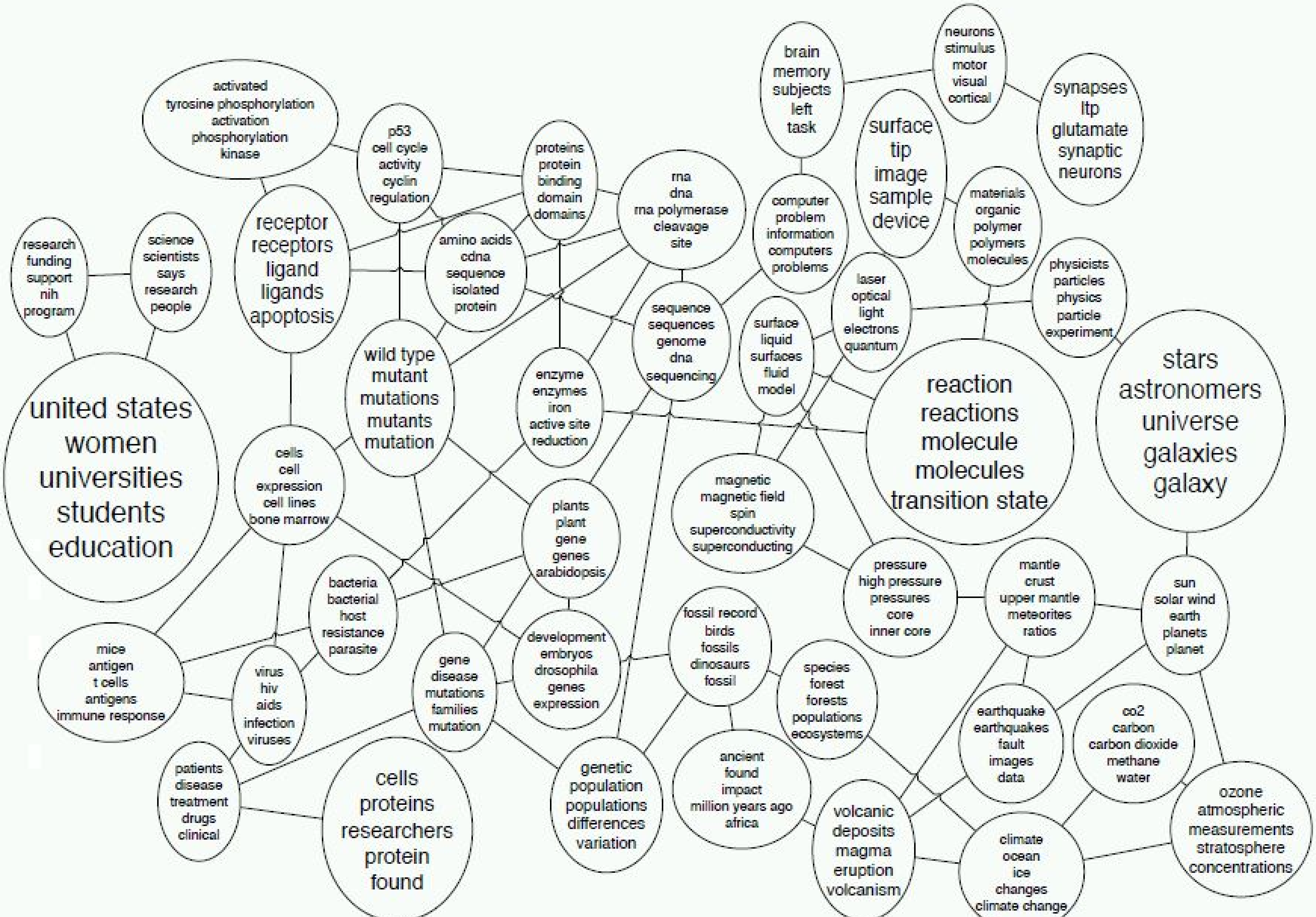
Probabilistic Topic Models and User Behavior

David M. Blei
Columbia University

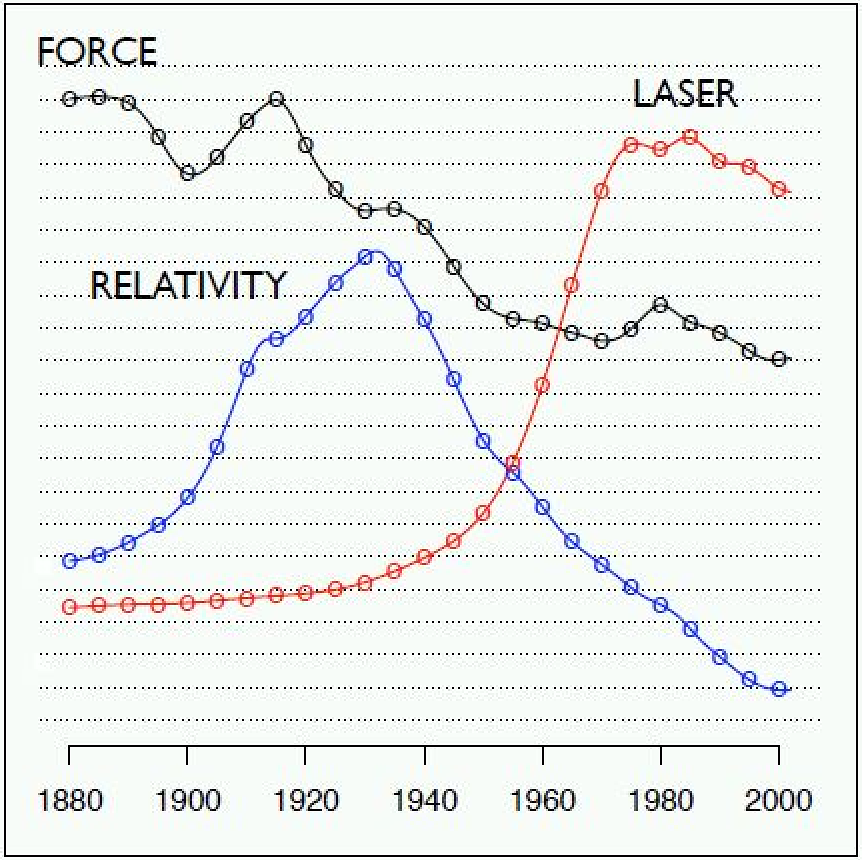
- ORGANIZE
- VISUALIZE
- SUMMARIZE
- SEARCH
- PREDICT
- UNDERSTAND

TOPIC MODELING

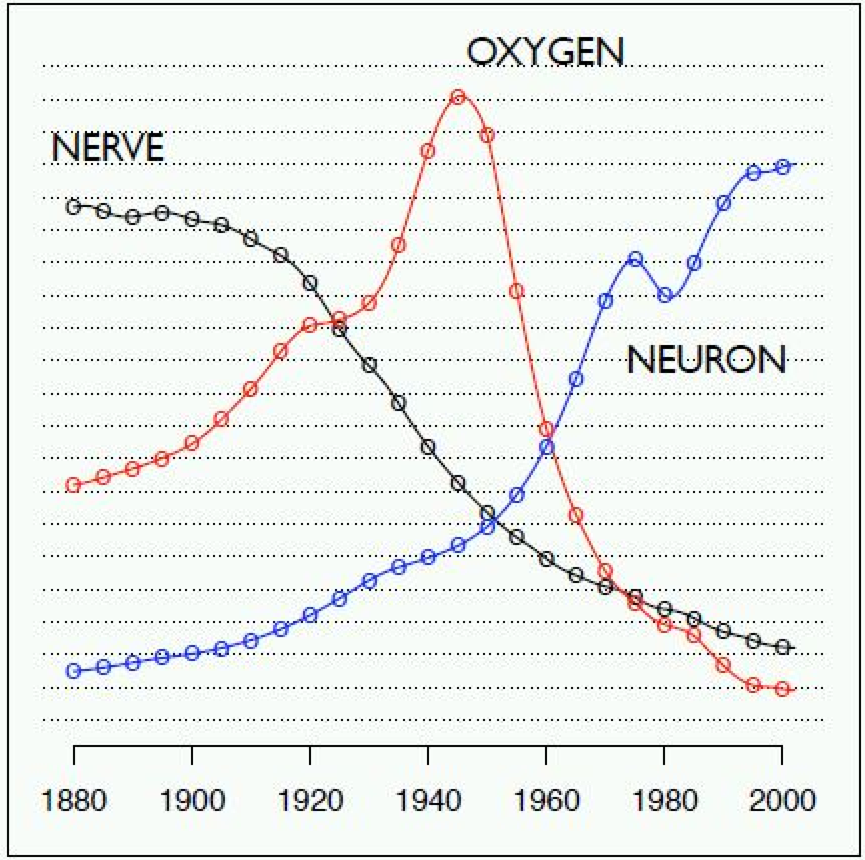
1. **Discover** the thematic structure
2. **Annotate** the documents
3. **Use** the annotations to visualize, organize, summarize, ...

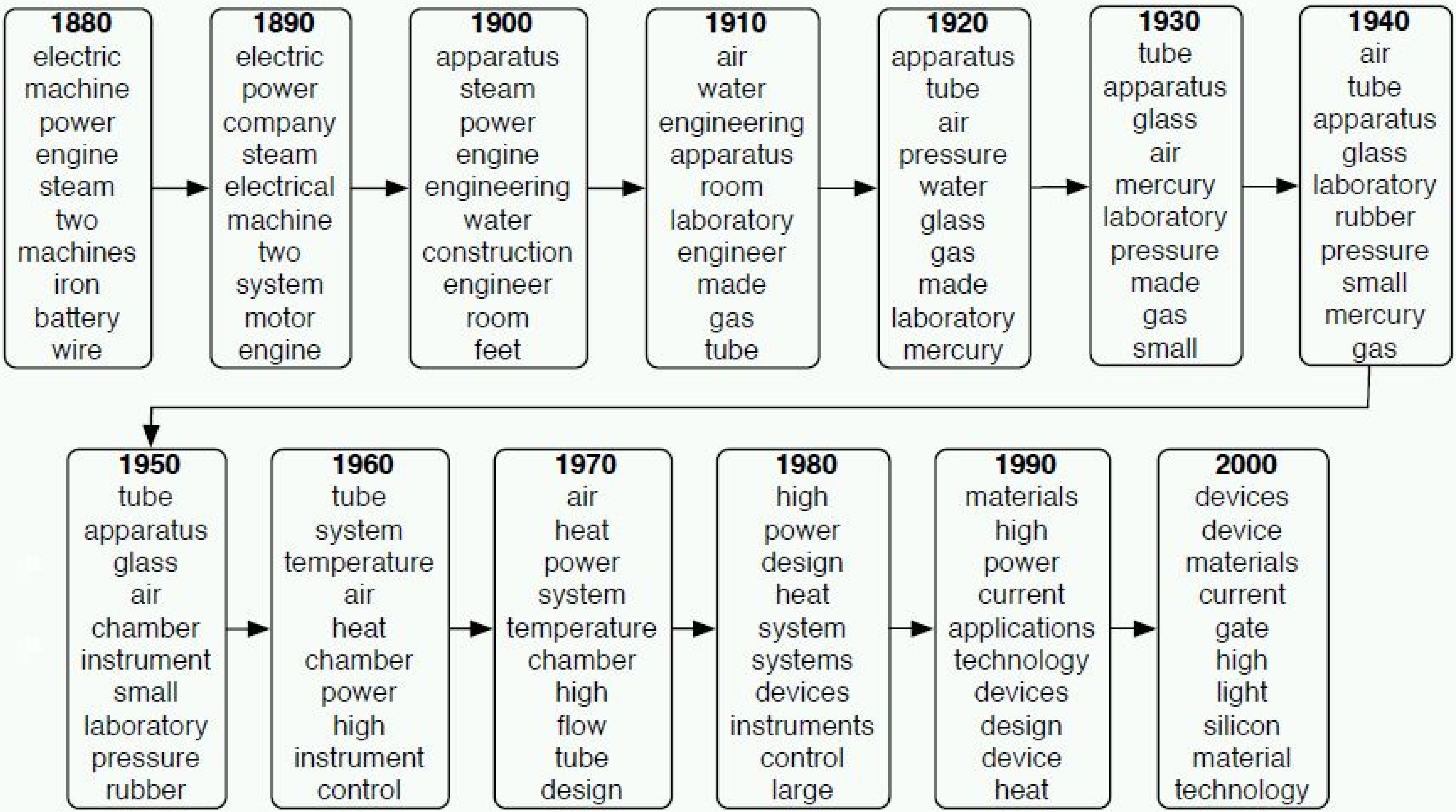

activated
tyrosine phosphorylation
activation
phosphorylation
kinase
p53
cell cycle
activity
cyclin
regulation
proteins
protein
binding
domain
domains
rna
dna
rna polymerase
cleavage
site
brain
memory
subjects
left
task
neurons
stimulus
motor
visual
cortical
synapses
ltp
glutamate
synaptic
neurons
surface
tip
image
sample
device
computer
problem
information
computers
problems
materials
organic
polymer
polymers
molecules
research
funding
support
nih
program
science
scientists
says
research
people
receptor
receptors
ligand
ligands
apoptosis
amino acids
cdna
sequence
isolated
protein
sequence
sequences
genome
dna
sequencing
surface
liquid
surfaces
fluid
model
laser
optical
light
electrons
quantum
physicists
particles
physics
particle
experiment
wild type
mutant
mutations
mutants
mutation
enzyme
enzymes
iron
active site
reduction
reaction
reactions
molecule
molecules
transition state
stars
astronomers
universe
galaxies
galaxy
united states
women
universities
students
education
cells
cell
expression
cell lines
bone marrow
magnetic
magnetic field
spin
superconductivity
superconducting
plants
plant
gene
genes
arabidopsis
pressure
high pressure
pressures
core
inner core
mantle
crust
upper mantle
meteorites
ratios
sun
solar wind
earth
planets
planet
bacteria
bacterial
host
resistance
parasite
development
embryos
drosophila
genes
expression
fossil record
birds
fossils
dinosaurs
fossil
species
forest
forests
populations
ecosystems
mice
antigen
t cells
antigens
immune response
virus
hiv
aids
infection
viruses
gene
disease
mutations
families
mutation
earthquake
earthquakes
fault
images
data
co2
carbon
carbon dioxide
methane
water
patients
disease
treatment
drugs
clinical
cells
proteins
researchers
protein
found
genetic
population
populations
differences
variation
ancient
found
impact
million years ago
africa
volcanic
deposits
magma
eruption
volcanism
climate
ocean
ice
changes
climate change
ozone
atmospheric
measurements
stratosphere
concentrations

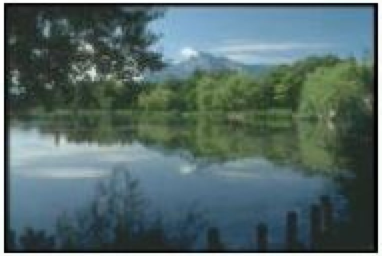
"Theoretical Physics"



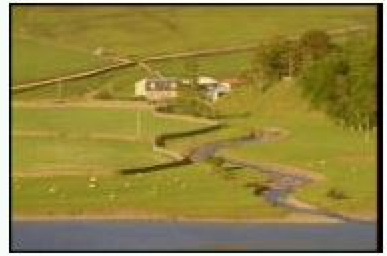
"Neuroscience"




1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire
1890
electric
power
company
steam
electrical
machine
two
system
motor
engine
1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet
1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube
1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury
1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small
1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas
1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber
1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control
1970
air
heat
power
system
temperature
chamber
high
flow
tube
design
1980
high
power
design
heat
system
systems
devices
instruments
control
large
1990
materials
high
power
current
applications
technology
devices
design
device
heat
2000
devices
device
materials
current
gate
high
light
silicon
material
technology



SKY WATER TREE
MOUNTAIN PEOPLE



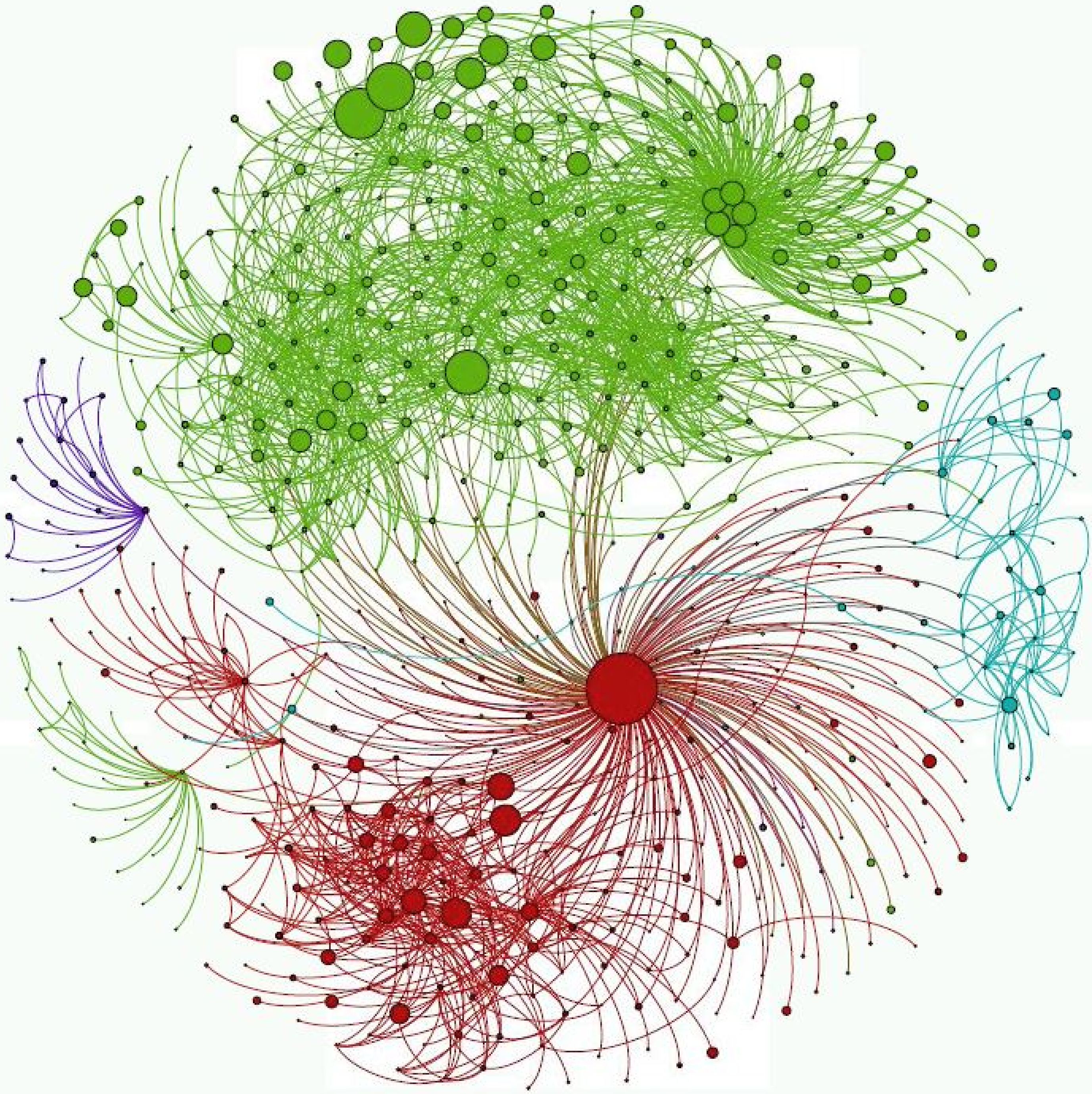
SCOTLAND WATER
FLOWER HILLS TREE

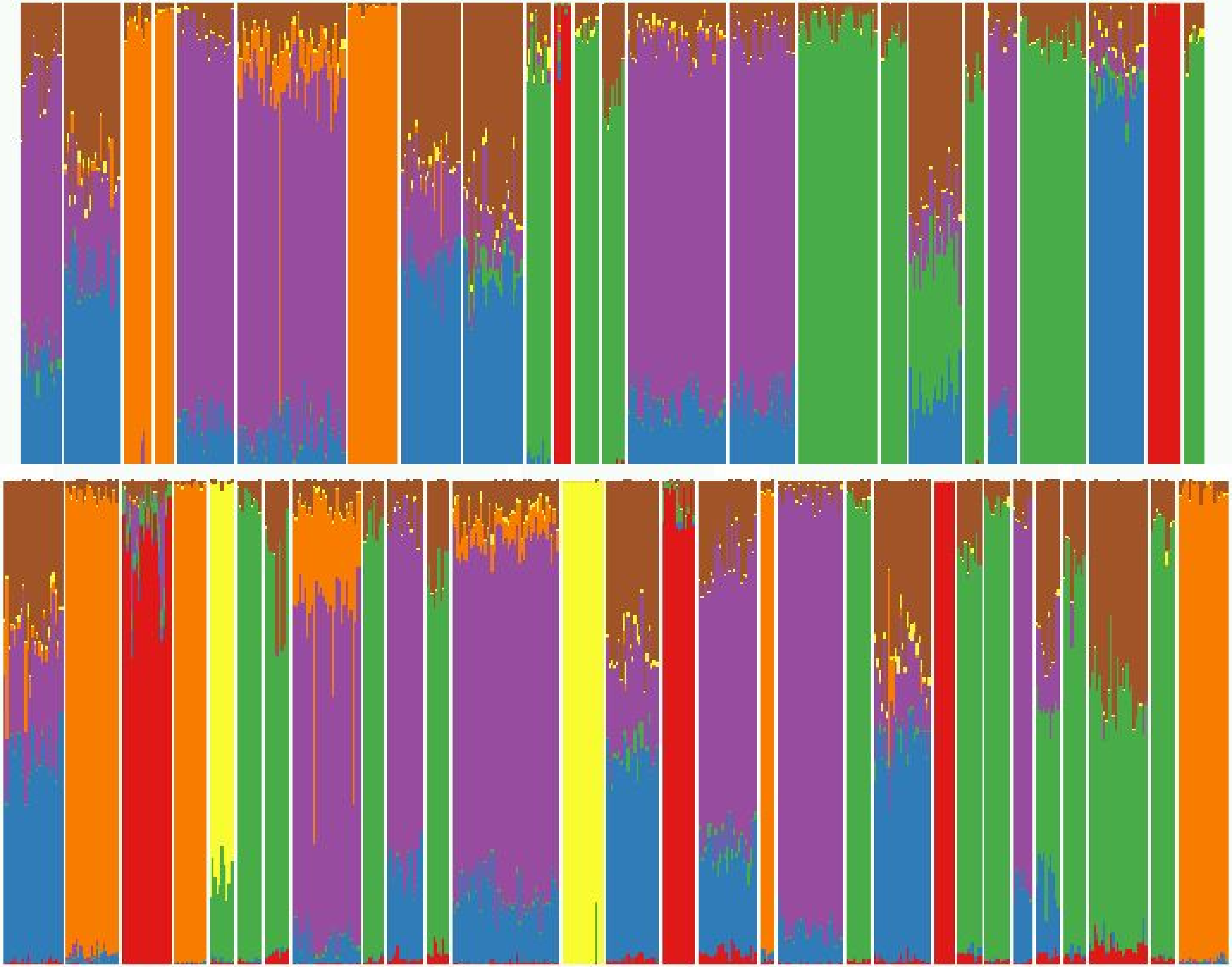
SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES





Charles Darwin's library

The NYC subway

- **People read documents.**
- These might be people for whom we want to form predictions.
- And, their behavior is an additional signal about the meaning of the documents and the organization of the collection.

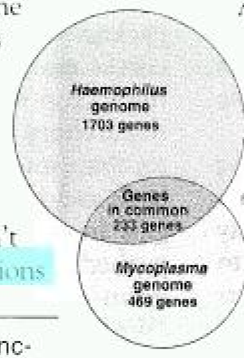
This talk

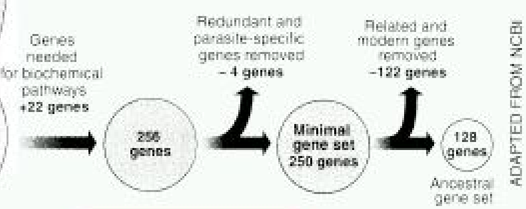
1. Introduction to topic modeling
2. Recommendation and exploration with collaborative topic models
3. The bigger picture: Using probability models to solve problems with data

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



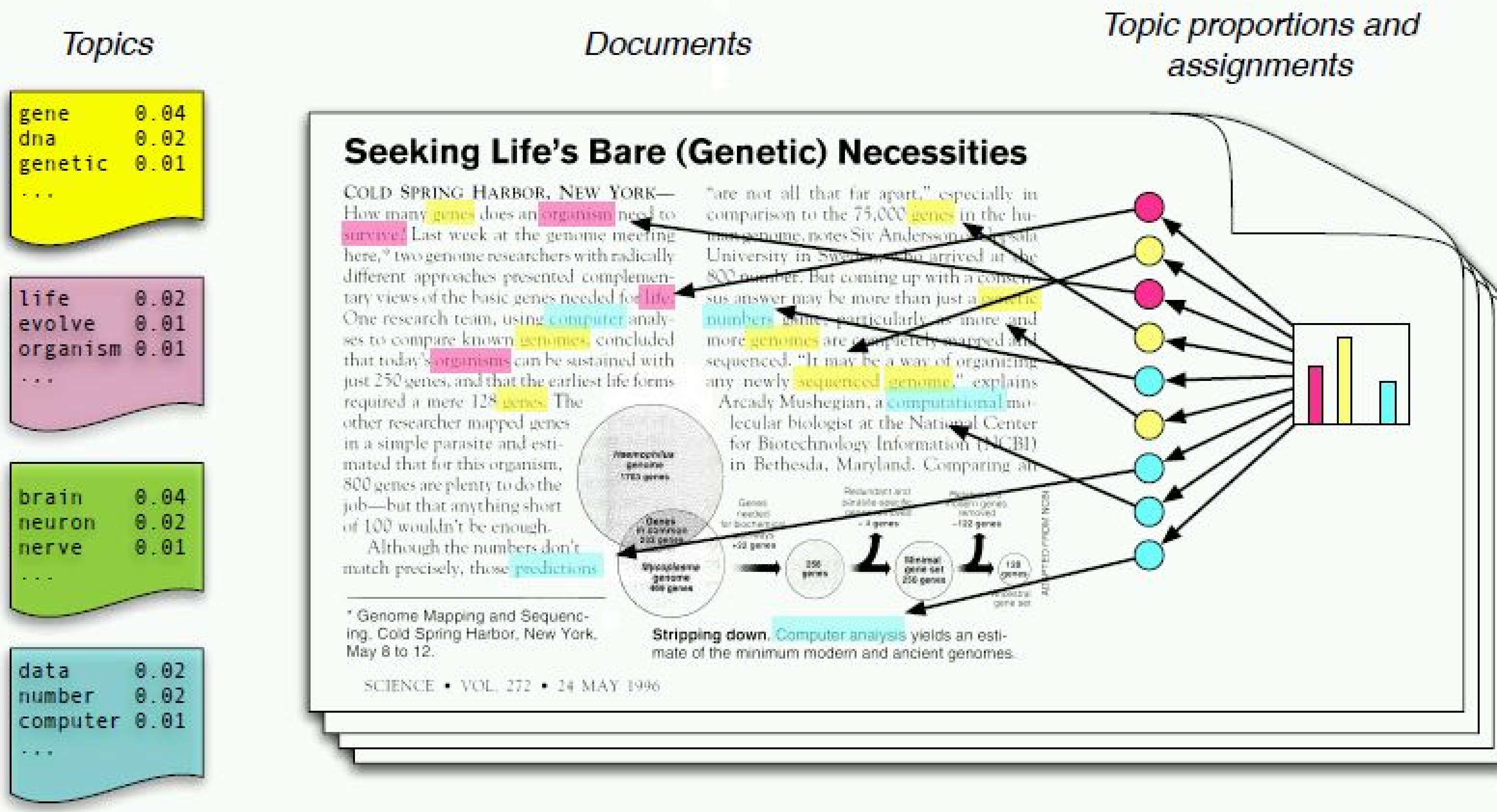


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

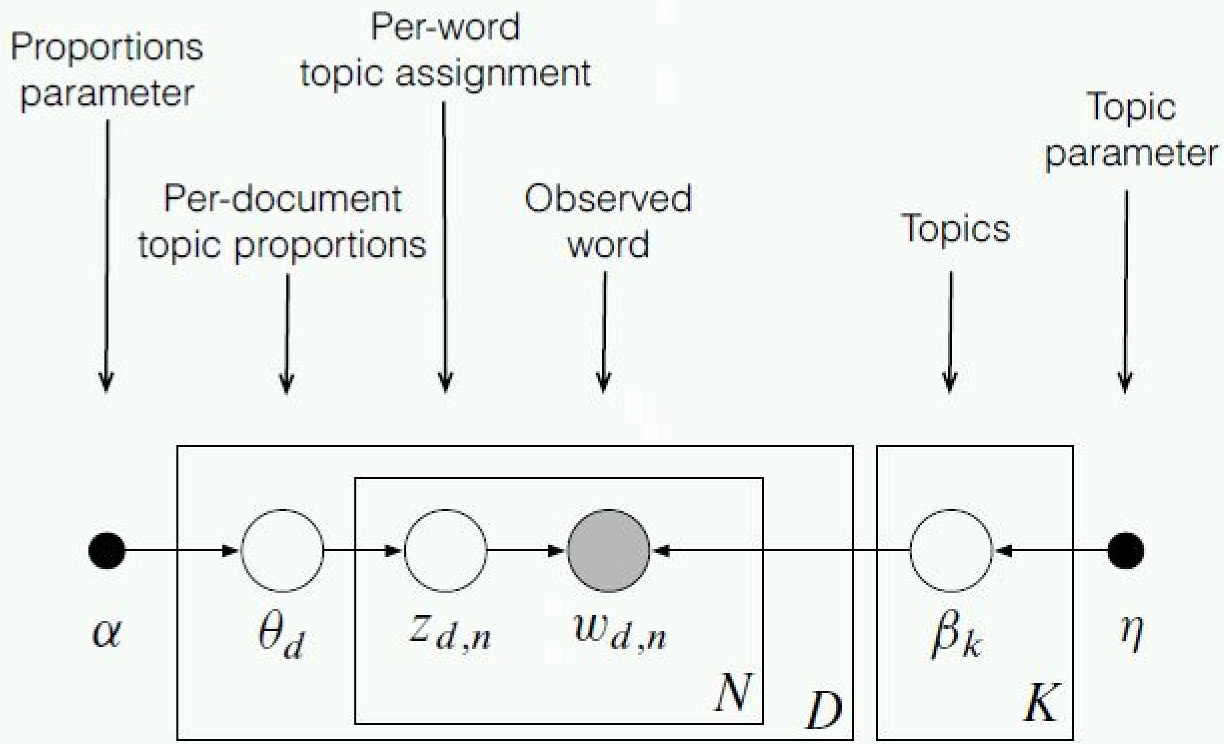
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Documents exhibit multiple topics.

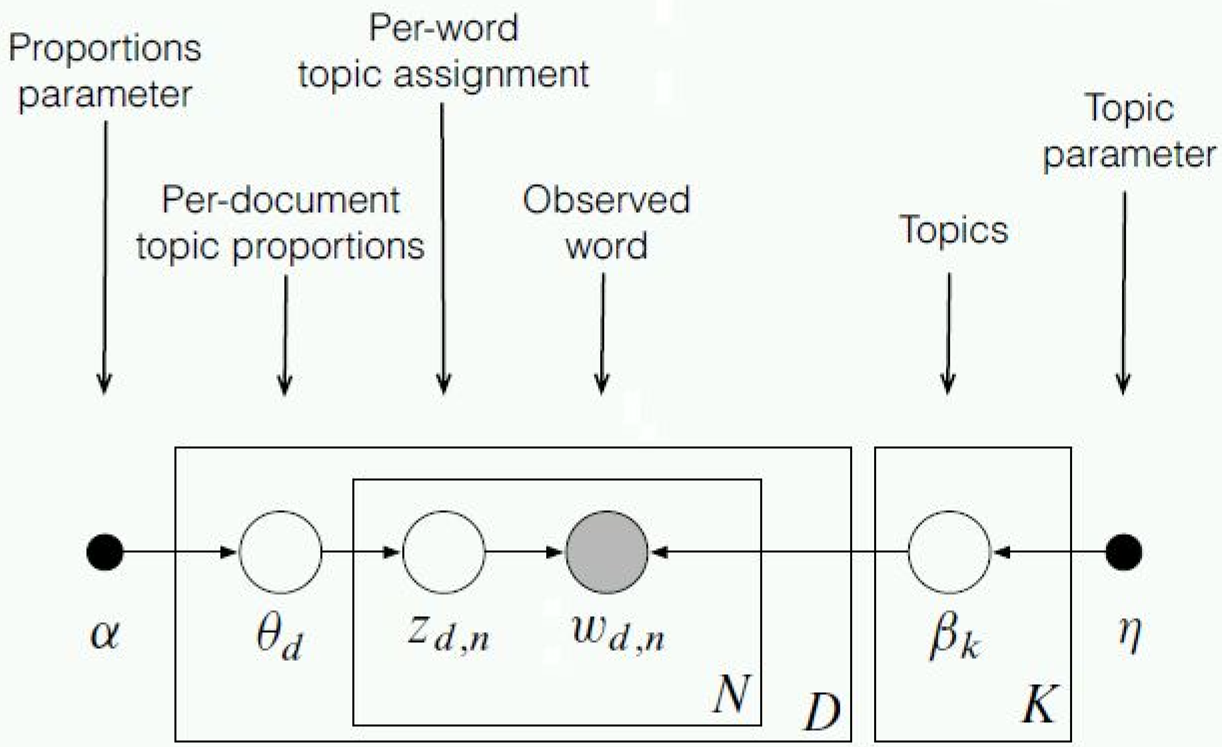


Latent Dirichlet Allocation



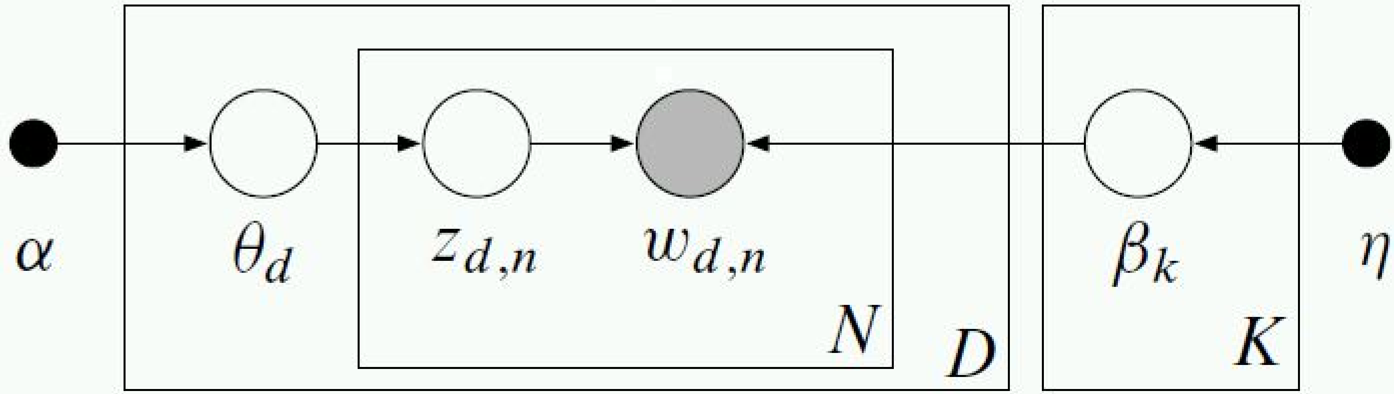
LDA as a graphical model

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

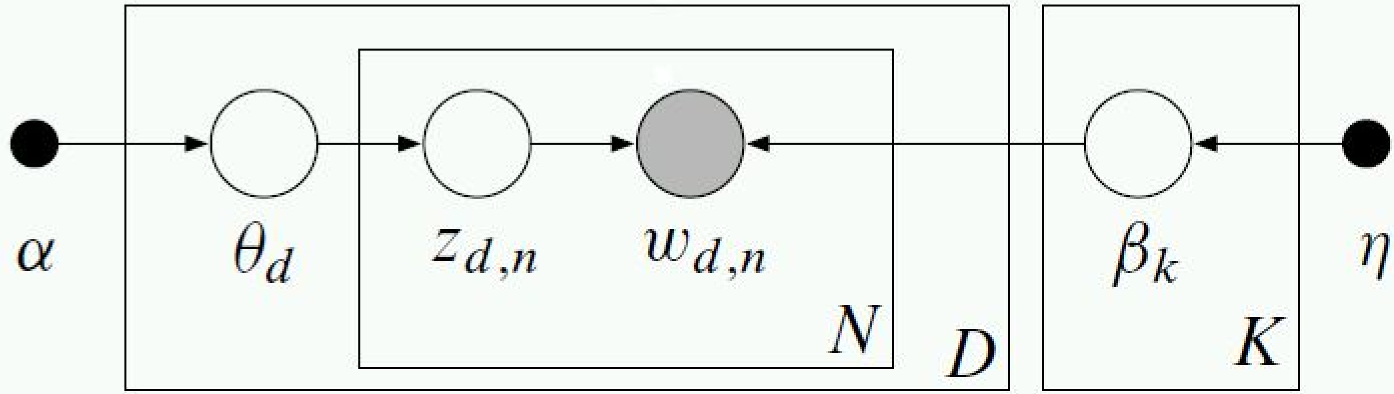


LDA as a graphical model

- Encodes independence assumptions about the variables
- Defines a factorization of the joint probability distribution
- Connects to algorithms for computing with data

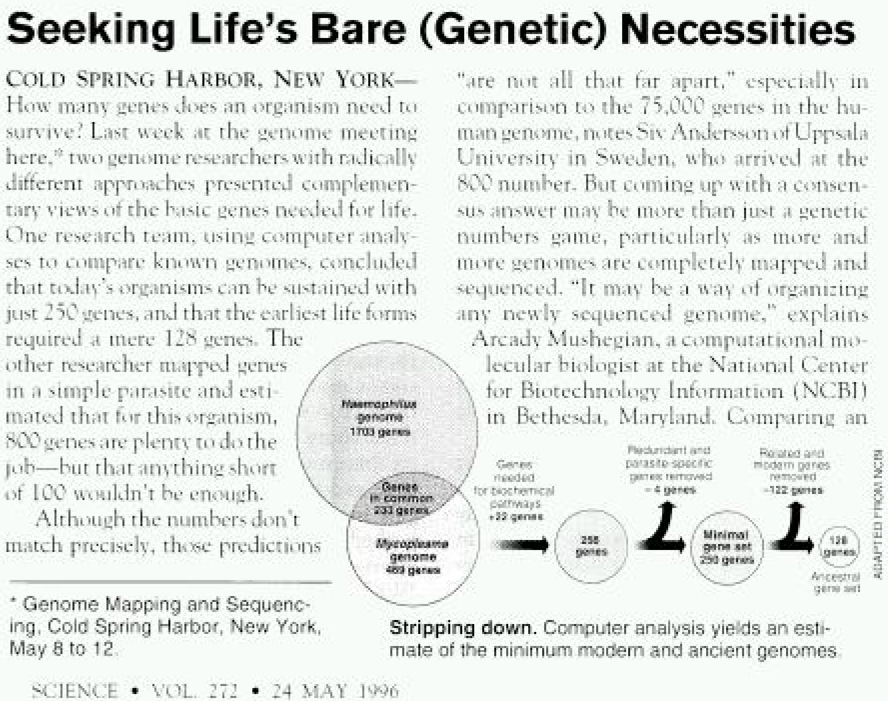


- The joint defines a posterior, $p(\theta, z, \beta \mid w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand: information retrieval,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



- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
- Collapsed variational inference (Teh et al., 2006)
- Stochastic inference (Hoffman et al., 2010, 2013; Mimno et al., 2012)
- Factorization inference (Arora et al., 2012; Anandkumar et al., 2012)

- **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

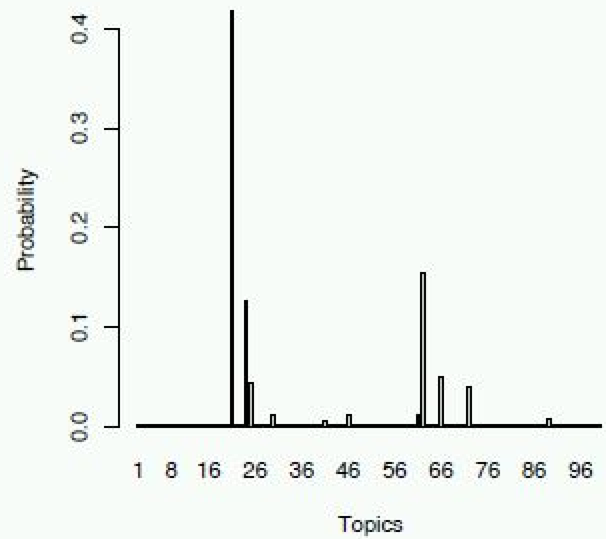
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

* 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.

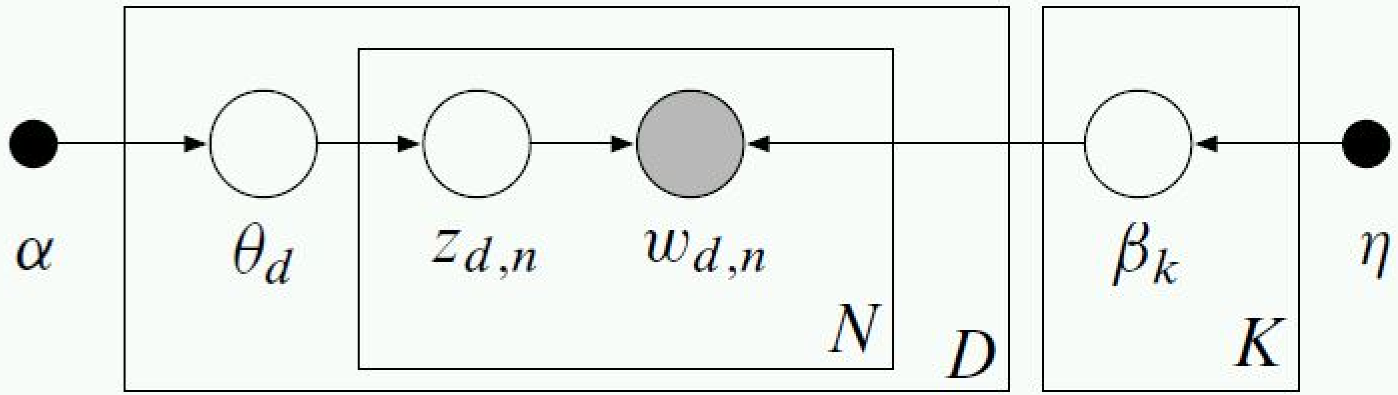
SCIENCE • VOL. 272 • 24 MAY 1996



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

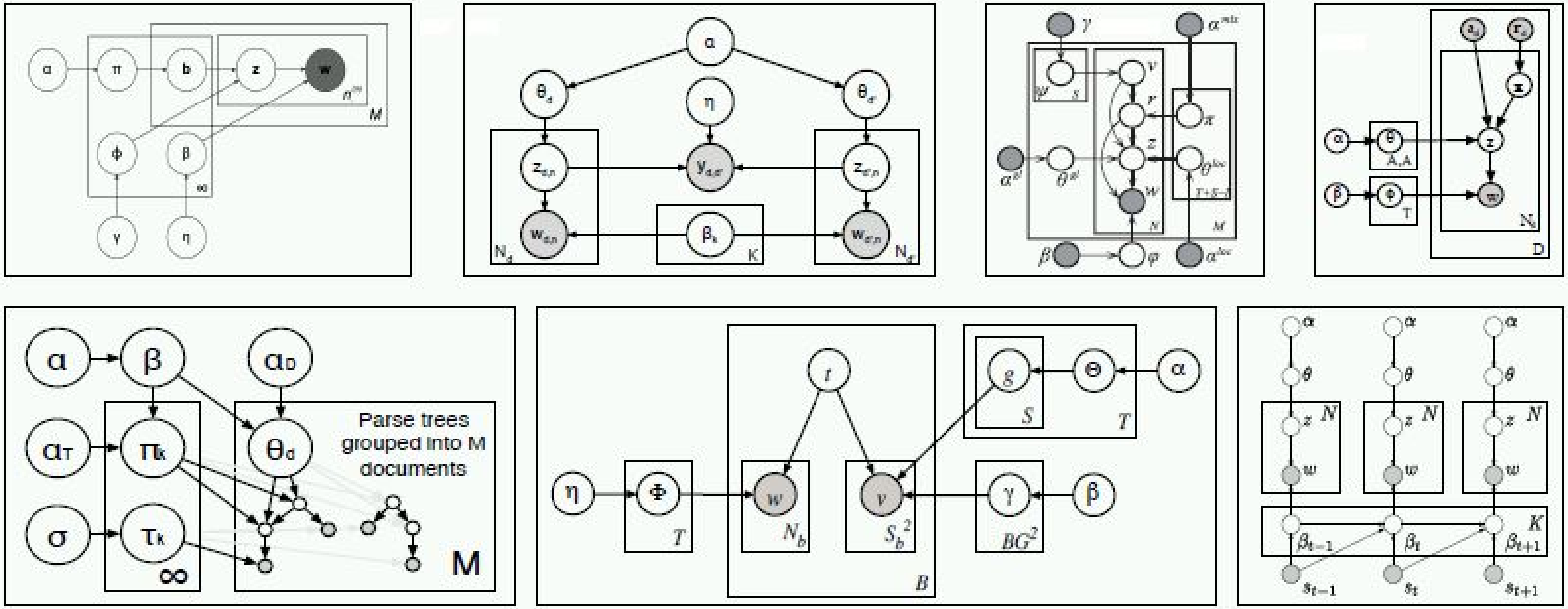
6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent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



- Summary: LDA discovers themes through posterior inference.

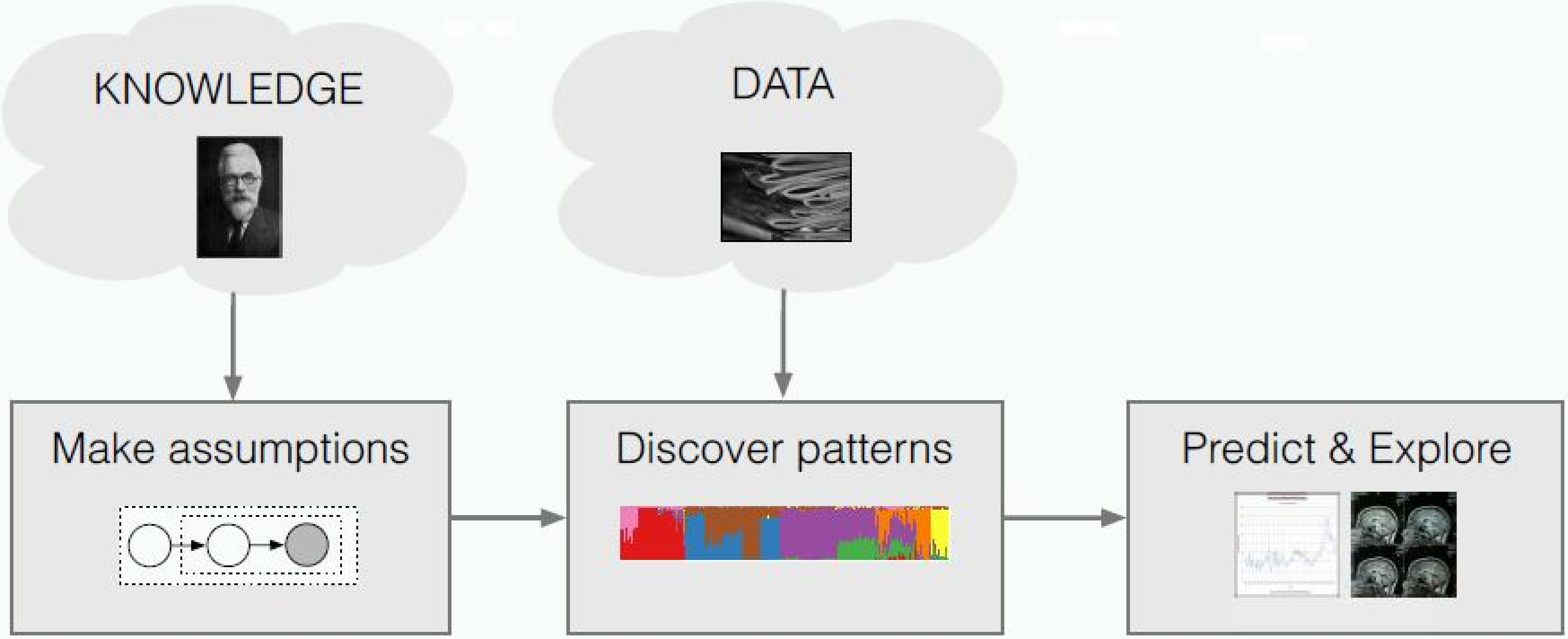
- Other perspectives
  - Latent semantic analysis [Deerwester et al., 1990; Hofmann, 1999]
  - A mixed-membership model [Erosheva, 2004]
  - PCA and matrix factorization [Jakulin and Buntine, 2002]
  - Was independently invented for genetics [Pritchard et al., 2000]



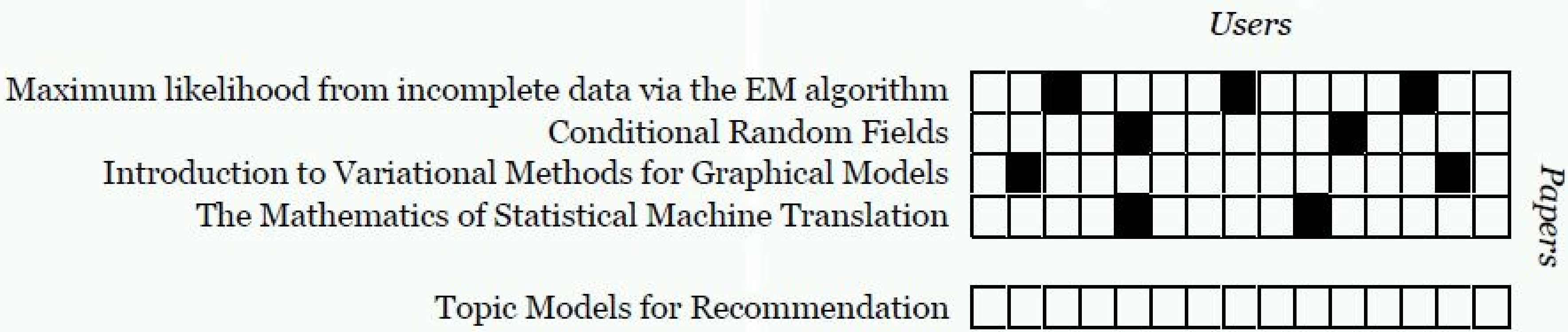
- Organizing and finding patterns in text is important in the sciences, humanities, industry, and culture.

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

- Algorithmic improvements let us fit models to massive data. (See VW, Gensim, Mallet, others.)

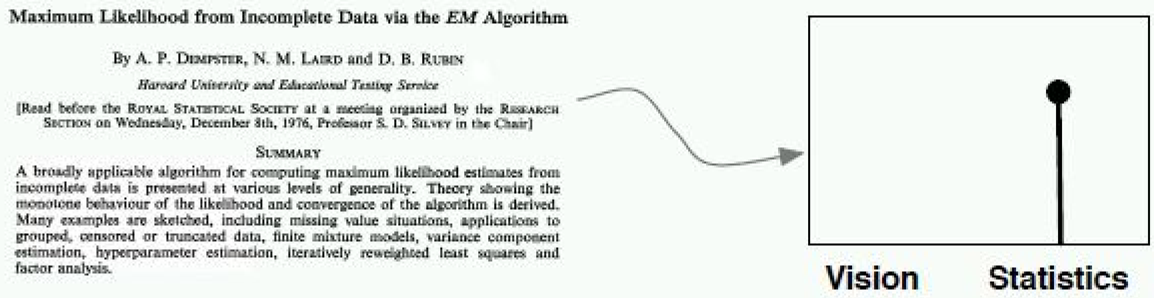


- Case study in **text analysis with probability models**
- Topic modeling research
  - develops new models.
  - develops new inference algorithms.
  - develops new applications, visualizations, tools.



- Example: Scientists share their research libraries.
- Collaborative topic models can
  - Helps readers discover documents, old and new.
  - Describe readers in terms of topical preferences
  - Identify documents that are impactful, interdisciplinary

- Consider EM (Dempster et al., 1977). We infer topics from its text:


Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. Dempster, N. M. Laird and D. B. Rubin

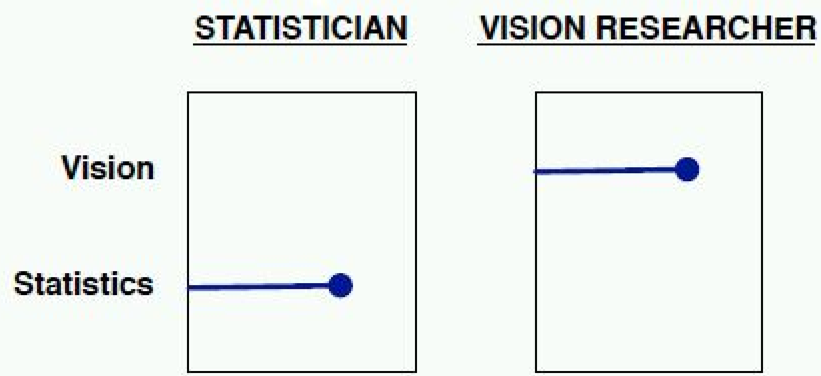
Harvard University and Educational Testing Service

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

Summary

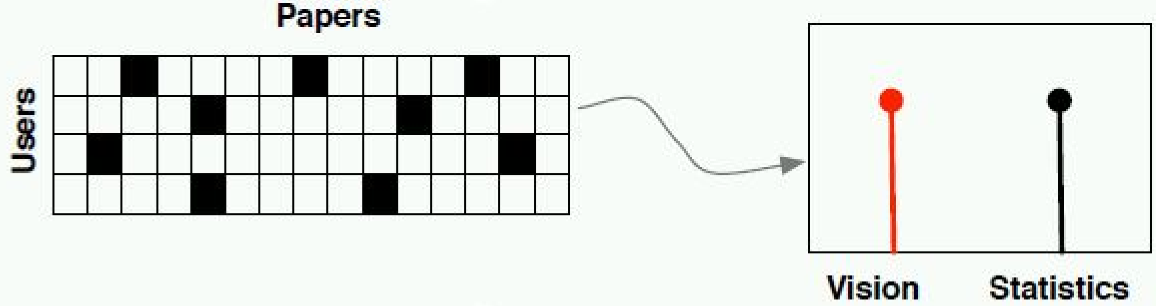
A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

- Suppose there are two types of scientists

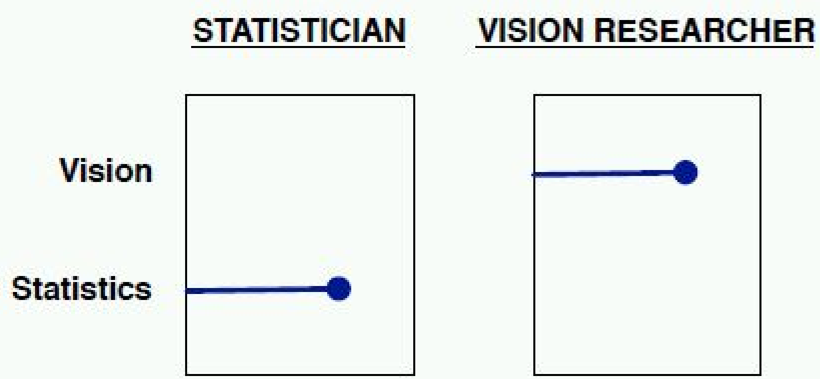


- We first recommend the EM paper to **statisticians**.

- With user data, we can adjust the topics to account for who liked it:



- Consider again the scientists



- We now recommend the EM paper to **vision researchers**.

AIG GRC Archer | AIG Asset Management Int... | Tools & Applications (Shar... | ServiceNow | em algorithm - Google Sea...

https://itservices.aig.net/com.glideapp.servicecatalog_checkout_view.do?v=1&sysparm_sys_id=aa1d2619c9886200f272cee7ef18299e&sysparm_new_request=true&sysparm_view=ess&sysparm_catalog=e0d08b13c3

Search

Order Status

Summary

Your request number is **REQ003423639**, which you can use to refer to this request in future interactions with the service desk.

You may also bookmark the following link to get back to REQ003423639.

Note that clicking on the bookmark link (above) will simply take you back to this screen.

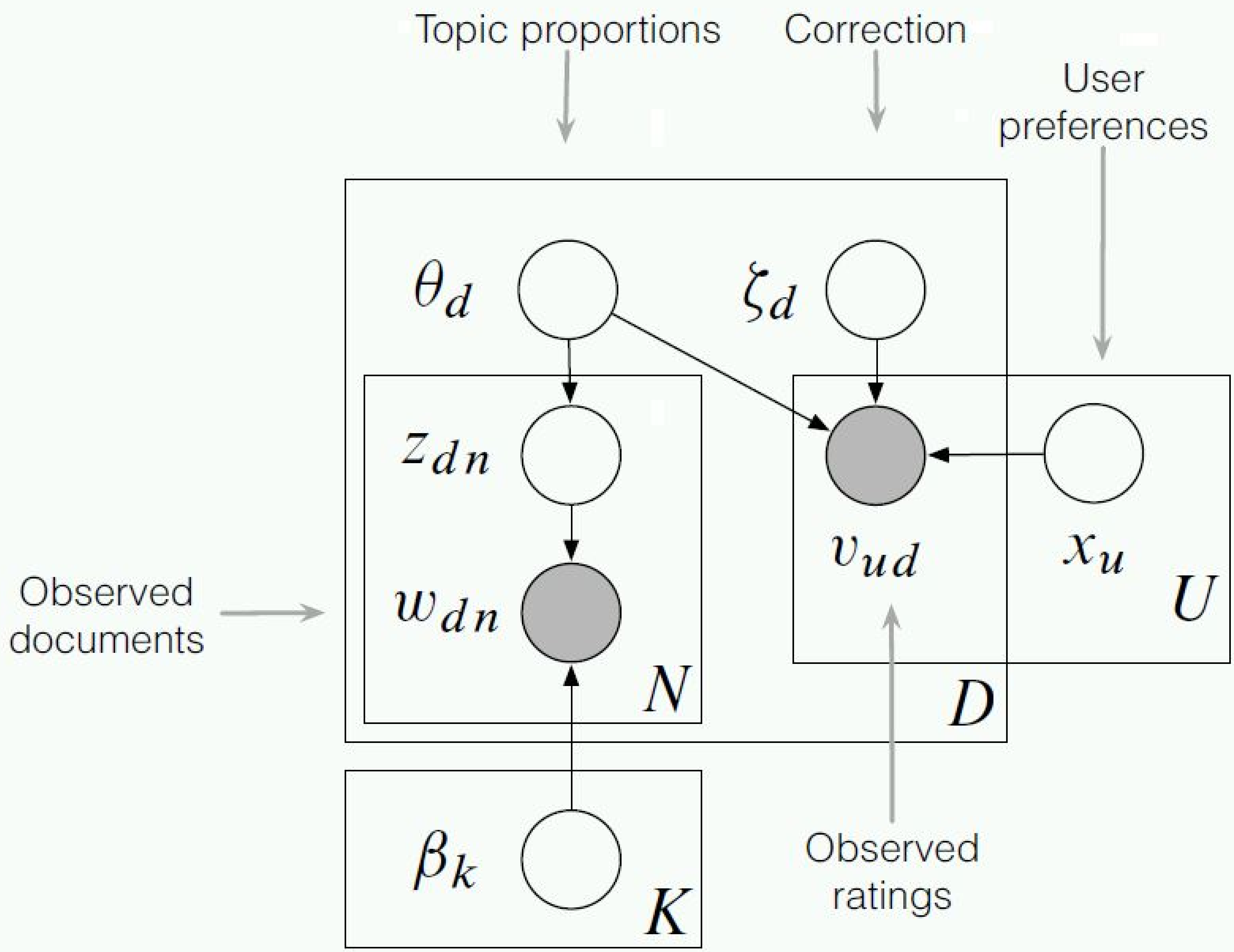
Description	Delivery Date	Stage	Price (ea.)	Qty	Total
Airwatch: New Account/Good Migration	2016-06-20		$0.00	1	$0.00
				Total:	**$0.00**

Delivery Information

Estimated Delivery Date of Complete Order: **2016-06-20**

Catalog

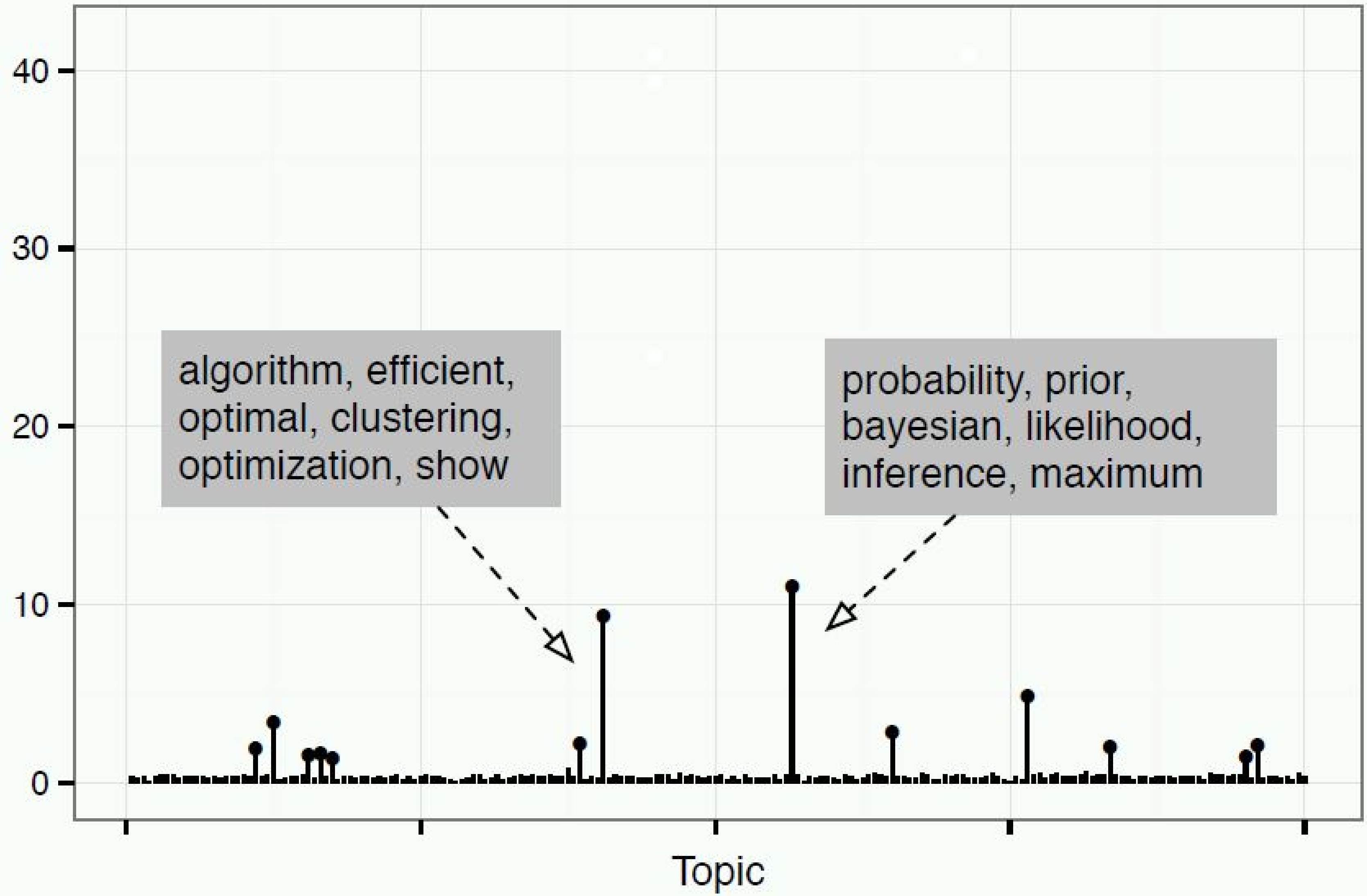
Home


Topic proportions
Correction
User preferences
θ_d
ζ_d
z_{dn}
w_{dn}
v_{ud}
x_u
U
N
D
β_k
K
Observed documents
Observed ratings

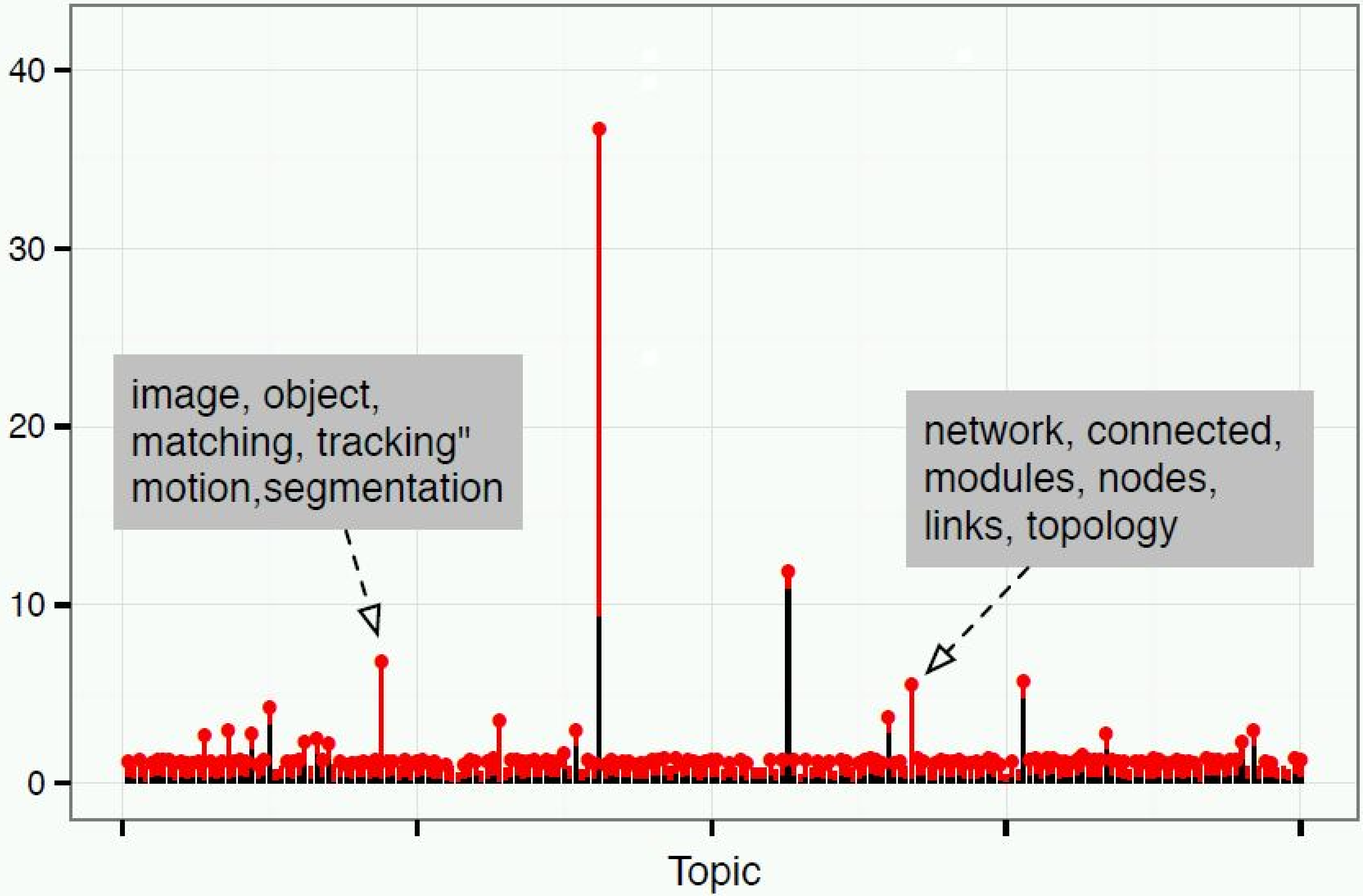
MENDELEY

- Big data set from Mendeley.com

- The data:
  - 261K documents
  - 80K users
  - 10K vocabulary terms
  - 25M observed words
  - 5.1M entries (sparsity is 0.02%)

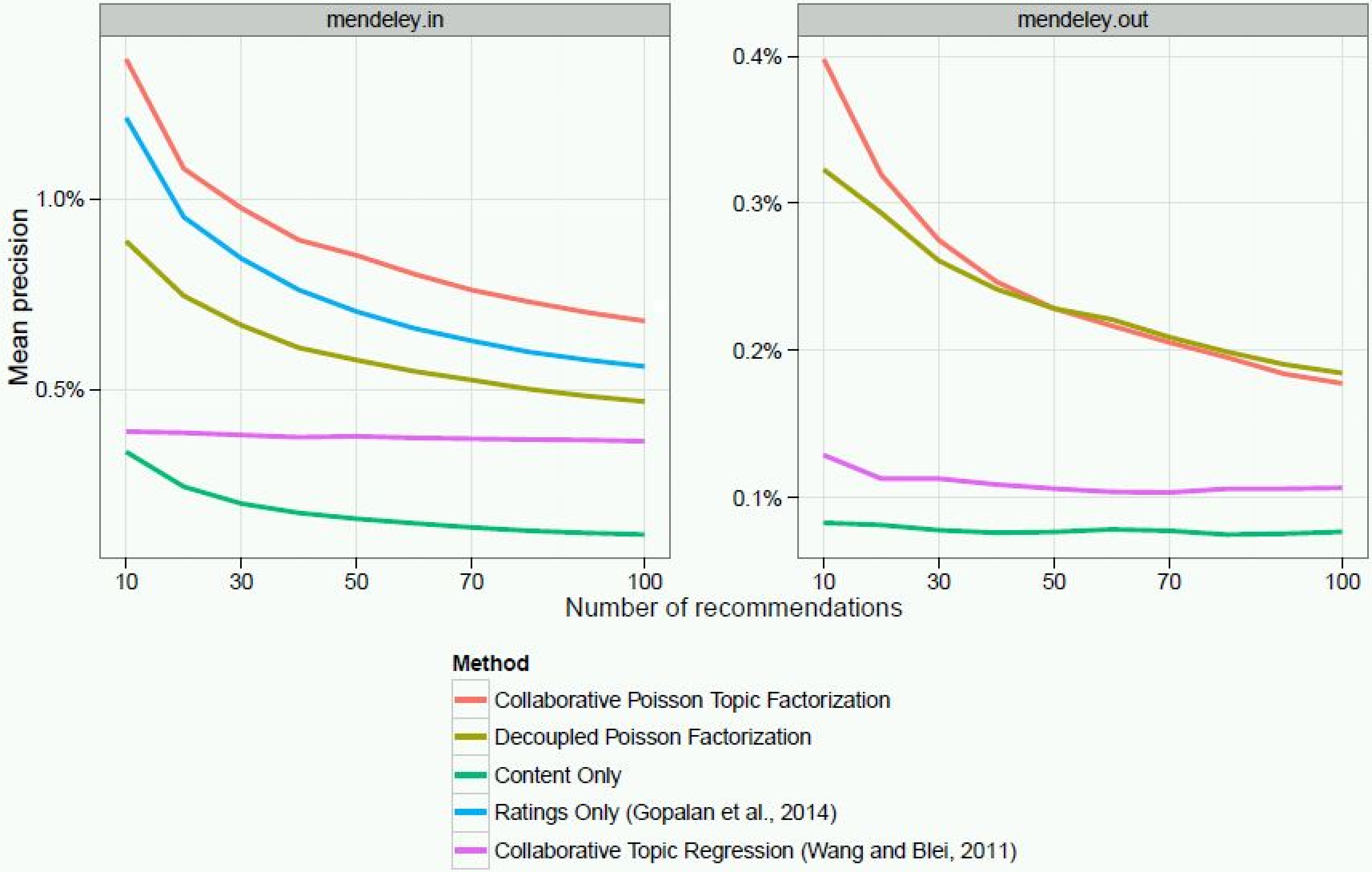


40
30
20
10
0
algorithm, efficient, optimal, clustering, optimization, show
probability, prior, bayesian, likelihood, inference, maximum
Topic



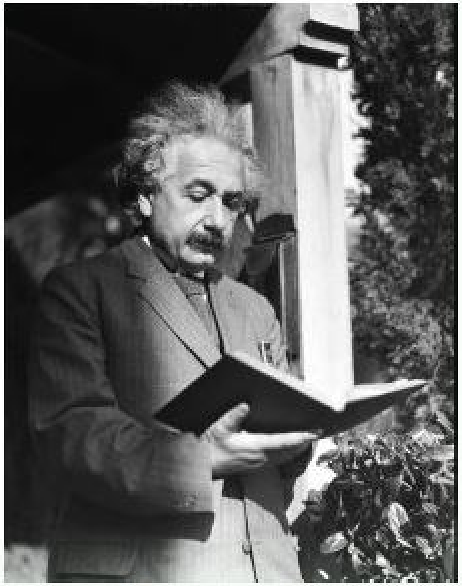
40
30
20
10
0
image, object, matching, tracking" motion,segmentation
network, connected, modules, nodes, links, topology
Topic

Mendeley



mendeley.in
mendeley.out
Mean precision
1.0%
0.5%
0.4%
0.3%
0.2%
0.1%
10
30
50
70
100
Number of recommendations
Method
Collaborative Poisson Topic Factorization
Decoupled Poisson Factorization
Content Only
Ratings Only (Gopalan et al., 2014)
Collaborative Topic Regression (Wang and Blei, 2011)

Darwin's library

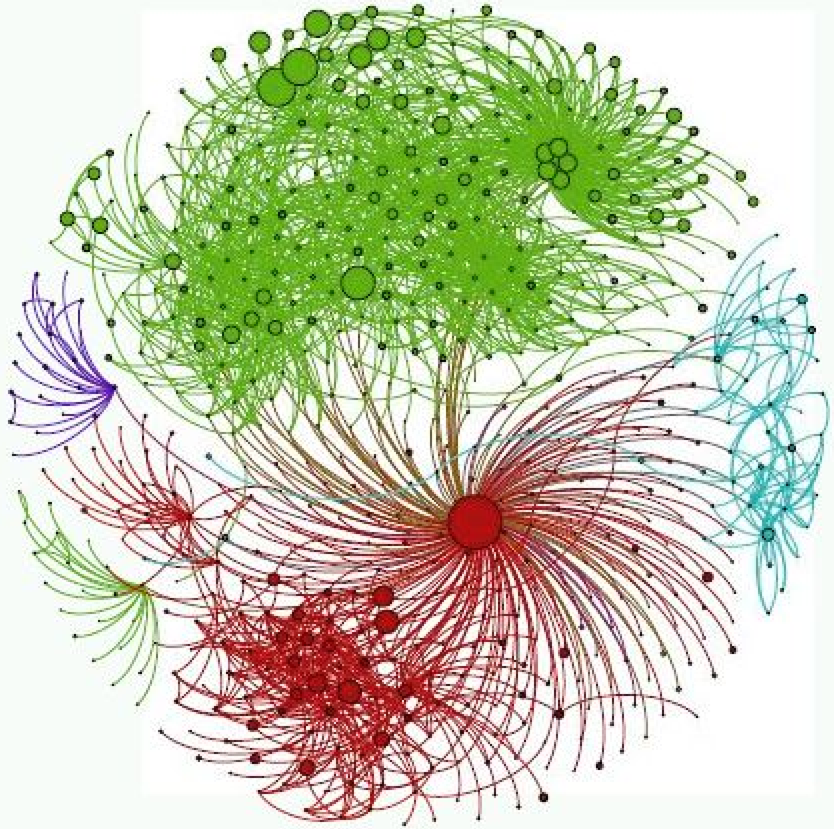


Einstein reading

Another scientist reading

- The readers also **tell us about the articles**.
- We can look at posterior estimates to find
  - Interdisciplinary articles
  - Influential articles within a field
  - Outside influences on a field

“Network Analysis”



network; connected; modules; nodes; links; topology; connectivity; graph; robustness; connections; modular; world; degree; properties

Assortative mixing in networks

M. E. J. Newman
Department of Physics, University of Michigan, Ann Arbor, MI 48109–1120 and
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

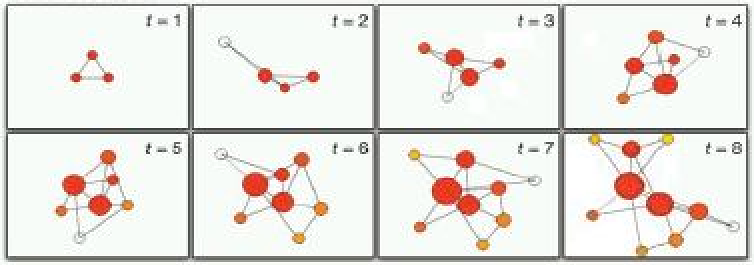
About networks

- Assortative mixing in networks

  (Newman, 2002)
- Mixing patterns in networks

  (Newman, 2002)
- Catastrophic cascade of failures in interdependent networks

  (Buldyrev et al., 2010)



About networks; for readers of networks

- Emergence of scaling in random networks
  (Barabassi and Albert, 1999)
- Statistical mechanics of complex networks
  (Albert and Barabassi, 2002)
- Complex networks: Structure and dynamics
  (Boccaletti et al., 2006)

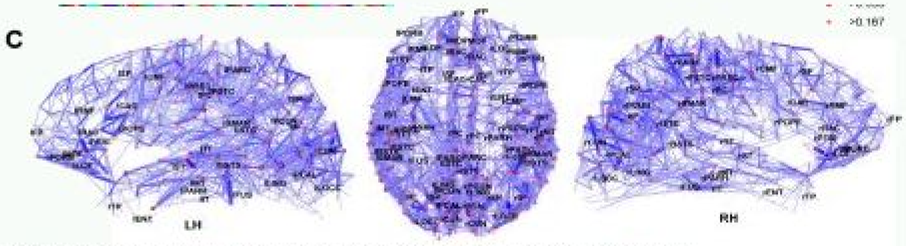
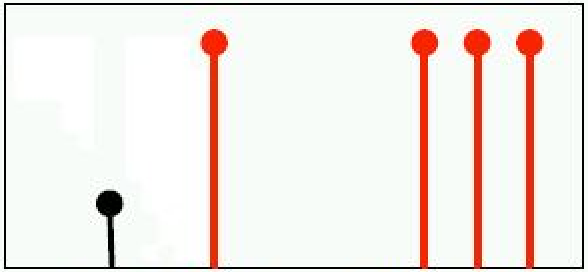
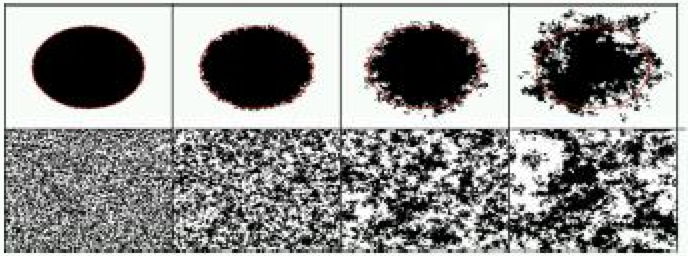


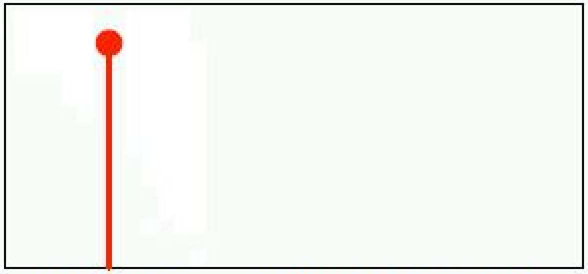
Figure 3. High-Resolution Connection Matrix, Network Layout and Connectivity Backbone (Participant A, scan 2)



About networks; for readers of other fields

- Mapping the Structural Core of Human Cerebral Cortex
  (Hagmann et al., 2008)
- Network thinking in ecology and evolution
  (Proulx et al., 2005)
- Linked: The New Science of Networks
  (Barabasi, 2002)





Not about networks; for readers of networks

- Power-law distributions in empirical data
  (Clauset et al., 2009)
- Statistical physics of social dynamics
  (Castellano et al., 2009)
- The origin of bursts and heavy tails in human dynamics
  (Barabasi, 2005)

“Statistical Modeling”

About this field; read by users in this field

- A Bayesian analysis of some nonparametric problems
- Bayesian measures of model complexity and fit
- Monte Carlo Methods in Bayesian Computation

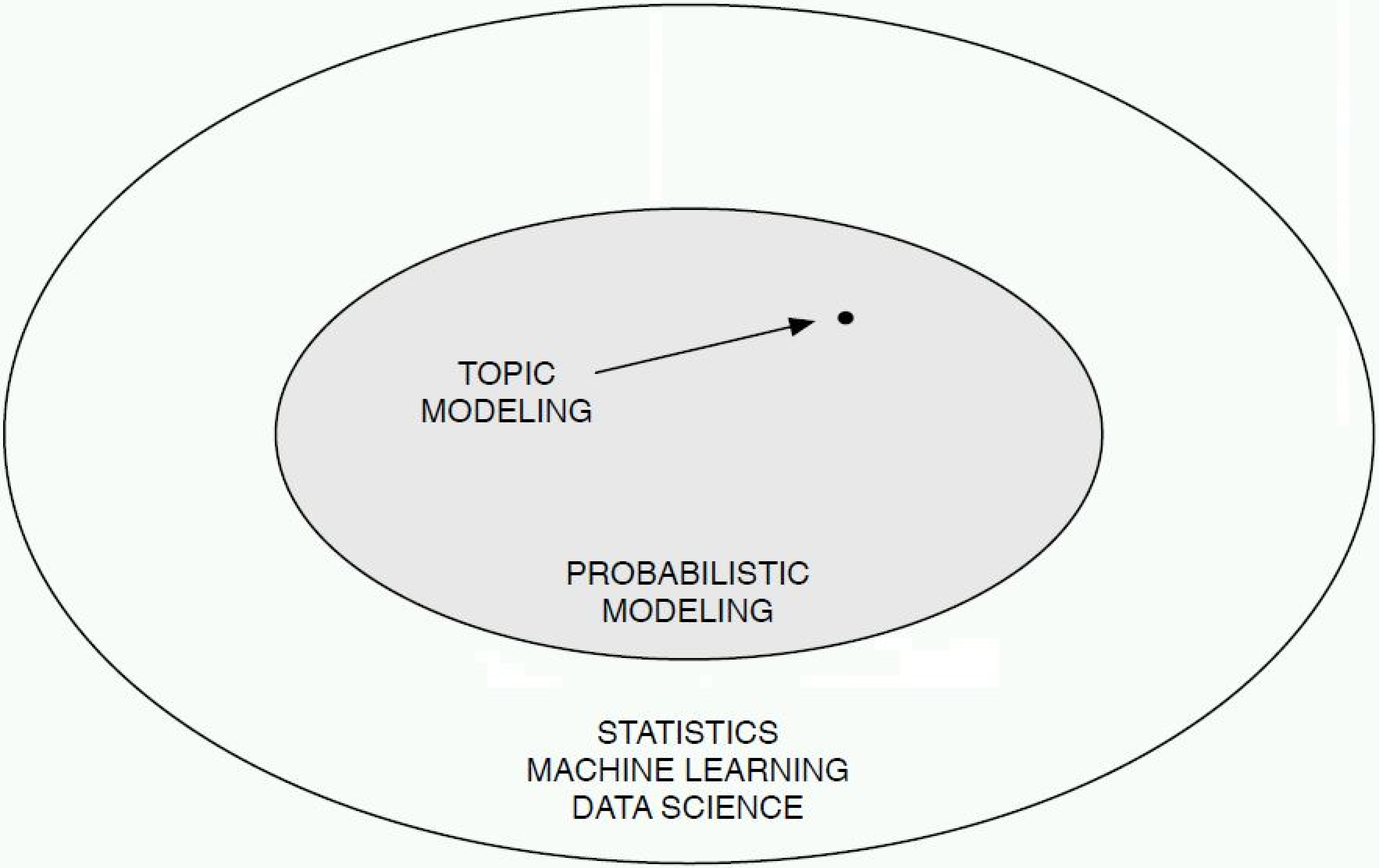
About this field; read by users in other fields

- A tutorial on HMMs and selected applications in speech recognition
- An Introduction to Bayesian Networks and Influence Diagrams
- Maximum likelihood from incomplete data via the EM algorithm

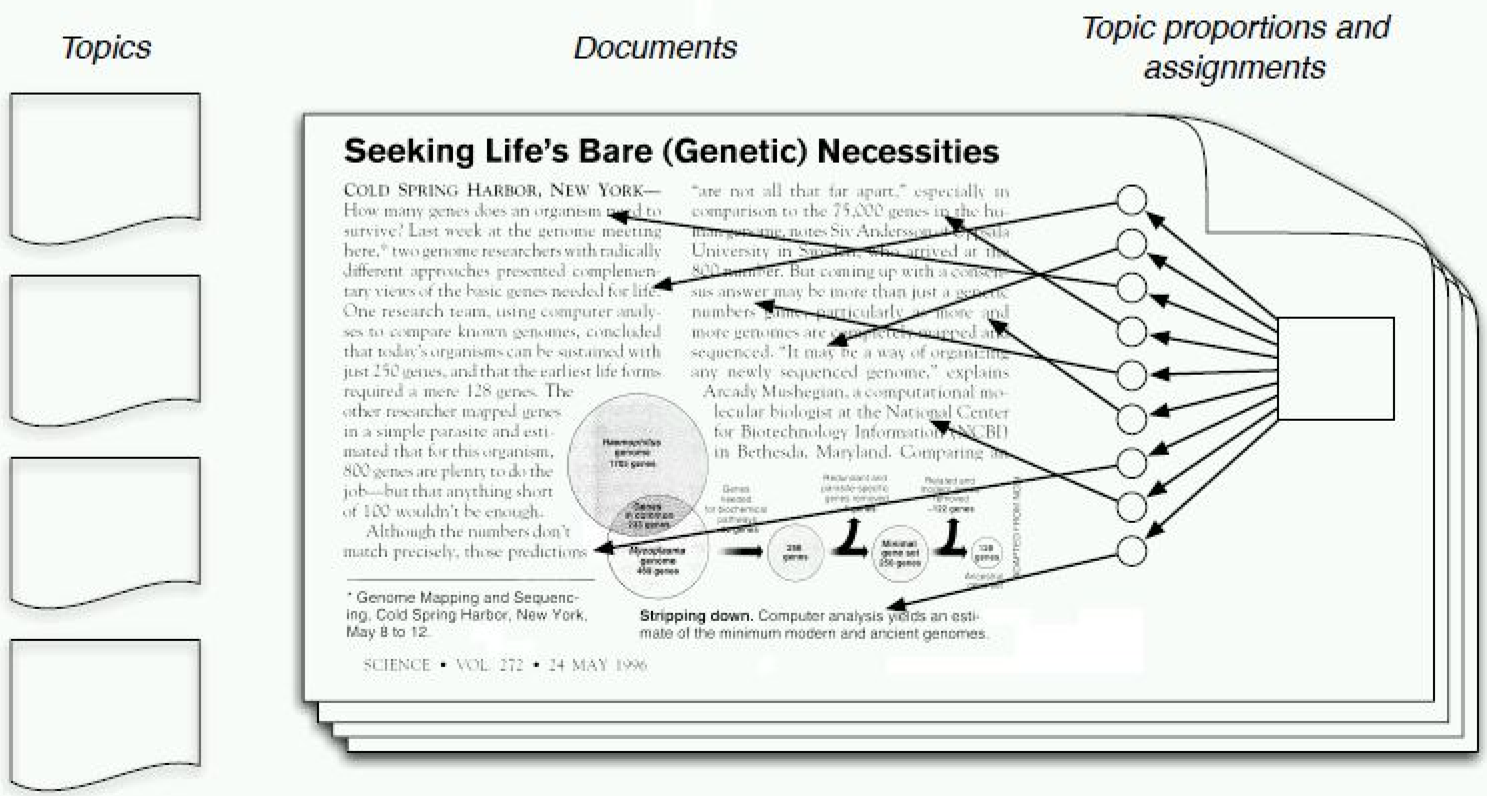
About other fields; read by users in this field

- Second Thoughts on the Bootstrap
- A guide to Eclipse and the R plug-in StatET
- Using Multivariate Statistics

- A decade of clicks on arXiv.org (2003–2013)
- The data:
  - 826K documents
  - 120K users
  - 14K vocabulary terms
  - 54M observed words
  - 43.6M entries (sparsity is 0.04%)


TOPIC
MODELING
PROBABILISTIC
MODELING
STATISTICS
MACHINE LEARNING
DATA SCIENCE

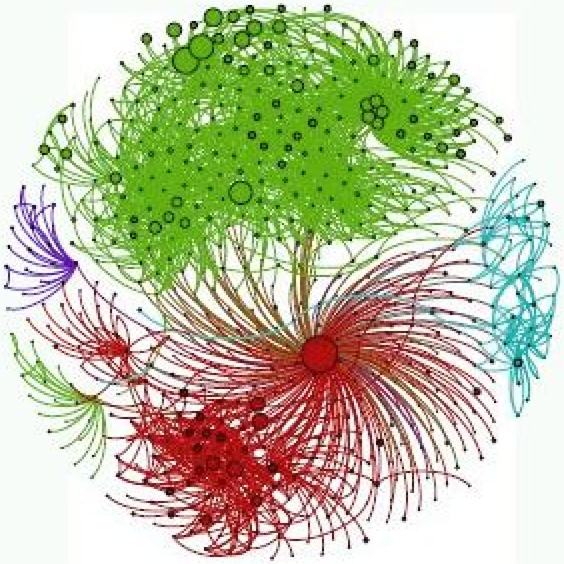
I. Assume our data come from a model with hidden patterns at work

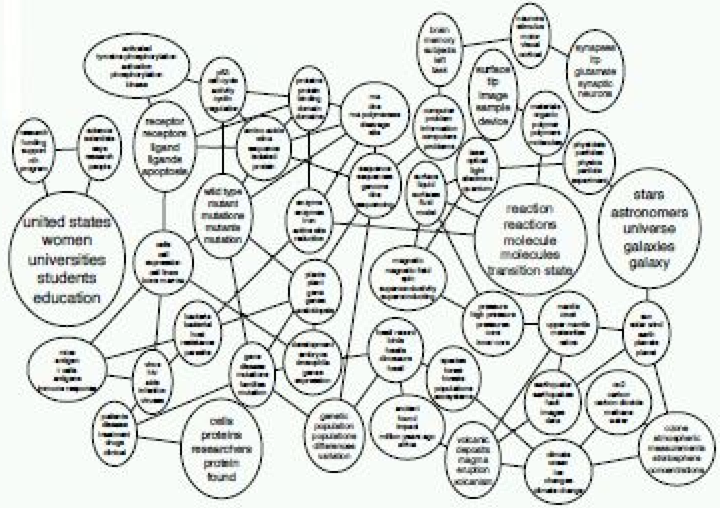


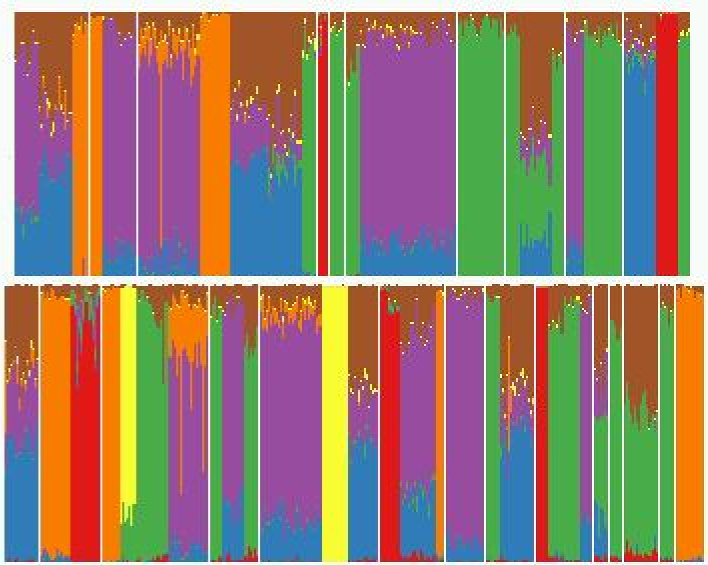
II. Discover those patterns from data

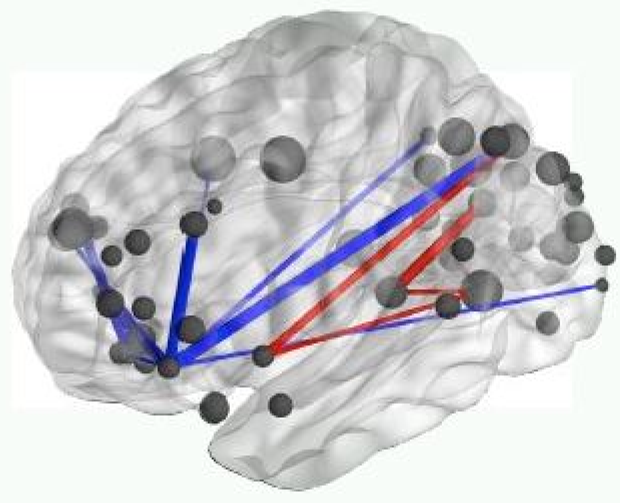
$$\boxed{\nu^* = \arg\max_{\nu} \mathbb{E}_q \left[\log p(x, z, \beta \mid \alpha)\right] + \mathbb{H}\left[q(z, \beta \mid \nu)\right]}$$

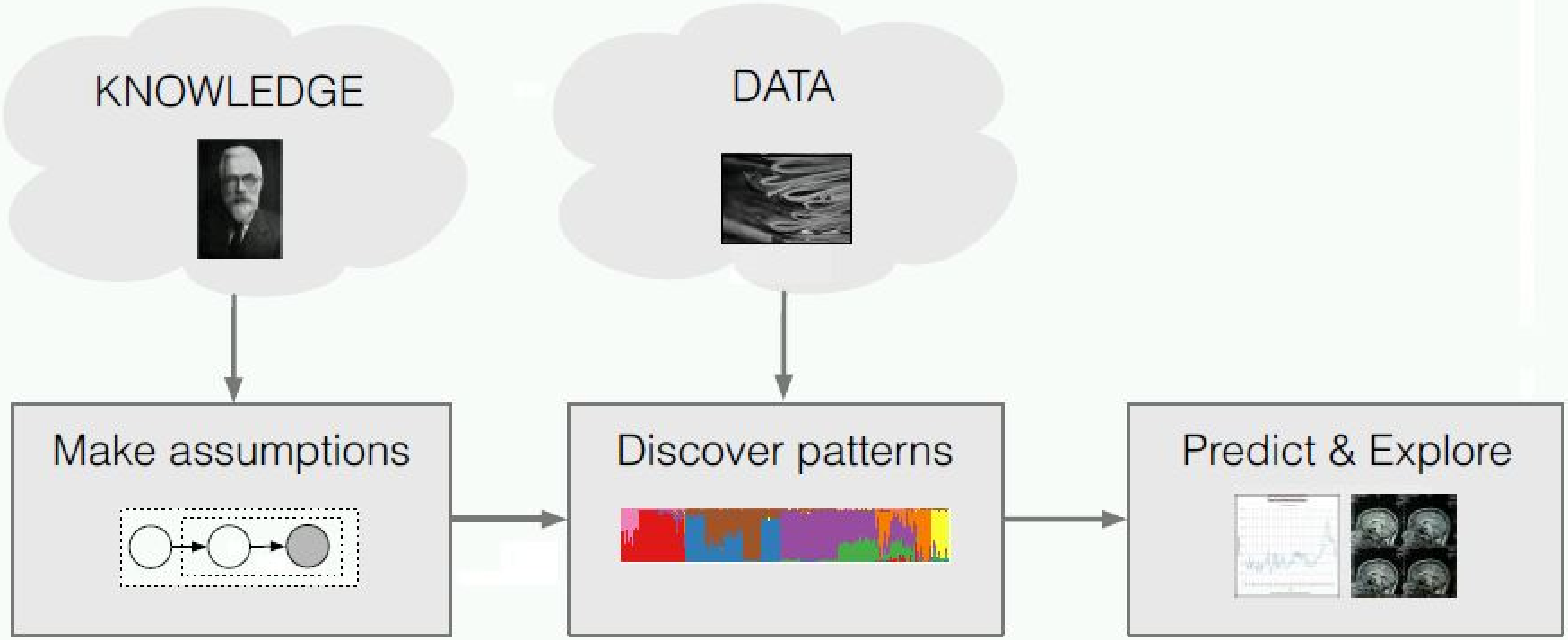
III. Use the discovered patterns to predict about and explore the data











What we need:

- **Flexible** and **expressive** components for building models
- **Scalable** and **generic** inference algorithms
- **Easy to use** software to stretch probabilistic modeling into new areas