1 Introduction

This Element develops a method for conducting automated impact analyses of court precedent and applies it to criminal sentencing. This topic has received much attention due to the massive build-up of prisons in the US criminal justice system. We apply methods from machine learning, natural language processing, and causal inference to measure the causal impact of criminal appeal decisions in circuit courts.

Legal theorists and historians have long debated the proper relationship between constitutional law and politics. While some have argued that judicial decision-making should be political (Schmitt 1969, 1985, 2005), most scholars have emphasized the importance of separation from political interests. Debates over the political role of the judiciary have intensified in recent years. This Element assesses the impacts of ideological motivations of United States US federal judges as reflected in their rulings and subsequent compliance by federal courts as evidence that this debate over judicial decision-making has consequences. We test the effects of legal precedent in criminal justice on subsequent sentencing decisions of district court judges and sentencing charges by federal prosecutors.

To conduct our analysis, we represent the text of judicial decisions as data. We then take these text features, along with metadata about the judges and case facts, to predict appeal court decisions (affirm/reverse) and district court sentencing decisions (length of sentence, in months). Using a high-dimensional instrumental variables approach, we measure the causal relations underlying these processes.

Our approach is based on Hartford and colleagues (2017). The prediction problem is divided into a two-stage model. In the first stage, we fit models that learn to predict appeal decisions of the circuit court as well as the vector representation of judge opinion text, where the instruments include characteristics of assigned judges. Intuitively, the Deep IV methods will be beneficial in predicting a high-dimensional embedding vector describing the text features of the written decision. In the second stage, we predict district court sentencing length decisions. These models use the first-stage predictions as inputs, so the resulting model parameters have a causal interpretation. We compare these Deep IV predictions to the noncausal Deep ordinary least squares (OLS) predictions and the Deep Reduced Form predictions that use only the judge characteristics as regressors. We also report feature importance and OLS coefficients. The reduced form model is used to substantiate causality and aid in interpretability.

We find that an appeal case that affirms a lower-court crime decision (i.e., a decision to be harsh) is followed by a statistically significant increase in
sentencing percentile relative to sentencing guidelines in the lower courts of that circuit. However, there is a statistically insignificant effect on sentence lengths. Sentence guidelines dictating the minimum and maximum are based on a formula using the prosecutor’s charge. We therefore interpret these results as being due to the interplay of prosecutors and judges, where prosecutors backlash to circuit rulings by issuing more lenient charges after a harsh ruling (or, conversely, harsh charges after a lenient ruling), yet district judges are largely obeying the circuit rulings. This is consistent with the growing attention to the large role for discretion in decision-making by prosecutors.

2 Theoretical Framework

There is an extensive research literature on the topic of judicial decision-making and sentencing. And it is clear that contextual factors related to political, judicial, and social environments affect prison sentences (Huang et al. 1996). Regional variation in sentencing has been documented in a lot of research, both at the local (Fearn 2007) and at the district or circuit level (Kautt 2002). This Element examines the casual link between legal rulings on appeal decisions in circuit courts and the subsequent sentencing decisions in the lower district courts within the circuit jurisdiction. We are unaware of any previous study of this causal question for sentencing and, more broadly, of how judicial writing style affects downstream outcomes.

In order to measure the causal impact, this Element considers sentencing lengths to be influenced by latent covariates from various political, social, and economic factors. At the core of our methodology is the use of features generated from a naturally occurring random process in our prediction task. We exploit the fact that judges of each case are randomly assigned, and we take judge characteristics as an instrumental variable (Chen et al. 2016).

2.1 Related Works on Law

Two decades ago, there were three main theories of judicial behavior – legal, attitudinal, and self-interested – the first posits that judges follow formal rules or legal philosophy (Kornhauser 1999). The latter two assume some form of bias: for example, the attitudinal model posits that judges follow political preferences (Cameron 1993) and the self-interested model posits that judges maximize their utility (Posner 1973). The distinction between legal and attitudinal is subtle: for instance, in a legal model, a judge can adhere to a strict interpretation of the Constitution, while, in an attitudinal model, the same behavior is interpreted as simply hewing to the preferences of a political party.
In recent years, the self-interested model has been reconceptualized to include identity accounts: for instance, one might gain identity utility from voting in a manner consistent with religious identity. On the other hand, self-interested decision-making can be attributed to seeking promotion. Finally, the behavioral economics revolution has entered into studies of judicial behavior. Thinking-fast judging would ascribe many of the cognitive and behavioral errors to the lack of slow, deliberate, intentional thinking.

It is a useful exercise to outline what the models of judicial behavior would predict in response to a precedent. In a legal model of judicial behavior, the judge makes decisions according to their perception of the law. Thus, a “legal” judge would follow the precedent because of the new legal rule. In an attitudinal model of judicial behavior, the precedent is only having effects on the political party’s preferences. Thus, an “attitudinal” judge would follow the precedent if the political party also shifts its preferences in accordance with the precedent. In an identity account of judicial behavior, the precedent has effects if their group identity has, as a group, responded to the precedent. Thus, an “identity”-motivated judge would respond to the precedent in the same manner as their group. In a labor market model of judicial behavior, the judge makes a decision to maximize the likelihood of promotion. Following the precedent reduces the risk of reversal, which would in typical circumstances increase the likelihood of promotion. Thus, a “labor market” judge, like the “legal” judge, would also choose to follow precedent. Finally, a thinking-fast judge would make decisions according to cognitive bias or error. A precedent would influence this judge for a couple of reasons. One reason could be that the judge relies on heuristics and follows the recent precedent as a heuristic.

None of these theories easily explain a potential backlash to the precedent. Some economists have offered a unified framework for understanding how humans respond to laws and regulations. Subordinate judges in a hierarchical court system are human. Their behavioral response to a law or regulation can fit under the unified framework.

On a theoretical level, it is widely presumed that the law can affect moral values and behavior simply through its expressive power. Formal models of law (e.g., Benabou and Tirole 2011) illustrate how laws can affect the morality of particular actions. This framework examines the implications of three motivations for human behavior: intrinsic motivations (i.e., values, including ideological or identity-based motives), extrinsic motivations (i.e., material incentives, including pecuniary incentives), and social motivations (i.e., norms). Social motivations arise from the honor or stigma attributed to an individual acting outside the norm. People would like to signal their type (i.e., values) and appear moral to gain honor or avoid stigma. Legal decisions inform people about social
norms. Prohibitions cause people to think that the government sees a problem. We call this an “expressive effect” when law causes what is viewed as moral to shift toward what the law values. Those who are motivated by intrinsic incentives have an easier time signaling to others as honorable. This expressive effect, however, only arises when a sufficient number of people do the stigmatized activity. When the normalizing effect exceeds the signaling effect, we call this a “backlash effect,” as the law causes what is viewed as moral to shift against what the law values. When few people do the stigmatized activity, the social perception of stigmatized activities can increase substantially if the shift in beliefs causes stigmatized activities to become normalized. We use the data and methodology to test this model.

The Benabou and Tirole (2011) model encompasses many existing theories of judicial behavior when this formal framework is applied to judicial decision-making in response to legal rules. Labor market motives fall under extrinsic motivations. Legal and attitudinal motives fall under intrinsic motivations. Group identity motives fall under social motivations. Strictness or leniency in criminal justice could be stigmatized. Cognitive error is not modeled.

Our analyses are restricted to testing the causal effects of legal rulings on judicial decisions. The existing evidence on compliance of judges in lower courts to higher court rulings is scant, but some quantitative evidence exists from the USA and from Norway (Bhueller and Sigstad 2021; Chen and Frankenreiter 2021).

2.2 Related Work on Machine Learning

Methodologically, previous work by Hartford and colleagues (2017) indicates that when doing counterfactual predictions, there is a benefit from a deep instrumental variable framework, which is a two-stage deep neural network instrumental variables method. The Deep IV framework can outperform both traditional two-stage OLS and standard feed-forward networks by significantly reducing counterfactual errors.

The field of counterfactual analysis has been developing fast and has started gaining more attention from the machine learning community in recent years. Recent work from Lewis and Syrgkanis (2018) uses generative adversarial networks (GANs) and finds that GANs have a similar or better performance compared to both direct models and other forms of two-stage models. Egami and colleagues (2017) also used a related method to measure treatment effects from text and showed applicability – however, in most of their papers, the model is tested on simulated data. In contrast, the focus of this Element is on a real, complex data environment. Other papers that connect machine
learning with estimating treatment effects in economics, law, and policy include Double ML (Chernozhukov et al. 2016), Causal Forest (Athey et al. 2019), and Orthogonal Random Forest (Oprescu et al. 2019).

### 3 Data Set

#### 3.1 Data Set Description

This Element constructs the final data set for analysis using four raw data sets. Here, we present brief descriptions for each of them.

**3.1.1 Cleaned Circuit Court Case Data**

First, we have raw text records of 253,164 Circuit Court Opinions collected from 1991 to 2013, organized by year, case identification number, opinion type, and author’s (judge’s) last name. They contain 82,635 unique cases, 3,288 unique judge names, and 14 unique opinion types. Seventy-five percent of the cases are stated as being affirmed, and 25 percent are stated as being reversed.

**3.1.2 Judge Biographical Characteristics**

Second, we have demographic and background information for about 714 unique judges. The information contains a mixture of 186 numerical, textual, and categorical features, including the judges’ name, age, and party affiliation, as well as their education and career backgrounds.

**3.1.3 District Courts Sentencing Data**

Third, we have the data set on district court sentencing information. The feature we use here is the sentencing length, which is a numerical feature ranging from 0 to 999. The number 999 represents the death sentence and hence will not be treated as a numeric value. We use interquartile range to detect outliers in the data set and thus consider data points with sentencing length greater than 152.5 as outliers. We eliminate those data from the analysis. The ones with missing values are also excluded from our analysis. The district courts’ sentencing data are later joined with circuit court data by using the US state and the date of sentencing.

**3.1.4 Circuit Cases Metadata**

Fourth, we have a data set containing rich metadata for each circuit case, including the case ID, decision, date, three concurring judges, and case type. We use these data to filter out criminal cases that can be matched with opinion records to extract case and judge information.
Table 1  Binary decision grouping rules

<table>
<thead>
<tr>
<th>Original category</th>
<th>Grouped as</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stay, petition, or motion granted</td>
<td>Reversed</td>
</tr>
<tr>
<td>Reversed (include reversed &amp; vacated)</td>
<td>Reversed</td>
</tr>
<tr>
<td>Reversed and remanded (or just remanded)</td>
<td>Reversed</td>
</tr>
<tr>
<td>Vacated &amp; remanded; set aside &amp; remanded; modified &amp; remanded</td>
<td>Reversed</td>
</tr>
<tr>
<td>Vacated</td>
<td>Reversed</td>
</tr>
<tr>
<td>Affirmed; or affirmed &amp; petition denied</td>
<td>Affirmed</td>
</tr>
<tr>
<td>Petition denied or appeal dismissed</td>
<td>Affirmed</td>
</tr>
<tr>
<td>Affirmed in part &amp; reversed in part; modified; Affirmed &amp; modified</td>
<td>Dropped</td>
</tr>
<tr>
<td>Affirmed in part, reversed in part, and remanded</td>
<td>Dropped</td>
</tr>
</tbody>
</table>

3.2 Data Preprocessing

3.2.1 Feature Engineering

Demeaning Features: Many features in our data are potentially endogenous to court and time. For example, the number of Democrats in the court may be different each year and could have a confounding trend with outcomes. Since our data span across twenty-three years, the changes over time might be significant. In addition, the cases are randomly assigned to judges conditional on the circuit and year. We therefore demean instruments by circuit-year to reduce the effects of confounding trends.

Target Calculation: We normalize the specified action string for the appeal decision to a binary variable, affirm or reverse. We group the seven action categories using the rules in Table 1.

We are interested in measuring the effect of an appeal decision. Therefore, we set the target variable as the change in the average sentencing length before and after an appeal decision. To do this, we measure the sentencing length changes followed by a circuit court decision using the three months before and after the decision. We subtract the average sentencing length of three months before the decision from the average sentencing length of three months after the decision. This can be seen as a first-differenced outcome by case.

3.2.2 Representing Case Text as Data

Apart from the binary appeal action (affirm or reverse), we are also interested in whether the explanation for that action – the written opinion – might have
a separate impact on sentencing decisions in the district court. To take account of this, we add text features to our treatment vector. The idea is that these embedded text features would represent some writing style characteristics that capture how judges reason toward sentencing decisions. We present two methods for representing textual features. First, we construct n-gram frequencies and reduce dimensionality using principal component analysis (PCA). Second, we use document embeddings.

**N-gram model with PCA**: Our first approach is to represent text using n-grams. An n-gram is a word sequence of length n. The n-gram model represents a text document with a collection of n-gram that appears in the text document.

There are multiple ways to featurize the n-gram representation into a numeric vector. One of the simplest ways would be to denote the presence of an n-gram using Boolean values of 0 and 1. Other simple ways include using the counts or frequencies of the n-gram. However, these methods come with well-known issues, such as not capturing the importance of an n-gram properly. Alternatively, we use term frequency-inverse document frequency (TF-IDF) to score each n-gram. The equations for calculating TF-IDF are shown in Eqs. (1)–(3):

\[
\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D),
\]

\[
\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},
\]

\[
\text{IDF}(t, D) = \log \left( \frac{N}{\text{count}(d \in D, t \in d)} \right),
\]

where \( t, d, \) and \( D \) denote the n-gram, the document, and the corpus that contains \( N \) documents. \( \text{TF}(t, d) \) measures how frequently the n-gram \( t \) occurs in current document \( d \), and \( \text{IDF}(t, D) \) measures how often the n-gram appears across all document \( d \) in the corpus \( D \) with the intuition that if an n-gram is common across many documents, then it is probably less informative about a particular document.

We convert the text documents into a TF-IDF matrix in Python and then apply PCA to reduce dimensionality and keep only the largest twenty-five principal components.

We also experiment with simple counts as scores for each n-gram. These counts featurize each document into a numeric vector. To compare the two methods, we use their resulting numeric representations of the document vectors to predict sentence length changes in an OLS regression model. We further experiment with using unigram (1-gram) or bigram (2-gram), in order to get better tradeoffs between representation power and computation cost. The result is shown in Table 2.
From the experiment, we saw that using TF-IDF is better than using simple counts. Furthermore, using bigrams gave better performance than using unigrams, which is intuitive. A further increase to 3-grams substantially increases computation burden, while the added benefit is slim. Therefore, for all the following experiments, we used bigrams with TF-IDF.

One of the limitations of this approach is the loss of information during dimensionality reduction. The information loss can be measured by the remaining explained variance of the selected principal components after PCA. We found that adding an additional principal component each time increased the explained variance by approximately 0.003, and even with 100 dimensions, the explained variance is just slightly above 10 percent of the total variance. This led us to seek a better method for representing text.

**Document Embeddings:** A better and more recent approach is to use document embeddings. Specifically, we used the Doc2Vec model proposed by Le and Mikolov (2014). Inspired by and similar to Word2Vec (Bengio et al. 2006; Collobert and Weston 2008; Mnih and Hinton 2008; Turian et al. 2010; Mikolov et al. 2013), Doc2Vec uses an unsupervised approach to learn feature representations from text. While the goal of Word2Vec methods is to learn representations for words, the goal of Doc2Vec is to learn representations for documents. Each document will be converted into a dense vector, where the distance between two vectors encodes the similarity between them. We trained our Doc2Vec model using a text corpus containing all cases’ opinion text and used it to generate document embeddings for each case’s opinion text. We used GenSim Doc2Vec implementation (Rehurek and Sojka 2010) with a context window of size 10 and generated a fixed size numeric vector of size 25 for
Figure 1 Projection of document embedding onto two-dimensional space. Each dot in the figure represents a case’s opinion text.

each case’s opinion text. We used t-SNE (Van der Maaten and Hinton 2008), a tool for visualizing high-dimensional data, to project the embedding vectors onto 2-D space. The scatter plot is shown in Figure 1.

An important property of this model is that the geometric location of the embedding vector in high-dimensional space encodes predictive information for the context-specific frequencies of words in the document. Intuitively, with Doc2Vec representation, similar cases’ opinion texts will be placed closer to each other in the embedding space. Le and Mikolov (2014) showed that the document vectors created with Doc2Vec outperformed other methods, including the popular bag-of-words model, for many natural language processing tasks. Figure 2 illustrates the idea of the Doc2Vec model.

3.2.3 Normalization and Splitting Data

Each circuit court has many judges, but three judges are randomly assigned to a case. We aggregate the characteristics of the three judges in each circuit court case. We normalize all columns based on mean and standard deviation. After that, we randomly split the data set into a training set, a validation set, and a test set.

The final data set has 7,388 cases as rows. Columns contain eighty-four different features for three different judges, twenty-five extracted text features from the case’s opinion, a binary column indicating appeal decision (affirm/reverse), and a target column indicating sentencing length changes.
Appendix A.2 contains a detailed description of the data set and features. The descriptive statistics are based on values after demeaning and before normalization.

4 Empirical Model

The statistical approach is mainly based on the two-stage Deep IV framework proposed by Hartford and colleagues (2017), which is a high-dimensional generalization of the reduced form causal analysis approach described by Angrist and colleagues (1996).

The Deep IV framework assumes the structural form shown in Eqs. (4) and (5) and defines the counterfactual prediction function as Eq. (6). The graphical model is illustrated in Figure 3.

\[ y = G(w,x) + e, \]  \hspace{1cm} (4)
\[ w = f(x,z,e), \]  \hspace{1cm} (5)
\[ h(w,x) := G(w,x) + E[e|x], \]  \hspace{1cm} (6)

Figure 3 Model illustration. Adapted from Hartford and colleagues (2017)