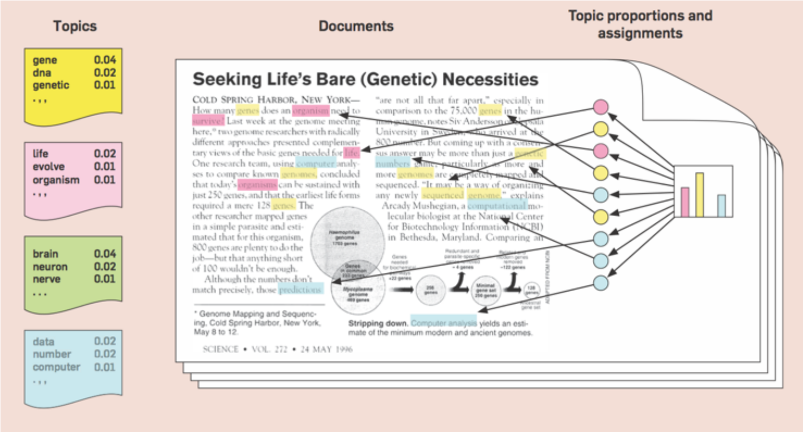

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Topic Modeling in Quant and Qual Research:

A Hands-On Approach

AoM Session 253
Saturday, 8-10 a.m.
Atlanta Marriott Marquis
Lobby L508

Co-Leads and Agenda

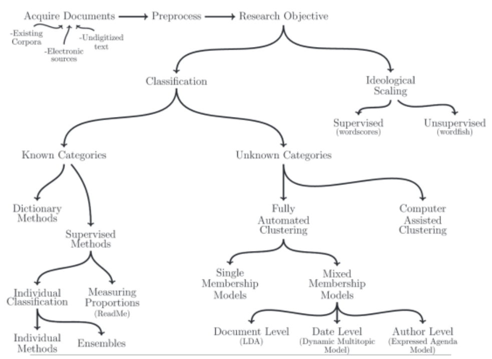


Fig. 1 An overview of text as data methods.

- Co-Organizers*
  - Dev Jennings
  - Sarah Kaplan

- Co-Leads*
  - Tim Hannigan
  - Richard Haan
  - Milo Wang
  - Hovig Tchalian
  - Keyvan Vakili

- Welcome

- TModel Opportunities & Cautions I

- Situating & Demo-ing Tmodels

- Beyond LDA

- TModels Opportunities & Cautions II

- Q&A

* In order of presentation

TOPIC MODELING: OPPORTUNITIES AND CAUTIONS

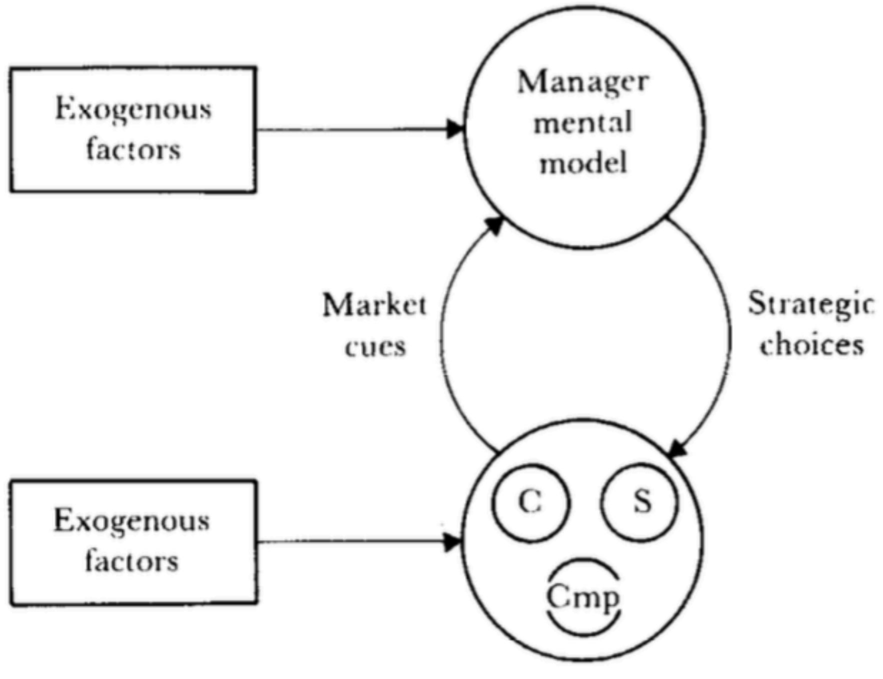
Sarah Kaplan
University of Toronto, Rotman School

TEXT ANALYSIS IN ORGANIZATION STUDIES

- Many organizational phenomena play out or are represented in written and spoken word.
  - Annual reports
  - Patents
  - Scientific publications
  - Press releases
  - Policy documents
  - Newspaper articles
  - Etc.
- Automated text analysis allows us to move from words to numbers
  - Capture ideas across large numbers of texts
  - Generate data that can be used in quantitative descriptive analysis or regression analysis

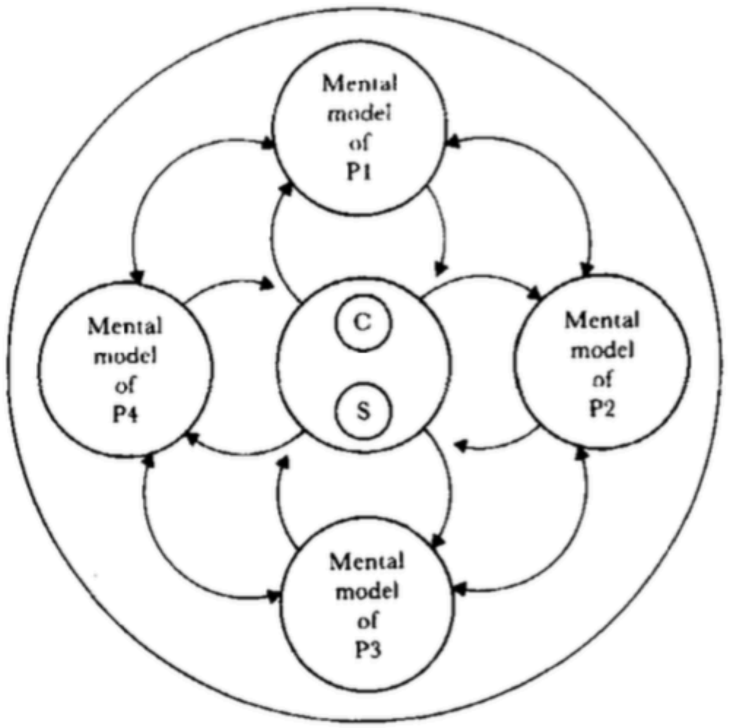
EXAMPLE: COGNITION IN FIRM ENVIRONMENTS

- A conversation started by scholars in the 1980's… most famously Porac, Thomas and Baden-Fuller (1989) (for a review of progress since then see Kaplan 2011 in *JMS*)
  - "structure of that industry both determines and is determined by managerial perceptions of the environment"



Note: C	=	Customers
S	=	Suppliers
Cmp	=	Competitors

Figure 1. Reciprocal influence of technical and cognitive levels of analysis

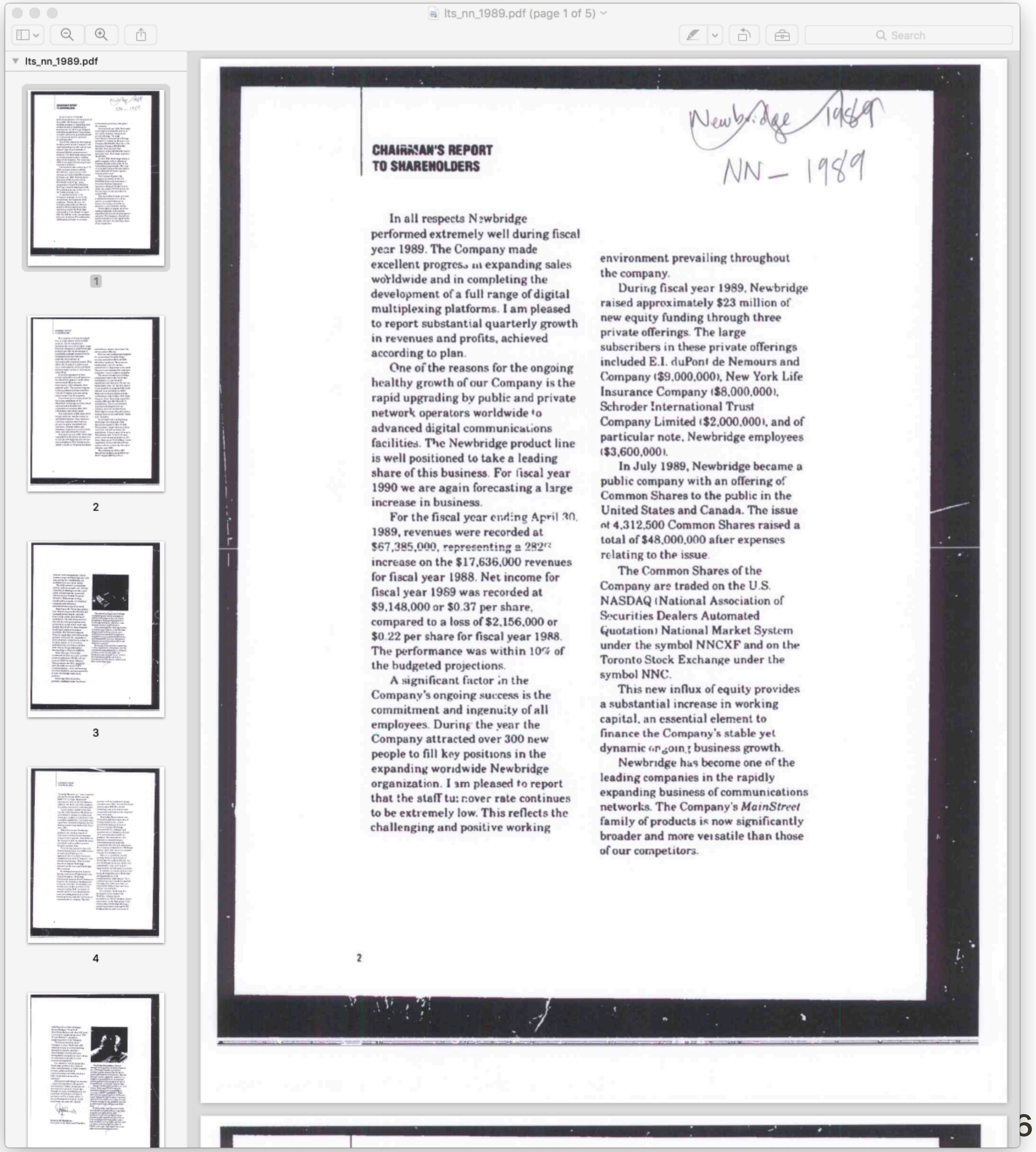


Note: C	=	Customers
S	=	Suppliers
Pn	=	Producers

Figure 2. Mutual enactment processes within an industrial sector

TEXT ANALYSIS IN ORGANIZATION STUDIES

- Historically, huge costs of analyzing collections of texts have blocked progress.
- Example: my own dissertation on firm response to technical change in the communications industry…
  - Hand coded off of print outs of microfiche
  - Or painstaking corrections of OCR from poor microfiche copies
- See Kaplan 2008 or Eggers & Kaplan 2009



Newbridge 1989
NN – 1989

CHAIRMAN'S REPORT
TO SHAREHOLDERS

In all respects Newbridge performed extremely well during fiscal year 1989. The Company made excellent progress in expanding sales worldwide and in completing the development of a full range of digital multiplexing platforms. I am pleased to report substantial quarterly growth in revenues and profits, achieved according to plan.

One of the reasons for the ongoing healthy growth of our Company is the rapid upgrading by public and private network operators worldwide to advanced digital communications facilities. The Newbridge product line is well positioned to take a leading share of this business. For fiscal year 1990 we are again forecasting a large increase in business.

For the fiscal year ending April 30, 1989, revenues were recorded at $67,385,000, representing a 282% increase on the $17,636,000 revenues for fiscal year 1988. Net income for fiscal year 1989 was recorded at $9,148,000 or $0.37 per share, compared to a loss of $2,156,000 or $0.22 per share for fiscal year 1988. The performance was within 10% of the budgeted projections.

A significant factor in the Company's ongoing success is the commitment and ingenuity of all employees. During the year the Company attracted over 300 new people to fill key positions in the expanding worldwide Newbridge organization. I am pleased to report that the staff turnover rate continues to be extremely low. This reflects the challenging and positive working environment prevailing throughout the company.

During fiscal year 1989, Newbridge raised approximately $23 million of new equity funding through three private offerings. The large subscribers in these private offerings included E.I. duPont de Nemours and Company ($9,000,000), New York Life Insurance Company ($8,000,000), Schroder International Trust Company Limited ($2,000,000), and of particular note, Newbridge employees ($3,600,000).

In July 1989, Newbridge became a public company with an offering of Common Shares to the public in the United States and Canada. The issue of 4,312,500 Common Shares raised a total of $48,000,000 after expenses relating to the issue.

The Common Shares of the Company are traded on the U.S. NASDAQ (National Association of Securities Dealers Automated Quotation) National Market System under the symbol NNCXF and on the Toronto Stock Exchange under the symbol NNC.

This new influx of equity provides a substantial increase in working capital, an essential element to finance the Company's stable yet dynamic ongoing business growth.

Newbridge has become one of the leading companies in the rapidly expanding business of communications networks. The Company's *MainStreet* family of products is now significantly broader and more versatile than those of our competitors.

2

PROMISE OF AUTOMATED TEXT ANALYSIS

- Promise of automated text analysis (including topic modeling):
  - Cost/time reduction.
- AND, reduction (in some ways) of human intervention
  - Do not need to specify topics/themes/count words in advance

TOPIC MODELING—METHOD "DU JOUR"

- For computer science: developed to improve search
- Use in social sciences, in last 6-7 years
- Key features:
  - "Bag of words" – no syntax (where syntax matters, there are better methods)
    - Best for identifying themes where categories are unknown
  - "Unsupervised" text analysis
    - But, sensitive to inputs to the algorithm
    - Often requires more "supervised" approaches to create semantically meaningful results
    - "Best fit" for computer scientists very different from "best fit" for social scientists

RECENT APPLICATION AREAS FOR TOPIC MODELING

- Using texts to analyze field-level logics (e.g., Jha & Beckman 2017)
  - Policy documents such as federal regulations, dept of education strategic plans, etc.
- Using texts to measure business unit attention to technological issues (e.g., Wilson & Joseph 2015)
  - “Background” section of patents
- Using texts to measure knowledge domains in patents (e.g., Kaplan & Vakili 2015)
  - Patent abstracts
- Using texts to identify policy framing (e.g., how government assistance to the arts has been framed, DiMaggio et al 2013)
  - Newspaper articles

CAUTIONS

- Topic modeling becoming “black boxed” in social science

- 4 Principles: from Grimmer and Stewart (2013)

1. All quantitative models of language are wrong but some are useful
2. Quantitative methods augment humans, but do not replace them
3. There is no one “best” method for automated text analysis
4. Validate, validate, validate

CAUTIONS

- Different approaches depending on research question and available documents

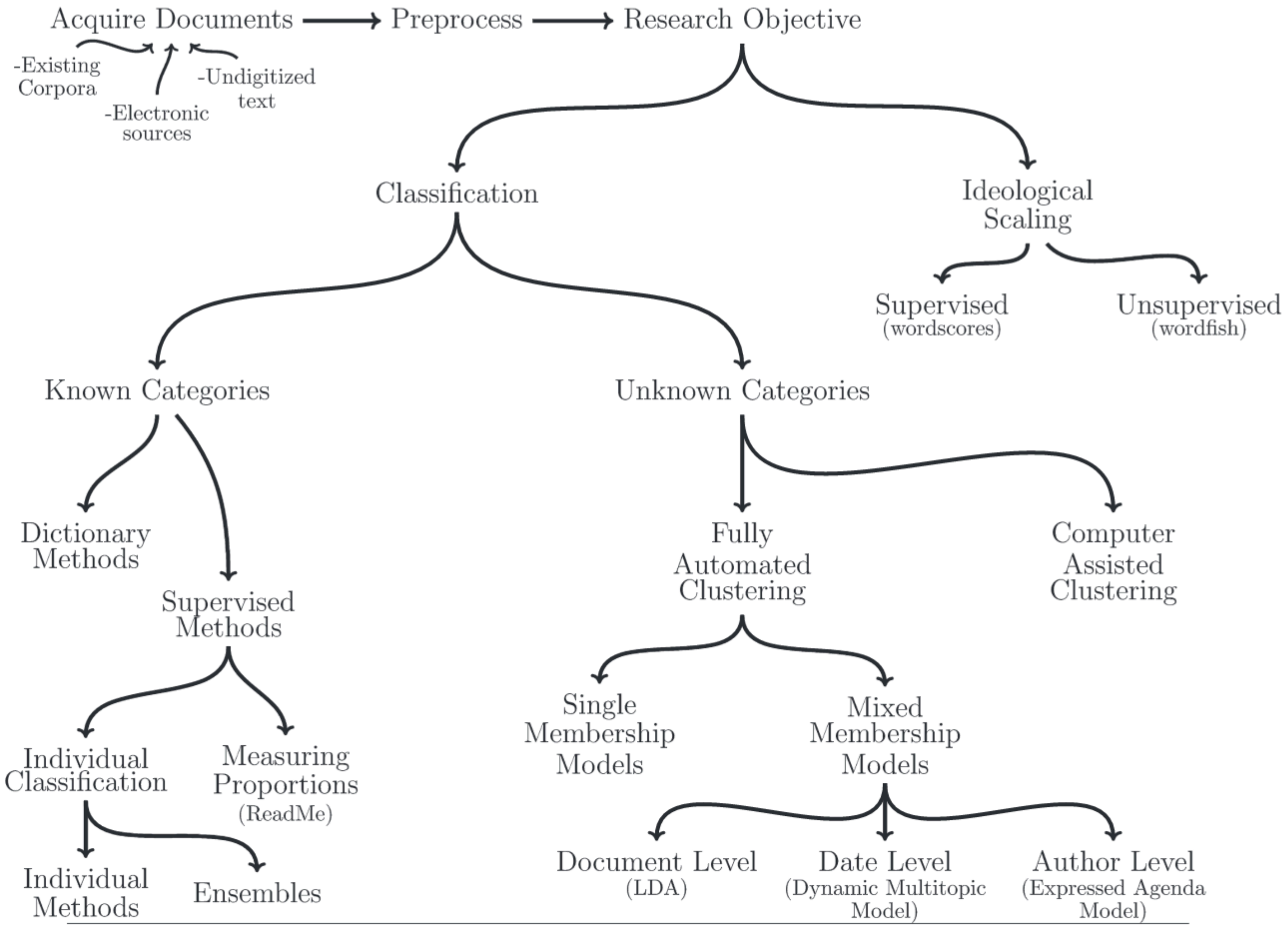


Fig. 1 An overview of text as data methods.

Source: from Grimmer and Stewart (2013) "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*

CAUTIONS

- Not all texts are amenable to automated text analysis. Works better when:
  - Focused text on a specific domain
  - Longer texts
    - Shorter texts such as tweets or open ended survey responses don’t provide enough information
- Automation does not replace deep understanding of the texts
  - Topic modeling still requires context-specific validation (see our approach in Kaplan & Vakili 2015)
    - Hand coding by researcher
    - Interviews
    - Expert coding and validation

Situating and demonstrating applications of Topic Modeling practice

AOM 2017 PDW – Topic Modeling
Aug 5, 2017

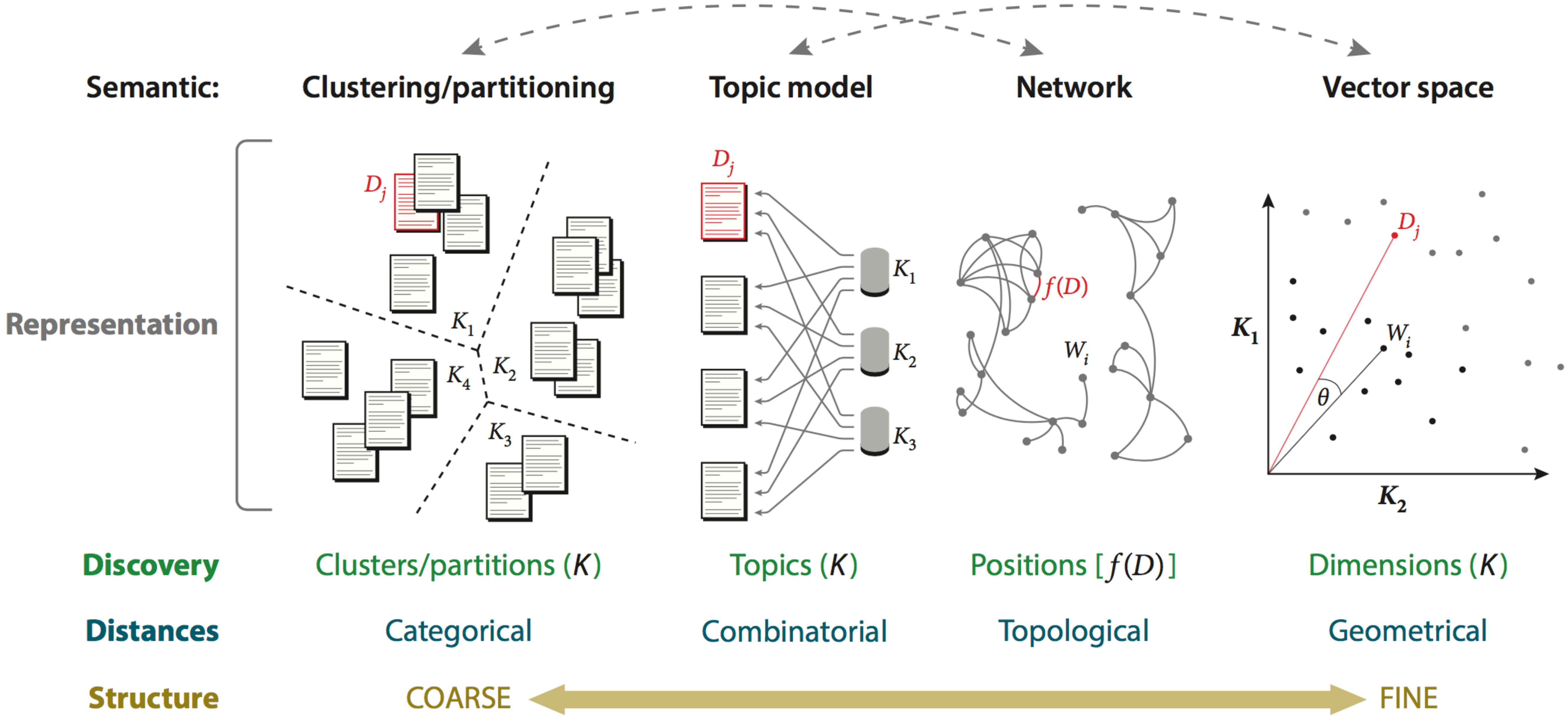
Dr. Tim Hannigan (tim.hannigan@ualberta.ca)
Richard F.J. Haans (haans@rsm.nl)
Milo Wang (swang7@ualberta.ca)

UNIVERSITY OF ALBERTA

RSM Erasmus ERASMUS UNIVERSITY | ROTTERDAM SCHOOL OF MANAGEMENT ERASMUS UNIVERSITY

Situating Topic Modeling in text analysis

- Overview of methodological approaches



Semantic:	Clustering/partitioning	Topic model	Network	Vector space
Representation				
Discovery	Clusters/partitions (K)	Topics (K)	Positions [$f(D)$]	Dimensions (K)
Distances	Categorical	Combinatorial	Topological	Geometrical
Structure	COARSE			FINE

Source: Evans & Aceves, 2016, ARS

What does Topic Modeling offer social scientists?

- meanings can be thought of as constellations of symbolic words that form latent constructs as topics (Mohr and Bogdanov, 2013)

- the key objective in topic modeling work is finding a simple representation of the symbolic complexity that preserves the structural integrity of a meaning system (Mohr, 1998).

  - provide us a reasonable automated content coding of large text corpora

  - enables us “to take the measure of large-scale social phenomena that we could not have previously been able to do” (Mohr et al., 2013)

Assumptions and practice

- bag of words assumption breaks down linguistic meaning structures (ie. sentences, parts of speech)
- LDA assumptions (stable topics, distribution over documents)
- how many topics? what sort of parameters?
  - validation involves “interpretive uncertainty” (DiMaggio, 2015)
- assume stable set of topics in corpus:
  - when texts are produced in organizational fields according to editorial standards (maybe a single topic) (eg. DiMaggio et al. 2013)
  - when knowledge base is well structured: ie. patents and literature may be more constituted using a variety of topics

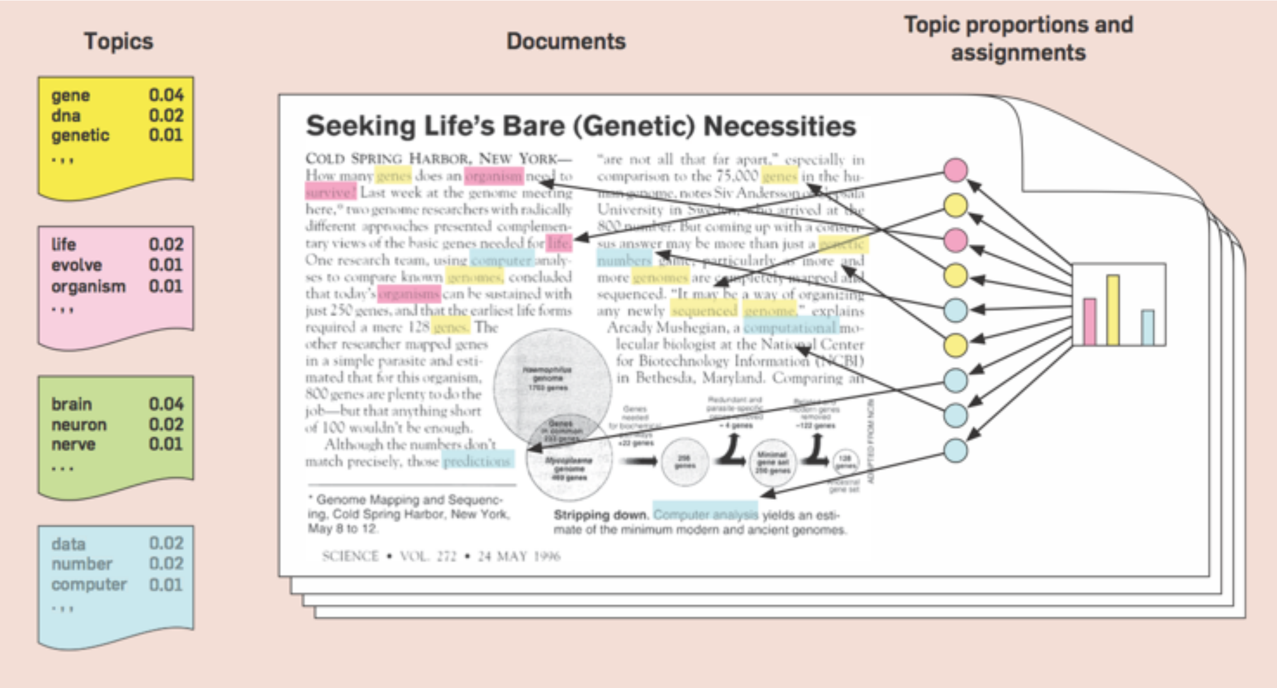
Deriving a topic using LDA Topic Modeling

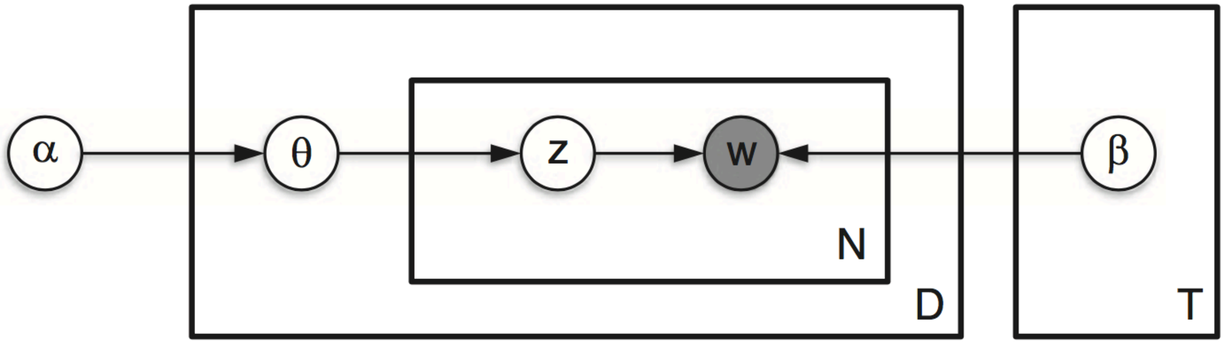
- in classic content analysis (Lasswell et al., 1952), the “goal was to find ways to measure ideas which were latent constructs indexed by constellations of word symbols”

- what is a topic in LDA?

  - a cluster of co-occurring words, a “word constellation” (Mohr et al., 2013) determined through the algorithm

  - the outcome of the analysis, available for researcher interpretation





Latent Dirichlet allocation in topic modeling

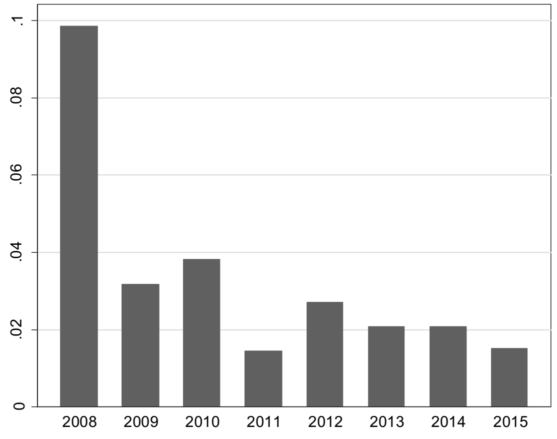
Source: Kaplan & Vakili, 2014

Applications of Topic Modeling

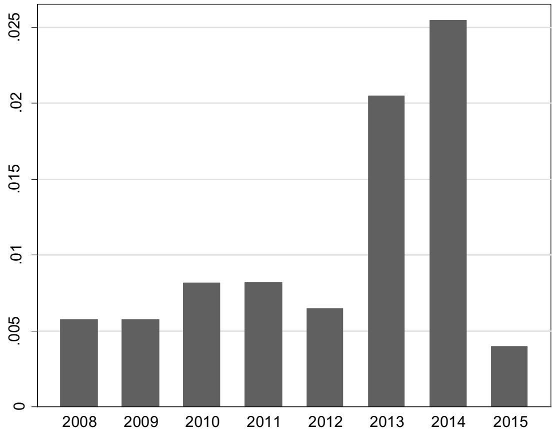
"The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are **useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments**."

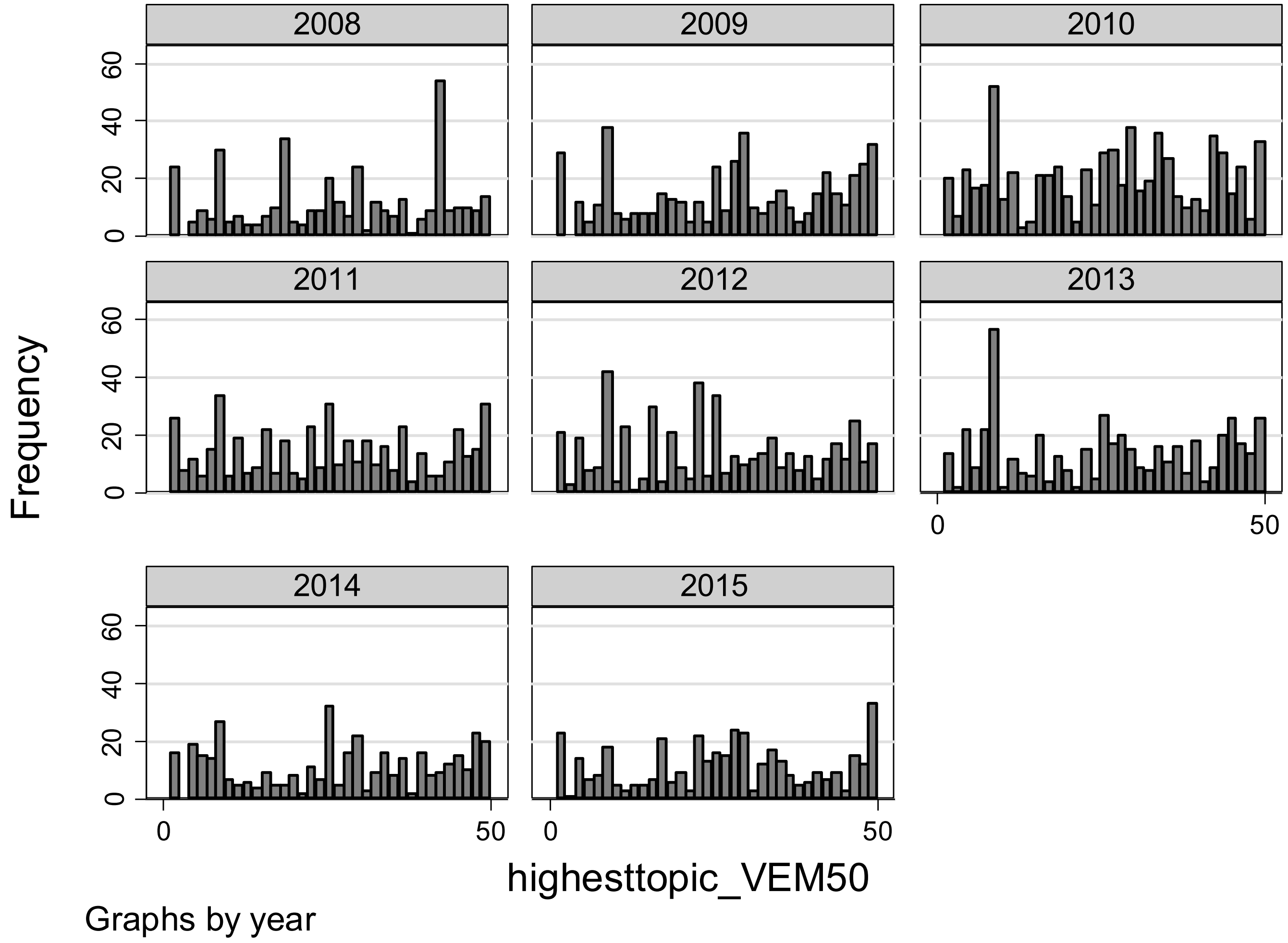
- Blei, Ng, and Jordan (2003)

Topic 42:
groups
political
million
campaign
money
spending
election



Topic 4:
law
state
religious
gay
marriage
rights





[demo in R]

Topic Modeling with Chinese Texts

- TM by default requires word boundaries (i.e., spaces between words)
- Unfortunately, no word boundaries in Chinese..
  - 主题建模是一项有用的技术。
  - Topic modeling is a useful technique.
  - ➢ “主题建模是一项有用的技术” (clause as a unit)
  - ➢ “主” “题” “建” “模” “是” “一” “项” “有” “用” “的” “技” “术” (char as a unit)

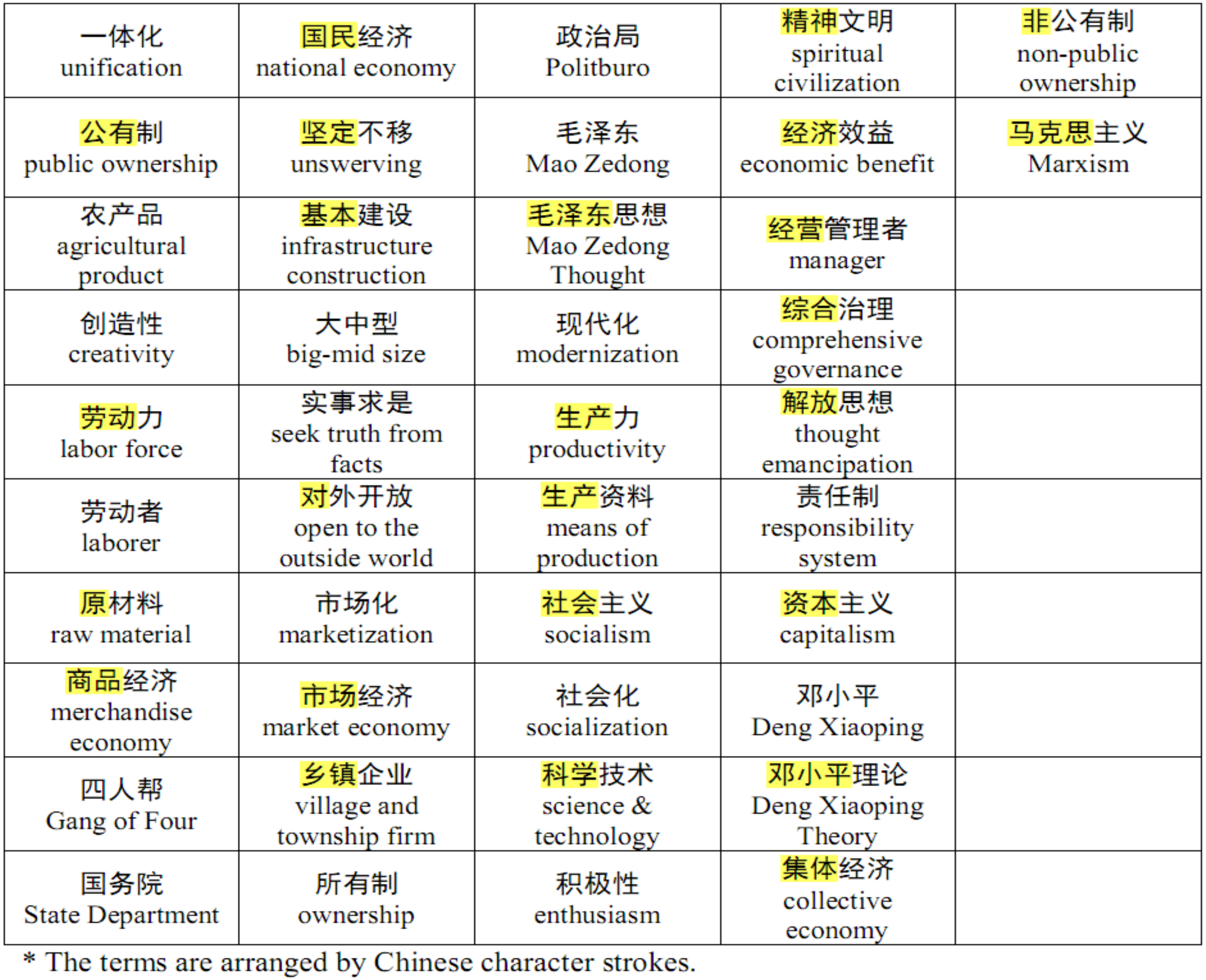
Topic Modeling with Chinese Texts

- Sometimes, char as a unit may be tolerated…
  - Miller. 2013. *Poetics*. Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach
- But, treating Chinese texts more seriously is warranted. Our task is to segment:
  - 主题建模是一项有用的技术
  - 主题 建模 是 一项 有用的 技术
  - Topic modeling is a useful technique

Segmenting Chinese

- Fortunately, there are multiple R-packages for automatic segmentation of Chinese texts:
  - *rmmseg4j*, *Ansj*, *jiebaR*, *Rwordseg*
  - ➢ Comparison: word vs. word + fixed phrase
    - “购买**力**” (purchasing power) as one unit; OR
    - “购买”(purchasing) “**力**”(power) as two units
- Yet, the default dictionary for Chinese stop words is much less satisfactory…

An Illustration of *Rwordseg*



一体化 unification	国民经济 national economy	政治局 Politburo	精神文明 spiritual civilization	非公有制 non-public ownership
公有制 public ownership	坚定不移 unswerving	毛泽东 Mao Zedong	经济效益 economic benefit	马克思主义 Marxism
农产品 agricultural product	基本建设 infrastructure construction	毛泽东思想 Mao Zedong Thought	经营管理者 manager	
创造性 creativity	大中型 big-mid size	现代化 modernization	综合治理 comprehensive governance	
劳动力 labor force	实事求是 seek truth from facts	生产力 productivity	解放思想 thought emancipation	
劳动者 laborer	对外开放 open to the outside world	生产资料 means of production	责任制 responsibility system	
原材料 raw material	市场化 marketization	社会主义 socialism	资本主义 capitalism	
商品经济 merchandise economy	市场经济 market economy	社会化 socialization	邓小平 Deng Xiaoping	
四人帮 Gang of Four	乡镇企业 village and township firm	科学技术 science & technology	邓小平理论 Deng Xiaoping Theory	
国务院 State Department	所有制 ownership	积极性 enthusiasm	集体经济 collective economy	

* The terms are arranged by Chinese character strokes.

Units with a higher than 20 frequency in all the CPC congress documents on economic reform

Other writing systems w/o word boundaries

- Japanese: ref. Japanese NLP Library
- Thai
- Lao
- Burmese
- …

DRUCKER SCHOOL of MANAGEMENT
CLAREMONT GRADUATE UNIVERSITY

Approaching Topic Modeling

Hovig Tchalian

OVERVIEW

Approaching Topic Modeling

- *Alternative Approaches*
- *Implementations*
- *Getting Started*

Two Alternative (and Distinct) *Supervised* Approaches

hLDA

- ***Supervised***, hierarchical (rank-ordered) topic generation
- Fewer parameters to choose
- Potentially more rigorous (Jordan)

Blei, Griffiths & Jordan, The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies (Journal of the ACM, Vol. 57, No. 2, Article 7 2010)

L-LDA

- ***Supervised*** (pre-labeled) topic generation
- Constrained to topics of interest
- Provides framework for apples-to-apples comparison

Ramage et al, Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora (*Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256)

hLDA (sunburst) example: *Malaysia Flight 370 (3.8.2014)*

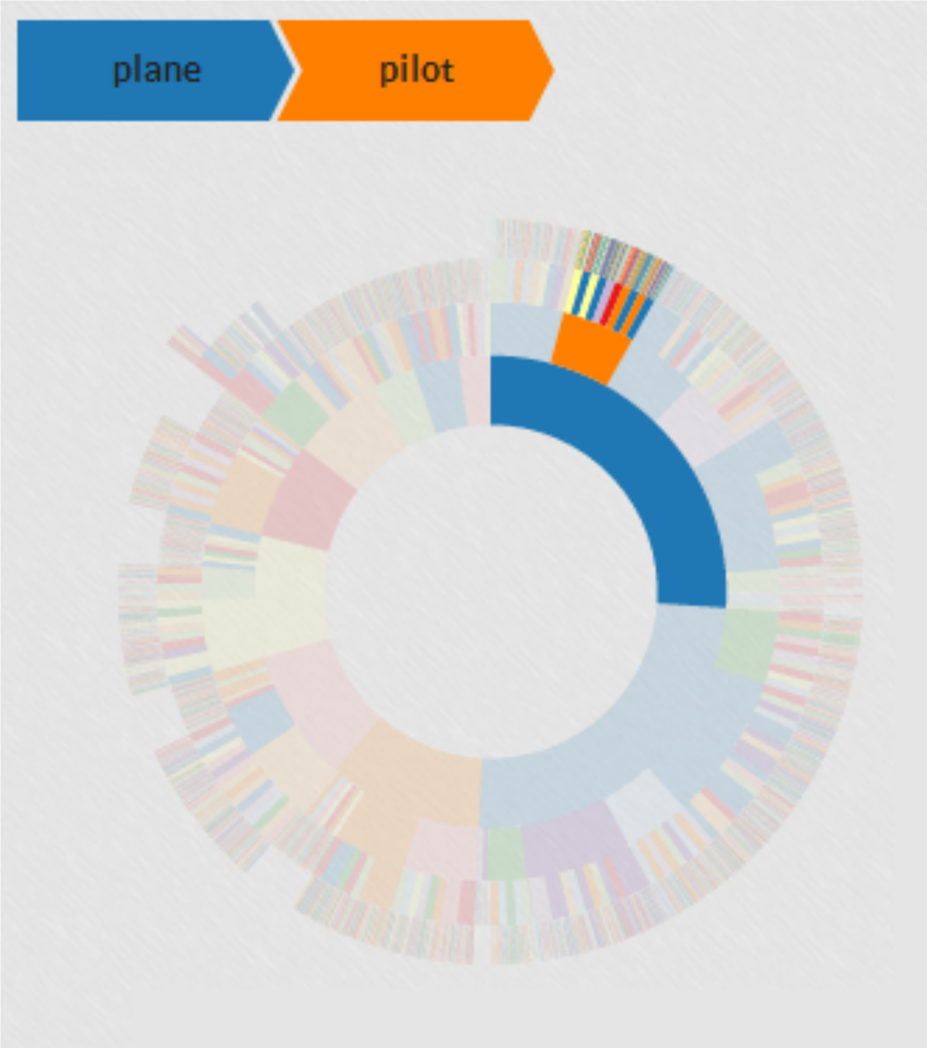


Figure 3: Our simple breadcrumb trail and contextual anchor offer constant context as the user explores the visualization. Highlighted slices within the contextual anchor are those currently displayed in the sunburst visualization.

plane, crash, crashed
plane, landed, land
plane, think, people
pilot, plane, hijacking
terrorist, terrorism, passports
suicide, pilot, ocean
Shah, Anwar, political
plane, China, world
phone, phones, cell
evidence, think, make

Table 1: The 10 high-level topics of the model generated from running HLDA on the Malaysia Flight MH-370 corpus. The bolded topics suggest specific theories regarding the status of the plane.

crash, water, crashed
failure, catastrophic, mayday
mechanical, failure, days
plane, ocean, did
plane, error, lost

L-LDA Example: *Cross-Disciplinary Dissertations*

Key Question: how well do cross-disciplinary dissertations (e.g., computer science and computational linguistics) fit their labels?

(– And secondarily, how close are corresponding departments?)

PROCESS

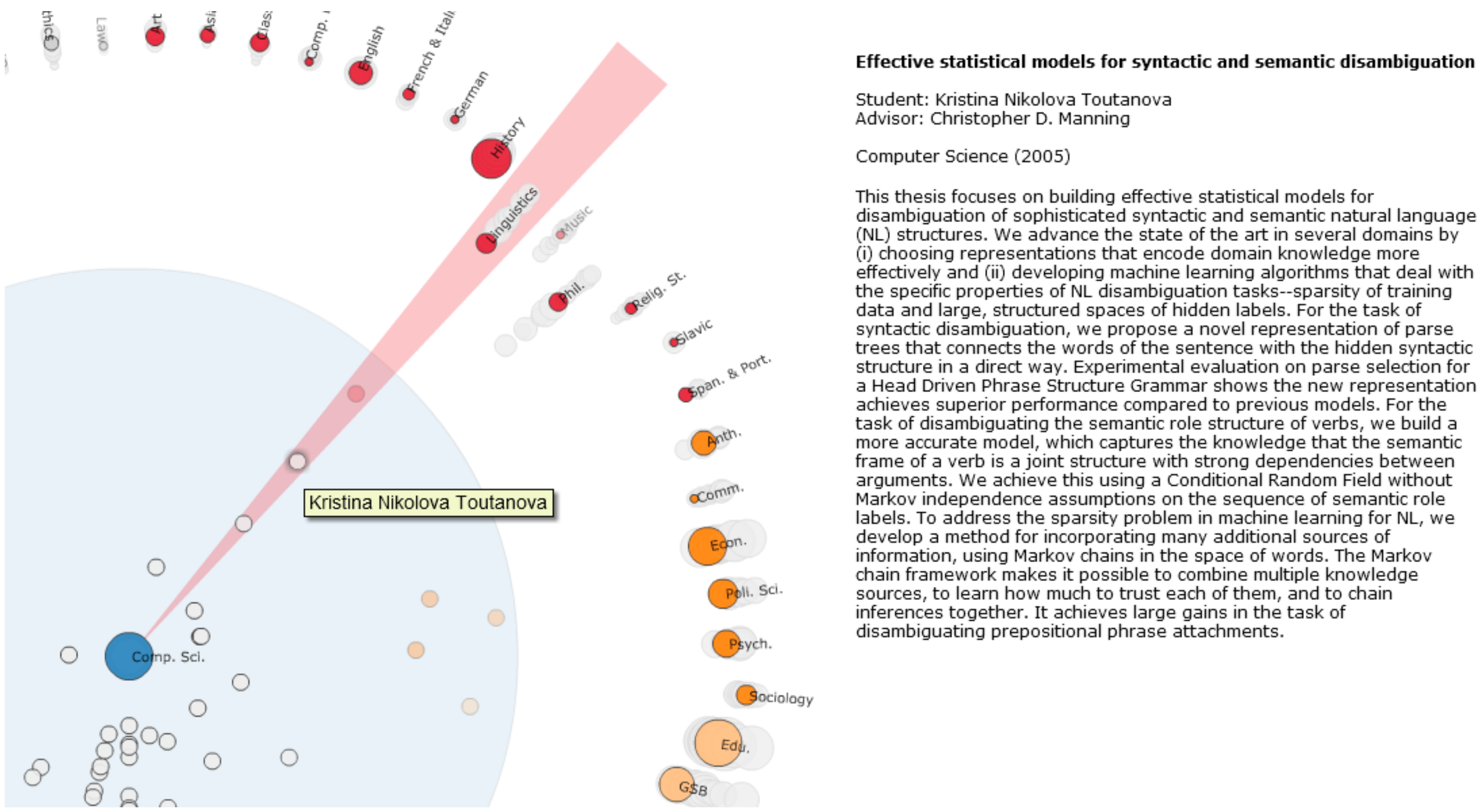
1. "Learn" topics based on department designations

2. Use departments as tags for L-LDA (i.e., departments = topics)

3. Ignore labels & rerun algorithm → compare results

Chuang et al, Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis (*CHI'12*, May 5–10, 2012)

L-LDA Output: *Cross-Disciplinary Dissertations*



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.

Figure 4. The Thesis View shows individual dissertations as small circles placed between the focus department and the next most similar department. Reading the original text of the dissertation enables experts to evaluate observed dept-dept similarities, and confirm the placement of three computational linguistics Ph.D.s that graduated in 2005.

Implementations

1. User-friendly / GUI tools – e.g., Topic Modeling Tool (TMT)
   - ✓*G Code Archive*: https://code.google.com/archive/p/topic-modeling-tool/
2. Mallet (Java) + Hierarchie for hLDA (*caveat*: Mallet hLDA in beta)
   - ✓*Mallet for Windows*: http://mallet.cs.umass.edu/
3. R and / or Python for "conventional" LDA and some variants
   - ✓R implementation covered in this PDW

Getting Started

- *Explore on your own, get a feel for output – start with GUI*
- *Partner with a technical expert – esp. Mallet implementation*
- *Experiment w R / Python – 6-mo learning curve but worth it*

An Excursus on Contextual Topic Models – T. Hannigan

- The practice of topic modeling can be blended within a combination of research designs
- Mohr et al. (2013, *Poetics*) used Burke’s 1941/1945 theoretical framing of dramatism in using modern computational methods to conduct a deep learning of text:
  - mapped Burke’s pentad of *Actors, Acts, Scenes* by using different modern techniques:
    - Actors (Natural Language Processing, Named Entity Recognition)
    - Acts (Semantic Network Analysis)
    - Scenes (LDA topic modeling)
  - LDA topic modeling determined the ’scenes’ underlying where actors act

An Excursus on Contextual Topic Models – T. Hannigan

- In practice, Mohr et al. (2013) Named Entity Recognition (NER) to identify nation states, organizations and people in texts

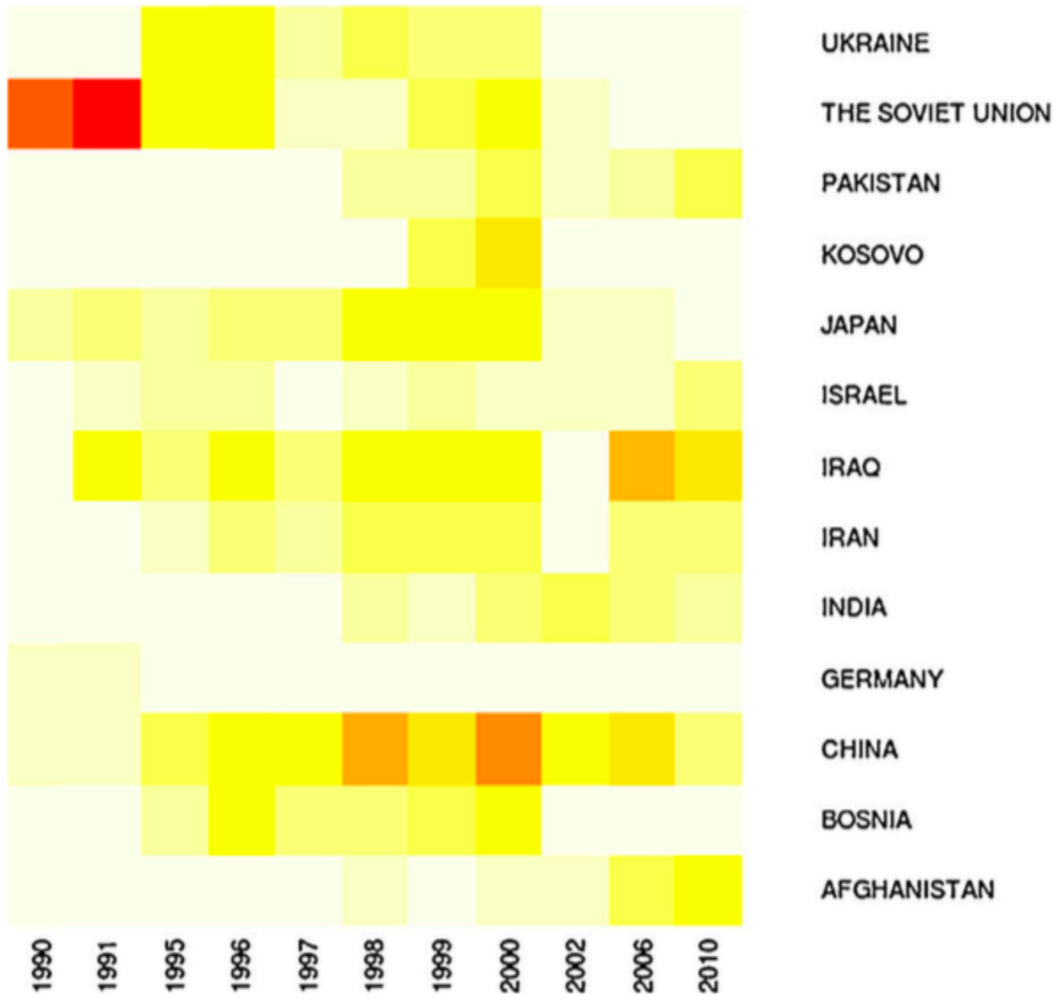


Fig. 1. Most frequent agentic terms—nation states (NSS, 1990–2010).

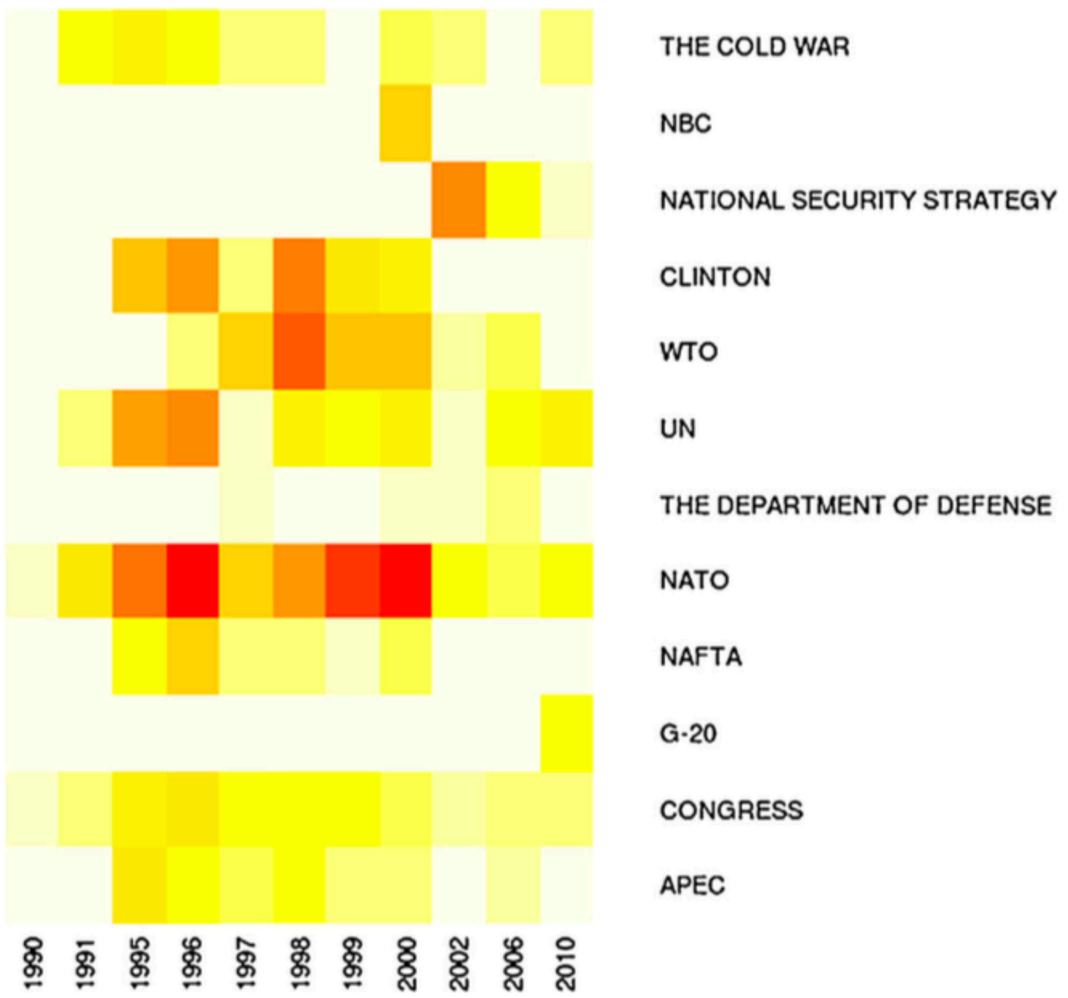


Fig. 3. Most frequent agentic terms—organizations and people (NSS, 1990–2010).

An Excursus on Contextual Topic Models – T. Hannigan

- In practice, Mohr et al. (2013) used LDA topic modeling to focus in on texts with high topic weighting to do deep reading analysis

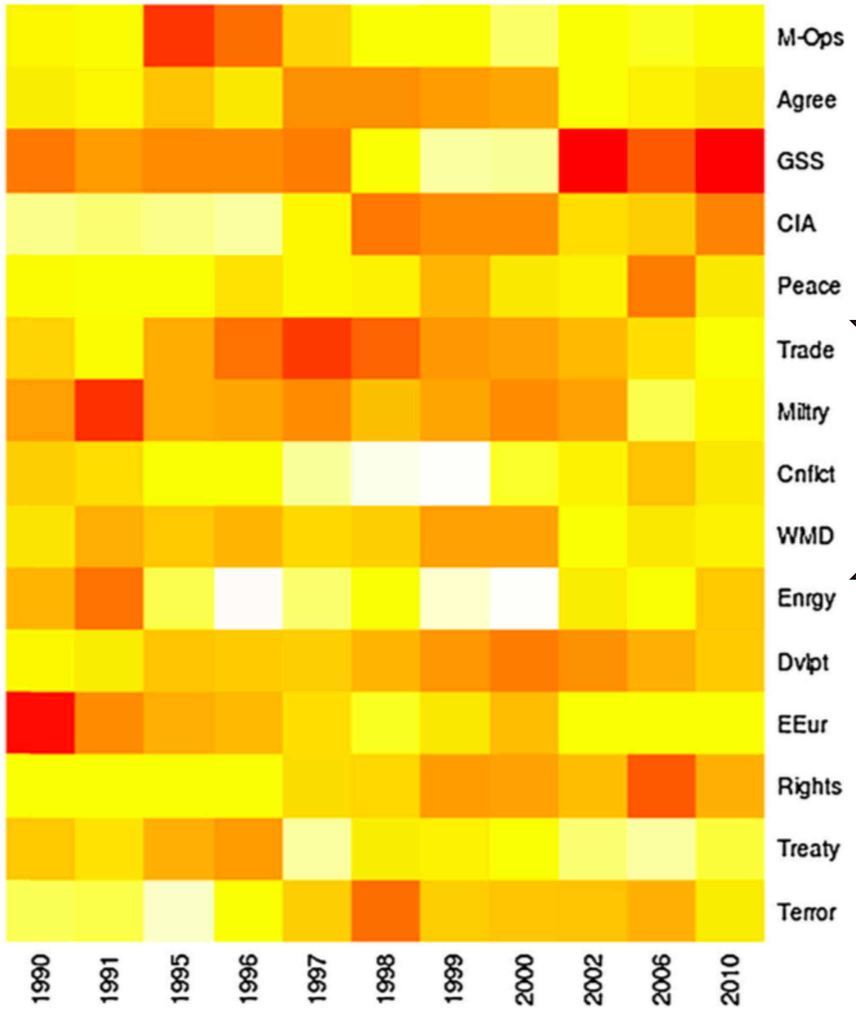


Fig. 11. 15 level topic model of NSS corpus—topic distribution across years.

Focus on topic 0, terrorism

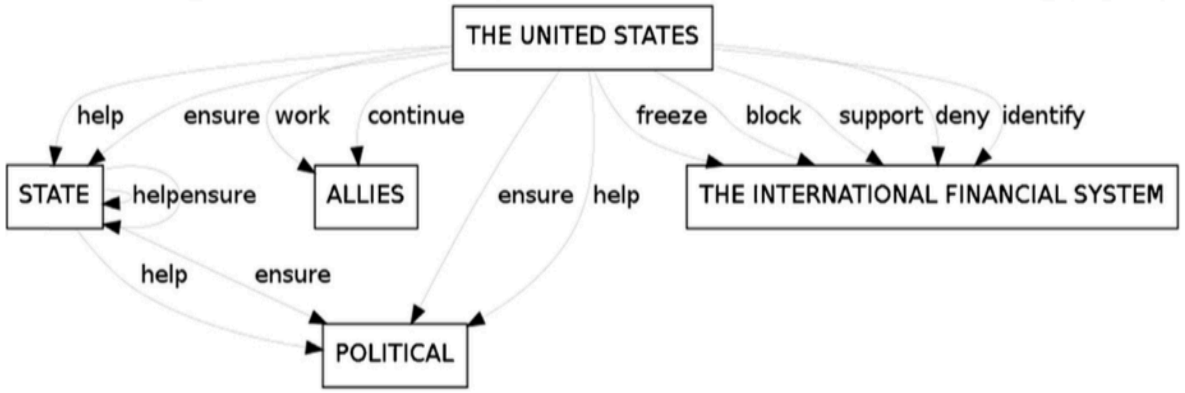


Fig. 14. 2002 NSS George W. Bush (Actor-Act-Actor); topic = 0 (terrorism).

QUESTIONS?

TOPIC MODELING: OPPORTUNITIES AND CAUTIONS

KEYVAN VAKILI
LONDON BUSINESS SCHOOL

AOM 2017

Have results; time to publish

- Still likely that editors and/or reviewers don't know much about topic modelling
- Some critical steps:
  - **Explain the method briefly and clearly (fortunately, there is prior art to cite)**
  - **Justify the use of topic modeling**
  - **Be transparent about all choices**
  - **Validate, validate, and validate**

Justify the use of topic modeling

- Topic modeling is only good for certain applications:
  - When you’re looking for latent topics
  - When you have large amount of text
  - When subjective intervention of human coders is a very costly

- Not so good for:
  - Analyzing narratives, semantics, tone, style, or anything that relies on the word sequences
    - There are specialized tools for each case
    - They can be combined with topic modeling
  - If you already have a preset categorization
    - Supervised classification would be a better choice

Be transparent

- Explain the data collection process clearly:
    - Data source
    - Which words are excluded
    - Which texts are excluded/included
        - Shorter than a certain length?
        - Duplicates?
    - Stemming?
    - Spelling errors?
- Explain how the main parameters are selected:
    - Number of topics
    - Topic smoothing and term smoothing parameters
- Show all the identified topics and their top 10 terms (preferably with term weightings), at least in appendix
- Show one or two representatives for each topic

Validation & Sensitivity Analysis

- Statistical techniques
  - Fit
- Semantic Validity using expert validation
  - Ask experts to verify that topics are meaningful and distinct
  - Use expert coding/labeling and inter-coding reliability
  - Evaluate/rate co-assignments of documents to same topics
  - Use experts to flag garbage topics
- Predictive validity
  - Use portion of data for modeling and the rest to measure prediction fitness
  - External validity assessment: certain events should increase or decrease the prominence of certain topics which should be visible
- Do sensitivity analysis around the input parameters
  - Results/Interpretations should be robust to small changes in the number of topics
  - Change in results due to change in the number of topics should make sense

Other quantitative methods of validation

- Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. "**Automatic evaluation of topic coherence**." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100-108. Association for Computational Linguistics, 2010.
- Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. "**Reading tea leaves: How humans interpret topic models.**" In *Advances in neural information processing systems*, pp. 288-296. 2009.
- Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. "**Optimizing semantic coherence in topic models.**" In *Proceedings of the conference on empirical methods in natural language processing*, pp. 262-272. Association for Computational Linguistics, 2011.

What can be considered a topic?

- Any language construct that can be signified with a set of words
- Cognitive frames (managerial, political, cultural, media, etc.)
- Technological or scientific domains/paths
- Product/Market/Industry categories
- Attention direction

Sky is the Limit

- Time trends
    - Category/theme emergence, decay, fads
- Aggregated associations
    - Revealed identity and identity changes
    - Locating actors in the content space; measuring distance
    - Multiple category memberships
    - Fuzzy categories
- Other ideas
    - Citation network among topics
    - Knowledge diffusion
    - Topic recombination
    - Refined, dynamic categorization (industry classification, patent classification)

Q&A