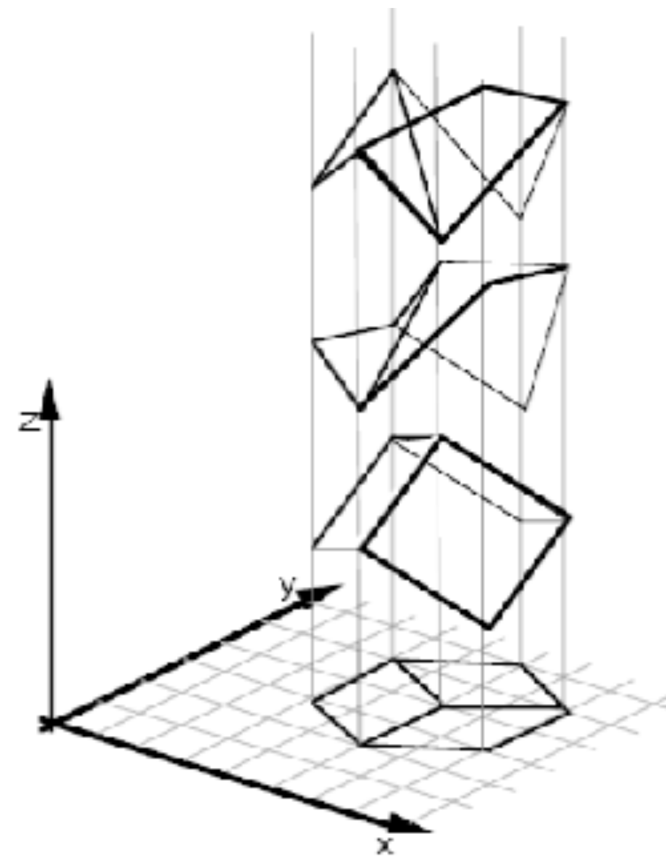
The Promise and Perils of Big Data

Some Slides from A. Efros and A. Torralba

Why do we need data?

Most problems in vision are ambiguous and hard.

- 2D -> 3D
- Segmentation/Edges



So, how do we solve these problems?

- Magic of data !
- Use data to learn better likelihoods: how things look like.
- Use data to learn priors of what is more likely than others.

But how much data do we need?

The extremes of learning

Extrapolation problem
Generalization
Transfer learning

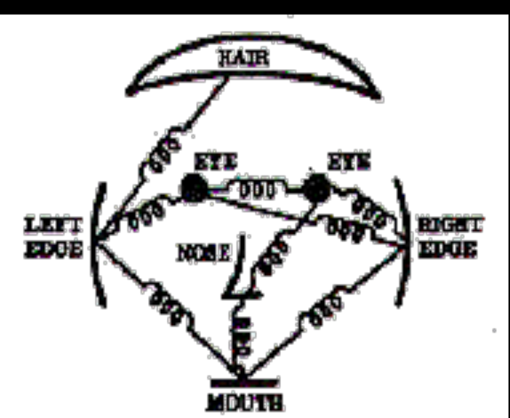
Interpolation problem
Correspondence
Finding the differences

Current Datasets

$1 \quad 10 \quad 10^2 \quad 10^3 \quad 10^4 \quad 10^{6\text{-}7} \quad 10^{10} \quad \infty$

Number of training samples

Datasets before 2012



So how much data does humans use?

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

High Fidelity Visual Memory is possible (Hollingworth 2004)

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


Slide by Aude Oliva

how far can we push the fidelity of visual LTM representation ?

Same object category, different instance

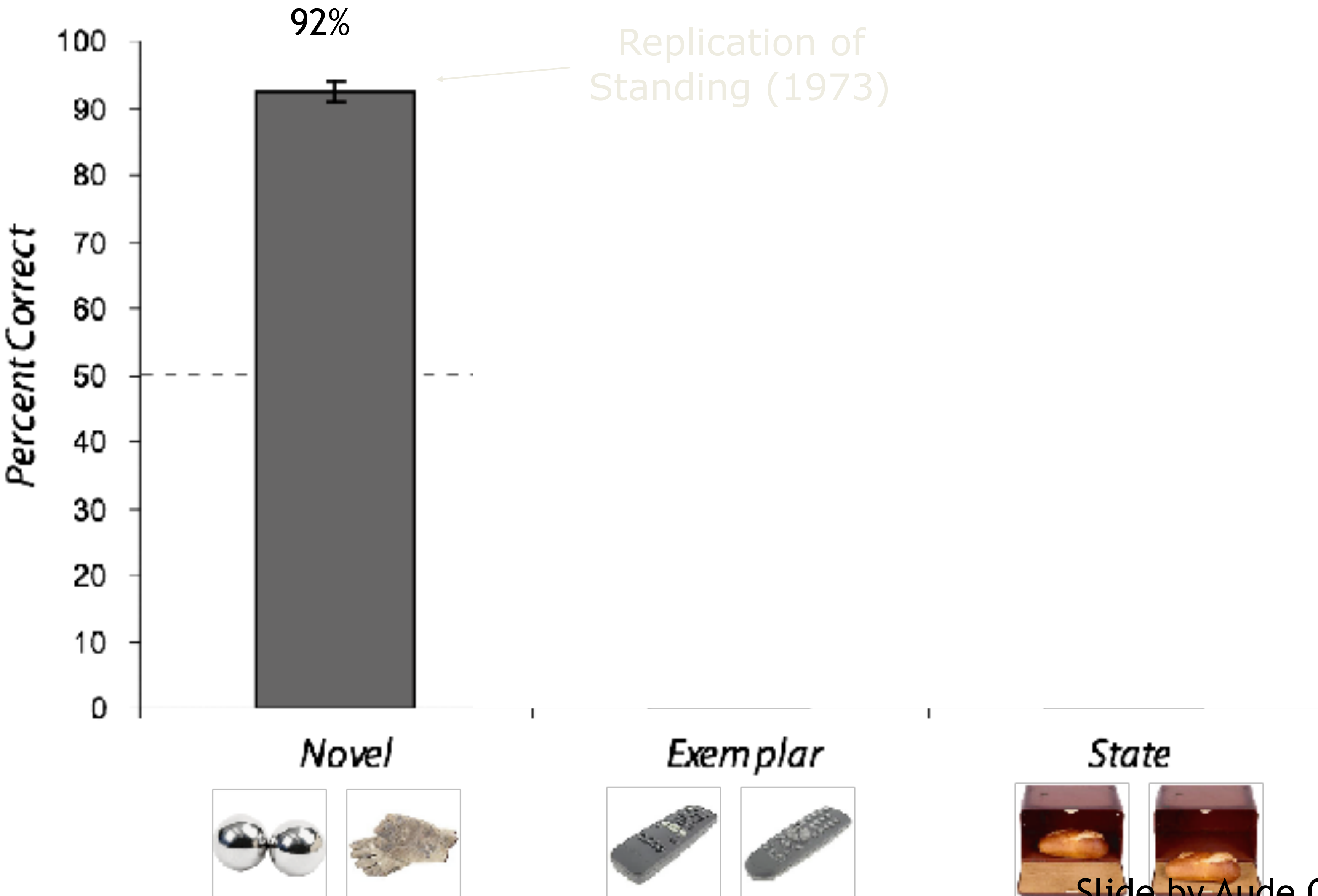
Slide by Aude Oliva

how far can we push the fidelity of visual LTM representation ?

Same object, different states

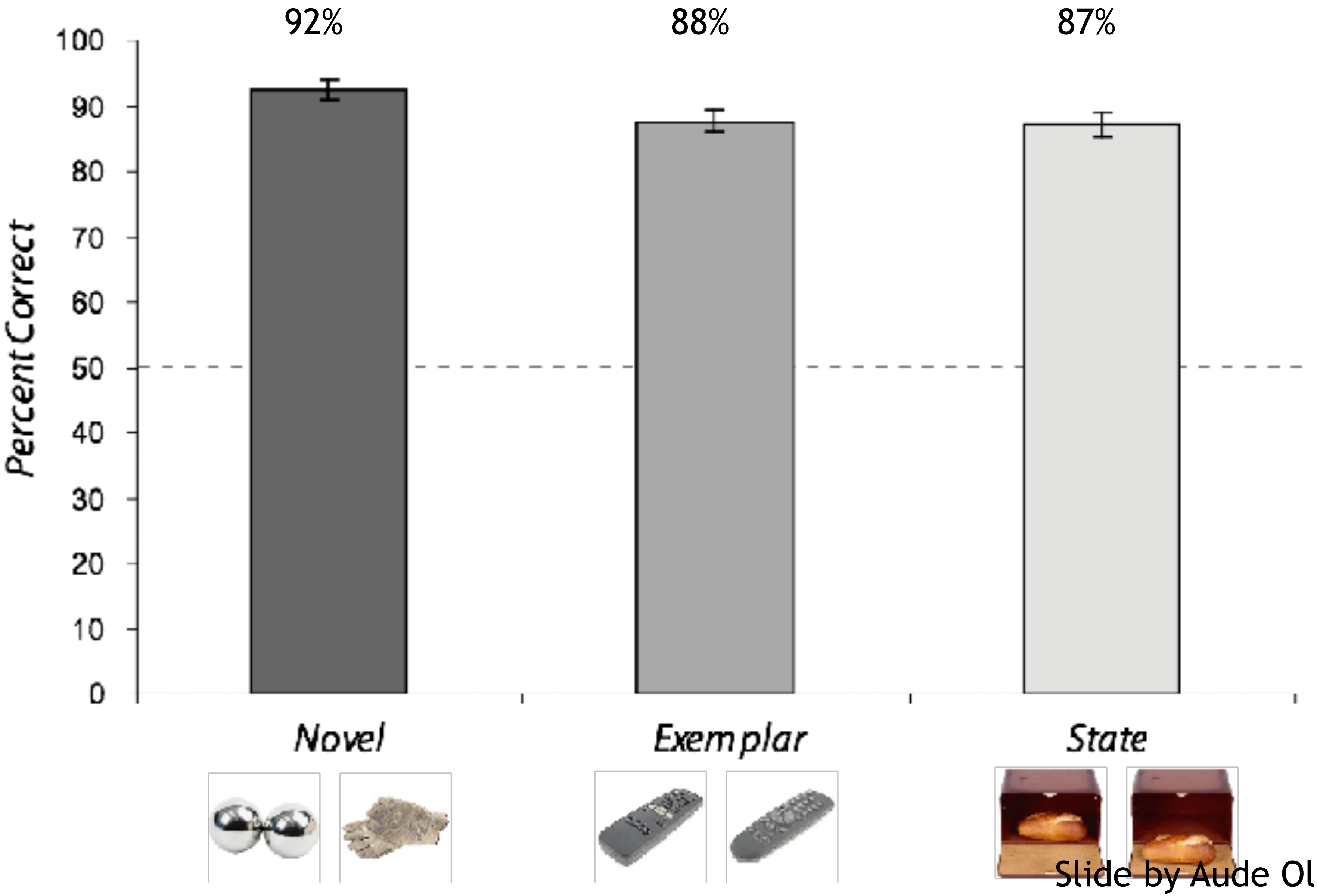
Slide by Aude Oliva

Massive Memory I: Recognition Memory Results



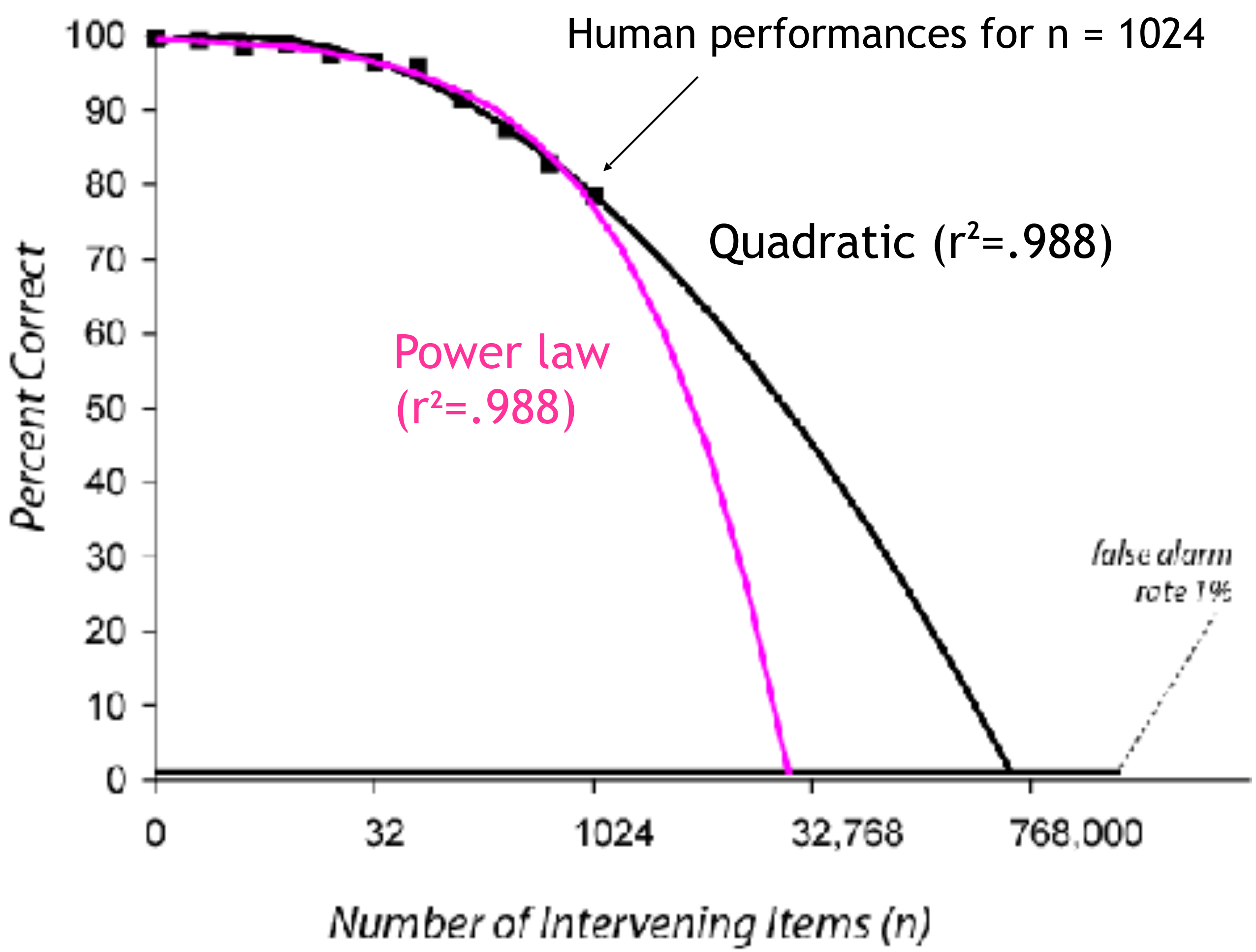
Slide by Aude Oliva

Massive Memory I: Recognition Memory Results



Slide by Aude Oliva

Extrapolation of Repeat Detection Data



Brady, Konkle, Alvarez, Oliva (submitted)

Slide by Aude Oliva

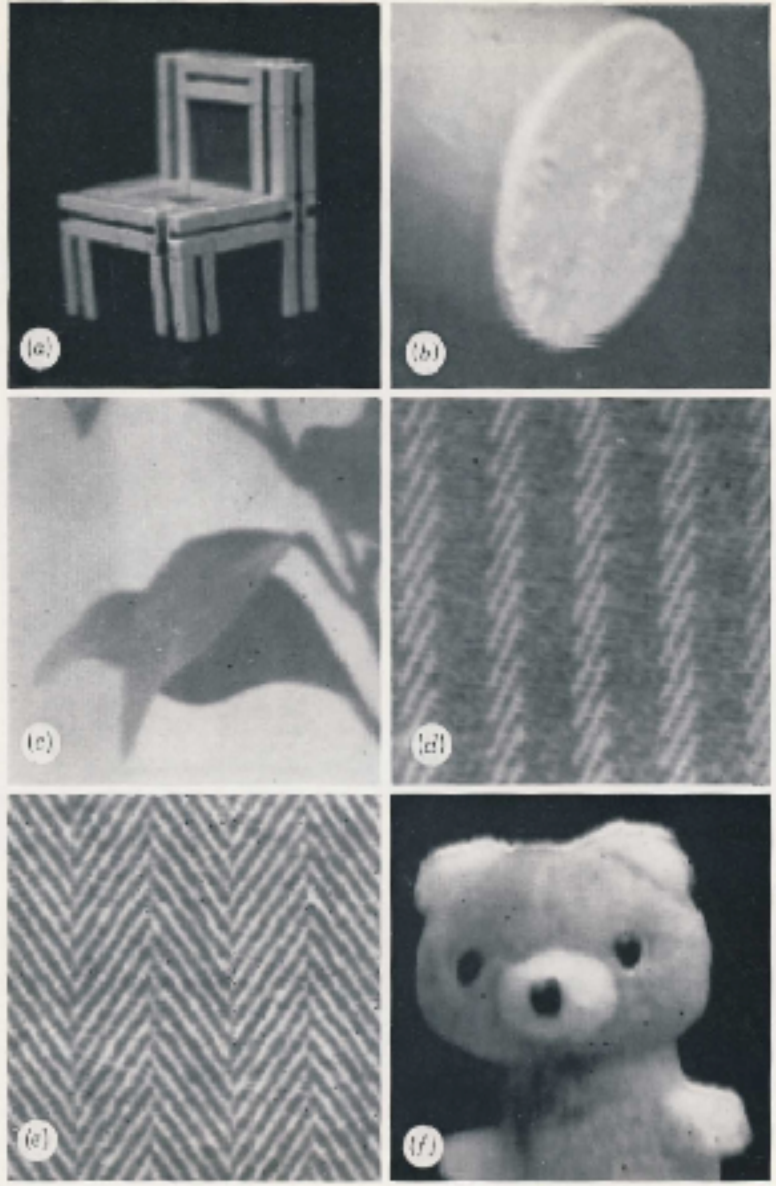
how much data does computer vision researchers use?

10^0
images

1972

10^1
images

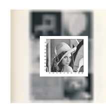
10^1 images



URE 4. This figure provides a high quality reproduction of the six images discussed in the text. (a) and (b) were taken with a considerably modified Information International Incorporated Vidissector, and the rest were taken with a Telemation TMC-2100 vidicon camera attached to a Spatial Data Systems digitizer (Camera Eye 108). The full dynamic range from black to white is represented by 256 grey-levels. The images reproduced here were created by an Optronics P1500hPhotowriter from intensity arrays that measured 128 elements square. This size of intensity array corresponds to viewing a 1 in square at 5 ft with the human retina. The image of the period at the end of this sentence probably covers more than 40 retinal receptors. The reader should view the images from a distance of about 5 ft when assessing the performance of the programs.

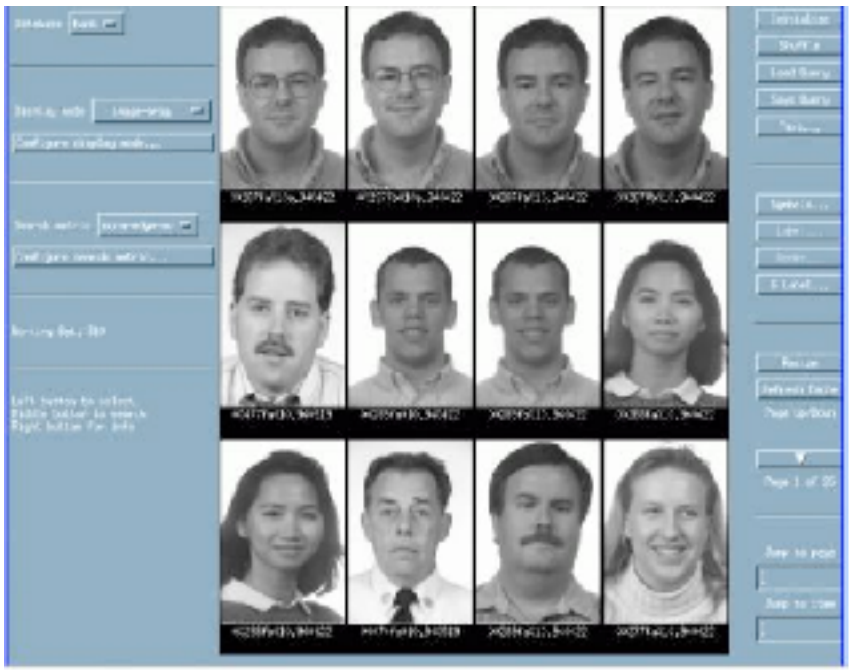
Marr, 1976

$10^{2\text{-}4}$
images

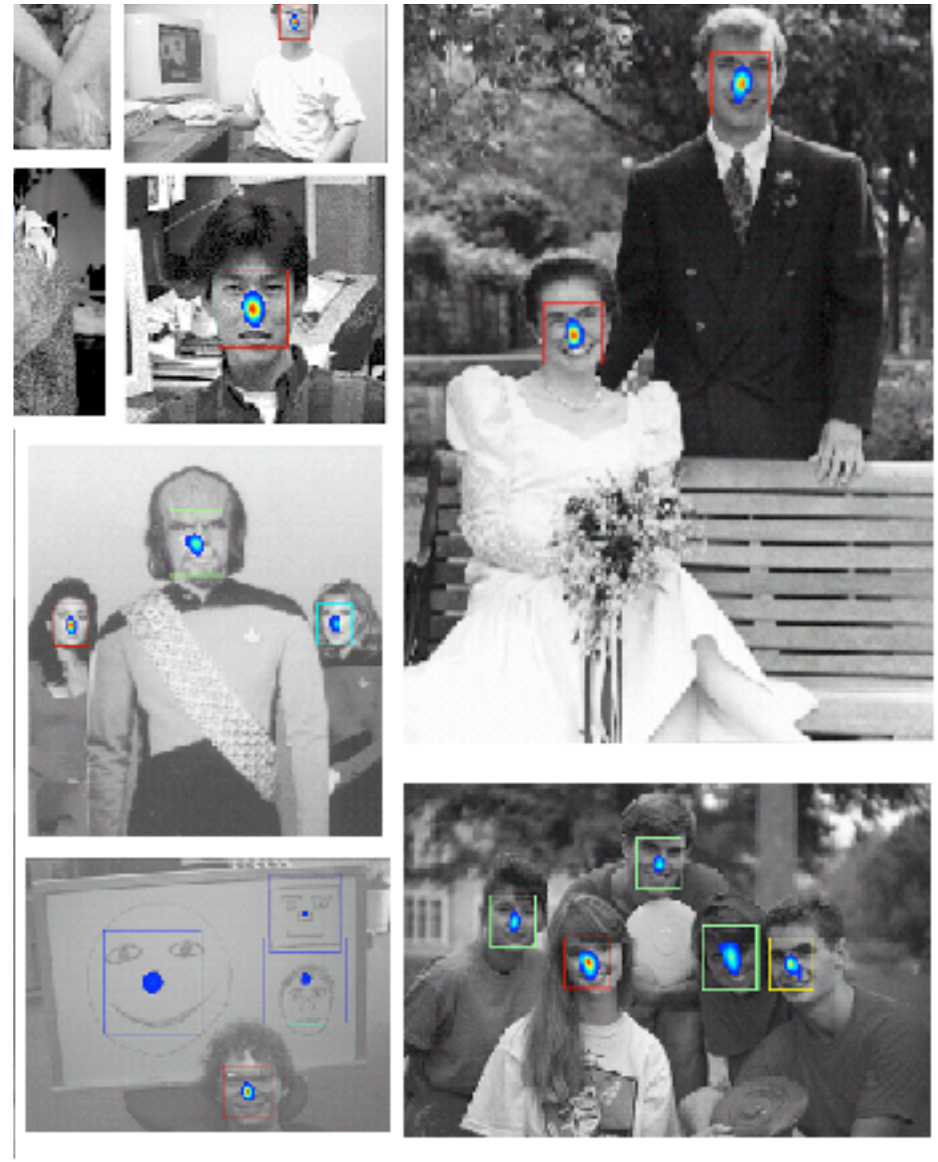


The faces and cars scale

10^{2-4} images



In 1996 DARPA released 14000 images, from over 1000 individuals.



The PASCAL Visual Object Classes

In 2007, the twenty object classes that have been selected are:

Person: person
Animal: bird, cat, cow, dog, horse, sheep
Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

M. Everingham, Luc van Gool , C. Williams, J. Winn, A. Zisserman 2007

$10^{2\text{-}4}$ images

10^5 images

Caltech 101 and 256

10^5 images

Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007

Lotus Hill Research Institute image corpus

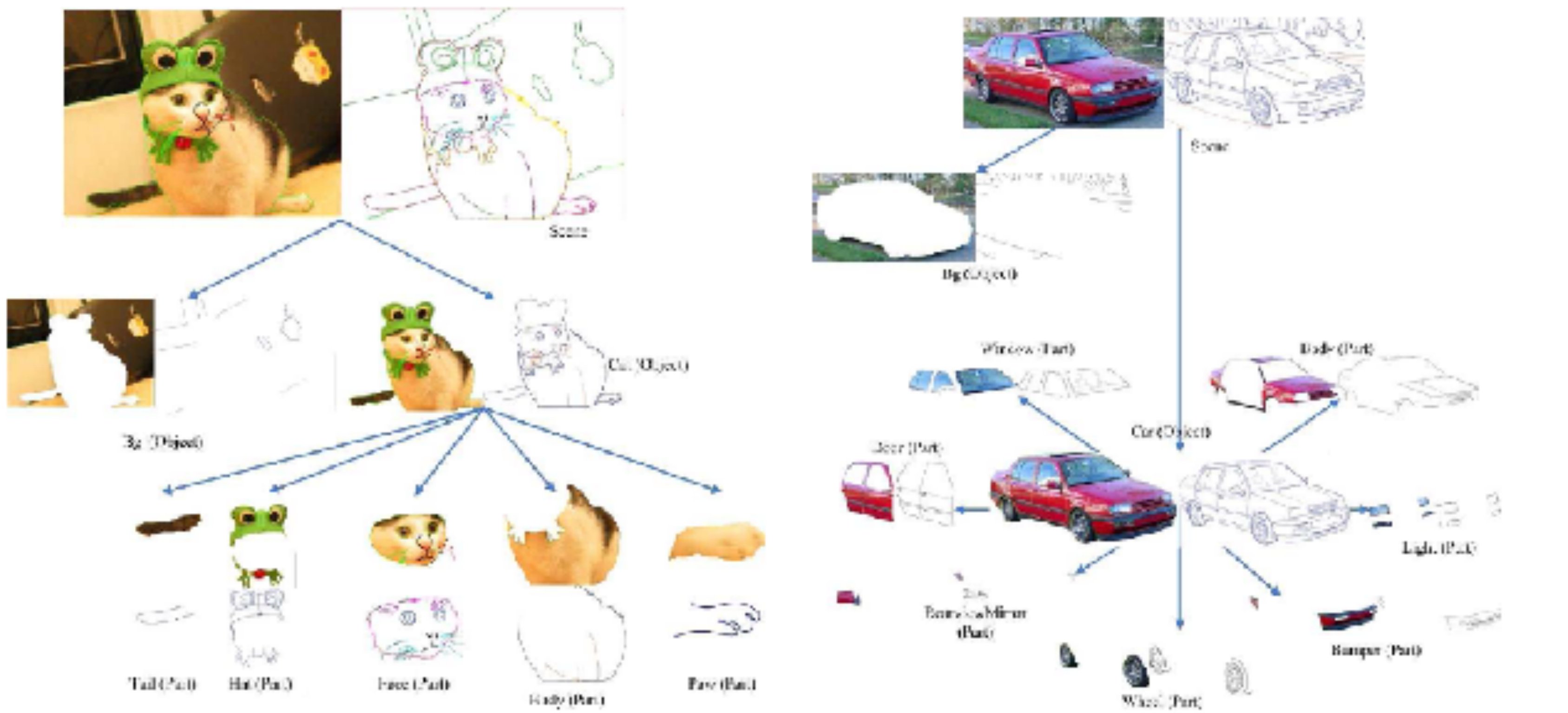
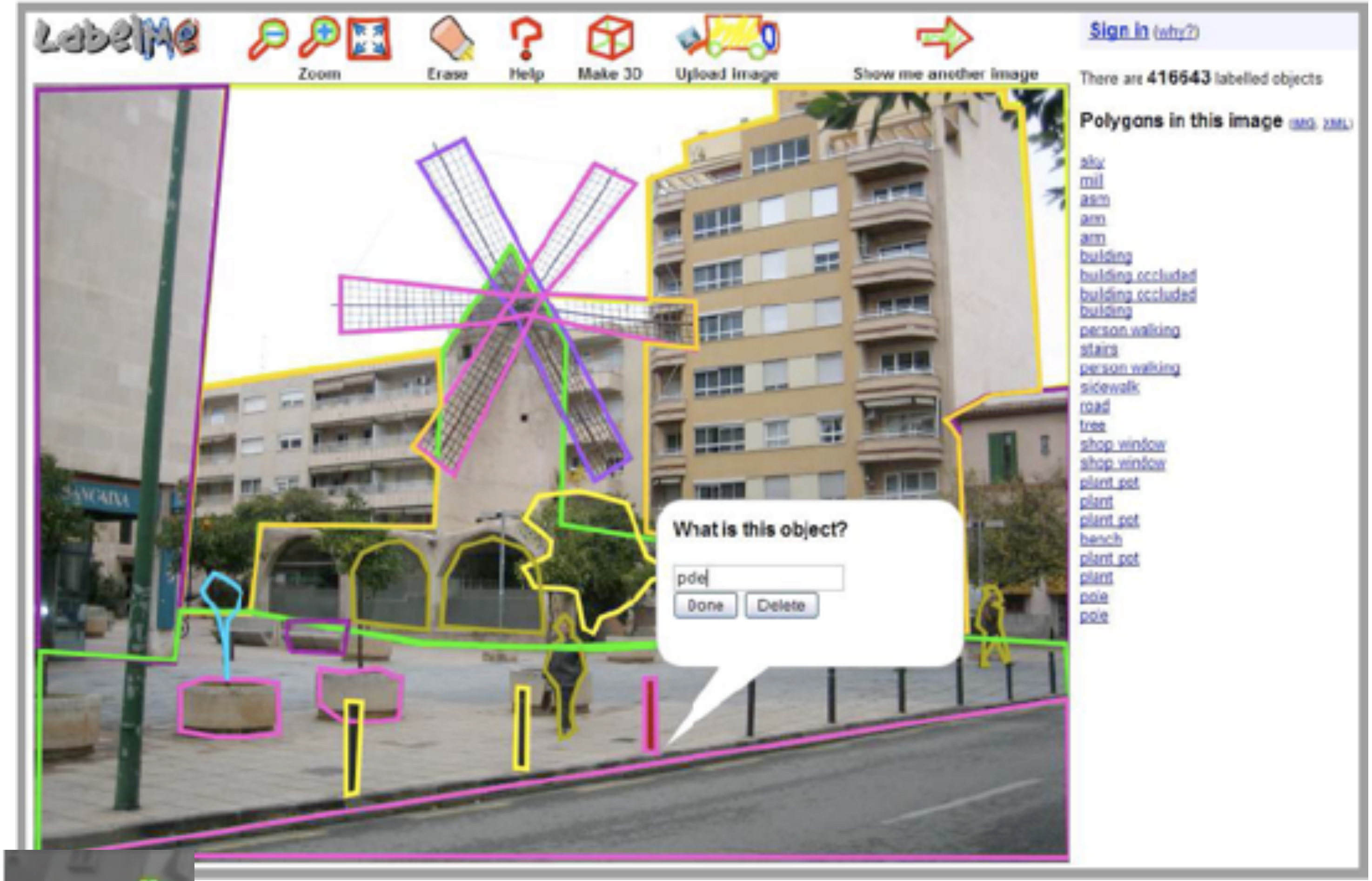


Figure 5: Two examples of the parse trees (cat and car) in the Lotus Hill Research Institute image corpus. From [87].

Z.Y. Yao, X. Yang, and S.C. Zhu, 2007

LabelMe

10^5
images

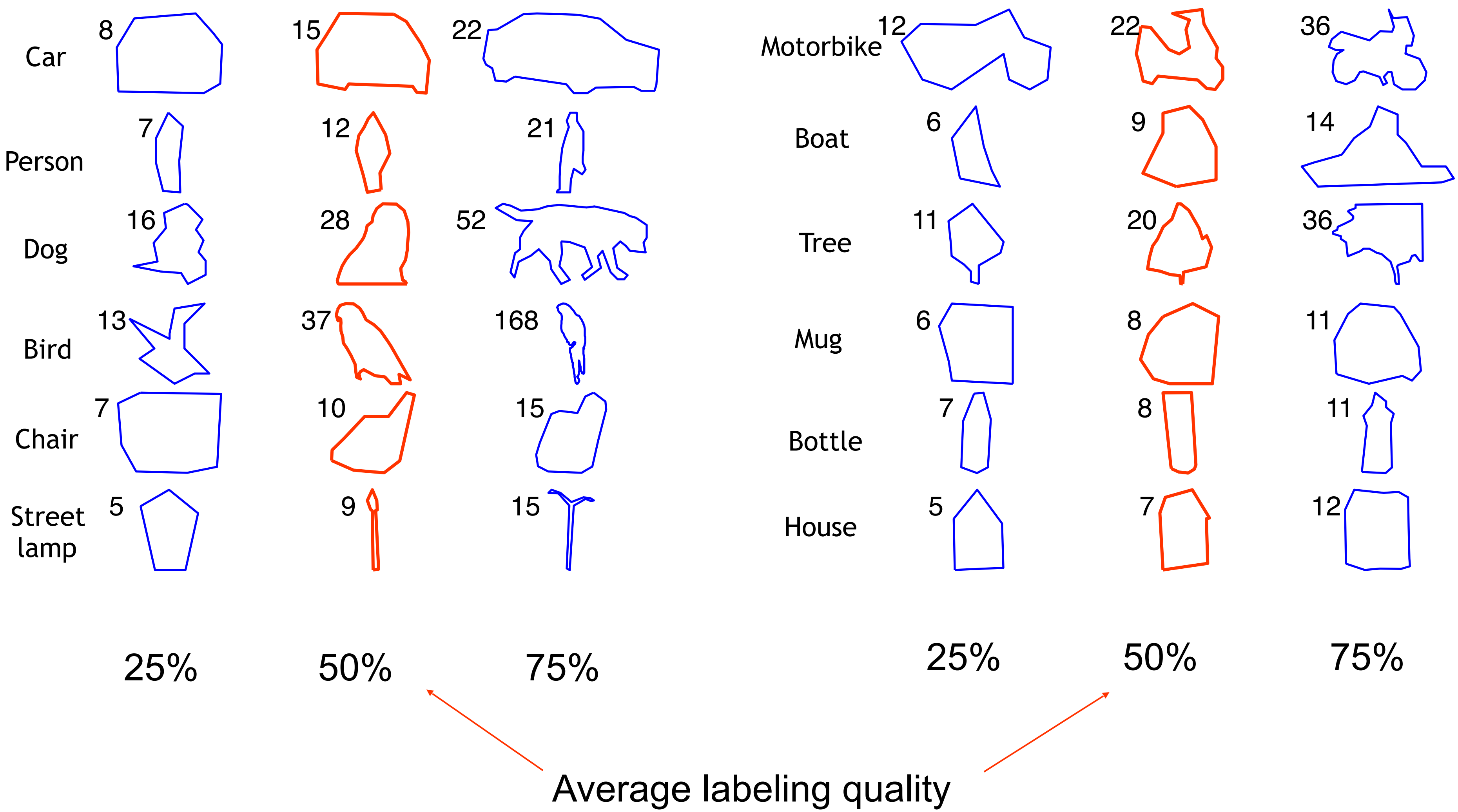


Tool went online July 1st, 2005
530,000 object annotations collected

Labelme.csail.mit.edu

B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, IJCV 2008

Quality of labeling



Extreme labeling



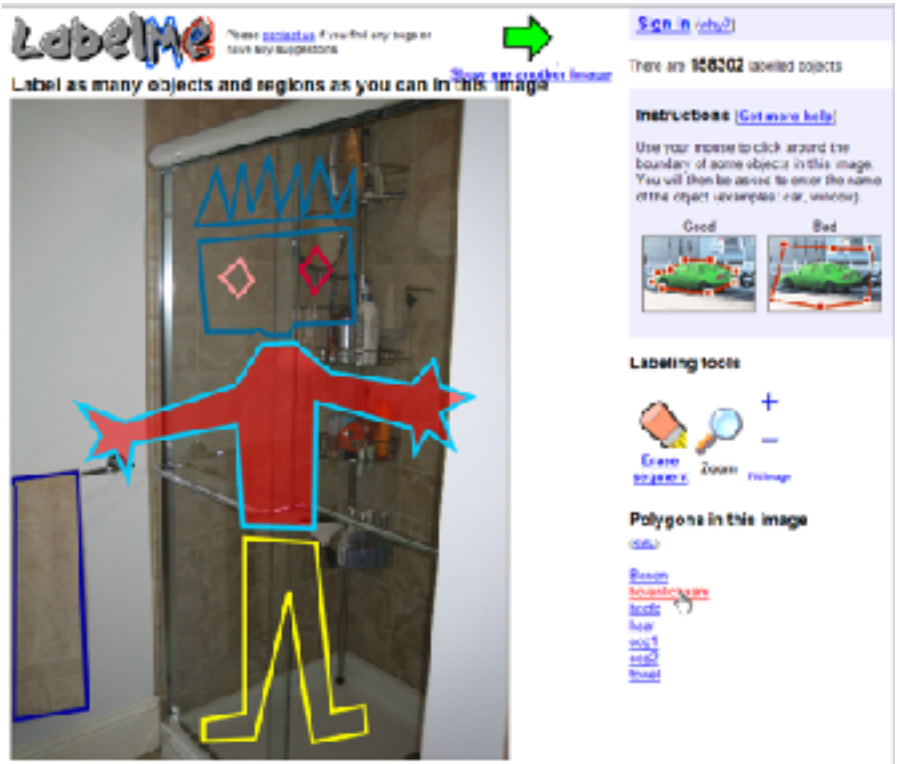




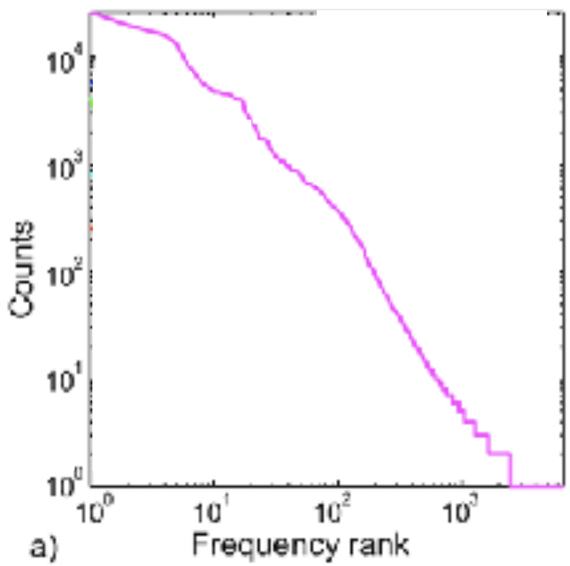

The other extreme of extreme labeling

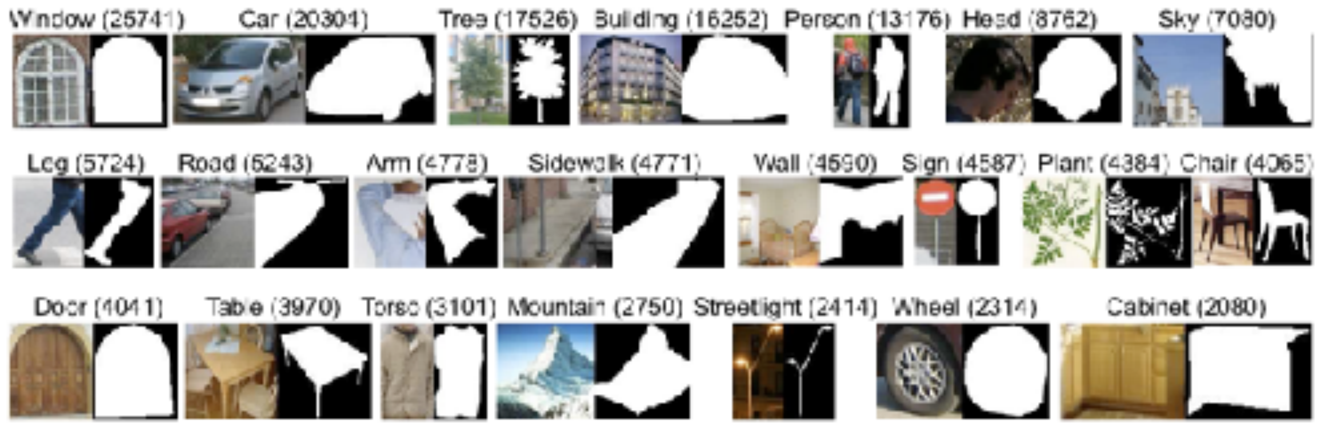
… things do not always look good…

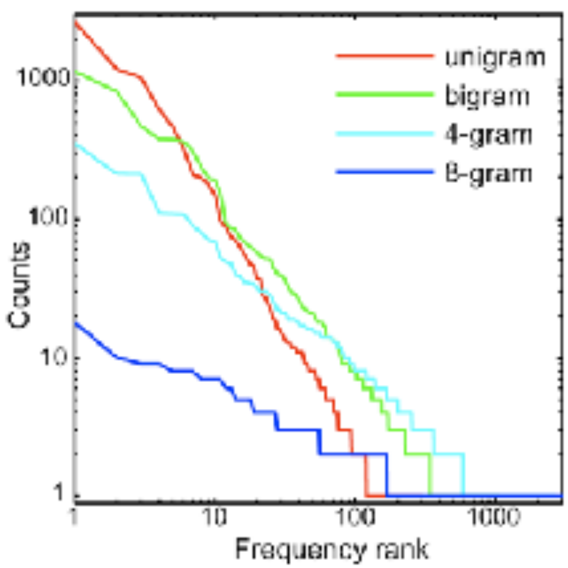
Creative testing

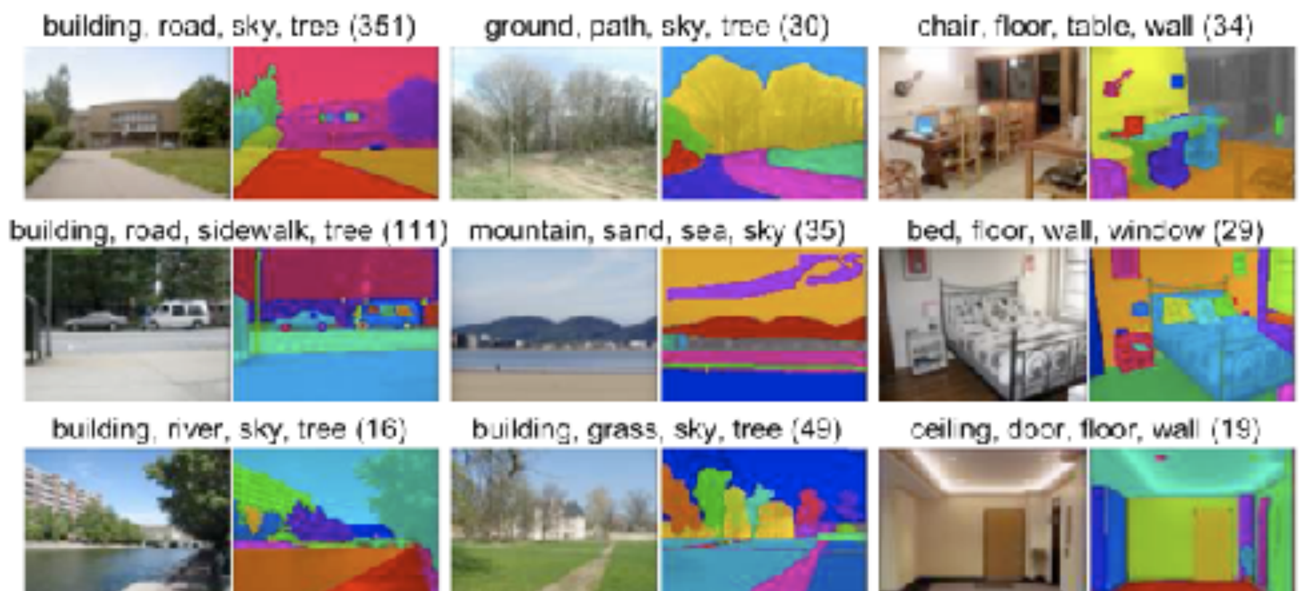


Scene and object biases



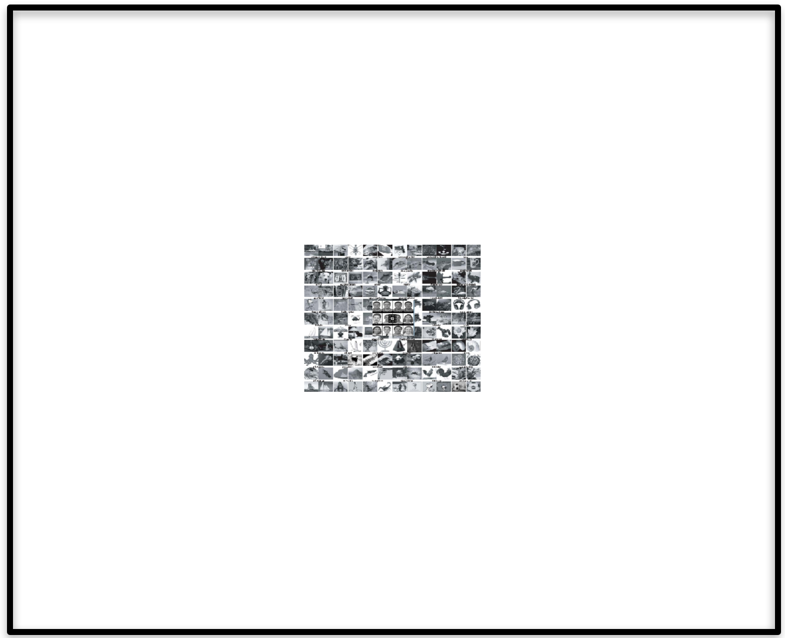






10^5 images

$10^{6\text{-}7}$
images



Things start getting out of hand

Collecting big datasets

$10^{6\text{-}7}$ images

- ESP game (CMU)
  Luis Von Ahn and Laura Dabbish 2004

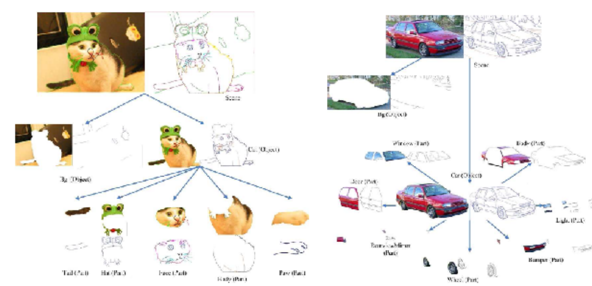
- LabelMe (MIT)
  Russell, Torralba, Freeman, 2005

- StreetScenes (CBCL-MIT)
  Bileschi, Poggio, 2006

- WhatWhere (Caltech)
  Perona et al, 2007

- PASCAL challenge
  2006, 2007

- Lotus Hill Institute
  Song-Chun Zhu et al, 2007



- 80 million images
  Torralba, Fergus, Freeman, 2007

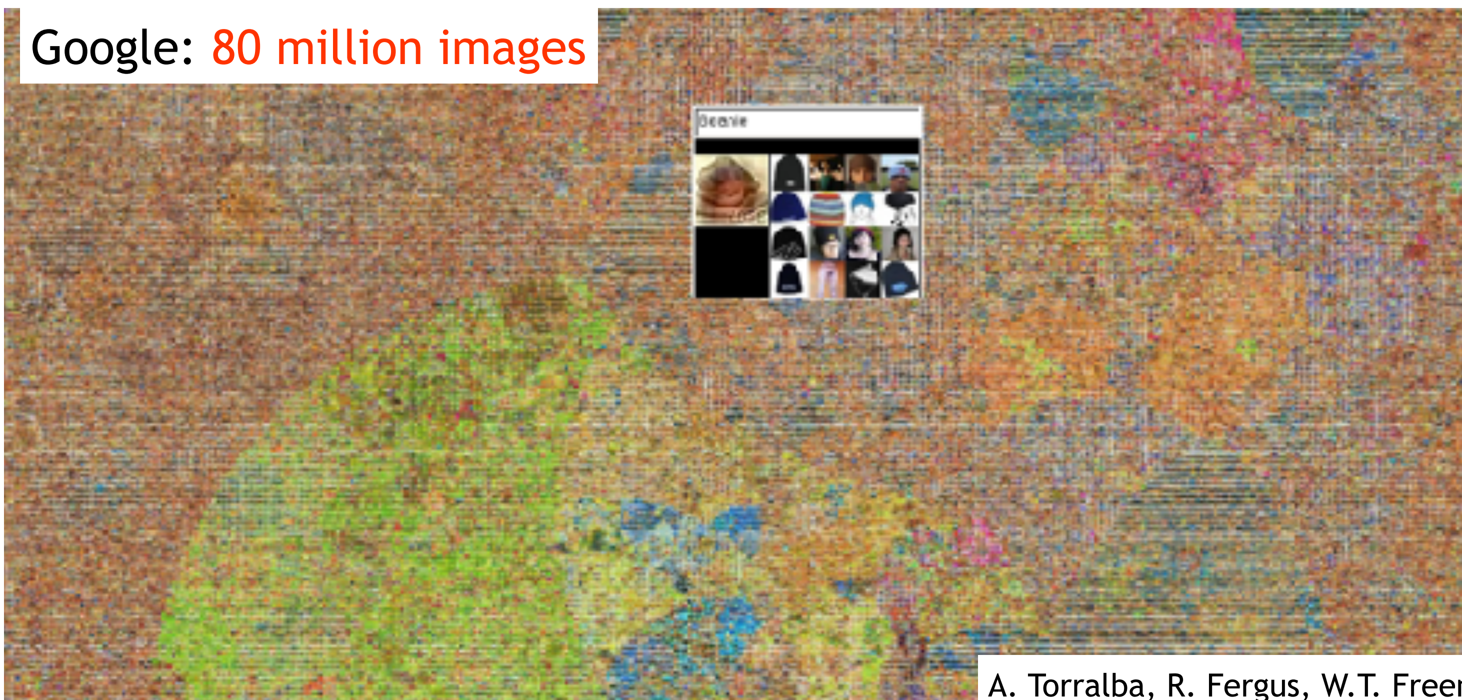
80.000.000 images

$10^{6\text{-}7}$ images

75.000 non-abstract nouns from WordNet

7 Online image search engines

And after 1 year downloading images

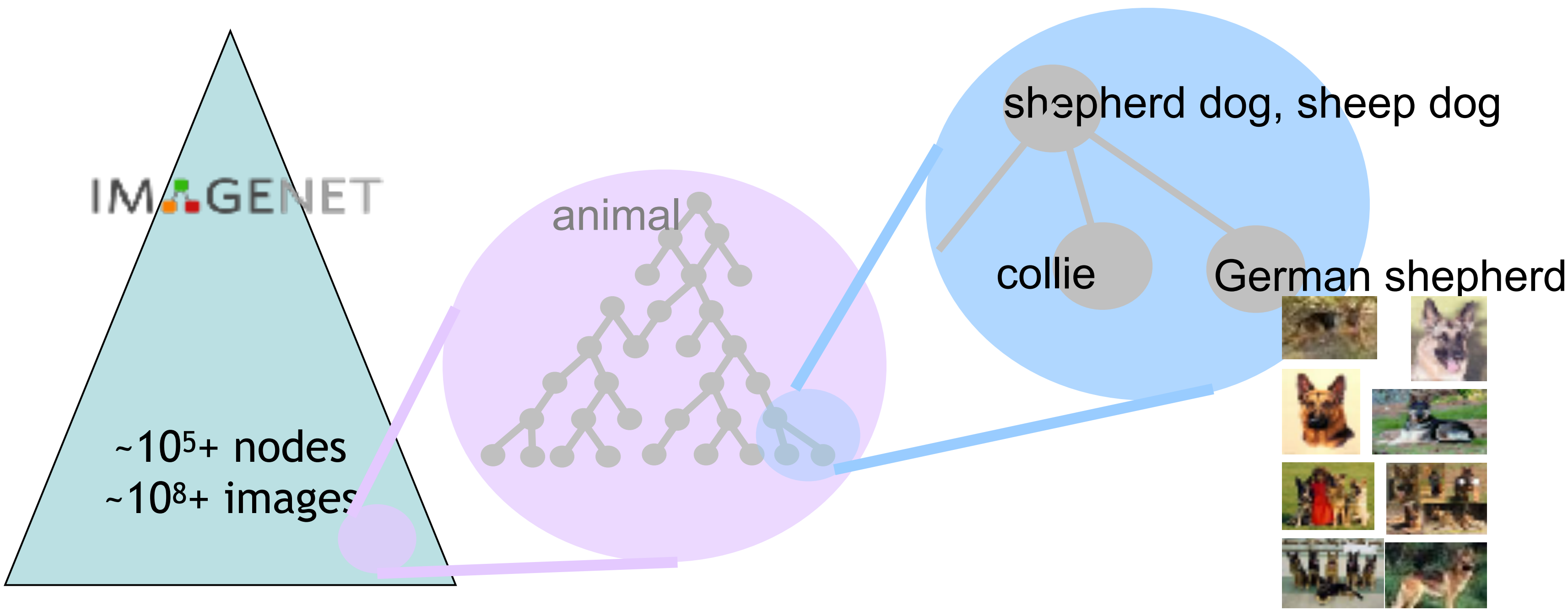


A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008

IMAGENET

$10^{6\text{-}7}$ images

- An ontology of images based on WordNet
- ImageNet currently has
  - 22,000+ categories of visual concepts
  - 15 million human-cleaned images (~700im/categ)
  - 1/3+ is released online @ www.image-net.org



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

amazon mechanical turk
Artificial Artificial Intelligence

Your Account | HITs | Qualifications

Already have an account?
Sign in as a Worker | Requester

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.

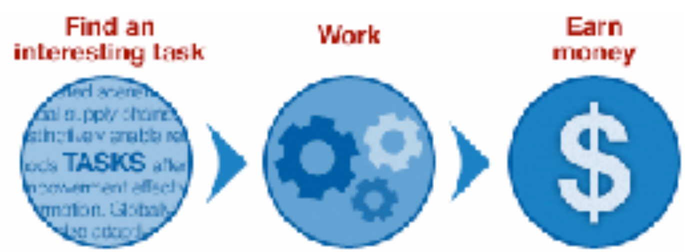
216,070 HITs available. View them now.

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 HITs Now

or learn more about being a **Worker**

Get Results
from Mechanical Turk Workers

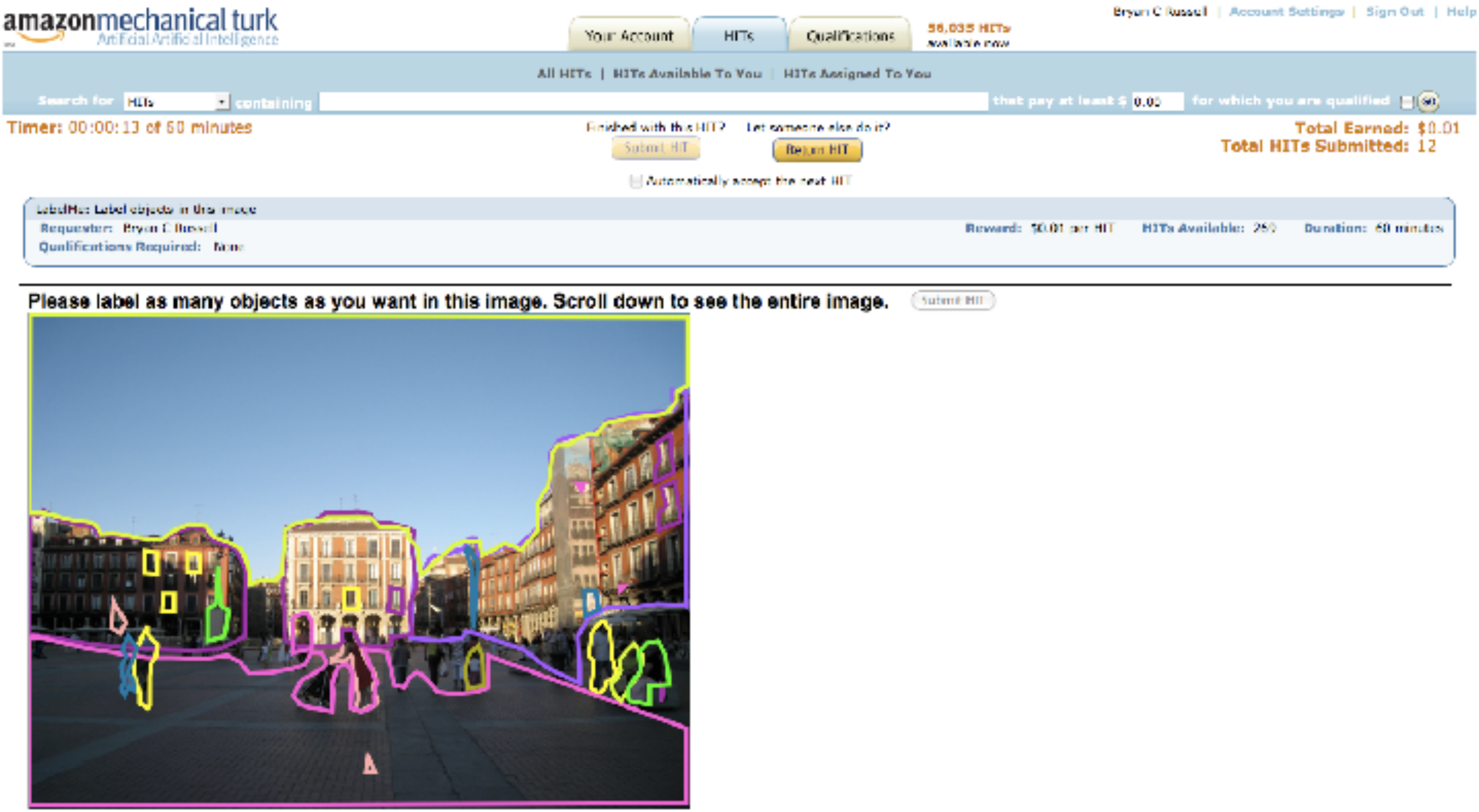
Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Register Now

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

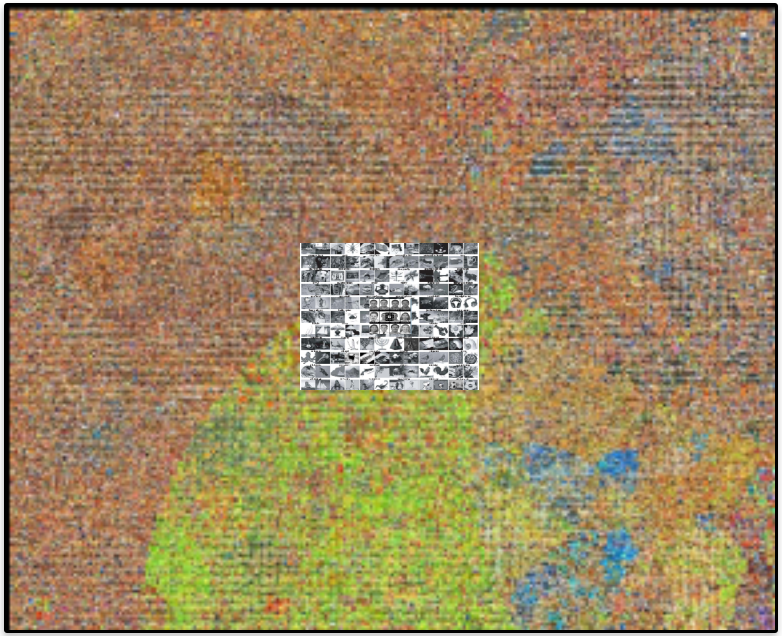
Get Started

Labeling for money



Alexander Sorokin, David Forsyth, "Utility data annotation with Amazon Mechanical Turk", First IEEE Workshop on Internet Vision at CVPR 08.

$10^{6\text{-}7}$
images



10^{8-11} images

Ask Images

Cydral

Google Image Search

webshots

picsearch

altavista

flickr

facebook

Datasets in perspective

Number of images on my hard drive: 10^5

Number of images seen during my first 10 years: 10^8
(3 images/second * 60 * 60 * 16 * 365 * 10 = 630720000)

Number of images seen by all humanity: 10^{20}
106,456,367,669 humans[1] * 100 years * 3 images/second * 60 * 60 * 16 * 365 =
1 from http://www.prb.org/Articles/2002/HowManyPeopleHaveEverLivedonEarth.aspx

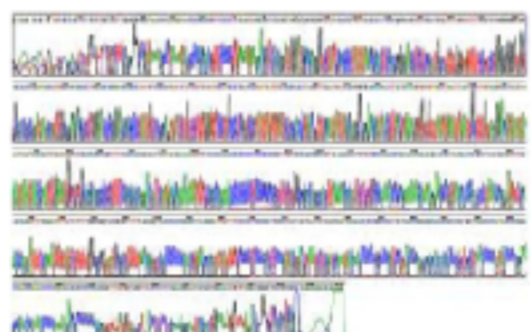
Number of all 32x32 images: 10^{7373}
$256^{32*32*3} \sim 10^{7373}$

PASCAL

Number of samples

When do we need big data?

Two Kinds of Things in the World

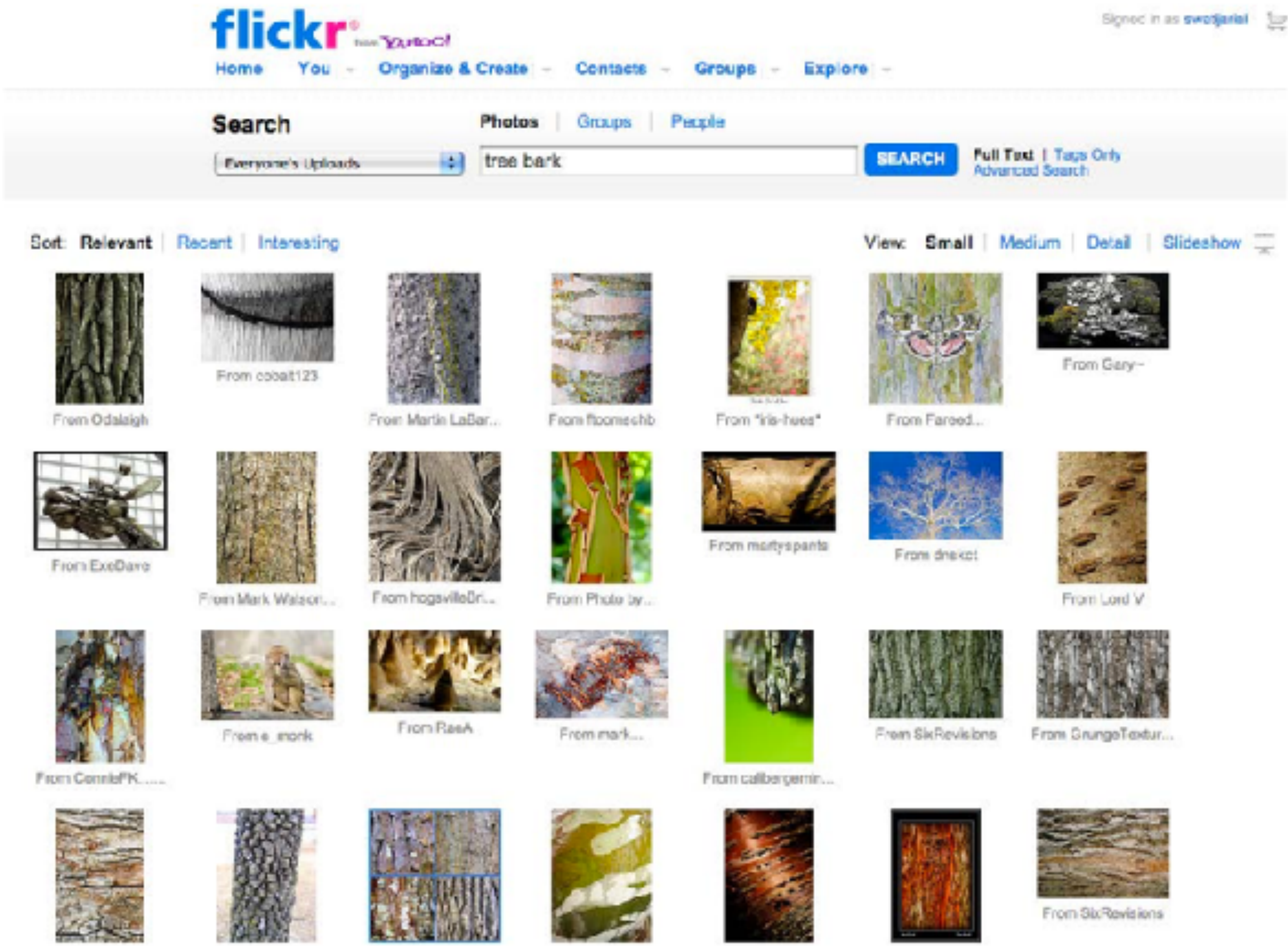


Navier-Stokes Equation

$$\frac{\partial \mathbf{u}}{\partial t} = -\left(\mathbf{u} \cdot \nabla\right)\mathbf{u} + v\nabla^2\mathbf{u} - \frac{1}{d}\nabla p + \mathbf{f}$$

+ weather
+ location
+ …

Lots of data available



“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,… visual understanding?
- Enter: The **<u>Magic of Data</u>**
  - Great advances in several fields:
    - e.g. speech recognition, machine translation, Google

Unreasonable Effectiveness of Data

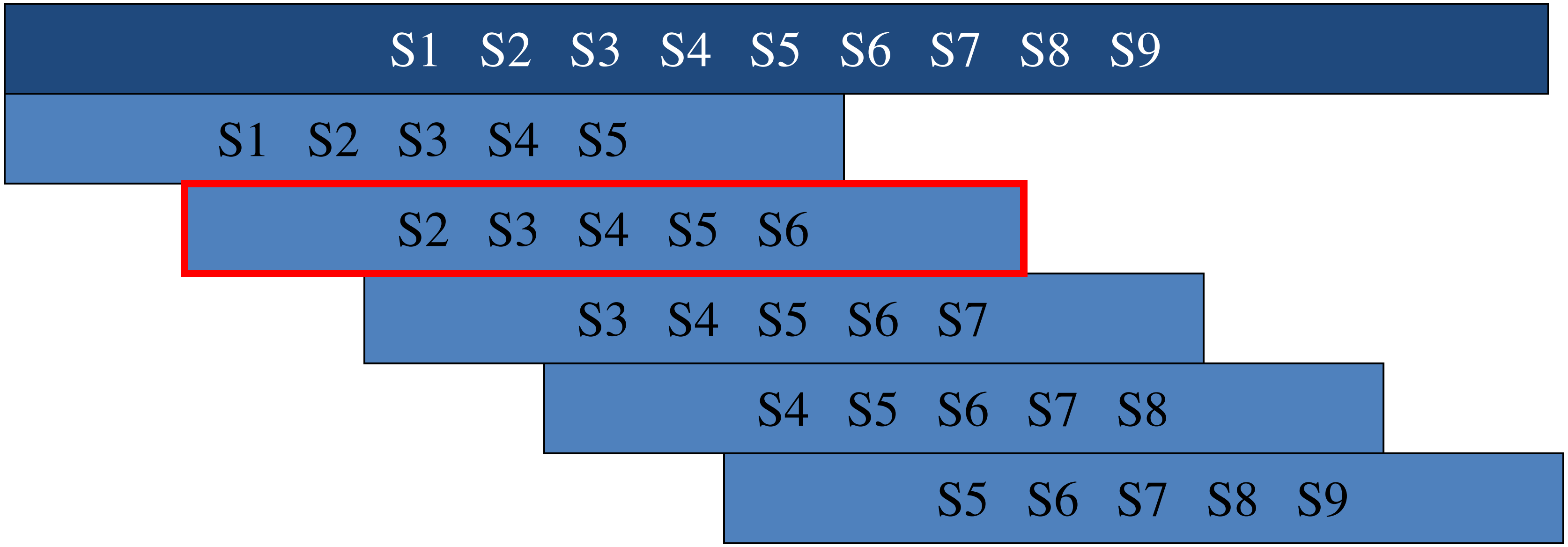
Simple Algorithms (Dumb) + Lot of Data are better than Complicated algorithms

Example: Machine Translation
Example: Texture Generation

Machine Translation

Step 1: Source Sentence Chunking

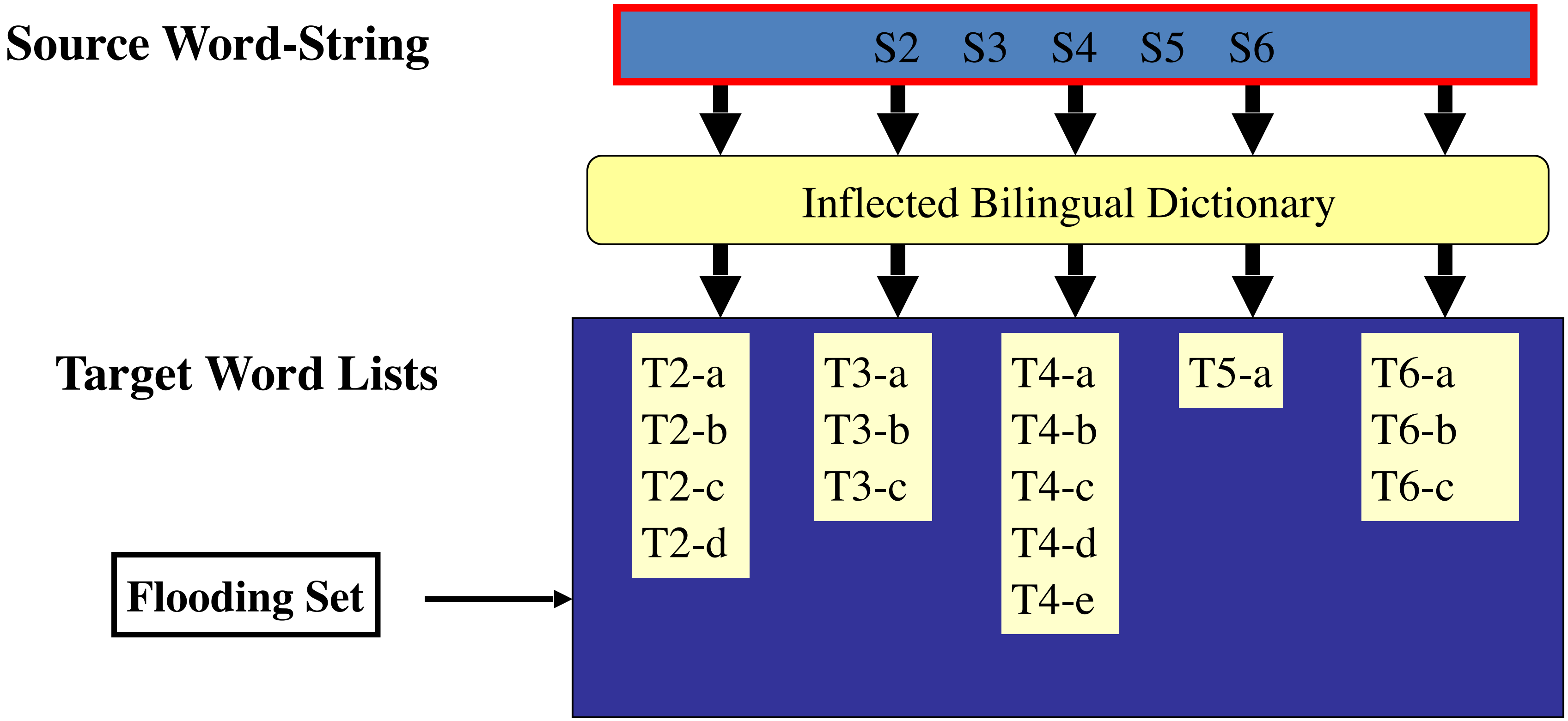
- Segment source sentence into overlapping n-grams via sliding window
- Typical n-gram length 4 to 9 terms
- Each term is a word or a known phrase
- Any sentence length



Slide by Jaime Carbonell

Step 2: Dictionary Lookup

- Using bilingual dictionary, list all possible target translations for each source word or phrase



Slide by Jaime Carbonell

Step 3: Search Target Text

- Using the Flooding Set, search target text for word-strings containing one word from each group

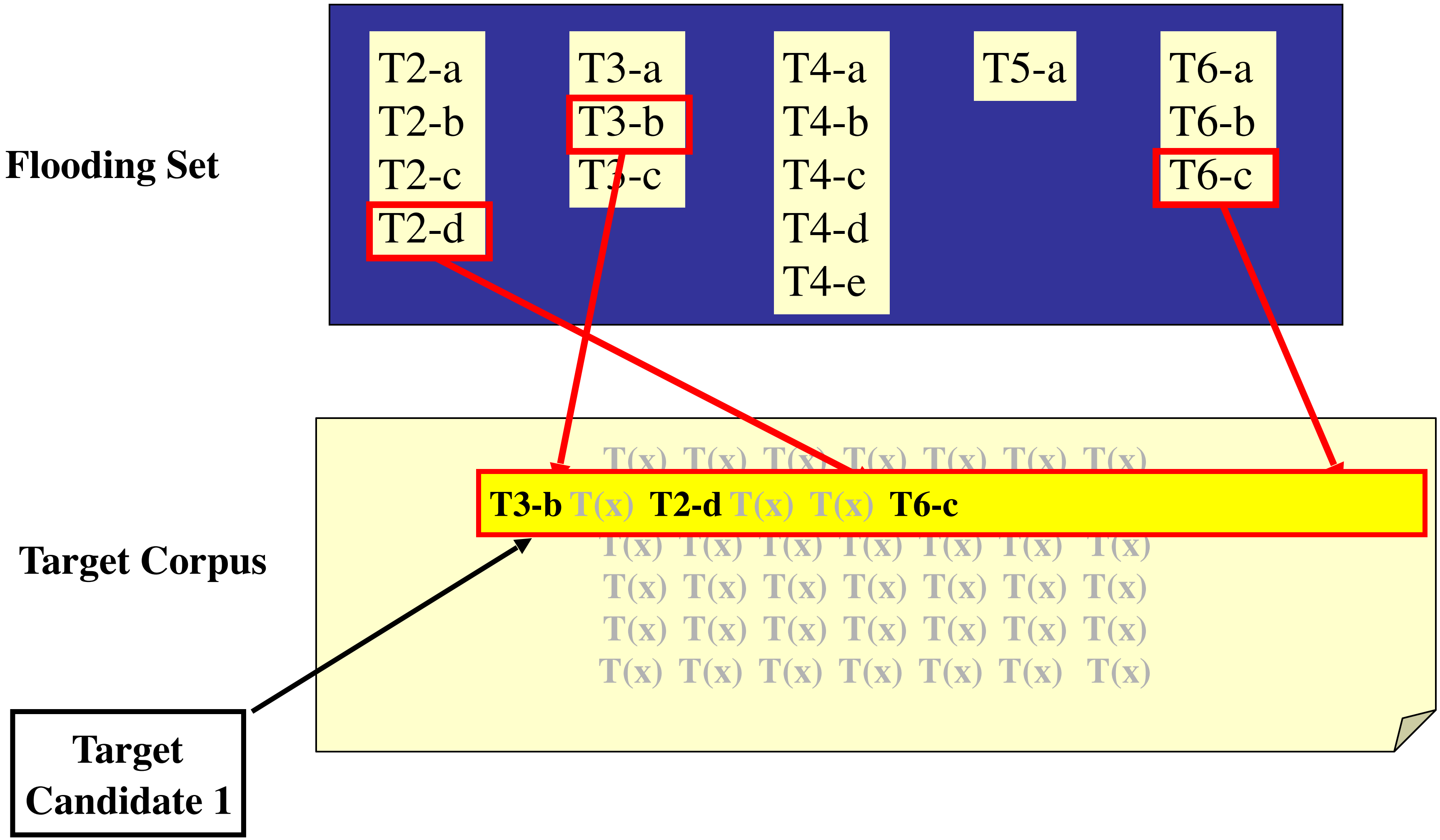
Flooding Set

T2-a	T3-a	T4-a	T5-a	T6-a
T2-b	T3-b	T4-b		T6-b
T2-c	T3-c	T4-c		T6-c
T2-d		T4-d		
		T4-e		

- Find maximum number of words from Flooding Set in minimum length word-string
  - *Words or phrases can be in any order*
  - *Ignore function words in initial step (T5 is a function word in this example)*

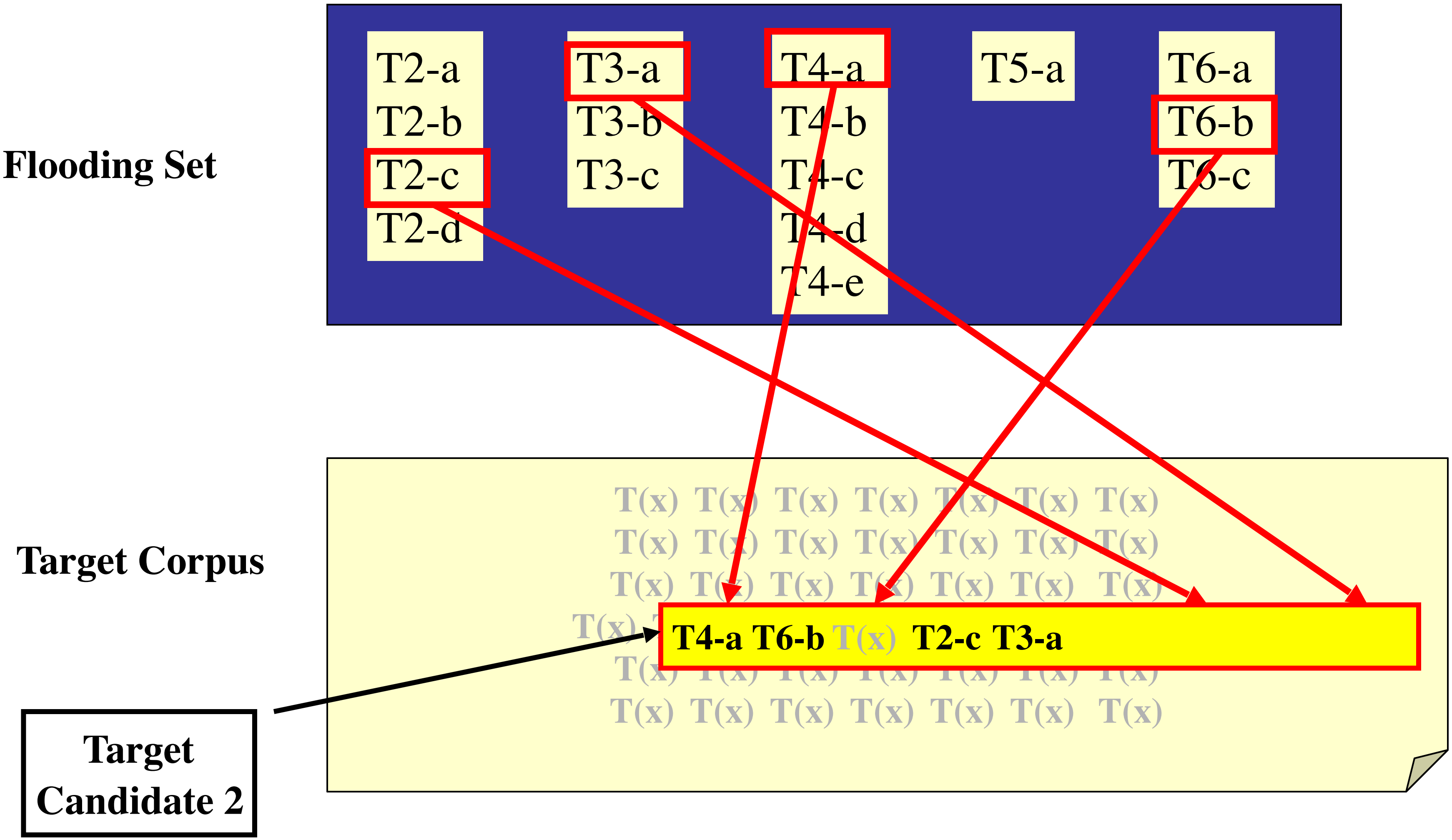
Slide by Jaime Carbonell

Step 3: Search Target Text (Example)



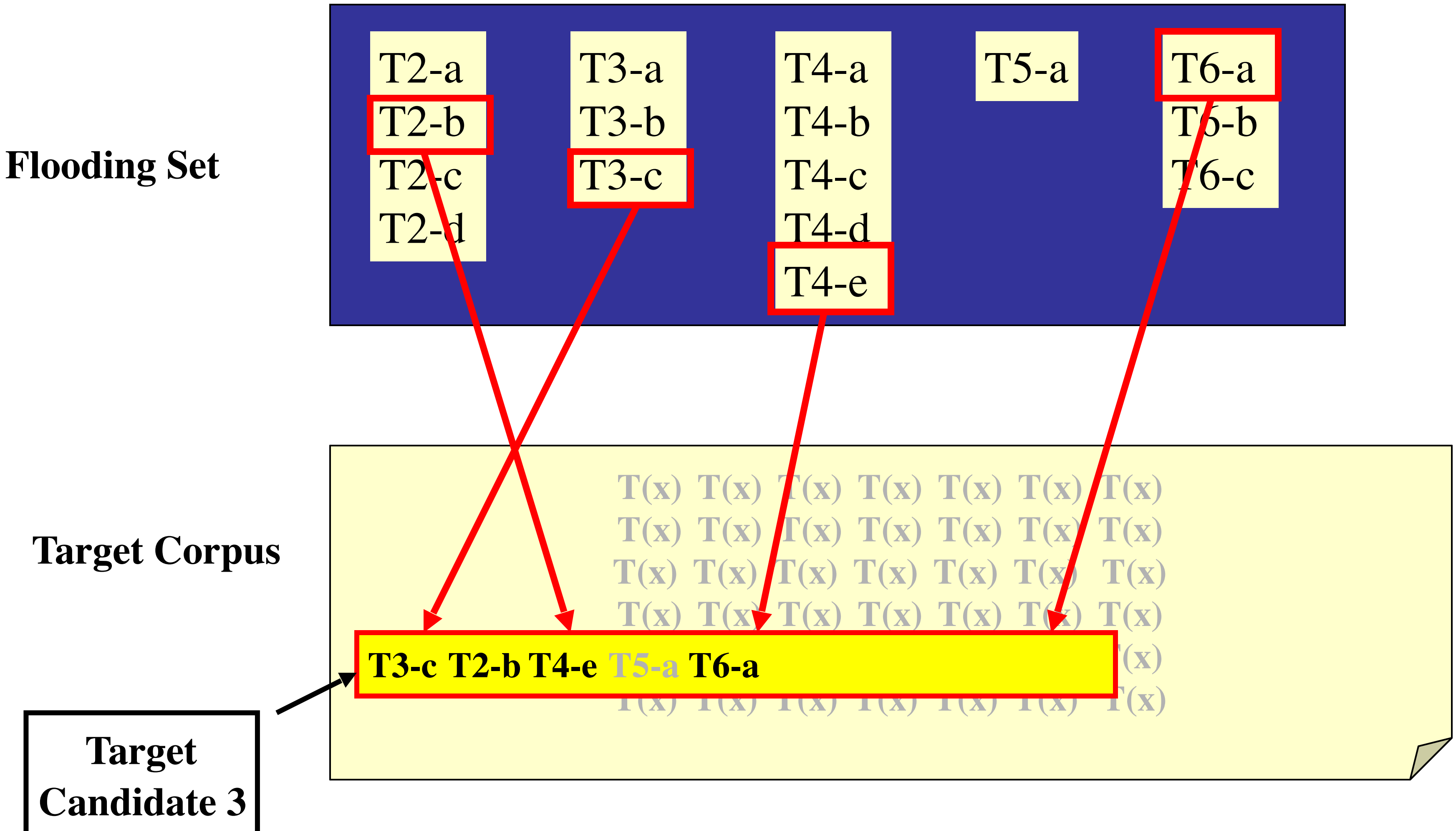
Slide by Jaime Carbonell

Step 3: Search Target Text (Example)



Slide by Jaime Carbonell

Step 3: Search Target Text (Example)

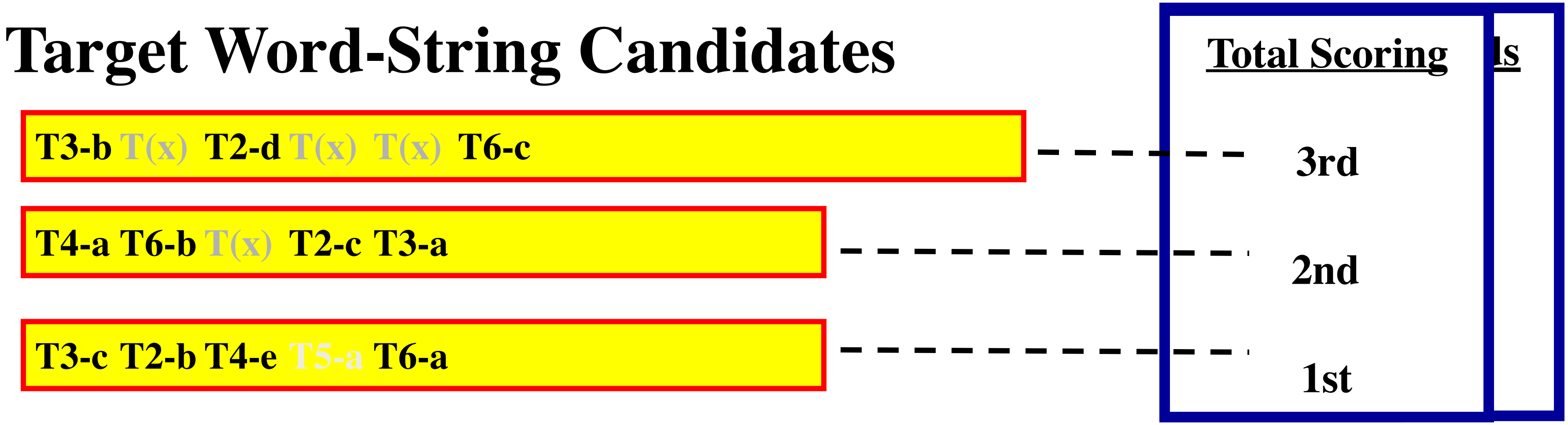


Reintroduce function words after initial match (e.g. T5)

Slide by Jaime Carbonell

Step 4: Score Word-String Candidates

- Scoring of candidates based on:
  - Proximity (minimize extraneous words in target n-gram ≈ precision)
  - Number of word matches (maximize coverage ≈ recall))
  - Regular words given more weight than function words
  - Combine results (e.g., optimize F_1 or p-norm or …)



Slide by Jaime Carbonell

Step 5: Select Candidates Using Overlap
(Propagate context over entire sentence)

Word-String 1 Candidates

T(x1) T2-d T3-c T(x2) T4-b

T(x1) T3-c T2-b T4-e

T(x2) T4-a T6-b T(x3) T2-c

Word-String 2 Candidates

T3-b T(x3) T2-d T(x5) T(x6) T6-c

T4-a T6-b T(x3) T2-c T3-a

T3-c T2-b T4-e T5-a T6-a

Word-String 3 Candidates

T2-b T4-e T5-a T6-a T(x8)

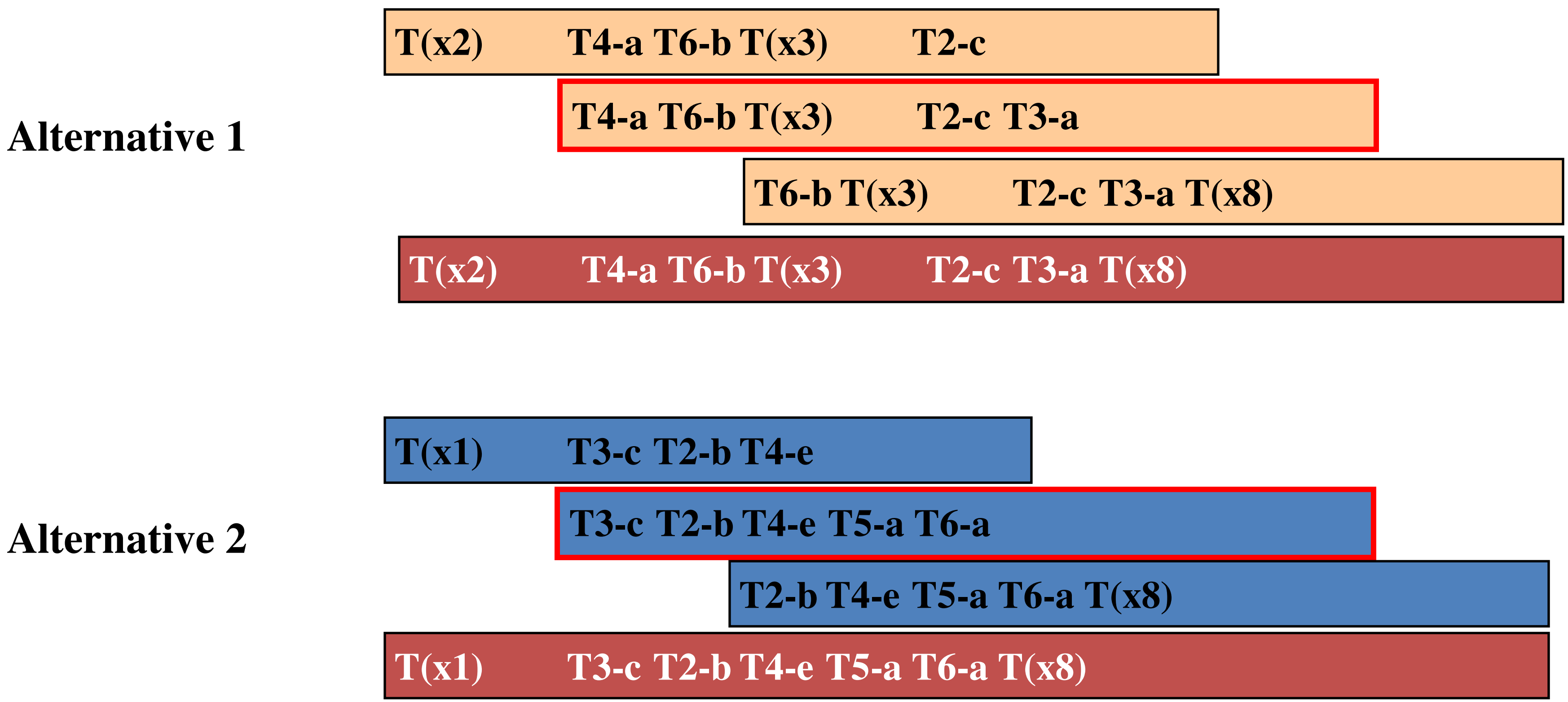
T6-b T(x11) T2-c T3-a T(x9)

T6-b T(x3) T2-c T3-a T(x8)

Slide by Jaime Carbonell

Step 5: Select Candidates Using Overlap

Best translations selected via maximal overlap



Slide by Jaime Carbonell

A (Simple) Real Example of Overlap

Flooding → N-gram fidelity

Overlap → Long range fidelity

N-grams generated from Flooding

a United States soldier

United States soldier died

soldier died and two others

died and two others were injured

two others were injured Monday

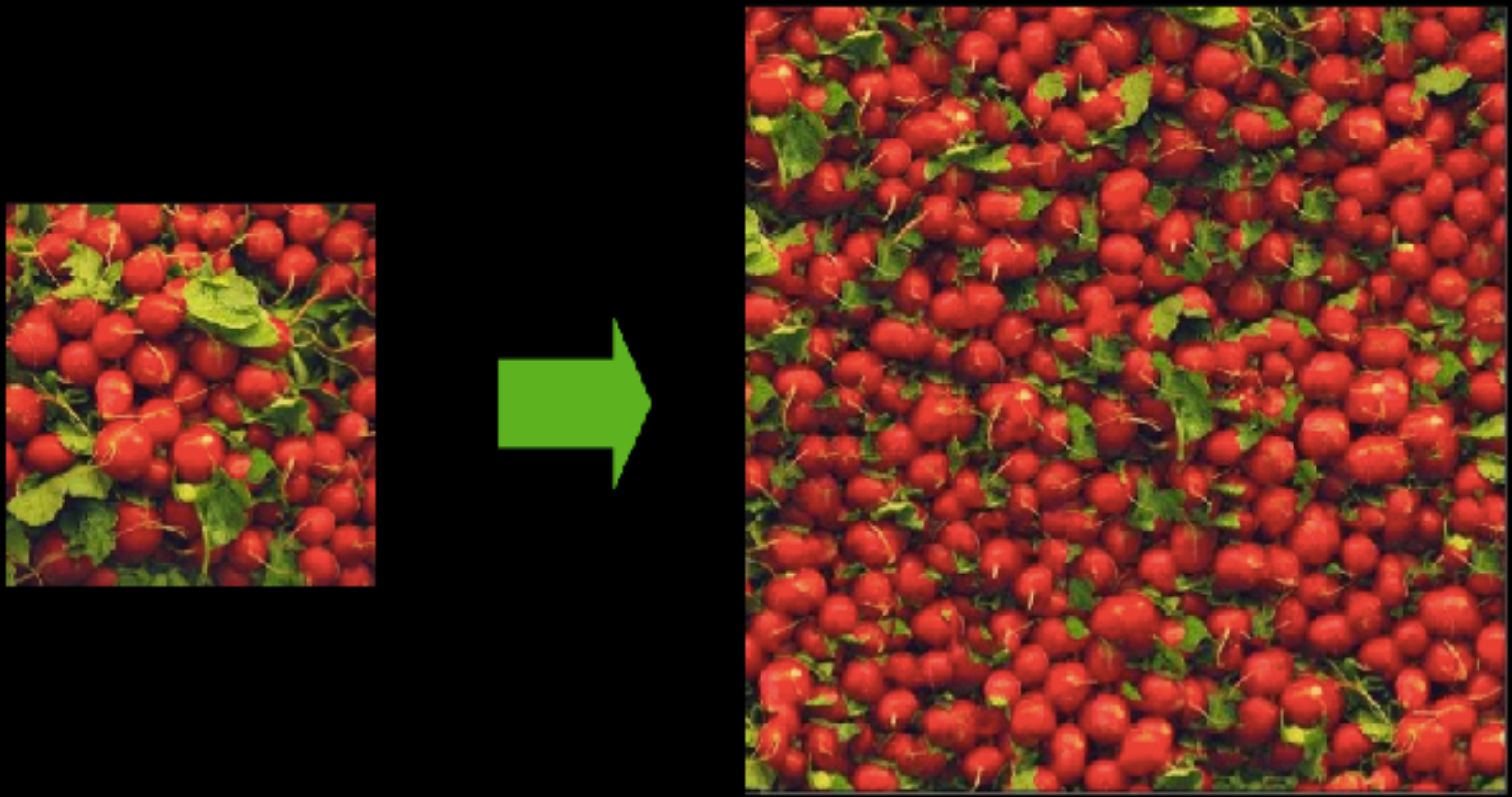
N-grams connected via Overlap

a United States soldier died and two others were injured Monday

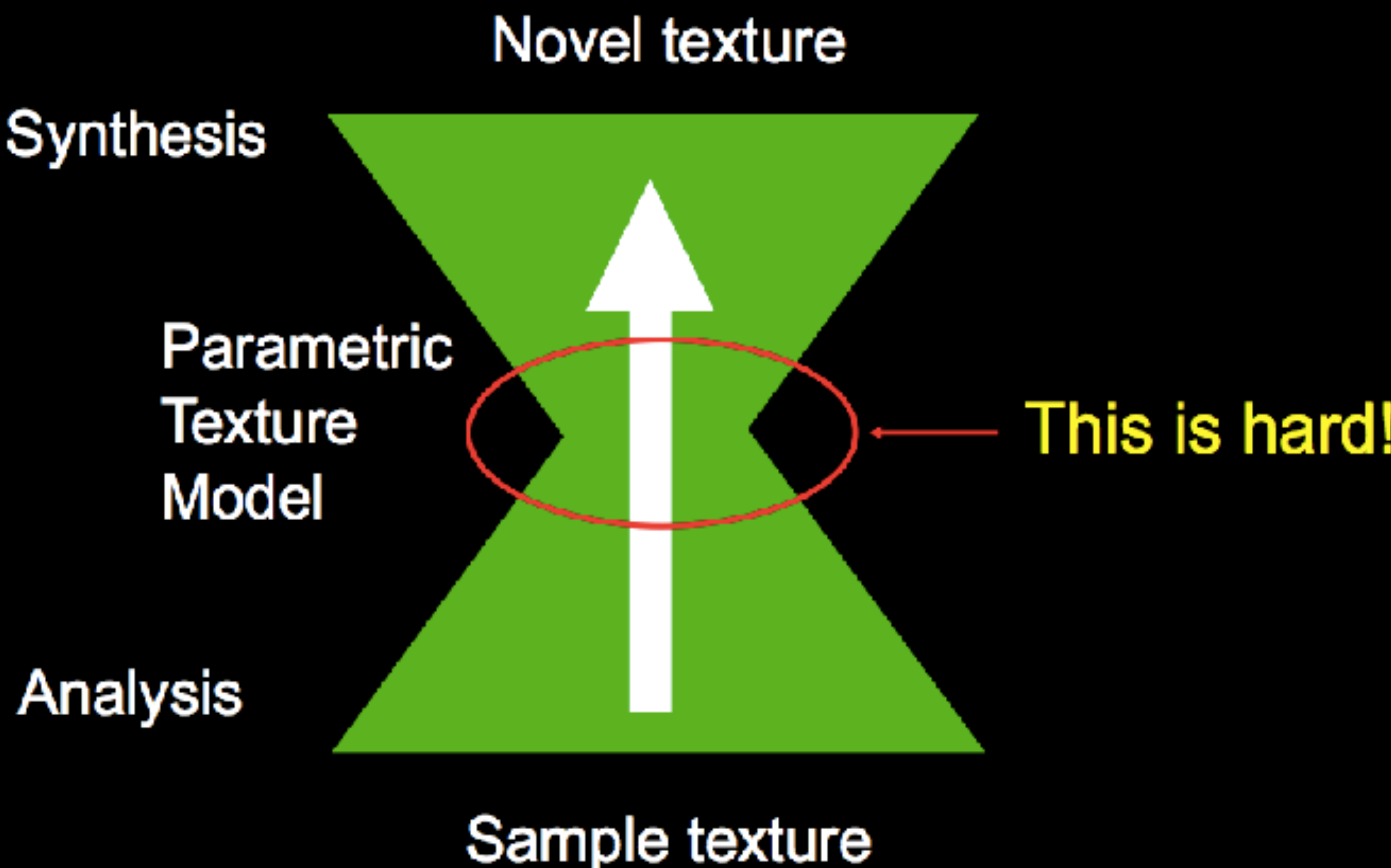
Slide by Jaime Carbonell

Texture Synthesis

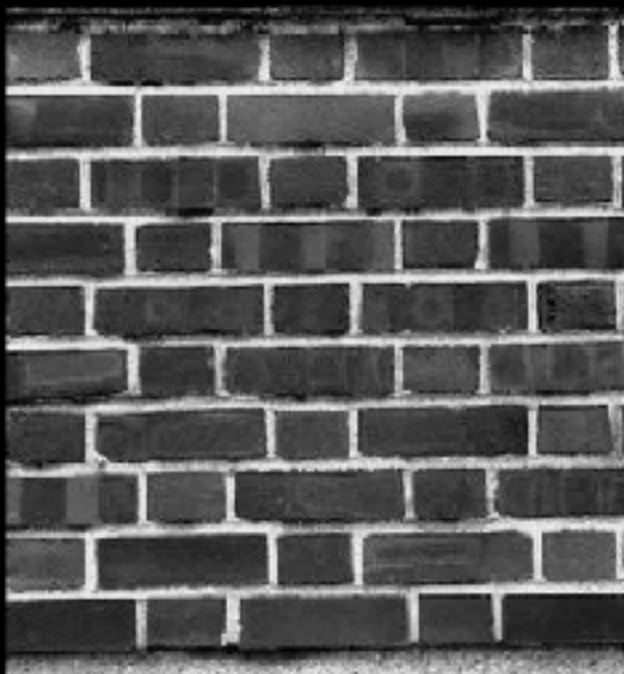
Texture Synthesis



Classical Texture Synthesis



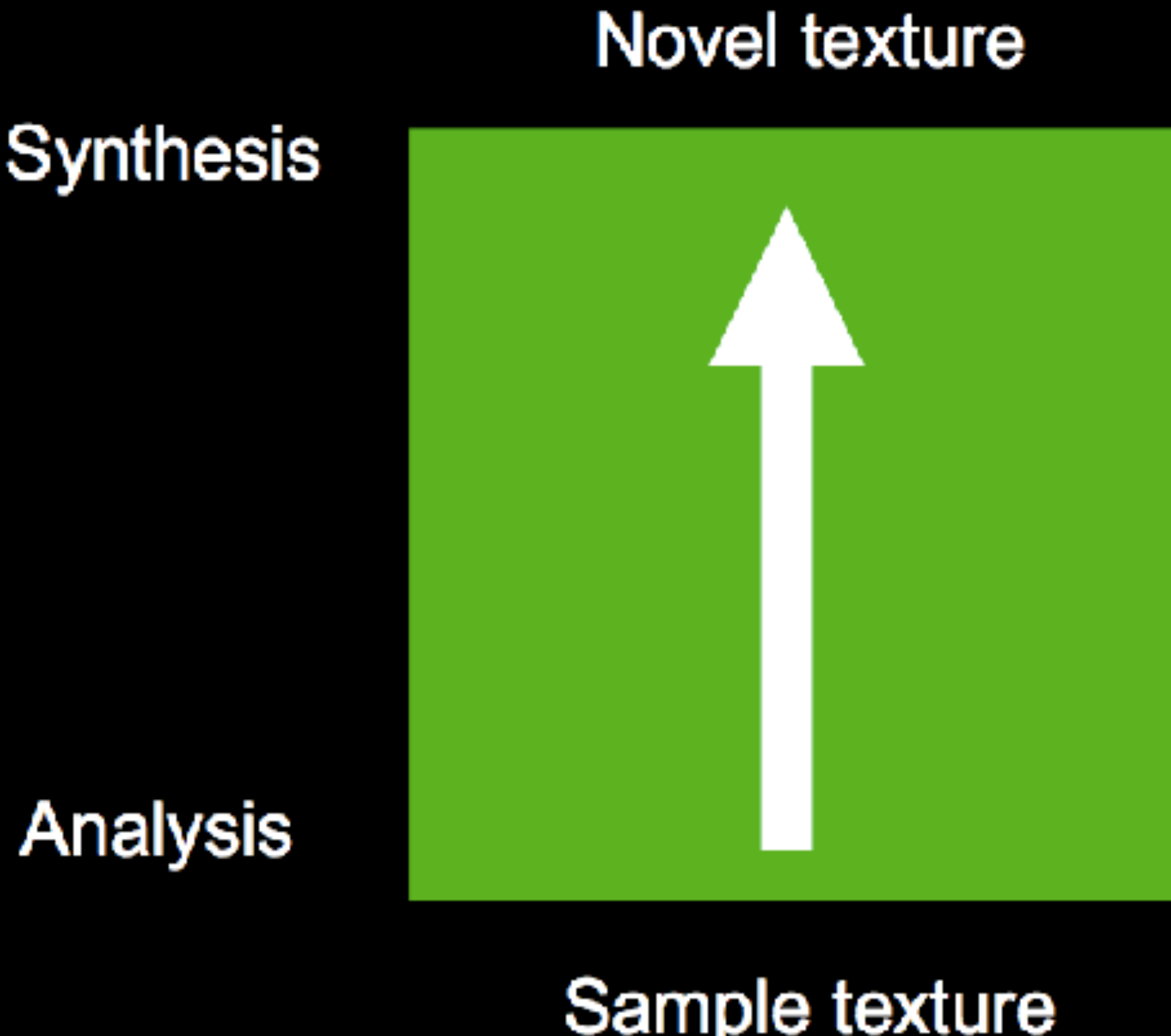
Throwing away too much too soon?



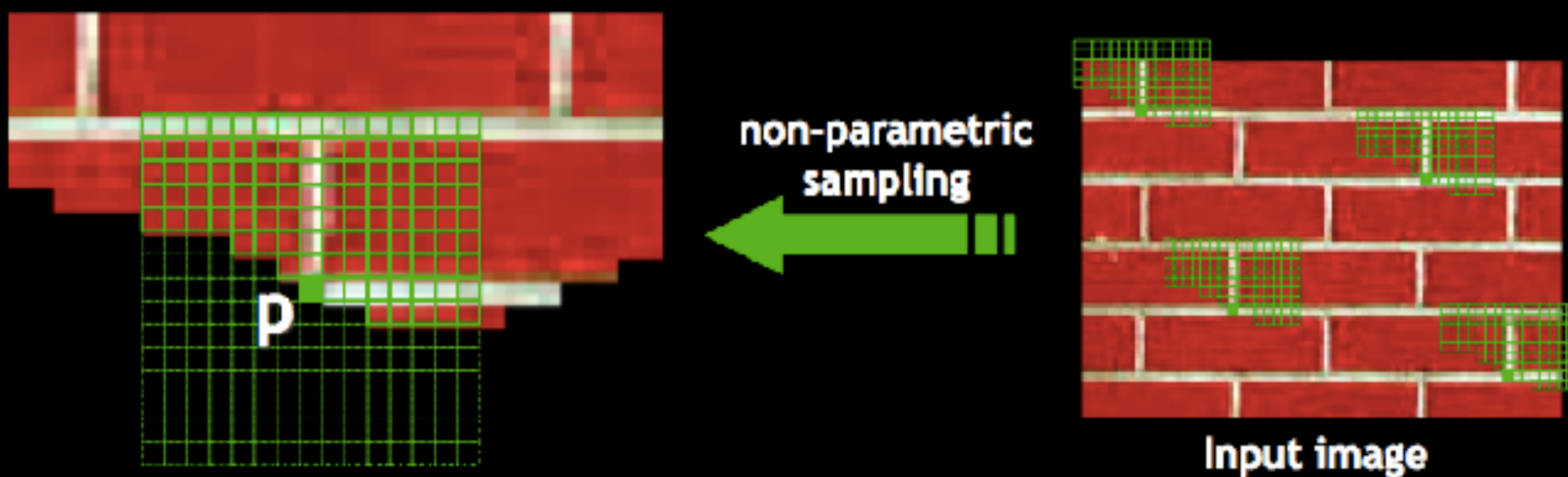
input texture

synthesized texture

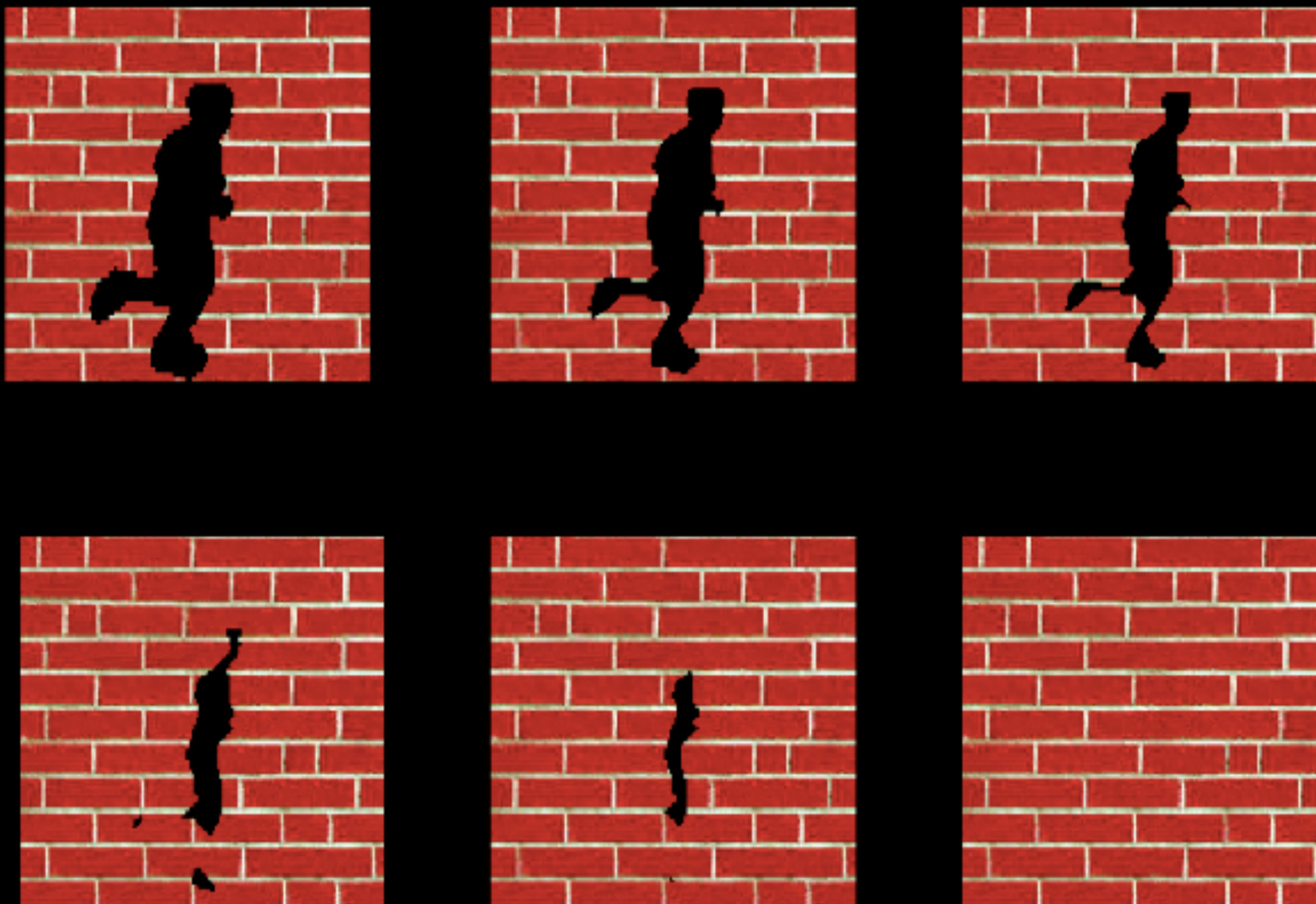
Non-parametric Approach



[Efros & Leung, '99, Efros & Freeman '01]

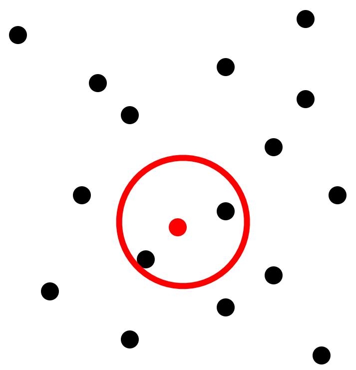


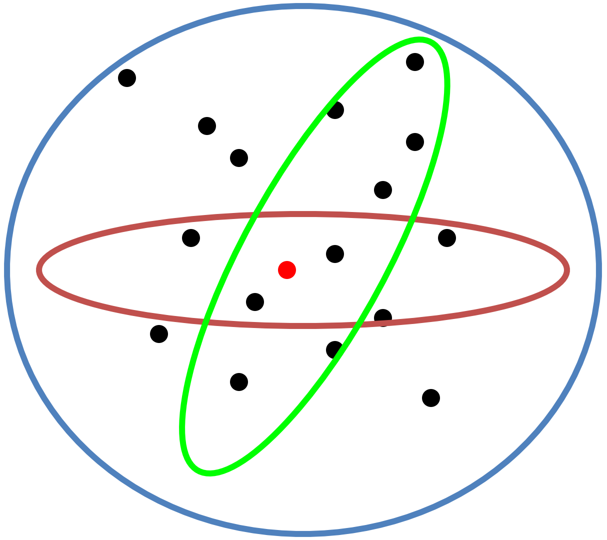
Texture Growing



So, how do we use big data?

Two ways to use Lots of Data



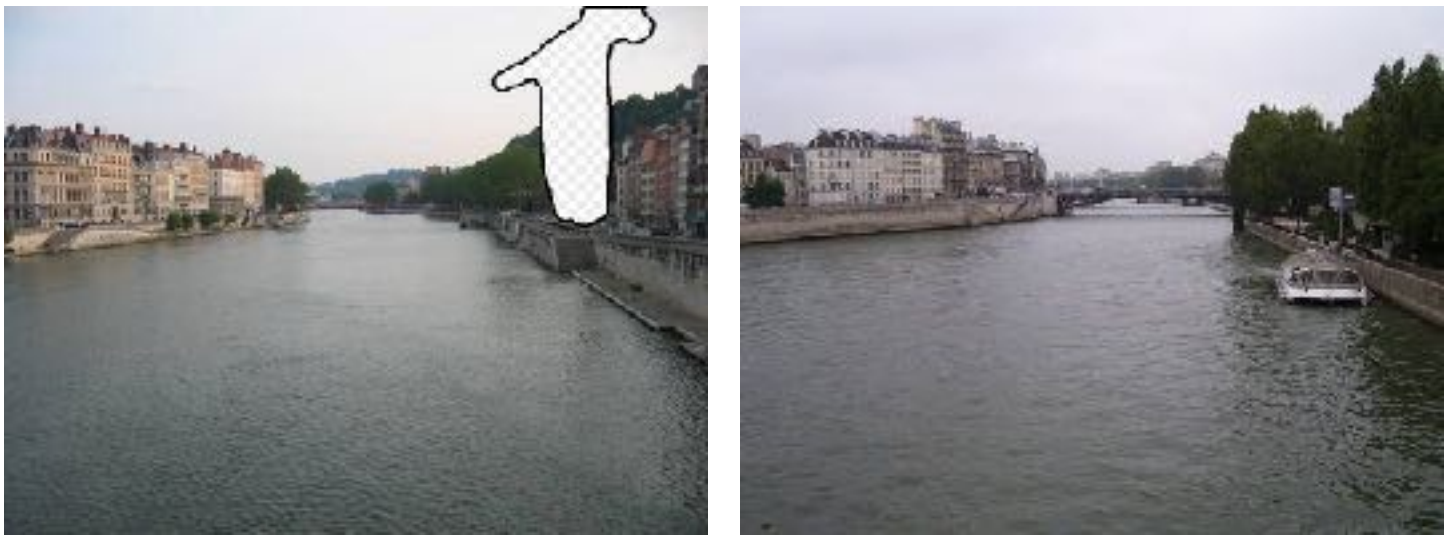


Brute Force Vision: Find that needle in the haystack and disregard the rest (a.k.a. kNN)

See what different subsets of data think of you

kNN matching is great…

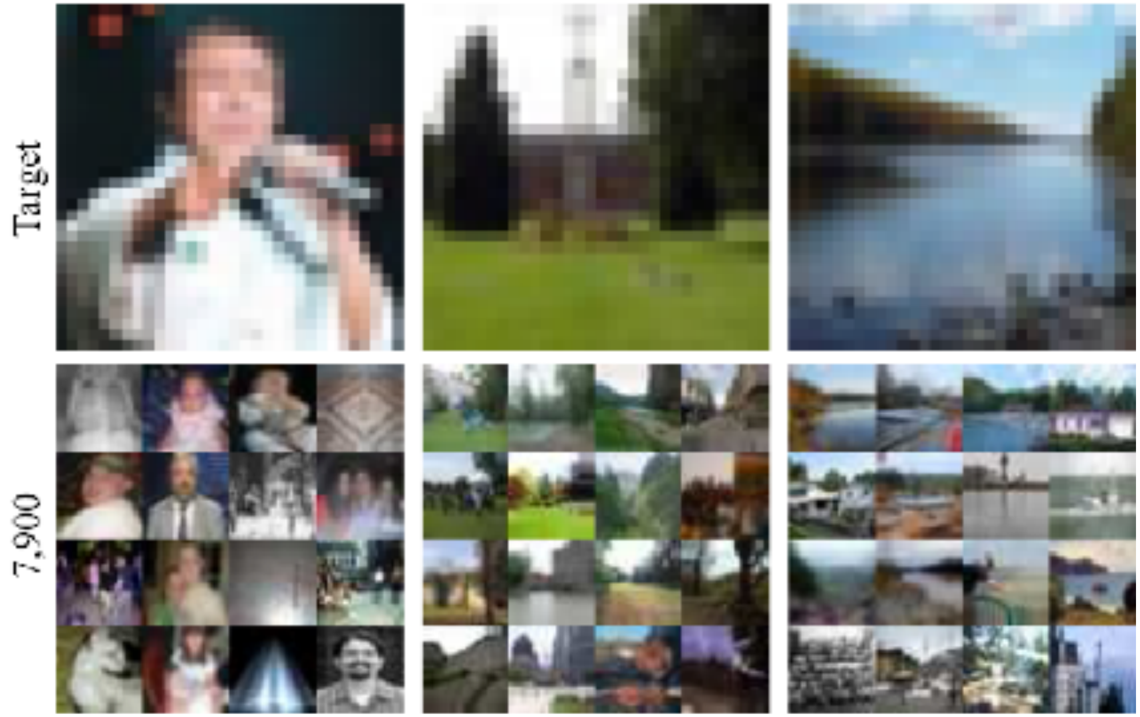
- **because we live in a (mostly) boring world!**



Lots

Of

Images

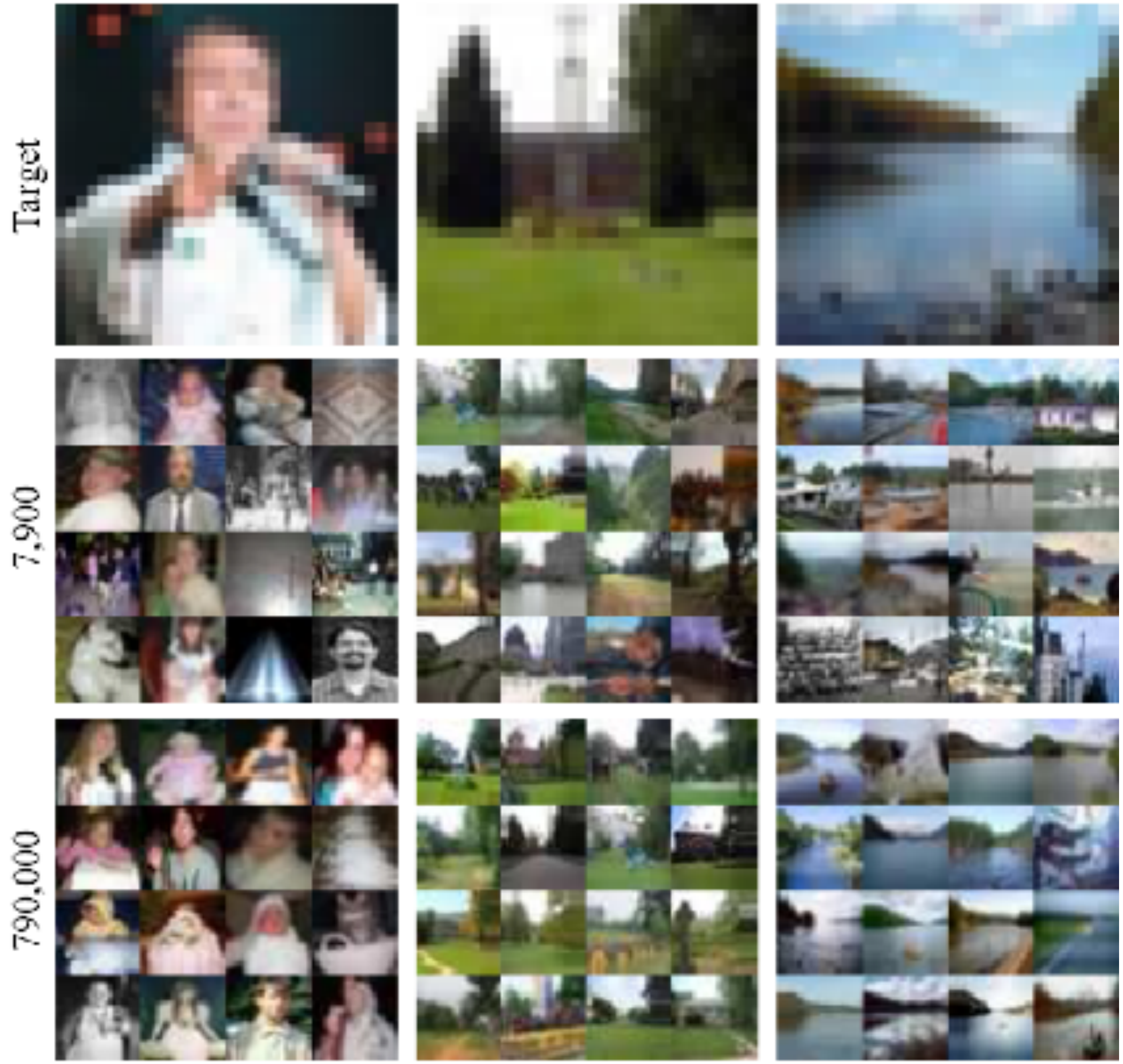


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images

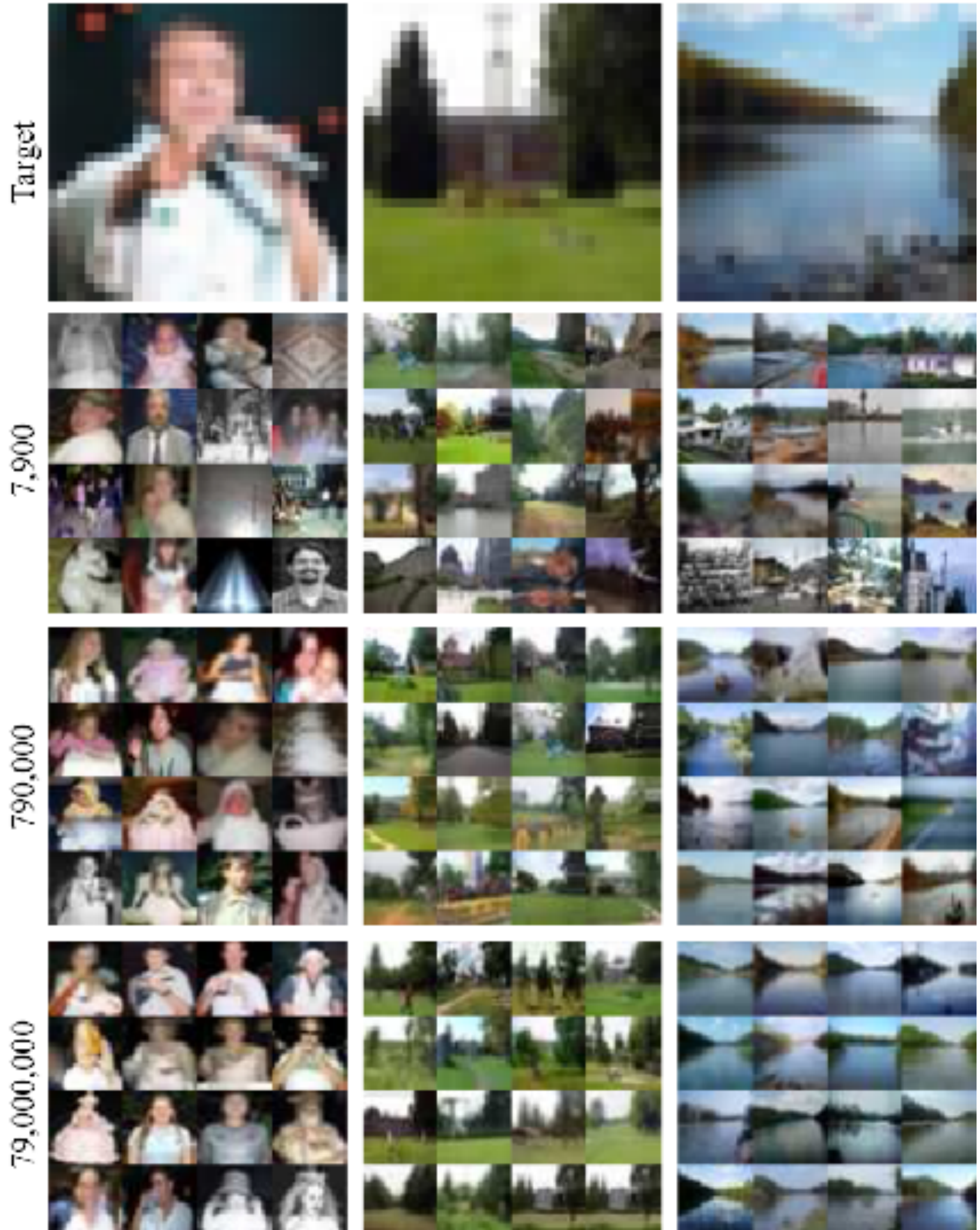


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

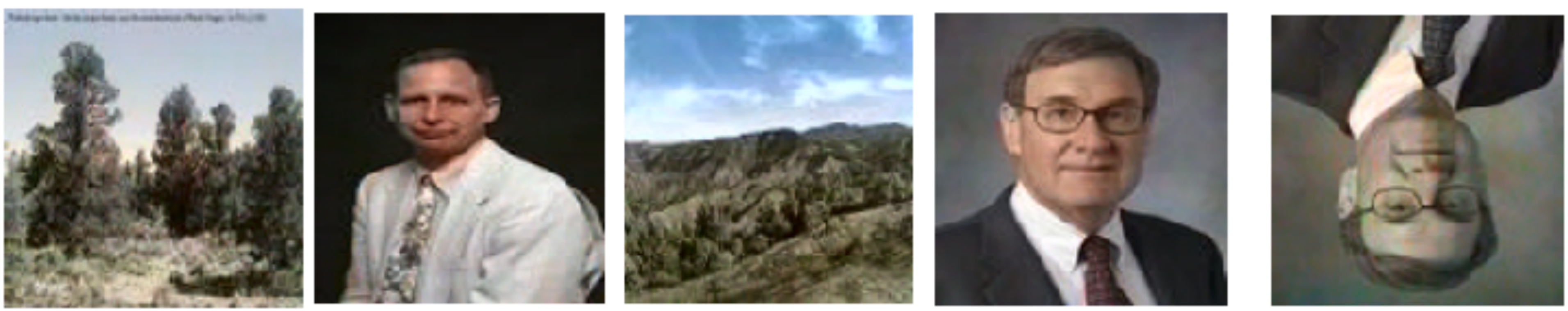
Images



Automatic Colorization Result

Grayscale input High resolution

Colorization of input using average



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

im2gps

Instead of using objects labels, the web provides other kinds of metadata associate to large collections of images

Figure 2. The distribution of photos in our database. Photo locations are cyan. Density is overlaid with the jet colormap (log scale).

20 million geotagged and geographic text-labeled images

Hays & Efros. CVPR 2008

im2gps

Hays & Efros. CVPR 2008

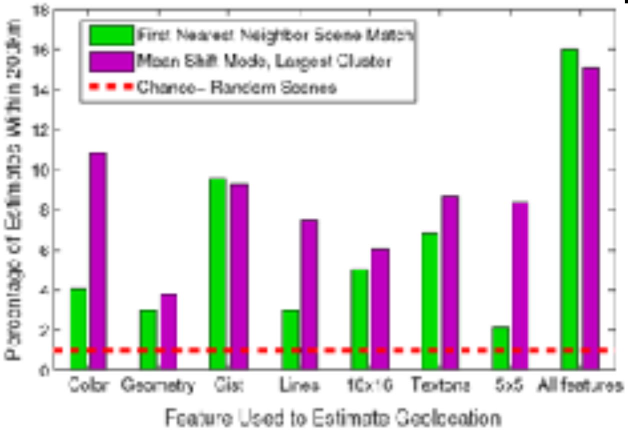


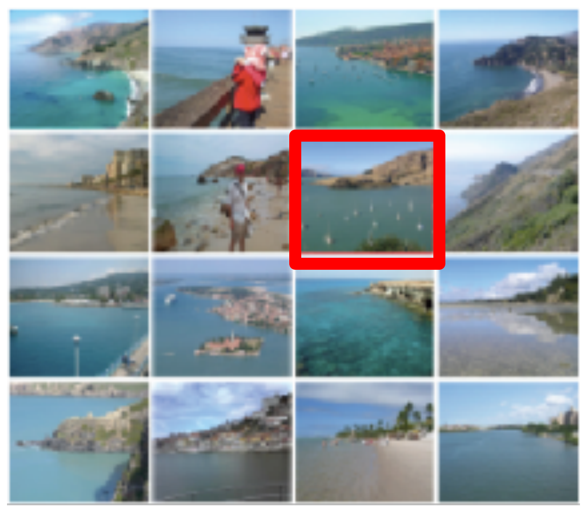
Figure 5. *Geolocation performance across features.* Percentage of test cases geolocated to within 200*km* for each feature. We compare geolocation by 1-NN vs. largest mean-shift mode.

Image completion

Original Image | Input | Criminisi et al. | MS *Smart Erase*

Instead, generate proposals using millions of images

Input

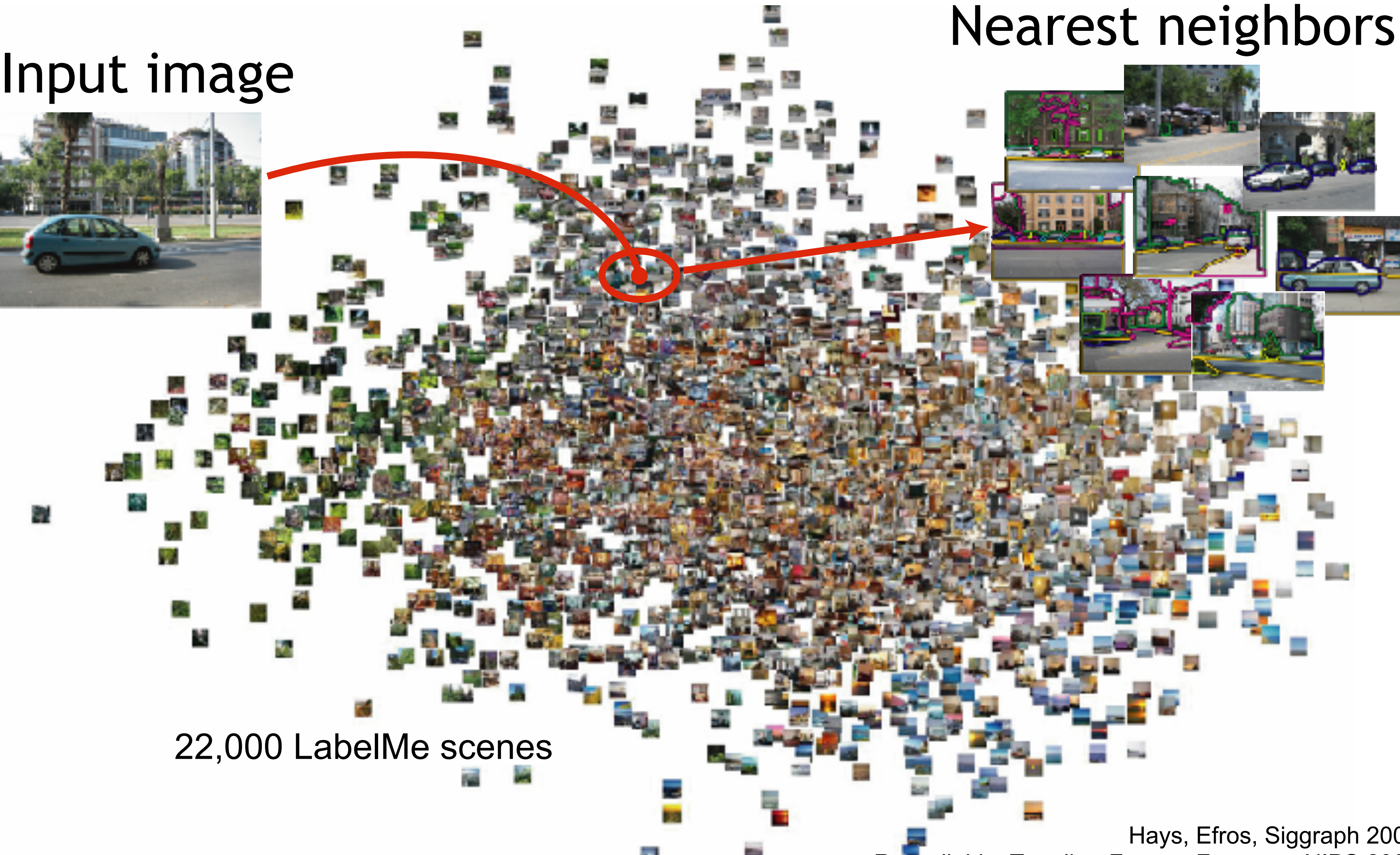


16 nearest neighbors
(gist+color matching)

output

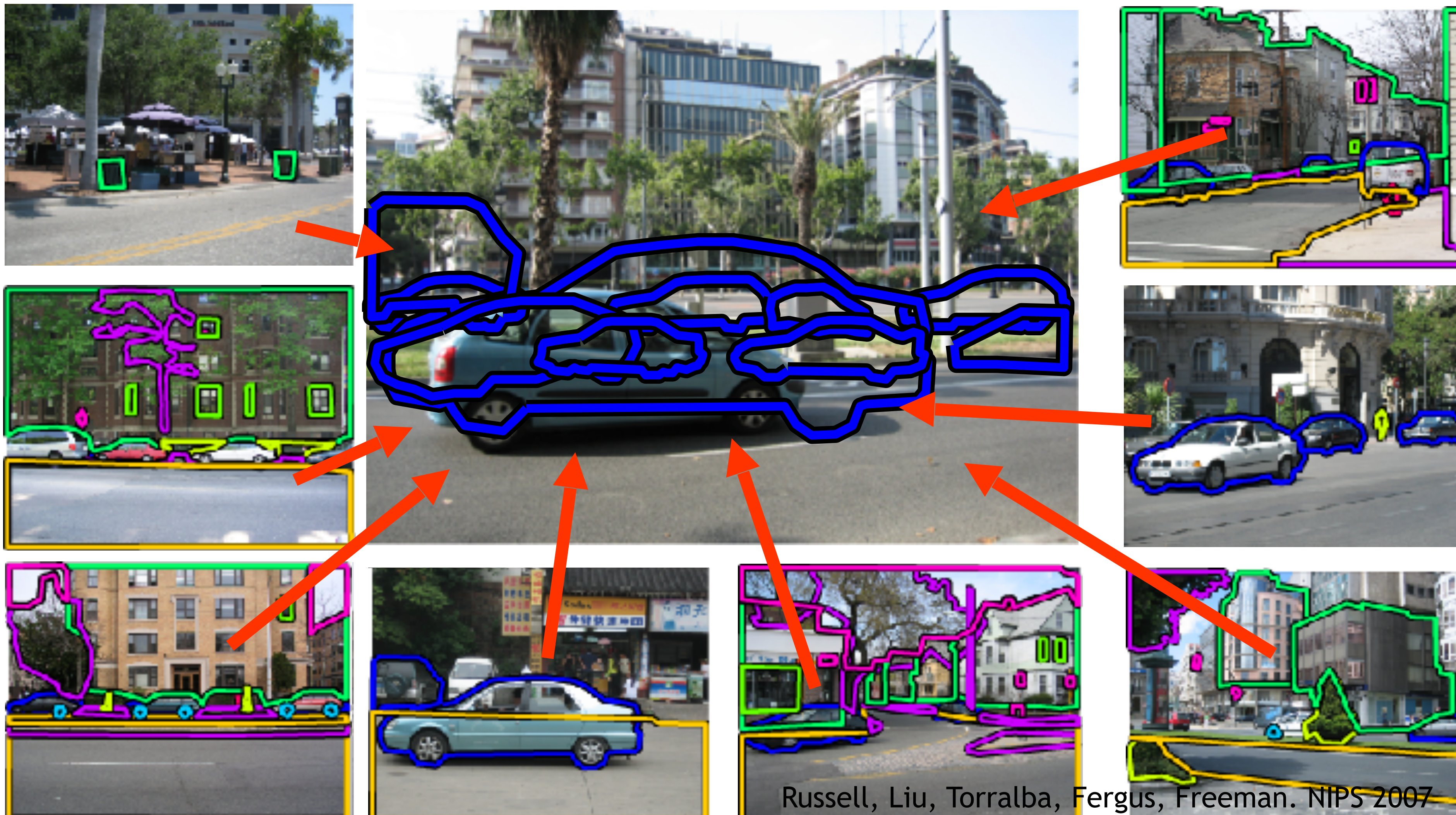
Hays, Efros, 2007

With a good image similarity and a lot of data…



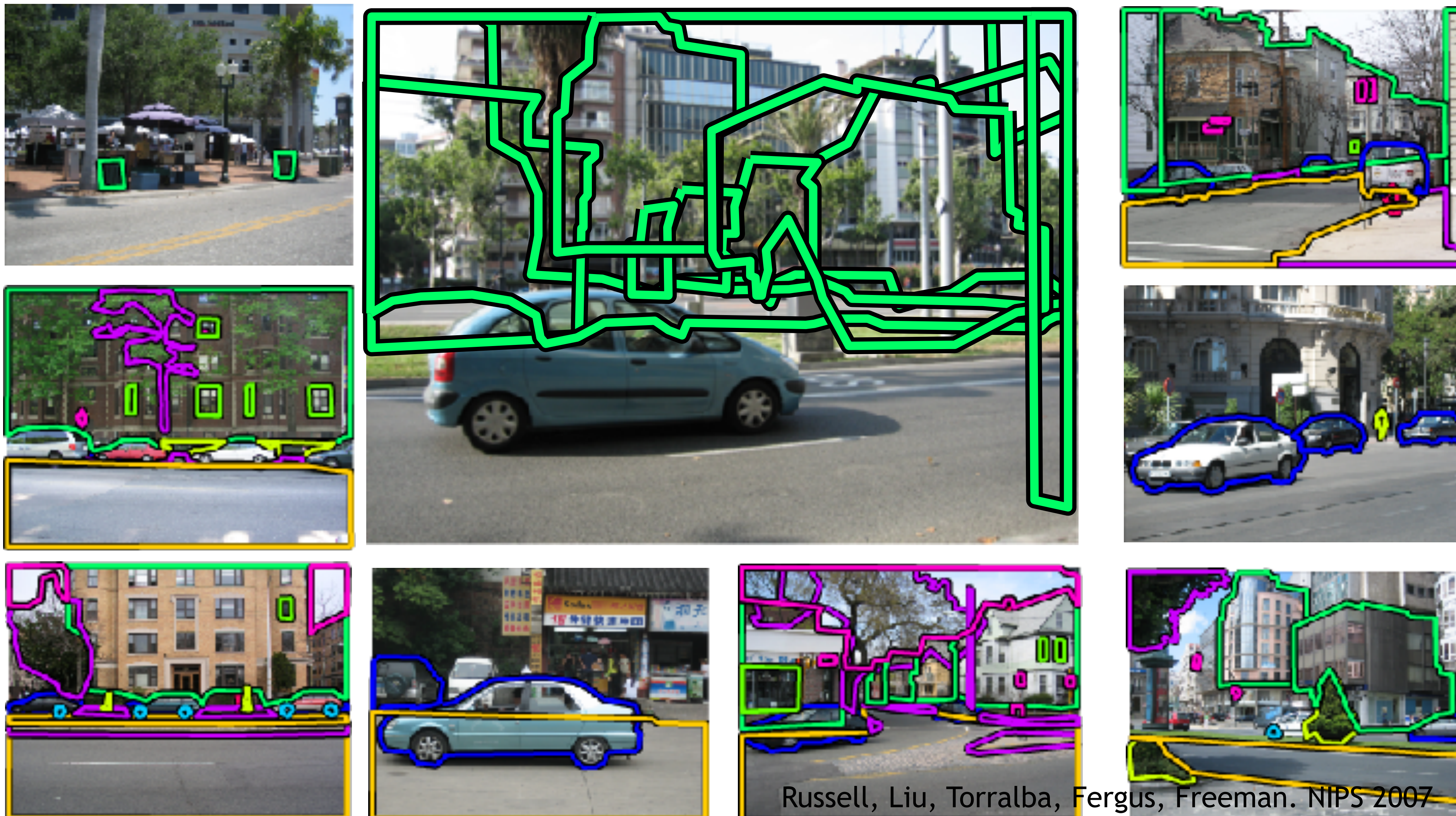
Hays, Efros, Siggraph 2006
Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007

With a good image similarity and a lot of data…



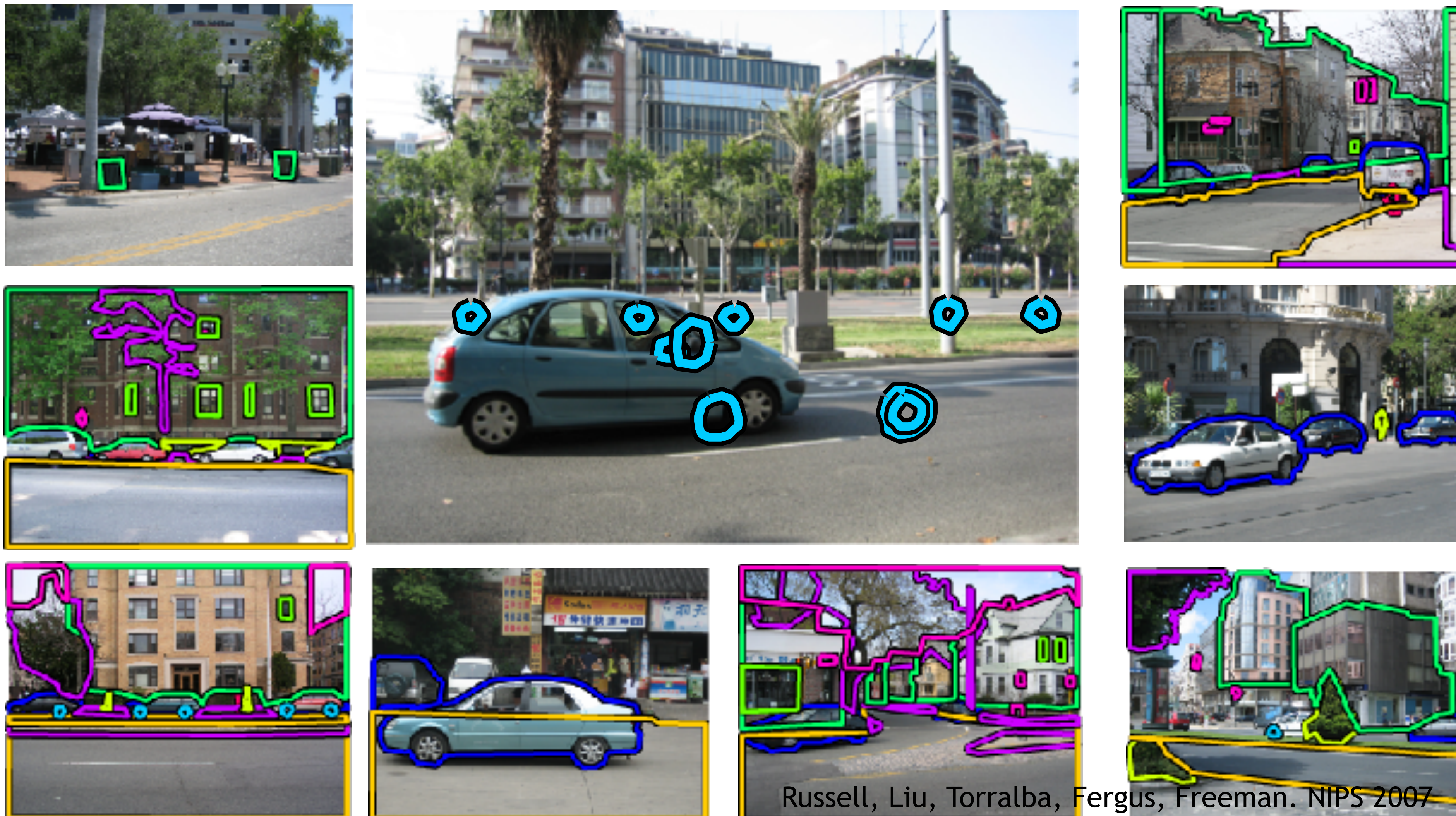
Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007

With a good image similarity and a lot of data…



Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007

With a good image similarity and a lot of data…

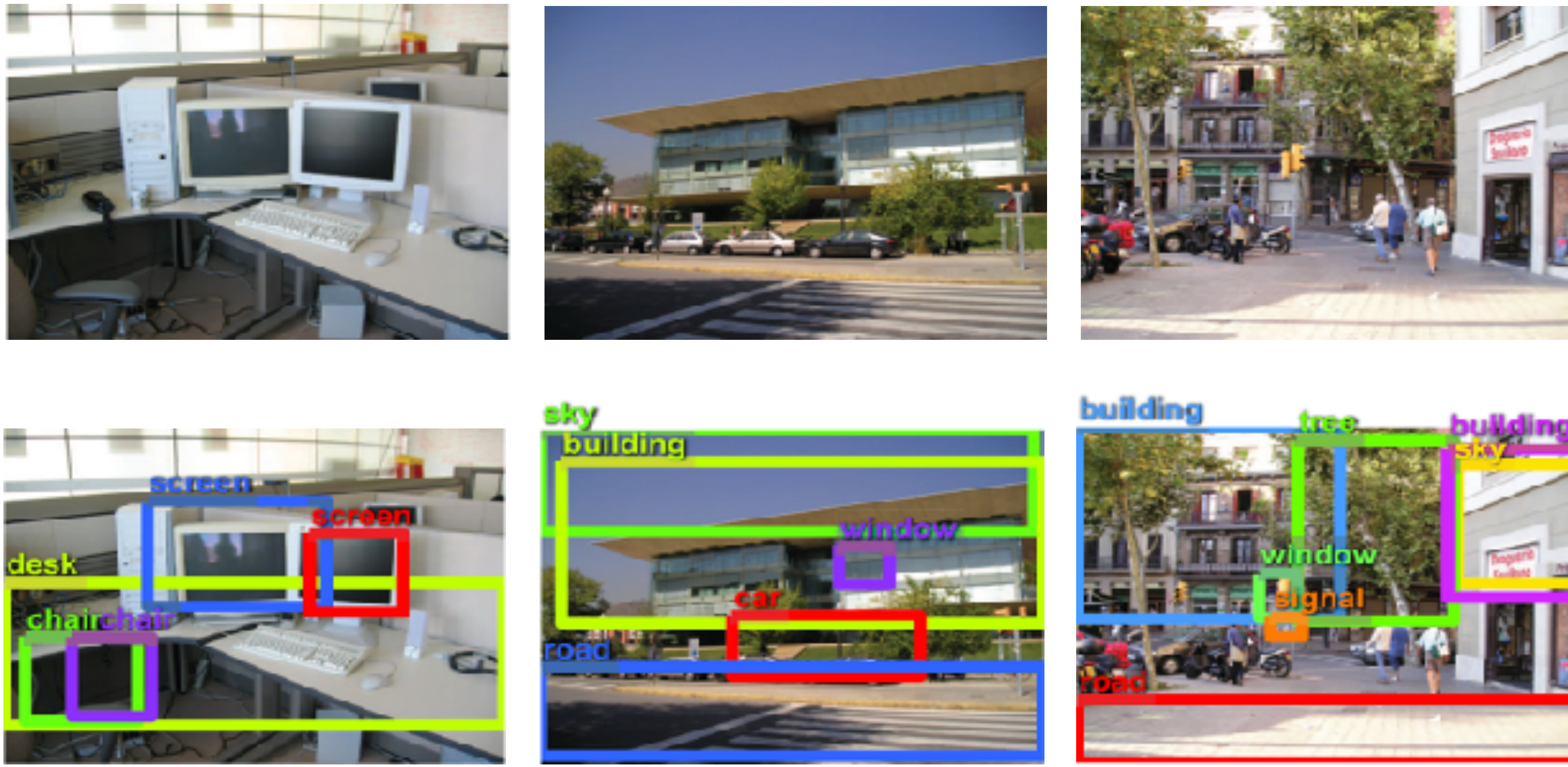


Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007

With a good image similarity and a lot of data…

Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007

Outputs



Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007

While many scenes are boring…

Slide by Antonio Torralba

Some scenes are unique



Slide by Antonio Torralba

Dealing with sparse data (rare scenes)

- better similarity

Medici Fountain, Paris

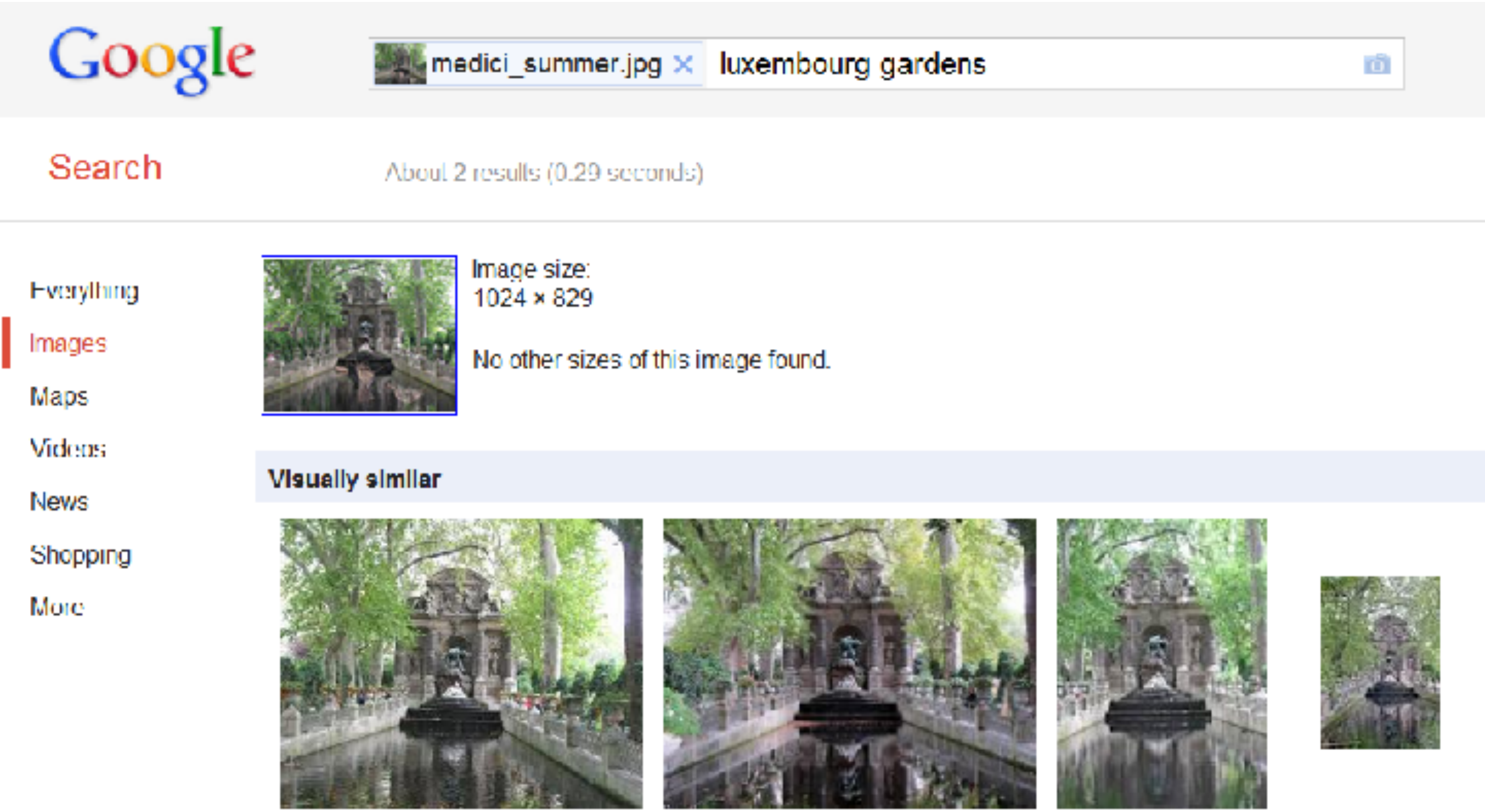
Google images

Search by image

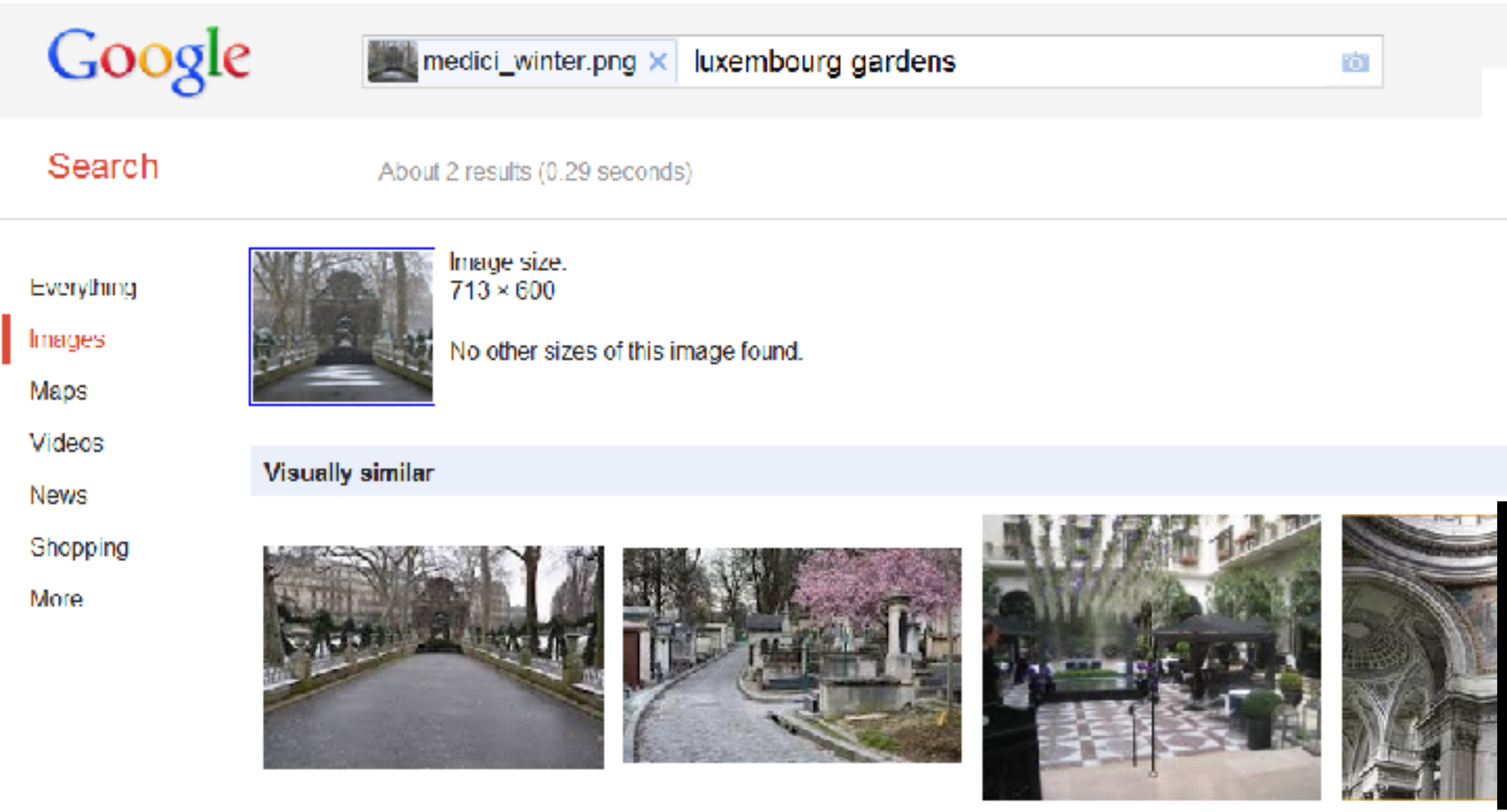
Drop image here

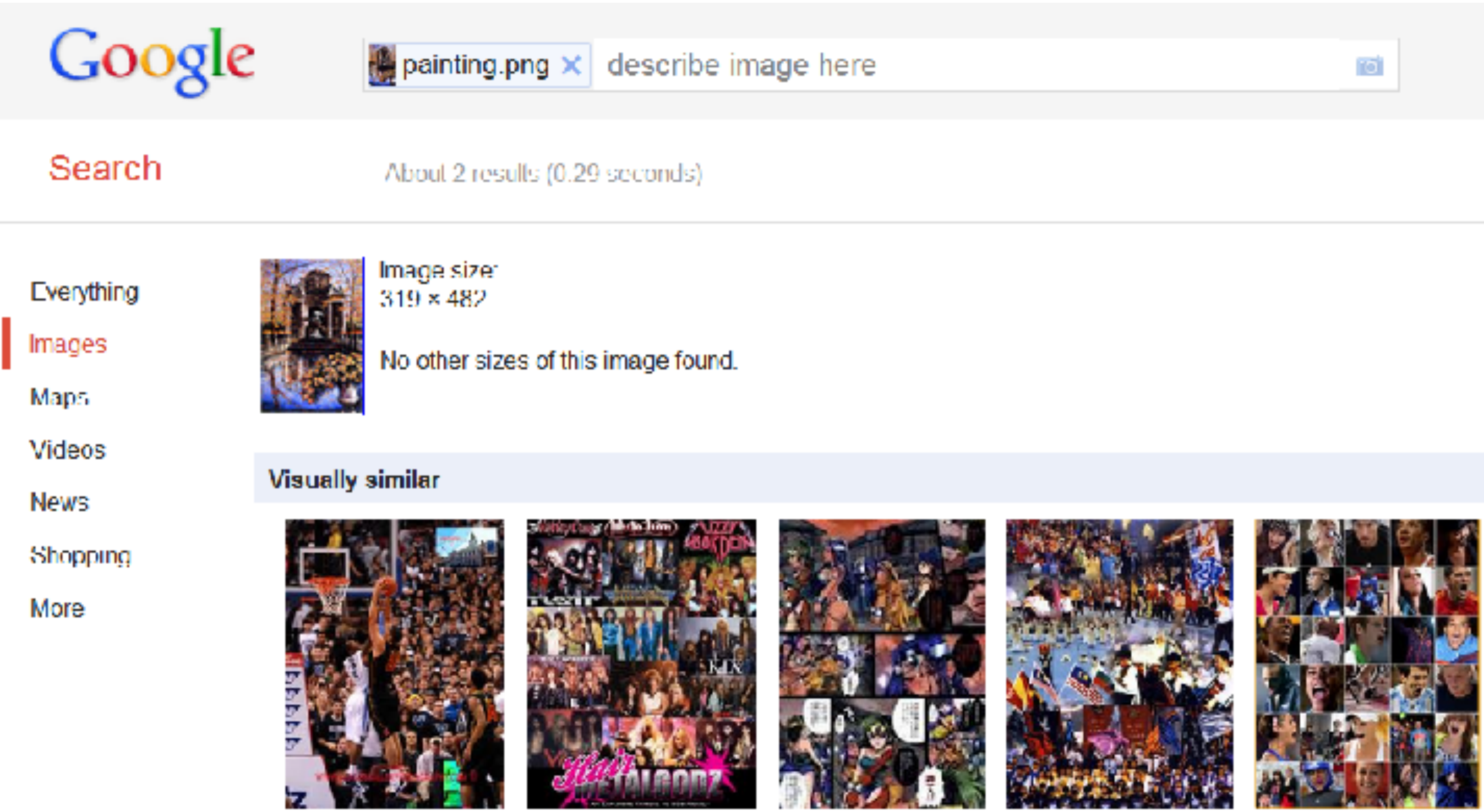
Move

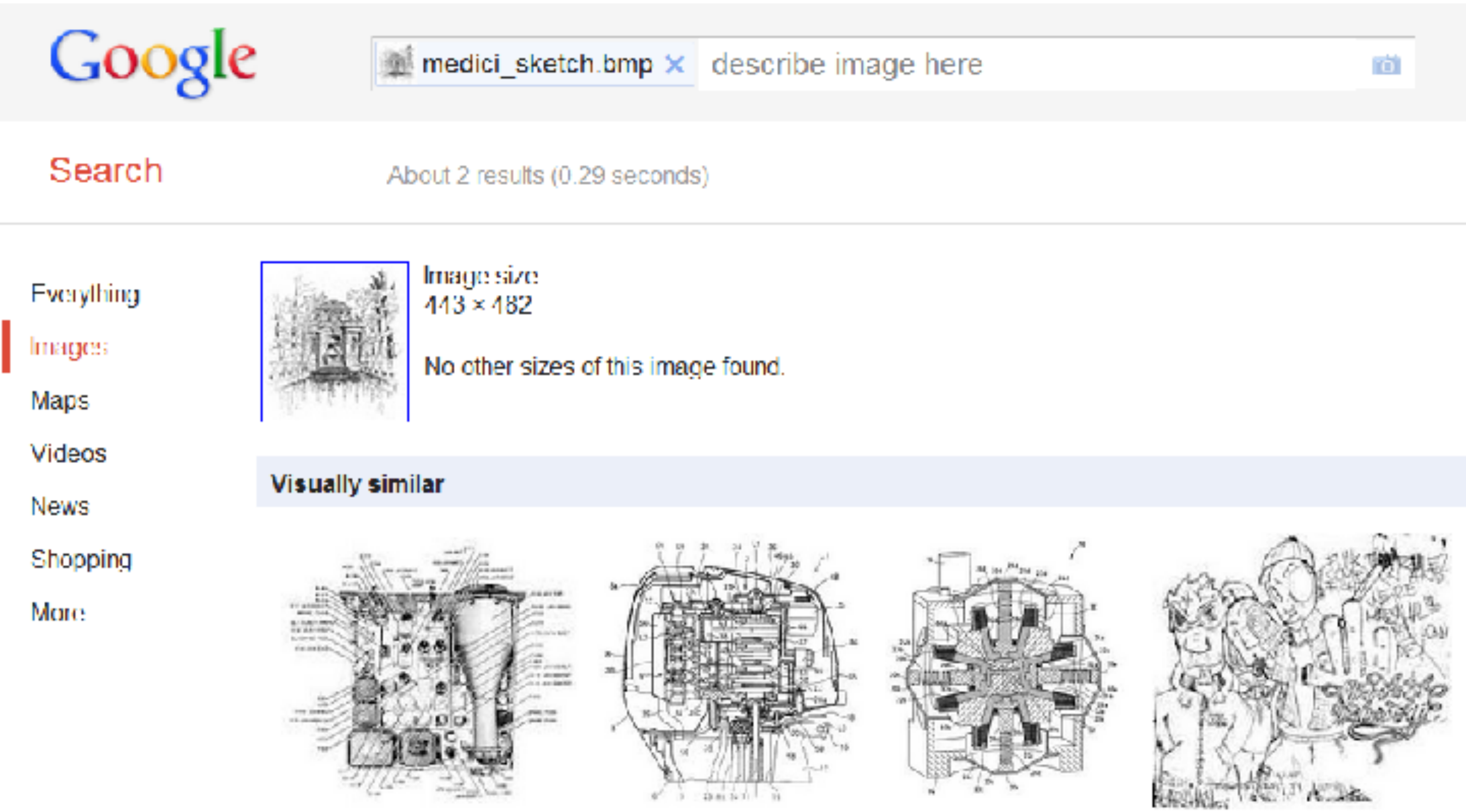
Watch a short video to learn more.


Google
medici_summer.jpg
luxembourg gardens
Search
About 2 results (0.29 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Image size:
1024 × 829
No other sizes of this image found.
Visually similar

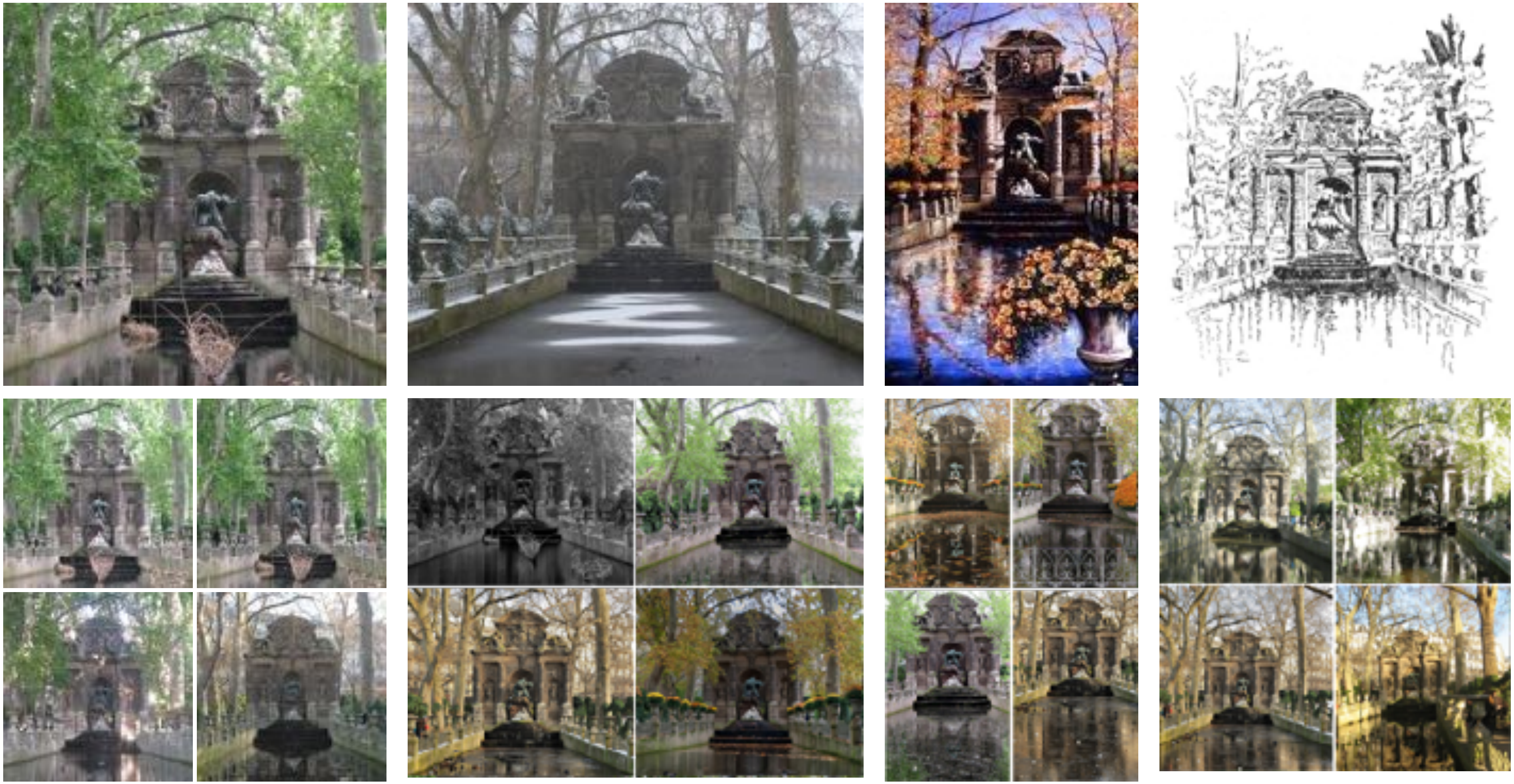
Medici Fountain, Paris (winter)

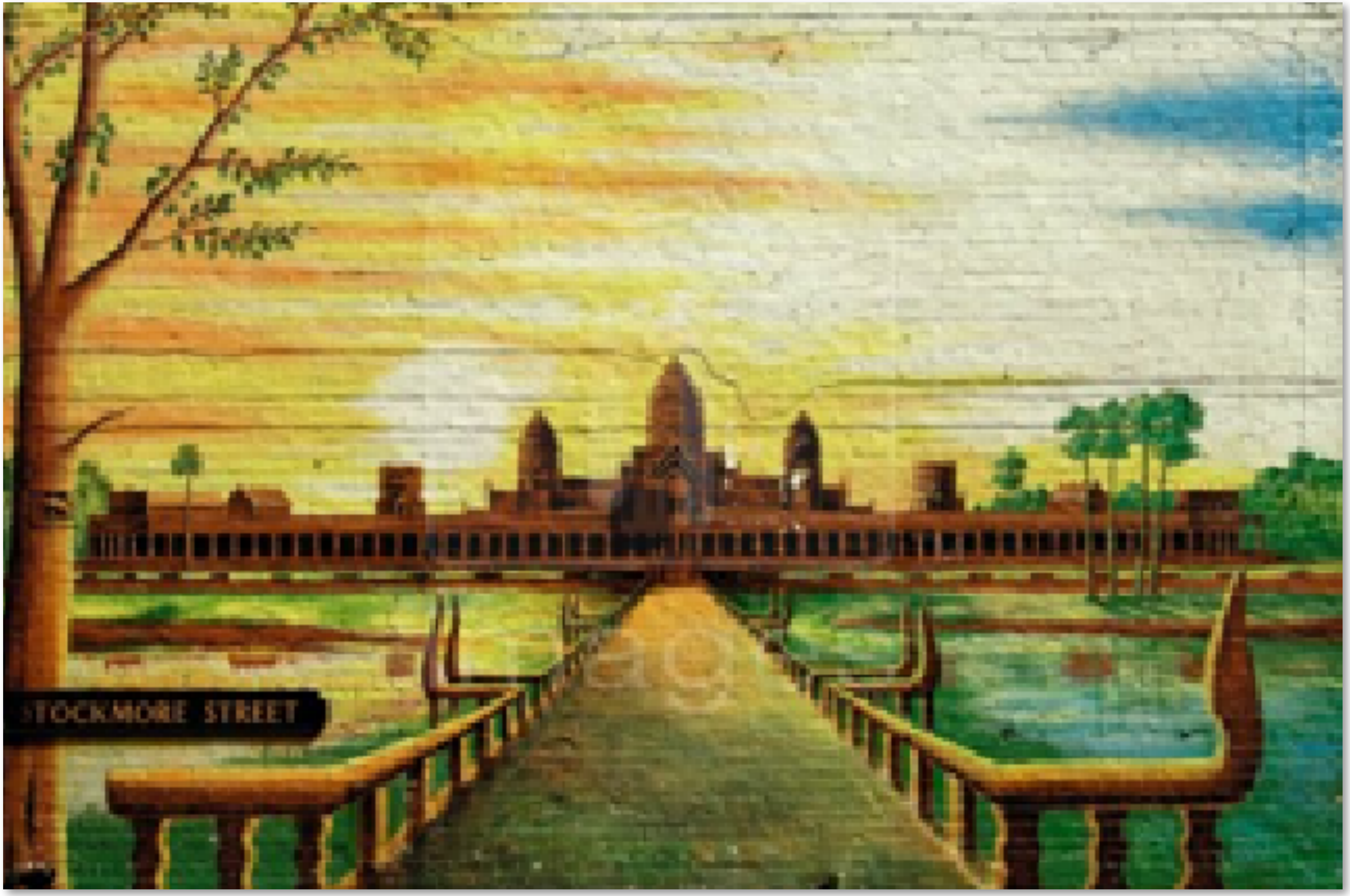

Google
medici_winter.png
luxembourg gardens
Search
About 2 results (0.29 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Image size:
713 × 600
No other sizes of this image found.
Visually similar


Google
painting.png ×
describe image here
Search
About 2 results (0.29 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Image size:
319 × 482
No other sizes of this image found.
Visually similar

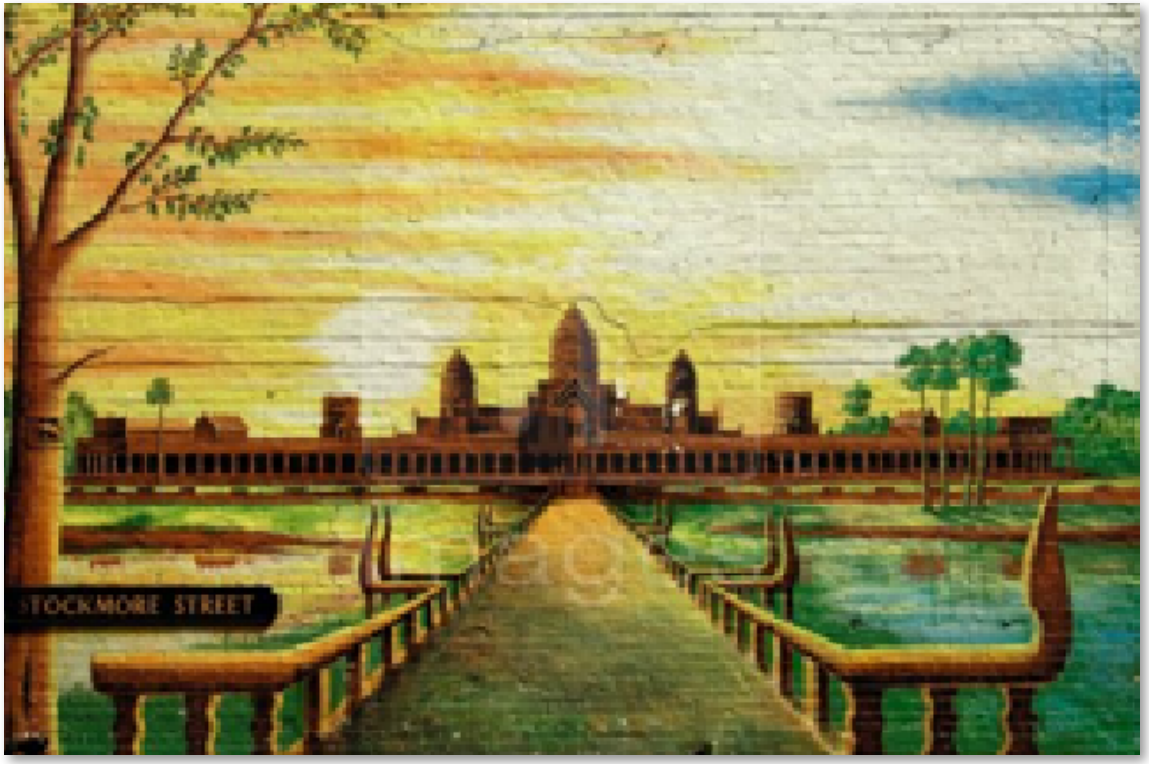

Google
medici_sketch.bmp
describe image here
Search
About 2 results (0.29 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Image size
443 × 482
No other sizes of this image found.
Visually similar

OUR GOAL



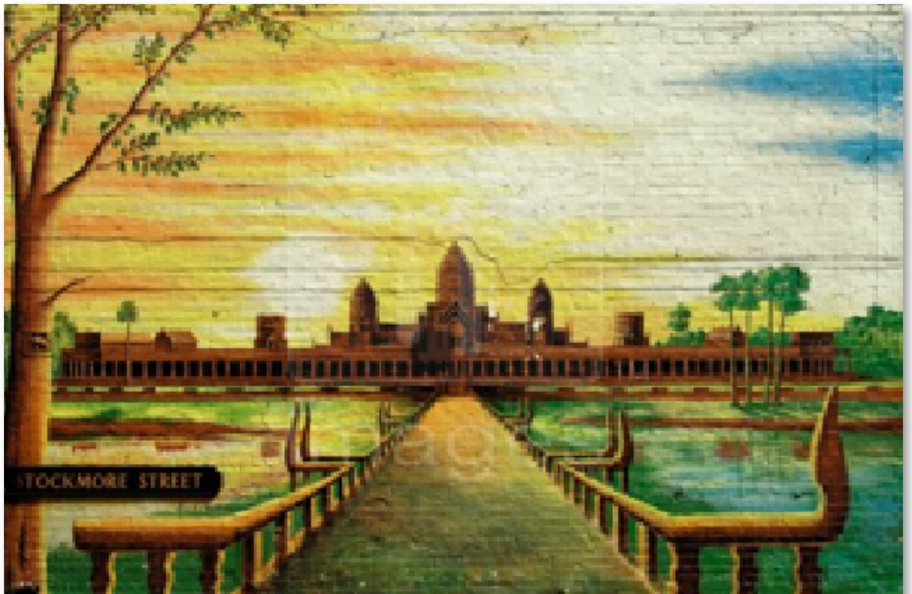

STOCKMORE STREET

Input Query



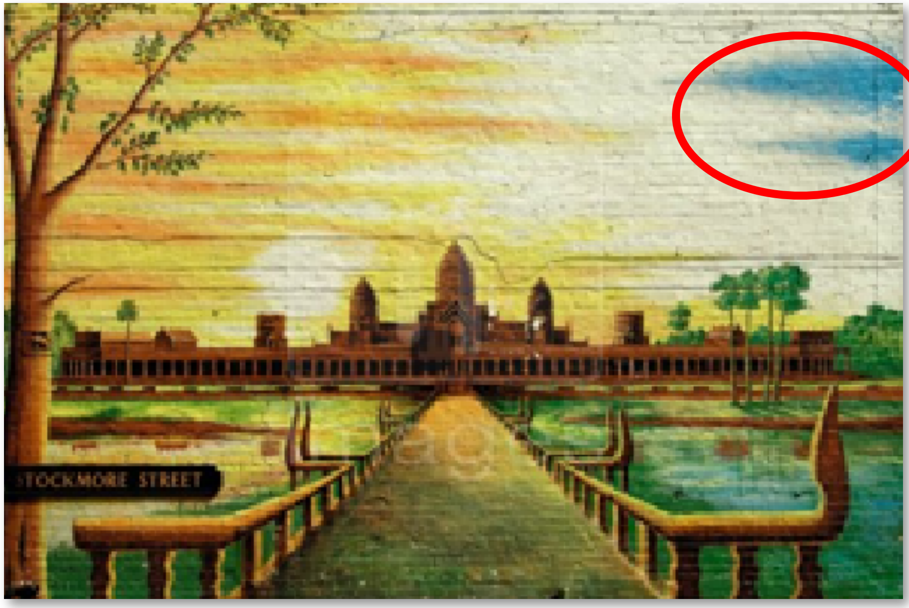
Top Matches

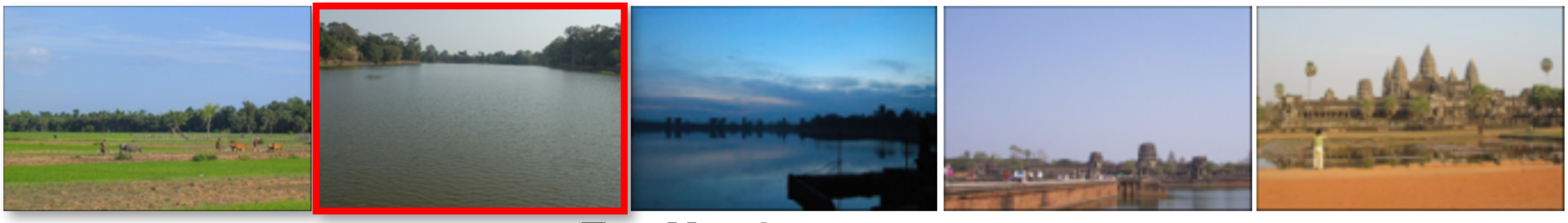
Input Query



Top Matches

Input Query

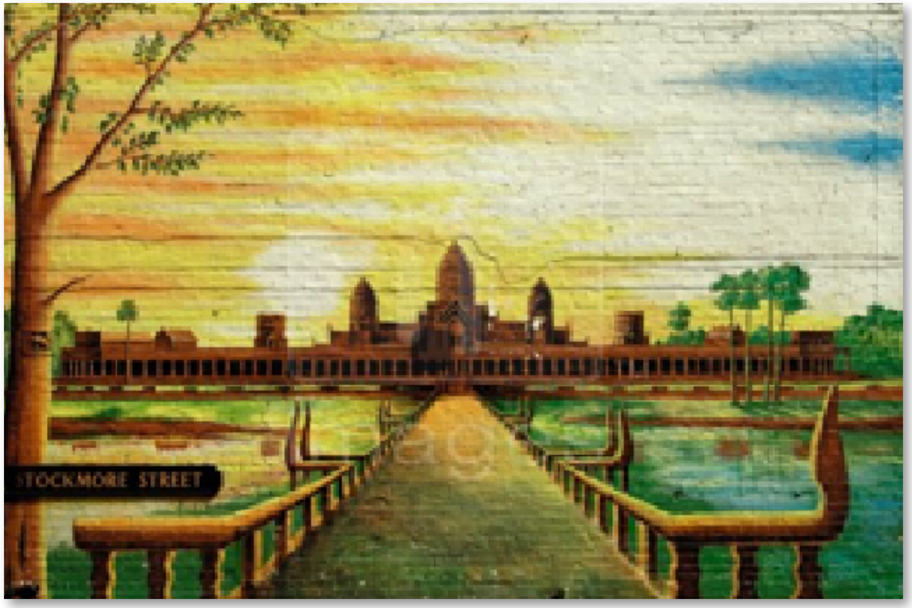




Top Matches

IMPORTANT PARTS?

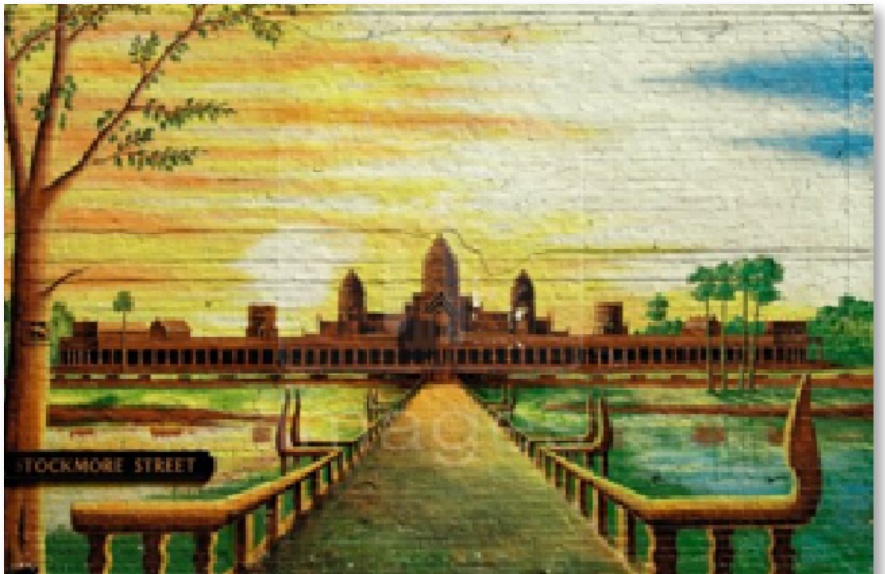
Input Query

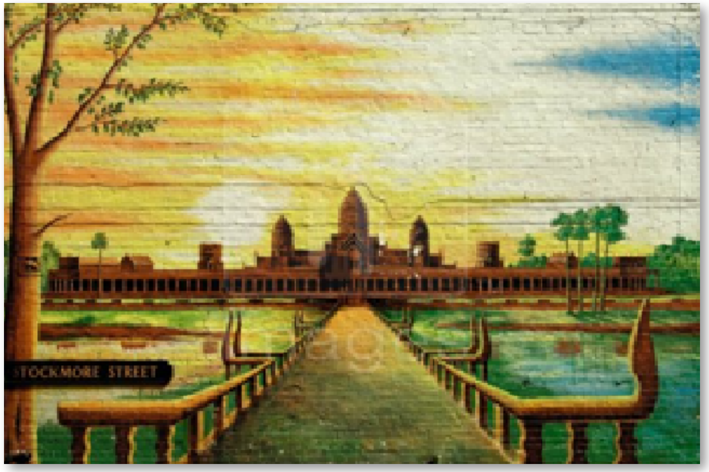


Important Parts

Top Matches

Input Query





“Data-driven Uniqueness”

Search using Images

Input Query

Top Matches

Search using Sketches

Search using Paintings

Input Painting

Top Matches

Search using Paintings

Input Painting

Top Matches

Dealing with sparse data (rare scenes)

- better similarity
- better alignment
  - e.g. reduce resolution, sifting, warping, etc.

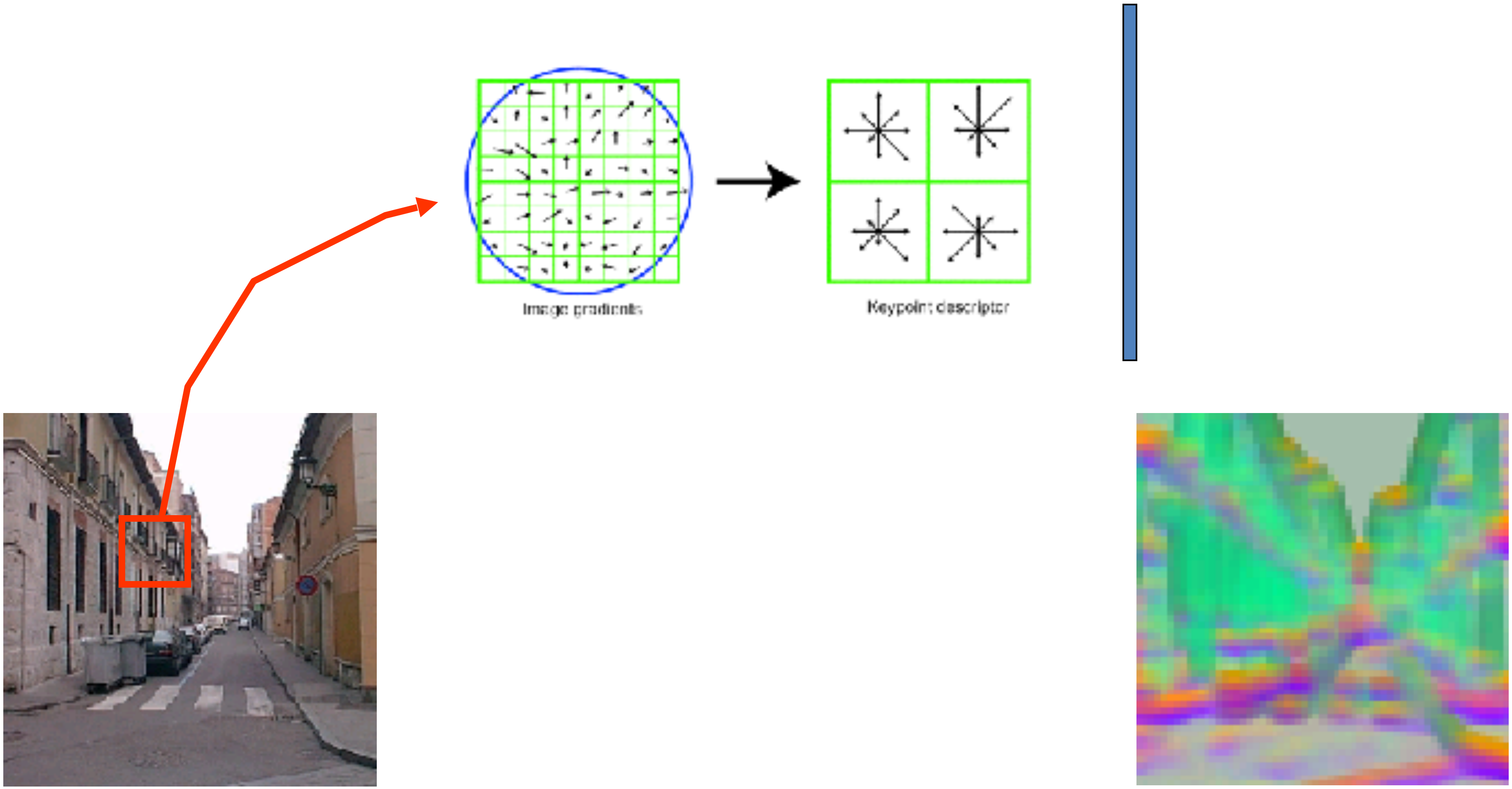
Matching scenes

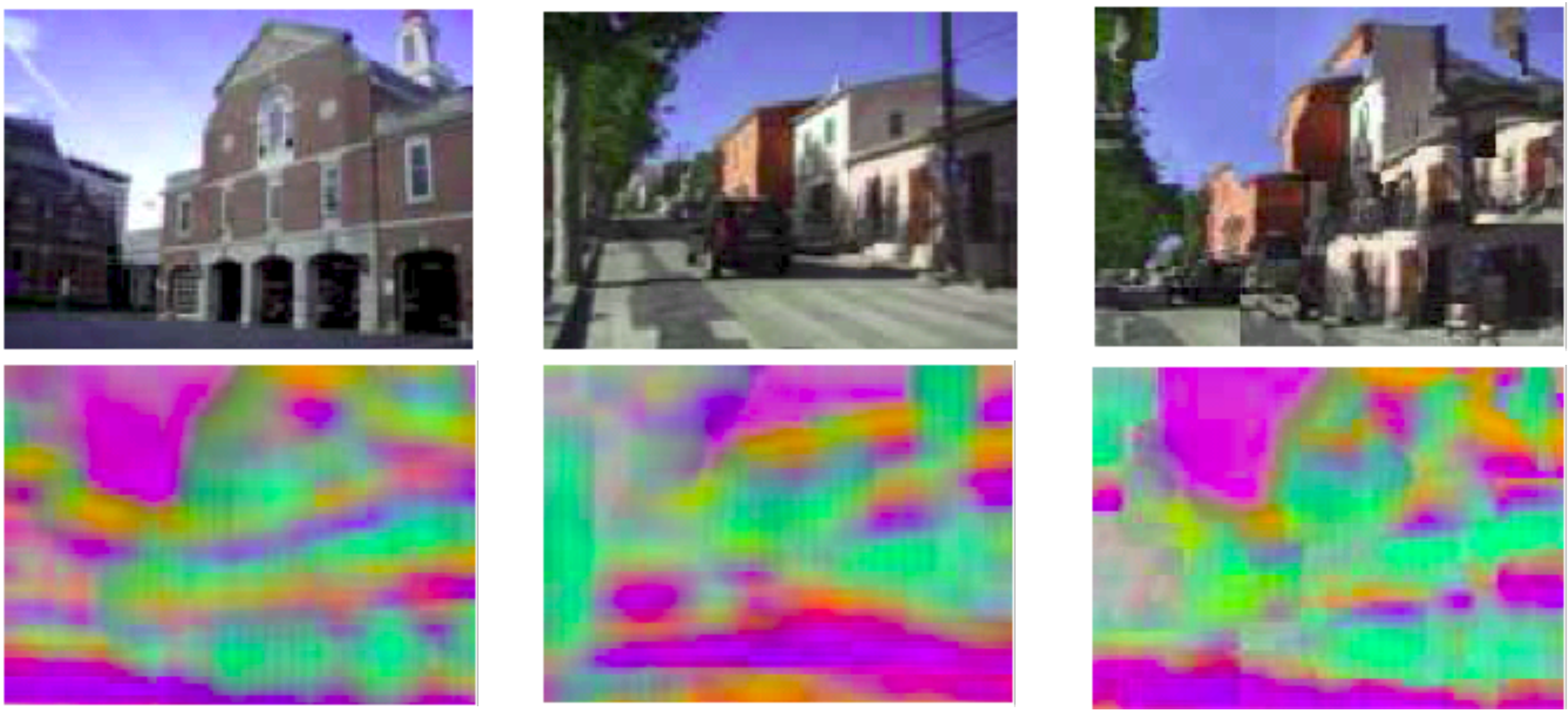
Two images taken from the same scene category, but different instances

- Contain different objects with different scales, perspectives and spatial location

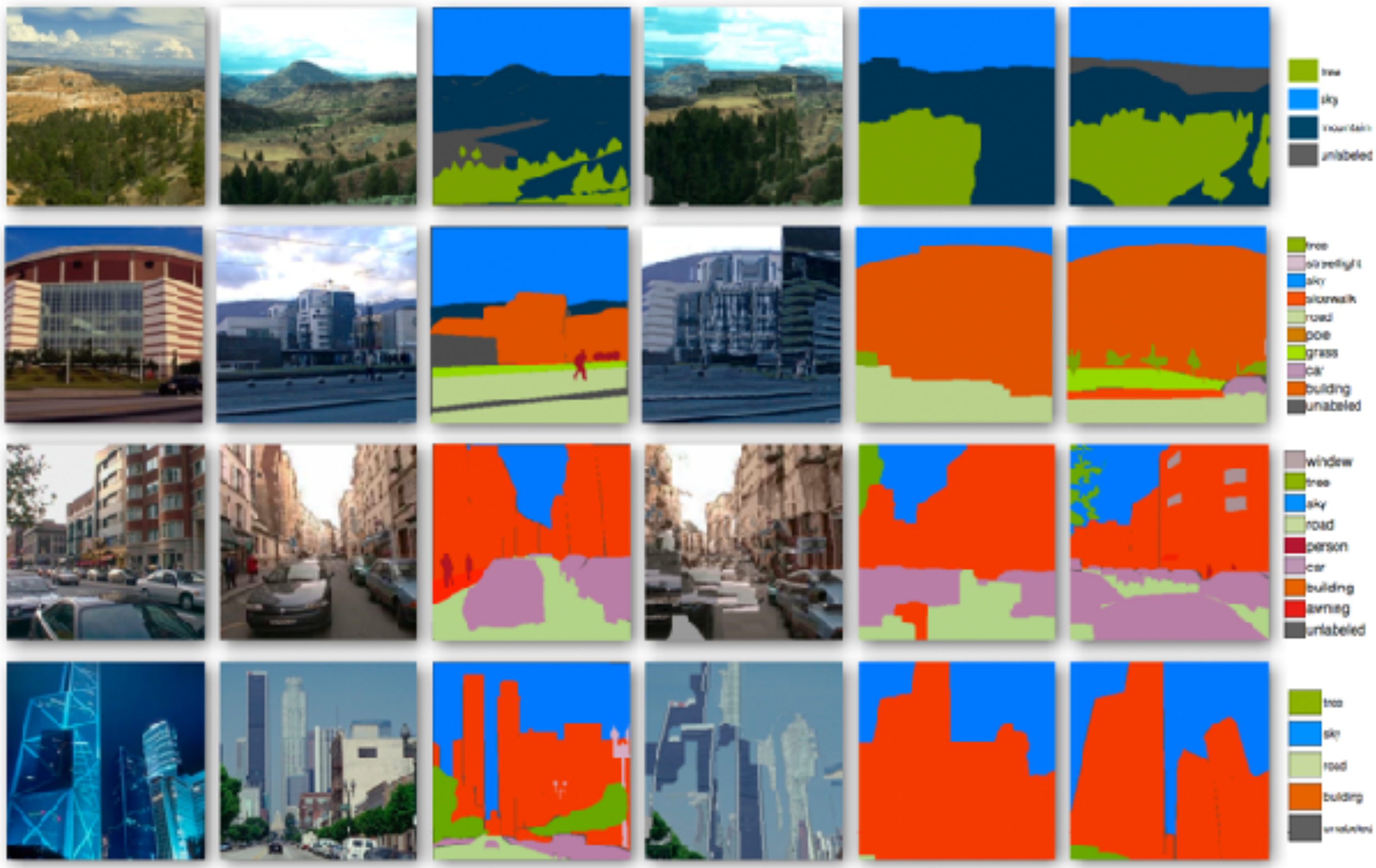
Liu, Yuen, Torralba, Sivic, Freeman. ECCV 08

Image representation





Scene parsing results



Query | Best match | Annotation of best match | Warped best match to query | Parsing result | Ground truth

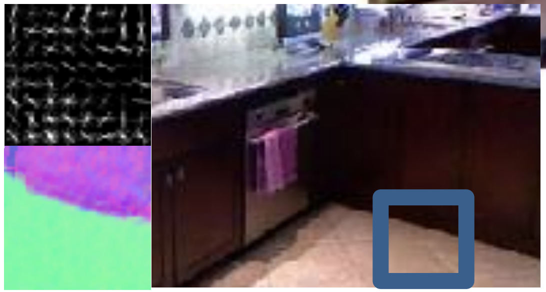
Dealing with sparse data (rare scenes)

- better similarity
- better alignment
  - e.g. reduce resolution, sifting, warping, etc.
- Use sub-images (primitives) to match
  - Allows matching from multiple images

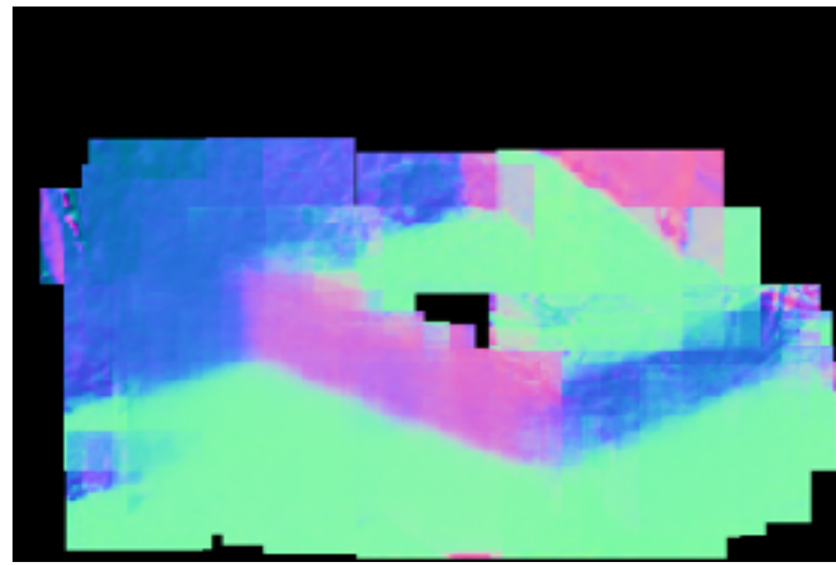
Predicting Surface Normals

Matching Parts

Matching Parts



Matching Parts

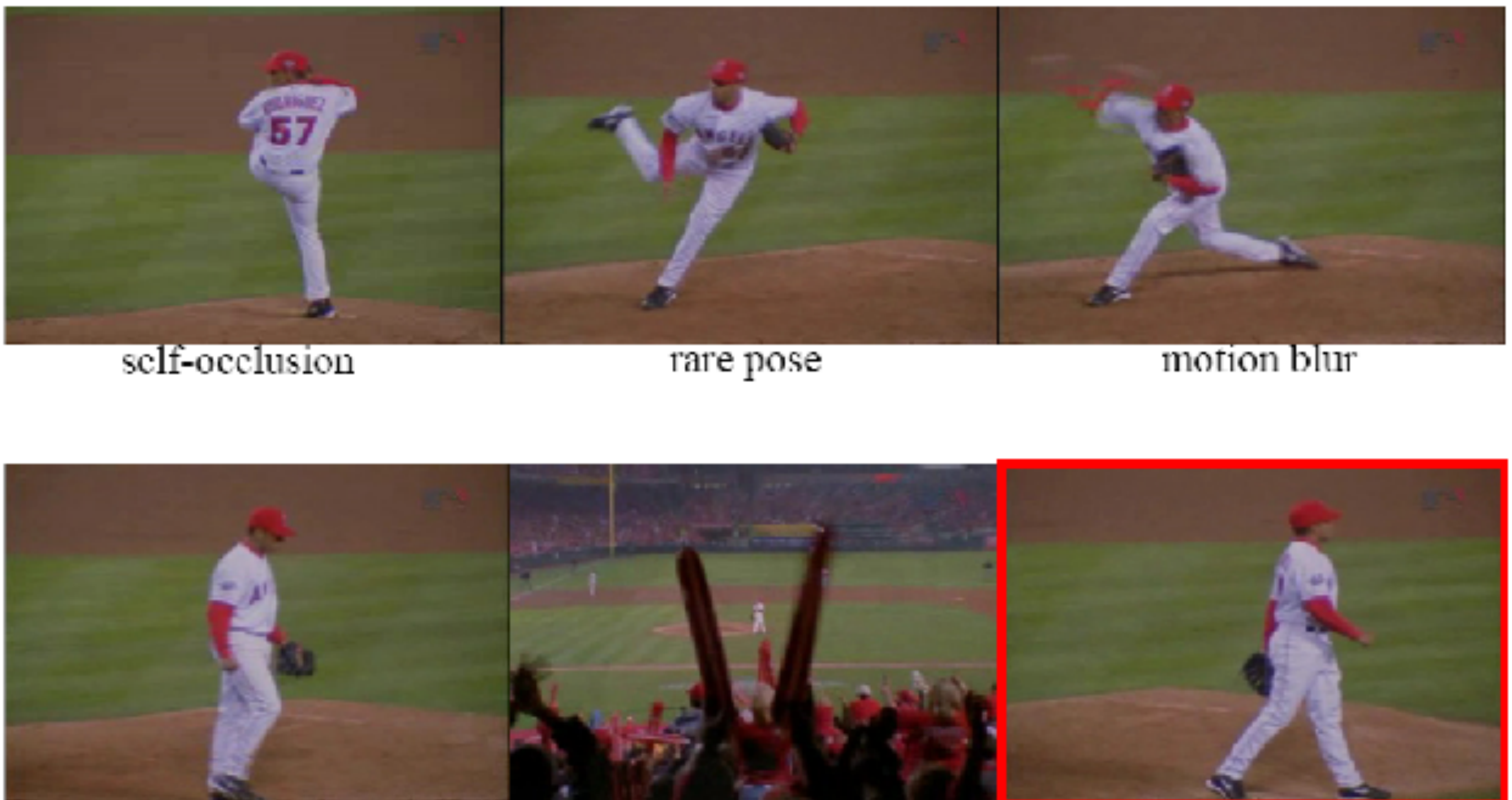


Dealing with sparse data (rare scenes)

- better similarity
- better alignment
- Use sub-images (primitives) to match
- Understand the simple stuff first
  - e.g. tracking via recognition, background subtraction, "object pop-out", etc.

Recognize when it's easy!

People take on a variety of poses, aspects, scales



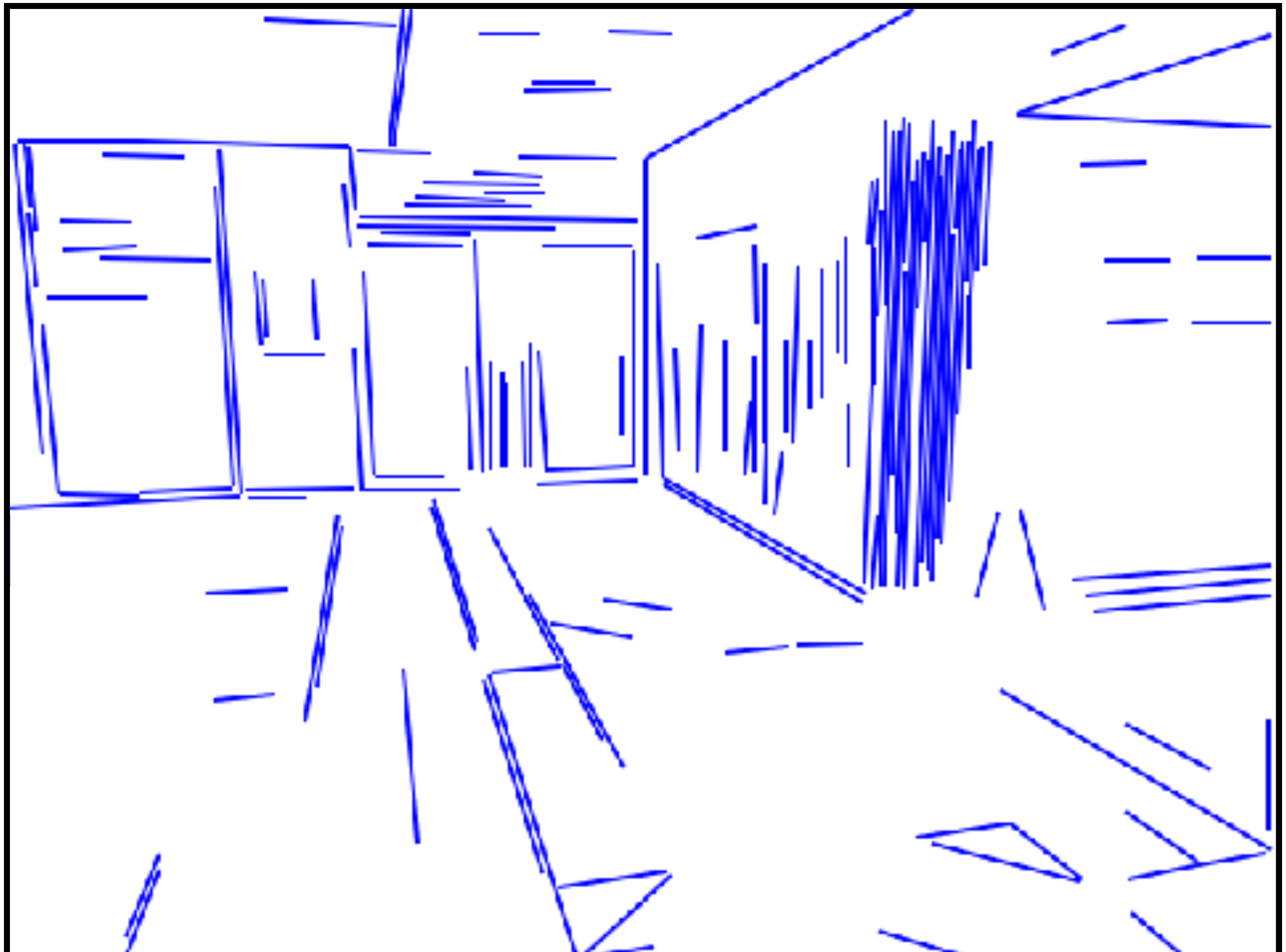
self-occlusion | rare pose | motion blur

non-distinctive pose | too small | just right detect this

27

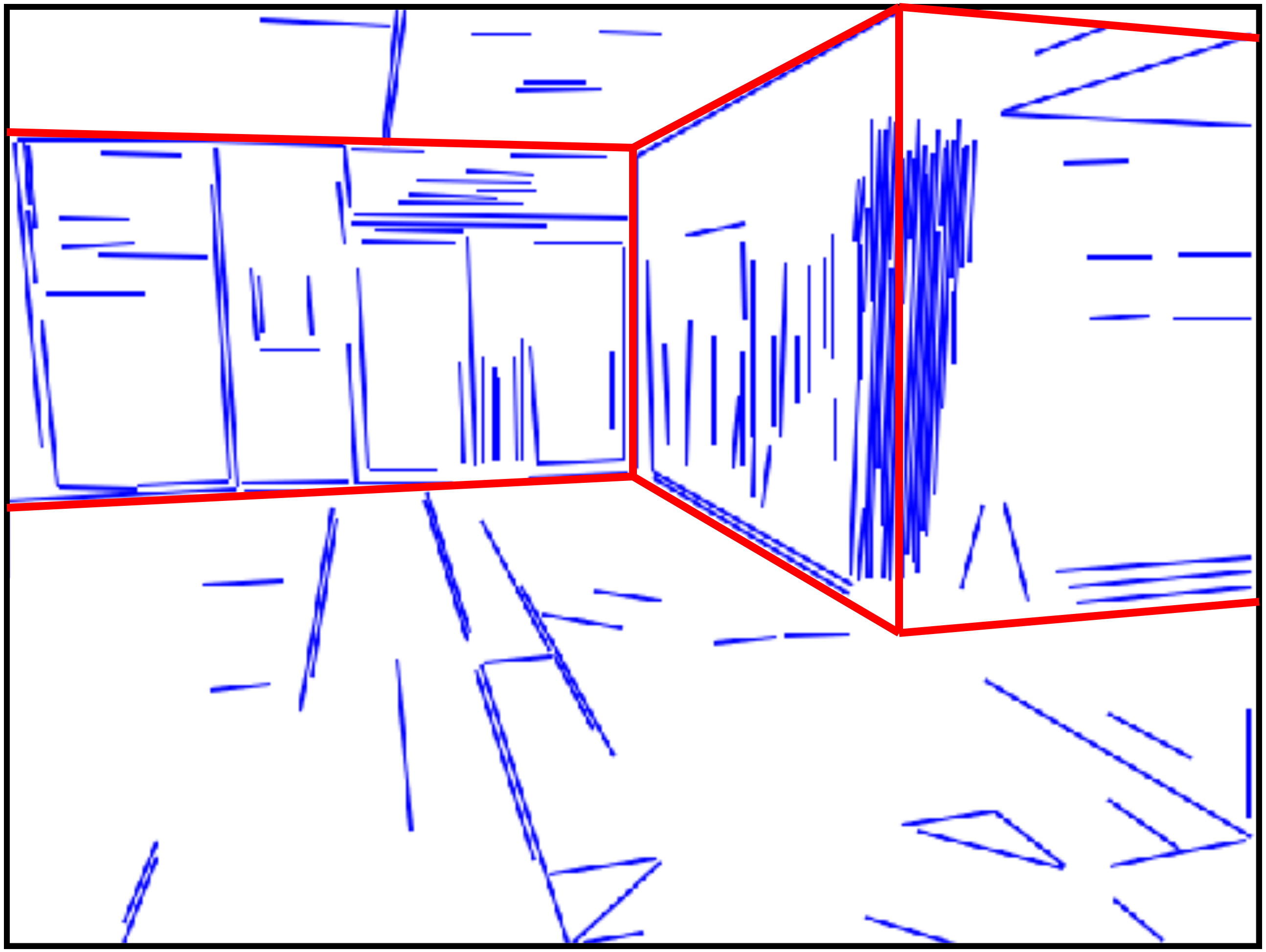
Ramanan, Forsyth, Zisserman, 2004

Guess structure



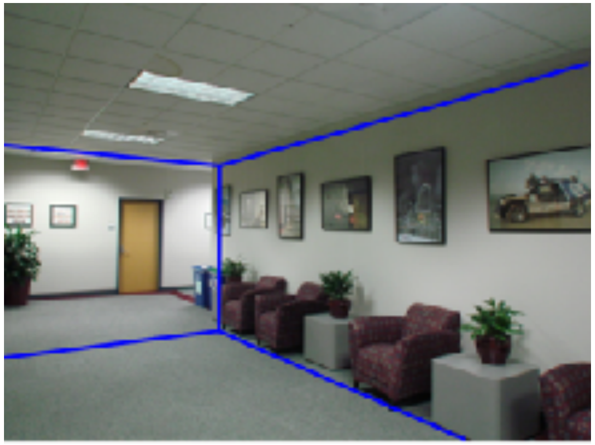
David C. Lee, Martial Hebert, Takeo Kanade, CVPR’09

Guess structure

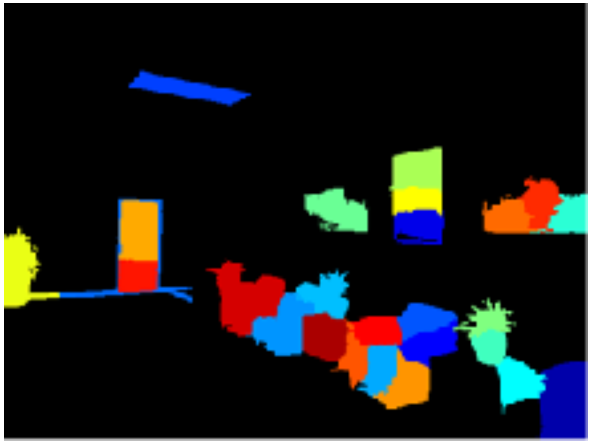


David C. Lee, Martial Hebert, Takeo Kanade, CVPR'09

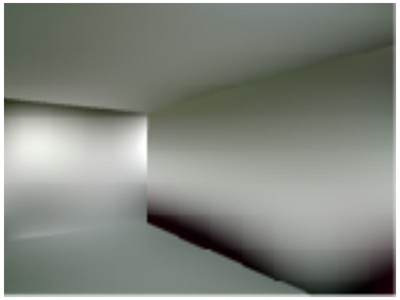
Subtracting away structure



Structure



Objects

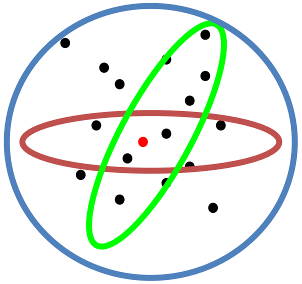


Wall appearance modeling

David C. Lee, Martial Hebert, Takeo Kanade, CVPR'09

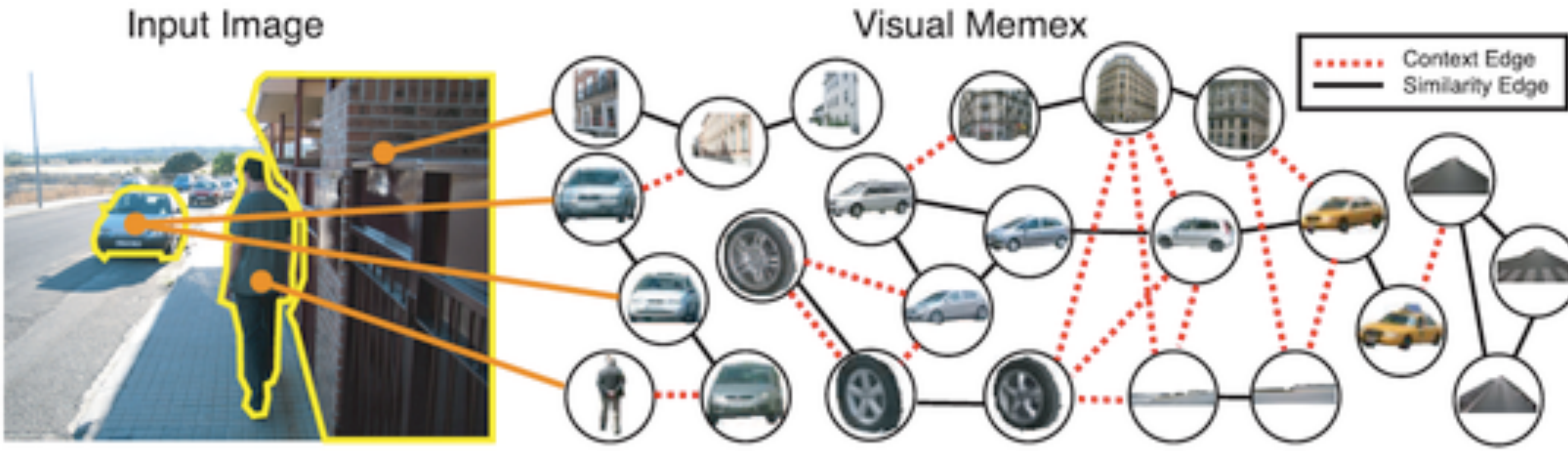
Dealing with sparse data (rare scenes)

- better similarity
- better alignment
  - e.g. reduce resolution, sifting, warping, etc.
- segment into chunks
  - e.g. segmentation for recognition approaches
- get rid of simple stuff first
  - e.g. background subtraction, “object pop-out”, etc.



- Moving away from kNN methodology…
- use data to make connections
  - e..g The Memex, manifold learning, data association, subpopulation means, etc.

Memex – Knowledge Graph



Manifolds

- Images are high dimensional: A 64x64 image is 4096 dimensional vector.
- But the possible images are much less!
- Is there a subspace where the set of images lie?

manifolds in vision

appearance variation

Slide by Dave Thompson

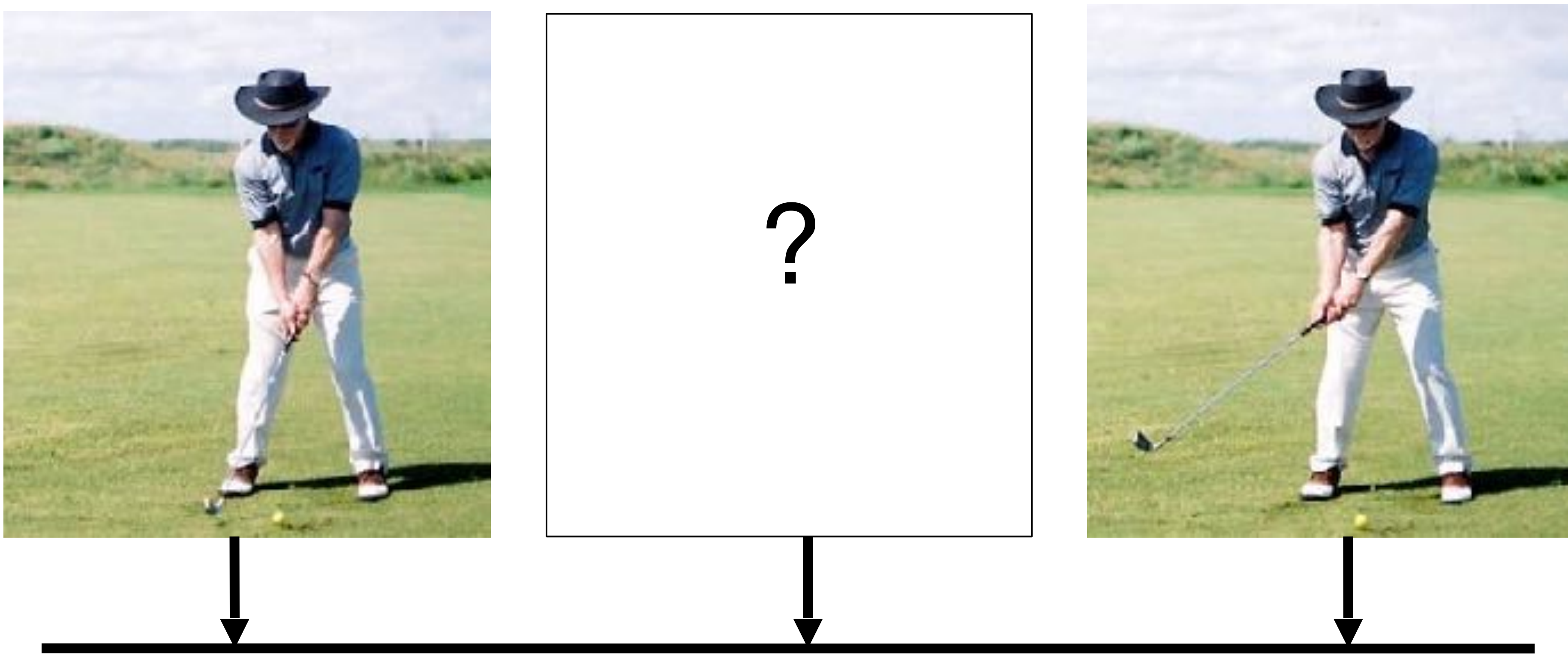
images from hormel corp.

manifolds in vision

Slide by Dave Thompson

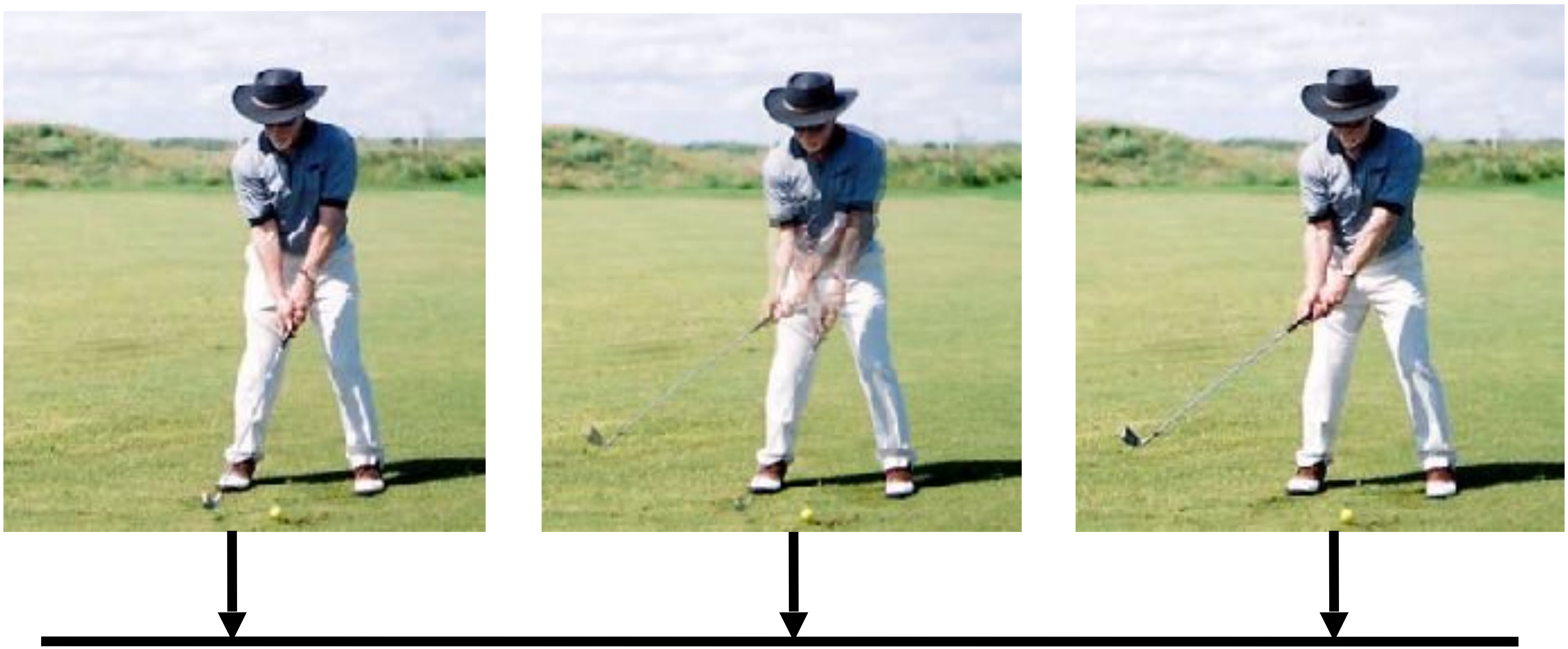
images from www.golfswingphotos.com

reasonable distance metrics



Slide by Dave Thompson

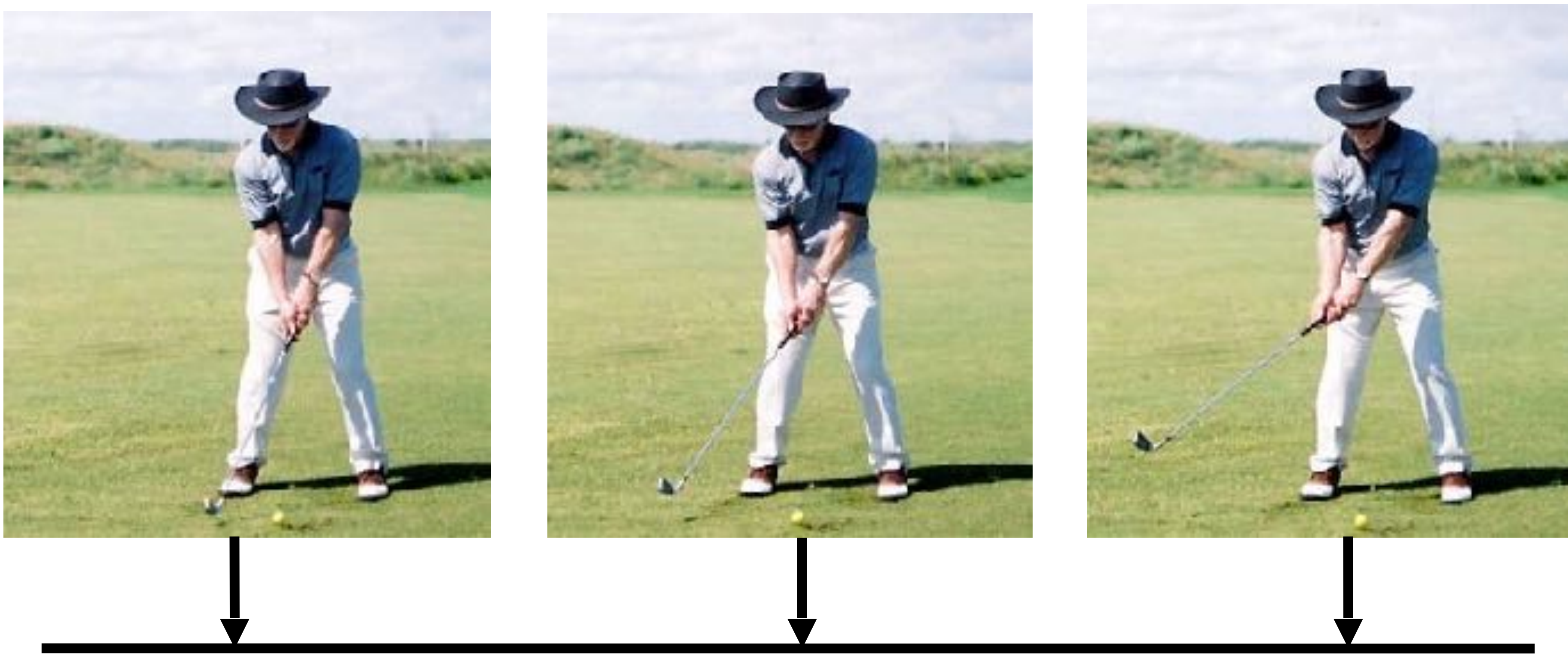
reasonable distance metrics



linear interpolation

Slide by Dave Thompson

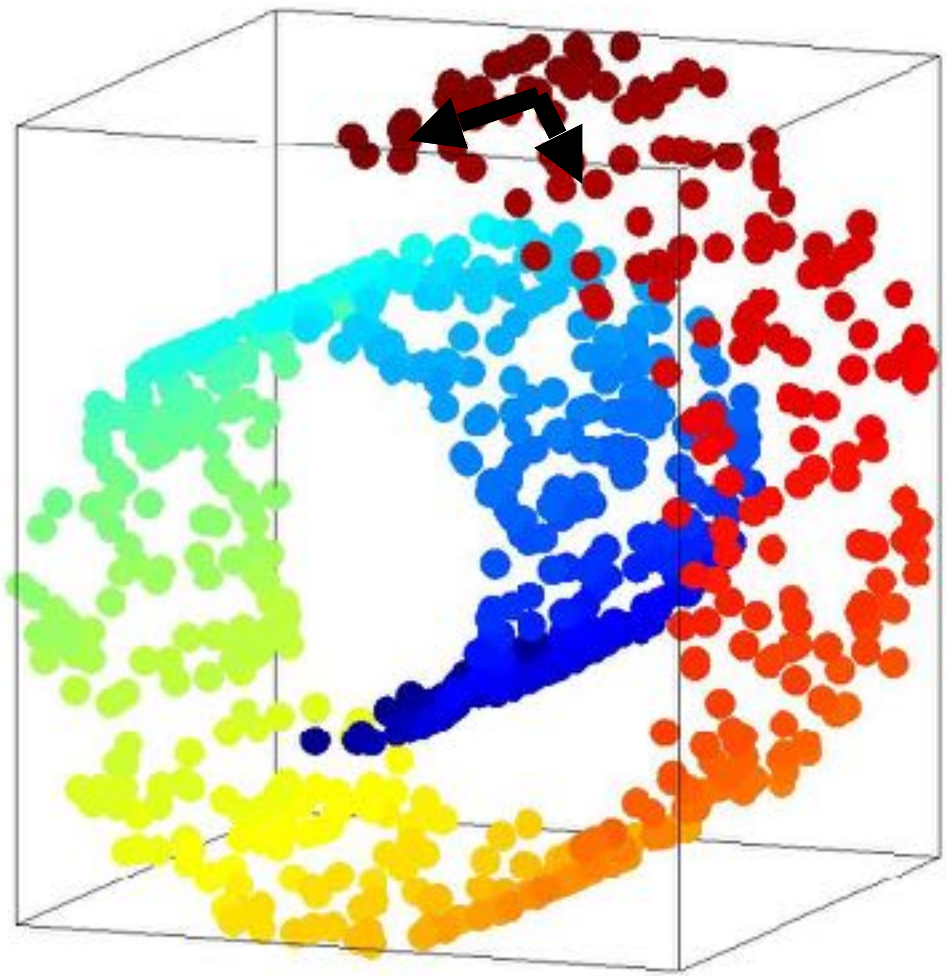
reasonable distance metrics



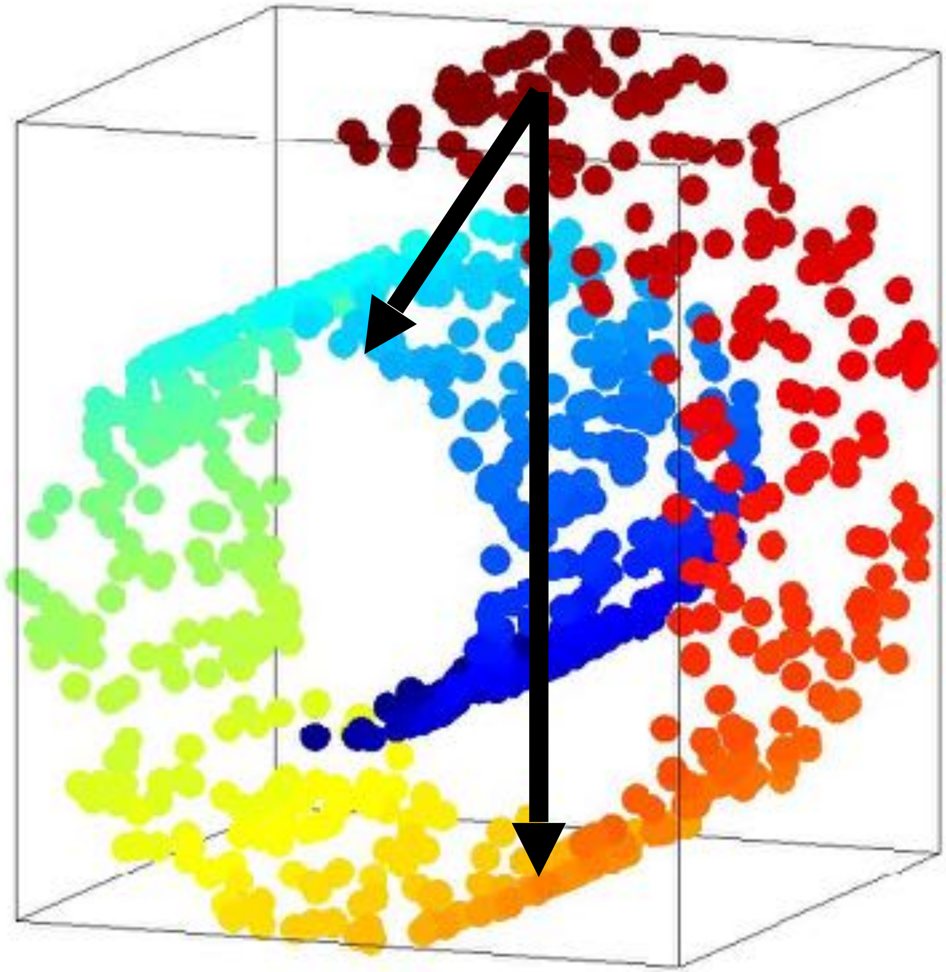
manifold interpolation

Slide by Dave Thompson

reasonable distance metrics

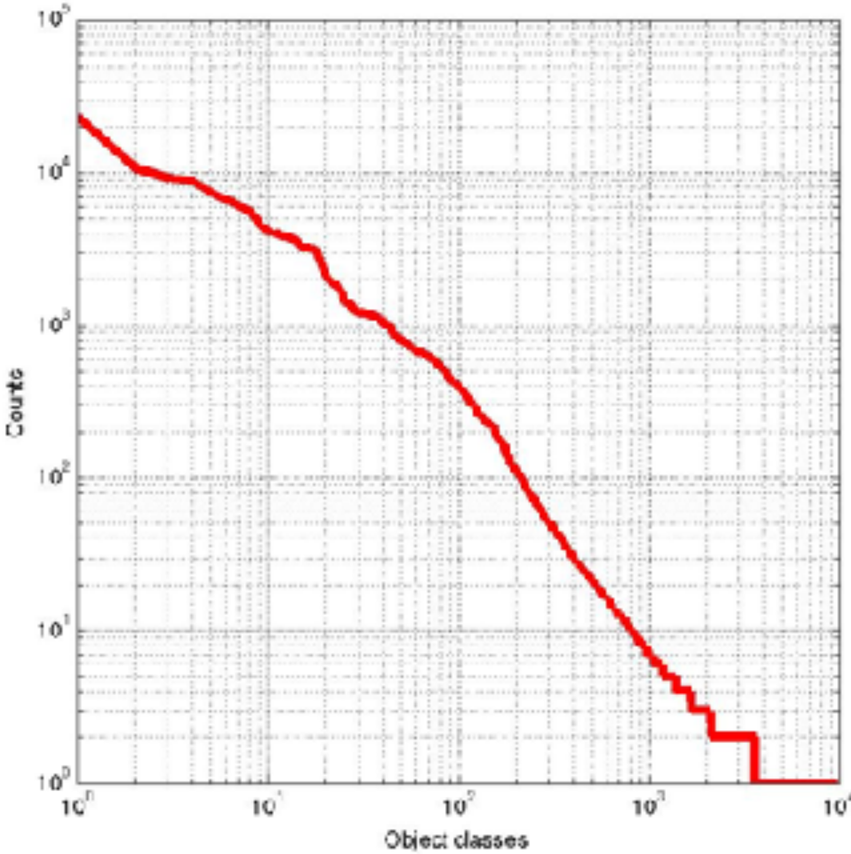


reasonable distance metrics

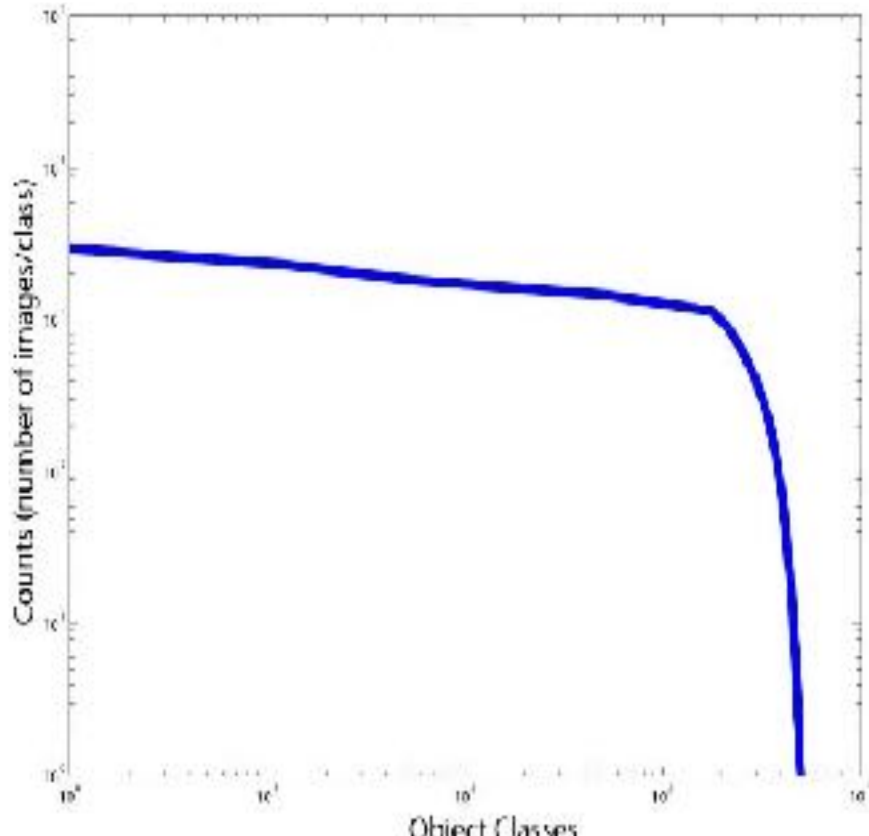


Some observations about data collection

Object distributions

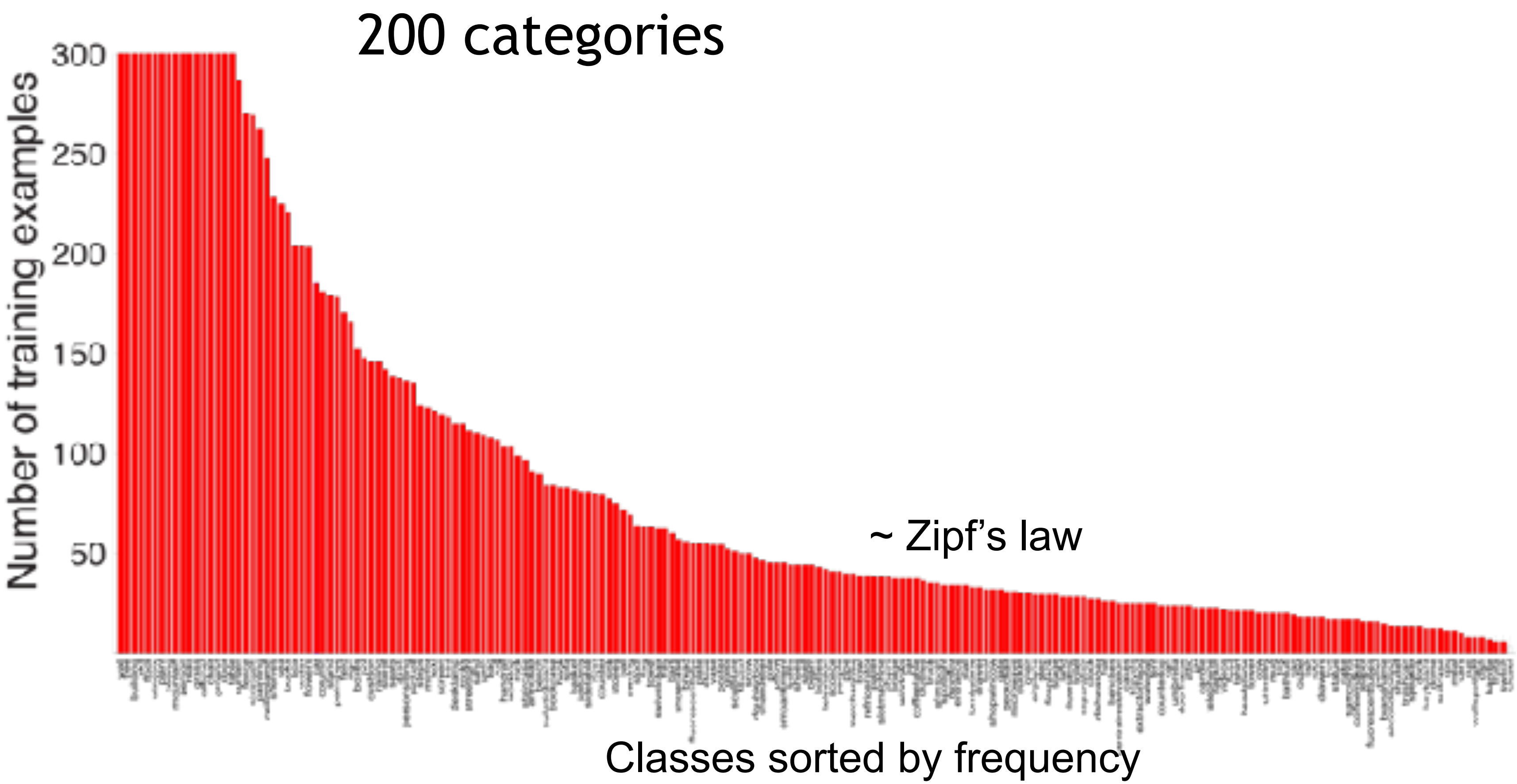


LabelMe



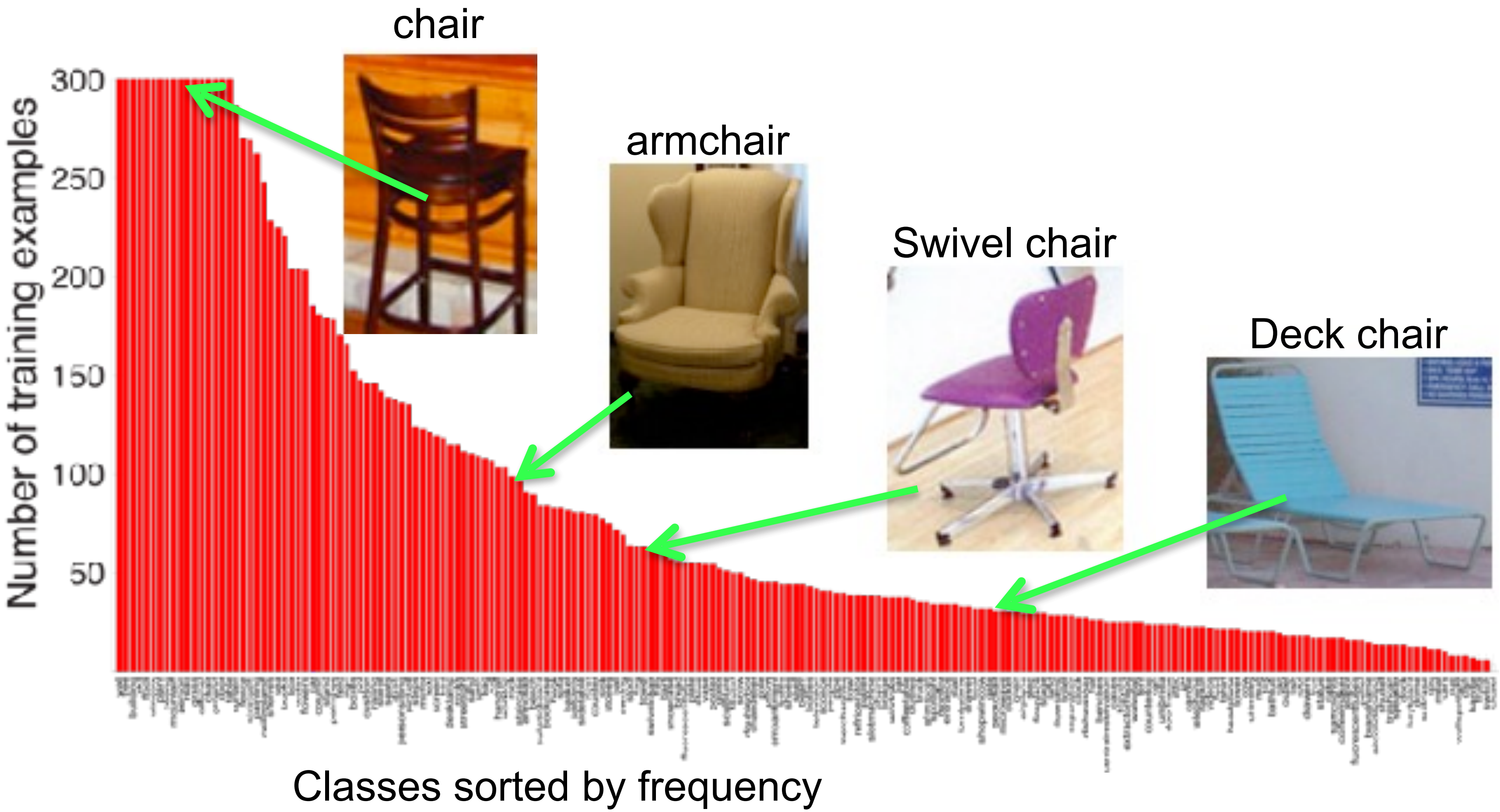
IMAGENET

SUN database



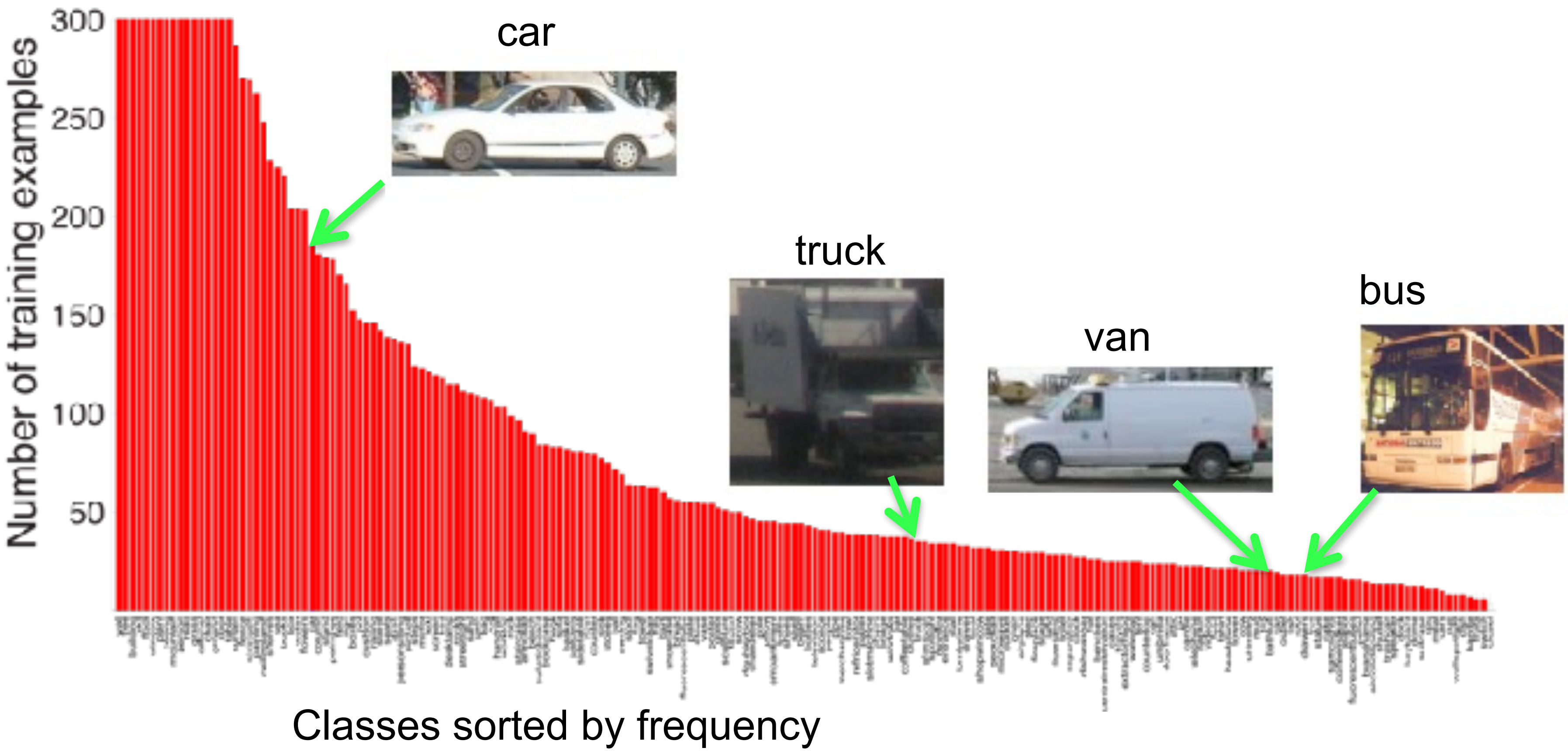
The first 9 objects account for 50% of all training examples
17 classes with more than 300 examples
109 classes with less than 50 examples

Rare objects are similar to frequent objects



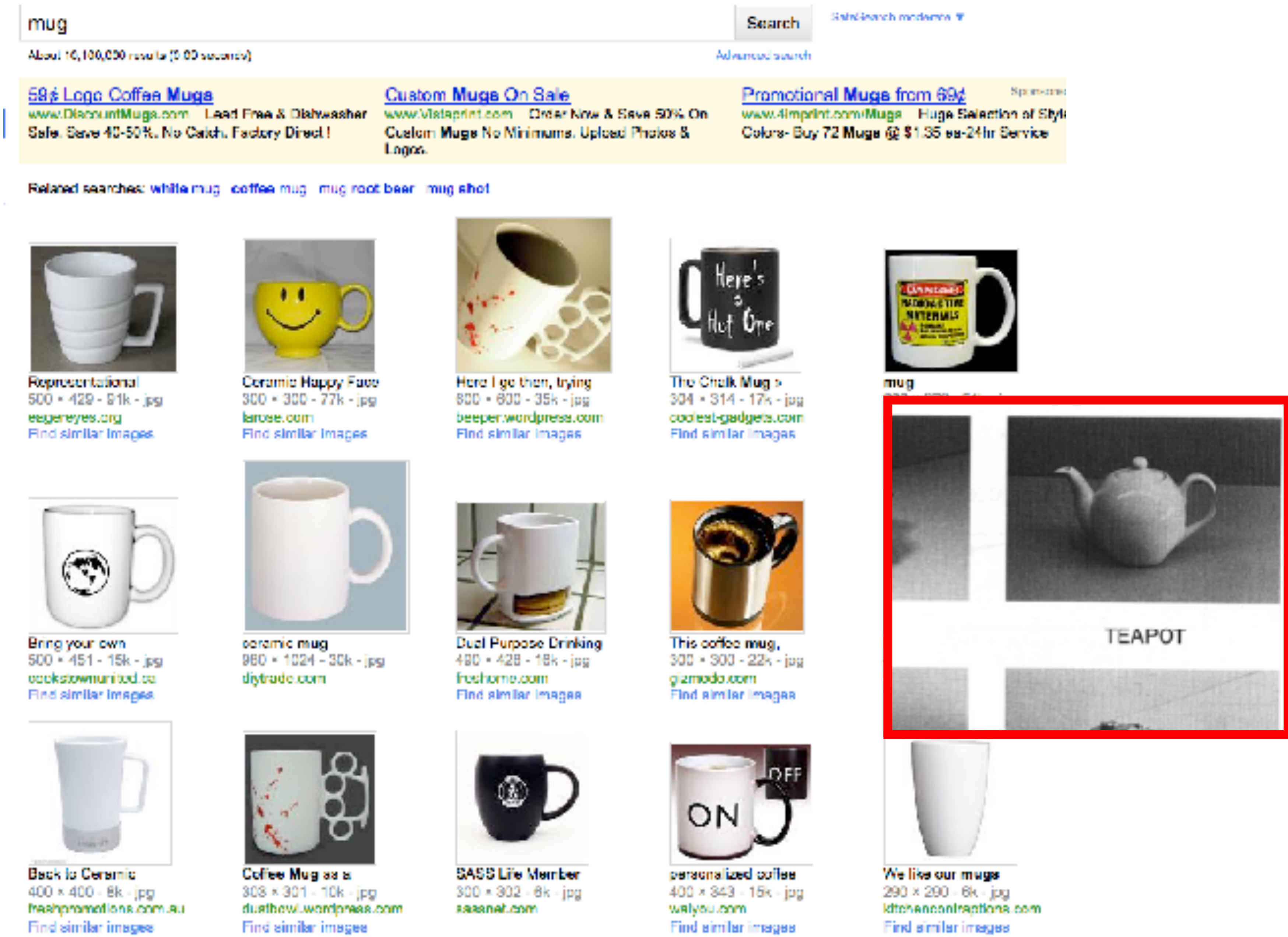
Salakhutdinov, Torralba, and Tenenbaum, CVPR, 2011

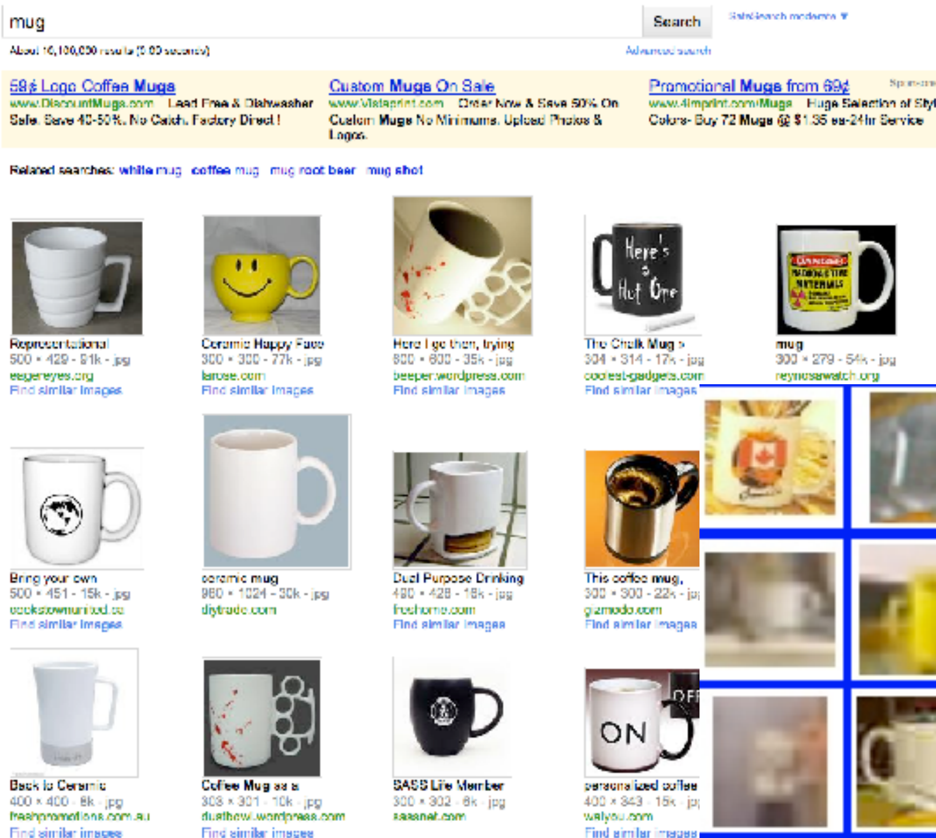
Rare objects are similar to frequent objects



Salakhutdinov, Torralba, and Tenenbaum, CVPR, 2011

Some bias comes from the way the data is collected

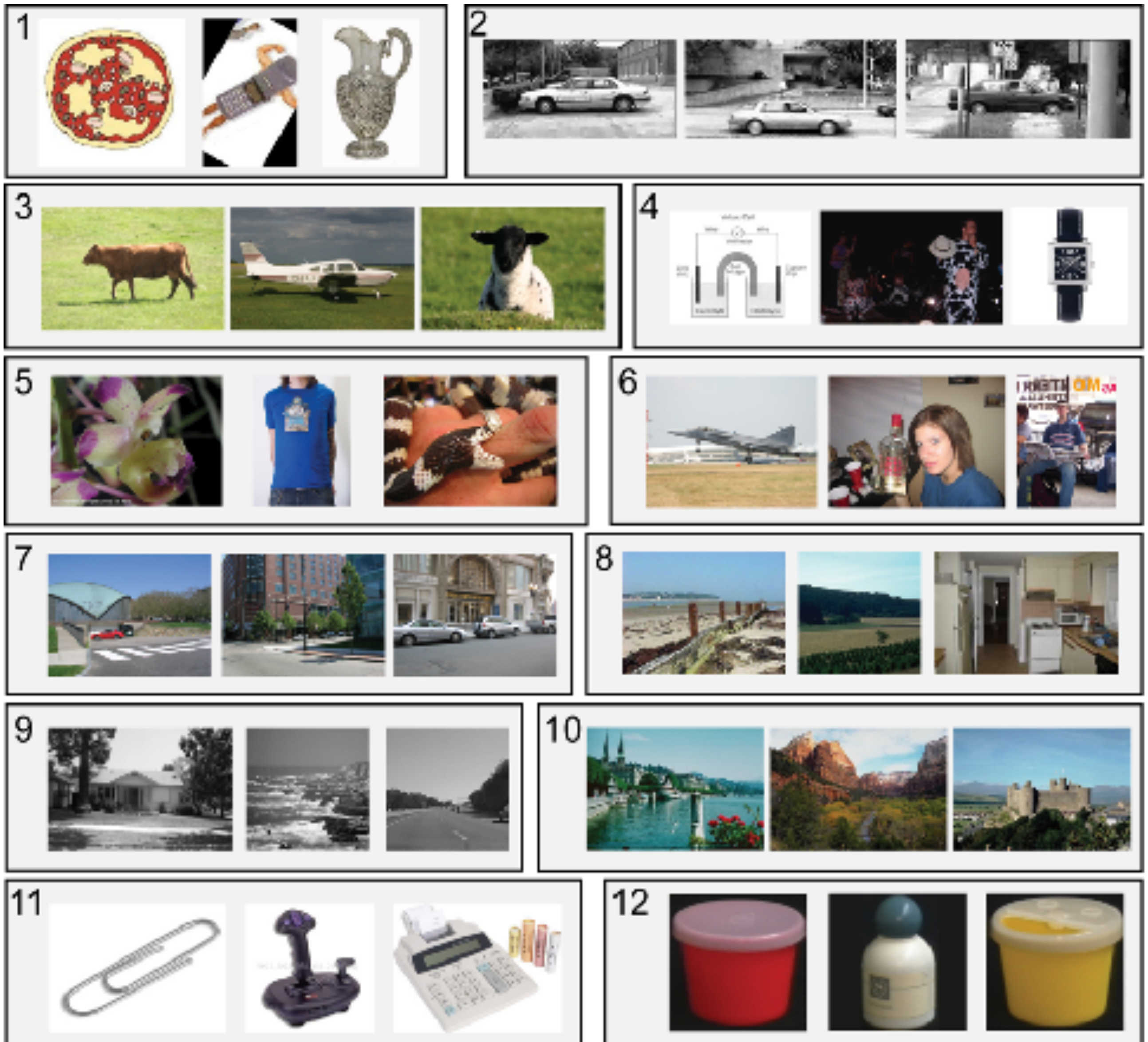




Google mugs

Mugs from LabelMe

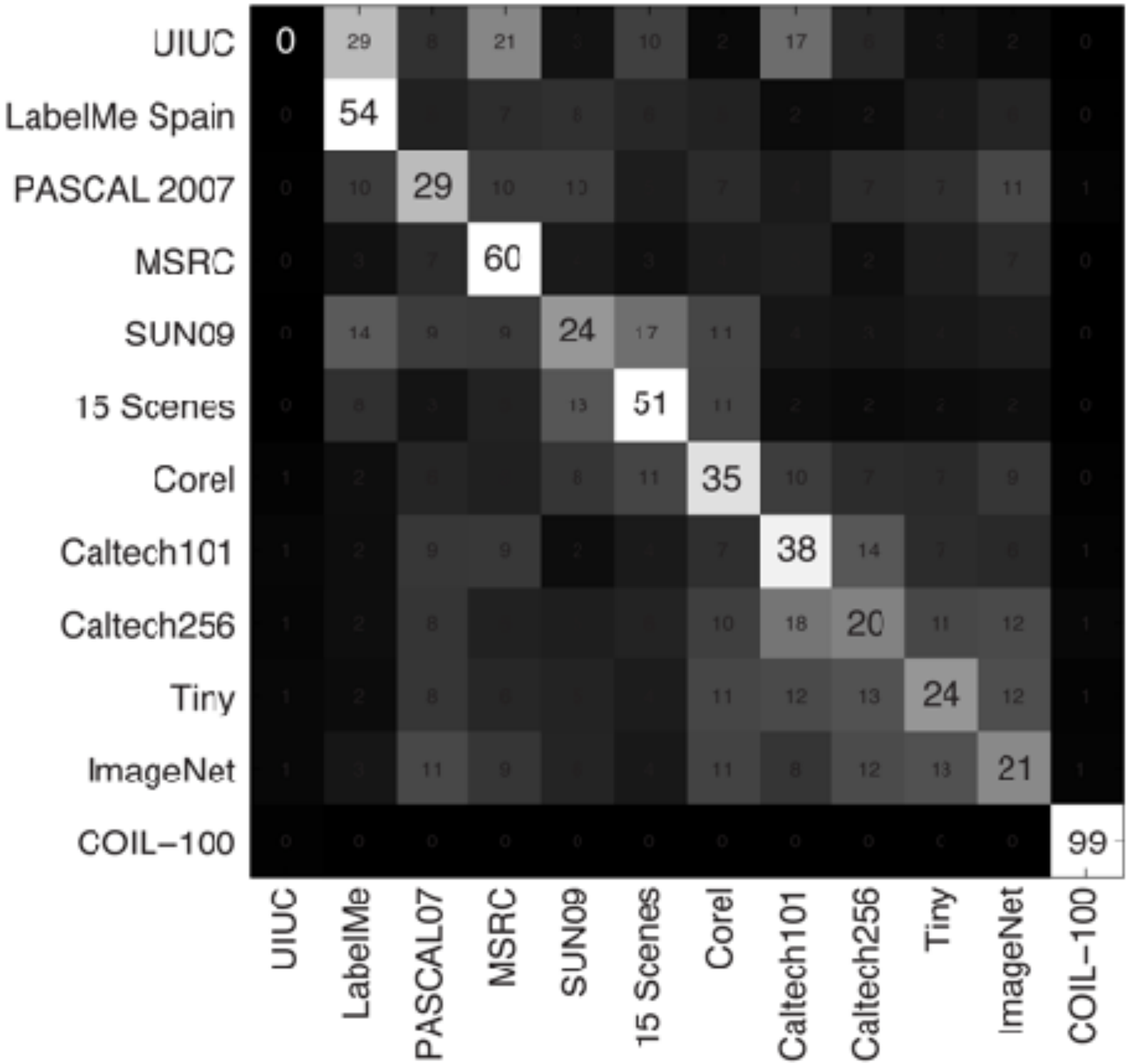
"*Name That Dataset!*" game



___ Caltech 101
___ Caltech 256
___ MSRC
___ UIUC cars
___ Tiny Images
___ Corel
___ PASCAL 2007
___ LabelMe
___ COIL-100
___ ImageNet
___ 15 Scenes
___ SUN'09

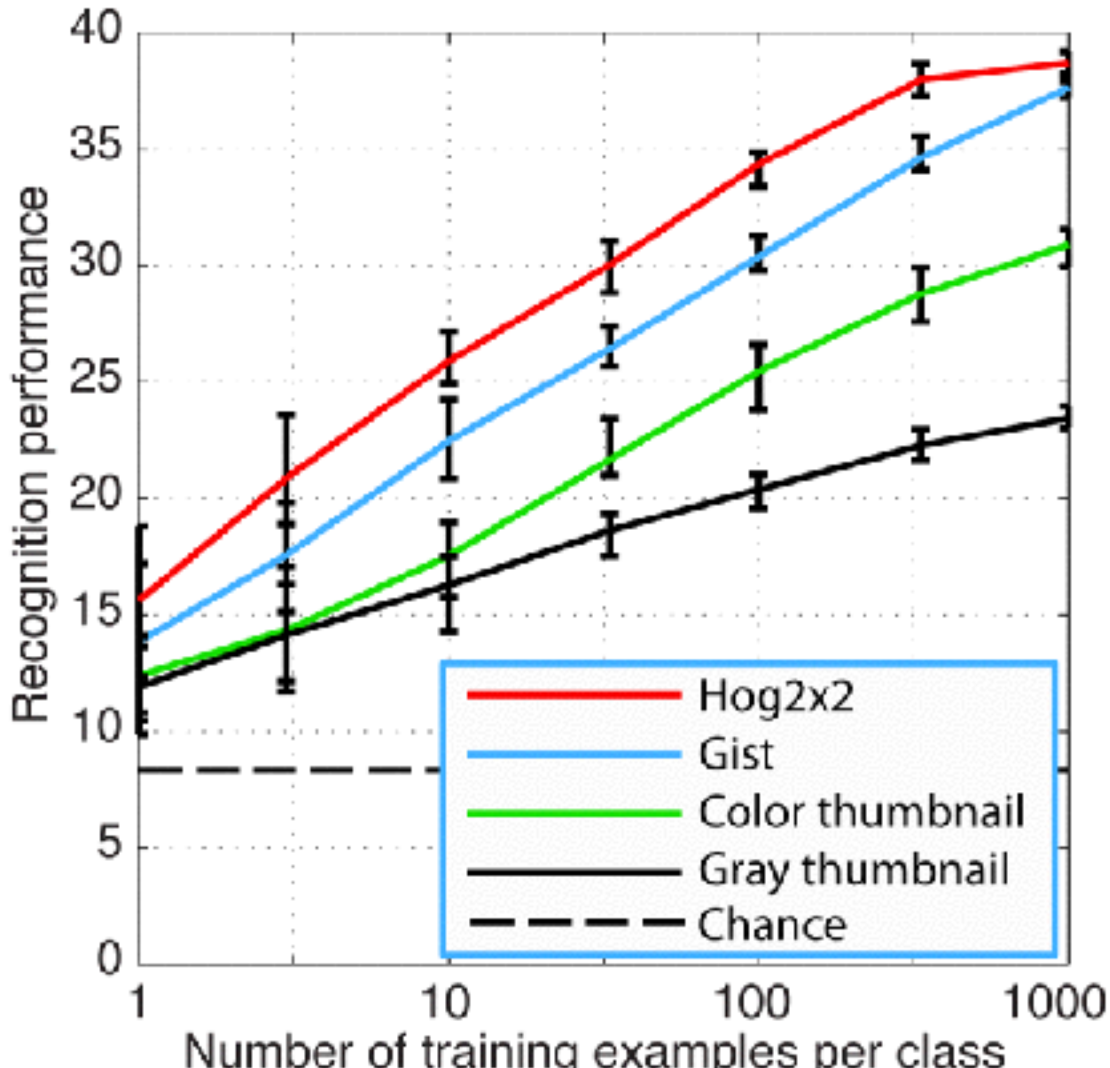
SVM plays “*Name that dataset!*”

SVM plays “*Name that dataset!*”



- 12 1-vs-all classifiers
- Standard full-image features
- 39% performance (chance is 8%)

SVM plays "*Name that dataset!*"



Datasets have different goals…

- Some are object-centric (e.g. Caltech, ImageNet)
- Otherwise are scene-centric (e.g. LabelMe, SUN’09)

- What about playing “*name that dataset*” on bounding boxes?

Cross-Dataset Generalization

MSRC

Classifier trained on MSRC cars

Mixing datasets