Big Data and the Regulation of Financial Markets

Sharyn O'Halloran, Sameer Maskey, Geraldine McAllister,
David K. Park and Kaiping Chen
Columbia University

Abstract—**The development of computational data science techniques in natural language processing (NLP) and machine learning (ML) algorithms to analyze large and complex textual information opens new avenues to study intricate processes, such as government regulation of financial markets, at a scale unimaginable even a few years ago. This paper develops scalable NLP and ML algorithms (classification, clustering and ranking methods) that automatically classify laws into various codes/labels, rank feature sets based on use case, and induce best structured representation of sentences for various types of computational analysis. The results provide standardized coding labels of policies to assist regulators to better understand how key policy features impact financial markets.**

Index Terms—**big data, natural language processing, machine learning, political economics, financial regulation**

I. Introduction

This paper combines observational methods with new data science techniques to understand the design of financial regulatory structure in the United States. The centerpiece of the analysis is a large-scale database encoding the text of financial regulation statutes from 1950 to 2010. Among other variables, we identify the amount of discretionary authority Congress delegates to executive agencies and the impact of this regulatory structure on financial markets. 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 manual data tagging project is time consuming, expensive and subject to potential measurement error.

To mitigate these limitations, we employ Natural Language Processing (NLP) and Machine Learning (ML) techniques to complement the observational study. These methods allow us to efficiently process large amounts of texts and represent them in feature vectors, taking into account words, phrases, syntax and semantics. These feature vectors can be easily paired with predefined policy features specified in the manual coding, thereby enabling us to build better predictive models.

The results show that combining observational methods with computational analysis greatly improves the accuracy of the estimates. The analysis offers policy makers a tool to automatically score policy features of financial regulation laws to understand their impact on financial markets. This paper thereby offers a new path, illustrating how triangulating different methods can enhance the understanding of important substantive public policy concerns.

II. The Political Economy of Financial Regulation

What explains the structure of financial regulation? Where, how and by whom policy is made significantly impacts market outcomes. When designing financial regulation laws, Congress specifies the rules and procedures that govern bureaucratic actions. The key is how much discretionary decision making authority Congress delegates to regulatory agencies. In some cases, Congress delegates broad authority, such as mandating the Federal Reserve to ensure the "safety and soundness of the financial system. Other times, Congress delegates limited authority, such as specifying interest rate caps on bank deposits.

A recurring theme in the political economy literature of regulatory design is that the structure of policy making is endogenous to the political environment in which it operates.[1] Epstein and O'Halloran (1999) show that Congress delegates policymaking authority to regulatory agencies when the policy preferences of Congress and the executive are closely aligned, policy uncertainty is low, and the cost (political and otherwise) of Congress setting policy itself is high. Conflict arises because of a downstream moral hazard problem between the agency and the regulated firm, which creates uncertainty over policy outcomes.[2]

Application of these theoretical insights to financial regulation is well-motivated. Banking is a complex policy area where bureaucratic expertise is valuable and market innovation makes outcomes uncertain. 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 the financial sector means that older regulations 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 new forms of financial firms and products, which it has shown neither the ability nor inclination to do.[3] Overall, then, we have the following testable hypotheses:[4] Congress delegates more discretion when: 1) The preferences of the President and

Correspondence should be addressed to Sharyn O'Halloran, Columbia University, New York, NY 10027

[1] For early work in this area, see, for example, McCubbins and Schwartz (1984) and McCubbins, Noll and Weingast (1987; 1989).

[2] Excellent technical work on the optimal type of discretion to offer agencies is provided by Melumad and Shibano (1991) and Alonso and Matouschek (2008), and Gailmard (2009). A series of studies examine 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. 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.

[3] Maskin and Tirole (2004) and Alesina and Tabellini (2007) also emphasize the benefits of delegation to bureaucrats and other non-accountable officials.

[4] For formal proofs of these propositions, the reader is referred to O'Halloran, McAllister and Chen (2014).

ASONAM '15, August 25-28, 2015, Paris, France

ACM ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2808841

Congress are more similar; and 2) Uncertainty over market outcomes (moral hazard) is higher.

A. Financial Regulation Database

To test the hypothesis that regulatory design responds to the political preferences of Congress and the executive, we create a new database comprising all U.S. federal laws enacted from 1950 to 2010 that regulate the financial sector. The unit of analysis is an individual law, which specifies the rules and producers that regulate the actions of financial market participants. The database contains 121 public laws. The average corpus of text of a each legislative summary is 6,278 words.[5]

The key variable of interest is the amount of discretionary authority Congress delegates to regulatory agencies to set policy (Discretion Index). Executive discretion depends not only on the amount of authority delegated (Delegation Ratio) but also on the administrative procedures that constrain executive actions (Constraint Ratio). Therefore, we construct a measure of agency discretion as a two-step process.

a) Delegation Ratio: Delegation is defined as authority granted to an executive branch actor to move policy away from the status quo.[6] For each law, we code if substantive authority is granted to executive agencies, the agency receiving authority (for example, the Securities and Exchange Commission, Treasury, etc.), and the location of the agency within the administrative hierarchy (for example, independent agency, cabinet, etc.).

To measure delegation, each law in our database was read independently, its provisions numbered, and all provisions that delegated substantive authority to the executive branch were identified and counted.[7]

From these tallies, we calculate the delegation ratio by dividing the number of provisions that delegate to the executive by the total number of provisions. In the database, each law contains an average of 27 provisions of which eleven delegate substantive authority to four executive agencies. The average delegation ratio across all laws then is 0.41 or $11/27$.

b) Constraint Ratio: Executive discretion depends not only on the amount of authority delegated but also on the administrative procedures that constrain executive actions. Accordingly, we identify 14 distinct procedural constraints associated with the delegation of authority and note every time one appears in a law.[8] Including all 14 categories in our analysis would be unwieldy, so we investigated the feasibility of using principal components analysis to analyze the correlation matrix of constraint categories. As only one factor was significant, first dimension factor scores for each law were calculated, converted to the [0,1] interval, and termed the constraint index. Each law on average contained three constraints of the possible 14, yielding an overall constraint ratio of 0.21.

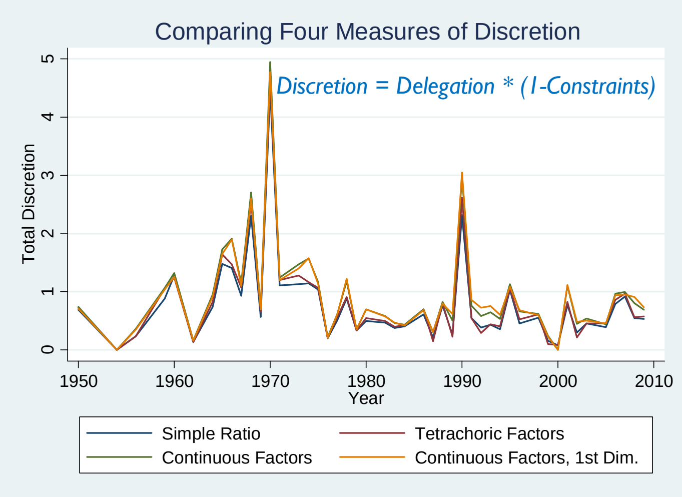

Fig. 1. Four measures of executive discretion.

c) Discretion Index: 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 total discretion is defined as $D*(1-C)$ — that is, the amount of unconstrained authority delegated to executive actors.[9] 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.

As an illustration for how this measure is calculated, the Dodd-Frank Act contains 636 provisions of which 314 delegate authority to 46 executive agencies, yielding a delegation ratio of 0.5. The law also indicated ten procedural constraints out of a possible 14, yielding a constraint index of 0.7 $(10/14)$. Combining delegation and constraints ratios produces a discretion index of $0.5*(1-0.7)=0.1$.

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 1 shows the average discretion index each year calculated four different ways. As the time series patterns are almost identical, our choice of method number four (continuous factors, first dimension) is not crucial to the analysis that follows.

B. An Observational Analysis of Financial Regulation

Having constructed the financial regulation database, we can now test the comparative statics hypothesis described in section II that Congress delegates greater levels of discretionary authority to executive branch actors with preferences closer to their own. As Barth, Caprio, and Levine (2006) report, policymaking in financial regulation tends to be uni-dimensional,

[5] The analysis relies on legislative summaries provided by *Congressional Quarterly* and contained in the Library of Congresss Thomas legislative database.

[6] For example, the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 (Dodd-Frank Act) delegated authority to the the Federal Deposit Insurance Corporation to provide for an orderly liquidation process for large, failing financial institutions.

[7] To ensure the reliability of our measures, each law was coded independently by two separate annotators. It was reviewed by a third independent annotator, who noted inconsistencies. Upon final entry, each law was then checked a fourth time by the authors. O'Halloran et. al. (2015) provides a detailed description of the coding method used in the analysis.

[8] Examples of procedural constraints include spending limits, and legislative action required, etc. See O'Halloran et. al (2015) for a detail description of these constraints.

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

separating actors with more pro-industry preferences from those placing more emphasis on consumer protection. In the United States over the period studied, Republicans have represented the former viewpoint and Democrats, the latter.[10] 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.

Two patterns of delegation are consistent with these constraints. If partisan differences are stronger than inter-branch differences, 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, though, delegation will 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, regulation will also respond to partisan control of Congress, so 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 provided in O'Halloran et. al. (2015) show that the cross-party partisan conflict variable[11] is consistently negative and significant in predicting discretion. The signs on Democratic control of Congress and the presidency are also as predicted and the cross-party effects hold constant even after a number of control variables are added to the regression.[12]

When predicting 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 as well. Restricting the sample to 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.

C. 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 impose a structure on the text, indicating 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 is 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 1 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 high-dimensional structure of the data. For example, rooted in the discretion index are measures of delegation and constraints. Embedded in the delegation and constraint ratios 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 ratio is a fourteen-sided polygon.[13]

III. Data Science Techniques and Financial Regulation

We complement the observational methods by applying NLP and ML algorithms to the financial regulation database detailed above. Let us consider again the example of the Dodd-Frank Act, which covers the activities financial institutions can undertake, how these institutions will be regulated, and the regulatory architecture itself. Recall the law contains 636 major provisions, of which 314 delegate authority to some 46 federal agencies. In addition, the Act has a total of 341 constraints across 11 different categories, with 22 new agencies created. If we process the text of this law by the coding method detailed above, 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 many novels. Unlike novels, 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.

A. Data Representation Using Natural Language Processing

We apply ML methods to the financial regulation database to automatically predict a law's discretion level based on the text. First, however, we need to represent the passages of the legal documents in a format that is suitable for ML methods.

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

[11] This variable equals 1 when Republicans control Congress and Democrats control the presidency, -1 when Democrats hold Congress and the president is Republican, and 0 otherwise.

[12] The analysis is conducted as an ordinary least squares regression, weighted by the number of total provisions associated with each law.

[13] 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.

We employ NLP techniques to convert the text of the laws into feature vectors. Some of the many different ways to encode text into features are listed here:

- 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: Documents or paragraphs of the documents can be represented in graphs where nodes can model sentences, entities, paragraphs and connections represent relations between them.
- Logical Forms: A sequence of words mapped into an organized structure that encodes 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 the computational experiments below.

B. Computational Experiments: Data Science Methods

Section II-A 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 in section II-A is a standard political economy approach to running experiments or, more commonly stated, testing hypotheses. In this section, we explore data science methods and identify the techniques best suited to address the limitations of traditional observational methods as outlined in section II-C. In particular, we seek to determine what factors or "features" of a law that predict agency discretion and also build a predictive model that can predict discretion with high accuracy. Identifying the key features, words or word patterns, that accurately predicts the level of agency discretion in a given law, helps refine and develop better proxies for institutional structure.

We next describe the computational model for predicting the level of agency discretion using NLP and ML techniques. We gain significant leverage building predictive models of agency discretion by employing advanced data science/big data methods, including:

- 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, 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.

1) Text Classification: We frame our problem of predicting the level of agency discretion in a given law as a classification problem. We denote the set of discretion classes as C_n, where n ranges from 0 to N. N is the total number of classes used to tag individual laws for the *Level of Discretion.*

The Level of Discretion D in a given law is a subjective measure of how much discretionary authority given to the agency in that law only. It is coded from 0 to 5, with 0 indicating that no discretionary authority was given to executive agencies to regulate financial markets and 5 meaning that significant discretionary authority was given.

a) Processing Raw Text Data of Individual Law: We must represent each individual law in a form suitable for ML algorithms to take as inputs. We first convert the raw text of an individual law in feature representation format. For the current experiment, we convert the text of the financial regulation laws to Word Vectors. We describe the process of converting text into feature vectors below.

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 metadata that may have been added by *CQ Almanac*.

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). This involves expanding abbreviations *(Mr. > Mister)*, expanding words *(I've > I have)*, splitting punctuation from adjoining words *(He said, > He said,)* and splitting text using a delimiter such as white space *(bill was submitted > [(bill) (was) (submitted)]).*

Step 3: Normalization - Once tokenized, we must then normalize the data. The normalization of data involves having consistent tokenization across the same set of words.

Step 4: Vocabulary - In order to represent text in the form of feature vectors we must find the total vocabulary of the corpus appended with the additional vocabulary of the language.

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.

Let D_i be the vector representation for document i. $D_i = w_1, w_2, ..., w_n$, where w_k represent the existence of word w_k in the document D_i. Let us take an example piece of text from the Dodd-Frank Act, contained in section 1506.

D_i=*"..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 document D_i will consist of a vector of length N where all values are set to zero except for the words that exist in document D_i. The total vocabulary size N tends to be significantly bigger than the number of unique words that exist in a given document so the vector tends to be very sparse. Hence, the vector V_i for document D_i is stored in sparse form such that only non-zero dimensions of the vector are actually stored. The vector of D_i will be

V$_i$ = $\{definition = 1.0, representation = 1.0, core = 1.0, purpose = 1.0, calculate = 1.0, insurance = 1.0, premium = 1.0, bank = 1.0\}$.

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

V$_q$ $= \{sec = 517.0, financial = 304.0, securities = 106.0, requires = 160.0, federal = 154.0, requirements = 114.0, ..., inspection = 2.0\}$

Step 6: $TF * IDF$ Transformation - Once we represent the document in raw word vector format, we can improve the vector representation format by weighting each dimension of the vector with a corresponding term known as *Inverse Document Frequency* (IDF) [20]. An IDF transformation takes account of giving less weight to words that occur across all documents. For example, if the word *SEC* occurs frequently in all laws then word *SEC* has less distinguishing power for a given class than *house*, which may occur less frequently, but is strongly tied to a given class. We re-weight all the dimensions of our vector D_q by multiplying them with the corresponding IDF score for the given word. We can obtain IDF scores for each word w_i by creating a IDF vector that can be computed by Equation 1.

$$IDF(w_i) = log(\frac{N}{count - of - Doc - with - w_i}), \quad (1)$$

where N is the total number of documents in the corpus and $count - of - Doc - with - w_i$ is the total number of documents with the word w_i. If the word w_i occurs in all documents then IDF score is 0.

2) *Naive Bayes Model:* Many different machine learning algorithms are used in document/text classification problems. One of the most commonly applied algorithms is a Naive Bayes method. We build a Naive Bayes Model for predicting discretion level for each of the laws y. As noted above, the *Discretion Level* that we are attempting to predict is based on subjectively labeled data for discretion. In contrast, the Discretion Index computed in section II-A is based on the delegation ratio and constraint index. The Discretion Level is a subjective label ranging from 0 to 5, where 0 represents no discretion while 5 represents the highest level of discretion. For ML models subjective judgement 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 Level (the gold standard subjective labels) instead of the Discretion Index. With this in mind, let C_i be the level of discretion that we are trying to predict for a given document (law) y.

We must compute $p(C_i|y)$ for each of the classes (discretion levels) and find the class C_i. $p(C_i|y)$ can be obtained by Equation 2

$$p(C_i|y) = \frac{p(C_i)p(y|C_i)}{p(y)} \quad (2)$$

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

$$i^* = arg \max_i p(C_i|y). \quad (3)$$

To compute $p(C_i|y)$, we use Bayes rule to obtain $p(C_i|y) = \frac{p(y|C_i)*p(C_i)}{p(y)}$. Since our task is to find argmax on C_i, we simply need to locate C_i with the highest probability that can be ignored. As the term $p(y)$ is constant across all different classes, it is typically ignored. Next, we describe how we can compute $p(y|C_i)$ and $p(C_i)$.

$p(C_i)$ is the prior probability of class C_i. This term is computed on the training set by counting the number of occurrences of each class. In other words, if N is the total number of documents in training and N_i is the number of documents from class i, then $P(C_i) = \frac{N_i}{N}$.

In order to compute the probability $p(y|C_i)$, we assume that document y is comprised of the following words $y = \{w_1, w_2, \ldots, w_n\}$, where n is the number of words in the document y. We make a conditional independence assumption that allows us to express $p(y|C_i) = p(w_1, \ldots, w_n|C_i)$ as

$$p(w_1, \ldots w_n|C_i) = \prod_{j=1}^{n} P(w_j|C_i). \quad (4)$$

We compute $P(w_j|C_i)$ by counting the number of times word w_j appears in all of the documents in the training corpus from class C_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 N_{ij} be the number of times word w_j is found in class C_i and let $P(w_j|C_i)$ be defined by equation 5, where V is the size of the vocabulary.

$$P(w_j|C_i) = \frac{N_{ij} + 1}{\sum_i N_{ij} + V} \quad (5)$$

Given a test document y, for each word w_j in y, we look up the probability $P(w_j|C_i)$ in this test document and substitute it into equation 5 to compute the probability of y being predicted as C_i. In the following sections, we describe the Naive Bayes Model built from different sets of features, thereby allowing us to compare the performance of our model in various settings.

IV. Comparing Methods

A. Naive Bayes Model 1

The first Naive Bayes Model is based on the document vectors where the data is 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 classes. We took the raw text of the laws and converted it into document vectors as described in the previous section and estimated the parameters of Naive Bayes Model. This model produced an accuracy of 37% with an F-Measure of 0.38.

Our baseline system is a model that predicts Class 0 for all documents. Absent any other information, the best prediction for a document is a class that has the highest prior probability, which is 0.26 for Class 0. We should note that the Naive Bayes Model 1 based solely on text features does better than the baseline model by 11%.

Table I shows the prior probabilities for the six classes of Discretion.

TABLE I
CLASS AND PRIOR PROBABILITY

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

B. Naive Bayes Model 2

We first compare the model with features extracted from the raw text derived from the coding rules outlined above. We take the same set of laws and their corresponding coding rules as features. We identified more than 40 features from the coding rules, including the Number of Provisions with Delegation, constraints such as Reporting Requirements, Time Limits, et cetera. We next created a second 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 using the same set of laws that was used to estimate the parameters for building Model 1, and produced an accuracy of 30.0% 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. When we build a Naive Bayes Models with manually hand coded features the model parameters are estimated in a similar fashion as stated in Equation 4 except instead of words w_j we have hand coded features h_k as described in Equation 6.

$$p(h_1, \ldots, h_m|C_i) = \prod_{k=1}^{m} P(h_k|C_i). \tag{6}$$

C. Naive Bayes Model 3

Naive Bayes Model 3 combines the purely raw text approach of examining all of the text and the manual approach in which we examine all the text from the coding rules. We again estimated the parameters as described in III-B2. This model produces an accuracy of 41% and an F-measure of 0.42. These results indicate that a combination of both raw text and manual approaches performs better than either individual approach. When we combine the features we are pooling both sets of w_j and h_k features into same pool. For the estimation of $p(w_i, \ldots, w_n, h_k, \ldots, h_m|C_i)$ we again assume conditional independence among features given the class allowing us to efficiently compute $p(w_i, \ldots, w_n, h_k, \ldots, h_m|C_i)$ using the following equation $\prod_{j=1}^{n} P(w_j|C_i). \prod_{k=1}^{m} P(h_k|C_i)$.

D. 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 on different constraints or on the maximization of a given function. We used a correlation-based feature selection algorithm which selects features that are highly correlated within a given class, but with low correlation across classes, as described in Hall (1998). The feature selection algorithm picked up a feature set containing 47 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 of Discretion Level include: *auditor, deficit, depository, executives, federal, prohibited, provisions, regulatory, and restrict*.

Model 4 produced the highest level of accuracy at 67% with an F-measure of 0.68. One of the reasons for such an increase in accuracy is that after discarding a number of word-based features, the smaller feature set that remained allowed us to better estimate the parameters with our data set of 121 laws reducing the data sparsity problem. The best model produced a high degree of accuracy only after careful feature selection and model design.

Table II summarizes the results of the four models.

TABLE II
NAIVE BAYES MODELS

Feature Type	Accuracy(%)	F-Measure
Model 1: Computer Generated Text Features (C)	36.66	0.38
Model 2: Manual 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

V. CONCLUSION

This paper develops scalable computational data science tools to understand a fundamental problem in political economy, the institutional structure of financial regulation. To improve our estimate of agency discretion and facilitate hypothesis testing, we employ both the observational method and computational data science 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 to test theories of regulatory design. These computational methods allow us to represent the text in a given law as a feature vector where each feature represents a word or weighted terms for words, thereby collating the relevant terms for different levels of discretion. Each of these techniques provide 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 can be seen as a complement to observational studies.

Combining both the observational studies and the new machines 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 high-dimensional datasets commonly found in political economy studies. The practical implications of the analysis are manifold. The analytical methods developed enable governments and financial market participants alike to: 1) automatically score policy choices and link them to various indicators of financial sector performance; 2) simulate the impact of various policies or combinations of policy under varying economic and political conditions; and 3) detect the rate of change of market innovation by comparing trends of policy efficacy overtime. The analysis will help governments to better evaluate the effect of the policy choices they confront, as well as assist business communities to better understand the impact of those choices on the competitive environment they face.

References

[1] Alesina, Alberto and Guido Tabellini. 2007. "Bureaucrats or Politicians? Part I: A Single Policy Task." *The American Economic Review* 97, no. 1: 169–179.

[2] Alonso, Ricardo and Niko Matouschek. 2008. "Optimal Delegation." *The Review of Economic Studies* 75, no. 1: 259–93.

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

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

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

[6] Epstein, David and Sharyn O'Halloran. 2009. "Avoiding Financial Katrinas: Systemic Risk as a Common Pool Problem." Working Paper, Columbia University.

[7] Gailmard, Sean. 2009. "Discretion Rather Than Rules: Choice of Instruments to Constrain Bureaucratic Policy-Making." *Political Analysis* 17(1): 25–44.

[8] Gailmard, Sean. 2009. "Multiple Principals and Oversight of Bureaucratic Policy-Making." *Journal of Theoretical Politics* 21(2): 161–86.

[9] 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.

[10] Gailmard, Sean and John Patty. 2012. "Formal Models of Bureaucracy." *Annual Review of Political Science* 15: 353–77.

[11] Hall, M.A. 1998. "Correlation-based Feature Subset Selection for Machine Learning." *Phd Thesis*. University of Waikato.

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

[13] McCubbins, Mathew D. and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28, no. 1: 165–79.

[14] McCubbins, Mathew D., Roger Noll and Barry Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics and Organization* 3: 243–77.

[15] McCubbins, Mathew D., Roger Noll and Barry Weingast. 1989. "Structure and Process, Politics and Policy: Administrative Arrangements and the Political Control of Agencies." *Virginia Law Review* 75: 431–82

[16] Melumad, Nahum D., and Toshiyuki Shibano. 1991. "Communication in settings with no transfers." *RAND Journal of Economics* 22(2): 173–98.

[17] Morgan, D.P.. 2002. "Rating Banks: Risk and Uncertainty in an Opaque Industry." *The American Economic Review* 92(4): 874-888.

[18] O'Halloran, Sharyn, Geraldine McAllister and Kaiping Chen. 2014. Working Paper, Columbia University.

[19] O'Halloran, Sharyn, et al. 2015. "Data Science and Political Economy: Application to Financial Regulatory Structure." Forthcoming. In Howard Rosenthal , ed. *Big Data and Political Economy.* New York: Russell Sage Press.

[20] Sparck Jones, K. 1972. "A statistical interpretation of term specificity and its application in retrieval" *Journal of Documentation* 28: 11-21.

[21] Volden, Craig. 2002. "A Formal Model of the Politics of Delegation in a Sep- aration of Powers System." *American Journal of Political Science* 46(1):111-133.

[22] Volden, Craig and Alan Wiseman. 2011. "Formal Approaches to the Study of Congress." In Eric Schickler and Frances Lee, eds. *Oxford Handbook of Congress.* Oxford: Oxford University Press pp. 36-65.

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