Editors’ Introduction:

“Topic models: What they are and why they matter.”
John Mohr (Sociology, UCSB) and Petko Bogdanov (Computer Science, UCSB)

In this short essay we provide a brief, non-technical introduction to the text mining methodology known as topic modeling. We start with the most basic question—what is a topic model? We review the theory behind the method and then focus in on the concept of a ‘topic.’ Here we especially look to see how the authors in this special issue have understood and interpreted the ‘topics’ that they have identified with these methods. We comment briefly on some of the demands, limitations and dilemmas of topic models and then proceed to the second question telegraphed by our title—why do topic models matter? We answer by describing some of the ways that we think these methods might change how scholars in the social and cultural sciences approach (and use) textual analysis. We end by introducing the eight research papers collected for this special issue. These are divided into three sections. Section 1 contains two papers that more self consciously highlight the question of how to utilize topic models to address social scientific questions. Section 2 has three papers that apply topic models to analyze the way that public discourse on threats to the state and broader cultural narratives of national security are constructed. Section 3 demonstrates several innovative ways to apply topic models to analyze cultural fields.

Section 1. Topic Models and the Sociology of Culture

Paul DiMaggio (Sociology, Princeton University), Manish Nag (Sociology, Princeton University), and David Blei (Computer Science, Princeton University).

Topic modeling provides a valuable method for identifying the linguistic contexts that surround social institutions or policy domains. This paper uses Latent Dirichlet Allocation (LDA) to analyze how one such policy domain, government assistance to artists and arts organizations, was framed in almost 8,000 articles. These comprised all articles that referred to government support for the arts in the U.S. published in five U.S. newspapers between 1986 and 1997, a period during which such assistance, once noncontroversial, became a focus of contention. We illustrate the strengths of topic modeling as a means of analyzing large text corpora, discuss the proper choice of models and interpretation of model results, describe means of validating topic-model solutions, and demonstrate the use of topic models in combination with more other statistical tools to estimate differences between newspapers in the prevalence of different frames. Throughout we emphasize affinities between the topic-modeling approach and such central concepts in the study of culture as framing, polysemy, heteroglossia, and the relationality of meaning.

Daniel A. McFarland (Education, Stanford), Daniel Ramage, Jason Chuang, Jeff Heer, Christopher D. Manning (Computer Science, Stanford) and Daniel Jurafsky (Linguistics, Stanford)

Sociologists wishing to employ topic models in their research need a helpful guide that describes the variety of topic modeling procedures, their issues, and various means of resolving them so as to convincingly answer sociological questions. We present this overview by recounting a series of
our prior collaborative projects that have employed and developed various forms of topic models to understand language differentiation in academe. With each project we encountered a variety of model-specific issues concerning the validity of topics and their suitability to our data and research questions. We developed variety of novel visualization techniques to make sense of topic-solutions and used a variety of techniques to validate our results. In addition, we created a variety of new topic modeling techniques and procedures suitable to different kinds of data and research questions.

Section 2. States, Threats and Discourses of National Security

Paper #3: “Rebellion, crime and violence in Qing China, 1722-1911: a topic modeling approach to the “great unread””
Ian Miller (History, Harvard University)

Banditry and rebellion in 18th and 19th century China have attracted substantial attention from several generations of researchers. Their studies, implicitly or explicitly, form typologies of violence and unrest as part of their arguments about the causes and consequences of these events. Often, they apply particular ontologies a priori to the source base. Given their reliance on many of the same sets of documents – most notably the Veritable Records of the Qing court – these studies fall prey to the particular biases and oversights of the record-keepers, as well as those of their theories of violence. To better understand the nature of crime, rebellion and unrest in the Qing Dynasty, it is important to come to terms with this bias – not because it renders the records useless, but rather because the bias itself reveals a great deal about the court, its agents, and their understandings of these phenomena. It is particularly difficult to apply fixed definitions to concepts like “banditry” and “unrest” - a problem that applies as much to modern researchers as to our historical informants. Therefore, rather than assuming a fixed set of categories, this study models Qing administrators' typologies of violence based on the frequencies of term co-occurrence using a Latent Dirichlet Allocation (LDA) topic model. In the model, reports of rebellion, crime and banditry are assigned to five topics – Crime, Unrest, Sedition, Rebellion, Major Rebellion and Border Rebellion. Each of these topics accounts for a particular statistical pattern of word use in the document; these in turn correspond to patterns of occurrence, observation and recording of related phenomena. These groupings give greater insight into the “crime rates” of the eighteenth and nineteenth centuries, as well as the incentives of state actors and their modes of record-keeping.

Tabitha Bonilla and Justin Grimmer (Political Science, Stanford).

A persistent concern in democracies is that terror threats make the public willing to restrict freedoms for increased safety. But a large literature has struggled to determine how terrorist threats affect the public's policy preferences. To more credibly estimate the effects of terrorist threats, we exploit elevations of the U.S. government's color coded alert system. Using this design, a new statistical model for texts, and a new collection of news stories, we show that media outlets allocate substantially more attention to terrorism after an alert. This sudden shift in media attention, though, has only limited effects on the public. The terror alerts raise the public's perceived likelihood of a terror attack, but opinion about President Bush's job performance, preferences for foreign intervention, or willingness to restrict civil liberties changes little in response to the alerts. Rather, the only consistent result is decreased economic expectations consistent with the strong economic downturn after the 9/11 attacks and the types of stories
published after the terror alerts are elevated. Terror alerts, then, did not exercise direct influence on the public's policy preferences. Instead, they changed the topic of conversation.

John W. Mohr (Sociology, UCSB), Robin Wagner-Pacifi (Sociology, The New School), Ron Breiger (Sociology, U of Arizona), Petko Bogdanov (Computer Science, UCSB).

The literary theorist Kenneth Burke (1945) outlined a methodology for identifying the basic “grammar of motives” that operate within texts. His strategy was to identify the logical form that is used for attributing meaning to human situations. We imagine how a variant of Burke's method might be applied in the era of automated text analysis, and then we explore an implementation of that variant (using a combination of natural language processing, semantic parsers and statistical topic models) in analyzing a corpus of eleven U.S. “National Security Strategy” documents that were produced between 1990 and 2010. This “automated” process for textual coding and analysis is shown to have much utility for analyzing these types of texts and to hold out the promise for being useful for other types of text corpora as well thereby opening up new possibilities for the scientific study of rhetoric.

**Section 3. Using Topic Models to Study Cultural Fields**

by Emily Marshall (Department of Sociology, Princeton University)

The content of academic journals provides insight into disciplinary boundaries and priorities. This paper uses correlated topic modeling (CTM), an innovative approach to textual analysis, for a cross-national comparison of the development of research agendas in the discipline of demography. Using articles from leading demographic journals from 1946-2005, CTM shows how the set of concepts relevant to the study of fertility was defined differently in France and Great Britain. Results indicate that demographic research agendas reflected both cultural and institutional differences that shaped different understandings of fertility decline. While British demography focused on high-fertility contexts, French demography focused on lower-fertility contexts. This difference reflects national intellectual traditions shaped by larger cultural discourses: the dominance of demographic transition theory and fears of overpopulation in Britain versus the co-existence in France of a second salient model, a theory of demographic “revolution” with sustained low fertility leading to depopulation. Relationships between expert concerns and broader public concerns are then examined in the British case by comparing journal publications to mass-media coverage of fertility and population issues. This comparison shows that British academic demography passed over some policy-relevant population issues, such as discussions of immigrant fertility, that were featured in the popular press.

Peter Leonard (University of Chicago) and Tim Tangherlini (Scandinavian Studies, UCLA)

Given a small, well-understood corpus that is of interest to a Humanities scholar, we propose *sub-corpus topic modeling* (STM) as a tool for discovering meaningful passages in a larger collection of less well understood texts. STM allows Humanities scholars to discover unknown passages from the vast sea of works that Moretti calls the “great unread,” and to significantly increase the researcher’s ability to discuss aspects of influence and the development of intellectual movements.
across a broader swath of the literary landscape. In this article, we test three typical Humanities research problems: in the first, a researcher wants to find text passages that exhibit similarities to a collection of influential non literary texts from a single author (here Darwin); in the second, a researcher wants to discover literary passages related to a well understood corpus of literary texts (here emblematic texts from the Modern Breakthrough); and in the third, a researcher hopes to understand the influence that a particular domain (here folklore) has had on the realm of literature over a series of decades. We explore these research challenges with three experiments, the first focused on the echoes of Darwin’s work in the broader Danish literary realm; the second focused on unknown authors from the “Modern Breakthrough,” a shift in Danish (and Nordic) literature away from Romanticism and toward Naturalism starting in the 1870s, and concomitant with the translation of Darwin’s works into Danish; and the third focused on the turn toward rural motifs in Danish literature from Romanticism through the progressive literature of the early twentieth century.

**Paper #8: Significant Themes in 19th-Century Literature**
by Matthew L. Jockers (Stanford Literary Lab, Stanford University) and David Mimno (Computer Science, Princeton University)

External factors such as author gender, author nationality, and date of publication affect both the choice of literary themes in novels and the expression of those themes, but the extent of this association is difficult to quantify. In this work, we apply statistical methods to identify and extract hundreds of "topics" from a corpus of 3,346 works of 19th-century British, Irish, and American fiction. We use these topics as a measurable, data-driven proxy for literary themes. External factors may predict fluctuations in the use of themes and the individual word choices within themes. We use topics to measure the evidence for these associations and whether that evidence is statistically significant.