Data Science and Political Economy: Application to Financial Regulatory Structure

SHARYN O'HALLORAN, SAMEER MASKEY, GERALDINE McALLISTER, DAVID K. PARK, AND KAIPING CHEN

The development of computational data science techniques in natural language processing and machine learning algorithms to analyze large and complex textual information opens new avenues for studying the interaction between economics and politics. We apply these techniques to analyze the design of financial regulatory structure in the United States since 1950. The analysis focuses on the delegation of discretionary authority to regulatory agencies in promulgating, implementing, and enforcing financial sector laws and overseeing compliance with them. Combining traditional studies with the new machine learning approaches enables us to go beyond the limitations of both methods and offer a more precise interpretation of the determinants of financial regulatory structure.

Keywords: big data, natural language processing, machine learning, political economics, financial regulation, banking and financial services sector

The development of computational techniques to analyze large and complex information, or big data, opens a window to studying the interaction between economics and politics. Natural language processing (NLP) and machine learning (ML) algorithms offer new approaches to examining intricate processes such as government's regulation of markets. For example, traditional observational studies of the design of regulatory structure rely on thousands of hours of well-trained annotators coding laws to extract information on the delegation of decision-making authority to agencies, the administrative procedures that circumscribe this authority, the scope of regulation, the subsequent rules promulgated, and the impact on financial market participants. Using big data methods to analyze this predominantly text-based information reduces the time and expense of data collection and improves the validity and efficiency of estimates. Fast and accurate processing of complex information in real time enables decision-makers to evaluate alternative theories of regulatory structure and, ultimately, to predict which institutional arrangements lead to more efficient markets and under what conditions.

Big data methods undoubtedly equip researchers with tools to study political economy questions that could not be addressed previously. As we have witnessed, however, the term "big data" has been thrust into the zeitgeist in

Sharyn O'Halloran is George Blumenthal Professor of Political Economy and professor of international and public affairs at Columbia University. **Sameer Maskey** is assistant professor at Columbia University. **Geraldine McAllister** is Senate director at Columbia University. **David K. Park** is dean of Strategic Initiatives at Columbia University. **Kaiping Chen** is a doctoral student at Stanford University.

Direct correspondence to: Sharyn O'Halloran at so33@columbia.edu, Department of Political Science, Columbia University, 420 West 118th St., New York, NY 10027; Sameer Maskey at smaskey@cs.columbia.edu, Geraldine McAllister at gam2116@columbia.edu, David K. Park at dkp7@columbia.edu; and Kaiping Chen at kpchen23@stanford.edu.

recent years with no consistent meaning or framework for interpreting results. Indeed, many computational analysts view big data as synonymous with causal inference: correlation supplants the need for explanation. As Rocío Titiunik (2015) explains, however, increasing the number of observations or variables in a data set does not resolve causation.[1]

We have always had “data,” and lots of it. So what is different about big data today? What is new this time around can be summarized along three dimensions: granularity, real time, and textual pattern recognition. With computational advances in the data sciences, researchers can now go beyond keyword searches and use more sophisticated word sequencing to construct measures, thereby reducing error and potential bias (Lewis 2014). Why is this important? Many public policy decisions rely on temporaneous data to predict impact and mitigate potential unintended consequences. Data science techniques thereby facilitate the management and processing of large quantities of information at rapid speeds, the availability of which can lead to better-informed policy.

The purpose of this paper is to illustrate how these new computational data science methods can enhance political economy research. We apply these tools to analyze the design of financial regulatory structure in the United States since 1950. The centerpiece of this work is a large database encoding the text of financial regulation laws. Among other variables, we code the amount of regulatory authority delegated to executive agencies and the procedural constraints associated with the use of that authority. The analysis requires aggregating measures from thousands of pages of text-based data sources with tens of thousands of provisions, containing millions of words. Such a large-scale data project is time-consuming, expensive, and subject to potential measurement error. To mitigate these limitations and demonstrate the robustness of the coding procedures, we employ data science techniques to complement the observational study of financial regulatory structure. The computational analyses conducted: (1) enable sensitivity analysis around manual rules-based coding, (2) identify the magnitude and location of potential error, and (3) allow for benchmarking. The results indicate that, while the manual coding rules perform better than unstructured text alone, the accuracy of the estimates improves significantly when both methods are combined. Thus, our results underscore the complementarities of computational sciences and traditional social sciences (rules-based coding) methods when examining important political economy questions.

The first section of the paper surveys the literature on delegation and agency design, highlighting the role of uncertainty and conflict as key determinants of regulatory architecture. The central hypothesis derived from this literature is that the closer the policy preferences of Congress and the executive, the more discretionary authority is delegated to agencies. To empirically test this hypothesis, the subsequent section details the rules and criteria used to construct the financial regulatory structure database. The statistical analysis reaffirms the political nature of financial market regulation: the closer the policy preferences of Congress and the executive, the more discretionary authority is delegated. To check the robustness of these findings, we recode the financial regulation laws using NLP, which converts the text into machine-readable form. We then apply both a naive and naive Bayes model to compare three coding schemes to predict agency discretion, noting that combined methods perform best. We conclude with a discussion of the implications of incorporating computational methods into text-based coding to improve the validity and robustness of the findings.

DELEGATION, DISCRETION, AND FINANCIAL REGULATORY DESIGN

As a necessary preamble, this section reviews the literature on delegation and agency design. The extensive corpus of work on the delegation of policymaking authority to administrative

[1] William R. Clark and Matt Golder (1995) review additional pitfalls of computational analysis, such as sampling populations, confounding variables, over-identification, and multiple hypothesis testing. See the symposium in the January 2015 issue of *PS*.

agencies can usefully be separated along three lines. First, why does Congress delegate regulatory authority? Second, how does Congress constrain agency decision-making, if at all? And third, given the answers to questions one and two, what drives the amount of substantive discretionary authority delegated by Congress?

The first strand of thought analyzes Congress's motivation to transfer authority to administrative agencies, noting key factors such as workload, political risk, bureaucratic expertise, and interest group politics, to name but a few. The aim of this line of inquiry is to describe, and at times even rationalize, the explosive growth of the federal bureaucracy and the corresponding implications for democratic institutions.[2] A second and related line of reasoning questions the constitutionality of Congress delegating expansive legislative authority to unelected bureaucrats. It contends that such unconstrained authority equates to congressional abrogation of its policymaking responsibilities and thereby fundamentally undermines the U.S. system of separate powers.[3] The counterpoint to these assertions recognizes that while Congress grants administrative functions to professional bureaucrats for many legitimate reasons, it would be foolhardy for reelection-minded legislators to hand over policy prerogatives without checks on agency action. Instead, when designing regulatory agencies, Congress specifies the criteria, rules, and administrative procedures that govern bureaucratic behavior. While this is not a perfect solution to the ubiquitous principal-agent problems of oversight and control (for example, bureaucratic drift), legislators can nonetheless retain both ex ante and ex post control over policy outcomes.[4]

Building upon the insights of these first two bodies of research, a growing literature recognizes that regulatory structure reflects the dynamics of an underlying principal-agent problem between Congress and the bureaucracy. Here the question shifts from why and how Congress delegates to what drives legislators' decision to give agencies substantive discretion in setting policy. What factors motivate Congress's choice? David Epstein and Sharyn O'Halloran (1999) show that more delegation occurs when Congress and the executive have aligned preferences, policy uncertainty is low, and the cost of Congress making policy itself is high. A recurring theme in much of the new political economy literature on agency design is that this conflict arises because of a downstream moral hazard problem between the agency and the regulated firm: that is, there is uncertainty over policy outcomes. Agency structure is thereby endogenous to the political environment in which it operates.[5] This trade-off between distributive losses and informational gains is further elaborated in a series of studies examining the politics of delegation with an executive veto (Volden 2002), civil service protections for bureaucrats (Gailmard and Patty 2007, 2012), and executive review of proposed regulations (Wiseman 2009), among others.[6]

The application of these models to the regulation of banking and financial services would seem to be well motivated. Banking is certainly a complex area where bureaucratic expertise would be valuable; Donald Morgan (2002), for instance, shows that rating agencies disagree significantly more over banks and insurance companies than over other types of firms. Furthermore, continual innovation in

[2] For examples of this logic, see Stigler (1971), Fiorina (1977, 1982), and McCubbins (1985).

[3] This view is articulated most clearly by Lowi (1979), Moe (1984), and Sundquist (1981).

[4] See, for example, McCubbins and Schwartz (1984) and McCubbins, Noll, and Weingast (1987, 1989).

[5] Some excellent technical work has been done on the optimal type of discretion to offer agencies. Nahum D. Melumad and Toshiyuki Shibano (1991) and Ricardo Alonso and Niko Matouschek (2008) provide instances where a principal would prefer to offer a menu of discontinuous choices to an agent receiving authority. Sean Gailmard (2009) demonstrates, however, that in situations where the principal cannot precommit to certain courses of action, interval-type delegation regimes are optimal.

[6] See also Bendor and Meirowitz (2004) for contributions to the spatial model of delegation and Volden and Wiseman (2011) for an overview of the development of this literature.

the financial sector causes older regulations to become less effective, or "decay," over time. If it did not delegate authority in this area, Congress would have to continually pass new legislation to deal with the new forms of financial firms and products, which it has shown neither the ability nor inclination to do.

These insights also overlap with the economic literature on the location of policymaking, as in Maskin and Tirole (2004) and Alesina and Tabellini (2007), both of which emphasize the benefits of delegation to bureaucrats or other non-accountable officials (such as courts) when presented with technical policy issues about which the public would have to pay high costs to become informed. We also draw parallels with the work of Yolande Hiriart and David Martimort (2012), who study the regulation of risky markets and show that when firms cannot be held individually responsible for the consequences of their actions, ex post regulators are faced with the ex ante moral hazard problem of firms engaging in overly risky behavior. Finally, we draw inspiration from agency-based models of corporate finance, as summarized in Tirole (2006).

Overall, then, we have the following testable hypotheses:[7]

1. *Allied principle:* Congress delegates more discretion when:
   a. The preferences of the president and Congress are more similar; and
   b. Uncertainty over market outcomes (moral hazard) is higher.
2. *Uncertainty principle:* The more risk-averse is Congress:
   a. The higher is the overall level of discretion; and
   b. The higher is the level of market regulation.

FINANCIAL REGULATORY STRUCTURE: AN OBSERVATIONAL STUDY

The logic and predictions derived from the theoretical literature described in the previous section inform the research design that we adopted and the subsequent financial regulation database that we constructed. Traditional methods used to test hypotheses rely on observational data to measure the dependent variable, such as financial regulatory structure, and the independent variables, such as differences in policy preferences, to make inferences regarding probable effect. The benefits of this research design are numerous; researchers can: (1) translate a model's theoretical propositions into testable hypotheses; (2) specify the mechanisms by which one variable impacts another; and (3) falsify hypotheses generated by alternative models. This exercise places theoretical arguments within an empirical context, highlighting important factors and thereby contributing to building better theory.

Two main challenges arise with observational studies: precision and validity.[8] To improve the precision of our estimates and mitigate any potential random error generated by compounding effects, we hold constant the issue area, focusing on financial regulatory structure, and employ multiple methods to check the robustness of our measures.[9] To improve the validity of our findings, we compare the current results with a cross-sectional study of all significant laws over the same time period.[10]

Constructing a Financial Regulation Database

Although many excellent histories of financial regulation are available,[11] and despite the popular argument that deregulation of the financial sector played a key role in the recent economic crisis, there is as yet no measure of

[7] For detail proofs of these propositions, see Groll, O'Halloran, and McAllister (2014).

[8] Melissa D.A. Carlson and R. Sean Morrison (1999) define "precision" as the lack of random error or random variation in a study's estimates. "Validity" refers to the extent to which the findings of a study can be generalized.

[9] By analyzing a single issue area, we control for the variance in market uncertainty and downstream (moral hazard) risks.

[10] Here we reference the work of Epstein and O'Halloran (1999).

[11] Here we reference the work of Epstein and O'Halloran (1999).

financial regulatory structure over time.[12] To test the hypotheses that agency discretion responds to the political preferences of Congress and the executive, we therefore created a new database comprising all federal laws and agency rules enacted from 1950 to 2009 that regulate the financial sector.[13]

The unit of analysis is an individual law regulating financial markets. While distinctions between the different types of financial institutions have become blurred over time, for the purposes of this research we define the universe of finance and financial institutions to include state-chartered and federally chartered banks, bank holding companies, thrifts and savings and loan associations, credit unions, investment banks, financial holding companies, securities, broker dealers, commodities, and mortgage lending institutions.

Sample Selection Criteria

Following David Mayhew (2005), we identify the relevant legislation in a three-sweep process. First, we include all laws mentioned in the policy tracker of the relevant issues of *Congressional Quarterly Almanac* (*CQ*) for the categories of banking, the savings and loan industry, the Federal Reserve, the stock market and financial services, insurance, and mortgages, yielding 69 laws. In the second sweep, we review the relevant secondary literature, such as *Banking Law and Regulation* (Macey, Miller, and Carnell 2001), reports by the Congressional Research Service, the websites of the federal banking regulators, and "Legislation in Current Congress" at the Library of Congress's THOMAS website. Any laws not already identified in the first sweep are included, thereby expanding our list by 81 additional laws. In the third sweep, we compare our list of key legislation against John Lapinski's (2008) 1,000 most significant U.S. laws to ensure that our sample covers all critical pieces of financial regulation legislation. Here we add another 5 laws. This process brings the total number of laws in our sample to 155. As our analysis focuses on regulatory design, we omit the mortgage lending laws, resulting in a sample size of 112 financial regulation laws.

The primary source for coding each law is *CQ*'s year-end summary of major legislation (80 laws). When data prove unavailable from *CQ,* we refer to the Library of Congress's THOMAS database (27 laws). When neither source contains sufficient detailed information on a specific law, we refer to the U.S. Statutes (5 laws). In omnibus legislation with a financial regulation subpart, we code only the relevant provisions (9 cases). Each law is then classified as belonging to one or more categories: depository institutions, securities, commodities, insurance, interest rate controls, consumer protection, mortgage lending or government-sponsored enterprises, and state-federal issues.

As a first cut into the analysis, the distribution of financial regulation laws by Congress is illustrated in figure 1, with unified and divided governments shown. At first blush, the figure does not indicate the influence of partisan factors in passing financial legislation; the average number of laws per Congress is almost identical under periods of unified and divided government.

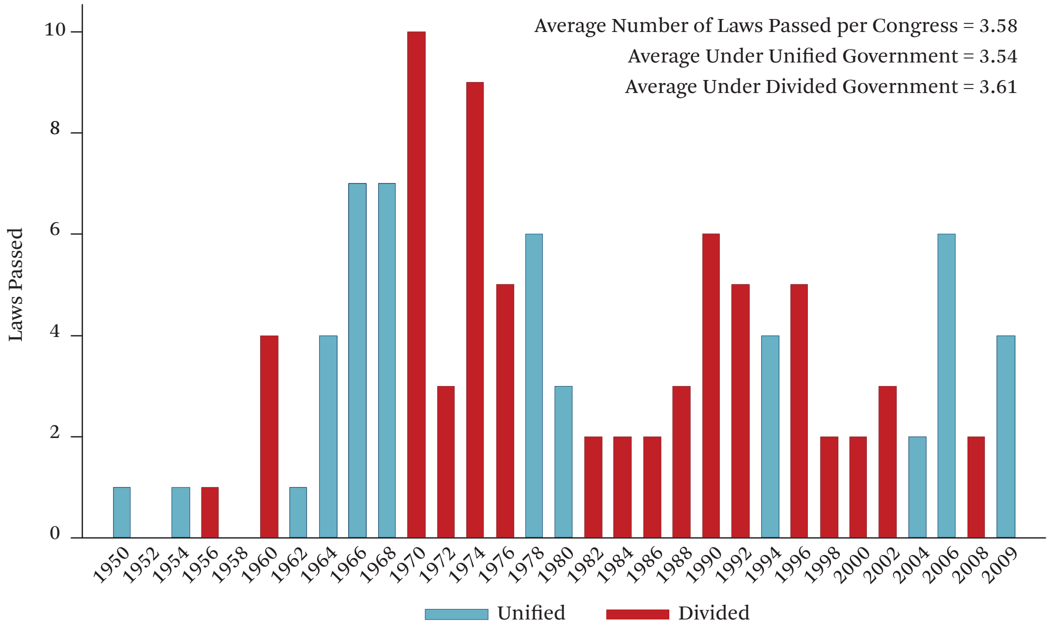
Coding Discretion

Agency discretion depends on both the authority delegated and the associated limits on its

[12] In a recent study of wages in the financial sector over time, Thomas Philippon and Ariell Reshef (2009) developed an index of deregulation, built around summary measures of bank branching restrictions, the separation of commercial and investment banks, interest rate ceilings, and the separation of banks and insurance companies. Unfortunately, their measure codes only for *deregulation* and omits the potential for increases in market regulation as witnessed in the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 (P.L. 111-203). In contrast, we analyze the political and economic determinants of regulatory structure and the subsequent impact on the financial sector. For a detailed discussion of Philippon and Reshef's measure, see the data appendix.

[13] The analysis begins in 1950 because in that year *Congressional Quarterly* started providing consistent reviews of the key provisions of enacted legislation. The major data sets compiled for the financial regulation database are summarized in the data appendix, which also provides the step-by-step manual coding process.

Figure 1. Financial Bills Passed per Congress, 1950–2009



Source: Authors' compilation.

use. Therefore, for each law we code for whether substantive authority is granted to executive agencies, the agency receiving authority (for example, the Securities and Exchange Commission [SEC], the Commodity Futures Trading Commission [CFTC], or the U.S. Treasury), and the location of the agency within the administrative hierarchy (for example, the Executive Office of the President, the cabinet, or an independent agency).

We then identify the procedural constraints circumscribing agency actions.[14] These data provide the bases from which we calculate law-by-law agency discretion.[15]

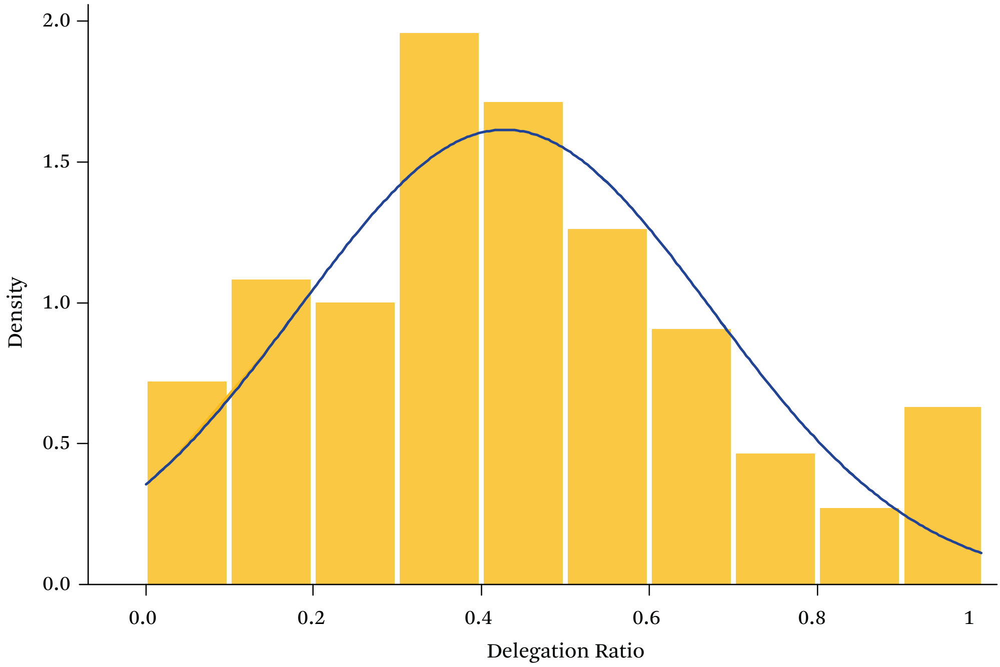
Delegation is defined as authority granted to an executive branch actor to move policy away from the status quo.[16] To measure delegation, then, we read each law in our database independently, number its provisions, and identify and count all provisions that delegate substantive authority to the executive branch. From these tallies, we calculate the delegation

[14] Additionally, we collect the number of regulatory agencies delegated authority per law; this shows the degree to which authority is being divided across executive branch actors. Regulators' degree of autonomy is measured by the relative mix of independent regulatory actors receiving authority, as opposed to actors and executive agencies under more direct presidential control. Each law is also coded for whether it increases, decreases, or leaves unchanged the regulatory stringency of financial markets based on disclosure rules, capital requirements, or increased oversight of products and firms. This enables us to construct a regulation-deregulation index, beginning in 1950 and running to 2010. Table 7 in the data appendix provides descriptive statistics on the key variables used in the analysis.

[15] To ensure the reliability of our measures, each law is coded independently by two separate annotators and reviewed by a third independent annotator, who notes inconsistencies. We then check each law a fourth and final time upon final entry. The data appendix provides a detailed description of the coding method used in the analysis.

[16] For example, the Dodd-Frank Act delegated authority to the Federal Deposit Insurance Corporation (FDIC) to provide for an orderly liquidation process for large, failing financial institutions. See P.L. 111-203, section 210; 124 Stat 1460.

Figure 2. Histogram of Delegation Ratio



Source: Authors' compilation.

ratio by dividing the number of provisions that delegate to the executive over the total number of provisions. A histogram of delegation ratios is shown in figure 2. As indicated, the distribution follows a more or less normal pattern, with a slight spike for those laws with 100 percent delegation. (These usually have a relatively small number of provisions.)

Executive discretion depends not only on the amount of authority delegated but also on the administrative procedures that constrain executive actions.[17] Accordingly, we identify fourteen distinct procedural constraints associated with the delegation of authority and note every time one appears in a law.[18] Including all fourteen categories in our analysis would be unwieldy, so we investigated the feasibility of using principal components analysis to analyze the correlation matrix of the constraint categories. Since only one factor was significant, we calculate first-dimension factor scores for each law, convert them to the [0,1] interval, and term these the "constraint index." Figure 3 displays the histogram of constraints present in each law: the majority of the laws contain four or fewer constraint categories.

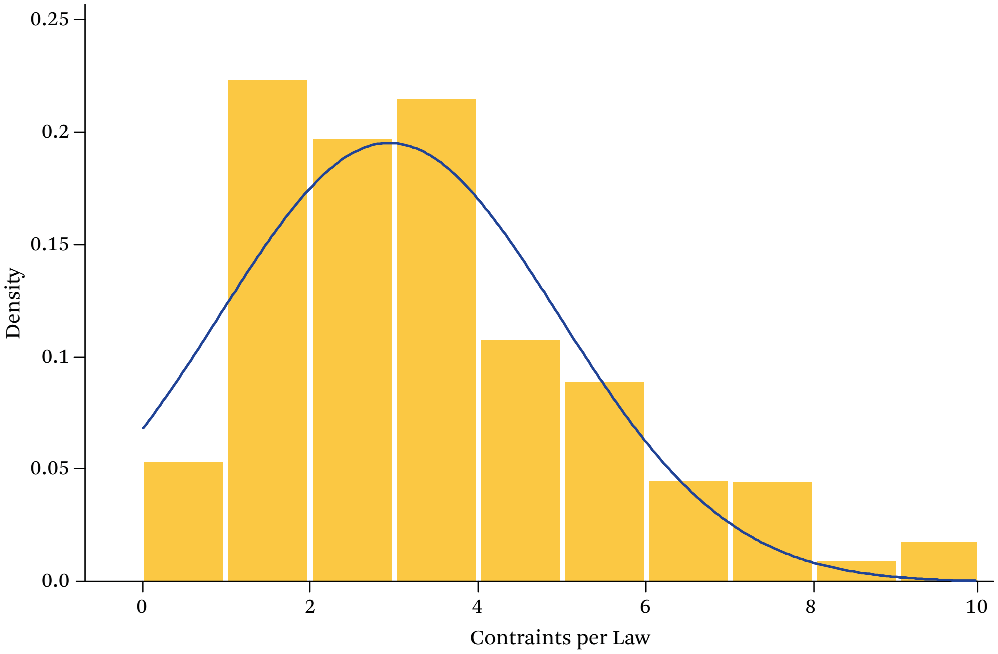
From these data, we calculate an overall "discretion index." For a given law, if the delegation ratio is D and the constraint index is C, both lying between 0 and 1, then discretion is defined as $D * (1 - C)$.[19] The more discretion an agency has to set policy, the greater the leeway it has to regulate market participants. Lower levels of agency discretion are associated with less regulation. Total discretion is thereby defined as delegation minus con-

[17] See McCubbins, Noll, and Weingast (1987).

[18] Each of these categories is coded as constraints above and beyond those required by the 1946 Administrative Procedure Act. For a detailed description of these administrative constraints and their definition, see the data appendix.

[19] See Epstein and O'Halloran (1999) for a complete discussion of this measure.

Figure 3. Histogram of Constraints per Law



Source: Authors' compilation.

straints—that is, the amount of unconstrained authority delegated to executive actors.

To verify the robustness of our estimates and confirm that our choice of aggregation methods for constraints does not unduly impact our discretion measure, figure 4 shows the average discretion index each year calculated four different ways. Since the time series patterns are almost identical, the fourth method (continuous factors, first dimension) is not crucial to the analysis that follows.

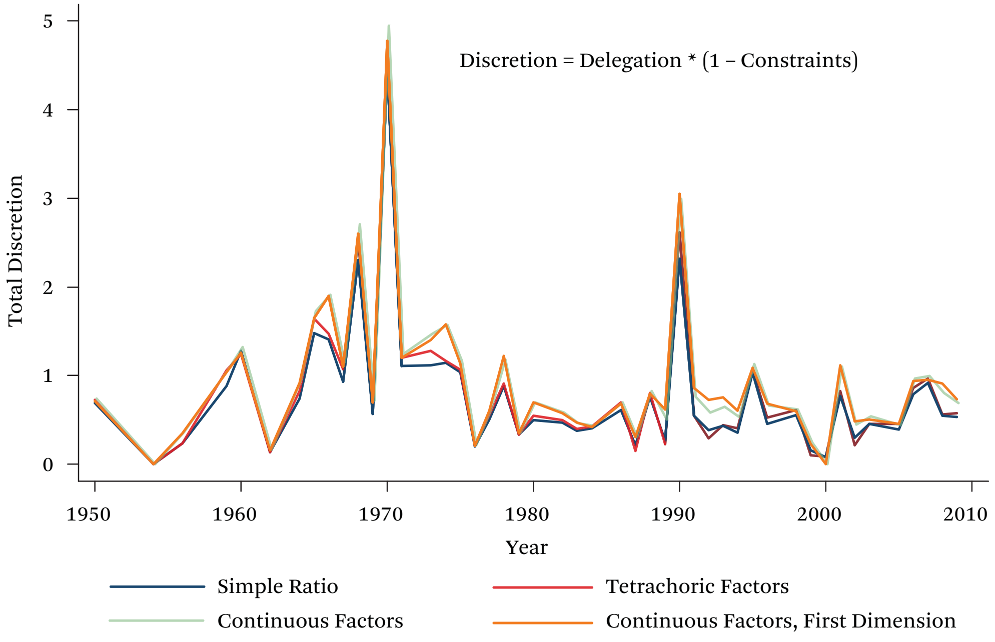
Trends in Agency Discretion

As a basic check on our coding of delegation and regulation, we compare the distribution of the discretion index for laws that regulate the financial industry overall and laws that deregulate it. We would expect from hypothesis 2 that laws regulating the industry would delegate more discretionary authority, and figure 5 shows that this is indeed the case. The average discretion index for the thirty-one laws that deregulate is 0.29, as opposed to 0.36 for the eighty-five laws that regulate. (Five laws neither regulate nor deregulate the industry but rather clarify or qualify a provision in an earlier law.)

These trends pose a puzzle: why was there a strong regulatory response to the spate of financial innovation in the 1960s, a decade that saw an explosion of credit in the economy, including the widespread use of credit cards, accompanied by an increase in the number of credit bureaus (which were unregulated), increased use of computers, and significant growth in both the number and membership of federal credit unions, but no such response to the most recent innovations—derivatives, nonbank lenders, and the rise of the shadow banking system? Both episodes developed under divided government, after all. We return to this question later.

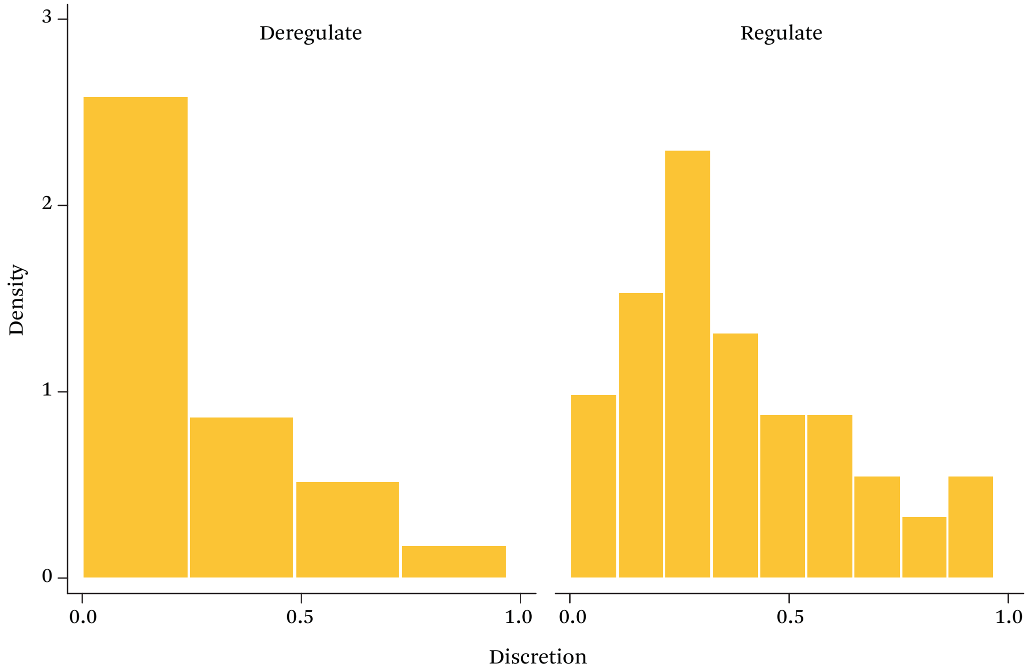
Figure 4 also indicates that the trend in recent decades has been for Congress to give executive branch actors less discretion in financial regulation. Since the Great Society era of the 1960s, and then on into the early 1970s, the total amount of new executive branch authority to regulate the financial sector has generally declined. The exceptions have been a few up-

Figure 4. Four Measures of Executive Discretion



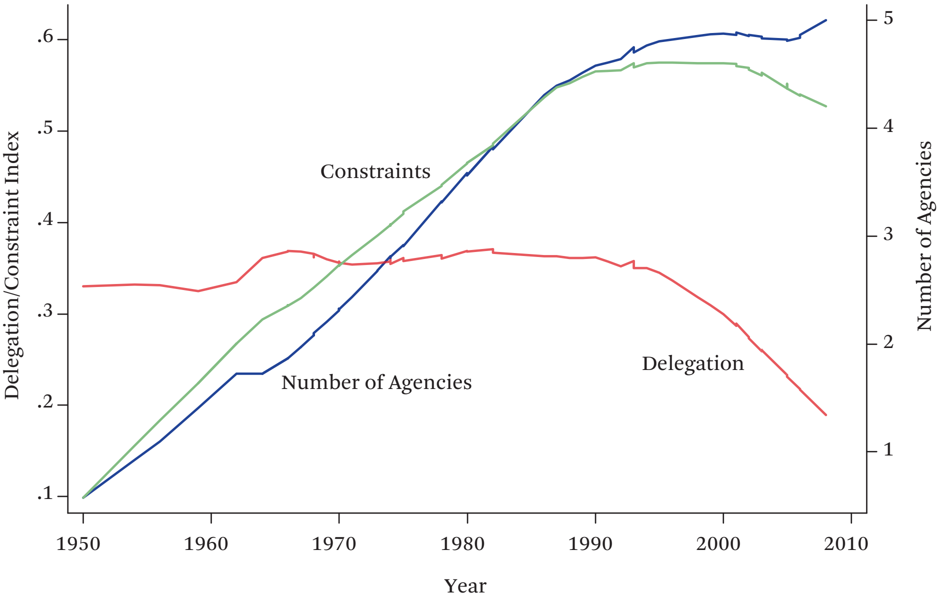
Source: Authors' compilation.

Figure 5. Distribution of Discretion Index for Deregulatory Laws and Regulatory Laws



Source: Authors' compilation.

Figure 6. Delegation, Constraints, and Agencies Receiving Authority, 1950–2008



Source: Authors' compilation.

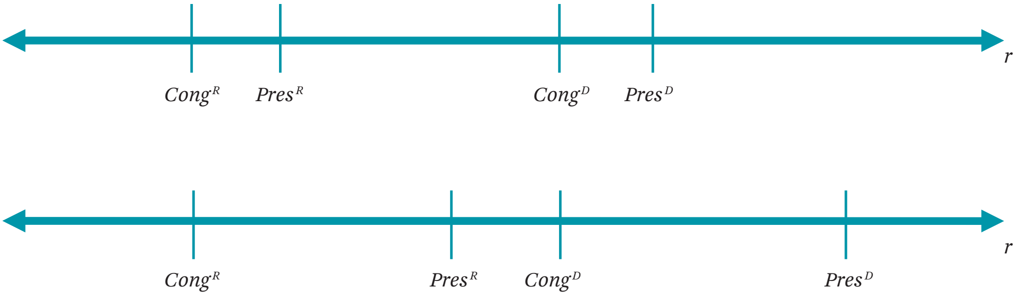
ticks in discretion that coincided with the aftermaths of well-publicized financial crises and scandals, including the savings and loan crisis of the 1980s and 1990s, the Asian crisis of the late 1990s, and the Enron scandal of 2001. Otherwise, the government has been given steadily less authority over time to regulate financial firms, even as innovations in that sector have made the need for regulation greater than ever and even as the importance of the financial sector in the national economy has greatly increased.[20]

What is the source of this decrease in discretion? As shown in figure 6, the amount of authority delegated to oversee the financial sector has remained fairly constant over time, perhaps decreasing slightly in the past decade. The trend in figure 4, then, is due mainly to a large and significant increase in the number of constraints placed on regulators' use of this authority. In addition, we find that the number of actors receiving authority has risen significantly over the time period studied, as also shown in figure 6, and that the location of these agencies in the executive hierarchy has changed, away from more independent agencies to those more directly under the president's control.

Overall, then, our preliminary analysis suggests that the current rules defining financial regulatory structure has created a web of interlocking and conflicting mandates, making it difficult for regulators to innovate in the rules and standards governing the financial industry, while at the same time opening up regulatory agencies to industry lobbying. The problem is not that there is too little regulation, then, but that regulators have too little discretion. Modern laws delegate less, constrain more, and split authority across more agencies than their predecessors. This has led to the heavy regulation of many areas of financial activity by the federal government even as those charged with oversight are hamstrung by overlapping jurisdictions, the need for other actors to sign off on their policies, or outright prohibitions on regulatory actions by Congress.

[20] The size of the financial services sector as a percentage of GDP rose from 3 percent in 1950 to over 8 percent in 2008.

Figure 7. Partisan Effects Captured by Divided Government (Top) and by Cross-Party Coalitions (Bottom)



Source: Authors' compilation.

Analyzing the Financial Regulation Database

Having constructed the financial regulatory structure database, we can now test the comparative statics hypothesis generated from the theoretical literature proposing that Congress delegates greater levels of discretionary authority to executive branch actors with preferences closer to their own. As James Barth, Gerard Caprio, and Ross Levine (2006) report, policymaking in financial regulation tends to be unidimensional, separating actors with more pro-industry preferences from those who place more emphasis on consumer protection.

In the United States over the period studied, Republicans have represented the former viewpoint and Democrats the latter.[21] We also posit that presidents will tend to be less pro-industry than legislators, as their national constituency would lead them to weigh more heavily consumer interests and the stability of the banking system at large.

As figure 7 shows, however, two patterns of delegation are consistent with these constraints. If partisan differences are stronger than interbranch differences, as in the top panel, then delegation should be higher under unified government as opposed to divided government; this was the pattern of delegation found in Epstein and O'Halloran (1999). If interbranch differences predominate, however, as in the bottom panel, then delegation will actually be highest from a Democratic Congress to a Republican president, lowest from a Republican Congress to a Democratic president, and intermediate for the other two combinations. Furthermore, in this "cross-party coalition" case, delegation should increase when Congress is controlled by Democrats, as opposed to Republicans, and when the presidency is controlled by Republicans, as opposed to Democrats.

We thus have the particular prediction that, when regressing discretion on partisan control of the branches, we should obtain a positive and significant coefficient on Democratic control of Congress and Republican control of the presidency. Further, hypothesis 2 predicts that the level of market regulation will also respond to partisan control of Congress: it should increase when Democrats control Congress, as opposed to Republicans, but the party controlling the presidency may or may not matter.

The estimation results are given in table 1. The cross-party partisan conflict variable is constructed to equal 1 when Republicans control Congress and Democrats control the presidency, −1 when Democrats hold Congress and the president is Republican, and 0 otherwise. As predicted, this variable is consistently negative and significant in predicting discretion, while the usual divided government variable is not significant. The signs on Democratic control of Congress and the presidency are also as predicted, as shown in model 3, and the cross-party effects holding constant a number of control variables are added to the regression in model 4.

Models 5 and 6 indicate that when predict-

[21] This is consistent with the findings of Kroszner and Strahan (1999), who analyze roll call votes on bank branching deregulation.

Table 1. Regression Analysis

	Discretion (1)	Discretion (2)	Discretion (3)	Discretion (4)	Regulation/ Deregulation (5)	Regulation/ Deregulation (6)
Cross-party	−0.084 (0.029)***			−0.080 (0.037)**		
Divided		0.043 (0.039)				
Democratic president			−0.066 (0.037)*		−0.573 (0.283)**	−0.637 (0.712)
Democratic congress			0.065 (0.024)***		0.764 (0.173)***	1.546 (0.566)***
Start of term				0.041 (0.042)		
Activist mood				0.012 (0.048)		
Budget deficit				0.027 (0.320)		
Δ DJIA				0.077 (0.133)		
Observations	121	121	121	108	121	23
R squared	0.071	0.011	0.091	0.074	0.169	0.425

Source: Authors' compilation.

Notes: Models 1-4 are OLS regressions with discretion as the independent variable. Models 5 and 6 are ordered probits with regulation/deregulation as the dependent variable. In model 6, only those laws with discretion indices under 0.2 are included in the sample.

ing whether a given law will regulate, deregulate, or leave unchanged the level of regulation of the financial industry, the coefficient on partisan control of Congress is significant in all cases, and in the predicted direction. The coefficient on control of the executive is significant in model 5 as well. Model 6 includes only those cases with a discretion index of 0.2 or under, as the regulation/deregulation relationship should hold most clearly when Congress does not delegate to the executive. Indeed, in these cases the coefficient on Congress remains positive and significant, while the coefficient on control of the presidency is no longer significant.[22]

Limitations of the Observational Method

The above analysis adopts a research design based on observational methods, which potentially suffer from a number of well-known shortcomings. First, observational studies assume that all variables of interest can be measured. For example, the analysis posits that discretion can be calculated as a combination of delegation and constraints. In constructing these measures, the coding rules invariably im-

[22] These results explain the different responses to financial sector innovation mentioned earlier. In the late 1960s and early 1970s, Congress was controlled by a Democratic majority and a Republican, Nixon, held the presidency. This is the cross-party scenario (bottom panel of figure 7) that leads to greater levels of agency discretion and therefore increases in market regulation. In contrast, during the late 1990s and the first decade of the twenty-first century, the Republicans controlled Congress in all but two of the twelve years, while the parties split control of the presidency. In this scenario, the cross-party effect would predict little or no discretion delegated to agencies.

pose a structure on the text, designating some words or phrases as delegation and others as constraints. Moreover, collecting original data is extremely time-consuming, especially when derived from disparate text-based sources, as we do here. The resources needed to extract the appropriate information, train annotators, and code the data can prove prohibitive and are prone to error.

Second, standard econometric techniques, upon which many political economic studies rely, including the one conducted here, face difficulty in analyzing high-dimensional variables that could theoretically be combined in a myriad of ways. For example, figure 4 shows four possible alternatives to calculate the discretion index by varying the weights assigned to the different categories of procedural constraints.

Third, amalgamating the panoply of independent variables into a single index would miss the embedded dimensional structure of the data. For example, rooted in the discretion index are measures of delegation and constraints. Embedded in the delegation ratio and constraint index are additional dimensions: the delegation ratio is a cube formed by the number of provisions that delegate authority to the executive over the total number of provisions; the constraint index is a fourteen-sided polygon.[23]

Our theory identifies specific factors that we expect to impact agency discretion. Of course, other theories might identify different subsets of variables acting through different political processes, which could also have significant impact on legislative and rule-making outcomes. Thus, the social science approach is to define a series of smaller, theory-driven empirical models rather than the more totalitarian kitchen sink models that typify much of big data analysis. This reduction in scope may indeed fail to incorporate certain variables that have surprising and significant impact on the phenomenon of interest. In return, however, the researcher is better able to infer important factors that drive the political process and hence evaluate alternative institutional structures.

NEW MACHINE LEARNING TECHNIQUES TO ANALYZE FINANCIAL REGULATION DATA

Our purpose is to apply computational data science methods, such as NLP and ML algorithms, to financial regulation in order to illustrate how these tools can be used to develop robust indicators of regulatory structure that previously have been limited by dependence on manual coding methods alone. Combining traditional methods with these new computational techniques offers a much richer process to both analyze and understand financial regulation.

Table 2 compares observational methods and data science techniques along four main criteria: coding legislation, structuring data sets, analysis, and internal validity. The table illustrates the limitations of manual rules–based coding methods and the ways in which these new techniques, when appropriately applied, can provide robustness checks on observational studies. Overall, computational analysis helps lessen error, reduce variance, find additional variables and patterns (data features), and add predictive power to models.[24]

Let us consider the example of the Dodd-Frank Act, which covers the activities that financial institutions can undertake, how these institutions will be regulated, and the regulatory architecture itself. The law contains 686 major provisions, of which 322 delegate authority to some 46 federal agencies. In addition, the act has a total of 341 constraints across 11 dif-

[23] Delegation provisions can be even further disaggregated into delegation to the executive, the states, or the courts. For our study, which focuses on only a subset of these data, a neural net trained on first-order interactive effects would yield over 15 million predictive variables.

[24] Lewis's (2014) study shows that manual coding using statistical quality control methods achieves higher levels of inter-annotator consistency, recall, and precision than it is commonly given credit for in the information retrieval literature. Nevertheless, he finds that text classification trained on fewer than 1,000 examples performs even better. When text classification is tuned to hit the same recall target as manual review, it allows fewer laws to be manually checked by lawyers, for a substantial cost reduction.

Table 2. Comparison of Observational Study and New Machine Learning Method

	Observational		Machine Learning	
	Process	Disadvantages	Process	Advantages
Coding congressional bills	Define coding rules	High labor and time costs	Use NLP to recode delegation and constrain provisions for each bill	Efficiency improved
	Assign multiple coders	Coder bias	Represent data in various feature representations as words, semantic units, relations, dependency structures, etc.	Consistency in coding improved
	Conduct multiple-round checks for coding consistency			Detect implicit/latent keywords Scalability
Structuring high-dimensional data set	Use factor analysis to construct discretion index	Limited in assigning meaningful and precise weights	Many ML algorithms can easily handle high-dimensional data	Feature selection algorithms allow us to reduce dimensions easily
	Use pivot table to structure raw data sets			

Analysis	Hypothesis testing, regression analysis, and finding correlations on dependent and independent variables	Limitation in ability to test many hypotheses	Compare the accuracy with text features, with human coding rules features, with the combination of human coding rules and text features, and with the feature selection algorithms	Not limited by the data volume
		Meaningful manual analysis of correlations on thousands of variable is difficult Highly dependent on scaling and sensitive to outliers		Go beyond coding rules to quantify each bill to build discretion model More precise feature selection We can optimize model complexity and predict capacity
Internal validity	Low	Miss the embedded structure of the data and important variables	High	No functional form imposed
	(combine the panoply of independent variables in a single analysis)		(Take account of the raw words of legal bills and explore word relations and other sets of features that otherwise would have been hard to encode manually)	Low generalization error No overfit

Source: Authors’ compilation.

ferent categories, and creates 22 new agencies. If we process the text of this law by the coding method detailed in the previous section, data annotators, trained in political economy theories, would read and code the provisions based on the rulebook provided. In effect, coders would have to read 30,000 words—the length of a short novel. Unlike a novel, however, legislation is written in complex legal language, which must be interpreted correctly and in painstaking detail. Consequently, there is the possibility that data annotators will introduce noise when coding laws.

Data Representation Using Natural Language Processing

Natural language processing is a subfield of computer science that deals with making machines process human (natural) language in the form of text and speech. The algorithms invented in NLP allow machines to better decipher the meaning of text (language understanding) and generate text that conforms with natural language grammar (language generation). For our purposes of processing legislation enacted by Congress, techniques of language understanding are relevant. One important topic in natural language understanding is data representation: how can we best and most appropriately represent text and speech data for machines to understand, and what information can we then extract from a given data structure?

The following text encoding and representation methods are used in NLP:

- *Bag of words:* A bag of words model represents text as a feature vector, where each feature is a word count or weighted word count.
- *Tag sequences:* Sentences or chunks of text are tagged with various information, such as parts of speech (POS) tags or named entities (NEs), which can be used to further process the text.
- *Graphs:* Laws or paragraphs of the laws can be represented in graphs where nodes can model sentences, entities, paragraphs, and connections that represent relations between them.
- *Logical forms:* This is a sequence of words mapped into an organized structure that encodes the semantics of the word sequence.

These methods can be applied to represent text, thereby allowing machines to extract additional information from the words (surface forms) of the documents. Depending on the problem being addressed, one or more of these tools may be useful. We next explain the representation form adopted for our computational analyses.

Computational Analyses: Data Science Methods

We have described the regression models and identified the key independent variables that correlate with the discretion index, defined as $D * (1 - C)$, where D is the delegation ratio and C is the constraint index. We should note that the process discussed earlier is a standard political economy approach to testing hypotheses. In this section, we explore data science methods and identify the techniques best suited to address the limitations of traditional observational methods. In particular, we seek not only to pinpoint important independent variables but also to determine the factors or "features" that predict agency discretion. Identifying the key features, words, or word patterns that predict the level of agency discretion in a given law helps refine and develop better proxies for institutional structure.[25]

We next describe the computational model for predicting the level of agency discretion using NLP and machine learning techniques. We gain significant leverage in building predictive models of agency discretion by employing ad-

[25] Unlike our earlier regression analysis, our purpose here is to find characteristics of the law itself that predict agency discretion. The computational analysis approach lets the data identify those policy features or attributes that most accurately predict outcomes rather than be limited to testing hypotheses about the impact of theoretically motivated independent explanatory variables. We argue that the two approaches—computational analysis and hypothesis testing—are opposite sides of the same coin.

vanced computational data science methods, including the following:

- We are not limited by the amount of data we can process.
- We are not limited to a handful of coding rules to quantify each law for building the discretion model.
- We can take account of the raw text of the law to explore word combinations and syntactic and dependency relations and identify other sets of features that otherwise would be difficult to encode manually.
- We can optimize model complexity and predictive capacity to obtain the optimal model for predicting agency discretion.

Text Classification

We frame the challenge of predicting the level of agency discretion in a given law as a classification problem. We denote the discretion rank as *Rn,* where *n* ranges from 0 to *N*. *N* is the total number of ranks used to tag individual laws for the discretion rank.

The discretion rank *R* in a given law is a subjective measure of how much discretionary authority is granted to the agency in that law only. It is coded from 0 to 5, with 0 indicating that no discretionary authority is given to executive agencies to regulate financial markets and 5 meaning that the law delegates significant discretionary authority.[26]

Processing the Raw Text Data of Individual Laws

We need to represent each individual law in a form suitable for machine learning algorithm to take as inputs. We first convert the raw text of an individual law in feature representation format. For the current analysis, we convert the text of the financial regulation laws into word vectors using a vector space model. We take the following steps to convert text into feature vectors:

Step 1—Data cleaning: For each law, we first clean the text to remove any words that do not represent core content, including meta-information such as dates, public law (P.L.) number, and other meta-data that may have been added by *CQ*.

Step 2—Tokenization: After cleaning the data, we tokenize the text. Tokenization in NLP involves splitting a block of text into a set of tokens (words) by expanding abbreviations ("Mr." becomes "Mister"), expanding words ("I've" becomes "I have"), splitting punctuation from adjoining words ("He said," becomes "He said"), and splitting text using a delimiter such as a white space ("bill was submitted" becomes "(bill) (was) (submitted)"). Tokenization is language-dependent and more difficult in those languages in which word segmentation is not as straightforward as splitting the text at white spaces.

Step 3—Normalization: Once the text is tokenized, we must then normalize the data. The normalization of data requires having consistent tokenization across the same set of words. For example, if we have three different tokens to represent the World Health Organization—"WHO," "W.H.O.," and "World Health Organization"—normalization will map all three into one tokenized form such as "World Health Organization." The normalization step also converts currency, dates, and times into standard formats, such as converting "$24.4 million" into "24 million and 400,000 U.S. dollars." Different representations of dates may be converted into a single canonical form.

Step 4—Vocabulary: To represent text in the form of feature vectors, we need to find the to-

[26] Note that the rank of discretion measure is distinct from the discretion index discussed earlier, which is constructed using detailed coding rules. The rank of discretion measure is determined to be significant when a law gives an agency or agencies authority in a sector or area of activity where none existed previously. Examples include the authority given to the Commodity Futures Trading Commission to regulate derivatives, or the creation of a single agency, such as the Consumer Financial Protection Bureau, to oversee consumer protection across the entire financial sector. The key criteria adopted in assigning a law to one of the five categories are: (1) the importance of the legislation, (2) the impact on the affected industry, and (3) the scope of applicable agency discretion. For example, the Bank Holding Company Act of 1956 gave the Federal Reserve authority to decide which companies could become a bank holding company. In this case, the act was assigned to category 3, as the agency's discretion applied only to a subset of firms.

tal vocabulary of the corpus appended with the additional vocabulary of the language. Any words not in the vocabulary will be considered out-of-vocabulary words, which tend to reduce the accuracy of the model. Hence, it is desirable to have the complete vocabulary of the domain for which we are building the model. We can excerpt vocabulary by extracting all unique tokens from the corpus of the text. If our corpus is small, we can also find pre-extracted vocabulary in a large set of English words, such as the Gigaword corpus.

Step 5—Vector representation: Once we have defined the vocabulary, we can treat each word as adding one dimension in the feature vector that represents a block of text. Thus, let *Li* be the vector representation for law *i*. *Li* = *w*1, *w*2, ... , *wn,* where *wk* represents the existence of word *wk* in the law *Li*. Let us take an example piece of text from the Dodd-Frank Act, contained in section 1506. *Li* = *“the definition of core deposits for the purpose of calculating the insurance premiums of banks.”* Let *N* be the total vocabulary size. The vector representation for this law *Li* will consist of a vector of length *N* where all values are set to zero except for the words that exist in law *Li*. The total vocabulary size *N* tends to be significantly bigger than the number of unique words that exist in a given law, so the vector tends to be very sparse. Hence, the vector *Vi* for law *Li* is stored in sparse form such that only non-zero dimensions of the vector are actually stored. The vector of *Li* will be

$$\begin{gathered} Vi = \{definition = 1.0,\ representation = 1.0, \\ core = 1.0,\ purpose = 1.0,\ calculate = 1.0, \\ insurance = 1.0,\ premium = 1.0, \\ bank = 1.0\}. \end{gathered} \tag{1}$$

This is a binary vector representation of the text *Li*. We can in fact keep track of the word count in the given law *Li* and store counts in the vector instead of storing the binary number representing whether the word is present in the law. Correspondingly, this generates a multinomial vector representation of the same text. If we take the entire Dodd-Frank Act as *Lq,* rather than sample text, and store counts for each word, we yield the vector representation of the act as:

$$\begin{gathered} Vq = \{sec = 517.0,\ financial = 304.0, \\ securities = 106.0,\ requires = 160.0, \\ federal = 154.0,\ requirements = 114.0,\ \ldots, \\ inspection = 2.0\}. \end{gathered} \tag{2}$$

Step 6—TF ∗ IDF *transformation:* Once we represent the laws containing the law in raw word vector format, we can improve the vector representation format by weighting each dimension of the vector with a corresponding inverse document frequency (IDF) (Robertson and Jones 1976). An IDF transformation takes account of giving less weight to words that occur across all laws. For example, if the word “house” occurs frequently in all laws, then it has less distinguishing power for a given class than “SEC,” which may occur less frequently but is strongly tied to a given rank of agency discretion level. We reweight all the dimensions of our vector *Lq* by multiplying them with the corresponding IDF score for the given word. We can obtain IDF scores for each word *wi* by creating an IDF vector that can be computed by equation 3.

N

$$\mathrm{IDF}(w_i) = \log \frac{N}{count - of - Doc - with - w_i} \tag{3}$$

where *N* is the total number of laws in the corpus and $count—of—Doc—with—w_i$ is the total number of laws with the word *wi*. If the word *wi* occurs in all laws, then the IDF score is 0.

Naive Bayes Model

Many different machine learning algorithms are used in text classification problems. One of the most commonly applied algorithms is a naive Bayes method. We build a naive Bayes model for predicting discretion rank for each of the laws *y*. As noted earlier, the discretion rank that we are attempting to predict is based on subjectively labeled data for discretion. In contrast, the discretion index computed earlier is based on the delegation ratio and the constraint index. The discretion rank is a subjective ranking of laws (*Ri*), ranging from 0 to 5, where 0 represents no discretion and 5 represents the highest level of discretion. For ML models, subjective judgment is the gold standard that algorithms have to predict (a standard practice when ML models are built).

Thus, we construct computational models to predict the discretion rank (the "true" subjective rankings [Ri]) instead of the discretion index. With this in mind, let Ri be the discretion rank that we are trying to predict for a given law y.[27]

We need to compute $p(Ri|y)$ for each of the ranks (discretion ranks) and find the rank Ri; we begin by obtaining $p(Ri|y)$ from equation 4:

$$p(R_i|y) = \frac{p(R_i)p(y|R_i)}{p(y)} \tag{4}$$

To find the best rank Ri, we compute the argmax on the class variable:

$$i* = \max p(Ri/y). \tag{5}$$

To compute $p(Ri|y)$, we use Bayes's rule to obtain $p(Ri|y) = (p(y|R_i)*p(R_i))/\mathrm{p(y)}$. Since our task is to find argmax on Ri, we simply need to locate Ri with the highest probability that can be ignored. Because the term $p(y)$ is constant across all different ranks of discretion, it is typically ignored.

Next, we describe how we can compute $p(y|Ri)$ and $p(Ri)$, which is the prior probability of class Ri. This term is computed on the training set by counting the number of occurrences of each discretion rank. In other words, if N is the total number of laws in training and Ni is the number of laws from a given discretion rank i, then $p(Ri) = Ni–Ni/N$.

To compute the probability $p(y|Ri)$, we assume that law y comprises the following words $y = \{w1, w2, \ldots, wn\}$, where n is the number of words in the law y. We make a conditional independence assumption that allows us to express $p(y|Ri) = p(w1, \ldots, wn|Ri)$ as

$$p(w_i, \ldots w_n|R_i = \Pi^n_{j=1} Pw_j|Ri) \tag{6}$$

We compute $P(w_j|R_i)$ by counting the number of times word w_j appears in all of the laws in the training corpus from rank R_i. Generally, add-one smoothing is used to address the words that never occur in the training document. Add-one smoothing is defined as follows: Let Nij be the number of times word wj is found in rank Ri and let $P(wj|Ri)$ be defined by equation 7, where V is the size of the vocabulary.

$$P(w_j|R_i) = \frac{N_{ij}+1}{\Sigma_i N_{ij}+v} \tag{7}$$

Given a test law y, for each word wj in y we look up the probability $P(wj|Ri)$ in the test laws and substitute it into equation 7 to compute the probability of y being predicted as Ri.

For the remainder of this section, we describe the naive Bayes model we built from different sets of features so as to be able to compare the performance of our model in various settings.

Naive Bayes model 1: The first naive Bayes model is based on the law vectors in which the data are all the text found in the financial regulatory laws, which includes more than 12,000 distinct words. Each word is a parameter that must be estimated across each of the six discretion ranks. We took the raw text of the laws and converted it into vectors, as described in the previous section, and estimated the parameters of the naive Bayes model. This model produced an accuracy of 37 percent with an *F*-measure of 0.38.

Our baseline system is a model that predicts rank 0 for all laws. Absent any other information, the best prediction for a law is a rank that has the highest prior probability, which is 0.26 for rank 0. We should note that naive Bayes model 1 based solely on text features did better

[27] The goal of the ML model is to learn patterns that generalize well for unseen data. To do this, we use the model to predict the answer on the evaluation data set and then compare the predicted target to the actual answer. To evaluate the performance of a given ML model in predicting agency discretion, for example, we first assign each law a label or rank *Ri*, ranging from 0 to 6 (ground truth). The value for the discretion rank is assigned by expert evaluators and is deemed the target answer. It is important to note that each law is assigned a category or rank level of discretion independent of the discretion index calculated earlier. Second, we compare the predictions yielded by the ML models against the baseline or target value. Finally, we compute a summary metric; here we use the *F*-statistic, which indicates the accuracy of alternative models in correctly classifying each law relative to the baseline. See Amazon Web Services, *Amazon Machine Learning Developer Guide,* http://docs.aws.amazon.com/machine-learning/latest/dg/what-is-amazon-machine-learning.html (accessed August 9, 2016).

Table 3. Class and Prior Probability for the Six Ranks of the Discretion Index

Class	Prior Probability
0	0.26
1	0.14
2	0.25
3	0.24
4	0.08
5	0.07

Source: Authors' compilation.

than the baseline model by 11 percent. Table 3 shows the prior probabilities for the six ranks of the discretion rank.

Naive Bayes model 2: We first compared the model with features extracted from the raw text derived from the coding rules outlined earlier. We took the same set of laws and their corresponding coding rules as features. We identified more than forty features from the coding rules, including the number of provisions dealing with delegation; constraints such as reporting requirements, exemptions, and appointment power limits; the number of major provisions, the total number of constraint types, and so on. These coding rules are detailed in the guidebook found in the data appendix.

We next created a naive Bayes model using these hand-labeled coding rules as features. Naive Bayes is a general classification algorithm that can take any type of feature vectors as inputs. For model 2, we again estimated the parameters employing the same set of laws that we used to estimate the parameters for building model 1 and produced an accuracy of 30 percent and *F*-measure of 0.40. Interestingly, the raw text model produced a higher level of accuracy than the model built solely from the coding rules.[28]

Naive Bayes model 3: The third model combines the purely raw text approach of examining all of the laws and the manual approach of examining all the laws from the coding rules. We again estimated the parameters described earlier. This model produced an accuracy of 41 percent and an *F*-measure of 0.42. These results indicate that a combination of raw text and manual approaches performs better than either individual approach.

Naive Bayes model 4: The number of parameters for model 1 is almost the same size as the vocabulary of the corpus, while the total number of parameters for model 2 equals the number of manually labeled coding rules. It is likely that the raw text-based features can be overwhelming for a small number of manually labeled features. Therefore, we built a fourth naive Bayes model where we ran a feature selection algorithm on the combined set of features.

Feature selection algorithms select a subset of features based either on different constraints or on the maximization of a given function. We used a correlation-based feature selection algorithm that selects features that are highly correlated with the given class but have low correlation among themselves, as described in Hall (1998). The feature selection algorithm picked up a feature set containing forty-seven features, including a few features from the manually produced coding rules and a few word-based features. Some of the words selected by the feature selection algorithm for discretion rank include: "auditor," "deficit," "depository," "executives," "federal," "prohibited," "provisions," "regulatory," and "restrict."

Model 4 produced the highest level of accuracy at 67 percent with an *F*-measure of 0.68. If the model had no predictive power, then the random assignment of each law to a given rank would be approximately 16 percent. The feature selection improved the accuracy of classifying each law into the correct rank by fourfold. A key reason for such an increase in accuracy was that after discarding a number of word-based features, the smaller feature selection set that remained allowed us to better estimate the parameters with our data set. The best model produced a high degree of accuracy

[28] However, when the data are as highly skewed as they are here, *F*-measure may be more appropriate, since it takes into account both precision and the recall or sensitivity of the analysis. In this case, the *F*-statistics for the rules-based manual coding method performed better than the unstructured computer-generated features.

Table 4. Naive Bayes Models

Feature Type	Accuracy (%)	F-Measure
Model 1: computer-generated text features (C)	36.66	0.38
Model 2: manually coded variables/features (M)	30.00	0.40
Model 3: C + M	40.83	0.42
Model 4: feature selection (C + M)	66.66	0.68

Source: Authors' compilation.

only after careful feature selection and careful model design.

Table 4 summarizes the results of the four models.

CONCLUSION

In this chapter, we have combined observational methods with new computational data science techniques to understand a fundamental problem in political economy—the institutional structure of financial sector regulation. The centerpiece of the study is a database of all financial regulation laws enacted since 1950. The analysis has focused on the delegation of discretionary authority to regulatory agencies with respect to financial sector laws. To improve our estimate of agency discretion and facilitate hypothesis testing, we employ both the observational method and data sciences techniques.

Computational data science captures complex patterns and interactions that are not easily recognized by coding rules. In particular, we apply new NLP and ML techniques to analyze text-based data on congressional legislation to test theories of regulatory design. For instance, these computational methods allow us to represent all the text in a given law as a feature vector where each feature represents a word or weighted terms for words, thereby collaring the relevant terms for different discretion ranks. Furthermore, we can use parsers to find syntactic and dependency parses of sentences that can help quantify intricate connections between the phrases of a sentence with respect to a given implied meaning of a provision. Each of these techniques provides potential improvements over manual coding from a set of defined rules. Yet these computational models rely on the critical data initially produced by subject matter experts to inform or "seed" the model and train complex algorithms. Therefore, big data techniques are not a replacement for observational studies; rather, they should be seen as complements.

Combining both the observational studies and the new machine learning approaches enables us to go beyond the limitations of both methods and offer a more precise interpretation of the determinants of financial regulatory structure. A research strategy that uses more than one technique of data collection can improve the validity of analyzing the high-dimensional data sets commonly found in political economy studies. By illustrating how triangulating different methods can enhance our understanding of important substantive public policy concerns, this paper offers a new path.

REFERENCES

Alesina, Alberto, and Guido Tabellini. 2007. "Bureaucrats or Politicians? Part I: A Single Policy Task." *American Economic Review* 97(1): 169–79.

Alonso, Ricardo, and Niko Matouschek. 2008. "Optimal Delegation." *Review of Economic Studies* 75(1): 259–93.

Barth, James R., Gerard Caprio Jr., and Ross Levine. 2006. *Rethinking Banking Regulation: Till Angels Govern.* New York: Cambridge University Press.

Bendor, Jonathan, and Adam Meirowitz. 2004. "Spatial Models of Delegation." *American Political Science Review* 98(2): 293–310.

Carlson, Melissa D. A., and R. Sean Morrison. 1999."Study Design, Precision, and Validity in Observational Studies." *Journal of Palliative Medicine* 12(1): 77–82.

Clark, William Roberts, and Matt Golder. 1995. "Big Data, Casual Inference, and Formal Theory: Contradictory Trends in Political Science?" *PS: Political Science & Politics APSC* 48(1): 65–70.

Epstein, David and Sharyn O'Halloran. 1999. *Delegating Powers.* New York: Cambridge University Press.

Fiorina, Morris P. 1977. "An Outline for a Model of Party Choice." *American Journal of Political Science* 21(3): 601–25.

———. 1982. "Legislative Choice of Regulatory Forms: Legal Process or Administrative Process?" *Public Choice* 39(1): 33–66.

Gailmard, Sean. 2009. "Discretion Rather than Rules: Choice of Instruments to Control Bureaucratic Policy making." *Political Analysis* 17(1): 25–44.

Gailmard, Sean, and John W. Patty. 2007. "Slackers and Zealots: Civil Service, Policy Discretion, and Bureaucratic Expertise." *American Journal of Political Science* 51(4): 873–89.

———. 2012. "Formal Models of Bureaucracy." *Annual Review of Political Science* 15(1): 353–77.

Groll, Thomas, Sharyn O'Halloran, and Geraldine McAllister. 2014. "Delegation and the Regulation of Finance in the United States Since 1950." Working paper. New York: Columbia University.

Hall, M. A. 1998. "Correlation-Based Feature Subset Selection for Machine Learning." PhD thesis. University of Waikato.

Hiriart, Yolande, and David Martimort. 2012. "How Much Discretion for Risk Regulators?" *The RAND Journal of Economics* 43(2): 283–314. doi: 10.1111/j.1756-2171.2012.001666.x.

Kroszner, Randall S., and Philip E. Strahan. 1999. "What Drives Deregulation? Economics and Politics of the Relaxation of Bank Branching Restrictions." *Quarterly Journal of Economics* 114(4): 1437–67.

Lapinski, John S. 2008. "Policy Substance and Performance in American Lawmaking, 1877–1994." *American Journal of Political Science* 52(2): 235–51.

Lewis, David. 2014. "Supervised Learning in Civil Litigation: A Case Study." Working paper. Washington, D.C.: American Association for the Advancement of Science.

Lowi, Theodore. 1979. *The End of Liberalism: The Second Republic of the United States.* 2nd ed. New York: W. W. Norton.

Macey, Jonathan R., Geoffrey P. Miller, and Richard Scott Carnell. 2001. *Banking Law and Regulation.* New York: Aspen Publishers.

Maskin, Eric, and Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *American Economic Review* 94(4): 1034–54.

Mayhew, David R. 2005. *Divided We Govern: Party Control, Lawmaking, and Investigations, 1946–2002.* New Haven, Conn.: Yale University Press.

McCubbins, Mathew D. 1985. "The Legislative Design of Regulatory Structure." *American Journal of Political Science* 29(4): 721.

McCubbins, Matthew D., Roger Noll, and Barry Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics, and Organization* 3(2): 243–77.

———. 1989. "Structure and Process, Politics and Policy: Administrative Arrangements and the Political Control of Agencies." *Virginia Law Review* 75(2): 431.

McCubbins, Matthew D., and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols Versus Fire Alarms." *American Journal of Science* 28(1): 165.

Melumad, Nahum D., and Toshiyuki Shibano. 1991. "Communication in Settings with No Transfers." *Economics* 22(2): 173.

Moe, Terry M. 1984. "The New Economics of Organization." *American Journal of Political Science* 28(4): 739.

Morgan, Donald. 2002. "Rating Banks: Risk and Uncertainty in an Opaque Industry." *American Economic Review* 92(4): 874–88.

Philippon, Thomas, and Ariell Reshef. 2009. "Wages and Human Capital in the U.S. Financial Industry: 1909–2006." Working Paper 14644. Cambridge, Mass.: National Bureau of Economic Research.

Robertson, S. E., and K. Sparck Jones. 1976. "Relevance Weighting of Search Terms." *Journal of the American Society for Information Science* 27(3): 129–46.

Stigler, George J. 1971. "The Theory of Economic Regulation." *Bell Journal of Economics* 2(1): 3–21.

Sundquist, James L. 1981. *The Decline and Resurgence of Congress.* Washington, D.C.: Brookings Institution.

Tirole, Jean. 2006. *The Theory of Corporate Finance.* Princeton, N.J.: Princeton University Press.

Titiunik, Rocío. 2015. "Can Big Data Solve the Fundamental Problem of Causal Inference?" *PS: Political Science & Politics* 48(1): 75–79.

Volden, Craig. 2002. "A Formal Model of the Politics of Delegation in a Separation of Powers System."

American Journal of Political Science 46(1): 111–33.

Volden, Craig, and Alan Wiseman. 2011. "Formal Approaches to the Study of Congress." In *The Oxford Handbook of the American Congress,* edited by Eric Schickler and Frances E. Lee. Oxford: Oxford University Press.

Wiseman, Alan E. 2009. "Delegation and Positive-Sum Bureaucracies." *Journal of Politics* 71(3): 998–1014.