Topic Modeling

A (SOMEWHAT) HANDS-ON APPROACH TO ANALYSIS AND THEORY BUILDING

The Topic Modeling Group:
Hannigan, Haans, Tchalian, Vakili, Glaser, Wang,
Kaplan & Jennings

Plan of the Session

- Noon - start
- 12:03 – Dev & Tim with welcome & plan of the session.
- 12:05 - Sarah on textual analysis and topic modeling
- 12:15 - Tim with a Primer on Rendering - a few questions
- 12:25 - Richard with Tim & Hovig doing a hands on training and example to show rendering (with support at tables from each person) - questions
- 1:15 - Vern, Hovig, Milo on pre-processing, processing, theorizing - some questions
- 1:40ish - Keyvan via SKYPE on issues with T.Models. Vern as moderator (and backup) – some questions
- 1:58 - Dev & Tim - quick thanks and wrap up.
- 2:00 end

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"

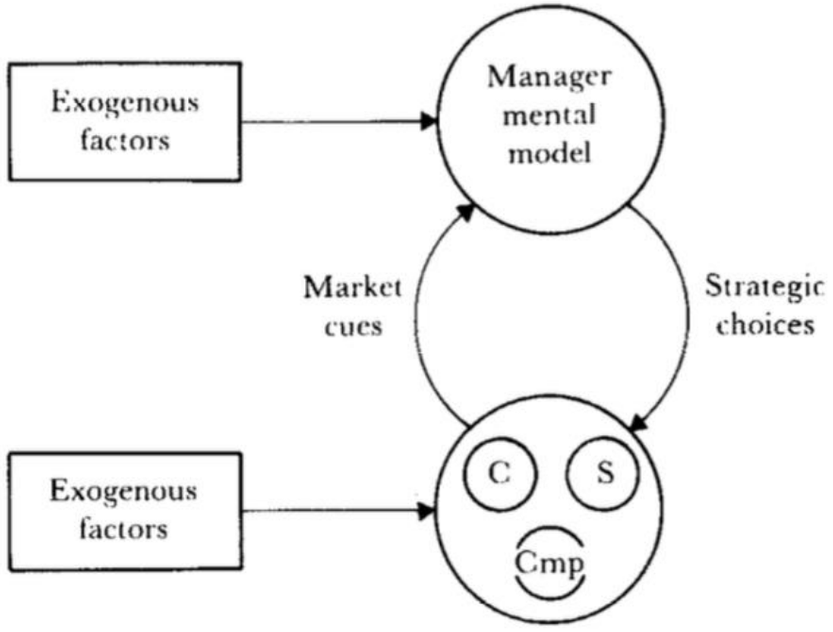


Figure 1. Reciprocal influence of technical and cognitive levels of analysis

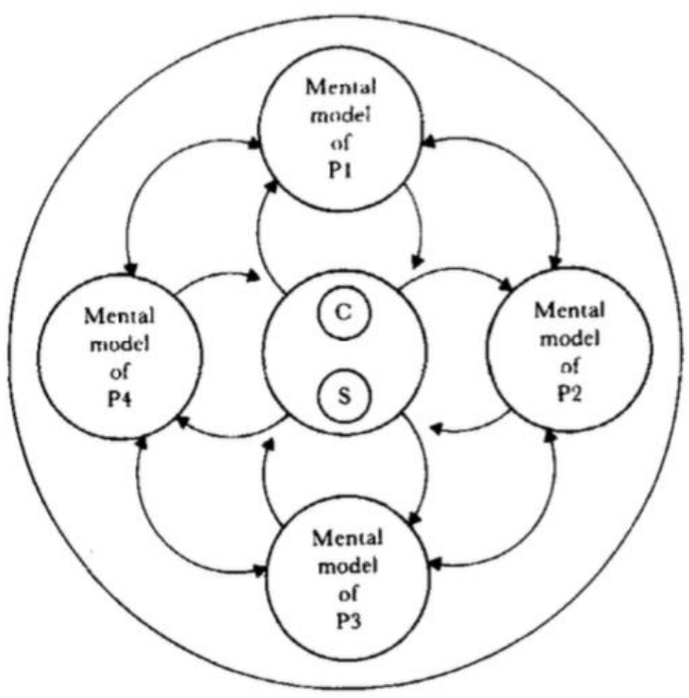
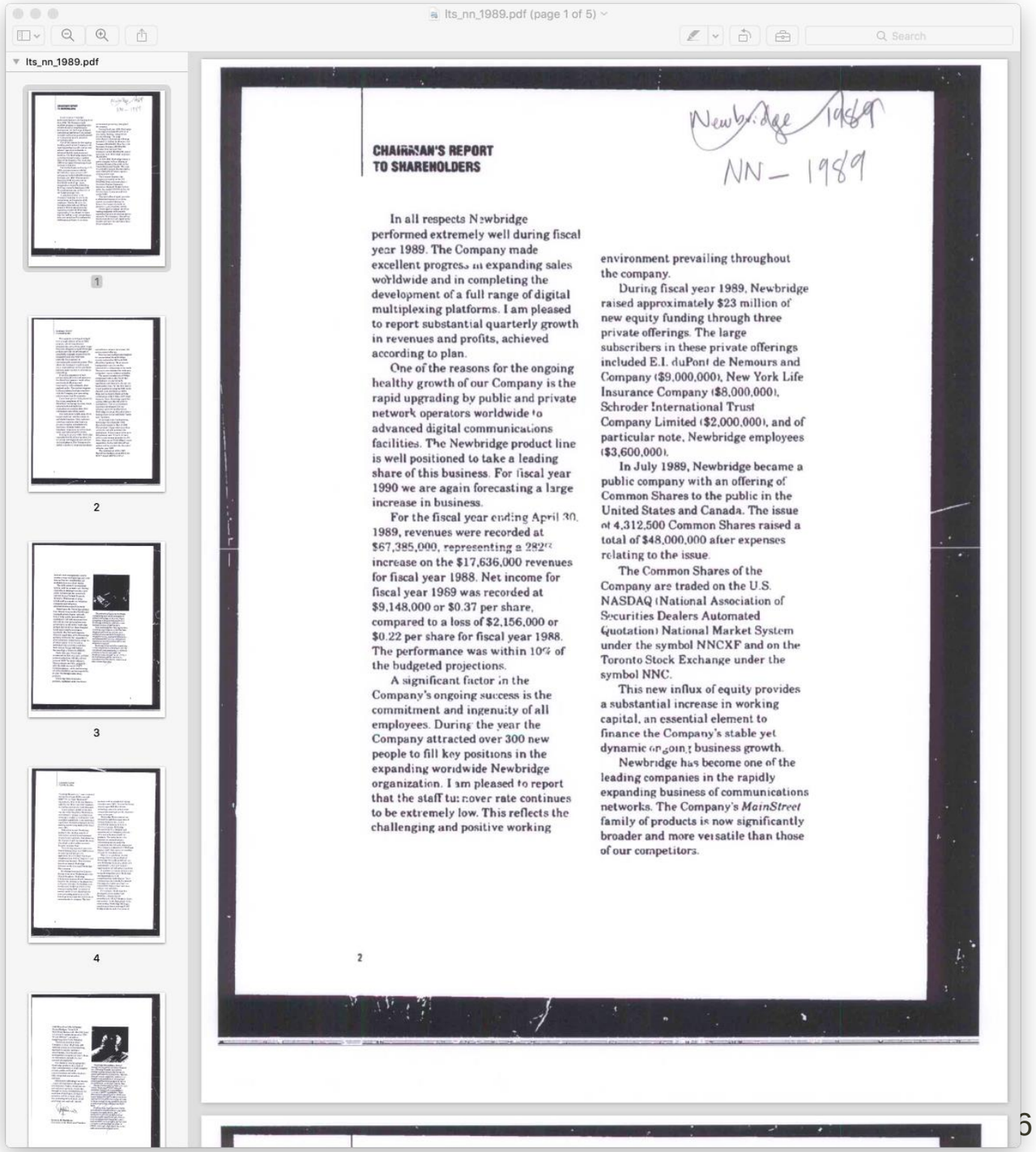


Figure 2. Mutual enactment processes within an industrial sector

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



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 decade
- 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
- And see our draft article in the *Academy of Management Annals* for many examples

CAUTIONS

- Different approaches depending on research question and available documents

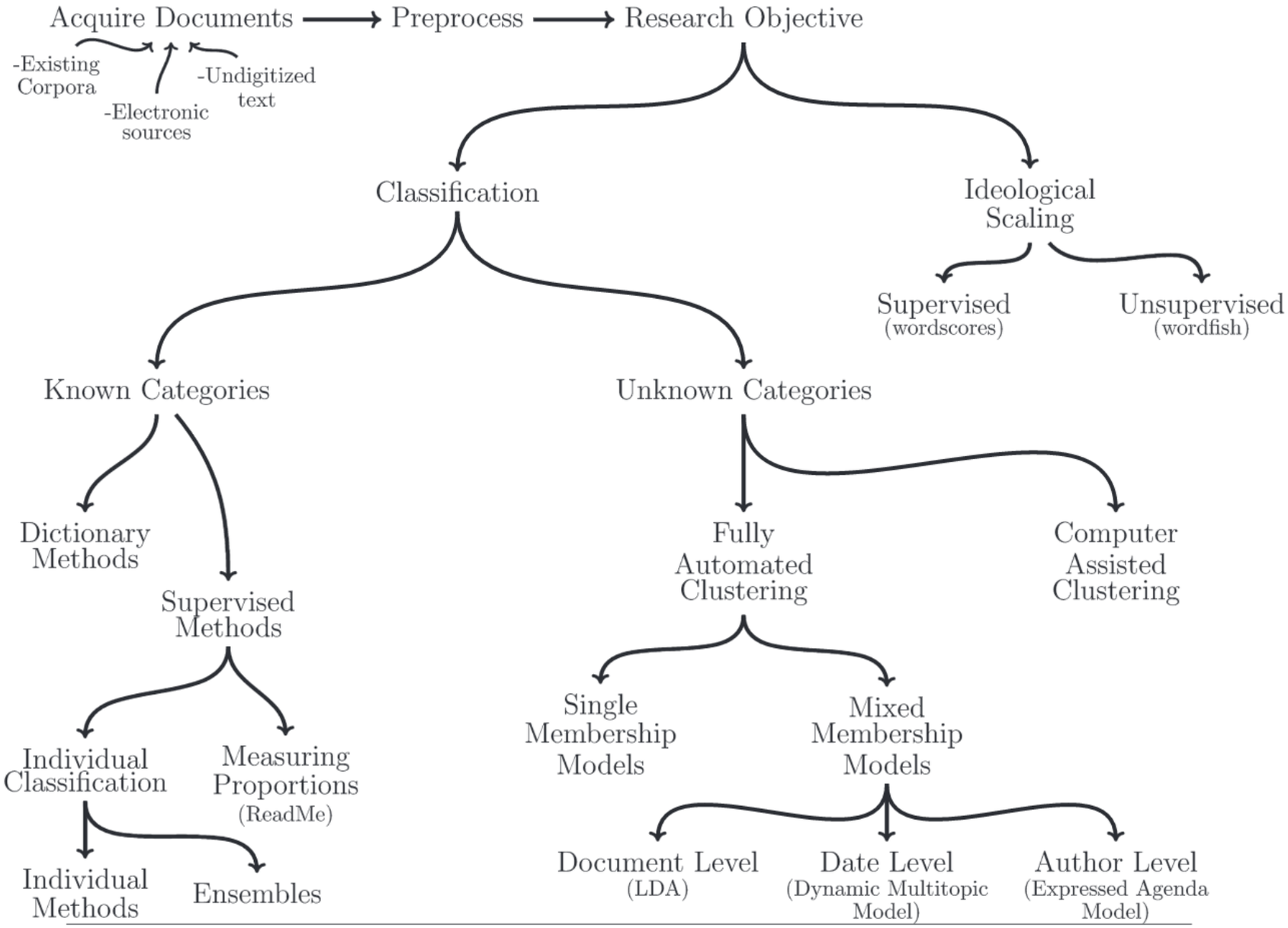


Fig. 1 An overview of text as data methods.

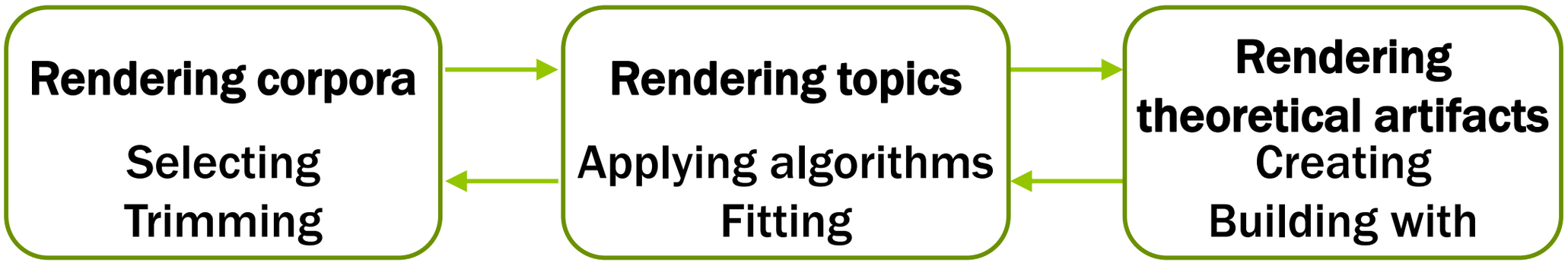
Source: from Grimmer and Stewart (2013) "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*

CAUTIONS

- 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

RENDERING PROCESS



Source: Hannigan, Tim, Richard F. J. Haans, Keyvan Vakili, Hovig Tchalian, Vern L. Glaser, Milo Wang, Sarah Kaplan, and J. Devereaux Jennings, Topic Models in Management Research: Rendering New Theory from Big Textual Data (Academy of Management Annals)

A Primer with Rendering:

topic modeling in practice

AOM 2018 PDW – Topic Modeling
Aug 11, 2018

Tim Hannigan
University of Alberta
(tim.hannigan@ualberta.ca)

Topic modeling is a non-linear process

Preprocessing our data. Remember: Garbage In Garbage Out

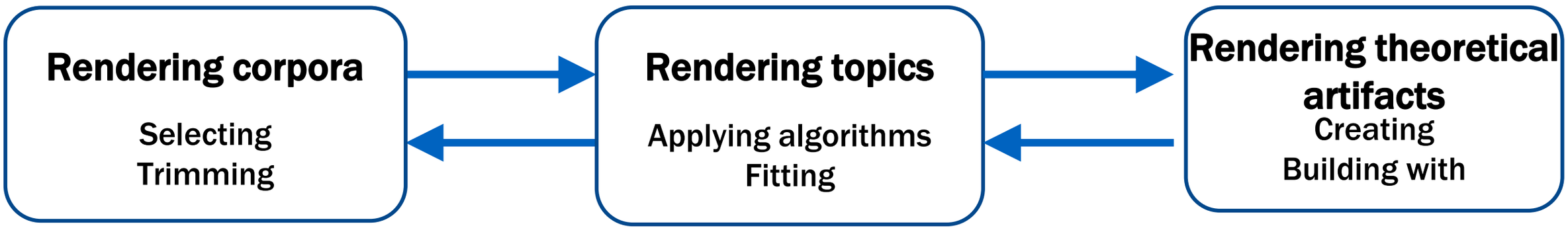
"NLP is 80% preprocessing."

-Lev Konstantinovskiy

Python Gensim community manager

- Computer scientists employing topic modeling have long recognized the need for preprocessing texts
  - This usually takes the form of standardizing words, removing words that contribute little meaning (known as *stopwords*)
- However, as management scholars, we recognize that this process is full of theoretical and methodological assumptions, each with implications for the work involved
  - *Thus, there is no set standard for preprocessing all texts, this requires thoughtful engagement with your data*

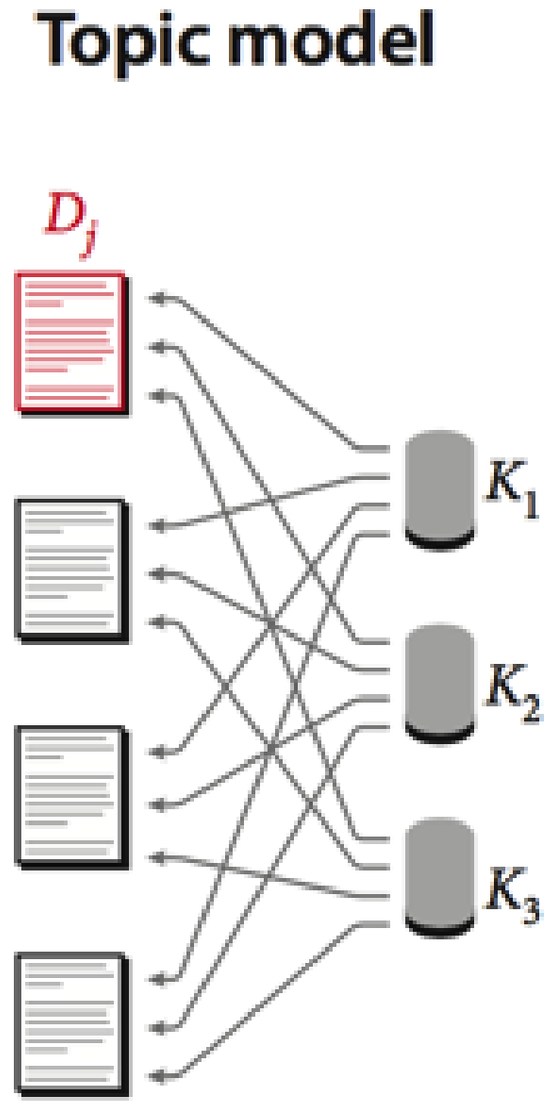
Characterizing Rendering as Topic Modeling practice



- **Rendering** is *a process of generating provisional knowledge by iterating between selecting and trimming raw textual data, applying algorithms and fitting criteria to surface topics, and creating and building with theoretical artifacts, such as processes, causal links or measures*
  (Hannigan, Haans, Vakili, Tchalian, Glaser, Wang, Kaplan & Jennings, 2018, Academy of Management Annals)

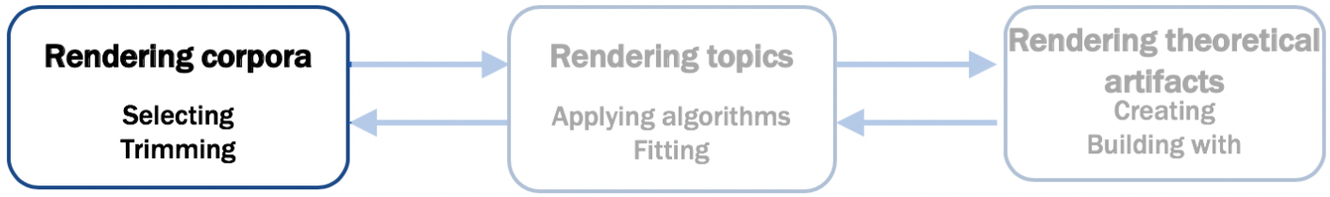
Approaching topic modeling

- 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).
  - Topic modeling enables distant reading, provides us with 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)
- Assumes the same generative processes behind texts in a corpora (ie. documents are created based on drawing from a set of topics)



Source: Evans & Aceves, 2016, ARS

Rendering Corpora



- This practice is guided by theoretical and empirical considerations, as the analyst curates a corpus
- *selecting* types of textual data
  - Which texts to include in the corpus? Need to account for **language**, **authoring**, **document sources**
- *trimming* that textual data is a process of shaping it into a corpus
  - Activities here include linguistic processing techniques such as: removing **stopwords**, and transforming/standardizing words through **stemming** (words paired down to stems) and **lemmatizing** (ie. converting to singular forms, or to high-level synonyms through a linguistic thesaurus such as WordNet)
- Conjoining words into meaningful *phrases* through transformation*;* based on knowledge of corpus or through an n-gram analysis

Rendering Topics

- The analyst *applies an algorithm* to identify appropriate topics
- The most well known topic modeling algorithm is called Latent dirichlet allocation, or **LDA**
- The analyst configures LDA by setting a number of topics and then inspecting output
  - There are quantitative measures of *fit* (ie. coherence metric)
  - But also *validity* ; are the topics meaningful to a domain expert?
- Running LDA gives two outputs:
  - A matrix of topic-documents, showing weights per topic
  - For each topic, a list of topic-words – with weights for each word in each topic

Rendering Topics: Outputs

- For each topic (recall that you set the number of topics for the LDA algorithm to compute), you will get a breakdown of weights for each word, often in descending order

- This can help you track the meaning of each topic and whether the model is valid or needs to be re-run

- Here is an example of a topic-word output (using topic #18) from a project on a British Parliamentary Expense Scandal

Topic # # of words to show

```
lda_model.show_topic(18, 20)
```

```
[('michael_martin', 0.064001233758723 06),
 ('mp', 0.05337934225238076 5),
 ('speaker', 0.04383698962871573),
 ('house', 0.03036203107529784),
 ('common', 0.02878127771137757),
 ('parliament', 0.01717623472259706),
 ('expense', 0.01370628831399159 5),
 ('confidence', 0.01060261402629448 2),
 ('office', 0.00913752554266106 4),
 ('motion', 0.00855920114122681 9),
 ('westminster', 0.00684350541697189 3),
 ('member', 0.00674711801673285 3),
 ('reputation', 0.00665073061649381 2),
 ('douglas_carswell', 0.00591818637467710 25),
 ('nick_clegg', 0.00580252149439025 3),
 ('expense_scandal', 0.00547480433357751 4),
 ('monday', 0.00535913945329066 6),
 ('time', 0.00518564213286039 3),
 ('statement', 0.00508925473262135 2),
 ('handling', 0.00481937001195203 7)]
```

Rendering Topics: Outputs

- You can validate topics by manually reading documents that feature a high topic for a particular topic
- Here is an article having a high weight for topic 18

```
lda_model.show_topic(18, 20)
```

```
[('michael_martin', 0.064001233758723 06),
 ('mp', 0.053379342252380765),
 ('speaker', 0.043836989628715 73),
 ('house', 0.030362031075297 84),
 ('common', 0.028781277711377 57),
 ('parliament', 0.017176234722597 06),
 ('expense', 0.013706288313991595),
 ('confidence', 0.010602614026294482),
 ('office', 0.009137525542661064),
 ('motion', 0.008559201141226819),
 ('westminster', 0.006843505416971893),
 ('member', 0.006747118016732853),
 ('reputation', 0.006650730616493812),
 ('douglas_carswell', 0.005918186374677 1025),
 ('nick_clegg', 0.005802521494390253),
 ('expense_scandal', 0.005474804333577514),
 ('monday', 0.005359139453290666),
 ('time', 0.005185642132860393),
 ('statement', 0.005089254732621352),
 ('handling', 0.004819370011952037)]
```

Speaker Michael Martin resigns: after the MPs' expenses scandal , one obvious candidate can restore dignity to the House

These are unprecedented times and only a Speaker of real stature will do, says Simon Heffer.

Image 1 of 2

Michael Martin lost the confidence of MPs across the House Photo: PA

By Simon Heffer
7:38PM BST 19 May 2009

MPs' expenses
Politics » Labour »
Comment »
Simon Heffer »
MPs expenses: rebuilding politics »

In MPs' Expenses

How the Daily Telegraph exposed the dry rot fraudster

Bizarre MPs' expense claims

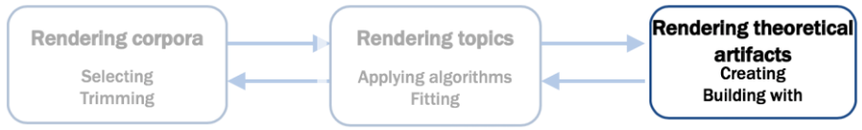
RSM Erasmus University
ROTTERDAM SCHOOL OF MANAGEMENT
ERASMUS UNIVERSITY

Rendering Topics: Outputs

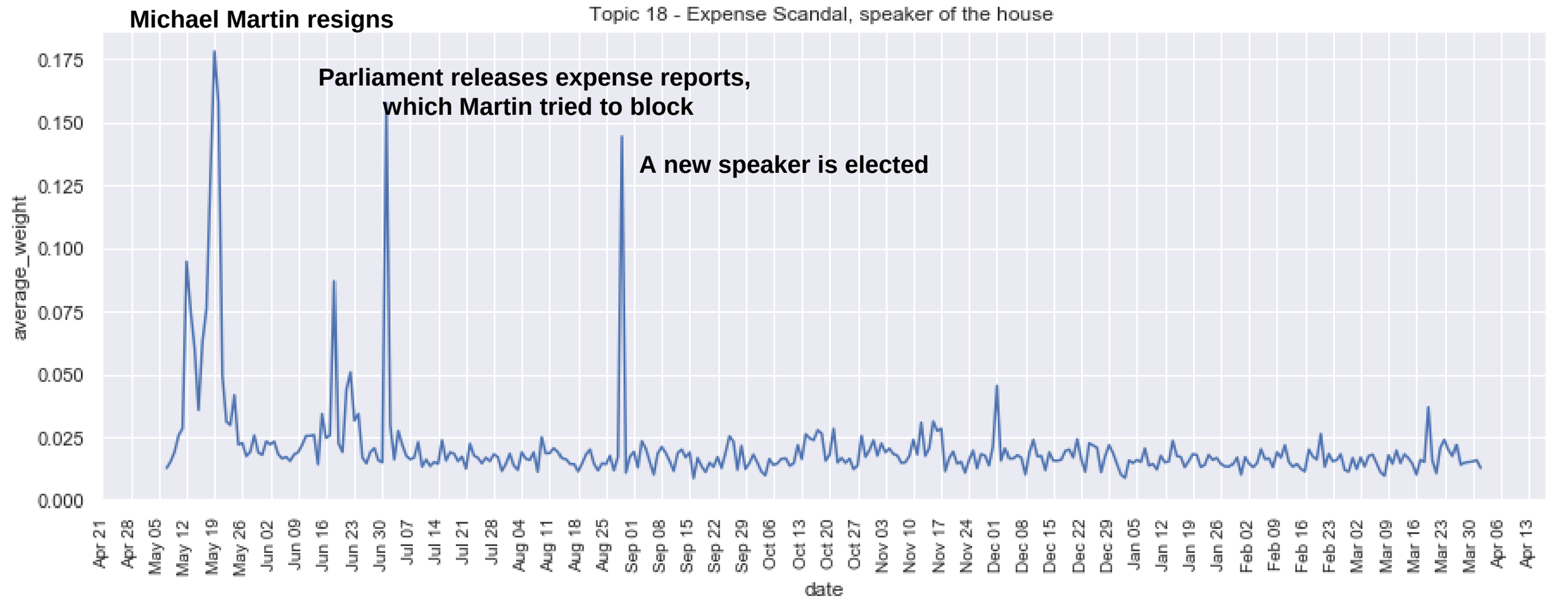
topic_0	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8	topic_9	...	topic_34	topic_35	topic_36	topic_37	topic_38	topic_39	dominant_topic	date	article_id
0.02	0.04	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.07	...	0.01	0.05	0.01	0.01	0.04	0.01	23	2009-05-07	313792
0.15	0.00	0.02	0.03	0.03	0.01	0.01	0.00	0.00	0.00	...	0.01	0.04	0.03	0.01	0.00	0.00	23	2009-05-07	313789
0.01	0.06	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.01	...	0.02	0.01	0.01	0.01	0.01	0.02	23	2009-05-07	313793
0.01	0.16	0.01	0.10	0.01	0.01	0.05	0.01	0.02	0.01	...	0.01	0.01	0.02	0.01	0.06	0.01	1	2009-05-07	313791
0.06	0.01	0.01	0.04	0.02	0.03	0.05	0.09	0.02	0.01	...	0.01	0.01	0.02	0.01	0.12	0.02	22	2009-05-07	313788

- A matrix of topic-documents, showing weights per topic
  - For each document, all topic weights add up to 1
- Can determine the prevalent, or dominant, topic per document
- Can combine with other information such as date
- **Working with these topic outputs, one can begin to build provisional knowledge structures**

Rendering Theoretical artifacts



- Once topic model has been curated and tuned, outputs can be rendered into theoretical artifacts (Whetten, 1989), and can be used in theory building
- The analyst can average out weights for a particular topic based on date information and track the prevalence trend over time
- *What does a topic correspond to theoretically? As you contextualize a topic within a case, you may modify an existing construct an enable novel understanding*



Topic Modeling–
A Demonstration

Richard Haans, Tim Hannigan, Hovig Tchalian

UNIVERSITY OF ALBERTA

RSM Erasmus University
ROTTERDAM SCHOOL OF MANAGEMENT
ERASMUS UNIVERSITY

Trends in Rendering Theory with Topic Modeling

Vern Glaser

Analyzing frames and framing

Topic modeling:

- Allows the identification and tracking of high level frames over large corpora
- Inductive nature allows for discover of unanticipated frames and audiences
- Algorithmic nature ensures replicability and can be paired with other methods
- Captures heteroglossia

Examples:

- Augustine & King 2017
- DiMaggio, Nag & Blei 2013
- **Fligstein, Stuart Brundage & Schultz 2017**
- Huang et al. 2017
- Jha & Beckman 2017
- Levy & Franklin 2014
- Weber et al. 2013
- Giorgi & Weber 2013

Framing and the financial crisis

(Fligstein et al 2017)

- Fligstein et al (2017) test hypotheses associated with frames of Federal Open Market Committee regulators
  - Primary frameworks either rested on macroeconomic or banking frames
  - Finance and banking frame was more prevalent among members who had a professional background in the private banking sector
  - Discordant facts raising negative scenarios were marginalized or re-interpreted as “normal” outcomes
  - The macroeconomic frame hindered the FOMC ability to perceive the sources and consequences of the financial crisis
- Key rendering moves: developing counts of topics in different time periods and drilling in on 2 specific “cases” to ensure validity

Framing and the financial crisis

(Fligstein et al 2017)

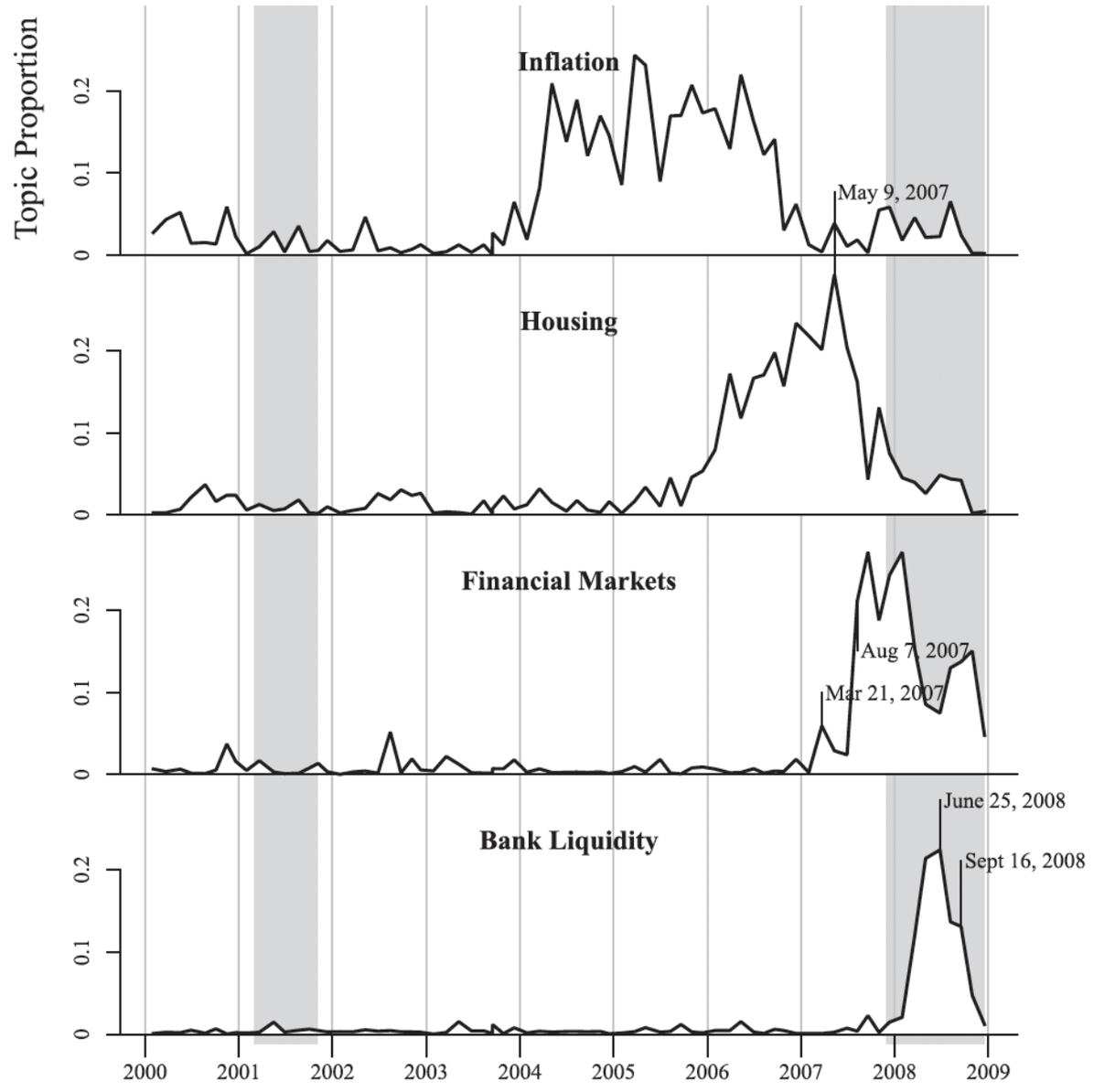


Figure 1b. Topic Proportions over Time; Framing Topics
Note: The height of each line represents the proportion of words in a given transcript assigned to that topic. Gray vertical bars indicate periods of recession.

Framing and the financial crisis

(Fligstein et al 2017)

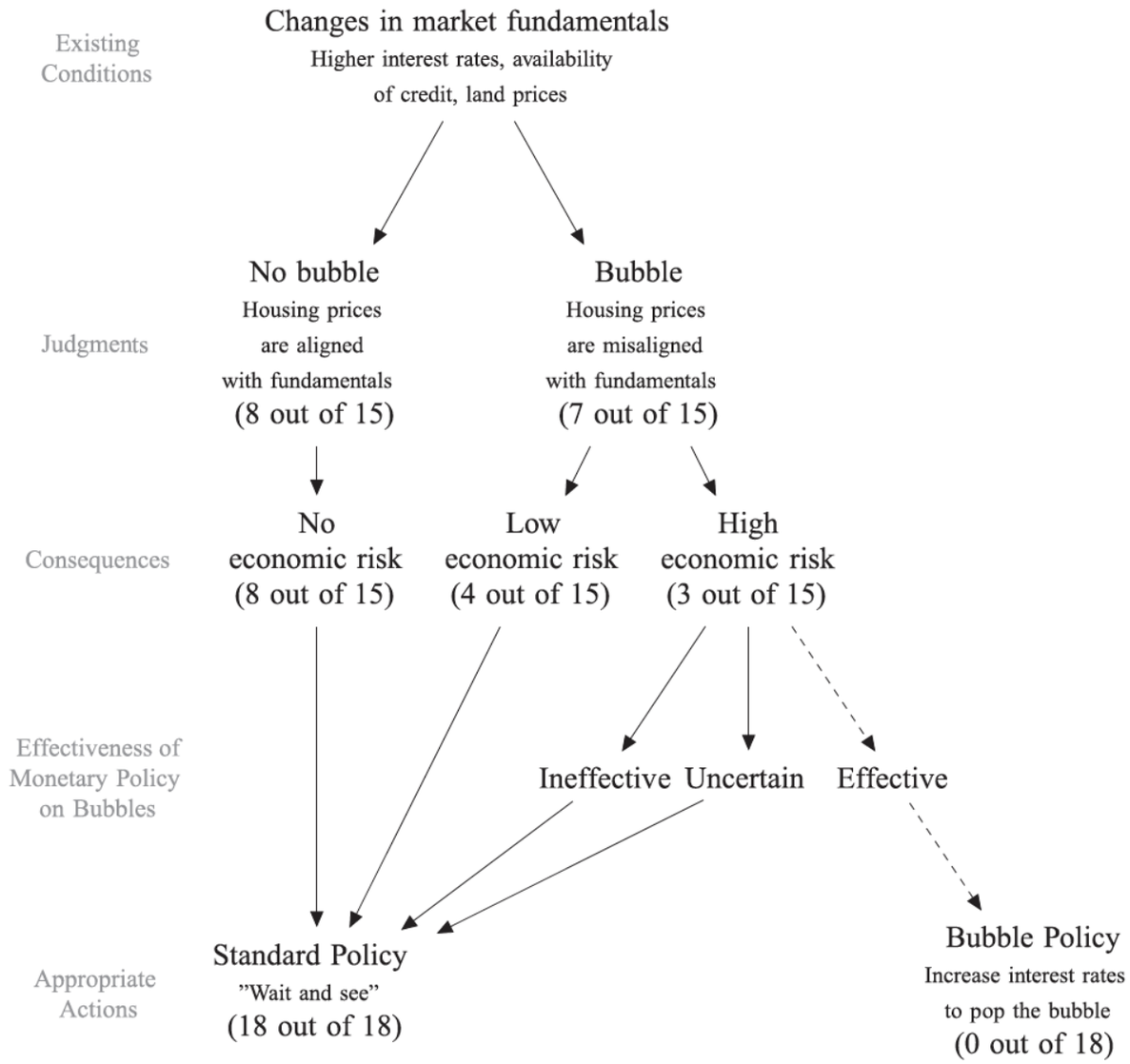


Figure 2. Reactions to Housing Overvaluation
Note: Positions taken during the FOMC meeting June 29 to 30, 2005. Counts shown indicate the number of participants who explicitly took that position.

Identifying themes and topics in historical corpora

Topic modeling:

- Accounts for the entire body of knowledge (the corpus) and also how it evolves over time in a field
- Highlights the emergent nature of topics independently of *ex post* impact or popularity, allowing the emergent and *ex post* impact to be compared
- Allows for simultaneous identification and tracking of topics
- Inductive nature allows for discovery of unanticipated (to the researcher) topics

Examples:

- Antons, Kleer & Salge 2016
- **Croidieu & Kim 2018**
- Jockers & Mimno 2013
- Marshall 2013
- Miller 2013
- Wang, Bendle, Mai & Cotte 2015
- Wang, Steele, Hinings, Ocasio, Marquis & Miller 2017

Lay-expertise legitimation (Croidieu & Kim 2018)

- Croidieu & Kim (2018) theorize how lay actors become recognized as legitimate experts
  - Building an advanced collective competence
  - Operating in an expanded public space
  - Providing transformational social contributions
  - Expanding an original collective role identity

- Key rendering move: applying the Gioia Method to topic modeling outputs in order to identify themes and topics in historical corpora

Lay-expertise legitimation (Croidieu & Kim 2018)

Illustrative topics vocabularies	First-order concepts (derived from topic vocabularies)	Second-order themes	Aggregate dimensions (or mechanisms)
Topic #391: hole screw brass drill fasten place strip thread solder requir make center machin wood mount support knob point give hold	Building a basic radio set	CC1. Forming a rudimentary competence	Mechanism 1: Build an advanced collective competence (CC)
Topic #96: book receiv includ practic work given radio instruct construct chapter principl treat modern time describ subject matter volum show design	Sharing instructional principles	CC2. Learning and sharing knowledge collectively	
Topic #47: station work amateur heard time citi mile relay ohio state coast denver distanc west record angel handl texa good district	Conducting radio relays	CC3. Developing a distinctive collective competence	

Lay-expertise legitimation (Croidieu & Kim 2018)

Second-order themes	# of topics	# of occurrences	Overall percentage of occurrences	Period 1							
				1899–1911	1912	1913	1914	1915	1916	1917	1918
Mechanism 1											
CC1. Forming a rudimentary knowledge	12	25	14%	17%	6%		14%	15%	25%	18%	40%
CC2. Learning and sharing knowledge collectively	7	14	8%	8%				8%	8%	18%	50%
CC3. Developing a distinctive collective competence	6	10	6%						8%	9%	10%

Other theoretical applications for topic modeling

- Determining clusters through similarity and coherence
- Developing inductive classification systems
- Detecting novelty and the emergence of new topics
- Making sense of online audiences

New trends in topic modeling

- Rendering corpora by using statistical natural language processing techniques (NLP) in pre-processing of data or analyzing non-English languages (e.g., Manning & Schuetze 1999)

- Rendering new theoretical artifacts by leveraging novel algorithms such as structural topic modeling (e.g., Schmiedel et al. 2018), plagiarism detection (Bail 2012) or hierarchical LDA (hLDA) (Blei et al. 2010)

DRUCKER SCHOOL of MANAGEMENT

CLAREMONT GRADUATE UNIVERSITY

Trends in Rendering with Analysis

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)*

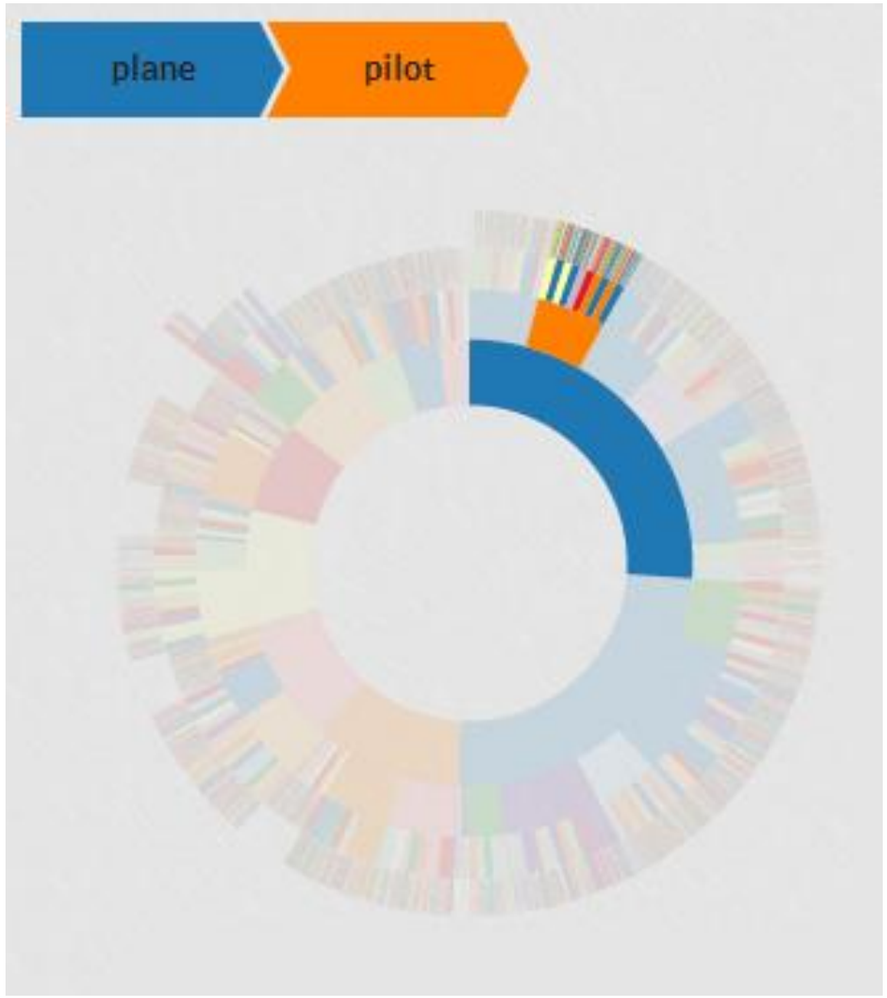


Figure 3: Our simple breadcrumb trail and contextual anchor offer constant context as the user explores the visualization. Highlighted slices within the contextual anchor are those currently displayed in the sunburst visualization.

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*

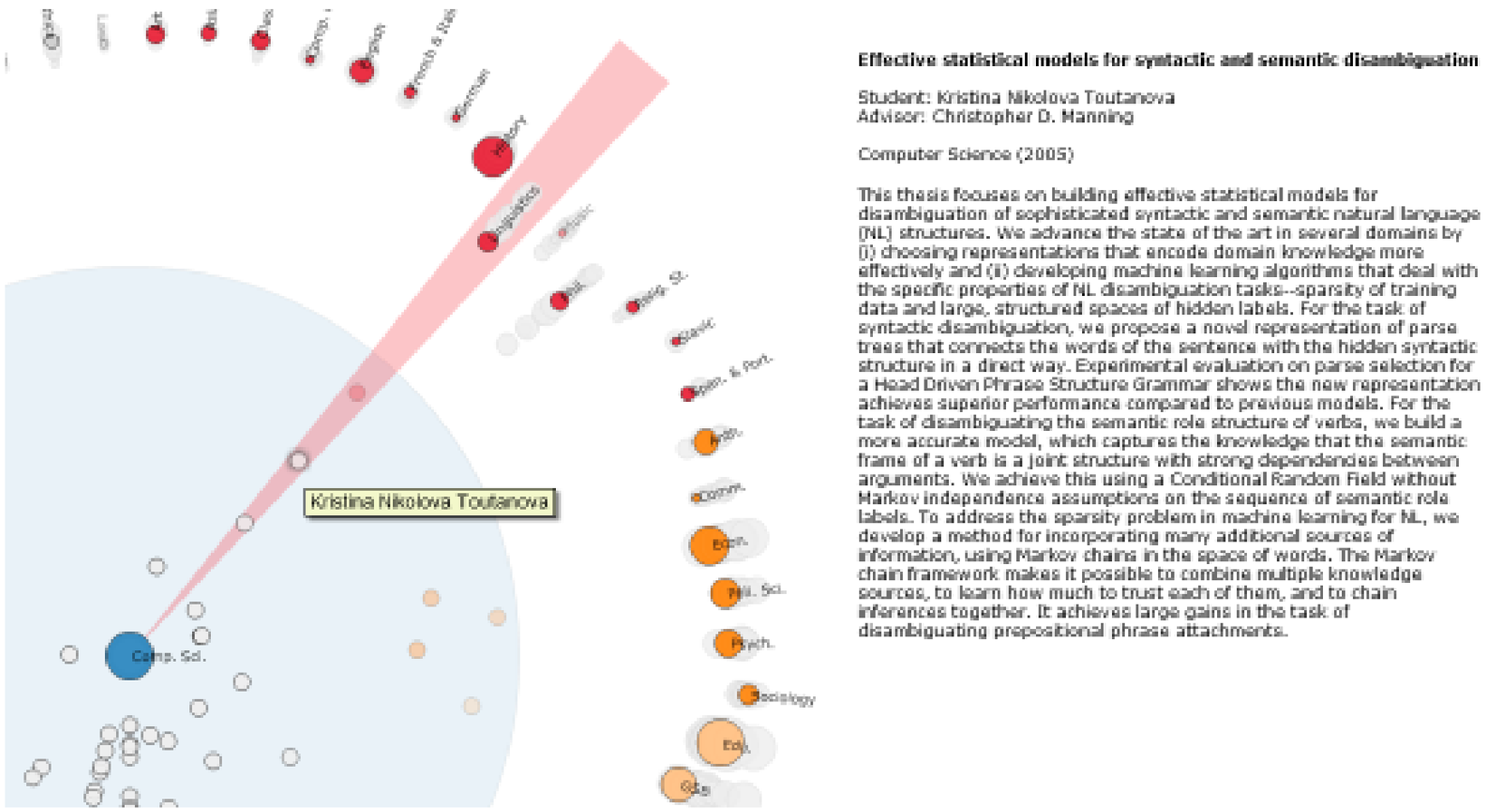


Figure 4. The Thesis View shows individual dissertations as small circles placed between the focus department and the next most similar department. Reading the original text of the dissertation enables experts to evaluate observed dept-dept similarities, and confirm the placement of three computational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

Trends in Rendering Corpora

Milo Wang

UNIVERSITY OF ALBERTA
ALBERTA SCHOOL OF BUSINESS

Morphological Difficulty

- In NLP, an English text can be regarded as a collection of **words**, which are the basic units, **separated by spaces**, in Topic Modeling.
- A Chinese syllable corresponds to a Chinese **character**; while some characters stand alone as an individual word, many words can consist of more than one character, with **no spaces** between words in a text.
  - For example: 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?)
  - ➢ **Solution**: **Segmentation**

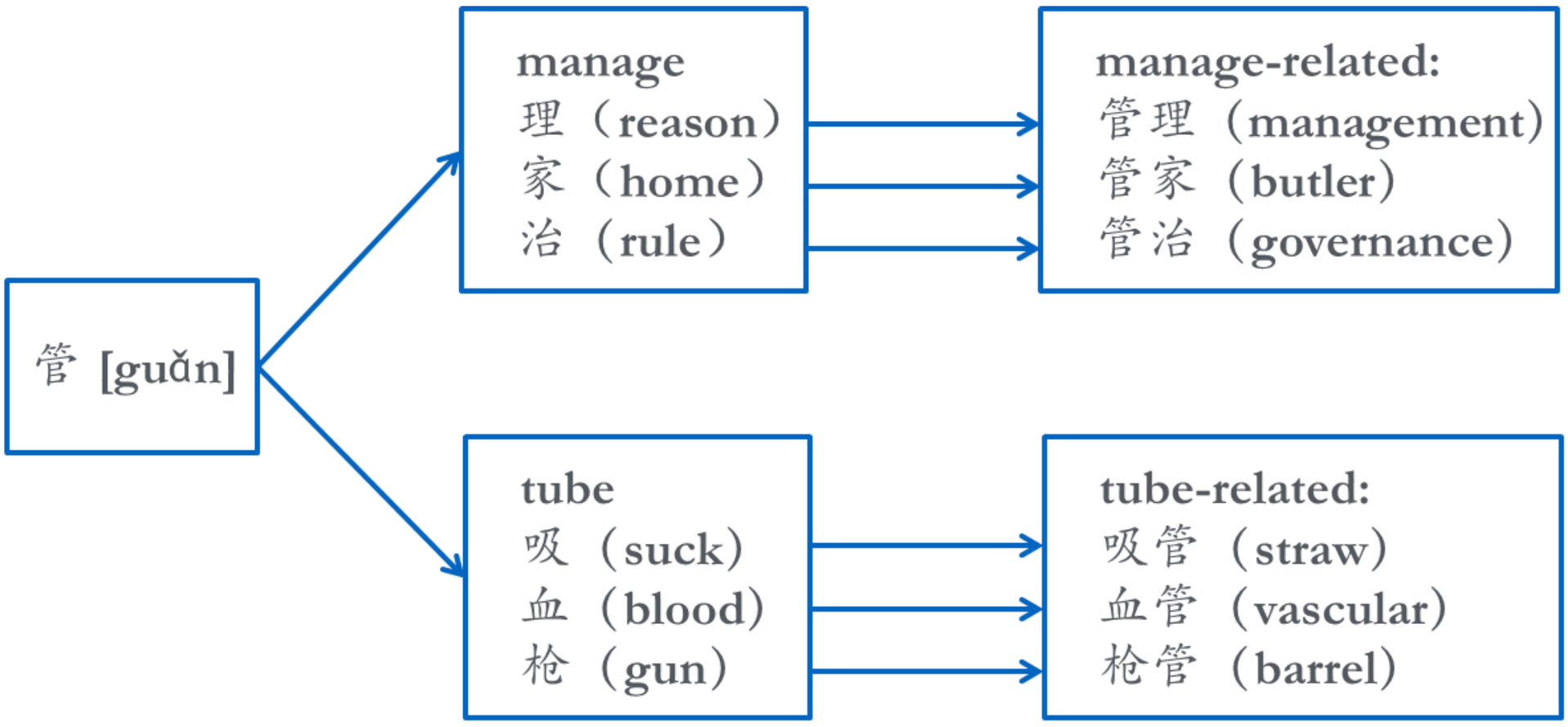
An Intuitive Choice: Word-based Segmentation

- **Word-based segmentation**
  - One word as a unit; just as in standard Topic Modeling of English text
  - **Limitations**:
    - ignore the fact that a Chinese word is composed of Chinese characters; words sharing the same character may have some semantic relations;
    - there are 20 thousand characters, constituting **123 thousand words**; and in the standard LDA model, all those words **unseen** in training data are assigned the **equal** (negligible) probability in a topic, regardless of their component characters having been **seen** and very **different** occurrence rates, such as **棒棒哒**, **城会玩**, **活久见**.

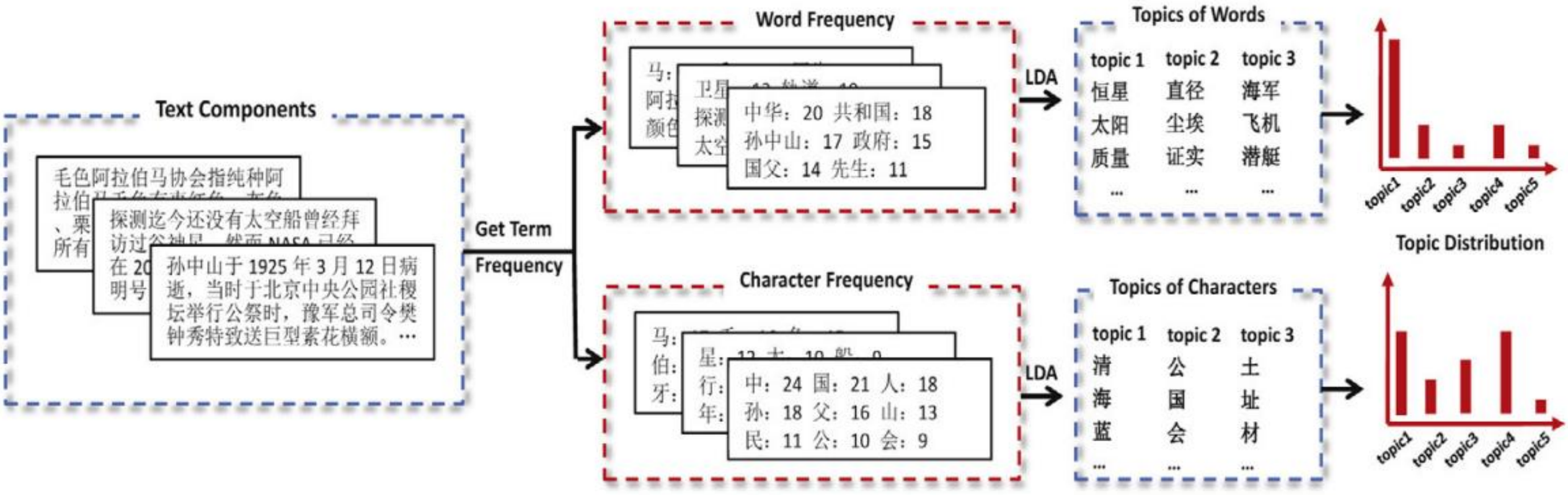
Character-based Segmentation

- **Character-based segmentation**
  - One character as a unit; more acceptable when combined with **local knowledge structure**
    - example: ancient Chinese text of criminal records (Miller, 2013)
  - **Efficient**: 99% of Chinese vocabulary (123 thousand words) can be reduced to **3500 characters**.
  - **Limitation**:
    - words in Chinese are highly **polysemous**; a considerable number of Chinese words have the semantic meaning **irrelevant** to the meanings of its component characters.

Using “guǎn” as the center character in words



Word-based vs. Character-based Segmentation

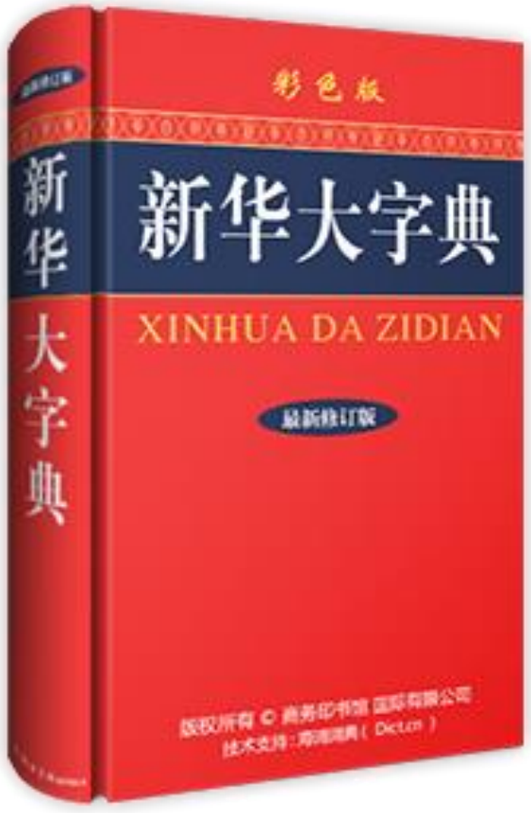


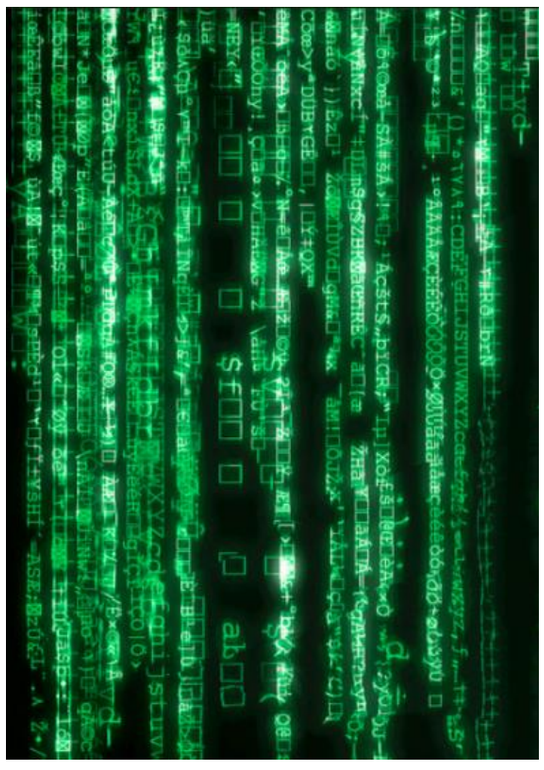
UNIVERSITY OF ALBERTA

Segmentation Methods

Two main categories:

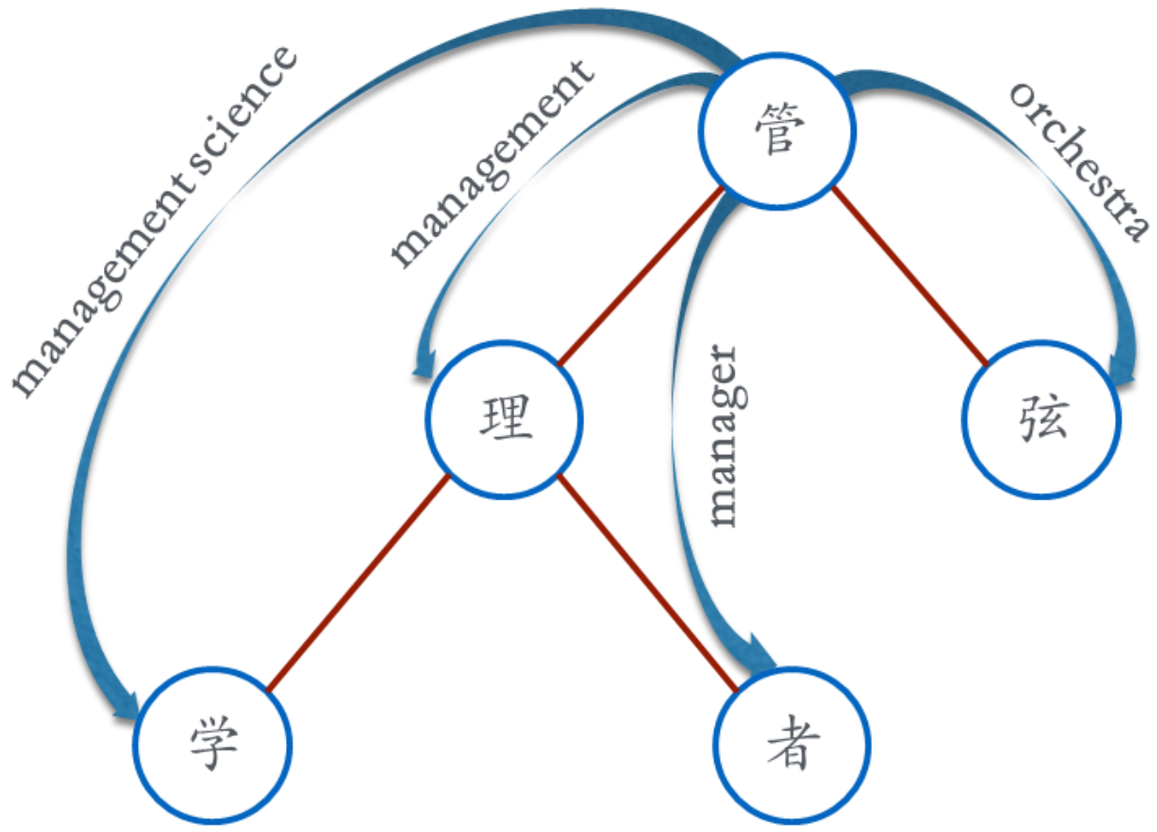
1. **Based on dictionaries and matching of character string**
2. **Based on statistics and machine learning (HMM, CRF, etc.)**

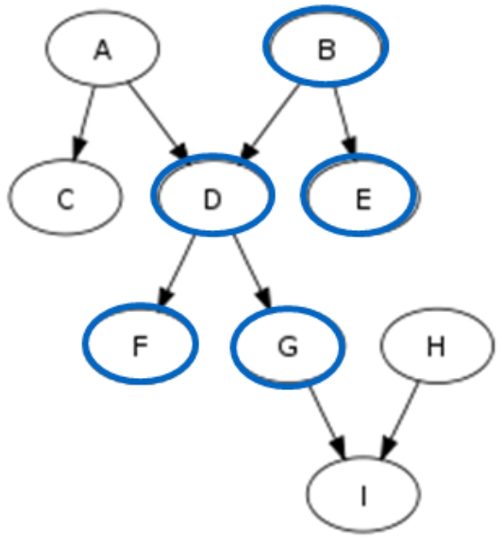




UNIVERSITY OF ALBERTA

Maximum matching segmentation model





From trie to directed acyclic graph (DAG)

A trie (digital tree) using *guǎn* as the center character

Maximum matching segmentation model

An example: **研究生命起源。**

Forward maximum matching: **研究生**/ **命**/ **起源**
graduate/ life, faith/ origin

Reverser maximum matching: **研究**/ **生命**/ **起源** ✓
study/ life/ origin

An example of **comparison algorithm**:

Non-recognized words:	0 (FMM) = 0 (RMM)
Single-character words:	1 (FMM) > 0 (RMM)
Total words:	3 (FMM) = 3 (RMM)
Total score (smaller, better):	4 (FMM) > 3 (RMM)

Conclusion: Reverser maximum matching is better.

UNIVERSITY OF ALBERTA

Maximum matching segmentation model

Two limitations:

1. **Ambiguity:** 1/169 of FMM and 1/245 of RMM are wrong.
2. **Out of vocabulary:**
    - Names: 爱德蒙·**唐泰斯** (Edmond Dantès)
    - Places: 热那亚 (Genoa)
    - Terminology: 线性回归 (linear regression)

➢ **Two solutions:**

1. **Customized dictionary/vocabulary**
2. **Statistical segmentation models**

Statistical segmentation model

Def: based on **language libraries** (that contain a large amount of segmented texts), using **statistical models** (machine learning) to learn the laws of segmentation (i.e., **training**) in order to segment **unknown** texts.

Mainstream models: Hidden Markov Model (HMM), Conditional Random Fields (CRF), N-gram, Maximum Entropy (ME)

UNIVERSITY OF ALBERTA

Hidden Markov model

Language library: e.g., *People's Daily*

Model: Hidden Markov Model (HMM) is based on four labels, B (Begin), M (Middle), E (End), S (Single). Given the observed set of words, the model learns to discover the hidden order (BMES) of each character in the string.

An example:	**I**	**study**		**marketization.**		
	我	**研**	**究**	**市**	**场**	**化。**
Hidden order:	/ S/	/B	E/	/B	M	E/

UNIVERSITY OF ALBERTA

Using segmentation packages

Good news: there are **quite a few** packages!!

Ansj, *Jieba*, *MMSEG*, *ICTCLAS*, *Rwordseg*, *THULAC*, *Standord*, *Hanlp*, *NLPIR*

Time

Precision

Recall

Using an R package: JiebaR

Four models of segmentation:

```
##  接受默认参数，建立分词引擎
mixseg = worker()

##  相当于
##       worker( type = "mix", dict = "somepath/dict/jieba.dict.utf8",
##               hmm  = "somepath/dict/hmm_model.utf8",  ### HMM模型数据
##               user = "somepath/dict/user.dict.utf8") ### 用户自定义词库

mixseg["江州市长江大桥参加了长江大桥的通车仪式"]  ### `<=` 或 `[` 为分词运算符

## 相当于 segment( "江州市长江大桥参加了长江大桥的通车仪式" , mixseg )
##或者 mixseg <= "江州市长江大桥参加了长江大桥的通车仪式"

[1] "江州"     "市长"     "江大桥"   "参加"     "了"       "长江大桥"
[7] "的"       "通车"     "仪式"
```

1. **MPSegment** = using Trie tree to construct a Directed Acyclic Graph (DAG) and using dynamic programming algorithm.
2. **HMMSegment** = using a Hidden Markov Model and viterbi algorithm to determine observed set of words. The default is based on *People's Daily* language library.
3. **MixSegment** = using both MP and HMM models to construct segmentation. The most effective model.
4. **QuerySegment** = using MixSegment to construct segmentation and then enumerates all the possible long words in the dictionary

UNIVERSITY OF ALBERTA

TOPIC MODELING: CAUTIONS AND OPPORTUNITIE

KEYVAN VAKILI
LONDON BUSINESS SCHOOL

AOM 2017

When to use topic modeling?

- In topic modeling:
  - Each document is a bag of words
  - Words are analyzed on how they appear, letter-by-letter

- 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 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 pre-set categorization
    - supervised classification usually works better
    - Some room for supervised topic modeling

What can be considered a topic?

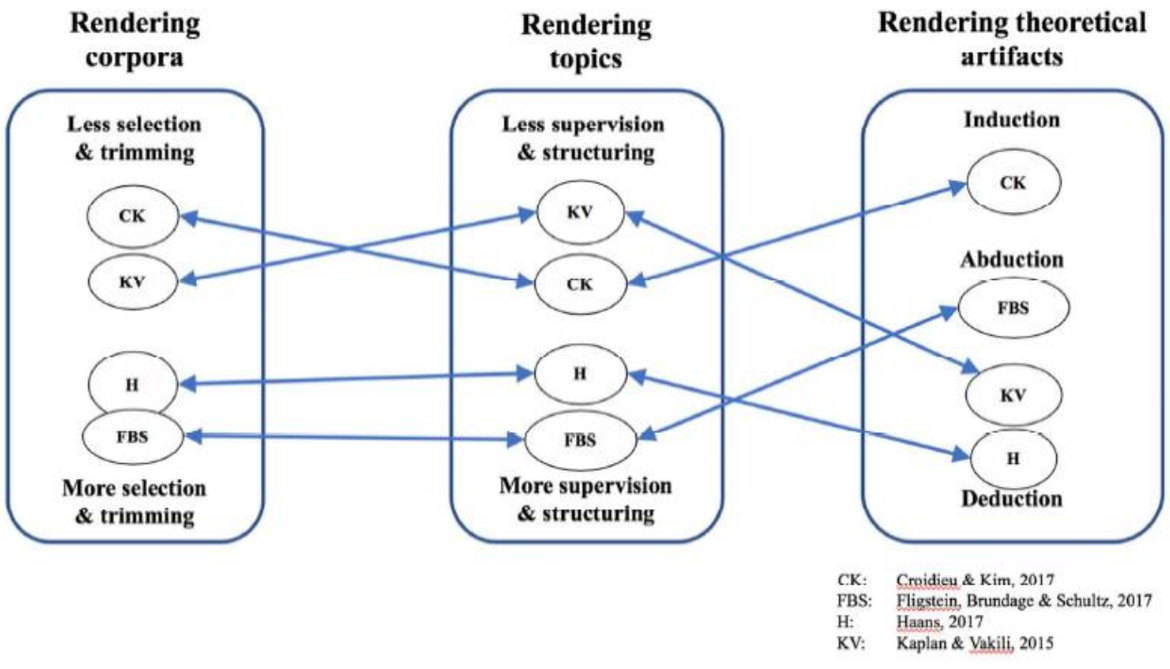
- Any language construct that can be signified with a set of words
- Cognitive frames (managerial, political, cultural, media, etc.)
- Themes in historical corpora
- Technological or scientific domains/paths
- Product/Market/Industry categories
- Attention direction

Avoid black-boxing the method

- ”Data have no value or meaning in isolation... They exist within a knowledge structure” (Borgman, 2015 p.4)
- Topic modeling is just part of the rendering process
- Explain all three steps of the rendering process clearly:
  - Rendering corpora
    Rendering topics
  - Rendering theoretical artifacts

Figure 4

Examples of Rendering Pathways in Management Research



Rendering corpora

- Explain the data collection process clearly:
  - Source and type of data
  - Which words are excluded (stop words)
  - Which texts are excluded/included
    - Shorter than a certain length?
    - Duplicates?
  - Stemming method?
  - Spelling errors?
  - Any other pre-processing

Rendering topics

- Explain how the main parameters are selected:
  - Number of topics
  - Topic smoothing and term smoothing parameters
- Justify the choice of algorithm
- Justify the use of topic modelling and the specific method used (LDA, hLDA, dynamic LDA,…)
- Explain the method briefly and clearly
- Show all the identified topics and their top terms preferably with term weightings (possibly in appendix)
- Show one or two representative pieces of text for each topic (possibly in appendix)
- Validate the results and do sensitivity analysis

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 in your topic modeling output
- 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.

Rendering theoretical artifacts

- Iterating between topics and theory to build theoretical artefacts based on
  - Attributes of topics (origin, prevalence, specificity, change over time, etc.)
  - Relationship between words and topics
  - Relationship between documents and topics
  - Relationship between topics themselves
  - Relationship between topics and other variables
  - …
- Explain clearly the path from topic modelling output to theoretical artefacts
  - It’s a process of simplification and structuring

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

From the Topic Modeling Group –

Thank You!

Please visit the site for new materials and posts:

https://github.com/RFJHaans/topicmodeling