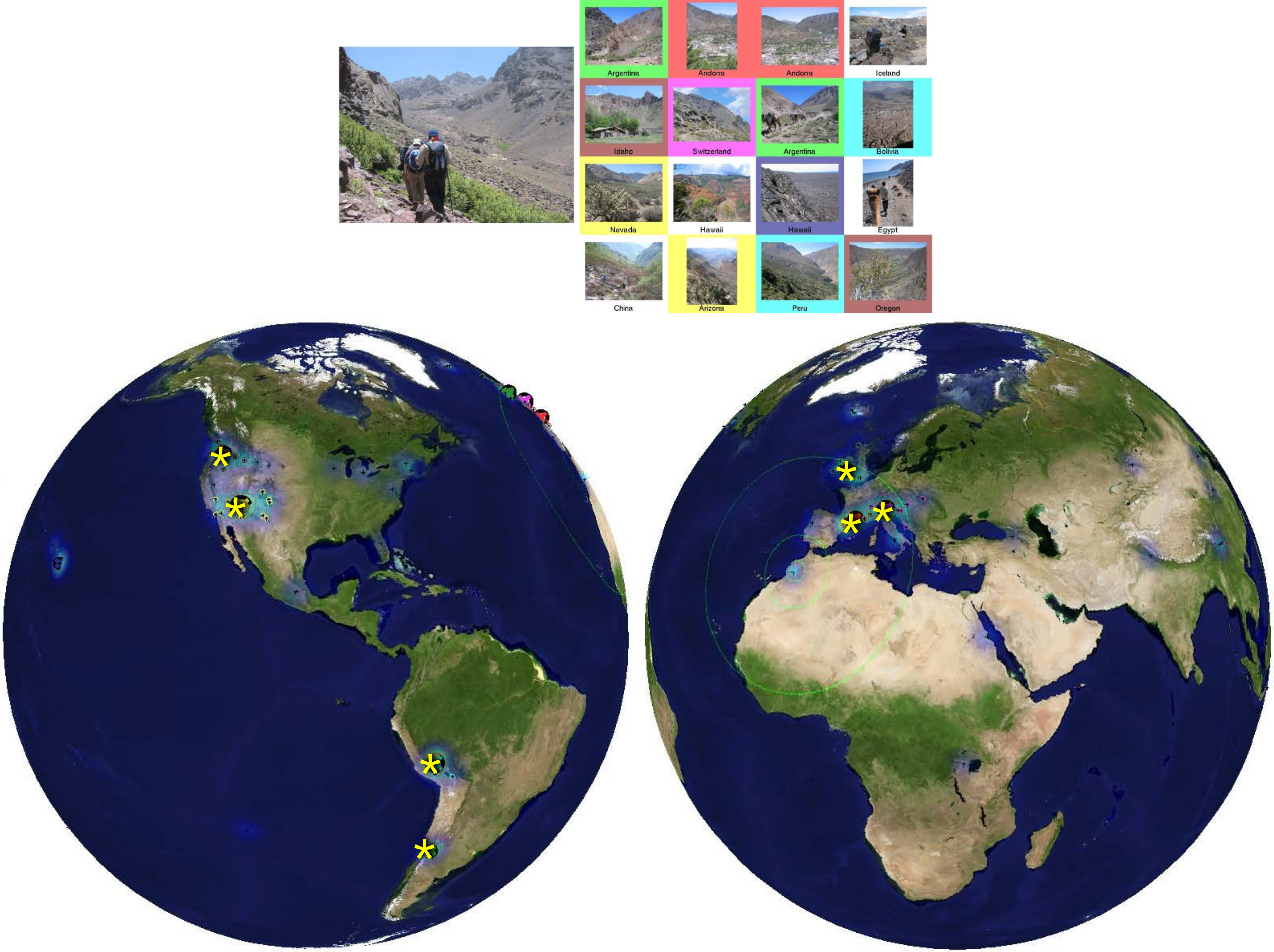
SouthDakota

USA

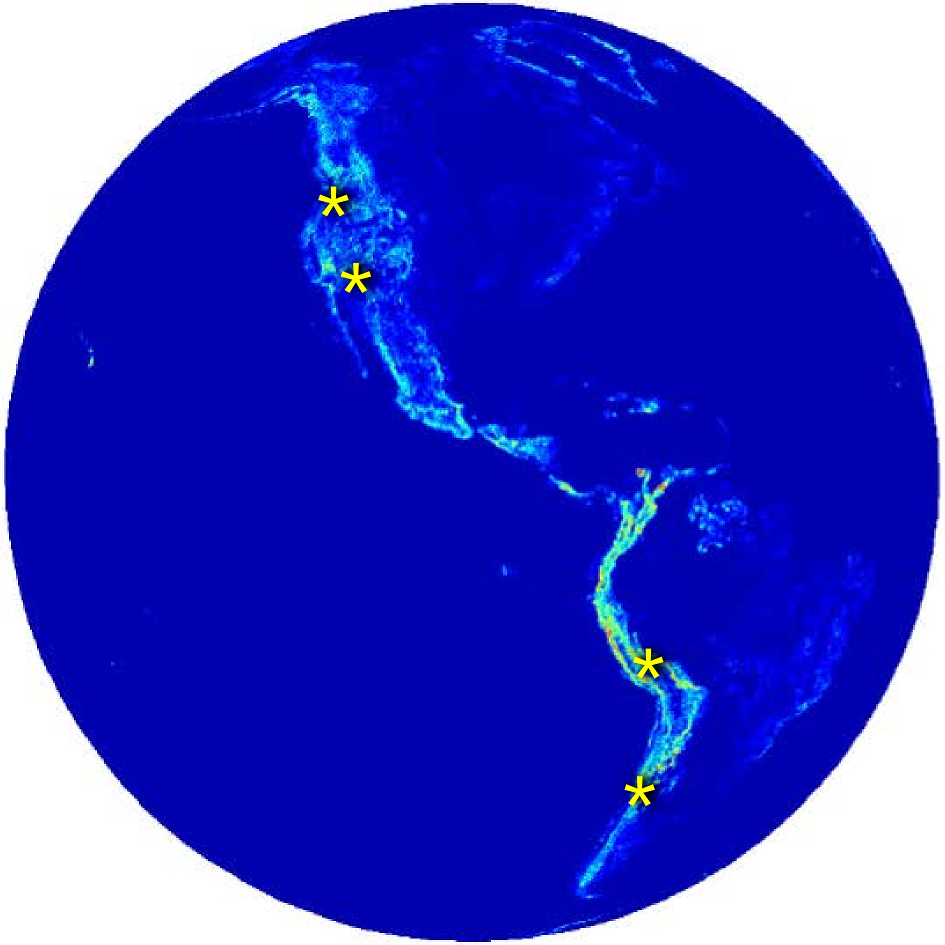
Bangkok

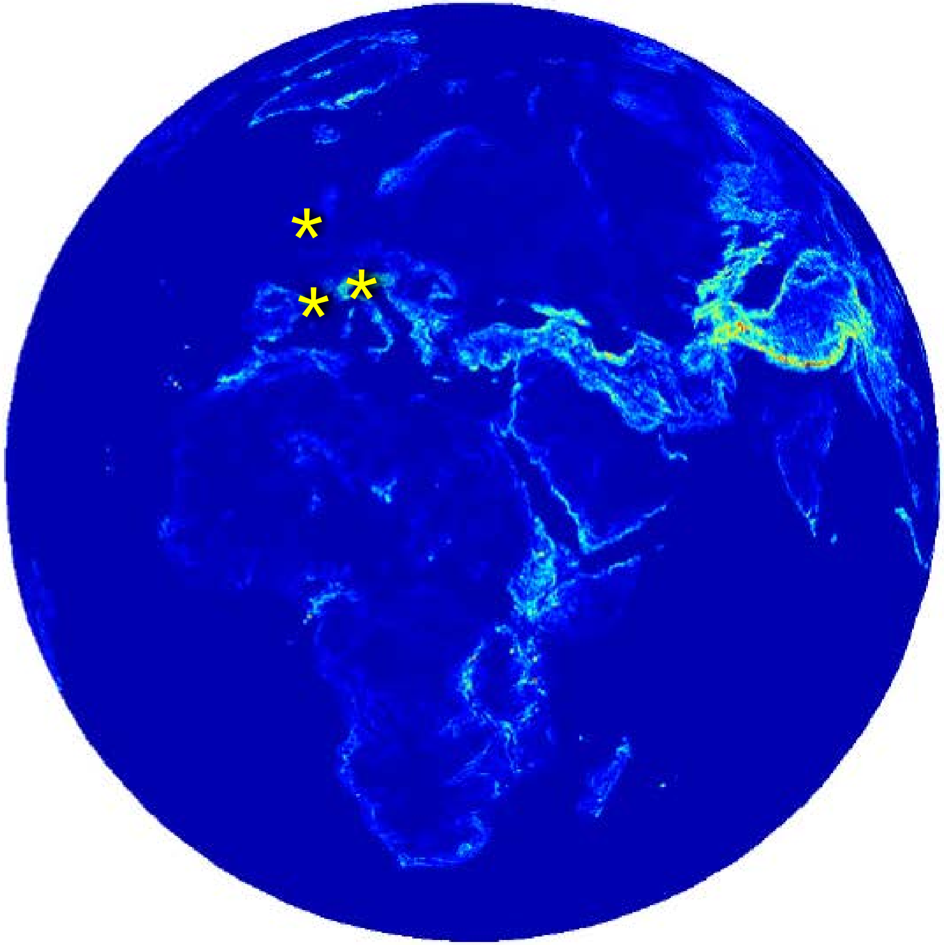
Italy


Lazy label transfer

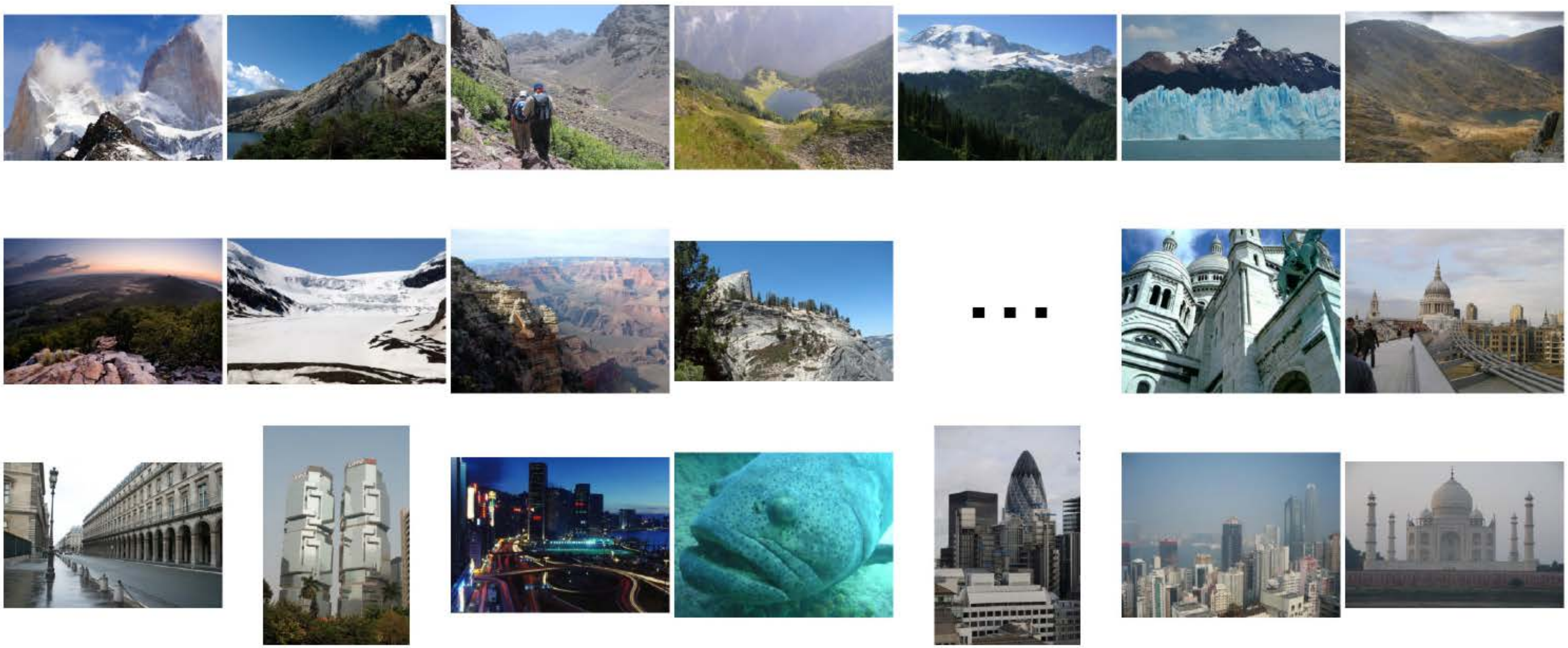

Argentina
Andorra
Andorra
Iceland
Idaho
Switzerland
Argentina
Bolivia
Nevada
Hawaii
Hawaii
Egypt
China
Arizona
Peru
Oregon

Elevation gradient =
112 m / km





Elevation gradient magnitude ranking



Population density map

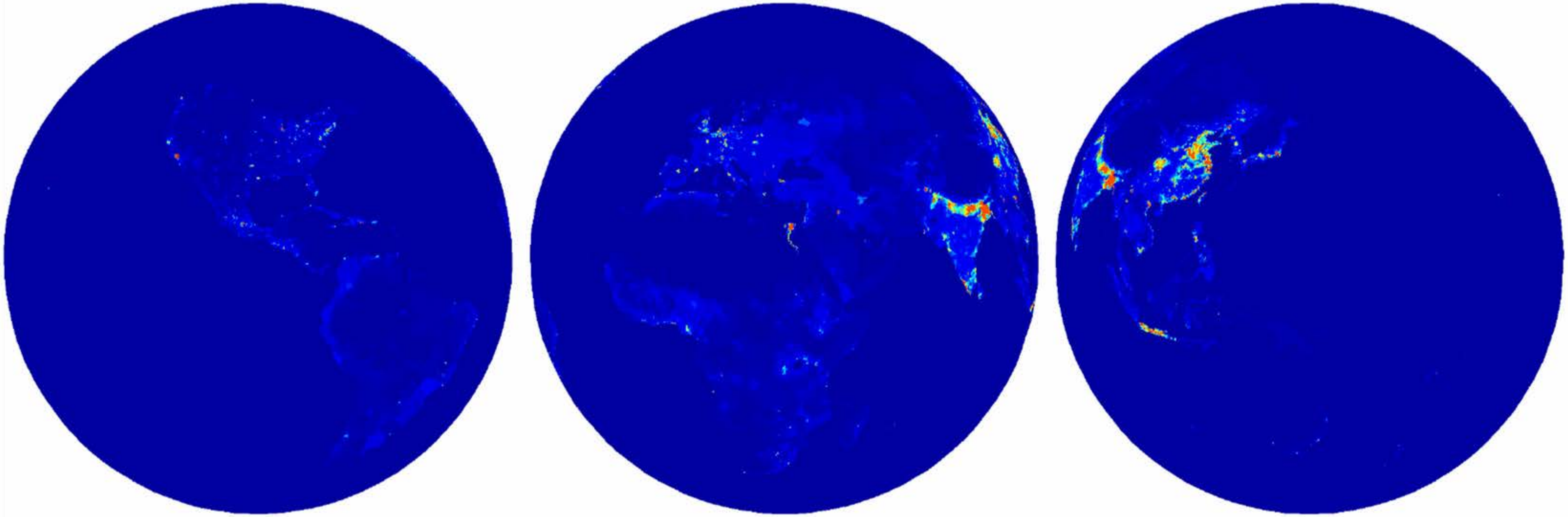
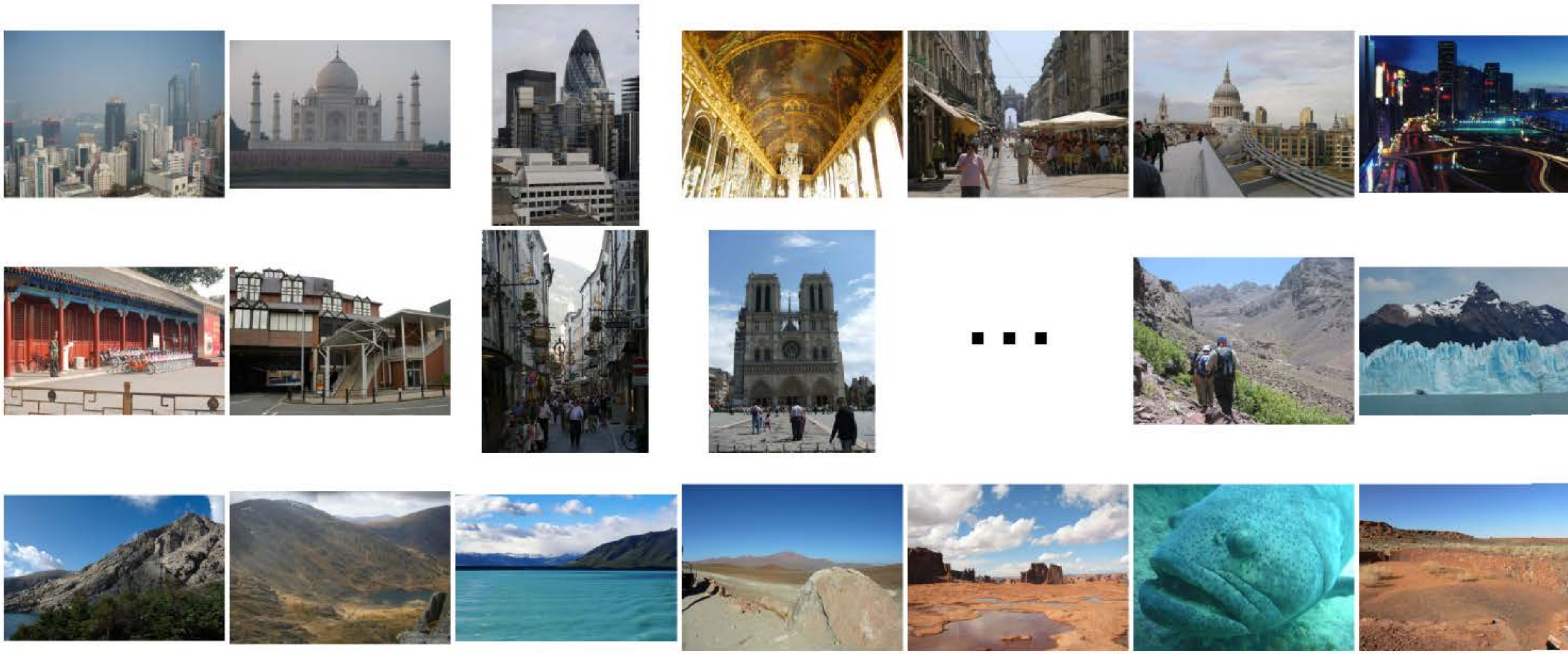


Figure 2. Global population density map.

Population density ranking



But surely the brain can’t remember this much!?

What's the Capacity of Visual Long Term Memory?

What we know…

Standing (1973)

10,000 images

83% Recognition

… people can remember thousands of images

What we don't know…

… what people are remembering for each item?

According to Standing

"Basically, my recollection is that we just separated the pictures into distinct thematic categories: e.g. cars, animals, single-person, 2-people, plants, etc.) Only a few slides were selected which fell into each category, and they were visually distinct."

Dogs
Playing Cards

"Gist" Only

Sparse Details

Highly Detailed

Slide by Aude Oliva

Massive Memory I: Methods

Showed 14 observers 2500 **categorically unique objects**

1 at a time, 3 seconds each

800 ms blank between items

Study session lasted about 5.5 hours

Repeat Detection task to maintain focus

Followed by 300 2-alternative forced choice tests

Slide by Aude Oliva

Massive Memory Experiment I

A stream of objects will be presented on the screen for
~ 3 second each.

Your primary task:

Remember them ALL!

afterwards you will be tested with…

Completely different objects...

Different exemplars of the same kind of object...

Different states of the same object...

Massive Memory Experiment I

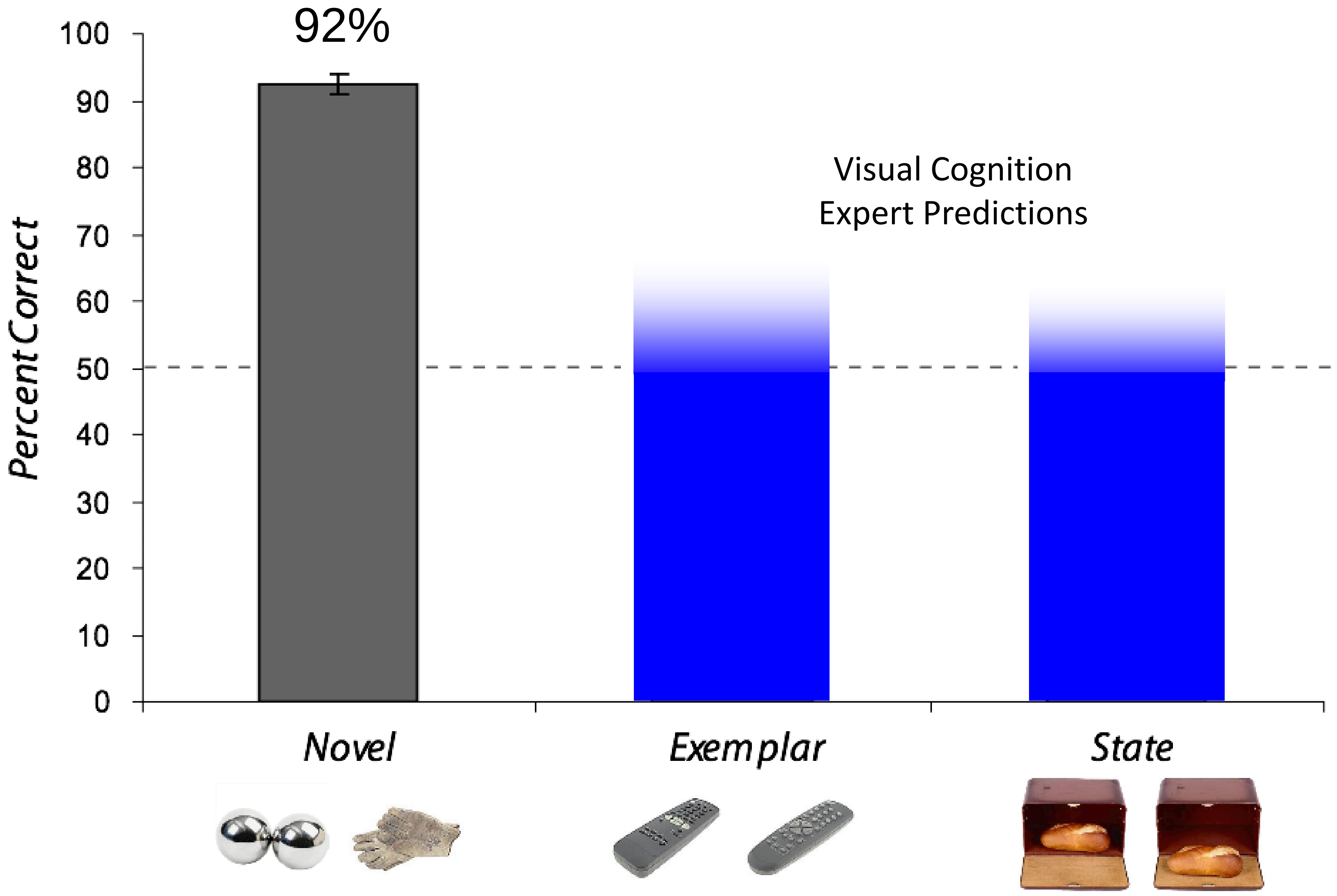
Your other task:

Detect exact repeats anywhere in the stream

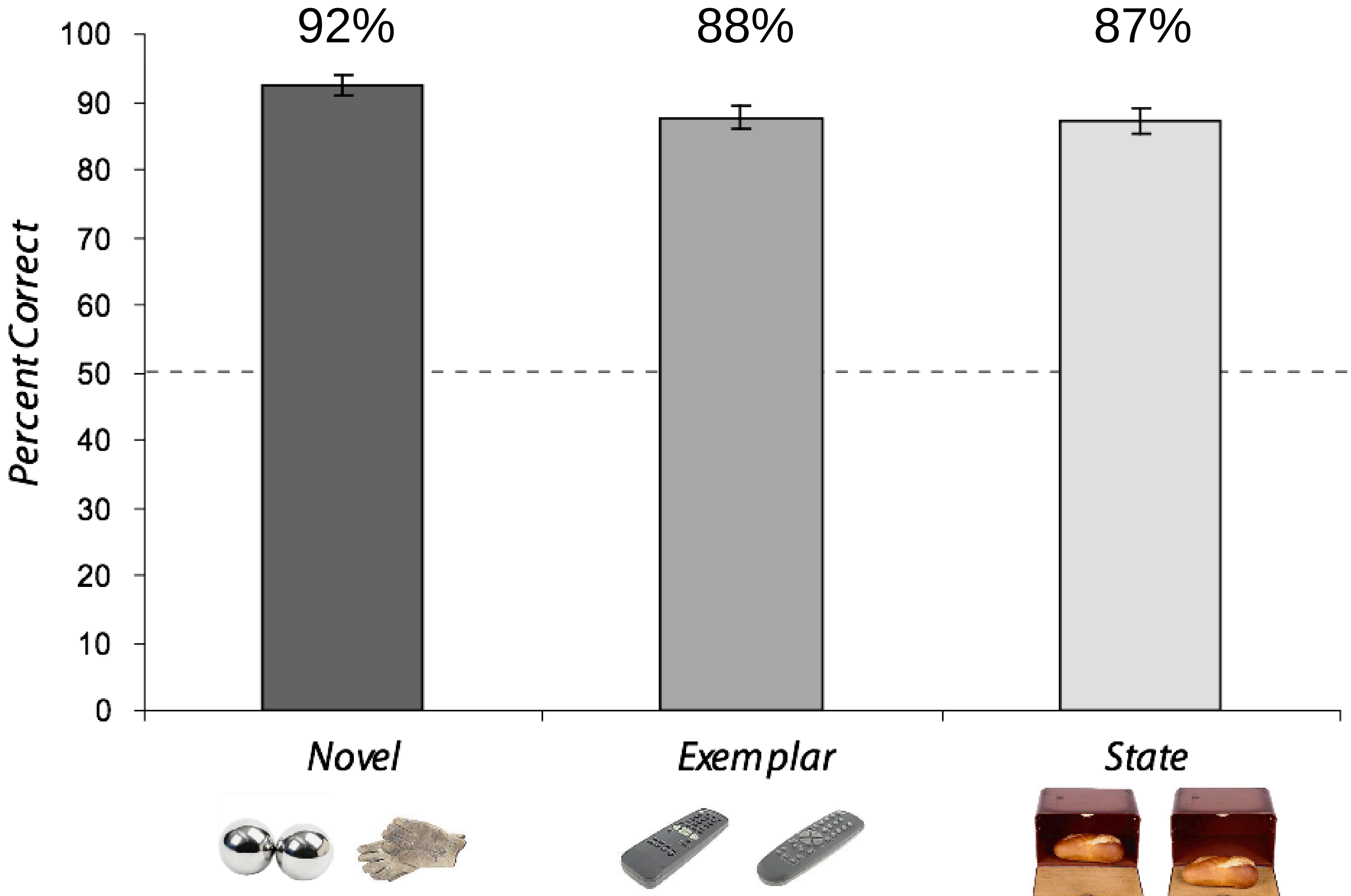
Examples of **State** memory test

All stimuli available at: cvcl.mit.edu/MM

Recognition Memory Results

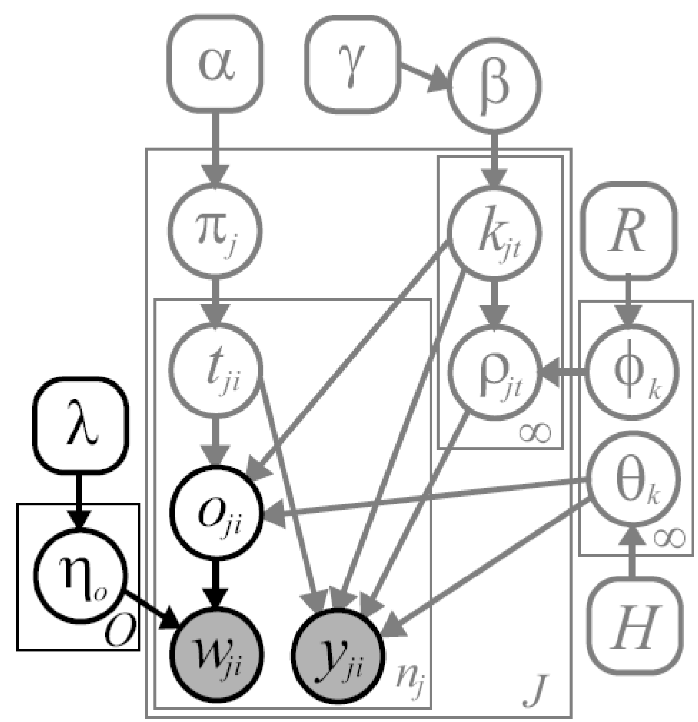


Recognition Memory Results



Brady, et al. (2008), *PNAS*

Part 2: Nearest Neighbors as a negative result



Word embeddings

- word2vec
- Matrix factorization
- (normalized) Nearest Neighbors
  - Omer Levy, Yoav Goldberg, “Linguistic regularities in sparse and explicit word representations.” CoNLL-2014.

Image captioning

- LSTMs
- Feed-forward CNNs
- Language models
- …

Easy to get fooled

“a car parked on the side of the road”









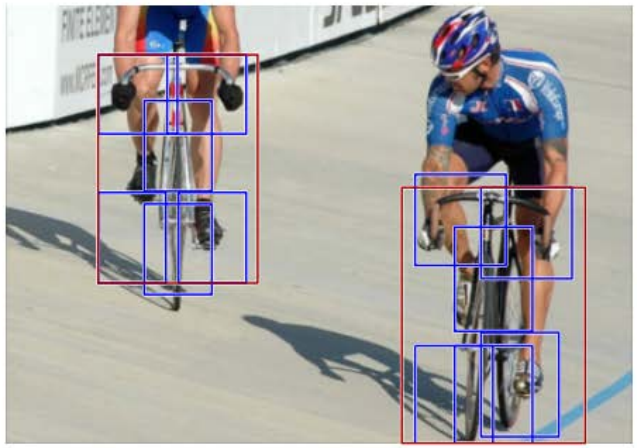
“a car parked on the side of the road”

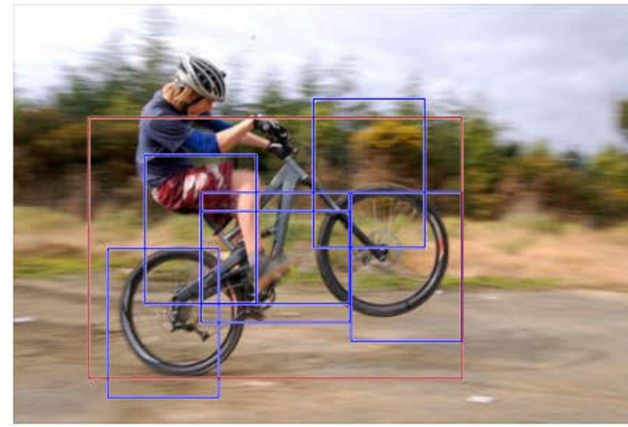
“a car parked on the side of the road”

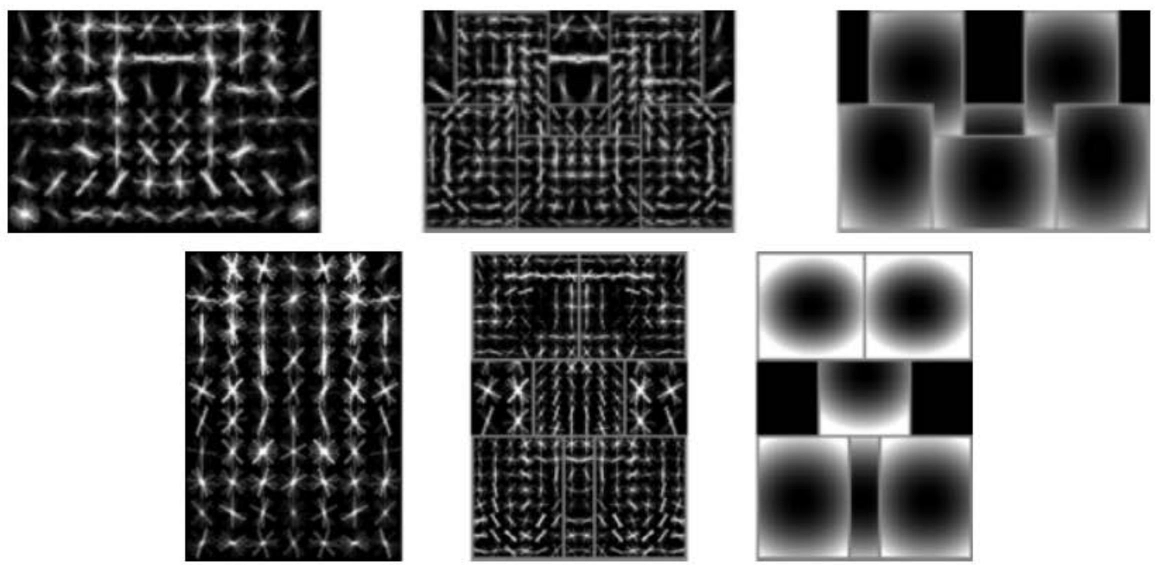
Image captioning

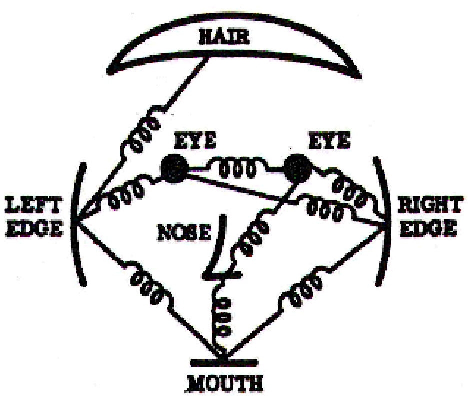
- LSTMs
- Feed-forward CNNs
- Language models
- …
- Nearest neighbors
  - “Language Models for Image Captioning: The Quirks and What Works”, Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, Margaret Mitchell, ACL 2015

Deformable Part Models







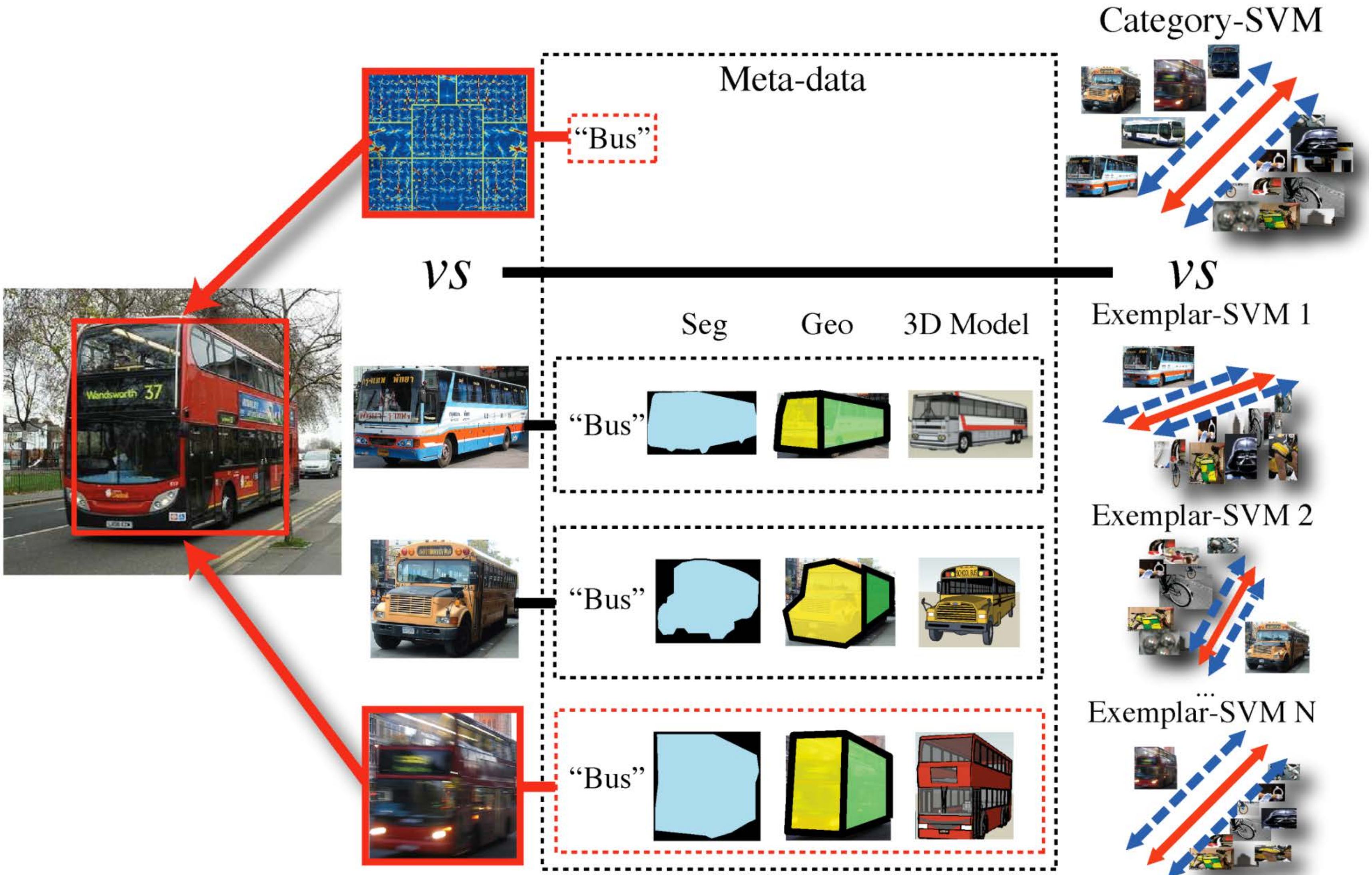


How important are "Deformable Parts" in the Deformable Parts Model?

Santosh K. Divvala, Alexei A. Efros, and Martial Hebert

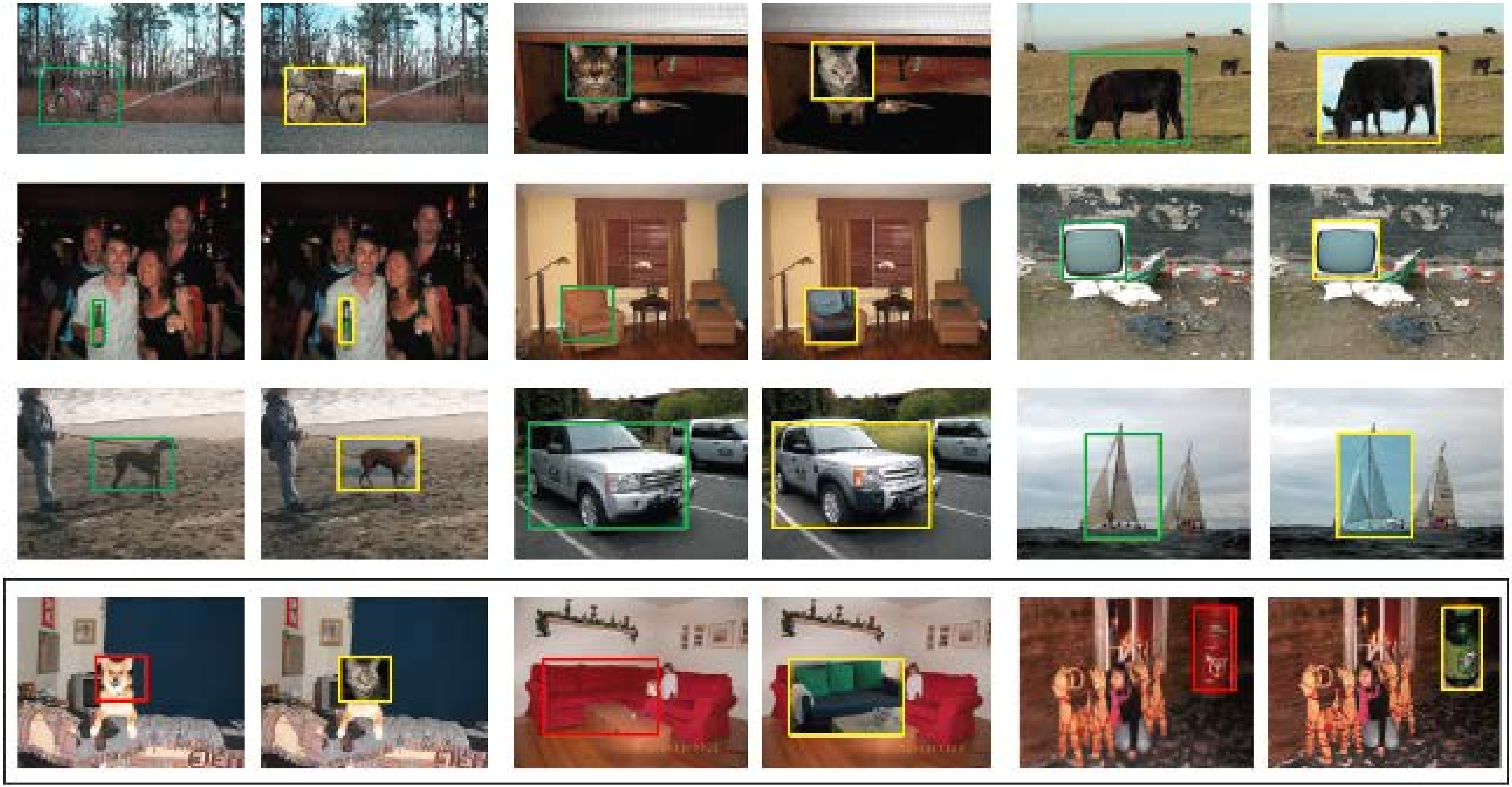
Robotics Institute, Carnegie Mellon University.

Exemplar-SVMs



Malisiewicz et al, ICCV'11

Showing off correspondences

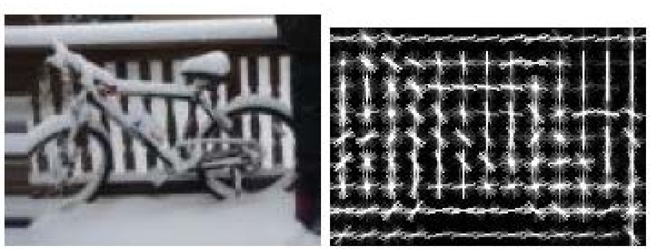


Malisiewicz et al, ICCV'11

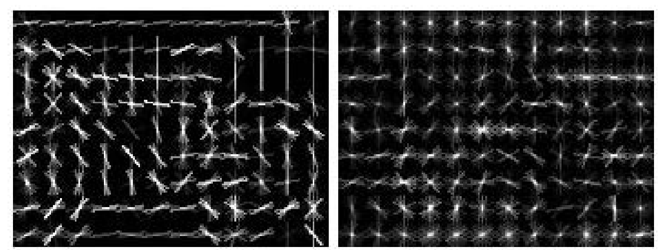
Discriminative Decorrelation for Clustering and Classification*

Bharath Hariharan[1], Jitendra Malik[1], and Deva Ramanan[2]

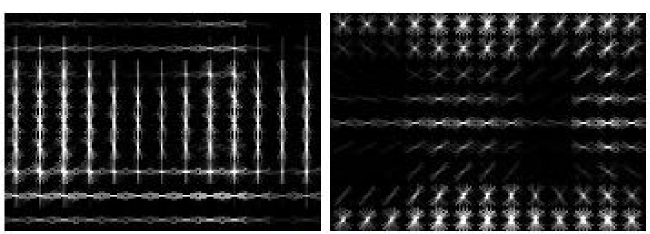
[1] Univerisity of California at Berkeley, Berkeley, CA, USA
{bharath2,malik}@cs.berkeley.edu
[2] University of California at Irvine, Irvine, CA, USA
dramanan@ics.uci.edu



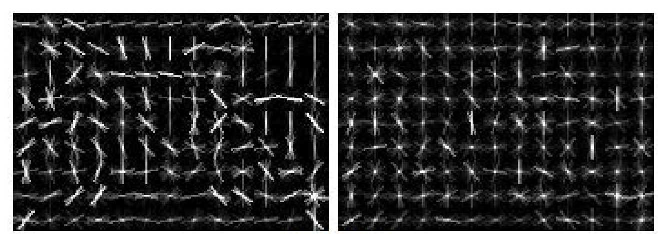
(a) Image (left) and HOG (right)



(b) SVM



(c) PCA



(d) LDA

im2GPS
(using 6 million GPS-tagged Flickr images)

Query Photograph

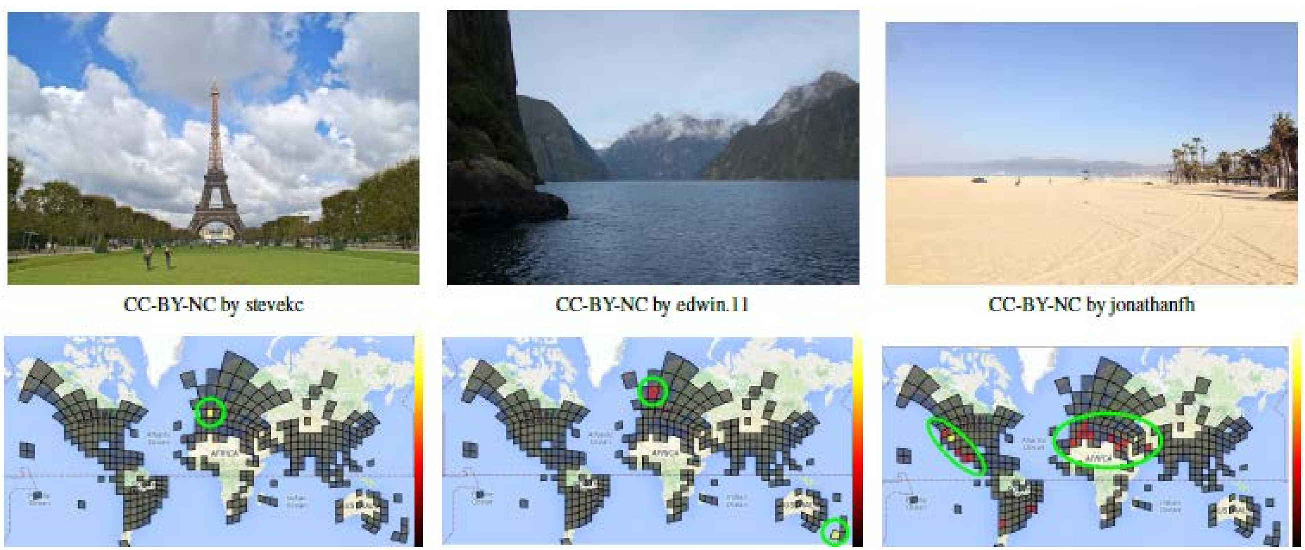
Im2gps [Hays & Efros, CVPR'08]

2006 to 2016

PlaNet - Photo Geolocation with Convolutional Neural Networks

Tobias Weyand[1], Ilya Kostrikov[2], James Philbin[3]

[1]Google, Los Angeles, USA
weyand@google.com
[2]RWTH Aachen University, Aachen, Germany*
ilya.kostrikov@rwth-aachen.de
[3]Zoox, Menlo Park, USA*
james@zoox.com

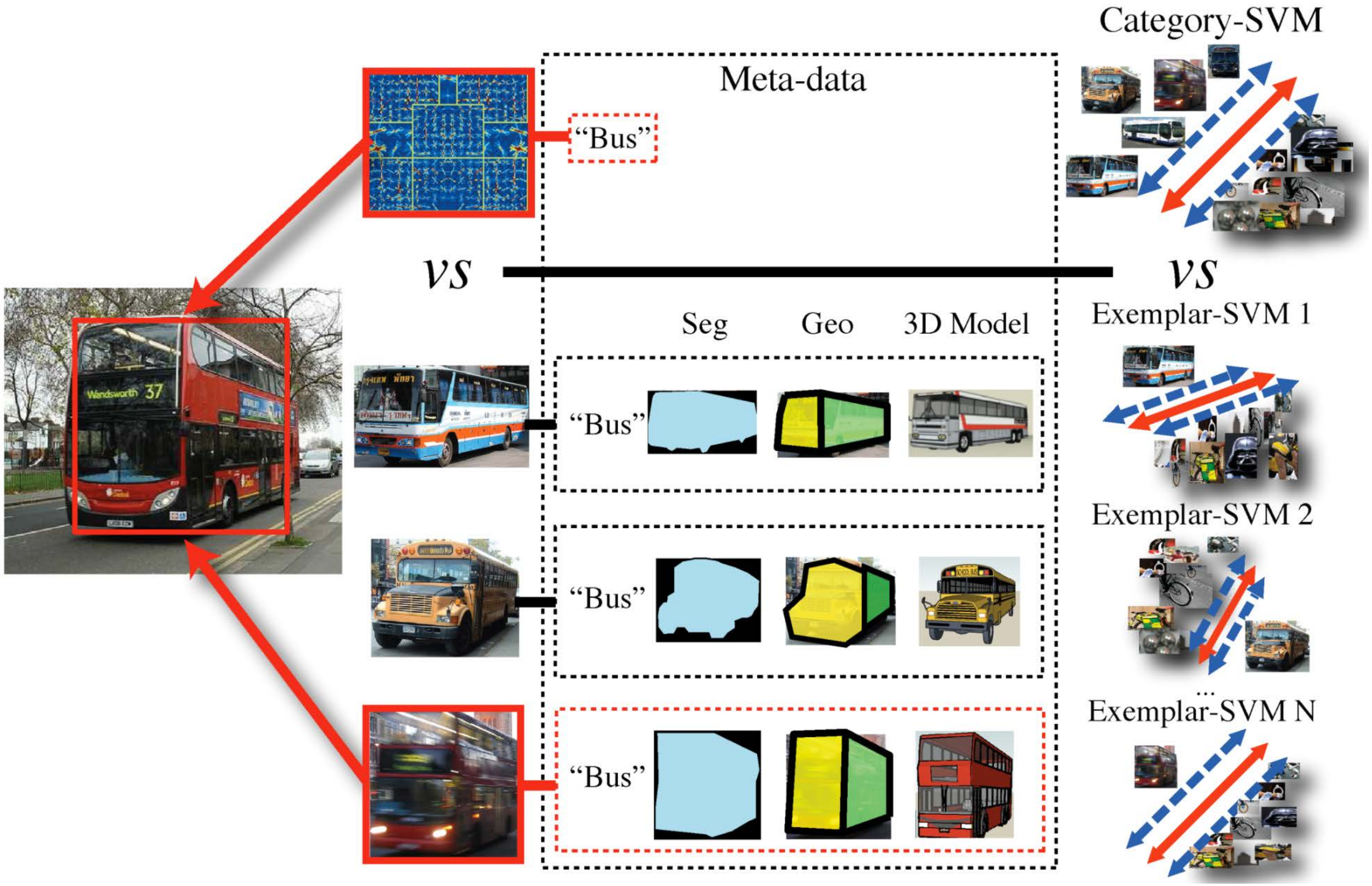


(a) (b) (c)

Deep Features vs. Data

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS (orig) [19]		12.0%	15.0%	23.0%	47.0%
Im2GPS (new) [20]	2.5%	21.9%	32.1%	35.4%	51.9%
PlaNet (900k)	0.4%	3.8%	7.6%	21.6%	43.5%
PlaNet (6.2M)	6.3%	18.1%	30.0%	45.6%	65.8%
PlaNet (91M)	**8.4%**	**24.5%**	**37.6%**	**53.6%**	**71.3%**

Exemplar-SVMs



Malisiewicz et al, ICCV'11

Part 3: Nearest Neighbors for category-free understanding

Understanding an Image

slide by Fei Fei, Fergus & Torralba

Object naming -> Object categorization

Object categorization

sky

building

flag

face

banner

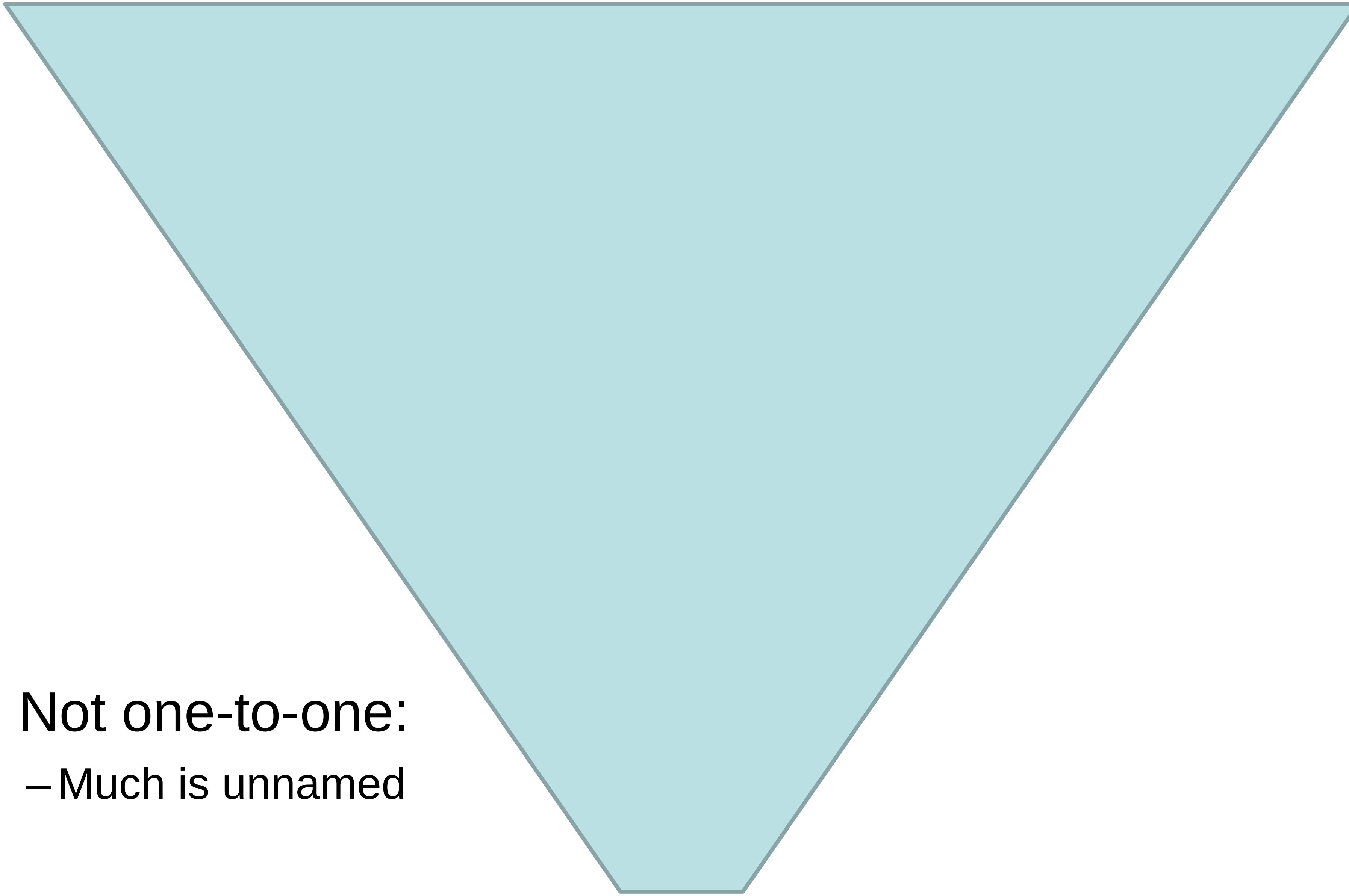
wall

street lamp

bus

bus

cars



- Not one-to-one:
  - Much is unnamed

Visual World

- Not one-to-one:
  - Much is unnamed

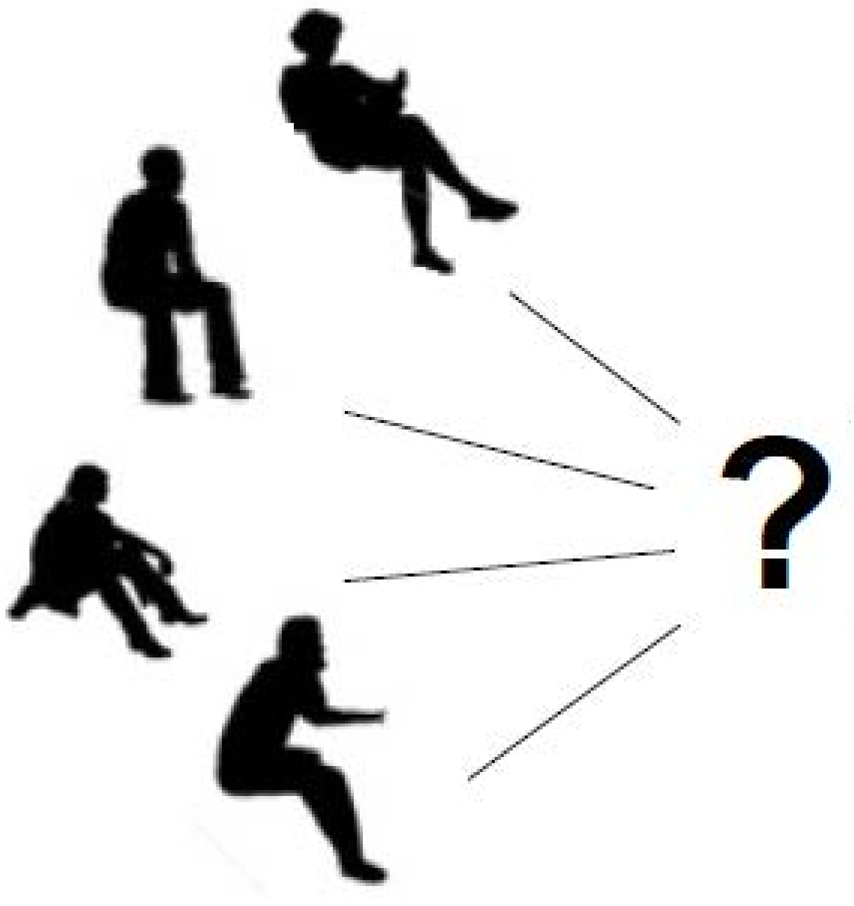
CITY

words

Verbs (actions)

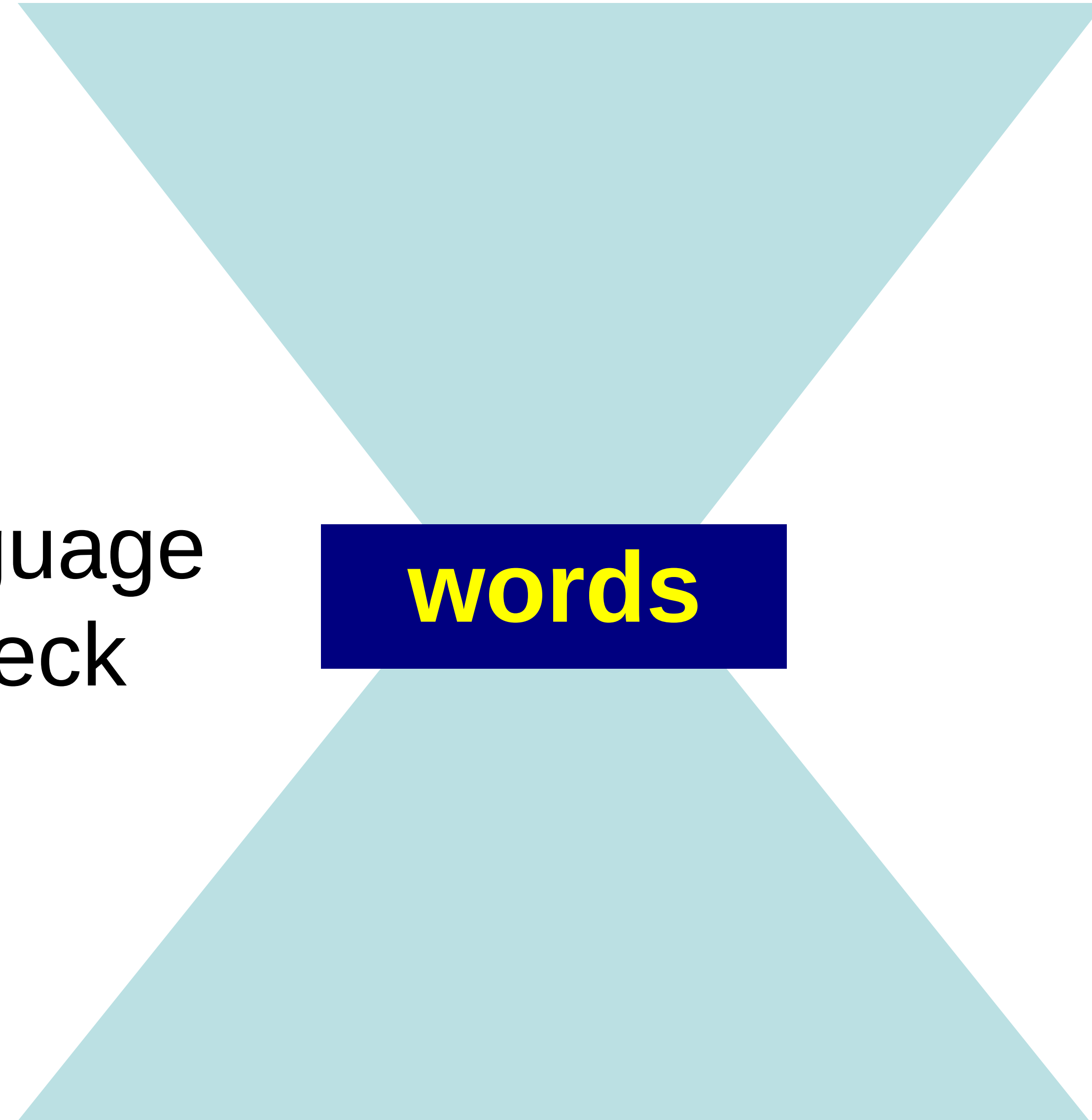
sitting

Visual “sitting”



Visual World

The Language Bottleneck



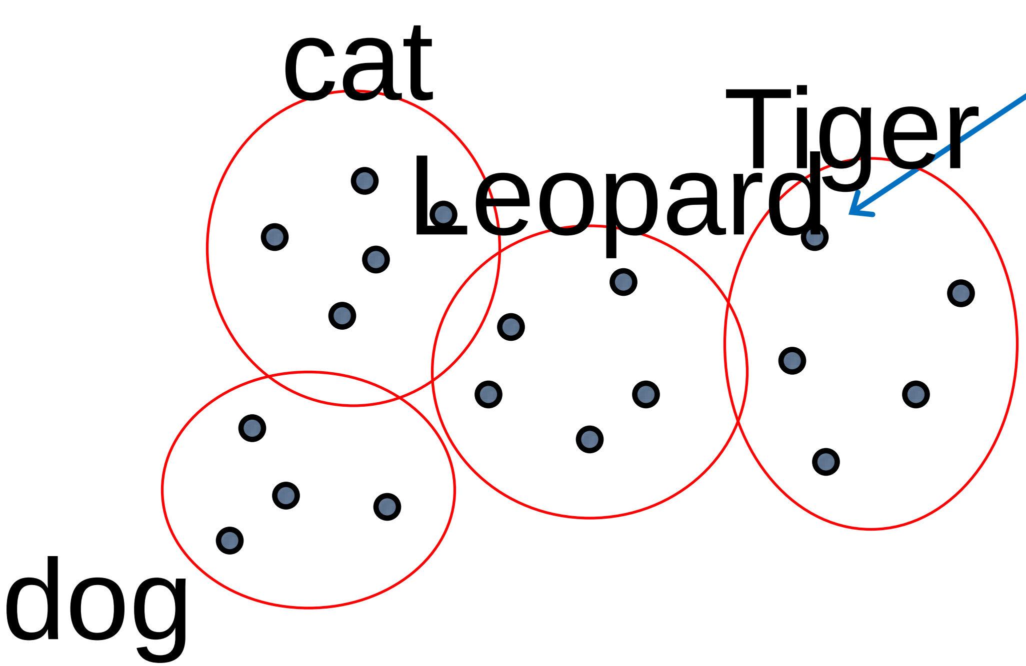
Scene understanding, spatial reasoning, prediction, image retrieval, image synthesis, etc.

Visual World

Scene understanding, spatial reasoning, prediction, image retrieval, image synthesis, etc.

Why Categorize?

1. Knowledge Transfer
2. Communication



Classical View of Categories

- Dates back to Plato & Aristotle
  1. Categories are defined by a list of properties shared by all elements in a category
  2. Category membership is binary
  3. Every member in the category is equal

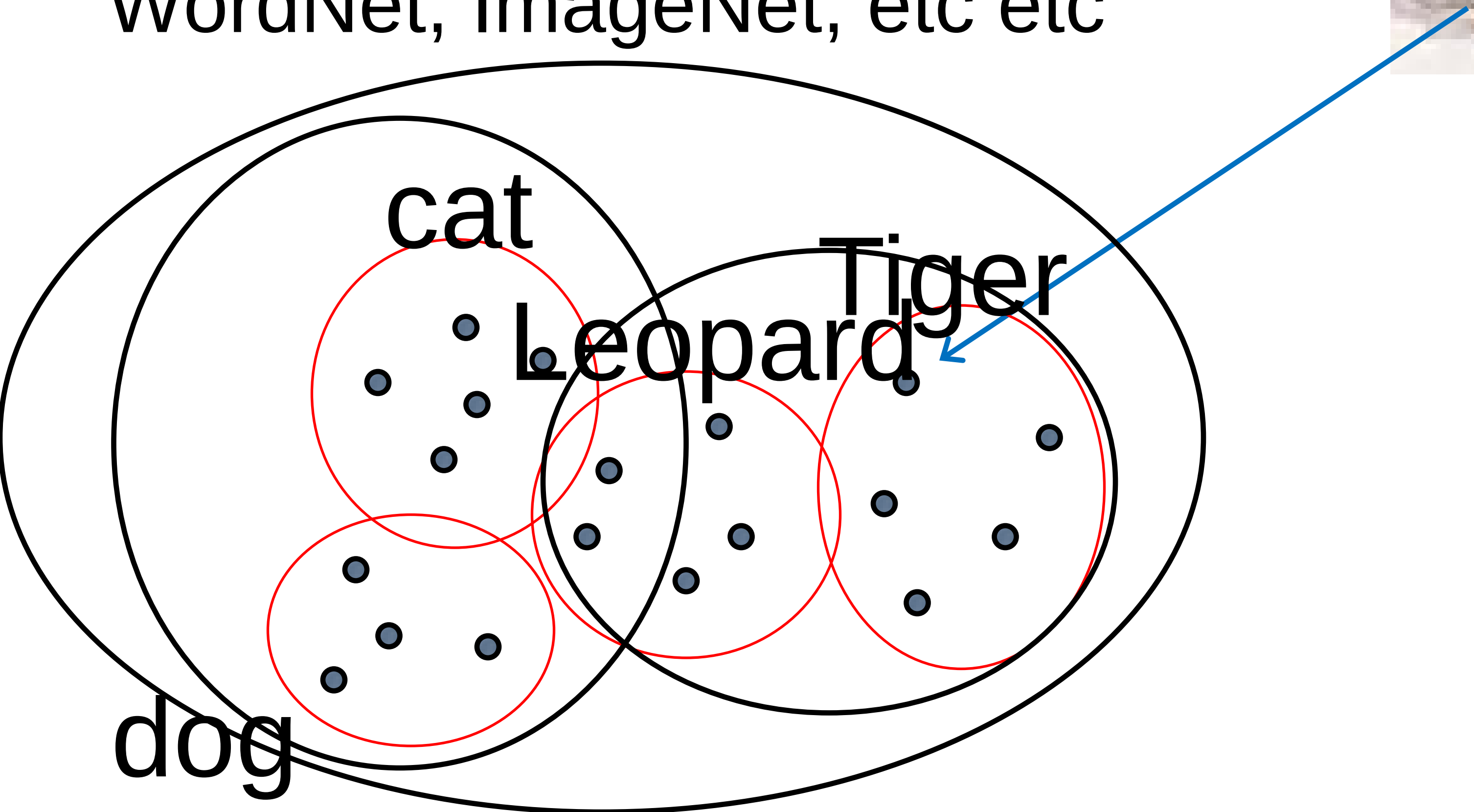
Problems with Classical View

- Humans don't do this!
  - People don't rely on abstract definitions / lists of shared properties (Wittgenstein 1953, Rosch 1973)
    - e.g. define the properties shared by all "games"
    - e.g. are curtains furniture? Are olives fruit?
  - Typicality
    - e.g. Chicken -> bird, but bird -> eagle, pigeon, etc.
  - Language-dependent
    - e.g. "Women, Fire, and Dangerous Things" category is Australian aboriginal language (Lakoff 1987)
  - Doesn't work even in human-defined domains
    - e.g. Is Pluto a planet?

Solution: hierarchy?

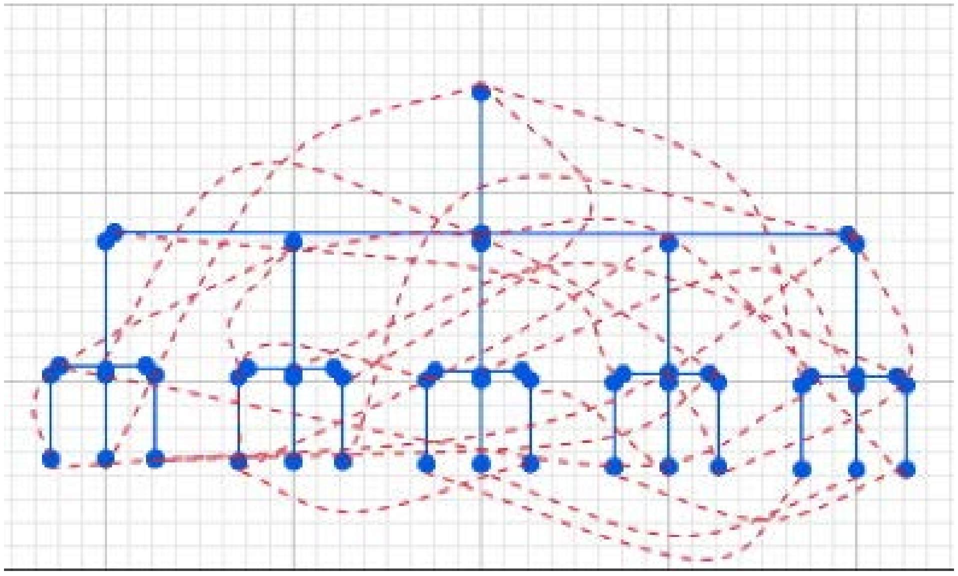
Ontologies, hierarchies, levels of categories (Rosch), etc.

WordNet, ImageNet, etc etc



Still Problematic!

- – Intransitivity
  - e.g. car seat is chair, chair is furniture, but …
- – Multiple category membership
  - it’s not a tree, it’s a forest!



Clay Shirky, “Ontologies are Overrated”

Fundamental Problem with Categorization

Making decisions too early!

Why not only categorize at run-time, once we know the task!

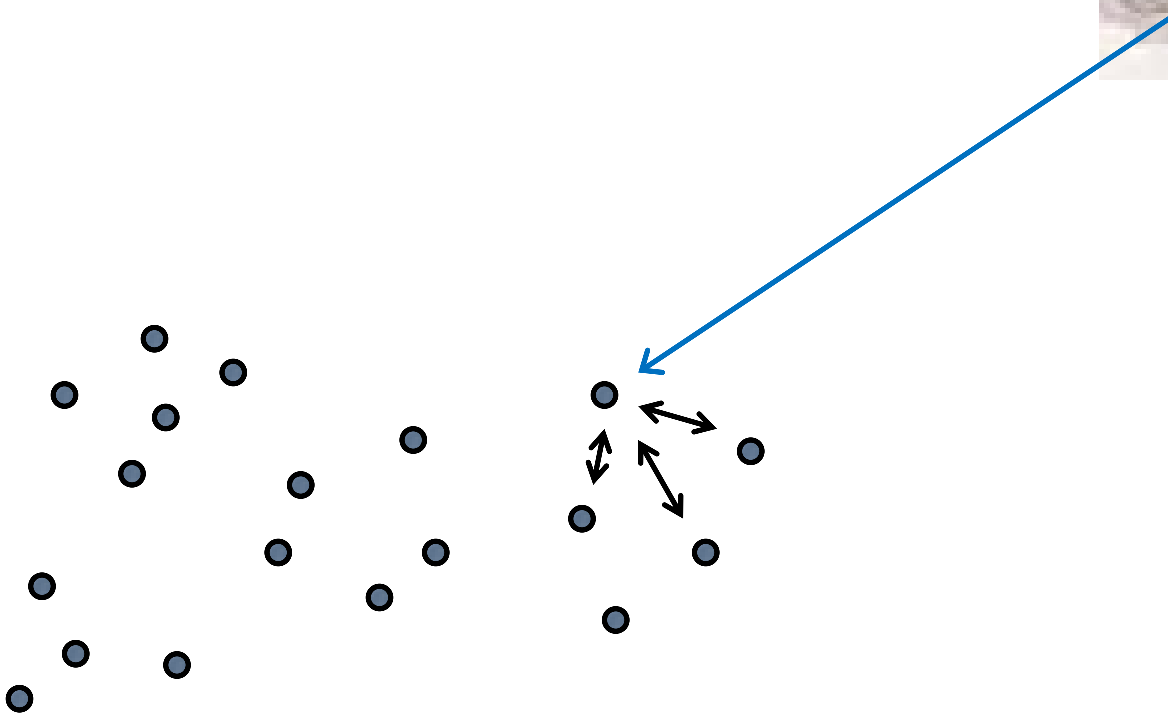
The Dictatorship of Librarians

categories are losing…

YAHOO! vs. Google™

On-the-fly Categorization?

1. Knowledge Transfer
2. ~~Communication~~



Association instead of categorization

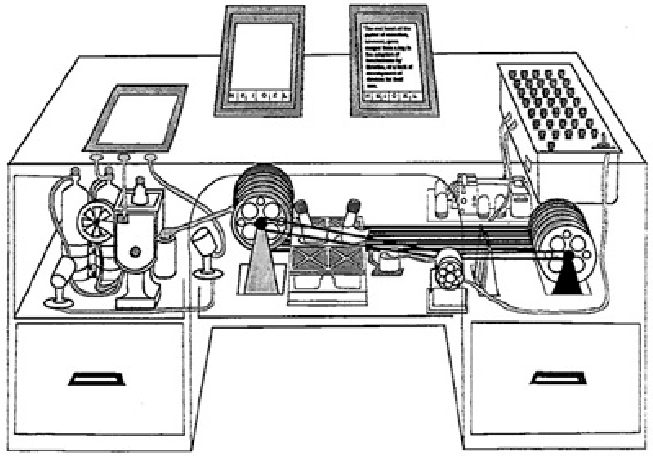
Ask not "what is this?", ask "what is this like"

– Moshe Bar

- Exemplar Theory (Medin & Schaffer 1978, Nosofsky 1986, Krushke 1992)
  - categories represented in terms of remembered objects (exemplars)
  - Similarity is measured between input and all exemplars
  - *think* non-parametric density estimation
- Vanevar Bush (1945), Memex (MEMory EXtender)
  - Inspired hypertext, WWW, Google…

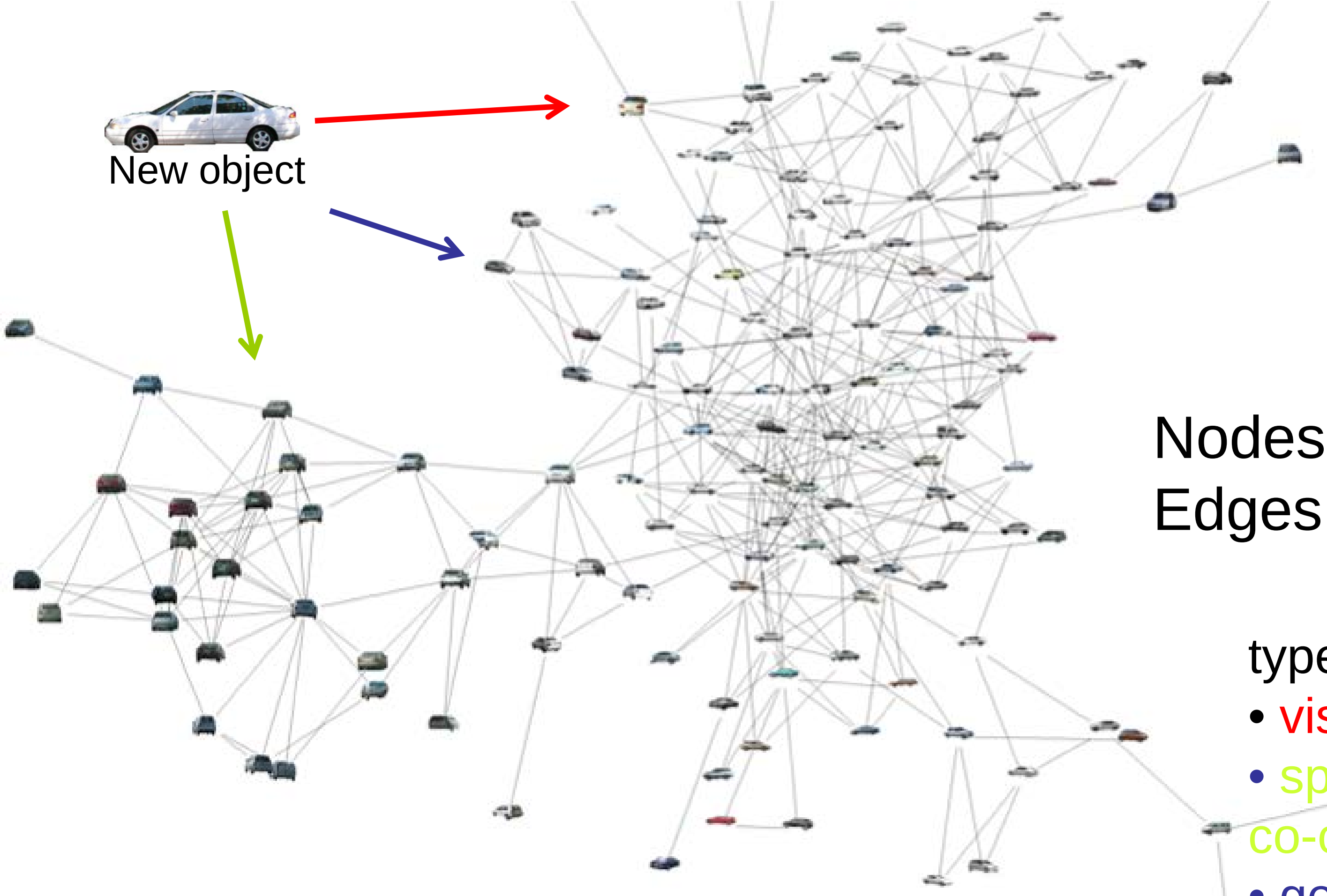
Bush’s Memex (1945)

- Store publications, correspondence, personal work, on microfilm
- Items retrieved rapidly using index codes
  - Builds on “rapid selector”
- Can annotate text with margin notes, comments
- Can construct a *trail* through the material and save it
  - Roots of hypertext
- Acts as an external memory



Visual Memex, a proposal

[Malisiewicz & Efros]



Nodes = instances
Edges = associations

types of edges:

- visual similarity
- spatial, temporal co-occurrence
- geometric structure
- language
- geography
- ..

Image Understanding via Memex

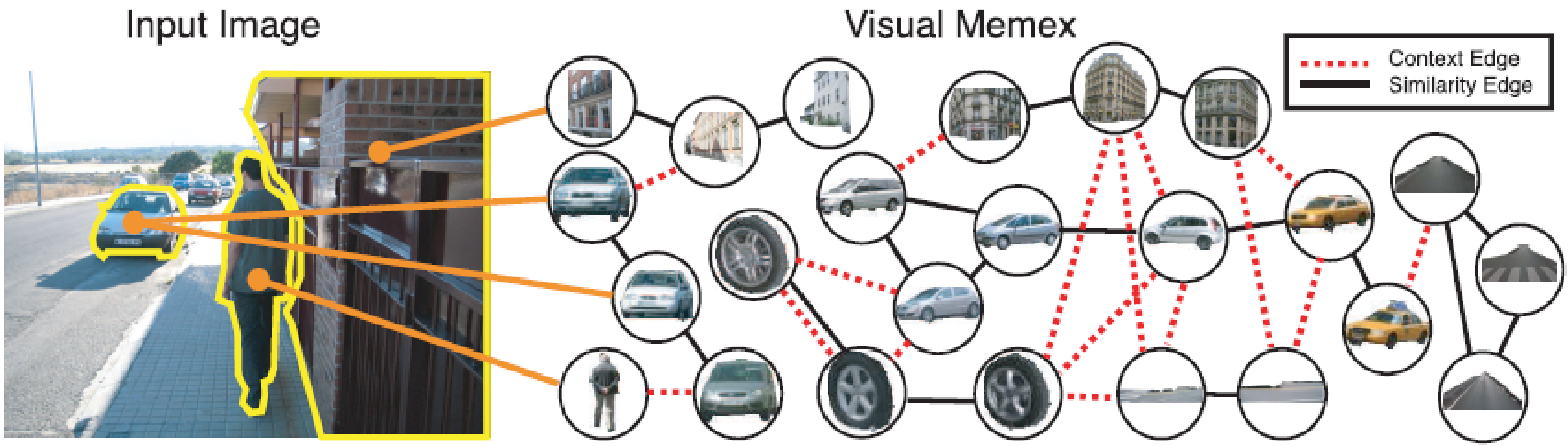


Figure 1: The **Visual Memex** graph encodes object similarity (solid black edge) and spatial context (dotted red edge) between pairs of object exemplars. A spatial context feature is stored for each context edge. The Memex graph can be used to interpret a new image (left) by associating image segments with exemplars in the graph (orange edges) and propagating the information.

Torralba’s Context Challenge

Torralba’s Context Challenge

Slide by Antonio Torralba

Torralba’s Context Challenge

Slide by Antonio Torralba

Our Challenge Setup

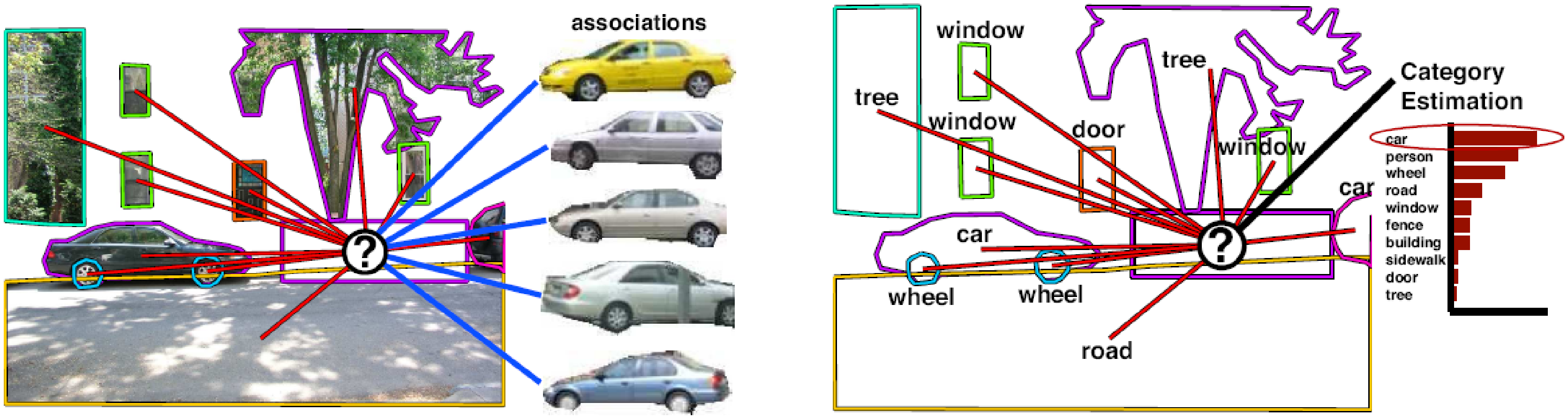


Figure 2: Torralba's Context Challenge: "How far can you go without running a local object detector?" The task is to reason about the identity of the hidden object (denoted by a "?") without local information. In our category-free Visual Memex model, object predictions are generated in the form of exemplar associations for the hidden object. In a category-based model, the category of the hidden object is directly estimated.

Malisiewicz & Efros, NIPS'09

3 models

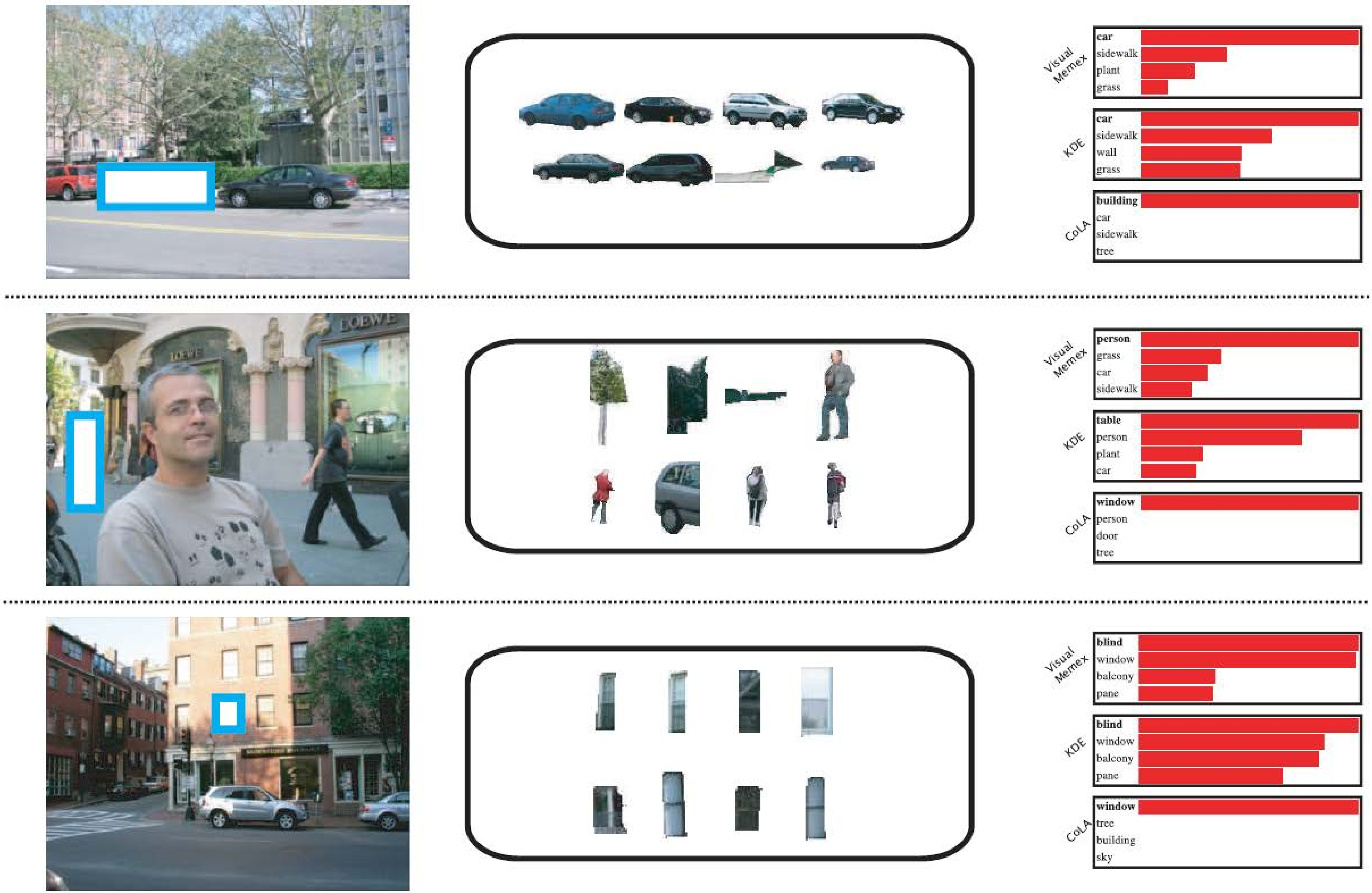
Visual Memex: exemplars, non-parametric object-object relationships

- Recurse through the graph

Baseline: CoLA: categories, parametric object-object relationships

Reduced Memex: categories, non-parametric relationships

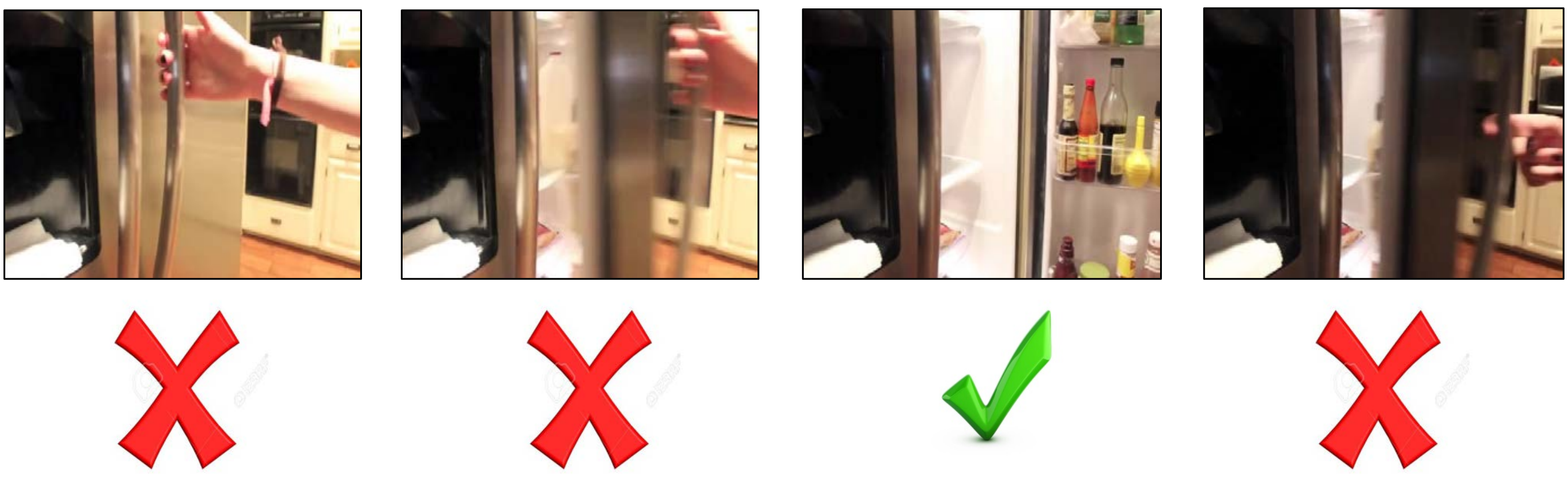
Qual. results


Visual Memex
car
sidewalk
plant
grass
KDE
car
sidewalk
wall
grass
CoLA
building
car
sidewalk
tree
Visual Memex
person
grass
car
sidewalk
KDE
table
person
plant
car
CoLA
window
person
door
tree
Visual Memex
blind
window
balcony
pane
KDE
blind
window
balcony
pane
CoLA
window
tree
building
sky

Part 4: Limitations of Nearest Neighbors

Are we fooling ourselves?

- E.g. action recognition
  - Very hard to improve on single frame classifiers
  - Consider “opening fridge” action:



Dataset bias is a problem, but so is our complacency

example by David Fouhey

Thank You

© Quint Buchholz