Foundations of Graphical Models

David M. Blei

Columbia University

Today's lecture

- What is this course about?
- Latent Dirichlet allocation: An example of a graphical model
- Other examples of applied probabilistic modeling
- Box's loop
- What will we cover?
- Prerequisites, requirements, and grades

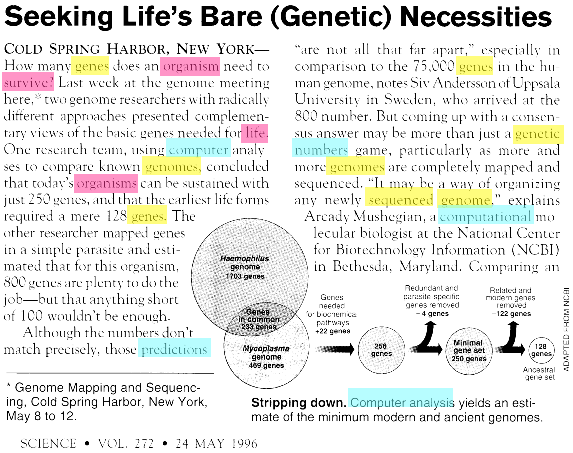
Announcements

- Go to the course website and fill out the survey.
- Sign up for Piazza

What is this course about?

Latent Dirichlet Allocation

(An example of a model that I know well)



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

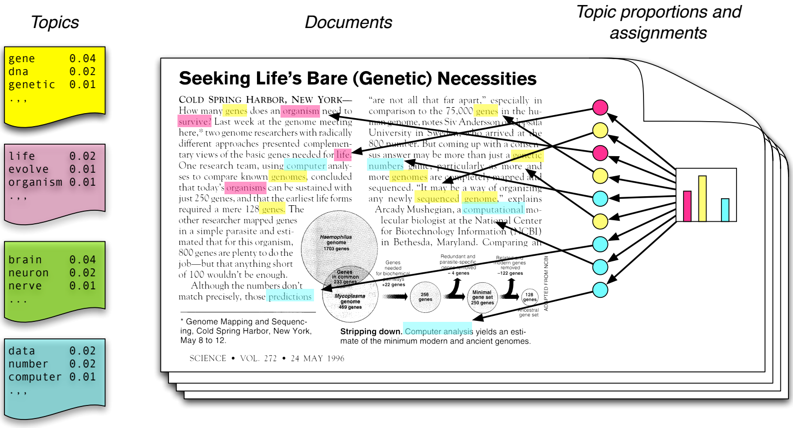
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

Documents exhibit multiple topics.



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

Topic proportions and assignments

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

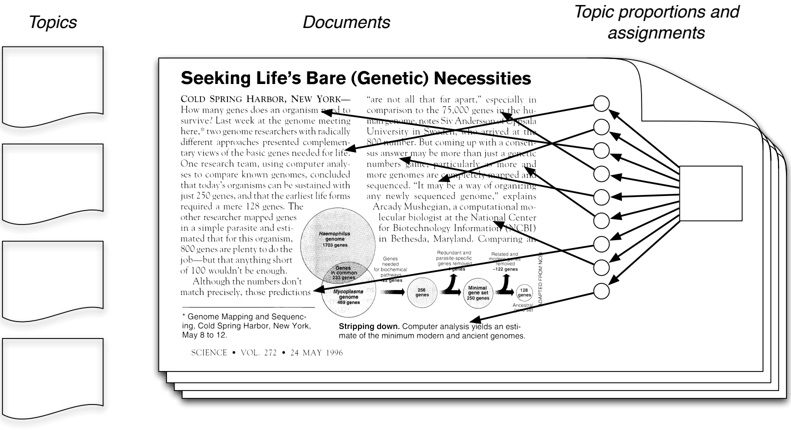
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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 Dirichlet Allocation



Topics

Documents

Topic proportions and assignments

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

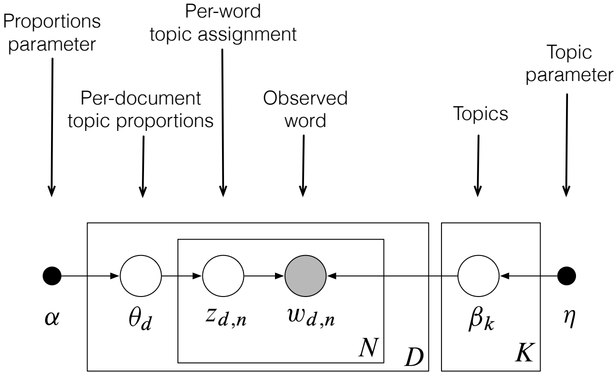
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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.

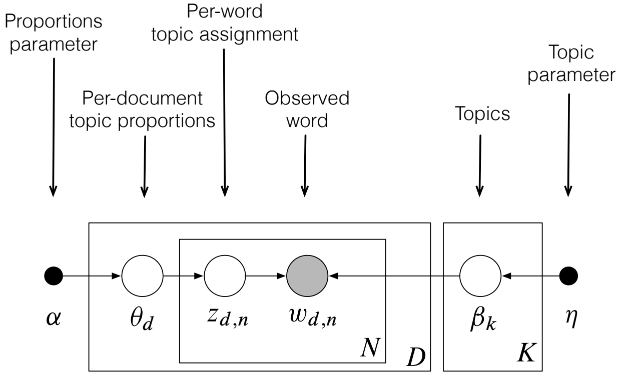
SCIENCE • VOL. 272 • 24 MAY 1996

Latent Dirichlet Allocation



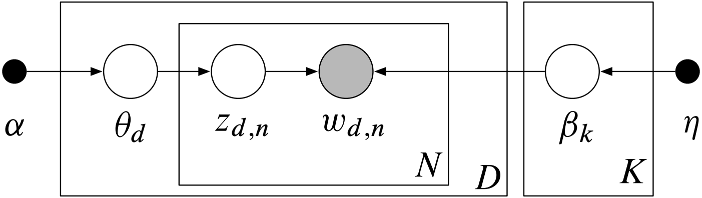
LDA as a graphical model

- 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
- Connects to algorithms for computing with data

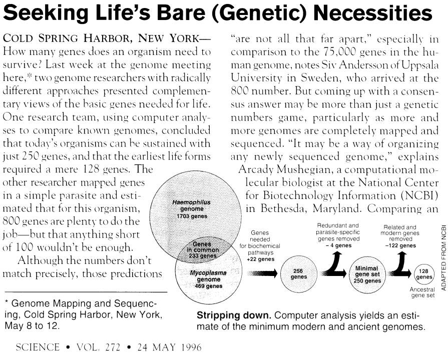


- The joint defines a posterior, $p(\theta, z, \beta \mid w)$.

- 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.

- **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,* 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.

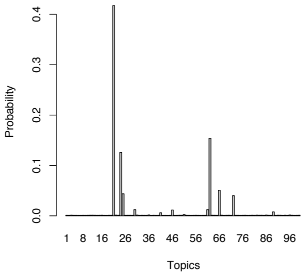
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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



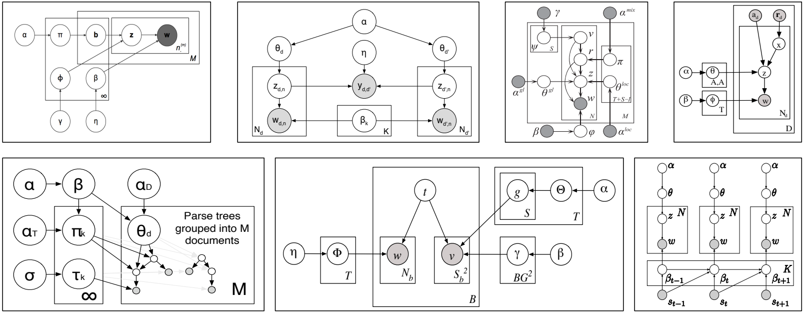
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	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	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

Topics found in 1.8M articles from the New York Times



- LDA is a simple building block that enables many applications.
  Topic modeling is an active field of research.

- Graphical models are a composable language for probability models.

- Each model connects to a set of assumptions
  and an algorithm for computing under them.

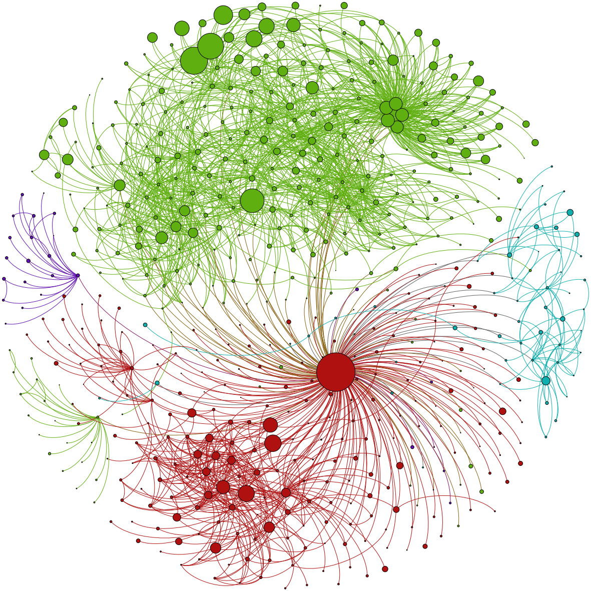
Edward: A library for probabilistic modeling, inference, and criticism

`github.com/blei-lab/edward`

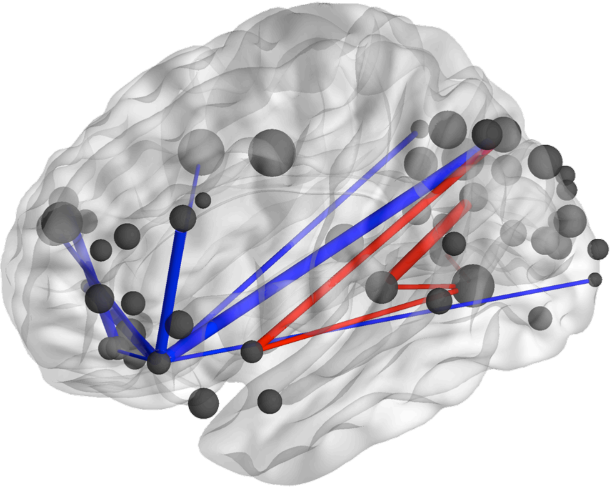
(lead by Dustin Tran)

Other examples of applied probabilistic modeling

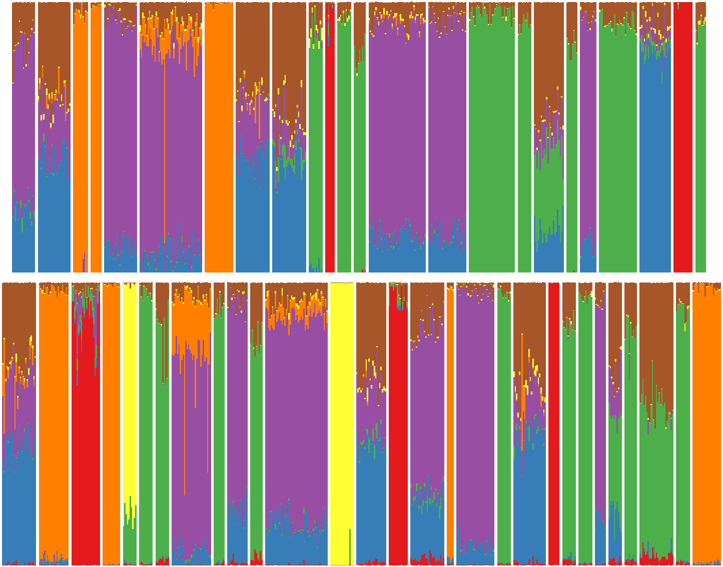
(from my research group and others)



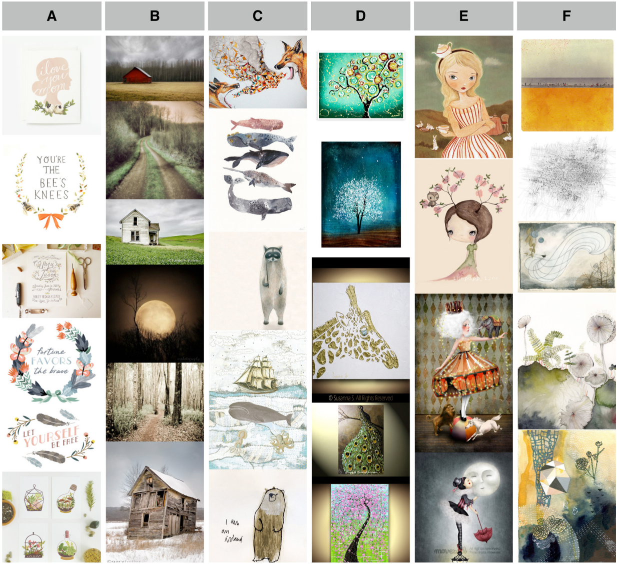
Communities discovered in a 3.7M node network of U.S. Patents



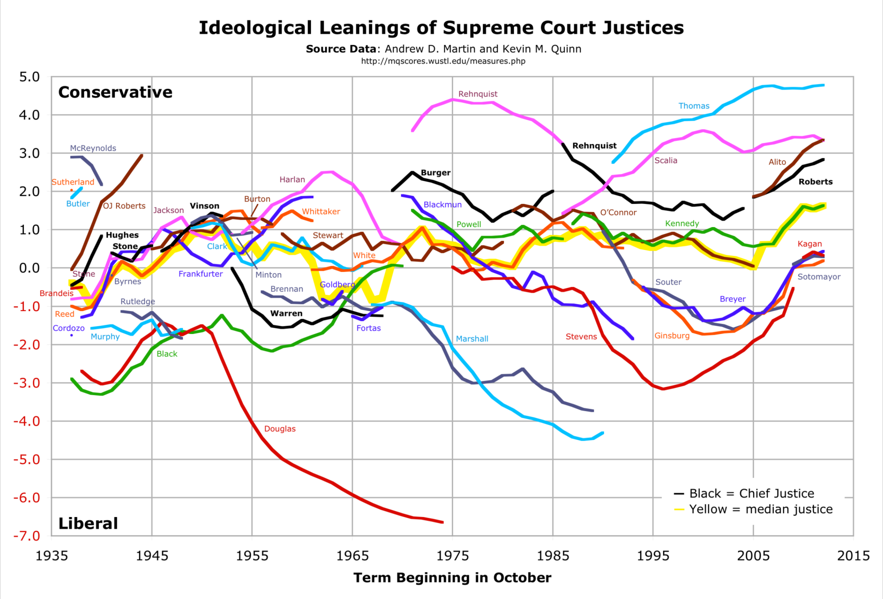
Neuroscience analysis of 220 million fMRI measurements



Population analysis of 2 billion genetic measurements



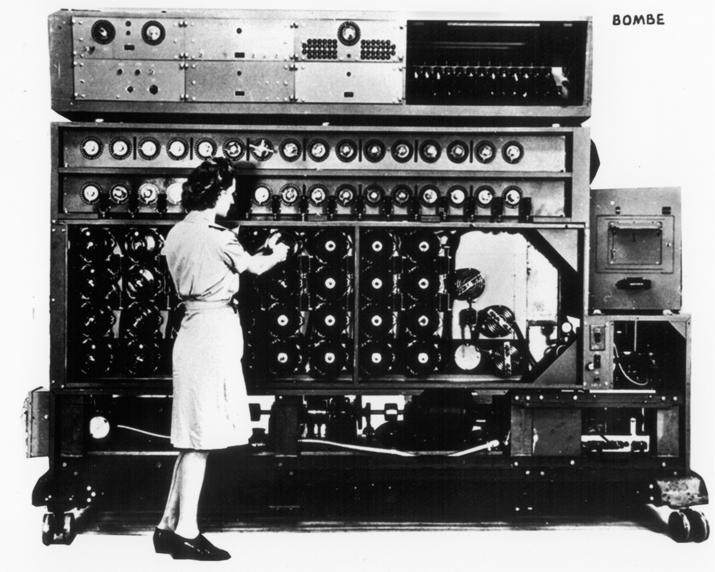
Patterns of preferences found at Etsy.com (Hu et al., 2014)



Ideological Leanings of Supreme Court Justices

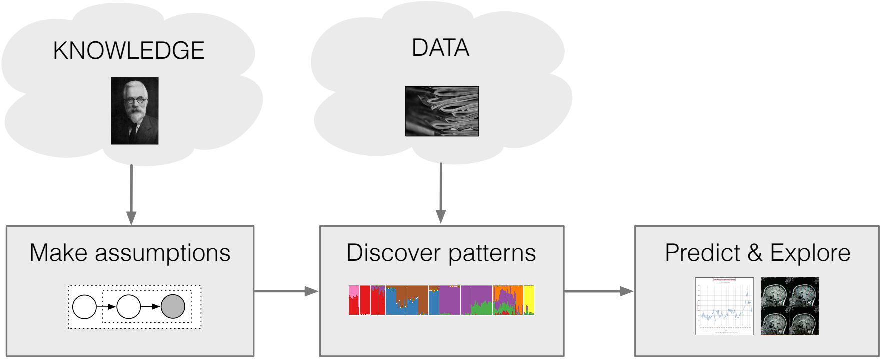
Source Data: Andrew D. Martin and Kevin M. Quinn
http://mqscores.wustl.edu/measures.php

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



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

Box's Loop



KNOWLEDGE

DATA

Make assumptions

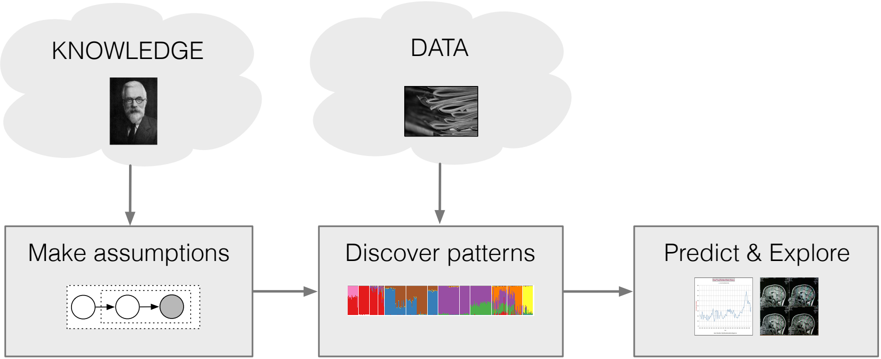
Discover patterns

Predict & Explore

Why we like this picture:

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

Box's Loop



KNOWLEDGE

DATA

Make assumptions

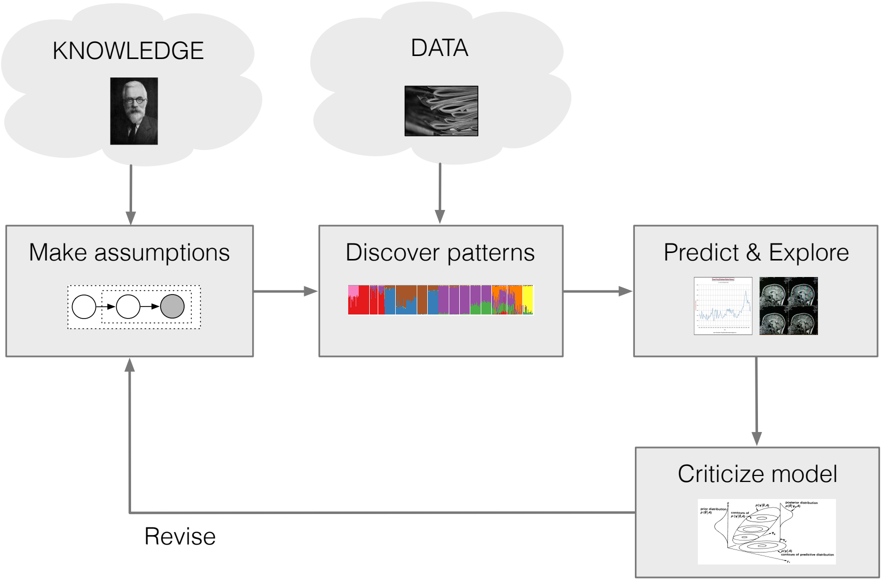
Discover patterns

Predict & Explore

What we need:

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

Box’s Loop



KNOWLEDGE

DATA

Make assumptions

Discover patterns

Predict & Explore

Criticize model

Revise

prior distribution $p(\theta|A)$
θ
contours of $p(\mathbf{y}|\theta,A)$
$p(\mathbf{y}|\theta,A)$
posterior distribution $p(\theta|\mathbf{y}_d,A)$
$\mathbf{y}_2$
$\mathbf{y}_d$
$p(\mathbf{y}|A)$
contours of predictive distribution
y_1

What will we cover?

The basics of graphical models

1. Probability: Basic concepts and review
2. Semantics of graphical models
3. D-separation and conditional independence
4. The elimination algorithm
5. Tree propagation and hidden Markov models

Latent variable models

1. Models, data, and statistical concepts
2. Bayesian mixtures of Gaussians and the Gibbs sampler
3. Exponential families, conjugacy, and mixtures of exponential families
4. Mixed-membership, topic models, and variational inference
5. Matrix factorization and recommendation systems

Conditional models

1. Regression: Linear and logistic
2. Generalized linear models
3. Regularized linear models
4. Hierarchical models, robust models, and empirical Bayes

Advanced ideas

1. Advanced Markov chain Monte Carlo
2. Advanced variational inference
3. An brief introduction to Bayesian nonparametrics

Some additional discussion

- Programming languages
- Applications
- Note: We will usually be at the board.

Prerequisites, Requirements, Grades, Etc.

- http://www.cs.columbia.edu/∼blei/fogm/
- Office hours: Tuesday 3:00-4:00PM, 912 SSW (but check the web!)
- Prerequisites
  - Probability and Statistics
  - Optimization
  - Programming
- Requirements
  - Weekly paper about the reading ($\leq$ 1 page)
  - Occasional homework
  - Final project
- Your grade: Mostly the final project