

STCS6701 : Probabilistic Models and Machine Learning

David M. Blei
Columbia University

Today's lecture

- ▶ Probabilistic machine learning
- ▶ Example of a probabilistic model : Latent Dirichlet allocation
- ▶ Box's loop
- ▶ What will we cover?
- ▶ Prerequisites, workload, course logistics



We have **complicated data**; we want to **make sense** of it.



What is **complicated data**?

- ▶ many data points; many dimensions
- ▶ complex structures: parse trees, networks of data, ...
- ▶ multimodal: images, links, text, clicks, ...



What is **making sense of data**?

- ▶ make predictions about the future
- ▶ uncover interpretable patterns
- ▶ do science: confirm, elaborate, form causal theories



PROBABILISTIC MACHINE LEARNING

- ▶ ML methods that **connect domain knowledge to data**
- ▶ Provides a computational methodology for scalable data analysis
- ▶ Goal: A methodology that is **expressive, scalable, easy to develop**



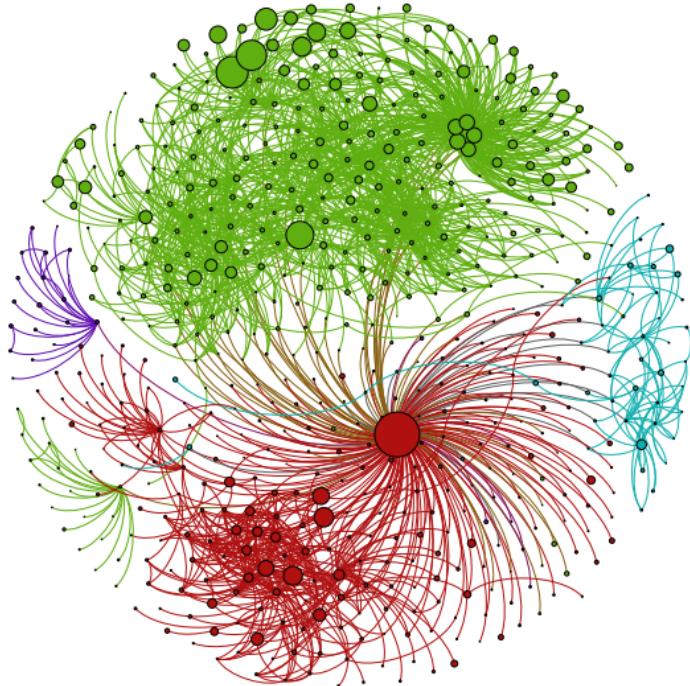
APPLIED BAYESIAN STATISTICS

- ▶ Statistical methods that **connect domain knowledge to data**.
- ▶ Provides a computational methodology for scalable data analysis
- ▶ Goal: A methodology that is **expressive, scalable, easy to develop**



MAIN IDEA

- ▶ Cast hidden and observed quantities into a **joint distribution**, a generative process that produces observations from a scaffolding of hidden variables.
- ▶ Make inferences about the hidden variables with the **posterior**.
- ▶ Make predictions with the **posterior predictive**.



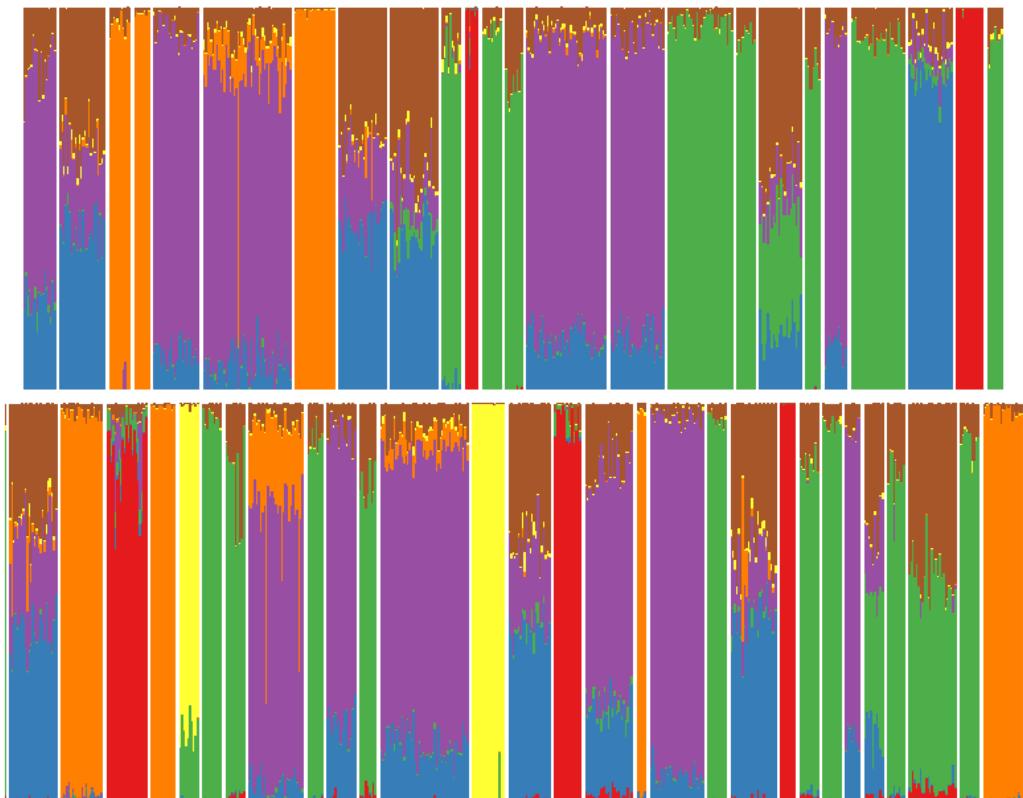
Communities discovered in a 3.7M node network of U.S. Patents

[Gopalan and Blei PNAS 2013]



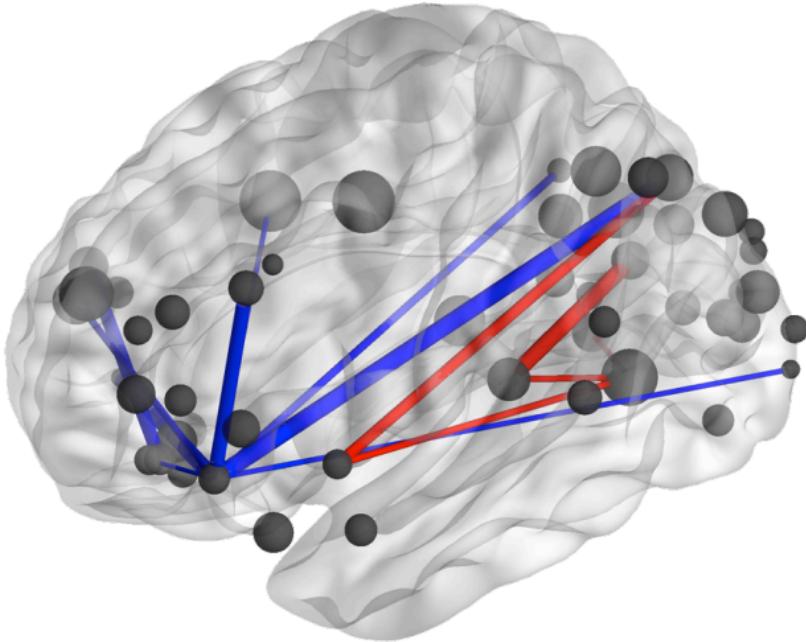
Topics found in 1.8M articles from the New York Times

[Hoffman+ JMLR 2013]



Population analysis of 2 billion genetic measurements

[Gopalan+ Nature Genetics 2016]



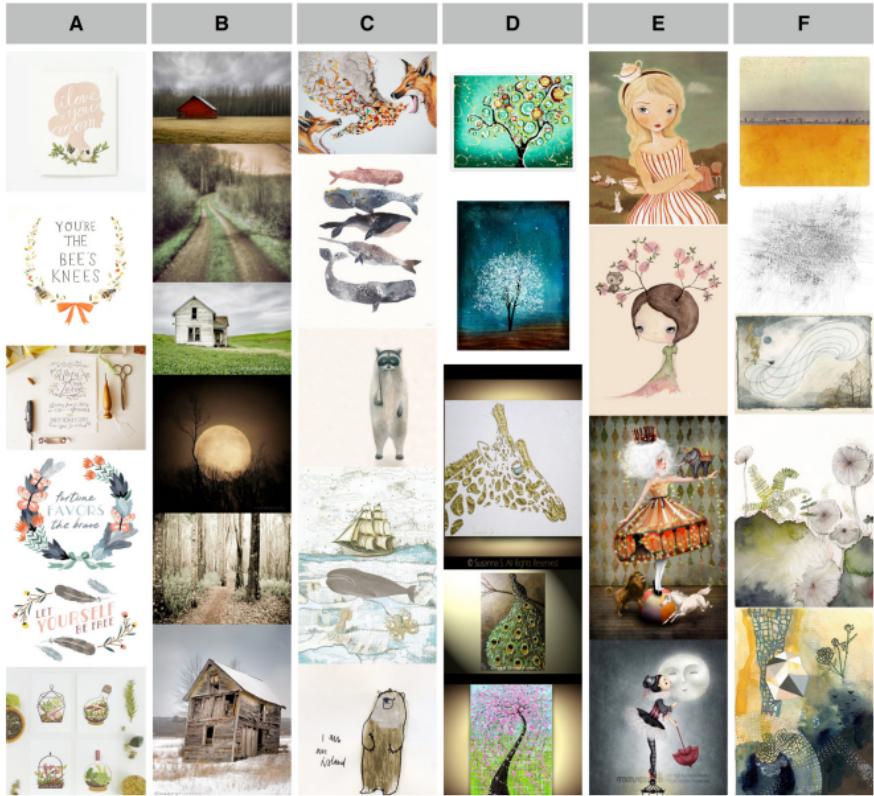
Neuroscience analysis of 220 million fMRI measurements

[Manning+ PLOS ONE 2014]



SHOPPER analysis of 5.7M purchases

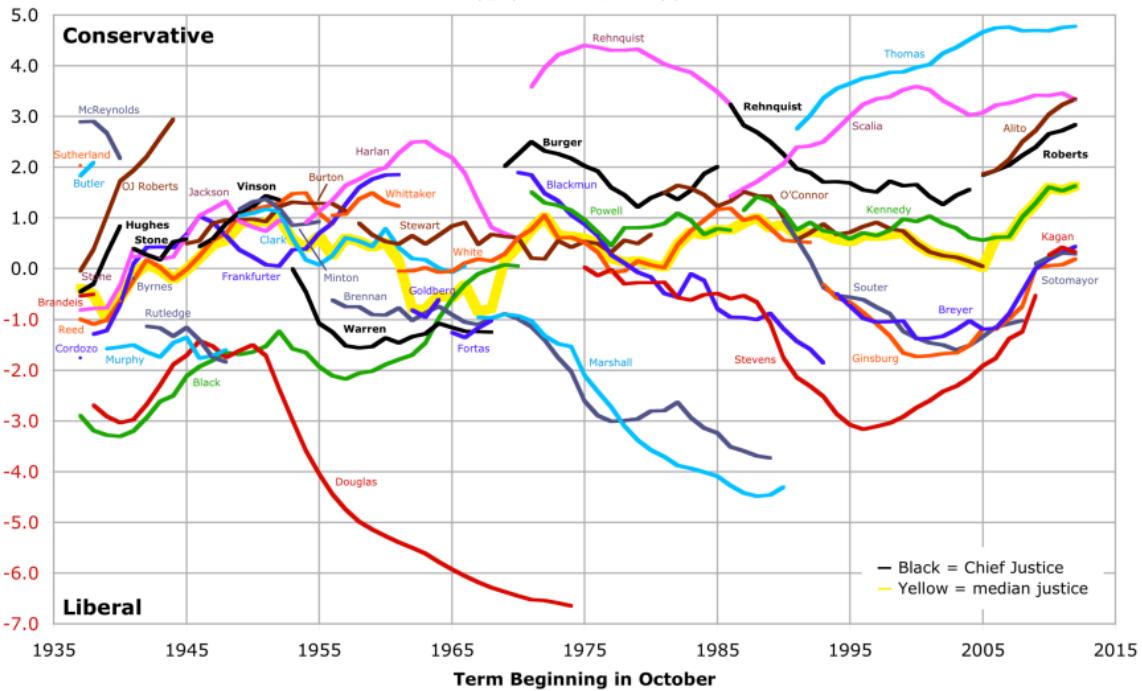
[Ruiz+ 2020]



Patterns of preferences found at Etsy.com (Hu et al., 2014)

Ideological Leanings of Supreme Court Justices

Source Data: Andrew D. Martin and Kevin M. Quinn
<http://mqscores.wustl.edu/measures.php>



Supreme Court Ideology over time (Martin and Quinn, 2001)

The image shows a mobile-style conversation interface. At the top, the large word "TIME" is visible, with "FEB. 27 / MARCH 6, 2023" printed above it. The conversation starts with a message from "Hi ChatGPT" (represented by a red square icon with a white "T") asking, "Hello! How can I assist you today?". The AI responds (represented by a green square icon with a white brain-like symbol) with, "What do you think of a TIME cover story with the title, 'The AI Arms Race Is Changing Everything?'". The AI continues, "As an AI language model, I don't have personal opinions or beliefs, but I can provide some context and information on the subject." Below this, a block of text discusses the potential impact of such a cover story on public awareness and responsible AI development. The text concludes with a note about the title's interpretation. On the right side of the interface, there is a vertical column with the author's name, "by ANDREW R. CHOW & BILLY PERRIGO", and their designation, "(humans)".

Hi ChatGPT

Hello! How can I assist you today?

What do you think of a TIME cover story with the title, "The AI Arms Race Is Changing Everything?"

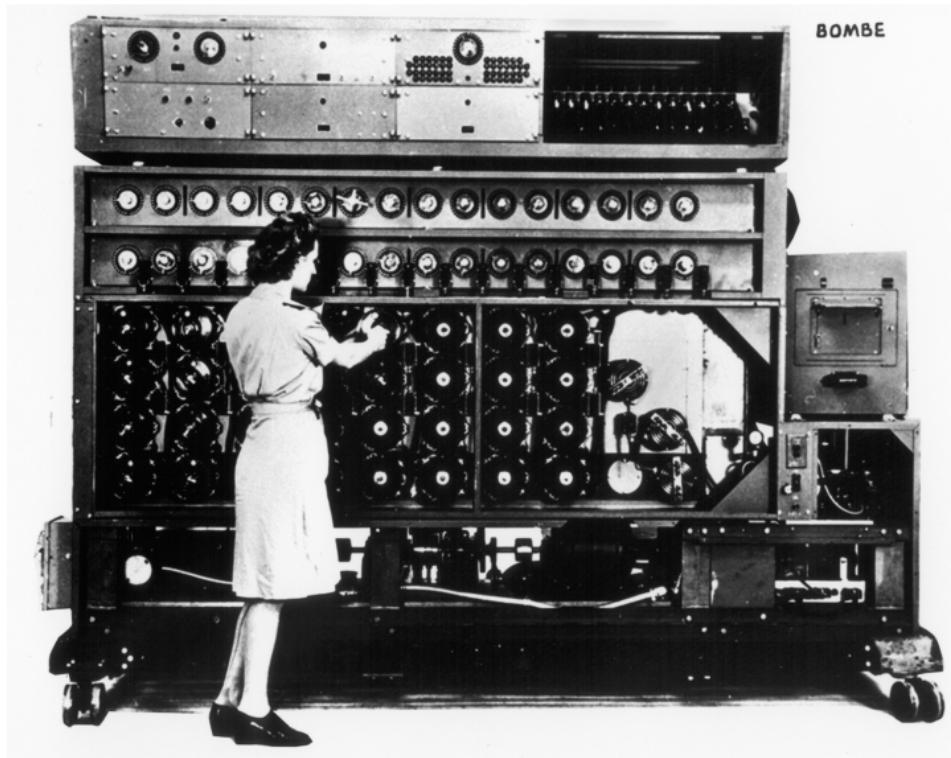
As an AI language model, I don't have personal opinions or beliefs, but I can provide some context and information on the subject.

"The AI Arms Race is Changing Everything" could be an attention-grabbing and thought-provoking title for a cover story on AI. A cover story in a widely-read publication like TIME could help raise public awareness about the potential risks and benefits of this trend, and stimulate discussions on how to ensure that AI is developed and deployed in a responsible and ethical manner.

It's important to note that the title of the story could also be interpreted as sensationalist or alarmist, and may not accurately reflect the complexities ■

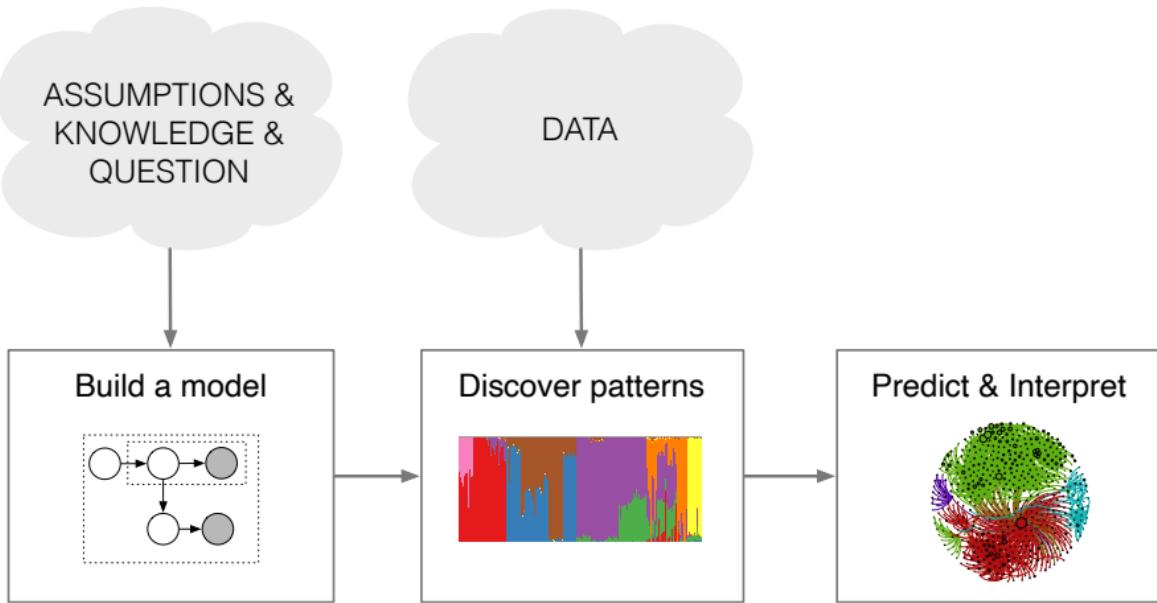
by
ANDREW R. CHOW & BILLY PERRIGO
(humans)

Large language models



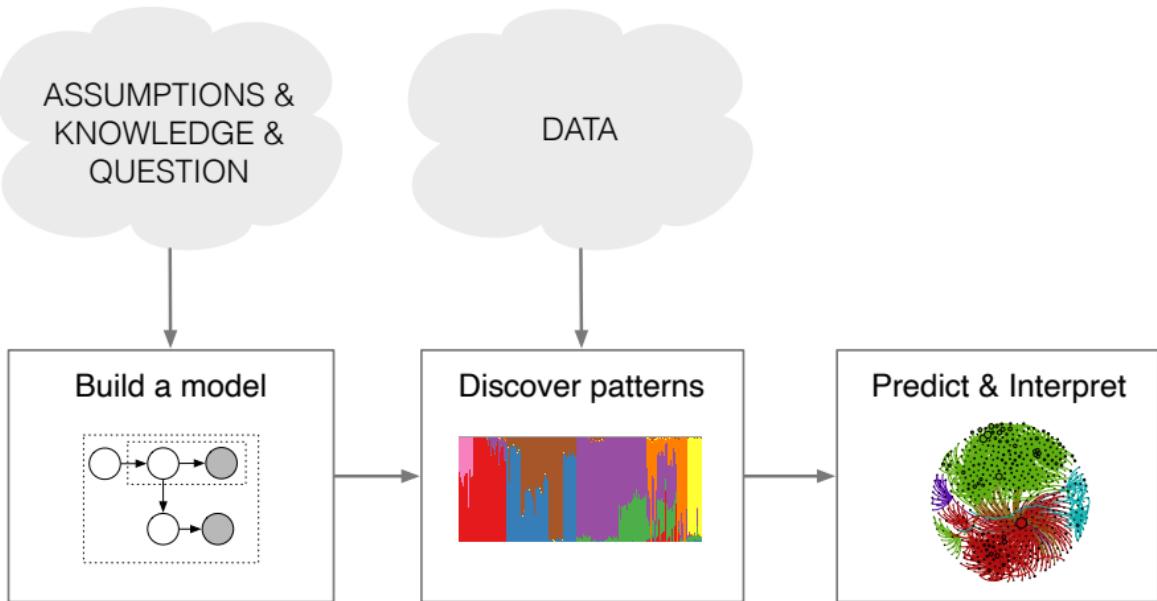
Breaking the Nazi code (Turing and Good, 194?)

Box's Loop



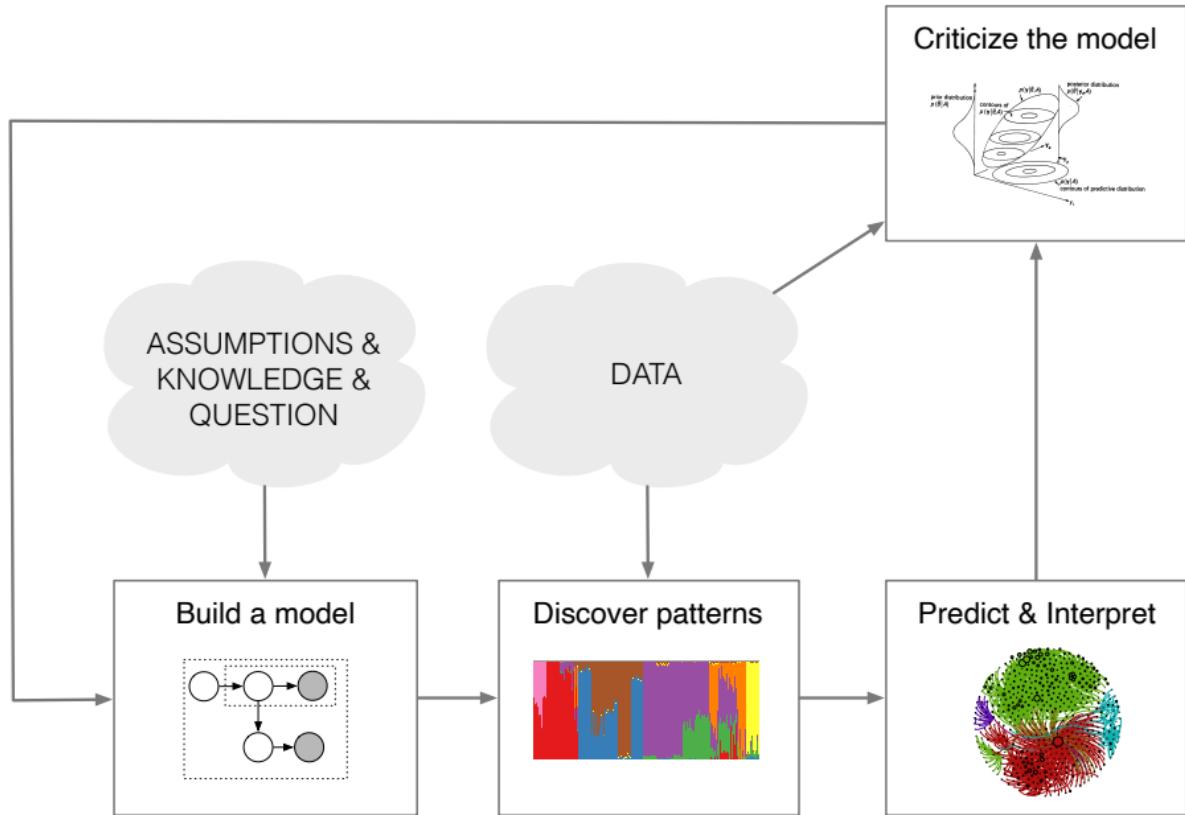
- ▶ Customized data analysis is important to many fields.
- ▶ This pipeline separates **assumptions, computation, application**.
- ▶ It facilitates solving data science problems.

Box's Loop



- ▶ Expressive components from which to build models
- ▶ Scalable and generic inference algorithms
- ▶ Stretch probabilistic modeling into new areas

Box's Loop



[Box 1980; Rubin 1984; Gelman+ 1996; Blei 2014]

Why study probabilistic ML?

- ▶ It is a unifying perspective around many ideas in ML
 - prediction, clustering, factorization, data fusion, ...
 - scalable ML, deep learning, representation learning, ...
- ▶ Some ideas require probabilistic ideas
 - Bayesian nonparametrics, hierarchical modeling, graphical models, causality, ...
- ▶ It provides a **compositional language** for expressing assumptions, capturing uncertainty, and deriving algorithms that find and use interpretable patterns in data.
- ▶ **Our premise: It is worth studying.**

Ask yourself

- ▶ What question do you want to answer?
- ▶ What data can help answer the question?
- ▶ What can you assume? What do you know?

Latent Dirichlet Allocation and Topic Modeling

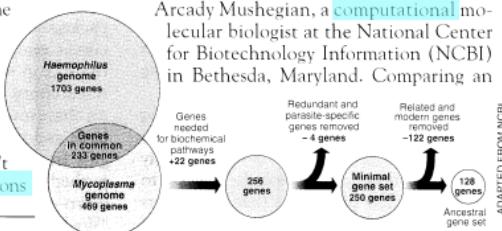
(An example of probabilistic machine learning)

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.

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

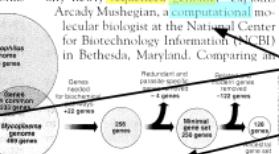
Seeking Life's Bare (Genetic) Necessities

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 350 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

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

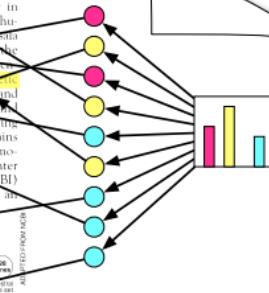
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson, a Spanish University in Stockholm who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game, particularly as more and more genomes are completely 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 all



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

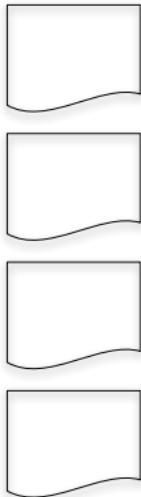
SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



Latent Dirichlet Allocation

Topics



Documents

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

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

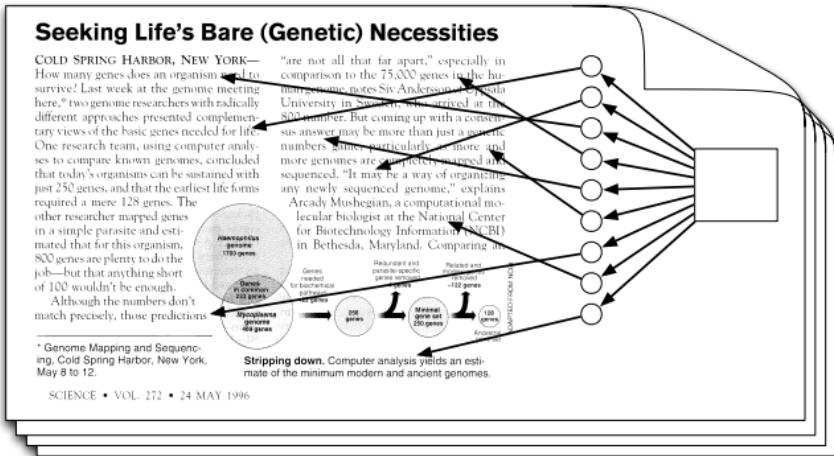
SCIENCE • VOL. 272 • 24 MAY 1996

"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 numeric 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 all

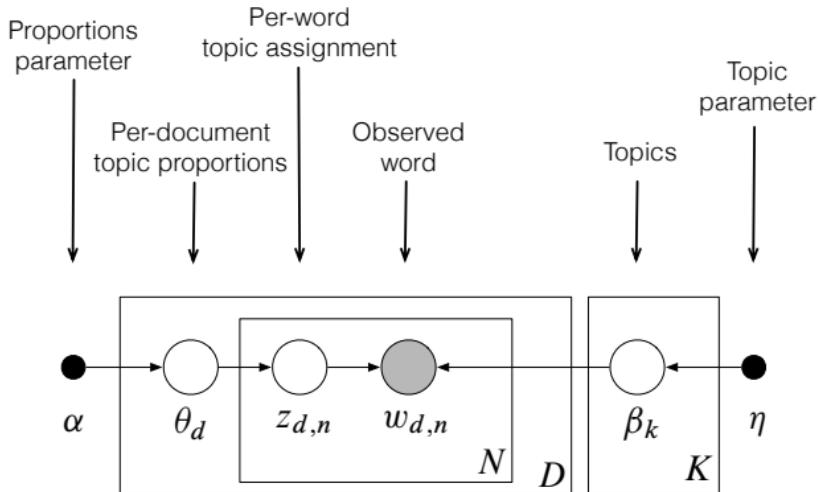


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Topic proportions and assignments

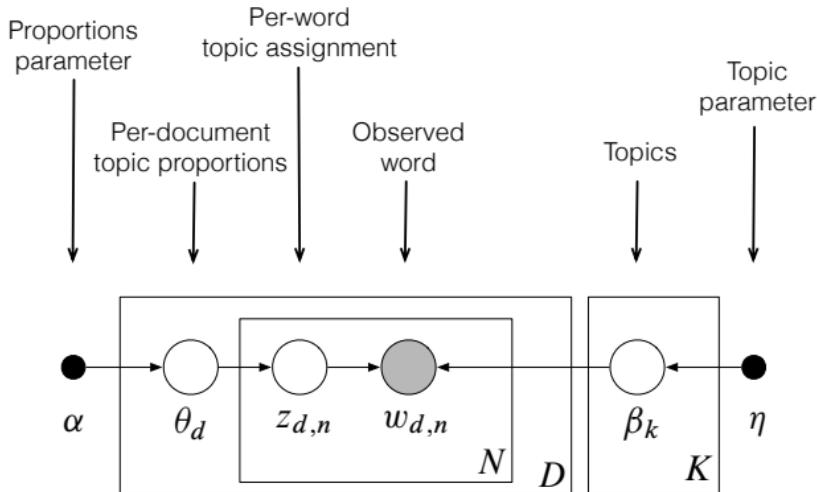


Latent Dirichlet Allocation



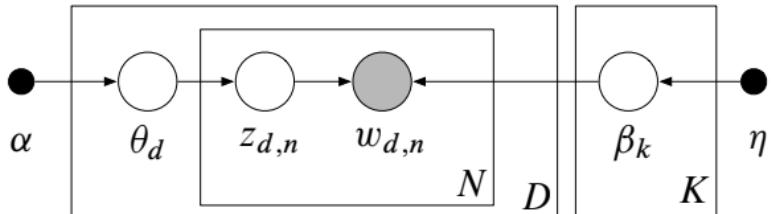
LDA as a graphical model

- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.



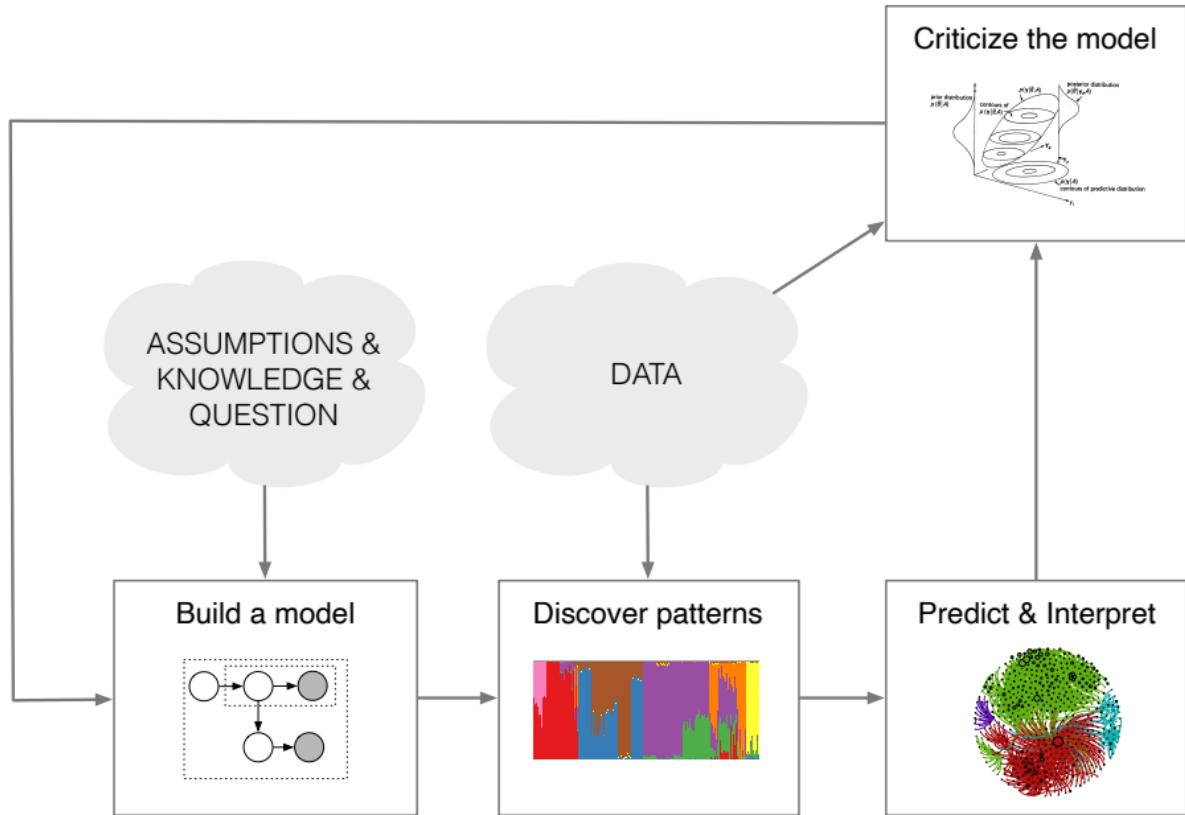
LDA as a graphical model

- ▶ Encodes independence assumptions
- ▶ Defines a factorization of the joint distribution, $p(\beta, \mathbf{z}, \boldsymbol{\theta}, \mathbf{w})$
- ▶ Connects to algorithms for computing with data



- ▶ The joint defines a posterior, $p(\beta, \theta, z | w)$.
- ▶ From a collection of documents, infer
 - Per-word topic assignment $z_{d,n}$
 - Per-document topic proportions θ_d
 - Per-corpus topic distributions β_k
- ▶ Then use posterior expectations to perform the task at hand:
information retrieval, document similarity, exploration, and others.

Box's Loop



[Box 1980; Rubin 1984; Gelman+ 1996; Blei 2014]



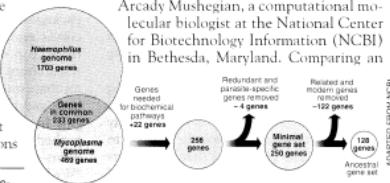
- ▶ **Data:** The OCR'ed collection of **Science** from 1990–2000
 - 17K documents
 - 11M words
 - 20K unique terms (stop words and rare words removed)
- ▶ **Model:** 100-topic LDA model using variational inference.

Seeking Life's Bare (Genetic) Necessities

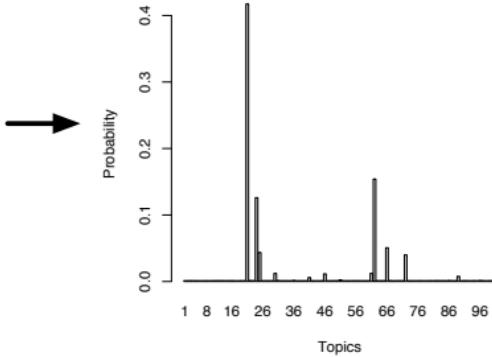
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,^{1,2} two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare 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.

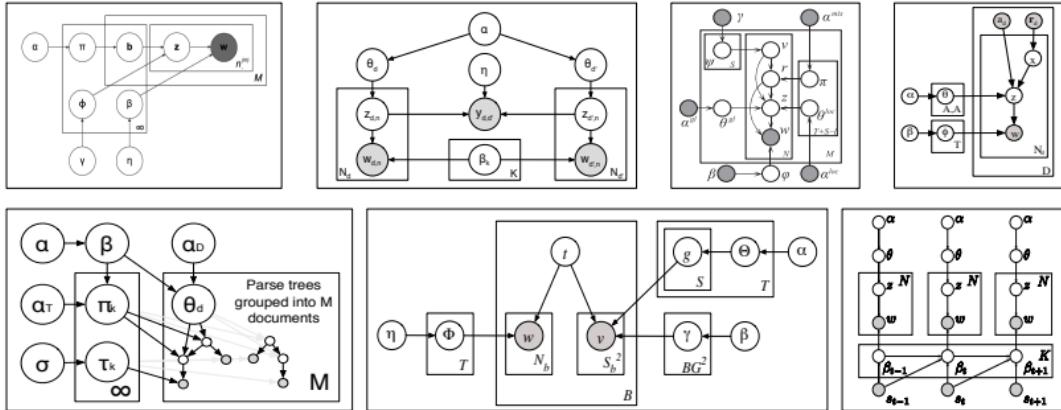


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

1	2	3	4	5
Game Season Team Coach Play Points Games Giants Second Players	Life Know School Street Man Family Says House Children Night	Film Movie Show Life Films Director Man Story Says	Book Life Books Novel Story Man Author House War Children	Wine Street Hotel House Room Night Place Restaurant Park Garden
6	7	8	9	10
Bush Campaign Clinton Republican House Party Democratic Political Democrats Senator	Building Street Square Housing House Buildings Development Space Percent Real	Won Team Second Race Round Cup Open Game Play Win	Yankees Game Mets Season Run League Baseball Team Games Hit	Government War Military Officials Iraq Forces Iraqi Army Troops Soldiers
11	12	13	14	15
Children School Women Family Parents Child Life Says Help Mother	Stock Percent Companies Fund Market Bank Investors Funds Financial Business	Church War Women Life Black Political Catholic Government Jewish Pope	Art Museum Show Gallery Works Artists Street Artist Paintings Exhibition	Police Yesterday Man Officer Officers Case Found Charged Street Shot

Topics found in 1.8M articles from the New York Times



- ▶ Probabilistic models are a **composable language** for methods that analyze data.
- ▶ LDA is a **simple building block** that enables many applications.
- ▶ Each model makes **assumptions**, and shows how to analyze data with them.

Ingredients of a probabilistic model

- ▶ **Random variables**

What we observe and what we want to discover

- ▶ **(Imaginary) Generative process**

How the RVs work together to make data

- ▶ **Posterior**

Conditional distribution of the hidden variables given the data

How the model “sees” the data

- ▶ **Predictive**

Conditional distribution of future data

What will we cover in this class?

The basic ideas

- ▶ The ingredients of probabilistic models
- ▶ Basic distributions, their conjugate priors
- ▶ Conditional distributions: Logistic and linear regression, and scalable inference

Models and algorithms

- ▶ Bayesian mixture models and the Gibbs sampler
- ▶ Mixed-membership models, topic models, and variational inference
- ▶ Matrix factorization, recommendation systems, and efficient MAP

Advanced ideas

- ▶ Exponential families, conjugacy, and generalized linear models
- ▶ Deep learning and probabilistic modeling
- ▶ Black box variational inference
- ▶ Model criticism and evaluation

If we still have time...

- ▶ Some theory of graphical models (d -separation, tree propagation)
- ▶ Diffusion models
- ▶ Large language models (and discrete sequence models)
- ▶ Bayesian nonparametrics
- ▶ Introduction to causality

Themes and thorny questions

Some recurring ideas in this course:

- ▶ Structural motifs in probabilistic models and how to compose them
- ▶ How to derive algorithms to approximating a posterior
- ▶ How to use the predictive distribution to evaluate and select models

Some important thorny questions to keep in mind:

- ▶ How should I select a model?
- ▶ Which inference strategy should I use? What are the trade-offs?
- ▶ Is my model good? What does it tell me about my data?

What this course is

- ▶ This course is about **doctoral research in probabilistic machine learning.**
 - methods, applications, theory
- ▶ (It is not about learning a cookbook of ML algorithms.)
- ▶ Side effects:
 - Learn how to navigate and read the PML literature
 - Learn (from experience) tools for doing PML
 - Learn (from experience) a pipeline for PML research
 - See PML as a unifying perspective on many aspects of ML

Logistics

- ▶ Tue/Thu 8:40-9:55
- ▶ DB office hour: Tuesday 10a-11a
- ▶ TAs: Casey Bradshaw, Justin Hong, Sebastian Salazar
- ▶ <http://www.cs.columbia.edu/~blei/fogm/2023F/>
- ▶ Communicate via slack (no email, please)

Prerequisites

- ▶ You are comfortable with basic probability
 - random variables, distributions
 - joint distributions, marginalization, conditionals
 - (conditional) expectations, mean, variance
- ▶ You are comfortable with optimization
 - setting an objective function; taking a gradient; following it
- ▶ You know how to program for data analysis (e.g., R, Python, ...)

Who can take the course?

- ▶ Open to **all** PhD students in **all** departments and schools
- ▶ After PhD students are enrolled, open to other students (subject to space)
- ▶ No pass/fail
- ▶ No auditors

But if you are bureaucratically unable to take the course then you are welcome to attend (subject to space). Please do all the work, including the final project.

- ▶ **Everyone:** Complete Homework 0 by EOD Wednesday

Coursework and grade

- ▶ Weekly: Choose a relevant reading and write a response
 - Hand it in **in the beginning of class only** (on Thursday)
 - We do not accept late reports (but it's OK to miss up to three)
 - If you read the book, please give me feedback!
- ▶ Three homeworks
 - 5 late days ; use them for legitimate and not-legitimate reasons for lateness.
- ▶ Final project (and milestones along the way)
 - Dive deeply into a problem in probabilistic ML
 - Many projects will be real applications of Box's loop
 - Some projects are more theoretical
- ▶ Your grade
 - reader reports (10%)
 - homeworks (30%)
 - final project (60%)

Other logistics

- ▶ Class is not recorded. But I will share lecture notes on Slack.
- ▶ If you have COVID symptoms please don't come to class.
(If you want to attend class, we try to arrange a private Zoom link.)
- ▶ Occasionally I have to miss class due to conference travel.
I will always hold (optional) make-up lectures.

A note on Chat-GPT

- ▶ Chat-GPT is a useful tool for programming and writing.
You can treat it as a tireless, smart roommate from whom you can ask for help.¹
- ▶ ***Please use good judgement to decide what is acceptable.***
- ▶ Examples of **acceptable** uses of Chat-GPT:
 - Ask to make suggestions on your writing
 - Ask to explain the arguments to a function in R or Python
 - Ask to fill in pseudocode with Python syntax (*)
 - Ask to write code to make a nice plot or preprocess raw data (*)
- ▶ Examples of **unacceptable** uses of Chat-GPT:
 - Ask to produce pseudocode and fill it in
 - Ask to complete a writing prompt from scratch (e.g., a reader report)
 - Ask to turn sketchy unreadable notes into fluent writing

¹To be clear, Chat-GPT is a high-order Markov model of language, trained to be able to chat. (It is not “intelligent.”) The roommate metaphor is simply helpful to define the boundaries of cheating.

Box's Loop

