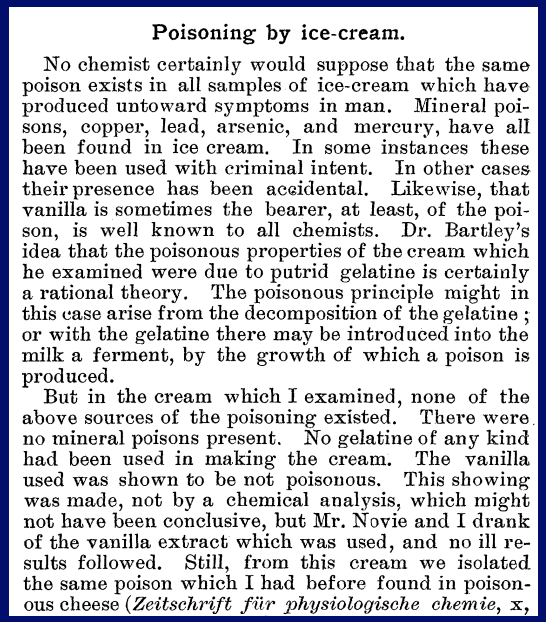
Modeling *Science*

Topic models of Scientific Journals and Other Large Text Databases

David M. Blei
Princeton University

(joint work with M. Jordan, A. Ng, and J. Lafferty)

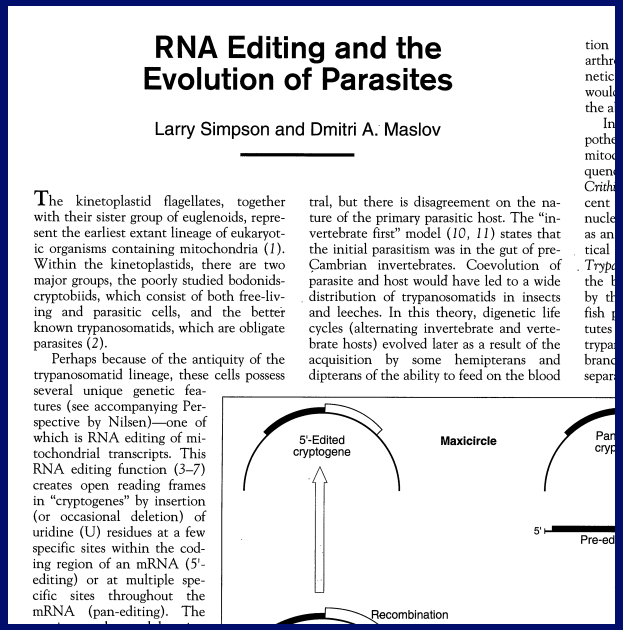
Modeling *Science*



Poisoning by ice-cream.

No chemist certainly would suppose that the same poison exists in all samples of ice-cream which have produced untoward symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this case arise from the decomposition of the gelatine; or with the gelatine there may be introduced into the milk a ferment, by the growth of which a poison is produced.

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but Mr. Novie and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (*Zeitschrift für physiologische chemie*, x,



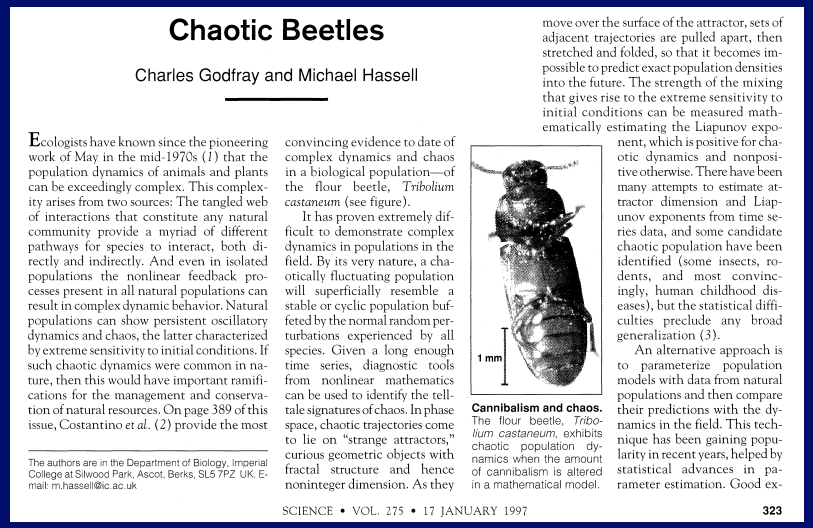
RNA Editing and the Evolution of Parasites

Larry Simpson and Dmitri A. Maslov

The kinetoplastid flagellates, together with their sister group of euglenoids, represent the earliest extant lineage of eukaryotic organisms containing mitochondria (*1*). Within the kinetoplastids, there are two major groups, the poorly studied bodonids-cryptobiids, which consist of both free-living and parasitic cells, and the better known trypanosomatids, which are obligate parasites (*2*).

Perhaps because of the antiquity of the trypanosomatid lineage, these cells possess several unique genetic features (see accompanying Perspective by Nilsen)—one of which is RNA editing of mitochondrial transcripts. This RNA editing function (*3*–*7*) creates open reading frames in "cryptogenes" by insertion (or occasional deletion) of uridine (U) residues at a few specific sites within the coding region of an mRNA (5′-editing) or at multiple specific sites throughout the mRNA (pan-editing). The

tral, but there is disagreement on the nature of the primary parasitic host. The "invertebrate first" model (*10*, *11*) states that the initial parasitism was in the gut of pre-Cambrian invertebrates. Coevolution of parasite and host would have led to a wide distribution of trypanosomatids in insects and leeches. In this theory, digenetic life cycles (alternating invertebrate and vertebrate hosts) evolved later as a result of the acquisition by some hemipterans and dipterans of the ability to feed on the blood



Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al*. (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

How can I automatically organize and browse a large, unstructured collection of OCR'ed docuuments?

Topic models

- Generative probabilistic models of text
- Use distributions over the vocabulary, called *topics*, to describe the collection
- Useful for many kinds of tasks
  - Organization
  - Classification
  - Collaborative filtering
  - Information retrieval

Discover topics from a corpus

"Genetics"	"Evolution"	"Research"	"Disease"	"Computers"
human	evolution	says	disease	computer
genome	evolutionary	researchers	host	models
dna	species	colleagues	bacteria	information
genetic	organisms	team	diseases	data
genes	life	just	resistance	computers
sequence	origin	like	bacterial	system
gene	biology	new	new	network
molecular	groups	work	strains	systems
sequencing	phylogenetic	years	control	model
map	living	called	infectious	parallel
information	diversity	dont	malaria	methods
genetics	group	say	parasite	networks
mapping	new	get	parasites	software
project	two	see	united	new
sequences	common	university	tuberculosis	simulations

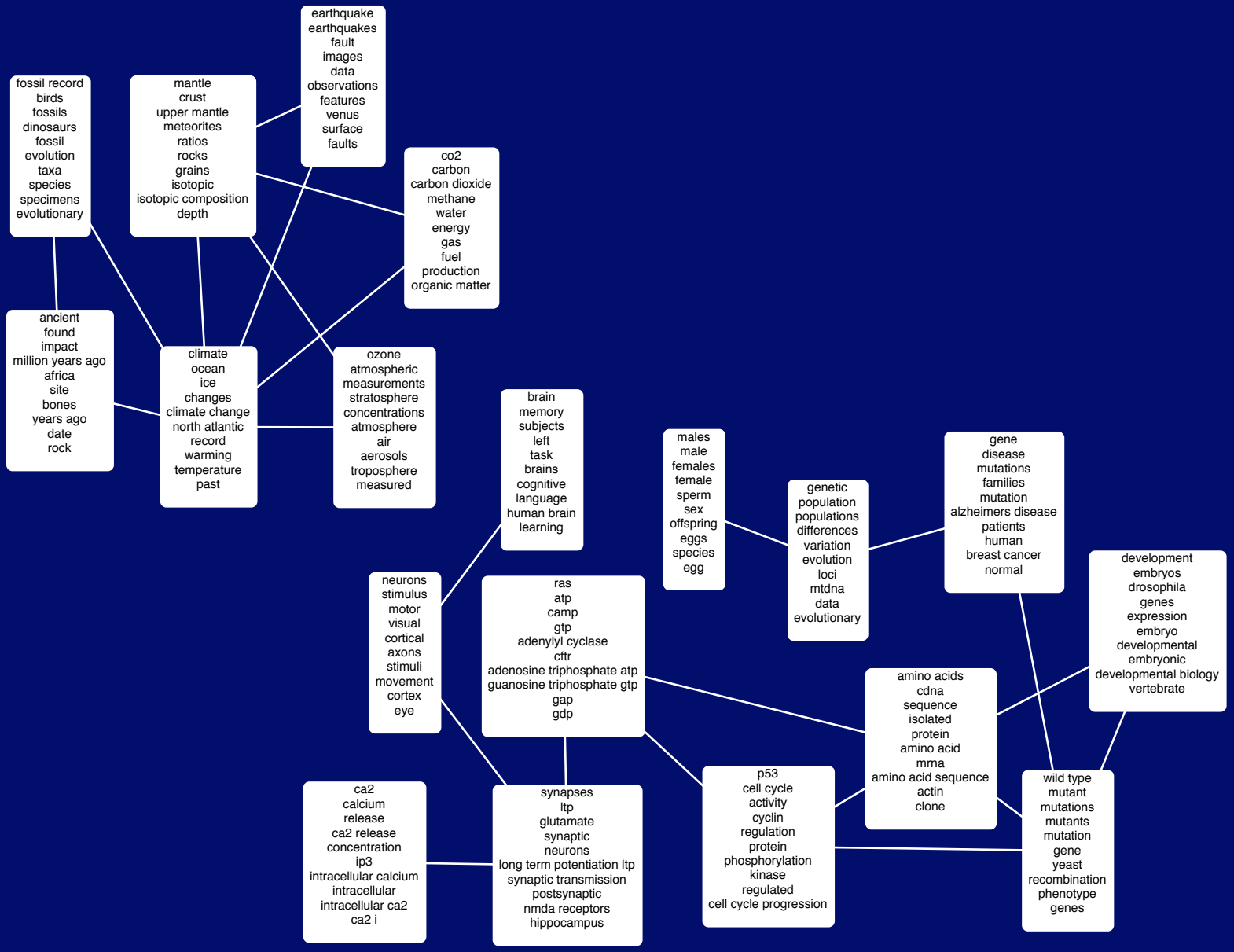
Annotate unlabeled images

scotland, water, flower, hills, tree

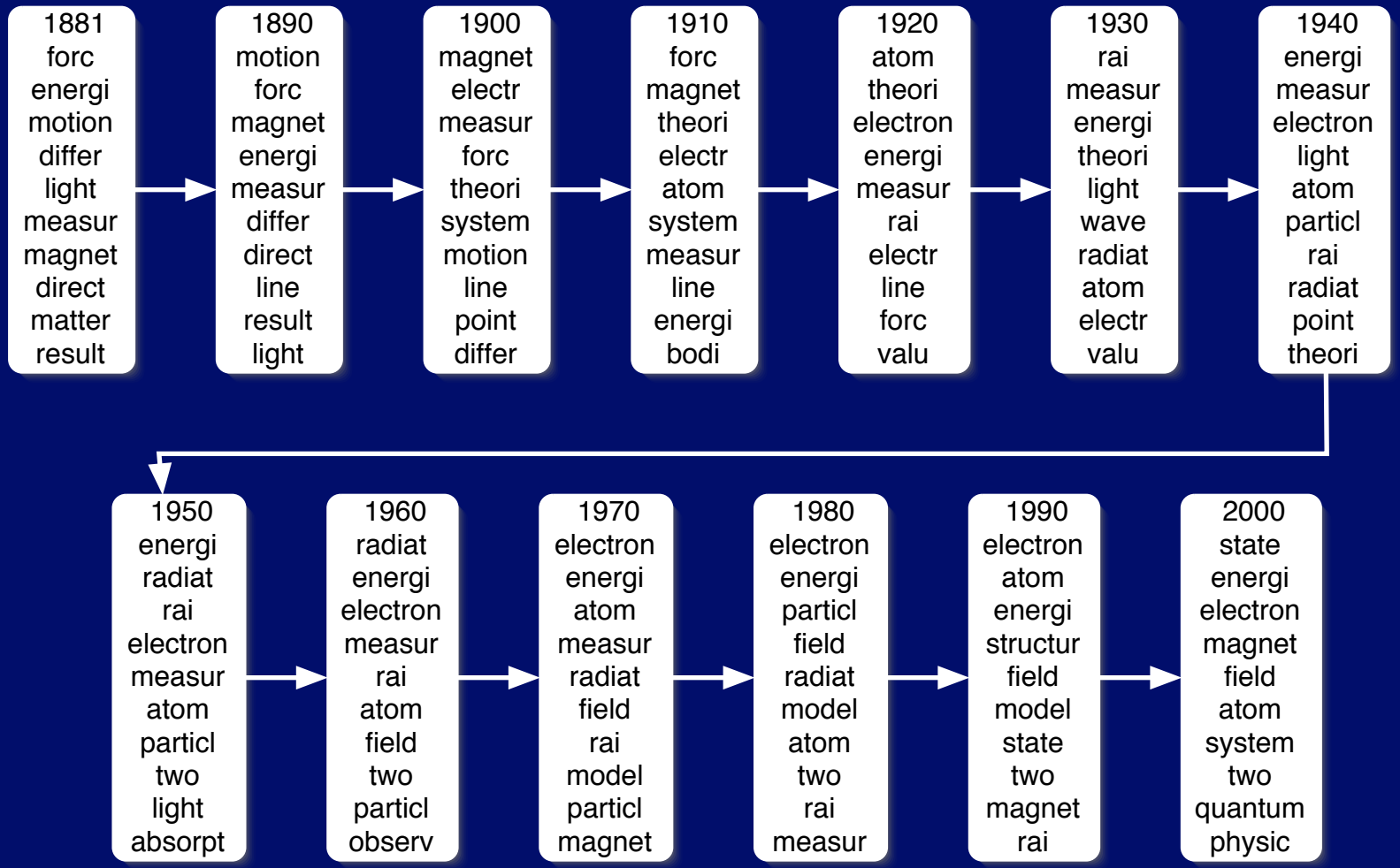
birds, nest, leaves, branch, tree

people, market, pattern, textile, display

Connections between topics



Topic evolution over time



Probabilistic modeling

- Probabilistic modeling is a mainstay in ML research
- Treat data as observations which arise from a generative probabilistic process that includes *hidden variables*
  - For documents, the hidden variables will reflect the themes or topics which the document is about

Probabilistic inference

- Learn the hidden structure based on data
  - For documents, inference amounts to learning the topics from a collection of documents
- Situate new data points into a learned model
  - E.g., how does this query or new document fit in to the discovered topic structure?

High level strategy

- Develop probabilistic models of observed data
- Cast tasks as statistical questions about the model
  - Classification
  - Information retrieval
- Pay attention to computational concerns
  - Apply to large datasets
  - Provide fast answers to desired questions

Outline

- Introduction
- The Latent Dirichlet allocation (LDA) topic model
- Posterior inference in LDA
- Correlated topic models
- Dynamic topic models
- Images and captions (if there's time)
- Summary

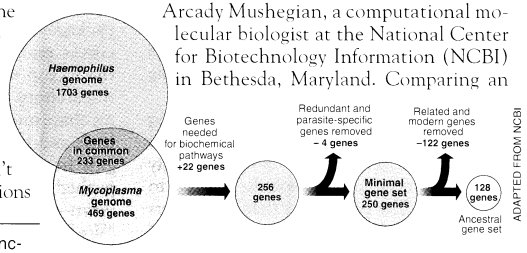
Latent Dirichlet allocation

Modeling document collections

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.

Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

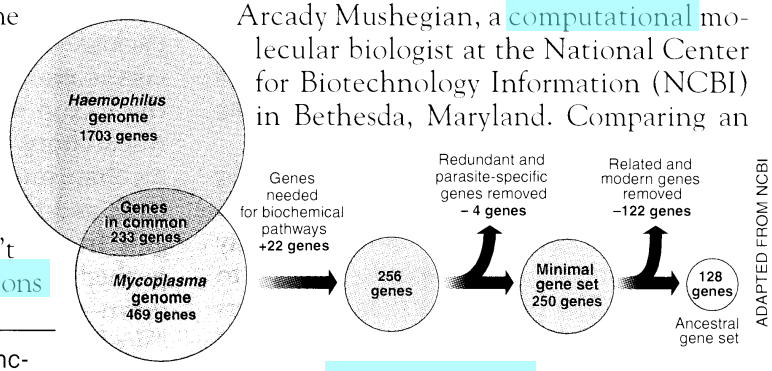
- Probabilistic model of document collections
  - Capture recurring patterns of word co-occurrences
- Facilitate activities such as classification, clustering, information retrieval, collaborative filtering

Combination of 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



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

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

SCIENCE • VOL. 272 • 24 MAY 1996

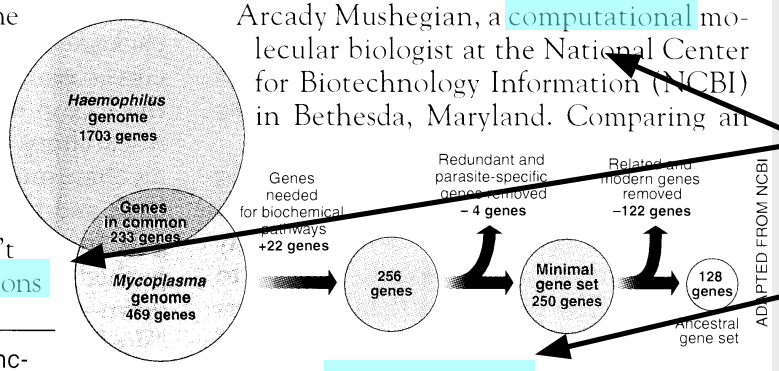
Different topics highlighted by different colors

Latent Dirichlet allocation (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.

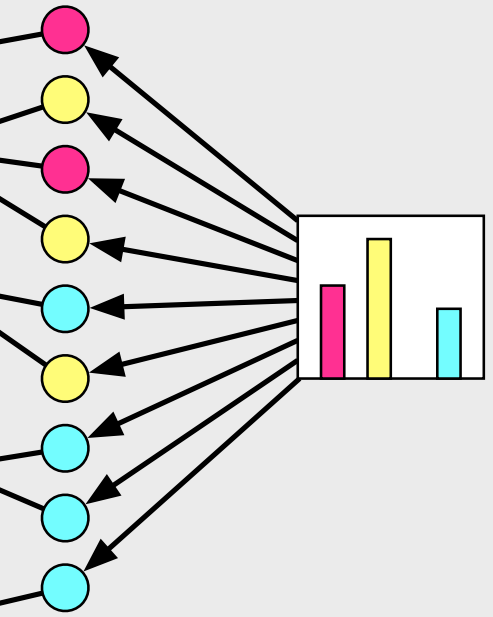
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

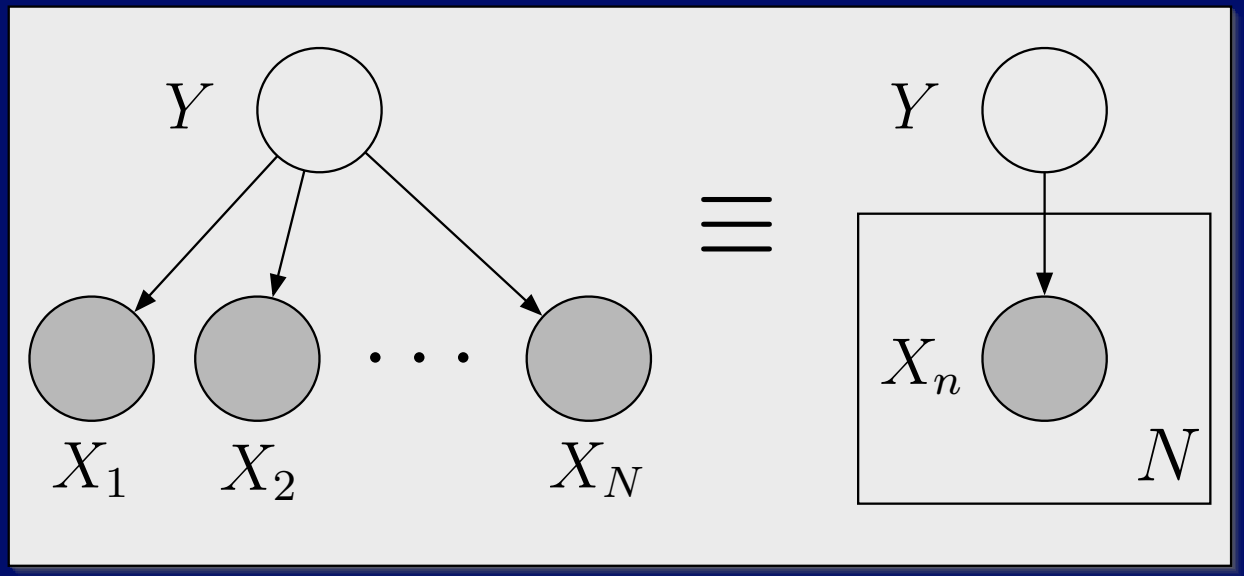
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



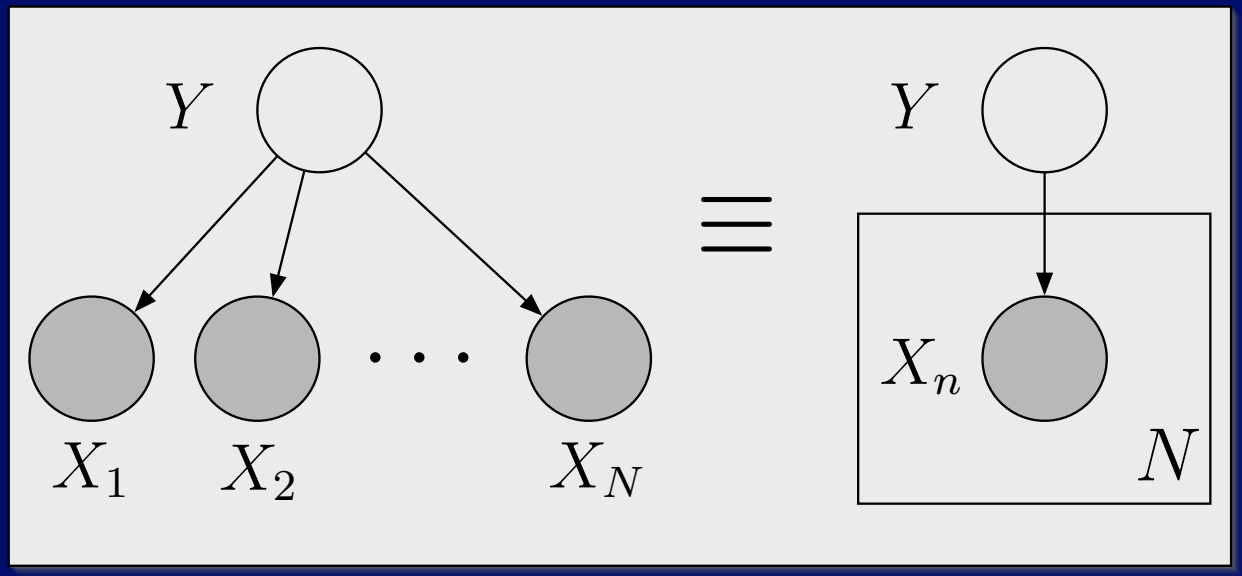
- Model the data with a probabilistic generative process
- Each document is a random mixture of topics

Graphical models



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

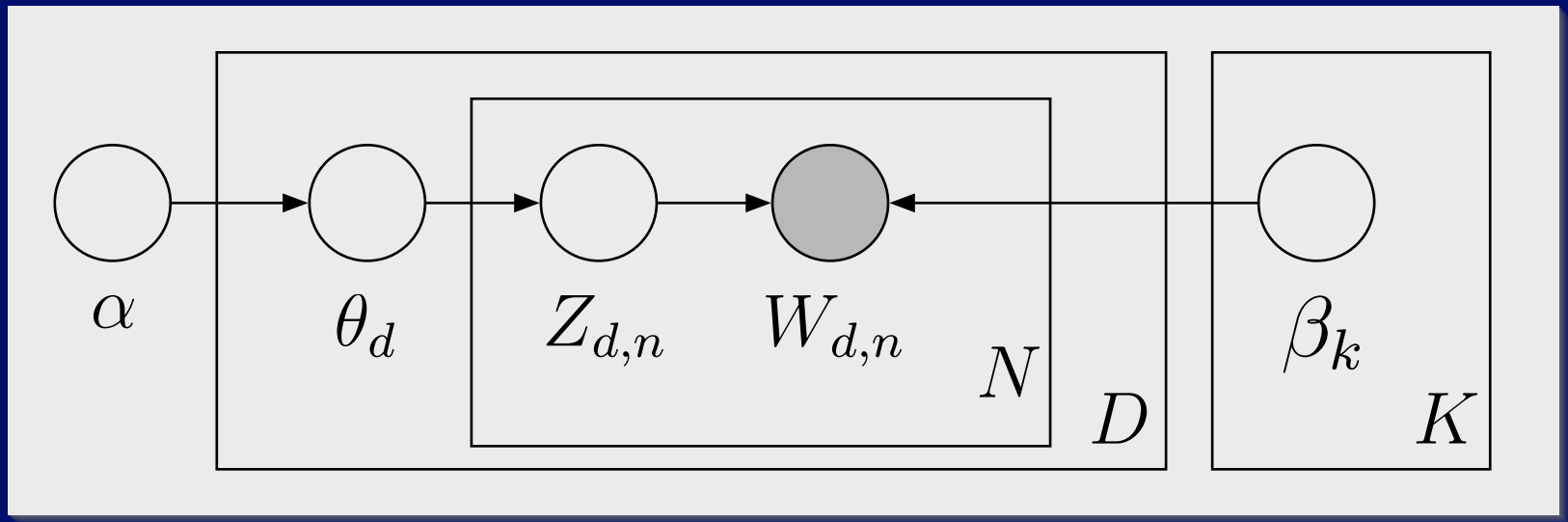
Graphical models



- Structure of the graph defines the pattern of conditional dependence between random variables
- Graph corresponds to a factorization of the joint

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

Latent Dirichlet allocation (LDA)



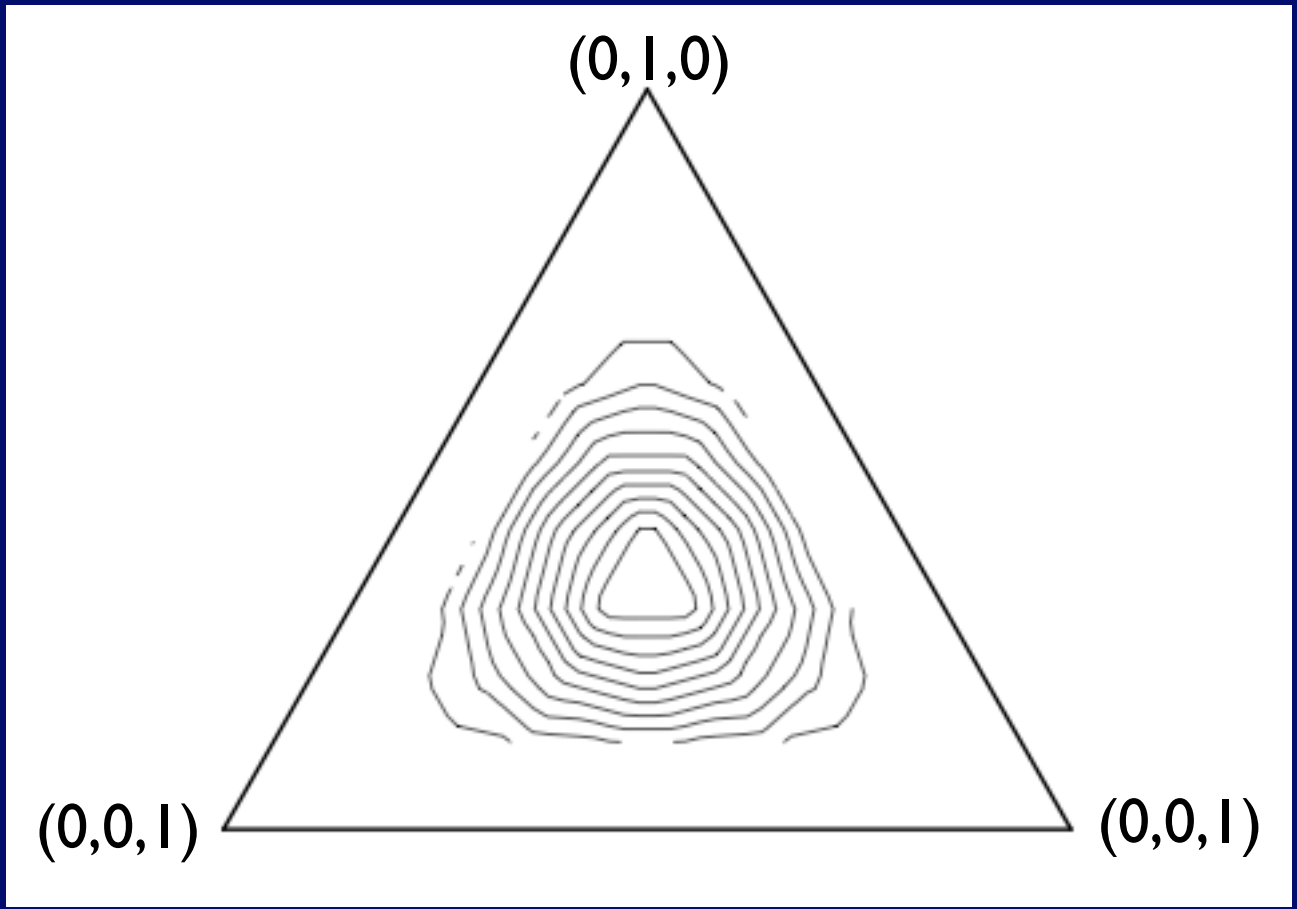
- For each document
  - Choose topic proportions $\theta \sim \mathrm{Dir}(\alpha)$
  - For each word
    - Choose a topic index $Z \sim \mathrm{Mult}(\theta)$
    - Choose a word $W \sim \mathrm{Mult}(\beta_z)$

The Dirichlet distribution

- The parameter to a multinomial distribution is a positive vector that sums to one
  - This space is called the *simplex*
- The Dirichlet distribution is a distribution on the simplex

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1},$$

The Dirichlet distribution



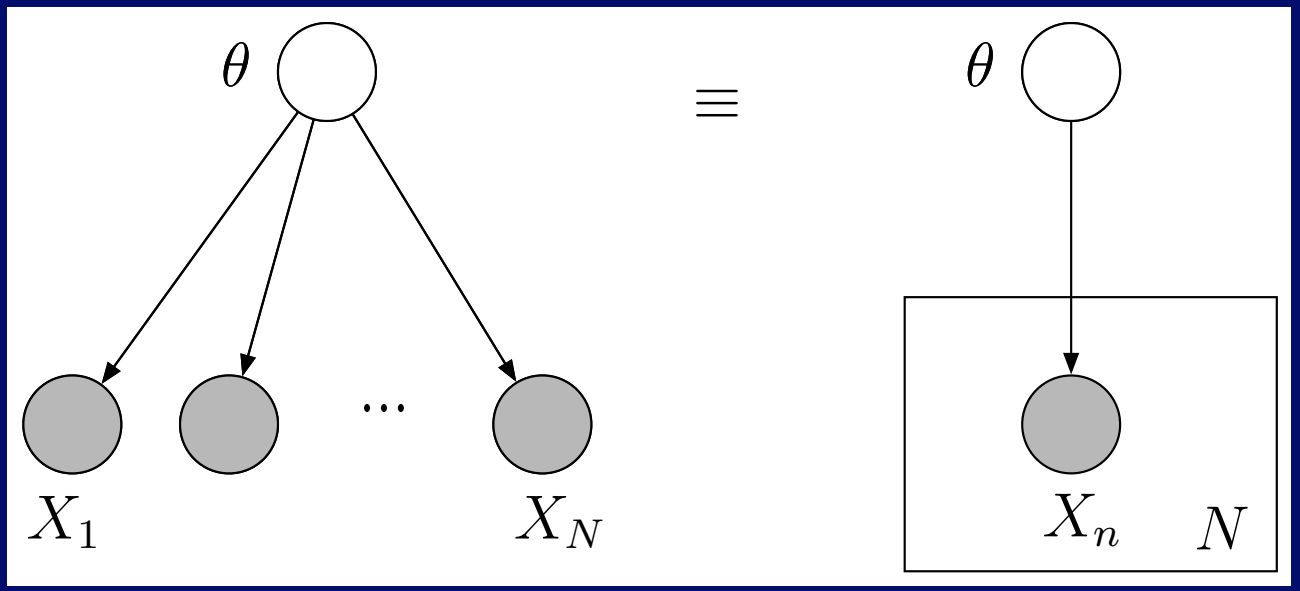
LDA and "bag of words"

- **Exchangeability**: statistical term for "bag of words"
- If $\{x_1, \ldots, x_N\}$ are exchangeable, then

$$p(x_1, \ldots, x_N) = p(x_{\pi(1)}, \ldots, x_{\pi(N)}).$$

  for any permutation π of $\{1, \ldots, N\}$
- *Not* the same as independent and identically distributed

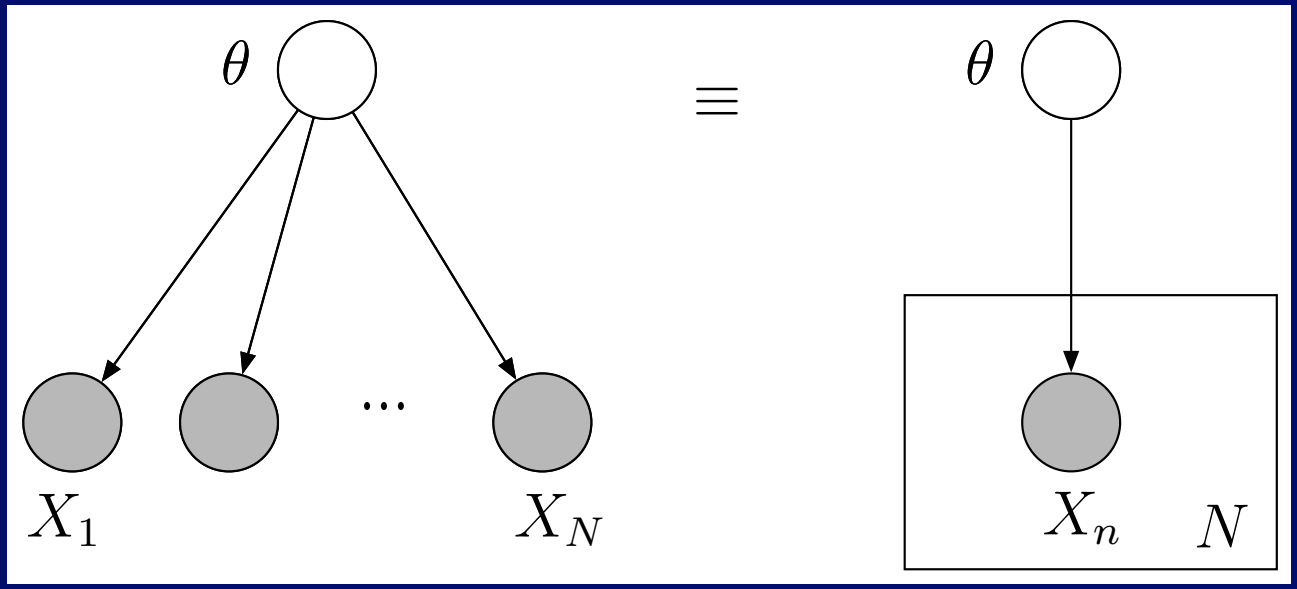
Representation theorem



- De Finetti's theorem says: it is *conditionally* IID
- If $\{x_1, \ldots, x_N\}$ are exchangeable, then the joint distribution can be represented as a continuous mixture

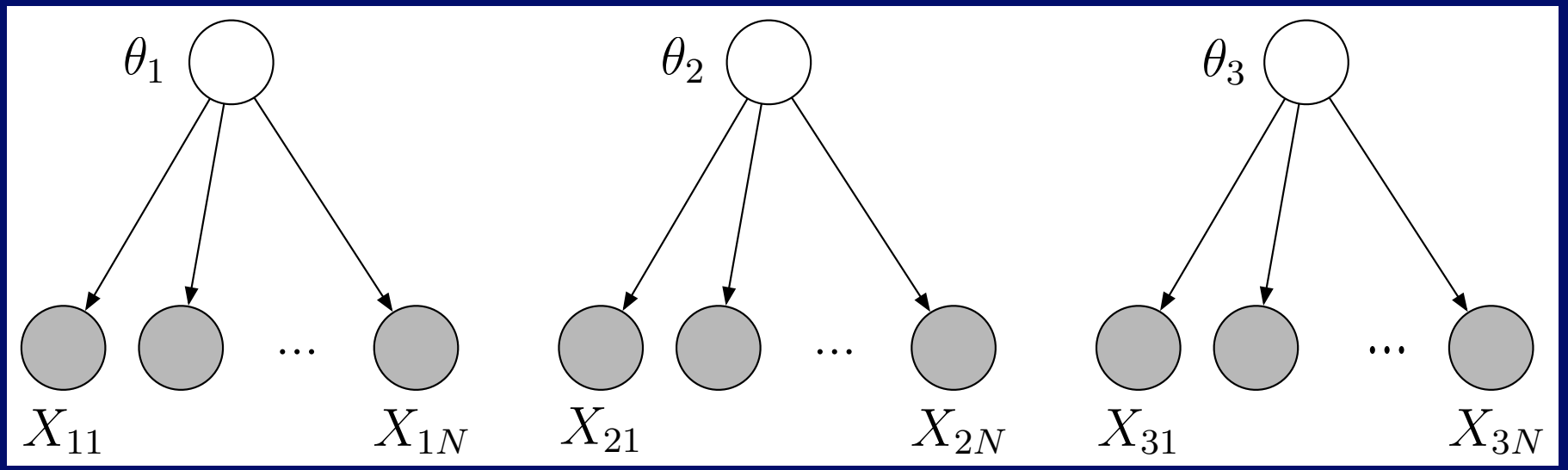
$$p(x_1, \ldots, x_N) = \int \prod_{n=1}^{N} p(x_n \mid \theta) P(d\theta)$$

Representation theorem



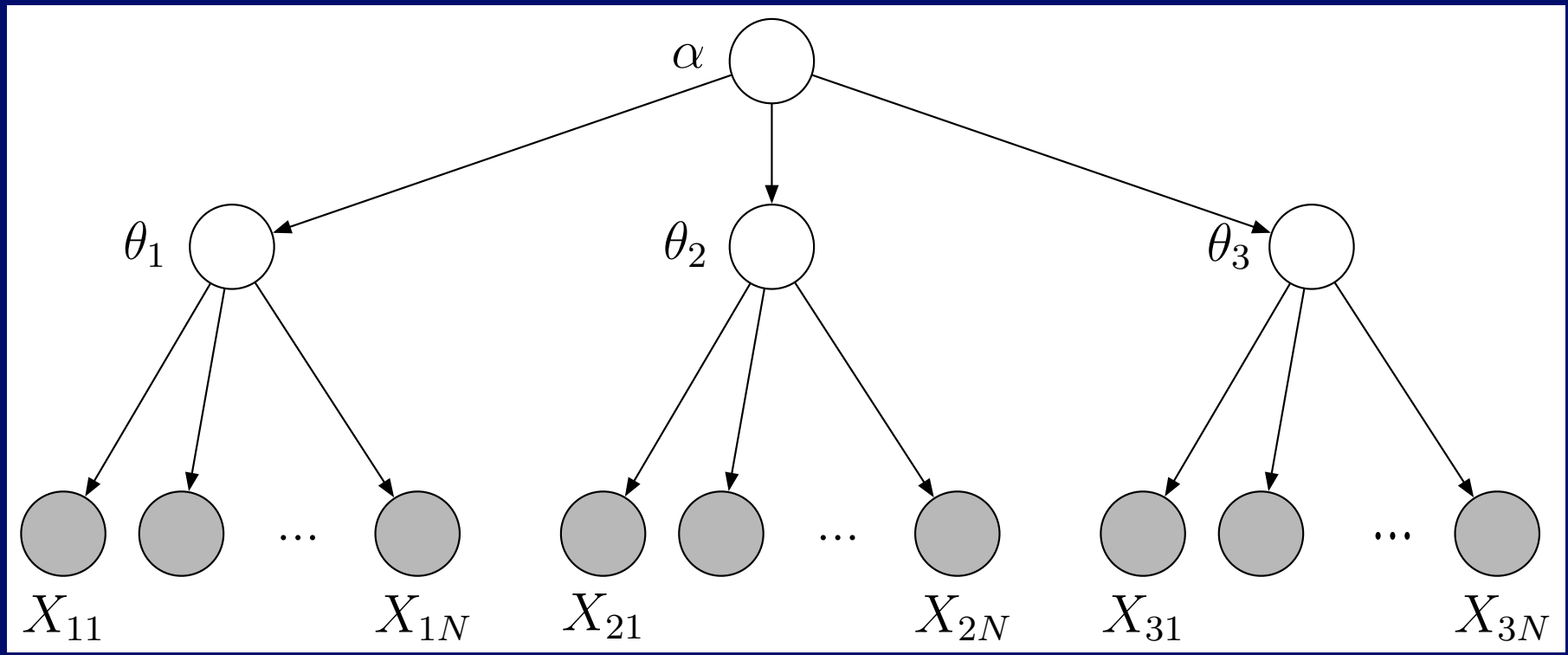
- Exchangeability is a weaker assumption than independent and identically distributed
- For many statisticians, de Finetti's theorem provides a solid justification of the Bayesian perspective

Hierarchical models



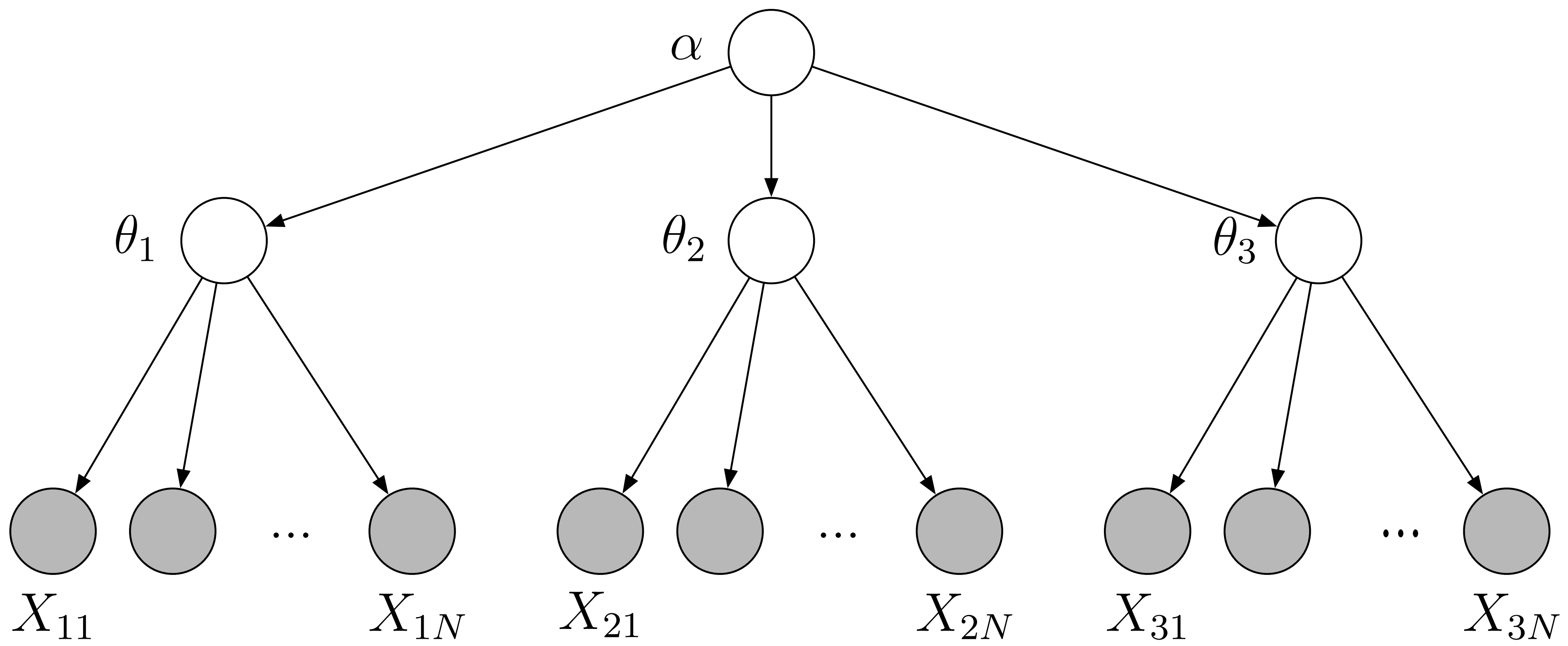
- Suppose our data is *grouped*
  - E.g., Gaussian data sets from a different means
- The data within each group is exchangeable
- Often we don't want to treat these groups independently

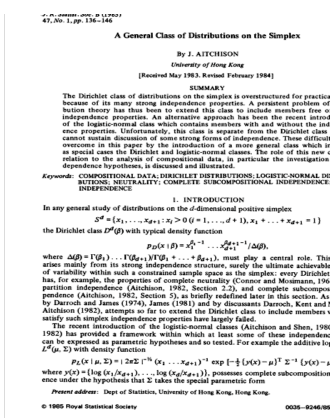
Hierarchical models



- A hyperparameter can be shared by all the groups
- Information about one group propagates to the other groups through the hyperparameter
- Groups are not independent, but they are exchangeable

Documents are grouped data!





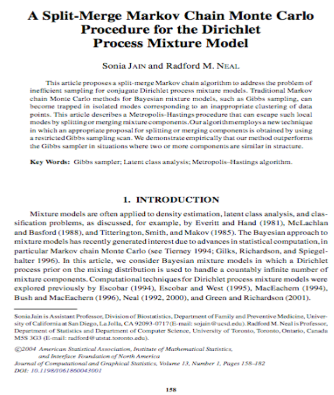
A General Class of Distributions on the Simplex

By J. AITCHISON

University of Hong Kong

Probabilistic Classification and Clustering in Relational Data

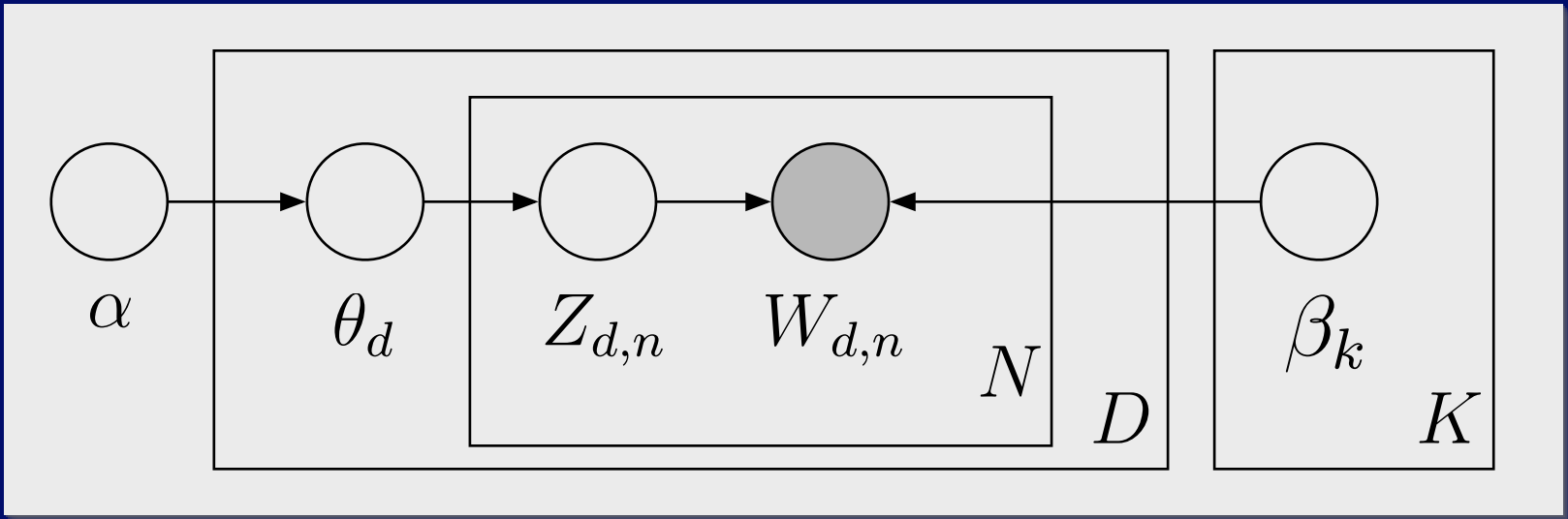
Ben Taskar, Eran Segal, Daphne Koller



A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model

Sonia JAIN and Radford M. NEAL

LDA posterior inference



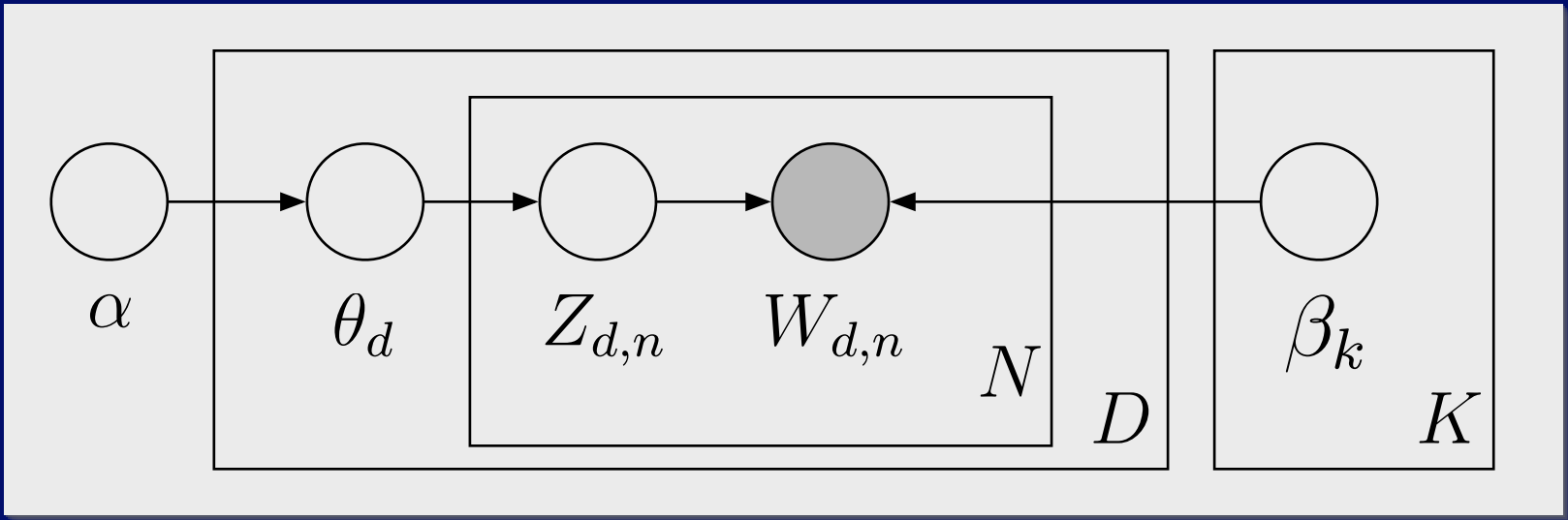
- Posterior inference : reverse the generative process
- For a collection, infer the topics which describe it
- For a document, infer proportions and assignments

$$p(\theta, z_{1:N} \mid w_{1:N}, \alpha, \beta)$$

Organizing a corpus

(Dave, switch to Firefox.)

LDA posterior inference



- Computing the posterior is intractable:

$$p(\theta, z_{1:N} \mid w_{1:N}, \alpha, \beta_{1:K}) = \frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_z p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

- (For now, assume the model is fixed.)

Variational inference

- Fast approximate technique for estimating posterior distributions (alternative to MCMC)
- Basic idea
  - Posit a factorized distribution of the latent variables with free parameters
  - Minimize the distance between the factorized distribution and the true posterior

Variational inference for LDA

- Define a factorized distribution of latent variables

$$q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

- Minimize KL divergence by iterating between

$$\begin{aligned} \phi_{ni} &\propto \beta_{iw_n} \exp\{\mathrm{E}_q[\log(\theta_i) \mid \gamma]\} \\ \gamma_i &= \alpha_i + \textstyle\sum_{n=1}^{N} \phi_{ni}. \end{aligned}$$

- Each iteration is O(N(K+1))

Modeling *Science*

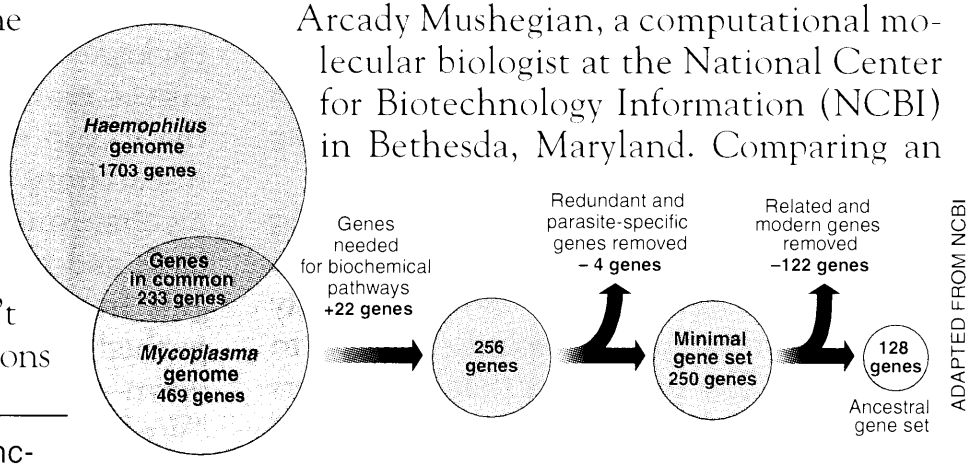
- Collection from JSTOR
  - 17,000 OCR'ed articles from the journal *Science*
  - 30,000 unique terms
- Estimated a 100-topic LDA model (~ 20 hours)
  - 100 distributions over words β_k
  - Dirichlet parameters α

An article from *Science*

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.

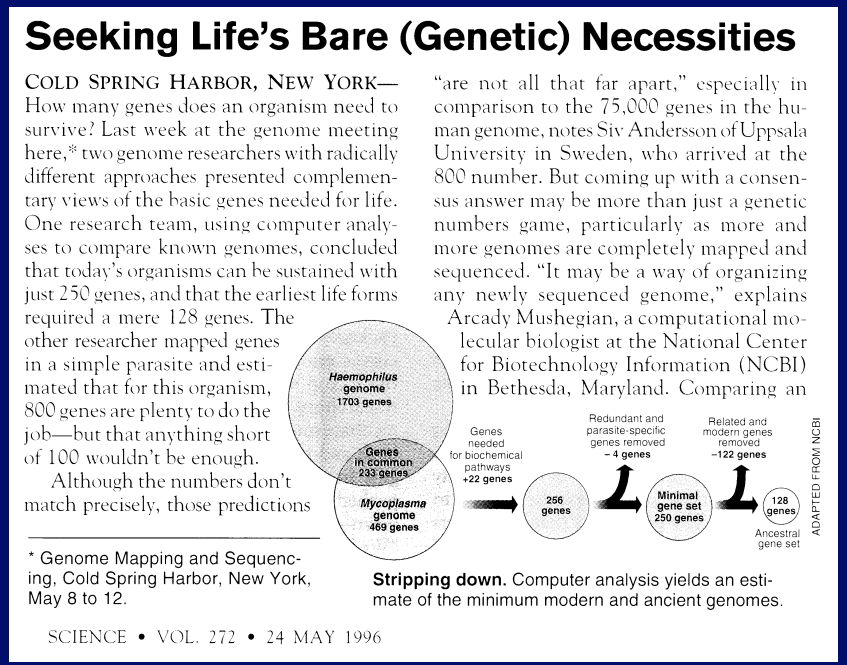
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

Approximate topic proportions



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.

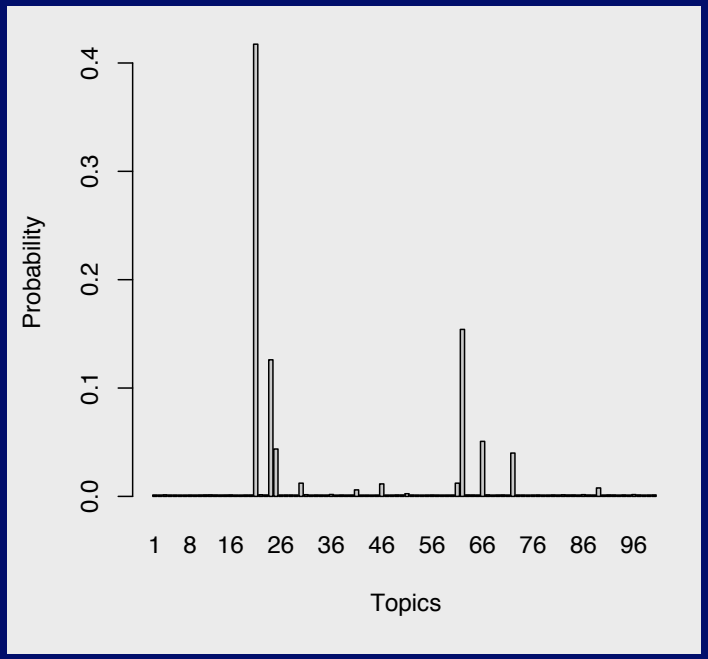
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

Posterior inference



- Low dimensional representation of the document
- Most probable words from the likely topics reflect themes of the document

"Genetics"	"Evolution"	"Research"	"Disease"	"Computers"
human	evolution	says	disease	computer
genome	evolutionary	researchers	host	models
dna	species	colleagues	bacteria	information
genetic	organisms	team	diseases	data
genes	life	just	resistance	computers
sequence	origin	like	bacterial	system
gene	biology	new	new	network
molecular	groups	work	strains	systems
sequencing	phylogenetic	years	control	model
map	living	called	infectious	parallel
information	diversity	dont	malaria	methods
genetics	group	say	parasite	networks
mapping	new	get	parasites	software
project	two	see	united	new
sequences	common	university	tuberculosis	simulations

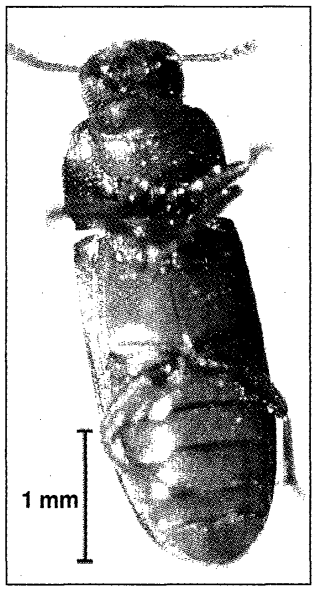
Another article

Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).



Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

"Mathematics"	"Modeling"	"Populations"	"Ecology"
problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

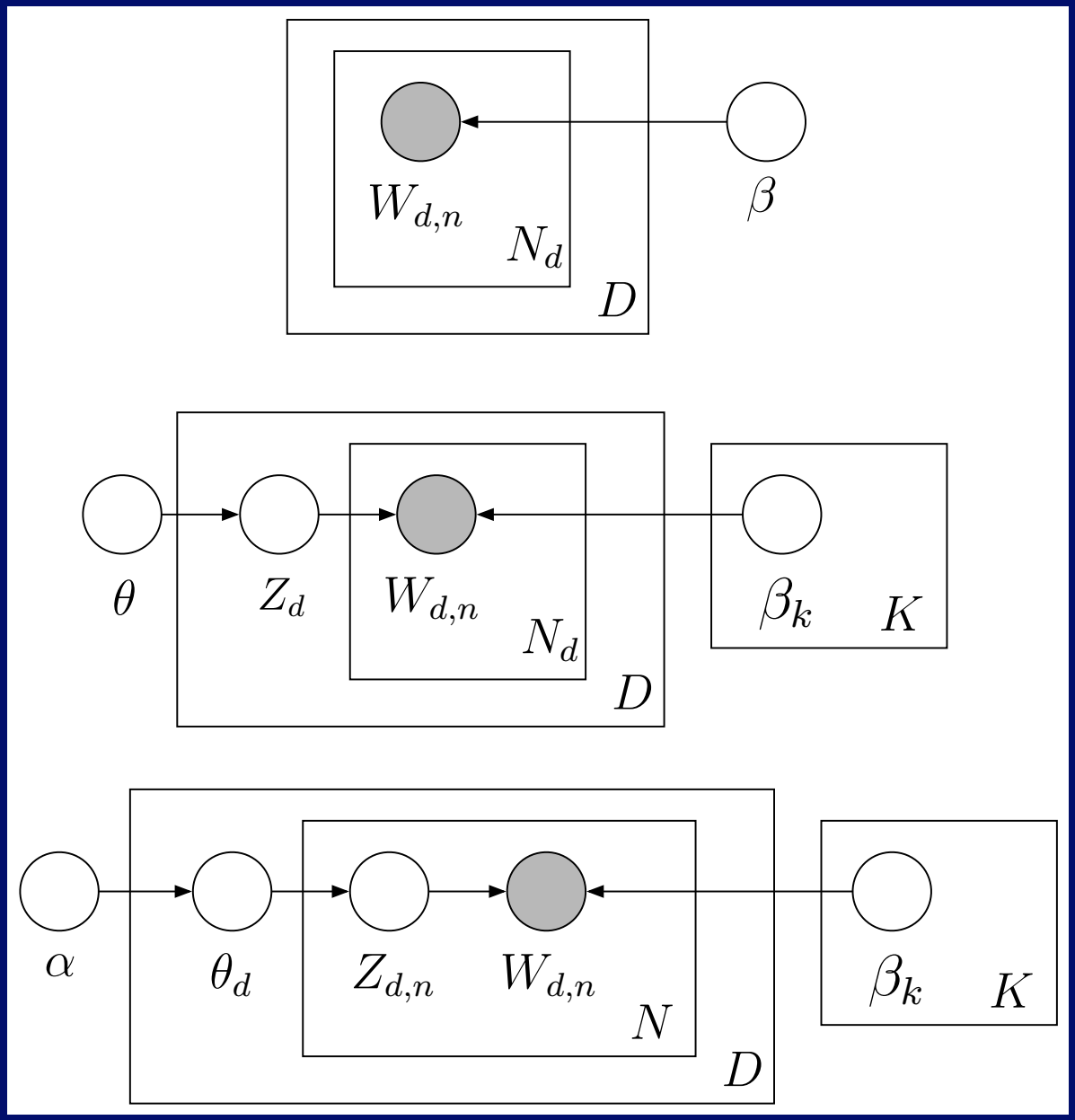
Learning the topics from data

- Variational expectation-maximization algorithm
- Iterate between
  - E step: For each document, estimate the posterior
  - M step: Estimate the topics from expected sufficient statistics, where the expectation is taken with respect to the variational distributions for each document

Quantitative evaluation

- Estimate latent variable models from a collection
- Compute likelihood of a held-out set of documents
  - Better models assign higher likelihood
- We use *perplexity*
  - Related to inverse likelihood, effective vocabulary size
  - (I.e., lower numbers are better)

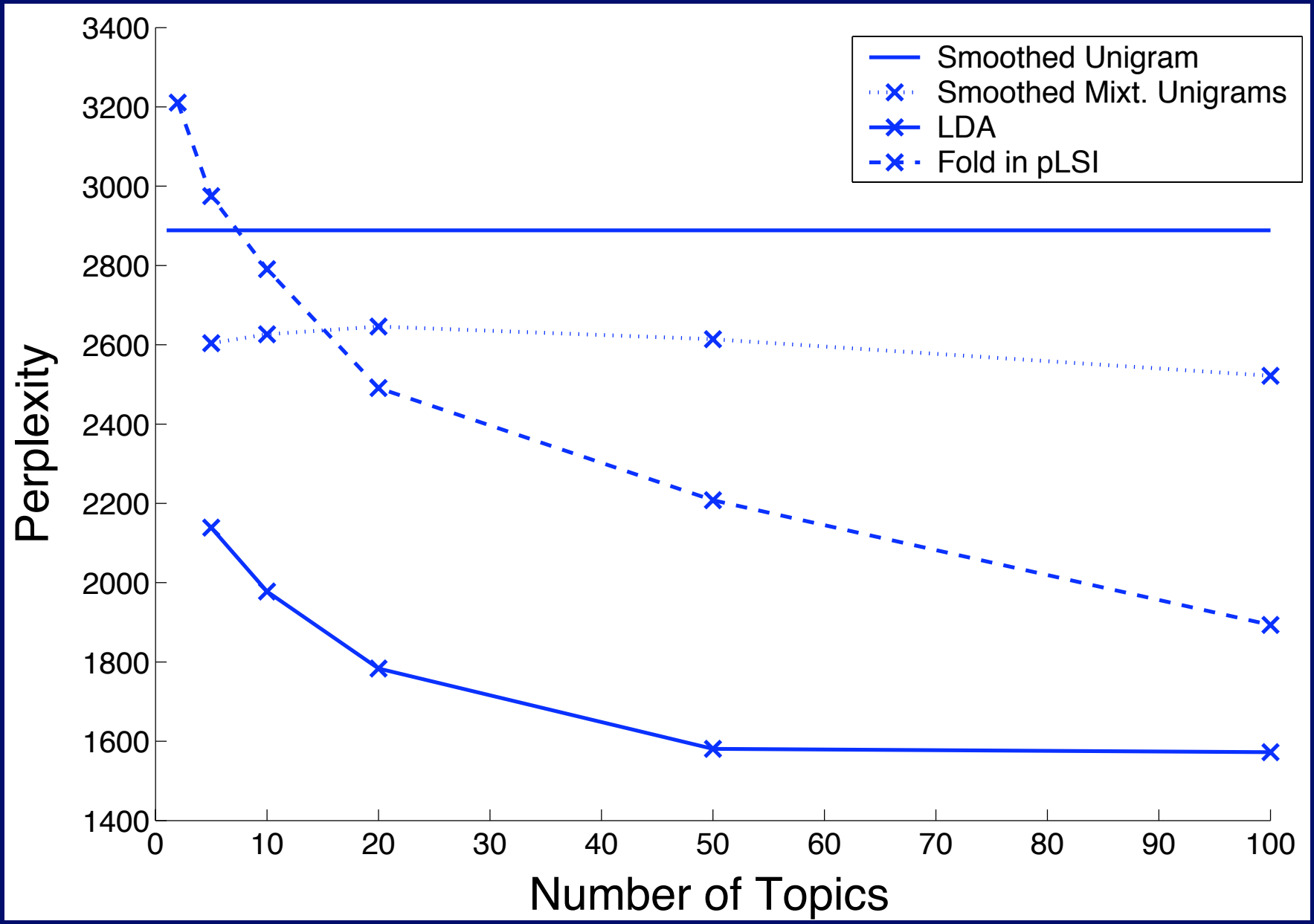
Empirical evaluation



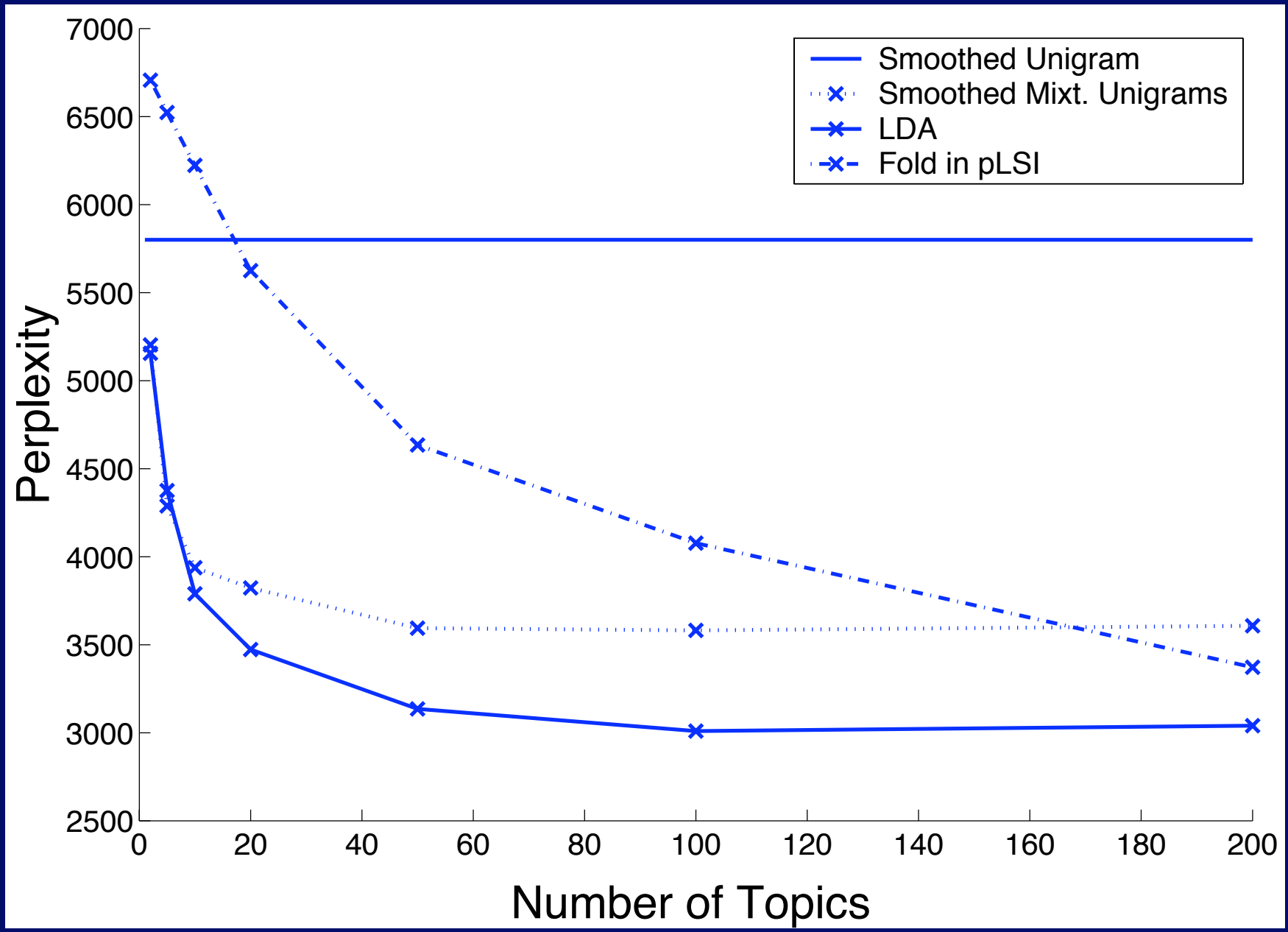
Corpora

- Biology collection
  - 5,000 abstracts about nematode biology
  - 500,000 words
  - 25,000 unique terms
- Associated Press collection
  - 16,000 document subset of the TREC AP collection
  - 775,000 words
  - 22,000 unique terms

Nematode held-out perplexity



AP held-out perplexity



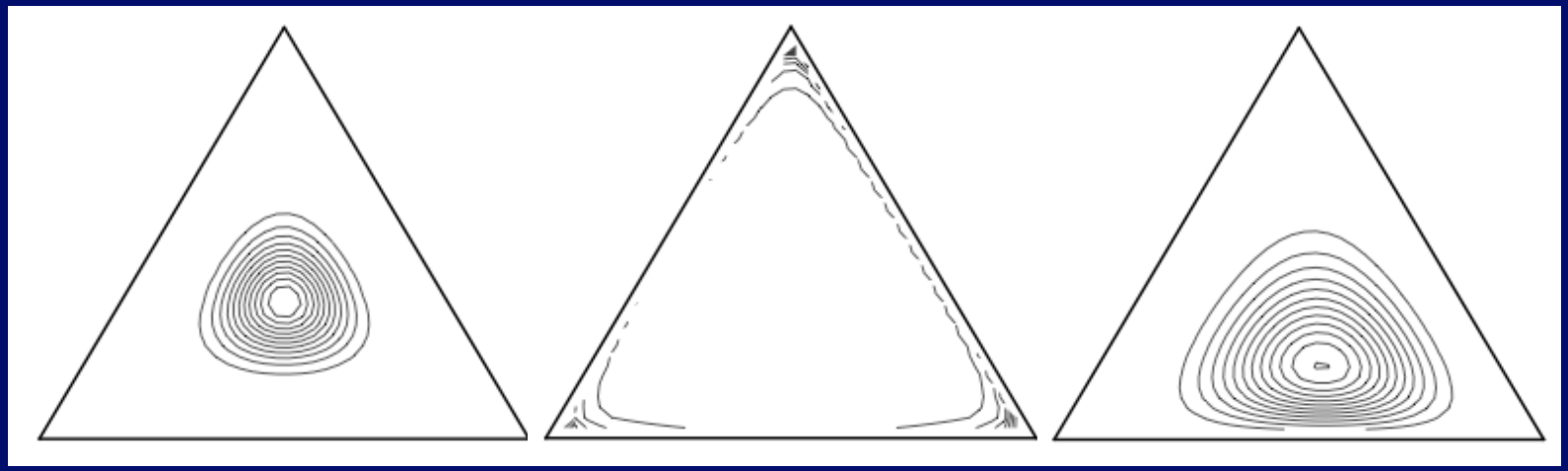
For further reading

- LDA (Blei et al., 2003)
- Latent semantic indexing (Deerwester et al., 1990)
- PLSI (Hofmann, 1999)
- Hierarchical Modeling (Gelman et al., 2001)
- Expectation propagation (Minka and Lafferty, 2002)
- Gibbs sampling (Griffiths and Steyvers, 2002)
- Sequential data (Girolami et al., 2004)
- Authors and topics model (Steyvers et al., 2004)
- Vision applications (Freeman et al., 2005)

Summary (so far)

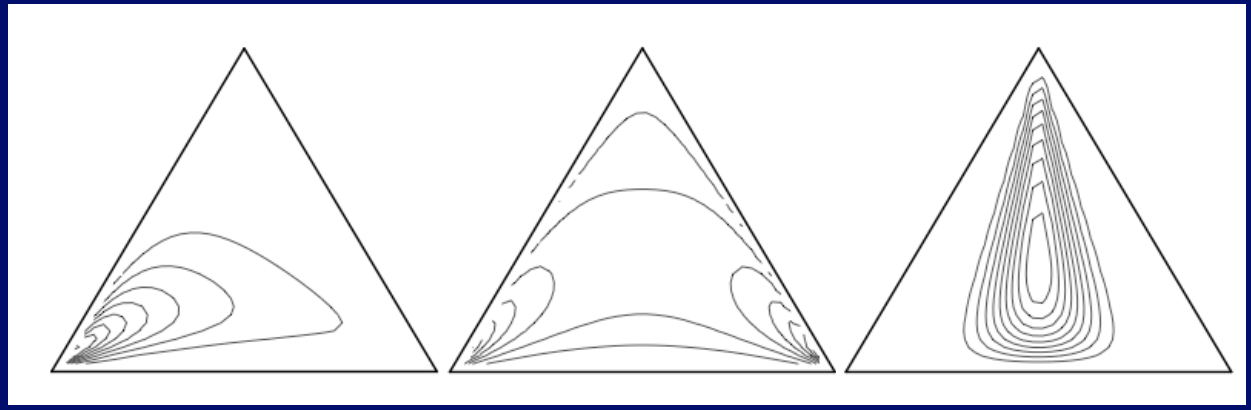
- LDA is a powerful model for
  - discovering structure in otherwise unstructured
  - generalizing new data to fit into that structure
- Graphical models are **modular**
  - Can be embedded in more complicated models
- Graphical models are (somewhat) **general**
  - LDA is primarily about the independence assumptions
  - Functional form of the data distribution can change
- Now, on to some extensions...

Correlated topics



- The Dirichlet is an exponential family distribution on the *simplex*, positive vectors which sum to one.
- However, the near independence of components makes it a poor choice for modeling topic proportions.
  - An article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

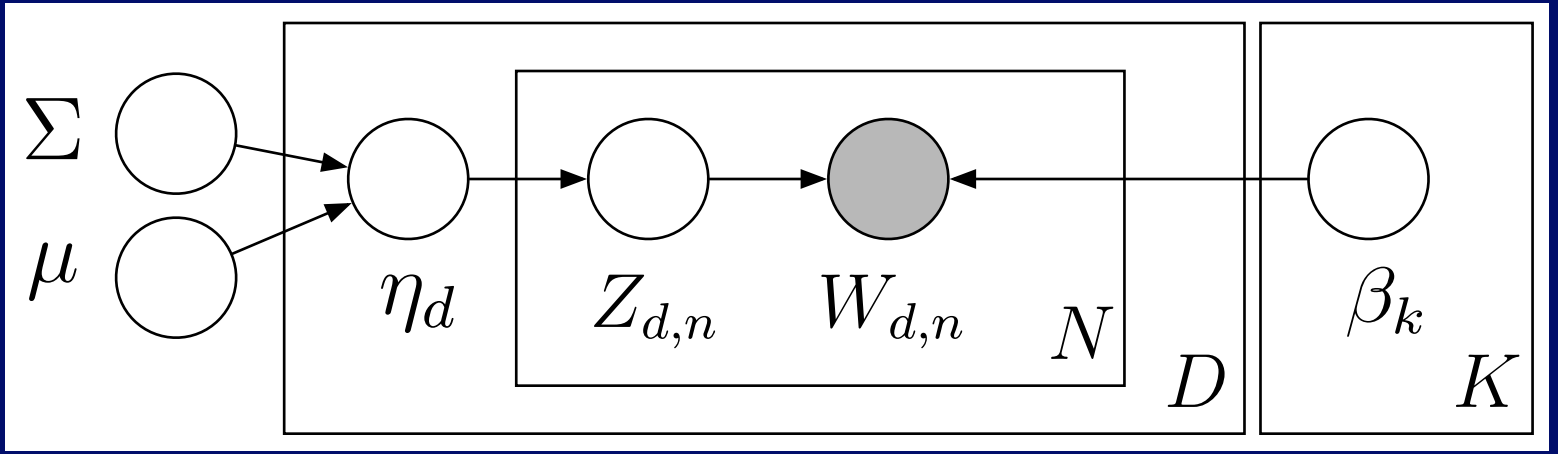
Logistic normal
(Aitchison and Shen, 1980)



- The logistic Normal is a distribution on the simplex, with a dependence between components.
- The natural parameters of the multinomial are drawn from a Gaussian distribution,

$$X \sim \mathcal{N}_{K-1}(\mu, \Sigma)$$
$$\theta_i = \exp\{x_i - \log(1 + \sum_{j=1}^{K-1} \exp\{x_j\})\}$$

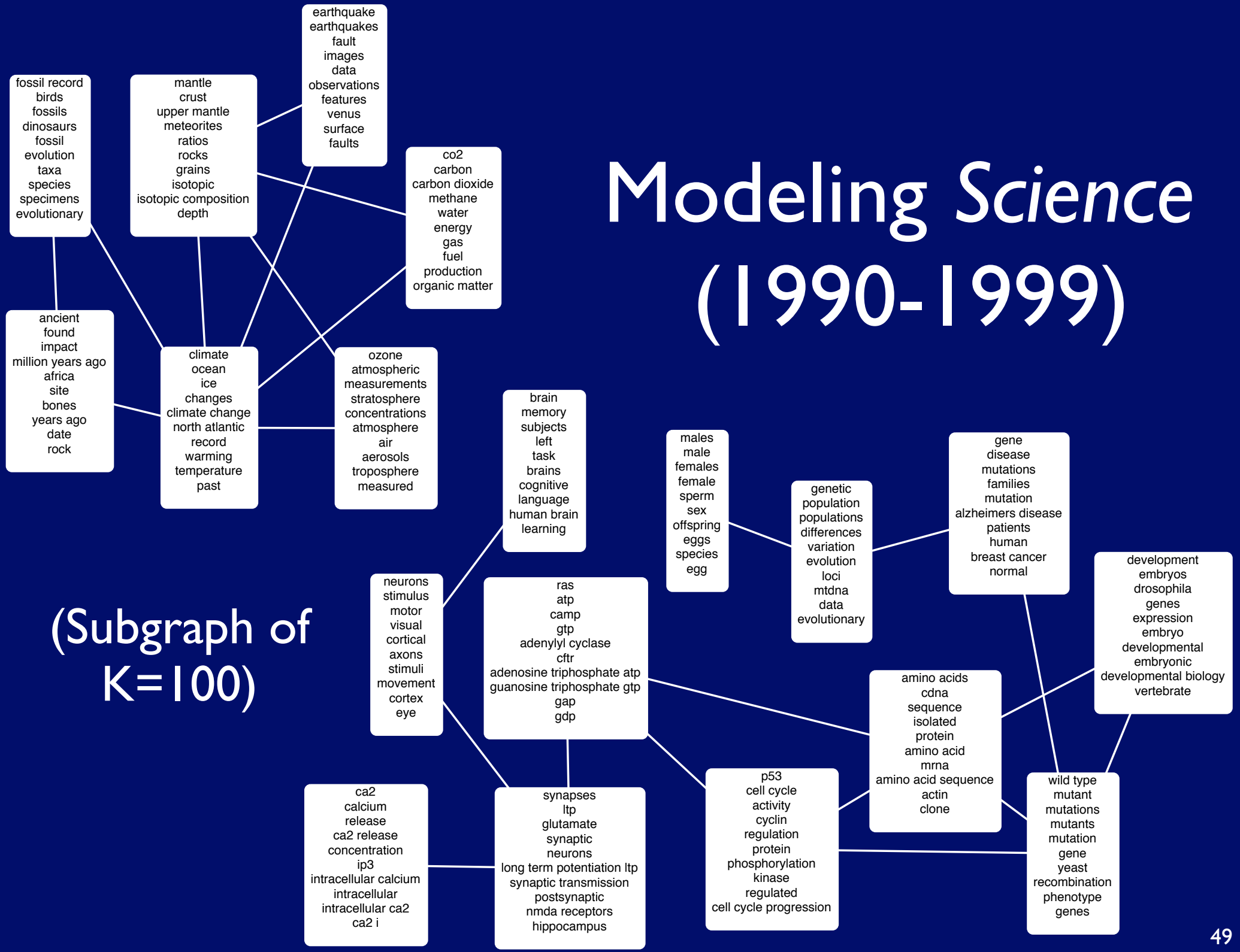
Correlated topic models



- Draw topic proportions from a logistic normal
- Useful for:
  - providing a "map" of topics and how they are related
  - better prediction via correlated topics
- We have lost conjugacy between the topic proportion model and the multinomial over topic indicators

Modeling *Science* (1990-1999)

(Subgraph of K=100)



Modeling the evolution of *Science*

- In LDA, documents are assumed to be exchangeable
  - Order doesn't matter
- This doesn't make sense; topics evolve over time
  - "Cleaning Birds" (1883)
  - "Interspecific Brood Parasitism in Blackbirds (Icterinae): A Phylogenetic Perspective" (1992)
- Many document collections have such dynamics
  - Emails, news articles, query logs, ...

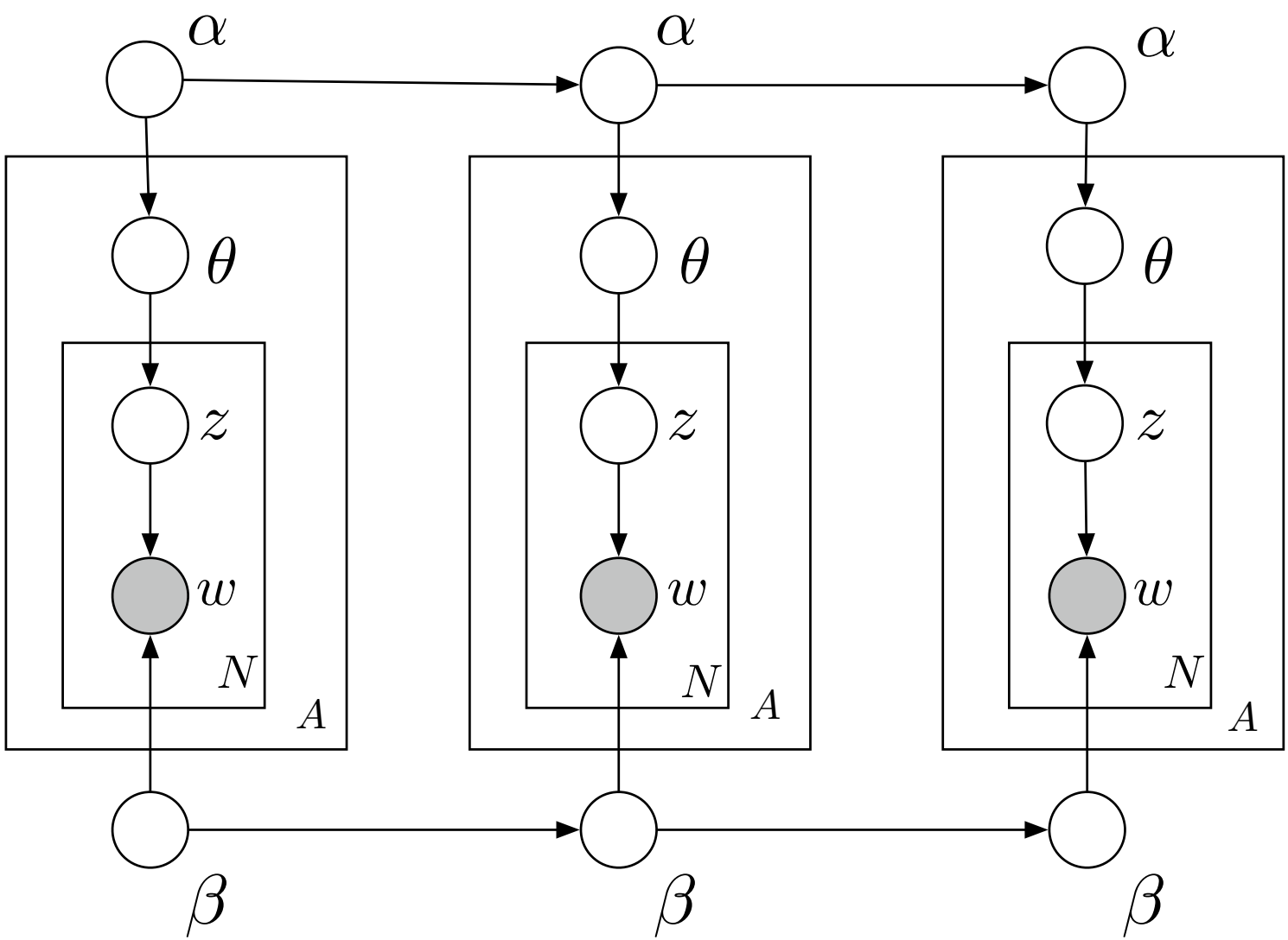
Science (1880-1883)

Administration	Limnology	Astronomy	Psychology
association	water	observatory	mind
meeting	lake	observations	nature
american	sea	stars	say
committee	waters	time	science
congress	lakes	made	psychology
members	gulf	astronomical	work
held	great	comet	knowledge
international	depth	star	truth
meetings	river	observed	religion
section	stream	telescope	human

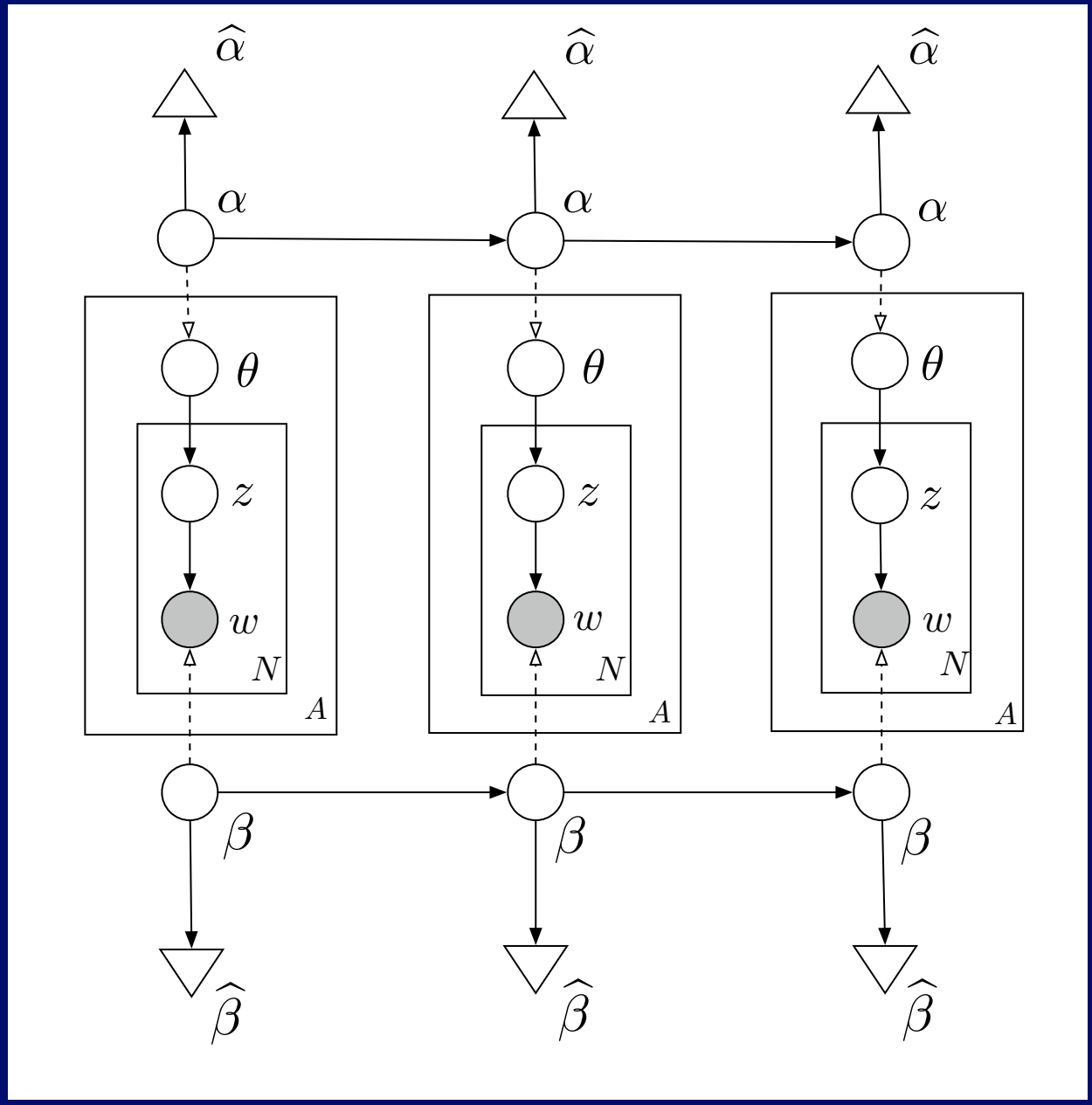
Science (1970-1976)

Administration	Limnology	Astronomy	Psychology
house	water	mass	human
congress	concentrations	radio	attempts
science	mercury	objects	theory
bill	fish	astronomy	learning
nsf	samples	xray	ideas
president	soil	stars	new
budget	lake	astronomical	memory
office	ppm	sources	psychology
committee	concentration	observations	behavior
new	waters	observatory	complex

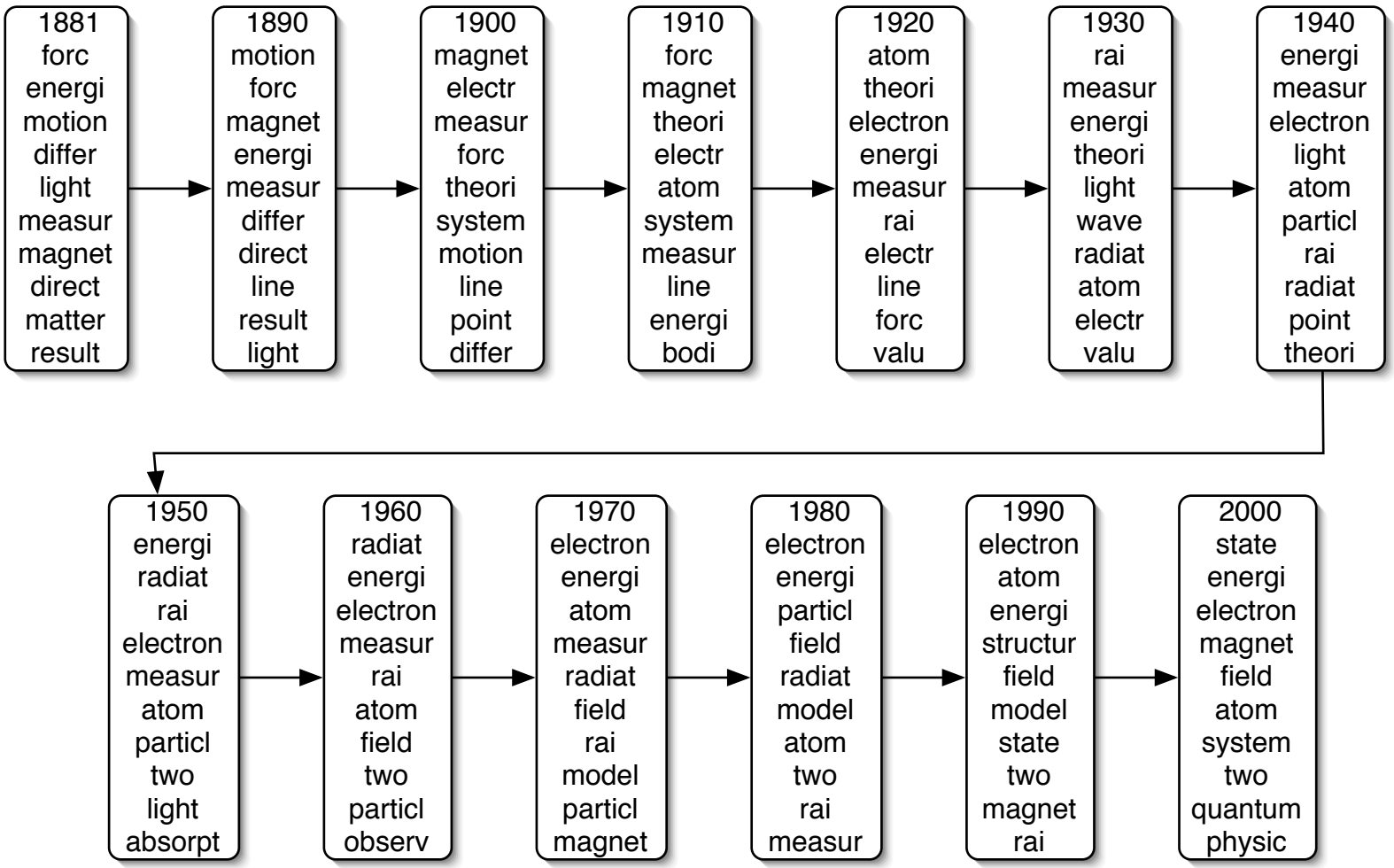
Chained topic models



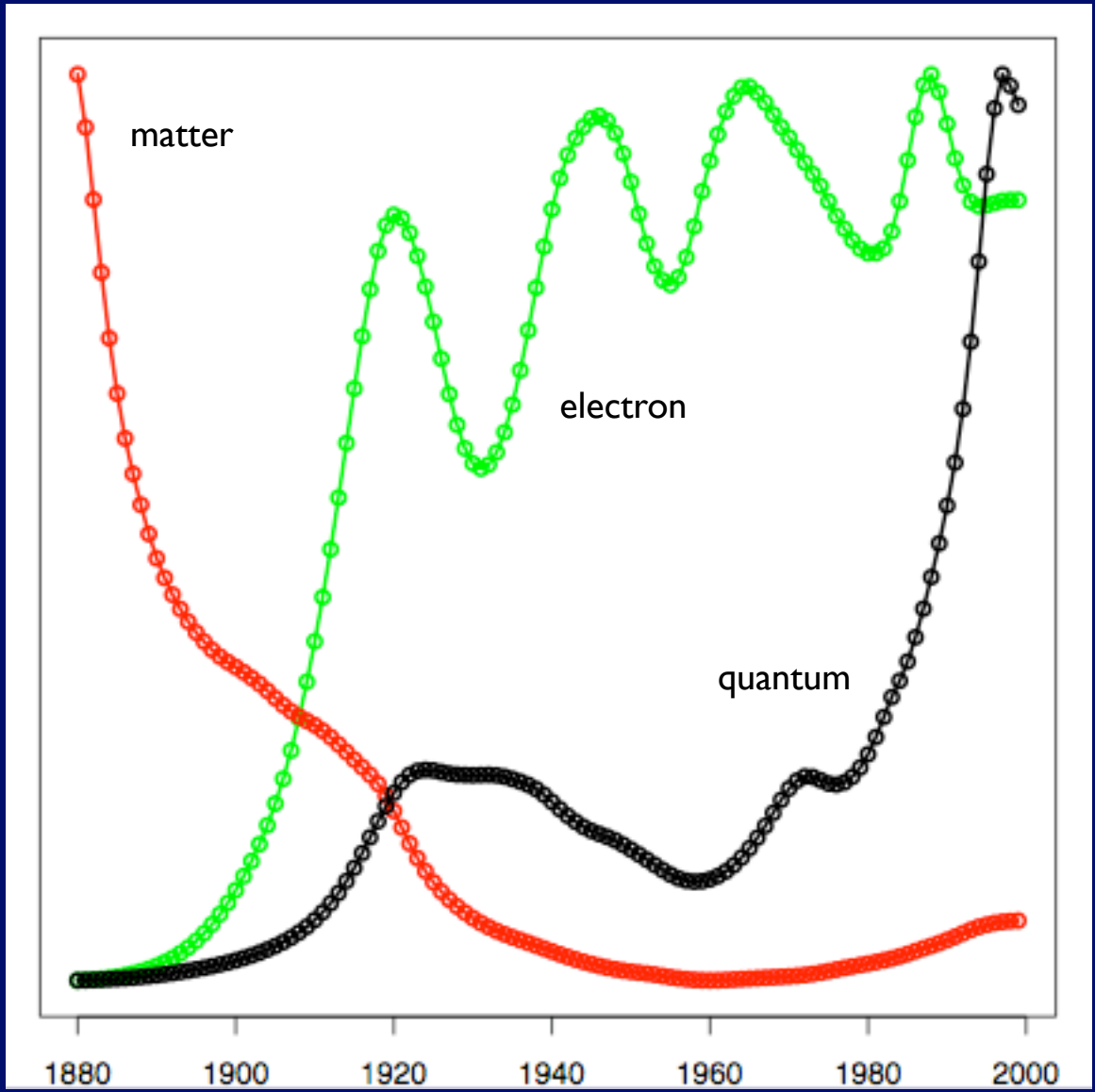
Variational distribution



“Atomic physics”



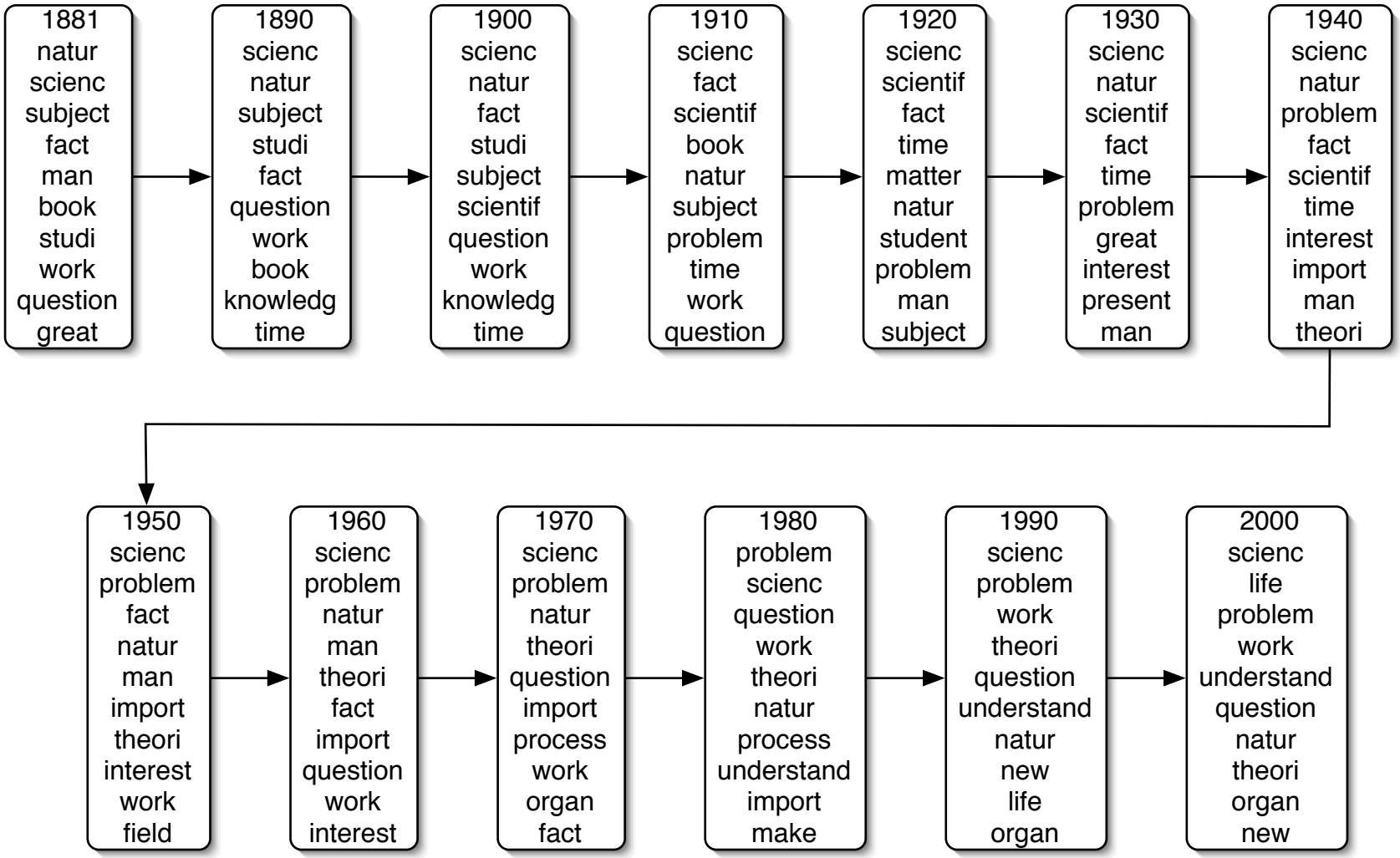
“Atomic physics”



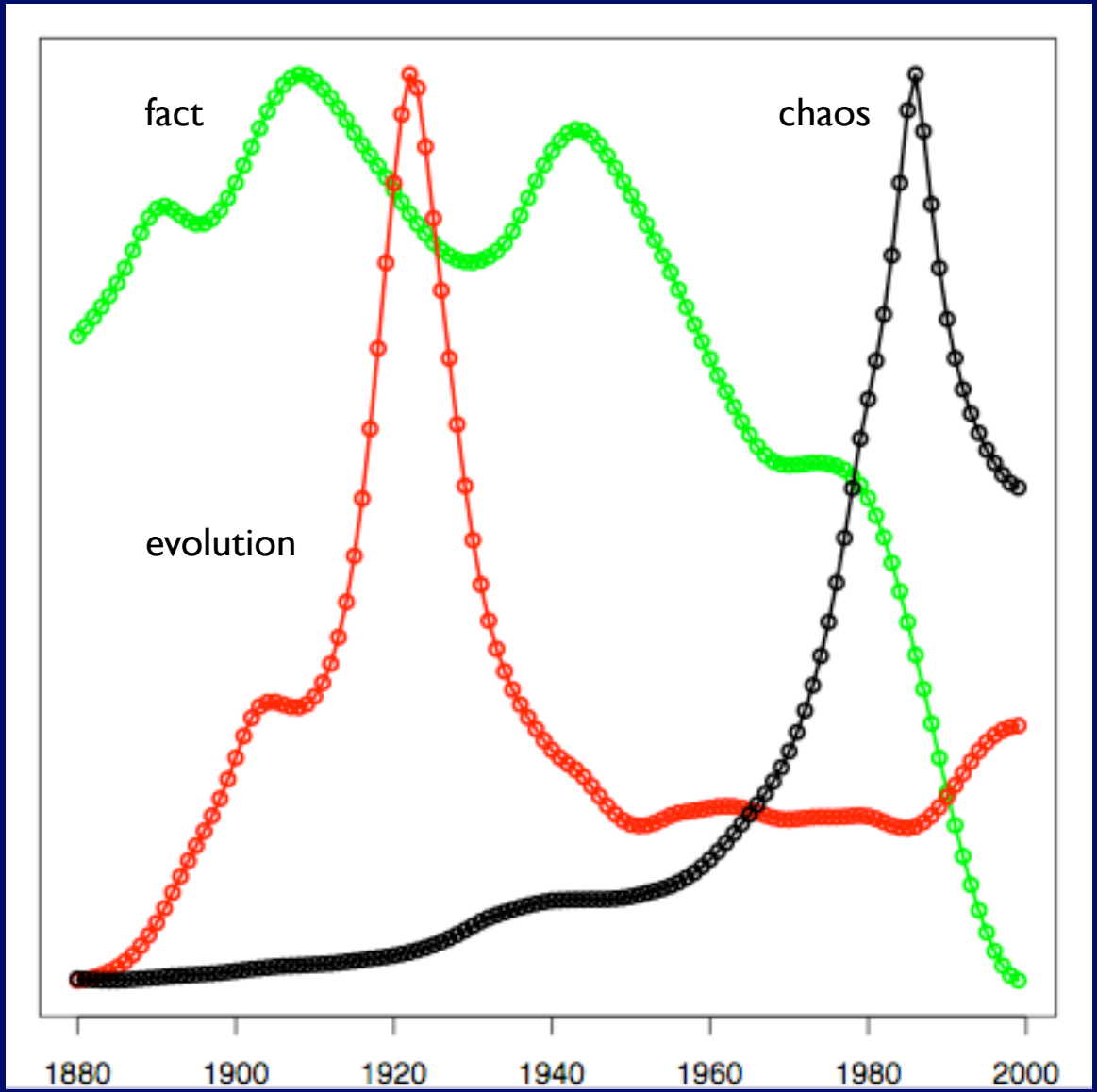
“Atomic physics”

- 1881 On Matter as a form of Energy
- 1892 Non-Euclidean Geometry
- 1900 On Kathode Rays and Some Related Phenomena
- 1917 ``Keep Your Eye on the Ball''
- 1920 The Arrangement of Atoms in Some Common Metals
- 1933 Studies in Nuclear Physics
- 1943 Aristotle, Newton, Einstein. II
- 1950 Instrumentation for Radioactivity
- 1965 Lasers
- 1975 Particle Physics: Evidence for Magnetic Monopole Obtained
- 1985 Fermilab Tests its Antiproton Factory
- 1999 Quantum Computing with Electrons Floating on Liquid Helium

“Philosophy”


1881
natur
scienc
subject
fact
man
book
studi
work
question
great
1890
scienc
natur
subject
studi
fact
question
work
book
knowledg
time
1900
scienc
natur
fact
studi
subject
scientif
question
work
knowledg
time
1910
scienc
fact
scientif
book
natur
subject
problem
time
work
question
1920
scienc
scientif
fact
time
matter
natur
student
problem
man
subject
1930
scienc
natur
scientif
fact
time
problem
great
interest
present
man
1940
scienc
natur
problem
fact
scientif
time
interest
import
man
theori
1950
scienc
problem
fact
natur
man
import
theori
interest
work
field
1960
scienc
problem
natur
man
theori
fact
import
question
work
interest
1970
scienc
problem
natur
theori
question
import
process
work
organ
fact
1980
problem
scienc
question
work
theori
natur
process
understand
import
make
1990
scienc
problem
work
theori
question
understand
natur
new
life
organ
2000
scienc
life
problem
work
understand
question
natur
theori
organ
new

“Philosophy”



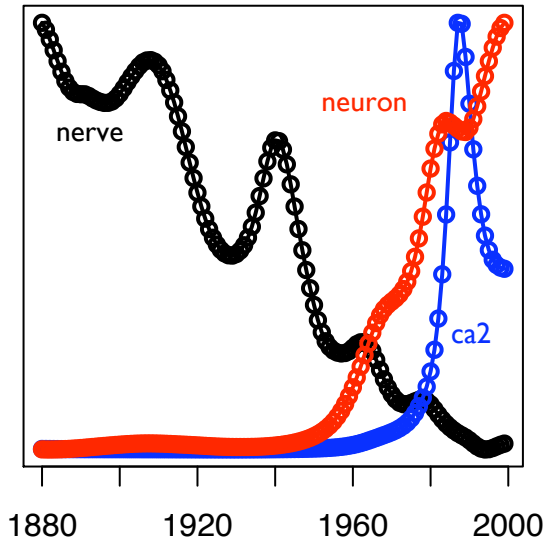
"Philosophy"

- 1881 On the Truthfulness of Human Knowledge Considered in the Light of the Unity of Nature
- 1890 The Method of Multiple Working Hypotheses
- 1901 A Final Word on Discord
- 1911 The Meaning of Vitalism
- 1921 The Spirit of Research
- 1930 The Organic World and the Causal Principle
- 1942 Evolution and Knowledge
- 1950 Social Responsibility in Science
- 1963 Decision Theory in Law, Science, and Technology
- 1974 Speaking of Science: Creativity: Can It Be Dissected? Can It Be Taught?
- 1989 Chaos Theory: How Big an Advance?
- 1991 Science, Slogans, and Civic Duty

“Neuroscience” and “Disease”

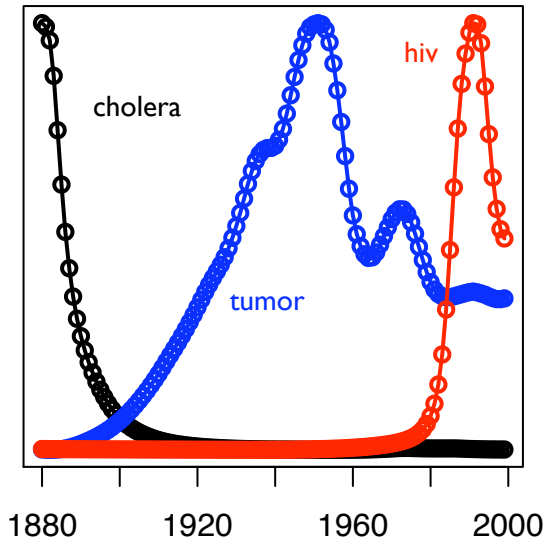
"Neuroscience"

1910 →	1920 →	1930 →	1940 →	1950 →	1960 →	1970 →	1980 →	1990 →	2000
brain	move	stimuli	record	response	response	response	cell	cell	neuron
eye	sound	muscle	nerve	record	stimuli	cell	neuron	channel	active
move	muscle	sound	stimuli	stimuli	record	potential	response	neuron	brain
right	active	move	response	nerve	condit	stimuli	active	ca2	cell
left	nerve	response	muscle	muscle	active	neuron	brain	active	fig
hand	stimuli	nerve	electrode	active	potential	active	stimuli	brain	response
nerve	fiber	frequency	active	frequency	stimuliu	nerve	muscle	receptor	channel
vision	reaction	fiber	brain	electrode	nerve	eye	system	muscle	receptor
sound	brain	active	fiber	potential	subject	record	nerve	response	synaptic
muscle	response	brain	potential	studi	eye	abstract	receptor	current	signal



"Disease"

1910 →	1920 →	1930 →	1940 →	1950 →	1960 →	1970 →	1980 →	1990 →	2000
disease	disease	disease	virus	virus	cell	cell	cell	cell	cell
medical	infect	virus	disease	tumor	virus	virus	human	disease	disease
medicine	medicine	infect	infect	cancer	serum	human	disease	infect	infect
cause	medical	inoculate	mice	serum	tumor	antigen	antibody	cancer	human
organ	inoculate	anim	tumor	tissue	mice	tumor	cancer	virus	immune
infect	organ	serum	serum	blood	antibody	infect	virus	human	virus
fever	clinic	immune	strain	mice	infect	serum	infect	hiv	cancer
blood	anim	tissue	tissue	infect	tissue	culture	tumor	antibody	tumor
health	patholog	mice	inoculate	disease	antigen	disease	patient	aids	host
inoculate	blood	strain	blood	human	human	normal	antigen	immune	vaccine



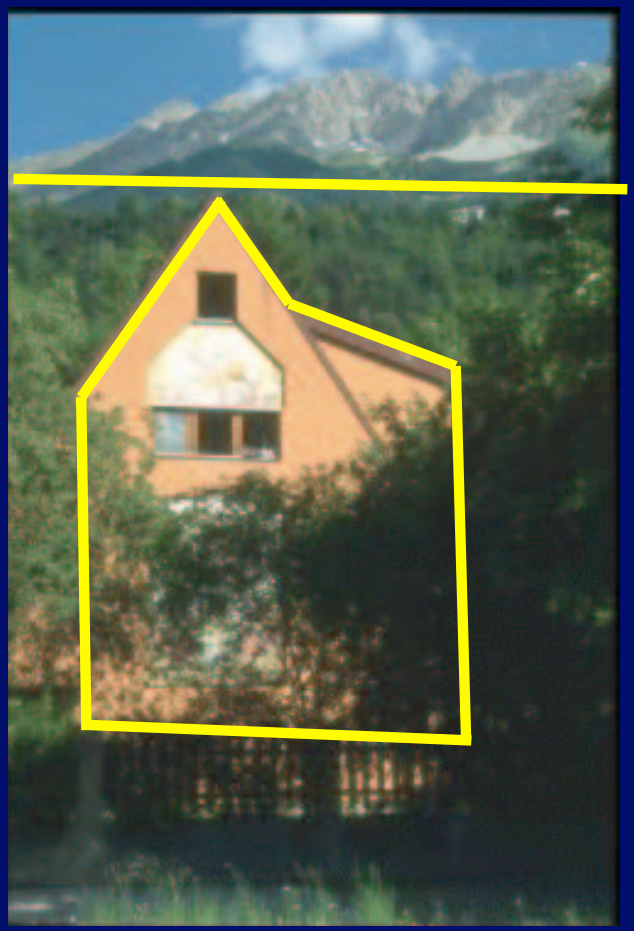
Annotated data

- Data with multiple types
  - One type is a *description*
- For example
  - Images and captions
  - Gene expression and function
- Tasks
  - Clustering, classification
  - Annotation, image retrieval

house, sky, trees

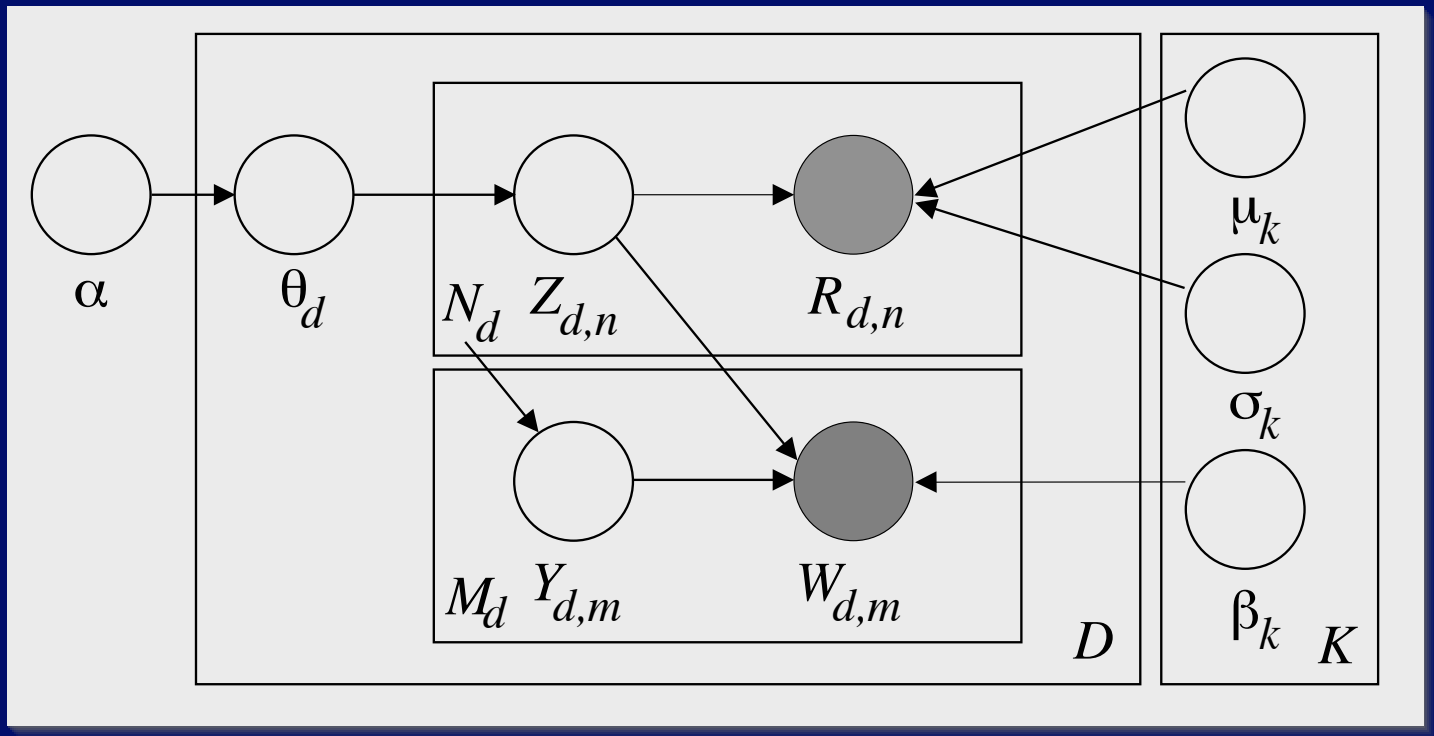
Documents to images

- *Document*: bag of words from a discrete vocabulary
  - multinomial data
- *Image:* bag of regions described by continuous feature vectors
  - Gaussian data



<-0.61 0.63 -0.31>
< 0.19 0.21 -0.48>
< 0.06 1.15 0.05>

Correspondence LDA



- Image is generated from a Gaussian LDA model
- For each caption word
  - Choose a *region* $Y \in \{1, \ldots, N_d\}$ uniformly at random
  - Choose the word from $W \sim \text{Mult}(\beta_{z_y})$

Corel data

sky, clouds, rock

tiger, ground

house, sky, trees

penguins, water

- 6,500 images and captions from the Corel database
  - Annotated with 2 - 6 words
  - Segmented with N-cuts (Shi and Malik, 2000)
  - Each region reduced to 45 continuous features
- 5000 training instances, 1500 held-out instances

Application: annotation






Application: annotation

sky water tree
mountain people

scotland water
flower hills tree

sky water buildings
people mountain

fish water ocean
tree coral

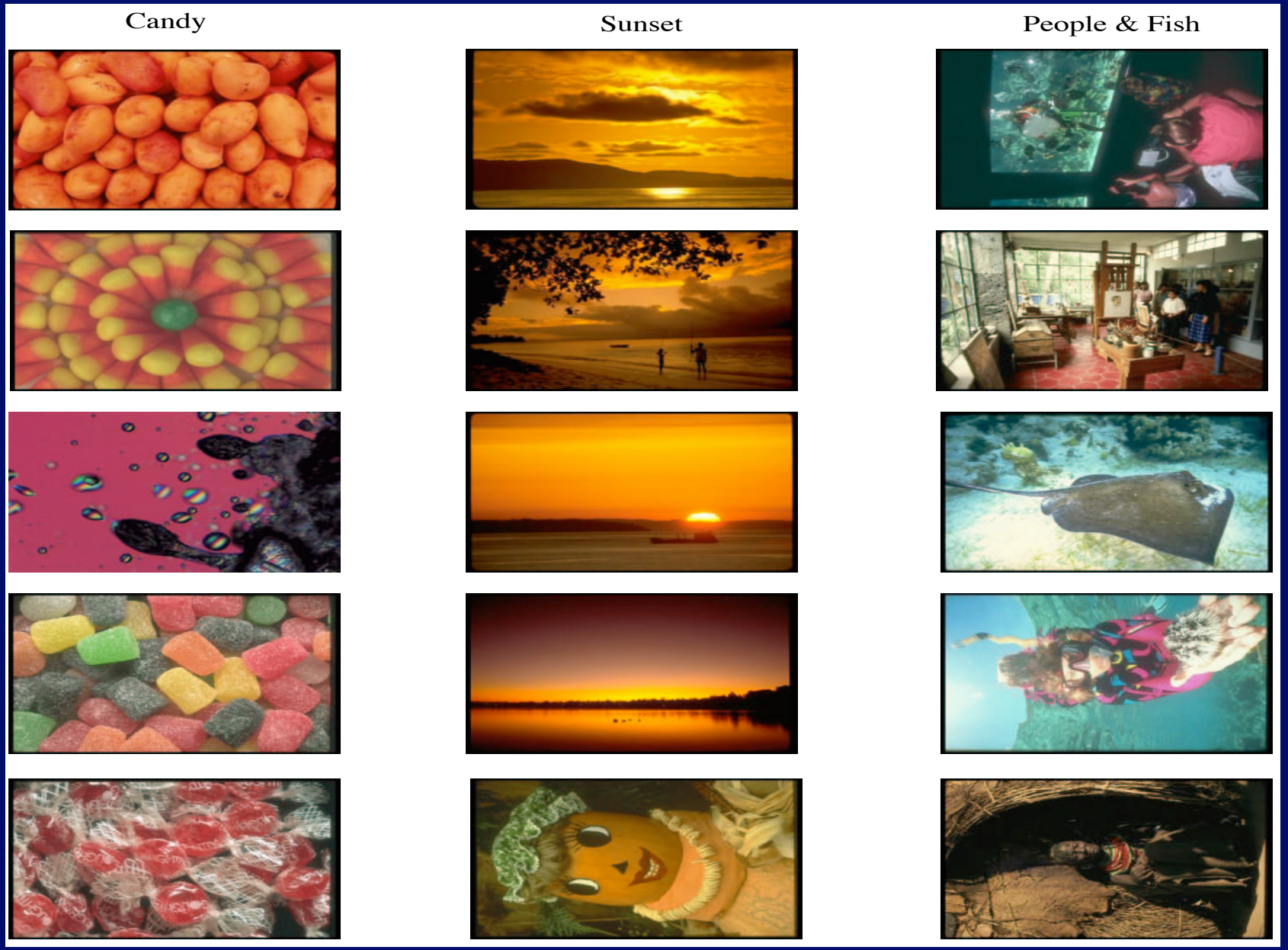
people market
pattern textile display

birds nest
leaves branch tree

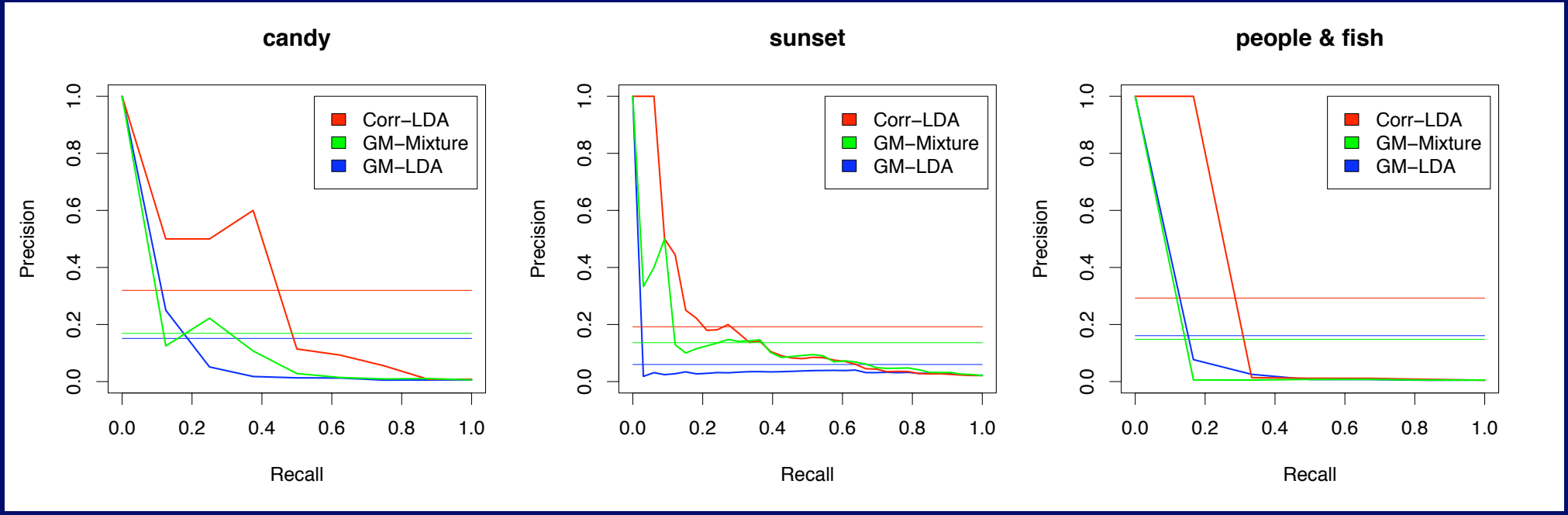
Application: image retrieval

- Given a collection of *unannotated* images
- Find images relevant to a text query
  - Rank according to p(query | image) for each image
- A form of the *language modeling* retrieval approach

Sample queries



Precision/recall curves



- For a ranking of N images (out of M images)
  - Precision = # relevant items / list size
  - Recall = # relevant items / total # relevant items

Summary

- Generative probabilistic models of collections that represents the heterogeneous nature of documents
  - Documents can exhibit multiple topics
- Inference
  - Variational inference allows large scale analysis
  - Variational EM is a natural map-reduce application
- Advantages of graphical models formalism
  - Modular: Can plug LDA into more complicated models
  - General: Applicable to different kinds of data

Summary

- C code to play with the basic LDA model
  - http://www.cs.cmu.edu/~lemur/science
- Explore a correlated topic model
  - http://www.cs.cmu.edu/~lemur/science/