topic modeling

hanna m. wallach
university of massachusetts amherst
wallach@cs.umass.edu

Ramona Blei-Gantz

Helen Moss (Dave's Grandma)

The Next 30 Minutes

- Motivations and a brief history:
  - Latent semantic analysis
  - Probabilistic latent semantic analysis
- Latent Dirichlet allocation:
  - Model structure and priors
  - Approximate inference algorithms
  - Evaluation (log probabilities, human interpretation)
- Post-LDA topic modeling...

The Problem with Information

www.betaversion.org/~stefano/linotype/news/26/

- Needle in a haystack: as more information becomes available, it is harder and harder to find what we are looking for
- Need new tools to help us organize, search and understand information

A Solution?

Candida Hofer

- Use topic models to discover hidden topic-based patterns
- Use discovered topics to annotate the collection
- Use annotations to organize, understand, summarize, search...

Topic (Concept) Models

- Topic models: LSA, PLSA, LDA
- Share 3 fundamental assumptions:
  - Documents have latent semantic structure (“topics”)
  - Can infer topics from word-document co-occurrences
  - Words are related to topics, topics to documents
- Use different mathematical frameworks
  - Linear algebra vs. probabilistic modeling

Topics and Words

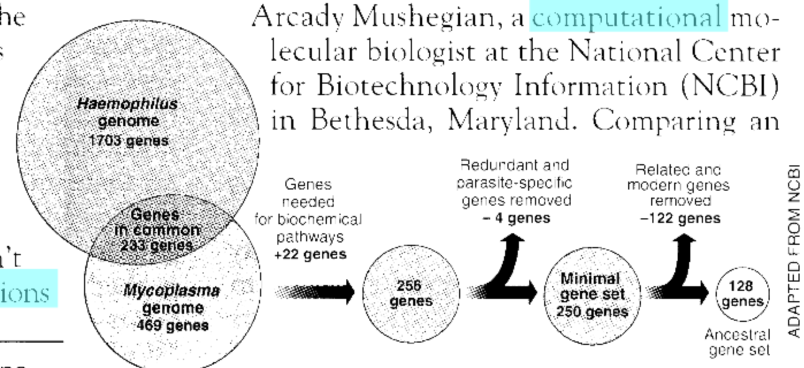
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

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

Latent Semantic Analysis

(Deerwester et al., 1990)

- Based on ideas from linear algebra
- Form sparse term–document co-occurrence matrix X
  - Raw counts or (more likely) TF-IDF weights
- Use SVD to decompose X into 3 matrices:
  - U relates terms to "concepts"
  - V relates "concepts" to documents
  - Σ is a diagonal matrix of singular values

Singular Value Decomposition

1. Latent semantic analysis (LSA) is a theory and method for ...
2. Probabilistic latent semantic analysis is a probabilistic ...
3. Latent Dirichlet allocation, a generative probabilistic model ...

	1	2	3
allocation	0	0	1
analysis	1	1	0
Dirichlet	0	0	1
generative	0	0	1
latent	1	1	1
LSA	1	0	1
probabilistic	0	2	1
semantic	1	1	0
...		...	

=

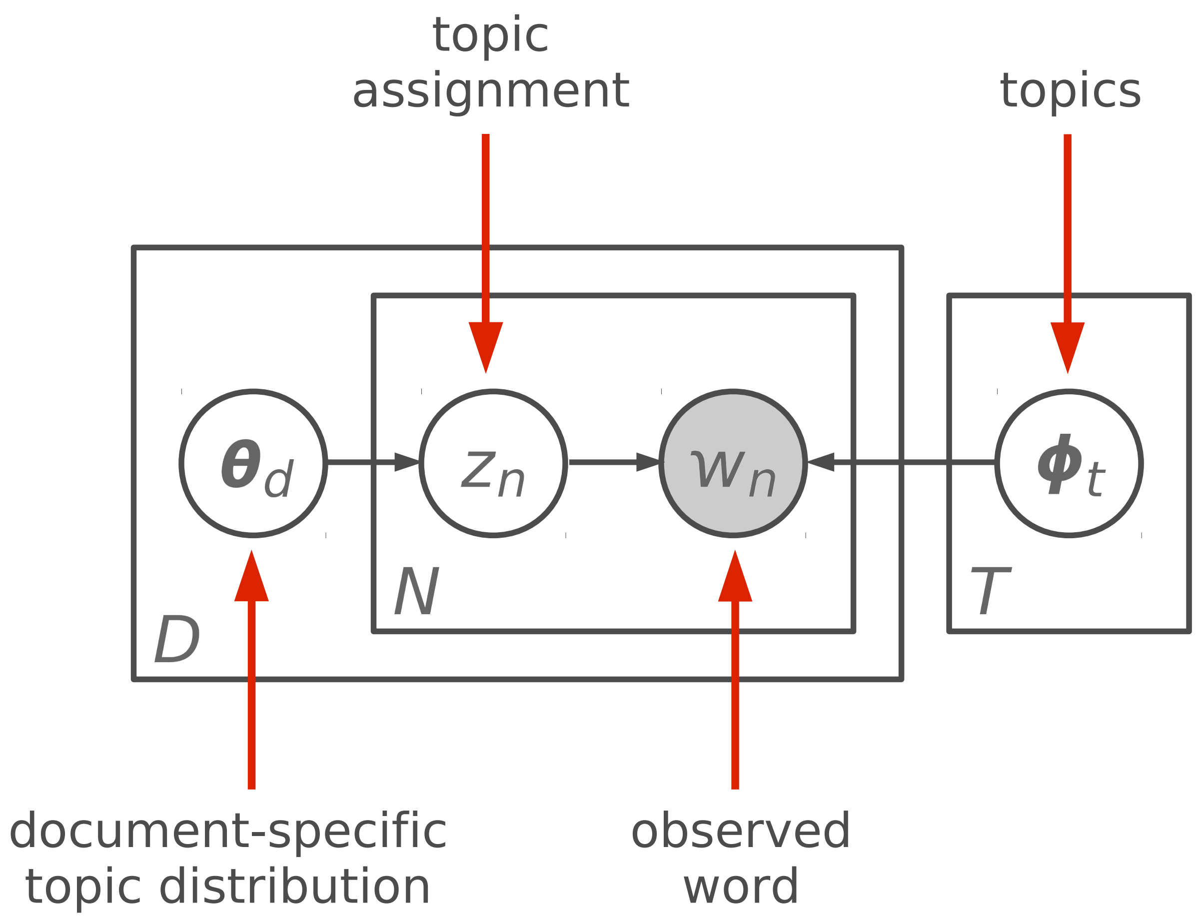
$$X = U \Sigma V^{\mathsf{T}}$$

Probabilistic Modeling

- Treat data as observations that arise from a generative probabilistic process that includes hidden variables
  - For documents, the hidden variables represent the thematic structure of the collection
- Infer the hidden structure using posterior inference
  - What are the topics that describe this collection?
- Situate new data into the estimated model

Probabilistic LSA

(Hofmann, 1999)



Advantages and Disadvantages

✔ Probabilistic model that can be easily extended and embedded in other more complicated models

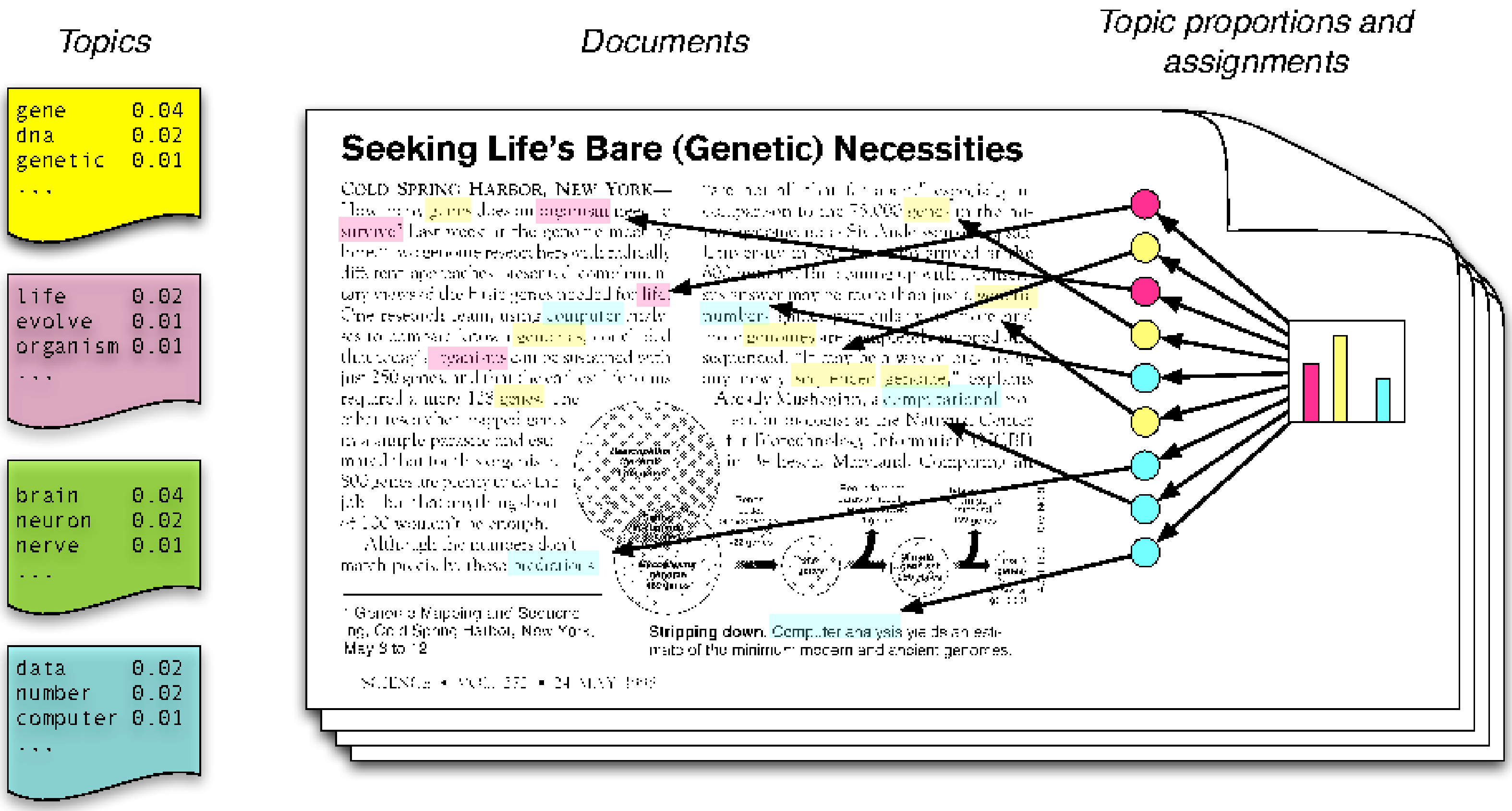
✗ Not a well-defined generative model: no way of generalizing to new, unseen documents

✗ Many free parameters (linear in # training documents)

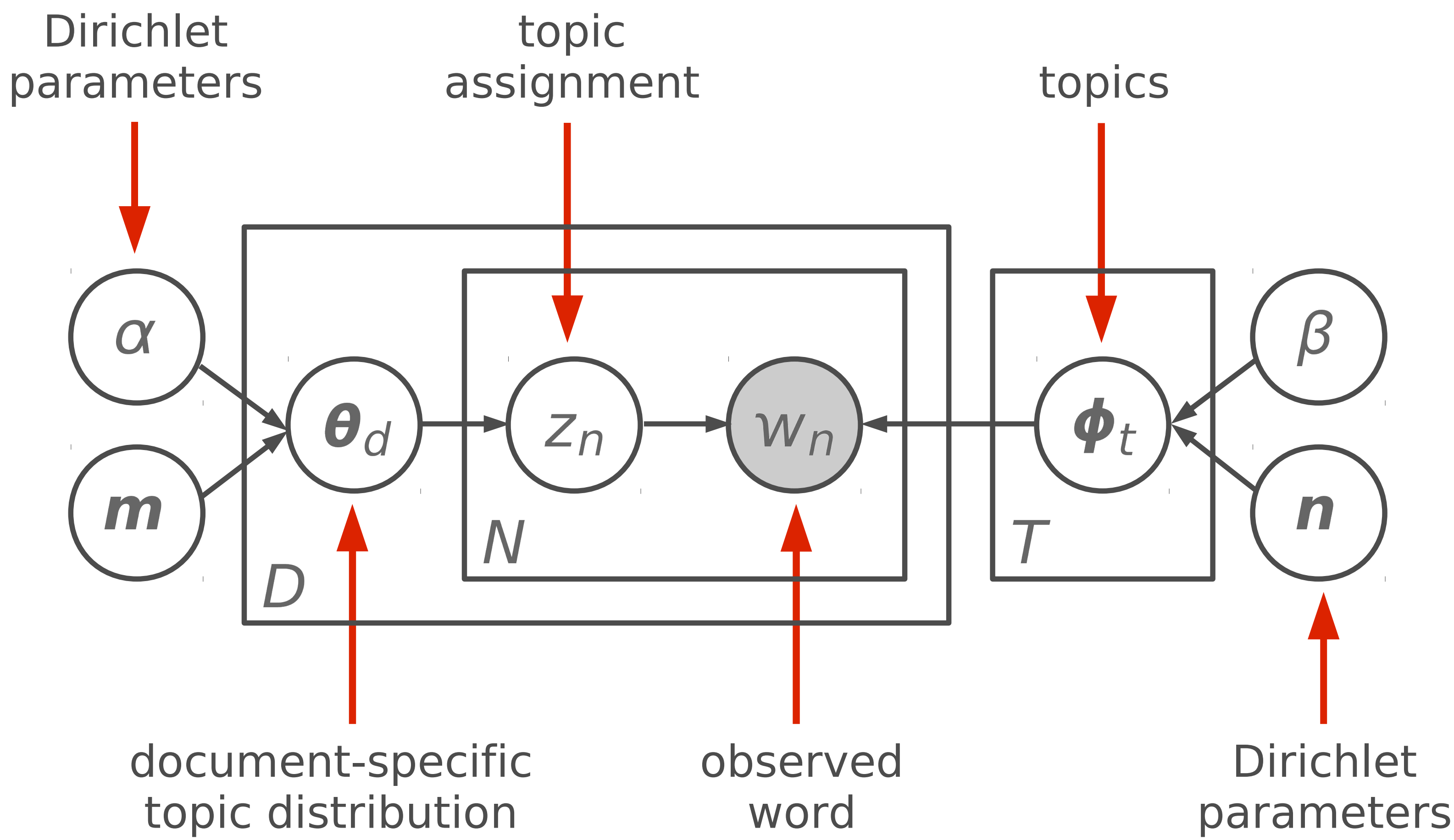
✗ Prone to overfitting (have to be careful when training)

Latent Dirichlet Allocation

(Blei et al., 2003)



Graphical Model



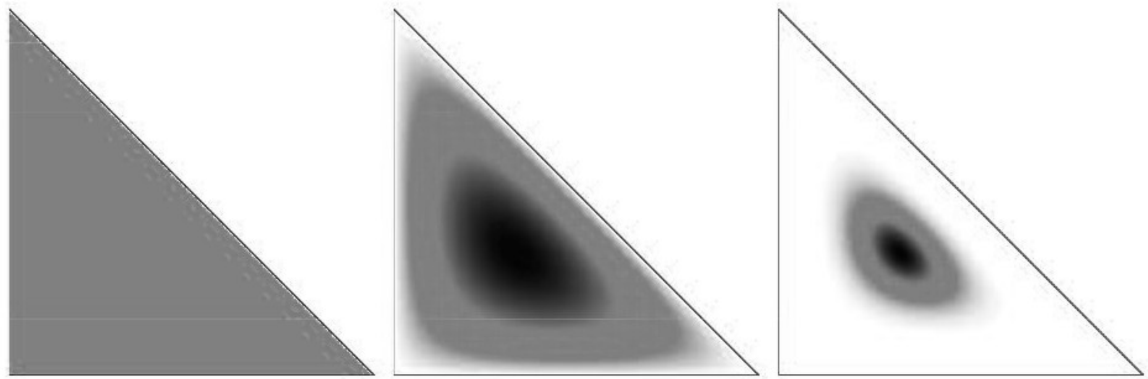
Dirichlet Distribution

- Distribution over K-dimensional positive vectors that sum to one (i.e., points on the probability simplex)

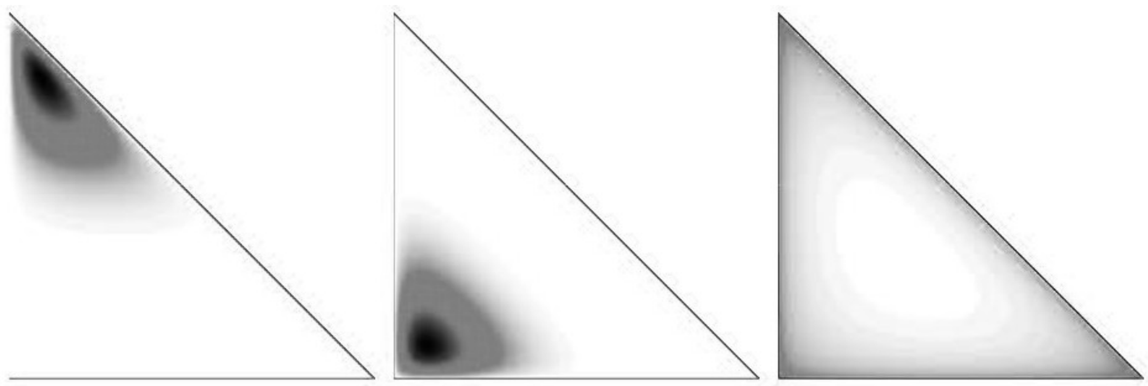
$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

- Two parameters:
  - Base measure, e.g., $\boldsymbol{m}$ (vector)
  - Concentration parameter, e.g., α (scalar)

Varying Parameters



$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 6, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 30, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$



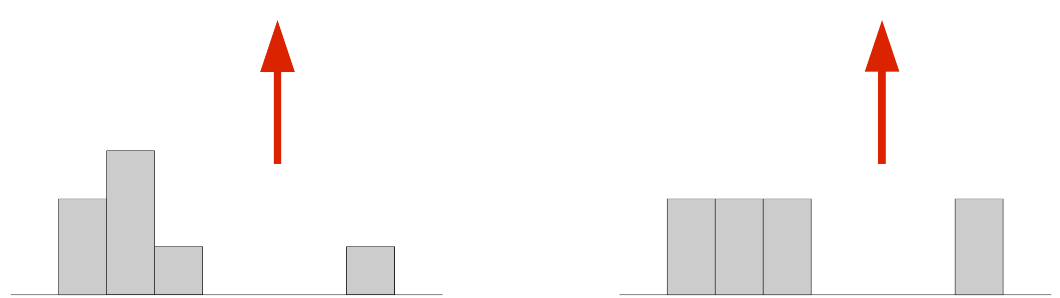
$\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7})$ $\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7})$ $\alpha = 2.7, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Asymmetric? Symmetric?

(Wallach et al., 2009)

- People (almost always) use symmetric Dirichlet priors with heuristically set concentration parameters
  - Simple, but is it the best modeling choice?
- Empirical comparison:

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha \boldsymbol{m}) \text{ and } \boldsymbol{\phi}_t \sim \mathrm{Dir}(\beta \boldsymbol{u})$$



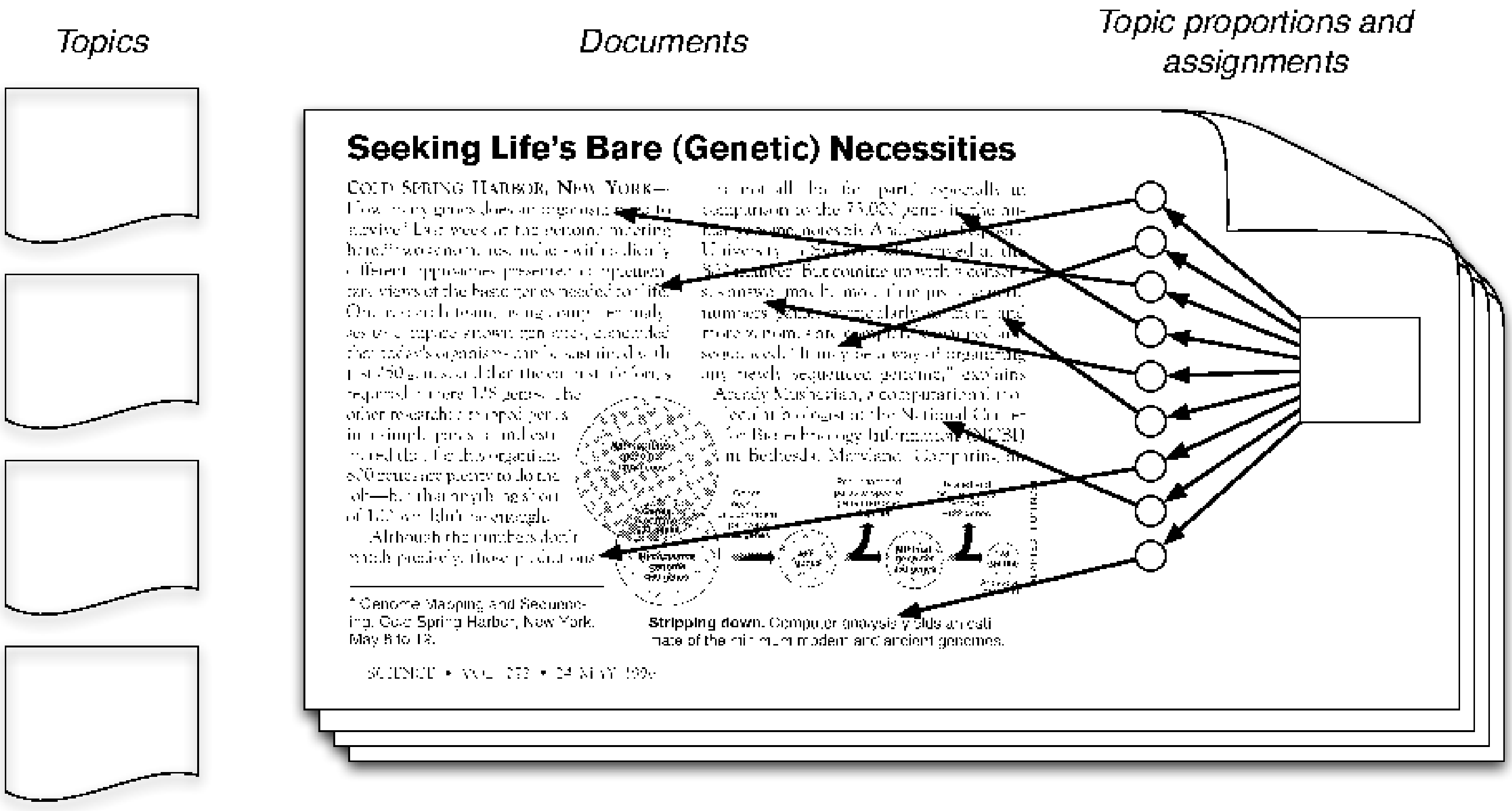
Priors and Stop Words

	symm. prior over Φ		asymm. prior over Φ	
symm. Θ	0.080	a **field emission** an **electron** the	0.042	a **field** the **emission** and **carbon** is
	0.080	a the **carbon** and **gas** to an	0.042	the **carbon catalyst** a **nanotubes**
	0.080	the of a to and about at	0.042	a the of **susbtrate** to **material** on
	0.080	of a **surface** the with in **contact**	0.042	**carbon single wall** the **nanotubes**
	0.080	the a and to is of **liquid**	0.042	the a **probe tip** and of to
asymm. Θ	0.895	the a of to and is in	1.300	the a of to and is in
	0.187	**carbon nanotubes nanotube catalyst**	0.257	and are of for in as such
	0.043	sub is c or and n sup	0.135	a **carbon material** as **structure nanotube**
	0.061	**fullerene compound fullerenes**	0.065	**diameter swnt** about **nm** than **fiber swnts**
	0.044	**material particles coating inorganic**	0.029	**compositions polymers polymer contain**

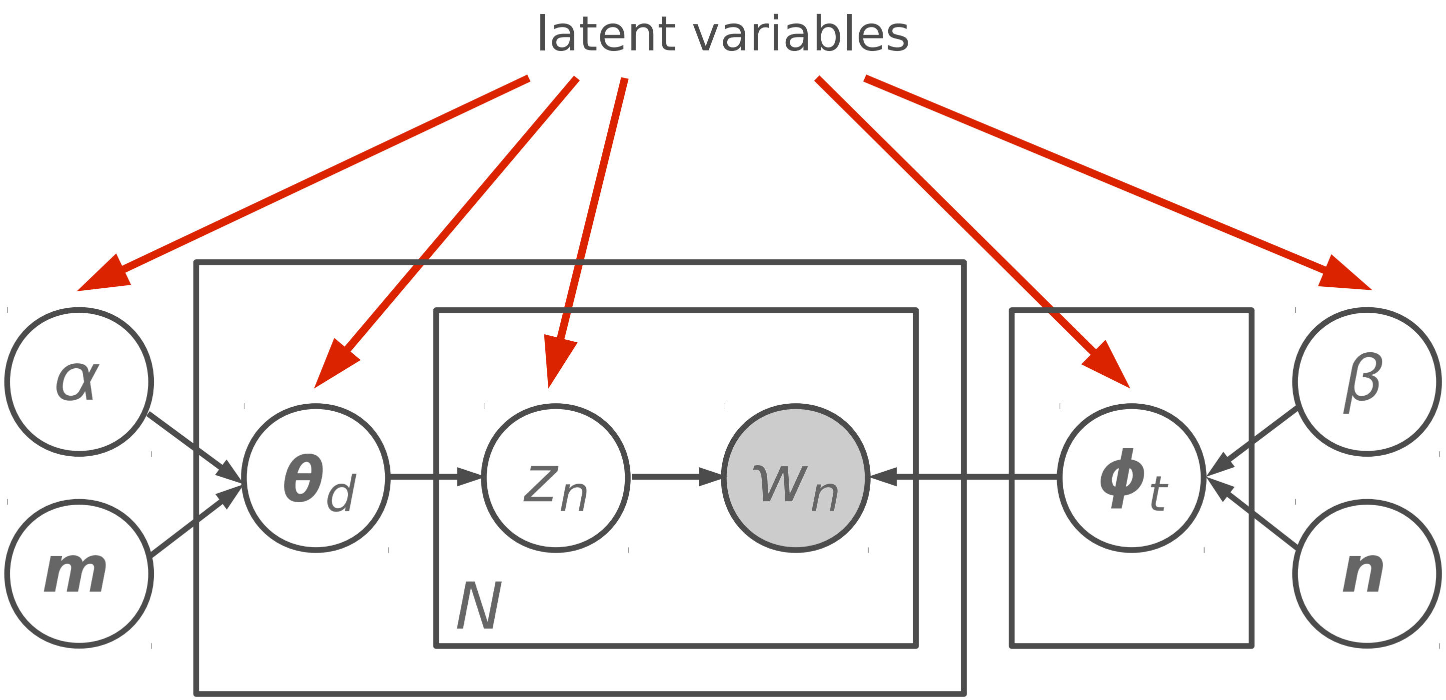
Intuition

- Topics are specialized distributions over words
  - Want topics to be as distinct as possible
  - Asymmetric prior over { $\boldsymbol{\phi}_t$ } makes topics more similar to each other (and to the corpus word frequencies)
  - Want a symmetric prior to preserve topic “distinctness”
- Still have to account for power-law word usage:
  - Asymmetric prior over { $\boldsymbol{\theta}_d$ } means some topics can be used much more often than others

Posterior Inference



Posterior Inference



- Infer (or integrate out) all latent variables, given tokens

Inference Algorithms

(Mukherjee & Blei, 2009; Asuncion et al., 2009)

- Exact inference in LDA is not tractable
- Approximate inference algorithms:
  - Mean field variational inference (Blei et al., 2001; 2003)
  - Expectation propagation (Minka & Lafferty, 2002)
  - Collapsed Gibbs sampling (Griffiths & Steyvers, 2002)
  - Collapsed variational inference (Teh et al., 2006)
- Each method has advantages and disadvantages

Evaluating LDA: Log Probability

- Unsupervised nature of LDA makes evaluation hard
- Compute probability of held-out documents:
  - Classic way of evaluating generative models
  - Often used to evaluate topic models
- Problem: have to approximate an intractable sum

$$P(\boldsymbol{w} \,|\, \boldsymbol{w}', \boldsymbol{z}', \alpha \boldsymbol{m}, \beta \boldsymbol{u}) = \sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z} \,|\, \boldsymbol{w}', \boldsymbol{z}', \alpha \boldsymbol{m}, \beta \boldsymbol{u})$$

Computing Log Probability

(Wallach et al., 2009)

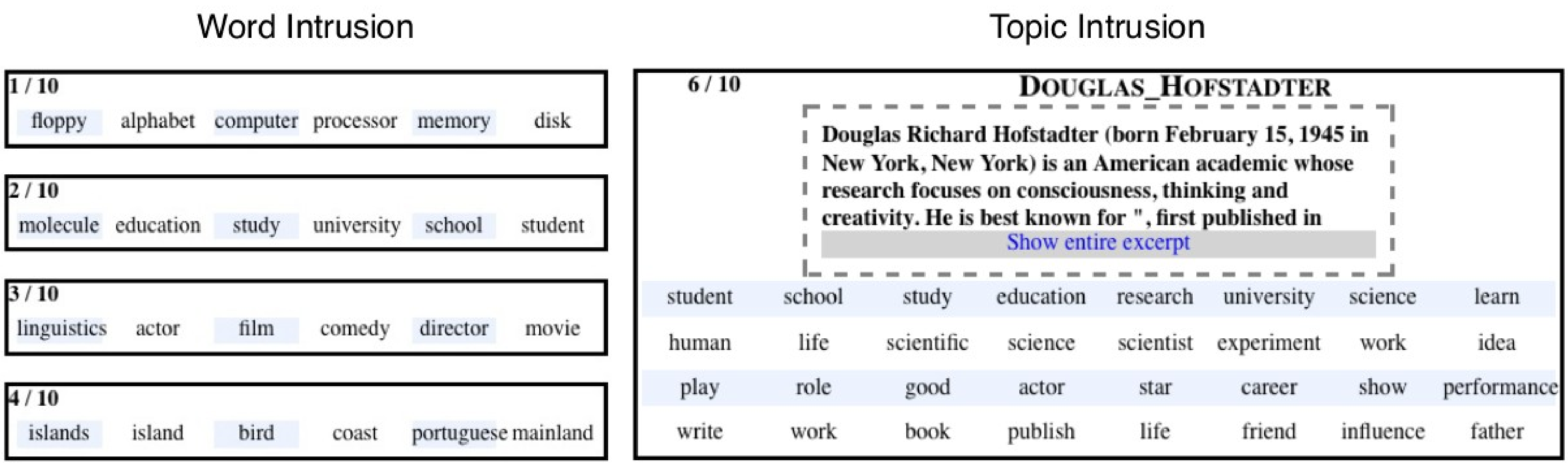
- Simple importance sampling methods
- The “harmonic mean” method (Newton & Raftery, 1994)
  - Known to overestimate, used anyway
- Annealed importance sampling (Neal, 2001)
  - Prohibitively slow for large collections of documents
- Chib-style method (Murray & Salakhutdinov, 2009)
- “Left-to-Right” method (Wallach, 2008)

Reading Tea Leaves

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
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

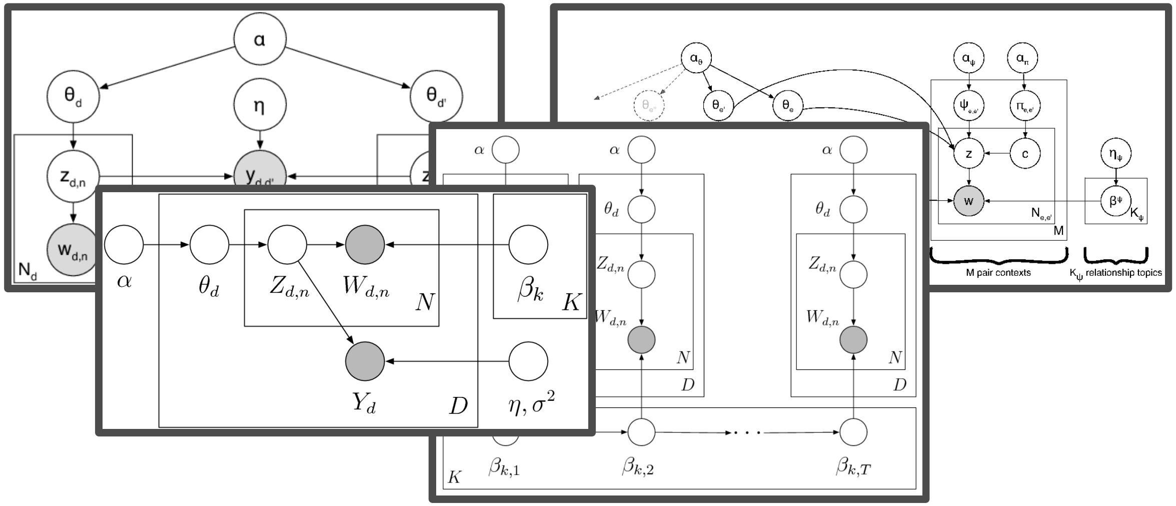
Word and Topic Intrusion

(Chang et al., 2009)



- Can humans find the “intruder” word/topic?

Post-LDA Topic Modeling



- LDA can be embedded in more complicated models
- Data-generating distribution can be changed

Today's Workshop

- Text and language (S. Gerrish & D. Blei; M. Johnson; T Landauer)
- Time-evolving networks (E. Xing)
- Visual recognition (L. Fei-Fei)
- Finance (G. Doyle & C. Elkan)
- Archeology (D. Mimno)
- Music analysis (D. Hu & L. Saul)
- ... even some theoretical work (D. Sontag & D. Roy)

questions?

(thanks to Dave Blei for letting me steal pictures/content etc.)