Machine Learning

10-701/15-781, Fall 2012

Latent Space Analysis

SVD and Topic Models

Eric Xing

Lecture 19, November 19, 2012

Reading: Tutorial on Topic Model @ ACL12

We are inundated with data …

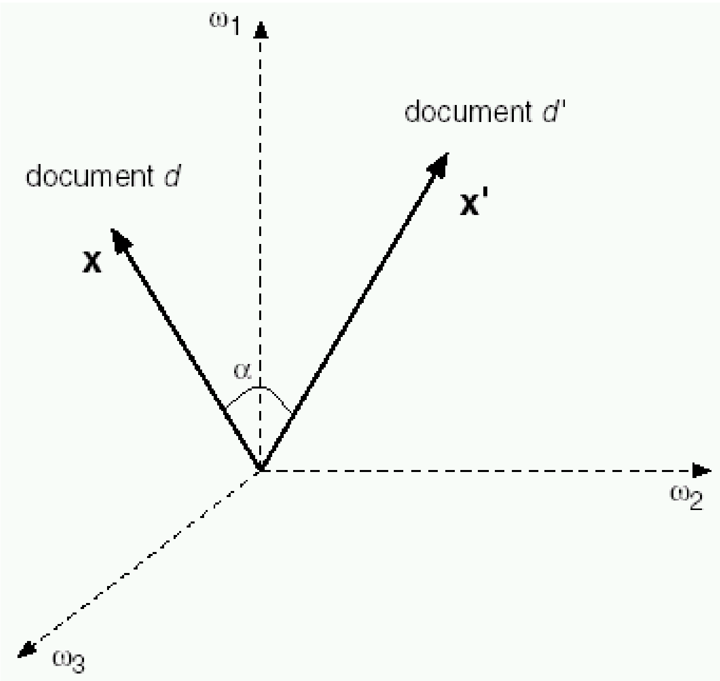
(from images.google.cn)

- Humans cannot afford to deal with (e.g., search, browse, or measure similarity) a huge number of text and media documents
- We need computers to help out …

A task:

- Say, we want to have a mapping …, so that

$\Rightarrow$



- Compare similarity
- Classify contents
- Cluster/group/categorize docs
- Distill semantics and perspectives
- ..

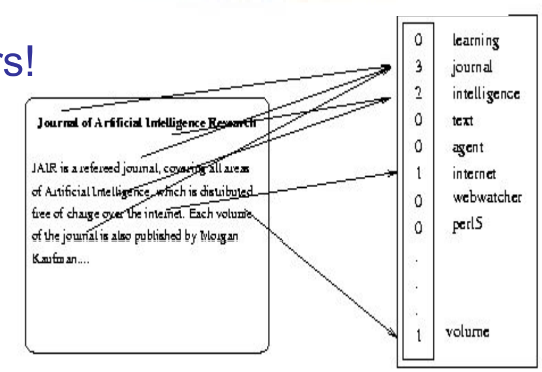
Representation:

- Data: **Bag of Words Representation**

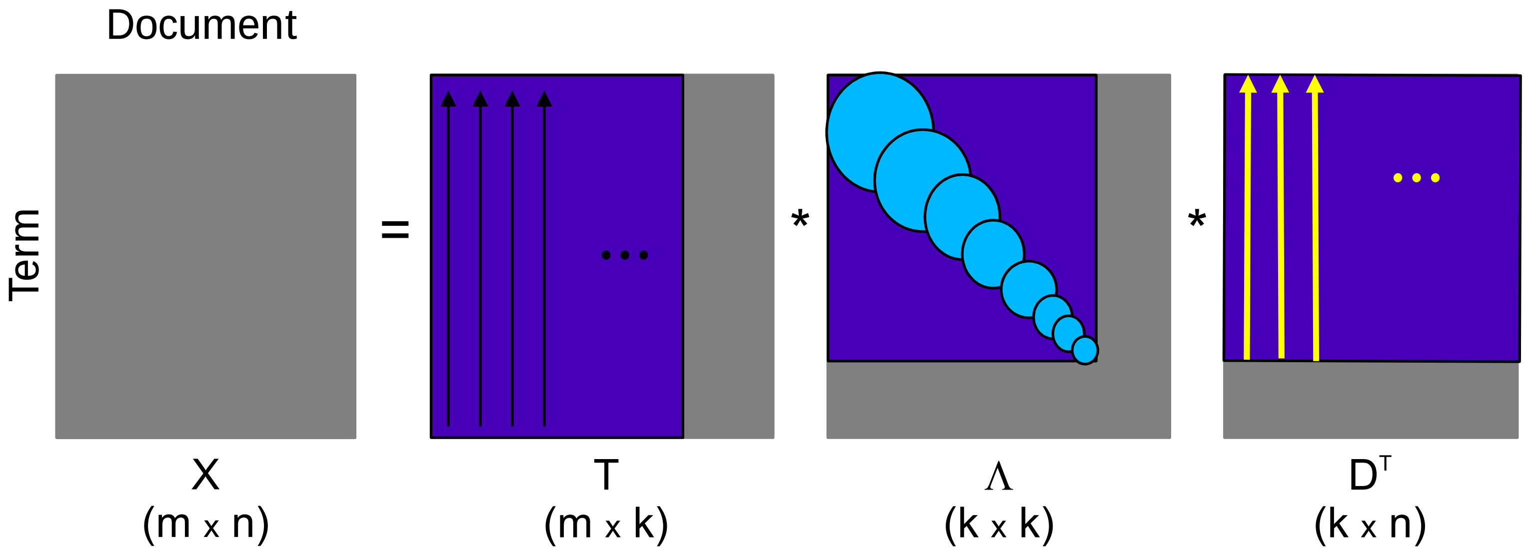
As for the Arabian and Palestinean voices that are against the current negotiations and the so-called peace process, they are not against peace per se, but rather for their well-founded predictions that Israel would NOT give an inch of the West bank (and most probably the same for Golan Heights) back to the Arabs. An 18 months of "negotiations" in Madrid, and Washington proved these predictions. Now many will jump on me saying why are you blaming israelis for no-result negotiations. I would say why would the Arabs stall the negotiations, what do they have to loose ?

- Each document is a vector in the word space
- Ignore the order of words in a document. Only count matters!

- A high-dimensional and sparse representation
  - Not efficient text processing tasks, e.g., search, document classification, or similarity measure
  - Not effective for browsing



Subspace analysis



cluster/topic/basis Distributions **(subspace)**

A priori **weights**

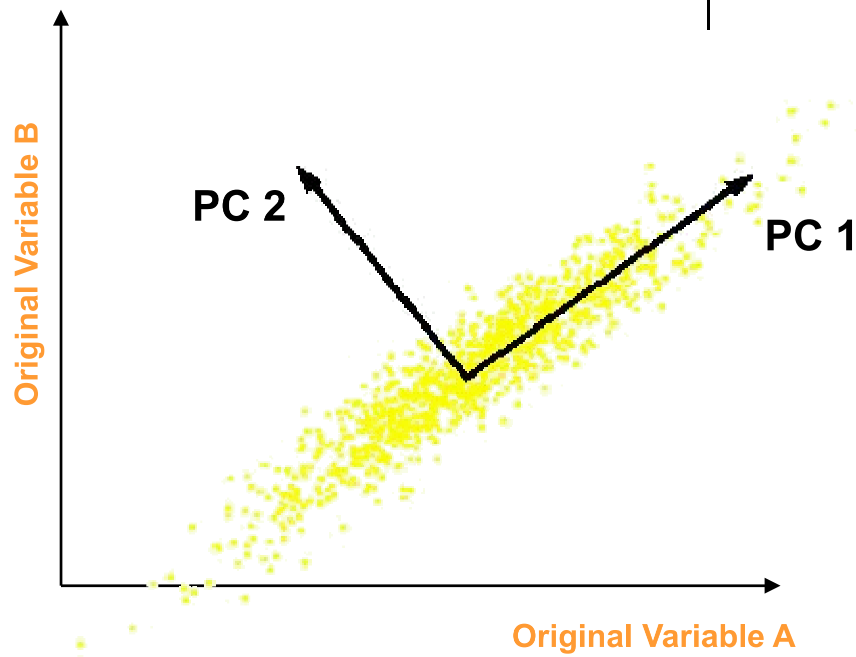
Memberships **(coordinates)**

- Clustering: (0,1) matrix
- LSI/NMF: “arbitrary” matrices
- **Topic Models: stochastic matrix**
- Sparse coding: “arbitrary” **sparse** matrices

An example:

Principal Component Analysis

- The new variables/dimensions
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
    - Orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components



- Orthogonal directions of greatest variance in data

- Projections along PC1 discriminate the data most along any one axis

- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on …

Computing the Components

- Projection of vector $\mathbf{x}$ onto an axis (dimension) $\mathbf{u}$ is $\mathbf{u}^T\mathbf{x}$
- Direction of greatest variability is that in which the average square of the projection is greatest:

$$\text{Maximize} \quad \mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u}$$

$$\text{s.t} \quad \mathbf{u}^T\mathbf{u} = 1$$

Construct Langrangian $\mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u} - \lambda\mathbf{u}^T\mathbf{u}$

Vector of partial derivatives set to zero

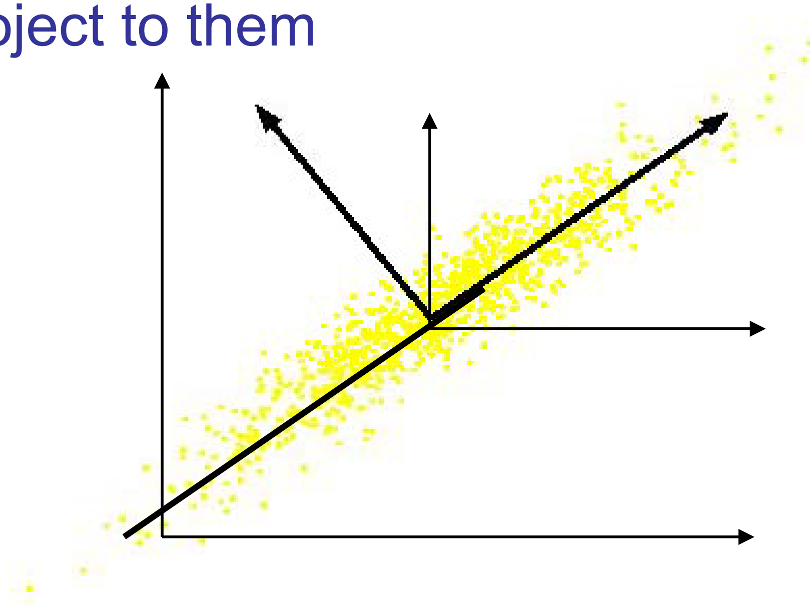
$$\mathbf{x}\mathbf{x}^T\mathbf{u} - \lambda\mathbf{u} = (\mathbf{x}\mathbf{x}^T - \lambda\mathbf{I})\,\mathbf{u} = 0$$

As $\mathbf{u} \neq \mathbf{0}$ then $\mathbf{u}$ must be an eigenvector of $\mathbf{X}\mathbf{X}^T$ with eigenvalue λ

- λ is the principal eigenvalue of the **correlation matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^T$**
- The eigenvalue denotes the amount of variability captured along that dimension

Computing the Components

- Similarly for the next axis, etc.
- So, the new axes are the eigenvectors of the matrix of correlations of the original variables, which captures the similarities of the original variables based on how data samples project to them



- Geometrically: centering followed by rotation
  - Linear transformation

Eigenvalues & Eigenvectors

- For symmetric matrices, eigenvectors for distinct eigenvalues are **orthogonal**

$$Sv_{\{1,2\}} = \lambda_{\{1,2\}} v_{\{1,2\}}, \text{ and } \lambda_1 \neq \lambda_2 \Rightarrow v_1 \bullet v_2 = 0$$

- All eigenvalues of a real symmetric matrix are **real**.

$$\text{if } |S - \lambda I| = 0 \text{ and } \mathrm{S} = \mathrm{S}^{\mathrm{T}} \Rightarrow \lambda \in \Re$$

- All eigenvalues of a positive semidefinite matrix are **non-negative**

$$\forall w \in \Re^n, w^T S w \geq 0, \text{ then if } \mathrm{Sv} = \lambda \mathrm{v} \Rightarrow \lambda \geq 0$$

Eigen/diagonal Decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with ***m*** **linearly independent eigenvectors** (a "non-defective" matrix)

- **Theorem**: Exists an **eigen decomposition**

$$\mathbf{S} = \mathbf{U}\underbrace{\mathbf{\Lambda}}_{\text{diagonal}}\mathbf{U}^{-1}$$

Unique for distinct eigen-values

(cf. matrix diagonalization theorem)

- Columns of ***U*** are **eigenvectors** of ***S***
- Diagonal elements of $\mathbf{\Lambda}$ are **eigenvalues** of $\mathbf{S}$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

PCs, Variance and Least-Squares

- The first PC retains the greatest amount of variation in the sample
- The k^{th} PC retains the kth greatest fraction of the variation in the sample
- The k^{th} largest eigenvalue of the correlation matrix C is the variance in the sample along the k^{th} PC
- The least-squares view: PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones

The Corpora Matrix

$X =$

	Doc 1	Doc 2	Doc 3	...
Word 1	3	0	0	...
Word 2	0	8	1	...
Word 3	0	1	3	...
Word 4	2	0	0	...
Word 5	12	0	0	...
...	0	0	0	...

Singular Value Decomposition

For an $m \times n$ matrix A of rank r there exists a factorization (Singular Value Decomposition = SVD) as follows:

$$A = U\Sigma V^T$$

$m \times m$ | $m \times n$ | V is $n \times n$

The columns of U are orthogonal eigenvectors of AA^T.

The columns of V are orthogonal eigenvectors of A^TA.

Eigenvalues $\lambda_1 \ldots \lambda_r$ of AA^T are the eigenvalues of A^TA.

$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag(\sigma_1 \ldots \sigma_r)$$

Singular values.

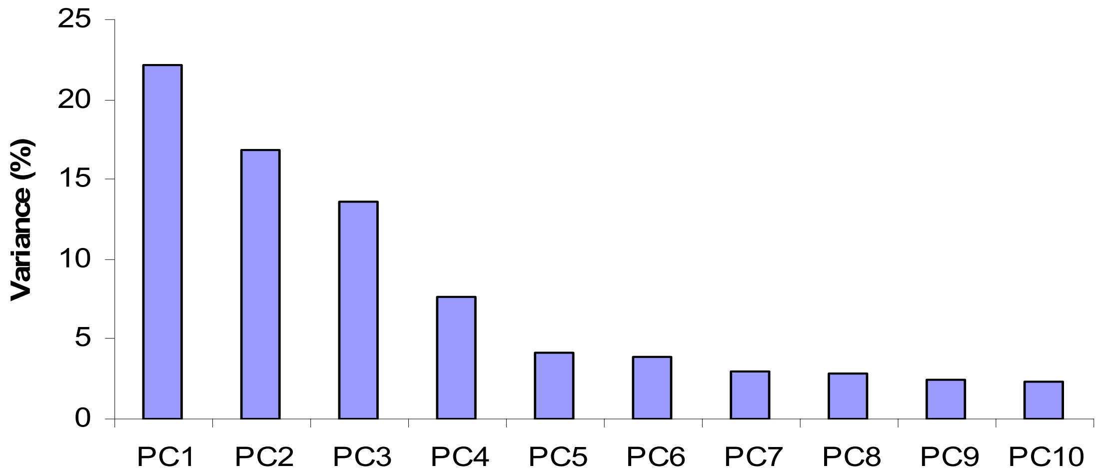
SVD and PCA

- The first root is called the prinicipal eigenvalue which has an associated orthonormal ($\mathbf{u}^T\mathbf{u} = 1$) *eigenvector* $\mathbf{u}$
- Subsequent roots are ordered such that $\lambda_1 > \lambda_2 > \ldots > \lambda_M$ with rank($\mathbf{D}$) non-zero values.
- Eigenvectors form an orthonormal basis i.e. $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$
- The eigenvalue decomposition of $\mathbf{XX}^T = \mathbf{U\Sigma U}^T$
- where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M]$ and $\mathbf{\Sigma} = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_M]$
- Similarly the eigenvalue decomposition of $\mathbf{X}^T\mathbf{X} = \mathbf{V\Sigma V}^T$
- The SVD is closely related to the above $\mathbf{X}=\mathbf{U}\,\mathbf{\Sigma}^{1/2}\,\mathbf{V}^T$
- The left eigenvectors **U,** right eigenvectors **V,**
- singular values = square root of eigenvalues**.**

How Many PCs?

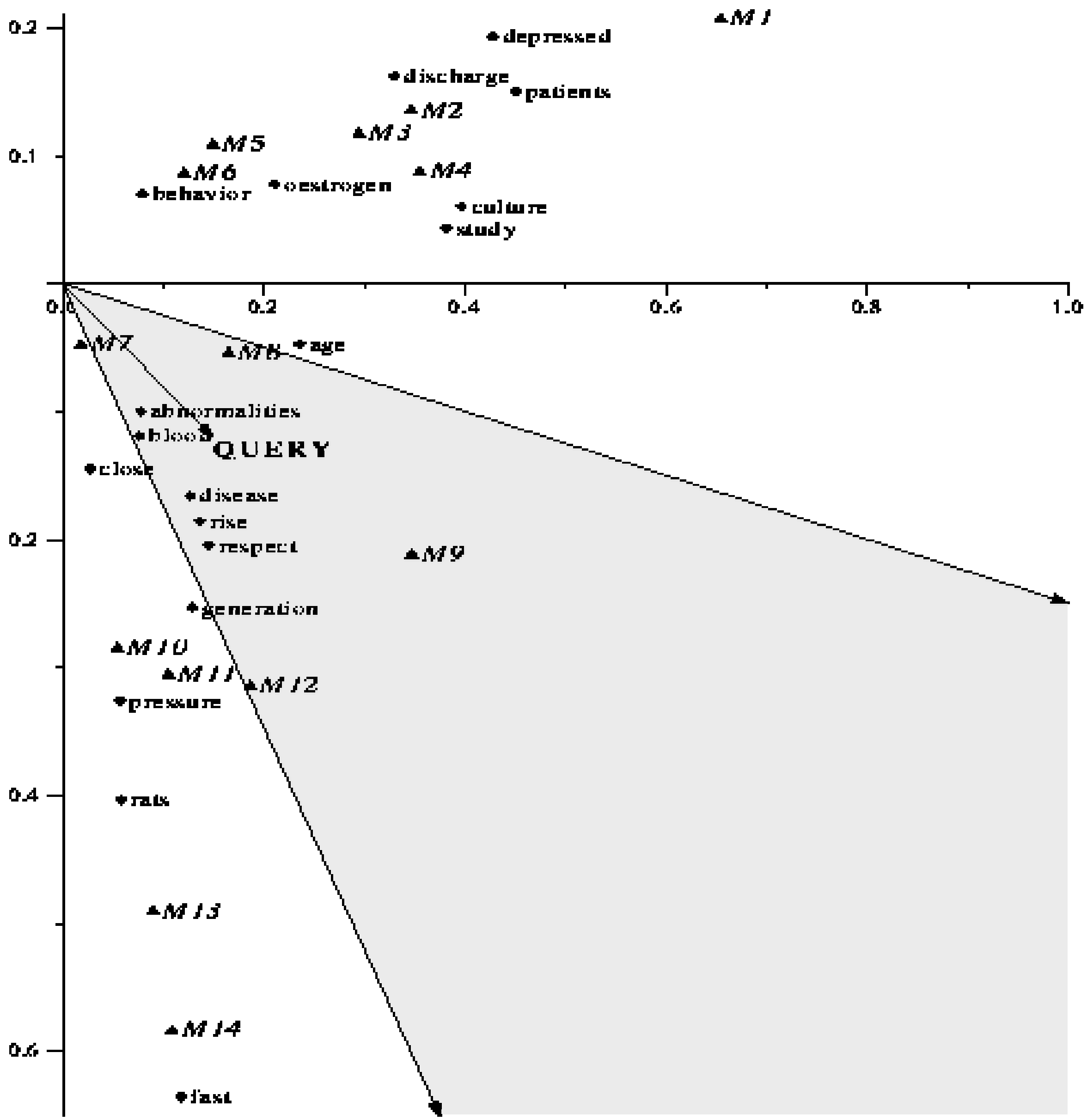
- For n original dimensions, sample covariance matrix is nxn, and has up to n eigenvectors. So n PCs.
- Where does dimensionality reduction come from?

  Can *ignore* the components of lesser significance.



You do lose some information, but if the eigenvalues are small, you don't lose much

- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions



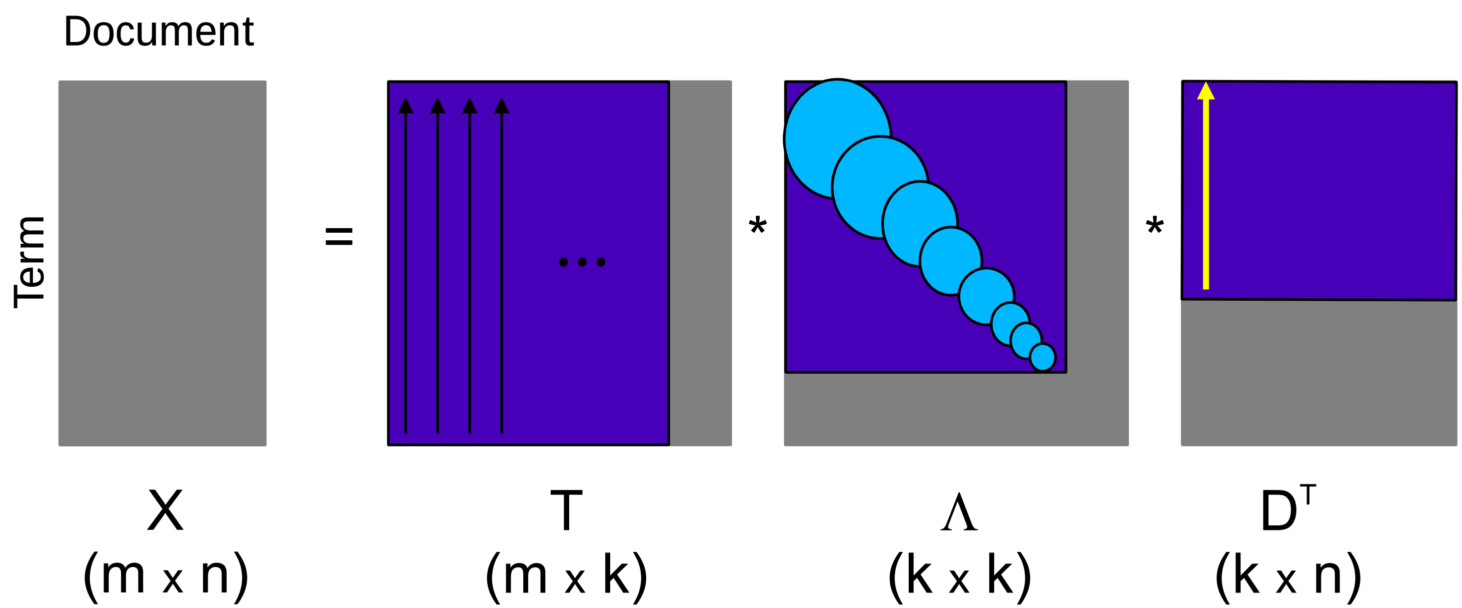
$$\begin{pmatrix} 0.1491 & -0.1199 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}^T \begin{pmatrix} 0.1623 & -0.1372 \\ 0.2068 & -0.0488 \\ 0.0597 & 0.0614 \\ 0.1663 & -0.1313 \\ 0.0258 & -0.1246 \\ 0.4534 & 0.0386 \\ 0.3579 & 0.1710 \\ 0.2931 & 0.1426 \\ 0.0690 & -0.1576 \\ 0.0940 & -0.6535 \\ 0.0599 & -0.2378 \\ 0.1560 & 0.0661 \\ 0.4948 & 0.1091 \\ 0.0460 & -0.3393 \\ 0.0369 & -0.4196 \\ 0.1797 & -0.1456 \\ 0.1067 & -0.2136 \\ 0.3814 & 0.0941 \end{pmatrix} \begin{pmatrix} 3.5919 & 0 \\ 0 & 2.6471 \end{pmatrix}^{-1}$$

Number of Factors					
$k=2$		$k=4$		$k=8$	
M 9	1.00	M 8	0.92	M 8	0.67
M12	0.88	M 9	0.89	M12	0.55
M 8	0.85	M 2	0.64	M10	0.54
M11	0.82	M10	0.48		
M10	0.79	M12	0.46		
M 7	0.74	M11	0.40		
M14	0.72				
M13	0.71				
M 4	0.67				
M 1	0.56				
M 2	0.42				

Within .40 threshold

K is the number of singular values used

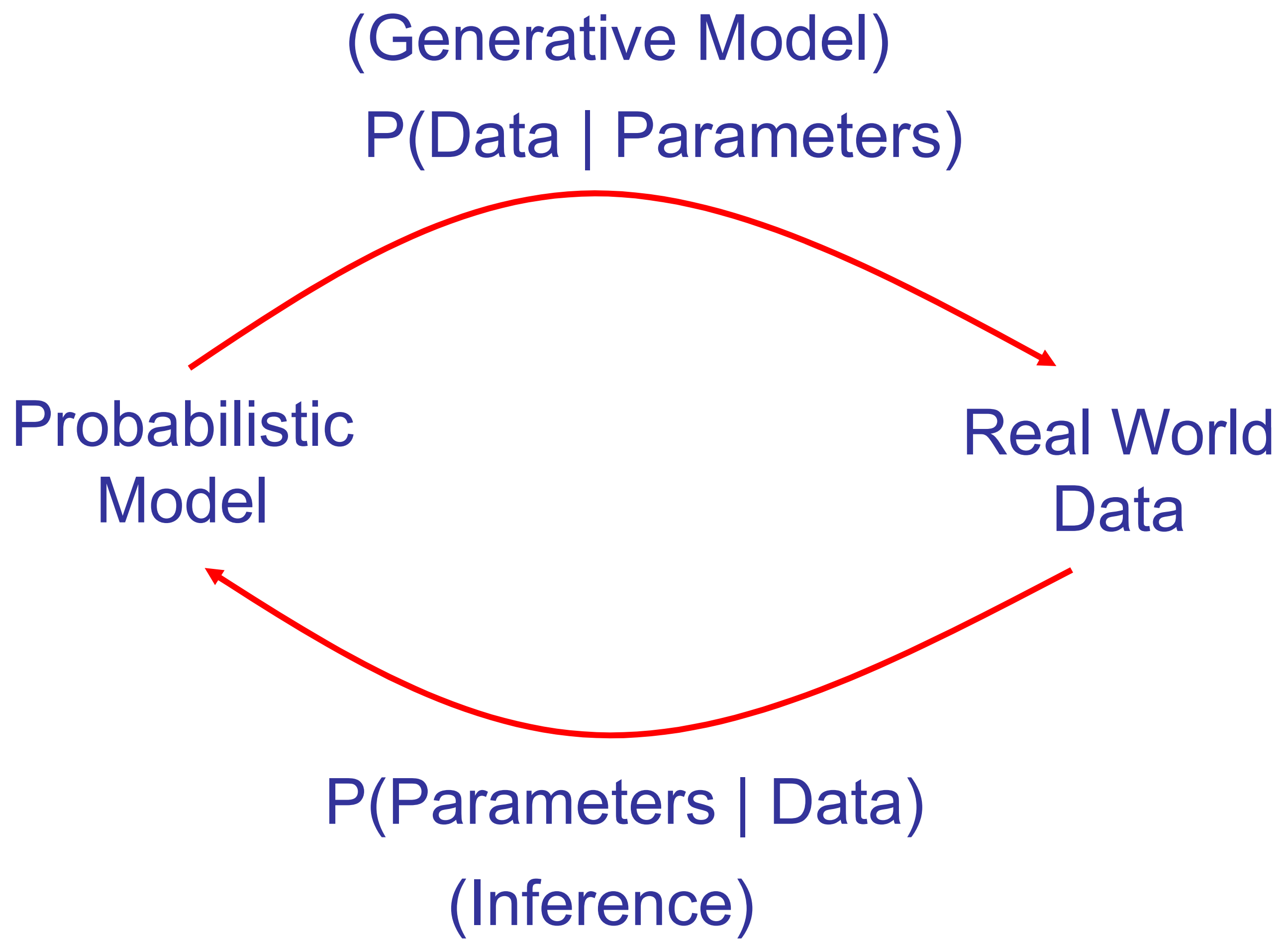
Summary: Latent Semantic Indexing (Deerwester et al., 1990)



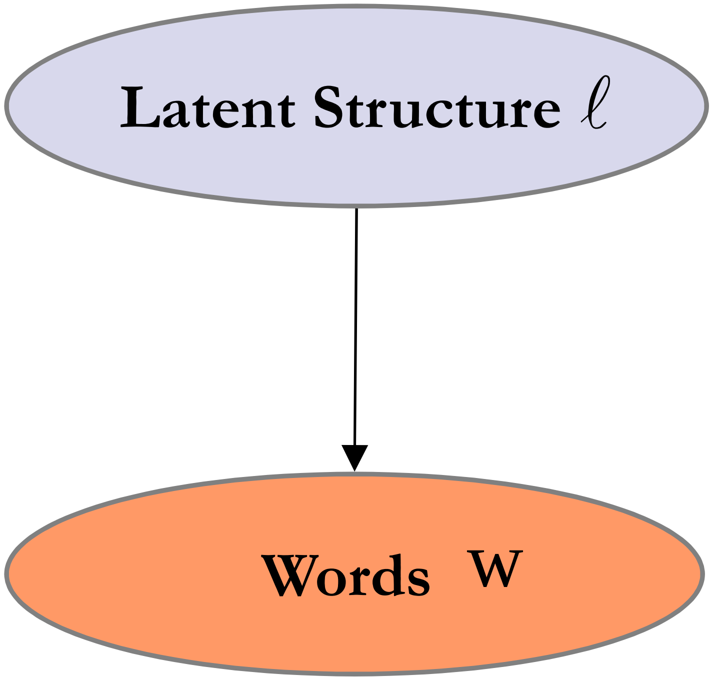
$$\vec{w} = \sum_{k=1}^{K} d_k \lambda_k \vec{T}_k$$

- LSI does not define a properly normalized probability distribution of observed and latent entities
  - Does not support probabilistic reasoning under uncertainty and data fusion

Connecting Probability Models to Data



Latent Semantic Structure in GM



Distribution over words

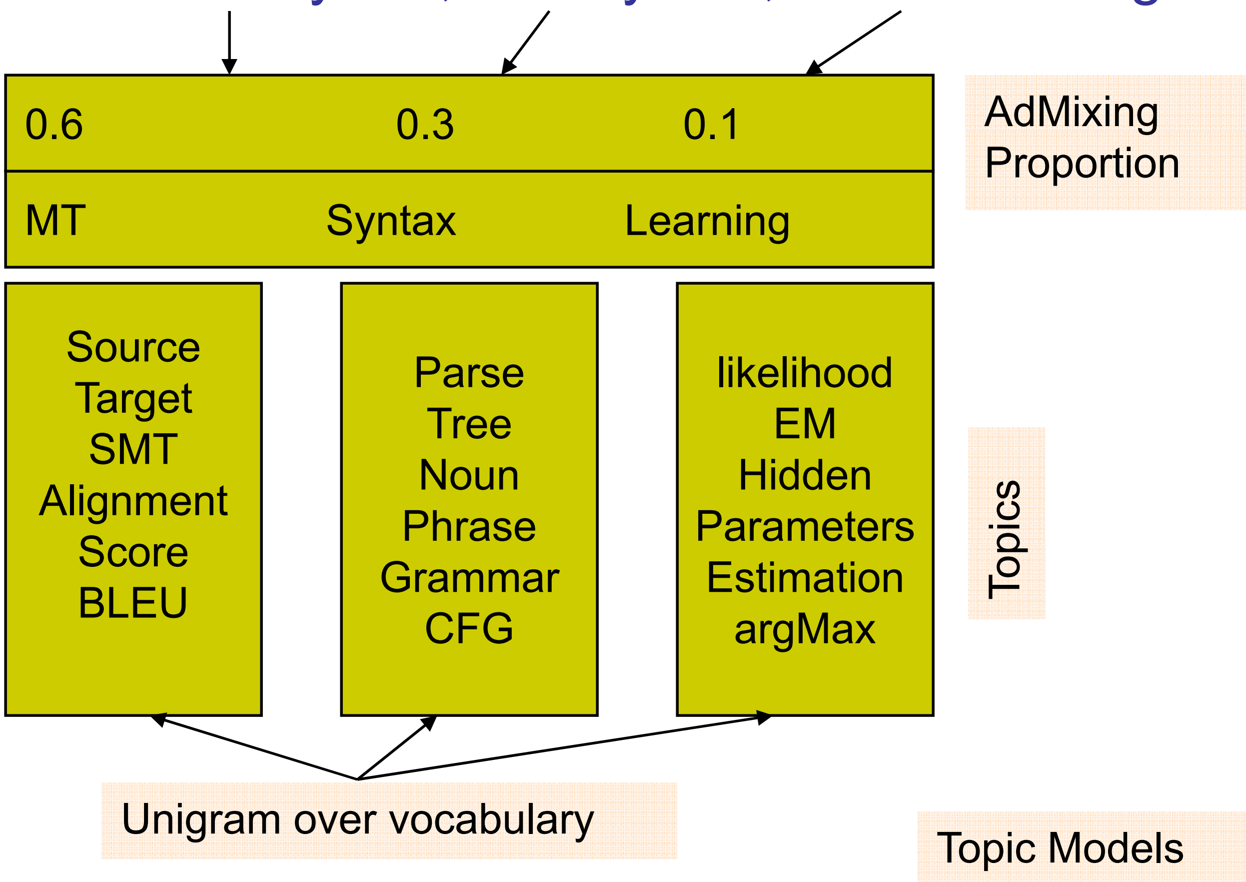
$$P(\mathbf{w}) = \sum_{\ell} P(\mathbf{w}, \ell)$$

Inferring latent structure

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

How to Model Semantics?

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



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.

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

AdMixing 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.

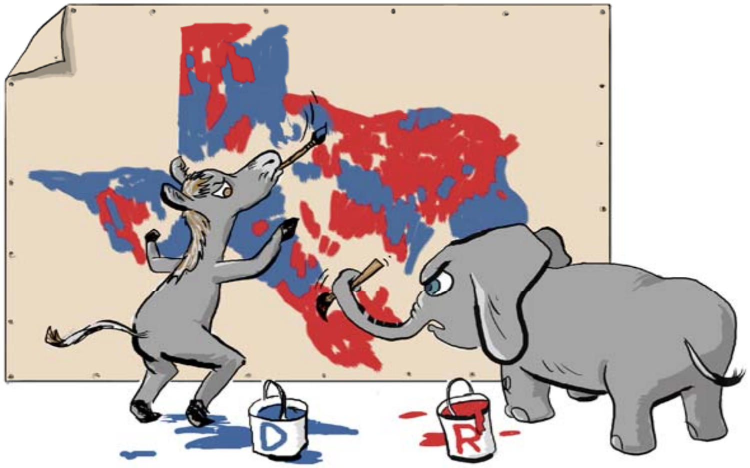
- Q: give me similar document?
  - Structured way of browsing the collection
- Other tasks
  - Dimensionality reduction
    - TF-IDF vs. topic mixing proportion
    - Classification, clustering, and more …

Words in Contexts

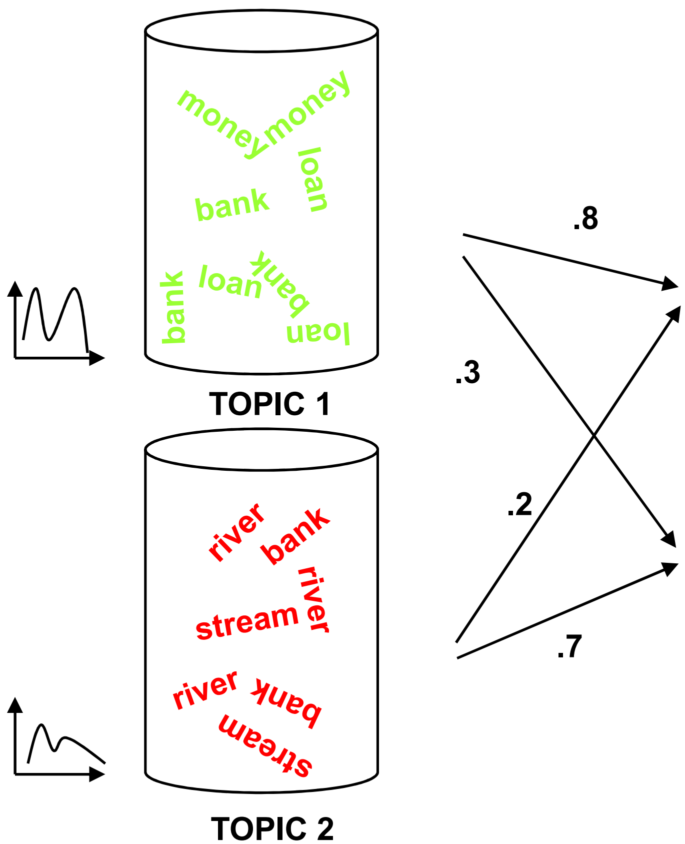
- “ It was a nice **shot**. ”

Words in Contexts (con'd)

- the opposition Labor **Party** fared even worse, with a predicted 35 **seats**, seven less than last **election**.



A possible generative process of a document



DOCUMENT 1: money^1 bank^1 bank^1 loan^1 river^2 stream^2 bank^1 money^1 river^2 bank^1 money^1 bank^1 loan^1 money^1 stream^2 bank^1 money^1 bank^1 bank^1 loan^1 river^2 stream^2 bank^1 money^1 river^2 bank^1 money^1 bank^1 loan^1 bank^1 money^1 stream^2

DOCUMENT 2: river^2 stream^2 bank^2 stream^2 bank^2 money^1 loan^1 river^2 stream^2 loan^1 bank^2 river^2 bank^2 bank^1 stream^2 river^2 loan^1 bank^2 stream^2 bank^2 money^1 loan^1 river^2 stream^2 bank^2 stream^2 bank^2 money^1 river^2 stream^2 loan^1 bank^2 river^2 bank^2 money^1 bank^1 stream^2 river^2 bank^2 stream^2 bank^2 money^1

Mixture Components (distributions over elements)

admixing weight vector θ (represents all components' contributions)

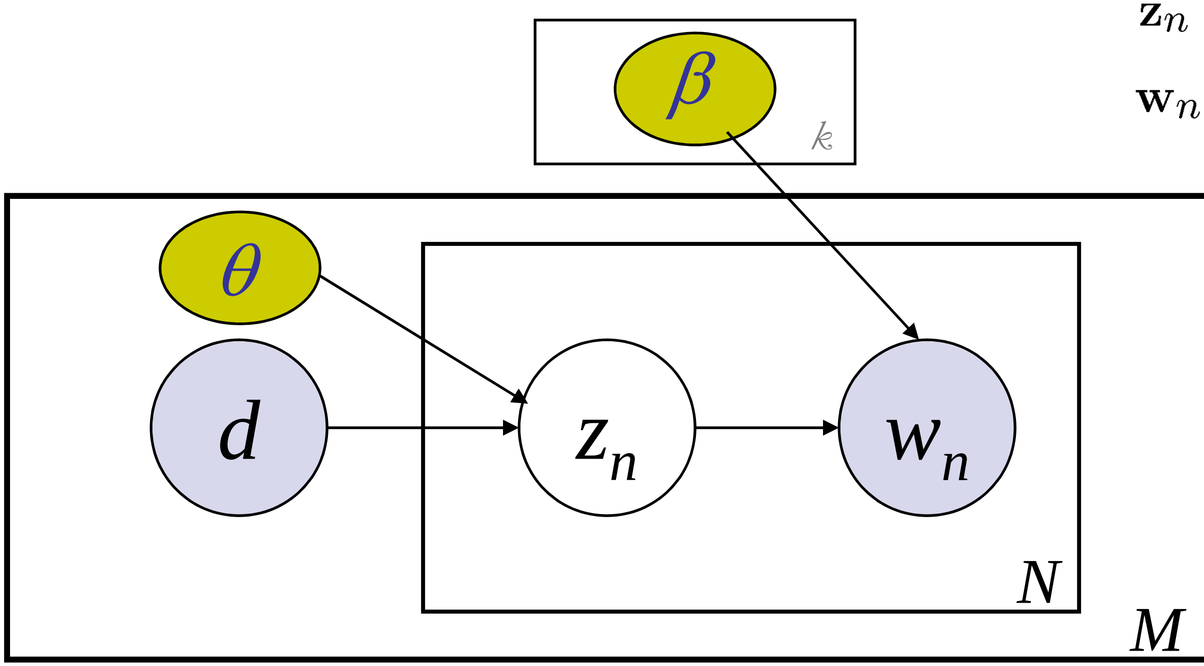
Bayesian approach: use priors

Admixture weights ~ Dirichlet(α)

Mixture components ~ Dirichlet(Γ)

Probabilistic LSI

Hoffman (1999)

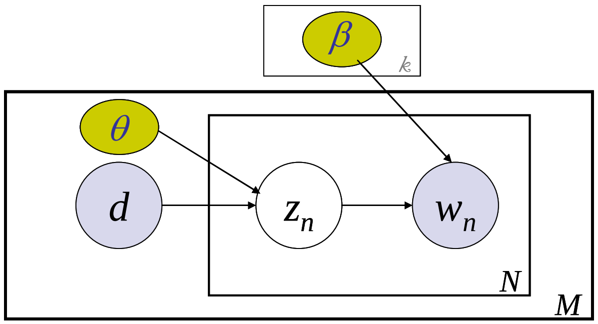


$$\mathbf{z}_n \sim \mathrm{Mult}(\boldsymbol{\theta})$$

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

Probabilistic LSI

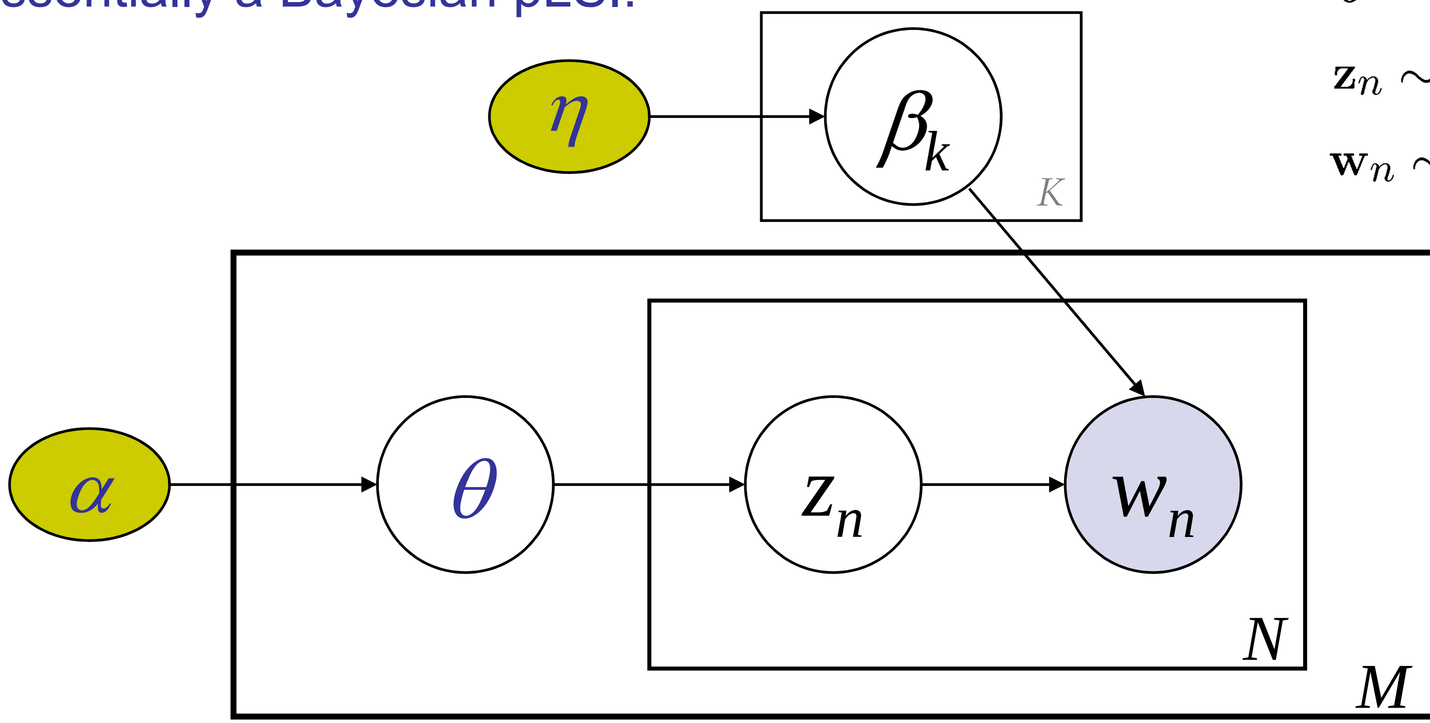
- A "generative" model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a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



Latent Dirichlet Allocation

Blei, Ng and Jordan (2003)

Essentially a Bayesian pLSI:



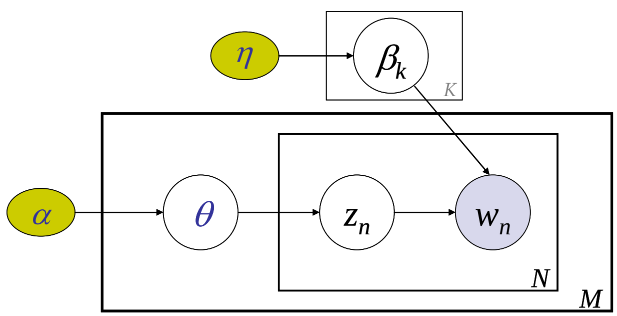
$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

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

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

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 admixing proportions for the components (i.e. topic vector).
- The topic vectors and the word rates each follows a Dirichlet prior --- essentially a Bayesian pLSI



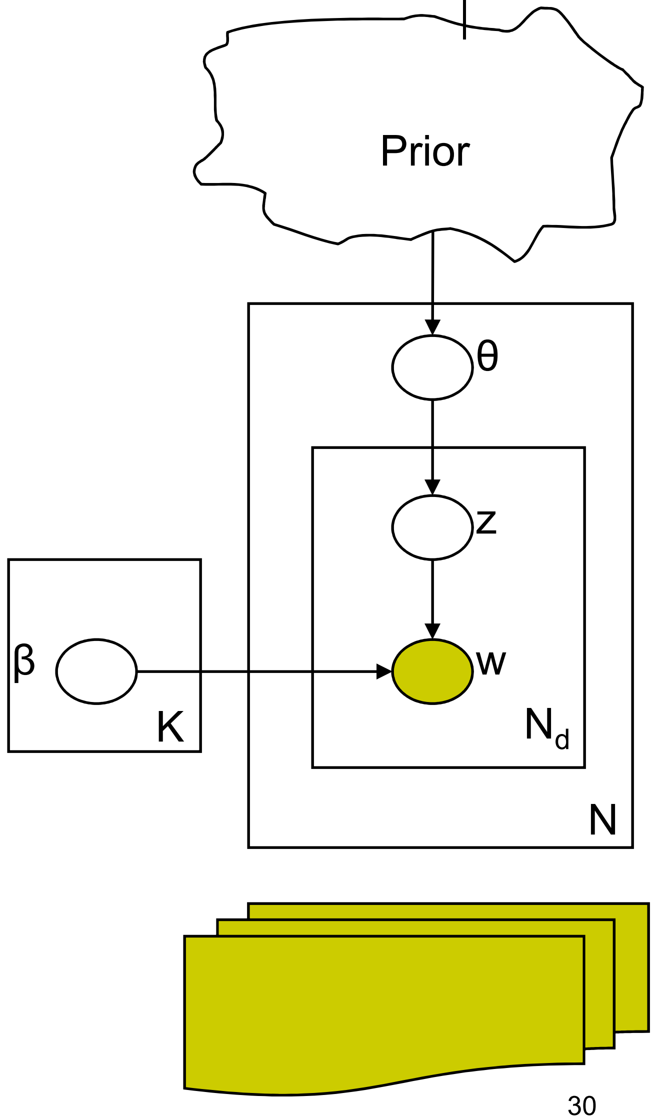
Topic Models = Mixed Membership Models = Admixture

Generating a document

– *Draw* θ from the prior

For each word n

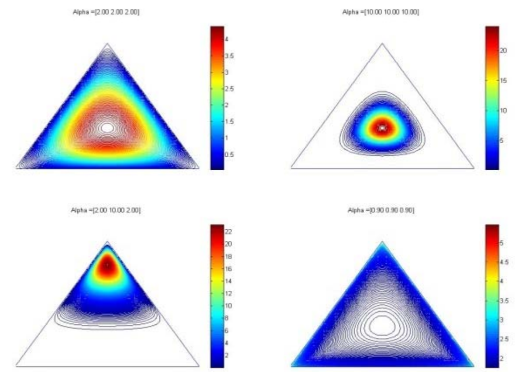
- Draw z_n from $multinomial(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$



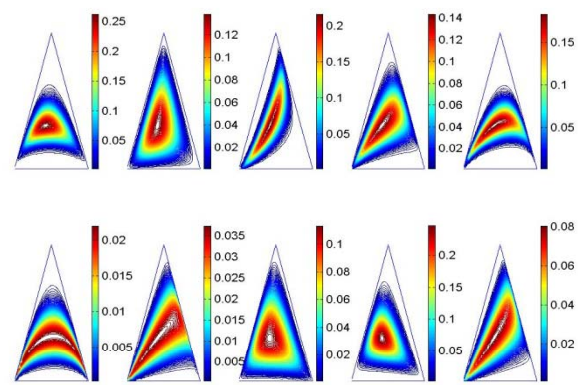
Which prior to use?

Choices of Priors

- 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



- Nested CRP (Blei et al 2005)
  - Defines hierarchy on topics
  - …

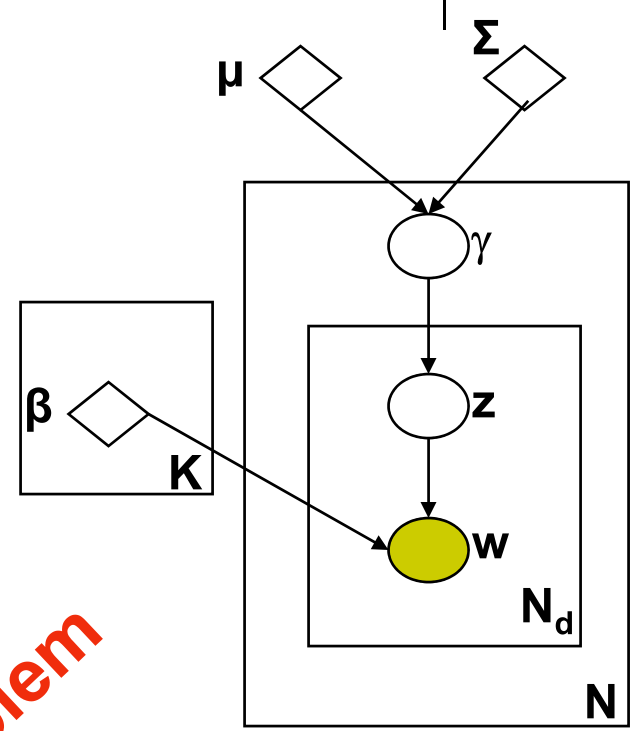
Generative Semantic of LoNTAM

Generating a document

– *Draw* θ from the prior

For each word n

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



$$\theta \sim LN_K(\mu, \Sigma)$$

$$\gamma \sim N_{K-1}(\mu, \Sigma) \qquad \gamma_K = 0$$

$$\theta_i = \exp\left\{\gamma_i - \log\left(1 + \sum_{i=1}^{K-1} e^{\gamma_i}\right)\right\}$$

$$C(\gamma) = \log\left(1 + \sum_{i=1}^{K-1} e^{\gamma_i}\right)$$

Problem

- **Log Partition Function**
- **Normalization Constant**

Outcomes from a topic model

- The "topics" β in a corpus:

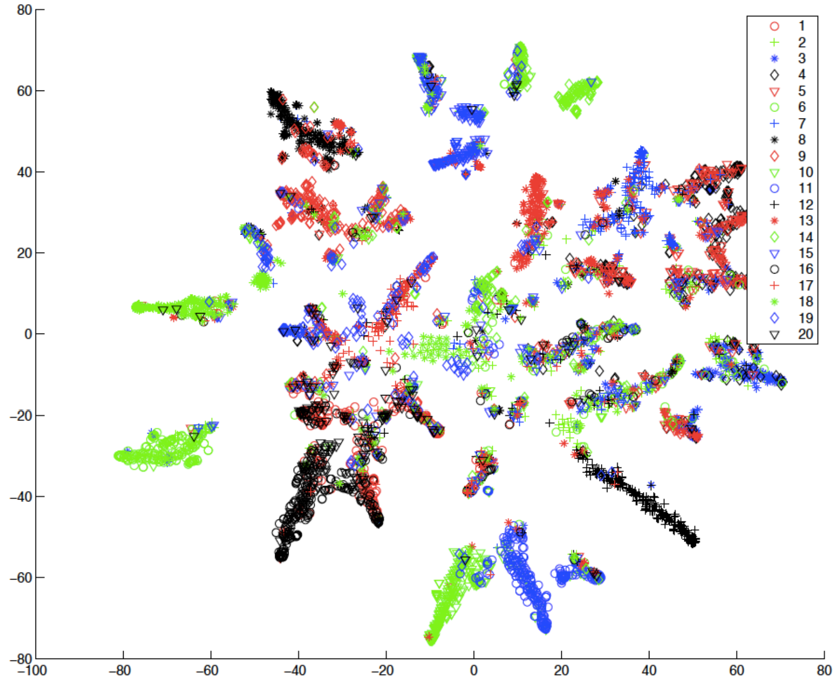
	T 59	T 104	T 31
comp.graphics	image	ftp	card
	jpeg	pub	monitor
	color	graphics	dos
	file	mail	video
	gif	version	apple
	images	tar	windows
	format	file	drivers
	bit	information	vga
	files	send	cards
	display	server	graphics
	T 30	**T 84**	**T 44**
sci.electronics	power	water	sale
	ground	energy	price
	wire	air	offer
	circuit	nuclear	shipping
	supply	loop	sell
	voltage	hot	interested
	current	cold	mail
	wiring	cooling	condition
	signal	heat	email
	cable	temperature	cd

	T 42	T 78	T 47
politics.mideast	israel	jews	armenian
	israeli	jewish	turkish
	peace	israel	armenians
	writes	israeli	armenia
	article	arab	turks
	arab	people	genocide
	war	arabs	russian
	lebanese	center	soviet
	lebanon	jew	people
	people	nazi	muslim
	T 44	**T 94**	**T 49**
misc.forsale	sale	don	drive
	price	mail	scsi
	offer	call	disk
	shipping	package	hard
	sell	writes	mb
	interested	send	drives
	mail	number	ide
	condition	ve	controller
	email	hotel	floppy
	cd	credit	system

- There is no name for each "topic", you need to name it!
- There is no objective measure of good/bad
- The shown topics are the "good" ones, there are many many trivial ones, meaningless ones, redundant ones, … you need to manually prune the results
- How many topics? …

Outcomes from a topic model

- The "topic vector" θ of each doc



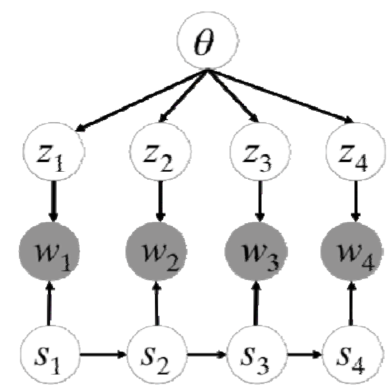
- Create an embedding of docs in a "topic space"
- Their no ground truth of θ to measure quality of inference
- But on θ it is possible to define an "objective" measure of goodness, such as classification error, retrieval of similar docs, clustering, etc., of documents
- But there is no consensus on whether these tasks bear the true value of topic models …

Outcomes from a topic model

- The per-word topic indicator z:

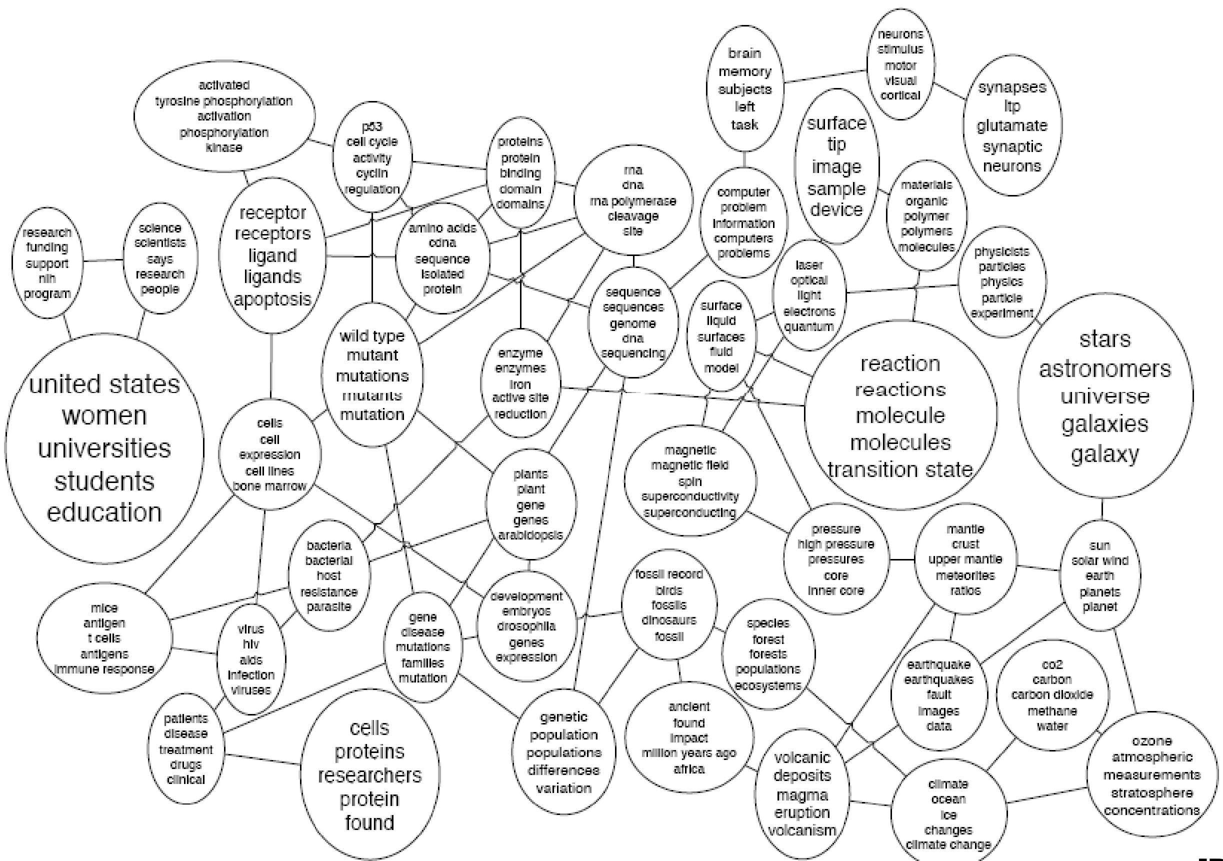
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

- Not very useful under the bag of word representation, because of loss of ordering
- But it is possible to define simple probabilistic linguistic constraints (e.g, bi-grams) over z and get potentially interesting results [Griffiths, Steyvers, Blei, & Tenenbaum, 2004]



Outcomes from a topic model

- The topic graph S (when using CTM):



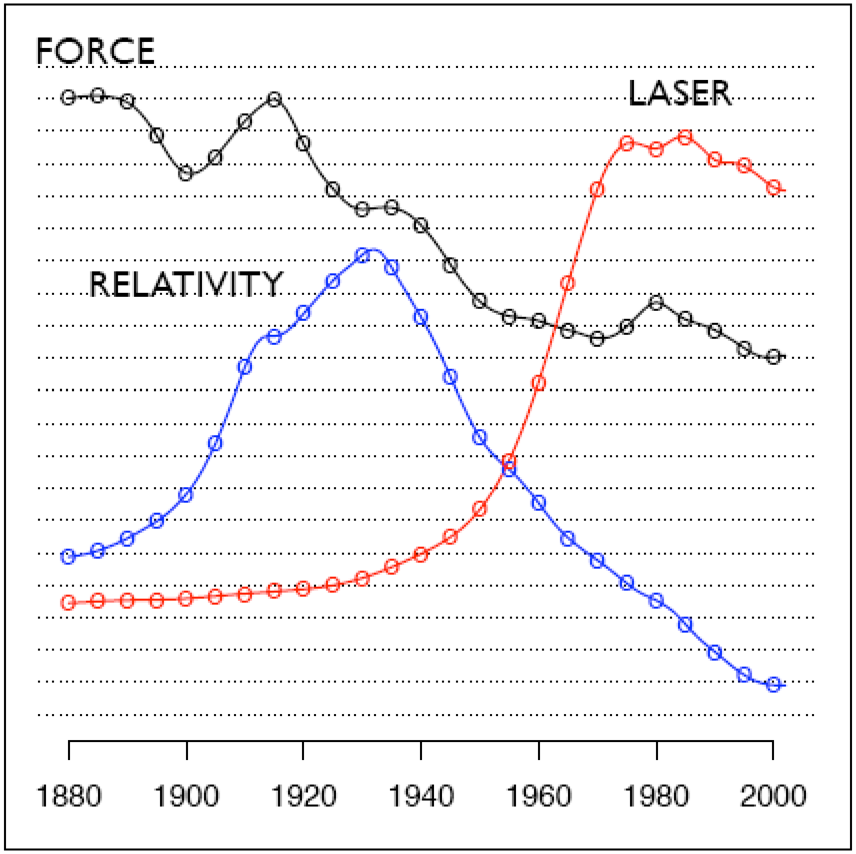
[David Blei, MLSS09]

- Kind of interesting for understanding/visualizing large corpora

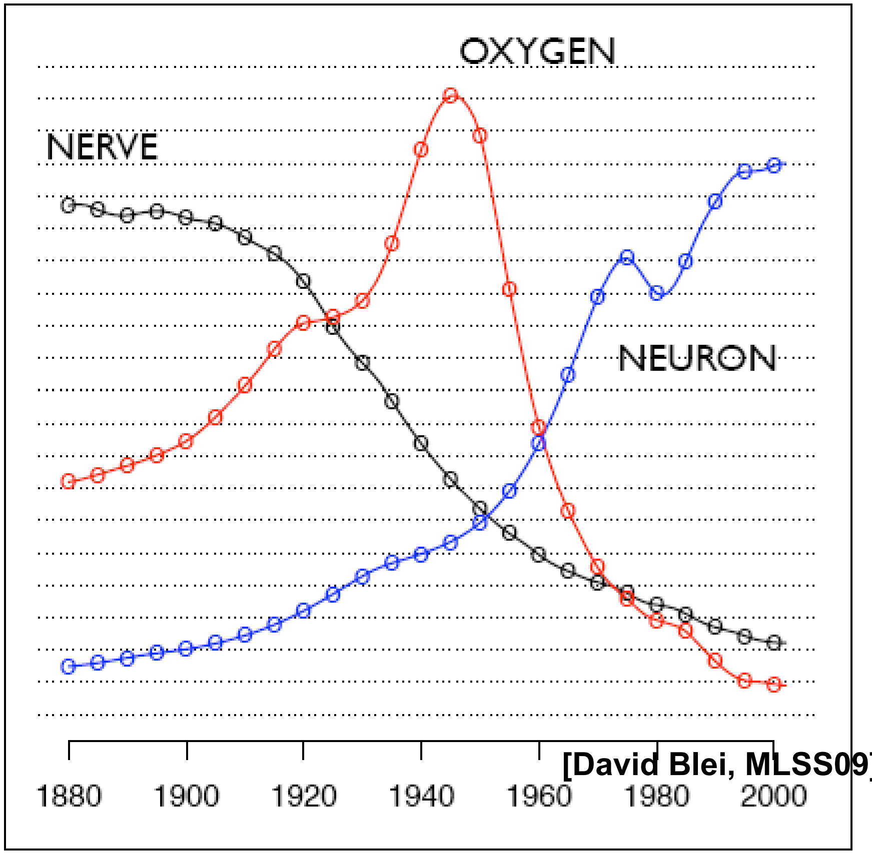
Outcomes from a topic model

- Topic change trends

"Theoretical Physics"

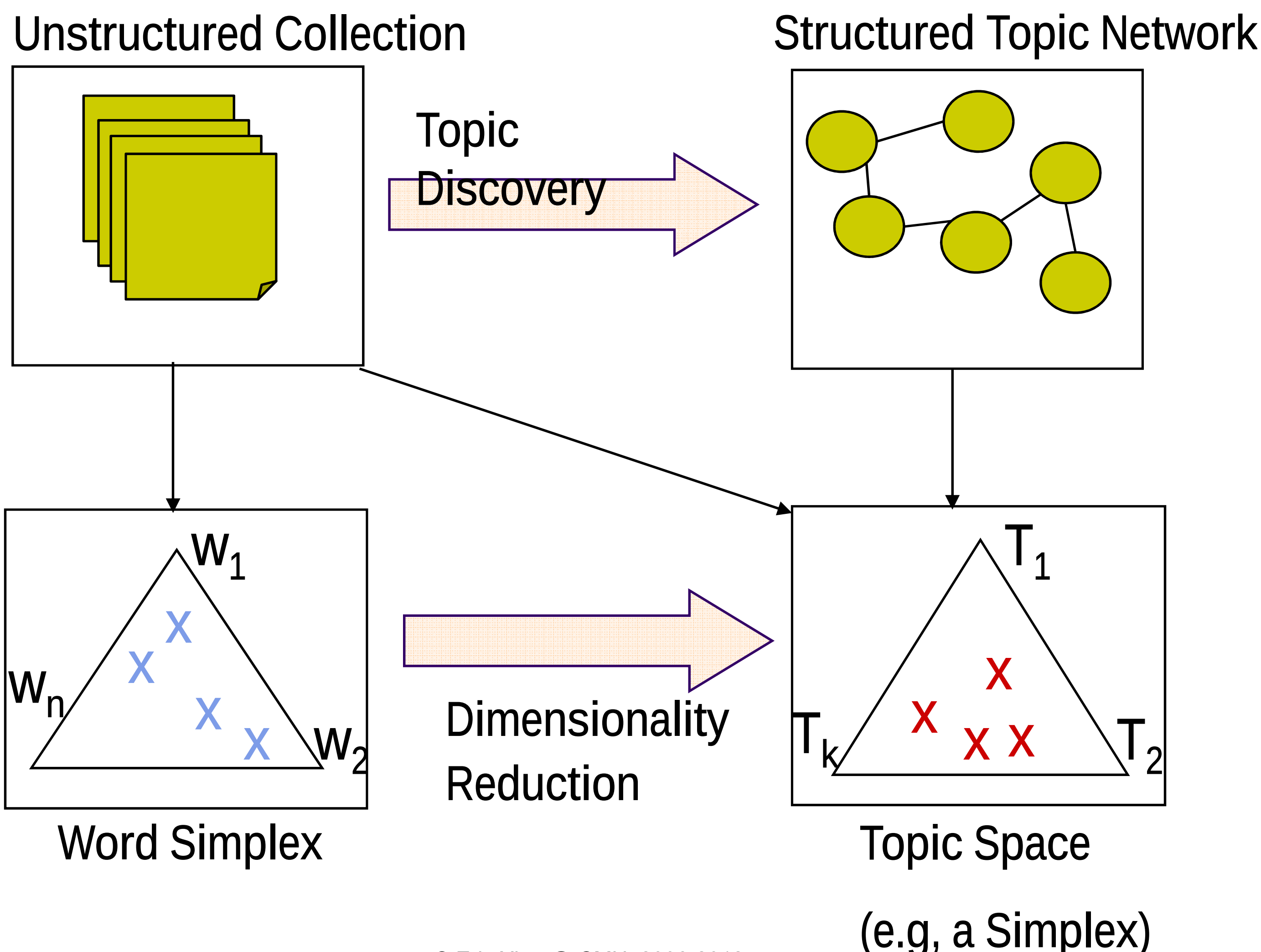


"Neuroscience"



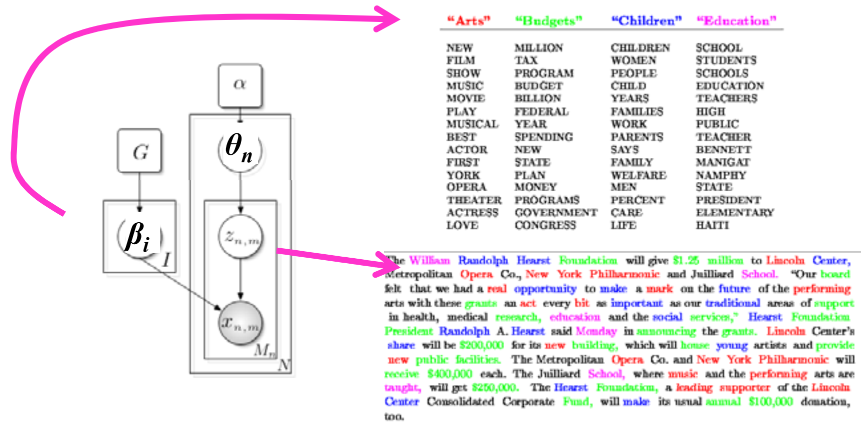
[David Blei, MLSS09]

The Big Picture



Computation on LDA

- Inference
  - Given a Document D
    - Posterior: P(Θ | μ,Σ, β ,D)
    - Evaluation: P(D| μ,Σ, β)



- Learning
  - Given a collection of documents $\{D_i\}$
    - Parameter estimation

$$\arg\max_{(\mu,\Sigma,\beta)} \sum \log\left(P\left(D_i \middle| \mu, \Sigma, \beta\right)\right)$$

Exact Bayesian inference on LDA is intractable

- A possible query:

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

- Close form solution?

$$p(\theta_n \mid D) = \frac{p(\theta_n, D)}{p(D)}$$
$$= \frac{\sum_{\{z_{n,m}\}} \int \left(\prod_n \left(\prod_m p(x_{n,m} \mid \beta_{z_n}) p(z_{n,m} \mid \theta_n) \right) p(\theta_n \mid \alpha) \right) p(\phi \mid G) d\theta_{-n} \, d\beta}{p(D)}$$

$$p(D) = \sum_{\{z_{n,m}\}} \int \cdots \int \left(\prod_n \left(\prod_m p(x_{n,m} \mid \beta_{z_n}) p(z_{n,m} \mid \theta_n) \right) p(\theta_n \mid \alpha) \right) p(\beta \mid G) d\theta_1 \cdots d\theta_N d\beta$$

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

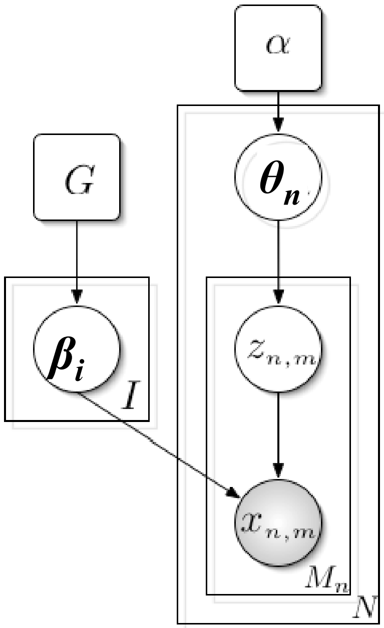
Approximate Inference

- Variational Inference
  - Mean field approximation (Blei et al)
  - Expectation propagation (Minka et al)
  - Variational 2^{nd}-order Taylor approximation (Ahmed and Xing)

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

Collapsed Gibbs sampling
(Tom Griffiths & Mark Steyvers)

- Collapsed Gibbs sampling
  - Integrate out θ



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

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

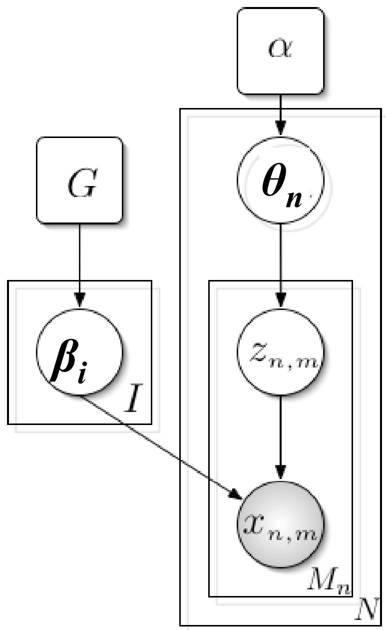
$\mathbf{z}_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_n^{(t)}$

$$\left\{z^{(1)}, z^{(2)}, \ldots, z^{(T)}\right\}$$

$$\theta = \frac{1}{T} \sum_t z^{(t)}$$

Gibbs sampling

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



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

$$= \frac{n_{-i,j}^{(w_i)} + G}{n_{-i,j}^{(\cdot)} + WG} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

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

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

Gibbs sampling

i	w_i	d_i	**iteration 1** 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

			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	

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

Gibbs sampling

			iteration 1	iteration 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	

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

Gibbs sampling

			iteration 1	iteration 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	

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

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	

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

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	

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

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	

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

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

$$\theta = \frac{1}{T} \sum_t z^{(t)}$$

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

Learning a TM

- Maximum likelihood estimation:

- Need statistics on topic-specific word assignment (due to z), topic vector distribution (due to θ), etc.
  - E.g,, this is the formula for topic k:

$$\beta_k = \frac{1}{\sum_d N_d} \sum_{d=1}^{D} \sum_{d_n=1}^{N_d} \delta(z_{d,d_n}, k) w_{d,d_n}$$

- These are hidden variables, therefore need an EM algorithm (also known as data augmentation, or DA, in Monte Carlo paradigm)

- This is a "reduce" step in parallel implementation

Conclusion

- GM-based topic models are cool
  - Flexible
  - Modular
  - Interactive
- There are many ways of implementing topic models
  - unsupervised
  - supervised
- Efficient Inference/learning algorithms
  - GMF, with Laplace approx. for non-conjugate dist.
  - MCMC
- Many applications
  - …
  - Word-sense disambiguation
  - Image understanding
  - Network inference