**MIT Computational Law Report •**

# Causal Inference with Legal Texts

**Jerrold Soh Tsin Howe**

**Published on:** Dec 07, 2021

**ABSTRACT**

Legal analysis requires inferring causality from legal texts. We often seek to identify, for instance, factors stated in judicial opinions that were dispositive of case outcome. Recent advances in causality theory and natural language processing could automate this process, further allowing causal legal questions to be examined more rigorously. This paper introduces causal text inference techniques to a generalist legal audience and illustrates how they apply to law. I consider legal use cases for text-as-outcome, treatment, and de-confounding methods, provide a detailed survey on text de-confounding methods, and identify a fundamental problem with inferring causal effects from judicial opinions. I further illustrate how causal text inference may be applied to a novel dataset of 7,046 Supreme Court certiorari petition briefs to examine whether petition origin causally affects certiorari grant rates. I show how covariate balancing on petition texts yield differing causal estimates depending on text embedding, balancing scheme, and estimation techniques used, suggesting that failing to account for text may yield spurious results.

*I thank Jim Greiner and Kevin Quinn for helpful advice, Ferrell Chee, David Costigan and Lily Lilliott for research support, and the United States Supreme Court for making petition briefs after 2017 freely accessible.*

---

## 1.0 Introduction

Some of the greatest legal debates center on causal inference. If indeed law is nothing more than “prophecies” of what courts will do in fact,[1] then lawyers should presumably be interested in the factors that truly, causally determine, rather than superficially correlate to, judicial decisions. Likewise, the debate between realism and formalism might be understood as a disagreement on whether predictive but purely associative factors can properly be termed ‘law’, as well as whether certain outcome-causal factors cannot be.[2]

Common law further assumes it possible for lawyers to extract these outcome-causal factors by reading preceding judicial opinions. Were opinions express and complete statements of all such factors, this would be a trivial comprehension exercise. But if expressivity and completeness characterized most opinions,[3] the attention (and financial reward) attributed to artful legal analysis would hardly be justified. Law schools devote significant attention, particularly in first year curricula, to teaching

students to analyze what opinions say. Legal analysis may thus be thought of, at least partly, as an exercise in drawing causal inferences from legal texts.

The law's persistent unfamiliarity with causality theory[4] has thus been described as both surprising and unhealthy.[5] Much of this criticism concerns the use of statistical techniques, particularly regression methods, to answer causal questions in the *civil rights context* (i.e. did the impugned law *cause* the disparate impact observed?).[6] Others have considered the implications of causality theory the assessment of epidemiological evidence in courtrooms, the success of criminal justice intervention programs, and the effectiveness of access to justice initiatives.[7] However, as the critique concerns *methodology*, it readily applies to the use of statistical techniques in legal studies generally.[8] Meanwhile, causal inference theory in the hard and social sciences has, in recent decades, undergone significant transformation.[9] The law's lagging adoption of modern causality theory, therefore, seems poised to widen.

Seeking to bridge this gap, this paper explores the implications of causality theory for answering causal questions in *legal analysis.* To clarify, I am not referring to the legal doctrine of causation which generally concerns whether the defendant's breach of duty is a "legal cause" of the plaintiff's injury.[10] Rather, I am interested in a recurrent question in legal analysis that can be expressed with the template: does a given factor T causally affect a legal outcome Y?

My interest here is further motivated by recent advances in empirical techniques for drawing causal inference from *text* data.[11] These techniques, which I refer to below as "causal text methods", allow causal effects to be identified in settings where text data is to be included as a key variable. To the extent that extracting causal effects from text is central to legal analysis, causal text methods have profound implications for lawyers.

Part 2 begins with a theoretical primer on causal inference theory. Part 3 provides background on causal text methods, in the process considering illustrative legal questions that each method may apply to. I focus on text de-confounding, a technique that attempts to hold texts constant across treatment and control settings, as that appears to have the broadest potential for legal application. In Part 4, I discuss challenges with applying text methods to the legal domain specifically and touch briefly on an important statistical limitation with inferring causality from *opinions*. Part 5 demonstrates how causal text methods might be applied to study how the Supreme Court grants certiorari and presents illustrative, albeit preliminary, findings on whether case origin matters.

## 2.0 Primer on Causal Inference

This Part covers the foundations of causal inference theory. Readers already familiar may wish to skip ahead.

Consider the following questions common in legal scholarship and practice:

1. Does law and economics training for judges affect how they decide cases? [12]
2. How does bench composition affect judicial decisions?[13]
3. What determines whether my client will be liable for injuries arising from a traffic accident in which they were involved?
4. Would pleading guilty reduce my client's sentence?
5. Would a longer or better written brief increase my client's odds?[14]

These questions can be reduced to or restated as questions of causality. We want to know whether a *different* pre-existing state of affairs leads to a different outcome and, if it does, how different that outcome might be. Further, we are often interested in isolating the outcome effect of one particular attribute of that pre-existing world. This leads to the following question: would a pre-existing world with that attribute yield a different outcome from one without it?[15]

### 2.1 The Potential Outcomes Framework

Causality theorists refer to the attribute of interest as the *treatment* variable, commonly denoted T, and to the outcome of interest as the outcome variable, denoted Y.[16] To illustrate, the 'treatment' for Question (4) above is pleading guilty while the outcome is sentence. For simplicity, assume the client can only choose whether to plead guilty or not (*i.e.* there are no partial choices like pleading guilty to *some* charges or an intermediate bargain), and the only possible sentence is a jail term ranging from zero to three months. We can express the treatment as a binary: T=1 if the client pleads guilty, and T=0 otherwise.[17] Y is then a continuous variable ranging between 0 to 3 months. Accordingly, Y(T=1) refers to the outcome jail term when the client pleads guilty, and Y(T=0) the jail term when the client does not.

It appears, then, that we may simply take the difference between Y(T=1) and Y(T=0) as the causal effect of T on Y. This, indeed, is the *definition* of the causal effect of T for that particular client (referred to as a single "unit" of analysis or one "observation"). In addition to this *unit* treatment effect, we might also be interested in the *average treatment effect* ("ATE") of a given policy on groups of individuals, derived by subtracting the group *average* Y(T=0) from the *average* Y(T=1).

In practice, however, this cannot be done. It is *physically* impossible to observe both Y(T=1) and Y(T=0) at the same time. Unlike Schrodinger's cat, the client may not at once plead both guilty and not guilty. Holland famously called this the Fundamental Problem of Causal Inference: for a given unit, we can only see either the treated or non-treated outcome, never both.[18] This reveals that causality is fundamentally, and inevitably, a *missing data problem*. To infer causality, we need a mechanism that reliably provides information on what the unobserved outcomes would potentially be if the treatment status changed.[19]

This insight underlies Imbens and Rubin's influential "potential outcomes" framework.[20] To work around the impossibility of observing both outcomes for one unit, we use observations on *multiple* units, some for whom we observe Y(T=1) (the "treatment group"), and others for whom we observe Y(T=0) (the "control group").[21] If the two groups are, loosely speaking, *comparable*, then observed outcomes for one can be used to 'fill in' missing potential outcomes for the other. That is, to estimate the control group's potential outcome *had they received the treatment*, and vice-versa. The control (treatment) outcomes for the treatment (control) group are conventionally denoted with $Y_T(T=0)$ and $Y_C(T=1)$ respectively, with the subscripts identifying which *observed* group the potential outcomes apply to. The comparability requirement can then be expressed as requiring $Y_C(T=0)$, which we observe, to be representative of $Y_T(T=0)$, which we do not, and vice-versa for $Y_T(T=1)$ and $Y_C(T=1)$.

What I loosely termed 'comparability' and 'representativeness' is formalized in causality theory under two assumptions, both of which must be fulfilled (or reasonably assumed) for legitimate causal inference: (1) the Stable Unit Treatment Value Assumption ("SUTVA"), and (2) unconfoundedness.[22] SUTVA requires both that (A) potential outcomes for any unit do not vary with treatment assignment to other units, and (B) there is only one form of the treatment and control.[23] Requirement (A) precludes interference across treatment and control groups. To illustrate, suppose two defendants $D_T$ and $D_C$ were charged for an identical crime. $D_T$ pleads guilty and gets 1 month in prison. $D_C$ does not and gets 3. We might expect the two months' difference in sentence to be the 'benefit' of $D_T$ pleading guilty. But if $D_T$ and $D_C$ were, in fact, partners in the same crime (which would explain the identical charges), then $D_T$'s guilty plea might have contributed to $D_C$'s longer sentence, and 2 months would probably overestimate the causal effect of $D_T$'s plea. Requirement (B) mirrors the earlier simplifying assumption that there are no 'intermediate' levels of pleading guilty.

If so, we would be wrongly comparing units which experienced different levels of treatment as if they had or had not received the same treatment.

While SUTVA relates primarily to the treatment variable T, unconfoundedness relates to all others (though we will see that it retraces to T). In taking the difference between Y(T=1) with Y(T=0), the two pre-existing worlds we are really interested in comparing are the worlds where T=1 and where T=0. In prior exposition I have implicitly assumed that all other attributes of those worlds remained constant, allowing us to isolate T's effect. If we compared worlds where T and another variable X both differ, we cannot tell if any resultant change in outcome is due to the variation in T or in X, *unless* we know *a priori* that X does *not* affect the outcome. This leads us to a definition of "confounders" (*i.e.* variables that confound identification of the causal effect of T on Y) that those trained in regression methods may find familiar: a confounder is a variable which varies with *both* T and Y.[24] Confounders must be held constant across the compared worlds for legitimate causal inference.

Denoting all confounders in a given pre-existing world as Z and their values as z, what we want, therefore, is to compare Y(T=1, Z=z) with Y(T=0, Z=z). Observe that it is unnecessary to hold constant every factor in the universe. Causal inference could be legitimate as long as *treatment assignment* is independent of potential outcomes. This is termed the "unconfoundedness" assumption,[25] and would explain why the Randomized Controlled Experiment ("RCT"), where treatment is randomly assigned to all units, is the gold standard for causal inference.[26] Random treatment assignment lets us safely expect confounders to be evenly distributed across treatment and control groups, such that they *effectively* do not differ across the worlds we want to compare. To be sure, this is *not* guaranteed, particularly for small sample sizes. Even if we randomly required either $D_T$ or $D_C$ to plead guilty, they would probably still differ in sentence-material aspects such as age.

## 2.2 Types of Causal Questions and How to Answer Them

"Treatment" connotes how causality requires performing some *intervention* to create the alternative, counterfactual world used for comparison. Pearl likens this to a "surgical procedure" in which we carefully intervene to set the treatment variable equal to its counterfactual value, leaving all else equal.[27] This indeed underpins the familiar *legal* doctrine of 'but for' causation where we examine if, but for the defendant's breach of duty, the damage would still have occurred.[28]

Legal causation is, in turn, one *type* of causal question that can be asked. Pearl and MacKenzie enumerate three levels of the so-called "Ladder of Causation".[29] At the bottom are questions of *association,* which relate to *observation.* Here, we are interested in informing our beliefs on Y upon *seeing* that T is true or otherwise not true. The second level comprises questions of *intervention,* which relate to *doing*. We want to know how *doing* or *not doing* T affects Y. Atop the ladder sits questions of *counterfactuals,* which relate to *imagination.* We want to know what Y would have been had T been different, and this can only be achieved in a hypothetical, alternative history. 'But for' causation, evidently, is a counterfactual question.

Identifying where a causal question sits on Pearl's Ladder is crucial because each level necessitates different approaches and tools. Most significantly, questions at a given level require inputs at or above that level.[30] For instance, interventional questions minimally require interventional input, but can also be answered with counterfactual input (though empirically observing counterfactuals is, as the Fundamental Problem of Causality reminds us, impossible).

To illustrate the subtle difference between levels one and two, consider the running question of whether pleading guilty reduces a client's sentence. This is an *interventional* question because we want to know how *doing* the act of pleading guilty affects sentence. But lawyers seem to treat this as an *associational* question. The Common lawyer's strategy for answering it would probably be to parse precedents for how sentence length historically varied across guilty pleas.[31] Strictly speaking, however, such observational data alone can only indicate whether we should expect different sentence outcomes when we *observe* (*i.e.* we are told that) a given defendant pleaded/did not plead guilty. Whether pleading guilty reduces a present client's sentence is different "because it involves not just seeing but changing what is".[32] The reason for this difference has, in fact, been explained when we considered confounders. There may have been some other factor X in the precedents parsed that (1) also varied with pleading guilty and (2) was causing differences in sentence. If so, our present client pleading guilty may not *cause* a reduction in sentence.[33]

Of course, to say that lawyers merely measure empirical correlations in precedents caricatures the legal method as it overlooks the crucial and complementary role of qualitative, doctrinal analysis. Most lawyers, or at least those worth their billables, know that precedents relied on must be *comparable* to the client's case and, by transitivity, to each other. Doctrinal analysis further identifies variables which are *a priori* unlikely to affect the outcome (that is, confound). To illustrate, contract doctrine

might suggest that two precedents interpreting an identical clause offers safe comparison, even if the disputes are set in different states, *provided* the two have similar contract interpretation doctrines. But it is probably unwise to compare two precedents interpreting an identical state constitution article if they come from different states.

In the vocabulary of causal inference, the legal method can be seen as an attempt to test observational legal data against what causality theorists often refer to as a "causal model".[34] Causal models are hypothesized relations, specified on the basis of prior domain knowledge, between the outcome, treatment, and other variables we are studying.[35] Such models are the critical interventional inputs that allow us to identify interventional causal effects, where possible.[36] If our causal model hypothesizes that T does affect Y but its effect is confounded by a single variable X and nothing else, then data on T, Y, and X is necessary but also sufficient for deriving causality. If we did not have data on X (or some variable which adequately proxies for it), or if X is impossible to measure or observe, then we also know it is impossible to identify causality, regardless of how sophisticated our statistical arsenal may get.

## 3.0 Causal Inference With Text

Once legal analysis is understood, at least partly, as an exercise in causal reasoning with legal texts, the promise that causal text methods hold for legal analysis becomes clear. Causal text inference extends an established, but growing, body of computational social science research that treats text as a form of quantitative *data*.[37] Such "text-as-data" techniques draw on fields such as corpus linguistics and the digital humanities.[38] There is a growing body of scholarship applying these techniques to law, [39] though to my knowledge none have directly considered causal *legal* text inference.[40]

Put simply, causal text methods 'plug-in' the text as one of the three key variables in the causal framework: outcome Y, treatment T, or confounder(s) Z. For instance, we may understand the guilty plea question above as case of text confounding: we want to know the causal effect of pleading guilty on sentence, but are concerned that other factors in preceding cases, as written in judicial opinions, confounds identification. Since opinion texts *are* observed, we could simply control for them by providing them as inputs into the statistical model. Conceptually, we are then examining how sentence lengths differ on two precedents with identical judgments, except the defendant in only one of them pleaded guilty.

Things are, of course, not so simple. While straightforward at a broad conceptual level, causal text inference quickly runs into theoretical and practical obstacles, both generally and when applied to law specifically.[41] Foremost, we cannot mathematically operate on text. The first step must therefore be to convert text to numbers or, in math-speak, to "encode" or "embed" text into a numerical space. Yet, as we shall see, the choice of text representation has significant implications for downstream inference because information in the text could be lost or mutated in translation.[42] Further, text-as-data does not obviate, and may in fact exacerbate, the pitfalls of statistical inference.[43]

In this light, this Part introduces causal text methods generally and highlights potential legal applications. I focus on text de-confounding because it arguably has the greatest scope for legal application.

## 3.1 The Codebook Function

Text-as-data first requires converting text to data. Methods for doing so are canonically known as the "codebook function", commonly denoted $g$, which can be broadly conceived of as an arbitrary algorithm that takes text as input and produces numbers as output.[44] $g$ derives its name from the physical books that empirical researchers use to specify how data should be encoded. Trusty research assistants hired to parse empirical data out of legal judgments are thus human examples of codebook functions. Manual codebooks are, of course, nothing new. But they are expensive, especially if legal experts (though perhaps not law students) must be hired to do it.

One of text-as-data's primary contributions lie in exploring how *automatic* content coding can, perhaps imperfectly, substitute for laborious human effort.[45] These techniques typically originate from the information and computer sciences, particularly a branch thereof known as natural language processing ("NLP"). Because computers can only perform numerical operations, NLP has developed (independently of causal inference) an array of methods for representing text as numbers. Numerical text representations are broadly known as "text embeddings".[46] This sub-Part provides background on automatic content coding to facilitate subsequent discussion of causal text methods generally. Readers familiar with this may skip ahead to Part 3.2.

### 3.1.1 Text as a Bag-of-Words

An illustrative and widely-used method for embedding documents are a family of encoding techniques built on the Bag-of-Words ("BOW") model of text which, as its

name suggests, conceives of documents simply as bags of words and/or phrases.[47] For example, a document with only the words “this is a document” is seen as nothing more than the sum of its parts: “this”, “is”, “a”, and “document”. Documents can then be converted, say, into binary variables that take the value 1 if a given word appears in the document and 0 otherwise. Table 1 illustrates:

TABLE 1. EXAMPLE BAG-OF-WORDS ENCODINGS

| Document Text | A | Another | Document | Is | Longer | That | This |
|---|---|---|---|---|---|---|---|
| Technique I: Binary Encoding | | | | | | | |
| This is a document. | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| This is another document. | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| This is a document that is longer. | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| Technique II: Term Frequency Encoding | | | | | | | |
| This is a document. | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| This is another document. | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| This is a document that is longer. | 1 | 0 | 1 | 2 | 1 | 1 | 1 |

*Table 1. Example Bag-of-Words Encodings*

This produces “document-term matrices”, or tables of numbers capturing the preponderance of each *term* in each *document*. Notice that such matrices tend to be extremely high-dimensional (there are many columns), yet sparse (many cells have ‘0’ values). Every word in the corpus requires one column to itself, even if it only appears once in one document. Further, words that appear in *every* document produce columns

with little variation, limiting their utility for downstream analysis. Both high dimensionality and sparsity hinder statistical inference.[48]

It is therefore standard NLP practice to, amongst other things, remove non-informative words (called stop words) from the corpus before embedding it. Standard stop word lists typically include what lawyers may call "glue words"[49] — syntactic sugar such as "the", "of", and "and".[50] Further, term frequency scores are often normalized by *document* frequency scores, capturing the intuition that documents are best characterized by words common *within* them, but rare in others.[51] Such a technique is known as the Term Frequency/Inverse Document Frequency algorithm, or "TFIDF".[52]

### 3.1.2 Topic Models

To further reduce dimensionality, statistical algorithms are often applied to compress document-term matrices into document-*topic* matrices. Consider an arbitrary compression algorithm that attempts to optimally compress the document-term matrices above from 7 to 3 columns. To minimize information loss (the hallmark of optimal compression), we might expect the algorithm to group similar words together. For simplicity, let us leave aside the technical complexity of defining what "similarity" means, and assume the algorithm understands language like we do. It might yield the following document-*topic* matrix:

| | Topic 1: This, That, Is | Topic 2: A, Another | Topic 3: Document, Longer |
|---|---|---|---|
| This is a document. | 1 | 0.5 | 0.5 |
| This is another document. | 1 | 0.5 | 0.5 |
| This is a document that is longer. | 1.5 | 0.5 | 1 |

*Table 2. Example Document-Topic Matrix*

Each column now represents a cluster of words, instead of just one. These word clusters are known in NLP literature as "topics". The entire algorithm is accordingly a "topic model". Notice that, in this stylized example, document 1's embedding is now indistinguishable from document 2's, while the two documents are distinguishable from document 3 on dimensions (*i.e.* topics) 1 and 3. These numbers were, of course, cherry-picked to illustrate that the compression, though not lossless, could nonetheless yield informative signals.

Importantly, topic models also report the numerical contribution of each term in the vocabulary to that topic. Such *topic-term* distributions may look as follows:

| Topic | A | Another | Document | Is | Longer | That | This |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0.8 | 0 | 0.9 | 1 |
| 2 | 0.9 | 0.8 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0.6 | 0 | 0.5 | 0 | 0 |

*Table 3. Example Topic-Term Distribution*

The above numbers are arbitrary. The mathematics behind how they are derived are rather involved and will not be presented here. In any event, going into details here would be unhelpful because the computations differ depending on the specific topic model deployed. An entire family of topic models exist that start from different document-term matrices and use different compression algorithms.[53] The discussion below will assume only general knowledge of topic models as algorithms that transform text inputs into (1) document-topic matrices reflecting the preponderance of a given 'topic' across the corpus, and (2) topic-term distributions reflecting how much a given term characterizes a given topic. I explain further details on topic models where appropriate. Part V will also illustrate the results of Latent Semantic Analysis ("LSA"), a classic topic model, on an actual legal corpus.[54] The main point here is that both naïve encoding techniques and the more sophisticated topic models are possible *automated* codebooks.[55]

### 3.1.3 Word Embeddings and Language Models

Because more sophisticated codebook functions have been used in causal text methods, brief mention must also be made here to two additional techniques. First, word *vector* algorithms[56] embed individual words into a high dimensional space by statistically modelling collocations between them.[57] Famously, word vectors were shown to permit semantically logical arithmetic: subtracting the vector for "man" from "king" and adding the result to "woman" produces "queen".[58] Once each document word is embedded, computing an overall *document* embedding can be as simple as taking the sum or average across all words though, once again, more sophisticated techniques exist.[59]

Another recent breakthrough involves embedding documents by training neural networks to predict arbitrarily-removed words in a document, an activity reminiscent of the cloze passages that children use to learn languages.[60] These "language models" have quickly become the new state-of-the-art in NLP, rewriting performance benchmarks across diverse tasks including summarization, translation, and question answering.[61] The cloze task (and other downstream optimizations) forces the neural network to create informative text representations within its internal layers.[62] In particular, studies on the "Bidirectional Encoder Representations from Transformers" model, an influential language model often called "BERT", note that its internal layers seem to capture syntactic properties of language.[63] The trained neural network can then be used to encode text.

As with topic models, word vectors and language models will not be explained here in detail. It suffices to note that these are likewise possible automated codebooks. Recall that a codebook function is, broadly speaking, anything which translates texts to computable numbers. Further, different codebook algorithms make different assumptions about the text they process, thus raising different technical considerations.[64] The next few sub-Parts explain how encoded text can be used as outcome, treatment, and finally de-confounding variables.

## 3.2 Text-as-Outcome

To adapt Question (1) in Part 1, we might want to know how law and economics training affects judicial *writing*. A simple approach to answering this would be to narrow down the question as follows: do they start using terms like "efficiency", and "transaction costs" more in their judgments? We might then represent judgments entirely by how frequently each term of interest occurs. Separate statistical models can then be used to test for significant changes in term use before/after training. A more sophisticated approach might compute the total difference across all these terms at once by performing linear algebra on the entire set of term frequencies.[65]

The example above provides an intuitive illustration of treating text-as-outcome. Notice, however, that it forces us to answer a narrower question which might not map perfectly onto our true question of interest. Moreover, results in the given example would be sensitive to the terms we choose to look for. This may work for law and economics, which has a clearly defined set of technical terms, but might not generalize to other contexts.

Egami et al. use an example with legal flavor to demonstrate a more general text-as-outcome framework.[66] They adapt Cohen et al.'s study of whether descriptions of past criminal behavior affects others' evaluations of how far they should be punished for a subsequent crime.[67] Experimental subjects were asked to read descriptions of an accused's crime. The control group was only given the crime description. The treatment group was given an identical crime description, followed by a description of prior crimes. Both were then asked (1) whether the accused should be jailed, and (2) to describe in at least two sentences why. The response texts generated by the latter question were taken as the outcome of interest and encoded with a topic model.[68] To illustrate, one of the topics derived was characterized by words like "deport", "think", "prison", "crime", and "imprison". Egami et al. interpreted this as a topic about deportation, so that a high preponderance of this topic in a response *suggested* the subject considered deportation as a suitable punishment. Other topics were interpreted to concern other punishments like incarceration.[69] Egami et al. then computed the ATE as the average difference in topic scores across the treatment and control groups and found that including criminal histories "significantly increases the likelihood that the respondent advocates for more severe punishment or deportation".[70]

## 3.3 Text as Treatment

Notice that the previous experiment was also an example of text-as-treatment: treatment group subjects were given a different text prompt from control group subjects. In that experiment, we were solely interested in *one* aspect of those prompts – whether it described the defendant's antecedents. A trivial codebook was implicitly used: we code the text 1 if it contains such a description, and 0 otherwise. In the law and economics example, we encoded judgments using the frequencies of specified law and economics terms.

Text-as-treatment has other legal applications. For instance, we might be interested in the effect of legal briefs on judicial decisions. Long and Christensen study whether more readable briefs improve the chances of winning an appeal.[71] As a codebook, they use the Flesch-Kincaid readability score as well as other indexes developed in linguistics. The effect of a statute or case law on some legal outcome can also be fit into a text-as-treatment approach, since all laws are written in text.[72] Imagine an experiment where subjects are randomly assigned different expressions of the same proposed law, and asked to evaluate them for fairness, and/or certainty.

Manual codebooks constrain us, however, if we are interested in multiple aspects of the treatment text. Further, the text features we want to investigate might be latent or multi-dimensional, such that they cannot be easily reduced into a defined set of word binaries or linguistic scores. That is, we might be interested in whether "the text" affects the outcome, but are unwilling or unable to reduce it to a set of known, specifiable attributes like readability. Automated codebooks are useful here because they may allow us to pinpoint multiple, latent treatment features. Egami et al. demonstrate this on a dataset likewise of legal interest.[73] They investigate whether the text of consumer complaints on financial products submitted to the Consumer Financial Protection Bureau affect how promptly the respondent business responds (Egami et al. do not distinguish between positive and negative responses). As above, the codebook outputs a set of topics that the authors then manually interpret as denoting complaints about "loans", "debt collection", "mortgage", "detailed complaint", and so on. They find that more detailed complaints, and complaints relating to loans in particular, receive prompter responses.[74]

An important premise with automatically discovering latent text treatment features is that, on top of SUTVA and unconfoundedness, we must further assume that the codebook is *sufficient*.[75] Formally, sufficiency requires that the codebook captures enough outcome-relevant information in the text that such any information left out is *orthogonal* to that representation. Simply put, the codebook must *not* leave out any text information capable of confounding the causal effects of the text features that *were* captured. Thus, the ability of the chosen codebook to capture text information becomes pivotal. This explains my earlier exposition on the more sophisticated word vectors and language models.

## 3.4 Text De-confounding

Text de-confounding has arguably the most potential application in law because it lets us ask questions of the form, "holding these texts constant, how does a given treatment T affect a given outcome Y?" A wide range of legal questions adhere to this template, including arguably the central question in case law analysis: holding the "facts" section of two judgments constant, how does a given legal factor, say, the accused pleading guilty, affect case dispositions?[76] Questions on judicial behavior may also be asked: given two similar case briefs, would judge 1 have decided differently from judge 2? Likewise, given two similar case descriptions, would lawyer 1's rather than lawyer 2's involvement have led to a different case outcome? We might also be

interested in legal-systemic questions, such as whether the originating lower court affects whether the Supreme Court grants certiorari, holding petition briefs constant.[77]

Given its wide applicability, the second half of this paper focuses primarily on text de-confounding. In this sub-Part, I delve into some detail on text de-confounding to establish the necessary background context. As text de-confounding implicates techniques I have yet to introduce, a brief detour into de-confounding itself is necessary.

### 3.4.1 A Primer on De-confounding

As mentioned in Part 2, the essence of de-confounding is ensuring comparability between two alternate pre-existing worlds. Technically speaking, what we are after is *covariate balance*: all "covariates" (*i.e.* non-treatment, non-outcome variables) should be similarly distributed in both the treatment and control groups.[78] In an RCT, random treatment assignment lets us expect this. With observational data, however, covariate balance is rarely guaranteed.[79] Covariate imbalance hinders causal inference because results from such data are sensitive to the specific statistical estimation techniques chosen and often imprecise.[80]

There are two general strategies to de-confounding. First, we can limit the data we use to regions where the treatment and control groups are most comparable. To recall the running example, to know whether pleading guilty reduces sentence, we might want to study only precedents involving defendants of the same gender as the client. This, of course, requires that we know, or at least have reasons to suspect, that defendants of one gender might be sentenced differently from defendants of another. The logical conclusion of this is to limit the dataset to precedents that are identical *on all non-treatment fronts.* Specifically, for every precedent in the control group, we look for an identical match in the treatment group. Put differently, we would be trying to compare legal *twins.*

As with biological twins, however, such perfectly-matched precedents are probably rare in legal practice. As a second-best solution, we might settle for finding the *closest available* correspondent in the opposing group, perhaps provided it meets a minimum closeness threshold. This then requires a measure of how *close* an observation is to another. In law, we tend to measure precedent closeness by zooming in on key legal and doctrinal factors — we look for cases with similar questions presented, issues raised, and so on.

The intuition that some factors matter more than others carries into causal inference methods. When attempting to match observations, we are most concerned about covariate balance. Closeness is therefore logically centered on the factors which vary most across treatment and control groups. Notice that this is equivalent to asking which factors best predict whether a unit receives the treatment (itself a level one causality question). To illustrate, if all, and *only*, male defendants pleaded guilty, then knowing whether the defendant was male would let us perfectly predict treatment status.

Causal inference theory actualizes this intuition in a technique known as propensity score matching ("PSM"). A statistical model is used to predict the probability that a unit receives treatment given the covariates in the dataset, being the "propensity score". By construction, the propensity score proxies for similarity across the treatment and control groups, weighting imbalanced covariates more highly. Units from different groups but with similar propensity scores may then be matched as quasi-legal twins.

Other than propensity score, it is also possible to simply take the distance between the covariates using one of many possible mathematical formulas for calculating the difference between two *sets* of numbers.[81] Armed with a closeness measure, we can then match observations that are close enough, or trim away observations too different from the rest, or both.[82]

The second de-confounding strategy applies after the dataset has been fixed (and preferably balanced), and would be familiar to those trained in regression methods. Specifically, it is to include potentially-confounding variables into the estimation model as so-called "controls" so that the model "adjusts" for those variables.[83] There are limitations to this approach, however.[84] Foremost, which variables to include or exclude falls entirely within the analyst's discretion. Assuming the analyst has specified the right causal model, the necessary controls to include are self-evident. But that provides little reassurance: if the analyst already knew the right causal model, there would be little point to studying the dataset; if the analyst did not know the right causal model, then she might simply have specified it correctly by luck. The more realistic expectation is for the analyst, whom we assumed conscientiously studied prior literature, to have specified a causal model which approximates, but does not perfectly reflect, the truth.

Including the *wrong* variables as controls may, in fact, worsen results.[85] A classic example is controlling for mediators.[86] Suppose T does causally change Y, but

primarily by first increasing a third variable M (short for “mediator”). To illustrate, suppose pleading guilty signals remorse, and remorse in turn lowers sentence outcomes as much as, if not more than, the guilty plea itself. Empirically, we would notice that M varies with *both* T and Y: setting T=1 causes M to increase; increasing M decreases Y. If M were included as a control, so that the model tried to vary T while holding M constant, we would end up underestimating the causal effect that T has on Y *through* M.[87]

Note that the two de-confounding techniques are complementary. Balancing the dataset reduces the need to include model controls; controls can address residual confounding that balancing might have missed.

### 3.4.2 A Review of Text De-confounding Methods

Leading techniques for text de-confounding essentially revolve around matching encoded texts.[88] An illustrative baseline procedure is what Roberts et al. call Topically Coarsened Exact Matching (“TCEM”), which in essence encodes text using a topic model before applying so-called coarsened exact matching on the topics. Coarsened exact matching (“CEM”) is a matching technique where the analyst manually reduces the granularity of certain covariates by specifying cut points for variables so that observations can be matched within the newly-cut categories rather than on specific numerical values. For example, instead of finding two students with the same raw exam scores, we might cut them into three categories based on percentiles: top 25%, middle 50%, and bottom 25%. CEM then attempts to find at least one pair of treatment/control units per category. Units in unpaired categories (which implies they are themselves unpaired) are then trimmed.

Roberts et al. further develop another technique, Topical Inverse Regression Matching (“TIRM”), which encodes text with a specialized topic model capable of taking the treatment variable as input. This allows the resulting topic-term distributions to accord greater weight to terms that predict treatment assignment, and for the model to thus produce a measure analogous to propensity scores. Roberts et al. then suggest using coarsened exact matching on both the propensity score analog and the document topic scores such that the treatment/control groups are matched on *both* textual content and treatment propensity. The authors test both TIRM and TCEM on three partially simulated datasets and find that they varyingly perform better/worse at improving covariate balance, depending on the metric used to measure balance.

Mozer et al. build on Roberts et al. above to propose a "general framework for constructing and evaluating text-matching methods". They first note that because text representations and closeness measures chosen might affect downstream inferences, these choices should be made in light of the specific confounders to be targeted in a given research question.[89] In this light, they consider a wider range of representations and measures.[90] They show that, depending on the dataset and causal question asked, simply matching on the document-term matrix itself might yield a better-matched sample.[91]

Mozer et al. then demonstrate how text matching can improve covariate balance in the context of an observational study using a *medical* setting. The causal question was whether a particular diagnostic heart scan improved adult patient survival rates for patients in critical care with sepsis. The texts were medical notes taken by hospital staff upon the patient's admission to critical care. Noting that the ideal text de-confounding scheme would "match documents on key medical concepts and prognostic factors that could both impact the choice of using [the heart scan] and the outcome",[92] they preferred matching on the basic document term matrix because this representation would "retain as much information in the text as possible".[93] This strategy produced a better covariate balance than propensity score matching on non-text variables alone.[94]

Veitch et al. recently proposed an even more computationally-sophisticated method for text de-confounding.[95] Of course, computational sophistication does not itself make a technique worth mentioning. I raise this primarily to highlight how text de-confounding is receiving attention across multiple disciplines and has much scope for development. As the technique is highly mathematically involved, I will only briefly describe the intuition. In essence, the authors extend the BERT language model introduced above to simultaneously produce (1) document embeddings, (2) propensity scores, and (3) predicted outcomes based on the embeddings. They then use these outputs to compute the ATE.[96] One key distinguishing factor from prior work is that, Veitch et al. encode the text with BERT instead of a topic model. They thus note that their approach "replace[s] the assumption that the topics capture confounding with the assumption that an embedding method [i.e. theirs] can effectively extract predictive information".[97]

Next, using BERT embeddings as a starting point, they alter the embeddings further by fitting them to the particular dataset they want to study, step known in NLP as 'fine-tuning'.[98] In this step, they task the embeddings model with predicting *both* the

treatment and outcome variables. To see why, recall that the original BERT is tasked with predicting missing cloze words, and thus develops an internal representation of general language properties. Since confounders are variables that vary across (i.e. predict) both treatment and outcome, when the model is given such an objective, it likewise begins internally representing confounding information in the dataset.

While these document embeddings could, like document-topic scores, then be piped through some matching algorithm, Veitch et al. simply use the same model to estimate causal effects. This is both logical and convenient since the model has already learnt to predict both treatment (*i.e.* what we do when estimating propensity scores) and outcome (*i.e.* what we do in the subsequent inference step).[99]

Given the expanse of algorithms and techniques covered in a relatively short exposition, it is useful to crystallize key themes in the text de-confounding literature. The central goal of text de-confounding is to keep the treatment and control groups comparable by exploiting information in relevant texts. This, to be sure, conceals much of the complexity behind defining what "comparable" is, especially when *texts* are involved. Because text is multi-faceted and multi-dimensional, applying traditional de-confounding strategies like PSM is difficult.[100] Obstacles begin to arise even when choosing a text representation (i.e. the codebook function). Though numerous methods of varying statistical sophistication have been proposed, there are as yet no clearly superior or inferior methods, only methods more or less suited to the particular causal questions we want to study and their hypothesized confounders. In the next Part, I consider challenges that may arise when applying text de-confounding to the legal setting.

## 4.0 De-confounding with Legal Text

As alluded to in Part 3.4, a wide inventory of legal questions might be answered with text de-confounding. This should not be over-stated. The legal setting raises unique challenges *on top of* those already identified above by general (text) de-confounding literature. This Part highlights this by considering how de-confounding might be used to extract what Falakmasir and Ashley call "legal factors", or "stereotypical patterns of fact that tend to strengthen or weaken a sides' argument in a legal claim".[101] This question is instructive because the corpus of case precedents is arguably the text corpus that Common lawyers work with most frequently.

## 4.1 A Fundamental Problem of Legal Analysis

Consider how we typically reason with precedent. First, we assemble a set of cases similar to the client's, trimming away those that differ on legally-material grounds. We then read the judgments, paying attention to each case's facts, procedure, dispositions, and *ratio decidendi*. If there is no express *ratio* covering the client's case (no doubt a common occurrence), we fall back to reasoning by analogy. The standard argument flows as follows:

> *In case A, legal factor T was present, and the defendant was held liable. This may be contrasted with case B, which is in all material respects similar to case B, except factor T was absent, and the defendant was not liable. Therefore, factor T is decisive for liability. Since T is absent in my client's case, he/she is probably not liable.*

Such reasoning adopts closely the language of causality.[102] A control unit (case B) is compared against a treatment unit (case A), and the difference in outcome indicates the treatment variable's outcome effect. But because precedents are *observational* inputs, where our intended causal question sits on the Ladder of Causation then determines the extent we can answer it.[103] If we are merely interested in predicting likely outcomes for the client based on what we *see* in the cases, this would likewise be an observational question that precedents themselves are sufficient to answer.

But if we want to advise the client on what to *do*, we would be interested in interventional causality, a question that requires interventional data to answer. While RCTs are a first-best solution to generate such data, RCTs on legal factors are practically impossible.[104] The challenge, therefore, is how we might *approximate* interventional data with observational precedents. This is a legal analog to Rubin's Fundamental Problem of Causal Inference. Just as we cannot observe both treatment and control outcomes for the same unit of analysis, we cannot observe both treatment and control outcomes for the same precedent. Nor can we randomize treatment assignment to simulate having observed it.

To overcome this challenge, we must first specify a causal model to test our observational data against.[105] This can be obtained from qualitative analysis of the cases. If we further want to derive statistical estimates of the outcome effects, however, then recourse to text de-confounding techniques is likely necessary. Put simply, we want to the data look *as if* we had randomly assigned legal factors independent of confounding case attributes.

With text de-confounding in mind, we might state the question as such: “holding judgment texts constant, how does varying some legal factor T affect some legal outcome Y?” This conveniently substitutes “potentially-confounding case characteristics” with the judgment texts we are used to working with. The next steps for causal inference seem clear: encode the judgments, match the cases, and calculate ATEs.

## 4.2 The Problem with De-confounding on Judgments

But this approach rests on the shaky assumption that judicial opinions are sufficient and accurate sources of information on potentially-confounding case characteristics.[106] Judgments are, after all, written to *justify* the adjudicator’s preferred outcome, so there is the danger of motivated writing.[107] Further, because opinions are inevitably written *after* the authoring judge knows the treatment status (e.g. whether the accused pleaded guilty), equating judgments would not, strictly speaking, equate *pre*-existing worlds.

In the language of causal inference, judgments are both *post-treatment* and *post-outcome* variables. Using such variables for causal inferences is problematic.[108] Broadly speaking, if judgments depended on both treatment status and outcome, then trying to keep the former judgments constant while varying treatment status is internally contradictory. Thus, as with controlling for mediators,[109] controlling for judgments might *introduce* bias into our results.[110] While a legally-trained human reader might be able to ‘read between the lines’ to discount any post-treatment or post-outcome information, an automated codebook probably cannot.[111]

Alternative text sources of confounding information are therefore necessary. Fortunately, if there is one thing the legal industry has in abundance, it is text. Indeed, *before* any case outcome can be determined, volumes of legal texts are often generated, taking the form of affidavits, briefs, evidence, and related filings. To the extent that case characteristics are confounders, documents which detail them might be useful de-confounders. Two document types are promising. First are case briefs, which we might expect to contain much case information. While parties might be expected to present only favorable factual or legal information, we could simply use the text of *both* sides’ briefs for de-confounding. An alternative is briefs from neutral parties such as amici (assuming they are actually neutral). In practice, briefs *are* generally accessible, though perhaps neither freely nor in bulk.[112]

Second are judgments from *lower* courts. To be sure, issues may change on appeal. Lower court judgments may also be outcome-motivated. Nonetheless, they are in practice slightly less costly to obtain,[113] and would represent an improvement over using judgment texts from the same courts. For these reasons, I focus on using *briefs* for de-confounding in the remaining third of this paper.

## 5.0 De-confounding the Certiorari Game

To enliven the techniques discussed above, this Part demonstrates how text de-confounding might be used to explore a question that has received both practical and scholastic attention: what determines how the Supreme Court grants certiorari (abbreviated "cert." for ease of reading)? I consider in particular whether cert. outcomes differ for cases from state supreme courts (the treatment group) vis-à-vis cases from circuit appeals courts (the control group). Effectively, text de-confounding is used to hypothetically conduct the following experiment: suppose we took two similar cases on petition, and randomly assigned whether they originated from circuit appeals or state supreme courts, would certiorari outcomes differ systematically? If they do, we would have causal evidence that case origin matters.

This question was chosen because it illustrates both the applications and limitations of legal text de-confounding. On applicability, cert. grant has received significant doctrinal and empirical study.[114] Further, petition *briefs* (and not judgments) provide a promising source of de-confounding. Finally, the certiorari question adheres to the template legal analysis question specified in Part 1 above and thus sheds light on how similar questions could be addressed.

On limitations, notice that whether petition briefs can be considered pre-treatment variables is a non-trivial question. Lawyers (or self-represented litigants) writing them might tailor their briefs to the court of origin. At the same time, a case's characteristics, particularly its' factual background, as described in those briefs *are* determined before court assignment. This tension forces us to be particularly careful when describing the relevant causal model.

Further, investigating this question illustrates what I argue are realistic legal data constraints. As will become clearer, although legal texts are generally accessible, dataset imbalance pervades law, necessitating specific countermeasures. Although I obtained data on more than 10,000 Supreme Court certiorari petitions, less than 100 of them successfully obtained the writ of certiorari. As one might expect, this imposed certain constraints on the estimation process that will be made clearer below.[115]

For these reasons, it must be emphasized that the causal estimates presented below are not intended to provide empirical confirmation of whether case origin causally affects certiorari grant. My focus is rather on illustrating the *procedure* for deriving them, rather than their substance. Note, however, that the general statistics presented do permit *associational* interpretation, and might thus be of interest nonetheless. This illustration can be seen as a proof-of-concept for a more conclusive empirical investigation in future work that makes use of further and better data.

The rest of this Part follows a standard chronology for a causal *text* inference study.[116] Part V.A analyses prior literature and identifies potential confounders. Part V.B describes the dataset and explores the initial covariate balance. Part V.C deals with dataset preparation which, for text methods, involve both text encoding and covariate balancing. Part V.D describes the estimation models and discusses results.

## 5.1 What Drives Certiorari Outcomes?

Scholarly interest on how the Supreme Court grants cert. can be traced back to Schubert's seminal paper on "The Certiorari Game".[117] A political scientist, Schubert hypothesized that the Justices' vote in "blocs" motivated by a law-development agenda. According to Schubert, a bloc of four Justices of the time were interested in ensuring the law favored railroad workers over employers. Using game theory, Schubert demonstrated that they would have the following pure strategy:[118]

> 1. *Never vote in favor of petitions from railroads;*
> 2. *Always vote in favor of petitions from workers where the lower appellate court reversed a trial judgment in their favor; and*
> 3. *Always vote for the petitioner on the merits.*

The qualification to point (2) arises because, under Schubert's assumptions, non-bloc Justices were more likely to vote in the workers' favor when the appellate court disagreed with the trial court than when both lower courts were in agreement. Thus, if the bloc voted to grant cert. in the latter case, cert. would be granted (since 4 is sufficient for cert.), but the bloc ran the risk of 'losing' the merits-stage vote 5-4 in the railroads' favor. 12 of the 13 railroad cases which the USSC heard when the bloc voted on cert. following this strategy were decided pro-worker, relative to 8 of 11 cases heard when the bloc did not vote accordingly.

In the Certiorari Game, therefore, cert. votes are a backwards-reasoned function of projected *merits* stage outcomes. If a Justice prefers that the Supreme Court affirms,

they would never vote for cert. If a Justice wants the Supreme Court to reverse, they would vote for cert. *provided* the Supreme Court was indeed likely to reverse.

Schubert's model continues to influence empirical analyses of the USSC's cert. behavior. Brenner tests an extension of the model on cert. votes cast between 1945 and 1957, obtained from Justice Harold Burton's private docket books, and found evidence suggesting that some Justices were, indeed "skillful players in the new certiorari game".[119] More recent work crystallizes this into an "outcome prediction" theory of cert. grants: Justices vote for cert. when they expect to win at the merits stage and against cert. when they expect to lose ('winning' or 'losing' defined in terms of a Justice's desired outcome).[120] Sommer takes this a step further, arguing that Justices further consider if they are likely to author the majority opinion if the case proceeds to oral argument.[121] The primary explanatory variables used in these studies were therefore Justice ideologies.

At the same time, outcome prediction theory provides an incomplete account of certiorari behavior. Brenner et al. point out that prior studies suffer from a "missing data problem" because cert. votes for *denied* petitions are *not* included in any of them.[122] They further criticize Caldeira et al.'s attempt to impute values to the missing variables for reasons beyond the scope of this article.[123] In their view, the outcome prediction theory *primarily* explains certiorari behavior for cases *actually* granted cert., and is less suited for the vast majority of cases *denied* cert.. The latter, they argue, might be better explained by the "error correction" theory: that justices only vote to grant cert. when they believe the lower court decision is legally (rather than ideologically) erroneous.[124]

The literature therefore distinguishes between "strategic" determinants of certiorari grants, being those motivated by the Certiorari Game, and "nonstrategic" determinants which revolve around the procedural and substantive characteristics of a case.[125] These include case salience, whether there was disagreement in the lower courts, whether the Solicitor General was involved, and the number of amici curiae involved at the petition stage.[126] Also relevant is the strength of the petition, in turn measured by whether it was filed *in forma pauperis*, whether it raises frivolous issues, whether the lower court handed down a written opinion, and whether the petition self-represents.[127]

The upshot of this brief literature review is that cert. grants are driven by numerous legal and political factors. Against this backdrop, we specify the causal model. Most crucially, this involves identifying potential confounders.

To fix ideas, let us express this in the language of potential outcomes. The outcome Y is 1 if the petition was granted and 0 if it was denied. The treatment (T=1) group comprises petitions on state supreme court judgments; the control group (T=0) comprises petitions on circuit appeals court judgments.[128] For ease of reference, I call the former state petitions and the latter circuit petitions. The ATE we are after is the difference in cert. outcomes for state petitions *had they originated from* circuit appeals courts, as well as for circuit petitions had they originated from state supreme courts. The potential counterfactual outcomes in both cases are, of course, unobserved.

Recall that confounders are variables which both influence potential and correlate with treatment assignment.[129] Applied here, this means whether cases are assigned to state or circuit courts should be independent of the other cert. relevant factors above. Given that the same Justices vote on petitions regardless of origin, judicial ideology and error correction factors are *conceivably* treatment independent.[130] However, simply as a matter of civil procedure, federal cases are more likely to involve interstate matters, matters involving the federal Constitution, and matters requiring the Department of Justice's ("DOJ"s) involvement. Circuit petitions might involve more salient public interest concerns. Apart from implicating more 'convincing' issues, circuit petitions might also be better drafted — either because parties are more likely to enlist the help of private attorneys, or because of the DOJ's involvement (the DOJ, as the statistics below suggest, has extensive experience with petition briefs).

While variables capturing whether attorneys/the DOJ were involved may alleviate confounding, they would be blunt instruments which implicitly assume that all attorney/DOJ-drafted briefs are of the same quality. Moreover, we would still fail to account for differences in case characteristics. Petition briefs are thus useful for de-confounding in two regards: their *form* de-confounds brief 'quality'; their *substance* de-confounds case characteristics including, but not limited to, salience.

## 5.2 The Dataset

To substantiate the intuitions above, I created a dataset from information retrieved from the Supreme Court website's docket search module.[131] The site provides information on every Supreme Court docket number from 2003 onwards. For cases filed in October 2017 and onwards, petition briefs are also available in PDFs. I downloaded them all and used PDF extraction software to (1) merge briefs that had to be downloaded in parts,[132] and (2) isolate the relevant brief text.[133] I excluded about 1500 of the 10,281 downloaded briefs that could not be processed this way as they had not been subject to optical character recognition.[134]

Using docket numbers as an index, I then linked the brief texts to metadata specified on the Supreme Court website. Variables extracted include case date, the name of the lower court implicated, and the list of litigants, attorneys, attorney offices, and amici curiae involved. I inspected the distribution of unique lower court names raised in the cases and manually mapped them to five main categories.[135] Using the list of attorney offices, I coded whether the Solicitor General was involved as petitioner/respondent.[136]

The final task was to determine petition outcomes. These could not be reliably automatically inferred from the Supreme Court website because where petition outcomes are specified in the docket varies with a case's procedural history. Instead, I extracted them from Supreme Court journals for the October 2017, 2018, and 2019 Terms[137] using a Python script created for this purpose.[138] The extraction algorithm reads the journals line by line while keeping a record of the last section header it came across (*e.g.* "Certiorari Denied"). If a docket number is found, it assigns an outcome to that number based on the section header if it appears under.[139] While the algorithm is not perfect, the number of cert grants/denies it detects are close to the actual numbers specified in the Journals.[140] Note that I do not include summary certiorari dispositions, also known as grant, vacate, and remand orders, as grants.[141]

To summarize, the above yields a dataset that captures, for each Supreme Court docket number from Oct 2017 onwards:

> 1. Whether cert. was granted
> 2. The date the case was docketed
> 3. The number of petitioners/respondents/amici involved through the *entire* case (this is not limited to the petition stage for cases that proceed beyond it)
> 4. The number of attorneys representing each side, again through the entire case
> 5. Whether the Solicitor General's office was one of those attorneys
> 6. The lower court it originates from, as well as the court's broad category
> 7. The text of the petition brief submitted.

For subsequent analysis, I use only data on cases from either state supreme or circuit appeals courts. These respectively comprised 1,058 (10.29%) and 7,618 (74.10%) of the 10,281 cases I obtained metadata on. Of these 8,676 total cases, 7,046 had useable brief texts.[142]

## 5.3 Improving Covariate Balance with Text Matching

### 5.3.1 Existing Covariate Imbalance

Table 4 summarizes the dataset. Note that the bottom five variables are binary, so their averages may be interpreted as percentages.

| | Ovr Mean | Ovr SD | Control Mean | Control SD | Treat Mean | Treat SD | SMD |
|---|---|---|---|---|---|---|---|
| No. Petrs | 1.002 | 0.039 | 1.002 | 0.04 | 1.001 | 0.034 | -0.012 |
| No. Resps | 1.065 | 0.386 | 1.066 | 0.396 | 1.058 | 0.305 | -0.021 |
| No. Amici | 0.211 | 2.2 | 0.206 | 2.225 | 0.243 | 2.013 | 0.017 |
| Petr Repd | 0.448 | 0.497 | 0.445 | 0.497 | 0.471 | 0.499 | 0.054 |
| Resp Repd | 0.788 | 0.409 | 0.802 | 0.399 | 0.694 | 0.461 | -0.265 |
| SG a Petr | 0.003 | 0.055 | 0.003 | 0.058 | 0 | 0 | -0.063 |
| SG a Resp | 0.448 | 0.497 | 0.512 | 0.5 | 0.002 | 0.048 | -1.09 |
| SG an Amicus | 0.006 | 0.074 | 0.006 | 0.075 | 0.005 | 0.067 | -0.016 |

*Table 4. Covariate Balance in the Unmatched Sample*

> Notes*: Statistics are based on all 8,676 state/circuit petitions. "Control" refers to circuit petitions while "Treat" refers to state petitions. "SD" stands for standard deviation. "SMD" refers to the standardized mean difference across treatment and control groups, also known as Cohen's d, which provides a natural, scale-free measure to assess the difference between treatment and control covariate distributions.*[143] *I use a pooled-variance denominator. As a general rule of thumb, SMDs 0.2 and below are 'small', around 0.5 are 'medium', and above 0.8 are large.*[144]

The average (state or circuit) petition involves about 1 petitioner, 1 respondent, and 0.22 amici. 44.8% of petitioners are represented by at least one attorney, with state petitioners being slightly more likely to be represented (47.1% versus 44.5%). A larger proportion (78.8%) of respondents are represented, with circuit respondents being noticeably more likely than state respondents to be represented (80.2% versus 69.4%).

The DOJ is almost never involved as a petitioner, having only appeared as such in 0.3% of circuit petitions and in exactly none of the state petitions. It also rarely appears as an amicus (about 0.5% of both circuit and state petitions). However, the DOJ is a very frequent respondent, appearing as such in 44.8% of all petitions since Oct 2017 onwards. This covers 51.2% of circuit petitions, but *only* 0.2% of state petitions. Taken together, these statistics further suggest the DOJ is seldom involved in state petitions in *any* capacity.

The preceding is consistent with our legal intuition above and reinforces concerns on covariate imbalance. While there does not appear to be wide disparities in the *number* of parties involved, the gaps in respondent representation and DOJ involvement correspond clearly to potential confounders.

### 5.3.2 Encoding and Matching Petition Briefs

Using this data, I demonstrate a simple document term matrix representation based on TFIDF as well as a topic model representation based on Latent Semantic Analysis.[145] Both are broadly based on the Bag-of-Words approach, explained in Part 3.1.

These methods were chosen primarily for their simplicity and transparency, particularly to non-experts. Both have been used to analyze *legal* corpora, albeit from a predictive perspective.[146] Note that whether they are conceptually the best text representations for this question is equivocal. As Part 3.4.2 explained, the 'right' text encoding method depends on the dataset, particularly the sources of confounding we mean to address. Here, textual form and substance both matter. While BOW representations are suitable for matching on substance,[147] because BOW discards syntax and word order, it seems inappropriate for matching on *form*. Language model representations which, as mentioned above, can account for syntax, may be more suitable for this context.[148] However, language models involve neural network methods which require special expertise even to interpret; practical application entails further technicalities.[149]

I thus chose to proceed with BOW models for now. For similar reasons, I follow standard NLP practice (outlined in Part 3.1) in preparing the text representations, even though a more tailored approach might be better suited to this setting.[150] Implementation details on both representations created may be found in Appendix 7.2.

For the TFIDF representation, I used Mozer et al.'s best performing model for the medical corpus, namely cosine similarity matching.[151] For the LSA representation, I used propensity score matching instead, as that allows me to demonstrate how text

and non-text methods compare on the same closeness metric.[152] As above, I followed standard PSM practices, save one deviation given that I had more control than treatment units: to allow more control observations to be used, I matched, with replacement, each treatment unit to *three* control units. Implementation details on matching can be found in Appendix 7.2.

### 5.3.3 Balancing Results

Four matching schemes were thus tested in total: TFIDF with cosine similarity (for ease of reference, "TFIDF+Cos"), propensity score matching with only non-text covariates ("non-text PSM"),[153] propensity score matching with only LSA-encoded texts ("LSA PSM"), and propensity score matching with both LSA-encoded texts and non-text variables ("LSA+non-text" or "Full" PSM).[154] For each scheme, Table 5 summarizes the number of petitions successfully matched as well as the number of *unique* petitions from each group used in the matched sample relative to the unmatched dataset.

| No. of Observations / Matching Scheme | All (non-unique) | Control (unique) | Treatment (unique) |
|---|---|---|---|
| Unmatched | 7046 | 6161 | 885 |
| Non-text PSM | 3540 | 84 | 885 |
| TFIDF + Cos | 3514 | 1049 | 880 |
| LSA PSM | 3540 | 829 | 885 |
| LSA + non-text PSM | 3540 | 722 | 885 |

*Table 5. Number of (Unique) Sample Observations Before and After Matching*

All matching schemes successfully identify sufficiently close (within the matching parameters I set) circuit petitions for all state petitions. However, the non-text PSM achieves this essentially by re-using only 84 unique circuit petitions to match the 885 state petitions. These 84 circuit petitions are, in other words, extremely 'state-like' in terms of the non-text variables. However, once we consider brief texts, a range of other circuit petitions are picked up.

Figure 1 illustrates the covariate balance achieved by each method on the *non-text variables* only. All matching schemes alleviate the potentially confounding imbalance in how often the solicitor general appears as respondent in state versus circuit petitions. Most also addressed imbalance in respondent representation, though the full PSM model seemingly swings it to the other side. This is noteworthy because the TFIDF and LSA only models were not directly supplied with such information, suggesting that information in the brief texts proxy for this. While the non-text PSM seemingly achieves a good balance, this is not surprising given that it attempted to match these same covariates *only.* Since only 84 unique circuit petitions were used, little weight should be placed on this.

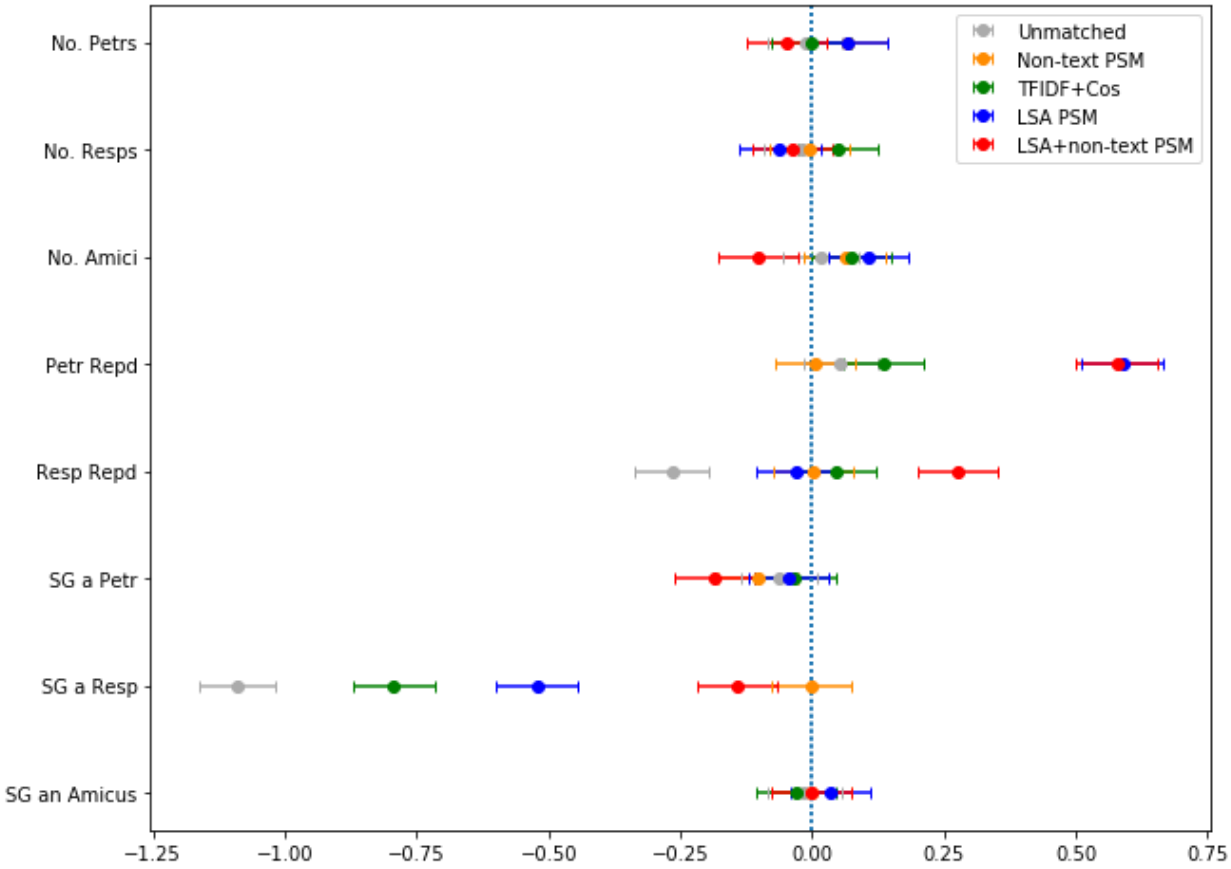


*Figure 1: Covariate Balance Under Various Text and Non-text Matching Schemes*

> Notes*: Dots are point estimates for the standardized mean difference between treatment and control means. Error bars represent 95% confidence intervals. If the intervals touch the dotted line at 0, the difference is not statistically significant at the 5% level. Notice that within the initial, unmatched sample (represented by grey intervals) there are significant differences for variables denoting respondent*

> *representation and Solicitor General involvement as respondent. For most variables, the colored intervals, each representing one matching scheme tested, fall closer to zero than the gray intervals, indicating a better covariate balance.*

Moreover, recall that the text *itself* should also be balanced. To illustrate how different the starting texts were, Figures 2 and 3 below present the difference in the distribution of linearized propensity scores[155] for the LSA and full PSM models respectively.

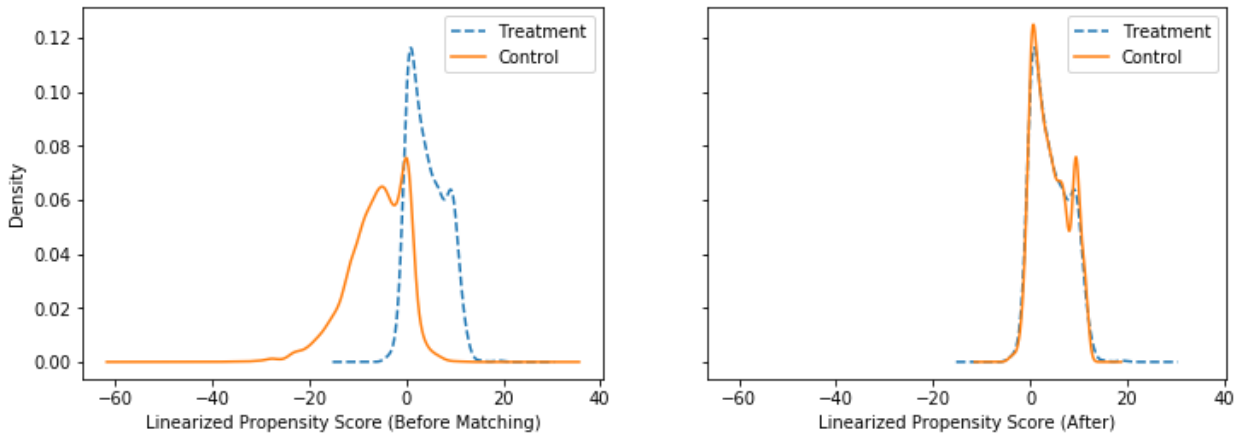


*Figure 2: Difference in Propensity Score Distributions Estimated by the LSA PSM, Before and After Matching*

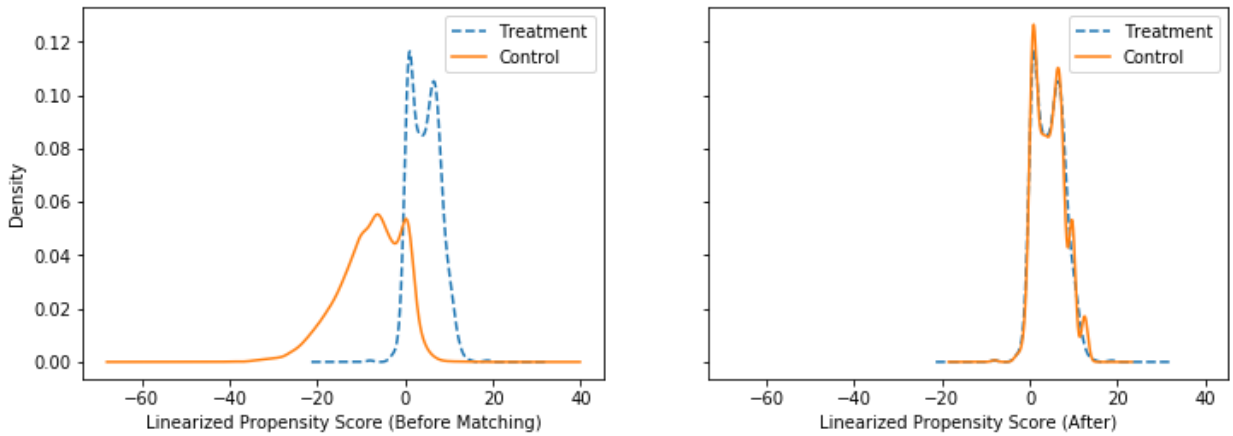


*Figure 3. Difference in Propensity Score Distributions Estimated by the Full PSM, Before and After Matching*

Recall that propensity scores are the probability that a given unit receives treatment given the data. The difference in propensity scores suggests that state and circuit petitions *are* written differently, though I cannot disentangle whether this is due to writing style or case substance (or both). Notably, LSA PSM correctly classifies 88.64%

of petition origins. Put differently, an algorithm given only the LSA-encoded texts could reliably separate them into state versus circuit petitions.[156] Accuracy on the non-text PSM and the Full PSM was 75.59% and 92.37% respectively. As explained above, a difference in brief texts is a possible confounder because it relates to factors such as case salience and petition strength. In this light, the next sub-Part demonstrates the difference that text matching makes on the causal estimates we derive.

## 5.4 Estimating Causal Effects

Finally, I estimated causal effects from the matched datasets. For illustration, I present results from two methods for calculating ATEs. The first simply takes the difference in treatment and control group means.[157] Conceptually, this is equivalent to taking the difference *within* each matched pair of treatment and control observations.[158] This method critically assumes that matching has accounted for all confounding, such that each matched pair effectively represents one set of potential outcomes (that is, both $Y(T=1|Z=z)$ and $Y(T=0|Z=z)$). However, text matching probably has not removed all confounding here. Some important non-text covariates remain unbalanced (see Figure 1 above). This is unsurprising since, as explained above, I chose possibly less appropriate text representations for the sake of illustration.

In this light, I use a second estimation method known as Peters-Belson regression.[159] Peters-Belson illustrates the potential outcomes framework well as it takes the Fundamental Problem of Causal Inference rather literally. The technique broadly involves three steps. First, a model is fit *only* on control group data, and used to impute missing potential outcomes for the *treatment* group. Second, a model is fit only on *treatment* group data and used to impute missing potential outcomes for the *control* group. Finally, with all missing potential outcomes 'filled in', the ATE can be calculated as the average difference between *all* treatment group outcomes and *all* control group outcomes.

The model used to impute treatment/control group outcomes can be varied, as any algorithm capable of classifying outcomes may be used. Inputs to the model can be tailored to account for residual confounders. Here, I use a Bayesian logistic regression to impute outcomes.[160] Across all matching schemes, I provide only three variables to the model: petitioner representation, respondent representation, and Solicitor-General involvement as respondent. These were the variables that generally remained imbalanced across all matching schemes. Keeping variable inputs constant also better isolates the effect of each scheme.

One caveat with the causal estimates that follow should be noted. Because the Supreme Court *rarely* grants cert. the outcome variable is highly imbalanced. Across the two-and-a-half Terms sampled, only 68 (0.89%) circuit petitions and 8 (0.76%) state petitions were granted.[161] Like covariate imbalance, outcome imbalance is problematic for statistical analysis, though in different ways. Just as statistics generally relies on having a sufficient number of observations, many statistical techniques, particularly more sophisticated analyses, relies on having sufficient *minority* outcome group observations as well.[162]

Note however that this does not prejudice the previous illustration of the *balancing* process because the outcome variable is not used therein. In fact, it is essential that outcome variables remain hidden to both algorithm and analyst, until *after* covariate balancing is complete, to prevent us from selecting observations in a manner that favors particular outcomes (thereby influencing causal estimates).[163] The results presented in Figure 4 below should be interpreted in this light.

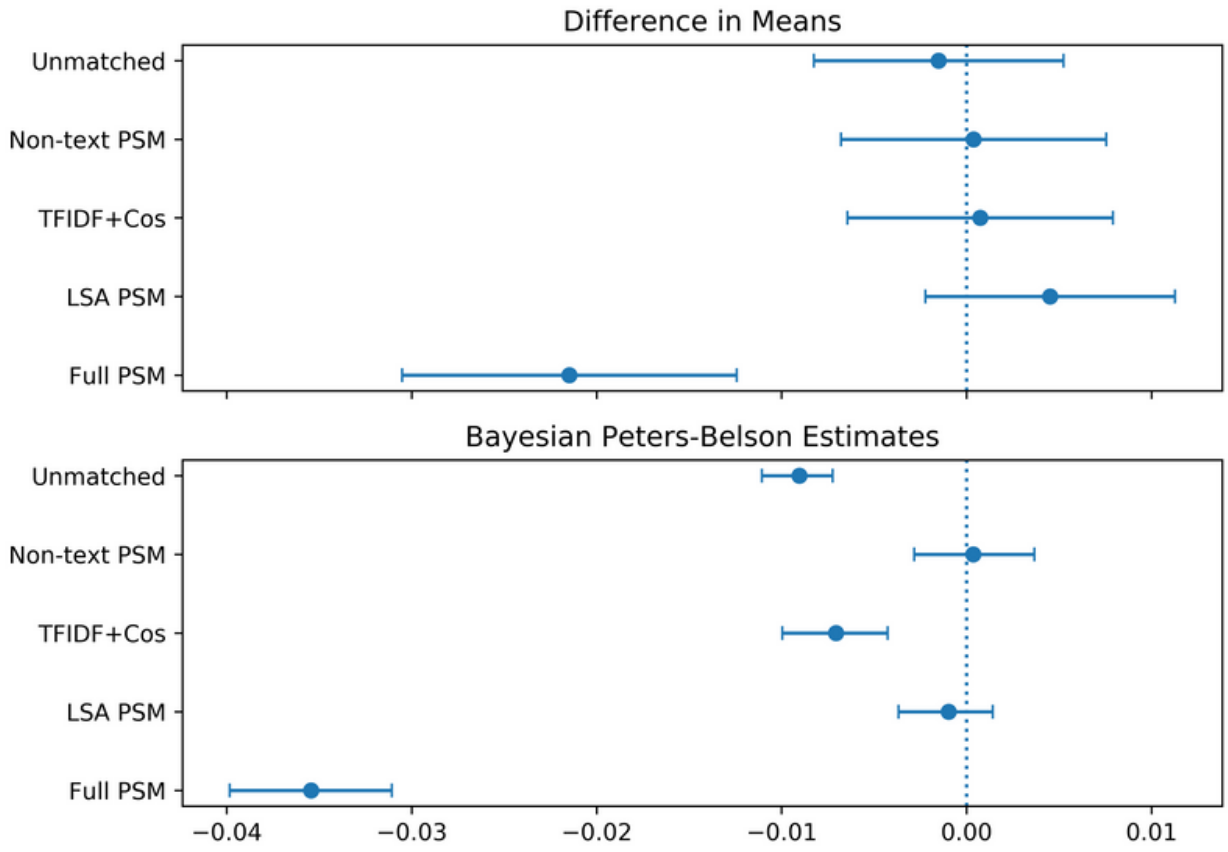


*Figure 4. Average Treatment Effect Estimates for Matched and Unmatched Samples*

> Notes*: Dots are point estimates for the ATE while error bars represent 95% confidence intervals. Where the intervals cross the zero mark, the ATE is not statistically significant at the 5% level. Intervals for the difference in means are calculated in the conventional manner,*[164] *while intervals for the Bayesian regression are based on the 0.025 and 0.975 percentiles of ATE estimates.*

These results yield conflicting signals on whether petition origin influences cert. grants. At face value, the naïve differences in means generally suggest there is generally no effect. But the Full PSM instead suggests a statistically significant 2% reduction in grant rates for state petitions. The theoretically superior Peters-Belson estimates are even less aligned. There, the unmatched and TFIDF+Cos estimates now suggest a significant reduction in grant rates of slightly less than 1%, while the Full PSM suggests an even greater 3+% reduction. The results, therefore, do not corroborate. Instead, ATE estimates differ significantly with and without matching, given different matching schemes, and given different estimation models.

Given the data and methodological constraints explained above, such volatility is unsurprising. It is important not to attach causal significance to any of these estimates. Instead, the primary takeaway of this illustration lies *precisely* in the volatility. Specifically, that text matching makes a significant difference suggests that failing to account for confounding texts might lead us to spurious estimates in this particular context. Thus, although insufficient to identify causality, these preliminary estimates provide an *indication*, which future work can build on, of the statistical countermeasures we likely need should a better dataset be available. These results further reinforce two central themes in this paper: (1) the importance of a causal framework, particularly a carefully-specified causal model, when working with observational legal data;[165] and (2) the importance of accounting for text confounding in causal legal analysis.[166]

To be sure, the observed estimate volatility is likely also driven by *dataset*, rather than methodological, factors, particularly the small number of granted petitions. To the extent that this sample illustrates a *realistic* legal dataset, however, these findings highlight the challenges with statistically deriving causal estimates from observational legal data. Text matching (when done optimally, not illustratively as I have done here) arguably represents a *best-efforts* attempt at doing so. Yet even such an attempt *might* not suffice given the additional challenges and tradeoffs required. For instance, because covariate balancing generally requires discarding observations, it reduces the

number of treatment/control units, possibly worsening the lack of (minority outcome) data that we can expect in law.[167]

## 6.0 Conclusion

Causal text inference methods hold great promise for law, a field which, possibly more than any other, relies on (manually) extracting predictive and causal information from specialized texts. Common legal questions might be re-cast as questions of text causality, including the fundamental legal question of whether a given legal factor affects case outcomes. As causal text methods in the computational (social) sciences become more established and sophisticated, its potential for legal applicability should grow. The case for causal *legal* text inference, which this paper aimed to make, is evident. Its adoption would not only allow us to more rigorously test old answers to old questions. We may find ourselves now able to answer new questions entirely.

There is some way to go, however. Text methods are relatively new, and far from canon even outside law. Adapting them to law might raise new challenges, and legal dataset may prove unwieldy for their purposes. Each step — choosing a text representation, computing a closeness measure, determining a balancing scheme — involves its own variables and parameters. We do not (yet) have clear guidance on what works for law. I have but sketched a framework to *begin* studying this, and illustrated a fraction of the ways in which legal text matching may be done.

If we are truly interested in capturing "a glimpse of [the law's] unfathomable process",[168] to understand what truly drives judicial decisions, and to accord in empirical legal studies the same primacy that text holds in doctrinal analysis, the thought and effort necessary to operationalize legal text inference should not deter us. We should, much to the contrary, question if de-confounding on text may yield results that challenge received wisdom, and bring us that much closer to the true, causal mechanisms of law.

---

## 7.0 Appendices

### 7.1 Distribution of Originating Courts

Table 6 below presents the number of petitions received by the Supreme Court from each originating court type. Data for this was retrieved from the Supreme Court's docket search feature and manually standardized into categories. It is subject to all

D.C circuit but excludes the federal circuit. The Federal Circuit contains *only* the United States Court of Appeals for the Federal Circuit and was intentionally kept separate from other circuit appeals courts due to its specialized docket. State supreme courts include apex courts for all fifty states. Territorial Supreme Court includes the apex courts of American Samoa, Northern Mariana Islands, Guam, Puerto Rico, Virgin Islands, and the District of Columbia.

| Court | Total No. | No. Granted | % Granted |
|---|---|---|---|
| Circuit Appeals Courts | 7618 | 68 | 0.8926 |
| State Supreme Court | 1026 | 8 | 0.7797 |
| Federal Circuit | 128 | 9 | 7.0312 |
| Territorial Supreme Court | 56 | 0 | 0.0000 |
| Other | 1453 | 4 | 0.2753 |

*Table 6. Originating Court Frequencies and Grant Rates, Oct 2017 – Dec 2019*

## 7.2 Implementation Details for Text Matching

All code used for preparing the dataset was written in Python 3 and is on file with the author. This Appendix outlines key implementation details for text matching. Note that all library code referenced in the following sub-sections were used with all default settings except those otherwise specified.

### 7.2.1 Pre-processing

I first used *spaCy*[170] to tokenize the texts and remove tokens that were entirely stop words, punctuation, whitespace, or digits. I further removed all tokens shorter than 3 characters. Remaining tokens were lemmatized and lowercased.

A second round of pre-processing was necessary because the above (standard) approach had trouble dealing with punctuation *within* words (as they would not be matched to stop words or lemmatized). Using basic string methods, I removed all hyphens, periods, and apostrophes occurring *within* tokens before again lemmatizing and lowercased them. I used the *NLTK*[171] WordNet lemmatizer for this step.

### 7.2.2 TFIDF + Cosine Similarity Matching

I used *scikit-learn's*[172] TFIDFVectorizer class with all default settings but two. First, I set *min_df*=3, thus removing words occurring in fewer than 3 documents. This was to remove lint tokens that arose from inevitable imperfections in the PDF extraction step.[173] Second, I set *max_df*=0.95, thus removing words than appear in more than 95% of all documents. This was primarily to reduce the dimensionality of the resulting document-term matrix and keep the computational load manageable. As explained in Part 3.1, such ubiquitous words should not in any event be highly informative of textual context.

This yielded a matrix of 73,086 token TFIDF scores across all 7,046 petition briefs. I fed this into the *sparse_dot_topn*[174] library which allows for efficient cosine similarity matching and wrote customized code to extract the matched docket numbers. For each state petition, I extracted the top 3 closest circuit petitions by cosine similarity, provided that the similarity scores differed by less than 0.0546, being the $0.1^{th}$ quantile the distribution of all cosine distances.[175]

### 7.2.3 Latent Semantic Analysis + Propensity Score Matching

For LSA, I fed the TFIDF matrix above into *scikit-learn's TruncatedSVD* (*i.e.* truncated singular value decomposition) class and extracted the top 100 principal components (*i.e.* topics). The topic scores were then normalized using the *Normalizer* class. Figure 5 visualizes the top 9 topics derived.

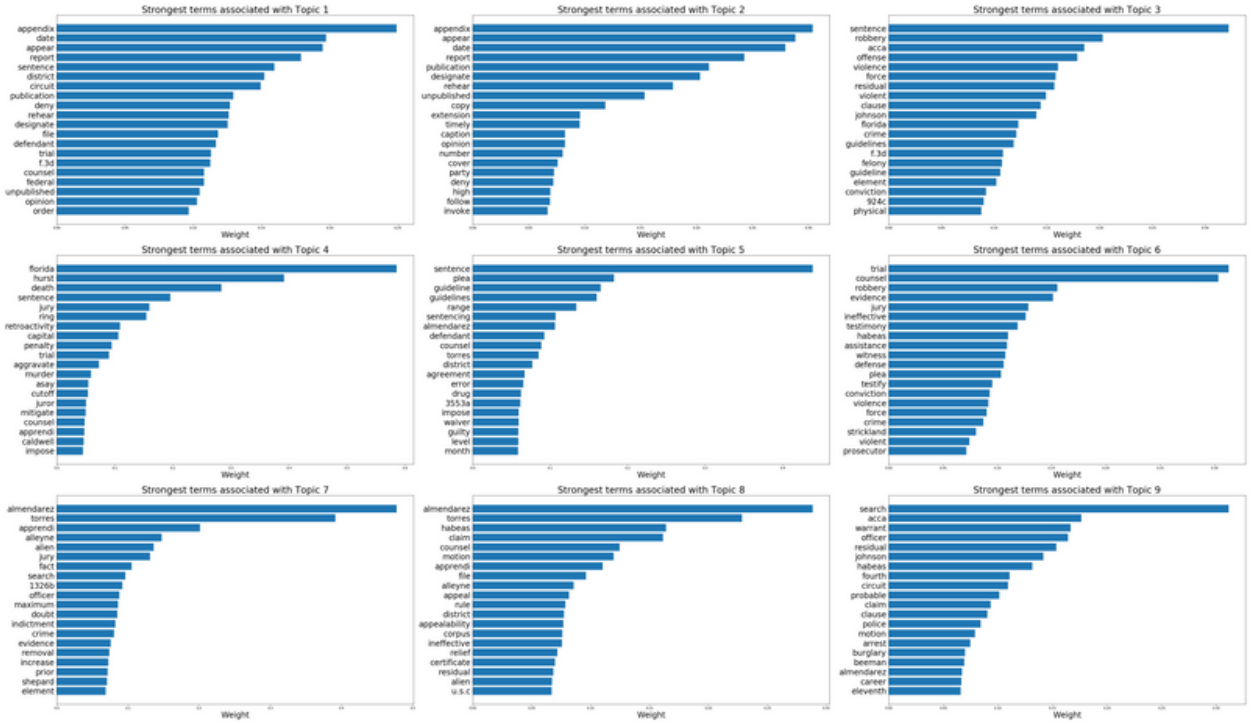


*Figure 5. Term Distributions for Top 9 Topics from LSA Representation of Petition Briefs*

The LSA document-topic matrix, along with base covariates such as the number of petitioner and respondents involved, were then fed varyingly into the *pymatch*[176] *Matcher* class to achieve the non-text, text-only, and LSA + text PSM schemes explained above.[177] Propensity scores were estimated through the library's *fit_scores* method with balanced samples set to true and using only one model. As with TFIDF+Cos, I matched each state petition with replacement to 3 circuit petitions.

---

*Header image generated with Wombo*

## Footnotes

1. Oliver W. Holmes Jr., *The Path of the Law*, 10 Harv. L. Rev. 457, 460 (1897). ↩
2. The formalists famously critiqued the realist account of law as being pre-occupied with "what the judge had for breakfast". Assuming the judge's morning menu is indeed either (1) predictive but not causal, or (2) both predictive *and* causal, would this affect whether we would call this 'law'? See, e.g., Alex Kozinski, What I Ate for Breakfast and Other Mysteries of Judicial Decision Making, 26 Loy. L.A. L. Rev. 993 (1993) (providing a strongly-expressed critique of the "frivolities" of the realist account); Dan Priel, *Law Is What the Judge Had for Breakfast: A Brief History of an*

*UnpalatableIdea*, 68 Buff. L. Rev. 899 (2020) (providing an overview of the debate). This debate persists today in the field of legal analytics. *See* Stefanie Bruninghaus & Kevin D. Ashley, *Predicting Outcomes of Case Based Legal Arguments*, Proc.9th Int'l Conf. on Artif Intell&L. 233, 233–34 (explaining the limitations of statistical approaches to predicting legal outcomes and proposing a rule-based approach); Andrew Stranieri et al., *A Hybrid Rule-Neural Approach for the Automation of Legal Reasoning in the Discretionary Domain of Family Law in Australia*, 7 Artif Intell &L. 153, 154 (1999) (noting the limitations of rule-based approaches in dealing with law's open texture); Andrew D. Martin et al., *Competing Approaches to Predicting Supreme Court Decision Making*, 2 Perspec. on Pol. 761, 761 (Cambridge University Press 2004) (discussing the difference between a legal formalist and political science statistical approach to predicting Supreme Court decisions). ↩

3. Posner takes a characteristically skeptical approach to the quality of judicial opinions. Richard A. Posner, Divergent Paths: The Academy and the Judiciary 16 (*in* JSTOR, 2016) ("[w]hoever writes them, the opinions are rhetorical statements rather than efforts at transparent communication, and both analysis and result may involve compromise among the judges of the appellate panel or even of the entire court, resulting in a blurring of focus. Readers of an appellate opinion, including law professors, may therefore have difficulty figuring out what the nominal author of the opinion was thinking—what moved him or her, what the real nub of the case is"). ↩

4. I will use "causality theory" and "causal inference theory" interchangeably. ↩

5. D. James Greiner, *Causal Inference In Civil Rights Litigation*, 122 Harv. L. Rev. 533, 587. ↩

6. For a sample of the controversy see Richard H. Sander, *A Systemic Analysis of Affirmative Action in American Law Schools*, 57 Stan. L. Rev. 367 (2004) (arguing via regressions that race-affirmative action in the law school admission process leads to Black students failing the bar at a disproportionate rate); Daniel E. Ho, *Why Affirmative Action Does Not Cause Black Students to Fail the Bar Scholarship Comment*, 114 Yale L.J. 1997, 1997 (2004–2005) (arguing in reply that the article "misapplies the basic principles of causal inference"); *see also* Greiner, *supra* note 5 (questioning heavy reliance on regression methods in civil rights cases and putting forth the potential outcomes framework as an important alternative); Steven L. Willborn & Ramona L. Paetzold, *Statistics is a Plural Word*, 122 Harv. L. Rev. F. 48 (2009) (arguing in reply that regression methods should minimally be one permissible statistical tool used); D. James Greiner, *Not All Statistics Are Created*

*Equal*, 122 Harv. L. Rev. F. 1, 2 (2010) (arguing in further reply that regression methods are inferior because they define causal effects in a way that requires "unrealistic, undesirable, and unnecessary assumptions"). ↩

7. Douglas L. Weed, *Truth, Epidemiology, and General Causation Symposium: A Cross-Disciplinary Look at Scientific Truth: What's the Law to Do*, 73 Brook. L. Rev. 943 (2007–2008); Richard A. Berk, *Causal Inference as a Prediction Problem*, 9 Crime & Just. 183 (1987); D. James Greiner, *The New Legal Empiricism & Its Application to Access-to-Justice Inquiries*, 148 Daedalus 64 (2019). ↩

8. *See generally* Greiner, *supra* note 7. This has particular implications for empirical legal studies, a field that frequently makes use of regression methods. ↩

9. Angrist and Pischke, one of the chief architects of the so-called "credibility revolution" in econometrics which advocates for empirical research designs more robust to varying assumptions, provides a good overview in Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics*, 24 J. Econ. Persp. 3 (2010). In a recent book, Pearl and MacKenzie describe advancements in causality theory from computer science as heralding "[a] new science of cause and effect". Judea Pearl & Dana Mackenzie, The Book of Why: The New Science of Cause and Effect (2018). ↩

10. Restatement (Second) of Torts § 430 (A.L.I. 1976). Tortious causation is not, of course, the only legal doctrine of causation, but arguably provides the archetype. ↩

11. *See, e.g.,* Justin Grimmer & Brandon M. Stewart, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, 21 Pol. Anal. 267 (2013); Margaret E. Roberts et al., Matching Methods for High-Dimensional Data with Applications to Text (Apr. 21, 2016) (unpublished manuscript) (on file with the University of California, San Diego); Naoki Egami et al., How to Make Causal Inferences Using Texts (Feb. 8, 2018) (unpublished manuscript), http://arxiv.org/abs/1802.02163; Reagan Mozer et al., *Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality*, Pol. Anal. 1 (2020); Victor Veitch et al., Using Text Embeddings for Causal Inference (May 29, 2019) (unpublished manuscript), https://arxiv.org/abs/1905.12741. ↩

12. Question 1 has been framed as a question of whether judges who have undergone Henry Manne's law and economics training program use more 'law and economics' language in their opinions. Elliott Ash et al., *Ideas Have Consequences: The Impact of Law and Economics on American Justice* (Ctr. L. & Econ., ETH Zurich, Working Paper No. 4, 2019), https://papers.ssrn.com/abstract=2992782. ↩
13. A large body of judicial behavior literature asks similar questions. *See, e.g.*, Jonathan P. Kastellec, *Panel Composition and Judicial Compliance on the US Courts of Appeals*, 23 J.L. Econ. & Org. 421 (2007) (examining how far sitting in a panel of three influences judicial compliance to higher court precedent). Research designs often rely on random case assignment to identify causal effects of judicial identity on legal outcomes. *See, e.g.*, Alma Cohen & Crystal S. Yang, *Judicial Politics and Sentencing Decisions*, 11 Am. Econ. J.: Econ. Pol'y 160 (2019) (exploiting random case assignment to study how far judge political affiliations influence sentencing decisions). If cases were not randomly assigned, as they sometimes may not be in certain courts, an obvious confounder (*i.e.* a factor that obstructs causal inference) would be case attributes such as the facts of the matter, the parties, involved, and the procedural history of the case. Information on these attributes, however, are usually stored in texts – legal briefs, affidavits, decisions, and other procedural documents. Thus, researchers would have to encode these documents into numerical data before any quantitative causal model may be built. ↩
14. See, e.g., Pamela C. Corley, The Supreme Court and Opinion Content: The Influence of Parties' Briefs, 61 Pol. Res. Q. 468; Lance N. Long & William F. Christensen, Does the Readability of Your Brief Affect Your Chance of Winning an Appeal?, 12 J. App. Prac. & Process 145 (2011). ↩
15. For a formal philosophical treatment see generally Boris Kment, *Counterfactuals and Explanation*, 115 Mind 261 (Oxford University Press 2006). ↩
16. The terminology and notation used throughout this Part originates from Imbens & Rubins' "potential outcomes" framework which will be explained shortly below. Guido W. Imbens & Donald B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction ch. 1 (2015). ↩
17. The terminology and notation used throughout this Part originates from Imbens & Rubin's "potential outcomes" framework which will be explained shortly below. Guido W. Imbens & Donald B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction ch. 1 (2015). ↩

18. Paul W. Holland, *Statistics and Causal Inference*, 81 J. Am. Stat. Ass'n 945, 947 (*in* JSTOR, [American Statistical Association, Taylor & Francis, Ltd.] 1986). ↩
19. *Id.* (noting that the "statistical solution" to the Fundamental Problem is to extract information on unobservable outcomes from the observed outcomes). ↩
20. Imbens & Rubin, *supra* note 15, at 3–22. ↩
21. *Id.* sec. 1.5. ↩
22. To be precise, there is a weaker version of unconfoundedness, "ignorability", under which causal inference is still legitimate. While unconfoundedness requires all potential outcomes to be independent of treatment assignment given the observed covariates, ignorability only requires the *unobserved* potential outcomes to be independent as such. Ignorability often suffices because causal inference only requires unobserved potential outcomes to be imputed. As the technicalities of unconfoundedness versus ignorability are presently immaterial and unconfoundedness is intuitively simpler, my exposition proceeds with the same. In any event, unconfoundedness is not frequently distinguished from ignorability in practice. For details on ignorability see *Id.* at 39–44. ↩
23. *Id.* at 10–12. ↩
24. This oversimplifies. For a formal treatment see Tyler J. VanderWeele & Ilya Shpitser, *On the Definition of a Confounder*, 41 Ann. Statist. 196, 196–220 (2013) (explaining the conventional view of a confounder as "a pre-exposure variable that was associated with exposure and associated also with the outcome conditional on the exposure, possibly conditional also on other covariates" while proposing an alternative). *See also generally* Judea Pearl & Azaria Paz, *Confounding Equivalence in Causal Inference*, 2 J. Causal Infer. 75 (2014) (describing formal criteria for deciding if a variable should be de-confounded for). ↩
25. Imbens & Rubin, *supra* note 15, at 38–39. *See generally Id.* at 257–80 (discussing confoundedness in more detail). ↩
26. Greiner, *supra* note 7. ↩
27. Judea Pearl, *On the Interpretation of Do(x)*, 7 J. Causal Infer. 1, 1–2 (2019); *see also* Judea Pearl, Causality, § 7.2.4 (*in* eBook Collection (EBSCOhost), 2000) (discussing how structural equation models may be thought of in terms of "local surgeries"). ↩

28. *See, e.g.*, Javid Gadirov, *Causal Responsibility in International Criminal Law*, 15 Int'l Crim. L. Rev. 970 (Brill Nijhoff 2015) (investigating parallels between Pearl's causality models and causality in international criminal law). ↩
29. Pearl & Mackenzie, *supra* note 9, at 23–51. ↩
30. *Id.* ↩
31. It would be surprising (and probably unethical) if the lawyer instead devised an experiment to test this by randomly advising certain clients to plead guilty. ↩
32. Pearl & Mackenzie, *supra* note 9, at 31 (also noting that "[w]e cannot answer questions about interventions with passively collected data, no matter how big the data set or how deep the neural network"). ↩
33. If changing T *causes* a change in X which in turn causally changes Y, then X is not a confounder but a mediating variable. This falls beyond my scope. For technical details Judea Pearl, *Interpretation and Identification of Causal Mediation*, 19 Psychol. Methods 459 (2014). ↩
34. For a gentle introduction to the concept see JoAnne M. Youngblut, *A Consumer's Guide to Causal Modeling: Part I*, 9 J. Pediatr Nurs. 268 (1994). *See also* Judea Pearl et al., Causal Inference in Statistics: A Primer 24–32 (2016) (providing a textbook introduction). Pearl & Mackenzie, *supra* note 9, at 1–15 (providing a non-technical overview in, without referring to the concept as a causal model). ↩
35. Youngblut, *supra* note 33, at 269. ↩
36. Judea Pearl, *On the Testability of Causal Models with Latent and Instrumental Variables,* UAI' 95: Proc. Eleventh Conf. on Uncertainty in Artif Intell 435 (1995) ("[i]t is well known that one cannot infer causation from statistical data unless one is willing to supplement the data with causal assumptions"). ↩
37. *See generally* Grimmer & Stewart, *supra* note 11 (providing an overview of "text-as-data" methods). ↩
38. Nina Varsava, *Computational Legal Studies, Digital Humanities, and Textual Analysis*, *in* Computational Legal Studies (Ryan Whalen ed., 2020). ↩
39. *See generally* Michael A. Livermore & Daniel N. Rockmore, Law as Data: Computation, Text, & the Future of Legal Analysis (2019); Computational Legal

Studies: The Promise and Challenge of Data-Driven Legal Research (Ryan Whalen ed., 2020). ↩

40. The closest work seems to be Marion Dumas & Jens Frankenreiter, *Text as Observational Data*, *in* Law as data (Michael A. Livermore & Daniel N. Rockmore eds., 2019) (pointing out that legal text is primarily observational, cautioning against using such data with machine learning methods that lack clear causal frameworks to study causal legal questions). ↩

41. *See generally id.* ↩

42. Michael A. Livermore & Daniel N. Rockmore, *Distant Reading the Law*, *in* Law as data 12 (Michael A. Livermore & Daniel N. Rockmore eds., 2019) ("translating the law into data capable of tractable analysis generally involves some loss of information"). ↩

43. *See generally* Dumas & Frankenreiter, *supra* note 39. ↩

44. Grimmer & Stewart, *supra* note 11. ↩

45. *See id.* ↩

46. ↩

47. G. Salton et al., *A Vector Space Model for Automatic Indexing*, 18 Comm. 613 (ACM 1975) (the seminal paper describing the Bag-of-Words model); Livermore & Rockmore, *supra* note 41, at 12 (explaining the Bag-of-Words for a legal audience). ↩

48. *See* Part 4 *infra* (explaining statistical challenges with causal legal inference). ↩

49. Richard Wydick, *Plain English for Lawyers*, 66 Calif. L. Rev. 727, 729 (1978). ↩

50. For a list of stop words used in "spaCy", a standard NLP code library, see Explosion.AI, *spaCy*, github.com, https://github.com/explosion/spaCy. ↩

51. Stephen Robertson, *Understanding Inverse Document Frequency: On Theoretical Arguments for IDF*, 60 J. Documentation. 503 (2004). ↩

52. Gerard Salton & Christopher Buckley, *Term-Weighting Approaches in Automatic Text Retrieval*, 24 Info. Processing & Mgmt. 513 (1988). ↩

53. Canonical examples include Latent Semantic Analysis and Latent Dirichlet Allocation. *See* T.K. Landauer et al., *Introduction to Latent Semantic Analysis*, 25

Discourse Processes 259 (1998); Scott Deerwester et al., *Indexing by Latent Semantic Analysis*, 41 J. Am. Soc'y Info. Sci. 391 (1990); David M. Blei et al., *Latent Dirichlet Allocation*, 3 J. Mach. LearningRes. 993 (2003). ↩

54. Deerwester et al., *supra* note 51. ↩

55. To be precise, topic models are not in themselves codebooks, but a method for *deriving* the codebook from the text itself. The general algorithm is first fit onto the text so it might learn attributes such as the vocabulary involved. The fitted algorithm is then used to transform the texts. This subtle difference is suppressed for ease of exposition. ↩

56. These are sometimes referred to in NLP literature as word *embedding* algorithms, though I prefer the alternative term "vector" to differentiate them from other methods discussed. ↩

57. Tomas Mikolov et al., *Distributed Representations of Words and Phrases and Their Compositionality*, Advances in Neural Info. Process. Syst. 26 3111 (C. J. C. Burges et al. eds., Curran Associates, Inc. 2013) (proposing the well-known "Word2Vec" model for computing word vectors); For exposition and application on word vectors in a legal setting see Dumas & Frankenreiter, *supra* note 39 (applying word vectors to opinion texts). ↩

58. Tomas Mikolov et al., *Linguistic Regularities in Continuous Space Word Representations*, Proc. 2013 Conf. N. Am. Chapter Ass'n for Computational Linguistics 746 (Jun. 2013). ↩

59. This is referred to as average or max *pooling*. *See* Denny Britz, *Understanding Convolutional Neural Networks for NLP*, WildML (Nov. 7, 2015), http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/ (providing a gentle introduction to pooling in neural methods); *see also* Qian Chen et al., *Enhancing Sentence Embedding with Generalized Pooling*, Proc. 27th Int'l Conf. on Computational Linguistics 1815 (Aug. 2018) (formally studying avg/max pooling and more sophisticated techniques). ↩

60. Jacob Devlin et al., *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, Proc. 2019 Conf. N. Am. Chapter Ass'n for Computational Linguistics 4171 (Jun. 2019). ↩

61. *Id.* ↩

62. William Fedus et al., *Maskgan: Better Text Generation via Filling in the __*, Proc. 6th Int'l Conf. on Learning Representations (2018); Devlin et al., *supra* note 58. ↩
63. Kevin Clark et al., *What Does BERT Look at? An Analysis of BERT's Attention*, Proc. 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP 276 (2019). ↩
64. See *infra* Part 3.4.1. ↩
65. Ash et al., *supra* note 12. ↩
66. Egami et al., *supra* note 11, at 18–21. ↩
67. .. . Mark A. Cohen et al., *Measuring Public Perceptions of Appropriate Prison Sentences: Report to National Institute of Justice*, No. 199365 (Washington, DC: USDeptof Justice, Office of Justice Programs, National Institute of Justice2002) ↩
68. Specifically, they used Roberts et al.'s "Structural Topic Model". *See* Roberts et al., *supra* note 11 (describing the Structural Topic Model). ↩
69. There is concededly some arbitrariness in the interpretation of word clusters. ↩
70. Egami et al., *supra* note 11, at 21. ↩
71. Lance N. Long & William F. Christensen, *Does the Readability of Your Brief Affect Your Chance of Winning an Appeal?*, 12 J. App. Prac. & Process. 145 (2011). There are practical applications as well. For instance, legal analytics companies tout how machine learning could help attorneys identify and write arguments that particular judges prefer. Barney Thompson, *Big Data: Legal Firms Play 'Moneyball,'* Financial Times (Feb. 6, 2019), https://www.ft.com/content/ca351ff6-1a4e-11e9-9e64-d150b3105d21. ↩
72. Empirical scholars have long been interested quantifying the effect of a case on the law. *See, e.g.*, William H.J. Hubbard, *The Effects of Twombly and Iqbal*, 14 J. Empirical Legal Stud. 474 (2017); Daniel L. Chen, *Judicial Analytics and the Great Transformation of American Law*, 27 Artif Intell & L. 15, 31–38 (2019) (explaining how the consequences of legal precedents might be empirically measured). ↩
73. Egami et al., *supra* note 11, at 21–24. For those interested in the technicalities, the specific algorithm is the supervised Indian Buffet Process introduced by Fong and Grimmer, 2016. Christian Fong & Justin Grimmer, *Discovery of Treatments from*

*Text Corpora*, Proc.54th Ann. MeetingAss'n for Computational Linguistics (Volume 1 1600) (Aug. 2016). Broadly, the algorithm assumes a set of latent features in the text *and* an underlying relationship between the text and the outcome. Given the text as well as outcome data as inputs, a nonparametric Bayesian approach is used to learn both a set of binary topic indicators as well as the effects of each topic on the outcome. Importantly, both Fong and Grimmer and Egami et al use a cross-validation approach to avoid leaking outcome data into the codebook. That is, they fit topic model on a separate 'training' data set before using it to encode and derive causal effects on a test set. ↩

74. They however clarify that the dataset used might not fulfil all the assumptions necessary for causal inference. Text-as-treatment further requires (on top of SUTVA and ignorability) the assumption that the codebook is *sufficient*, in the sense that the representation created by the codebook must capture enough outcome-relevant information in the text that any outcome-relevant text features left out of the representation must be orthogonal to that representation. Egami et al., *supra* note 11, at 23. ↩

75. *Id.* at 13. ↩

76. I revisit this question in detail below. *See infra* Part V. ↩

77. *See infra* Part V. ↩

78. *See* Imbens & Rubin, *supra* note 15, at 309–10. ↩

79. On the contrary, one of the fundamental implications of Zipf's law is that nature tends toward *imbalance*. *See* G.K. Zipf, The Psycho-Biology of Language: An Introduction to Dynamic Philology 254–61 (2013) (explaining why words might be expected to follow an exponential frequency curve); *see generally also* David M.W. Powers, *Applications and Explanations of Zipf's Law*, *in* New Methods in Language Process. & Computational Natural Language Learning (1998) (surveying potential applications of Zipf's law beyond word frequencies). Somewhat relevantly, Zipf's law was formulated from the observation at word frequencies in a corpus tend to follow a power distribution: the most frequent words occur exponentially more than the least. Katz et al's study of law professor hiring and placement networks further provides evidence of "the highly skewed and/or fractal properties of legal systems". Daniel Martin Katz et al., *Reproduction of Hierarchy? A Social Network Analysis of the American Law Professoriate*, 61 J. Legal Educ. 76, 94 (2011) (the paper generally

identifies how JD graduates from only two institutions, Harvard and Yale, dominate most of the American legal professoriate). ↩

80. Imbens & Rubin, *supra* note 15, at 337–38. ↩

81. *See, e.g.*, *id.* at 339–43 (illustrating Mahalanobis distance matching). ↩

82. Each of these techniques involve their own considerations and specificities that are too extensive to comprehensively cover here. *See generally id.* at 337–74 (detailing matching and trimming methods). ↩

83. Pearl & Paz, *supra* note 23, at 75. ↩

84. *See generally supra* Part 1 and sources cited there (discussing the limitations of regression). ↩

85. *See* Joshua David Angrist & Jörn-Steffen Pischke, Mostly Harmless Econometrics : An Empiricist's Companion 64–68 (*in* eBook Collection (EBSCOhost), 2009) (discussing the problem of "bad controls"). ↩

86. *See* Pearl, *supra* note 32 (formally detailing causal mediation analysis). ↩

87. *See generally ibid.* ↩

88. Roberts et al., *supra* note 11; Mozer et al., *supra* note 11. ↩

89. I consider how these choices might be made in a legal context below. *See infra* Parts IV and V. ↩

90. They also consider a range of distance metrics, but these are not germane to my present discussion. ↩

91. Mozer et al., *supra* note 11, at 13. They assessed "better" by comparing the matched sample against a human-matched standard. ↩

92. *Id.* at 18. ↩

93. *Id.* at 7. ↩

94. *Id.* at 19–20. ↩

95. Veitch et al., *supra* note 11. ↩

96. *See id.* at 6 (providing the formula used). ↩

97. *Id.* at 2. ↩
98. *Id.* at 6. ↩
99. A caveat is that a cross-validation approach must be used because the model would have already seen the outcomes once. Thus, the model has to learn the embeddings from a different set of data from that used to actually estimate causal effects. *See* Egami et al., *supra* note 11, at 9–16 (explaining a cross-validation procedure for causal text estimation). ↩
100. *See also* Roberts et al., *supra* note 11, at 1 (explaining PSM's limitations in a high-dimensional setting). ↩
101. Mohammad Hassan Falakmasir & Kevin D. Ashley, *Utilizing Vector Space Models for Identifying Legal Factors from Text*, Proc. 30th Int'l Conf. on Legal Knowledge & Info. Sys. – JURIX 2017 183 (Adam Wyner & Giovanni Casini eds., Frontiers in Artificial Intelligence and Applications, 2017). ↩
102. *See also supra* Part 2.1 (illustrating how varying legal questions might be seen as questions of causality). ↩
103. *See supra* Part 2.2. ↩
104. Chen, *supra* note 70, at 34 ("[i]n law, we cannot randomize judicial decisions, since doing so would undermine the notion of justice and equal treatment before the law"). ↩
105. *See supra* Part 2.1 (explaining how data helps to overcome the Fundamental Problem of Causal Inference). ↩
106. For Posnerian skepticism on the comprehensiveness of opinions see *supra* note 3. ↩
107. Thus, facts and other case attributes may be, consciously or unconsciously, selectively presented to suit the judge's preferred outcome. *See generally* Elliott Ash et al., *Motivated Reasoning in the Field* (Toulouse Sch. Econ., Working Paper n. 18–976, 2018), https://papers.ssrn.com/abstract=3205116 (explaining motivated reasoning by judges and text mining appellate opinions for phrases that judges may use to express ideological preferences). ↩

108. *See* Angrist & Pischke, *supra* note 83, at 64 (noting that "bad controls are variables that are themselves outcome variables in the notional experiment at hand [while g]ood controls are variables that we can think of as having been fixed at the time the [treatment] of interest was determined"); *see also* Jacob M. Montgomery et al., *How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It*, 62 Am. J. Pol. Sci. 760 (2018) (explaining the dangers of post-treatment bias even in experimental settings). ↩
109. *See supra* III.D.1 (explaining the problem with controlling for mediators). ↩
110. Early work in empirical legal studies attempted to identify the outcome effects of legal factors by coding information from judgments. In light of modern causality theory, these should be interpreted as uncovering *associational* or predictive, but not necessarily causal, legal factors. *See, e.g.*, Jeffrey A. Segal, *Predicting Supreme Court Cases Probabilistically: The Search and Seizure Cases, 1962-1981*, 78 Am. Pol. Sci. Rev. 891 (1984) (using logistic regression on a dataset of manually-encoded search and seizure cases to identify how the court decides). ↩
111. An avenue for future research might thus be to identify whether such information could be pre-processed or fine-tuned *out* of a text representation. I speculate that this would be difficult since it means penalizing the model for being able to predict treatments and/or outcomes, when in order to learn confounding information we aim to do the exact opposite. ↩
112. Controversy surrounds the fees charged to access court documents on PACER. *See* Seamus Hughes, *The Federal Courts Are Running An Online Scam*, POLITICO Mag. (Mar. 20, 2019), https://politi.co/2HJss6Y. ↩
113. The Harvard Law Library has digitized substantially all of U.S. case law up till 2018, and has made them freely available online (albeit subject to certain restrictions). Harvard Law Library, Caselaw Access Project, https://case.law/ (last visited July 30, 2020). ↩
114. *See infra* Part V.A. ↩
115. *See infra* Part V.B (presenting statistics on cert. grant rates by origin). ↩
116. As one might expect, there is no one 'perfect' order of steps, but a range of possible ones. ↩

117. Glendon A. Schubert, *The Study of Judicial Decision-Making as an Aspect of Political Behavior*, 52 Am. Pol. Sci. Rev. 1007 (*in* JSTOR, [American Political Science Association, Cambridge University Press] 1958). ↩
118. A pure strategy is a set of moves which completely defines how a given player will play the game. It is usefully contrasted against a mixed strategy, where a player has incentives to alternate between moves probabilistically. ↩
119. Saul Brenner, *The New Certiorari Game*, 41 J. Pol. 649, 655 (*in* JSTOR, [University of Chicago Press, Southern Political Science Association] 1979). ↩
120. Lawrence Baum, The Puzzle of Judicial Behavior 79 (*in* JSTOR, 1997). ↩
121. Udi Sommer, *How Rational Are Justices on the Supreme Court of the United States? Doctrinal Considerations during Agenda Setting*, 23 Rationality & Soc'y 452 (2011). ↩
122. Saul Brenner et al., *The Outcome-Prediction Strategy in Cases Denied Certiorari by the U.S. Supreme Court*, 130 Pub. Choice 225 (*in* JSTOR, Springer 2007). This in fact reinforces the importance of having a clear causal framework underlying the inquiry. ↩
123. Gregory A. Caldeira et al., *Sophisticated Voting and Gate-Keeping in the Supreme Court*, 15 J. L. Econ. & Org. 549 (1999). ↩
124. Baum, *supra* note 118, at 79 (explaining the error correction literature); Brenner et al., *supra* note 120, at 235–36 (suggesting that error correction applies to "most cases denied cert"). ↩
125. Caldeira et al., *supra* note 121, at 549–50; Sara C. Benesh et al., *Aggressive Grants by Affirm-Minded Justices*, 30 Am. Pol. Res. 219, 220–21 (SAGE Publications Inc 2002). ↩
126. Caldeira et al., *supra* note 121, at 549–50; Benesh et al., *supra* note 123, at 220–21. ↩
127. Brenner et al., *supra* note 120, at 226–27. ↩
128. For the courts included in each category see Part V.B *infra*. Petitions do reach the Supreme Court from other courts, but these two originating courts types are the

most frequent. *Seeinfra* Appendix VII.A (presenting statistics on originating courts). ↩
129. *See supra* Part 2.1. ↩
130. This does rest on the premise that the Justices do not see their law development/error correction roles differently for state supreme vis-à-vis circuit appeals courts. I will assume this for now, though this premise should be researched more rigorously. ↩
131. Supreme Court of the United States, *Docket Search*, supremecourt.gov, https://www.supremecourt.gov/docket/docket.aspx. (last visited July 30, 2020). ↩
132. The Supreme Court website occasionally serves the briefs in separately hyperlinked parts. My scraper names them by docket number, so I know if the same case's brief was downloaded in more than one part. ↩
133. After manually studying a sample of the briefs, I noted that the PDFs often contained appendices laying out reference documents such as the lower court judgment. The main petition text almost always occurred first, though it was occasionally pre-empted by an application to file the writ *in forma pauperis*. To isolate just the briefs, I created a script that searched for the first occurrence of variations of the word "appendix" (case insensitive) appearing *on its own line.* All proceeding text was removed. ↩
134. I inspected a random sample of excluded briefs and found that many were extremely short handwritten briefs. Others were type-written briefs that had been uploaded as scanned documents. I did not find any systematic pattern on which briefs were excluded this way. ↩
135. The categories are: (1) circuit appeals courts (which include the D.C. circuit appeals court), (2) state supreme courts, (3) the federal circuit court (which includes *only* the Court of Appeals for the Federal Circuit), (4) territorial supreme courts (e.g. the D.C. Court of Appeals and the Supreme Court of Guam), and (5) others (including the U.S. Court of Appeals for the Armed Forces, and many *state* intermediate appellate, circuit, and district courts). Those familiar with the U.S.' court naming 'system' would know that this process is rather involved. I first matched all court names to their official names. There were typographical errors, and court names were not always uniformly presented (e.g. the "Supreme Court of State X" is often also referred to as the "State X Supreme Court" and vice-versa). I picked out the

state supreme courts essentially by guessing based on the names and checking this online. The manual codebook used can be made available on request. ↩

136. I achieve this with a case-insensitive search for search for "Noel J. Franscisco" in the attorneys list of each party type. I corroborate this against a search for whether "Solicitor General" or "United States Department of Justice" appears in the list of attorney offices. Both yield nearly identical results and differ only in 16 cases where a *state* solicitor general's office was labelled "Solicitor General" (e.g. docket 18-391). I preferred the former method as it was more specific. Note that it only works because my dataset only extends back to October 2017 – Solicitor General Francisco was appointed in September 2017. ↩

137. The 2019 Term is still ongoing when I wrote this, thus the 2019 Journal is incomplete. ↩

138. All code used for this paper is on file and may be made available on request. ↩

139. The actual algorithm used is more sophisticated and takes care of edge cases like docket numbers breaking across lines. ↩

140. Actual cert. grant numbers from the 2017 and 2018 Journals are 77 and 83 respectively. Statistics are not available for 2019 as the Term is ongoing. It is not clear if the reported numbers include summary dispositions, though there is a separate row item for that. The cert. grant numbers I derived from parsing the Journals, which exclude summary dispositions, are 68 and 82. Some non-alignment is to be expected because I extract outcomes for the *entire* Journal, which states outcomes for all cases that term, while Supreme Court statistics are reported as of June 29 of the given year. Supreme Court of the United States, *Journal of the Supreme Court of the United States: October Term 2017* (2018) at II; Supreme Court of the United States, *Journal of the Supreme Court of the United States: October Term 2018* (2019) at II. ↩

141. Such orders, or "GVR"s, are summarily granted in exceptional cases, such as when a Supreme Court ruling issued after the lower judgment may have superseded it. *See generally* Aaron-Andrew P. Bruhl, *The Supreme Court's Controversial GVRs - And an Alternative*, 107 Mich. L. Rev. 711 (2009). ↩

142. This comprises 6,161 circuit and 885 state petitions. *See infra* Table 5. ↩

143. Imbens & Rubin, *supra* note 15, at 310–11 (explaining that SMDs are preferable to ordinary t-statistics when assessing covariate balance because the latter decreases as sample size increases, even though improving balance should be easier with a larger sample). ↩
144. Shlomo S. Sawilowsky, *New Effect Size Rules of Thumb*, 8 J. Mod. Applied Stat. Methods 597, 598 (2009). ↩
145. *See* Part 3.1.2 *supra* (explaining topic models); *see also* note 51 *infra* (citing LSA literature). ↩
146. *See, e.g.*, Eric Talley & Drew O'Kane, *The Measure of a MAC: A Machine-Learning Protocol for Analyzing Force Majeure Clauses in M&A Agreements*, 168 J. Institutional. & Theoretical Econ. 202 (2012) (using LSA-encoded contracts to predict if a clause is a material adverse change clause); Nikolaos Aletras et al., *Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective*, 2 PeerJ Computer Sci. (2016) (using LSA-encoded opinions to predict European Court of Human Rights decisions). ↩
147. To recall, Mozer et al. found that matching on particular medical words and phrases was best achieved by simple (albeit, high-dimensional) document term matrix representations built on the Bag-of-Words model. This finding arguably carries over into legal concepts, since law has its own highly precise vocabulary of art. ↩
148. Mozer et al. do not benchmark language models in their study. *See generally* Mozer et al., *supra* note 11. ↩
149. Explaining what language models do is itself an NLP a sub-field known as "probing". Clark et al., *supra* note 61, at 276. ↩
150. Recall that standard NLP practice excludes stop words like "the", "of", and "for" from the texts before preparing the representation. *See* Part 3.1 *supra*. This allows the representation to focus on substantive words, but further removes stylistic information. Future work could therefore experiment with text representations that do not preclude stop words. To clarify, language model inputs are not subject to stop word removal. ↩
151. Recall that PSM is unfeasible for TFIDF representations as such representations have extremely high dimensions. *See* Part 3.2.2 *supra*. For my corpus, the TFIDF

matrix had 73,086 columns (one for each unique word in the vocabulary). ↩
152. *Cf.* Mozer et al., who compare propensity score matching on non-text variables to cosine similarity matching on text variables. Mozer et al., *supra* note 11, at 18–21. ↩
153. The propensity score here was estimated with a logistic regression of treatment status on all of the covariates presented in Table 4 above *except* whether the Solicitor General was a petitioner. This had to be excluded from the regression as it had almost no variation. ↩
154. Note that the Full PSM's approach to incorporating both text and non-text variables differs from Mozer et al., who suggest first trimming observations with non-text propensity scores more than 0.1 standard deviations away from the average (what the literature calls "propensity score calipers") before cosine similarity matching on text variables only. Mozer et al., *supra* note 11, at 18. This was unfeasible here as doing so only leaves 9 observations (implying that most propensity scores are at the extreme ends). ↩
155. Linearized propensity scores are essentially propensity scores put through a logistic transformation. Imbens and Rubin explain that the difference in propensity scores can be used to assess covariate balance, and further that linearizing the scores allows better visual comparison. For the formula and details see Imbens & Rubin, *supra* note 15, at 314–17. ↩
156. To be precise, "accuracy" refers to the fraction correctly classified, including both true positives and true negatives. Note that this was calculated based on in-sample fit. No claims are made about out-of-sample predictions. ↩
157. This being the ATEs presented in prior work. *See* Mozer et al., *supra* note 11, at 20. ↩
158. This relies on the mathematical fact that the average of differences equals the difference in averages. ↩
159. For a more detailed explanation of Peters-Belson regression against a legal context (civil rights) see H. Hikawa et al., *Local Linear Logistic Peters-Belson Regression and Its Application in Employment Discrimination Cases*, 3 Stat. & Its Interface 125 (2010); Efstathia Bura et al., *The Use of Peters-Belson Regression in Legal Cases*, *in* Nonparametric Statistical Methods And Related Topics 213 (2012). ↩

160. Unlike conventional regression methods which estimate only one set of parameters, a Bayesian regression assumes that model parameters come from some prior distribution, and samples that distribution repeatedly to produce various estimated parameters. At risk of oversimplification, this is similar to fitting the model multiple times with different random initializations. Doing so better accounts for the uncertainty in model parameters. When used in a Peters-Belson framework, this then produces a range of estimated potential outcomes, which in turn yields a range of ATEs. Confidence intervals for the ATEs can then be derived based on lower and higher ATE percentiles. For details on the technique see D. James Greiner & Cassandra Wolos Pattanayak, *Randomized Evaluation in Legal Assistance: What Difference Does Representation (Offer and Actual Use) Make?*, Yale L.J. 2118, nn. 2149–2152 (2012) (applying Bayesian Peters-Belson to estimate causal effects of an offer of legal aid representation on beneficiary legal outcomes). Following Greiner and Pattanayak, I used the "MCMCpack" library in R to run these numbers. Andrew D. Martin et al., *MCMCpack: Markov Chain Monte Carlo in R*, 42 J. Stat. Software 1, 1–21 (2011). ↩

161. This is based on all 10,281 circuit and state cert. petitions. Percentages are rounded to two decimal places. ↩

162. Two particular problems are "separation", where one or a few variables perfectly predict minority outcome and thus preclude many statistical models from running to fruition, and the "k>>n" problem, which is when the number of variables we must consider far exceed the number of (minority outcome) observations we have. *See generally* Hui Zou & Trevor Hastie, *Regularization and Variable Selection via the Elastic Net*, 67 J. Royal Stat. Soc'y 301 ([Royal Statistical Society, Wiley] 2005) (describing k>>n and statistical countermeasures); Christopher Zorn, *A Solution to Separation in Binary Response Models*, 13 Pol. Anal. 157 (2005) (describing separation and statistical countermeasures). ↩

163. Elizabeth A. Stuart, *Matching Methods for Causal Inference: A Review and a Look Forward*, 25 Stat. Sci. 1, 2 (2010) ("[e]ven if the outcome values are available at the time of the matching, the outcome values should not be used in the matching process, to preclude the selection of a matched sample that leads to a desired result – or even the appearance of doing so") (citation omitted). ↩

164. That is, taking an interval scaled by a t-value and the two-sample standard deviation around the difference in means. ↩

165. *See supra* Part 2.2 (noting that interventional causal questions cannot be answered by observational data *alone*); *supra* Part 4 (noting the problems with extracting causality from judicial opinions). ↩

166. *See supra* note 6 and accompanying next. My results support the critiques of unbalanced regressions raised there. ↩

167. Notice that even if the Supreme Court had made all petition briefs freely available, that would not change how the vast majority of cert. petitions get denied. A larger dataset, therefore, does not escape outcome imbalance. However, that the greater number of granted petitions that should come with a larger sample *might* alleviate the constraints I faced here to the extent that causal inference becomes legitimate. I aim to obtain and analyze a larger petition dataset in future work. ↩

168. Jr, *supra* note 1, at 477. ↩

169. *See supra* note 133 and accompanying text. ↩

170. Matthew Honnibal & Ines Montani, SpaCy 2 2017. ↩

171. Steven Bird et al., Natural Language Processing with Python (2009). ↩

172. Fabian Pedregosa et al., *Scikit-Learn: Machine Learning in Python*, 12 J. Mach. Learning. Res. 2825 (2011). ↩

173. Even the most cutting edge PDF extraction software makes mistakes, as text in PDFs are not stored as text but pixel coordinates. For example, “court” might be wrongly extracted as “èourt” due to visual similarities. Setting a minimum document frequency helps remove most (though not all) of these processing artifacts. ↩

174. Zhe Sun, Sparse-Dot-Topn, “0.2.9,” GitHub, https://github.com/ing-bank/sparse_dot_topn. (last visited July 30, 2020). ↩

175. *See* Mozer et al., *supra* note 11, at 9. ↩

176. Ben Miroglio, Pymatch: Matching Techniques for Observational Studies, “0.3.4” (2019), GitHub, https://github.com/benmiroglio/pymatch. (last visited July 30, 2020). ↩

177. *See supra* Part V.C.3. ↩