

Topic Modeling in Management Research


Timothy R. Hannigan

Timothy (Tim) Hannigan

- BA (hons) in Computer Science and Economics
- MSc in Information Systems
- DPhil (PhD) in Management Research
- Associate Professor of Strategy and Organization at Telfer School of Management, University of Ottawa

Biography



- 
1. Discuss what topic modeling is all about
 2. Discuss how to use the rendering framework to govern how to effectively use topic modeling in practice
 3. Discuss two different topic modeling approaches (LDA and STM)
 4. Discuss best practices for publishing quality research using Topic Modeling

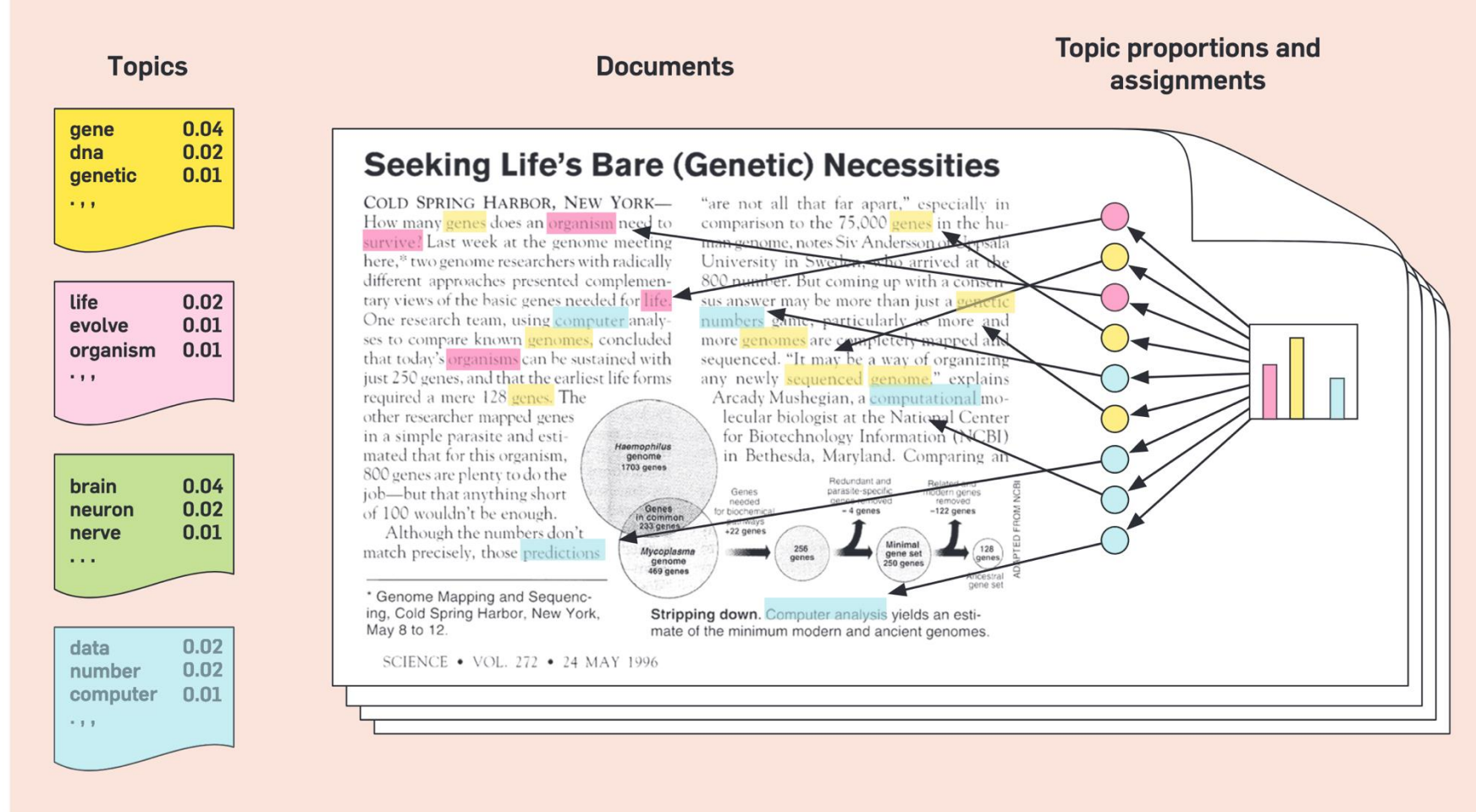
Agenda

What is topic modeling all about?

What is a topic model?

- In the abstract, a topic model is a computational model that generates coding categories in a corpus of texts/documents
- This is an extended form of computational content analysis
- *“The most distinctive feature of topic models is that they provide an automated procedure for coding the content of a corpus of texts (including very large corpora) into a set of substantively meaningful coding categories called “topics.” The algorithms can do this with a minimum of human intervention, and this makes the method more inductive than traditional approaches to text analysis in the social and human sciences.” (Mohr & Bogdanov, 2013)*

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of “topics,” which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



Main steps in generating a topic model

- 1) Rather than starting with pre-defined codes or categories of meaning (like those we generate when we start to hand-code a text), the researcher begins by **specifying the number of topics for the algorithm to find**
- 2) Based on this parameter (number of topics), the algorithm then fits a model based on this, then returns the probabilities of words being used in a topic (**a topic-word matrix**), as well as the distribution of those topics across the corpus of texts (**a topic-document matrix**)

Theoretical background

- meanings are relational (Saussure, 1959); the meanings that define a coherent topic of conversation are constructed from a set of word clusters
- a topic is then a constellation of words that tend to come up in a discussion
- this captures co-occurrences regardless of these words' embeddedness within other complexities of language—such as syntax, narrative, or location within the text
 - each document is treated as if it were a so-called “bag of words”

Background

- the simplest and most widely used model is Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003)
- this implies a generative process of how documents are written:
 - each document (text) within a corpus is viewed as a bag-of-words produced according to a mixture of themes that the author of the text intended to discuss
 - each theme (or topic) is a distribution over all observed words in the corpus (words that are strongly associated with the document's dominant topics have a higher chance of being selected and placed in the document bag)
 - this assumes that the author repeatedly picks a topic, then a word and places them in the bag until a document is complete

Basic LDA

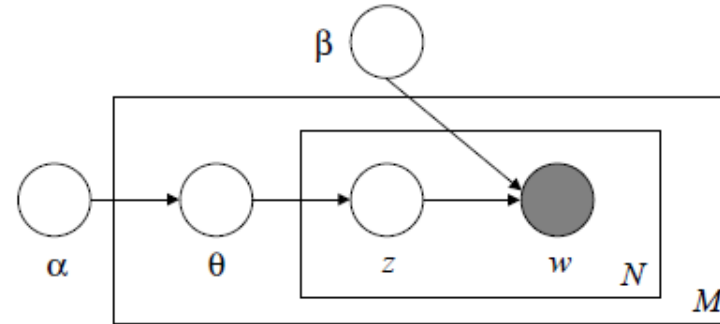
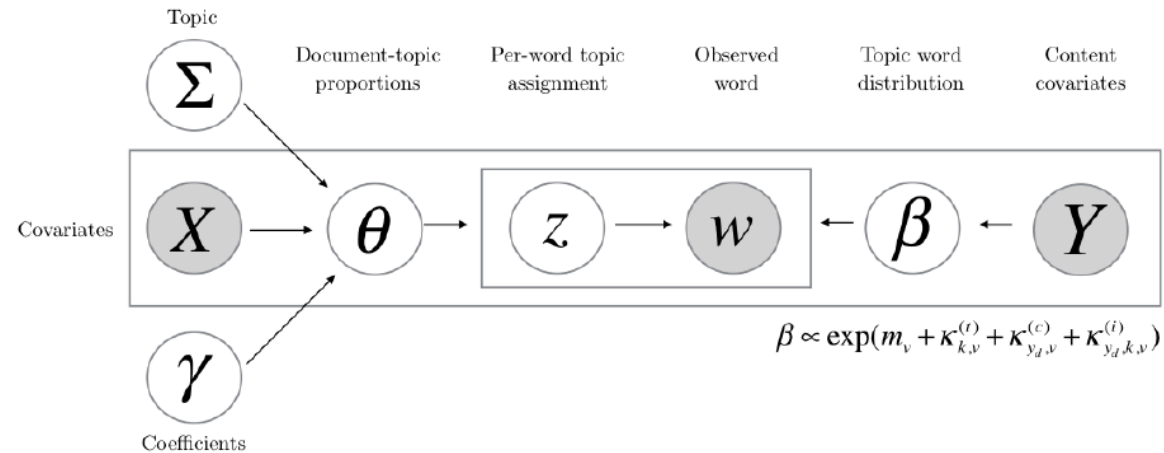


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

- The topic model learns only from the observed words co-occurrences in documents
- **Assumption:** identical generative processes behind texts in a corpora: documents are created based on drawing from a fixed set of topics—unchanging over time, independent of who generated the topics, etc

Structural Topic Models



- A recent innovation is the ability to incorporate information from metadata into the estimation of the topic-word distribution or the document-topic proportions
- This enables understanding how e.g. characteristics of the document producer or contextual factors shape the extent to which topics are used in documents

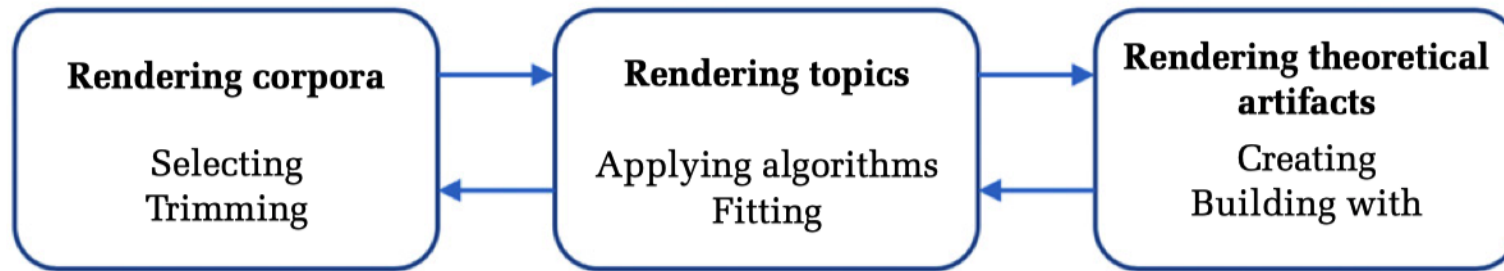
How do I use topic modeling in practice?

Topic modeling practices

- Topic modeling is a flexible method and can be paired with a number of different social science research designs (i.e. regression framework, grounded theory)
- To build theory with topic modeling, we need to consider three main practices:
 - (1) building a corpus of documents/texts
 - (2) using algorithms (such as LDA) to fit topics in a model
 - (3) using the model, to then create artifacts that can enable us to generate theoretical insights
- This can become unwieldy to stay on top of, particularly because surprises sometimes crop up in steps 2 or 3 that force you to revisit step 1

Rendering as a framework for organizing topic modeling practice

FIGURE 2
Topic Modeling Rendering in Theory-Building Spaces



- Hannigan et al., (2019) defined rendering in topic modeling as *a three-part 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, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*, 13(2), 586–632.

Rendering a corpus

Rendering corpora

Selecting
Trimming

- Guided by theoretical and empirical considerations, you need to select types of textual data
 - consider how a corpus is a meaningful collection of texts that were generated from the same set of meaning structures
- **Selection:**
 - need to account for language, authoring, and document sources
 - consider the logical fit with the research question of your study
 - also consider representativeness, levels of analysis, temporal considerations (e.g., longitudinal vs. cross-sectional data)
- **Trimming and cleaning**
 - important practices in rendering a corpus includes preprocessing and cleaning texts
 - the goal is to have your data in a **data frame** (i.e. a spreadsheet with column titles [id, document title, document body] , where every row and column has a value)
 - no best practices for this, but some guidance is available (Hickman et al., 2020 ORM)

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, 1094428120971683.

Rendering a corpus: key steps

- *[Take notes on your process]*
- Consider what your **data frame** must be
- Identifying moderate to large textual data that might yield insight into your Research Question
- Scraping those data from a variety of source – e.g., Factiva, Lexus-Nexus, etc
- *[Take notes on your process]*
- Pre-processing the scraped data block to clean it
- *[Take notes on your process]*
- Run analytics to further render the block
- Create **an actual data frame** – likely a .csv file – that capture the rows of observations and columns of attributes
- Double-check the data's structure and cleanliness in Python and/or pass it to R for checking

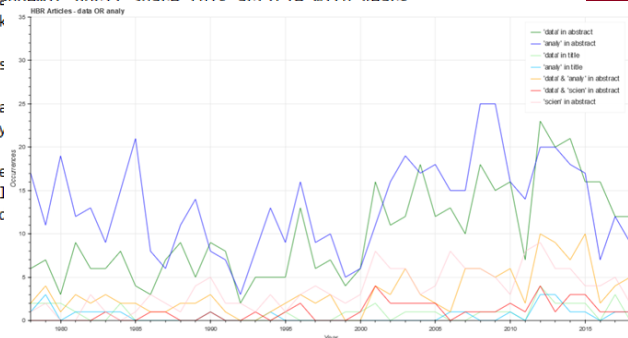
Rendering corpora

Selecting
Trimming

SOURCE:
Factiva, Twitter,
Dow Jones, etc.

Attributes: *ID, Journal, Volume, Number, Page range, Authors, Year, Title, DOI, Abstract, journal_code*

Merck Home Responsibility A day in the family during Hurricane Katrina Explore up as himself Merck communications execu d vaccines that may help millions of pe th impact Corporate responsibility Supp o Inc All rights reserved Forwardlookin ovel coronavirus disease COVID the imp report on Form K and the companys other mation you need to guide you on your he Access Program may be able to help pati gth inspire you From isolation to empowerment Stories from n is to deliver innovations that extend and improve the li verse Event or Product Quality Complaint with a specific M worth NJ USA the company includes forwardlooking statement: ates and internationally global trends toward health care t site No Duty to Update The information contained in this o combat the pandemic Annil Share this article with your so they can tak and tested for re of while hes ent counseling s may differ me ons The company anada where we help combat the bal health chal ll we have to c



Rendering a corpus: a data frame

Rendering corpora

Selecting
Trimming

Data Frame (Rules)

- Thinking about which algorithm and work through it as well
- The rules followed will curate an object with which you will work (rendering topic models)
- Often thinking in terms of blocks, which might represent matrix elements; but in data science take it apart by column as blocks (rather than looping across rows to columns). (Loop vs. column processing.)
- Note that the corpus refers to the unstructured data that will be inside the data frame via processing

	Periodical	Volume	Number	Page range	editor or organ	Year	Title	Reference typ	ID	DOI	Abstract	journal_c
0	The Academy of Management Annals	1	1	119-179	Adler, Paul S.	2007	Critical Management Studies	journalArtic	09TVAM2C	10.1080/07	Critical management studi	ama
1	The Academy of Management Annals	1	1	1-64	Dalton, Dan R.	2007	The Fundamental Agency Prok	journalArtic	TFKLJ4TP	10.1080/07	A central tenet of agency t	ama
2	The Academy of Management Annals	1	1	181-224	Elsbach, Kimbri	2007	The Physical Environment in O	journalArtic	ZPVK6MH2	10.1080/07	We review empirical resea	ama
3	The Academy of Management Annals	1	1	225-267	Kelman, Steven	2007	Public Administration and Org	journalArtic	AKCX6M8NC	10.1080/07	The study of public organiz	ama
4	The Academy of Management Annals	1	1	269-314	Edmondson, Am	2007	Three Perspectives on Team L	journalArtic	2PKFWRX3	10.1080/07	The emergence of a resear	ama
5	The Academy of Management Annals	1	1	315-386	Elfenbein, Hillan	2007	Emotion in Organizations	journalArtic	UFUFEUZE	10.1080/07	Emotion has become one o	ama
6	The Academy of Management Annals	1	1	387-438	Gilmartin, Matti	2007	Leadership Research in Healt	journalArtic	FCYN7BW6	10.1080/07	This chapter's purpose is to	ama
7	The Academy of Management Annals	1	1	439-477	George, Jennifer	2007	Creativity in Organizations	journalArtic	EB4EZVWC	10.1080/07	In this chapter, I review coi	ama
8	The Academy of Management Annals	1	1	479-511	Inkpen, Andrew	2007	Learning and Strategic Allianc	journalArtic	B4LN754Y	10.1080/07	Various researchers have s	ama
9	The Academy of Management Annals	1	1	513-547	Berchicci, Luca	2007	Postcards from the Edge	journalArtic	F8FQ7D92	10.1080/07	Environmental issues, while	ama
10	The Academy of Management Annals	1	1	549-615	MacDuffie, John	2007	HRM and Distributed Work	journalArtic	VY58WWS5	10.1080/07	The phenomenon of manag	ama
11	The Academy of Management Annals	1	1	617-650	Roberson, Lorian	2007	When Group Identities Matte	journalArtic	7KIUZXQ6	10.1080/07	Performance appraisals ari	ama
12	The Academy of Management Annals	1	1	65-117	Ashford, Susan J.	2007	Old Assumptions, New Work	journalArtic	AH448T76	10.1080/07	We review the literature o	ama
13	The Academy of Management Annals	2	1	133-165	Cascio, Wayne F.	2008	Staffing Twenty-first-century	journalArtic	F2884JFN	10.1080/15	We highlight important difi	ama

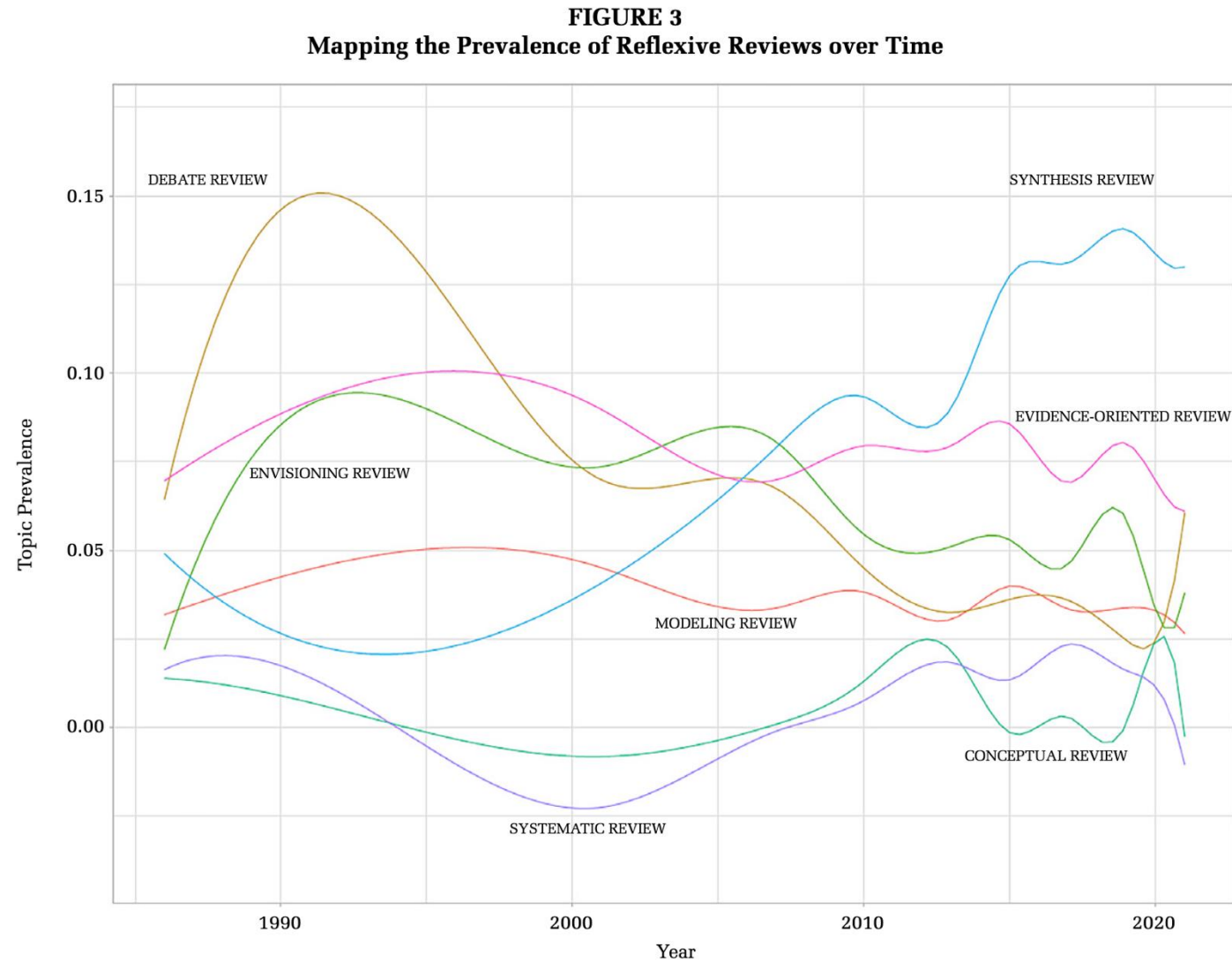
- Example of a Data Frame & Corpus

Data Frame from 2025 Annals Paper

	Periodical	Volume	Number	Page range	editor or organ	Year	Title	reference ty	ID	DOI	Abstract	journal_code	Abstract_processed	
0	The Academy of Management Annals	1	1	119-179	Adler, Paul S.	2007	Critical Management Studies	journalArtic	D9TVAM2C	10.1080/07	Critical management studi	ama	critical management studies cms offe	
1	The Academy of Management Annals	1	1	1-64	Dalton, Dan R.	2007	The Fundamental Agency Prot	journalArtic	TFKLJ4TP	10.1080/07	A central tenet of agency t	ama	a central tenet of agency theory is thi	
2	The Academy of Management Annals	1	1	181-224	Elsbach, Kimberl	2007	The Physical Environment in O	journalArtic	ZPVK6MH2	10.1080/07	We review empirical rese	ama	we review empirical research on the	
3	The Academy of Management Annals	1	1	225-267	Kelman, Steven	2007	Public Administration and Org	journalArtic	AKCXM8NC	10.1080/07	The study of public organiz	ama	the study of public organizations has v	
4	The Academy of Management Annals	1	1	269-314	Edmondson, Am	2007	Three Perspectives on Team L	journalArtic	2PXFWRX3	10.1080/07	The emergence of a resear	ama	the emergence of a research literatur	
5	The Academy of Management Annals	1	1	315-386	Elfenbein, Hillary	2007	Emotion in Organizations	journalArtic	UFUEFUZE	10.1080/07	Emotion has become one o	ama	emotion has become one of the most	
6	The Academy of Management Annals	1	1	387-438	Gilmartin, Matti	2007	Leadership Research in Health	journalArtic	FCYN7BW6	10.1080/07	This chapter's purpose is to	ama	this chapter purpose is to advance lea	
7	The Academy of Management Annals	1	1	439-477	George, Jennifer	2007	Creativity in Organizations	journalArtic	EB4E2VWC	10.1080/07	In this chapter, I review co	ama	in this chapter i review contemporary	
8	The Academy of Management Annals	1	1	479-511	Inkpen, Andrew	2007	Learning and Strategic Allianc	journalArtic	B4LM754Y	10.1080/07	Various researchers have s	ama	various researchers have suggested tl	
9	The Academy of Management Annals	1	1	513-547	Berchicci, Luca	2007	Postcards from the Edge	journalArtic	F8FQ7D9Z	10.1080/07	Environmental issues, whil	ama	environmental issues while of growing	
10	The Academy of Management Annals	1	1	549-615	MacDuffie, John	2007	HRM and Distributed Work	journalArtic	VYSBWWSV	10.1080/07	The phenomenon of manag	ama	the phenomenon of managing work t	
11	The Academy of Management Annals	1	1	617-650	Roberson, Loria	2007	When Group Identities Matter	journalArtic	7KJUXZQ6	10.1080/07	Performance appraisals ar	ama	performance appraisals are a critical	
12	The Academy of Management Annals	1	1	65-117	Ashford, Susan J.	2007	Old Assumptions, New Work	journalArtic	AH44BT76	10.1080/07	We review the literature o	ama	we review the literature on nonstand	
13	The Academy of Management Annals	2	1	133-165	Cascio, Wayne F.	2008	Staffing Twenty-first-century C	journalArtic	FZ884JFN	10.1080/19	We highlight important dif	ama	we highlight important differences be	

Krlev, G., Hannigan, T., & Spicer, A. (2025). What Makes a Good Review Article? Empirical Evidence From Management and Organization Research. *Academy of Management Annals*, annals.2021.0051. <https://doi.org/10.5465/annals.2021.0051>

Using a data frame columns in STM



Krlev, G., Hannigan, T., & Spicer, A. (2025). What Makes a Good Review Article? Empirical Evidence From Management and Organization Research. *Academy of Management Annals*, annals.2021.0051. <https://doi.org/10.5465/annals.2021.0051>

Rendering a corpus: Pre-Processing in R vs. Python

Rendering corpora

Selecting
Trimming

- Much of pre-processing work stems from data science, and many data scientists are multi-lingual
- Regardless of what programming language you use (or if LLM), the goal is a tidy* data frame
- You could do it all manually to create the frame (e.g., an Excel .csv file)
- The libraries in Python and R give a lot of affordances – and hard to see without looking into them
- Beneath R is the “tidyverse”*
- Pandas are in Python, designed to replicate the tidyverse

```
1. Rendering Corpus.ipynb
1A. Scraping Tweets.ipynb
2. STM_rendering_topics.RData
2. STM_rendering_topics.Rmd
3. Rendering Artifacts.ipynb
4. Preparation for Content Analysis.ipynb
5. Content Analysis.ipynb
6. Event History Analysis.nb.html
6. Event History Analysis.Rmd
```



**Tidy (vs messy) data – every column is a variable, every row is a single case, and every cell is single value.*

Rendering a corpus: covariates in Structural topic modeling (STM)

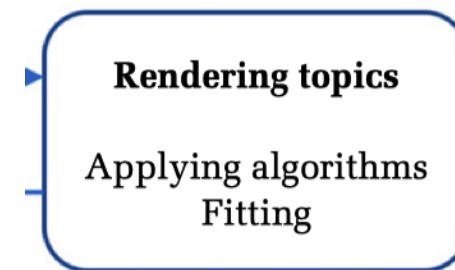
Rendering corpora

Selecting
Trimming

- It is worth noting at this point that you are rendering a corpus with an idea of what algorithm you will be using to render the topics
- In Structural Topic Modeling (STM) The columns in the data frame can become covariates, as we see with sources and time; these will need to be interpreted and then coded. (These are “transformed” within corpus data.. For example, time needs to be transformed into an integer representing number of units from a date of origin, can be days, months, years)
- There are data that you could collect in the corpus that could be used for covariate analysis. (These are matched and merged in covariates)
- STM also produces meta-data based on the columns (these are STM internal ones)
- Covariates should be created within the context of the project and creating the corpus. Please try to avoid just throwing in covariates without inspecting them and thinking about their linkage to the corpus

Lindstedt, N. C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017. *Social Currents*, 6(4), 307–318. <https://doi.org/10.1177/2329496519846505>

Rendering topics



- Guided by affordances of different topic modeling algorithms (i.e. LDA, or STM) you apply an algorithm to generate a topic model (i.e. identify appropriate topics for analysis)
- Applying algorithms:
 - first, decide on an algorithm, then use your data frame; each algorithm provides you a pre-programmed set of rules; in effect this reduces the dimensionality of your data frame
 - LDA is most popular and is straightforward; does not work well on short texts
 - STM brings in covariates (i.e. source name, or date), which enables you to track topic prevalence as a smoothed curve over time
 - BTM is appropriate for topic modeling tweets
- Fitting
 - fitting a model is most often about determining the optimal number of topics to use
 - one view of fit is based on a logic of accuracy and using statistical measures to diagnose possibilities (e.g. perplexity, semantic coherence, exclusivity); these are built into algorithms
 - you can generate a number of different topic model specifications within a range (i.e. in steps of 5 between 50 and 75, so: one model based on 50, another on 55, another on 60, and so on...)
 - another is based on interpretability: DiMaggio et al. (2013) identified two key forms of validity: semantic or internal validity, and predictive or external validity

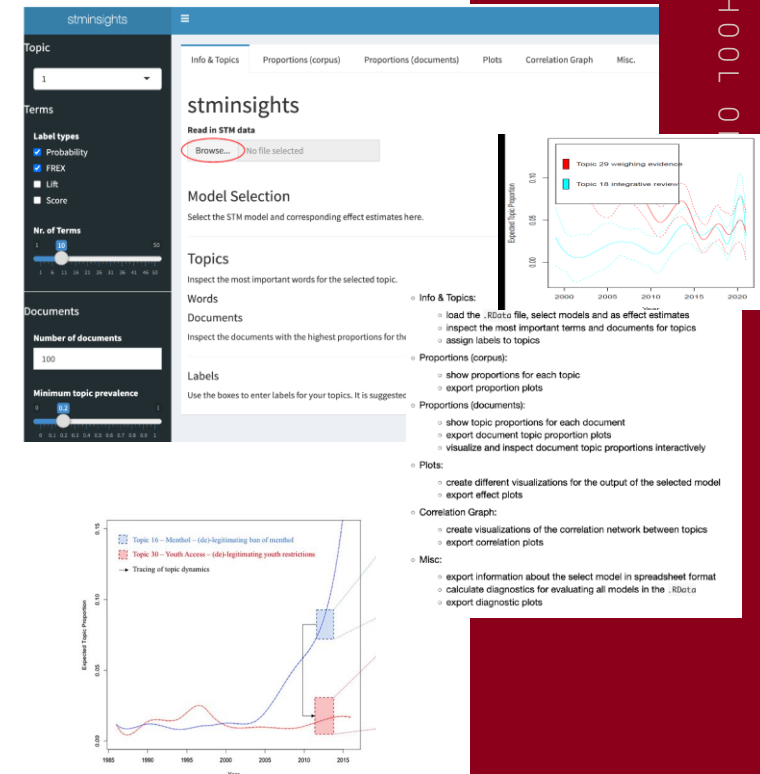
Rendering topics: key steps (STM version)

- Read in the framed data (.csv file) from Python or from another external source into R
- Do some pre-processing of (e.g., in R) to clean and organize last bits of data.
- Run **STM algorithm within R** with variety of topic settings, looking for coherence scores and exclusivity score balance to determine the topic #
- Settle on the topic number and then move to topic interpretation via features such as **STM Insights** and **LDavis**. Going back a step may be necessary
 - This requires reading top words and also checking the top loading documents with those words
- Decide on stable topic set and interpretation.
- Now move to consider what else those documents tell us about mechanisms co-occurring with those words
 - Might be temporal, or actor-based covariates
 - Code those
 - Run topics by covariates over time. Interpret the meaning.

You may also want to go further and then dig into the document data with a qualitative analysis program. e.g., using Atlas.TI

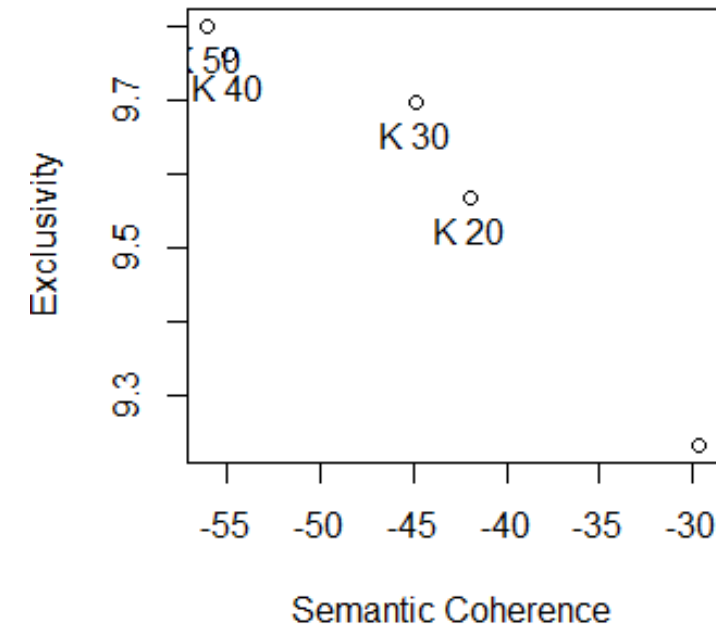
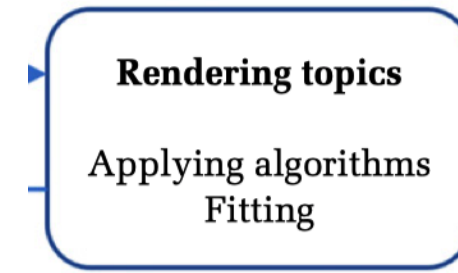
Rendering topics

Applying algorithms
Fitting



Rendering topics: key steps (STM version)

- An important tip is to use both the logic of accuracy and the logic of interpretability; i) use diagnostics to find a subset of topic model specifications, ii) use interpretability to make the final decision, iii) validate, validate, validate ! (Stewart et al., 2022)
- when finding a subset of topic specifications to zoom in on is to use the concept of the exclusivity-coherence frontier (Roberts et al., 2015)
- In this, you render a number of different topic model specifications, then use STM to graph the statistic diagnostics of **exclusivity** and **semantic coherence**
- Based on this, look for the model(s) that jointly maximize (i.e. the far top-right-hand-corner)
- For the subset of models, closely examine artifacts produced by each to validate and make a decision about which is most interpretable given your domain expertise (and RQ)

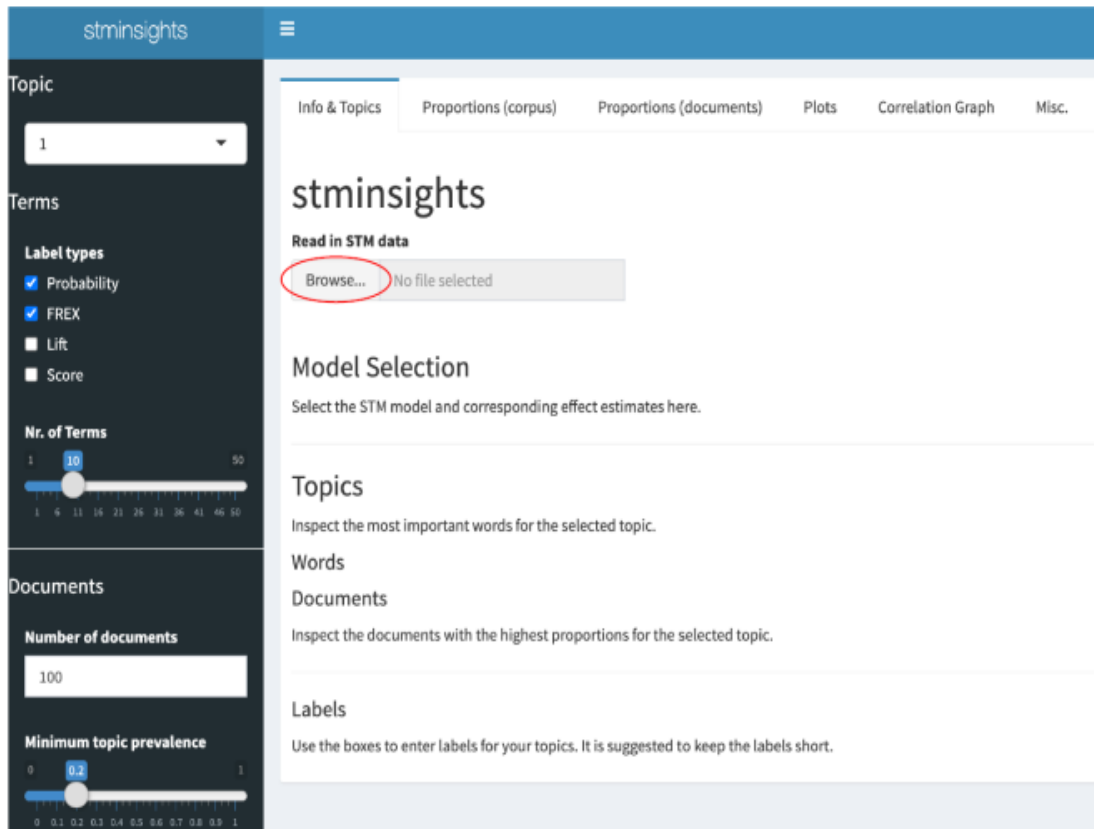


Rendering topics: key steps (STM version)

- Within STM is a tool called STM Insights; this can be used to validate different model specifications

Rendering topics

Applying algorithms
Fitting



◦ Info & Topics:

- load the `.RData` file, select models and as effect estimates
- inspect the most important terms and documents for topics
- assign labels to topics

◦ Proportions (corpus):

- show proportions for each topic
- export proportion plots

◦ Proportions (documents):

- show topic proportions for each document
- export document topic proportion plots
- visualize and inspect document topic proportions interactively

◦ Plots:

- create different visualizations for the output of the selected model
- export effect plots

◦ Correlation Graph:

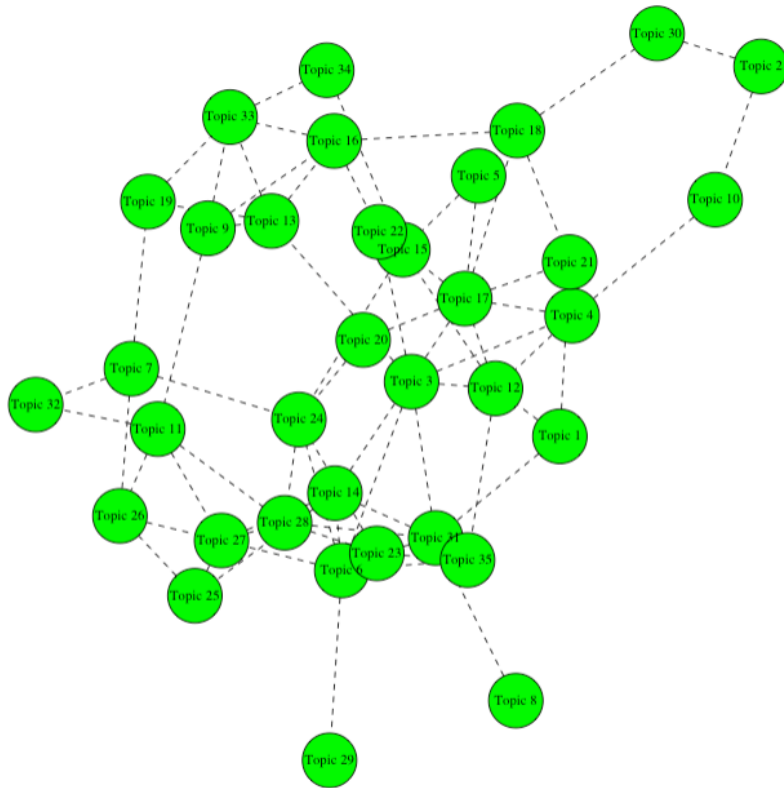
- create visualizations of the correlation network between topics
- export correlation plots

◦ Misc:

- export information about the select model in spreadsheet format
- calculate diagnostics for evaluating all models in the `.RData`
- export diagnostic plots

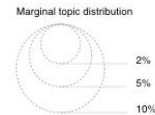
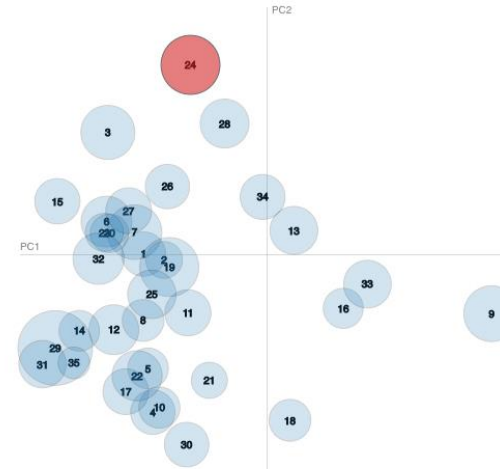
Rendering topics: key steps (STM version)

- STM easily enables use to generate Topic Maps to help validate– Correlational Network and LDAvis' Topic MDS



Selected Topic: 24 Previous Topic Next Topic Clear Topic

Intertopic Distance Map (via multidimensional scaling)

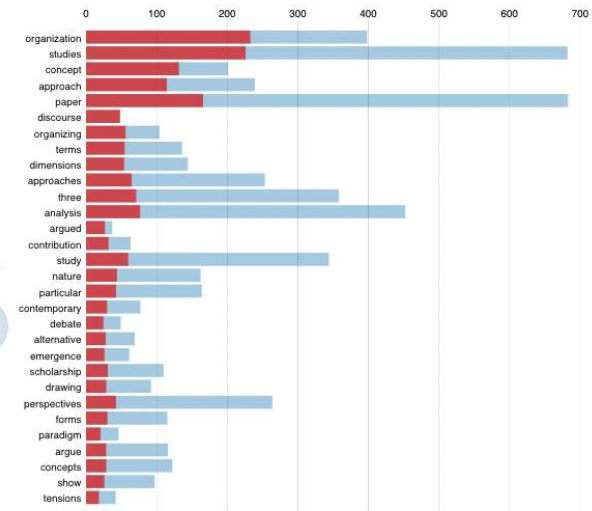


Slide to adjust relevance metric:⁽²⁾

$\lambda = 0.6$

0.0 0.2 0.4 0.6 0.8 1.0

Top-30 Most Relevant Terms for Topic 24 (4.5% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

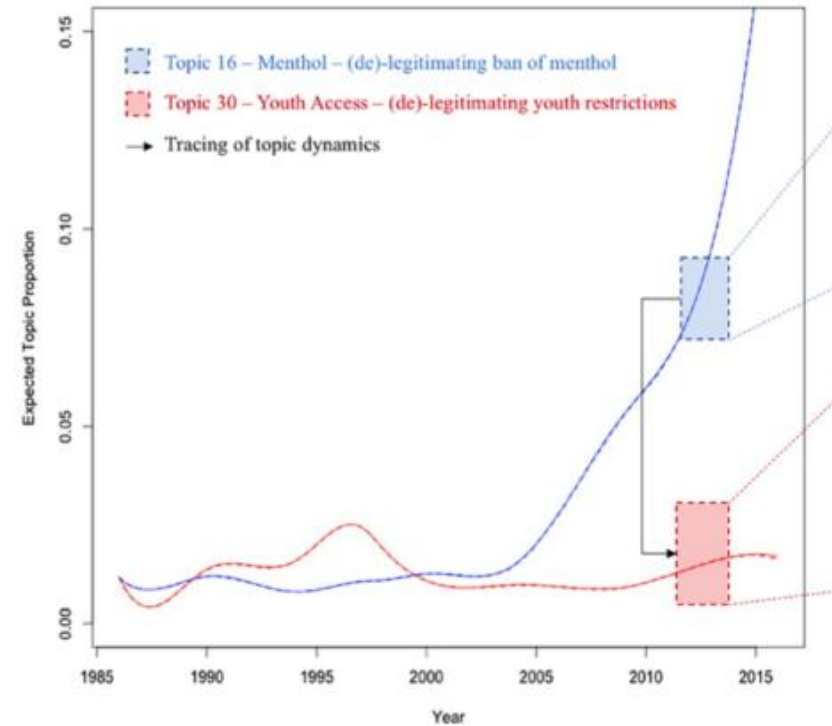
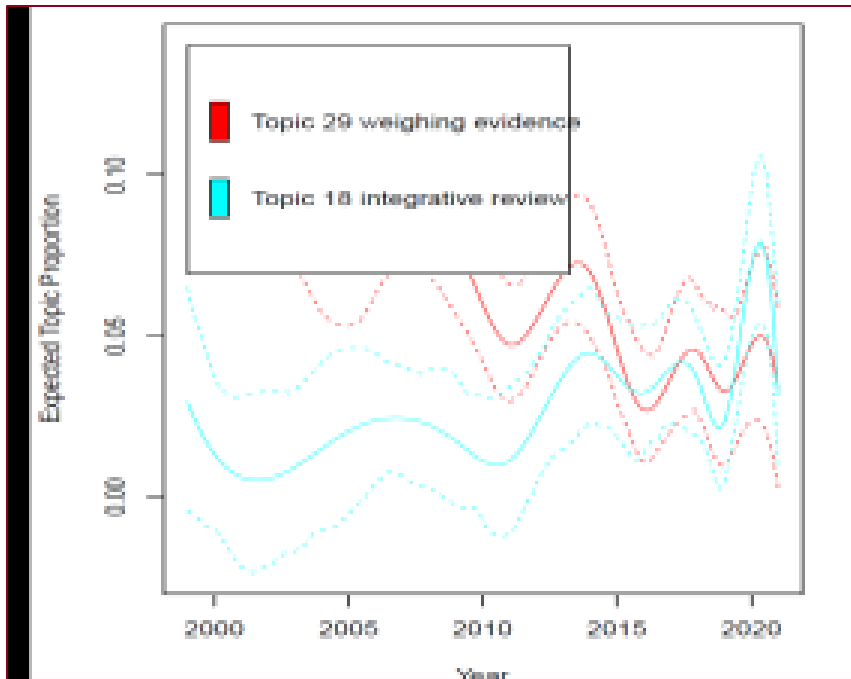
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

Rendering topics: key steps (STM version)

- STM also renders prevalence plots which we can use to validate

Rendering topics

Applying algorithms
Fitting



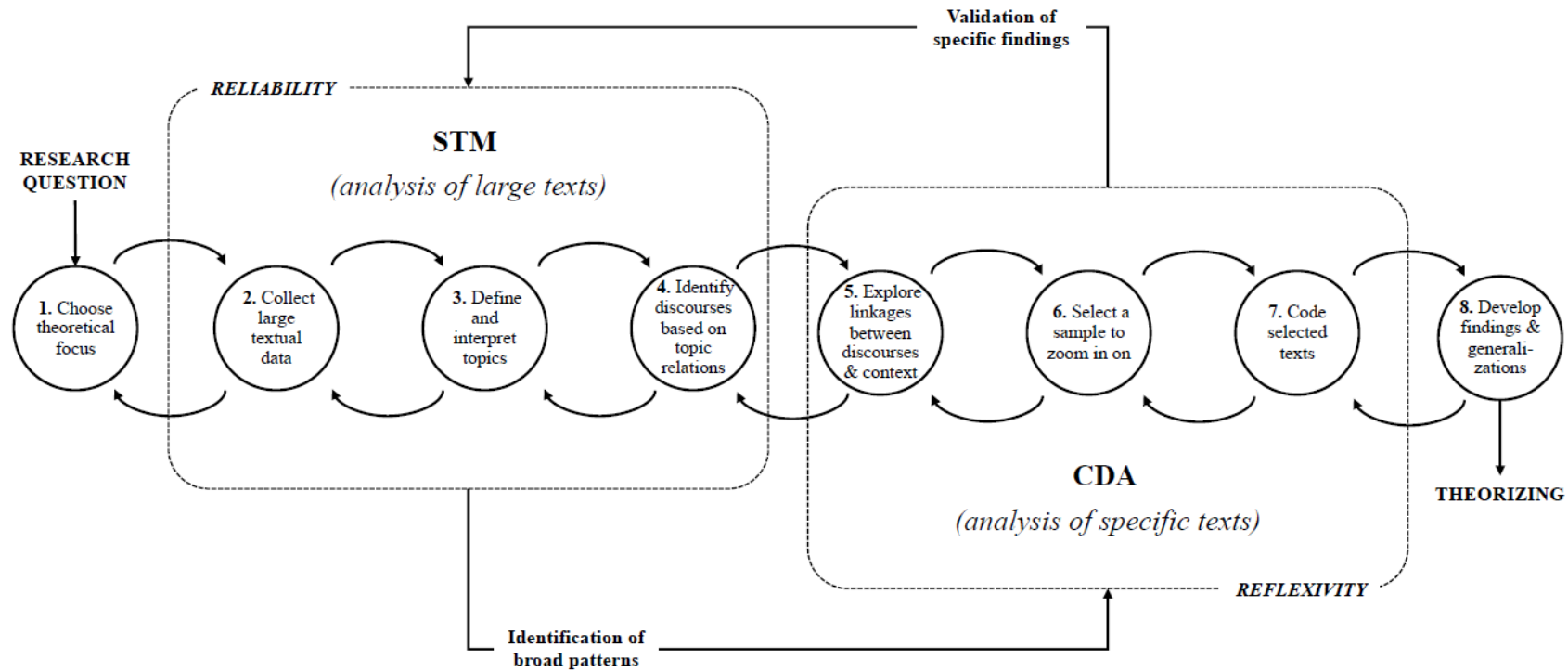
Rendering theoretical artifacts

Rendering theoretical artifacts

Creating
Building with

- You need to iterate between theory and the topics that emerge from your shortlist of model specifications to create new theoretical artifacts
- Once you settle on the specification that fits best, you can use these artifacts to build theory (Whetten, 1989)
- In LDA and STM: the document-word matrix and topic-document offer a wide range of opportunities for the researcher to build artifacts
 - You can export these as CSV files to Excel
- STM offers additional affordances for constructing artifacts based on your topic model
- This is also where you can integrate topic modeling into more traditional research designs:
 - For example, Croidieu & Kim (2018) used the topic-word matrix as “Illustrative topics vocabularies” in an axial coding (Grounded Theory) analysis; based on these, they derived First-Order concepts
 - Aranda et al. (2021) combined STM with Critical Discourse Analysis (CDA)

A STEPWISE MODEL TO INTEGRATE CDA WITH STM



Aranda, A. M., Sele, K., Etchanchu, H., Guyt, J. Y., & Vaara, E. (2021). From big data to rich theory: Integrating critical discourse analysis with structural topic modeling. *European Management Review*, 18(3), 197-214.

EMPIRICAL ILLUSTRATION: THE US TOBACCO INDUSTRY 1986-2016

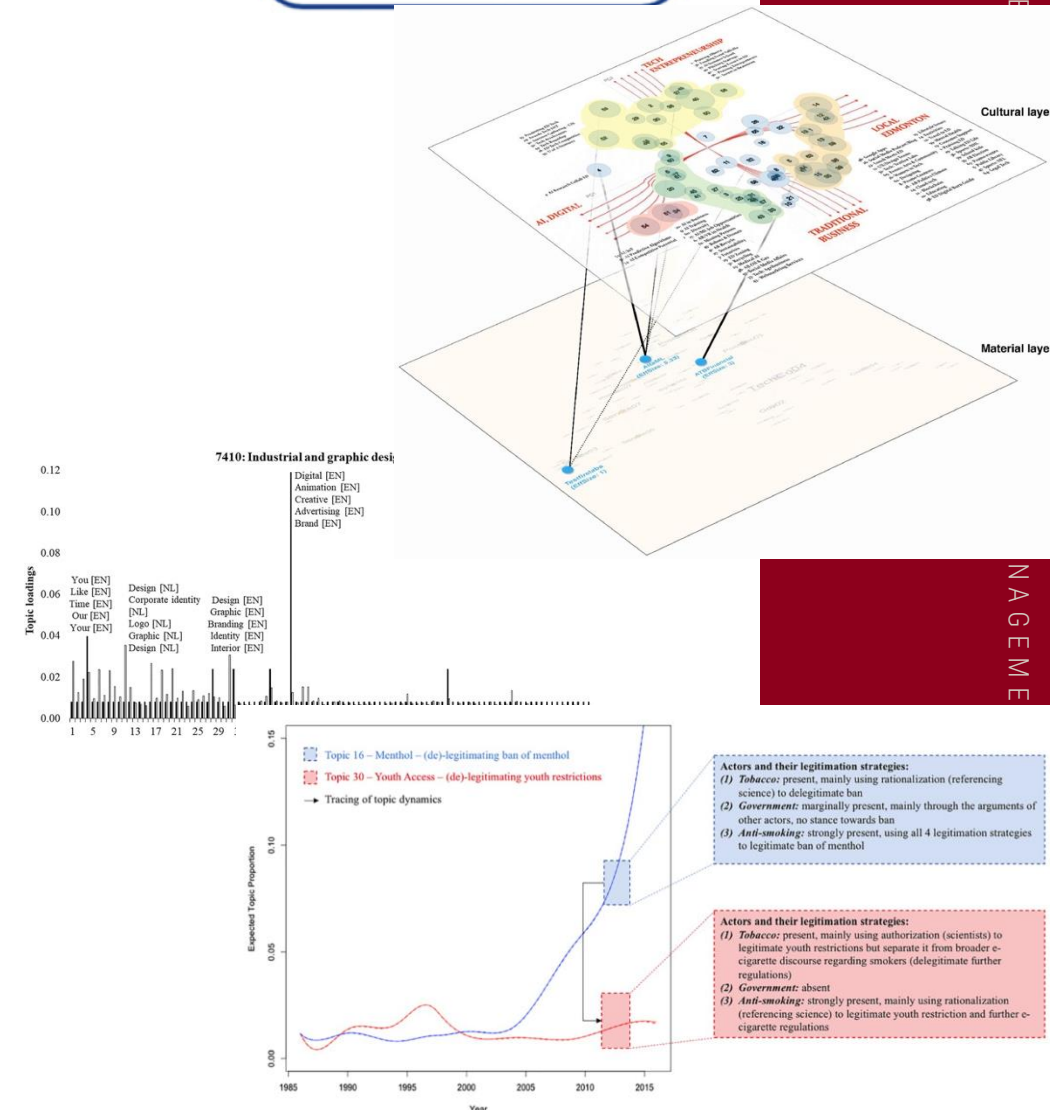
	Steps	CDA	STM
1	Choose theoretical focus	Actor-based approach focused on discursive legitimation struggles	
2	Collect large textual data	<i>New York Times (NYT) newspaper articles between 1986 and 2016</i>	Data cleaning and preparation of the corpus
3	Define and interpret topics	<i>Inductively derive detailed and distinct labels for each of the topics</i>	Grid search over a feasible range of topics – 43 topics solution
4	Identify discourses based on topic relations	<i>Identify four main discourses: health, marketing, legal, regulatory.</i>	Network graph reveals clusters of topics
5	Explore linkages between discourses and context	Link actors to discourses: Anti-smoking groups - health. Government and tobacco industry - legal. All - regulatory and marketing	<i>Evolution of topic proportions over time to identify milestones or key moments of interest (e.g., 1998 MSA).</i>
6	Select a sample to zoom in on	Focus on the menthol in cigarettes increase in the last decade	<i>Identify specific relevant newspaper articles on menthol published in the last decade</i>
7	Code selected texts	Discursive legitimation strategies (Van Leeuwen, 2007)	<i>Codes from metadata (i.e., actor, time)</i>
8	Develop findings and generalizations	Integrate actors' legitimation strategies and time dynamics	

Aranda, A. M., Sele, K., Etchanchu, H., Guyt, J. Y., & Vaara, E. (2021). From big data to rich theory: Integrating critical discourse analysis with structural topic modeling. *European Management Review*, 18(3), 197-214.

Rendering theoretical artifacts: key steps

- The theoretical artifacts can be derived from the maps that are created by STM Insights, such as network analysis (see Hannigan et al. 2021) – but only if match ones constructs and lead to key processes
- As Laura Nelson (2019) says, “don’t just give me a map”... we need to be thoughtful about what can be uncovered as a sets of mechanisms or processes via topic models
- There may be the creation of a more refined measure (e.g., of differentiation in Haans et al., 2019; or novelty in Kaplan & Vakili, 2015)
- There may also be a more textured interpretation of what actors are doing to create fames, as we see in the example from Aranda et al., (2021).

Rendering theoretical artifacts
Creating
Building with



How do I publish topic modeling papers in quality journals?

Publishing with Topic Modeling

- It is important to frame the topic modeling technique properly within a research design
 - Are you using it to facilitate Grounded Theory?
 - Are you using it to generate an additional IV to help with your regression model?
- This is a semi-supervised machine learning technique for conducting computer-aided content analysis– to use it with excellence, you need to show transparency using the right levers
 - Use the rendering framework (Hannigan et al., 2019) to organize your work
 - Consider key questions such as:
 - Is your corpus representative of the data that matches your RQ?
 - Does the chosen algorithm match the data? (i.e. LDA with Tweets?)
 - How did you decide on your particular model specification? Can you show a combination of artifacts to document your process?
 - How did you build upon the theoretical artifacts that your topic model generated?

Publishing with Topic Modeling

- Topic Modeling is not a panacea that is magic – or shortcuts – for doing research
- There are a number of different ways to use it
- In Hannigan et al., (2019) we found topic modeling that has enhanced management theory in five subject areas:
 - *detecting novelty and emergence*
 - *developing inductive classification systems*
 - *understanding online audiences and markets*
 - *analyzing frames and social movements*
 - *understanding cultural dynamics*
- There are likely new ways that it has been used since then as well

Publishing with Topic Modeling

- Most often, I see (as a reviewer and editor) issues with topic modeling in manuscripts appear around:
 1. Too much Black boxing and not enough showing of the technique application
 2. Insufficient explanation behind analytic choices (i.e. just saying “we followed best practices of 100 topics” is not enough)
 3. Poor justification of topic specification (# of topics)
 4. Poor fit with research design (i.e. sufficient space dedicated to discussing topic modeling, but not really integrating it into the research design)
 5. Not providing enough artifacts to help the reader along- use appendices!
 6. Glossing over fundamental tensions in topic modeling; for example, there is no perfect number of topics

What's next?

Topic Modeling Is An Interpretive Data Science

- LDA and other implementations of topic modeling identify latent structure, based on a (dirichlet) probability distribution
 - However, generating insights requires a fair amount of interpretation
 - Good topic modeling practice combines quantitative and qualitative insights
- Methods more advanced than LDA such as STM allow for a more dynamic, iterative process of theorization
 - Here, visualization can act as a critical aid to the theorization process

Topic Modeling Is An Interpretive Data Science

- Topic modeling is an approach based on some key assumptions about: i) a corpus as a socially and culturally meaningful set of documents generated from same meaning structures (coding categories), ii) dimensionality reduction as a means for finding those meaning structures and forming representations, iii) ignoring syntax, sentiment, and other grammatical information
 - Not always the right method to use!
- Topic modeling continues to evolve; recent techniques are combining with word embedding methods (BERTopic), AI, and particular innovation on the application side within social science
 - Brandtner, C., Ashur, P., & Srinivasa Desikan, B. (2025). Dynamic persistence of institutions: Modeling the historical endurance of Red Vienna's public housing utopia. *Organization Studies*

The Interpretative Data Science (IDeaS) Group



- Much of this work has stemmed from a collective effect by the Interpretative Data Science (IDeaS) Group (co-founders Dev Jennings, Tim Hannigan, Vern Glaser)
- We are an informal collection of researchers from different academic disciplines and organizational nodes who share an interest in interpretive approaches to curating and employing textual and visual data in management, organizations, and entrepreneurial theory-building – with policy implications for ecosystems, governance and embedded action.
- (<https://www.interpretivedatascience.com/>)

We wish to use topic model rendering

- To Create
- a useful variety interpretive analyses and insights that are both zoomed out further and zoomed in more closely.

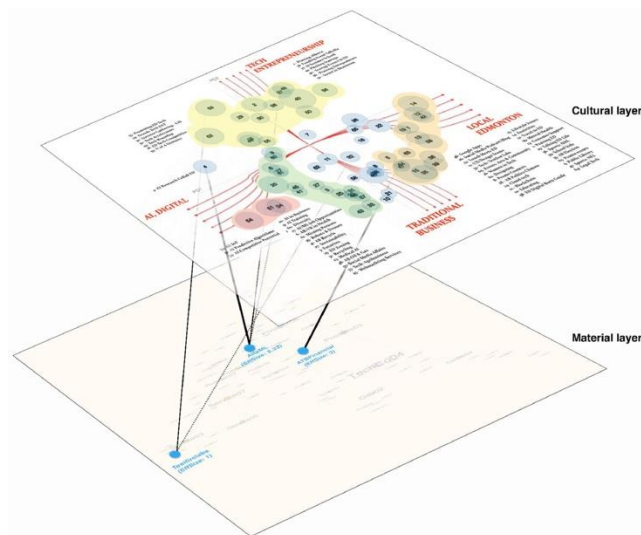
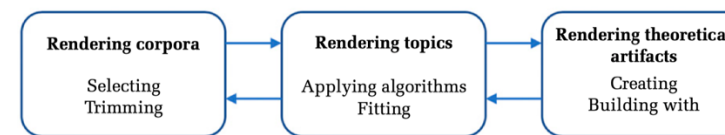
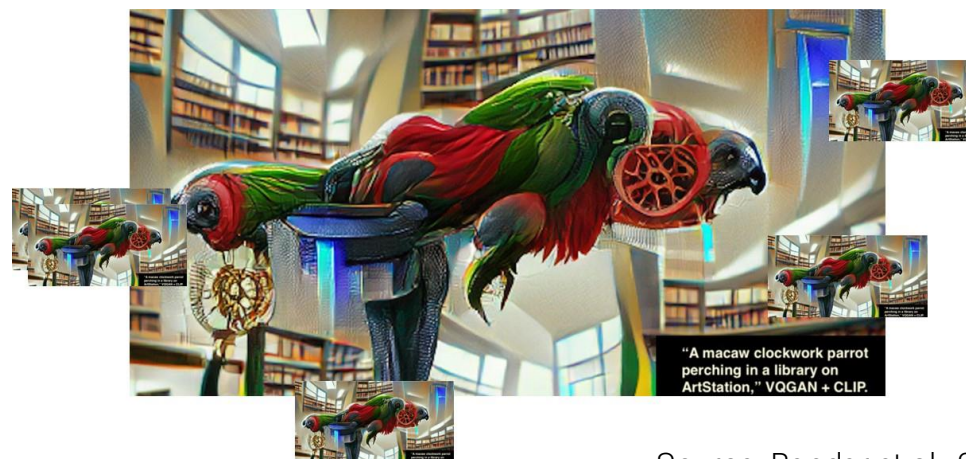


FIGURE 2
Topic Modeling Rendering in Theory-Building Spaces



- Yet Avoid
- unbelievable “stochastic parrots”
- AI hallucinations



Source: Bender et al., 2021



TELFER

ÉCOLE DE GESTION **TELFER** SCHOOL OF MANAGEMENT

Thanks

Email: tim.hannigan@telfer.uottawa.ca