Probabilistic Topic Models: Origins and Challenges

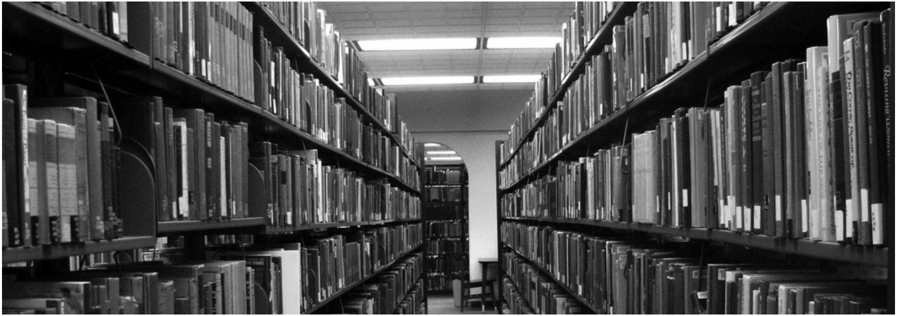
David M. Blei

Department of Computer Science
Princeton University

December 9, 2013

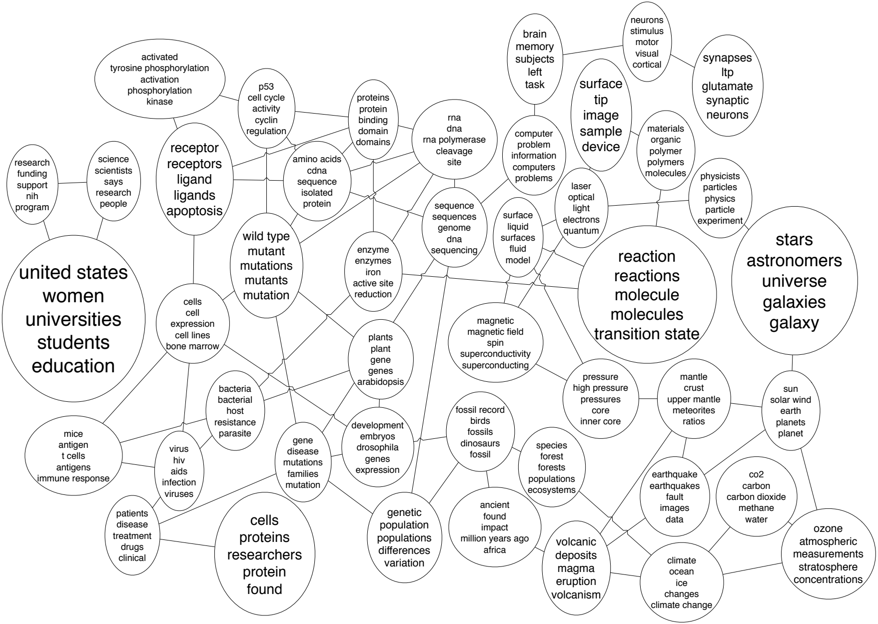
- **ORGANIZE**
- **VISUALIZE**
- **SUMMARIZE**
- **SEARCH**
- **PREDICT**
- **UNDERSTAND**

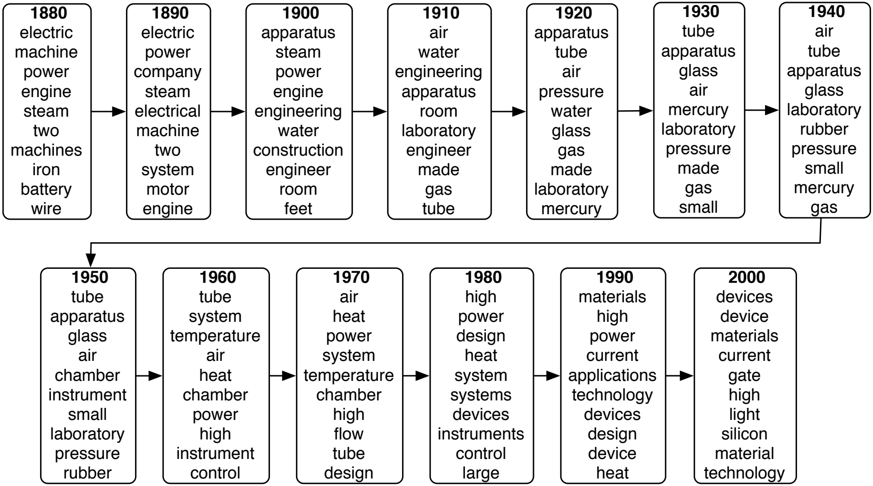
Probabilistic Topic Modeling



Input: An unorganized collection of documents

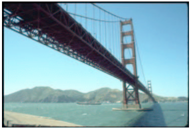
Output: An organized collection, and a description of how


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


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

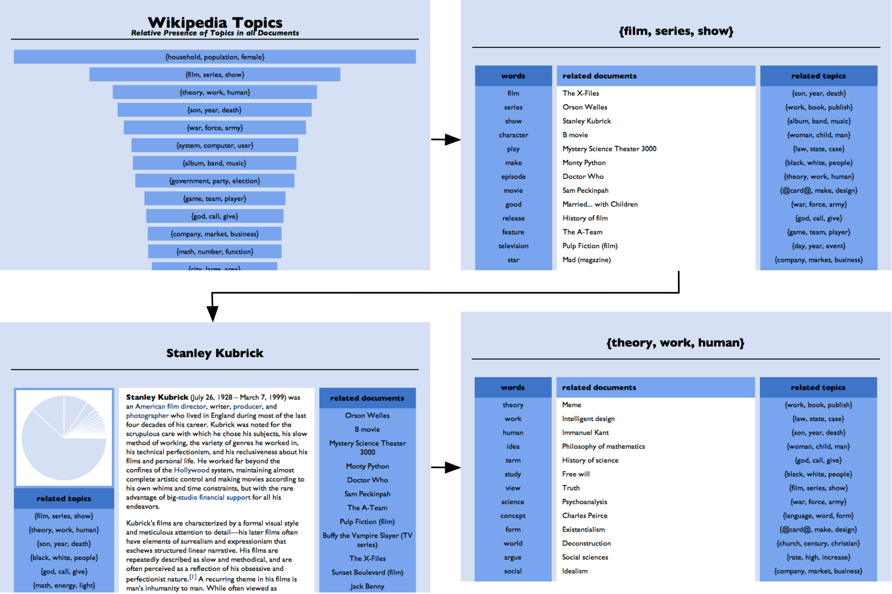


SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES



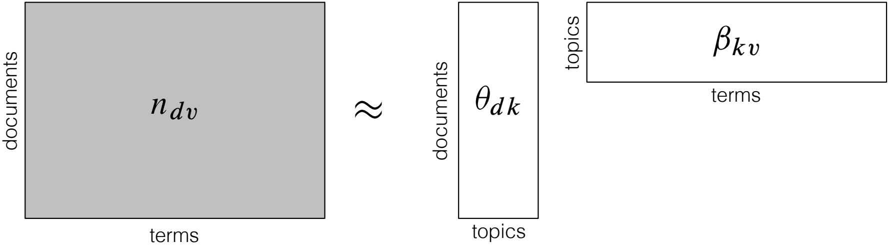
Wikipedia Topics
Relative Presence of Topics in all Documents
{household, population, female}
{film, series, show}
{theory, work, human}
{son, year, death}
{war, force, army}
{system, computer, user}
{album, band, music}
{government, party, election}
{game, team, player}
{god, call, give}
{company, market, business}
{math, number, function}
{city, large, area}
{film, series, show}
words
film
series
show
character
play
make
episode
movie
good
release
feature
television
star
related documents
The X-Files
Orson Welles
Stanley Kubrick
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
Married... with Children
History of film
The A-Team
Pulp Fiction (film)
Mad (magazine)
related topics
{son, year, death}
{work, book, publish}
{album, band, music}
{woman, child, man}
{law, state, case}
{black, white, people}
{theory, work, human}
{@card@, make, design}
{war, force, army}
{god, call, give}
{game, team, player}
{day, year, event}
{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
theory
work
human
idea
term
study
view
science
concept
form
world
argue
social
related documents
Meme
Intelligent design
Immanuel Kant
Philosophy of mathematics
History of science
Free will
Truth
Psychoanalysis
Charles Peirce
Existentialism
Deconstruction
Social sciences
Idealism
related topics
{work, book, publish}
{law, state, case}
{son, year, death}
{woman, child, man}
{god, call, give}
{black, white, people}
{film, series, show}
{war, force, army}
{language, word, form}
{@card@, make, design}
{church, century, christian}
{rate, high, increase}
{company, market, business}

This talk

1. The origins of probabilistic topic modeling
2. The basics of latent Dirichlet allocation
3. A couple ideas that we are exicted about in my group
4. Open questions, challenges, and discussion

Latent Semantic Analysis (LSA)

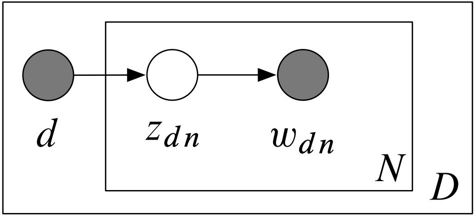
(Deerwester et al., 1990)



- This is the seminal work that launched topic modeling.
- Treat a collection as a document by term matrix of TFIDF scores.
- Choose a number of topics, and run SVD on the matrix.
- This results in
    - a matrix of per-document topic weights
    - a matrix of per-topic term weights

Probabilistic Latent Semantic Analysis (pLSA)

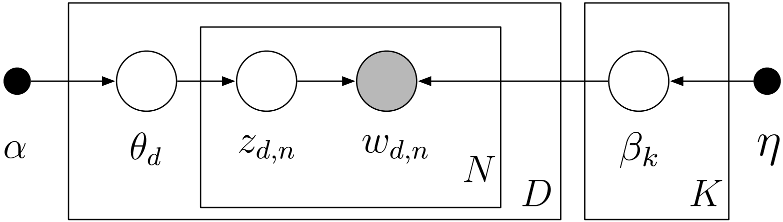
(Hofmann, 1999)



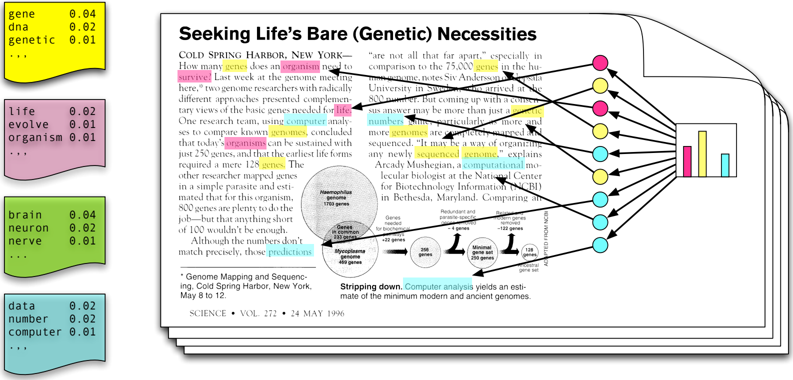
- A probabilistic model based on the main ideas of LSA
- Define a **topic** as a distribution over terms.
- Describe each document as a distribution over topics.
- Learn these two sets of parameters with EM.
- Note: This model was also defined in Papadimitriou et al., 1998

Latent Dirichlet Allocation (LDA)

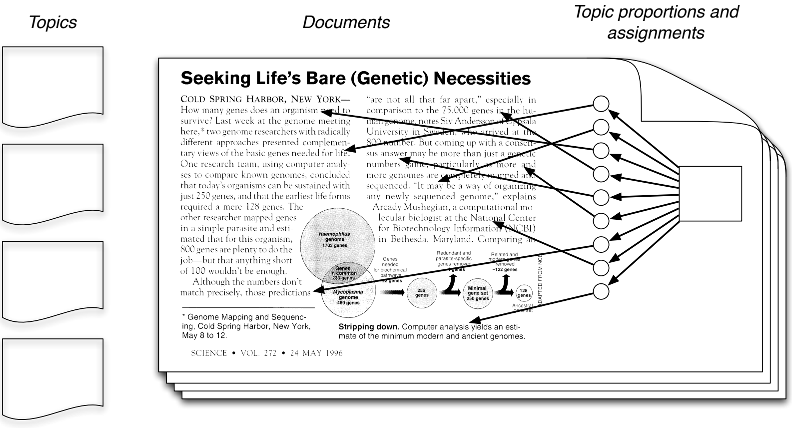
(Blei et al., 2001; Blei et al., 2003)



Topics *Documents* *Topic proportions and assignments*



Generative process

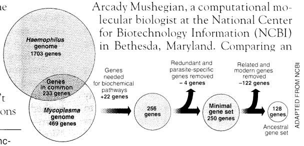


Posterior inference

Seeking Life's Bare (Genetic) Necessities

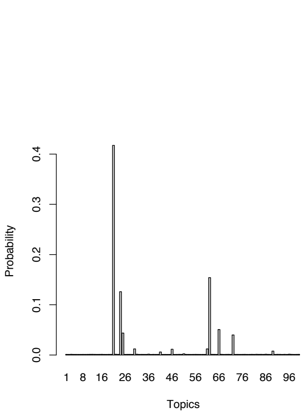
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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



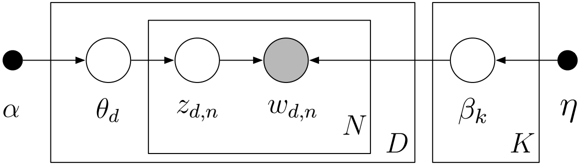
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* 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

Why does LDA "work"?



- LDA trades off two goals

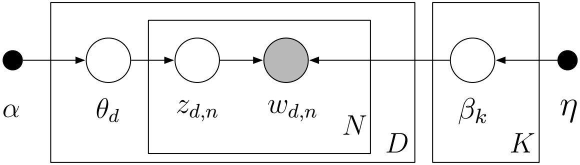
  1. In each **document**, allocate its words to **few topics**.
  2. In each **topic**, assign high probability to **few terms**.

- We see this from the joint

$$\log p(\cdot) = \ldots + \textstyle\sum_d \sum_n \log p(z_{dn} \mid \theta_d) + \log p(w_{dn} \mid \beta_{z_{dn}}) + \ldots$$

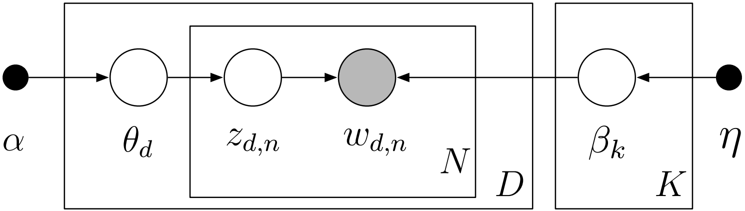
- Sparse proportions come from the 1st term.
  Sparse topics come from the 2nd term.

Why does LDA "work"?

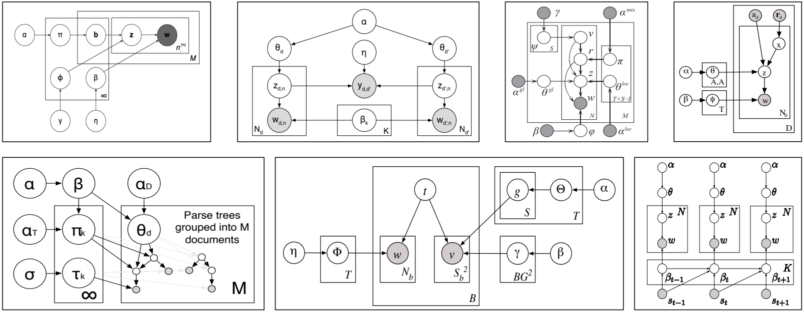


- LDA trades off two goals
  1. In each **document**, allocate its words to **few topics**.
  2. In each **topic**, assign high probability to **few terms**.
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard.
  - Putting very few words in each topic makes #1 hard.
- Trading off these goals finds groups of tightly co-occurring words.

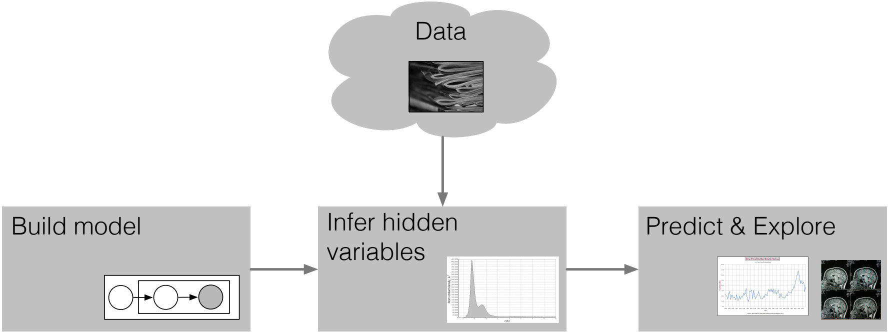
Summary and other perspectives



- Disovers topics through posterior inference
- Can be seen as *multinomial PCA* (Buntine and Jakulin, 2004)
- Is a type of *mixed-membership model* (Erosheva, 2004)
- Independently invented in population genetics (Pritchard et al., 2000)



- LDA is a simple building block that enables many applications.
- Organizing and finding patterns in data has become important in the sciences, humanties, industry, and culture.
- Algorithmic improvements let us fit models to massive data.



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

Some ideas we are excited about in my research group

Idea #1: User behavior data

Charles Darwin's library

Reading on the New York subway

- **People use documents.**
- This information can be used to
  - Help people find documents that they are interested in
  - Learn about how the documents are implicitly organized
  - Learn about the people reading the documents

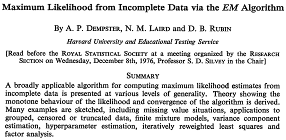
Idea #1: User behavior data

Charles Darwin's library

Reading on the New York subway

- **Collaborative topic models** analyze text and user data.
- They can be used to
  - recommend articles to readers: old and new
  - describe users in terms of their preferences
  - identify impactful, interdisciplinary articles

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



Maximum Likelihood from Incomplete Data via the *EM* Algorithm

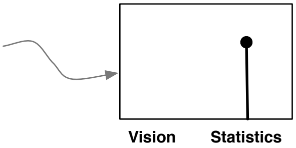
By A. P. Dempster, N. M. Laird and D. B. Rubin

Harvard University and Educational Testing Service

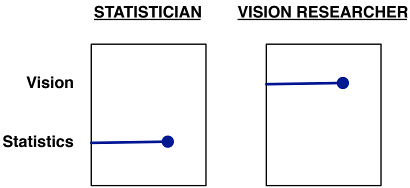
[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

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.

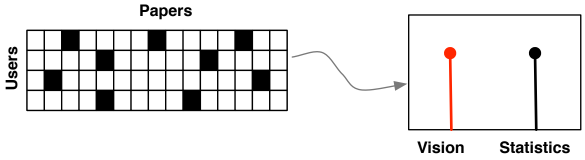


- Suppose there are two types of scientists

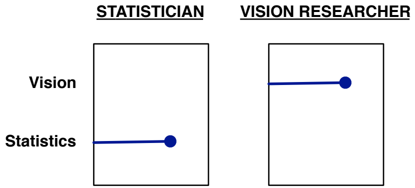


- We first recommend the EM paper to **statisticians**.

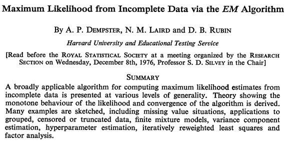
- With user data, we can adjust the topics to account for who liked it:



- Consider again the scientists



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



Maximum Likelihood from Incomplete Data via the *EM* Algorithm

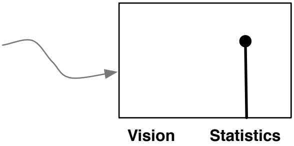
By A. P. Dempster, N. M. Laird and D. B. Rubin

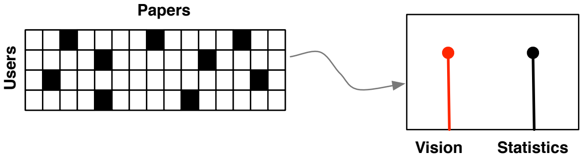
Harvard University and Educational Testing Service

[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

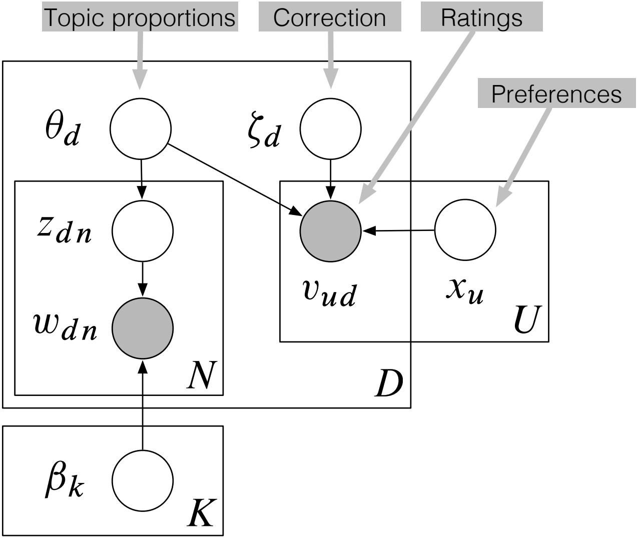
A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations,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



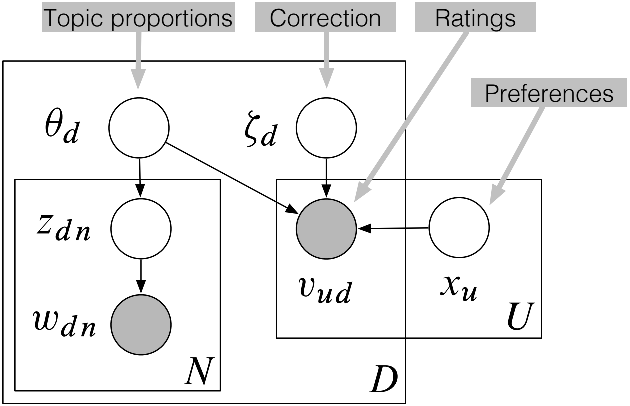


1. **Without text, we cannot initially recommend to anyone.**
2. **Without user data, we cannot recommend to vision researchers.**
3. **We learned about the special interdiscplinary status of the EM paper.**

The collaborative topic model



(Wang and Blei, 2011)



$$v_{ud} \sim f((\theta_d + \zeta_d)^\top x_u)$$

- Trades off matrix factorization and content recommendation
- The dimensions of user preferences also explain the text.
- Thus, they are interpretable.

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

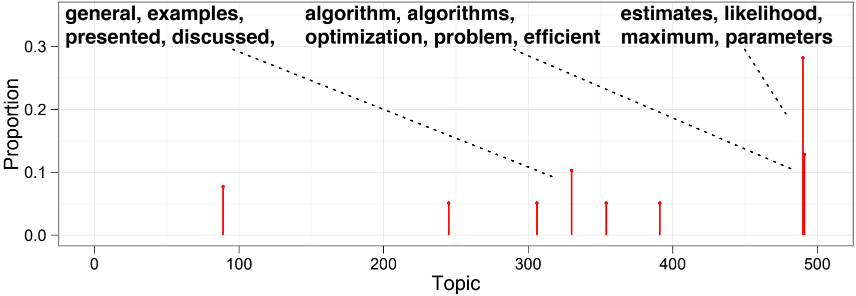
By A. P. Dempster, N. M. Laird and D. B. Rubin

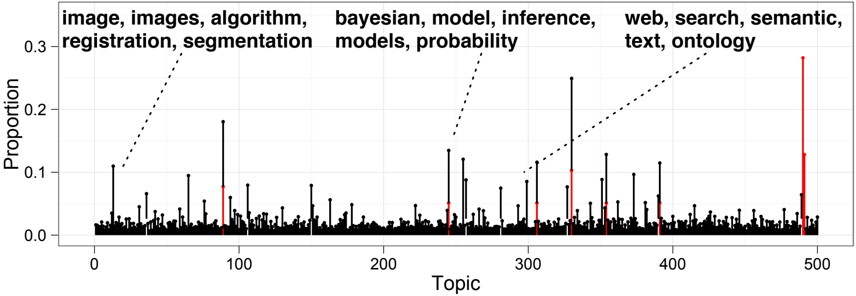
Harvard University and Educational Testing Service

[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.


general, examples,
presented, discussed,
algorithm, algorithms,
optimization, problem, efficient
estimates, likelihood,
maximum, parameters
Proportion
0.3
0.2
0.1
0.0
0
100
200
300
400
500
Topic


image, images, algorithm,
registration, segmentation
bayesian, model, inference,
models, probability
web, search, semantic,
text, ontology
Proportion
0.0
0.1
0.2
0.3
Topic
0
100
200
300
400
500

Maximum Likelihood Estimation

{Estimates, Likelihood, Maximum, Parameters, Method}

Widely read

- Maximum Likelihood Estimation of Population Parameters
- Bootstrap Methods: Another Look at the Jackknife
- R. A. Fisher and the Making of Maximum Likelihood

Interdisciplinary MLE articles

- Maximum Likelihood from Incomplete Data with the EM Algorithm
- Bootstrap Methods: Another Look at the Jackknife
- Tutorial on Maximum Likelihood Estimation

Outside influences

- Random Forests
- Identification of Causal Effects Using Instrumental Variables
- Matrix Computations

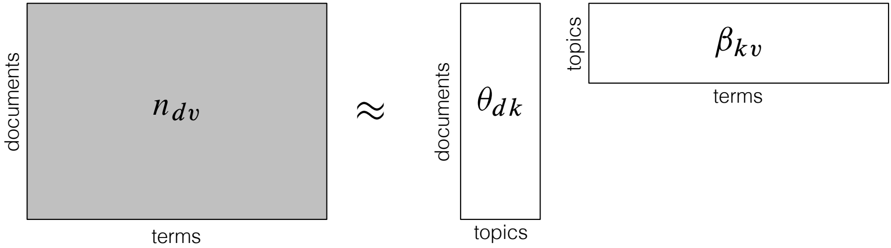
Idea #1: User behavior data

Charles Darwin's library

Reading on the New York subway

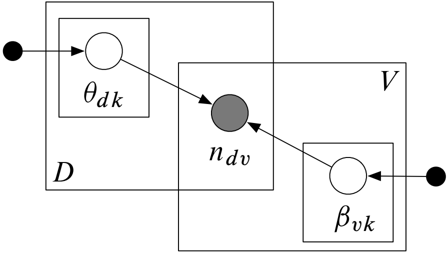
- Collaborative topic models give good recommendations.
- User behavior data give us a new window into the collection.
- Q: What if the users are in a network?
- Q: What if the users write reviews?

Idea #2: Poisson factorization



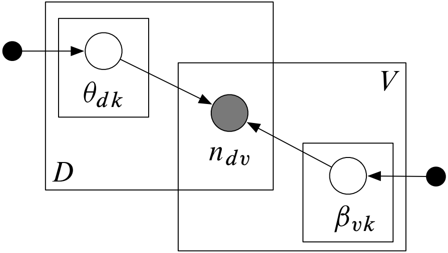
1. For each term v and topic k: draw $\beta_{kv} \sim \text{Gamma}(a, b)$
2. For each document d:
   a. For each topic k: draw $\theta_{dk} \sim \text{Gamma}(c, d)$.
   b. For each term v: draw $n_{dv} \sim \text{Poisson}(\theta_d^\top \beta_v)$.

Idea #2: Poisson factorization



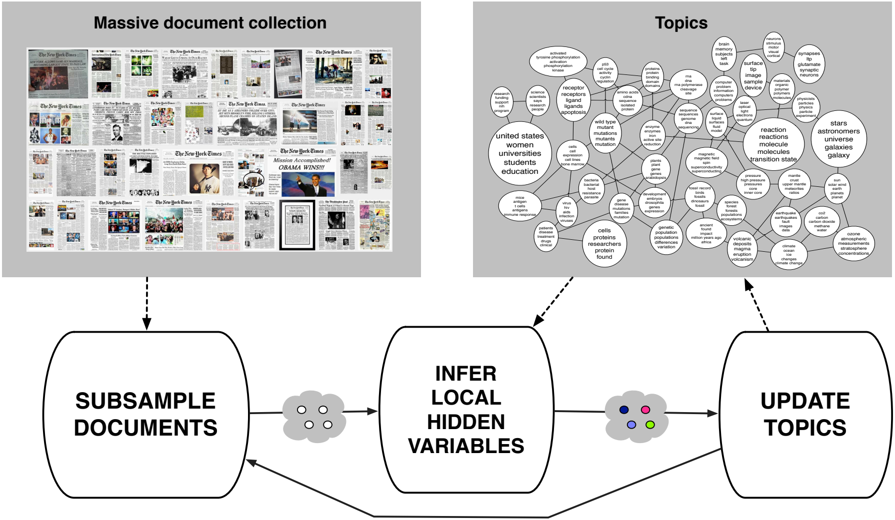
- Shows better perplexity than LDA. (Canny, 2004)
- Easy to fit with auxiliary variables
- Easy to extend the Poisson additive model on word counts
- Equivalent to LDA when we condition on document length
  (It is multinomial PCA.)
- Is a Bayesian form of NMF with "KL loss" (Lee and Seung, 2000)

Idea #2: Poisson factorization



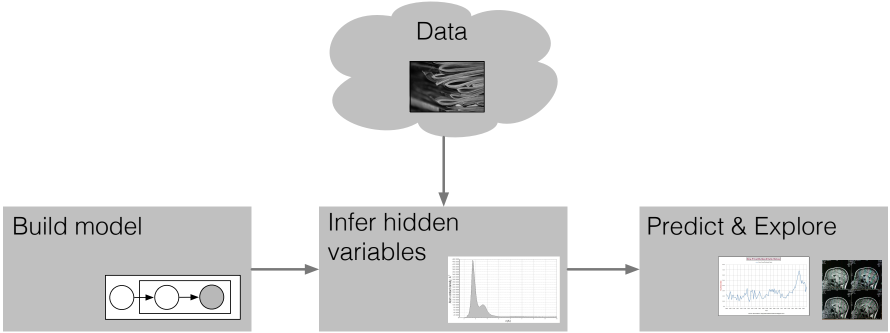
- Works well in other settings
  - networks (Ball et al., 2012) ; recommendation (Gopalan et al., 2013)
- We can build Bayesian nonparametric versions (Gopalan et al., yesterday)
- Why is it better than LDA?
  - Explicitly models document length?
  - Avoids pesky normalizations?

Idea #3: Stochastic Variational Inference



(Hoffman et al., 2010, 2013)

Challenges to topic modeling

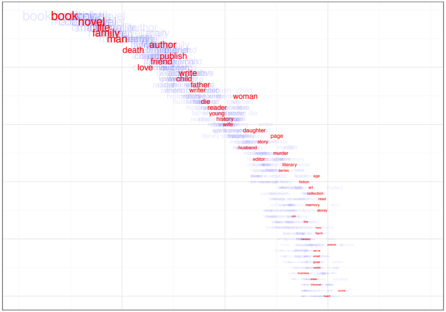


- Topic modeling research
  - develops new models.
  - develops new inference algorithms.
  - develops new applications, visualizations, tools.
- Workshops are also for half-baked ideas and difficult-to-articulate problems.

How do we explore?

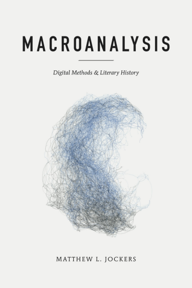
- Topic models are used to explore collections.
- How can we build and evaluate models with this goal?
- Brings to focus thorny issues
  - Visualization, Interpretability
  - Interactivity, Never-ending collections
- Theory of exploration (Tukey, 1962; Good, 1983; Diaconis, 1985)

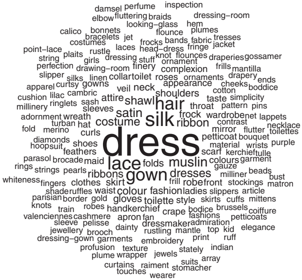
How do we select and revise?



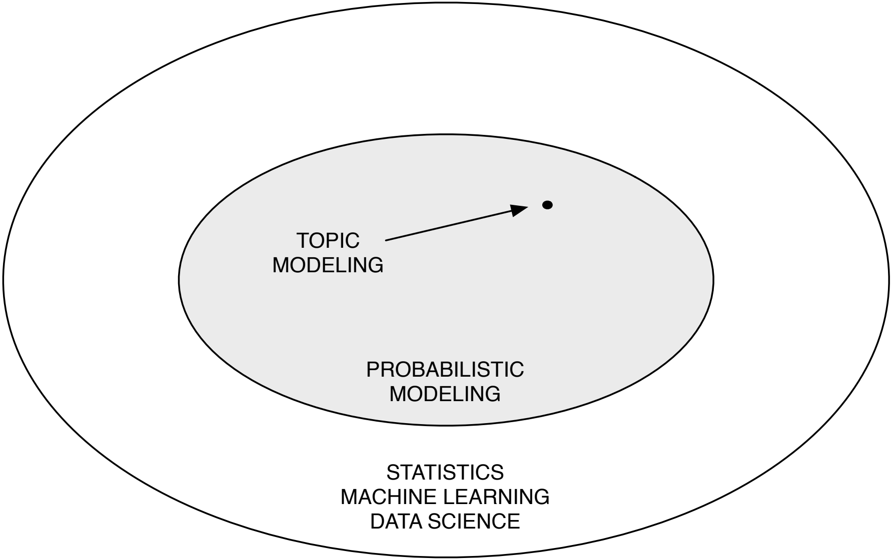
- Which model should I choose for my problem?
- Where does my model go right? Where does it go wrong?
- More thorny issues
  - Model evaluation
  - Posterior predictive checks (Box, 1980; Rubin, 1984; Gelman et al., 1996)

How do we apply?

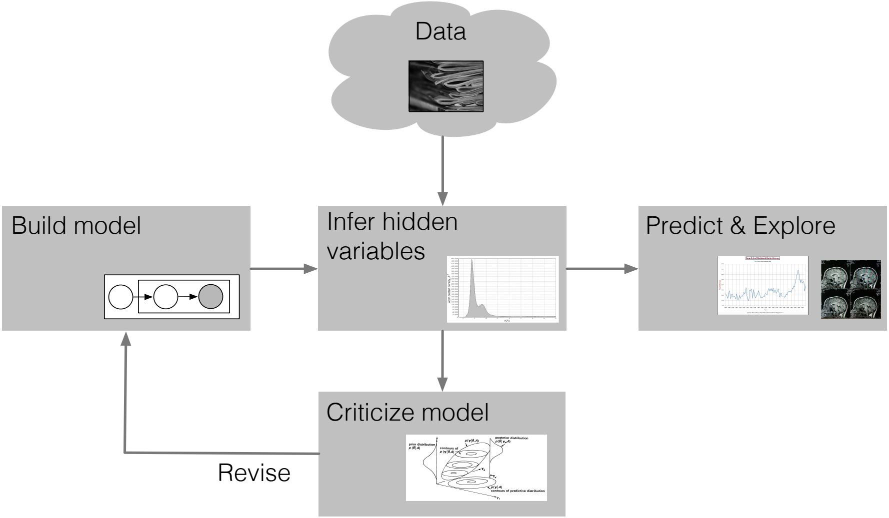




- Topic modeling moves in useful directions when we solve real problems.
- Collaborate with scientists/scholars that want to analyze texts
  - E.g., History, Comparative Literature, Political Science The Law, Cognitive Science, Sociology, Media Theory, Linguistics, Biology
- Create usable open-source tools for topic modeling.
- Success story: MALLET and the digital humanities.


TOPIC
MODELING
PROBABILISTIC
MODELING
STATISTICS
MACHINE LEARNING
DATA SCIENCE

Box's loop


Data
Build model
Infer hidden variables
Predict & Explore
Criticize model
Revise