A Data-Driven Voter Guide for U.S. Elections: Adapting Quantitative Measures of the Preferences and Priorities of Political Elites to Help Voters Learn About Candidates

ADAM BONICA

Internet-based voter advice applications have experienced tremendous growth across Europe in recent years but have yet to be widely adopted in the United States. By comparison, the candidate-centered U.S. electoral system, which routinely requires voters to consider dozens of candidates across a dizzying array of local, state, and federal offices each time they cast a ballot, introduces challenges of scale to the systematic provision of information. Only recently have methodological advances combined with the rapid growth in publicly available data on candidates and their supporters to bring a comprehensive data-driven voter guide within reach. This paper introduces a set of newly developed software tools for collecting, disambiguating, and merging large amounts of data on candidates and other political elites. It then demonstrates how statistical methods developed by political scientists to measure the preferences and expressed priorities of politicians can be adapted to help voters learn about candidates.

Keywords: ideal point estimation, text-as-data, supervised machine learning, voting advice applications

The onset and proliferation of web applications that help voters identify the party that best represents their policy preferences, commonly known as "voter advice applications," is among the most exciting recent developments in the practice and study of electoral politics (Alvarez et al. 2014; Louwerse and Rosema 2013; Rosema, Anderson, and Walgrave 2014). After their emergence in the early 2000s, they quickly spread throughout Europe and beyond and have since become increasingly popular among voters. In recent elections in Germany, the Netherlands, and Switzerland, upwards of 30 to 40 percent of the electorates used these tools to vote (Ladner, Felder, and Fivaz 2010). Despite their growing popularity, voter advice applications have yet to make significant headway in the United States. While voter advice applications have excelled in parliamentary democracies, which require data on the issue positions for a small number of parties, the multi-tiered, candidate-centered U.S. electoral system introduces challenges of size, scale, and complexity to the systematic provision of information.

Reformers have long advocated for greater disclosure and government transparency as a means to inform voters and enhance electoral accountability. In justifying the value of disclosure in *Buckley v. Valeo,* the Supreme Court wrote that "disclosure provides the electorate with information 'as to where political campaign money comes from and how it is spent by the candidate' in order to aid the voters in evaluating those who seek federal office. It allows voters to place each candidate in the political spectrum more precisely than is often

Adam Bonica is assistant professor of political science at Stanford University. He is also co-founder at Crowdpac Inc.

Direct correspondence to: Adam Bonica at bonica@stanford.edu, Stanford University, Encina Hall West, Room 307, 616 Serra St., Stanford, CA 94305-6044.

possible solely on the basis of party labels and campaign speeches."[1] Disclosure requirements have long been a central component of campaign finance regulation, churning out millions upon millions of records each election cycle. Yet despite the stringent disclosure requirements and reporting standards, making data transparent and freely available is seldom sufficient on its own. More is needed to translate this raw information into a truly useful resource for voters.

Thus far, the use of data-intensive applications in U.S. politics has primarily been in service of parties and campaigns. This is perhaps best exemplified by the Obama campaign's success in leveraging large-scale databases to learn about voters and predict their behavior, which was widely lauded following the 2012 elections (Issenberg 2012). However, the true potential of the data revolution in U.S. politics might very well be realized by harnessing its power to help voters, donors, and other consumers of politics learn about candidates. As political scientists are well aware, the information available on political elites—through what they say, how they vote, and how they network and fund-raise—is much richer and of higher quality than the information available on the mass public. Delivering on the promise of disclosure requires (1) a means of summarizing the information contained in the raw data into a format that is more easily interpreted but still highly informative, and (2) an intuitive platform for accessing data quickly and efficiently. The first is a familiar problem to social scientists, who have spent decades developing numerous data reduction methods to summarize revealed preference data. Only more recently has the possibility of developing a platform to enable voters to interact with the data come within reach.

This paper introduces a new database and modeling framework developed to power Crowdpac's new political information platform (Willis 2014). I begin with a discussion of how methods developed by political scientists to measure the policy preferences and expressed priorities of politicians can be adapted to help voters learn about candidates. For many of the same reasons they have proven useful to political scientists, there could be significant value in retooling these quantitative measures of political preferences for a wider audience. After providing an overview of the automated data collection and entity resolution techniques used to build and maintain the database, I introduce a modeling framework developed to generate issue-specific measures of policy preferences incorporating established methods for analyzing political text, voting records, and campaign contributions.

DEMOCRATIZING POLITICAL DATA

The U.S. electoral system imposes considerable informational costs on voters. Even for the most sophisticated of voters, filling out a ballot is a daunting task. Depending on the state, a typical ballot might ask voters to select candidates in dozens of races and decide on multiple ballot measures. The informational costs are particularly high in primary elections and other contests where voters are unable to rely on partisan cues and other informational shortcuts and find themselves unsure about which candidate is best aligned with their preferences.

Information is crucial to effective political participation. In the context of elections, an informed vote is a matter of being confident in a set of predictions about how the candidates under consideration would behave in office. The uncertainty experienced by voters in the polling booth can arise from many sources, but much of it is a consequence of information asymmetries rather than the capriciousness of politicians. Most politicians have well-defined policy preferences but often lack either the means or incentives to communicate them clearly to voters. Politicians rarely behave unpredictably when presented with familiar choices.

Even though disclosure data has been sold as providing a service to voters, those best positioned to utilize it have been campaigns, lobbyists, and special-interest groups. This is reflected in the market for political information. The most sophisticated data collection and analytics have gone into subscription fee–based services (such as Legistorm and Catalist) and

1. *Buckley v. Valeo* 96 S. Ct. 612; 46 L. Ed. 2d 659; 1976 U.S. LEXIS 16; 76-1 U.S. Tax Cas. (CCH) P9189.2.

are targeted at these groups. Insider knowledge about which politicians are power brokers, which races are likely to be competitive, and where to direct contributions is a service that lobbyists provide to high-powered clients.

Past efforts to build a comprehensive political information platform for voters in U.S. elections have all struggled with the problem of conveying meaningful signals about the policy preferences of non- incumbents and officeholders beyond a select group of legislative bodies. In the 1990s, Project Vote Smart adopted an innovative strategy for dealing with this asymmetry with respect to incumbency status. The National Political Awareness Test (NPAT) was a major effort to compile policy positions by surveying candidates. Unfortunately, after a decline in response rates, due in part to active discouragement by party leaders, NPAT achieved only limited success. Project Vote Smart has since shifted strategies and begun to code issue positions manually, based on candidates' public statements. This approach has shown some promise but is limited by issues of scalability, completeness, and coder reliability.

There are three main challenges in creating such a resource for the public: (1) automating the collection and maintenance of a large-scale database drawn from numerous sources; (2) devising effective ways to summarize and visualize data on candidates; and (3) designing a user interface that is easy to understand and follow for those with varying levels of political sophistication. This paper introduces a set of data-driven strategies to address these challenges.

DATA ARCHITECTURE

This section introduces the new database that serves as a central repository for data on candidates and political elites. The database draws on three main sources of data: political text, voting and legislative behavior, and campaign contributions.

A system of automated scrapers is used to collect and process new data as they become available. To ensure scalability, a new data source is not included until the feasibility of maintaining it with minimal human supervision has been established. Beyond automating the compiling and updating of the database, transforming the raw data into a usable format presented its own challenges. In particular, a solution was needed for merging and disambiguating data drawn from difference sources. This was managed with customized automated identity resolution and record-linkage algorithms supplemented by strategic use of human-assisted coding when identifying personal contributions made by candidates. Each of the three data sources is described in this section.

Political Text

Political text is any written or transcribed public statement by a political actor. In its current state, the database of text largely comprises documents originating from legislation and the *Congressional Record,* which contains transcripts of all proceedings, floor debates, and extensions of remarks in Congress. Congressional bill text is taken from Congress.gov. Additional contextual data on legislation, such as information on sponsorship, co-sponsorship, and committee activity, are also collected. Importantly, the Congressional Research Service (CRS) provides subject codes for each bill. These tags are used to train the topic model discussed later. The *Congressional Record* is taken from the Federal Digital System (FDsys) of the U.S. Government Printing Office (GPO). Each document in the database is linked to a candidate ID and, when applicable, a bill ID. Bill authorship is attributed to the sponsor(s). Speeches made during floor debates are attributed to the speaker and, when applicable, any bills specifically referenced during the speech. The text database currently includes over half a million documents.

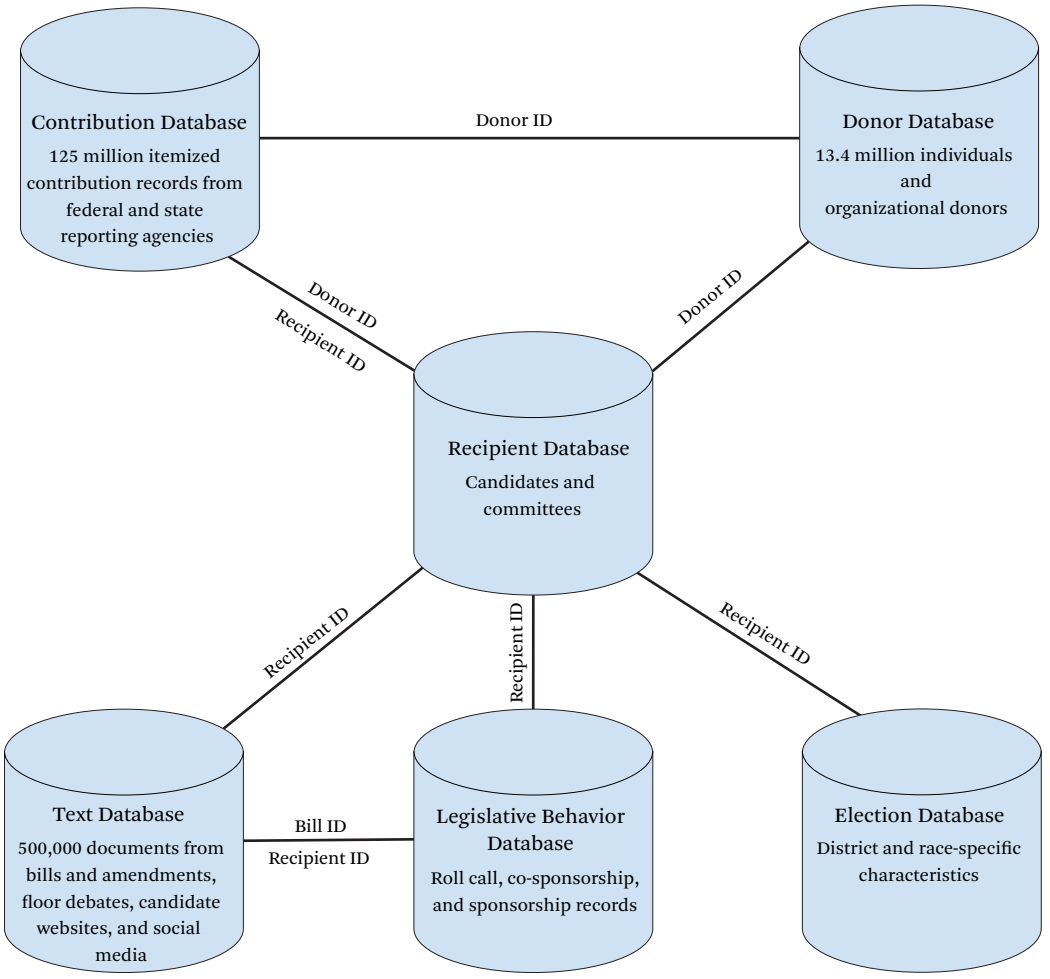
Legislative Voting

Congressional voting records are downloaded from voteview.com via the W-NOMINATE R package (Poole et al. 2011). Bills and amendments are assigned unique identifiers that provide a crosswalk to other tables in the database.

Campaign Contributions

Contribution records are drawn from an augmented version of the Database on Ideology,

Figure 1. Data Architecture of Database on Ideology, Money, and Elections (DIME)



Source: Author's calculations.

Money, and Elections (DIME) (Bonica 2014).[2] Since nearly every serious candidate for state or federal office engages in fund-raising (either as a recipient or a donor), campaign finance data provide the scaffolding for constructing the recipient database. Figure 1 presents a visual representation of the data architecture. The database consists of six tables corresponding to the different record types. The unique identifiers for candidates, donors, and bills serve as crosswalks between the tables. Each line in the figure indicate a crosswalk between two tables.

The recipient table plays a central role in structuring the data. It can be mapped onto each of the other databases by one or more crosswalks. It contains variables for numerous characteristics, including the office sought, bi-

2. For information on access to the database and reference documentation, see Adam Bonica, Database on Ideology, Money in Politics, and Elections (DIME): public version 1.0 (computer file) (Stanford, Calif.: Stanford University Libraries, 2013), http://data.stanford.edu/dime (accessed May 31, 2016).

ographical profiles, past campaigns and offices held, fund-raising statistics (for example, totals by source or amounts raised from donors within the district), committee assignments, and various other data rendered on the site. Each row represents a candidate-cycle observation. The recipient table currently includes 360,173 rows extending back to 1979, covering 105,967 distinct candidates and 38,689 political committees. Additional identity resolution processing is applied to candidates who have run for state and federal office to ensure that each one is linked to a single identifier.

The contribution table contains records of more than 125 million itemized contributions to state and federal elections. Each record maps onto the recipient database via the corresponding recipient ID. Contribution records can also be linked to the originating candidate or committee for the set of recipients who have donated via the contributor IDs. The donor table summarizes and standardizes the information in the contribution database into a more usable format with a single row per donor.

The text table includes documents scraped from legislative text for bills and amendments, floor speeches, candidate web pages, and social media accounts. Every document is linked to either a candidate from the recipient table or a bill or amendment from the legislative table—or both in the case of sponsored legislation.

By combining these data sources, a single database query can return a wealth of information on a candidate, including information on the candidate's ideology, fund-raising activity, and donors, his or her personal donation history, sponsored and co-sponsored legislation, written and spoken text, voting records, electoral history, personal and political biographies, and more. All of the data sources needed to replicate the database schema are available for download as part of DIME or as part of a supplemental database of legislative text and votes titled DIME+.[3]

The remaining sections explain the modeling framework applied to the database.

OVERALL MEASURES OF CANDIDATE IDEOLOGY

Beginning in the late 1970s, political scientists began combining techniques from econometrics and psychometrics to study the preferences of political elites (Aldrich and McKelvey 1977; Poole and Rosenthal 1985). This pioneering work found that low-dimensional mapping is highly predictive of congressional roll call voting. With the exception of a few periods in American history, a single dimension explains the lion's share of congressional voting outcomes (Poole and Rosenthal 1997). Ideal point estimation methods have since been used to measure the preferences of political actors serving in institutions other than Congress, including the courts (Epstein et al. 2007; Martin and Quinn 2002) and state legislatures (Shor and McCarty 2011).

The various applications have revealed elite preferences to be low-dimensional across a wide range of political institutions and demonstrated that positions along a liberal-conservative dimension are informative signals about policy preferences. However, relying on voting records to measure preferences precludes generating ideal points for nonincumbent candidates and most nonlegislative office-holders.

A particular challenge has been in comparing ideal points of actors across voting institutions (Bailey 2007; Shor and McCarty 2011). In recent years, political scientists have developed methods to measure preferences from various other sources of data, including candidate surveys (Ansolabehere, Snyder, and Stewart 2001; Burden 2004), campaign contributions (McCarty and Poole 1998), political text (Laver, Benoit, and Garry 2003; Monroe and Maeda 2004; Monroe, Colaresi, and Quinn 2008; Slapin and Proksch 2008), co-sponsorship networks (Peress 2013), voter evaluations (Hare et al. 2014), and Twitter follower networks (Barberá 2015).

The model used here to generate scores for candidates overcomes this problem by scaling campaign contributions using the common-

3. See note 2. The supplemental legislative database is hosted on Harvard Dataverse at http://dx.doi.org/10.7910/DVN/BO7WOW (accessed May 31, 2016).

space DIME methodology (Bonica 2014). The key advantages of this approach are its inclusiveness and scalability. The vast interconnected flows of campaign dollars tie American politics together. The resulting data make it possible to track a broad range of candidates, including non-incumbent candidates who have not previously held elected office, and to reach much further down the ballot. The data also provide estimates of how liberal or conservative individual donors are and place them in a common space with other candidates and organizations spanning local, state, and federal politics.

Campaign Finance Measures

Ideal point estimates for donors and candidates are recovered from campaign finance data using the common-space DIME methodology (Bonica 2014). I refer to Bonica (2014) for a complete treatment of the methodology. Here I provide a general overview of the measurement strategy and validation.

The measurement strategy is relatively straightforward. It relies on donors' collective assessments of candidates as revealed by their contribution patterns. The core assumption is that donors prefer candidates who share their policy views to those who do not. As a result, contributors are assumed—at least in part—to distribute funds in accordance with their evaluations of candidate ideology. As a result, by researching and seeking out candidates who share their policy preferences, donors provide information about the preferences of candidates.

Bonica (2014) offers three main pieces of evidence to validate the measures. First, the DIME scores are strongly correlated with vote-based measures of ideology such as DW-NOMINATE scores, providing strong evidence of their external validity. Second, there is a strong correspondence between contributor and recipient scores for candidates who have both fund-raised and made donations to other candidates, indicating that independently estimated sets of ideal points reveal similar information about an individual's ideology. For the 1,638 federal candidates who ran in the 2014 congressional elections and have scores as both donors and recipients, the correlations between contributor and recipient ideal points are $\rho = 0.97$ overall, $\rho = 0.92$ among Democrats, and $\rho = 0.94$ among Republicans. Third, the scores for individual donors and recipients are robust to controlling for candidate characteristics related to theories of strategic giving, such as incumbency status and electoral competitiveness.

An important claim made here is that the fund-raising activities of non-incumbents are predictive of how they will behave if elected to office. One way to assess the non-incumbent estimates is to compare scores recovered for successful challenger and open-seat candidates with their future scores as incumbents. The correlations between non-incumbent and incumbent CFscores is $r = 0.96$ overall, $r = 0.93$ for Republicans, and $r = 0.88$ for Democrats. This is consistent across candidates for state and federal office (Bonica 2014).

In order for the model to estimate a score for a candidate, the candidate must have received contributions from at least two distinct donors who also gave to at least one other candidate. This covers the vast majority of candidates running for state and federal offices. The model also assigns scores to all donors who contributed to at least two candidates. The donor scores are estimated independently of the recipient scores and exclude any contributions made to one's own campaign.

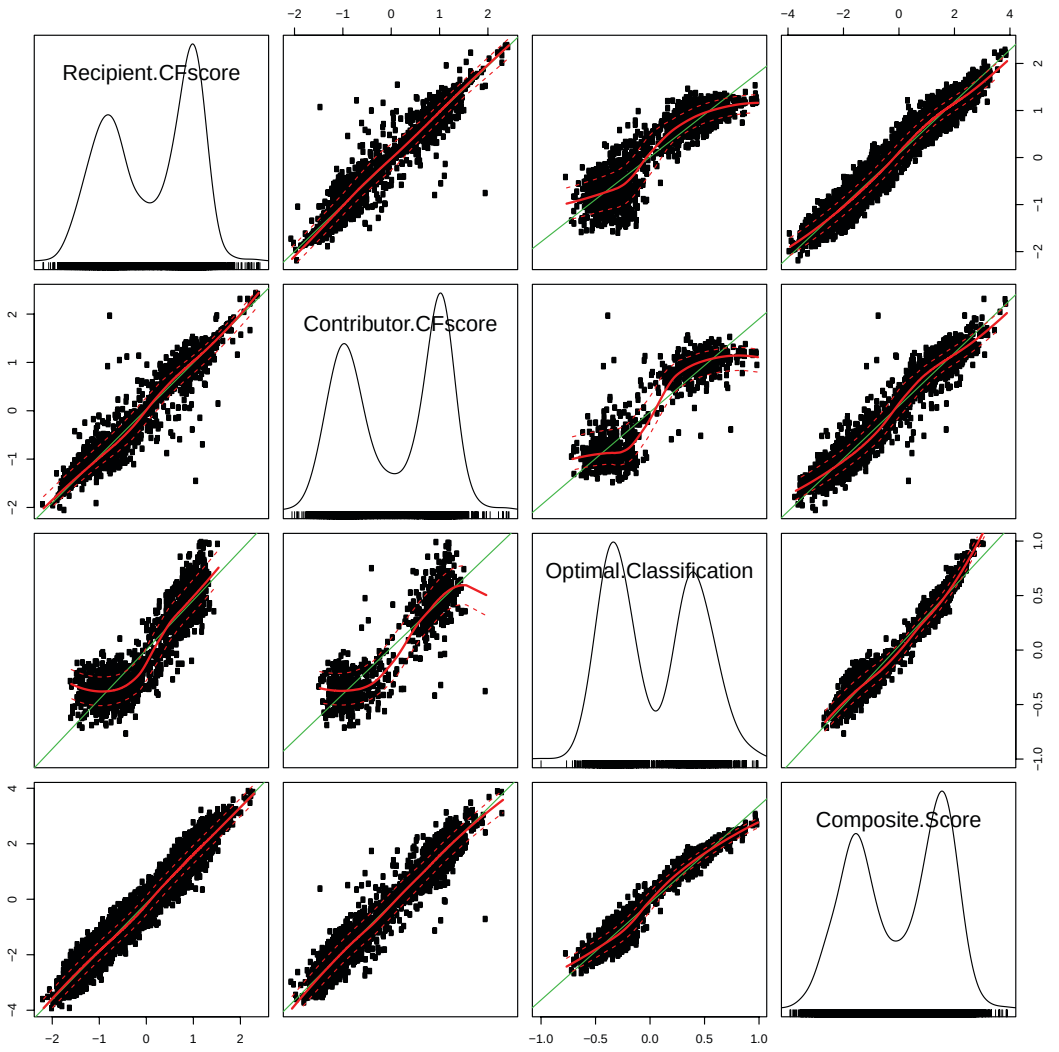
Roll Call Measures

Roll call measures are estimated for candidates who have served in Congress using optimal classification (OC) (Poole 2000). OC is a nonparametric unfolding procedure built directly on the geometry of spatial voting. The scores are from the first dimension of a two-dimensional joint scaling of the House and the Senate based on votes cast during the 108th to 113th Congresses. The roll call–based measures are nearly identical to the common-space DW-NOMINATE scores.

Combining Information Across Measures

Measures derived from distinct data sources may differ in the extent to which they condition on certain areas of politics and types of concerns. For example, Congress rarely votes on socially charged issues such as abortion and same-sex marriage. Yet such issues often fea-

Figure 2. Comparison of Scores for Candidate Ideology Generated from Different Data Sources



Source: Author's calculations.

ture prominently in campaign rhetoric and are a frequent subject of ballot initiatives. PACs and ballot committees that focus on these issues consistently draw large numbers of donors. This suggests value in combining information across measures.

Given the availability of multiple measures of candidate ideology, I average information across different sets of scores. I utilize a multiple over-imputation framework designed to handle multiple continuous variables with measurement error and missing data (Blackwell, Honaker, and King 2010). After imputing five sets of scores, I run a separate principle component analysis (PCA) on each data set. The overall scores are calculated by averaging over candidate scores from the first dimension recovered in each of the runs.

Figure 2 provides a comparison of scores generated from different data sources and the averaged scores. Given that the sets of scores are highly correlated, the first PCA dimension explains most of the variance. The averaged scores correlate with the recipient scores at $\rho = 0.98$, the contributor scores at $\rho = 0.96$, and the roll call scores at $\rho = 0.98$.

A MODEL TO MEASURE CANDIDATE PRIORITIES AND POSITIONS ACROSS ISSUES

Scoring candidates along a single dimension provides highly informative summaries of their policy preferences. Many voters and donors might also be interested in seeing how the preferences and expressed priorities of candidates vary by issue. The following sections outline a three-stage modeling strategy for measuring preferences and expressed priorities across issue dimensions that combines topic modeling, ideal point estimation, and machine learning methods. The first stage applies a topic model to the database of political text. The second stage estimates issue-specific ideal points for legislators based on past voting records using the estimated topic weights to identify the dimensionality of roll calls. The third stage trains a support-vector machine to predict issue scores for a wider set of candidates by conditioning on shared sources of data.

A Topic Model for Political Text

Topic models in their various forms have been extensively used in the study of politics (Grimmer 2010; Grimmer and Stewart 2013; Lauderdale and Clark 2014; Roberts et al. 2014). Political text is particularly well suited to the task of categorizing documents by issue area, whether it be bills, press releases, public speeches, or debates. Topic models offer a computational approach to automating the process of organizing large corpuses of documents into a set of issue categories. This is accomplished by breaking down each document into a set of words or phrases (n-grams) that then can be analyzed as text-as-data. The relative word frequencies found in each document contain information about which documents are most closely associated with which topics. In cases where the set of topics is reasonably well understood prior to the analysis, as is the case here, supervised methods can be used. These methods typically rely on a sample of human-coded documents to train a model that can then be used to infer topics for other documents.

The type of topic model used here is a partially labeled dirichlet allocation (PLDA) model (Ramage, Manning, and Dumais 2011). The PLDA model is a partially supervised topic model designed for use with corpuses where topic labels are assigned to documents in an unstructured or incomplete manner. An important feature of the model is that it allows for documents that address multiple or overlapping issue areas to be tagged with more than one topic. In addition to the specified issue categories, the model allows for a latent category that acts as a catchall or background category.

Issue Labels

The model makes use of issue labels assigned by the Congressional Research Service as a starting point in training the PLDA model. For each bill introduced, the CRS assigns one or more labels from a wide range of potential categories. Although the CRS labels have the advantage of being specific to the task at hand, they are neither well structured nor assigned based on a systematic coding scheme. The raw data include a total of 4,386 issue codes, and it is not uncommon for coders to tag a bill with a dozen or more labels. Many of these issue codes are overly idiosyncratic (for example, "dyslexia" and "grapes"), closely related or overlapping ("oil and gas," "oil-well drilling," "natural gas," "gasoline," and "oil shales"), or subcategorizations. To streamline the issue labels, a secondary layer of normalization is applied on top of the CRS issue codes. This is done by mapping issue labels onto a more general set of categories. CRS issue labels that overlap two larger categories are tagged accordingly (for example, "minority employment" ⇒ "civil rights" and "jobs and the economy"). CRS issue labels that are either too idiosyncratic (for example, "noise pollution") or too ambiguous (for example, "competition") to cleanly map onto a category are removed. All other documents, including those scraped from social media feeds and candidate websites, are used only during the inference stage.

Constructing the Training Set

The training set consists of all documents that can be linked to legislation with CRS issue tags. Since the CRS issue tags are derived from the content of the legislation, bills are espe-

cially important during the training stage. Documents that contain floor speeches made in relation to a specific bill, usually as part of the floor debate, are also included as part of the training set. Such inclusion assumes that the CRS categories assigned to a bill also apply to its floor debate. As such, topic loadings for a bill can reflect both its official language and the floor speeches of members during debate. This is intended as a way to capture how legislators (both supporters and opponents) speak about a bill and better grasp the types of concerns raised during debate. For example, the official language of a health care bill might mostly speak to policy related to health care, but often a single paragraph or specific provision amounting to a small fraction of the bill's language (for example, a provision relating to abortion or reproductive rights) is seized on and becomes the focus of the floor debate. The coding scheme should take this into account by giving more weight to the types of issues that legislators emphasize when speaking about the bill.

Linking Documents to Bills and Legislators

Floor speeches transcribed in the *Congressional Record* are organized into documents based on the identity of the speaker and, if applicable, related legislation. A customized parser was used to extract the speaker's identity, filter on the relevant body of text, and link floor speeches to bill numbers. In order for a document to be linked to a bill, the bill number must be included somewhere in the heading or the speaker must directly reference the name or number of the legislation in the text. Not all floor speeches are related to specific legislation. Legislators are routinely given the opportunity to make commemorations or generally address an issue of their choosing. These speeches often are used as position-taking exercises and are thus informative signals about the legislator's expressed priorities.

PLDA Model and Results

The PLDA model was fit using the Stanford Topic Model Toolkit (Ramage et al. 2009). Terms were organized as both unigrams and bigrams.[4] In addition to the typical list of stopwords included in the Natural Language Took Kit (NLTK) Python package, several terms specific to congressional proceedings and legislation were removed from the text. Stemming was performed using the WordNet lemmatizer, again provided by the NTLK Python package. Rare terms found in fewer than one hundred documents were filtered out. Documents that did not meet the minimum threshold of ten terms were excluded. The model was iterated five thousand times to ensure convergence.

To give a sense of which words are associated with which topics, table 1 reports the top eight words identified by the model as being most closely associated with each topic.

In addition to estimating topic loadings for bills, it is possible to construct measures of the expressed issue priorities of candidates by combining the set of documents linked to an individual, including sponsored legislation. As a way to validate the expressed priorities of legislators, Justin Grimmer (2010) argues that leaders of congressional committees should allocate more attention to the issue topics under their jurisdiction. Figure 3 replicates an analysis found in Grimmer and Stewart (2013) that compares the average attention paid to each topic by Senate committee leaders to the average attention allocated by the rest of the Senate. Note that the analysis here differs in that it compares all members of House and Senate committees with jurisdiction over an issue, not just committee chairs and ranking members. The figure reveals that for every included category, the topic model results indicate that committee members devote significantly more attention to related issues.

Issue-Specific Optimal Classification

In this section, I introduce an issue-specific optimal classification scaling model. The OC scaling model is an attractive option for this application because of its computational efficiency, robustness to missing values, and ability to jointly scale members of the House and Senate in a common-space by using those who

4. A unigram is a single word in a document (for example, "taxes"), and a bigram is the combination of two consecutive words ("cut taxes").

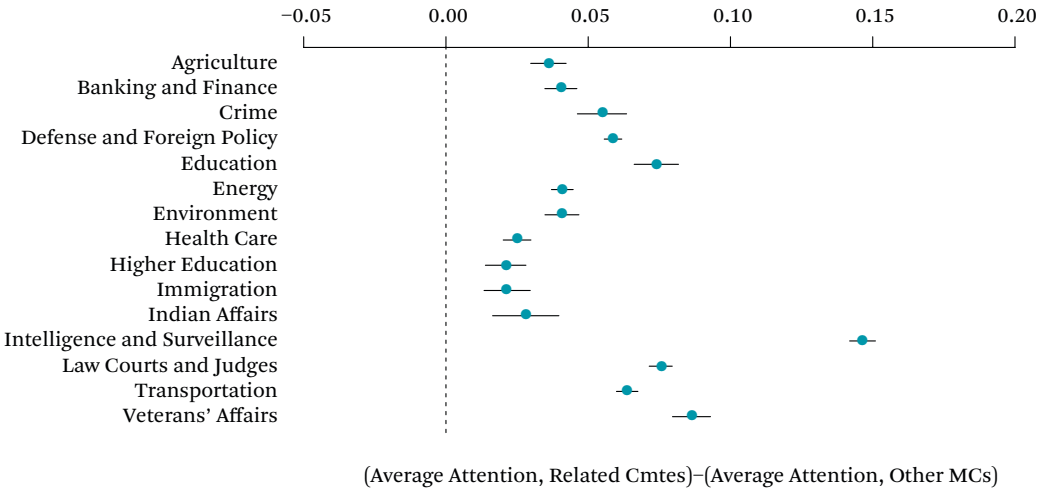
Table 1. Top Terms by Issue Category

	term 1	term 2	term 3	term 4	term 5	term 6	term 7	term 8
Latent	work	make	congress	million	important	country	nation	going
Federal agencies and regulation	commission	sec	office	activity	requirement	director	government	development
Economy	tax	code	credit	revenue	taxable	respect	qualified	benefit
Health care	care	drug	medical	medicare	coverage	disease	patient	insurance
Education	school	education	grant	student	educational	child	local	eligible
Defense and foreign policy	country	government	international	foreign	war	world	force	right
Banking and finance	financial	loan	insurance	housing	credit	mortgage	business	company
Law courts and judges	court	action	person	product	violation	claim	civil	employee
Energy	energy	fuel	oil	gas	vehicle	renewable	facility	production
Procedural	fiscal	budget	appropriation	available	provided	authority	office	congress
Parks and recreation	land	area	management	forest	water	river	project	park
Environment	water	administrator	environmental	species	protection	system	control	project
Veterans' affairs	veteran	defense	military	force	armed	affair	operation	code
Crime	enforcement	child	crime	criminal	attorney	justice	general	grant
Agriculture	food	agricultural	agriculture	farm	producer	crop	payment	assistance
Transportation	transportation	project	safety	system	vehicle	highway	air	funding
Immigration	alien	immigration	border	homeland	status	employer	visa	nationality
Intelligence and surveillance	intelligence	general	internet	person	surveillance	electronic	foreign	privacy
Higher education	education	student	institution	college	higher	science	university	loan
Civil rights	election	right	candidate	vote	voting	voter	political	civil
Emergency	emergency	line	disaster	page	flood	SA	proposed	hurricane
Indian affairs	indian	tribe	native	tribal	land	water	agreement	hawaiian
Women's issues	woman	violence	sexual	assault	domestic	victim	child	prevention
Abortion and social conservatism	right	abortion	religious	cell	woman	human	stem	research
Guns	firearm	person	gun	general	attorney	model	code	ammunition

Source: Author's calculations.

Note: Topics are listed in descending order based on their relative weights. Bigrams are excluded from the table to enhance readability.

Figure 3. Average Attention to Topics by Senate Committee Leaders Compared to Average Attention by Other Senate Members



Source: Author's calculations.

served in both chambers as bridge observations (Poole 2000). The model follows recent methodological developments in multidimensional ideal point estimation (Clark and Lauderdale 2012; Gerrish and Blei 2012; Lauderdale and Clark 2014). Borrowing from the issue-adjusted ideal point model developed by Sean Gerrish and David Blei (2012), the dimensionality of roll calls is identified using a topic model trained on issue tags provided by the CRS. The issue-specific OC model differs in its approach to mapping the results from the topic model onto the dimensionality of roll calls. Gerrish and Blei incorporate a vector of issue adjustment parameters that in effect serve as dimension-specific utility shocks. The issue-specific OC model instead utilizes the basic geometry of spatial voting through the parameterization of the normal vectors. This approach distinguishes the issue-specific OC model from the approach taken by Tom Clark and Benjamin Lauderdale (2012), who similarly extend OC to generate issue-varying ideal points for U.S. Supreme Court justices by kernel-weighting errors based on substantive similarity. The approach is actually most similar to related work by Lauderdale and Clark (2014) that combines latent dirichlet allocation with an item response theory model.

In the standard OC model, the dimensionality of bill j is determined by a heuristic cutting plane algorithm that searches the parameter space for the normal vector N_j and corresponding cutting line c_j, which minimize classification errors. The issue-specific OC model instead differs by calculating the normal vectors based on the parameters recovered from the PLDA model. Given a k-length vector λ_j of topic weights for roll call j, the normal vector is calculated as $N_{jk} = \lambda_{ik}/||\lambda_j||$. Legislator ideal points are then projected onto the projection line: $w_i = \theta_i' N_j$. Given the mapping onto w, finding the optimal cutting point c_j is identical to a one-dimensional classification problem. Given the estimated roll call parameters, issue-specific ideal points can be recovered dimension by dimension. Holding parameters for $\theta_{i\text{-}k}$ constant, classification errors are minimized by finding the optimal value of θ_{ik} given c_j and the projected values $w_{ij} = \theta'_{i\text{-}k} N_{j\text{-}k} + \theta'_{ik} N_{jk}$. As an identification assumption, $\theta_{k=1}$ is fixed at its starting value.

A further extension to the OC model is the incorporation of kernel methods to capture the relative importance of bills to legislators. A member's sponsorship of a bill or contribution to the floor debate suggests that the bill has greater significance to her than other bills on which she is silent. The inputs to the kernel-weighting function are status as a sponsor or

Table 2. Roll Call Classification, 108th to 113th Congresses

	Correct Classification	Aggregate Proportional Reduction in Error	Errors	Weighted CC	Weighted APRE	Weighted Errors
One-dimensional OC	0.936	0.825	154569	0.938	0.818	179,598
Issue-specific OC	0.940	0.835	145430	0.943	0.832	166,126

Source: Author's calculations.

co-sponsor and the total word count devoted to the legislation. The weight matrix is constructed as follows:

$$\omega_{ij} = 1 + \gamma_1 sponsor_{ij} + \gamma_2 cosponsor_{ij} + \gamma_3 log(wordcount_{ij}) \quad (1)$$

The γ parameters are calibrated using a cross-validation scheme. Given a set of parameter values, the model is subjected to repeated runs with a fraction of observed vote choices held out. After the model run has converged, the total errors are calculated for a held-out sample based on the recovered estimates. Values are typically somewhere in the region of $\gamma_1 = 5$, $\gamma_2 = 2$, and $\gamma_3 = 1$.

Starting values are estimated separately for each dimension using a one-dimensional OC scaling with issue-weighted errors. Given an issue dimension k, errors on each roll call are weighted by the proportion of the related text associated with the issue. A classification error on a roll call where $\lambda_{jk} = 0.5$ is weighted 50 times that of an error on a roll call where $\lambda_{jk} = 0.01$. After dropping roll calls where $\lambda_{jk} < 0.01$, the model is run to convergence.

Table 2 reports the classification statistics for the issue-specific OC model. The issue-specific model increases correct classification (CC) over the one-dimensional model, but only marginally. Congressional voting has become so unidimensional that only a small fraction of voting behavior is left unexplained by a one-dimensional model. The issue-specific model explains a nontrivial percentage of the remaining error. However, this is slightly less than the reduction in error associated with adding a second dimension to the standard OC model.

The marginal increase in fit occurs largely by design and is explained by constraints built into the issue-specific OC model. Classifying roll call votes in multiple dimensions can be highly sensitive to slight changes to the position or angle of the cutting line. The cutting-plane search is free to precisely position the cutting line by simultaneously manipulating the normal vector and cutting line. Hard-coding the dimensionality of bills based on the topic loading constrains normal vectors and limits the search to c_j. These effects are further compounded by a modeling assumption, made largely in the interest of reducing computational costs, that constrains the values for $N_{jk} \geq 0$, corresponding to the vector of topic loadings for each bill from which they are calculated. This means that bill proposals must move policy on all relevant dimensions in the same direction (that is, toward the ideological left or right). For example, the model does not allow for a bill to move economic policy to the right but immigration policy to the left.[5]

To assess the extent to which holding the normal vectors fixed explains the marginal reduction in error, I ran the cutting-plane search algorithm with the legislator ideal points set at values recovered from the issue-specific model. Relaxing the constraint on the normal vectors resulted in an appreciable reduction in error: correct classification was boosted to 96.4 percent.

5. For a two-dimensional model, this would constrain the normal vector to the upper-right quadrant. This constraint could be relaxed by the addition of a sign vector, which would allow values in the normal vector to take on negative or positive values. For an in-depth discussion of this issue, see Lauderdale and Clark (2014).

Figures 4 and 5 display a series of parallel plots that compare ideal points from standard OC and issue-specific OC for members of the 108th and 113th Congresses. The points on the top are ideal points from a standard one-dimensional OC scaling. The points on the bottom are the corresponding issue-specific ideal points. The line segments trace changes in ideal points between models.

In contrast to the near-perfect separation between the parties in Congress in the one-dimensional OC model during the period under analysis, the issue-specific model does show increased partisan overlap for most issues. The issues for which this overlap is most apparent are abortion and social conservatism, agriculture, guns, immigration, Indian affairs, intelligence and surveillance, and women's issues.

Where the issue-specific model excels is in identifying key legislators who broke ranks on one or more issue dimensions. For example, the sole legislator to cross over on defense and foreign policy was Representative Jim Leach (R-IA), who was known for his progressive views on foreign affairs. Of the legislators to cross over on abortion and social conservatism, pro-life advocates Senator Ben Nelson (D-NE) and Representatives John Breaux (D-LA) and Bobby Bright (D-ALA) were the three most conservative Democrats, and pro-choice advocates Representatives Sherry Boehlert (R-NY) and Rob Simmons (R-CT) and Senator Olympia Snowe (R-ME) were the three most liberal Republicans. Although few legislators break with their party on any given issue dimension, the ones who do are often noteworthy and highly visible players on the issue who stand out as examples of either cross- pressured bipartisans or uncompromising hard-liners. Often the largest differences are associated with legislators who are active on an issue. On immigration, for example, the legislators whose issue-specific ideal points shifted them the most from their overall score were Senators Chuck Hagel (R-NE) and Jeff Flake (R-AZ), both of whom had co-sponsored bipartisan immigration reform bills at different points in time.

The issue-specific ideal points on the intelligence and surveillance dimension are especially revealing. Four of the most conservative Republicans—Representatives Ron Paul (R-TX) and Justin Amash (R-MI) and Senators Rand Paul (R-KY) and Mike Lee (R-UT)—voted so consistently against their party that they flipped to have some of the most liberal ideal points on the issue. This fits with the libertarian leanings of these candidates as well as their public and vocal opposition to government surveillance.

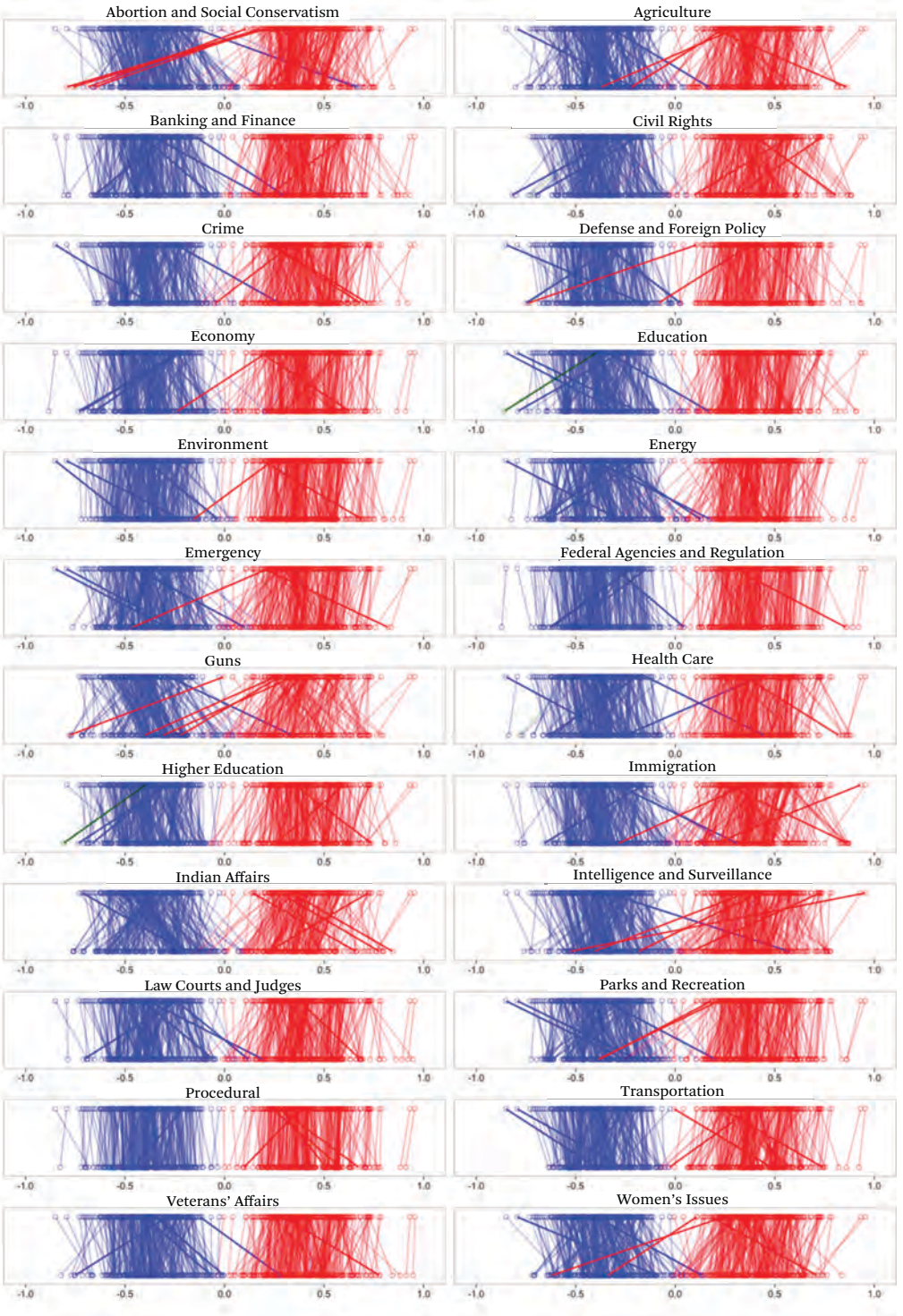
Changes in patterns of partisan overlap from the 108th Congress to the 113th can also be revealing. In the 108th, the issue-specific ideal points for a handful of Republicans, including Senators Lincoln Chafee (R-RI), George Voinovich (R-OH), Mike Dewine (R-OH), and John Warner (R-VA), accurately place them well to left of center on guns. By the 113th Congress, the only remaining Republican crossover was Senator Mark Kirk (R-IL), whereas the number of Democrats breaking with their party over gun policy had grown to include Senators Byron Dorgan (D-ND), Max Baucus (D-MT), and Mark Pryor (D-AR), Representatives Henry Cuellar (D-TX) and Kurt Schrader (D-TX), and several others.

Support Vector Regression

The final stage in the model integrates campaign contributions. The objective is to produce issue-specific ideal points for the vast majority of candidates who lack voting records. Ideally, the model would seamlessly integrate voting and contribution records to estimate issue-specific ideal points for the entire population of candidates simultaneously. Unfortunately, such an approach is out of reach. I instead rely on supervised machine learning methods.

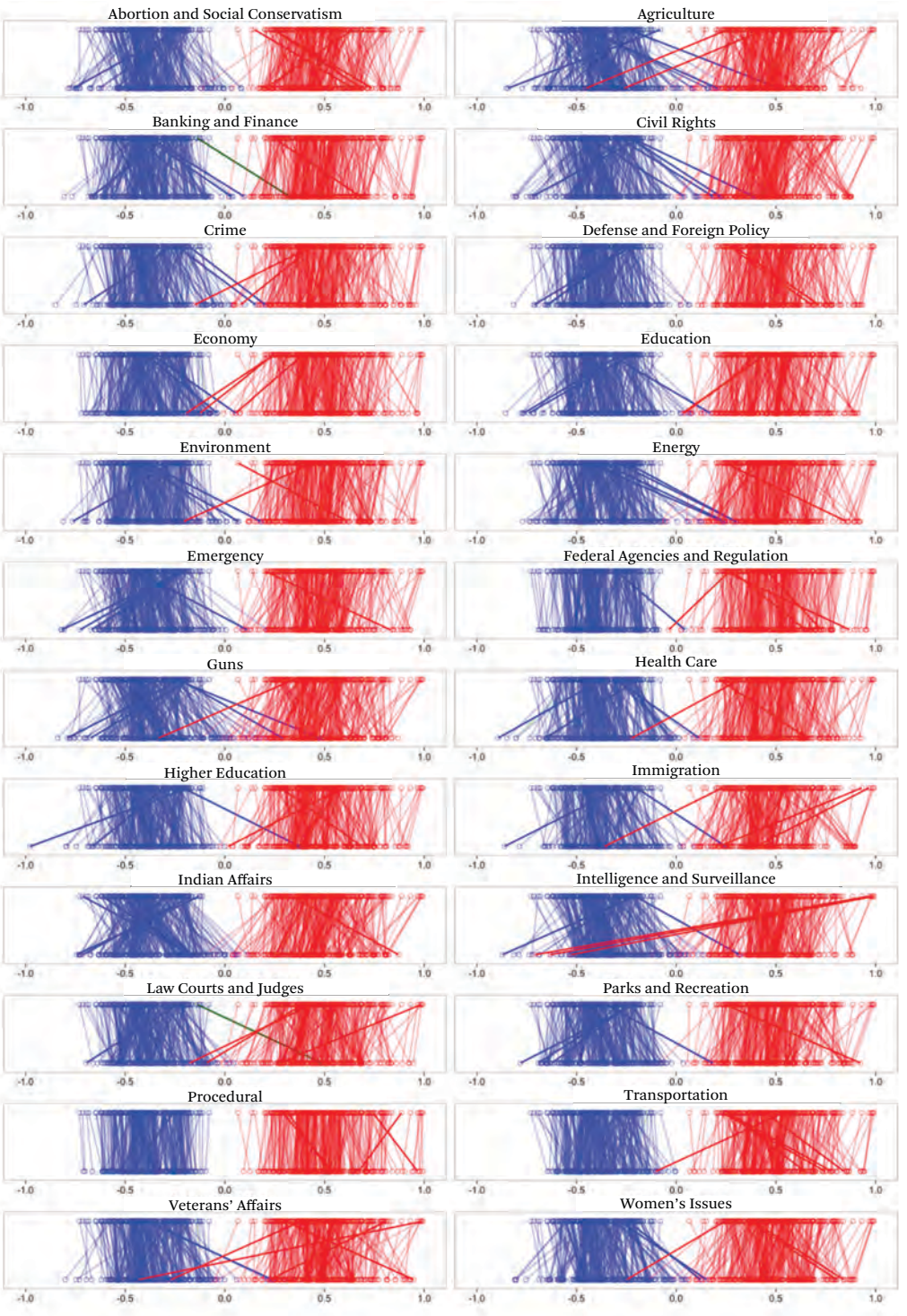
The structure of campaign contributions has many similarities to text-as-data. The contingency matrix of donors and recipients is functionally similar to a document-term matrix, only with shorter documents and more highly informative words. As such, translating models originally designed for political text for use with campaign contributions is relatively straightforward. Although several classes of the models typically applied to textual analysis could be used here, I focus on support vector

Figure 4. Legislator Ideal Points from Classical OC and Issue-Specific OC (108th Congress)



Source: Author's calculations.

Figure 5. Legislator Ideal Points from Classical OC and Issue-Specific OC (113th Congress)



Source: Author's calculations.

regression (SVR) (Drucker et al. 1997; Smola and Schölkopf 2004).[6]

The SVR approach has several advantages. What this approach lacks in elegance is made up for by its extensibility and generalizability. In theory, there is no reason why other types of data could not be included alongside the contribution data as additional features. The model presented here combines contribution records with word frequencies from the document-term matrix for use as the predictor matrix. Although contribution data perform much better than text-as-data when modeled separately, including both data sources boosts cross-validated R-squared by one to two percentage points for most issue dimensions over the contribution matrix alone.

A downside to this approach is that it takes the roll call estimates as known quantities despite the presence of measurement error. Assessing model fit thus becomes somewhat problematic, as the extent to which cross-validation error actually reflects attenuation bias is unclear. Although not ideal, I proceed by treating the roll call estimates as though they are measured without error.[7]

The SVR model is fit using a linear kernel and recursive feature selection. To help the model handle the sparsity in the contribution matrix, I construct an n-by-k matrix that summarizes the percentage of funds a candidate raised from donors within different ideological deciles. This is done by calculating contributor coordinates from the weighted average of contributions made to the set of candidates with roll call estimates for the target issue scale and then binning the coordinates into deciles. The candidate decile shares are then calculated as the proportion of total funds raised from contributors located within each decile. When calculating the contributor coordinates, contributions made to candidates in the test set are excluded so as not to contaminate the cross-validation results. This simple trick helps to augment feature selection. As is typical with support vector machines, the modeling parameters require careful calibration. The ε and cost parameters are tuned separately for each issue dimension.

Table 3 reports fit statistics for fifteen issue dimensions for members of the 113th Congress. The cross-validated correlation coefficients are above 0.95 for every issue. The within-party correlations are generally above 0.60, indicating that the model can explain variation in the scores of co-partisans.

The SVR model demonstrates the viability of training a machine learning model to learn about candidate issue positions from contribution records and text. The SVR as presented performs quite well for its intended purpose but leaves room for improvement. In most other contexts, the cross-validation results would be a resounding success. In this context, however, the historically high level of issue constraint causes the model to suffer from a "curse of unidimensionality." Candidate positions across issues are so strongly correlated that it becomes a challenge to train a model that is nuanced enough to pick up on variation revealed by the issue-specific OC model, which is often driven by a small fraction of legislators who deviate from their positions on one or two given issue dimensions. Moving forward, ensemble methods that build on the SVR model—k-nearest neighbors methods in particular—show promise for improving predictive performance. It also remains to be seen whether similarly high levels of issue constraint are present in the state legislatures.

A DATA-DRIVEN VOTER GUIDE

In this section, I provide an overview of the design and development of Crowdpac's data-driven voter guide. The initial motivation was to build a tool capable of providing users with objective information on the policy preferences and expressed priorities of a comprehen-

6. For a complete treatment of the application of supervised machine learning methods to infer roll call ideology from campaign contributions, see Bonica (2016).

7. An alternative approach worth exploring would be to train a binary classifier on individual vote choices on bills and then scale the predicted vote choices for candidates using the roll call parameters recovered from OC. Although this approach would sidestep issues with measurement error, it would probably present additional challenges.

Table 3. Fit Measures from Cross-Validation on Fifteen Issue Dimensions, 113th Congress

	All		Democrats		Republicans	
	Pearson R	RMSE	Pearson R	RMSE	Pearson R	RMSE
Latent	0.979	0.074	0.819	0.06	0.775	0.085
Defense and foreign policy	0.973	0.085	0.732	0.073	0.740	0.094
Banking and finance	0.973	0.081	0.700	0.076	0.751	0.085
Energy	0.971	0.084	0.711	0.074	0.722	0.092
Health care	0.970	0.091	0.760	0.078	0.741	0.100
Economy	0.968	0.089	0.687	0.081	0.721	0.095
Environment	0.966	0.094	0.680	0.089	0.732	0.095
Women's issues	0.964	0.094	0.619	0.083	0.687	0.101
Education	0.963	0.099	0.679	0.087	0.678	0.108
Abortion and social conservatism	0.961	0.102	0.637	0.096	0.691	0.107
Higher education	0.958	0.104	0.698	0.090	0.697	0.115
Immigration	0.957	0.110	0.643	0.103	0.699	0.115
Fair elections	0.956	0.117	0.626	0.099	0.659	0.139
Intelligence and surveillance	0.952	0.108	0.705	0.088	0.543	0.126
Labor	0.952	0.122	0.603	0.123	0.663	0.123
Guns	0.951	0.116	0.680	0.089	0.560	0.137

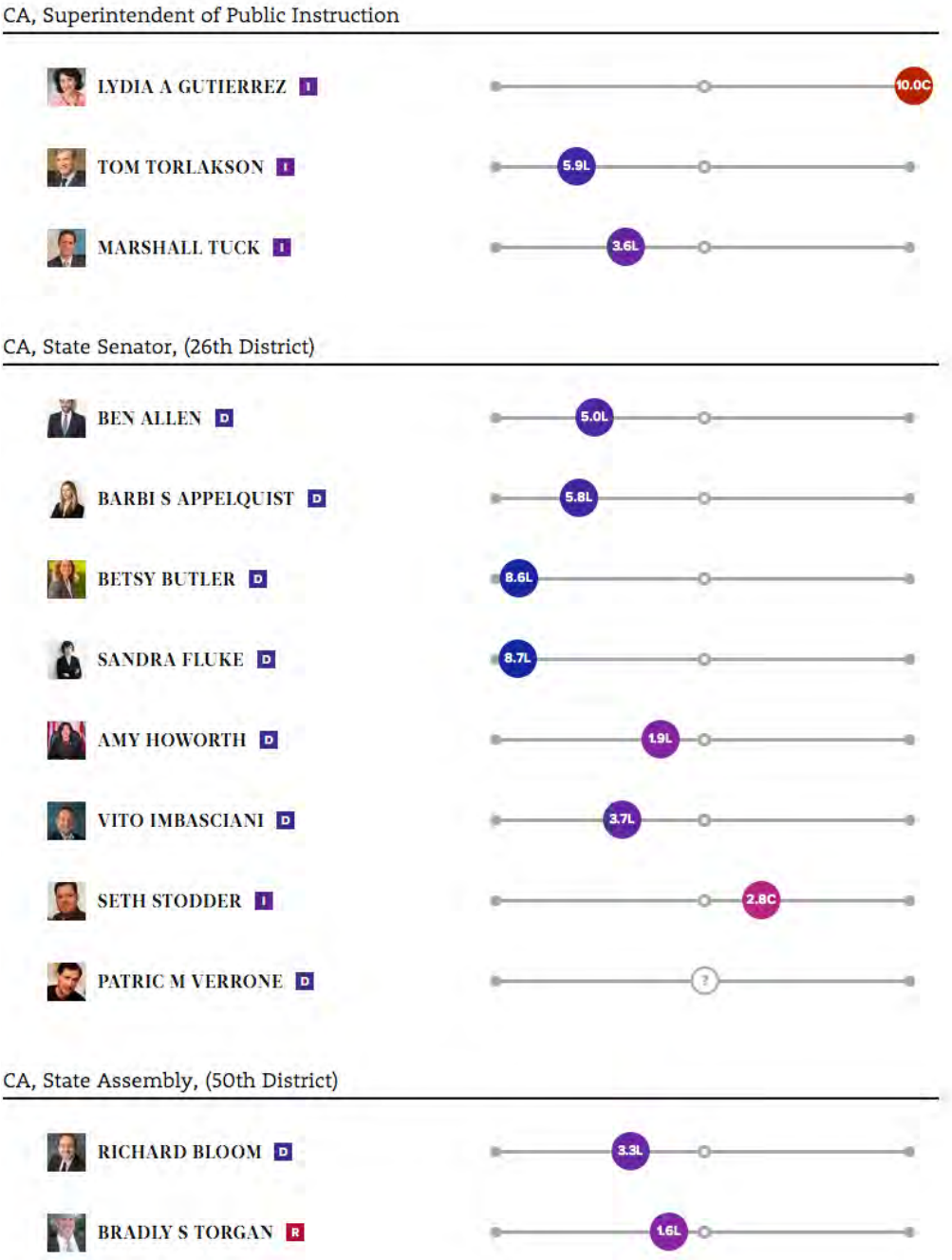
Source: Author's calculations.

sive set of candidates. While Crowdpac's voter guide provides an illustrative example of such a tool, the data and techniques employed here are quite flexible and could be extended to produce different types of voter guides.

Figure 6 displays a screenshot that captures three of the eleven primary races appearing on the sample ballot from the Crowdpac voter guide for the 2014 California primary elections. Each candidate in the contest is assigned an overall ideological score ranging from 10L for candidates on the far left to 10C for candidates on the far right. The scores are rescaled to enhance interpretability for users. The rescaling function is identified using the historical averages for the parties in Congress over the past two decades. First, the historical party means are calculated by aggregating over the ideal points of the members from each party serving in each Congress from 1992 to 2012. The scores are then rescaled such that the midpoint between the party means is set to 0 and the historical party means are positioned at 5L and 5C. The scores are windsorized at 10L and 10C. The user interface was designed to scope with respect to the level of detail displayed about a candidate.

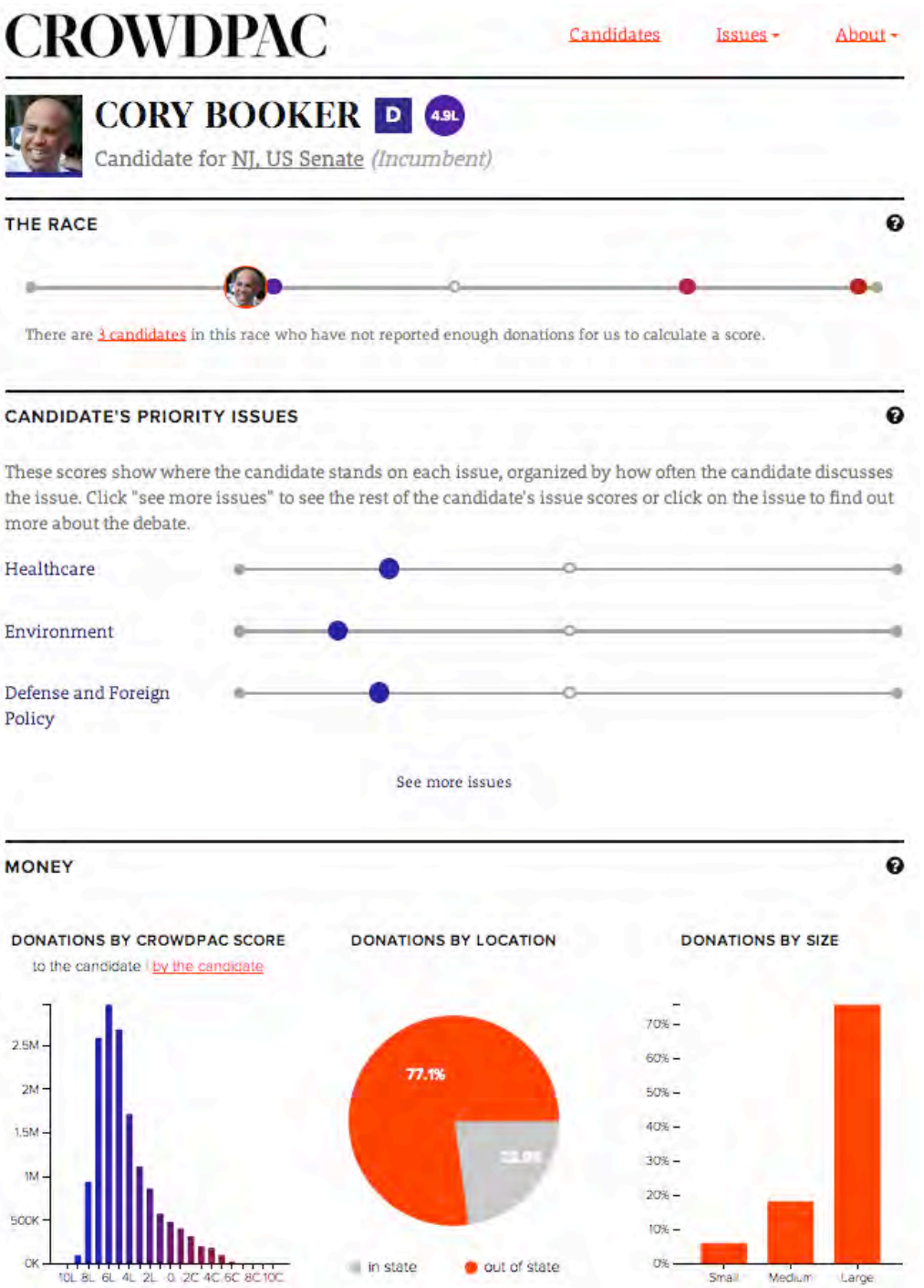
The unidimensional scores for candidates are top-level summaries that serve as jumping-off points for exploring more detailed data on them. More inquiring users are given the option to further explore the data by clicking through to the "data details" pages provided for each candidate. Figure 7 displays a screenshot for the data details page for Cory Booker (D-NJ) as an example. The module on the top displays the candidate's ideal point with respect to his opponents in the upcoming election. While the voter guide makes extensive use of scores along a liberal-to-conservative dimension, issue-specific ideal points are also available for a large percentage of candidates who meet the minimum data requirement of raising funds from at least 100 distinct donors who have also donated to one or more other candidates. The bottom modules summarize the candidate's fund-raising activity by showing the distribution of ideal points of donors to his campaign along with other general fund-raising statistics. For candidates who have made personal donations to other candidates and committees, there is a toggle option that shows the ideological distribution of the recipients weighted by amount. Other modules not shown include (1) a visualization of the candidate's fund-raising network accompa-

Figure 6. Screenshot of Sample Ballot from Crowdpac Voter Guide to 2014 California Primary Elections



Source: www.crowdpac.com (accessed September 29, 2014).

Figure 7. Screenshot of Data Details Page for Crowdpac's Voter Guide to the Candidate Cory Booker (D-NJ)



Source: www.crowdpac.com (accessed September 29, 2014).

nied by a listing of the candidate's nearest neighbors (that is, the donors who gave to the candidate and also gave to candidates X, Y, Z); (2) a summary of the candidate's text showing his expressed priorities and a word cloud of top terms; (3) a video of the candidate from YouTube or another video sharing service; (4) biographical information, including past political experience and offices held; and (5) for sitting members of Congress, a summary of recent voting behavior and interest group ratings.

CONCLUSIONS

This paper proposes a scalable strategy for collecting and modeling data on U.S. political elites as a means of measuring the positions and priorities for a comprehensive set of candidates. The project hinges on the ability to collect, process, and organize large amounts of data on candidates and other political elites. Many of the needed support structures for data provision have been institutionalized by disclosure regimes. The initial fixed costs associated with building the tools for automating the process of collecting and processing new data as they become available are considerable, but once paid, the database should yield continued benefits with much reduced maintenance costs.

Although more work remains, the model is able to reliably position candidates along a liberal-to-conservative dimension and capture meaningful variation in legislator ideal points across issue dimensions. By training on the set of ideal points recovered from the issue-specific OC model, a support vector regression model is used to infer scores for other candidates based on shared sources of data. This modeling strategy demonstrates the viability of training a model to predict how candidates would vote on an issue if they were in office.

The potential benefits are twofold. First, the model offers a valuable new data resource for social scientists. In addition to compiling and standardizing data on political candidates, legislative behavior, political text, campaign contributions, and election outcomes in an accessible format, considerable effort has gone into automating data collection and merging and disambiguating data drawn from different sources. The result is a unified data resource on American political elites unprecedented in its size, scope, and variety. Moreover, the data architecture is designed to accommodate the addition of new data sources centered on candidates and political organizations—for example, Twitter follower networks (Barberá 2015) or interest group ratings and endorsements—which would then be automatically linked to each of the other included data sources.

Second, the model provides a means of democratizing political disclosure data by making such data more accessible to citizens. The website design has been built around the founding principle that, as with almost any activity, most citizens want to minimize the time and effort they spend on politics while maximizing their effectiveness. If successful, the model could promote political engagement by lowering information costs, reducing uncertainty, and enhancing efficacy.

REFERENCES

Aldrich, John H., and Richard D. McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(1): 111–30.

Alvarez, R. Michael, Ines Levin, Peter Mair, and Alexander H. Trechsel. 2014. "Party Preferences in the Digital Age: The Impact of Voting Advice Applications." *Party Politics* 20(2): 227–36.

Ansolabehere, Stephen, Jr., James M. Snyder, and Charles Stewart III. 2001. "Candidate Positioning in U.S. House Elections." *American Journal of Political Science* 45(1): 136–59.

Bailey, Michael A. 2007. "Comparable Preference Estimates Across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51(3): 433–48.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1): 76–91.

Blackwell, Matthew, James Honaker, and Gary King. 2010. "Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data." Working paper. July 19. Available at: http://polmeth.wustl.edu/files/polmeth/measure.pdf (accessed May 31, 2016).

Bonica, Adam. 2014. "Mapping the Ideological Mar-

ketplace." *American Journal of Political Science* 58(2): 367–87.

———. 2016. "Inferring Roll Call Scores from Campaign Contributions Using Supervised Machine Learning." Working paper. Stanford, Calif.: Stanford University (Mach 12). Available at: SSRN: http://ssrn.com/abstract=2732913 (accessed May 31, 2016).

Burden, Barry C. 2004. "Candidate Positioning in U.S. Congressional Elections." *British Journal of Political Science* 34(2): 211–27.

Clark, Tom S., and Benjamin Lauderdale. 2012. "The Supreme Court's Many Median Justices." *American Political Science Review* 106(4): 847–66.

Drucker, Harris, Chris J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. "Support Vector Regression Machines." *Advances in Neural Information Processing Systems* 9: 155–61.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal, and Chad Westerland. 2007. "The Judicial Common Space." *Journal of Law, Economics, and Organization* 23(2): 303–25.

Gerrish, Sean, and David M. Blei. 2012. "How They Vote: Issue-Adjusted Models of Legislative Behavior." *Advances in Neural Information Processing Systems* 25: 2753–61.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1): 1–35.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis,* 1–31. doi: 10.1093/pan/mps028.

Hare, Christopher, David A. Armstrong, Ryan Bakker, Royce Carroll, and Keith T. Poole. 2014. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3): 759–74.

Issenberg, Sasha. 2012. *The Victory Lab: The Secret Science of Winning Campaigns.* New York: Random House.

Ladner, Andreas, Gabriela Felder, and Jan Fivaz. 2010. "More Than Toys? A First Assessment of Voting Advice Applications in Switzerland." In *Voting Advice Applications in Europe: The State of the Art,* edited by Lorella Cedroni and Diego Garcia. Naples: ScriptaWeb.

Lauderdale, Benjamin, and Tom S. Clark. 2014. "Scaling Politically Meaningful Dimensions Using Texts and Votes." *American Journal of Political Science* 58(3): 754–71.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 92(2): 311–32.

Louwerse, Tom, and Martin Rosema. 2013. "The Design Effects of Voting Advice Applications: Comparing Methods of Calculating Matches." *Acta Politica* (October 18). doi:10.1057/ap.2013.30.

Martin, Andrew D., and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10(2): 134–53.

McCarty, Nolan M., and Keith T. Poole. 1998. "An Empirical Spatial Model of Congressional Campaigns." *Political Analysis* 7(1): 1–30.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4): 372–403.

Monroe, Burt L., and Ko Maeda. 2004. "Talk's Cheap: Text Based Estimation of Rhetorical Ideal-Points." Paper presented to the Twenty-First Annual Summer Meeting of the Society for Political Methodology. Stanford University (July 29–31).

Peress, Michael. 2013. "Estimating Proposal and Status Quo Locations Using Voting and Cosponsorship Data." *Journal of Politics* 75(3): 613–31.

Poole, Keith T. 2000. "Nonparametric Unfolding of Binary Choice Data." *Political Analysis* 8(3): 211–37.

Poole, Keith, Jeffrey Lewis, James Lo, and Royce Carroll. 2011. "Scaling Roll Call Votes with WNOMINATE in R." *Journal of Statistical Software* 42(14): 1–21.

Poole, Keith T., and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2): 357–84.

———. 1997. *Congress: A Political-Economic History of Roll Call Voting.* New York: Oxford University Press.

Ramage, Daniel, Christopher D. Manning, and Susan Dumais. 2011. "Partially Labeled Topic Models for Interpretable Text Mining." In Association for Computing Machinery (ACM), *Proceedings of the 17th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining (San Diego, Calif., August 21–24), 457–65.

Ramage, Daniel, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland. 2009. "Topic Modeling for the Social Sciences." Presented at Neural Information Processing Systems (NIPS) 2009 Workshop on Applications for Topic Models: Text and Beyond. Whistler, Canada (December).

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4): 1064–82.

Rosema, Martin, Joel Anderson, and Stefaan Walgrave. 2014. "The Design, Purpose, and Effects of Voting Advice Applications." *Electoral Studies* (36): 240–43.

Shor, Boris, and Nolan M. McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105(3): 530–51.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3): 705–22.

Smola, Alex J., and Bernhard Schölkopf. 2004. "A Tutorial on Support Vector Regression." *Statistics and Computing* 14(3): 199–222.

Willis, Derek. 2014. "New Voter Guide Follows the Money." *New York Times,* September 1.