

LBS Research Online

T Hannigan, R F J Haans, [K Vakili](#), H Tchalian, V Glaser, M Wang, S Kaplan and P D Jennings
Topic modeling in management research: rendering new theory from textual data
Article

This version is available in the LBS Research Online repository: <https://lbsresearch.london.edu/id/eprint/1136/>

Hannigan, T, Haans, R F J, [Vakili, K](#), Tchalian, H, Glaser, V, Wang, M, Kaplan, S and Jennings, P
D

(2019)

Topic modeling in management research: rendering new theory from textual data.

Academy of Management Annals, 13 (2). pp. 586-632. ISSN 1941-6520

DOI: <https://doi.org/10.5465/annals.2017.0099>

Taylor & Francis (Routledge)

<https://journals.aom.org/doi/10.5465/annals.2017.0...>

Users may download and/or print one copy of any article(s) in LBS Research Online for purposes of research and/or private study. Further distribution of the material, or use for any commercial gain, is not permitted.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

TOPIC MODELING IN MANAGEMENT RESEARCH: RENDERING NEW THEORY FROM TEXTUAL DATA

Submission to the *Academy of Management Annals*

Tim Hannigan (U. of Alberta: tim.hannigan@ualberta.ca), Richard F.J. Haans (Rotterdam School of Management, Erasmus University: haans@rsm.nl), Keyvan Vakili (London Business School: kvakili@london.edu), Hovig Tchalian (Claremont Graduate University: hovig.tchalian@cgu.edu), Vern L. Glaser (U. of Alberta: vglaser@ualberta.ca), Milo Wang (U. of Alberta: swang7@ualberta.ca), Sarah Kaplan (U. of Toronto: sarah.kaplan@rotman.utoronto.ca), and P. Devereaux Jennings (U. of Alberta: dev.jennings@ualberta.ca)

Corresponding Author:

Dev Jennings
Alberta School of Business
University of Alberta
dev.jennings@ualberta.ca
780-492-3998

Approx. Word Count = 27,500 + 3,000 for appendix (25,000 normal max.)

Key words: topic modeling, management theory, rendering, text analysis, big data, theory building, qualitative analysis, mixed methods

49
50
51
52
53
54
55
56
57
58
59
60

We would like to thank the editors of the Academy of Management Annals for their support and helpful comments. We also thank the participants in our various topic modeling presentations and reviewers and division organizers (specifically Peer Fiss and Renate Meyer) at the Academy of Management meetings. In addition, we would like to recognize Marc-David Seidel and Christopher Steele from the Interpretive Data Science (IDeaS) group for their role in germinating these ideas, Mike Pfarrer for his comments on a later draft of the paper, and Kara Gehman for her fine-grained edits on next-to-final drafts. Finally, we wish to express our appreciation to our life partners for not only putting up with, but actively discussing this paper as it evolved.

ABSTRACT

Increasingly, management researchers are using topic modeling, a new method borrowed from computer science, to reveal phenomenon-based constructs and grounded conceptual relationships in textual data. By conceptualizing topic modeling as the process of *rendering* constructs and conceptual relationships from textual data, we demonstrate how this new method can advance management scholarship without turning topic modeling into a black box of complex computer-driven algorithms. We begin by comparing features of topic modeling to related techniques (content analysis, grounded theorizing, and natural language processing). We then walk through the steps of rendering with topic modeling and apply rendering to management articles that draw on topic modeling. Doing so enables us to identify and discuss how topic modeling has advanced management theory in five areas: detecting novelty and emergence, developing inductive classification systems, understanding online audiences and products, analyzing frames and social movements, and understanding cultural dynamics. We conclude with a review of new topic modeling trends and revisit the role of researcher interpretation in a world of computer-driven textual analysis.

N = 168 words

TOPIC MODELING IN MANAGEMENT RESEARCH: RENDERING NEW THEORY FROM TEXTUAL DATA

New methods can have profound impacts on management scholarship (Arora, Gittelman, Kaplan, Lynch, Mitchell, & Siggelkow, 2016), as they enable scholars to take fresh approaches to theory and re-examine previously intractable problems and old questions (Timmermans & Tavory, 2012). For example, the introduction of event history analysis helped advance both population ecology (Hannan & Carroll, 1992) and institutional analysis (Tolbert & Zucker, 1996) research; the introduction of the case comparison method aided the development of strategy process research (Eisenhardt, 1989); and the introduction of set theoretic methods and qualitative comparative analysis (QCA) led to renewed investigations of configurations (Fiss, 2007; Ragin, 2008). Recently, the management field's understandings of cognition, meaning, and interpretation have been dramatically reshaped by the emergence of new computer-based language processing techniques (DiMaggio, 2015), which have amplified and sharpened the linguistic turn in management research (Alvesson & Kärreman, 2000). In our review, we focus on one of the most commonly used new techniques: topic modeling.

During the last decade, social scientists have increasingly used topic modeling to analyze textual data. Borrowed from computer science, this method involves using algorithms to analyze a corpus (a set of textual documents) to generate a representation of the latent topics discussed therein (Mohr & Bogdanov, 2013; Schmiedel, Müller, & vom Brocke, 2018). It has helped scholars unpack conundrums in management theory, such as how critics' framings of corporate activities simultaneously affect and are affected by their audiences (Giorgi & Weber, 2015), and how knowledge recombination is a double-edged sword with opposite impacts on an innovation's degree of novelty and its usefulness (Kaplan & Vakili, 2015). Similarly, topic

1
2
3 modeling has been used to generate new conceptual linkages, such as how a particular topic
4 appearing in media statements impacted departures of British parliament members (Hannigan,
5 Porac, Bundy, Wade, & Graffin, 2019), and to refine older constructs such as strategic
6 differentiation (Haans, 2019). Because of its features, topic modeling can serve as a bridge in the
7 social sciences, for it sits at the interfaces between case studies and big data, unstructured and
8 structured analysis, and induction and deduction (DiMaggio, Nag, & Blei, 2013; Grimmer &
9 Stewart, 2013; Mützel, 2015). Not surprisingly, its use in social science, and in management
10 theory more specifically, has increased greatly over the last decade.
11
12
13
14
15
16
17
18
19
20

21 As with all new methods, topic modeling techniques continue to be refined. In the current
22 emergent phase of its employment, scholars are still learning the best ways to reveal constructs
23 and develop theory (Evans & Aceves, 2016; Grimmer & Stewart, 2013)—which implies a need
24 for deeper insights into how topic modeling can inform new theories. There are also many
25 technical issues to resolve around topic modeling, such as how to collect and prepare data (Evans
26 & Aceves, 2016), how much supervision should be involved in topic creation (DiMaggio, 2015;
27 Schmiedel et al., 2018), which algorithms are most useful (Bail, 2014), and how new constructs
28 and conceptual linkages can be derived when developing theories from big data (Nelson, 2017,
29 Timmermans & Tavory, 2012). This review addresses these questions with the aim of expanding
30 its use and effectiveness.
31
32
33
34
35
36
37
38
39
40
41
42
43

44 We begin by comparing topic modeling's technical and theory-building features to those
45 of close methodological cousins: content analysis, grounded theorizing, and general natural
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 language processing (NLP) of text.¹ Topic modeling’s attractive features and ease of use are
4
5 generating increased interest across the social sciences—raising the disconcerting possibility that
6
7 the method will become a technical “black box” without an appropriate appreciation of topic
8
9 modeling’s statistical and theoretical underpinnings and implications. In this review, we show
10
11 that topic modeling is best conceptualized as a “rendering process,” which can be understood as
12
13 a means to juxtapose data and theory (Charmaz, 2014) in order to generate new theoretical
14
15 artifacts such as constructs and the links between them (Whetten, 1989). This process involves
16
17 the rendering of corpora (preparing the sets of texts to be analyzed), the rendering of topics
18
19 (making analytical choices that determine how topics are identified within those texts), and the
20
21 rendering of theoretical artifacts (crafting topics into constructs, causal links or mechanisms). By
22
23 articulating this rendering process, we show that using the machine learning algorithms of topic
24
25 modeling do not reduce textual analysis to a mechanistic process, but actually foreground and
26
27 inform the analyst’s interpretive decisions and theory work.
28
29
30
31
32

33 Our own topic modeling analysis of topic modeling articles created or routinely used by
34
35 management researchers reveals five theoretical subject areas to which the technique has
36
37 contributed: detecting novelty and emergence, developing inductive classification systems,
38
39 understanding online audiences and products, analyzing frames and social movements, and
40
41 understanding cultural dynamics. For each subject area, we review key concepts and theoretical
42
43 relationships that have surfaced from the use of topic modeling and identify articles that
44
45
46
47
48
49

50 ¹ Topic modeling can be seen both as a specific NLP approach and as something distinct from NLP. Topic modeling
51 relies on interpretation and language-oriented rules, but is also unique in its emphasis on the role of human
52 researchers in generating and interpreting specific groups of topics based on the social contexts in which they are
53 embedded. Recent developments have also moved topic modeling further away from NLP, as researchers have
54 applied it to images (Cao & Fei-Fei, 2007) and music (Hu & Saul, 2009) rather than natural language.
55
56
57
58
59
60

1
2
3 exemplify its application. We then turn to new trends in topic modeling in the rendering of
4
5 corpora, topics, and theoretical artifacts. Our review demonstrates that topic modeling not only
6
7 appeals to diverse management audiences—those interested in topic, content, and category
8
9 models as well as mixed methods—but also can play a part in cultural structuralism (Lounsbury
10
11 & Ventresca, 2003), new archivalism (Ventresca & Mohr, 2002), and interpretative data science
12
13 (Breiger et al., 2018; Mattmann, 2013).
14
15
16
17
18

19 **SITUATING TOPIC MODELING AS A TECHNIQUE**

20
21 Thanks to widespread availability of digitized textual data from a variety of sources and
22
23 significant increases in computational power, it is now possible for social scientists to study large
24
25 collections of text (Alvesson & Kärreman, 2000; Langley & Abdallah, 2011; Vaara, 2010). Not
26
27 surprisingly, a variety of methods for textual analysis—often from neighboring disciplines—
28
29 have appeared as part of this “linguistic turn.” To distinguish the key characteristics of topic
30
31 modeling and situate it among this wider set of techniques, we first briefly examine three closely
32
33 related methods: content analysis (Duriau, Reger, & Pfarrer, 2007; Krippendorf, 1980, 2004;
34
35 Lasswell, 1948), grounded theorizing with textual data (Gioia, Corley, & Hamilton, 2013; Locke,
36
37 2001), and interpretive analysis using the broad class of NLP approaches. These three are
38
39 particularly useful for elucidating topic modeling’s features because they capture the extremes
40
41 from highly contextualized, careful assessment of smaller batches of selected texts to broader,
42
43 more algorithmic and systematic assessment of text from large corpora.
44
45
46
47
48

49 **Content analysis.** Social scientists have long been interested in using texts to understand
50
51 social phenomena (see Krippendorf, 1980 for a review). Content analysis, “a research technique
52
53 for the objective, systematic, and quantitative description of the manifest content of
54
55
56
57
58
59
60

1
2
3 communication” (Berelson, 1952, p. 18) represents arguably the most prominent and mainstream
4 approach in this domain (Nelson, 2017; Tirunillai & Tellis, 2014). It relies on the creation of
5 dictionaries or indices comprised of mutually exclusive lists of words that can then be applied to
6 texts to isolate meanings and systematically measure specific constructs of interest to the
7 researcher (Krippendorff, 2004). Since its introduction to management theory, scholars have
8 employed content analysis in flexible ways, using a range of data sources in areas as varied as
9 the study of management fads (Abrahamson & Fairchild, 1999), industry categories and CEO
10 compensation (Porac, Wade, & Pollock, 1999), corporate reputation (Pfarrer, Pollock, &
11 Rindova, 2010), and technology strategy (Kaplan, 2008a).

12
13 From its inception, content analysis scholars have been particularly concerned with the
14 reliability and validity of its various methods (Weber, 1990), advocating the use of protocols and
15 multiple coders to guide text selection and analysis. In recent years, those who employ content
16 analysis have increasingly relied on computer-aided text analysis using software and general
17 dictionaries such as General Inquirer and Linguistic Inquiry and Word Count (LIWC) to further
18 improve its scalability and systematic nature. At the same time, the mutually exclusive nature of
19 dictionaries precludes “polysemy” (DiMaggio et al., 2013, p. 578)—an important concept in
20 linguistics where the same word may have a different meaning based on the context in which it
21 appears. A common critique of content analysis has therefore been that it yields decontextualized
22 results by reducing complex theoretical constructs into overly general and simple indices (Dey,
23 1995; Prein & Kelle, 1995).

24
25 **Grounded theorizing with textual data.** To develop theory, scholars often use a highly
26 contextualized approach whereby they gather and engage intensively with texts and then use
27 comparative coding to identify higher-order constructs (Charmaz, 2014). By engaging in such
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 grounded theorizing with textual data, a researcher demonstrates a commitment to “‘discovery’
4 through direct contact with the social world studied coupled with a rejection of a priori
5 theorizing” (Locke, 2001, p. 34). Proponents of this approach urge researchers to start with a
6 loosely scoped research question and phenomenon of interest, with the researcher subsequently
7 identifying recurring patterns, ideas, or elements that emerge directly from the data. Doing so
8 often requires culling primary observations and key points and then using axial coding to identify
9 constructs or relationships (Denzin & Lincoln, 2011). Researchers then iteratively group codes
10 into higher-order categories to develop general theory. Rather than measurement, grounded
11 theorizing is thus fundamentally concerned with identifying deeper structures embedded in data
12 to attain a rich understanding of social processes.
13
14
15
16
17
18
19
20
21
22
23
24
25

26 During the last two decades, grounded theorizing has been used by many groups of
27 management scholars (Charmaz, 2014), including those interested in analyzing language in
28 organizations (Alvesson & Kärreman, 2000), organizational processes and routines (Langley,
29 1999; Pentland & Feldman, 2005), and culture and identity (Hatch & Schultz, 2017; Nelsen &
30 Barley, 1997). Its theoretical flexibility also makes it the target of some critiques, because the
31 role and primacy of meaning, discourse, and understanding typically are not made explicit in
32 research studies (Locke, 2001). Practically speaking, the method also requires great knowledge
33 of context and expertise to apply; it can be not only time- and resource-intensive, but also
34 difficult to use with large scale textual data (Baumer, Mimno, Guha, Quan, & Gay, 2017;
35 Gehman, Glaser, Eisenhardt, Gioia, Langley, & Corley, 2018).
36
37
38
39
40
41
42
43
44
45
46
47
48

49 **Interpretive analysis using NLP.** Researchers in linguistics have long employed
50 computerization to enable systematized analysis of natural language informed by linguistic rules,
51 with NLP emerging in the 1980s as a way to combine dictionary-based data processing with
52
53
54
55
56
57
58
59
60

1
2
3 semantic use to map out likely interpretations of text (Manning & Schütze, 1999). Early versions
4
5 of NLP relied heavily on grammatical rules from language structure, but have given way to more
6
7 flexible, stochastic approaches to language use (especially as machine learning-based approaches
8
9 evolved with increased computing power). In management research, scholars often leverage NLP
10
11 tools to perform semantic parsing on big data and then interpret emerging patterns using
12
13 computer-aided recognition tools. Kennedy (2005, 2008) was one of the first to analyze media
14
15 data and sort through evaluations of firms using these tools. Recently, Mollick and others have
16
17 studied linguistic patterns in crowdfunding and other contexts involving pitches (Kaminski,
18
19 Jiang, Piller, & Hopp, 2017; Mollick, 2014).
20
21
22
23

24 Consistent with its roots in computer science, NLP has been developed to optimize
25
26 specific tasks or solve particular problems, such as part-of-speech tagging, word segmentation,
27
28 machine translation, and automatic text summarization. This has resulted in a rich and varied
29
30 toolkit that is deeply informed by linguistic rules and a firm appreciation for the complexities
31
32 underpinning human language. At the same time, a single unifying theory does not link the
33
34 various NLP tools, nor are there standard practices or rules about engaging in NLP-based work.
35
36 This has created certain challenges for management researchers in applying technical or
37
38 descriptive tools for theoretically informed purposes. Indeed, scholars have noted that
39
40 “cooperation between linguistics and the social sciences with regard to text analysis has always
41
42 been meager” (Pollach, 2012: 264); however, this does not imply that NLP approaches are, by
43
44 definition, unable to inform management theory.
45
46
47
48

49 **Topic modeling.** In the early 2000s, topic modeling was developed as a unique NLP-like
50
51 approach to information retrieval and the classification of large bodies of text (Blei, Ng, &
52
53 Jordan, 2003). Topic modeling uses statistical associations of words in a text to generate latent
54
55
56
57
58
59
60

1
2
3 topics—clusters of co-occurring words that jointly represent higher-order concepts—but without
4
5 the aid of pre-defined, explicit dictionaries or interpretive rules. In a pivotal article, Blei et al.
6
7 (2003) introduced a Bayesian probabilistic model using latent Dirichlet allocation (LDA) to
8
9 uncover latent structures in texts. LDA is a “statistical model of language” (DiMaggio et al.,
10
11 2013, p. 577) and is the simplest of several possible generative models available for topic
12
13 modeling (Blei, 2012). It focuses on words that co-occur in documents, viewing documents as
14
15 random mixtures of latent topics, where each topic is itself a distribution among words (Blei et
16
17 al., 2003). Importantly, an assumption of topic modeling is that documents are “bags of words”
18
19 without syntax, which defines meaning as relational (Saussure, 1959) and emerging from co-
20
21 occurrence patterns independent of syntax, narrative, or location within the documents (Mohr,
22
23 Wagner-Pacifici, Breiger, & Bogdanov, 2013).
24
25
26
27

28
29 Generating topics using statistical probabilities has three key benefits. First, researchers
30
31 do not have to impose dictionaries and interpretive rules on the data. Second, the method enables
32
33 the identification of important themes that human readers are unable to discern. Third, it allows
34
35 for polysemy because topics are not mutually exclusive; individual words appear across topics
36
37 with differing probabilities, and topics themselves may overlap or cluster (DiMaggio et al., 2013,
38
39 p. 578).
40
41

42 **A comparison of text analysis techniques in management research.** Figure 1
43
44 compares the use of topic modeling in social science and management research to the use of
45
46 grounded theory, content analysis, and general NLP approaches in articles listed in the Web of
47
48 Science and Scopus published between 2003 (the year Blei and colleagues’ foundational article
49
50 was published) and 2017. We included articles for topic modeling if “topic mod*” appears in
51
52 their titles, abstracts, keywords, or automated indexed keywords. We included articles for
53
54
55
56
57
58
59
60

1
2
3 grounded theorization, content analysis and NLP if they contain “ground theor*,” “content
4 analys*,” and “natural language process*,” respectively.² The bar charts in each panel represent
5 the cumulative number of articles in each year, with black bars showing the number of articles in
6 business and economics specifically, and white bars showing articles in the social sciences more
7 generally.
8
9
10
11
12
13

14
15 --- Insert Figure 1 about here ---
16

17 As a group, the four panels highlight the linguistic turn in social science, with increased
18 use of all of these approaches reflecting the increasing appetite in the field to study the structure
19 and meaning underpinning collections of text. By 2017, 1,000 topic modeling articles had been
20 published, with around 300 in the management domain specifically. Although this is just a
21 fraction of the literature relative to studies based on more established approaches, Figure 1 does
22 suggest that the use of topic modeling has been particularly high in the management domain.
23
24 Indeed, 29.8% of all articles based on topic modeling published between 2003 and 2017 fall
25 within the management domain, compared to 13.4%, 22.0%, and 22.9% for NLP, grounded
26 theorization, and content analysis, respectively. Figure 1 also reveals that topic modeling has
27 been adopted at an exceptionally rapid rate in recent years, with a compound annual growth rate
28 of 34.4% since 2010, versus 11.1% for NLP, 15.1% for grounded theory, and 16.5% for content
29 analysis. We suggest that topic modeling’s appeal primarily lies in its unique position at the
30 intersection of the other three approaches, a point that we elaborate in the conclusion.
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

51
52 ² Although these may under-count articles that do not mention the methodologies and over-count articles without
53 textual data, we suspect that these issues are equally salient for each approach. For illustration, adding “Linguistic
54 Inquiry and Word Count” and “LIWC” adds just 271 articles to the set of over 20,000 for content analysis.
55
56
57
58
59
60

RENDERING THEORY FROM DATA IN TOPIC MODELING

Given its increasing importance in the social sciences and its unique location between human-based and machine-learned analysis of discourse, a more careful consideration of the nature of topic modeling and the topic modeling process is useful for management researchers. To date, much of the work on topic modeling has focused on issues of algorithm selection (e.g., Blei et al., 2003; Schmiedel et al., 2018) and its application to curated texts. We think it is important to discuss the use of topic modeling from the pre-processing to theorization stages to illustrate its possibilities for theory building.

We use the term “rendering” to describe the iterative creation of theory from corpora through topic modeling. In the social sciences, Charmaz (2014, pp. 216, 369) employed the term rendering to describe the process of “juxtaposing data and concept” and “categorizing data” for interpretation, while computer scientists use rendering to create photorealistic or non-photorealistic images in two or three dimensions via automated analysis and specific algorithms (Strothotte & Schlechtweg, 2002). Drawing on these descriptions for inspiration, we define 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*. These three steps are displayed in Figure 2. To provide readers with background information, we present definitions of common terms used in topic modeling in Table 1.

--- Insert Figure 2 and Table 1 about here ---

Rendering corpora

In the first process—rendering corpora—an analyst, guided by theoretical and empirical

1
2
3 considerations, *selects* types of textual data. As with any form of empirical analysis, selection of
4 the sample (in our context, texts) is a crucial step that fundamentally shapes all subsequent steps.
5
6 For textual data in particular, selection needs to account for language, authoring, and document
7 sources—ensuring a logical fit with the research question being investigated while
8
9 simultaneously considering common issues such as representativeness, levels of analysis, and
10 temporal considerations (e.g., longitudinal vs. cross-sectional data). The analyst then compiles
11 such data for further pre-processing and cleaning. If the data are from one primary source, the
12 compiled text is considered a corpus; if from different sources, corpora.
13
14
15
16
17
18
19
20
21

22 On the whole, topic modeling tends to be applied more frequently to sampled corpora
23 than to a single, homogenous corpus (Borgman, 2015; Kitchin & McArdle, 2016). As a result,
24 topic modeling relies on a great deal of pre-processing with various techniques and rules of
25 practice to prepare texts for analysis (Nelson, 2017; Schmiedel et al., 2018). During pre-
26 processing, the texts are sorted, disassembled, and then trimmed according to broader content
27 analysis principles such as ignoring “stop words” (for example: “the” and “a”) and focusing on
28 nouns rather than verbs, adjectives, or adverbs. Topic modelers also often standardize word
29 forms, using *stemming* and *lemmatizing* (see Table 1) to transform words into their roots
30 (Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2018). Recently, more refined techniques
31 such as WordNet have been developed to convert words to their singular forms or to use higher-
32 level synonyms (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). These considerations are
33 all crucial, as most topic modeling algorithms analyze words based on how they appear, letter-
34 by-letter (e.g., “firm” is not the same as “firms”). As such, these cleaning steps represent a form
35 of systematic, normatively-guided trimming to standardize words to allow the capture of
36 constellations of words that represent deeper socio-cultural structures (Mohr, 1998).
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Rendering topics

During the second process—rendering topics—the analyst applies an algorithm to identify appropriate topics. An algorithm provides an analyst with the ability to use a pre-programmed set of rules to automatically reduce the dimensions of the corpora (e.g., Mohr, 1998). The most well-known algorithm, as discussed above, is LDA. According to Blei et al. (2003, p. 994), the key assumption in LDA is that “each word in a document [is modeled] as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of ‘topics.’” The major theoretical and methodological insight here is that documents are assumed to draw content from a latent set of topics with probability-based parameters that can be adjusted to determine those topics. This implies that words are generated from a topic, yet can also be used in different topics with different probabilities. Because documents belong to the same corpus, the algorithm assumes that they were generated from the same process, and thus each document constitutes a mixture of the same set of “topics” in different proportions. Topics are a weighted vector of words and each topic corresponds to a distinct concept (Grimmer & Stewart, 2013). However, unlike the dictionaries used in content analysis, which are comprised of mutually exclusive lists of words (Krippendorff, 2004, p. 132), in topic modeling, the same words can appear in different topics (DiMaggio et al., 2013, p. 578), though likely in very different proportions and juxtaposed with different words.

The inputs to the LDA algorithm include: (a) a set of documents that can be represented as a document-word matrix—with rows representing each document in the corpora, columns representing each unique word in the corpus, and cells indicating the number of times each word

1
2
3 occurs in each document—and (b) the number of topics to be estimated by the algorithm.
4
5 Importantly, most topic modeling algorithms (such as LDA) require probability draws for each
6
7 document, such that each document is considered “a bag of words” with no syntax. The outputs
8
9 from LDA include a topic-word matrix (vectors of the weights of words in each topic) and a
10
11 topic-document matrix (vectors of the weights of topics in each document). In subsequent
12
13 analyses, math (i.e., vector space calculations) can be applied to these outputs to classify texts
14
15 into categories, analyze themes, or compare corpora based on similarities.
16
17
18

19 Each successfully computed model is based on different parameters (e.g., number of
20
21 topics) and generates a distribution of topics over documents and/or words, which can be used by
22
23 the researcher to identify the eventual model that will be used in the study. The notion of *fit* is
24
25 typically invoked to decide how many topics are derived, how they are related, and what they
26
27 might mean. A researcher can focus on one of two notions of fit—rooted in a logic of either
28
29 accuracy or validity—and this focus has important implications for which topic model is judged
30
31 to provide the most appropriate fit given the research question.
32
33
34

35 One version of fit is based on a logic of accuracy, a central focus of computer scientists
36
37 who rely on metrics such as *perplexity*, *log-likelihood* and *coherence* (defined in Table 1) to
38
39 determine the number of topics and their salience (Azzopardi, Girolami, & van Risjbergen, 2003;
40
41 Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009; Mimno, Wallach, Talley, Leenders, &
42
43 McCallum, 2011). However, Chang et al. (2009) pointed to disparities between some
44
45 quantitative metrics and how people interpret topics: topic models that perform better on
46
47 quantitative metrics tend to infer topics that humans judge to be semantically less meaningful.
48
49 Indeed, DiMaggio et al. (2013, p. 582) suggested that “there is no statistical test for the optimal
50
51 number of topics or for the quality of a solution” and that “the point is not to estimate population
52
53
54
55
56
57
58
59
60

1
2
3 parameters correctly, but to identify the lens through which one can see the data most clearly.”
4

5 Therefore, social scientists tend to focus more on the logic of fit as validity (DiMaggio,
6 2015). DiMaggio et al. (2013) identified two key forms of validity: semantic or internal validity,
7 and predictive or external validity. To demonstrate internal validity, the researcher must confirm
8 that the model meaningfully discriminates between different senses of the same or similar terms.
9
10 To demonstrate external validity, the researcher must determine whether particular topics
11 correspond to information external to the topic model (e.g., by confirming that certain topics
12 became more salient when an external event relevant to those topics occurred) (DiMaggio et al.,
13 2013). For example, Kaplan and Vakili (2015) identified models with 50, 75 and 100 topics for a
14 corpora of nanotechnology patent abstracts and then used three expert evaluators to determine
15 that the 100-topic model was the most semantically meaningful. Jointly, these two forms of
16 validity are concerned with confirming that the topic model’s outputs are semantically
17 meaningful—a process that entails substantial interpretive uncertainty (DiMaggio, 2015). Due to
18 the uncertainty involved in the rendering of topics, most scholars in the social sciences attempt to
19 locate the optimal balance between the two logics of accuracy and validity to identify the “best”
20 topic model to be used in further theorizing.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 In sum, topic modeling has advanced how we think about and interpret topics in textual
41 data by enabling researchers to uncover latent topics rather than imposing pre-established
42 categories on the data. It is superior to word-count techniques because it identifies ideas or
43 concepts based on constellations of words used across documents in a corpus. It is thus sensitive
44 to semiotic principles of polysemy (words with multiple meanings or uses), heteroglossia (uses
45 predicated on audiences and authors, as described by Bakhtin, 1982), and the relationality of
46 meaning (which is contextually dependent) (DiMaggio et al., 2013). As a result, topic model
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 outputs, after some interpretation and theoretical defense, are useful in generating theoretical
4 artifacts, especially in large and otherwise unmanageable data sets.
5
6
7
8
9

10 **Rendering theoretical artifacts**

11
12 In the third process—rendering theoretical artifacts—researchers iterate between theory
13 and the topics that emerge from the chosen model to *create* new theoretical artifacts or to *build*
14 theory with them (Whetten, 1989). The word- and topic-vectors offer a wide range of
15 opportunities for the researcher to build artifacts. The artifacts may be multi-dimensional
16 constructs, such as novelty (Kaplan & Vakili, 2015) or differentiation (Haans, 2019), captured by
17 a set of topics clustered or scaled around words or concepts. The artifacts may also be relational
18 (correlational, causal or process-based), thereby allowing researchers to uncover mechanisms.
19
20
21
22
23
24
25
26
27

28 For instance, Croidieu and Kim (2018, p. 11) used an “iterative, multi-step process” to
29 interpret the outputs of the topic model in order to discover concepts related to lay expertise
30 legitimation and the mechanisms underpinning it. They described their process for creating
31 theoretical artifacts from their algorithmic output in detail.
32
33
34
35
36
37

38 First, we started with the raw topics as descriptive codes. Second, we labeled these topics
39 as first-order concepts. We coded all labels separately and together as an author team,
40 extensively discussed the results, and recoded the topics when necessary. Third, we
41 grouped these topics into more abstract and general second-order themes. Fourth, we
42 analyzed the distribution of these second-order themes per year and iteratively developed
43 four aggregate dimensions, which we present in the following sections as the mechanisms
44 for expertise legitimation. Fifth, we refined the labeling and theorizing of these aggregate
45 dimensions by dividing our analysis into two periods... We chose these periods both for
46 their historical significance and because they are anchored by a central empirical puzzle
47 related to our theoretical framework... Last, we repeated this procedure multiple times to
48 ensure tight correspondence between our raw-topic data and our coding interpretations.
49 From this iterative coding work, we produced our findings and constructed our process
50 model. (Croidieu & Kim, 2018, p. 11)
51
52

53 The inherent flexibility of the rendering process has enabled topic modeling researchers
54
55
56
57
58
59
60

1
2
3 to develop better measures and clever extensions of existing theoretical constructs and
4 relationships, and to induce novel concepts, processes, and mechanisms. As such, topic modeling
5 can be used for either deductive or inductive theorizing. Indeed, during the rendering process,
6 different choices arise (e.g., around selection, fit, and the form of artifact) based on whether one
7 uses more deductive versus inductive theorizing. The many paths defined by these choices
8 provide further evidence of topic modeling's flexibility and potential. Not surprisingly, topic
9 modeling is contributing to a wide array of management theory subjects, some arising from more
10 mature theory, some from emerging areas.
11
12
13
14
15
16
17
18
19
20
21
22
23

24 **BUILDING MANAGEMENT KNOWLEDGE THROUGH TOPIC MODELING**

25
26 During the 15 years since topic modeling was first employed in management research, its
27 use through rendering has enabled management scholars to explore subjects in new ways,
28 thereby building management knowledge. To systematically identify the subjects enhanced by
29 such rendering, we applied the topic modeling rendering process depicted in Figure 2 to topic
30 modeling articles in the literature (for similar meta-theorizing moves, see Mohr & Bogdanov,
31 2013, or Wang, Bendle, Mai, & Cotte, 2015). Although our rendering process was iterative and
32 recursive, we present our methodological approach as a series of sequential steps, as outlined in
33 Figure 1 (e.g., rendering our corpus, topics, and theoretical artifacts).
34
35
36
37
38
39
40
41
42
43

44 We began our analysis by curating a corpus consisting of all relevant topic modeling
45 articles from the Web of Science and Scopus. We winnowed those articles down by focusing on
46 management journals (e.g., *ASQ*, *SMJ*, etc.) and other journals that management scholars read.
47 We identified these journals based on both our first-hand experience and citations of articles that
48 have influenced management scholars. Following the procedure employed by Mohr and
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Bogdanov (2013), we divided the articles into paragraphs to form 5,362 documents and used the
4
5 Stanford CoreNLP software (Manning et al., 2014) to lemmatize the words, yielding 351,786
6
7 distinct words for analysis. During our analysis, we sharpened our criteria for including and
8
9 excluding particular articles in our analysis as we interpreted the output of topic modeling
10
11 algorithms. Our final corpus contained 66 articles (for details, consult Table A1 in the
12
13 Appendix). We organized these procedures using the Jupyter Notebook software in Python,
14
15 which enabled us to track and visually annotate our process.
16
17
18

19 We continued our analysis by applying a collapsed Gibbs sampler with the LDA
20
21 algorithm to our corpus to render topics. Collapsed Gibbs sampling (Griffiths & Steyvers, 2004)
22
23 is an approach from the Markov Chain Monte Carlo framework that iteratively steps through
24
25 configurations to estimate optimal model fit. When combined with the LDA algorithm (Blei et
26
27 al., 2003), topics can be estimated with minimal configuration by the user. As is common
28
29 practice (e.g. Mohr & Bogdanov, 2013; Jha & Beckman, 2017), we used the MALLET software
30
31 tool (McCallum, 2002) to conduct this procedure. We approached the critical task of determining
32
33 the optimal number of topics by computing a variety of topic models. For each model, we
34
35 graphed the average coherence score across topics (Mimno et al., 2011), which revealed a
36
37 plateau value; we used this evidence as guidance and observed several models (i.e., those with
38
39 30, 35, 40, 45, and 50 topics) more closely from an interpretive perspective. Fligstein et al.
40
41 (2017) followed a similar procedure, moving from collapsed Gibbs sampling through various
42
43 models, using coherence and interpretability to narrow in on stable sets of topics. Finally,
44
45 following Mohr and Bogdanov (2013), we applied our 35-topic model (derived from separate
46
47 paragraphs) to each document to generate a distribution of topic weights (i.e., the topic-document
48
49 matrix where each row is a document and each column is a topic weight, with all weights adding
50
51
52
53
54
55
56
57
58
59
60

1
2
3 up to 1). We then sorted topics for salience based on average topic weights and word relevance
4
5 to identify 35 ordered topics.
6

7
8 Three co-authors then independently used the algorithmic output of the topic models to
9
10 render theoretical artifacts. Specifically, we each created a summary document for each topic
11
12 that contained three visualizations generated by the topic modeling algorithm: a weighted word
13
14 list, a weighted document list, and a multidimensional scaling visualization (Sievert & Shirley,
15
16 2014) that showed each topic in relation to other topics (see Appendix, Figure 2, for an example
17
18 of this theoretical artifact). The three authors then independently analyzed these documents to
19
20 generate first- and second-order codes (e.g., Bansal & Corley, 2014; Denzin & Lincoln, 2011;
21
22 Gioia et al., 2013; Pratt, 2009; Strauss & Corbin, 1998). Through a series of independent coding
23
24 exercises and interactive conversations, the authors then aggregated these first- and second-
25
26 order codes into broader management subject areas (e.g., Gioia et al., 2013). In other words, in
27
28 keeping with rendering practice, we tried not to impose too much meaning on the set of topics;
29
30 instead, we let the insights and themes for management theorizing emerge from them.
31
32
33
34

35 Our bottom-up, inductive analysis suggests that topic modeling has enhanced our
36
37 management theory knowledge in five subject areas: detecting novelty and emergence,
38
39 developing inductive classification systems, understanding online audiences and markets,
40
41 analyzing frames and social movements, and understanding cultural dynamics.³ This specific
42
43 ordering of subjects is not determined by topic weights; moreover, the timing of their
44
45 identification in the model's convergence does not reflect a strict ordering. In fact, our
46
47
48
49
50

51
52 ³ In addition, some topics corresponded specifically to the method of performing topic modeling, and given our
53
54 interest in the rendering of management theory, we purposefully backgrounded these topics (see Appendix Table 2
55
56 for details).
57
58
59
60

1
2
3 preliminary analyses of the wider corpora in the field and understanding of the field's evolution
4
5 reveal how analyses of novelty, classification and online audiences developed in parallel with
6
7 analyses of framing and cultural dynamics. In the sections that follow, we focus on how
8
9 theoretical knowledge in each subject area has been extended by rendering with topic modeling.
10
11 Subject areas, topic-based themes, exemplary articles, and theoretical contributions are
12
13 summarized in Table 2.
14
15

16
17 --- Insert Table 2 about here ---
18

19 **Detecting novelty and emergence.** Management researchers are interested in topics of
20
21 novelty and emergence because they apply to a variety of research streams, such as categories
22
23 (e.g., Durand & Khaire, 2017; Hannan et al., 2007; Kennedy & Fiss, 2013), cultural
24
25 entrepreneurship (e.g., Lounsbury & Glynn, 2001, 2019), innovation (e.g., Fleming, 2001;
26
27 Sørensen & Stuart, 2000), organizational forms (e.g., Rao et al., 2003), and changes in
28
29 managerial cognition and attention (e.g., Ocasio, 1997). Novelty is a key concern within
30
31 innovation studies (Kline & Rosenberg, 1986; Trajtenberg, 1990), but measures typically are
32
33 indirect. For instance, as noted by Kaplan and Vakili (2015), many studies identify emergence
34
35 based on the successful introduction of new innovations, thus raising concerns of endogeneity
36
37 and lack of causal identification.
38
39
40

41
42 Topic modeling offers a solution to fundamental challenges faced in these broad research
43
44 streams. Specifically, topic modeling can be applied to documents to generate theoretical insights
45
46 because: (a) the language used in documents represents their cognitive content (Whorf, 1956);
47
48 and (b) actors use vocabularies to describe similar ideas (Loewenstein, Ocasio, & Jones, 2012).
49
50 Thus, topic modeling can be used to discern the cognitive content of documents that describe
51
52 cases of novelty and emergence (i.e., innovation contexts) and assess the extent to which such
53
54
55
56
57
58
59
60

1
2
3 content is similar or different across documents. Topics rendered in our analysis include:
4
5 explaining shifts in patent citations (#25), understanding innovation (#24), managerial cognition
6
7 (#1), understanding knowledge dynamics (#14), and emerging organizational forms (#10).
8
9

10 The first topic in this subject area relates to the use of topic modeling to measure the
11
12 novelty of ideas in patents—an arena in which novelty has been heavily studied under the rubric
13
14 of recombination and innovation (Fleming, 2001). For instance, Kaplan and Vakili (2015)
15
16 applied topic modeling techniques to create representations of ideas in documents that can be
17
18 compared using mathematical distance to determine cognitive novelty. This measure of novelty
19
20 based on the actual cognitive content of documents provides several advantages over more
21
22 traditional measures of novelty based on citations in subsequent patents or publications
23
24 (Trajtenberg, 1990). In the popular citation-based approach, a patent is flagged as a breakthrough
25
26 if it has a substantial impact on subsequent technologies. However, citation-based measures of
27
28 technological novelty often confound novelty and impact (Momeni & Rost, 2016); consequently,
29
30 novel ideas may not be recognized as important precursors due to the processes by which
31
32 citations are produced (false negatives), and incremental ideas may be incorrectly identified as
33
34 novel when they generate substantial impact for reasons other than novelty (false positives).
35
36
37
38
39

40 In contrast to simple counts of citations or patent classes, a measure based on the
41
42 cognitive content of a document enables researchers to gauge the novelty of the idea(s)
43
44 presented, independent of their ex-post economic value. Kaplan and Vakili (2015) used topic
45
46 modeling to distinguish cognitive novelty from economic value. In their analysis of nanotube
47
48 patents, they reported a very small correlation between topics identified by LDA and patent
49
50 classes assigned by the U.S. Patent and Trademark Office (USPTO). Often, truly novel ideas are
51
52 assigned to classes that may not reflect their actual cognitive content. Their study has
53
54
55
56
57
58
59
60

1
2
3 implications for teasing out longstanding debates in management around contrasting theories of
4 creative processes surrounding the sources of innovative breakthroughs. In a related study,
5
6 Ruckman and McCarthy (2017) used topic modeling to analyze patents in an attempt to explain
7
8 why some patents are licensed over others. Their goal was to address conflicting findings in prior
9
10 research: some scholars have advocated a “status model” (Podolny, 1993), whereas others have
11
12 supported organizational learning explanations based on optimizing knowledge transfer in
13
14 licensing contracts (Arora, 1995). Ruckman and McCarthy used topic modeling to directly
15
16 measure cognitive content, enabling them to construct a set of “alternate patents” that could have
17
18 been licensed based on content, but were not. Thus, by controlling for cognitive content, they
19
20 were able to isolate other variables such as the licensor’s technological prestige and experience at
21
22 licensing, and characteristics of the patent itself such as combined technological breadth and
23
24 depth. Using better controls when comparing similar patents enabled them to produce a
25
26 contingent model of patent licensing likelihood based on licensor attributions and the
27
28 combination of technological breadth and depth as an attractive signal. Topic modeling has thus
29
30 enabled researchers who study patents and innovation to not only increase the precision of their
31
32 analyses, but also develop new theory about the role of knowledge dynamics on economic
33
34 outcomes.
35
36
37
38
39
40
41

42 A second topic in this subject area that is closely related to explaining shifts in patent
43 citations is the use of texts more generally as a means to measure innovation and creativity.
44
45 Toubia and Netzer (2016) proposed that creative and novel ideas should have some type of
46
47 structural signature that can be found in cognitive representations. Drawing on literature related
48
49 to cognitive creative processes in science (i.e., Rothenberg, 2014; Uzzi et al., 2013), they
50
51 explored this proposition as an optimal balance of familiarity and novelty. Toubia and Netzer
52
53
54
55
56
57
58
59
60

1
2
3 (2016) primarily adopted a semantic network analysis approach to explore the structural
4
5 argument of familiarity, showing how co-occurrences of word stems can constitute a common
6
7 substructure, what they called a “structural prototype.” In turn, they argued that creativity is a
8
9 function of a semantic network structure with a core substructure corresponding to a familiar
10
11 prototype, and novelty dimensions reflected as sufficient semantic distance in the overall
12
13 structure. They demonstrated this argument empirically across eight studies and 4,000 different
14
15 ideas in multiple domains that were coded by expert judges. They used LDA as a robustness
16
17 check to show that creativity was not simply a function of semantic distance. Interestingly, both
18
19 Toubia and Netzer (2016) and Kaplan and Vakili (2015) featured in this topic: in different
20
21 domains, the authors leveraged topic modeling techniques to theorize how to identify innovation
22
23 in documents through the direct measurement of cognitive representations.
24
25
26
27

28 The third and fourth topics—using topic models to understand managerial cognition and
29
30 knowledge dynamics—relate to actors detecting novelty within a body of knowledge. The core
31
32 idea of employing topic modeling to study knowledge dynamics is based on two related insights:
33
34 first, the language used in documents represents their cognitive content (Whorf, 1956); and
35
36 second, actors use similar vocabularies to describe similar ideas (Loewenstein, Ocasio, & Jones,
37
38 2012). In our analysis, the third topic reveals that topic models can be used to understand
39
40 changing cognition over time through varying managerial attention (Ocasio, 1997). When a
41
42 corpus covers the body of knowledge in a specific domain (e.g., scientific papers or patents in
43
44 the technology field), topic modeling can reveal an accurate depiction of the idea space in that
45
46 body of knowledge. However, topic modeling can also reveal how actors, as producers of
47
48 documents, can attend to ideas in the latent idea space. As Kaplan and Vakili (2015)
49
50 demonstrated, to the extent that describing a truly novel (or disruptive) idea requires using a new
51
52
53
54
55
56
57
58
59
60

1
2
3 vocabulary, one can identify the level of cognitive novelty in a document by measuring how
4 much it conforms to or deviates from previously established topics and their constitutive
5 vocabularies in the corresponding body of knowledge. Wilson and Joseph (2015, p. 417)
6 employed topic modeling to render the “patent background” as a “representation of a technical
7 problem” at a particular point in time. Because managerial attention is scarce, it is allocated
8 across a small set of technological problems, particularly at the level of a business unit (Argote
9 & Greve, 2007). Thus, the rise and fall of topics as technological problems reflect not only
10 managerial attention within a firm, but also novelty within the broader field or patent class.
11
12
13
14
15
16
17
18
19
20
21

22 Topic modeling has also been used to study knowledge dynamics in science by tracking
23 the novelty of ideas in journals over time. Conceptualizing scientific communities as “thought
24 collectives with distinct thought styles,” Antons, Joshi, and Salge (2018, p. 1) used topic
25 modeling to break down articles in terms of topical and rhetorical attributes. They demonstrated
26 that topical newness is not only associated with a paper “citation premium” in a scientific
27 community, but also significantly increases with a rhetorical stance of tentativeness rather than
28 certainty. Similarly, Wang et al. (2015) used topic modeling to discover emerging trends in
29 knowledge fields, noting that citation analyses and LDA together can be used to narrate a story
30 about novelty and progress against a broader backdrop of social structure, including niche topical
31 areas and author status dynamics. Both articles in this topic contextualize traditional citation
32 based measures of article impact against cognitive dynamics in topic analyses.
33
34
35
36
37
38
39
40
41
42
43
44
45
46

47 A final topic revealed by our analysis of this subject area reflects the use of topic
48 modeling to understand emerging organizational forms. This approach provides a method to
49 trace how meanings of organizational forms emerge longitudinally. Jha and Beckman (2017)
50 used topic modeling to show how field-level logics moderated actors’ attempts to carve out
51
52
53
54
55
56
57
58
59
60

1
2
3 organizational identities around charter schools. Topic modeling enabled the authors to connect
4
5 two traditionally distinct theoretical concepts—institutional logics and organizational
6
7 identities—and explain the relationships between them. Given how meaning has typically been
8
9 studied in organizational theory using concepts such as identity, institutional logics, and frames,
10
11 studying the emergence of meanings in spaces such as organizational fields and categories may
12
13 become an increasingly relevant application of topic modeling methods.
14
15

16
17 Topic modeling has increased precision and enabled deeper insights in studies of novelty
18
19 and knowledge dynamics, thereby facilitating the generation of new theory in a variety of
20
21 innovation-related contexts. Topic modeling provides considerable advantages over traditional
22
23 methods such as counts of patent filings or subsequent citations, which rely on existing
24
25 classification methods that were not designed to capture novel and emergent ideas. By directly
26
27 leveraging the cognitive content of texts (such as patents or papers), topic modeling augments
28
29 traditional measures of impact in knowledge fields. Furthermore, by separating measures of
30
31 impact from those of knowledge itself, topic modeling has advanced theory by empowering
32
33 researchers to invent more precise means to empirically test competing theoretical mechanisms.
34
35 In the bigger picture, these uses of topic modeling may help scholars address longstanding
36
37 questions in the management literature by conceptualizing the role of novelty with institutional
38
39 logics (Thornton et al., 2012), or delineating the roles of innovation and boundaries with
40
41 paradigms (Kuhn, 1996).
42
43
44
45

46
47 **Developing inductive classification systems.** Management researchers routinely use
48
49 topic modeling to develop inductive classification systems. Such systems are particularly
50
51 important in a variety of theoretical research streams, including studies of competitive dynamics
52
53 and optimal distinctiveness (Deepphouse, 1999; Zhao et al., 2017), and the evaluation of risk
54
55
56
57
58
59
60

1
2
3 factors in corporate disclosures to investors (e.g., Fama & French, 1993). More generally, these
4
5 research streams are exploring classification as shared structures of meaning that are not
6
7 formally materialized. For example, studying institutional logics (Thornton et al., 2012) or
8
9 implicit understandings of early industry structure (Forbes & Kirsch, 2011) requires researchers
10
11 to develop inductive understandings of shared meanings that have categorical imperatives.
12
13
14 Researchers in each of these traditions who seek to identify categories of meaning in text face
15
16 challenges of analyzing large quantities of data without introducing researcher bias. Our analysis
17
18 reveals six topics in this subject area: understanding dynamics of meanings and networks in
19
20 knowledge fields (#34), understanding how categories affect competitive dynamics (#18),
21
22 understanding the relationships between risk and investment (#31), inducing underlying
23
24 meanings associated with cultural events (#32), and classifying sets of data and consumers (#4).
25
26
27

28
29 The first topic reveals how researchers use topic modeling to compare hidden meaning
30
31 structures in knowledge fields with networks of relationships among articles, journals, scholars
32
33 and citations. One approach has been to track the development of a journal or field by combining
34
35 historical topic modeling analyses with bibliometrics and authorship networks (Cho, Fu, & Wu,
36
37 2017; Wang et al., 2015) to confirm field-level insights using patterns of dominant topics while
38
39 rendering “hidden structures and development trajectories” (Antons et al., 2016, p. 726). This
40
41 approach has been applied in science to track the rise and fall of meanings within a journal
42
43 (Antons et al., 2016; Wang et al., 2015). For instance, Antons et al. (2016) used a semi-
44
45 automated topic model combining both inductive (machine) analysis and abductive (human)
46
47 labeling and generalization to add fine-grained detail to prior reviews of literature in the *Journal*
48
49 *of Product Innovation Management*. Their topic model revealed latent meaning structures not
50
51 identified in earlier reviews because the journal’s interdisciplinary character made it difficult to
52
53
54
55
56
57
58
59
60

1
2
3 identify and properly assess the breadth of papers published during its 30-year history.
4

5 A major benefit of Antons et al.'s (2016) approach is the ability to compare and contrast
6 content according to classification schemes in the field and then induce categories of topics.
7
8 They first applied the topic model analysis using LDA. After employing methodological best
9 practices and ensuring inter-rater reliability across 14 researchers, they clustered related topics
10 into six semantically-meaningful groups, including new ones the authors identified and labeled
11 (once again, inductively) in correspondence with the interpretation and theory-generation stages
12 depicted in Figure 3. The authors then made an abductive, conceptual link to disciplinary
13 trends—that is, they modeled “topic dynamics” by creating a weighting scheme. Finally, the
14 authors combined this human-centered approach with a final and more automated deductive
15 move, regressing topics that appeared more frequently than the median topics (those with a topic
16 loading greater than 10%) for each year of their analysis, tracing topic development by
17 comparing each of the topics against the mean, and in a final abductive iteration, classifying
18 them according to trajectory shape (“hot,” “cold,” “revival” and “evergreen”). The result is a
19 large-scale, many-to-many classification scheme across the entire study period that serves as a
20 comprehensive semi-automated literature review, balancing meaningful knowledge categories
21 with abductively rendered topics.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 In another form of rendering in the classification of science, scholars have used topics as
43 intermediate artifacts to perform social network analyses of authorship behavior. Cho et al.
44 (2017) used topic modeling to augment co-authorship network data from 25 marketing journals
45 over a 25-year period. Building on the work of Wang et al. (2015), who used topic modeling to
46 map topic usage over time in the *Journal of Consumer Research* to predict promising research
47 topics for the future, Cho et al. (2017) showed that social network analysis revealed two major
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 communities of co-authors, whereas topic modeling analysis revealed three. They then used
4 these intermediate analyses to show that communities of highly-cited papers corresponded to
5 these intermediate analyses to show that communities of highly-cited papers corresponded to
6 heterogeneous clusters of related topics, but that the communities identified by each method had
7 different features. In combining topic modeling with network analysis, Cho et al. (2017) showed
8 how journals comprise the ecology of a field, but the structures constituting it (communities) can
9 be seen at the levels of both citations and topics. Management scholars are not alone in
10 employing topic modeling analysis to advance field-level bibliometric studies, as it is being
11 adopted in psychology (Oh, Stewart, & Phelps, 2017) and the humanities (Mimno, 2012) as well.
12 Topic modeling has thus provided scholars with a way to both develop new understandings of
13 cultural meanings and to connect those understandings with network and other structural features
14 of fields.
15
16
17
18
19
20
21
22
23
24
25
26
27

28 A second topic relates to the role of categories in shaping competitive dynamics.
29 Questions around optimal distinctiveness have long been of interest to management scholars
30 (Deepphouse, 1999; Navis & Glynn, 2011; Zhao, Fisher, Lounsbury, & Miller, 2017), but this line
31 of research is contingent upon the ability to measure coherence and variation of strategic action
32 against the backdrop of a category. How to delineate categorical boundaries is thus a key
33 concern. Haans (2019) explored the optimal distinctiveness of firm positioning relative to
34 industry categories. He used topic modeling on texts from organizational websites to uncover the
35 strategic positioning of firms in Dutch creative industries. The method enabled him to calculate
36 both industry average and distinctiveness measures for individual firms. By using topic modeling
37 to induce bottom-up, positioning-based classifications, Haans (2019) was able to generate new
38 theoretical insights that diverged from prior research by suggesting that optimal distinctiveness
39 for organizations depends on the distinctiveness of other organizations. Thus, positioning-based
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 classification, as identified through topical analysis has strategic implications. In related work,
4
5 scholars have used topic modeling to develop important conceptual infrastructure in the form of
6
7 inductive classifications for research on industry intelligence and competitive dynamics (Guo,
8
9 Sharma, Yin, Lu, & Rong, 2017; Shi, Lee, & Whinston, 2016).

12 A third topic in this area identifies topic modeling as a means to derive categories of risk
13
14 perception in finance. Such studies build on a long history of debates about the impact of
15
16 corporate disclosures on investor behavior (Fama & French, 1993). Researchers have struggled
17
18 to classify how risk factors are communicated and perceived by companies, analysts, and
19
20 investors. In contrast to the established method of using predefined dictionaries for content
21
22 analysis to quantify risk types (e.g., Campbell, Chen, Dhaliwal, Lu, & Steele, 2014 using the
23
24 schema: idiosyncratic, systematic, financial, tax, litigation), researchers have applied
25
26 unsupervised learning methods to financial texts to inductively classify risk factors. For example,
27
28 Bao and Datta (2014) applied LDA to induce risk types from corporate 10-K forms, and then
29
30 tested these against risk perceptions of investors, advancing theory by showing that the topic
31
32 modeling-induced risk meanings better predicted investor perceptions of risk. Huang, Lehavy,
33
34 Zang, and Zheng (2017) were able to extend this analysis to inductively identify risk factors and
35
36 other economically interpretable topics within analyst reports and corporate conference calls,
37
38 providing additional insights into how analysts both discover relevant information and interpret it
39
40 on behalf of investors. In both of these papers, scholars used topic modeling to extend textual
41
42 analyses of corporate financial disclosures by moving beyond the “how” (i.e., volume, sentiment,
43
44 and length) to the level of topical meaning in terms of “what is the meaning of what is being
45
46 said.” Topic modeling thus has enabled researchers to develop better classification systems based
47
48 on the textual data being sampled.
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Another topic focuses on meanings associated with cultural events that are not captured
4
5 by formal documents and artifacts. Miller (2013) used topic modeling to capture meanings
6
7 around the nature of violence during the Qing Dynasty in China. Instead of relying on a fixed set
8
9 of categories, the method enabled him to induce an original typology of violence based on the
10
11 Qing administrator's perceptions of unrest. Similarly, Ahonen (2015) applied topic modeling
12
13 techniques to challenge existing theory by inductively identifying the sources of legal traditions
14
15 across countries. The author considered differences in legal language in government budgeting
16
17 legislation as a basis for distinguishing between legal traditions. Both studies offer an approach
18
19 to overcome biases associated with interpreting cultural events.
20
21
22
23

24 In similar articles, scholars have used topic modeling to study topic-based classifications
25
26 in patent data (Kaplan & Vakili, 2015; Suominen, Toivanen, & Seppänen, 2017; Venugopalan &
27
28 Rai, 2015). The practice of mapping knowledge structures in science is in its infancy, and the use
29
30 of topic modeling has the potential to change how scientific fields are classified (see Song, Heo,
31
32 & Lee, 2015; Song & Kim, 2013; Yau, Porter, Newman, & Suominen, 2014) since topic
33
34 modeling analyses do not perfectly correspond to formal systems of classification (Cho et al.,
35
36 2017; Kaplan & Vakili, 2015). Topic modeling analyses also may reveal insights when used in
37
38 conjunction with other forms of analysis such as citation and co-authorship patterns. As such,
39
40 topic modeling can yield more fine-grained classifications and extend classic bibliometric and
41
42 content analysis methods.
43
44
45
46

47 The papers we reviewed in this section map the knowledge spaces and dynamics of
48
49 academic fields. Topic modeling enables scholars to compare latent topics in particular
50
51 documents with pre-existing bodies of knowledge and quantitatively measure broad trends in
52
53 meaning, thus providing a counterpoint or corroboration of coding performed exclusively by
54
55
56
57
58
59
60

1
2
3 humans. Because topic modeling is a rendering process based on human and algorithmic efforts,
4 employing it to map knowledge spaces uncovers latent classification systems that may or may
5 not overlap with more formal classifications. Our review of papers in this subject area has
6 resulted in the discovery of new concepts that can be used to better understand phenomena in a
7 variety of management research streams.
8
9
10
11
12
13

14 **Understanding online audiences and products.** For the last two decades, management
15 theorists have been particularly interested in understanding how audiences evaluate firms and
16 products in research on cultural entrepreneurship (Martens, Jennings, & Jennings, 2007; Navis &
17 Glynn, 2010, 2011), status (Podolny, 1993), categories (Hannan et al., 2007; Zuckerman, 1999),
18 and now, with the expansion of the Internet, understanding how these dynamics may change in
19 online contexts (Mollick, 2014). These scholars have sought to understand the deeper patterns
20 and meanings of producer communications and theorize audiences' reactions (e.g., Cornelissen
21 et al., 2015). Nevertheless, isolating nuances both in the meanings of sensegiving
22 communications (e.g., about products) and the responses of heterogeneous audiences remains
23 difficult.
24
25
26
27
28
29
30
31
32
33
34
35
36

37 Topic modeling has been taken up by researchers—particularly in marketing—
38 to analyze the cognitive content of online discourse about products and the behavior of online
39 consumers as audiences. This subject area of understanding online audiences and products has
40 emerged out of four topics: the nature of online consumer profiles (#12), online consumer brand
41 recognition and preferences (#23), online customer evaluations and responses to them (#29), and
42 enhanced topic modeling techniques on products and audiences (#13).
43
44
45
46
47
48
49
50

51 The first topic, the nature of online consumer profiles, has been advanced by
52 conceptualizing consumers based on the clicking patterns of different online groups (Trusov, Ma,
53
54
55
56
57
58
59
60

1
2
3 & Jamal, 2016), the network of related brands and brand tags clicked on by consumers (Netzer et
4 al., 2012), and communities of consumers defined based on common virtual market participation
5 (e.g., portals) or similar patterns of geo-location markers (Zhang, Moe, & Schweidel, 2017). In
6 these studies, topics were rendered not just from a “bag of words” across a corpus of documents,
7 but from a “bag of behaviors” across a corpus of activities. This conceptual pivot maps roles to
8 “topics” of behaviors. For example, click patterns for a group across diverse products/services
9 during a particular time period offer unobtrusive measures of both a latent set of consumer
10 profiles and their associated behaviors. Marketing studies using topic modeling have also
11 uncovered evaluations by consumers in new ways. For instance, Zhang et al.’s work (2017) on
12 elite universities revealed that the willingness to tweet—and, even more importantly, retweet—
13 about topics associated with a university reinforces the elite university status hierarchy.
14 Ironically, the most elite of the elites receive more tweet outs and retweets, not only from their
15 own members, but also from members of other universities. Management scholars interested in
16 categories (Durrand & Paoella, 2015; Vergne & Wry, 2014) and communities (Marquis &
17 Davis, 2007) might use these re-conceptualized online consumer communities to broaden
18 theorization and measures of their core constructs. Scholars might also use online endorsements
19 (clicks and tweets) to complement other forms of analyst assessments (Giorgi & Weber, 2015;
20 Zuckerman, 2001).

21
22 A second topic is online brand recognition and preference. Here, scholars conceptualize
23 brands not just as specific offerings with cachet, but as the associated networks of audiences
24 linked to those products along with the sets of user-generated tags employed by audiences to
25 identify brand groups. For example, Nam, Joshi, and Kannan (2017) used topic modeling to
26 render representative topics based on user-generated “social tags” from the shared bookmarking
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 service Delicious. They then examined how Apple customers linked and endorsed Apple
4 products via product tags, such as, “mac,” “phone,” and “Apple,” all of which were linked to
5
6 “Apple Corporation.” The brand in its fullest form (Apple), then, was the overall network of
7
8 linked tags used by customers. Similarly, Netzer, Feldman, Goldenberg, and Fresko (2012) used
9
10 car brand clicks on the online forum Edmonds.com to identify co-occurring words in topics
11
12 about different car brands. The clusters of words (topics) revealed overlaps, evolving brand
13
14 clusters, and “semantic networks” (i.e., meaningful text-based attributes) that differentiated
15
16 brands. In addition, Netzer et al. (2012) were able to anticipate brand switches within and across
17
18 these topic-based networks. They did so by studying changes in discussions about and
19
20 associations among brands in these topic networks (also see Tirunillai & Tellis, 2014). These
21
22 rendering moves do not differ significantly from management theory approaches to fashion and
23
24 design (Dalpiaz, Rindova, & Ravasi, 2016) and exemplar categories (Zhao, Ishihara, Jennings, &
25
26 Lounsbury, 2018); management scholars working in this vein might broaden their
27
28 understandings of how meaning is associated with brands and use topic modeling to augment
29
30 their measures of templates and categories. In addition, given the association of brand and
31
32 identity (Navis & Glynn, 2010; Raffaelli, 2018), management scholars might use group brand
33
34 identification (as measured by topic preferences) to track identity formation and evolution.
35
36
37
38
39
40
41

42 A third topic focuses on the dynamics of influencing online consumers, or in other words,
43 how agency is exercised online and with what effects. Marketing scholars, by and large, believe
44 that online consumers are more difficult to understand and influence because they are
45
46 decentralized, diverse, and switch often. Research identified as related to the topic of online
47
48 consumer responses suggests that learning adjustment is due to latent structural modifications
49
50 around topics captured by analyzing online forum data. For example, Puranam, Narayan and
51
52
53
54
55
56
57
58
59
60

1
2
3 Kadiyali (2016) used topic modeling to analyze all New York City restaurant reviews before and
4 after the implementation of a regulation that required posting calorie counts; their results
5
6 demonstrate a shift in online consumer evaluations, and in their view, food consumption patterns
7
8 in New York City. More recently, Wang and Chaudhry (2018) examined online hotel ratings,
9
10 and the effects of managers' responses to positive and negative customer reviews. They used
11
12 LDA to generate a measure of response tailoring by comparing the content of managers'
13
14 responses to a baseline value. Highly tailored managerial responses to negative reviews were
15
16 considered by customers to be a form of high-quality complaint management; in contrast,
17
18 tailored responses to positive reviews were considered to be overly promotional (hence,
19
20 backfired on management). The use of topic modeling techniques to capture consumer
21
22 evaluations and adjustments is of interest to management scholars engaged in cultural analysis
23
24 and neo-structuralism research (DiMaggio, 2015; Lounsbury & Ventresca, 2003; Mohr &
25
26 Bogdanov, 2013), because a bedrock assumption in these culture-oriented approaches is that
27
28 agency is less observable and more distributed. Topic modeling of online reviews across
29
30 audiences can also help capture actor adjustments around latent structures (e.g., see Hannigan et
31
32 al., 2019; Heugens & Lander, 2009). In addition, longitudinal, affect-based topic modeling might
33
34 enrich studies of performance adjustment (Greve, 2003), anchoring (Ballinger & Rockman,
35
36 2010), and event analysis (Morgeson, Mitchell, & Liu, 2015).
37
38
39
40
41
42
43

44 A final topic in this subject area is focused on improving topic modeling of online
45 audiences and products to capture nuances of communication and audience responses (#13). The
46
47 groundbreaking and oft-cited work by Lee and Bradlow (2011) regarding automated online
48
49 reviews has several features that have become norms for rendering with topic modeling, such as
50
51 using triangulation (e.g., with *k*-means clustering and MDS), mapping structures, thinking about
52
53
54
55
56
57
58
59
60

1
2
3 “fit” with algorithms, and examining change over time. Recently, Guerreiro, Rita, and Trigueros
4 (2016) and Jacobs, Donkers, and Fek (2016) introduced correlational topic models, sentence-
5 based models, and hierarchical topic models to demonstrate the utility of using some supervision
6 and structure in topic model rendering. Along similar lines, Büschken and Allenby (2016) used
7 sentences and phrases rather than words as inputs for LDA to show that topics based on them
8 might exhibit less change (i.e., be “sticky”) over time. Because management researchers are
9 currenly interested in understanding the interface of such methods and derived topics and
10 meaning (DiMaggio, 2015; Schmeidel et al., 2018), Büschken and Allenby’s (2016) work poses
11 an interesting rendering question for management researchers: Is stickiness a product of using
12 sentences (the method) or is it due to linguistic meaning being constructed at the sentence-
13 (rather than word-) level by online consumers?
14
15
16
17
18
19
20
21
22
23
24
25
26
27

28 To summarize, using topic modeling to analyze online audiences and products enables
29 management scholars to think more deeply about the nature of online audiences (e.g., as click-
30 based profiles, virtual networks, and computer-mediated communities); to reconceptualize
31 products as distributed brands tied to evolving individual and category identities; and to capture
32 the more subtle means by which audiences evaluate online products, and correspondingly
33 understand how organizations might adjust in real time to those evaluations. In addition, the
34 refinement of topic models of online audiences creates modeling standards for other topic
35 modeling research, and encourages scholars to think more deeply about the meaning given to
36 products by online audiences.
37
38
39
40
41
42
43
44
45
46
47
48

49 **Analyzing frames and social movements.** Topic modeling also has been used to analyze
50 frames and understand the dynamics of social movements. Management scholars have long been
51 interested in symbolic management (Zajac & Fiss, 2006; Zajac & Westphal, 1994; Zott & Huy,
52
53
54
55
56
57
58
59
60

1
2
3 2007), such as understanding how investors respond to organizational framing efforts (Giorgi &
4 Weber, 2015; Rhee & Fiss, 2014), theorizing the political dynamics associated with different
5 framing strategies within firms (Kaplan, 2008b), and understanding the dynamics of social
6 movements (e.g., Benford & Snow, 2000). This research requires scholars to identify frames—
7 epistemological devices that actors use to organize experiences by answering the question posed
8 by Goffman (1974, p. 8): “What is it that’s going on here?”
9

10
11
12
13
14
15
16
17 Topic modeling methods have helped scholars expand theoretical boundaries in this area
18 by providing an empirical method for inductively uncovering latent frames and then
19 understanding the dynamics associated with frame proliferation and effectiveness. Our topic
20 modeling analysis revealed four topics in this subject area: understanding how frames influence
21 political processes (#27); the relationship between frames, context, and audience (#6);
22 understanding field-level relationships between organizations, discourses, and strategies (#17);
23 and social movement strategies, networks and actions (#11).
24
25
26
27
28
29
30
31
32

33 The first topic relates to how frames influence political processes. Frames enable actors
34 to “render what would otherwise be a meaningless aspect...into something that is meaningful”
35 (Goffman, 1974, p. 21). Scholars are particularly interested in the often political and contested
36 dynamics associated with framing (e.g., Fiss & Hirsch, 2005; Kaplan, 2008b). An exemplar
37 article showing how topic modeling can contribute to this research stream is Fligstein et al.’s
38 (2017) study of the Federal Open Market Committee’s decision-making processes in public
39 meetings. Specifically, they sought to develop a theory to explain how the committee failed to
40 appropriately perceive the risks to the economy in the months leading up to the financial crisis.
41 In addition to confirming the existence of macroeconomics as a master frame, their topic
42 modeling approach revealed the existence and application of a banking frame and a finance
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 frame. By focusing on the specific events—the housing bubble and the financial crisis—the
4
5 researchers were able to track which frames came to dominate Fed committee discussions at the
6
7 time of each event. The authors thus used topic modeling to develop a theory that explains how a
8
9 predominant frame can blind actors involved in decision-making processes.
10
11

12 A second topic explores the relationship between frames, context, and audience. Actors
13
14 use distinct frames to advance their interests (Kaplan, 2008b) and seek to create effective frames
15
16 through mechanisms such as frame alignment (Snow, Rochford, Word, & Benford, 1986) or
17
18 frame resonance (Snow & Benford, 1988). In an exemplar article, Levy and Franklin (2014) used
19
20 topic modeling as a means of identifying distinct discursive frames. Specifically, they used a
21
22 study of political contention in the U.S. trucking industry regarding hours of service to
23
24 inductively analyze the frames that emerged from a study of comments on a public website. They
25
26 were able to use topic modeling to uncover distinct differences between individual and
27
28 organizational uses of frames in the debate, showing how different parties used different frames
29
30 to promote their interests. Uncovering nuanced distinctions in framing content deployed by
31
32 different parties over time can help researchers generate new theory about the influence of
33
34 communication content and techniques on political processes.
35
36
37
38
39

40 The third topic relates to research on field-level relationships between organizations,
41
42 discourse, and strategy. Specifically, to understand framing effects, it is often necessary to move
43
44 beyond the content of a specific frame. To illustrate, Bail, Brown, and Mann (2017) explored the
45
46 relationship between conversational and emotional styles in advocacy work—seeking to
47
48 incorporate sentiment analysis into our understanding of frames. The authors used topic
49
50 modeling to classify the types of topics raised by autism advocates and used LIWC to capture
51
52 sentiment and bias in normalized spaces. This unique combination of topic modeling and LIWC
53
54
55
56
57
58
59
60

1
2
3 sentiment analysis enabled them to reveal the cognitive and emotional “currents” running
4 through advocacy groups, and to show how the ability to “dispatch messages that contribute to a
5 phase shift [between emotional and cognitive-focused communication]” ultimately leads to more
6 effective results (Bail et al., 2017, p. 1205). Thus, topic modeling has enhanced our
7 understanding of frame effectiveness in the context of broad field-level relationships between
8 organizations, discourse, and strategy.
9

10
11 Similarly, the fourth topic relates to researchers’ attempts to understand the relationship
12 between social movement strategies, networks, and actions. For example, Almquist and Bagozzi
13 (2017) sought to understand the network relationships between radical environmental activists in
14 the United Kingdom. Based on a longitudinal corpus of a radical social movement’s texts, they
15 identified the centrality of network ties and then used structural topic modeling to locate the
16 groups and the positions they took on various radical issues, thereby enabling them “to evaluate
17 whether the presence of a given group tie (or cluster member) significantly increases the
18 attention dedicated to a given topic” (Almquist & Bagozzi, 2017, p. 26). By combining structural
19 topic modeling and network analysis, the authors were able to classify subnetworks of actors to
20 develop a better theoretical account of the discursive actions and network relationships of social
21 movements by mapping unseen or hidden ties. Put another way, topic modeling generates
22 theoretical artifacts that facilitate researchers’ efforts to connect the content of communications
23 with other theoretical constructs.
24
25

26
27 In summary, topic modeling provides several benefits that have led to significant
28 theoretical advancements related to frames and framing. First, topic modeling has helped
29 researchers strengthen their understanding of frames. For example, scholars can use topic
30 modeling to track the prominence of researcher-derived high-level frames for large corpora over
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 an extended period of time. Additionally, the algorithmic nature of topic modeling approaches
4 ensures the replicability of identified frames. Second, the inductive nature of many topic
5 modeling techniques enables the discovery of unanticipated frames and audiences that use them,
6 providing a powerful opportunity for scholars to generate new theory. Specifically, topic
7 modeling methods enable researchers to understand the dynamics associated with the co-
8 presence of competing voices within a single text (i.e., heteroglossia, Bakhtin, 1982), which
9 provides researchers with a way to study multiple competing or collaborative frames. Finally,
10 topic modeling facilitates the creation of new theory since it produces theoretical artifacts that
11 can be paired with other forms of analysis such as sentiment analysis or network analysis.
12
13
14
15
16
17
18
19
20
21
22
23

24 **Understanding cultural dynamics.** Management scholars have sought to leverage
25 psychological and sociological research on culture—“the interaction of shared cognitive
26 structures and supra-individual cultural phenomena (material culture, media messages, or
27 conversation, for example) that activate those structures” (DiMaggio, 1997, p. 264)—to explain
28 diverse phenomena. For example, in research on institutional logics (e.g., Thornton, Ocasio, &
29 Lounsbury, 2012), strategic action fields (e.g., Fligstein & McAdam, 2011), and professions
30 (e.g., Abbott, 1988), scholars have theorized the evolution and impact of cultural meanings at the
31 level of an institutional field. In research on organizational culture (e.g., Hatch, 1993) and
32 organizational identity (e.g., Gioia & Thomas, 1996), scholars have theorized the evolution and
33 impact of cultural meanings at the level of the organization. In research on cultural
34 entrepreneurship (e.g., Lounsbury & Glynn, 2001, 2019; Martens et al., 2007) and institutional
35 work (e.g., Lawrence, Suddaby, & Leca, 2009), scholars have attempted to understand how
36 individuals leverage cultural material to achieve strategic objectives. In all of these areas,
37 researchers have attempted to theorize both the dynamics of cultural influences and the evolution
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 of cultural concepts.
4

5 Overall, this research on culture has faced significant challenges. One such challenge
6 relates to the measurement of cultural constructs. For example, scholars have defined
7 institutional logics as “the socially constructed, historical pattern of material practices,
8 assumptions, values, beliefs, and rules by which individuals produce and reproduce their material
9 subsistence, organize time and space, and provide meaning to their social reality” (Thornton &
10 Ocasio, 1999, p. 804). But in empirical studies, it has been harder to specify them. A second
11 challenge is to understand the temporal dynamics associated with culture. For example, in
12 cultural entrepreneurship research, scholars attempt to understand how entrepreneurial
13 organizations are able to legitimate a new market category over an extended period of time (e.g.,
14 Navis & Glynn, 2010). Researchers also attempt to connect cultural meanings with events and
15 actions, for example, by connecting the content of organizational discourse with changes in
16 organizational networks and broader social discourse (Bail, 2012).
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

33 Scholars have used topic modeling methods to push the boundaries of our understanding
34 of such cultural dynamics. Our analysis reveals five themes in this research: understanding the
35 professionalization of a field (#2), using topic modeling to analyze big data to understand
36 cultural trends (#5), understanding dynamics associated with literary meanings (#9),
37 understanding how cultural meanings change over time (#19), and understanding the evolution
38 of cultural trends (#28). Topic modeling has enabled scholars to generate novel theory by
39 providing an operational means to identify cultural concepts and then trace the evolution of those
40 concepts over time and across different locations of social space.
41
42
43
44
45
46
47
48
49
50

51 The first topic in this area revolves around developing new theory about the
52 professionalization of fields. Specifically, Croidieu and Kim (2018) theorized the rise of alternate
53
54
55
56
57
58
59
60

1
2
3 fields and quasi-professions by studying the emergence of U.S. wireless radio broadcasting field
4 and the “lay professional legitimation” of amateur radio operators from 1899 to 1927. To
5 understand the legitimation process for amateur operators, the authors had to gather a wide,
6 diverse constellation of documents from various archival sources: U.S. government regulations,
7 radio operators from the era, radio corporations, and the *New York Times*. They analyzed the
8 distribution of topics over time and by audience to determine the meanings of those patterns
9 using historical (or case) records. This process enabled the authors to identify first- and second-
10 order mechanisms by period. They paired topic modeling of diverse archival materials with
11 standard historical reading and complementary content analysis to create and defend a theoretical
12 account of professionalization based on historical data.
13
14
15
16
17
18
19
20
21
22
23
24
25

26 A second topic focuses on how big data can be used to understand cultural trends. These
27 articles describe and illustrate nuances of the processes scholars use to extract meanings from
28 large corpora. For example, Wagner-Pacifici, Mohr, and Breiger (2015) summarized a special
29 issue in *Big Data & Society* on assumptions of sociality that synthesized the results of several
30 other subjects. First, they highlighted the importance of recognizing that big data methods,
31 unreflexively applied, can lead to biased results. Second, they discussed the importance of the
32 interpretive role of analysts who use big data and related methods to generate theory. Third, they
33 emphasized how big data methods require a move away from traditional deductive science,
34 highlighting their inherently inductive and abductive nature. Finally, they showed how analyzing
35 big data requires scholars to ask fundamental questions such as “What is a thing? What is an
36 agent? What is time? What is context? What is cause?” (Wagner-Pacifici et al., 2015, p. 5). Thus,
37 scholars must reflexively consider the cultural implications of studying big data.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54 Interestingly, in sociological research that has provided analogical inspiration for
55
56
57
58
59
60

1
2
3 management scholars, Mohr and Bogdanov (2013) used topic modeling to analyze literary
4 meanings. In the humanities, Tangherlini and Leonard (2013) introduced a technique called sub-
5 corpus topic modeling to compare canonical texts with broader literature and societal discourse.
6 Specifically, they used the technique to “develop a well-curated topic model of a sub-corpus”
7 and then used “the ensuing model to discover passages from the large, unlabeled corpus”
8 (Tangherlini & Leonard, 2013, p. 728). To illustrate the utility of their method, they showed how
9 topics associated with Charles Darwin’s intellectual ideas penetrated “into the broader literary
10 world” (Tangherlini & Leonard, 2013, p. 735). They thus used topic modeling to understand
11 topics associated with well-known texts and then applied the outputs to analyze other, less well-
12 known cultural meanings.
13
14
15
16
17
18
19
20
21
22
23
24
25

26 Another evident topic focuses on how cultural meanings evolve over time. An example of
27 this can be seen in the work of DiMaggio et al. (2013), who identified the frames invoked and
28 crafted by news outlets in their coverage of the public controversy surrounding the U.S.
29 government’s support of artists and art organizations. The authors rendered corpora using data
30 from five mainstream media outlets; after applying unsupervised LDA to isolate and link topics,
31 they inductively identified different frames. Their results reveal not only the differences across
32 frames by time period, but also how a single text produced by these media outlets might use
33 multiple frames. Applying a fractional multinomial logit analysis, they calculated the expected
34 relative prominence of topics based on their LDA analysis. By further aggregating those topics
35 into particular topic groupings, then classifying them as conflict or comparison frames, they were
36 able to reveal the likely link between the relative increase in conflict topics that accompanied the
37 growing sentiment against public funding for U.S. arts organizations starting in the 1980s. The
38 authors thus used topic modeling to identify different frames of cultural meaning in the public
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 sphere and then showed how these meanings changed over time.
4

5 A final topic looks at the impact of cultural meanings on societal actions. For instance,
6
7 Marshall (2013) sought to understand the evolution of cultural trends by contrasting how
8
9 different academic theories of demography unfolded over a 60-year period in Great Britain and
10
11 France. Specifically, she used correlated topic modeling (to account for the assumption that
12
13 topics in her corpus might be correlated across documents) to understand how concepts
14
15 associated with fertility were understood (and unfolded) differently in different cultural contexts.
16
17 She used topic modeling to identify topics, measure the prevalence of those topics in the corpus,
18
19 and then connect those topics to the dominant theories of demography in effect during that time.
20
21 The topic modeling analysis enabled her to identify differences between the responses of French
22
23 and British academics to changing demographics during the study period. Topic modeling thus
24
25 enables scholars to trace the evolution of cultural trends by connecting the prevalence of themes
26
27 in discourse to historical events.
28
29
30
31
32

33 Overall, topic modeling has provided management scholars with a new methodology for
34
35 generating novel insights about cultural dynamics. First, topic modeling provides a means to
36
37 develop an unbiased understanding of the prevalence of distinct cultural concepts over an
38
39 extended period of time, thereby enabling scholars measure cultural concepts more precisely.
40
41 Second, topic modeling enables scholars to compare a well-known subset of knowledge to
42
43 broader corpora that might reflect that knowledge structure more generally, thereby enabling
44
45 scholars to develop new theories and link constructs that previously had been difficult to
46
47 connect, both empirically and theoretically. Similarly, topic modeling enables researchers to see
48
49 how different meanings within the discourse surrounding a particular topic exist and shift over
50
51 time. Finally, topic modeling can connect shifts in discourse to broader cultural trends.
52
53
54
55
56
57
58
59
60

NEW TRENDS RELATED TO TOPIC MODELING AND RENDERING

Many new trends in management and computer science research are relevant to management scholars' use of topic modeling to render corpora, topics, and theoretical artifacts (see Figure 2). Each trend within a rendering process has a unique trajectory that is important to discuss and respect. For instance, some trends broaden specific rendering processes (e.g., creating corpora), whereas others deepen them (e.g., fitting topic models). Trends also involve some of the aforementioned management subject areas. In this section, we discuss not only trends, but also their implications for rendering and building management knowledge.

Trends in Rendering Corpora

As management researchers embrace approaches that move beyond dictionary-centric content analysis, corpus selection becomes an even more critical step in topic modeling research. Recent methods papers on text analysis reveal a broad effort to engage more closely both with computational linguistics and NLP (Kobayashi et al., 2018; Schmeidel et al., 2018). These efforts were precipitated by an important shift toward conceptualizing corporal dimensions to enable comparison.

Corpus linguistics. Within management, this trend of engaging with computational linguistics is most evident in a recent special issue of *Organizational Research Methods* (Tonidandel, King, & Cortina, 2018) on big data and modern data analytics. This special issue demonstrates the arc of pre-processing corpora as a precursor to higher order text analyses with big data (Kobayashi et al., 2018; Schmeidel et al., 2018). However, many of these pre-processing techniques were highlighted several years earlier by Pollach (2012), who pointed management

1
2
3 researchers to a branch of linguistics known as “corpus linguistics” to show how word patterns
4 can lead to meaningful insights by virtue of the corpora in which they appear. Techniques for
5 analyzing corpora themselves—both qualitatively and quantitatively—include word frequency
6 lists, keyword-in-context searches, the comparison of corpora, word collocations, and statistical
7 methods for assessing word-frequency patterns.
8
9

10
11
12
13
14
15 Pollach (2012) originally positioned corpus linguistics techniques as methodological
16 innovations for content analysis. In very recent work, Kobayashi et al. (2018, p. 1) took a
17 broader approach, suggesting that such pre-processing considerations represent a “fundamental
18 logic” of mining “text data.” As part of that mining, papers in this vein have stressed the
19 imperative of pre-processing as “wrangling” text data into a corpus (Braun, Kuljanin & DeShon,
20 2018). Schmiedel et al. (2018) have laid out some steps that recognize the fundamental
21 importance of data collection and cleaning in topic modeling analysis. Theoretically speaking,
22 these papers draw on core ideas from linguistics, such as the famous *distributional hypothesis*
23 (Firth, 1957)—that is, “words that occur in the same contexts tend to have similar meanings”
24 (Turney & Pantel, 2010, p. 142). Inferring meanings, in other words, depends on the context
25 created by the corpus. As a result, these recent papers are raising the bar in terms of the level of
26 sophistication and reporting standards required for scholars who use topic modeling and other
27 text analysis methods.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

44
45 In fact, we built our rendering process on the insight that corpora curation has
46 implications for theoretical work because meaning is inferred from context. A source corpus
47 begins as natural language, which can be messy and thus requires selecting and trimming. These
48 two steps standardize documents, which then enable topics in the corpus to be rendered at a
49 higher level of abstraction. Moretti (2013) called this “distant reading,” where a corpus can be
50
51
52
53
54
55
56
57
58
59
60

1
2
3 fully and adequately represented in terms of topics. Sharpening this reading requires iteration; for
4
5 this reason, our rendering process has an arrow pointing back from rendering topics to rendering
6
7 corpora. The trends we identified in pre-processing point to the adaptation of techniques from
8
9 corpus linguistics for the purposes of corpus curation, thereby expanding the toolkit for
10
11 rendering.
12
13

14 **NLP.** Innovations in NLP are advancing how scholars prepare and preprocess the words
15
16 in corpora. NLP research highlights two key concerns: first, as the base unit of meaning, a token
17
18 (a word, parts of words, or phrase combining words) is a function of grammar; and, second,
19
20 structures of grammar are embedded in sentences, which have co-dependencies across words and
21
22 paragraphs within a document. Uttered meanings correspond to parts of speech. For example, the
23
24 meaning of the token *Google* changes based on whether it is a noun (i.e., referring to the
25
26 company or software), or a verb (i.e., referring to use of the search engine), and can be referred
27
28 to in a similar manner through a pronoun in a subsequent sentence. Thus, a token as a unit of
29
30 meaning may be a word or multiple words (i.e., a phrase) (Chomsky, 1956).
31
32
33
34

35 NLP research suggests that latent meaning in texts can be captured by bigrams, or two-
36
37 word units rather than individual words, as in the standard “bag of words” approach (Manning,
38
39 Raghavan, & Schütze, 2010). Some management researchers have therefore shifted the unit of
40
41 analysis to a “bag of sentences” (Bao & Datta, 2014; Büschken & Allenby, 2016). Determining
42
43 the boundary of analysis is technically tricky. For example, because a sentence break is not just a
44
45 function of searching for the full stop character (i.e., “.”), researchers have developed NLP
46
47 methods to determine sentence boundaries in a common task called sentence segmentation (Kiss
48
49 & Strunk, 2006). Moreover, advanced deep learning algorithms (e.g., neural networks) are being
50
51 introduced that go beyond “bag of words” approaches altogether to consider syntactic position
52
53
54
55
56
57
58
59
60

1
2
3 and context when identifying linguistic structures such as constituency and dependency parsing
4 representations (Manning et al., 2014). Deep learning is an unsupervised algorithm that can be
5 trained on large text corpora to “learn” latent structures, including semantic compositionality
6 (Socher et al., 2013) within texts (or other kinds of data) that can then be used for explanatory or
7 predictive purposes.
8
9

10
11
12
13
14
15 Additional advances have improved the precision of identifying tokens. For example,
16 mentions of individual actors may be standardized by employing NLP technologies such as
17 Named Entity Recognition (Mohr et al., 2013) and co-reference resolution (Manning et al.,
18 2014). The former is an NLP method that can automatically identify entities based on their
19 appearance in texts and can annotate analytical codes as actors, organizations, and countries. The
20 latter is an NLP tool that can extend Named Entity Recognition to pronouns and other references
21 to entities across sentences. Standardizing entities to resolve ambiguities inherent in manifested
22 natural language facilitates machine-based reading.
23
24
25
26
27
28
29
30
31
32

33 Approaches to making such transformations are particularly salient in topic modeling
34 because this trimming determines the token unit upon which topics are established (Schmiedel et
35 al., 2018). These decisions regarding rendering corpora have theoretical implications. The NLP
36 methods discussed here are largely inductive tools, with machine learning algorithms annotating
37 texts. While inductive methods have become more widely accepted in management journals,
38 there is still considerable risk of over-fitting findings to the data if scholars generalize too
39 quickly (i.e., engage in “theoretical over-fitting”) (Tchalian, 2019). Thus, researchers must
40 continue to check the validity of such annotating.
41
42
43
44
45
46
47
48
49
50

51 **Non-Western languages.** Another new corpus-rendering trend that touches upon these
52 developments in corpus linguistics is the treatment of languages that are structurally dissimilar to
53
54
55
56
57
58
59
60

1
2
3 most Western languages—in particular, languages without spaces between words (or, *scriptio*
4 *continua*), including many Southeast Asian writing systems (e.g., Thai, Burmese, Lao) and those
5
6 that use Chinese characters (i.e., Chinese and Japanese). Treatment of these languages is not
7
8 straightforward. For example, each Chinese clause can be recognized as a group of characters.
9
10 Each Chinese character corresponds to a syllable; although some characters represent individual
11
12 (i.e., one-syllable) words, many words consist of more than one character. These linguistic
13
14 features make pre-processing necessary to ensure effective topic modeling and theorizing,
15
16 thereby enabling the algorithm to identify the tokens that comprise the texts.
17
18
19
20
21

22 The traditional content-analytical method of using pre-set dictionaries to match characters
23
24 with possible words in the corpus confronts computational problems, and the permutations and
25
26 ambiguities of language often lead to poor results. Customized dictionaries improve fit, but still
27
28 yield substantial inaccuracies (Allen et al., 2017; Slingerland, Nichols, Neilbo, & Logan, 2017).
29
30 Today, statistical and machine learning models are complementing, if not replacing, pre-set
31
32 dictionaries. These models build internal lists of words by training algorithms through iterative
33
34 learning. This training can be performed using extant language libraries (e.g., the People’s Daily
35
36 Language Library) to segment unknown texts.
37
38
39

40 The introduction and development of these methods has opened the door to employing
41
42 topic models to investigate a wide range of novel data sources and cultures. For example, Huang
43
44 et al. (2015) used topic modeling to analyze one of China’s biggest online social network
45
46 platforms, *Weibo*, to track the real-time ideation process of suicide, which is traditionally
47
48 assessed by surveys and interviews and thus suffers self-reporting and retrospective biases. Their
49
50 approach has shed new light on future studies of various ideation processes such as
51
52 entrepreneurial ideation.
53
54
55
56
57
58
59
60

1
2
3 Such word segmentation processes also make comparative analysis and theorization of
4 multiple-language corpora feasible. In particular, with appropriate pre-processing, topic models
5 can be used to analyze the diffusion and translation of new ideas, frames, and categories crossing
6 national borders. For example, the cross-national diffusion of CSR has attracted scholarly
7 attention (Kim & Bae, 2016; Lim & Tsutsui, 2012). But identifying the extent to which CSR has
8 been locally translated and innovated would require fine-grained analysis of multiple-language
9 corpora, which topic modeling can facilitate. Because the topic outputs from non-English
10 corpora must be translated into their English equivalents to be used in comparison and
11 theorization, and because the cultural context still matters for those identified topics, such
12 comparative projects are best developed by teams with at least one researcher who knows the
13 language and culture and can apply that knowledge to help validate the rendering of the corpora.
14
15
16
17
18
19
20
21
22
23
24
25
26
27

28 **Summary.** New trends in rendering corpora hold great promise for addressing the
29 technical and theoretical limitations of current topic modeling approaches. They show that
30 corpus selection as well as lemmatizing and other forms of corpus preparation have theoretical
31 implications, and therefore must be explicitly discussed in methods sections of papers, likely
32 under the aegis of “data pre-processing.” The use of foreign languages only magnifies these
33 challenges, just as they do in any form of archivalism applied to other cultures.
34
35
36
37
38
39
40
41
42
43
44

45 **Trends in Rendering Topics**

46
47 Researchers are continuing to refine how topics are rendered in an effort to manage the
48 degree of supervision required and how fit can be defined. In Figure 2, we show how the
49 rendering of topics revolves around the criteria for identifying robust, applicable topics (i.e.,
50 around supervision and fit criteria). Supervision and fitting, in turn, depend on the form of
51
52
53
54
55
56
57
58
59
60

1
2
3 theorizing taken—inductive, abductive, or deductive—with induction aligning with less
4 supervision and fitting than deduction.
5
6

7 **Integrating topic rendering with other approaches.** Many scholars today are finding
8 that topic modeling works best when integrated with other methods of analysis, which has
9 implications for the rendering of topics. One recent style of work covered by labels such as “big
10 qual” (Davidson, Edwards, Jamieson, & Weller, 2019) and “RiCH (Reader in Control of
11 Hermeneutics)” (Breiger, Wagner-Pacifici, & Mohr, 2018) gives the interpretive human reader
12 primacy, but leans on the affordances of computational tools for forming rich representations of
13 topics. Other styles in recent work integrate topic modeling with more traditional deductive
14 methods (Haans, 2019; Hankammer, Antons, Kleer, & Piller, 2016; Roberts et al., 2014), where
15 topics are rendered according to a logic of variable coherence. Topic modeling in these
16 correlational analyses seems to rely on a parsimony principle, where topics are presented in
17 papers as tables with applied labels and fewer than 10 highly associated words per topic (i.e.,
18 Schmiedel et al., 2018). Our reading of this trend reveals that the dominant method in the
19 research design affects how topics are rendered.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

37 Recent trends in topic modeling within management research have also shifted attention
38 toward alternate ways of capturing latent patterns to reveal new (sometimes provisional)
39 meaning structures that change over time. The LDA-based analyses we reviewed in this paper
40 mostly followed a pattern of rendering one set of topics in a corpus. Through iterative steps in
41 the rendering process, Hannigan et al. (2019) found that a key topic in a scandal’s media
42 coverage was changing due to the disclosure of a social control agent’s judgements of
43 wrongdoing. To overcome this challenge, they split their corpus in two, rendering topics across
44 each sub-corpus. They used the word-topic matrices from both models to find comparable topics,
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 which they subsequently used as independent variables representing media effects of a scandal in
4
5 event history models at different time periods. Similar efforts to periodize data can be seen in
6
7 work by Croidieu and Kim (2018). We see such efforts as contextualizing topics in ongoing
8
9 theoretical concerns.
10

11
12 As another example, Cho et al. (2017) embedded topic modeling with other commonly
13
14 used methods of conducting a literature review. The concept of topic was used to approximate an
15
16 “author community” of researchers exhibiting certain topics prominently in their work. This
17
18 framing affected the logic of how they rendered topics. They rendered latent author communities
19
20 using topic modeling against those derived using bibliometric network analysis to show
21
22 similarities and differences in approaches, but this comparison governed the validity of topics
23
24 rendered. Alternative analytical approaches that help generate theory (Bail, 2012; Kennedy,
25
26 2008), especially emergence processes, also promise the ability to better articulate latent patterns
27
28 to reveal hierarchical linguistic structure (Mohr et al., 2013). Therefore, the rendering of topics is
29
30 part of the overall theory generation process itself.
31
32
33

34
35 **Structural topic modeling.** Just as LDA disrupted latent semantic indexing (LSI),
36
37 scholars are attempting to modify LDA by improving fit algorithms and making it more
38
39 structured and systematic. One major development is structural topic modeling (STM) (Bail et
40
41 al., 2017; Roberts et al., 2014; Schmiedel et al., 2018), which extends LDA by incorporating
42
43 meta-data about documents, such as who wrote each text and when or where they were written.
44
45 This information can be re-applied to the topic estimation procedure and help improve model fit.
46
47 In so doing, STM enables researchers to identify relationships not just between topics and
48
49 documents, but also between the producers of documents and the texts and topics. It can be used
50
51 in a linear regression framework to analyze specific meta-data (as covariates) to identify
52
53
54
55
56
57
58
59
60

1
2
3 statistically significant relationships to each topic. It can also be used in mixed methods
4
5 approaches such as with critical discourse analysis to tie textual data analyzed using topic models
6
7 with richer qualitative analysis (Vaara, Aranda, Etchanchu, Guyt, & Sele, 2019).
8
9

10 In recent working papers appearing in Academy of Management Annual Meeting
11
12 Proceedings, researchers have adopted mixed STM approaches. For instance, Aggarwal, Lee,
13
14 and Hwang (2017) used topic modeling to operationalize review diversity in Yelp reviews to
15
16 show that status gains are correlated with higher-quality reviews and non-elite conformity to
17
18 those same reviews. Likewise, Karanovic, Berends, and Engel (2018) used topic modeling to
19
20 study actors' perceptions of "platform capitalism" (Davis, 2016) in a popular online forum for
21
22 Uber drivers. Their analysis reveals consistent patterns in a large corpus representing over
23
24 120,000 forum posts and shows that drivers' reactions can both contribute to and critically
25
26 evaluate the legitimacy of a new organizational form, despite being imposed from above.
27
28
29

30
31 **Hierarchical LDA.** Another promising extension to LDA topic modeling is hierarchical
32
33 LDA (hLDA) (Blei, Griffiths, & Jordan, 2010). While LDA traditionally requires that a
34
35 researcher set the number of topics (the k parameter), hLDA can generate the optimal number of
36
37 topics based on other researcher-defined parameters, such as the number of hierarchical levels
38
39 and number of terms per topic. While different software implementations of hLDA use different
40
41 algorithms to generate the hierarchical models, generally speaking, the hLDA algorithm
42
43 generates a set of sub-topics after identifying an aggregate topic. The algorithm then "reshuffles
44
45 the deck" by reclassifying documents or document segments into synthetic document groupings
46
47 and rerunning the algorithm for each grouping to generate additional sub-topics. The result is a
48
49 hierarchy representing the topics and sub-topics, or sub-dimensions, of the texts being analyzed.
50
51
52

53
54 The ability to generate a hierarchical representation of the internal structure of a
55
56
57
58
59
60

1
2
3 discourse can provide substantial theoretical insights. Tchalian, Glaser, Hannigan, and
4
5 Lounsbury (2019) are using hLDA to identify the competing and complementary messaging
6
7 efforts of stakeholders in the emergent electric vehicle (EV) industry: automobile manufacturers,
8
9 newspaper reporters, automotive experts, and government officials. The hierarchical structure of
10
11 the hLDA output is enabling Tchalian et al. (2019) to trace both the longitudinal appearance of
12
13 different topics involved with the construction of the emergent EV category and their
14
15 prominence within the discourse. This approach allows them to define the theoretical concept of
16
17 “institutional attention”—the field-level convergence that both isolates and aggregates the
18
19 various interests involved in the social construction of the EV as a market category. The
20
21 hierarchical arrangement of topics in their paper and others (e.g., Smith, Hawes, & Myers, 2014),
22
23 reveals not only the primacy of ideas over time, but also the socio-cognitive meaning structures
24
25 emphasized in cultural sociology (Mohr, 1998) and content analysis (Duriau et al., 2007), thus
26
27 highlighting the great potential of topic modeling approaches for generating novel theoretical
28
29 insights.
30
31
32
33
34

35 **Summary.** Advances in rendering topics have broadened topic modeling’s use by pairing
36
37 it with other techniques, and deepened its use by creating variants that structure topics (e.g.,
38
39 hLDA). Rendering topics, at least for the near future, appears sufficiently robust to work with
40
41 developments in near variants such as NLP and specific machine processing algorithms (i.e.,
42
43 “trained” algorithms in specific domains). These trends have the potential to extend the
44
45 theoretical deltas we identified in our analysis of management subject areas. However, applying
46
47 new algorithms for topic modeling and determining proper logics of fit and validity also raises
48
49 important questions about research design. For example, use of STM reinforces critical decisions
50
51 about appropriate measurement and variation in econometric based approaches, and hLDA
52
53
54
55
56
57
58
59
60

1
2
3 simply shifts a researcher's interpretive choices from determining the number of topics to
4
5 deciding the number of levels and words per topic. These advances demonstrate that the most
6
7 powerful path of development in topic modeling is not to displace, but rather complement
8
9 traditional research designs by enabling the use of different approaches to abstract and measure
10
11 phenomena using text.
12
13

14 15 16 17 **Trends in Rendering Theoretical Artifacts** 18

19 Trends in rendering theoretical artifacts may offer the richest, most open-ended area of
20
21 development in the field. Three trends are of particular interest: delineating latent structures,
22
23 mapping new meaning, and blending AI with human supervision to generate new artifacts. Each
24
25 trend has been pursued using a range of theorizing approaches from inductive to deductive, and
26
27 each has the ability to both extend and build theory, as indicated by the iterative arrows in Figure
28
29 2.
30
31
32

33 **Latent structures and the "new structuralism."** Increasingly, scholars are using topic
34
35 modeling to assess structural relations in fields (Bail, 2014; Jha & Beckman, 2017; McFarland et
36
37 al., 2013). Structural artifacts formed through rendering may enable theorists to identify new
38
39 mechanisms for uncovering organizational or institutional structures, including those flexible
40
41 enough to allow for a variety of instantiations in studies of fields (Lounsbury & Ventresca,
42
43 2003). The central thread relates to the use of topic modeling to map cultural dynamics around
44
45 social structures. A macro approach involves mapping the meaning structures that comprise
46
47 business environments (Pröllochs & Feuerriegel, 2018), knowledge profiles of firms (Suominen
48
49 et al., 2017), emerging fields (Hannigan & Casanovas, 2019), and political issues (Kim et al.,
50
51 2018). Researchers have modeled the topics and rhetorical attributes of scientific articles, in turn
52
53
54
55
56
57
58
59
60

1
2
3 finding links between the hidden topic structure of scientific communities as “thought
4 collectives” and impacts on knowledge consumption patterns (Antons et al., 2018). Others have
5 identified the “backstage” influences of stakeholder groups in the sustainability movement in
6 higher education and have used measures of discursive distance to identify field-level coherence
7 (Augustine & King, 2017).
8
9

10
11
12
13
14
15 More micro approaches involve modeling the formation of social network ties using
16 topic-based proximity measures (Lee, Qui, & Whinston, 2016), or tracking the signatures of
17 content authorship using author-topic models (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004).
18
19 Scholars are using these micro approaches to revisit a classic question in social science: How are
20 social structures and meanings co-constituted? Lee et al. (2016) considered the mechanism of
21 homophily in network formation by topic modeling texts of user-generated biographies and their
22 associated tweets. In turn, they found that people with similar topic vectors were more likely to
23 check-in to the same locations and form similar online social network ties. Rosen-Zvi et al.
24 (2004) used an extension to LDA to model the contents of documents and authors’ interests.
25 They created the “author-topic model” artifact which can be used to compare documents for
26 similarity and applied to automatically match paper authors to reviewers. In each of these papers,
27 researchers used topic modeling to render and theorize structural dimensions as artifacts.
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42
43 Scholars are extending the new structuralist approach by using topic modeling to analyze
44 dynamics of culture and meaning (Lounsbury & Glynn, 2019; Mohr & Bogdanov, 2013). The
45 simultaneous rendering of topics and contents of identified topic clusters reveals how social
46 structure and meanings can be co-constituted at the field level. An example of a classic approach
47 in this style of work is an exploration of “grass-fed beef” (Weber, Patel, & Heinze, 2013) as a
48 construct that conveys particular meanings and describes the evolving structure of a market.
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Topic modeling enables social structures and meanings to be studied in new ways. Hannigan and
4 Casasnovas (2019) used topic modeling and Named Entity Recognition to map the co-occurrence
5 of actors and topics appearing in media coverage to identify the spatial and temporal
6 arrangements of an emerging field. Following classic works in the new structuralist tradition
7 (i.e., Mohr & Duquenne, 1997), Hannigan and Casasnovas created incidence matrices of topic
8 and actor co-occurrence and used them to generate maps of hierarchical Galois lattice structures.
9 These lattice artifacts are visual maps that demonstrate co-constitution by showing the nesting of
10 substructures formed through two modes of analysis. Mohr and Duquenne (1997) used lattices to
11 show how practices and meanings co-constituted institutional logics, whereas Hannigan and
12 Casasnovas (2019) used lattices to reveal the types of actors and topics co-constituting spatial
13 and temporal arrangements in field formation. Advances in relational topic modeling (RTM)
14 (Chang & Blei, 2009; Gerlach, Peixoto, & Altmann, 2018) that identify document networks are
15 also being used to render more document-based theoretical artifacts, perhaps representing
16 different audience perspectives. These audience perspectives, including those captured using
17 STM, enable latent structures among knowledge creators to be identified.

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38 **Bringing back meaning.** Whilst topic modeling provides tools for extracting and
39 presenting constellations of words and phrases that appear in patterns across documents in
40 corpora, the question of whether such topics represent *meaning structures* is an important one
41 (Mohr, 1998). During the initial analytical stage, analysts interpret topics based on logics of fit
42 and interpretability. However, presenting topics without careful concern for theoretical artifacts
43 risks presenting disembodied arguments about meaning. Thus, a naive machine learning analysis
44 may omit important distinctions if applied crudely. An important topic modeling trend thus
45 centers on how to capture meaning and meaning structures.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Organizational scholars have long been interested in studying meanings, particularly in
4 light of recent concerns about measuring the construction and deployment of culture (i.e.,
5
6 Gehman & Soublière, 2017; Lounsbury & Glynn, 2019; Weber & Dacin, 2011). Whilst topic
7
8 modeling-based research promises the potential to study cultural dynamics with increased scale
9
10 and precision, scholars acknowledge that the technique must be paried with a respect for
11
12 symbolic and social boundaries (Lounsbury & Glynn, 2019; Mohr et al., 2013). For example,
13
14 Mohr et al. (2013) pointed to Burke’s (1945) classic analytical structure of the *pentad* to study
15
16 scenes of action. They used topic modeling and NLP to study the pentad in a corpus of U.S.
17
18 national security documents. Analytically, they used named entity recognition to map *actors*,
19
20 topic modeling to identify *scenes*, and NLP-based semantic grammar parsers to identify *acts*.
21
22
23
24
25
26 Other scholars have described the utility of applying related computational methods such as
27
28 semantic network analysis to contextualize topic modeling through theoretical artifacts (Carley
29
30 & Kaufer, 1993; Diesner & Carley, 2005). Combined with a concern for theoretical artifacts,
31
32 topic modeling thus opens the door to rendering modes of meaning, such as observing
33
34 connotations and denotations of an institutional field.
35
36

37
38 **Blending topic modeling and AI.** A third fertile area of enhancing the theoretical
39
40 artifacts built with topic modeling lies at the intersection of artifacts derived from artificial
41
42 intelligence (AI) and those derived from topic model rendering. AI and the deep learning models
43
44 on which it is built can be blended with topic models in at least two ways. First, in the class of AI
45
46 models known as “deep neural networks,” two relevant methods enable blending with topic
47
48 modeling: convolutional neural network (CNN) methods and recurrent neural network (RNN)
49
50 methods. Unlike machine learning models such as LDA that use minimal inferences about
51
52 context, these models retain more contextual information and thus are becoming increasingly
53
54
55
56
57
58
59
60

1
2
3 relevant for social science researchers. They are more appropriate for dealing with streaming
4 data such as Facebook updates and Amazon reviews, in which local contexts (e.g., prior words in
5 a word sequence) affect the position of each topic term (Jin, Luo, Zhu, & Zhuo, 2018).
6
7

8
9
10 Combining these methods with topic models may enable a more complex and dynamic rendering
11 of theoretical artifacts such as frames, logics and the latent value orientations discussed above.
12

13
14 When applied to large text corpora, both CNN and RNN are particularly effective in managing
15 the tradeoff of specificity, enabling the analysis and modeling of latent structures that better
16 balance under- and over-fitting. Moreover, they may help generate entirely new theoretical
17 artifacts to help identify and explain social and role structure, partisanship, ideological
18 contestation, discursive fields, and other socio-cultural structures and institutional regimes more
19 dynamically.
20
21
22
23
24
25
26
27

28
29 Second, deep learning can be integrated with topic models to analyze images—alone or
30 along with verbal text—which opens a new path to rendering theoretical artifacts. Whereas
31 verbal text is descriptive, linear, additive and temporal, images and visual features are embodied,
32 spatial, holistic and simultaneous, which defies conventional analytical techniques. The
33 integration of deep learning into topic models creates potential for future theoretical development
34 that considers both visual features and verbal text (Krizhevsky, Sutskever, & Hinton, 2012). In
35 particular, scholars have argued that the role of visual features in the process of
36 institutionalization is significant, but largely under-examined (Meyer, Jancsary, Höllerer, &
37 Boxenbaum, 2017).
38
39
40
41
42
43
44
45
46
47
48

49 In other words, deep learning helps manage tradeoffs around specificity and
50 configuration, and represents an effective solution to the ever-present issue of theoretical
51 parsimony, but it also comes with a caution. Because deep learning is a computationally
52
53
54
55
56
57
58
59
60

1
2
3 inductive modeling tool, many of its operationalizations are “black boxed,” making its feature
4 permutations challenging to reconstruct mathematically. It ironically highlights the tradeoff of
5 human supervision and reinforces the need to apply it along with other analytical techniques
6 within a mixed-methods approach to generating theoretical artifacts.
7
8
9
10

11
12 **Summary.** All three new trends in topic modeling—eliciting latent structures, capturing
13 meaning, and using AI to help generate theoretical artifacts—open up new avenues for theory
14 building. They complement the agnostic assumptions about meaning that are embedded in the
15 LDA algorithm and, in this way, echo how trends related to corpora selection and trimming and
16 to supervising and fitting topics are helping scholars overcome some of topic modeling’s foibles
17 while preserving its power. In particular, by revealing latent patterns and meaning structures,
18 topic modeling is increasingly able to generate social, cultural, and political constructs that
19 define evolving cultural meanings, discursive fields, and political action.
20
21
22
23
24
25
26
27
28
29
30

31 32 33 **FROM THE BALCONY**

34
35 Topic modeling, a method adapted from computer science, “represents a novel tool for
36 analyzing large collections of qualitative data in a scalable and reproducible way” (Schmiedel et
37 al., 2018, p. 3; see also Kobayashi et al., 2018). Our review reveals that topic modeling has been
38 used in surprisingly diverse ways by management scholars, demonstrating that it is a malleable
39 methodological and theoretical tool for tackling a variety of research questions. Although many
40 papers we examined described the technical underpinnings of the LDA algorithm, we found that
41 topic modeling practices are part of an often-implicit process of *rendering* corpora, topic models,
42 and theoretical artifacts from raw data. We applied topic model rendering in this review to curate
43 and make sense of the topic modeling corpus in the management literature. Our analysis reveals
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 that topic modeling is gaining steam in management research (see Figure 1), particularly in five
4 areas: detecting novelty and emergence, developing inductive classification systems,
5
6 understanding online audiences and products, analyzing frames and social movements, and
7
8 understanding cultural dynamics. Topic modeling has both strengthened knowledge in each area
9
10 and enabled scholars to explore subjects in new ways. The current trends in rendering with topic
11
12 modeling have only increased the value added by the technique. We now wish to briefly consider
13
14 the topic modeling field in management research from a broader perspective, touching on
15
16 important challenges and debates that will shape the direction of research and the evolution of
17
18 the domain.
19
20
21
22
23
24
25

26 **Challenges and Debates**

27
28 Perhaps the biggest challenge in the near future stems from how topic modeling has
29
30 helped open the door to a plethora of work based on the quantitative structural study of meaning
31
32 (Mohr, 1998; Ventresca & Mohr, 2002). Emergent classification systems based on meaning
33
34 structures, such as those we have examined in topic modeling research, provide a reflexive
35
36 contrast to others recognized and used to parse meaning in materialized structures, such as patent
37
38 classification, risk typification, and industrial categorization. In this sense, we see management
39
40 moving in a direction that reflects current trends in cultural sociology, political science, and
41
42 linguistics; a machine learning approach like topic modeling can reveal shared cultural meanings
43
44 that in turn can be integrated into the analytical process alongside traditional socio-cultural
45
46 variables and constructs. Our identified trends in topic modeling reveal that this integration is
47
48 indeed occurring. Thus, topic modeling is *not necessarily disrupting or displacing existing*
49
50 *methods, so much as augmenting and extending them.*
51
52
53
54
55
56
57
58
59
60

1
2
3 By highlighting the different modes of studying meaning (Mohr et al., 2013), we also
4 acknowledge to the views of semiotics and qualitatively-oriented scholars who have long
5 recognized that meanings are grounded in practice and take on different levels of ambiguity. In
6 the debates around semiotics and modeling, it is important to recognize that topic modeling
7 combines the poetic (or connotative) with the semantic (or denotative) meanings of words in
8 topics and subjects; although the words in “bags” are independent, they are combined in
9 proximity and recognized in context. Integrating machine reading within studies of meaning
10 necessitates a discussion around the trade-offs of standardizing content and linking to theoretical
11 artifacts. This also highlights that topic modeling practice in management is a deeply theoretical
12 endeavor. Now that topic modeling algorithms are becoming more readily available through
13 toolkits in R, Python, and other open source software, we worry that topic modeling risks being
14 pigeon-holed as an LDA algorithm and “black boxed” as just another textual analysis technique.
15 By attending to the rendering process, we hope we have helped scholars understand the choices
16 inherent in the creation and pre-processing of corpora, the parameters used in the topic models
17 themselves, and in the creation of theoretical artifacts from the analysis. Indeed, by articulating
18 the rendering process, we have highlighted how topic modeling using machine learning
19 algorithms actually foregrounds analysts’ interpretive decisions and theory work.

20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Ultimately, theory is paramount for grounding claims around meaning. Our review has emphasized that incorporating topic modeling in a theoretical manner entails careful engagement with the cultural ecology of a social space. Our definition of the rendering process was created along these lines; particularly when employing topic modeling to study the meanings of a social space, one cannot neglect its structural foundations. The ecology imagery evokes connotations of a structured space, contoured by theoretical concerns of social structure, such as boundaries,

1
2
3 stratification, and reputations of actors. This also invokes the imagery by philosophers of science
4
5 in assemblage theory, where a socio-cultural ecology is constituted by relationships formed
6
7 through processes of encoding meanings, such as stratification and territory (DeLanda, 2006).
8
9

10 The assemblage theory approach to conceptualizing knowledge-based fields is relevant to
11
12 our consideration of the researcher generating knowledge alongside algorithms with machine
13
14 learning. Such work is not performed by the human or the machine alone; rather, it is a combined
15
16 effort. We reflect on how assemblage theory has illustrated the institution of science operating
17
18 against the backdrop of two ideal styles of action—"nomadic" versus "state"—where the former
19
20 is paradigm breaking and smooth, concerned with variation and problematization, and the latter
21
22 is striated and contoured, concerned with precision and advances in structured fields of
23
24 knowledge (Jensen & Rødje, 2010; Deleuze & Guattari, 1987). Machine learning approaches
25
26 that are not configured with contextual structural knowledge may be nomadic—that is, overly
27
28 fluid and rendering meaning structures across fields, only looking for what is statistically
29
30 significant, but not necessarily socially or culturally significant. Understanding these ideal
31
32 "nomadic" and "state" approaches to scientific endeavors can help us understand the ideal types
33
34 of machine learning reading (nomadic: naive, fast, fluid, distant) and human-only reading (state:
35
36 careful, slow, narrowly focused, deep). Our hope is that by delineating the rendering process, we
37
38 are striking a middle ground between the two; in reflexively using machine learning tools in this
39
40 manner, the analyst can see possibilities (latent meaning structures) against materialized social
41
42 structures (formal classification systems).
43
44
45
46
47
48

49 To render meaning in this manner is to engender engagement with data, where the
50
51 researcher zooms in and zooms out based on distant reading (Moretti, 2013) and representations
52
53 of meaning structures. By conceptualizing topic modeling as part of a rendering process, we
54
55
56
57
58
59
60

1
2
3 hope that we have also avoided the fear that social science researchers are just “squeezing [their]
4 unstructured texts, sounds, or images into some special-purpose data model” (Underwood, 2015,
5 p. 1). Instead, researchers employ rendering processes for topic modeling as a “discovery
6 strategy” to infer meaning. This blending of formal analytical methodologies with an interpretive
7 focus helps reveal meanings and is echoed in an emerging stream of work in organizational
8 theory that Ventresca and Mohr (2002) labeled “new archivalism.”
9

10
11
12
13
14
15
16
17 Nevertheless, one challenge remains: as topic modeling has diffused into management
18 research, the practices for applying it have not remained static. Indeed, by adapting this method,
19 management scholars have contributed the rendering process itself. We see this contribution as
20 being aligned with movements that draw upon formal methods to generate representations of
21 meanings, which can then be analyzed in a plethora of ways (Brieger et al., 2018; Davidson et
22 al., 2019; Ventresca & Mohr, 2002). We found that many authors did indeed use computational
23 modeling tools in a manner similar what Ventresca and Mohr described in 2002; however, we
24 also found that the process of rendering goes further, particularly as it relates to rendering
25 meanings. In our opinion, topic modeling tends to naturally ally more with mixed approaches to
26 studying text (Brieger et al., 2018; Davidson et al., 2019; Ventresca & Mohr, 2002). Moreover,
27 because meaning schema (i.e., dictionaries, coding categories, etc.) are rejected *a priori*, the
28 technique often seems to be more inductive in nature.
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

44
45 Of course, this is by no means the only mode of theorizing enabled through topic
46 modeling. Other work has been more abductive in nature. For example, Fligstein et al.’s (2017)
47 frame analysis helps explain how the Federal Open Market Committee underestimated the risks
48 to the economy leading up to the 2008 financial crisis; their research design enabled them to use
49 topic modeling to connect hypotheses to texts via a combination of qualitative and quantitative
50
51
52
53
54
55
56
57
58
59
60

1
2
3 techniques. Indeed, topic modeling has also been used with partially deductive forms of
4
5 theorizing (e.g., see Haans, 2019; Kaplan & Vakili, 2015).
6

7
8 As a final, cumulative point, we think that the flexibility of topic modeling—its utility in
9
10 creating corpora, its ability to be paired with different quantitative and qualitative methods, and
11
12 its applicability in variety of theoretical approaches—underpins its power and promise for
13
14 management research. By surfacing topic modeling’s flexibility, we hope our detailed
15
16 exploration of the rendering process has persuaded the reader, at least to some extent, to consider
17
18 engaging with topic modeling in order to build new management theory.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

REFERENCES

- Abbott, A. (1988). *The system of professions: An essay on the division of expert labor* (1st ed.). Chicago, IL: University of Chicago Press.
- Abrahamson, E. (1996). Management fashion. *Academy of Management Review*, 21(1), 254–285.
- Abrahamson, E., & Fairchild, G. (1999). Management fashion: Lifecycles, triggers, and collective learning processes. *Administrative Science Quarterly*, 44(4), 708–740.
- Aggarwal, V., Lee, M. K., & Hwang, E. (2017). Status gains and subsequent effects on evaluations. *Academy of Management Proceedings*, 2017(1), 16355.
- Ahonen, P. (2015). Institutionalizing big data methods in social and political research. *Big Data & Society*, 2(2), 1–12.
- Al Sumait, L., Barbará, D., Gentle, J., & Domeniconi, C. 2009. Topic significance ranking of LDA generative models. In *Machine learning and knowledge discovery in databases*, vol. 5781(pp. 67–82). Berlin, Heidelberg: Springer.
- Allen, C., Luo, H., Murdock, J., Pu, J., Wang, X., Zhai, Y., & Zhao, K. (2017). *Topic modeling the Han dian ancient classics*. ArXiv preprint arXiv:1702.00860. Retrieved from <https://arxiv.org/abs/1702.00860>
- Almquist, Z. W., & Bagozzi, B. E. (2017). Using radical environmentalist texts to uncover network structure and network features. *Sociological Methods & Research*. <https://doi.org/10.1177/0049124117729696>
- Alvesson, M., & Kärreman, D. (2000). Taking the linguistic turn in organizational research: Challenges, responses, consequences. *Journal of Applied Behavioral Science*, 36(2), 136–158.
- Alvesson, M., & Kärreman, D. (2000). Varieties of discourse: On the study of organizations through discourse analysis. *Human Relations*, 53(9), 1125–1149.
- Anand, N., & Peterson, R. A. (2000). When market information constitutes fields: Sensemaking of markets in the commercial music industry. *Organization Science*, 11(3), 270–284.
- Antons, D., Joshi, A. M., & Salge, T. O. (2018). Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management*. <https://doi.org/10.1177/0149206318774619>
- Antons, D., Kleer, R., & Salge, T. O. (2016). Mapping the topic landscape of JPIM, 1984–2013: In search of hidden structures and development trajectories. *Journal of Product Innovation Management*, 33(6), 726–749.
- Argote, L., & Greve, H. R. (2007). “A behavioral theory of the firm”: 40 years and counting: Introduction and impact. *Organization Science*, 18(3), 337–349.
- Arora, A. (1995). Licensing tacit knowledge: Intellectual property rights and the market for know-how. *Economics of Innovation and New Technology*, 4(1), 41–60.
- Arora, A., Gittelman, M., Kaplan, S., Lynch, J., Mitchell, W., & Siggelkow, N. (2016). Question-based innovations in strategy research methods. *Strategic Management Journal*, 37(1), 3–9.
- Augustine, G., & King, B. G. (2017). Behind the scenes: A backstage look at field formation within sustainability in higher education. *Academy of Management Proceedings*, 2017(1), 15788.

- 1
2
3 Azzopardi, L., Girolami, M., & van Risjbergen, K. (2003). *Investigating the relationship*
4 *between language model perplexity and IR precision-recall measures*. Presented at the
5 26th Annual International ACM SIGIR Conference on Research and Development in
6 Information Retrieval. Association for Computing Machinery. Toronto, Ontario.
- 7
8 Bail, C. A. (2012). The fringe effect: Civil society organizations and the evolution of media
9 discourse about Islam since the September 11th attacks. *American Sociological Review*,
10 77(6), 855–879.
- 11
12 Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and*
13 *Society*, 43(3–4), 465–482.
- 14
15 Bail, C. A., Brown, T. W., & Mann, M. (2017). Channeling hearts and minds: Advocacy
16 organizations, Cognitive-emotional currents, and public conversation. *American*
17 *Sociological Review*, 82(6), 1188–1213.
- 18
19 Bakhtin, M. M. (1982). *The dialogic imagination: Four essays*. Austin, TX: University of Texas
20 Press.
- 21
22 Ballinger, G. A., & Rockmann, K. W. (2010). Chutes versus ladders: Anchoring events and a
23 punctuated-equilibrium perspective on social exchange relationships. *Academy of*
24 *Management Review*, 35(3), 373–391.
- 25
26 Bansal, P., & Corley, K. (2011). The coming of age for qualitative research: Embracing the
27 diversity of qualitative methods. *Academy of Management Journal*, 54(2), 233–237.
- 28
29 Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual
30 risk disclosures. *Management Science*, 60(6), 1371–1391.
- 31
32 Barley, S. R. (1983). Semiotics and the study of occupational and organizational cultures.
33 *Administrative Science Quarterly*, 28(3), 393–413.
- 34
35 Barry, D. (1997). Telling changes: From narrative family therapy to organizational change and
36 development. *Journal of Organizational Change Management*, 10(1), 30–46.
- 37
38 Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded
39 theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the*
40 *Association for Information Science and Technology*, 68(6), 1397–1410.
- 41
42 Bendle, N. T., & Wang, X. (2016). Uncovering the message from the mess of big data. *Business*
43 *Horizons*, 59(1), 115–124.
- 44
45 Benford, R. D., & Snow, D. A. (2000). Framing processes and social movements: An overview
46 and assessment. *Annual Review of Sociology*, 26, 611–639.
- 47
48 Bennardo, G., & de Munck, V. C. (2014). *Cultural models: Genesis, methods, and experiences*.
49 Oxford, UK: Oxford University Press.
- 50
51 Berelson, B. (1952). *Content analysis in communication research*. New York, NY: Free Press.
- 52
53 Blanchard, S. J., Aloise, D., & Desarbo, W. S. (2017). Extracting summary piles from sorting
54 task data. *Journal of Marketing Research*, 54(3), 398–414.
- 55
56 Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- 57
58 Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and
59 Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2), 7:1–
60 7:30.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied*
Statistics, 1(1), 17–35.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine*
Learning Research, 3(Jan), 993–1022.

- 1
2
3 Boden, D., & Zimmerman, D. H. (Eds.). (1991). *Talk and social structure: Studies in*
4 *ethnomethodology and conversation analysis*. Berkeley, CA: University of California
5 Press.
- 6 Boje, D. M. (1995). Stories of the storytelling organization: A postmodern analysis of Disney as
7 “Tamara-Land.” *Academy of Management Journal*, 38(4), 997–1035.
- 8 Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*.
9 Cambridge, MA: MIT Press.
- 10 Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer &
11 G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York, NY:
12 Academic Press.
- 13 Brannen, M. Y. (2004). When Mickey loses face: Recontextualization, semantic fit, and the
14 semiotics of foreignness. *Academy of Management Review*, 29(4), 593–616.
- 15 Breiger, R. L., Wagner-Pacifici, R., & Mohr, J. W. (2018). Capturing distinctions while mining
16 text data: Toward low-tech formalization for text analysis. *Poetics*, 68, 104–119.
- 17 Bucheli, M., & Wadhvani, R. D. (Eds.). (2014). *Organizations in time: History, theory,*
18 *methods*. Oxford, UK: Oxford University Press.
- 19 Burke, K., 1945. *A grammar of motives*. Berkeley, CA: University of California Press.
- 20 Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews.
21 *Marketing Science*, 35(6), 953–975.
- 22 Buurma, R. S. (2015). The fictionality of topic modeling: Machine reading Anthony Trollope’s
23 Barsetshire series. *Big Data & Society*. <https://doi.org/10.1177/2053951715610591>
- 24 Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H., & Steele, L. B. (2014). The information
25 content of mandatory risk factor disclosures in corporate filings. *Review of Accounting*
26 *Studies*, 19(1), 396–455.
- 27 Cao, L., & Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation
28 and classification of objects and scenes. In *Computer Vision, 2007. IEEE 11th*
29 *International Conference on ICCV* (pp. 1-8). Piscataway, NJ: IEEE.
- 30 Carley, K.M., Kaufer, D., 1993. Semantic connectivity: An approach for analyzing semantic
31 networks. *Communication Theory*, 3(3), 183–213.
- 32 Chang, J., & Blei, D. (2009). Relational topic models for document networks. In *Artificial*
33 *intelligence and statistics* (pp. 81–88). Retrieved from
34 <http://proceedings.mlr.press/v5/chang09a.html>
- 35 Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves:
36 How humans interpret topic models. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I.
37 Williams, & A. Culotta (Eds.), *Advances in neural information processing systems*, vol.
38 22. Retrieved from [https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-](https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf)
39 [interpret-topic-models.pdf](https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf)
- 40 Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage.
- 41 Cho, Y.-J., Fu, P.-W., & Wu, C.-C. (2017). Popular research topics in marketing journals, 1995–
42 2014. *Journal of Interactive Marketing*, 40, 52–72.
- 43 Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on*
44 *Information Theory*, 2(3), 113–124.
- 45 Corley, K. G., & Gioia, D. A. (2004). Identity ambiguity and change in the wake of a corporate
46 spin-off. *Administrative Science Quarterly*, 49(2), 173–208.
- 47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Cornelissen, J. P., Durand, R., Fiss, P. C., Lammers, J. C., & Vaara, E. (2015). Putting
4 communication front and center in institutional theory and analysis. *Academy of*
5 *Management Review*, 40(1), 10-27.
- 6
7 Croidieu, G., & Kim, P. H. (2018). Labor of love: Amateurs and lay-expertise legitimation in the
8 early U.S. radio field. *Administrative Science Quarterly*, 63(1), 1–42.
- 9
10 Crossley, S. A., Dascalu, M., & McNamarac, D. S. (2017). How important is size? An
11 investigation of corpus size and meaning in both latent semantic analysis and latent
12 Dirichlet allocation. In *30th International Florida Artificial Intelligence Research Society*
13 *Conference, FLAIRS 2017*. Menlo Park, CA: AAAI Press.
- 14
15 Dalpiaz, E., Rindova, V., & Ravasi, D. (2016). Combining logics to transform organizational
16 agency: Blending industry and art at Alessi. *Administrative Science Quarterly*, 61(3),
17 347–392.
- 18
19 Davidson, E., Edwards, R., Jamieson, L., & Weller, S. (2019). Big data, qualitative style: A
20 breadth-and-depth method for working with large amounts of secondary qualitative data.
21 *Quality & Quantity*, 53(1), 363-376.
- 22
23 Davis, G. F. (2016). Can an economy survive without corporations? Technology and robust
24 organizational alternatives. *Academy of Management Perspectives*, 30(2), 129–140.
- 25
26 Deephouse, D. L. (1999). To be different, or to be the same? It's a question (and theory) of
27 strategic balance. *Strategic Management Journal*, 20(2), 147–166.
- 28
29 Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing
30 by latent semantic analysis. *Journal of the American Society for Information Science*,
31 41(6), 391.
- 32
33 DeLanda, M. (2006). *A new philosophy of society: Assemblage theory and social complexity*.
34 London, UK: A&C Black.
- 35
36 Deleuze, G., & Guattari, F. (1987). *A Thousand Plateaus*. New York, NY: Continuum.
- 37
38 Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research*. Thousand Oaks, CA:
39 Sage.
- 40
41 Denzin, N. K., & Lincoln, Y. S. (2011). *The SAGE handbook of qualitative research*. Thousand
42 Oaks, CA: Sage.
- 43
44 Dey, I. (1995). Reducing fragmentation in qualitative research. In U. Kelle (Ed.), *Computer-*
45 *aided qualitative data analysis: Theory, methods, and practice* (pp. 69–79). London:
46 Sage.
- 47
48 Diesner, J., Carley, K.M. (2005). Revealing social structure from texts: Meta-matrix text analysis
49 as a novel method for network text analysis. In V. K. Narayanan & D. J. Armstrong
50 (Eds.), *Causal mapping for research in information technology* (pp. 81–108). Hershey,
51 PA: IGI Global.
- 52
53 Dilthey, W., & Jameson, F. (1972). The rise of hermeneutics. *New Literary History*, 3(2), 229–
54 244.
- 55
56 DiMaggio, P. J. (1997). Culture and cognition. *Annual Review of Sociology*, 23, 263–287.
- 57
58 DiMaggio, P. J. (2015). Adapting computational text analysis to social science (and vice versa).
59 *Big Data & Society*, 2(2), 2053951715602908.
- 60
61 DiMaggio, P. J., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and
62 the sociological perspective on culture: Application to newspaper coverage of U.S.
63 government arts funding. *Poetics*, 41(6), 570–606.

- 1
2
3 DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the
4 sociological perspective on culture: Application to newspaper coverage of U.S.
5 government arts funding. *Poetics*, *41*(6), 570–606.
- 6 Durand, R., & Khaire, M. (2017). Where do market categories come from and how?
7 Distinguishing category creation from category emergence. *Journal of Management*,
8 *43*(1), 87–110.
- 9 Durand, R., & Paoletta, L. (2013). Category stretching: Reorienting research on categories in
10 strategy, entrepreneurship, and organization theory. *Journal of Management*
11 *Studies*, *50*(6), 1100–1123.
- 12 Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis
13 literature in organization studies: Research themes, data sources, and methodological
14 refinements. *Organizational Research Methods*, *10*(1), 5–34.
- 15 Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure:
16 Evidence from latent Dirichlet allocation. *Journal of Accounting and Economics*, *64*(2),
17 221–245. <https://doi.org/10.1016/j.jacceco.2017.07.002>
- 18 Edmondson, A. C., & Mcmanus, S. E. (2007). Methodological fit in management field research.
19 *Academy of Management Review*, *32*(4), 1246–1264.
- 20 Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management*
21 *Review*, *14*(4), 532–550.
- 22 Etter, M., Ravasi, D., & Colleoni, E. (2018). Social media and the formation of organizational
23 reputation. *Academy of Management Review*.
- 24 Evans, J. A., & Aceves, P. (2016). Machine translation: Mining text for social theory. *Annual*
25 *Review of Sociology*, *42*(1), 21–50.
- 26 Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds.
27 *Journal of Financial Economics*, *33*(1), 3–56.
- 28 Firth, J. R. (1957). *Papers in linguistics*. London, UK: Oxford University Press.
- 29 Fiss, P. C. (2007). A set-theoretic approach to organizational configurations. *Academy of*
30 *Management Review*, *32*(4), 1180–1198.
- 31 Fiss, P. C., & Hirsch, P. M. (2005). The discourse of globalization: Framing and sensemaking of
32 an emerging concept. *American Sociological Review*, *70*(1), 29–52.
- 33 Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*,
34 *47*(1), 117–132.
- 35 Fligstein, N., & McAdam, D. (2011). Toward a general theory of strategic action fields.
36 *Sociological Theory*, *29*(1), 1–26.
- 37 Fligstein, N., Stuart Brundage, J., & Schultz, M. (2017). Seeing like the Fed: Culture, cognition,
38 and framing in the failure to anticipate the financial crisis of 2008. *American Sociological*
39 *Review*, *82*(5), 879–909.
- 40 Forbes, D. P., & Kirsch, D. A. (2011). The study of emerging industries: Recognizing and
41 responding to some central problems. *Journal of Business Venturing*, *26*, 589–602.
- 42 Gehman, J., Glaser, V. L., Eisenhardt, K. M., Gioia, D., Langley, A., & Corley, K. G. (2018).
43 Finding theory-method fit: A comparison of three qualitative approaches to theory
44 building. *Journal of Management Inquiry*, *27*(3), 284–300.
- 45 Gehman, J., & Soublière, J.-F. (2017). Cultural entrepreneurship: From making culture to
46 cultural making. *Innovation*, *19*(1), 61–73.
- 47 Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models.
48 *Science Advances*, *4*(7), eaaq1360.
- 49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Gioia, D. A., & Chittipeddi, K. (1991). Sensemaking and sensegiving in strategic change
4 initiation. *Strategic Management Journal*, *12*(6), 433–448.
- 5 Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive
6 research: Notes on the Gioia methodology. *Organizational Research Methods*, *16*(1), 15–
7 31.
- 8
9 Gioia, D. A., & Thomas, J. B. (1996). Identity, image, and issue interpretation: Sensemaking
10 during strategic change in academia. *Administrative Science Quarterly*, *41*(3), 370–403.
- 11 Giorgi, S., & Weber, K. (2015). Marks of distinction: Framing and audience appreciation in the
12 context of investment advice. *Administrative Science Quarterly*, *60*(2), 333–367.
- 13 Glaser, B. G., & Strauss, A. L. (1967). *Discovery of grounded theory: Strategies for qualitative*
14 *research*. New York, NY: Routledge.
- 15
16 Glaser, V., Fiss, P. C., & Kennedy, M. T. (2011). Rhetoric and resonance: Framing strategies for
17 institutionalizing new market conceptions. *Academy of Management Proceedings*,
18 *2011*(1), 1–6.
- 19 Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge,
20 MA: Harvard University Press.
- 21 Greve, H. R. (2003). *Organizational learning from performance feedback: A behavioral*
22 *perspective on innovation and change*. Cambridge, UK: Cambridge University Press.
- 23 Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic
24 content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.
- 25 Guerreiro, J., Rita, P., & Trigueiros, D. (2016). A text mining-based review of cause-related
26 marketing literature. *Journal of Business Ethics*, *139*(1), 111–128.
- 27 Guo, L., Sharma, R., Yin, L., Lu, R., & Rong, K. (2017). Automated competitor analysis using
28 big data analytics: Evidence from the fitness mobile app business. *Business Process*
29 *Management Journal*, *23*(3), 735–762.
- 30
31 Haans, R. F. J. (2019). What's the value of being different when everyone is? The effects of
32 distinctiveness on performance in homogeneous versus heterogeneous categories.
33 *Strategic Management Journal*, *40*(1), 3–27.
- 34 Hankammer, S., Antons, D., Kleer, R., & Piller, F. (2016). Researching mass customization:
35 Mapping hidden structures and development trajectories. In *Academy of Management*
36 *Proceedings*, *2016*(1), 10900.
- 37 Hannan, M. T., & Carroll, G. R. (1992). *Dynamics of organizational populations: Density,*
38 *legitimation, and competition*. Oxford, UK: Oxford University Press.
- 39 Hannan, M. T., & Freeman, J. (1977). The population ecology of organizations. *American*
40 *Journal of Sociology*, *82*(5), 929–964.
- 41 Hannan, M. T., Pólos, L., & Carroll, G. R. (2007). *Logics of organization theory: Audiences,*
42 *codes, and ecologies*. Princeton, NJ: Princeton University Press.
- 43 Hannigan, T. R. & Casanovas, G. (2019). New structuralism and field emergence: The co-
44 constitution of meanings and actors in the early moments of impact investing. *Research*
45 *in the Sociology of Organizations*.
- 46 Hannigan, T. R., Bundy, J., Graffin, S. D., Wade, J. B., & Porac, J. (2015). The social
47 construction of scandal: The role of media in the British parliamentary expense affair.
48 *Academy of Management Proceedings*, *2015*(1), 14966.
- 49 Hannigan, T. R., Porac, J. F., Bundy, J., Wade, J. B., & Graffin, S. D. (2019). *Crossing the line*
50 *or creating the line: Media effects in the 2009 British MP expense scandal*. Working
51 paper.
52
53
54
55
56
57
58
59
60

- 1
2
3 Hardy, C. (2001). Researching organizational discourse. *International Studies of Management &*
4 *Organization*, 31(3), 25–47.
- 5 Hatch, M. J. (1993). The dynamics of organizational culture. *Academy of Management Review*,
6 18(4), 657–693.
- 7
8 Hatch, M. J., & Schultz, M. (2017). Toward a theory of using history authentically: Historicizing
9 in the Carlsberg Group. *Administrative Science Quarterly*, 62(4), 657–697.
- 10 Heugens, P. P., & Lander, M. W. (2009). Structure! Agency! (and other quarrels): A meta-
11 analysis of institutional theories of organization. *Academy of Management Journal*,
12 52(1), 61–85.
- 13 Houghton, J. P., Siegel, M., Madnick, S., Tounaka, N., Nakamura, K., Sugiyama, T., Nakagawa,
14 D., & Shirnen, B. (2017). Beyond keywords: Tracking the evolution of conversational
15 clusters in social media. *Sociological Methods & Research*.
16 <https://doi.org/10.1177/0049124117729705>
- 17 Hu, D. J., & Saul, L. K. (2009). A probabilistic topic model for music analysis. In *Proceedings of*
18 *NIPS*, vol. 9 (pp. 1–4). Retrieved from
19 http://cseweb.ucsd.edu/~dhu/docs/nips09_abstract.pdf
- 20 Huang, A. H., Lehavey, R., Zang, A. Y., & Zheng, R. (2017). Analyst information discovery and
21 interpretation roles: A topic modeling approach. *Management Science*, 64(6), 2833–2855.
- 22 Huang, X., Li, X., Zhang, L., Liu, T., Chiu, D., & Zhu, T. (2015). Topic model for identifying
23 suicidal ideation in Chinese microblog. In *29th Pacific Asia Conference on Language,*
24 *Information and Computation* (pp. 553–562).
- 25 Humphreys, A., & Wang, R. J.-H. (2018). Automated text analysis for consumer research.
26 *Journal of Consumer Research*, 44(6), 1274–1306.
- 27 Ignatow, G., & Robinson, L. (2017). Pierre Bourdieu: Theorizing the digital. *Information,*
28 *Communication & Society*, 20(7), 950–966.
- 29 Jacobs, B. J. D., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large
30 assortments. *Marketing Science*, 35(3), 389–404.
- 31 Jensen, C. B. & Rødje, K. (2010). *Deleuzian Intersections: Science, Technology, Anthropology*.
32 New York, NY: Berghahn Books
- 33 Jha, H. K., & Beckman, C. M. (2017). A patchwork of identities: Emergence of charter schools
34 as a new organizational form. In *Emergence*, vol. 50 (pp. 69–107). Bingley, UK: Emerald
35 Publishing.
- 36 Jin, M., Luo, X., Zhu, H., & Zhuo, H. H. (2018). Combining deep learning and topic modeling
37 for review understanding in context-aware recommendation. In *Proceedings of NAACL-*
38 *HLT 2018* (pp. 1605–1614). doi: 10.18653/v1/N18-1145
- 39 Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*,
40 41(6), 750–769.
- 41 Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing* (2nd ed.). Upper Saddle
42 River, N.J: Pearson.
- 43 Kaminski, J., Jiang, Y., Piller, F., & Hopp, C. (2017). Do user entrepreneurs speak different?
44 Applying natural language processing to crowdfunding videos. In *Proceedings of the*
45 *2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp.
46 2683–2689). New York, NY: ACM.
- 47 Kaplan, S. (2008a). Cognition, capabilities, and incentives: Assessing firm response to the fiber-
48 optic revolution. *Academy of Management Journal*, 51(4), 672–695.
- 49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Kaplan, S. (2008b). Framing contests: Strategy making under uncertainty. *Organization Science*,
4 19(5), 729–752.
- 5 Kaplan, S., & Vakili, K. (2015). The double-edged word of recombination in breakthrough
6 innovation. *Strategic Management Journal*, 36(10), 1435–1457.
- 7 Karanovic, J., Berends, H., & Engel, Y. (2018). Is platform capitalism legit? Ask the workers.
8 *Academy of Management Proceedings*, 2018(1), 17038.
- 9 Kennedy, M. T. (2005). Behind the one-way mirror: Refraction in the construction of product
10 market categories. *Poetics*, 33(3–4), 201–226.
- 11 Kennedy, M. T. (2008). Getting counted: Markets, media, and reality. *American Sociological*
12 *Review*, 73(2), 270–295.
- 13 Kennedy, M. T., Chok, J. I., & Liu, J. (2012). What does it mean to be green? The emergence of
14 new criteria for assessing corporate reputation. In T. G. Pollock & M. L. Barnett (Eds.),
15 *Oxford Handbook of Corporate Reputation*. Oxford, UK: Oxford University Press. doi:
16 10.1093/oxfordhb/9780199596706.013.0004
- 17 Kennedy, M. T., & Fiss, P. C. (2013). An ontological turn in categories research: From standards
18 of legitimacy to evidence of actuality. *Journal of Management Studies*, 50(6), 1138–
19 1154.
- 20 Kim, H., Ahn, S.-J., & Jung, W.-S. (2018). Horizon scanning in policy research database with a
21 probabilistic topic model. *Technological Forecasting and Social Change*.
22 <https://doi.org/10.1016/j.techfore.2018.02.007>
- 23 Kim, S., & Bae, J. 2016. Cross-cultural differences in concrete and abstract corporate social
24 responsibility (CSR) campaigns: Perceived message clarity and perceived CSR as
25 mediators. *International Journal of Corporate Social Responsibility*, 1, 1–14.
- 26 Kinney, A. B., Davis, A. P., & Zhang, Y. (2018). Theming for terror: Organizational adornment
27 in terrorist propaganda. *Poetics*, 69, 27–40.
- 28 Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection.
29 *Computational Linguistics*, 32(4), 485–525.
- 30 Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological
31 characteristics of 26 datasets. *Big Data & Society*, 3(1), 1–10.
- 32 Kline, S. J., & Rosenberg, N. (1986). An overview of innovation. In R. Landau & N. Rosenberg
33 (Eds.), *The positive sum strategy: Harnessing technology for economic growth* (pp. 275–
34 306). Washington, DC: National Academy of Sciences.
- 35 Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Corlay, S.
36 (2016). Jupyter Notebooks—a publishing format for reproducible computational
37 workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and power in academic*
38 *publishing: Players, agents and agendas* (pp. 87–90). Clifton, VA: IOS Press. doi:
39 10.3233/978-1-61499-649-1-87
- 40 Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text
41 classification for organizational researchers: A tutorial. *Organizational Research*
42 *Methods*, 21(3), 766–799.
- 43 Krippendorff, K. (1980). *Content analysis* (1st ed.). Beverly Hills, CA: Sage.
- 44 Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.).
45 Thousand Oaks, CA: Sage.
- 46 Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (3rd ed.).
47 Thousand Oaks, CA: Sage.
- 48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep
4 convolutional neural networks. In *Proceedings of the 25th International Conference on*
5 *Neural Information Processing Systems*, vol. 1 (pp. 1097–1105). Red Hook, NY: Curran
6 Associates Inc.
7
8 Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, IL: University of
9 Chicago Press.
10
11 Lakoff, G. (1970). Linguistics and natural logic. *Synthese*, 22(1–2), 151–271.
12
13 Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management*
14 *Review*, 24(4), 691–710.
15
16 Langley, A., & Abdallah, C. (2011). Templates and turns in qualitative studies of strategy and
17 management. In *Building methodological bridges* (pp. 201–235). Bingley, UK: Emerald
18 Group Publishing.
19
20 Lasswell, H. D. (1948). *Power and personality*. New York, NY: W.W. Norton.
21
22 Lasswell, H. D., & Lerner, D. (1952). *The comparative study of symbols*. Palo Alto, CA:
23 Stanford University Press.
24
25 Lawrence, T. B., Suddaby, R., & Leca, B. (2009). *Institutional work* (1st ed.). Cambridge, UK:
26 Cambridge University Press.
27
28 Lee, G. M., Qiu, L., & Whinston, A. B. (2016). A friend like me: Modeling network formation in
29 a location-based social network. *Journal of Management Information Systems*, 33(4),
30 1008–1033.
31
32 Lee, H., Kwak, J., Song, M., & Kim, C. O. (2015). Coherence analysis of research and education
33 using topic modeling. *Scientometrics*, 102(2), 1119–1137.
34
35 Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer
36 reviews. *Journal of Marketing Research*, 48(5), 881–894.
37
38 Levy, K. E. C., & Franklin, M. (2014). Driving regulation: Using topic models to examine
39 political contention in the U.S. trucking industry. *Social Science Computer Review*, 32(2),
40 182–194.
41
42 Lim, A., & Tsutsui, K. 2012. Globalization and commitment in corporate social responsibility:
43 Cross-national analyses of institutional and political economy effects. *American*
44 *Sociological Review*, 77(1): 69-98.
45
46 Liu, Y., Mai, F., & MacDonald, C. (2018). A big-data approach to understanding the thematic
47 landscape of the field of business ethics, 1982–2016. *Journal of Business Ethics*.
48 <https://doi.org/10.1007/s10551-018-3806-5>
49
50 Locke, K. D. (2001). *Grounded theory in management research*. London: Sage.
51
52 Loewenstein, J., Ocasio, W., & Jones, C. (2012). Vocabularies and vocabulary structure: A new
53 approach linking categories, practices, and institutions. *Academy of Management Annals*,
54 6(1), 41–86.
55
56 Lounsbury, M., & Glynn, M. A. (2001). Cultural entrepreneurship: Stories, legitimacy, and the
57 acquisition of resources. *Strategic Management Journal*, 22(6/7), 545–564.
58
59 Lounsbury, M., & Glynn, M. A. (2019). *Cultural entrepreneurship: A new agenda for the study*
60 *of entrepreneurial processes and possibilities*. Cambridge, UK: Cambridge University
Press.
- Lounsbury, M., & Ventresca, M. (2003). The new structuralism in organizational theory.
Organization, 10(3), 457–480.
- McCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. Retrieved from
<http://mallet.cs.umass.edu>.

- 1
2
3 Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural*
4 *Language Engineering*, 16, 100–103.
- 5 Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*
6 (1st ed.). Cambridge, MA: MIT Press.
- 7
8 Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The
9 Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd*
10 *Annual Meeting of the Association for Computational Linguistics: System*
11 *Demonstrations* (pp. 55-60). <https://www.aclweb.org/anthology/P14-5010>
- 12 Mantere, S., & Ketokivi, M. (2013). Reasoning in organization science. *Academy of*
13 *Management Review*, 38(1), 70–89.
- 14
15 Marciniak, D. (2016). Computational text analysis: Thoughts on the contingencies of an evolving
16 method. *Big Data & Society*, 3(2).
- 17 Marquis, C., Glynn, M. A., & Davis, G. F. (2007). Community isomorphism and corporate social
18 action. *Academy of Management Review*, 32(3), 925–945.
- 19 Marshall, E. A. (2013). Defining population problems: Using topic models for cross-national
20 comparison of disciplinary development. *Poetics*, 41(6), 701–724.
- 21
22 Martens, M. L., Jennings, J. E., & Jennings, P. D. (2007). Do the stories they tell get them the
23 money they need? The role of entrepreneurial narratives in resource acquisition. *Academy*
24 *of Management Journal*, 50(5), 1107–1132.
- 25 Mattmann, C. A. (2013). Computing: A vision for data science. *Nature*, 493(7433), 473.
- 26 McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013).
27 Differentiating language usage through topic models. *Poetics*, 41(6), 607–625.
- 28 Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for
29 microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th*
30 *International ACM SIGIR Conference on Research and Development in Information*
31 *Retrieval* (pp. 889–892). New York, NY: ACM.
- 32
33 Meyer, J., & Hannan, M. T. (1979). *National development and the world system: Educational,*
34 *economic, and political change, 1950-1970*. Chicago, IL: University of Chicago Press.
- 35 Meyer, R. E., Jancsary, D., Höllerer, M. A., & Boxenbaum, E. (2017). The role of verbal and
36 visual text in the process of institutionalization. *Academy of Management Review*, 43(3),
37 392–418.
- 38
39 Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*.
40 Thousand Oaks, CA: Sage.
- 41 Miles, M. B., Huberman, A. M., & Saldaña, J. (2013). *Qualitative data analysis* (3rd ed.).
42 Thousand Oaks, CA: Sage.
- 43
44 Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to
45 WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235–
46 234.
- 47
48 Miller, I. M. (2013). Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling
49 approach. *Poetics*, 41(6), 626–649.
- 50
51 Mimno, D. (2012). Computational historiography: Data mining in a century of classics journals.
52 *Journal on Computing and Cultural Heritage*, 5(1), 3:1–3:19.
- 53
54 Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing
55 semantic coherence in topic models. In *Proceedings of the Conference on Empirical*
56 *Methods in Natural Language Processing* (pp. 262–272). Stroudsburg, PA: Association
57 for Computational Linguistics.
- 58
59
60

- 1
2
3 Moe, W. W., Netzer, O., & Schweidel, D. A. (2017). Social media analytics. In B. Wierenga &
4 R. van der Lans (Eds.), *Handbook of Marketing Decision Models* (pp. 483–504). New
5 York, NY: Springer International.
6
7 Moe, W. W., & Schweidel, D. A. (2017). Opportunities for innovation in social media analytics.
8 *Journal of Product Innovation Management*, 34(5), 697–702.
9
10 Mohr, J. W. (1998). Measuring meaning structures. *Annual Review of Sociology*, 24, 345–370.
11
12 Mohr, J. W., & Bogdanov, P. (2013). Introduction—topic models: What they are and why they
13 matter. *Poetics*, 41(6), 545–569.
14
15 Mohr, J. W., & Duquette, V. (1997). The duality of culture and practice: Poverty relief in
16 New York City, 1888–1917. *Theory and Society*, 26(2/3), 305–356.
17
18 Mohr, J. W., Wagner-Pacifici, R., Breiger, R. L., & Bogdanov, P. (2013). Graphing the grammar
19 of motives in national security strategies: Cultural interpretation, automated text analysis
20 and the drama of global politics. *Poetics*, 41(6), 670–700.
21
22 Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business*
23 *Venturing*, 29(1), 1–16.
24
25 Momeni, A., & Rost, K. (2016). Identification and monitoring of possible disruptive
26 technologies by patent-development paths and topic modeling. *Technological*
27 *Forecasting and Social Change*, 104, 16–29.
28
29 Moretti, F. (2013). *Distant reading*. New York, NY: Verso.
30
31 Morgeson, F. P., Mitchell, T. R., & Liu, D. (2015). Event system theory: An event-oriented
32 approach to the organizational sciences. *Academy of Management Review*, 40(4), 515–
33 537.
34
35 Mützel, S. (2015). Facing big data: Making sociology relevant. *Big Data & Society*, 2(2),
36 2053951715599179.
37
38 Nag, M. (2015). *Meaning is relational: The changing contexts of the keyword “risk” in the New*
39 *York Times using a bag-of-tuples topic model*. Presented at the Text II Conference,
40 Princeton, NJ.
41
42 Nag, R., & Gioia, D. A. (2012). From common to uncommon knowledge: Foundations of firm-
43 specific use of knowledge as a resource. *Academy of Management Journal*, 55(2), 421–
44 457.
45
46 Nam, H., Joshi, Y. V., & Kannan, P. K. (2017). Harvesting brand information from social tags.
47 *Journal of Marketing*, 81(4), 88–108.
48
49 Navis, C., & Glynn, M. A. (2010). How new market categories emerge: Temporal dynamics of
50 legitimacy, identity, and entrepreneurship in satellite radio, 1990–2005. *Administrative*
51 *Science Quarterly*, 55(3), 439–471.
52
53 Navis, C., & Glynn, M. A. (2011). Legitimate distinctiveness and the entrepreneurial identity:
54 Influence on investor judgments of new venture plausibility. *Academy of Management*
55 *Review*, 36(3), 477–499.
56
57 Nelsen, B. J., & Barley, S. R. (1997). For love or money? Commodification and the construction
58 of an occupational mandate. *Administrative Science Quarterly*, 42(4), 619–653.
59
60 Nelson, L. K. (2017). Computational grounded theory: A methodological framework.
Sociological Methods & Research, 0049124117729703.
Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-
structure surveillance through text mining. *Marketing Science*, 31(3), 521–543.

- 1
2
3 Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for
4 digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*
5 (pp. 215–224). New York, NY: ACM.
- 6 Ocasio, W. (1997). Towards an attention-based view of the firm. *Strategic Management Journal*,
7 18, 187–206.
- 8 Oh, J., Stewart, A. E., & Phelps, R. E. (2017). Topics in the Journal of Counseling Psychology,
9 1963–2015. *Journal of Counseling Psychology*, 64(6), 604–615.
- 10 Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce*. (A. W. Burks, Ed.).
11 Cambridge, MA: Harvard University Press.
- 12 Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and*
13 *psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.
- 14 Pentland, B. T., & Feldman, M. S. (2005). Organizational routines as a unit of analysis.
15 *Industrial and Corporate Change*, 14(5), 793–815.
- 16 Pfarrer, M., Pollock, T., & Rindova, V. (2010). A tale of two assets: The effects of firm
17 reputation and celebrity on earnings surprises and investors' reactions. *Academy of*
18 *Management Journal*, 53(5), 1131–1152.
- 19 Phillips, N., Lawrence, T. B., & Hardy, C. (2004). Discourse and institutions. *Academy of*
20 *Management Review*, 29(4), 635–652.
- 21 Podolny, J. M. (1993). A status-based model of market competition. *American Journal of*
22 *Sociology*, 98(4), 829–872.
- 23 Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided
24 text analysis. *Organizational Research Methods*, 15(2), 263–287.
- 25 Porac, J. F., Wade, J. B., & Pollock, T. G. (1999). Industry categories and the politics of the
26 comparable firm in CEO compensation. *Administrative Science Quarterly*, 44(1), 112–
27 144.
- 28 Pratt, M. G. (2009). From the editors: For the lack of a boilerplate: Tips on writing up (and
29 reviewing) qualitative research. *Academy of Management Journal*, 52(5), 856–862.
- 30 Prein, G., & Kelle, U. (1995). Using linkages and networks for theory building. In U. Kelle
31 (Ed.), *Computer qualitative data analysis: Theory, methods, and practice* (pp. 62–68).
32 London, UK: Sage Publications.
- 33 Pröllochs, N., & Feuerriegel, S. (2018). Business analytics for strategic management: Identifying
34 and assessing corporate challenges via topic modeling. *Information & Management*.
35 <https://doi.org/10.1016/j.im.2018.05.003>
- 36 Puranam, D., Narayan, V., & Kadiyali, V. (2017). The effect of calorie posting regulation on
37 consumer opinion: A flexible latent Dirichlet allocation model with informative priors.
38 *Marketing Science*, 36(5), 726–746.
- 39 Raffaelli, R. (2018). Technology reemergence: Creating new value for old technologies in Swiss
40 mechanical watchmaking, 1970–2008. *Administrative Science Quarterly*,
41 0001839218778505. <https://doi.org/10.1177/0001839218778505>
- 42 Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago, IL: University
43 of Chicago Press.
- 44 Rao, H., Monin, P., & Durand, R. (2003). Institutional change in Toque Ville: Nouvelle cuisine
45 as an identity movement in French gastronomy. *American Journal of Sociology*, 108(4),
46 795–843.
- 47
48
49
50
51
52
53
54
55
56
57
58
59
60

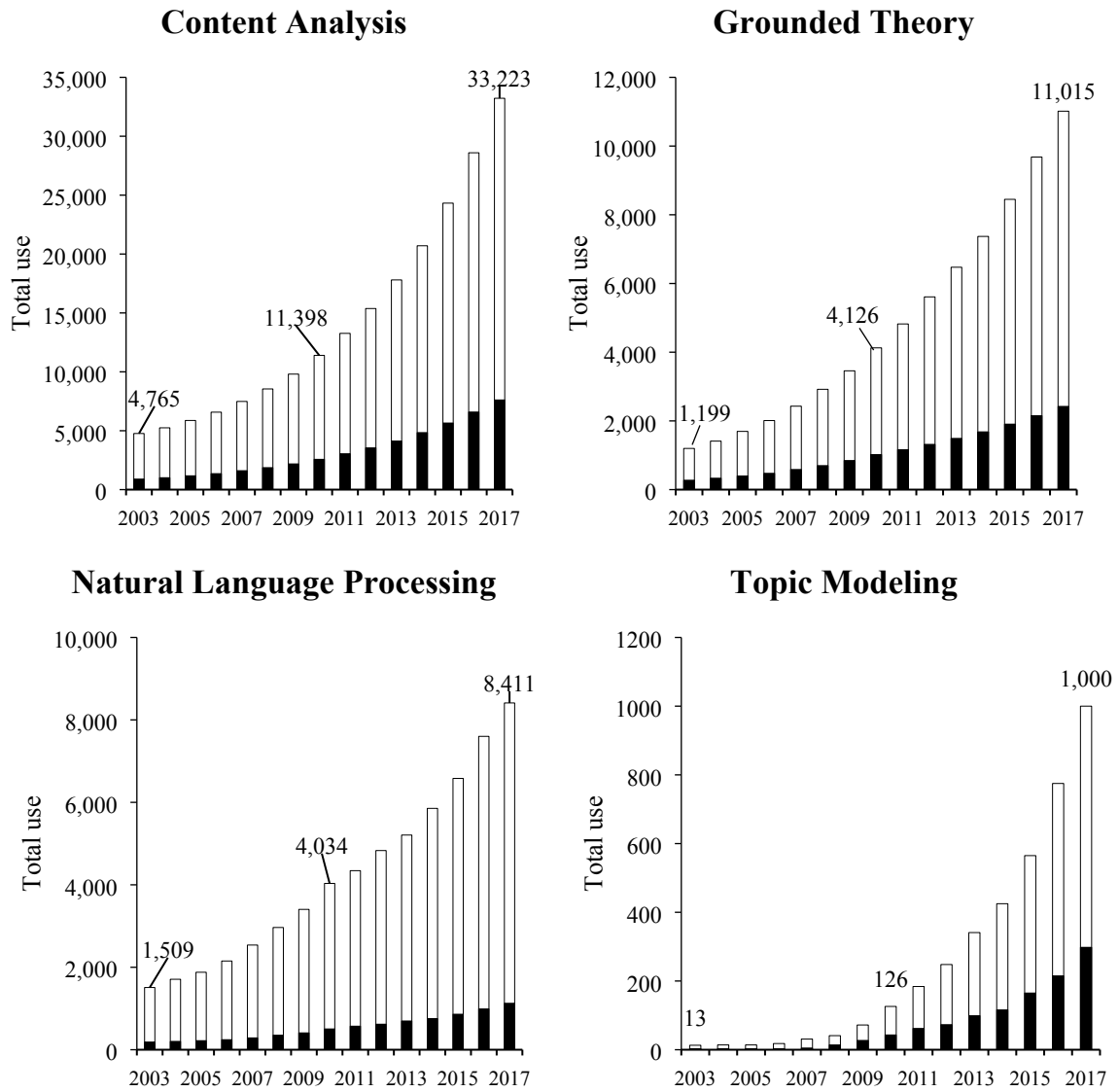
- 1
2
3 Rhee, E. Y., & Fiss, P. C. (2014). Framing controversial actions: Regulatory focus, source
4 credibility, and stock market reaction to poison pill adoption. *Academy of Management*
5 *Journal*, 57(6), 1734–1758.
- 6
7 Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K.,
8 Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey
9 responses. *American Journal of Political Science*, 58(4), 1064–1082.
- 10 Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures.
11 In *Proceedings of the Eighth ACM International Conference on Web search and Data*
12 *Mining* (pp. 399–408). New York, NY: ACM.
- 13
14 Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for
15 authors and documents. In *Proceedings of the 20th Conference on Uncertainty in*
16 *Artificial Intelligence* (pp. 487–494). Arlington, VA: AUAI Press.
- 17 Rothenberg, A. (2014) *Flight from wonder—An investigation of scientific creativity*. New York,
18 NY: Oxford University Press.
- 19 Ruckman, K., & McCarthy, I. (2017). Why do some patents get licensed while others do not?
20 *Industrial and Corporate Change*, 26(4), 667–688.
- 21 Saussure, F. de (1959). *Course in general linguistics*. New York, NY: McGraw-Hill.
- 22 Schmiedel, T., Müller, O., & vom Brocke, J. (2018). Topic modeling as a strategy of inquiry in
23 organizational research. *Organizational Research Methods*, 3(1).
24 <https://doi.org/10.1177/1094428118773858>
- 25
26 Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a better measure of business proximity:
27 Topic modeling for industry intelligence. *MIS Quarterly*, 40(4), 1035–1056.
- 28 Shimizu, T. (2018). *Coordinating collective framing process with heterogeneous actors in*
29 *technology standard development*. Presented at the 78th Annual Meeting of the Academy
30 of Management, Chicago, IL.
- 31 Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In
32 *Proceedings of the Workshop on Interactive Language Learning, Visualization, and*
33 *Interfaces* (pp. 63–70). Stroudsburg, PA: Association for Computational Linguistics.
- 34
35 Slingerland, E., Nichols, R., Neilbo, K., & Logan, C. (2017). The distant reading of religious
36 texts: A “big data” approach to mind-body concepts in early China. *Journal of the*
37 *American Academy of Religion*, 85(4), 985–1016.
- 38
39 Smith, A., Hawes, T., & Myers, M. (2014). Hiéarchie: Visualization for hierarchical topic
40 models. In *Proceedings of the Workshop on Interactive Language Learning,*
41 *Visualization, and Interfaces* (pp. 71–78). Baltimore, MD: Association for Computational
42 Linguistics.
- 43
44 Snow, D. A., & Benford, R. D. (1988). Ideology, frame resonance, and participant mobilization.
45 In *International Social Movement Research*, vol. 1 (pp. 197–218). Greenwich, CT: JAI
46 Press.
- 47
48 Snow, D. A., Rochford, E. B., Worden, S. K., & Benford, R. D. (1986). Frame alignment
49 processes, micromobilization, and movement participation. *American Sociological*
50 *Review*, 51(4), 464–481.
- 51
52 Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013).
53 Recursive deep models for semantic compositionality over a sentiment treebank. In
54 *Proceedings of the 2013 Conference on Empirical Methods in Natural Language*
55 *Processing* (pp. 1631–1642). Retrived from
56 https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf
57
58
59
60

- 1
2
3 Song, M., Heo, G. E., & Lee, D. (2015). Identifying the landscape of Alzheimer's disease
4 research with network and content analysis. *Scientometrics*, *102*(1), 905–927.
- 5 Song, M., & Kim, S. Y. (2013). Detecting the knowledge structure of bioinformatics by mining
6 full-text collections. *Scientometrics*, *96*(1), 183–201.
- 7 Sørensen, J. B., & Stuart, T. E. (2000). Aging, obsolescence, and organizational innovation.
8 *Administrative Science Quarterly*, *45*(1), 81–112.
- 9 Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic
10 Coherence over Many Models and Many Topics. Proceedings of the 2012 Joint Conference
11 on Empirical Methods in *Natural Language Processing and Computational Natural*
12 *Language Learning*, 952–961.
- 13 Strauss, A. C., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and*
14 *procedures for developing grounded theory* (2nd ed.). Los Angeles, CA: Sage.
- 15 Strothotte, T., & Schlechtweg, S. (2002). *Non-photorealistic computer graphics: Modeling,*
16 *rendering, and animation*. San Francisco, CA: Morgan Kaufmann.
- 17 Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping
18 patent data with unsupervised learning. *Technological Forecasting and Social Change*,
19 *115*, 131–142.
- 20 Tangherlini, T. R., & Leonard, P. (2013). Trawling in the sea of the great unread: Sub-corpus
21 topic modeling and humanities research. *Poetics*, *41*(6), 725–749.
- 22 Tchalian, H. (2019). Microfoundations and recursive analysis: A mixed-methods framework for
23 language-based research, computational methods, and theory development. In
24 *Microfoundations of Institutions*.
- 25 Tchalian, H., Alsudais, A., & Ocasio, W. (2018). *Bringing values back in: Cultural persistence*
26 *in institutional vocabularies*. Working paper.
- 27 Tchalian, H., Glaser, V. L., Hannigan, T. R., & Lounsbury, M. (2019). *Institutional attention:*
28 *Cultural entrepreneurship and the dynamics of category construction*. Working paper.
- 29 Thornton, P. H., & Ocasio, W. (1999). Institutional logics and the historical contingency of
30 power in organizations: Executive succession in the higher education publishing industry,
31 1958-1990. *American Journal of Sociology*, *105*(3), 801–843.
- 32 Thornton, P. H., Ocasio, W., & Lounsbury, M. (2012). *The institutional logics perspective: A*
33 *new approach to culture, structure and process*. New York, NY: Oxford University
34 Press.
- 35 Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From
36 grounded theory to abductive analysis. *Sociological Theory*, *30*(3), 167–186.
- 37 Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic
38 brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing*
39 *Research*, *51*(4), 463–479.
- 40 Tolbert, P. S., & Zucker, L. G. (1996). The institutionalization of institutional theory. In S.
41 Clegg, C. Hardy, & W. Nord (Eds.), *Handbook of organization studies* (pp. 175–190).
42 London: Sage.
- 43 Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data
44 analytic techniques to build organizational science. *Organizational Research Methods*,
45 *21*(3), 525-547.
- 46 Toubia, O., & Netzer, O. (2016). Idea generation, creativity, and prototypicality. *Marketing*
47 *Science*, *36*(1), 1–20.
- 48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations.
4 *RAND Journal of Economics*, 21(1), 172–187.
- 5 Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the cookie: User profiling in customer-base
6 analysis and behavioral targeting. *Marketing Science*, 35(3), 405–426.
- 7 Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of
8 semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.
- 9 Tzoukermann, E., Klavans, J. L., & Strzalkowski, T. (2005). Information retrieval. In R. Mitkov
10 (Ed.), *Oxford Handbook of Computational Linguistics* (pp. 1–12). Oxford, UK: Oxford
11 University Press.
- 12 Underwood, T. (2015). The literary uses of high-dimensional space. *Big Data & Society*, 2(2), 1–
13 6.
- 14 Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013) Atypical combinations and scientific
15 impact. *Science*, 342(6157):468–472.
- 16 Vaara, E. (2010). Taking the linguistic turn seriously: Strategy as a multifaceted and
17 interdiscursive phenomenon. In J. Baum & J. B. Lampel (Eds.), *The globalization of*
18 *strategy research* (pp. 29–50). Bingley, UK: Emerald Group Publishing.
- 19 Vaara, E., Aranda, A., Etchanchu, H., Guyt, J., and Sele, K. (2019). How to make use of
20 structural topic modeling in critical discourse analysis? Working paper.
- 21 Ventresca, M. J., & Mohr, J. W. (2002). Archival research methods. In J. A. C. Baum (Ed.),
22 *Blackwell companion to organizations* (pp. 805–828). Oxford, UK: Blackwell.
- 23 Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in
24 patents. *Technological Forecasting and Social Change*, 94, 236–250.
- 25 Vergne, J.-P., & Wry, T. (2014). Categorizing categorization research: Review, integration, and
26 future directions. *Journal of Management Studies*, 51(1), 56–94.
- 27 Wagner-Pacifici, R., Mohr, J. W., & Breiger, R. L. (2015). Ontologies, methodologies, and new
28 uses of big data in the social and cultural sciences. *Big Data & Society*, 2(2),
29 2053951715613810.
- 30 Wang, X., Bendle, N. T., Mai, F., & Cotte, J. (2015). The Journal of Consumer Research at 40:
31 A historical analysis. *Journal of Consumer Research*, 42(1), 5–18.
- 32 Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with
33 an application to information retrieval. In *Proceedings of the Seventh IEEE International*
34 *Conference on Data Mining (ICDM 2007)* (pp. 697–702). Piscataway, NJ: IEEE. doi:
35 10.1109/ICDM.2007.86
- 36 Wang, Y., & Chaudhry, A. (2018). When and how managers' responses to online reviews affect
37 subsequent reviews. *Journal of Marketing Research*, 55(2), 163–177.
- 38 Washington, M., & Zajac, E. J. (2005). Status evolution and competition: Theory and evidence.
39 *Academy of Management Journal*, 48(2), 282–296.
- 40 Weber, K., Patel, H., & Heinze, K. L. (2013). From cultural repertoires to institutional logics: A
41 content-analytic method. In M. Lounsbury & E. Boxenbaum (Eds.), *Institutional Logics*
42 *in Action*, Vol. 39B (pp. 351–382). Bingley, UK: Emerald Group Publishing.
- 43 Weber, K., & Dacin, M. T. (2011). The cultural construction of organizational life:
44 Introduction to the special issue. *Organization Science*, 22(2), 287–298.
- 45 Weber, R. P. (1990). *Basic content analysis* (2nd ed.). London: Sage.
- 46 Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management*
47 *Review*, 14(4), 490–495.
- 48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*.
4 Cambridge, MA: Technology Press of Massachusetts Institute of Technology.
5
6 Wilson, A. J., & Joseph, J. (2015). *Organizational attention and technological search in the*
7 *multibusiness firm: Motorola from 1974 to 1997*. Bingley, UK: Emerald Group
8 Publishing.
9
10 Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents
11 with topic modeling. *Scientometrics*, *100*(3), 767–786.
12
13 Zajac, E. J., & Fiss, P. C. (2006). The symbolic management of strategic change: Sensegiving
14 via framing and decoupling. *Academy of Management Journal*, *49*(6), 1173–1193.
15
16 Zajac, E. J., & Westphal, J. D. (1994). The costs and benefits of managerial incentives and
17 monitoring in large U.S. corporations: When is more not better? *Strategic Management*
18 *Journal*, *15*(S1), 121–142.
19
20 Zhang, Y., Moe, W. W., & Schweidel, D. A. (2017). Modeling the role of message content and
21 influencers in social media rebroadcasting. *International Journal of Research in*
22 *Marketing*, *34*(1), 100–119.
23
24 Zhao, E. Y., Fisher, G., Lounsbury, M., & Miller, D. (2017). Optimal distinctiveness:
25 Broadening the interface between institutional theory and strategic management.
26 *Strategic Management Journal*, *38*(1), 93–113.
27
28 Zhao, E. Y., Ishihara, M., Jennings, P. D., & Lounsbury, M. (2018). Optimal distinctiveness in
29 the console video game industry: An exemplar-based model of proto-category evolution.
30 *Organization Science*, *29*(4), 588–611.
31
32 Zott, C., & Huy, Q. N. (2007). How entrepreneurs use symbolic management to acquire
33 resources. *Administrative Science Quarterly*, *52*(1), 70–105.
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 1
A Comparative Assessment of Topic Modeling's Use



Note: The charts show the number of unique articles published in the social sciences (white bars) and the business/management literature (black bars) in Scopus and the Web of Science.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 2
Topic Modeling Rendering in Theory-Building Spaces

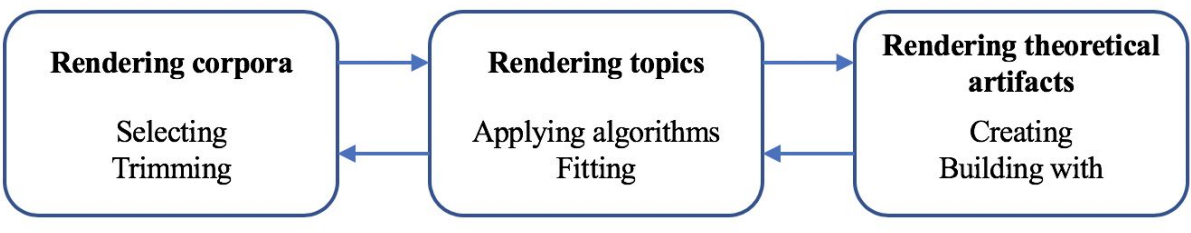


Table 1
Topic Modeling Conceptual Terms

Conceptual Terms	Definition in the Context of Rendering with Topic Modeling
Algorithm	A process or unambiguous set of rules to be followed, usually by a computer. An automated processing technique for distilling data inputs into topic modeling elements (clusters, weights, similarities).
Big textual data	Data characterized by large volume (a million or more words), high variety (diverse sources), and high temporality (many periods).
Coherence	A quantitative metric for topic quality. Clear and well-bounded topic(s) with evident criteria for classification of other text or topics within it. Based on pairs of words in a topic that have high co-document frequencies.
Dictionary	The set of meaningful key words to be used to assess the content and meaning of a corpus. The basis for annotating words in a text as a code category.
Disambiguation	A process of using the context to adjudicate between different meanings (or readings) of a word beyond its literal definition.
Fit	Criteria for how many topics are derived, how they are related, and what they might mean.
Heteroglossia	Multiple styles of word-use in a single text reflecting different perspectives or styles of expression.
hLDA	Hierarchical latent Dirichlet allocation—a form of structured LDA.
KWIC	Key words in context; embedding or considering words in their relationship with other words in a corpus and in a socio-cultural condition.
Lemmatizing	Transforming a word into its dictionary form. In practice, different lemmatizing methods convert words to their singular forms or by using a higher-level synonym from a linguistic thesaurus.
LDA	Latent Dirichlet allocation, in which documents are assumed to draw content from a latent set of topics with probability-based parameters that can be adjusted to determine those topics.
LIWC	Linguistic inquiry and word count (aka “Luke”) is a dictionary-based, positive- and negative-affect word frequency program designed to capture content and affective meaning.
LSI	Latent semantic indexing (LSI) is an algorithm which uses linear algebra to perform dimensionality reduction and convert texts to a matrix form.
LSVDs	Lasswell Value Dictionary tags
Perplexity	A quantitative metric for the quality of a topic model based on the number of topics selected. In general, perplexity is a statistical measure of how well a model fits based on splitting data into a training set and test set. In LDA topic modeling, it is a relative measure of topic fit; better models have lower perplexity scores.
Polysemy	Words that have multiple meanings or uses.
Relationality	Words whose meanings are contextually dependent.
Rendering	The process of generating provisional knowledge by iterating between selecting and trimming raw textual data, applying algorithms and fit criteria

Conceptual Terms	Definition in the Context of Rendering with Topic Modeling
	to surface topics, and creating and building with theoretical artifacts, such as processes, causal links or measures.
Selecting	Selecting documents (e.g., using sampling) and forms of text to be assessed.
Smoothing	Applying LDA-related algorithms to reduce the number of and disparity among topics, normally through iteration.
Stemming	The conversion of text segments (words) to their root word forms.
Stop words	Words that serve a less important role in meaning construction (i.e., articles such as “the” or “a”).
Theoretical artifact	A construct, conceptual association, process, causal linkage, mechanism or measure.
Token	The smallest, disaggregated, distinct bit of textual data (normally a noun) used in analysis.
Topic	A bag of words that frequently appear together across documents; the derived word(s) from a topic in topic modeling representing word tokens.
Trimming	Reducing textual data and specific words into useful tokens, normally by lemmatizing and/or stemming; a form of text normalization.

Table 2
Management Subject Areas Enhanced by Topic Modeling Research

Subject Area	Topics	Exemplars	Key Contributions
Detecting novelty and emergence	Understanding shifts in patent citations (#25: <i>patent, technological, knowledge, citation, identify, path, base, cite, highly</i>)	Kaplan & Vakili (2015)	Provides a means to disentangle the cognitive content of novel innovations from the outcomes associated with innovations
	Measuring topics to understand innovation (#24: <i>idea, weight, distribution, edge, measure, base, node, combination, average, semantic</i>)	Toubia & Netzer (2016)	Provides a means of empirically measuring different theoretical dimensions of creativity to develop new understandings of idea generation
	Using topic models to understand managerial cognition through technology problems, search and attention (#1: <i>problem, search, structure, attention, concept, process, exist, unit, create, general</i>)	Wilson & Joseph (2015)	Provides a way for researchers to understand the dynamics of managerial attention relative to background knowledge.
	Understanding knowledge dynamics (#14: <i>scientific, impact, focus, app, knowledge, article, content, find, rhetorical, attribute</i>)	Antons et al. (2018)	Provides a means to theorize how latent knowledge structures undergird innovative activities
	Understanding emerging organizational forms (#10: <i>form, identity, community, logic, organizational, actor, institutional, application, distinct, school</i>)	Jha & Beckman (2017)	Provides a method for theorizing the relationships between constructs at different levels of analysis, such as organizational identity and institutional logics

Subject Area	Topics	Exemplars	Key Contributions
Developing inductive classification systems	Understanding dynamics of meanings and networks in knowledge fields (#34: <i>article, journal, field, publish, year, citation, scholar, papers, author, paper</i>)	Wang et al. (2015)	Provides a means to discover emerging trends in knowledge fields by enabling researchers to identify different dimensions of knowledge and connect these dimensions with other theoretical constructs
	Understanding how categories affect competitive dynamics (#18: <i>firm, category, industry, performance, position, distinctiveness, competitor, show, level, competitive</i>)	Haans (2019)	Provides a means to measure differentiation associated with cultural concepts in strategic action
	Understanding the relationships between risk and investment (#31: <i>information, analyst, report, investor, risk, discovery, interpretation, manager, role, find</i>)	Huang et al. (2017)	Provides a way for researchers to compare disparate forms of data such as written reports and transcripts of conference calls
	Inducing underlying meanings associated with cultural events (#32: <i>major, rebellion, job, event, state, report, case, crime, level, related</i>)	Miller (2013)	Provides a way to overcome human biases associated with interpreting cultural events
	Classifying sets of data and consumers (#4: <i>make, pile, task, datum, set, summary, consumer, sort, propose, item</i>)	Blanchard, Aloise, & Desarbo (2017)	Introduces a new technique that can be used to address a classic consumer behavior problem of sorting
Understanding online audiences and products	The nature of online consumer profiles (#12: <i>user, content, message, social-media, consumer, influence, individual, role, activity, platform</i>)	Trusov et al. (2016)	Provides a means for conceptualizing customers as click groups, networks, and online communities

Subject Area	Topics	Exemplars	Key Contributions
	Online brand recognition and preference (#23: <i>brand, approach, car, text-mining, map, keyword, association, mention, tag, consumer</i>)	Netzer et al. (2012)	Helps capture brand network attributes and evolving brand linkages
	Online customer evaluations and responses to them (#29: <i>review, response, rating, health, restaurant, post, hotel, regulation, find, treatment</i>)	Wang & Chaudry (2018)	Maps the co-occurrence of reviews and responses in real time to understand performance adjustment effects
	Improving topic modeling of online audiences and products (#13: <i>product, dimension, customer, consumer, attribute, purchase, market, prediction, review, online</i>)	Jacobs et al. (2016)	Refines topic selection and supervision criteria, as well as fit criteria (e.g., smoothing, correlation, and hierarchy across topics)
Analyzing frames and social movements	Understanding how frames influence political processes (#27: <i>financial, fomc, economy, price, market, hypothesis, macroeconomic, primary, discussion, real</i>)	Fligstein et al. (2017)	Provides a means to identify and measure the deployment of different frames in political activities
	The relationship between frames, context, and audience (#6: <i>frame, context, audience, important, framing, make, process, give, individual, part</i>)	Levy & Franklin (2014)	Enables researchers to identify distinct discursive frames
	Understanding field-level relationships between organizations, discourse, and strategies (#17: <i>organization, theme, individual, effort, people, comment, strategy, day, term, field</i>)	Bail et al. (2017)	Provides a means to capture sentiment and bias in normalized spaces

Subject Area	Topics	Exemplars	Key Contributions
	Social movement strategies, networks, and actions (#11: <i>group, network, identify, radical, movement, pair, environmental, action, strategy, finding</i>)	Almquist & Bagozzi (2017)	Provides a means to map unseen or hidden ties
Understanding cultural dynamics	Understanding the professionalization of a field (#2: <i>amateur, field, professional, public, space, radio, actor, theme, expertise, expert</i>)	Croidieu & Kim (2018)	Provides a method for inductively analyzing a corpus as part of a longitudinal case study
	Using topic modeling to analyze big data to understand cultural trends (#5: <i>social, conversation, big-data, language, theory, cognitive, public, shift, meaning, emotional</i>)	Wagner-Pacifici et al. (2013)	Articles that explicitly describe and illustrate how to use topic modeling to extract meanings from large corpora
	Understanding dynamics associated with literary meanings (#9: <i>work, author, write, literary, passage, read, corpus, series, gender, stm</i>)	Tangherlini & Leonard (2013)	Enables researchers to identify and compare meanings across different sub-corpora over time
	Understanding how cultural meanings change over time (#19: <i>art, support, term, percent, view, recombination, newspaper, assign, agency, grant</i>)	DiMaggio et al. (2013)	Enables researchers to analyze shifts in cultural meanings over time
	Understanding the evolution of cultural trends (#28: <i>time, period, trend, change, fertility, population, country, context, british, demographic</i>)	Marshall (2013)	Uses methods such as correlated topic modeling to connect changes in cultural meaning over time with quantitative data

APPENDIX

Topic Modeling Topic Modeling Research in Management

Following recent efforts by scholars using topic modeling to map literatures (e.g., Antons et al., 2016, 2018; Cho et al., 2017; Guerreiro et. al. 2016, Liu et al., 2018, Oh et al., 2017), we utilized the method to inductively analyze our topic modeling corpus. In this appendix, we provide additional details about our rendering process (see Figure 2 in the main text) that we did not have the space to discuss in the body of the paper. In order to do so sensibly, we need to provide those details within the context of the rendering steps that we discussed in the body. As a result, this appendix represents a standalone description of our topic modeling effort.

Rendering a Corpus

As highlighted in the main text, in order to identify management subjects on which topic modeling has been making an impact, we first curated relevant journal articles that leveraged topic modeling methods - not a simple task, for it required rounds of selection and trimming. Specifically, we created a corpus by conducting a computerized text search in Scopus and the Web of Science for article abstracts with keywords signaling topic modeling: “topic model*”, “LDA”, “Latent Dirichlet Allocation”. After pruning articles containing false positives for the LDA acronym (such as “linear discriminant analysis” or “loss distribution approach”), and duplicates, this yielded a vast set of articles (N= 1466 in 639 publications). Many articles were from computer or information science, so we narrowed out the corpus by curating only include articles from publications that were identified by Scopus and Web of Science as “business” (N=566 articles in 219 publications). We analyzed this preliminary corpus using topic modeling techniques; we found that were still many topics that were about algorithms, big textual data,

1
2
3 computer science, logistics, MIS - or just not very interpretable. We continued to narrow our
4 analysis by selecting a sub-set of articles published in mainstream management journals (e.g.,
5 ASQ, SMJ, etc.) and journals from related disciplines that management scholars using topic
6 modeling methods read and cited. For example, we found that many management scholars were
7 influenced by and referenced articles from the special issue in *Poetics* (e.g., Mohr & Bogdanov,
8 2013). Using this approach, we ultimately trimmed the corpus to 66 manuscripts that were
9 directly relevant to management theory.

10
11
12 More specifically, to effectively manage our rendering process in one place, we used
13 Jupyter Notebooks with Python (Kluyver et al., 2016) alongside the libraries Gensim, Pandas,
14 and the Natural Language Toolkit (NLTK). We also used Python to interface (using shell
15 commands) with the Java software packages Mallet and Stanford CoreNLP. In our initial
16 analysis, we relied on abstracts and titles for topic modeling. However, following on Mohr &
17 Bogdanov (2013)—particularly in light of Crossley et al.'s (2017) caution to use over 1,000
18 documents and 20,000 words for good convergence—we downloaded the full content of articles
19 as PDFs, then used Python to break them down into paragraphs and clean the text. Our paragraph
20 tokenization process was custom-written in Python and based on regular expressions
21 corresponding to common patterns manually found in improper paragraph breaks. This analysis
22 was applied across all 66 papers and resulted in 5362 paragraphs, the latter serving as the
23 “documents” for LDA.

24
25 Before doing detailed cleaning of the text, we first attempted to identify common phrases.
26 Followed the procedure from Antons et al. (2016) to identify and replace n-grams in each
27 paragraph, we employed an algorithm from NLTK that analyzed common bigrams and trigrams
28 appearing in each paragraph. We then manually coded each phrase as interpretable, given our
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 domain expertise. For all phrases coded as interpretable, we collapsed them into a single token
4
5 by substituting a “-” character for space characters (ie. “big data” became “big-data”). The
6
7 insight here was to collapse common phrases such as “social media” that have interpretable
8
9 meaning which would be lost when LDA scrambles word order in the bag of words projection
10
11 (Wang, McCallum, & Wei, 2007). We also examined high and low relevance and common
12
13 phrases to be sure that we had stable and unique keywords for our topics, thus removing phrases
14
15 such as “latent Dirichlet allocation”.
16
17

18
19 After processing phrases, we cleaned each paragraph using the NLP parsing approach
20
21 with the Stanford CoreNLP software. This computational linguistics/NLP tool broke down each
22
23 paragraph into constituent sentences, removed punctuation, then analyzed each word according
24
25 to their Part-Of-Speech to determine an adequate lemma. For the collapsed phrases, this analysis
26
27 just reported the full phrase (i.e., “big-data”). Each paragraph was thus converted into a single
28
29 unordered list of lemmatized words and n-gram phrases. We then assessed that corpus using
30
31 LDA (applying the Gibbs algorithm for its convergence method) with the number of topics based
32
33 on the coherence measure data and interpretability. This final corpus used for the LDA contained
34
35 5362 documents with 351,786 distinct words. Appendix Table 1 summarizes the end result of
36
37 our rendered corpus by detailing our final list of 66 articles.
38
39
40

41
42 --- Insert Appendix Table 1 about here (or put online) ---
43

44 **Rendering Topics**

45
46 In order to render topics from this corpus, we used the LDA algorithm in two major
47
48 steps): first, we derived an LDA model from the paragraph dataset, and, second, we applied that
49
50 model to the corpus of 66 articles to derive a topic document matrix. This two-step approach was
51
52 used by Mohr & Bogdanov (2013) to analyze the paragraph as a unit of analysis in deriving the
53
54
55
56
57
58
59
60

1
2
3 model, where the corpus needs to be sufficiently large to confidently project a specification for
4 the LDA algorithm that converges. Statistical significance and convergence are functions of the
5 model specification, but this model can then be applied to individual documents to derive a topic
6 probability distribution. The major analytical move here is in using individual paragraphs from
7 all papers (N=5362) generate the model, but then applying it back on the full papers (N=66) to
8 determine the topic document matrix.
9

10
11 The LDA procedure was executed by the software tool Mallet (McCallum, 2002). A key
12 concern in conducting this procedure is determining the proper number of topics; i.e., fitting the
13 topic model. In this process we initially built upon quantitative evidence, using the popular
14 “UMass” measure of topic coherence (Mimno et al., 2011). Topic coherence is a metric done at
15 the level of a topic, developed to match human evaluations of topic quality (see Chang et al.,
16 2009 for a discussion on intrinsic measures of topics not correlating with human judgements).
17 The UMass metric of coherence considers high scoring words in a topic, tracking the semantic
18 similarity of documents in which they co-occur (see Mimno et al, 2011 for full description).
19 Stevens et al. (2012) extended this coherence score as a measure of overall topic model quality.
20 They generated different topic models based on specifications varying the number of topics (ie.
21 across a reasonable range generating models in steps of 5 or 10). They then graphed the average
22 topic coherence in each model and looked for evidence of a plateau. We conducted a similar
23 analysis, generating nine different models in Mallet ranging from 10 topics to 50, in steps of 5.
24 We followed the procedures from Mallet documentation, setting the hyper-parameters at
25 recommended values and computing diagnostic files for each model. Each diagnostic file was
26 processed in Python to compute average coherence scores. In summary, we projected different
27 LDA models for a range of topics k , graphing the coherence measure for each value of k between
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 5 and 50 topics (in increments of 5, so 5, 10, 15, topics and so on). The coherence graph
4
5 indicated that 35 topics was ideal as a plateau. For models two steps away on each side of 35
6
7 (i.e., 20, 25, 40, 45 topics) we manually inspected the top topic words for interpretability and
8
9 confirmed that 35 was adequate.
10

11
12 --- Insert Appendix Table 2 and Appendix Figure 1 about here (or put online) ---
13
14

15 16 17 **Rendering Theoretical Artifacts** 18

19 To render theoretical artifacts from the topic output, inspired by manuscripts such as
20
21 Croidieu and Kim (2017), Antons et al. (2016) and Mohr et al. (2013), we sought to approach
22
23 this visually using tools such as LDAvis (Sievert & Shirley, 2014). From this, we developed a
24
25 four-step process. First, for each topic, we analyzed the MDS plot, reordering the top words
26
27 according to the relevance metric in Sievert & Shirley (2014), which altered the order between
28
29 extremes of common words across topics and those uniquely within. We also tracked linkages
30
31 between topics and documents, using topic weights to form a Topic Significance Ranking (Al
32
33 Sumait, Barbará, Gentle, & Domeniconi, 2009) to sense the meaning of topics based on domain
34
35 expertise of papers. Second, we created a “rendering artifact” that synthesized critical
36
37 information about each topic on one page (see Appendix Figure 2). Specifically, we showed the
38
39 words in the topic (along with the weight of the words), the documents the topic was found in
40
41 (along with topic weights in documents), and the MDS chart.
42
43
44
45

46
47 --- Insert Appendix Figure 2 about here (or put online) ---
48

49 Third, three of the co-authors went through each topic and independently assessed the
50
51 theoretical meaning of these topics and their keywords. Each examined the words and weighted
52
53 documents (paragraphs in articles) by topic and created first and second-order codes of the
54
55
56
57
58
59
60

1
2
3 topics, which the authors then aggregated into management subject areas. Fourth, the authors
4
5 compared codes to determine level of agreement and generated a master spreadsheet of words,
6
7 topics, articles, key contributions and subjects (see Table 2). In keeping with theoretical
8
9 rendering, we paid particular attention to how subject areas were signaled and extended by
10
11 particular topics, as well as the ways in which topic modeling research introduced new
12
13 constructs, relationships, and mechanisms into those areas. Both represented the theoretical
14
15 “delta” of using topic modeling. Such grounded theorizing using axial codes, employed by
16
17 trained experts is relatively standard in management theory today (Bansal & Corley, 2014;
18
19 Denzin & Lincoln, 2011; Gioia et al., 2013; Pratt, 2009; see also, Croidieu & Kim, 2018).
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Appendix Table 1
Rendering the Corpus

Authors	Year	Article Title	Journal
Ahonen, P	2015	Institutionalizing Big Data methods in social and political research	Big Data & Society
Almquist Z.W.; Bagozzi B.E.	2017	Using radical environmentalist texts to uncover network structure and network features	Sociological Methods and Research
Antons D.; Joshi A.M.; Salge T.O.	2018	Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts	Journal of Management
Antons, D; Kleer, R; Salge, TO	2016	Mapping the topic landscape of JPIM, 1984-2013: In search of hidden structures and development trajectories	Journal of Product Innovation Management
Bail, CA; Brown, TW; Mann, M	2017	Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation	American Sociological Review
Bao, Y; Datta, A	2014	Simultaneously discovering and quantifying risk types from textual risk disclosures	Management Science
Bendle, NT; Wang, X	2016	Uncovering the message from the mess of Big Data	Business Horizons
Blanchard, SJ; Aloise, D; DeSarbo, WS	2017	Extracting summary piles from sorting task data	Journal of Marketing Research
Büschken, J; Allenby, GM	2016	Sentence-based text analysis for customer reviews	Marketing Science
Buurma, RS	2015	The fictionality of topic modeling: Machine reading Anthony Trollope's Barsetshire series	Big Data & Society
Cho, YJ; Fu, PW; Wu, CC	2017	Popular Research Topics in Marketing Journals, 1995-2014	Journal of Interactive Marketing
Croidieu, G; Kim, PH	2018	Labor of love: Amateurs and lay-expertise legitimation in the early US radio field	Administrative Science Quarterly
DiMaggio, P	2015	Adapting computational text analysis to social science (and vice versa)	Big Data & Society
DiMaggio, P; Nag, M; Blei, D	2013	Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding	Poetics
Evans, JA; Aceves, P	2016	Machine translation: Mining text for social theory	Annual Review of Sociology
Fligstein, N; Stuart Brundage, J; Schultz, M	2017	Seeing like the Fed: Culture, cognition, and framing in the failure to anticipate the Financial Crisis of 2008	American Sociological Review

1				
2				
3				
4	Giorgi, S; Weber, K	2015	Marks of distinction: Framing and audience appreciation in the context of investment advice	Administrative Science Quarterly
5	Grimmer, J.; Stewart, B.M.	2013	Text as data: The promise and pitfalls of automatic content analysis methods for political texts	Political Analysis
6				
7				
8	Guerreiro, J; Rita, P; Trigueiros, D	2016	A text mining-based review of cause-related marketing literature	Journal of Business Ethics
9				
10	Guo, L., Sharma, R., Yin, L., Lu, R., & Rong, K.	2017	Automated competitor analysis using big data analytics	Business Process Management Journal
11				
12	Guo, XH; Wei, Q; Chen, GQ; Zhang, J; Qiao, DD	2017	Extracting representative information on intra-organizational blogging	MIS Quarterly
13				
14	Haans, R.	2019	What's the value of being different when everyone is? (Move to in press? No clean text to topic model)	Strategic Management Journal
15				
16	Houghton J.P., Siegel M., Madnick S., Tounaka N., Nakamura K., Sugiyama T., Nakagawa D., Shirmen B.	2017	Beyond keywords: Tracking the evolution of conversational clusters in social media	Sociological Methods and Research
17				
18				
19				
20				
21	Huang A.H., Leheavy R., Zang A.Y., Zheng R.	2017	Analyst information discovery and interpretation roles: A topic modeling approach	Management Science
22				
23	Humphreys, A; Wang, RJH	2018	Automated text analysis for consumer research	Journal of Consumer Research
24	Jacobs, BJD; Donkers, B; Fok, D.	2016	Model-based purchase predictions for large assortments	Marketing Science
25				
26	Jha, HK; Beckman, CM	2017	A patchwork of identities: Emergence of charter schools as a new organizational form	Research in the Sociology of Organizations
27				
28	Jockers, ML; Mimno, D	2013	Significant themes in 19th-century literature	Poetics
29				
30	Kaplan, S; Vakili, K	2015	The double-edged sword of recombination in breakthrough innovation	Strategic Management Journal
31				
32	Kinney A.B., Davis A.P., Zhang Y.	2018	Theming for terror: Organizational adornment in terrorist propaganda	Poetics
33				
34	Kobayashi V.B., Mol S.T., Berkers H.A., Kismihók G., Den Hartog D.N.	2018	Text mining in organizational research	Organizational Research Methods
35				
36				
37	Lee, H.; Kwak, J., Song, M. Kim, C.	2015	Coherence analysis of research and education using topic modeling	Scientometrics
38				
39	Lee, T., & Bradlow, E.	2011	Automated marketing research using online customer reviews	Journal of Marketing Research
40				
41				
42				
43				
44				
45				
46				
47				

1	Levy, K.E.C.; Franklin, M.	2014	Driving regulation: Using topic models to examine political contention in the U.S. trucking industry	Social Science Computer Review
2				
3	Liu Y.; Mai F.; MacDonald, C.	2018	A Big-Data approach to understanding the thematic landscape of the field of business ethics, 1982–2016	Journal of Business Ethics
4				
5	Luo, JH; Pan, XW; Zhu, XY	2015	Identifying digital traces for business marketing through topic probabilistic model	Technology Analysis & Strategic Management
6				
7	Marciniak, D	2016	Computational text analysis: Thoughts on the contingencies of an evolving method	Big Data & Society
8				
9	Marshall, EA	2013	Defining population problems: Using topic models for cross-national comparison of disciplinary development	Poetics
10				
11	McFarland, DA; Ramage, D; Chuang, J; Heer, J; Manning, CD; Jurafsky, D	2013	Differentiating language usage through topic models	Poetics
12				
13	Miller, IM	2013	Rebellion, crime and violence in Qing China, 1722-1911: A topic modeling approach	Poetics
14				
15	Moe, WW; Schweidel, DA	2017	Opportunities for innovation in social media analytics	Journal of Product Innovation Management
16				
17	Mohr, JW; Bogdanov, P	2013	Introduction-Topic models: What they are and why they matter	Poetics
18				
19	Mohr, JW; Wagner-Pacifici, R; Breiger, RL; Bogdanov, P	2013	Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics	Poetics
20				
21	Momeni, A.; Rost, K.	2016	Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling	Technological Forecasting and Social Change
22				
23	Mützel, S	2015	Facing Big Data: Making sociology relevant	Big Data & Society
24				
25	Nam, H; Joshi, YV; Kannan, PK	2017	Harvesting brand information from social tags	Journal of Marketing
26				
27	Netzer, O.; Feldman, R.; Goldenberg, J.; Fresko, M.	2012	Mine your own business	Marketing Science
28				
29	Oh, J.; Stewart, A.; Phelps, R.	2017	Topic modeling journal topics	Journal of Counseling Psychology
30				
31	Puranam, D; Narayan, V; Kadiyali, V	2017	The effect of calorie posting regulation on consumer opinion: A flexible Latent Dirichlet Allocation model with informative priors	Marketing Science
32				
33	Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-	2014	Structural topic models for open-ended survey responses	American Journal of Political Science
34				
35				
36				
37				
38				
39				
40				
41				
42				
43				
44				
45				
46				
47				

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Luis, S. Gadarian, B. Albertson, and D. Rand			
Ruckman, K; McCarthy, I	2017	Why do some patents get licensed while others do not?	Industrial and Corporate Change
Schmiedel T., Müller O., vom Brocke J.	2018	Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture	Organizational Research Methods
Shi, Z; Lee, GM; Whinston, AB	2016	Toward a better measure of business proximity: Topic modeling for industry intelligence	MIS Quarterly
Suominen, A., Toivanen, H., & Seppänen, M.	2017	Firm's knowledge profiles	Technological Forecasting and Social Change
Tangherlini, TR; Leonard, P	2013	Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research	Poetics
Tirunillai, S; Tellis, GJ	2014	Mining marketing meaning from online chatter: Strategic brand analysis of Big Data using Latent Dirichlet Allocation	Journal of Marketing Research
Toubia, O; Netzer, O	2017	Idea generation, creativity, and prototypicality	Marketing Science
Trusov, M; Ma, LY; Jamal, Z	2016	Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting	Marketing Science
Underwood, T	2015	The literary uses of high-dimensional space	Big Data & Society
Venugopalan, S.; Rai, V.	2015	Topic based classification and pattern identification in patents	Technological Forecasting and Social Change
Wagner-Pacifici, R; Mohr, JW; Breiger, RL	2015	Ontologies, methodologies, and new uses of Big Data in the social and cultural sciences	Big Data & Society
Wang, X; Bendle, NT; Mai, F; Cotte, J	2015	The Journal of Consumer Research at 40: A historical analysis	Journal of Consumer Research
Wang, Y; Chaudhry, A	2018	When and how managers' responses to online reviews affect subsequent reviews	Journal of Marketing Research
Wilson, AJ; Joseph, J	2015	Organizational attention and technological search in the multibusiness firm: Motorola from 1974 to 1997	Advances in Strategic Management
Yau, C., Porter, A., Newman, N., & Suominen, A.	2014	Clustering scientific documents with topic modeling	Scientometrics
Zhang, YC; Moe, WW; Schweidel, DA	2017	Modeling the role of message content and influencers in social media rebroadcasting	International Journal of Research Marketing

Appendix Table 2
Rendering Topics

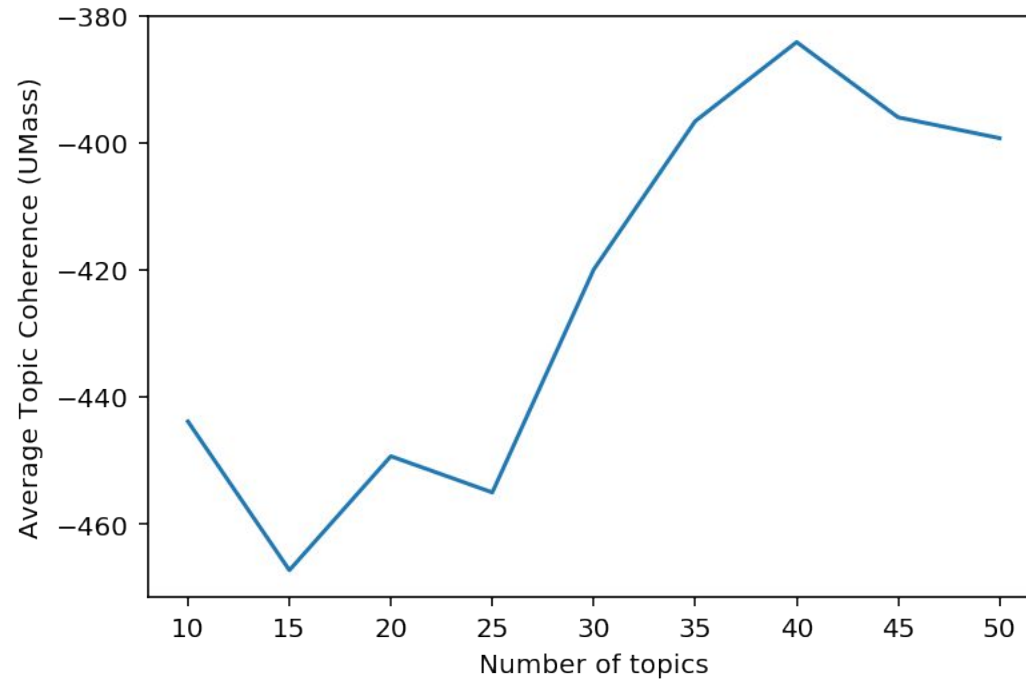
Topic #	Topic Weight (Rank)	Raw Topics
1	14	problem, search, structure, attention, concept, process, exist, unit, create, general
2	32	amateur, field, professional, public, space, radio, actor, theme, expertise, expert
3	20	sample, company, set, select, point, follow, test, dataset, describe, section
4	33	make, pile, task, datum, set, summary, consumer, sort, propose, item
5	1	social, conversation, big-data, language, theory, cognitive, public, shift, meaning, emotional
6	27	frame, context, audience, important, framing, make, process, give, individual, part
7	9	researcher, identify, discuss, insight, decision, subject, culture, specific, approach, organizational
8	12	show, figure, table, top, average, represent, high, present, compare, higher
9	8	work, author, write, literary, passage, read, corpus, series, gender, stm
10	24	form, identity, community, logic, organizational, actor, institutional, application, distinct, school
11	30	group, network, identify, radical, movement, pair, environmental, action, strategy, finding
12	3	user, content, message, social-media, consumer, influence, individual, role, activity, platform
13	10	product, dimension, customer, consumer, attribute, purchase, market, prediction, review, online
14	29	scientific, impact, focus, app, knowledge, article, content, find, rhetorical, attribute
15	5	document, corpus, label, identify, blei, process, algorithm, collection, text, latent
16	4	model, distribution, probability, parameter, observe, estimate, give, latent, assume, fit
17	26	organization, theme, individual, effort, people, comment, strategy, day, term, field
18	23	firm, category, industry, performance, position, distinctiveness, competitor, show, level, competitive
19	35	art, support, term, percent, view, recombination, newspaper, assign, agency, grant
20	11	text, category, approach, human, researcher, code, text-analysis, classification, construct, automate
21	13	effect, variable, significant, increase, estimate, coefficient, test, positive, regression, control
23	21	brand, approach, car, text-mining, map, keyword, association, mention, tag, consumer
24	28	idea, weight, distribution, edge, measure, base, node, combination, average, semantic
25	22	patent, technology, knowledge, technological, citation, identify, path, base, cite, highly

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

26	6	word, term, sentence, frequency, assign, matrix, common, represent, meaning, count
27	34	financial, fomc, economy, price, market, hypothesis, macroeconomic, primary, discussion, real
28	15	time, period, trend, change, fertility, population, country, context, british, demographic
29	19	review, response, rating, health, restaurant, post, hotel, regulation, find, treatment
30	18	relationship, licensor, characteristic, increase, similar, find, size, licensing, licensee, choice
31	31	information, analyst, report, investor, risk, discovery, interpretation, manager, role, find
32	25	major, rebellion, job, event, state, report, case, crime, level, related
33	2	datum, text, information, analyze, application, collect, tool, amount, online, extract
34	7	article, journal, field, publish, year, citation, scholar, papers, author, paper
35	16	model, text, unsupervised, assumption, political, apply, make, scale, grimmer, learn

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Appendix Figure 1
Rendering Topics with Coherence Scores



Appendix Figure 2 Rendering Theoretical artifact based on topic output

Topic 25	Words	patent	technology	knowledge	technological	citation	identify	path	base	cite	highly	
	1st Order Code	patents, knowledge, technology, and citation patterns										
	2nd Order Theme	strategy and innovation										
	Theory or Method?	theory										

0.128*"patent" + 0.067*"technology" + 0.034*"knowledge" + 0.028*"technological" + 0.020*"citation" + 0.020*"identify" + 0.019*"path" + 0.018*"base" + 0.016*"cite" + 0.015*"highly" + 0.013*"classification" + C

--weight: 0.56 title: Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling - Momeni, A., & Rost, K., 2016 Technological Forecasting and
 --weight: 0.33 title: Topic based classification and pattern identification in patents - Venugopalan, S., & Rai, V., 2015 Technological Forecasting and
 --weight: 0.23 title: The double-edged sword of recombination in breakthrough innovation - Kaplan, S; Vakili, K, 2015 STRATEGIC MANAGEMENT JOURNAL
 --weight: 0.19 title: Firm's knowledge profiles - Suominen, A., Toivanen, H., & Seppanen, M., 2017 Technological Forecasting and
 --weight: 0.12 title: Why do some patents get licensed while others do not? - Ruckman, K; McCarthy, I., 2017 INDUSTRIAL AND CORPORATE CHANG
 --weight: 0.08 title: ORGANIZATIONAL ATTENTION AND TECHNOLOGICAL SEARCH IN THE MULTIBUSINESS FIRM: MOTOROLA FROM 1974 TO 1997 - Wilson, AJ; Joseph, J., 2015 Advances in Strategic Managemen

