# Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges

John Wilkerson and Andreu Casas

Department of Political Science, University of Washington, Seattle, Washington 98195; email: jwilker@uw.edu

Annu. Rev. Polit. Sci. 2017. 20:529–44

The *Annual Review of Political Science* is online at polisci.annualreviews.org

https://doi.org/10.1146/annurev-polisci-052615-025542



## Keywords

text as data, automatic coding, machine learning, computational social sciences

## Abstract

Text has always been an important data source in political science. What has changed in recent years is the feasibility of investigating large amounts of text quantitatively. The internet provides political scientists with more data than their mentors could have imagined, and the research community is providing accessible text analysis software packages, along with training and support. As a result, text-as-data research is becoming mainstream in political science. Scholars are tapping new data sources, they are employing more diverse methods, and they are becoming critical consumers of findings based on those methods. In this article, we first describe the four stages of a typical text-as-data project. We then review recent political science applications and explore one important methodological challenge—topic model instability—in greater detail.

## INTRODUCTION

Words are an integral part of politics. Officials and citizens use words to express opinions, make proposals, and defend their actions. Laws and regulations are also largely codified in words. Political scientists have always been interested in words, but a revolution has occurred that is creating unprecedented research opportunities (Cardie & Wilkerson 2008, Monroe & Schrodt 2008, Alvarez 2016). The internet is providing an avalanche of data related to politics. For example, all departments and agencies of the US federal government must now post their public records on the internet, and many other governments adhere to similar practices. Most major newspapers offer online access to their archives. Project Gutenberg and Google Books offer free access to the complete texts of millions of books. Social media sites such as Twitter and Facebook encourage researchers to use their data. The Internet Archive offers archival information about millions of government web pages dating back to 1996.

The research community has responded to this surfeit of data by developing accessible open-source text analysis libraries in R, Python, and other programming languages (e.g., Munzert et al. 2014). The combination of so many untapped research opportunities and accessible tools and training makes this an excellent time for specialists in all areas to invest in text. Legislative scholars can now systematically investigate floor speeches, constituent communications, revisions to laws and regulations, and much more. International relations scholars can systematically compare final treaties or agreements to hundreds of proposals made at earlier stages. Political theorists can explore political thought by searching across centuries of published works. This newfound ability to computationally investigate text (as well as many other innovative data sources, such as images and sound) will transform political science research as scholars become more adept at exploiting the available opportunities.

Because not all readers may be familiar with text-as-data research, we first provide an overview of the four stages of a typical project. This overview highlights key considerations for potential projects and provides context for appreciating recent developments and methodological challenges. We then review recent political science applications and explore one important methodological challenge—topic model instability—in greater detail.

## FOUR STAGES OF A TEXT AS DATA PROJECT

Text-as-data methods expand research opportunities for political scientists in two ways. First, they leverage the power of computing to make ambitious data collection tasks feasible. Second, they offer a growing number of options for analyzing large volumes of text quantitatively. A typical text-as-data project proceeds through four stages. Text must be obtained, converted to quantitative data, analyzed, and validated.

### Obtaining Text

The first stage of a project usually entails downloading digitized content. For many projects, this is now a fairly minor step. However, it is probably wise to investigate what will be required before committing to a project. Some sources make it easy for researchers to get exactly what they need, whereas extracting relevant information from other sources can be difficult and time consuming. Application user interfaces (APIs) enable users to "request" selected content from an underlying structured database using a single line of code. APIs are ideal when they include options that serve the needs of a project. Examples include the multiple APIs offered by the *New York Times* (e.g.,

Article Search API, Congress API), the Sunlight Foundation (e.g., Open States, Capitol Words), and prominent social media sites (e.g., Twitter, Facebook).

If an API is not available, the next best option in terms of ease of use is obtaining documents that are similarly formatted. Identical formatting makes it possible to write a single script to extract more specific content from many documents at once, such as the thousands of congressional bill texts available through the Government Printing Office. Almost all documents contain hidden formatting language that may also be helpful for systematically extracting more specific content. The look and feel of a web page come from embedded HTML or XML tags. These tags may do little more than format the visible text, but they can be used to isolate desired content [see, e.g., the @unitedstates project (**https://theunitedstates.io/**)]. Other types of documents (.doc, .docx, .txt, .pdf) also contain hidden formatting that may provide unique markers to facilitate splitting. Even text formatting can be helpful. In transcripts of Federal Reserve Board meetings, only the speaker's name is printed in all capital letters ("MS. YELLEN") and can be used to easily split transcripts by speaker statement.

The most challenging text extraction or "scraping" projects are those that draw content from diverse sources. For example, extracting the same content from many different candidate websites is challenging because each website has a different structure. One option is to write multiple scripts. The OpenStates project (**https://openstates.org**) recruited volunteer programmers to write scores of scripts to extract information about legislative bills for different state government websites. For less ambitious projects, crowdsourcing may be more practical. Sites such as Mechanical Turk and Crowdflower farm out small tasks to thousands of workers around the world. For a small fee (often a few cents), these workers will (for example) copy and paste website content. Another option is to collect simpler metrics at the source, such as counts of keywords, a common approach of many "big data" projects (Carneiro & Mylonakis 2009, Leskovec et al. 2009, Schmidt 2015).

## From Text to Data

The content of each document must then be converted to quantitative data. Frequently, the objective is to create a term–document or term–frequency matrix where each row is a document and each column is a feature found in at least one of those documents.[1] Thus, at this stage researchers need to decide on the appropriate unit of analysis. For example, US presidential State of the Union addresses (SOUs) are lengthy and cover many different subjects; a project that examines SOU policy topics will probably be improved by splitting each address into more focused paragraphs or sentences.

The next step is to specify which features within each document will be used in the quantitative analysis. The starting point is usually to treat every unique word as a separate feature. Researchers then exclude document content that is thought to be irrelevant to the analysis and potentially misleading. Standard options include removing punctuation, common words (stopwords), very infrequent words (sparse terms), and word suffixes (stemming). However, each of these actions deserves careful consideration. For example, standard stopwords such as "can't" and "cannot" might be relevant features for a study of presidential address tone. The next step may be to create features beyond the basic bag of words. One common practice is to include word pairs (bigrams) as additional features. But the possibilities are truly endless. Instead of treating synonyms as separate

[1]The cell values indicate whether a feature is present (0,1) in a term–document matrix, or how often it is found (0,N) in a term–frequency matrix.

words, researchers might combine them into a single feature. They might also assign more weight to features that are thought to be especially informative, or create new features from outside information. Roberts et al. (2016) find that incorporating information about whether a blog has liberal or conservative leanings helps to predict its topics.

## Quantitative Analysis of Text

Simple metrics can be very useful and have the added virtues of transparency and replicability. Eggers & Spirling (2017) study parliamentary dynamics by examining frequencies of specific word usage across time. Casas et al. (2016) use lists of positive and negative words to study how the media portrays protestors. However, today much of the focus (some would say hype) is on statistical machine learning methods. Scholars continue to debate, water-cooler style, the differences between machine learning and statistics. We are certainly not going to settle that debate, but we do think that the distinction can help to highlight general differences in approach. Political scientists are accustomed to using statistical methods to test theories. They choose the best model for the data (ordinary least squares, logistic regression, etc.) before testing model specifications that include a limited number of theoretically derived input (independent) variables. The focus is typically on the coefficients or parameters for the input variables—e.g., other things equal, are women significantly more likely to identify as Democrats than men? Whether the model accurately predicts the partisan identification of each voter is usually of secondary concern.

In machine learning research, the focus is usually on the outputs rather than the inputs. Instead of asking whether women are more likely to identify as Democrats, a more typical objective would be to predict state-level political opinion using Twitter (Beauchamp 2017). This focus on outputs leads researchers to be more concerned with prediction accuracy and less concerned with explanation. Beauchamp reports the features most associated with pro-Obama and pro-Romney poll shifts but does not try to explain why (for example) the most important predictor for Obama support is "75" and the most important for Romney is "cia." The focus on prediction also encourages more experimentation with different algorithms and features (Domingos 2015). We review some of the most relevant machine learning applications later in this article.

## Evaluating Performance

Validation is a critical component of every text-as-data project (Saldana 2009, Grimmer & Stewart 2013). For some methods validation is straightforward. Supervised machine learning results are validated by comparing an algorithm's predictions to pre-existing "gold standard" results. These may be documents labeled by human annotators, but there are many other possibilities. The gold standard for Beauchamp (2017) are state-level public opinion polls. In computer science, researchers frequently take advantage of online ratings and reviews to train and validate algorithms capturing sentiment. To guard against overfitting, researchers typically train the algorithm on one set of labeled examples before testing accuracy using a different, held-out, set.[2] Whether the gold standard labels validly capture the phenomenon of interest is a separate (and important) question. For other methods, where no gold standard is available, validation is typically multifaceted. For unsupervised machine learning methods, scholars have delved into specific examples within

[2] Repeating this process several times, using different training and testing sets, and then aggregating the validation results (N-fold cross-validation), is an even better approach (Kohavi 1995, Arlot 2010).

topics to show that the topics make sense; demonstrated that different algorithms produce similar clusters; and established that variations in topic emphasis across time or venues correlate with real-world events (Blei & Lafferty 2009, Quinn et al. 2010, Grimmer & King 2011, Roberts et al. 2014).

## RECENT DEVELOPMENTS IN POLITICAL SCIENCE

The purpose of this section is to provide a sense of the research opportunities available for political scientists. We make no attempt to be comprehensive but instead focus on four general research objectives. Two (classification and scaling) will be familiar to many readers (Grimmer & Stewart 2013). The other two (text reuse and semantics) have received less attention to date.

### Classification

Classification is a popular objective of text-as-data projects. Unsupervised machine learning methods [e.g., K-means, principal components analysis (PCA), latent Dirichlet allocation (LDA)] compare the similarity of documents based on co-occurring features. Despite their name, unsupervised methods require a lot of input from the user, who must (among other things) specify the number of topics in advance and interpret their meaning. In one of the earliest applications by political scientists, Quinn et al. (2010) used an unsupervised learner to classify Senate speeches by policy topic. They then validated their results by showing that their topics were similar to those developed using more time-consuming methods. Bousaills & Coan (2016) and Farrell (2016) use topic modeling to investigate climate change "skepticism" in reports and communications by think tanks and interest groups. Grimmer & King (2011) demonstrate how unsupervised methods can lead to new discoveries. They find that congressional press releases cluster in ways that match Mayhew's (1974) typology of constituent advertising, position taking, and credit claiming, but they also observe an additional cluster they label "partisan taunting" (see also Grimmer 2013). Roberts et al. (2014) show how incorporating additional information about documents (beyond the bag of words) into topic models can aid in interpretation of open-ended survey responses.

Whereas unsupervised methods are often used for discovery, supervised learning methods are primarily used as a labor-saving device. For example, Workman (2015) and Collingwood & Wilkerson (2011) use supervised methods to apply a well-established Policy Agendas topic-coding system to new research domains (federal regulations and congressional bills). Boydstun et al. (2016) are currently labeling thousands of newspaper articles for issue frame with the long-term goal of developing a supervised learner that can predict frames in other articles. The fact that supervised methods often require thousands of training examples makes them a nonstarter for many researchers and projects. However, there are often creative ways to reduce the effort required. Examining 250,000 Enron emails, Drutman & Hopkins (2013) use simple identification techniques to first exclude the 99% that were not political in nature. Crowdsourcing is also frequently used to build training sets in computer science. When a project does not require individual document labels, ReadMe is a supervised method that reliably predicts class proportions using a much smaller number of training examples (Hopkins & King 2010). King et al. (2013) use ReadMe to classify millions of social media posts by topic in a study of government censorship in China. Ceron et al. (2014) use it to study citizens' policy preferences in Italy and France.

Sentiment analysis is another important area of classification research where supervised and unsupervised methods are often used. The objective is to classify text ordinally (from negative to positive, for example) rather than categorically. Because businesses care about how consumers are responding to their products online, sentiment analysis is a well-funded area of research in

computer science. As a result, political scientists can take advantage of many pre-existing training corpora for a wide variety of research domains.[3]

## Scaling

Some of the earliest applications of automated text analysis in political science focused on using speeches and manifestos to locate European political parties in continuous ideological space (Laver et al. 2003, Lowe 2008, Slapin & Proksh 2008). Subsequent research has extended this by employing new methods and investigating new domains. In a pathbreaking study, Benoit et al. (2016) show that crowdsourcing can be a viable, even preferable, alternative to expert-based approaches to locating parties on policy dimensions. Kluver (2009) uses statements by interest groups and EU regulators to estimate ideological positions and gauge influence. Diermeier et al. (2012) test several different approaches to estimating legislator ideology from statements in the *Congressional Record* (see also Lauderdale & Herzog 2016). Barbera (2015) uses Twitter data and information about posters' followers to estimate the ideological positions of politicians, parties, and individual citizens. Lauderdale & Clark (2015) combine past votes with topic modeling of judicial opinions to critique single-dimensional scaling of justices and to develop separate estimates of judicial ideology for different issue areas.

## Text Reuse

Text reuse, as the name implies, is about discovering instances of similar language usage. The distinctive feature of text reuse algorithms is that they explicitly value word sequencing in judging document similarity. Political scientists have recently employed them to trace the origins of policy proposals in legislation (Wilkerson et al. 2015), to study the influence of interest groups in state legislatures (Hertel-Fernandez & Kashin 2015),[4] and to study party messaging strategies (Jansa et al. 2015). Other possibilities yet to be exploited by political scientists include studying the diffusion of political memes and contagion effects in new and old media (Leskovec et al. 2009, Smith et al. 2013). Different algorithms also support different types of analyses. Global alignment approaches (e.g., Needleman-Wunsch 1970) measure the overall similarity of documents whereas local alignment approaches (e.g., Smith & Waterman 1981) identify and score shared word sequences within documents. Thus, in a study of lawmaking or treaty negotiations, a global alignment approach might be used to see how much the entire proposal changes as it moves from one stage of the process to the next, whereas a local alignment approach could be used to investigate the fates of more specific provisions or proposals.

## Natural Language Processing

Social network analysis often employs text to investigate relationships among actors (Ward et al. 2011). Natural language processing (NLP) makes it possible to go beyond simply establishing connections to investigating the state of relationships—moving from "whom?" to "who did what to whom?" (Van Atteveldt et al. 2016). For example, political event data analysis draws on media reports to systematically monitor interactions between international actors. Instead of simply counting the number of times two actors are mentioned in reports, event data analysis incorporates

[3] Examples include **http://www.cs.cornell.edu/home/llee/data/** and **http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/**.

[4] See also the Legislative Influence Detector project (**https://dssg.uchicago.edu/lid/**).

syntax (sentence structure) and semantics (word meaning) to systematically track whether a relationship is improving or worsening and (possibly) to attribute credit or blame for developments.

Early event data research relied on human annotators to develop dictionaries of named entities and actions (Schrodt & Gerner 1994, Gerner et al. 2014). More recent research seeks to dramatically expand the scope of this research by taking advantage of extensive NLP resources developed by computer scientists and linguists (Leetaru & Schrodt 2013; see Ward et al. 2013 for an overview). For example, the Stanford Parser and the Stanford Named Entity Recognizer can be used to automatically extract specific parts of speech from documents and to tag different references to the same entity (e.g., USA, America, United States). Other valuable resources such as Wordnet can be used to identify synonyms for similar actions or sentiments. Denny et al. (2015) demonstrate how NLP methods can be used to systematically isolate the substantive provisions in legislation that typically includes lots of irrelevant "boilerplate" language. The creative possibilities are extensive, and Bird et al. (2009) provide an excellent primer on available NLP resources.

## TOPIC MODEL INSTABILITY AND A CALL FOR GREATER ATTENTION TO ROBUSTNESS IN TEXT-AS-DATA RESEARCH

In this final section, we shift from providing an overview of the field to delving into one contemporary challenge in more detail. Unsupervised machine learning methods (topic models) are very popular in political science in part because they classify documents without the extensive labeling efforts often required for supervised learning methods. The common practice has been to report and validate a single topic model after comparing results for several different models that vary by the number of topics specified by the researcher. This choice is usually based on the researcher's subjective judgment about which model's clusters best reflect the substantive goals of the project.

Above, we noted that the absence of a gold standard makes validation more challenging for these methods. A second challenge is model instability. Chuang et al. (2015, figure 1) estimate the same structural topic model 50 times to find that only two of 25 topics persist across all of the estimations. This can happen because different estimations can converge at different local maxima (Roberts et al. 2015). In a second experiment, the same authors find that manipulating just one feature of a structural topic model also leads to very different results (**Figure 1**). Many machine learning packages remove rarely used words by default to reduce processing time and avoid overfitting. In **Figure 1**, varying only this feature leads to important differences in terms of the topics that emerge from different estimations of the same model.

A number of recent studies have proposed different ways to assess and respond to topic-model instability (Grimmer & King 2011, Schmidt 2012, Boyd-Graber et al. 2014, Roberts et al. 2014). However, the focus, as far as we are aware, continues to be on selecting and validating a single best model. In conventional statistical studies, researchers try to demonstrate that their results are robust by reporting results for multiple model specifications. A study examining gender and voting will test and report several combinations of theoretically derived independent variables to demonstrate that the central findings persist. Supervised machine learning analyses also commonly address robustness by basing results on the consensus prediction of an ensemble rather than a single algorithm. Grimmer & King (2011) propose comparing topic model results for different algorithms but do not incorporate those differences into their findings.

Robustness can be evaluated with respect to methods, parameters, features, and data partitions. No study can consider all permutations, but we do think that political scientists using text-as-data methods should explicitly address robustness in their results. Do the central findings stand up to reasonable variations in modeling choices? Where topic models are concerned, one option is to move away from the current convention of reporting results for a single model.

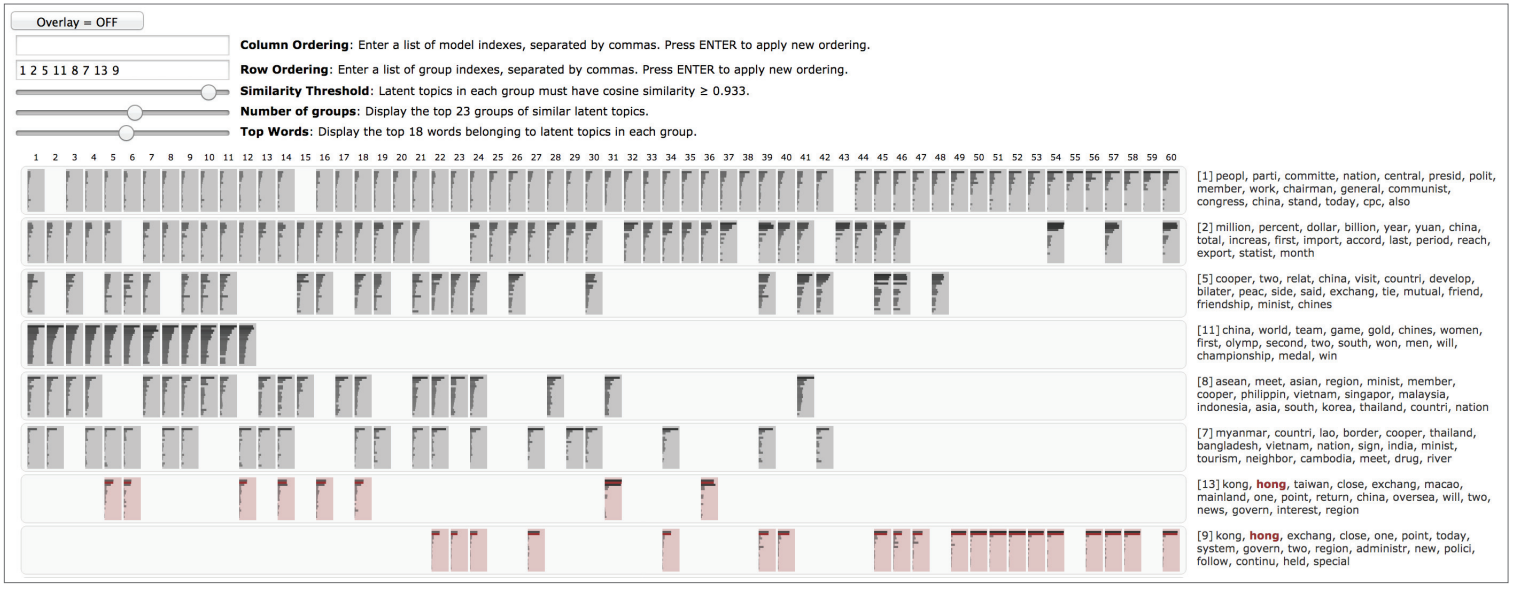


**Figure 1**

Impact of a feature on topic stability (from Chuang et al. 2015). Each of the 50 columns is a 25-topic latent Dirichlet allocation model where the only difference is the threshold used to exclude sparse terms. Each row is a topic. The shaded cells indicate when a model includes the topic. Chuang et al. assume that two models share the same topic if the cosine similarity of the topic terms is greater than 0.9. Darker shades indicate higher similarity.

## Exploring the Topics of Legislators' Floor Speeches

In this section, we illustrate how topic robustness can inform a study of congressional floor speeches.[5] Members of the US House of Representatives gave almost 10,000 "one-minute" floor speeches during the 113th Congress (2013–2014). These speeches are given before ordinary business and are primarily intended for public consumption (Schneider 2015; **https://www.fas.org/sgp/crs/misc/RL30135.pdf**). A quick review indicates that their subjects are often quite diverse. Some honor constituent accomplishments (such as a state basketball championship), whereas others address political and legislative issues. However, to our knowledge, no one has systematically investigated what members talk about in these speeches. What topics are covered and which are the most common? Do Republicans and Democrats tend to talk about the same issues or emphasize different ones?

To examine these questions, we first used the Sunlight Foundation's Capitol Words API to download all member statements from the *Congressional Record* of the 113th Congress. We then removed statements that did not begin with the opening phrase of a one-minute speech: "Mr. Speaker, I rise today . . . ." This produced a corpus of 5,346 one-minute speeches given by 179 Democrats and 4,358 given by 213 Republicans. We converted the words in each speech to lower case and removed punctuation, stopwords, word stems, and words of two characters or fewer. Finally, we constructed a term–document matrix where each row is a one-minute speech and each column is a vector indicating whether a feature/word is present in a given speech.

The next step was to estimate a series of latent Dirichlet allocation (LDA) models where the number of topics ($k$) ranges from 10 to 90 in five topic increments (Blei et al. 2003). These 17 models yield 850 topics ($10 + 15 + 20 \ldots + 90$). To determine which topics were robust, we first calculated cosine similarity[6] for all topic pairs (resulting in 722,500 similarity scores) and then used the Spectral Clustering algorithm to group the 850 topics based on cosine similarity. The Spectral Clustering algorithm does this by maximizing average intra-cluster cosine similarity for a given number of clusters $c$. The substance of a given cluster can then be investigated by examining the most predictive words ("top terms") in each cluster.

Ten thousand speeches by 435 lawmakers should cover a diverse set of topics. On the other hand, dividing the speeches into too many clusters may complicate the analysis without improving the overall fit (average intra-cluster similarity) of the model to the data. **Figure 2** displays how fit improves as the number of speech clusters ($c$) is varied from five to 100. More clusters improve overall fit until approximately 50 clusters. The average similarity of the clusters excluding the largest catch-all "melting pot" cluster is also quite volatile until approximately 50 clusters. We therefore base our analysis on the robust topics from a 50-cluster model.

After clustering the 850 topics from 17 models into 50 clusters (see **Figure 3**), we grouped some of the clusters into what we will call metatopics. For example, the education metatopic includes three clusters about more specific aspects of education. All of the topics in which we were unable to discern a consistent theme were excluded by assignment to one "unclear" metatopic. Thus, the results presented are based on 697 of the original 850 topics from 16 of the 17 original topic models. In **Figure 4**, the education metatopic, for example, includes 37 topics found in 14 different topic models. In our view, the figure underscores the drawbacks of presenting results

[5]Supplemental and replication materials for this section can be found in the **Supplemental Materials** section of the Annual Reviews website (**http://www.annualreviews.org/db/suppl**) and at **https://github.com/CasAndreu/wilkerson_casas_2016_TAD**. These materials include a Python module, `rlda`, to apply the robust latent Dirichlet allocation models used here (**https://github.com/CasAndreu/rlda**).

Supplemental Material

[6]For each possible pair of topics, $cos(\theta) = \frac{a \cdot b}{||a|| ||b||}$, where $a$ and $b$ are vectors of counts recording topic-word assignments in the final estimation iteration.

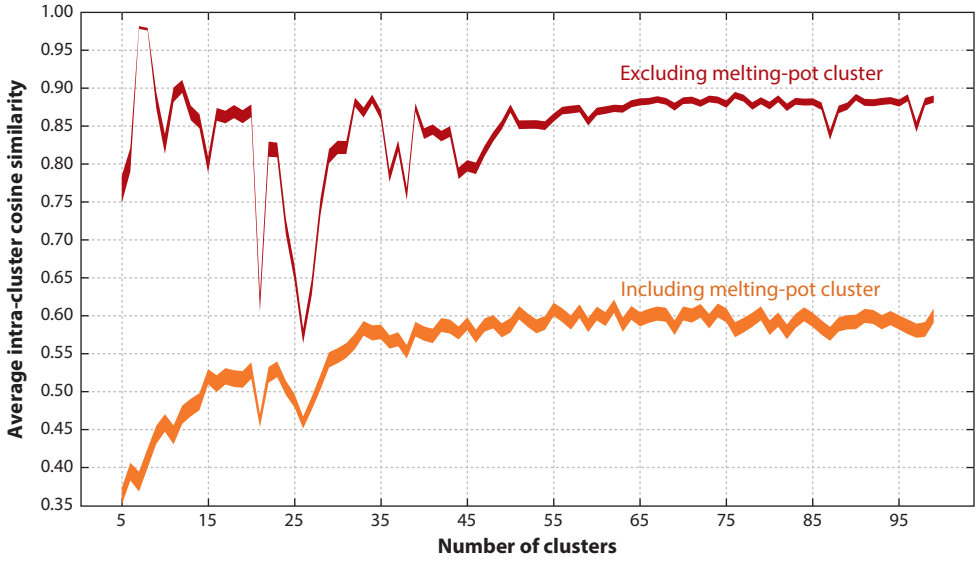


**Figure 2**

Number of clusters and average intra-cluster similarity. The lower line indicates that more clusters improve overall fit until approximately 50 clusters. The upper line indicates that the average similarity of the clusters excluding the largest catch-all "melting pot" cluster is also quite volatile until approximately 50 clusters.

based on a single model. Topics that are common to many models are often missing from any one of them.

## Topic Attention in One-Minute Speeches

In an LDA model, the topics exist before the documents (see Blei et al. 2003, pp. 996–97). Each document is assumed to be about each topic with some positive probability. To study speech attention, we must first label individual speeches for primary topic. We assume each speech is about its most probable topic. Thus, we classify 9,704 speeches for each of 16 topic models. We then report results for only those topics from each model that are part of the 21 metatopics.

**Figure 5** displays those results. For example, for education, the consensus of the different topic models is that Democrats gave more speeches about education than Republicans did. It is reassuring

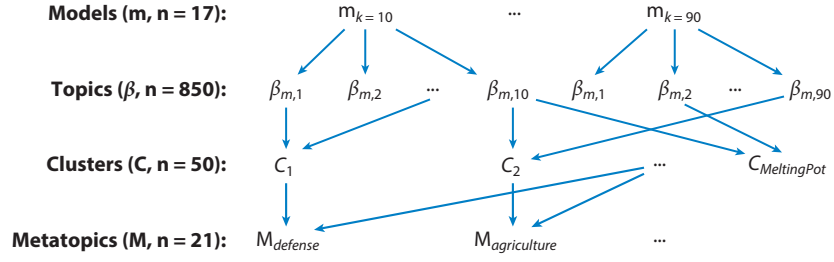


**Figure 3**

Workflow of moving from 17 topic models to 21 metatopics.

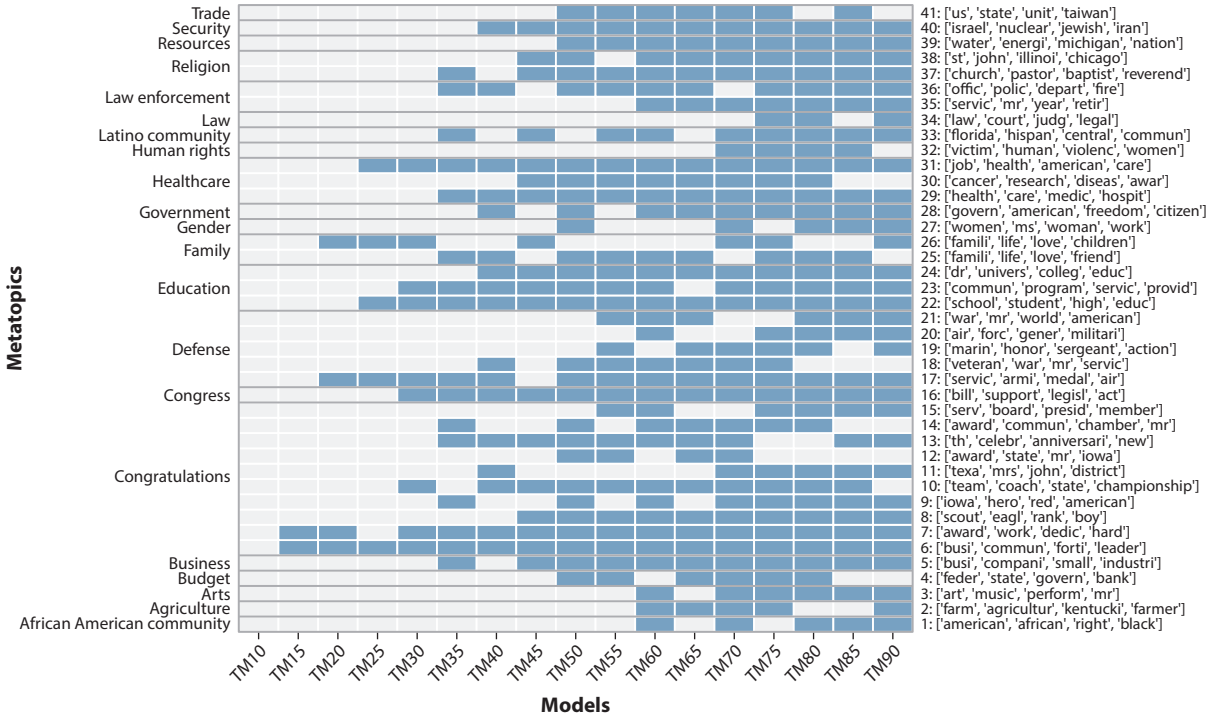


**Figure 4**

The 21 metatopics of a 50-cluster model. Each column is one of the 17 LDA topic models (ranging from 10 to 90 topics), and each row is a topic cluster. The 21 substantive metatopics are listed on the left. The shaded cells indicate where the topics in each cluster or metatopic originate.

that the models generally agree concerning partisan emphasis for most of the metatopics. However, there is often considerable disagreement regarding the size of the difference.

The average amount of attention (across the models) given to different metatopics (such as education) by Republicans and Democrats is displayed on the far right. The results indicate that Republicans were most likely to give "Congrats" speeches (17%), followed by speeches about "Defense" (8%), "Education" (7%), and "Family" (6%). Democrats also gave lots of congratulatory speeches (9%), but were as likely to give speeches about education (9%), followed by health care (7%) and family (5%).

## Validation

We think that similarity of estimates of topic emphasis across different models is an important type of validation. **Figure 5** should inspire confidence in the robustness of general differences in speech topic emphasis but less confidence in the amount of difference in many cases.

Our results generally support Petrocik's (1996) "issue ownership" argument. The main exception seems to be agriculture. According to Petrocik, Democrats own the issue of agriculture, whereas most of the models of our analysis indicate that Republicans own it. We therefore took a closer look at who was giving speeches about agriculture and found a strong correlation between the proportion of a member's speeches that were about agriculture and the number of district

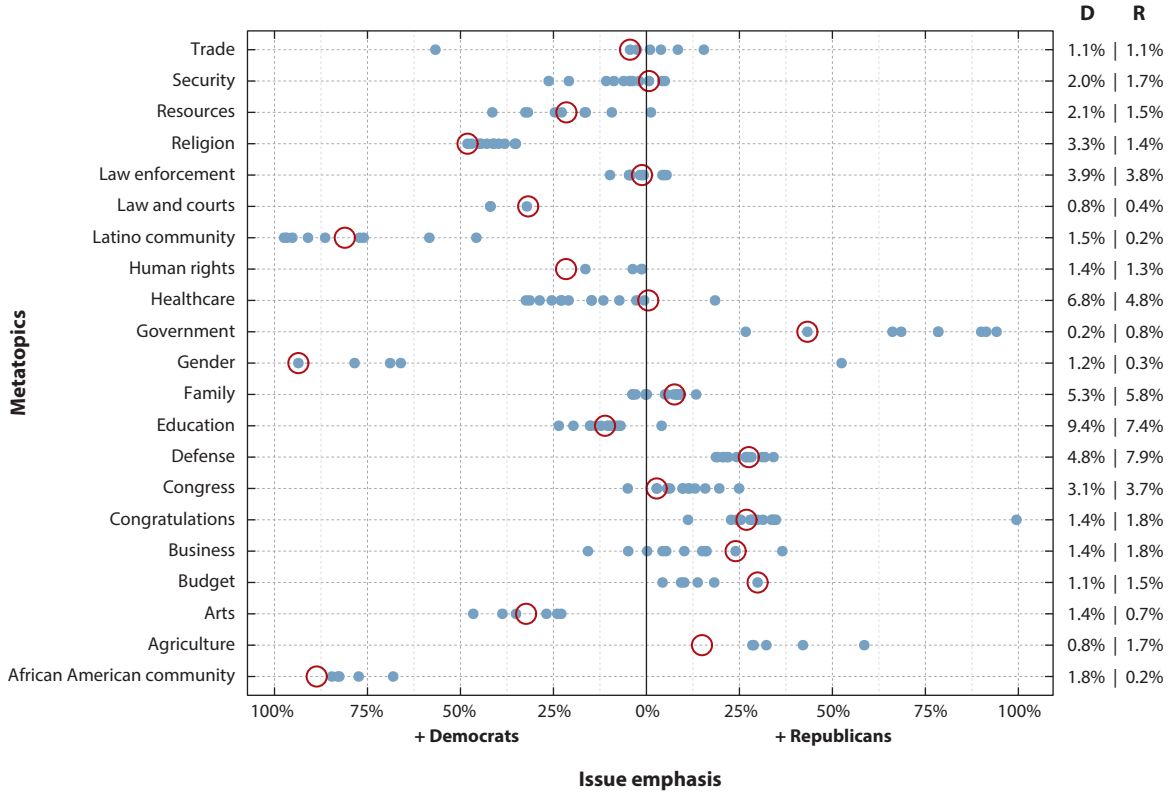


**Figure 5**

A robust examination of issue emphasis in one-minute speeches. Each row is one of the 21 metatopics. Each dot is a result for one topic model. The average amount of attention (across the models) given to different metatopics by Republicans and Democrats is displayed on the far right. The red circles display relative topic attention for a single ($k = 50$) topic model.

workers employed in the agriculture, forestry, fishing, hunting, and mining industries (Pearson's $r = 0.4$).[7] Thus, it seems likely that there has been a transfer of ownership on this issue since Petrocik's article was published 20 years ago.

## DISCUSSION

Computerized text analysis is transforming political science research because scholars now have the ability to explore massive amounts of politically relevant text using increasingly sophisticated tools. These developments have already produced important advances in research methods (Hopkins & King 2010, Benoit et al. 2016), opened the door to new research questions (Wilkerson et al. 2015), and altered current understandings (Lauderdale et al. 2015). We have argued that researchers do not need to be computer programmers or statistical methodologists to use text-as-data methods in their research. They do need to be attentive to the same concerns about validity and reliability that apply to all methods.

Supplemental Material

[7] See the appendix (**Supplemental Materials**; **http://www.annualreviews.org/db/suppl**) for more details.

The other area where political scientists are making important advances is in assessing the quality of findings. Recent studies are examining the sensitivity of findings to alternative feature specifications (Spirling & Denny 2017) and proposing new approaches to explicitly incorporate information about reliability into research findings (Grimmer et al. 2016). Unsupervised learning methods (such as topic models) are among the more popular methods used in political science. A central attraction for many researchers is that they do not require labeled training sets. To be sure, supervised learning methods have their own limitations. However, the absence of any gold standard makes choosing and validating a model even more challenging for unsupervised methods.

Scholars have recently proposed new ways of selecting among alternative topic models, for example by examining the cohesiveness and distinctiveness of the topic words (Roberts et al. 2014). A robust approach to reporting topic-model results takes advantage of the information provided by alternative specifications. This approach has its own limits, but in our view it is informative and transparent and adheres to current conventions that lead researchers to explicitly address robustness in statistical studies.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The authors thank Jeffrey Arnold, Noah Smith, Jason Chuang, and an anonymous reviewer for helpful feedback.

## LITERATURE CITED

Alvarez RM, ed. 2016. *Computational Social Science: Discovery and Prediction. Analytical Methods for Social Research*. New York: Cambridge Univ. Press

Arlot C. 2010. A survey of cross-validation procedures for model selection. *Stat. Surv.* 4:40–79

Barbera P. 2015. Birds of the same feather tweet together. Bayesian ideal point estimation using Twitter data. *Polit. Anal.* 23(1):76–91

Beauchamp N. 2017. Predicting and interpolating state-level polls using Twitter textual data. *Am. J. Polit. Sci.* In press. doi: 10.1111/ajps.12274

Benoit K, Conway D, Lauderdale B, Laver M, Mikhaylov S. 2016. Crowd-sourced text analysis: reproducible and agile production of political data. *Am. Polit. Sci. Rev.* 110(2):278–95

Bird S, Klein E, Loper E. 2009. *Natural Language Processing with Python—Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media

Blei D, Lafferty J. 2009. Topic models. In *Text Mining: Classification, Clustering, and Applications*, ed. AN Srivastava, M Sahami, pp. 71–94. Data Mining and Knowledge Discovery Ser. Boca Raton, FL: Chapman & Hall/CRC

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Boussalis C, Coan TG. 2016. Text-mining the signals of climate change doubt. *Glob. Environ. Change* 36:89–100

Boyd-Graber J, Mimno D, Newman D. 2014. Care and feeding of topic models: problems, diagnostics, and improvements. In *Handbook of Mixed Membership Models and Their Applications*, pp. 3–34. Boca Raton, FL: CRC Press

Boydstun A, Butters R, Card D, Gross J, Resnik P, Smith N. 2016. *Under what conditions does media framing influence public opinion on immigration?* Presented at Annu. Meet. Midwest Polit. Sci. Assoc., Chicago, IL, Apr. 7–9

Cardie C, Wilkerson J. 2008. Text annotation for political science research. *J. Inf. Technol. Polit.* 5(1):1–6

Carneiro HA, Mylonakis E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.* 49(10):1557–64

Casas A, Davesa F, Congosto M. 2016. The media coverage of a connective action: the interaction between the 15-M Movement and the mass media. *Rev. Espan. Investig. Sociol.* 155:73–96

Ceron A, Curini L, Iacus SM, Porro G. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media Soc.* 16(2):340–58

Chang J, Boyd-Graber J, Wang C, Gerrish S, Blei DM. 2009. Reading tea leaves: how humans interpret topic models. In *Advances in Neural Information Processing Systems*, ed. Y Bengio, D Schuurmans, J Lafferty, CKI Williams, A Culotta, pp. 288–96. Cambridge, MA: MIT Press

Chuang J, Roberts M, Stewart B, Weiss R, Tingley D, et al. 2015. TopicCheck: interactive alignment for assessing topic model stability. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pp. 175–84. Denver, CO: Assoc. Comput. Linguist.

Chuang J, Wilkerson JD, Weiss R, Tingley D, Stewart BM, et al. 2014. *Computer-assisted content analysis: topic models for exploring multiple subjective interpretations.* Presented at Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning, Montreal, Dec. 8–13

Collingwood L, Wilkerson J. 2011. Tradeoffs in accuracy and efficiency in supervised learning methods. *J. Inf. Technol. Polit.* 4:1–28

Denny MJ, O'Connor B, Wallach H. 2015. *A little bit of NLP goes a long way: finding meaning in legislative texts with phrase extraction*. Presented at Annu. Meet. Midwest Polit. Sci. Assoc., 73rd, Apr. 16–19

Denny MJ, Spirling A. 2017. *Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it.* Unpublished manuscript, Dep. Polit. Sci, Stanford Univ and Inst. Quant. Soc. Sci., Harvard Univ. **https://ssrn.com/abstract=2849145**

Diermeier D, Yu B, Kaufmann S, Godbout JE. 2012. Language and ideology in Congress. *Br. J. Polit. Sci.* 42(1):31–55

Domingos P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books

Drutman L, Hopkins DJ. 2013. The inside view: using the Enron email archive to understand corporate political attention. *Legis. Stud. Q.* 38(1):5–30

Eggers A, Spirling A. 2017. The shadow cabinet in Westminster systems: modeling opposition agenda setting in the House of Commons, 1832–1915. *Br. J. Polit. Sci.* In press

Farrell J. 2016. Corporate funding and ideological polarization about climate change. *PNAS* 113(1):92–97

Gerner DJ, Schrodt PA, Francisco RA, Weddle JL. 2014. Machine coding of event data using regional and international sources. *Int. Stud. Q.* 38(1):91

Grimmer J. 2013. Appropriators not position takers: the distorting effects of electoral incentives on congressional representation. *Am. J. Polit. Sci.* 57(3):624–42

Grimmer J, King G. 2011. General purpose computer-assisted clustering and conceptualization. *PNAS* 108(7):2643–50

Grimmer J, King G, Superti C. 2016. *The unreliability of measures of intercoder reliability, and what to do about it.* Unpublished manuscript, Dep. Polit. Sci., Stanford Univ. **http://web.stanford.edu/~jgrimmer/Handbib.pdf**

Grimmer J, Stewart BM. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21(3):267–97

Hertel-Fernandez A, Kashin K. 2015. *Capturing business power across the states with text reuse*. Presented at Annu. Meet. Midwest Polit. Sci. Assoc., Chicago, IL, Apr. 16–19

Hopkins DJ, King G. 2010. A method of automated nonparametric content analysis for social science. *Am. J. Polit. Sci.* 54(1):229–47

Huang A. 2008. Similarity measures for text document clustering. In *Proc. Sixth New Zealand Computer Science Research Student Conference*, pp. 49–56. Christchurch, New Zealand: NZCSRSC

Jansa J, Hansen E, Gray V. 2015. *Copy and paste lawmaking: the diffusion of policy language across American state legislatures*. Work. Pap., Dep. Polit. Sci., Univ. North Carolina, Chapel Hill

Jockers ML. 2014. *Text Analysis with R for Students of Literature*. New York: Springer

King G, Pan J, Roberts ME. 2013. How censorship in China allows government criticism but silences collective expression. *Am. Polit. Sci. Rev.* 107(2):326–43

Kluver H. 2009. Measuring interest group influence using quantitative text analysis. *Eur. Union Polit.* 10(4):535–49

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. Int. Joint Conf. Artificial Intelligence*, pp. 1137–43. San Francisco: Morgan Kaufmann

Lauderdale BE, Clark TS. 2014. Scaling politically meaningful dimensions using texts and votes. *Am. J. Polit. Sci.* 58(3):754–71

Lauderdale BE, Herzog A. 2016. Measuring political positions from legislative speech. *Polit. Anal.* 26:374–94

Laver M, Benoit K, Garry J. 2003. Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.* 2:311–31

Leetaru K, Schrodt P. 2013. *GDELT: global data on events, location, and tone, 1979–2012*. Presented at International Studies Association Annu. Conv., San Francisco, CA, Apr.

Leskovec J, Backstrom L, Kleinberg J. 2009. *Memetracking and the dynamics of the news cycle*. Presented at ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), Paris, June

Lowe W. 2008. Understanding Wordscores. *Polit. Anal.* 16(4):356–71

Mayhew DR. 1974. *Congress: The Electoral Connection*. New Haven, CT: Yale Univ. Press

Monroe BL, Schrodt PA. 2008. Introduction to the special issue: the statistical analysis of political text. *Polit. Anal.* 16(4):351–55

Munzert S, Rubba C, Meissner P, Nyhuis D. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Hoboken, NJ/Chichester, UK: Wiley & Sons

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48(3):443–53

Petrocik JR. 1996. Issue ownership in presidential elections, with a 1980 case study. *Am. J. Polit. Sci.* 40(3):825–50

Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR. 2010. How to analyze political attention with minimal assumptions and costs. *Am. J. Polit. Sci.* 54(1):209–28

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, et al. 2014. Structural topic models for open-ended survey responses: structural topic models for survey responses. *Am. J. Polit. Sci.* 58(4):1064–82

Roberts M, Stewart B, Tingley D. 2016. Navigating the local modes of big data: the case of topic models. In *Computational Social Sciences*, ed. RM Alvarez, pp. 51–97. New York: Cambridge Univ. Press

Saldana J. 2009. *The Coding Manual for Qualitative Researchers*. Los Angeles: Sage

Schmidt BM. 2012. Words alone: dismantling topic models in the humanities. *J. Digit. Humanit.* (2)1. **http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/**

Schmidt B. 2015. Is it fair to rate professors online? *New York Times*, Dec. 16, Sec. Room for Debate

Schneider J. 2015. *One-minute speeches: current house practices*. Congr. Res. Serv. Rep. 7-5700, 1-7

Schrodt PA, Gerner DJ. 1994. Validity assessment of a machine-coded event data set for the Middle East, 1982–92. *Am. J. Polit. Sci.* 38(3):825

Slapin JB, Proksch S-O. 2008. A scaling model for estimating time-series party positions from texts. *Am. J. Polit. Sci.* 52(3):705–22

Smith DA, Cordell R, Dillon EM. 2013. Infectious texts: modeling text reuse in nineteenth-century newspapers. In *Proc. IEEE Int. Conf. Big Data*, pp. 86–94. Santa Clara, CA: Inst. Electrical and Electronics Engineers

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147(1):195–97

Van Atteveldt W, Shenhav SR, Fogel-Dror Y. 2017. Clause analysis: using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 Gaza War. *Polit. Anal.* In press

Wallach H, Dicker L, Jensen S. 2010. An alternative prior for nonparametric Bayesian clustering. In *Proc. Thirteenth International Conference on Artificial Intelligence and Statistics, May 13–15, 2010, Chia Laguna Resort, Sardinia, Italy*, ed. YW Teh, M Titterington, 9:892–99. **http://www.jmlr.org/proceedings/papers/v9/**

Ward M, Beger A, Josh C, Dickenson M, Dorff C, Radford B. 2013. Comparing GDELT and ICEWS event data. *Analysis* 21:267–97

Ward M, Stovel K, Sacks A. 2011. Network analysis and political science. *Annu. Rev. Polit. Sci.* 14:245–64

Wilkerson J, Smith D, Stramp N. 2015. Tracing the flow of policy ideas in legislatures: a text reuse approach. *Am. J. Polit. Sci.* 59(4):943–56

Workman S. 2015. *The Dynamics of Bureaucracy in the US Government: How Congress and Federal Agencies Process Information and Solve Problems*. Cambridge, UK: Cambridge Univ. Press