6.869 Advances in Computer Vision Spring 2010

http://people.csail.mit.edu/torralba/courses/6.869/6.869. computervision.htm

Lecture 21

Bayes

Project presentations

May 5 1pm – 2:30pm

2:30pm – 4pm

Complex motion

International Journal of Computer Vision 51(2), 91–109, 2003

Dynamic Textures

GIANFRANCO DORETTO
Computer Science Department, University of California, Los Angeles, CA 90095
doretto@cs.ucla.edu

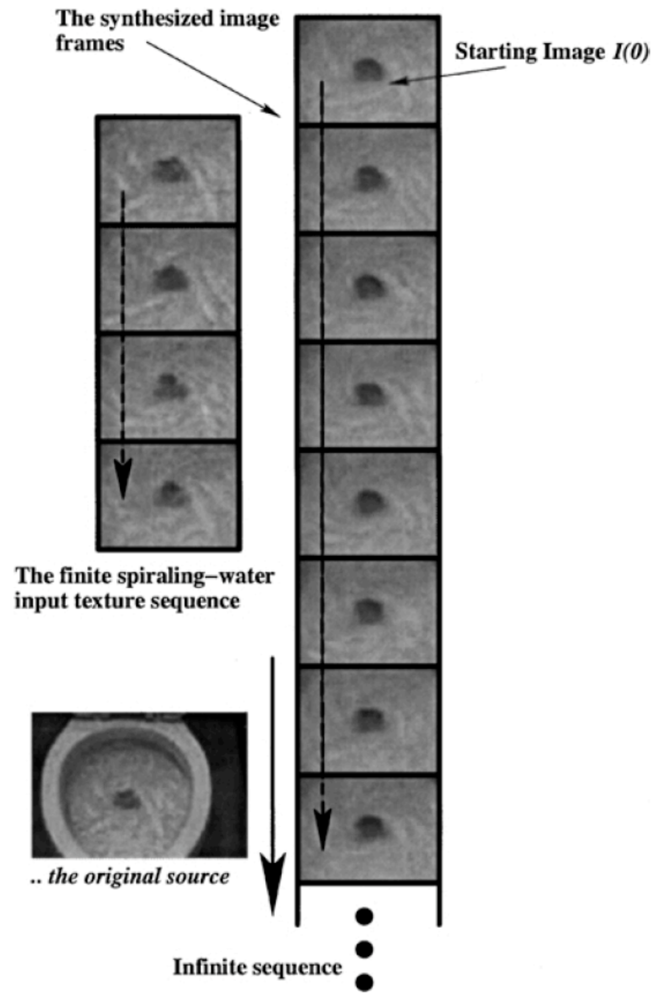
ALESSANDRO CHIUSO
Dipartimento di Ingegneria dell'Informazione, Università di Padova, Italy 35131
chiuso@dei.unipd.it

YING NIAN WU
Statistics Department, University of California, Los Angeles, CA 90095
ywu@stat.ucla.edu

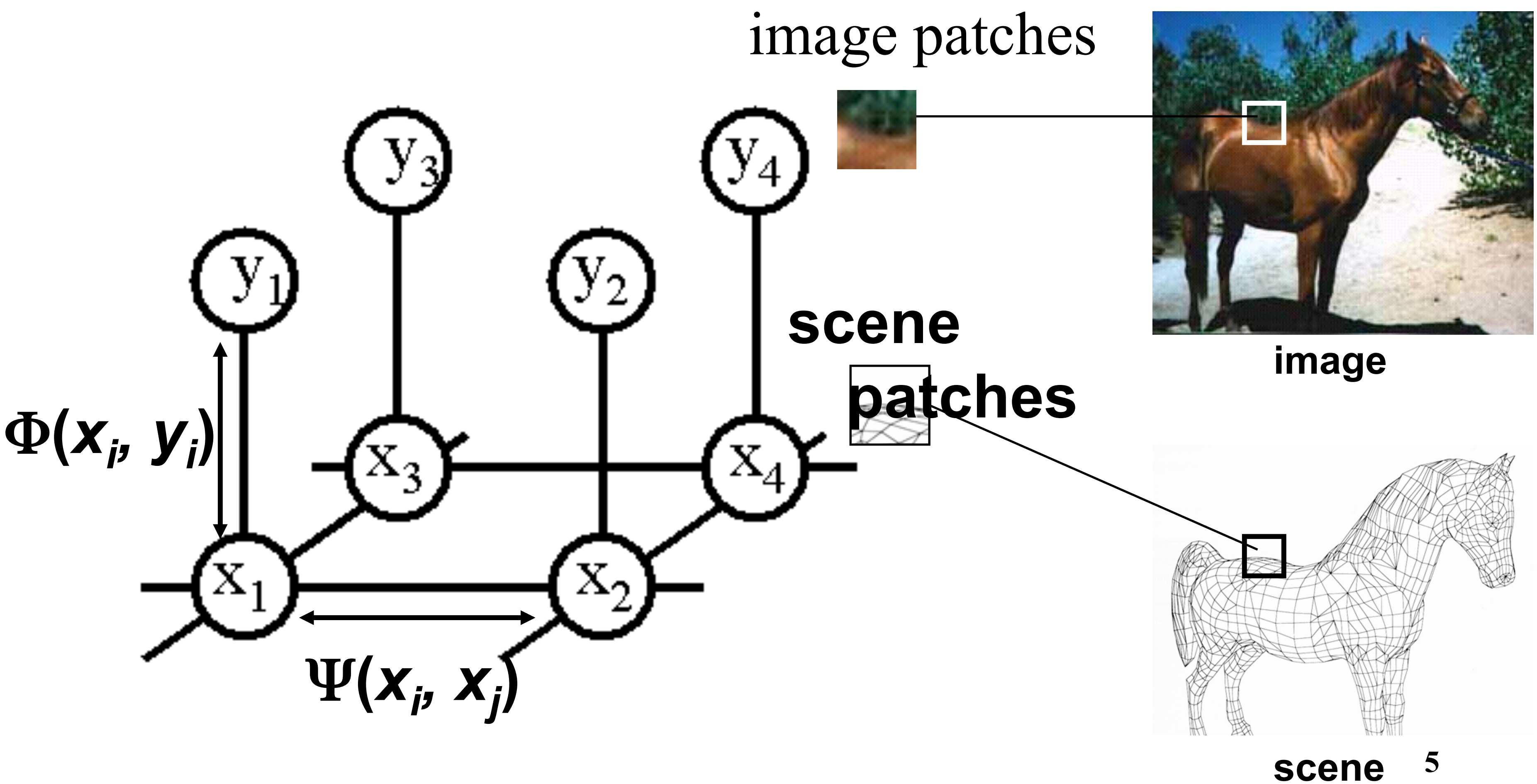
STEFANO SOATTO
Computer Science Department, University of California, Los Angeles, CA 90095
soatto@ucla.edu

Received May 1, 2001; Revised February 21, 2002; Accepted July 3, 2002

$$\begin{cases} x(t+1) = Ax(t) + v(t) & v(t) \sim \mathcal{N}(0, Q); \quad x(0) = x_0 \\ y(t) = Cx(t) + w(t) & w(t) \sim \mathcal{N}(0, R) \end{cases}$$



MRF nodes as patches



Network joint probability

$$P(x, y) = \frac{1}{Z} \prod_{i,j} \Psi(\underbrace{x_i, x_j}_{\text{neighboring scene nodes}}) \prod_i \Phi(x_i, y_i)$$

scene

image

Scene-scene compatibility function

neighboring scene nodes

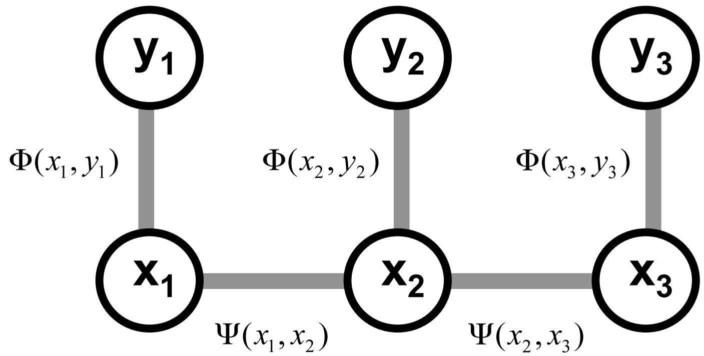
Image-scene compatibility function

local observations

In order to use MRFs:

- Given observations y, and the parameters of the MRF, how <u>infer</u> the hidden variables, x?
- How <u>learn</u> the parameters of the MRF?

Derivation of belief propagation



minimum mean square error (MMSE)

$$x_{1MMSE} = \underset{x_1}{\text{mean}} \; \underset{x_2}{\text{sum}} \; \underset{x_3}{\text{sum}} \; P(x_1, x_2, x_3, y_1, y_2, y_3)$$

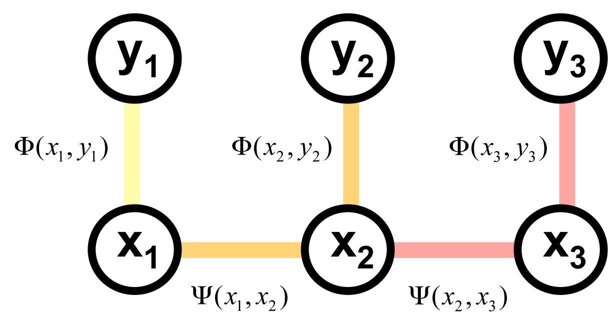
The posterior factorizes

$$x_{1MMSE} = \underset{x_1}{\text{mean}}\ \underset{x_2}{\text{sum}}\ \underset{x_3}{\text{sum}}\ P(x_1, x_2, x_3, y_1, y_2, y_3)$$

$$= \underset{x_1}{\text{mean}}\ \underset{x_2}{\text{sum}}\ \underset{x_3}{\text{sum}}\ \Phi(x_1, y_1)$$

$$\Phi(x_2, y_2)\ \Psi(x_1, x_2)$$

$$\Phi(x_3, y_3)\ \Psi(x_2, x_3)$$

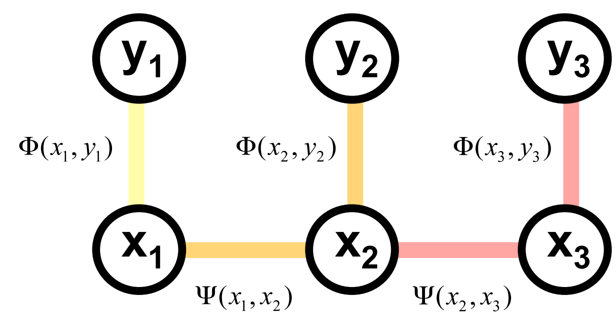


Propagation rules

$$x_{1MMSE} = \operatorname*{mean}_{x_1} \operatorname*{sum}_{x_2} \operatorname*{sum}_{x_3} P(x_1, x_2, x_3, y_1, y_2, y_3)$$

$$x_{1MMSE} = \operatorname*{mean}_{x_1} \operatorname*{sum}_{x_2} \operatorname*{sum}_{x_3} \Phi(x_1, y_1) \Phi(x_2, y_2) \Psi(x_1, x_2) \Phi(x_3, y_3) \Psi(x_2, x_3)$$

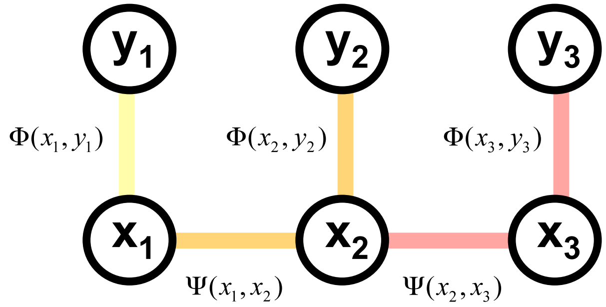
$$x_{1MMSE} = \operatorname*{mean}_{x_1} \Phi(x_1, y_1) \operatorname*{sum}_{x_2} \Phi(x_2, y_2) \Psi(x_1, x_2) \operatorname*{sum}_{x_3} \Phi(x_3, y_3) \Psi(x_2, x_3)$$



Propagation rules

$$x_{1MMSE} = \operatorname*{mean}_{x_1} \Phi(x_1, y_1)$$
$$\operatorname*{sum}_{x_2} \Phi(x_2, y_2)\, \Psi(x_1, x_2)$$
$$\operatorname*{sum}_{x_3} \Phi(x_3, y_3)\, \Psi(x_2, x_3)$$

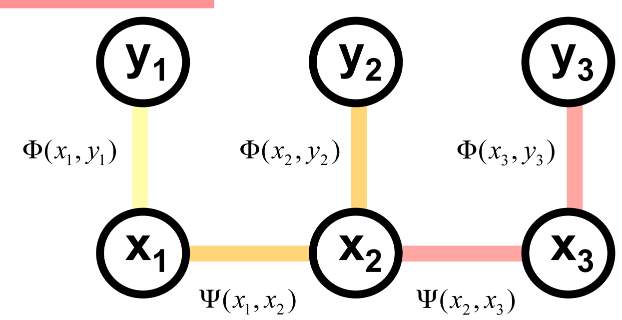
$$M_1^2(x_1) = \operatorname*{sum}_{x_2} \Psi(x_1, x_2)\; \Phi(x_2, y_2)\; M_2^3(x_2)$$



Propagation rules

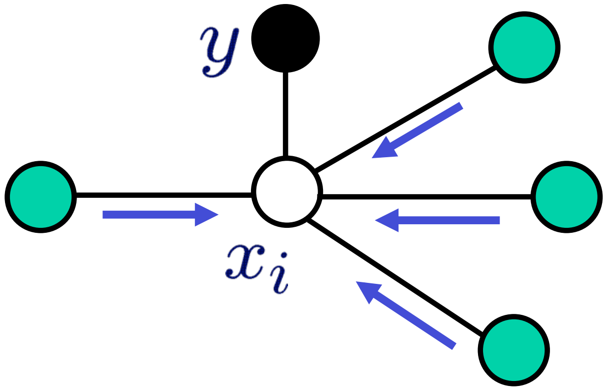
$$x_{1MMSE} = \operatorname*{mean}_{x_1} \Phi(x_1, y_1)$$
$$\operatorname*{sum}_{x_2} \Phi(x_2, y_2)\, \Psi(x_1, x_2)$$
$$\operatorname*{sum}_{x_3} \Phi(x_3, y_3)\, \Psi(x_2, x_3)$$

$$\boxed{M_1^2(x_1)} = \operatorname*{sum}_{x_2} \Psi(x_1, x_2)\ \Phi(x_2, y_2)\ \boxed{M_2^3(x_2)}$$



Belief Propagation

BELIEFS: Approximate posterior marginal distributions

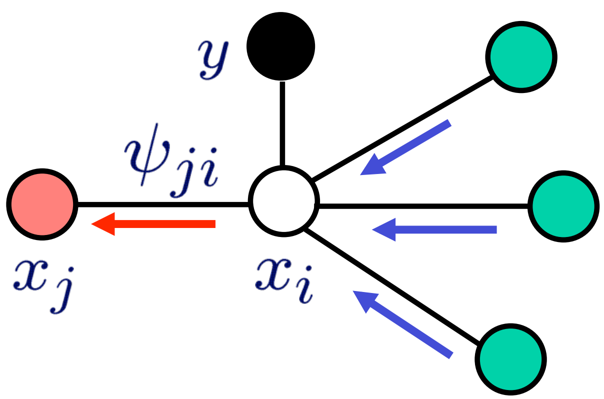


$$\hat{p}(x_i \mid y) \propto \psi_i(x_i, y) \prod_{k \in \Gamma(i)} m_{ki}(x_i)$$

$\Gamma(i) \longrightarrow$ ***neighborhood*** **of node *i***

MESSAGES: Approximate sufficient statistics

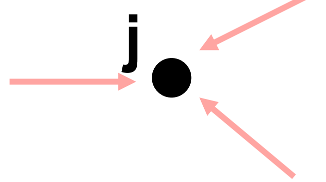
$$m_{ij}(x_j) \propto \int_{x_i} \psi_{j,i}(x_j, x_i) \psi_i(x_i, y) \prod_{k \in \Gamma(i) \setminus j} m_{ki}(x_i)\, dx_i$$



I. Belief Update (Message Product)

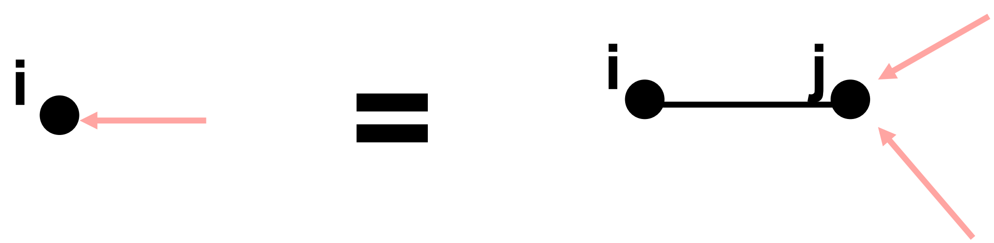
II. Message Propagation (Convolution)

Belief, and message updates

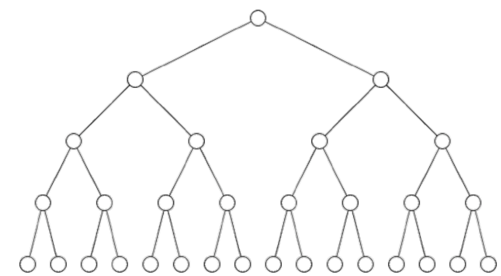


$$b_j(x_j) = \prod_{k \in N(j)} M_j^k(x_j)$$

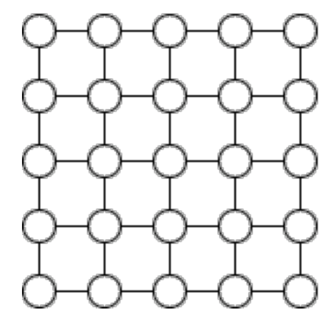
$$M_i^j(x_i) = \sum_{x_j} \psi_{\mathrm{ij}}(x_i, x_j) \prod_{k \in N(j) \setminus i} M_j^k(x_j)$$



Justifications for BP



- **Gives *exact* marginals for trees**
  - → ***Optimal estimates***
  - → ***Confidence measures***



- For general graphs, *loopy BP* has excellent empirical performance in many applications
- Recent theory provides some guarantees:
  - Statisical physics: *variational method*
    (Yedidia, Freeman, & Weiss)
  - BP as reparameterization: *error bounds*
    (Wainwright, Jaakkola, & Willsky)
  - Many others…

Belief propagation: the nosey neighbor rule

"Given everything that I know, here's what I think you should think"

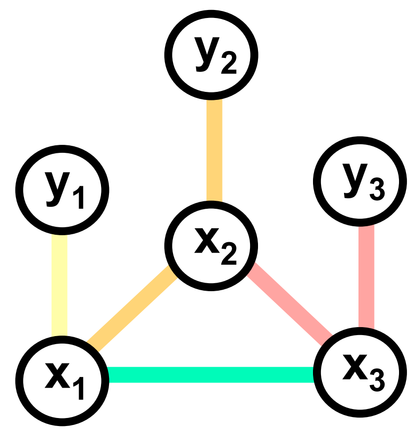
(Given the probabilities of my being in different states, and how my states relate to your states, here's what I think the probabilities of your states should be)

No factorization with loops!

$$x_{1MMSE} = \boxed{\operatorname*{mean}_{x_1}} \Phi(x_1, y_1)$$

$$\boxed{\operatorname*{sum}_{x_2}} \Phi(x_2, y_2)\, \Psi(x_1, x_2)$$

$$\boxed{\operatorname*{sum}_{x_3}} \Phi(x_3, y_3)\, \Psi(x_2, x_3) \;\; \boldsymbol{\Psi(x_1, x_3)}$$



References on BP and GBP

- J. Pearl, 1985
  - classic
- Y. Weiss, NIPS 1998
  - Inspires application of BP to vision
- W. Freeman et al learning low-level vision, IJCV 1999
  - Applications in super-resolution, motion, shading/paint discrimination
- H. Shum et al, ECCV 2002
  - Application to stereo
- M. Wainwright, T. Jaakkola, A. Willsky
  - Reparameterization version
- J. Yedidia, AAAI 2000
  - The clearest place to read about BP and GBP.

Interpreting images by propagating Bayesian beliefs

Yair Weiss
Dept. of Brain and Cognitive Sciences
Massachusetts Institute of Technology
E10-120, Cambridge, MA 02139, USA

In this paper we show that an architecture in which *Bayesian Beliefs* about image properties are propagated between neighboring units yields convergence times which are several orders of magnitude faster than traditional methods and avoids local minima. In particular our architecture is non-iterative in the sense of Marr [5]: at every time step, the local estimates at a given location are optimal given the information which has already been propagated to that location. We illustrate the algorithm's performance on real images and compare it to several existing methods.

$$J(Y) = \sum_k w_k (y_k - y_k^*)^2 + \lambda \sum_i (y_i - y_{i+1})^2$$

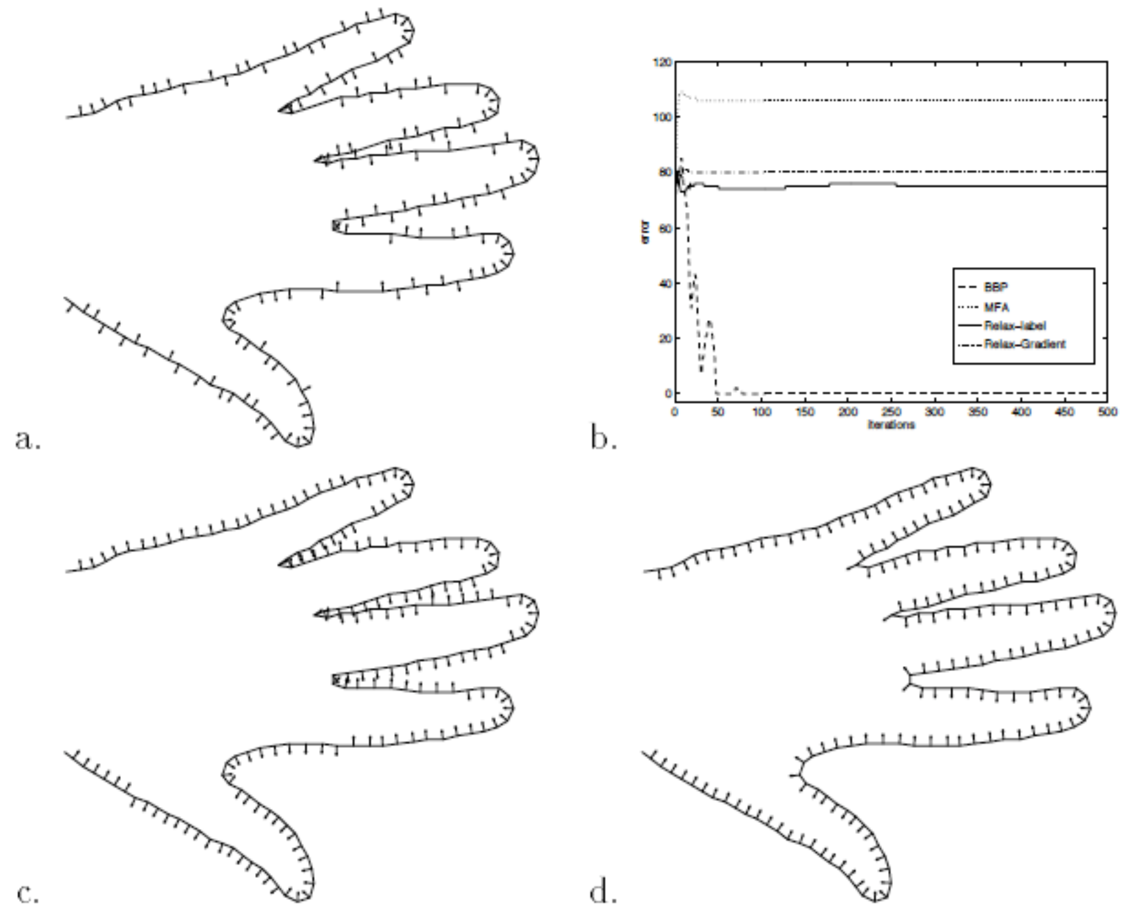


Figure 4: **a.** Local estimate of DOF along the contour. **b.** Performance of Hopfield,gradient descent, relaxation labeling and BBP as a function of time. BBP is the only method that converges to the global minimum. **c.** DOF estimate of Hopfield net after convergence. **d.** DOF estimate of BBP after convergence.

Random Fields for segmentation

I = Image pixels (observed)
h = foreground/background labels (hidden) – one label per pixel
θ = Parameters

$$\underbrace{p(h \mid I, \theta)}_{\textbf{Posterior}}$$

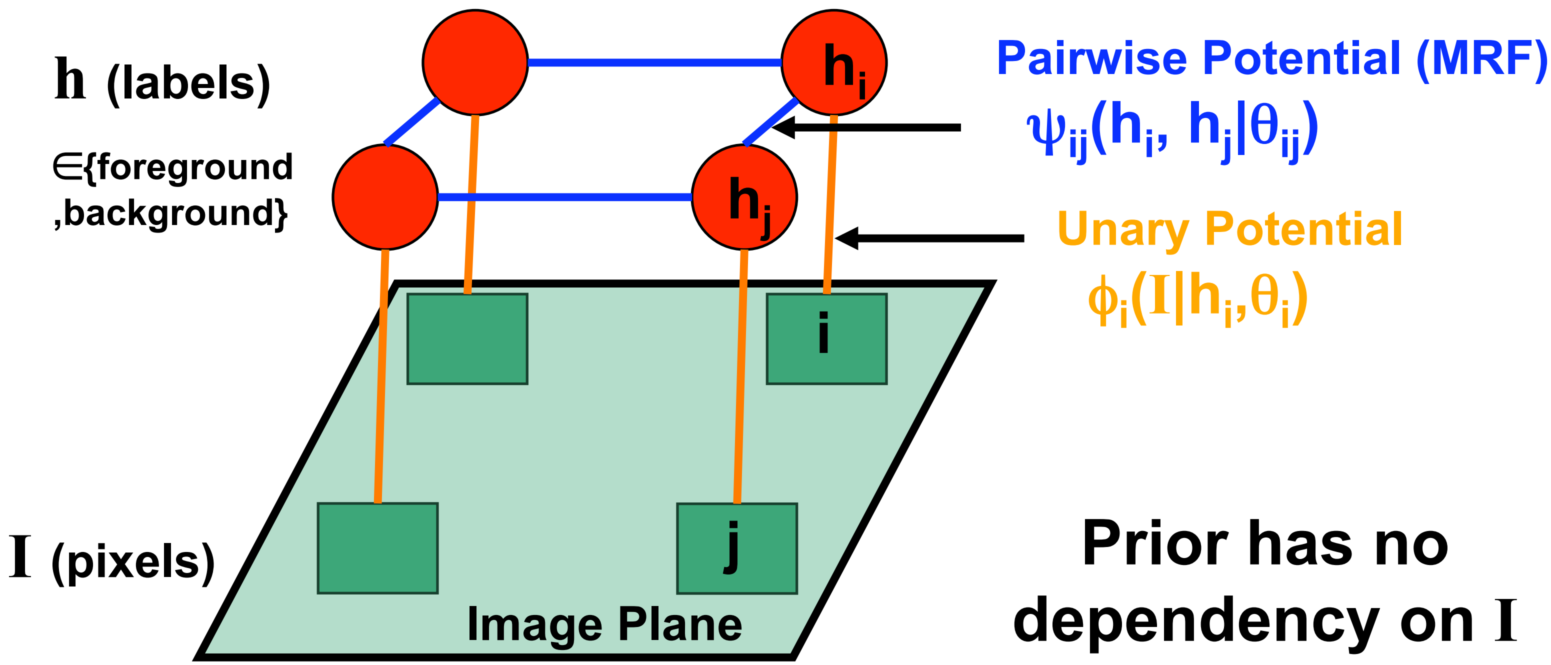
1. **Generative approach models joint**
   → **Markov random field (MRF)**

2. **Discriminative approach models posterior directly**
   → **Conditional random field (CRF)**

Generative Markov Random Field

$$p(h, I \mid \theta) = \boxed{p(I \mid h, \theta)}\,\boxed{p(h \mid \theta)}$$

$$= \frac{1}{Z(\theta)}\left[\underbrace{\prod_i \phi_i(I \mid h_i, \theta_i)}_{\text{Likelihood}}\underbrace{\prod_{ij} \psi_{ij}(h_i, h_j \mid \theta_{ij})}_{\text{MRF Prior}}\right]$$



Prior has no dependency on I

Conditional Random Field

Lafferty, McCallum and Pereira 2001

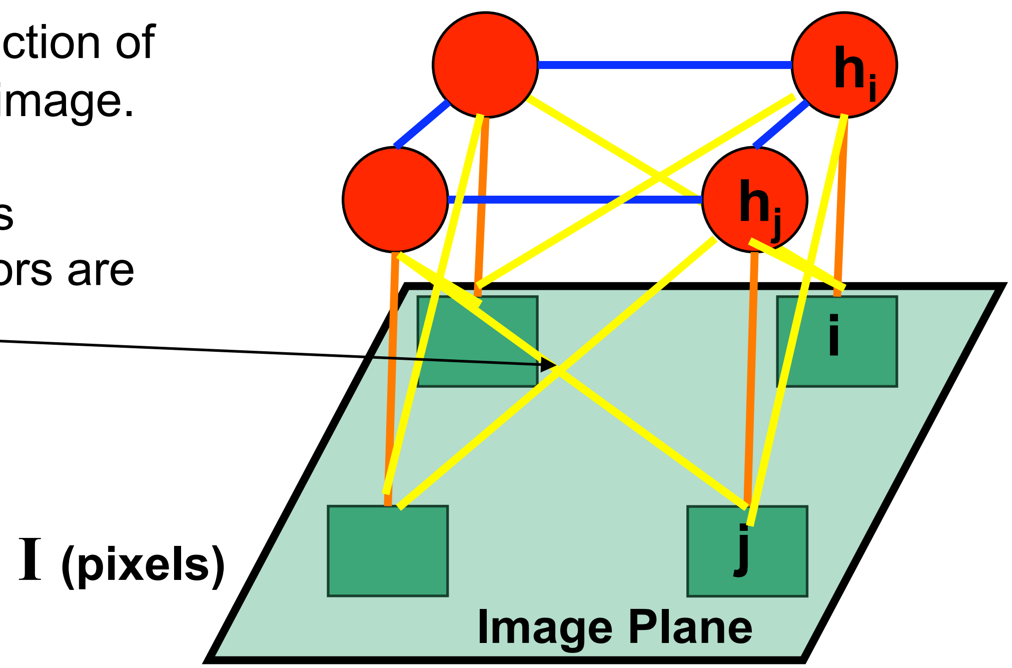
Discriminative approach

$$p(h \mid I, \theta) = \frac{1}{Z(I,\theta)} \left[\prod_i \underbrace{\phi_i(h_i, I \mid \theta_i)}_{\textbf{Unary}} \prod_{ij} \underbrace{\psi_{ij}(h_i, h_j, I \mid \theta_{ij})}_{\textbf{Pairwise}} \right]$$

• Dependency on I allows introduction of pairwise terms that make use of image.

• For example, neighboring labels should be similar only if pixel colors are similar → Contrast term

e.g Kumar and Hebert 2003



OBJCUT

Kumar, Torr & Zisserman 2005

$$p(h \mid \Omega, I, \theta) \propto \left[\prod_i \overbrace{\underbrace{\phi_i^1(I \mid h_i, \theta_i)}_{\text{Color Likelihood}} \cdot \underbrace{\phi_i^2(h_i \mid \Omega)}_{\text{Distance from } \Omega}}^{\text{Unary}} \prod_{ij} \overbrace{\underbrace{\psi_{ij}^1(h_i, h_j \mid \theta_{ij})}_{\text{Label smoothness}} \cdot \underbrace{\psi_{ij}^2(I \mid h_i, h_j, \theta_{ij})}_{\text{Contrast}}}^{\text{Pairwise}} \right]$$

- Ω is a shape prior on the labels from a Layered Pictorial Structure (LPS) model

- Segmentation by:

  - Match LPS model to image (get number of samples, each with a different pose

  - Marginalize over the samples using a single graph cut [Boykov & Jolly, 2001]

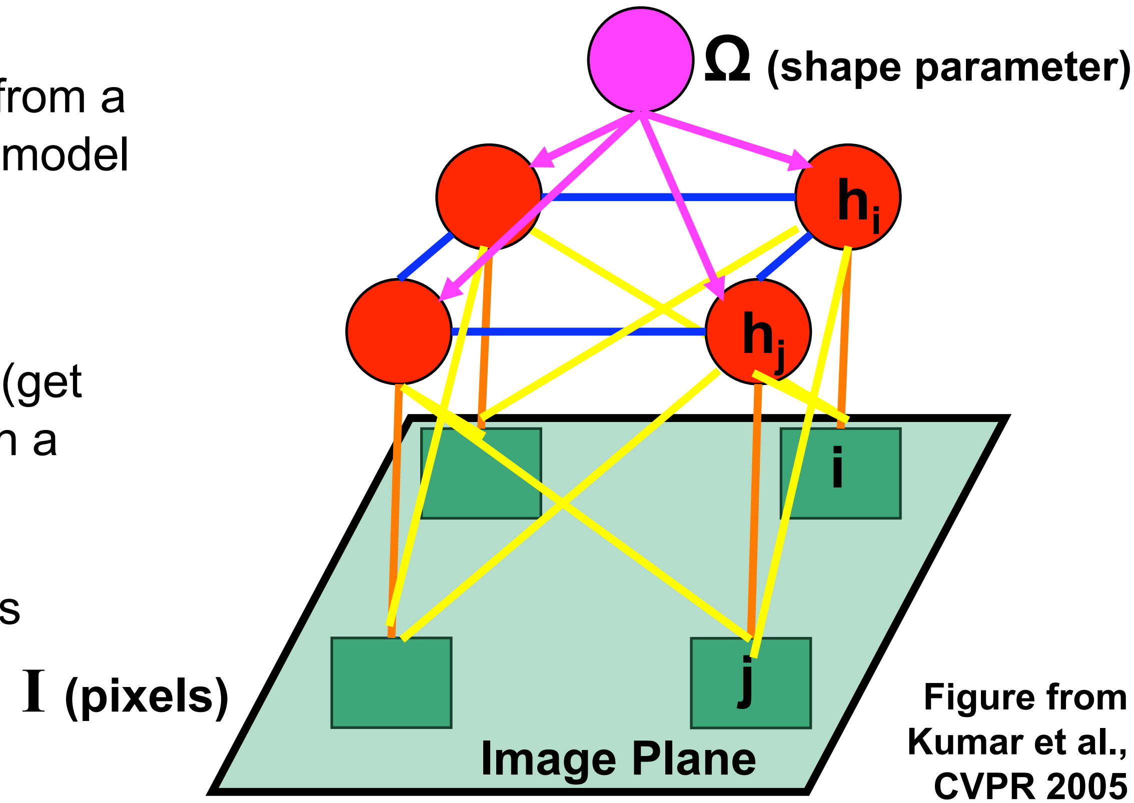
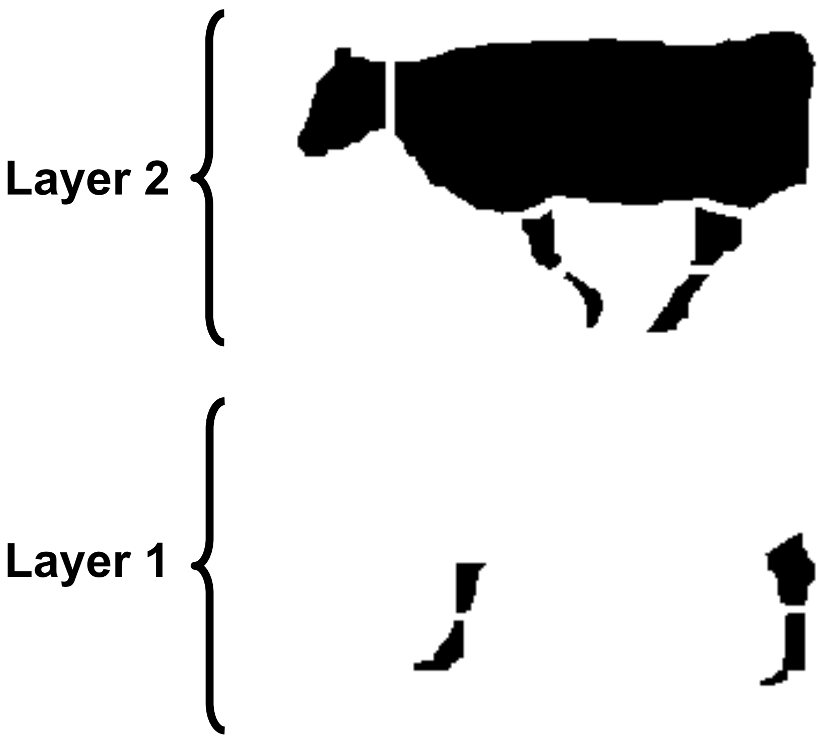
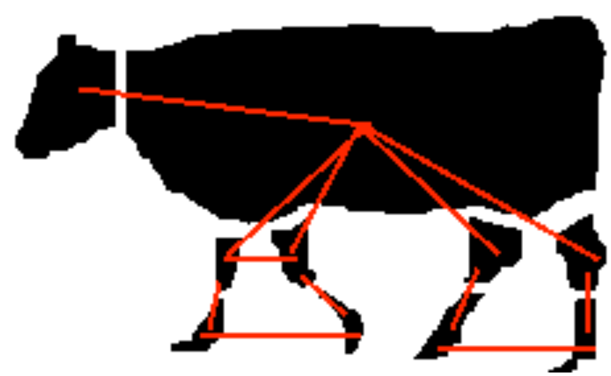


Figure from Kumar et al., CVPR 2005

OBJCUT:
Shape prior - **Ω** - Layered Pictorial Structures (LPS)

- Generative model
- Composition of parts + spatial layout





Spatial Layout
(Pairwise Configuration)

Parts in Layer 2 can occlude parts in Layer 1

Kumar, et al. 2004, 2005

OBJCUT: Results

Using LPS Model for Cow

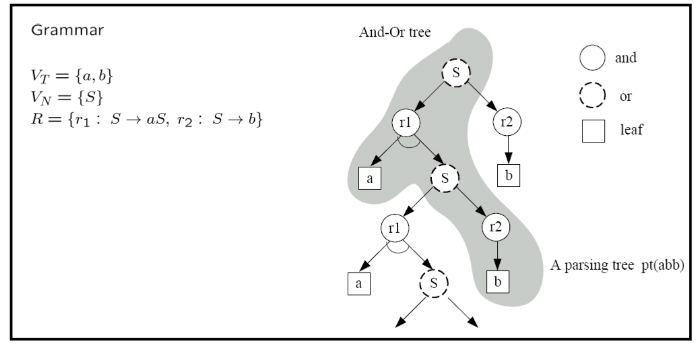
In the absence of a clear boundary between object and background

Image	**Segmentation**
	
	

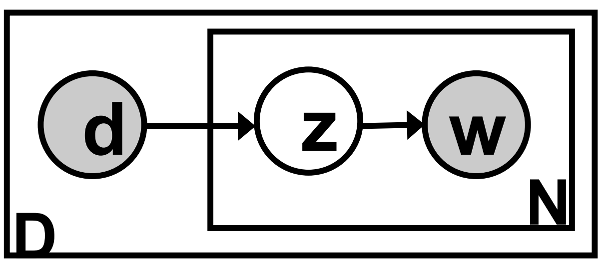
Generative models

Two big families:

- Grammar based models



- Topic models



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.

Grammars

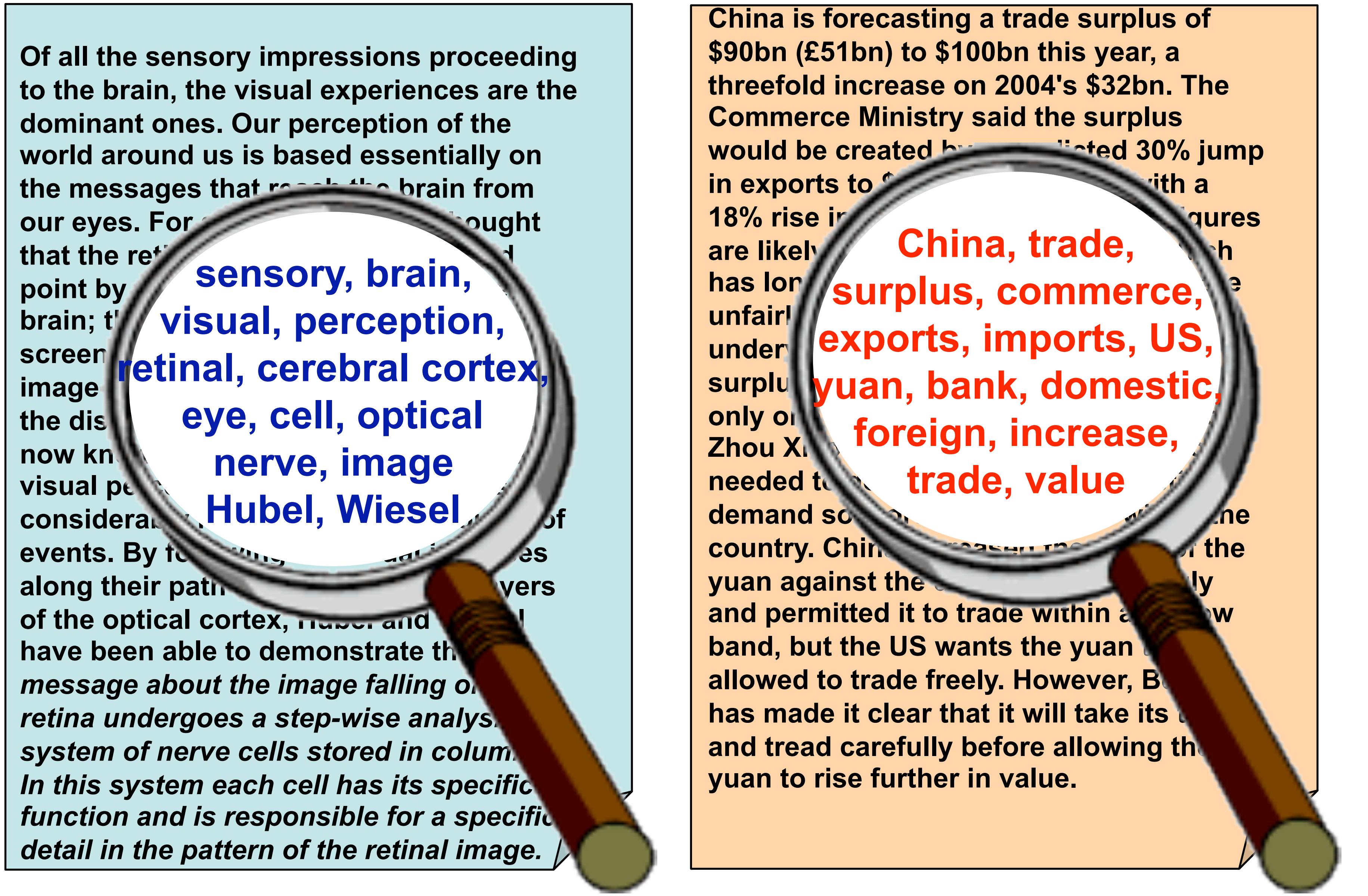
"A common framework for visual knowledge representation and object categorization. Gram-mars, studied mostly in language, are known for their expressive power in generating a very large set of configurations or instances, i.e. their language, by composing a relatively much smaller set of words, i.e. shared and reusable elements, using production rules."

A Stochastic Grammar of Images
Song-Chun Zhu and David Mumford

Object → **Bag of ‘words’**

Slide credit: Fei fei

Analogy to documents



Slide credit: Fei fei

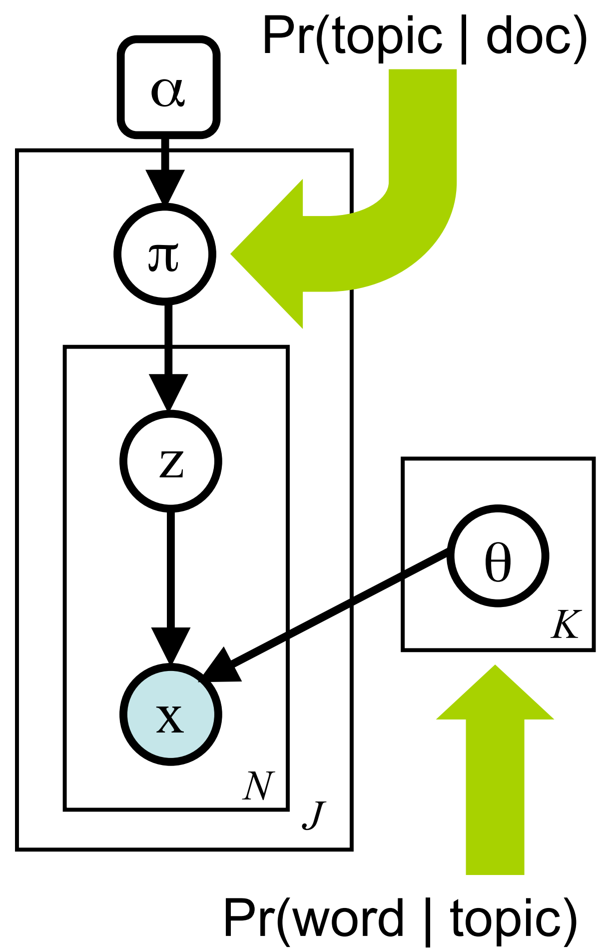
Related works

- Early “bag of words” models: mostly texture recognition
  – Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003;
- Hierarchical Bayesian models for documents (pLSA, LDA, etc.)
  – Hoffman 1999; Blei, Ng & Jordan, 2004; Teh, Jordan, Beal & Blei, 2004
- Object categorization
  – Csurka, Bray, Dance & Fan, 2004; Sivic, Russell, Efros, Freeman & Zisserman, 2005; Sudderth, Torralba, Freeman & Willsky, 2005;
- Natural scene categorization
  – Vogel & Schiele, 2004; Fei-Fei & Perona, 2005; Bosch, Zisserman & Munoz, 2006

Slide credit: Fei fei

Hierarchical Topic Models

- Topic models typically use a *“bag of words”* approx.:
  - Learning topics allows transfer of information within a corpus of related documents
  - Mixing proportions capture the distinctive features of particular documents



Latent Dirichlet Allocation (LDA)

Blei, Ng, & Jordan, JMLR 2003

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**

2 generative models

1. Naïve Bayes classifier
   – Csurka Bray, Dance & Fan, 2004

2. Hierarchical Bayesian text models (pLSA and LDA)
   – Background: Hoffman 2001, Blei, Ng & Jordan, 2004
   – Object categorization: Sivic et al. 2005, Sudderth et al. 2005
   – Natural scene categorization: Fei-Fei et al. 2005

Slide credit: Fei fei

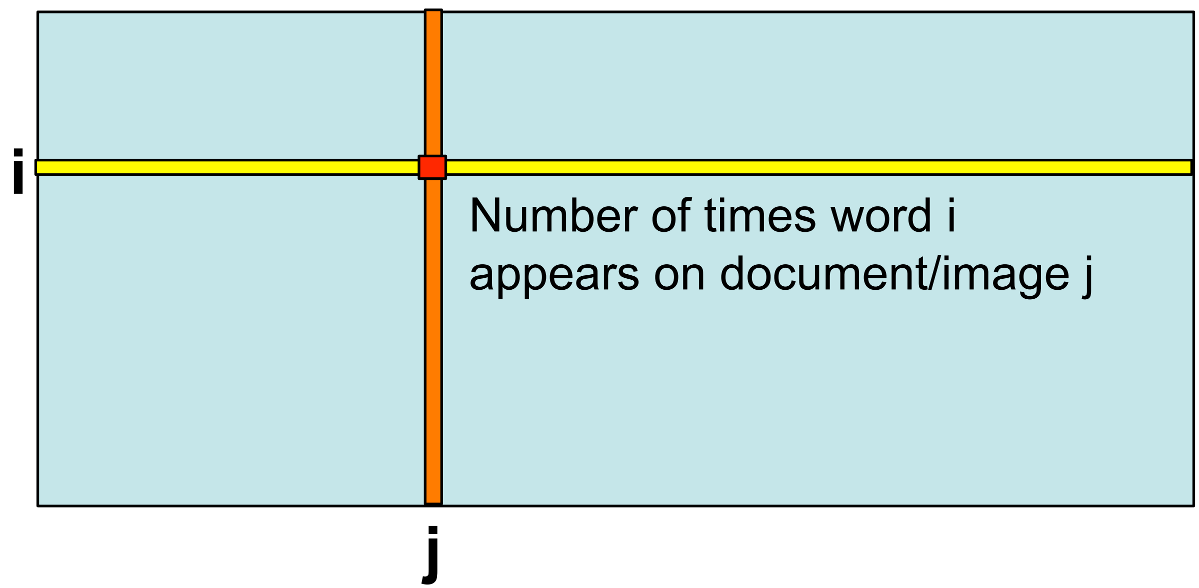
First, some notations

- w_n: each patch in an image
  - $w_n = [0,0,\ldots 1,\ldots,0,0]^T$
- **w**: a collection of all N patches in an image
  - $\mathbf{w} = [w_1, w_2, \ldots, w_N]$
- d_j: the j^{th} image in an image collection
- c: category of the image
- z: theme or topic of the patch

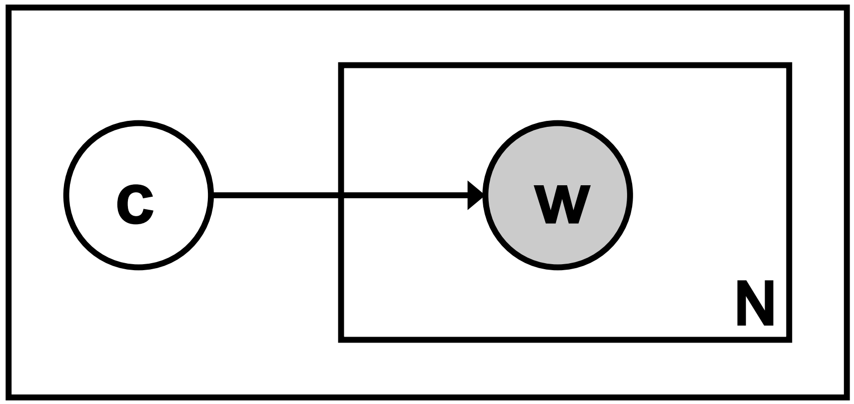
Slide credit: Fei fei

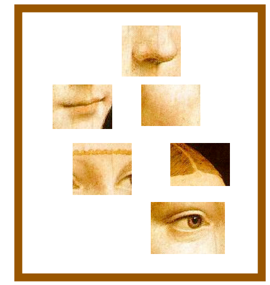
Documents collection

Co-ocurrence table:



Case #1: the Naïve Bayes model





$$c^* = \arg\max_c \; p(c \mid w) \propto p(c)\,p(w \mid c) = p(c)\prod_{n=1}^{N} p(w_n \mid c)$$

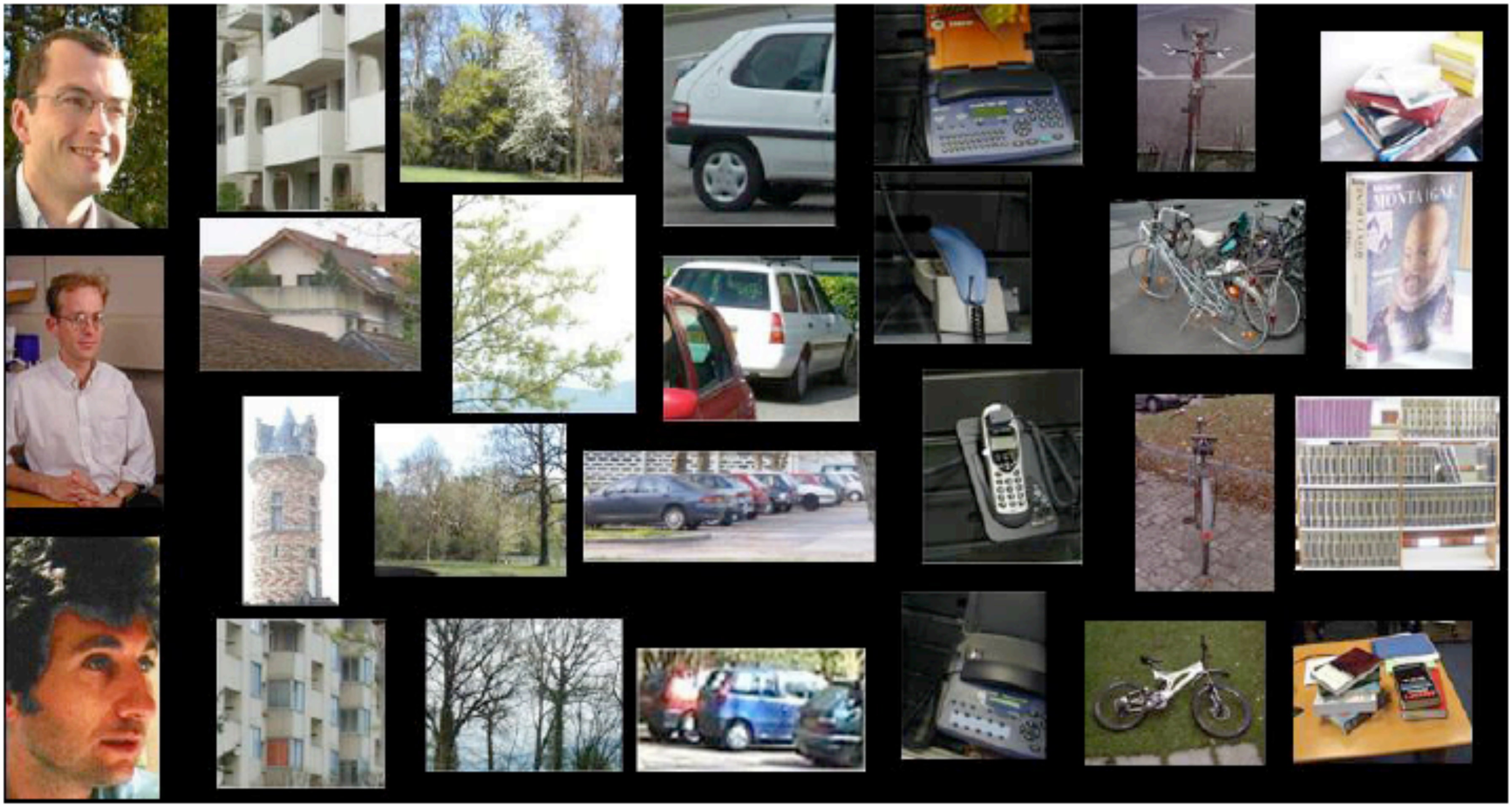
Object class decision

Prior prob. of the object classes

Image likelihood given the class

Csurka et al. 2004

Our in-house database contains 1776 images in seven classes[1]: faces, buildings, trees, cars, phones, bikes and books. Fig. 2 shows some examples from this dataset.



Csurka et al. 2004

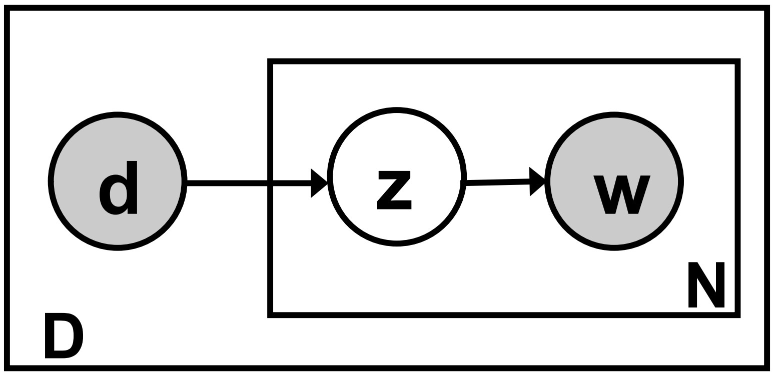
Table 1. Confusion matrix and the mean rank for the best vocabulary (*k=1000)*.

True classes →	*faces*	*buildings*	*trees*	*cars*	*phones*	*bikes*	*books*
faces	**76**	4	2	3	4	4	13
buildings	2	**44**	5	0	5	1	3
trees	3	2	**80**	0	0	5	0
cars	4	1	0	**75**	3	1	4
phones	9	15	1	16	**70**	14	11
bikes	2	15	12	0	8	**73**	0
books	4	19	0	6	7	2	**69**
Mean ranks	1.49	1.88	1.33	1.33	1.63	1.57	1.57

Csurka et al. 2004

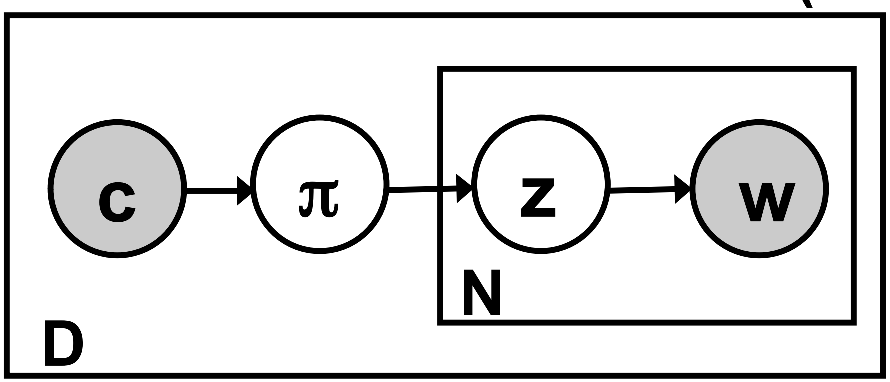
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 2001

Latent Dirichlet Allocation (LDA)

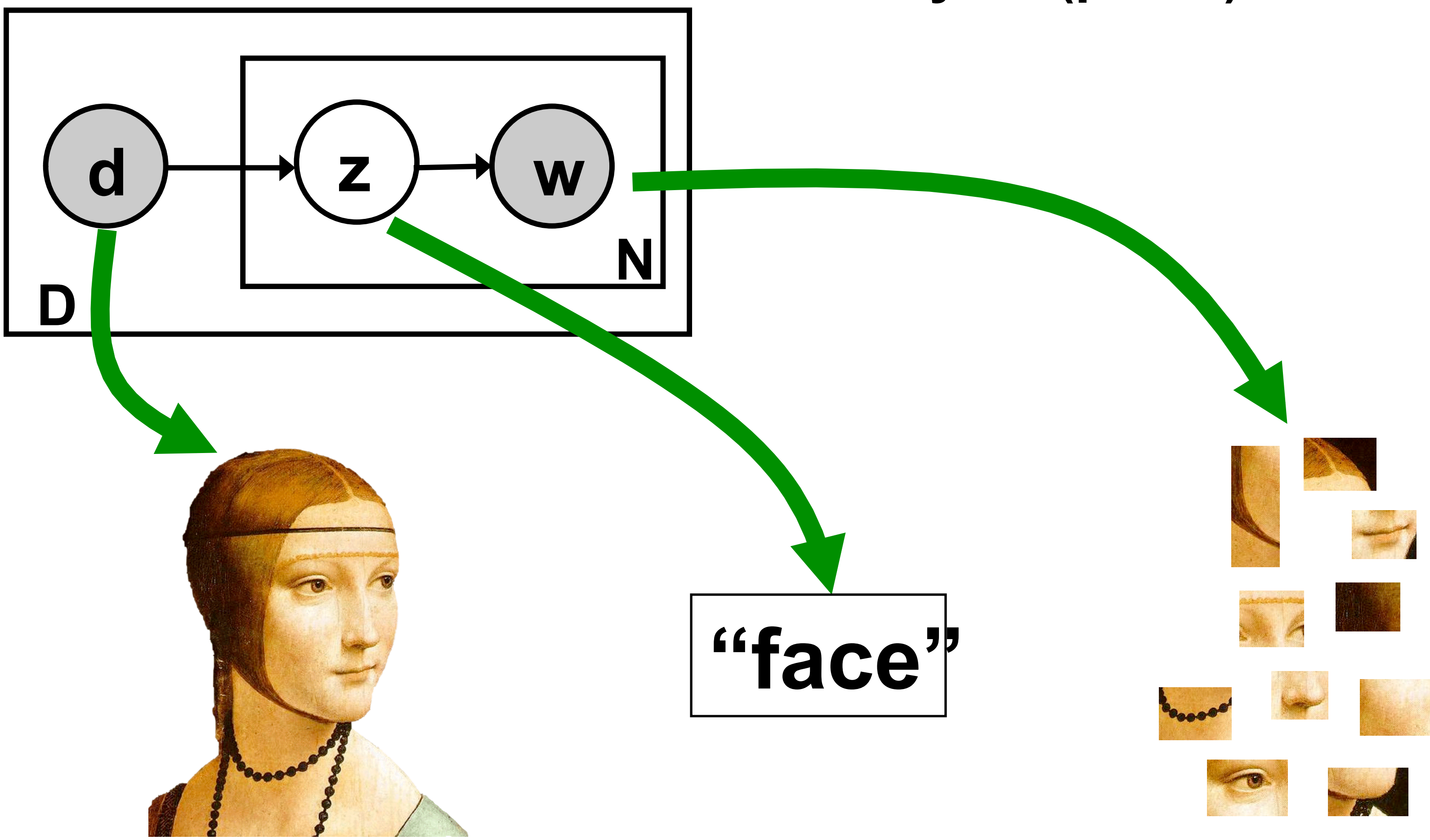


Blei et al., 2001

Slide credit: Fei fei

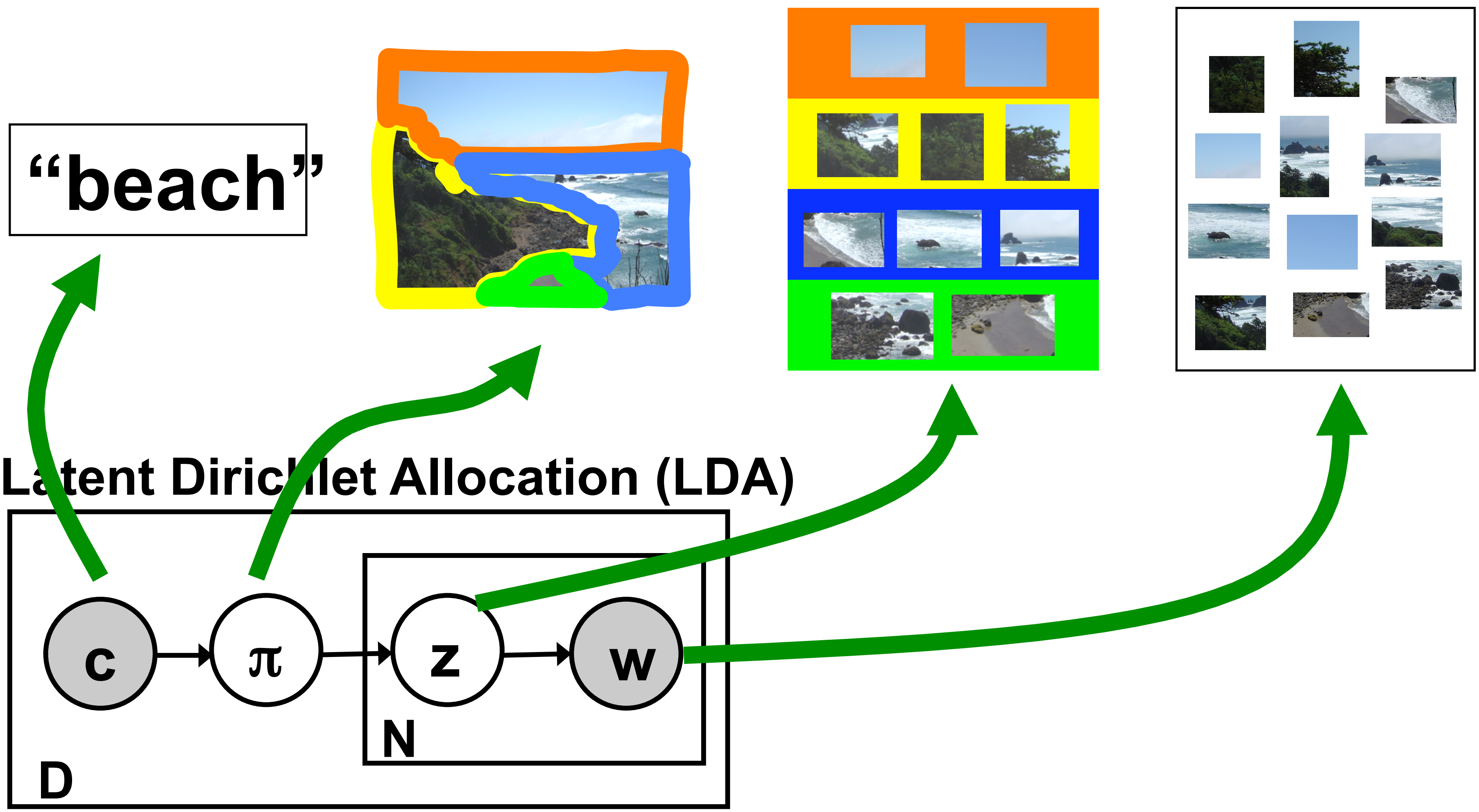
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



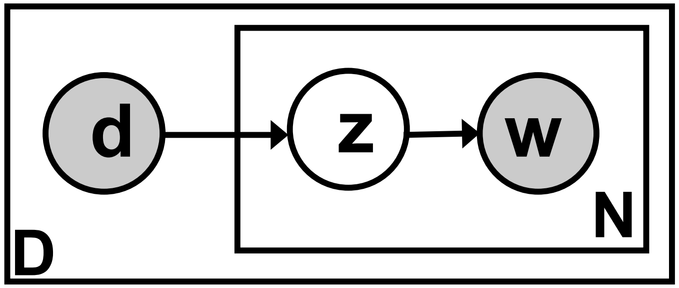
Sivic et al. ICCV 2005

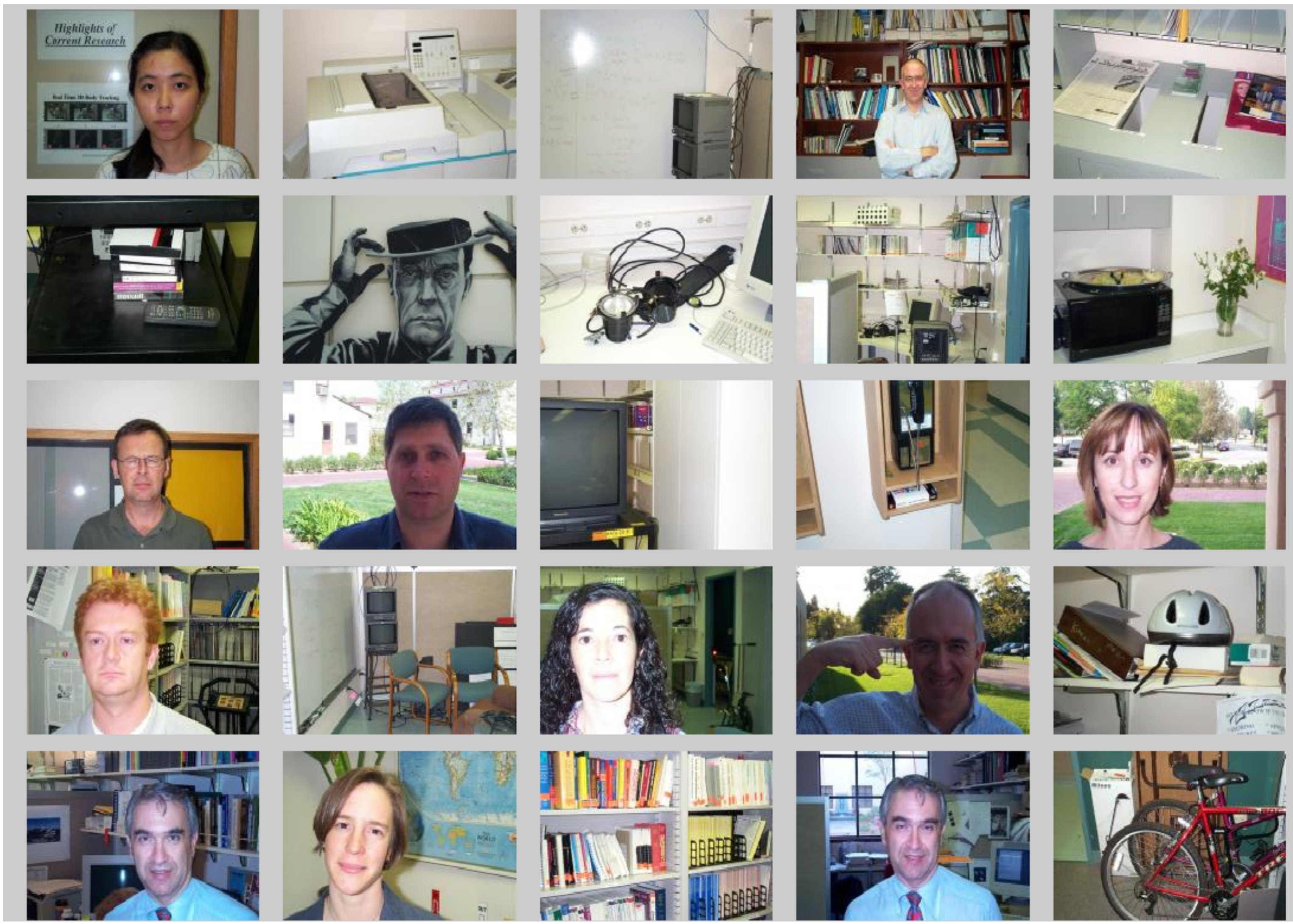
Case #2: Hierarchical Bayesian text models

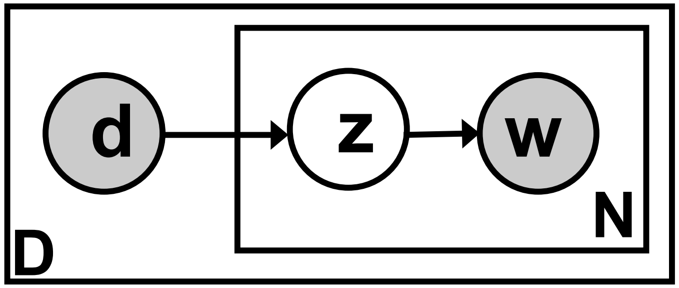


Fei-Fei et al. ICCV 2005

Case #2: the pLSA model

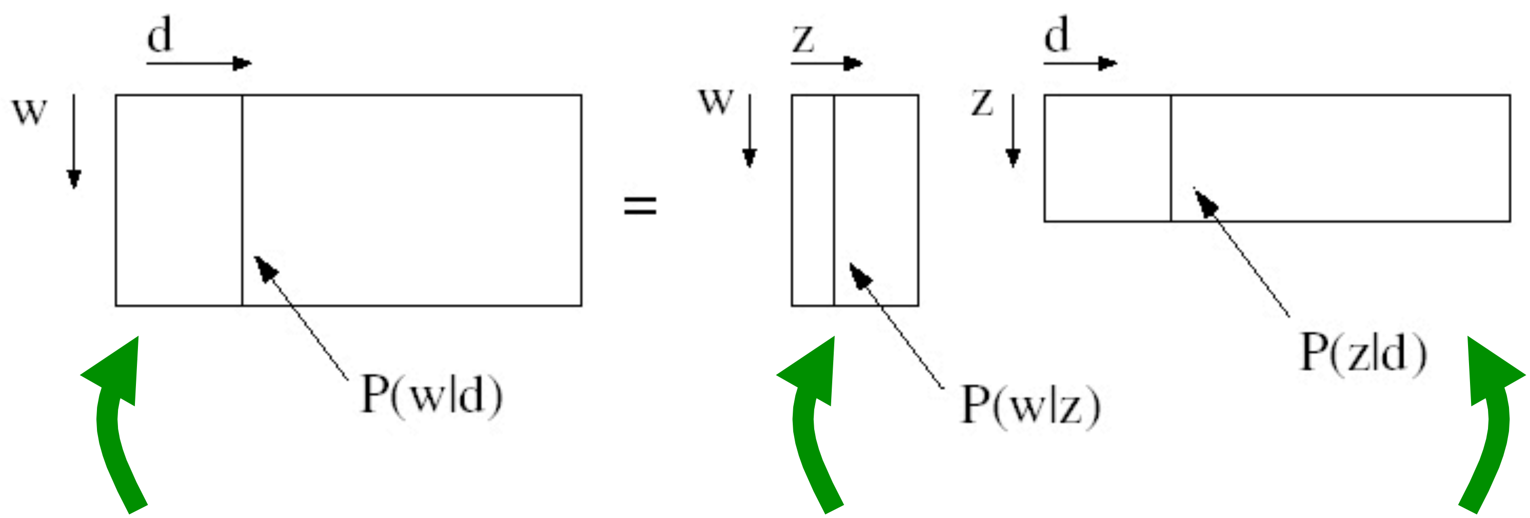






Case #2: the pLSA model

$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k) p(z_k \mid d_j)$$



Observed codeword distributions

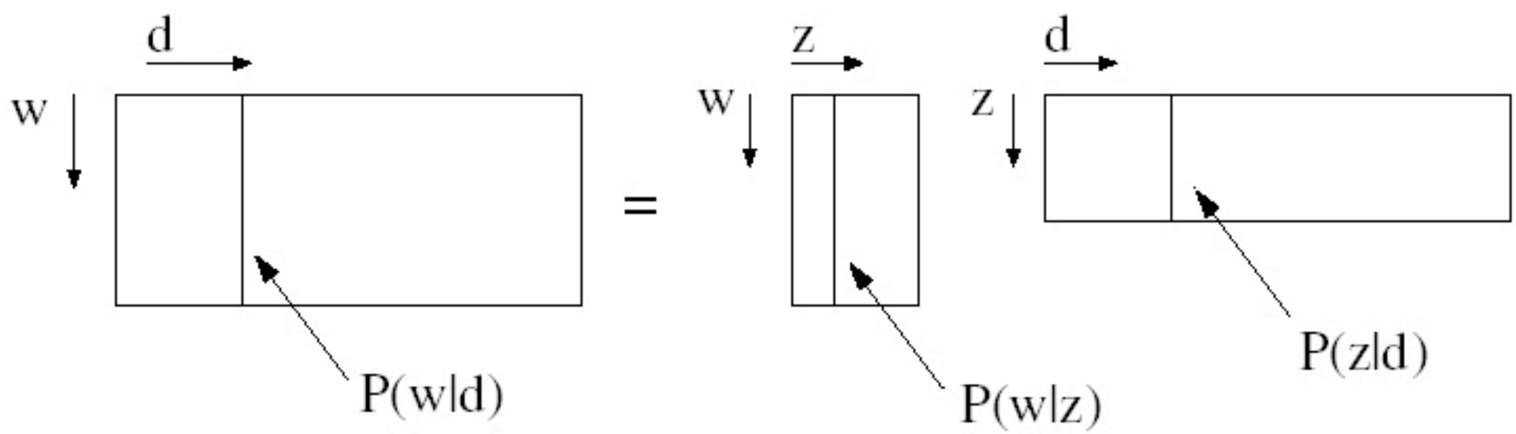
Codeword distributions per theme (topic)

Theme distributions per image

Slide credit: Josef Sivic

Case #2: Recognition using pLSA

$$z^* = \arg\max_z p(z \mid d)$$



Slide credit: Josef Sivic

Case #2: Learning the pLSA parameters

Observed counts of word *i* in document *j*

$$L = \prod_{i=1}^{M} \prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)}$$

$$\sum_{k=1}^{K} P(z_k|d_j) P(w_i|z_k)$$

Maximize likelihood of data using EM

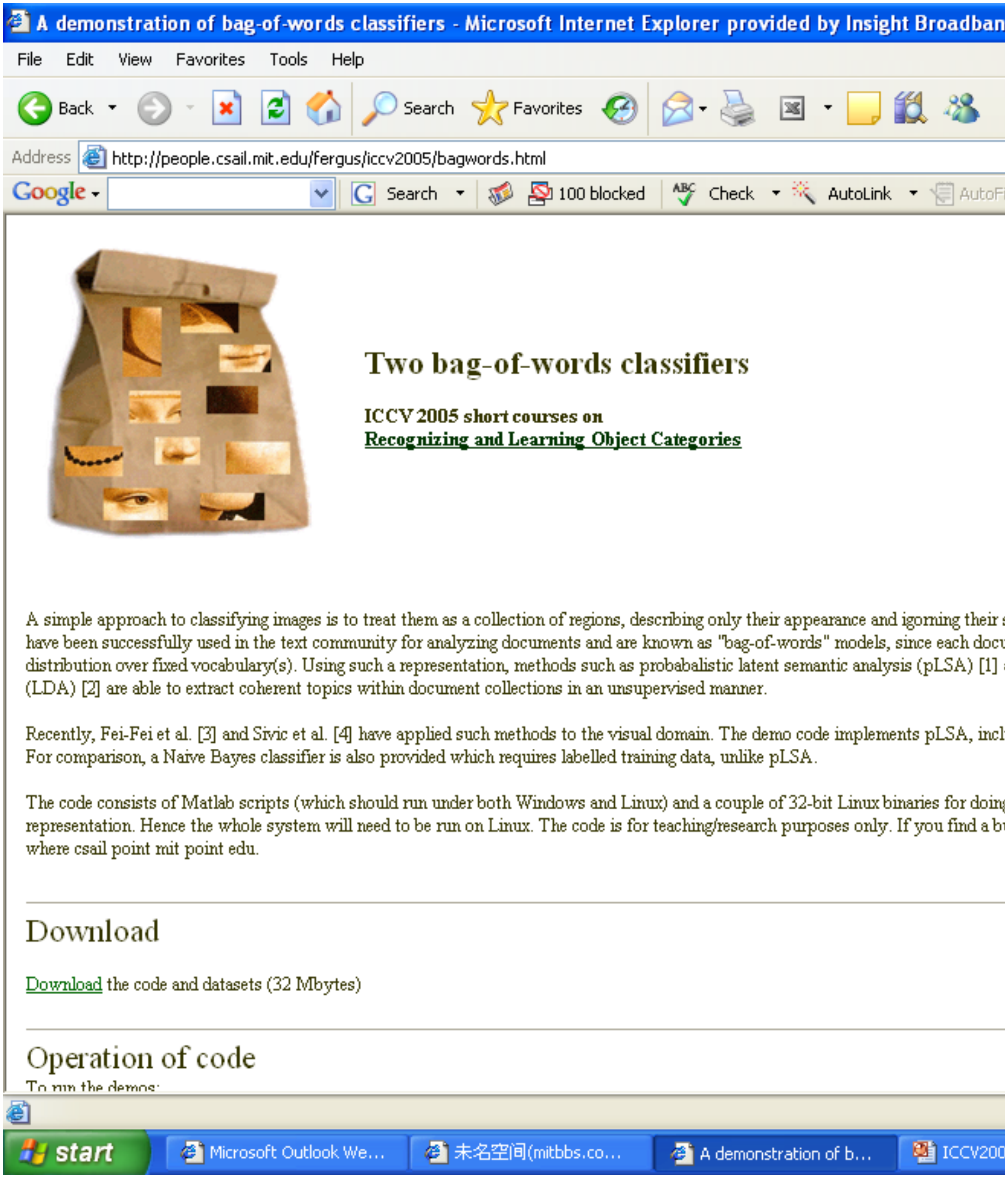
M … number of codewords

N … number of images

Slide credit: Josef Sivic

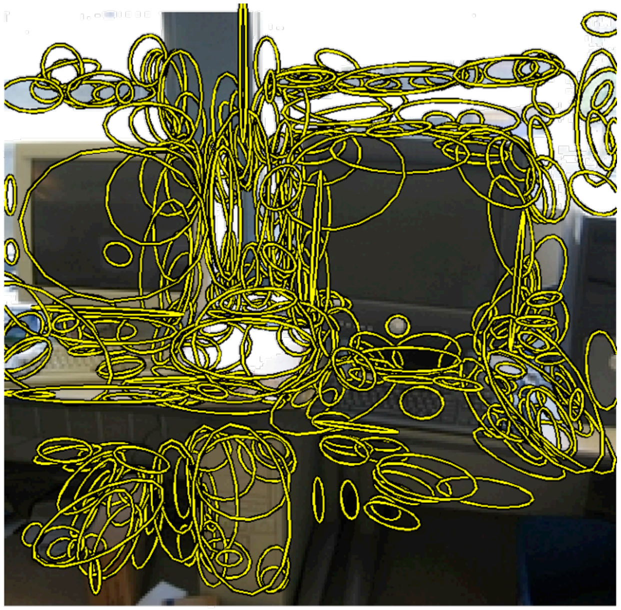
Demo

- Course website



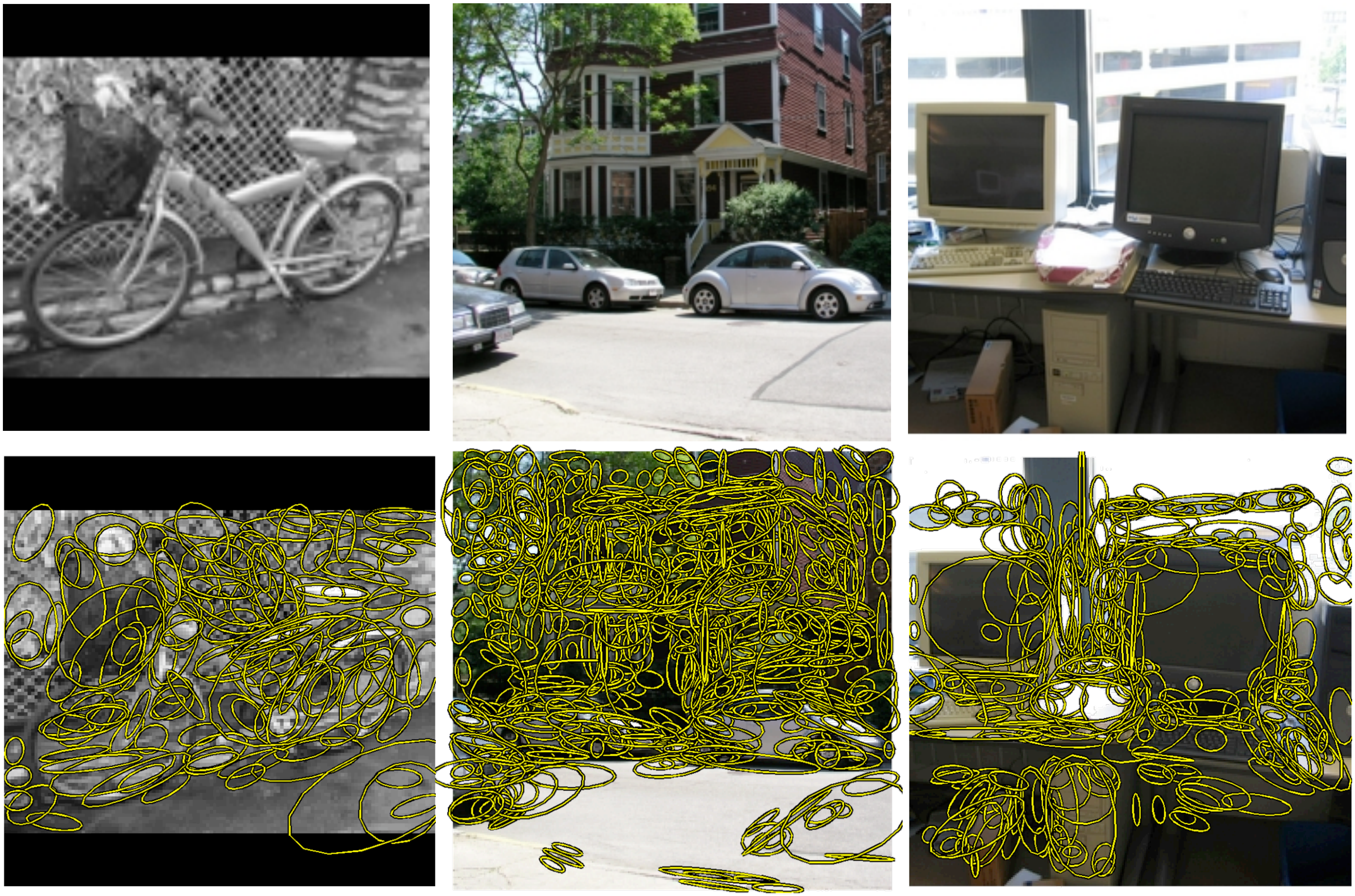
From Images to Features

- Pixels are very sensitive to changes in lighting & pose
- Instead represent image as *affine covariant regions:*
  – Harris affine invariant regions (corners & edges)
  – Maximally stable extremal regions (segmentation)



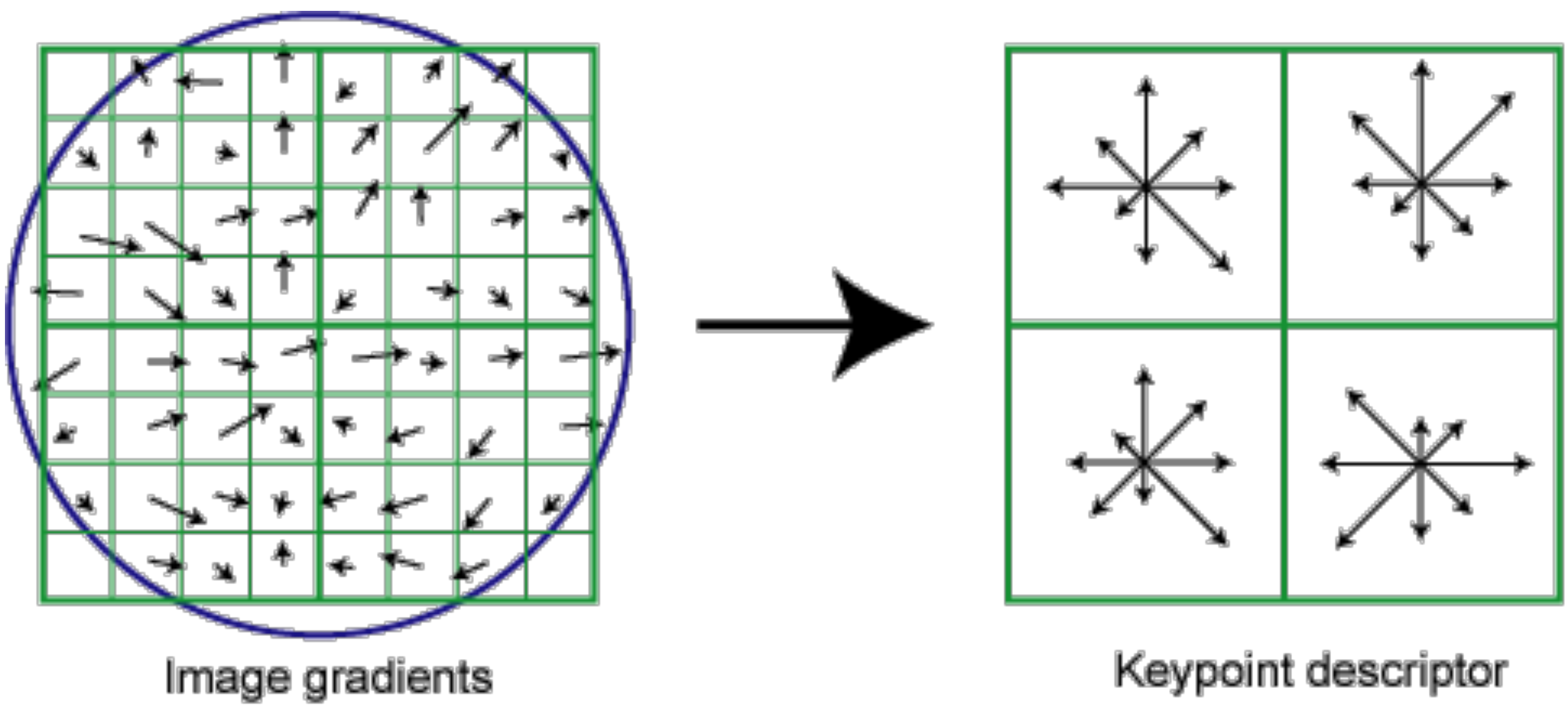
Software provided by
Oxford Visual Geometry Group

Sample Detected Features



Describing Feature Appearance

- **SIFT:** Scale Invariant Feature Transform
- Normalized histogram of orientation energy in each affinely adapted region (128-dim.)



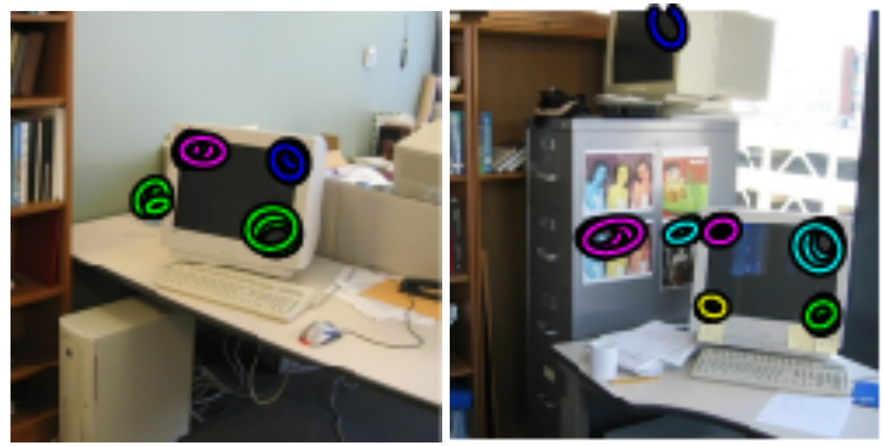
D. Lowe, IJCV 2004

A Discrete Feature Vocabulary

- Using all training images, build a dictionary via K-means clustering (~1000 words)
- Map each SIFT descriptor to nearest word

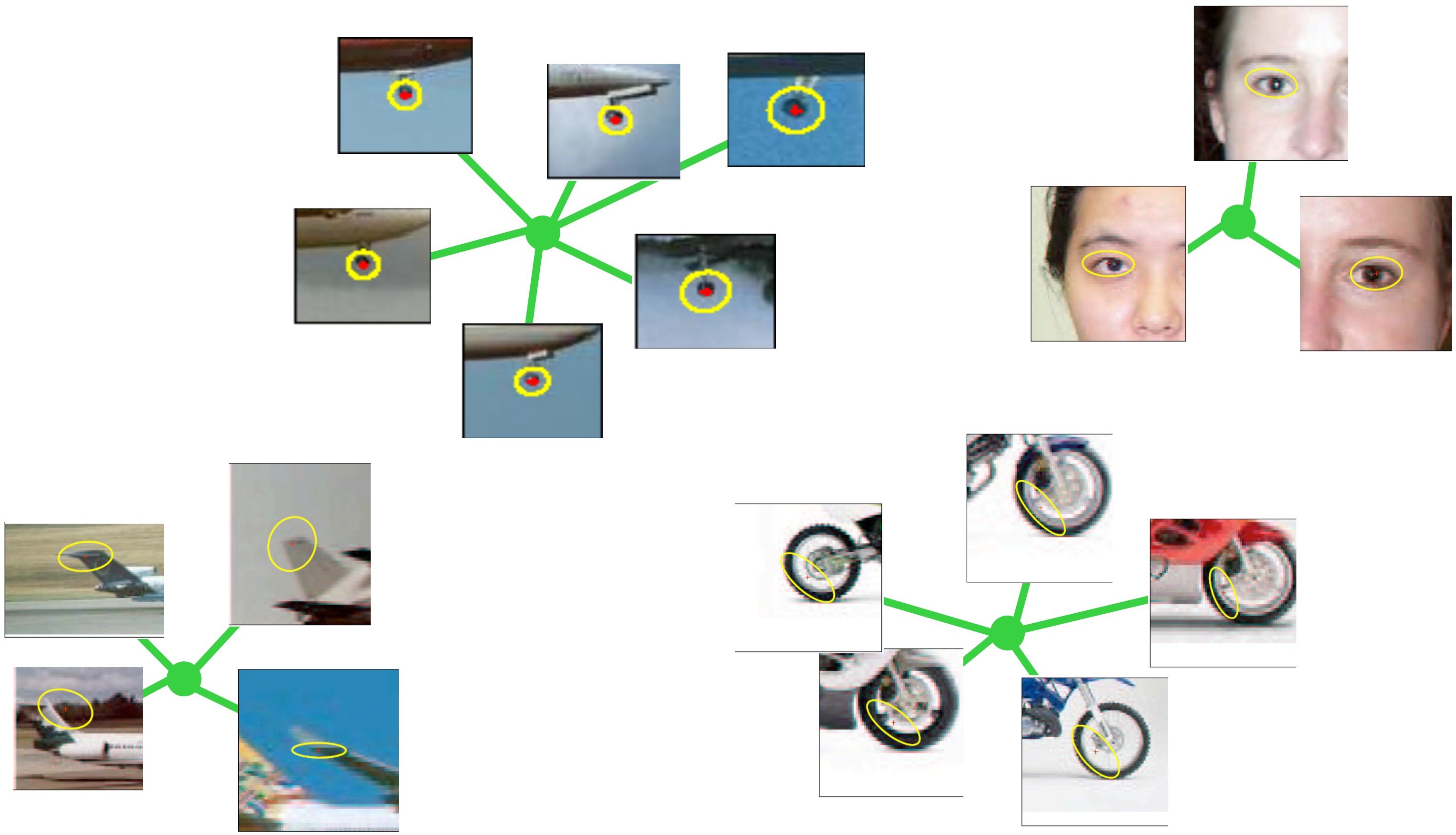
$w_{ji} \longrightarrow$ **appearance of feature i in image j**

$y_{ji} \longrightarrow$ **2D position of feature i in image j**



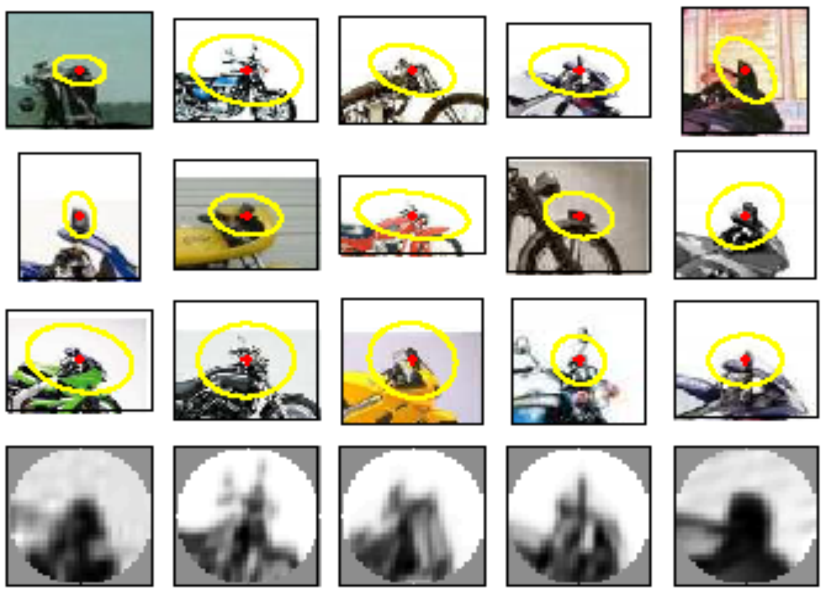
Form dictionary

Build visual vocabulary by k-means clustering SIFT descriptors (K~2,000)

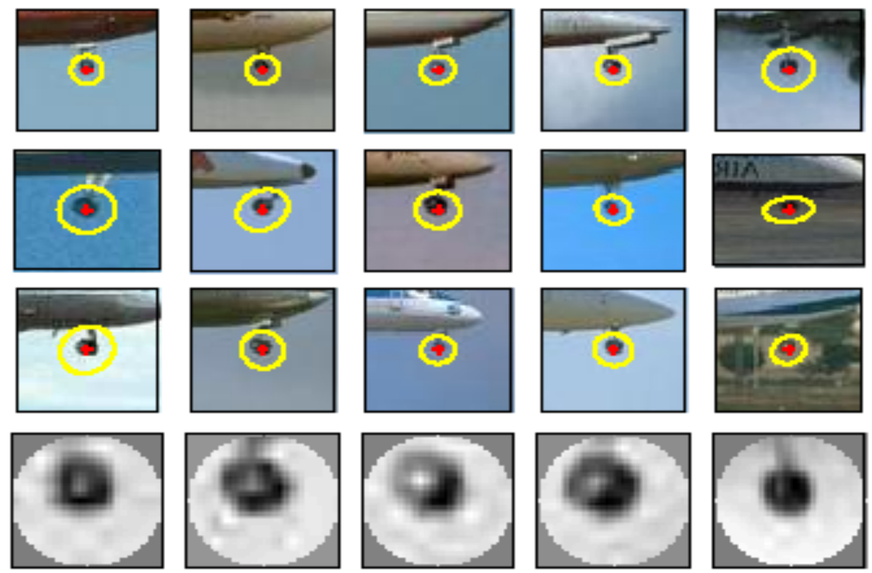


Slide credit: Bryan Russell & Josef Sivic

Example regions assigned to the same dictionary cluster



Cluster 1



Cluster 2

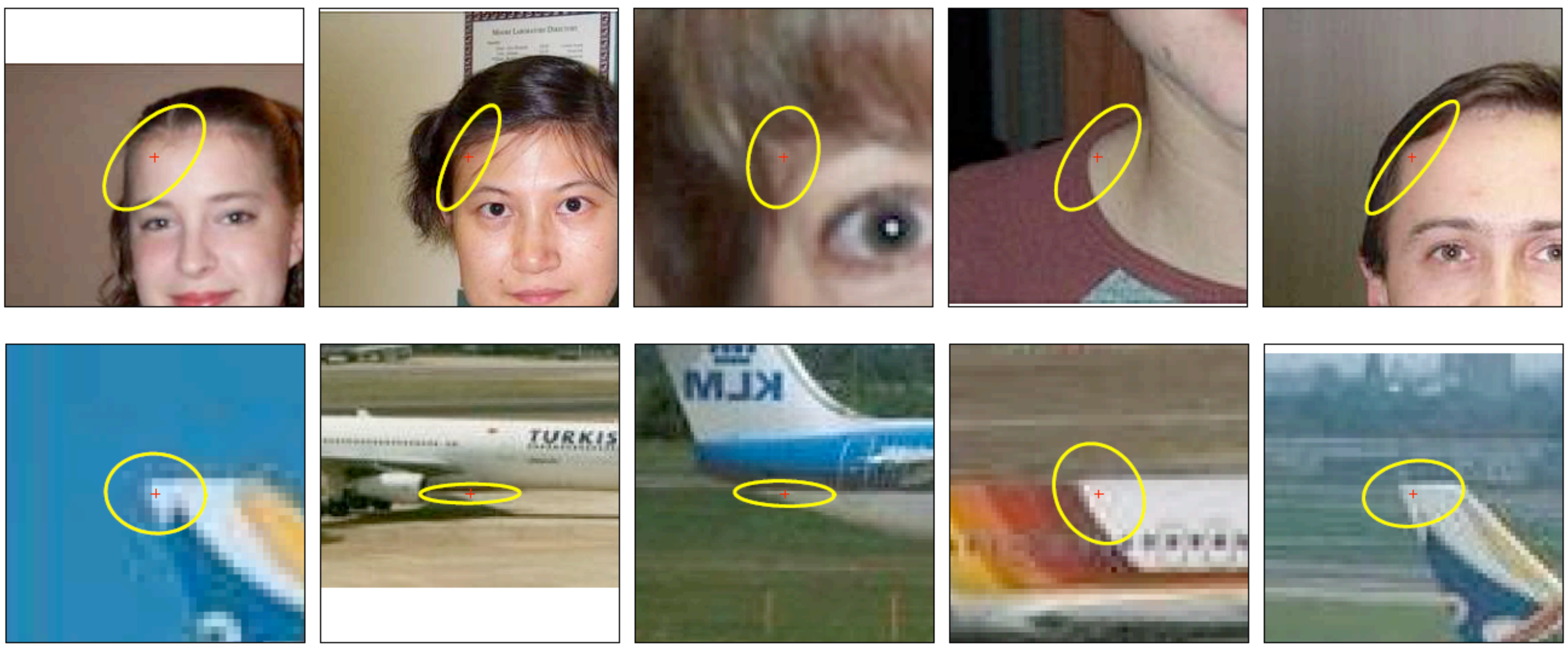
Slide credit: Bryan Russell & Josef Sivic

Polysemy

In English, “bank” refers to:

1. **a institution that handle money**
2. **the side of a river**

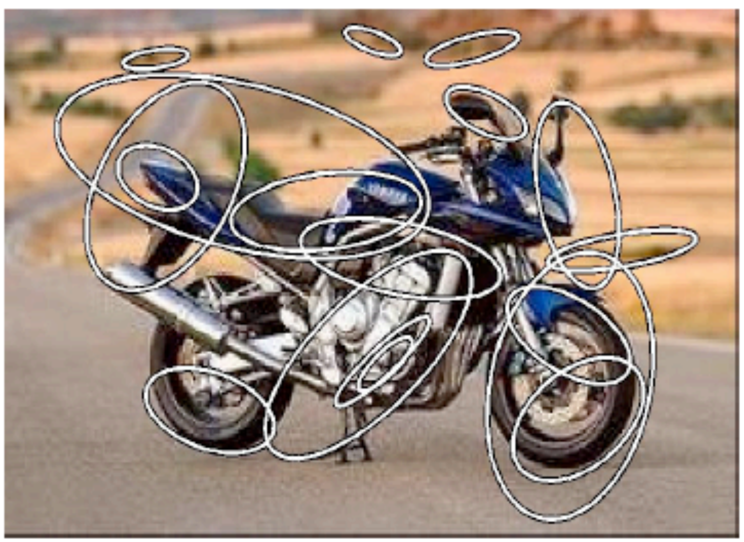
Regions that map to the same visual word:

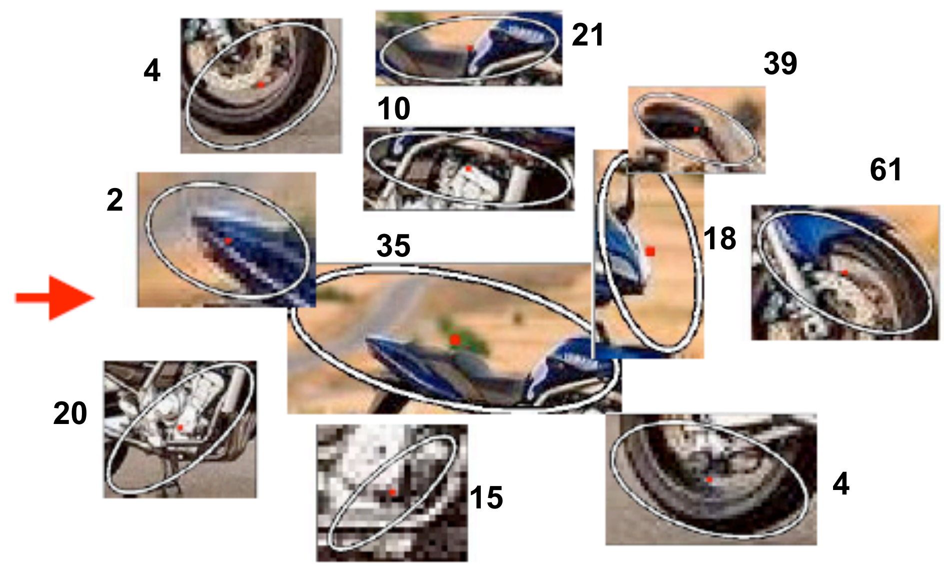


Slide credit: Bryan Russell & Josef Sivic

Representing an image with visual words

Sivic & Zisserman ’03





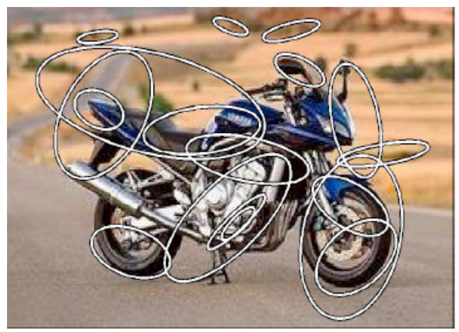
Interest regions

Visual words

Slide credit: Bryan Russell & Josef Sivic

System overview

Input image

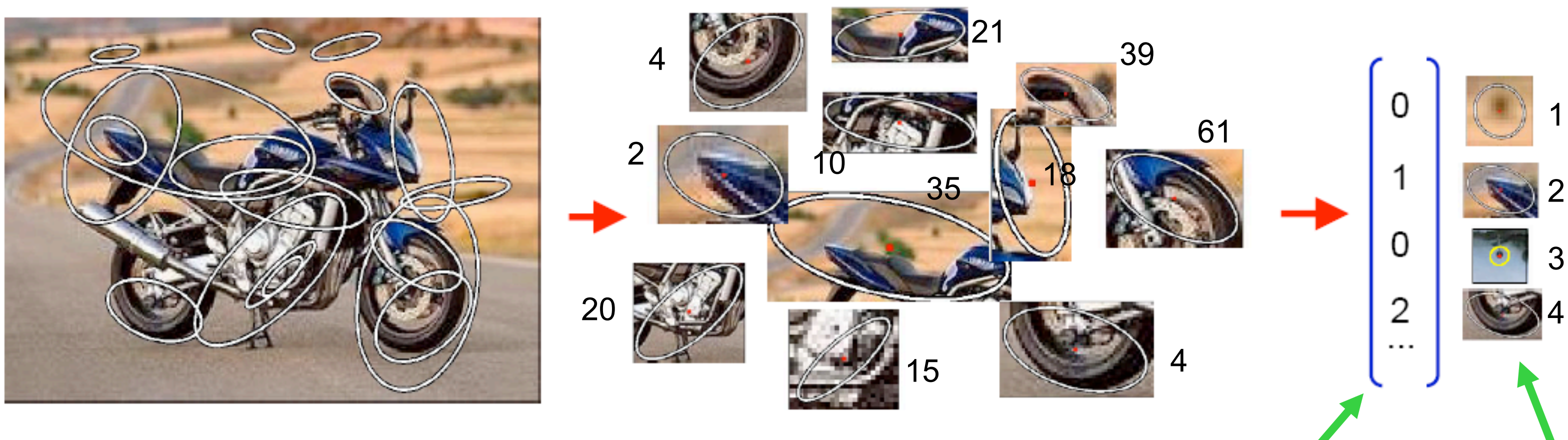


Compute visual words

Discover visual topics

Slide credit: Bryan Russell & Josef Sivic

Bag of words



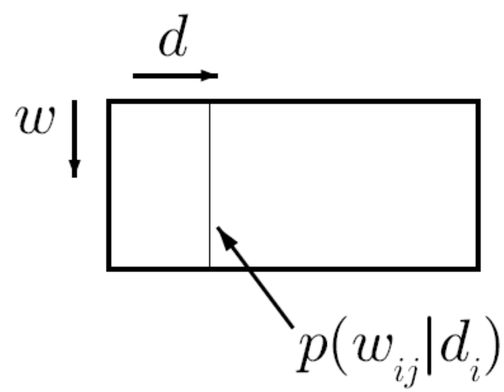
Interest regions

Visual words

Histogram

Dictionary

Stack visual word histograms
as columns in matrix



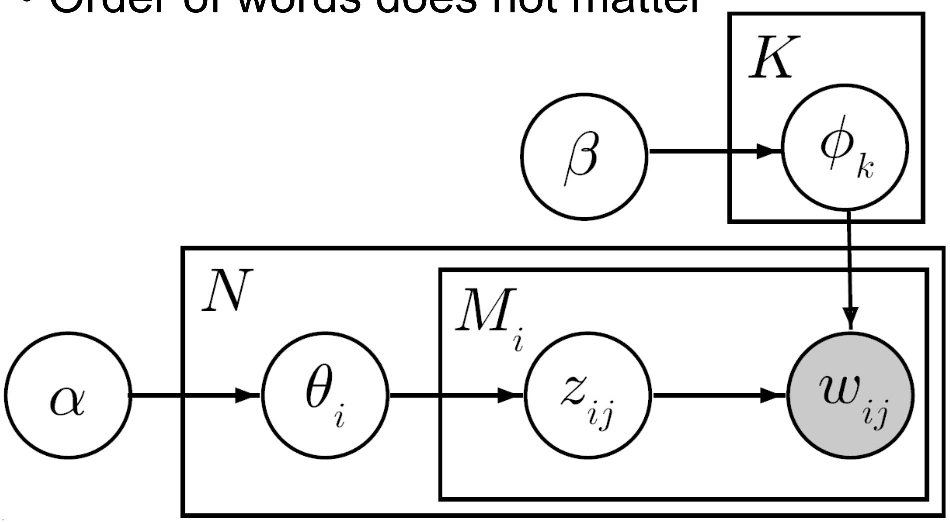
Throw away spatial information!

Slide credit: Bryan Russell & Josef Sivic

Latent Dirichlet Allocation (LDA)

Blei, et al. 2003

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



w_{ij} - words

z_{ij} - topic assignments

μ_i - topic mixing weights

Φ_k - word mixing weights

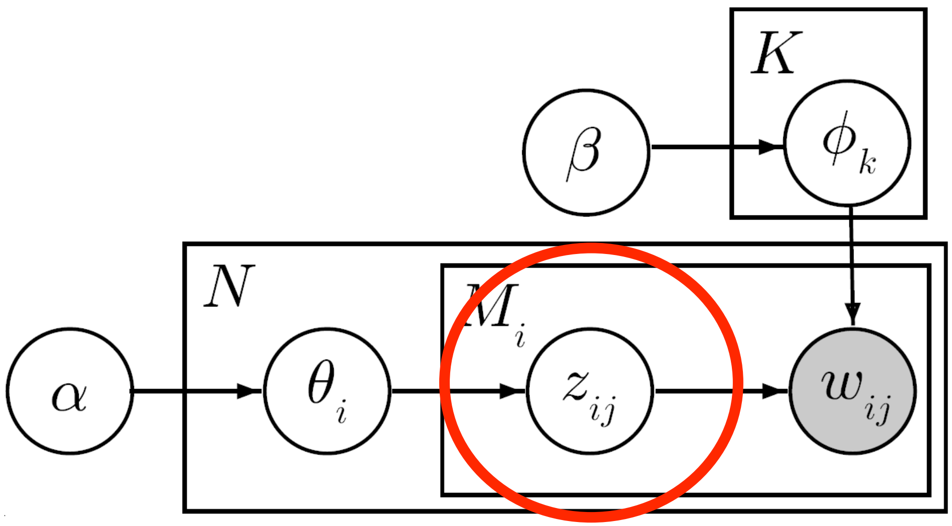
$$z_{ij}|\theta_i \sim \theta_i \qquad \theta_i|\alpha \sim Dirichlet(\alpha)$$

$$w_{ij}|z_{ij} = k, \phi \sim \phi_k \qquad \phi_k|\beta \sim Dirichlet(\beta)$$

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

Slide credit: Bryan Russell & Josef Sivic

Inference



w_{ij} - words

z_{ij} - topic assignments

μ_i - topic mixing weights

$Á_k$ - word mixing weights

Use Gibbs sampler to sample topic assignments

[Griffiths & Steyvers 2004]

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

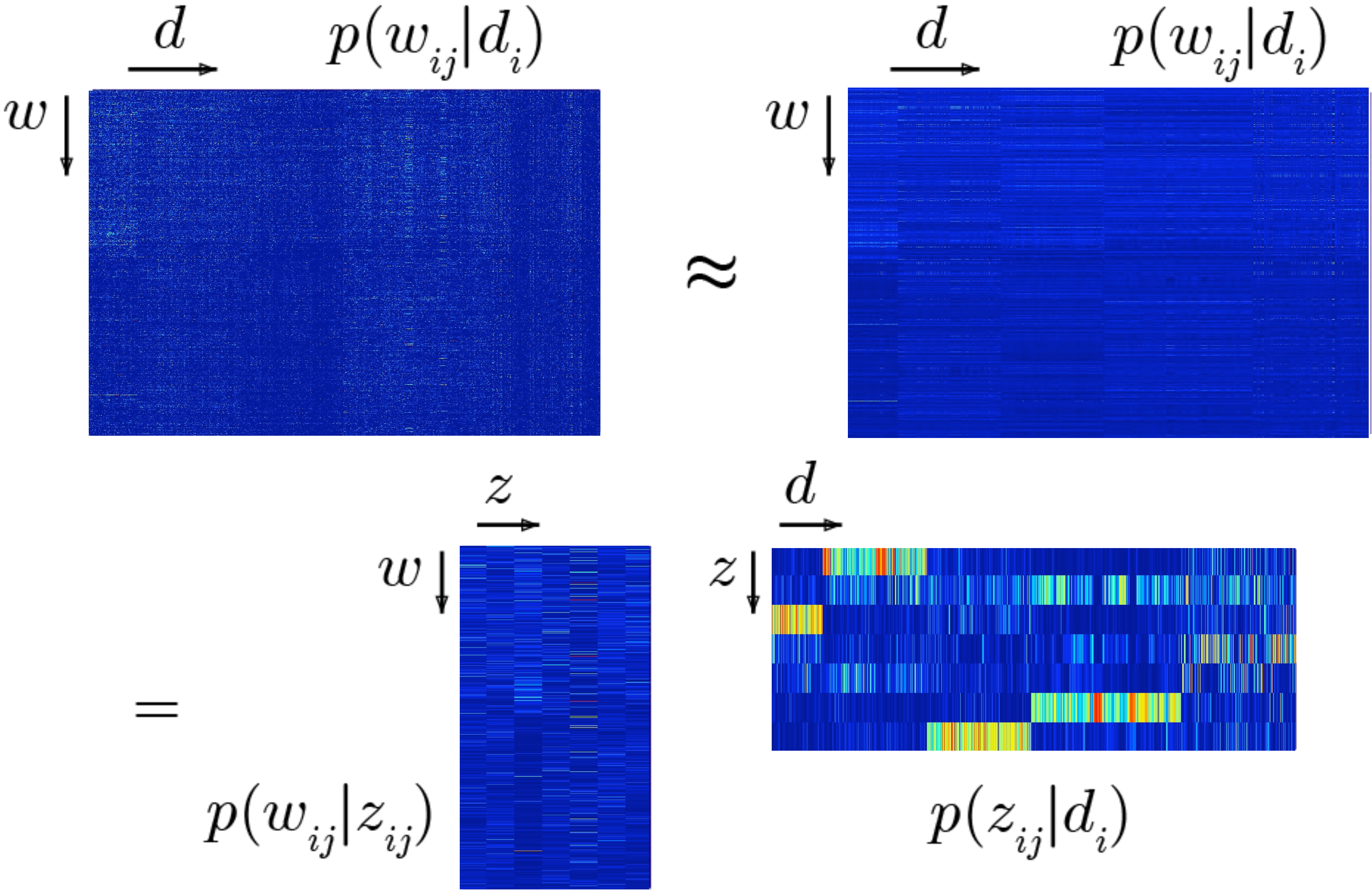
- Only need to maintain counts of topic assignments
- Sampler typically converges in less than 50 iterations
- Run time is less than an hour

Slide credit: Bryan Russell & Josef Sivic

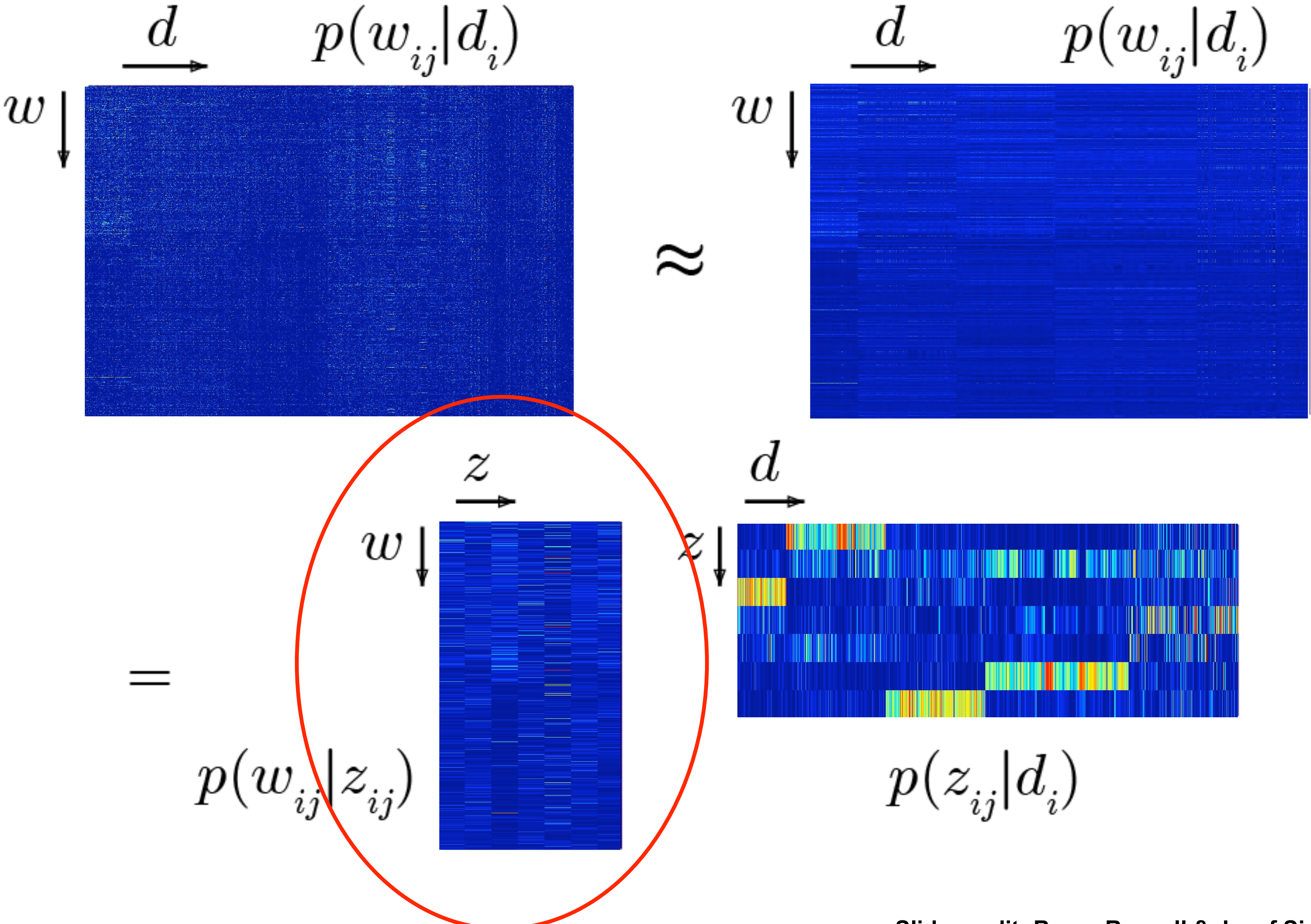
Apply to Caltech 4 + background images

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

Slide credit: Bryan Russell & Josef Sivic

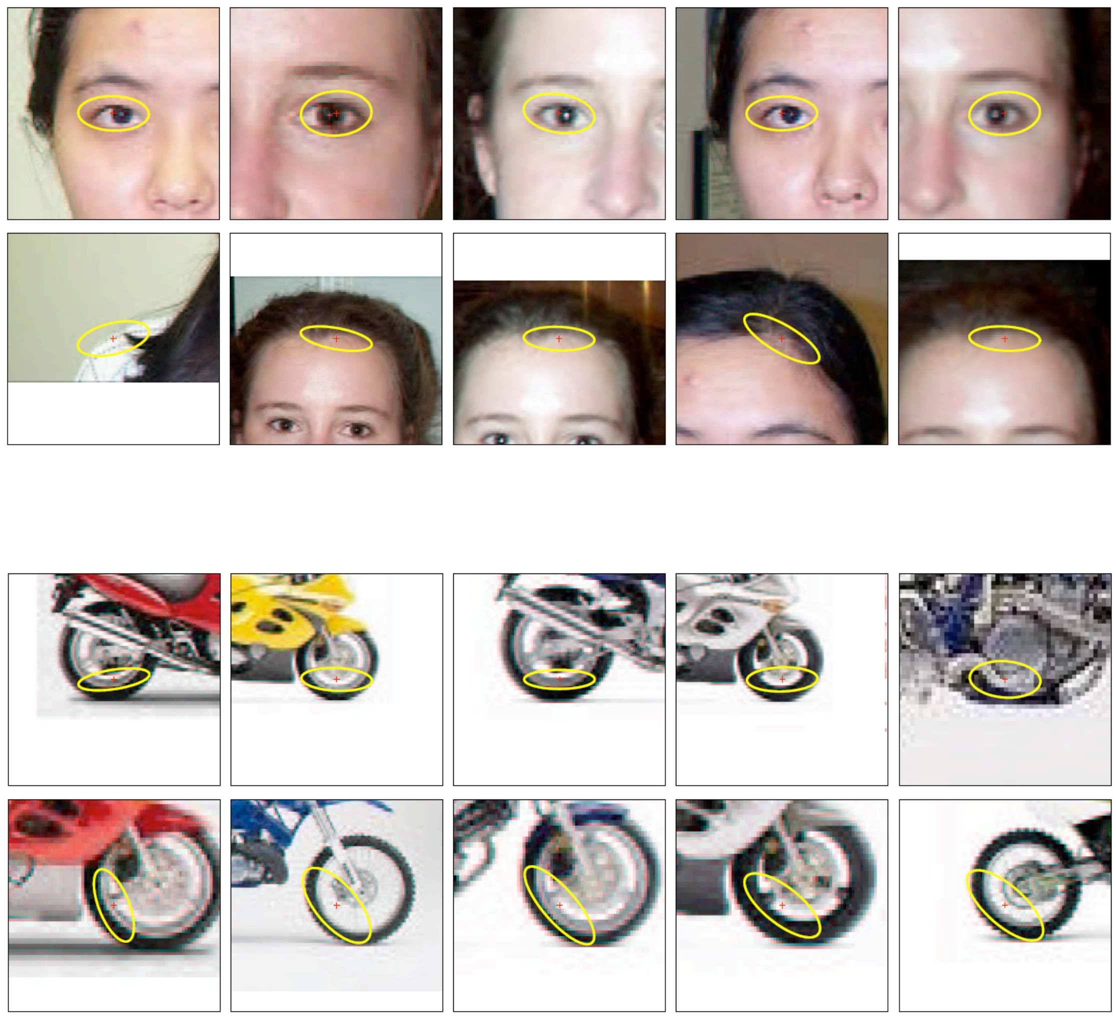


Slide credit: Bryan Russell & Josef Sivic



Slide credit: Bryan Russell & Josef Sivic

Most likely words given topic



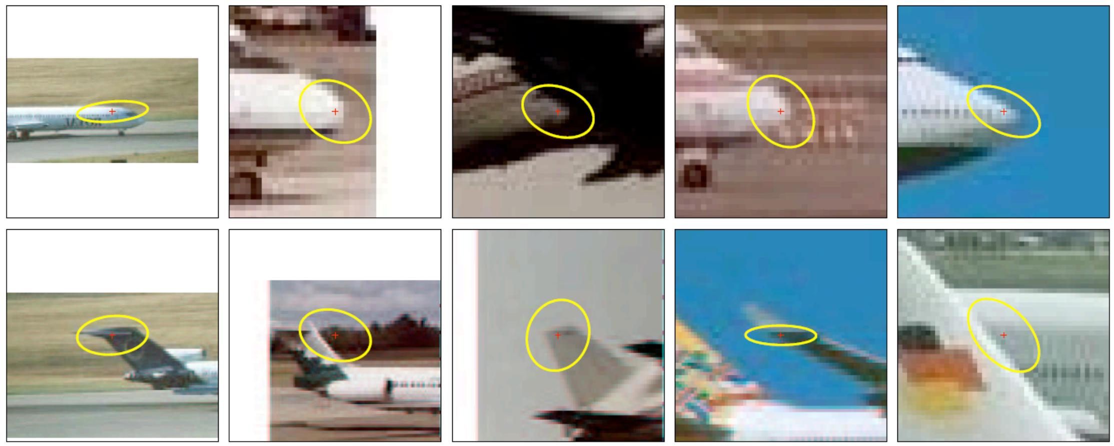
Topic 1: Word 1, Word 2

Topic 2: Word 1, Word 2

Slide credit: Bryan Russell & Josef Sivic

Most likely words given topic

Topic 3



Word 1

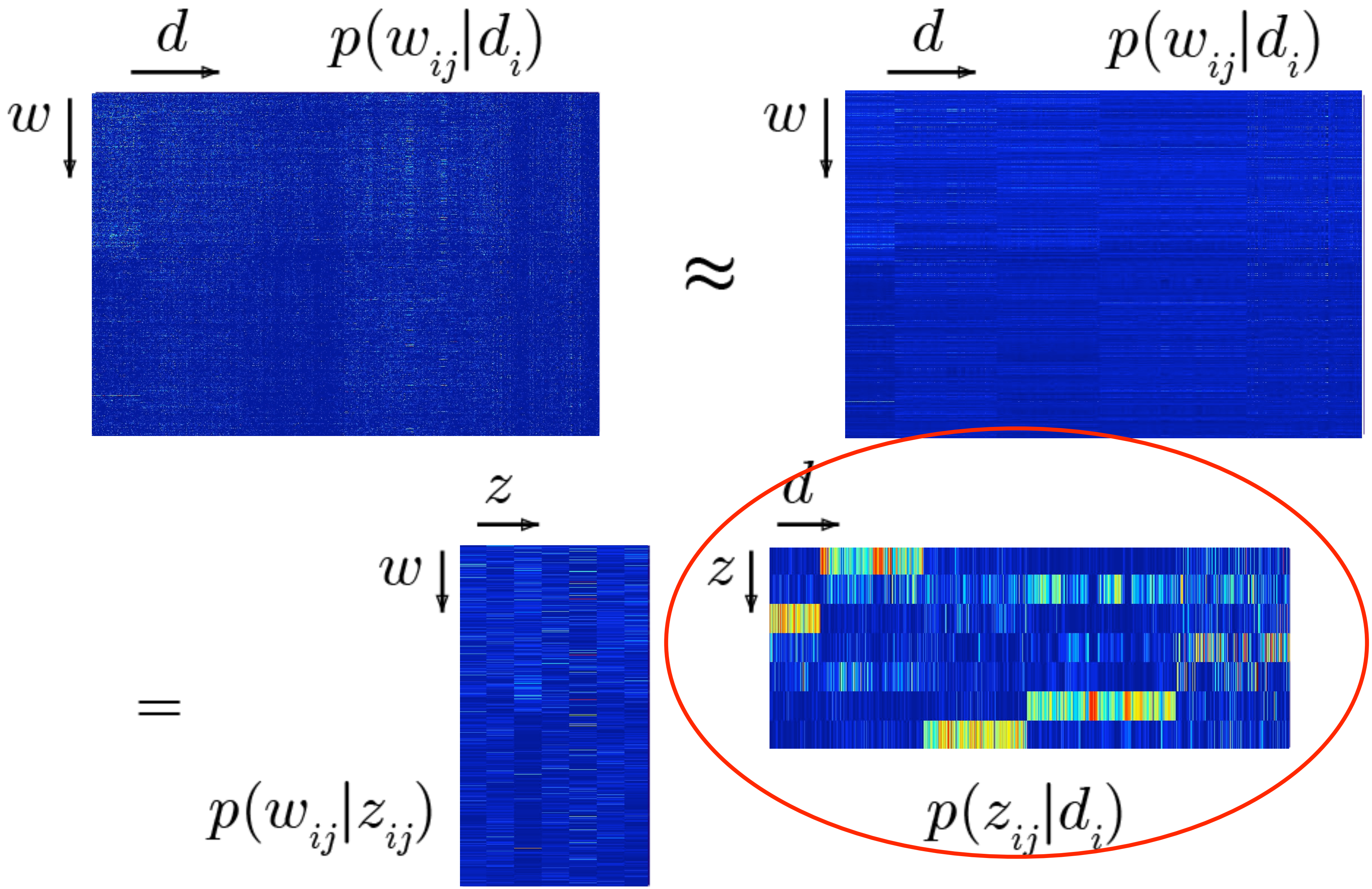
Word 2

Topic 4

Word 1

Word 2

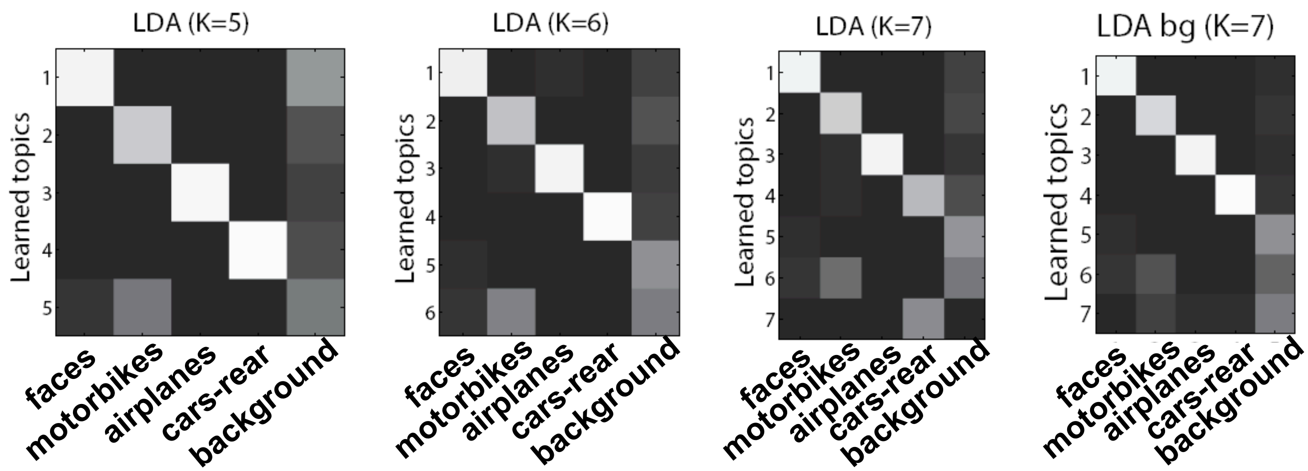
Slide credit: Bryan Russell & Josef Sivic



Slide credit: Bryan Russell & Josef Sivic

Image clustering

Confusion matrices:



Average confusion:

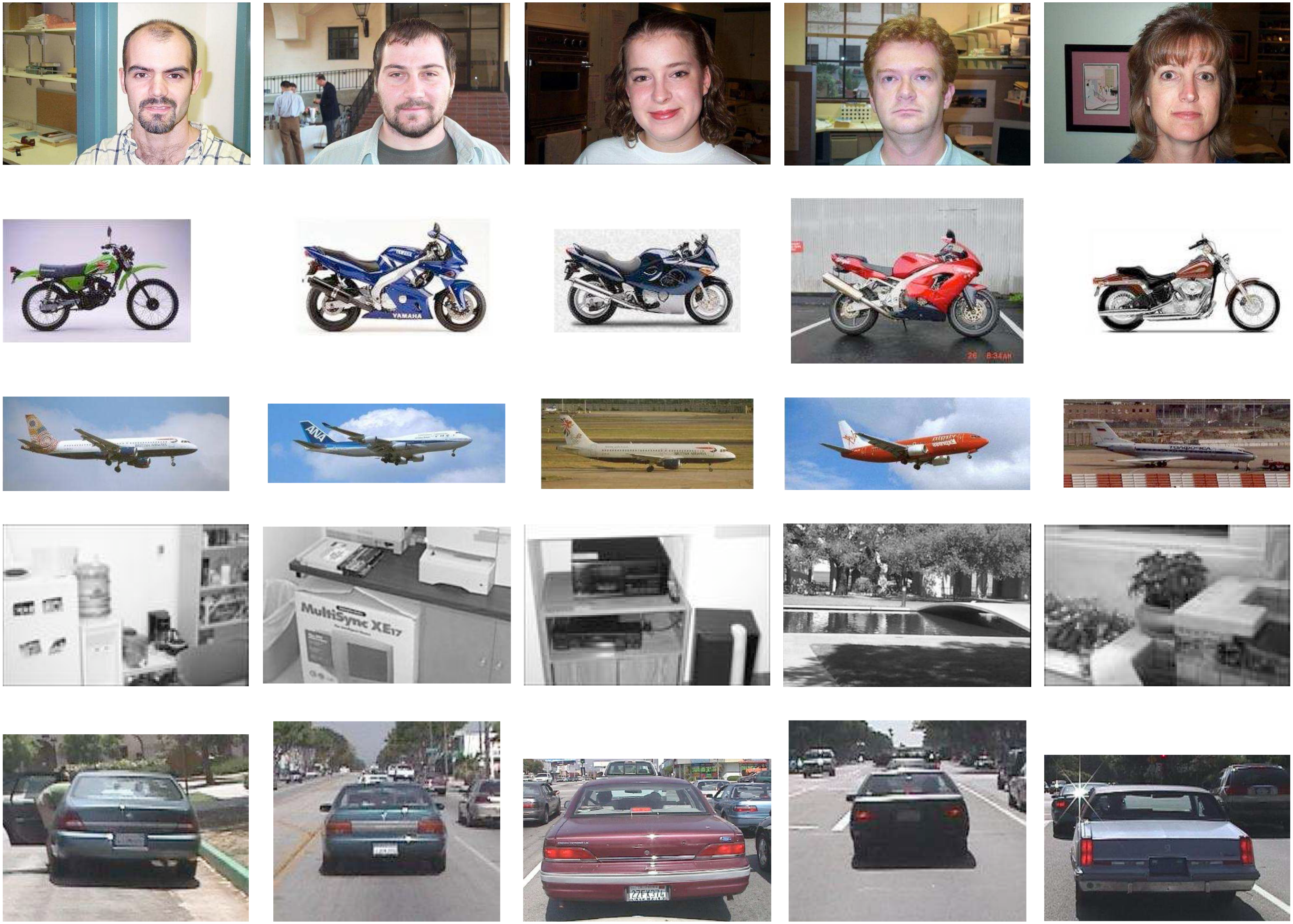
Expt.	Categories	T	LDA %	LDA #	pLSA %	pLSA #	KM baseline %	KM baseline #
(1)	4	4	97	86	98	70	72	908
(2)	4 + bg	5	78	931	78	931	56	1820
(2)*	4 + bg	6	84	656	76	1072	–	–
(2)*	4 + bg	7	78	1007	83	768	–	–
(2)*	4 + bg-fxd	7	90	330	93	238	–	–

Slide credit: Bryan Russell & Josef Sivic

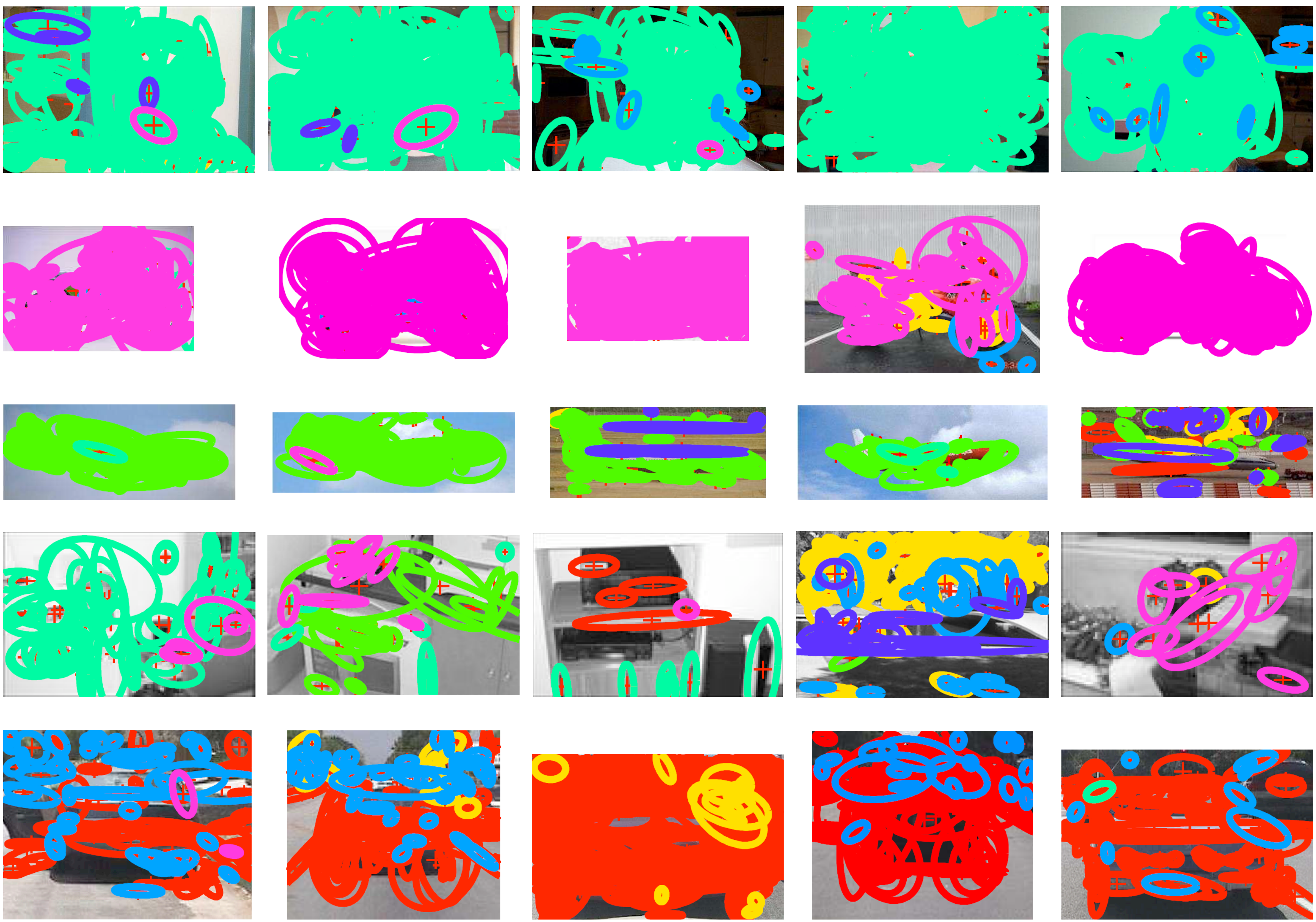
Image as a mixture of topics (objects)

Slide credit: Bryan Russell & Josef Sivic



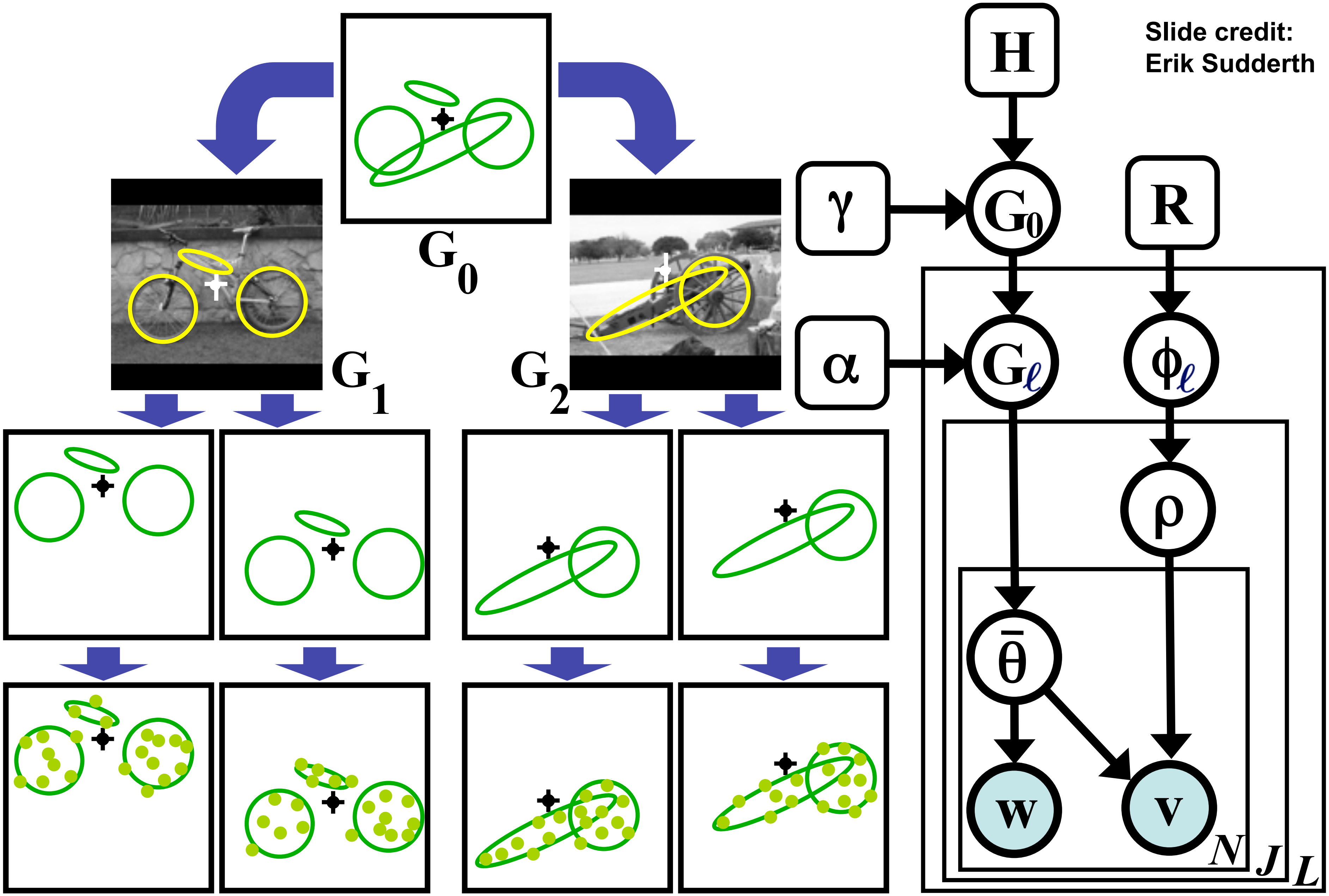
Slide credit: Bryan Russell & Josef Sivic



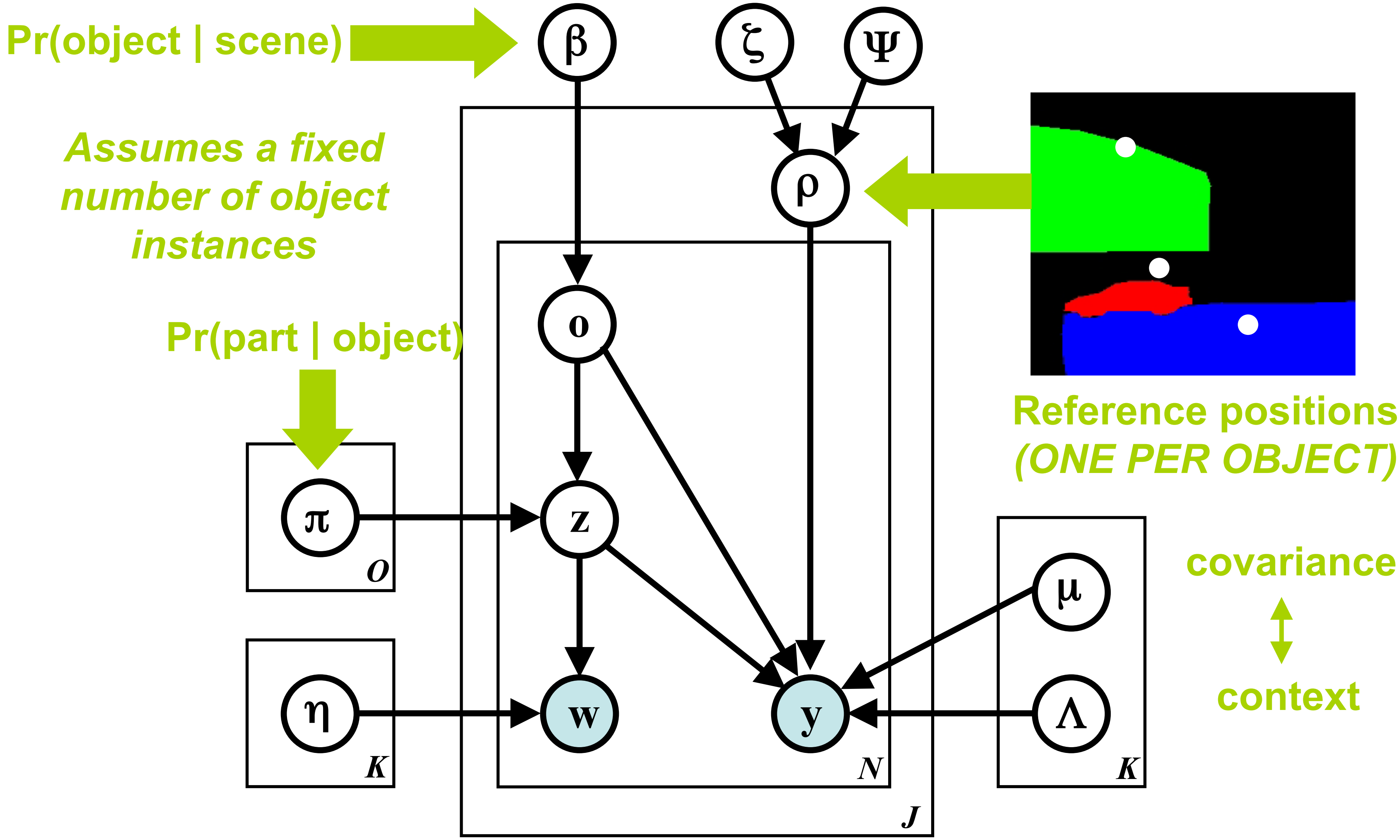
Slide credit: Bryan Russell & Josef Sivic

Hierarchical DP Object Model

Slide credit:
Erik Sudderth



Scenes of Fixed Sets of Objects



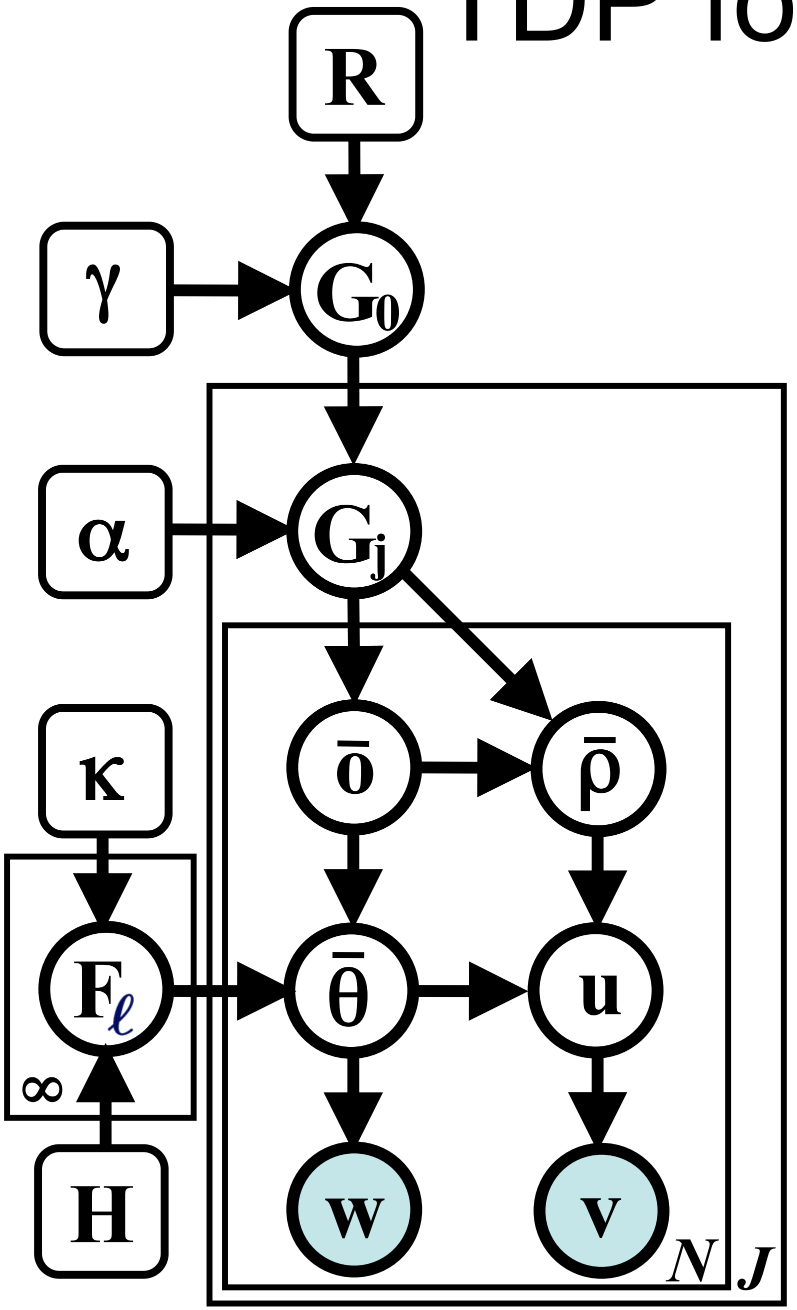
Slide credit: Erik Sudderth

Street Scene Segmentations

1-2 minutes Gibbs sampling per image

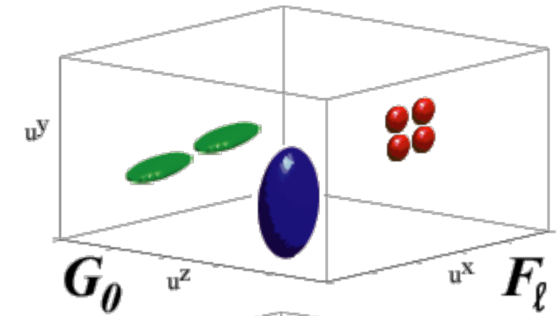
Slide credit: Erik Sudderth

TDP for 3D Scenes



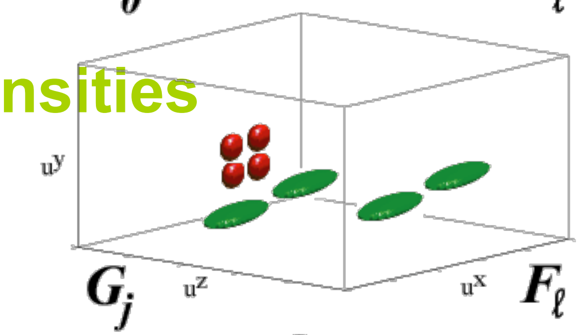
Global Density

Object category
Part size & shape
Transformation prior



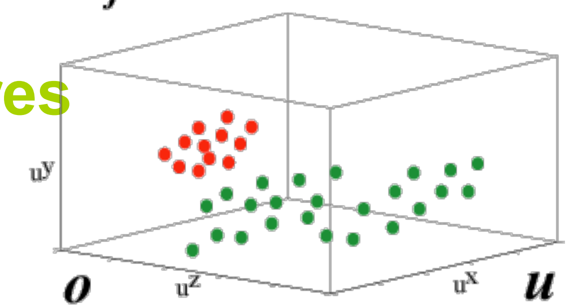
Transformed Densities

Object category
Part size & shape
Transformed locations



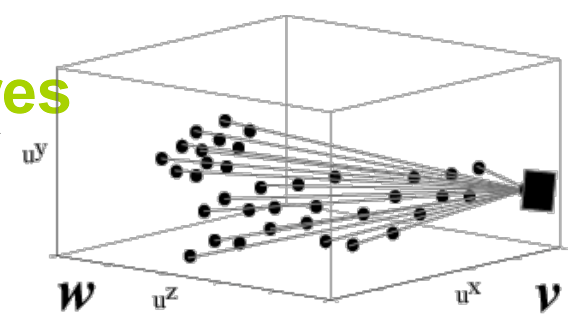
3D Scene Features

Object category
3D Location

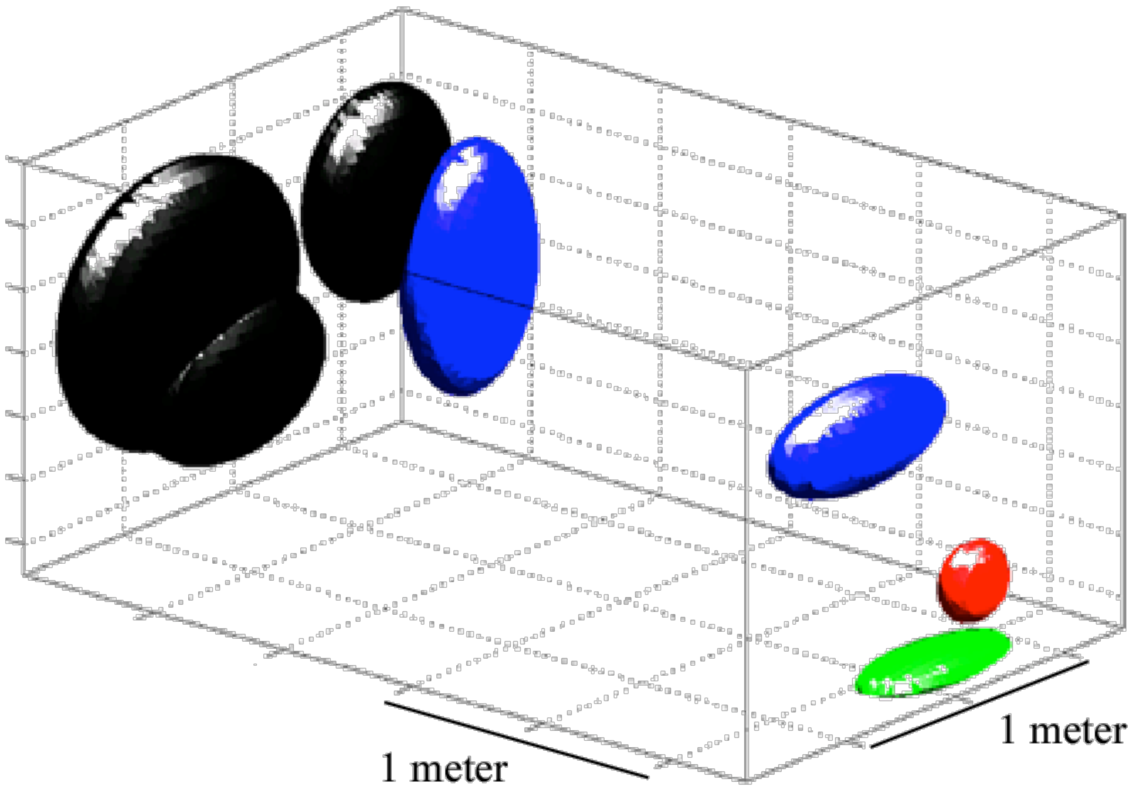


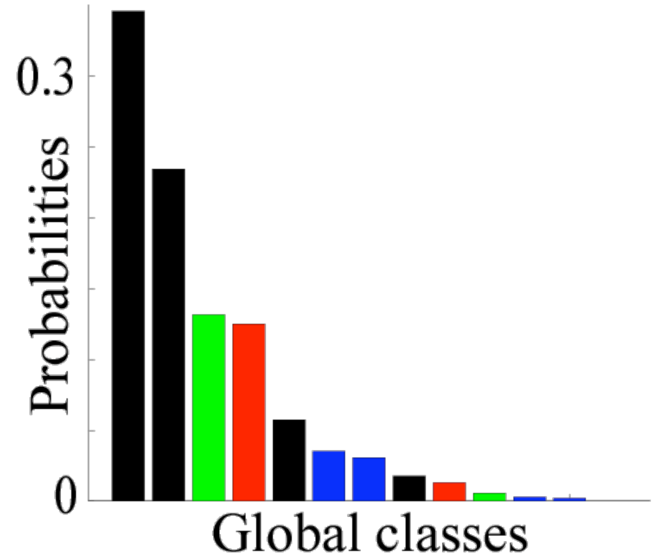
2D Image Features

Appearance Descriptors
2D Pixel Coordinates

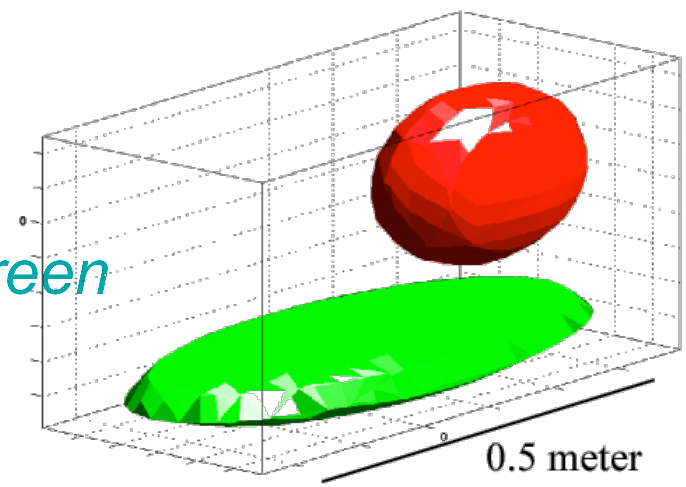


Single-Part Office Scene Model





Background *Bookshelves* *Computer Screen*
Desk



Slide credit: Erik Sudderth