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: to ensure that science and technology provide the greatest possible benefit to society by identifying, funding, and managing high quality science
- Administrative, financial, and political actions
- Actions chosen to have impact on, e.g.,
  - Stimulating breakthrough research
  - Increasing economic prosperity
  - Broadening participation

Science Policy Actions

- Funding-related actions:
  - Using federal funds for research on human stem cells
  - “People not projects” vs. pre-defined deliverables
- Intellectual property actions:
  - Allowing patents claiming natural gene sequences
- Educational actions:
  - Providing outreach for under-represented groups
  - Running training and mentoring programs

Making Policy Decisions

- Responsibility of many federal government organizations
- … but also involves private sector, education
- Core tasks: identify actions, estimate impact, understand outcomes

The Problem...

"A time-consuming, domain-specific, expert-intensive process, frequently done under severe time constraints without a systematic, reproducible audit trail or bias control using limited tools against an overwhelming information deluge."

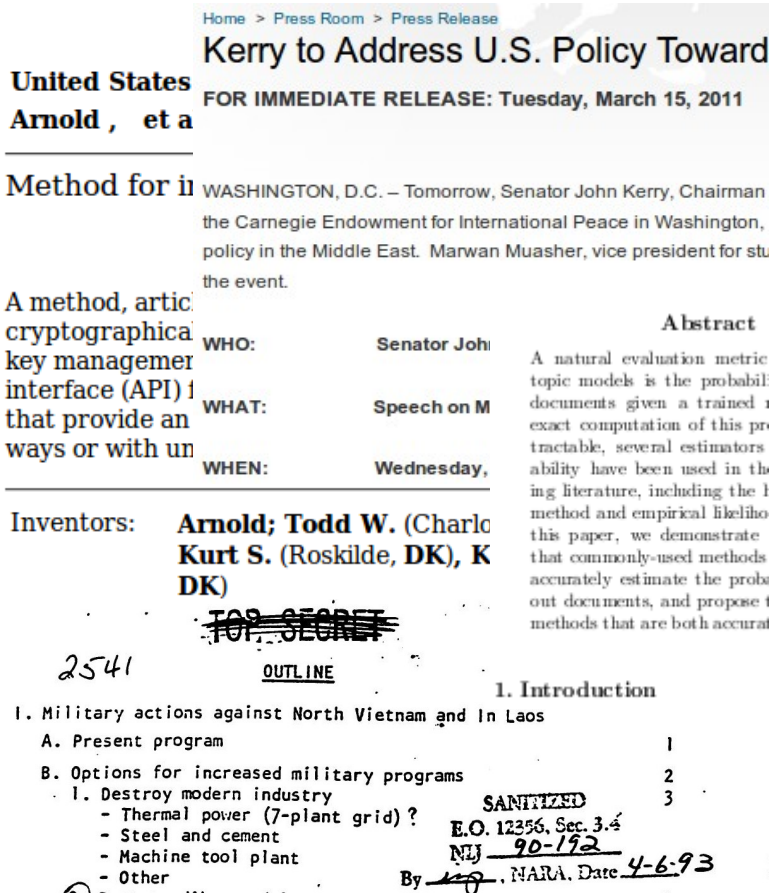
— Dewey Murdick, IARPA, 2010

Data-Driven Decision-Making

“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

Text as Data



- Structured and formal: e.g., publications, patents, press releases
- Messy and unstructured: e.g., OCRed documents, transcripts, blog posts

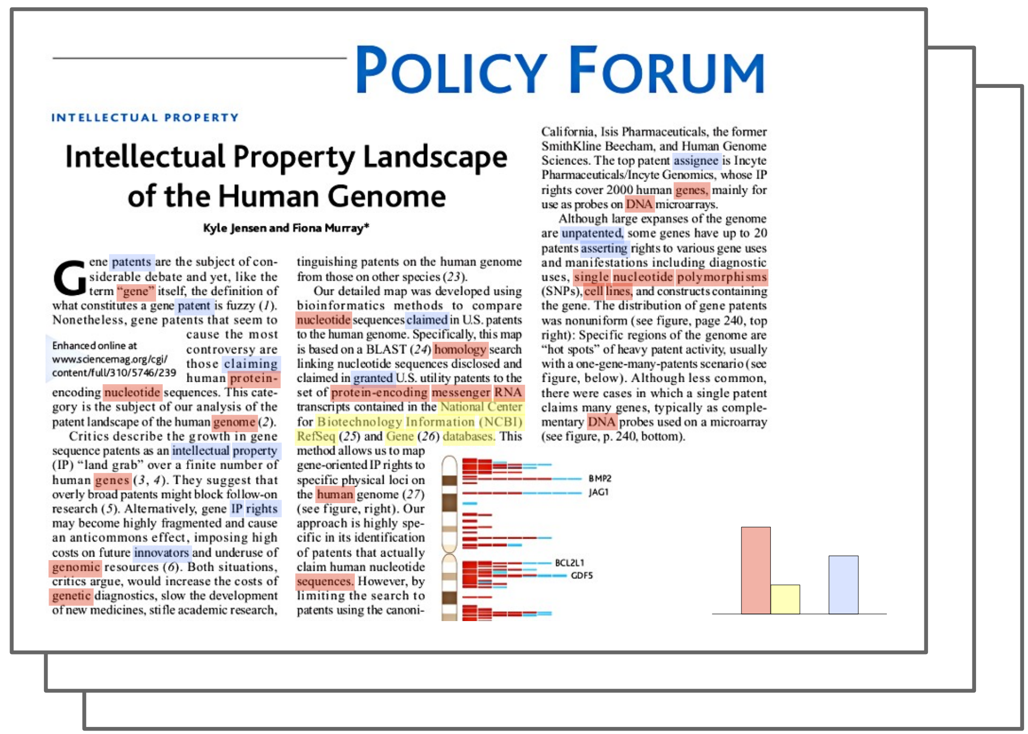
⇒ Large-scale, robust methods for analyzing text

Statistical Topic Models

- Goal: large-scale, exploratory data analysis:
  - “What do these data tell us that we don't already know?”
  - i.e., characterizing the shape of the haystack
- Topic models excel at discovering hidden thematic structure in large, unstructured document collections
- Given a document collection, topic models
  - Learn the composition of the topics that best represent it
  - Learn which topics are used in each document

Latent Dirichlet Allocation (LDA)

[Blei, Ng & Jordan, '03]



POLICY FORUM

INTELLECTUAL PROPERTY

Intellectual Property Landscape of the Human Genome

Kyle Jensen and Fiona Murray*

Gene patents are the subject of considerable debate and yet, like the term "gene" itself, the definition of what constitutes a gene patent is fuzzy (1). Nonetheless, gene patents that seem to cause the most controversy are those claiming human protein-encoding nucleotide sequences. This category is the subject of our analysis of the patent landscape of the human genome (2).

Enhanced online at www.sciencemag.org/cgi/content/full/310/5746/239

Critics describe the growth in gene sequence patents as an intellectual property (IP) "land grab" over a finite number of human genes (3, 4). They suggest that overly broad patents might block follow-on research (5). Alternatively, gene IP rights may become highly fragmented and cause an anticommons effect, imposing high costs on future innovators and underuse of genomic resources (6). Both situations, critics argue, would increase the costs of genetic diagnostics, slow the development of new medicines, stifle academic research, tinguishing patents on the human genome from those on other species (23).

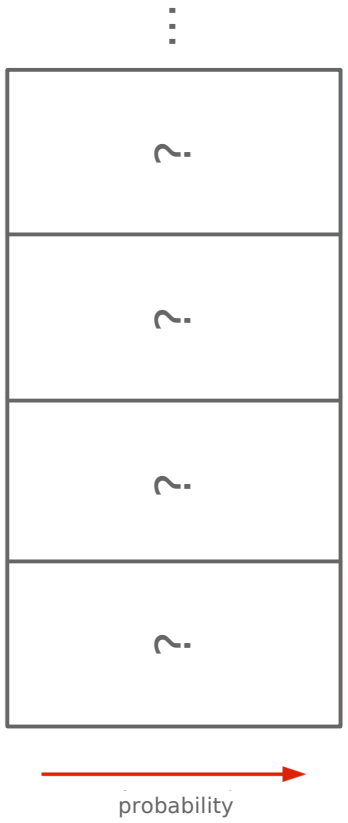
Our detailed map was developed using bioinformatics methods to compare nucleotide sequences claimed in U.S. patents to the human genome. Specifically, this map is based on a BLAST (24) homology search linking nucleotide sequences disclosed and claimed in granted U.S. utility patents to the set of protein-encoding messenger RNA transcripts contained in the National Center for Biotechnology Information (NCBI) RefSeq (25) and Gene (26) databases. This method allows us to map gene-oriented IP rights to specific physical loci on the human genome (27) (see figure, right). Our approach is highly specific in its identification of patents that actually claim human nucleotide sequences. However, by limiting the search to patents using the canoni-

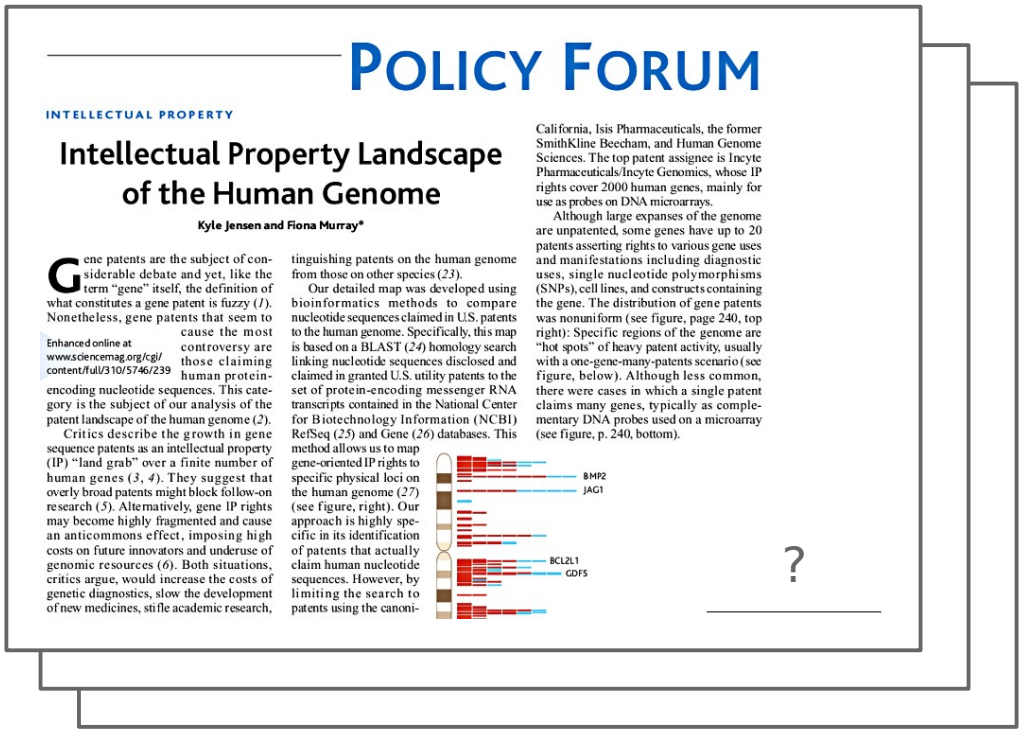
BMP2
JAG1
BCL2L1
GDF5

California, Isis Pharmaceuticals, the former SmithKline Beecham, and Human Genome Sciences. The top patent assignee is Incyte Pharmaceuticals/Incyte Genomics, whose IP rights cover 2000 human genes, mainly for use as probes on DNA microarrays.

Although large expanses of the genome are unpatented, some genes have up to 20 patents asserting rights to various gene uses and manifestations including diagnostic uses, single nucleotide polymorphisms (SNPs), cell lines, and constructs containing the gene. The distribution of gene patents was nonuniform (see figure, page 240, top right): Specific regions of the genome are "hot spots" of heavy patent activity, usually with a one-gene-many-patents scenario (see figure, below). Although less common, there were cases in which a single patent claims many genes, typically as complementary DNA probes used on a microarray (see figure, p. 240, bottom).

Real Data: Statistical Inference





POLICY FORUM

INTELLECTUAL PROPERTY

Intellectual Property Landscape of the Human Genome

Kyle Jensen and Fiona Murray*

Gene patents are the subject of considerable debate and yet, like the term "gene" itself, the definition of what constitutes a gene patent is fuzzy (*1*). Nonetheless, gene patents that seem to cause the most controversy are those claiming human protein-encoding nucleotide sequences. This category is the subject of our analysis of the patent landscape of the human genome (*2*).

Enhanced online at www.sciencemag.org/cgi/content/full/310/5746/239

Critics describe the growth in gene sequence patents as an intellectual property (IP) "land grab" over a finite number of human genes (*3*, *4*). They suggest that overly broad patents might block follow-on research (*5*). Alternatively, gene IP rights may become highly fragmented and cause an anticommons effect, imposing high costs on future innovators and underuse of genomic resources (*6*). Both situations, critics argue, would increase the costs of genetic diagnostics, slow the development of new medicines, stifle academic research, tinguishing patents on the human genome from those on other species (*23*).

Our detailed map was developed using bioinformatics methods to compare nucleotide sequences claimed in U.S. patents to the human genome. Specifically, this map is based on a BLAST (*24*) homology search linking nucleotide sequences disclosed and claimed in granted U.S. utility patents to the set of protein-encoding messenger RNA transcripts contained in the National Center for Biotechnology Information (NCBI) RefSeq (*25*) and Gene (*26*) databases. This method allows us to map gene-oriented IP rights to specific physical loci on the human genome (*27*) (see figure, right). Our approach is highly specific in its identification of patents that actually claim human nucleotide sequences. However, by limiting the search to patents using the canoni-

California, Isis Pharmaceuticals, the former SmithKline Beecham, and Human Genome Sciences. The top patent assignee is Incyte Pharmaceuticals/Incyte Genomics, whose IP rights cover 2000 human genes, mainly for use as probes on DNA microarrays.

Although large expanses of the genome are unpatented, some genes have up to 20 patents asserting rights to various gene uses and manifestations including diagnostic uses, single nucleotide polymorphisms (SNPs), cell lines, and constructs containing the gene. The distribution of gene patents was nonuniform (see figure, page 240, top right): Specific regions of the genome are "hot spots" of heavy patent activity, usually with a one-gene-many-patents scenario (see figure, below). Although less common, there were cases in which a single patent claims many genes, typically as complementary DNA probes used on a microarray (see figure, p. 240, bottom).

BMP2
JAG1
BCL2L1
GDF5

?

This Talk

- Statistical topic models for science policy-makers
- “Off-the-shelf” topic models: priors, stop words
- A database of National Institutes of Health grants

The Reality...

- Decision-makers are eager to use topic models as a strategic asset in their daily routines
- ... but topic models aren't always usable by non-experts
- Need to bridge this gap between producers and consumers of topic modeling technology:
  - Let practitioners' needs guide the research
  - Explore the interplay between theory and practice
  - Question unquestioned assumptions

“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 ...”

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

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

Preprocess your data to remove stop words...

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

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

Why It Matters

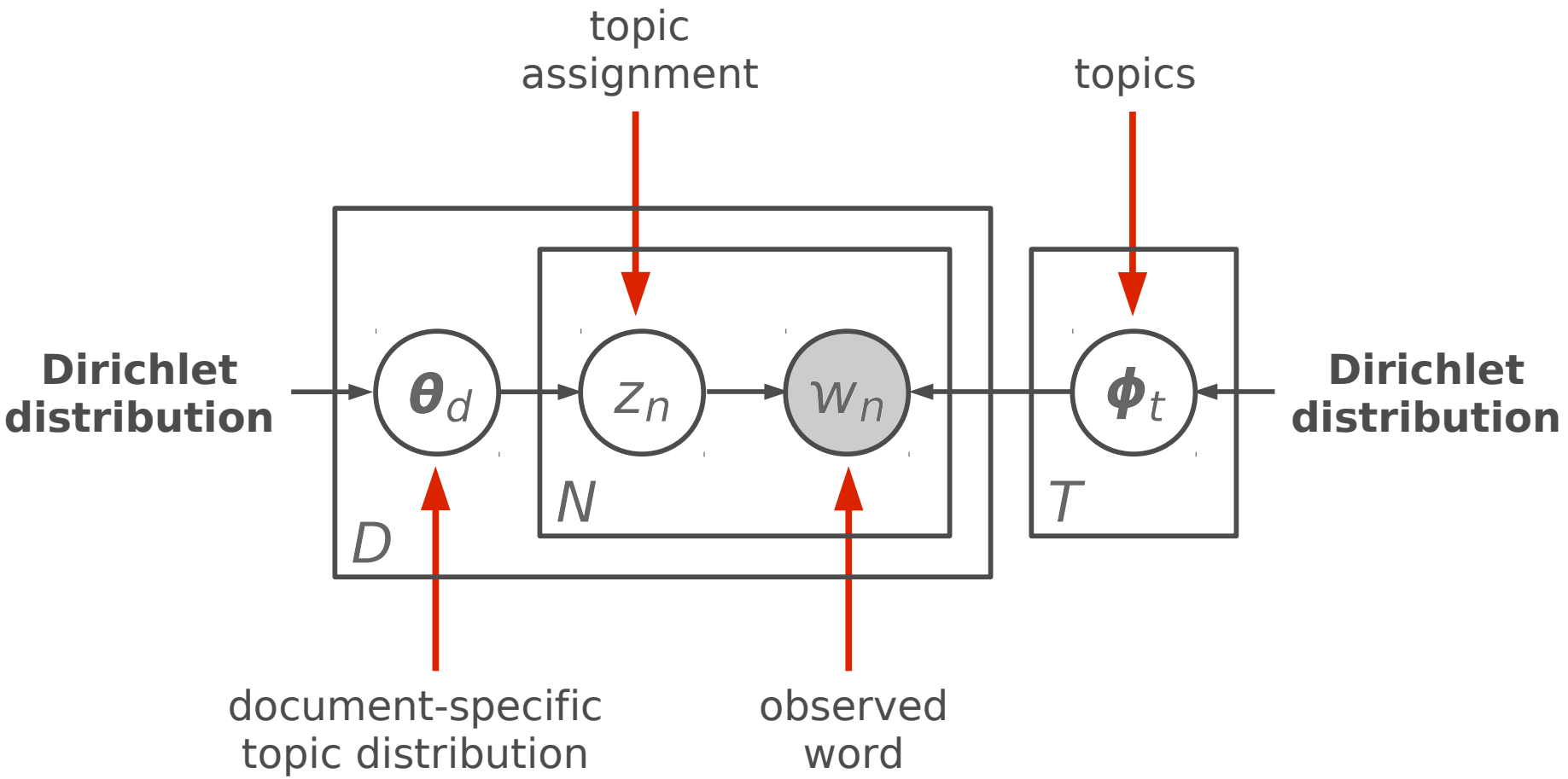
- More tolerant of mistakes if we understand them
- If we don't understand why mistakes occur, it's much harder to predict when they will occur:
  - Unpredictability → loss of confidence
- Goal: minimize pre-analysis effort:
  - Want to run systems without repeated intervention
  - Decision-makers are busy people with specialized skills whose time is better invested in post-analysis

This Talk

- Statistical topic models for science policy-makers
- “Off-the-shelf” topic models: priors, stop words
- A database of National Institutes of Health grants

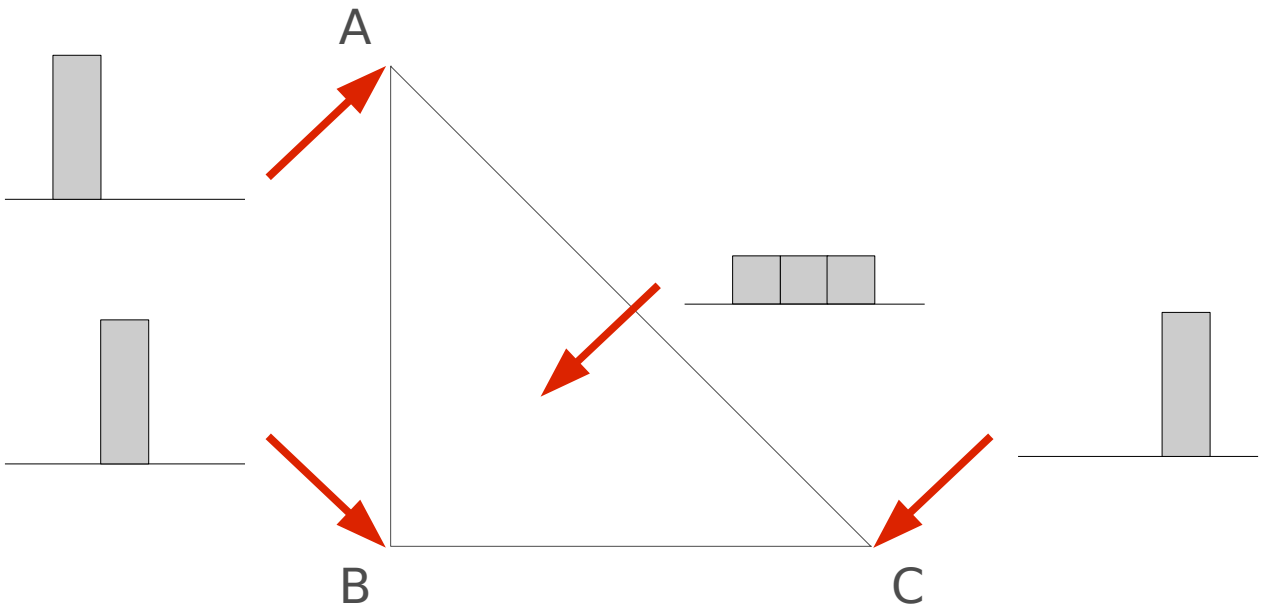
Latent Dirichlet Allocation (LDA)

[Blei, Ng & Jordan, '03]



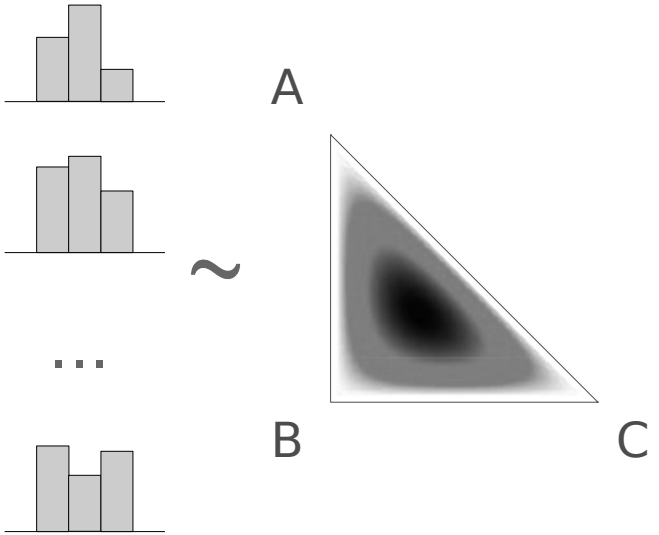
Discrete Probability Distributions

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



Dirichlet Distribution

- Distribution over discrete probability distributions:

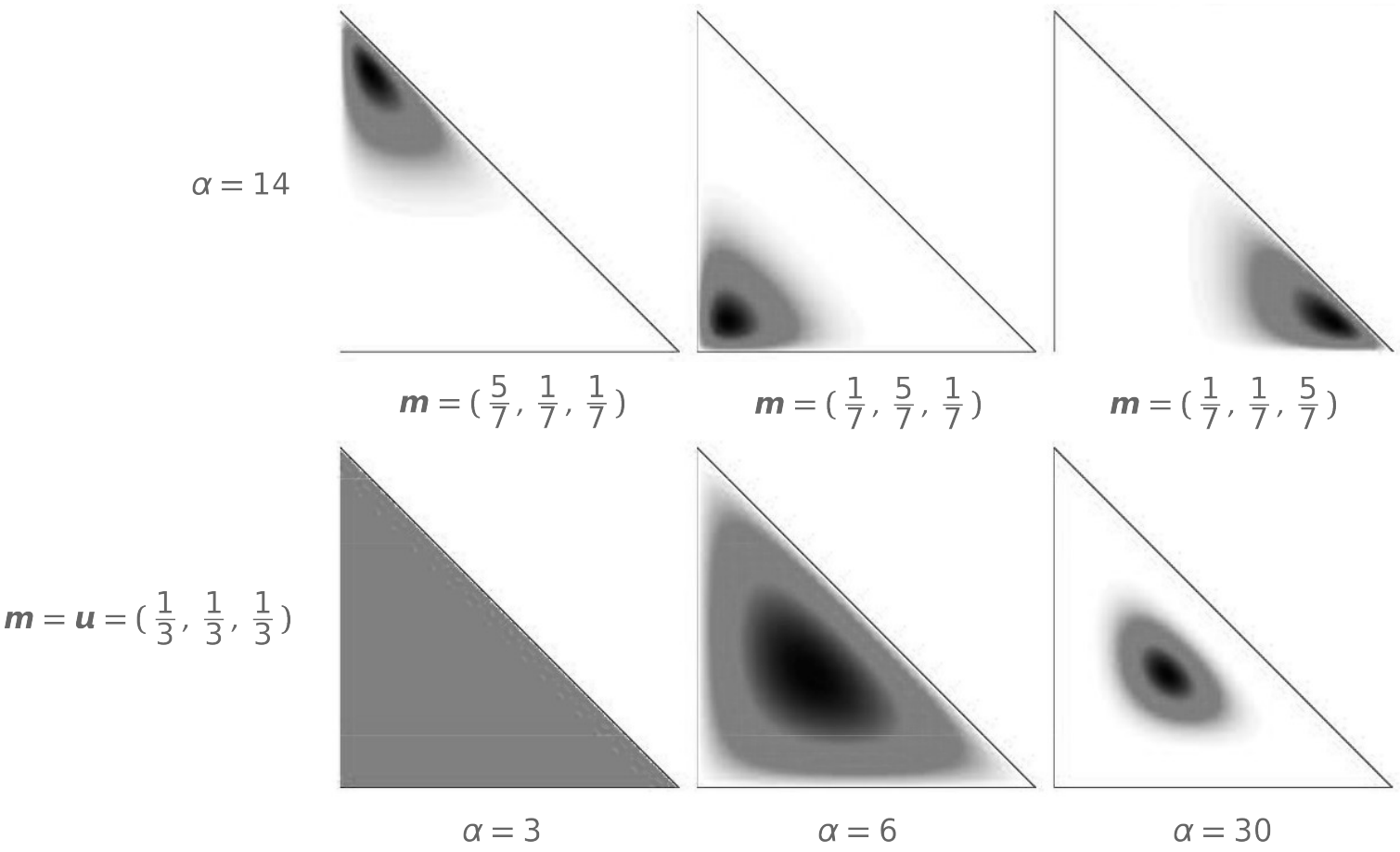


$\sim$ $\equiv$ $\boldsymbol{p} \sim \mathrm{Dir}(\alpha \boldsymbol{m})$

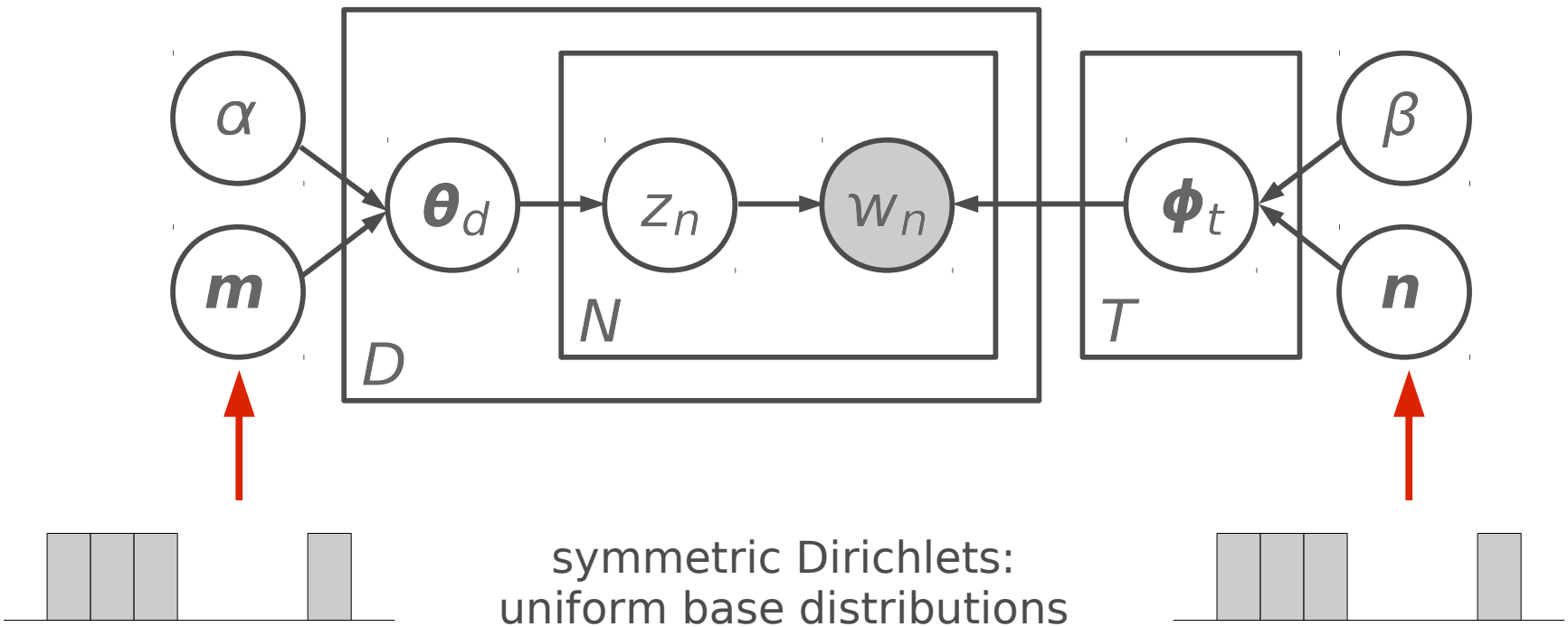
base distribution (mean)

concentration parameter

Dirichlet Parameters



Dirichlet Priors for LDA



Dirichlet Priors for LDA

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

Symmetric → Asymmetric

[Wallach et al., '09]

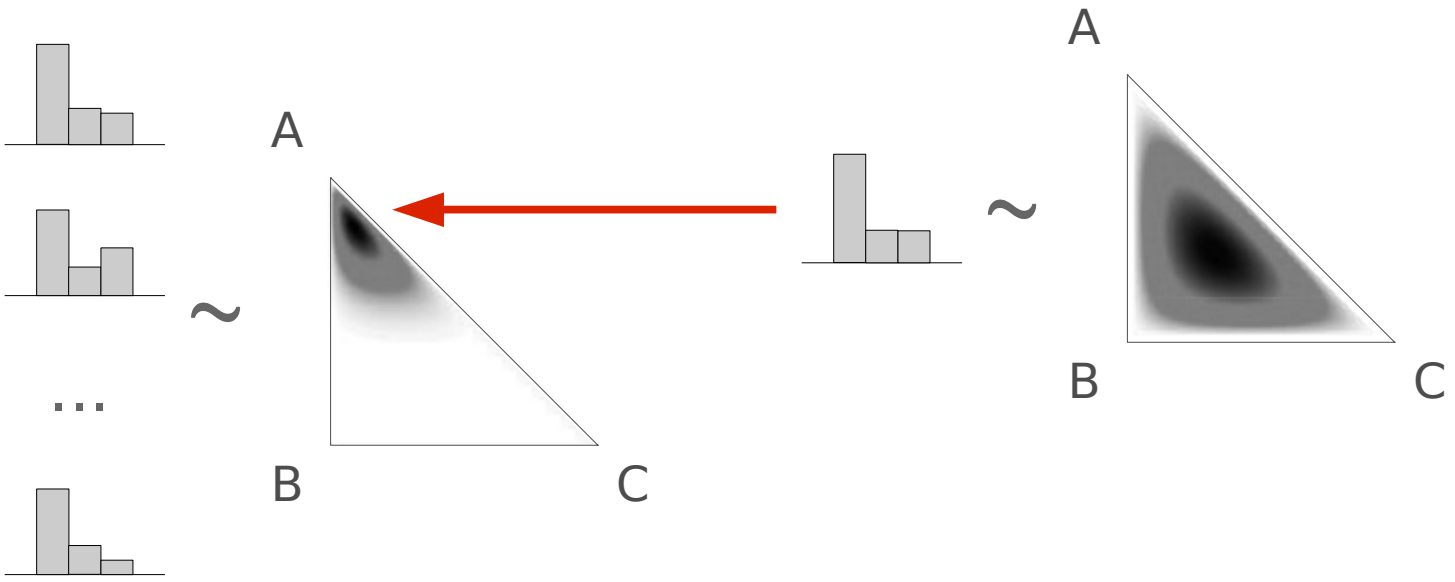
- Use prior over $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$ as a running example
- Uniform base distribution → nonuniform distribution

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

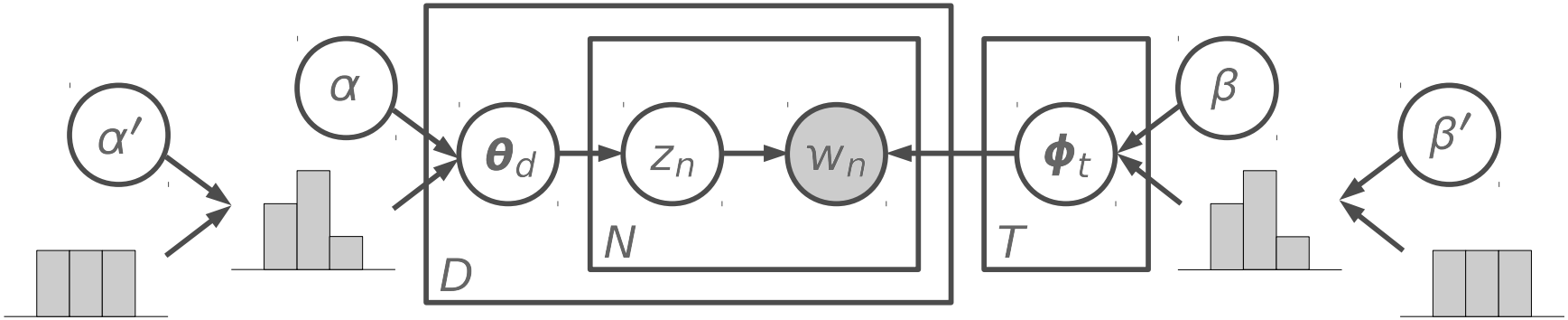
- Asymmetric prior: some topics more likely a priori

Hierarchical Asymmetric Dirichlet

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



Putting Everything Together



- Asymmetric hierarchical Dirichlet priors
- Integrate out Θ, Φ and base distributions
- 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

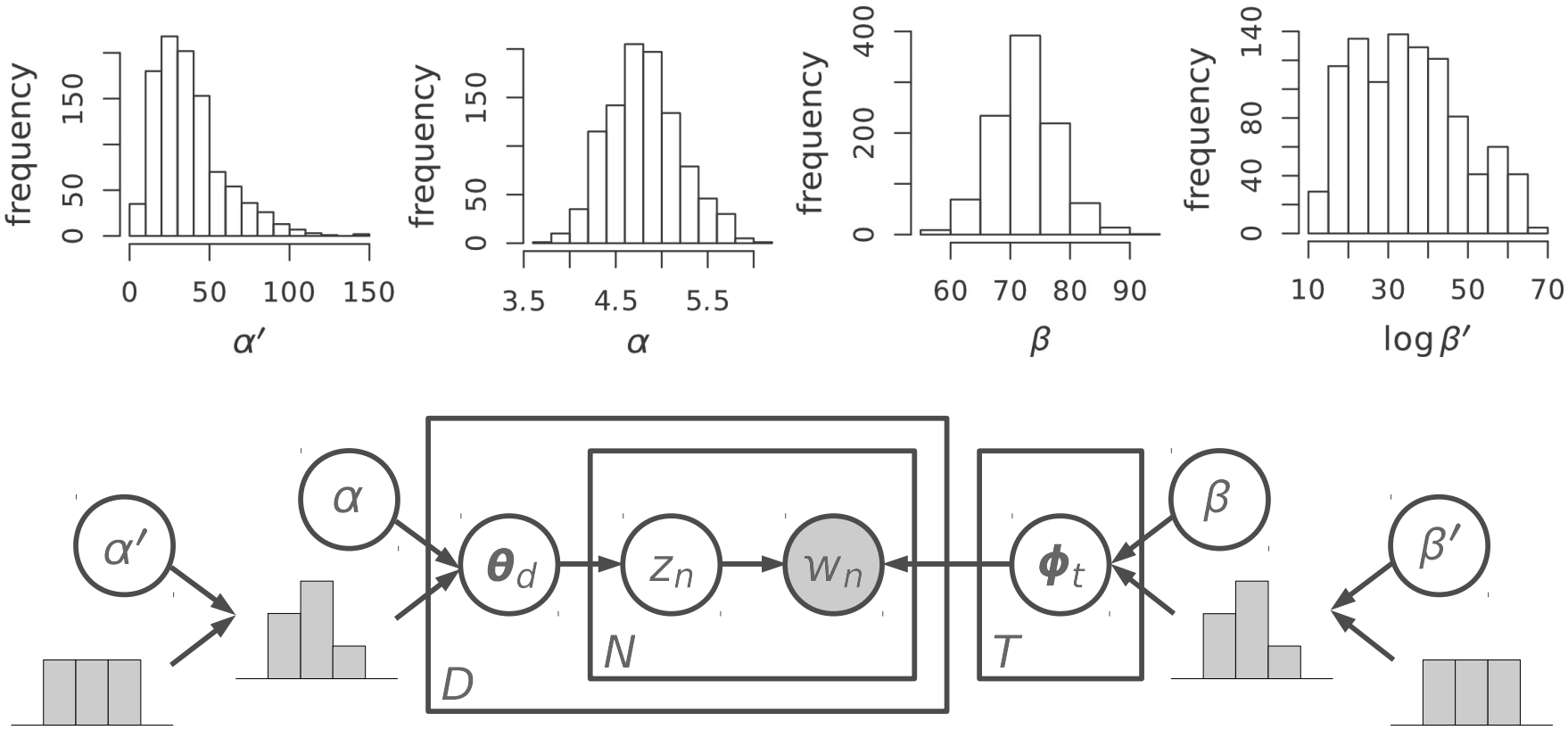
before →

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

after →

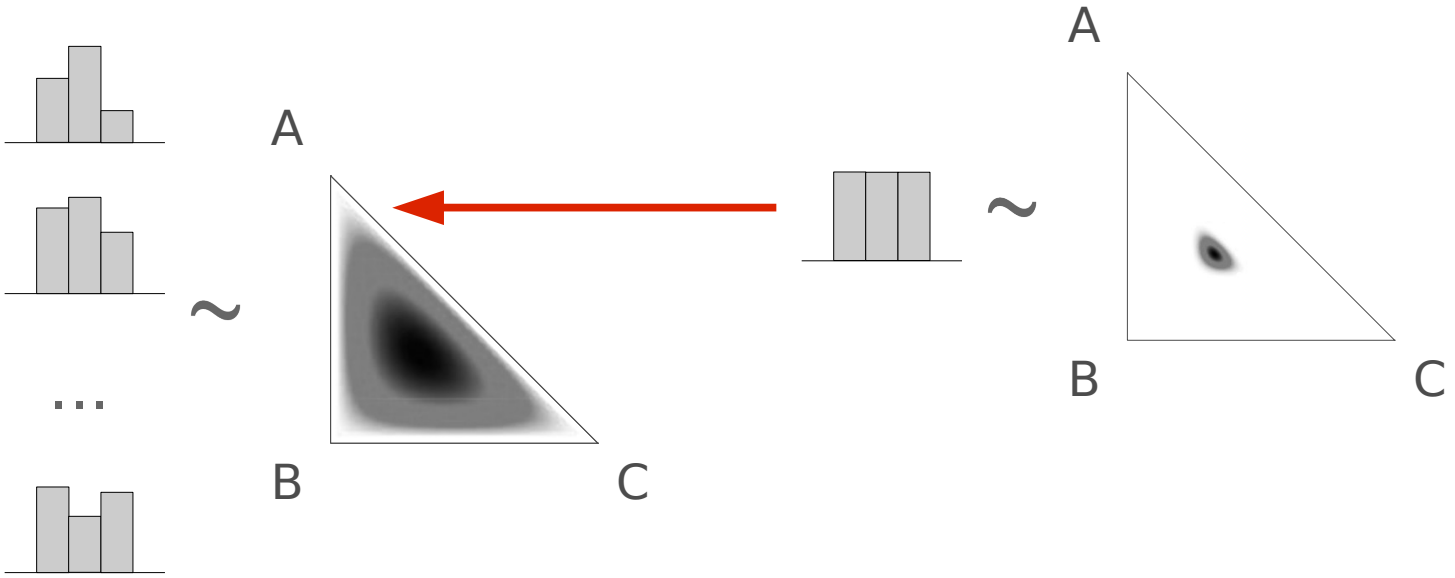
the	**carbon**	**metal**	**composite**
a	**nanotubes**	**catalytic**	**polymer**
of	**nanotube**	**transition**	**matrix**
to	**catalyst**	**catalyst**	**weight**
and	**substrate**	from	**fiber**
...	...	...	...

Sampled Concentration Parameters

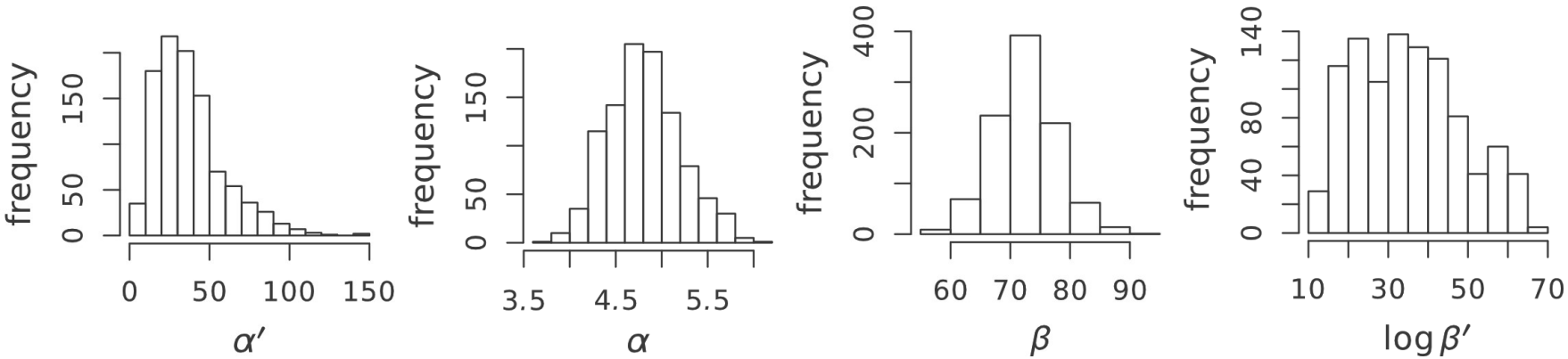


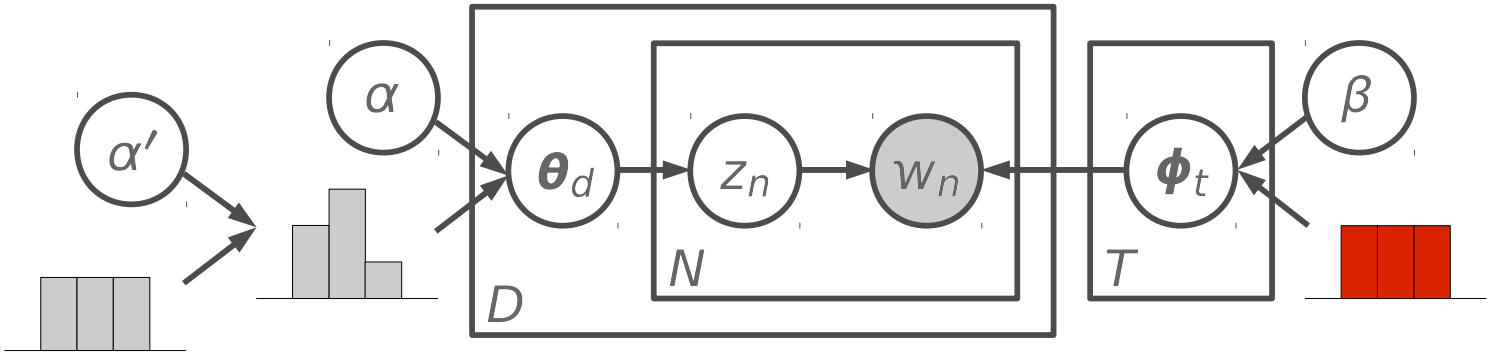
A Theoretical Observation...

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



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)
  - Symmetric prior preserves 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**
...	...	...	...

This Talk

- Statistical topic models for science policy-makers
- “Off-the-shelf” topic models: priors, stop words
- A database of National Institutes of Health grants

National Institutes of Health

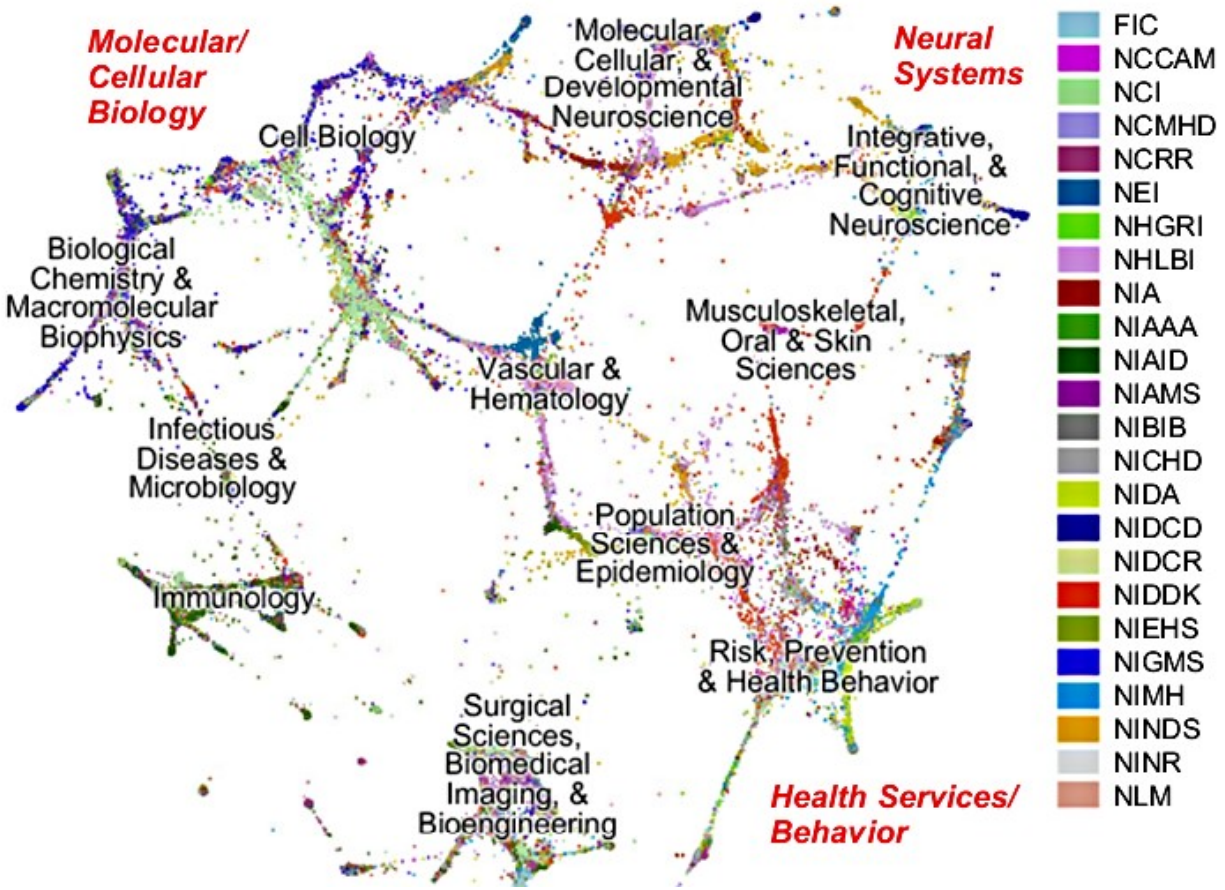
- Funds biomedical research (~80,000 awards per year)
- 27 institutes and centers:
  - Often disease-focused (e.g., cancer, diabetes)
  - … but complicated by politics and expediency
  - Diseases cross scientific boundaries
  - Significant overlap in the research funded
- Daunting landscape for choosing research directions, funding allocations, and policy actions

NIH Information Access

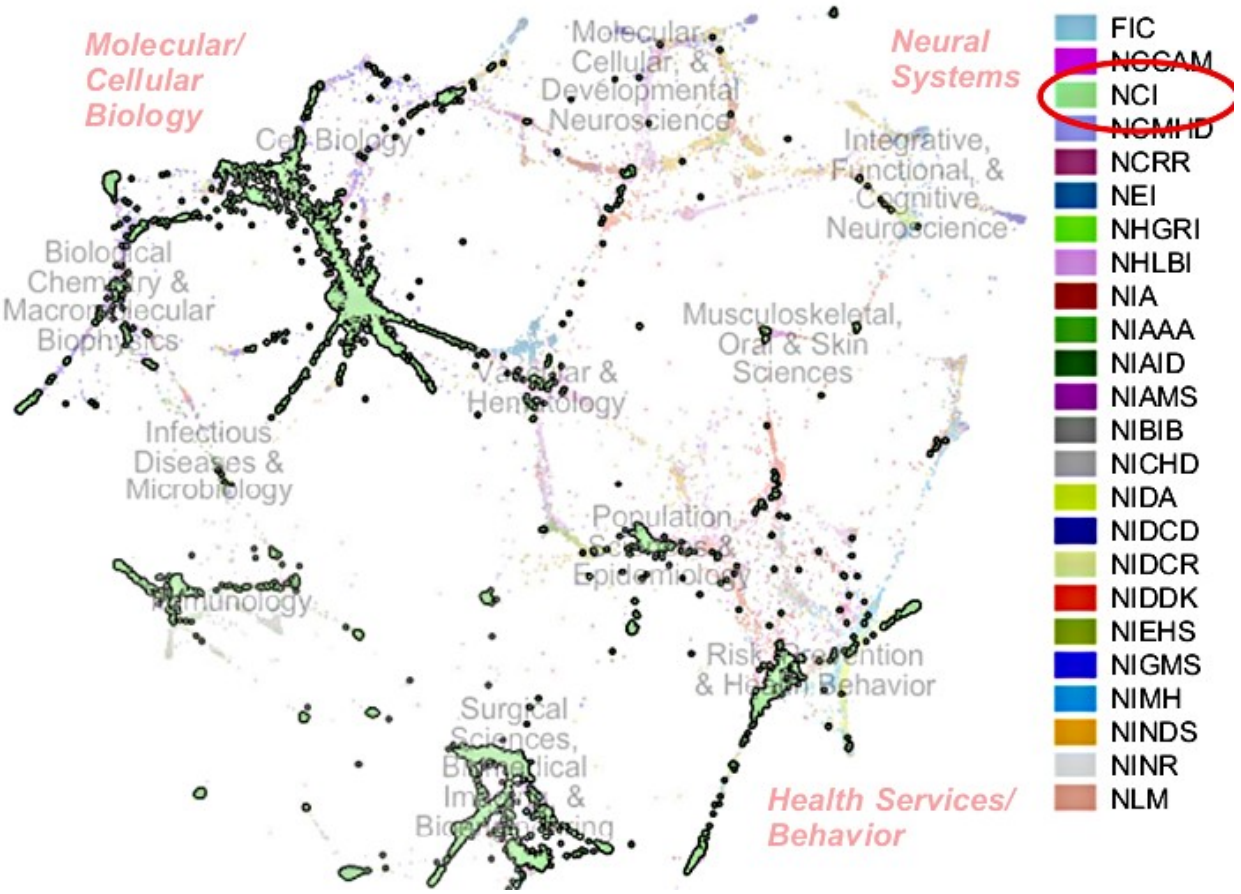
- 1972: CRISP database
  - Awards manually annotated with thesaurus terms
  - Expensive to maintain, limited search capabilities
- 2009: RePORTER and RCDC
  - Partially automated: 229 categories, preset keywords
  - Categories chosen to meet NIH reporting requirements
- 2011: NIHMaps database
  - Topic modeling + graph-based clustering

NIHMaps

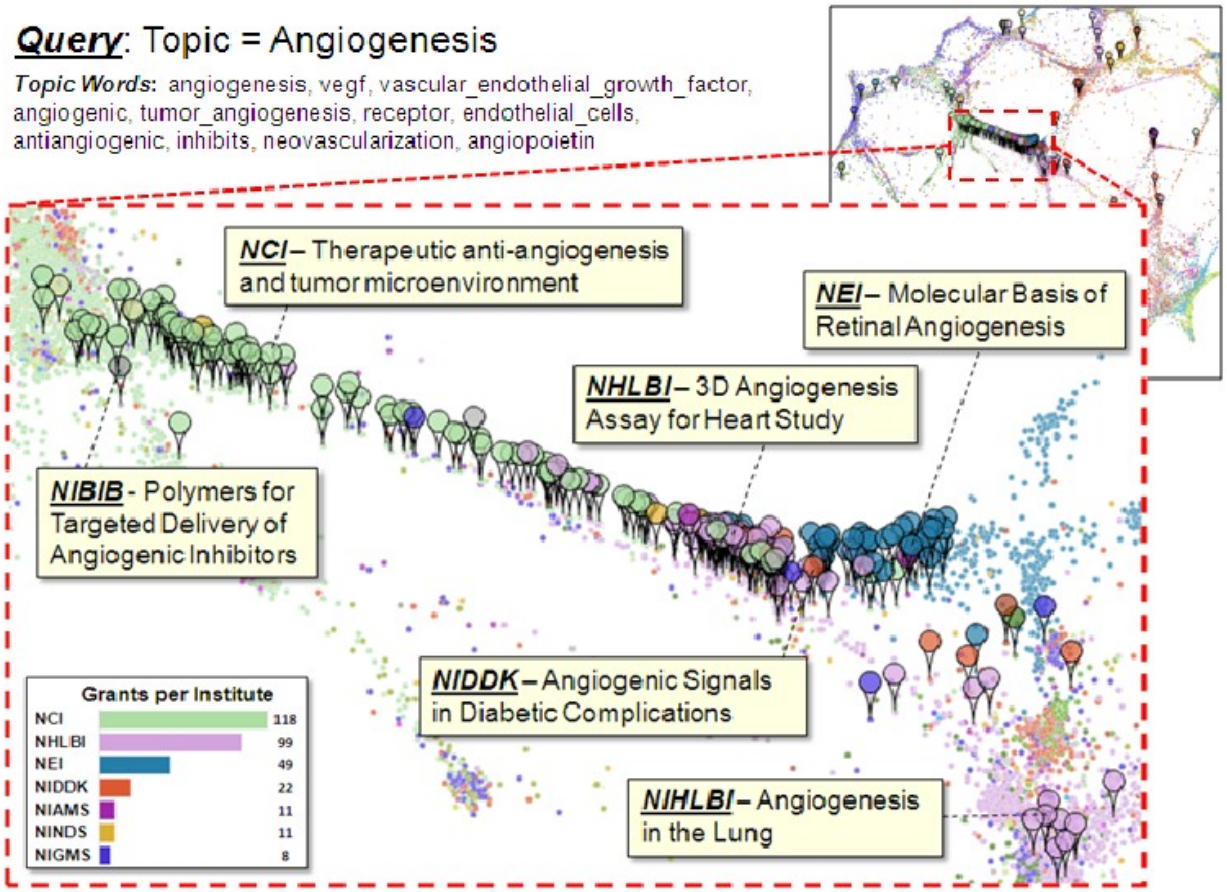
[Talley et al., '11]



Institute Organization



Topic-Based Queries



RCDC Category Exploration

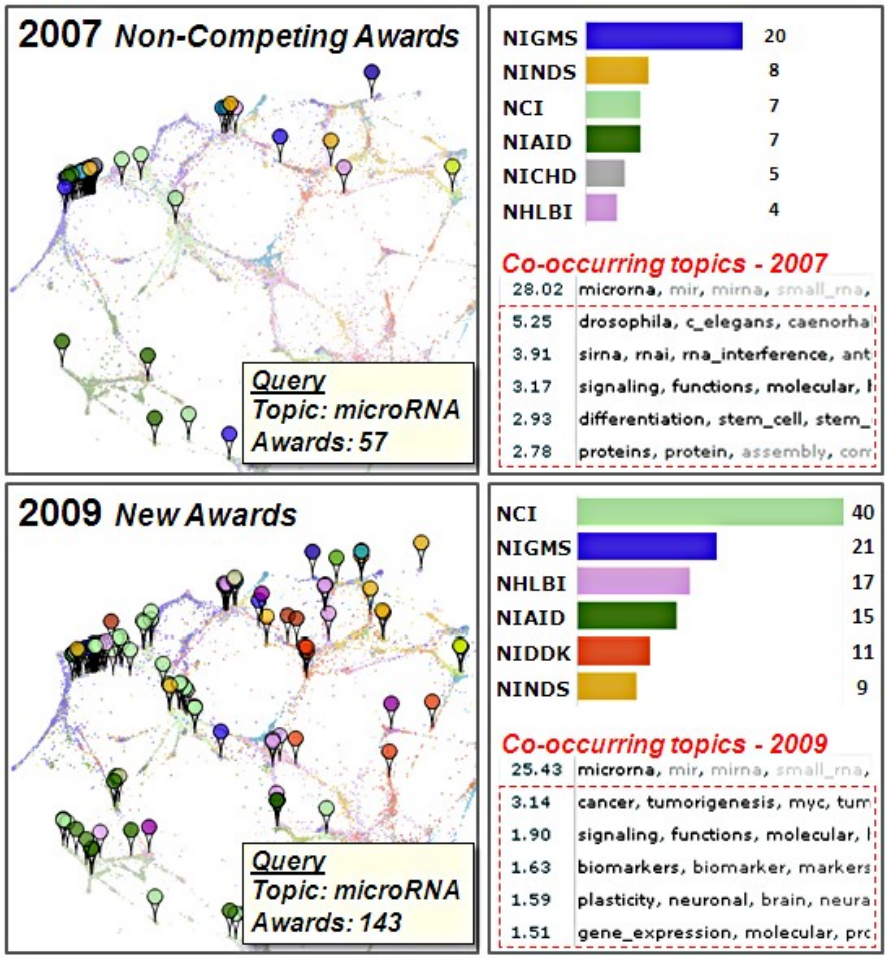
NIH Sleep Research

1	*Circadian Rhythms.....*	**circadian**, clock, rhythms, suprachiasmatic, melatonin, light, rhythm, drosophila,
2	*Sleep Disorders.....*	**sleep**, fatigue, insomnia, older, disturbances, disturbance, syndrome, restless,
3	*Neurobiology Sleep/Arousal..*	**sleep**, hypocretin, orexin, sleep_deprivation, rem_sleep, wakefulness, sleep_wa
4	*Sleep Disordered Breathing...*	**sleep_apnea**, **obstructive**, **respiratory**, intermittent_hypoxia, breathing, sleep, sl

	NHLBI	NINDS	NIMH	NIA	NCRR	NICHD	NIGMS	NIDDK	NINR	NIDA	NEI	NCCAM
1	31-32	51-58	28-32	13-18	13-15	4-6	50-52	9-10	1	9	11-14	1-4
2	72-73	17	27	36-37	27	9-10	2	3	16	5	1	10
3	17-31	50-55	35-39	21-25	7-10	3	6	1-2	0-2	3-4	1	0-1
4	82-89	6-7	1-2	3-7	12-14	21-29		0-2	2-4	1-2		
Total:	192	128	101	73	68	60	56	24	23	21	19	15

grants (*estimated*)
0-4
3-7
7-18
17-32
35-58
72-89

Topic-Based Trend Analysis



Summary

- Significant need for data-driven science policy
- Decision-makers are eager to use topic models as a strategic asset in their daily routines
- Fantastic opportunities for researchers:
  - Let practitioners' needs guide the research
  - Explore the interplay between theory and practice
  - Question unquestioned assumptions
  - Produce tools that will transform science policy

Thanks!

Acknowledgements: A. McCallum, D. Mimno, E. Talley, M. Leenders, D. Newman, Bruce Herr, Gully Burns

wallach@cs.umass.edu
http://www.cs.umass.edu/~wallach/