

Topic Models for Data Mining

Chenghua Lin
Dept. of Computing Science
University of Aberdeen

Probabilistic topic models



As more information becomes available, it becomes more difficult to find and discover what we need.

We need new tools to help us organize, search, and understand these vast amounts of information.

Probabilistic topic models



David M. Blei



Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

Topic Models

A topic model is a generative probabilistic model for discrete data with latent structure

- Generative model for capturing semantic properties of text documents
- Can be applied to a wide variety of data
 - images, purchase logs, social network, music
- Easy to extend and implement

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

Topic Extraction

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

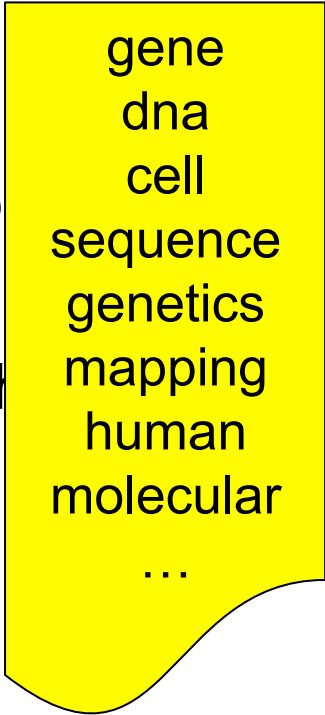
Input:
Documents

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Latent Dirichlet Allocation (LDA)

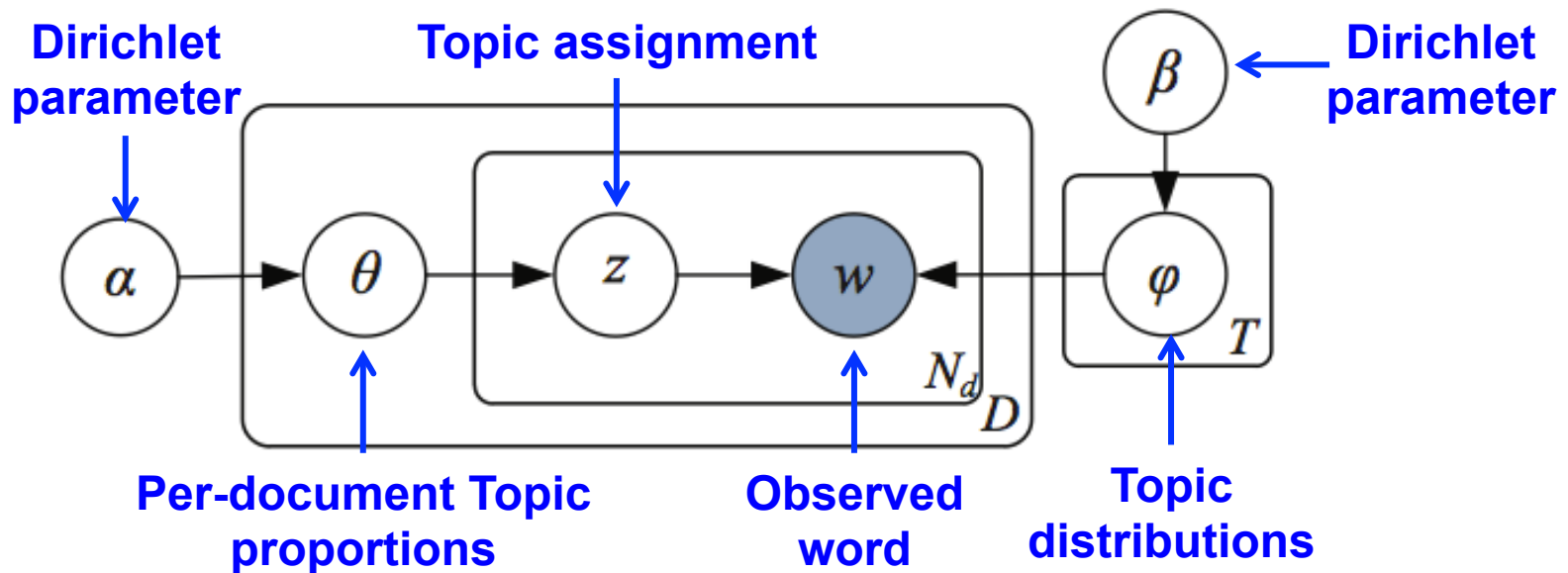
- LDA (Blei, Ng & Jordan, 2003)
 - A fully unsupervised Bayesian model
 - Assumes that documents exhibit multiple topics known as “**theme**” or “**gist**”)
 - Each topic is a distribution over words which have semantic relation with one another



gene
dna
cell
sequence
genetics
mapping
human
molecular

...

LDA Model



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

Dir
para



• Int

—

—

—

IDA Model			
“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

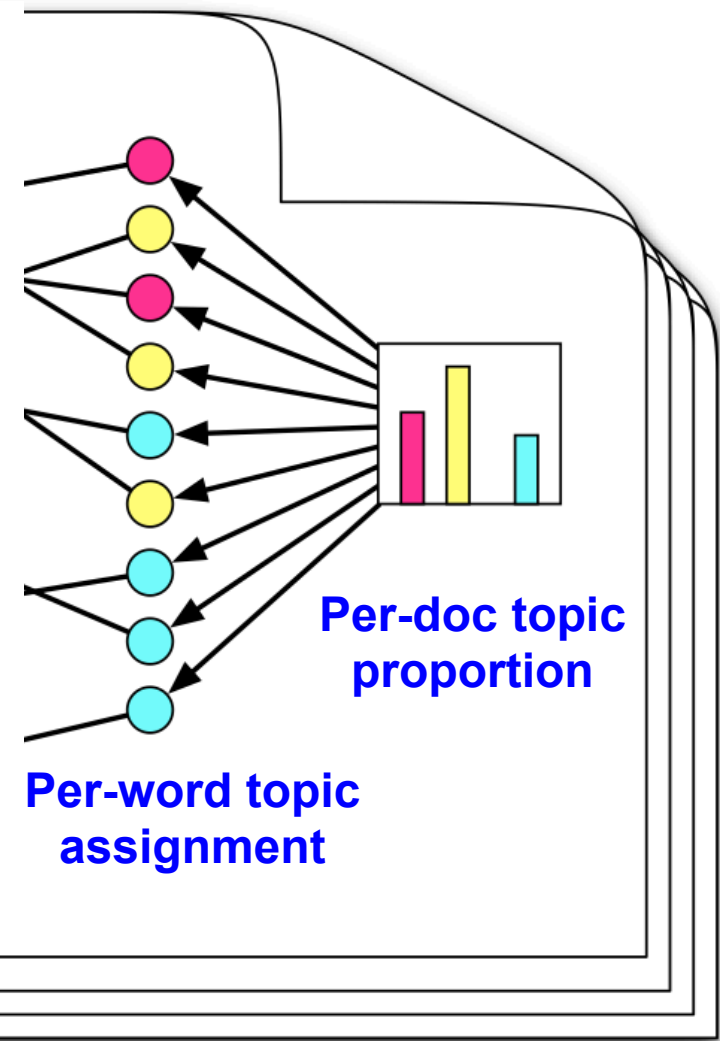
irichlet
rameter

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

w1 w2 w3 w4 ?? ...

Generate a document with a bulk
of words ...



Topics:

gene	0.04
dna	0.02
cell	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

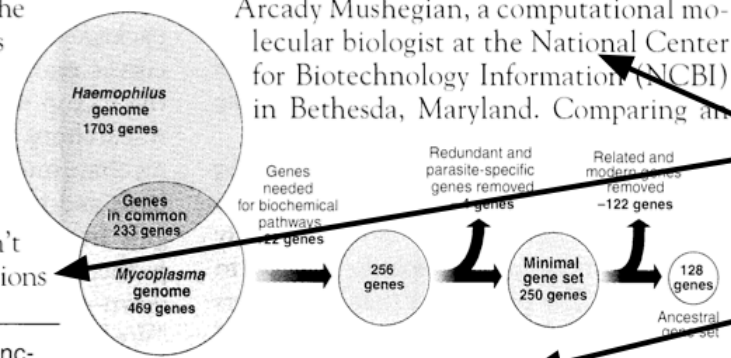
data	0.04
number	0.04
computer	0.04
...	

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.

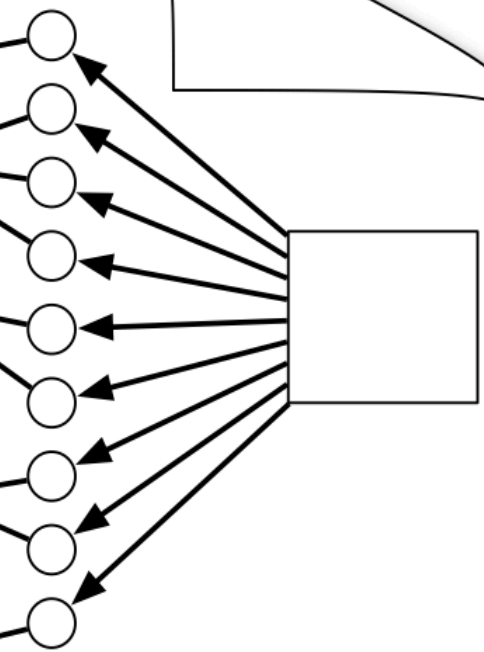
Although the numbers don't match precisely, those predictions

"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

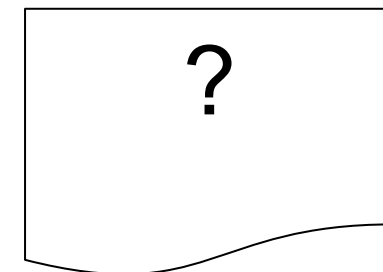
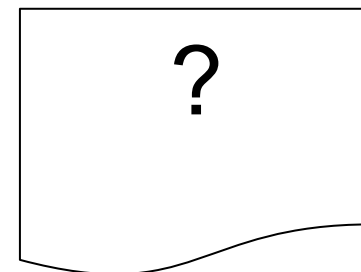
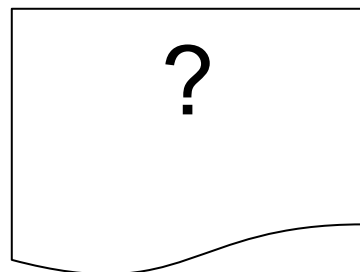
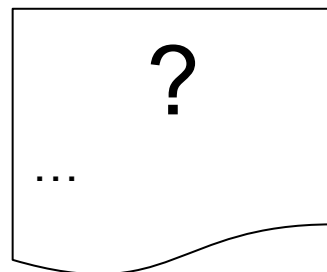


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

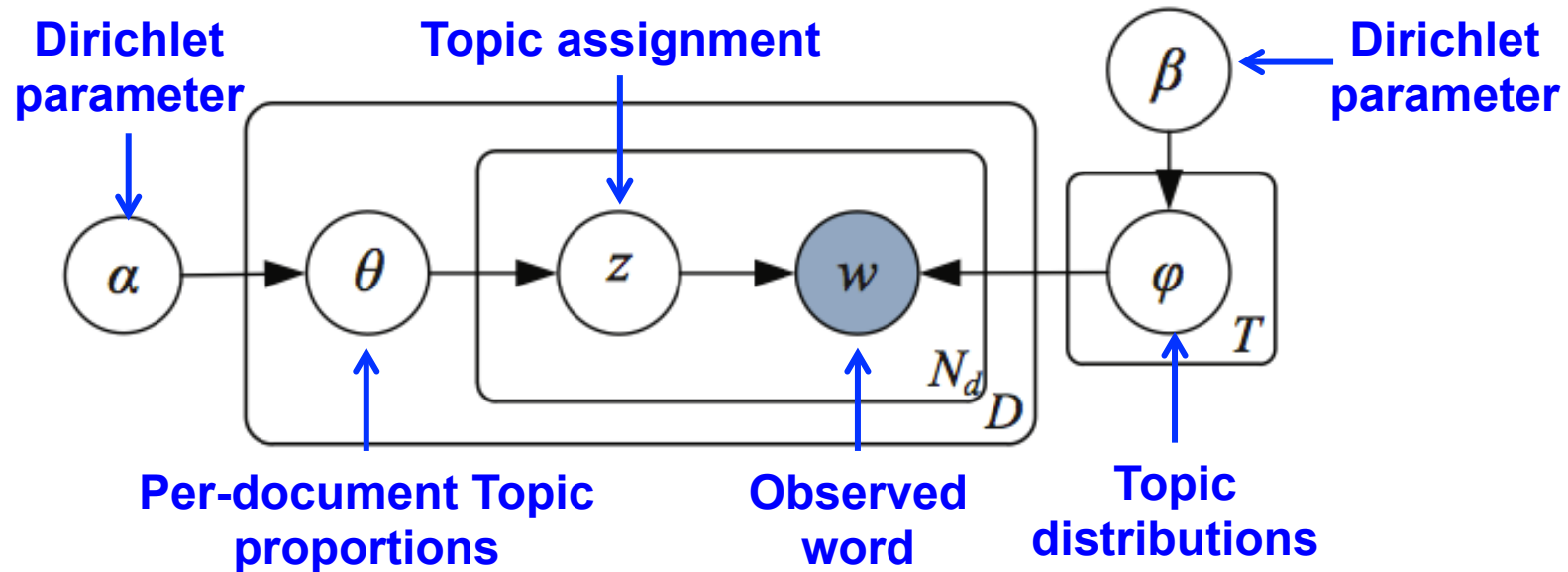
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.



Topics:



LDA Model



From a collection of documents, we need to infer

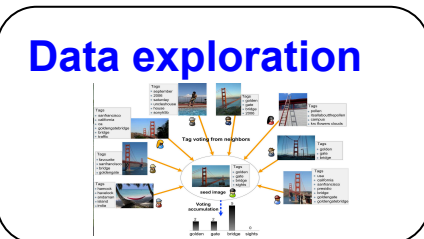
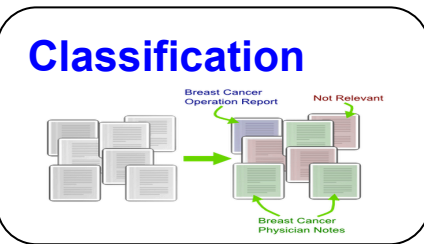
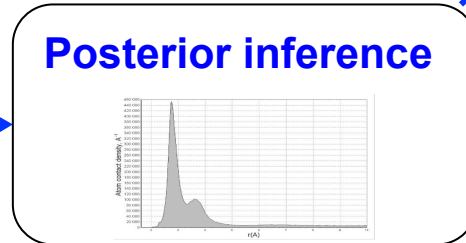
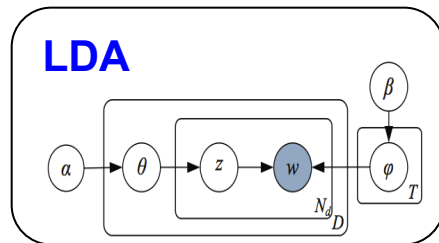
- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions θ_d
- Per-corpus topic distributions ϕ_k

Estimate a posterior, $p(\theta, z, \beta | w)$.

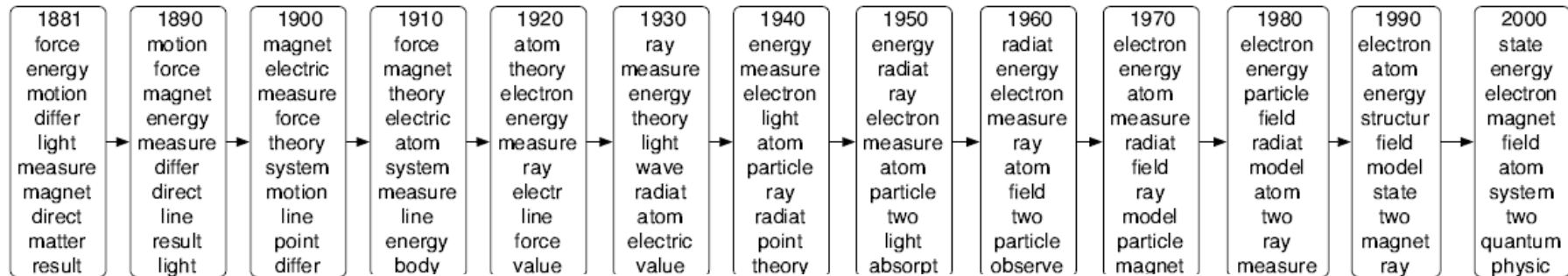
- Gibbs sampling
- Variational Bayes inference

Applications

$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^K p(\beta_i | \eta) \right) \left(\prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

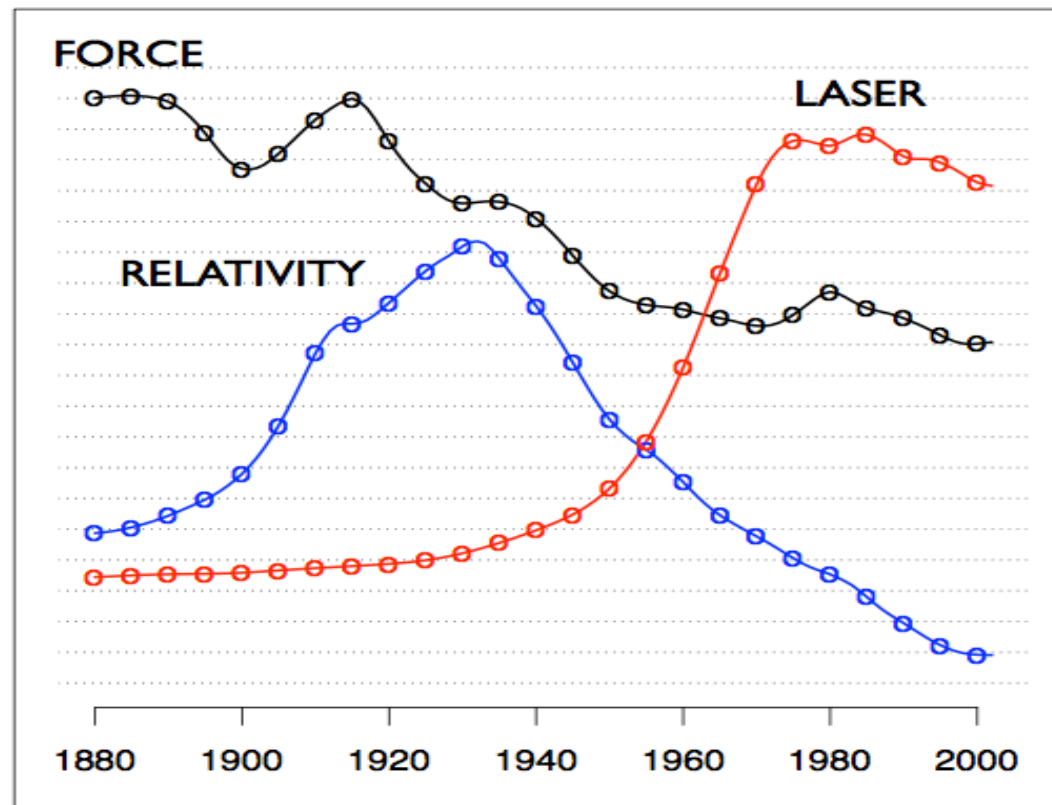


Dynamic Topic Analysis



"Atomic Physics"

Input:
Documents
with time info



ated Phenomena





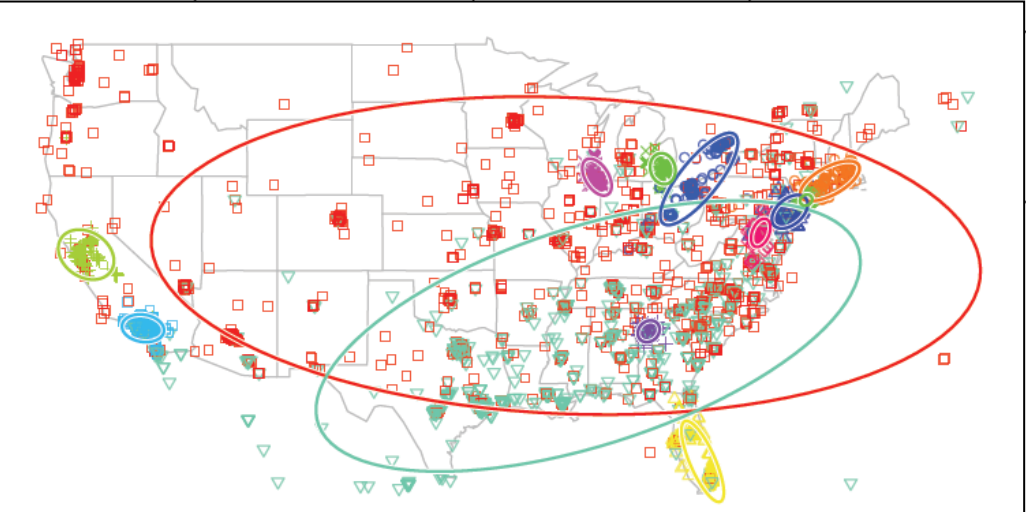
ne Common Metals

I

gnetic Monopole Obtained
ory
ns Floating on Liquid Helium

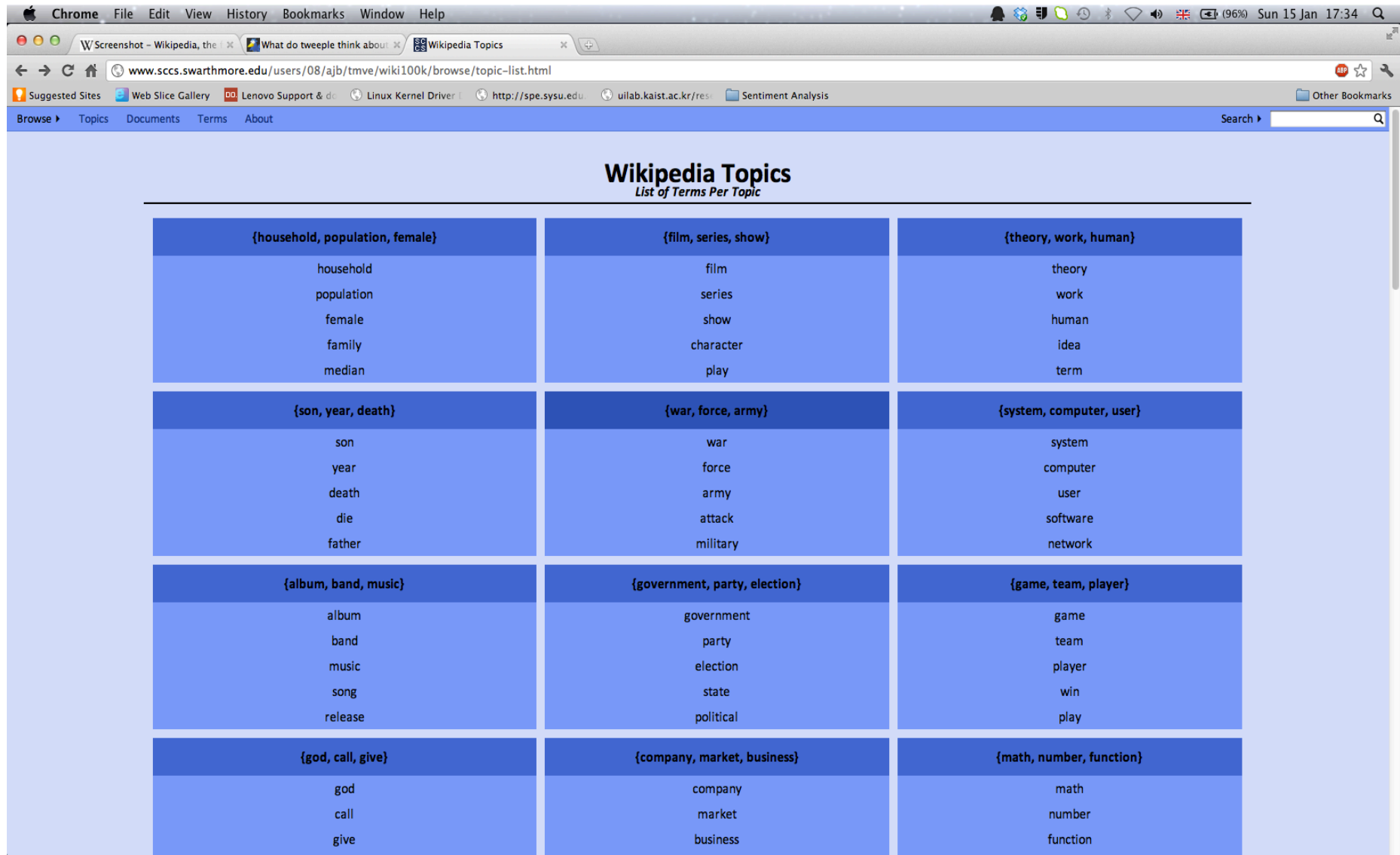
2006

Geo-location topic analysis

Input: geotaged tweets	“basketball” PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	“popular music” album music beats artist video #LAKERS ITUNES tour produced vol	“daily life” tonight shop weekend getting going chilling ready discount waiting iam	“emoticons”) haha :d :(;) :p xd :/ hahaha hahah	“chit chat” lol smh jk yea wyd coo ima wassup somethin jp
Boston 	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	<i>ese</i> exam suttin sippin
N. California 	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	<i>pues</i> hella koo SAN fckn	hella flirt hut iono OAKLAND
New York 	NETS KNICKS	BRONX	iam cab	oww	wasssup nm
Los Angeles 	#KOBE #LAKERS AUSTIN	#LA HO im			

J. Eisenstein, et al., A Latent Variable Model for Geographic Lexical Variation, EMNLP2010

A Wikipedia Article Browser



The screenshot shows a web browser window with the address bar displaying `www.sccs.swarthmore.edu/users/08/ajb/tmve/wiki100k/browse/topic-list.html`. The page title is "Wikipedia Topics" and the subtitle is "List of Terms Per Topic". The page content is organized into a 4x3 grid of topic categories, each with a list of related terms.

{household, population, female}	{film, series, show}	{theory, work, human}
household	film	theory
population	series	work
female	show	human
family	character	idea
median	play	term

{son, year, death}	{war, force, army}	{system, computer, user}
son	war	system
year	force	computer
death	army	user
die	attack	software
father	military	network

{album, band, music}	{government, party, election}	{game, team, player}
album	government	game
band	party	team
music	election	player
song	state	win
release	political	play

{god, call, give}	{company, market, business}	{math, number, function}
god	company	math
call	market	number
give	business	function

Topic Model for Opinion Mining

Motivating Example

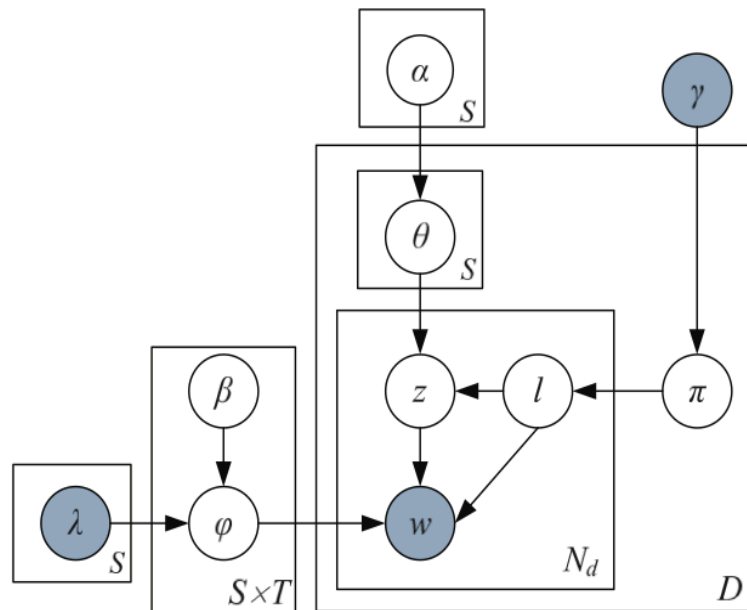
Customer Reviews

Amazon Kindle Keyboard Leather Cover, Black



- Rating does not tell you everything.
- Why many people gave 4 and 3 stars, because of quality, design, or what?
- Need to quickly gain insights into what people are talking about the product and the opinions

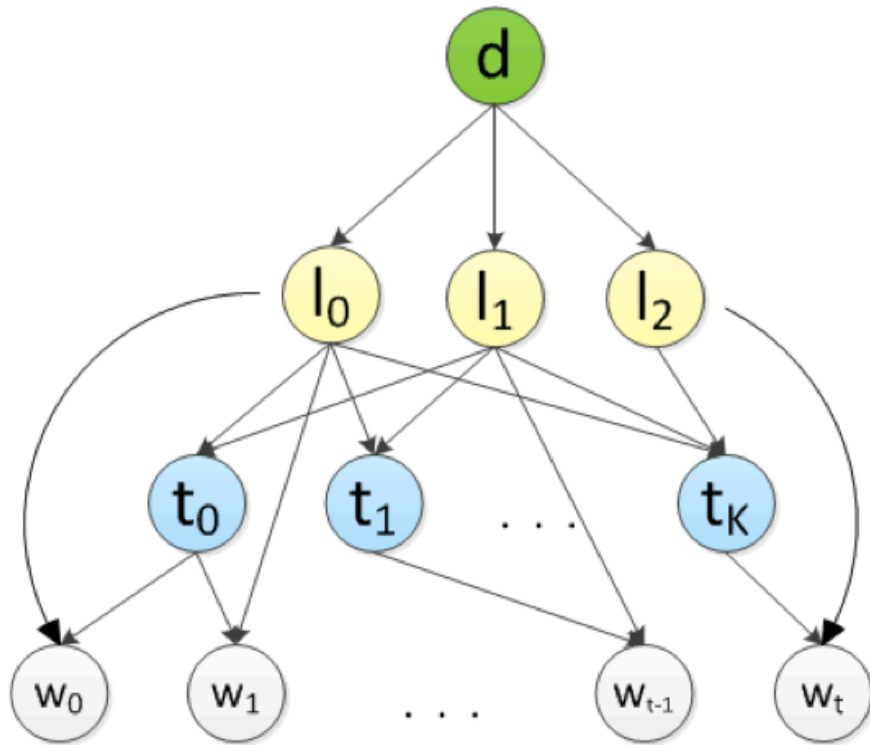
Joint Sentiment-Topic (JST) Model



JST is a weakly-supervised hierarchical Bayesian model which can:

- Detect the **overall sentiment** of a review/document.
- Automatically extract **sentiment-bearing topics** from a large number of reviews as succinct summaries.

JST Generative Process



For each document d

- Draw $\pi_d \sim \text{Dir}(\gamma \times \varepsilon_d)$
- For each topic label k , draw $\theta_{d,k} \sim \text{Dir}(\alpha_k)$

For each word w in d

- Draw a class label $l_i \sim \text{Mult}(\pi_d)$
- Draw a topic $z_i \sim \text{Mult}(\Theta_{d,l_i})$
- Draw a word $w_i \sim \text{Mult}(\Phi_{l_i,z_i})$

Applying JST to Review Data

Customer Reviews

Amazon Kindle Keyboard Leather Cover, Black



Negative sentiment topic

price
cover
high
expensive
kindle
overpriced
...

46 of 48 people found the following review helpful:

★★★★☆ Nice quality but....., 4 Sep 2010

By [C. Knight](#) (UK) - [See all my reviews](#)

REAL NAME

This review is from: Kindle Leather Cover, Black (Fits 6" Display, Latest Generation Kindle) (Accessory)

Not a bad cover. Nice leather, neat design , but **overpriced**, particularly when you consider the kindle is only 3 times the prices of this cover. And wouldnt it have been nice to make it left or right handed. The kindle is great having the page switches on both sides. but the cover makes the left hand switches harder to use. I do think the kindle should have come packaged with a cheap sleeve and feel this is a bit of a marketing rip off , (there is no way I can walk round with the kindle unprotected). But, have to say, it does the job. Just well **overpriced**.

Summary

- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a posterior inference problem.
- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.
- LDA is easy to extend for other applications, e.g. sentiment analysis, etc.
- What you should know
 - What are topic models?
 - What kinds of things can they do?