Statistical Topic Models for Studying Collaborative Processes

Hanna M. Wallach

University of Massachusetts Amherst

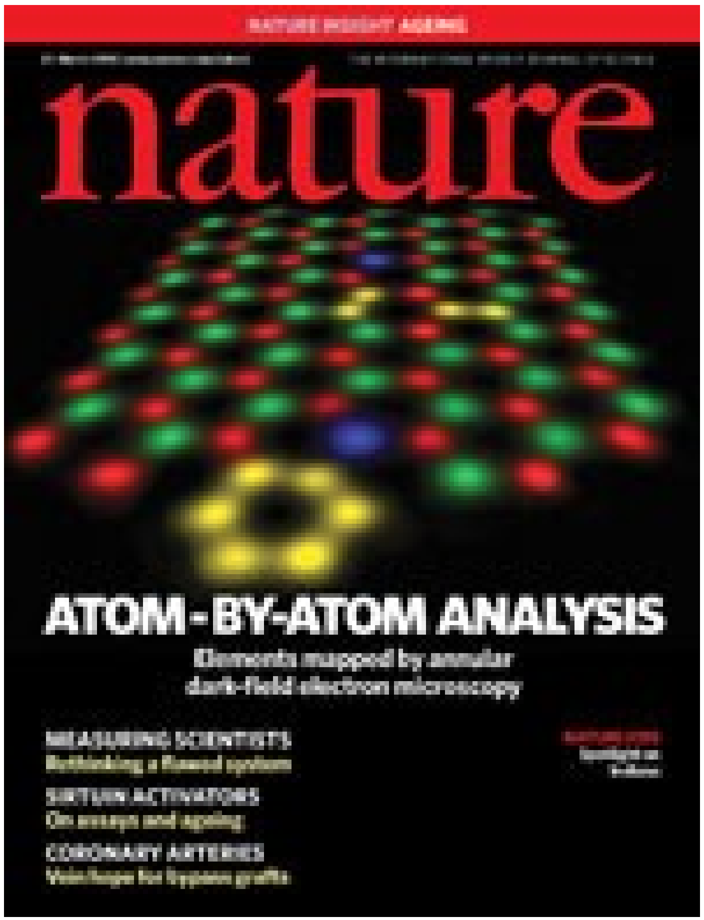
wallach@cs.umass.edu

Products of Collaboration

"Scientific information is both the basic raw material for, and one of the principal products of, scientific research [...] Scientists find out what other scientists are accomplishing through [...] journals, books, abstracts and indexes, bibliographies, reviews."

— NSF Brochure, 1962

Collaborate to Study Collaboration



“There needs to be a greater focus on what these [science interaction] data mean [...] This requires the input of social scientists, rather than just those more traditionally involved in data capture, such as computer scientists.”

— Julia Lane, NSF, 24 March 2010

Approach: Statistical Models

- Modeling challenges:
  - Aggregating and representing large data sets
  - Handling data from sources with disparate emphases
  - Reasoning under uncertain information
  - Performing efficient inference
- Bayesian latent (hidden) variable models:
  - Powerful and flexible [Wallach et al. & Adams et al., AISTATS '10]
  - This talk: statistical topic models

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Some current and future projects:
  - Analyzing free software development communities
  - Predicting when to declassify documents

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Some current and future projects:
  - Analyzing free software development communities
  - Predicting when to declassify documents

Statistical Topic Modeling

- Three fundamental assumptions:
  - Documents have latent semantic structure (“topics”)
  - Can infer topics from word–document co-occurrences
  - Words are related to topics, topics to documents
- Given a data set, the goal is to
  - Learn the composition of the topics for that data set
  - Learn which topics are used in each document

Why Topic Models?

kriging
covariance
mean
estimate
weight
random
mse
matrix
conditional
point

vs.

gaussian
regression
covariance
prediction
function
bayesian
process
prior
distribution
matrix

Topics and Words

probability

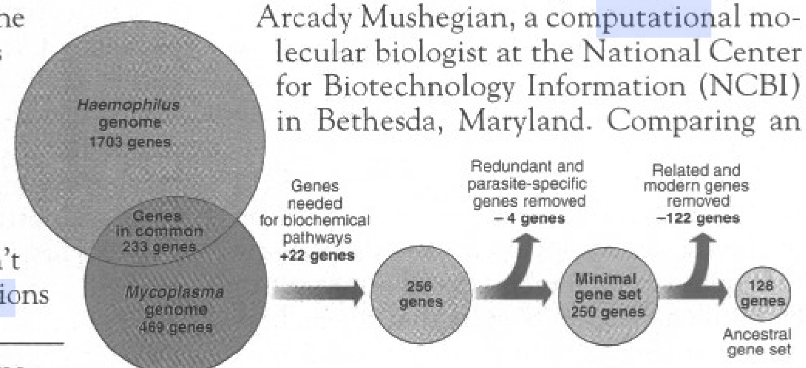
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
...	...	...	...

Documents and Topics

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.

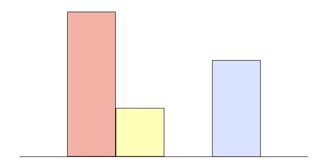
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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.

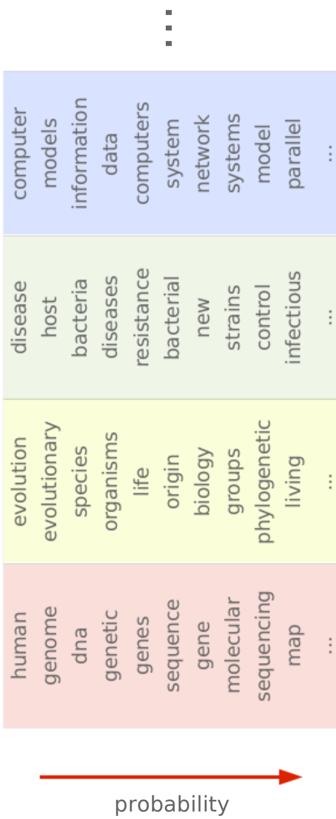
SCIENCE • VOL. 272 • 24 MAY 1996



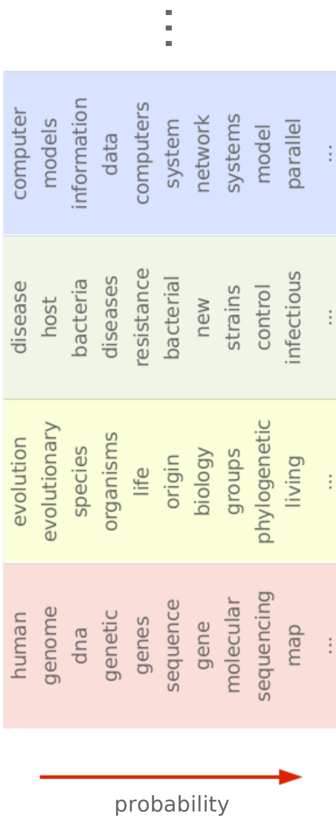
Generative Statistical Modeling

- Assume data was generated by a probabilistic model:
  - Model may have hidden structure (latent variables)
  - Model defines a joint distribution over all variables
  - Model parameters are unknown
- Infer hidden structure and model parameters from data
- Situate new data in estimated model

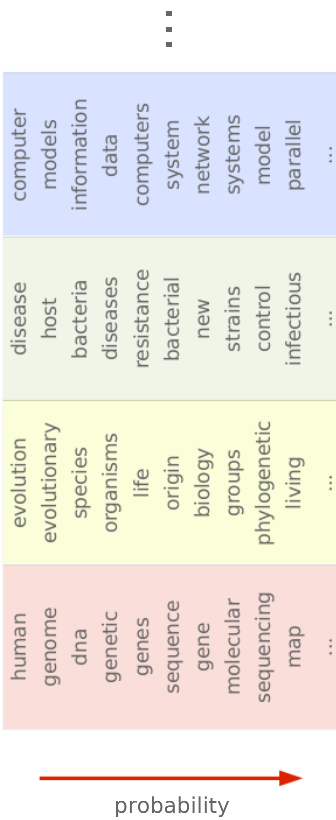
Generative Process

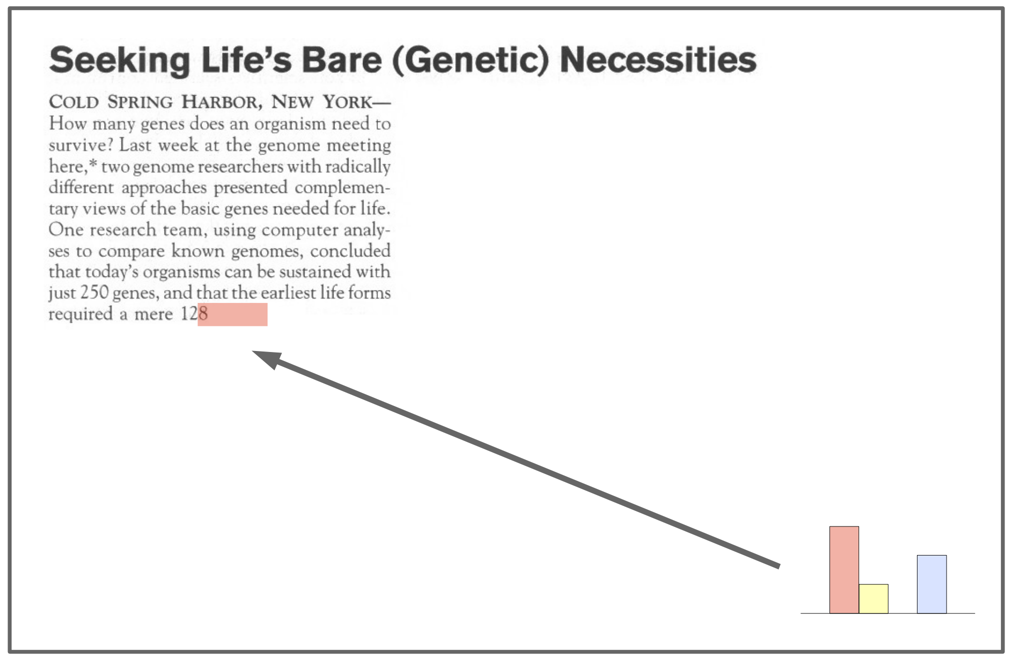


Choose a Distribution Over Topics

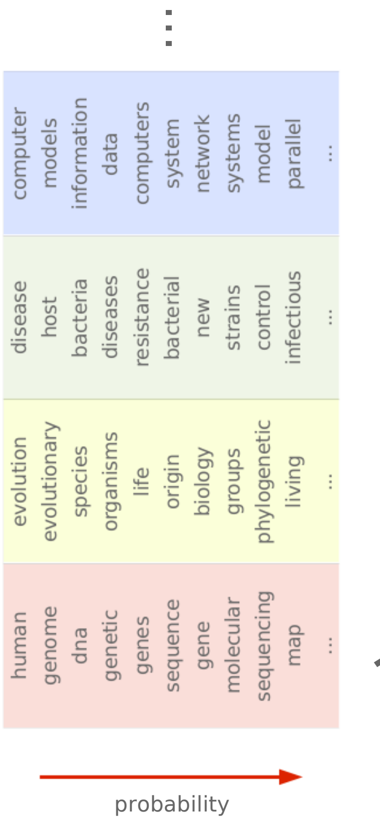


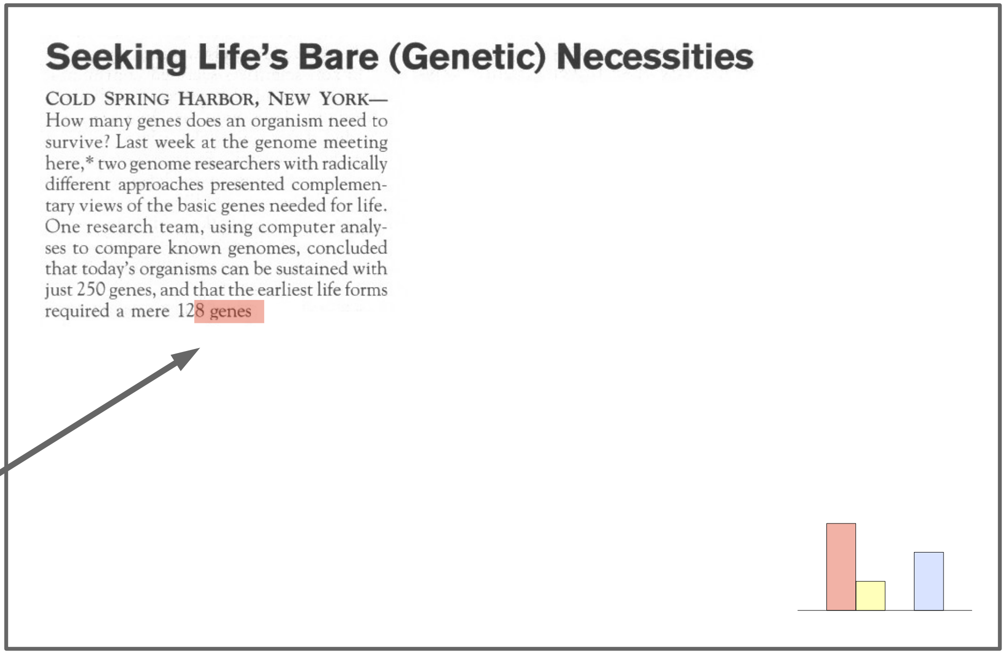
Choose a Topic





Choose a Word

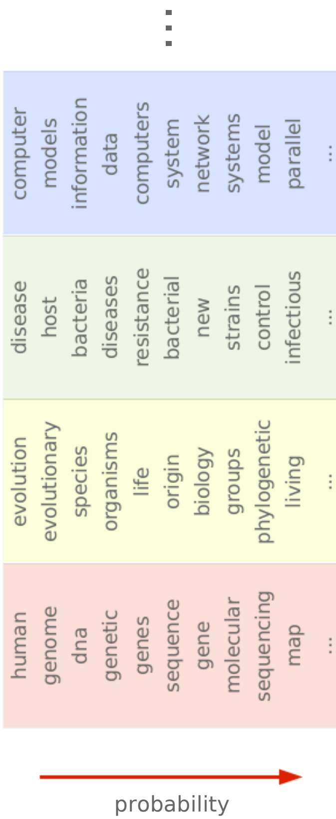


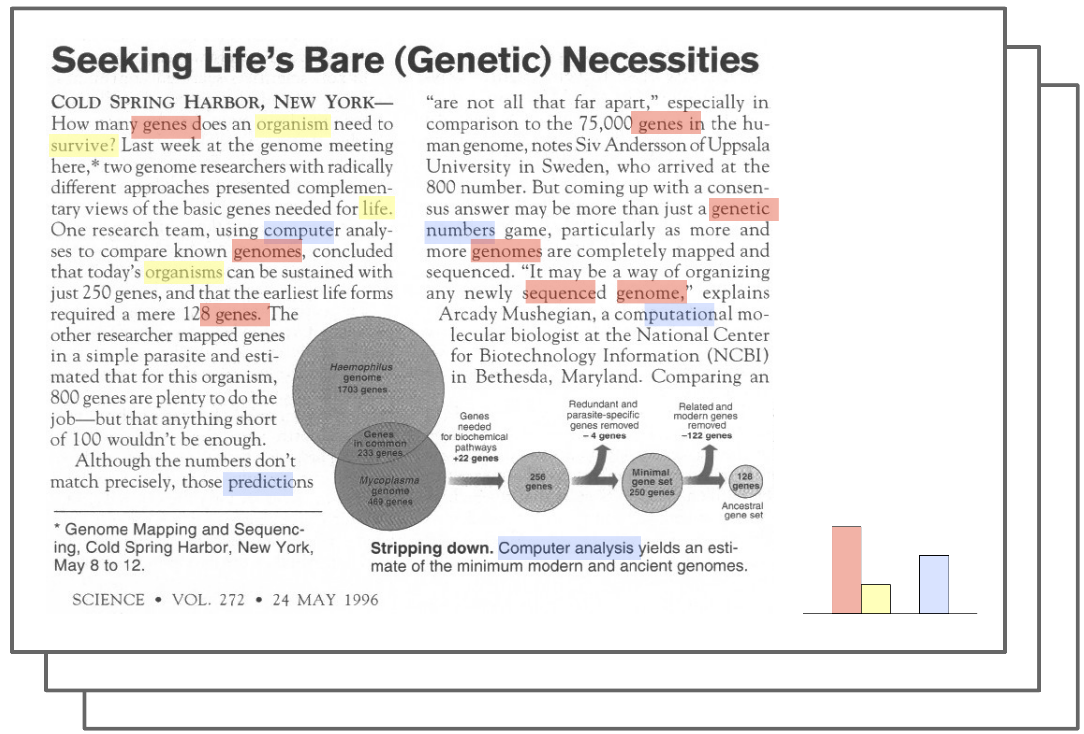


Seeking Life's Bare (Genetic) Necessities

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

… And So On





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.

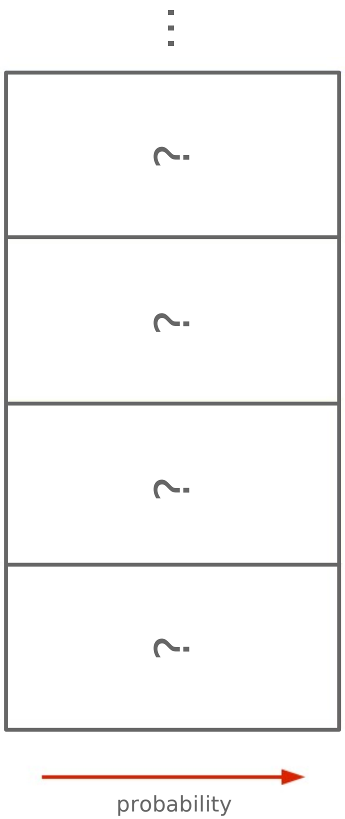
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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

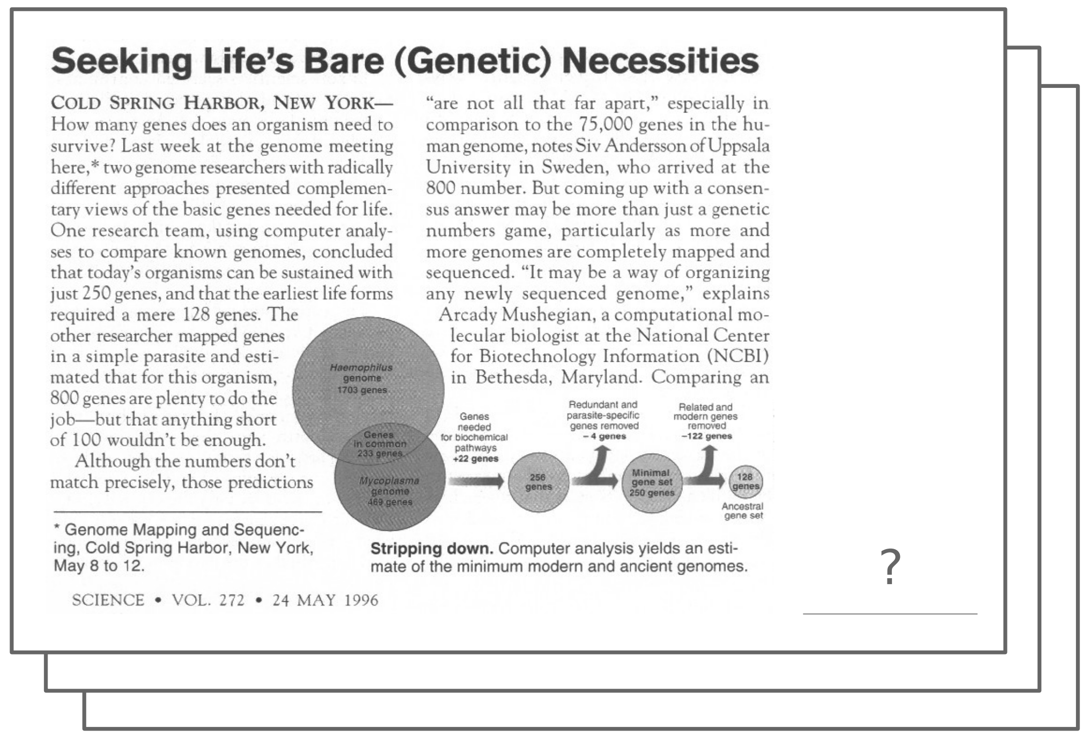
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Real Data: Statistical Inference





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.

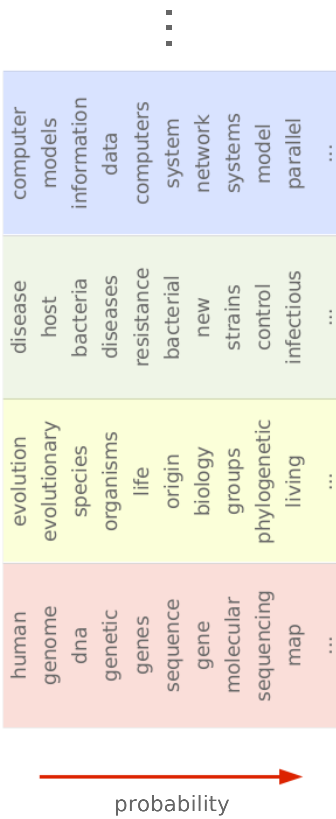
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

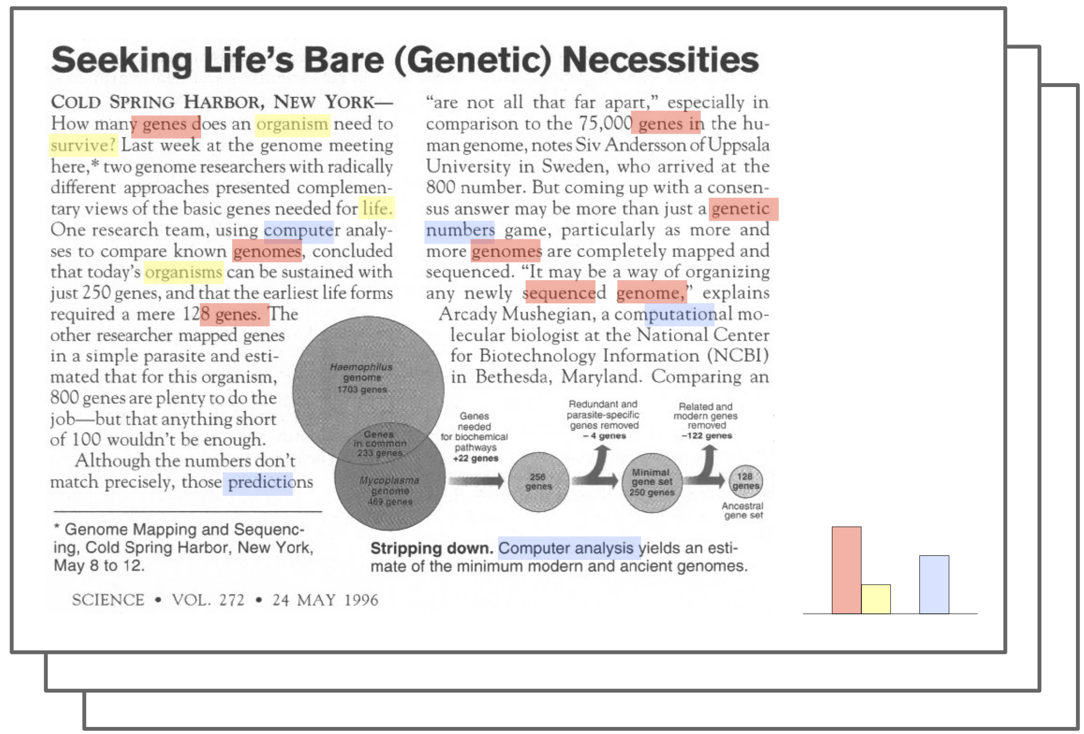
SCIENCE • VOL. 272 • 24 MAY 1996

Inference: Gibbs Sampling

- Randomly guess which topic “generated” each word:
- Given a set of guesses, can estimate probabilities
  - Initially the probabilities will be random
- Repeatedly refine the guess for each word:
  - Probability of guessing topic t for word w in document d is proportional to # of times topic t has been guessed for other words in document d and # of times topic t has been guessed for all other occurrences of word w

The End Result...





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.

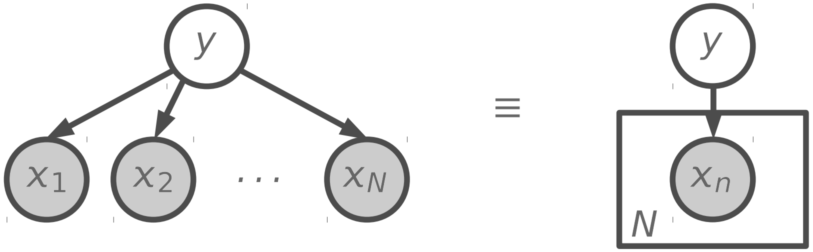
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Directed Graphical Models

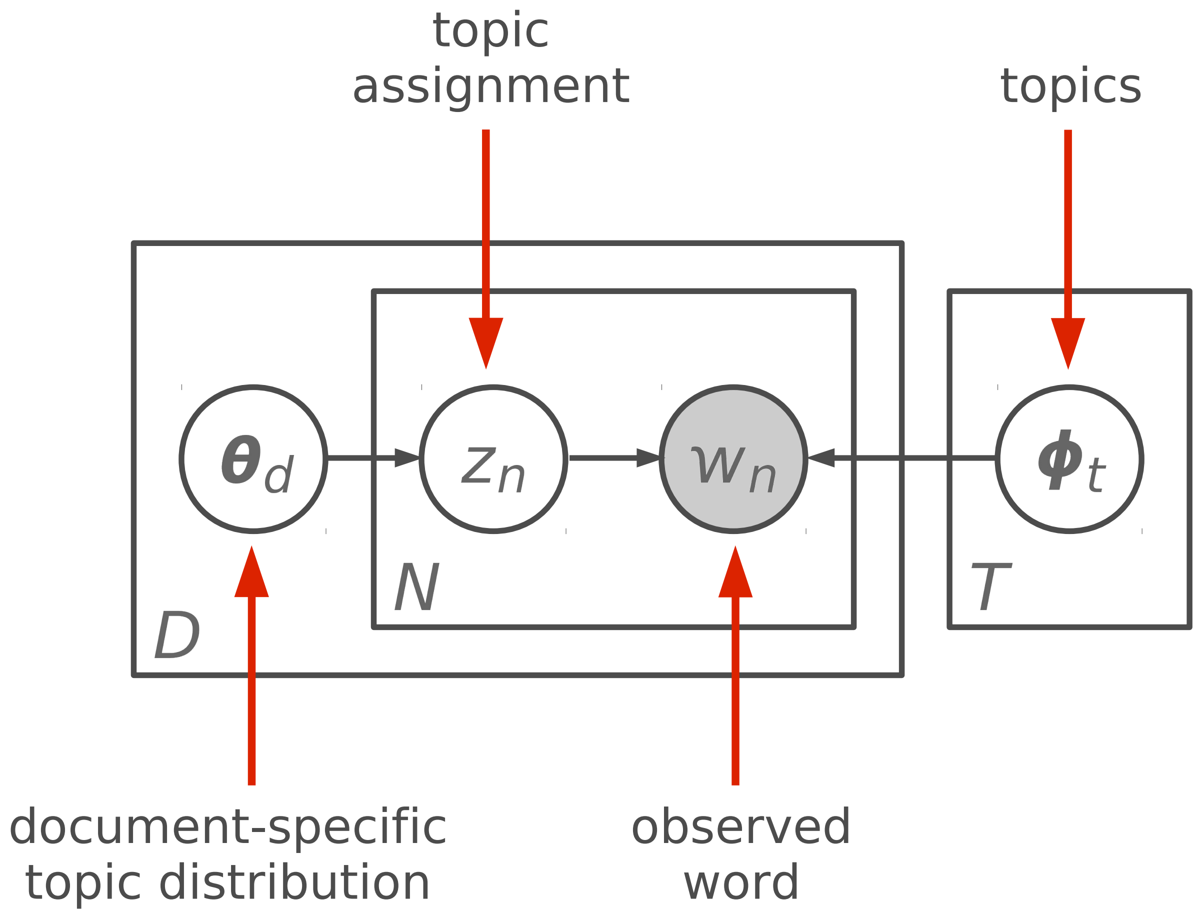
$$P(y, x_1, \ldots, x_N) = P(y) \prod_{n=1}^{N} P(x_n \,|\, y)$$

- Nodes: random variables (latent or observed)
- Edges: probabilistic dependencies between variables
- Plates: “macros” that allow subgraphs to be replicated



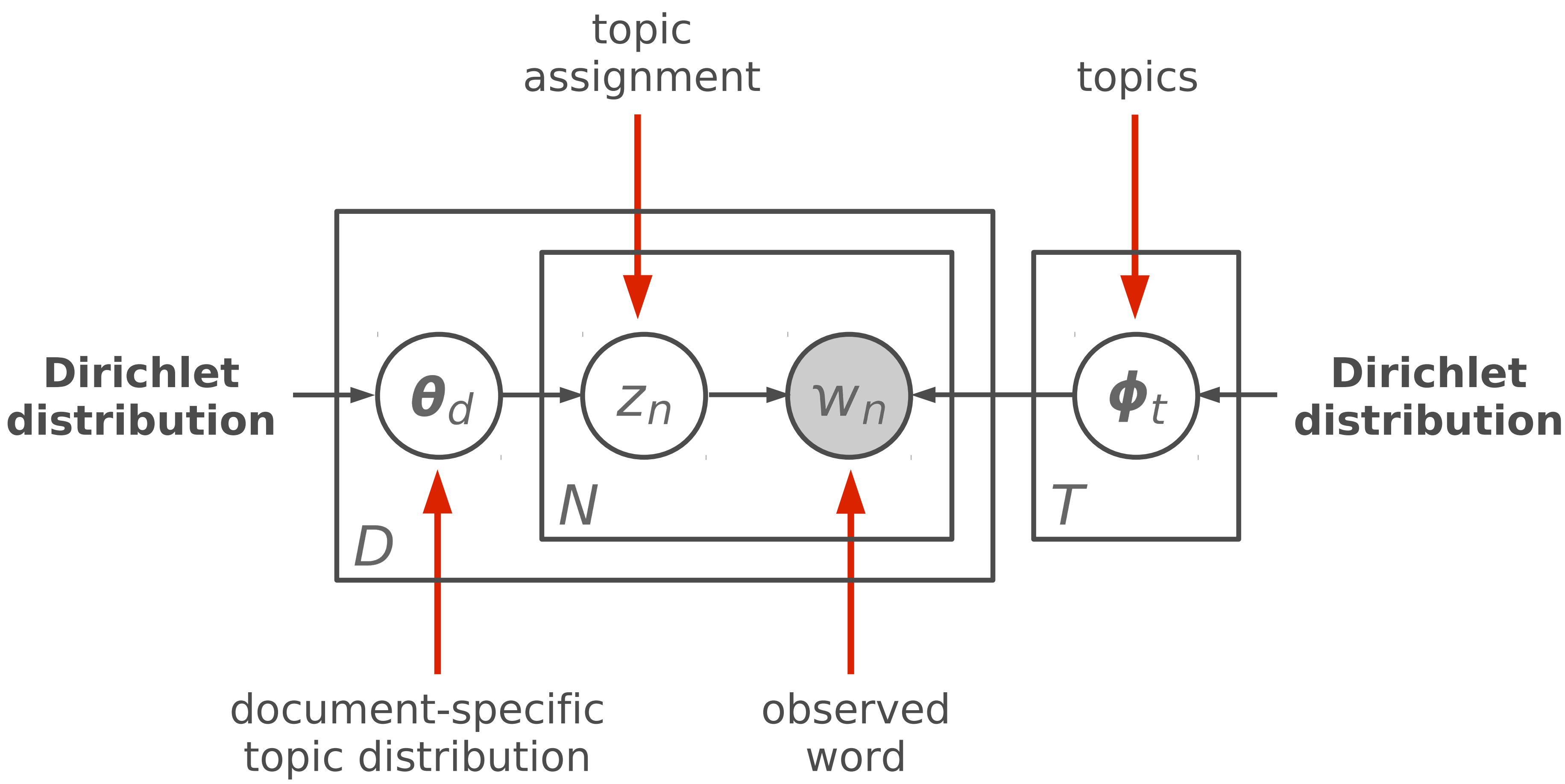
Statistical Topic Modeling

[Hofmann, '99]



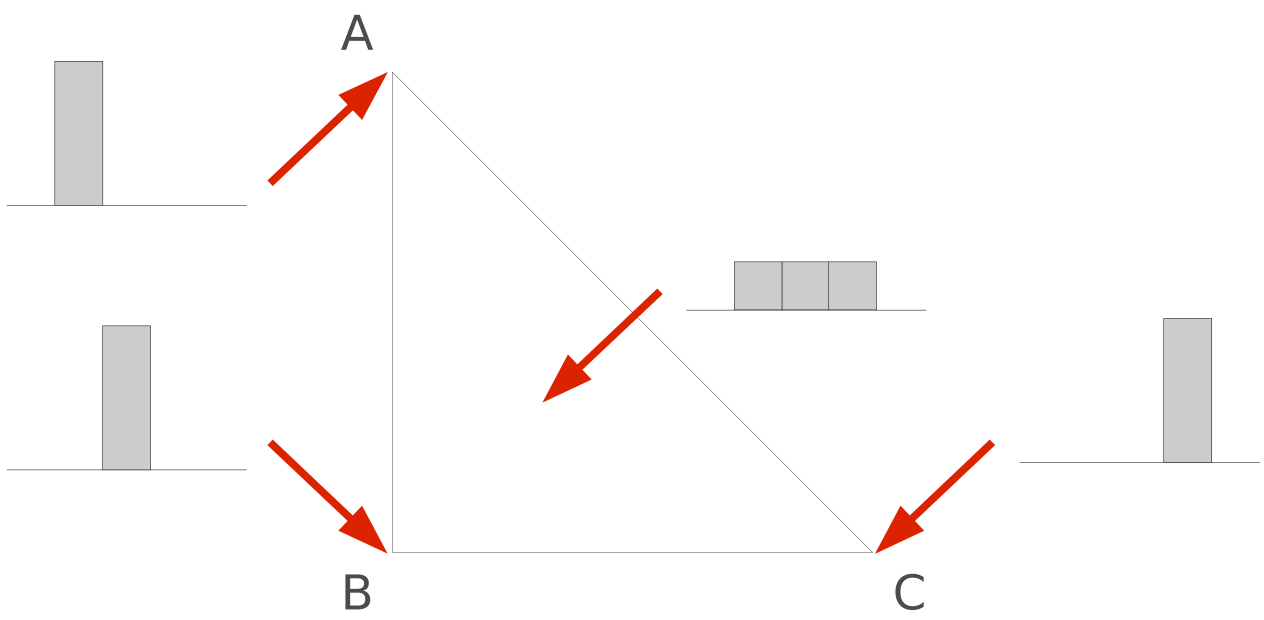
Latent Dirichlet Allocation (LDA)

[Blei, Ng & Jordan, '03]



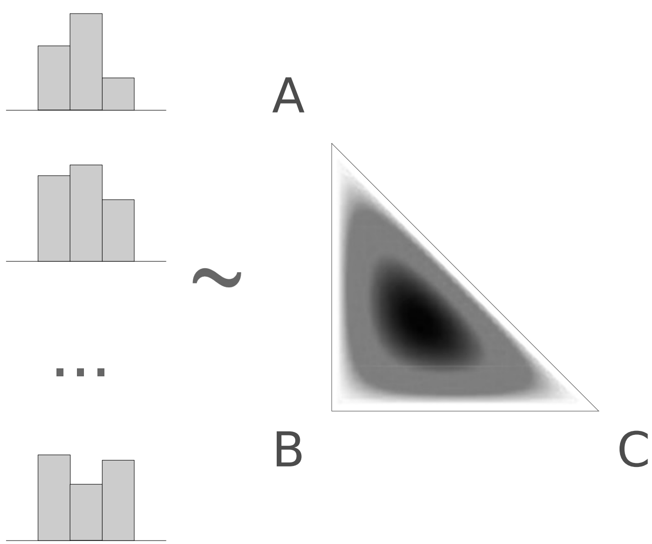
Discrete Probability Distributions

- 3-dimensional discrete probability distributions can be visually represented in 2-dimensional space:



Dirichlet Distribution

- Distribution over discrete probability distributions:

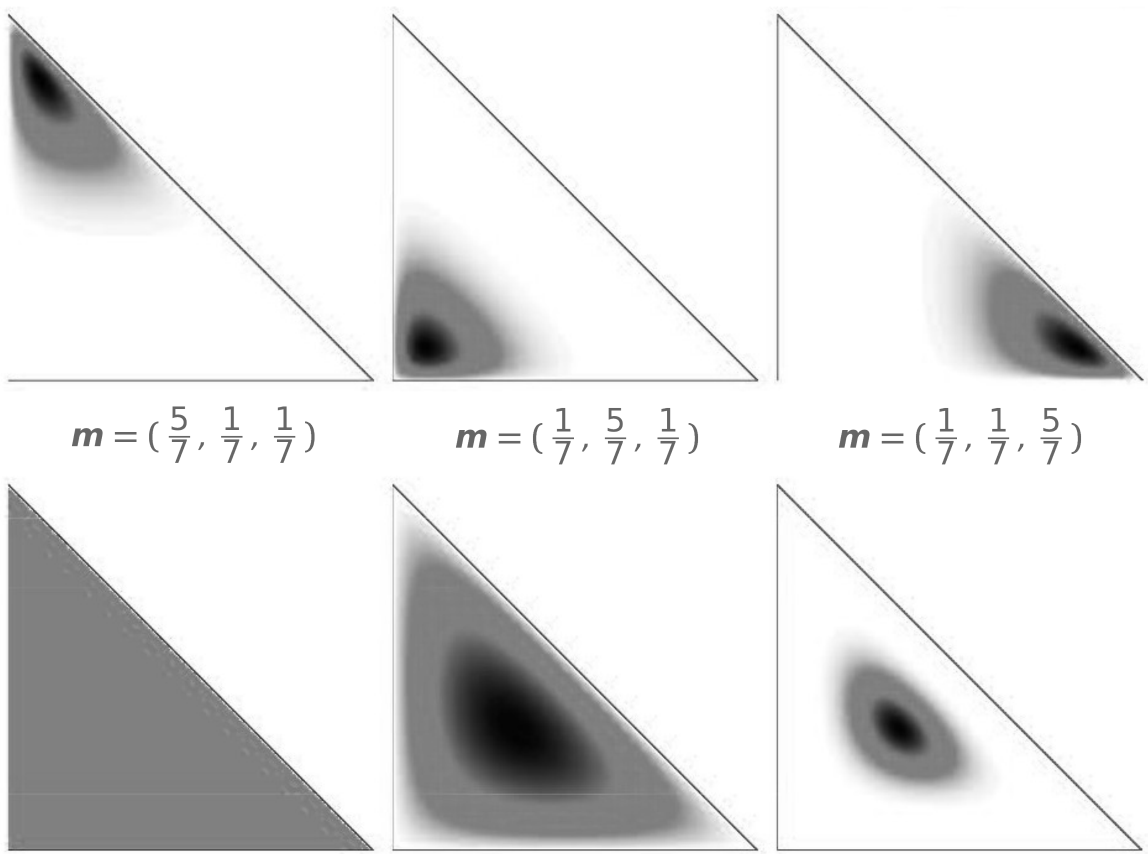


base measure (mean)

$$\equiv \quad \boldsymbol{p} \sim \mathrm{Dir}(\alpha \boldsymbol{m})$$

concentration parameter

Dirichlet Parameters



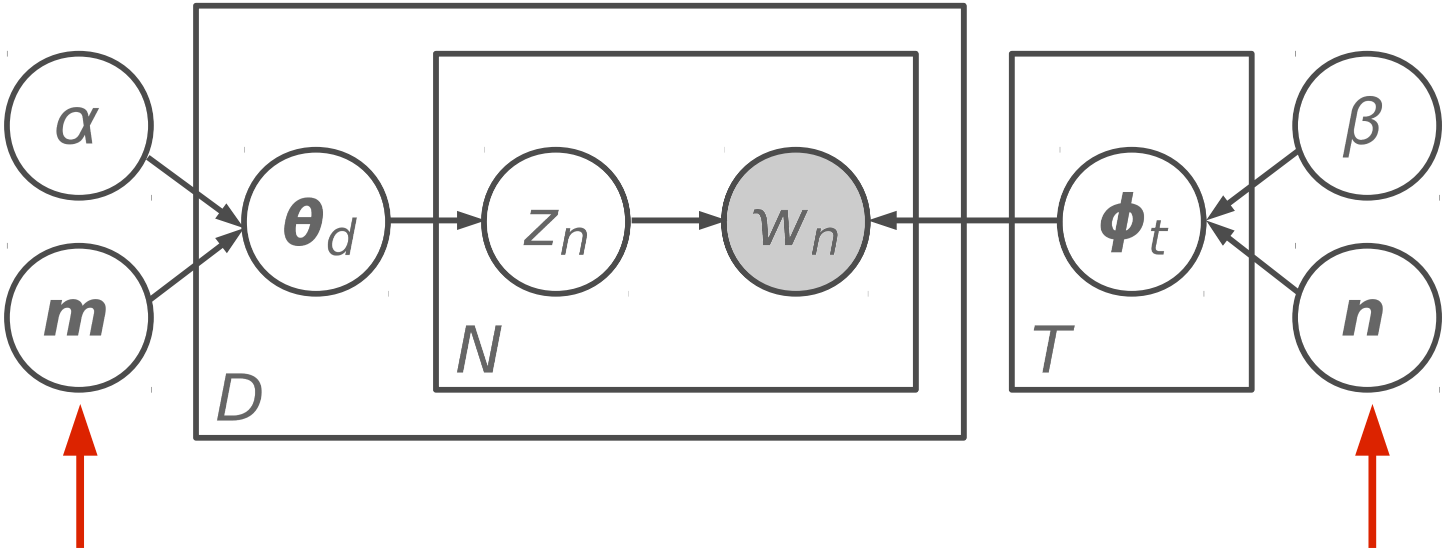
$\alpha = 14$

$\boldsymbol{m} = (\frac{5}{7}, \frac{1}{7}, \frac{1}{7})$ $\boldsymbol{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7})$ $\boldsymbol{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7})$

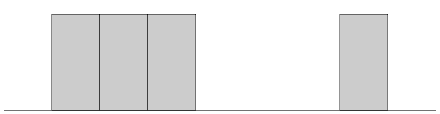
$\boldsymbol{m} = \boldsymbol{u} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

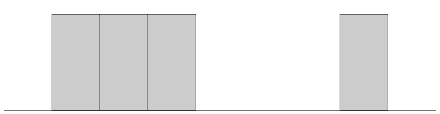
$\alpha = 3$ $\alpha = 6$ $\alpha = 30$

Dirichlet Priors for LDA



symmetric priors:
uniform base measures





This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Some current and future projects:
  - Analyzing free software development communities
  - Predicting when to declassify documents

The State of The Art

- Topic models are extremely popular
- … but they're not always usable by non-experts
- Need to bridge this gap between producers and consumers of topic modeling technology:
  - Address problems/challenges faced by practitioners
  - Question unquestioned assumptions
  - Explore the interplay between theory and practice

“Off-the-Shelf” Topic Modeling

I want to model technology emergence by analyzing patent abstracts...

I have a statistical model that you can use...

“Off-the-Shelf” Topic Modeling

I want to model technology emergence by analyzing patent abstracts...

I have a statistical model that you can use...

a	a	the	the
field	the	of	**invention**
emission	**carbon**	a	of
an	and	to	to
electron	**gas**	and	**present**
...	...	...	...

“Off-the-Shelf” Topic Modeling?

Help! All my topics consist of “the, and of, to, a …”

Preprocess your data to remove stop words…

Now they all consist of “invention, present, thereof ...”

Make a domain-specific list of stop words...

Wait, but how do I choose the right number of topics?

Evaluate the probability of unseen data for different numbers...

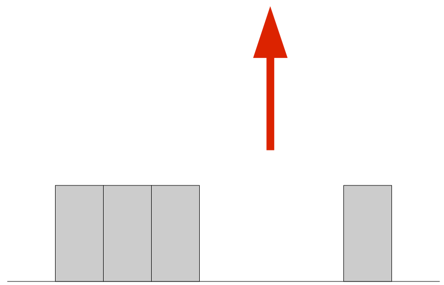
Dirichlet Priors for LDA

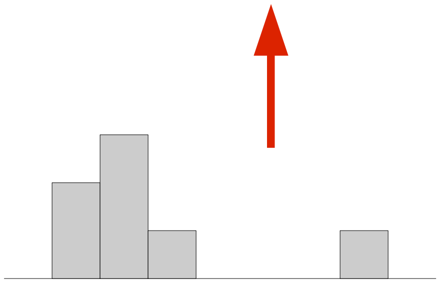
- Two scalar concentration parameters: α and β
- Concentration parameters are usually set heuristically
  - e.g., $\alpha = 50$ and $\beta = 0.01W$
- Some recent work on learning optimal values for the concentration parameters from data
- No rigorous study of the Dirichlet priors:
  - e.g., asymmetric vs. symmetric base measures
  - Effects of the base measures on the inferred topics

Symmetric → Asymmetric

- Use prior over $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$ as a running example
- Uniform base measure → nonuniform base measure

$$\Theta \sim \mathrm{Dir}\,(\alpha \boldsymbol{m}) \qquad\qquad \Theta \sim \mathrm{Dir}\,(\alpha \boldsymbol{m})$$



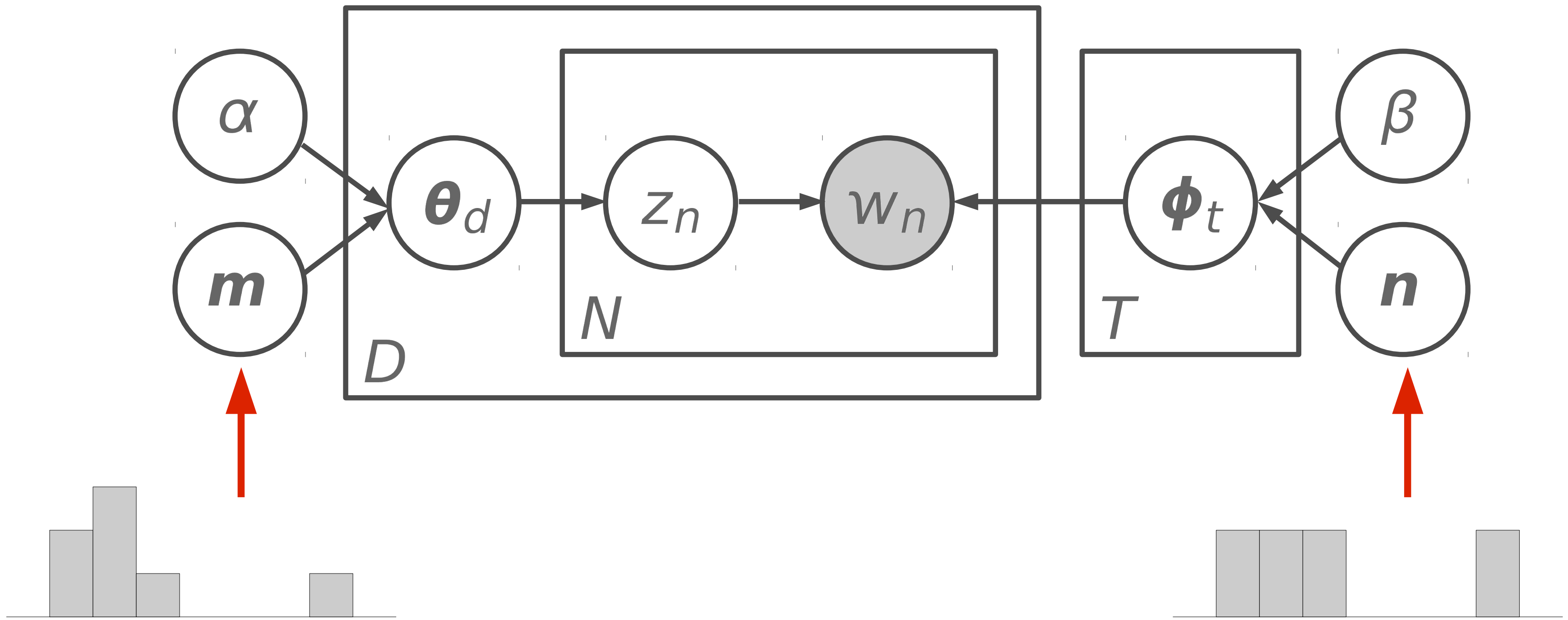


- Asymmetric prior: some topics more likely a priori

Data Sets

- Carbon nanotechnology patents:
  - Ultimate goal: track innovation and emergence
  - Fullerene and carbon nanotube patents
  - 1,016 abstracts (~100 words each)
  - 103,499 total words; 6,068 unique words
- 20 Newsgroups data (80,012 total words)
- New York Times articles (477,465 total words)

The Result



α
$\boldsymbol{m}$
$\boldsymbol{\theta}_d$
z_n
w_n
$\boldsymbol{\phi}_t$
β
$\boldsymbol{n}$
D
N
T

Inferred Topics

before

a	a	the	the
field	the	of	**invention**
emission	**carbon**	a	of
an	and	to	to
electron	**gas**	and	**present**
...	...	...	...

after

the	**carbon**	**metal**	**composite**
a	**nanotubes**	**catalytic**	**polymer**
of	**nanotube**	**transition**	**matrix**
to	**catalyst**	**catalyst**	**weight**
and	**substrate**	from	**fiber**
...	...	...	...

Intuition

- Topics should be distinct from each other:
  - Asymmetric prior over topics makes topics more similar to each other (and to corpus-wide word frequencies)
  - Want a symmetric prior to preserve topic “distinctness”
- Still have to account for power-law word usage:
  - Asymmetric prior over document-specific topic distributions means some topics (e.g., “the, a, of, to ...”) can be used more often than others in all documents

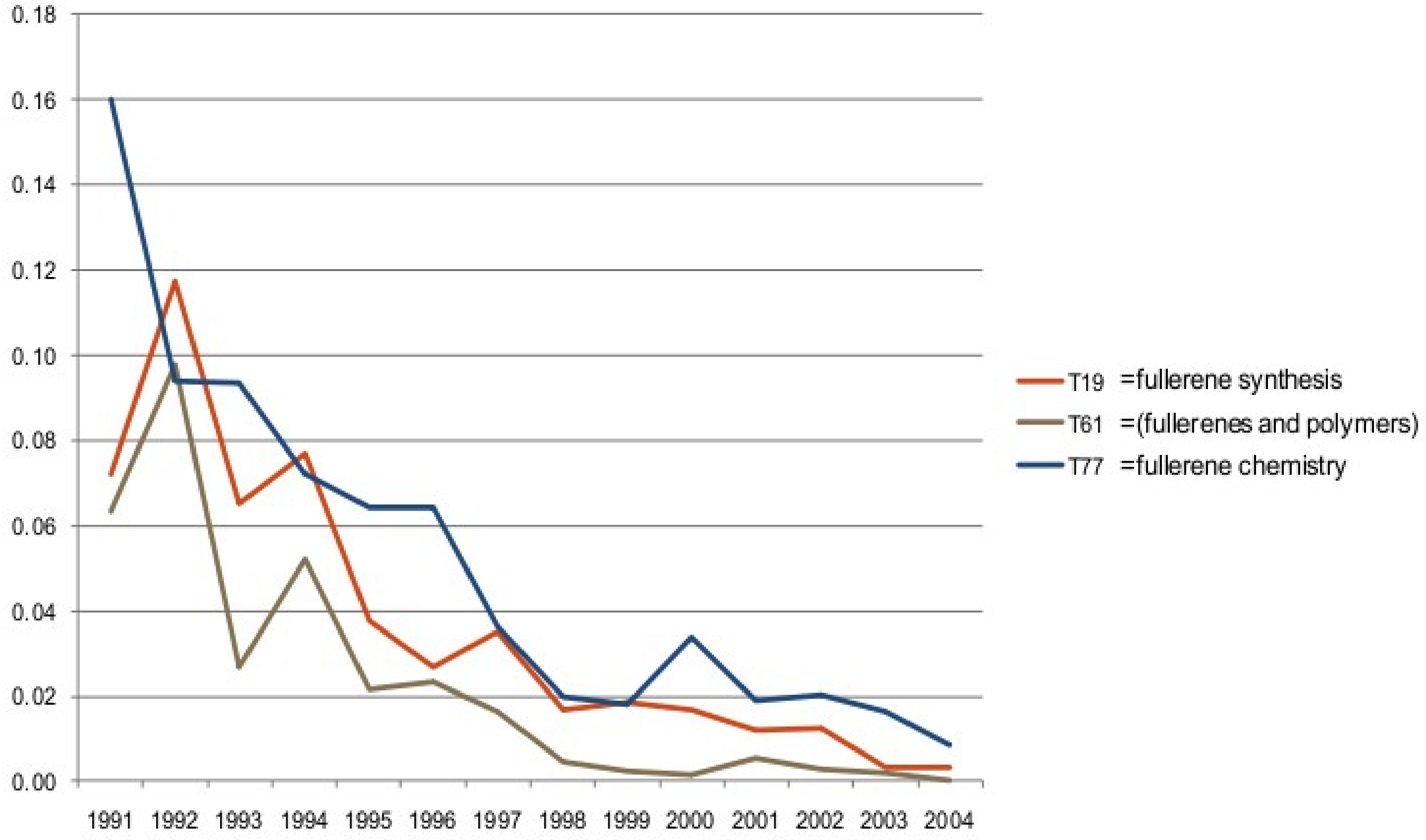
“Off-the-Shelf” Topic Modeling

I can model technology emergence by analyzing patent abstracts!

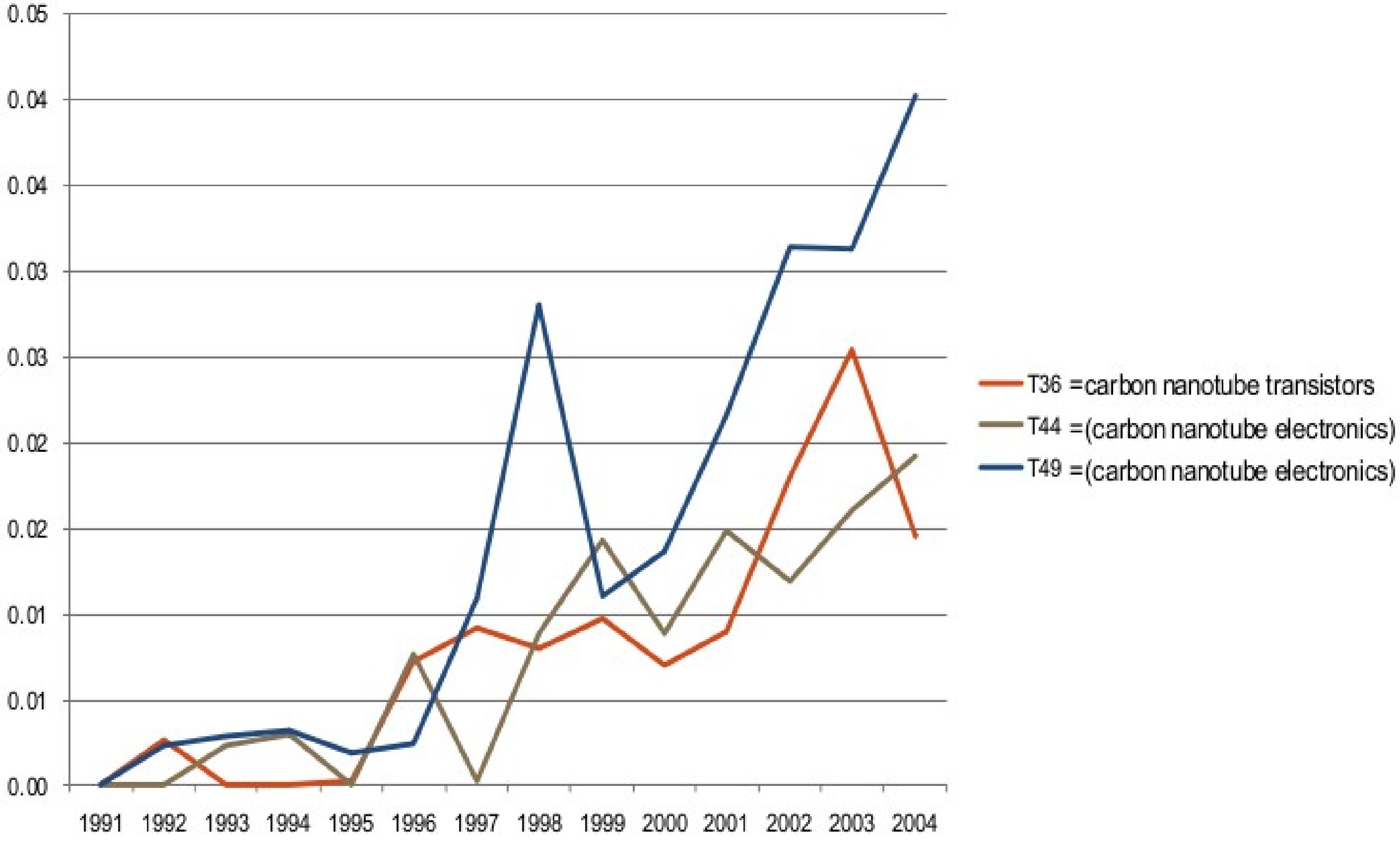
Great! Let me know if you need any more help!

the	**carbon**	**metal**	**composite**
a	**nanotubes**	**catalytic**	**polymer**
of	**nanotube**	**transition**	**matrix**
to	**catalyst**	**catalyst**	**weight**
and	**substrate**	from	**fiber**
...	...	...	...

Declining Topics



Rising Topics



Topic Model Output in Practice

- Common to use a single random set of topic assignments (typically the last sample drawn)
- Not necessarily representative
- Other (better) approaches:
  - Average summary statistics over multiple samples
  - Use the posterior mode
  - Use hierarchical agglomerative clustering to construct a new aggregate set of topic assignments

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Some current and future projects:
  - Analyzing free software development communities
  - Predicting when to declassify documents

FOSS Development Communities

- Considerable commercial, noncommercial, academic interest in FOSS development communities:
  - Complex technological, legal, social structures
  - Geographically distributed collaboration
- Organizational and social processes underlying collaborative FOSS development are largely unknown:
  - Area of study for social and computer scientists

FOSS Collaboration Data

- Most FOSS collaboration data are publicly available:
  - Mailing lists, IRC channels
  - Commit messages, bug reports
  - Comments in source code, documentation
  - GPG keysigning records

⇒ Use these collaboration data to study organizational and social processes underlying FOSS development

Data Challenges

- Informal, messy, and often highly unstructured data:
  - Developers use different identifiers in different fora
  - IRC channels have multiple interleaved conversations
  - Mix of highly technical and “off-topic” discussion
  - Conversational style is often casual

⇒ Significant text analysis is required prior to developing models for answering social science questions

Analyzing Debian Mailing Lists

- Quoted text and signatures stripped
- Debian-project mailing list:
  - 19,347 messages
  - 1225797 words (max. 7,916 per message)
- Debian-women mailing list:
  - 4,124 messages
  - 228,076 words (max. 1,524 per message)

100 Topics

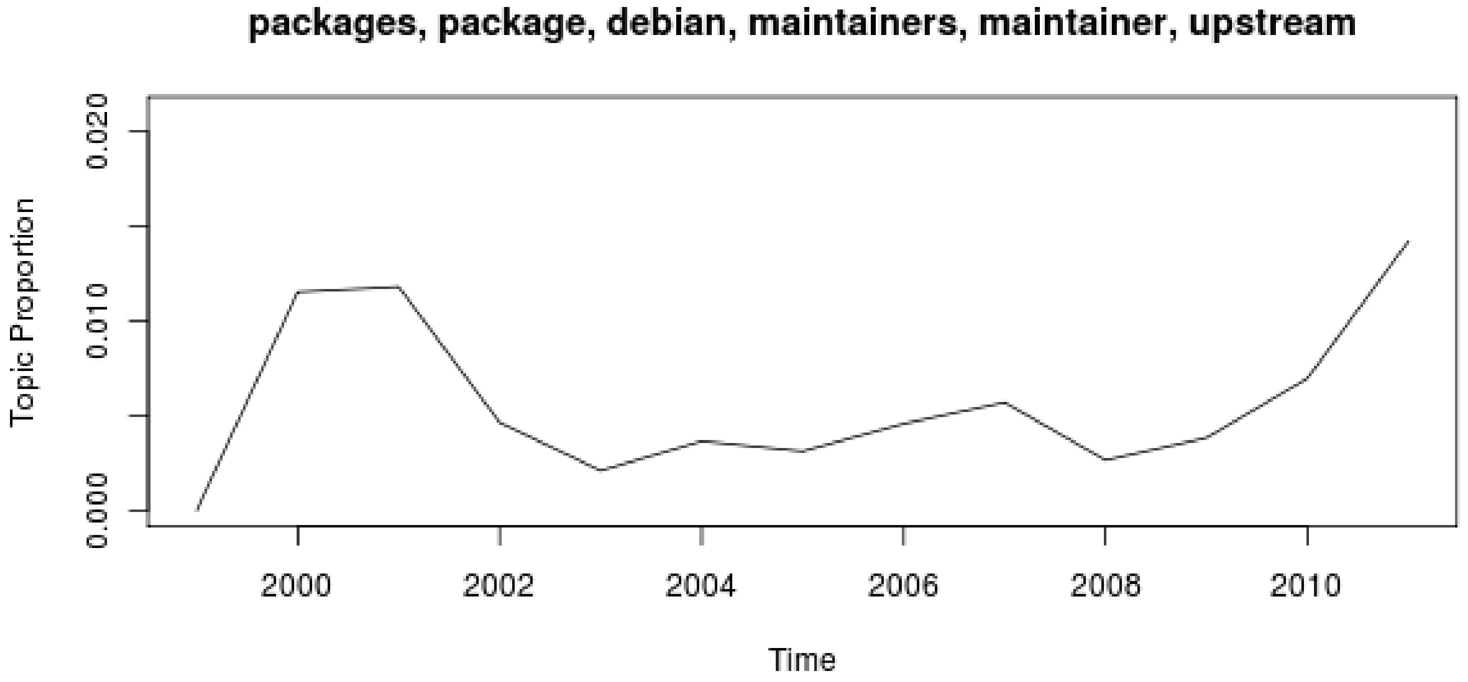
d-project →

package	ubuntu	nm	ftp-master
packages	debian	process	queue
install	patches	applicant	packages
apt-get	derivatives	dam	upload
apt	lts	fd	team
...	...	...	...

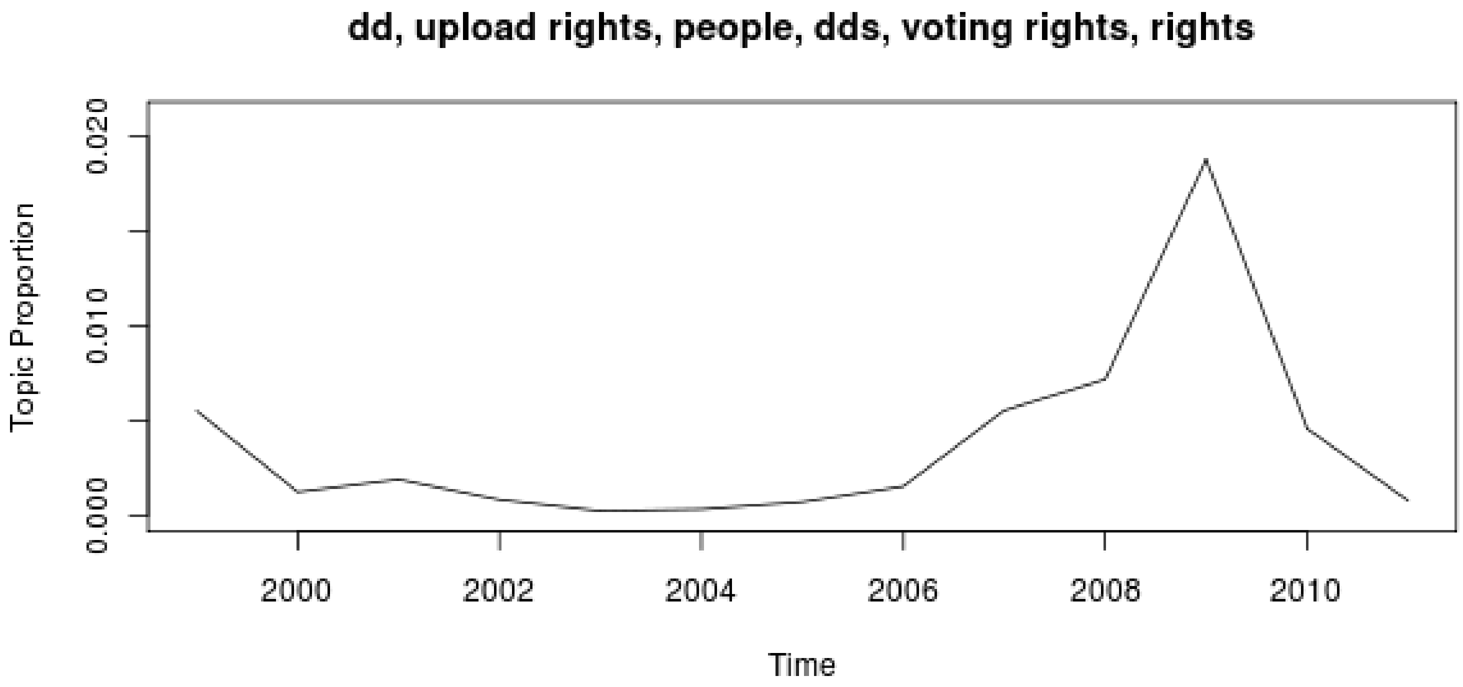
d-women →

women	website	post	nm
men	page	culture	debian
female	site	response	process
male	work	posts	dd
man	d-w	behavior	packages
...	...	...	...

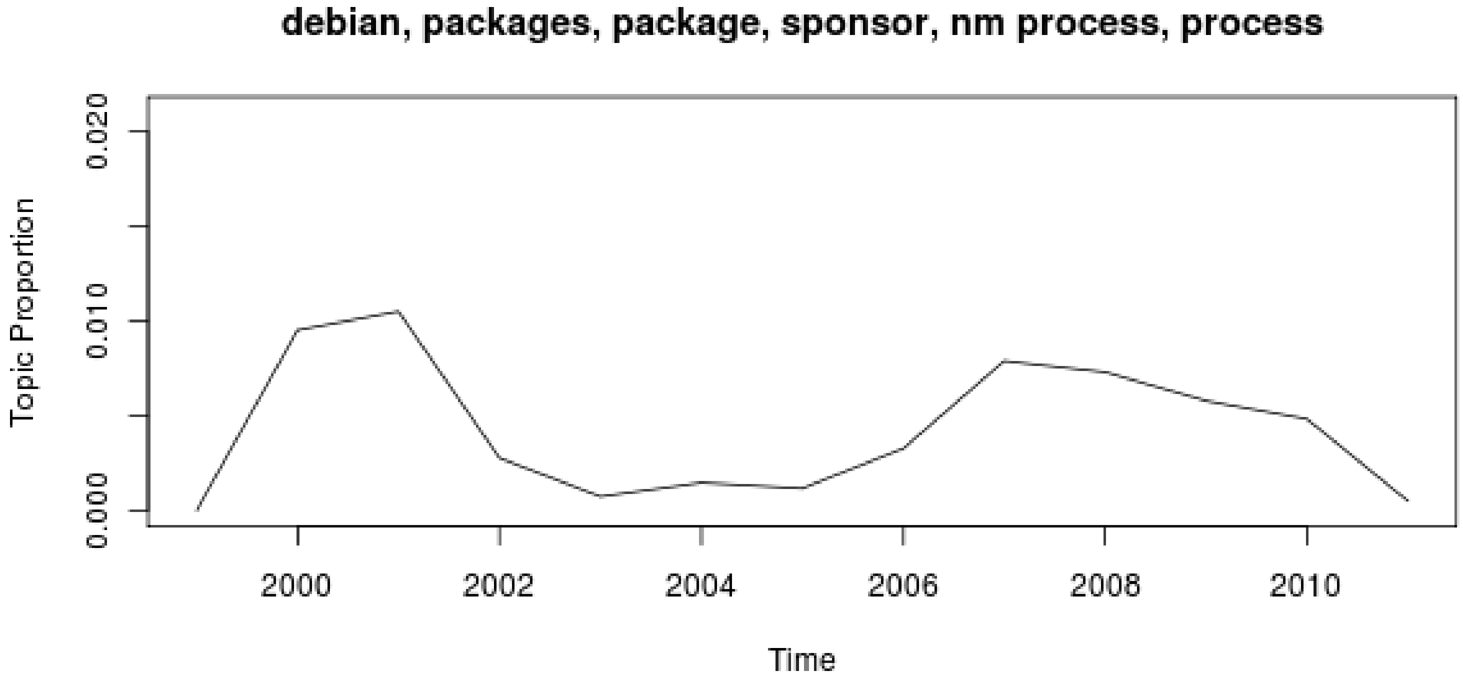
Topic Usage Over Time



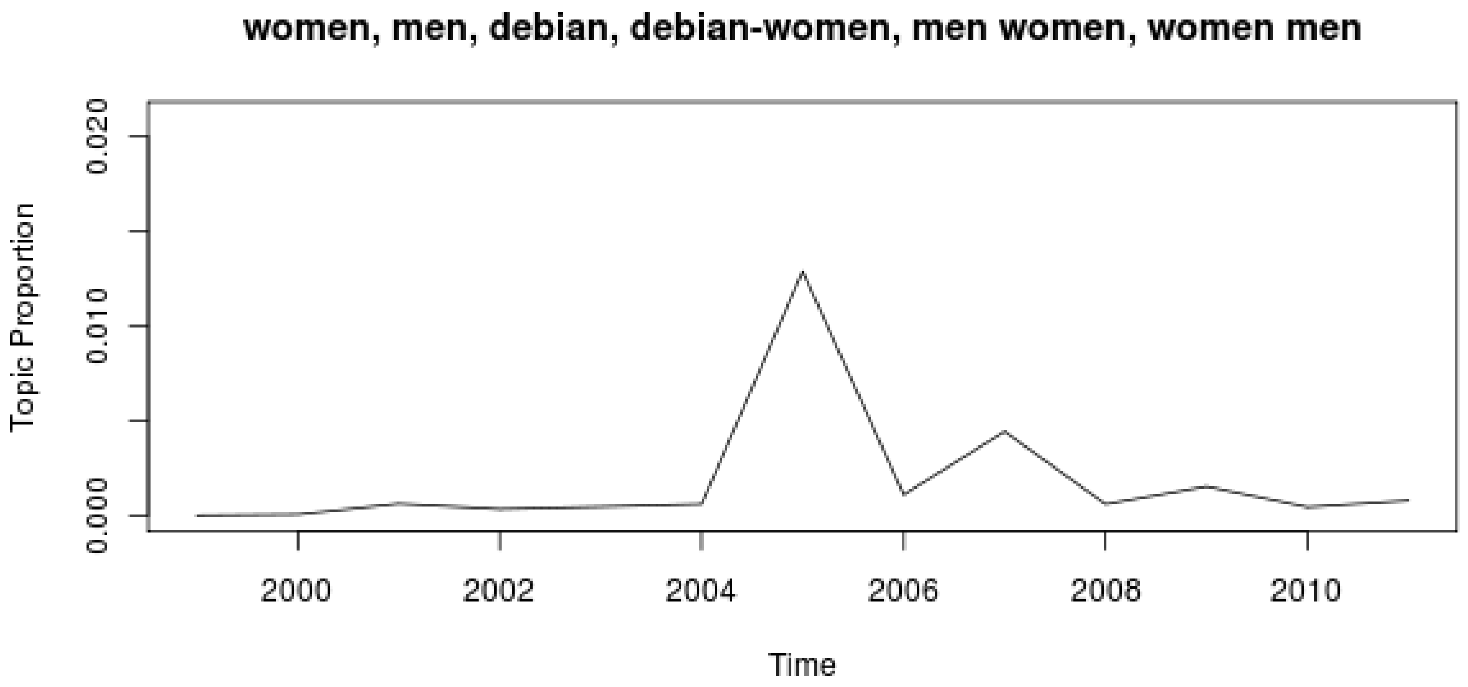
Topic Usage Over Time



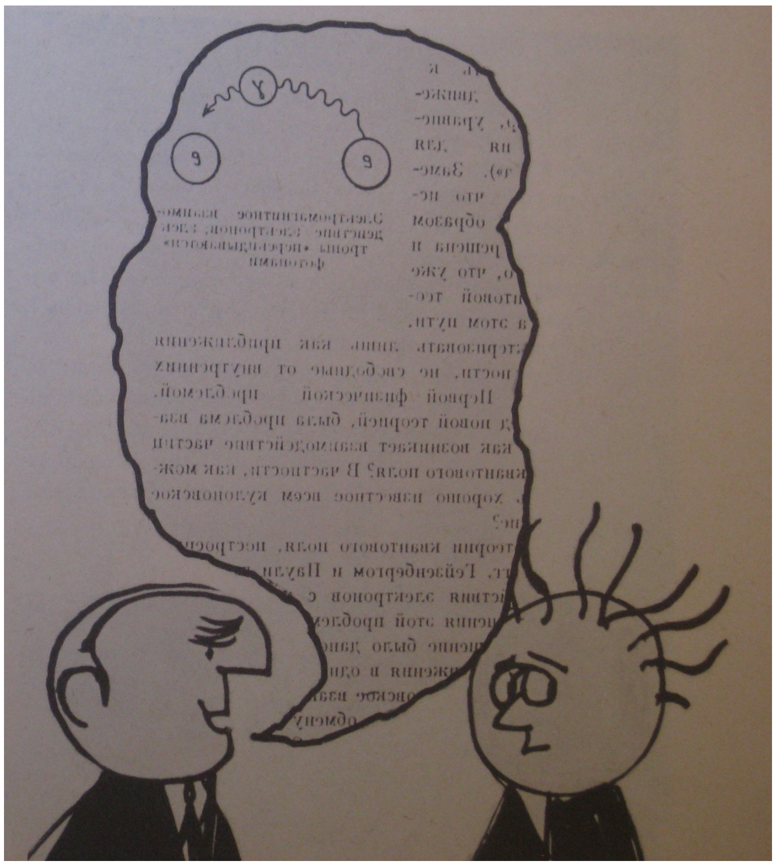
Topic Usage Over Time



Topic Usage Over Time



Cross-language Analysis



"He may know one language backwards and forward, but he can't communicate with a scientist who only knows another: a graphic illustration of the need for translation of foreign scientific documents."

— NSF Brochure, 1962

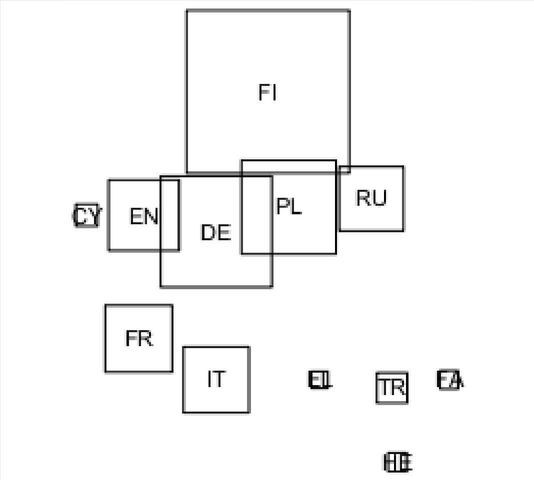
Polylingual Topics

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضايي ماموريت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

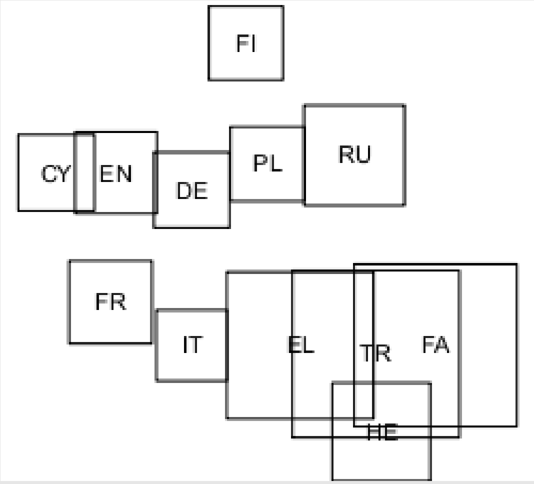
Polylingual Topics

CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

Differences in Topic Emphasis



world
ski
km
won
...



ottoman
empire
khan
byzantine
...

Aligned Corpora

- Fully parallel corpora: direct translations
  - Expensive to produce, relatively rare
- Partially parallel corpora: few parallel “glue” tuples
  - $<$ 25% is sufficient to obtain aligned topics
- Comparable corpora: documents have similar content
  - e.g., Wikipedia in English, Portuguese, French, ...
  - e.g., documentation in multiple languages

Cross-cultural Study of FOSS

- Use Wikipedia pages, FOSS websites, documentation in different languages as aligned document tuples
- Use resultant topics to analyze mailing lists in order to study FOSS culture in different parts of the world
- Specifically interested in:
  - Brazil (governmental adoption of FOSS)
  - European countries (e.g., France)
  - United States

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Some current and future projects:
  - Analyzing free software development communities
  - Predicting when to declassify documents

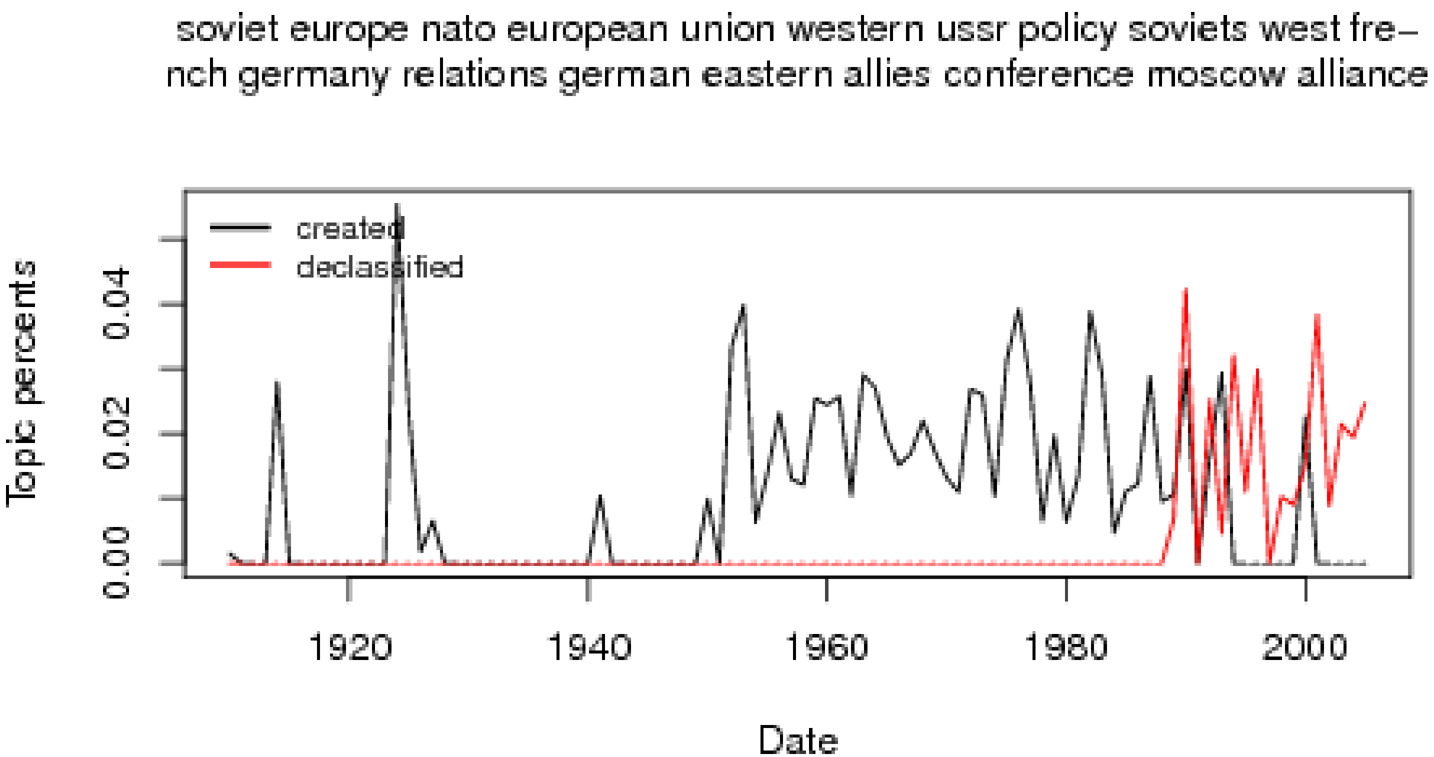
Document Declassification

- Massive amount of secret data is protected as part of the United States Government Classification System
- Human readers manually declassified almost 29,000,000 pages of information in 2009
- Need automated tools:
  - Prioritize documents for human review
  - Academic study of (de)classification patterns

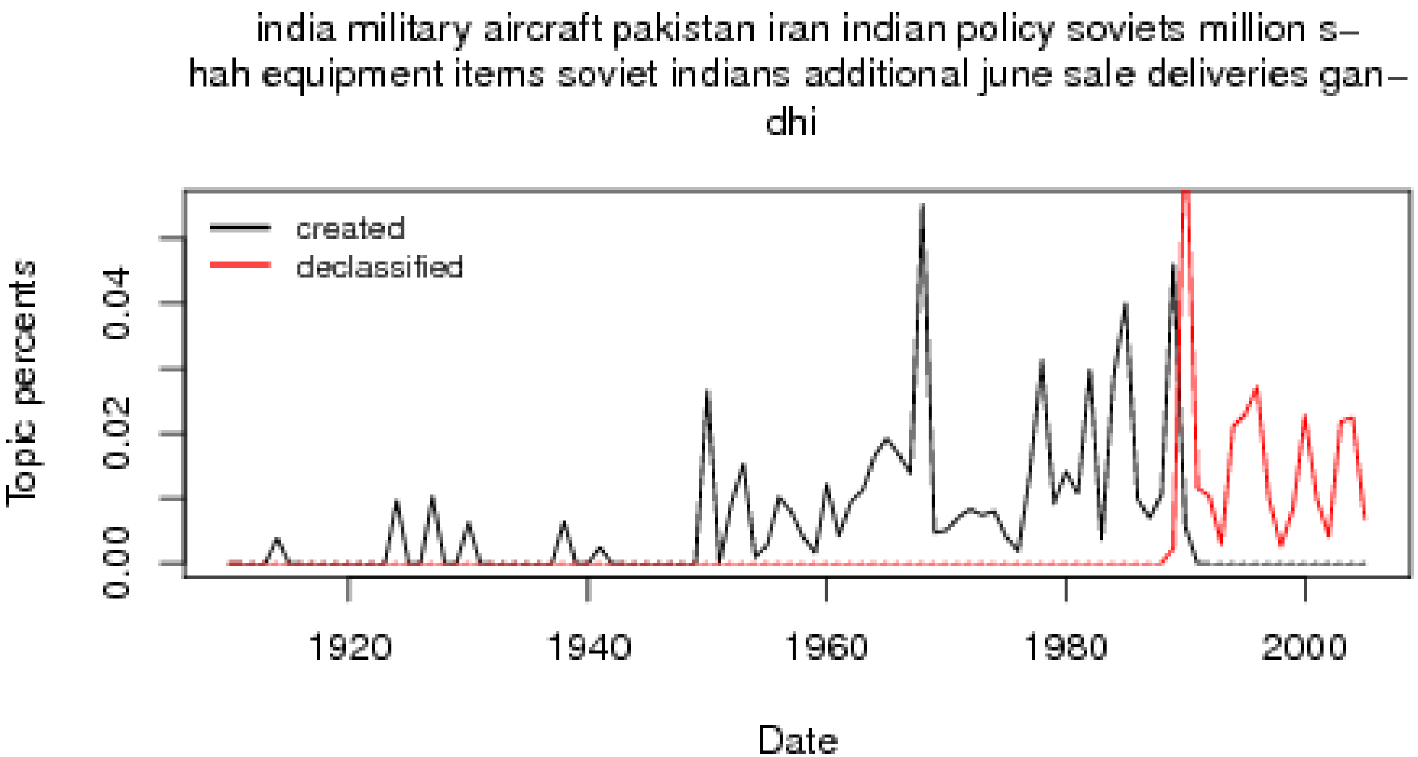
Declassified Documents

- Study classification patterns by looking at classification and declassification dates of declassified documents
- Model temporal patterns and document content
- Declassified Documents Reference System
  - 85,000 declassified documents
  - Classification, declassification dates
  - Issuer (e.g., White House)

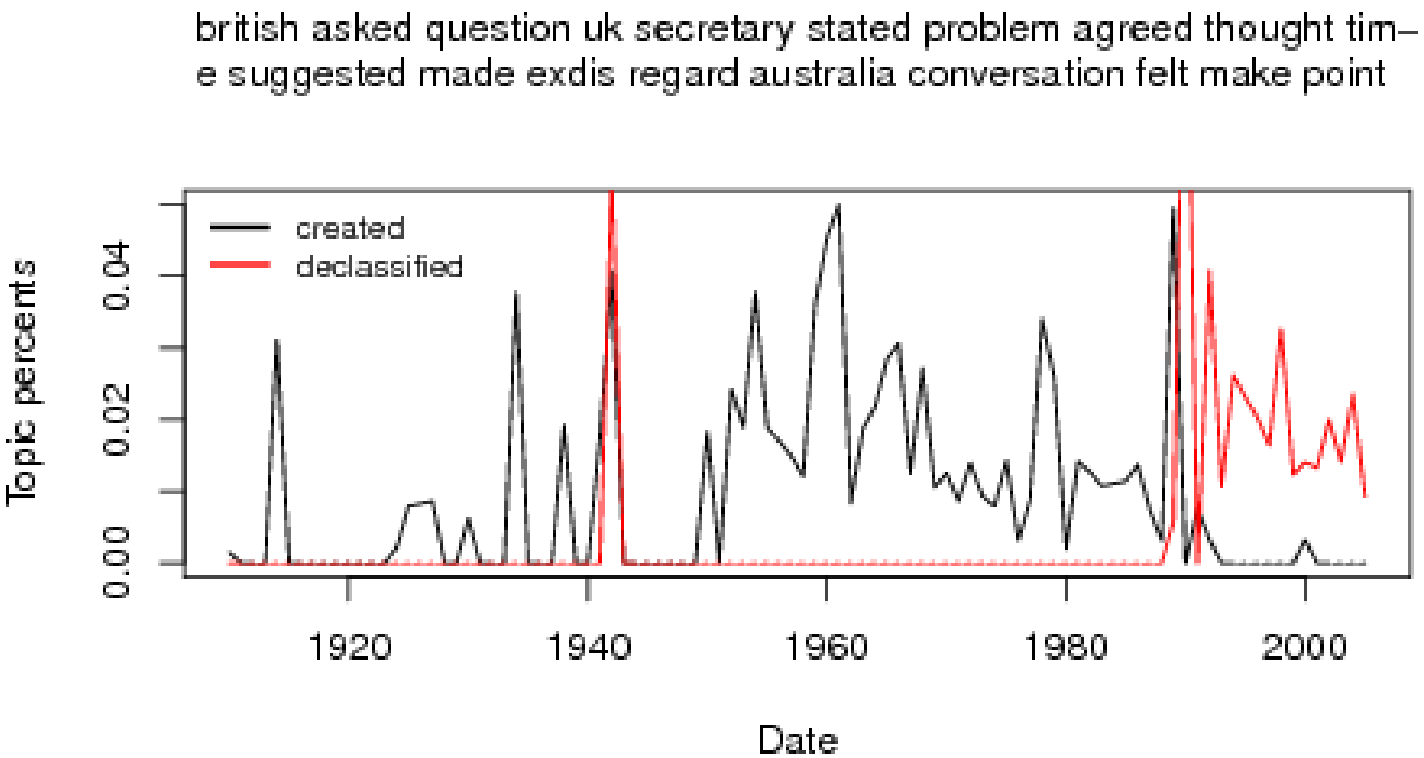
Declassification by Topic



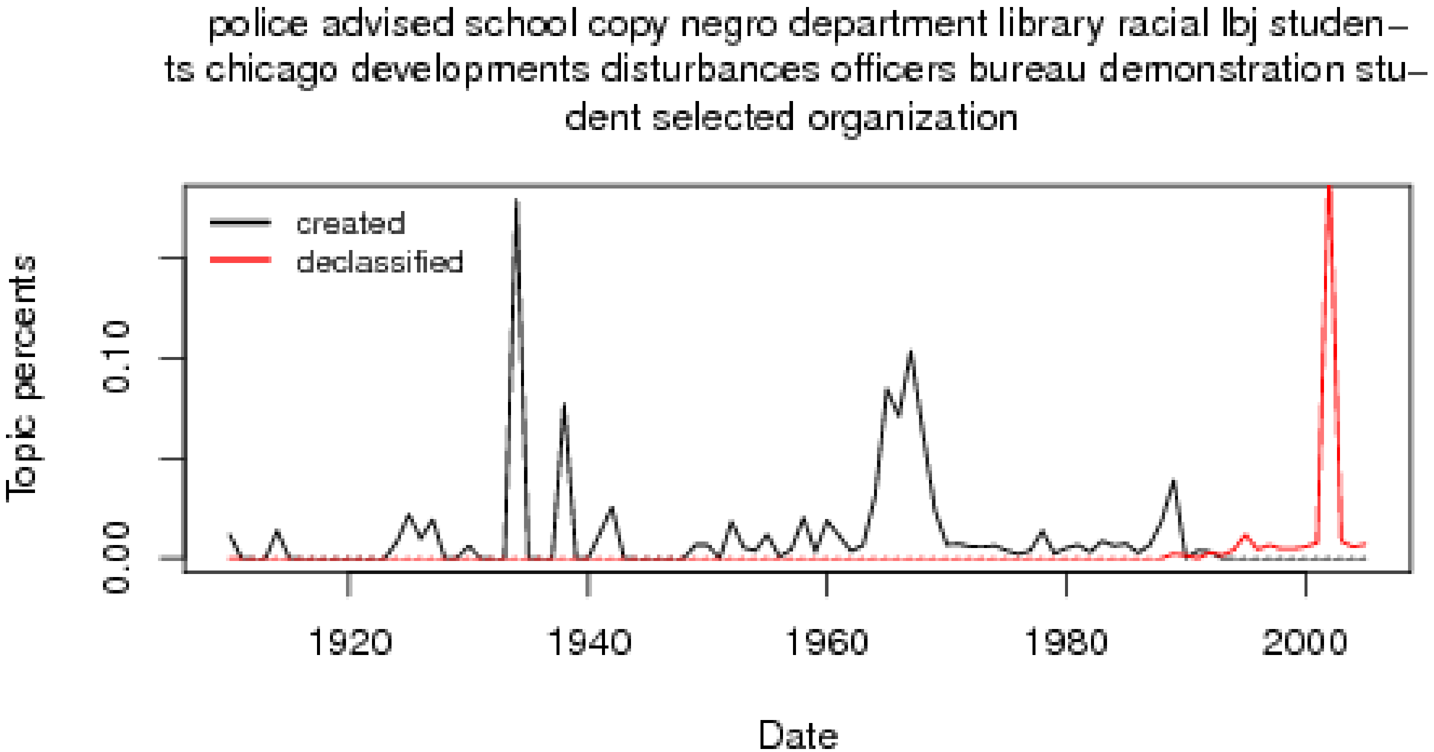
Topic Declassification



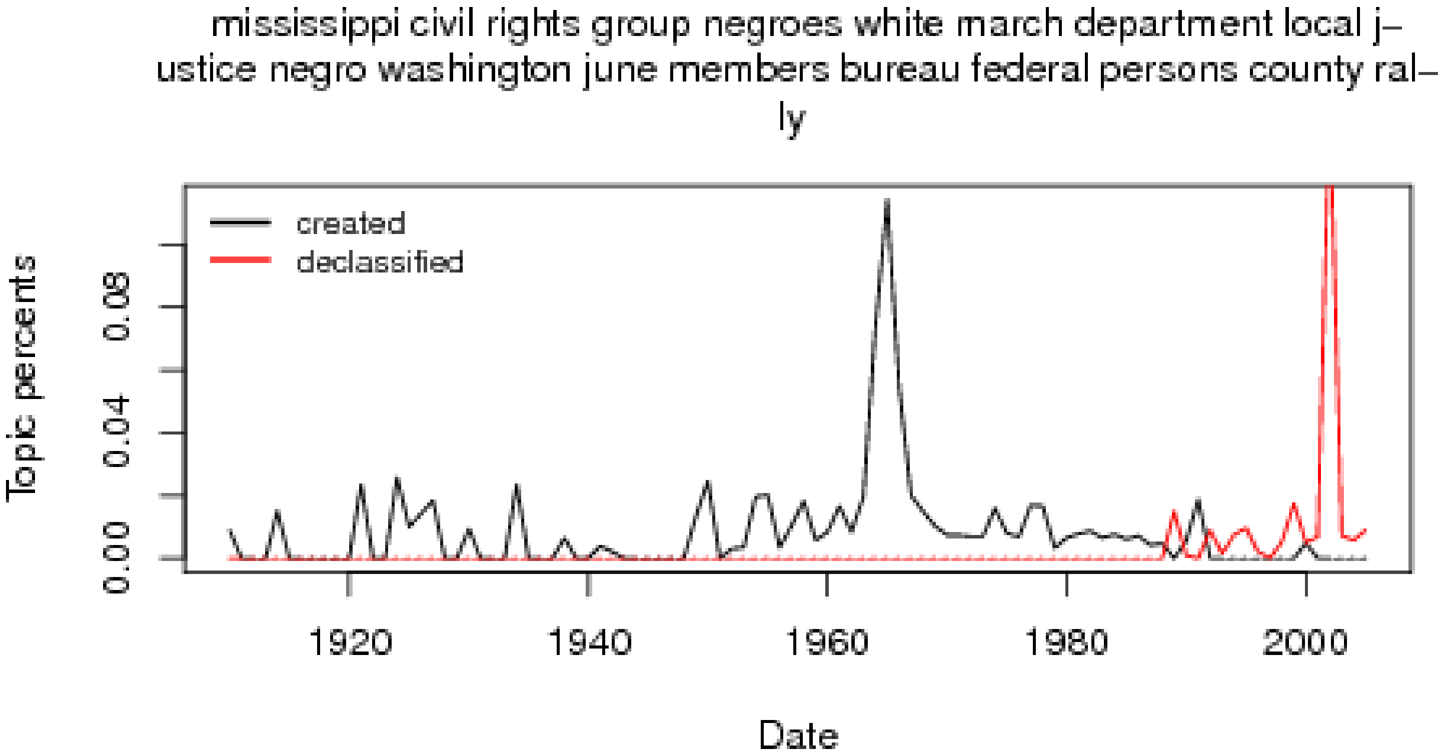
Topic Declassification



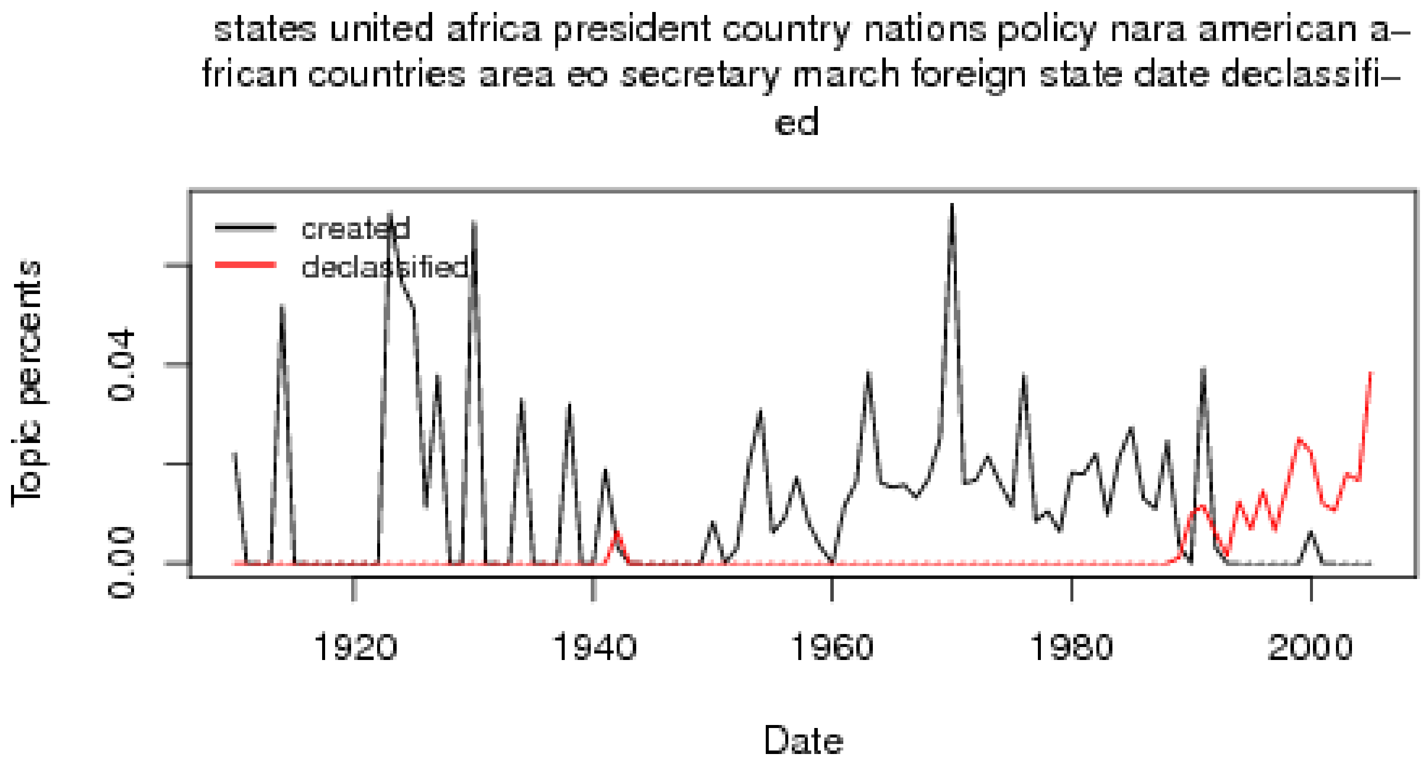
Topic Declassification



Topic Declassification



Topic Declassification



Thanks!

Acknowledgements: Dafydd Harries, Sarah Kaplan, Andrew McCallum,
David Mimno, David Nusinow, Rachel Shorey