Foundations of Graphical Models

David M. Blei

Columbia University

Today's lecture

- Logistics, especially pertaining to enrollment
- 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

Enrollment

Logistics

- This room holds 54 people.
- Please fill out the Google form at http://www.cs.columbia.edu/∼blei/fogm/

Note

- This course will be offered every Fall.
- We will consider expanding it in the future.

What is this course about?

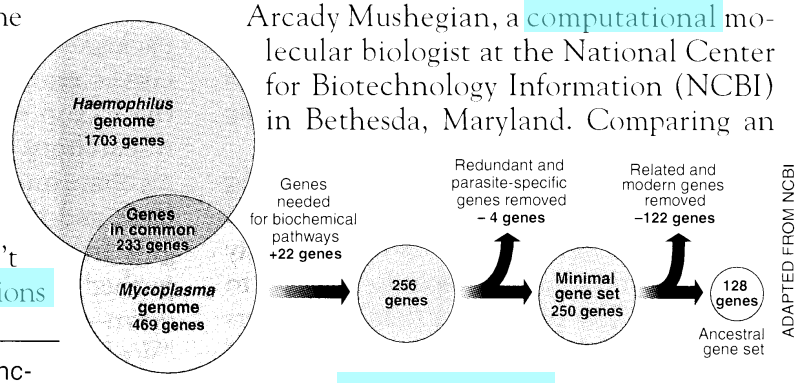
Latent Dirichlet Allocation

(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



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.

Documents exhibit multiple topics.

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

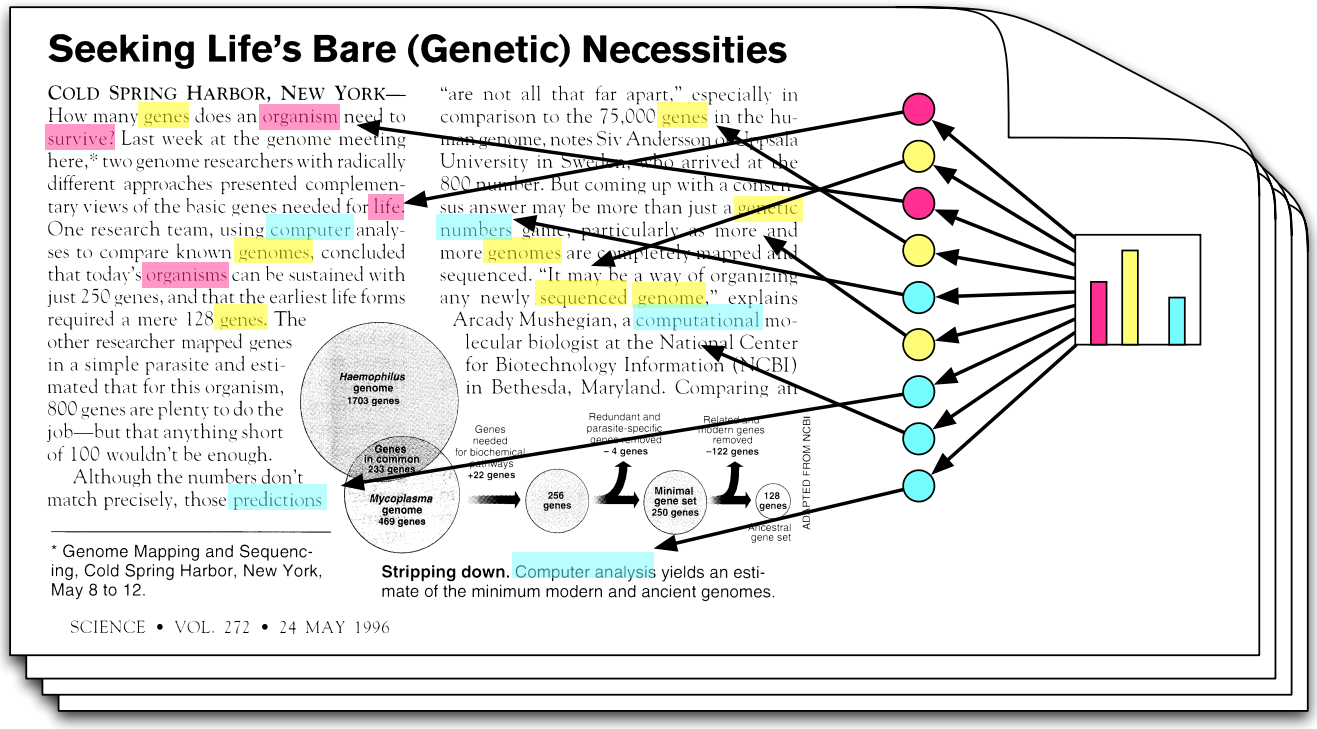
life	0.02
evolve	0.01
organism	0.01
...	

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

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

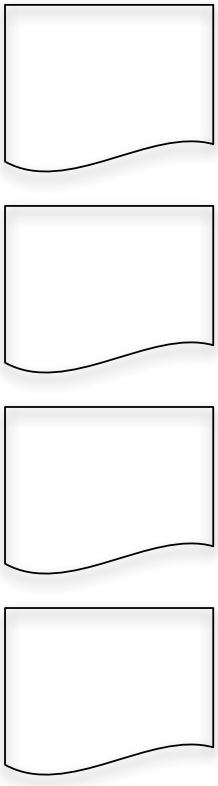
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

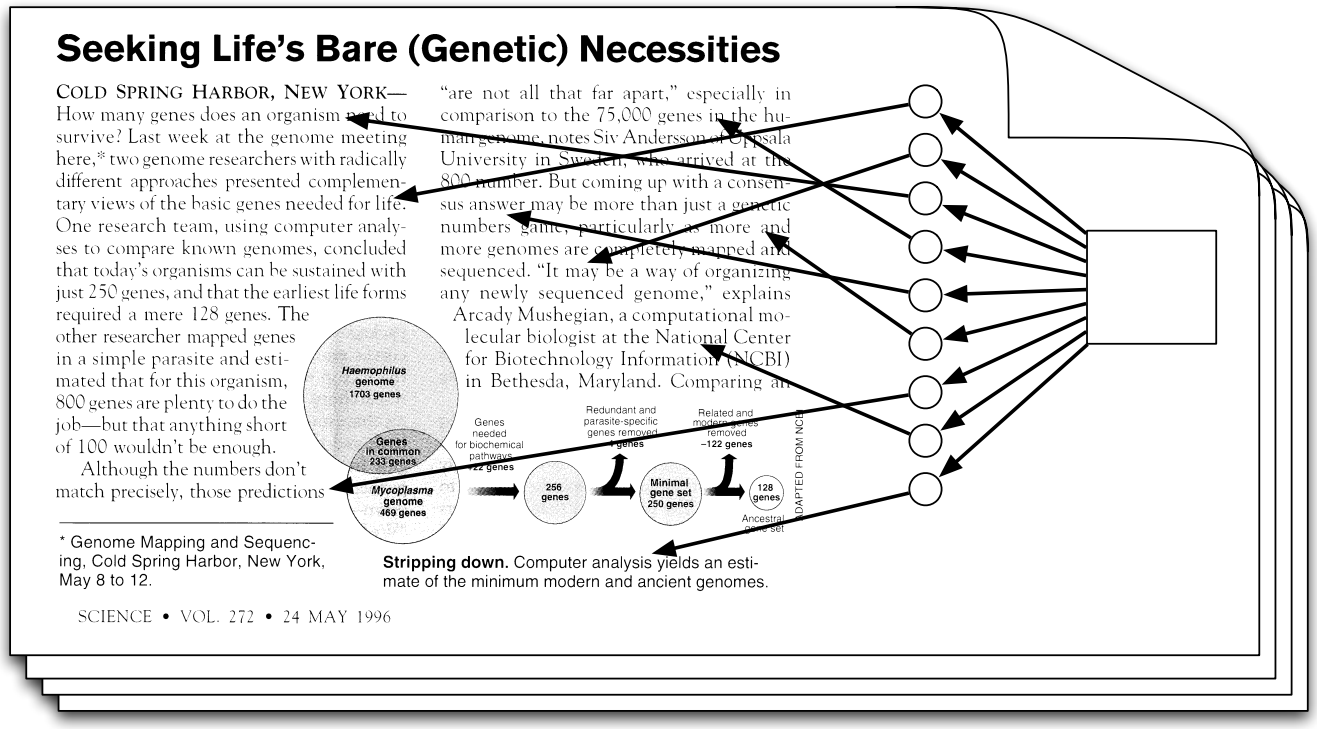
Latent Dirichlet Allocation

Topics

Documents

Topic proportions and assignments





Seeking Life's Bare (Genetic) Necessities

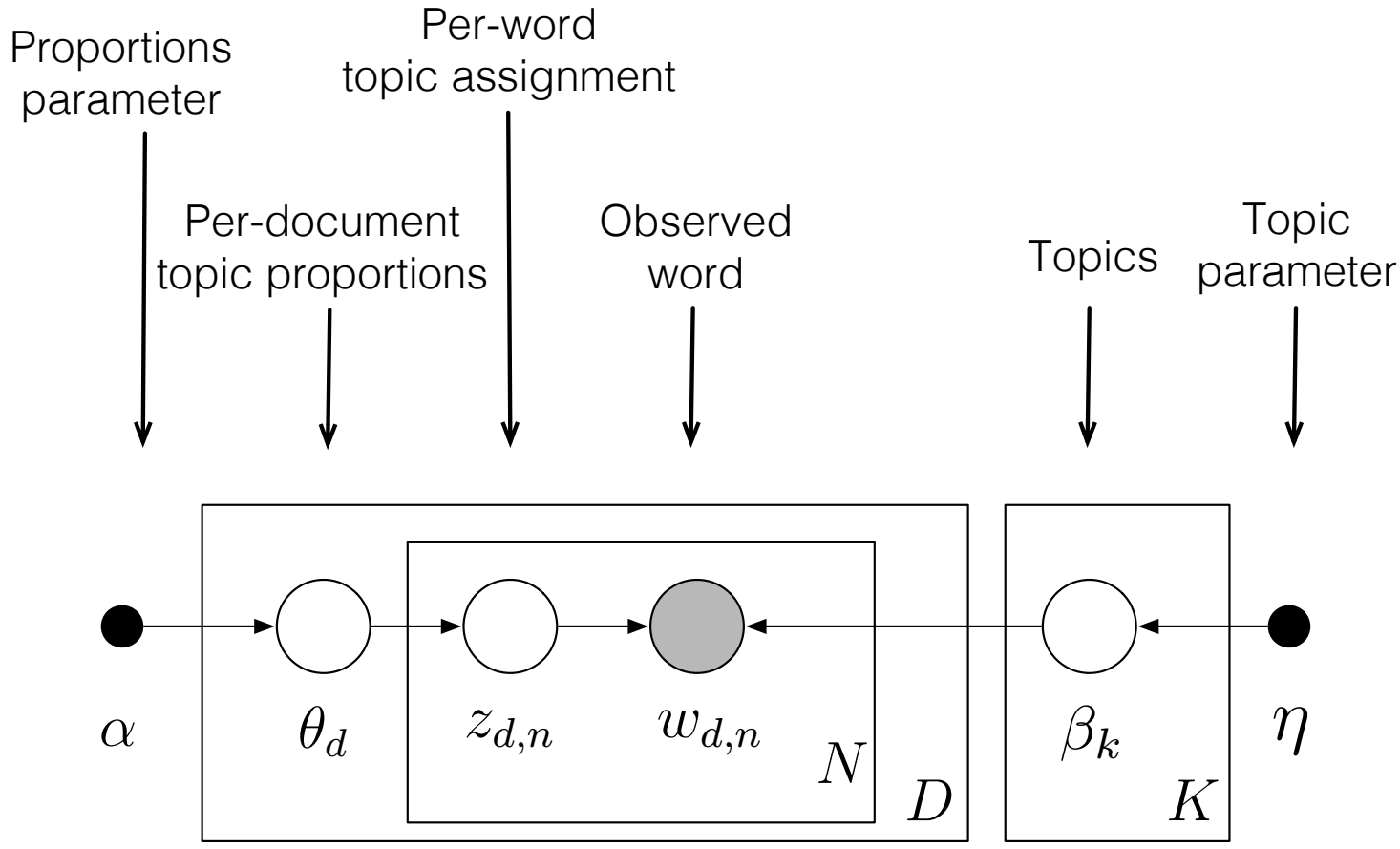
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

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

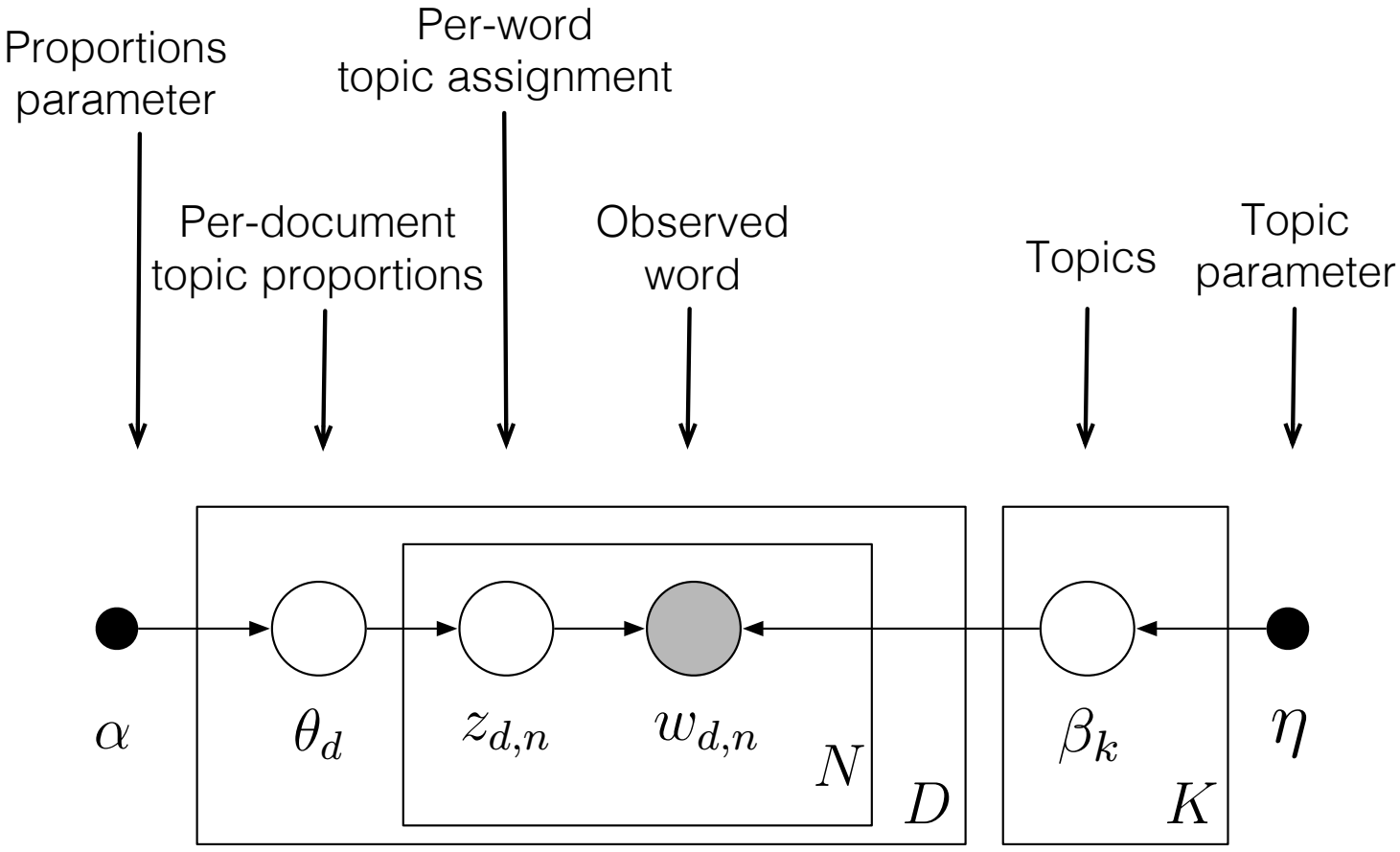
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Latent Dirichlet Allocation



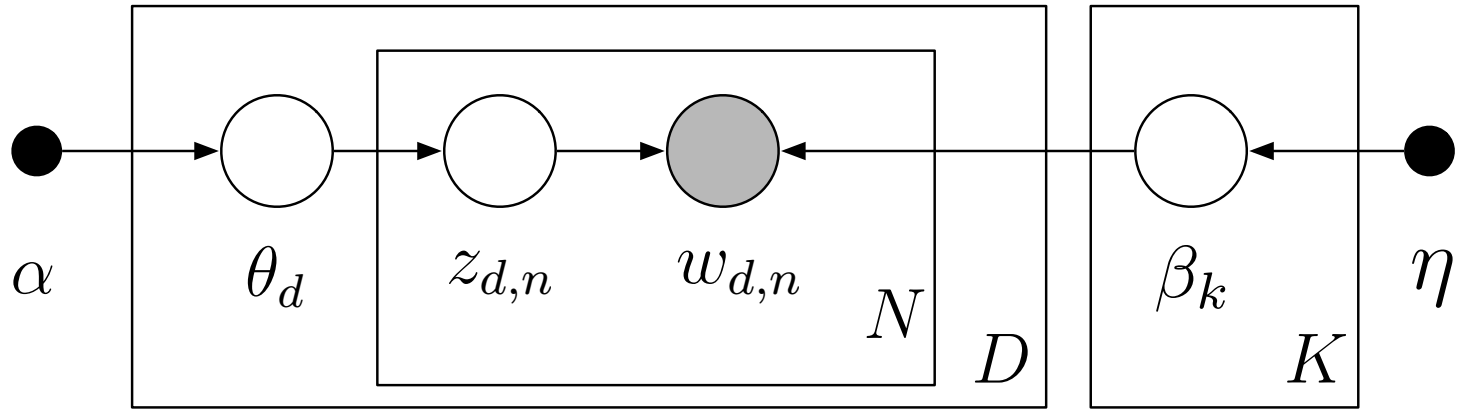
LDA as a graphical model

- 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



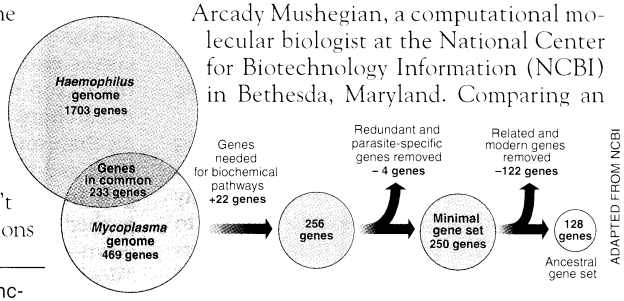
- The joint defines a posterior, $p(\theta, z, \beta \mid w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand: information retrieval,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

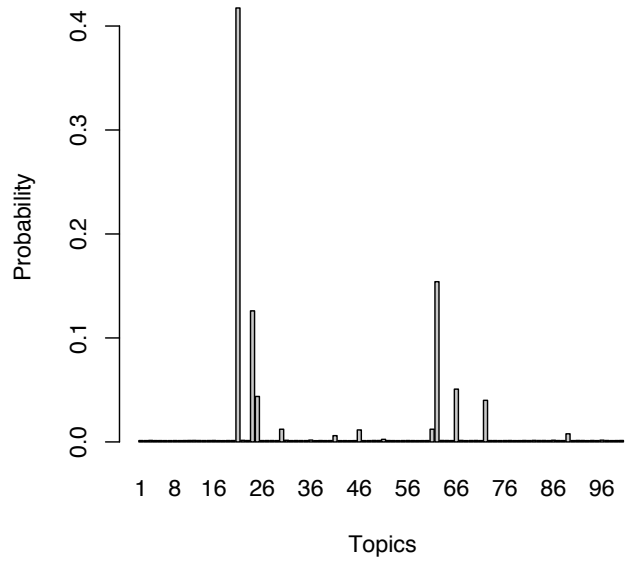
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

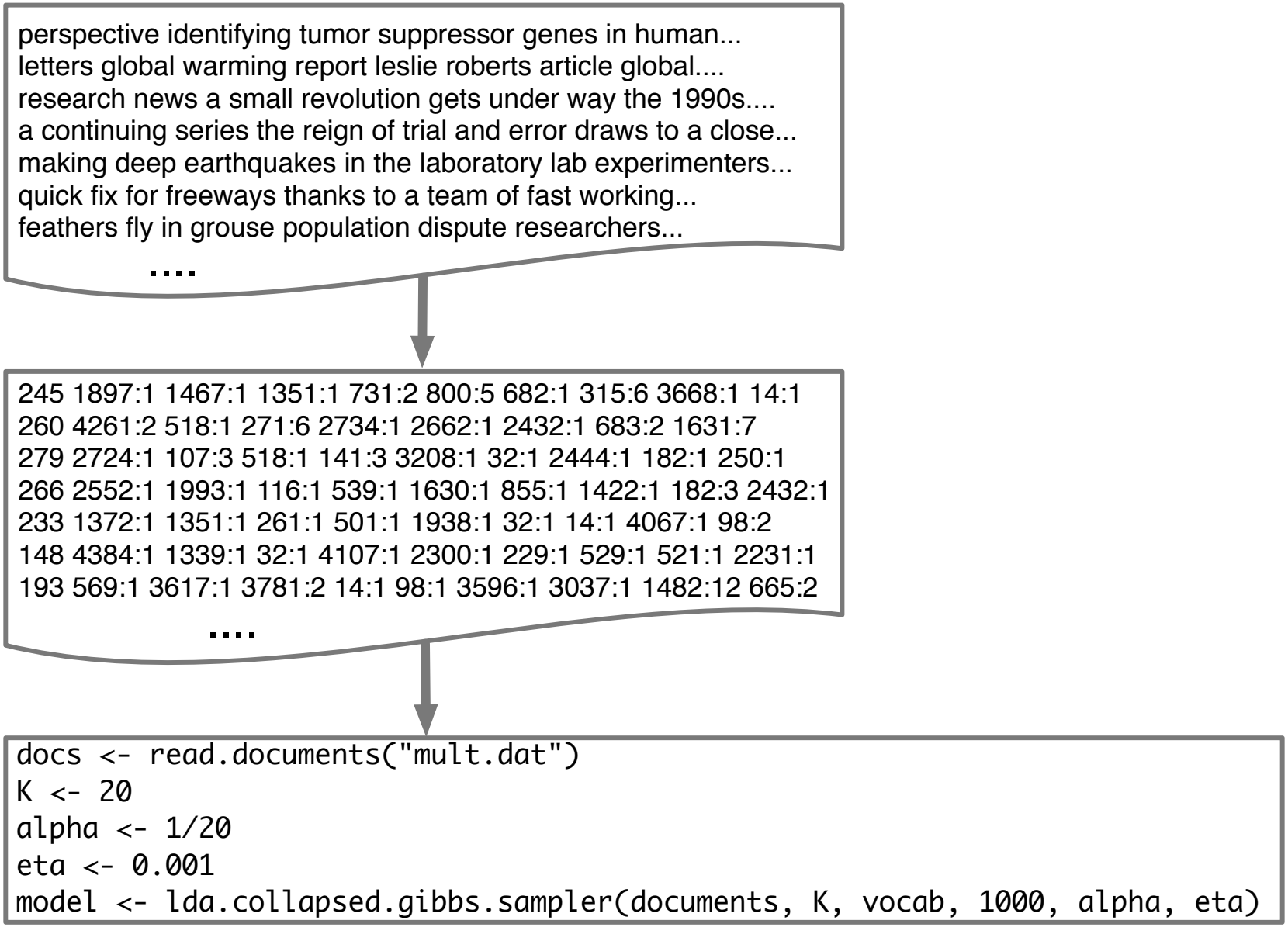


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* 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


perspective identifying tumor suppressor genes in human...
letters global warming report leslie roberts article global....
research news a small revolution gets under way the 1990s....
a continuing series the reign of trial and error draws to a close...
making deep earthquakes in the laboratory lab experimenters...
quick fix for freeways thanks to a team of fast working...
feathers fly in grouse population dispute researchers...
....
245 1897:1 1467:1 1351:1 731:2 800:5 682:1 315:6 3668:1 14:1
260 4261:2 518:1 271:6 2734:1 2662:1 2432:1 683:2 1631:7
279 2724:1 107:3 518:1 141:3 3208:1 32:1 2444:1 182:1 250:1
266 2552:1 1993:1 116:1 539:1 1630:1 855:1 1422:1 182:3 2432:1
233 1372:1 1351:1 261:1 501:1 1938:1 32:1 14:1 4067:1 98:2
148 4384:1 1339:1 32:1 4107:1 2300:1 229:1 529:1 521:1 2231:1
193 569:1 3617:1 3781:2 14:1 98:1 3596:1 3037:1 1482:12 665:2
....
docs <- read.documents("mult.dat")
K <- 20
alpha <- 1/20
eta <- 0.001
model <- lda.collapsed.gibbs.sampler(documents, K, vocab, 1000, alpha, eta)

LDA in R


1
dna
gene
sequence
genes
sequences
human
genome
genetic
analysis
two
2
protein
cell
cells
proteins
receptor
fig
binding
activity
activation
kinase
3
water
climate
atmospheric
temperature
global
surface
ocean
carbon
atmosphere
changes
4
says
researchers
new
university
just
science
like
work
first
years
5
mantle
high
earth
pressure
seismic
crust
temperature
earths
lower
earthquakes
6
end
article
start
science
readers
service
news
card
circle
letters
7
time
data
two
model
fig
system
number
different
results
rate
8
materials
surface
high
structure
temperature
molecules
chemical
molecular
fig
university
9
dna
rna
transcription
protein
site
binding
sequence
proteins
specific
sequences
10
disease
cancer
patients
human
gene
medical
studies
drug
normal
drugs
11
years
million
ago
age
university
north
early
fig
evidence
record
12
species
evolution
population
evolutionary
university
populations
natural
studies
genetic
biology
13
protein
structure
proteins
two
amino
binding
acid
residues
molecular
structural
14
cells
cell
virus
hiv
infection
immune
human
antigen
infected
viral
15
space
solar
observations
earth
stars
university
mass
sun
astronomers
telescope
16
fax
manager
science
aaas
advertising
sales
member
recruitment
associate
washington
17
cells
cell
gene
genes
expression
development
mutant
mice
fig
biology
18
energy
electron
state
light
quantum
physics
electrons
high
laser
magnetic
19
research
science
national
scientific
scientists
new
states
university
united
health
20
neurons
brain
cells
activity
fig
channels
university
cortex
neuronal
visual

Wikipedia Topics

Relative Presence of Topics in all Documents

{household, population, female}
{film, series, show}
{theory, work, human}
{son, year, death}
{war, force, army}
{system, computer, user}
{album, band, music}
{government, party, election}
{game, team, player}
{god, call, give}
{company, market, business}
{math, number, function}
{city, large, area}

{film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}

Stanley Kubrick

related topics

{film, series, show}
{theory, work, human}
{son, year, death}
{black, white, people}
{god, call, give}
{math, energy, light}

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.[1] A recurring theme in his films is man's inhumanity to man. While often viewed as

related documents

Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire Slayer (TV series)
The X-Files
Sunset Boulevard (film)
Jack Benny

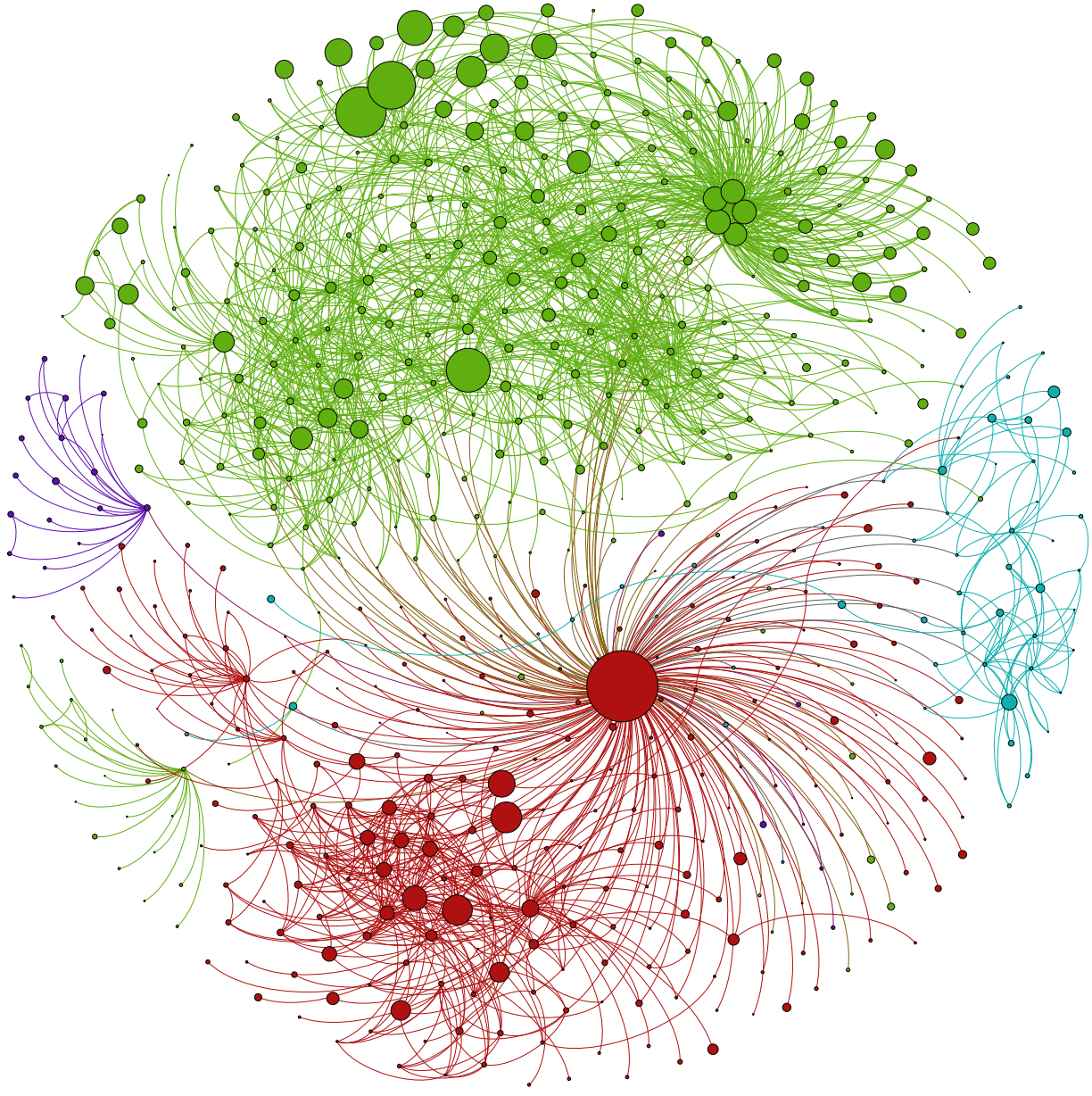
{theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

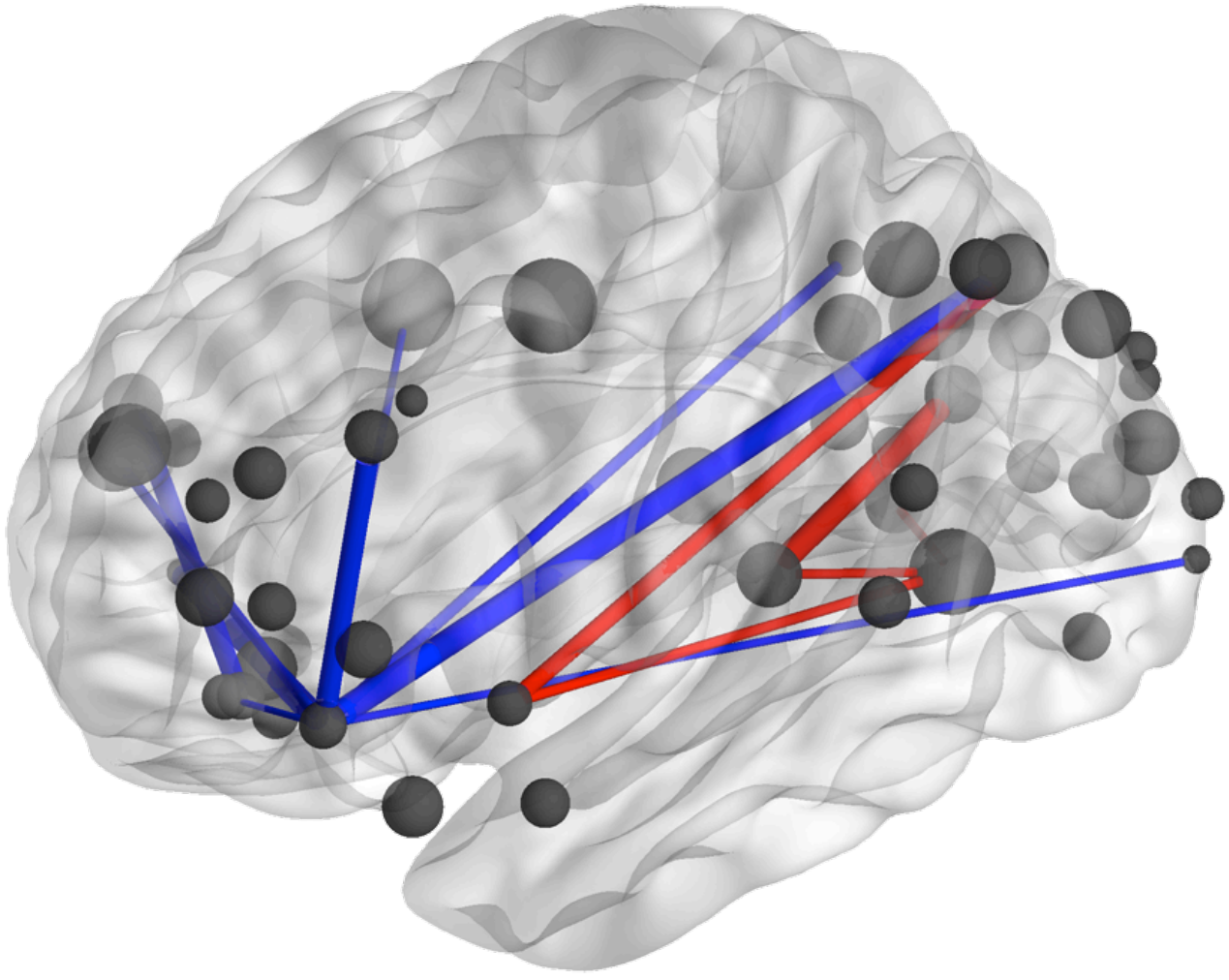
Topics found in 1.8M articles from the New York Times

Other examples of applied probabilistic modeling

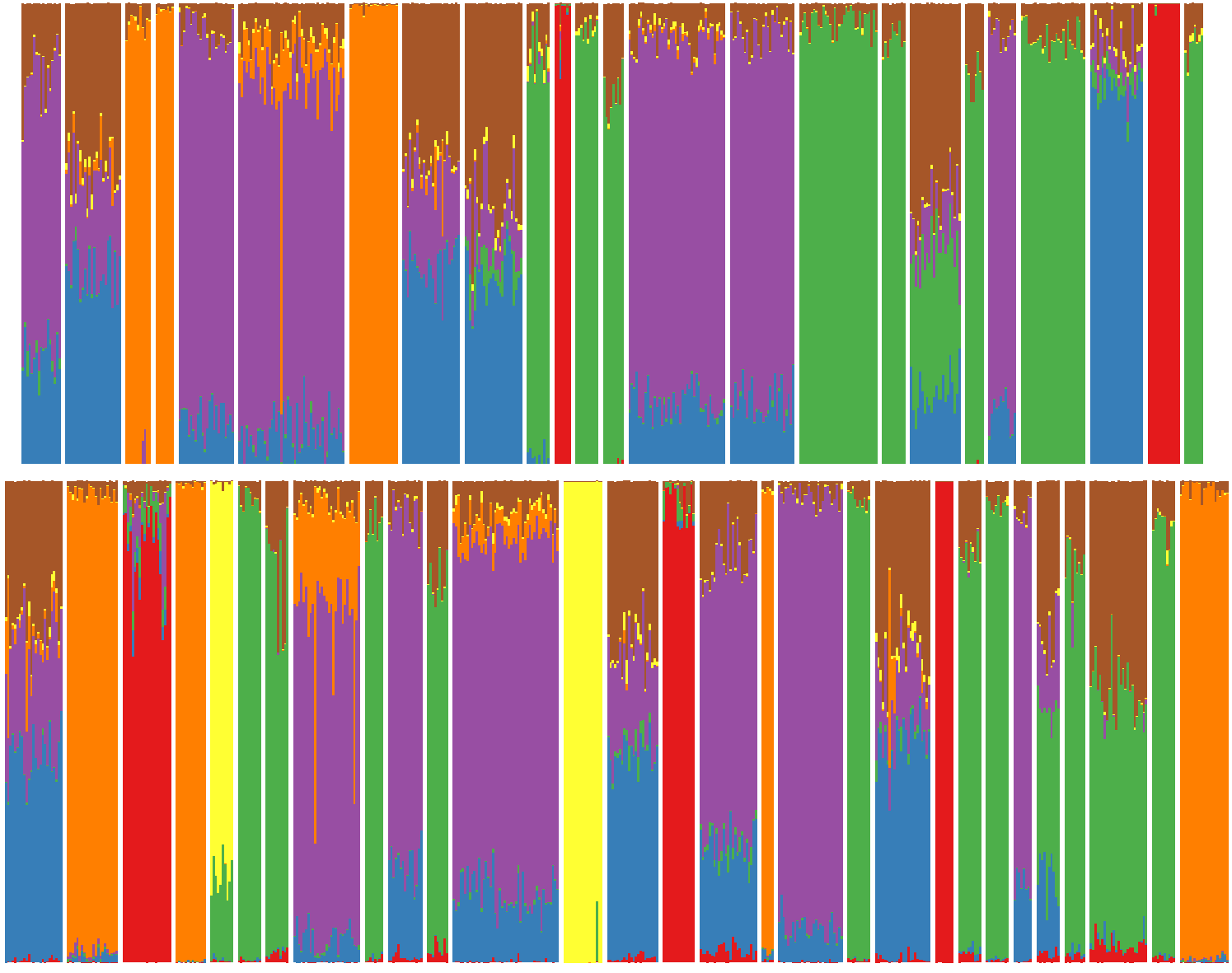
(from my research group and others)



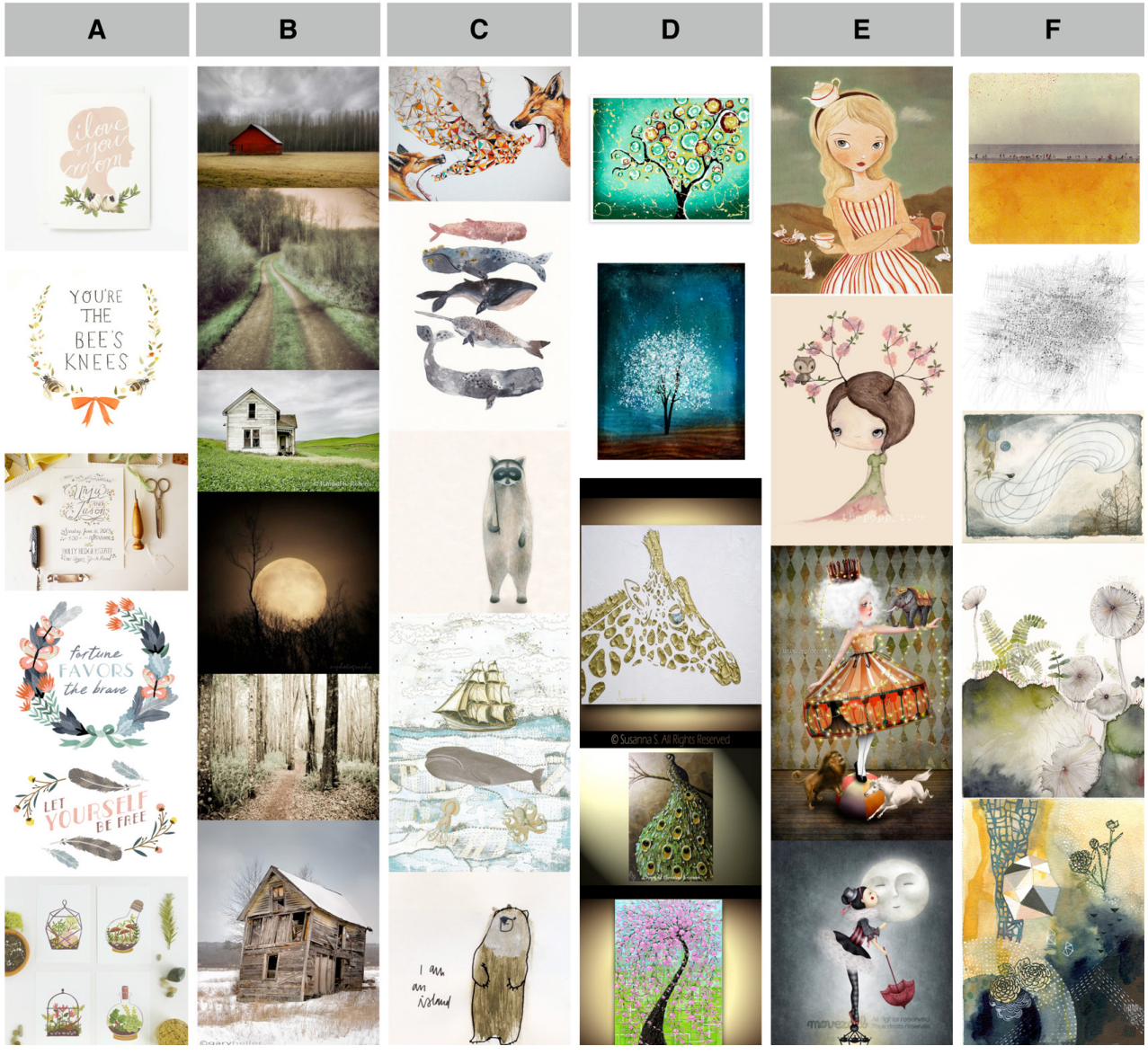
Communities discovered in a 3.7M node network of U.S. Patents



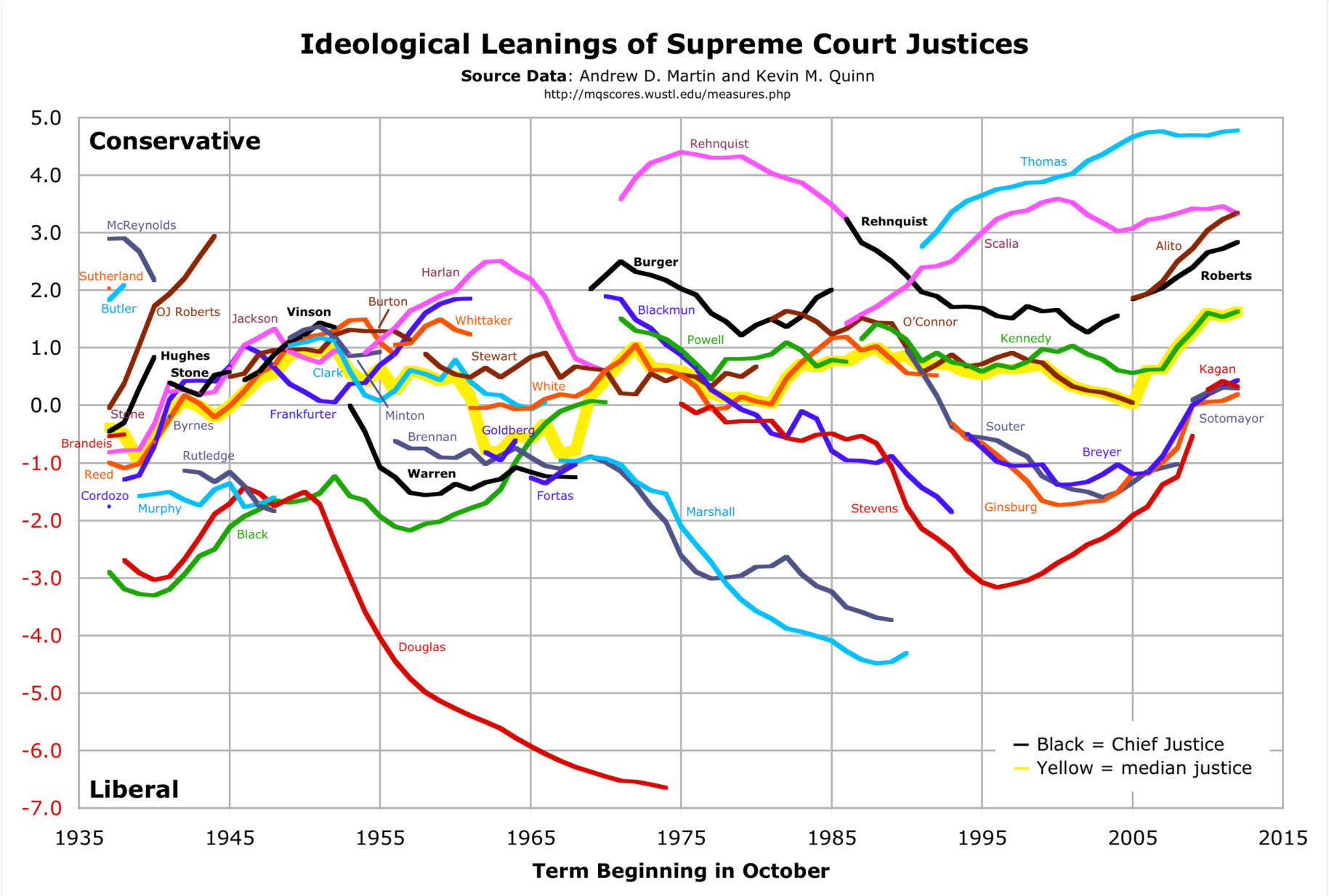
Neuroscience analysis of 220 million fMRI measurements



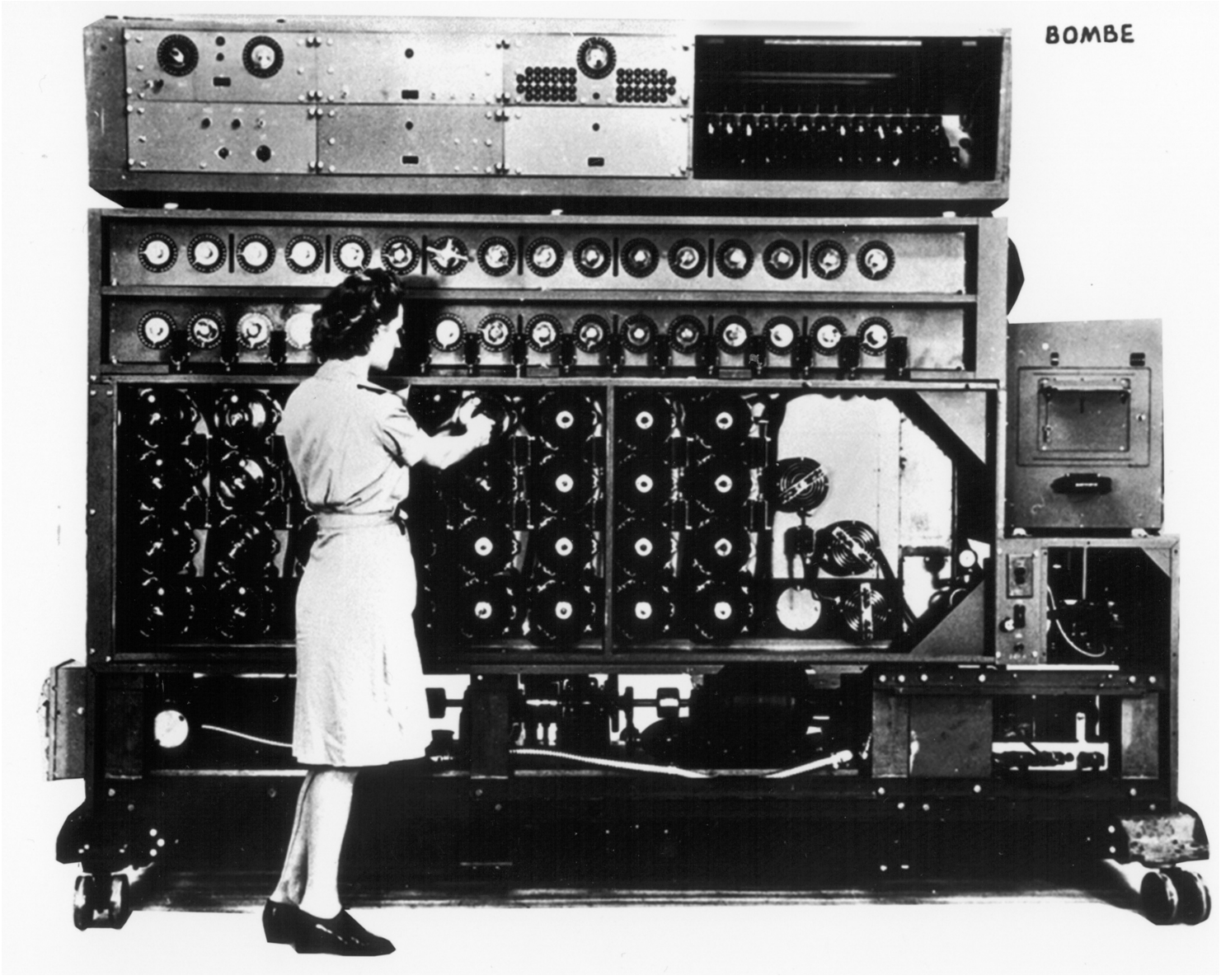
Population analysis of 2 billion genetic measurements



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

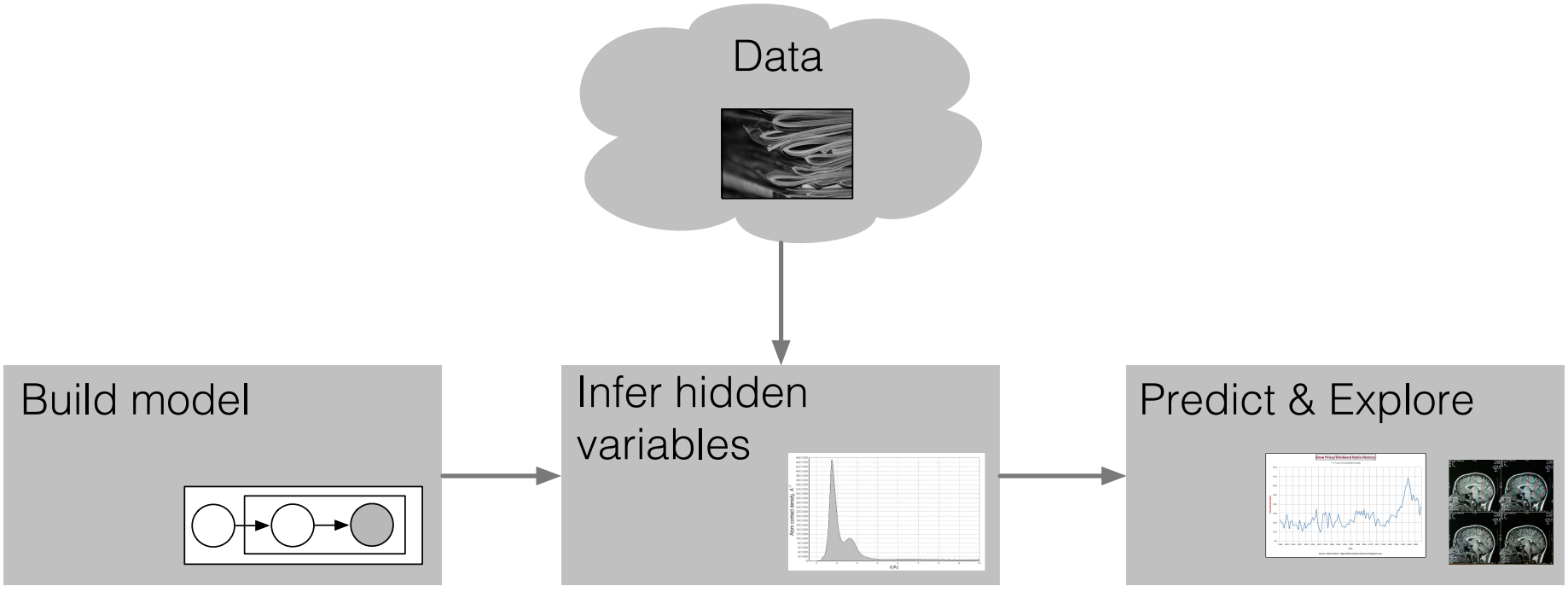


Supreme Court Ideology over time (Martin and Quinn, 2001)



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

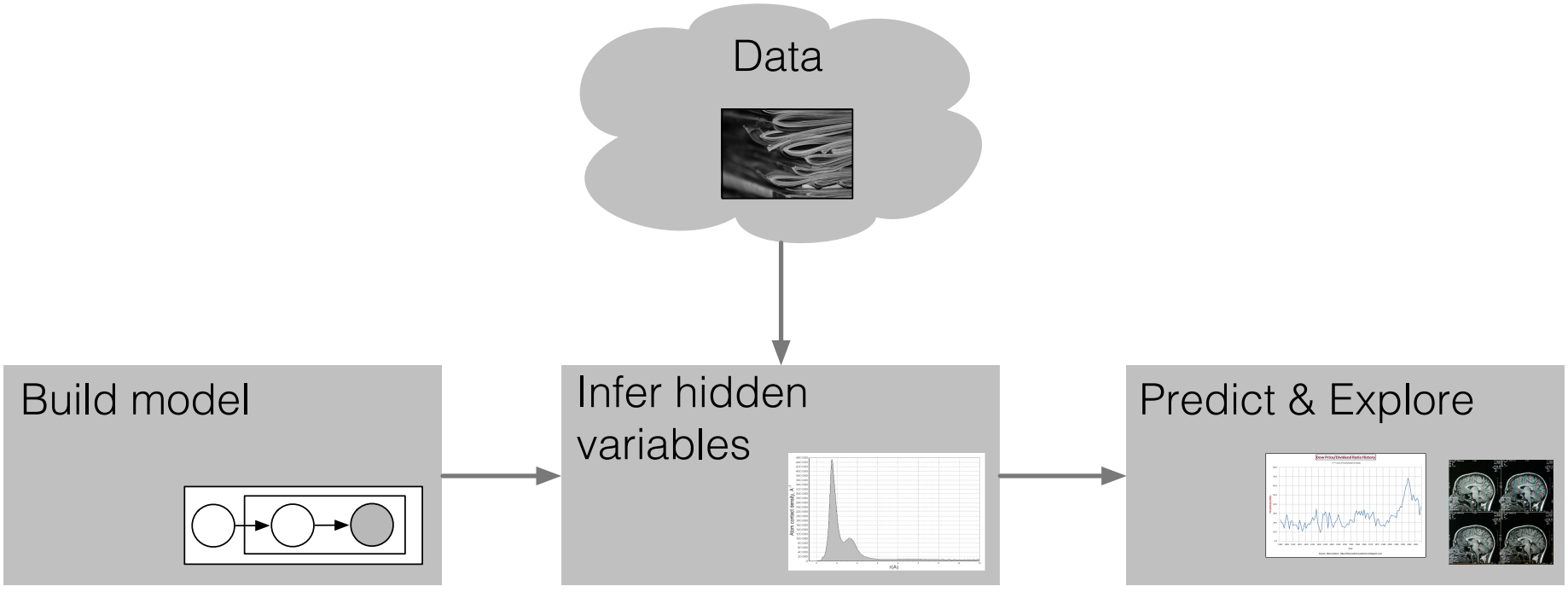
Box's Loop



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.

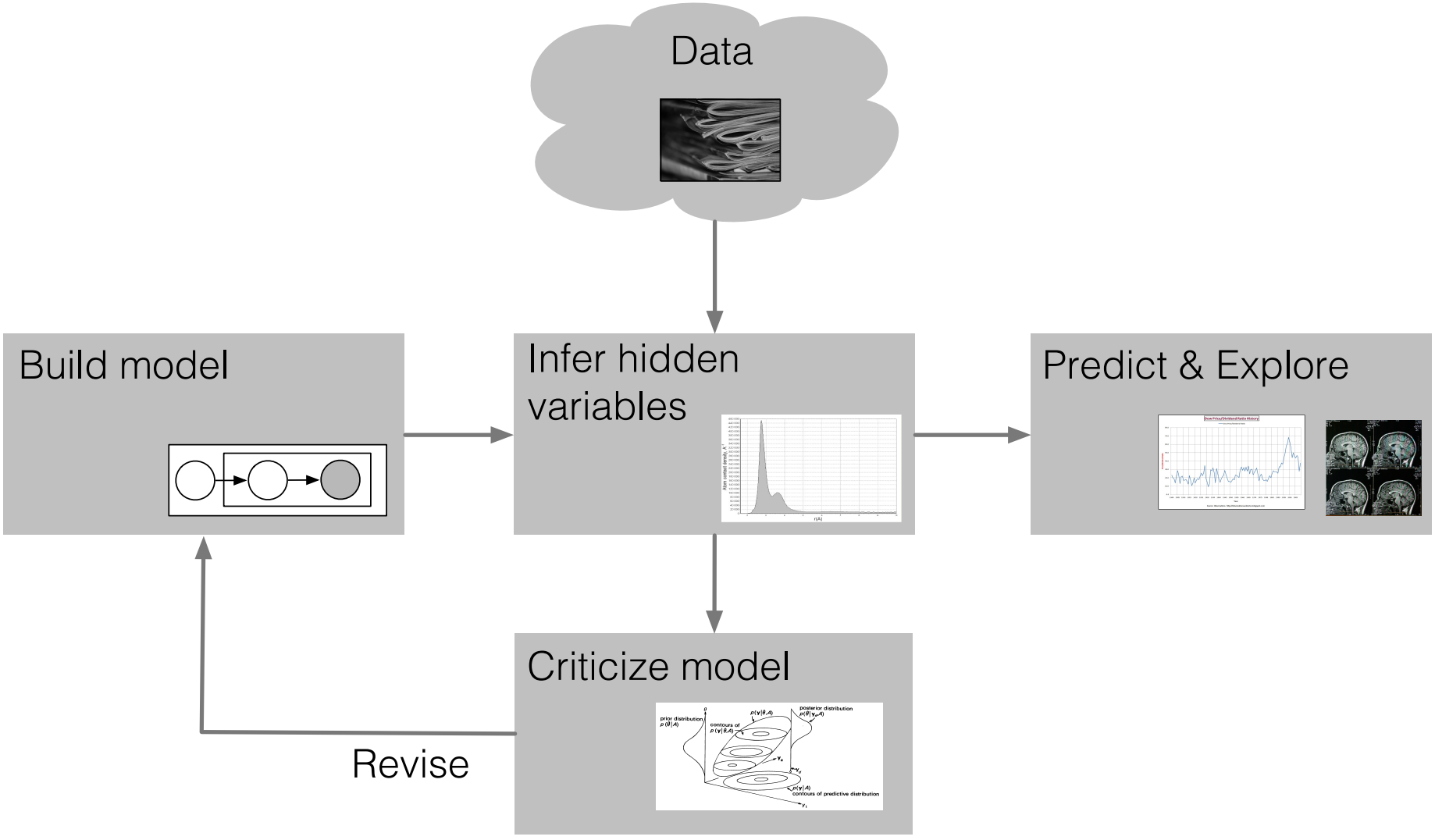
Box's Loop



What we need:

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

Box's Loop


Data
Build model
Infer hidden variables
Predict & Explore
Criticize model
Revise
prior distribution
contours of
posterior distribution
contours of predictive distribution

What will we cover?

The basics of graphical models

- Basic concepts in probability; the semantics of graphical models
- D-separation and conditional independence in graphical models
- Message passing, tree propagation, and a word about the junction tree

Using probability models to analyze data

- Probability models, data, and statistical concepts
- Bayesian mixtures of Gaussians and why we need approximate inference
- Markov chain Monte Carlo sampling and the Gibbs sampler
- The exponential family, conjugacy, and mixtures of exponential families
- Variational inference (and a word about expectation maximization)

The building blocks of complex models

- Mixtures and mixed membership models (including topic models)
- Matrix factorization: Gaussian, Poisson, exponential family
- Time series models: Hidden Markov models and state-space models
- Spatial models
- Regression: Linear and logistic, generalized linear models, regularization
- Bayesian nonparametrics I: Clustering models
- Bayesian nonparametrics II: Latent feature models
- Bayesian nonparametrics III: Gaussian processes

Advanced topics

- Scalable inference with stochastic variational inference
- Model checking and revision with posterior predictive checks
- Hierarchical models, multi-level models, empirical Bayes
- Causal inference and probabilistic modeling (a rabbit hole)

Some additional discussion

- Programming languages
- Applications
- Box's loop, again
- Note: We will usually be at the board.

Prerequisites, Requirements, Grades, Etc.

- http://www.cs.columbia.edu/∼blei/fogm/
- Office hours: Wednesdays 2:30-4:00, 703 CEPSR
- Prerequisites
  - Probability and Statistics
  - Optimization
  - Programming
- Requirements
  - Weekly paper about the reading ($<$ 1 page)
  - Final project
- Your grade: Mostly the final project