C280, Computer Vision

Prof. Trevor Darrell
trevor@eecs.berkeley.edu

Lecture 13: Topic Models for Recognition

Last Lecture

- Scanning window paradigm
- GIST
- HOG
- Boosted Face Detection
- Local-feature Alignment; from Roberts to Lowe...
- BOW Indexing

Next three lectures

- Today: learning object categories from the web
  - LSA and LDA models
  - Harvesting training data from the web
  - Exploiting image and text
- Tues. Oct. 20th: Generative models
  - Condensation
  - ISM
  - Transformed-HDPs
  - More Context…
- Thurs. Oct. 22nd: Advanced BOW kernels
  - Pyramid and spatial-pyramid match
  - Multi-kernel learning
  - Latent-part SVM models

Today: Topic Models for Recognition

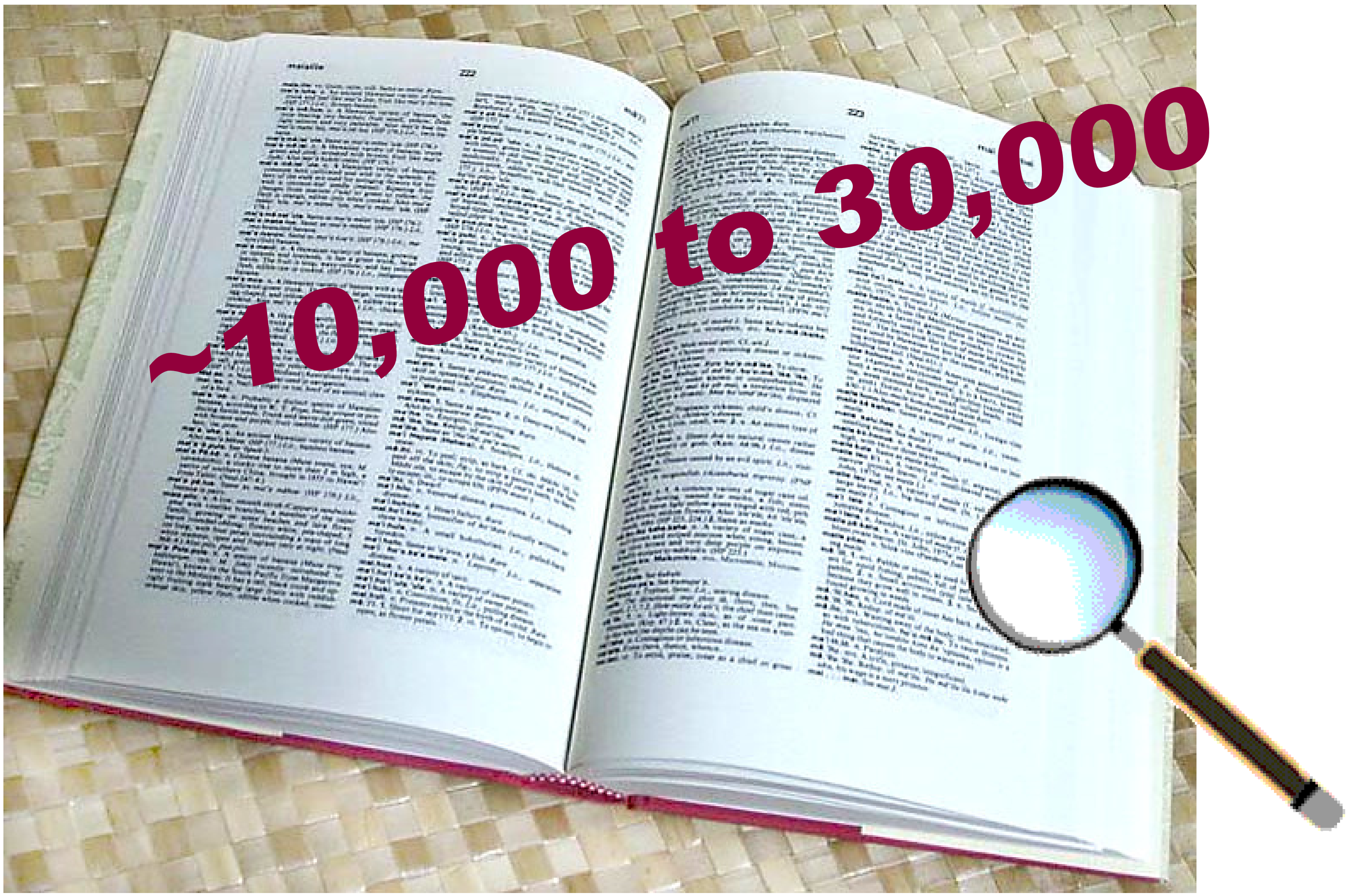
Guest lecture by Kate Saenko:

- Dataset issues
- Topic models for category discovery [Sivic05]
- Category discovery from web [Fergus05]
- Bootstrapping a category model [Li07]
- Using text in addition to image [Berg06]
- Learning objects from a dictionary [Saenko08]

Topic Models for Object Recognition

slide credit Fei-Fei et. al.

How many object categories are there?



slide credit Fei-Fei et. al.

Biederman 1987

Dataset Issues In Object Recognition

Datasets

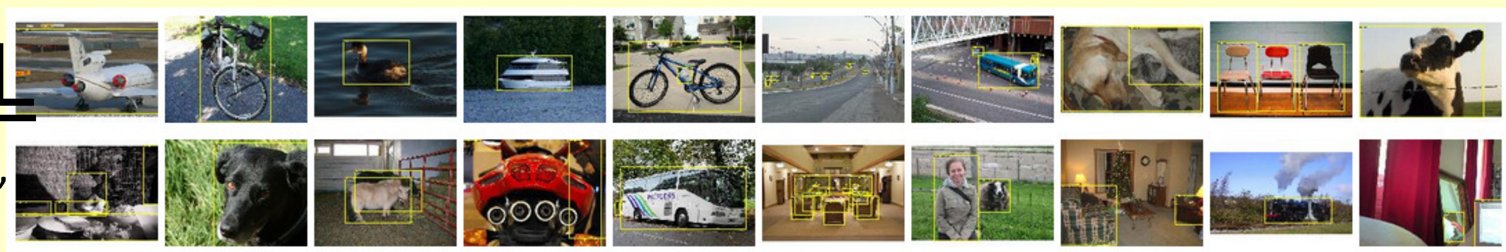
Caltech101/256

[Fei-Fei et al, 2004]

[Griffin et al, 2007]

PASCAL

[Everingham et al, 2009]



MSRC

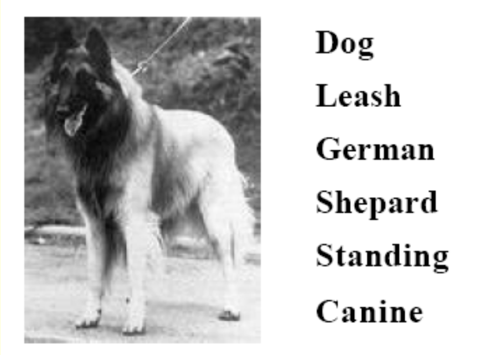
[Shotton et al. 2006]

slide credit Fei-Fei et. al.

Datasets

ESP

[Ahn et al, 2006]

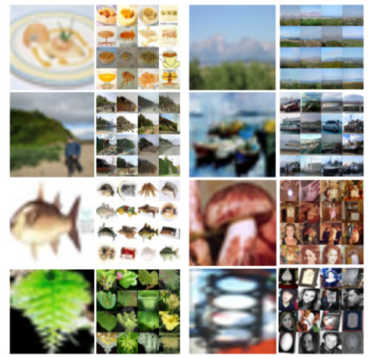


LabelMe

[Russell et al, 2005]

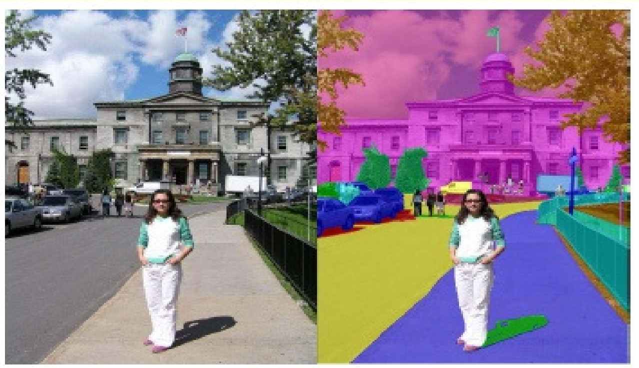
TinyImage

Torralba et al. 2007



Lotus Hill

[Yao et al, 2007]



slide credit Fei-Fei et. al.

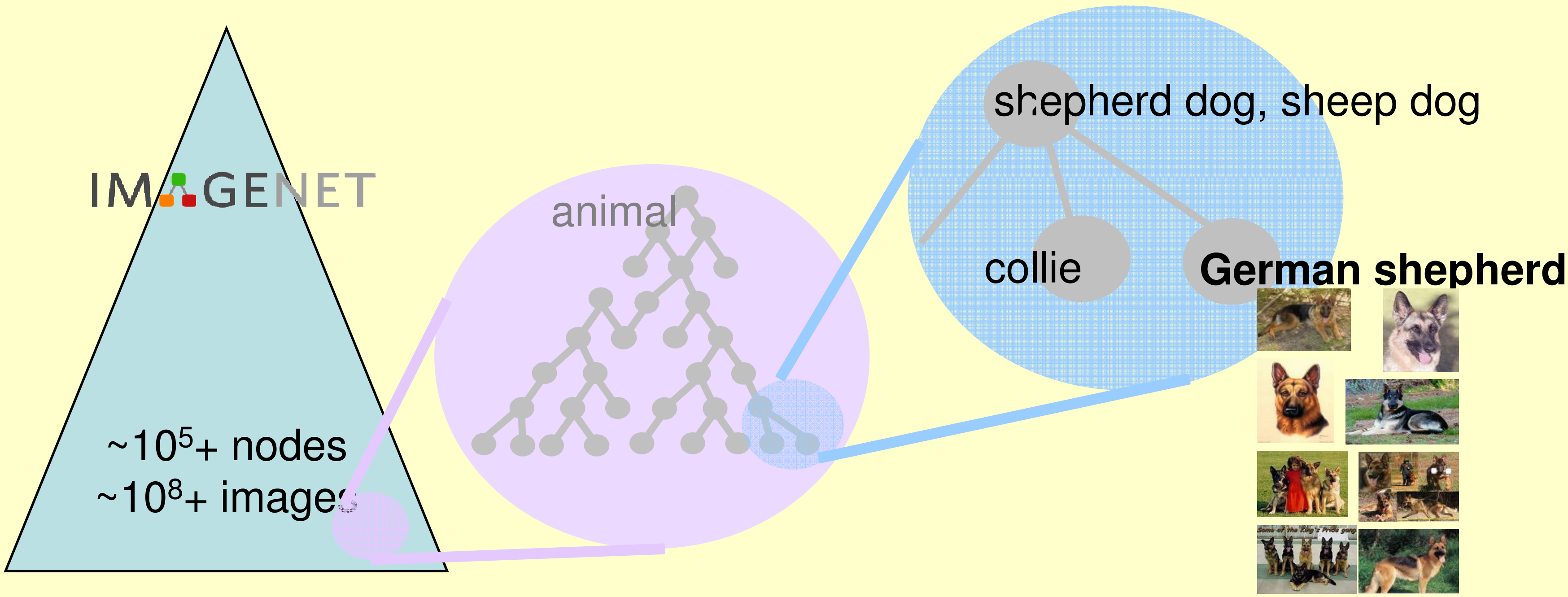
Size of existing datasets

Datasets	# of categories	# of images per category	# of total images	Collected by
Caltech101	101	~100	~10K	Human
Lotus Hill	~300	~ 500	~150K	Human
LabelMe	183	~200	~30K	Human
Ideal	~30K	>>10^2	A LOT	Machine

slide credit Fei-Fei et. al.

IMAGENET

- An ontology of images based on WordNet
- Collected using Amazon Mechanical Turk



slide credit Fei-Fei et. al.

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

IMAGENET

14,847 categories, 9,349,136 images

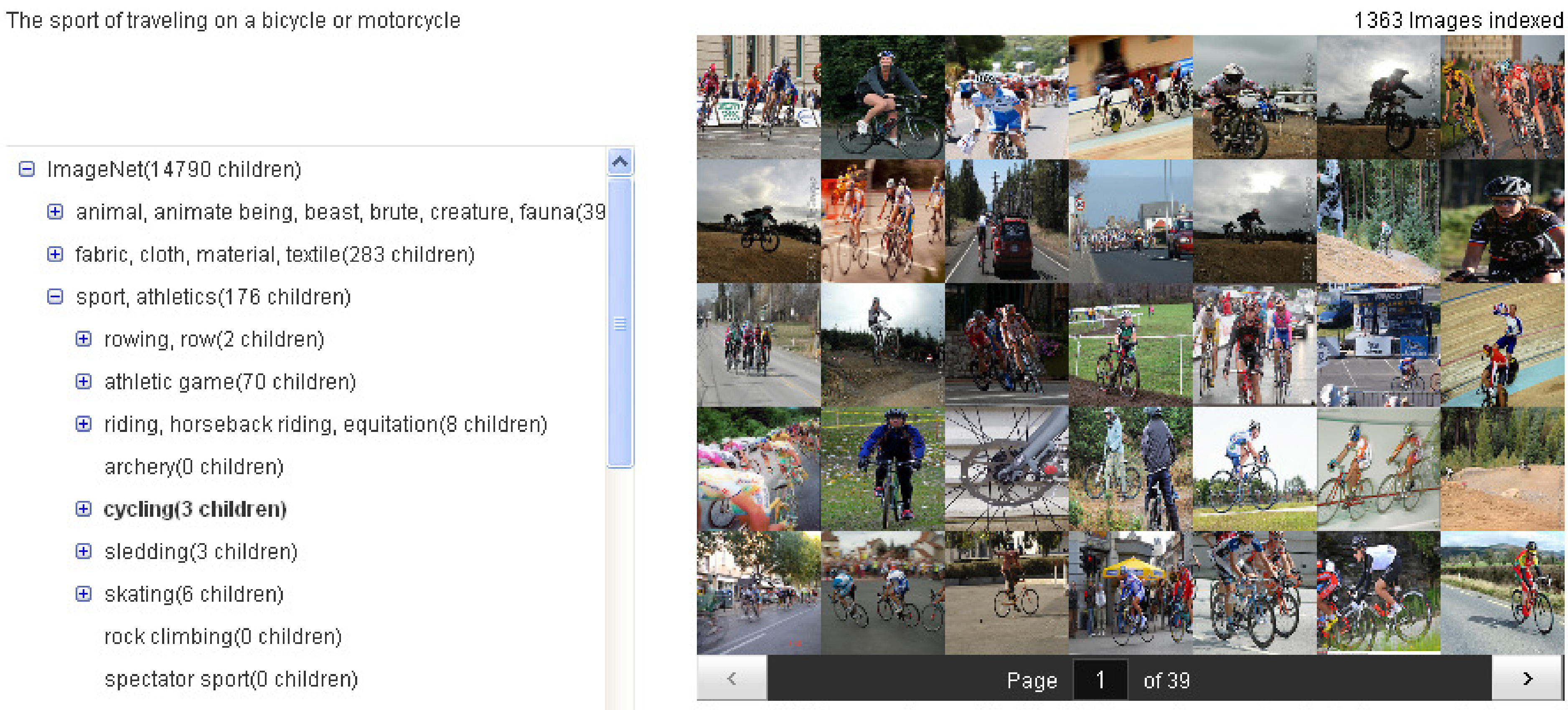
- Animals
  - Fish
  - Bird
  - Mammal
  - Invertebrate
- Scenes
  - Indoors
  - Geological formations
- Sport Activities
- Fabric Materials
- Instrumentation
  - Tool
  - Appliances
  - …
- Plants
  - …

slide credit Fei-Fei et. al.

Deng, Wei, Socher, Li, Li, Fei-Fei, CVPR 2009

IMAGENET

“Cycling”

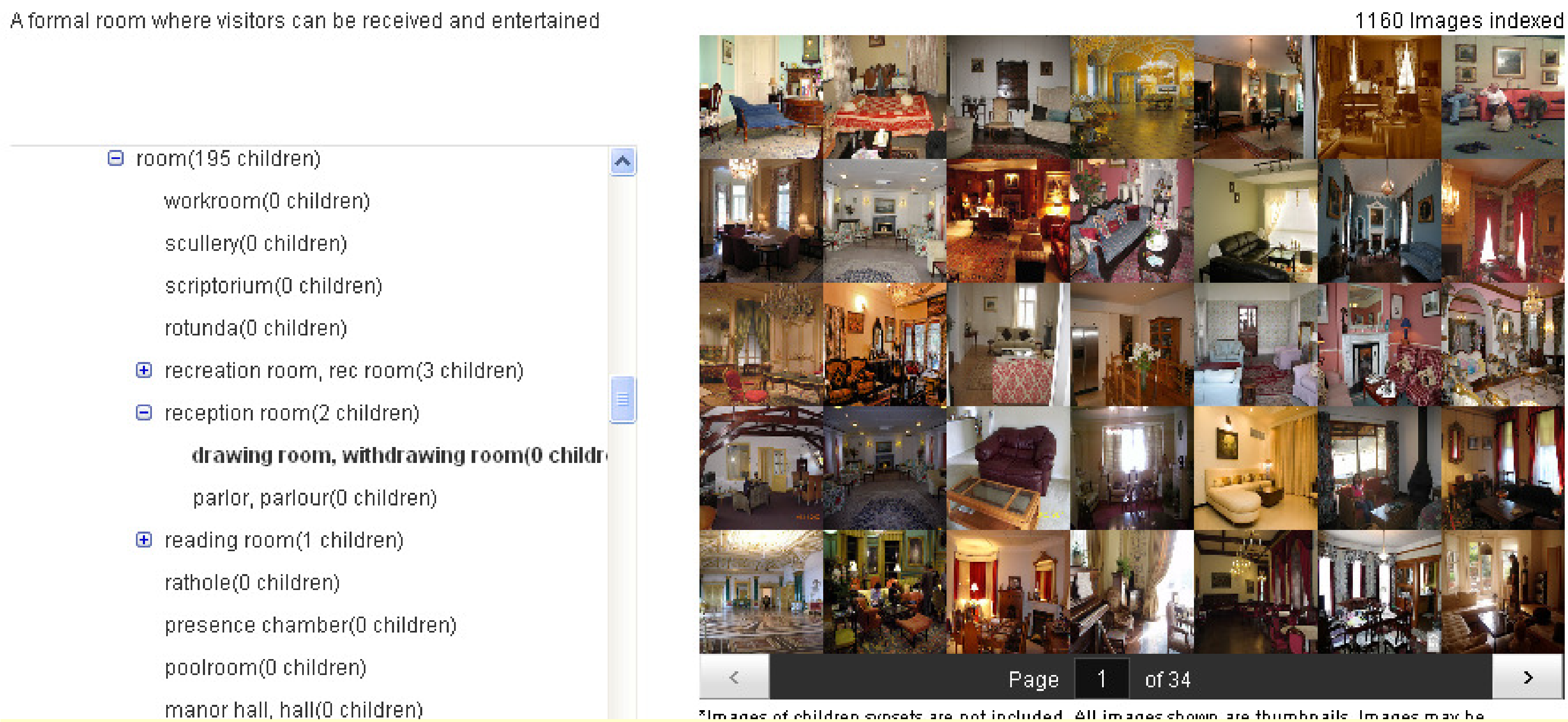


slide credit Fei-Fei et. al.

Deng, Wei, Socher, Li, Li, Fei-Fei, CVPR 2009

IMAGENET

"Drawing room, withdrawing room"



slide credit Fei-Fei et. al.

Deng, Wei, Socher, Li, Li, Fei-Fei, CVPR 2009

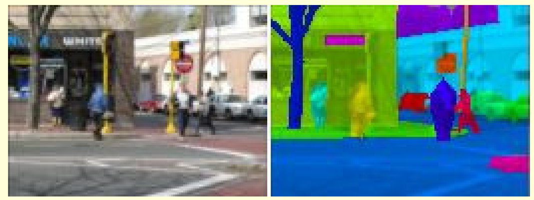
How much supervision do you need to learn models of objects?

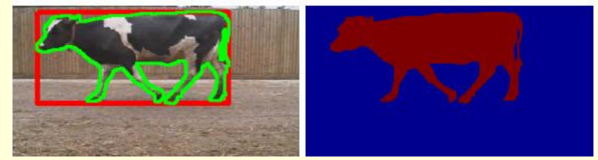
Object label + segmentation

MIT+CMU frontal faces

Viola & Jones ’01
Rowley et al. ’98

LabelMe, PASCAL, TU Darmstadt,
MIT scenes and objects





Agarwal & Roth ’02, Leibe &
Schiele ’03, Torralba et al. ’05

Object appears somewhere in the image

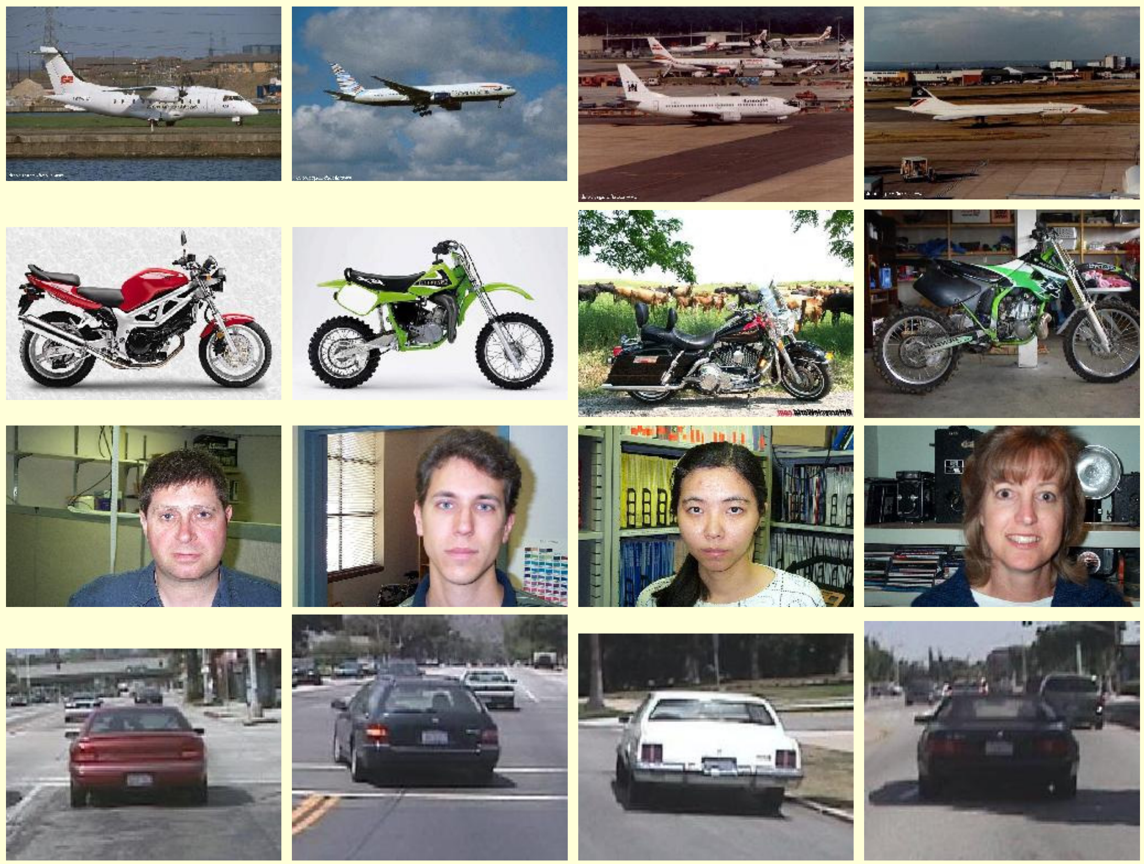
Caltech 101, PASCAL, MSRC

airplane

motorbike

face

car



Fergus et al. ’03, Csurka et al. ’04,
Dorko & Schmid ’05

Image + text caption

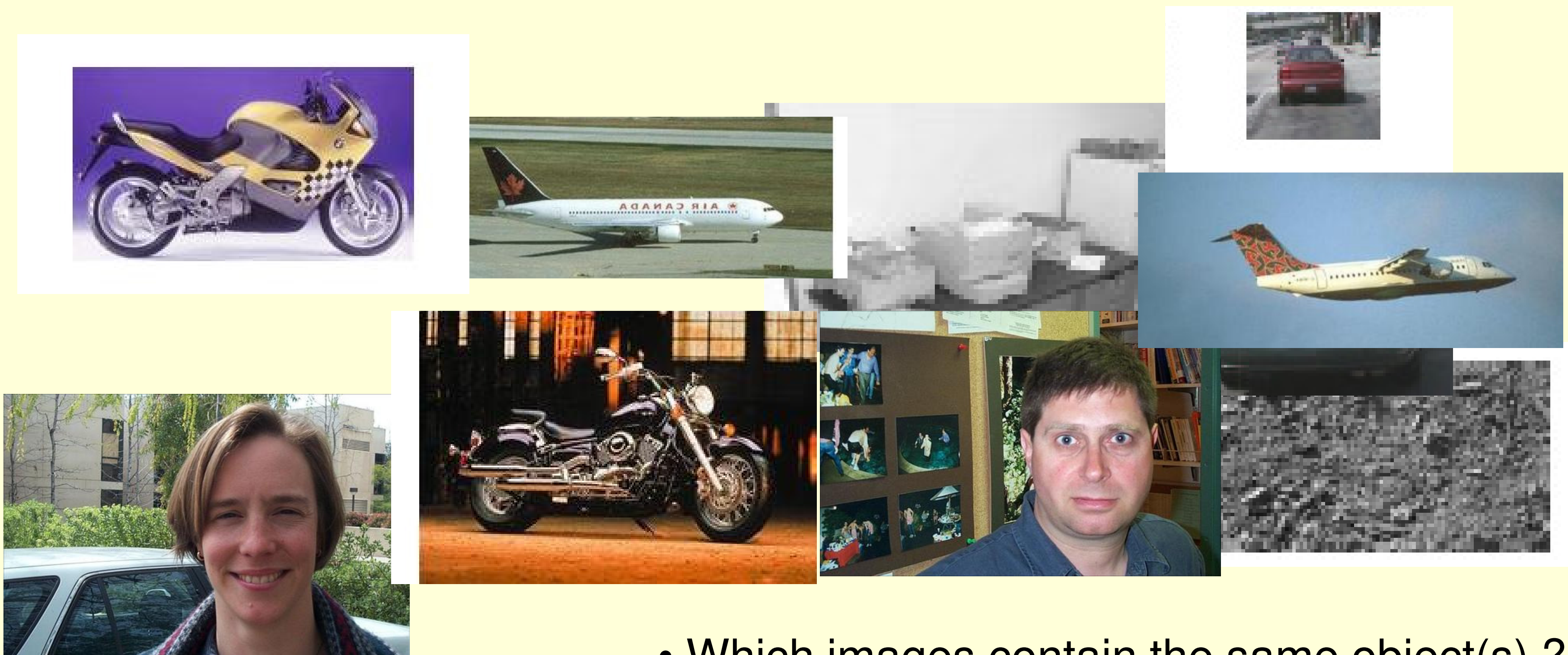
Corel, Flickr, Names+faces, ESP game

British director **Sam Mendes** and his partner actress **Kate Winslet** arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars **Tom Hanks** as a Chicago hit man who has a separate family life and co-stars **Paul Newman** and Jude Law. REUTERS/Dan Chung

Barnard et al. '03, Berg et al. '04

Images only

Given a collection of unlabeled images, discover visual object categories and their segmentation



- Which images contain the same object(s) ?
- Where is the object in the image?

Discovering Objects and Their Location in Images

J. Sivic, B. C. Russell, A. A. Efros,
A. Zisserman, W. T. Freeman.

Presented at the International Conference on Computer Vision, 2005.

Analogy: Discovering topics in text collections

Text document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Discovered topics

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

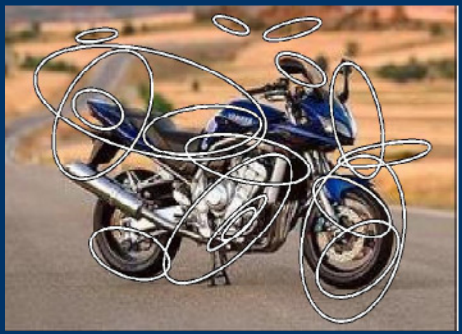
Blei, et al. 2003

Visual analogy

document	-	image
word	-	visual word
topics	-	objects

System overview

Input image

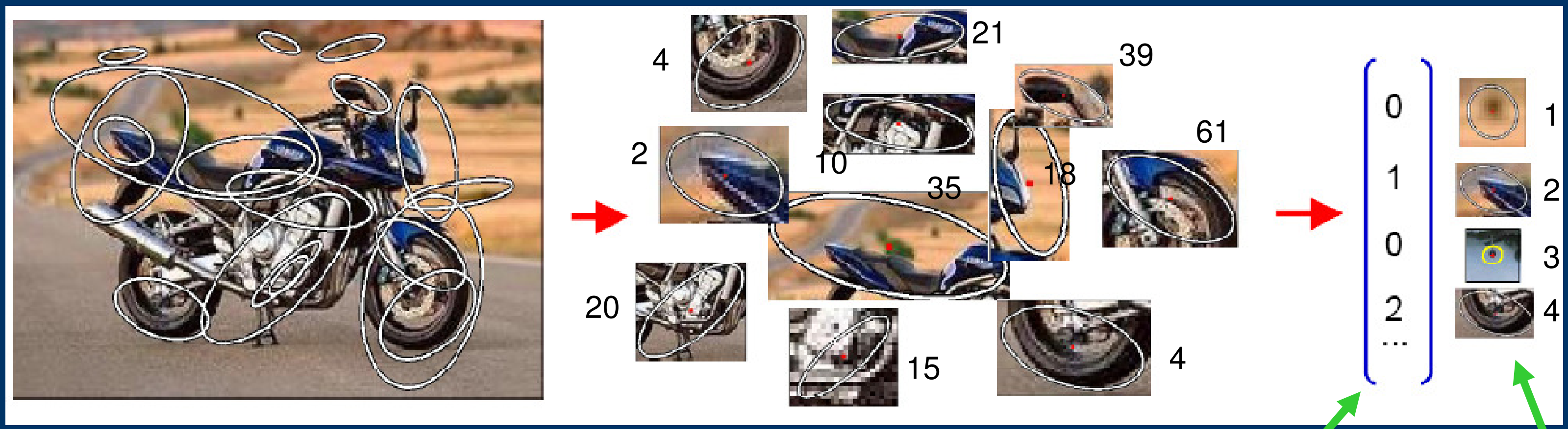


Compute visual words

Discover visual topics

Bag of words

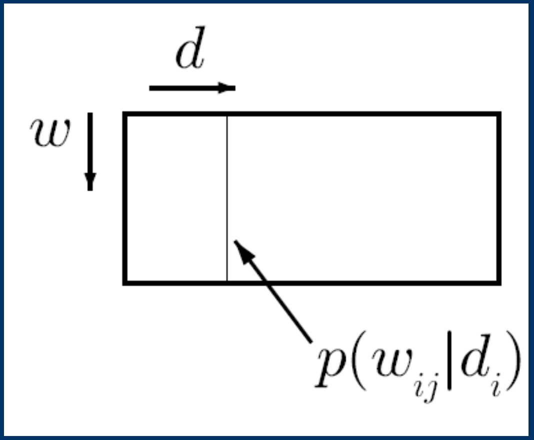
- LDA model assumes exchangeability
- Order of words does not matter



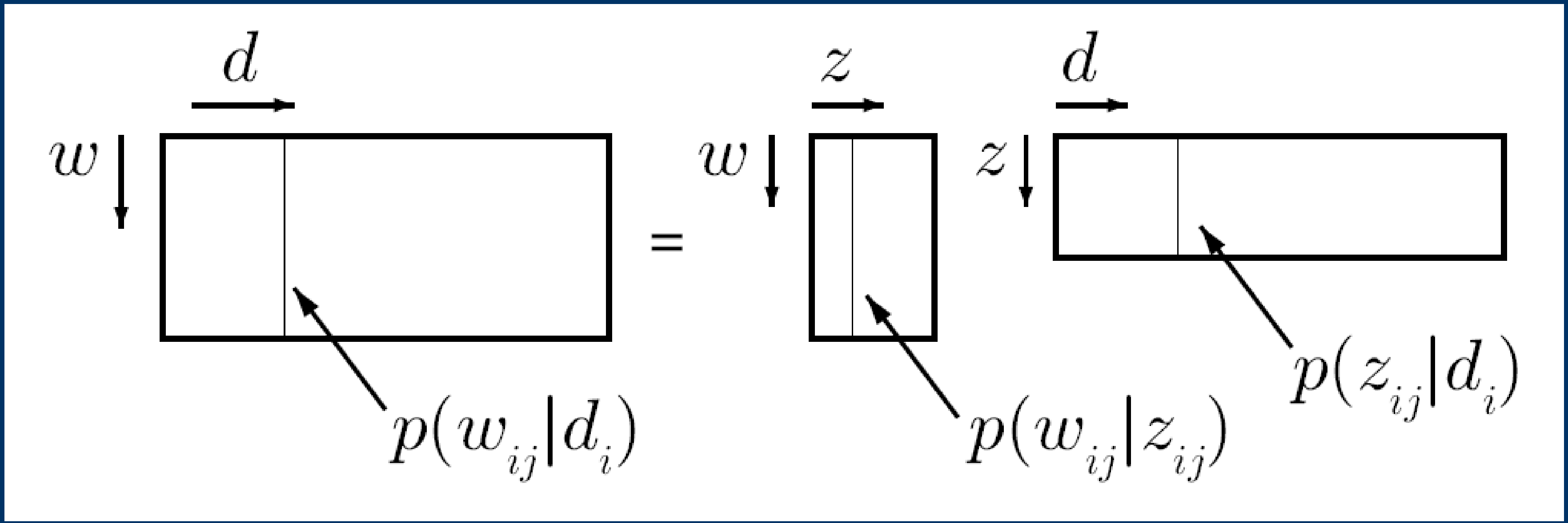
Interest regions Visual words Histogram Dictionary

Stack visual word histograms as columns in matrix

Throw away spatial information!



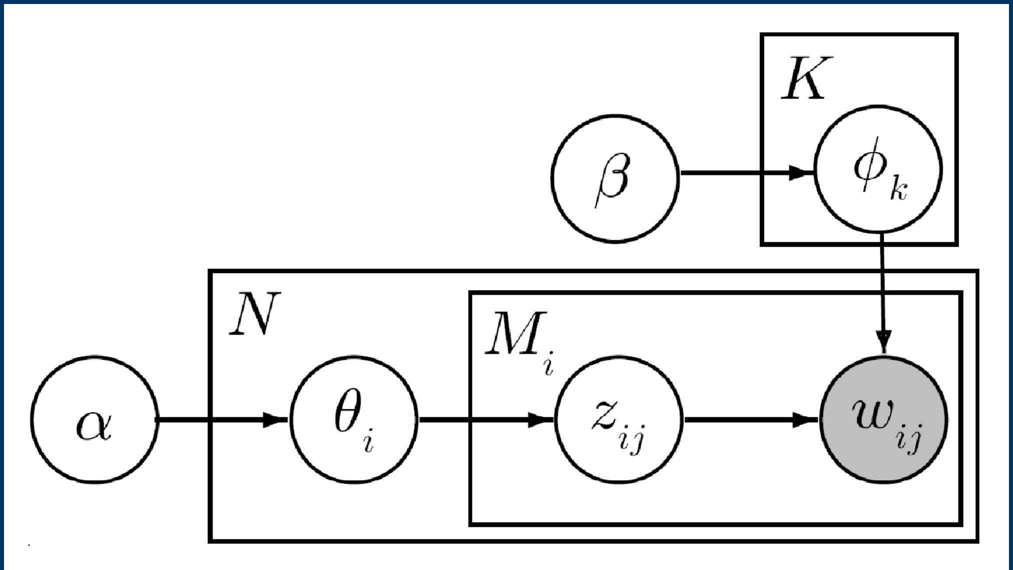
Low-rank matrix factorization



- Latent Semantic Analysis (Deerwester, et al. 1990)
- Probabilistic Latent Semantic Analysis (Hofmann 2001)

Latent Dirichlet Allocation (LDA)

Blei, et al. 2003



w_{ij} - words

z_{ij} - topic assignments

θ_i - topic mixing weights

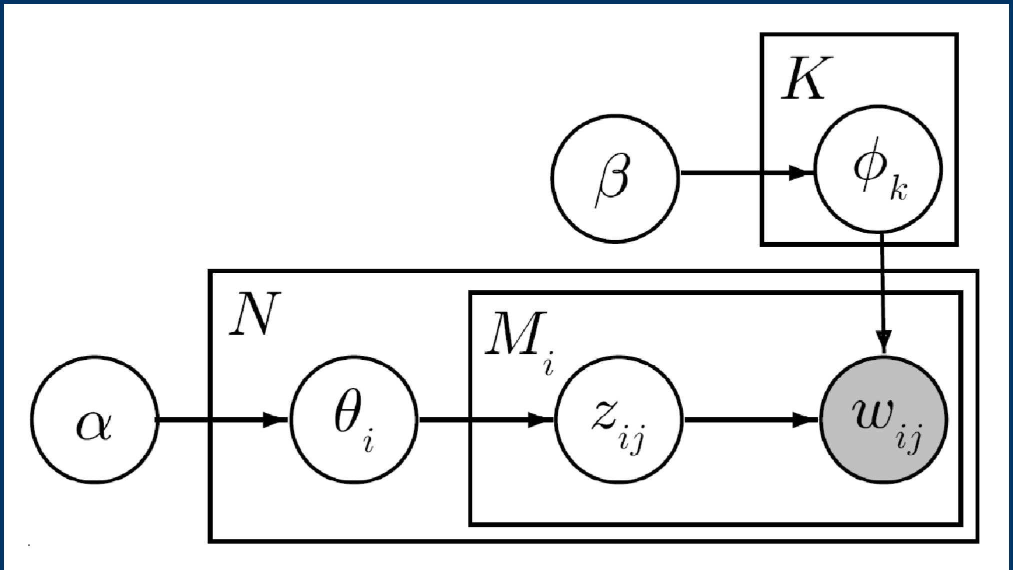
φ_k - word mixing weights

$$
\begin{aligned}
z_{ij}|\theta_i &\sim \theta_i \\
w_{ij}|z_{ij}=k,\phi &\sim \phi_k
\end{aligned}
$$

$$
\begin{aligned}
\theta_i|\alpha &\sim Dirichlet(\alpha) \\
\phi_k|\beta &\sim Dirichlet(\beta)
\end{aligned}
$$

Latent Dirichlet Allocation (LDA)

Blei, et al. 2003



w_{ij} - words

z_{ij} - topic assignments

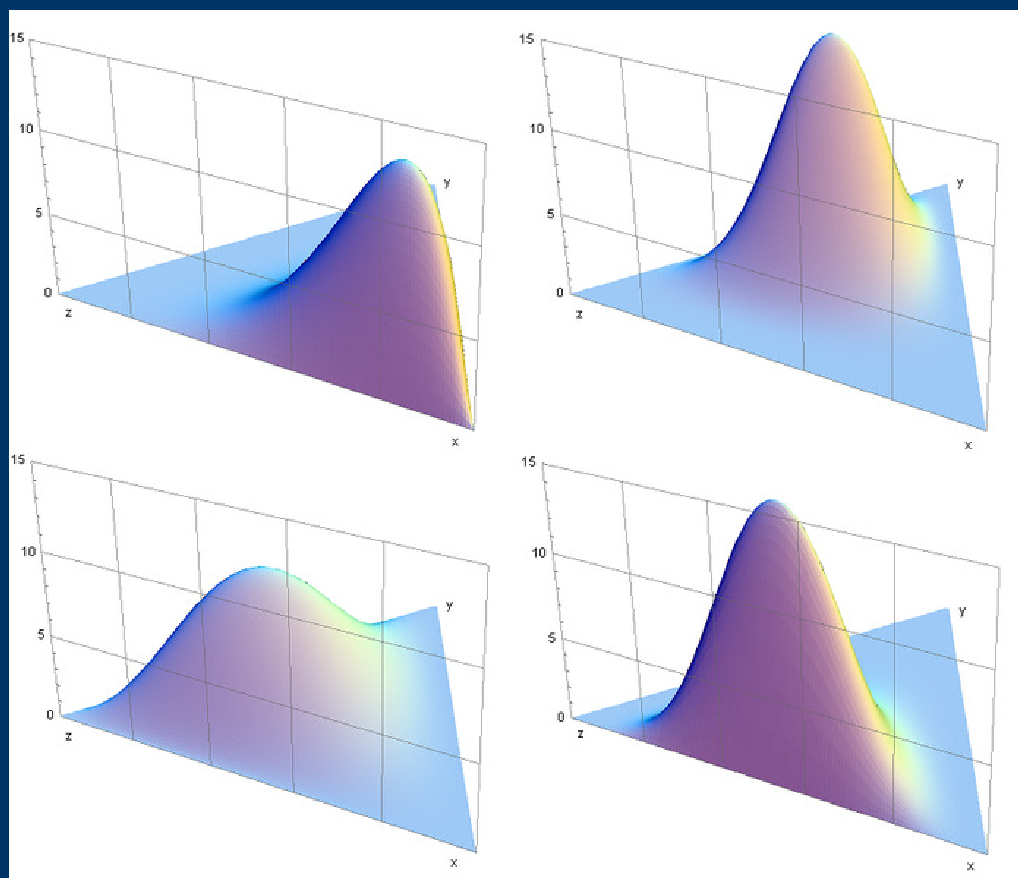
θ_i - topic mixing weights

φ_k - word mixing weights

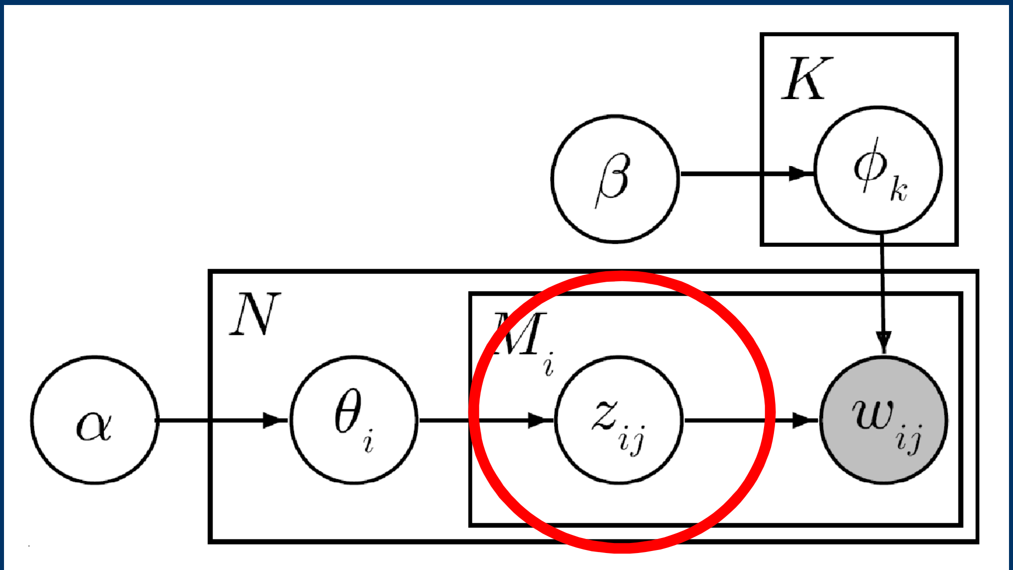
$$p(w_{ij}) \propto \sum_{k=1}^{K} p(w_{ij} | z_{ij} = k, \phi_k) \; p(z_{ij} = k | \theta_i)$$

Dirichlet Distribution

$$f(x_1, \dots, x_{K-1}; \alpha_1, \dots, \alpha_K) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}.$$



Inference



w_{ij} - words

z_{ij} - topic assignments

θ_i - topic mixing weights

φ_k - word mixing weights

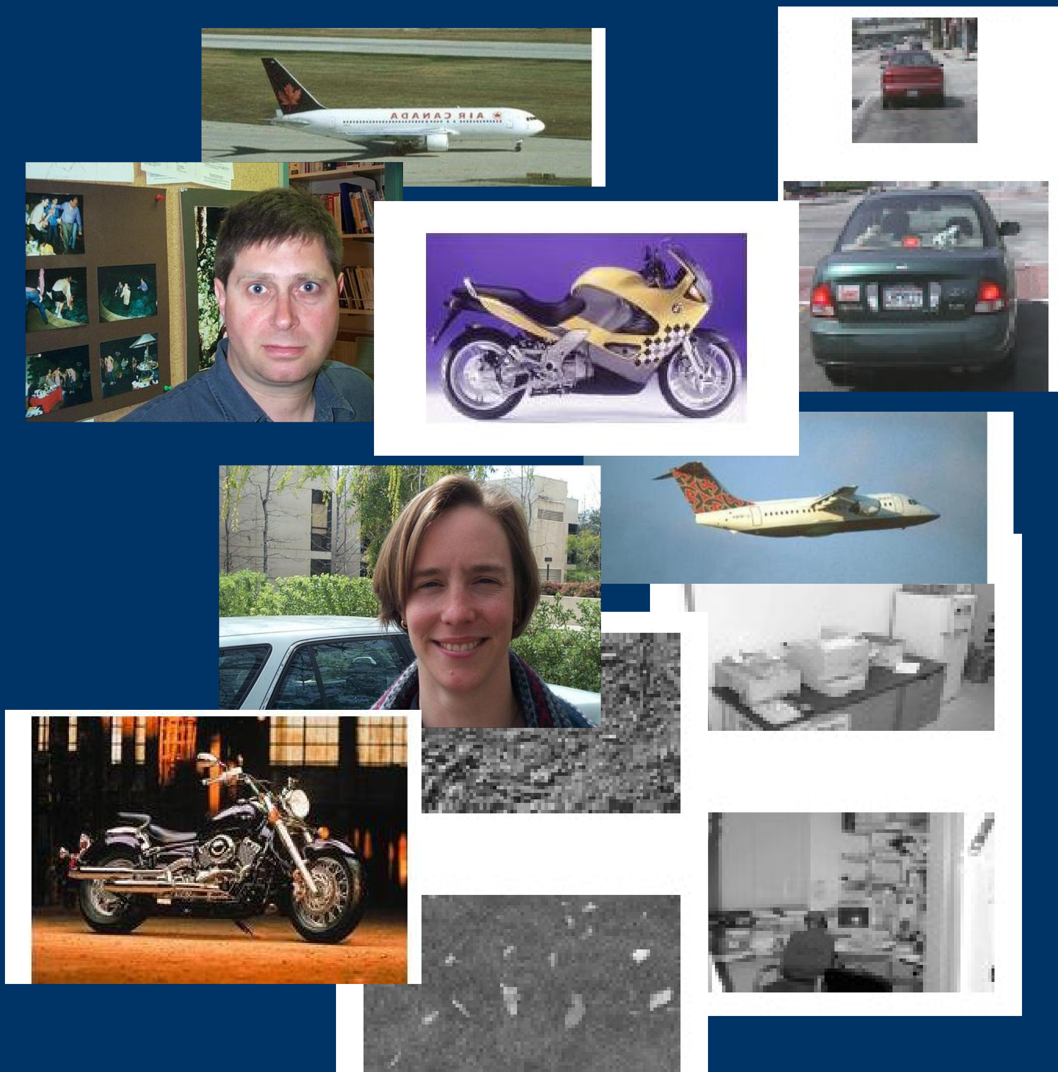
Use Gibbs sampler to sample topic assignments

[Griffiths & Steyvers 2004]

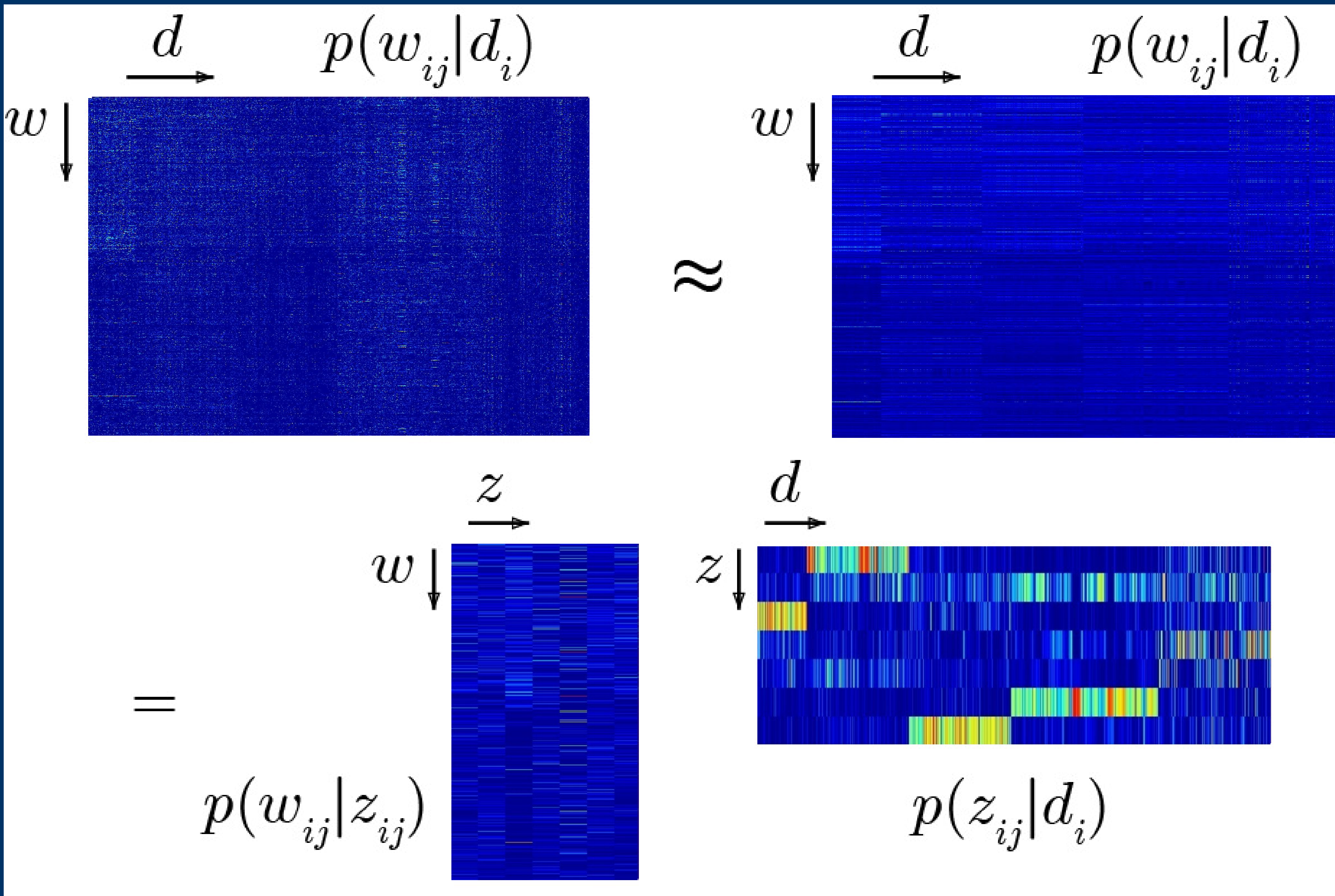
$$z_{ij} \sim p(z_{ij} = k | w_{ij} = v, w_{\setminus(ij)}, z_{\setminus(ij)}, \alpha, \beta)$$

- Only need to maintain counts of topic assignments
- Sampler typically converges in less than 50 iterations
- Run time is less than an hour

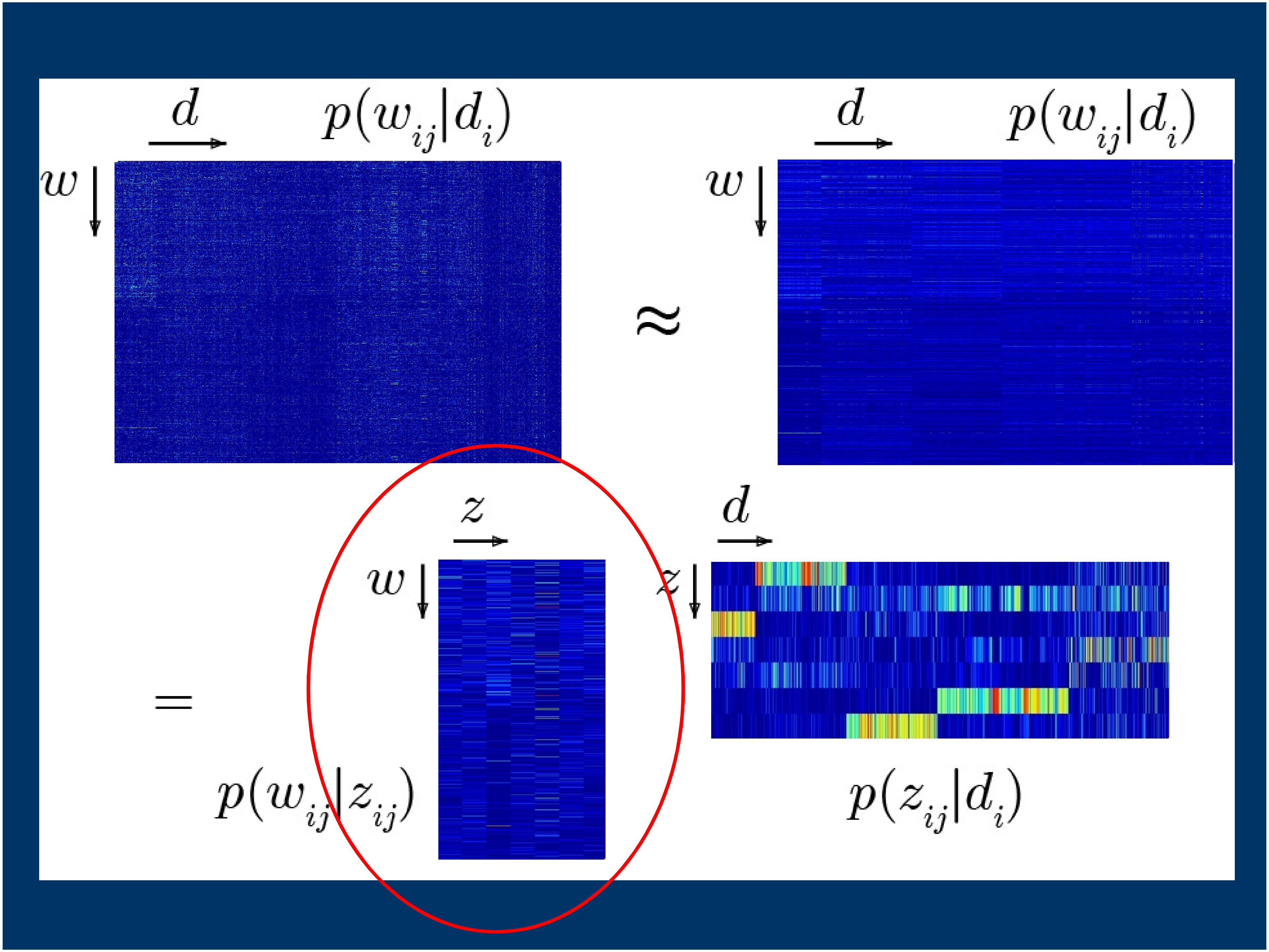
Apply to Caltech 4 + background images



Faces	435
Motorbikes	800
Airplanes	800
Cars (rear)	1155
Background	900
Total:	**4090**

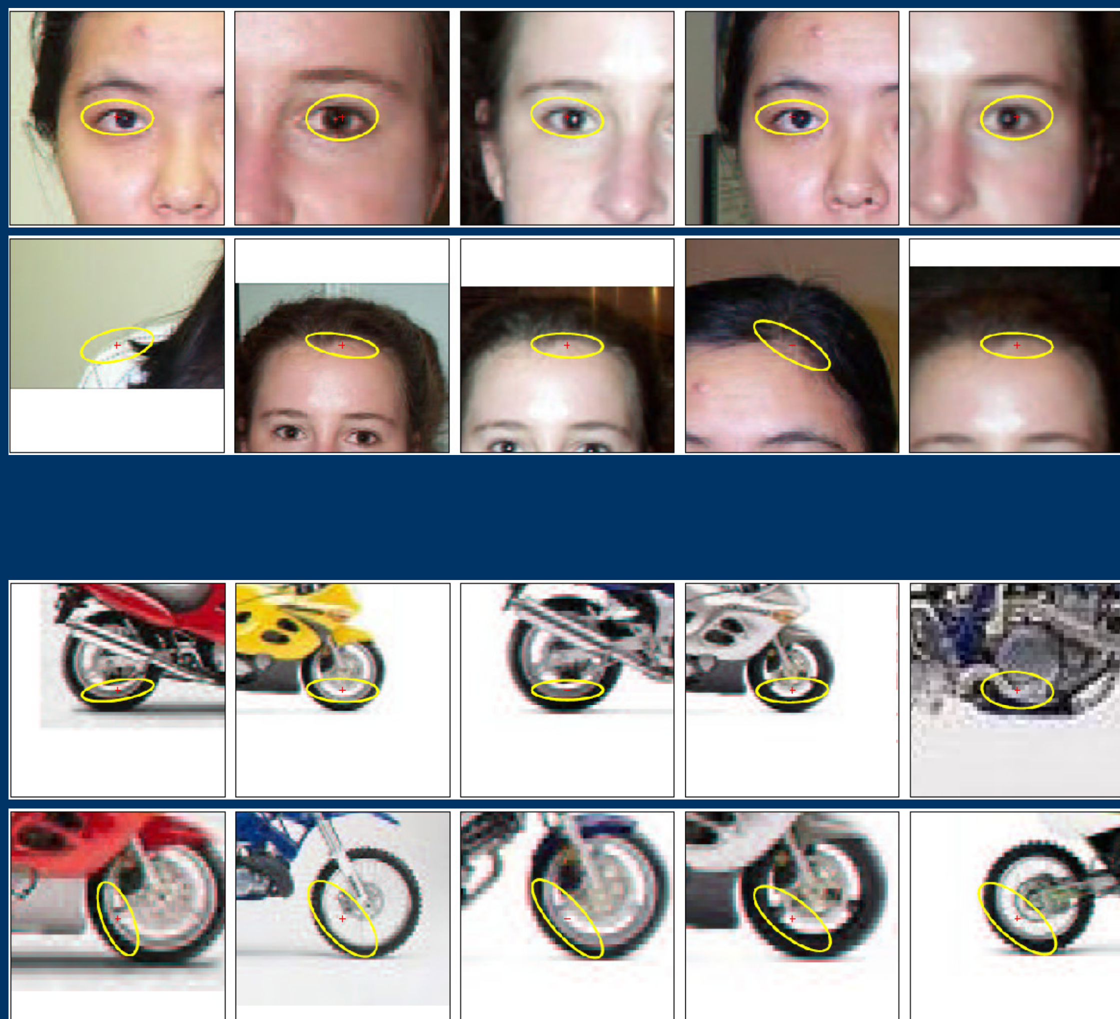


d
$p(w_{ij}|d_i)$
w
$\approx$
d
$p(w_{ij}|d_i)$
w
$=$
z
w
$p(w_{ij}|z_{ij})$
d
z
$p(z_{ij}|d_i)$



d
$p(w_{ij}|d_i)$
w
$\approx$
d
$p(w_{ij}|d_i)$
w
$=$
z
w
$p(w_{ij}|z_{ij})$
d
z
$p(z_{ij}|d_i)$

Most likely words given topic



Topic 1

Word 1

Word 2

Topic 2

Word 1

Word 2

Most likely words given topic

Topic 3

Word 1

Word 2

Topic 4

Word 1

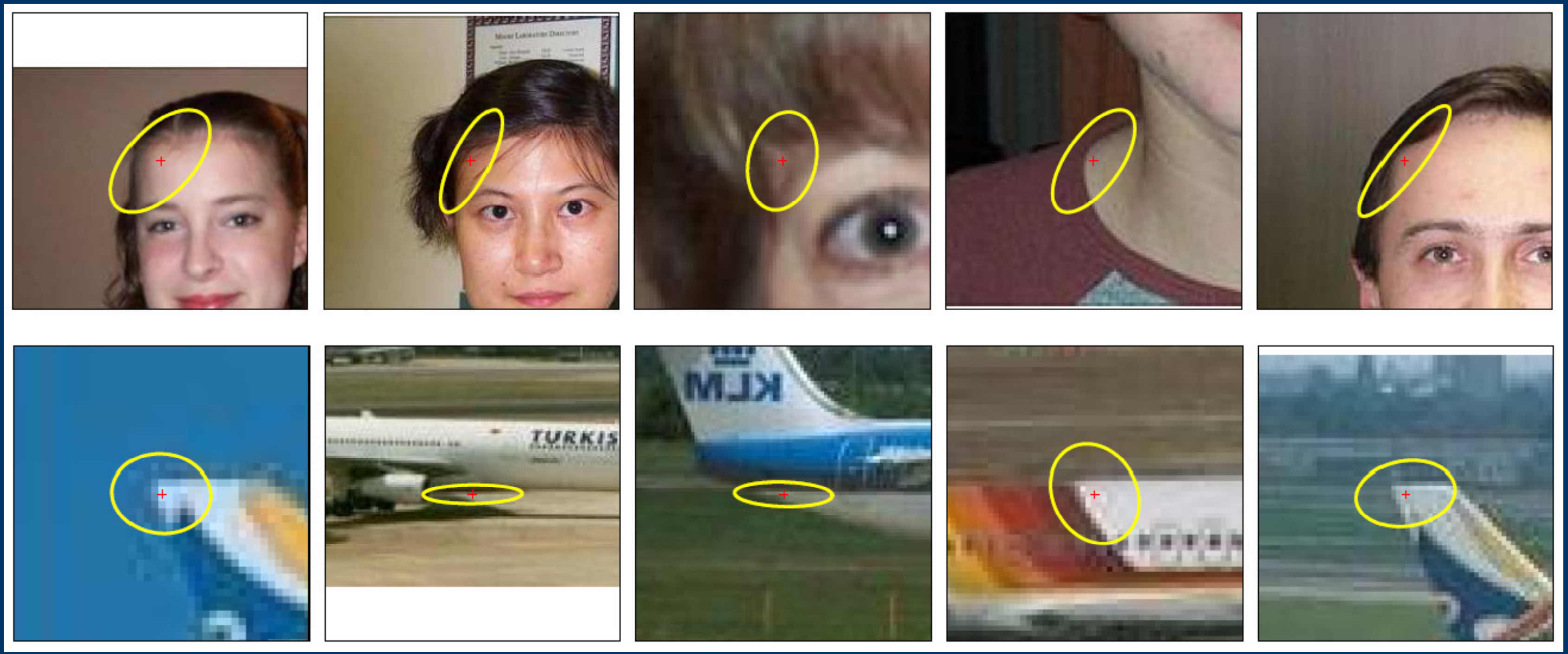
Word 2

Polysemy

In English, "bank" refers to:

1. a institution that handles money
2. the side of a river

Regions that map to the same visual word:



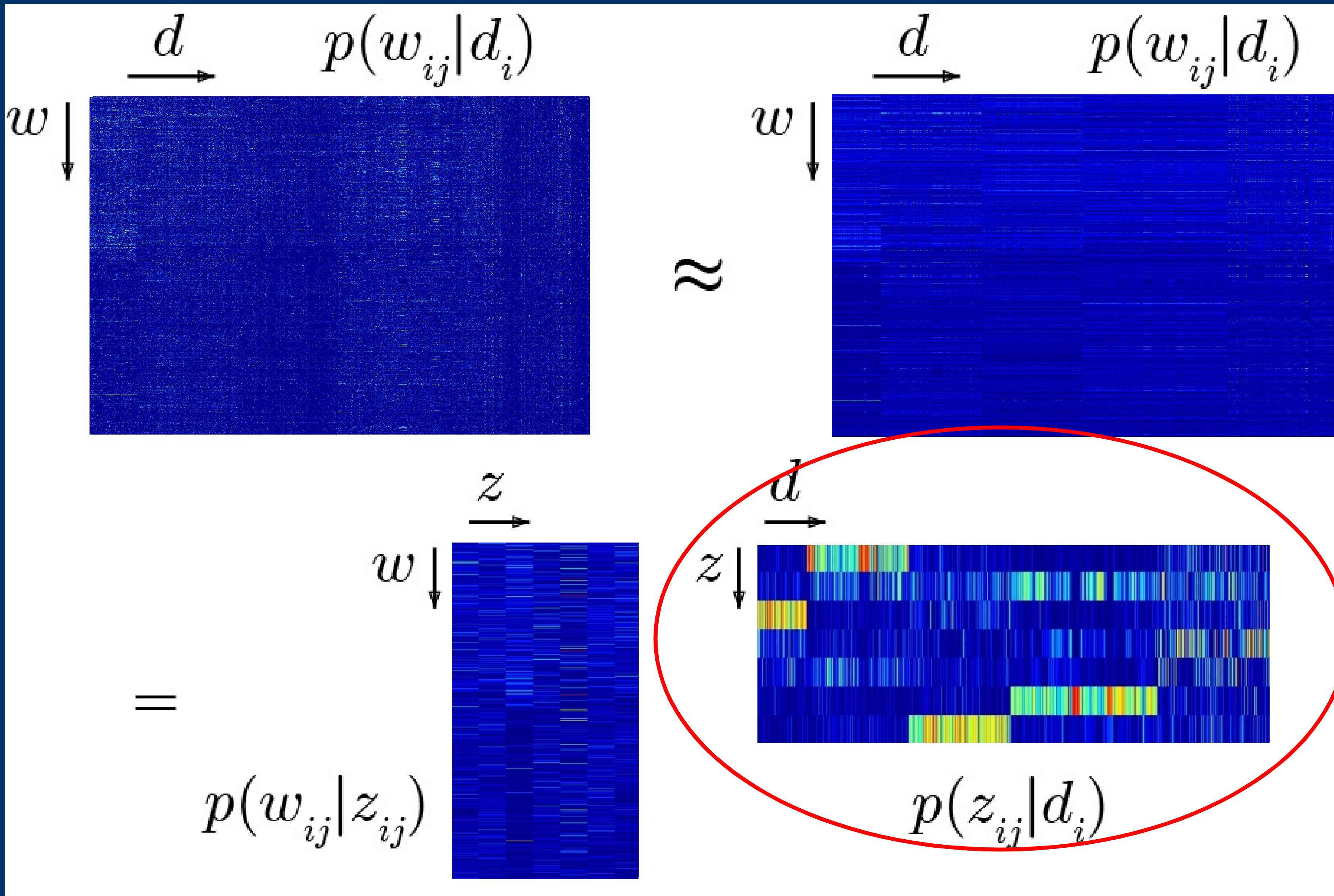
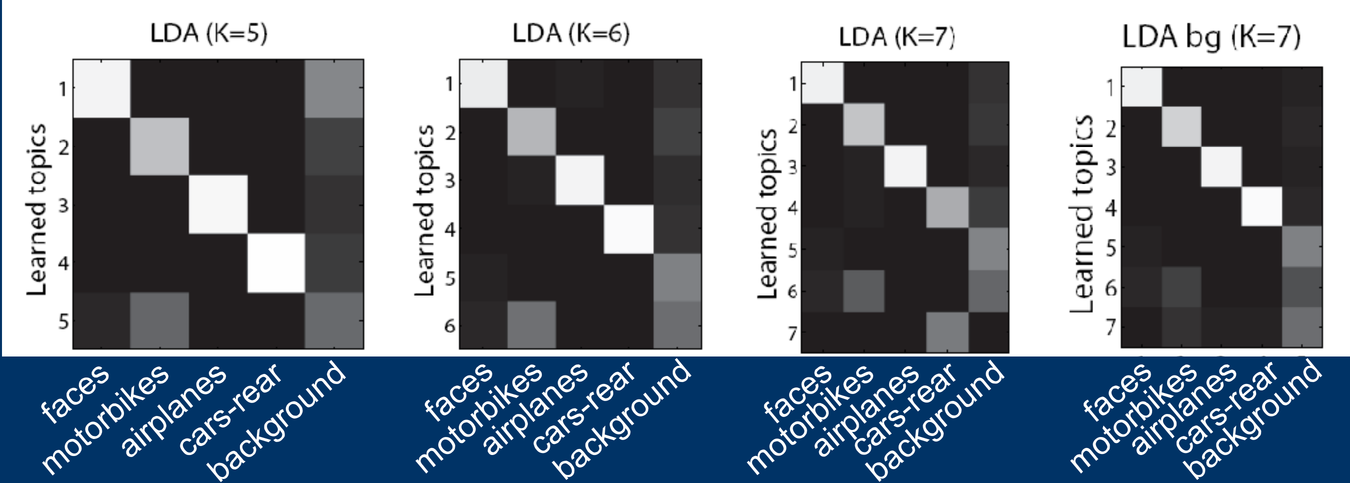

d
$p(w_{ij}|d_i)$
w
$\approx$
d
$p(w_{ij}|d_i)$
w
$=$
z
w
$p(w_{ij}|z_{ij})$
d
z
$p(z_{ij}|d_i)$

Image clustering

Confusion matrices:



Average confusion:

Expt.	Categories	T	LDA		pLSA		KM baseline	
			%	#	%	#	%	#
(1)	4	4	97	86	98	70	72	908
(2)	4 + bg	5	78	931	78	931	56	1820
(2)*	4 + bg	6	84	656	76	1072	–	–
(2)*	4 + bg	7	78	1007	83	768	–	–
(2)*	4 + bg-fxd	7	90	330	93	238	–	–

Comparison with supervised model

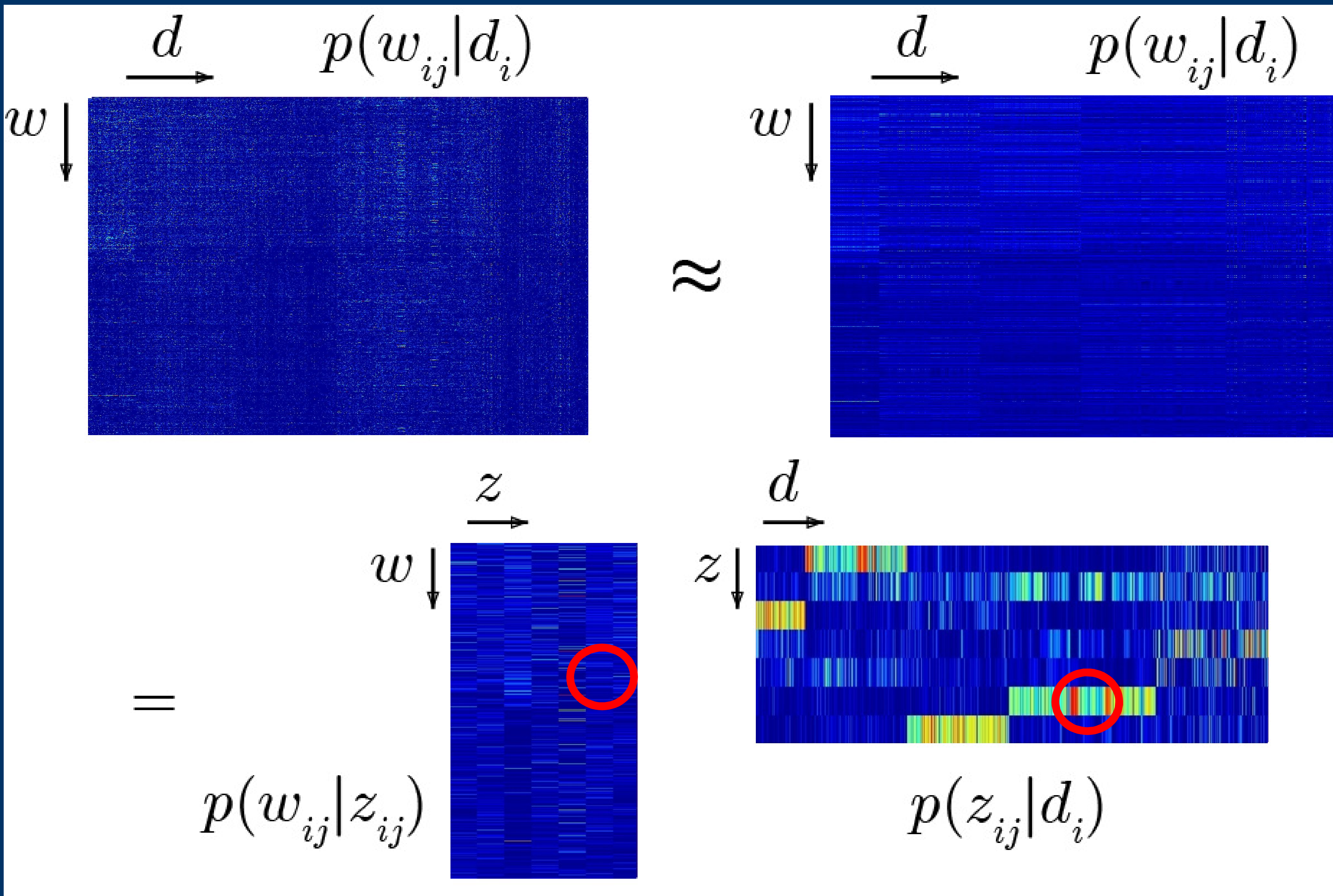
Percent ROC equal error rate

	LDA	Constellation model [Fergus et al. '03]
Faces	7.8	3.6
Motorbikes	9.9	6.7
Airplanes	2.5	7.0
Cars rear	8.5	9.7

- Comparable performance to constellation model
- Level of supervision:

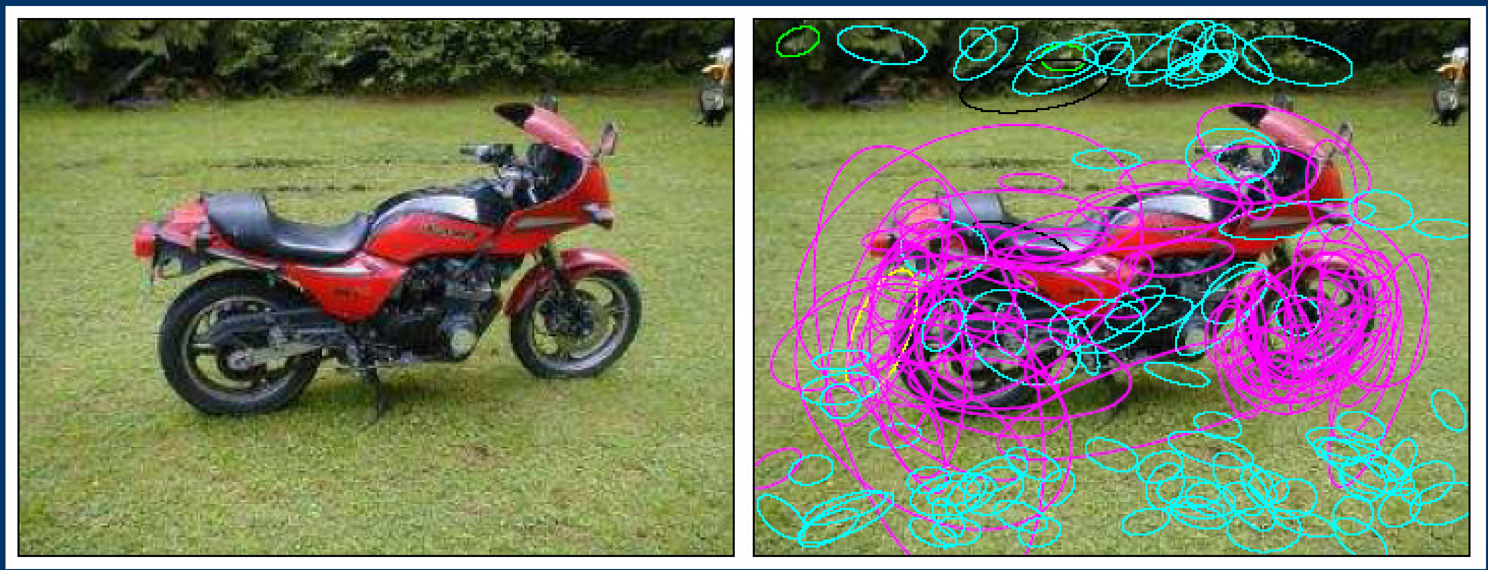
  LDA: one number (of topics)

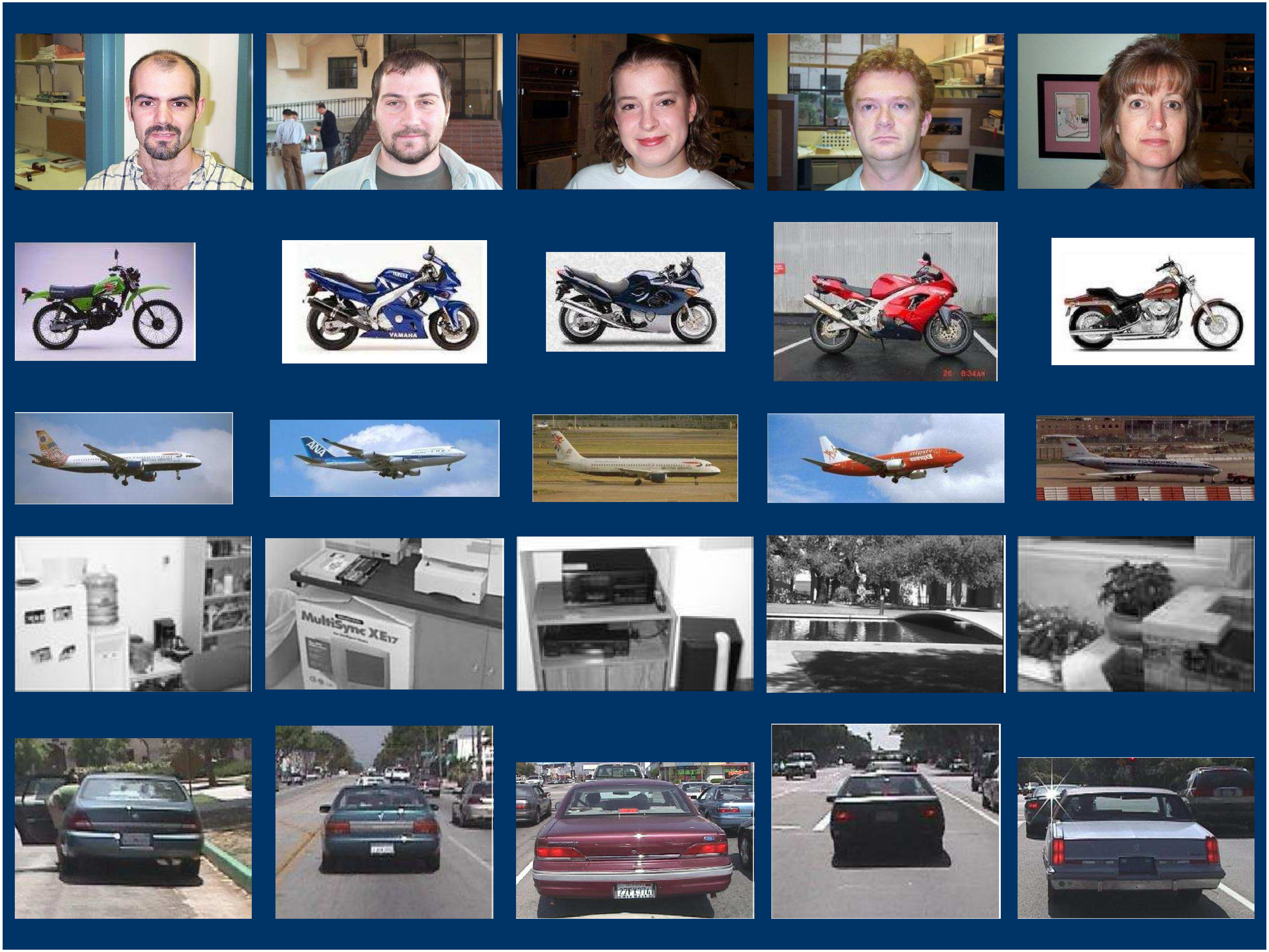
  Constellation model: 400 labels for each category
- Also an indication of the level of difficulty of the Caltech 5 dataset



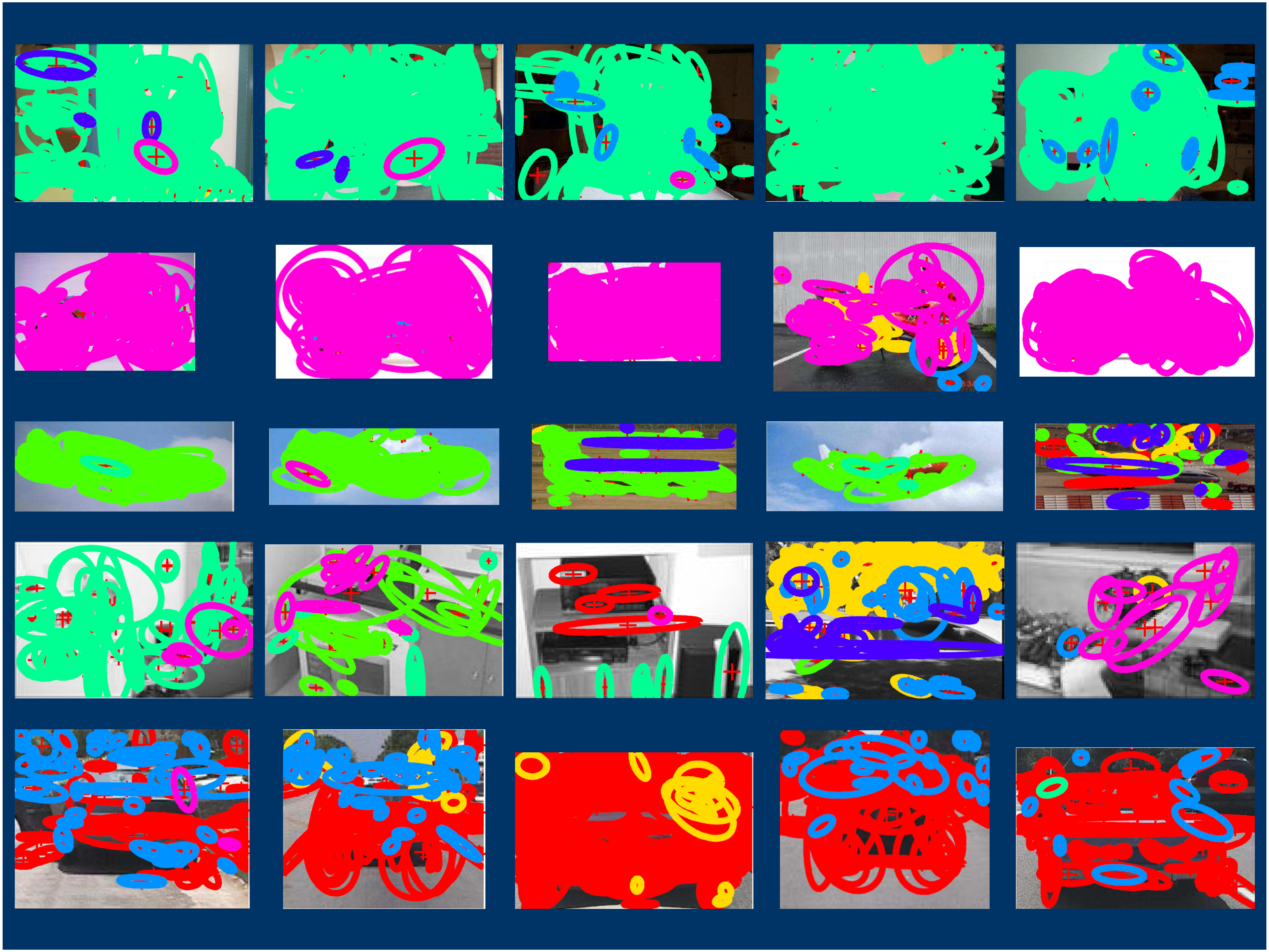
d
$p(w_{ij}|d_i)$
w
$\approx$
d
$p(w_{ij}|d_i)$
w
$=$
z
w
$p(w_{ij}|z_{ij})$
d
z
$p(z_{ij}|d_i)$

Image as a mixture of topics (objects)





YAMAHA
ANA
MultiSync XE17



Summary -- Sivic

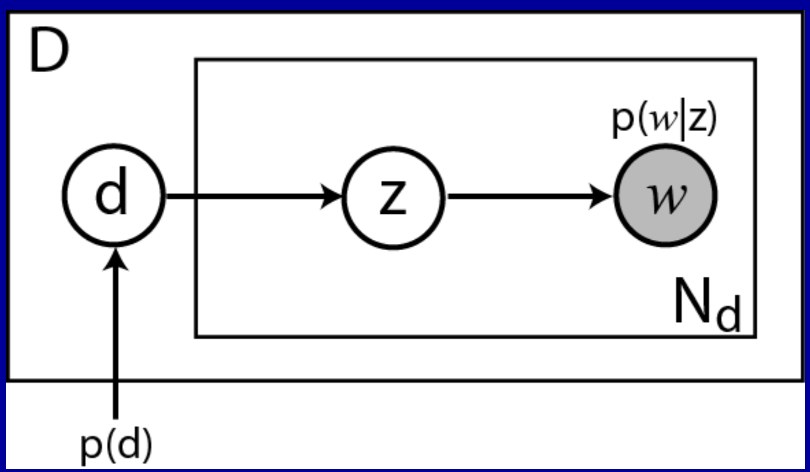
- Discovered visual topics corresponding to object categories from a corpus of unlabeled images
- Used visual words representation and topic discovery models from the text understanding community
- Classification on unseen images is comparable to supervised methods on Caltech 5 dataset
- The discovered categories can be localized within an image

Learning Object Categories from contaminated data

Rob Fergus
Li Fei-Fei
Pietro Perona
Andrew Zisserman

Probabilistic Latent Semantic Analysis (pLSA)

- Introduced by Hofmann in text analysis field
- Latent Dirichlet Allocation (LDA) – Blei and Jordan
- Adapted to visual data by:
  - Sivic, Russell et al. (Unsupervised object category discovery) *ICCV '05*
  - Fei-Fei and Perona (Scene analysis using LDA) *CVPR '05*

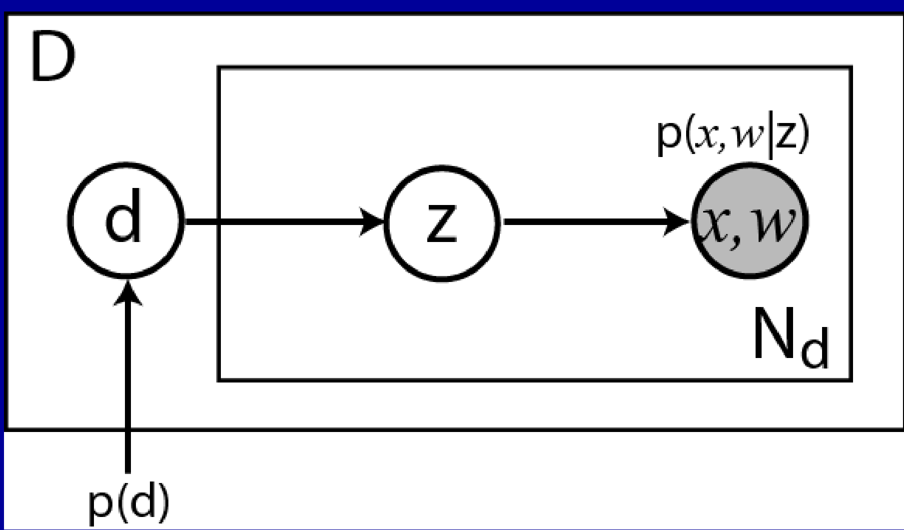


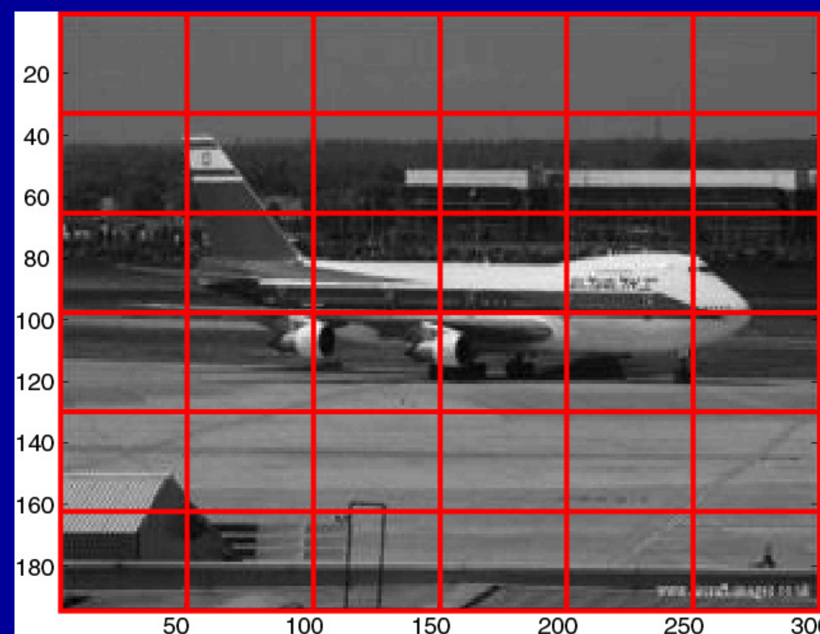
- Need to choose # topics (Z)

	Text domain	Image domain
d	Document	Image
z	Topic	Object
w	Word	VQ'd appearance of region

1. Improvements on pLSA: ABS-pLSA

- No spatial information used in pLSA
- Simplest form of spatial model:
- Joint spatial/word model
  - Quantize location of region within image
  - Absolute coordinate frame

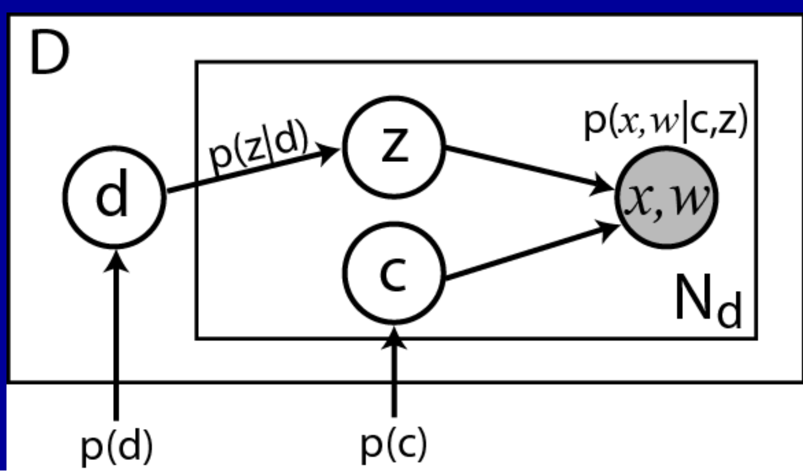


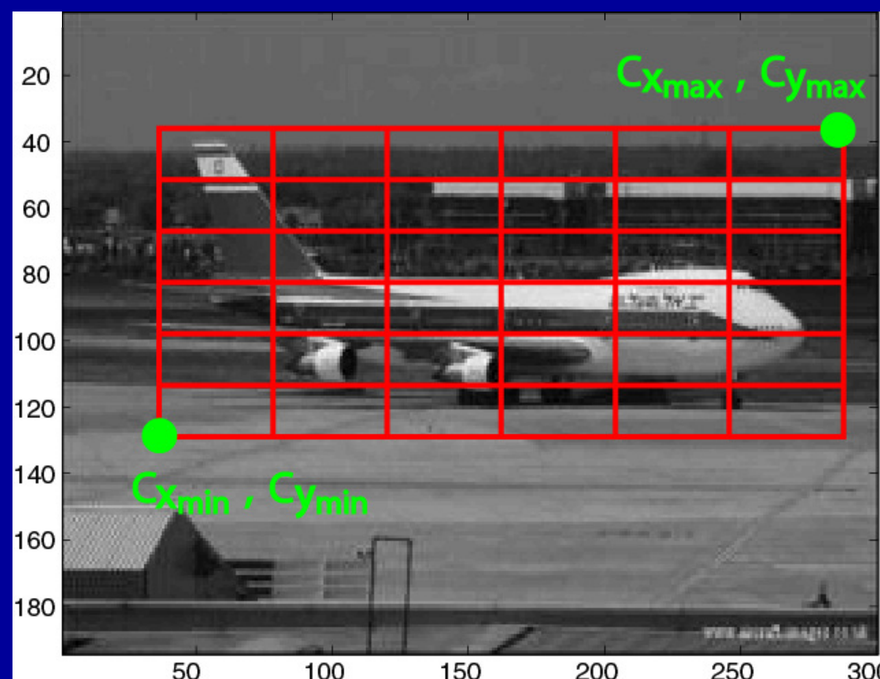


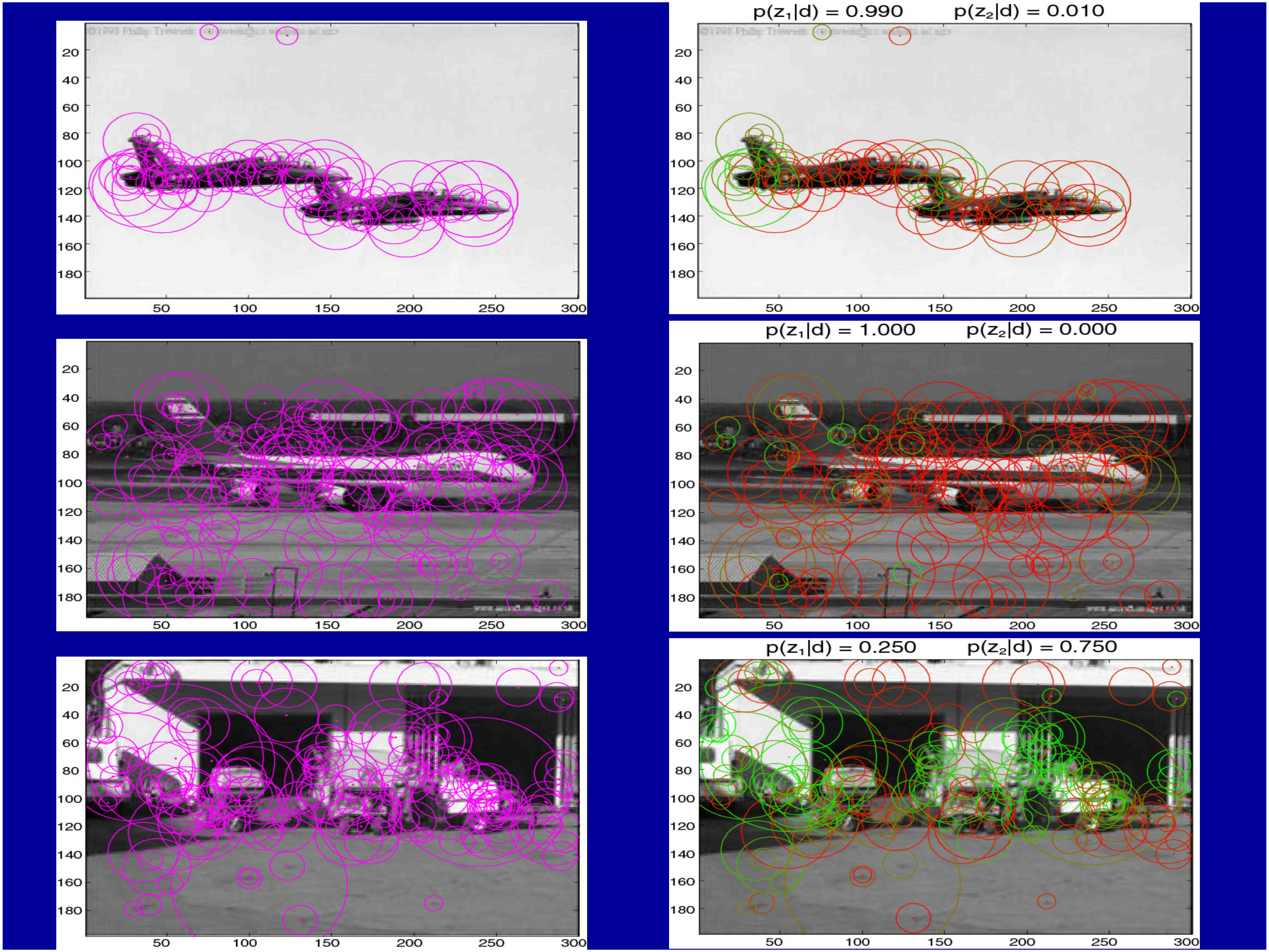
2. Improvements on pLSA: TSI-pLSA

- ABS-pLSA uses absolute coordinate frame
  - Cannot handle translation or scaling
- Introduce sub-window conditioned on hidden variable c:

- c is a 4-d vector – gives bounding box of object
- Gives (T)ranslation and (S)cale (I)nvariance.





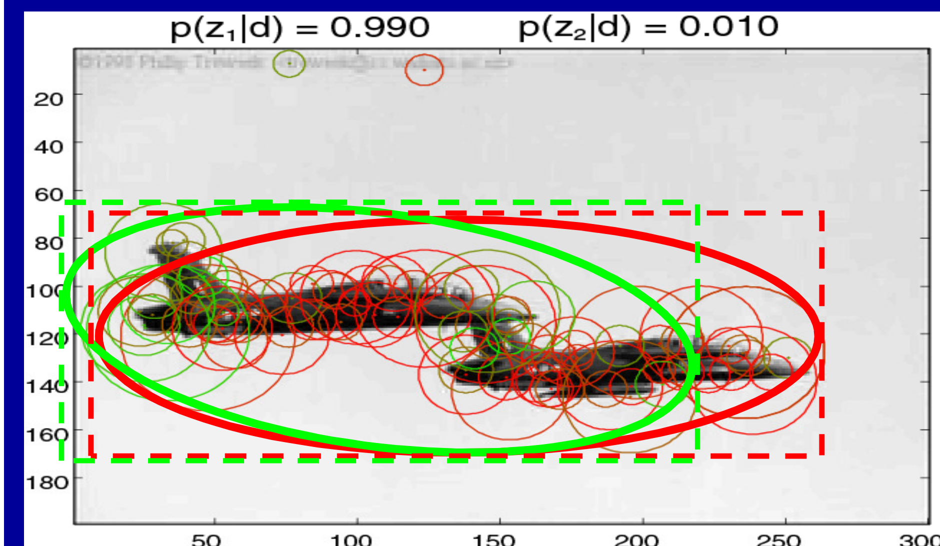


p(z1|d) = 0.990 p(z2|d) = 0.010
p(z1|d) = 1.000 p(z2|d) = 0.000
p(z1|d) = 0.250 p(z2|d) = 0.750

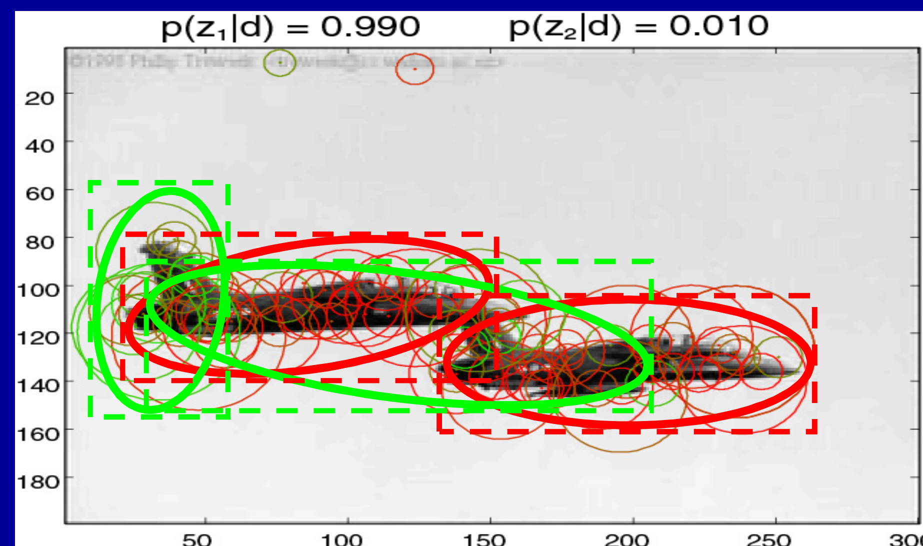
Proposing bounding boxes

- Use pLSA to propose bounding boxes in a bottom-up manner
- Use regions weighted by P(w|z).
- Fit Gaussian mixture model with (C=1 & C= 2) components for each topic:

C = 1 component



C = 2 components



- Gives us a set of possible bounding boxes

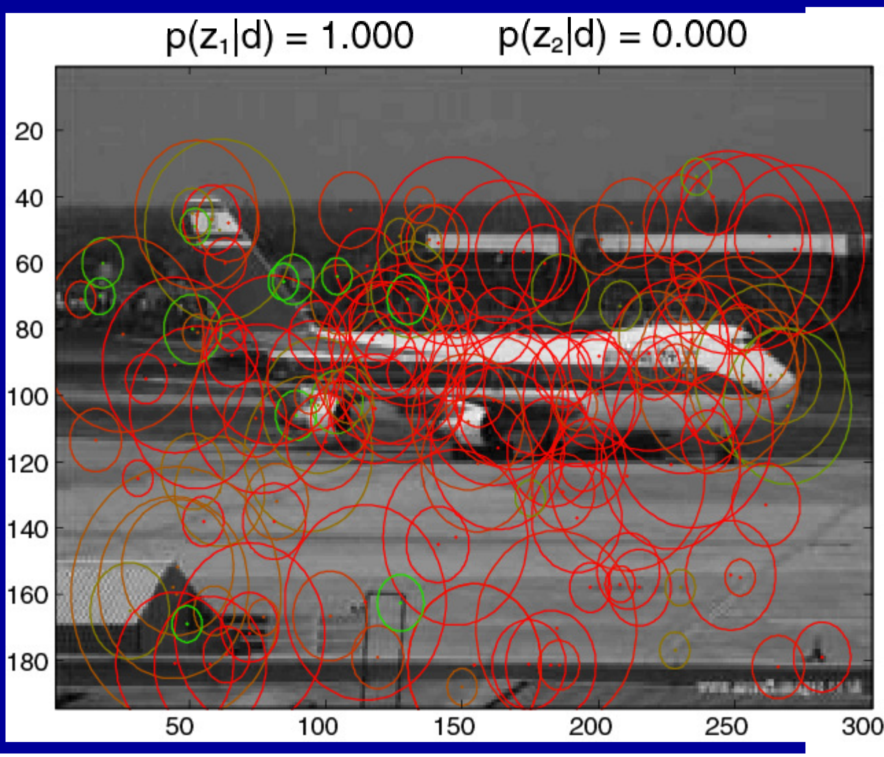
Obtaining bounding box estimates

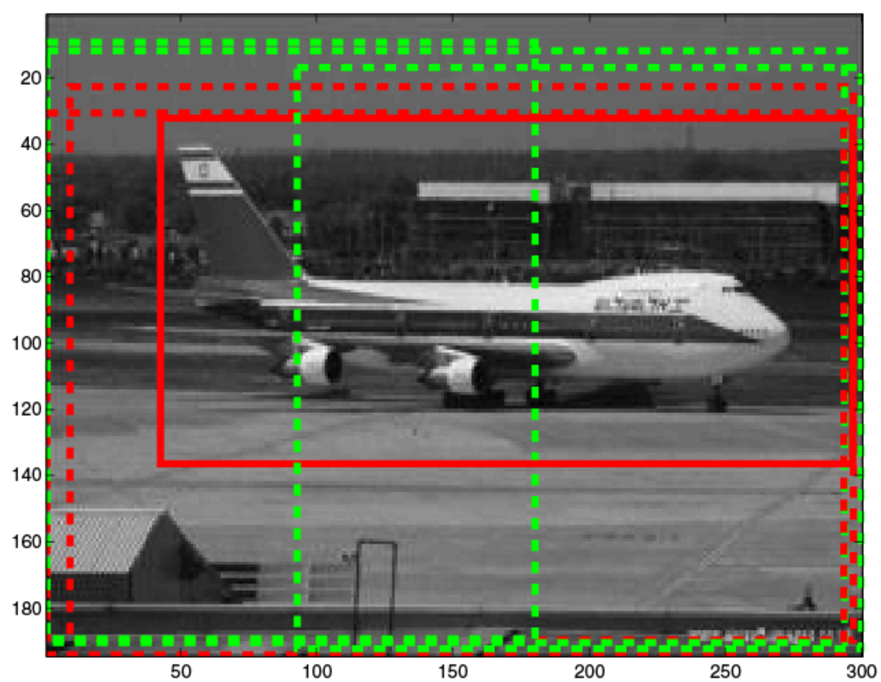
Learning:

- Use p(w|z) density from plain pLSA model, learn from training data
- Sum over different sub-windows in learning

Recognition:

- Average p(w|z) over sub-window of learnt TSI-pLSA model
- Drawback: Only use appearance information
- Restrict choice of bounding boxes to those belonging to best topic





Comparison between pLSA models

- PASCAL object recognition challenge datasets
- Classification task

	Plain pLSA	ABS – pLSA	TSI-pLSA
PASCAL Cars	31.7	30.8	25.8
PASCAL Motorbikes	33.7	30.2	25.7

Training pLSA models from Google images

- Multiple topics can handle polluted data
  - Each topic models a visually consistent component of data

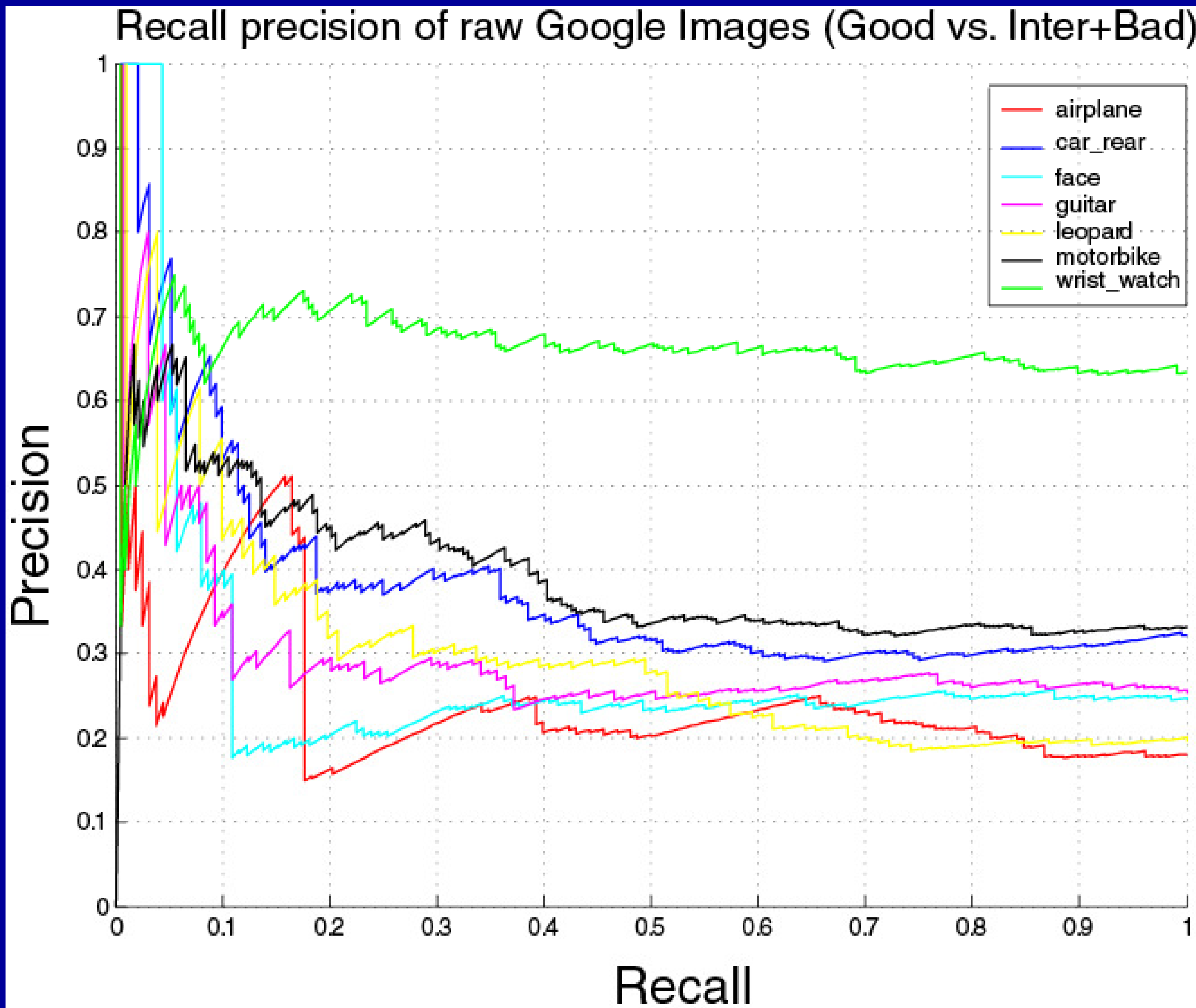
- Different aspects handled by
  - Different topics
  - Multimodal nature of densities

- TSI-pLSA can handle translation and scaling of object within image

- Everything automatic except for:

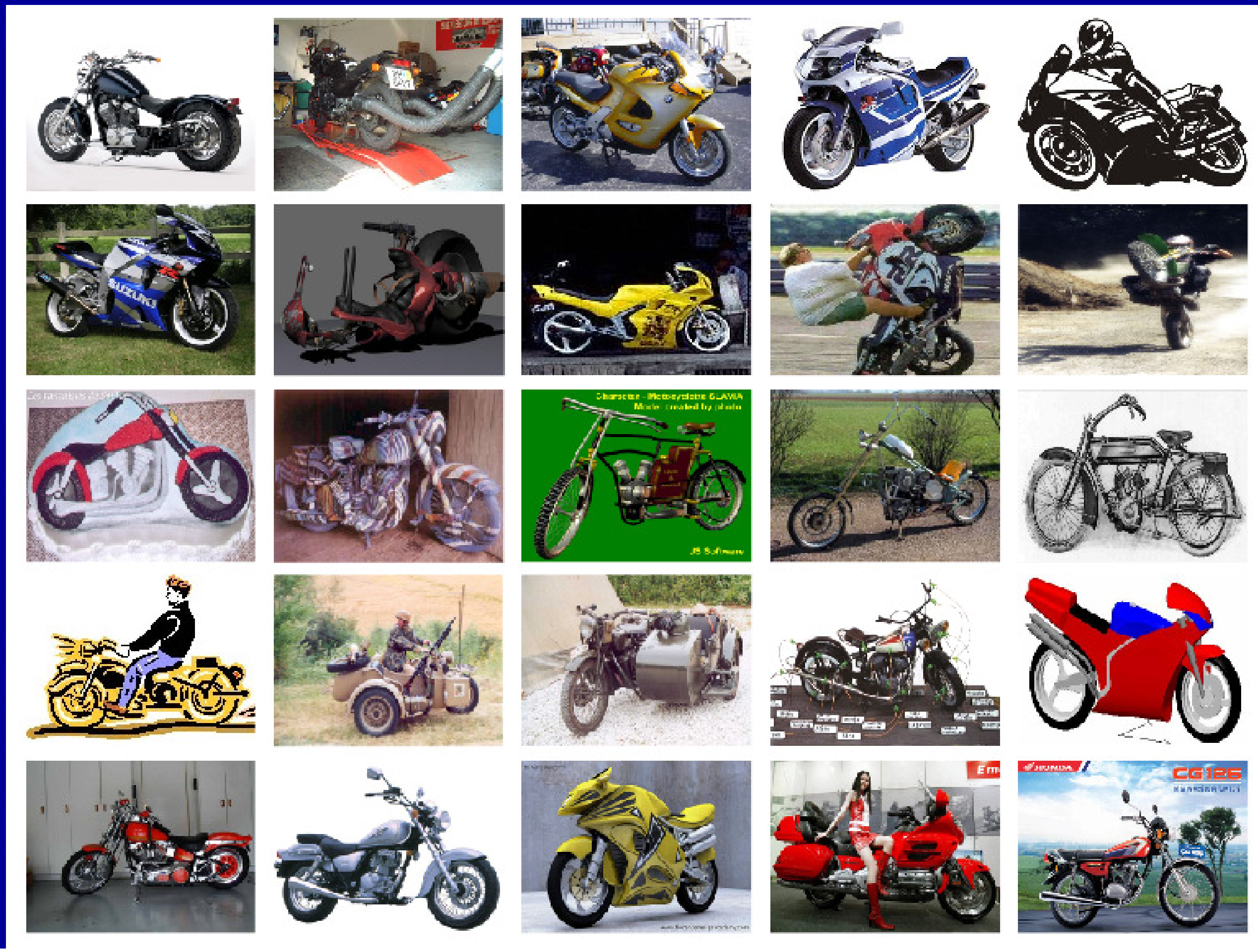
1. Number of topics to use (Z)? Fix Z =8
2. How to pick topic belonging to good images?

Google's variable search performance



Picking the best topic

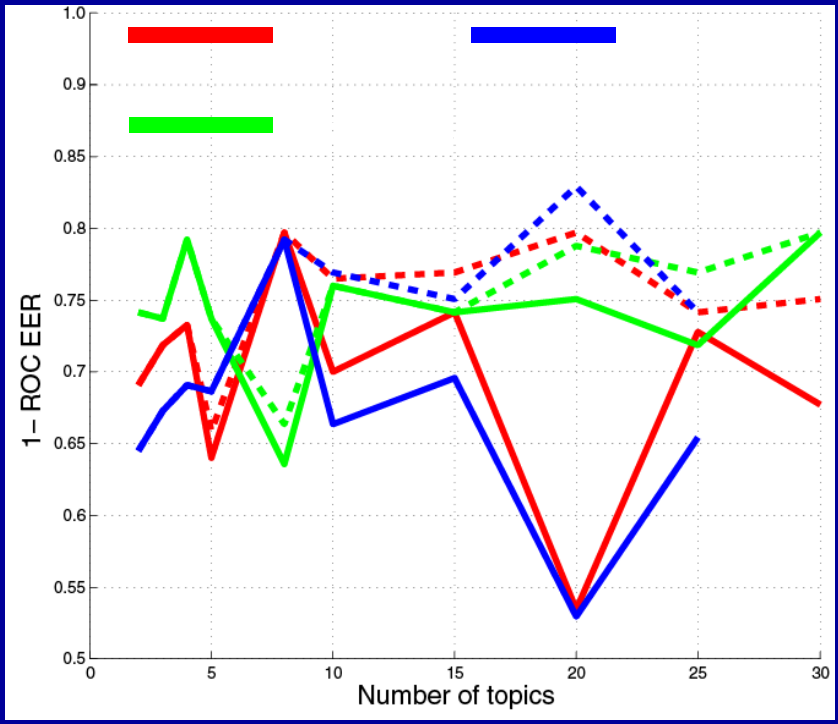
- Use Google's automatic translation tool to translate keyword
- Use: German, French, Italian, Spanish, Chinese, English
- Take first 5 images returned using translated keywords to give validation set

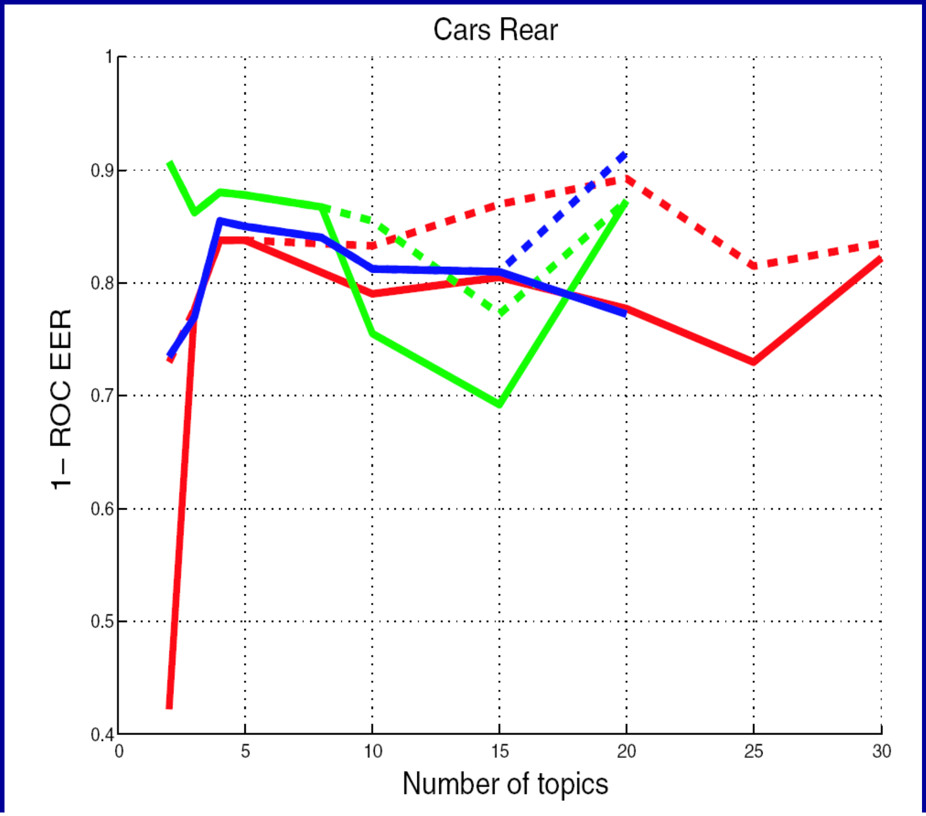


2. Picking the number of topics

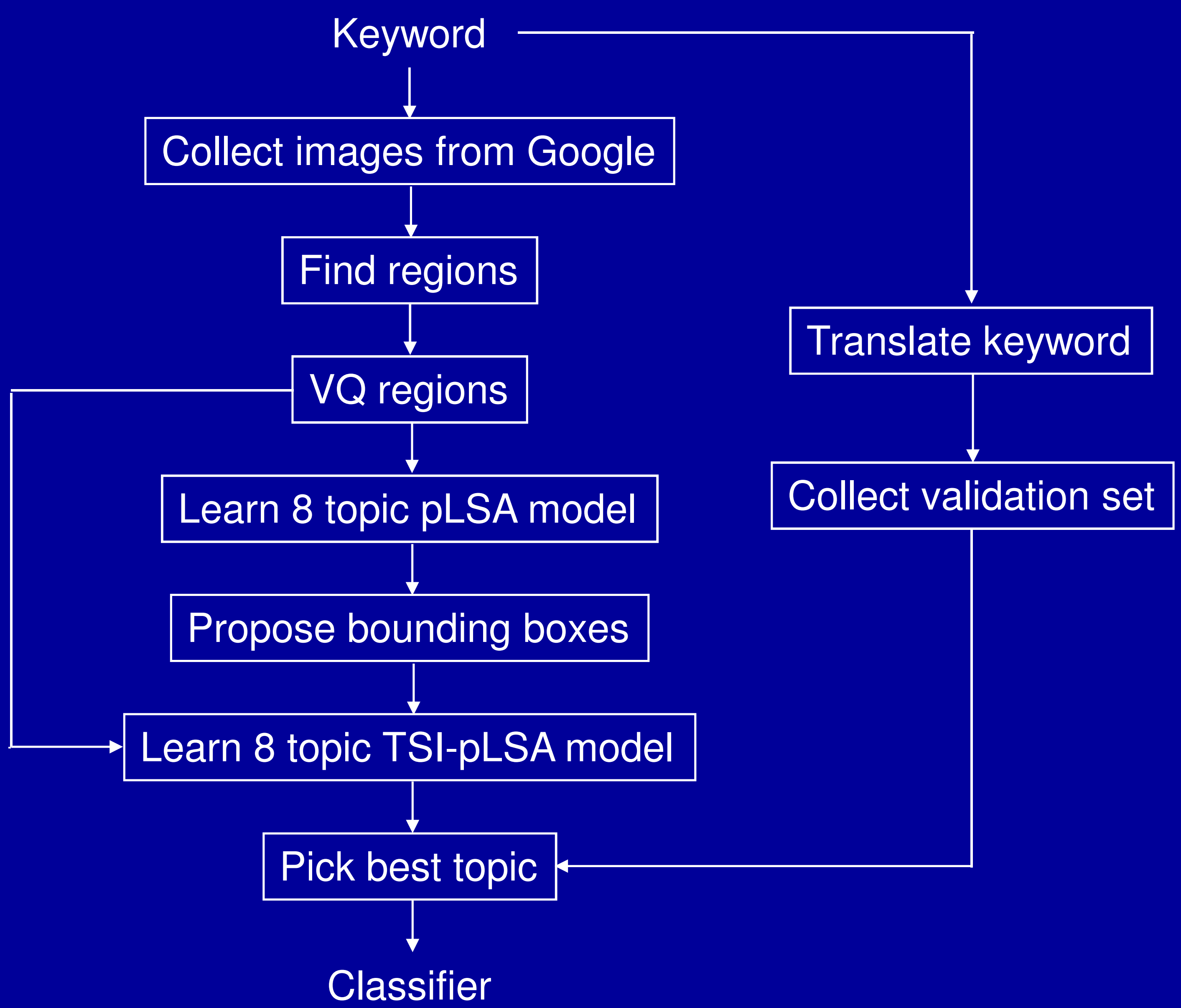
- Small number gives very general topics
- Large number gives highly specific topics
  - Overfitting problems
  - Difficult to automatically pick best one
- Use Z=8, chosen empirically using face and airplane classes

Face



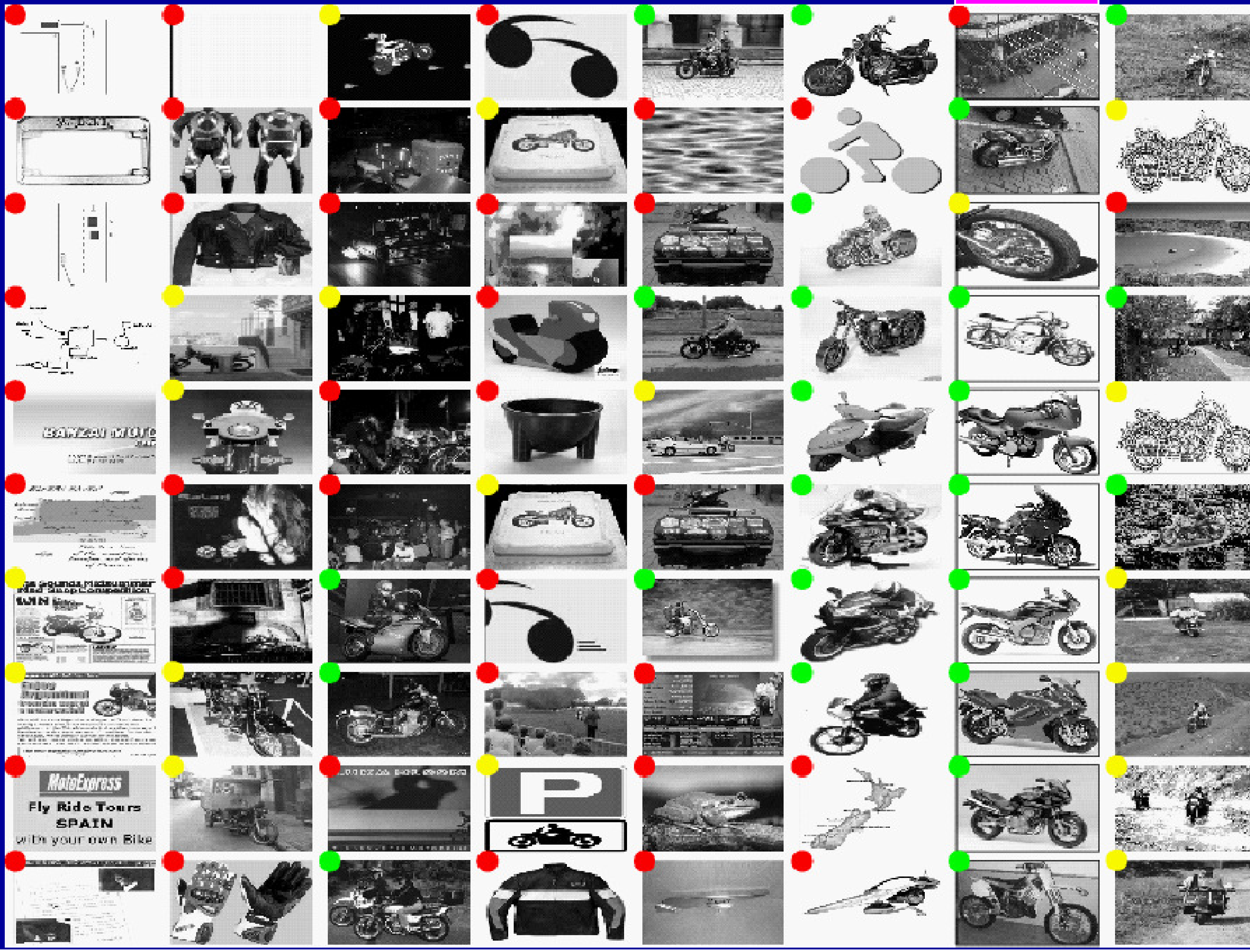


Overall learning scheme



Motorbike – pLSA

Motorbike – TSI-pLSA



Car Rear – pLSA

Car Rear – TSI-pLSA

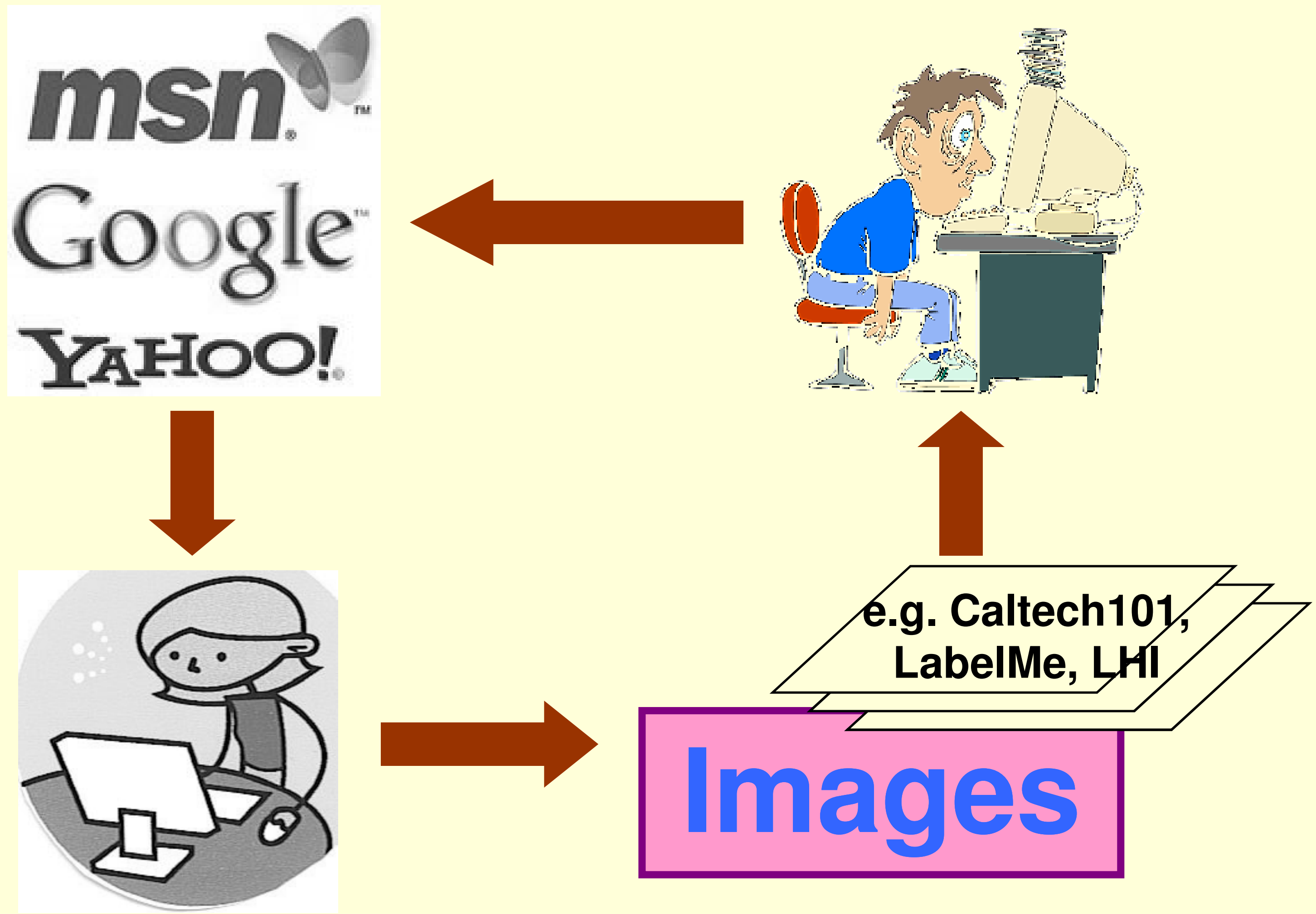
OPTIMOL: automatic Object Picture collecTion via Incremental MOdel Learning

L.-J. Li
G. Wang
L. Fei-Fei

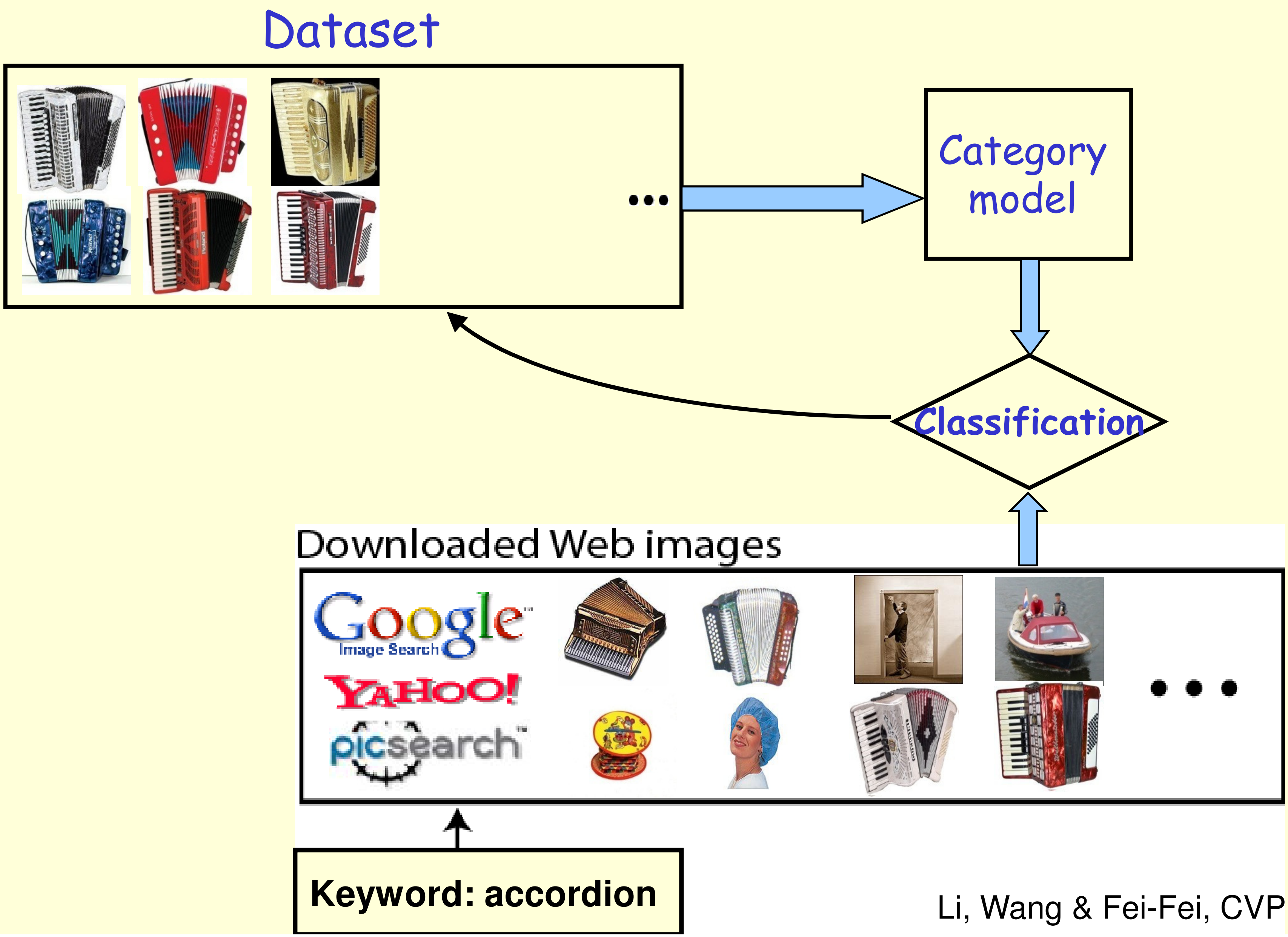
Presented at CVPR '07

a chicken and egg problem…

…among users, researchers, and data

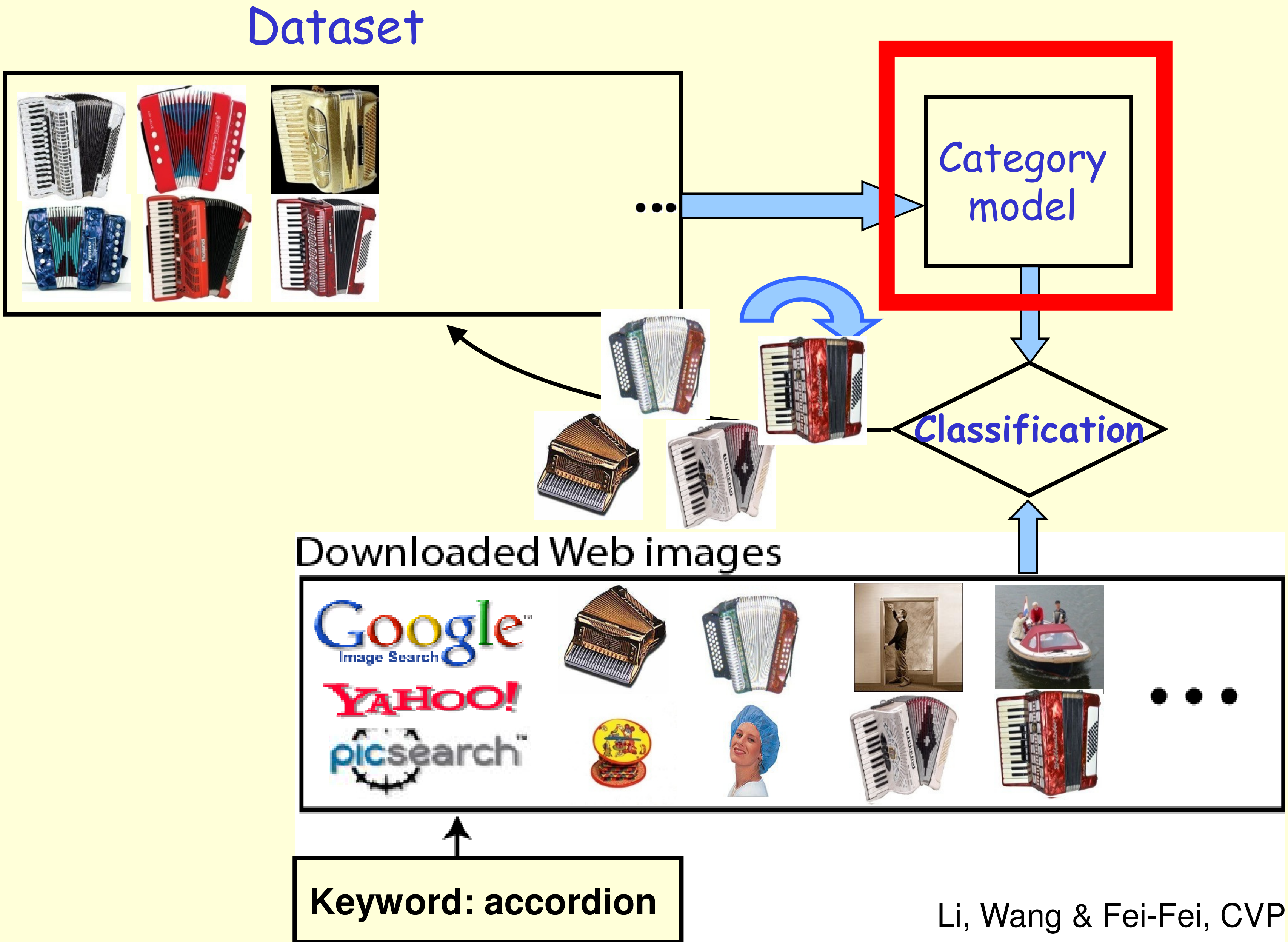


Framework



Li, Wang & Fei-Fei, CVPR 2007

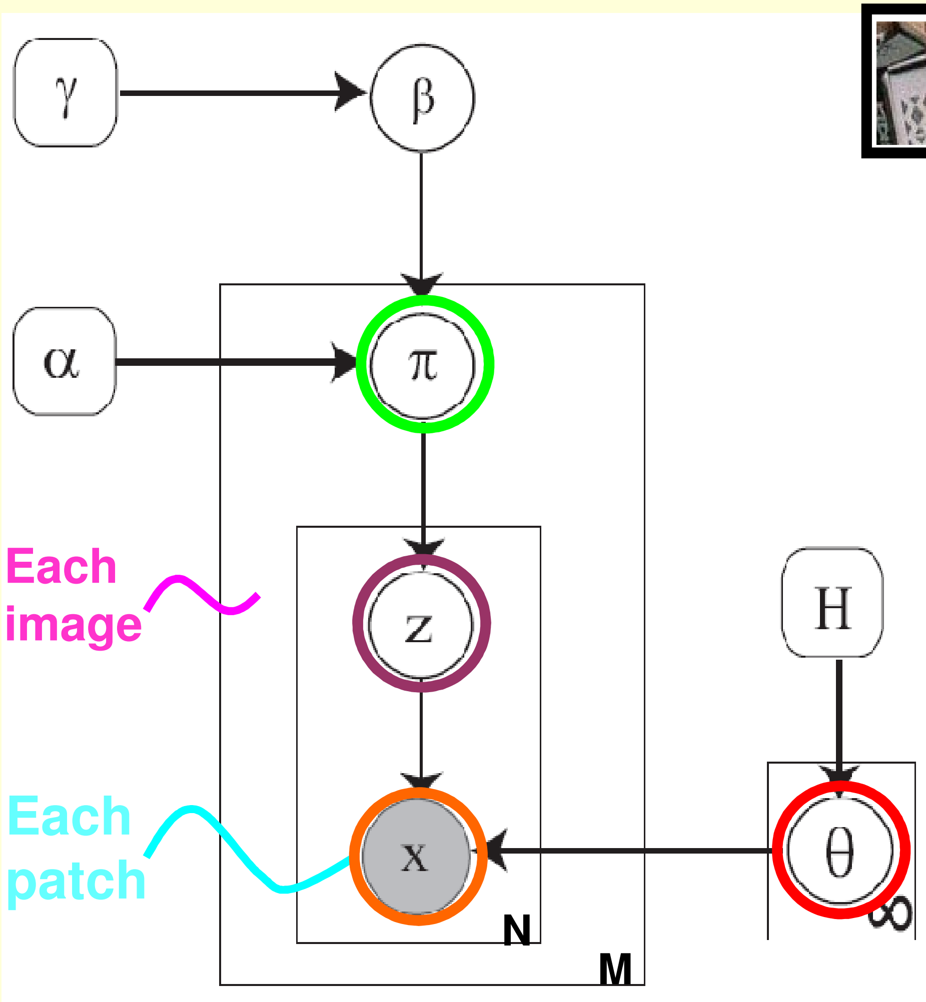
Framework

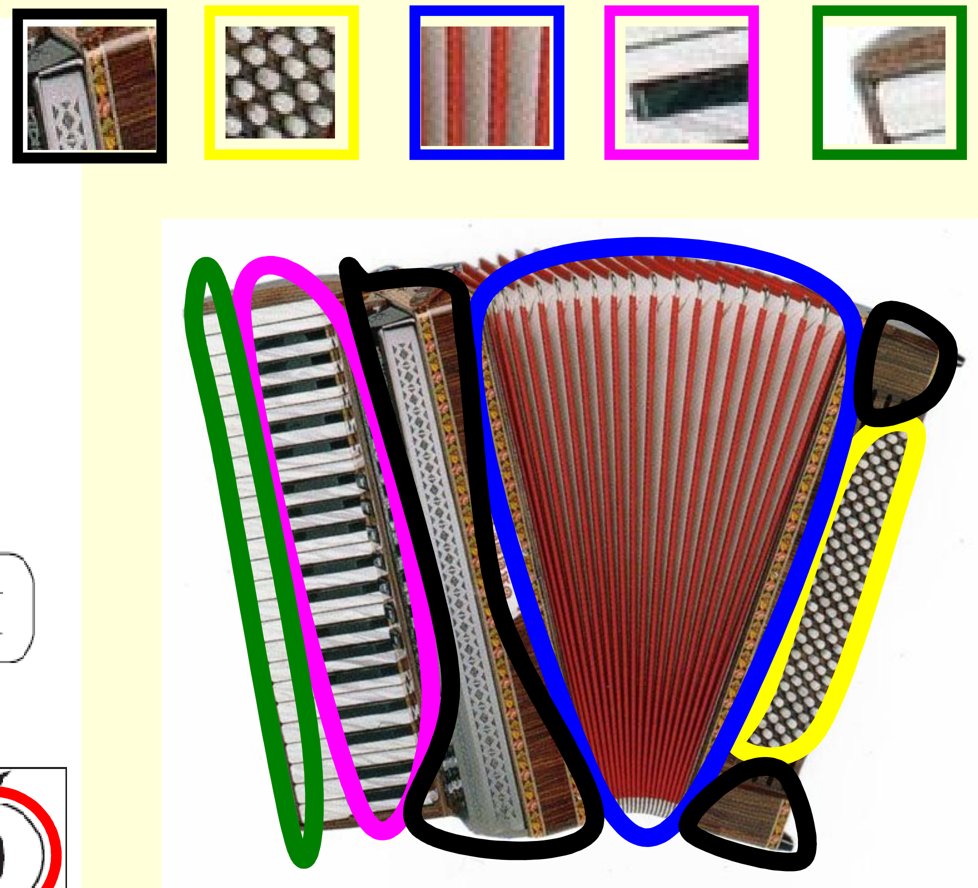


Li, Wang & Fei-Fei, CVPR 2007

Nonparametric topic model

-Hierarchical Dirichlet Process (HDP)

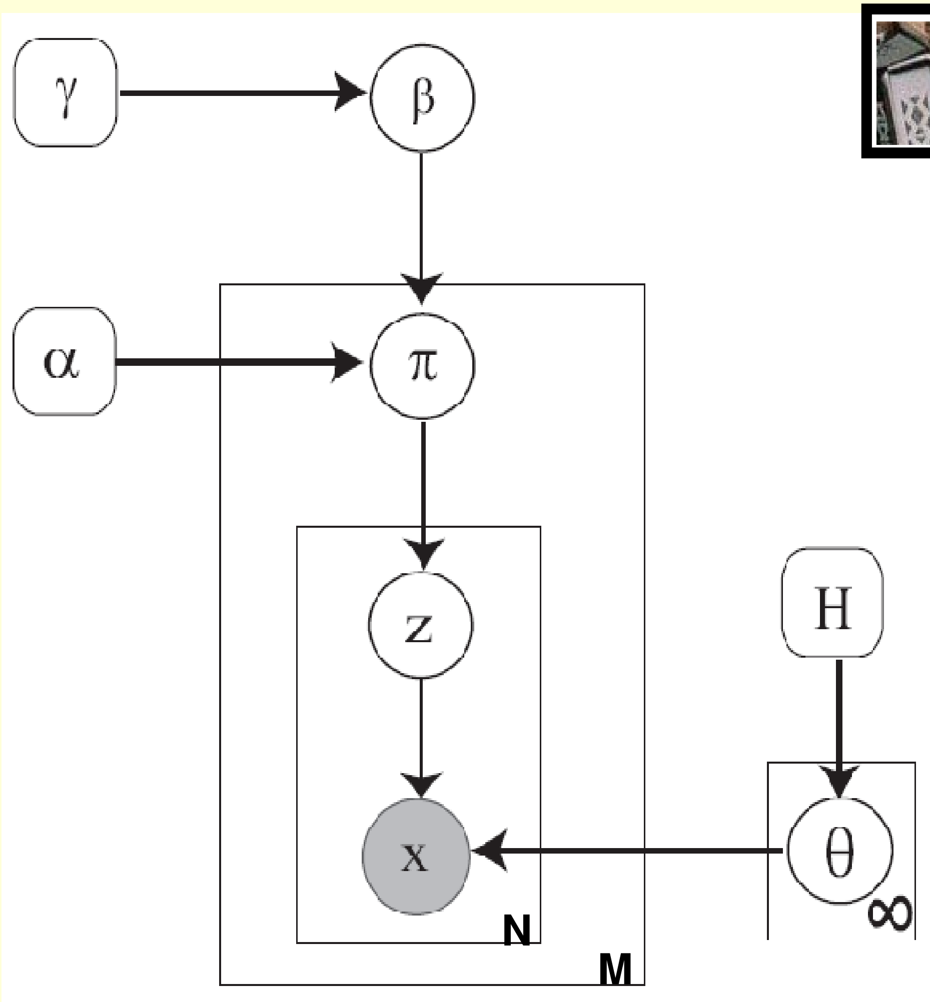


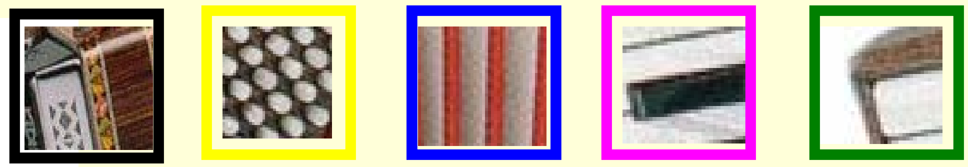


Teh, et al. 2004; Sudderth et al. CVPR 2006; Wang, Zhang & Fei-Fei, CVPR 2006

Nonparametric topic model

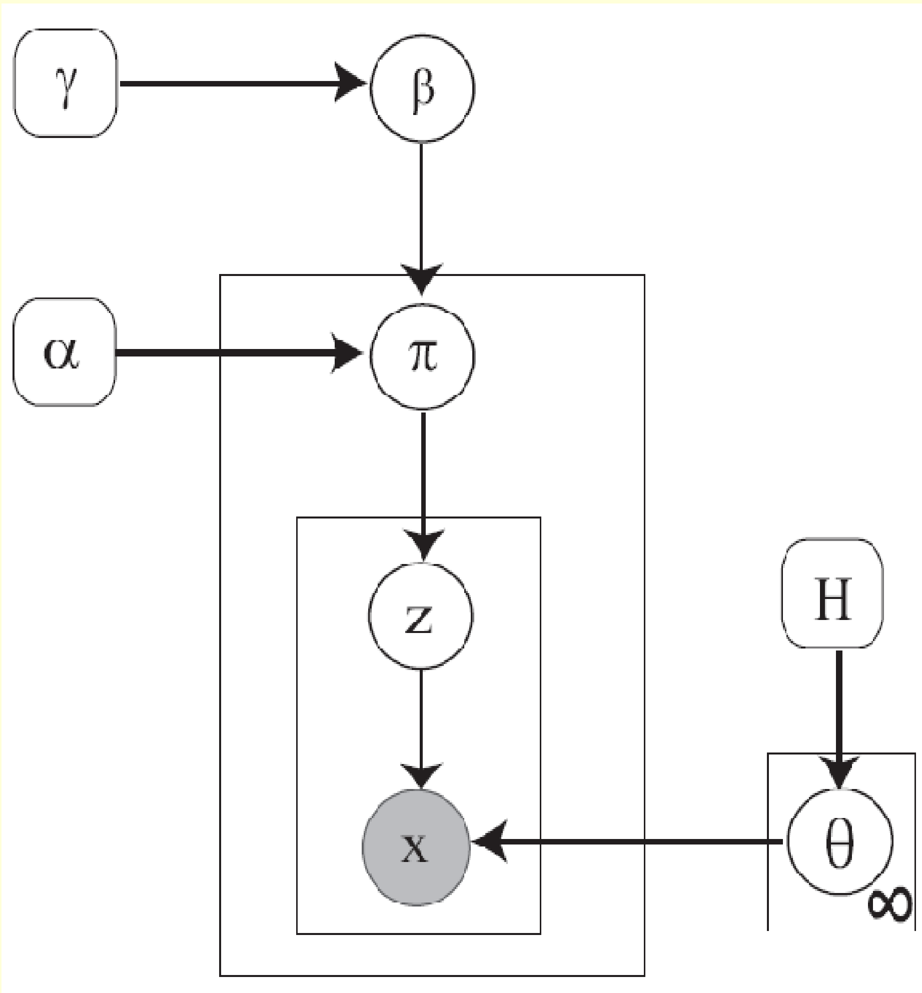
-Hierarchical Dirichlet Process (HDP)





Teh, et al. 2004; Sudderth et al. CVPR 2006; Wang, Zhang & Fei-Fei, CVPR 2006

Classification



Category likelihood for I:

$$(I|c) = \prod_{i} \sum_{j} P(x_i|z_j, c)P(z_j|c)$$

Likelihood ratio for decision:

$$\frac{P(I|c_f)}{P(I|c_b)}$$

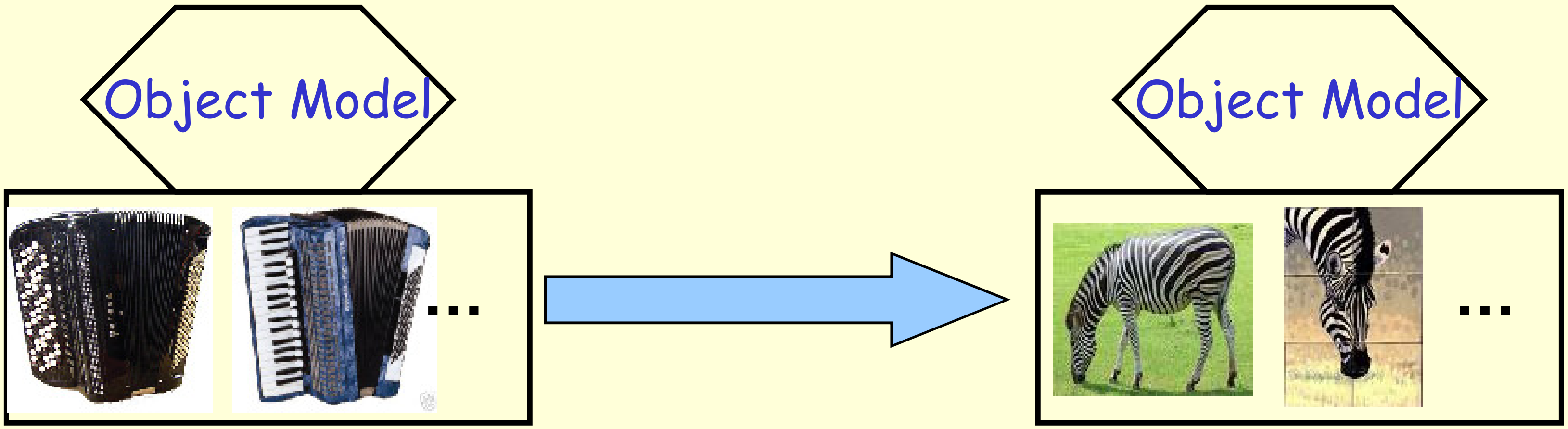
Li, Wang & Fei-Fei, CVPR 2007

Annotation

$$p(x|c_f) = \sum_i p(x|z_i, c_f)p(z_i|c_f)$$

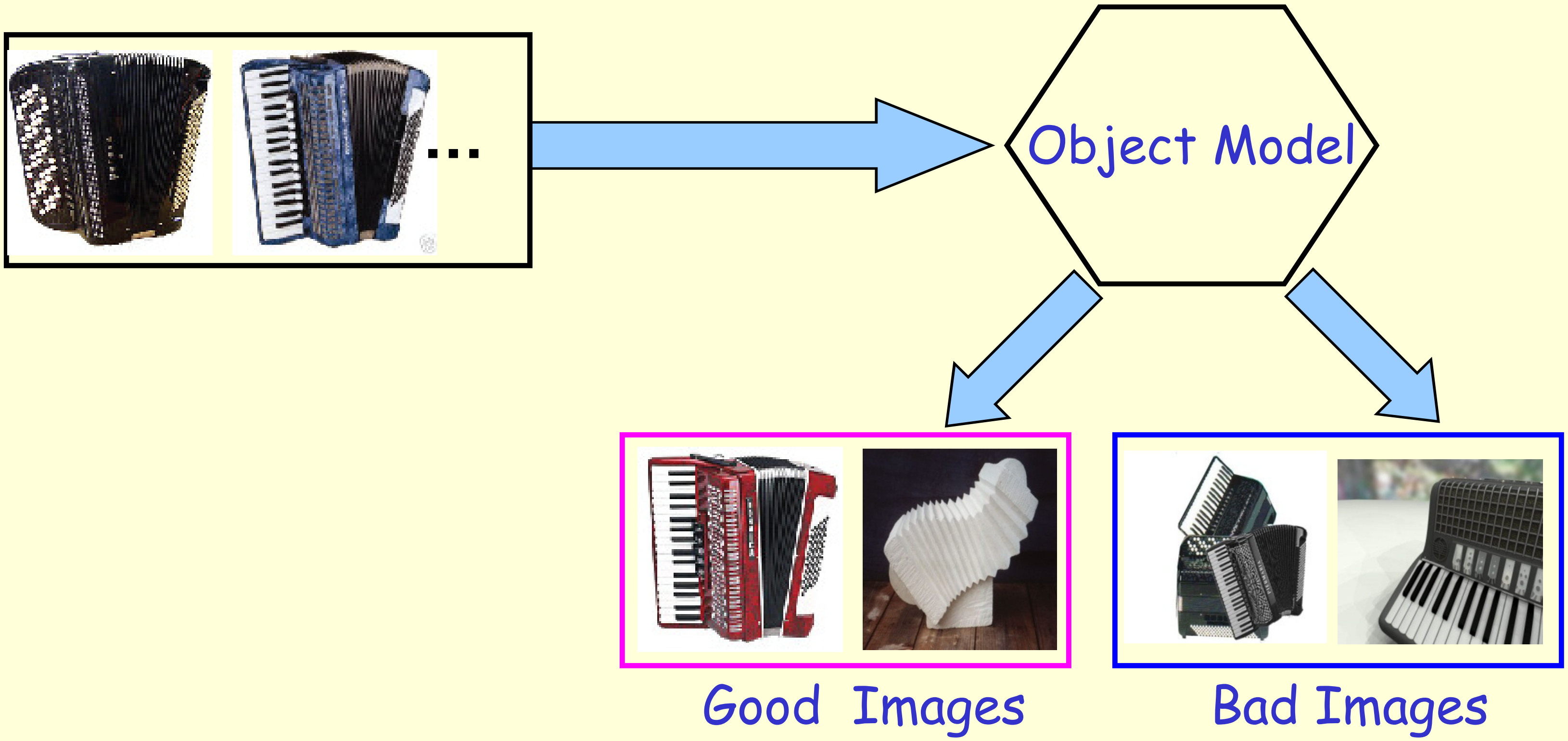
Li, Wang & Fei-Fei, CVPR 2007

Pitfall #1: model drift



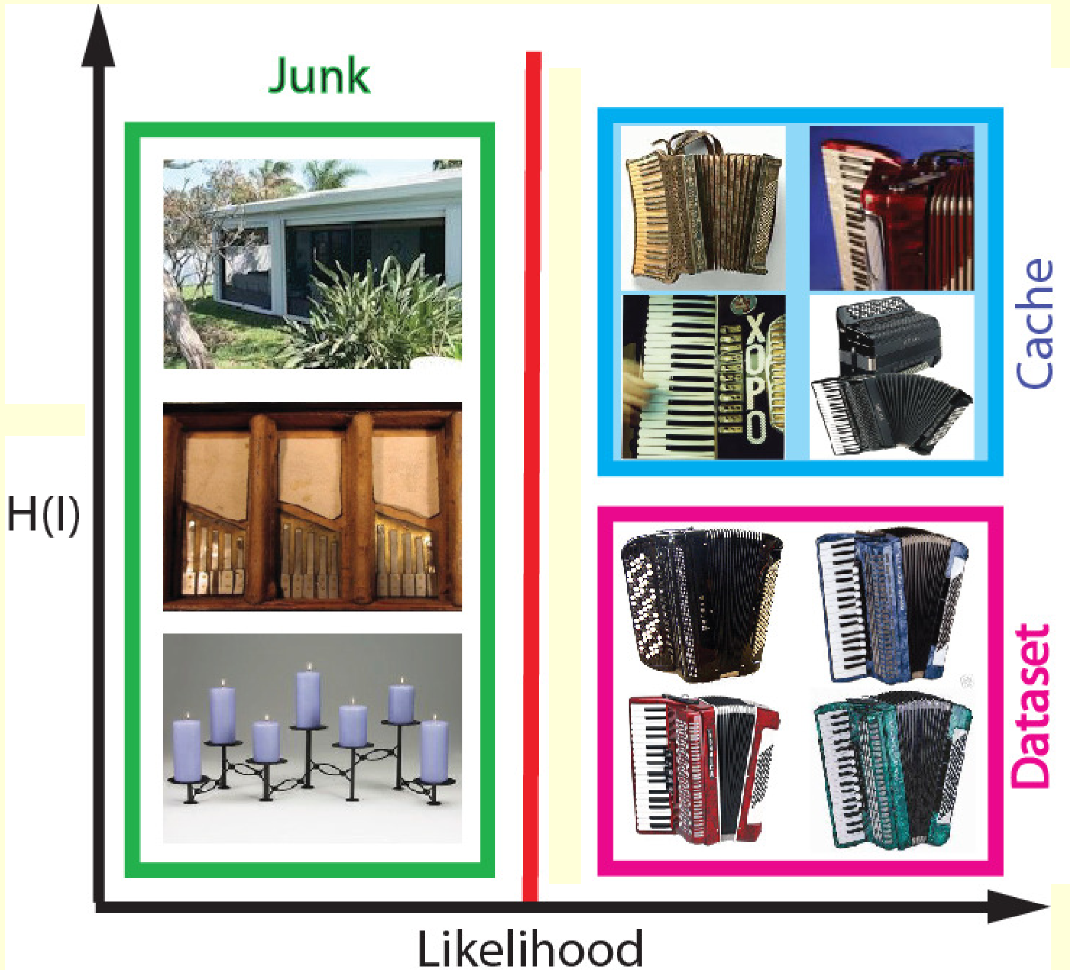
Li, Wang & Fei-Fei, CVPR 2007

Pitfall #2: model diversity



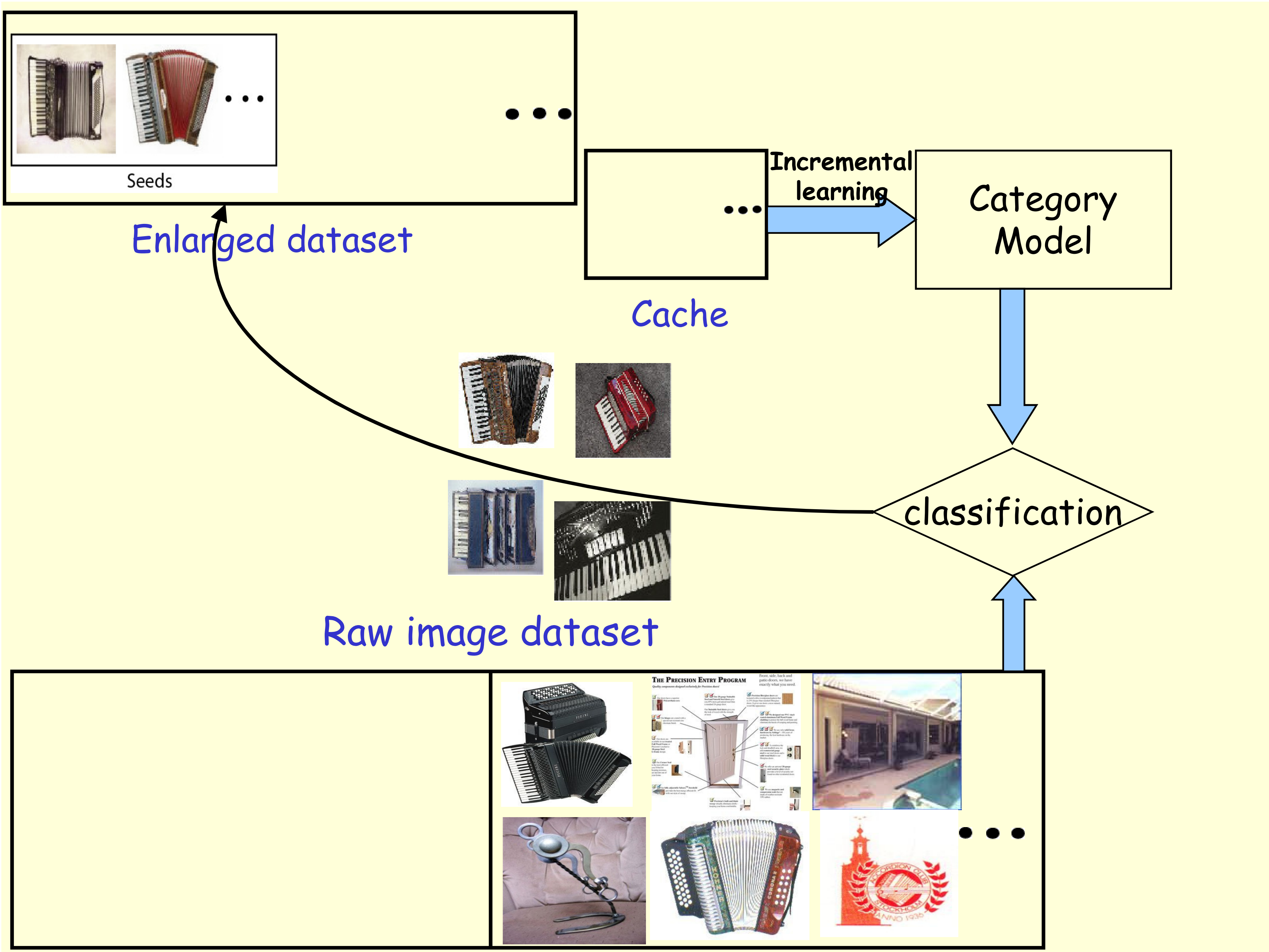
Li, Wang & Fei-Fei, CVPR 2007

The “cache set”

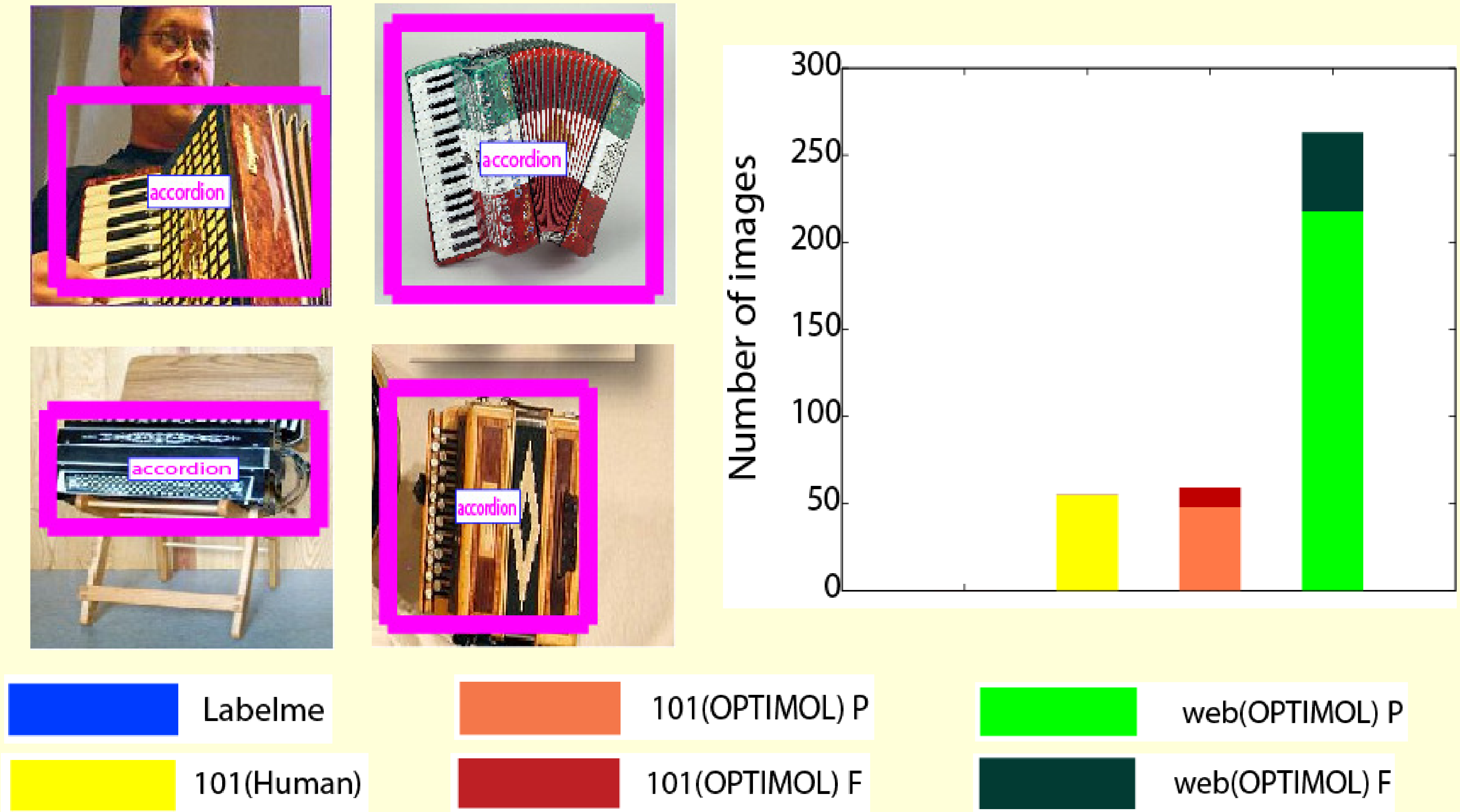


$$H(I) = -\sum_z p(z|I) \ln p(z|I)$$

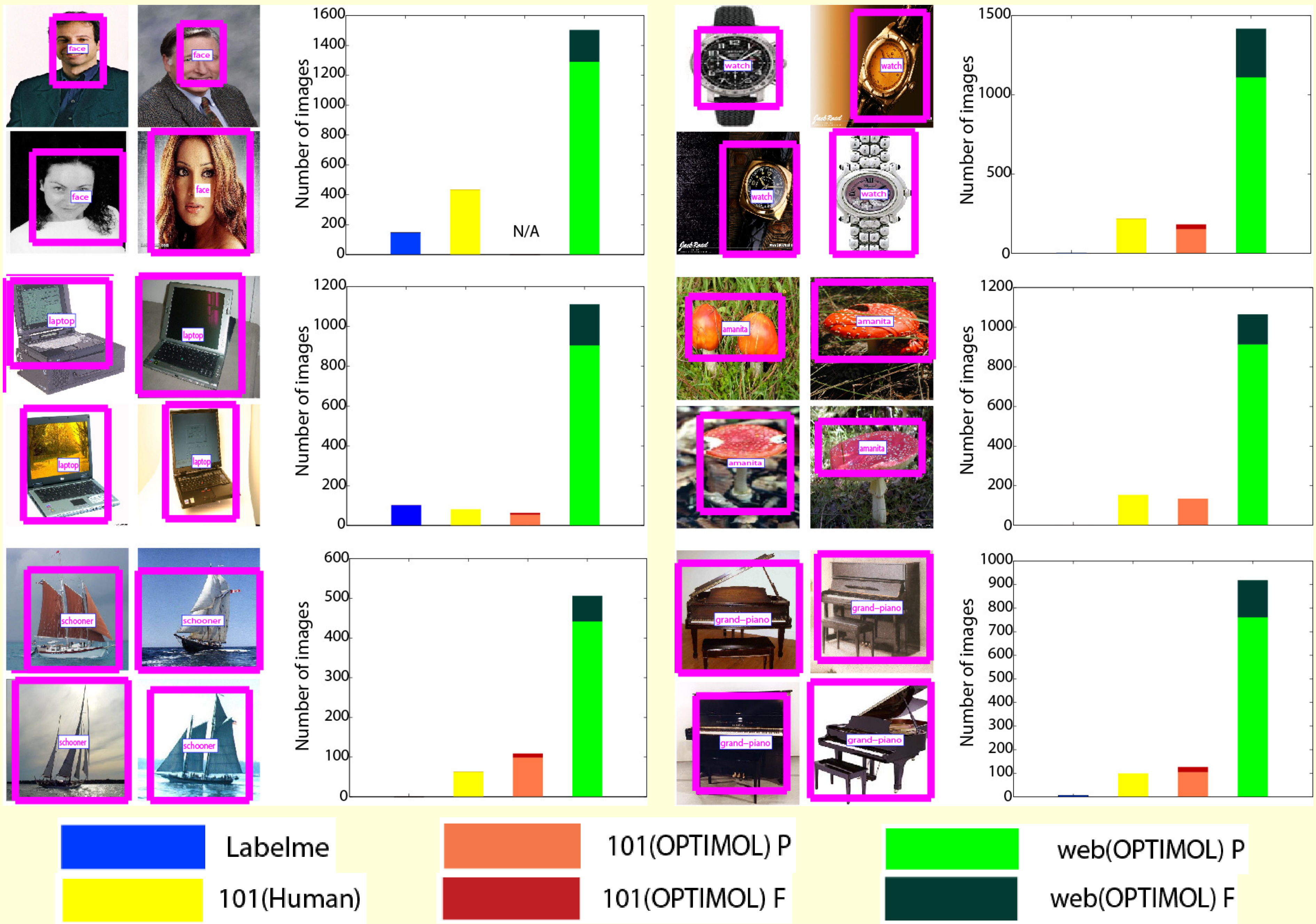
ei-Fei, CVPR 2007


Seeds
Enlarged dataset
Cache
Incremental learning
Category Model
classification
Raw image dataset

Result



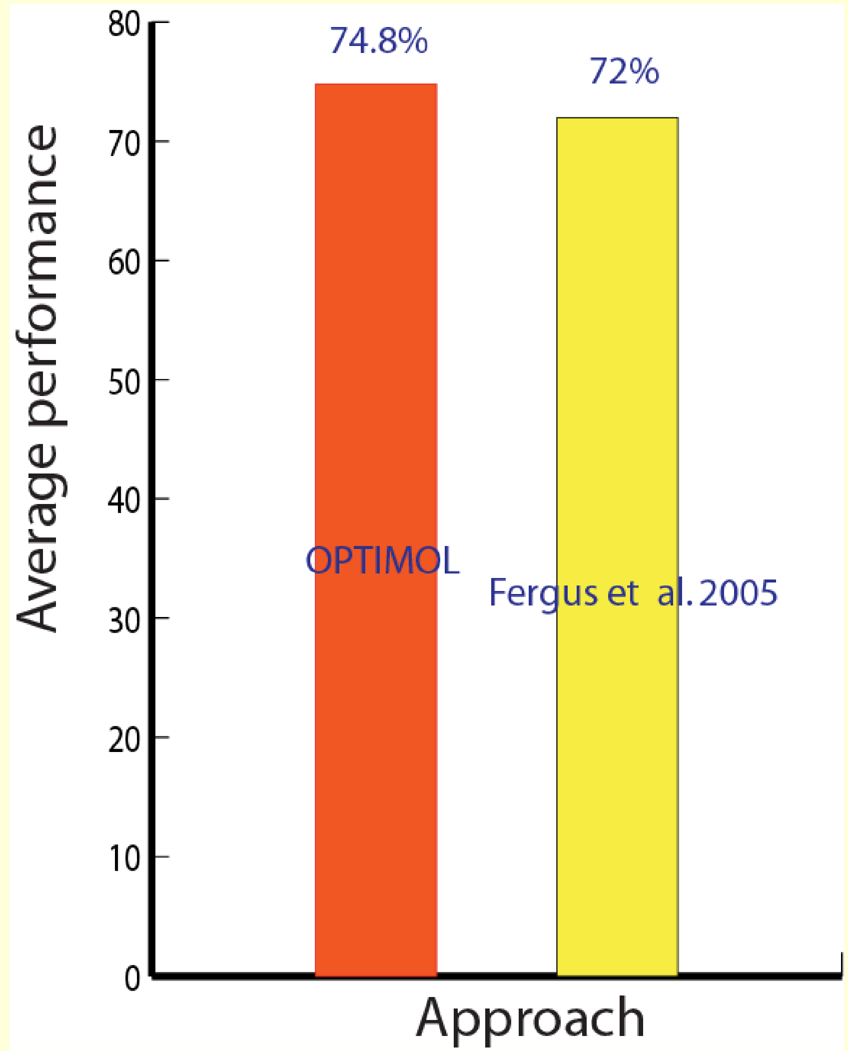
Li, Wang & Fei-Fei, CVPR 2007



Li, Wang & Fei-Fei, CVPR 2007

OPTIMOL also learns good models

	a	c	f	g	l	m	w
airplane	76.0	14.0	0.3	5.3	0.3	0.3	4.8
car	1.0	94.5	0.3	4.5	0.3	0.3	0.3
face	0.5	1.4	82.9	3.7	0.5	0.5	11.5
guita	2.2	4.9	5.6	60.4	13.3	0.2	13.3
leopard	1.0	2.0	1.0	5.0	89.0	1.0	2.0
motorbike	0.3	5.5	0.3	5.5	1.0	67.3	20.5
watch	1.7	5.5	17.7	11.0	5.5	5.0	53.6



Li, Wang & Fei-Fei, CVPR 2007

Animals on the Web

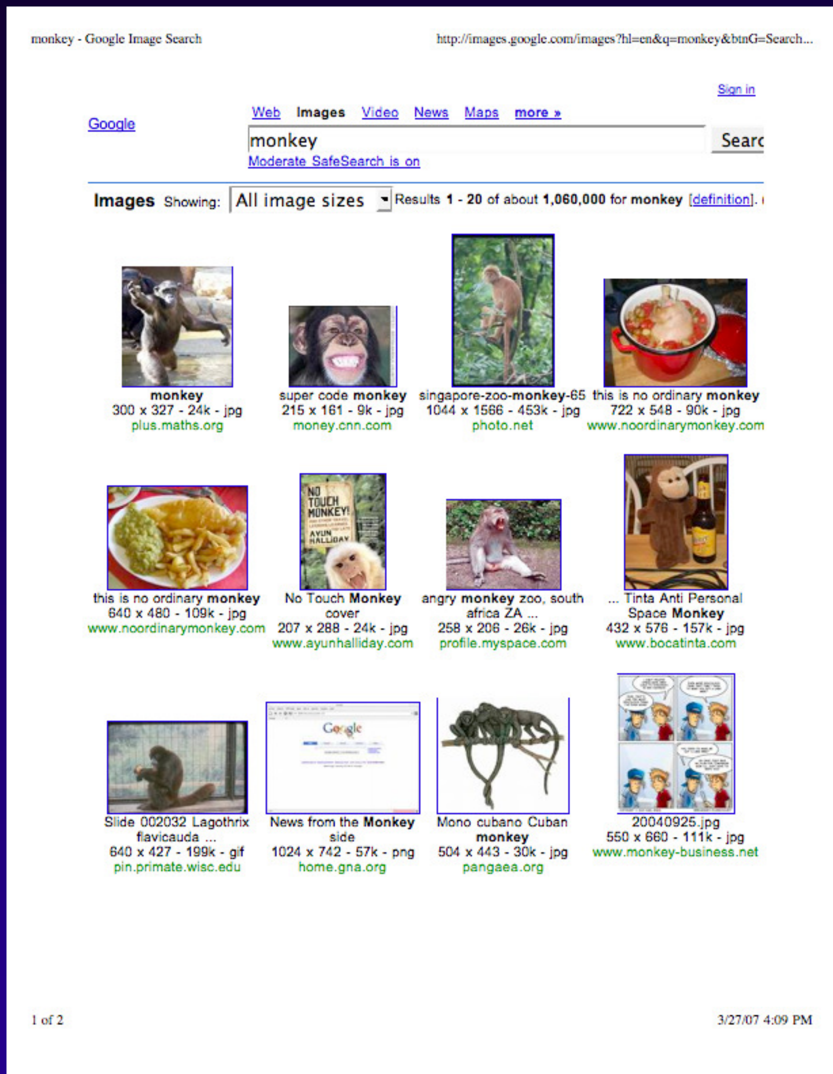
Tamara L. Berg
D. Forsyth

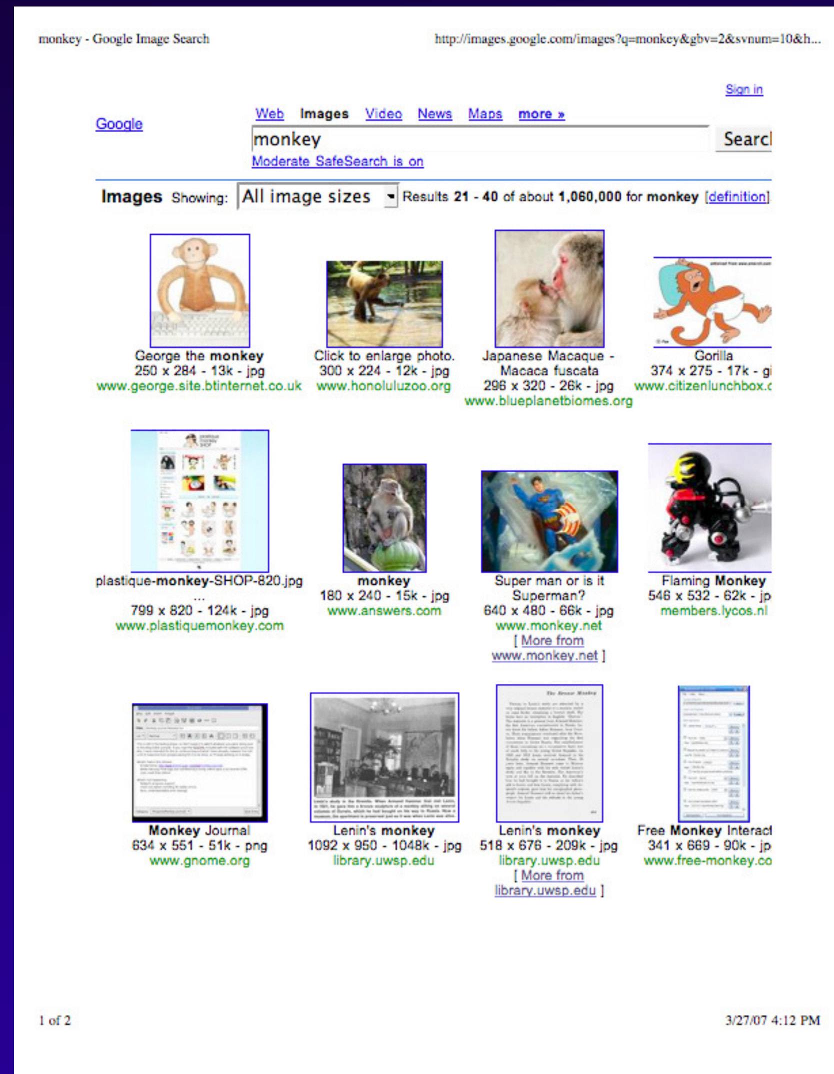
Presented at CVPR '06.

I want to find lots of good pictures of monkeys…

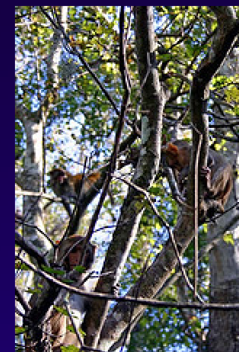
What can I do?

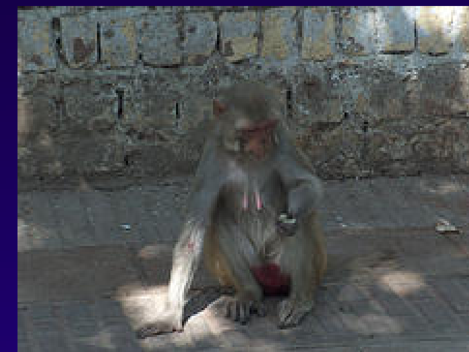
Google Image Search -- monkey





Flickr Search - monkey





Even with humans doing the labeling, the data is extremely noisy -- context, polysemy, photo sets

Words alone still won’t work!

“Animals on the Web” Results

General Approach

- Vision alone won't solve the problem.
- Text alone won't solve the problem.
  -> Combine the two!

Consumer Photo Collections

Flickr – 3 billion photographs, several million uploaded per day

Over the hills and far away

Road, Hills, Germany, Hoffenheim, Outstanding Shots, specland, Baden-Wuerttemberg

Heavenly

Peacock, AlbinoPeacock, WhiteBeauty, Birds, Wildlife, FeathredaleWildlifePark, PictureAustralia, ImpressedBeauty

End of the world - Verdens Ende - The lighthouse 1

Verdens ende, end of the world, norway, lighthouse, ABigFave, vippefyr, wood, coal

Museum and Library Collections

Fine Arts Museum of San Francisco (82,000 images)

bowl stemmed small Irridescent glass

Woman of Head Howard H G Mrs Gift America North bust States United Sculpture marble

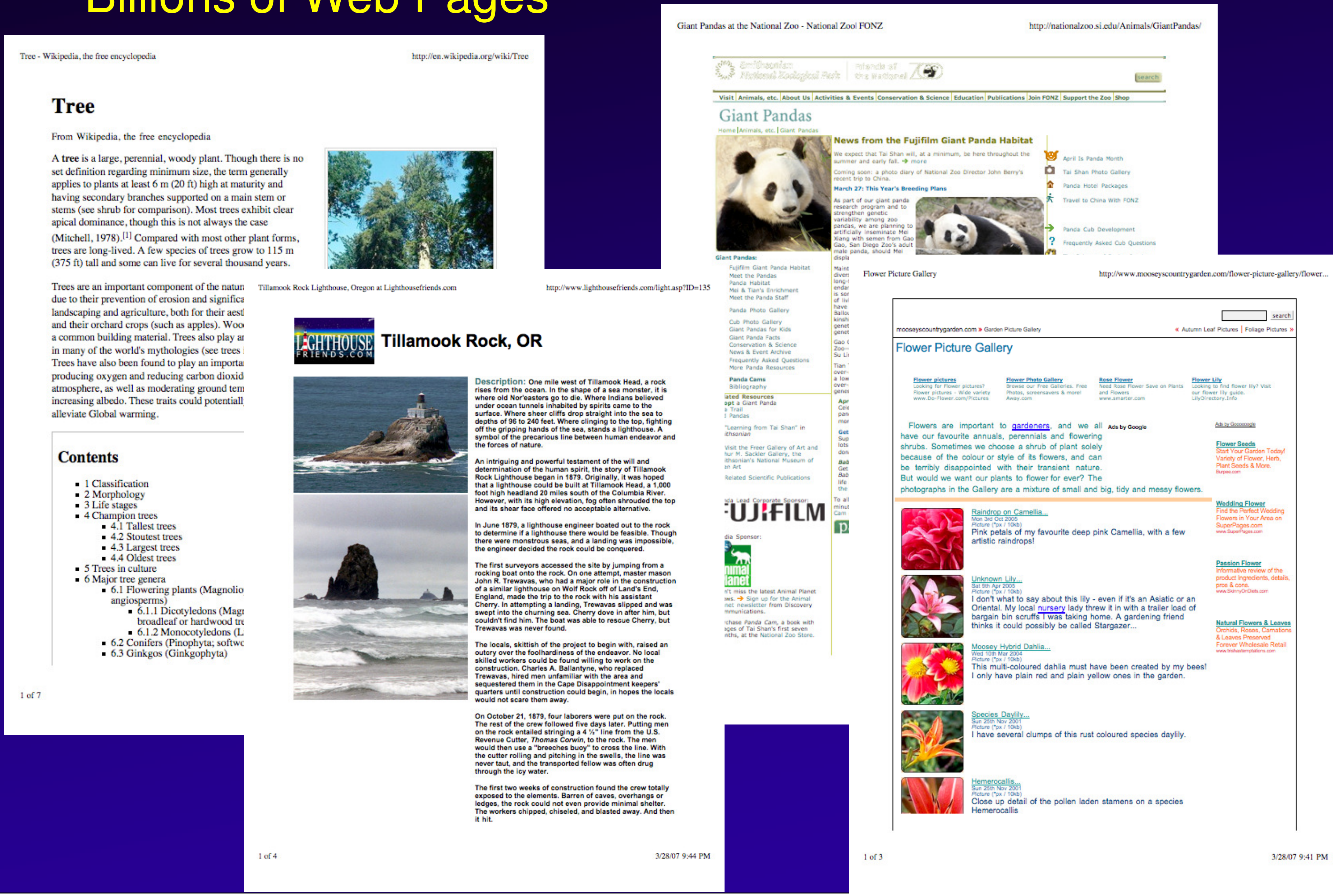
New York Public Library Digital Collection

The new board walk, Rockaway, Long Island

Part of New England, New York, east New Iarsey and Long Iland.

Web Collections

Billions of Web Pages



Video

OUTSIDE IN THE RAIN THE SENATOR WEARING HIS UH BASEBALL CAP A BOSTON RED SOX CAP AS HE TALKED TO HIS SUPPORTERS HERE IN THE RAIN THE UH SENATOR THEY'RE DOING HIS BEST TO TRY TO MAKE HIS CASE THAT HE WILL BE THE MAN FOR THE MIDDLE CLASS AND UH TRY TO CONVINCE HIS SUPPORTERS TO EXPRESS THEIR SUPPORT THROUGH A VOTE ON TUESDAY IN THERE WE ARE TWENTY FOUR HOURS FROM THE GREAT MOMENT THAT THE WORLD IN AMERICA IS WAITING FOR IT I NEED TO YOU IN THESE HOURS TO GO OUT AND DO THE HARD WORK NOT ON THOSE DOORS MAKE THOSE PHONE CALLS TO TALK TO FRIENDS TAKE PEOPLE TO THE POLLS HELP US CHANGE THE DIRECTION OF THIS GREAT NATION FOR THE BETTER CAN YOU IMAGINE A UH SENATOR BEGINNING HIS DAY IN FLORIDA TODAY

TrecVid 2006

Consumer Products

Marc by Marc Jacobs
Adorable peep-toe pumps, great for any occasion. Available in an array of uppers. Metallic fabric trim and bow detail. Metallic leather lined footbed. Lined printed design. Leather sole. 3 3/4" heel.

Zappos.com

soft and glassy patent calfskin trimmed with natural vachetta cowhide, open top satchel for daytime and weekends, interior double slide pockets and zip pocket, seersucker stripe cotton twill lining, kate spade leather license plate logo, imported
2.8" drop length
14"h x 14.2"w x 6.9"d

Katespade.com

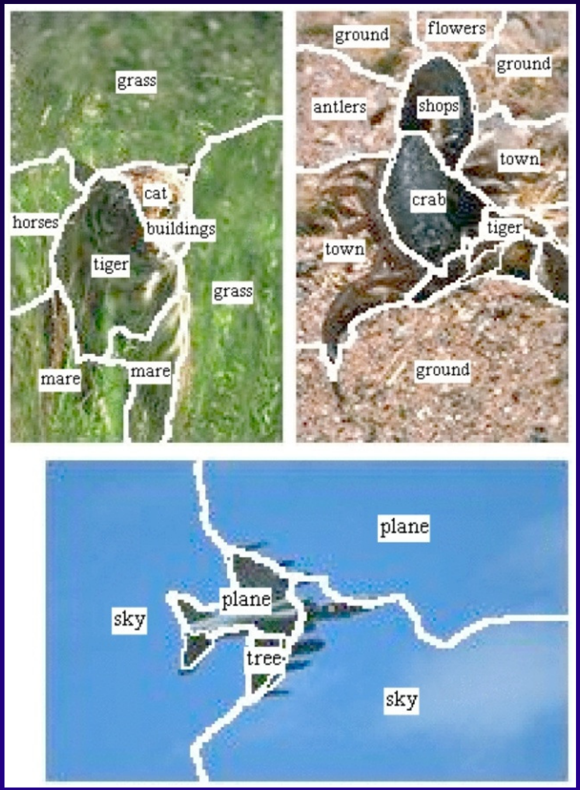
It's the perfect party dress. With distinctly feminine details such as a wide sash bow around an empire waist and a deep scoopneck, this linen dress will keep you comfortable and feeling elegant all evening long.
* Measures 38" from center back, hits at the knee.
* Scoopneck, full skirt.
* Hidden side zip, fully lined.
* 100% Linen. Dry clean.

bananarepublic.com

E-commerce transactions in 2004, 2005, 2006 of $145 billion, $168 billion, and $198 billion (Forrester Research).

Previous Work - Words & Pictures

Labeling Regions



Barnard et al,
JMLR 2003

Clustering Art

Barnard et al,
CVPR 2001

Auto-Annotation

Li and Wang, PAMI 2003

Image Classification

Yanai et al, MIR 2005

Animals on the Web Outline:

Harvest pictures of animals from the web using Google Text Search.

Select visual exemplars using text based information.

Use visual and textual cues to extend to similar images.

Harvested Pictures

14,051 images for 10 animal categories.

12,886 additional images for monkey category using related monkey queries (primate, species, old world, science…)

Text Model

Latent Dirichlet Allocation (LDA) on the words in collected web pages to discover 10 latent topics for each category.

Each topic defines a distribution over words. Select the 50 most likely words for each topic.

Example Frog Topics:

1.) frog frogs water tree toad leopard green southern music king irish eggs folk princess river ball range eyes game species legs golden bullfrog session head spring book deep spotted de am free mouse information round poison yellow upon collection nature paper pond re lived center talk buy arrow common prince

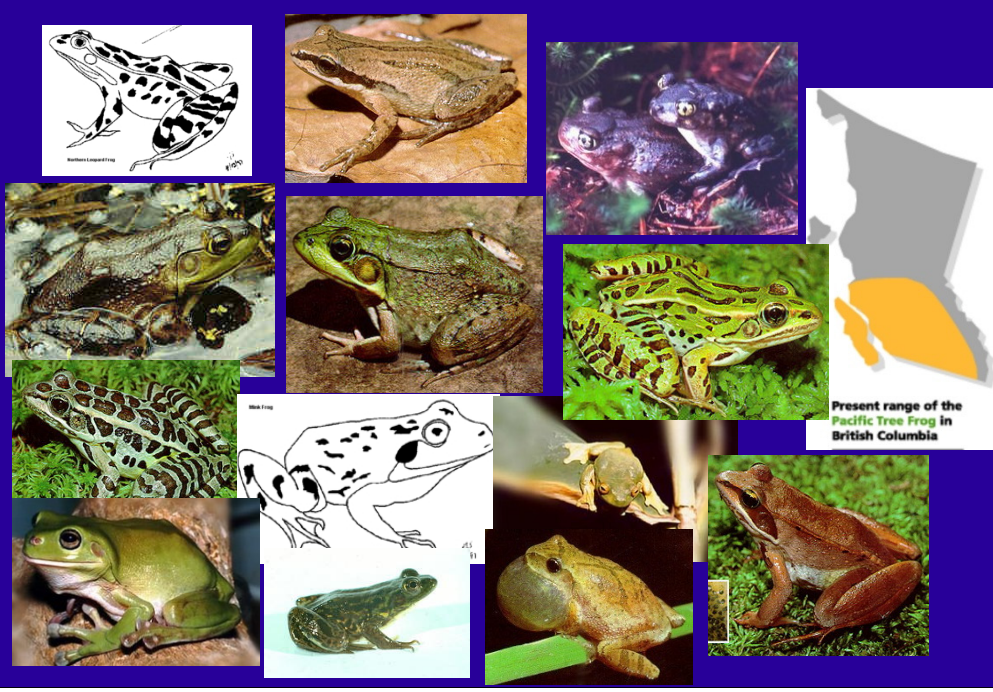
2.) frog information january links common red transparent music king water hop tree pictures pond green people available book call press toad funny pottery toads section eggs bullet photo nature march movies commercial november re clear eyed survey link news boston list frogs bull sites butterfly court legs type dot blue

Select Exemplars

Rank images according to whether they have these likely words near the image in the associated page (word score)

Select up to 30 images per topic as exemplars.

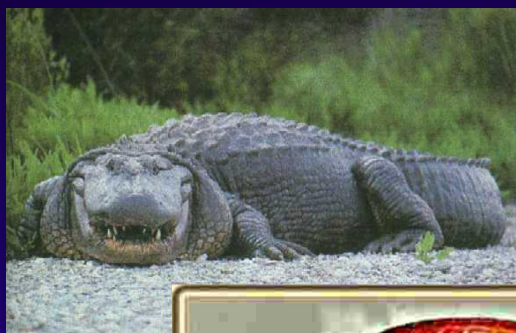
1.) frog frogs water tree toad leopard green southern music king irish eggs folk princess river ball range eyes game species legs golden bullfrog session head ...

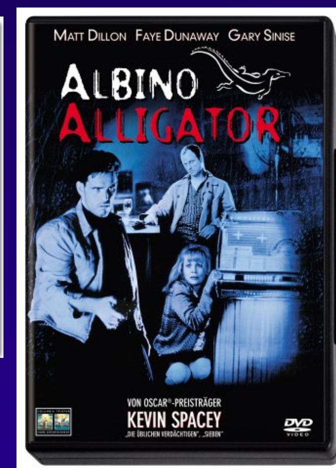


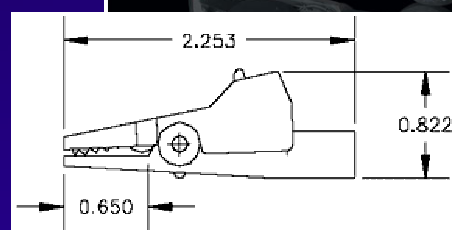
2.) frog information january links common red transparent music king water hop tree pictures pond green people available book call press ...

Senses

There are multiple senses of a category within the Google search results.



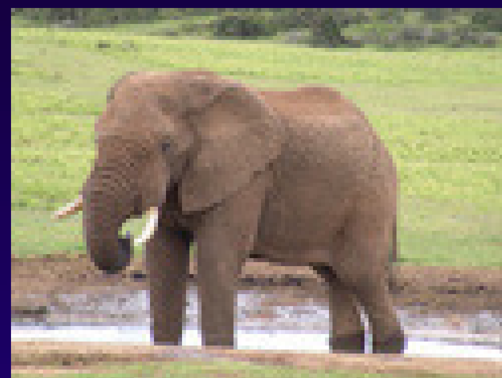




Ask the user to identify which of the 10 topics are relevant to their search. Merge.

Optional second step of supervision – ask user to mark erroneously labeled exemplars.

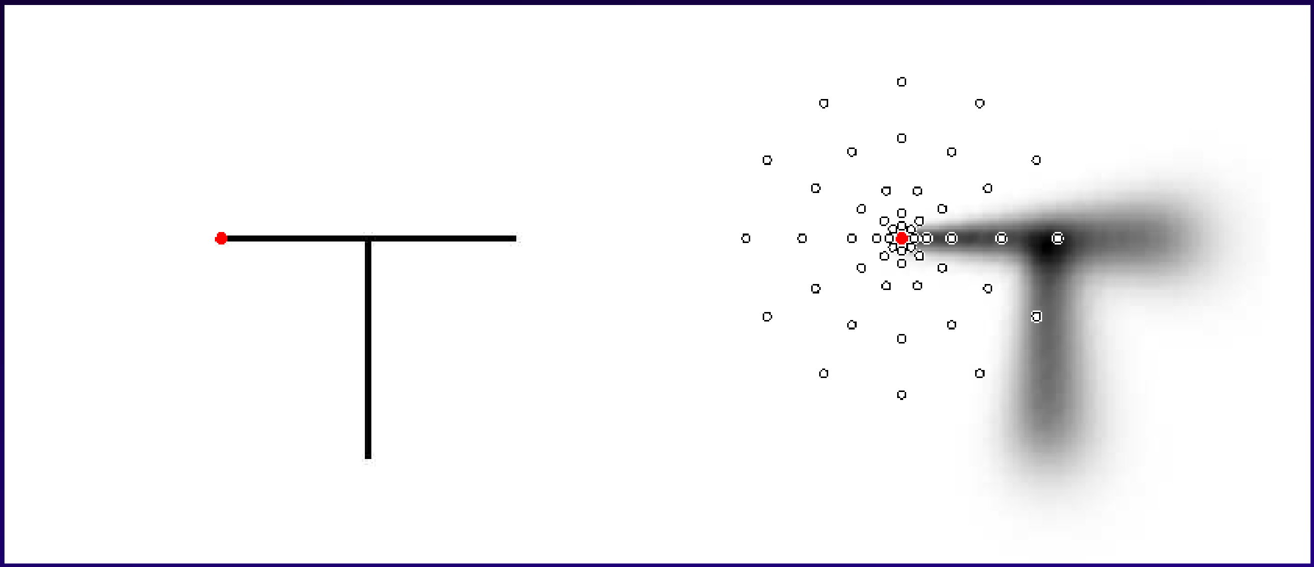
Image Model



Match Pictures of a category

Geometric Blur Shape Feature



(A.) Berg & Malik '01

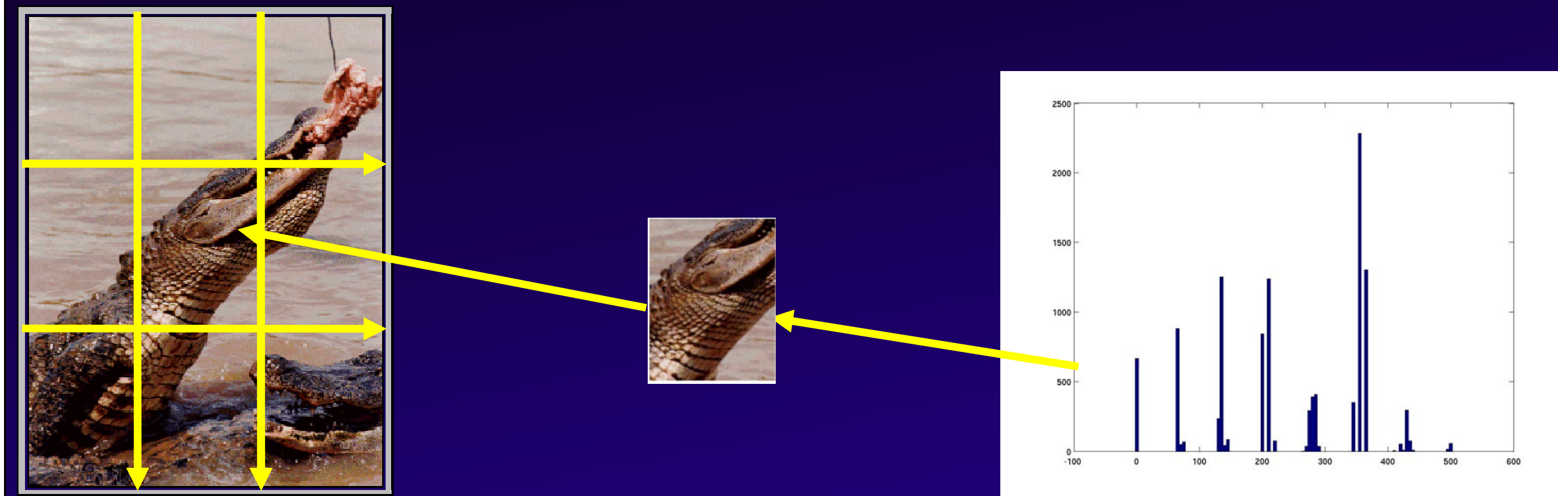
Sparse Signal Geometric Blur

Captures local shape, but allows for some deformation.
Robust to differences in intra category object shape.

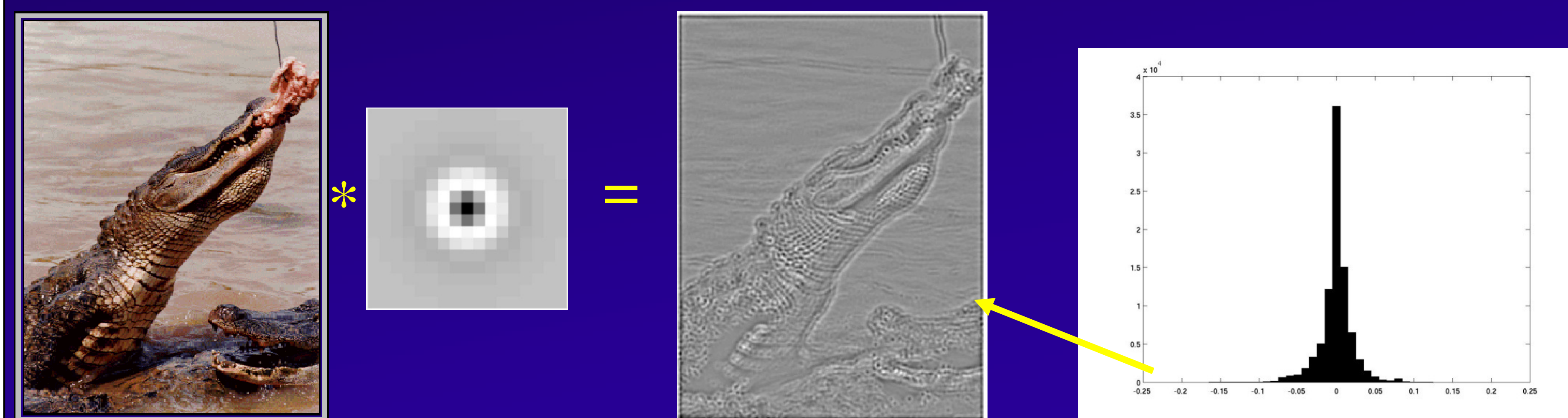
Used in current best object recognition systems
Zhang et al, CVPR 2006
Frome et al, NIPS 2006

Image Model (cont.)

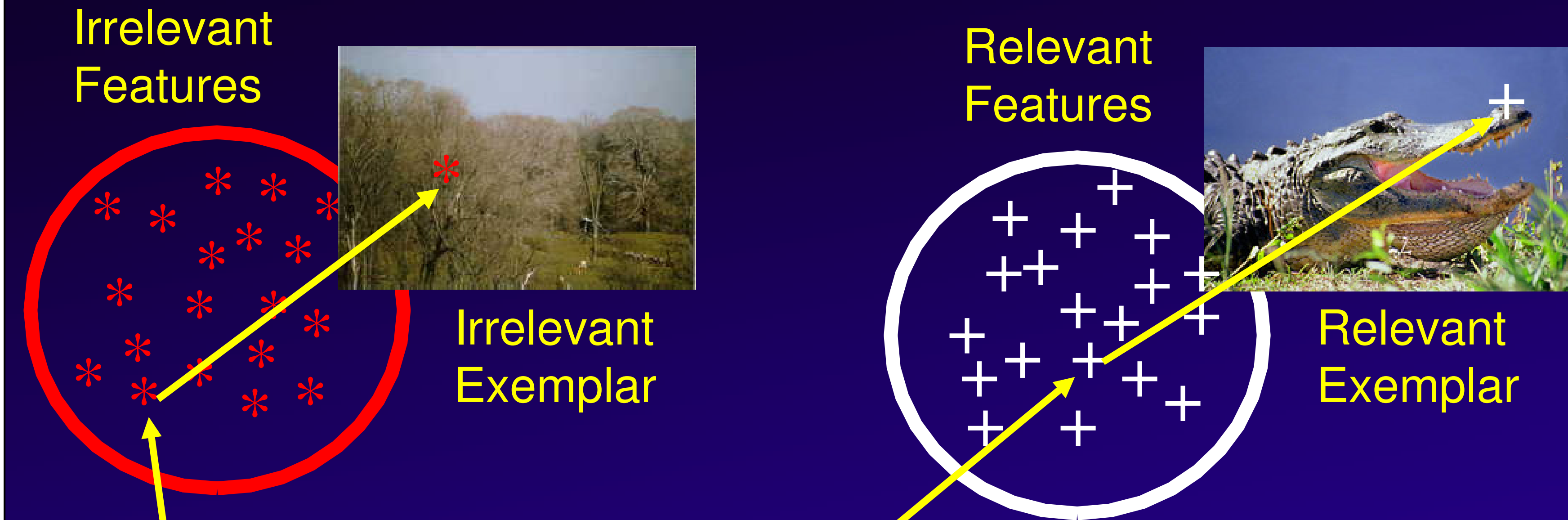
Color Features: Histogram of what colors appear in the image

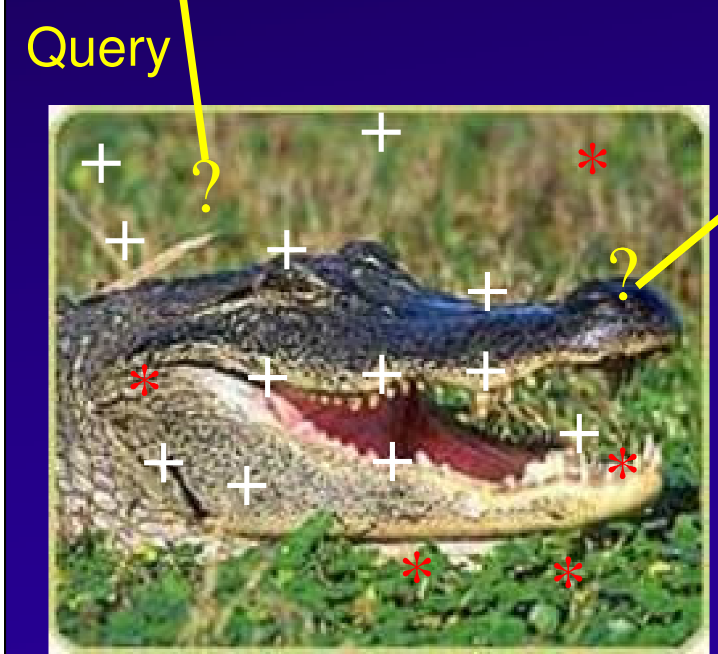


Texture Features: Histograms of 16 filters



Scoring Images





For each query feature apply a 1-nearest neighbor classifier. Sum votes for relevant class. Normalize.

Combine 4 cue scores (word, shape, color, texture) using a linear combination.

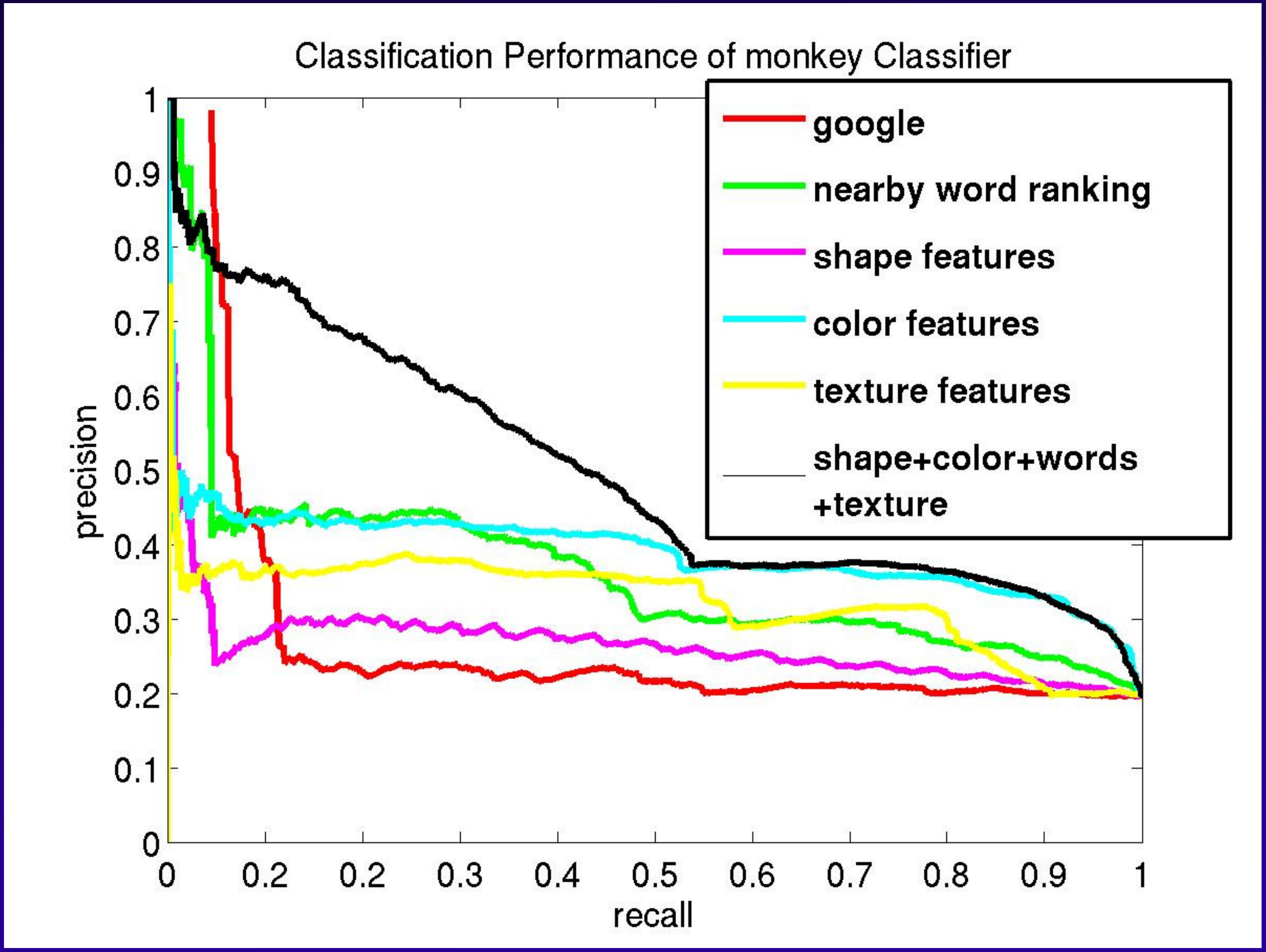
Classification Comparison

Words

Words + Picture

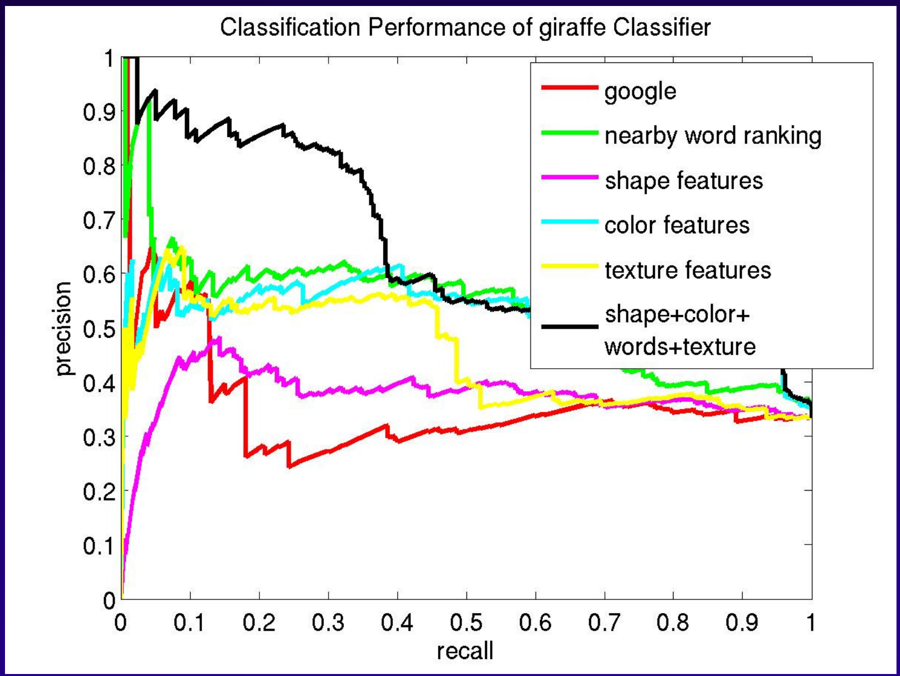
Cue Combination:

Monkey

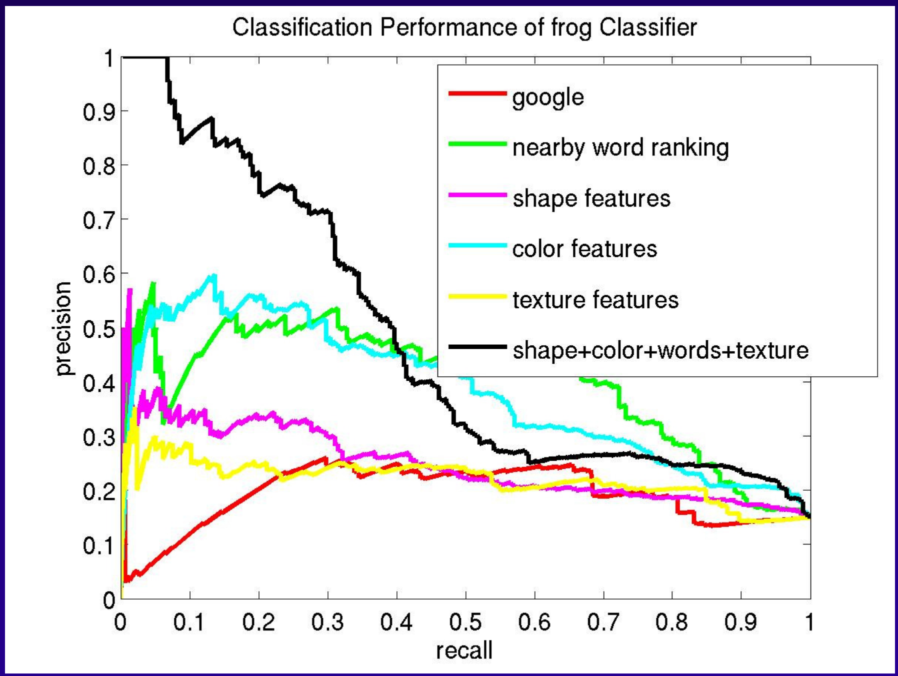


Cue Combination:

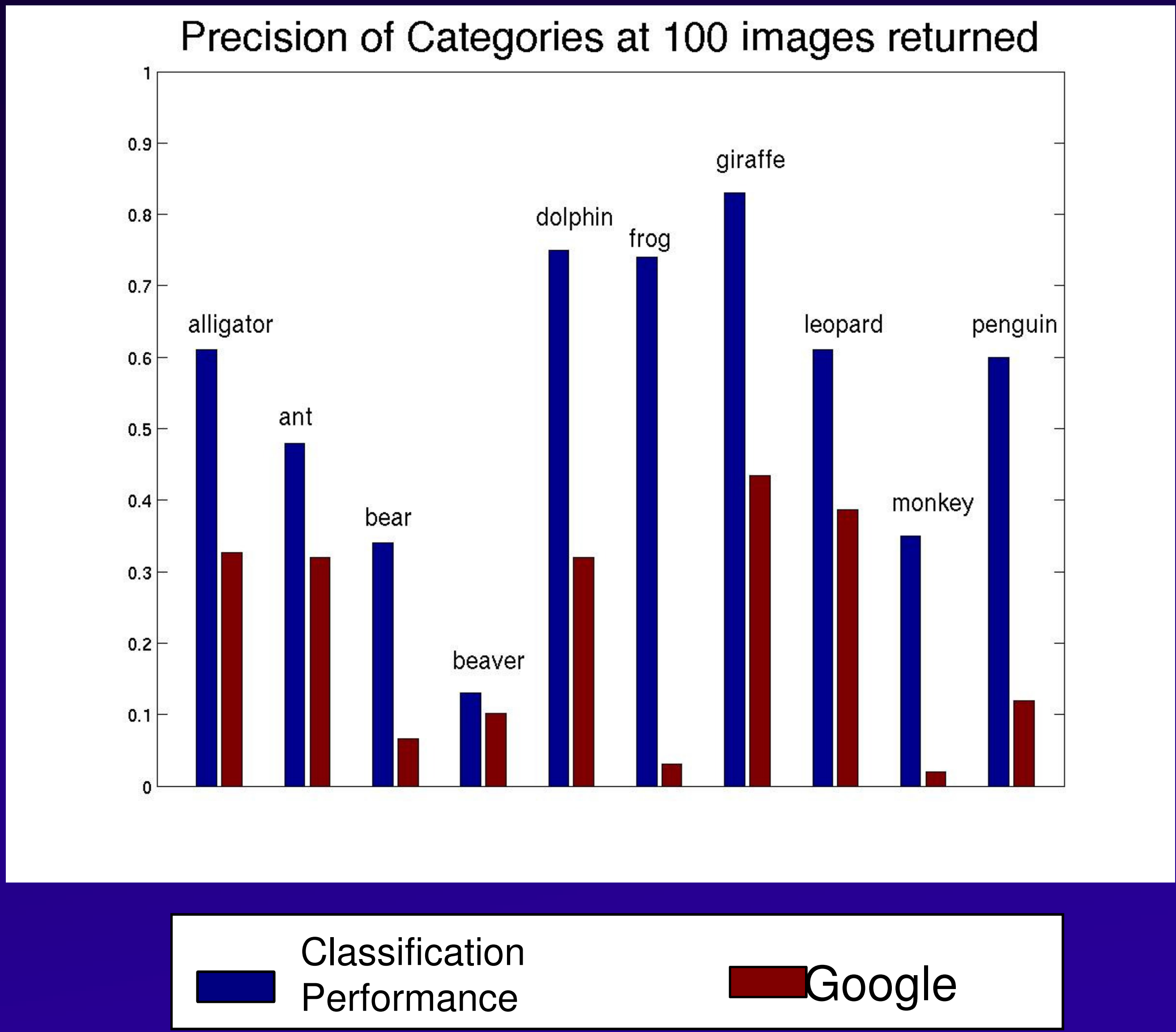
Giraffe



Frog

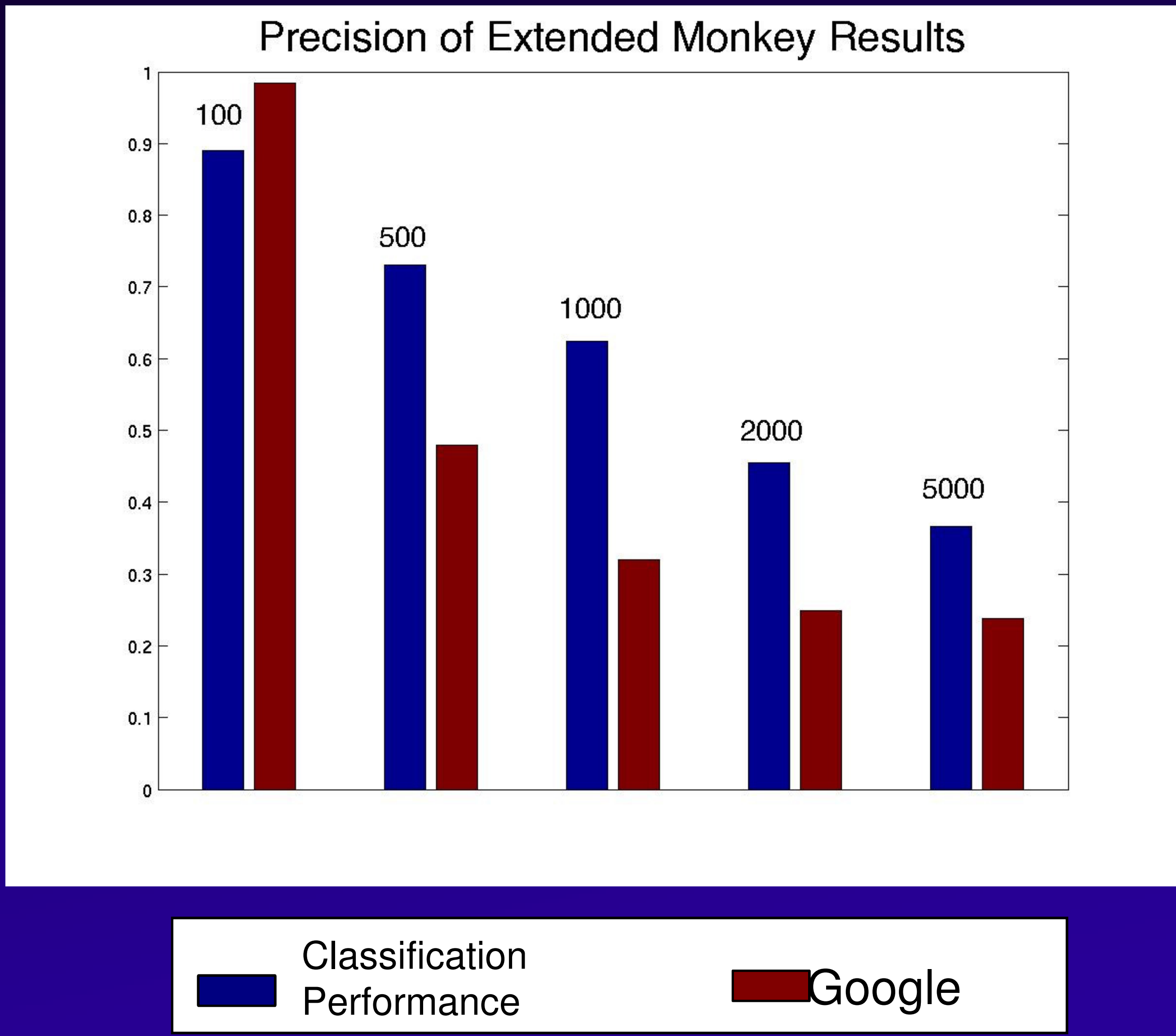


Re-ranking Precision



Re-ranking Precision

Monkey



Summary - Berg

Enormous amounts of data.

How to deal with it is still an open question.

We should combine words & pictures

Data opens up lots of new research problems!

WISDOM: An Unsupervised Model of Image Sense

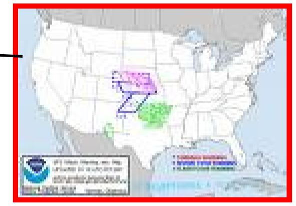
Kate Saenko and Trevor Darrell

Saenko and Darrell. Unsupervised learning of visual sense models for polysemous words. NIPS 2008,
Saenko and Darrell, Filtering Abstract Senses From Image Search Results, NIPS 2009.


bag
bag
bag
bag
... The Tote is the perfect example of two handbag design principles that ... The lines of this tote are incredibly sleek, but ... The semi buckles that form the handle attachments are ...
Noun
• bag, container (a flexible container with a single opening)
• bag, handbag, pocketbook, purse (a container used for carrying money and small personal items or accessories (especially by women))
• bag, travelling bag, suitcase (a rectangular container for carrying clothes)

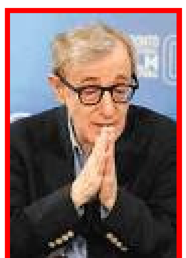
Image Sense Disambiguation

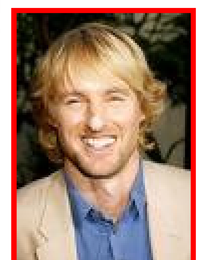
Hurricane, tornado watch



Celebrity watch

Watch out!

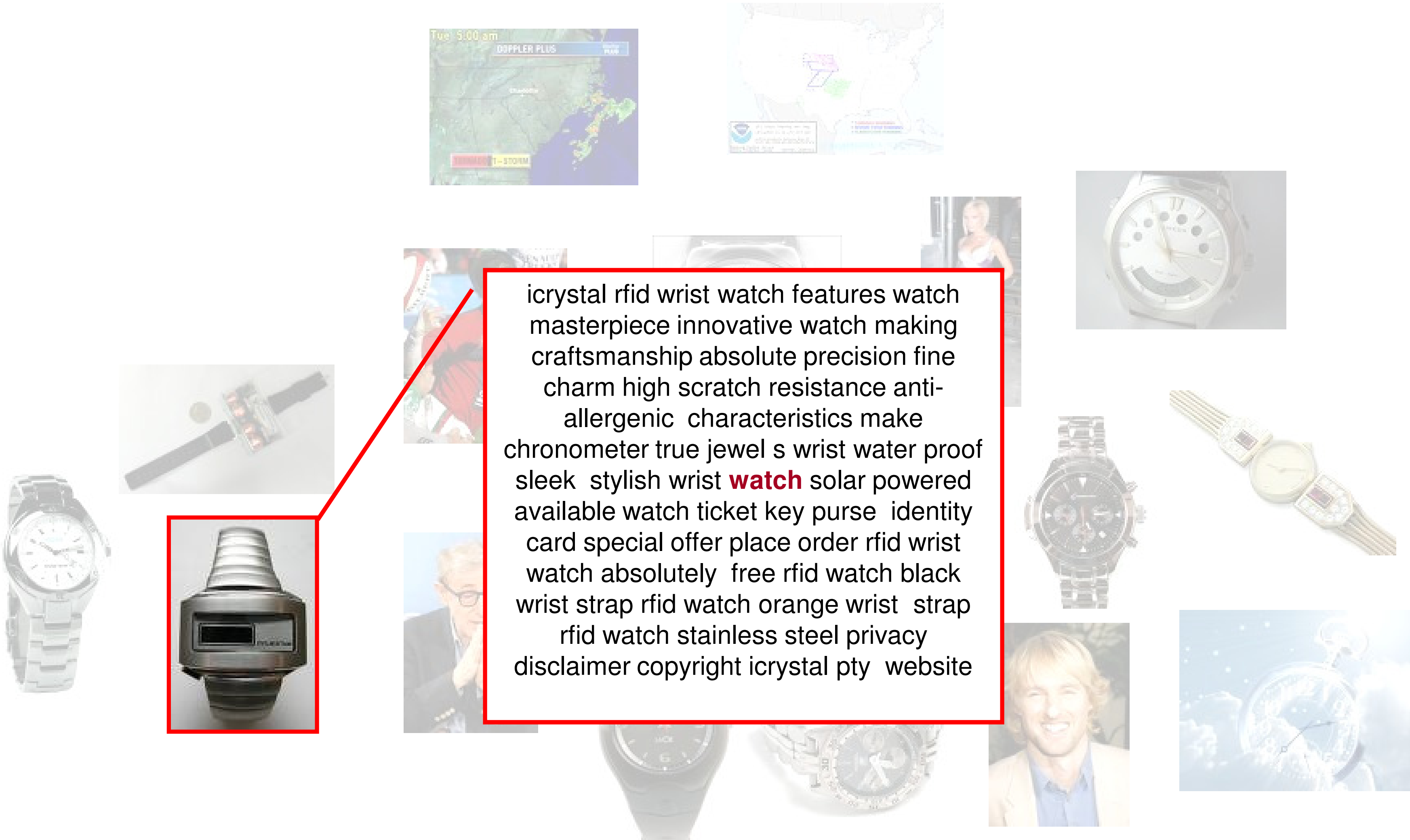




Would rather watch…

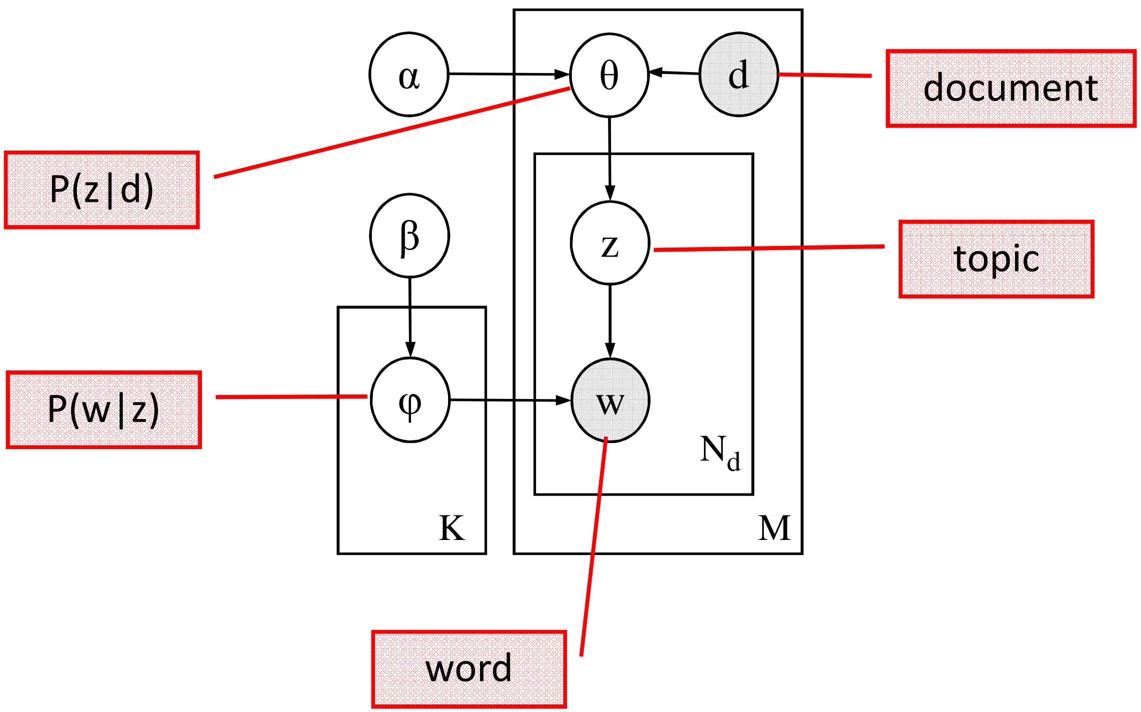
Suicide watch

Text contexts



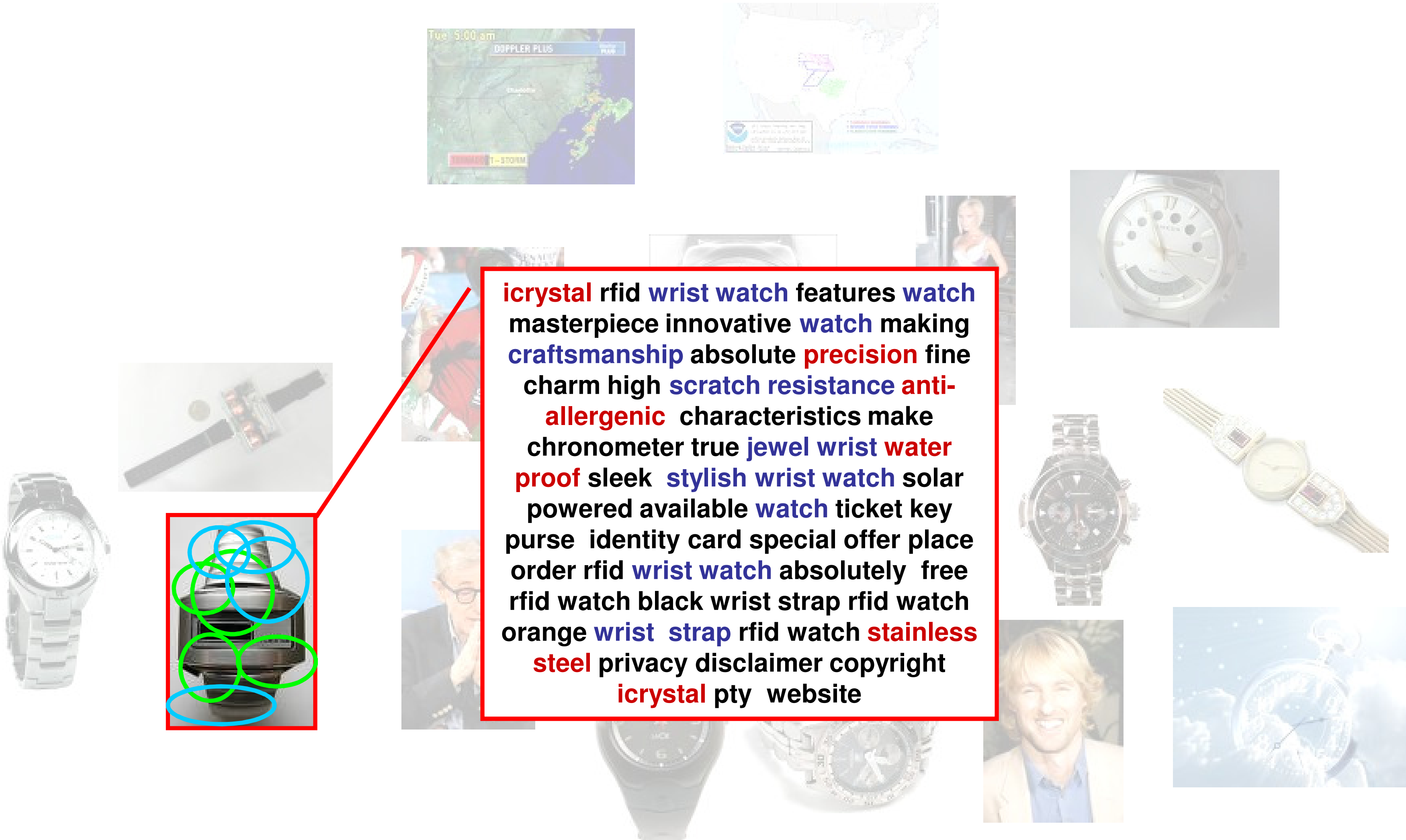
Latent Dirichlet allocation (LDA) (Blei et al. '03)

- One of several techniques for discovering latent dimensions in bag-of-words data



Topic 2
new
world
media
right
said
house
april
obama
islam
march
bush
war
american
time
…

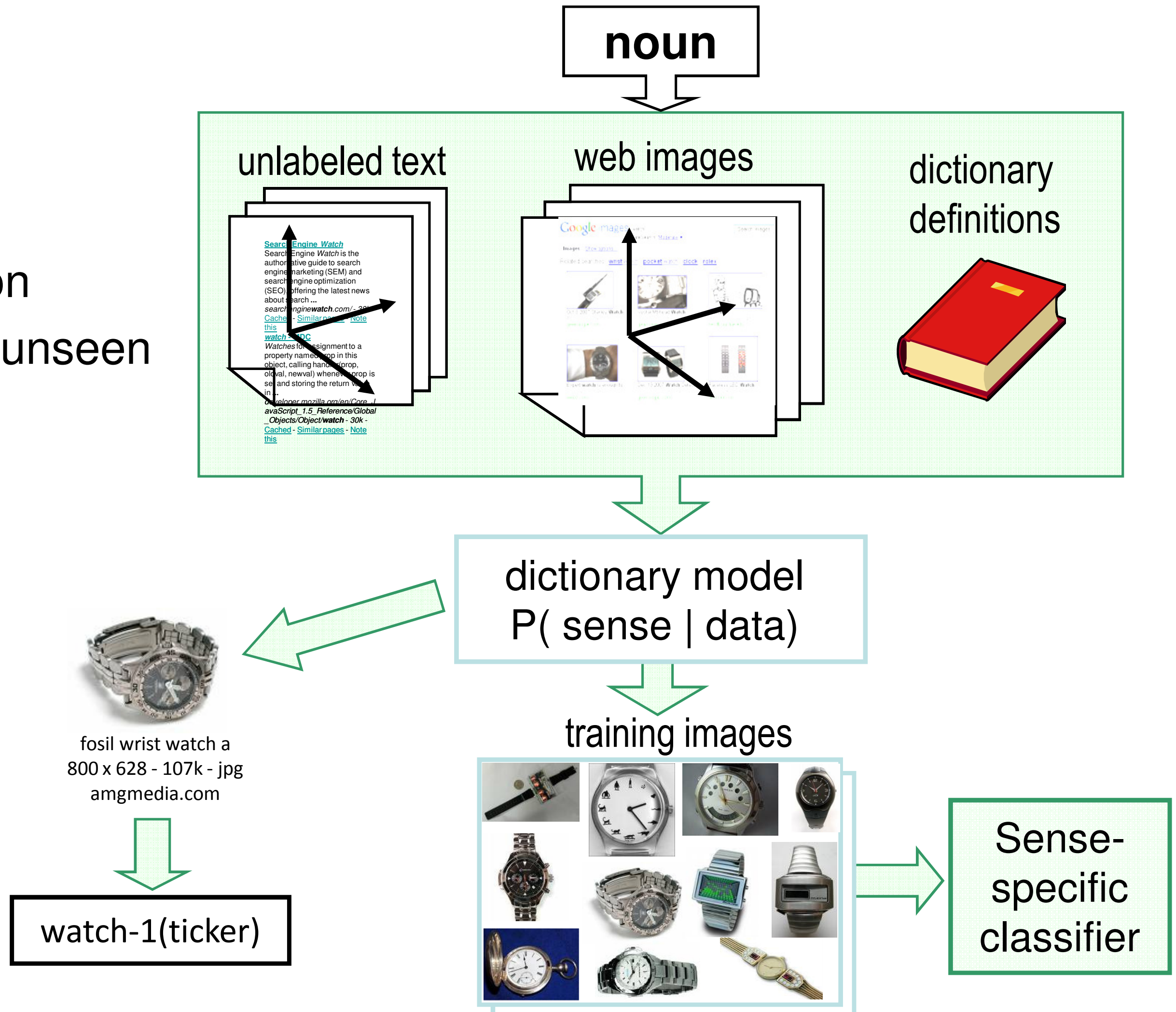
Latent Topics



Web Image Sense DictiOnary Model

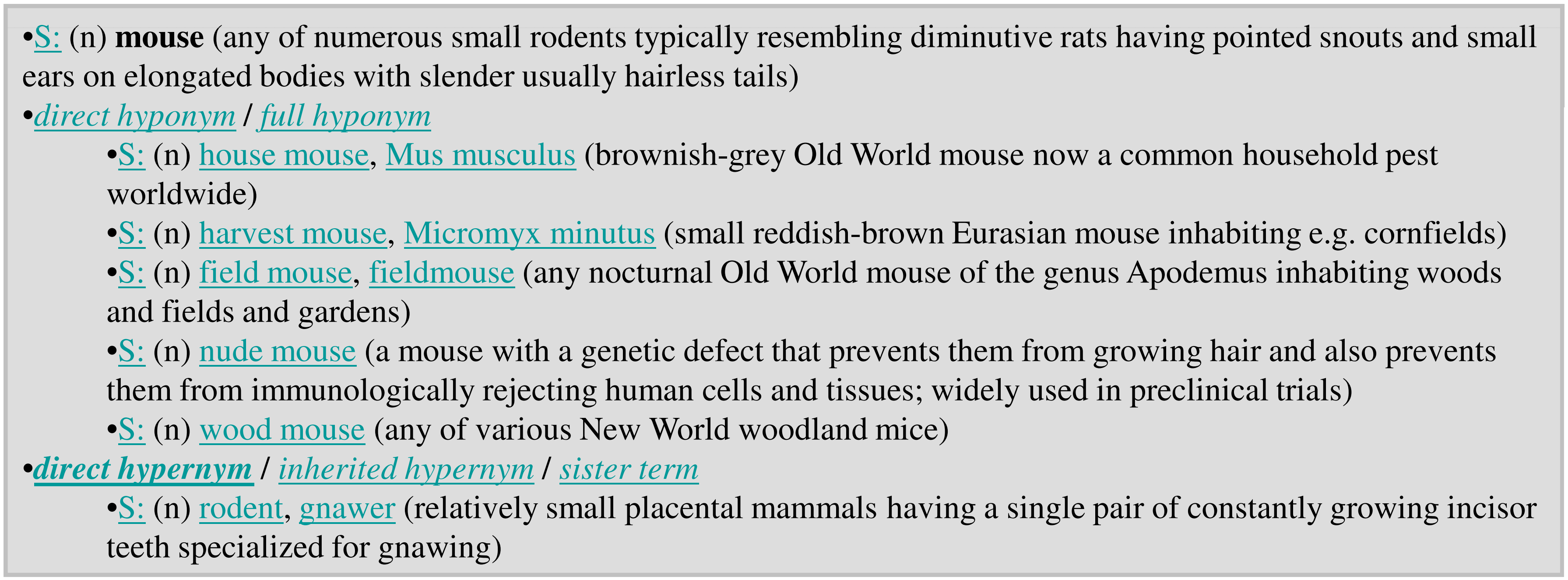
WISDOM does:

1. image sense disambiguation
2. dataset collection
3. classification of unseen images



WISDOM: Using dictionary entries to ground senses

- Use entry text to learn a probability distribution over words for that sense
- Problem: entries contain very little text
  - Expand by adding synonyms, example sentences, etc.
  - Still, very few words are covered!



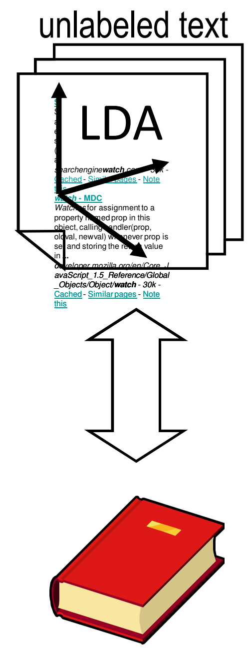
WISDOM: Probabilistic dictionary-based model

Main idea:

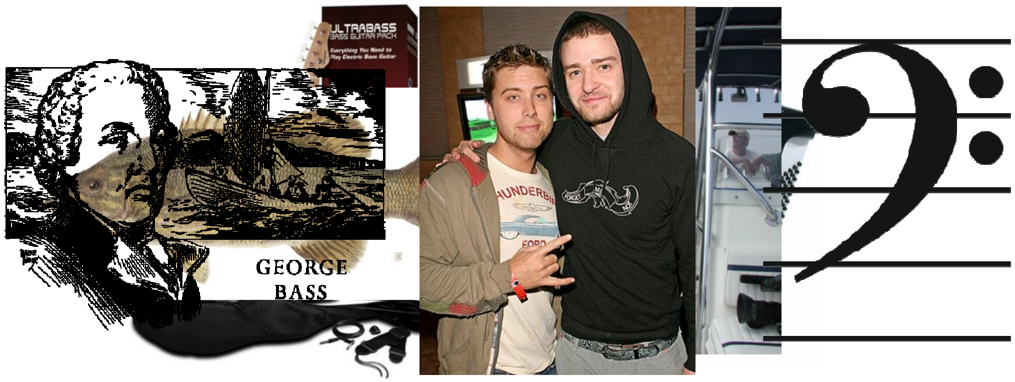
– Using LDA, learn latent sense-like dimensions on a large amount of related text,

– Model dictionary senses in LDA space:

  • Map image contexts to topics
  • Map topics to senses



Evaluation datasets

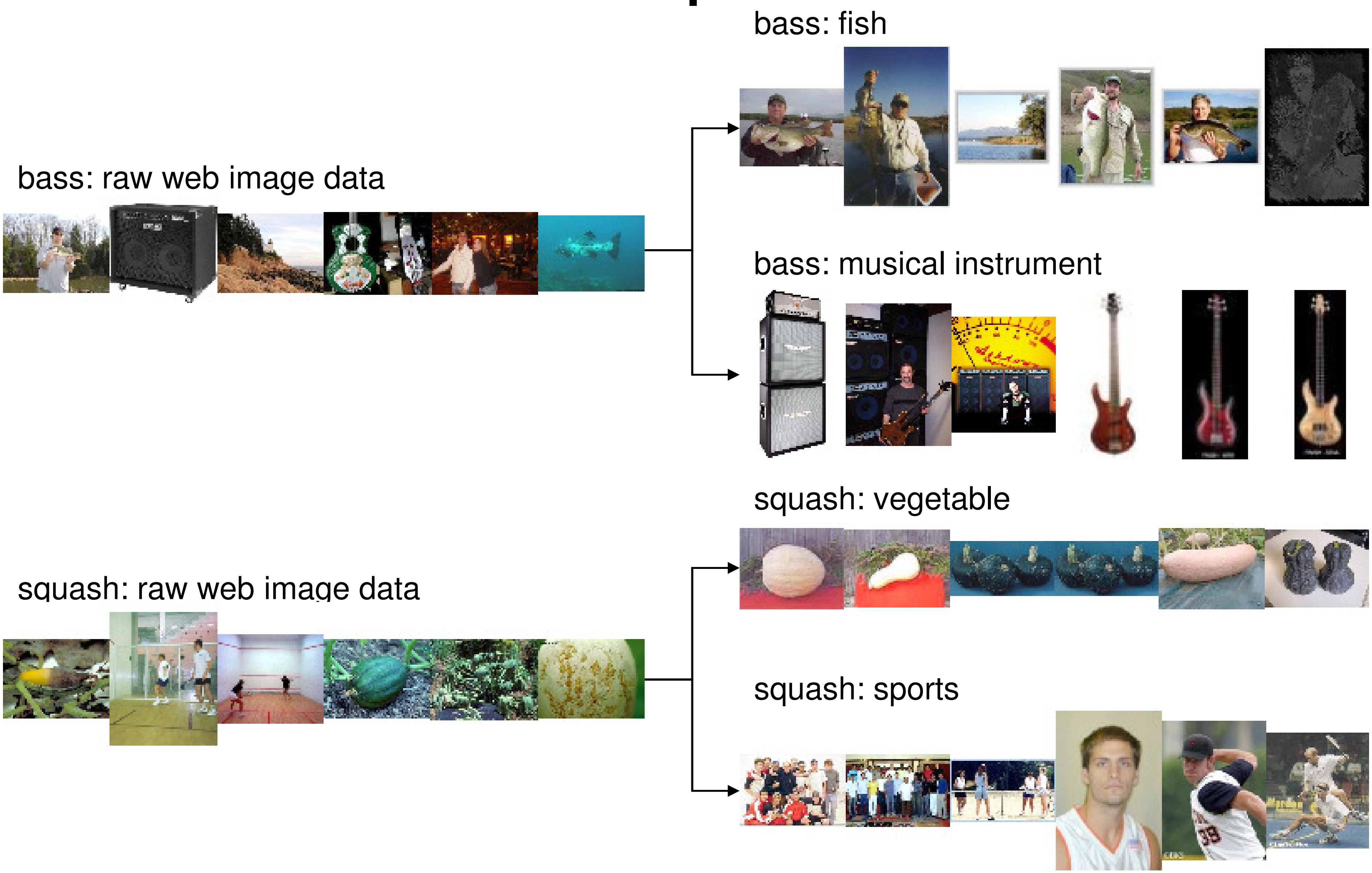


unrelated

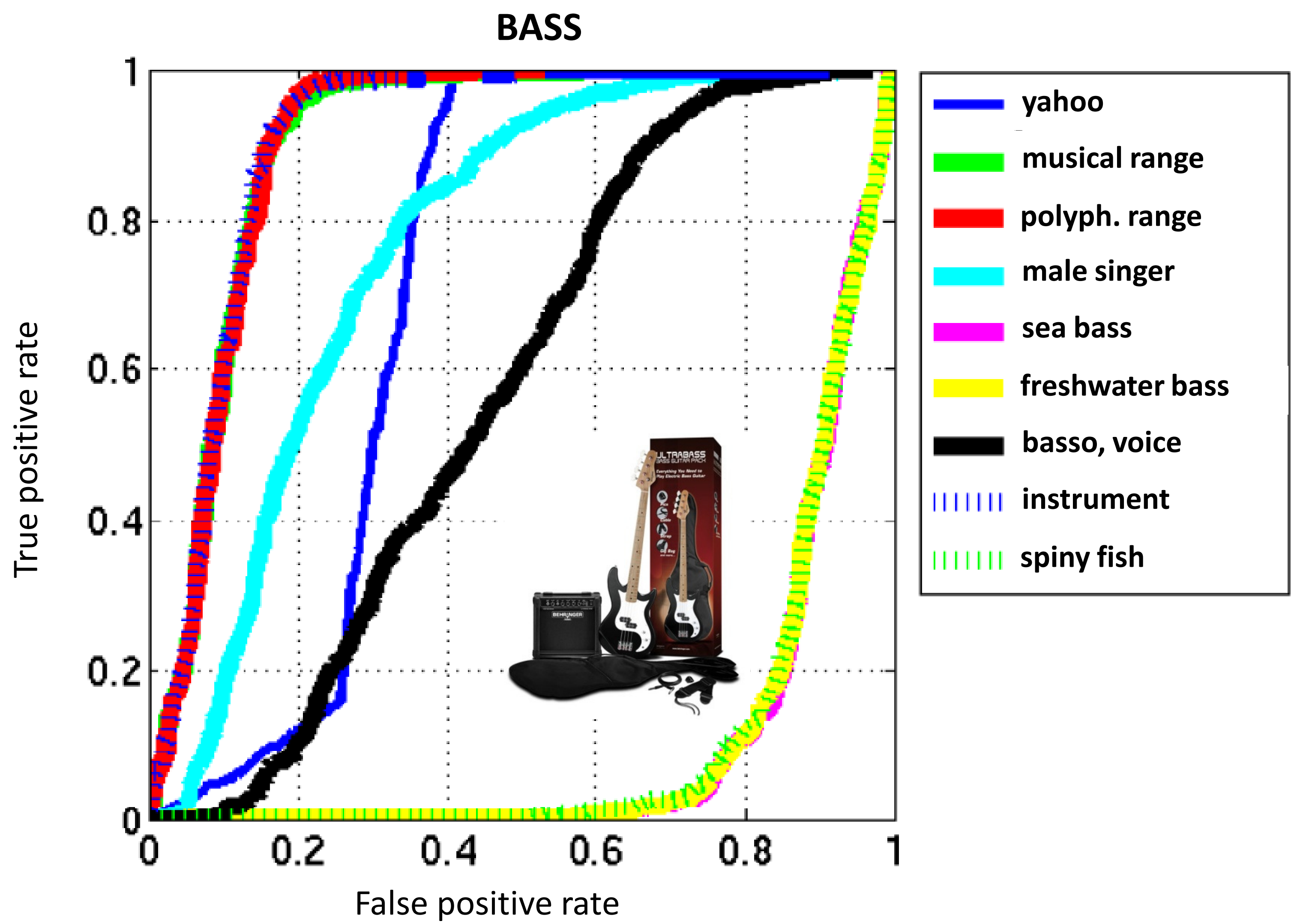
???

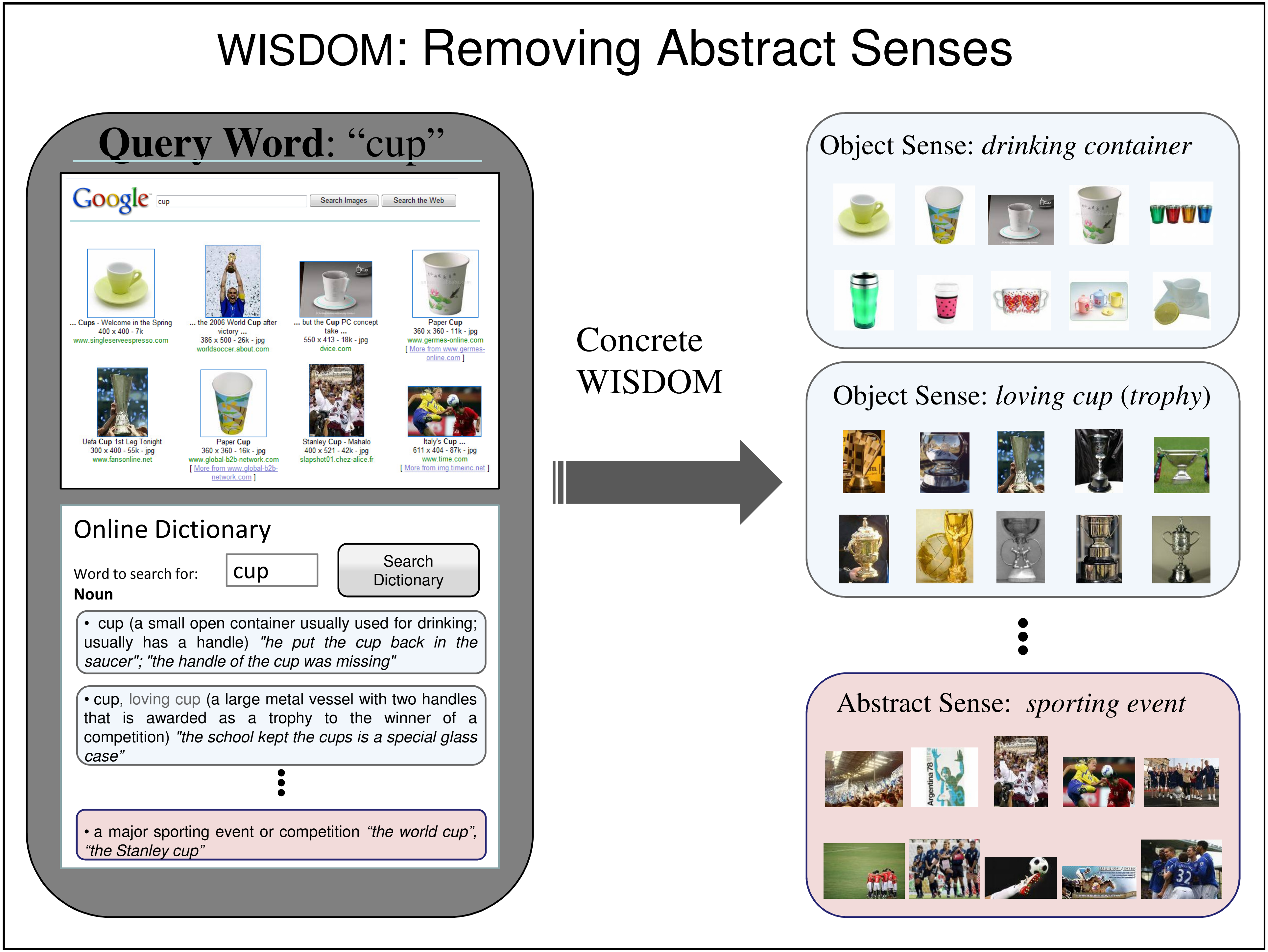
- Collected by querying YAHOO! Image Search
  - MIT-ISD: bass, face, mouse, speaker, watch
  - MIT-OFFICE: cellphone, fork, hammer, keyboard, mug, pliers, scissors, stapler, telephone, watch
  - UIUC-ISD: bass, crane, squash

ISD example results



ISD Results: ROC using each WordNet sense for BASS




WISDOM: Removing Abstract Senses
Query Word: "cup"
Google
cup
Search Images
Search the Web
... Cups - Welcome in the Spring
400 x 400 - 7k
www.singleserveespresso.com
... the 2006 World Cup after victory ...
386 x 500 - 26k - jpg
worldsoccer.about.com
... but the Cup PC concept take ...
550 x 413 - 18k - jpg
dvice.com
Paper Cup
360 x 360 - 11k - jpg
www.germes-online.com
[More from www.germes-online.com]
Uefa Cup 1st Leg Tonight
300 x 400 - 55k - jpg
www.fansonline.net
Paper Cup
360 x 360 - 16k - jpg
www.global-b2b-network.com
[More from www.global-b2b-network.com]
Stanley Cup - Mahalo
400 x 521 - 42k - jpg
slapshot01.chez-alice.fr
Italy's Cup ...
611 x 404 - 87k - jpg
www.time.com
[More from img.timeinc.net]
Online Dictionary
Word to search for:
cup
Search Dictionary
Noun
• cup (a small open container usually used for drinking; usually has a handle) "he put the cup back in the saucer"; "the handle of the cup was missing"
• cup, loving cup (a large metal vessel with two handles that is awarded as a trophy to the winner of a competition) "the school kept the cups is a special glass case"
• a major sporting event or competition "the world cup", "the Stanley cup"
Concrete WISDOM
Object Sense: drinking container
Object Sense: loving cup (trophy)
Abstract Sense: sporting event
Argentina 78

Results: Filtering visual senses

Yahoo Search: "telephone"

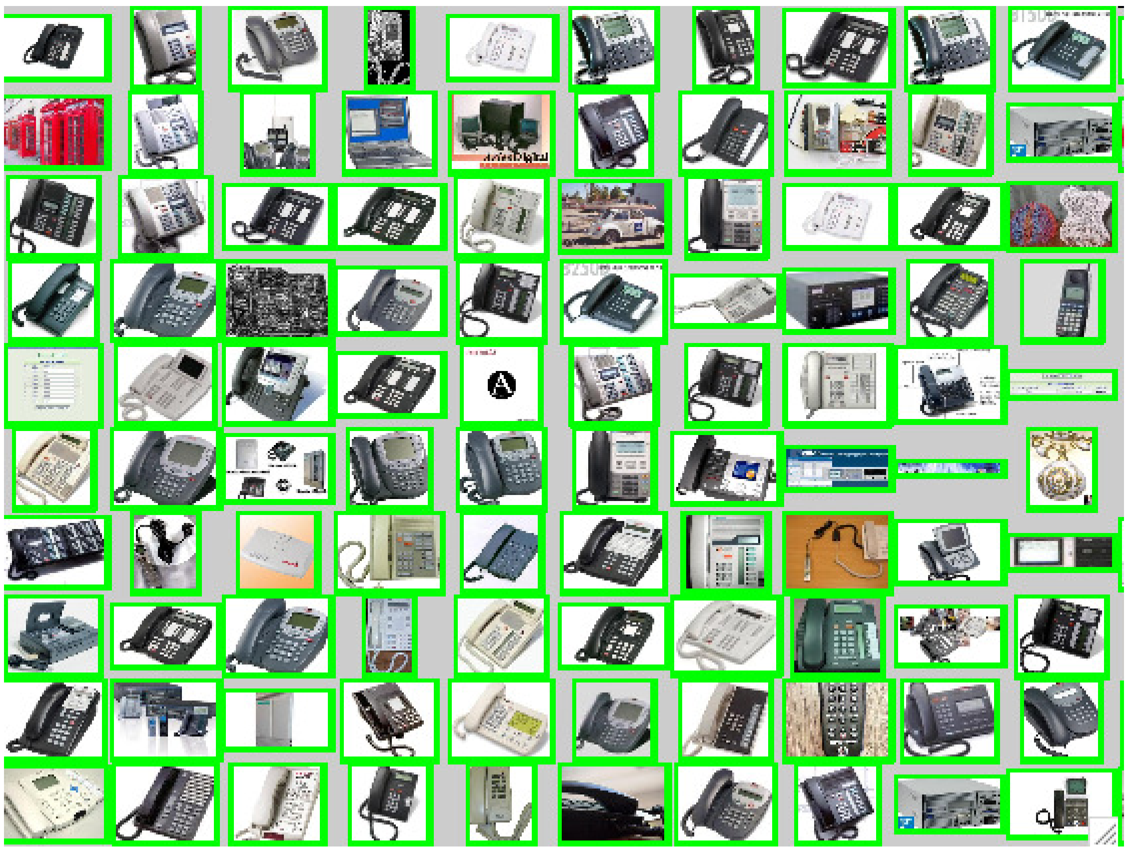
DICTIONARY

1: (n) **telephone**, phone, telephone set (electronic equipment that converts sound into electrical signals that can be transmitted over distances and then converts received signals back into sounds)

2: (n) **telephone**, telephony (transmitting speech at a distance)

Results: Filtering visual senses

Artifact sense: "telephone"



DICTIONARY

1: (n) **telephone**, phone, telephone set (electronic equipment that converts sound into electrical signals that can be transmitted over distances and then converts received signals back into sounds)

2: (n) **telephone**, telephony (transmitting speech at a distance)

Lecture Summary

- The web contains unlimited, but extremely noisy object category data
- The text surrounding the image on the web page is an important recognition cue
- Topic models (pLSA, LDA, HDP, etc.) are useful for discovering objects in images and object senses in text
- Bootstrap model from small amount of labeled or weakly labeled data
- Still an open research problem!

Next Lectures

- Today: learning object categories from the web
  - LSA and LDA models
  - Harvesting training data from the web
  - Exploiting image and text
- Tues. Oct. 20th: Generative models
  - Condensation
  - ISM
  - Transformed-HDPs
  - More Context…
- Thurs. Oct. 22nd: Advanced BOW kernels
  - Pyramid and spatial-pyramid match
  - Multi-kernel learning
  - Latent-part SVM models

Bibliography

- R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," ICCV vol. 2, 2005, pp.1816-1823 Vol. 2. http://dx.doi.org/10.1109/ICCV.2005.142
- Sivic, Russell et al. “Unsupervised object category discovery” ICCV ‘05
- Fei-Fei and Perona “Scene analysis using LDA” CVPR ‘05
- T. Berg and D. Forsyth, "Animals on the Web". In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Washington, DC, 1463-1470. http://dx.doi.org/10.1109/CVPR.2006.57
- L.-J. Li, G. Wang, and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp.1-8. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4270073
- K. Saenko and T. Darrell, "Unsupervised Learning of Visual Sense Models for Polysemous Words". Proc. NIPS, December 2008, Vancouver, Canada. http://people.csail.mit.edu/saenko/saenko_nips08.pdf

Additional reading

- N.Loeff, C.O. Alm, D.A. Forsyth, “Discriminating image senses by clustering with multimodal features”. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 547–554, Sydney, July 2006 [PDF]
- G. Wang and D. Forsyth, "Object image retrieval by exploiting online knowledge resources". *IEEE Computer Vision and Pattern Recognition (CVPR).* 2008. [PDF]
- F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1-8. http://dx.doi.org/10.1109/ICCV.2007.4409099
- D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM Press, 2003, pp. 127-134. http://dx.doi.org/10.1145/860435.860460
- P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*. London, UK: Springer-Verlag, 2002, pp. 97-112. http://portal.acm.org/citation.cfm?id=645318.649254

Slide Credits

- As attributed…