Object Recognition as Machine Translation: Learning a Lexicon for a Fixed image Vocabulary

Pinar Duygulu
Middle East Technical University, Turkey

Joint work with Kobus Barnard,
Nando de Freitas and David Forsyth
as a part of
UC Berkeley Digital Library Project

Problems in Object Recognition

- What is an object ?
- How to model?

- Scale

Our Approach

Object recognition on a large scale is linking words with image regions

Use joint probability of words and pictures in large datasets

tiger grass cat

Auto-Annotating Images

Finding words for the images

tiger grass cat

Barnard, Forsyth (ICCV 2001) , Barnard, Duygulu, Forsyth (CVPR 2001)

Other related work : Maron 98, Mori 99

Annotation vs Recognition

?

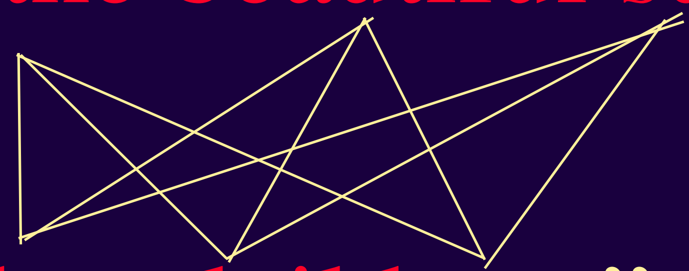
tiger cat grass

Cannot be solved with one example

Statistical Machine Translation

Data: Aligned sentences, but word correspondences are unknown

"the beautiful sun"



"le soleil beau"

Brown, Della Pietra, Della Pietra & Mercer 93

Statistical Machine Translation

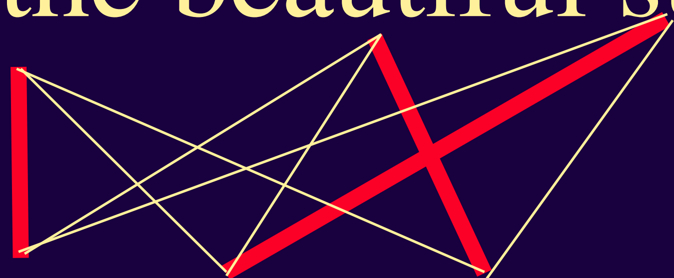
Given the correspondences, we can estimate the translation p(sun|soleil)

Given the probabilities, we can estimate the correspondences

Statistical Machine Translation

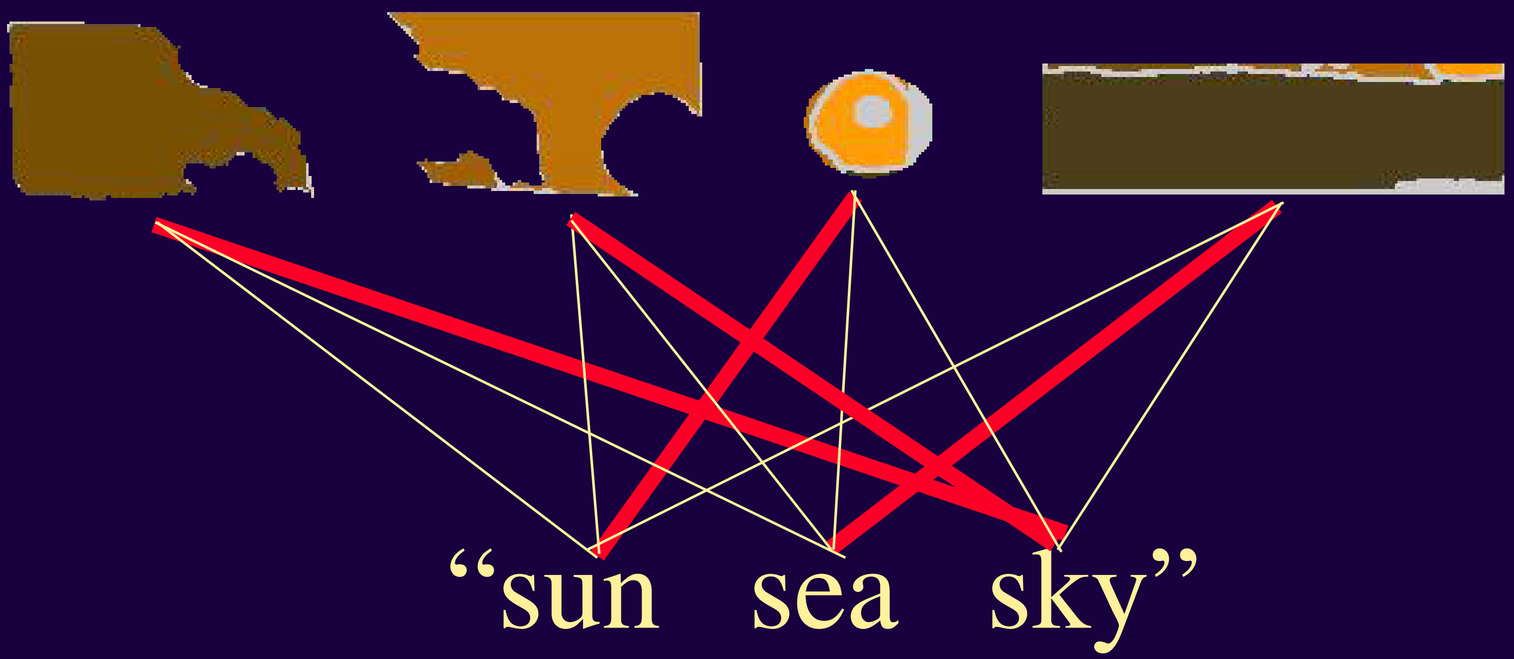
Enough data + EM, we can obtain the translation p(sun|soleil)=1

"the beautiful sun"



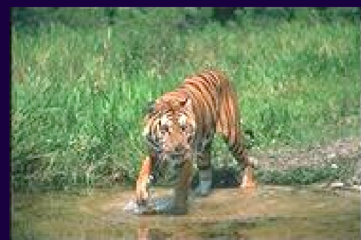
"le soleil beau"

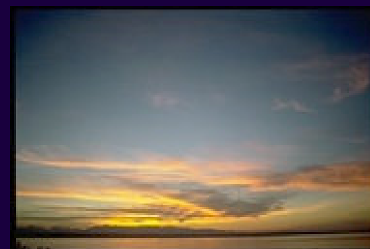
Multimedia Translation

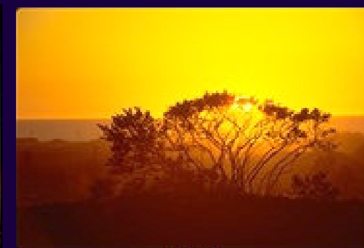


Corel Database


118011
WATER HARBOR
SKY CLOUDS


TIGER CAT WATER GRASS


1090
SUN CLOUDS
WATER SKY


1015
SUN TREE
PLAIN SKY


143078
MOUNTAINS TREES
aspens VALLEY


102042
MUSEUM memorial
FLAGS GRASS


119094
GARDEN BUILDING
FLOWERS TREES


131007
GARDEN FLOWERS
HOUSE TREES

392 CD’s, each consisting of 100 annotated images.

Input

segmentation*

sun sky waves sea

Each blob is a large vector of features

- Region size
- Position
- Color
- Oriented energy (12 filters)
- Simple shape features

* Thanks to Blobworld team [Carson, Belongie, Greenspan, Malik], N-cuts team [Shi, Tal, Malik]

Tokenization

- Words → word tokens

- Image segments

  •represented by 30 features
  (size, position, color, texture and shape)

  •k-means to cluster features

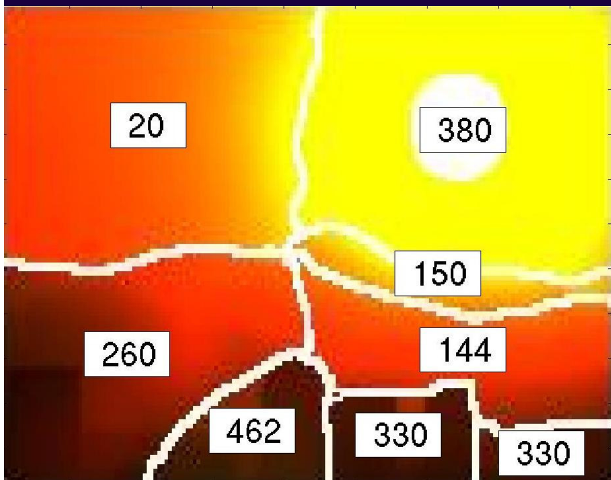
  •best cluster for the blob → blob tokens

Data

160 CD’s from Corel Data Set
100 images in each

10 sets
each :
  randomly selected 80 CD’s
  ~6000 training
  ~2000 test
  150-200 word tokens
  500 blob tokens

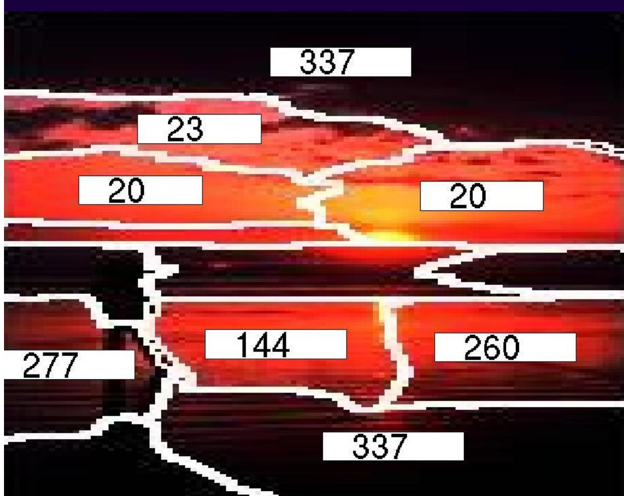
Segmentation (using Ncuts)
  about a month



city mountain sky sun

jet plane sky

cat forest grass tiger

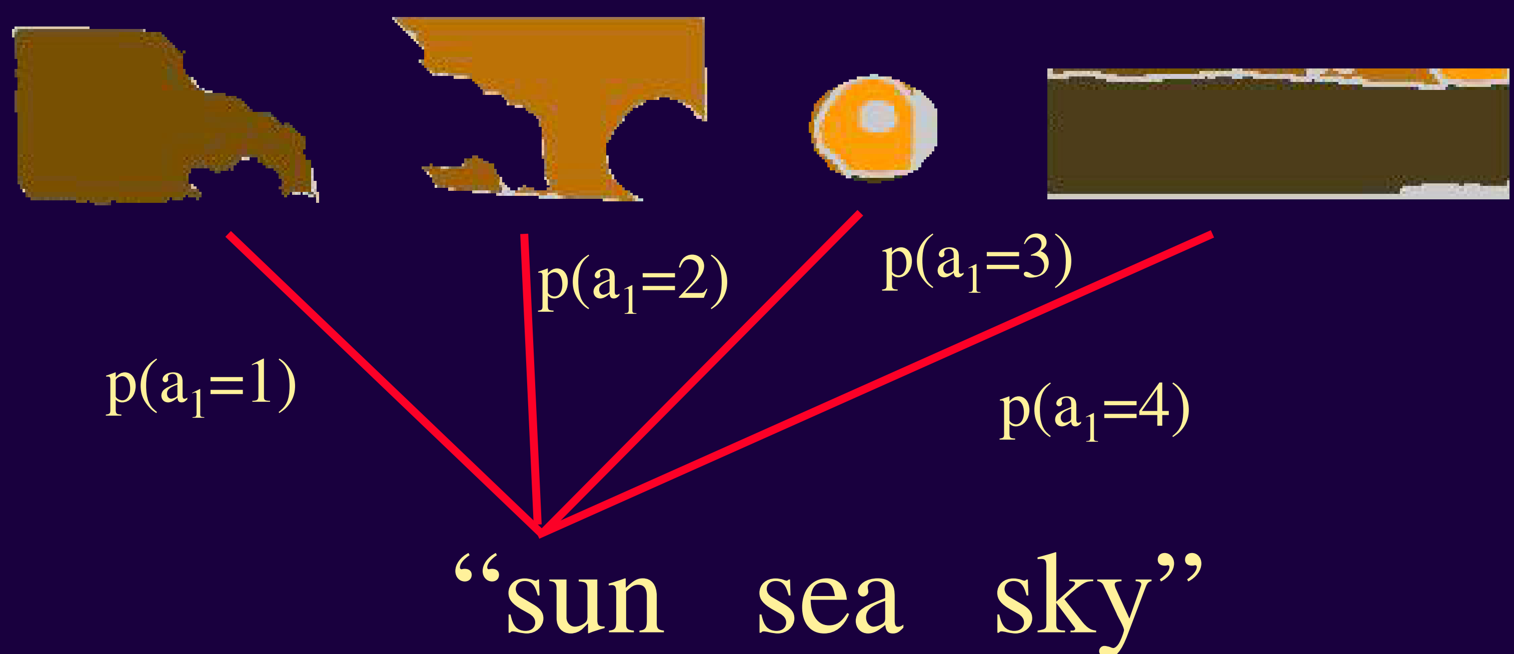


beach people sun water

jet plane sky

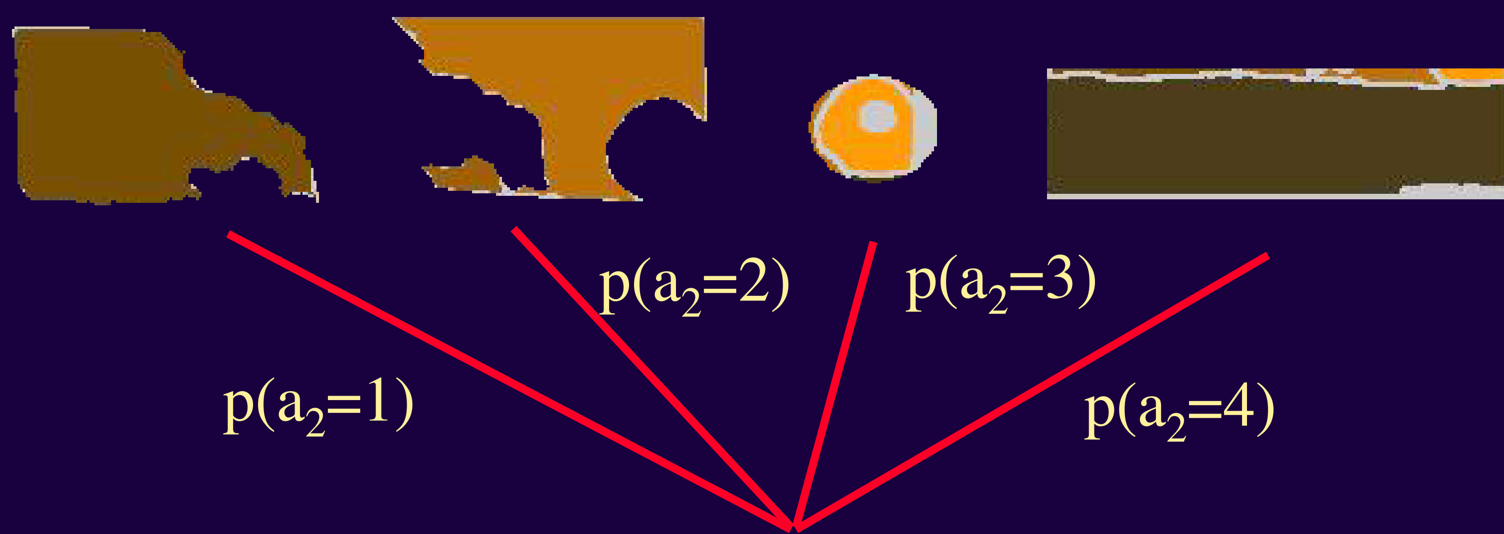
cat grass tiger water

Assignments



$$\sum_{i=1}^{B_n} p(a_1 = i) = 1$$

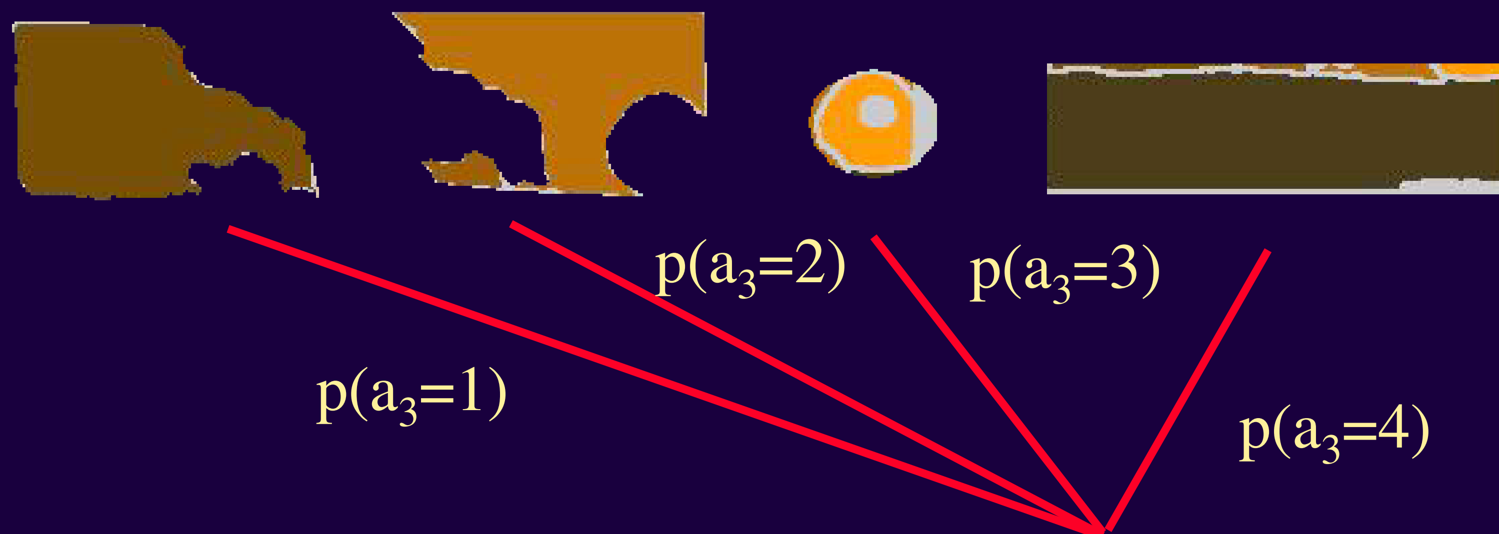
Assignments



"sun sea sky"

$$\sum_{i=1}^{B_n} p(a_2 = i) = 1$$

Assignments

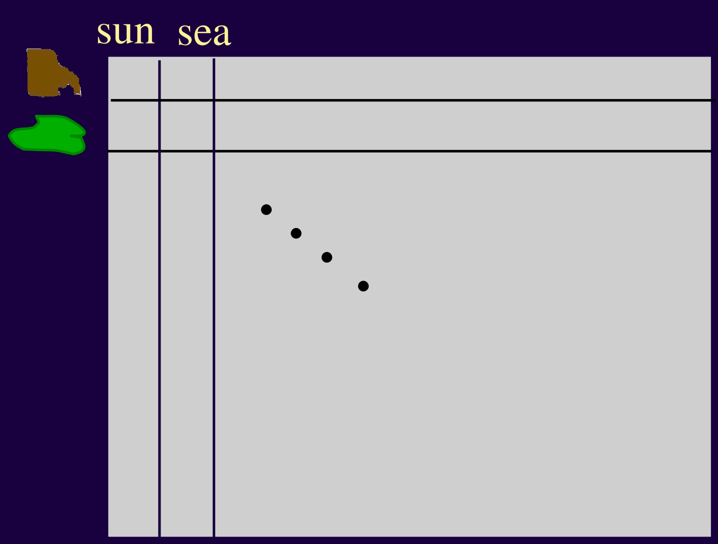


"sun sea sky"

$$\sum_{i=1}^{B_n} p(a_3 = i) = 1$$

Initialization

Initialize translation table to blob-word cooccurences (emprical joint distribution of blobs and words)



Using Expectation Maximization

$$p(w|b) = \prod_{n=1}^{N} \prod_{j=1}^{Mn} \sum_{i=1}^{Ln} p(a_{nj} = i)\, t(w = w_{nj}, b = b_{ni})$$

Given the translation probabilities estimate the correspondences

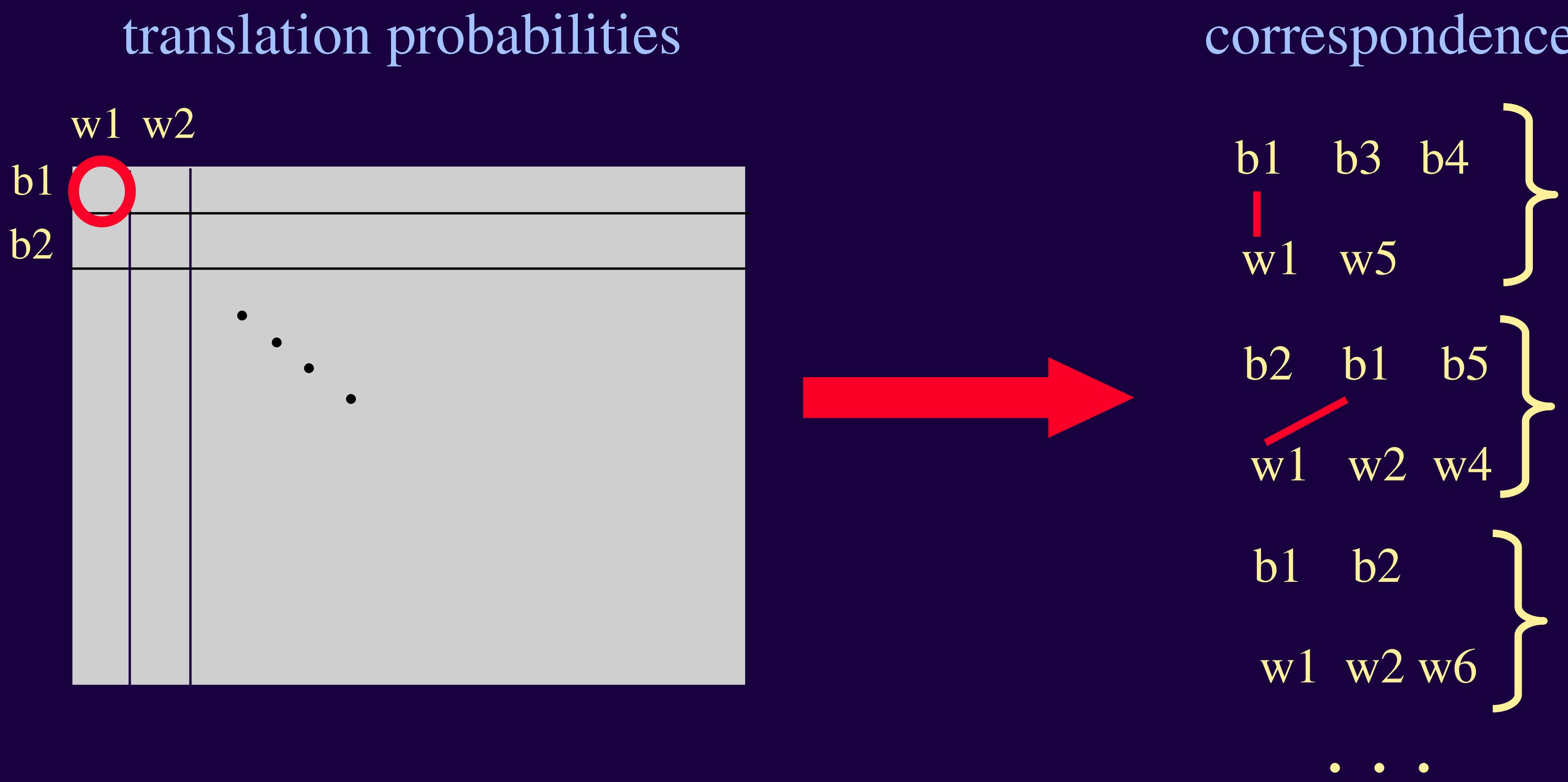
Given the correspondences estimate the translation probabilities

Dempster et al., 77

EM algorithm

E step : Predicting correspondences from translation probabilities

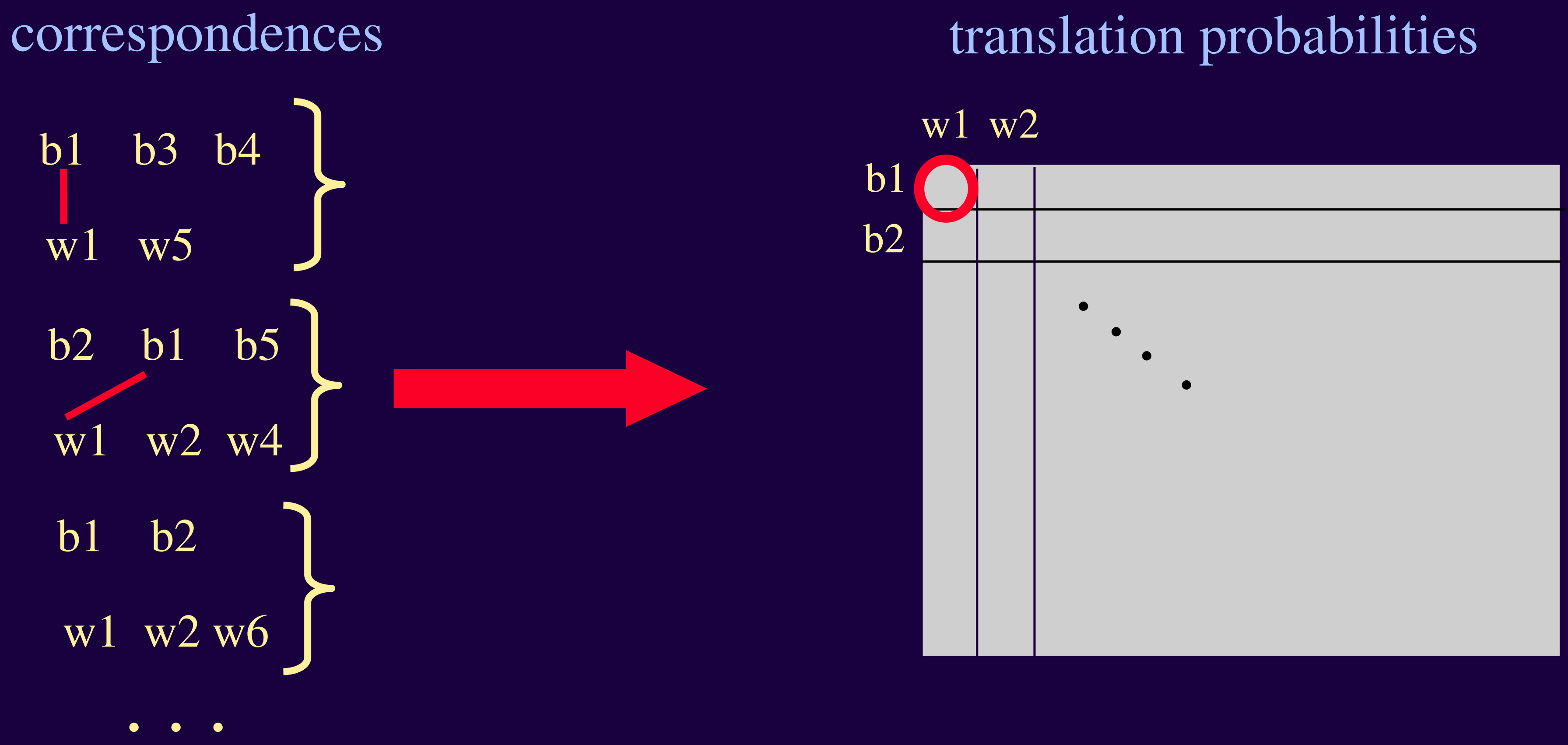
(for one pair)



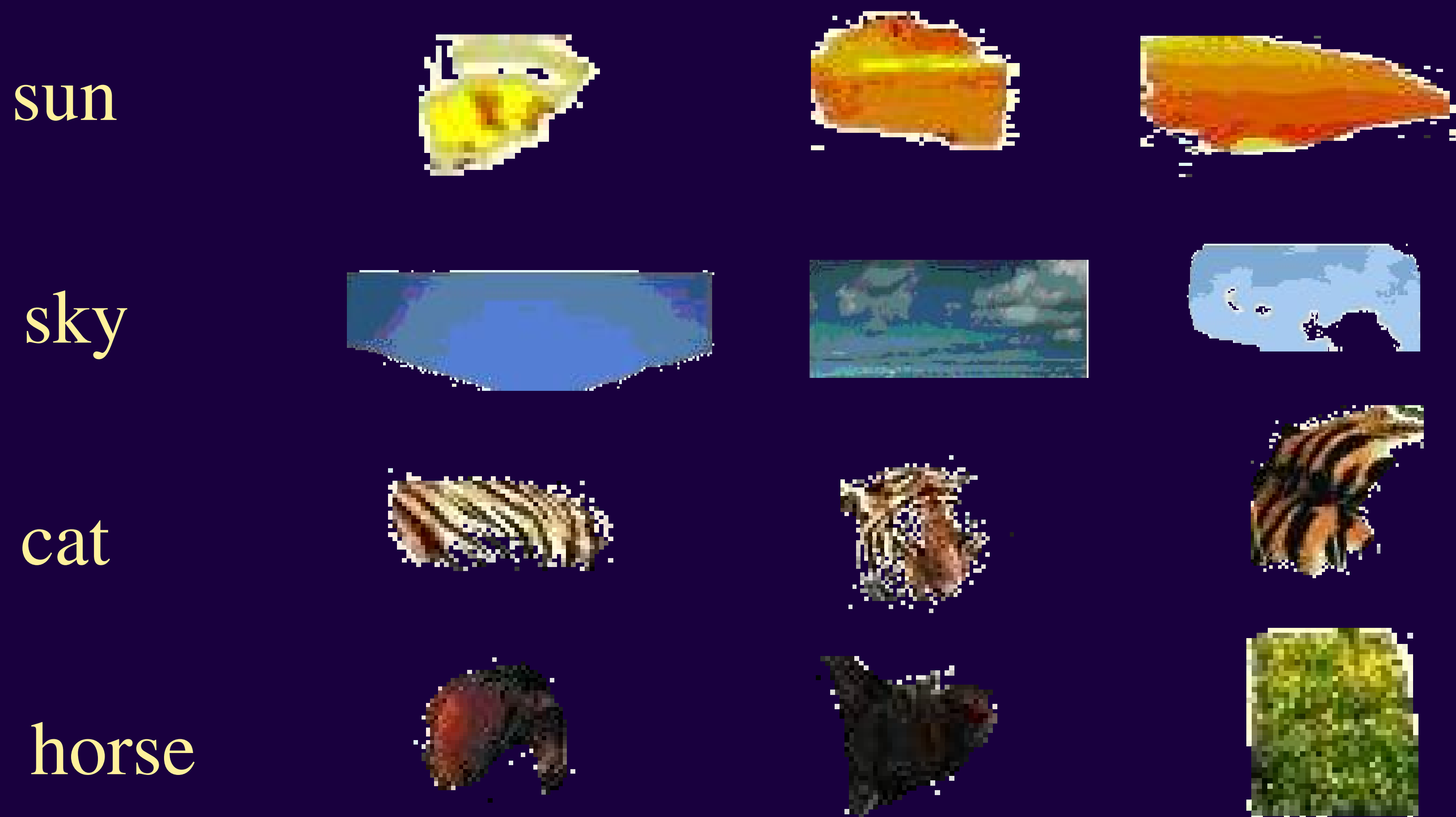
EM algorithm

M step : Predicting translation probabilities from correspondences

(for one pair)



Dictionary

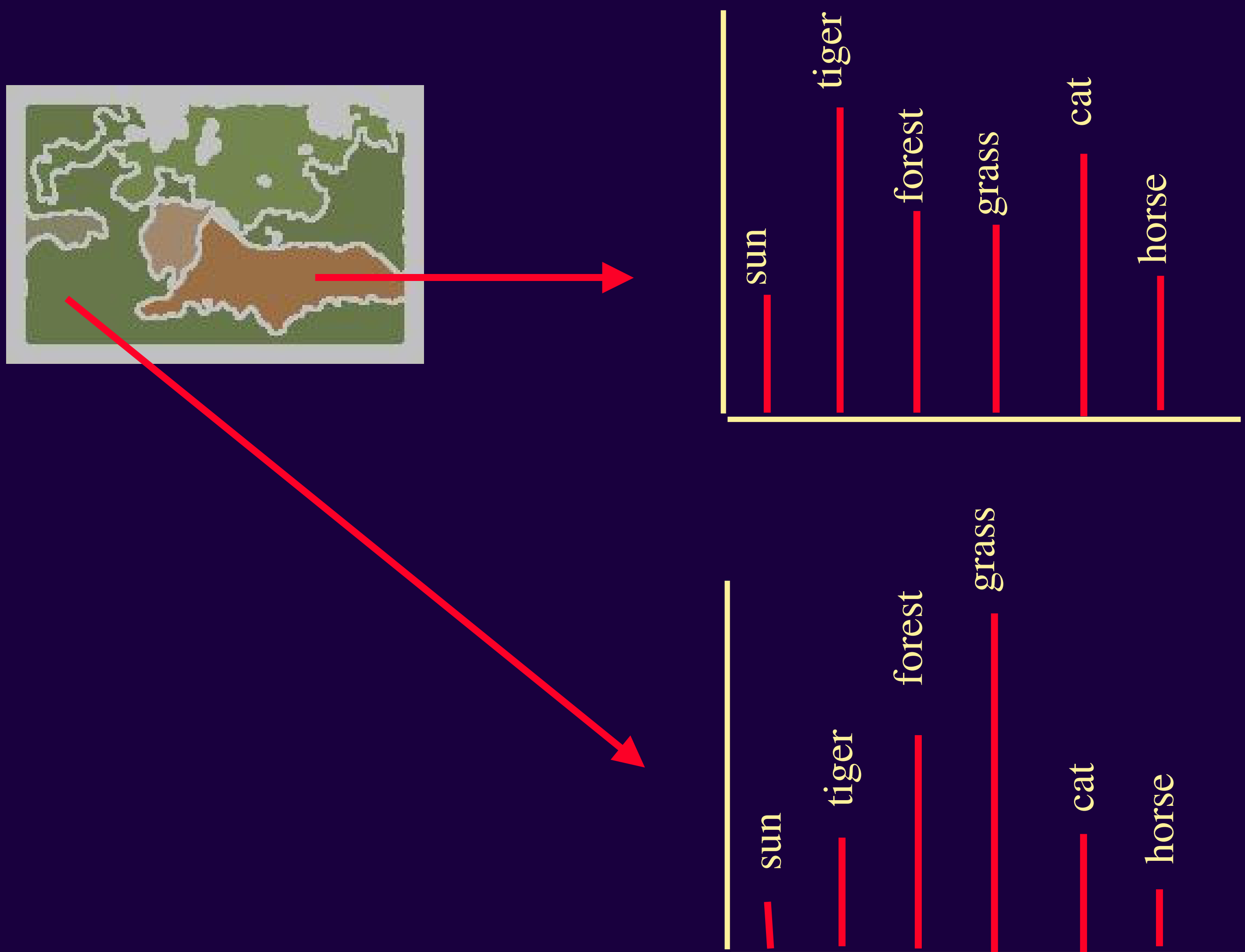


Labeling Regions

On a new image

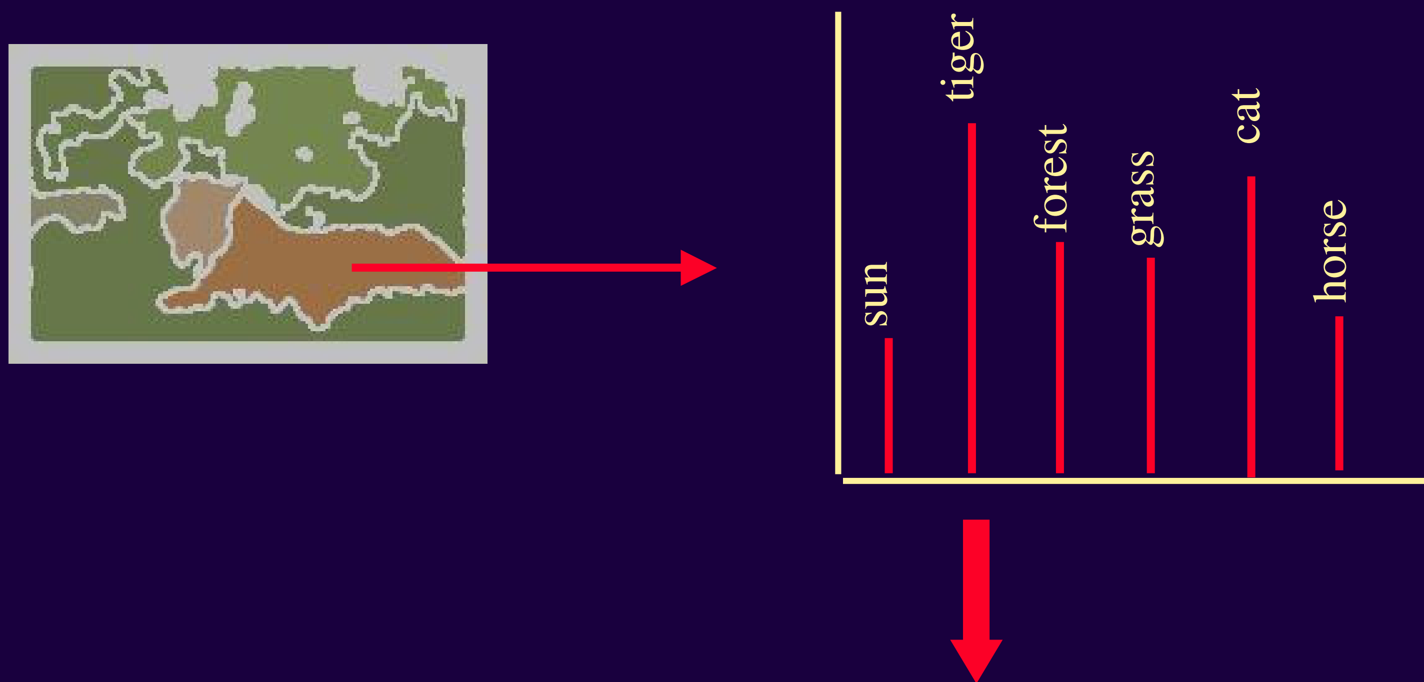
- Segment the image
- For each region
  - Find the blob token
  - Look at the word posterior given the blob

Labeling Regions

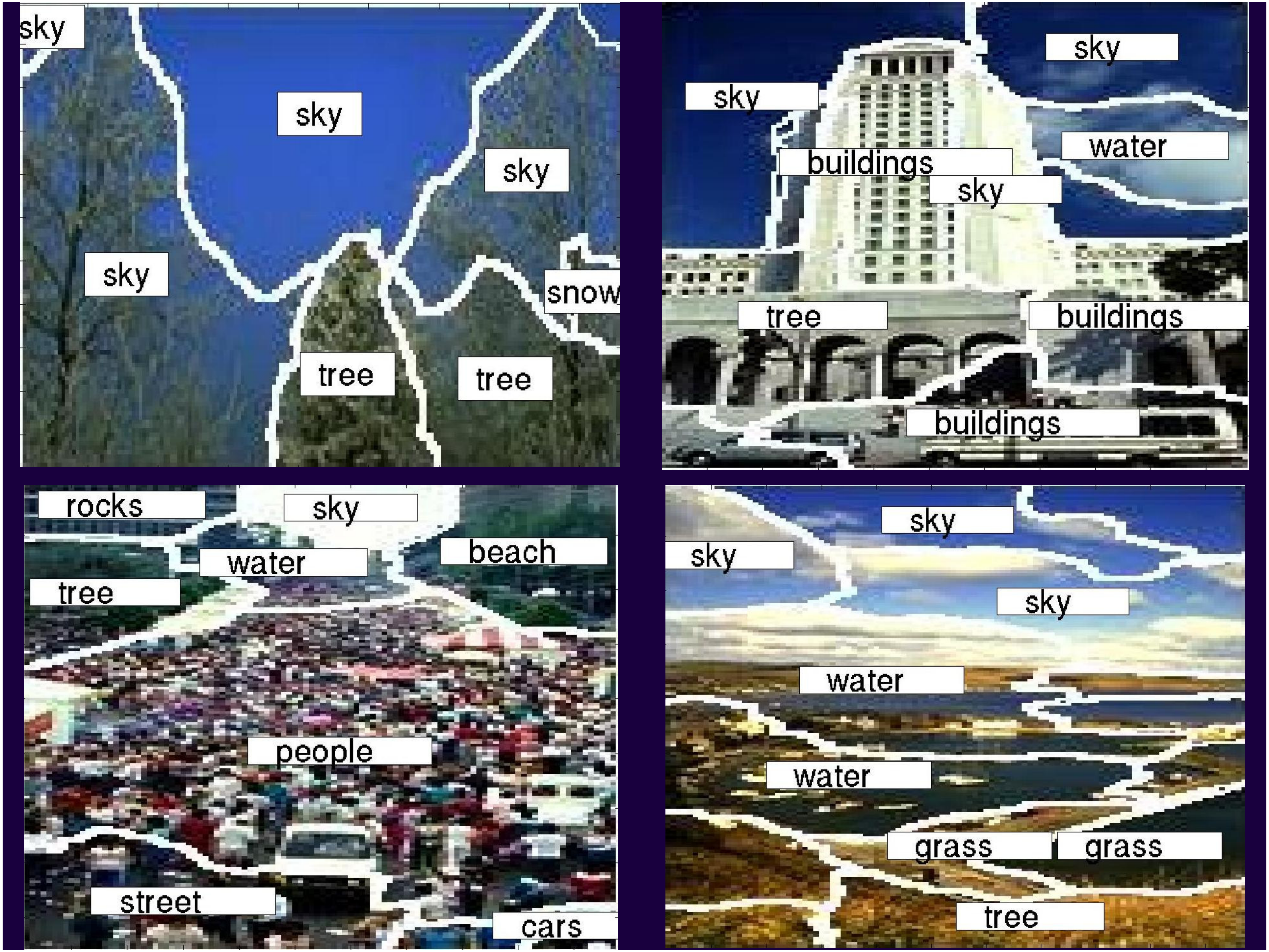


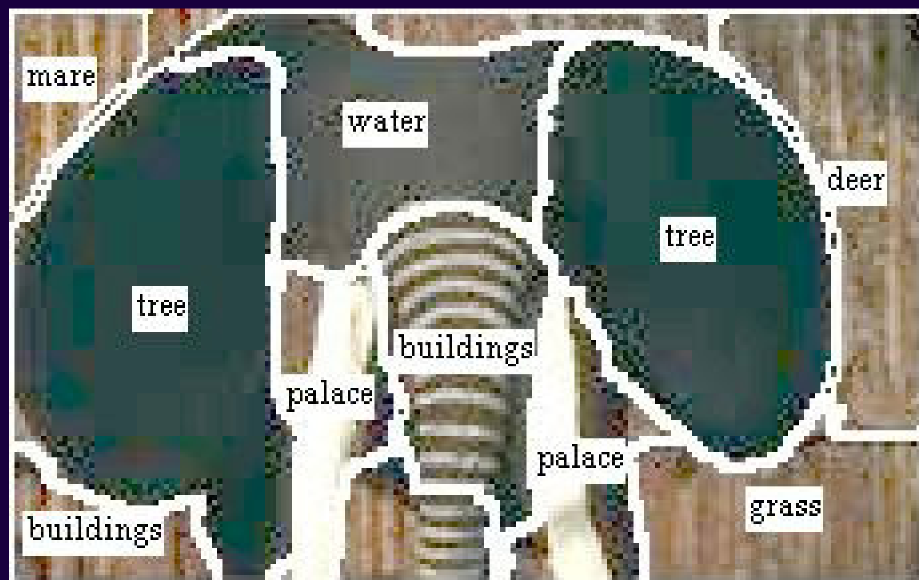
Labeling Regions

Display only maximal probable word



tiger


sky
sky
sky
sky
snow
tree
tree
sky
sky
buildings
water
sky
tree
buildings
buildings
rocks
sky
water
beach
tree
people
street
cars
sky
sky
sky
water
water
grass
grass
tree



mare
water
deer
tree
tree
buildings
palace
palace
grass
buildings

Measuring Performance

First strategy--score by hand

Second strategy--use annotation performance as a proxy.

First Strategy: Score by hand

Average performance is four times better than guessing the most common word

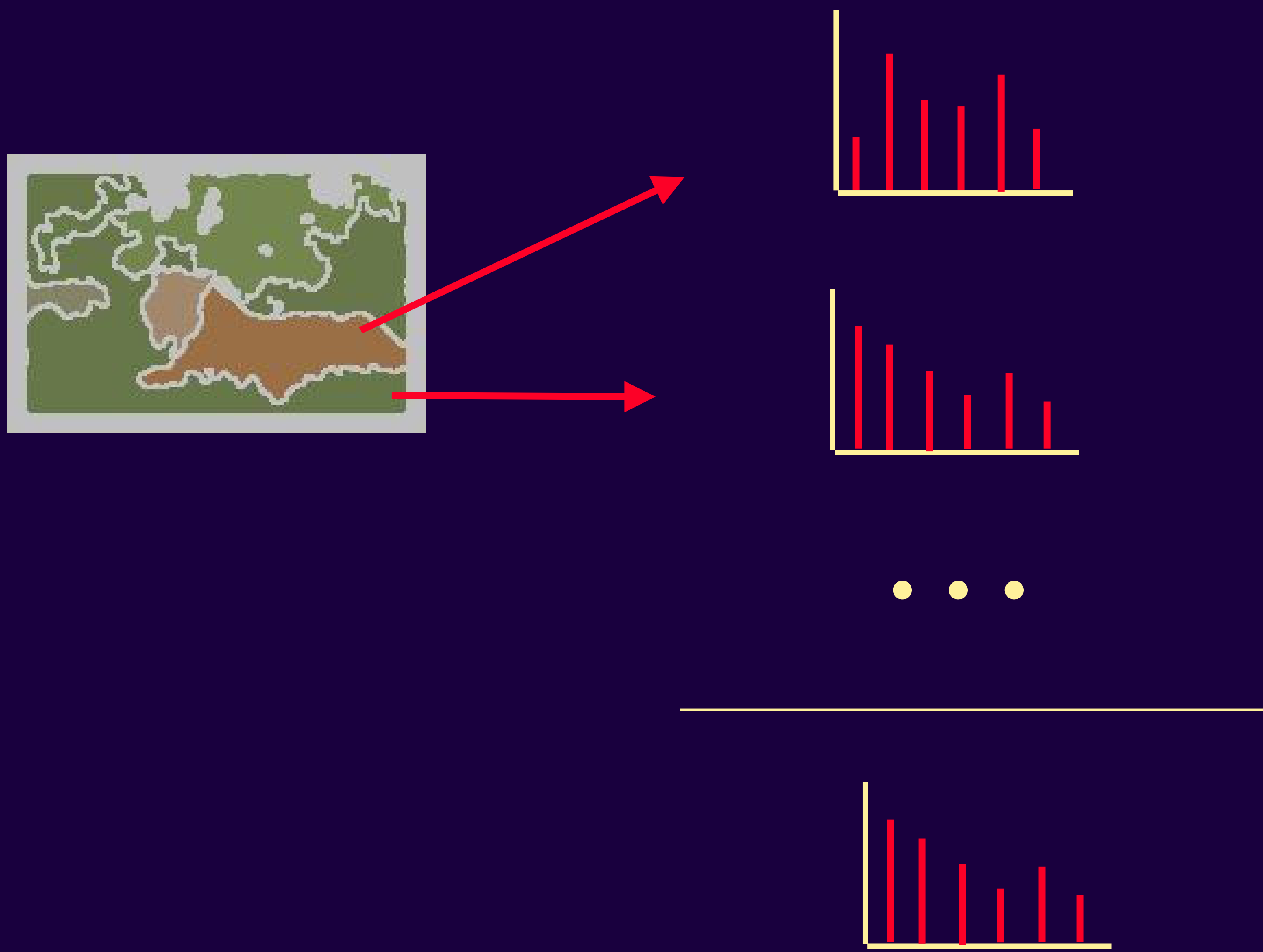
("water")

Second Strategy:
Use Annotation

tiger cat grass water

Automatic : Don’t need to do by hand

Annotating Images



Measuring Annotation Performance

Actual Keywords

GRASS TIGER CAT FOREST

Predicted Words

CAT HORSE GRASS WATER

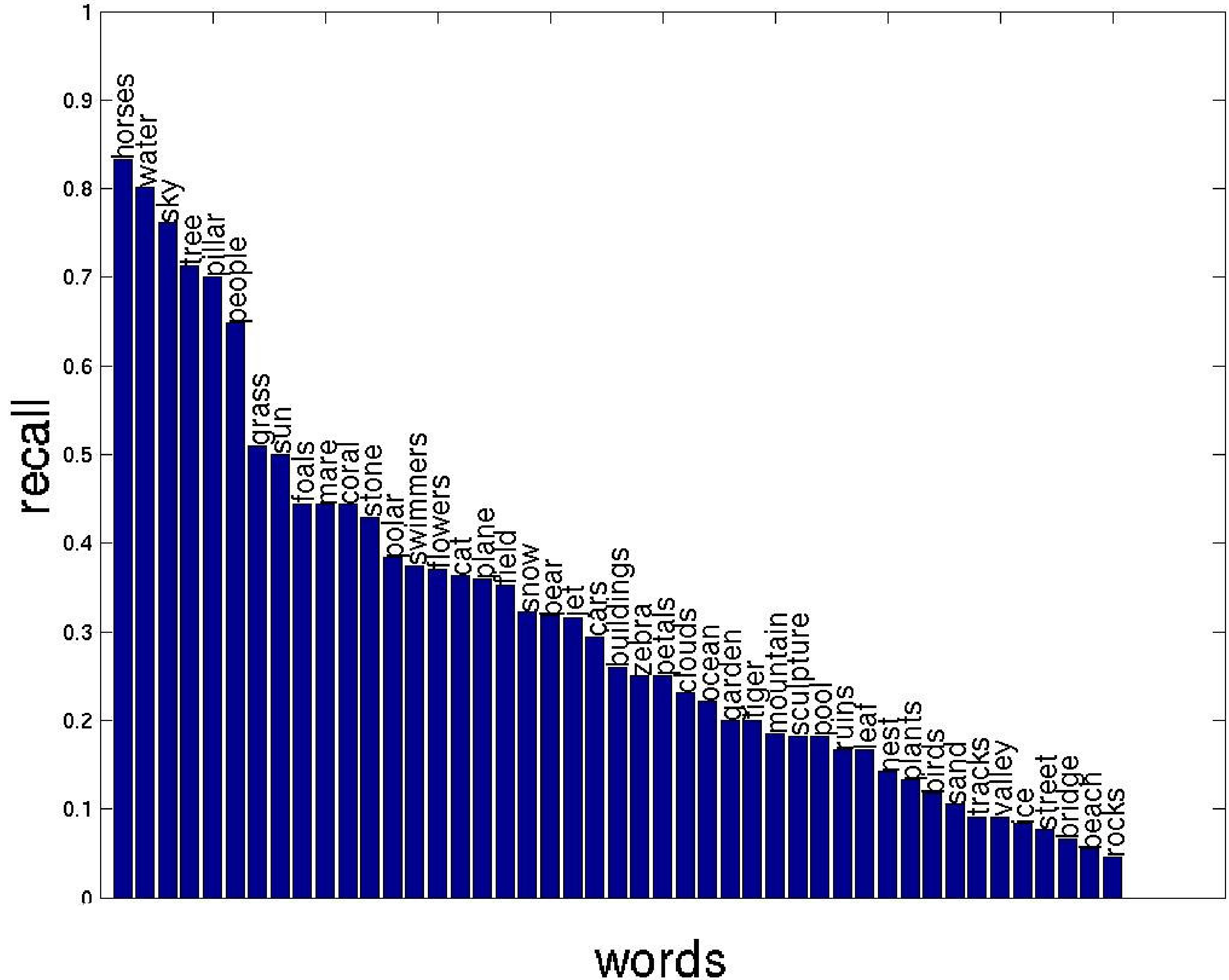
Measuring Annotation Performance

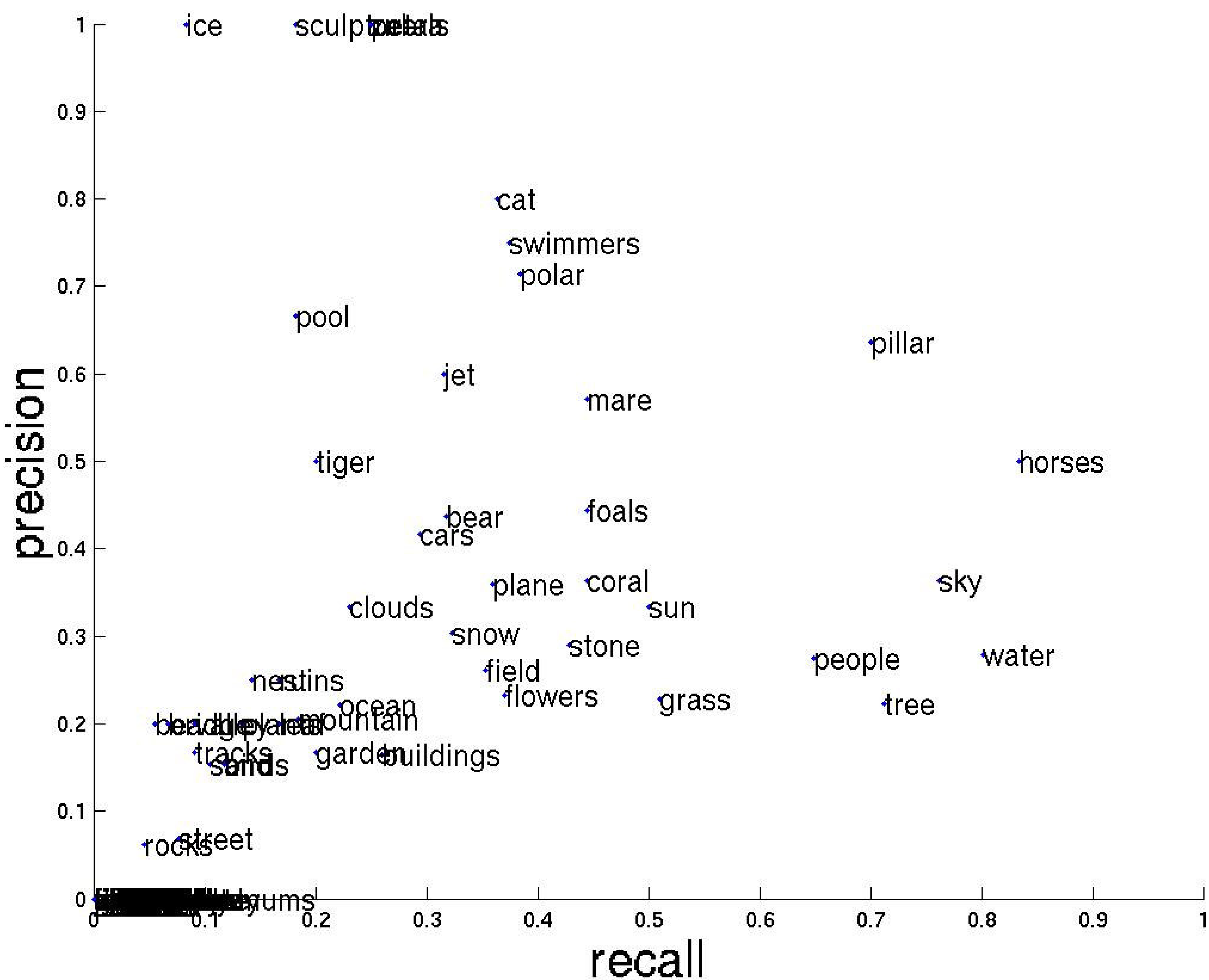
Actual Keywords

GRASS TIGER CAT FOREST

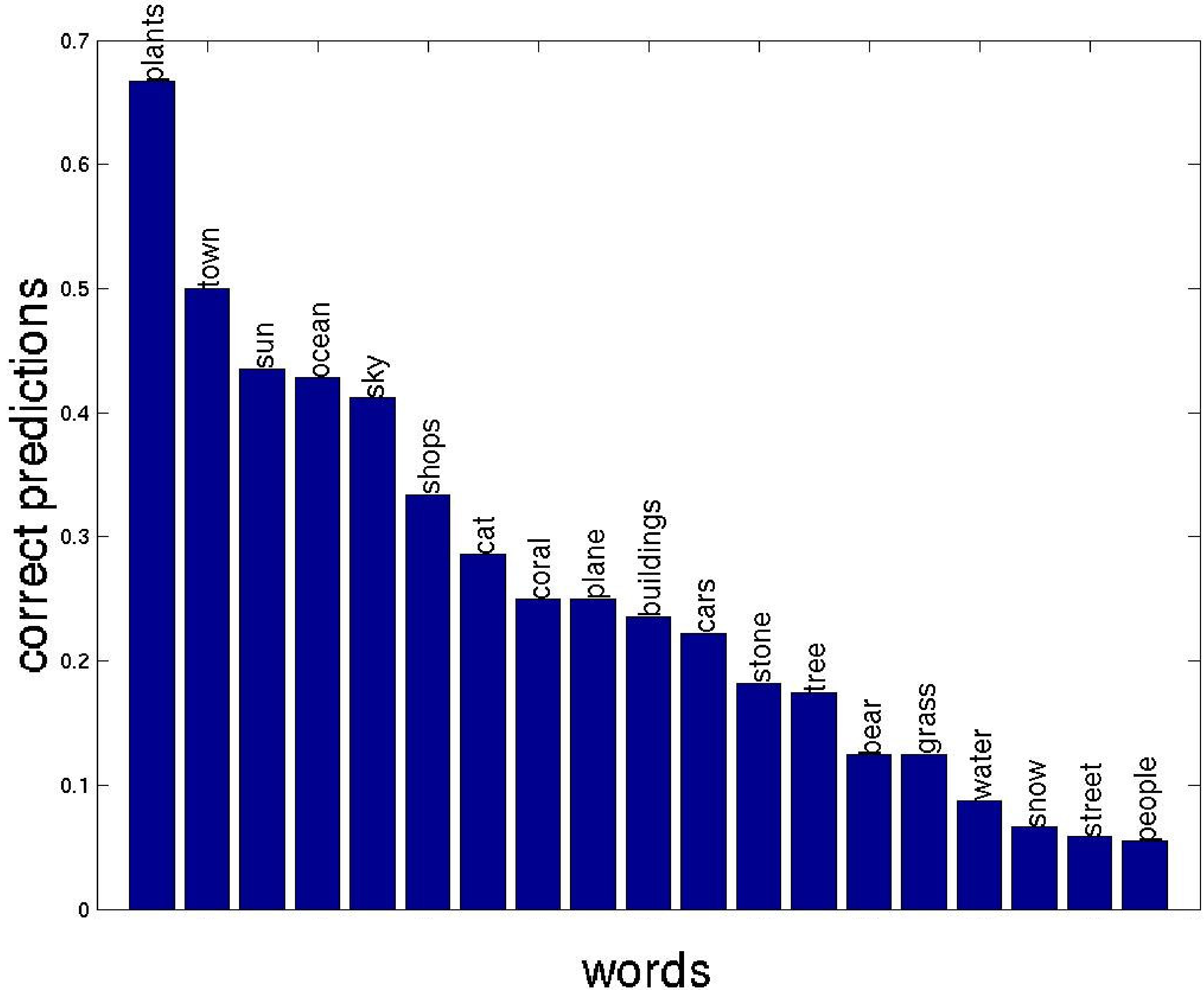
Predicted Words

CAT HORSE GRASS WATER


recall
words
horses
water
sky
tree
pillar
people
grass
sun
foals
mare
coral
stone
polar
swimmers
flowers
cat
plane
field
snow
bear
jet
cars
buildings
zebra
petals
clouds
ocean
garden
tiger
mountain
sculpture
pool
ruins
leaf
nest
plants
birds
sand
tracks
valley
ice
street
bridge
beach
rocks
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1



precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
ice
cat
swimmers
polar
pool
pillar
jet
mare
tiger
horses
bear
foals
cars
plane
coral
sky
clouds
sun
snow
stone
people
water
field
flowers
grass
tree
ocean
mountain
tracks
garden
buildings
street
rocks


correct predictions
words
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
plants
town
sun
ocean
sky
shops
cat
coral
plane
buildings
cars
stone
tree
bear
grass
water
snow
street
people

Improving the System

- Refusing to predict
- Merging indistinguishable words

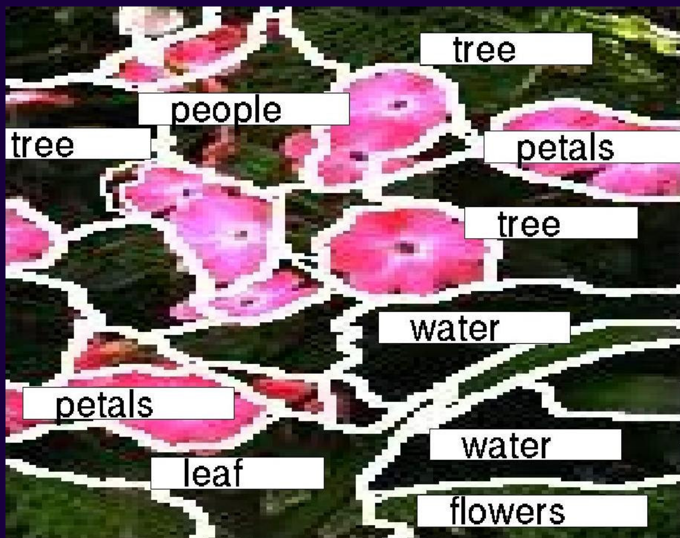
Refusing to predict

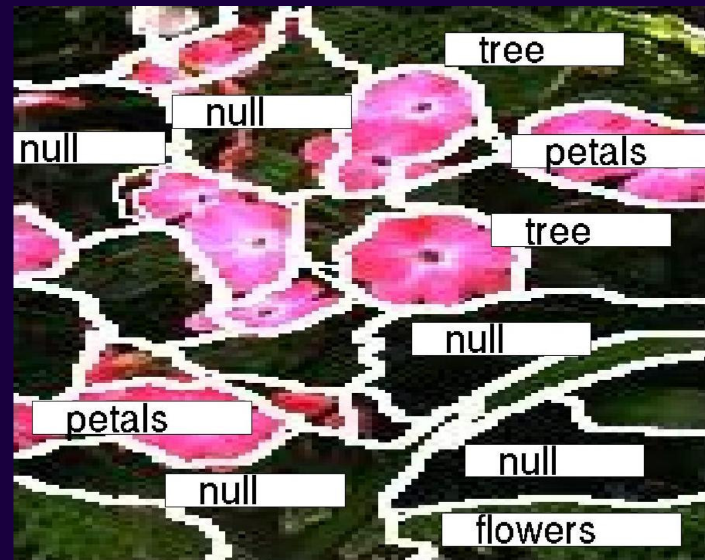
Null and fertility problems
simple solution to null - refusing to predict

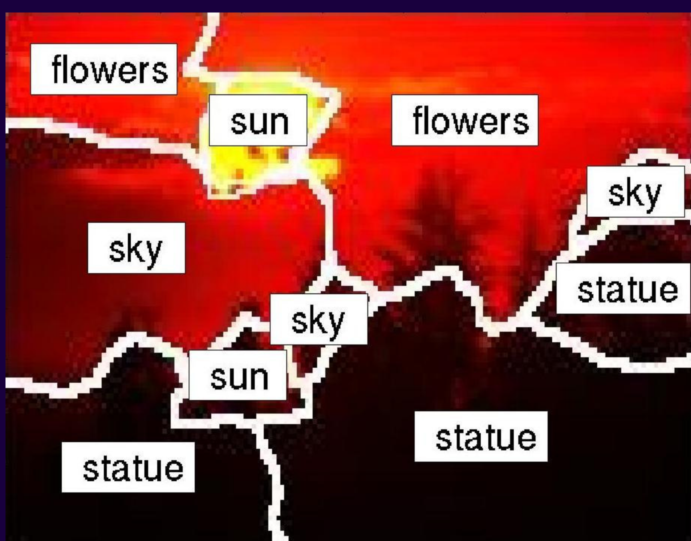
if p(word | blob) > threshold

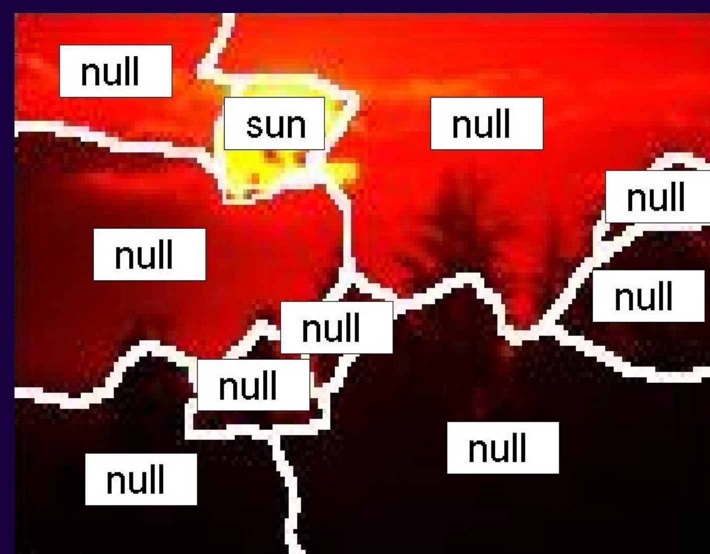
predict a word
otherwise
assign null

Examples
(null threshold = 0.2)



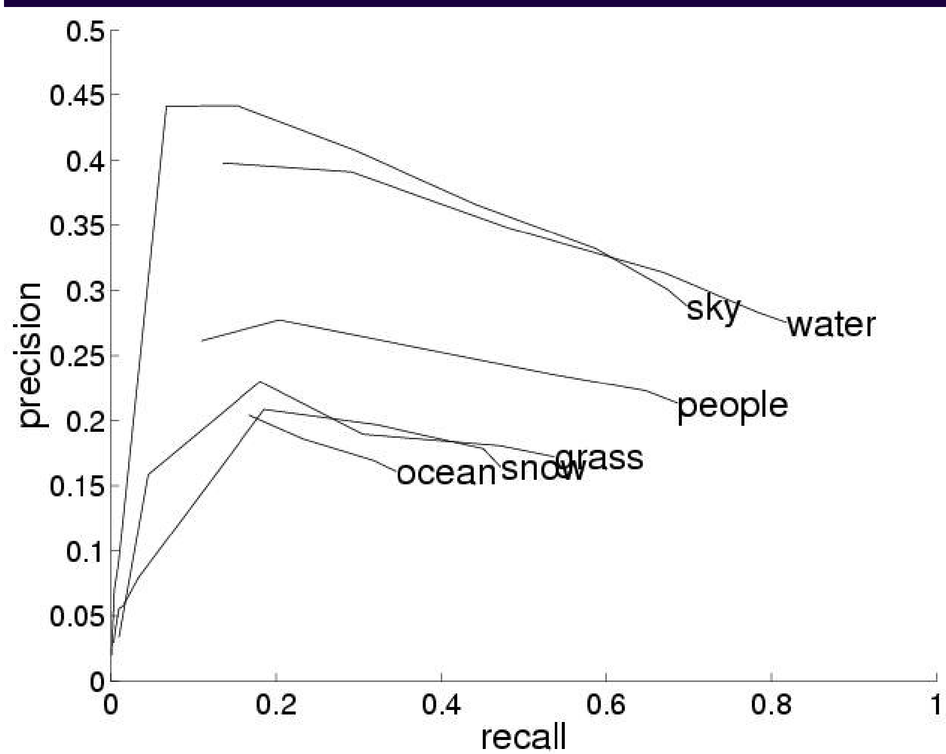




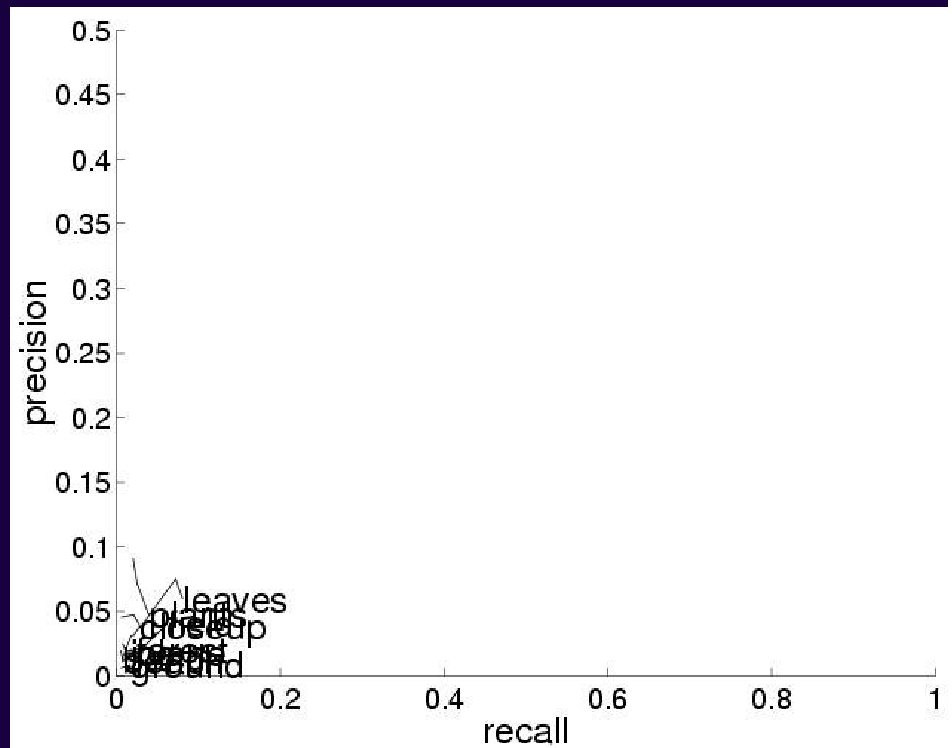


Recall and Precision

(for null threshold from 0 to 0.5)



selected good words



selected bad words

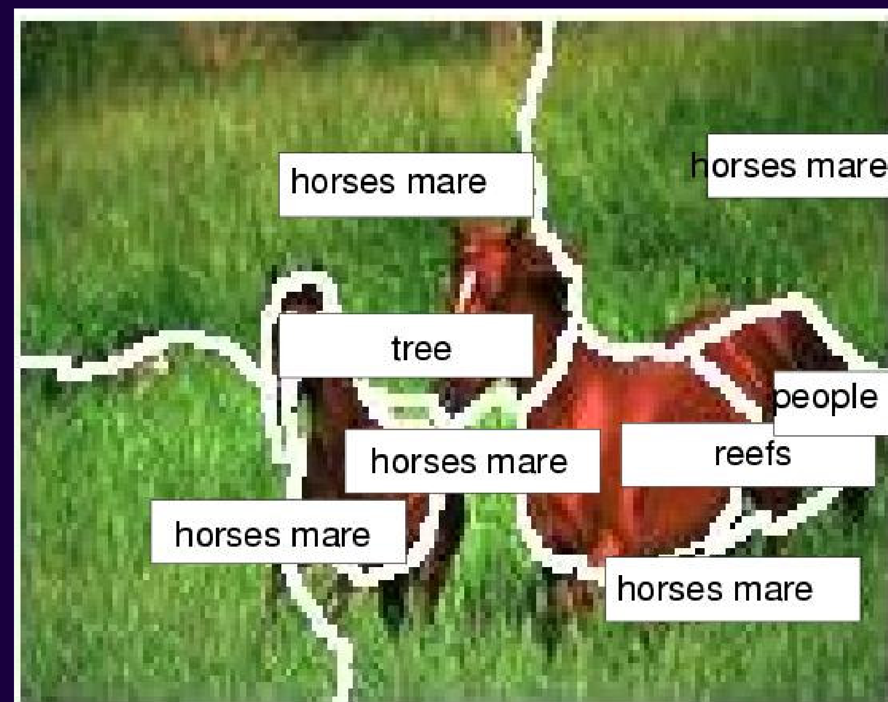
Clustering Indistinguishable Words

merge words which can’t be told apart

e.g. locomotive vs. train

Examples


sky beach
train
train locomotive
train locomotive
tree
tree
reflection
horses mare
stone
tree


horses mare
horses mare
tree
people
horses mare
reefs
horses mare
horses mare

Applying Performance Measurement

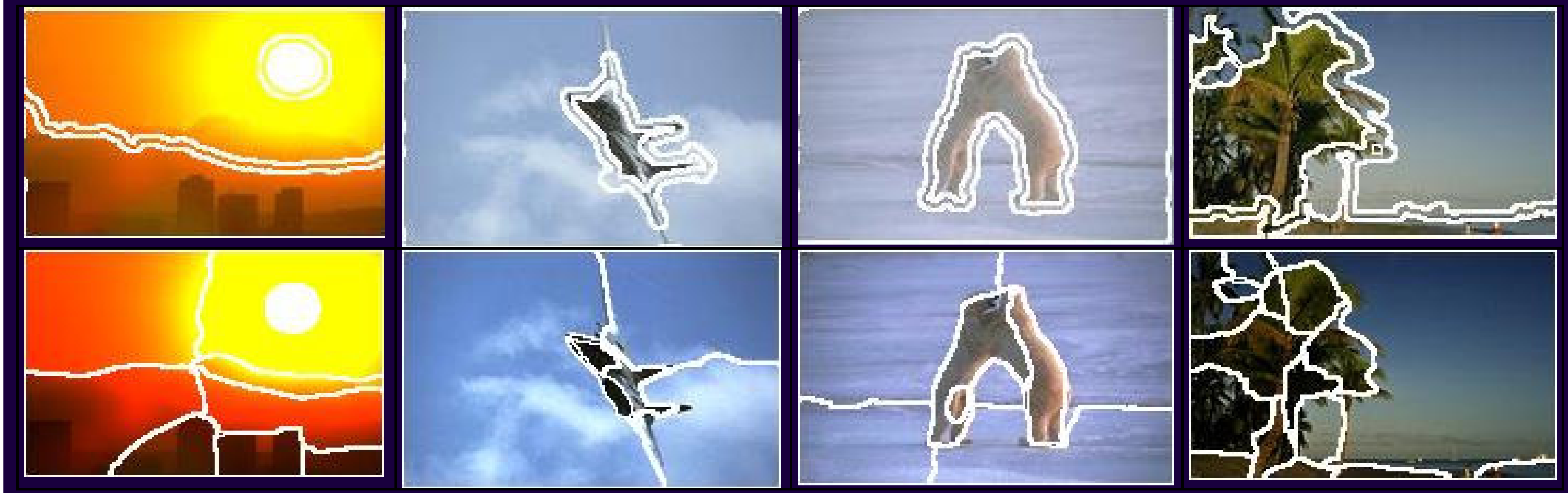
- Feature Selection
- Segmentation Comparison
- Model Selection

Feature Selection

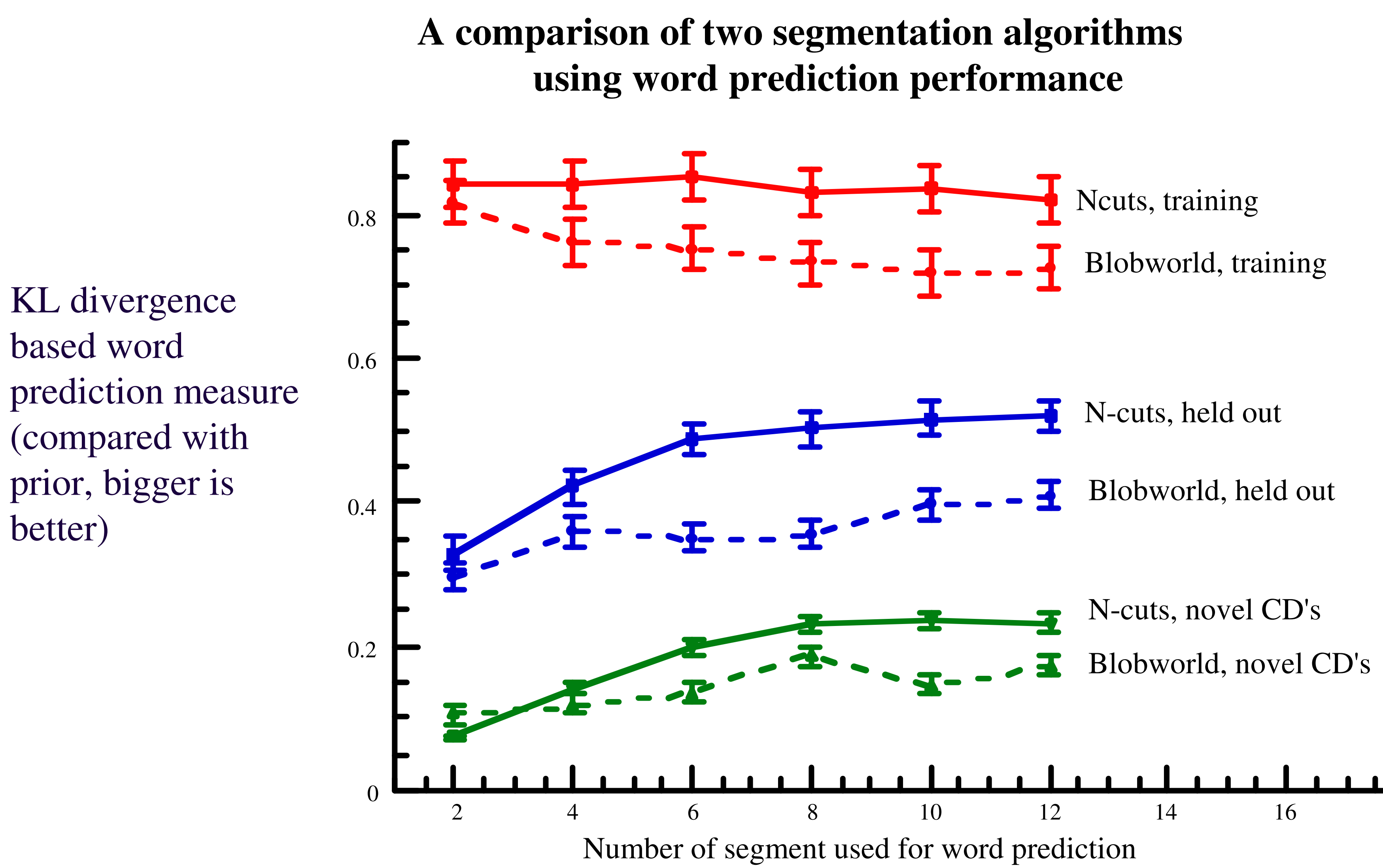
Propose good features to differentiate words that are not distinguishable (e.g., eagle and jet)

Segmentation Comparison

Blobworld segmentations



N-cuts segmentations


A comparison of two segmentation algorithms
using word prediction performance
KL divergence based word prediction measure (compared with prior, bigger is better)
0
0.2
0.4
0.6
0.8
2
4
6
8
10
12
14
16
18
Number of segment used for word prediction
Ncuts, training
Blobworld, training
N-cuts, held out
Blobworld, held out
N-cuts, novel CD's
Blobworld, novel CD's

Model Selection

Model for joint probability of text and blobs

- Clustering models
- Aspect models
- Hierarchical models
- Bayesian models
- Co-occurrence models

Many of these based on models proposed for text [Brown, Della Pietra, Della Pietra & Mercer 93; Hofmann 98; Hofmann & Puzicha 98]

A comparison paper is submitted to JMLR
'Matching words and Pictures', Barnard, Duygulu, Forsyth, Freitas, Blei, Jordan

Discussion

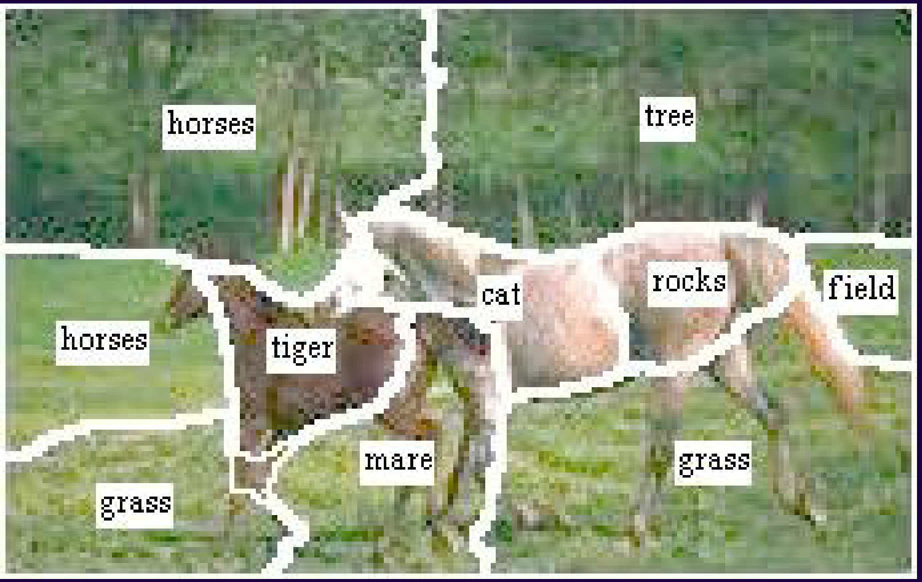
Recognition on the large scale

Unsupervised - using the available data efficiently

Learn what to recognize

Future Directions

Estimate where a minimal amount of supervision can be most helpful (and provide it)



Using labelled data

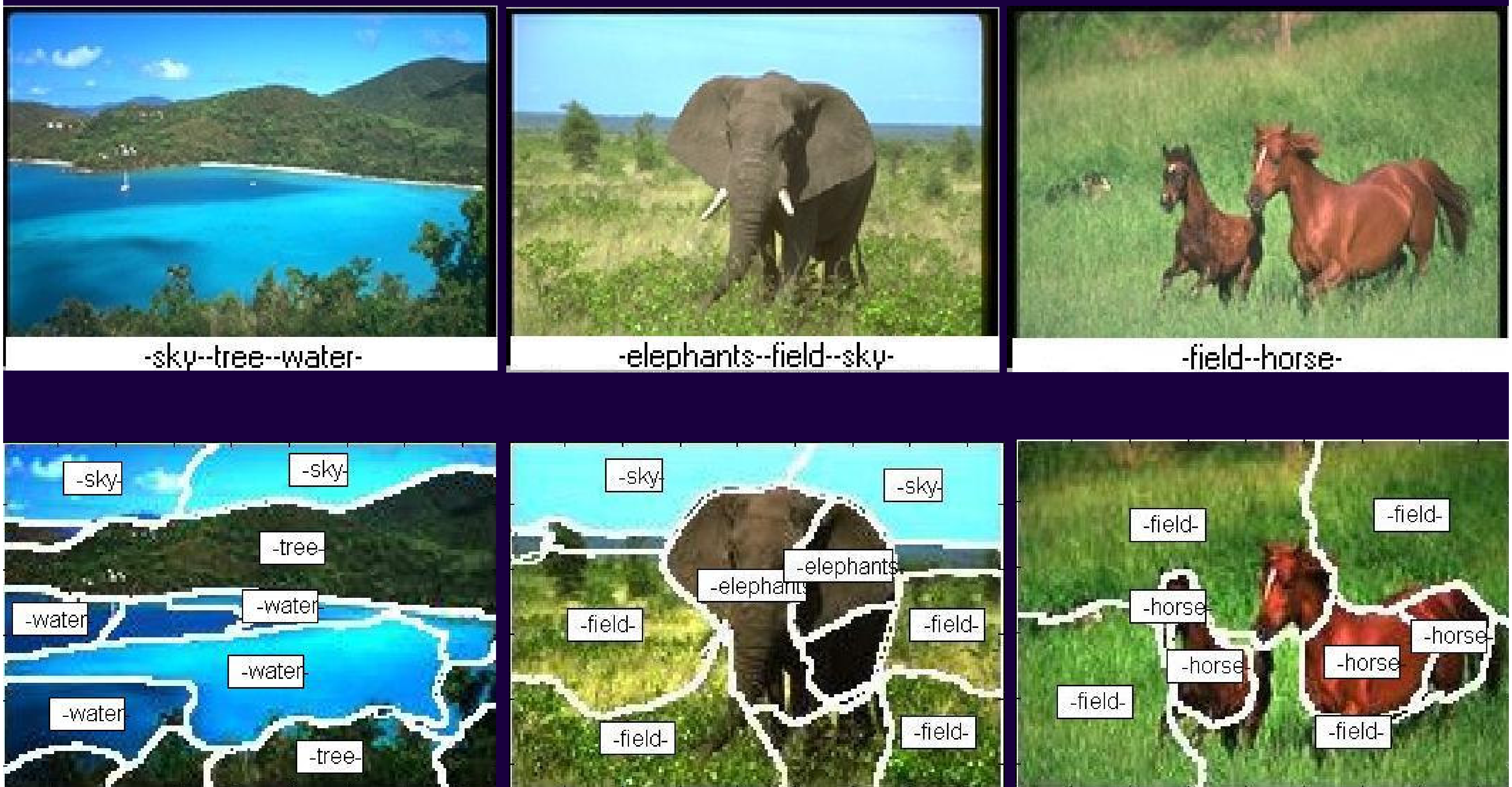
500 hand labeled images

Modified to be added to each of 10 sets

very hard !!!

- takes a lot of time
- large vocabulary
- cheetah, leopard or cat

Using labelled data



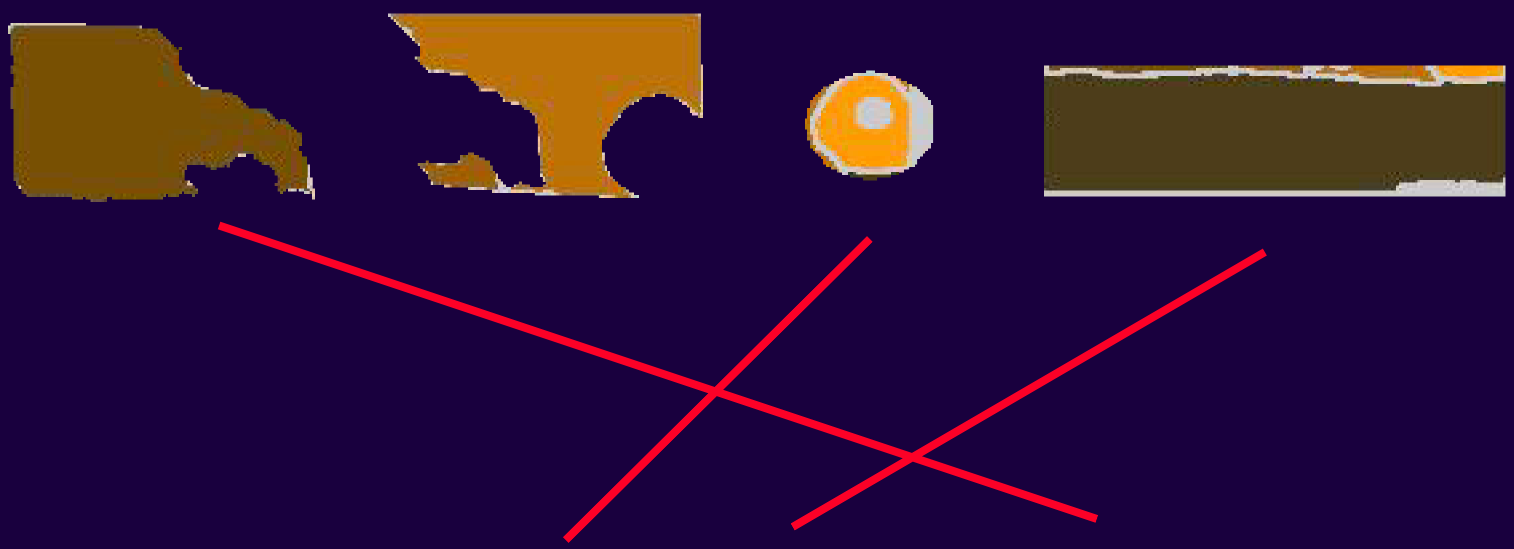
Using labelled data

use them to supervise

-add to data

-fix correspondences

-retrain



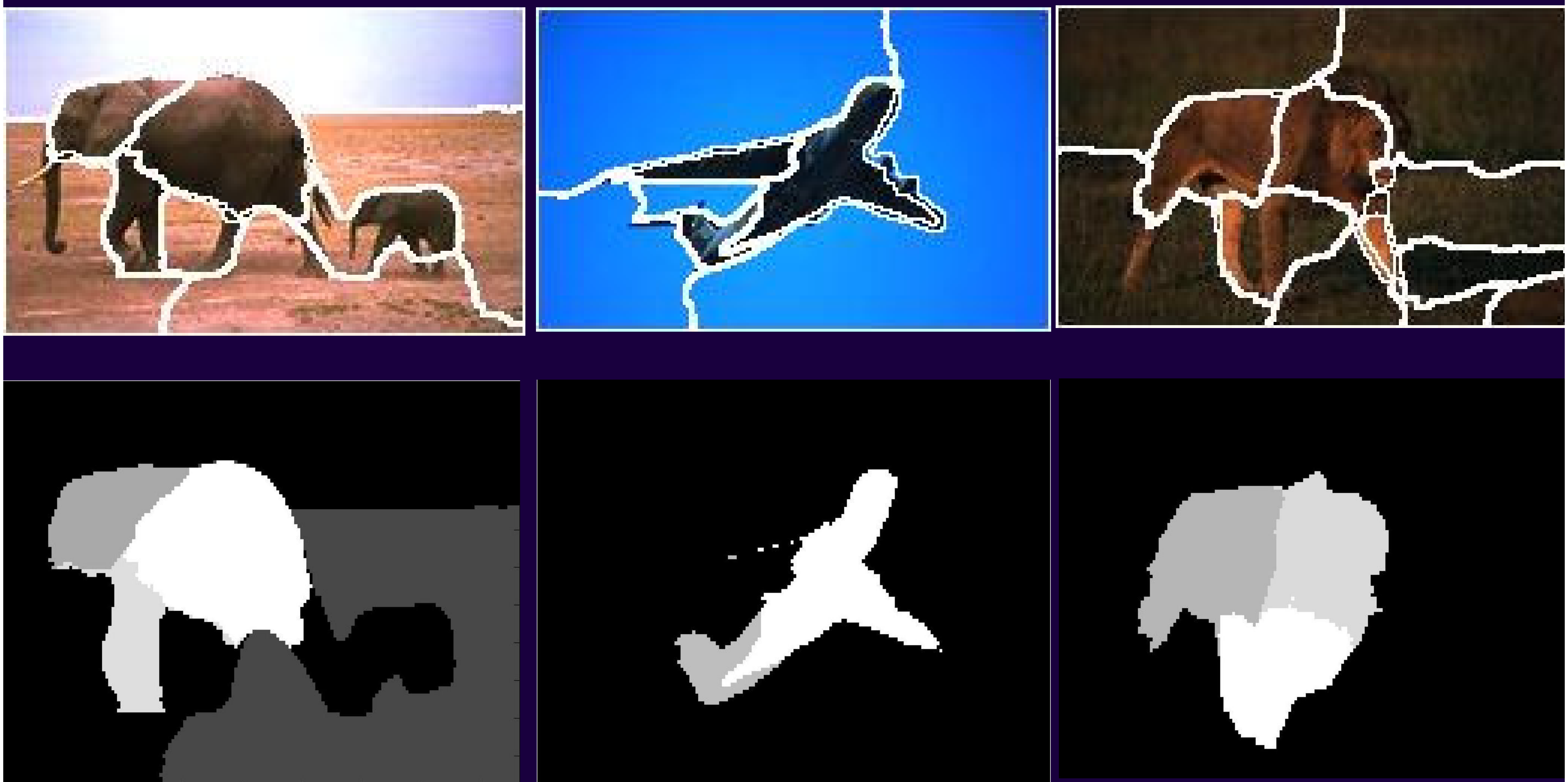
“sun sea sky”

Future Directions

Propose region merging based on posterior word probabilities

Propose merging

Preliminary Results



elephant plane cat

Future Directions (other data)

Corel Image Data	40,000 images
Fine Arts Museum of San Francisco	83,000 images online
Cal-flora	20,000 images, species information
News photos with captions (yahoo.com)	1,500 images per day available from yahoo.com
Hulton Archive	40,000,000 images (only 230,000 online)
internet.archive.org	1,000 movies with no copyright
TV news archives (televisionarchive.org, informedia.cs.cmu.edu)	Several terabytes already available
Google Image Crawl	>330,000,000 images (with nearby text)
Satellite images (terrarserver.com, nasa.gov, usgs.gov)	(And associated demographic information)
Medial images	(And associated with clinical information)

FAMSF Data
(83,000 images online)

Web number: 4359202410830012

rec number: 2

Title: Le Matin

Primary class: Print

Artist: Tissot

Description:
serving woman stands in a dressing room, in front of vanity with chair, mirror and mantle, holding a tray with tea and toast

Display date: 1886

Country: France

Natural Language Processing

- Parts of speech* (prefer nouns for now)
- Sense Disambiguation
- Expand semantics using WordNet †

* We use Eric Brill's parts of speech tagger (available on-line)

† WordNet is an on-line lexical reference system from Princeton (Miller et.al)

Multiple Senses

26078 water grass trees **bank**s

125090 **bank** machine money currency bills

125084 piggy **bank** coins currency money

212001 **bank** buildings trees city

173044 mink rodent **bank** grass

151096 snow **bank**s hills winter

News data

News photos with captions
(1500 images per day available from yahoo.com)

learn topic structure using both images and text

different pictures for the same topic

different stories that use the same picture

Other Applications

- Auto Annotation
- Auto Illustration
- Organizing Image Collections for Browsing

Words from Pictures (Auto-annotation)

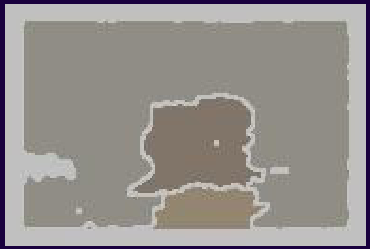
Keywords

GRASS TIGER CAT FOREST

Predicted Words (rank order)

tiger cat grass people water bengal buildings ocean forest reef

---



Keywords

HIPPO BULL mouth walk

Predicted Words (rank order)

water hippos rhino river grass reflection one-horned head plain sand

---

Keywords

FLOWER coralberry LEAVES PLANT

Predicted Words (rank order)

fish reef church wall people water landscape coral sand trees

Pictures from Words (Auto-illustration)

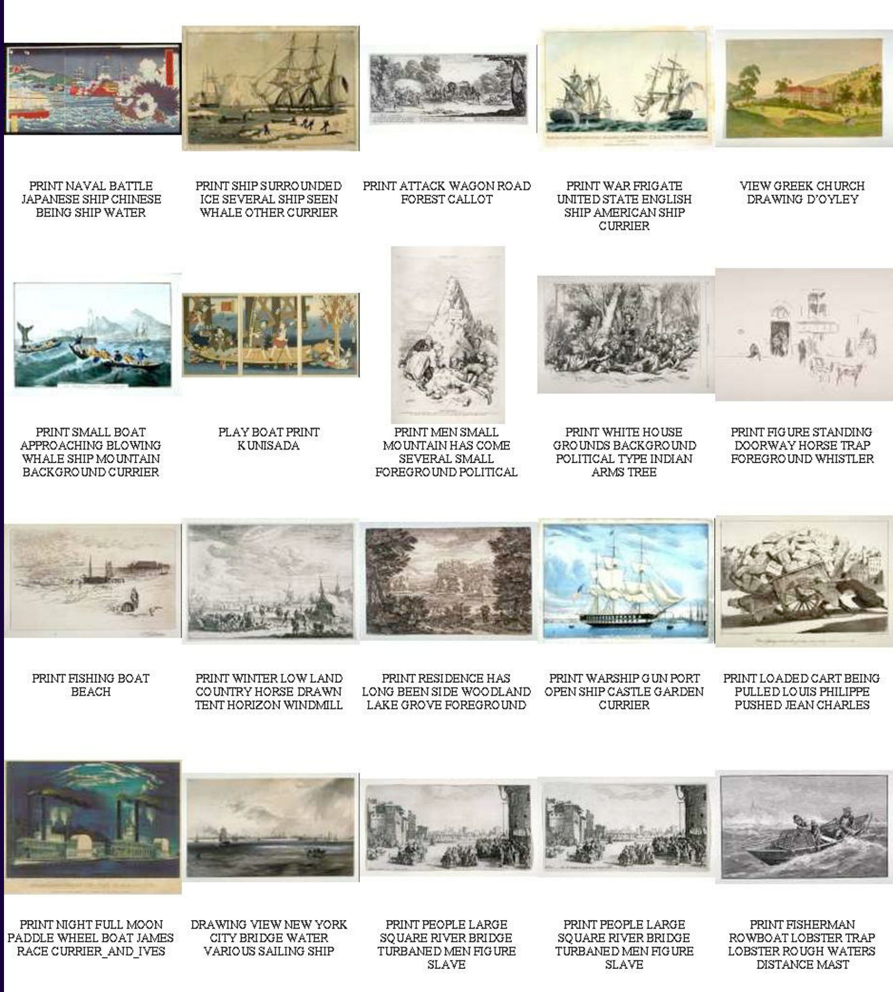
Text Passage (Moby Dick)

"The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship …"

Extracted Query

large importance attached fact
old dutch century more command
whale ship was person was
divided officer word means fat
cutter time made days was
general vessel whale hunting
concern british title old dutch ...

Retrieved Images



PRINT NAVAL BATTLE JAPANESE SHIP CHINESE BEING SHIP WATER

PRINT SHIP SURROUNDED ICE SEVERAL SHIP SEEN WHALE OTHER CURRIER

PRINT ATTACK WAGON ROAD FOREST CALLOT

PRINT WAR FRIGATE UNITED STATE ENGLISH SHIP AMERICAN SHIP CURRIER

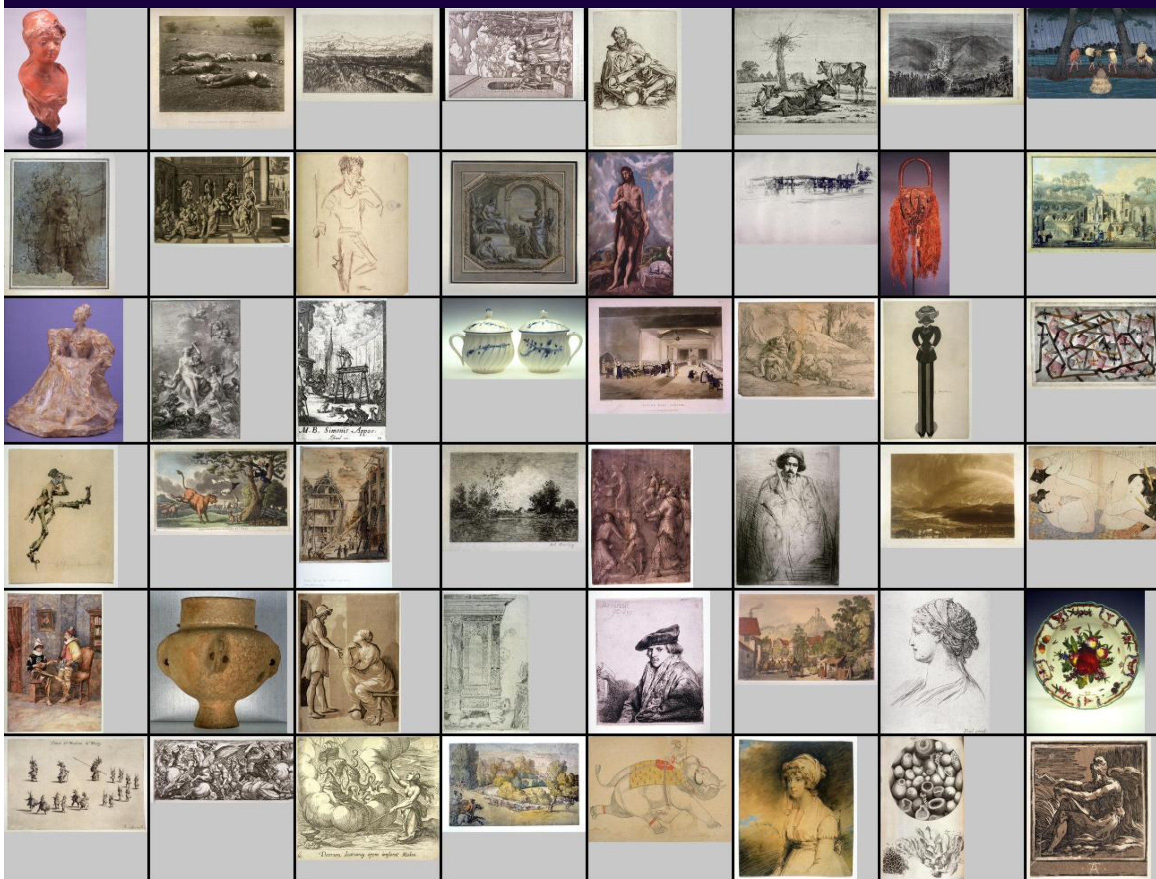
PRINT SMALL BOAT APPROACHING BLOWING WHALE SHIP MOUNTAIN BACKGROUND CURRIER

PLAY BOAT PRINT KUNISADA

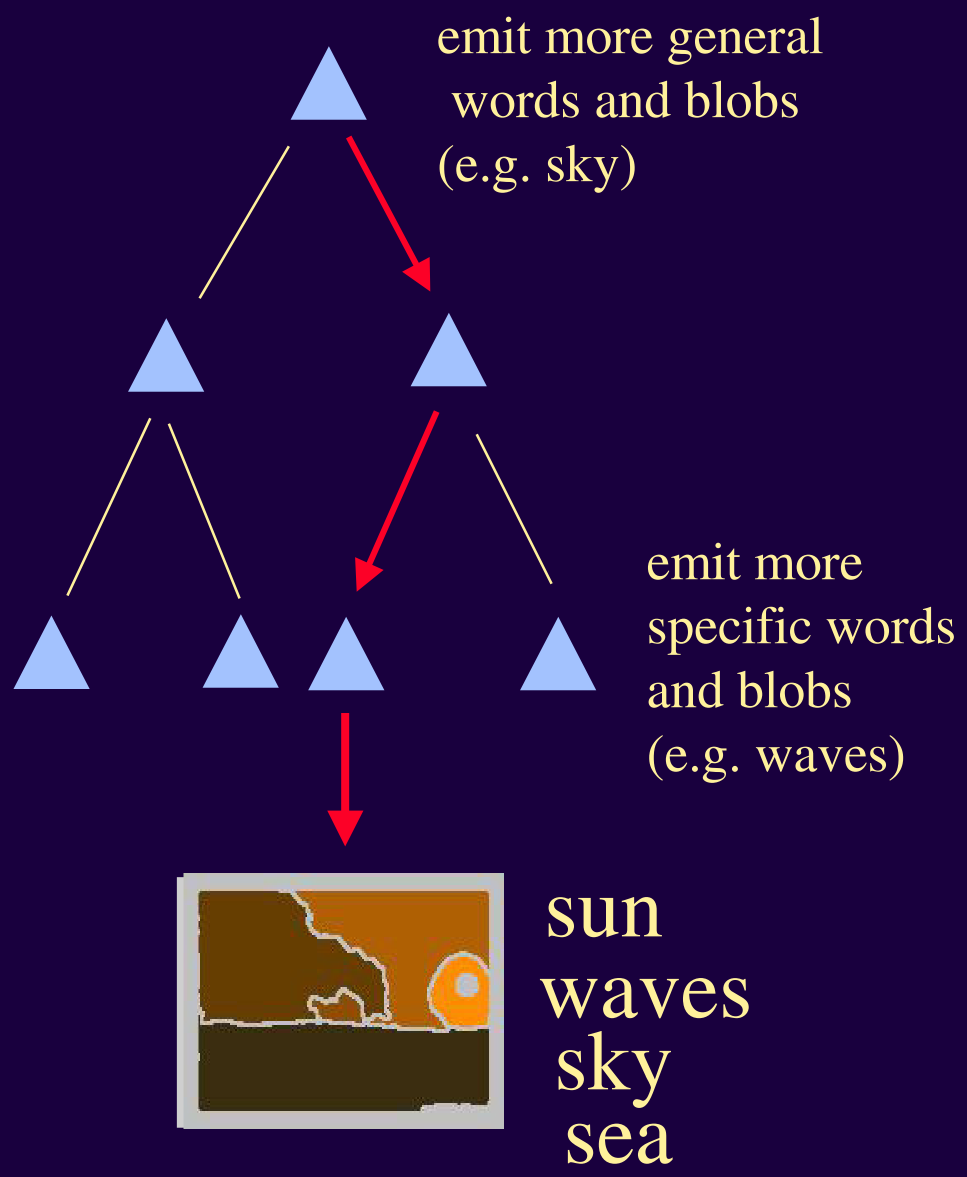
PRINT MEN SMALL MOUNTAIN HAS COME SEVERAL SMALL FOREGROUND POLITICAL

PRINT WHITE HOUSE GROUNDS BACKGROUND POLITICAL TYPE INDIAN ARMS TREE

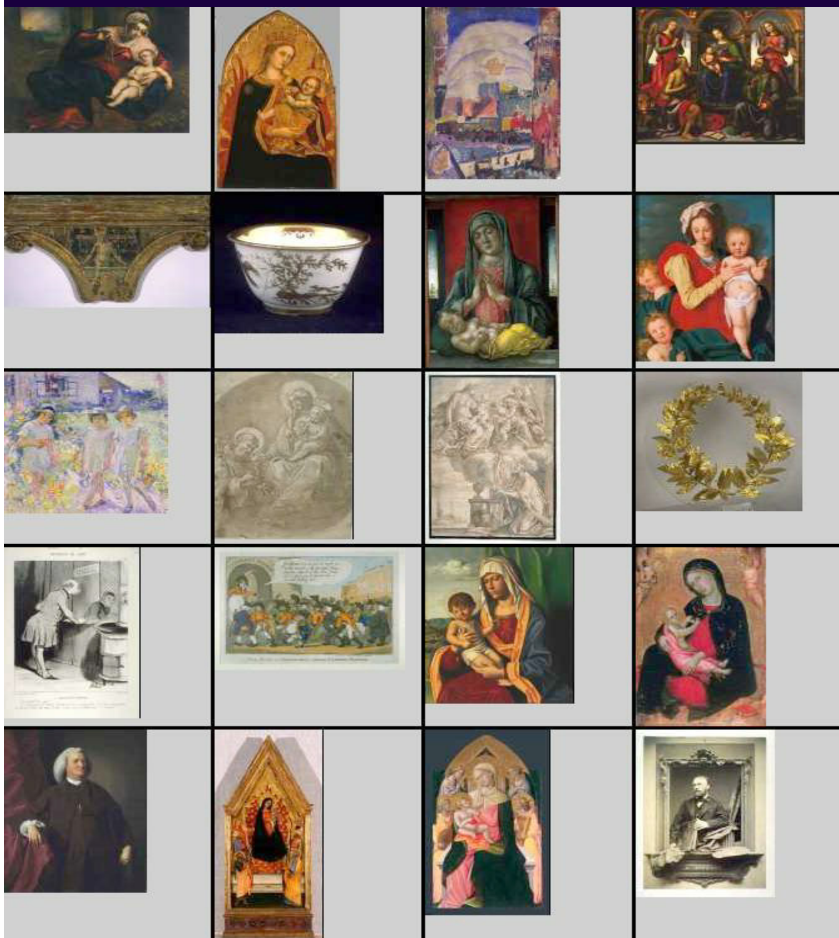
Organizing Image Collections

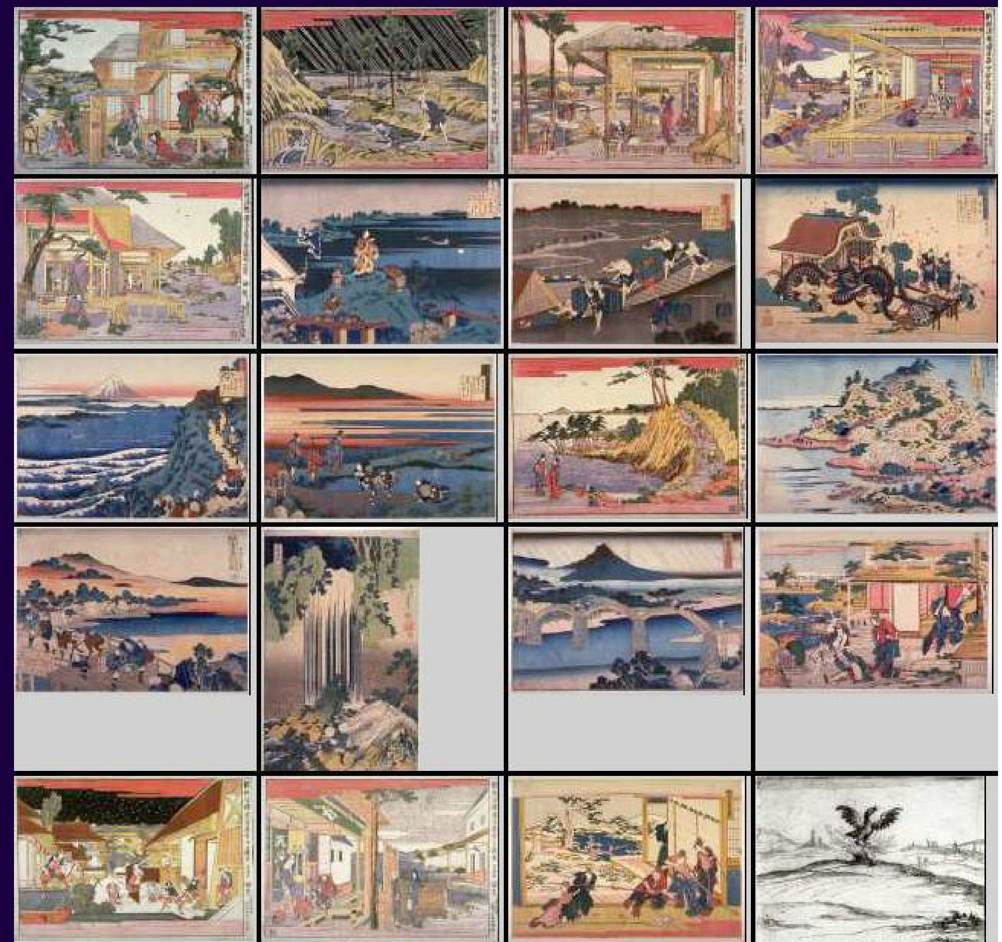


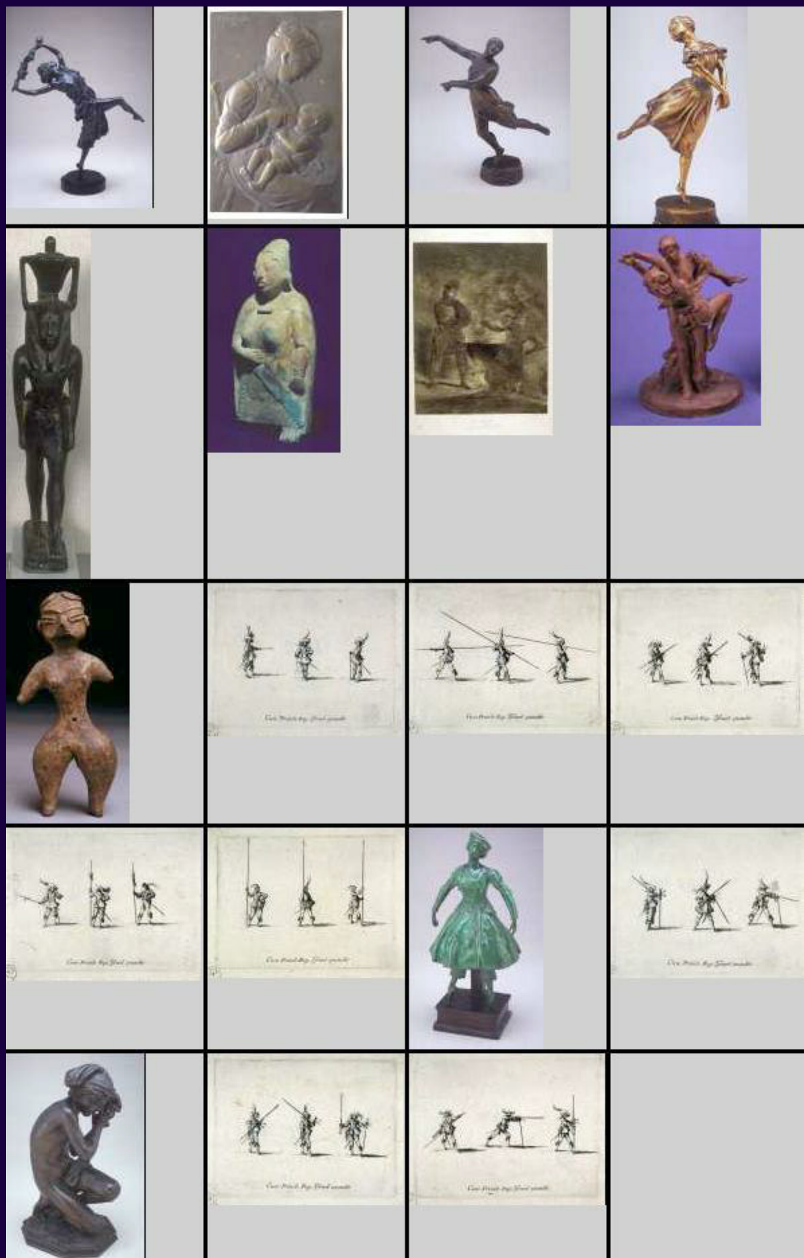
Hierarchical model

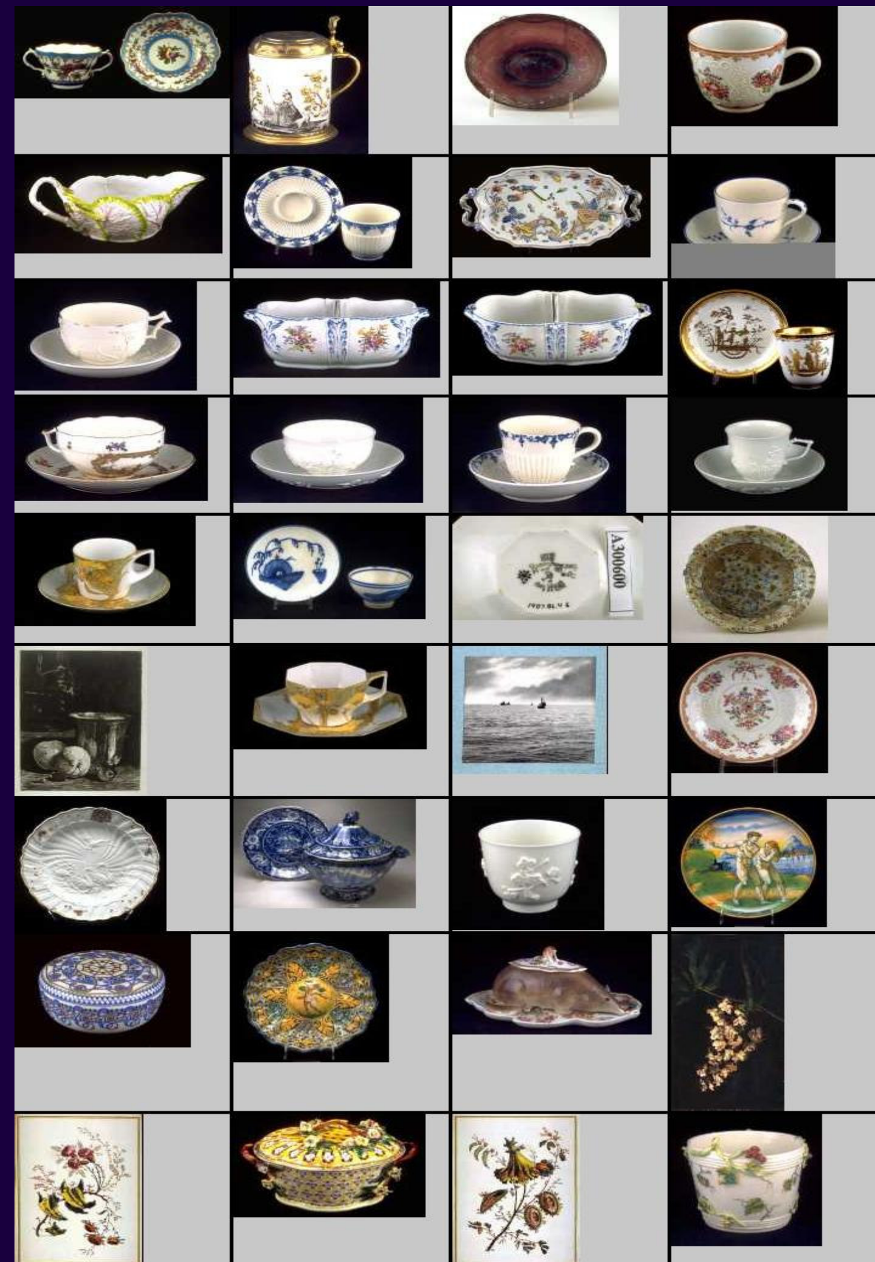


[Hofmann 98; Hofmann & Puzicha 98]







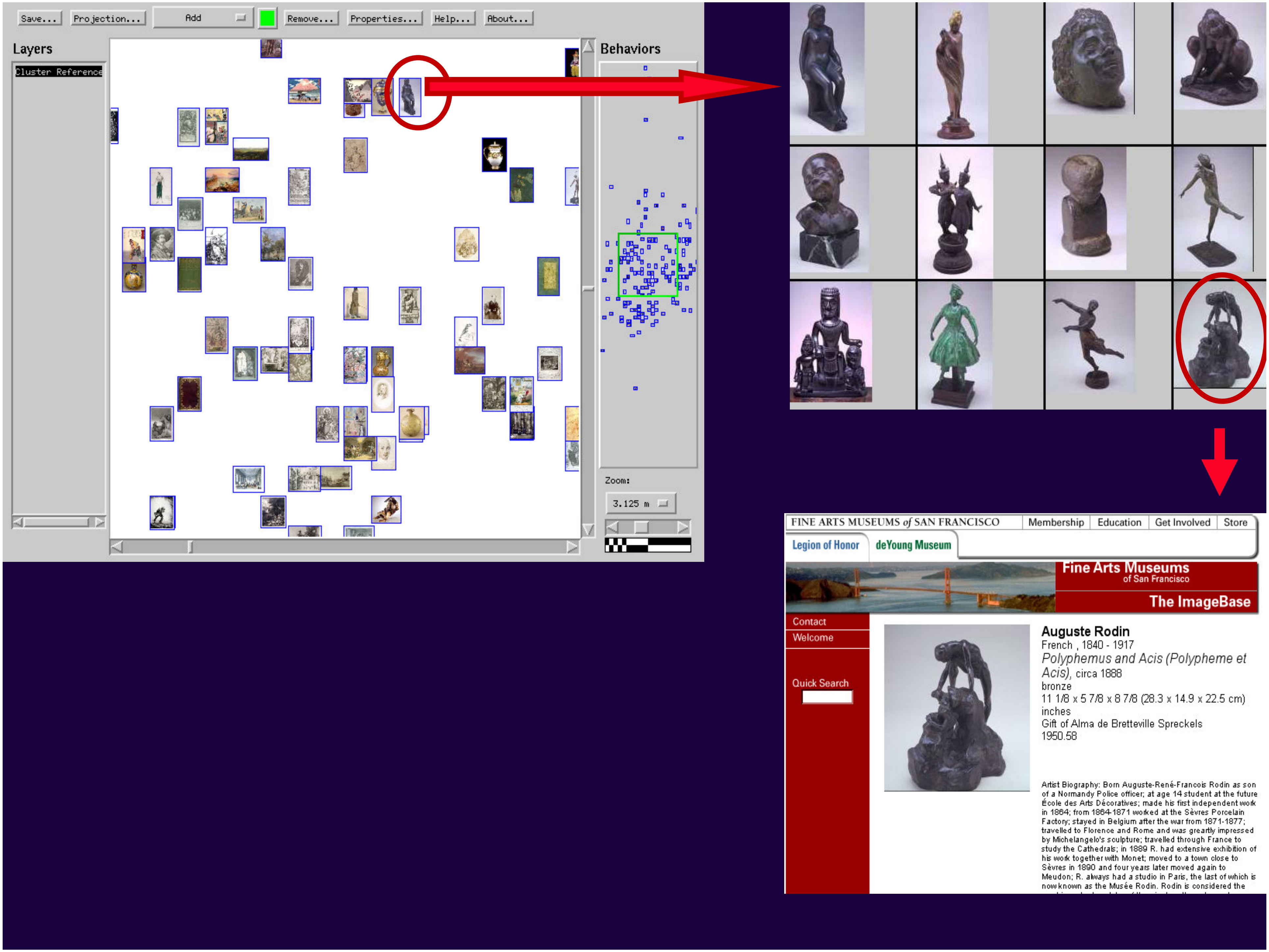

A300600

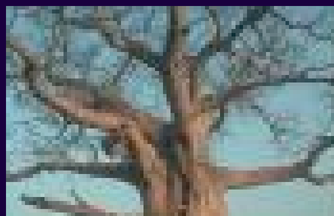
Browsing

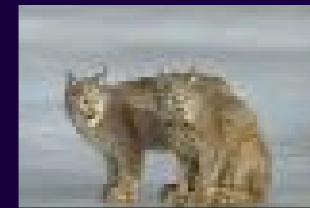
Browsing gives users an overall understanding of what is in a collection--a prerequisite for effective searching.

Need to organize images in a way that is relevant to humans

related studies---Sclaroff, Taycher, and La Cascia, 98; Rubner, Tomasi, and Guibas, 00; Smith Kanade, 97.


Save...
Projection...
Add
Remove...
Properties...
Help...
About...
Layers
Cluster Reference
Behaviors
Zoom:
3.125 m
FINE ARTS MUSEUMS of SAN FRANCISCO
Membership
Education
Get Involved
Store
Legion of Honor
deYoung Museum
Fine Arts Museums
of San Francisco
The ImageBase
Contact
Welcome
Quick Search
Auguste Rodin
French , 1840 - 1917
Polyphemus and Acis (Polypheme et Acis), circa 1888
bronze
11 1/8 x 5 7/8 x 8 7/8 (28.3 x 14.9 x 22.5 cm) inches
Gift of Alma de Bretteville Spreckels
1950.58
Artist Biography: Born Auguste-René-Francois Rodin as son of a Normandy Police officer; at age 14 student at the future École des Arts Décoratives; made his first independent work in 1864; from 1864-1871 worked at the Sèvres Porcelain Factory; stayed in Belgium after the war from 1871-1877; travelled to Florence and Rome and was greatly impressed by Michelangelo's sculpture; travelled through France to study the Cathedrals; in 1889 R. had extensive exhibition of his work together with Monet; moved to a town close to Sèvres in 1890 and four years later moved again to Meudon; R. always had a studio in Paris, the last of which is now known as the Musée Rodin. Rodin is considered the






The End