CS 277, Data Mining

Topic Modeling for Text Documents

Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Project Progress Report 1

- Grade distribution
  - A few 9’s: excellent reports
  - Many 8’s: good reports, nothing outstanding
  - Many 7’s: good, but could be improved, e.g., missing key information
  - Below 7: needs improvement

Project Progress Report 1

- Grade distribution
  - A few 9's: excellent reports
  - Many 8's: good reports, nothing outstanding
  - Many 7's: good, but could be improved, e.g., missing key information
  - Below 7: needs improvement

- What is required to get a top score?
  1. Technical progress on your project
     - Not just summarizing well-known knowledge (e.g., known algorithms)
     - Insights, figures, details: show me what you have learned so far
     - Clear that you have spent time on this project
       (note that spending time alone is not sufficient to get a high score)
  2. Clearly written report
     - Well organized, each section builds on the preceding section
     - Good use of figures and graphs
     - Goals, methods, results are clearly explained: reader is not guessing what you did
     - Appropriate level of detail (not too much…but some details are good)

TimeTable (Updated)

- Progress Report 2: now due Wednesday Feb 26th (1 week away)
  - Hand in hardcopy in class
  - AND submit electronic copy online via EEE
  - You can re-use a small amount of figures/text that you think is necessary/helpful from your earlier report, but clearly point this out. At least 80% of what you report should be new.

- Final Report
  - Due noon Friday March 14th (electronically to EEE)
  - Will discuss the format and expectations in more detail later in the quarter

Basis of Final Grade

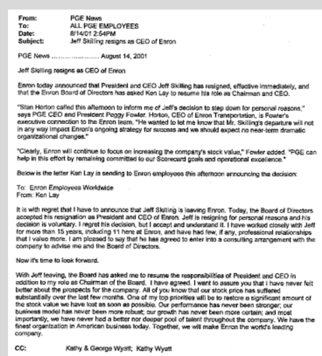
- Weighted combination of Assignments and Reports
  - 10% for Assignment 1
  - 20% for each of the Progress Reports
  - 30% for the Final Report

Topic Models for Text

Examples of Large Text Corpora

NYT: 1.5 million news articles

NIH: 1 million grant abstracts



From:	PGE News
To:	ALL PGE EMPLOYEES
Date:	8/14/01 2:54PM
Subject:	Jeff Skilling resigns as CEO of Enron

PGE News .... August 14, 2001

Jeff Skilling resigns as CEO of Enron

Enron today announced that President and CEO Jeff Skilling has resigned, effective immediately, and that the Enron Board of Directors has asked Ken Lay to resume his role as Chairman and CEO.

"Stan Horton called this afternoon to inform me of Jeff's decision to step down for personal reasons," says PGE CEO and President Peggy Fowler. Horton, CEO of Enron Transportation, is Fowler's executive connection to the Enron team. "He wanted to let me know that Mr. Skilling's departure will not in any way impact Enron's ongoing strategy for success and we should expect no near-term dramatic organizational changes."

"Clearly, Enron will continue to focus on increasing the company's stock value," Fowler added. "PGE can help in this effort by remaining committed to our Scorecard goals and operational excellence."

Below is the letter Ken Lay is sending to Enron employees this afternoon announcing the decision:

To: Enron Employees Worldwide
From: Ken Lay

It is with regret that I have to announce that Jeff Skilling is leaving Enron. Today, the Board of Directors accepted his resignation as President and CEO of Enron. Jeff is resigning for personal reasons and his decision is voluntary. I regret his decision, but I accept and understand it. I have worked closely with Jeff for more than 15 years, including 11 here at Enron, and have had few, if any, professional relationships that I value more. I am pleased to say that he has agreed to enter into a consulting arrangement with the company to advise me and the Board of Directors.

Now it's time to look forward.

With Jeff leaving, the Board has asked me to resume the responsibilities of President and CEO in addition to my role as Chairman of the Board. I have agreed. I want to assure you that I have never felt better about the prospects for the company. All of you know that our stock price has suffered substantially over the last few months. One of my top priorities will be to restore a significant amount of the stock value we have lost as soon as possible. Our performance has never been stronger; our business model has never been more robust; our growth has never been more certain; and most importantly, we have never had a better nor deeper pool of talent throughout the company. We have the finest organization in American business today. Together, we will make Enron the world's leading company.

CC: Kathy & George Wyatt; Kathy Wyatt

Enron:

250,000 emails

PubMed: 20 million biomedical articles

Unsupervised Learning from Text

- Large collections of unlabeled documents..
  - Web
  - Digital libraries
  - Email archives, etc

- Often wish to organize/summarize/index/tag these documents automatically

- We will look at probabilistic techniques for clustering and topic extraction from sets of documents

Problems of Interest

- What topics do these documents "span"?
- Which documents are about a particular topic?
- How have topics changed over time?
- What does author X write about?
- Who is likely to write about topic Y?
- Who wrote this specific document?
- and so on.....

"Bag of Words" Matrix for Documents

	water	rights	cattle	hunting	land	use	dust	maize	chumash	navajo	reservation	war	disease
doc1	1	1			2								
doc2			1	4	1								1
doc3	2	1	1		1	1	1						
doc4			2		1						1		
doc5	1	1	1	1	3	1	1						
doc6		1		2									
doc7									1	1	1	3	
doc8					1			1	1				
doc9										1	1		
doc10									1		1		1

Statistical Topic Models for Count Data

- Simple hypothetical "generative" models for sparse counts
  - A description of how the data might have been generated
  - Simple in nature ("all models are wrong but some are useful")
  - Can handle counts, metadata, etc

- Learning the parameters given the data
  - Generative model = P(D | θ): how likely data D are given the parameters θ
  - Use Bayes rule to get P(θ | D): how likely parameters θ are given data D

Statistical Topic Models for Count Data

- Simple hypothetical "generative" models for sparse counts
  - A description of how the data might have been generated
  - Simple in nature ("all models are wrong but some are useful")
  - Can handle counts, metadata, etc
- Learning the parameters given the data
  - Generative model = $P(D \mid \theta)$: how likely data D are given the parameters θ
  - Use Bayes rule to get $P(\theta \mid D)$: how likely parameters θ are given data D
- Key Features
  - Multimembership: rows can "belong" to multiple factors
  - Leverage sparsity: computational advantages
  - Can build in dependence on metadata (e.g., document authors)

Modeling Word Frequencies given Count Data

Tossing a die: 6 sides, equally likely, memoryless

Parameters of a "model" for a die:

A vector of 6 probabilities $\theta_1,..\theta_6$ sum to 1, $\Sigma\,\theta = 1$

Modeling Word Frequencies given Count Data

Tossing a die: 6 sides, equally likely, memoryless

Parameters of a "model" for a die:

A vector of 6 probabilities $\theta_1,..\theta_6$ sum to 1, $\Sigma\ \theta\ = 1$

Same model for text?

Now we have a K-sided die, where K could be 100,000

A vector of K probabilities $\underline{\theta}$, sum to 1, $\Sigma\ \theta\ = 1$

Modeling Word Frequencies given Count Data

Tossing a die: 6 sides, equally likely, memoryless

Parameters of a "model" for a die:

A vector of 6 probabilities $\theta_1,..\theta_6$ sum to 1, $\Sigma\,\theta = 1$

Same model for text?

Now we have a K-sided die, where K could be 100,000

A vector of K probabilities $\underline{\theta}$, sum to 1, $\Sigma\,\theta = 1$

Can learn these probabilities from a corpus, via smoothed frequency counts

Applications? detecting shifts in language usage in scientific literature, in customer complaints, in social media, etc

Topic = "Focused" Probability Distribution over Words

Word	Probability
president	0.129
roosevelt	0.032
congress	0.030
johnson	0.026
office	0.021
wilson	0.021
nixon	0.020
reagan	0.018
kennedy	0.018
…	…

Different Topics for Different Semantic Concepts

Word	Probability
red	0.202
blue	0.099
green	0.096
yellow	0.073
white	0.048
color	0.030
bright	0.029
colors	0.027
brown	0.027
….	….

Word	Probability
president	0.129
roosevelt	0.032
congress	0.030
johnson	0.026
office	0.021
wilson	0.021
nixon	0.020
reagan	0.018
kennedy	0.018
….	….

Key Idea: Documents as Mixtures of Topics

Topic 1: search_query (0.4), precision (0.3), retrieval (0.3)

Topic 2: classification (0.5), neural_network (0.3), labels (0.2)

Topic 3: experiment (0.6), result (0.2), significance (0.2)

Key Idea: Documents as Mixtures of Topics

Topic 1: search_query (0.4), precision (0.3), retrieval (0.3)

Topic 2: classification (0.5), neural_network (0.3), labels (0.2)

Topic 3: experiment (0.6), result (0.2), significance (0.2)

Topic model: documents = convex combinations of Topics 1, 2, 3: e.g.,

P(Words) for Doc 1 = 0.4 * Topic 1 + 0.4 * Topic 2 + 0.2 * Topic 3

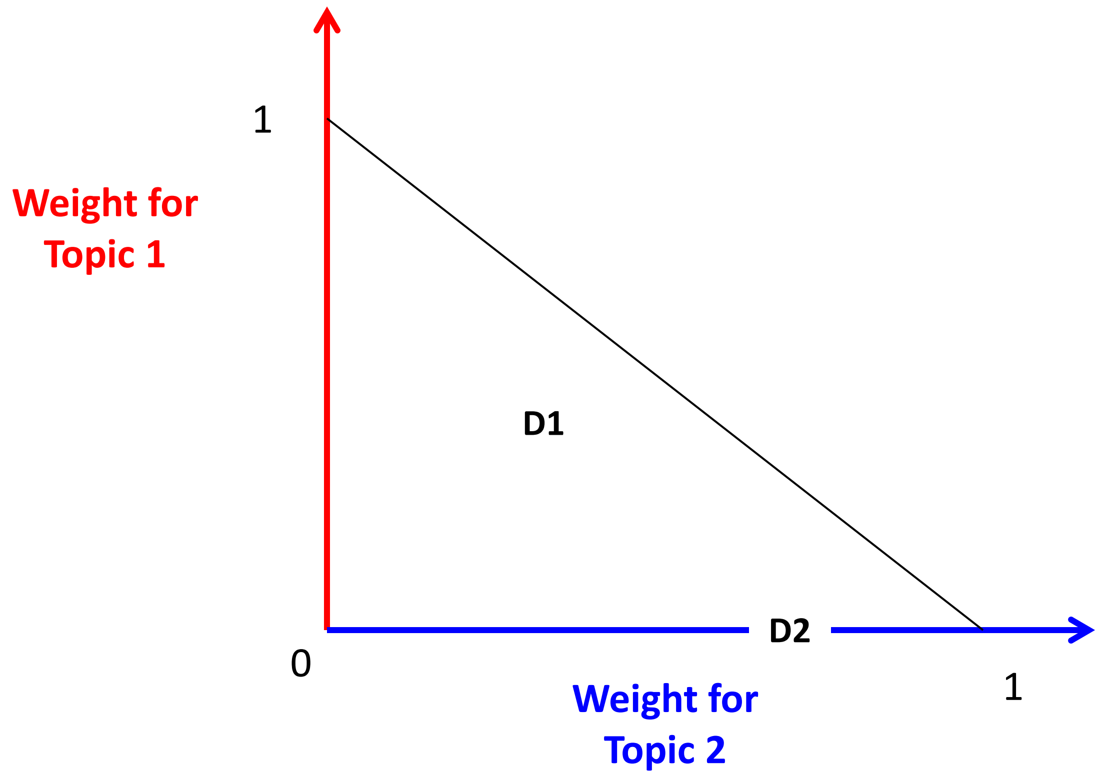
P(Words) for Doc 2 = 0.0 * Topic 1 + 0.8 * Topic 2 + 0.2 * Topic 3

Key Idea: Documents as Mixtures of Topics

Topic 1: search_query (0.4), precision (0.3), retrieval (0.3)

Topic 2: classification (0.5), neural_network (0.3), labels (0.2)

Topic 3: experiment (0.6), result (0.2), significance (0.2)



A Generative Model for Documents

- Our topic model is a simple "forward generative" model for observed data (i.e., counts of words in documents)

- We can think of it as a simulator (with pseudocode)

```
For each document in our corpus
    For each word in our document
        Sample a topic from P(topics | document)
        Given the topic, sample a word from P(words | topic)
    End
End
```

A Generative Model for Documents

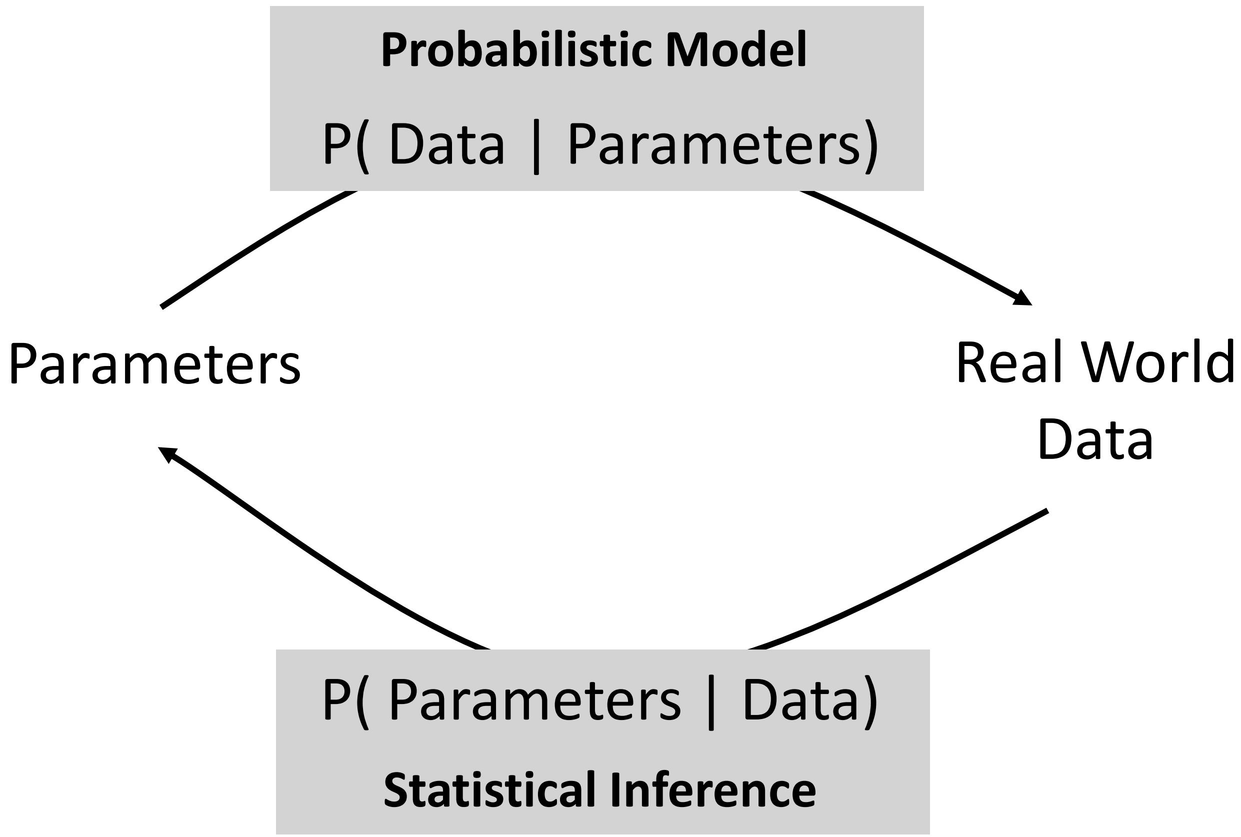
- Our topic model is a simple "forward generative" model for observed data (i.e., counts of words in documents)

- We can think of it as a simulator (with pseudocode)

```
For each document in our corpus
    For each word in our document
        Sample a topic from P(topics | document)
        Given the topic, sample a word from P(words | topic)
    End
End
```

- Key points:
  - Many useful statistical models in machine learning have a very simple "forward mechanism" (a few lines of pseudocode)
  - Learning this model is essentially "inverting" this code via Bayes rule
  - The reverse "inference" step is usually much harder than the forward step!

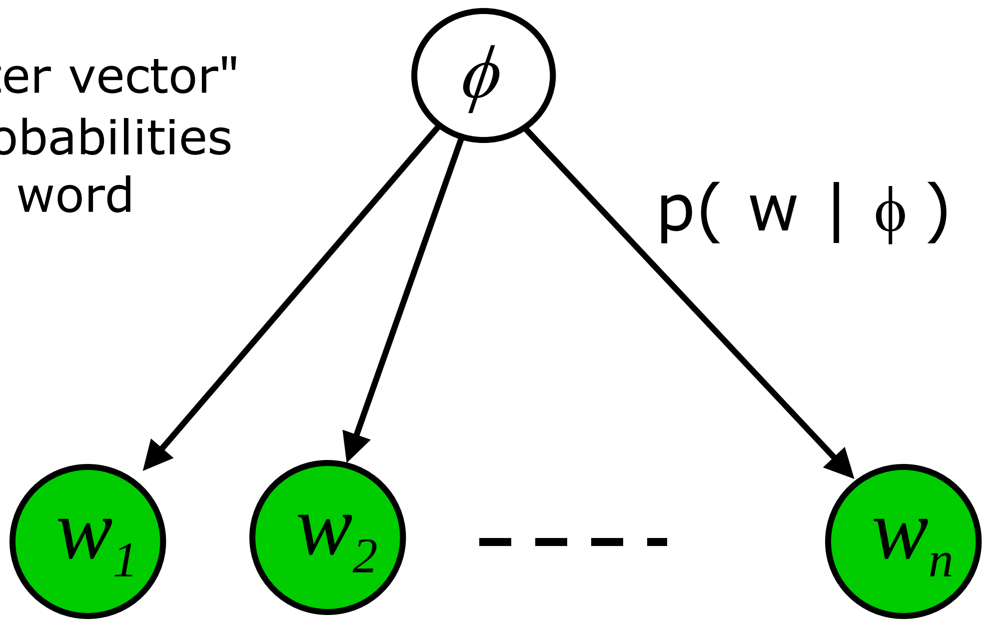
Generative Statistical Models for Data



A Graphical Model

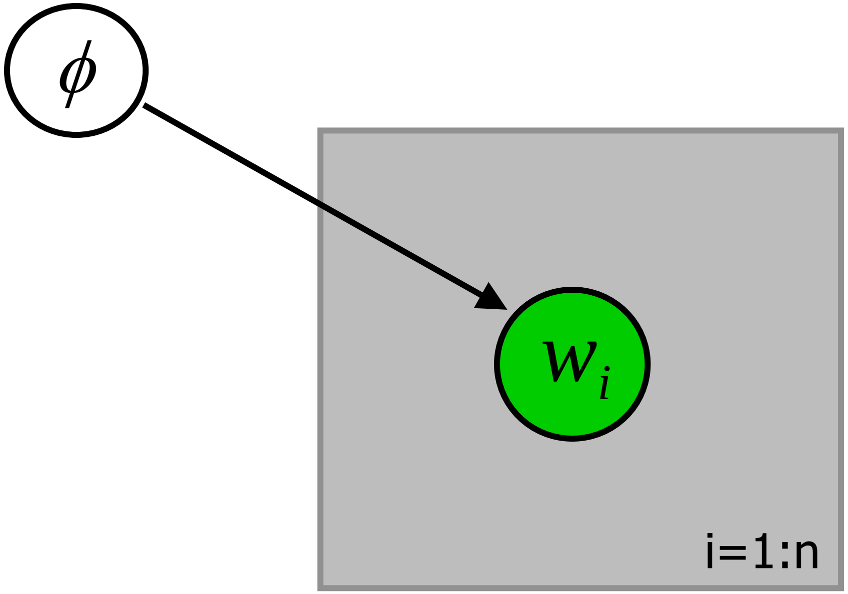
$$p(\text{ doc } | \phi) = \prod p(w_i | \phi)$$

ϕ = "parameter vector"
= set of probabilities one per word



Another view....

$$p(\text{ doc } | \, \phi \,) = \prod \, p(w_i \,|\, \phi \,)$$



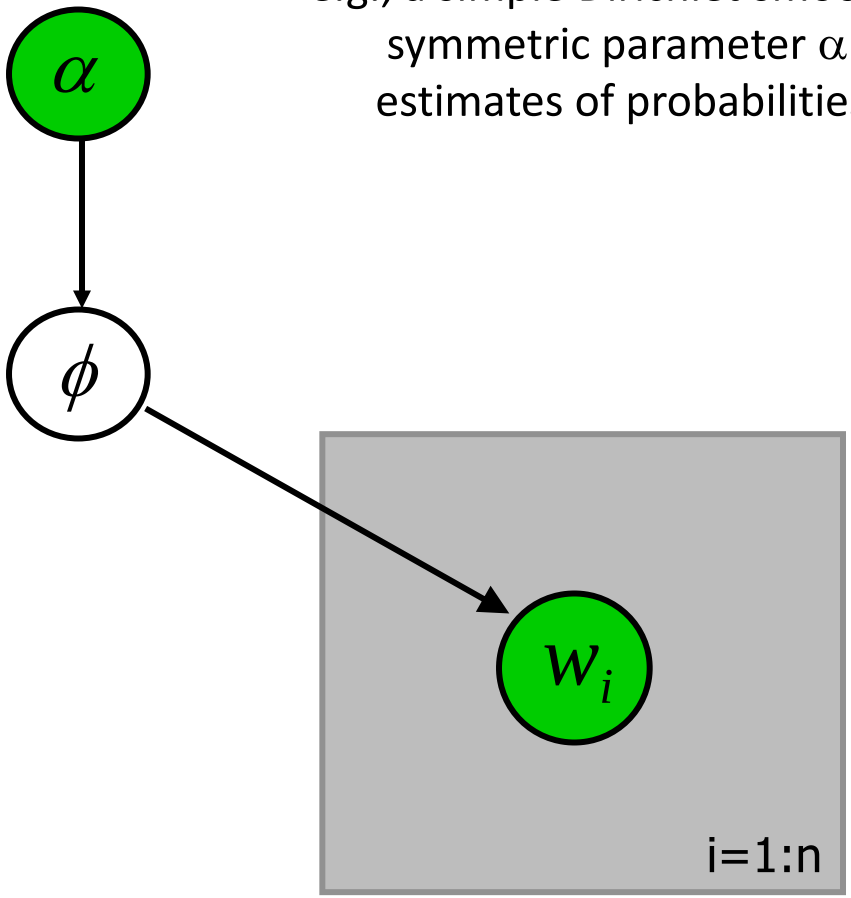
This is "plate notation"

Items inside the plate
are conditionally independent
given the variable outside
the plate

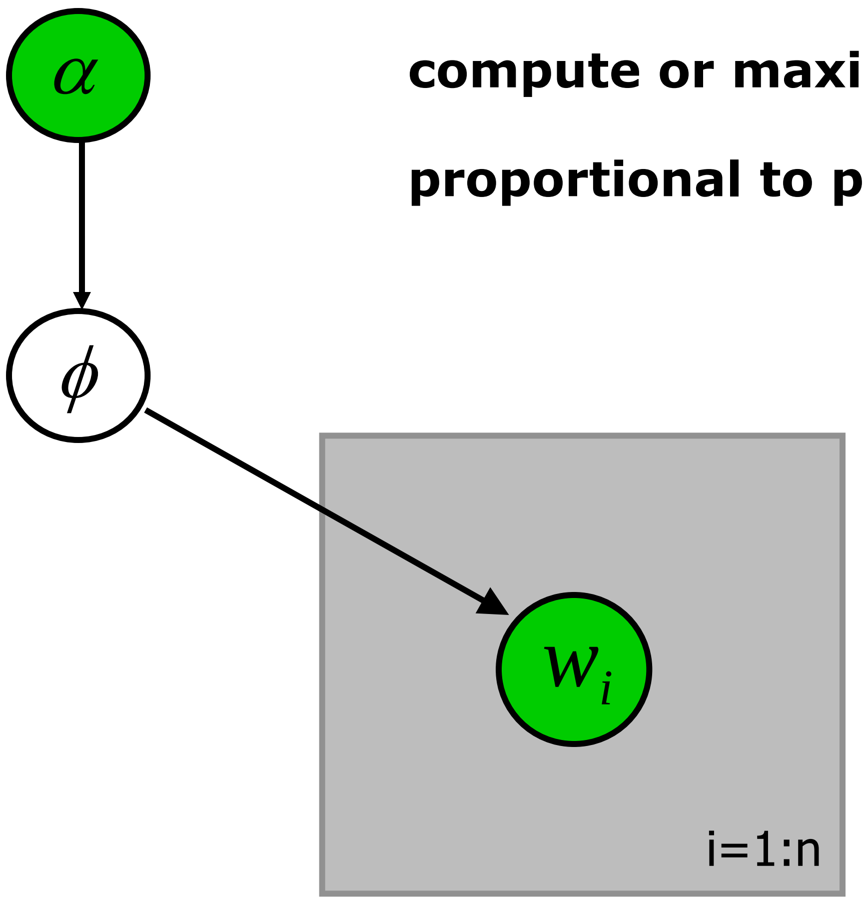
There are "n" conditionally
independent replicates
represented by the plate

Being Bayesian....

This is a prior distribution on our multinomial parameters, e.g., a simple Dirichlet smoothing prior with symmetric parameter α to avoid estimates of probabilities that are 0



Being Bayesian....



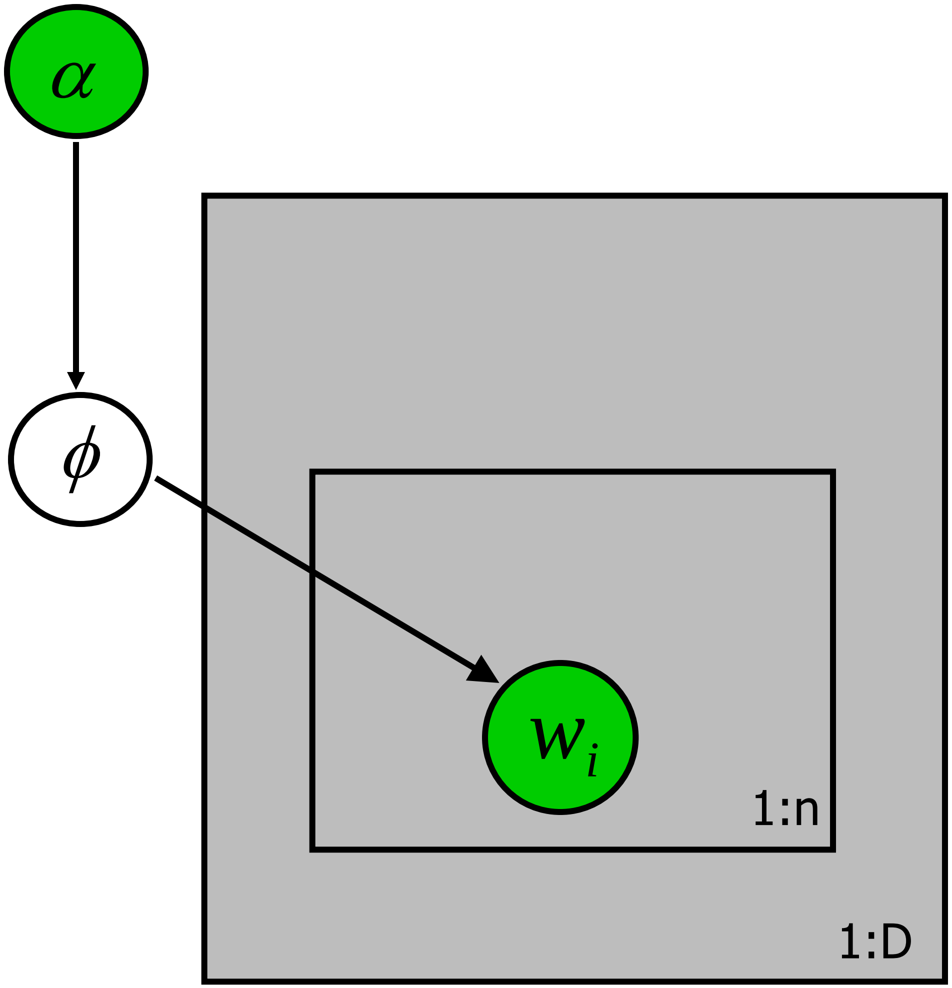
Learning:

compute or maximize p(ϕ | words, α)

proportional to p(words | ϕ) p(ϕ|α)

Multiple Documents

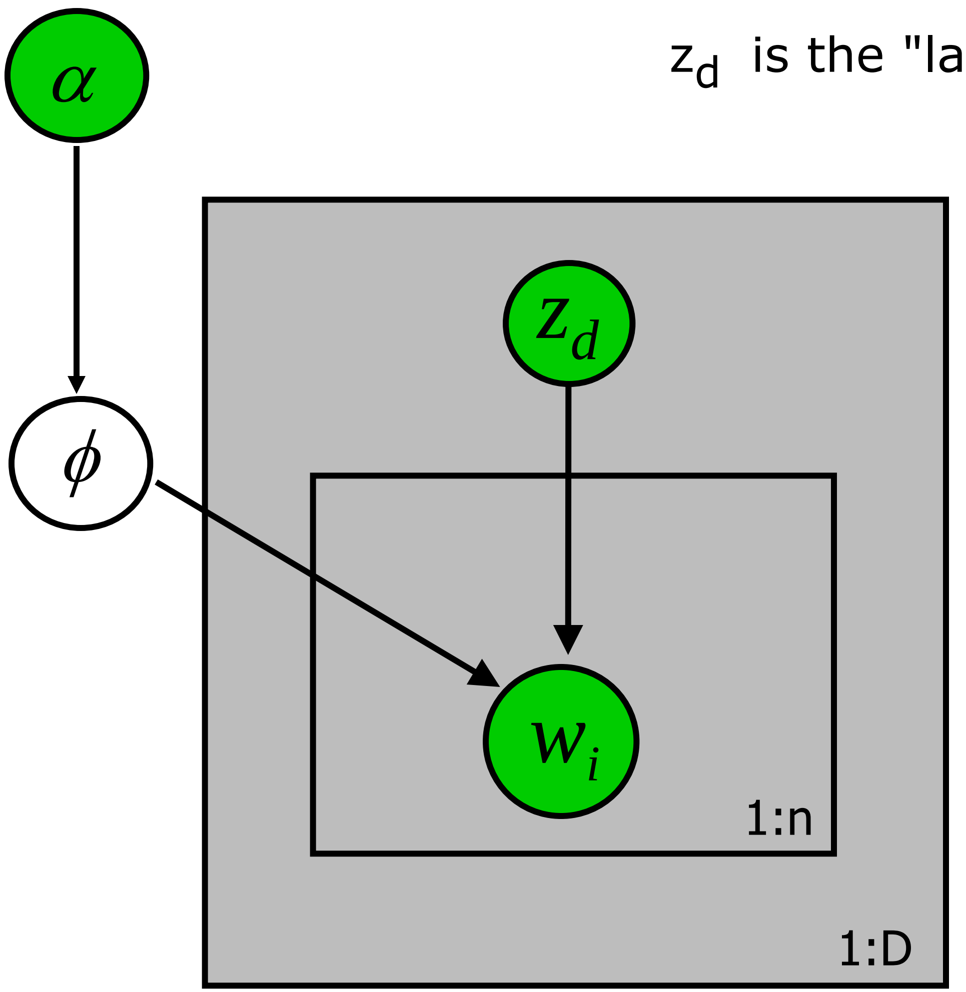
$$p(\text{ corpus } | \phi) = \Pi \; p(\text{ doc } | \phi)$$



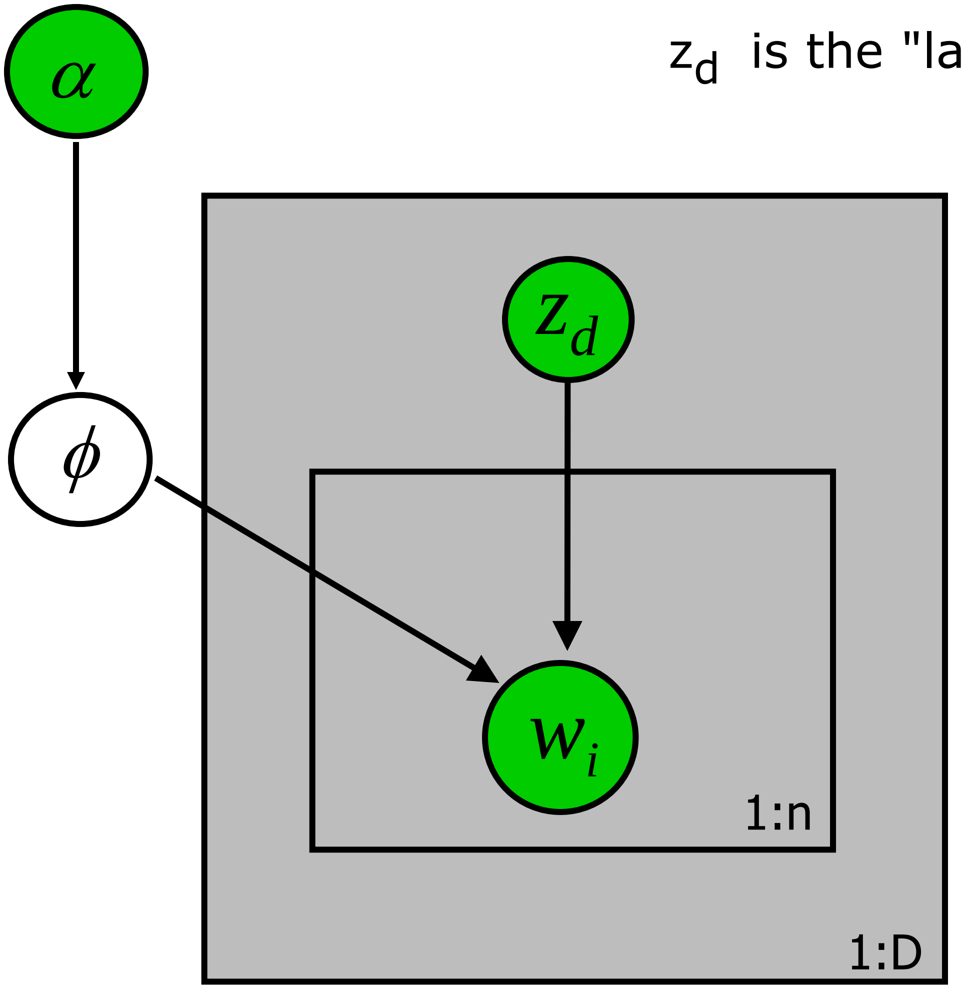
Different Document Types

p(w | ϕ , z_d) is a multinomial over words

z_d is the "label" for each doc



Different Document Types



$p(w \mid \phi, z_d)$ is a multinomial over words

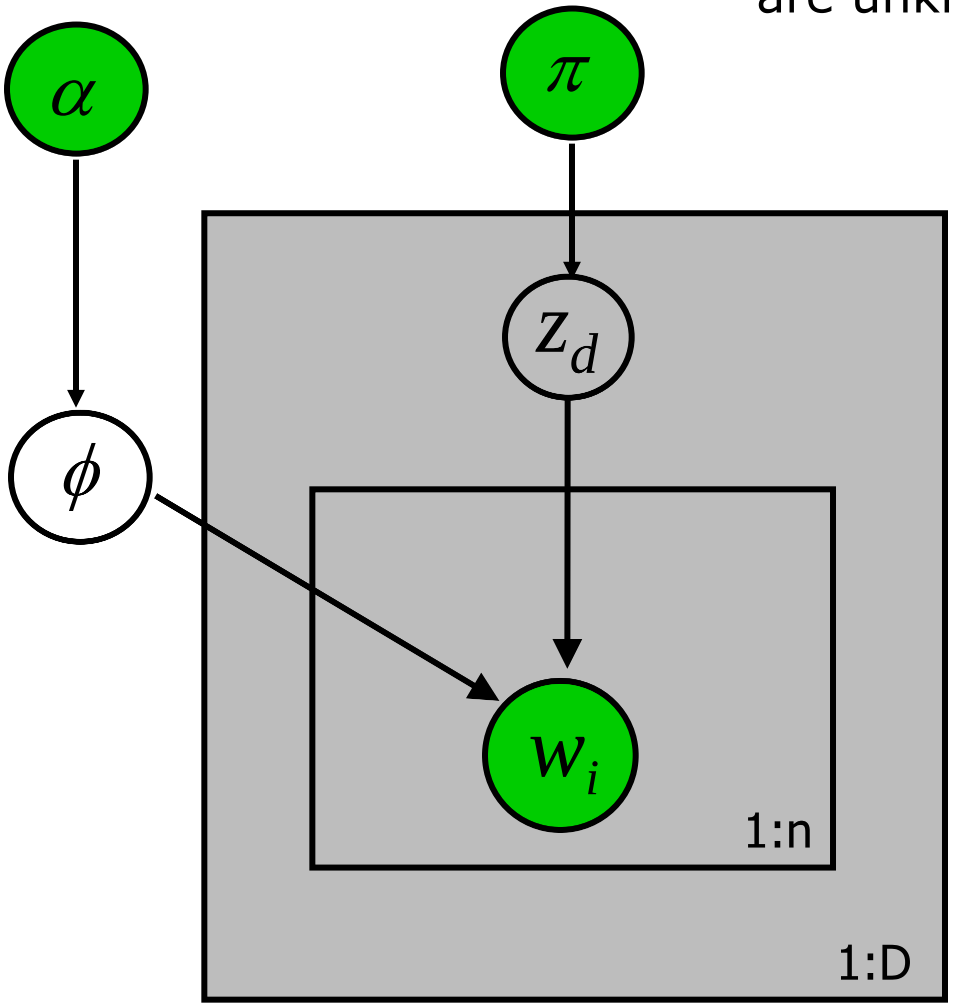
z_d is the "label" for each doc

Different multinomials, depending on the value of z_d (discrete)

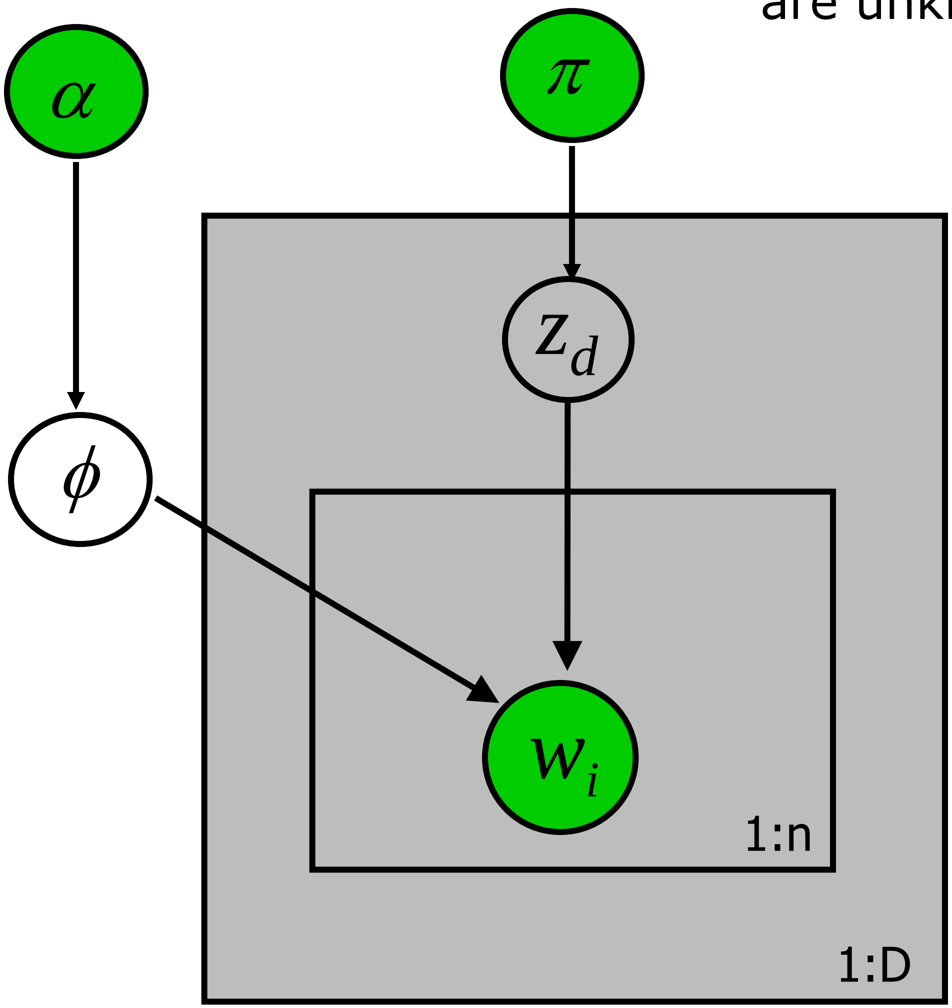
ϕ now represents $|z|$ different multinomials

Unknown Document Types

Now the values of z for each document are unknown - hopeless?



Unknown Document Types



Now the values of z for each document are unknown - hopeless?

Not hopeless :)

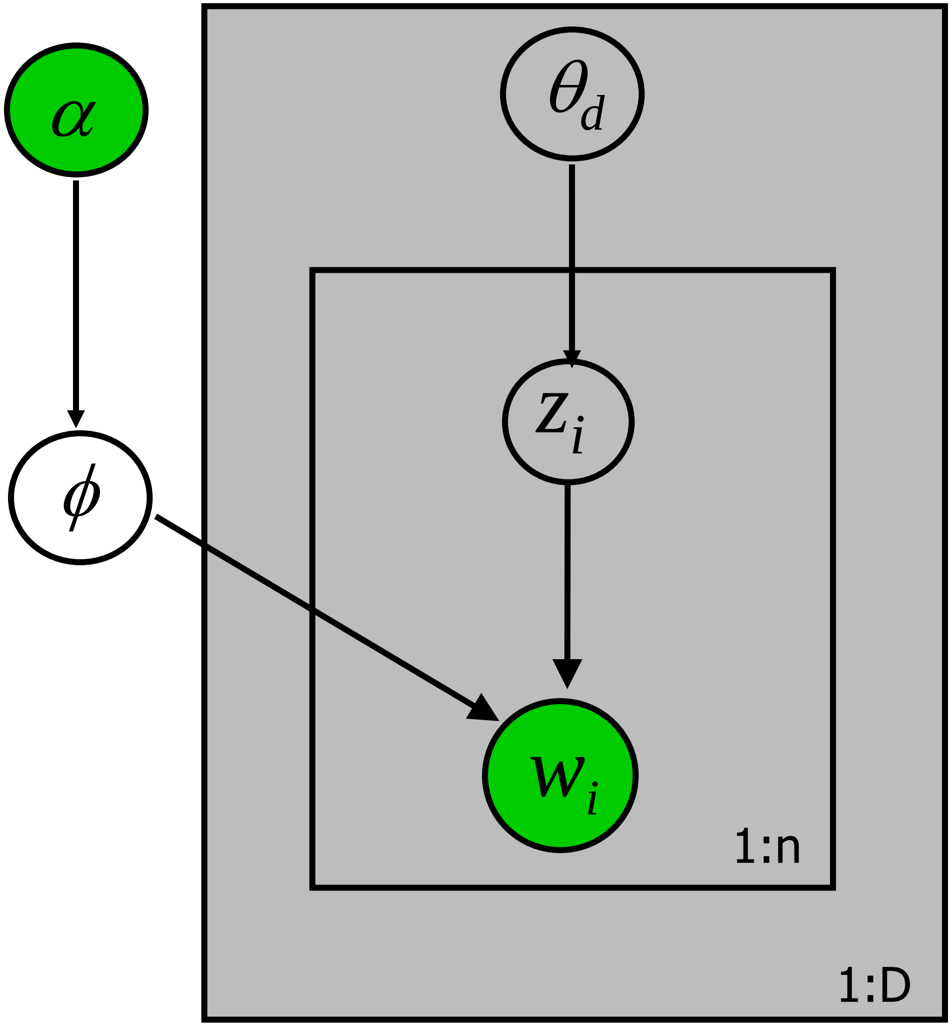
Can learn about both z and θ

e.g., EM algorithm

This gives probabilistic clustering

p(w | z=k , ϕ) is the kth multinomial over words

Topic Model



z_i is a "label" for each <u>word</u>

p(w | φ , z_i = k)
= multinomial over words

= a "topic"

p(z_i | θ_d) = distribution over topics that is document specific

Mixture Model Equation for Topic Models

$$p(w_i|d) = \sum_{j=1}^{T} p(w_i|z_j)p(z_j|d)$$

Multinomial over words for topic z (the ϕ's)

Multinomial over topics for document d (the θ's)

Key Features of Topic Models

- Generative model for documents in form of bags of words

- Allows a document to be composed of multiple topics
  - More flexible than clustering which assumes 1 cluster for each document

- Completely unsupervised
  - Topics learned directly from data
  - Leverages strong dependencies at word level AND large data sets

- Learning algorithm
  - Collapsed Gibbs sampling is the method of choice

- Scalable
  - Linear in number of word tokens
  - Can be run on millions of documents

Learning the Model

- Three sets of latent variables we can learn
  - topic-word distributions ϕ
  - document-topic distributions ϑ
  - topic assignments for each word z

- Options:
  - EM algorithm to find point estimates of ϕ and θ
    - e.g., Chien and Wu, IEEE Trans ASLP, 2008
  - Gibbs sampling
    - Find $p(\phi \mid \text{data})$, $p(\theta \mid \text{data})$, $p(z \mid \text{data})$
    - Can be slow to converge
  - Collapsed Gibbs sampling
    - Most widely used method

[See also Asuncion, Welling, Smyth, Teh, UAI 2009 for additional discussion]

Gibbs Sampling

- Say we have 3 parameters x,y,z, and some data

- Bayesian learning:
  - We want to compute p(x, y, z | data)
  - But frequently it is impossible to compute this exactly
  - However, often we can compute conditionals for individual variables, e.g.,

    p(x | y, z, data)

  - Not clear how this is useful yet, since it assumes y and z are known (i.e., we condition on them).

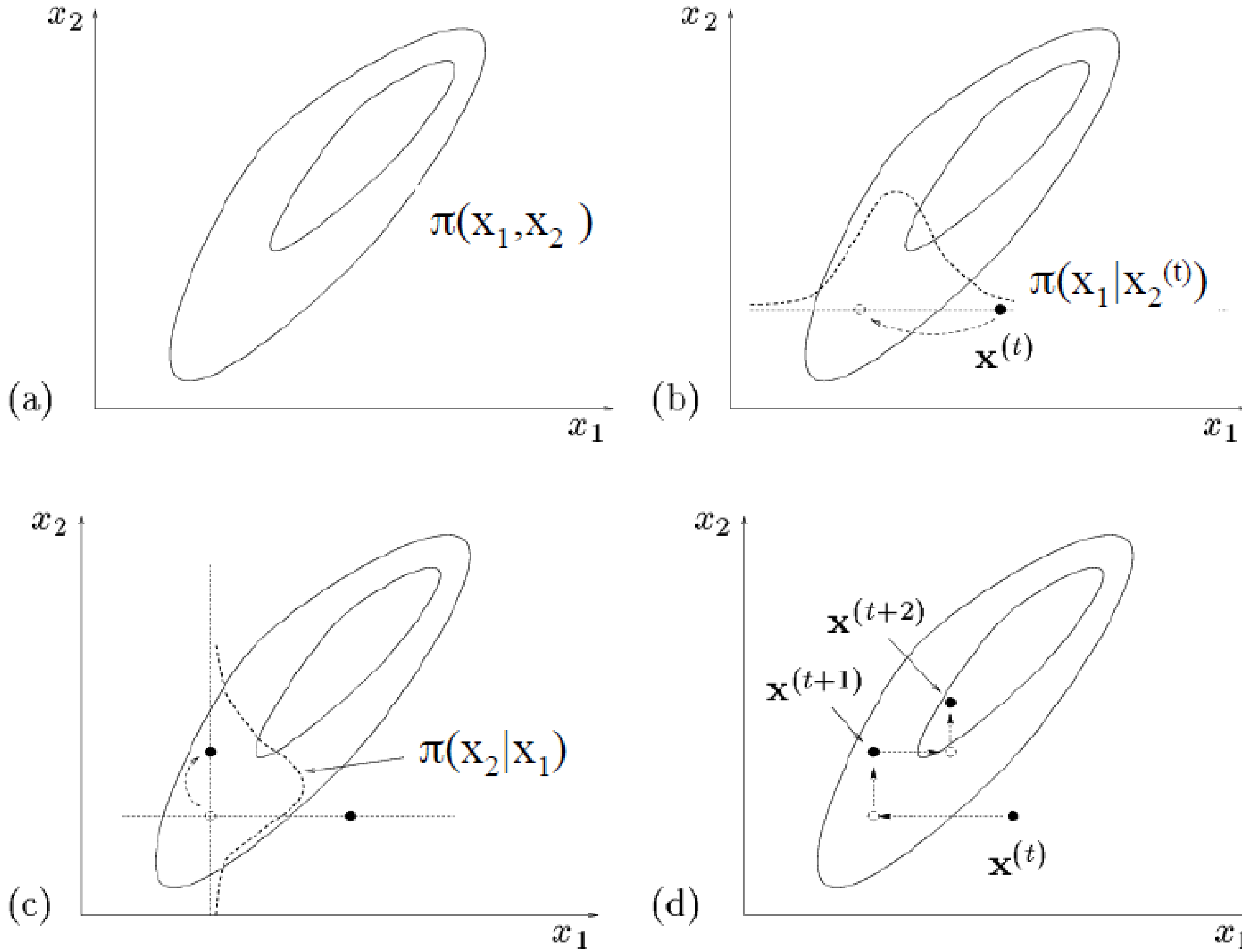
Gibbs Sampling 2

- Example of Gibbs sampling:
  - Initialize with x', y', z' (e.g., randomly)
  - Iterate:
    - Sample new x' ~ P(x | y', z', data)
    - Sample new y' ~ P(y | x', z', data)
    - Sample new z' ~ P(z | x', y', data)
  - Continue for some number (large) of iterations
  - Each iteration consists of a sweep through the hidden variables or parameters (here, x, y, and z)
  - Gibbs = a Markov Chain Monte Carlo (MCMC) method

In the limit, the samples x', y', z' will be samples from the true joint distribution P(x , y, z | data)

This gives us an *empirical estimate* of P(x , y, z | data)

Example of Gibbs Sampling in 2d



From online MCMC tutorial notes by Frank Dellaert, Georgia Tech

Gibbs Sampling for the Topic Model

- Recall: 3 sets of latent variables we can learn
  - topic-word distributions ϕ
  - document-topic distributions ϑ
  - topic assignments for each word z

- Gibbs sampling algorithm
  - Initialize all the z's randomly to a topic, z_1, ….. z_N
  - Iteration
    - For i = 1,…. N
      - Sample $z_i \sim p(z_i$ | all other z's, data)
  - Continue for a fixed number of iterations or convergence

  - Note that this is collapsed Gibbs sampling
    - Sample from $p(z_1, ….. z_N$ | data), "collapsing" over ϕ and θ

Review of Topic Models.....

- A topic model represents:
  - Documents as probability distributions over topics
  - Topics as probability distributions over words

- The model is learned from bag-of-words data (docs x words count matrix) using unsupervised learning

- The most commonly used learning algorithm is based on a technique called Gibbs sampling

Topic Model Learning Algorithm

- Input:
  - N documents as count vectors, number of topics T

- Output:
  - T topics, i.e., T topic-word probability vectors
  - N sets of topic weights, one per document
  - Assignment of each word in each document to 1 of T topics

- Algorithm (based on Gibbs Sampling)
  - Randomly initialize all word tokens to a topic (a number from 1 to T)
  - For each iteration
    - Sample a new topic for each word token, keeping all other assignments fixed
    - Iterate through all word tokens in all documents
  - Typically converges quickly (20 to 100 iterations)
  - Each iteration is linear in the number of word tokens

Collapsed Gibbs Sampling Equation for Topic Modeling

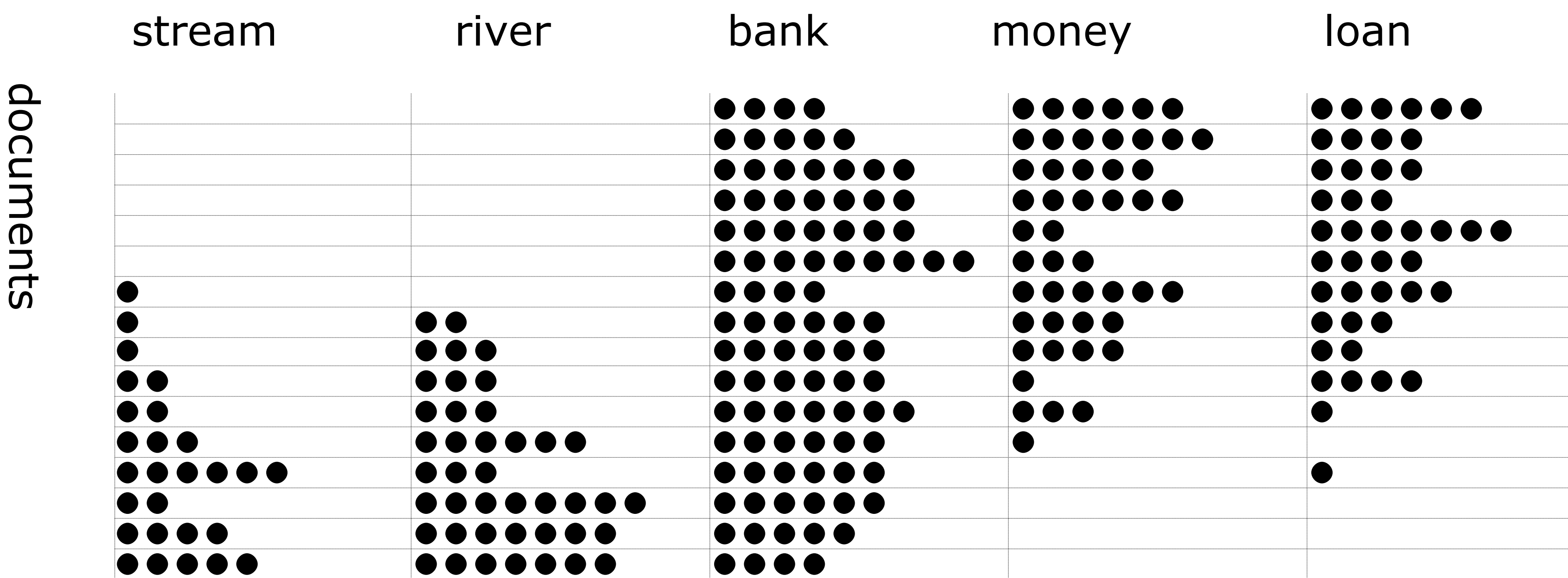
count of topic ***t*** assigned to doc ***d***

count of word ***w*** assigned to topic ***t***

$$p(z_i = t \mid z_{-i}) = \frac{n_{td}^{-i} + \alpha}{\sum_{t'} n_{t'd}^{-i} + T\alpha} \times \frac{n_{wt}^{-i} + \beta}{\sum_{w'} n_{w't}^{-i} + W\beta}$$

probability that word ***i*** is assigned to topic ***t***

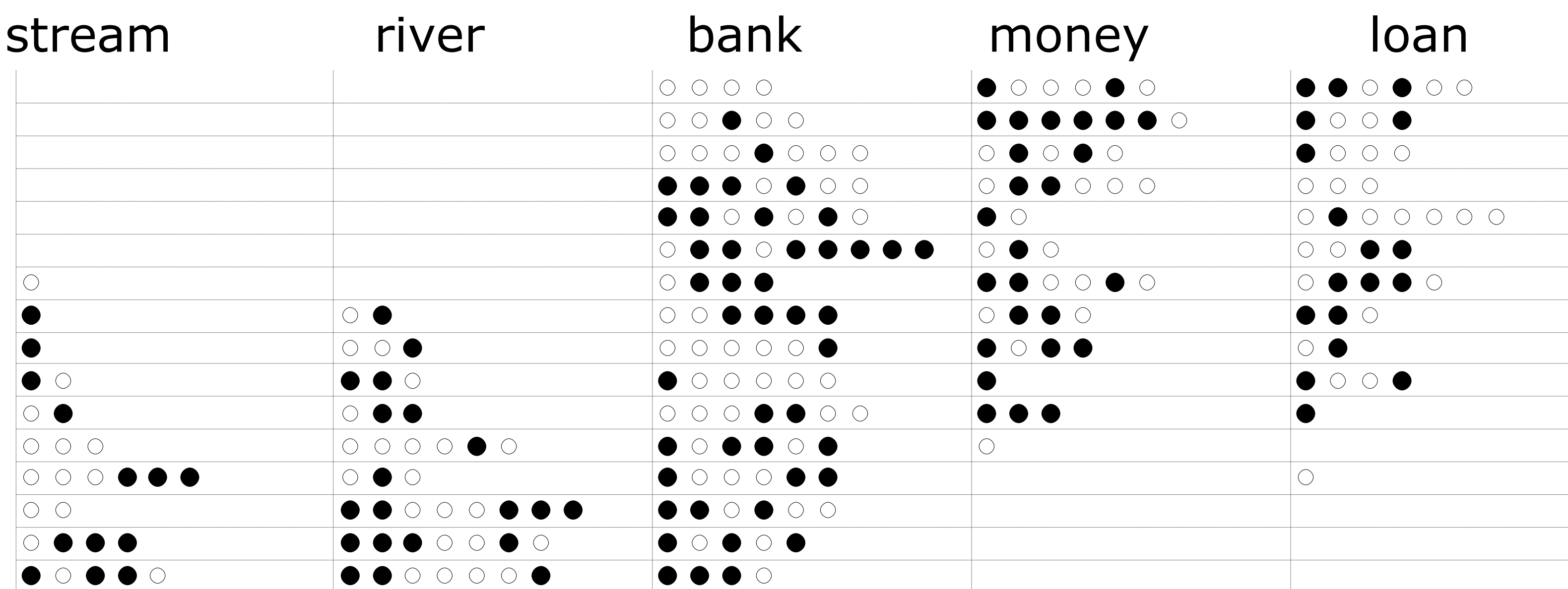
Word/Document counts for 16 Artificial Documents



Can we recover the original topics and topic mixtures from this data?

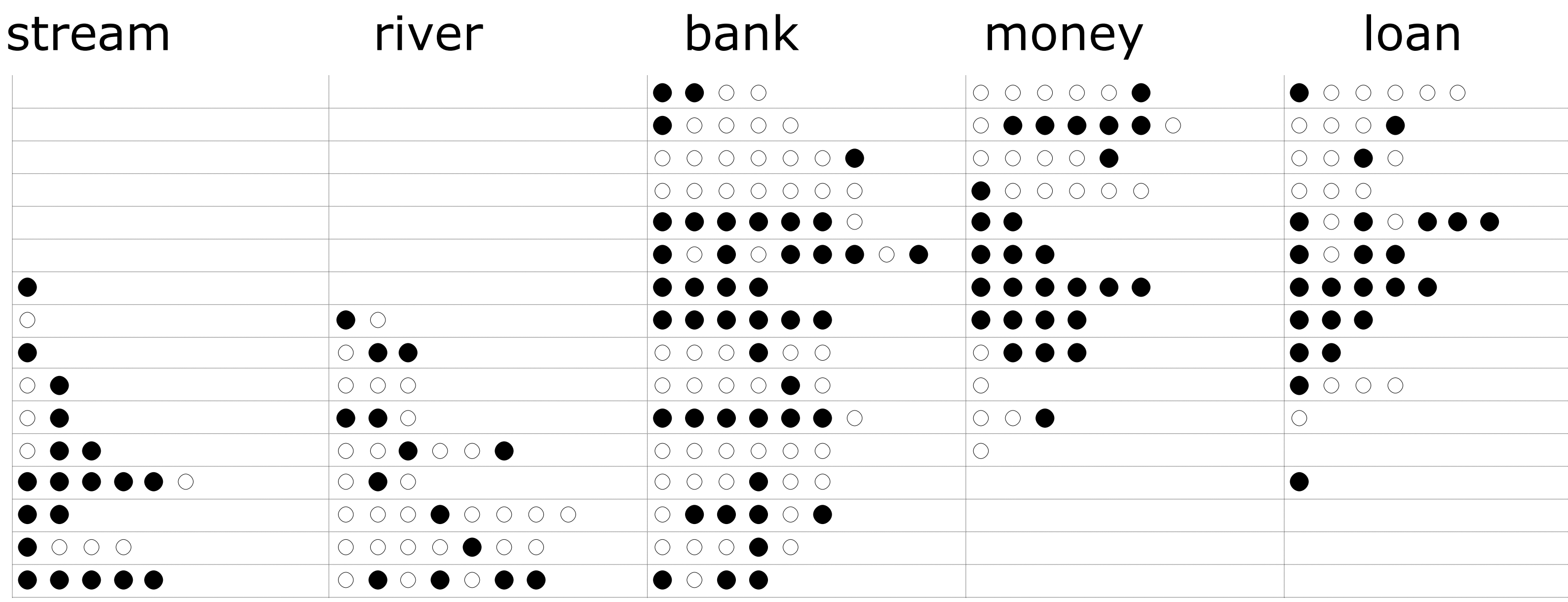
Example of Collapsed Gibbs Sampling

- Assign word tokens randomly to 2 topics:

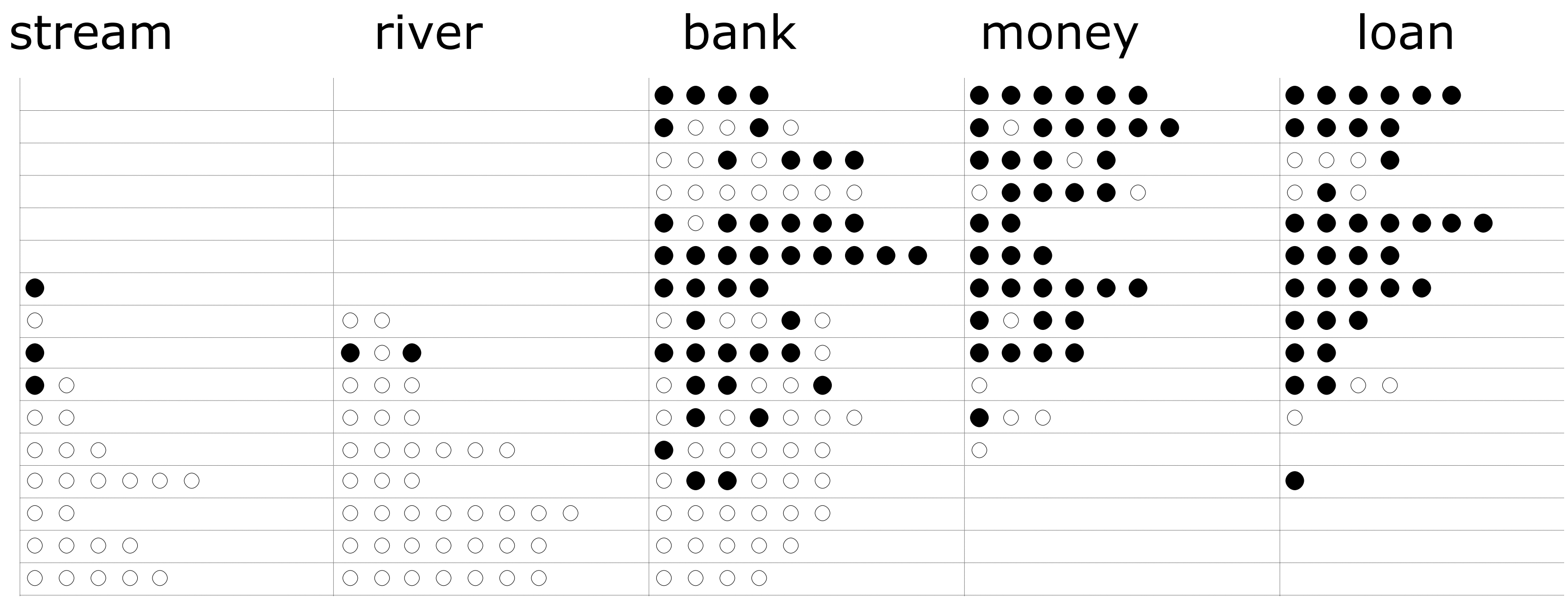


UCIrvine
University of California, Irvine

After 1 iteration



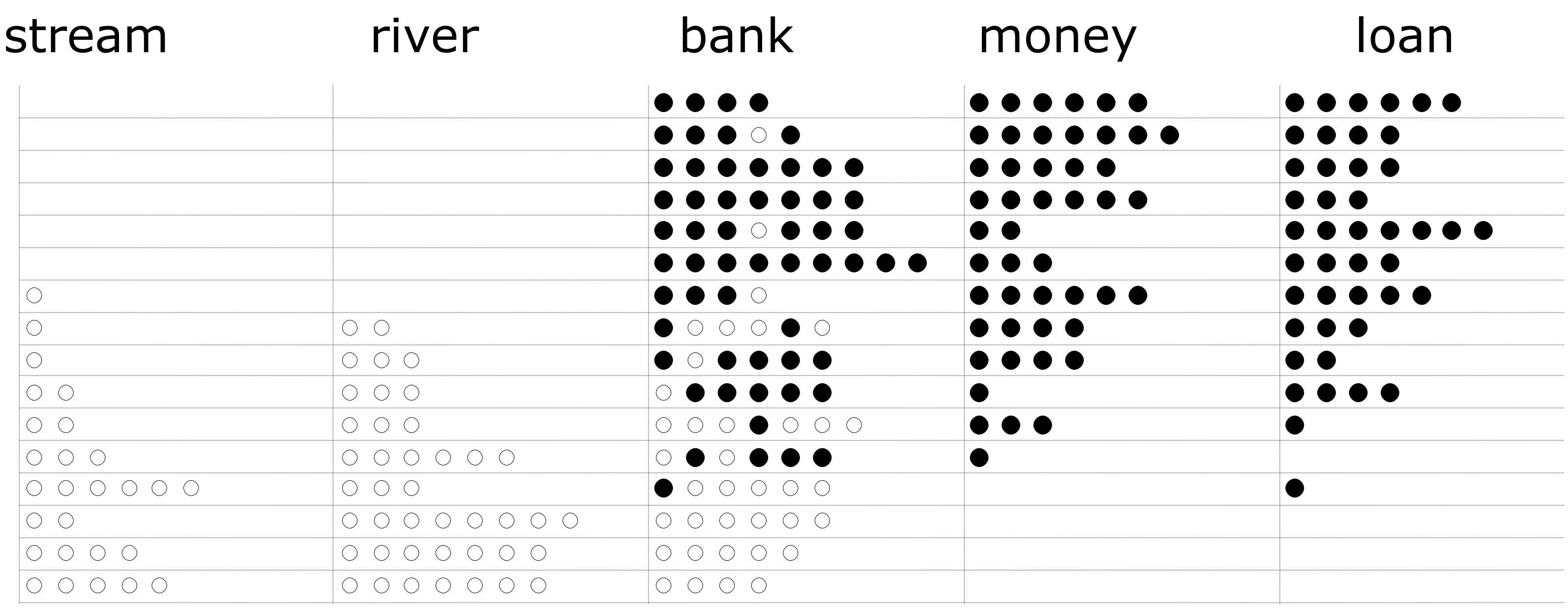
After 4 iterations



After 32 iterations

topic 1	
stream	.40
bank	.35
river	.25

topic 2	
bank	.39
money	.32
loan	.29



Computational Aspects

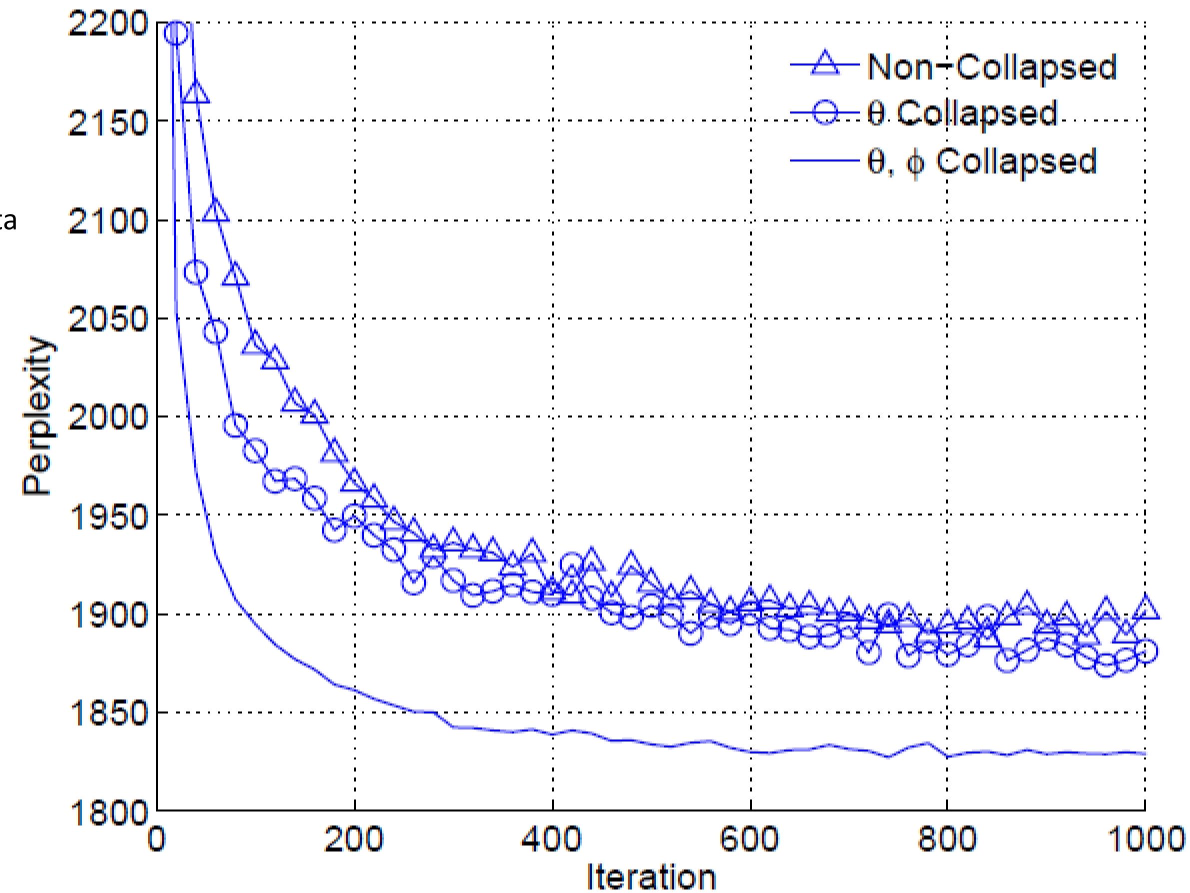
- Convergence
  - In the limit, samples x', y', z' are from P(x , y, z | data)
  - How many iterations are needed?
    - Cannot be computed ahead of time
    - Early iterations are discarded ("burn-in")
    - Typically monitor some quantities of interest to monitor convergence
    - Detecting convergence in Gibbs/MCMC is a tricky issue!

- Complexity per iteration
  - Linear in number of hidden variables and parameters
  - Times the complexity of generating a sample each time

Convergence Example

(from Newman et al, JMLR, 2009)

A measure of how well
the model is fitting the data



The New York Times

music	book	art	game	show
band	life	museum	knicks	film
songs	novel	show	nets	television
rock	story	exhibition	points	movie
album	books	artist	team	series
jazz	man	artists	season	says
pop	stories	paintings	play	life
song	love	painting	games	man
singer	children	century	night	character
night	family	works	coach	know

theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

Each box contains the high-probability words from each topic

Note that this is entirely unsupervised – no human labeling, just the words

Example from Hoffman et al, 2013

3 of 300 example topics (from TASA Corpus)

Example from Mark Steyvers

TOPIC 82		TOPIC 77		TOPIC 166	
WORD	PROB.	WORD	PROB.	WORD	PROB.
PLAY	0.0601	MUSIC	0.0903	PLAY	0.1358
PLAYS	0.0362	DANCE	0.0345	BALL	0.1288
STAGE	0.0305	SONG	0.0329	GAME	0.0654
MOVIE	0.0288	PLAY	0.0301	PLAYING	0.0418
SCENE	0.0253	SING	0.0265	HIT	0.0324
ROLE	0.0245	SINGING	0.0264	PLAYED	0.0312
AUDIENCE	0.0197	BAND	0.0260	BASEBALL	0.0274
THEATER	0.0186	PLAYED	0.0229	GAMES	0.0250
PART	0.0178	SANG	0.0224	BAT	0.0193
FILM	0.0148	SONGS	0.0208	RUN	0.0186
ACTORS	0.0145	DANCING	0.0198	THROW	0.0158
DRAMA	0.0136	PIANO	0.0169	BALLS	0.0154
REAL	0.0128	PLAYING	0.0159	TENNIS	0.0107
CHARACTER	0.0122	RHYTHM	0.0145	HOME	0.0099
ACTOR	0.0116	ALBERT	0.0134	CATCH	0.0098
ACT	0.0114	MUSICAL	0.0134	FIELD	0.0097
MOVIES	0.0114	DRUM	0.0129	PLAYER	0.0096
ACTION	0.0101	GUITAR	0.0098	FUN	0.0092
SET	0.0097	BEAT	0.0097	THROWING	0.0083
SCENES	0.0094	BALLET	0.0096	PITCHER	0.0080

Topic Modeling on different text sources …

Collection	# docs	Description
New York Times	1,500,000	News articles from New York Times
Austen	1,400	The six Jane Austen novels, broken up into 100-line sections
Blogs	4,000	Blog entries harvested from Daily Kos
Bible	1,200	Chapters in the bible (KJV)
Police Reports	250,000	Police accident reports from North Carolina
CiteSeer	750,000	Abstracts from research publications in computer science and engineering
Search Queries	1,000,000	Queries issued to web search engine
Enron	250,000	Enron emails seized by the US Government for the federal case against the company

… sample topics

Collection	Sample Topic
New York Times	[WMD] IRAQ iraqi weapon war SADDAM_HUSSEIN SADDAM resolution UNITED_STATES military inspector U_N UNITED_NATION BAGHDAD inspection action SECURITY_COUNCIL
Austen	[SENTIMENT] felt comfort feeling feel spirit mind heart ill evil fear impossible hope poor distress end loss relief suffering concern dreadful misery unhappy
Blogs	[ELECTIONS] november poll house electoral governor polls account ground republicans trouble
Bible	[COMMANDS] thou thy thee shalt thine lord god hast unto not shall
Police Reports	[RAN OFF ROAD] v1 off road ran came rest ditch traveling struck side shoulder tree overturned control lost
CiteSeer	[GRAPH THEORY] graph edge vertices edges vertex number directed connected degree coloring subgraph set drawing
Search Queries	[CREDIT] credit card loans bill loan report bad visa debt score
Enron	[ENERGY CRISIS] state california power electricity utilities davis energy prices generators edison public deregulation billion governor federal consumers commission plants companies electric wholesale crisis summer

Enron Email Topics

TOPIC 36	
WORD	**PROB.**
FEEDBACK	0.0781
PERFORMANCE	0.0462
PROCESS	0.0455
PEP	0.0446
MANAGEMENT	0.03
COMPLETE	0.0205
QUESTIONS	0.0203
SELECTED	0.0187
COMPLETED	0.0146
SYSTEM	0.0146

TOPIC 72	
WORD	**PROB.**
PROJECT	0.0514
PLANT	0.028
COST	0.0182
CONSTRUCTION	0.0169
UNIT	0.0166
FACILITY	0.0165
SITE	0.0136
PROJECTS	0.0117
CONTRACT	0.011
UNITS	0.0106

TOPIC 54	
WORD	**PROB.**
FERC	0.0554
MARKET	0.0328
ISO	0.0226
COMMISSION	0.0215
ORDER	0.0212
FILING	0.0149
COMMENTS	0.0116
PRICE	0.0116
CALIFORNIA	0.0110
FILED	0.0110

TOPIC 23	
WORD	**PROB.**
ENVIRONMENTAL	0.0291
AIR	0.0232
MTBE	0.019
EMISSIONS	0.017
CLEAN	0.0143
EPA	0.0133
PENDING	0.0129
SAFETY	0.0104
WATER	0.0092
GASOLINE	0.0086

"Personal" Topics…

TOPIC 66	
WORD	**PROB.**
HOLIDAY	0.0857
PARTY	0.0368
YEAR	0.0316
SEASON	0.0305
COMPANY	0.0255
CELEBRATION	0.0199
ENRON	0.0198
TIME	0.0194
RECOGNIZE	0.019
MONTH	0.018

TOPIC 182	
WORD	**PROB.**
TEXANS	0.0145
WIN	0.0143
FOOTBALL	0.0137
FANTASY	0.0129
SPORTSLINE	0.0129
PLAY	0.0123
TEAM	0.0114
GAME	0.0112
SPORTS	0.011
GAMES	0.0109

TOPIC 113	
WORD	**PROB.**
GOD	0.0357
LIFE	0.0272
MAN	0.0116
PEOPLE	0.0103
CHRIST	0.0092
FAITH	0.0083
LORD	0.0079
JESUS	0.0075
SPIRITUAL	0.0066
VISIT	0.0065

TOPIC 109	
WORD	**PROB.**
AMAZON	0.0312
GIFT	0.0226
CLICK	0.0193
SAVE	0.0147
SHOPPING	0.0140
OFFER	0.0124
HOLIDAY	0.0122
RECEIVE	0.0102
SHIPPING	0.0100
FLOWERS	0.0099

Political Topics

TOPIC 18	
WORD	PROB.
POWER	0.0915
CALIFORNIA	0.0756
ELECTRICITY	0.0331
UTILITIES	0.0253
PRICES	0.0249
MARKET	0.0244
PRICE	0.0207
UTILITY	0.0140
CUSTOMERS	0.0134
ELECTRIC	0.0120

TOPIC 22	
WORD	PROB.
STATE	0.0253
PLAN	0.0245
CALIFORNIA	0.0137
POLITICIAN Y	0.0137
RATE	0.0131
BANKRUPTCY	0.0126
SOCAL	0.0119
POWER	0.0114
BONDS	0.0109
MOU	0.0107

TOPIC 114	
WORD	PROB.
COMMITTEE	0.0197
BILL	0.0189
HOUSE	0.0169
WASHINGTON	0.0140
SENATE	0.0135
POLITICIAN X	0.0114
CONGRESS	0.0112
PRESIDENT	0.0105
LEGISLATION	0.0099
DC	0.0093

TOPIC 194	
WORD	PROB.
LAW	0.0380
TESTIMONY	0.0201
ATTORNEY	0.0164
SETTLEMENT	0.0131
LEGAL	0.0100
EXHIBIT	0.0098
CLE	0.0093
SOCALGAS	0.0093
METALS	0.0091
PERSON Z	0.0083

Automated Tagging of Words

(numbers & colors → topic assignments)

Example from Mark Steyvers

A **Play**[082] is written[082] to be performed[082] on a stage[082] before a live[093] audience[082] or before motion[270] picture[004] or television[004] cameras[004] (for later[054] viewing[004] by large[202] audiences[082]). A **Play**[082] is written[082] because playwrights[082] have something

He was listening[077] to music[077] coming[009] from a passing[043] riverboat. The music[077] had already captured[006] his heart[157] as well as his ear[119]. It was jazz[077]. Bix beiderbecke had already had music[077] lessons[077]. He wanted[268] to **play**[077] the cornet. And he wanted[268] to **play**[077] jazz[077]

Jim[296] **plays**[166] the game[166]. Jim[296] likes[081] the game[166] for one. The game[166] book[254] helps[081] jim[296]. Don[180] comes[040] into the house[038]. Don[180] and jim[296] read[254] the game[166] book[254]. The boys[020] see a game[166] for two. The two boys[020] **play**[166] the game[166].

What is this paper about?

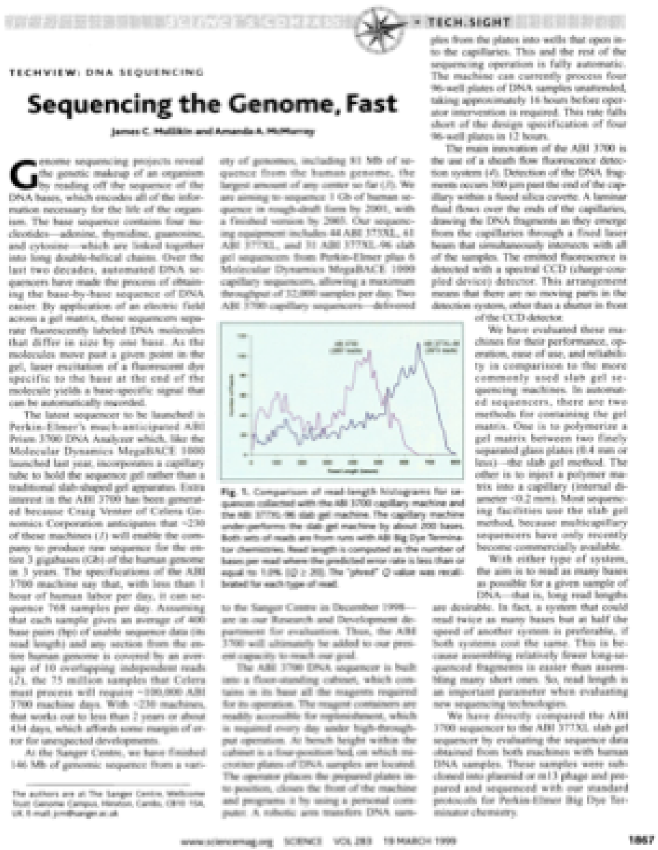
Empirical Bayes screening for multi-item associations

Bill DuMouchel and Daryl Pregibon, ACM SIGKDD 2001

Most likely topics according to the model are…

1. data, mining, discovery, association, attribute..
2. set, subset, maximal, minimal, complete,…
3. measurements, correlation, statistical, variation,
4. Bayesian, model, prior, data, mixture,…..

Original article


TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

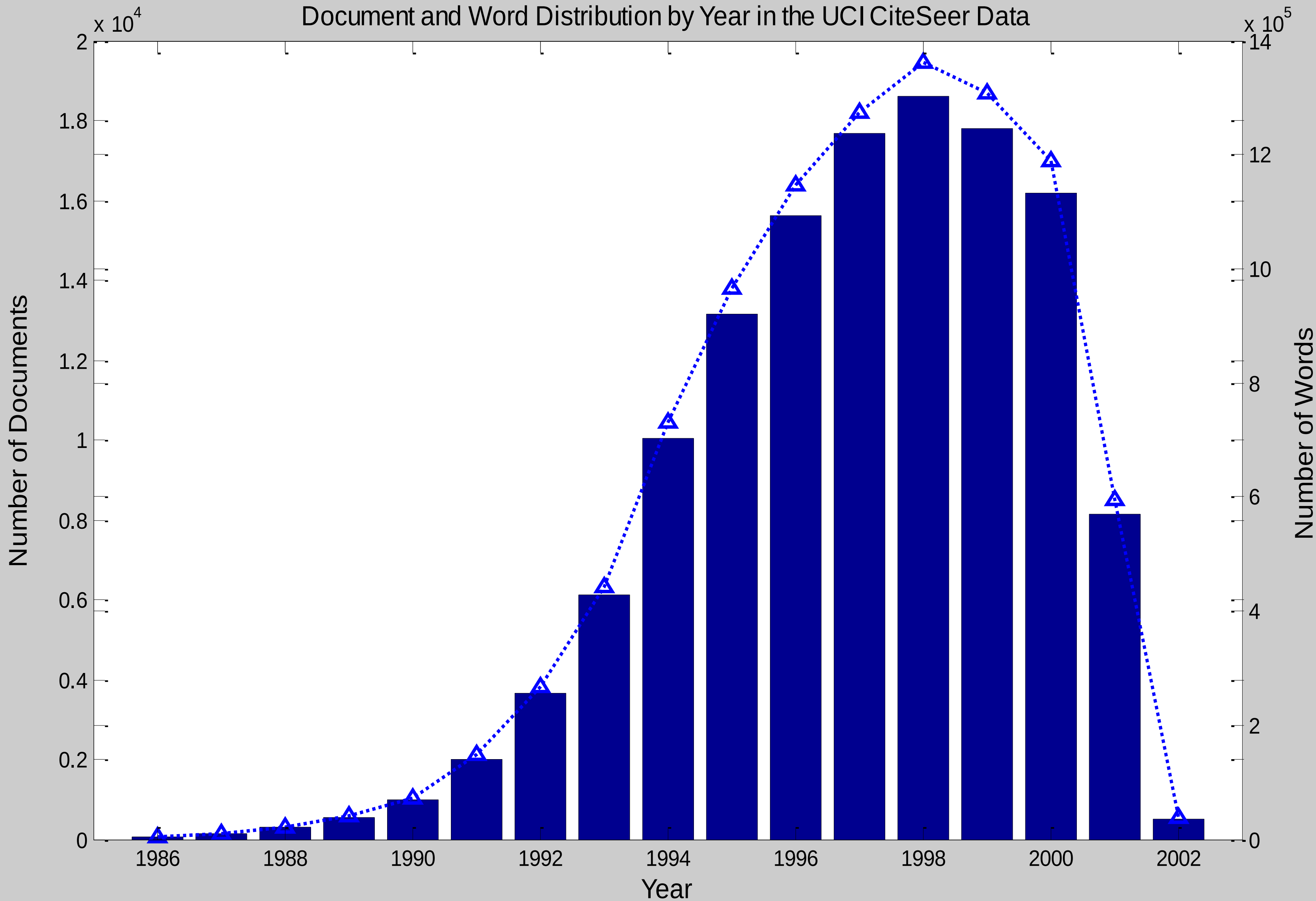
Most likely words from top topics

sequence	devices	data
genome	device	information
genes	materials	network
sequences	current	web
human	high	computer
gene	gate	language
dna	light	networks
sequencing	silicon	time
chromosome	material	software
regions	technology	system
analysis	electrical	words
data	fiber	algorithm
genomic	power	number
number	based	internet

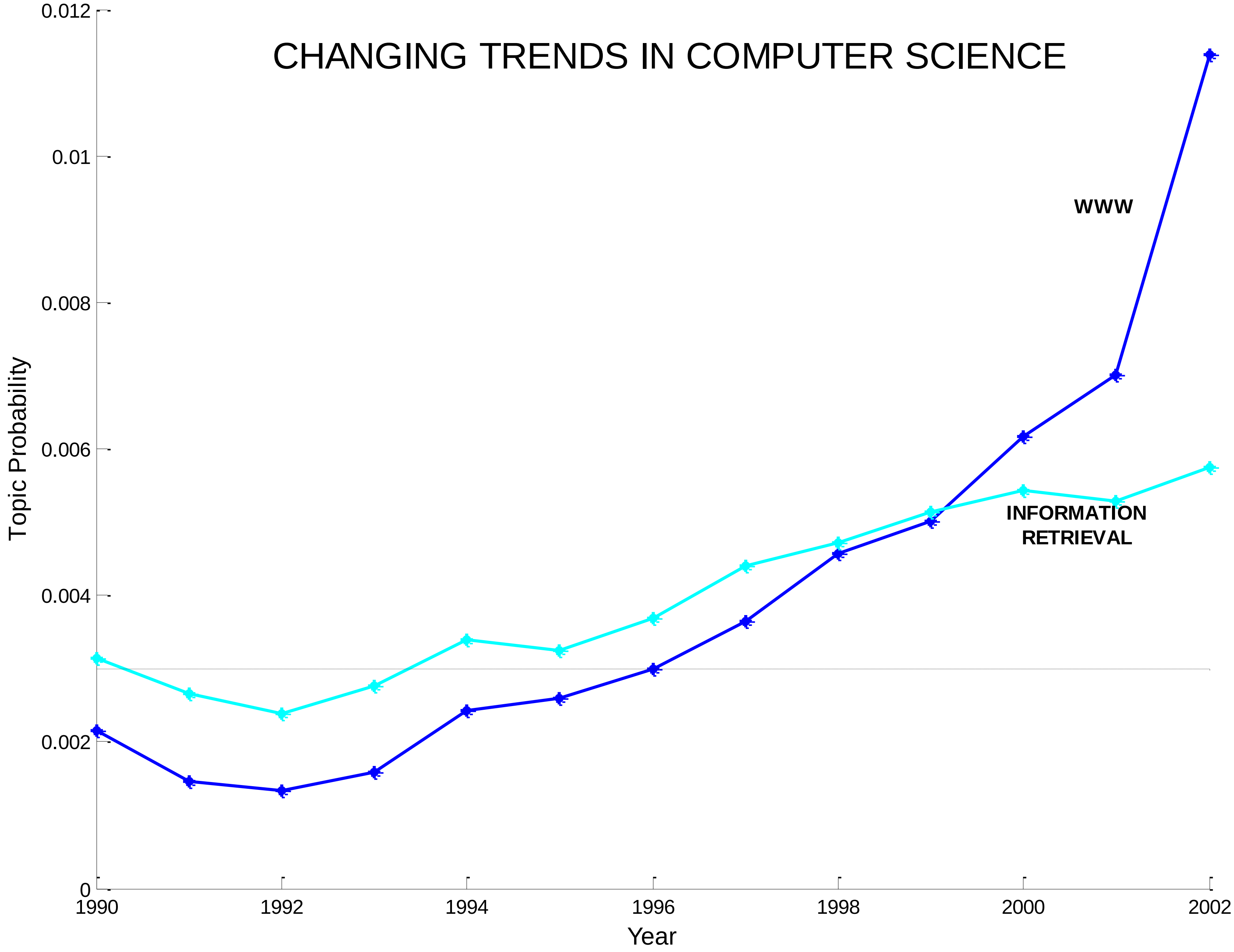
Example courtesy of David Blei, Princeton

Temporal patterns in Topics: Hot and Cold Topics

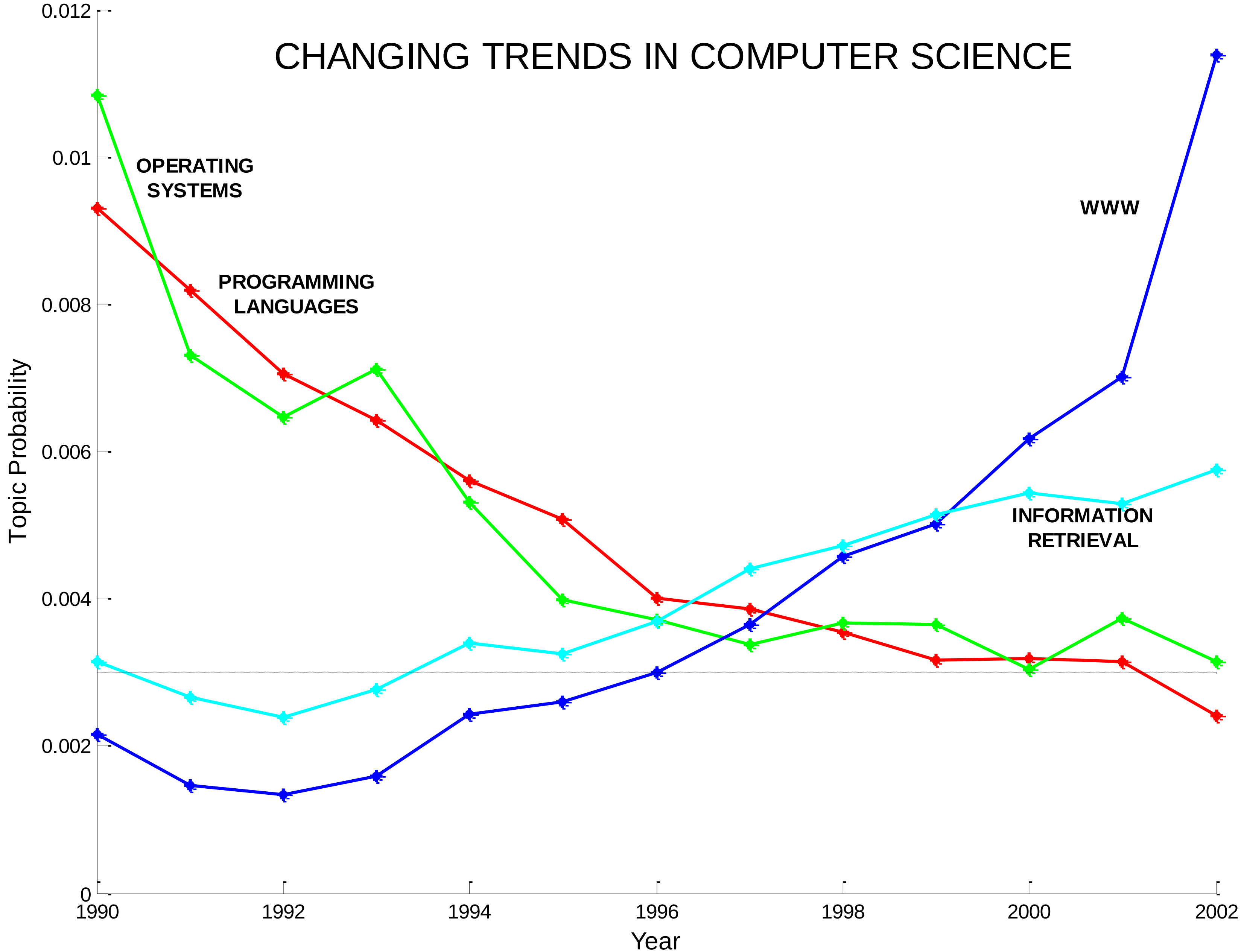
- CiteSeer papers from 1986-2002, about 200k papers
- For each year, calculate the fraction of words assigned to each topic
- This gives us time-series for topics
  - Hot topics become more prevalent
  - Cold topics become less prevalent

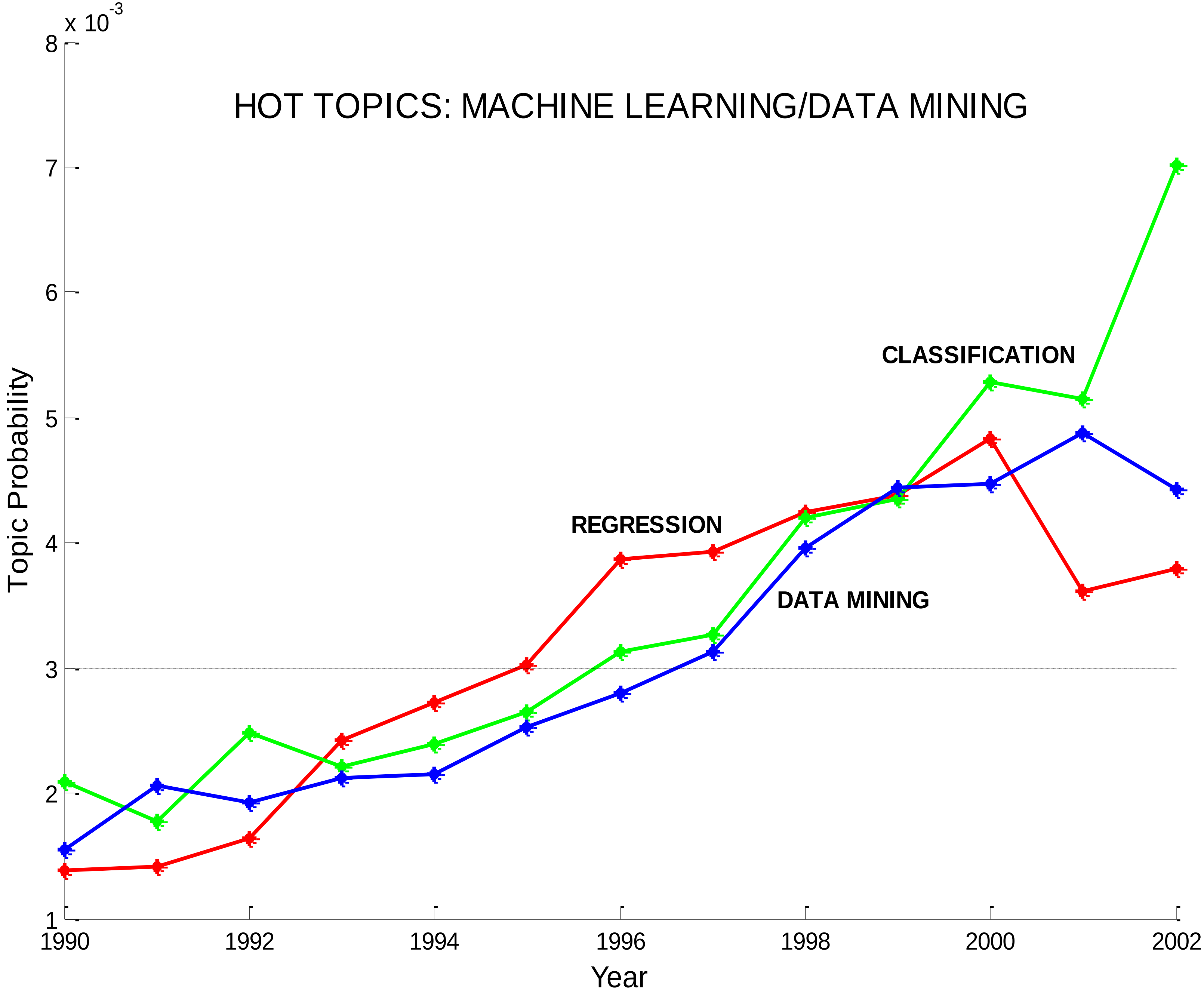


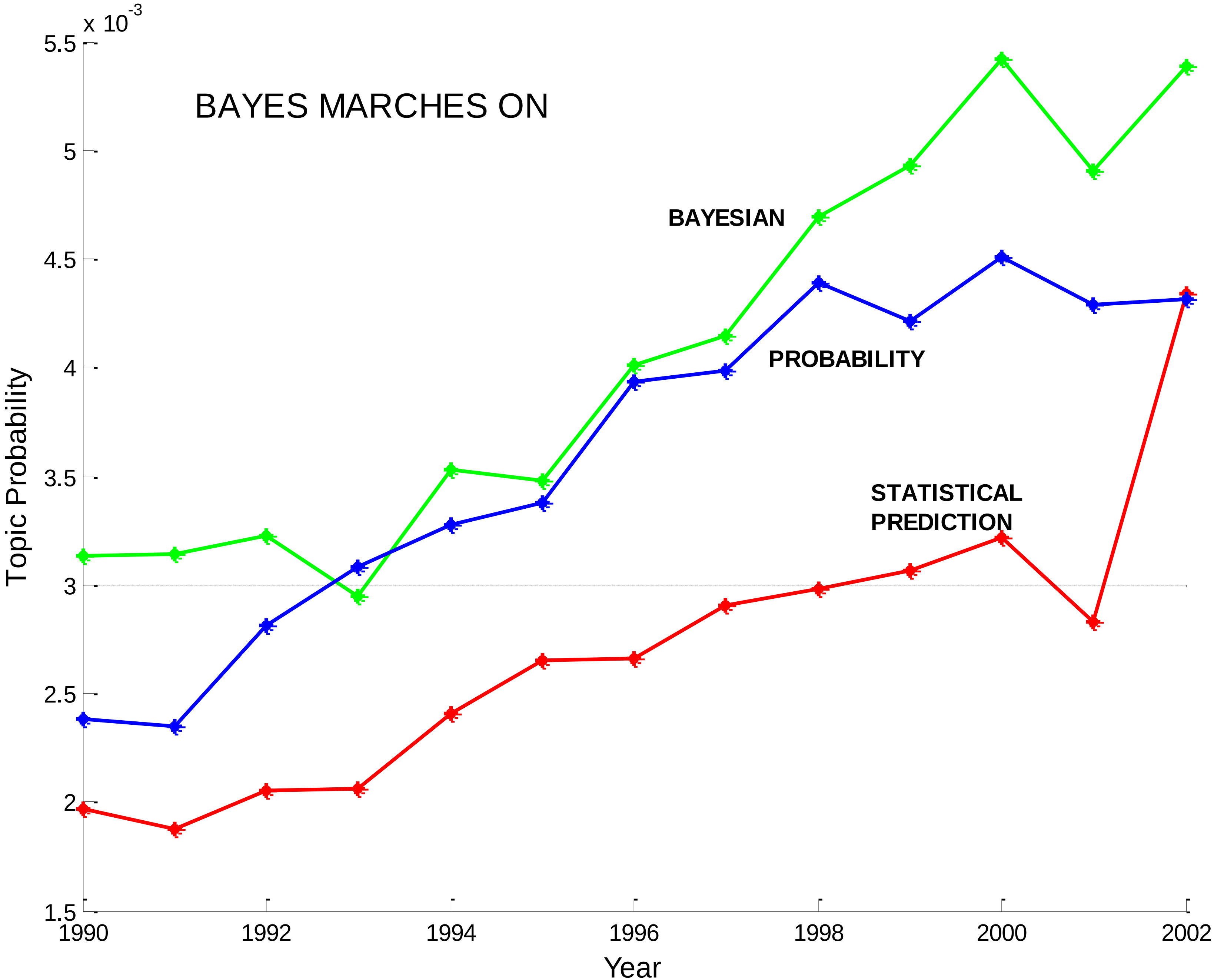
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

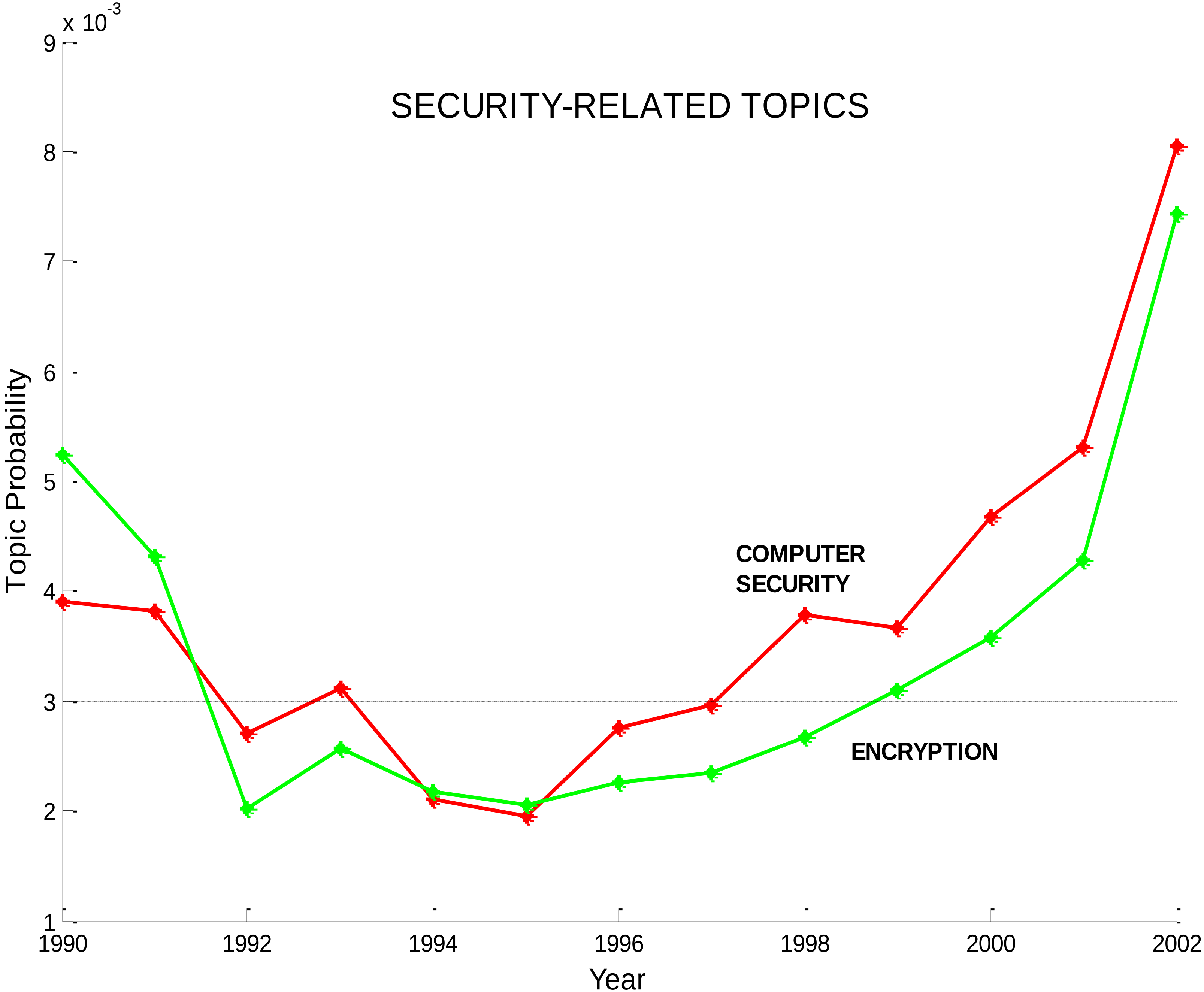


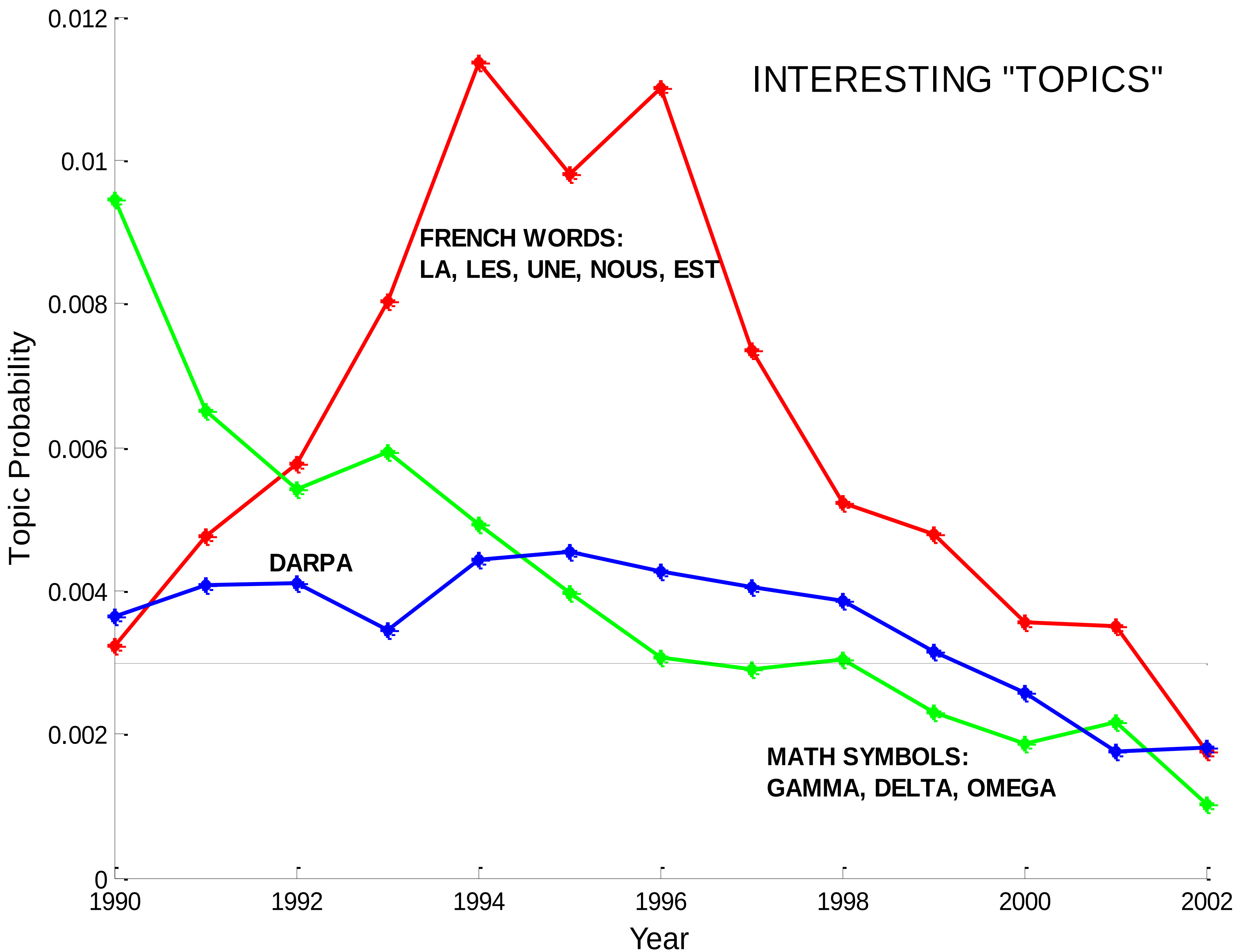
CHANGING TRENDS IN COMPUTER SCIENCE
WWW
INFORMATION RETRIEVAL
Topic Probability
Year
0.012
0.01
0.008
0.006
0.004
0.002
0
1990
1992
1994
1996
1998
2000
2002


CHANGING TRENDS IN COMPUTER SCIENCE
OPERATING SYSTEMS
PROGRAMMING LANGUAGES
WWW
INFORMATION RETRIEVAL
Topic Probability
Year
0
0.002
0.004
0.006
0.008
0.01
0.012
1990
1992
1994
1996
1998
2000
2002


x 10-3
8
7
6
5
4
3
2
1
HOT TOPICS: MACHINE LEARNING/DATA MINING
CLASSIFICATION
REGRESSION
DATA MINING
Topic Probability
1990
1992
1994
1996
1998
2000
2002
Year

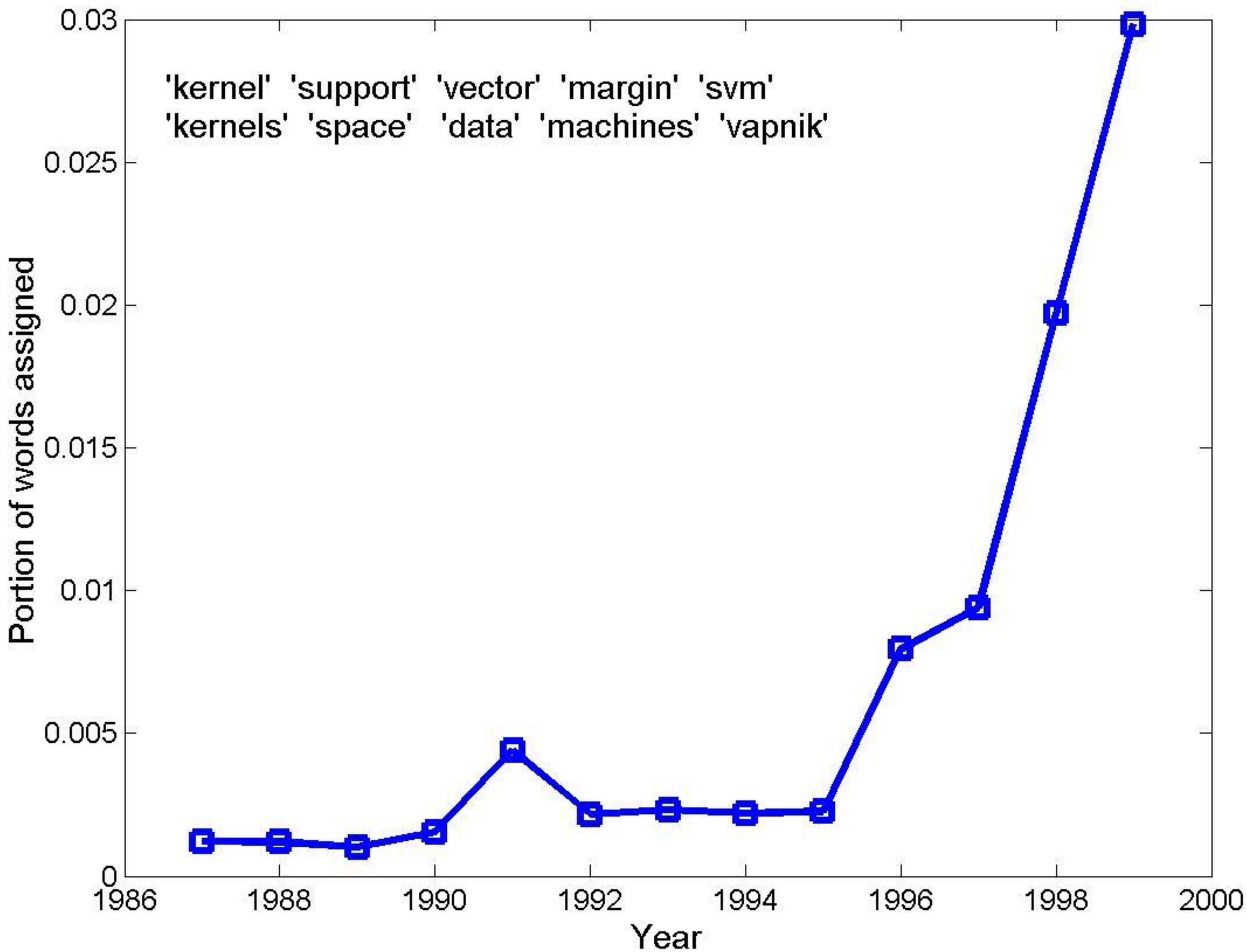

x 10^-3
BAYES MARCHES ON
BAYESIAN
PROBABILITY
STATISTICAL
PREDICTION
Topic Probability
Year
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1990
1992
1994
1996
1998
2000
2002


SECURITY-RELATED TOPICS
x 10^-3
9
8
7
6
5
4
3
2
1
Topic Probability
COMPUTER
SECURITY
ENCRYPTION
1990
1992
1994
1996
1998
2000
2002
Year

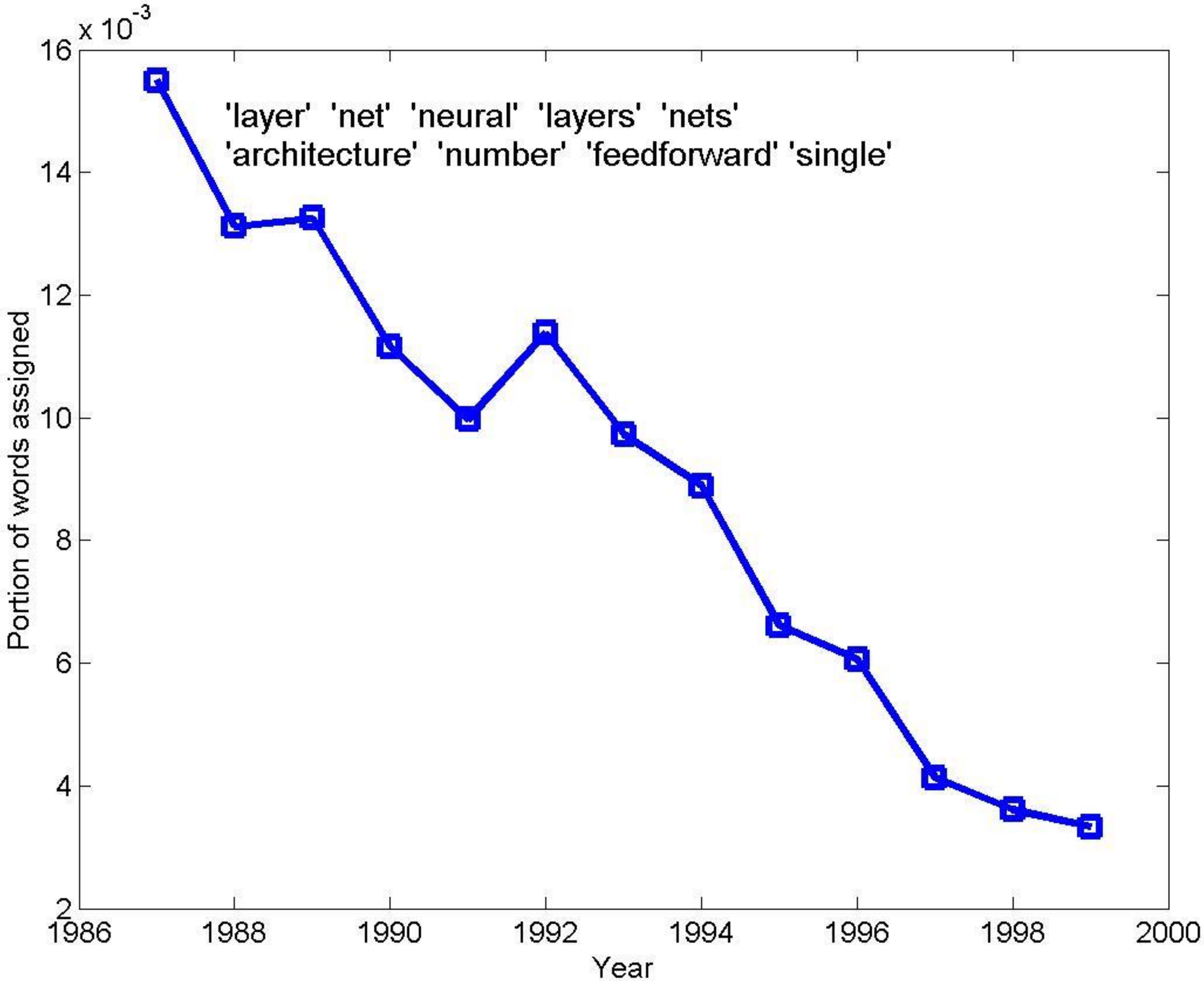


INTERESTING "TOPICS"
FRENCH WORDS:
LA, LES, UNE, NOUS, EST
DARPA
MATH SYMBOLS:
GAMMA, DELTA, OMEGA
Topic Probability
0.012
0.01
0.008
0.006
0.004
0.002
0
1990
1992
1994
1996
1998
2000
2002
Year

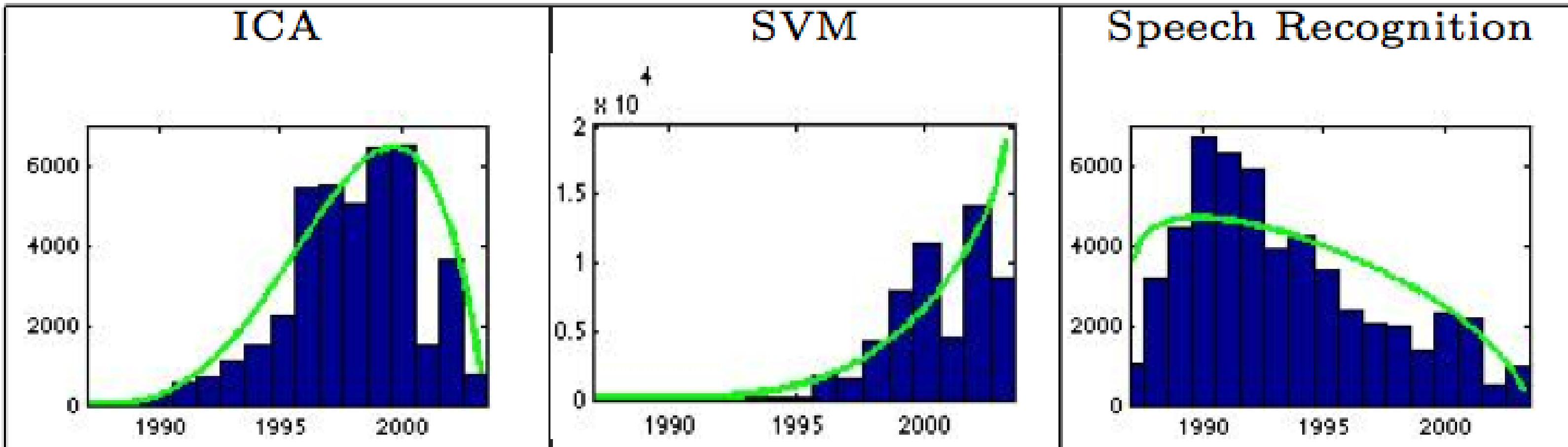
NIPS: SVM Topic



NIPS: neural network topic



Topics over Time for NIPS Proceedings



ICA		SVM		Speech Recognition	
ica	0.023	kernel	0.063	speech	0.055
independent	0.021	svm	0.025	recognition	0.039
natural	0.018	support	0.025	word	0.023
source	0.018	kernels	0.022	hmm	0.019
basis	0.017	vector	0.020	training	0.018
sources	0.015	training	0.020	system	0.017
separation	0.014	space	0.013	speaker	0.016
component	0.014	machines	0.011	context	0.013
information	0.014	feature	0.011	phoneme	0.009
matrix	0.013	data	0.010	state	0.008

Figure courtesy of Xuerie Wang and Andrew McCallum, U Mass Amherst

All NIPS Topics over Time

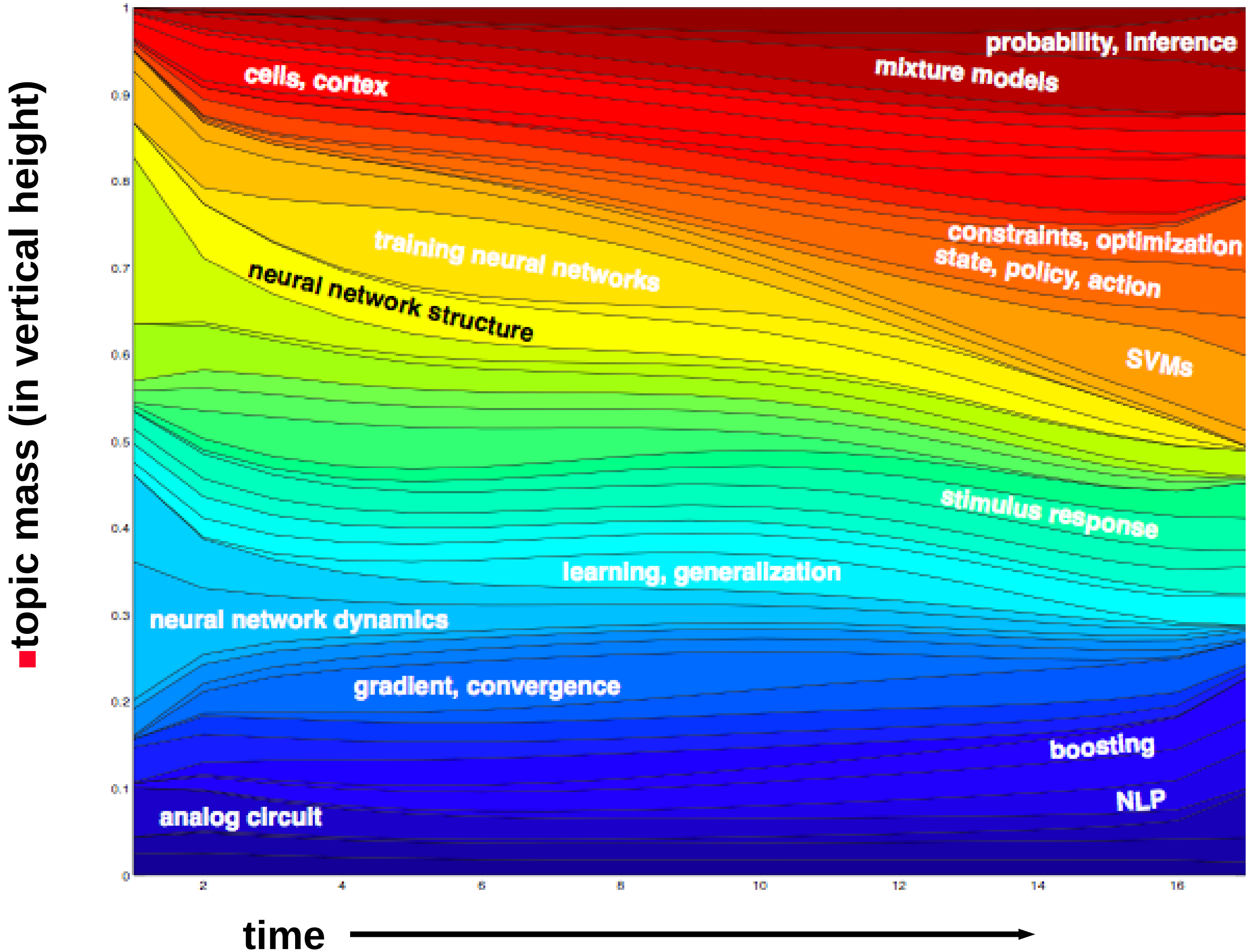


Figure courtesy of Xuerie Wang and Andrew McCallum, U Mass Amherst

Topics from the *Pennsylvania Gazette*


July 20. 1749.

Numb. 1075.

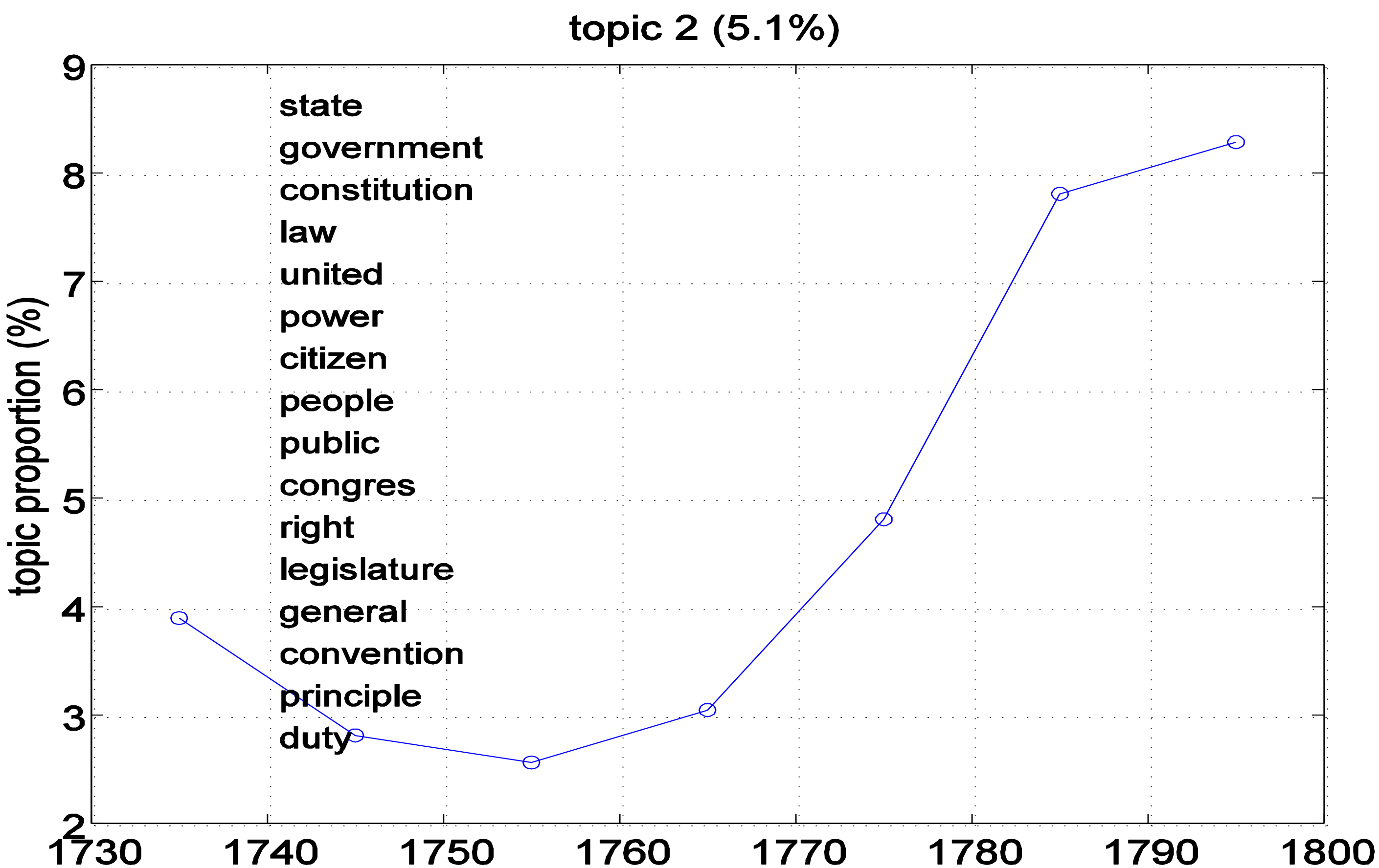
The Pennsylvania GAZETTE.

Containing the freshest Advices, Foreign and Domestick.

Size	Most likely words in topic
6%	away reward servant **named feet jacket** high paid hair coat run inches master
5%	state government constitution **law united** power citizen people public congress
5%	good house acre **sold land meadow** mile premise plantation stone mill dwelling
4%	silk cotton ditto **white black linen cloth** women blue worsted fine thread plain
2%	church life god **society great friend christian** good virtue religion minister rev

(from Dave Newman and Sharon Block, UC Irvine)

Historical Trends in Pennsylvania Gazette Data



Topics from DNA Microarray Literature

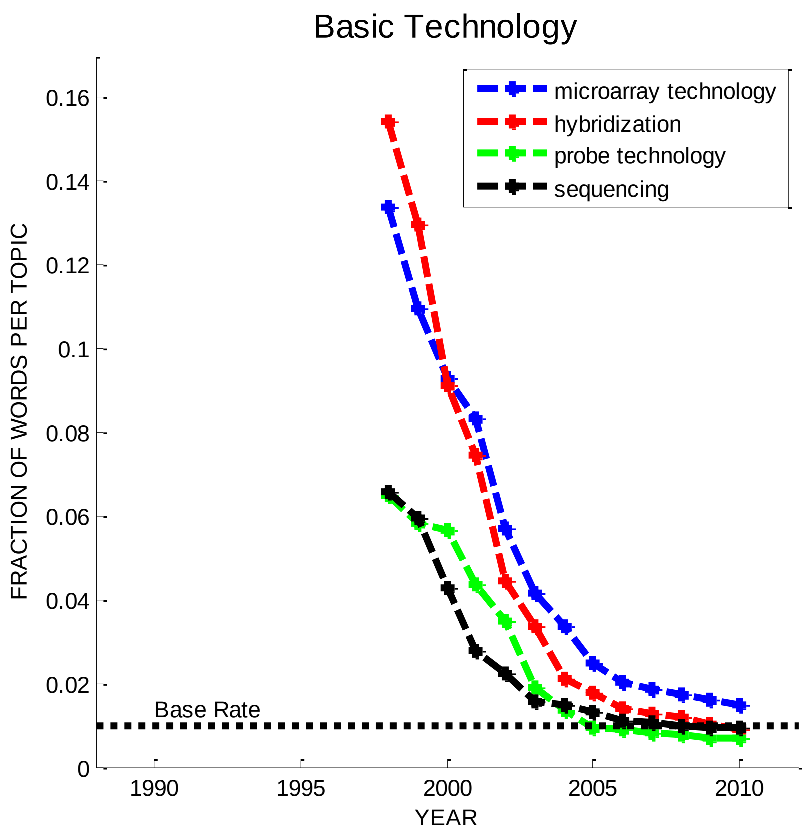
49,000 PubMed abstracts related to DNA Microarrays

Displayed below are the top 5 highest probability words for 5 selected topics

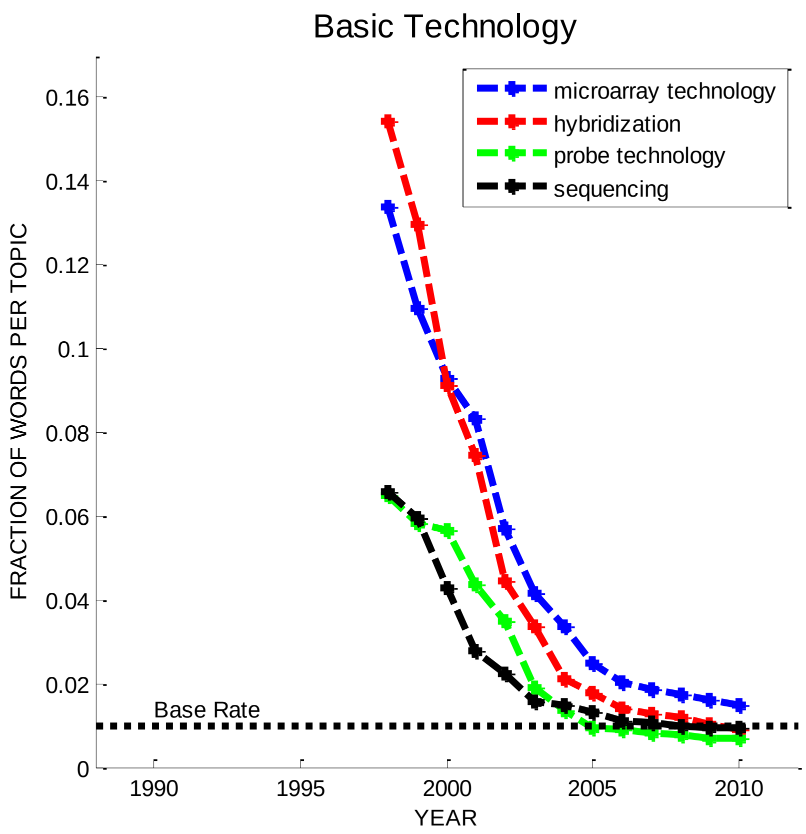
Microarray Chip Technology	Classification Methods	Databases and Annotation	Regulatory Networks	Cancer
detection	classification	databases	network	patient
surface	selection	tool	regulatory	tumor
fluorescence	cancer	annotation	pathway	cancer
hybridization	algorithm	data set	interaction	survival
array	feature	web	transcriptional	prognostic

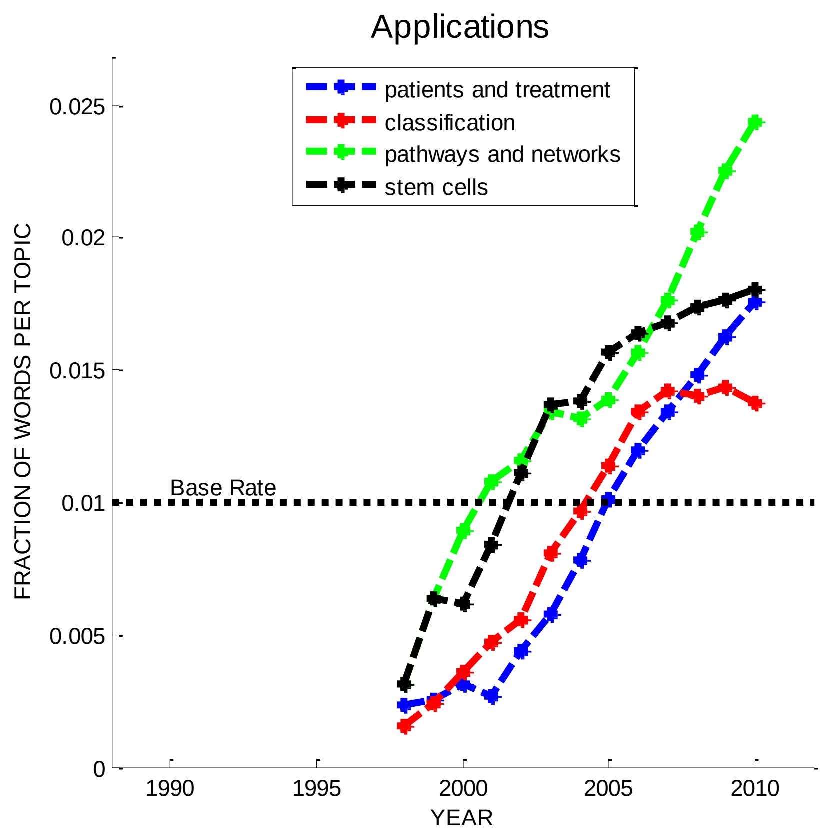
From basic technology to applications

Technology v. Application Topics

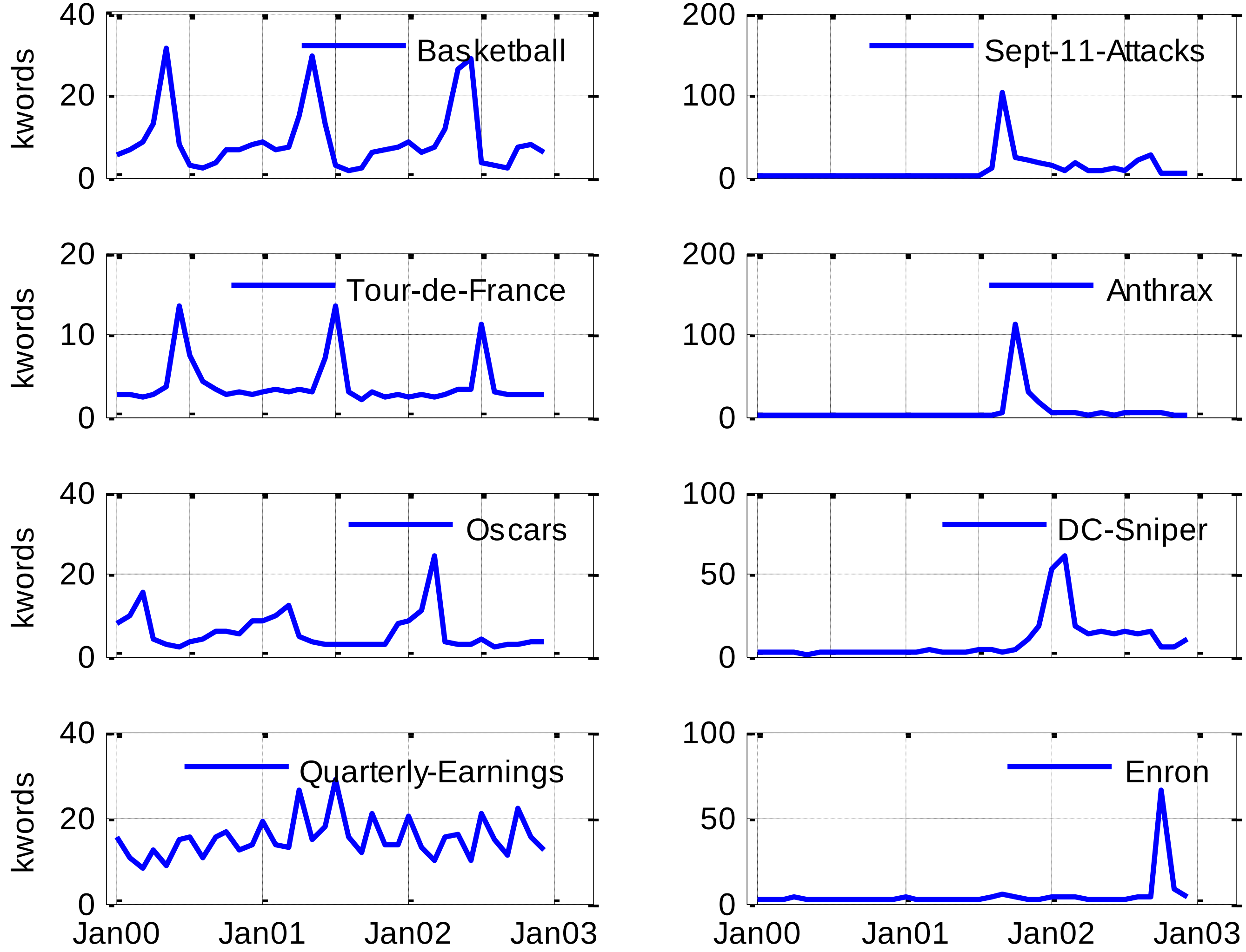


Technology v. Application Topics





Topic Trends (New York Times articles)



Topic Models and Related Approaches

A Brief History of Topic Modeling

Latent Semantic Analysis
(Deerwester et al, JASIS, 1990)

Aspect Model
(Hofmann, 1999, SIGIR)

Latent Dirichlet Allocation (LDA)
(Blei, Ng, Jordan, 2003, JMLR)

Gibbs Sampling Algorithm
(Griffiths and Steyvers, 2004, PNAS)

New Models

e.g, Author-Topic Model
(Rosen-Zvi et al, 2004)

Fast Algorithms

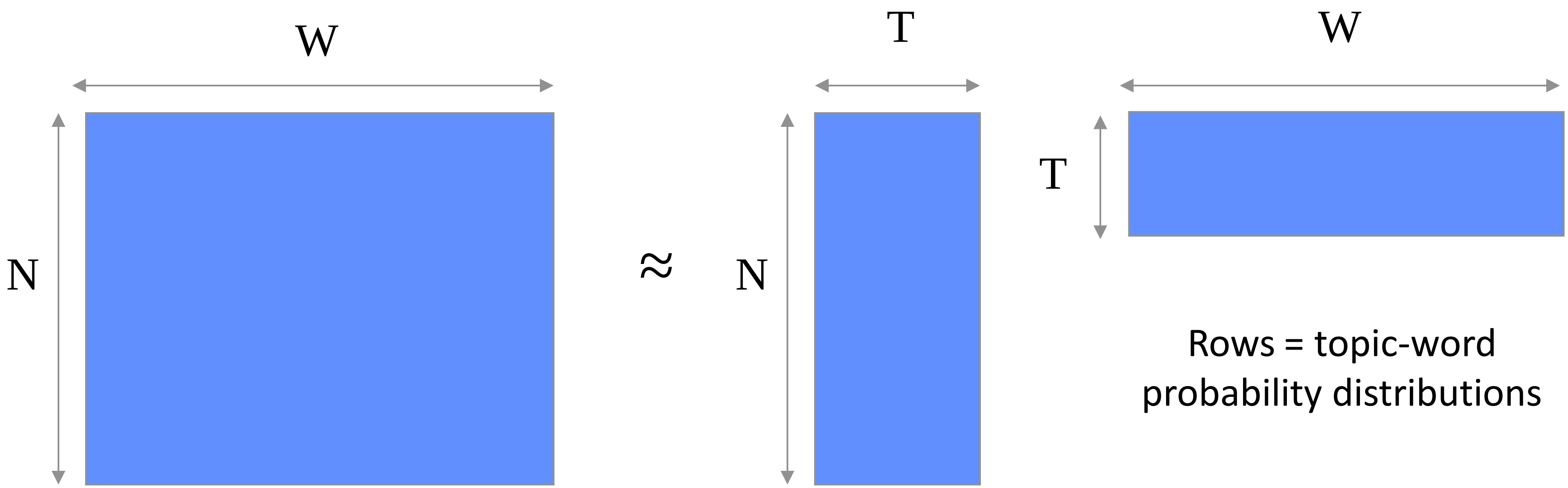
e.g, Distributed Learning
(Newman et al, 2009)

e.g, Fast Approximations
(Hofmann et al, 2013)

New Application Areas

Sociology, Political Science,
Scientometrics, Humanities,
History, and more

Topics as Matrix Factorization

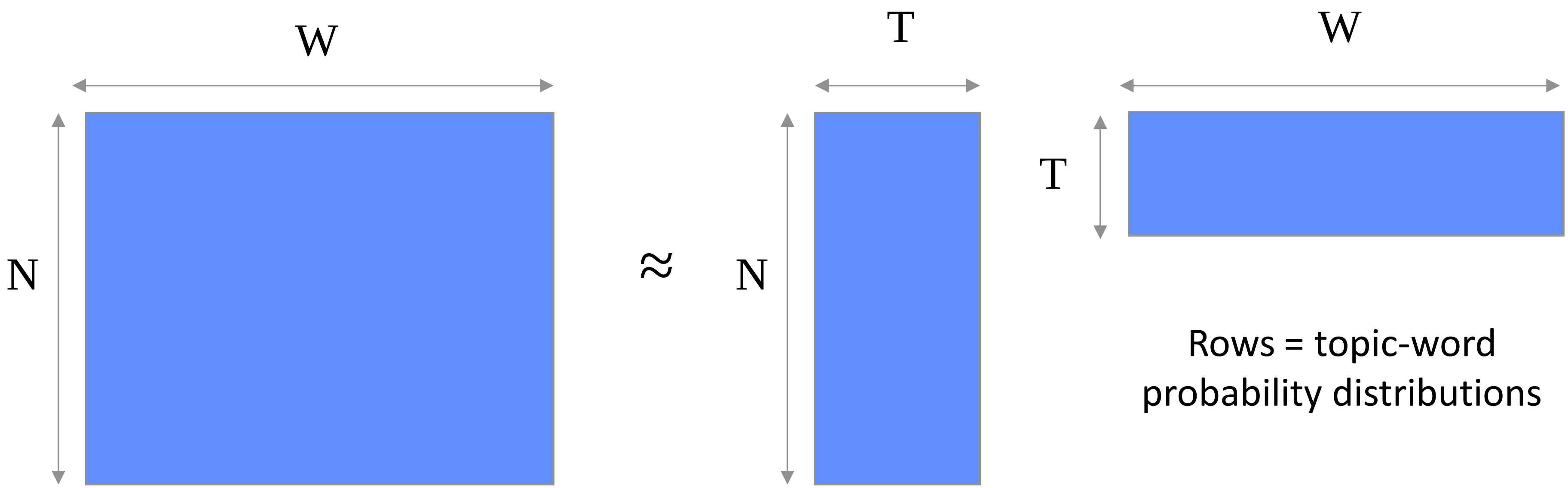


Rows = topic-word probability distributions

Probability of Observed Words
Rows = Docs, Cols = Words

Rows = T topic weights for each document

Topics as Matrix Factorization



Rows = topic-word
probability distributions

Probability of Observed Words
Rows = Docs, Cols = Words

Rows = T topic weights
for each document

Directly analogous to principal components and factor models:

data matrix ≈ weights * basis functions

Clusters v. Topics

Original Document

Hidden Markov Models in Molecular Biology: New Algorithms and Applications

Pierre Baldi, Yves C Hauvin, Tim Hunkapiller, Marcella A. McClure

Hidden Markov Models (HMMs) can be applied to several important problems in molecular biology. We introduce a new convergent learning algorithm for HMMs that, unlike the classical Baum-Welch algorithm is smooth and can be applied on-line or in batch mode, with or without the usual Viterbi most likely path approximation. Left-right HMMs with insertion and deletion states are then trained to represent several protein families including immunoglobulins and kinases. In all cases, the models derived capture all the important statistical properties of the families and can be used efficiently in a number of important tasks such as multiple alignment, motif detection, and classification.

Clusters v. Topics

Original Document

Hidden Markov Models in Molecular Biology: New Algorithms and Applications

Pierre Baldi, Yves C Hauvin, Tim Hunkapiller, Marcella A. McClure

Hidden Markov Models (HMMs) can be applied to several important problems in molecular biology. We introduce a new convergent learning algorithm for HMMs that, unlike the classical Baum-Welch algorithm is smooth and can be applied on-line or in batch mode, with or without the usual Viterbi most likely path approximation. Left-right HMMs with insertion and deletion states are then trained to represent several protein families including immunoglobulins and kinases. In all cases, the models derived capture all the important statistical properties of the families and can be used efficiently in a number of important tasks such as multiple alignment, motif detection, and classification.

One Cluster

[cluster 88]

model data
models time
neural figure state
learning set
parameters
network
probability
number networks
training function
system algorithm
hidden markov

Clusters v. Topics

Original Document

Hidden Markov Models in Molecular Biology: New Algorithms and Applications

Pierre Baldi, Yves C Hauvin, Tim Hunkapiller, Marcella A. McClure

Hidden Markov Models (HMMs) can be applied to several important problems in molecular biology. We introduce a new convergent learning algorithm for HMMs that, unlike the classical Baum-Welch algorithm is smooth and can be applied on-line or in batch mode, with or without the usual Viterbi most likely path approximation. Left-right HMMs with insertion and deletion states are then trained to represent several protein families including immunoglobulins and kinases. In all cases, the models derived capture all the important statistical properties of the families and can be used efficiently in a number of important tasks such as multiple alignment, motif detection, and classification.

One Cluster

[cluster 88] model data models time neural figure state learning set parameters network probability number networks training function system algorithm hidden markov

Multiple Topics

[topic 10] state hmm markov sequence models hidden states probabilities sequences parameters transition probability training hmms hybrid model likelihood modeling

[topic 37] genetic structure chain protein population region algorithms human mouse selection fitness proteins search evolution generation function sequence sequences genes

Methodological Issues

- Selection of smoothing parameters/priors
- Convergence of the sampler?
- Consistency of results across sampling runs?
- Selecting the numbers of topics
- Publicly available code?
  - Mallet from U Mass
  - See also David Blei's Web page

Methods for Evaluating Topic Models

- Human inspection
  - e.g., coherence of high probability words
- Log Probability of Test Documents
  - Better models assign higher probability to unseen documents
- Performance on specific tasks
  - Information retrieval
  - Document classification

Extensions to Topic Models

Adding Metadata to Topic Models?

- **Topic models with metadata? With time? With spatial information?**
  - Yes: can write down simple generative models and then "invert"

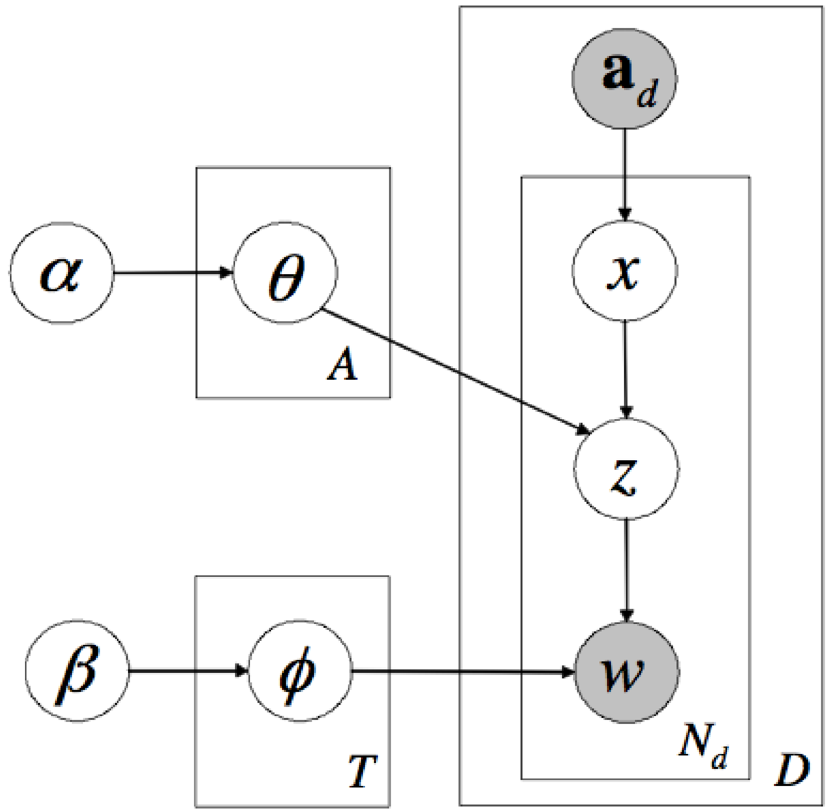
- **Example: Probabilistic Pseudocode for Author-Topic Model**

```
For each document in our corpus
    For each word in our document
        Randomly select an author of the document
        Sample a topic from P(topics | author)
        Given the topic, sample a word from P(words | topic)
    End
End
```

- **We can write down many interesting models in a similar way**
  - Much harder to do inference, i.e., "invert" these models given data

The Author-Topic Model

Steyvers et al, 2004; Rosen-Zvi et al, 2010



Examples of Author-Topic Models

Learned using an author-topic model applied to 20 years of papers from the NIPS conference on machine learning

TOPIC 4	
WORD	PROB.
LIGHT	.0306
RESPONSE	.0282
INTENSITY	.0252
RETINA	.0241
OPTICAL	.0233
KOCH	.0190
BACKGROUND	.0162
CONTRAST	.0145
CENTER	.0124
FEEDBACK	.0118
AUTHOR	PROB.
Koch_C	.0903
Boahen_K	.0320
Skrzypek_J	.0283
Liu_S	.0250
Delbruck_T	.0232
Etienne-C._R	.0210
Bair_W	.0178
Bialek_W	.0133
Yasui_S	.0106
Hsu_K	.0103

TOPIC 13	
WORD	PROB.
RECOGNITION	.0500
CHARACTER	.0334
TANGENT	.0246
CHARACTERS	.0232
DISTANCE	.0197
HANDWRITTEN	.0166
DIGITS	.0154
SEGMENTATION	.0142
DIGIT	.0124
IMAGE	.0111
AUTHOR	PROB.
Simard_P	.0602
Martin_G	.0340
LeCun_Y	.0339
Henderson_D	.0289
Denker_J	.0245
Revow_M	.0206
Rashid_M	.0205
Rumelhart_D	.0185
Sackinger_E	.0181
Flann_N	.0142

TOPIC 28	
WORD	PROB.
KERNEL	.0547
VECTOR	.0293
SUPPORT	.0293
MARGIN	.0239
SVM	.0196
DATA	.0165
SPACE	.0161
KERNELS	.0160
SET	.0146
MACHINES	.0132
AUTHOR	PROB.
Scholkopf_B	.0774
Smola_A	.0685
Vapnik_V	.0487
Burges_C	.0411
Ratsch_G	.0296
Mason_L	.0232
Platt_J	.0225
Cristianini_N	.0179
Laskov_P	.0160
Chapelle_O	.0152

TOPIC 9	
WORD	PROB.
SOURCE	.0389
INDEPENDENT	.0376
SOURCES	.0344
SEPARATION	.0322
INFORMATION	.0319
ICA	.0276
BLIND	.0227
COMPONENT	.0226
SEJNOWSKI	.0224
NATURAL	.0183
AUTHOR	PROB.
Sejnowski_T	.0627
Bell_A	.0378
Yang_H	.0349
Lee_T	.0348
Attias_H	.0290
Parra_L	.0271
Cichocki_A	.0262
Hyvarinen_A	.0242
Amari_S	.0160
Oja_E	.0143

Author-Topic Models for CiteSeer

TOPIC 205	
WORD	**PROB.**
DATA	0.1563
MINING	0.0674
ATTRIBUTES	0.0462
DISCOVERY	0.0401
ASSOCIATION	0.0335
LARGE	0.0280
KNOWLEDGE	0.0260
DATABASES	0.0210
ATTRIBUTE	0.0188
DATASETS	0.0165
AUTHOR	**PROB.**
Han_J	0.0196
Rastogi_R	0.0094
Zaki_M	0.0084
Shim_K	0.0077
Ng_R	0.0060
Liu_B	0.0058
Mannila_H	0.0056
Brin_S	0.0054
Liu_H	0.0047
Holder_L	0.0044

TOPIC 209	
WORD	**PROB.**
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	**PROB.**
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048

TOPIC 289	
WORD	**PROB.**
RETRIEVAL	0.1179
TEXT	0.0853
DOCUMENTS	0.0527
INFORMATION	0.0504
DOCUMENT	0.0441
CONTENT	0.0242
INDEXING	0.0205
RELEVANCE	0.0159
COLLECTION	0.0146
RELEVANT	0.0136
AUTHOR	**PROB.**
Oard_D	0.0110
Croft_W	0.0056
Jones_K	0.0053
Schauble_P	0.0051
Voorhees_E	0.0050
Singhal_A	0.0048
Hawking_D	0.0048
Merkl_D	0.0042
Allan_J	0.0040
Doermann_D	0.0039

TOPIC 10	
WORD	**PROB.**
QUERY	0.1848
QUERIES	0.1367
INDEX	0.0488
DATA	0.0368
JOIN	0.0260
INDEXING	0.0180
PROCESSING	0.0113
AGGREGATE	0.0110
ACCESS	0.0102
PRESENT	0.0095
AUTHOR	**PROB.**
Suciu_D	0.0102
Naughton_J	0.0095
Levy_A	0.0071
DeWitt_D	0.0068
Wong_L	0.0067
Chakrabarti_K	0.0064
Ross_K	0.0061
Hellerstein_J	0.0059
Lenzerini_M	0.0054
Moerkotte_G	0.0053

Author-Profiles

Author = Andrew McCallum, U Mass:.

- Topic 1: classification, training, generalization, decision, data,…
- Topic 2: learning, machine, examples, reinforcement, inductive,….
- Topic 3: retrieval, text, document, information, content,…

Author = Hector Garcia-Molina, Stanford:

- Topic 1: query, index, data, join, processing, aggregate….
- Topic 2: transaction, concurrency, copy, permission, distributed….
- Topic 3: source, separation, paper, heterogeneous, merging…..

Author = Jerry Friedman, Stanford:

- Topic 1: regression, estimate, variance, data, series,…
- Topic 2: classification, training, accuracy, decision, data,….
- Topic 3: distance, metric, similarity, measure, nearest,…


The New York Times
Japanese Commuter Train Kills at Least 71

Learning Word Combinations

Terrorism	Wall Street Firms	Stock Market	Bankruptcy
SEPT_11	WALL_STREET	WEEK	BANKRUPTCY
WAR	ANALYSTS	DOW_JONES	CREDITORS
SECURITY	INVESTORS	POINTS	BANKRUPTCY_PROTECTION
IRAQ	FIRM	10_YR_TREASURY_YIELD	ASSETS
TERRORISM	GOLDMAN_SACHS	PERCENT	COMPANY
NATION	FIRMS	CLOSE	FILED
KILLED	INVESTMENT	NASDAQ_COMPOSITE	BANKRUPTCY_FILING
AFGHANISTAN	MERRILL_LYNCH	STANDARD_POOR	ENRON
ATTACKS	COMPANIES	CHANGE	BANKRUPTCY_COURT
OSAMA_BIN_LADEN	SECURITIES	FRIDAY	KMART
AMERICAN	RESEARCH	DOW_INDUSTRIALS	CHAPTER_11
ATTACK	STOCK	GRAPH_TRACKS	FILING
NEW_YORK_REGION	BUSINESS	EXPECTED	COOPER
NEW	ANALYST	BILLION	BILLIONS
MILITARY	WALL_STREET_FIRMS	NASDAQ_COMPOSITE_INDEX	COMPANIES
NEW_YORK	SALOMON_SMITH_BARNEY	EST_02	BANKRUPTCY_PROCEEDINGS
WORLD	CLIENTS	PHOTO_YESTERDAY	DEBTS
NATIONAL	INVESTMENT_BANKING	YEN	RESTRUCTURING
QAEDA	INVESTMENT_BANKERS	10	CASE
TERRORIST_ATTACKS	INVESTMENT_BANKS	500_STOCK_INDEX	GROUP

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Reinforcement Learning Topic with Topical N-grams

LDA	Topical N-grams (2+)
state	reinforcement learning
learning	optimal policy
policy	dynamic programming
action	optimal control
reinforcement	function approximator
states	prioritized sweeping
time	finite-state controller
optimal	learning system
actions	reinforcement learning RL
function	function approximators
algorithm	markov decision problems
reward	markov decision processes
step	local search
dynamic	state-action pair
control	markov decision process
sutton	belief states
rl	stochastic policy
decision	action selection
algorithms	upright position
agent	reinforcement learning methods

Figure courtesy of Xuerie Wang and Andrew McCallum, U Mass Amherst

Support Vector Machine Topic with Topical N-grams

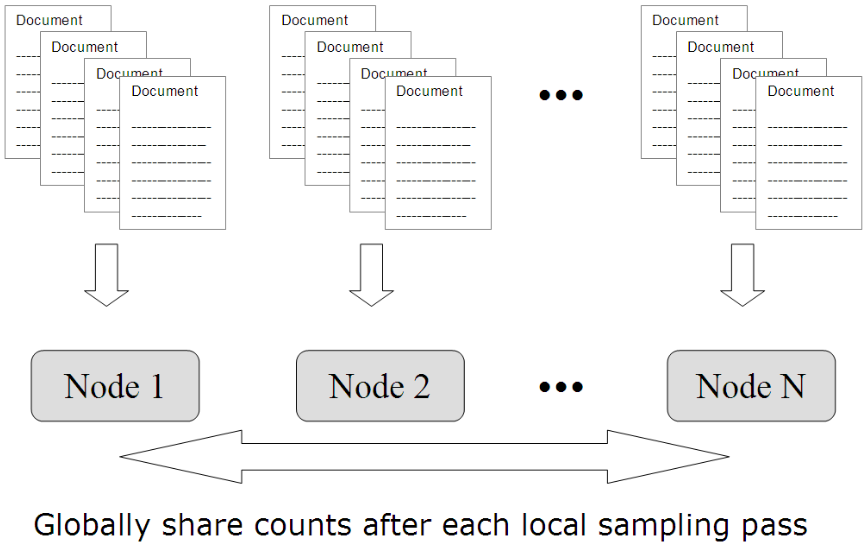
LDA	Topical N-grams (2+)
kernel	support vectors
linear	test error
vector	support vector machines
support	training error
set	feature space
nonlinear	training examples
data	decision function
algorithm	cost functions
space	test inputs
pca	kkt conditions
function	leave-one-out procedure
problem	soft margin
margin	bayesian transduction
vectors	training patterns
solution	training points
training	maximum margin
svm	strictly convex
kernels	regularization operators
matrix	base classifiers
machines	convex optimization

Figure courtesy of Xuerie Wang and Andrew McCallum, U Mass Amherst

Scaling to Large Corpora

- Time complexity: linear in number of word tokens and topics
  - But even this can be slow on millions of documents
- Distributed algorithms
  - Distribute documents across multiple processors (Newman et al, 2009)
  - Approximate, but works very well in practice
- Fast sampling tricks
  - Inner sampling loop is computed billions of times
  - Can re-order operations to get order of magnitude speedup (Porteous et al, 2006)
- Stochastic gradient methods
  - Standard: 1 iteration = full sweep through all words in the corpus
  - Stochastic/online: update parameters after every few documents
    - Algorithm can converge even before a single iteration is complete!

Distributed Topic Modeling

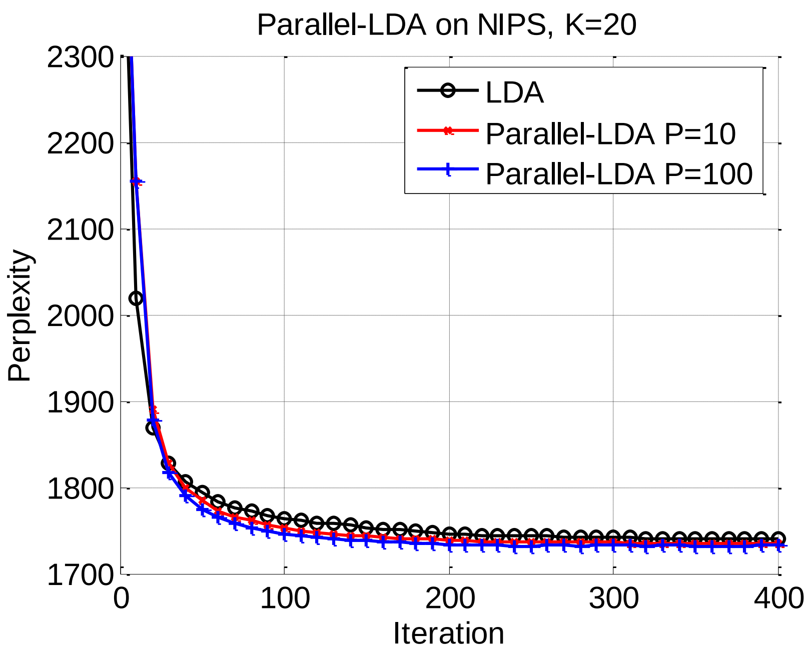


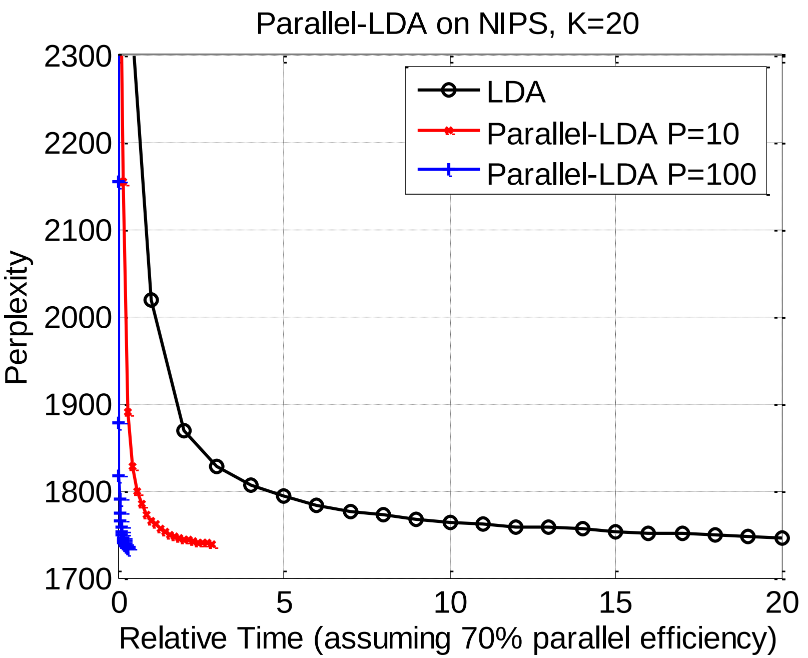
Parallel-LDA [Newman, Asuncion, Smyth, Welling, NIPS 2007, JMLR 2009]

- Each processor performs Gibbs sampling over local set of documents
- At the end of each iteration, all processors combine topic counts to create global model

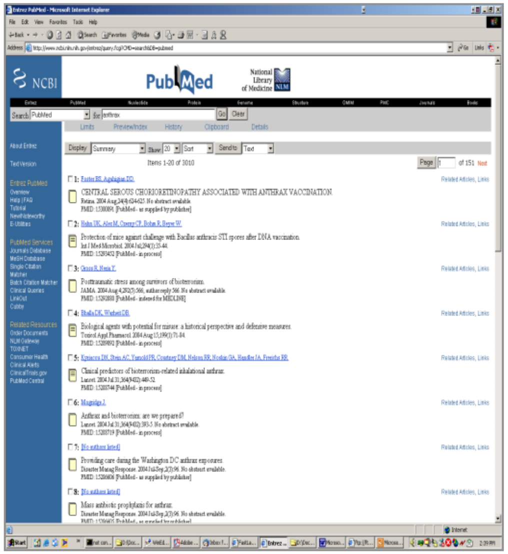
Parallel-LDA Results

[Newman, Asuncion, Smyth, Welling, NIPS 2007, JMLR 2009]

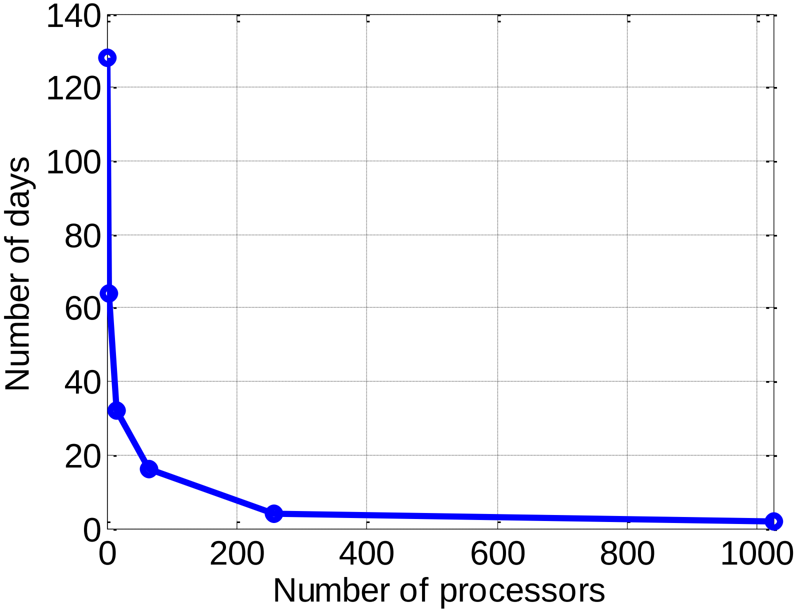




Experiments with 8 Million Documents



All of MEDLINE
8 million abstracts
1 billion words
2000 topics

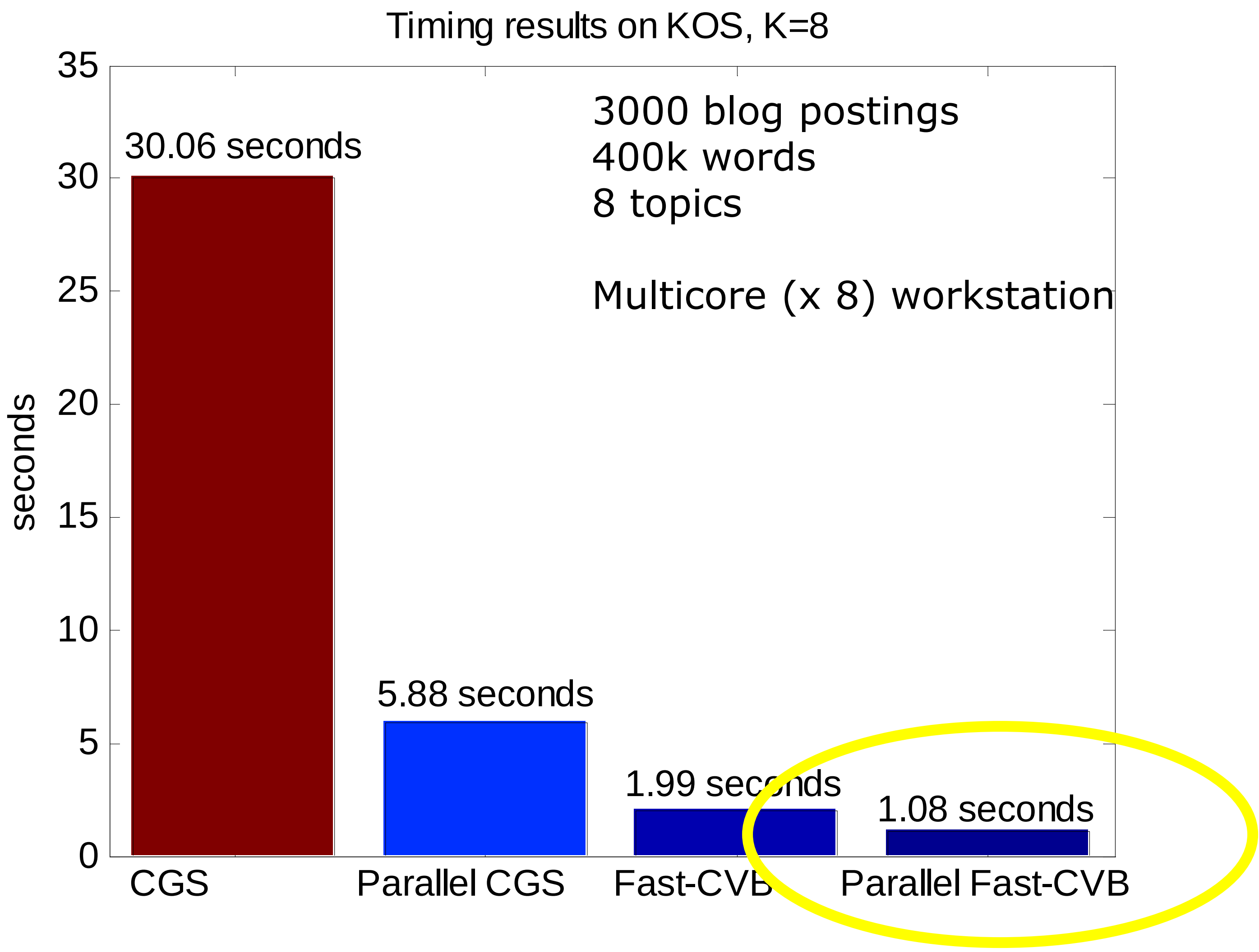


Experiments with 1000 processors at the San Diego Supercomputing Center (SDSC)

Time goes from 4 months to a few hours

Real-Time Topic Modeling

Asuncion et al, 2009



Real-Time Topic Modeling of Search Results

TASER

Theme-Assisted Search Engine in Real-time

mouse [Search] ○ Web ◉ News

Learned Topics

Themes found:

T1: mouse school cartilage skin cell team mice worship pyramid messages
[order by] [zoom]

T2: mouse ces multitouch show consumer microsoft electronic scanner vegas las
[order by] [zoom]

T3: model cell stem gene blood disorder inherited children embryonic lines
[order by] [zoom]

T4: mouse haifa prweb disney big amusement park january mickey home
[order by] [zoom]

T5: mouse click computer control surface coming xbox super cost world
[order by] [zoom]

T6: mouse touch microsoft gestures based looking research window top june
[order by] [zoom]

T7: mouse magic microsoft apple touch air multi function finally sensitive
[order by] [zoom]

The $130 TRON mouse and mouse surface: beautiful overkill

It's going to take a lot of convincing to get us to give our blessing to a mouse and surface that cost $130 . It is covered with TRON branding and style, so fans of the franchise have a little extra reason to pick it up, but that's a very expensive mouse, accessory, collectible, or whatever it is. The packaging is suitably high-end, as the mouse and the surface are displayed in heavy cardboard ...

http://arstechnica.com/gaming/reviews/2011/01/the-130-tron-mouse-and-mouse-surface-beautiful-overkill.ars?utm_source=rss&utm_mediu m=rss&utm_campaign=rss -- 0kb -- 2011/01/10

Similar Pages Thematic Markup

Microsoft Touch Mouse Announced for June Release

Microsoft is looking to enhance your Window 7 navigating experience with its new gesture-based Touch Mouse.

http://www.pcmag.com/article2/0,2817,2375277,00.asp?kc=PCRSS03069TX1K0001121 -- 0kb -- 2011/01/06

Similar Pages Thematic Markup

Microsoft Touch Mouse due midyear

Microsoft has announced the Touch Mouse, the company's answer to Apple's Magic Mouse. Designed for use with Windows 7, it uses capacitative multitouch technology.

http://www.itwire.com/business-it-news/technology/44261-microsoft-touch-mouse-due-midyear -- 0kb -- 2011/01/09

Similar Pages Thematic Markup

See how they run (amok): Mouse calls are up this year

Big snowdrifts and recent mild winters are being blamed for more mouse calls from Twin Cities residents. The best mouse trap? Prevention.

http://www.startribune.com/local/113177959.html -- 0kb -- 2011/01/10

Similar Pages Thematic Markup

Julia Gillard's 'Mouse Pack' and other dumb stuff

Julia Gillard is "not well informed" and is part of a Melbourne-based gang called The Mouse Pack, ,while Tony Abbott has "good manners", is "formidable" and possessed of a "first-class mind".

http://www.brisbanetimes.com.au/entertainment/books/julia-gillards-mouse-pack-and-other-dumb-stuff-20110110-19kbe.html -- 0kb -- 2011/01/10

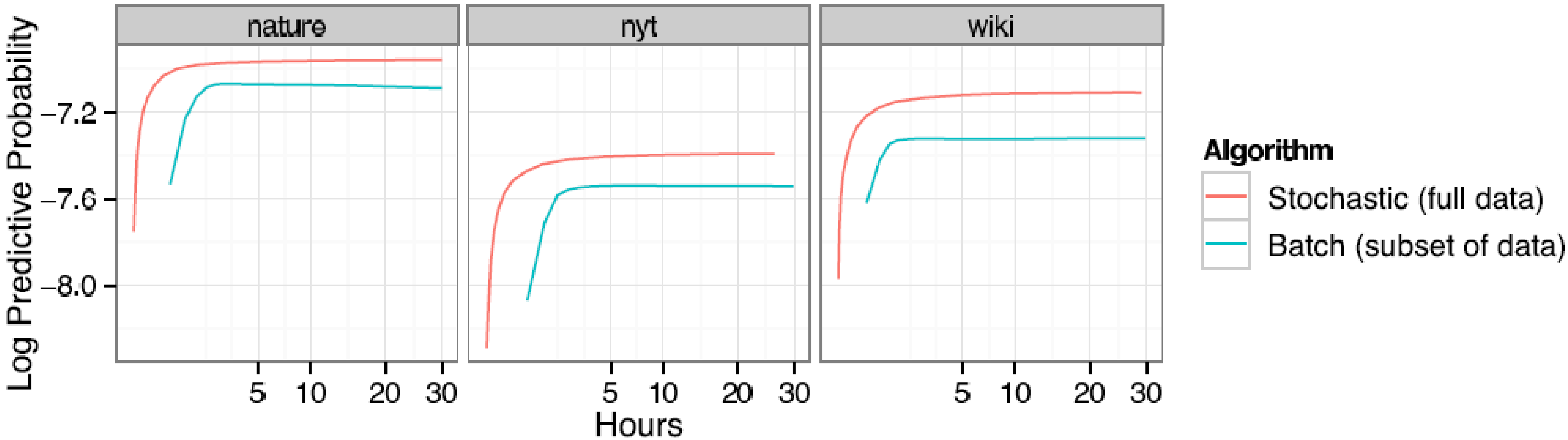
Similar Pages Thematic Markup

Topic Mixtures

Fast Topic Learning with Stochastic Methods

(From Hoffmann et al, 2013)

350k documents	1.8M documents	3.8 M documents
58M words	461M words	482M words



Y-axis = predictive log probability on test documents: higher is better
X-axis = time on a log-scale

Stochastic algorithm learns a better model in minutes than batch algorithm does in hours

Applications of Topic Modeling

Application: Calit2 Research Browser

- System crawled UCI/UCSD faculty websites
- Browser built on topic model learned from faculty papers
- Query-answering = computation of conditional probabilities

home | researchers | research topics

one topic

neural network models and algorithms

network input unit learning output training pattern neural_network representation weig
grammar class structure connectionist learn net performance simple prediction connec
elman classes experiment features architecture modeling training_set recognition initial
vowel mit_press chaotic epoch mapping rules dynamical feature label

Other researchers in neural network models and algorithms (UCSD,UCI):

most prolific researchers for this topic

(19%) DE SA, VIRGINIA
(11%) COTTRELL, GARRISON
(11%) ELMAN, JEFFREY L.
(5%) MJOLSNESS, ERIC D.
(4%) BELEW, RICHARD K.
(4%) YOUSEFIZADEH, HOMAYOUN
(3%) GRANGER, RICHARD H.
(3%) BALDI, PIERRE F.
(2%) WELLING, MAX
(2%) ABARBANEL, HENRY D.
(2%) BORK, ALFRED
(1%) KIBLER, DENNIS F.
(1%) CHANCE, FRANCES S.
(1%) TRIESCH, JOCHEN
(1%) STEYVERS, MARK
(1%) TODOROV, EMANUEL
(1%) BATALI, JOHN D.
(1%) ESKIN, ELEAZAR

one researcher

COTTRELL, GARRISON

COG SCI
DIVISION OF SOCIAL SCIENCES
UCSD
email: gary@ucsd.edu
publications URL: http://www-cse.ucsd.edu/users/gary/ (53 papers collected)

topics this researcher works on

Research topics:

(28%) [neural network models and algorithms] network input unit learning outpu
(14%) [image and vision modeling] image images face recognition pixel features
(7%) [information retrieval] query retrieval feature image user document syster
(7%) [cognitive experiments] subject word memory experiment task participant
(4%) [data analysis] data correlation analysis sample average estimates param
(4%) [cognition and EEG] word erp processing brain sentence language semant
(4%) [language modeling] language verb theory sense structure word meaning
(4%) [human learning and development] children word development learning ag
(3%) [modeling] model simulation parameter modeling process

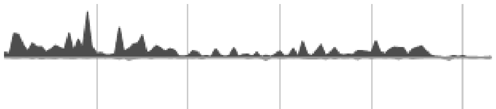
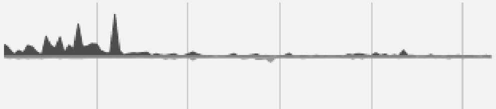
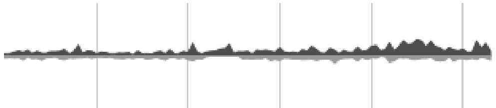
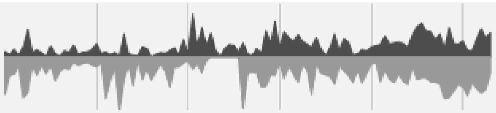
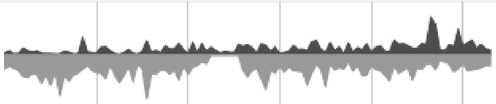
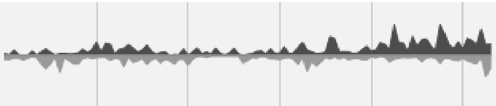
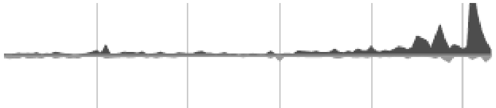
other researchers with similar topical interests

Related researchers (UCSD,UCI) :

(0.9) DE SA, VIRGINIA
(0.7) ELMAN, JEFFREY L.
(0.6) MJOLSNESS, ERIC D.
(0.5) BELONGIE, SERGE J.
(0.5) VASCONCELOS, NUNO
(0.5) BELEW, RICHARD K.
(0.5) TRIESCH, JOCHEN
(0.4) KREIGMAN, DAVID
(0.4) WELLING, MAX
(0.3) STEYVERS, MARK
(0.3) ESKIN, ELEAZAR
(0.3) KIRSH, DAVID J.
(0.3) BROWN, SCOTT D.
(0.3) GRANGER, RICHARD H.
(0.3) JAIN, RAMESH CHANDRA

From David Mimno, Computational Historiography: Data Mining in a Century of Classics Journals, *Journal on Computing and Cultural Heritage (JOCCH)*, 2012

VII: Aligned topics in English (AJP), in dark gray extending upwards, and German (Hermes), in light gray extending downwards. Lines are at 1920, 1940, 1960, 1980 and 2000.

case languages language example nominative english genitive means cases object dative subject grammatical accusative article finnish like noun feminine	sprachen sprache kasus beispiel fall akkusativ haus genitiv grammatik deutschen dativ latein beispiele kennt deutsch person ausdrücken präposition regel	
sanskrit self vedic mind buddhist three nature buddhism hindu type means through buddha reality meditation practice veda seeing compound	sanskrit drei vier buddhismus indischen lehre bedeutet prakriti purusha dukkha wesen hinduismus zustand buddhistischen buddha meditation yoga	
linguistics context words use specific type example particular meanings linguistic same terms different lexical analysis sound function literal within	bedeutung zwischen beispiel verschiedene linguistik also beispielsweise wort verschiedenen zeichen lassen wörtern kontext sprachwissenschaft englisch gegenstand siehe unterscheidung spricht	
rhetoric demosthenes cicero against speeches isocrates speech orator others style public along should teacher important private orators ten funeral	cicero rede rhetorik reden redner allem ersten demosthenes ciceros bedeutung tod prozess seinem letzte wobei damit rhetor auseinandersetzung meidias	
law legal court case criminal laws under right rule principle will act crime civil cases decision parties judge courts	recht gesetz wenn also lat law keine ohne wegen liegt tat non sog strafe gemäß deutschen grundsatz vertrag deutschland	
she daughter wife married mother sister marriage husband child children became queen gave bore woman named father birth herself	tochter mutter ihr frau schwester vater ihrem ihre ihrer ihres heiratete ehe gattin kind verheiratet ihren kinder ehefrau geboren	
women woman men man female young male considered famous seen than sexual feminine beautiful role make reports sex parallel	frauen männer gehört mädchen männlichen frau anderen personen jungen junge ihre weiblichen mann männern bestand tritt bringen weiblicher männliche	

From Nguyen et al., Modeling topic control to detect influence in conversations, *Machine Learning Journal*, 2013

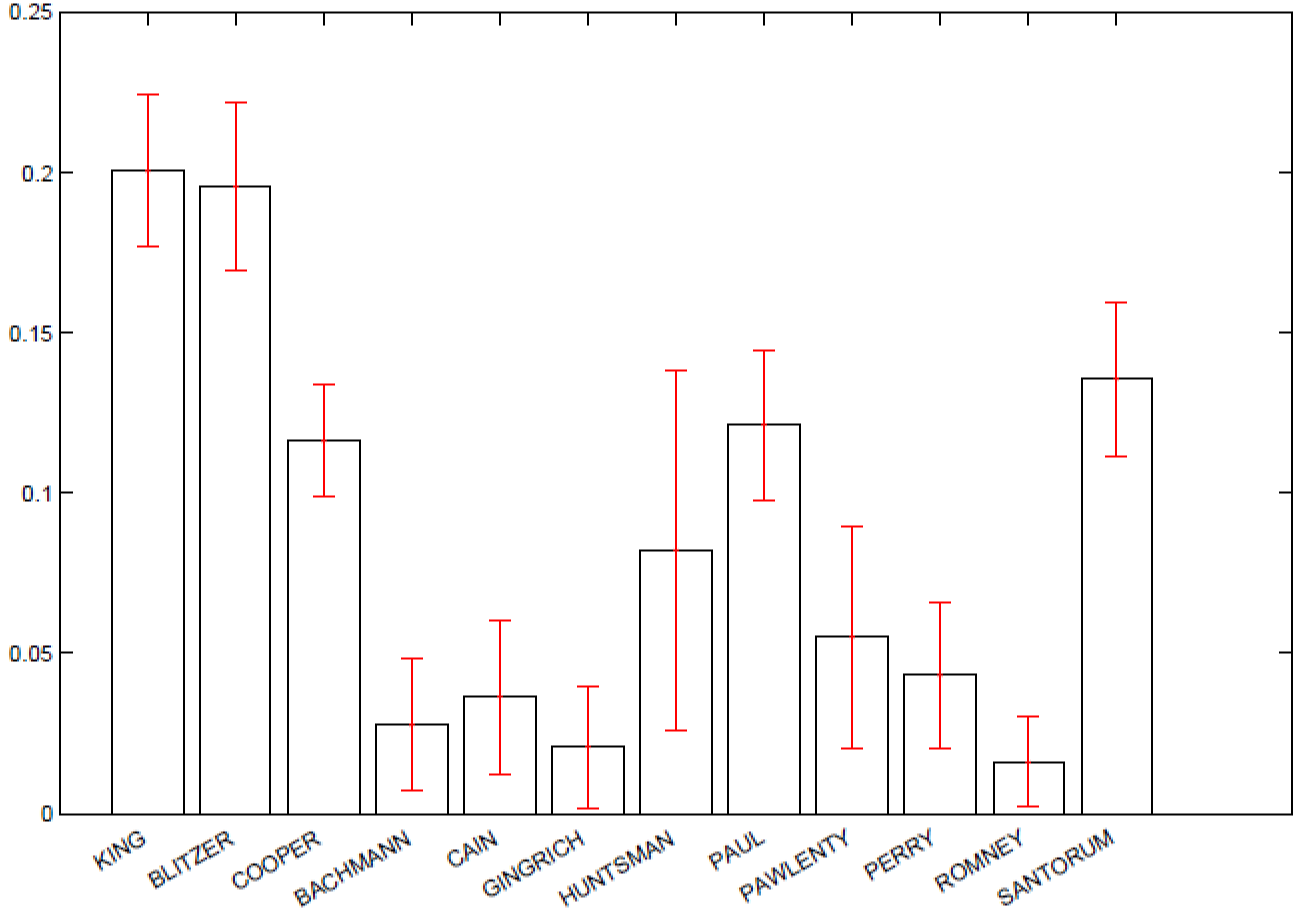


Fig. 6: Topic shift tendency π of speakers in the 2012 Republican Primary Debates (larger means greater tendency). KING, BLITZER and COOPER are moderators in these debates; the rest are candidates.

Analyzing Psychotherapy Trancripts

(joint work with Mark Steyvers, Cognitive Science, UCI)

Each transcript treated as a document

“Subject” and “Symptom” labels manually assigned to some transcripts

Can build in sequential dependence, talk turns, etc

Label Type	Label	High Probability Words
Subject	Medications	dose, mg, medicine, need, wellbutrin, lamictal, mood, sleep, medicines, prescription, medication, use, xanax, klonopin, lexapro, morning, blood, zoloft
	Spousal relationships	wife, married, marriage, home, husband, children, relationship, situation, talked, love, guy, problems, course, work, accept, divorce, couple, meet, girl, attitude, happy, type
Symptom	Fatigue	sleep, blood, depression, energy, tired, low, thyroid, pressure, hormone, fatigue, problems, pituitary, months, growth, level, exercise, treatment, depakote, pharmacy
	Depression	depressed, depression, doctor, sleep, worse, pain, problems, upset, tired, afraid, care, sad, crying, medication, feels, help, sorry, left, understand, hurt, lexapro, remember

Analyzing Psychotherapy Trancripts

(joint work with Mark Steyvers, Cognitive Science, UCI)

Each transcript treated as a document

"Subject" and "Symptom" labels manually assigned to some transcripts

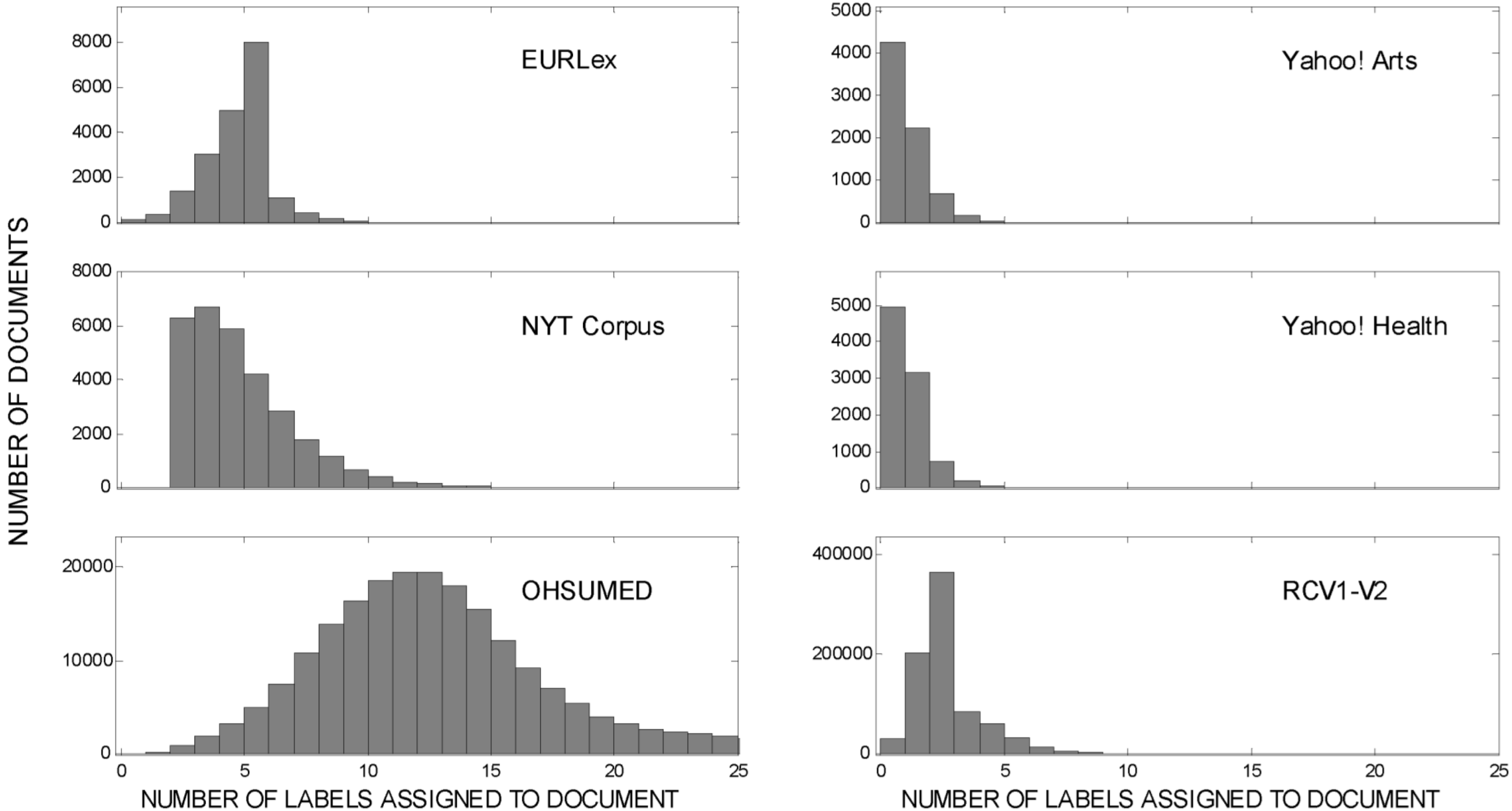
Can build in sequential dependence, talk turns, etc

Label Type	Label	High Probability Words	Example Talk Turn Assigned by Model
Subject	Medications	dose, mg, medicine, need, wellbutrin, lamictal, mood, sleep, medicines, prescription, medication, use, xanax, klonopin, lexapro, morning, blood, zoloft	[THERAPIST] so risperdal, geodon, i like geodon, you haven't been on that. so, we got risperdal, geodon and invega.
	Spousal relationships	wife, married, marriage, home, husband, children, relationship, situation, talked, love, guy, problems, course, work, accept, divorce, couple, meet, girl, attitude, happy, type	[PATIENT] see, because by saying like, `` yes, i'll go home with my wife and i'll stay home, " then i reaffirm my way of life and my relationship with my wife.
Symptom	Fatigue	sleep, blood, depression, energy, tired, low, thyroid, pressure, hormone, fatigue, problems, pituitary, months, growth, level, exercise, treatment, depakote, pharmacy	[PATIENT] pretty good, i'm still having such trouble with being able to have energy. my energy level is down and i am just sleepy, sleepy all day long.
	Depression	depressed, depression, doctor, sleep, worse, pain, problems, upset, tired, afraid, care, sad, crying, medication, feels, help, sorry, left, understand, hurt, lexapro, remember	[PATIENT] and it's sort of - i don't know. i just really don't want to be stuck, like i really don't want to wake up one morning and realize that i hate my life completely. that i'm really miserable where i am or where i decided to go.

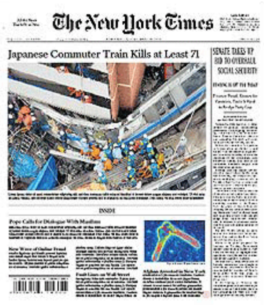
UCIrvine
University of California, Irvine

Topic Models for Supervised Learning with Multilabel Documents

Multilabel Document Data Sets



Real-World Multilabel Data Sets


The New York Times
Japanese Commuter Train Kills at Least 71

- Annotated New York Times
  - 1.5 million news articles
  - Thousands of topics
  - Tagged by team of librarians

- Open Directory Project (ODP)
  - Science subtree
  - 10,817 categories
  - 78k Web pages
  - 11-level hierarchy

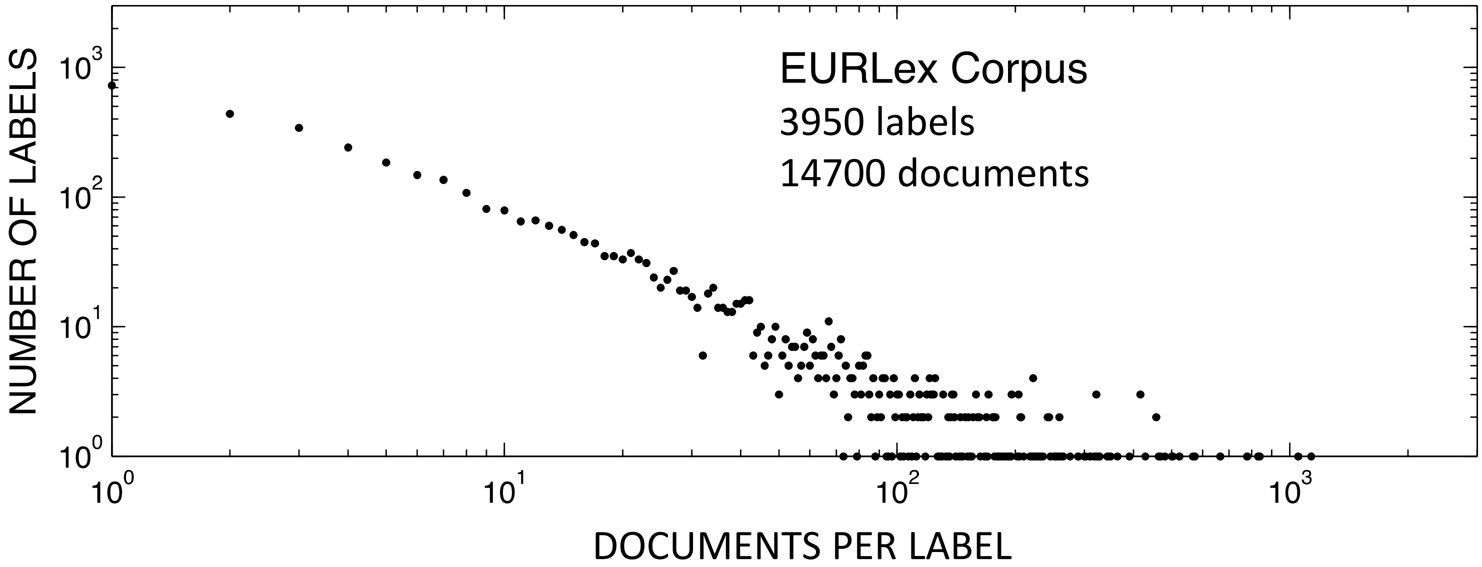
- Wikipedia
  - O(100k) categories
  - Network of relationships
  - Wiki pages tagged with categories

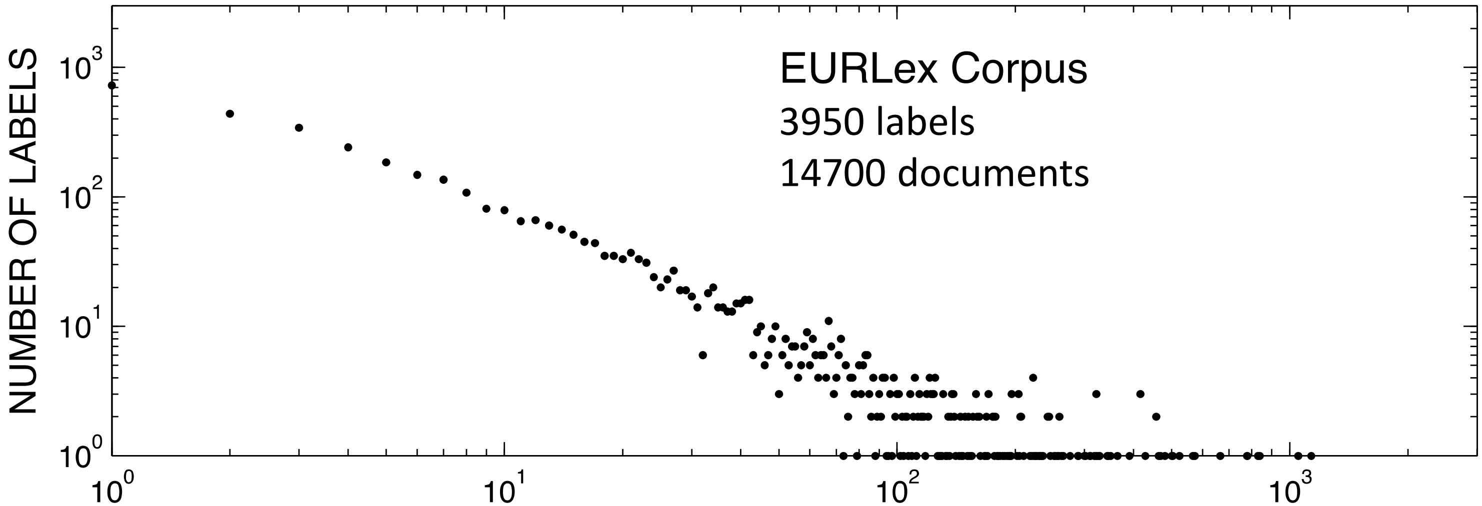
MultiLabel Document Data Sets

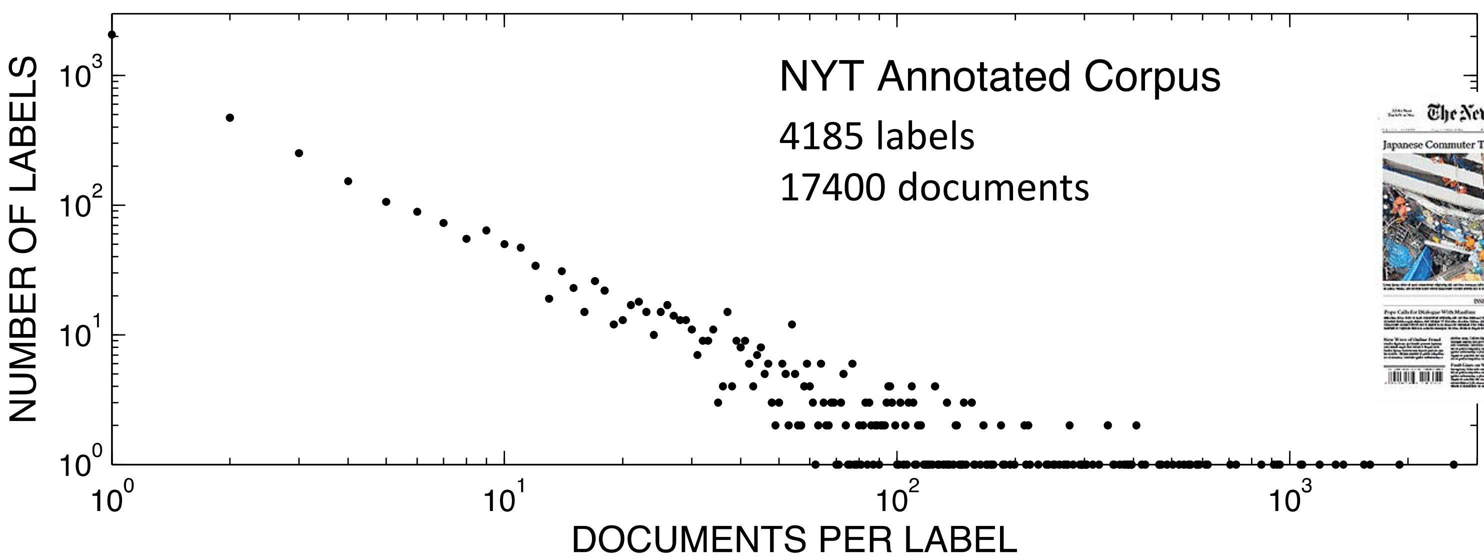
Data Set	Number of Unique Labels	Median Number of Documents per Label
RCV1-V2	103	7410
Yahoo! Arts	14	530
Yahoo! Health	19	500

MultiLabel Document Data Sets

Data Set	Number of Unique Labels	Median Number of Documents per Label
RCV1-V2	103	7410
Yahoo! Arts	14	530
Yahoo! Health	19	500
EUR-Lex	3993	6
New York Times	4185	3


EURLex Corpus
3950 labels
14700 documents
NUMBER OF LABELS
DOCUMENTS PER LABEL
10^0
10^1
10^2
10^3


EURLex Corpus
3950 labels
14700 documents
NUMBER OF LABELS
10^0
10^1
10^2
10^3


NYT Annotated Corpus
4185 labels
17400 documents
NUMBER OF LABELS
DOCUMENTS PER LABEL
10^0
10^1
10^2
10^3


The New York Times
Japanese Commuter Train Kills at Least 71

Prior Work in Multilabel Text Classification

Unrealistically small numbers of labels are often used, e.g.,

- Yahoo! directories: ~20 labels
- RCV-1 ~ 100 labels
- OHSUMED ~ 100 labels

SVMs do well on common labels, but do poorly when there are few documents per label

- See study by Liu et al (*SIGKDD Explorations*, 2005)

Room for new approaches for data sets with many labels

Discriminative Learning

- “One-versus-all” discriminative learning
  - Learn a binary classifier for each label
  - e.g., SVMs, logistic regression
- Potential limitations of discriminative approach
  - documents with many labels
  - labels with few documents
- Differences with generative (topic) model
  - Discriminative: assigns labels at the document level
  - Generative: assigns labels at the word level

Applying Topic Models to Multilabel Classification

Rubin, Chambers, Smyth, Steyvers, MLJ 2012

- Simple idea:
  - Associate each label with a topic (see Ramage et al, EMNLP, 2009)
  - During learning, restrict the sampler to the known labels for the document
  - Algorithm learns a distribution over words for each label
  - Key difference with discriminative methods: labels are assigned per word, not per document
- Modeling label dependencies
  - Extend standard LDA to allow for label (topic) dependencies – significantly improves performance

Topic Modeling with Labels

- Say our documents can have multiple labels (supervised data)

- Simple observation:
  - Labels and topics: 1-1 correspondence
  - When sampling with Gibbs sampler, we can restrict sampling to "topics" (labels) assigned to that document

- Algorithm learns
  - Which labels are associated with each word within a document
  - Probability distribution over words for each label

- Probabilistic basis for multi-label document classification

NY Times Article

Document Labels	Label Freq.	SVM (weight)
ANTITRUST ACTIONS AND LAWS	19	nintendo
SUITS AND LITIGATION	67	mcgowan
VIDEO GAMES	1	futuristic
		compatible
		illusion
		shrewd
		inception
		truthful
		profiles
		billionayear
		suing
		infringement
		architecture
		handheld
		tantamount
		payoff

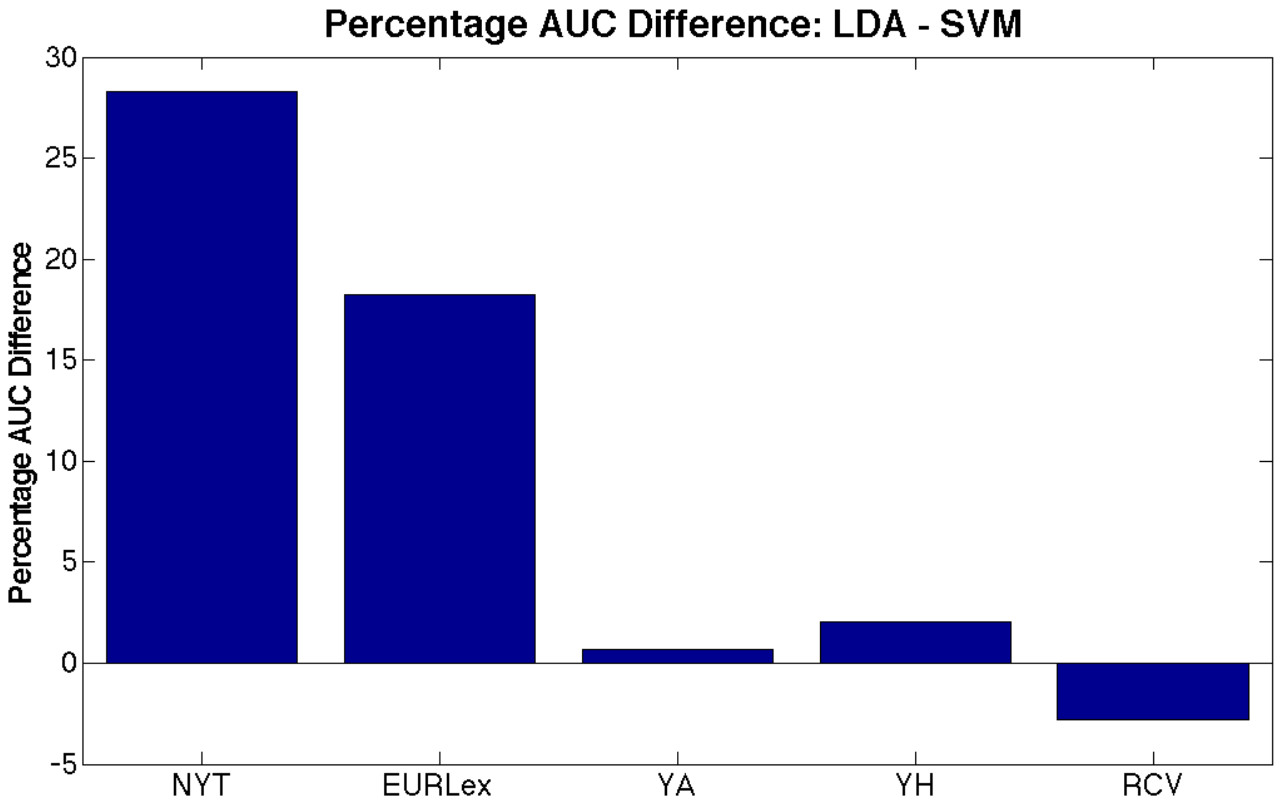
Document Excerpt

A flurry of lawsuits, started by a small American software developer, now surrounds the Nintendo Entertainment System, the best-selling toy in the United States last year...Atari Games argues that Nintendo's high degree of control is tantamount to monopoly, and is suing Nintendo for antitrust violations...

NY Times Article		Models for VIDEO GAMES	
Document Labels	**Label Freq.**	**SVM (weight)**	**LDA (prob.)**
ANTITRUST ACTIONS AND LAWS	19	nintendo	nintendo
SUITS AND LITIGATION	67	mcgowan	games
VIDEO GAMES	1	futuristic	software
		compatible	video
		illusion	system
		shrewd	game
		inception	chip
		truthful	control
		profiles	market
		billionayear	home
		suing	computer
		infringement	shortage
		architecture	say
		handheld	buy
		tantamount	demand
		payoff	developer

Document Excerpt

A flurry of lawsuits, started by a small American software developer, now surrounds the Nintendo Entertainment System, the best-selling toy in the United States last year...Atari Games argues that Nintendo's high degree of control is tantamount to monopoly, and is suing Nintendo for antitrust violations...



From Rubin, Chambers, Smyth, Steyvers, MLJ 2012

Data Set	Median Number of Documents per Label	Metrics where Topics were better	Metrics where SVMs were better
RCV1-V2	7410	1	24
Yahoo! Arts	530	11	13
Yahoo! Health	500	13	12
EUR-Lex	6	18	6
New York Times	3	22	1

From Rubin, Chambers, Smyth, Steyvers, MLJ 2012

Background Reading on Topic Models

David Blei's Topic Modeling Web page:

https://www.cs.princeton.edu/~blei/topicmodeling.html

See introductry papers and slides from various tutorials

See also code, browsers, visualizations, discussion list, etc

Original paper on topic modeling,

Latent Dirichlet allocation, David Blei, Andrew Y. Ng and Michael Jordan. *Journal of Machine Learning Research*, 3:993-1022, 2003.

Probabilistic topic models, Steyvers, M. & Griffiths, T. (2006). In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum

Distributed algorithms for topic models, D. Newman, A. Asuncion, P. Smyth, and M. Welling, *Journal of Machine Learning Research*, 10(Aug):1801-1828, 2009.