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
- 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
- Do Homework 0 (due Friday)

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

What is ***complicated data****?*

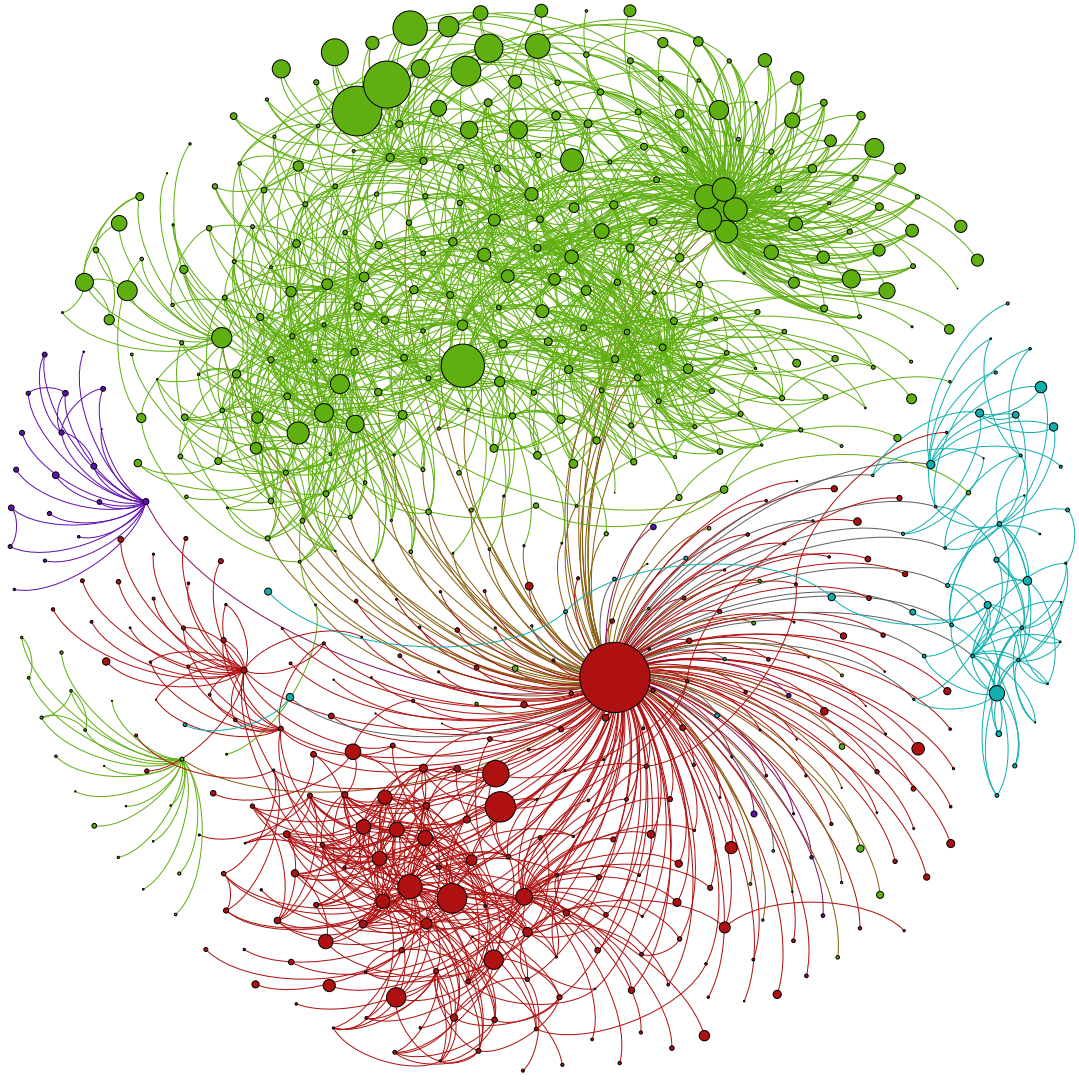
- many data points; many dimensions
- unstructured (e.g. text)
- multimodal and interconnected (e.g., images, links, text, clicks)

What is ***making sense of data***?

- make predictions about the future
- identify 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 modeling
- goal: A methodology that is *expressive, scalable, easy to develop*

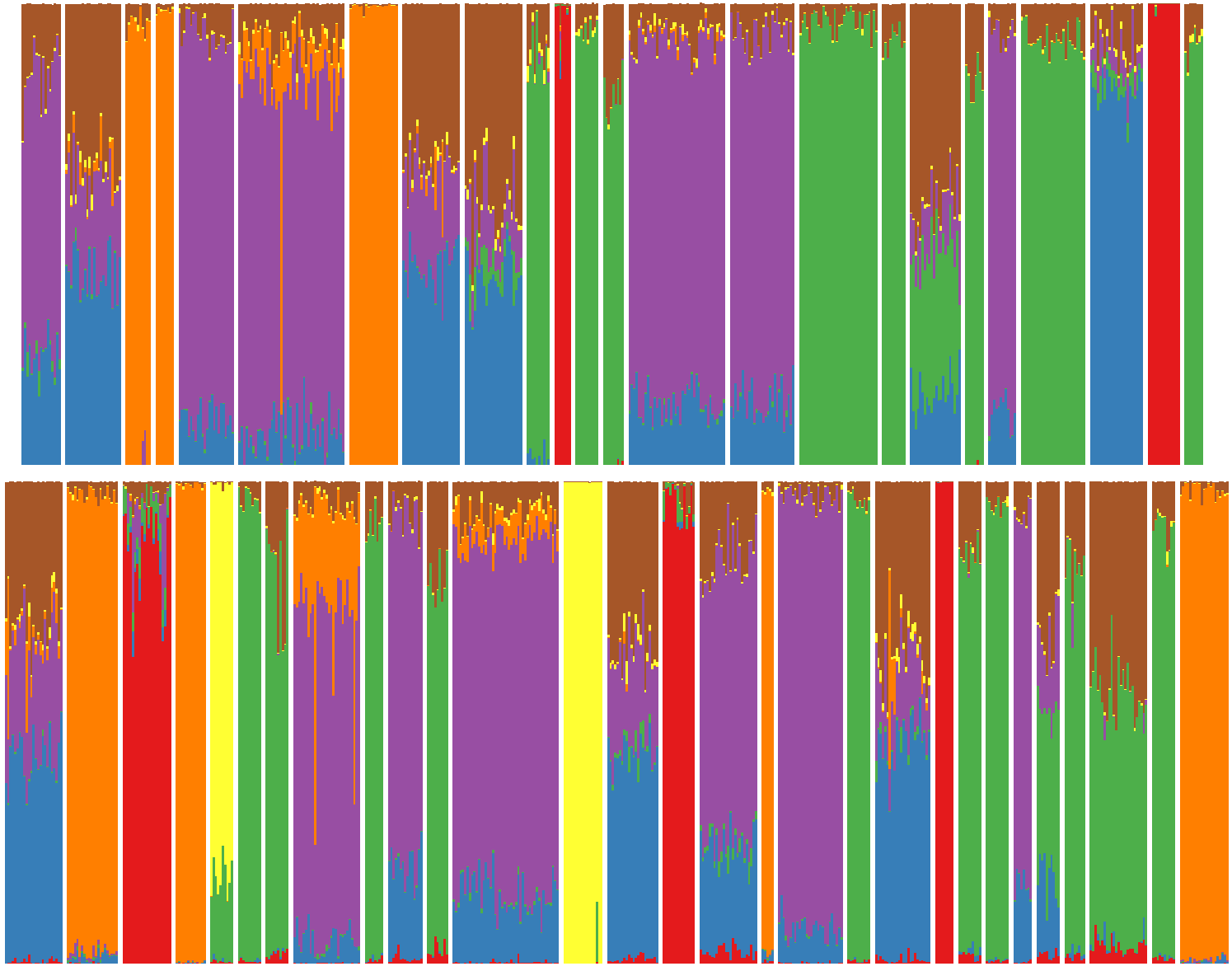


Communities discovered in a 3.7M node network of U.S. Patents

[Gopalan and Blei PNAS 2013]

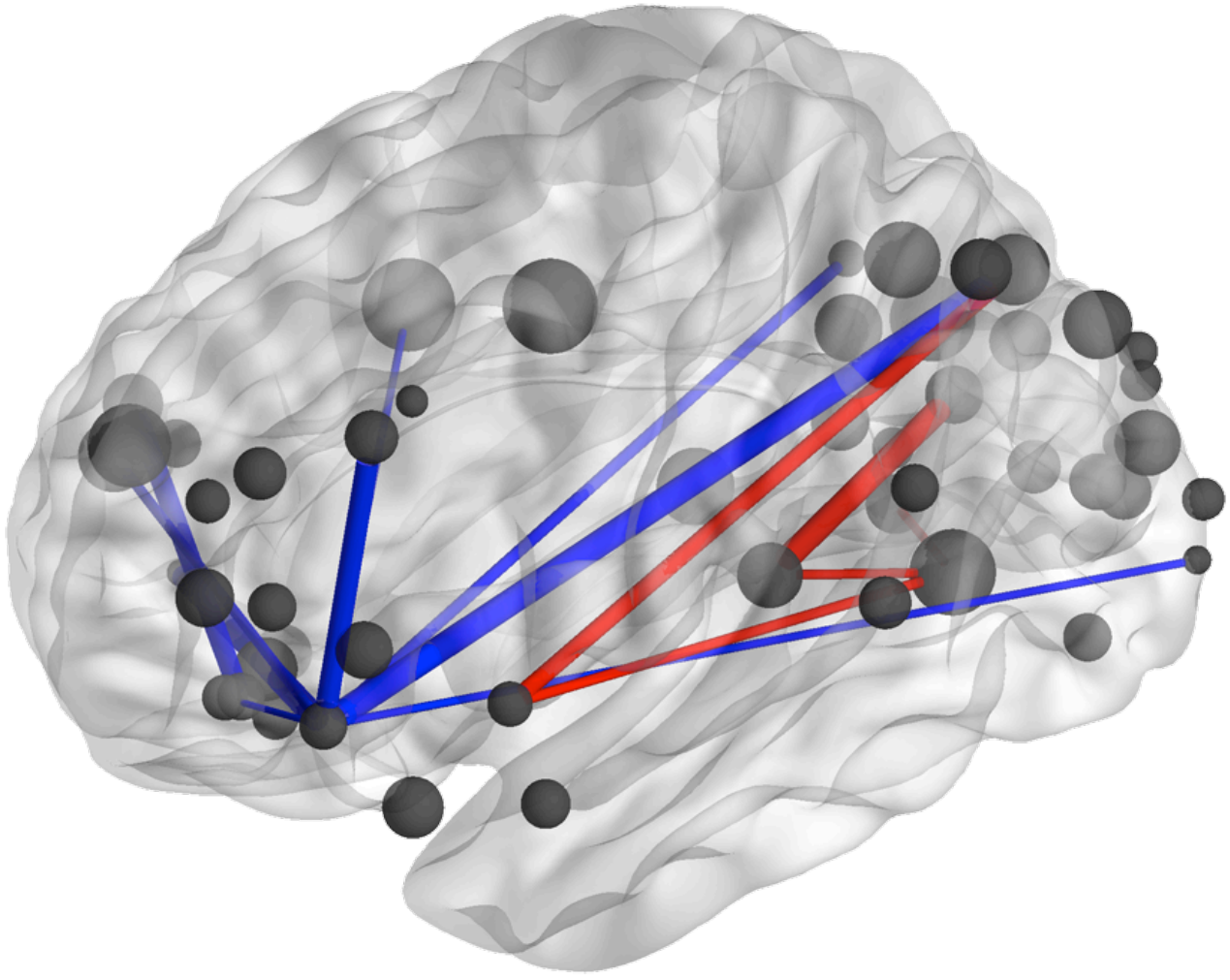
Topics found in 1.8M articles from the New York Times

[Hoffman+ JMLR 2013]



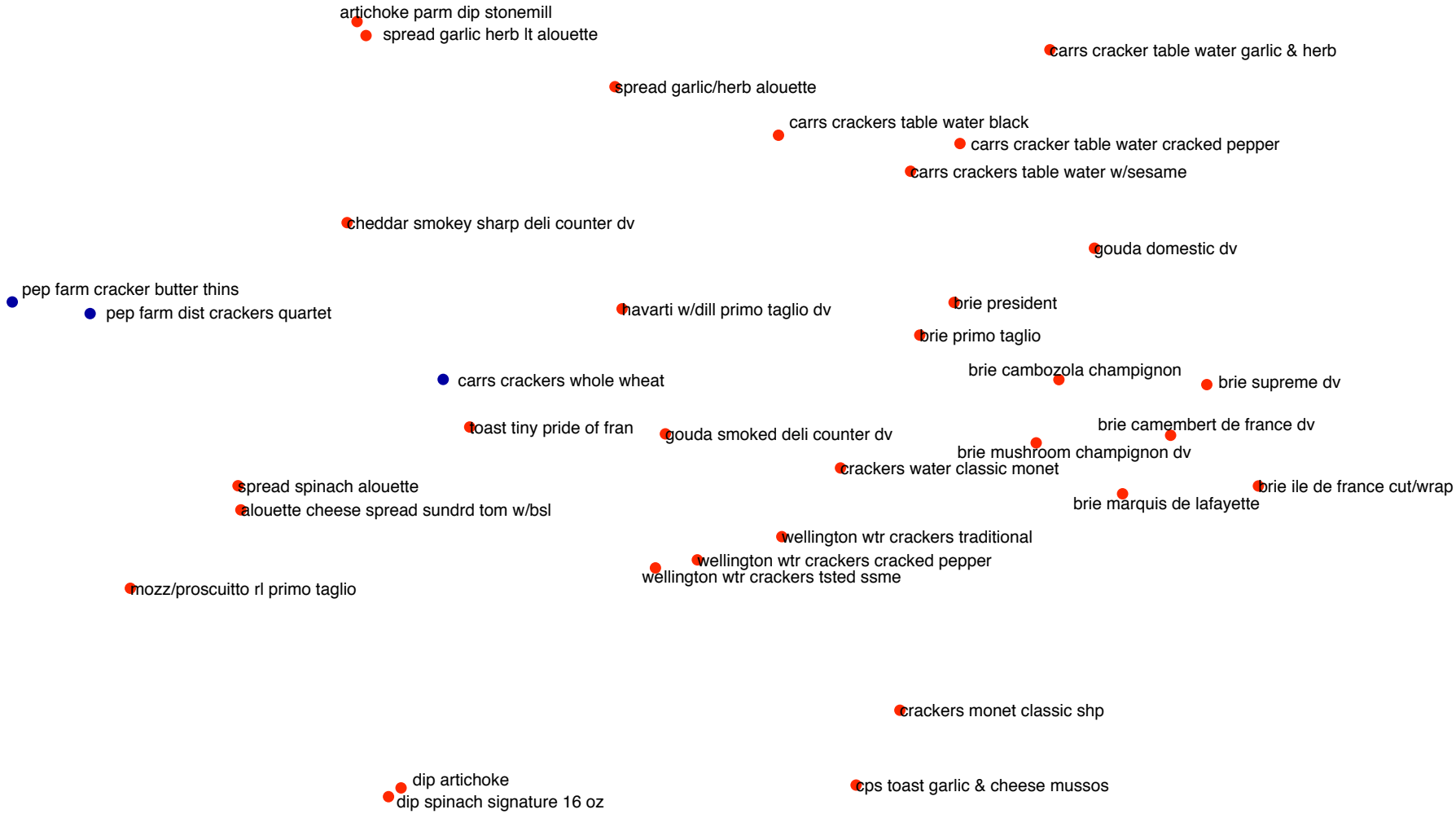
Population analysis of 2 billion genetic measurements

[Gopalan+ Nature Genetics 2016]



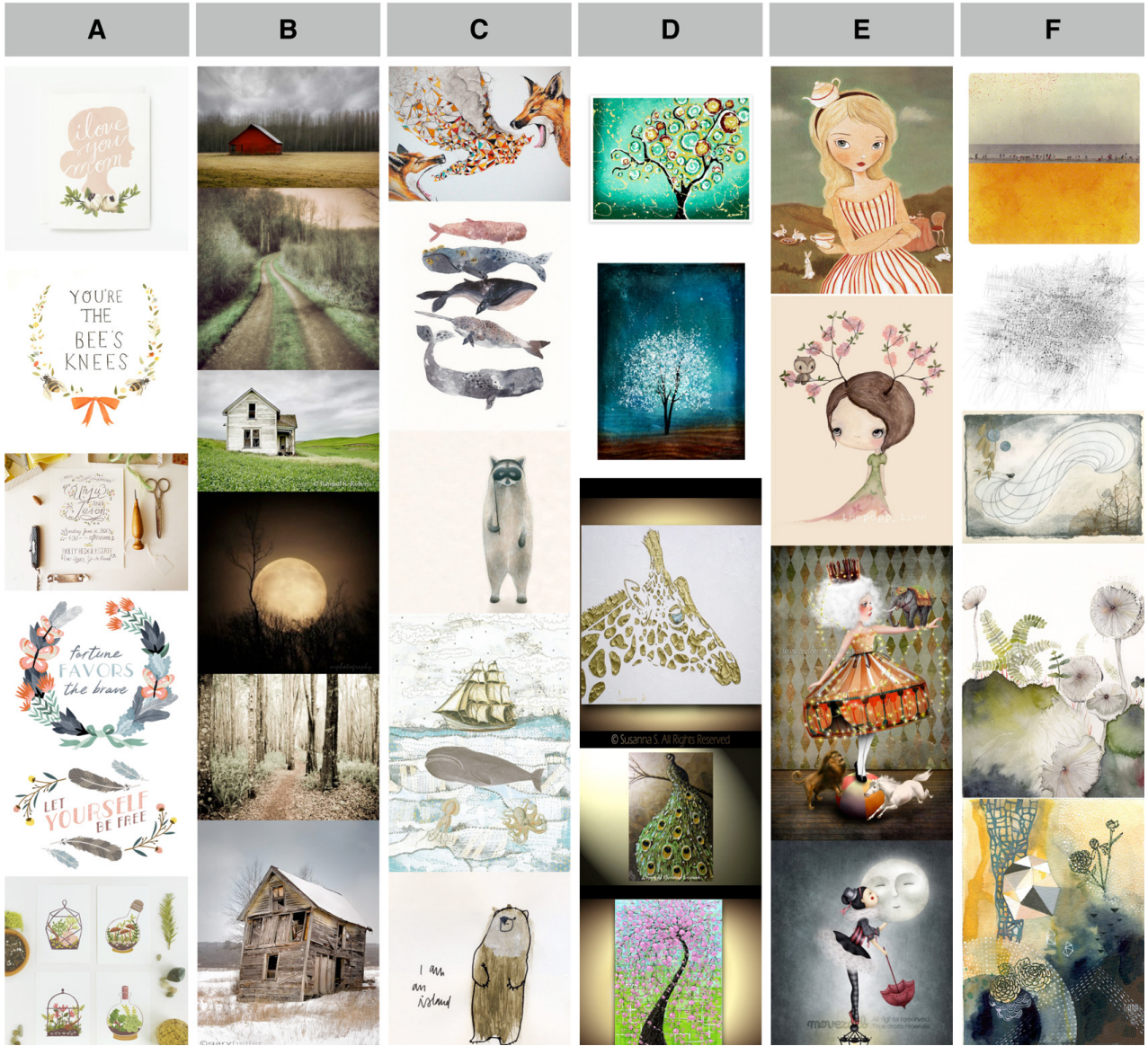
Neuroscience analysis of 220 million fMRI measurements

[Manning+ PLOS ONE 2014]

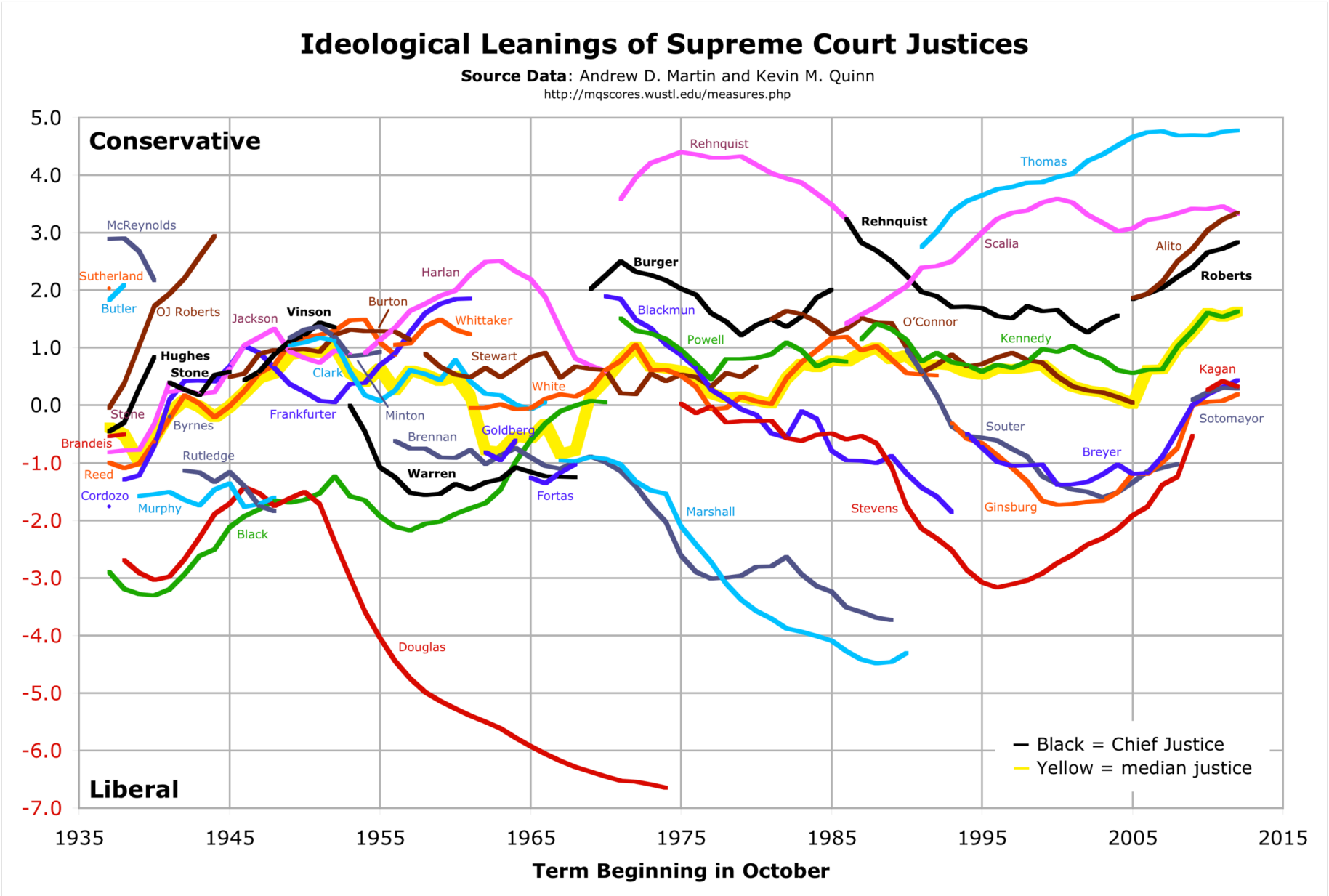


SHOPPER analysis of 5.7M purchases

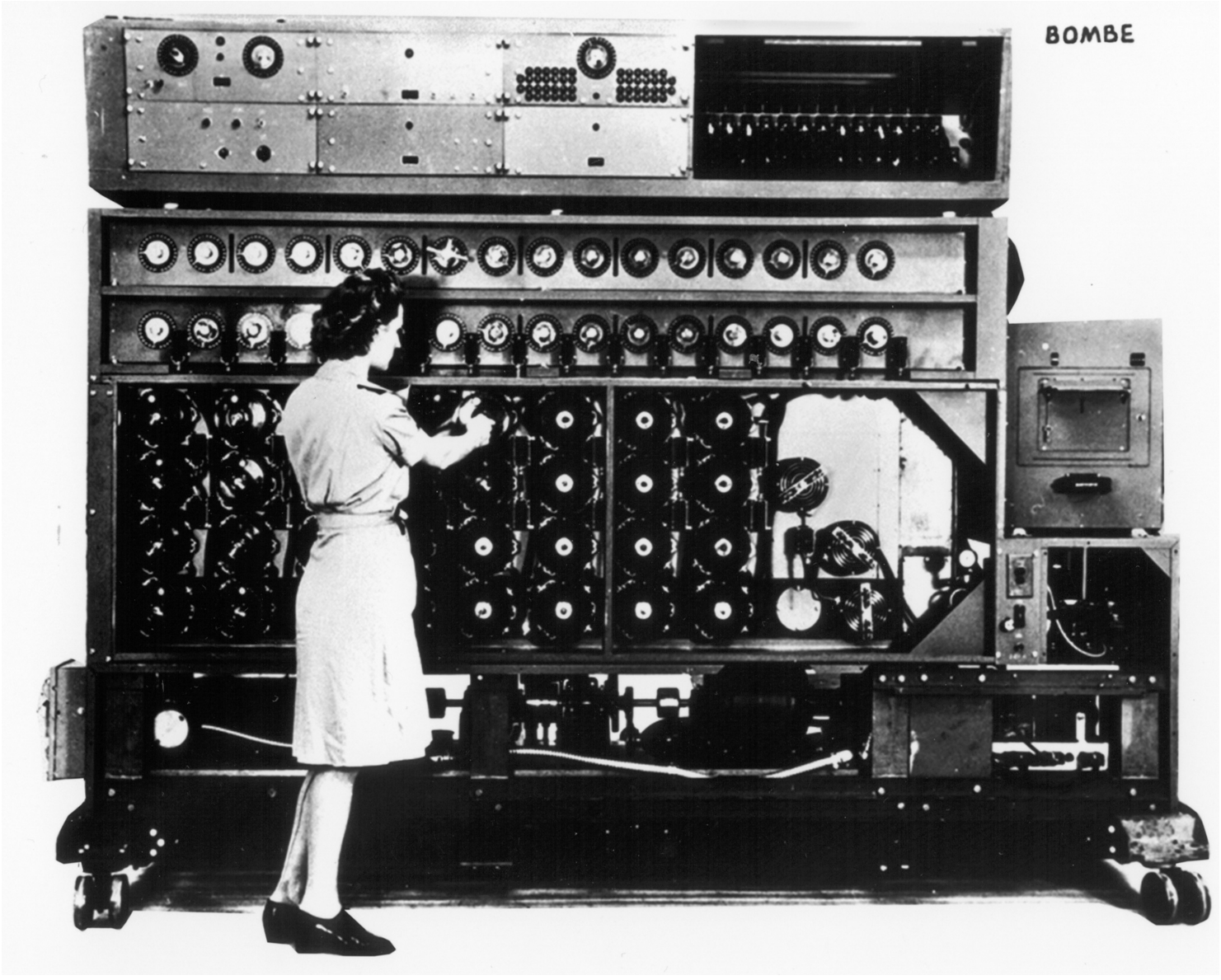
[Ruiz+ 2017]



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

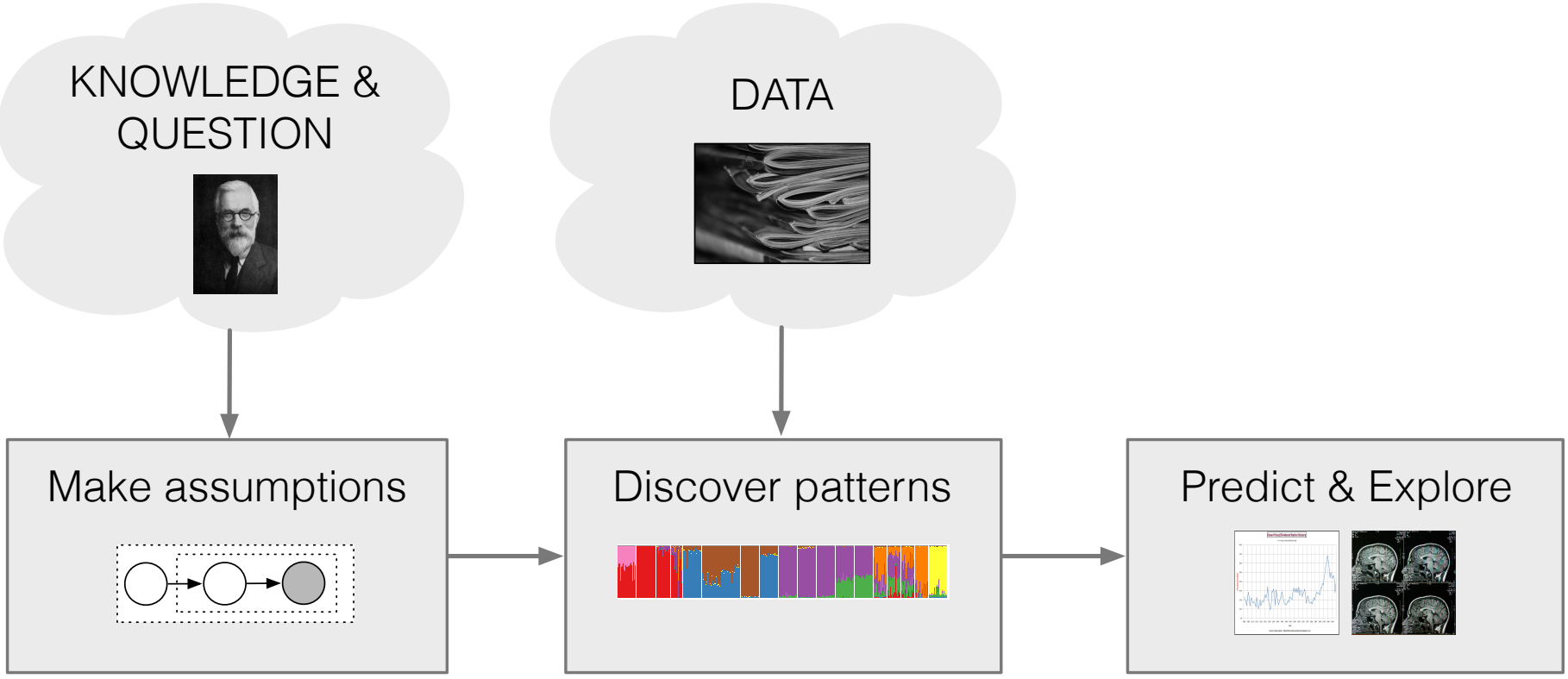


Supreme Court Ideology over time (Martin and Quinn, 2001)



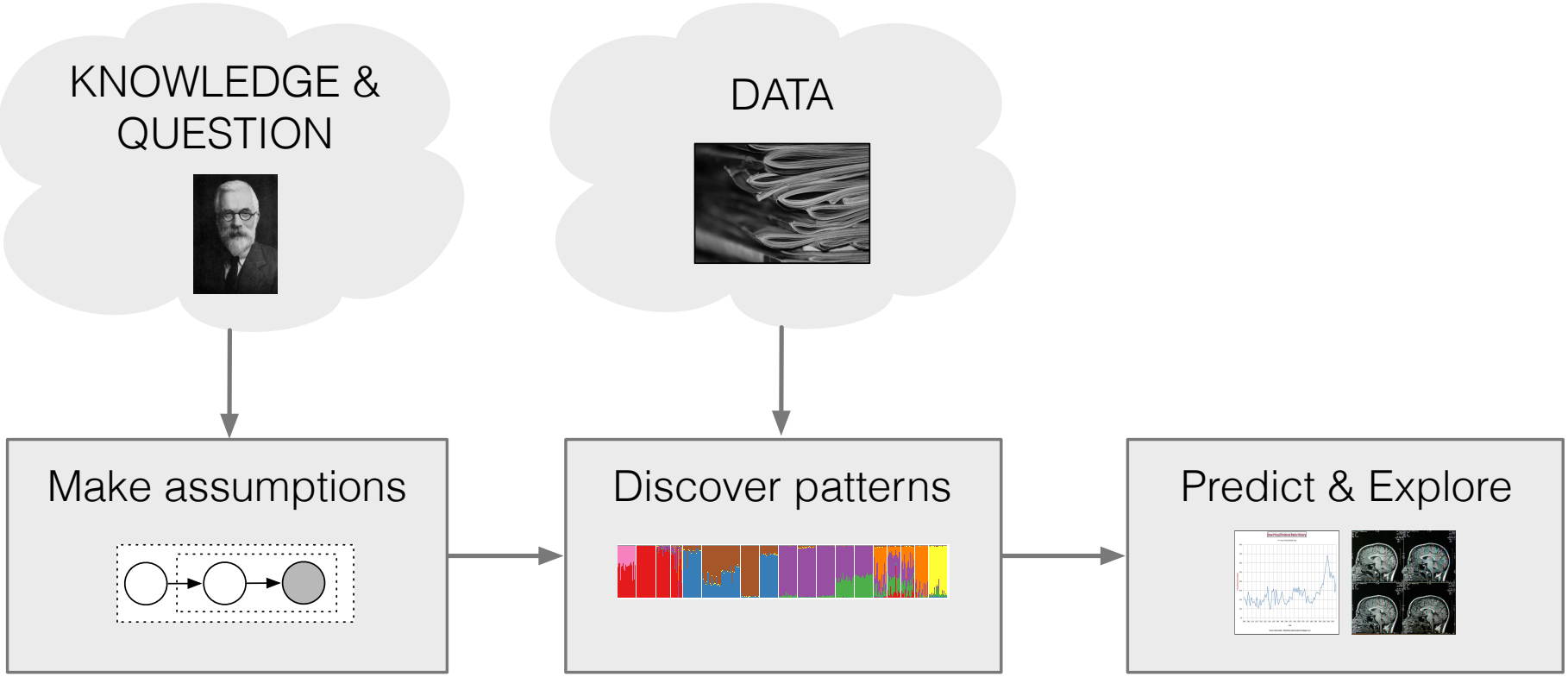
Breaking the Nazi code (Turing and Good, 194?)

Box's Loop



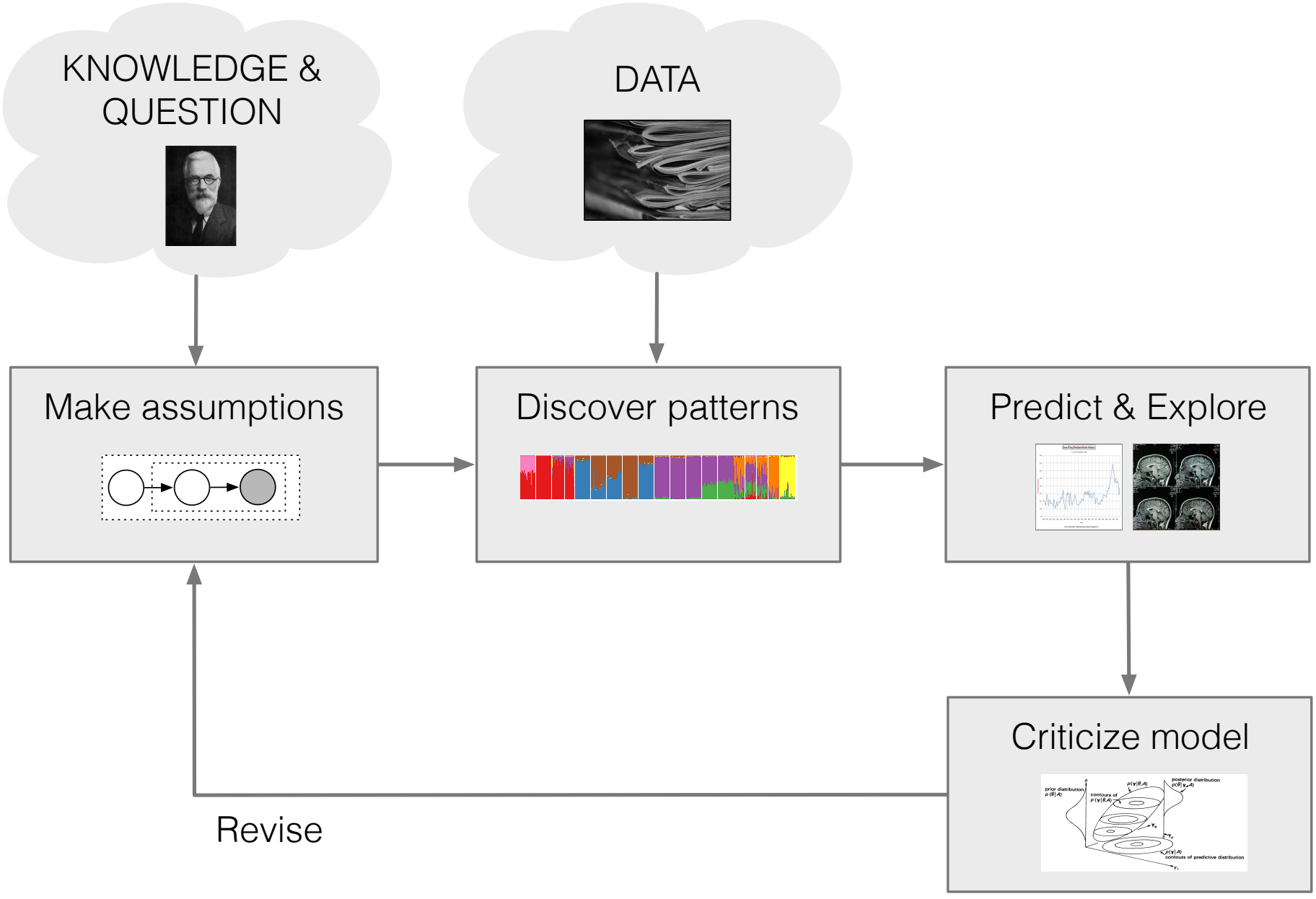
- Customized data analysis is important to many fields.
- This pipeline separates *assumptions*, *computation*, *application*.
- It facilitates solving data science problems.

Box's Loop



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

Box's Loop


KNOWLEDGE & QUESTION
DATA
Make assumptions
Discover patterns
Predict & Explore
Criticize model
Revise

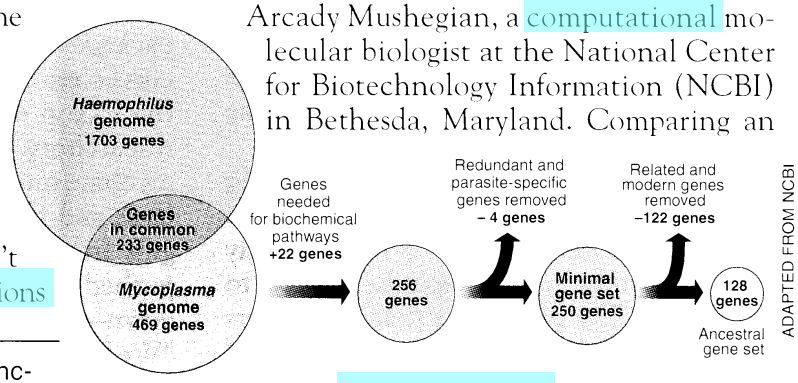
Latent Dirichlet Allocation

(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.

Documents exhibit multiple topics.

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

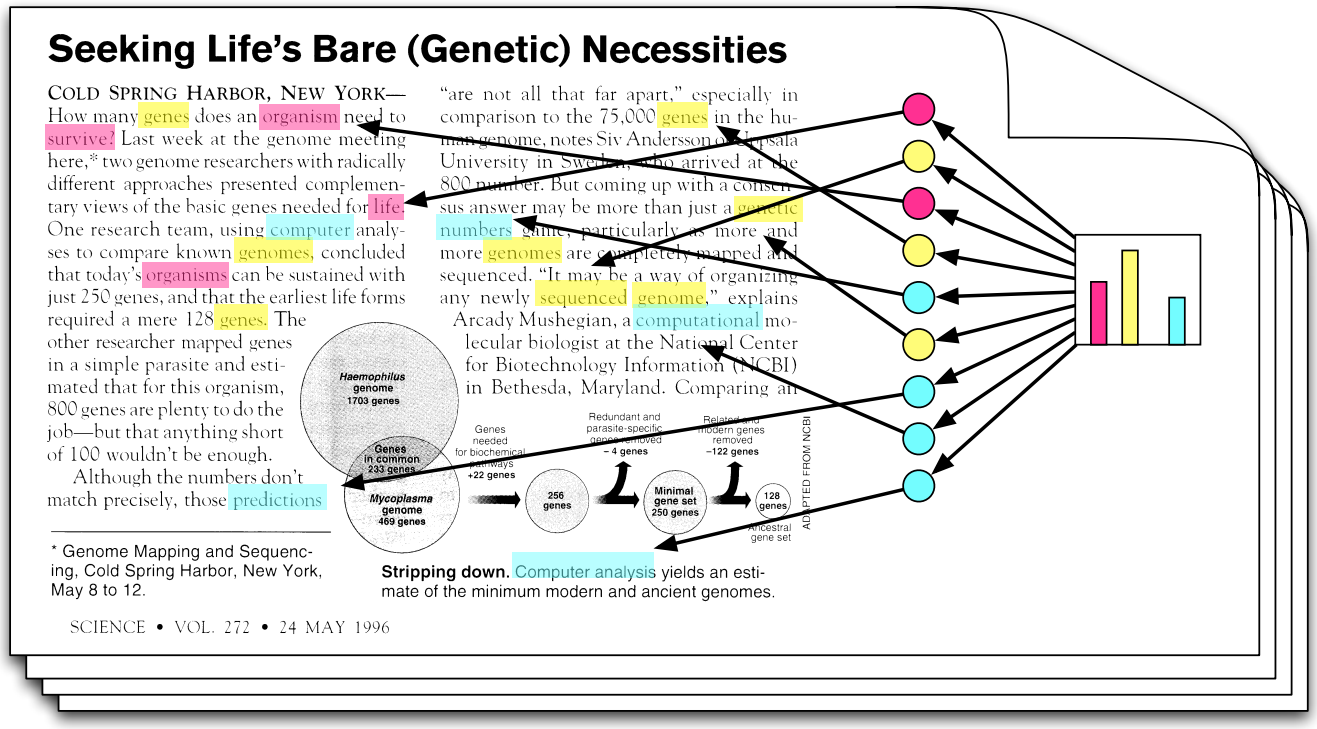
life	0.02
evolve	0.01
organism	0.01
...	

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.

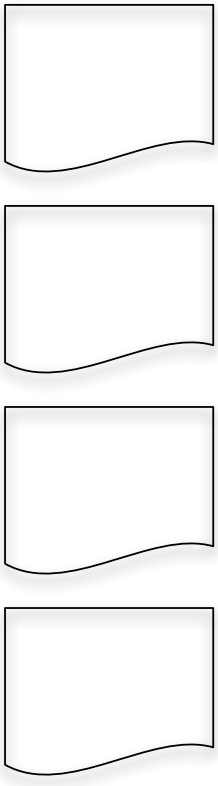
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

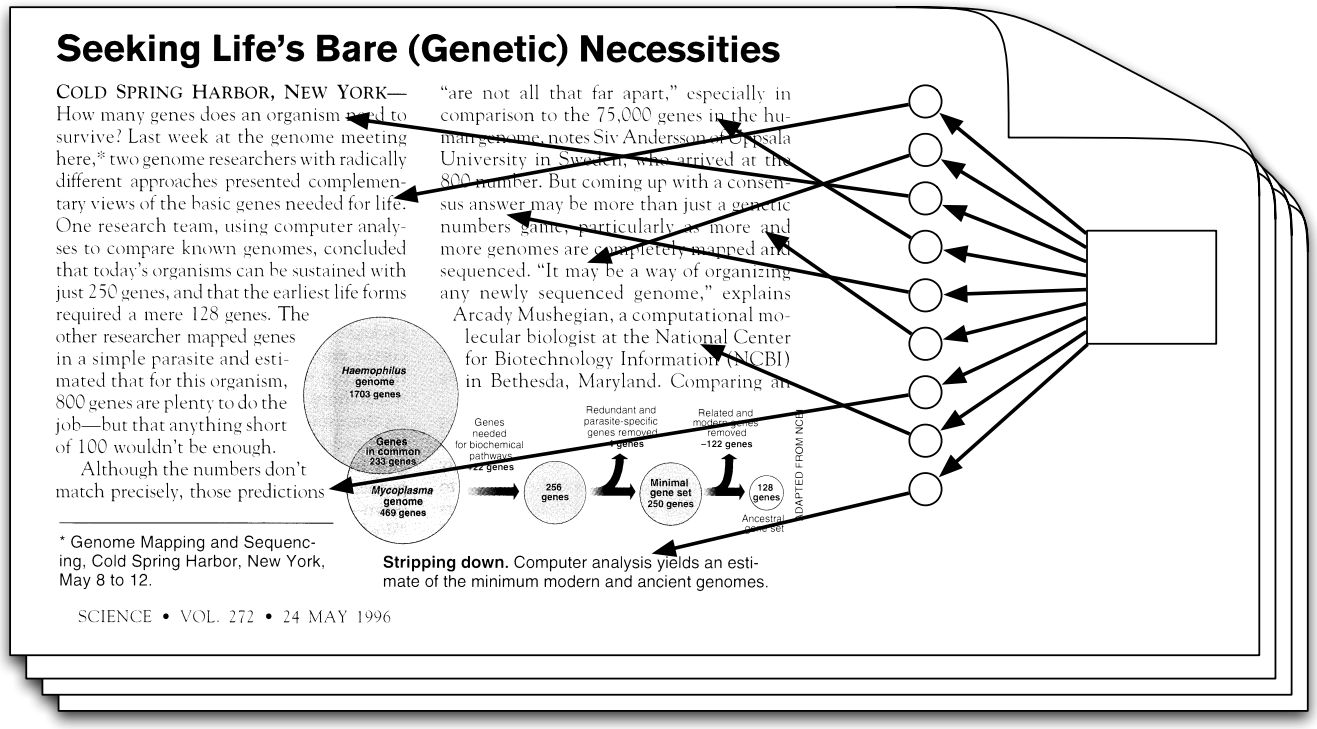
Latent Dirichlet Allocation

Topics

Documents

Topic proportions and assignments





Seeking Life's Bare (Genetic) Necessities

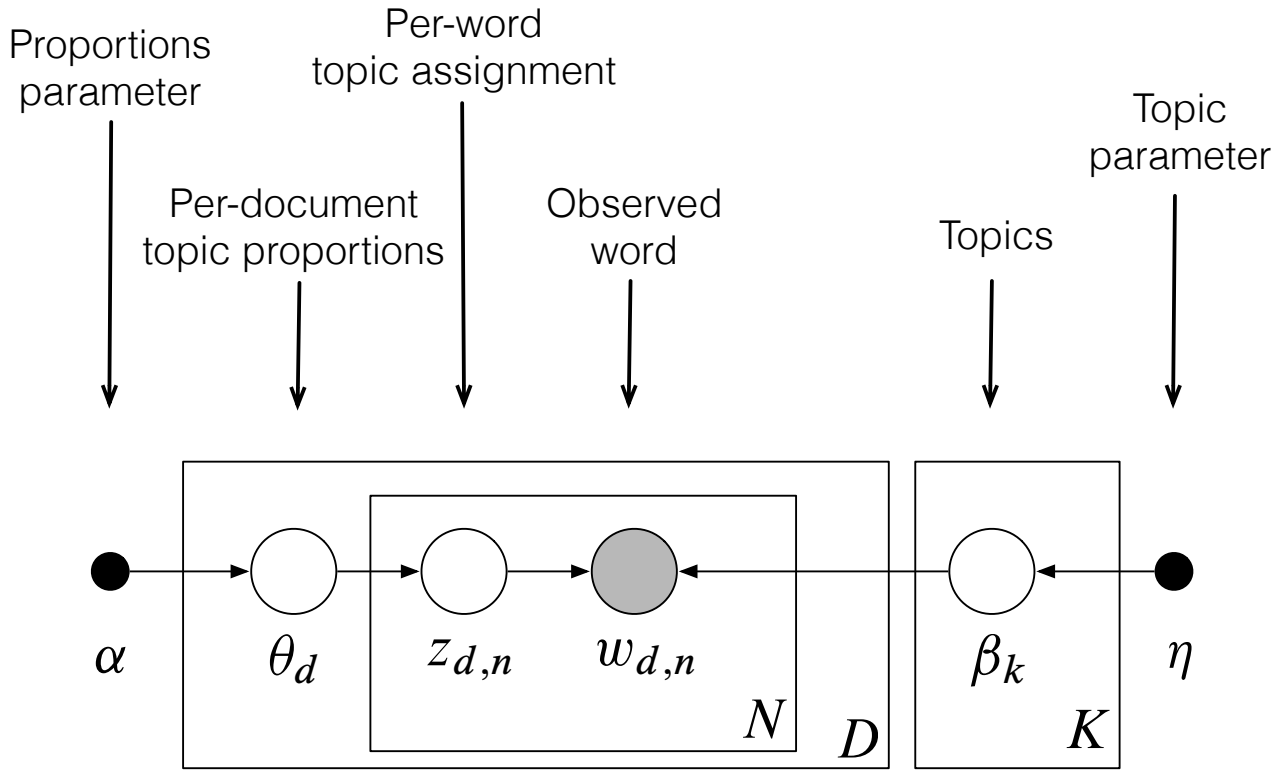
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

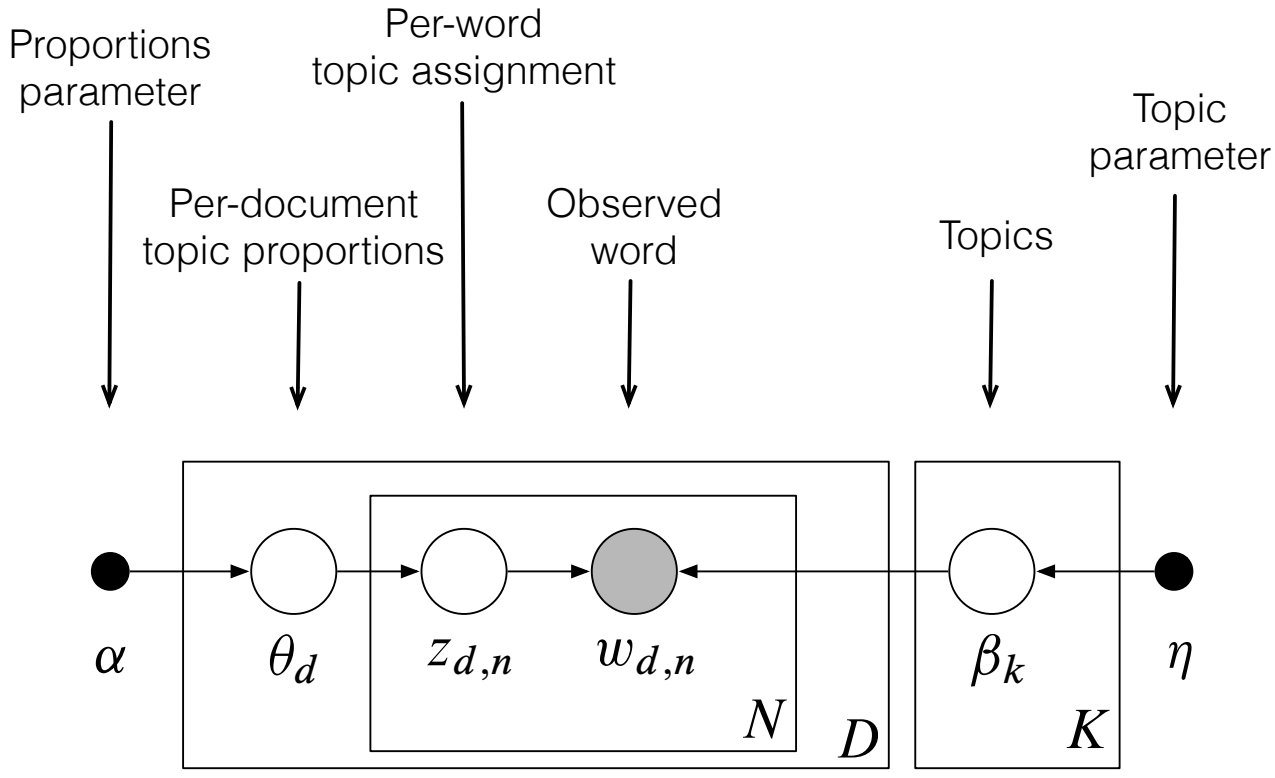
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Latent Dirichlet Allocation



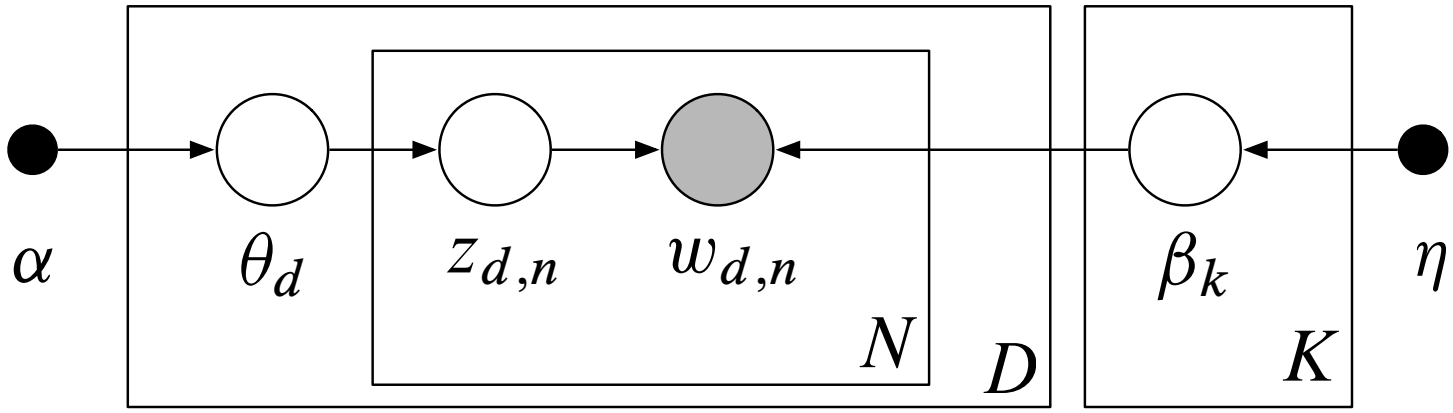
LDA as a graphical model

- 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



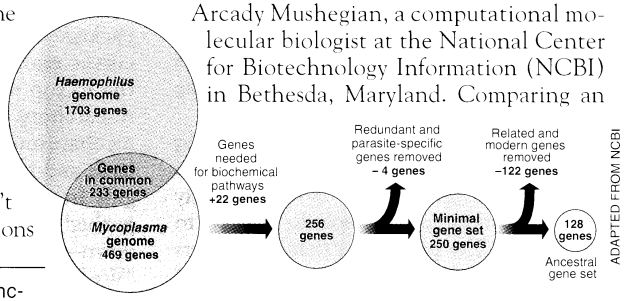
- The joint defines a posterior, $p(\theta, z, \beta \mid w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand: information retrieval,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

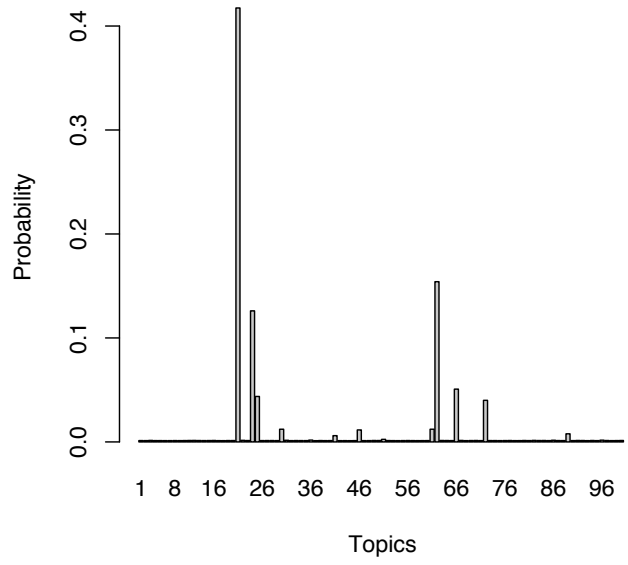
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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



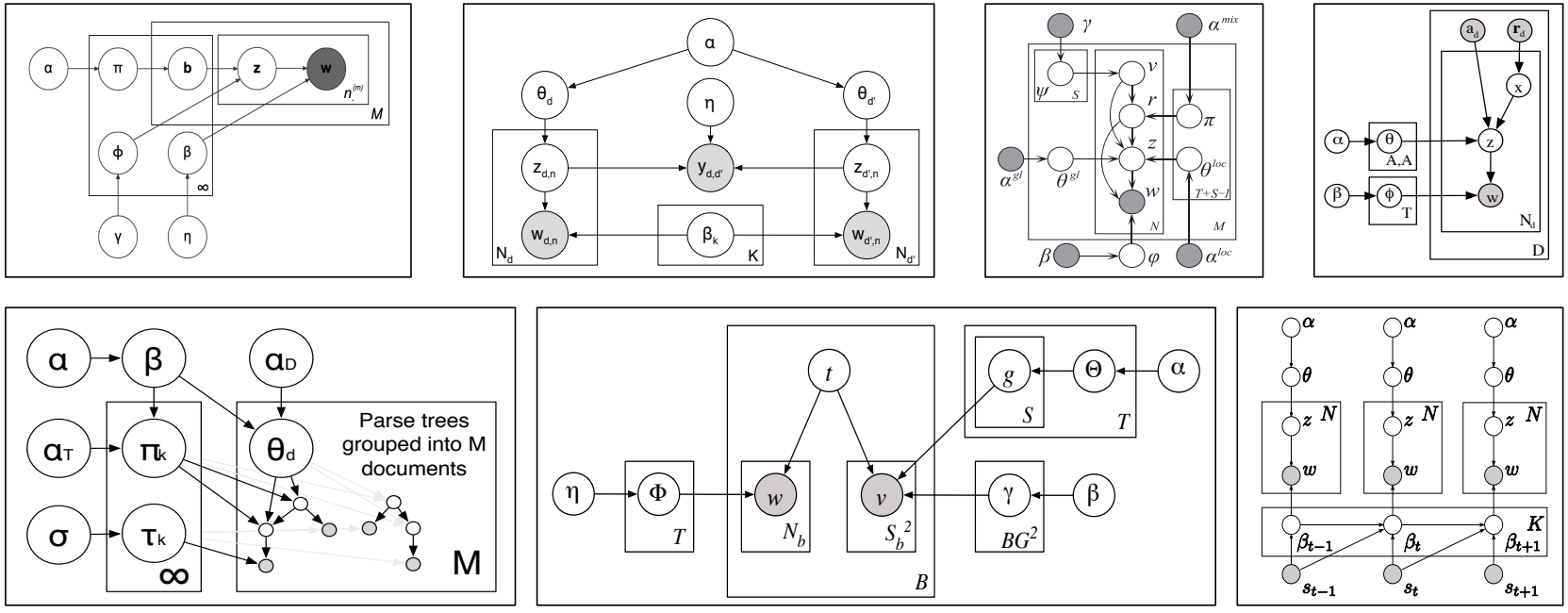
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* 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

Topics found in 1.8M articles from the New York Times



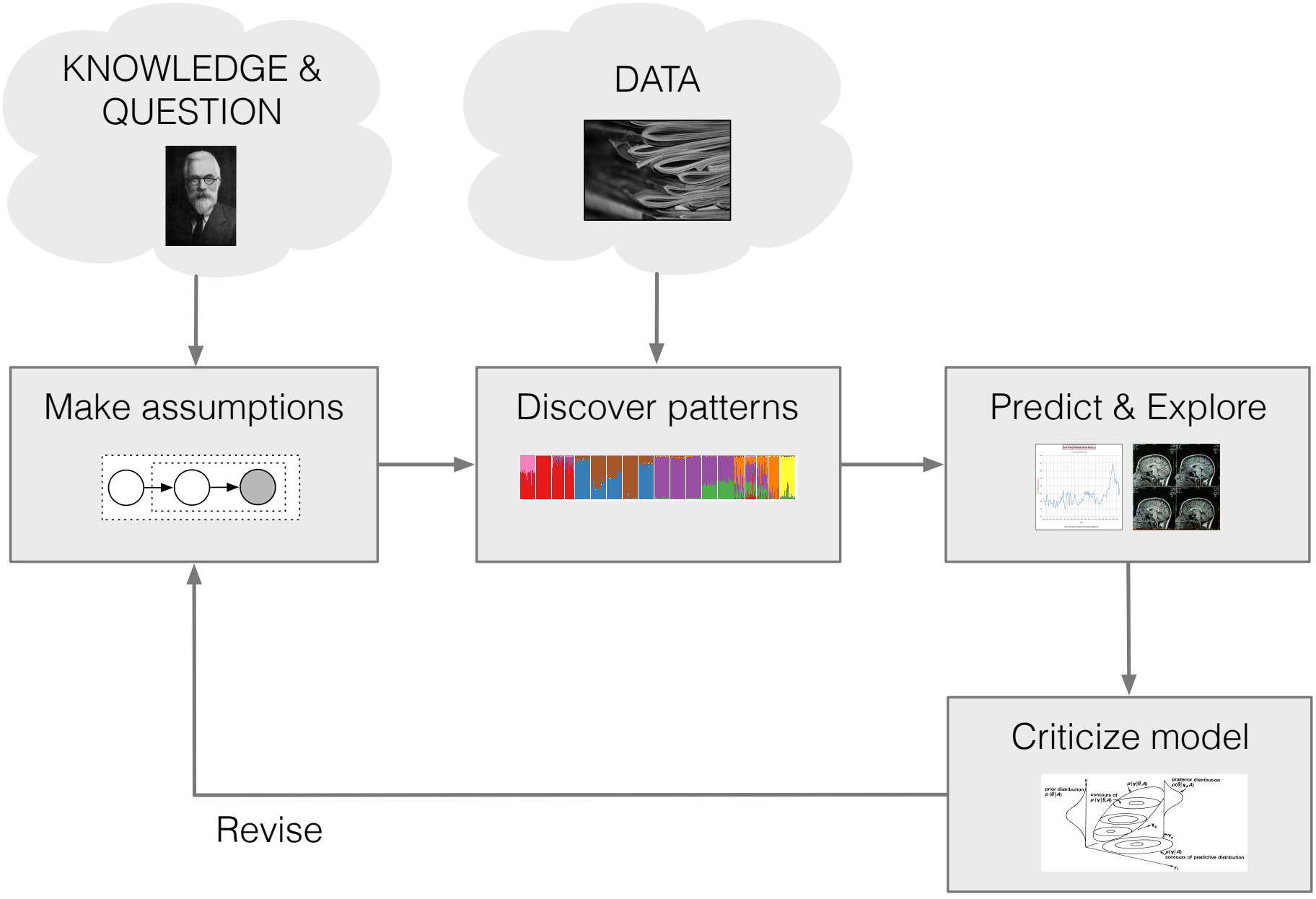
- LDA is a simple building block that enables many applications.
  Topic modeling is an active field of research.

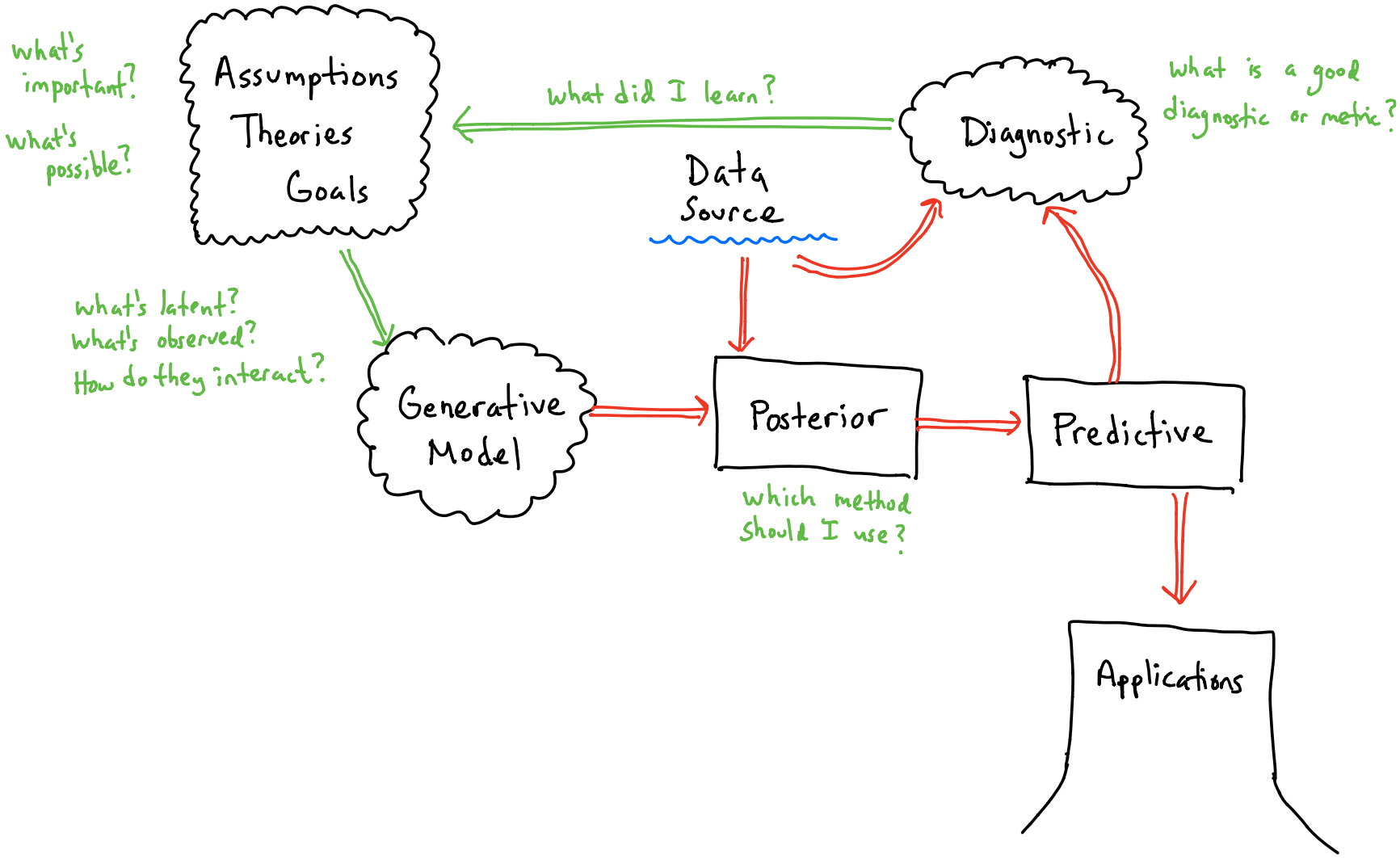
- Graphical models are a composable language for probability models.

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

What will we cover in this class?

Box's Loop


KNOWLEDGE & QUESTION
DATA
Make assumptions
Discover patterns
Predict & Explore
Criticize model
Revise



A problem that involves data → a solution that involves a probability model

Preliminaries

- Probability: Basic concepts and review
- The ingredients of probabilistic models

Latent variable models

- Bayesian mixtures and the Gibbs sampler
- Mixed-membership, topic models, and variational inference
- Matrix factorization, recommendation systems, and efficient MAP
- Deep generative models and black box variational inference
- Time series and sequence models
- Exponential families, conjugacy, and generalized linear models
- Hierarchical models, robust models, and empirical Bayes

Advanced ideas

- The semantics of graphical models and *d*-separation
- Tree propagation and hidden Markov models
- Advanced methods in variational inference
- Model checking and model diagnosis
- An introduction to causality

What this course is

- This course is about *doctoral research in probabilistic machine learning*.
  - New applications
  - New methods
  - New theories
- (It is not about learning a cookbook of ML methods to get a job at Facebook.)
- Side effects:
  - Learn how to navigate and read the PML literature
  - Learn (from experience) tools for doing PML
  - Understand the relationship between PML and deep learning
  - Learn (from experience) a pipeline for ML research

Prerequisites, Requirements, Grades, Etc.

- http://www.cs.columbia.edu/∼blei/fogm/
- Office hours: Check my website
- TA: Keyon Vafa, Office hours TBD
- Prerequisites
  - You are comfortable with probability and statistics
  - You are comfortable with optimization
  - You know how to program for data analysis (e.g. R, Python, C)
- Requirements
  - Weekly: Choose a reading and write a response
    (Please give me feedback on book chapters!)
  - Occasional homework
  - Final project (and milestones along the way)
- Your grade: Mostly the final project