Statistical Topic Models for Science and Innovation Policy

Hanna M. Wallach

University of Massachusetts Amherst

wallach@cs.umass.edu

Science and Innovation

"Whether it's improving our health or harnessing clean energy, protecting our security or succeeding in the global economy, our future depends on reaffirming America's role as the world's engine of scientific discovery and technological innovation."

— President Barack Obama

… Behind the Scenes

"The public has generally treated this progress as something that just happened, without recognizing that it is, in fact, largely the result of a sustained federal commitment to support science through science policies."

— http://science-policy.net

Science and Innovation Policy

- Goal: identify administrative, financial, political actions
- Actions chosen to have impact on, e.g.,
  - Stimulating breakthrough research
  - Increasing economic prosperity
  - Broadening participation
- Government, private sector, education
- This talk: statistical models for facilitating efficient, data-driven science policy decisions

Examples of Policy Actions

- Funding actions:
  - Using federal funds for research on human stem cells
  - “People not projects” vs. pre-defined deliverables
- Patenting actions:
  - Granting software patents
- Educational actions:
  - Running high school outreach activities
  - Providing mentoring programs

Data-Driven Policy Decisions

Candida Hofer

- Discovery: identifying possible policy actions
- Prediction: estimating expected impact
- Evaluation: assessing observed outcomes

⇒ Automated data analysis

Data: 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

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

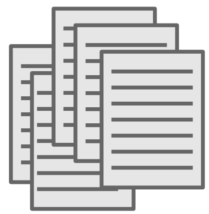
My Research Goal

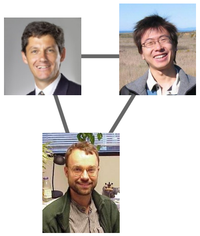
```
    $line .= <CASEBOOKS>;
    redo unless eof(CASEBOOKS);
}

$line =~ s/\\\t/xyzdrptmpxyz/g;
@columns = split("\t", $line);
$columns[3] = uc $columns[3];
$line = join("\t", @columns);
$line =~ s/xyzdrptmpxyz/\\\t/g;
```

$$\prod_t \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \frac{\prod_w \Gamma(N_{w|t}+\beta)}{\Gamma(N_{\cdot|t}+W\beta)}$$

+

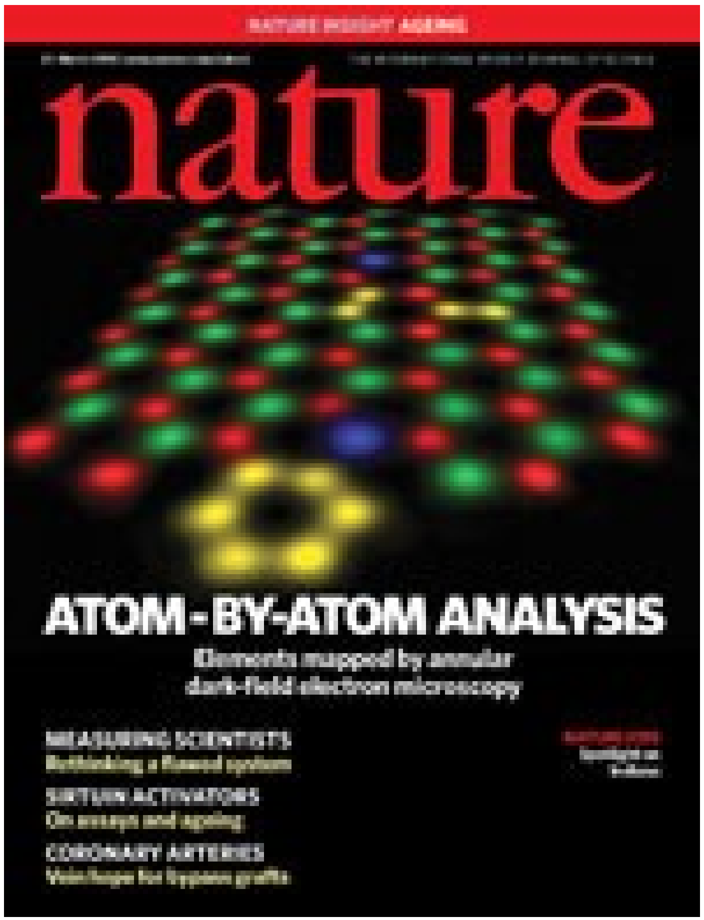




→

To develop new statistical models and computational tools for representing and analyzing large quantities of complex data in order to better enable scientific policy-makers to identify and evaluate high-impact policy actions and advance the study of science and innovation policy.

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

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

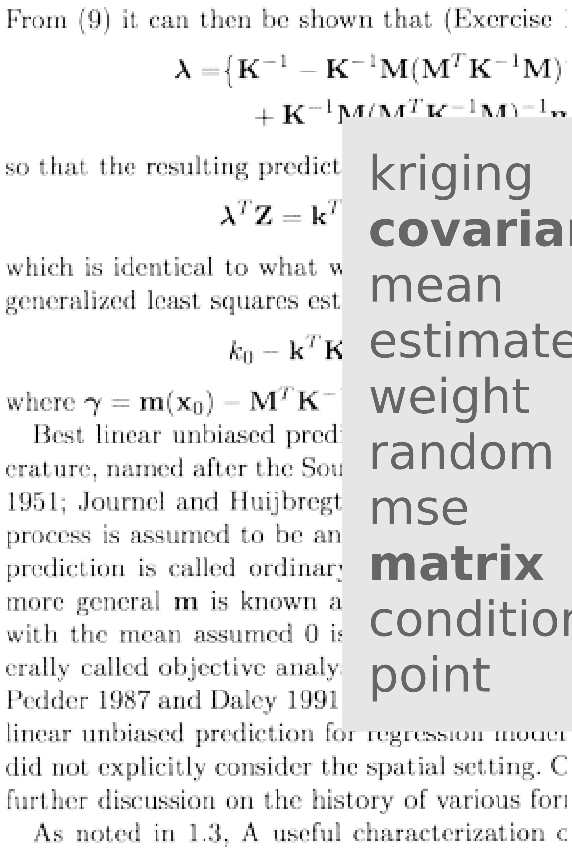
Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH

Why Topic Models?



From (9) it can then be shown that (Exercise

$\boldsymbol{\lambda} = \{\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{M}(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})$

so that the resulting predict

$\boldsymbol{\lambda}^T\mathbf{Z} = \mathbf{k}^T$

which is identical to what w
generalized least squares est

$k_0 - \mathbf{k}^T\mathbf{K}$

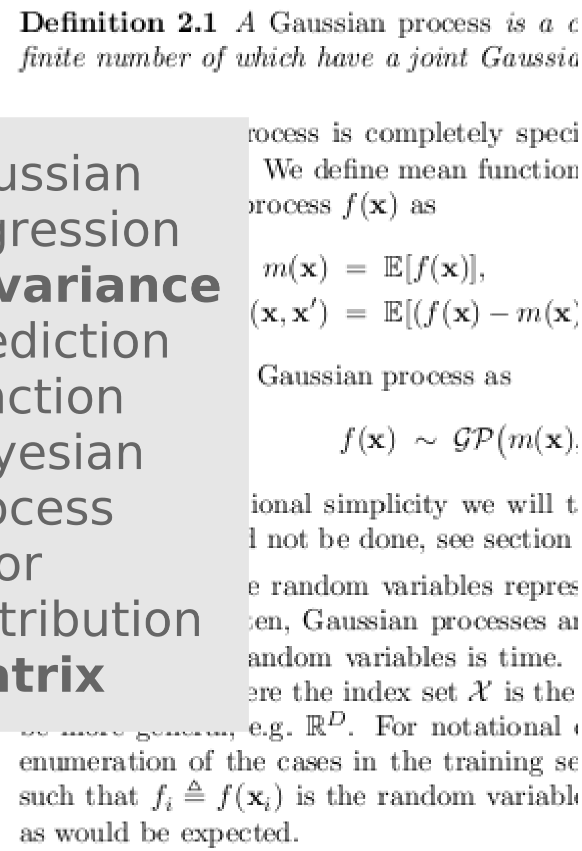
where $\boldsymbol{\gamma} = \mathbf{m}(\mathbf{x}_0) - \mathbf{M}^T\mathbf{K}^{-1}$

Best linear unbiased predi
erature, named after the Sou
1951; Journel and Huijbregt
process is assumed to be an
prediction is called ordinary
more general **m** is known a
with the mean assumed 0 is
erally called objective analys
Pedder 1987 and Daley 1991
linear unbiased prediction for regression model
did not explicitly consider the spatial setting. C
further discussion on the history of various for

As noted in 1.3, A useful characterization c

kriging
covariance
mean
estimate
weight
random
mse
matrix
conditional
point

vs.



Definition 2.1 *A Gaussian process is a c
finite number of which have a joint Gaussia*

rocess is completely speci
We define mean function
process $f(\mathbf{x})$ as

$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})],$

$(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x})$

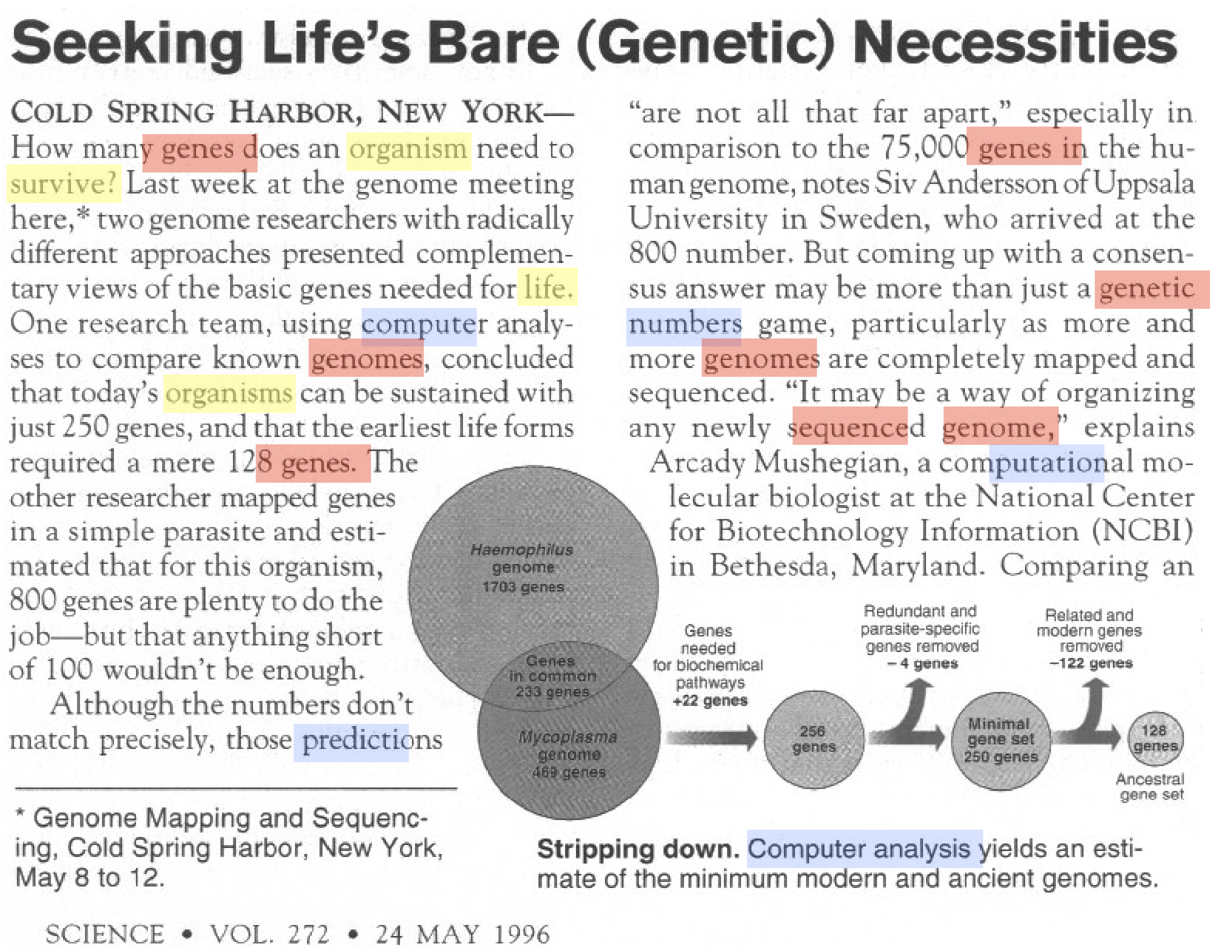
Gaussian process as

$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}),$

ional simplicity we will t
l not be done, see section
e random variables repres
ten, Gaussian processes a
andom variables is time.
ere the index set $\mathcal{X}$ is the
e.g. $\mathbb{R}^D$. For notational
enumeration of the cases in the training se
such that $f_i \triangleq f(\mathbf{x}_i)$ is the random variabl
as would be expected.

gaussian
regression
covariance
prediction
function
bayesian
process
prior
distribution
matrix

Documents and Topics



Seeking Life's Bare (Genetic) Necessities

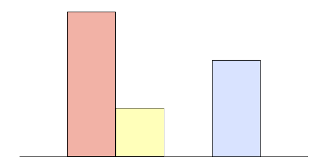
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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



Topics and Words

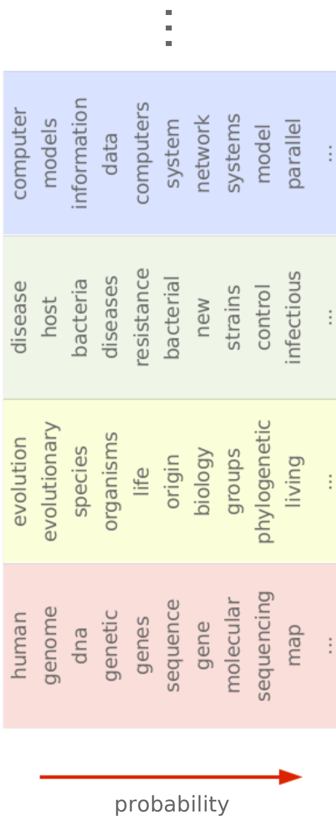
probability

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
...	...	...	...

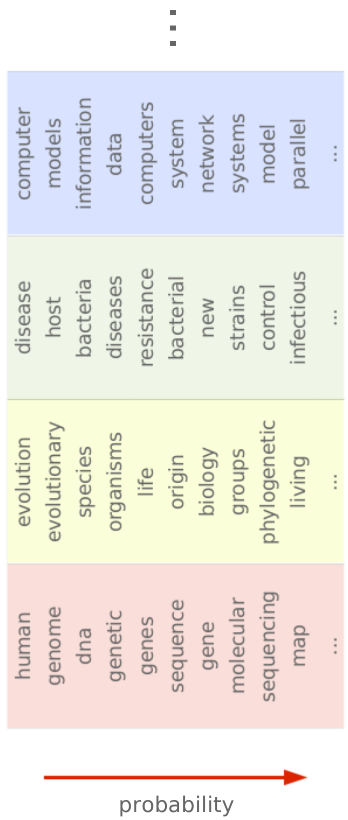
Generative Statistical Modeling

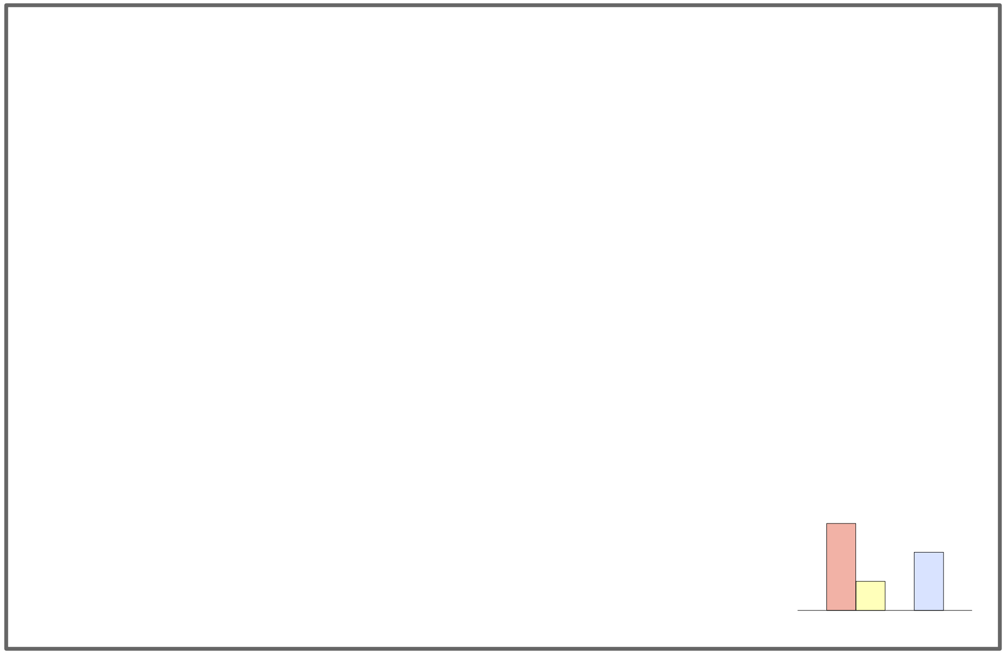
- Assume data was generated by a probabilistic model:
  - Model may have hidden structure (latent variables)
  - Model defines a joint distribution over all variables
  - Model parameters are unknown
- Infer hidden structure and model parameters from data
- Situate new data into estimated model

Generative Process

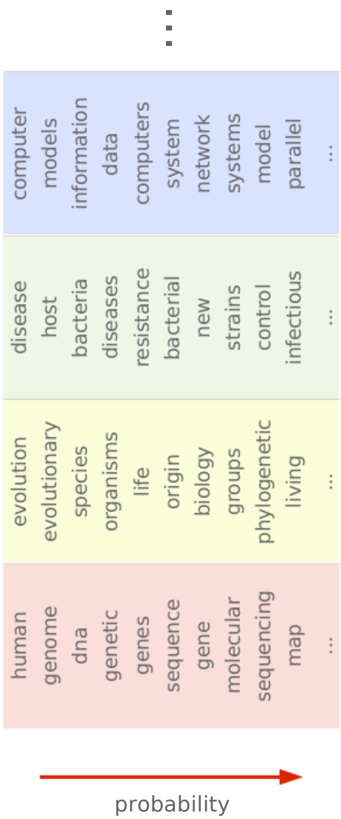


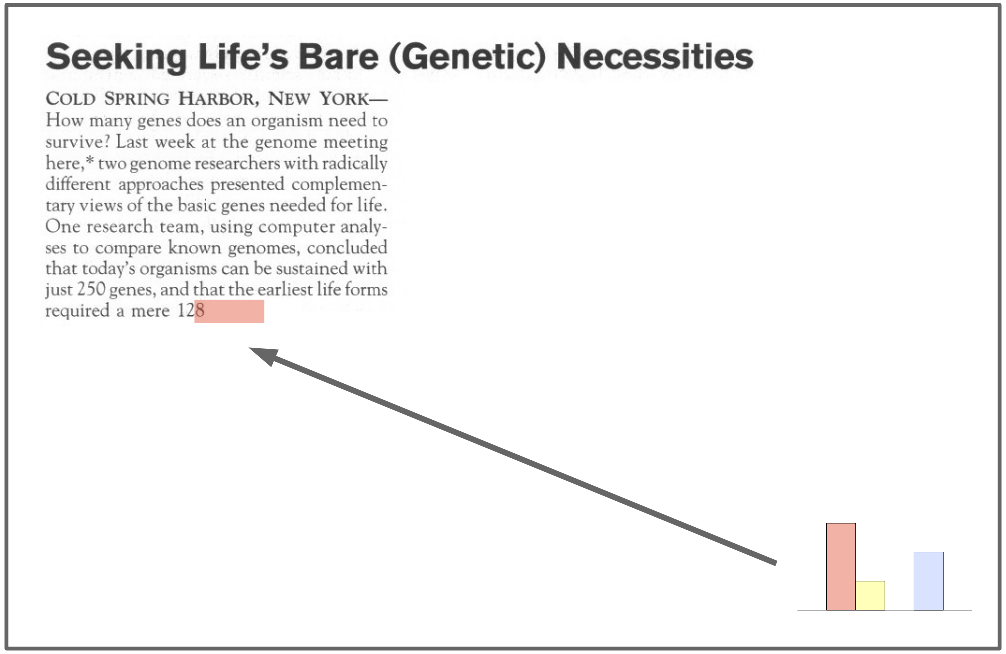
Choose a Distribution Over Topics



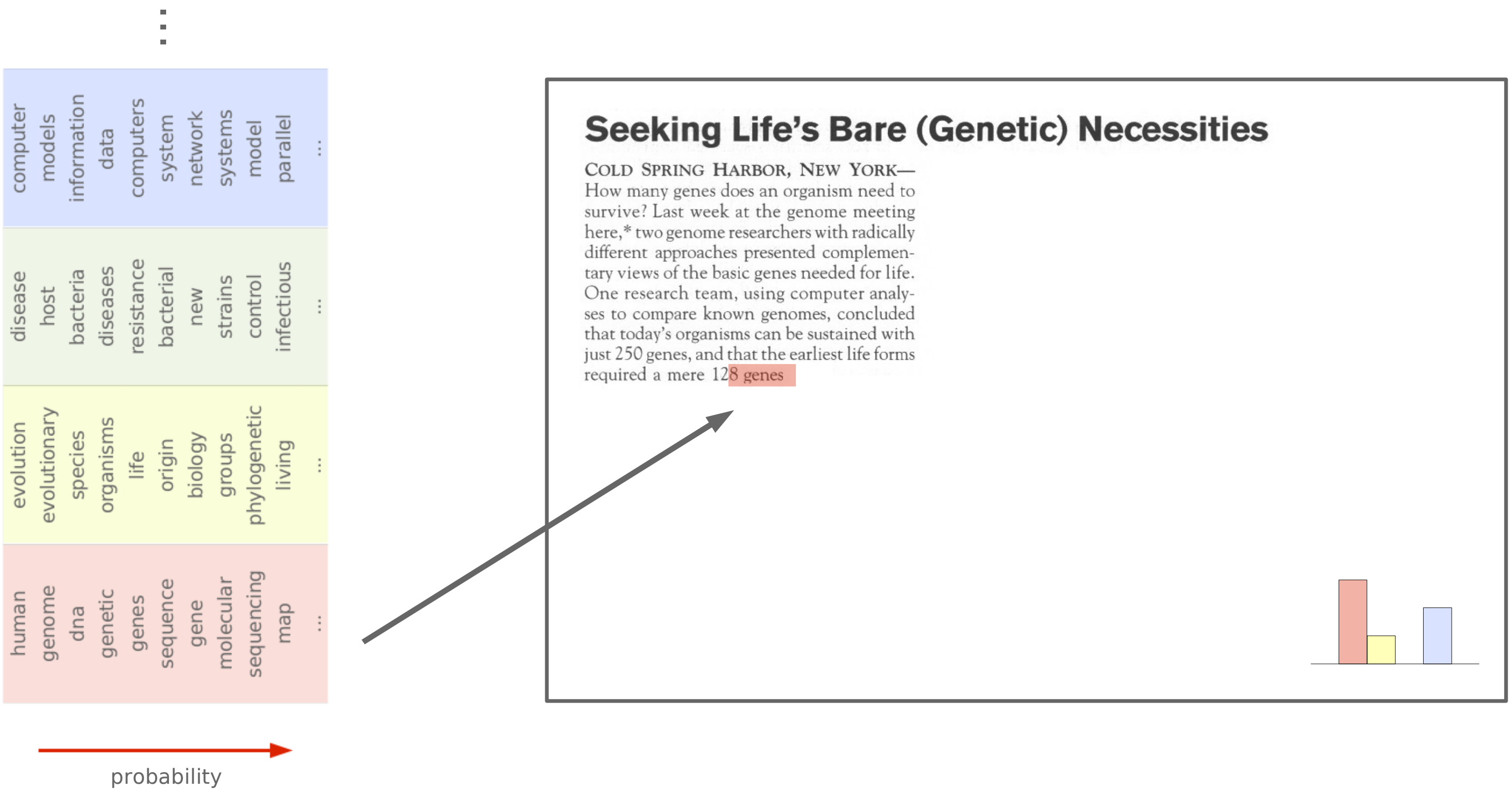


Choose a Topic

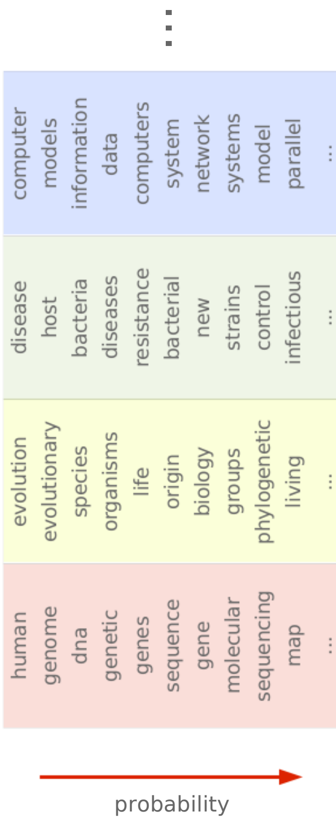


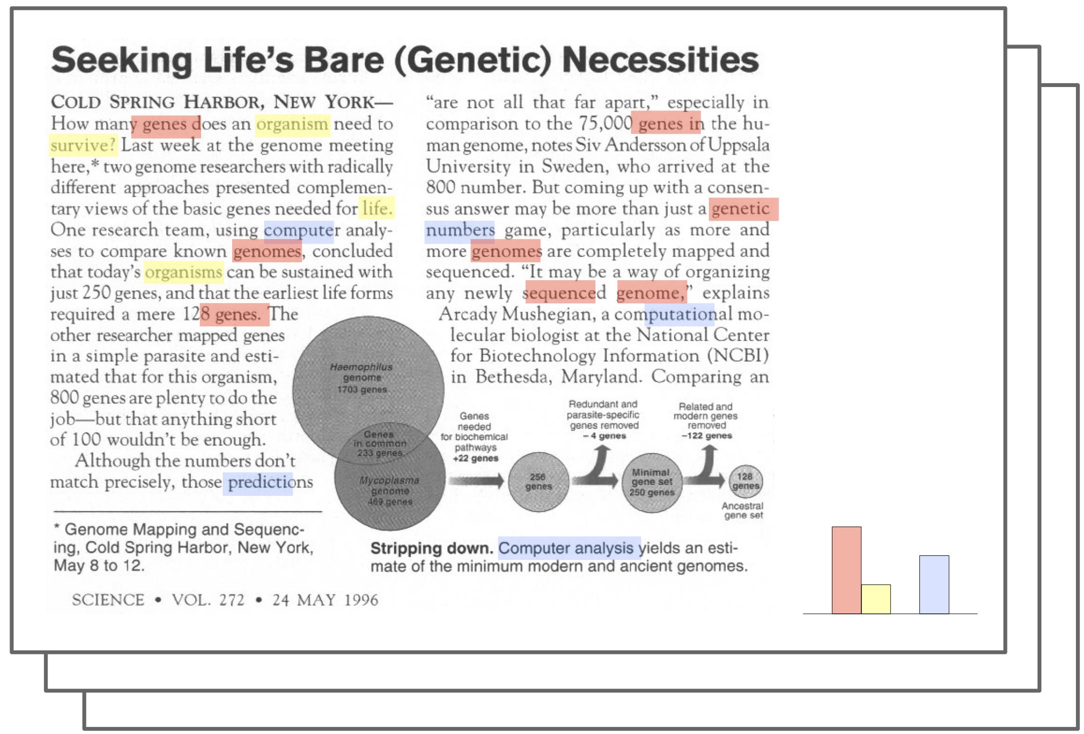


Choose a Word



… 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.

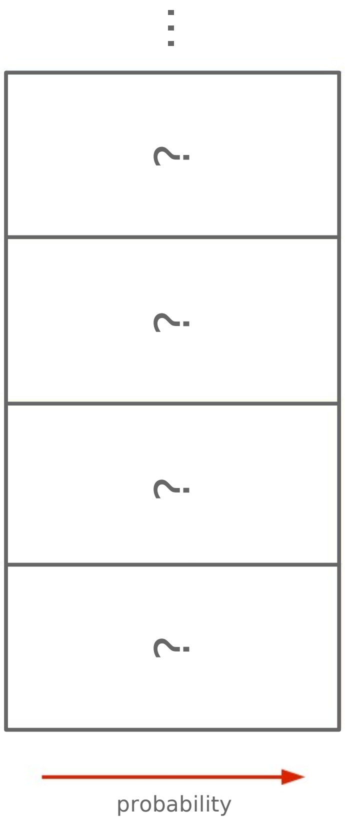
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

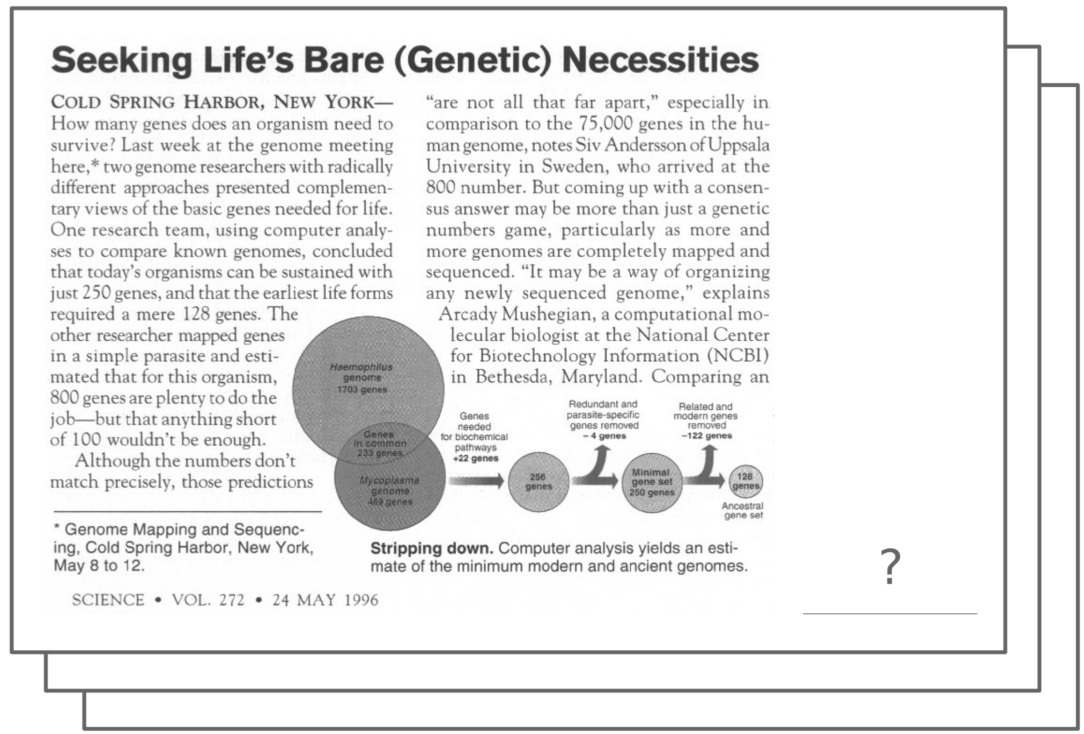
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* 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.

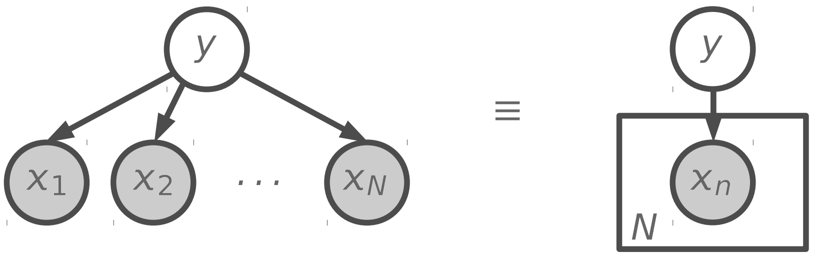
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Directed Graphical Models

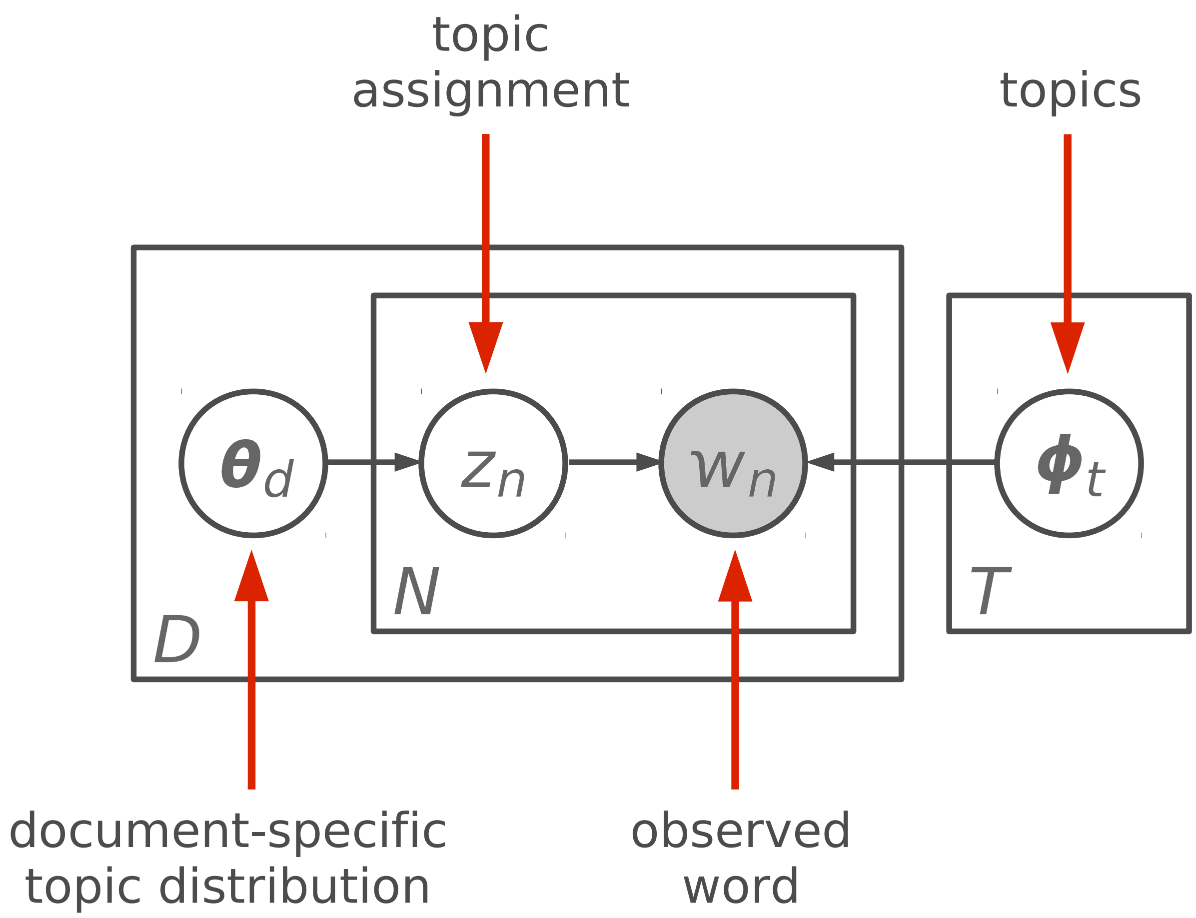
$$P(y, x_1, \ldots, x_N) = P(y) \prod_{n=1}^{N} P(x_n \mid y)$$

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



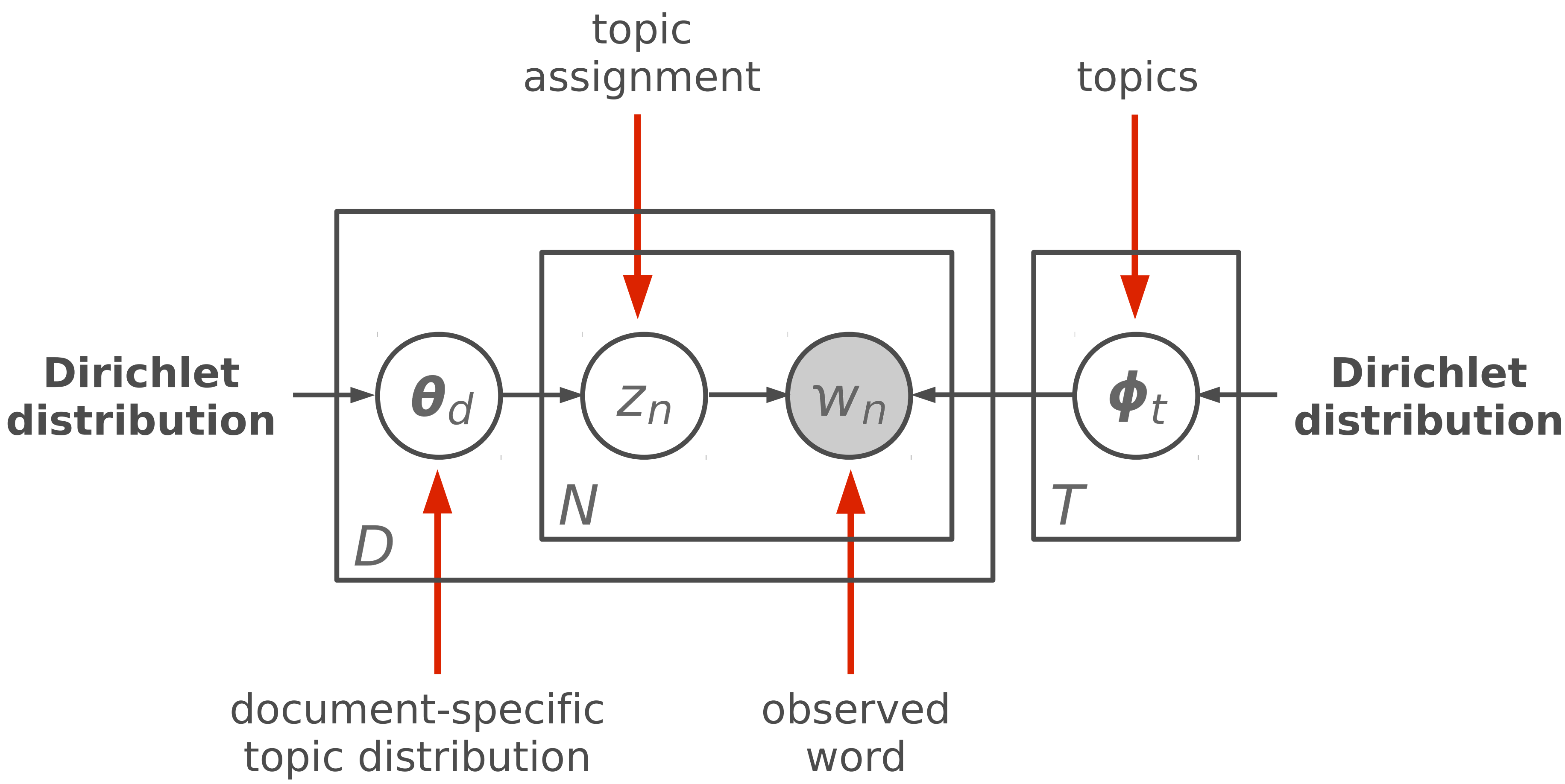
Statistical Topic Modeling

[Hofmann, '99]



Latent Dirichlet Allocation (LDA)

[Blei, Ng & Jordan, '03]



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

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH

“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 ...”

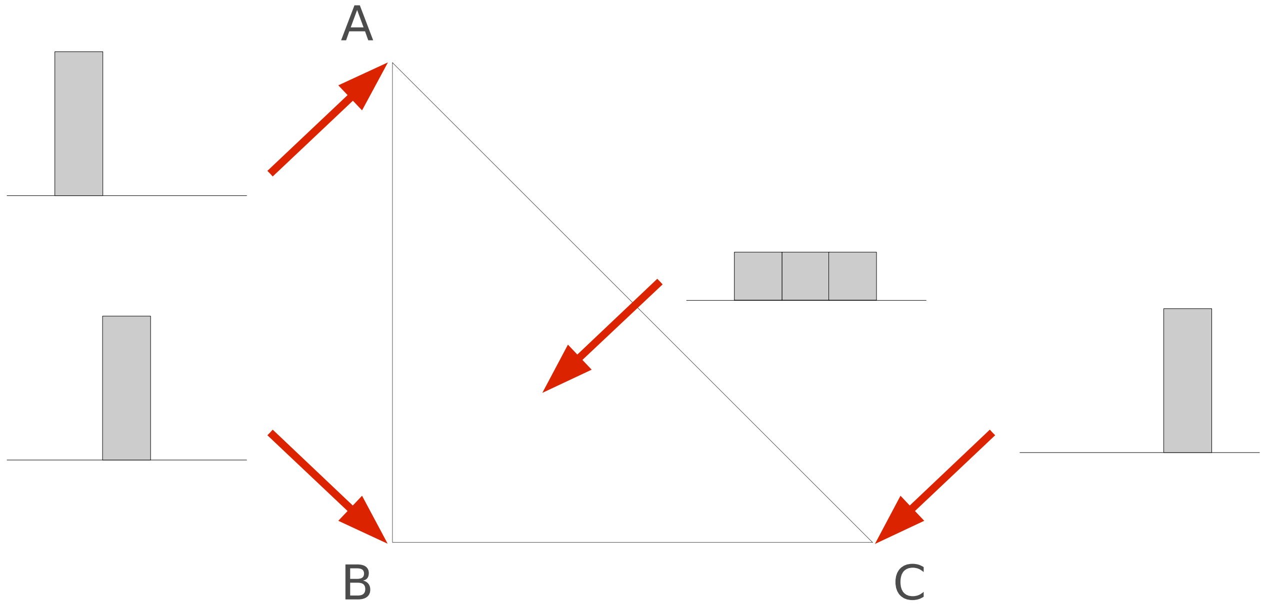
Make a domain-specific list of stop words...

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

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

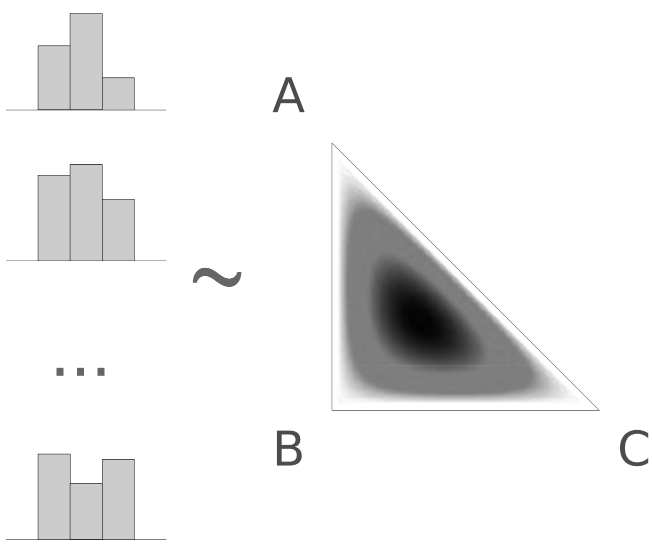
Discrete Probability Distributions

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



Dirichlet Distribution

- Distribution over discrete probability distributions:

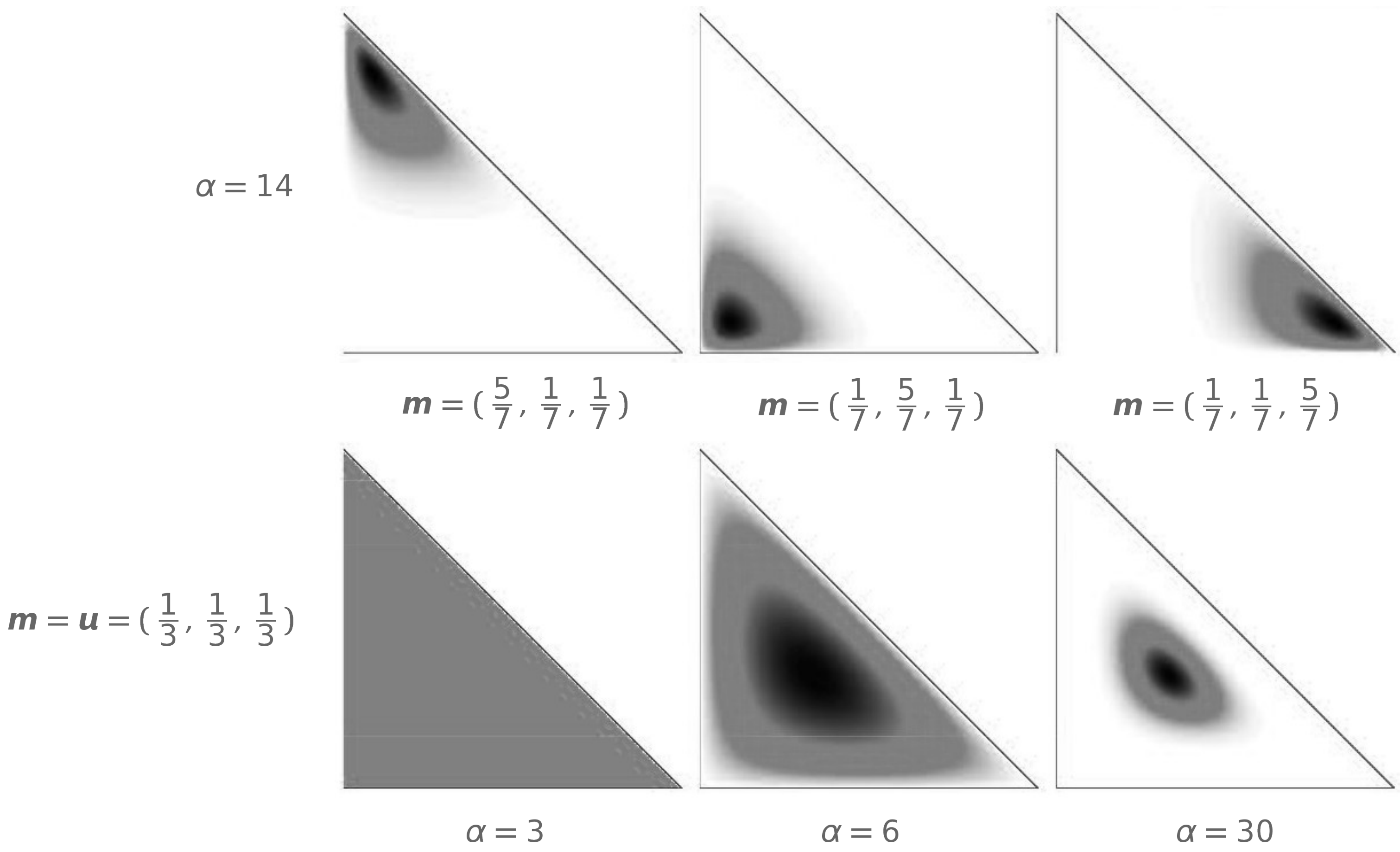


base measure (mean)

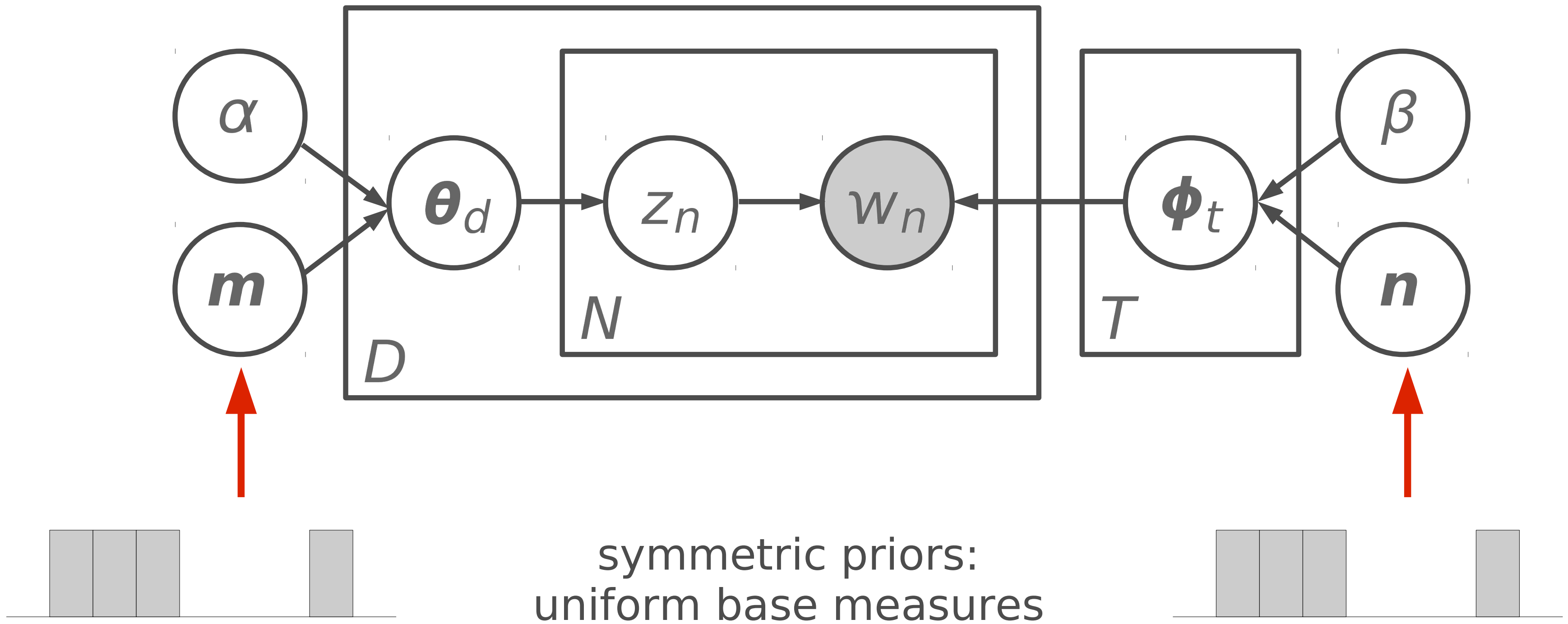
$$\equiv \quad \boldsymbol{p} \sim \mathrm{Dir}\,(\alpha \boldsymbol{m})$$

concentration parameter

Dirichlet Parameters



Dirichlet Priors for LDA



symmetric priors:
uniform base measures

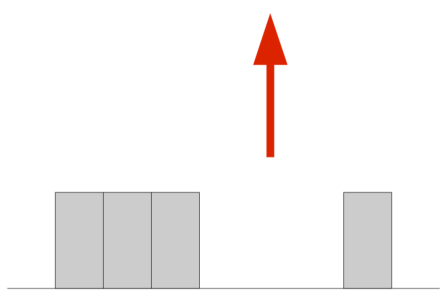
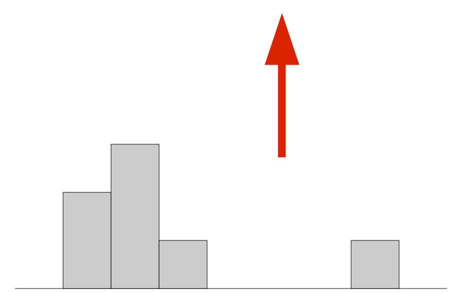
Dirichlet Priors for LDA

- 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

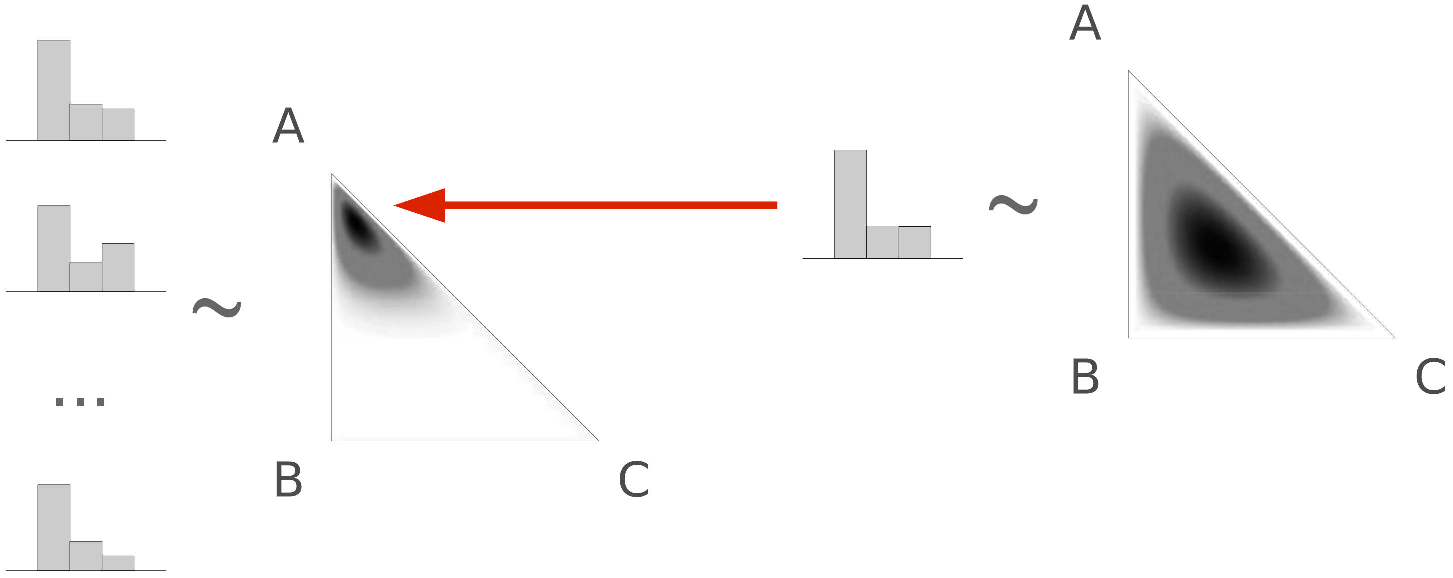
$\Theta \sim \mathrm{Dir}(\alpha \boldsymbol{m})$ $\Theta \sim \mathrm{Dir}(\alpha \boldsymbol{m})$

- Asymmetric prior: some topics more likely a priori

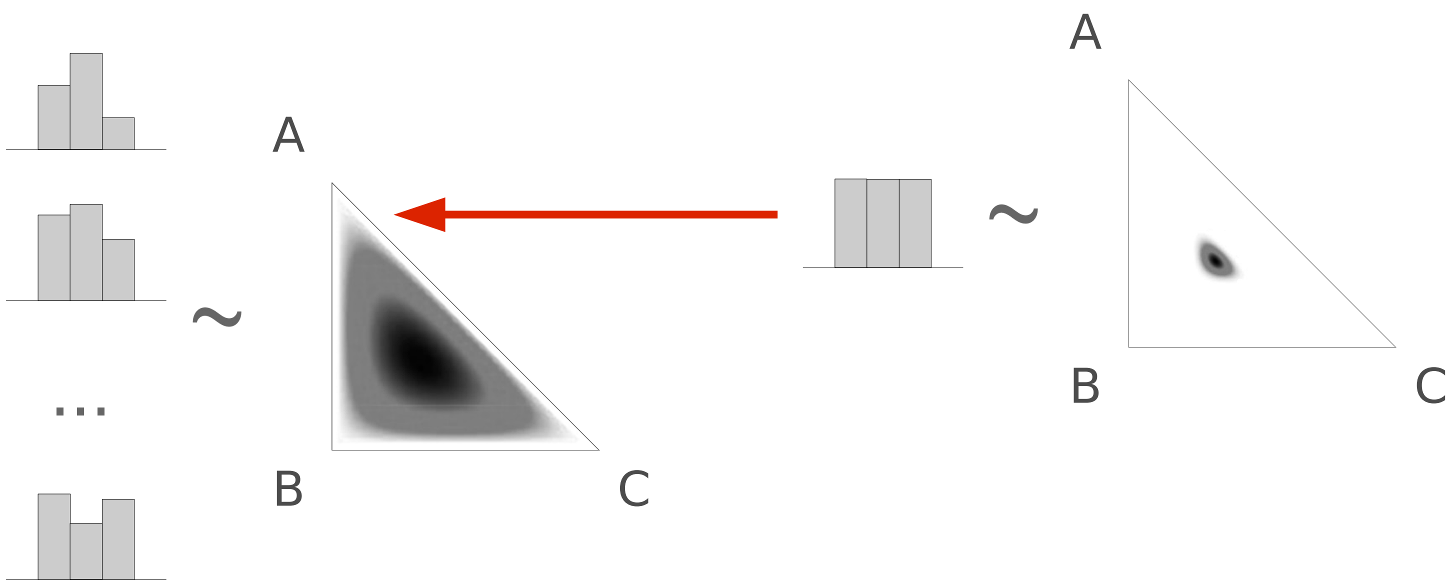
Hierarchical Asymmetric Dirichlet

- Which topics should be more probable a priori?
  - Draw ***m*** from a Dirichlet distribution:

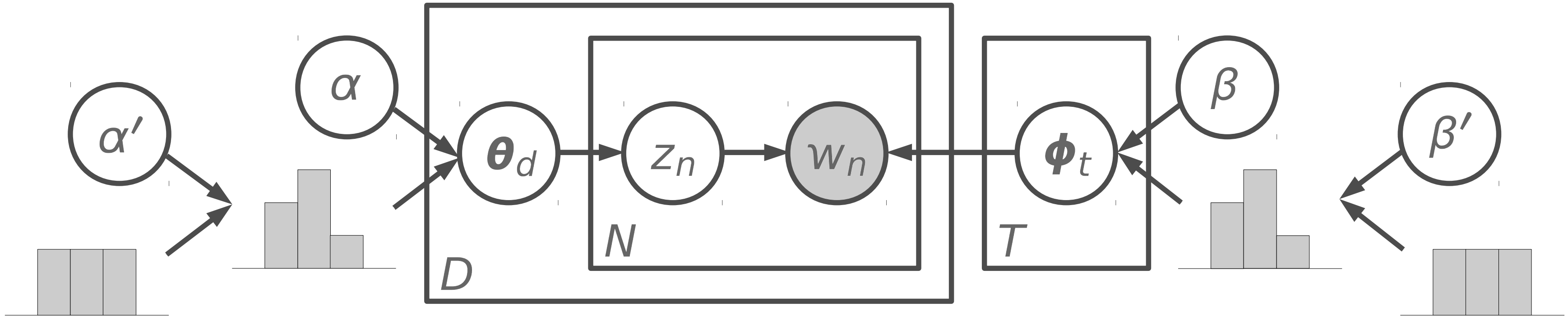


A Theoretical Observation...

- Symmetric Dirichlet is a special case of the hierarchical asymmetric Dirichlet (large concentration parameter)



Putting Everything Together



- Asymmetric hierarchical Dirichlet priors
- Integrate out Θ, Φ and base measures
- Learn $\boldsymbol{z}$ and concentration parameters from data

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)

Inferred Topics

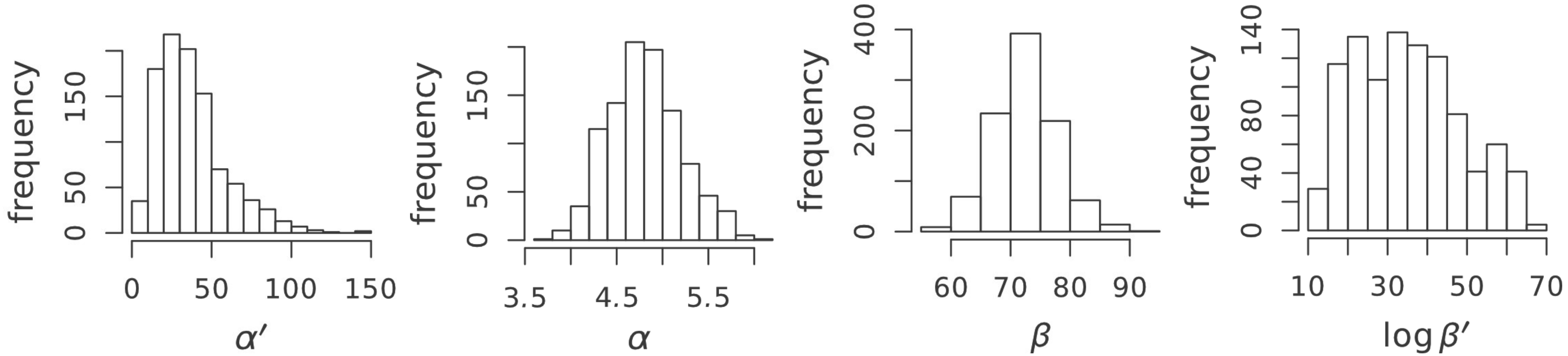
before

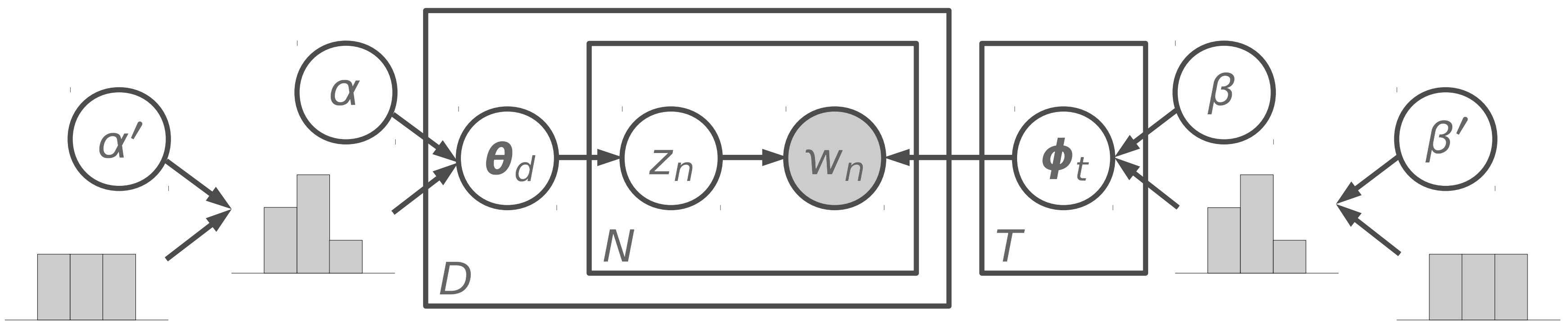
a	a	the	the
field	the	of	**invention**
emission	**carbon**	a	of
an	and	to	to
electron	**gas**	and	**present**
...	...	...	...

after

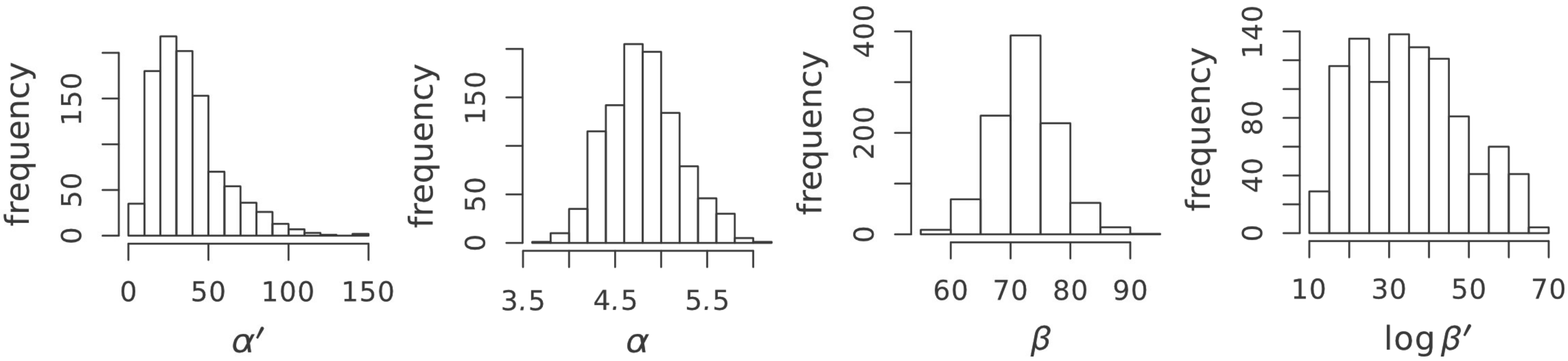
the	**carbon**	**metal**	**composite**
a	**nanotubes**	**catalytic**	**polymer**
of	**nanotube**	**transition**	**matrix**
to	**catalyst**	**catalyst**	**weight**
and	**substrate**	from	**fiber**
...	...	...	...

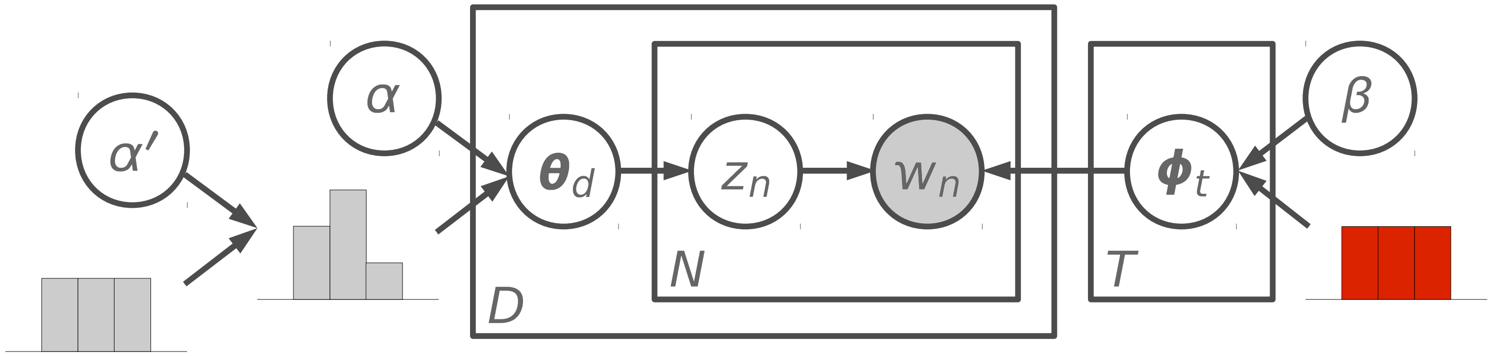
Sampled Concentration Parameters





Sampled Concentration Parameters





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

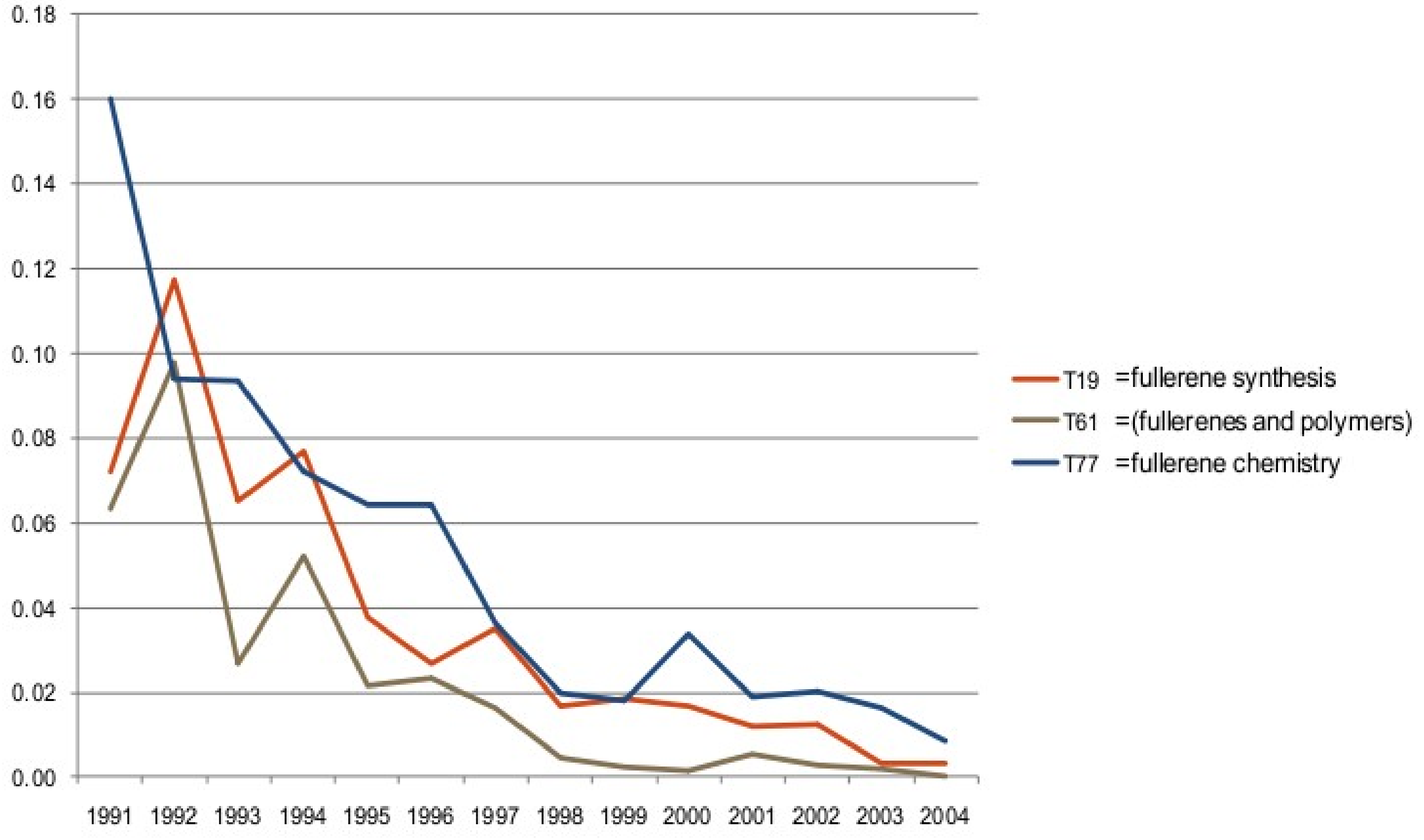
“Off-the-Shelf” Topic Modeling

I can model technology emergence by analyzing patent abstracts!

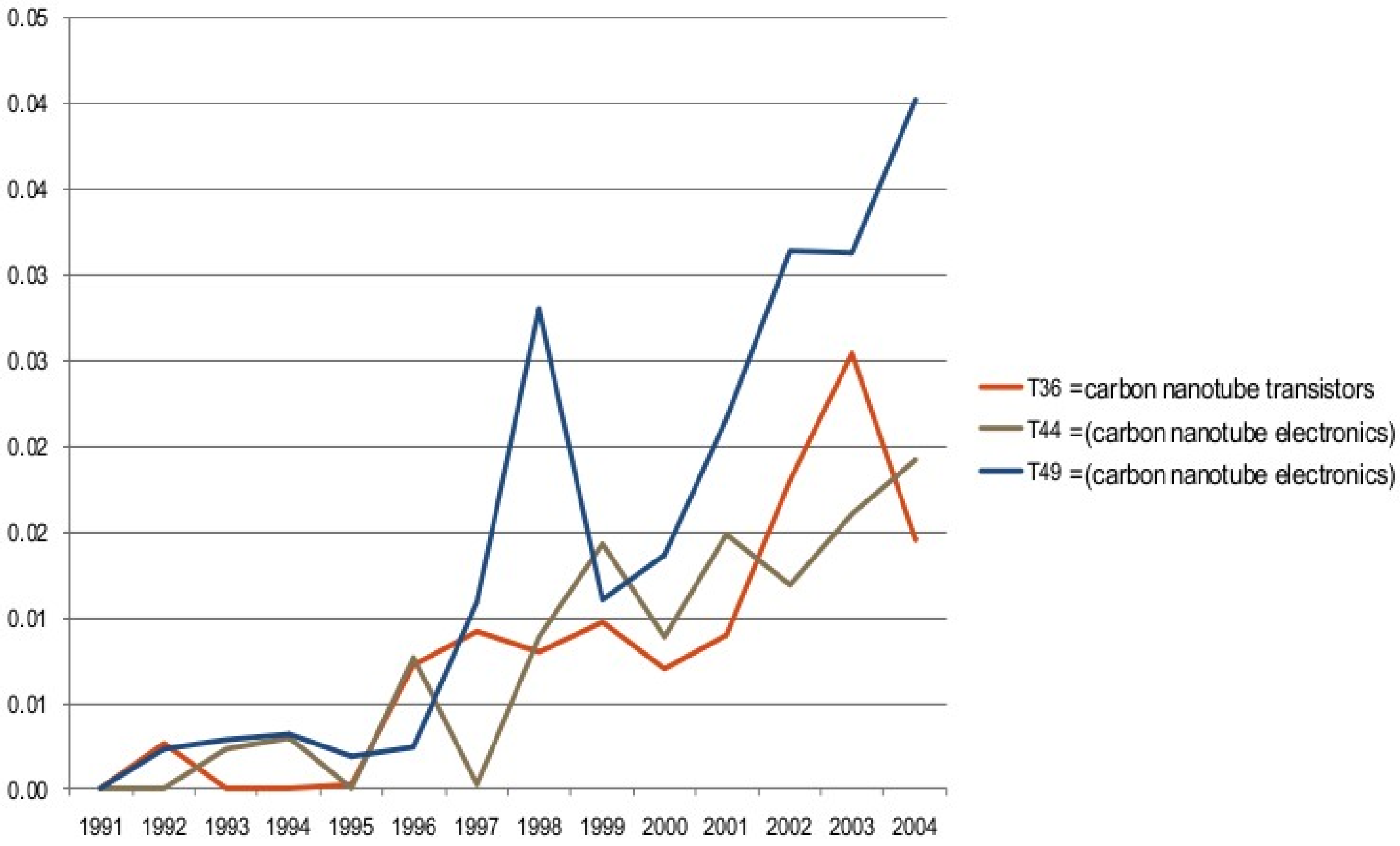
Great! Let me know if you need any more help!

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

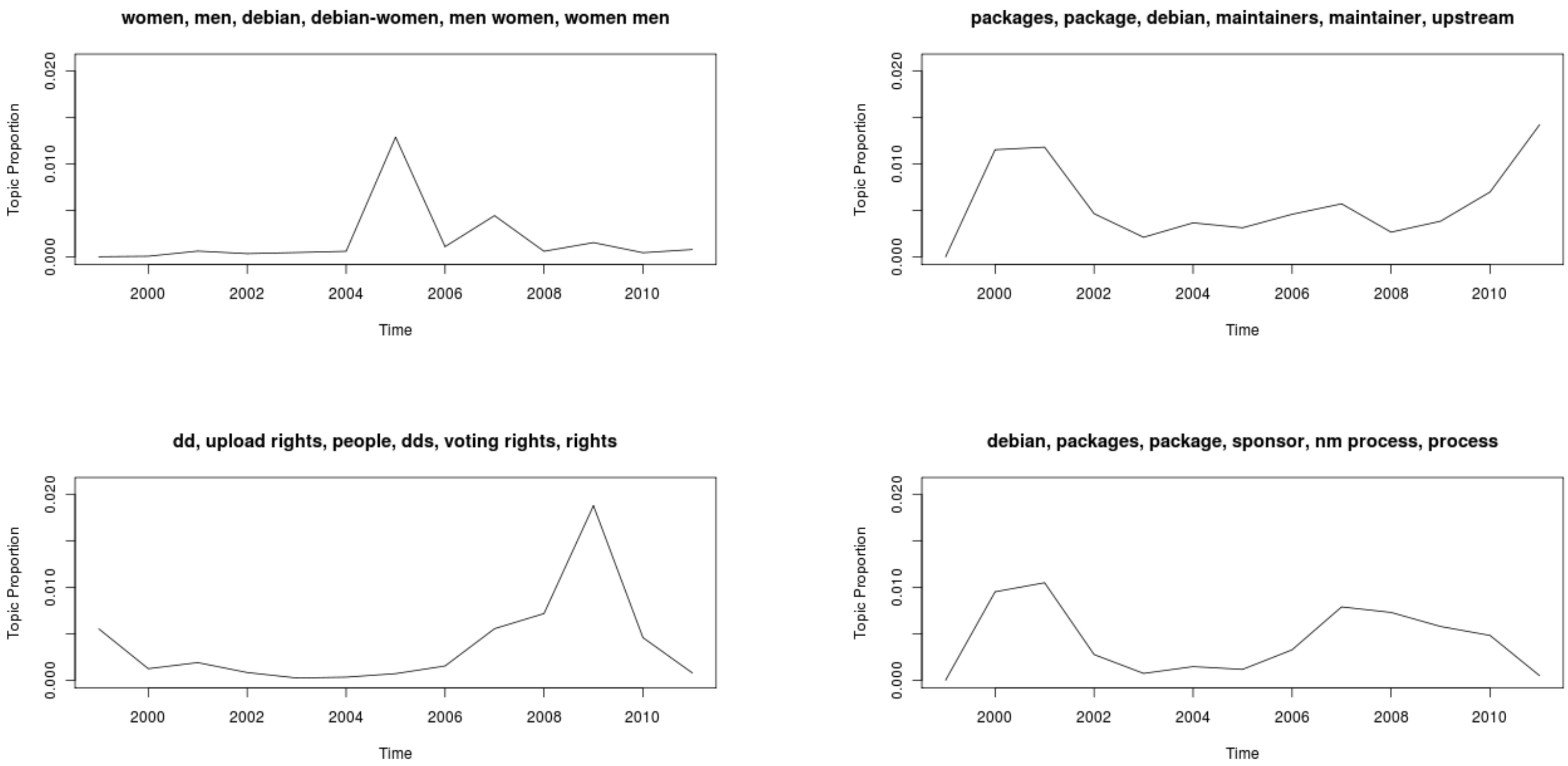
Declining Topics



Rising Topics



Analyzing Debian Mailing Lists



Building Other Tools

- Topic-based language modeling [Wallach, ICML '06]
  - Predict the next word given previous words
  - Have to model stop words
- Polylingual topic modeling [Mimno et al., EMNLP '09]
  - Track scientific progress in other countries
  - Simultaneously model text in many languages
  - Need robustness to word usage in many languages

This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH

Evaluating Topic Models

- Topic models are unsupervised so evaluation is hard
- A lot of topic modeling research has skirted this issue
- Easy to get a sense of topics from “eyeballing” output
  - … but this isn't rigorous evaluation
- One common evaluation metric is the probability of held-out documents [Wallach et al., ICML '09]
- Also need expert-driven evaluation

Expert-Driven Evaluation

- Scientific policy-makers know their own domains
- Invaluable resource for model evaluation:
  - Identification of good/poor quality topics
  - Characterization of different types of topics
- Collaborative research:
  - Automated evaluation metrics
  - Prior distributions that influence model output

Evaluation of NIH Topics

- 2 experts from NIH, 150 topics (NINDS coverage)
- Collaboratively developed 3-stage evaluation protocol
- 4 classes of poor quality topics:
  - Intruded: 2 or more unrelated concepts
  - Chained: e.g., “fatty acids” → “acids” → “nucleic acids”
  - Unbalanced: mix of general and specific terms
  - Random: no clear concept represented

Evaluation Metrics

- Number of words assigned to each topic (topic size)
- Within-document co-occurrence of the top words

Intruded	Chained
sleep	cerebellar
sars	cerebellum
insomnia	pb
cov	purkinje
disturbances	ag
...	...

cerebellar	1149	499	1	318	2
cerebellum	499	1283	2	228	1
pb	1	2	372	0	3
purkinje	318	228	0	479	0
ag	2	1	3	0	1321
cell	269	248	55	253	198

Automated Evaluation

- Word co-occurrence-based metric:
  - 17 of 20 worst-scoring topics are “bad”
  - 18 of 20 best-scoring topics are “good”
- Goal: incorporate co-occurrence information directly into the model to prevent poor quality topics:
  - Words that do not co-occur in documents should not have high probability within the a single topic

Generalized Polya Urns

- The topic–word component of LDA is a Polya urn
- Can be replaced with a generalized Polya urn
  - Can then incorporate co-occurrence statistics directly into the model via the generalized Polya urn schema
- Relatively little computational cost beyond LDA
- Resultant topics are more coherent:
  - Much better evaluation scores (automated, humans)

Thanks!

Acknowledgements: Sarah Kaplan, University of Toronto; Andrew McCallum, University of Massachusetts Amherst; David Mimno, University of Massachusetts Amherst; Ned Talley, NIH