Advanced Machine Learning

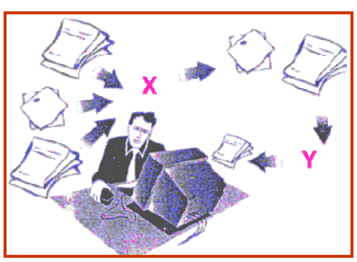
Applications: Topic Models

Eric Xing

Lecture 18, August 14, 2008

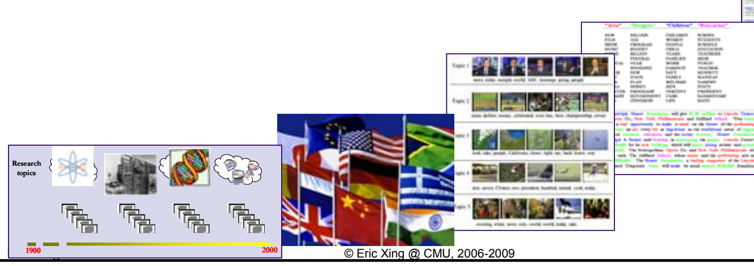
Reading: see class homepage

The Problem: NLP and Data Mining



We want:

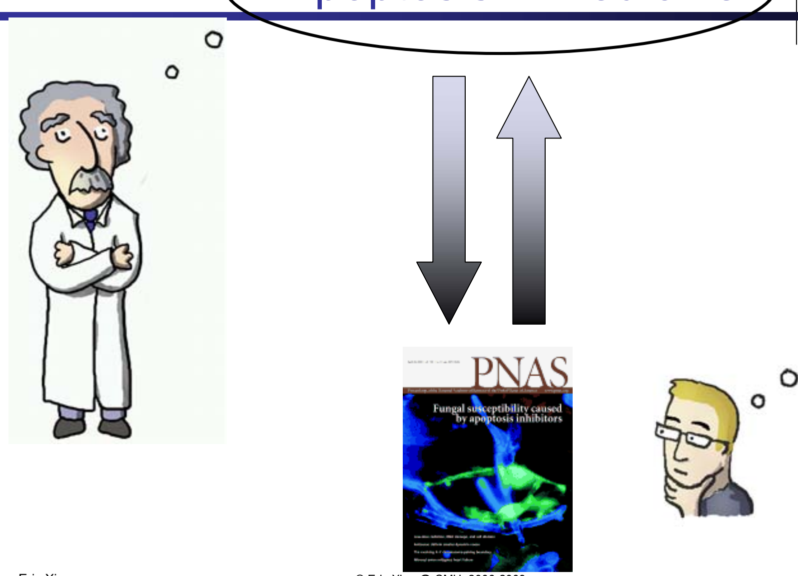
- Semantic-based search
- infer topics and categorize documents
- Multimedia inference
- Automatic translation
- Predict how topics evolve
- …



Modeling document collections

- A document collection is a dataset where each data point is itself a collection of simpler data.
  - Text documents are collections of words.
  - Segmented images are collections of regions.
  - User histories are collections of purchased items.
- Many modern problems ask questions of such data.
  - Is this text document relevant to my query?
  - Which documents are about a particular topic?
  - How have topics changed over time?
  - What does author X write about? Who is likely to write about topic Y? Who wrote this specific document?
  - Which category is this image in? Create a caption for this image.
  - What movies would I probably like?
  - and so on…..

Apoptosis + Medicine



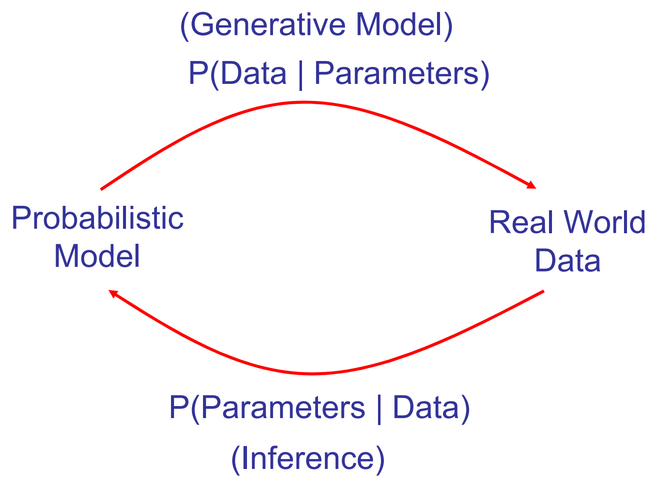
Apoptosis + Medicine

probabilistic
generative
model

Apoptosis + Medicine

statistical
inference

Connecting Probability Models to Data



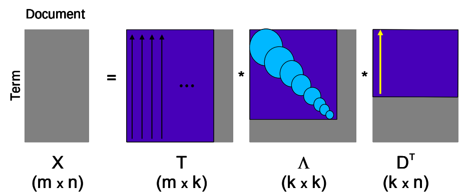
Motivation for modeling latent topical aspects

- Dimensionality reduction
  - A VSM lives in a very high-dimensional feature space (usually larger vocabulary, V)
  - Sparse representation of documents (|V| >> actual number of appeared words in any given document) --- often too spurious for many IR tasks
- Semantic analysis and comprehension
  - A need to define conceptual closeness,
  - to capture relation between features,
  - to distinguish and infer features from heterogeneous sources ...

Latent Semantic Indexing

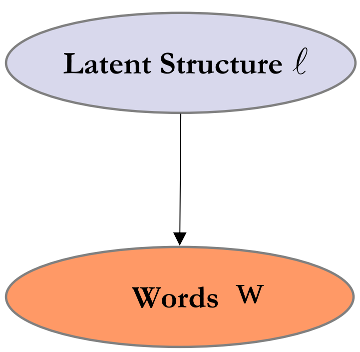
(Deerwester et al., 1990)

- Classic attempt at solving this problem in information retrieval



- Uses SVD to reduce document representations
- Models synonymy and polysemy
- Computing SVD is slow
- Non-probabilistic model

Latent Semantic Structure



Distribution over words

$$P(\mathrm{w}) = \sum_{\ell} P(\mathrm{w}, \ell)$$

Inferring latent structure

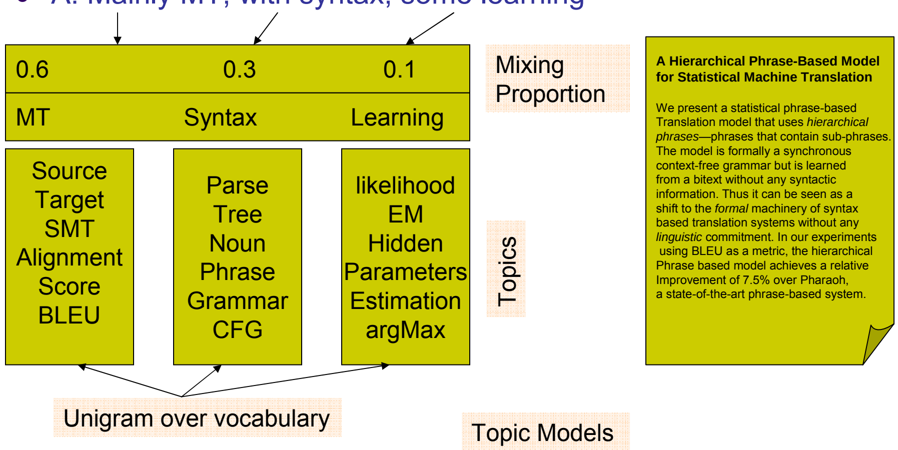
$$P(\ell \mid \mathrm{w}) = \frac{P(\mathrm{w} \mid \ell) P(\ell)}{P(\mathrm{w})}$$

Prediction

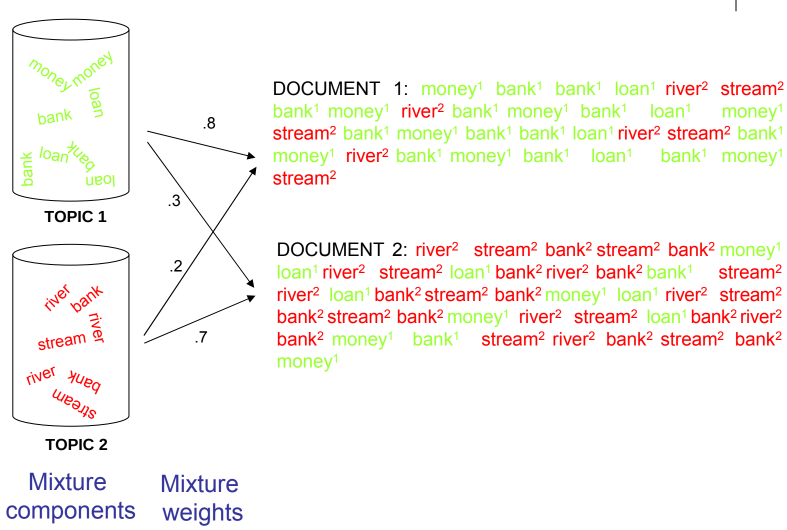
$$P(w_{n+1} \mid \mathrm{w}) = \ldots$$

How to Model Semantic?

- Q: What is it about?
- A: Mainly MT, with syntax, some learning



GENERATIVE PROCESS



Why this is Useful?

- Q: What is it about?
- A: Mainly MT, with syntax, some learning

0.6	0.3	0.1
MT	Syntax	Learning

Mixing Proportion

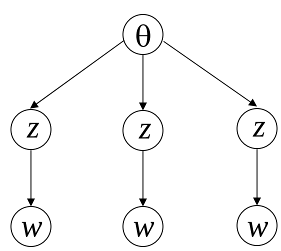
A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses *hierarchical phrases*—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the *formal* machinery of syntax based translation systems without any *linguistic* commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

- Q: give me similar document?
  - Structured way of browsing the collection
- Other tasks
  - Dimensionality reduction
    - TF-IDF vs. topic mixing proportion
    - Classification, clustering, and more …

A generative model for documents

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

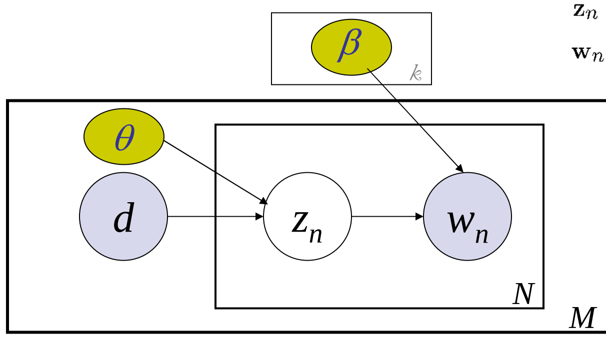


Probabilistic LSI

Hoffman (1999)

$$\mathbf{z}_n \sim \text{Mult}(\boldsymbol{\theta})$$

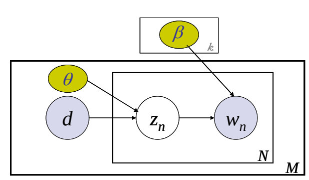
$$\mathbf{w}_n \sim p(\mathbf{w}_n | \mathbf{z}_n, \boldsymbol{\beta})$$



$$p(d, w_n) = p(d) \sum_{\mathbf{z}} \left(\prod_{n=1}^{N} p(w_n | z_n) p(z_n | d) \right)$$

Probabilistic LSI

- A "generative" model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a list of mixing proportions for the components (i.e. topic vector θ).



Latent Dirichlet Allocation

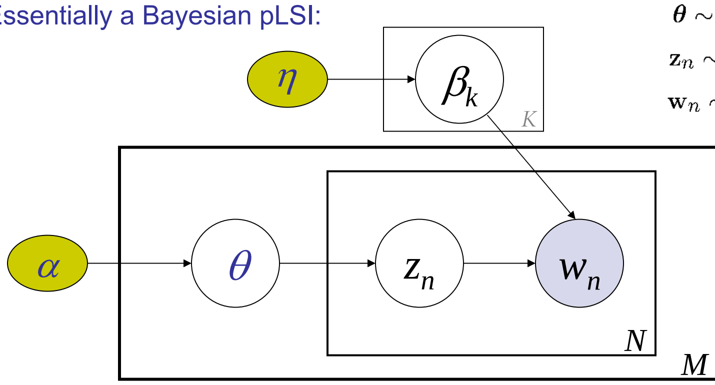
Blei, Ng and Jordan (2003)

Essentially a Bayesian pLSI:

$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$

$\mathbf{z}_n \sim \mathrm{Mult}(\boldsymbol{\theta})$

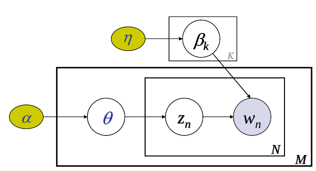
$\mathbf{w}_n \sim p(\mathbf{w}_n|\mathbf{z}_n, \boldsymbol{\beta})$



$$p(\mathbf{w}) = \sum_{\mathbf{z}} \int p(\theta) p(\beta) \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|\beta_{z_n}) \right) d\theta \, d\beta$$

LDA

- Generative model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a list of mixing proportions for the components (i.e. topic vector).
- The topic vectors and the word rates each follows a Dirichlet prior --- essentially a Bayesian pLSI



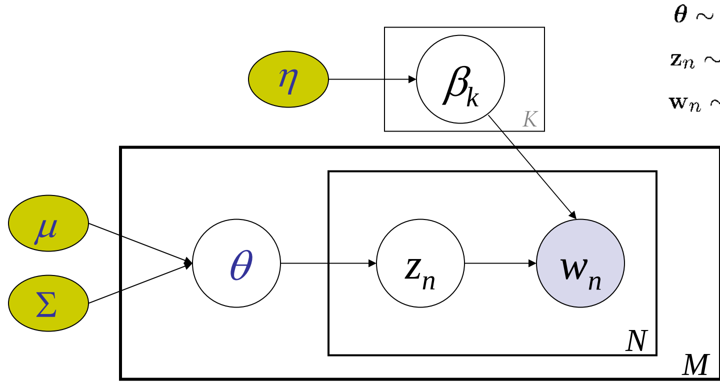
Correlated Topic Model

Blei & Lafferty (2005)

$\theta \sim \mathrm{Dir}(\alpha)$

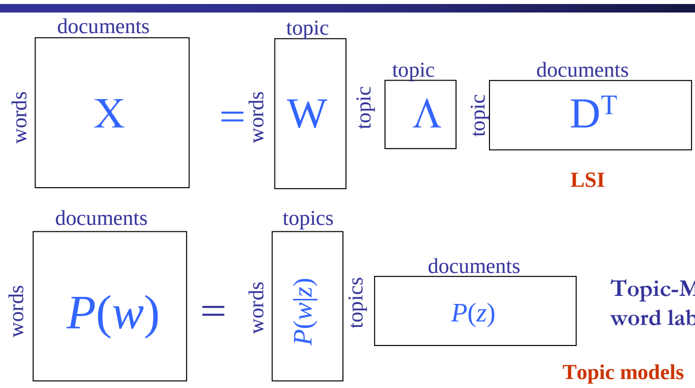
$\mathbf{z}_n \sim \mathrm{Mult}(\theta)$

$\mathbf{w}_n \sim p(\mathbf{w}_n | \mathbf{z}_n, \beta)$



$$p(\mathbf{w}) = \sum_{\mathbf{z}} \int p(\theta) p(\beta) \left(\prod_{n=1}^{N} p(z_n | \theta) p(w_n | \beta_{z_n}) \right) d\theta \, d\beta$$

Comparison of model semantics



$\vec{x} = W' \vec{d}$

Topic-Mixing is via repeated word labeling

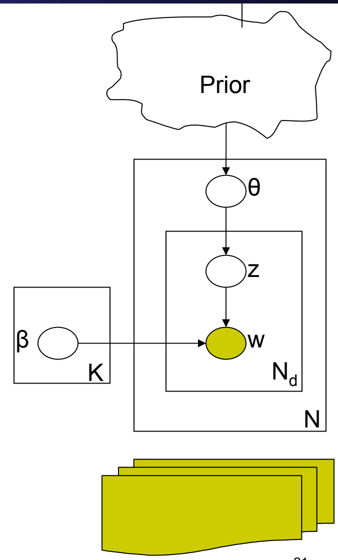
Topic Models = Mixture Membership Models

Generating a document

– Draw θ from the prior

For each word n

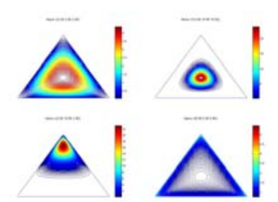
- Draw z_n from $\mathit{multinomial}(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $\mathit{multinomial}(\beta_{z_n})$



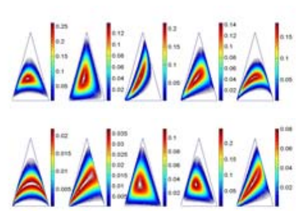
Which prior to use?

Prior Comparison

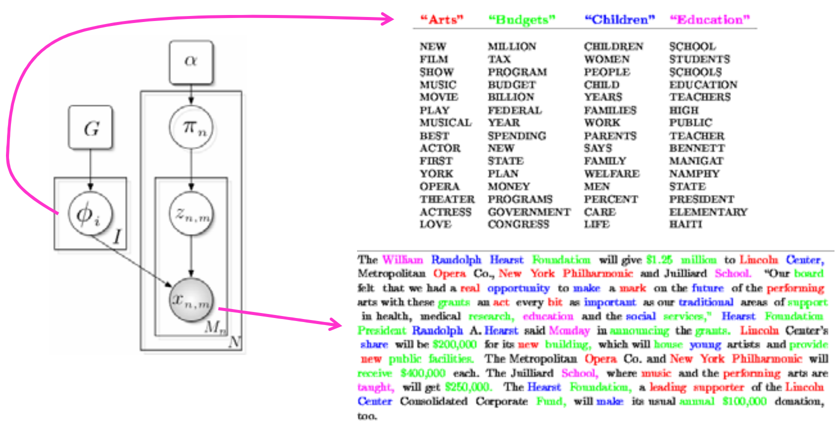
- Dirichlet (LDA) (Blei et al. 2003)
  - Conjugate prior means efficient inference
  - Can only capture variations in each topic's intensity independently



- Logistic Normal (CTM=LoNTAM) (Blei & Lafferty 2005, Ahmed & Xing 2006)
  - Capture the intuition that some topics are highly correlated and can rise up in intensity together
  - Not a conjugate prior implies hard inference



Inference Tasks



Bayesian inference

- A possible query:

$$p(\pi_n \mid D) = ?$$
$$p(z_{n,m} \mid D) = ?$$

- Close form solution? $p(\pi_n \mid D) = \dfrac{p(\pi_n, D)}{p(D)}$

$$= \frac{\sum_{\{z_{n,m}\}} \int \left(\prod_n \left(\prod_m p(x_{n,m} \mid \phi_{z_n}) p(z_{n,m} \mid \pi_n) \right) p(\pi_n \mid \alpha) \right) p(\phi \mid G) d\pi_{-i} d\phi}{p(D)}$$

$$p(D) = \sum_{\{z_{n,m}\}} \int \cdots \int \left(\prod_n \left(\prod_m p(x_{n,m} \mid \phi_{z_n}) p(z_{n,m} \mid \pi_n) \right) p(\pi_n \mid \alpha) \right) p(\phi \mid G) d\pi_1 \cdots d\pi_N d\phi$$

- Sum in the denominator over T^n terms, and integrate over n k-dimensional topic vectors

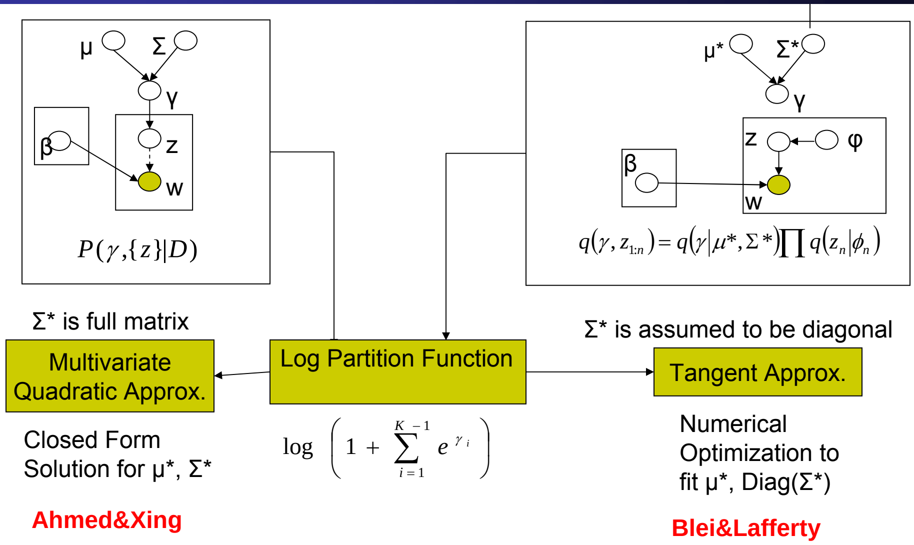
Approximate Inference

- Variational Inference
  - Mean field approximation (Blei et al)
  - Expectation propagation (Minka et al)
  - Variational 2nd-order Taylor approximation (Xing)

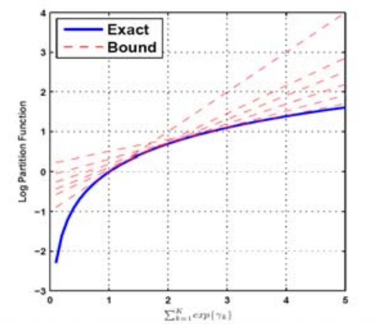
- Markov Chain Monte Carlo
  - Gibbs sampling (Griffiths et al)

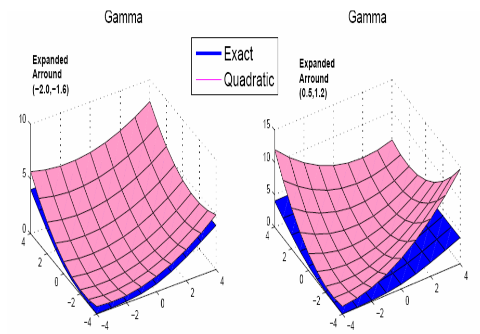
Variational Inference

(e.g., MF, Jordan et al 1999, GMF, Xing et al 2004)

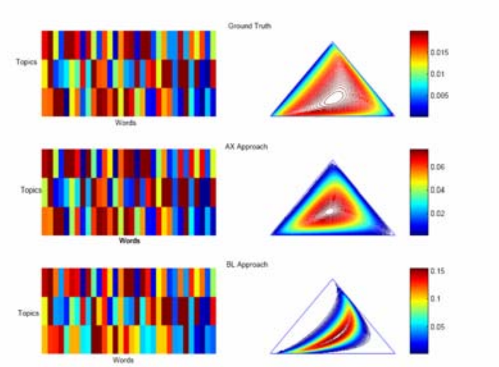


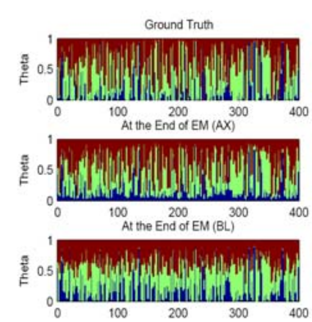
Tangent Approximation





Test on Synthetic Text

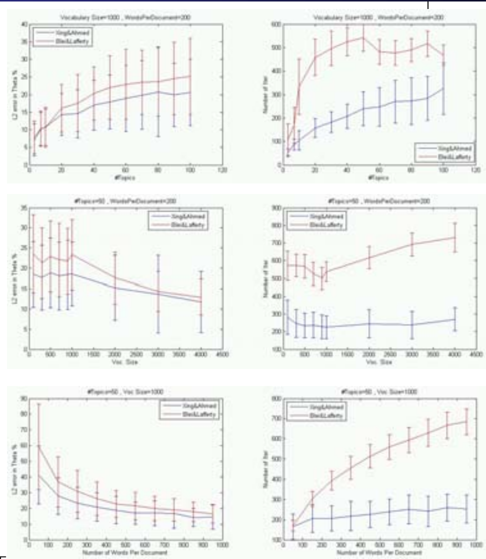




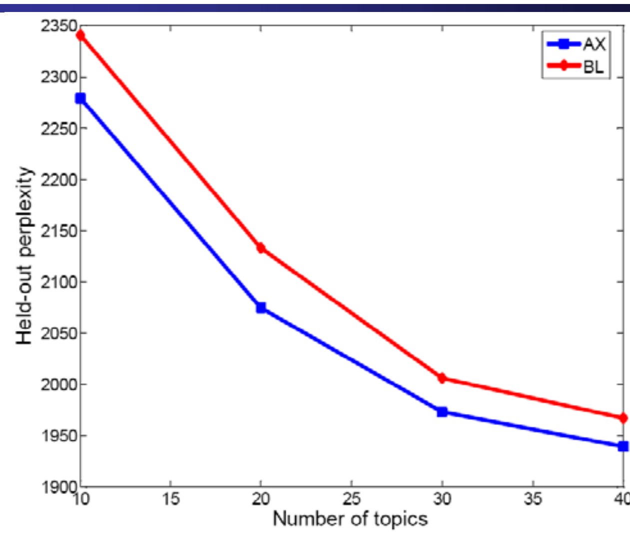
Comparison: accuracy and speed

L2 error in topic vector est. and # of iterations

- Varying Num. of Topics
- Varying Voc. Size
- Varying Num. Words Per Document



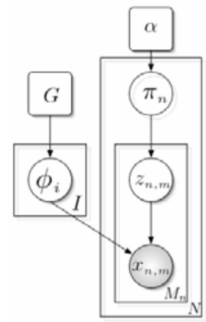
Comparison: perplexity



Collapsed Gibbs sampling

(Tom Griffiths & Mark Steyvers)

- Collapsed Gibbs sampling
  - Integrate out π



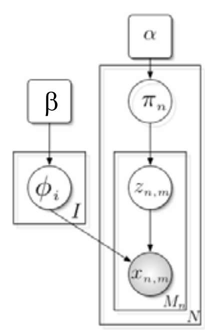
For variables $\mathbf{z} = z_1, z_2, \ldots, z_n$

Draw $z_i^{(t+1)}$ from $P(z_i|\mathbf{z}_{-i}, \boldsymbol{w})$

$\mathbf{z}_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_n^{(t)}$

Gibbs sampling

- Need full conditional distributions for variables
- Since we only sample z we need



$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})P(z_i = j|\mathbf{z}_{-i})$$

$$= \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + W\beta} \frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,\cdot} + T\alpha}$$

$n_j^{(w)}$ number of times word w assigned to topic j

$n_j^{(d)}$ number of times topic j used in document d

Gibbs sampling

			iteration 1
i	w_i	d_i	z_i
1	MATHEMATICS	1	2
2	KNOWLEDGE	1	2
3	RESEARCH	1	1
4	WORK	1	2
5	MATHEMATICS	1	1
6	RESEARCH	1	2
7	WORK	1	2
8	SCIENTIFIC	1	1
9	MATHEMATICS	1	2
10	WORK	1	1
11	SCIENTIFIC	2	1
12	KNOWLEDGE	2	1
.	.	.	.
.	.	.	.
.	.	.	.
50	JOY	5	2

Gibbs sampling

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

Gibbs sampling

i	w_i	d_i	iteration 1 z_i	iteration 2 z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

i	w_i	d_i	iteration 1 z_i	iteration 2 z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	?
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	?
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	?
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	2
5	MATHEMATICS	1	1	?
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

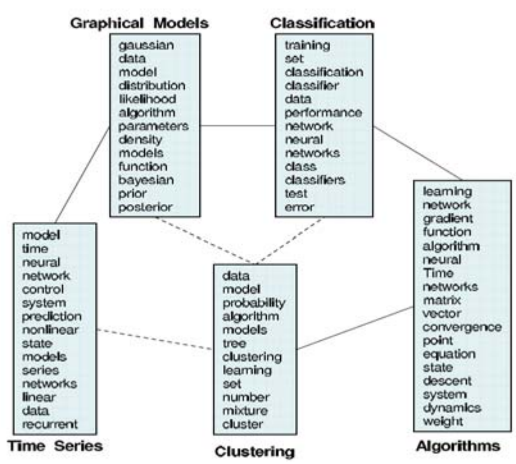
$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration			
			1	2	…	1000
i	w_i	d_i	z_i	z_i		z_i
1	MATHEMATICS	1	2	2		2
2	KNOWLEDGE	1	2	1		2
3	RESEARCH	1	1	1		2
4	WORK	1	2	2		1
5	MATHEMATICS	1	1	2		2
6	RESEARCH	1	2	2		2
7	WORK	1	2	2		2
8	SCIENTIFIC	1	1	1	…	1
9	MATHEMATICS	1	2	2		2
10	WORK	1	1	2		2
11	SCIENTIFIC	2	1	1		2
12	KNOWLEDGE	2	1	2		2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
50	JOY	5	2	1		1

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{-i,j}+\beta}{n^{(\cdot)}_{-i,j}+W\beta}\,\frac{n^{(d_i)}_{-i,j}+\alpha}{n^{(d_i)}_{-i,\cdot}+T\alpha}$$

Topics and topic graphs



Result on PNAS collection

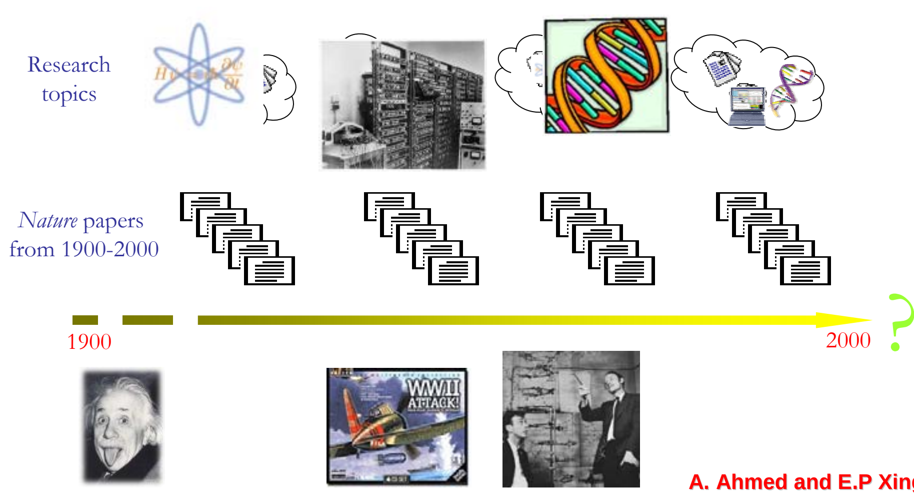
- PNAS abstracts from 1997-2002
  - 2500 documents
  - Average of 170 words per document
- Fitted 40-topics model using both approaches
- Use low dimensional representation to predict the abstract category
  - Use SVM classifier
  - 85% for training and 15% for testing

Classification Accuracy

Category	Doc	BL	AX
Genetics	21	61.9	61.9
Biochemistry	86	65.1	77.9
Immunology	24	70.8	66.6
Biophysics	15	53.3	66.6
Total	146	64.3	72.6

-Notable Difference
-Examine the low dimensional representations below

Extension 1: topic evolution?



A. Ahmed and E.P Xing, Submitted 2007

How to Model Topic Evolution

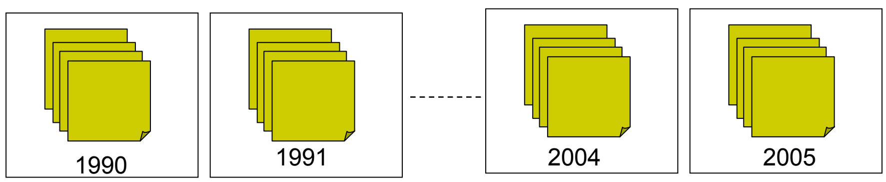
Topic Trends

Topic Keywords

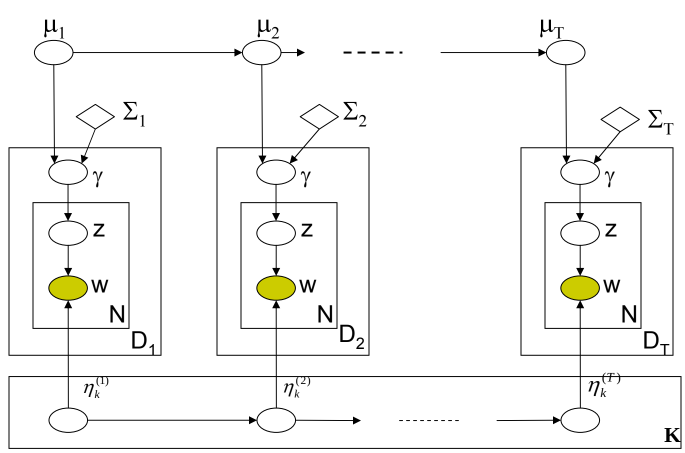
Topic correlations

~~Number of topics~~

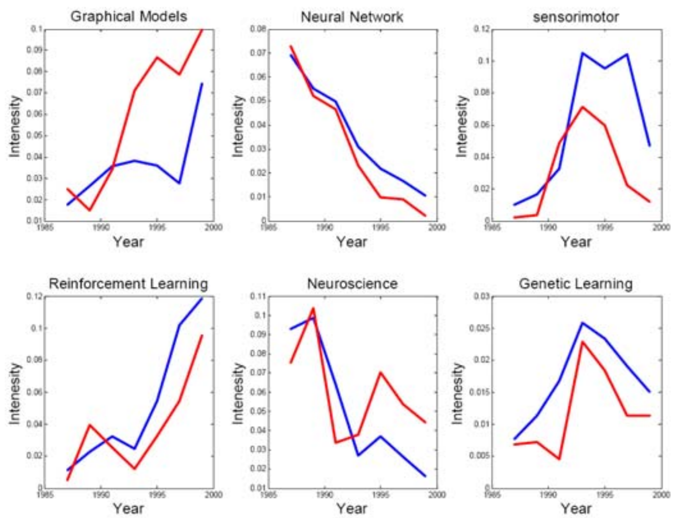
The Dynamic Correlated Topic model



The Dynamic CTM



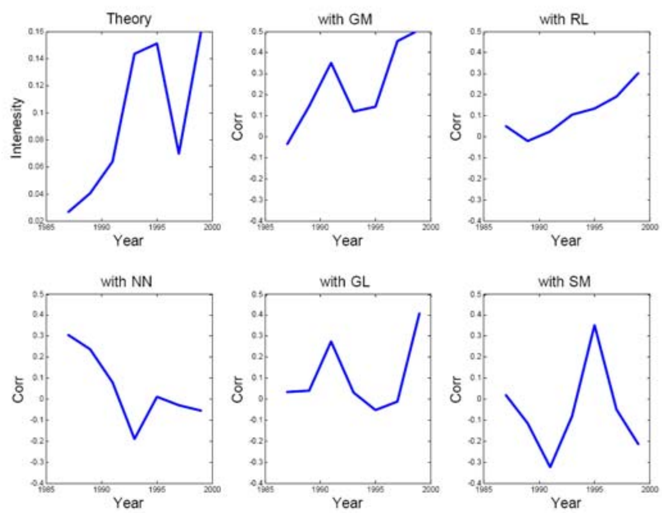
Topic Trends



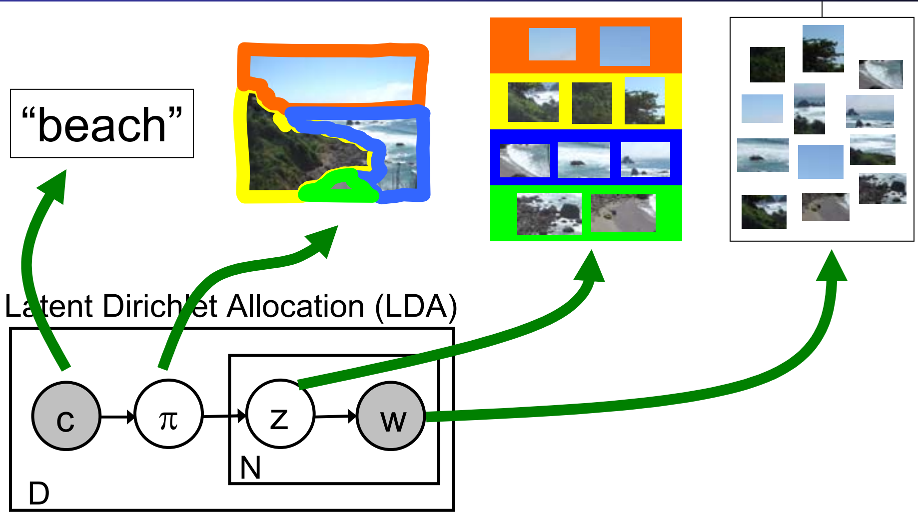
Topic Words over Time

Leg Insect Controller Sensory sweeping	learning Sweeping state prioritized time control	learning State neural control reinforcement policy	Policy learning State reinforcement Optimal action Reward markov nonlinear	RL
brain information neurons fibers potential model synaptic	model activity network neurons cell phase firing	Model time neurons Spike input Cells head information Rate synaptic temporal	Synergy synaptic Pruning neurons adaptation Spike neuron degradation Irregular hebbian	Neuroscience

Topic Correlations Over Time

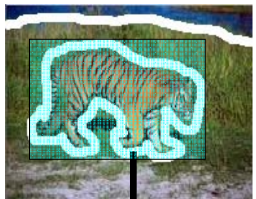


Extension 2:
Topic Models for Images



Fei-Fei et al. ICCV 2005

Image Representation



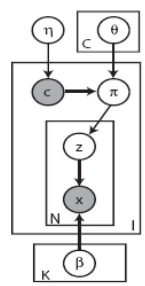
cat, grass, tiger, water

$[r_{11} \cdots r_{1d}]$, $[w_1 \cdots w_{|V|}]$

representation vector (real, 1 per image segment) : **annotation vector** (binary, same for each segment)

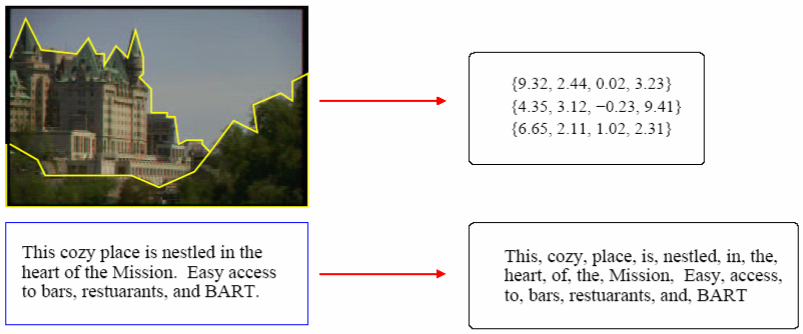
$$[r_{n1} \cdots r_{nd}] \text{ , } [w_1 \cdots w_{|V|}]$$

To Generate an Image …



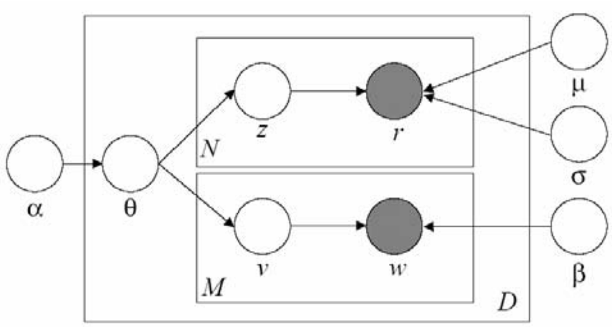
$$\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, c|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\beta}) &= p(c|\boldsymbol{\eta})p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) \cdot \\
&\prod_{n=1}^{N} p(z_n|\boldsymbol{\pi})p(x_n|z_n, \boldsymbol{\beta}) \\
p(c|\boldsymbol{\eta}) &= \text{Mult}(c|\boldsymbol{\eta}) \\
p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) &= \prod_{j=1}^{C} Dir(\boldsymbol{\pi}|\boldsymbol{\theta}_{j\cdot})^{\delta(c,j)} \\
p(z_n|\boldsymbol{\pi}) &= \text{Mult}(z_n|\boldsymbol{\pi}) \\
p(x_n|z_n, \boldsymbol{\beta}) &= \prod_{k=1}^{K} p(x_n|\boldsymbol{\beta}_{k\cdot})^{\delta(z_n^k,1)}
\end{aligned}$$

Annotated images



- Forsyth et. al. (2001): images as documents where region-specific feature vectors are like visual words.
- A captioned image can be thought of as annotated data: two documents, one of which describes the other.

Gaussian-multinomial LDA



- A natural next step is to glue two LDA models together.
- Bottom: a traditional LDA model on captions
- Top: a Gaussian-LDA model on images
  - each region is a multivariate Gaussian
- Does not work well

Automatic annotation

True caption
birds tree

Corr–LDA
birds nest leaves branch tree

GM–LDA
water birds nest tree sky

GM–Mixture
tree ocean fungus mushrooms coral

True caption
fish reefs water

Corr–LDA
fish water ocean tree coral

GM–LDA
water sky vegetables tree people

GM–Mixture
fungus mushrooms tree flowers leaves

Conclusion

- GM-based topic models are cool
  - Flexible
  - Modular
  - Interactive
- Efficient Inference/learning algorithms
  - GMF, with Laplace approx. for non-conjugate dist.
  - MCMC
- Many applications
  - …
  - Word-sense disambiguation
  - Word-net
  - Network inference