Topic Models
(Generative Clustering Models)

Roman Stanchak and Prithviraj Sen

CMSC828G, Instructor: Prof. Lise Getoor

24^{th} April, 2008.

Outline

Outline

Motivating Applications

Mixed membership clustering of document copora:

- e.g., document $\longrightarrow$ words

Modeling consumer behaviour for marketing data:

- e.g., households $\longrightarrow$ trips $\longrightarrow$ products

Fraud detection in telecommunications:

- e.g., users $\longrightarrow$ call features

Protein function prediction:

- e.g., mixed membership of proteins to functional modules

Object detection/recognition in images:

- e.g., images $\longrightarrow$ feature patches

Connections to other Surveys

Collective classification:

- discriminative vs. generative
- Edo's talk, missing link model [Cohn and Hofmann, 2001]

Entity resolution:

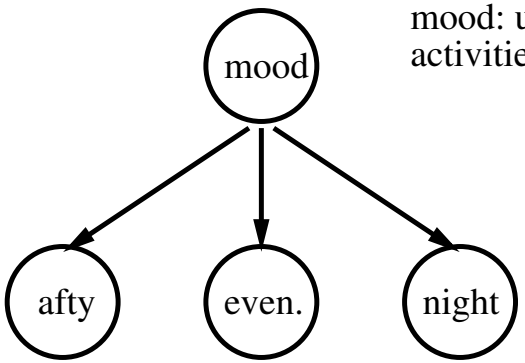
- LDA-ER

Group Detection Surveys:

- Stochastic Block Models
- Clustering in Relational Data/Community Detection

Outline

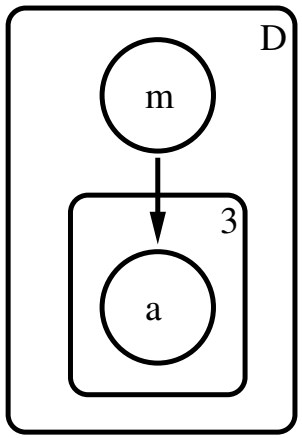
Plate Notation: A Slacker's Day Planner



mood: upbeat, bored, sad
activities: go to sleep, watch TV, go to pub, go to beach, go bowling

0.4	0.3	0.1	0.1	0.1

upbeat

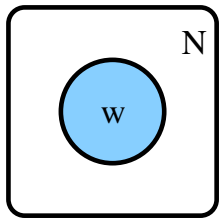


nodes
random variables

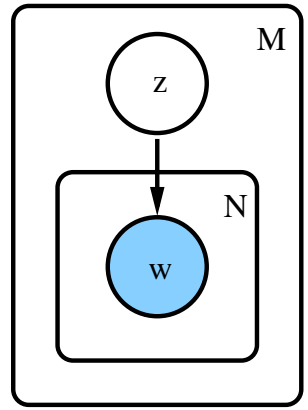
edges
dependencies

plates
repetitions

Unigram Model and Mixture of Unigrams



Unigram Model



Mixture of Unigrams

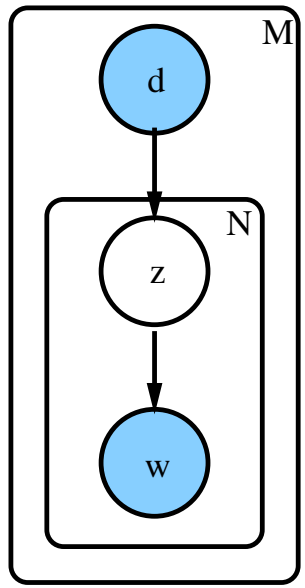
Disadvantages:

- Does not model documents dealing with a mixture of topics.

Mixture of Unigrams:

- Also known as, *naive bayes* model [McCallum and Nigam, 1998]
- Generative single class classification model

PLSI: Mixture Model for Text [Hofmann, 1999]



Advantage:

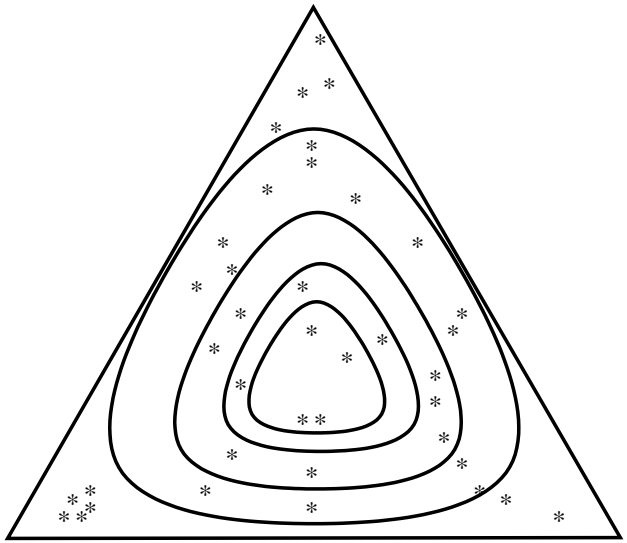
- First mixture model for documents

Disadvantage:

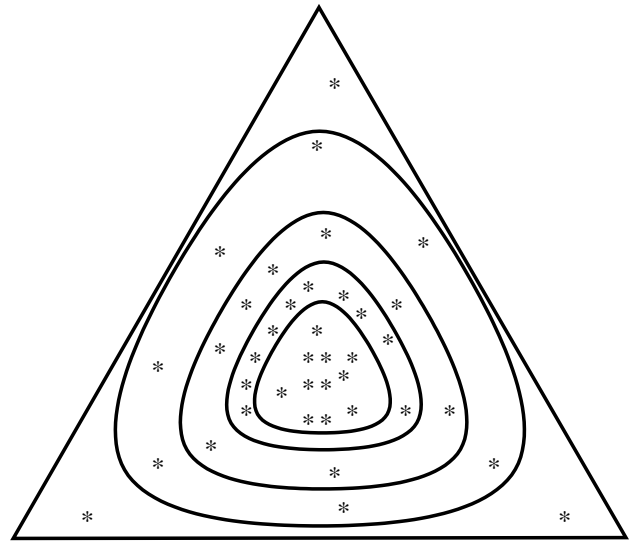
- Mixture parameters for each document, too many parameters
- Poor generalization properties

Problems with PLSI

2-D simplex showing the space of document mixtures for 3 topics

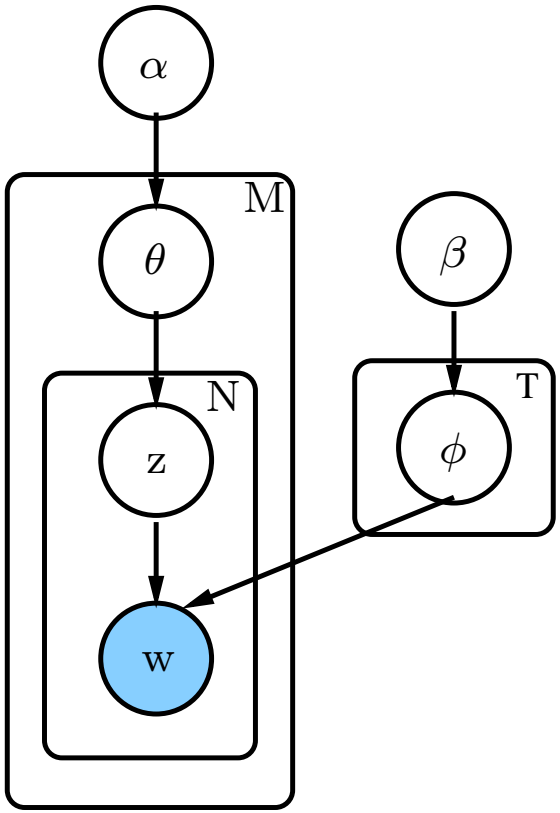


PLSI



LDA

Latent Dirichlet Allocation [Blei et al, 2003]



Generative process:

- Choose $\theta \sim Dir(\alpha)$
- For each word in doc:
  - Choose topic $z \sim mult(\theta)$
  - Choose word $w \sim mult(\phi_z)$

M	# of Documents
N	# of Words
T	# of Topics
w	Generated word
z	Topic of word w
θ	Distribution of topics
ϕ_z	Distribution of words given topic z
α	Dirichlet parameter
β	Dirichlet parameter

Discriminative vs. Generative

Word topics

arts	budget	education
new	million	school
film	tax	students
show	program	schools
music	budget	education
movie	billion	teachers
play	federal	high
musical	year	public
best	spending	teacher
⋮	⋮	⋮

Document mixtures

- θ_{29795}: ..... wanted to play jazz ....
- θ_{1883}: .... play ... performed ... stage ....
- θ_{21359}: ..... don and jim play the game ....
- The θ's estimated for each document can be used as a low dim. rep. for the doc., can be used to classify the docs.

Gibbs Sampling for LDA [Griffiths and Steyvers, 2004]

$$P(z_i = j|\mathbf{z}_{-j}, \mathbf{w}) = \underbrace{\frac{n^{w_i}_{-i,j} + \beta}{\sum_{w_i} n^{w_i}_{-i,j} + W\beta}}_{\text{prob. of } w_i \text{ under topic } j} \overbrace{\frac{n^{d_i}_{-i,j} + \alpha}{\sum_j n^{d_i}_{-i,j} + T\alpha}}^{\text{prob. of } z_i \text{ in doc containing } w_i}$$

- Perform burn-in
- Run iterations of the Gibbs sampler collecting samples after regular intervals
- For each iteration:
  - For word w_i in corpus, sample z_i from $P(z_i = j|\mathbf{z}_{-i}, \mathbf{w})$
- Straightforward to recover θ's and ϕ's after Gibbs sampler has converged

About LDA and Gibbs Sampling

Why dirichlet?

- Conjugate prior of multinomial. Lets you analytically integrate over θ and ϕ.

Why multinomial?

- Legacy reasons.
- Multinomial does not model bursty nature of text [Madsen et al, 2005].

Gibbs sampling vs. variational methods:

- Gibbs sampling is slower (takes days for mod.-sized datasets), variational inference takes a few hours.
- Gibbs sampling is more accurate.
- Gibbs sampling convergence is difficult to test, although quite a few machine learning approximate inference techniques also have the same problem.
- More sophisticated Gibbs Sampling based on split/merge techniques are available (see [Jain and Neal, 2000]).

Outline

The Missing Link [Cohn and Hofmann, 2001]

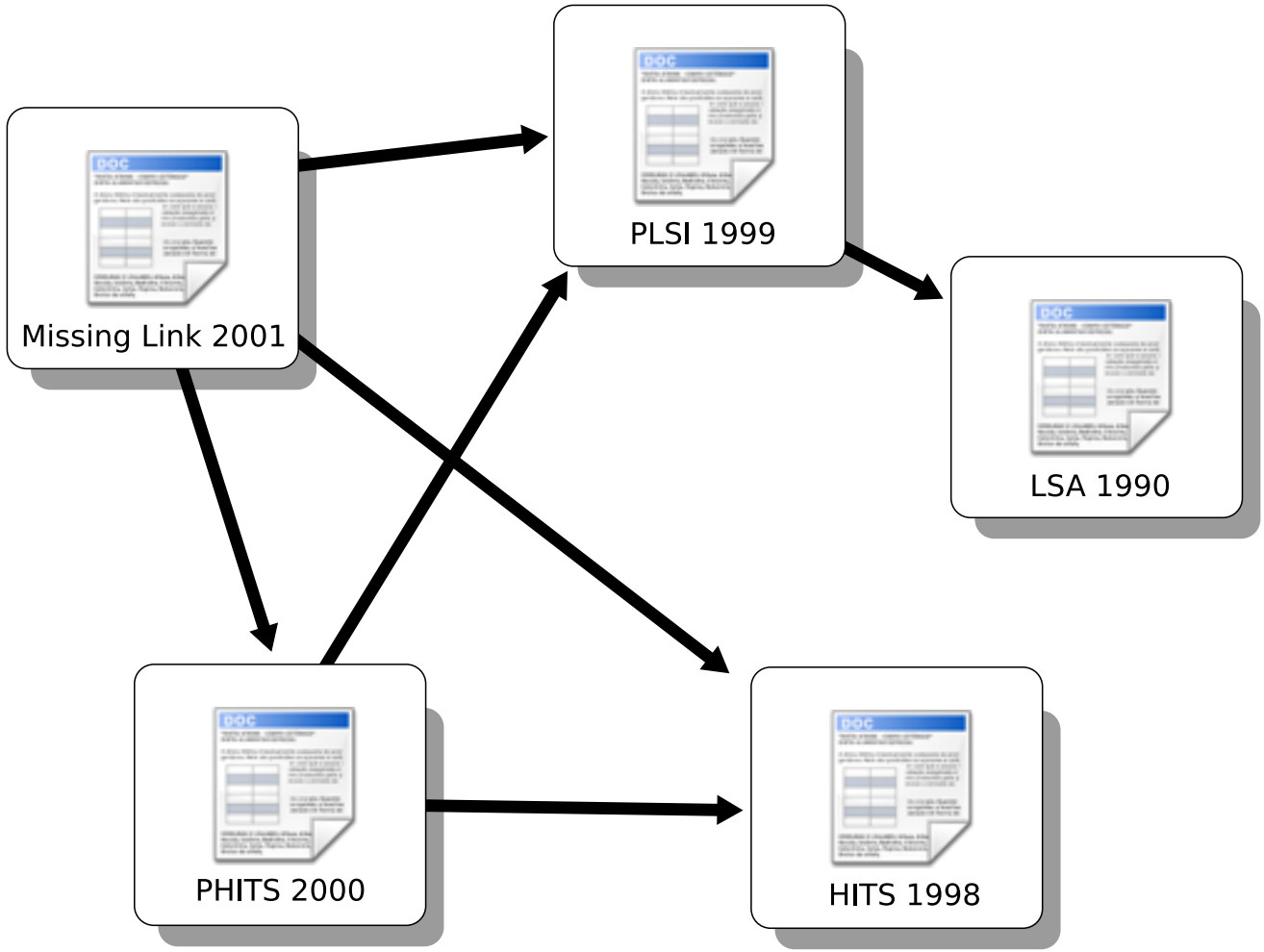
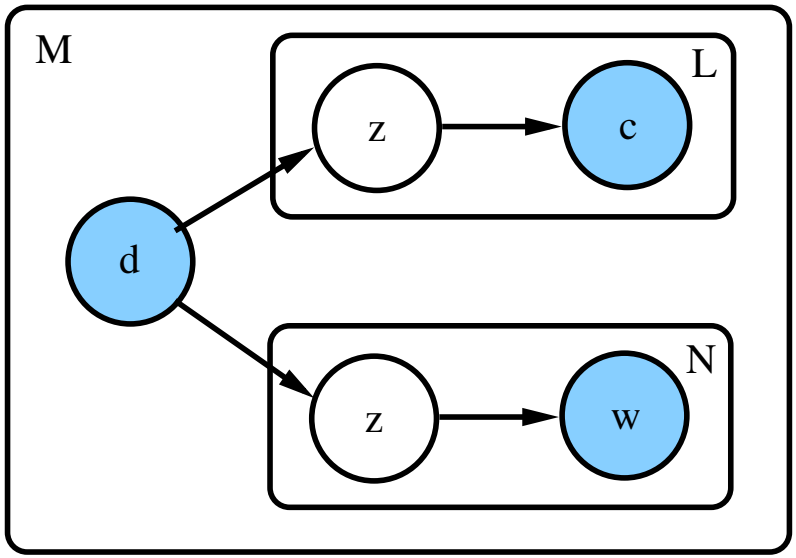


Figure: Document topics are influenced by citations as well as content.

The Missing Link [Cohn and Hofmann, 2001]



w	Generated word
z_w	Topic of word w
c	Generated link
z_c	Topic of link c
N	# of Words
L	# of Links
M	# of Documents

Generative Process

For each of M documents d,

- For each of N words in document d, draw:
  - Topic z_w from $P(\textbf{topic}|\textbf{doc})$
  - Word w from $P(\textbf{word}|\textbf{topic})$
- For each of L links in document d, draw:
  - Topic z_c from $P(\textbf{topic}|\textbf{doc})$
  - Link c from $P(\textbf{link}|\textbf{topic})$

The Missing Link [Cohn and Hofmann, 2001]

Summary

- Joint probabilistic model for content and links.
- *Interpolates* between PLSA and PHITS
- Improves classification accuracy over standard PLSA and PHITS on Cora and WebKB.

Limitations

- Suffers from same over-fitting problems of PLSA
- Performance is dependent on α weighting term

The Missing Link

Questions?

Author Topic Model [Rosen-Zvi, et al. 2004]

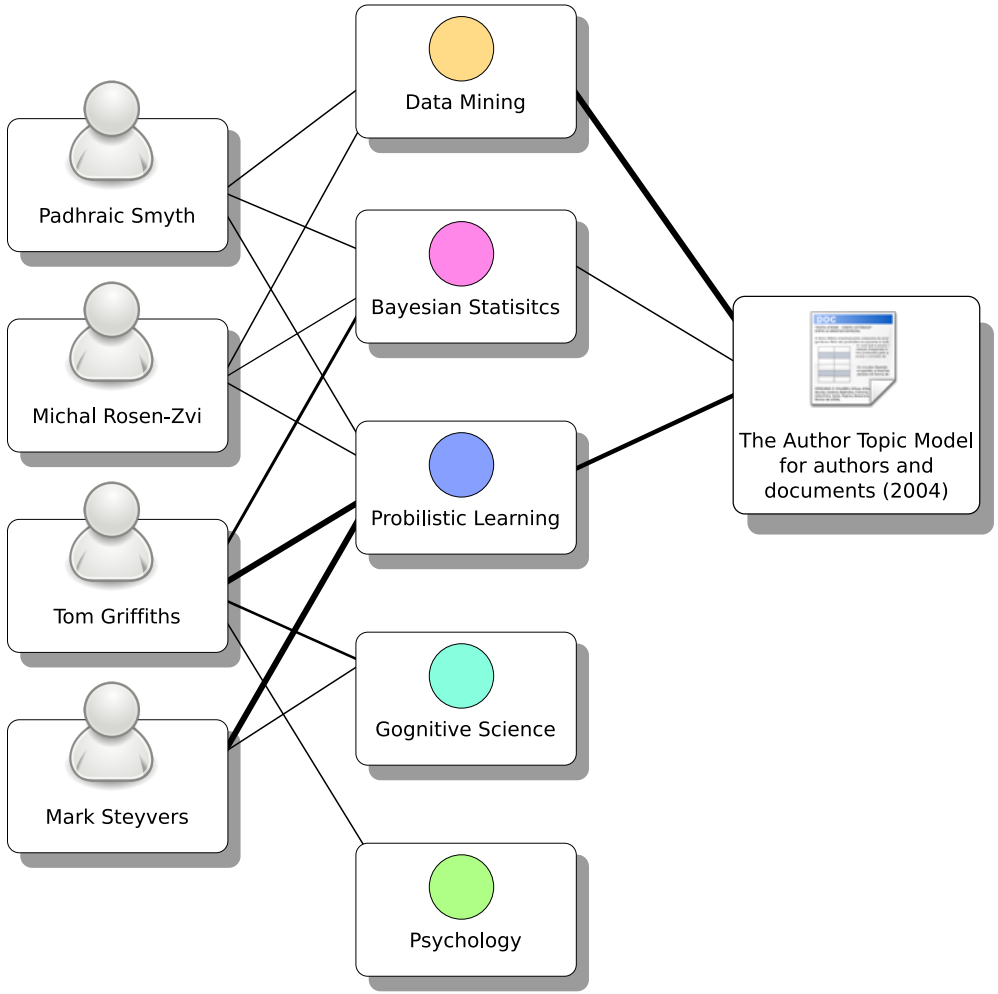
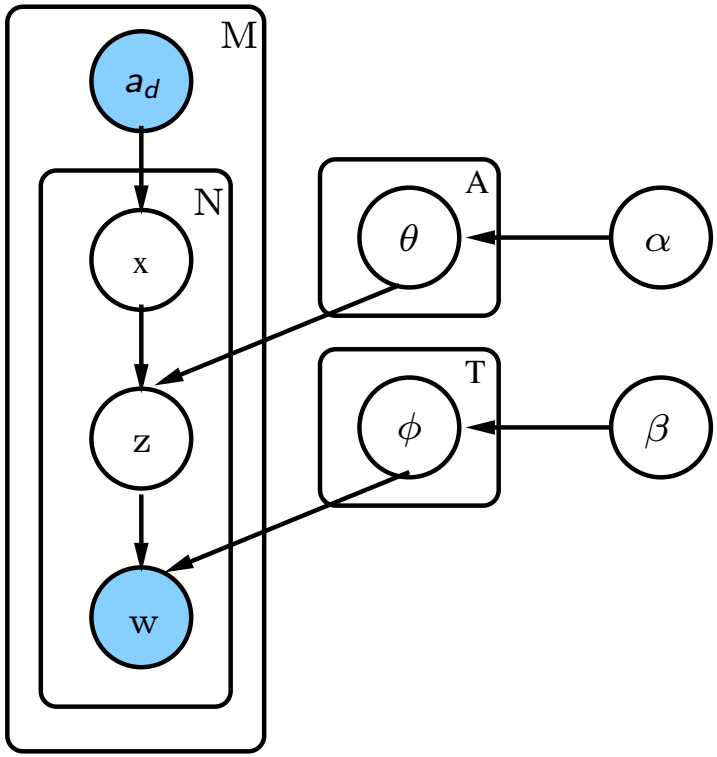


Figure: Authors influence topic selection

Author Topic Model [Rosen-Zvi, et al. 2004]



w	Generated word	M	# of Documents
z	Topic of word w	N	# of Words
x	Author of word w	A	# of Authors
$\mathbf{a_d}$	Authors of document d	T	# of Topics
θ_x	Distribution of topics given author x		
ϕ_z	Distribution of words given topic z		
α	Dirichlet parameter		
β	Dirichlet parameter		

Generative process:

- Choose $\theta \sim Dir(\alpha)$
- Choose $\phi \sim Dir(\beta)$
- For each word w in doc d:
  - Given the set of authors, $\mathbf{a_d}$, choose an author x uniformly from $\mathbf{a_d}$.
  - Choose topic $z \sim mult(\theta_x)$ θ_x *is author specific*
  - Choose word $w \sim mult(\phi_z)$ ϕ_z *is topic specific*

Author Topic Model

TOPIC 10	
WORD	PROB.
SPEECH	0.1134
RECOGNITION	0.0349
WORD	0.0295
SPEAKER	0.0227
ACOUSTIC	0.0205
RATE	0.0134
SPOKEN	0.0132
SOUND	0.0127
TRAINING	0.0104
MUSIC	0.0102
AUTHOR	PROB.
Waibel_A	0.0156
Gauvain_J	0.0133
Lamel_L	0.0128
Woodland_P	0.0124
Ney_H	0.0080
Hansen_J	0.0078
Renals_S	0.0072
Noth_E	0.0071
Boves_L	0.0070
Young_S	0.0069

TOPIC 209	
WORD	PROB.
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	PROB.
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048

TOPIC 87	
WORD	PROB.
USER	0.2541
INTERFACE	0.1080
USERS	0.0788
INTERFACES	0.0433
GRAPHICAL	0.0392
INTERACTIVE	0.0354
INTERACTION	0.0261
VISUAL	0.0203
DISPLAY	0.0128
MANIPULATION	0.0099
AUTHOR	PROB.
Shneiderman_B	0.0060
Rauterberg_M	0.0031
Lavana_H	0.0024
Pentland_A	0.0021
Myers_B	0.0021
Minas_M	0.0021
Burnett_M	0.0021
Winiwarter_W	0.0020
Chang_S	0.0019
Korvemaker_B	0.0019

TOPIC 20	
WORD	PROB.
STARS	0.0164
OBSERVATIONS	0.0150
SOLAR	0.0150
MAGNETIC	0.0145
RAY	0.0144
EMISSION	0.0134
GALAXIES	0.0124
OBSERVED	0.0108
SUBJECT	0.0101
STAR	0.0087
AUTHOR	PROB.
Linsky_J	0.0143
Falcke_H	0.0131
Mursula_K	0.0089
Butler_R	0.0083
Bjorkman_K	0.0078
Knapp_G	0.0067
Kundu_M	0.0063
Christensen-J	0.0059
Cranmer_S	0.0055
Nagar_N	0.0050

Figure: An illustration of 4 topics from a 300-topic solution for the CiteSeer collection. Each topic is shown with the 10 words and authors that have the highest probability conditioned on that topic [Rosen-Zvi, et al. 2004].

Author Topic Model

Summary

- Similar to LDA, but assumes that a topic z is generated by author x from the author-specific topic distribution θ_x.
- Increased descriptive ability in applications using authorship information.
  - Automated reviewer recommendation for research papers
- Predictive ability is better than LDA with small training sets.
  - But LDA improved with a larger training set and more topics

Hierarchical Topic Models

Observation

Topics aren't independent.

Example

- The topic of CS consists of AI, Systems, Theory, etc.
- AI consists of NLP, Machine Learning, Robotics, Vision, etc.

Question

How to encode dependencies between topics?

Pachinko Allocation Model[Li et al, 2006]

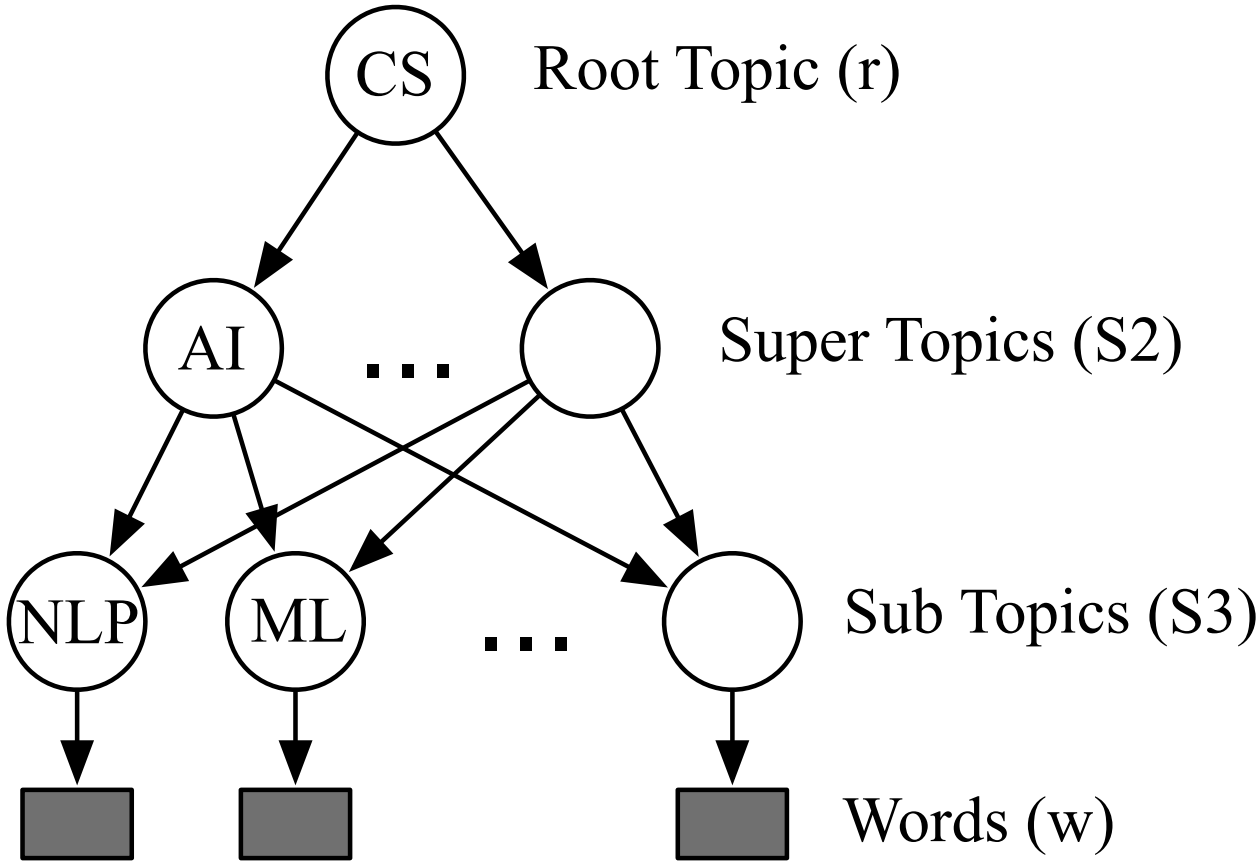


Figure: Four-level Pachinko Model

Pachinko Allocation

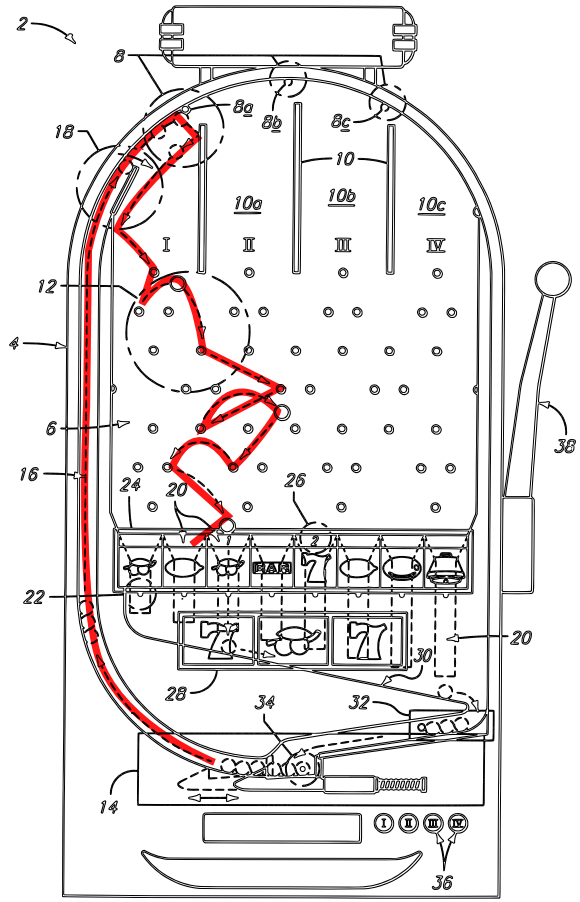


Figure: Pachinko Machine – A path of the ball is shown in red.

From `http://www.freepatentsonline.com/6619659.html`

Pachinko Allocation Model

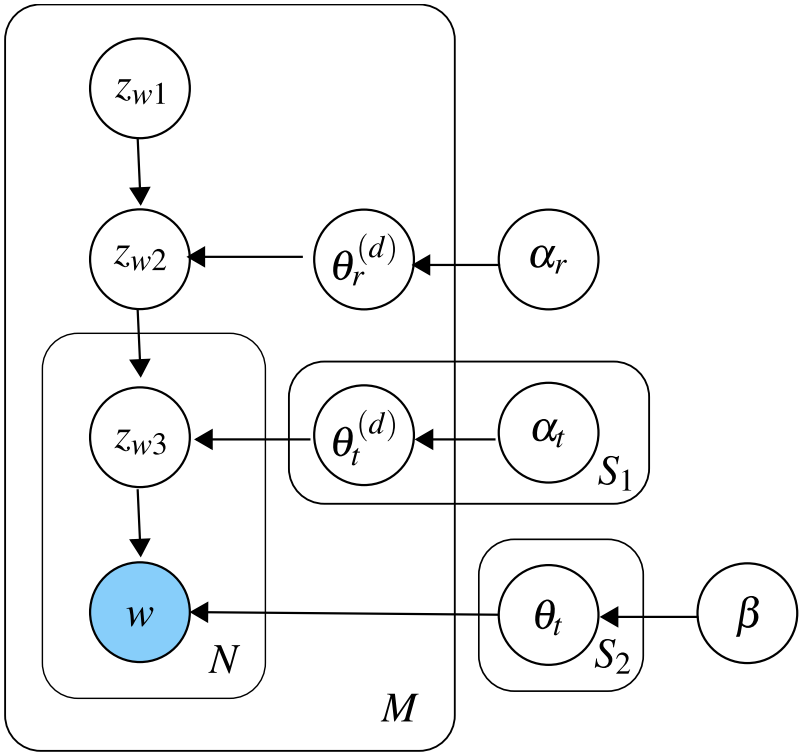


Figure: 4-level Pachinko Model

w	Generated word	M	# of Documents
z_{w1}	Root topic	N	# of Words
z_{w2}	Super topic	S_1	# of Super topics
z_{w3}	Sub topic	S_2	# of Sub topics

Generative Process

- For each topic, sample $\theta \sim Dir(\alpha)$
- For each word w in the document,
  - Sample topic path z_w starting at the root topic node and terminating at a leaf node. Each $z_i \sim mult(\theta)$.
  - Sample word w from $mult(\theta)$ of the last last topic along the path

Pachinko Allocation Model

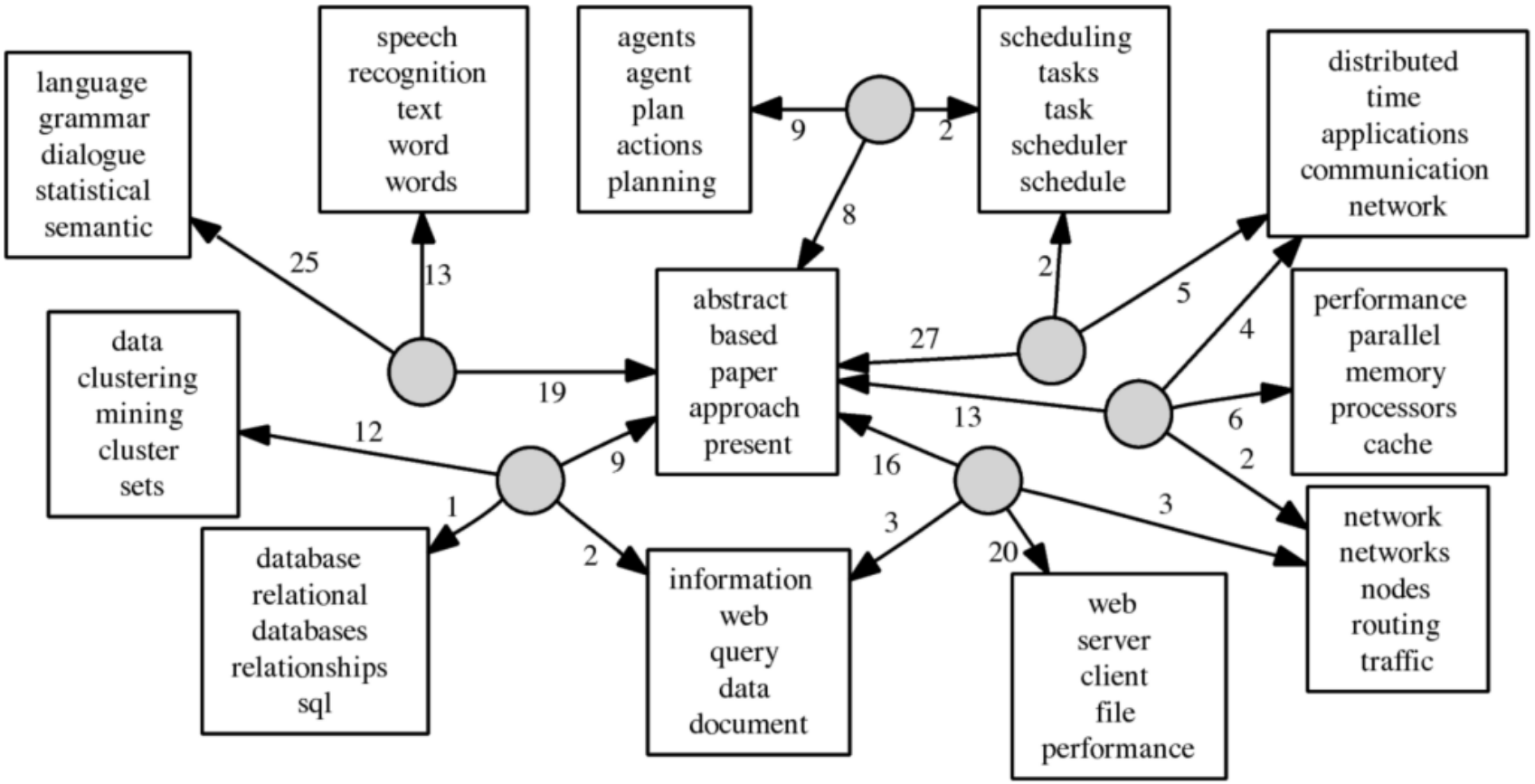


Figure: Discovered topics (circles), sub-topics (squares), and their dependencies (Figure from [Li et al, 2006]).

Pachinko Allocation Model

Summary

- Fixed tree of topics, word distributions as leaves
- Captures arbitrary, sparse and nested correlations between topics.
- Use Gibbs Sampling for inference and parameter estimation.
- Better performance than competing models:
  - Derived more intuitive topics than LDA on NIPS dataset (according to human judges)
  - Higher likelihood than LDA, CTM and HDP on NIPS dataset
  - Higher document classification accuracy than LDA on 20 newsgroup dataset.

Limitations

- Number of topics is fixed
- Depth of tree must be pre-specified

Other Hierarchical Models

Hierarchical LDA[Blei, et al. 2003]

- A document is generated by sampling words from the topics along *a single path* from the root to leaf node of a topic tree.
- Tree depth L is fixed, the # of topics is inferred using a nested CRP.

Correlated Topic Model[Blei and Laferty, 2006]

- Similar to LDA, but uses Logistic Gaussian prior instead of Dirichlet.
  - Not really hierarchical
  - Covariance matrix Σ models pair-wise correlation
- Many parameters to estimate – Σ grows with the square of the number of topics $\rightarrow$ slow inference.

Nonparametric Bayes Pachinko Allocation[Li et al, 2007]

- Similar to PAM, uses Hierarchical Dirichlet Process to infer # of topics

Beyond Bag of Words

Bag of Words Assumption

Assumes that words order in a document is irrelevant.

- It is mathematically convenient, but not strictly true!!!

Problem

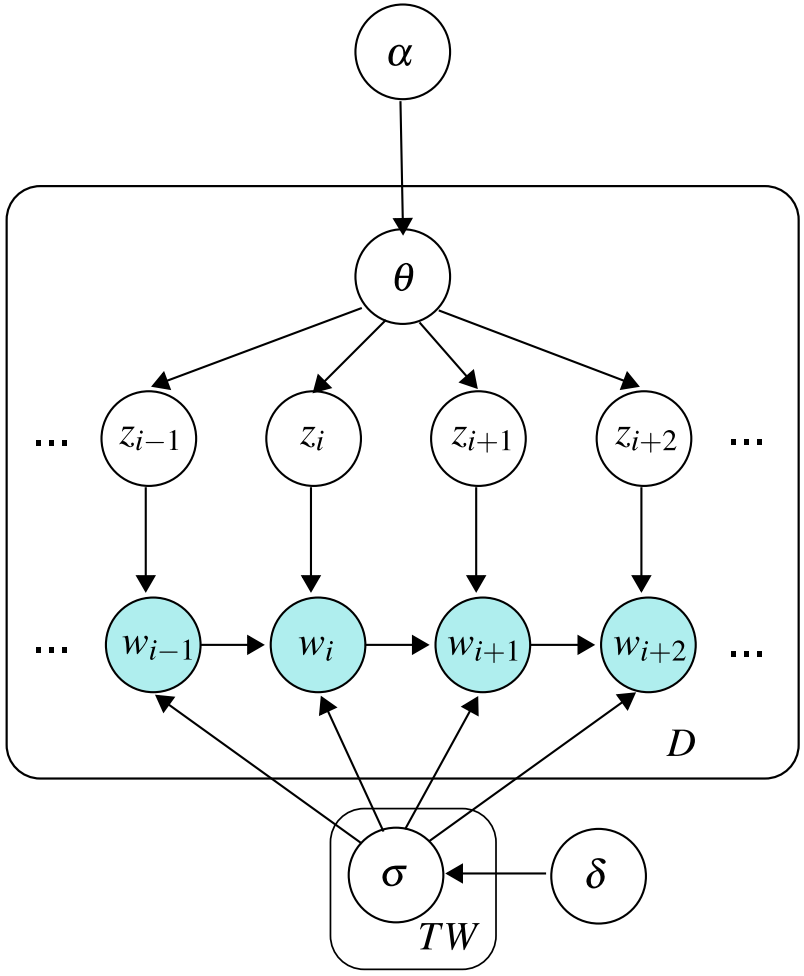
Under these models all of the following sentences are equally likely:

- *the department chair couches offers*
- *the department chair offers couches*
- *couches the chair department offers*

Solution

Explicitly incorporate word order into graphical model.

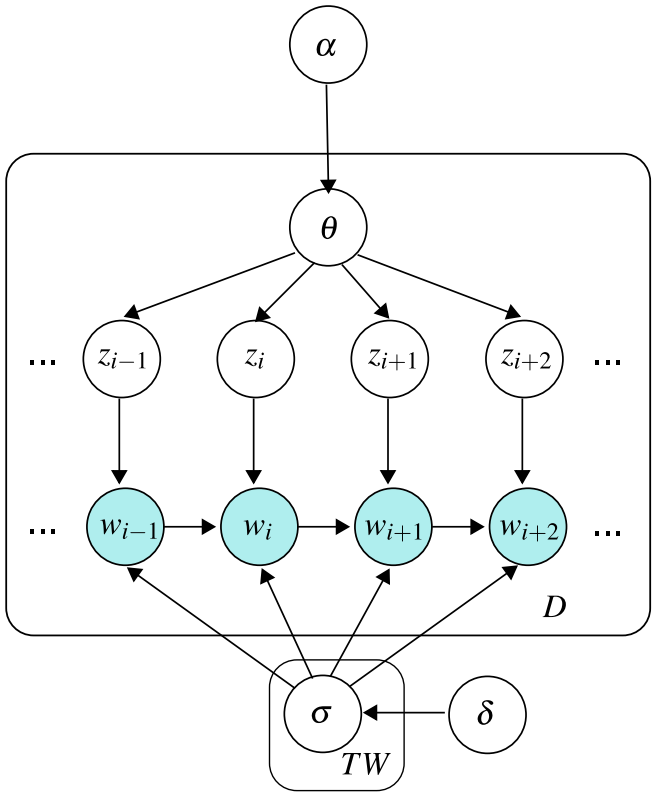
Bigram Topic Model [Wallach, 2006]



Summary

- Similar to LDA, except distribution of word w_i is dependent on the topic and the previous word w_{i-1}.

Bigram Topic Model [Wallach, 2006]



Generative Process

- for each topic, word pair (z, w), draw a discrete distribution σ_{zw} from a Dirichlet prior δ
- for each document d, draw a discrete distribution $\theta^{(d)}$
- For each position i in document d, draw:
  a topic $z_i^{(d)}$ from Discrete($\theta^{(d)}$)
  a word $w_i^{(d)}$ from Discrete(σ_{zw})

Bigram Topic Model

LDA Topic Model

the	i	that	easter
"number"	is	proteins	ishtar
in	satan	the	a
to	the	of	the
espn	which	to	have
hockey	and	i	with
a	of	if	but
this	metaphorical	"number"	english
as	evil	you	and
run	there	fact	is

Bigram Topic Model

party	god	"number"	the
arab	believe	the	to
power	about	tower	a
as	atheism	clock	and
arabs	gods	a	of
political	before	power	i
are	see	motherboard	is
rolling	atheist	mhz	"number"
london	most	socket	it
security	shafts	plastic	that

Figure: Comparison of discovered topics between LDA and Bigram model (From [Wallach, 2006])

Bigram Topic Model

Performance

- Lower *Information Rate* than LDA for Psychology Abstracts dataset and 20 Newsgroups Dataset
- 10-20s per Gibbs iteration (at 60 topics)

Limitations

- Simple model, always generates a bigram.
- Many parameters to infer

LDA Composite Model [Griffiths et al, 2004]

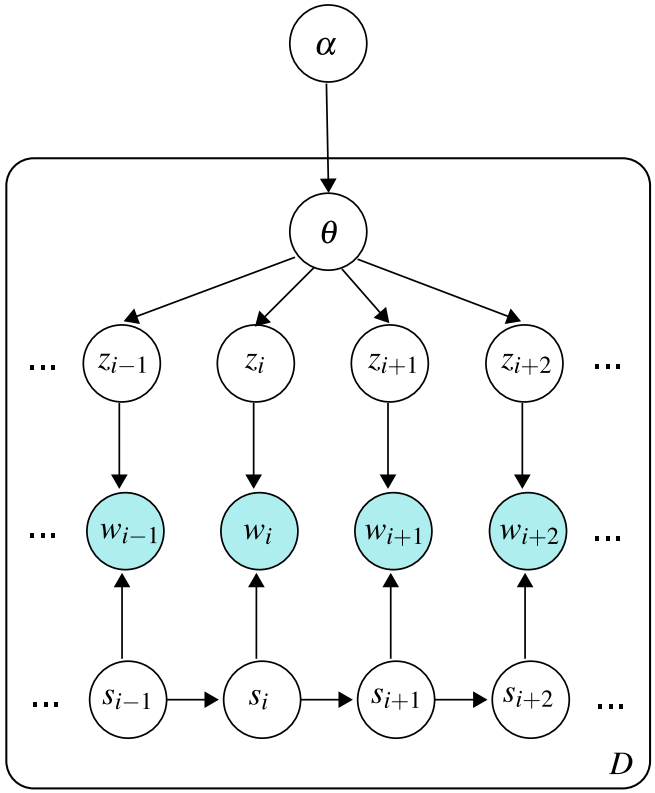


Figure: LDA Composite Plate Model

Summary

- Similar to Bigram model, but overlays an HMM over the word sequence.
  - Allows integration of syntactic models.
- Empirical Performance:
  - Higher quality topics than LDA
  - Likelihood of held out data is higher than LDA
  - Part of speech tagging significantly better than HMM and Distributional Clustering for 10 high-level tags.
  - Somewhat *worse* performance on document classification task than LDA.

Topic Models: Extensions

Questions?

Application: Object Recognition in Images

General Goal

Given an image, determine if it contains a particular object

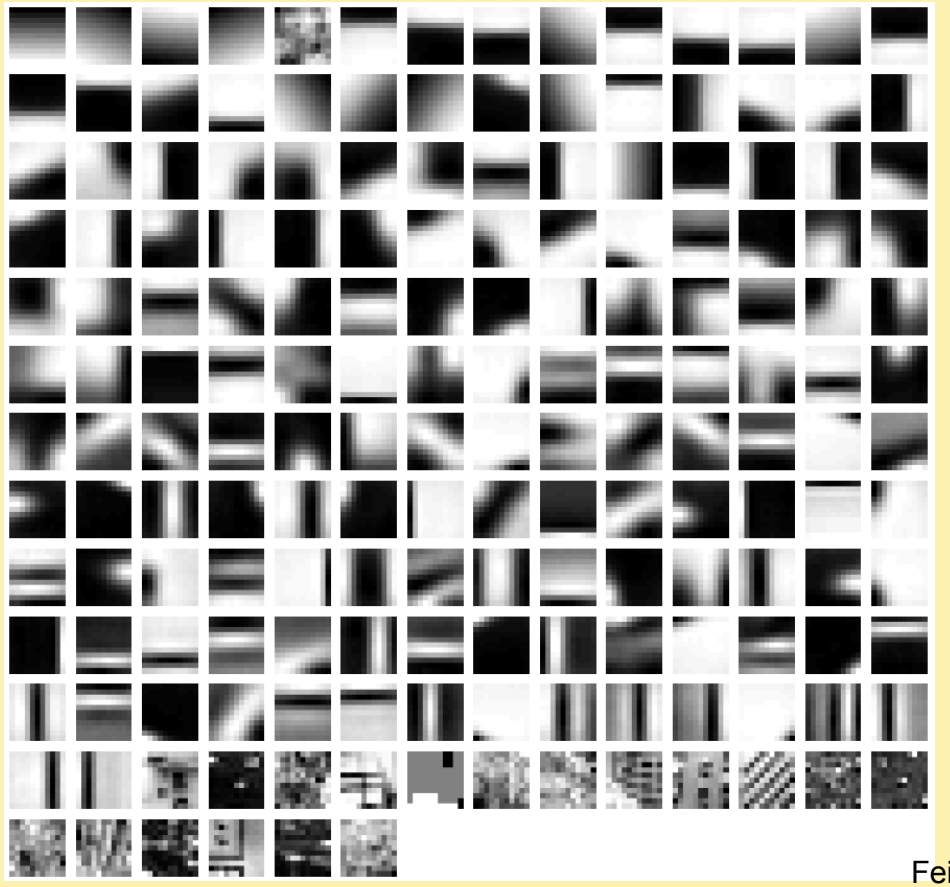
Approach

Model a database of labeled images using mixtures of topics, where:

- Each image is a document
- Image feature patches correspond to *visual words*
- Each object class label corresponds to a distribution of topics.

Application: Object Recognition in Images

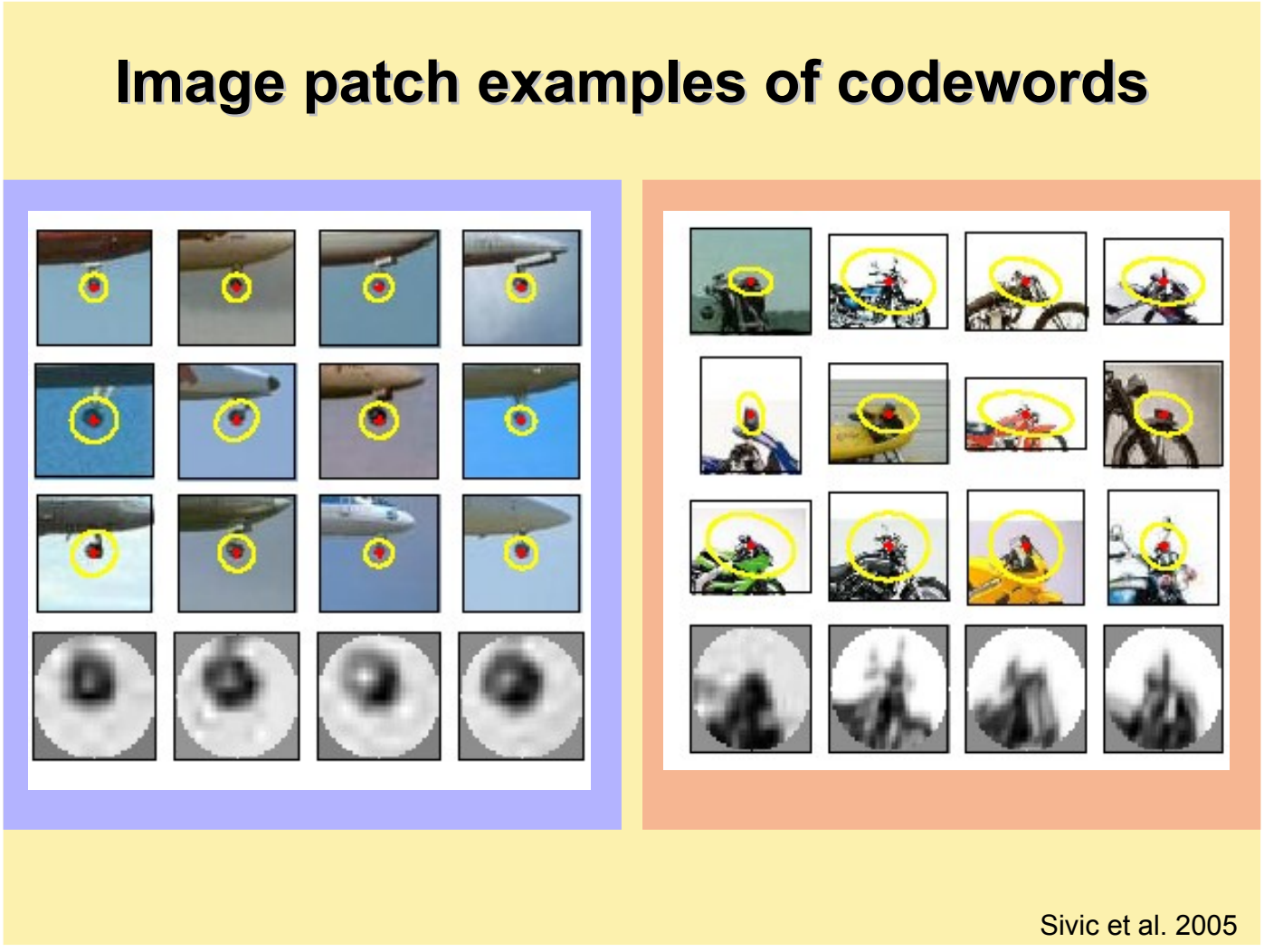
2. Codewords dictionary formation



Fei-Fei et al. 2005

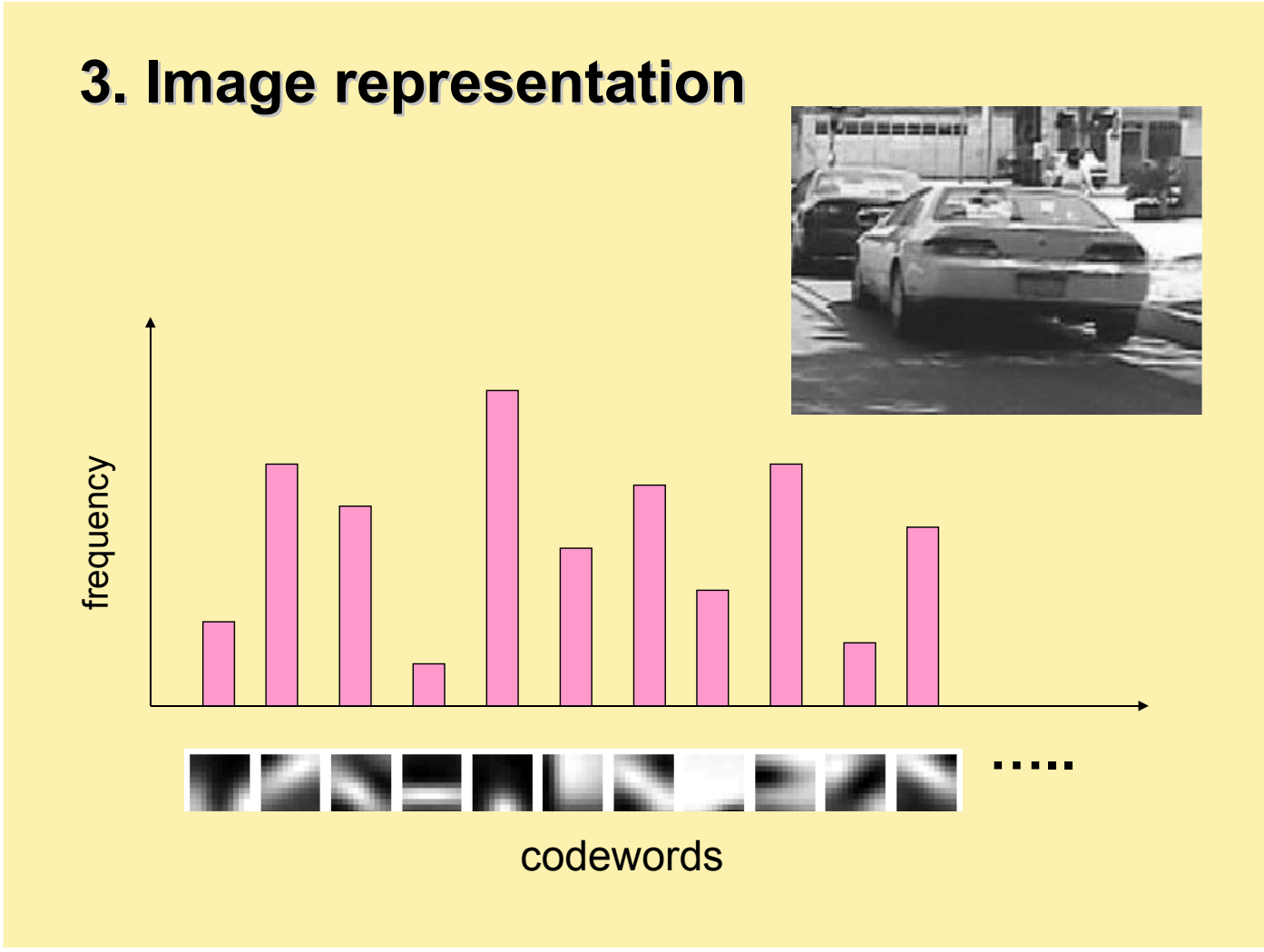
Slides from [CVPR 2007 Short Course on Object Recognition]

Application: Object Recognition in Images



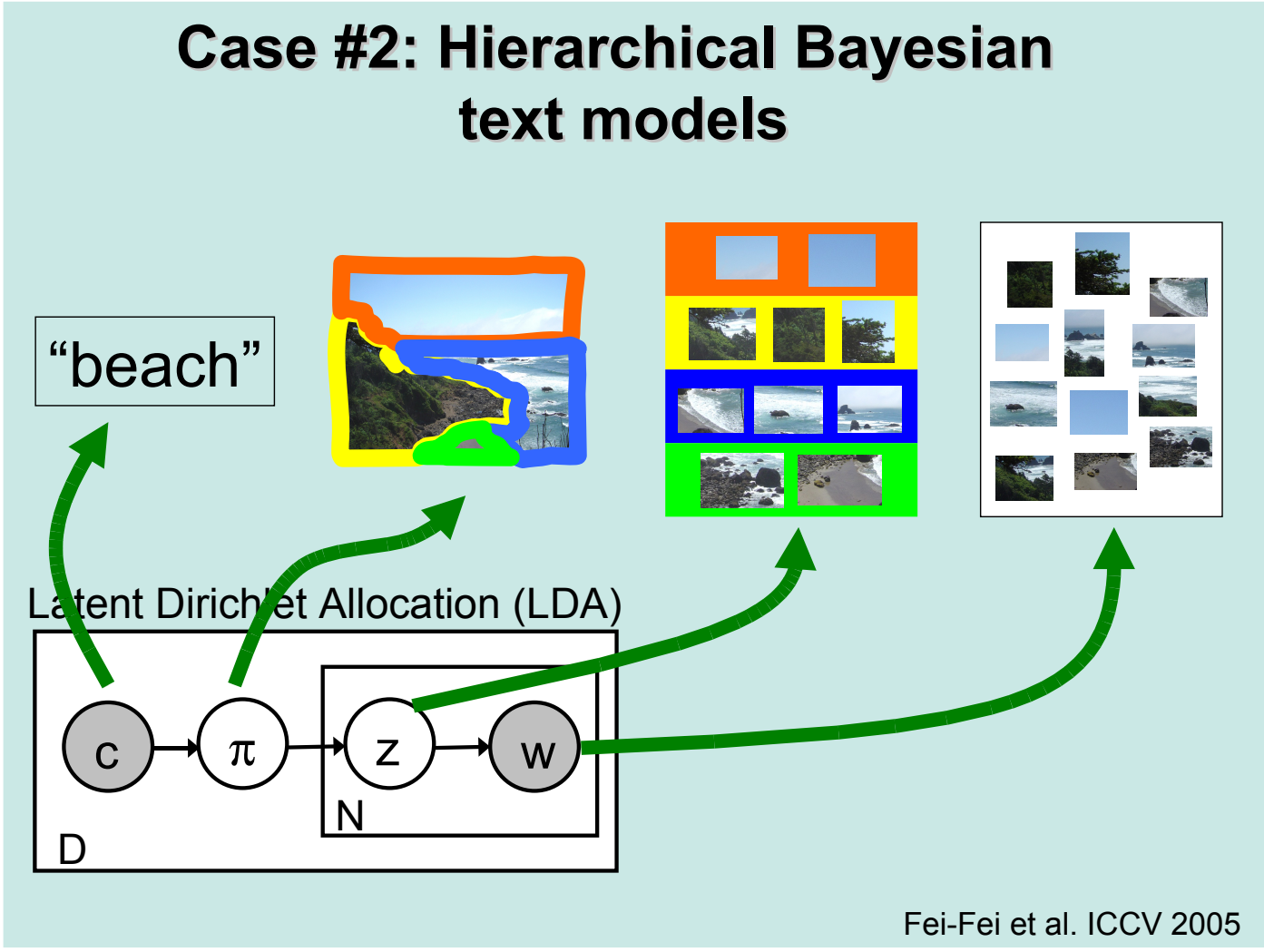
Slides from [CVPR 2007 Short Course on Object Recognition]

Application: Object Recognition in Images



Slides from [CVPR 2007 Short Course on Object Recognition]

Application: Object Recognition in Images



Slides from [CVPR 2007 Short Course on Object Recognition]

Application: Object Recognition in Images

Learning

Use variational bayes or MCMC to learn:

- β - a matrix which encodes the probability of observing a codeword w conditioned on a topic z.
- θ - a matrix which encodes the Dirichlet parameters for each image class.

Classification

For an unknown image x, want to determine the image class c that has the highest likelihood of generating x:

$$\text{Image class } c = \mathit{argmax}_c p(x|c, \theta, \beta)$$

- Must integrate over hidden variables π, z
- Intractable $\rightarrow$ must resort to approximate methods (again)

Application: Object Recognition in Images

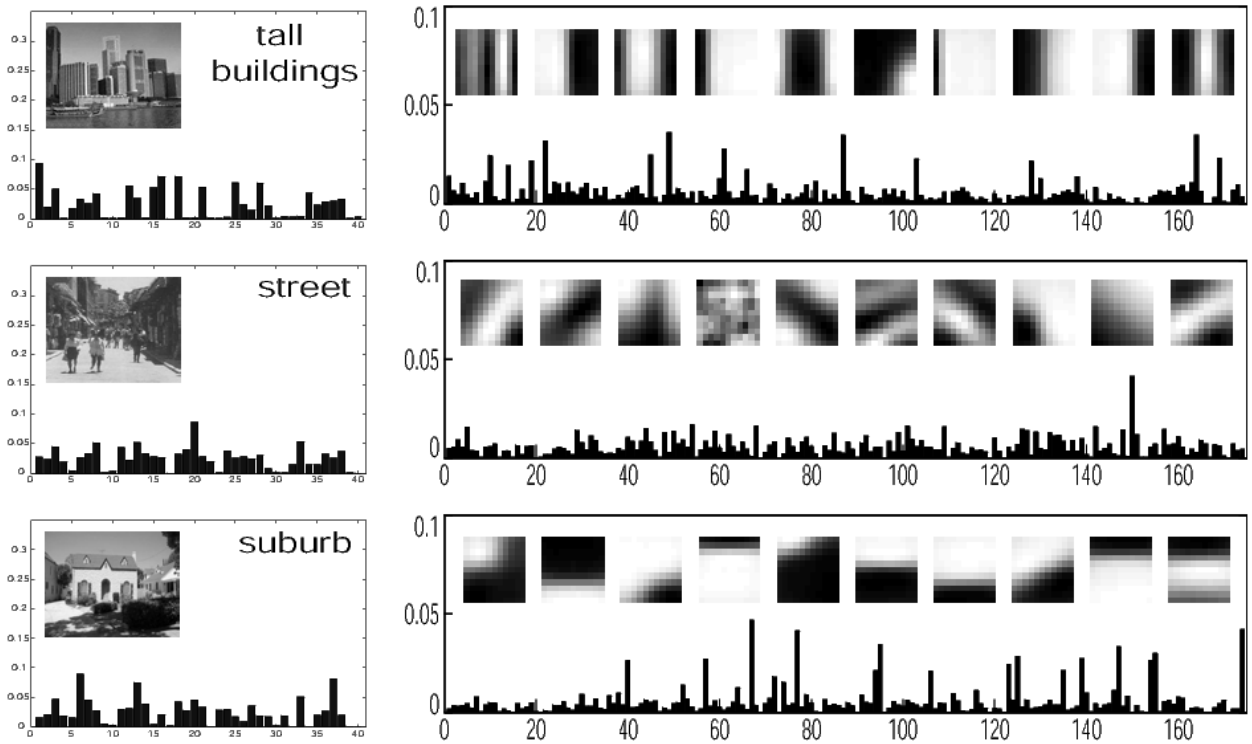


Figure: Models of 3 image categories. From [Fei-Fei and Perona, 2005]

Application: Object Recognition in Images

Figure: Examples of testing images for each category. From [Fei-Fei and Perona, 2005]

Questions?

References


D. M. Blei, A. Y. Ng and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 2003.


D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. NIPS 2003.


D. Blei and J. Laferty. Correlated Topic Models. NIPS 2006.


D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. *NIPS*, 2001.


T. Hofmann. Probabilistic Latent Semantic Indexing. *SIGIR*, 1999.


M. Steyvers and T. L. Griffiths. Probabilistic Topic Models. In *Latent Semantic Analysis: A Road to Meaning*.


T. L. Griffiths and M. Steyvers. Finding Scientific Topics. *PNAS*, 2004.


TL Griffiths, M Steyvers, D Blei, JB Tenenbaum. Integrating Topics and Syntax. *NIPS* 2004.


A. Mccallum and K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

References (cont)


L. Fei-Fei, P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. *CVPR 2005*.


R. Madsen, D. Kauchak and C. Elkan. Modeling Word Burstiness using the Dirichlet Distribution. *ICML*, 2005.


Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. W. Li and A. McCallum. *ICML 2006.*


Nonparametric Bayesian Pachinko Allocation. W. Li, D. Blei and A. McCallum. UAI 2007.


M Rosen-Zvi, T Griffiths, M Steyvers and P Smyth. The Author-Topic Model for Authors and Documents. UAI 2004.


H. Wallach. Topic modeling: beyond bag-of-words. ICML 2006.


L. Fei Fei. Bag of words models. CVPR 2007 Short Course. Presentation Slides. `http://vision.cs.princeton.edu/documents/CVPR2007_tutorial_bag_of_words.ppt`