Text Mining:
Text Clustering

Laura Brown

Some slides adapted from P. Smyth; Han, Kamber, & Pei;
Tan, Steinbach, & Kumar; C. Volinsky; R. Tibshirani;

Latent Semantic Indexing

- Criticism: queries can be posed in many ways, but still mean the same
  - Data mining and knowledge discovery
  - Car and automobile
  - Beet and beetroot
- *Semantically*, these are the same, and documents with either term are relevant.
- Using synonym lists or thesauri are solutions, but messy and difficult.
- Latent Semantic Indexing (LSI): tries to extract hidden semantic structure in the documents
- Search what I meant, not what I said!

LSI

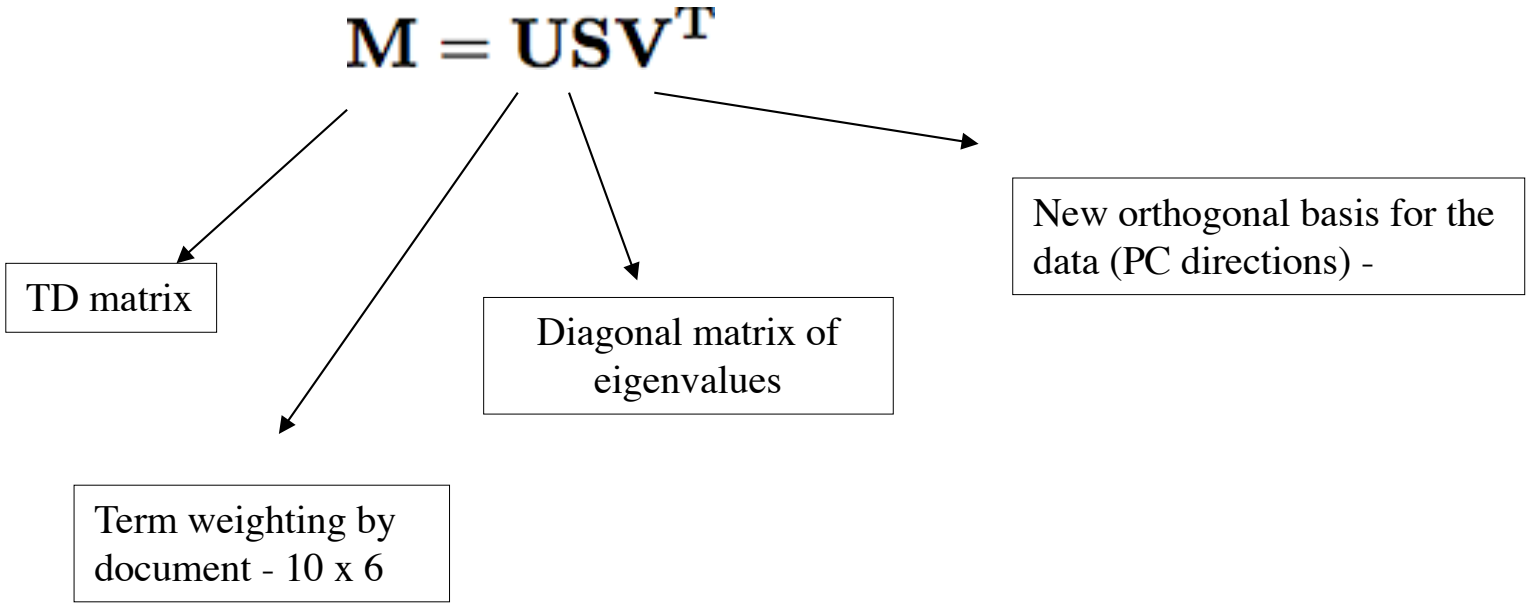
- Approximate the T-dimensional term space using principle components calculated from the TD matrix
- The first k PC directions provide the best set of k orthogonal basis vectors - these explain the most variance in the data.
  - Data is reduced to an N x k matrix, without much loss of information
- Each "direction" is a linear combination of the input terms, and define a clustering of "topics" in the data.
- What does this mean for our toy example?

	Database	SQL	Index	Regression	Likelihood	linear
D1	24	21	9	0	0	3
D2	32	10	5	0	3	0
D3	12	16	5	0	0	0
D4	6	7	2	0	0	0
D5	**43**	31	20	0	3	0
D6	2	0	0	18	7	6
D7	0	0	1	32	12	0
D8	3	0	0	22	4	4
D9	1	0	0	34	27	25
D10	6	0	0	17	4	23

	Database	SQL	Index	Regression	Likelihood	linear
D1	2.53	14.6	4.6	0	0	2.1
D2	3.3	6.7	2.6	0	1.0	0
D3	1.3	11.1	2.6	0	0	0
D4	0.7	4.9	1.0	0	0	0
D5	4.5	**21.5**	10.2	0	1.0	0
D6	0.2	0	0	12.5	2.5	11.1
D7	0	0	0.5	**22.2**	4.3	0
D8	0.3	0	0	15.2	1.4	1.4
D9	0.1	0	0	**23.56**	9.6	17.3
D10	0.6	0	0	11.8	1.4	16.0

LSI

- Typically done using Singular Value Decomposition (SVD) to find principal components



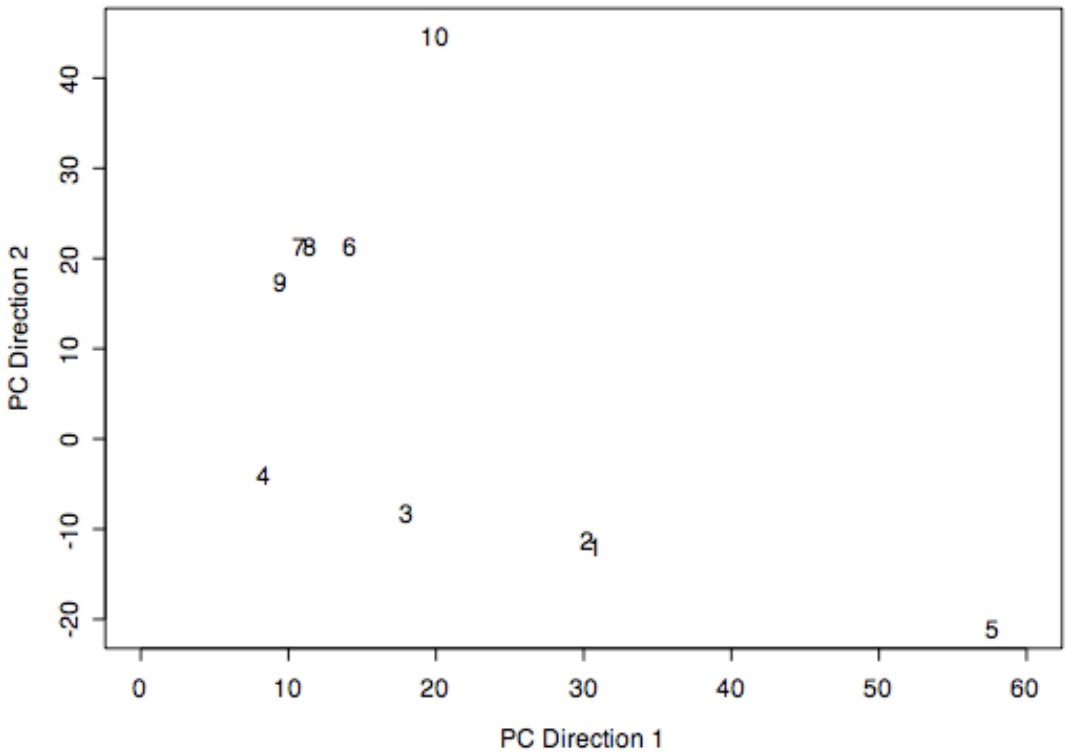
For our example: **S** = (77.4,69.5,22.9,13.5,12.1,4.8)

Fraction of the variance explained (PC1&2) $= \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{6} \lambda_i}$ = 92.5%

LSI

```
> pc2
        [,1]  [,2]
 [1,] 30.90 -11.5
 [2,] 30.30 -10.8
 [3,] 18.00  -7.7
 [4,]  8.37  -3.5
 [5,] 57.70 -20.6
 [6,] 14.20  21.8
 [7,] 10.80  21.9
 [8,] 11.50  21.8
 [9,]  9.50  17.8
[10,] 20.00  45.1
```

Top 2 PC make new pseudo-terms to define documents…



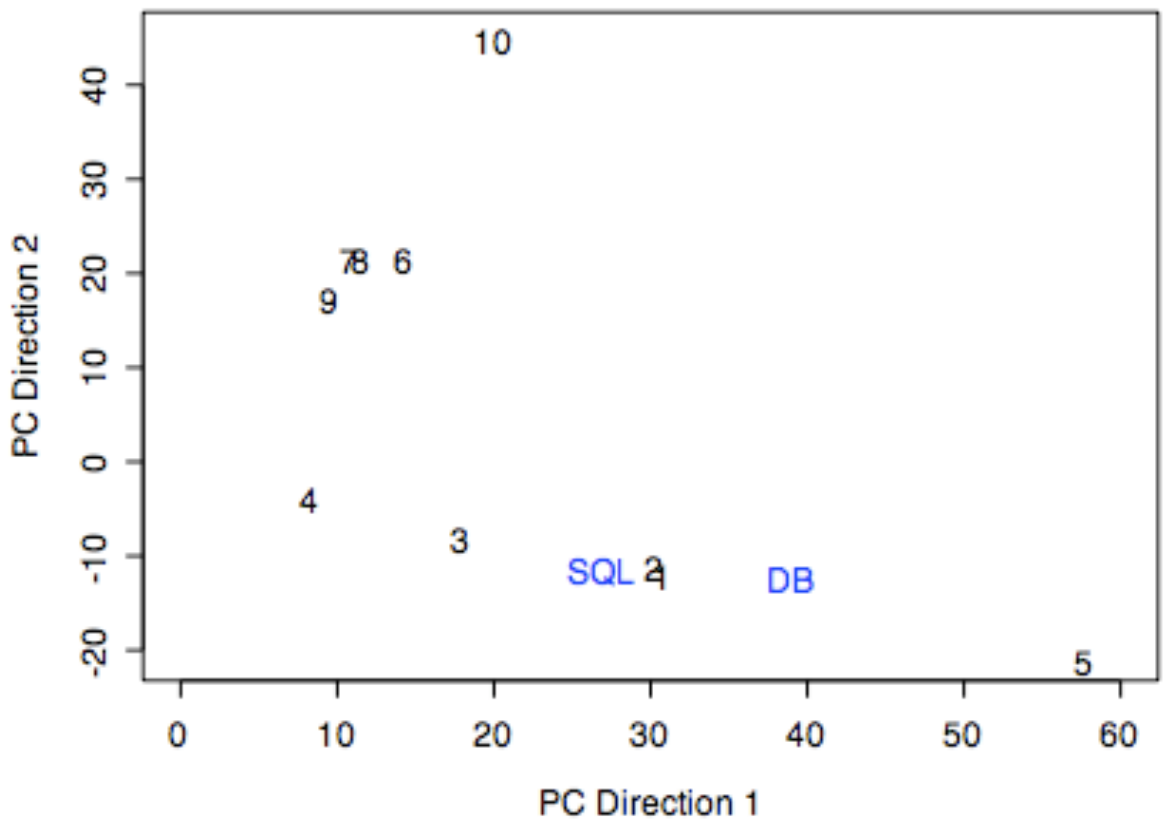
Also, can look at first two Principal components:

```
(0.74,0.49, 0.27,0.28,0.18,0.19)   -> emphasizes first two terms
(-0.28,-0.24,-0.12,0.74,0.37,0.31) -> separates the two clusters

Note how distance from the origin shows number of terms,
And angle (from the origin) shows similarity as well
```

LSI

- Here we show the same plot, but with two new documents, one with the term "SQL" 50 times, another with the term "Databases" 50 times.
- Even though they have no phrases in common, they are close in LSI space



Textual analysis

- Once we have the data into a nice matrix representation (TD, TDxIDF, or LSI), we can throw the data mining toolbox at it:
  - Classification of documents
    - If we have training data for classes
  - Clustering of documents
    - unsupervised

Document Clustering

- Can also do clustering, or unsupervised learning of docs.
- Automatically group related documents based on their content.
- Require no training sets or predetermined taxonomies.
- Major steps
  - Preprocessing
    - Remove stop words, stem, feature extraction, lexical analysis, …
  - Hierarchical clustering
    - Compute similarities applying clustering algorithms, …
  - Slicing
    - Fan out controls, flatten the tree to desired number of levels.

- Like all clustering examples, success is relative

Document Clustering

- To Cluster:
  - Can use LSI
  - Another model: Latent Dirichlet Allocation (LDA)
  - LDA is a generative probabilistic model of a corpus. Documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words.
- LDA:
  - Three concepts: words, topics, and documents
  - Documents are a collection of words and have a probability distribution over topics
  - Topics have a probability distribution over words
  - Fully Bayesian Model

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
  - Can handle counts, meta-data, 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 are given data D

Example from: P. Smyth

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 $\theta_1, \theta_2, \ldots, \theta_6$ sum to 1, $\Sigma\theta = 1$
- Same model for text?
  - Now consider a *k*-sided dies where *k* could be 100,000
  - A vector of *k* probabilities θ, sum to 1, $\Sigma\theta = 1$
  - Can learn these probabilities from a corpus, via smoothed frequencies

Example from: P. Smyth

Topics, 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
…	…

Example from: P. Smyth

Different Topics for Different Sematic 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
….	….

Example from: P. Smyth

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)

Example from: P. Smyth

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

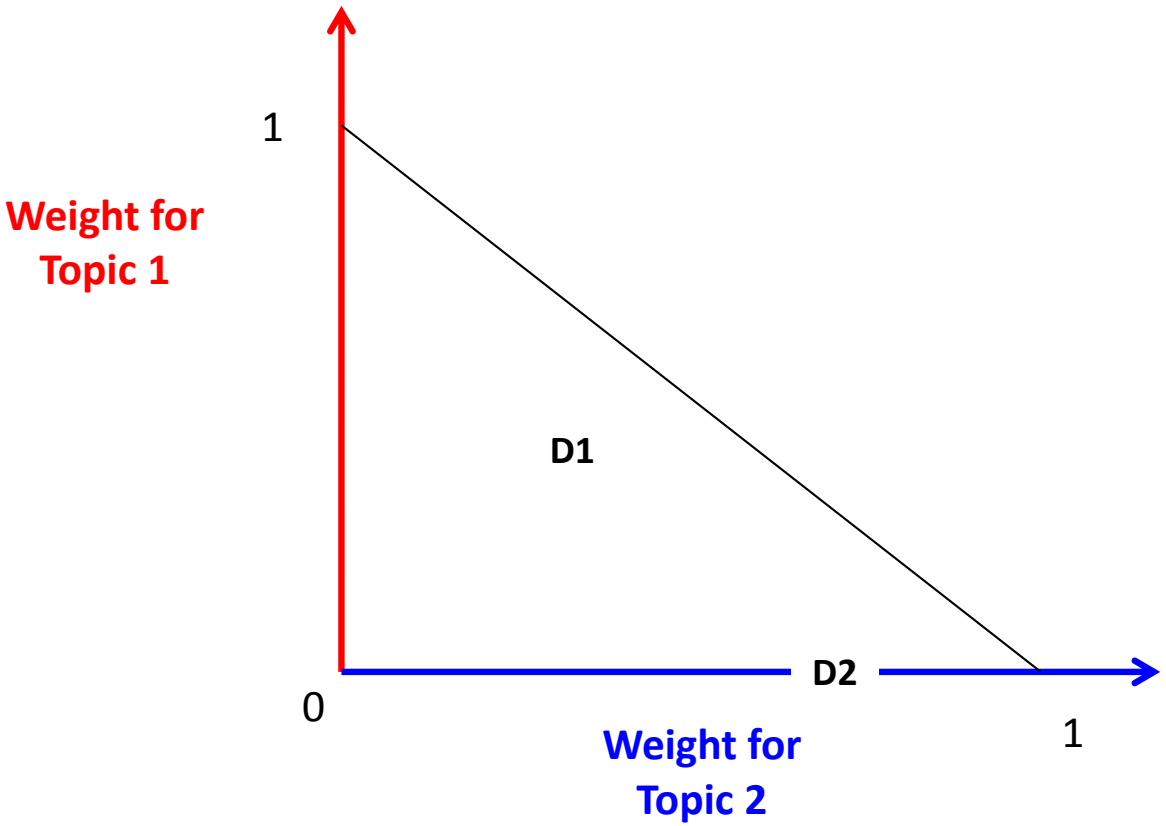
Example from: P. Smyth

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)



Example from: P. Smyth

Generative Model for Documents

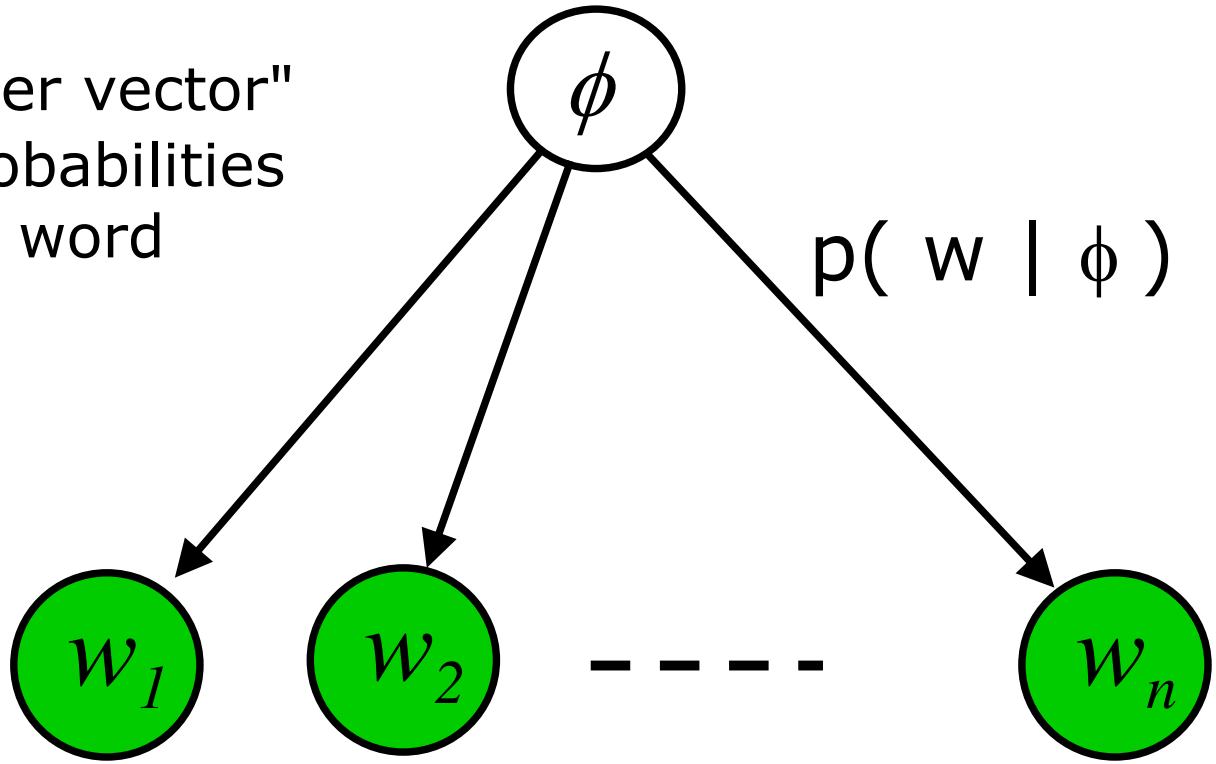
- Topic model is a simple “forward generative” model for observed data (counts of words in docs)
- For each document in corpus
  - For each word in document
    - Sample a topic from P(topics | document)
    - Given the topic, sample a word from P(words | topic)
  - End
- End
- Learning this model uses Bayes rule

Example from: P. Smyth

Graphical Model

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

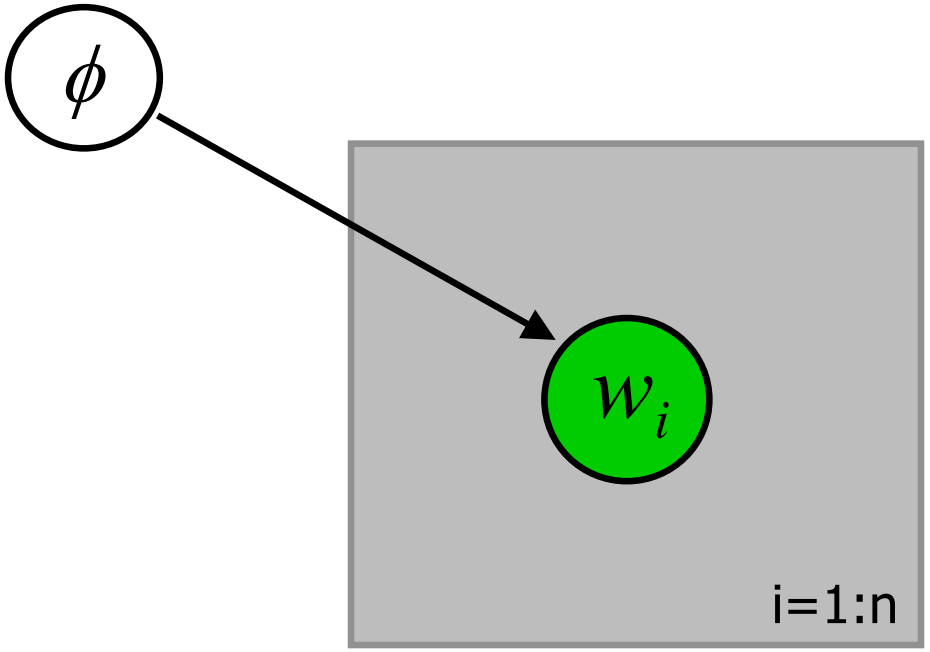
ϕ = "parameter vector"
= set of probabilities
one per word



Example from: P. Smyth

Graphical Model – another view

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



This is “plate notation”

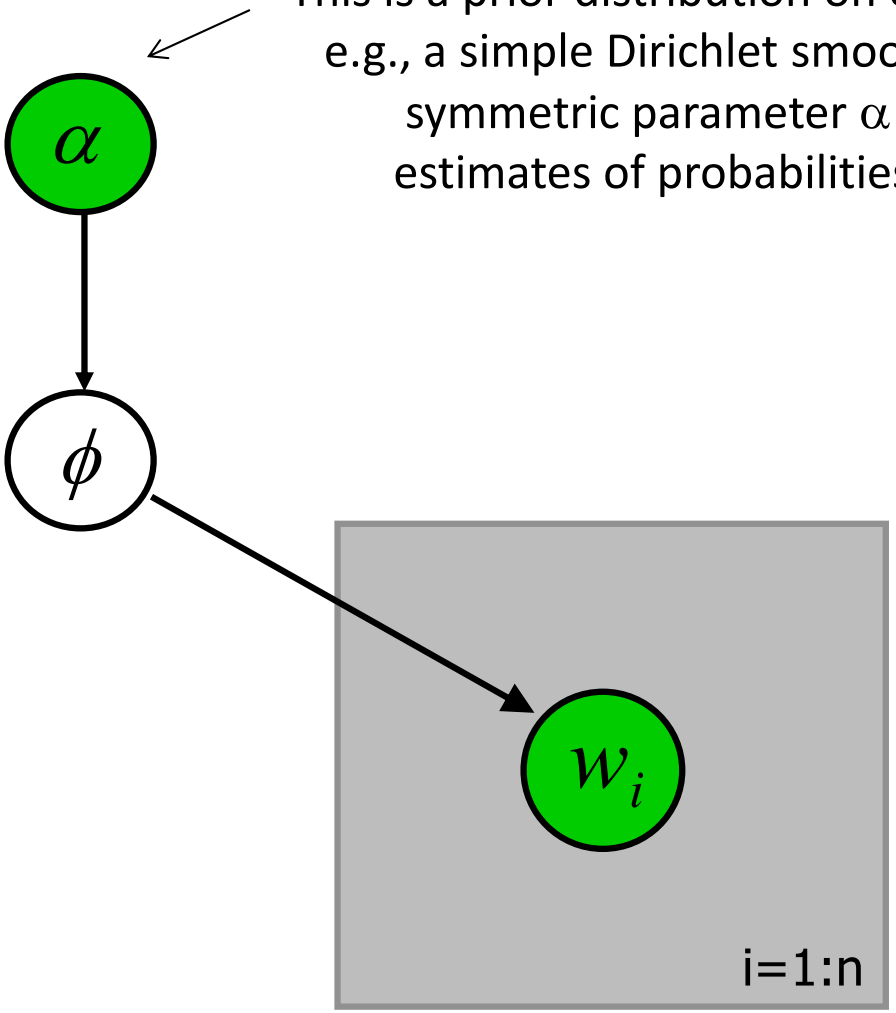
Items inside the plate
are conditionally independent
given the variable outside
the plate

There are “n” conditionally
independent replicates
represented by the plate

Example from: P. Smyth

Graphical Model - extended

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



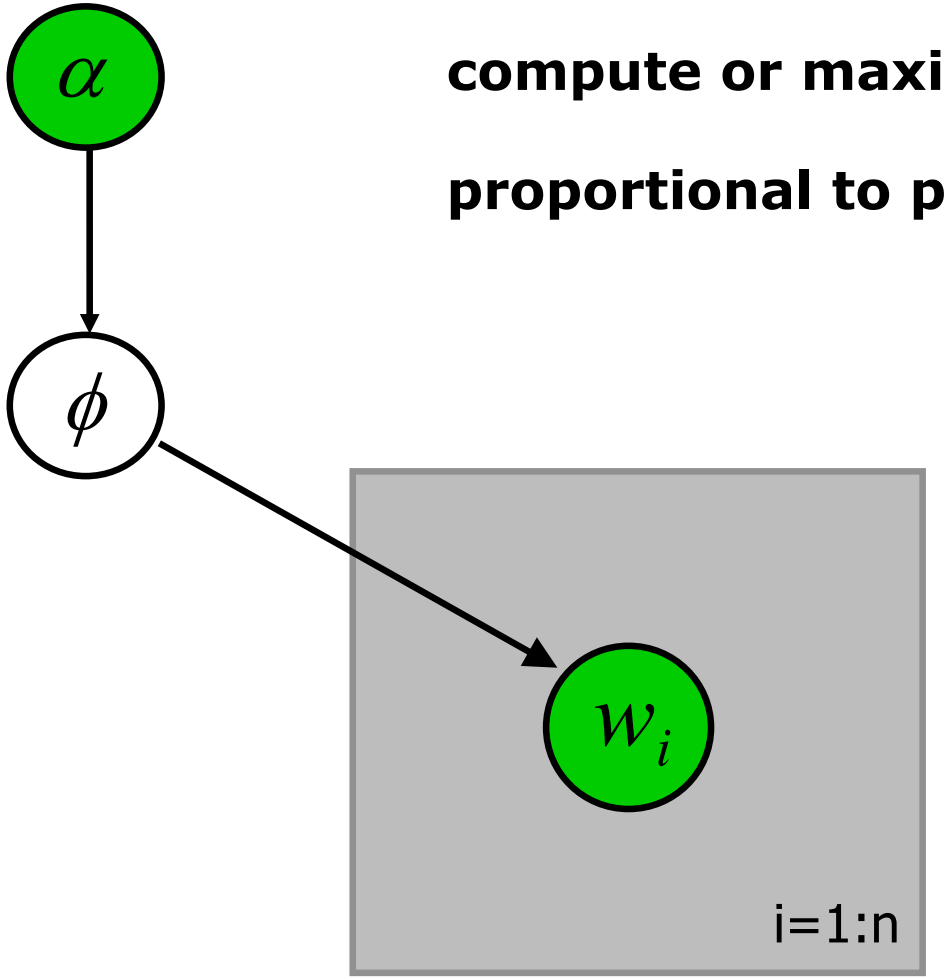
Example from: P. Smyth

Graphical Model, cont.

Learning:

compute or maximize p(ϕ | words, α)

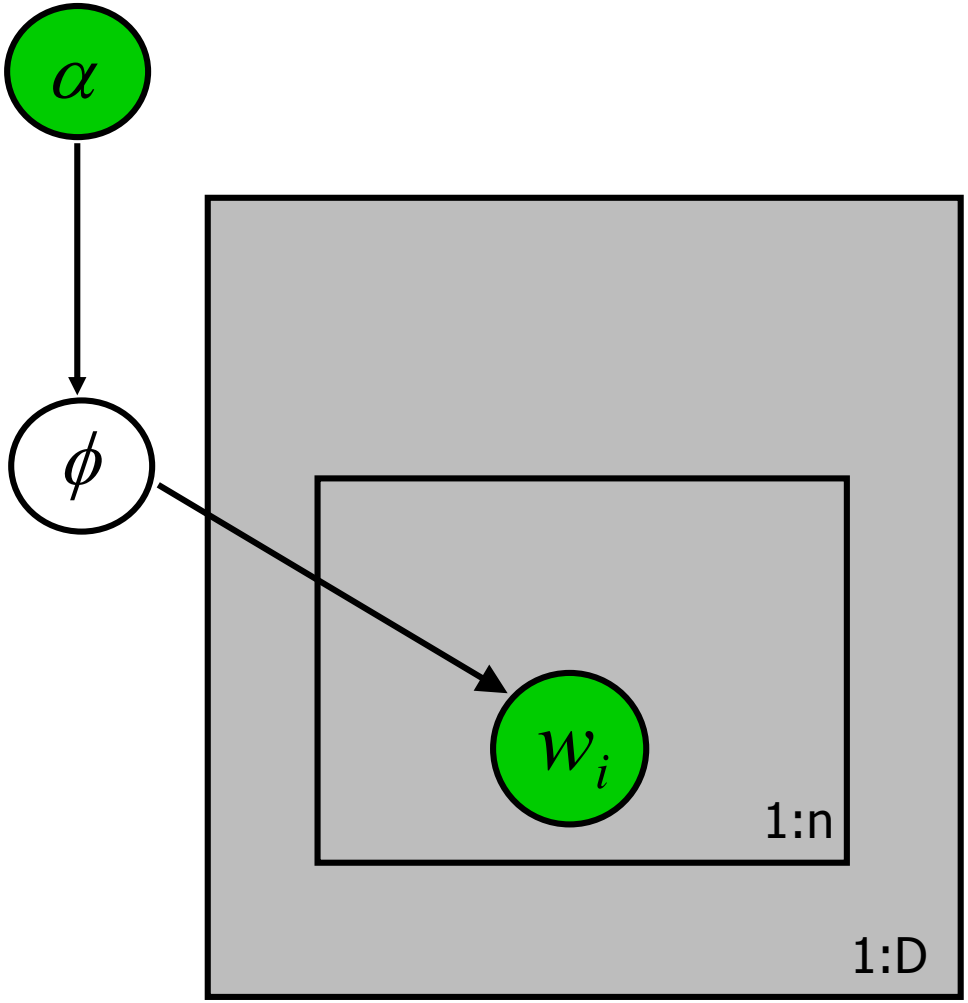
proportional to p(words | ϕ) p($\phi|\alpha$)



Example from: P. Smyth

Graphical Model, multiple documents

p(corpus | ϕ) = Π p(doc | ϕ)

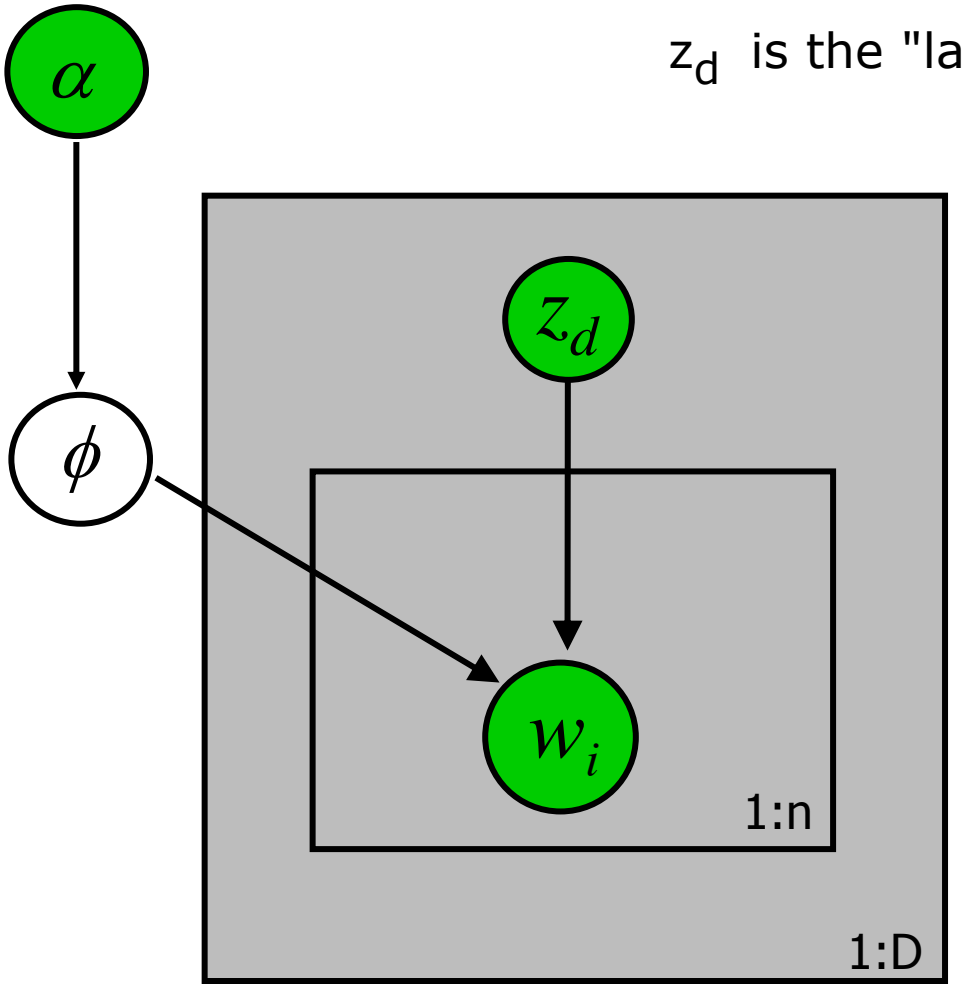


Example from: P. Smyth

Graphical Model, different document types

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

z_d is the "label" for each doc

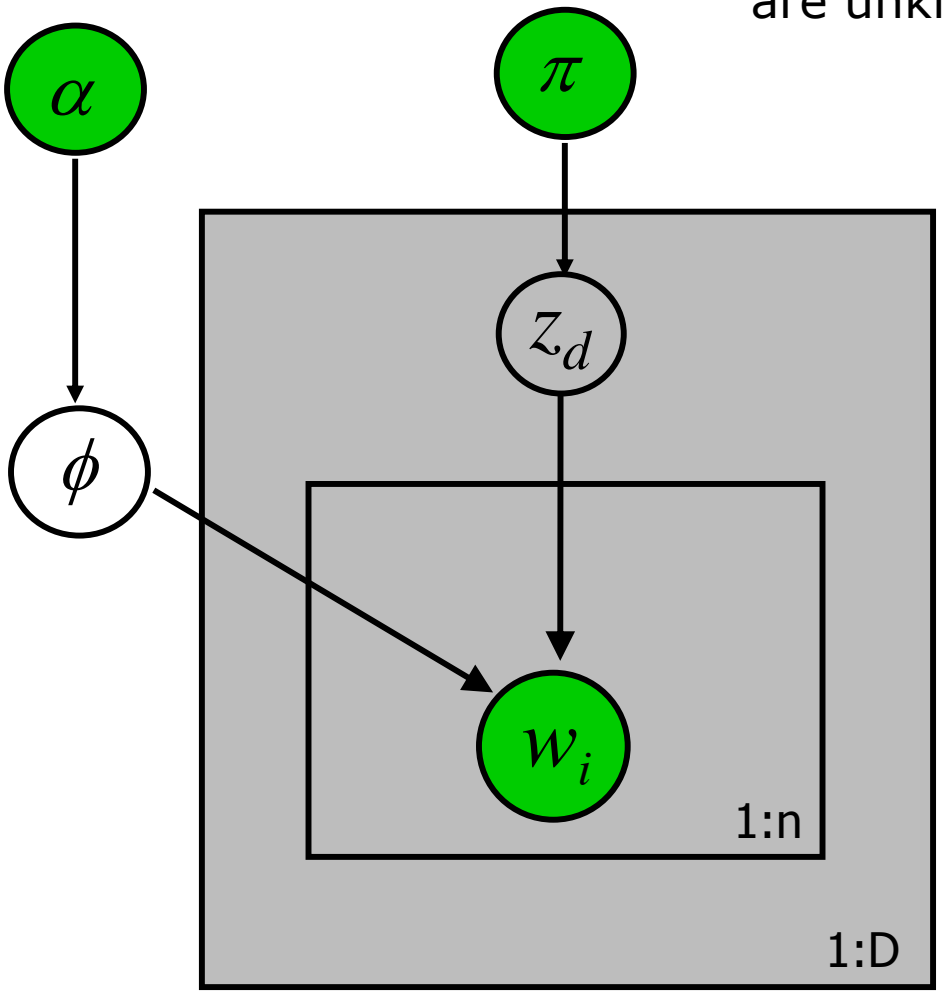


Different multinomials, depending on the value of z_d (discrete)

ϕ now represents |z| different multinomials

Example from: P. Smyth

Graphical Model, unknown document types



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

Not hopeless :)

Can learn about both z and θ

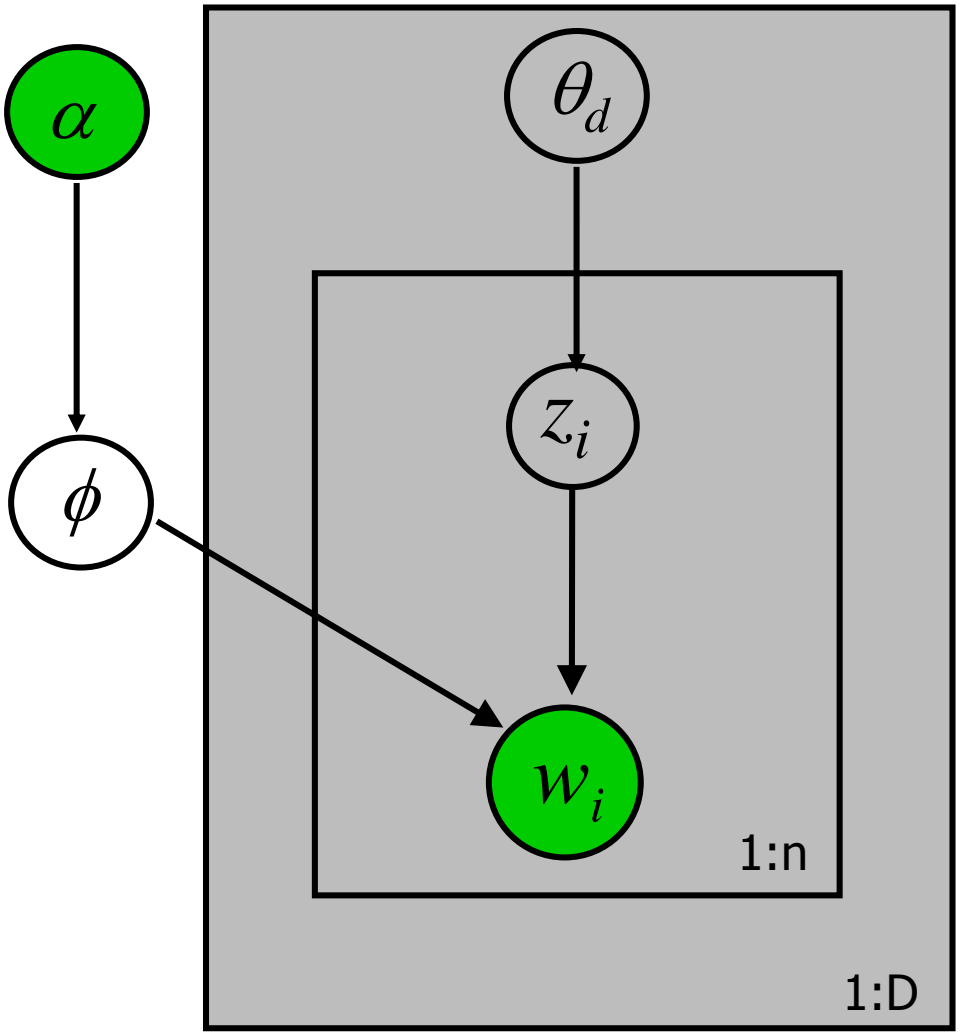
e.g., EM algorithm

This gives probabilistic clustering

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

Example from: P. Smyth

Topic Model



z_i is a "label" for each word

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

= a "topic"

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

Example from: P. Smyth

Mixture Model Equation

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

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

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

Example from: P. Smyth

LDA

- Assume data was generated by a generative process:
- q is a document - made up from topics from a probability distribution
- z is a topic made up from words from a probability distribution
- w is a word, the only real observables (N=number of words in all documents)

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n

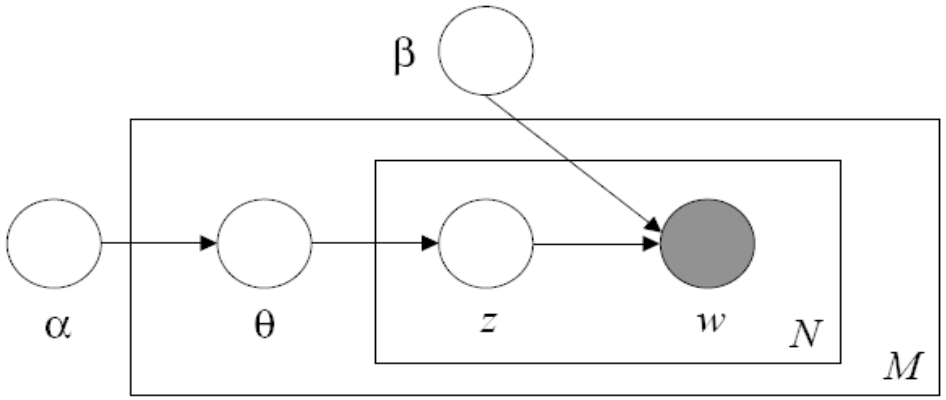
α=per-document topic distributions

The LDA equations

$$(2)\ p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta)$$

$$(3)\ p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)\, p(w_n|z_n, \beta)\right) d^k\theta$$

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)\, p(w_{dn}|z_{dn}, \beta)\right) d^k\theta_d$$



Which can be solved via advance computational techniques

see Blei, et al 2003

LDA output

The result can be an often-useful classification of documents into topics, and a distribution of each topic across words:

"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

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.

Another look at LDA

- Model: Topics made up of words used to generate documents

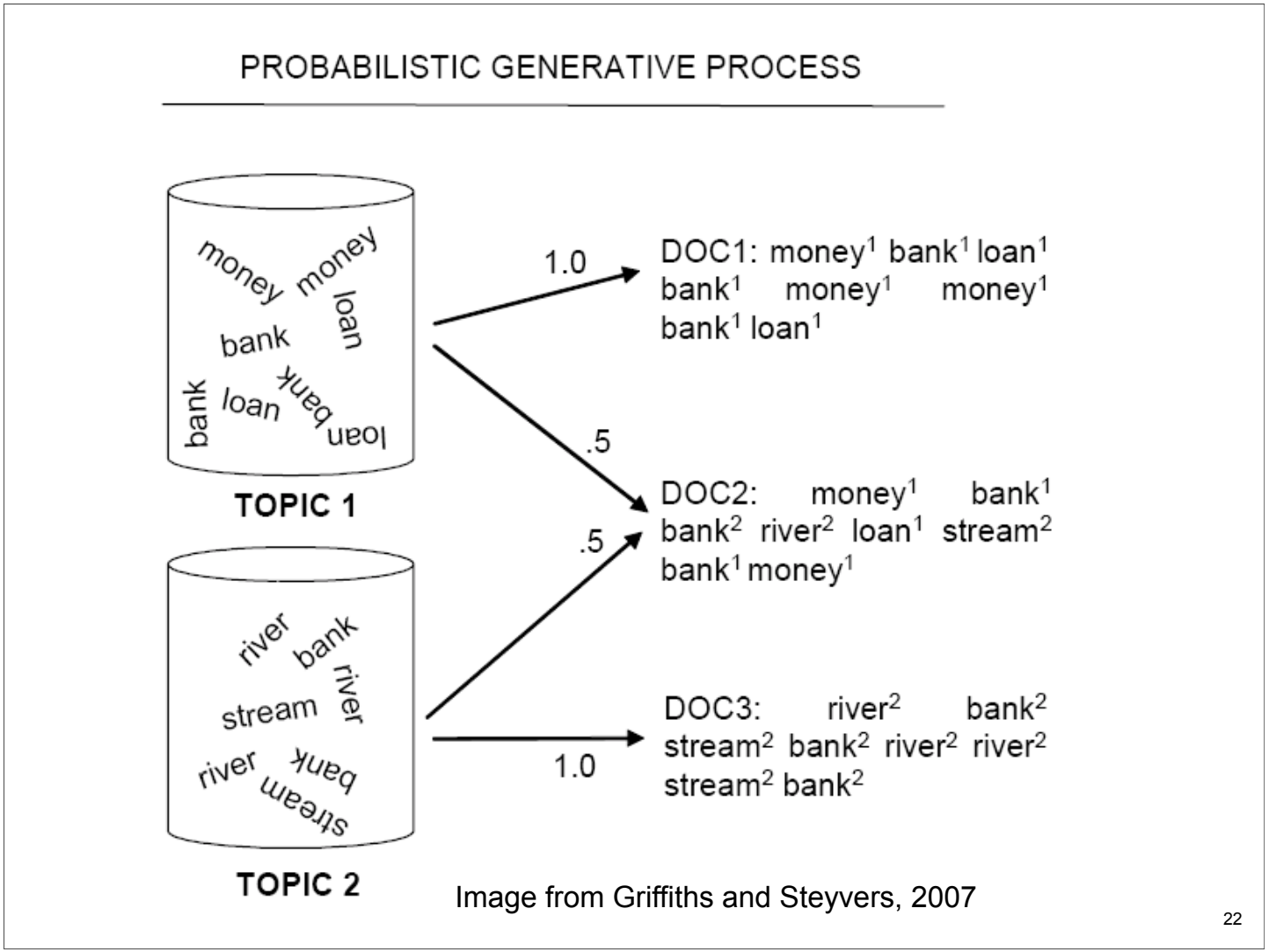
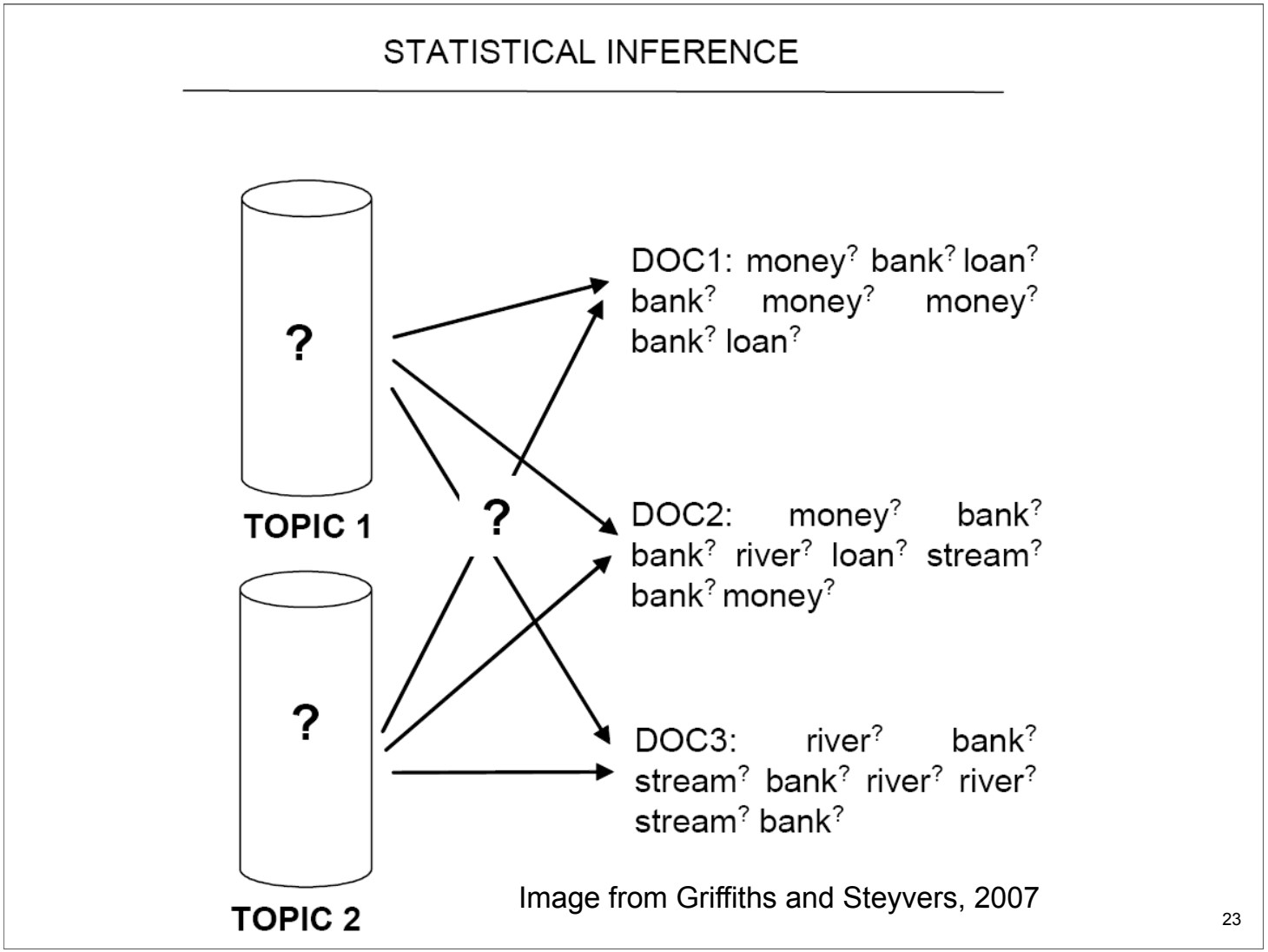


Image from Griffiths and Steyvers, 2007

Another Look at LDA

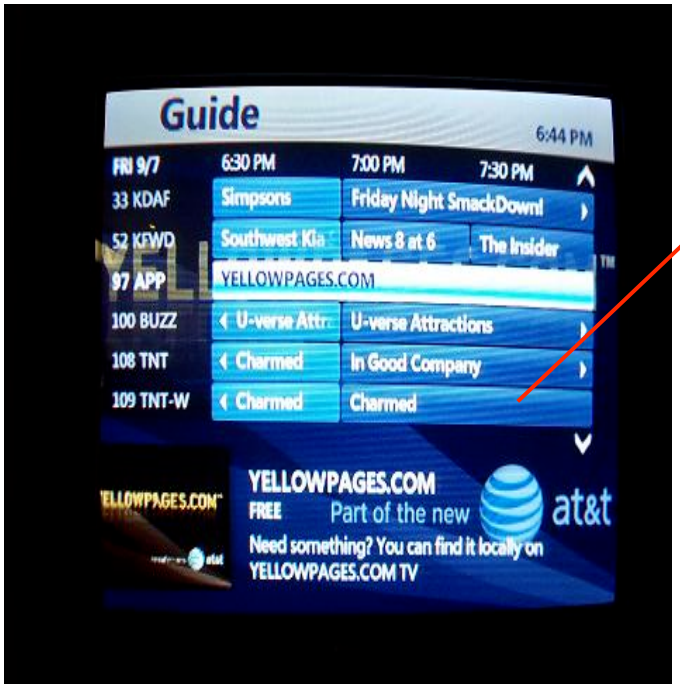
- Reality: Documents observed, infer topics



Case Study: TV Listings

- Use text to make recommendations on TV shows

The program guide data



Charmed - When three evil sisters magically steal the Charmed Ones' identities and powers, Piper, Phoebe and Paige must convince Chris that they are the real Charmed Ones

Data Issues

- 10013|In Harm's Way|In Harm's Way|A tough Naval officer faces the enemy while fighting in the South Pacific during World War II.|A tough Naval officer faces the enemy while fighting in the South Pacific during World War II.|en-US| Movie,NR Rating|Movies:Drama|||165|1965|USA||||||STARS-3||NR|John Wayne, Kirk Douglas, Patricia Neal, Tom Tryon, Paula Prentis s, Burgess Meredith|Otto Preminger||||Otto Preminger|

Parsed Program Guide entries – 2 weeks, ~66,000 programs, 19,000 words

- Collapse on series (syndicated shows are still a problem)
- Stopwords/stemming, duplication, paid programming, length normalization

Data Preprocessing

- Combine shows from one series into a ‘canonical’ f

300726 charmed Series Drama
Alyssa_Milano Brian_Krause
Holly_Marie_Combs Rose_McGowan
Shannen_Doherty

+

When three evil sisters magically steal the Charmed Ones' identities and powers, Piper, Phoebe and Paige must convince Chris that they are the real Charmed Ones
A warlock makes a pact with Tempus, a demon who will turn back time until the warlock can succeed in killing all the charmed ones.
On their first anniversary of becoming witches, the sisters face a cloven-hoofed demon called Abraxas, that steals the Book of Shadows and undoes their spells.
The Cleaners kidnap Wyatt and erase his existence after the child accidentally brings a dragon to life; Paige is sidelined by sexual harrassment issues.

Using standard information Retrieval procedures

→

Charmed Series Drama
Rose_McGowan,
Holly_Marie_Combs, Alyssa_Milano,
Brian_Krause, Shannen_Doherty
angel attack attempt babi billi book cast charm chri cole curs death demon destroi discov elder evil face feel find free futur ghost goe help innoc kidnap kill learn leo live love magic murder mysteri name paig piper plan power premonit protect return save school seek shadow sister sourc spell steal stop take trap tri try turn wed witch wyatt

Corpus consists of these TV program “documents”

19

Results

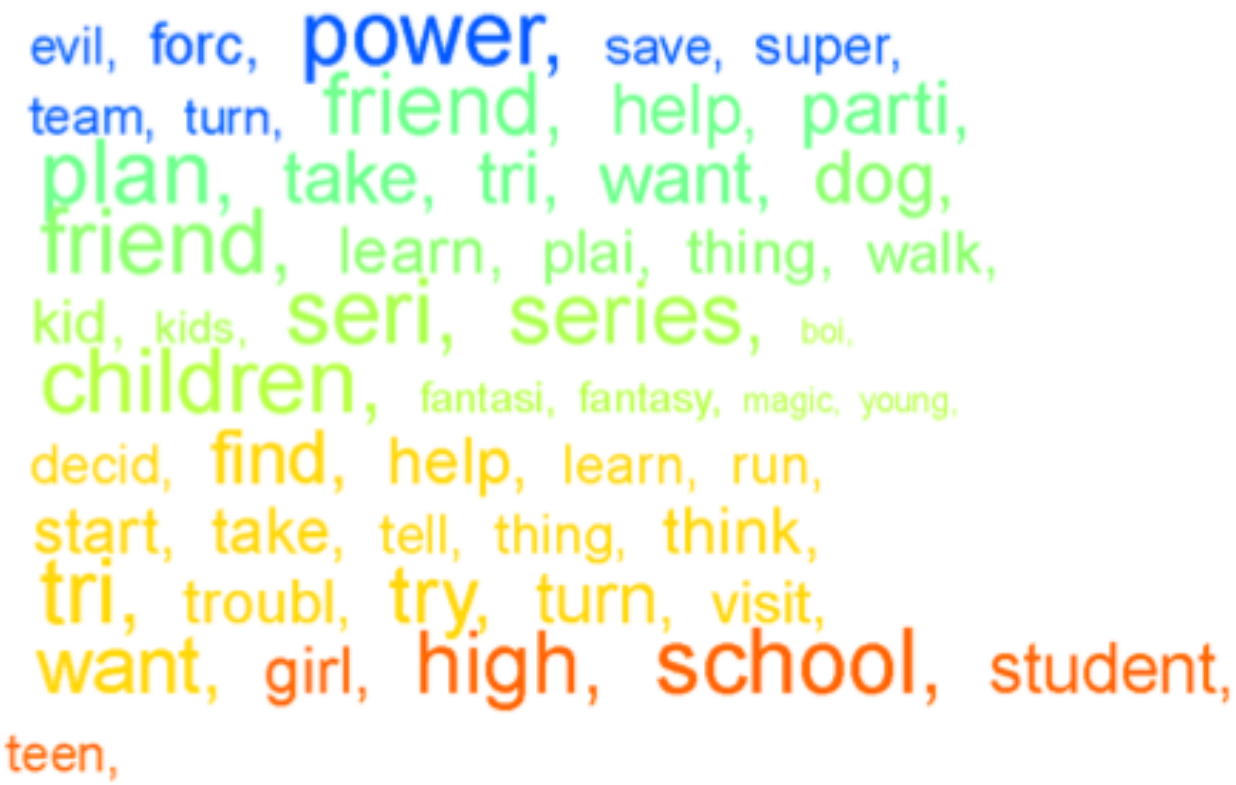
- We fit LDA
  - Results in a full distribution of words, topics and documents
  - Topics are unveiled which are a collection of words

TOPIC_45	prob(w|t_45)	TOPIC_46	prob(w|t_46)	TOPIC_47	prob(w|t_47)
parentprogram	0.08749	music	0.33400	nrrate	0.03364
series	0.04938	perform	0.06735	nrrating	0.03302
seri	0.04382	video	0.04795	josã	0.02490
god	0.03907	musical	0.04595	ndez	0.01873
church	0.03676	artist	0.04056	n	0.01651
christian	0.03473	song	0.03656	lez	0.01602
faith	0.03066	concert	0.02592	joaquã	0.01505
word	0.02672	featur	0.02191	guez	0.01207
with	0.02618	band	0.01640	marã	0.01082
bibl	0.01981	singer	0.01390	el	0.01006
jesu	0.01967	gospel	0.01252	renã	0.00923
pastor	0.01886	musician	0.01214	joven	0.00853
christ	0.01791	mtv	0.01202	hombr	0.00756
ministri	0.01479	sing	0.01127	germã	0.00742
inspir	0.01424	jazz	0.01077	delgado	0.00687
servic	0.01411	award	0.01039	ramã	0.00687
messag	0.01357	countri	0.00939	ntintan	0.00680
teach	0.01357	pop	0.00889	valdã	0.00673
biblic	0.01221	latin	0.00826	npardavã	0.00666
todai	0.01167	sound	0.00789	miguelm	0.00652

Results

- For user modelling, consider the collection of shows a single user watches as a ‘document’ – then look to see what topics (and hence, words) make up that document

Another Use


evil, forc, power, save, super,
team, turn, friend, help, parti,
plan, take, tri, want, dog,
friend, learn, plai, thing, walk,
kid, kids, seri, series, boi,
children, fantasi, fantasy, magic, young,
decid, find, help, learn, run,
start, take, tell, thing, think,
tri, troubl, try, turn, visit,
want, girl, high, school, student,
teen,

Show mining via text

Input: democratic presidential candidate (the first three shows)

226.158	democratic presidential candidate debate post-debate program
217.297	democratic presidential candidates debate analysis
215.648	democratic presidential candidates debate
179.264	republican presidential candidate debate post-debate program
166.694	democratic presidential debate 2008
155.663	cnn/nevada state democratic party presidential primary debate
149.004	democratic presidential post debate analysis
139.174	nevada democratic party presidential primary debate
139.174	des moines register democratic presidential debate
128.194	republican presidential candidates debate
118.745	hdnet presents: the iowa brown & black presidential forum
114.448	presidential forum
111.286	washington week with gwen ifill and national journal
111.13	late edition with wolf blitzer
108.66	des moines register presidential debates
104.562	the candidates 2008
100.643	meet the press

Show mining via text

Input: Battlestar galactica & The twilight zone

505.035	battlestar galactica
451.205	the twilight zone
217.191	the outer limits
200.673	star trek: enterprise
184.948	smallville
176.778	the x-files
175.978	steven spielberg presents pinky & the brain
171.027	star trek: voyager
169.716	star trek: the next generation
169.418	gargoyles
162.568	stargate atlantis
158.798	gene roddenberry's andromeda
154.304	mega science
153.998	doctor who
153.592	the planets
152.436	futurama

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

Topic Modeling: 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 E-mail 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

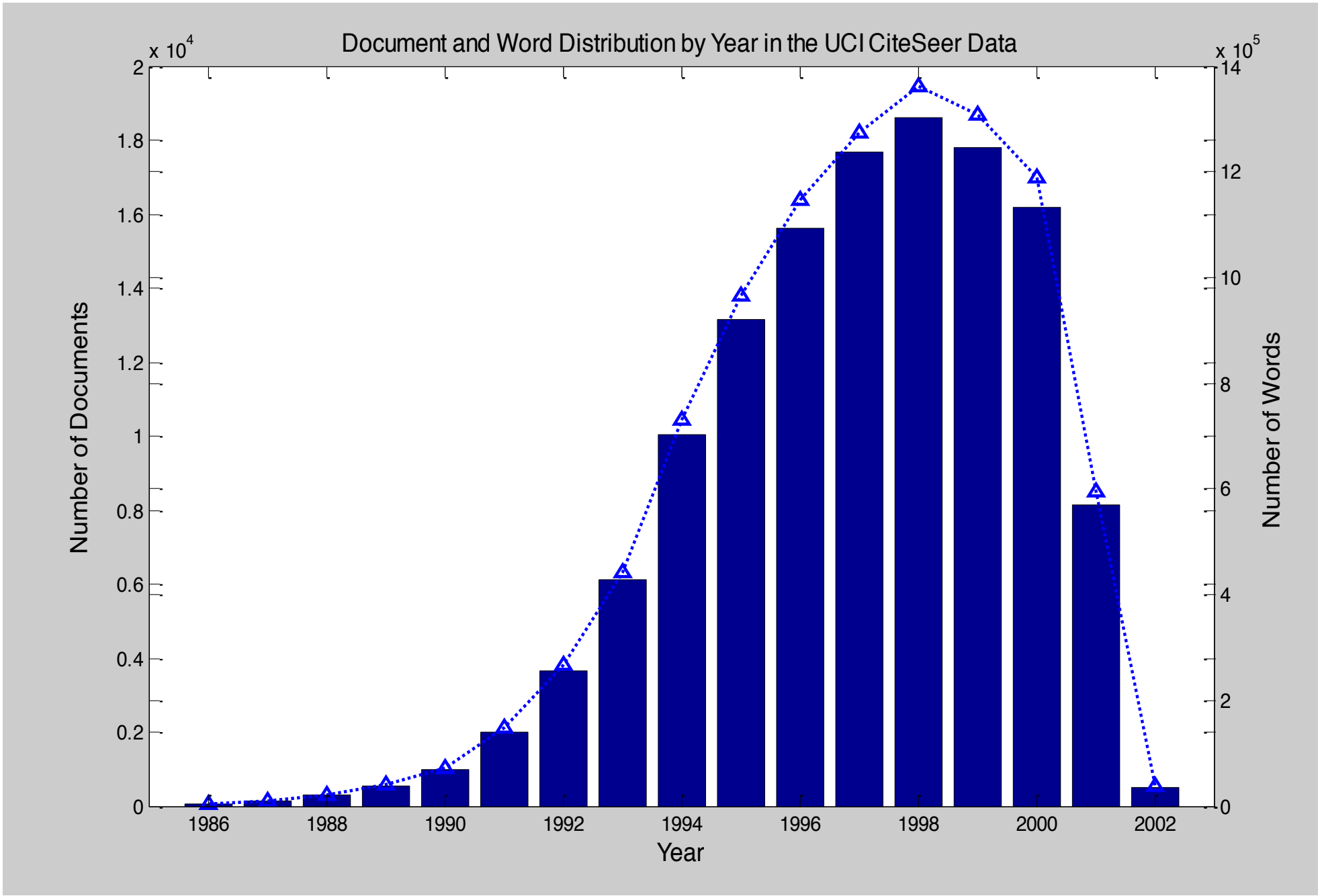
TOPIC 22	
WORD	**PROB.**
STATE	0.0253
PLAN	0.0245
CALIFORNIA	0.0137
POLITICIAN Y	0.0137
RATE	0.0131
BANKRUPTCY	0.0126
SOCAL	0.0119
POWER	0.0114
BONDS	0.0109
MOU	0.0107

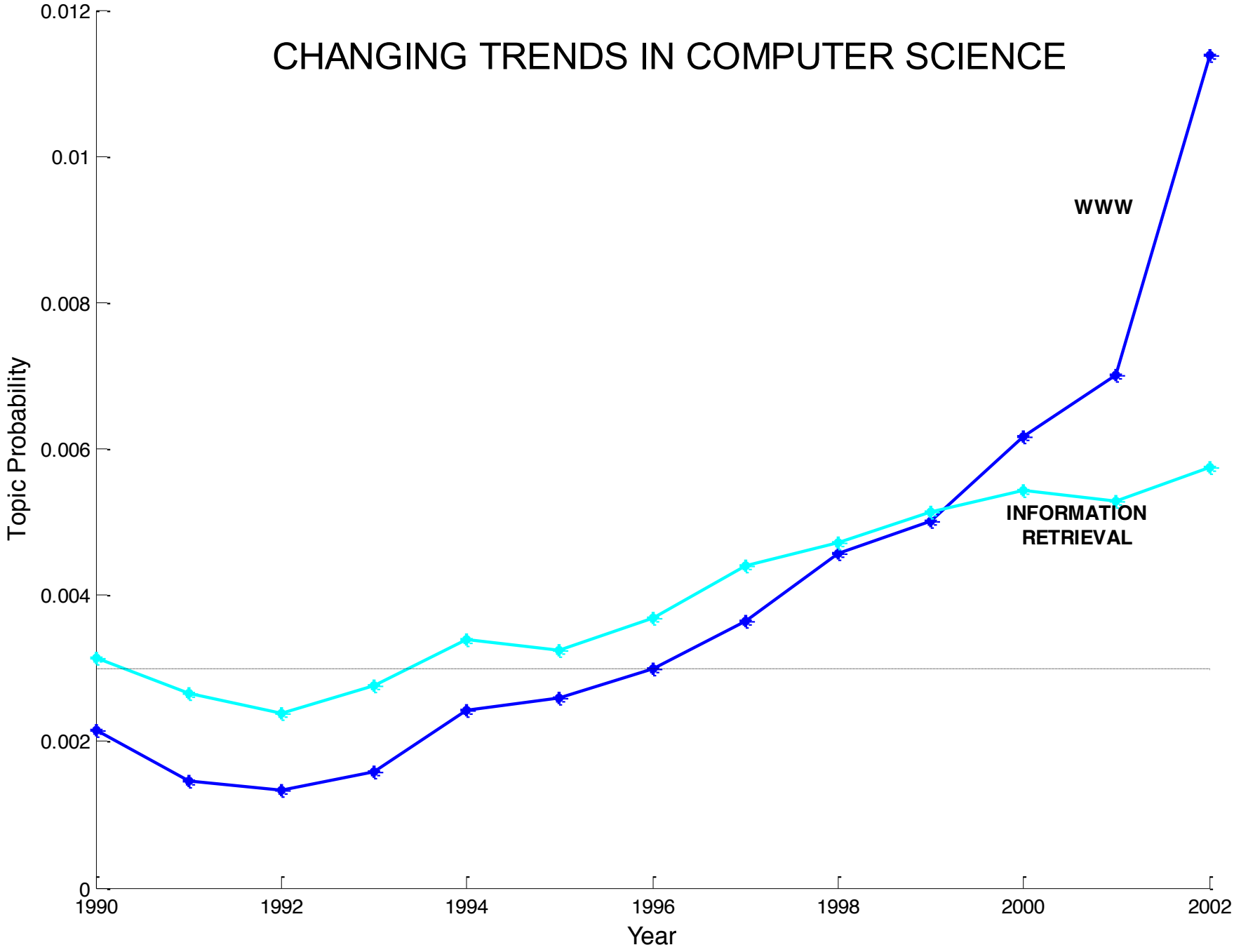
TOPIC 114	
WORD	**PROB.**
COMMITTEE	0.0197
BILL	0.0189
HOUSE	0.0169
WASHINGTON	0.0140
SENATE	0.0135
POLITICIAN X	0.0114
CONGRESS	0.0112
PRESIDENT	0.0105
LEGISLATION	0.0099
DC	0.0093

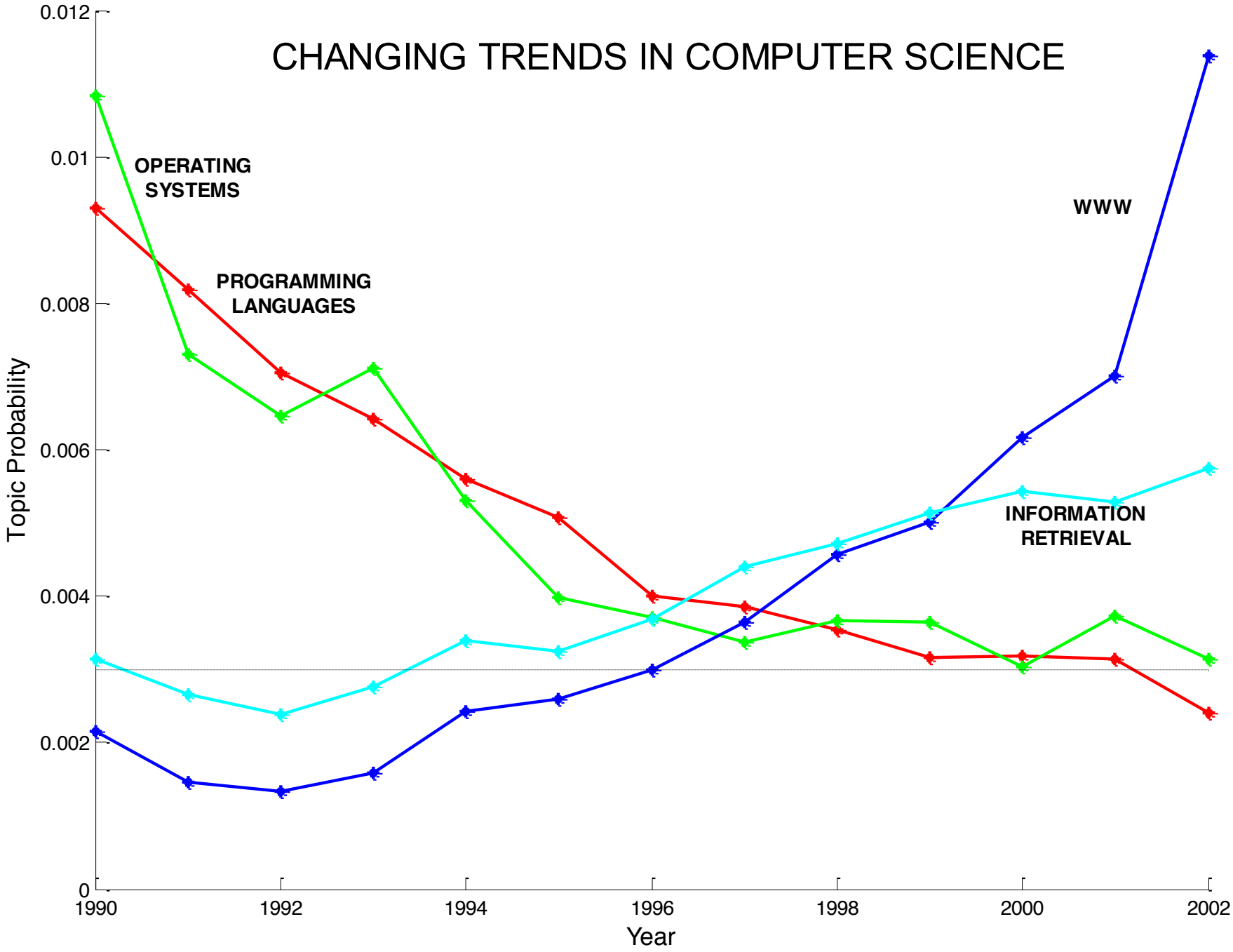
TOPIC 194	
WORD	**PROB.**
LAW	0.0380
TESTIMONY	0.0201
ATTORNEY	0.0164
SETTLEMENT	0.0131
LEGAL	0.0100
EXHIBIT	0.0098
CLE	0.0093
SOCALGAS	0.0093
METALS	0.0091
PERSON Z	0.0083

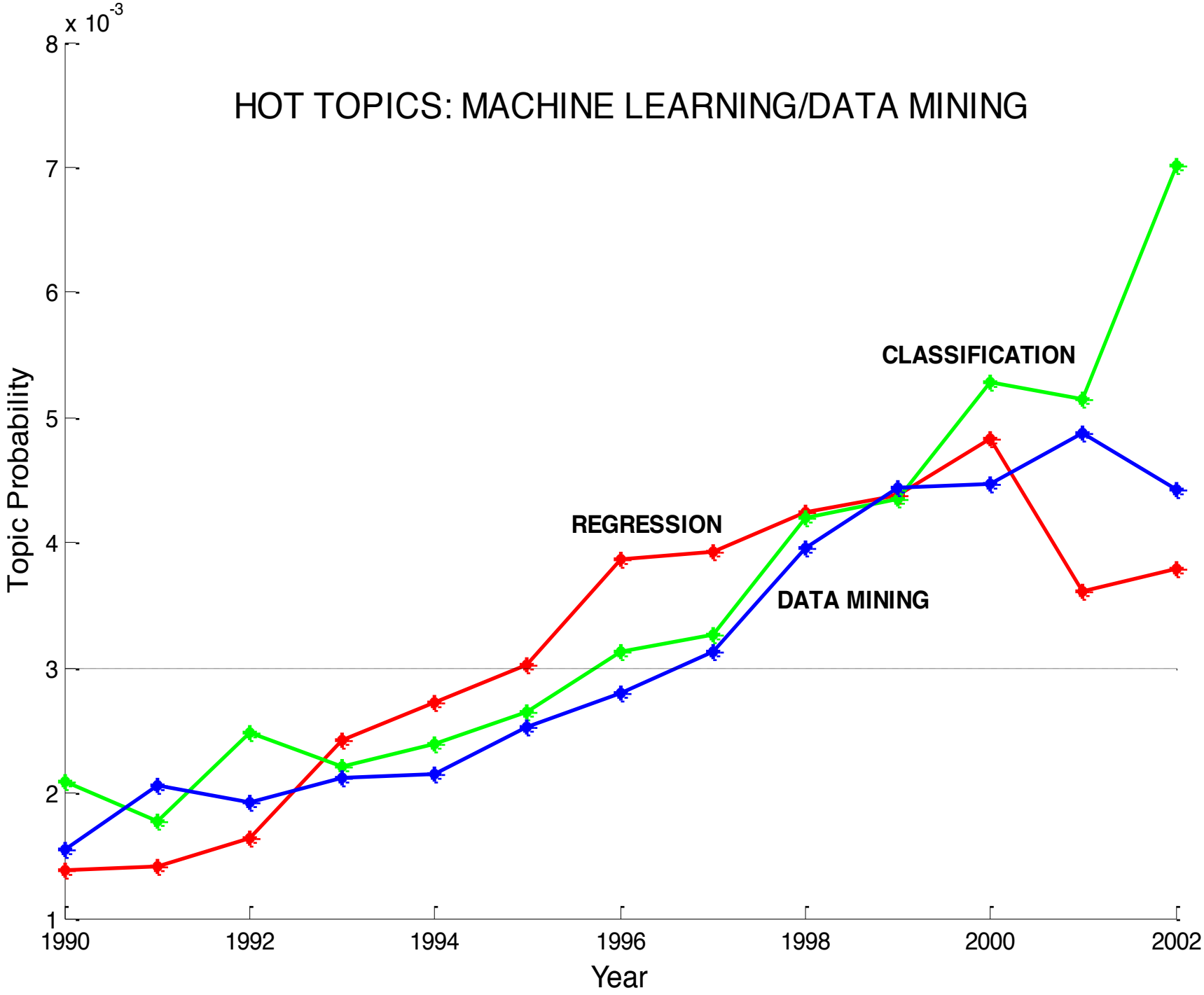
Temporal Patterns in Topics

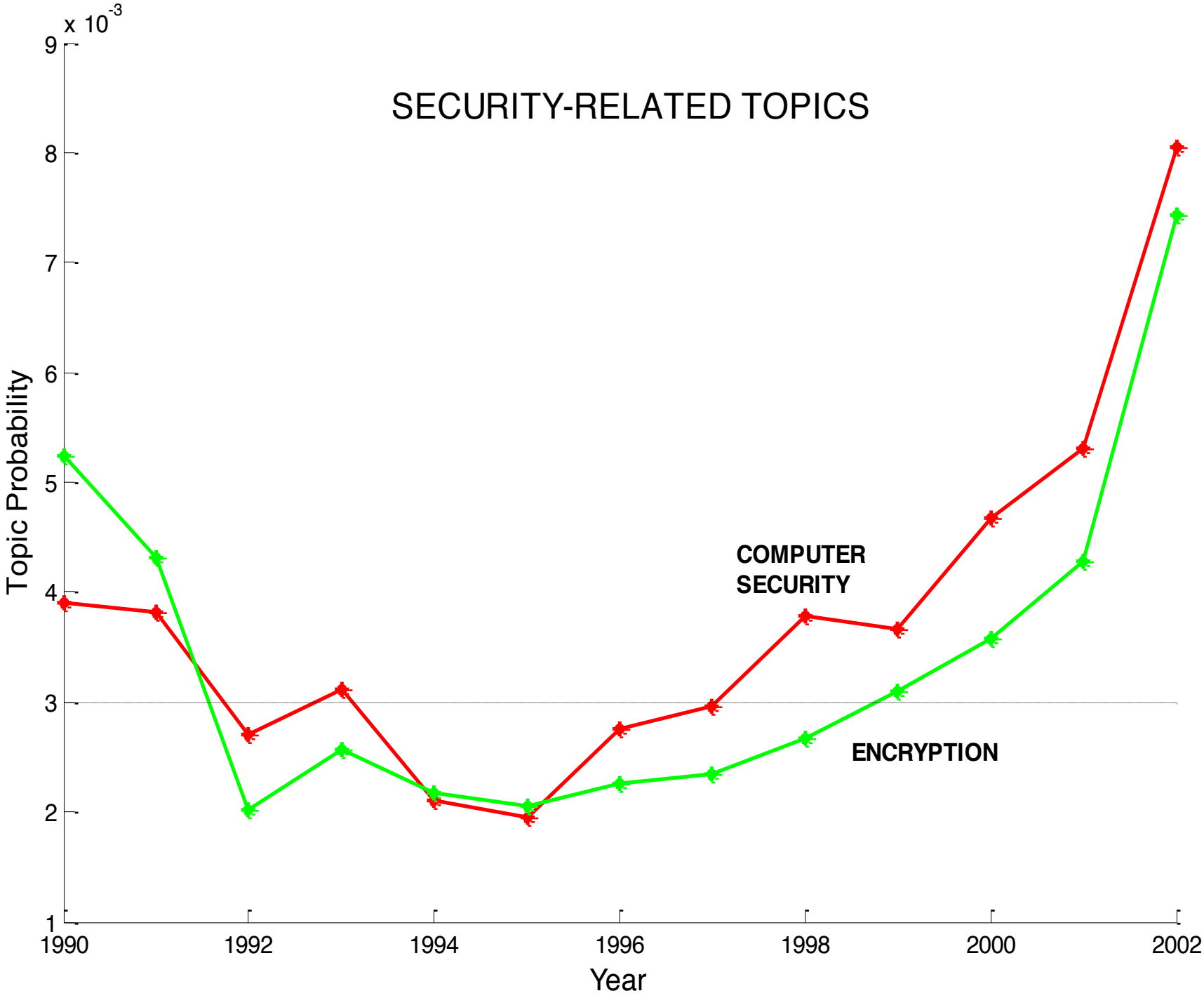
- Hot and Cold Topics
- CiteSeer Papers from 1986-2002, about 200K papers
- For each year, calculate the fraction of words assigned to each topic
- This results in a time-series for topics
  - Hot topics become more prevalent
  - Cold topics become less prevalent


Document and Word Distribution by Year in the UCI CiteSeer Data
x 10^4
x 10^5
Number of Documents
Number of Words
Year
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
2
4
6
8
10
12
14
1986
1988
1990
1992
1994
1996
1998
2000
2002

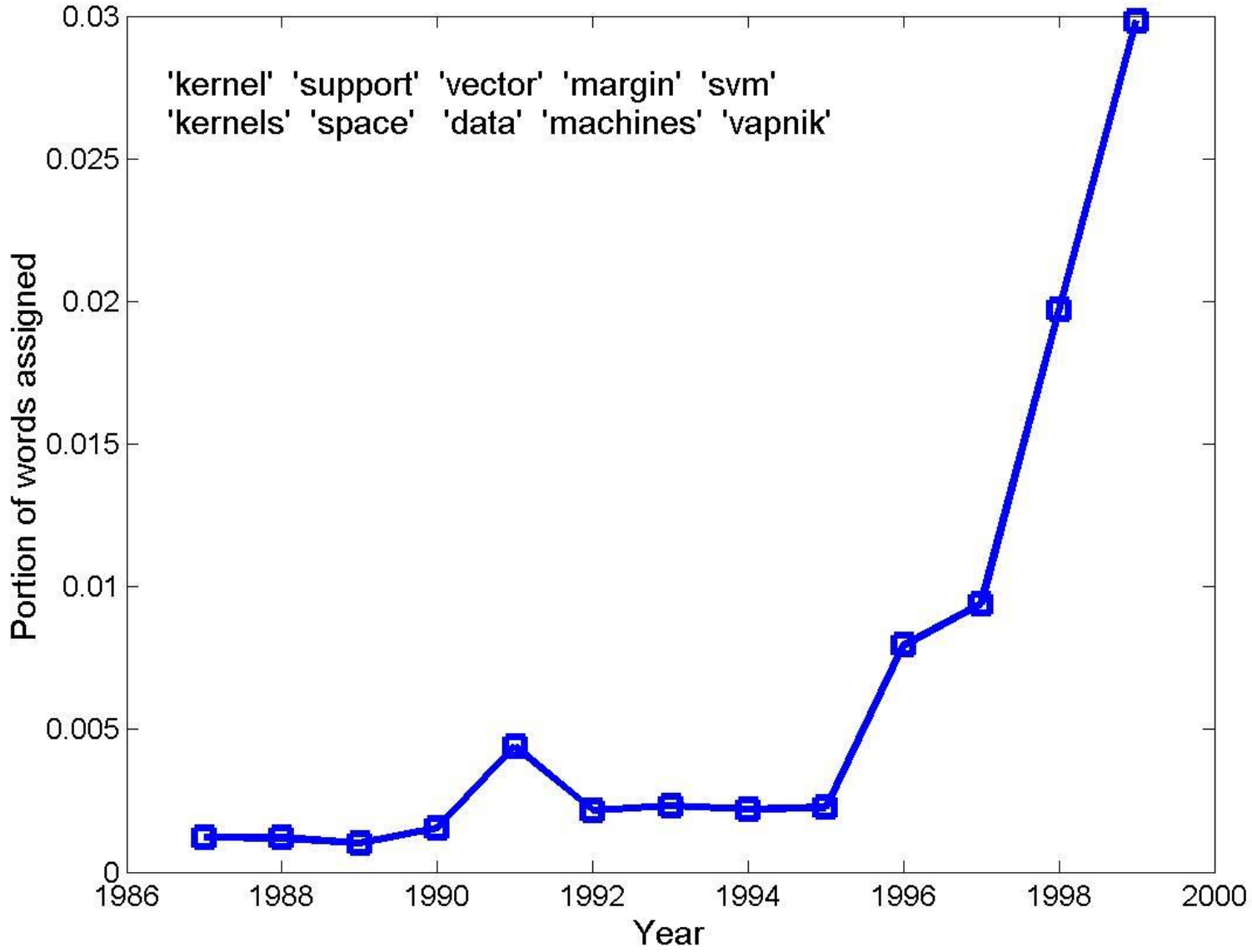

CHANGING TRENDS IN COMPUTER SCIENCE
Topic Probability
Year
0.012
0.01
0.008
0.006
0.004
0.002
0
1990
1992
1994
1996
1998
2000
2002
WWW
INFORMATION
RETRIEVAL


CHANGING TRENDS IN COMPUTER SCIENCE
OPERATING SYSTEMS
PROGRAMMING LANGUAGES
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

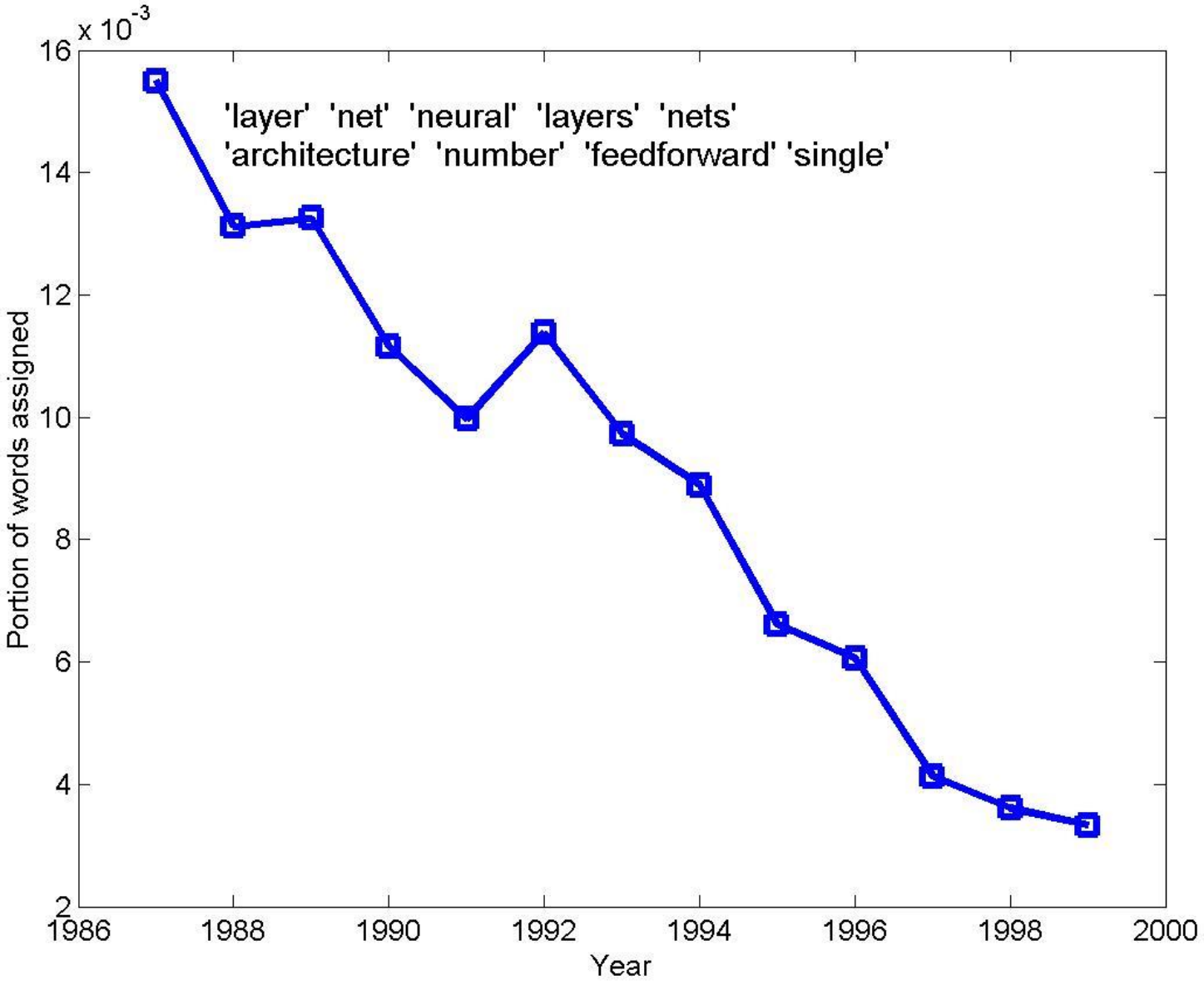

HOT TOPICS: MACHINE LEARNING/DATA MINING
x 10^-3
Topic Probability
Year
CLASSIFICATION
REGRESSION
DATA MINING
1
2
3
4
5
6
7
8
1990
1992
1994
1996
1998
2000
2002


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

NIPS conference, SVM Topic



NIPS conference: Neural Network topic



Topics over Time at NIPS

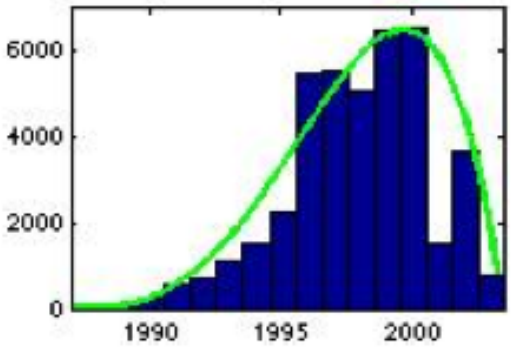
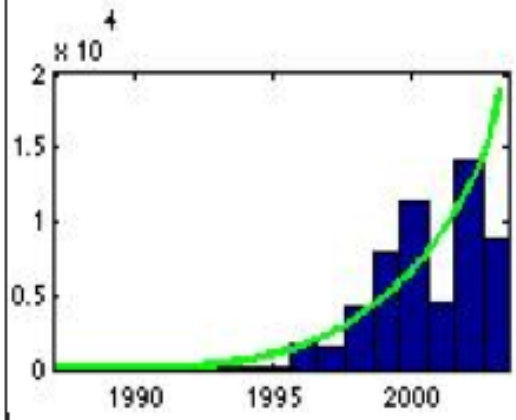
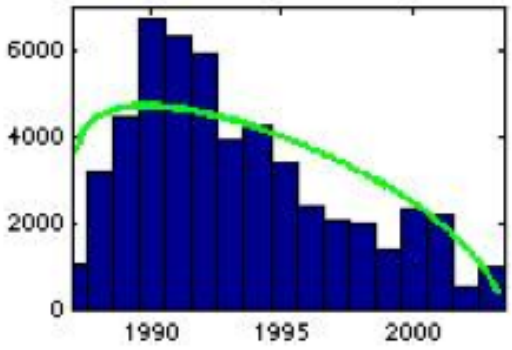
ICA		SVM		Speech Recognition	
					
ica	0.023	kernel	0.063	speech	0.055
independent	0.021	svm	0.025	recognition	0.039
natural	0.018	support	0.025	word	0.023
source	0.018	kernels	0.022	hmm	0.019
basis	0.017	vector	0.020	training	0.018
sources	0.015	training	0.020	system	0.017
separation	0.014	space	0.013	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 from Xuerie Wang and Andrew McCallum

All NIPS Topics over Time

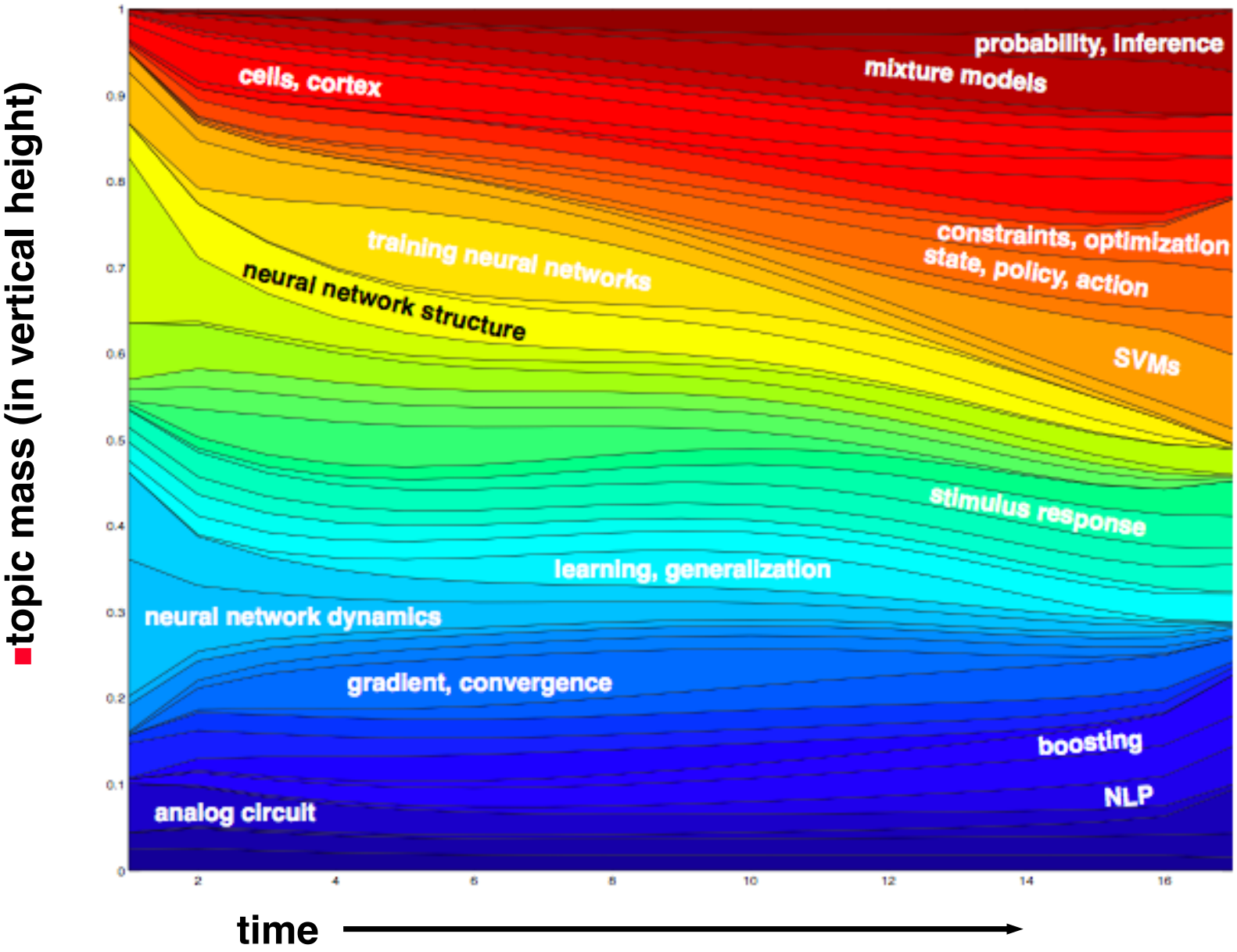
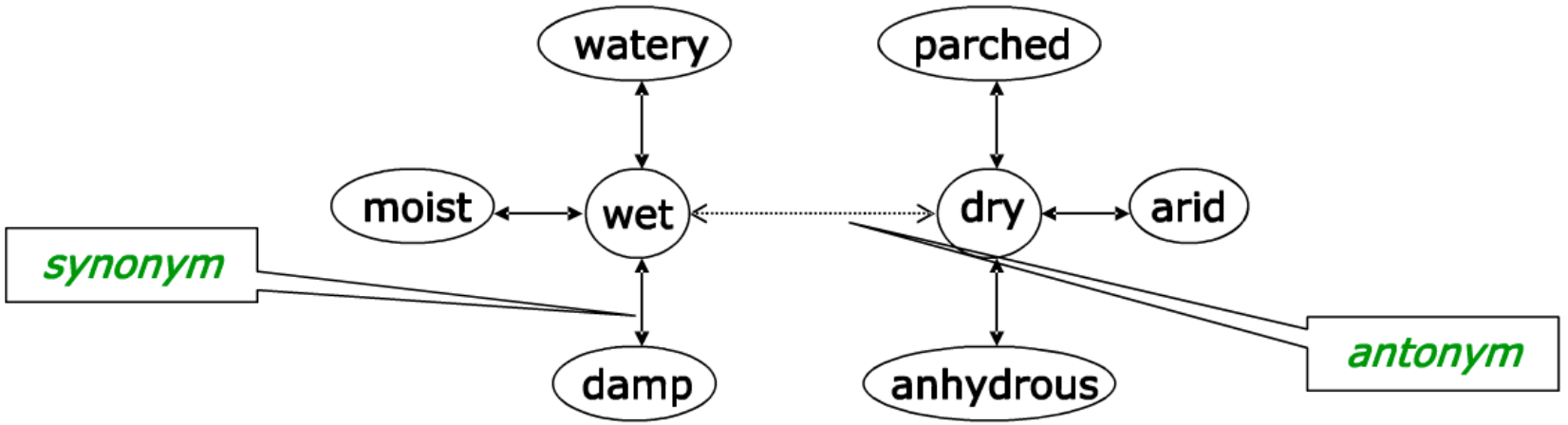


Figure from Xuerie Wang and Andrew McCallum

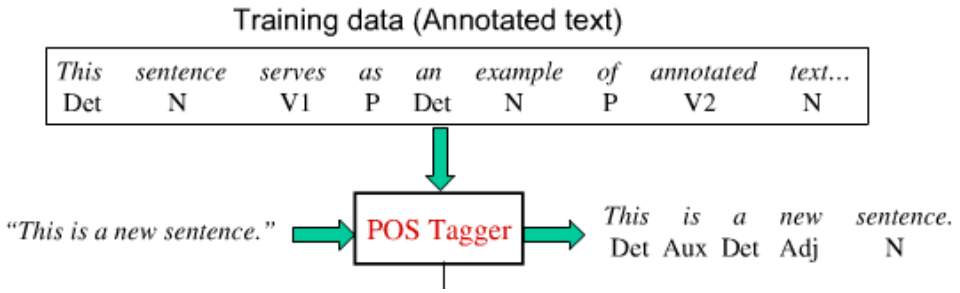
Text Mining: Helpful Data

- WordNet
  - An extensive lexical network for the English language
  - Contains over 138,838 words.
  - Several graphs, one for each part-of-speech.
  - Synsets (synonym sets), each defining a semantic sense.
  - Relationship information (antonym, hyponym, meronym …)
  - Downloadable for free (UNIX, Windows)
  - Expanding to other languages (Global WordNet Association)
  - Funded >$3 million, mainly government (translation interest)
  - Founder George Miller, National Medal of Science, 1991.

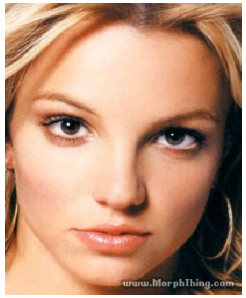


Text Mining – Other Topics

- Part of Speech Tagging
  - Assign grammatical tags to words (verb, noun, etc)
  - Helps in understanding documents : uses Hidden Markov Models



- Named Entity Classification
  - Classification task: can we automatically detect proper nouns and tag them
  - "Mr. Jones" is a person; "Madison" is a town.
  - Helps with dis-ambiguation: *spears*



Named Entity-Extraction

- Often a combination of
  - Knowledge-based approach (rules, parsers)
  - Machine learning (e.g., hidden Markov model)
  - Dictionary

- Non-trivial since entity-names can be confused with real names
  - E.g., gene name ABS and abbreviation ABS

- Also can look for co-references
  - E.g., “IBM today…… Later, the company announced…..”

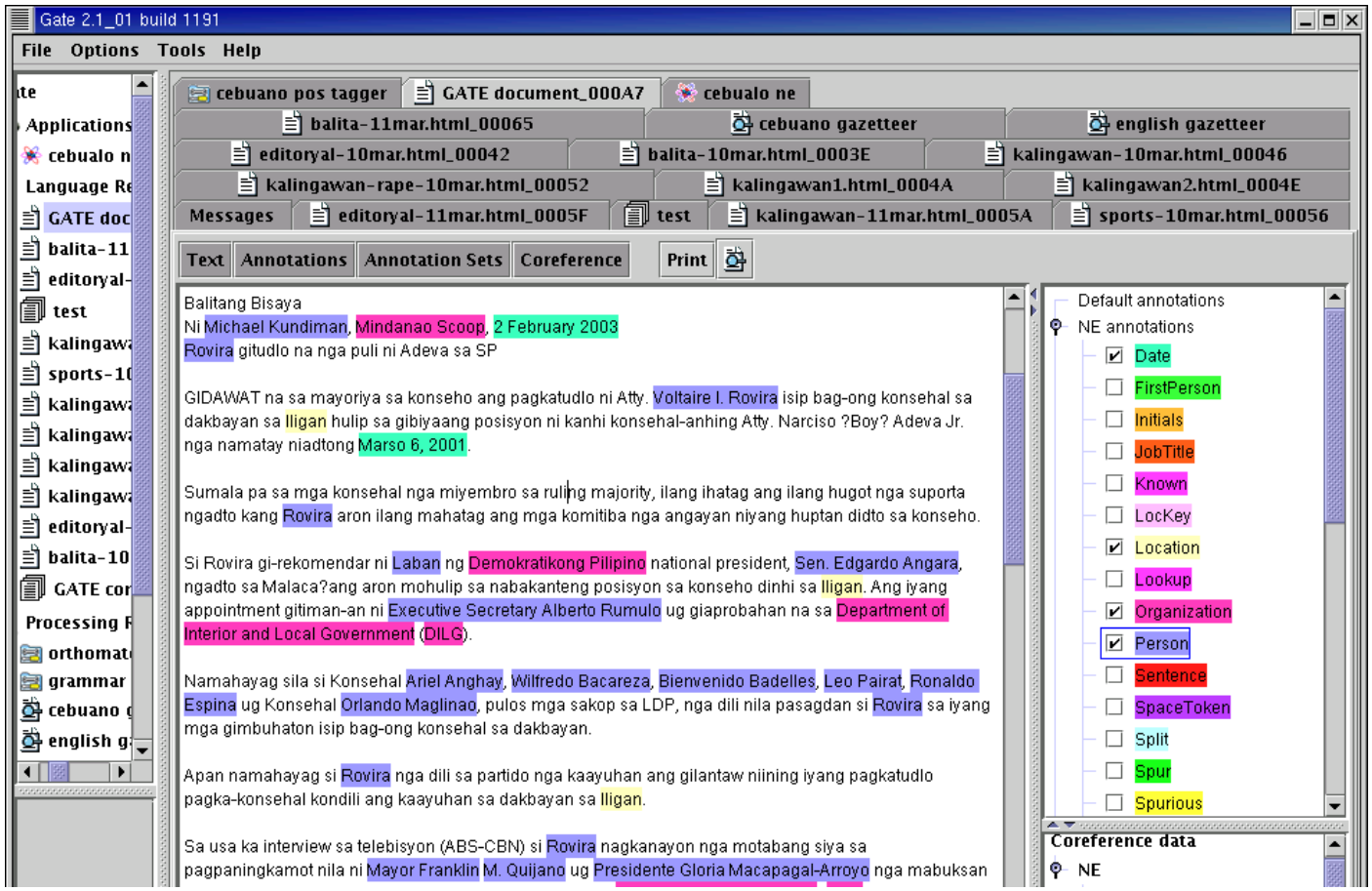
- Very useful as a preprocessing step for data mining,

  e.g., use entity-names to train a classifier to predict the category of an article

Example: GATE/ANNIE extractor

- GATE: free software infrastructure for text analysis (University of Sheffield, UK)
- ANNIE: widely used entity-recognizer, part of GATE
  http://www.gate.ac.uk/annie/

http://gate.ac.uk/sale/images/cebuano1.png



Information Extraction

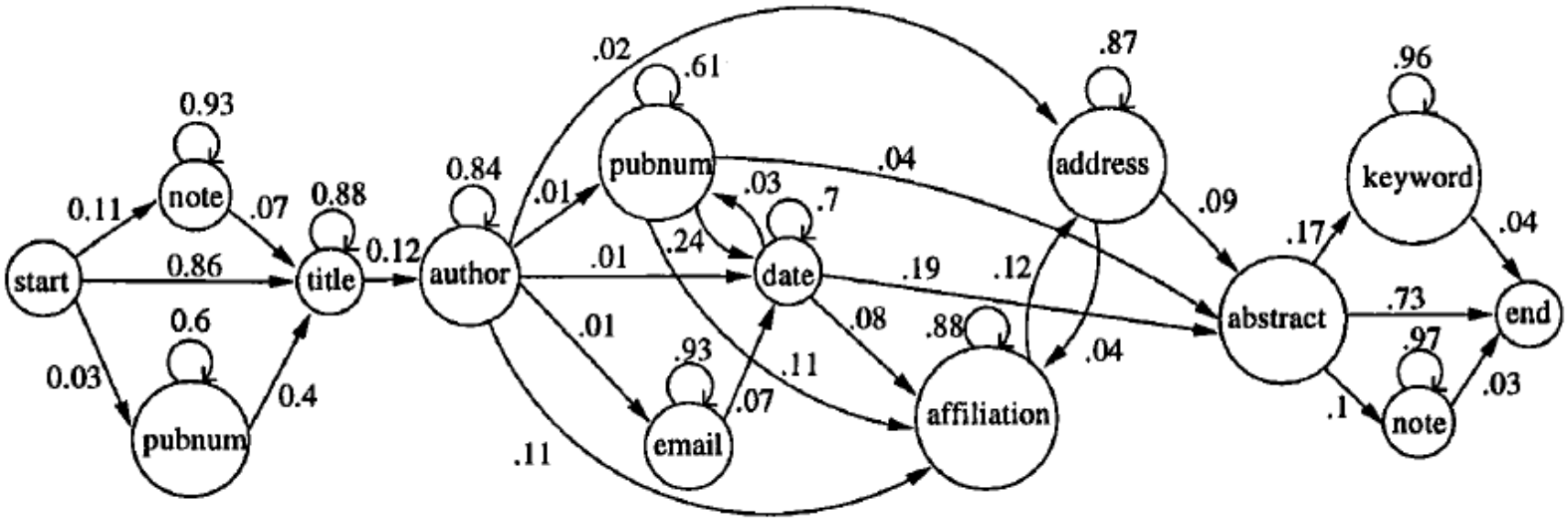


Figure 1: Example HMM. Each state emits words from a class-specific multinomial distribution.

From Seymore, McCallum, Rosenfeld, Learning Hidden Markov Model Structure for Information Extration, AAAI 1999

Text Mining – Other Topics

- Sentiment Analysis
  - Automatically determine tone in text: positive, negative or neutral
  - Typically uses collections of good and bad words
  - *"While the traditional media is slowly starting to take John McCain's straight talking image with increasingly large grains of salt, his base isn't quite ready to give up on their favorite son. Jonathan Alter's bizarre defense of McCain after he was caught telling an outright lie, perfectly captures that reluctance[.]"*
  - Often fit using Naïve Bayes

- There are sentiment word lists out there:
  - See http://neuro.imm.dtu.dk/wiki/Text_sentiment_analysis