A Language-based Approach to Measuring Scholarly Impact

Sean Gerrish and David Blei
Princeton University

22 June 2010

Identifying influential documents

- In large collections of text documents, which ones are important?

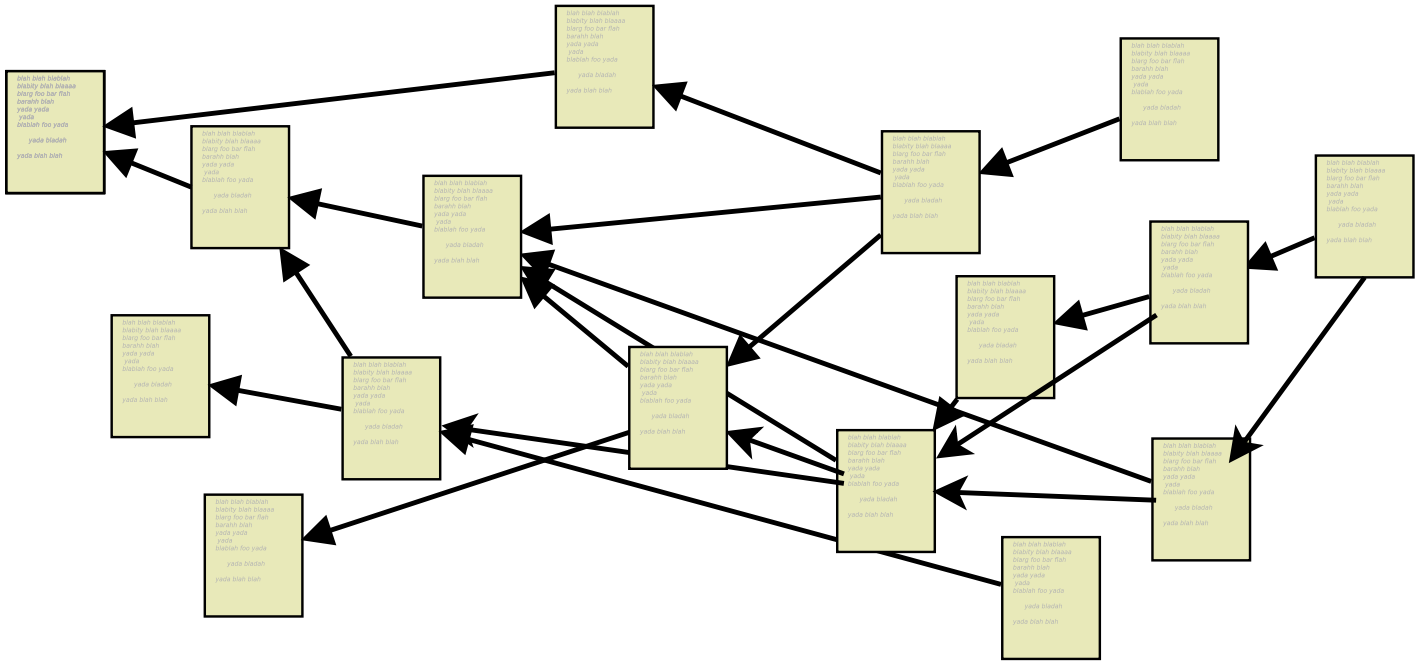
Image source: `http://jenslapinski.files.wordpress.com/2008/06/documents.jpg`

- We have developed a statistical model to measure the influence of text documents.
- Based only on the language of the documents, this measure is significantly correlated to citation counts.

Identifying influential documents

- Goal: Develop a measure of influence based only on language.
- There are many specific application areas.
  - News articles
  - Legal opinions
  - Scientific impact
- Entire fields of study rely on work like this.
  - History
  - Academic research
  - Much of Bibliometrics

Common approach: citations

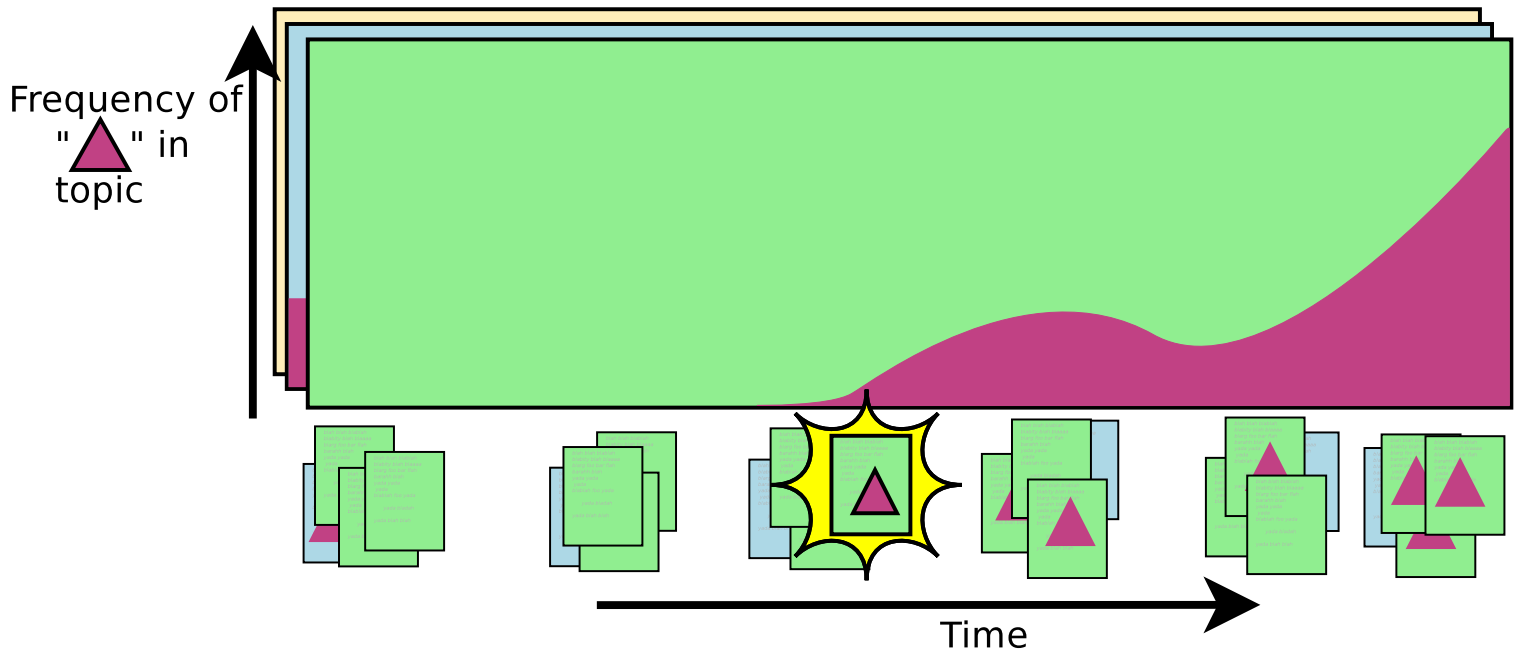


- Traditionally, citations are used to identify influential documents.
- E.g.: *A novel progressive spongiform encephalopathy in cattle*, [**?**] (Cited by 709)

Common approach: citations

- Citations are limited
  - May not exist
  - Hard to get
  - Describe only one kind of influence
- A language-based approach enables us to measure the influence of new kinds of documents
  - News articles
  - Historic manuscripts

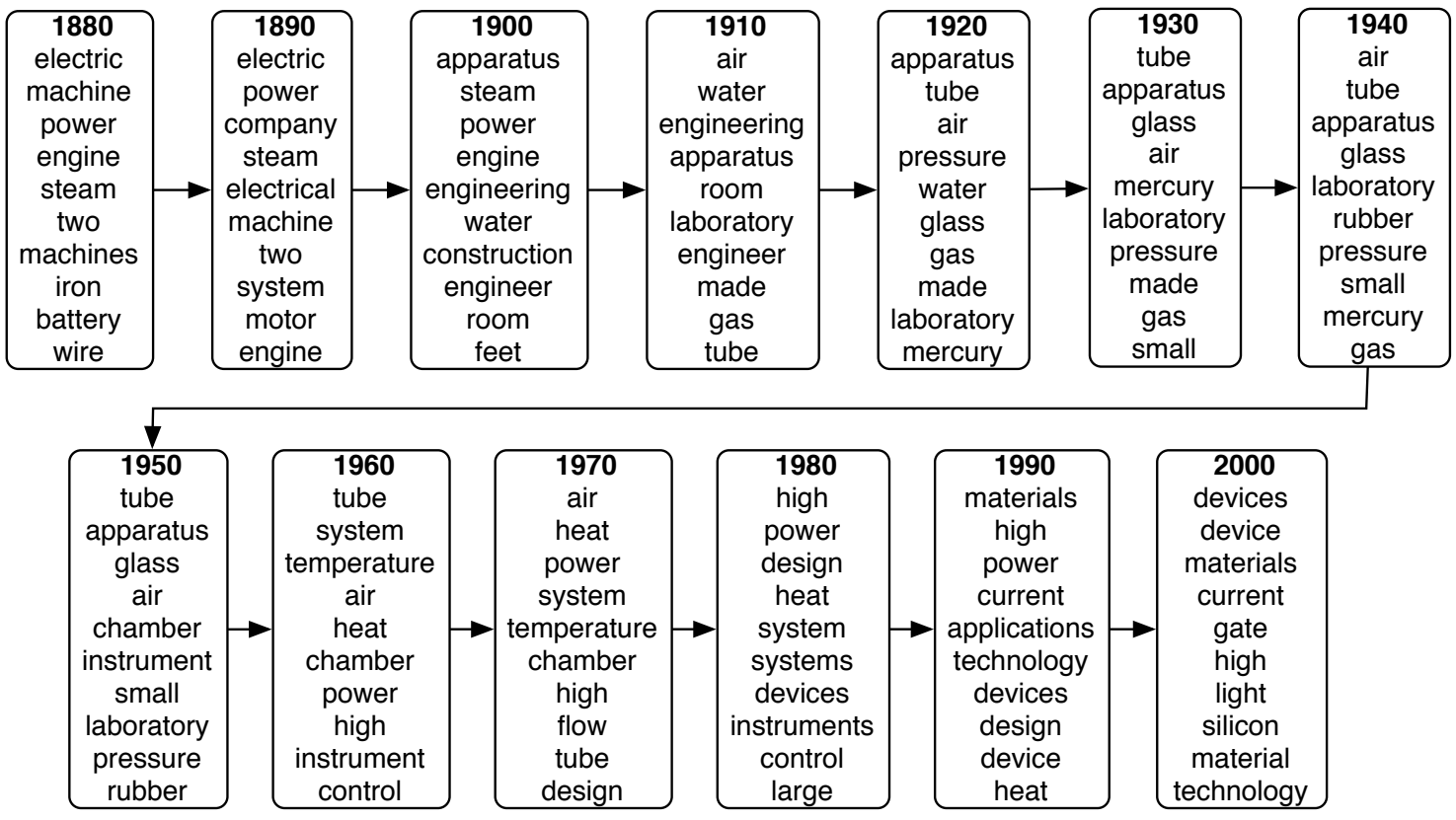
Methodology



- We develop a probabilistic model to captures the influence of documents.
- The intuition: influential documents use language that becomes more popular in later years.
- With posterior inference, we retrospectively see which documents have been influential on the corpus.

Changing topics

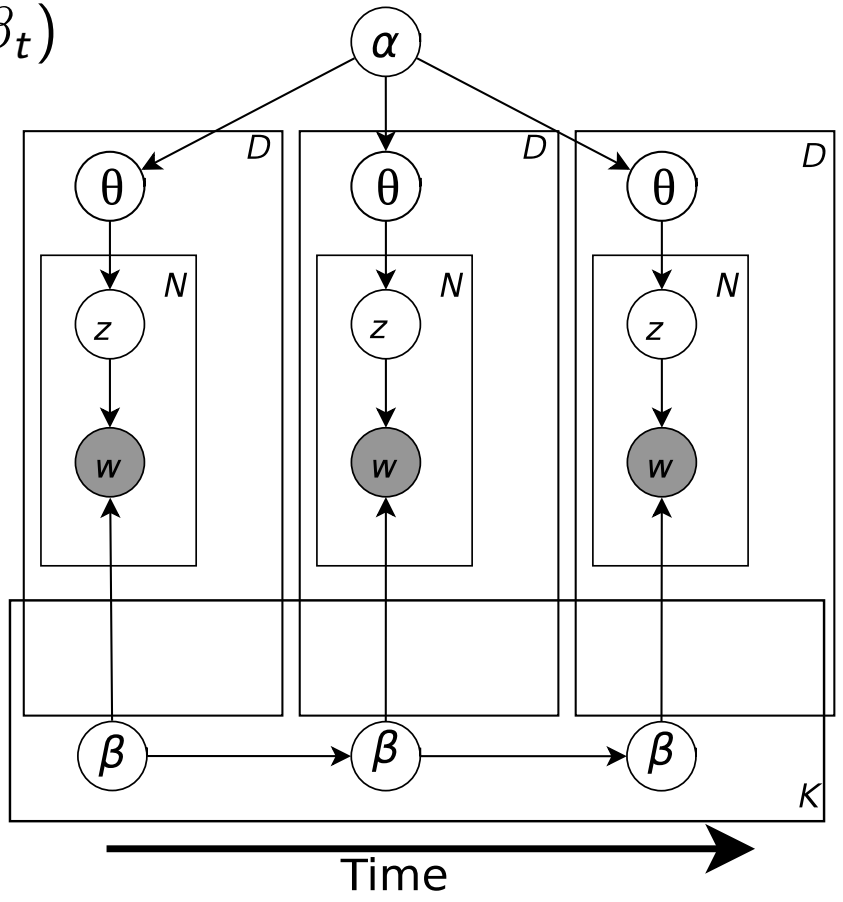
- Topic models decompose a corpus into a set of topics, i.e. distributions over terms.
- The Dynamic Topic Model allows topics to change over time [**?**].



The Dynamic Topic Model

- Topics drift in a Markov chain:
  $\beta_t \sim \mathcal{N}(\beta_{t-1}, \sigma^2)$

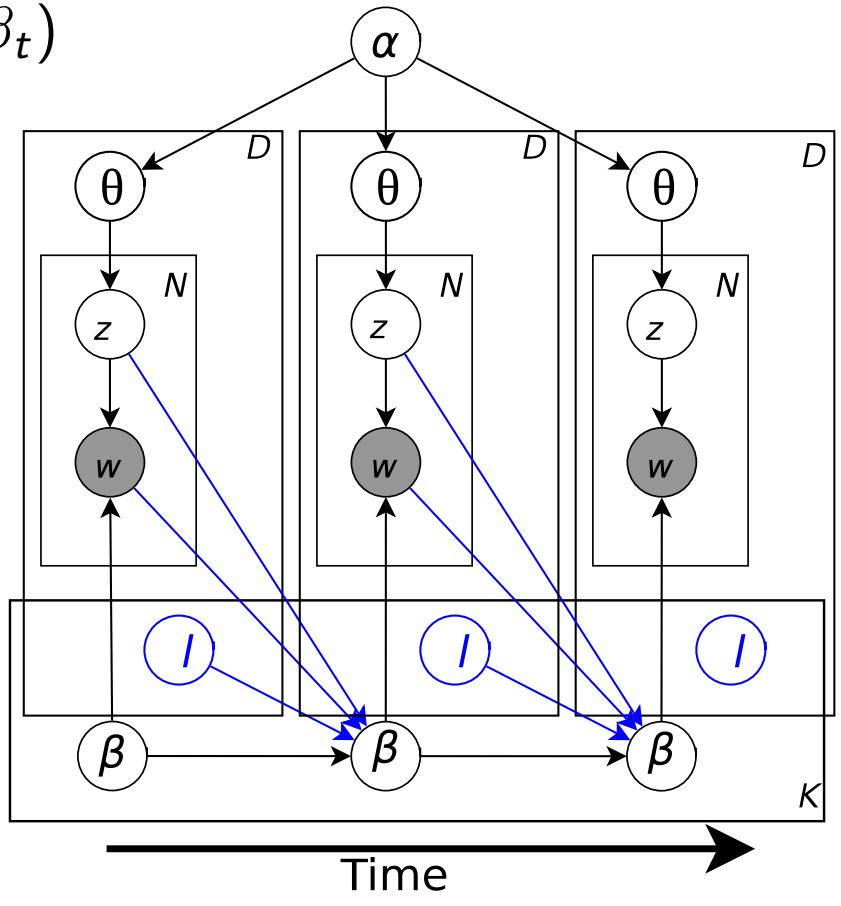
- Documents are generated from Latent Dirichlet Allocation
  $D_t \sim \text{LDA}(\alpha_t, \beta_t)$



The Document Influence Model

- Topics drift in a Markov chain:
  $\beta_t \sim \mathcal{N}(\beta_{t-1} + \text{Infl}(...), \sigma^2)$

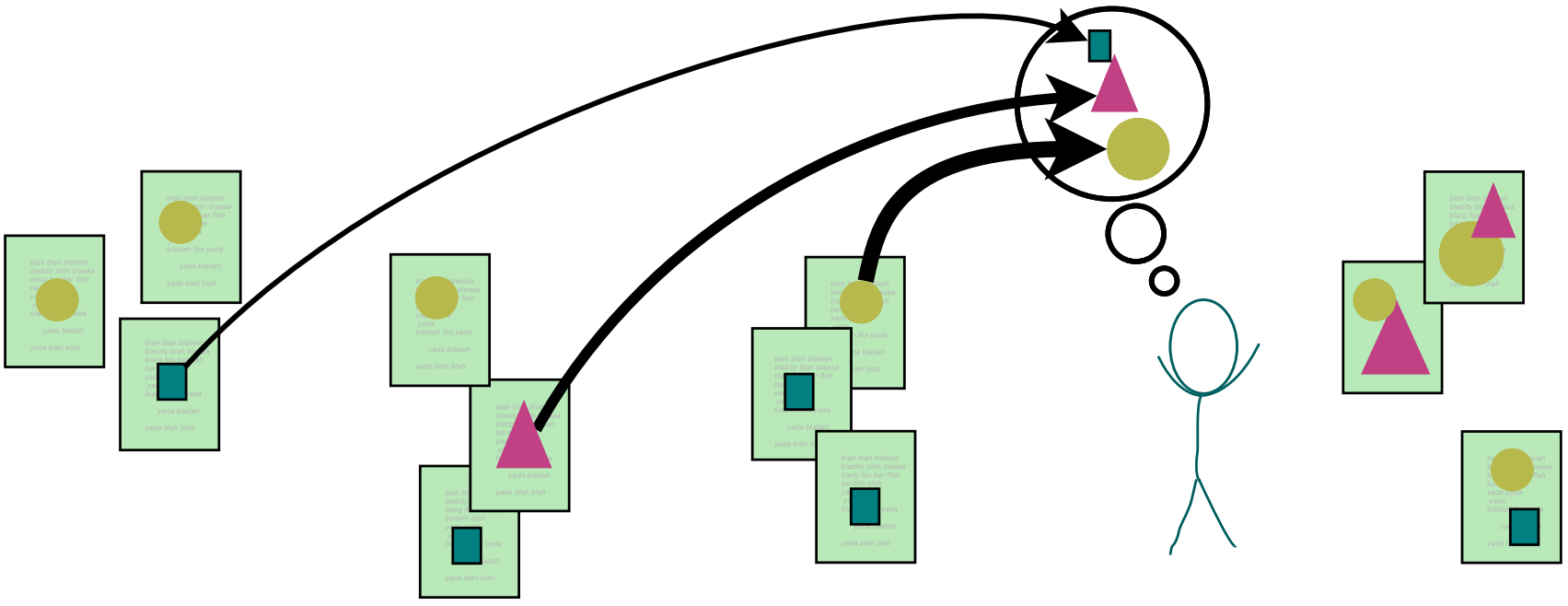
- Documents are generated from Latent Dirichlet Allocation
  $D_t \sim \text{LDA}(\alpha_t, \beta_t)$



The DIM influence function

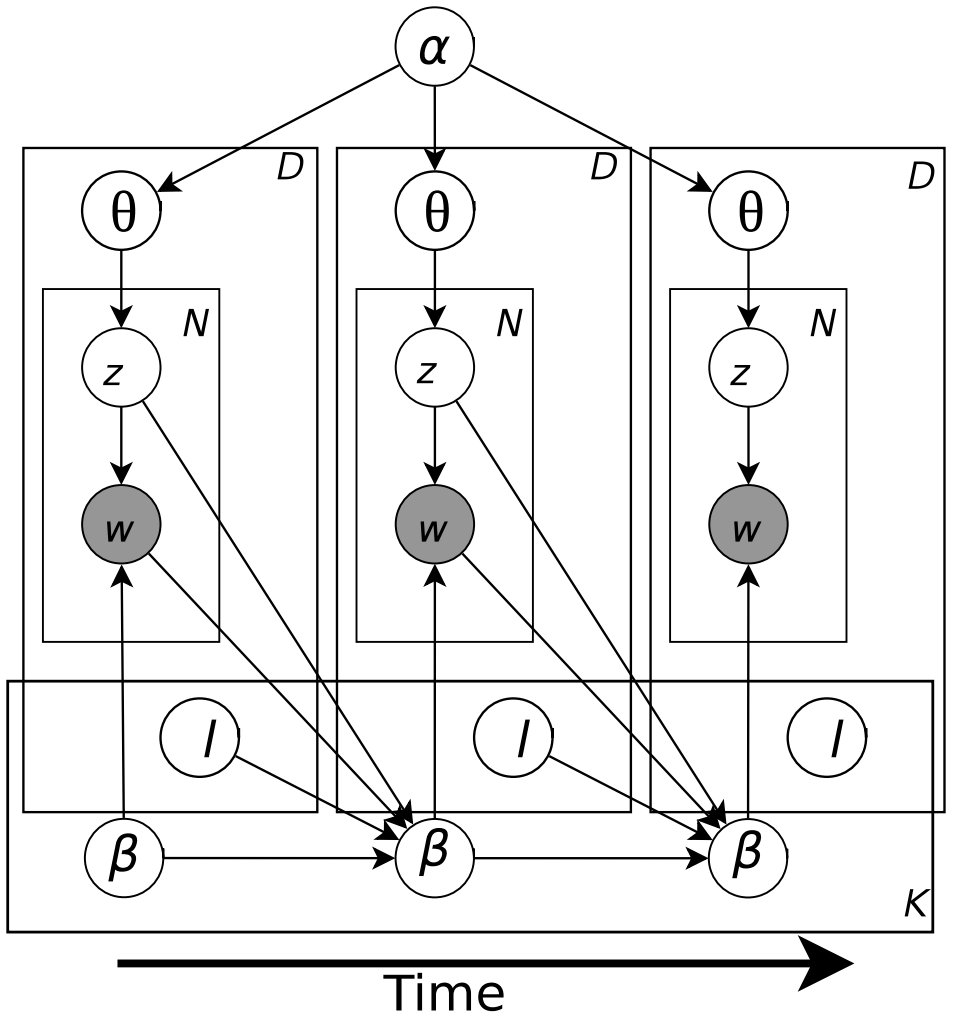
Markov step: $\beta_{t,k} \sim \mathcal{N}(\beta_{t-1,k} + \mathit{Infl}(t,k), \sigma^2 I)$,
We developed the model to have certain characteristics:

- More recent documents have greater influence.
- A document only influences relevant topics.



How do we find the model parameters?

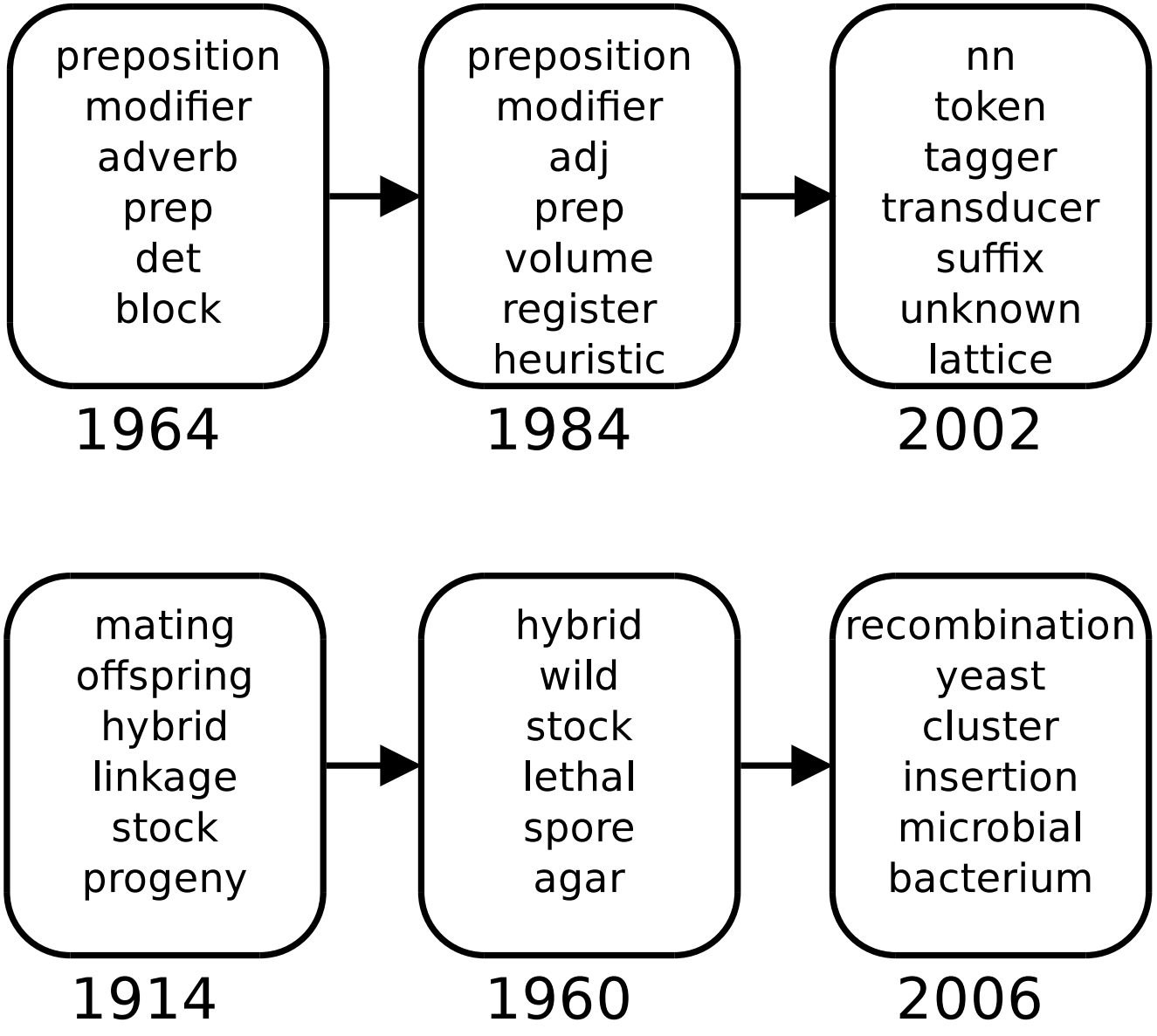
- We observe only the words $\mathbf{w}$.
- We want to find the posterior $p(I, \theta, z, \beta | \mathbf{w})$.
- We use variational methods [**?**].



Experiments

- We analyzed three corpora with the DIM.
  - The *ACL Anthology*: 7561 documents
  - *Nature*: 34454 documents
  - *PNAS*: 11855 documents
- This provides estimates of the influence of each article.
- We computed the Spearman rank correlation with citations.
  - ACL Anthology Network [**?**]
  - Google Scholar (PNAS and Nature)

Topics in ACL and PNAS



A closer look

- We can inspect documents with high or low scores.
  - High influence and high citations
  - Low influence and high citations
  - High influence and low citations

High influence and high citations

The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

We describe a series of five statistical models of the translation process and give algorithms for estimating the parameters of these models given a set of pairs of sentences that are translations of one another. We define a concept of word-by-word alignment between such pairs of sentences. For any given pair of such sentences each of our models assigns a probability to each of the possible word-by-word alignments. We give an algorithm for seeking the most probable of these alignments. Although the algorithm is suboptimal, the alignment thus obtained accounts well for the word-by-word relationships in the pair of sentences. We have a great deal of data in French and English from the proceedings of the Canadian Parliament. Accordingly, we have restricted our work to these two languages; but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages. We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus.

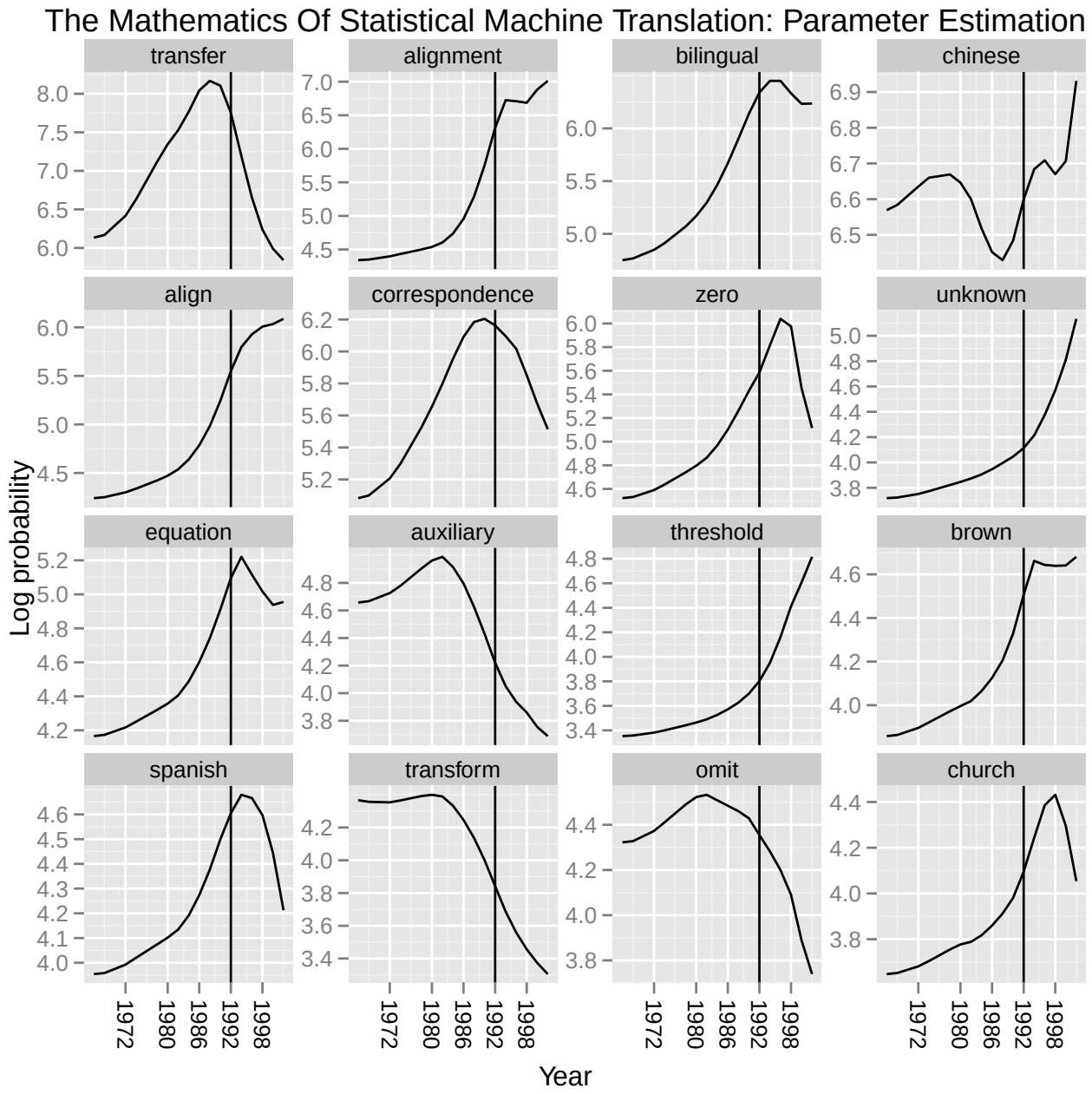
1. Introduction

The growing availability of bilingual, machine-readable texts has stimulated interest in methods for extracting linguistically valuable information from such texts. For example, a number of recent papers deal with the problem of automatically obtaining pairs of aligned sentences from parallel corpora (Warwick and Russell 1990; Brown, Lai, and Mercer 1991; Gale and Church 1991b; Kay 1991). Brown et al. (1990) assert, and Brown, Lai, and Mercer (1991) and Gale and Church (1991b) both show, that it is possible to obtain such aligned pairs of sentences without inspecting the words that the sentences contain. Brown, Lai, and Mercer base their algorithm on the number of words that the sentences contain, while Gale and Church base a similar algorithm on the number of characters that the sentences contain. The lesson to be learned from these two efforts is that simple, statistical methods can be surprisingly successful in achieving linguistically interesting goals. Here, we address a natural extension of that work: matching up the words within pairs of aligned sentences.

In recent papers, Brown et al. (1988, 1990) propose a statistical approach to machine translation from French to English. In the latter of these papers, they sketch an algorithm for estimating the probability that an English word will be translated into any particular French word and show that such probabilities, once estimated, can be used together with a statistical model of the translation process to align the words in an English sentence with the words in its French translation (see their Figure 3).

* IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

ACL citations: 7561



Low influence and high citations

Building a Large Annotated Corpus of English: The Penn Treebank

Mitchell P. Marcus*
University of Pennsylvania

Beatrice Santorini†
Northwestern University

Mary Ann Marcinkiewicz‡
University of Pennsylvania

1. Introduction

There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora. Such corpora are beginning to serve as important research tools for investigators in natural language processing, speech recognition, and integrated spoken language systems, as well as in theoretical linguistics. Annotated corpora promise to be valuable for enterprises as diverse as the automatic construction of statistical models for the grammar of the written and the colloquial spoken language, the development of explicit formal theories of the differing grammars of writing and speech, the investigation of prosodic phenomena in speech, and the evaluation and comparison of the adequacy of parsing models.

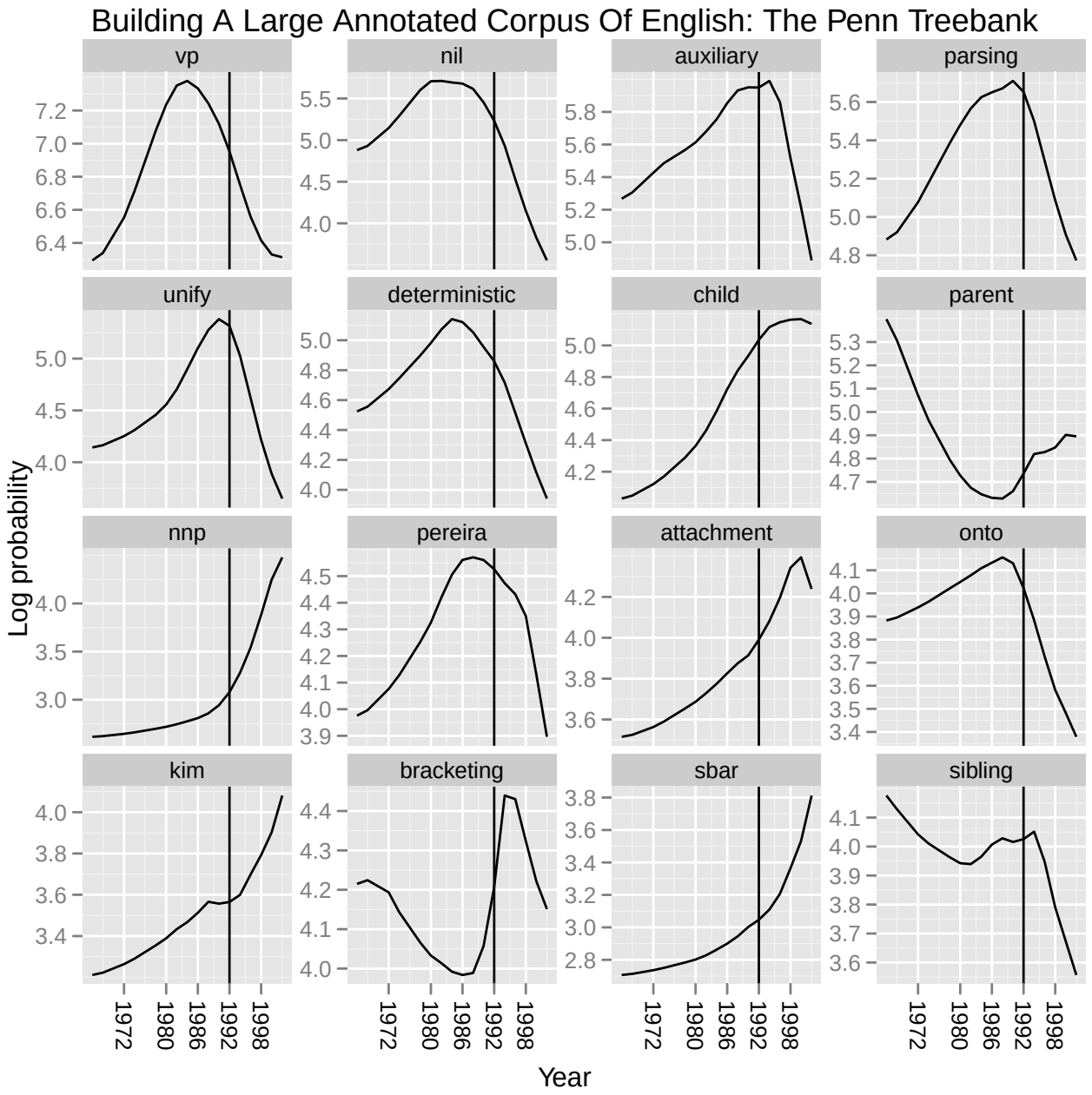
In this paper, we review our experience with constructing one such large annotated corpus—the Penn Treebank, a corpus[1] consisting of over 4.5 million words of American English. During the first three-year phase of the Penn Treebank Project (1989–1992), this corpus has been annotated for part-of-speech (POS) information. In addition, over half of it has been annotated for skeletal syntactic structure. These materials are available to members of the Linguistic Data Consortium; for details, see Section 5.1.

The paper is organized as follows. Section 2 discusses the POS tagging task. After outlining the considerations that informed the design of our POS tagset and presenting the tagset itself, we describe our two-stage tagging process, in which text is first assigned POS tags automatically and then corrected by human annotators. Section 3 briefly presents the results of a comparison between entirely manual and semi-automated tagging, with the latter being shown to be superior on three counts: speed, consistency, and accuracy. In Section 4, we turn to the bracketing task. Just as with the tagging task, we have partially automated the bracketing task: the output of

* Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA 19104.
† Department of Linguistics, Northwestern University, Evanston, IL 60208.
‡ Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA 19104.

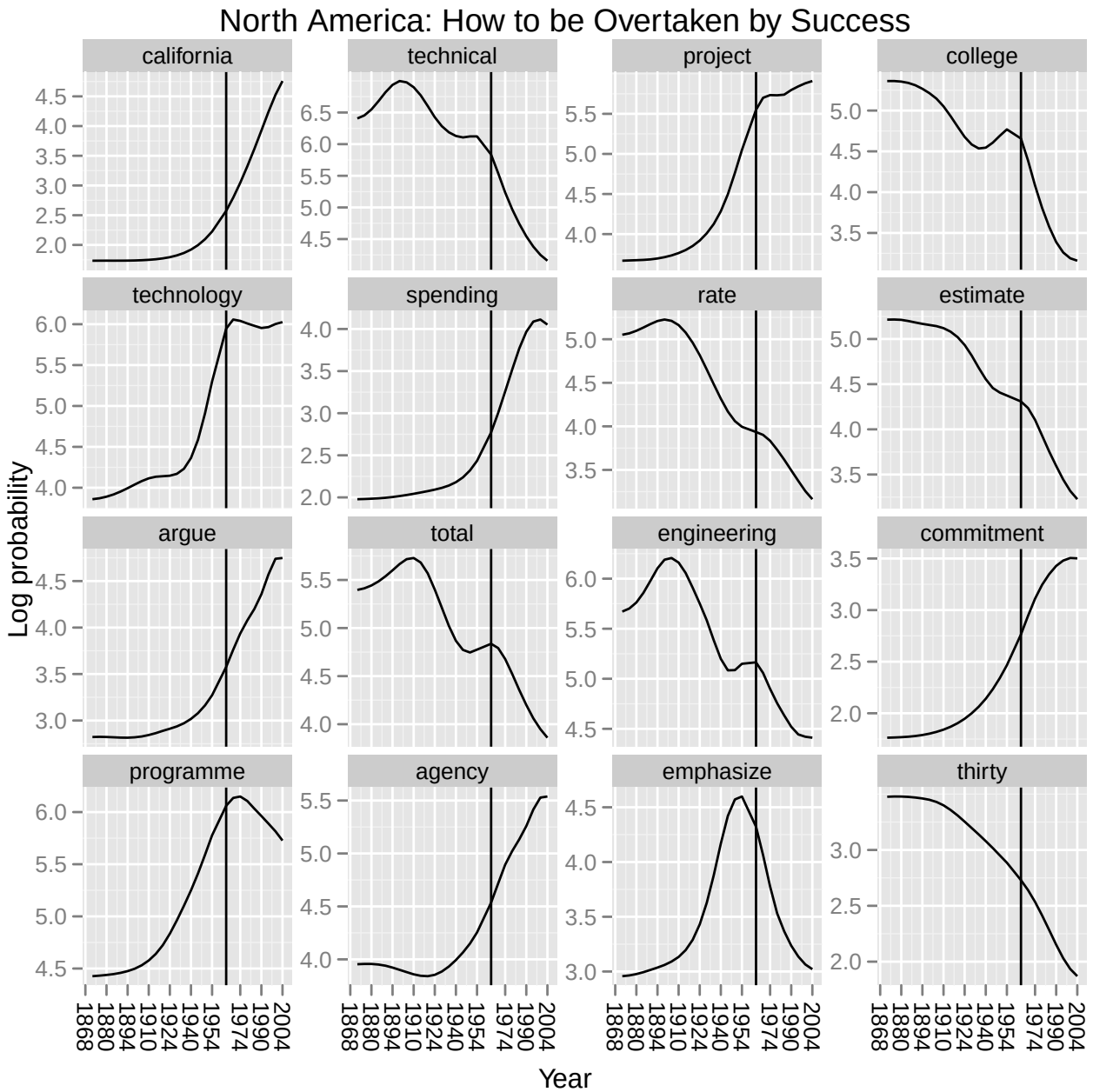
1 A distinction is sometimes made between a *corpus* as a carefully structured set of materials gathered together to jointly meet some design principles, and a *collection*, which may be much more opportunistic in construction. We acknowledge that from this point of view, the raw materials of the Penn Treebank form a collection.

ACL citations: 2180



High influence and low citations

Citations: NA

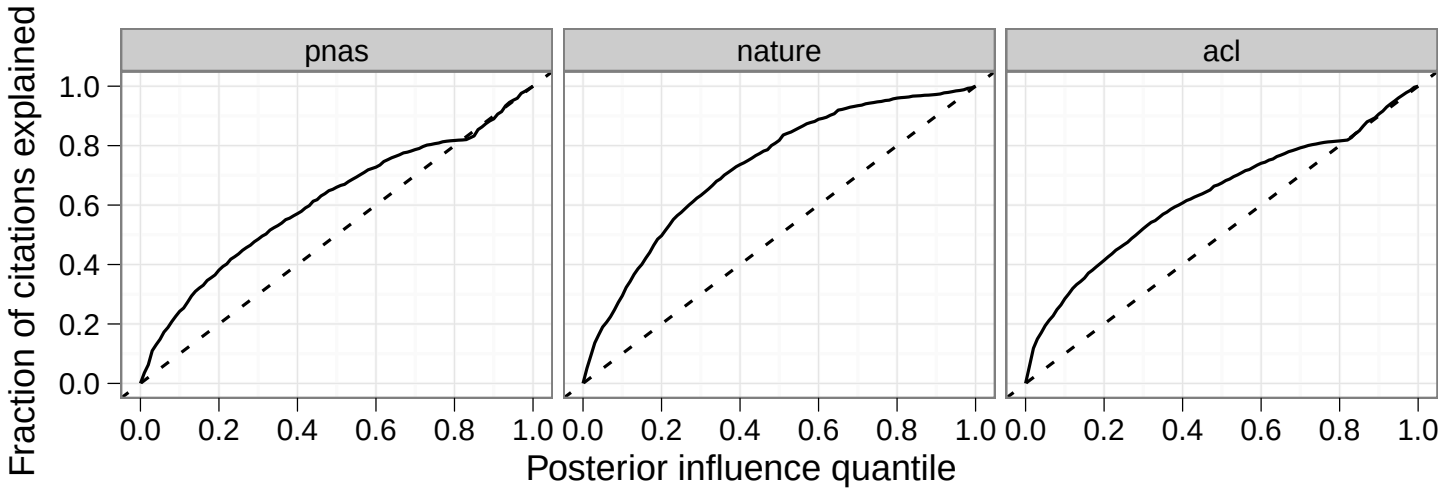


Results

- Significant correlation with citations ($p \leq 1e-4$)
- Window sizes are 4 years for *ACL* and *PNAS* and 5 years for *Nature*
- Spearman rank correlation 0.37 (ACL); 0.28 (Nature); 0.20 (PNAS)
- Also developed a simple baseline which does well: 0.2 (ACL, PNAS) and 0.26 (Nature)

Results

- We can sort articles by decreasing influence and inspect how many citations the top N have.
- E.g., the top 20% of *Nature* articles receive 50% of *Nature*'s citations.



Summary

- We developed an algorithm that identifies influential documents by analyzing their texts
- The Document Influence Model
  - Divides the words of a corpus into topics
  - Infers which articles "influenced" their topics
- Based only on analyzing the language, our model finds a measure of influence that correlates significantly with citation.
- Some future directions
  - Per-document influence envelope
  - Validation on non-citation data (e.g., usage data)
  - Application to corpora without citations

Contact

- Sean Gerrish (sgerrish@cs.princeton.edu)
  `http://www.cs.princeton.edu/~sgerrish`
- David Blei (blei@cs.princeton.edu)
  `http://www.cs.princeton.edu/~blei`

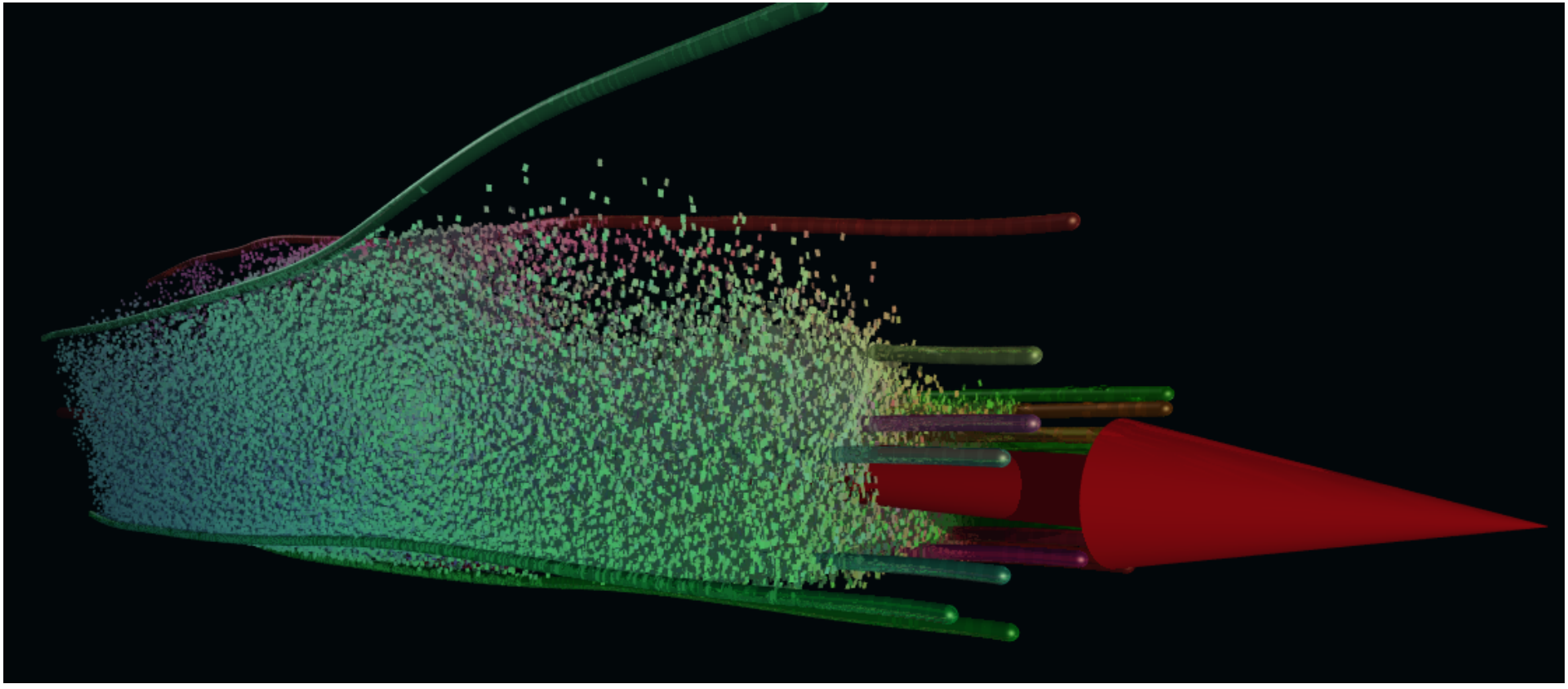


Figure: *Nature* articles and their dynamic topics

Bibliography I

Baseline

One possible heuristic is simple:

- Define a word's weight at time t as:

$$w_t := \frac{\text{Frequency of } w \text{ in } [t, t+f]}{\text{Frequency of } w \text{ in [t-b, t]}}$$

- Document **D**'s score is the weighted average of these:

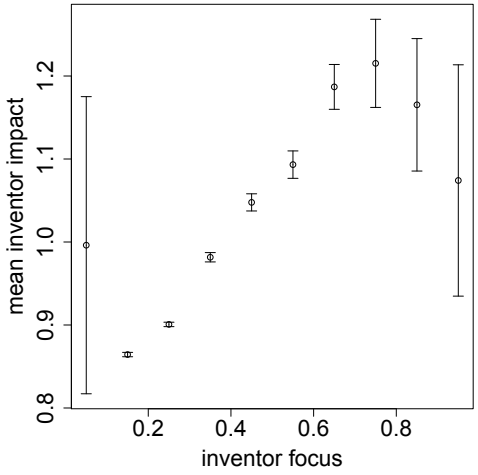
$$\mathcal{I}(\mathbf{D}) := \frac{\sum_{w \in \mathbf{D}} \text{Count}(w) w_t}{\sum_{w \in \mathbf{D}} \text{Count}(w)}$$

Heuristic solution

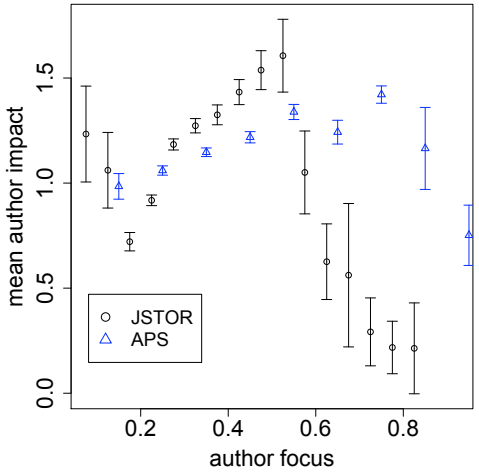
- Fast
- Easy to implement
- Not "optimal" in any obvious sense
- Does not incorporate information about documents' semantics
- Not focused at the level of research contributions
  - Too large a hammer
  - E.g.: cows and health policy around 1986

Focus and knowledge contribution

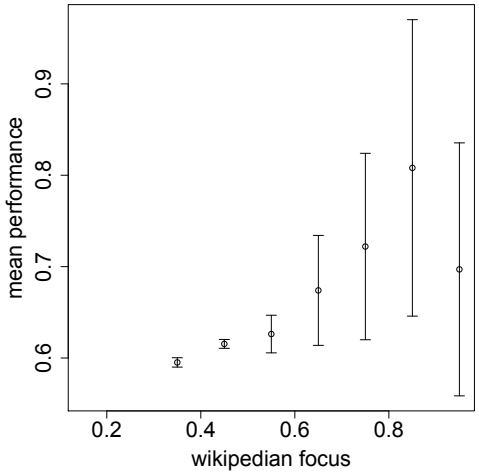
The quality of researchers' contributions increases with focus.



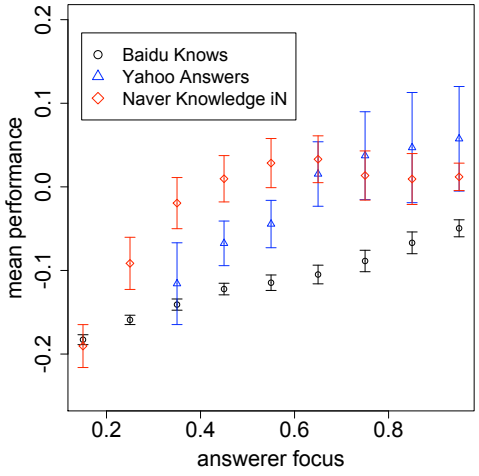
Patents



Research articles



Wikipedia



Q&A forums

Motivation for $\exp(-\beta)$ coefficient in $\mathit{Infl}(t, l, z, w)$

$$
\begin{aligned}
& \exp(\beta_t) = \exp(\beta_{t-1}) + \mathit{Infl}_t \\
\iff \quad & 1 = \exp(\beta_{t-1} - \beta_t) + \exp(-\beta_t)\mathit{Infl}_t \\
\iff \quad & 1 - \exp(-\beta_t)\mathit{Infl}_t = \exp(\beta_{t-1} - \beta_t) \\
\iff \quad & \log(1 - \exp(-\beta_t)\mathit{Infl}_t) = \beta_{t-1} - \beta_t \\
\iff \quad & \beta_t = \beta_{t-1} - \log(1 - \exp(-\beta_t)\mathit{Infl}_t)
\end{aligned}
\tag{1}
$$

Note that when $\exp(-\beta_t)\mathit{Infl}_t$ is small, we have $\beta_t \approx \beta_{t-1} + \exp(-\beta_t)\mathit{Infl}_t$.

Regularized linear regression for $\tilde{l}$ updates

$$g(s,q) := \Lambda_{\exp(-\tilde{m}_{q,k}+\tilde{V}_{q,k}/2)}(\mathbf{W}_{s,k} \circ \phi_{s,k}) \tag{2}$$

$$h(s,q) := ((\mathbf{W}_{s,k} \circ \phi_{s,k})^T \Lambda_{\exp(-2\tilde{m}_q+2\tilde{V}_q)+\exp(-2\tilde{m}_q+\tilde{V}_q)}(\mathbf{W}_{s,k} \circ \phi_{s,k}) \tag{3}$$

$$+\Lambda_{(\mathbf{W}_{s,k}\circ\mathbf{W}_{s,k}\circ(\phi_{s,k}-\phi_{s,k}\circ\phi_{s,k}))^T(\exp(-2\tilde{m}_q+2\tilde{V}_q)+\exp(-2\tilde{m}_q+\tilde{V}_q))}) \tag{4}$$

$$\tilde{l}_{t,k} \leftarrow \left(\frac{\sigma^2}{\sigma_d^2} I + \left(\sum_{i=t}^{T-1} r(i-t)^2 h(t,i)\right)\right)^{-1}$$
$$\left(\sum_{i=t}^{T-1} r(i-t) g(t,i)^T (\tilde{m}_{i+1,k} - \tilde{m}_{i,k} + \tilde{V}_{i,k} - \sum_{j=0\ldots i, j\neq t} r(i-j) g(j,i) \tilde{l}_{j,k})\right) \tag{5}$$

Dimensionality reduction

Bag-of-words model

- Only worry about the word counts in each document
- So a document is basically a sparse list of word counts:

$$\text{“the cat in the hat”} \rightarrow (0, 0, 1, 0, 2, \ldots, 0)$$

Doing anything with these huge lists is hard

- Popular statistics tools like Principal Component Analysis help us to reduce this to a smaller number
- Topic models accomplish a similar thing:

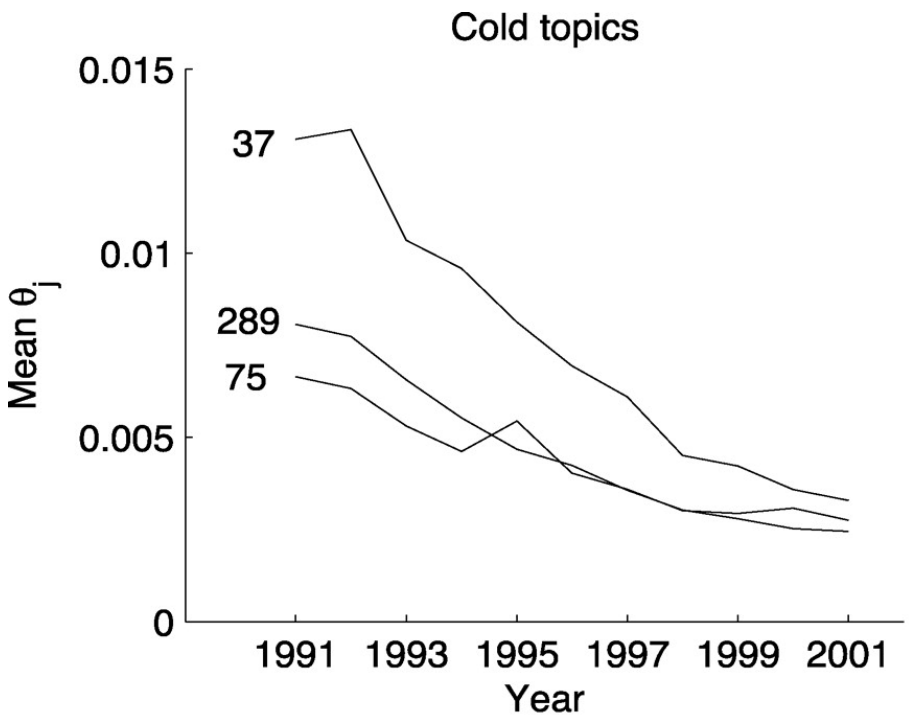
$$10^4 \text{ words} \rightarrow 50 \text{ topics}$$

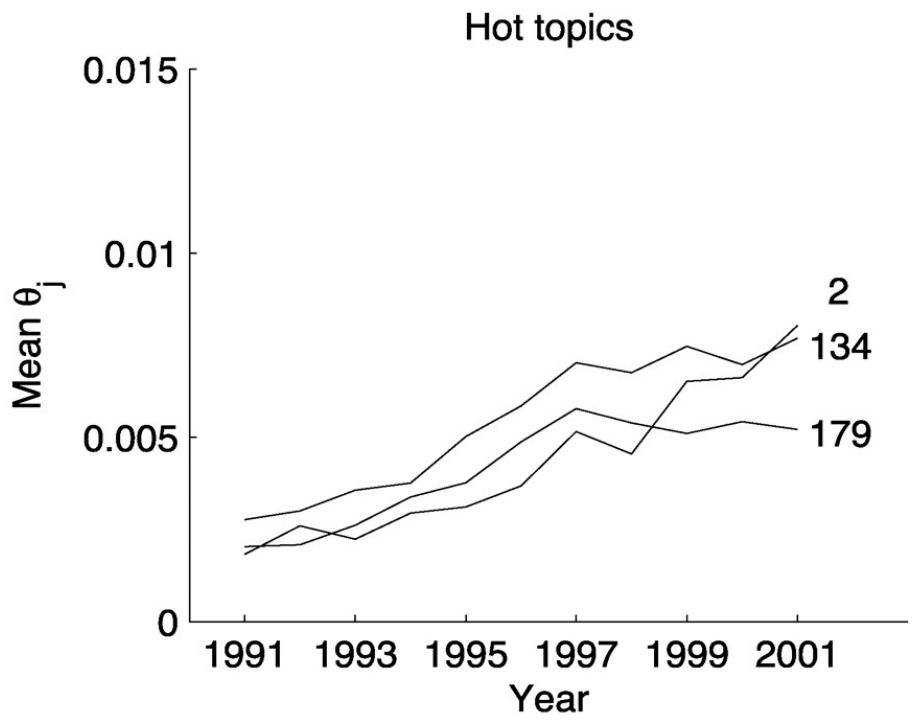
The DIM generative model

For time $t = 1, \ldots, T$:

- For topic $k = 1, \ldots, K$:
  Draw natural parameters
  $\beta_{t,k} | \beta_{t-1,k}, \mathbf{z}_{s<t}, l_{s<t} \sim \mathcal{N}(\beta_{t-1,k} + \mathit{Infl}(t,k), \sigma^2 I)$
- For each document d_t:
  - Generate document d_t using traditional LDA with parameters α_t and β_t.
  - For topic $k = 1, \ldots, K$, draw document weight $l_{d,k} \sim \mathcal{N}(\mathbf{0}, \sigma_d^2 I)$;

Topic models - applications





37	289	75
CDNA	KDA	ANTIBODY
AMINO	PROTEIN	ANTIBODIES
SEQUENCE	PURIFIED	MONOCLONAL
ACID	MOLECULAR	ANTIGEN
PROTEIN	MASS	IGG
ISOLATED	CHROMATOGRAPHY	MAB
ENCODING	POLYPEPTIDE	SPECIFIC
CLONED	GEL	EPITOPE
ACIDS	SDS	HUMAN
IDENTITY	BAND	MABS
CLONE	APPARENT	RECOGNIZED
EXPRESSED	LABELED	SERA

2	134	179
SPECIES	MICE	APOPTOSIS
GLOBAL	DEFICIENT	DEATH
CLIMATE	NORMAL	CELL
CO2	GENE	INDUCED
WATER	NULL	BCL
ENVIRONMENTAL	MOUSE	CELLS
YEARS	TYPE	APOPTOTIC
MARINE	HOMOZYGOUS	CASPASE
CARBON	ROLE	FAS
DIVERSITY	KNOCKOUT	SURVIVAL
OCEAN	DEVELOPMENT	PROGRAMMED
EXTINCTION	GENERATED	MEDIATED

Image source: [?]; Available 2010-01-15 http://www.pnas.org/content/101/suppl.1/5228/F5.large.jpg

Sterling similarity

We aimed to use a metric that captures three qualities: *variety*, or how many different areas an individual contributes to; *balance*, or how evenly their efforts are distributed among these areas; and *similarity*, or how related those areas are. We use the Stirling measure $\mathcal{F}$, which captures all three aspects:

$$\mathcal{F} := \sum_{i,j} s_{ij} p_i p_j,$$

where p_i is the proportion of the individual's contributions in category i and $s_{ij} = n_{ij}/n_j$ is a measure of similarity between categories i and j, inferred from the number of joint contributors n_{ij} between two categories i and j.

Exploration

Understand the themes in a collection of documents and how they relate to one another