On the Tyranny of "Elegant" ideas

or

why I like Nearest Neighbors

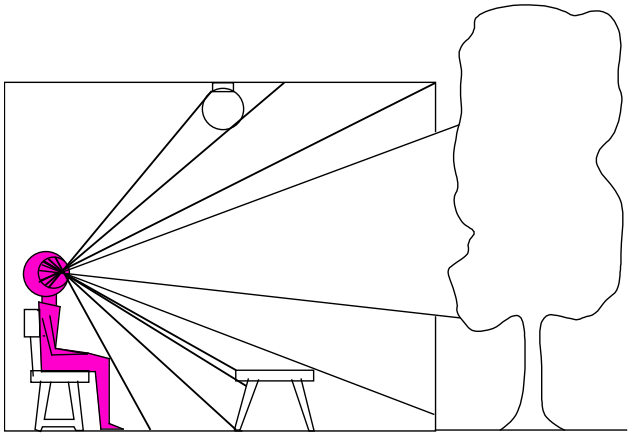
"For every complex problem there is an answer that is clear, simple, and wrong."
-- H. L. Mencken

Alyosha Efros

UC Berkeley

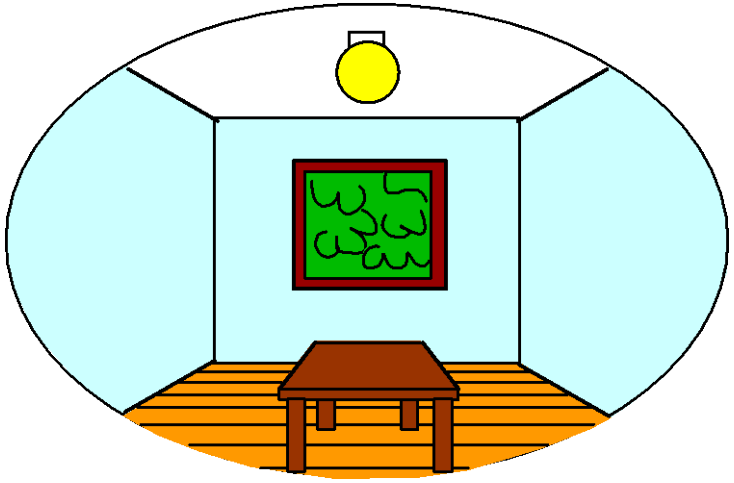
How do humans see 3D?

3D world



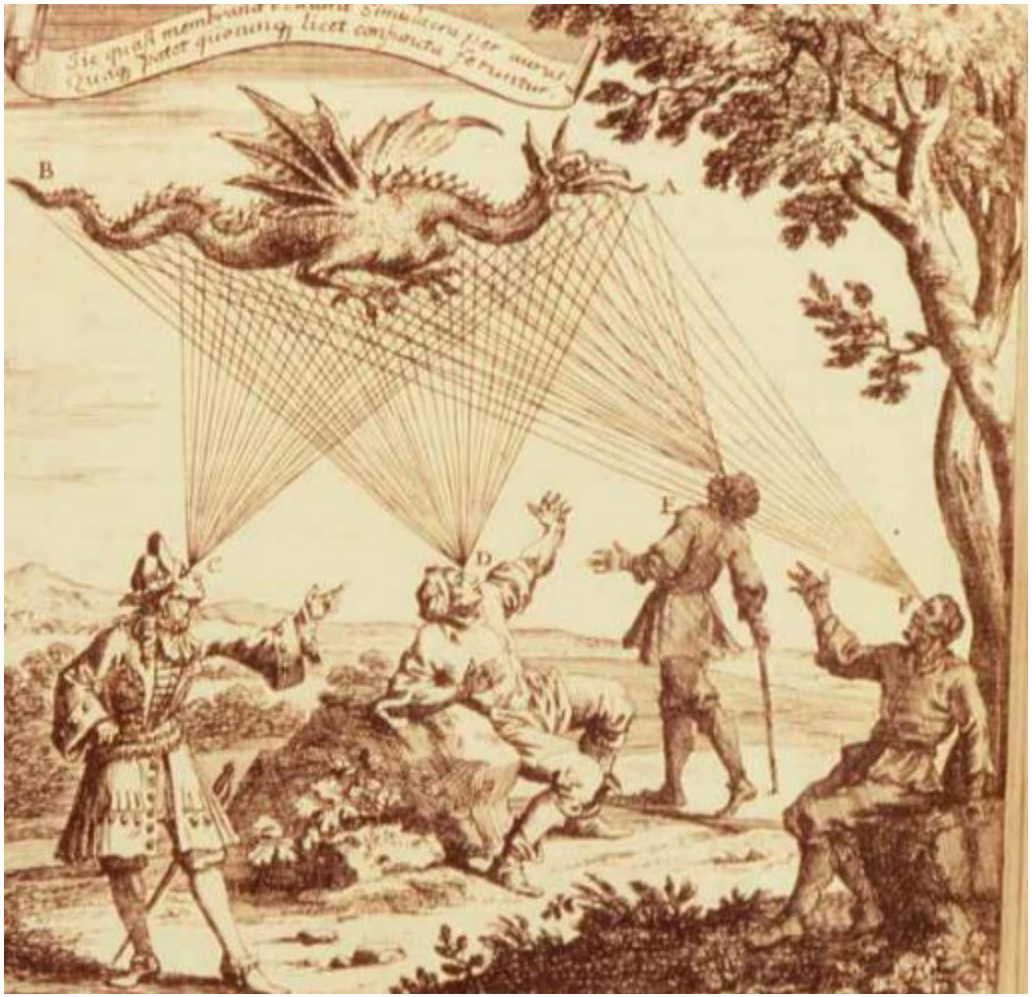
Point of observation

2D image



But there is a problem…

Emission Theory of Vision



Eyes send out “feeling rays” into the world

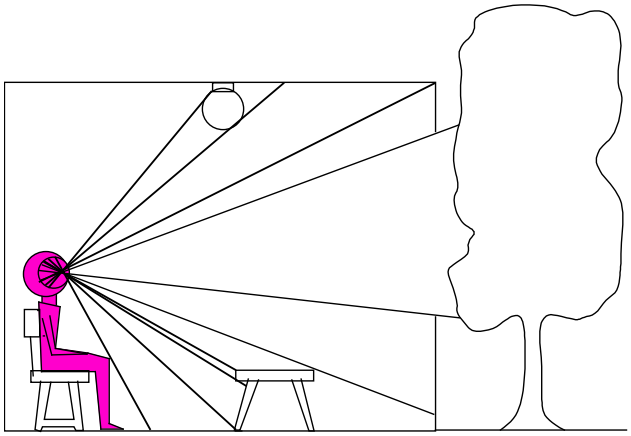
Supported by:

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

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

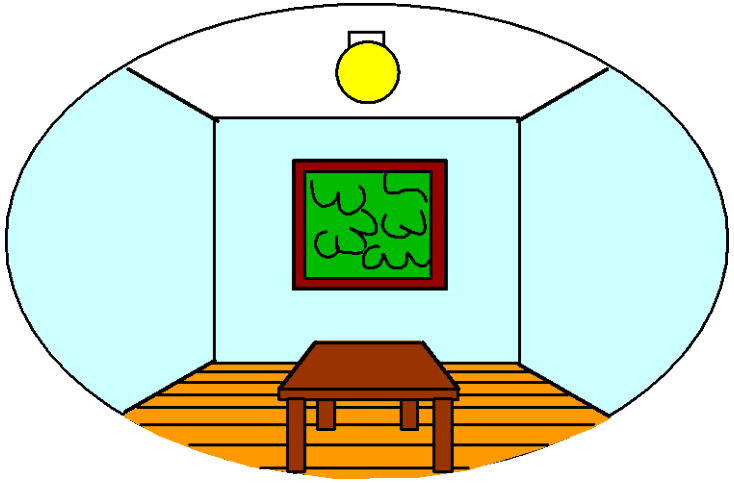
Truth is often ugly

3D world



Point of observation

2D image



But we still prefer elegant answers…

- As CVPR reviewer, which paper would you rather accept?

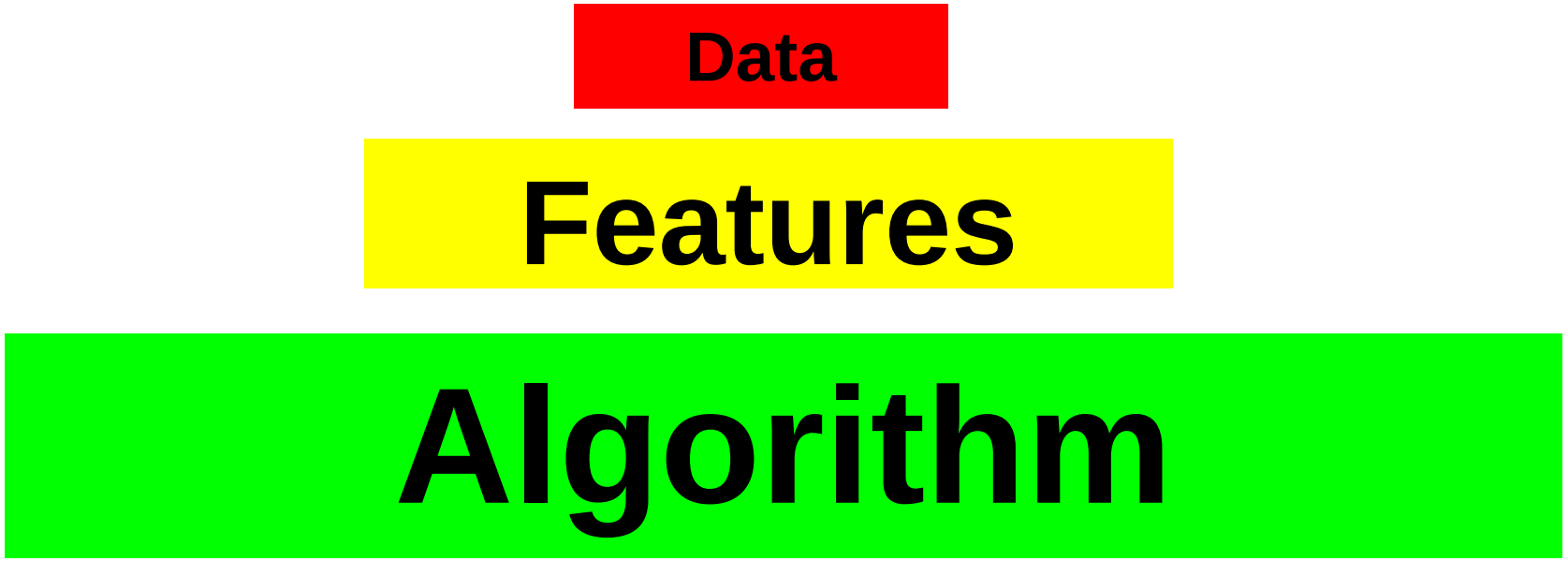
- Emission theory:
  - “mathematically sound formulation”
  - “Sophisticated approach”
  - “principled”

- What actually happens:
  - “inelegant”
  - “not principled”
  - “just a dataset”

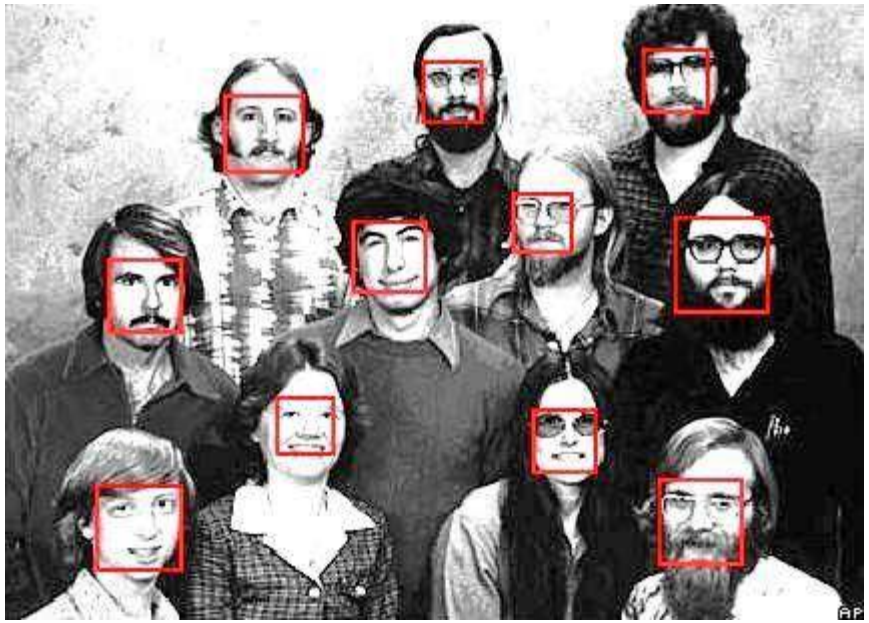
Our Scientific Narcissism

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**

Word embeddings

- word2vec
- Matrix factorization
- Normalized Nearest Neighbors

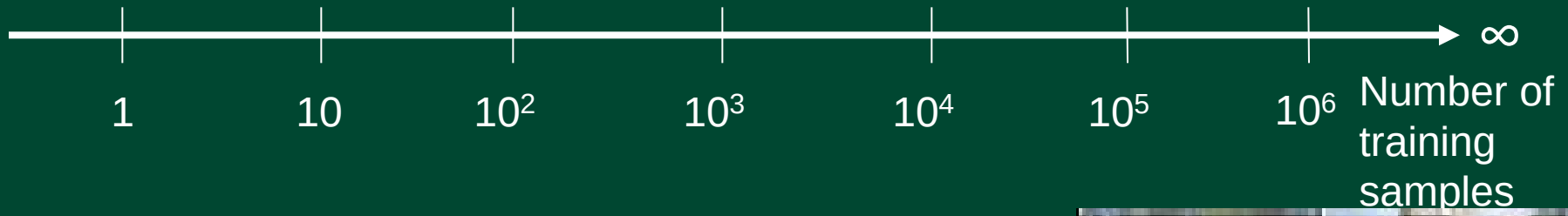
Image captioning

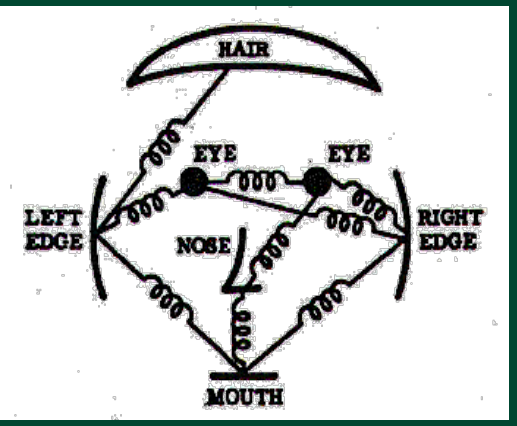
- LSTMs
- Feed-forward CNNs
- Language models
- …
- Nearest neighbors

Learning Spectrum

Extrapolation problem
Generalization

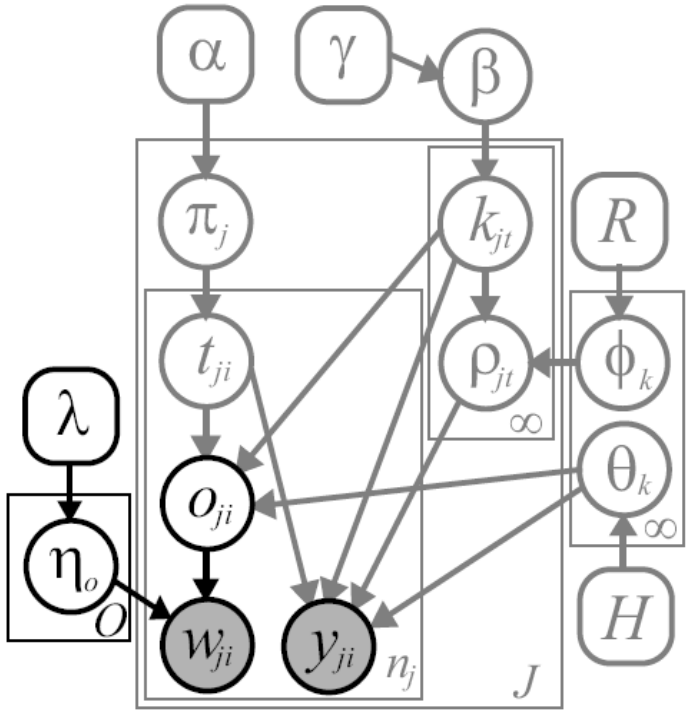
Interpolation problem
Correspondence



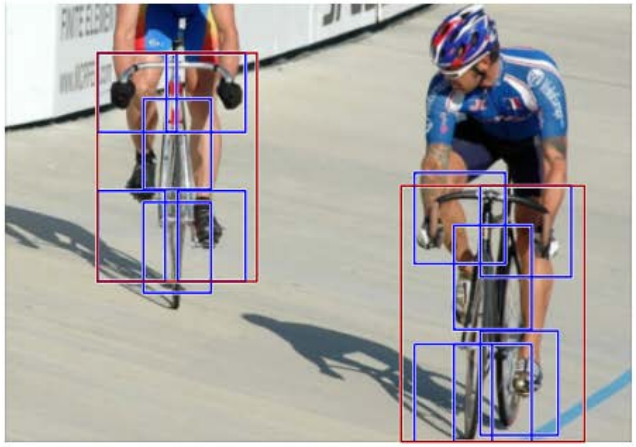


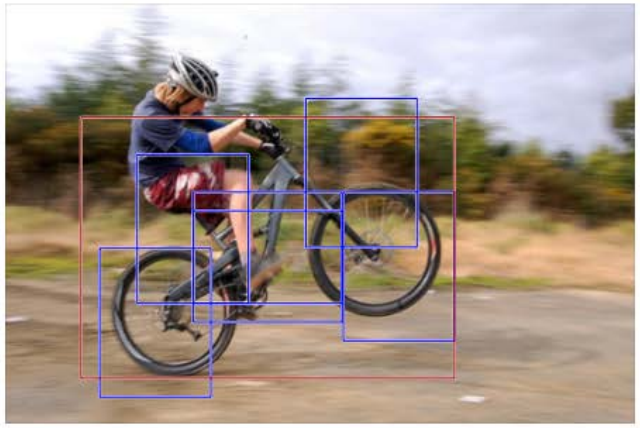
Slide by Antonio Torralba

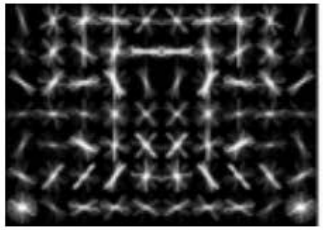
Part 1: Nearest Neighbors as a negative result

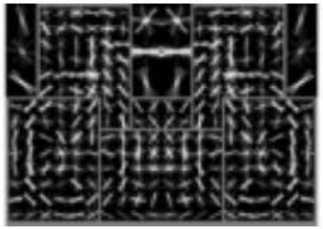


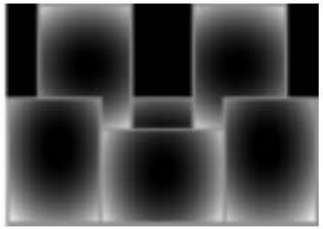
Deformable Part Models

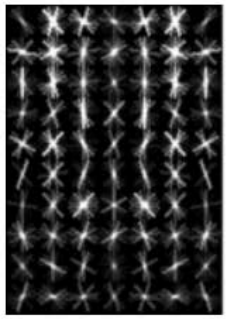


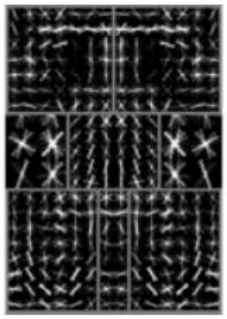


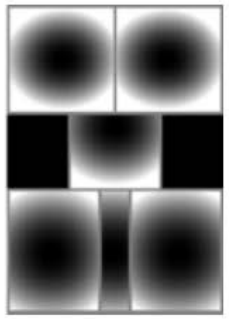


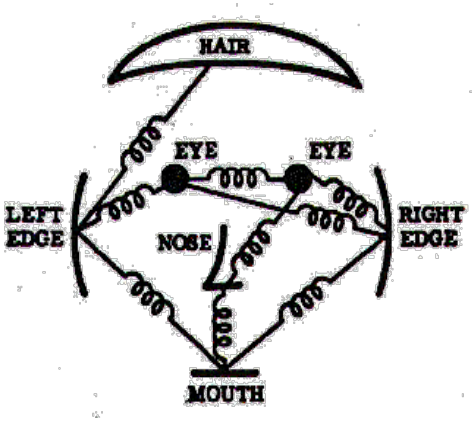










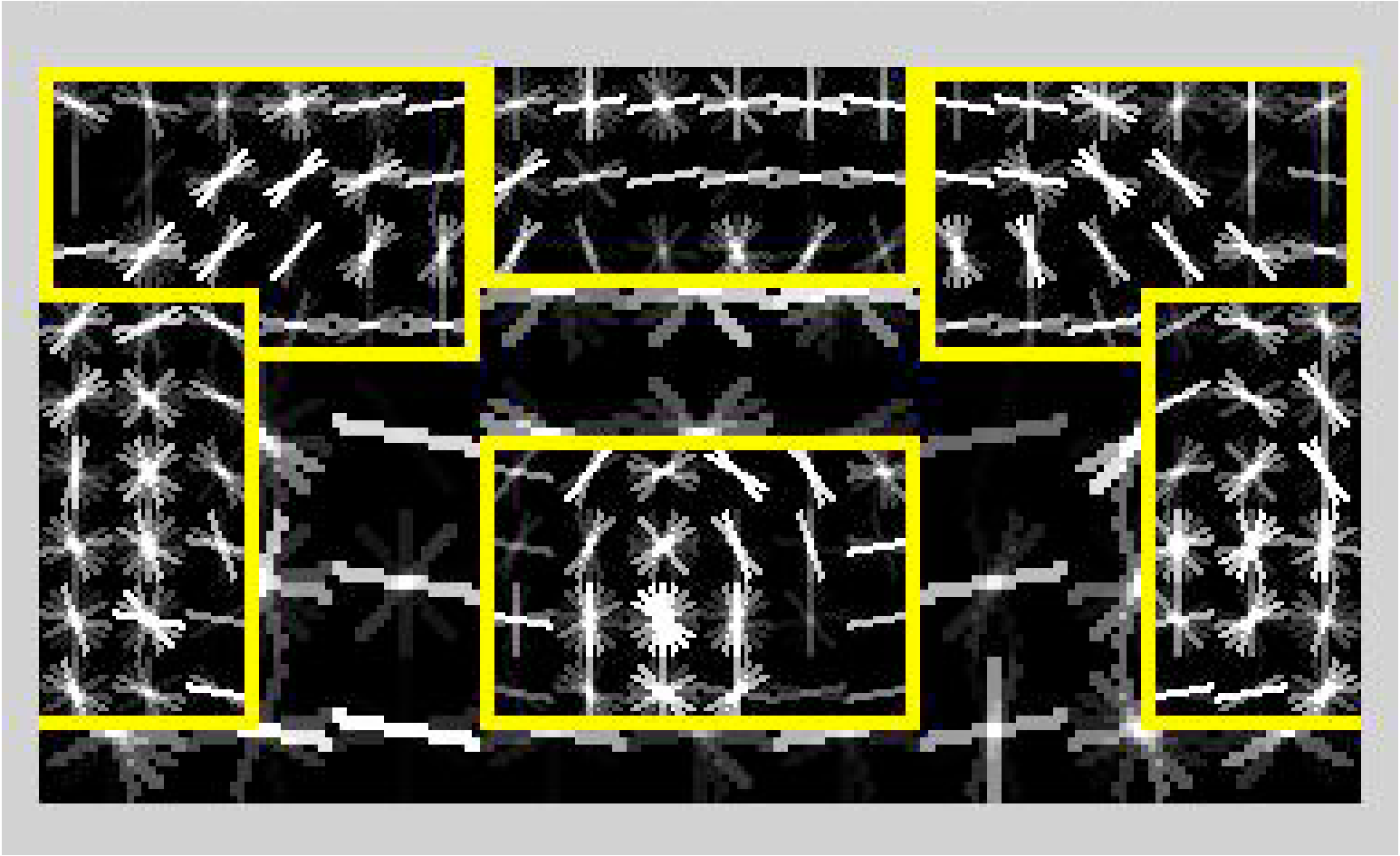


How important are "Deformable Parts" in the Deformable Parts Model?

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

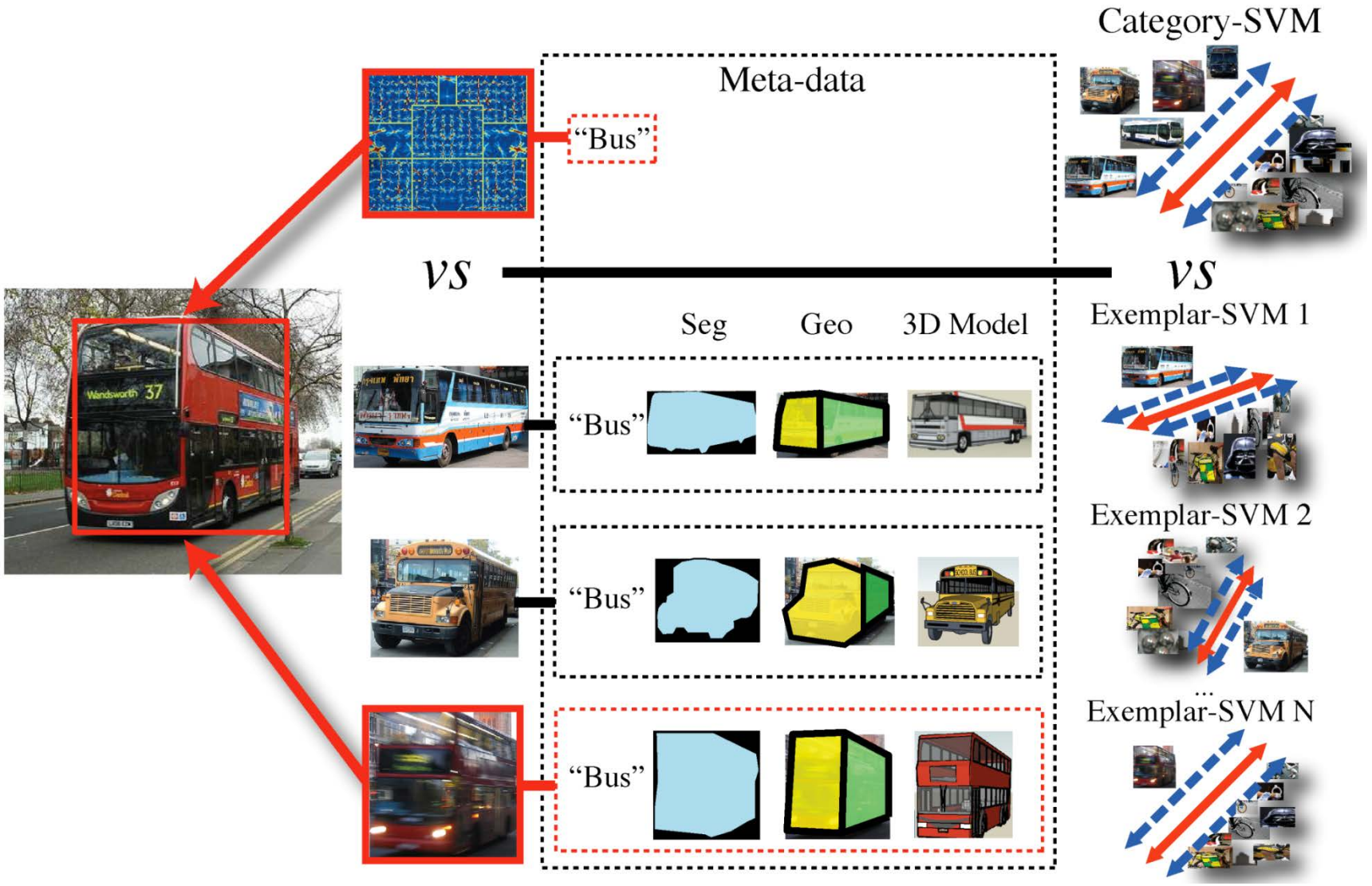
Robotics Institute, Carnegie Mellon University.

Typical HOG car detector



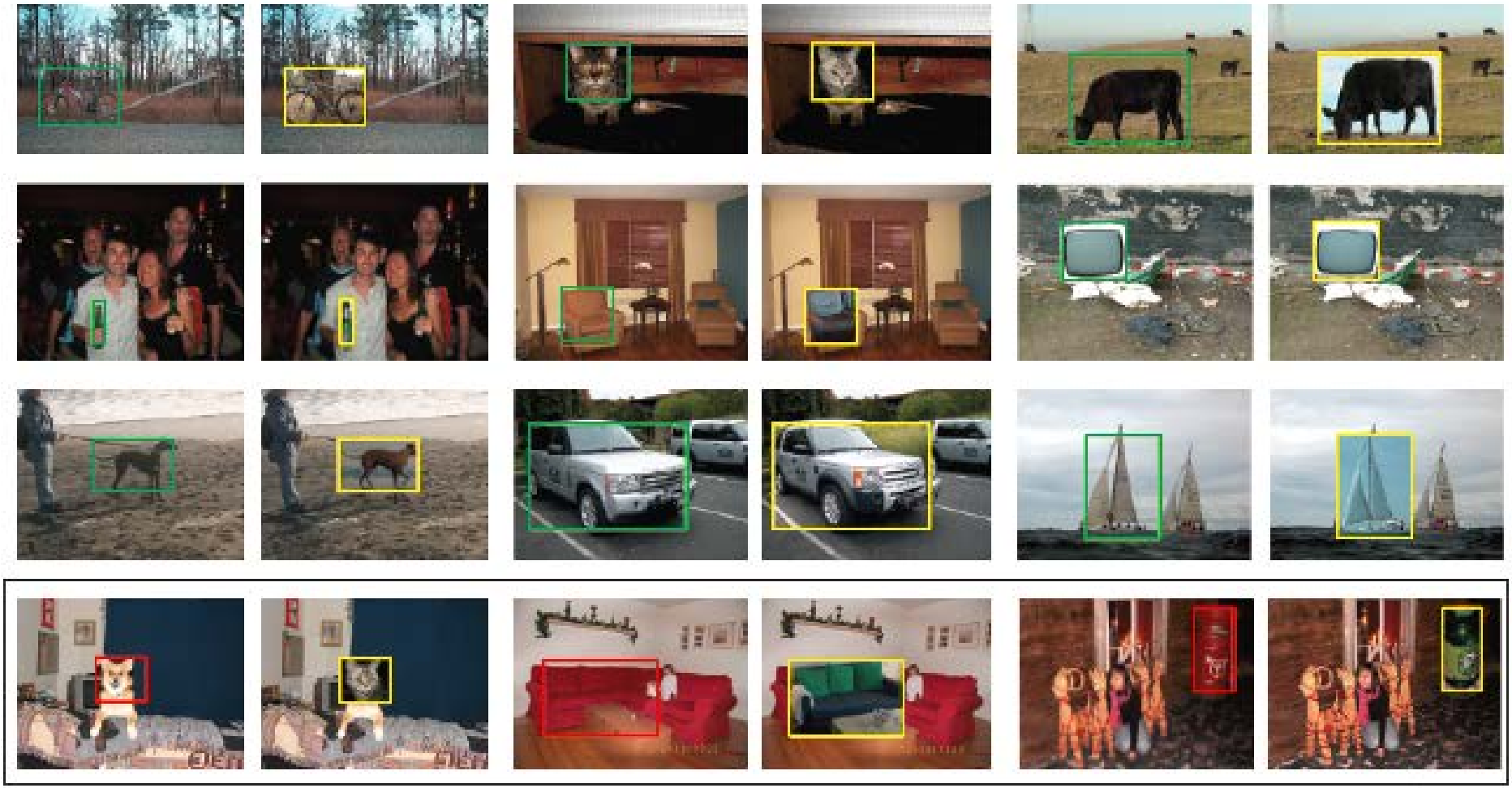
Felzenszwalb et al, PASCAL 2007

Exemplar-SVMs



Malisiewicz et al, ICCV'11

Showing off correspondences

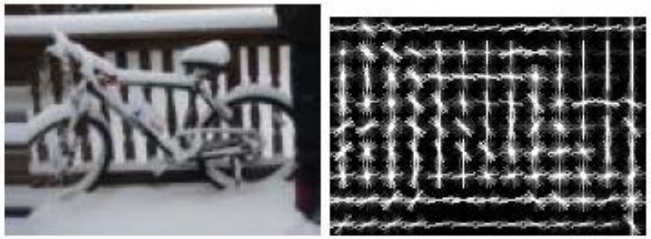


Malisiewicz et al, ICCV'11

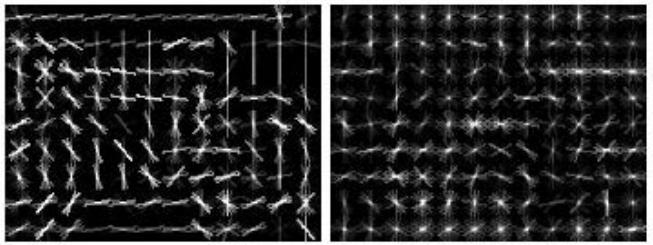
Discriminative Decorrelation for Clustering and Classification*

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

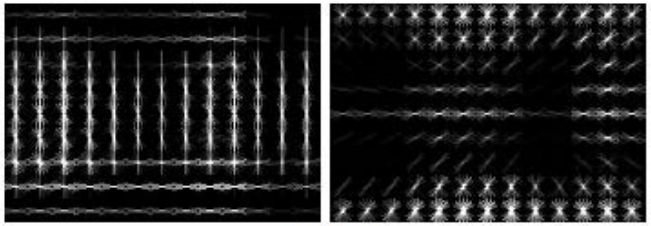
[1] Univerisity of California at Berkeley, Berkeley, CA, USA
{bharath2,malik}@cs.berkeley.edu
[2] University of California at Irvine, Irvine, CA, USA
dramanan@ics.uci.edu



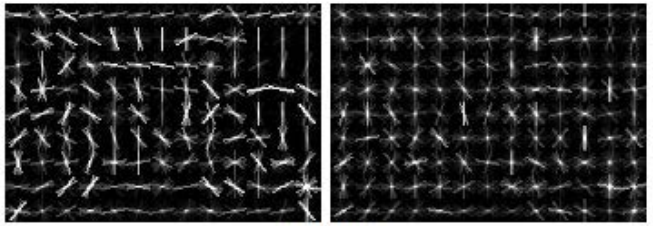
(a) Image (left) and HOG (right)



(b) SVM

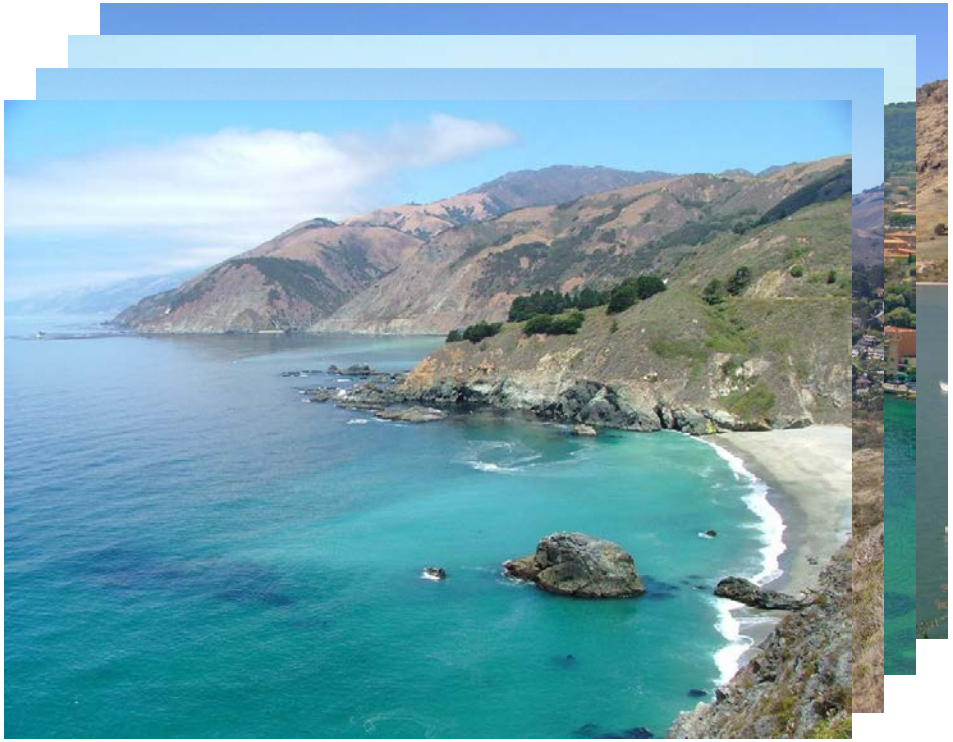


(c) PCA



(d) LDA

Brute Force Scene Understanding



Label Transfer

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

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

CC-BY-NC by stevekc

CC-BY-NC by edwin.11

CC-BY-NC by jonathanfh

(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%**

“Unreasonable Effectiveness of Data”

[Halevy, Norvig, Pereira 2009]

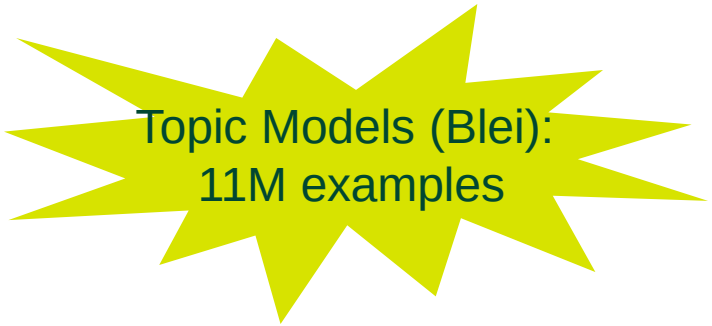
- Parts of our world can be explained by elegant mathematics:
  - physics, chemistry, astronomy, etc.
- But much cannot:
  - psychology, genetics, economics, etc.

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

Everything else being equal…

… the natural world is just much **richer**!

- 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!)**

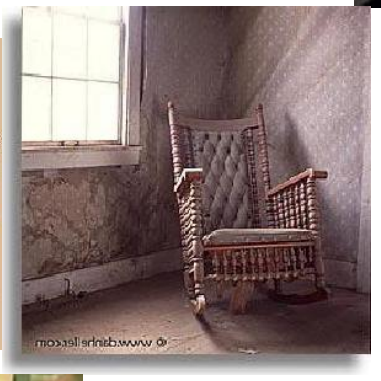


Yet, we train on 15 examples?!











Caltech 101

Google™

A.I. for the postmodern world

The Good News

Really stupid algorithms + Lots of Data

= *“Unreasonable Effectiveness”*

Lessons learned so far…

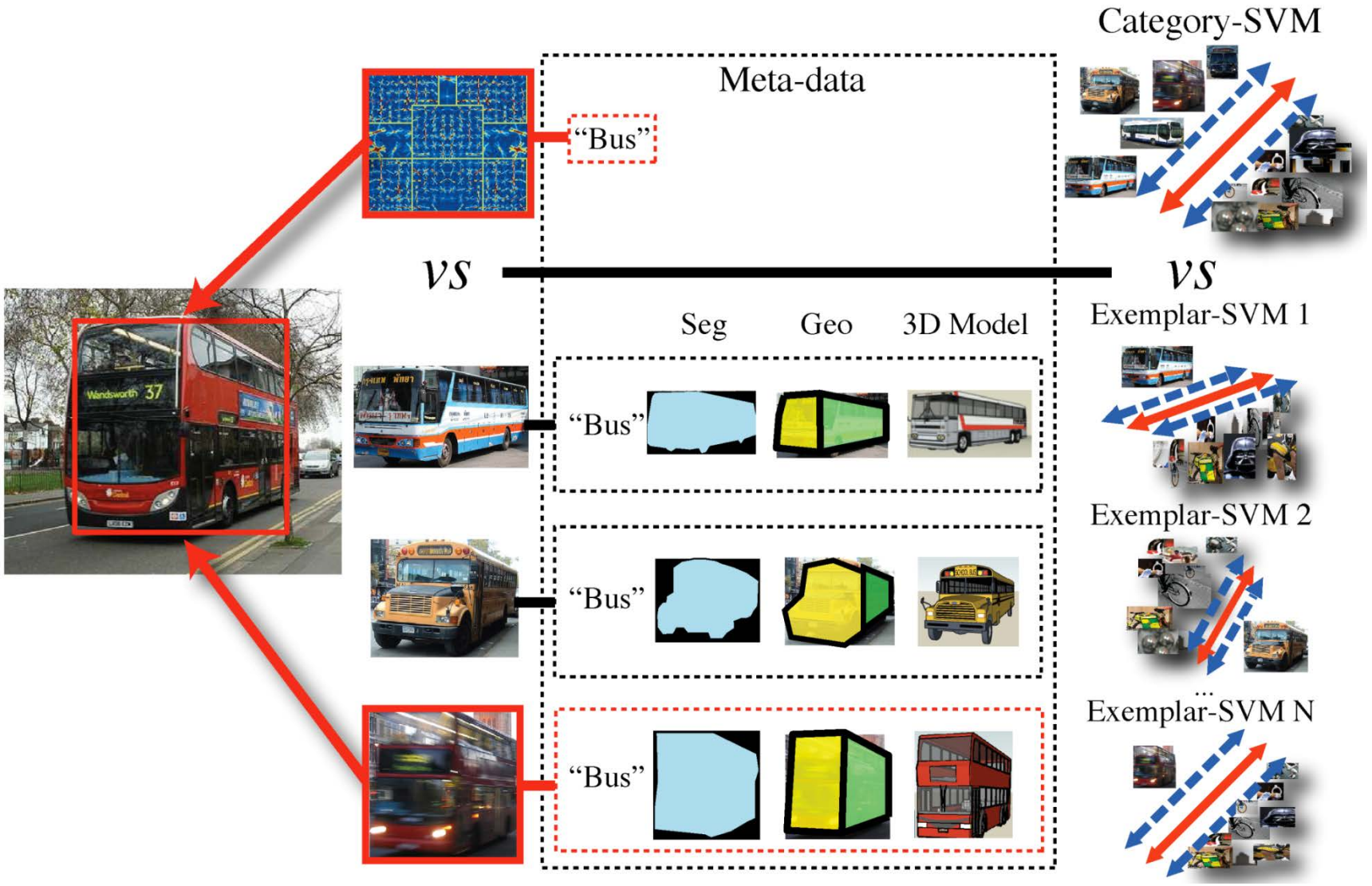
1. “Please don’t explain – Show me!”

Takeo Kanade's top 3 vision problems:

- Correspondence!
- Correspondence!!
- Correspondence!!!*

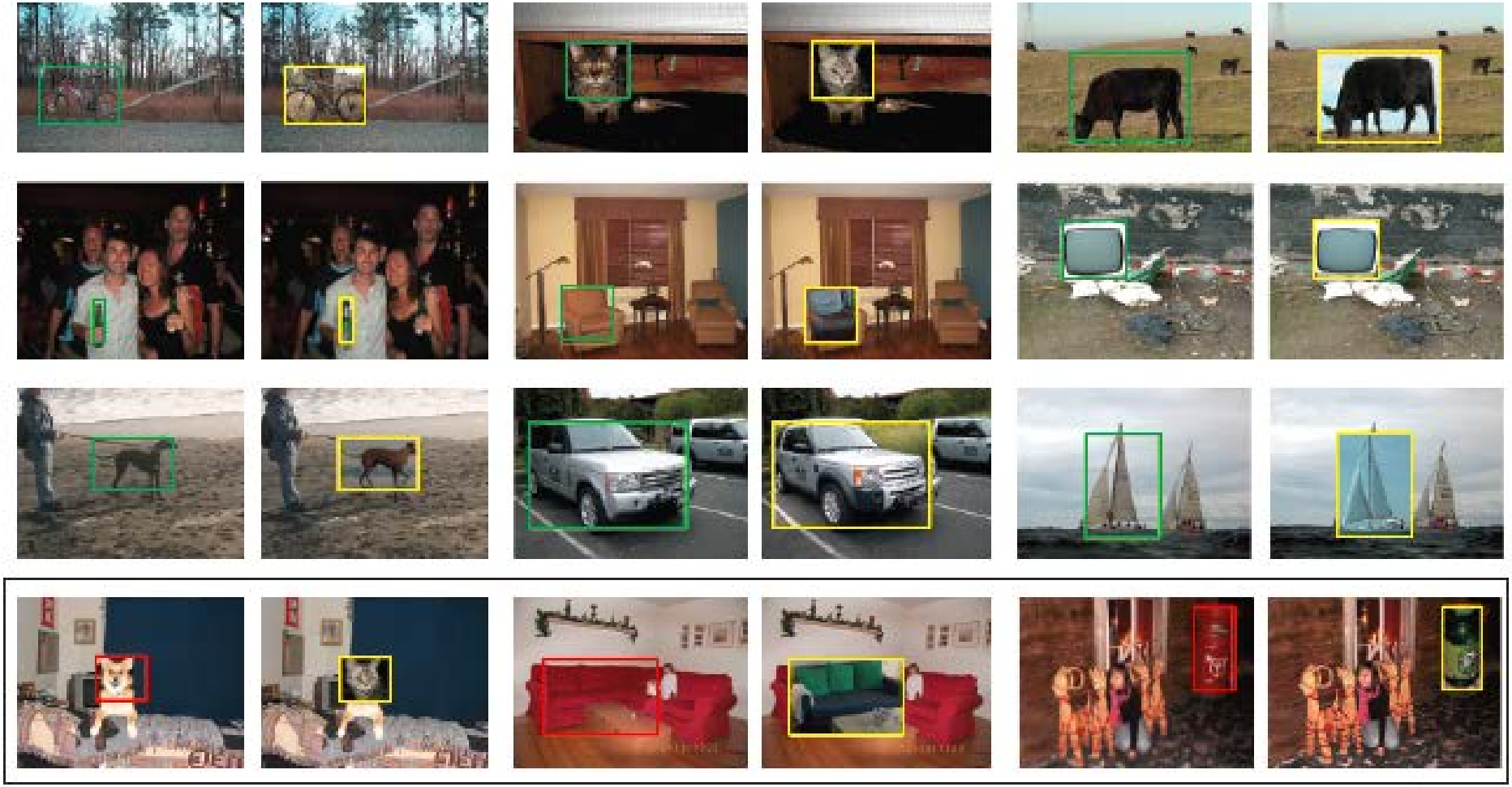
** actually, back in the day, it was "registration", but we argue that correspondence, alignment, and registration are all talking about the same issue*

Exemplar-SVMs



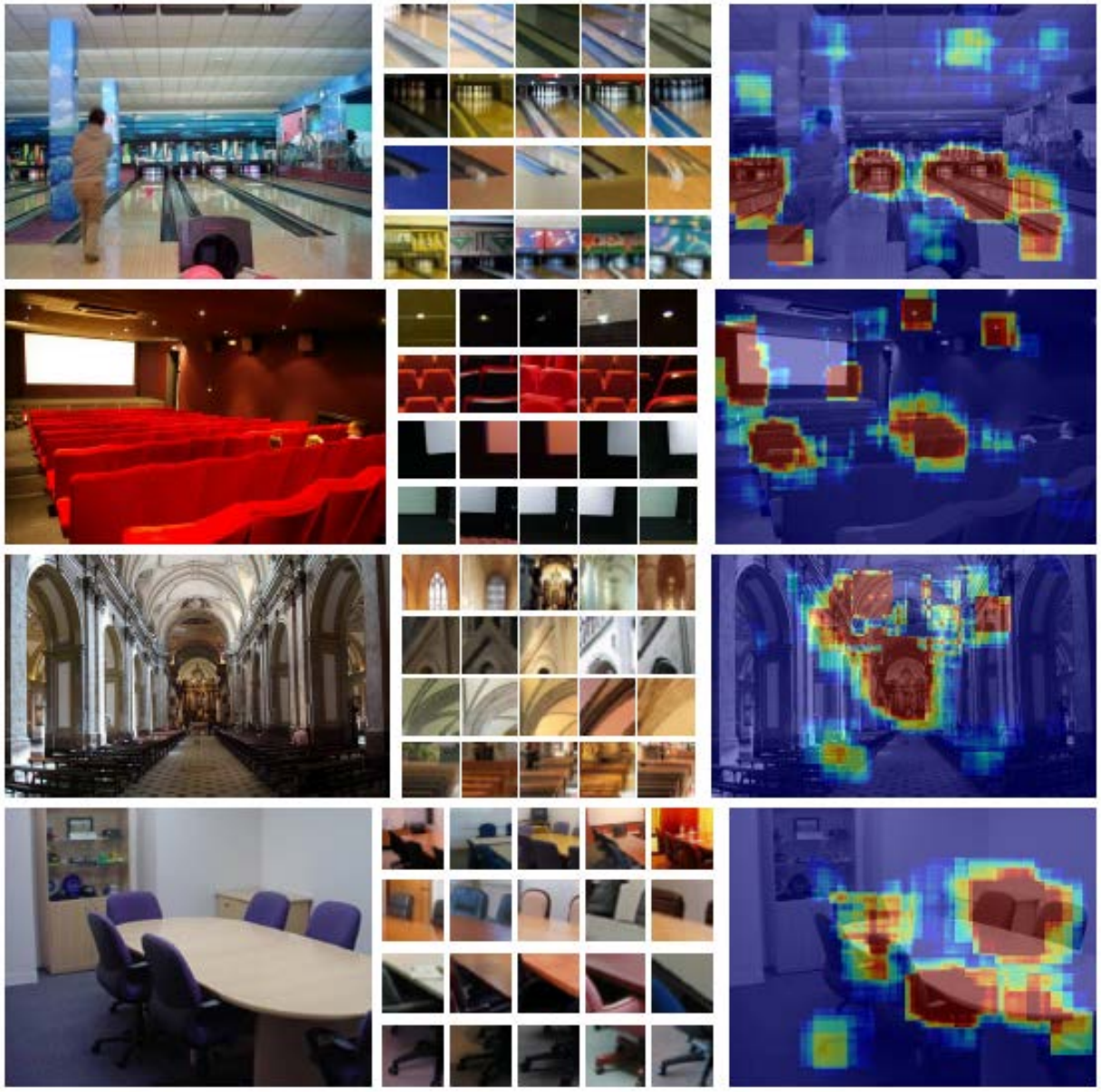
Malisiewicz et al, ICCV'11

Showing off correspondences



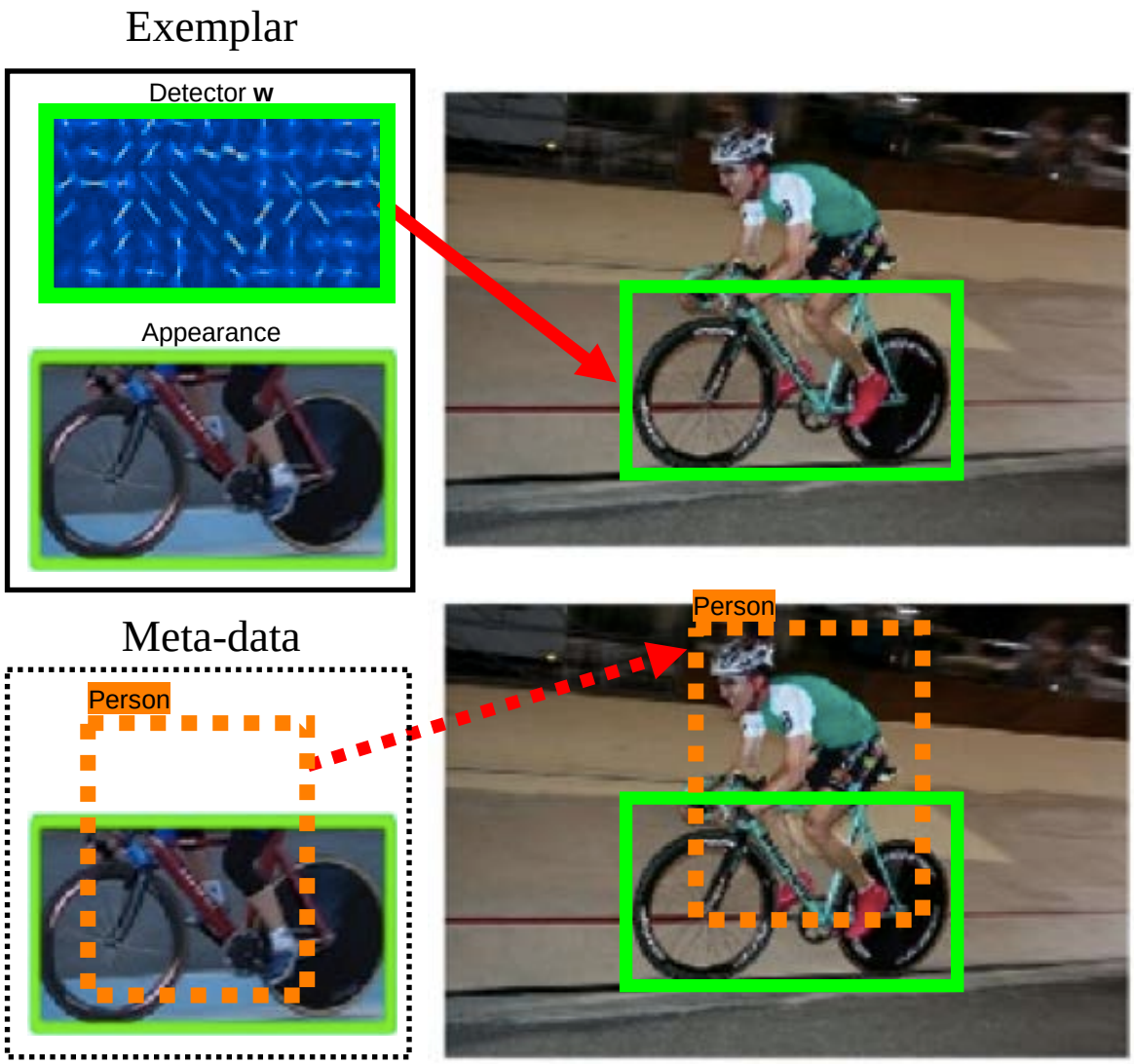
Malisiewicz et al, ICCV'11

Scene Classification, demystified



Doersh et al, NIPS 2013

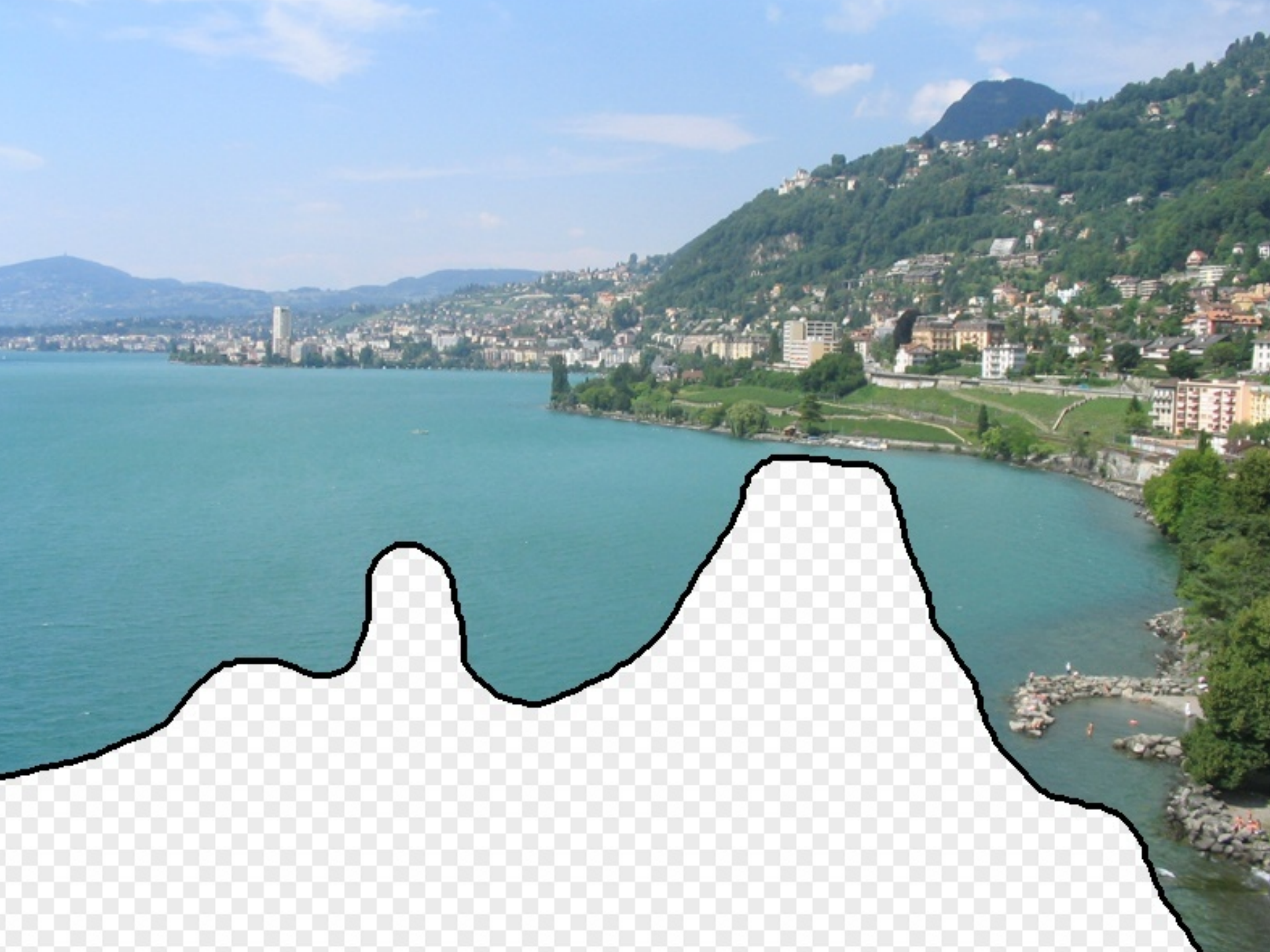
Can make predictions

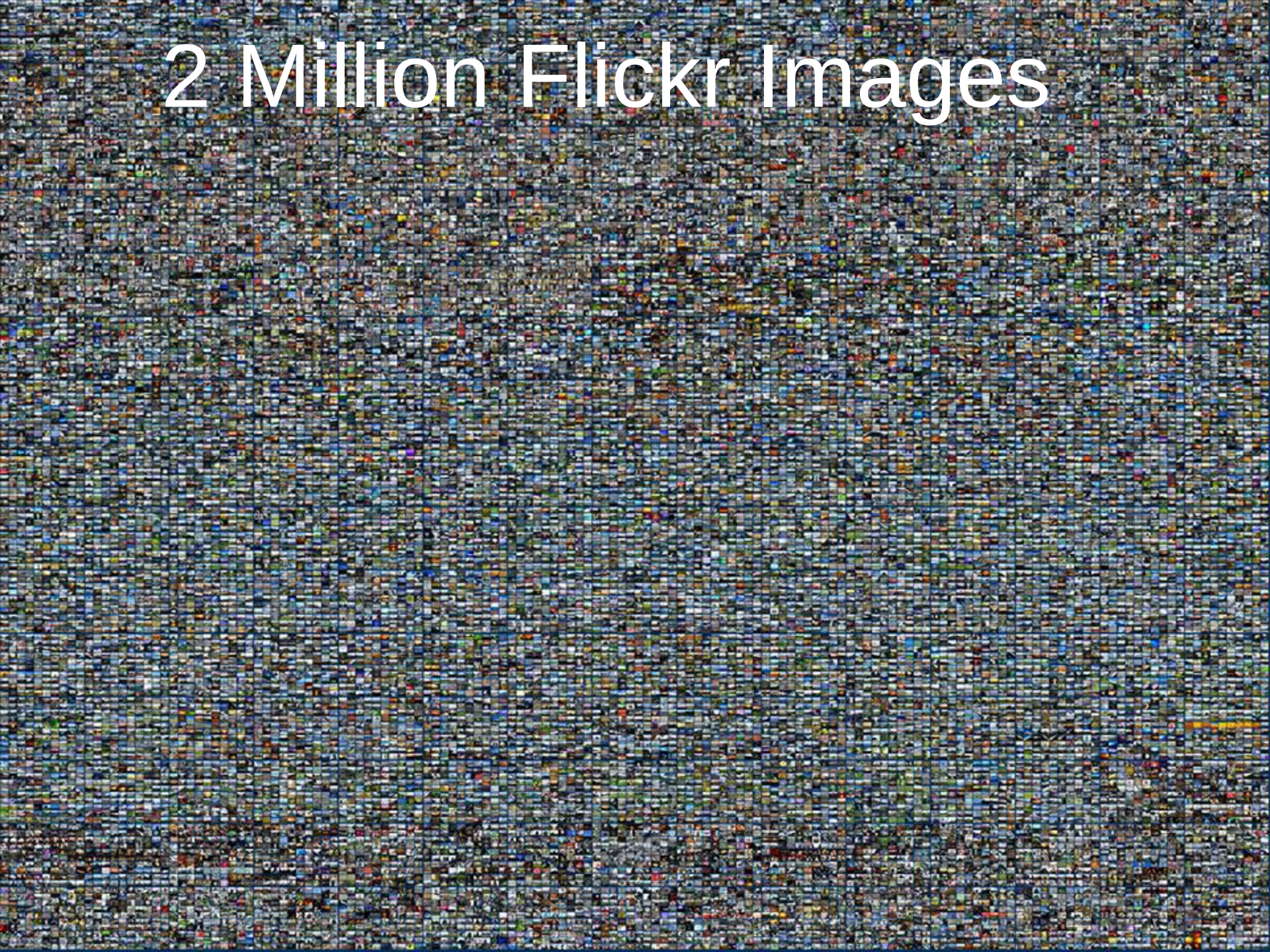


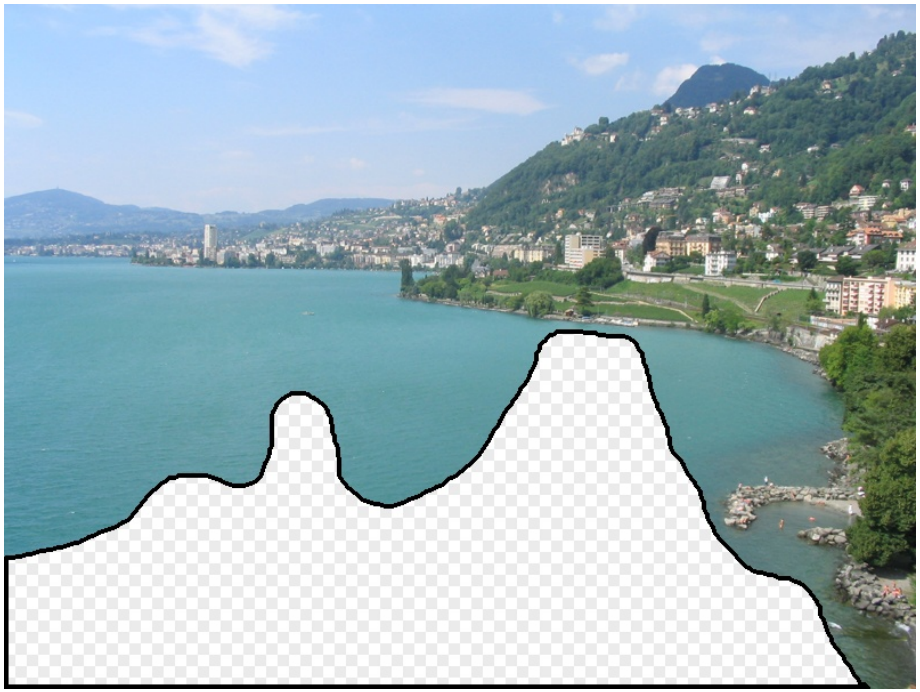
Malisiewicz et al, ICCV'11

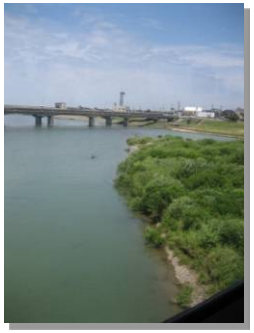
2. “Size Does Matter”

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



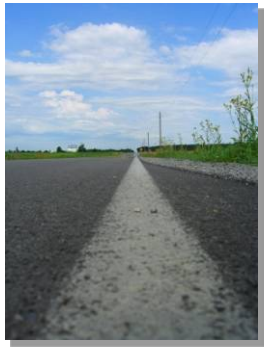

2 Million Flickr Images



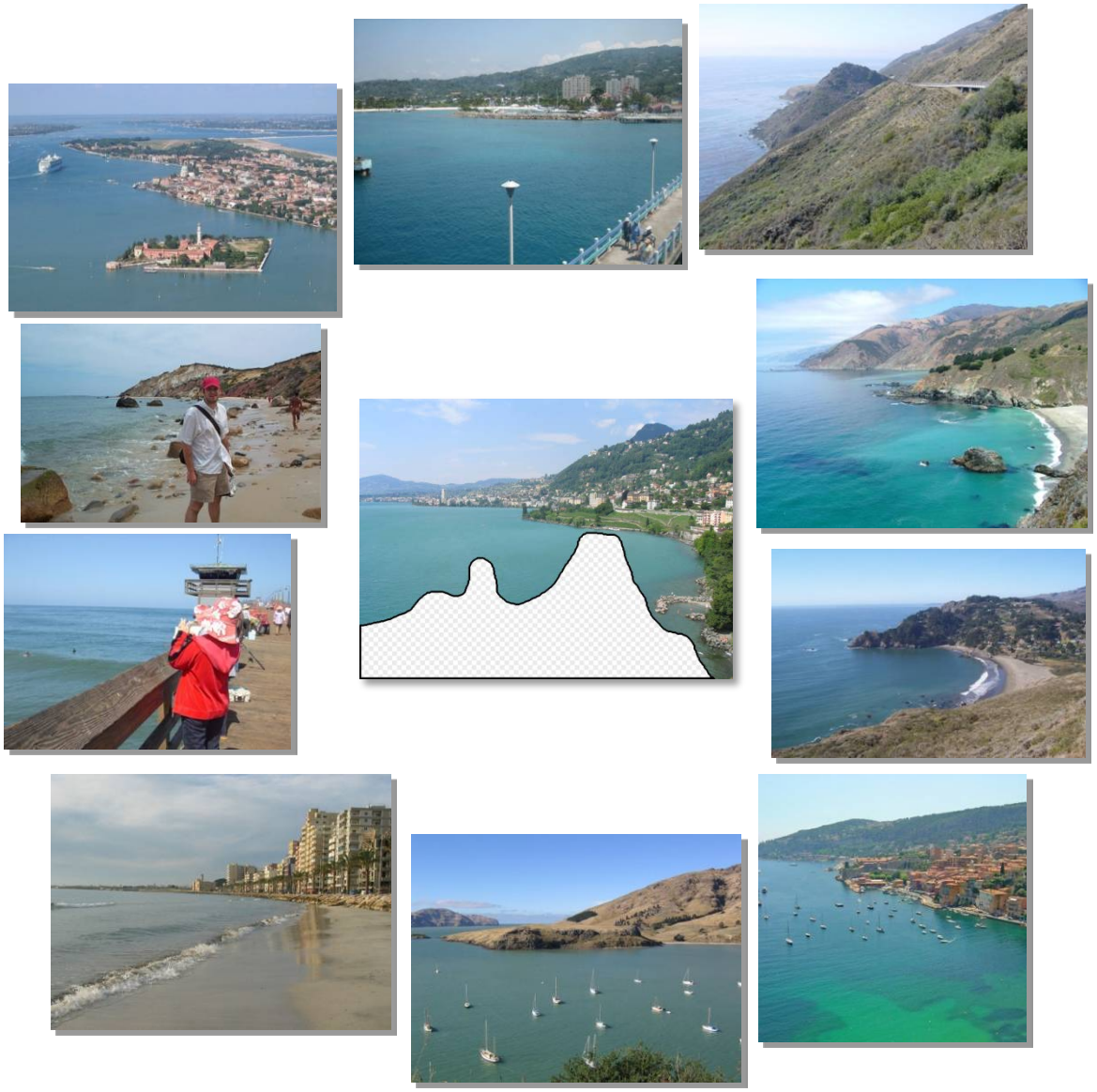






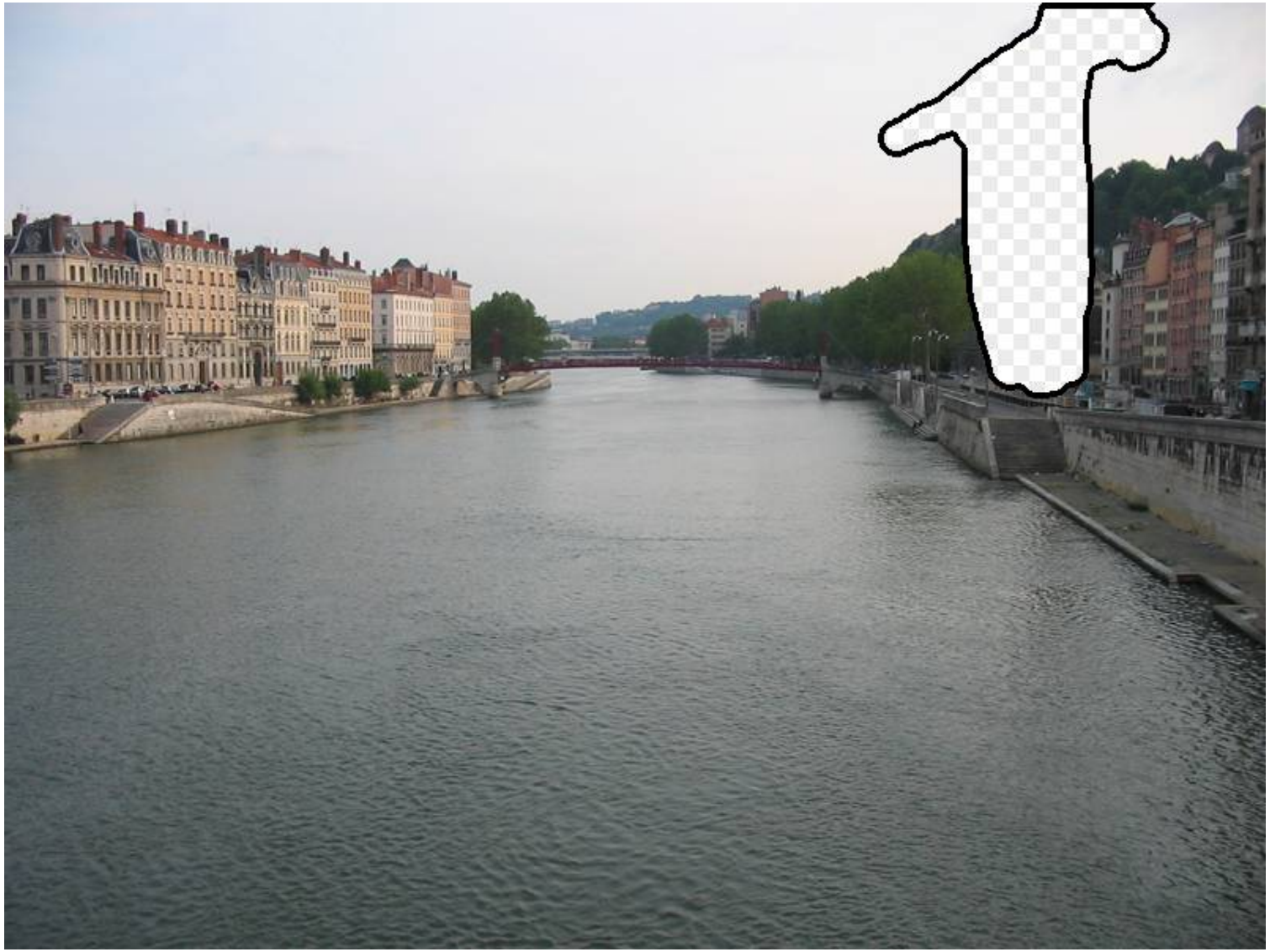

Nearest neighbors from a
collection of 20 thousand images

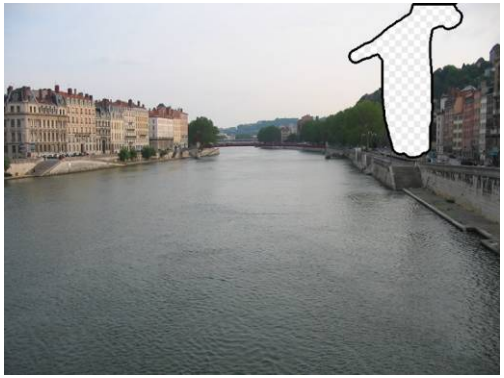


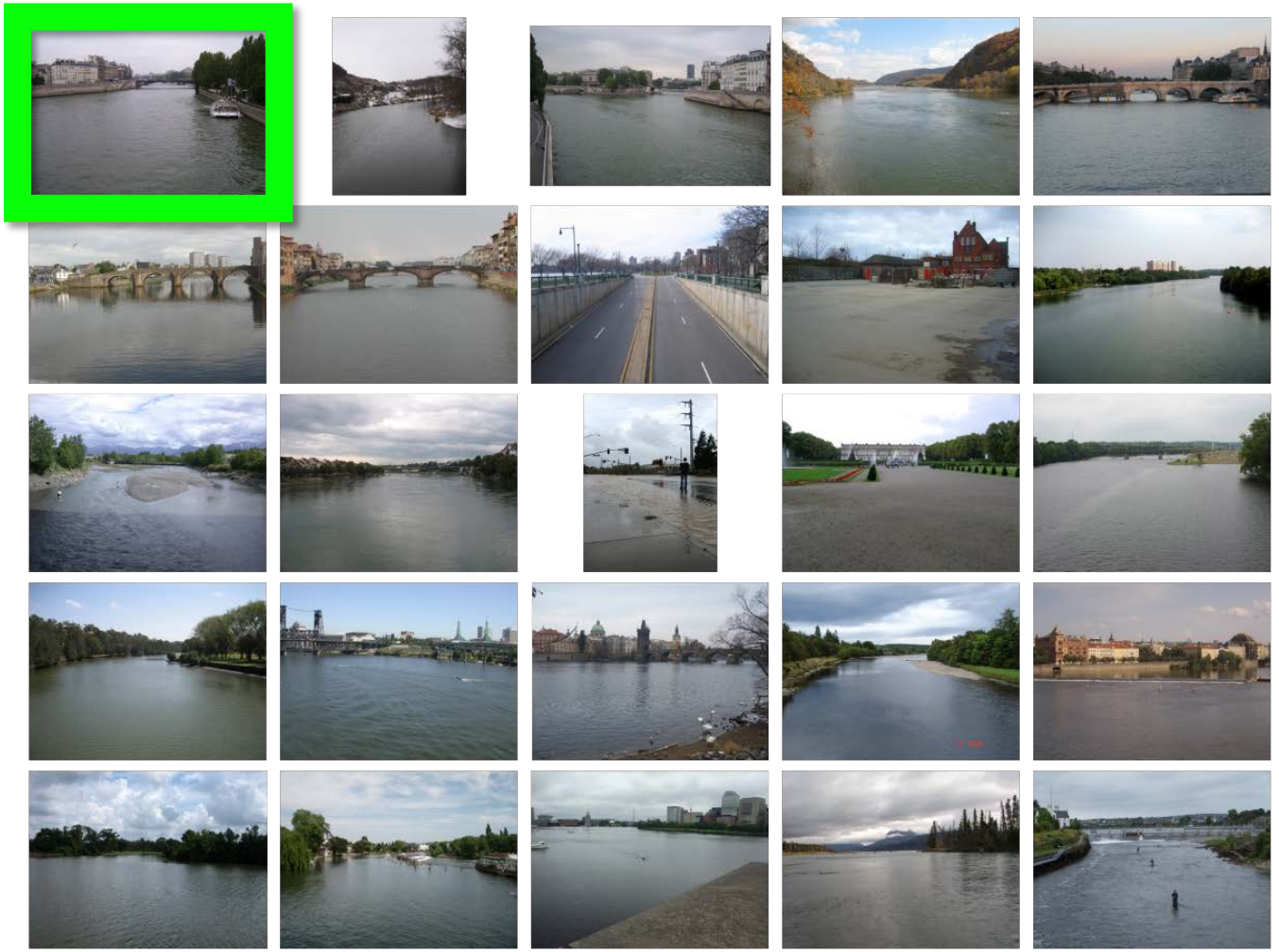
Nearest neighbors from a
collection of 2 million images

3. “Want a bigger party? Shrink the venue!”

Making Closed Worlds

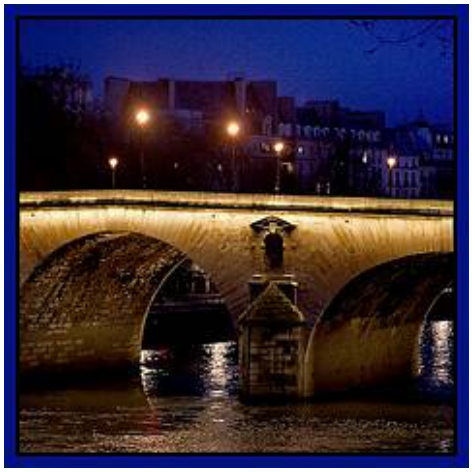






… 200 scene matches

Flickr Paris



Google StreetView Paris

Other obvious examples

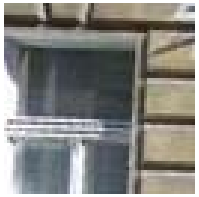
- Webcams
- First-person (robot) data
- Memes
  - E.g. http://thehairpin.com/2011/01/women-laughing-alone-with-salad

4. “Don’t give a man a fish – teach him how to fish”

Don’t give the computer the answer (label) right away, may it suffer first…

Paris vs. Not Paris



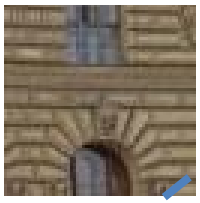










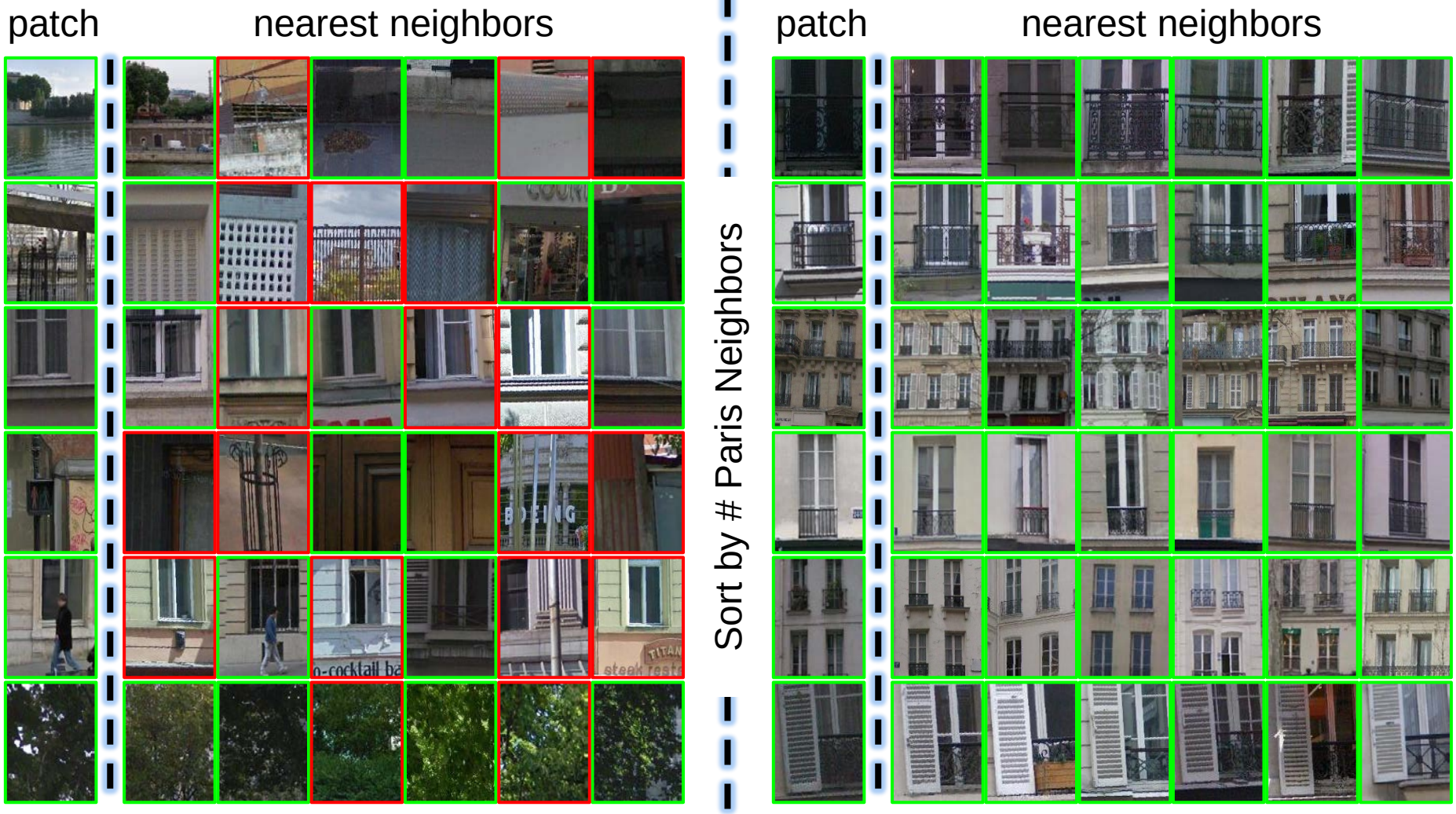
Step 1: Nearest Neighbors for Every Patch

Using normalized correlation of HOG features as a distance metric

Paris

Not Paris

Step 2: Find the Parisian Clusters by Sorting



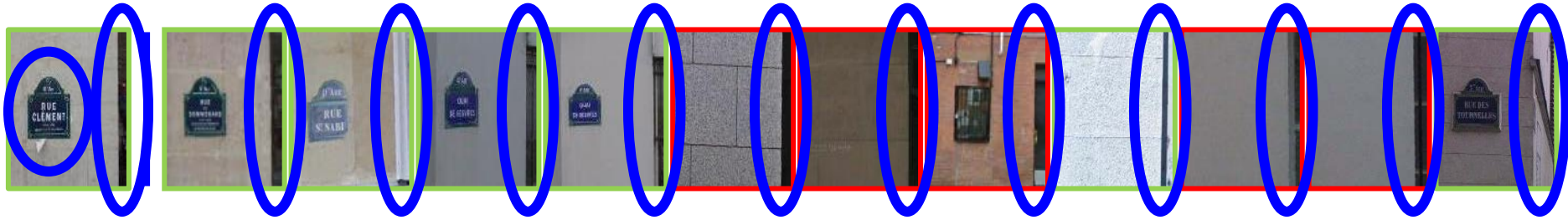
5. Data is the glue trying everything together

Features / Data / Algorithms

Feature + Data = better feature

Good Patches may have Bad Neighbors!

patch matches



- The naïve distance metric gives equal weight to the vertical bar and the sign.

Paris
Not Paris

Iterate using the new matches

7. “Keep the computer sufficiently stressed”

Latent SVMs getting too cozy…

Initial

Final

Initial

Final

Cross-validation Training

KMeans

Iter 1

Iter 2

Iter 3

Iter 4

Cross-validation Training

KMeans

Iter 1

Iter 2

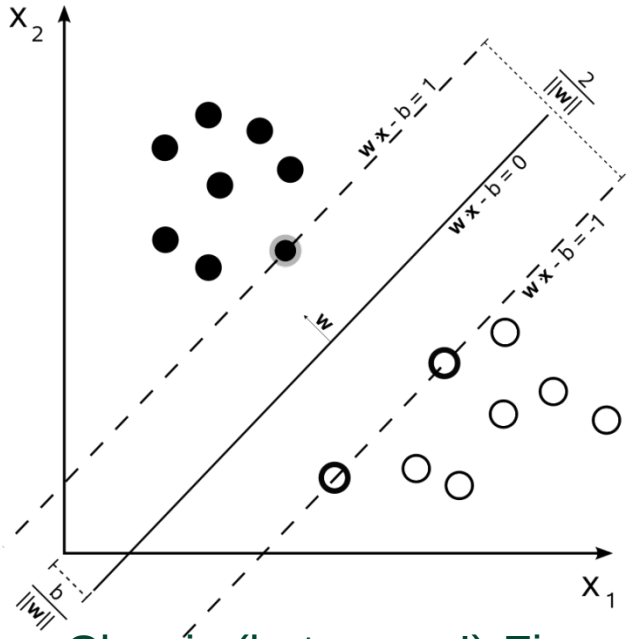
Iter 3

Iter 4

8. Trust in Data – don’t categorize

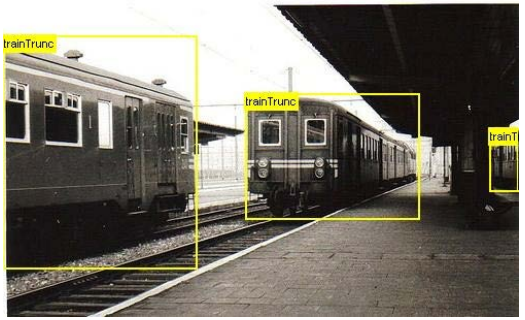
The Myth of Semantic Category Generalization

- In non-linear SVMs:
  - In textbooks, ~10% of data are support vectors
  - In recognition, up to 100% of data are support vectors!!!
- In linear SVMs:
  - Typical setup: 10,000D feature. but only 300 “chair” examples



Classic (but wrong!) Figure

Do we really need this linguistic “helper”



“train” category (PASCAL VOC)

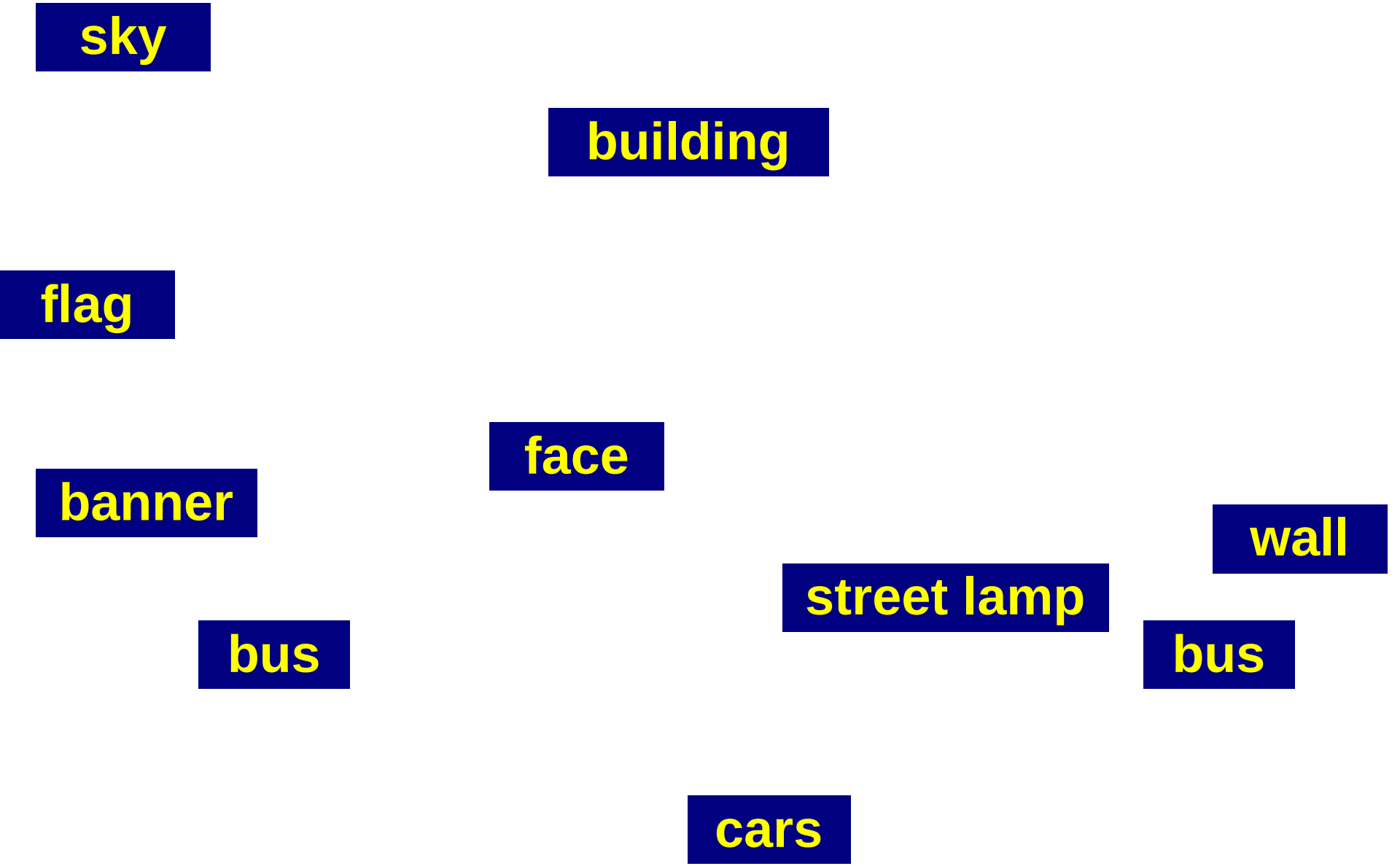
Understanding an Image

slide by Fei Fei, Fergus & Torralba

Object naming -> Object categorization

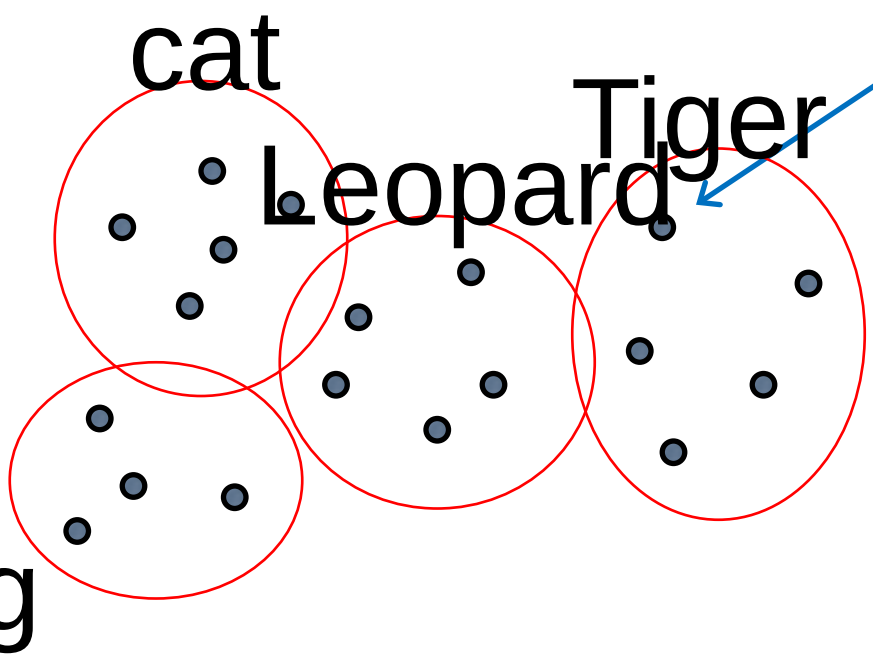
Slide by Fei Fei, Fergus & Torralba

Object categorization



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?

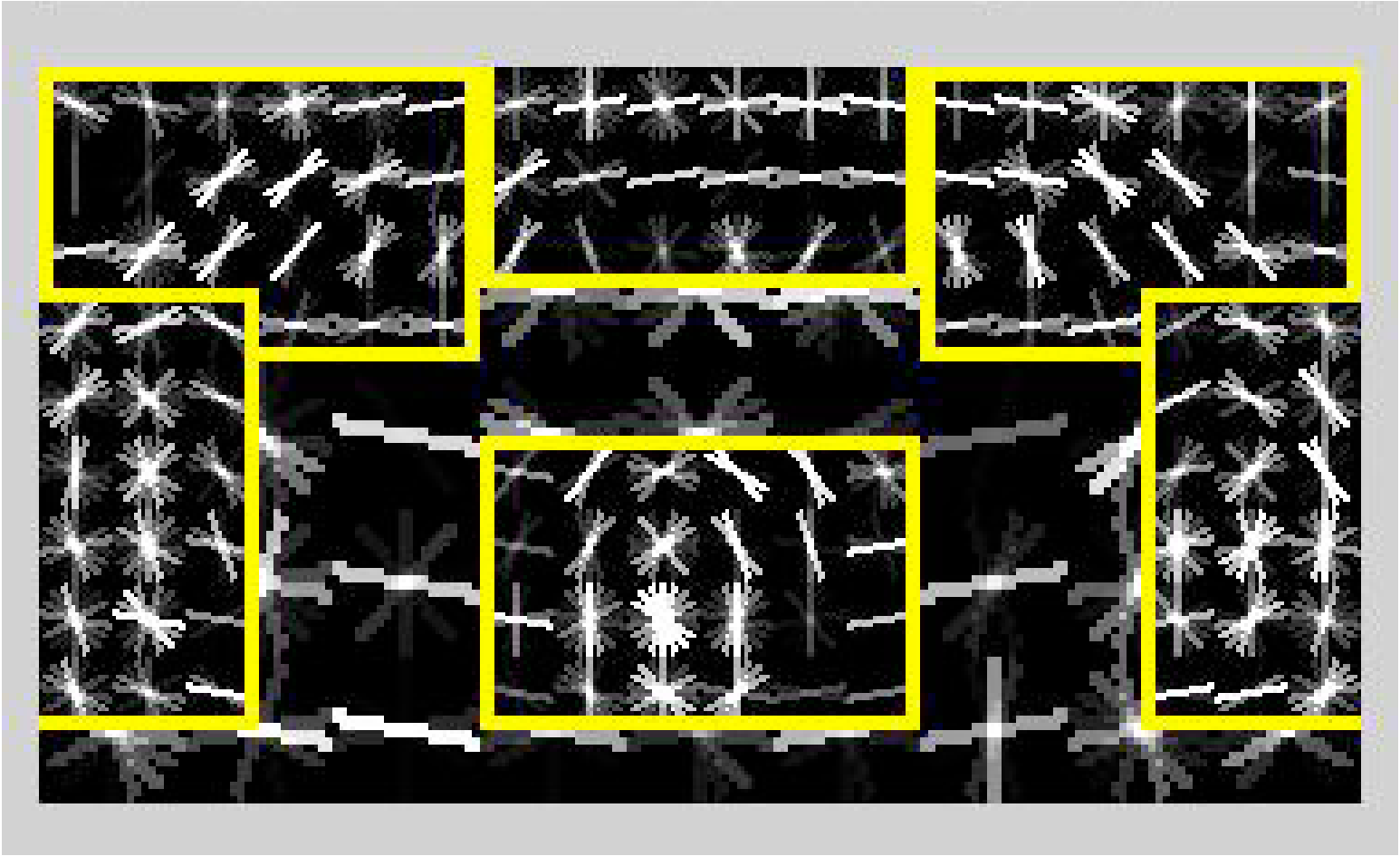
Problems with Visual Categories

- A lot of categories are functional

Chair

- World is too varied

Typical HOG car detector



Felzenszwalb et al, PASCAL 2007

Why not?

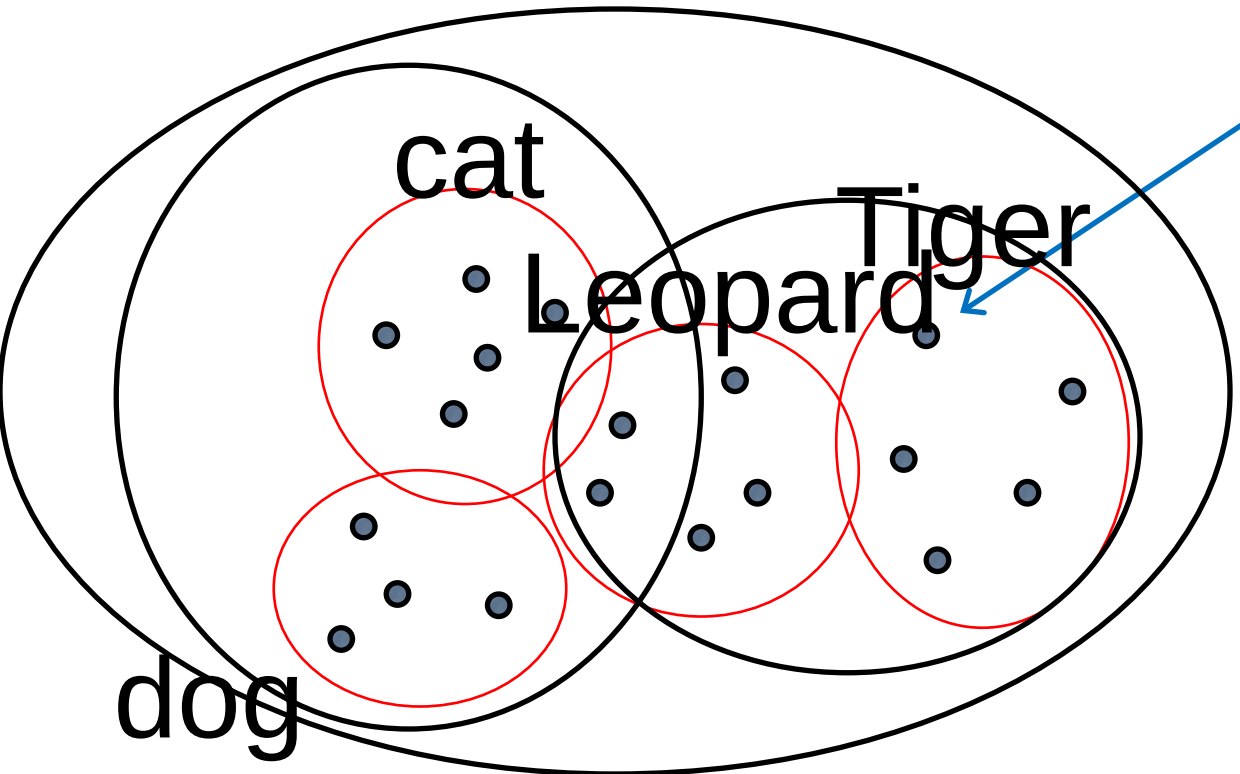

ABSOLUT
Country of Sweden
VODKA
This superb vodka
was distilled from grain grown
in the rich fields of southern Sweden.
It has been produced at the famous
old distilleries near Åhus
in accordance with more than
400 years of Swedish tradition.
Vodka has been sold under the name
Absolut since 1879.
40% ALC./VOL. (80 PROOF) 1 LITER
IMPORTED
PRODUCED AND BOTTLED IN ÅHUS, SWEDEN
BY THE ABSOLUT COMPANY
A DIVISION OF V&S VIN&SPRIT AB

+

Solution: hierarchy?

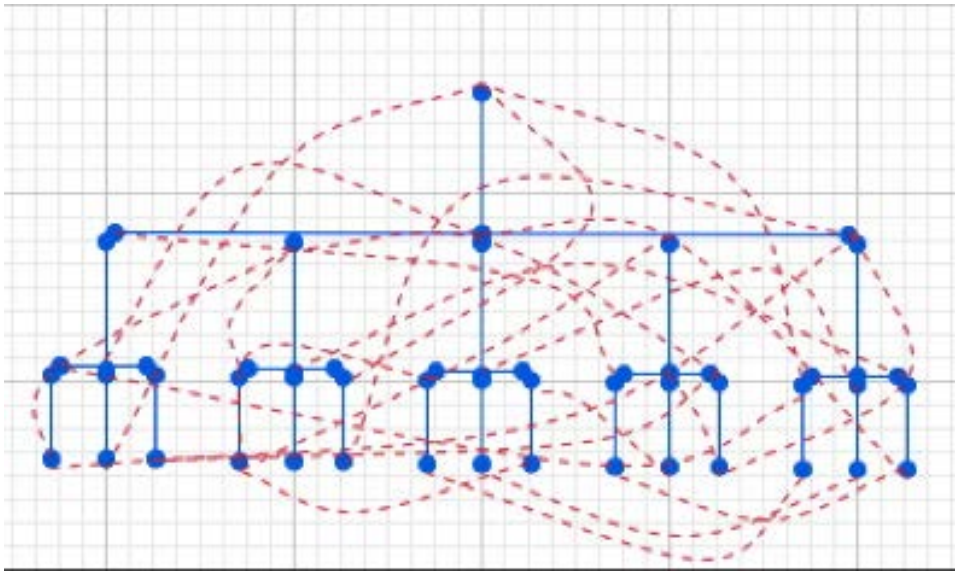
Ontologies, hierarchies, levels of categories (Rosch), etc.

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!

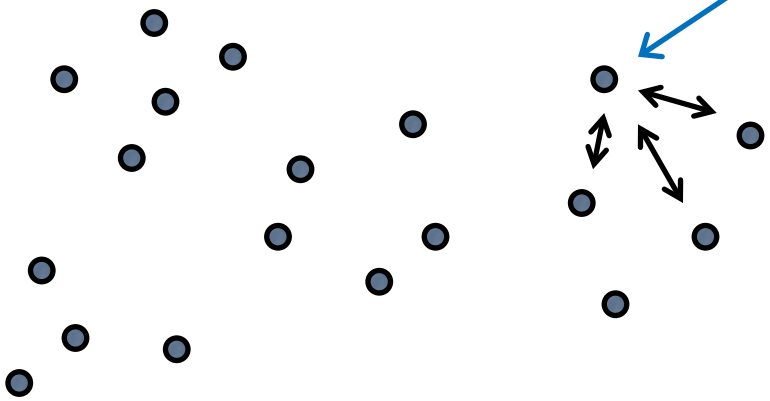
The Dictatorship of Librarians

categories are losing…

vs.

On-the-fly Categorization?

1. Knowledge Transfer
2. ~~Communication~~



Association instead of categorization

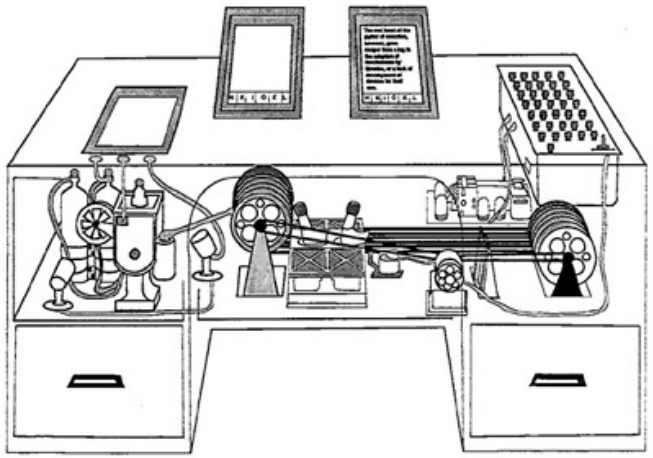
Ask not "what is this?", ask "what is this like"

– 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]

New object

Nodes = instances
Edges = associations

types of edges:
- visual similarity
- spatial, temporal co-occurrence
- geometric structure
- language
- geography
- ..

Image Understanding via Memex

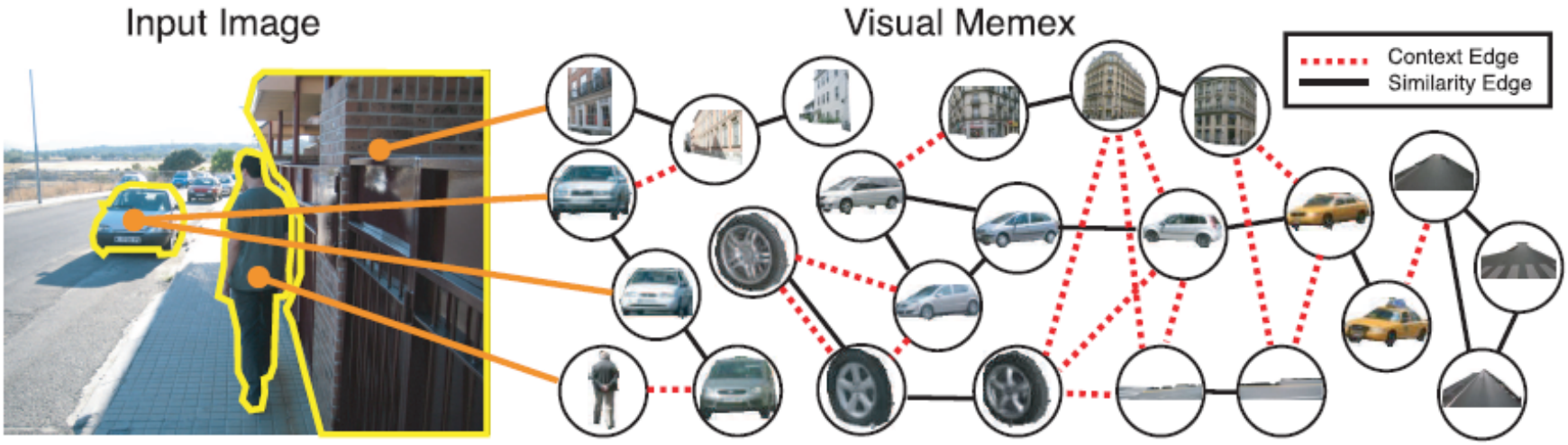


Figure 1: The **Visual Memex** graph encodes object similarity (solid black edge) and spatial context (dotted red edge) between pairs of object exemplars. A spatial context feature is stored for each context edge. The Memex graph can be used to interpret a new image (left) by associating image segments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

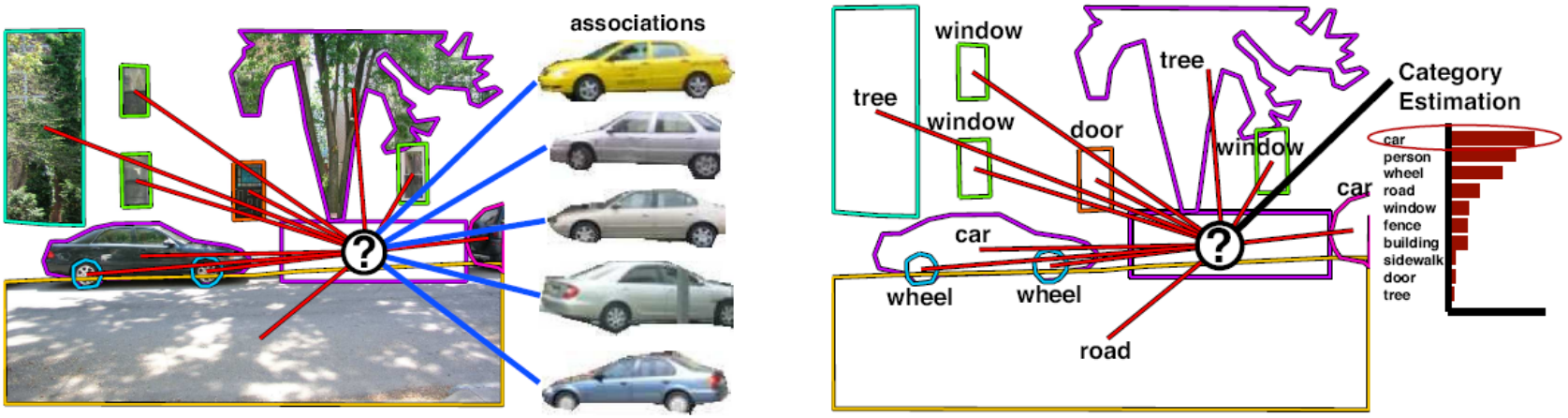


Figure 2: Torralba's Context Challenge: "How far can you go without running a local object detector?" The task is to reason about the identity of the hidden object (denoted by a "?") without local information. In our category-free Visual Memex model, object predictions are generated in the form of exemplar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

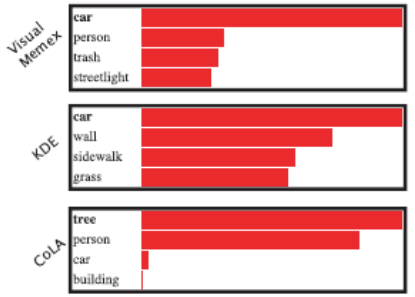
Reduced Memex: categories, non-parametric relationships

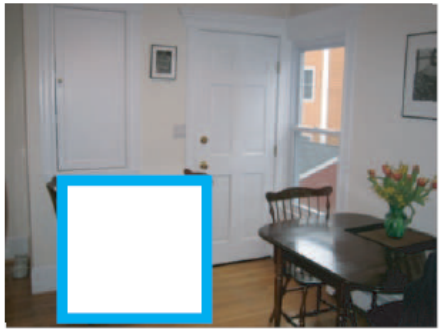
Qual. results

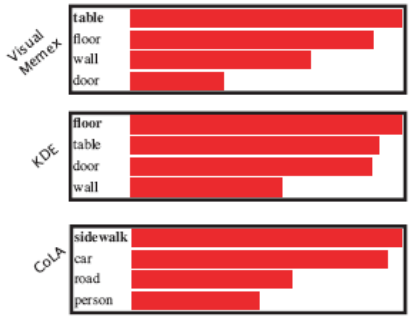
Input Image + Hidden Region

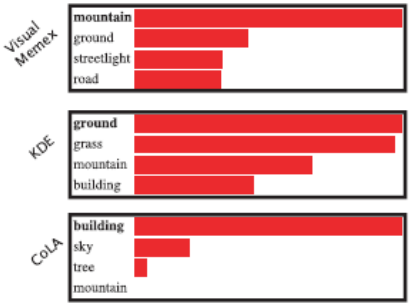
Visual Memex Exemplar Predictions

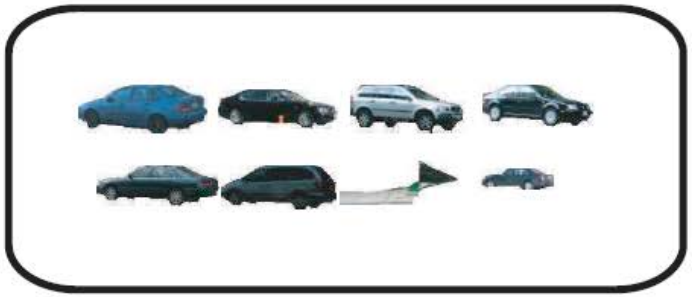
Categorization Results

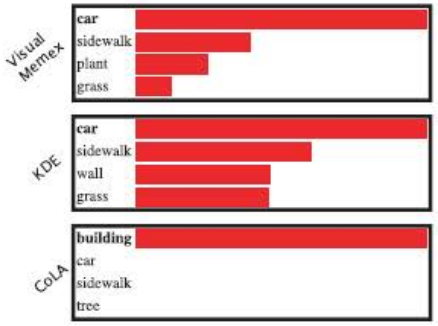


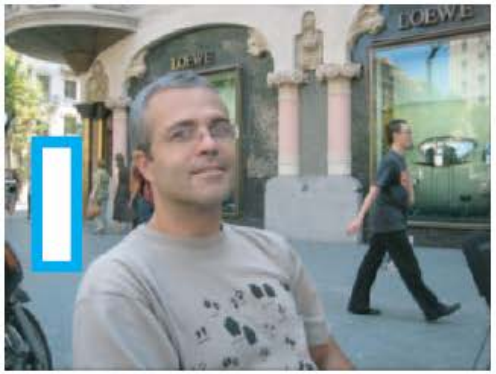


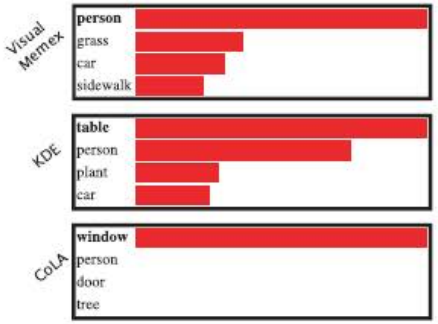







Visual Memex
car
sidewalk
plant
grass
KDE
car
sidewalk
wall
grass
CoLA
building
car
sidewalk
tree




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

9. Use other modalities as labels

9. Use other modalities as weak labels

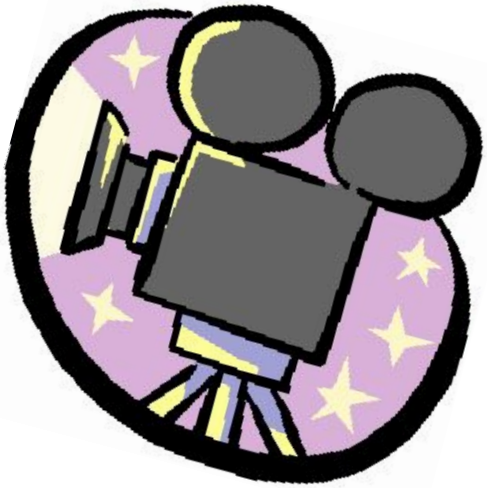
- GPS
- Date
- Artist
- …

Visual + haptic

Passive Data-gathering machine

Active Data-gathering machine

24/7 Poke-bot Farm



Conclusions

- *“If you torture data long enough, it might confess”* -- Ronald Coase

Thank You

© Quint Buchholz