Targeting Inconsistency: The Why and How of Studying Disagreement in Adjudication[*]

Ryan Copus[†] and Ryan Hübert[‡]

July 21, 2016

Abstract

Much of the judicial decision making research has focused on inter-judge inconsistency. How much more likely is a panel of all Democratic appointees to decide in favor of civil rights plaintiffs than is a panel of all Republican appointees? How does the presence of a black judge on a panel affect the likelihood that an affirmative action plan will be found constitutional? How large are the inter-judge disparities in asylum grant rates? How inconsistent are social security disability decisions? But despite the attention paid to disagreement, the vast majority of research shares a peculiar feature: it ignores much of judicial disagreement. By coding decisions on a single dimension (*e.g.*, liberal versus conservative or grant versus deny) and grouping judges according to demographic characteristics, researchers are mediating the study of inconsistency through intuitions about the nature of judicial disagreement. For reasons of both theory and policy, we argue for an agnostic, data-driven approach to the study of inconsistency. Using a newly collected dataset of appellate review in the Ninth Circuit, we show how machine-learning methods can help researchers make progress on previously identified challenges to the study of inconsistency. We also briefly discuss how our approach to inconsistency could be used to evaluate adjudication systems and changes to those systems as well as address debates between legal realists and formalists.

DRAFT PREPARED FOR
2016 CONFERENCE ON EMPIRICAL LEGAL STUDIES

[*] We thank participants at the 2016 meeting of the Midwest Political Science Association and the 2016 Political Economy and Public Law conference as well as Joy Milligan, Alan Kluegel, and Aaron Smyth for helpful comments. All errors and omissions are our own.

[†] Ryan Copus is a Ph.D. candidate in Jurisprudence and Social Policy and an M.A. candidate in biostatistics at the University of California, Berkeley. Email: `rwcopus@gmail.com`.

[‡] Ryan Hübert is an assistant professor of political science at the University of California, Davis. Email: `rhubert@ucdavis.edu`.

> *A number of blind men came to an elephant. Somebody told them that it was an elephant. The blind men asked, 'What is the elephant like?' and they began to touch its body. One of them said: 'It is like a pillar.' This blind man had only touched its leg. Another man said, 'The elephant is like a husking basket.' This person had only touched its ears. Similarly, he who touched its trunk or its belly talked of it differently.*
>
> – Ramakrishna Paramahamsa

Consistency in judicial decision making has been a central theme of legal scholarship. It is a staple in debates about the rule of law, legal certainty, stare decisis, rules versus standards, realism versus formalism, critical legal studies, *ex post facto* laws, fairness in the criminal justice system, and the very structure of courts. It is intimately tied to our notions of justice, to accuracy in decision making, and to predictability in law that allows parties to plan and coordinate their activities.

The empirical literature on inconsistency in adjudication is rapidly accumulating. We now have studies of inter-judge disparities in asylum cases, in social security disability awards, criminal sentencing in the federal courts, the Patent and Trademark Office, and in nursing home inspections. Political scientists have also amassed a vast literature on the United States Courts of Appeals, detailing how judges' political party, race, and gender are associated with decisions. Some of the findings have been unsettling. Evidence of inconsistency in criminal sentencing, for example, facilitated the development of the Federal Sentencing Guidelines (Stith and Cabranes 1998). And large disparities in asylum grant rates have given rise to calls for institutional reform (Ramji-Nogales, Schoenholtz, and Schrag 2007), as have the findings of inconsistency in the Courts of Appeals (Tiller and Cross 1999).

But these studies understate, and sometimes misrepresent, the true extent and nature of inconsistency—they touch only small parts of the elephant. By comparing decision rates on a single, intuitively specified dimension, scholars are potentially missing much of inter-judge disagreement (Fischman 2014). In brief, a judge's decision making may vary across different kinds of cases, and the nature of that variation might also depend on the judge she is being compared to.[1] For example, two judges might have identical reversal rates overall but have very different reversal rates in subsets of cases: one judge may reverse more often when the plaintiff prevails, while the other may reverse more often when the defendant prevails. We could, as some researchers have done, subjectively code certain types of reversals differently from other types (*e.g.*, in employment discrimination cases code a reversal as liberal/conservative if a defendant/plaintiff won at the trial level). And while we might fully capture the disagreement between Judge A and Judge B by changing the analyzed outcome to liberal rather than reversed, that coding scheme might poorly describe the form of disagreement between Judge A and Judge C. Moreover, such a time-consuming effort may miss important distinctions or stress distinctions that are not actually relevant.

If we are to understand inconsistency, monitor it, and intelligently reduce it, we will need to uncover as much of it as possible. In this paper, we show how rapidly expanding data can be coupled with machine learning to aggressively reveal inconsistency. Our trick, ultimately, is simple. We take the daunting problem of uncovering inconsistency and transform it into a task that statisticians and computer scientists are getting very good at: prediction. For any comparison of two judges, Judge A and Judge B, we show that the ultimate goal is to partition cases into two sets—the set of cases that Judge A is more likely to reverse and the set of cases that Judge B is more likely to reverse. By using validation-set predictions generated with machine-learning algorithms, researchers can make the best partition possible. This approach has the additional benefit of eliminating finite sample bias from measures of inconsistency that involve multiple comparisons of judges (Fischman 2014). We also provide a way to handle those contexts where the number of cases decided by individual judges is too small to allow for individual comparisons. Our method groups judges by the similarity of their actual voting patterns rather than their *ex ante* shared traits, aggregating judges so as to capture the dimensions of disagreement that are present in the data.

[1] In essence, there are heterogeneous treatment effects.

We illustrate our approach with an analysis of a newly created dataset of Ninth Circuit civil cases. With the population of docket sheets between 1995 and 2014, we extracted a large collection of case characteristics, including the judicial votes, the outcome, party information, and case type. We show how our machine-learning approach reveals more inconsistency than either a simple comparison of reversal rates or an approach that utilizes the common intuition that judges appointed by Democratic Presidents are more likely to favor plaintiffs in civil cases.

1 Legal Inconsistency

While consistency in decision making is sometimes treated like an end in itself, its importance is better understood through its connection to the values of comparative justice, accuracy, and predictability. These values can be helpfully thought of as truth-by-coherence, truth-by-correspondence, and truth-neutral values, respectively. We briefly discuss each, but we note that our discussion is necessarily limited; there is a vast and sophisticated literature on the normative foundations of consistency.

Comparative justice is perhaps the commonly cited benefit of consistency in decision making. It can be summed up in the adage that we should "treat like cases alike." But this pithy way of describing the benefits of consistency is merely a way of re-describing consistency, thus treating it as a good in itself. Instead, the substance behind the intuition in favor of treating like cases alike (apart from the benefits to accuracy and predictability discussed below) is probably best located in the desire to prevent discrimination. While the absence of inconsistency in no way guarantees an absence of discrimination (*e.g.*, all judges might consistently disfavor black litigants), inconsistency provides evidence that judges have room for mischief. We worry that judges can use that discretion to advance their biases or parochial interests, often in ways that are hard to detect.

Consistency is also associated with accuracy. This is perhaps the most under-appreciated benefit of consistency, probably due to the awkwardness in speaking of many types of legal decisions as "correct." But there is no need to enter that thicket to understand the benefits of consistency to accuracy if we think of it simply as treating better cases better: lower-risk inmates should be paroled before higher risk inmates, more negligent defendants should be found liable before less negligent defendants, and those refugees more likely to be persecuted should be granted asylum before those who are less likely to be persecuted. Again, consistency does not guarantee that judges are treating better cases better, but there are good reasons to treat it as a good indicator of such. If, as we must hope is the case, judges tend to successfully identify better cases, then inconsistencies between judges suggest an adjudication system is failing to treat better cases better (see Laqueur and Copus 2016).

Finally, consistency in adjudication makes the application of law more predictable. As Justice Brandeis famously stated, "in most matters it is more important that the applicable rule of law be settled than that it be settled right." If outcomes are invariant to the judge assigned to hear a case, then people and businesses could better and more easily predict how the law would be applied, allowing them to better and more cheaply plan and coordinate their activities. The lawyer's task can be understood as "the prediction of the incidence of the public force through the instrumentality of the courts," and inconsistency makes that task more difficult (Holmes Jr. 1897). Rather than trying to understand the patterns, precedent, and reasoning of a single, coherent decision making unit (the court), the lawyer must understand the idiosyncracies of the system's constituent parts (the judges). Moreover, the problem posed to prediction by inconsistency is not only one of complexity. Not even an omniscient (and free) lawyer can eliminate legal uncertainty in the presence of judicial disagreement: one could be assigned to the judge who always says yes or the judge who always says no. This ineradicable uncertainty is costly for risk-averse clients. Finally, legal uncertainty also has unsettling political implications. We might like to believe that the composition of judicial policy preferences in courts tends to aggregate in such a way that the composition is fairly reflected in social policy. But for most areas of litigation, a finding that one has failed to comply with the law is accompanied by a sharp discontinuity in costs: liability for tort damages or the need to suddenly abandon an affirmative action plan. Thus, extremist judges who also favor liability in an area will tend to have outsized policy influence in a world where potential litigants will have an incentive to tailor their actions in accordance with the extremist so as to avoid that discontinuity in costs.[2]

There are also, of course, costs to pursuing consistency in decision making (Legomsky 2007). The Federal Sentencing Guidelines stand as a prime example. While we express no opinion as to whether the benefits of consistency exceed the costs, it is clear that there are costs. Critics complain that the Guidelines are too rigid, giving judges too little discretion to shape sentences in accordance with the nuances of the case. Critics also argue that the Guidelines are simply too harsh, and that thus require judges to make systematically bad decisions. Efforts to increase consistency might also offend fairness intuitions (Lind and Tyler 1988), cement or inflate biases Laqueur and Copus (2016), or have other unintended consequences (*e.g.*, critics argue that the Guidelines simply placed discretion with prosecutors instead of judges).

Despite the importance of and extensive literature on inconsistency, there was no systematic framework for its empirical study before Fischman (2014). Most importantly, the paper makes clear that estimates of inconsistency can only be lower bounds on the true level of inconsistency. For example, if one judge reverses 20% of cases and another reverses 40% of cases, they *at least* disagree on 20% of cases,[3] but they could disagree on more if they tended to reverse different types of cases. As Fischman points out, this poses a serious problem if one wants to evaluate the effects of institutional reforms, as any changes in lower bounds

[2] This is an application of the Steve Shavell's analysis of uncertainty in legal standards (Shavell 2007).

[3] For simplicity, we assume infinite sample size and randomization of cases to judges.

may not reflect changes in the true level of inconsistency.

Fischman suggests ways of improving the bounds. Most similar to our approach, he proposes estimating disparities in subsets of cases in order to capture more heterogeneous treatment effects. But it is not clear how one would efficiently choose these subsets so as to maximize the lower bound of inconsistency. Moreover, it is not clear which variables one would use to subset the data or how fine the subsets should be—too fine and the problem of finite sample bias would be unmanageable, too coarse and substantial amounts of inconsistency could remain masked. He also suggests using surveys to measure judicial inconsistency, as they would allow simultaneous observation of judges' decisions. Of course, as he acknowledges, surveys have their own serious limitations, primarily external validity.

In this paper, we show how scholars can use expanding data to aggressively uncover additional inconsistency. We show how the advances in machine learning and data collection can help us better push measures of inconsistency toward the truth. Nevertheless, we acknowledge that precisely tracking inconsistency—even with the aid of big data and machine learning—is daunting. But it is not futile. If we are to intelligently improve our adjudication systems, we will need the most informative measures of inconsistency as possible, and even if observational studies of inconsistency are never by themselves informative enough, they will undoubtedly be part of a triangulated approach to understanding inconsistency.

2 Our Approach

A decision making body can be defined by four components. First, there is a finite set of decision makers, which we label $\mathcal{J} = \{1, ..., J\}$, and index by j. We allow decision makers to be groups, setting aside the micro-foundations of group decision making.[4] Second, there is a finite set of cases about which decisions are made, which we label $\mathcal{N} = \{1, ..., N\}$, and

[4] For an overview of the issues raised by group decision making, see chapter 2 of Persson and Tabellini (2000). For an application to appellate courts, see Landa and Lax (2009).

index by i. For a subset of cases decided by a specific decision maker $j \in \mathcal{J}$, we write $\mathcal{N}_j = \{1, ..., N_j\}$. Third, there is a set of possible decisions that could be made for each case, which we label Y.[5] For example, Y could be dichotomous, such as plaintiff/defendant or reverse/affirm, or continuous, such as the length of criminal sentences. Finally, there is a decision making function, which is a function mapping sets of decision makers and cases into outcomes, which we denote as $\mathcal{C} : \mathcal{J} \times \mathcal{N} \to Y$. For example, the decision making function for a set of district judges would specify how each judge would rule on each case. Since we only observe actual decisions, we treat $\mathcal{C}$ as a black box and focus on measuring observable patterns in decision making.

Given these components, we can formally define inconsistency in a decision making body. We first must differentiate between two distinct concepts: disagreement and inconsistency. We define *disagreement* to be the proportion of cases where two decision making units (individual decision makers or sets of decision makers) would come to different decisions:

$$\delta(j, k) = \mathrm{E}\Big[\mathbf{1}_{(D,\infty)}\big[d(Y_i(j), Y_i(k))\big]\Big] \tag{1}$$

where $\mathbf{1}$ is the indicator function, $j, k \in \mathcal{J}$ are two different decision makers, $d(\cdot)$ is a metric on Y, and D is a scalar.[6] Since we implement our method using data from the Ninth Circuit, where we treat decisions as binary (*i.e.*, affirm or reverse the lower court's decision), we will assume that $Y = \{0, 1\}$ and we will use the usual Euclidean metric on $\mathbb{R}$. We can thus rewrite equation (1) as:

$$\delta(j, k) = \mathrm{E}\Big[|Y_i(j) - Y_i(k)|\Big] \tag{2}$$

[5] We could allow the set of decisions to depend explicitly on the case, but here we assume that the set of possible decisions that could be taken for each case is constant.

[6] For example, suppose $Y = [0, 100]$. Then, we might consider two decisions different from one another if they are more than ten units apart. Formally, $d(Y_i(j), Y_i(k)) = |Y_i(j) - Y_i(k)| \leq 10 = D$.

In a decision making body with two decision makers, $\delta(\cdot)$ would completely characterize the amount of inconsistency that exists among the decision makers. However, with more than two decision makers, we must define a composite measure based on the disagreement between each pair. There are many ways to characterize this quantity, but following Fischman (2014), we focus on two: average inconsistency and extreme inconsistency. Define $\mathcal{P}$ to be the set of decision maker pairs: $\mathcal{P} \equiv \mathcal{J} \times \mathcal{J}$. Then, *average inconsistency* is defined by

$$\Delta_a = \mathrm{E}\left[\delta(j,k)\right] \tag{3}$$

and represents the average level of disagreement between the decision makers. Intuitively, how many decisions would be made differently if all the cases were re-randomized?[7] *Extreme inconsistency* is formally defined by

$$\Delta_e = \max\{\delta(j,k) : (j,k) \in \mathcal{P}\} \tag{4}$$

and represents the disagreement between the two decision makers who are the most dissimilar in their decision making. This quantity can be viewed as bookend normative benchmark, as it captures how high disagreement *could* be.

Limits of Traditional Estimands

As is apparent from equation (1) we express disagreement and inconsistency in idealized terms. In particular, disagreement between two decision makers, j and k, is measured across all cases, even though only only one decision maker can sit on a particular case i. To put this in terms of the Neyman-Rubin causal model, $Y_i(j)$ and $Y_i(k)$ are potential outcomes, where

[7] As we describe in greater detail below, we characterize average inconsistency differently than Fischman (2014). Briefly, we allow that cases could be re-randomized to *the same* panel, whereas Fischman (2014) assumes that cases are re-randomized to different panels.

one decision maker is considered the “control” condition and one decision maker is considered the “treatment” condition. As long as the assignment of decision makers is random, we have sufficient sample size, and we invoke the Stable Unit Treatment Value Assumption,[8] then standard methods allow us to generate an unbiased estimate of the average treatment effect (ATE, which we denote as $\phi(j,k)$):

$$\phi(j,k) = \mathrm{E}\big[Y(j) - Y(k)\big]$$

In fact, the major *methodological* benefit of studying ATEs is that they are relatively easy to estimate once we satisfy these few assumptions. Specifically, due to the linearity of the estimand, we can decompose it into $\mathrm{E}[Y(j)] - \mathrm{E}[Y(k)]$, which allows us to simply compare the means of the treatment and control groups.

Indeed, when comparing the decisions of decision makers, researchers usually use this analytical framework to estimate an ATE (or some related quantity, such as regression coefficients).[9] For example, how many more pro-civil rights decisions does an all Democratic panel of judges make as compared to an all Republican panel of judges? Or, how many more asylum applications does Asylum Officer 1 grant than Asylum Officer 2? Such research questions are at least implicitly concerned with measuring the extent of disagreement between decision makers or types of decision makers. However, ATEs can systematically understate actual disagreement, as well as the extent of inconsistency in a decision making system.

Consider an example comparing two panels in the Ninth Circuit. To make the example more salient, suppose one panel is an all Democratic-appointee panel and the other is an all Republican-appointee panel. Moreover, suppose an analyst is interested in studying how

[8] Assignment of treatment to one unit does not affect the outcomes for an entirely different unit. Essentially, the treatment effect does not spill across units.

[9] There are more examples than we can reasonably list here, but several recent ones are particularly noteworthy. See, for example, Revesz (1997), Anderson, Kling, and Stith (1999), Cockburn, Kortum, and Stern (2003), Farhang and Wawro (2004), Ramji-Nogales, Schoenholtz, and Schrag (2007), Boyd, Epstein, and Martin (2010), and Kastellec (2013).

Democratic appointees and Republican appointees decide civil rights cases differently. If she were to try to measure the extent of disagreement between these two panels, the appropriate estimand would be derived directly from equation (2). We label the estimand for equation (2) by $\delta(j, k)$, and in this example it is:

$$\delta(DDD, RRR) = \mathrm{E}\big[|Y(DDD) - Y(RRR)|\big]$$

Of course, estimation of this quantity is complicated by the fact that the expectation operator cannot be linearly decomposed. If instead, the analyst estimates an ATE—which is easier to estimate—then her estimand is

$$\phi(DDD, RRR) = \mathrm{E}[Y(DDD)] - \mathrm{E}[Y(RRR)]$$

Unfortunately, $\phi(DDD, RRR)$ would be a downward biased estimand for disagreement. In Proposition 1, we show generally that[10]

$$\phi(j, k) \leq \delta(j, k). \tag{5}$$

Moreover, equation (5) holds strictly whenever there are strong heterogeneous treatment effects as defined in Definition 1 in the appendix. Intuitively, an ATE will always understate disagreement if the treatment has a positive effect in some cases and a negative effect in others. In our example, if Y is whether a civil rights case is reversed, then an ATE comparing DDD and RRR panels would understate the level of disagreement between Democratic panels and Republican panels if Democrats are more likely to reverse than Republicans when the defendant won in the lower court but less likely to reverse than Republicans when the plaintiff won in the lower court.

[10] Fischman (2014) demonstrates how measures of inconsistency are always interval-defined, so Proposition 1 can be seen as alternative expression of results from that paper.

This problem does not disappear by re-coding the outcome variable. For example, an analyst might recode the outcome variable to be pro-plaintiff/pro-defendant instead of reverse/affirm. While this generates a valid measure of the difference in rate of pro-plaintiff decisions for the two types of panels, it still does not capture disagreement. Suppose, for example, that the two panels differ in their propensity to reverse lower court decisions: DDD panels always reverse the lower court decision, and RRR panels never do. Moreover, suppose cases won by the plaintiff are appealed as often as cases won by the defendant. Then, the rate of pro-plaintiff decision making by both types of panel is 0.5. An analyst would observe an ATE of zero, potentially concluding that the two panels disagree very little. In fact, they perfectly disagree: *in every case, the panels rule differently.*

As a general principle, estimating ATEs using different outcome variables will reveal different amounts of disagreement between decision makers. The reason is fairly straightforward: each outcome variable reflects disagreement on different dimensions of the decision makers' utility functions. If, for example, preferences about deferring to lower courts is the primary dimension on which Ninth Circuit judges disagree, then reversal rates will be a better measure of disagreement than whether the plaintiff or defendant ultimately prevail. But, analysts almost never know *ex ante* which outcome best captures disagreement. The estimation of a specific ATE represents a small, and specific, bite of the apple, and may even lead researchers to draw faulty theoretical conclusions. Yet, some ATEs may do a better job of capturing disagreement. For example, the treatment could partition the decision makers into groups that are "most like-minded" and the outcome of interest could be the issue on which the groups of judges disagree most. But, since decision making differs across many possible dimensions, an ATE based on a particular treatment and particular outcome derived intuitively will yield a poor proxy for the overall level of disagreement between judges.

In light of this problem, we reformulate the analysis of decision making as a prediction problem with the goal of backing out the dimension characterizing the most disagreement.

Getting Around The Problem: Heterogeneous Treatment Effects

One way to view the problem we identify is that there are heterogeneous treatment effects (HTEs, see Athey and Imbens 2015; Grimmer, Messing, and Westwood 2016; Bullock, Green, and Ha 2010). In the context of medicine, for example, a doctor who knows that a particular drug has a positive average treatment effect, may also wish to know which patients respond positively, which respond negatively, and which do not respond at all. Such information, which is thrown away by the particular way that treatment effects are aggregated, has important clinical applications and can help doctors understand better how a drug works. Define $\mathcal{M}$ to be a partition of $\mathcal{N}$ that represents a partition of the covariate space and where each $M \in \mathcal{M}$ is nonempty. Then, the estimand of interest is the conditional average treatment effect (CATE):[11]

$$\phi(j,k,M) = \mathrm{E}_M[Y(j) - Y(k)] \tag{6}$$

As Grimmer, Messing, and Westwood (2016) point out, if $\phi(j,k,M)$ varies as M does, then there are heterogeneous treatment effects. In our context, such variation is informative because it allows us to observe how often treatment effects are non-zero, which maps into disagreement. Disagreement for a specific partition $\mathcal{M}$ is:

$$\begin{aligned} \delta(j,k,\mathcal{M}) &\equiv \mathrm{E}_{\mathcal{M}}\big[|\mathrm{E}_M[Y(j) - Y(k)]|\big] \\ &= \mathrm{E}_{\mathcal{M}}\big[|\phi(j,k,M)|\big] \end{aligned}$$

This approach helps researchers avoid making problematic assumptions on the joint dis-

[11] Another equivalent way to write the CATE is

$$\phi(Y,T,x) = \mathrm{E}[Y(T=1) - Y(T=0)|X=x]$$

where X is a vector of covariates and x is a particular value of covariates.

tribution of the potential outcomes by bringing the absolute value outside the expectation operator, thus allowing for "traditional" estimation of average treatment effects. The downside to this procedure is that it is highly dependant on the method used to partition the parameter space (*i.e.*, the choice of $\mathcal{M}$). At the limit, $\delta(j, k, \mathcal{M})$ becomes $\delta(j, k)$ as the partition becomes fine enough such that the average treatment effect is estimated for each unit separately (*i.e.*, as $\mathcal{M}$ approaches $\mathcal{N}$). Of course, the fundamental problem of causal inference rules out this possibility (Holland 1986), but one could implement matching to find the closest match for every treated (or control) unit. With a large enough sample size, one could find suitable matches, but the estimation of the average treatment effect for each matched pair would introduce unmanageable finite sample bias.

For practical reasons, an analyst must choose a partition $\mathcal{M}$. Usually, when they are interested in studying heterogeneous treatment effects, the choice of $\mathcal{M}$ is driven by *a priori* theoretical concerns. In fact, recent work on estimating HTEs in the context of randomized experiments has even cautioned against post hoc subgroup analyses (see Pocock et al. 2002; Imai and Strauss 2011). However, we are not interested in HTEs *per se.* What we wish to know is not *what is the treatment effect (among subgroups)?*, but rather *is there a treatment effect (among subgroups)?* Thus, our fundamental task is to pick $\mathcal{M}$ to maximize $\delta(j, k, \mathcal{M})$,[12] and since $\delta(j, k, \mathcal{M})$ is always biased downward due to the averaging of heterogeneous effects, we can optimally partition the parameter space into a partition $\mathcal{M}^*$ with exactly two subsets:

$$M^+ = \{i \in \mathcal{N} : Y_i(j) \geq Y_i(k)\} \qquad\qquad M^- = \{i \in \mathcal{N} : Y_i(j) < Y_i(k)\}$$

[12] To see why, see Proposition 2 in the appendix.

Then, our estimand for disagreement can be written as

$$\begin{aligned}\delta(j,k,\mathcal{M}^*) &= \mathrm{E}_{\mathcal{M}^*}\left[|\phi(Y,T,M)|\right] \\ &= \mathrm{E}\left[Y(j)-Y(k)|Y(j)\geq Y(k)\right]\Pr[Y(j)\geq Y(k)] \\ &\quad + \mathrm{E}\left[Y(k)-Y(j)|Y(j)<Y(k)\right]\Pr[Y(j)<Y(k)]\end{aligned} \tag{7}$$

Finally, given that our ultimate goal is to estimate inconsistency in an entire decision making body and disagreement only measures inconsistency between two decision makers, the appropriate estimands for average and extreme inconsistency can be written as follows:[13]

$$\Delta_a = \mathrm{E}_{(j,k)\in\mathcal{P}}\left[\delta(j,k,\mathcal{M}^*)\right] \qquad \Delta_e = \max\{\delta(j,k,\mathcal{M}^*) : (j,k)\in\mathcal{P}\} \tag{8}$$

Estimation

We have shown that ATE-based estimands of disagreement will always be biased downward. Our goal is to significantly reduce this bias to provide better measures of inconsistency. However, as is apparent from the foregoing discussion, we can only reduce bias by subsetting the parameter space, thus reducing sample sizes and increasing variance. Our estimation challenge is therefore to find the optimal bias-variance trade-off. We face three specific problems in estimation, what we refer to as the problems of *partitioning*, *clustering* and *finite sample bias.*

The partitioning problem refers to the challenge of optimally selecting $\mathcal{M}$ to reduce bias in the estimates for $\delta(j,k)$. The problem is both theoretical and practical. In the previous section, we derive an estimand with the most efficient partition $\delta(j,k,\mathcal{M}^*)$, see equation (7). Because we need only divide our sample into observations where $Y(j)\geq Y(k)$

[13] Our estimand for average inconsistency differs slightly from the estimand in Fischman (2014). He estimates the percentage of cases that would be decided differently if all cases were reassigned to *different* decision makers. We think a more accurate measure of inconsistency is one that allows for the possibility that a case could have been reassigned to the same decision maker or type of decision maker.

and $Y(j) < Y(k)$, our partition is as coarse as possible thus increasing variance by as little as possible (relative to the baseline ATE). The practical problem is how to classify observations into the two sets of the partition. We treat this as a prediction problem and use novel machine learning methods to generate estimates of $Y_i(j)$ and $Y_i(k)$ for all i that were decided by either j or k. We label these $\widehat{Y}_i(j)$ and $\widehat{Y}_i(k)$. Advances in machine learning allow us to generate predictions with less error than previous approaches. In our estimation, we use `Super Learner`, which is an ensemble method that uses a set of constituent algorithms to predict outcomes in the data. Specifically, it constructs a weighted model of these constituent algorithms that generates the best predictions, as measured by mean squared error. As we show in Table 3, our `Super Learner` model outperforms each constituent algorithm. See Appendix D for more information.

`[Table 3 about here.]`

Until now, our discussion has focused heavily on the estimation of disagreement, $\delta(j,k)$, but not inconsistency. To measure inconsistency, we need to define the set $\mathcal{P}$, which is the set of pairwise comparisons we wish to study. In the context of experiments, this is known as the choice of the treatment arms. This is what we call the clustering problem. In some contexts, the clustering problem is not actually a problem. For example, suppose we have a decision making body with three decision makers, A, B, and C, who each make decisions on 1,000 cases. If we had data from all 3,000 decisions, we would have sufficient sample size to efficiently estimate disagreement between each pair of decision makers: $\delta(A,B)$, $\delta(A,C)$ and $\delta(B,C)$. However, suppose instead that each decision maker makes decisions on just 50 cases (and we have 15 covariates in our prediction model). Then, our predictions would be extremely noisy and uninformative.

In the context of the Ninth Circuit (and many other decision making bodies), this poses a more serious problem since there are a small number of cases assigned to some of the treatments. For example, if we define each decision making unit as a specific three-judge panel, then even ignoring senior and designated judges, with 28 active judges there are $3,276$ possible panels. The population of cases is too thinly split among such a large number of decision-making units to allow for meaningful analysis. We therefore cluster judges into larger groupings to increase sample sizes in our prediction models. Essentially, we are re-defining the "treatment" and "control" to be panels of different *types* of judges, rather than specific judges. To be clear, our solution to the clustering problem explicitly opts for increased bias in order to decrease variance. We do so out of necessity, but are aware that the extent to which an analyst trades off bias for variance is context-dependent and discretionary. Below, we describe the details of our clustering procedure and note that different clustering procedures could yield more or less biased and more or less noisy estimates.

Finally, our approach solves the problem of finite sample bias. Variance can artificially inflate estimates of inconsistency. Imagine, for example, that two judges are actually identical in their decision-making, each reversing 30% of cases, but due to noise, one has a sample reversal rate of 25% and another of 35%. Traditional estimates of disagreement between judges, because they treat all observed differences as reflecting true differences, overstate inconsistency. Moreover, the problem can become more severe as researchers increase variance by subsetting in the search of more inconsistency. Fischman (2014) describes a method for adjusting inconsistency estimates for finite sample bias. Our approach, rather than correcting for finite sample bias, avoids introducing it in the first place. By using training sets to set our expectations for the direction of inter-judge differences *ex ante*, we allow noise to result in negative estimates of disagreement, eliminating variance's contribution to bias.

Given that we live in a world of finite data, our main contribution is our methodological solutions to the problems of partitioning, clustering, and finite sample bias. However, a secondary contribution is that we bring much more data to our analysis than scholars of the federal appellate courts usually do. The partitioning and clustering problems are more or less severe depending on the nature of the data. As a general principle, the more data available, the less severe the bias-variance trade-off is. The trade-off is also less severe when a decision making body has only a small number of decision makers who each decide a relatively large number of cases.[14] In our case, we estimate inconsistency in the Ninth Circuit using a newly and extensively coded dataset of *all* civil cases filed in the circuit over a period of nineteen years. By offering methodological solutions to the partitioning and clustering problems, we hope that scholars will more focused on aggressively assembling large and high quality datasets.

3 Ninth Circuit Data

To conduct our analysis, we constructed an original dataset on judicial decision making in the Ninth Circuit Court of Appeals. The Ninth Circuit is one of thirteen Courts of Appeals in the U.S. federal court system and it contains nine states and two overseas territories (see Figure 1). It hears appeals originating from the district courts located within these states and territories, as well as appeals of some agency decisions originating from within its borders (*e.g.*, decisions on immigration cases).

`[Figure 1 about here.]`

We compiled our dataset from the circuit's docket sheets covering every case filed in the court between 1995 and 2013, which we collected directly from PACER. In most U.S. courts, docket sheets are used to track the progress of cases, and they therefore serve as a court's administrative record of the procedural developments in each case. In the Ninth Circuit, a docket sheet contains background information on the case—such as area of law, trial judge and parties—and separate entries for each event in the case. In Appendix C, we present

[14] For example, 281,608 civil cases were filed in the U.S. District Courts in 2015 (United States Courts 2015, Table C). This is an average of roughly 415 cases for each authorized district judgeship per year.

an example docket sheet. They are an incredibly rich source of information and allow us to perform more detailed analyses than have been done in studies of judicial decision making in the federal courts.

Using standard text parsing methods implemented in `python`,[15] we extracted key information from each docket sheet: background information about the case (area of law and number of parties), information about the district court proceedings (judge, court and disposition), and information about the appeal (judge panel, disposition and opinion publication). We began with 217,273 docket sheets, of which 51,729 of them were appeals of civil cases. In this analysis we limit our attention to civil appeals.

Our data represents a significant improvement over other available datasets. Firstly, we have data for every single case filed in the Ninth Circuit for nearly twenty years, as opposed to (a) a much smaller random sample or (b) a potentially biased sample, such as published opinions. This first issue has limited scholars' ability to study heterogeneous effects (due to small sample sizes), but in principle it should not affect the validity of the effects that are estimated. However, the second issue poses a major challenge for empirical studies of courts. As Fischman (2015) points out: "[s]ome studies may also introduce correlated effects by selecting cases on the basis of endogenous characteristics, such as whether an opinion was published or whether it cited a particular precedent" (p. 812-813). Our data contains the universe of civil cases in the Ninth Circuit, which are (in theory) randomly assigned.[16]

Secondly, since we derive our data from docket sheets, we have access to a wide range of case-related data. Moreover, our machine-assisted coding methods allowed us to accurately and aggressively code variables not previously available to scholars of the Courts of Appeals. For example, our dataset contains information on the parties and their attorneys. We go beyond simple counts and even code *types* of parties.

[15] Code available upon request.

[16] As we discuss in Section 4, we find evidence that there are deviations from random assignment in the Ninth Circuit. We account for this in two ways, which we describe in detail.

Finally, our data is coded directly from court records, thus avoiding some of the potential problems associated with data collected from commercial database services, such as Westlaw or LexisNexis. In particular, these databases' primary clientele is practitioners, so the data is likely to be incomplete. With respect to docket sheets in particular, our cursory comparison between Westlaw's data and ours reveals that they do not keep all of their docket sheets up-to-date. One possible reason for this could be that these services stop updating docket sheets for cases that they determine to be unimportant for their customers. While scholars rarely assess the completeness of commercial databases, our unique dataset allows for such comparisons.[17]

For our analysis, we reduced our sample in two important ways. Firstly, because our main dependent variable is negative treatment of lower court decisions, we drop all cases that are not terminated on the merits, such as dismissed appeals. Second, because some of the cases may be related to one another (see Section 4), we batched similar cases together. To generate each observation in the batched dataset, we simply average over the constituent cases. These steps left us with a sample of 16,723 batched cases, on which our analysis is based.

In Table 4 and Table 5, we present some basic descriptive statistics for our sample. The second and third columns of each table provide summary statistics from our original, unbatched dataset while the fourth and fifth columns provide summary statistics for the batched dataset. Table 4 presents basic characteristics about the appeals, including outcomes (merits review, negative treatment, published opinions, etc.) and characteristics of the reviewing panels (party, race and sex). Since the negative treatment rate is our main dependent variable, we also present the negative treatment rate subset for each variable in each dataset. For example, appeals with all-Democrat panels reversed the district court's decision 37.7% of the time in the original unbatched dataset and 38.6% of the time in the

[17] This is an interesting avenue for future research.

batched dataset. Table 5 presents summary statistics about the trial characteristics.

`[Table 4 and Table 5 about here.]`

4 Method

In this section we describe a general method for estimating inconsistency and apply it to our Ninth Circuit dataset. As we describe in detail in Section 2, we treat the estimation of disagreement and inconsistency as a prediction problem. To generate the best possible prediction models of our data, we use novel ensemble machine learning methods, implemented with `Super Learner` (Laan, Polley, and Hubbard 2007). Our method consists of six basic steps.

1. Build Training-Set Models of Each Judge.
2. Cluster Judges and Apply Categorization to the Training and Test Sets.
3. Use Training Set to Generate Panel-Specific Predictions for the Test Set.
4. Code Test-Set Decisions for Each Pairwise Panel Comparison.
5. Identification Strategy.
6. Estimate Inconsistency Using the Test Set.

Steps 1 and 2 address the clustering problem, using the training set to cluster judges in accordance with their voting patterns rather than their demographic characteristics. Steps 3 and 4 address the partitioning problem, using estimated differences in panel-type voting patterns to partition our data. We note that Steps 1 and 2 will be unnecessary in systems where decision-making units each decide a large number of cases. Where this is true, there is no reason to waste data by using a designated training set to generate predictions for the test set. Researchers can instead conserve data via cross-validation, allowing parts of the data to successively serve as temporary test sets.

Step 1: Build Training-Set Models of Each Judge

We randomly sample 70% of the data for inclusion in the training set. The remaining 30% is reserved as a test set, which we use to undertake our main analysis. We contribute a greater share of the data to the training set because the tasks we use it for are more data intensive.

Using the training set, we construct a `Super Learner` model of voting for each appellate judge that sat on more than 70 (batched) cases. Each judge's model takes as inputs data about each case they sat on, and returns a predicted probability that the judge would vote to reverse the trial court decision. In a sense, the model allows us to characterize each judge's behavior over all their cases leveraging information about the panels they sat on. We include 6 candidate models in the `Super Learner`: a LASSO regression, the mean reversal rate, two user-specified linear regressions that we thought could capture voting patterns, a random forest, and boosted CARTs. Details are available in Appendix D.

It is worth noting that the researcher can be aggressive in this step. If we were presenting predictions from these models as *results* (*i.e.*, as claims about the state of the world), we would have to be concerned that they were an artifact of data mining, intentional or not. But partitioning the test set from the training set allows the researcher to combine the power of clinical judgment and mechanical algorithms—if we get too aggressive and find patterns in the data that reflect chance rather than reality, then applying that "finding" to the test set we set aside will tend to yield uninformative and null estimates. Since we will eventually use these judge-specific models to group judges, if they capture noise rather than signal, then they should not prove useful in the test set.

Figure 2 visualizes judge-specific models for some prominent jurists in the Ninth Circuit. To illustrate the potential benefits of a machine-learning approach, we highlight the relationship between the trial court winner and reversal likelihood. The models suggest that Judge Reinhardt and Judge Leavy decide cases in a highly inconsistent manner but that the magnitude of the overall inconsistency is entirely captured by a comparison of reversal rates—our model predicts that Judge Reinhardt is always more likely to reverse a case than is Judge Leavy. On the other hand, although Judge Pregerson and Judge Kleinfeld have similar overall reversal rates, we predict that they frequently reverse different types of cases. We predict that Pregerson is more likely to reverse when a defendant won in the lower court but less likely to reverse when a plaintiff won.

[Figure 2 about here.]

Step 2: Cluster Judges and Apply to Training and Test Sets

We use each judge's judge-specific model of voting (from Step 1) to generate a predicted probability for how they would have voted in each of the training-set cases, even for cases they did not sit on. Then, for each pair of judges, we calculate the mean absolute distance between the two judges' predicted votes. This is roughly interpretable as an estimate (albeit a very noisy estimate) of the percentage of cases for which that pair of judges would cast different votes. Using these pairwise distances between judges, we use standard cluster analysis to group judges.[18] Figure 3 shows the results of the cluster analysis.

[Figure 3 about here.]

The judges cluster pretty clearly into six groups. With more data, it might make sense to use a finer clustering, but anything more than six groups begins to stretch the data too thinly, leaving too few observations of each panel type to make statistically relevant comparisons.

[Tables 6 to 11 about here.]

We take the liberty of using the names of well known judges to label the clusters: judges are thus each identified as being part of the Reinhardt Cluster, the Leavy Cluster, the Kozinski Cluster, the Pregerson Cluster or the O'Scannlain Cluster. The exact membership and

[18] Specifically, we use the `hclust` package for hierarchical clustering in `R`. We use the `ward.D` method for its tendency to generate clusters of relatively equal size.

demographic characteristics of the groups are available in Tables 6 to 11. We also add a cluster that we label the Visiting Cluster. This cluster consists of judges who had fewer than 70 observations in the training set, most of whom are judges sitting by designation. Throughout the text, we refer to particular "types" of judges using formatted labels corresponding to the cluster names: `R`, `L`, `K`, `P`, `O` and `V`. For example, we refer to a judge from the O'Scannlain Cluster as a `O`-judge, and a panel of such judges as an `OOO`-panel.

Step 3: Generate Panel-Specific Predictions and Apply to Test Set

We reuse the training set and the same candidate models that we used to build the judge-specific models (Step 1) to generate *panel-specific* predictions for each case in the test set.[19] Figure 4 visualizes predictions for four of the panel types and highlights the potential for machine-learning approaches. The left panel suggests that the main source of disagreement between `LLL`-panels and `RRR`-panels is something related to judges' reversal proclivity. Even though `LLL`-panels appear to exhibit a pro-defendant leaning, `RRR`-panels do not. `RRR`-panels simply reverse *a lot* more than `LLL`-panels, regardless of whether the plaintiff or defendant won in the lower court. An analyst concerned about the mechanisms driving differences between `LLL`-panels and `RRR`-panels would infer that willingness to reverse is what actually differentiates decision making between `RRR`- and `LLL`-panels.

On the other hand, the right panel of the figure illustrates that substituting a `KKK`-panel for a `PPP`-panel is predicted to decrease the probability of reversal where a defendant won in lower court but increase the probability where a defendant lost. The figure suggest that a simple comparison of average reversal rates of `PPP`-panels and `KKK`-panels would understate

[19] Ideally, one would use a new training set to construct panel-specific predictions, as any noise that contributed to the grouping of judges could be compounded in the panel model, leading us to over-estimate differences between panel types. But we think it was more important to "spend" our data on the judge-specific models. Furthermore, any over-estimating of differences between panels will ultimately bias our test-set estimates of inconsistency downward.

Moreover, if Steps 1 and 2 are not necessary (as explained above), the panel-specific predictions can be estimated on the entire dataset.

the extent to which the two types of panels decide cases differently because they tend to reverse *different types of cases.* In other words, judges who are more similar to Kozinski tend to reverse defendant-wins less often than judges more similar to Pregerson, and vice versa. Next, in Step 4, we show how these predictions can be used to recode decisions in the test set so that we can more accurately estimate the actual level of inconsistency in adjudication.

`[Figure 4 about here.]`

Step 4: Autocode Test Set Outcomes for Each Pairwise Panel Comparison

As we discussed in Section 2, the presence of heterogeneous treatment effects means that a comparison of overall reversal rates between types of panels will lead one to underestimate the level of inconsistency in adjudication. For example, all-Republican panels and all-Democratic panels might have identical reversal rates, but all-Republican panels may be more likely to reverse civil rights cases when plaintiffs won in the lower court and all-Democratic panels may be more likely to reverse civil rights cases when defendants won. As a result, our challenge is to optimally partition the parameter space to capture the most amount of disagreement.

We seek to estimate $\delta(j, k, \mathcal{M}^*)$ for each pair of decision makers, $(j, k) \in \mathcal{P}$. Restating equation (7), our estimand is:

$$\begin{aligned}\delta(j, k, \mathcal{M}^*) = {} & \mathrm{E}\left[Y(j) - Y(k) | Y(j) \geq Y(k)\right] \Pr[Y(j) \geq Y(k)] \\ & + \mathrm{E}\left[Y(k) - Y(j) | Y(j) < Y(k)\right] \Pr[Y(j) < Y(k)]\end{aligned}$$

We allow `Super Learner` to acquire knowledge about how panel types decide cases. Specifically, we estimate potential outcomes on our training set, $\widehat{Y}_i(j)$ for all $i \in \mathcal{N}$ and all $j \in \mathcal{P}$. Then, for a given pairwise comparison (j, k), we code the outcome of a case as a j-decision or a k-decision depending on whether a j panel or k panel is predicted as more likely to have made the decision that was actually made (according to the model that we constructed with the training set). To implement this, we treat either j or k as the "treatment" and code a new outcome variable, which we refer to as "autocoded" and label $\widetilde{Y}_i^{j,k}(\cdot)$. Suppose we define j to be the treatment, then:

$$\widetilde{Y}_i^{j,k}(\cdot) = \begin{cases} Y_i & \text{if } \widehat{Y}_i(j) \geq \widehat{Y}_i(k) \\ 1 - Y_i & \text{if } \widehat{Y}_i(j) < \widehat{Y}_i(k) \end{cases} \tag{9}$$

One can interpret $\widetilde{Y}_i^{j,k}(\cdot)$ to be whether or not observation i had a j-like outcome. For example, if j is an LLL-panel and k is an RRR-panel, $\widetilde{Y}_i^{\texttt{LLL},\texttt{RRR}}(\cdot) = 1$ implies that observation i featured an outcome Y_i that was consistent with a model of LLL decision making, but not RRR decision making.

In Table 1, we illustrate the logic using a example with six hypothetical cases. The example assumes there are only two types of panels: RRR, and LLL. The table displays predicted probabilities for each panel type (which would be generated by applying our Step 3 model to the test set). Then, we automatically code an outcome variable based on the actual outcome and the predicted probabilities. For example, on Case 3 we predict a higher probability of reversal by an RRR panel than an LLL panel, so we code it as a RRR-leaning decision (*i.e.*, 1).

Table 1: Example: Calculating Inconsistency

			RRR	LLL	
	Actual	Actual	Model	Model	Autocoded
Case	Outcome	Panel	Preds.	Preds.	Outcome
i	Y_i	j	$\widehat{Y}_i(\texttt{RRR})$	$\widehat{Y}_i(\texttt{LLL})$	$\widetilde{Y}_i^{\texttt{RRR,LLL}}$
1	1	LLL	0.2	0.4	0
2	0	LLL	0.8	0.1	0
3	1	LLL	0.7	0.4	1
4	0	LLL	0.9	0.4	0
5	1	LLL	0.1	0.7	0
6	1	RRR	0.4	0.4	1
7	0	RRR	0.9	0.6	0
8	1	RRR	0.9	0.7	1
9	1	RRR	0.7	0.2	1
10	1	RRR	0.7	0.7	0

In our original dataset, we see that LLL panels have a reversal rate of 0.6, whereas RRR panels have a reversal rate of 0.8. However, using our autocoded outcome variable, we now see that these rates change to 0.2 and 0.6 respectively, underscoring more extreme differences between RRR and LLL panels in their decision making. Our autocoded outcomes are no longer interpretable as reversal rate, but instead incorporate information about the "expected" decision that would have been made (based on our model in Step 3).

Notice that our autocoded outcome estimates $\delta(j, k, \mathcal{M}^*)$ by splitting our data for each pairwise comparison of panels into four groups:

$$\widehat{M}_j^+ \equiv \{i \in \mathcal{N}_j : \widehat{Y}_i(j) \geq \widehat{Y}_i(k)\} \qquad \widehat{M}_j^- \equiv \{i \in \mathcal{N}_j : \widehat{Y}_i(j) < \widehat{Y}_i(k)\}$$

$$\widehat{M}_k^+ \equiv \{i \in \mathcal{N}_k : \widehat{Y}_i(j) \geq \widehat{Y}_i(k)\} \qquad \widehat{M}_k^- \equiv \{i \in \mathcal{N}_k : \widehat{Y}_i(j) < \widehat{Y}_i(k)\}$$

Our estimator is therefore:[20]

$$\widehat{\delta}(j,k,\mathcal{M}^*) = \frac{1}{N_j}\left(\sum_{\widehat{M}_j^+} Y_i + \sum_{\widehat{M}_j^-}(1-Y_i)\right) - \frac{1}{N_k}\left(\sum_{\widehat{M}_k^+} Y_i + \sum_{\widehat{M}_k^-}(1-Y_i)\right)$$

In Step 6, we go into more detail about how we estimate our composite measures of inconsistency, Δ_a and Δ_e, using this procedure.[21] Moreover, we show that the automated partitioning that we adopt in this paper substantially outperforms both an approach that simply compares overall reversal rates and an approach that relies on an intuition that liberals generally favor plaintiffs.

Step 5: Identification Strategy

Our analysis relies on an assumption of statistical independence between cases and judge panels. We require this assumption since our goal is to isolate the *specific* effect of judges on outcomes. If this were not the case, our estimates of each judge's effect on case outcomes could simply reflect differences in the types of cases that judges are assigned.

One potential methodological advantage of studying decision making in federal courts is that cases are generally assigned to judges at random. However, many past studies of decision making in appellate courts are effectively unable to exploit this randomization due to biases introduced in the selection of their samples. For example, studies of published opinions can no longer treat panel assignment as if it were random, since the judges themselves decide whether to publish their opinions. Since our data includes the entire population of cases filed in the Ninth Circuit, we avoid some of the pitfalls of previous studies and better exploit the assignment of judge panels in the circuit.

[20] See Proposition 3, where we prove that our autocode procedure generates an equivalent estimand to $\delta(j,k,\mathcal{M}^*)$.

[21] Recall, the estimate of $\widehat{\delta}(j,k,\mathcal{M}^*)$ is for a pairwise comparison of j and k type panels. We seek an *overall* measure of inconsistency, where we incorporate the $\widehat{\delta}(j,k,\mathcal{M}^*)$ for each pairwise comparison.

However, examination of our data and conversations with a former clerk in the Ninth Circuit revealed two potential threats to identification. As is standard practice in other courts, related cases may be grouped together and assigned to the same panel.[22] To control for this possibility, we batched all of our cases by panel, area of law and year. For example, if Judges Reinhardt, Kozinski and Paez served together on four Fair Labor Standards Act cases in 2004, we would batch these cases into a single observation. Ideally, we would batch cases we *know* to have been batched after randomization, but our data does not allow us to observe this. However, the batching rule we used is conservative, in the sense that we may be *over*batching but we are not *under*batching. The former simply reduces our sample size beyond what is necessary, whereas the latter would undermine the panel randomization.

Since litigants in a suit may settle at any point, a second threat to identification could be strategic settlement by litigants after a panel is randomly chosen and revealed to the litigants. Others have argued that this is not likely to affect randomization significantly since panels are drawn shortly before litigants are expected to present their cases to the court (Fischman 2015). But because settlement may occur anytime before an opinion is actually released, and because opinions are released, on average, 18 months after appeals are filed in the Ninth Circuit,[23] we consider settlement behavior a plausible threat to randomization. We think the threat is particularly serious when a case was orally argued, as judges may reveal their intentions through questioning, thereby altering settlement behavior.

In order to guard against threats to randomization, we move beyond raw comparisons of panel decision rates (that would rely on random assignment) and account for the possibility that some panels may be more or less likely to issue decisions in cases that are, as a general

[22] This practice was at the center of a recent ethics controversy involving federal Judge Shira Scheindlin in the Southern District of New York. In 2013, whe was removed from a high profile stop-and-frisk case by the Second Circuit, who noted (among other things) that Judge Scheindlin had abused her district's "related case" rule that allows judges to bypass random assignment and take cases reasonably related to cases already before them.

[23] Ideally, we would like to know how many months after *panel revelation* the opinion is released, but our current dataset does not contain the date that the panel was selected.

matter, more or less likely to be reversed. Because bias occurs when a confounding variable is correlated with both the treatment and the outcome, we use a prognostic score correction, which aims to make the confounding variable orthogonal to the outcome (Hansen 2008).[24] To do this, we use machine learning and the training set to estimate each case's predicted probability of reversal under the "control" condition, which we have been denoting by k. We label the predicted probability $\widehat{\psi}_i(k, \mathbf{X}_i)$, which is commonly referred to as the "prognostic score" (Hansen 2008). In our main analysis, we incorporate these prognosis scores directly into our outcome variable. That is, instead of using the actual decision,

$$Y_i = \begin{cases} 1 & \text{if panel reverses} \\ 0 & \text{if panel affirms,} \end{cases}$$

we use the difference between the predicted probability of a reversal under the control condition (estimated with the training set, and referred to as k) and the actual outcome,

$$Z_i^{j,k} \equiv Y_i - \widehat{\psi}_i(k, \mathbf{X}_i).$$

Since $Z_i^{j,k} \in [-1, 1]$, the bias corrected autocode is

$$\dot{Y}_i^{j,k} = \begin{cases} Z_i^{j,k} & \text{if } \widehat{Y}_i(j) \geq \widehat{Y}_i(k) \\ -Z_i^{j,k} & \text{if } \widehat{Y}_i(j) < \widehat{Y}_i(k) \end{cases}$$

Thus, if, for example, some panels are more likely to issue decisions in cases with high reversal probabilities (due to breakdowns in randomization), that fact is accounted for (as best as possible) when making comparisons.[25] Our modified estimator, now resistant to

[24] Researchers have traditionally used propensity scores (or some other technique) to force independence between the confounding variable and the treatment.

[25] Two details are worth mentioning. First, we estimate the prognosis scores with and without party variables for fear they could be post-treatment. The results do not change significantly. Second, since the identification of the "treatment" and "control" groups is arbitrary when comparing two panels, all of our

breakdowns in the randomization procedure, is:

$$\widehat{\delta}(j,k,\mathcal{M}^*|\mathbf{X}) = \frac{1}{N_j}\left(\sum_{\widehat{M}_j^+} Z_i^{j,k} - \sum_{\widehat{M}_j^-} Z_i^{j,k}\right) - \frac{1}{N_k}\left(\sum_{\widehat{M}_k^+} Z_i^{j,k} - \sum_{\widehat{M}_k^-} Z_i^{j,k}\right)$$

Step 6: Estimate Inconsistency with the Test Set

Our final step is to estimate average inconsistency and extreme inconsistency. Our estimator for extreme inconsistency is:

$$\widehat{\Delta}_e \equiv \max\left\{\widehat{\delta}(j,k,\mathcal{M}^*|\mathbf{X}) : (j,k) \in \mathcal{P}\right\} \tag{10}$$

Recall that average inconsistency is an estimate of the percentage of cases that would have been decided differently if the court had re-randomized the assignment of cases to panels. In calculating average inconsistency, we account for the fact that different panel types hear greater or fewer cases using a re-randomization weighting. Consider panels of type j and k. Then, the probability that a j panel is re-randomized an k panel (or vice versa) is:

$$\widehat{w}(j,k) = \frac{N_j}{N} \cdot \frac{N_k}{N}$$

where N_j and N_k are the number of cases seen by a type j panel and type k panel, respectively. Our estimand for average inconsistency is therefore:

$$\widehat{\Delta}_a \equiv \sum_{(j,k)\in\mathcal{P}} \widehat{w}(j,k)\ \widehat{\delta}(j,k,\mathcal{M}^*|\mathbf{X}) \tag{11}$$

To illustrate how we measure inconsistency, consider again the example in Table 1. If we were to compare the reversal rates of `RRR` panels and of `LLL` panels to estimate an ATE and

analyses with prognosis scores is completed twice, with each panel being regarded as the "control" group. Results are not sensitive to the arbitrary choice of the "control" group, but we nevertheless average results.

(naïvely) use it as a proxy of disagreement, we would get an estimate of 0.5. Specifically, since the RRR reversal rate is 1 and the LLL reversal rate is 0.5, the ATE would be the difference in these rates: $E[Y_{\texttt{RRR}}] - E[Y_{\texttt{LLL}}] = 0.5$. However, by using the panel-specific predictions to code RRR and LLL decisions, we have essentially allowed ourselves to estimate more than just an ATE. In fact, we are estimating a heterogeneous effect: how do the reversal rates of the different types of panels vary across different types of cases. In our example, our automatic coding transformed the outcome on Case 3 since it appears to be a LLL-leaning decision, despite the fact that it entailed a reversal and RRR panels reverse more often (*e.g.*, a reversal of a pro-plaintiff lower court employment discrimination judgment). Using our automatically coded outcome, we would now estimate a difference between RRR and LLL panels of $E[\dot{Y}_{\texttt{RRR}}] - E[\dot{Y}_{\texttt{LLL}}] = 1$, which is significantly higher than the ATE of 0.5.

Using our test set, we estimate extreme inconsistency at 40%. Specifically, we estimate that RRR-panels would decide 40% of cases differently than LLL-panels. Furthermore, disagreement in monotonically increasing as more L judges replace R judges on a RRR panel, assuring that the degree of disagreement is unlikely to be explained away by regression to the mean. We also estimate an average inconsistency of 9%. For purposes of comparison, we show how our estimates compare with naive measures of inconsistency that use party of appointing president and differences in reversal or pro-plaintiff decision rates. Our approach uncovers considerably more inconsistency than the naive measures.

Table 2: Comparing Measures of Inconsistency

	Extreme	Average
Outcome Measure	Inconsistency	Inconsistency
Cluster and Automated Coding	40%	9%
Dems v. Reps on Reversal Rates	12%	4%
Dems v. Reps on Pro-Plaintiff Rates	2%	1%

5 Discussion: Applications for Inconsistency Estimates

Despite its long history (*e.g.*, Everson 1919), the empirical study of inconsistency is still young. We have provided a new method for finding the best possible estimates of disagreement, but usefully employing those estimates presents another challenge. In this section, we offer suggestions for a way forward. First, we briefly discuss four general issues that researchers should be aware of. Second, with a series of brief discussions and examples, we show how researchers might productively proceed in awareness of those issues.

The Difficulty of Interpretation. We have shown that Ninth Circuit panels could disagree on at least 40% of cases and that at least 9% of cases would be decided differently if they were randomly reassigned. A response might be "So what?" We empathize. Reports of inconsistency are difficult to interpret without context. How much inconsistency is disconcerting? Is it worth addressing with institutional reform? The answers will be highly dependent on the details of adjudication system being considered, and there are (at least yet) no analytically tractable optimization problems that can deliver answers; a formal cost-benefit analysis that incorporates the benefits of predictability, comparative justice, and accuracy is beyond our current capabilities. Instead, we think scholars will often be reliant on the holistic judgment of system insiders—the judges, repeat litigants, and administrators who have the deepest experience with the system will often have the best ability to interpret inconsistency levels. But we might also be able to make better sense of them by locating external benchmarks, such as inconsistency levels in comparable systems.

Estimates are a Lower Bound. Studies of inconsistency using observational data will only be able to identify a lower bound on inter-judge disagreement.[26] As a technical

[26] Fischman also describes a method for identifying an upper bound. We do not pursue an upper bound for two reasons. First, the problem becomes considerably more complex once we allow for the dimensions of disagreement to depend on the particular two judges being compared. Second, we think that upper bounds will generally be of less interest than lower bounds. Upper bounds will often be implausibly high and less informative than lower bounds.

matter, we will only ever be to say that judges in a particular system disagree in *at least* $X\%$ of cases. In short, while other methods may expose only the elephant's toenail, more aggressive methods like ours might still only expose the leg. Moreover, there is no way to know with certainty how much inconsistency we could still be missing. While more data and more variables can push us closer and closer to the true levels of inconsistency, we may never get there. Judges' decisions might differentially depend, for example, on litigant eye color. Without a measure of litigant eye color (or a measure of some variable that strongly correlates with eye color), estimates of inconsistency will not capture that dimension of disagreement.

The Selection of Disputes for Litigation. In most systems of adjudication, the cases that judges dispose of are only a fraction of the total number of cases that are settled (or never even initiated) via anticipation of what judges would decide (Priest and Klein 1984). The extent to which this is true is dependent on context. For example, the US federal court system quite certainly dispenses with the vast majority of potential disputes by maintaining a body of precedent that is clear enough to keep potential litigants from seeking clarification through a judicial ruling, but the California Prison Parole system disposes of almost all potential cases with an actual judicial decision. The selection of disputes for litigation contributes to the difficulty of interpreting inconsistency estimates.[27] But more importantly, it complicates attempts to measure the effect of legal changes (Hubbard 2013; Gelbach 2014). Because the set of litigated cases can change in response to legal reform, apparent differences in judicial decision making patterns may reflect a change in the cases being decided rather than an actual change in judicial decision making.

Comparisons are Sensitive to Sample Size. As sample size increases, the predictive power of algorithms is also likely to increase, thus leading to higher and more accurate esti-

[27] As Judge Easterbrook once said "Given selection pressure in litigation, the puzzling feature of the judicial system is *agreement*." (Swarthmore graduation speech)

mates of inconsistency. Comparisons between systems will therefore be sensitive to sample size, and it will generally make sense to use similar sample sizes when estimating inconsistency for the purposes of comparison.

The above four issues are serious, but analysis of inconsistency can nonetheless prove useful. Below, we suggest some of the ways that estimates of inconsistency might be employed.

Assessing Systems

Scholars have long used measures of inter-judge disparities to assess the quality of adjudication systems. Although they have so far used only unidimensional measures of disagreement that likely understate the extent of inconsistency, some systems—such as those for making decisions in asylum and social security disability awards—are home to such large and easily detectable disparities that most people seem to agree: something is not right. But in many contexts, it will not be immediately clear that there is a problem worth trying to address. It is not obvious to us, for example, that something is amiss with the Ninth Circuit given our findings that panels could disagree on at least 40% of cases and that at least 9% of cases would be decided differently if they were randomly reassigned. Is that level of inconsistency disconcerting?

Answering the question is important. Inconsistency has been at the heart of the debate over whether the Ninth Circuit should be split into two or more smaller circuits. Ninth Circuit Judge Tallman, for example complained about the inconsistency in the Ninth Circuit in Congressional testimony:

> [I]t gets back to what we talked about at the last hearing, which is the importance of maintaining consistency and predictability in the law. The problem that we have now with 50 judges resident, active and senior, and 150 to 200 visiting judges is that it is like going to Las Vegas in terms of what the outcome is going to be. Tell me who

> the three judges are going to be on the panel and I might be able to predict how that particular panel is going to go.

But others have vigorously disagreed with Judge Tallman's perspective. The Federal Judicial Center, for example, has claimed that "despite concerns about the proliferation of precedent as the courts of appeals grow, there is currently little evidence that intracircuit inconsistency is a significant problem" (McKenna 1993).

Can measures of inconsistency help to illuminate this debate? We think so, and we propose three potential benchmarks for aid in determining whether estimates of inconsistency are alarming. First, insider intuitions might be useful: empirical evidence of inconsistency might convince judges that inconsistency is more prevalent and serious than they think it is. Second, we can use public claims about the percentage of "hard cases" to gauge whether the Ninth Circuit has a problem. Prominent legal figures like Judge Harry Edwards and Judge Cardozo have claimed that between 5-15% of cases are legally indeterminate. Our evidence shows that some panels can disagree on at least 40% of cases, which suggests a much higher level of indeterminacy in the Ninth Circuit.[28] Finally, we could compare estimates of inconsistency in the Ninth Circuit to the other circuits. Unfortunately, due to restrictions placed on access to the court's public records system (PACER) we do not have data for other circuits. While these comparisons would not be definitive—the differences in estimates might reflect different case composition and/or differences in the detectability of inter-judge disagreement—much higher levels of detectable inconsistency in the Ninth Circuit would be cause for concern.

[28] It is also possible that 40% of cases are not legally indeterminate: some fraction of that disagreement could stem from judicial error. But we find the indeterminacy/error dichotomy generally unhelpful—one judge's error is often another judge's indeterminacy.

Evaluating Institutional Reform

We might want to know whether institutional reforms increase or decrease inconsistency. As Fischman (2014) has pointed out, there a challenges:

> If one goal of an institutional reform is to reduce inconsistency, it will be difficult to assess whether the reform is successful. Estimates of inconsistency pre- and post-intervention will be interval-identified, and these intervals will typically overlap. When this occurs, it will be impossible to determine whether the intervention increased or decreased inconsistency. Thus, evaluation of institutional changes will typically require additional data and assumptions.

Our approach to measuring inconsistency efficiently uses all available data in an effort to make the required additional assumptions as unobjectionable as possible. And we think those additional assumptions will often be reasonable. In brief, in order for a reduction in estimated inconsistency to reflect an actual decrease in inconsistency levels, it must be the case that the reform did not increase inconsistency on new, undetectable dimensions. While it may always be possible that there are substantial deviations from that assumption (*i.e.*, some significant increases in disagreement are undetectable), our data-adaptive approach minimizes the possibility. A second issue is that changes in inconsistency levels are artifacts of changing case composition due to the selection of disputes for litigation. For this reason, insofar as possible, randomized controlled trials would generally be preferable to before and after studies.

Realism v. Formalism

The debate between realists and formalists, though it may never have existed in its caricatured form (Tamanaha 2009), is still central in legal practice and theory. Most visibly, Judge Richard Posner is not without his detractors when he argues for a more pragmatic, policy oriented approach to judicial decision making. And a substantial part of his argument is that judges are basically doing what he recommends already, even if often unwittingly.

If the current legalist approach to decision making is not delivering consistent decisions, his argument carries more water. As we noted above, judges and scholars, generally in defense of a legalist approach, have made claims that the percentage of hard cases is between 10 and 15%. Are these claims compatible with the evidence? At least in the Ninth Circuit, they are not—some of these elite, well-trained judges can disagree in at least 40% of all civil cases. Whether this is unique to the Ninth Circuit should be the subject of future research.

Less abstractly, the realist v. formalist divide is also at the heart of the debate over so called "unpublished" judicial opinions. In the U.S. Courts of Appeals, judges have discretion to designate opinions as "unpublished." [29] Crucially, unpublished opinions are not binding on future cases, and before 2007 some Circuits even forbid litigants from citing them as persuasive authority. The practice of designating opinions as unpublished is now more the rule than the exception: an estimated 80% of opinions in the U.S. Courts of Appeals are unpublished. There has been extensive debate about the merit of this practice, with one panel even going so far as to declare it unconstitutional (*Anastasoff v. United States* 2000, 8th Cir.). Defenders of unpublished opinions often rely on claims that the practice is reserved for cases that are easy and routine. Legal scholars have also contributed to the narrative that the cases decided with unpublished opinions are "widely agreed to be simple and straightforward and to involve no difficult or complex issues of law" (Sunstein, Schkade, and Ellman 2004) and "quotidian" (Miles and Sunstein 2008). But the claim that unpublished opinions are just for easy cases is simply incompatible with the evidence, at least in the Ninth Circuit. Let us even pretend that there is some stable collection of cases that judges would generally agree deserve to be unpublished. Even ignoring this possibility that judges frequently disagree on which decisions should be designated as unpublished in the first place, it is clear that unpublished opinions do not dispose of simple and straightforward

[29] The term is now a misnomer, as in contemporary times almost all unpublished opinions are in fact published in books and online legal databases.

cases. While we can detect significantly higher levels of inconsistency in published opinions (extreme inconsistency of 50% and average inconsistency of 12%), inter-judge disagreement is non-negligible in unpublished opinions as well: we estimate that panels could disagree on the outcome in at least 27% of unpublished opinions and that least 7.4% of cases would be decided differently if they were randomly reassigned.[30] While we express no opinion as to the ultimate value of designating opinions as unpublished, we do think it's clear that its defenders—at least in the Ninth Circuit—should abandon the "unpublished cases are easy" narrative.

6 Conclusion

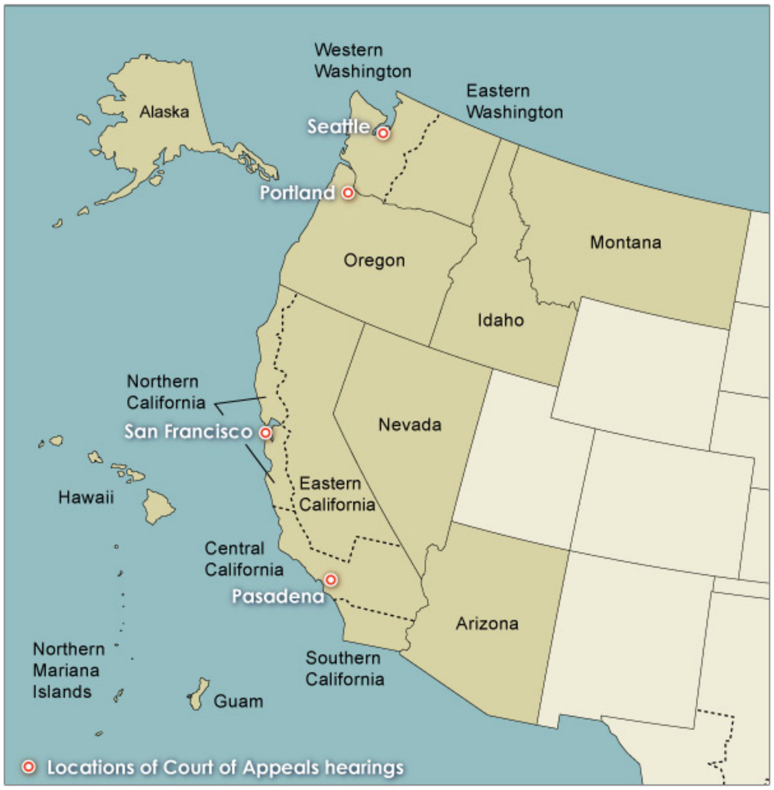
We find it helpful to think of inconsistency as a biomarker for adjudication systems. In medical research, biomarkers—measurements like blood pressure, cholesterol levels, prostate-specific antigen, mutations of p53 genes—play a major role in the development of drugs and the diagnosis of diseases. Although improvements or changes on these biomarkers is usually not important in and of themselves, they can be useful when the ultimate outcomes of interest—incidence of strokes, heart attacks, prostate cancer, or bone cancer—are slow to develop and/or difficult to detect. By using biomarkers as surrogate endpoints or diagnostic tools, medical researchers have made major drug and diagnostic advancements. Inconsistency measurements can play a similar role for adjudication systems. We rarely have the ability to measure the more foundational values that our judges serve: it is not often that we can objectively identify legal decisions as correct or not (DNA exoneration is an exception), we as yet have no good way of measuring predictability, and detecting discrimination with observational data is notoriously difficult. But inconsistency is intimately connected to these values, and in this paper we have shown how advancements in machine learning can allow us

[30] For the measures of extreme inconsistency, in order to guard against regression to the mean, we use the comparison between the panel types for which patterns in the training set suggest the largest difference.

to better study it, helping us to diagnose problems and improve outcomes in our decision-making systems.

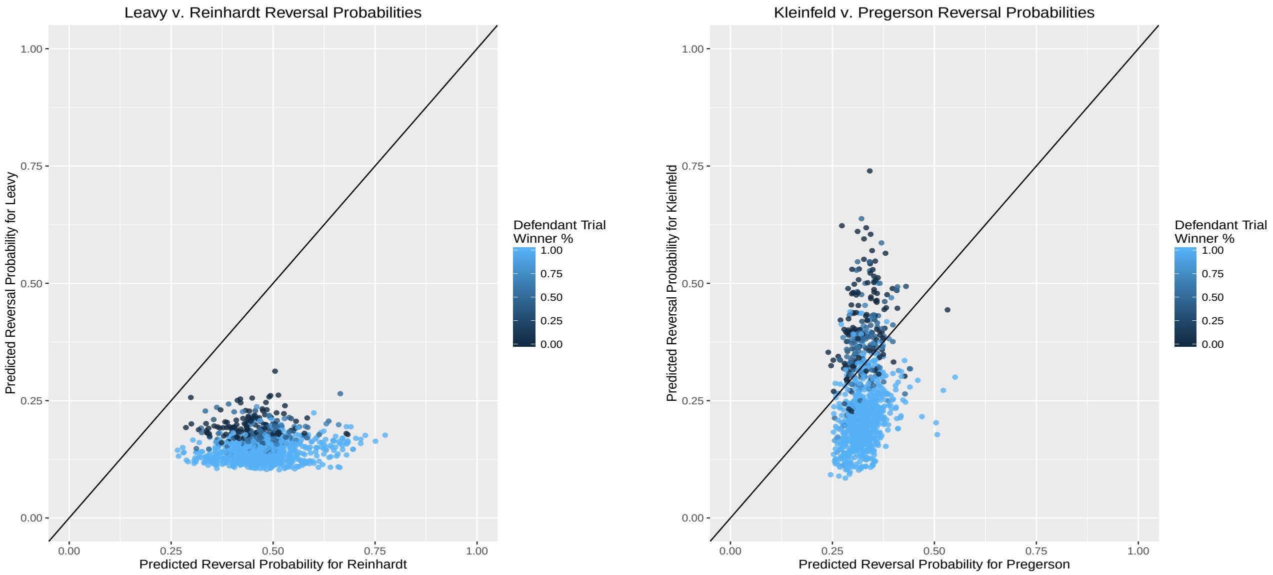
A Figures

Figure 1: The Ninth Circuit



Source: http://www.ca9.uscourts.gov/

Figure 2: Comparing Judge Predictions



Leavy v. Reinhardt Reversal Probabilities

Kleinfeld v. Pregerson Reversal Probabilities

Figure 3: Cluster Plot

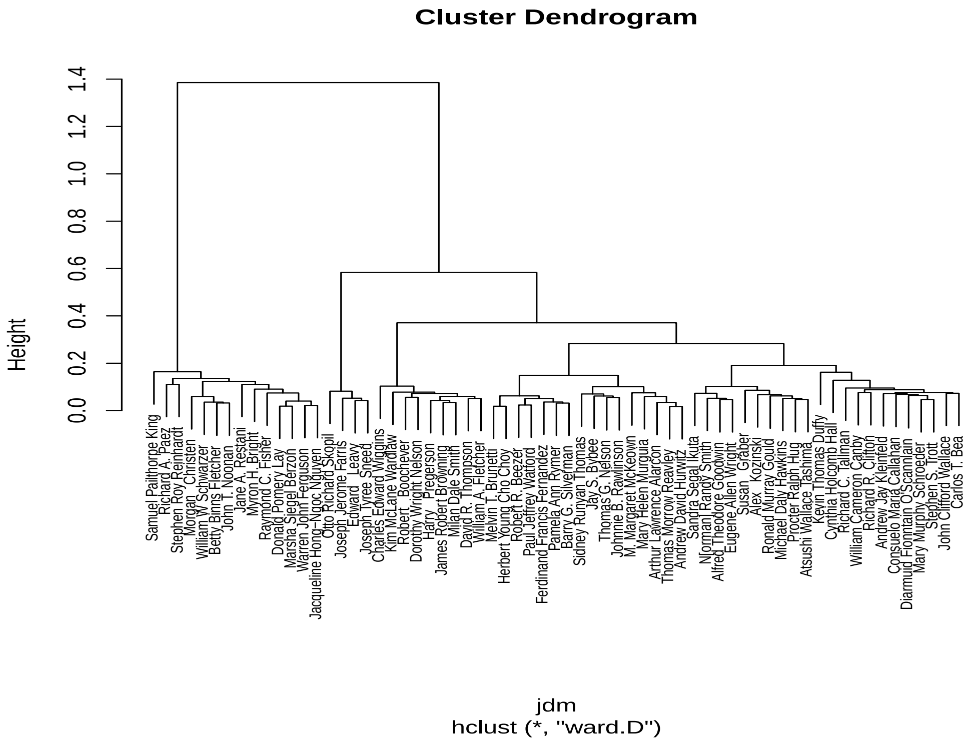
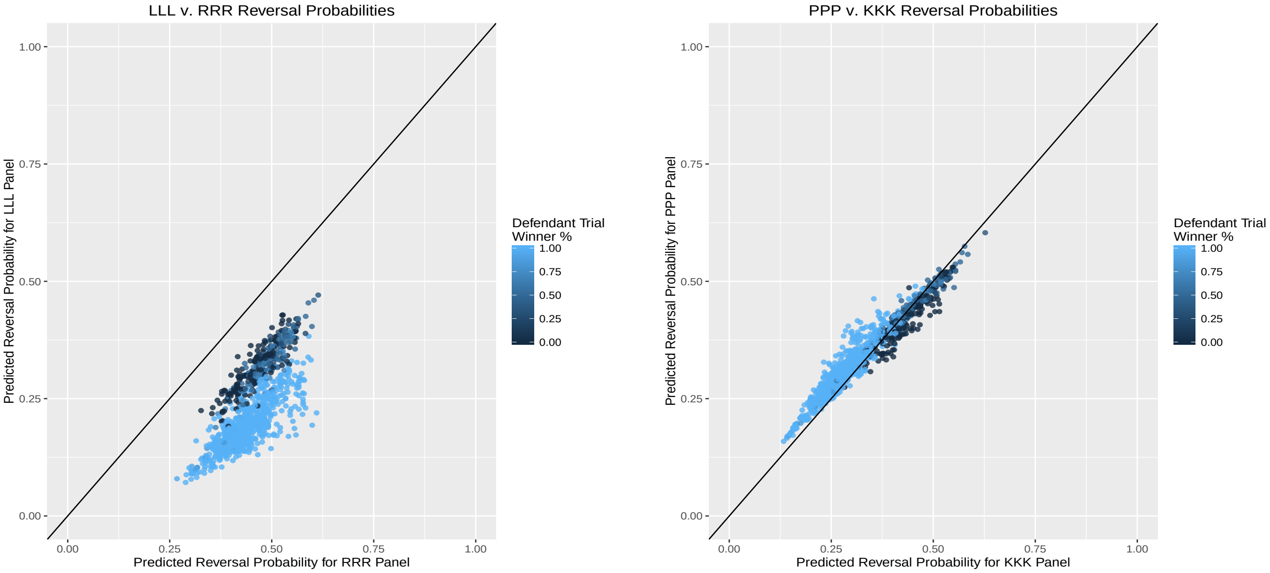


Figure 4: Comparing Panel Predictions



B Tables

Table 3: Model Performance

Model	MSE	Weight
Super Learner	0.179	–
Boosted Trees	0.181	0.51
Random Forest	0.181	0.45
LASSO	0.184	0.00
Regression 1	0.189	0.00
Regression 2	0.189	0.00
Regression 3	0.185	0.04
Mean	0.195	0.00

Table 4: Descriptive Statistics: Appeal Characteristics

	Civil Cases		*Batched*	
Variable	*Proportion*	*Negative Treatment*	*Proportion*	*Negative Treatment*
Appeal Characteristics				
Termination on Merits	0.521	0.310	1.000	0.310
Negative Treatment	0.162	1.000	0.310	1.000
Published Opinion	0.177	0.432	0.309	0.452
Dissent	0.018	0.557	0.038	0.564
Concurrence	0.017	0.537	0.026	0.616
Appellate Panels				
Party: DDD	0.095	0.377	0.175	0.386
Party: DDR	0.222	0.306	0.414	0.317
Party: DRR	0.168	0.270	0.313	0.270
Party: RRR	0.046	0.241	0.083	0.262
Race: WWW	0.305	0.304	0.553	0.309
Race: WWN	0.186	0.301	0.349	0.316
Race: WNN	0.040	0.270	0.076	0.284
Race: NNN	0.003	0.263	0.006	0.288
Sex: FFF	0.005	0.308	0.009	0.306
Sex: FFM	0.069	0.322	0.135	0.328
Sex: FMM	0.234	0.301	0.436	0.310
Sex: MMM	0.227	0.294	0.405	0.304
N	51,729		16,723	

Table 5: Descriptive Statistics: Trial Characteristics

	Civil Cases		*Batched*	
Variable	*Proportion*	*Negative Treatment*	*Proportion*	*Negative Treatment*
Trial Judges				
Magistrate Judge	0.067	0.140	0.069	0.281
Democratic Appointee	0.416	0.158	0.402	0.318
Republican Appointee	0.486	0.168	0.497	0.318
Non-white	0.218	0.164	0.129	0.352
White	0.684	0.163	0.413	0.311
Woman	0.211	0.146	0.200	0.301
Man	0.691	0.169	0.700	0.321
Case Characteristics				
Private Suit	0.772	0.162	0.728	0.323
U.S. Party	0.228	0.160	0.272	0.288
Plaintiff Won	0.238	0.192	0.192	0.377
Defendant Won	0.727	0.152	0.778	0.274
District Court				
Alaska	0.019	0.171	0.022	0.330
Arizona	0.081	0.139	0.081	0.305
California - Central	0.290	0.166	0.280	0.340
California - Eastern	0.072	0.140	0.070	0.288
California - Northern	0.157	0.148	0.156	0.285
California - Southern	0.053	0.166	0.053	0.343
Hawaii	0.033	0.113	0.028	0.252
Idaho	0.018	0.181	0.018	0.376
Montana	0.024	0.182	0.031	0.294
Nevada	0.072	0.170	0.070	0.327
Oregon	0.072	0.192	0.077	0.339
Washington - Eastern	0.018	0.181	0.019	0.310
Washington - Western	0.083	0.176	0.090	0.323
N	51,729		16,723	

Table 6: Reinhardt Cluster

	Name	Party	Sex	Race	Senior	Termination
1	Marsha Siegel Berzon	D	F	White		
2	Myron H. Bright	D	M	White	1985-06-01	
3	Morgan Christen	D	F	White		
4	Warren John Ferguson	D	M	White	1986-07-31	2008-06-25
5	Raymond C. Fisher	D	M	White	2013-03-31	
6	Betty Binns Fletcher	D	F	White	1998-11-01	2012-10-22
7	Samuel Pailthorpe King	R	M	Pac. Isl./White	1984-11-30	2010-12-07
8	Donald Pomery Lay	D	M	White	1992-01-07	2007-04-29
9	Jacqueline Hong-Ngoc Nguyen	D	F	Asian American		
10	John T. Noonan	R	M	White	1996-12-27	
11	Richard A. Paez	D	M	Hispanic		
12	Stephen Roy Reinhardt	D	M	White		
13	Jane A. Restani	R	F	White	2015-03-01	
14	William W Schwarzer	R	M	White	1991-04-30	

Table 7: Kozinski Cluster

	Name	Party	Sex	Race	Senior	Termination
1	Alfred Theodore Goodwin	R	M	White	1991-01-31	
2	Ronald Murray Gould	D	M	White		
3	Susan Graber	D	F	White		
4	Michael Daly Hawkins	D	M	White	2010-02-12	
5	Procter Ralph Hug	D	M	White	2002-01-01	
6	Sandra Segal Ikuta	R	F	White		
7	Alex Kozinski	R	M	White		
8	N. Randy Smith	R	M	White		
9	Atsushi Wallace Tashima	D	M	Asian American	2004-06-30	
10	Eugene Allen Wright	R	M	White	1983-09-15	2002-09-03

Table 8: Leavy Cluster

	Name	Party	Sex	Race	Senior	Termination
1	Arthur Lawrence Alarcon	D	M	Hispanic	1992-11-21	2015-01-28
2	Robert R. Beezer	R	M	White	1996-07-31	2012-03-30
3	Melvin T. Brunetti	R	M	White	1999-11-11	2009-10-30
4	Jay S. Bybee	R	M	White		
5	Herbert Young Cho Choy	R	M	Asian American	1984-10-03	2004-03-10
6	Joseph Jerome Farris	D	M	African American	1995-03-04	
7	Ferdinand Francis Fernandez	R	M	Hispanic	2002-06-01	
8	Andrew David Hurwitz	D	M	White		
9	Edward Leavy	R	M	White	1997-05-19	
10	M. Margaret McKeown	D	F	White		
11	Mary Helen Murguia	D	F	Hispanic		
12	Thomas G. Nelson	R	M	White	2003-11-14	2011-05-04
13	Johnnie B. Rawlinson	D	F	African American		
14	Thomas Morrow Reavley	D	M	White	1990-08-01	
15	Pamela Ann Rymer	R	F	White	2011-09-21	2011-09-21
16	Barry G. Silverman	D	M	White		
17	Otto Richard Skopil	D	M	White	1986-06-30	2012-10-18
18	Joseph Tyree Sneed	R	M	White	1987-07-21	2008-02-09
19	Sidney Runyan Thomas	D	M	White		
20	Paul Jeffrey Watford	D	M	African American		

Table 9: O'Scannlain Cluster

	Name	Party	Sex	Race	Senior	Termination
1	Carlos T. Bea	R	M	Hispanic		
2	Consuelo Maria Callahan	R	F	Hispanic		
3	William Cameron Canby	D	M	White	1996-05-23	
4	Richard R. Clifton	R	M	White		
5	Kevin Thomas Duffy	R	M	White	1998-01-10	
6	Cynthia Holcomb Hall	R	F	White	1997-08-31	2011-02-26
7	Andrew Jay Kleinfeld	R	M	White	2010-06-12	
8	Diarmuid Fionntain O'Scannlain	R	M	White		
9	Mary Murphy Schroeder	D	F	White	2011-12-31	
10	Richard C. Tallman	D	M	White		
11	Stephen S. Trott	R	M	White	2004-12-31	
12	John Clifford Wallace	R	M	White	1996-04-08	

Table 10: Pregerson Cluster

	Name	Party	Sex	Race	Senior	Termination
1	Robert Boochever	D	M	White	1986-06-10	2011-10-09
2	James Robert Browning	D	M	White	2000-09-01	2012-05-06
3	William A. Fletcher	D	M	White		
4	Dorothy Wright Nelson	D	F	White	1995-01-01	
5	Harry Pregerson	D	M	White		
6	Milan Dale Smith	R	M	White		
7	David R. Thompson	R	M	White	1998-12-31	2011-02-19
8	Kim McLane Wardlaw	D	F	Hispanic		
9	Charles Edward Wiggins	R	M	White	1996-12-31	2000-03-02

Table 11: Visiting Cluster

	Name	Party	Sex	Race	Senior	Termination
1	Absent From Federal Biography					
2	Judges with Fewer Than 100 Observations					

C Example Docket Sheet

```
|05-35459|05/13/2005|1110 Insurance|03/27/2007|Siaperas v. Montana State|U.S. District
   Court for Montana, Butte|
|civil|united states|null|
|ORIG_CRT|0977|2|CV-04-00013-RWA||02/17/2004|https://ecf.mtd.uscourts.gov/cgi-
   bin/DktRpt.pl?caseNumber=CV-04-00013-RWA||
|ORIG_PER|Trial Judge|Richard|W.|Anderson||Magistrate Judge|
|ORIG_DAT|04/07/2005||05/04/2005||||
|PARTY|PATSY SIAPERAS||Plaintiff - Appellant,|
|ATTORNEY|Robert|Lee|Kelleher|Sr.|Attorney|||PO Box 397|||Kelleher Law Office||406-299-
   2581|Butte|MT|59703||[COR LD NTC Retained]|||Esquire|
|PARTY|STATE OF MONTANA MUTUAL INSURANCE FUND||Defendant - Appellee,|
|ATTORNEY|Charles|G.|Adams||Attorney|||50 South Last Chance Gulch|P.O. Box 598||KELLER,
   REYNOLDS, DRAKE, JOHNSON & GILLESPIE||406/442-0230|Helena|MT|59624||[COR LD NTC
   Retained]||3rd Floor|Esquire|
|ATTORNEY|Thomas|E.|Martello||Attorney|||P.O. Box 4759|||MONTANA STATE FUND||406/444-
   6480|Helena|MT|59604-4759||[COR LD NTC Please Select]|||Esquire|
|CAPTION|PATSY SIAPERAS, Plaintiff - Appellant, v. STATE OF MONTANA MUTUAL INSURANCE FUND,
   Defendant - Appellee.|
|ENTRY|05/13/2005|DOCKETED CAUSE AND ENTERED APPEARANCES OF COUNSEL. CADS SENT (Y/N): Y.
   setting schedule as follows: Fee payment is due 5/27/05; CADS is 5/20/05; appellant's
   designation of RT is due 5/16/05; appellee's designation of RT is due appellant order
   transcript by 6/3/05; court reporter shall file transcript in DC by 7/5/05; certificate
   of record be filed by 7/11/05; appellant's opening brief is due 8/22/05; appellees'
   brief is due 9/19/05; appellants' reply brief is due 10/3/05. [05-35459] (GV)||
|ENTRY|06/03/2005|Filed order CONFATT (em) The court of appeals' records do not indicate
   that aplt has filed a CADS in accordance with Cir.R. 3-4... Within 14 days of the filing
   of this order, aplt shall file a CADS, contact the clerk and assist location of the
   docketing stmt if one is already filed, or dismiss the appeal voluntarily.... [05-
   35459] (HH)||
|ENTRY|07/05/2005|Filed Robert C. Kelleher for Appellant Patsy Siaperas Civil Appeals
   Docketing Statement; served on 7/1/05 (to CONFATT) [05-35459] [05-35459] (HH)||
|ENTRY|07/08/2005|Filed order CONFATT (em) Aplts have failed to file a CADS in accordance
   with the court's order of 6/3/05. Within 7 days of the filing date of this order, aplts
   shall file a CADS or a motion to dismiss the appeal voluntarily under FRAP 42(b), or
   shall show cause in writing why this appeal should not be dismissed. Failure to comply
   with this order will result in dismissal. See 9th Cir.R. 42-1. [05-35459] (HH)||
|ENTRY|08/01/2005|14 day oral extension by phone of time to file Appellant brief. [05-
   35459] appellants' brief due 9/6/05; appellees' brief due 10/6/05; Appellant's optional
   reply brief is due within 14 days of service of Appellee's brief. (LW)||
|ENTRY|08/19/2005|Received aplt's 5 copies of transcript of proceedings of 12/16/03, in 1
   vol. (SHELF) [05-35459] (HH)||
|ENTRY|09/06/2005|Received Appellant Patsy Siaperas's brief in 15 copies 25 pages (
   Informal: n) deficient : lacks excerpts: notified counsel. Served on 9/2/05 [05- 35459]
   (GR)||
|ENTRY|09/19/2005|14 day oral extension by phone of time to file Appellee's brief. [05-
   35459] appellees' brief due 10/20/05; appellants' reply brief due n 14 days of svc of
   aple's brief. (KM)||
|ENTRY|10/21/2005|Filed order CONFATT (bls) Case referred to Confatt for assessment
   conference only. Conference to be on 11/17/05 at 2:00pm Pacific (S.F.) Time. By
   telephone (y/n): y. The briefing schedule prevsly set by the court remains in effect.
   [05-35459] (HH)||
|ENTRY|10/24/2005|Rec'd orig. &amp;amp; 15 copies aple's brf of 16 pages and 5 copies suppl
   excerpts of record in 1 vol; served on 10/20/05. Deficient; excerpts need index and
   white covers. Notified csl. [05-35459] (XX)||
|ENTRY|10/27/2005|Received Appellant Patsy Siaperas's satisfaction of (major) brief
   deficiency: Excerpts of Record. Aplt's fee remains due. [05-35459] (GR)||
|05-35459|05/13/2005|1110 Insurance|03/27/2007|Siaperas v. Montana State|U.S. District
   Court for Montana, Butte|
|civil|united states|null|
|ORIG_CRT|0977|2|CV-04-00013-RWA||02/17/2004|https://ecf.mtd.uscourts.gov/cgi-
   bin/DktRpt.pl?caseNumber=CV-04-00013-RWA||
|ORIG_PER|Trial Judge|Richard|W.|Anderson||Magistrate Judge|
|ORIG_DAT|04/07/2005||05/04/2005||||
|PARTY|PATSY SIAPERAS||Plaintiff - Appellant,|
|ATTORNEY|Robert|Lee|Kelleher|Sr.|Attorney|||PO Box 397|||Kelleher Law Office||406-299-
   2581|Butte|MT|59703||[COR LD NTC Retained]|||Esquire|
|PARTY|STATE OF MONTANA MUTUAL INSURANCE FUND||Defendant - Appellee,|
|ATTORNEY|Charles|G.|Adams||Attorney|||50 South Last Chance Gulch|P.O. Box 598||KELLER,
   REYNOLDS, DRAKE, JOHNSON & GILLESPIE||406/442-0230|Helena|MT|59624||[COR LD NTC
   Retained]||3rd Floor|Esquire|
|ATTORNEY|Thomas|E.|Martello||Attorney|||P.O. Box 4759|||MONTANA STATE FUND||406/444-
   6480|Helena|MT|59604-4759||[COR LD NTC Please Select]|||Esquire|
|CAPTION|PATSY SIAPERAS, Plaintiff - Appellant, v. STATE OF MONTANA MUTUAL INSURANCE FUND,
   Defendant - Appellee.|
|ENTRY|05/13/2005|DOCKETED CAUSE AND ENTERED APPEARANCES OF COUNSEL. CADS SENT (Y/N): Y.
   setting schedule as follows: Fee payment is due 5/27/05; CADS is 5/20/05; appellant's
   designation of RT is due 5/16/05; appellee's designation of RT is due appellant order
   transcript by 6/3/05; court reporter shall file transcript in DC by 7/5/05; certificate
   of record be filed by 7/11/05; appellant's opening brief is due 8/22/05; appellees'
   brief is due 9/19/05; appellants' reply brief is due 10/3/05. [05-35459] (GV)||
|ENTRY|06/03/2005|Filed order CONFATT (em) The court of appeals' records do not indicate
   that aplt has filed a CADS in accordance with Cir.R. 3-4... Within 14 days of the filing
   of this order, aplt shall file a CADS, contact the clerk and assist location of the
   docketing stmt if one is already filed, or dismiss the appeal voluntarily.... [05-
   35459] (HH)||
|ENTRY|07/05/2005|Filed Robert C. Kelleher for Appellant Patsy Siaperas Civil Appeals
   Docketing Statement; served on 7/1/05 (to CONFATT) [05-35459] [05-35459] (HH)||
|ENTRY|07/08/2005|Filed order CONFATT (em) Aplts have failed to file a CADS in accordance
   with the court's order of 6/3/05. Within 7 days of the filing date of this order, aplts
   shall file a CADS or a motion to dismiss the appeal voluntarily under FRAP 42(b), or
   shall show cause in writing why this appeal should not be dismissed. Failure to comply
   with this order will result in dismissal. See 9th Cir.R. 42-1. [05-35459] (HH)||
|ENTRY|08/01/2005|14 day oral extension by phone of time to file Appellant brief. [05-
   35459] appellants' brief due 9/6/05; appellees' brief due 10/6/05; Appellant's optional
   reply brief is due within 14 days of service of Appellee's brief. (LW)||
|ENTRY|08/19/2005|Received aplt's 5 copies of transcript of proceedings of 12/16/03, in 1
   vol. (SHELF) [05-35459] (HH)||
|ENTRY|09/06/2005|Received Appellant Patsy Siaperas's brief in 15 copies 25 pages (
   Informal: n) deficient : lacks excerpts: notified counsel. Served on 9/2/05 [05- 35459]
   (GR)||
|ENTRY|09/19/2005|14 day oral extension by phone of time to file Appellee's brief. [05-
   35459] appellees' brief due 10/20/05; appellants' reply brief due n 14 days of svc of
   aple's brief. (KM)||
|ENTRY|10/21/2005|Filed order CONFATT (bls) Case referred to Confatt for assessment
   conference only. Conference to be on 11/17/05 at 2:00pm Pacific (S.F.) Time. By
   telephone (y/n): y. The briefing schedule prevsly set by the court remains in effect.
   [05-35459] (HH)||
|ENTRY|10/24/2005|Rec'd orig. &amp;amp; 15 copies aple's brf of 16 pages and 5 copies suppl
   excerpts of record in 1 vol; served on 10/20/05. Deficient; excerpts need index and
   white covers. Notified csl. [05-35459] (XX)||
|ENTRY|10/27/2005|Received Appellant Patsy Siaperas's satisfaction of (major) brief
   deficiency: Excerpts of Record. Aplt's fee remains due. [05-35459] (GR)||
```

D Machine Learning

Throughout the steps of our analysis, we construct most of our predictive models using `Super Learner`, an ensemble machine-learning method developed in University of California Berkeley's Biostatistics Department (Laan, Polley, and Hubbard 2007). `Super Learner` takes as input any number of user-supplied models (*e.g.*, a parametric linear regression, random forest, LASSO, etc.) and combines those models' predictions to generate "super" predictions. Specifically, the `Super Learner` proceeds in two steps: first, validation-set predictions are generated for each candidate model; second, the true outcome is regressed on the candidate models' predictions to assign each model's predictions a weight.

In order to generate validation-set predictions, `Super Learner` breaks whatever data it is given into ten separate random "chunks." Ten-fold cross-validation is the default and is generally regarded as an appropriate choice. The first chunk, the first tenth of the data, is then set aside and the underlying models are built using the remaining nine tenths of the data. The left-out tenth of the data, the "validation set," is then plugged into the underlying models and used to generate model predictions. The same process is repeated for each of the remaining chunks. That is, the second tenth of the data is set aside, and `Super Learner` builds the models on the remaining nine tenths of the data (the first chunk is now being used to help build the model) and then generates validation set predictions for the second chunk. And so on for all ten chunks. The appeal of these validation set predictions is that they allow us to estimate how the underlying model would perform on data it has never seen.

The first step generates validation set predictions for each data point for each underlying model. In the second step, `Super Learner` then leverages the cross-validation information on model performance to assign weights to each model according to how well their predictions match the true outcome. It does this by regressing the true outcome on the underlying model predictions. As a default, `Super Learner` runs a non-negative least squares regression.

E Proofs

Definition 1. A CATE exhibits **strongly heterogeneous treatment effects** if and only if there exists $M \in \mathcal{M}$ such that $\phi(j,k,M) > 0$ and $M' \in \mathcal{M}$ such that $\phi(j,k,M) < 0$.

Proposition 1. Suppose that $Y \subseteq \mathbb{R}$. Every ATE-based estimand will be a lower bound of disagreement. That is, it will be biased downward: $\phi(j,k) \leq \delta(j,k)$.

Proof of Proposition 1. Because $d(\cdot)$ is a metric on Y, it follows that

$$d(x,z) \leq d(x,y) + d(y,z)$$

Moreover, by the properties of expectations,

$$E[d(x,z)] \leq E[d(x,y)] + E[d(y,z)]$$

Now, consider three points in the set Y: $Y(j)$, $Y(k)$ and 0. We can express the triangle inequality as follows:

$$E[d(Y(j),0)] \leq E[d(Y(j),Y(k))] + E[d(Y(k),0)]$$

Rearranging terms yields

$$E[d(Y(j),0)] - E[d(Y(k),0)] \leq E[d(Y(j),Y(k))]$$

and by the linearity of the expectation operator,

$$E[d(Y(j),0) - d(Y(k),0)] \leq E[d(Y(j),Y(k))]$$

Finally, note that in $\mathbb{R}$, $d(Y(j), 0) = Y(j)$, so that

$$E[Y(j) - Y(k)] \leq E[d(Y(j), Y(k))]$$

By the definitions of disagreement and ATE, we have directly shown that all ATEs will be weakly smaller than disagreement. □

Corollary 1. An ATE-based estimand will be *strictly* lower than disagreement if there are strongly heterogeneous treatment effects in the sense of Definition 1.

Proposition 2. For all $\mathcal{M} \neq \mathcal{N}$, $\delta(j, k, \mathcal{M}) \leq \delta(j, k)$.

Proof. By contradiction, suppose that there exists some $\mathcal{M} \neq \mathcal{N}$ such that $\delta(j, k, \mathcal{M}) > \delta(j, k)$. We can rewrite this condition as

$$E_{M \in \mathcal{M}}\big[|E_{i \in M}[Y_i(j) - Y_i(k)]|\big] > E_{i \in \mathcal{N}}[|Y_i(j) - Y_i(k)|]$$

By the law of iterated expectations, this can be further re-written as

$$E_{M \in \mathcal{M}}\big[|E_{i \in M}[Y_i(j) - Y_i(k)]|\big] > E_{M \in \mathcal{M}}\left[E_{i \in M}[|Y_i(j) - Y_i(k)|]\right]$$

By the definition of $\mathcal{M}$, since $\mathcal{M} \neq \mathcal{N}$, there must be at least one element $M \in \mathcal{M}$ such that $|M| > 1$. Denote that element by M'. By Jensen's inequality,

$$|E_{i \in M'}[Y_i(j) - Y_i(k)]| \leq E_{i \in M'}[|Y_i(j) - Y_i(k)|]$$

It follows, then, that

$$E_{M \in \mathcal{M}}\left[E_{i \in M}[Y_i(j) - Y_i(k)]|\right] \leq E_{M \in \mathcal{M}}\left[E_{i \in M}[|Y_i(j) - Y_i(k)|]\right]$$

This contradicts our assertion that $\delta(j, k, \mathcal{M}) > \delta(j, k)$. □

Definition 2. Let $A_i^{j,k}(j)$ and $A_i^{j,k}(k)$ be defined as follows:

$$A_i^{j,k}(j) = \begin{cases} Y_i & \text{if } Y_i(j) \geq Y_i(k) \text{ and } i \in \mathcal{N}_j \\ 1 - Y_i & \text{if } Y_i(j) < Y_i(k) \text{ and } i \in \mathcal{N}_j \end{cases}$$
$$A_i^{j,k}(k) = \begin{cases} Y_i & \text{if } Y_i(j) \geq Y_i(k) \text{ and } i \in \mathcal{N}_k \\ 1 - Y_i & \text{if } Y_i(j) < Y_i(k) \text{ and } i \in \mathcal{N}_k \end{cases}$$

Proposition 3. $\delta(j, k, \mathcal{M}^*) = E\left[A^{j,k}(j)\right] - E\left[A^{j,k}(k)\right]$.

Proof. We show this directly:

$$\begin{aligned}
\delta(j,k,\mathcal{M}^*) &= E\left[Y(j) - Y(k)|\widehat{Y}(j) \geq \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) \geq \widehat{Y}(k)\right) + E\left[Y(k) - Y(j)|\widehat{Y}(j) < \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) \\
&= E\left[Y(j) - Y(k)|\widehat{Y}(j) \geq \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) \geq \widehat{Y}(k)\right) + E\left[Y(k) - Y(j)|\widehat{Y}(j) < \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) \\
&\quad + \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) - \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) \\
&= E\left[Y(j)|\widehat{Y}(j) \geq \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) \geq \widehat{Y}(k)\right) - E\left[Y(k)|\widehat{Y}(j) \geq \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) \geq \widehat{Y}(k)\right) \\
&\quad - E\left[Y(j)|\widehat{Y}(j) < \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) + E\left[Y(k)|\widehat{Y}(j) < \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) \\
&\quad + \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) - \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) \\
&= E\left[Y(j)|\widehat{Y}(j) \geq \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) \geq \widehat{Y}(k)\right) + \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) - E\left[Y(j)|\widehat{Y}(j) < \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) \\
&\quad - E\left[Y(k)|\widehat{Y}(j) \geq \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) \geq \widehat{Y}(k)\right) - \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) + E\left[Y(k)|\widehat{Y}(j) < \widehat{Y}(k)\right] \Pr\left(\widehat{Y}(j) < \widehat{Y}(k)\right) \\
&= E\left[A_i^{j,k}(j)\right] - E\left[A_i^{j,k}(k)\right]
\end{aligned}$$

Therefore, $\delta(j, k, \mathcal{M}^*) = E\left[A^{j,k}(j)\right] - E\left[A^{j,k}(k)\right]$. □

References

Anastasoff v. United States. 2000. 223 F.3d 898.

Anderson, James M., Jeffrey R. Kling, and Kate Stith. 1999. "Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines." *Journal of Law and Economics* 42 (S1): 271–308.

Athey, Susan, and Guido W. Imbens. 2015. "Machine Learning Methods for Estimating Heterogeneous Causal Effects."

Boyd, Christina L., Lee Epstein, and Andrew D. Martin. 2010. "Untangling the Causal Effects of Sex on Judging." *American Journal of Political Science* 54 (2): 389–411.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. "Yes, But What's the Mechanism? (Don't Expect an Easy Answer)." *Journal of Personality and Social Psychology* 98 (4): 550–558.

Cockburn, Ian, Samuel Kortum, and Scott Stern. 2003. "Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes." In *Patents in the Knowledge-Based Economy,* 19–53. Washington, DC: The National Academic Press.

Everson, George. 1919. "The Human Element in Justice." *Journal of the American Institute of Criminal Law and Criminology* 10:90–99.

Farhang, Sean, and Gregory J. Wawro. 2004. "Institutional Dynamics on the U.S. Court of Appeals: Minority Representation Under Panel Decision Making." *Journal of Law, Economics, and Organization* 20 (2): 299–330.

Fischman, Joshua B. 2014. "Measuring Inconsistency, Indeterminacy, and Error in Adjudication." *American Law and Economics Review* 16 (1): 40–85.

Fischman, Joshua B. 2015. “Interpreting Circuit Court Voting Patterns: A Social Interactions Framework.” *Journal of Law, Economics, and Organization* 31 (4): 808–842.

Gelbach, Jonah B. 2014. “Can the Dark Arts of the Dismal Science Shed Light on the Empirical Reality of Civil Procedure.” *Stanford Journal of Complex Litigation* 2 (2): 223–296.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2016. “Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods.”

Hansen, Ben B. 2008. “The Prognostic Analogue of the Propensity Score.” *Biometrika* 95 (2): 481–488.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 (396): 945–960.

Holmes Jr., Oliver Wendell. 1897. “The Path of Law.” *Harvard Law Review* 10 (457).

Hubbard, William HJ. 2013. “Testing for Change in Procedural Standards, with Application to Bell Atlantic v. Twombly.” *The Journal of Legal Studies* 42 (1): 35–68.

Imai, Kosuke, and Aaron Strauss. 2011. “Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign.” *Political Analysis* 19 (1): 1–19.

Kastellec, Jonathan P. 2013. “Racial Diversity and Judicial Influence on Appellate Courts.” *American Journal of Political Science* 57 (1): 167–183.

Laan, Mark J. van der, Eric C. Polley, and Alan E. Hubbard. 2007. “Super Learner.” `http://biostats.bepress.com/ucbbiostat/paper222/`.

Landa, Dimitri, and Jeffrey R. Lax. 2009. “Legal Doctrine on Collegial Courts.” *Journal of Politics* 71 (3): 946–963.

Laqueur, Hannah, and Ryan Copus. 2016. "Machines Learning Justice: The Case for Judgmental Bootstrapping of Legal Decisions."

Legomsky, Stephen H. 2007. "Learning to Live with Unequal Justice: Asylum and the Limits to Consistency." *Stanford Law Review* 60 (473).

Lind, E. Allan, and Tom R. Tyler. 1988. *The Social Psychology of Procedural Justice.* New York: Plenum Press.

McKenna, Judith A. 1993. *Structural and Other Alternatives for the Federal Courts of Appeals.* Technical report. Washington, DC: Federal Judicial Center.

Miles, Thomas J., and Cass R. Sunstein. 2008. "The New Legal Realism." *University of Chicago Law Review* 75 (2): 831–851.

Persson, Torsten, and Guido Tabellini. 2000. *Political Economics: Explaining Economic Policy.* Cambridge, MA: The MIT Press.

Pocock, Stuart J., Susan E. Assmann, Laura E. Enos, and Linda E. Kasten. 2002. "Subgroup Analysis, Covariate Adjustment and Baseline Comparisons in Clinical Trial Reporting: Current Practice and Problems." *Statistics in Medicine* 21 (19): 2917–2930.

Priest, George L., and Benjamin Klein. 1984. "The Selection of Disputes for Litigation." *Journal of Legal Studies* 13:1–55.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Phillip G. Schrag. 2007. "Refugee Roulette: Disparities in Asylum Adjudication." *Stanford Law Review* 60:295–412.

Revesz, Richard L. 1997. "Environmental Regulation, Ideology, and the D.C. Circuit." *Virginia Law Review* 83 (8): 1717–1772.

Shavell, Steven. 2007. *Economic Analysis of Accident Law.* 320. Cambridge, MA: Harvard University Press.

Stith, Kate, and José A. Cabranes. 1998. *Fear of Judging: Sentencing Guidelines in the Federal Courts.* Chicago: University of Chicago Press.

Sunstein, Cass R., David Schkade, and Lisa Michelle Ellman. 2004. "Ideological Voting on Federal Courts of Appeals: A Preliminary Investigation." *Virginia Law Review* 90 (1): 301–354.

Tamanaha, Brian Z. 2009. *Beyond the Formalist-Realist Divide: The Role of Politics in Judging.* Princeton, NJ: Princeton University Press.

Tiller, Emerson H., and Frank B. Cross. 1999. "A Modest Proposal for Improving American Justice." *Columbia Law Review* 99 (1): 215–234.

United States Courts. 2015. *Federal Judicial Caseload Statistics.* Technical report. `http://www.uscourts.gov/report-names/federal-judicial-caseload-statistics`.