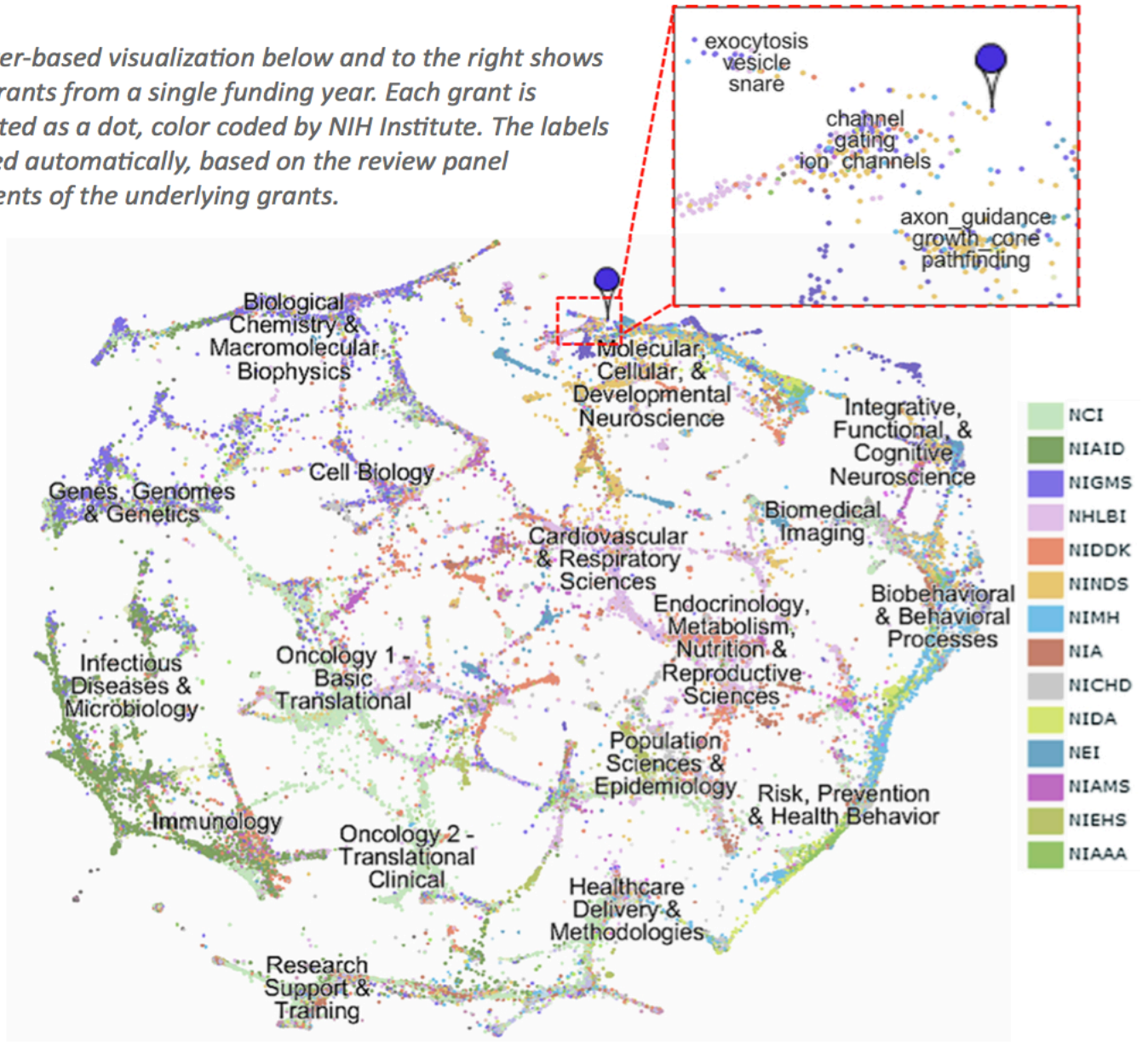
CS544: Topic Models
Latent Dirichlet Allocation

March 14, 2013

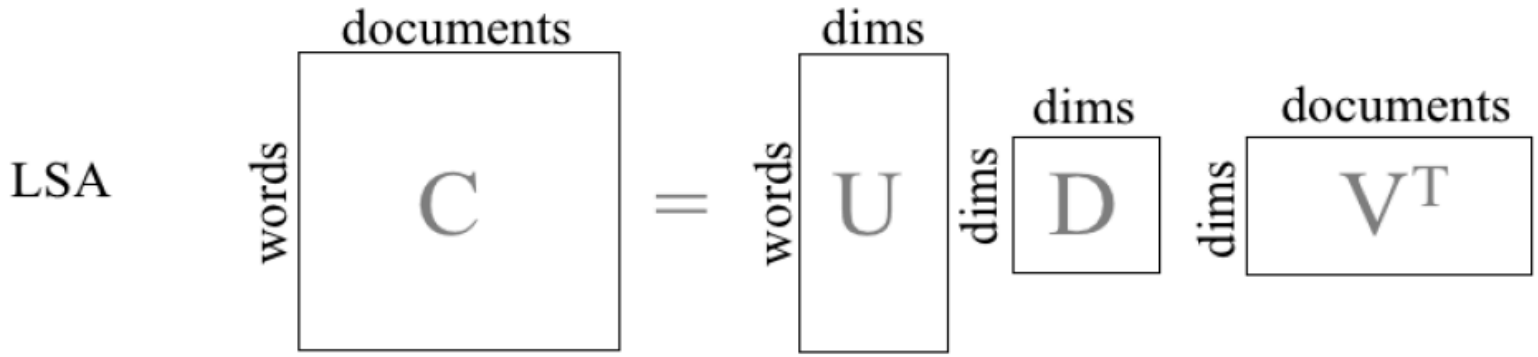
Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

The cluster-based visualization below and to the right shows all NIH grants from a single funding year. Each grant is represented as a dot, color coded by NIH Institute. The labels are placed automatically, based on the review panel assignments of the underlying grants.



https://app.nihmaps.org/public/browser/#center=0.000000%2C0.000000;data=nih.active;showViz=true;zoom=2.05;
http://www.nihmaps.org/movies/introduction.mov

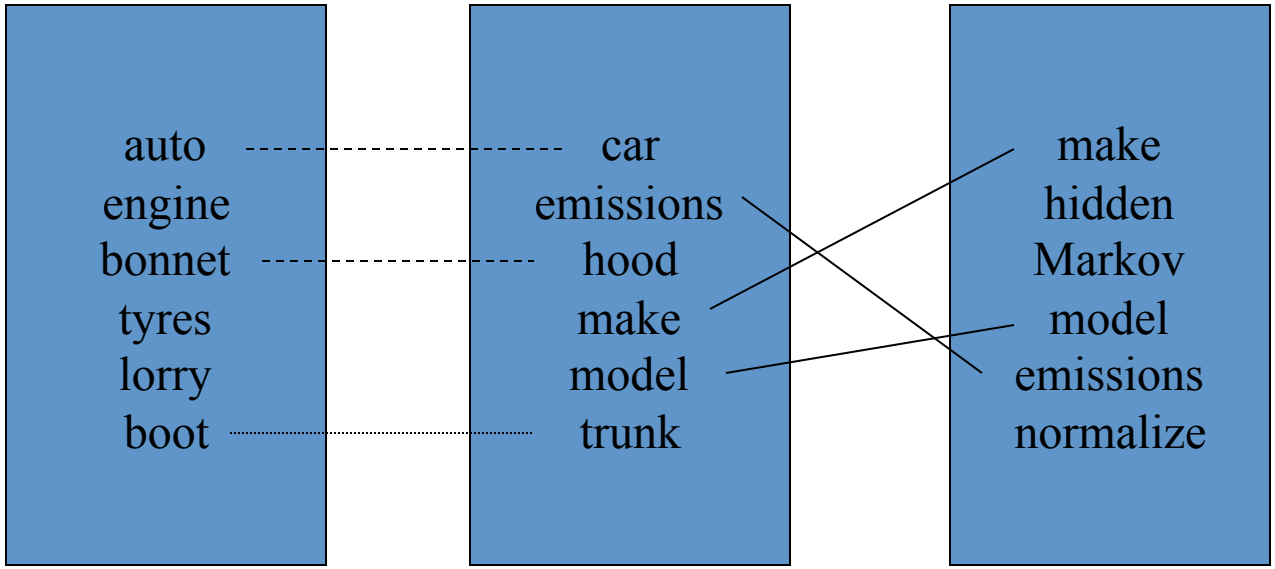
LSA


LSA
documents
words
C
=
dims
words
U
dims
dims
D
documents
dims
V^T

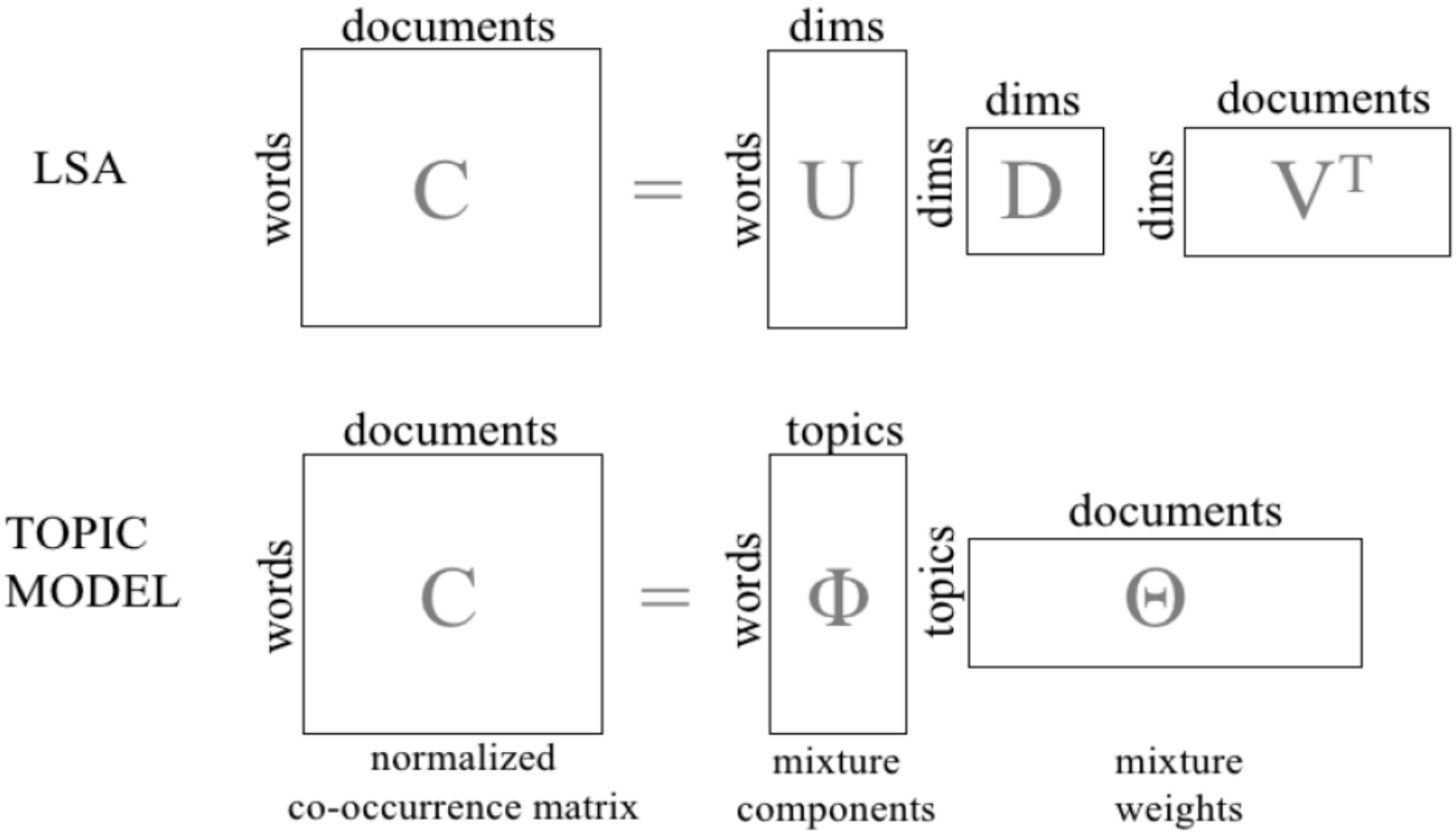
Problem with LSA

Problem with LSA

- Solves **Synonymy**, but does not solve **Polysemy**



LSA and Topic Models



Probabilistic Topic Models

David M. Blei

Andrew Ng

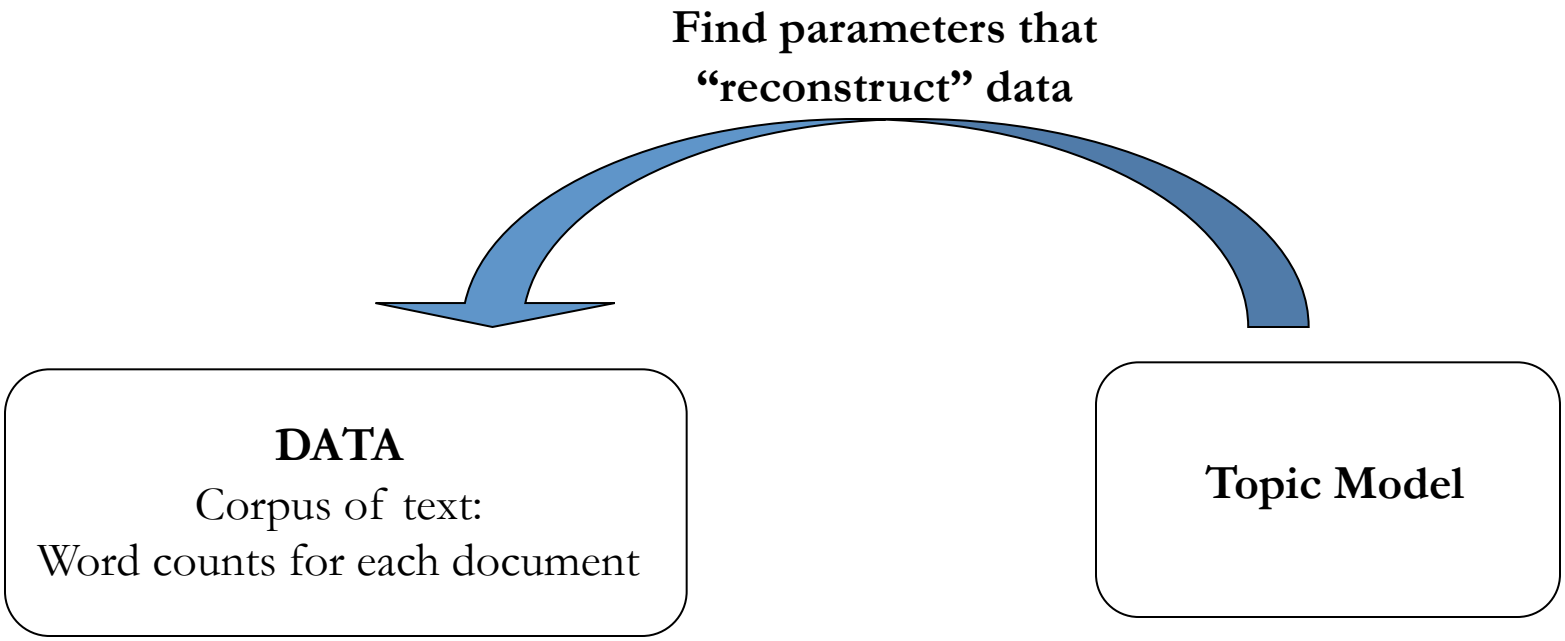
Michael Jordon

• Originated in domain of statistics & machine learning (e.g., Hoffman, 2001; Blei, Ng, Jordan, 2003)

Probabilistic Topic Models

- Extracts **topics** from large collections of text
- Topics are **interpretable** unlike the arbitrary dimensions of LSA

Model is Generative



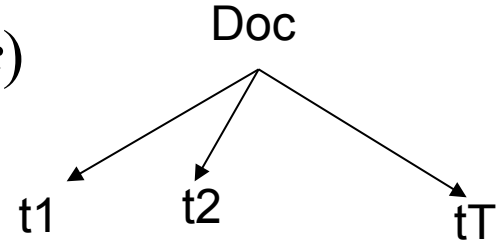
Probabilistic Topic Models

- Each document is a probability distribution over topics
- Each topic is a probability distribution over words

Probabilistic Latent Semantic Analysis

$$P(doc) = P(term_1 \mid doc)P(term_2 \mid doc)...P(term_L \mid doc)$$

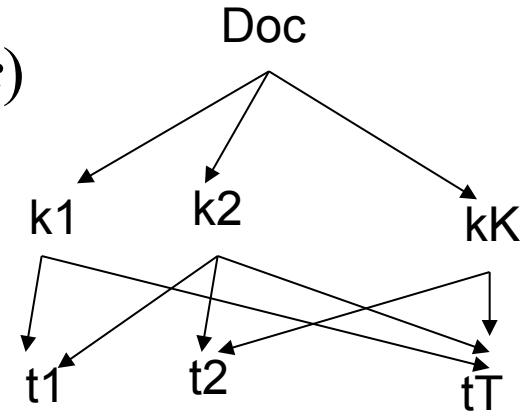
$$= \prod_{l=1}^{L} P(term_l \mid doc) = \prod_{t=1}^{T} P(term_t \mid doc)^{X(term_t, doc)}$$



Probabilistic Latent Semantic Analysis

- Now let us have K topics:

$$P(term_t \mid doc) = \sum_{k=1}^{K} P(term_t \mid topic_k) P(topic_k \mid doc)$$



The same, written using shorthands :

$$P(t \mid doc) = \sum_{k=1}^{K} P(t \mid k) P(k \mid doc)$$

So by replacing this, for any doc in the collection,

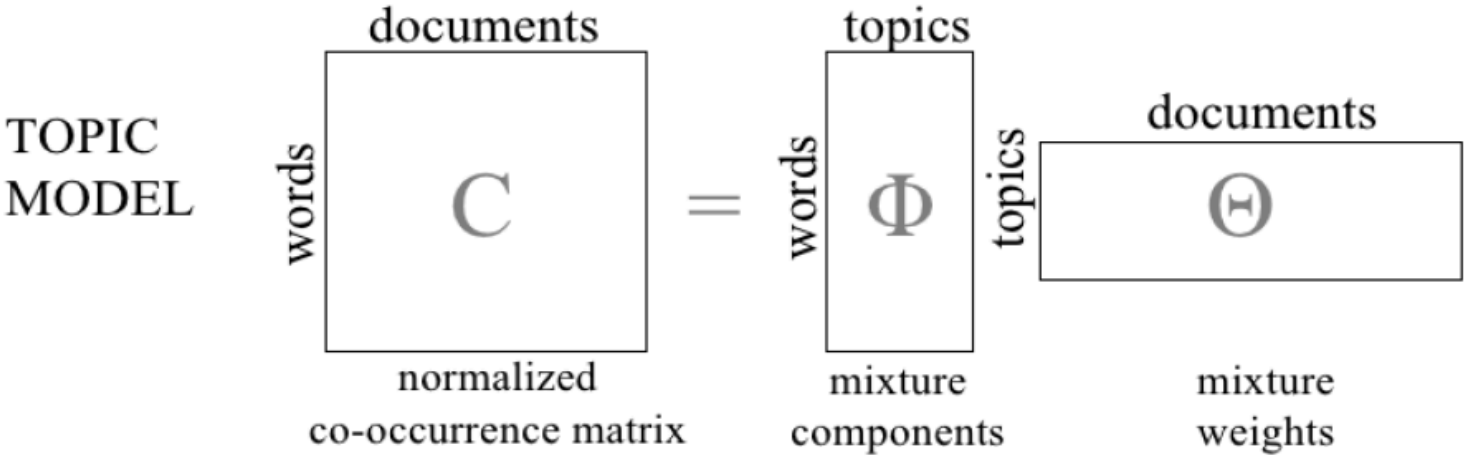
$$P(doc) = \prod_{t=1}^{T} \{ \sum_{k=1}^{K} P(t \mid k) P(k \mid doc) \}^{X(t,doc)}$$

"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

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.

Topic Model

- P(tlk) for all t and k, is a term by topic matrix (gives which terms make up a topic)
- P(kldoc) for all k and doc, is a topic by document matrix (gives which topics are in a document)



EXAMPLE

Analysis of TASA Corpus

- Given a text collection written by first grade to college students

- Data has following characteristics
  - 26,000+ word types (stop words removed)
  - 37,000+ documents
  - 6,000,000+ word tokens

- Find topics in the data

Topics in the Educational Corpus (TASA)

- 37K docs, 26K words
- 1700 topics, e.g.:

PRINTING	PLAY	TEAM	JUDGE	HYPOTHESIS	STUDY
PAPER	PLAYS	GAME	TRIAL	EXPERIMENT	TEST
PRINT	STAGE	BASKETBALL	COURT	SCIENTIFIC	STUDYING
PRINTED	AUDIENCE	PLAYERS	CASE	OBSERVATIONS	HOMEWORK
TYPE	THEATER	PLAYER	JURY	SCIENTISTS	NEED
PROCESS	ACTORS	PLAY	ACCUSED	EXPERIMENTS	CLASS
INK	DRAMA	PLAYING	GUILTY	SCIENTIST	MATH
PRESS	SHAKESPEARE	SOCCER	DEFENDANT	EXPERIMENTAL	TRY
IMAGE	ACTOR	PLAYED	JUSTICE	TEST	TEACHER
PRINTER	THEATRE	BALL	EVIDENCE	METHOD	WRITE
PRINTS	PLAYWRIGHT	TEAMS	WITNESSES	HYPOTHESES	PLAN
PRINTERS	PERFORMANCE	BASKET	CRIME	TESTED	ARITHMETIC
COPY	DRAMATIC	FOOTBALL	LAWYER	EVIDENCE	ASSIGNMENT
COPIES	COSTUMES	SCORE	WITNESS	BASED	PLACE
FORM	COMEDY	COURT	ATTORNEY	OBSERVATION	STUDIED
OFFSET	TRAGEDY	GAMES	HEARING	SCIENCE	CAREFULLY
GRAPHIC	CHARACTERS	TRY	INNOCENT	FACTS	DECIDE
SURFACE	SCENES	COACH	DEFENSE	DATA	IMPORTANT
PRODUCED	OPERA	GYM	CHARGE	RESULTS	NOTEBOOK
CHARACTERS	PERFORMED	SHOT	CRIMINAL	EXPLANATION	REVIEW

Polysemy

PRINTING	**PLAY**	TEAM	JUDGE	HYPOTHESIS	STUDY
PAPER	PLAYS	GAME	TRIAL	EXPERIMENT	**TEST**
PRINT	STAGE	BASKETBALL	**COURT**	SCIENTIFIC	STUDYING
PRINTED	AUDIENCE	PLAYERS	CASE	OBSERVATIONS	HOMEWORK
TYPE	THEATER	PLAYER	JURY	SCIENTISTS	NEED
PROCESS	ACTORS	**PLAY**	ACCUSED	EXPERIMENTS	CLASS
INK	DRAMA	PLAYING	GUILTY	SCIENTIST	MATH
PRESS	SHAKESPEARE	SOCCER	DEFENDANT	EXPERIMENTAL	TRY
IMAGE	ACTOR	PLAYED	JUSTICE	**TEST**	TEACHER
PRINTER	THEATRE	BALL	**EVIDENCE**	METHOD	WRITE
PRINTS	PLAYWRIGHT	TEAMS	WITNESSES	HYPOTHESES	PLAN
PRINTERS	PERFORMANCE	BASKET	CRIME	TESTED	ARITHMETIC
COPY	DRAMATIC	FOOTBALL	LAWYER	**EVIDENCE**	ASSIGNMENT
COPIES	COSTUMES	SCORE	WITNESS	BASED	PLACE
FORM	COMEDY	**COURT**	ATTORNEY	OBSERVATION	STUDIED
OFFSET	TRAGEDY	GAMES	HEARING	SCIENCE	CAREFULLY
GRAPHIC	**CHARACTERS**	TRY	INNOCENT	FACTS	DECIDE
SURFACE	SCENES	COACH	DEFENSE	DATA	IMPORTANT
PRODUCED	OPERA	GYM	CHARGE	RESULTS	NOTEBOOK
CHARACTERS	PERFORMED	SHOT	CRIMINAL	EXPLANATION	REVIEW

Three Documents with the word “play”

(numbers & colors → topic assignments)

A **Play**[082] is written[082] to be performed[082] on a stage[082] before a live[093] audience[082] or before motion[270] picture[004] or television[004] cameras[004] (for later[054] viewing[004] by large[202] audiences[082]). A **Play**[082] is written[082] because playwrights[082] have something

He was listening[077] to music[077] coming[009] from a passing[043] riverboat. The music[077] had already captured[006] his heart[157] as well as his ear[119]. It was jazz[077]. Bix beiderbecke had already had music[077] lessons[077]. He wanted[268] to **play**[077] the cornet. And he wanted[268] to **play**[077] jazz[077]

Jim[296] **plays**[166] the game[166]. Jim[296] likes[081] the game[166] for one. The game[166] book[254] helps[081] jim[296]. Don[180] comes[040] into the house[038]. Don[180] and jim[296] read[254] the game[166] book[254]. The boys[020] see a game[166] for two. The two boys[020] **play**[166] the game[166].

Enron Email Data

From: PGE News
To: ALL PGE EMPLOYEES
Date: 8/14/01 2:54PM
Subject: Jeff Skilling resigns as CEO of Enron

PGE News ..... August 14, 2001

Jeff Skilling resigns as CEO of Enron

Enron today announced that President and CEO Jeff Skilling has resigned, effective immediately, and that the Enron Board of Directors has asked Ken Lay to resume his role as Chairman and CEO.

"Stan Horton called this afternoon to inform me of Jeff's decision to step down for personal reasons," says PGE CEO and President Peggy Fowler. Horton, CEO of Enron Transportation, is Fowler's executive connection to the Enron team. "He wanted to let me know that Mr. Skilling's departure will not in any way impact Enron's ongoing strategy for success and we should expect no near-term dramatic organizational changes."

"Clearly, Enron will continue to focus on increasing the company's stock value," Fowler added. "PGE can help in this effort by remaining committed to our Scorecard goals and operational excellence."

Below is the letter Ken Lay is sending to Enron employees this afternoon announcing the decision:

To: Enron Employees Worldwide
From: Ken Lay

It is with regret that I have to announce that Jeff Skilling is leaving Enron. Today, the Board of Directors accepted his resignation as President and CEO of Enron. Jeff is resigning for personal reasons and his decision is voluntary. I regret his decision, but I accept and understand it. I have worked closely with Jeff for more than 15 years, including 11 here at Enron, and have had few, if any, professional relationships that I value more. I am pleased to say that he has agreed to enter into a consulting arrangement with the company to advise me and the Board of Directors.

Now it's time to look forward.

With Jeff leaving, the Board has asked me to resume the responsibilities of President and CEO in addition to my role as Chairman of the Board. I have agreed. I want to assure you that I have never felt

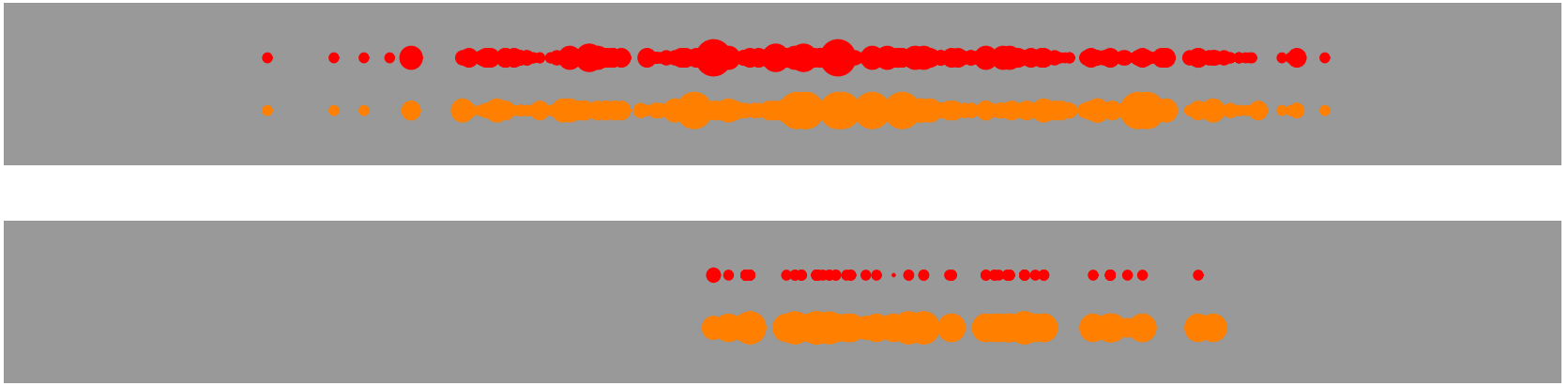
500,000 emails

5000 authors

1999-2002

Enron Topics

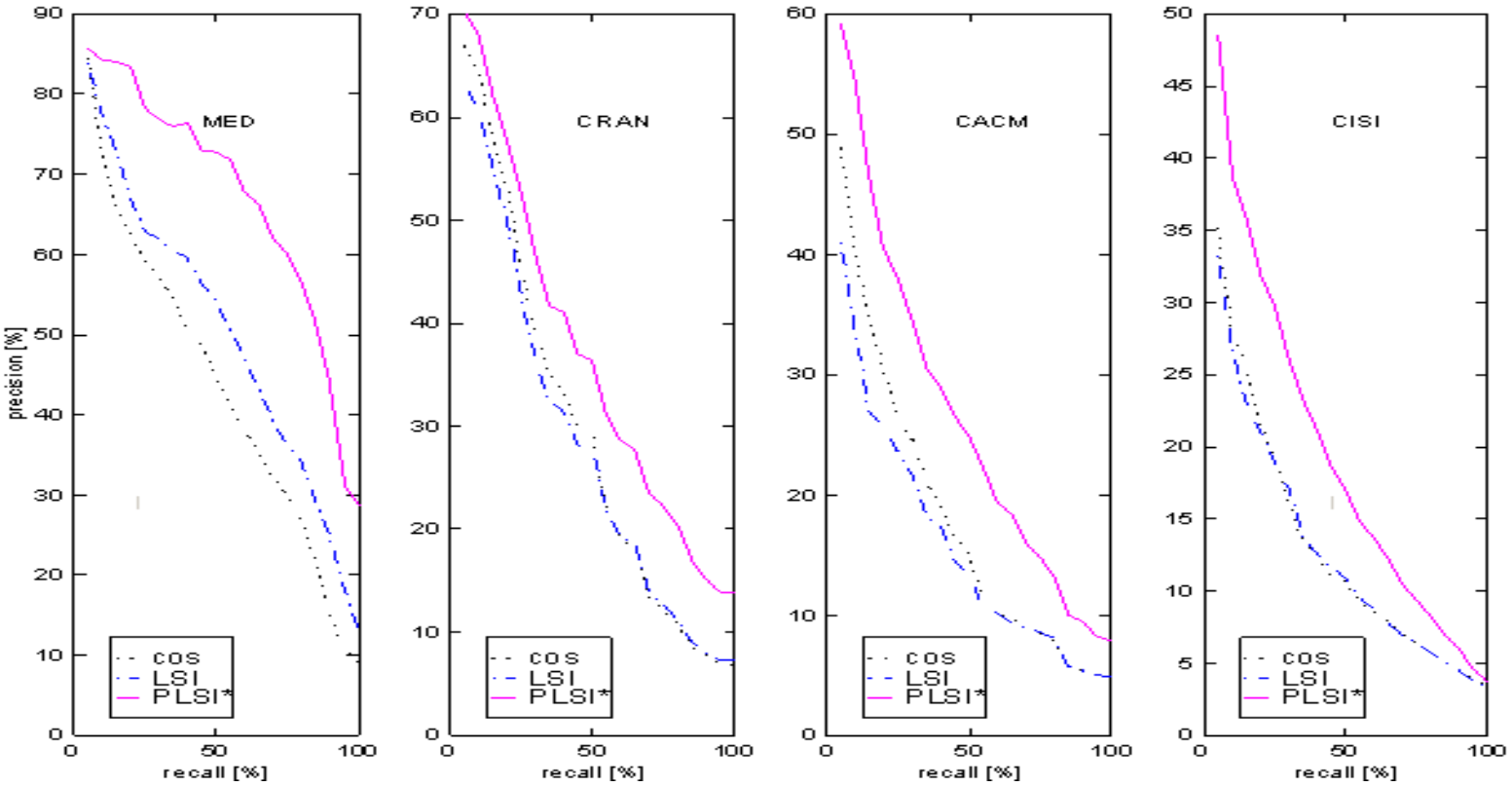
TEXANS	GOD	ENVIRONMENTAL	FERC	POWER	STATE
WIN	LIFE	AIR	MARKET	CALIFORNIA	PLAN
FOOTBALL	MAN	MTBE	ISO	ELECTRICITY	CALIFORNIA
FANTASY	PEOPLE	EMISSIONS	COMMISSION	UTILITIES	DAVIS
SPORTSLINE	CHRIST	CLEAN	ORDER	PRICES	RATE
PLAY	FAITH	EPA	FILING	MARKET	BANKRUPTCY
TEAM	LORD	PENDING	COMMENTS	PRICE	SOCAL
GAME	JESUS	SAFETY	PRICE	UTILITY	POWER
SPORTS	SPIRITUAL	WATER	CALIFORNIA	CUSTOMERS	BONDS
GAMES	VISIT	GASOLINE	FILED	ELECTRIC	MOU



May 22, 2000
Start of California energy crisis

TIMELINE

The performance of a retrieval system based on this model (PLSI) was found superior to that of both the vector space based similarity (cos) and a non-probabilistic latent semantic indexing (LSI) method. (We skip details here.)



From Th. Hofmann, 2000

Comparing PLSA and LSA

- LSA and PLSA perform dimensionality reduction
  - In LSA, by keeping only K singular values
  - In PLSA, by having K aspects

- Comparison to SVD
  - U Matrix related to P(d|z) (doc to aspect)
  - V Matrix related to P(z|w) (aspect to term)
  - E Matrix related to P(z) (aspect strength)

- The main difference is the way the approximation is done
  - PLSA generates a model (aspect model) and maximizes its predictive power
  - Selecting the proper value of K is heuristic in LSA

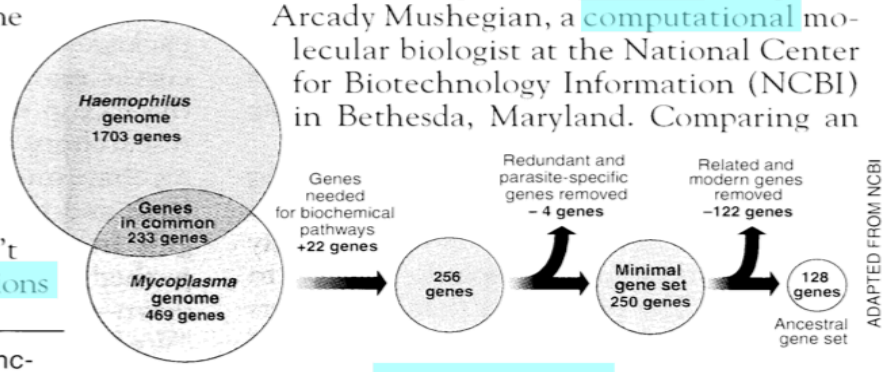
Latent Dirichlet Allocation

Intuition Behind LDA

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

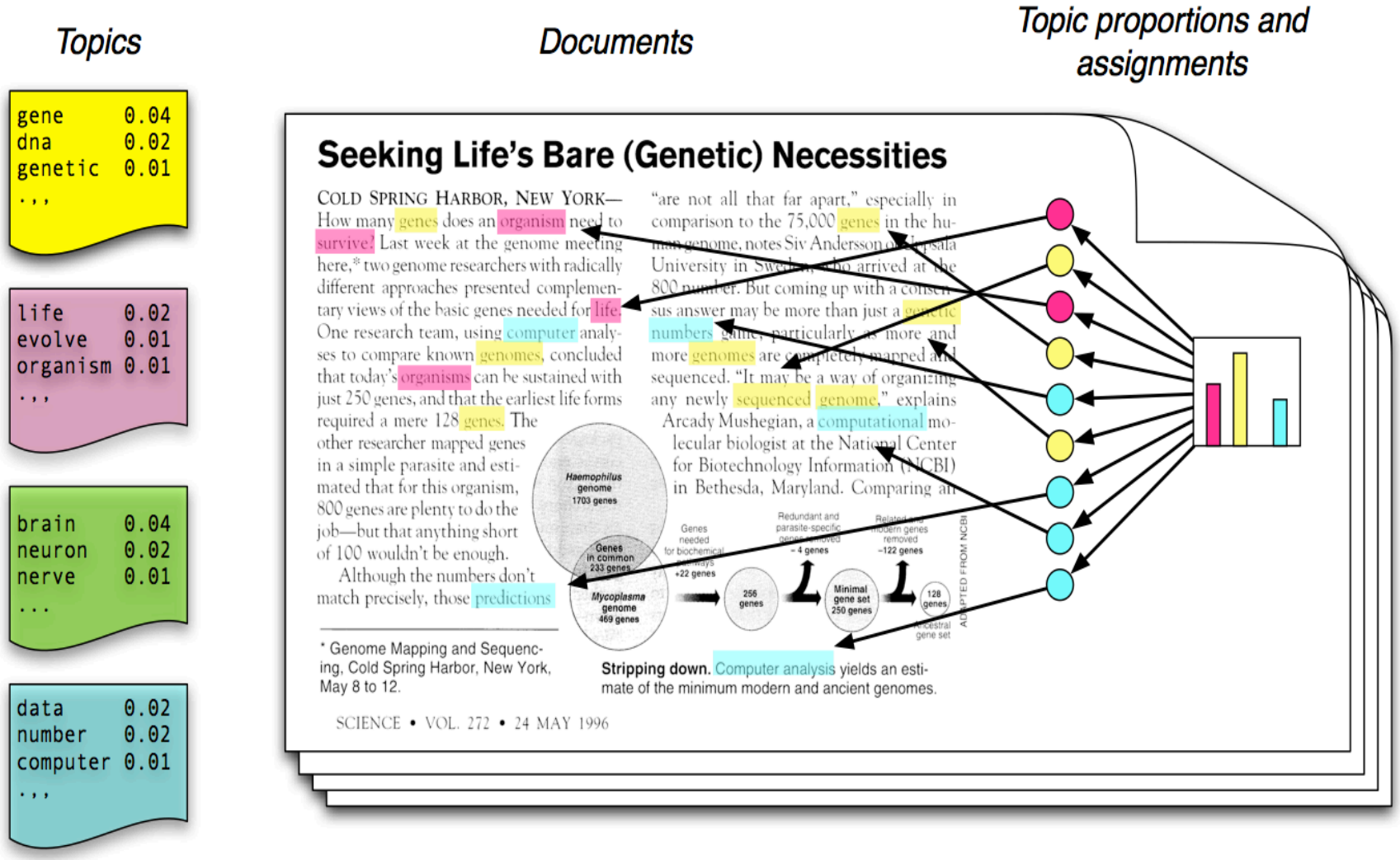
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Documents exhibit multiple topics

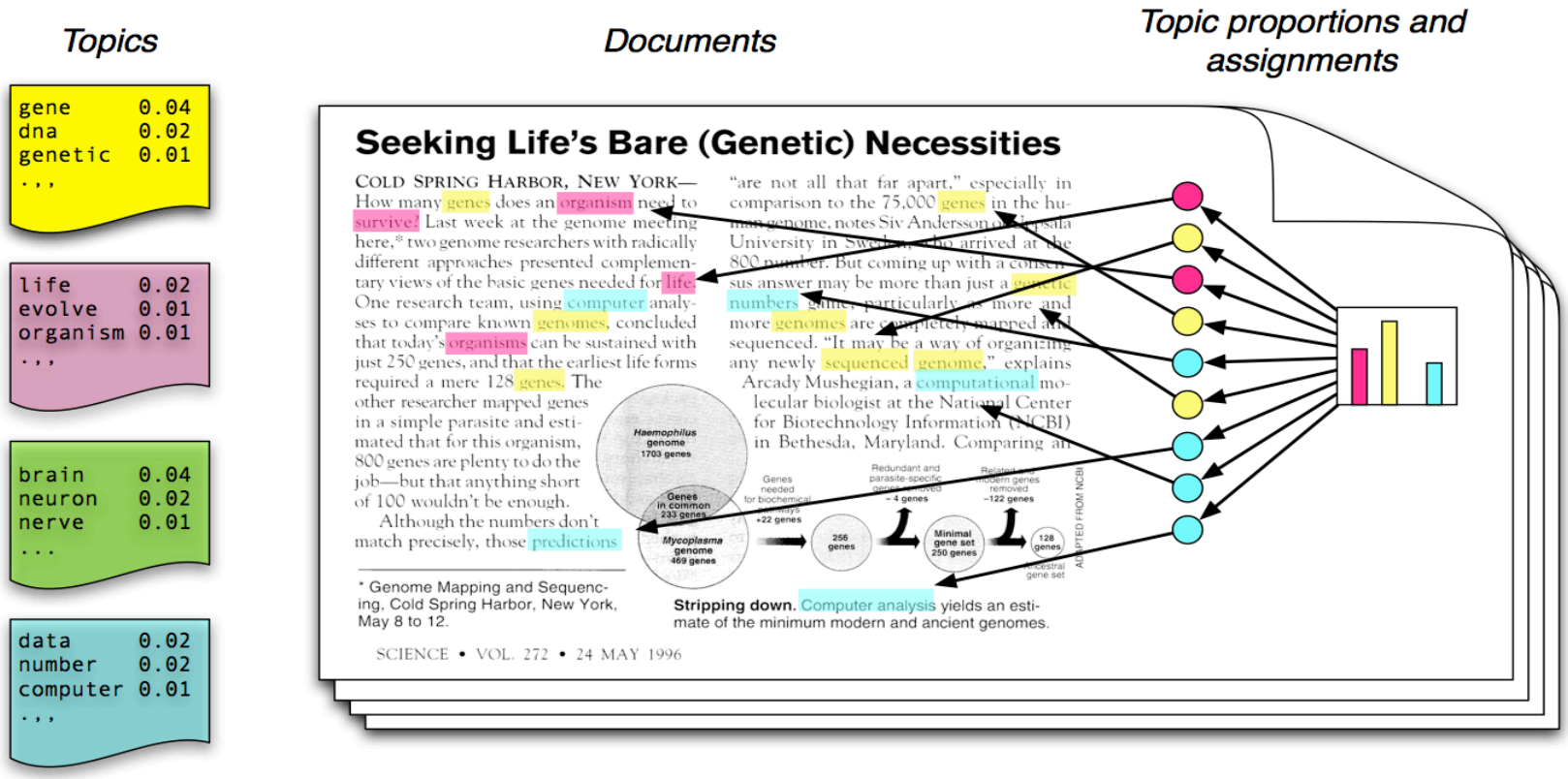
Slides by David Blei

Generative Model



Slides by David Blei

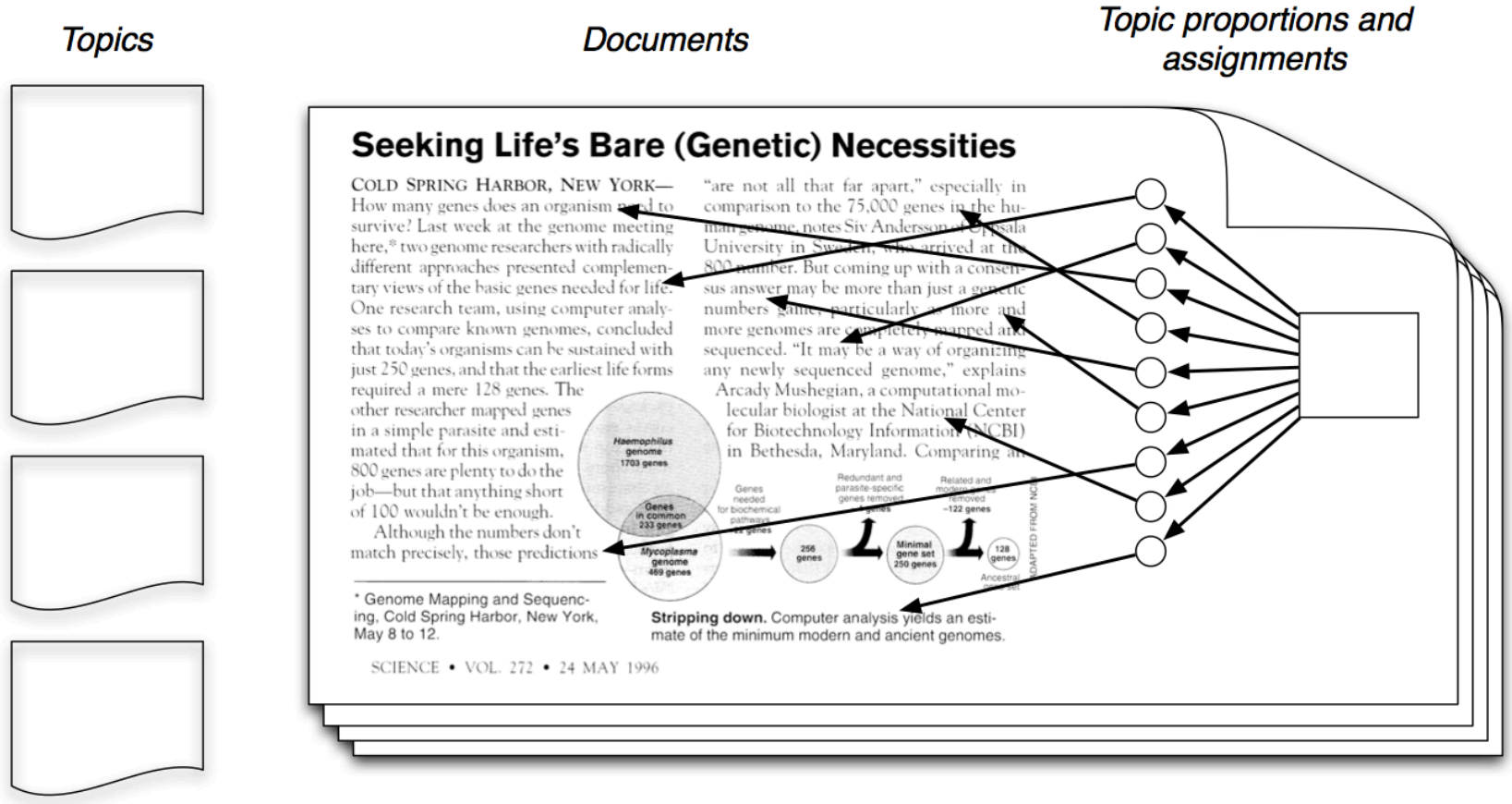
Generative Model



- Each document is a random mixture of topics found in the corpus
- Each word is drawn from one of those topics

Slides by David Blei

The posterior distribution



- We only observe the documents
- Our goal is to **infer** the underlying topic structure

Slides by David Blei

Probabilistic Model

- The observations are generated from a generative probabilistic process that includes hidden variables

- Infer the hidden structure using posterior inference.
  - What are the topics that describe this collection?

- Situate new data into the estimated model.
  - How does this query or new document fit into the estimated topic structure?

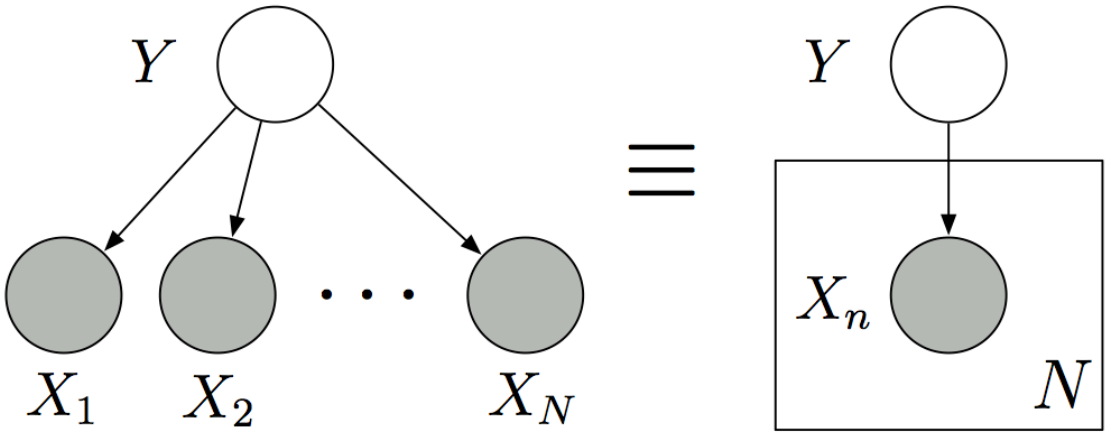
Slides by David Blei

Notation

- Word: 1..V
- Document: w=(w1,w2,…,wn) sequence of *N* words
- Corpus: D={w1,…,wm) collection of *M* documents

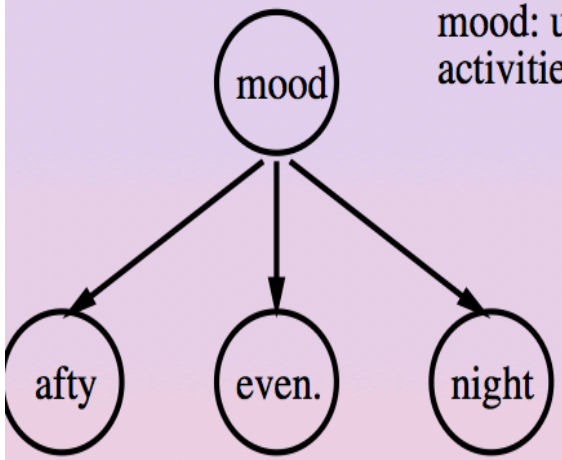
Slides by David Blei

Graphical Models



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

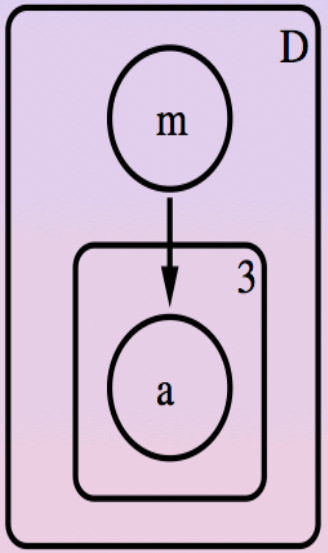
Slides by David Blei



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

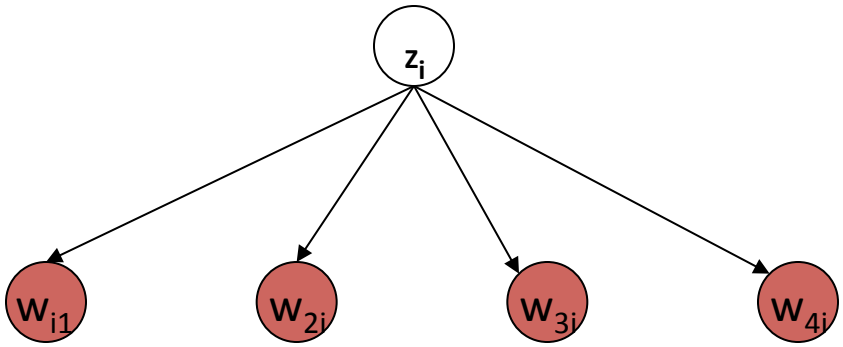


nodes
random variables

edges
dependencies

plates
repetitions

Mixture of Unigrams



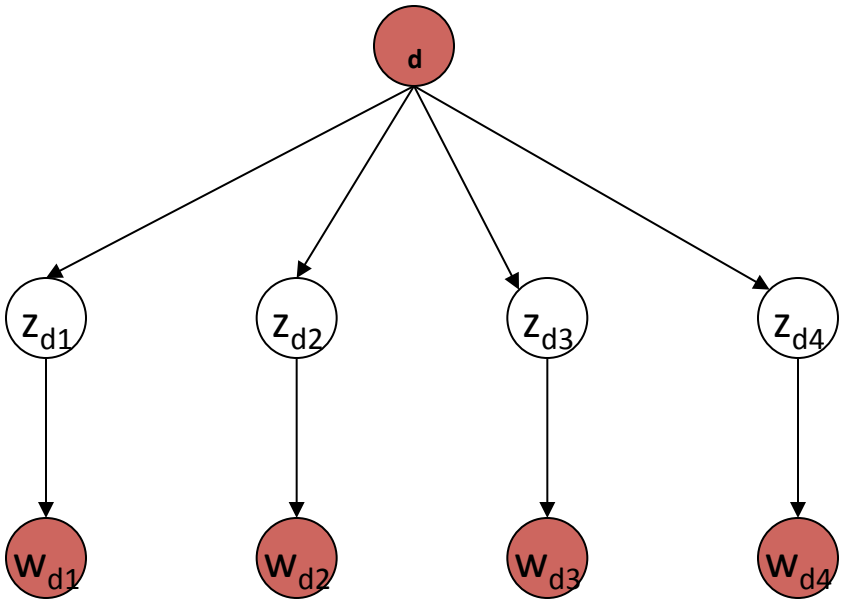
Mixture of Unigrams Model (this is just Naïve Bayes)

For each of M documents,

- Choose a topic z.
- Choose N words by drawing each one independently from a multinomial conditioned on z.

In the Mixture of Unigrams model, we can only have one topic per document!

The pLSI Model



Probabilistic Latent Semantic Indexing (pLSI) Model

For each word of document d in the training set,

- Choose a topic z according to a multinomial conditioned on the index d.
- Generate the word by drawing from a multinomial conditioned on z.

In pLSI, documents can have multiple topics.

Motivations for LDA

- In pLSI, the observed variable d is an index into some training set. There is no natural way for the model to handle previously unseen documents.
- The number of parameters for pLSI grows linearly with M (the number of documents in the training set).
- We would like to be Bayesian about our topic mixture proportions.

Dirichlet Distributions

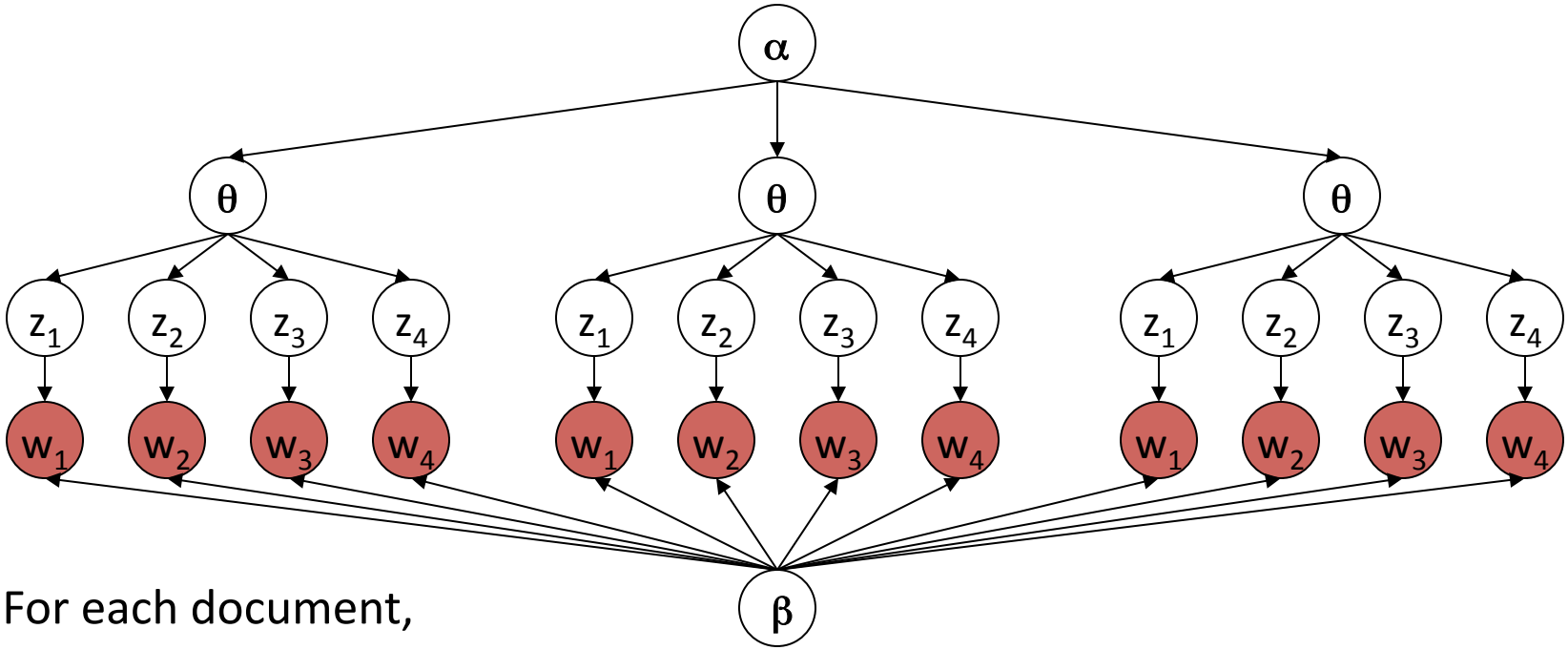
- In the LDA model, we would like to say that the *topic mixture proportions* for each document are drawn from some distribution.
- So, we want to put a distribution on multinomials. That is, k-tuples of non-negative numbers that sum to one.
- The space is of all of these multinomials has a nice geometric interpretation as a (k-1)-*simplex*, which is just a generalization of a triangle to (k-1) dimensions.
- Criteria for selecting our prior:
  - It needs to be defined for a (k-1)-simplex.
  - Algebraically speaking, we would like it to play nice with the multinomial distribution.

Dirichlet Distributions

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

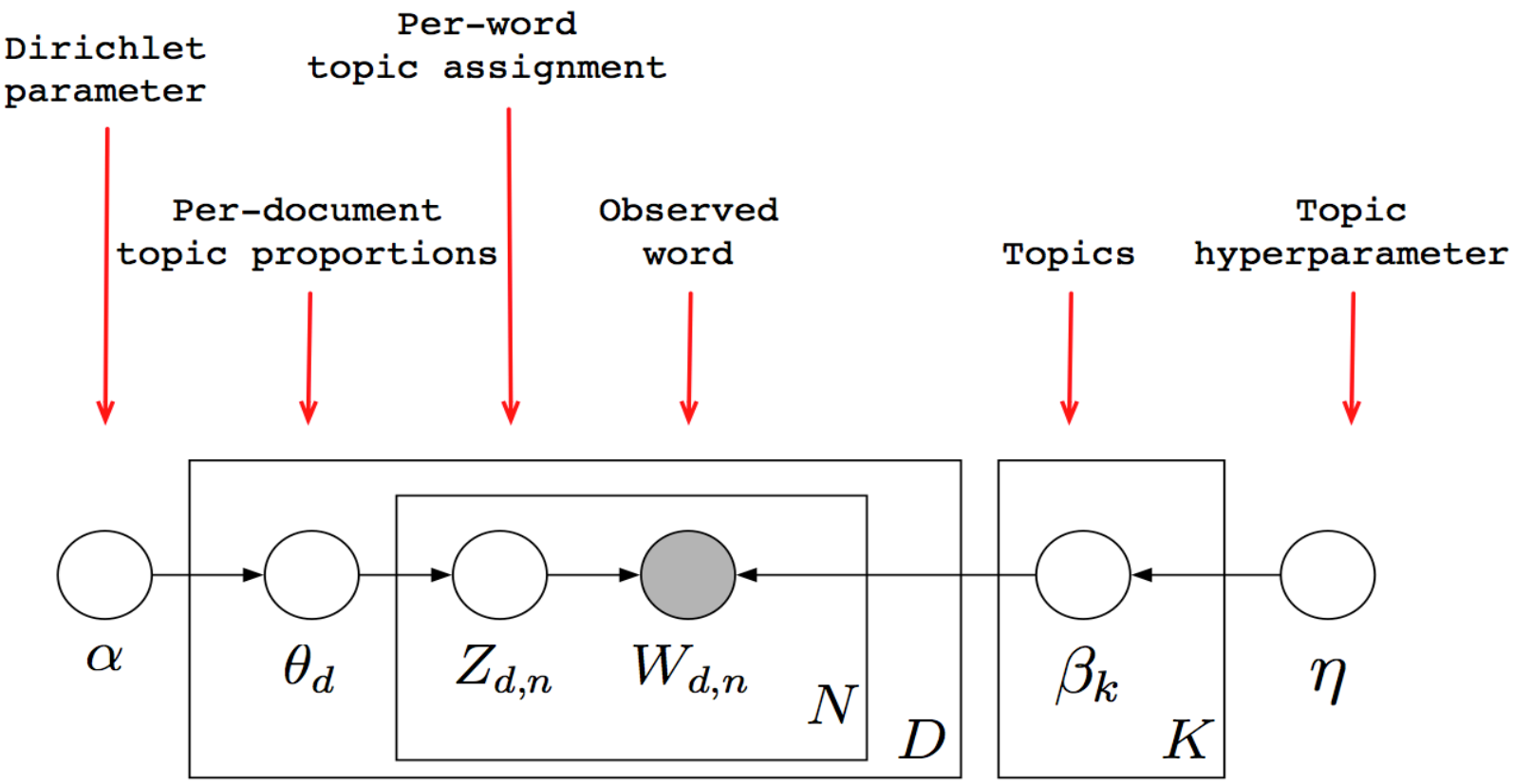
- Useful Facts:
  - This distribution is defined over a (k-1)-simplex. That is, it takes k non-negative arguments which sum to one. Consequently it is a natural distribution to use over multinomial distributions.
  - In fact, the Dirichlet distribution is the conjugate prior to the multinomial distribution. (This means that if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet!)
  - The Dirichlet parameter α_i can be thought of as a prior count of the i^{th} class.

The LDA Model



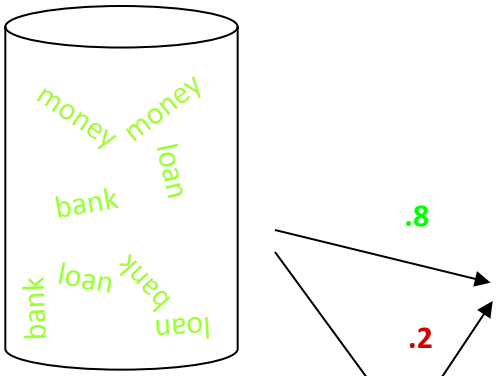
- For each document,
- Choose θ~Dirichlet(α)
- For each of the N words wn:
  - Choose a topic z_n» Multinomial(θ)
  - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

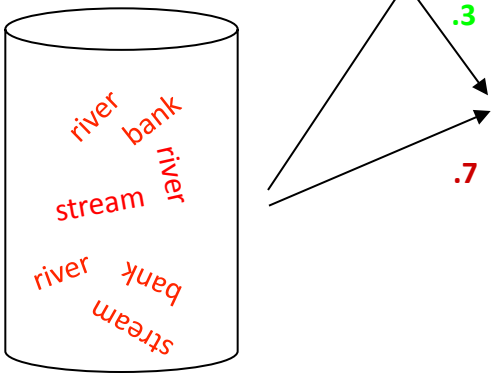
Latent Dirichlet Allocation

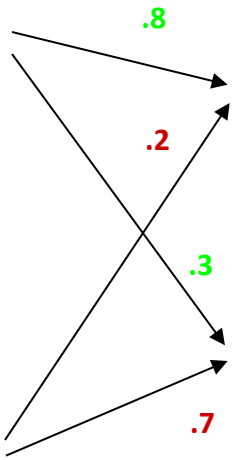


Slides by David Blei

Example







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

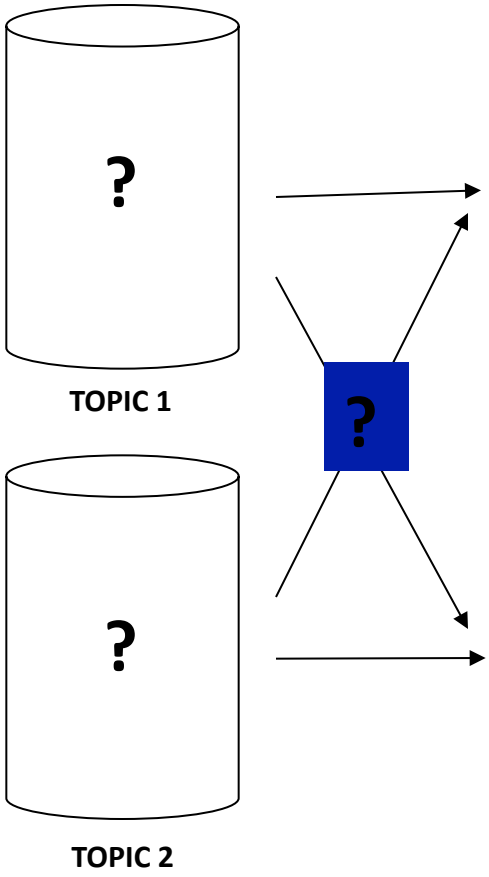
Mixture weights

Bayesian approach: use priors

Mixture weights ~ Dirichlet(α)

Mixture components ~ Dirichlet(β)

Inverting ("fitting") the model



DOCUMENT 1: money? bank? bank? loan? river? stream? bank? money? river? bank? money? bank? loan? money? stream? bank? money? bank? bank? loan? river? stream? bank? money? river? bank? money? bank? loan? bank? money? stream?

DOCUMENT 2: river? stream? bank? stream? bank? money? loan? river? stream? loan? bank? river? bank? bank? stream? river? loan? bank? stream? bank? money? loan? river? stream? bank? stream? bank? money? river? stream? loan? bank? river? bank? money? bank? stream? river? bank? stream? bank? money?

Mixture components

Mixture weights

Illustrative Example

- S1: I like to eat broccoli and bananas.
- S2: I ate a banana and spinach smoothie for breakfast.
- S3: Chinchillas and kittens are cute.
- S4: My sister adopted a kitten yesterday.
- S5: Look at this cute hamster munching on a piece of broccoli.

What is LDA doing?

What is LDA doing?

- A way to automatically discover the topics contained in the sentences.

Can you discover the topics on your own?

Illustrative Example

- S1: I like to eat broccoli and bananas.
- S2: I ate a banana and spinach smoothie for breakfast.
- S3: Chinchillas and kittens are cute.
- S4: My sister adopted a kitten yesterday.
- S5: Look at this cute hamster munching on a piece of broccoli.

What is LDA doing?

- Given the sentences and 2 topics, LDA might produce something like

  Sentences 1 and **2**: 100% Topic A

  Sentences 3 and **4**: 100% Topic B

  Sentence 5: 60% Topic A, 40% Topic B

  Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching

  Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster

LDA Model

- Decide on the number of words N the document will have (say, according to a Poisson distribution).

- Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics).

For example, assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

LDA Model

- Generate each word w_i in the document by:
  - first picking a topic (according to the multinomial distribution that you sampled above)

    for example, you might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability.

LDA Model

- Generate each word w_i in the document by:
  - first picking a topic (according to the multinomial distribution that you sampled above)
  - use the topic to generate the word itself (according to the topic's multinomial distribution)

for example, if we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on.

Example Document Generation

- Pick 5 to be the number of words in D.
- Decide that D will be 1/2 about food and 1/2 about cute animals.
- Pick the first word to come from the food topic, which then gives you the word "**broccoli**".
- Pick the second word to come from the cute animals topic, which gives you "**panda**".
- Pick the third word to come from the cute animals topic, giving you "**adorable**".
- Pick the fourth word to come from the food topic, giving you "**cherries**".
- Pick the fifth word to come from the food topic, giving you "**eating**".

- Given a set of documents and fixed number of K topics to discover, use LDA to learn the topic representation of each document and the words associated with each topic.

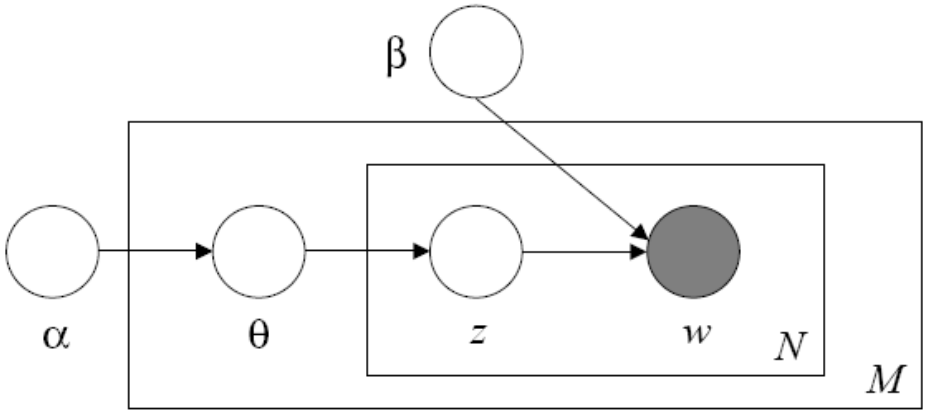
- Given a set of documents and fixed number of K topics to discover, use LDA to learn the topic representation of each document and the words associated with each topic.
- Go through each document d, and randomly assign each word in the document to one of the K topics.

- Go through each word w in d and calculate
  - p(topic t | document d) the proportion of words in document d that are currently assigned to topic t
  - p(word w | topic t) the proportion of assignments to topic t over all documents that come from this word w.

- Reassign w a new topic, where we choose topic t with probability

  p(topic t | document d) * p(word w | topic t)
- According to this generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current word's topic with this probability
- Assume that all topic assignments except for the current word in question are correct, and then update the assignment of the current word using our model of how documents are generated.

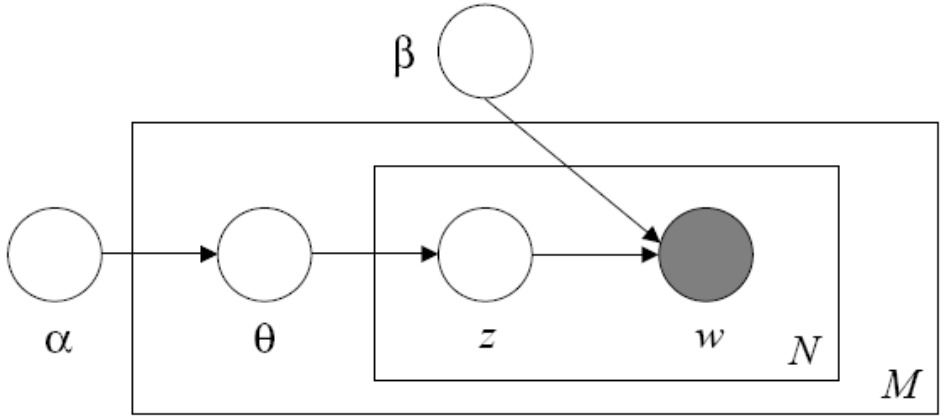
The LDA Model



For each document,

- Choose θ» Dirichlet(α)
- For each of the N words w_n:
  - Choose a topic z_n» Multinomial(θ)
  - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

Inference



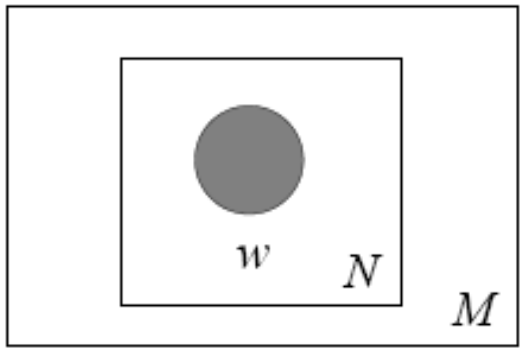
•The inference problem in LDA is to compute the posterior of the hidden variables given a document and corpus parameters α and β. That is, compute $p(\theta,z|w,\alpha,\beta)$.

•Unfortunately, exact inference is intractable, so we turn to alternatives…

Relationship with Other Latent Variable Models

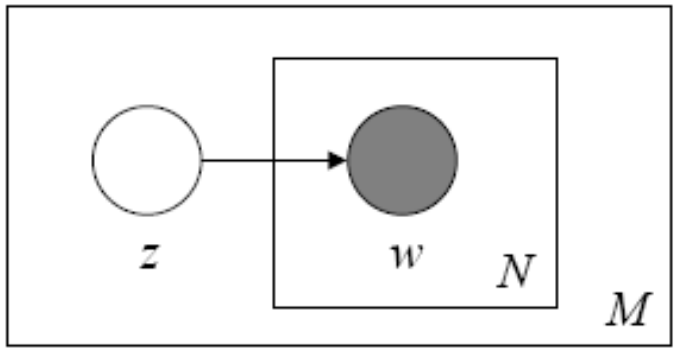
- Unigram model

$$p(w) = \prod_{n=1}^{N} p(w_n)$$



- Mixture of unigrams
  - Each document is generated by first choosing a topic z and then generating N words independently form conditional multinomial
  - k-1 parameters

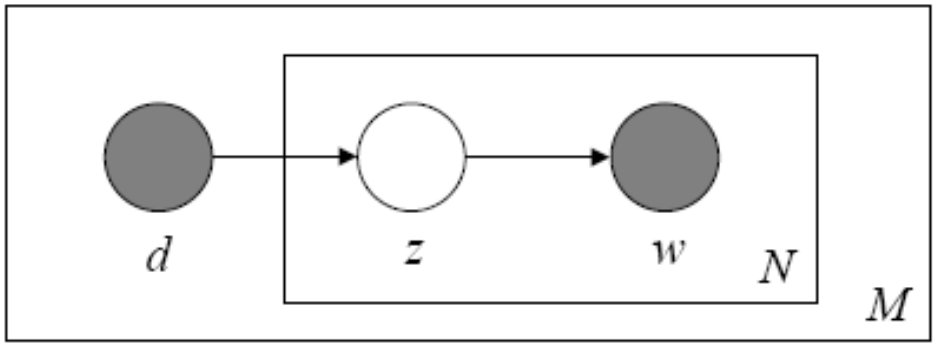
$$p(w) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z)$$



Relationship with Other Latent Variable Models

- Probabilistic Latent Semantic Analysis
  - Attempt to relax the simplifying assumption made in the mixture of unigrams models
  - In a sense, it does capture the possibility that a document may contain multiple topics
  - kv+kM parameters and linear growth in M

$$p(d, w_n) = p(d) \sum_z p(w_n \mid z) p(z \mid d)$$



Relationship with Other Latent Variable Models

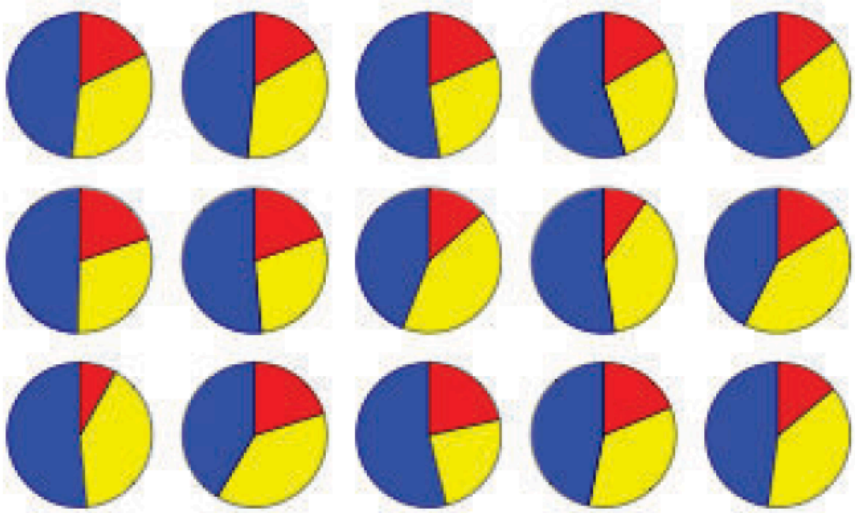
- **Unigram model** find a single point on the word simplex and posits that all word in the corpus come from the corresponding distribution.
- **Mixture of unigram model** posits that for each documents, one of the k points on the word simplex is chosen randomly and all the words of the document are drawn from the distribution
- **pLSA model** posits that each word of a training documents comes from a randomly chosen topic. The topics are themselves drawn from a document-specific distribution over topics.
- **LDA** posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter

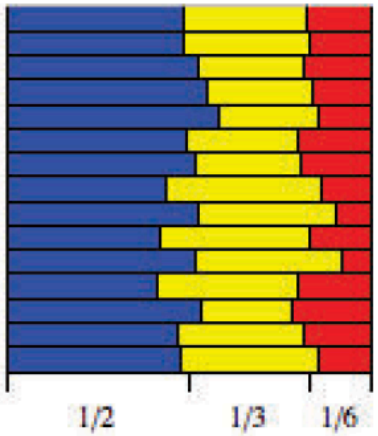
Latent Dirichlet Allocation

LDA assumes the following generative process:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of N words w_n:
   (a) Choose topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose word $w_n \sim$ from $P(w_n|z_n, \beta)$

Dirichlet Distribution





$Dir(\alpha)$; $\alpha = (3, 2, 1)$

Cut strings (each of initial length 1.0) into K pieces with different lengths

(from Wikipedia)

Dirichlet Distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

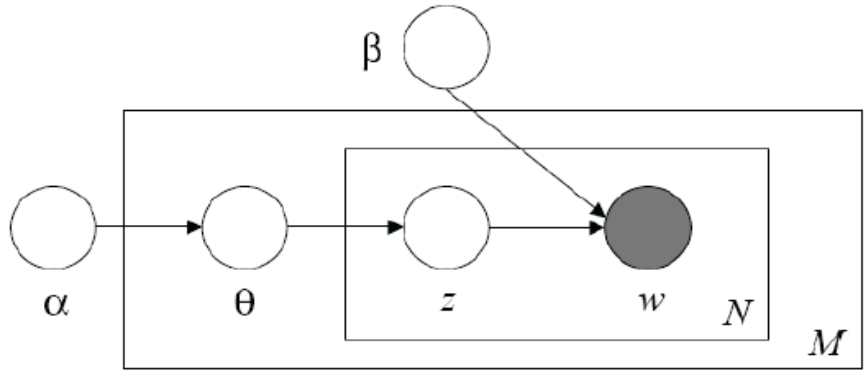
- The Dirichlet is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ.
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

Dirichlet Distribution

From a collection of documents, **infer**

- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions θ_d
- Per-corpus topic distributions β_k

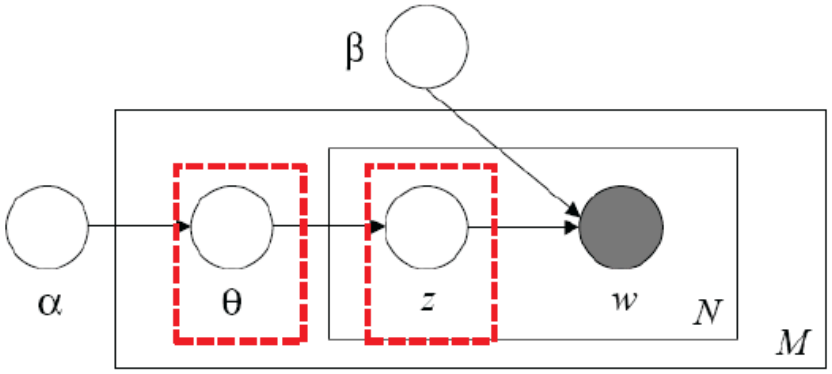
Inference



"Arts" "Budgets" "Children" "Education"

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.

Inference



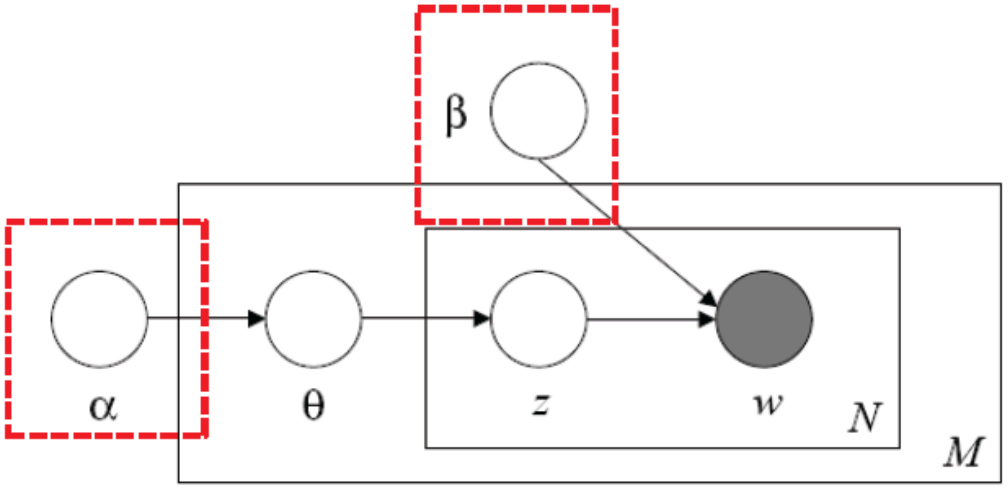
θ: Per-document topic proportions

z: Per-word topic assignment

"Arts" "Budgets" "Children" "Education"

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.

Parameter Estimation



α : Dirichlet parameter

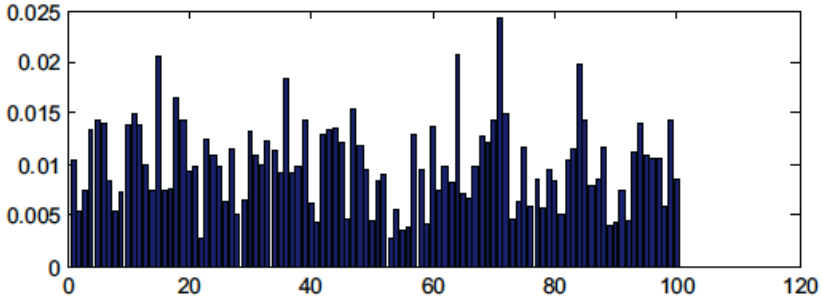
β : Topics

- Try to estimate parameters (α, β), given corpus {w}.
- EM algorithm:

Parameter estimation

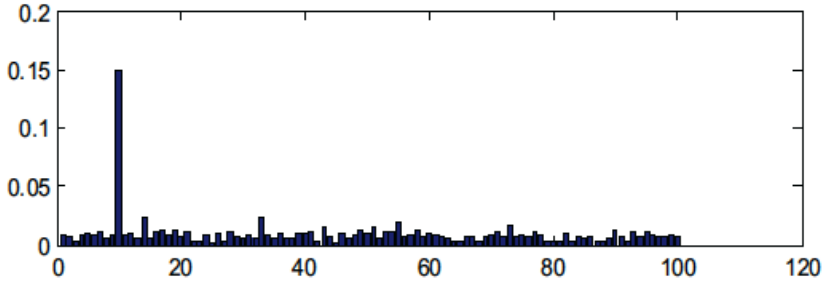
- α controls proportion distribution of topics in one document.

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ equally large



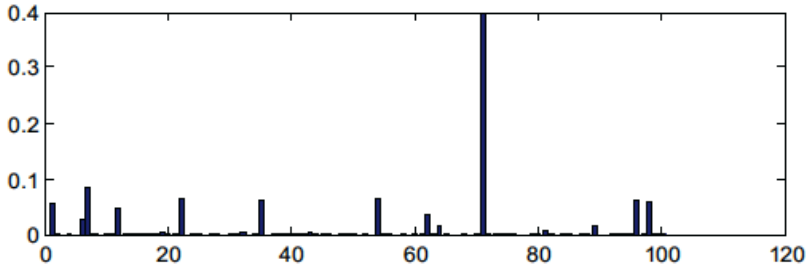
Topics are almost equally likely

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ equal, but α_{10} is larger



10th topic is more likely to appear

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ are equally small



Topics distribution is sparse (few topics in one document), with one random peak

- β is the probability matrix of topics and words

Inference example

"Arts" "Budgets" "Children" "Education"

q(z|w)>0.9

Bag-of-words assumption

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.

TOPIC MODELS
FOR TEXT CLASSIFICATION

LDA/pLSA for Text Classification

- Topic models are easy to incorporate into text classification:
  1. Train a topic model using a big corpus
  2. Decode the topic model (find best topic/cluster for each word) on a training set
  3. Train classifier using the topic/cluster as a feature
  4. On a test document, first decode the topic model, then make a prediction classifier

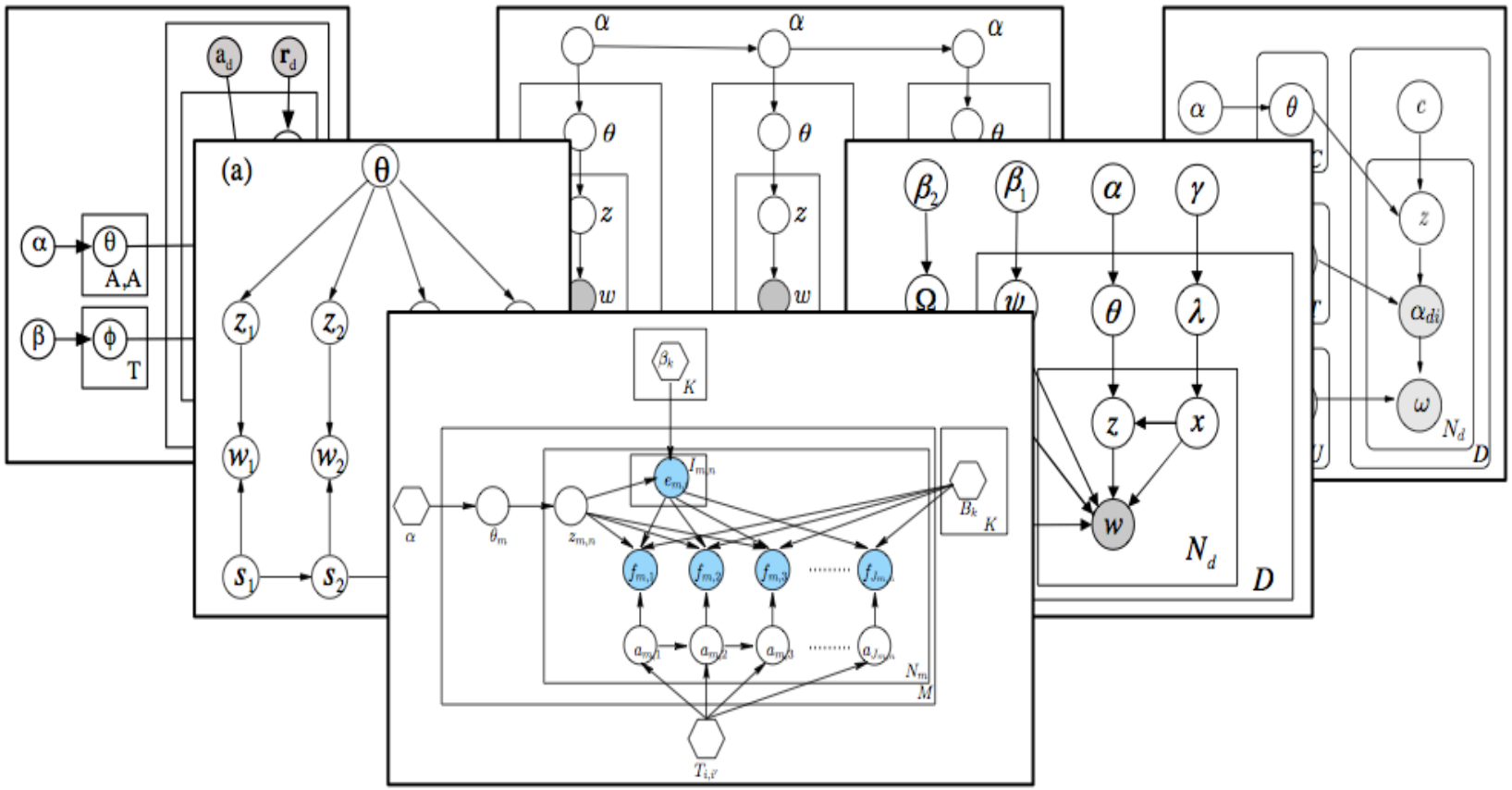
Why use a topic model for classification?

- Topic models help handle polysemy and synonymy
  - The count for a topic in a document can be much more informative than the count of individual words belonging to that topic.

- Topic models help combat data sparsity
  - You can control the number of topics
  - At a reasonable choice for this number, you'll observe the topics many times in training data

    (unlike individual words, which may be very sparse)

Other Topic Models



Slides by David Blei

EXAMPLE

LDA Topics from Emails

- **Trig/Family/Inspiration:** family, web, mail, god, son, from, congratulations, children, life, child, down, trig, baby, birth, love, you, syndrome, very, special, bless, old, husband, years, thank, best
- **Wildlife/BP Corrosion:** game, fish, moose, wildlife, hunting, bears, polar, bear, subsistence, management, area, board, hunt, wolves, control, department, year, use, wolf, habitat, hunters, caribou, program, denby
- **Energy/Fuel/Oil/Mining:** energy, fuel, costs, oil, alaskans, prices, cost, nome, now, high, being, home, public, power, mine, crisis, price, resource, need, community, fairbanks, rebate, use, mining, villages
- **Gas:** gas, oil, pipeline, agia, project, natural, north, producers, companies, tax, company, energy, development, slope, production, resources, line, gasline, transcanada, said, billion, plan, administration, million, industry, …
- **Education/Waste:** school, waste, education, students, schools, million, read, email, market, policy, student, year, high, news, states, program, first, report, business, management, bulletin, information, reports, 2008, quarter
- **Presidential Campaign/Elections:** mail, web, from, thank, you, box, mccain, sarah, very, good, great, john, hope, president, sincerely, wasilla, work, keep, make, add, family, republican, support, doing, p.o

Hello Governor Palin, Our **family** wanted to congratulate **you** and your **family** on the **birth** of your **son**, **Trig**. Our fourth **child**, Daniel, was **born** with **Down Syndrome**, and we can't imagine our **family** without him. Recently, I met a mom with a 34-year-old **daughter** with DS and she said it best: "Don't **you** feel like you've been chosen to be a member of a **very special** club?" **God** bless your **family**, what a **beautiful** example of **love** you are to all who see you! the Paul & Tricia Pietig **family**, Des Moines, Iowa

In this email, 99% of the words feel into the Trig/Family/Inspiration category

We understand that you have been discussed as a possible choice for the **Vice Presidency**.

As **people** who **support** the democratic process and care about protecting our **wildlife** for future generations, we want **you** to know that we don't believe **people** in our states would vote for **you** for any office if they knew your record on these issues.

It is troubling that **you** are **now** working to deny more than 50,000 Alaskans a vote on **aerial** killing of **wolves** and **bears** with legislation now **being** considered in the Alaska legislature.

In this email, 10% of the words feel into the Presidential Campaign/Election category (in red) and 90% of the words feel into into the Wildlife/BP Corrosion category (in green)

TOPIC MODELING TOOLKIT

Mallet

- **Command line scripts:**
  – *bin/mallet [command] --[option] [value]*

- **Direct Java API**
  – *http://mallet.cs.umass.edu/api*

- **Download**
  - http://mallet.cs.umass.edu/topics.php

Topic Modeling with Mallet

- **Importing Documents:**
  bin/mallet train-topics --input topic-input.mallet \
  --num-topics 100 --output-state topic-state.gz

  --*input* [File] specifies the collection
  --*num-topics* [Number] is the number of topics to use

Topic Modeling with Mallet

- **Model Output:**

  --*output-model* [File] a file to write Mallet topic trainer

  --*output-state* [File] a compressed file containing the word in the corpus with their topic assignments

  --*output-doc-topics* [File] writes the topic composition of the documents

  --*output-topic-keys* [File] consist of top "*k*" words for each topic

For More Details Check

- Mallet's tutorial

http://mallet.cs.umass.edu/mallet-tutorial.pdf