Title

Machine Learning and the Reliability of Adjudication

Permalink

https://escholarship.org/uc/item/7t80m17d

Author

Copus, Ryan W

Publication Date

2017

Peer reviewed|Thesis/dissertation

Machine Learning and the Reliability of Adjudication

by

Ryan W. Copus

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Jurisprudence and Social Policy

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Justin McCrary, Co-Chair
Professor Kevin Quinn, Co-Chair
Professor Anne Joseph O'Connell
Professor Sean Farhang

Fall 2017

Machine Learning and the Reliability of Adjudication

Abstract

Machine Learning and the Reliability of Adjudication

by

Ryan W. Copus

Doctor of Philosophy in Jurisprudence and Social Policy

University of California, Berkeley

Professor Justin McCrary, Co-Chair
Professor Kevin Quinn, Co-Chair

Machine learning can be used to help guide and regulate adjudicator decisions, increasing the reliability and overall quality of decision making. The first chapter provides an analytic and normative overview of what I refer to as "statistical precedent." It explains how statistical models of previous decisions can help assess and improve the reliability of an adjudication system. The subsequent chapters elaborate on and empirically illustrate two of the techniques introduced in the first chapter. Chapter two, using an original dataset of Ninth Circuit Court of Appeals decisions, presents a method for estimating the amount of inter-judge disagreement. Chapter three, using an original dataset of California parole hearings, demonstrates the potential of synthetically crowdsourced decision making.

To Maryana Pinchuk

My Archimedean Point. I am thankful for her intellectual, financial, and emotional support of “Robo-Judge.”

Chapter 1: Statistical Precedent

1 Introduction

Adjudication is omnipresent in the regulatory state. In addition to the federal, state, magistrate and administrative law judges who we employ to apply rules and standards to particular claims, we assign cases to armies of probation officers, food safety inspectors, asylum officers, tax auditors, Medicare claims processors, nursing home inspectors, parole commissioners, teaching evaluators, police detectives, veteran's disability claims processers, patent examiners, IRB compliance officers—the list is almost endless. With rare exception, these decentralized decision-making systems operate in much the same way as they always have: various human judges[1] using their human judgments to make human decisions. But those decisions are often all too human. Due to inter-judge and intra-judge inconsistency as well as systematic biases, cleaner restaurants are shut down while dirtier restaurants remain open, dangerous inmates are released into society while safer inmates are kept in prison, less deserving sexual discrimination plaintiffs win cases while more deserving plaintiffs lose, and bad patents are issued while good patents are denied.

This article argues that machine learning has a major role to play in improving our adjudication systems. Advanced analytics is transforming industries: finance, sales, medicine, transportation, elections, sports, and even fantasy sports have been deeply impacted by improvements in predictive technology. It can do the same for adjudication. In brief, I argue that it can help us determine what the law is. Does that immigrant qualify for asylum? Do those acts constitute sexual harassment? Is that invention novel? Does the claimant qualify for social security disability benefits? While answers to such questions do, and, for the foreseeable future, will continue to depend on the subtlety and flexibility of human reasoning, predictive technologies can supplement that reasoning. This article explains how.

In any effort to determine what the law is, it is important to distinguish between what legal philosophers have termed the "internal" and "external" legal perspectives (Shapiro 2006). Broadly speaking, the internal perspective is a normative one. By taking the internal perspective, we address ourselves to the proper interpretation and application of law. The archetype is a judge deciding a case. The external perspective, in contrast, is primarily a descriptive one. The inquiries we conduct from the internal perspective eventually yield decisions, and these are the data points on which the external perspective primarily relies. We might adopt the external perspective as social scientists, attempting to trace the genealogy of law and offering causal explanations for its content. Or we might engage with the external perspective as Holmesian lawyers, interested in merely predicting what a court will decide so as to better advise and represent our clients.

Machines, while of value to the internal perspective, are severely limited in their ability to engage in the normative reasoning that the internal perspective requires. Without ex-

[1] I use the terms "judges" and "adjudicators" interchangeably to refer decision-makers in decentralized decision-making systems.

traordinary advancements in artificial intelligence that are unlikely to materialize anytime soon, they cannot even begin to replicate the complexity of human judgment that the internal perspective so often requires. Computer programmers write code that can automatically apply clearly articulated rules, but that code cannot yet engage normatively or interpret the meaning of rules. Statisticians and economist can estimate the factors that inform the internal perspective, letting us better understand the effects decisions may have (e.g., they may provide estimates of recidivism risk that can inform sentencing decisions). But again, their statistics cannot tell us what the law means and requires. Both tasks can prove useful, but they will often fall far short of answering the internal perspective's version of "What is the law?"

Machines are more at home in the external perspective. Freed from the need to replicate the complexity of the human reasoning process, they can simply aim to predict its results, and machines excel at the task of prediction. It is the reason for the excitement in the private industry, as companies like Lex Machina and Legalist build models to predict the outcomes of lawsuits.

But for those of us who are neither pure social scientists nor Holmes's bad man, machine competence in the external perspective may fail to inspire hope in the ability of machines to answer the "What is law?" that we care most about. But it should. The external perspective is of critical importance to forming and defending our internal perspective. The reason is twofold. First is humility. Lacking the expertise, talent, or time to fully engage with the internal perspective, we may defer to judicial decisions out of humility, thus partially adopting as our own the results of an adjudication system's tangling with the internal perspective. For most of us most of the time, our best answer to the internal perspective's "What is the law?" is the same as the answer to the external perspective's – it is whatever judges decide it is. Even judges, high on talent and expertise if not time, routinely resort to the external perspective in defense of their judgments when they cite non-binding precedent in string cites. A second justification for internalizing the external perspective is the second-order values of fairness and predictability. Appreciation for these pragmatic benefits might reasonably lead someone to modify their first-order interpretation of the law. In fact, these are the values that can justify the rule of precedent – the formal incorporation of the external into the internal perspective. But the interests of fairness and predictability may support incorporation even in the absence of a formal rule of precedent. As pithily captured by Justice Brandeis, "in most matters it is more important that the applicable rule of law be settled than that it be settled right." Thus, whether out of humility or for fairness and predictability, there are important reasons for our internal perspective to defer to the external perspective.

But there are serious barriers to internalizing the external perspective. First, the humility we owe to the external perspective is partially a function of how reliable legal decisions are. In other words, how much we trust in the wisdom of a particular judicial decision depends on how often a different judge – or even the same judge at a different time – would decide the case differently. But reliability is difficult to assess. With rare exception, we only get to see one decision per case, so we don't get to observe judicial disagreement directly. Researchers and administrators have sought to overcome the problem in two ways (Grunwald 2015). In one approach, judges are asked to respond to simulated materials so that we can observe different judges' responses to the same stimuli. The main problem with these "inter-rater reliability" studies is external validity, as disagreement in the simulated environment may poorly describe real-world disagreement. The other solution has been to identify disparities in rates at which judges making actual, real-world decisions. While these disparity studies have high external validity, they suffer from problems with internal validity because they can miss important dimensions of disagreement. Two decision makers might, for example, each decide 50% of cases in favor of claimants. A simple comparison of averages would detect no difference between the decision makers, but it's possible that they would actually disagree in 100% of cases (Fischman 2014a).

The second barrier to internalization is the difficulty of mapping the external perspective in a timely manner. While we can simply wait for the outcome of particular cases, we'll often want to form our internal perspective before the outcome is decided. Obviously, this is at least true for the judge charged with deciding the outcome. A formal or informal rule of precedent partially solves the problem. With publication of outcomes and reasoning in prior cases, judges can look to past results in similar cases to guide and constrain their decisions in current cases. But the shortcoming of precedent as a solution to inconsistency is in the ambiguity of "similar." If interpreted narrowly, such that even minor differences in fact patterns can distinguish cases, precedent fails to constrain decision-making. If interpreted broadly, such that cases with more substantive differences are still deemed similar, precedent suffers from the familiar problem of the over- and under-inclusiveness of rules. The problem is particularly acute in legal areas where factual idiosyncrasies are important to merits, and thus we generally see only limited use of traditional precedent in areas like social security disability, parole, or asylum. Moreover, even if precedent can successfully constrain decision-making with minimal over- and under-inclusiveness, it may not be reliably produced: in a system where the outcome and rule are largely dependent on the preferences of the judge assigned to decide a case, precedent may deserve little humility-induced deference.

Statistical precedent can substitute or complement traditional rules of precedent, helping to overcome the ambiguity of "similar." Statistical precedent, rather than locating similarity in a few particular cases, tethers decision-making to statistical patterns in a dataset of historical decisions. But it is still in its infancy. To date, proposals and implementations of statistical precedent—such as disseminating information about peer decision rates, establishing decision quotas, or targeting for review decisions from judges whose rates are abnormal—have addressed reliability with little precision. In brief, because they only target raw inter-judge disparities, they fail to address multi-dimensional forms of inconsistency.

In this article, I explain how predictive technology can help to overcome the problems of assessing reliability and of mapping the external perspective so that statistical precedent can be used to intelligently guide and constrain decision-making. More specifically, predictive models of individual judges can be used to dramatically improve upon the estimates of reliability offered by disparity studies. And predictive models of adjudication systems—models that treat an adjudication system as a unitary entity—can smooth over the noise and inconsistency in decision making, helping to bring statistical precedent into maturity.

The paper proceeds as follows: Part I introduces readers to the basics of machine learning. If predictive technology is to play a vibrant role in the improvement of adjudication systems, machine learning will have to be demystified and understood by legal administrators and scholars. Part I is a contribution to that effort. Part II briefly discusses the applications of machine learning to the internal perspective's interpretation of law, as well as their shortcomings. Part III provides a basic sketch as to how predictive models of decision making expand the ability to assess reliability in adjudication systems and to establish high-quality statistical precedent. Part IV considers barriers to and problems with using machine learning to pursue reliability in decisions, including legal and data-based hurdles. Part V briefly concludes.

2 An Intuitive Introduction to Machine Learning

Imagine a newly appointed parole board commissioner. Her previous experience as a sheriff means she has some experience with the inmate population, but it certainly doesn't make her an expert in recidivism. It's time to make her first decision. For whatever reasons, a few characteristics of the potential parolee keep coming to her mind as she contemplates her decision: the inmate is a 39-year-old white male, currently serving year 20 for a rape and murder, has two prior robbery convictions, one previous assault conviction, a small facial tattoo, a 10th grade education, an acceptance letter from a half-way house, no offspring, two write-ups in prison for cell phone use, a history of moderate alcoholism, and a two-year sobriety token from Alcoholics Anonymous. During the hearing, the inmate seemed sincere in his regret and intent to avoid criminal activity, but having been fooled many times as a sheriff, she has learned to question those instincts. The Parole Board keeps extensive, computerized historical data on recidivism. She decides to take a look to see if it can help. First, she looks up last year's violent-crime recidivism rate: 15%. It's lower than she would have thought, so that's useful, but she wants to know more about this particular inmate's likelihood of recidivism. So, she searches for the recidivism rate of 39-year-old white males who served 20 years for a rape and murder, had two prior robbery convictions, one previous assault conviction, a small facial tattoo, a 10th grade education, an acceptance letter from a half-way house, no offspring, has two write-ups in prison for cell phone use, a history of moderate alcoholism, and a two-year sobriety token from Alcoholics Anonymous. No results. That was obviously too specific. So instead she constructs a search that is more moderate on the specific/general scale: males between thirty and forty who were serving time for rape and murder, had at least one prior assault conviction, did not graduate high school, and had a facial tattoo. There are 8 such cases, and five of those individuals (63%) went on to commit another violent crime – the sample size seems too small to trust. Facial tattoos aren't very common: maybe she can get a bigger sample by eliminating that as a search parameter. She gets 213 results with 21 cases of violent recidivism (10%). But is that the best estimate? Is a facial tattoo an important correlate of crime that this estimate ignores? Or should she maybe try additional searches. For example, maybe she should have tried dropping the education parameter rather than the facial tattoo – and maybe that even lets her put the cell-phone write-ups back into the search?

A branch of machine learning, called supervised learning, attempts to provide optimal solutions to problems like the above: with a supply of predictor variables (e.g., age, facial tattoo, education) and outcome variables (e.g., violent recidivism), we can let a machine train itself to identify which combinations of predictor variables are most helpful in predicting the outcome. This part of the article provides an introduction to that process, introducing readers to some of the basic vocabulary and concepts of machine learning.

2.1 The Bias-Variance Tradeoff

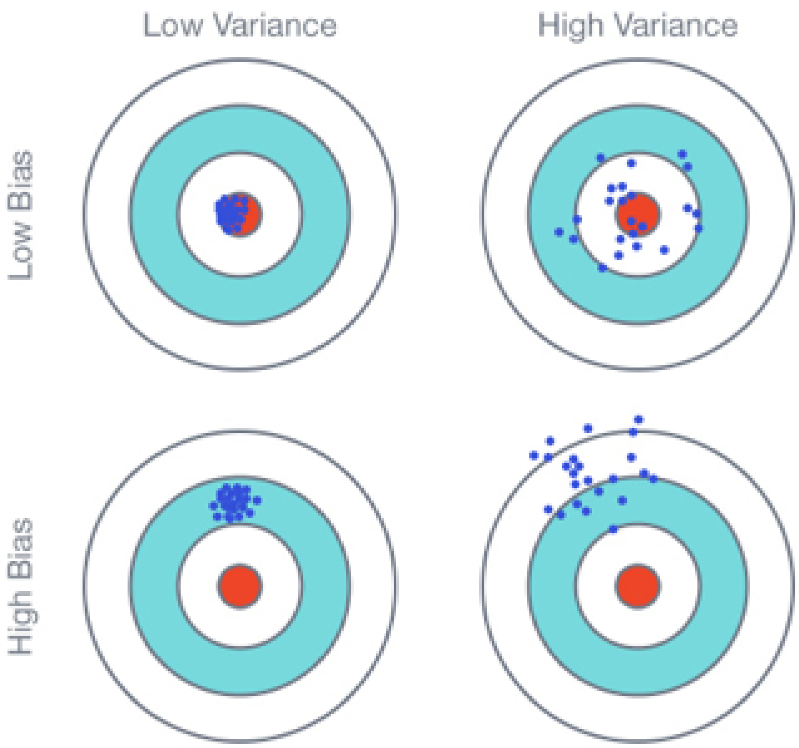
The above scenario captures the key issue addressed by predictive technology: the battle between bias and variance. Image 1[2] helps to convey the concepts. We want a low bias, low variance estimate, as represented by the target in the upper left corner. Unfortunately, lower bias generally means higher variance, and lower variance generally means higher bias. Why? Consider again the experience of the newly appointed parole commissioner. An unbiased estimate of an individual's likelihood of violent recidivism uses all information about that individual—it aims for the center of the target. But by using all of the information, the number of comparable individuals (i.e., individuals with the same characteristics) dwindles, and any estimate based on such a small number of people is likely to be unreliable—our dart player is aiming for the center, but she has a shaky (high variance) hand. By ignoring information about the individual of interest, say by leaving the facial tattoo out of the search query, we increase the number of individuals we are basing an estimate on, but we move the aim away from the center of the target, towards individuals without a facial tattoo. The dart player's hand is steadier, but it is no longer aiming at the center.

Even worse, variance increases exponentially as more characteristics are added to the search. In a world with extensive electronic records, the list of available characteristics can be almost endless, so this "curse of dimensionality" can be a serious problem. For example, even with only 10 dichotomous variables (e.g., male or female, previous assault conviction or not, history of alcoholism or not), there are $2^{10} = 1024$ different types of people. Even with a moderately sized dataset of ten thousand, we'd expect only ten of each type of person. With such small sample sizes, estimates would have extremely high variance.

Fortunately, we don't have to choose between adding a characteristic to the search inquiry or simply ignoring it. With techniques like multiple regression, we can partially add characteristics to the "search inquiry" (the quotes are now necessary because the partial addition of characteristics involves mathematical operations that are more sophisticated than a simple search inquiry, and we'd be more accurate to now call it a statistical model). Rather than observing the recidivism rate for the rare 30 to 40-year-old white male convicted of murder who also has a facial tattoo, we could instead start with the recidivism rate for the much more common 30 to 40-year-old white males convicted of murder (with or without a facial tattoo). Worried that we've disregarded an important predictor of recidivism (i.e., worried that we've taken on too much bias in the effort to reduce variance), we could try

[2] Image from http://scott.fortmann-roe.com/docs/BiasVariance.html

Figure 1: The Bias-Variance Tradeoff



different methods of incorporating the facial tattoo as a predictor. We might, for example, see how facial tattoos are associated with recidivism rates for all inmates and add that to our baseline estimate for 30 to 40-year-old white males convicted of murder. Or perhaps we suspect that the association is special for those convicted of murder, so we instead check how facial tattoos are associated with recidivism for that subgroup.

The problem is now even starker: with all of the choices about which variables to add, which to add partially, and how to add them partially, how can we possibly figure out the "search query"—the statistical model—with the best bias/variance balance? In other words, how do we find the dart player with the right mix of aim and steadiness?

2.1.1 Training, Validation, Testing, and Application Sets

Finding the right model, the model with the right mix of bias and variance, requires a good testing procedure. Allow me to switch metaphors. Imagine that a school principal is trying to find the student that will get the highest scores on next year's SAT so that can nominate the student for a national SAT competition. She has a large set of SAT questions from previous years' tests. What should she do with them? Most of her students have never even seen an SAT question, so she knows there isn't any point in simply holding a test now and choosing the highest scorer—she'd miss out on talented students who only performed poorly because they were unfamiliar with the SAT. She could release all of the questions to let her students get the most possible practice with the SAT, but then how could she assess the students – the students with the highest scores might just be the students who memorized the answers to all the questions. As a general matter, it's clear what she should do: use most of the questions to let the students train, but save some the questions so that student performance can be evaluated later.

Machine learning practitioners do the same thing as the school principal. They use some of the dataset, called the "training set," to fit each model. For a simple search inquiry model like that contemplated by the newly appointed parole commissioner, that training would be as simple recording at the average recidivism rates for various subgroups in the training set. They then test to see how well the trained models perform on the remaining data, called the "validation set." This helps to make sure that the models aren't "overfitting" the data—performance on the validation set makes sure that the models are truly learning something about the world and not just "memorizing the exam answers." It helps us to distinguish the models that contain useful information (the signal) from the models that contain useless information (the noise). The use of a validation set is one of the key moves in finding a model with a good mix of bias and variance.

But might the school principal do even better? Perhaps. Her role has been passive so far, but she could get more active in helping students to reach their potential. More specifically, she might take students who perform well on the validation test and encourage them to tweak their study habits. In the same way, machine learning practitioners often identify promising models and change them, or "tune" them, for better predictions. They then validate the models again with the validation set, and retune them, and so on. But this raises a concern. If we're changing the models according to their performance on the validation set, we reencounter the problem of overfitting that inspired the use of a validation set in the first place: we may now be overfitting models not just on the training set, but on the validation set as well. To continue the student testing metaphor, if the principal keeps retesting students with the same test questions, she should worry that the high performing students are really just the students who have memorized the test questions. To solve this problem, machine learning practitioners sometimes reserve another portion of the data, called the "test set," to serve as a final source of validation for the most promising models.

A testing procedure also needs to be coupled with a grading method. We need a metric for choosing a good statistical model, and an oft-used and intuitively appealing metric is mean squared error, or MSE. The MSE for a predictive technique is simply the average of the squared differences between the technique's predictions and the actual outcomes. So, for example, assume that a statistical model predicts that the likelihood or recidivism for three individuals is 68%, 22%, and 4%. Further assume that only the second individual recidivates, thus getting a score of 1 while the other two get scores of zero. The MSE is $((0 - .68)^2 + (1 - .22)^2 + (0 - .04)^2)/3 = .36$. With machine learning, we are often trying to find a statistical model that will minimize the MSE. Why is minimizing the MSE appealing? Because of the MSE's intimate connection with the average. For some intuition, consider the following puzzle: you are given a list of numbers, and you have to choose some value that, when subtracted from each number in the list, squared, and averaged, produces the smallest number. The value that solves this puzzle is simply the average of the list of numbers. In other words, the average minimizes the MSE. The puzzle also works in reverse: minimizing the MSE gets us closest to the average. So, one can think of the effort to find a statistical model that minimizes the MSE as an effort to most accurately estimate group averages (e.g., recidivism rates).

In summary, machine learning practitioners often divide their dataset into (1) a training set for the purpose of training models, (2) a validation set for the purpose of evaluating, tuning, and identifying the best models (with a metric like mean squared error) and (3) a test set for final evaluation of the top model(s). But with this setup, data taken for one purpose is data taken from another purpose, and data can be extremely valuable. The technique of cross-validation, described in the next section, is a powerful trick for avoiding the tradeoffs between consuming data for training and validating.

2.1.2 K-Fold Cross-Validation

The school principal, concerned with getting her student a lot of practice, might reserve only 10% of the test questions for validating her students. But she might have difficulty getting accurate student assessments with such a small test. Of course, were she to dedicate a higher percentage of the test questions to validation, the students would have less to practice with. In brief, while reserving more questions would give her a better chance of finding the best students, it would also – by limiting the practice students get – reduce the chances of even creating the best students in the first place. It is a seemingly inescapable tradeoff between training and validation.

But the tradeoff can be avoided with the training and validation of statistical models through k-fold cross validation. In the effort to find the best models, we can use almost all of the data to both train and validate. The technique works as follows. For illustration, let k equal 10:

- Randomly split the data (less a possible test set) into k=10 sets.
- Train models on sets 1-9. Validate them on set 10.
- Train models on sets 1-8 and set 10. Validate them on set 9.
- Train models on sets 1-7 and sets 9-10. Validate them on set 8.
- Continue the training process until all ten sets have been used for validation.
- Select the best model.
- Train the best model on the full data for application or further testing.

With k-fold cross-validation, all of the data can be used to validate the models. At the same time, the models are validated after being trained with a large portion of the data, assuring that strong models are not passed over simply because they have not had sufficient opportunity for practice. The trick works because, unlike a student, a model's memory can be erased – after a round of training and validation, memory of the both the training and validation can be erased, allowing for a new round of training and validation.

The choice of the value of k is largely a tension between training and computing time. By increasing k, models are allowed more training before validation. With k=10, the models are trained with 90% of the data. With k=20, 95%. With k=100, 99%. But increasing k also increases the rounds of training and validation, a process that can take a significant amount of time. Because the returns from increasing k diminish rapidly (e.g., increasing the rounds of training/validation from 10 to 20 to 100 only increases the percentage of data used for training from 90% to 95% to 99%), it often makes sense to set k at a number near 10.

2.1.3 Choosing the Candidate Models

The previous sections have yielded the goal: choose a statistical model that we think will have the lowest MSE in application, using as our criteria the model that has the lowest cross-validated MSE. But what should the candidate models be? It's often impossible, and almost always unwise, to validate every possible model—the possibilities are near infinite, and even if they weren't, validating all of the possibilities would generally take too much time and undermine the effectiveness of the validation procedure (if the principal tests 1 billion students, the best performing students probably had a lot of luck on their side).

So how do we identify a reasonable number of models to enter into the validation process? A human could create some. The parole commissioner might, for example, chose to enter an assortment of different search queries that pop into her mind, and she might enter some set of regressions into the validation process as well. But "just choose some" isn't much of answer.

Researchers in machine learning have been making advances in the development of self-constructing models. Freely available statistical programs now provide anyone with access to machine-learning algorithms like random forests, LASSO regressions, gradient boosted trees, neural networks, and support vector machines. These types of algorithms can automatically select predictive variables for inclusion in a model as well as select how to combine those variables in the model (e.g., whether to interact two variables). Moreover, those algorithms can easily be tweaked, or "tuned," so that they construct multiple models that can be included in the cross-validation process.

To help readers better understand how algorithms can select variables and how they interact, this section provides a very brief introduction to just one type of such an algorithm, random forests. But in order to understand random forests, one has to first understand Classification and Regression Trees (CARTs). The core idea behind CART is simple and is perhaps best conveyed by an example. Suppose the newly appointed parole commissioner wants to use a CART to estimate the likelihood that inmates will violently commit recidivism. Label the dataset she is using as "Node A." CART's first step will be to find the single characteristic that can best split Node A between those who committed a violent crime upon release and those who did not. Call these two new sets of data Node B and Node C, respectively. Perhaps age was the characteristic that accomplished the best split: parolees over 30 years of age are much less likely to violently recidivate. CART the proceeds to make the best splits of the new nodes. Perhaps Node B, consisting of parolees under 30, is best divided between those who committed a violent crime upon release and those who did not by splitting on prior violent crimes: those who had two or more previous convictions for violent crime are much more likely to commit a violent crime in the future. Call these two new datasets Node D and Node E. And perhaps Node C, consisting of parolees 30 or older, are best split into recidivators and non-recidivators by their disciplinary record in prison: those who were free of disciplinary infractions for the previous three years were much less likely to recidivate. Call these two new datasets Node F and Node G. CART continues this process, best splitting each node into two new nodes until there are no more non-trivial splits to make. At the culmination of this process, each historical inmate will be in an end node that is defined by the splits that got it there. For example, one end node could be parolees under 30 who had fewer than two previous convictions for violent crime, were older than 25, had a facial tattoo, had no children, and had a history of drug abuse. Perhaps four of the five parolees in this end node violently recidivated, so an inmate with these characteristics would have a predicted 80% of violent recidivism. A different end node might consist of parolees who were over 30, were free of disciplinary infractions for the previous three years, had a less than a high school education, had an acceptance letter from a half-way house, had a history of moderate alcoholism, and had a two-year sobriety token from Alcoholics Anonymous. Perhaps only one of the eight parolees in this end node recidivated, so an inmate with these characteristics would have a predicted 12.5% chance of violently recidivating.

But there's a major problem with using CARTs for prediction: they tend to overfit the data, taking on too little variance and being too unbiased. They're similar to the student who simply memorizes the exam answers, and they're thus unlikely to perform well in the validation process. First, the end nodes they create are generally too small for reliable estimates – an end node with five data points represents a highly variable estimate. Second, CARTs track the noisy features of the data too closely, making early splits that aren't holistically optimal (e.g., maybe splitting on age is just barely the better as first split than splitting on previous convictions, and splitting on previous convictions would have allowed for better downstream splits). Random forests are a solution to the low variance of CARTs, adding variance through two means. First, rather than grow one CART with all of the training data, random forests consist of many CARTs, each grown with only a random sample of the training data. This helps avoid too closely tracing the unimportant details of the data. Second, instead of splitting each node based on the single best variable of all of the variables, a CART grown for a random forest splits each node with the single best variable of a random selection of the variables. Again, this helps prevent overfitting. The random forest is now the collection of these CARTs, and each CART can be understood as casting a vote for each parolee depending on whether it classified the parolee as likely to recidivate or not. If, for example, 60% of the CARTs classified parolees with a certain set of characteristics as likely to recidivate, an inmate with those characteristics would have a predicted 60% chance of violently recidivating according to the random forest. The random forest algorithm can thus automatically create a strong candidate model for testing in validation process.

2.1.4 Ensemble Learning

Imagine that our parole commissioner has identified a set of candidate models. Perhaps she has used her intuition to build some of them (e.g., some basic search queries that she thinks could yield helpful estimates), but she also used an assortment of algorithms like random forests that can self-construct. She uses 10-fold cross validation and calculates the MSE for each candidate models' predictions. The winner (i.e., the model with the lowest mean squared error) is a random forest that splits nodes based on a randomly selected 30% of the variables. She might then decide to use that random forest to help her estimate recidivism risks.

But our parole commissioner can probably do better with "ensemble" learning. Instead of just using the one best model, she could use some a collection of the models. Perhaps, for example, an average of the three best models' predictions would be best. Or maybe a random forest predicts really well for some type of parolees, but a Lasso regression generates better predictions for other types. Or maybe the models should be weighted according to their cross-validation performance. There are all sorts of ways that different models' predictions could be combined to create ensemble predictions, and each different method of combining the predictions is yet a new predictive model.

How does the parole commissioner identify which combination of the models she should use? She faces a very familiar problem: she wants to identify the candidate combination that has the right mix of bias and variance! How does she find the best candidate combination? By doing the same thing she did to identify the best candidate model: cross-validating. The details of ensemble cross-validation, which involve two layers of cross-validation, are a little beyond the scope of this introduction, but the idea shouldn't be objectionable. Combinations of cross-validated models are just models themselves, and as explained above, models can be cross-validated.

2.1.5 Final Evaluation of the Winning Model/Combination

With the best model or combination of models chosen, there remains a key question: how good is it? There are various measures for how well it performs on the test set (or the validation set if it hasn't been overused), but two of the most common and intuitive measures are the correct classification rate (CCR) and the area under the curve (AUC). The correct classification rate is exactly what it sounds like: what percentage of cases does the model correctly classify? But this can be unsatisfying. For example, if there is low base rate of recidivism like 20%, a model could have a high CCR (80%) simply by predicting that everyone has a 20% chance of recidivating (and thus classifying everyone as non-recidivators). Such a model would be worthless – it would have no predictive power. On the other hand, another model could have considerable predictive power but perform no better in terms of CCR. Such a model could succeed in identifying inmates that are significantly more likely and less likely to recidivate – perhaps it correctly predicts that half of inmates have zero percent chance of recidivating and the other half have a 40% chance. AUC does a better job of capturing this difference in predictive power. Intuitively, the AUC can be understood as the probability that a model rates randomly selected 1 (e.g., a recidivator) as more likely to recidivate than a randomly selected 0 (e.g., a non-recidivator). Of course, what we ultimately care about is whether the model is useful for its intended purpose. While measures like CCR and AUC can be helpful indicators of usefulness, more direct tests are often possible. For example, one might judge a model of recidivism by whether its decisions outperform the decisions of human judges (Lakkaraju et al. 2017).

3 The Limits of Machine Learning in the Internal Perspective

Machines, in general, are of limited use in interpreting the law. While platforms like Turbo Tax can fruitfully apply if-then logical statements for clear applications of law, as soon as things get complicated, if then statements are no longer useful. They cannot hermeneutically engage with the law.

But machine learning can inform the internal perspective through evidence-based decision making (EBDM). With EBDM, rather than relying on unaided judgment, decision-makers use predictions of some important outcome to assist in making decisions. The above introduction to machine learning was also an introduction to EBDM. In that scenario, that outcome is recidivism, the primary issue for many decisions made in the criminal justice system, including parole and bail decisions. Research in EBDM is progressing rapidly in the criminal justice context, with Richard Berk's work on using machine learning to estimate criminal risk leading the way (Berk 2012).

Despite the progress in criminal justice EBDM, it is a poor fit for many, if not most, of our adjudication systems. The problem is that we lack objectively measurable outcomes that could be used to guide many of our most important adjudications. How do we know if an employer's behavior constituted sexual harassment? If a nursing home is adequately treating its residents? If a child is being abused or neglected? If a patent should or shouldn't be issued? If a disability claim should be granted? If asylum should be granted? We have yet to find outcome metrics that could reliably answer these questions. Without such metrics, EBDM has little to offer. At least as of now, our collective judgment is the best we have.

And even when it we do seem to have objectively measurable outcomes, the objectivity may be an illusion. Measures of recidivism, for example, may be a poor approximation of actual recidivism. If certain types of people (e.g., racial minorities) are policed more heavily than others, then the more heavily policed individuals will tend to have higher measured rates of recidivism even if they commit crimes at the same rate. Or some parolees may simply be better at evading detection. If these gaps between measures of recidivism and actual recidivism are large, then than predictive models that rely on measures of recidivism will be inaccurate. As the computer science lingo states, "Garbage in garbage out."

Finally, even if we do truly have objectively measurable outcomes, there is still the problem of selection bias. Legal decisions, by their nature, generally have important effects. For example, in the criminal justice system we only see outcomes for the released. While the released and non-released may look similar on the variables we have access to, we have good reason to suspect that the released inmates are different than the non-released inmates in ways we can't detect—judges, who presumably get to consider more characteristics of the inmates than the algorithm does, decided to release the former but not the latter. If the suspicion is correct, an algorithm that was constructed using data on released inmates would not be expected to perform well in predicting the criminal risk of non-released inmates. The problem extends to the most adjudication systems: because decisions are an important determinant of the outcome data we see, it is precarious to use that data to inform decisions (Lakkaraju et al. 2017). We might imagine, for example, the Social Security Disability tracking employment outcomes of those who submitted a disability claim and using that data to build a predictive model of employment. Given that disability claims are only supposed to be granted where an individual is not "able to engage in any substantial gainful activity because of a medically-determinable physical or mental impairment," predictions about the individual. But, of course, the decision on a disability claim dramatically affects whether an individual finds gainful employment: those denied disability, it is reasonable to presume, are much more likely to seek employment in order to offset the denial of their disability claim. But if the disability decision affects the outcome, the outcome cannot safely guide the decision.

4 Existing Approaches to Internalizing the External Perspective

The empirical analysis and pursuit of reliability in adjudication systems is not new, but current approaches are severely limited. Below, I briefly review the state of existing research into reliability and the primary methods used to promote it.

4.1 Assessing Reliability

Empirical assessment of reliability has generally taken one of two forms.[3] In the first, adjudicators are asked to make separate decisions after reviewing the same case materials or simulated scenarios. The main problems with this approach are expense and the difficulty of simulating realistic decision-making environments and materials. It is costly and difficult to have adjudicators spend time on fake cases and to create realistic and representative scenarios. In the second from of analysis, researchers and administrators identify disparities in actual decisions. The primary problem with this approach is that it has so far been limited to observing differences in averages between judges, and this can dramatically misrepresent the level of inconsistency.

[3] A third approach, aimed directly at error rates rather than inconsistency, is to use some proxy for whether decisions are correct (Benitez-Silva, Buchinsky, and Rust 2004).

4.1.1 Inter-rater Reliability Studies

In inter-rater reliability studies, adjudicators are asked to make decisions after reviewing identical case materials. Researchers can then analyze disagreement rates to assess the consistency of the adjudication system. In some contexts, the case materials are actual cases simply assigned to multiple adjudicators. Daniel Ho, for example, showed that Washington state food safety inspection unit disagreed in 60% of cases when assigned to evaluate the same establishments (Daniel E. Ho 2017). But because that can be time-consuming for adjudicators and an unreasonable burden on parties/claimants in systems where they dynamically participate in the process, inter-rater reliability studies can also be conducted with some subset or simulation of case materials. The Veterans Disability Administration, for example, evaluates reliability of its claims processing with a questionnaire that describes a brief scenario.[4]. Although more common in highly administrative settings , inter-rater reliability studies have also been used to assess reliability in traditional court systems. These studies are uniformly conducted with simulated cases rather than actual cases (Austin and Williams III 1977; Partridge and Eldridge 1974; Van Koppen and Kate 1984).

In addition to the expense, the main problem with using inter-rater reliability studies in adjudication systems is the lack of external validity. The scenarios that are often used may not be representative of actual cases, and adjudicators might treat cases differently when they are aware their decision will have no real-world effect or when they know that reliability is being evaluated. It would be difficult, for example, to entice federal judges to spend the time and energy on a simulated employment discrimination case as they would on an actual case. And even if it were possible, those judges might seek to protect their or court's reputation, putting their ideological preferences aside if they knew that their decision was part of an effort to measure levels of disagreement among judges.

4.1.2 Disparity Studies

In light of the limitations and expense of inter-rater reliability studies, it is not surprising that reliability is often studied through disparities in actual decisions. For example, adjudicators differ in the rates at which they grant asylum to refugees (Ramji-Nogales, Schoenholtz, and Phillip G. Schrag 2007), social security disability benefits (Nakosteen and Zimmer 2014a), and extended prison sentences (Anderson, Kling, and Stith 1999). Generally, these disparity studies leverage the fact that adjudication systems frequently make use of random or as-if random assignment of cases to judges, allowing them to attribute the cause of disparities to differences in adjudicators' preferences.

By analyzing real decisions in real cases, this approach to the study of reliability avoids the problems with external validity that plague inter-rater reliability studies, but it suffers instead from problems with internal validity. Because disparity studies rely on decisions in different cases (that are only the same on average) rather than decisions in identical cases, it is difficult to detect the true amount of disagreement (Fischman 2014a). In brief, differences in averages may leave some disagreement uncovered. Consider an example where one judge

[4] http://www.gao.gov/assets/670/667027.pdf

reverses 20% of cases and another judge reverses 30% of cases. Assuming randomization of case assignment and a sufficiently large sample of cases, the judges would obviously disagree in 10% of cases. But the level of this counterfactual disagreement may be much higher: if they had decided the same cases, the two judges might always reverse different cases, meaning they would disagree in a full 50% of cases. The concern is not merely theoretical. In one of the early inter-rater reliability studies on federal sentencing (Partridge and Eldridge 1974), a key finding was that the "vast majority of the judges are sometimes severe relative to their colleagues and sometimes lenient." A simple comparison of actual sentence lengths given by judges would have missed this and vastly overstated the reliability of federal sentencing.

Recoding of decisions along lines that better capture disagreement or subsetting cases for independent analysis can help overcome these problems (Fischman 2014a). For example, a comparison of reversal rates in employment discrimination cases might reveal less disagreement than a comparison of the rates at which judges decide in favor of the plaintiff. In effect, by recoding outcomes as pro/anti plaintiff rather than reverse/affirm, the analysis of disagreement is conducted on two subsets of cases: those where the plaintiff appealed and those where the defendant appealed. The difficulty, of course, is identifying the appropriate subsets of cases. Adjudicators may disagree along an assortment of dimensions, and even if an experienced participant could articulate those dimensions, there may be no useful measurements (e.g., some judges might find articulate litigants more persuasive, but in many instances, we may have not recorded raw speech data, much less quantified "articulateness" with it). Moreover, the dimensions that judges disagree over may vary between pairs of judges. While different preferences for plaintiffs prevailing might capture the majority of disagreement between Judge A and Judge B, it could be that different levels of aversion to reversing cases captures the main source of disagreement between Judge A and Judge C, while different preferences for well-written briefs captures the disagreement between Judge C and Judge B.

Despite the shortcomings of disparity studies, their findings can and have been instructive – even if they do overstate the reliability of an adjudication system, those overstatements can be alarmingly low, generating calls for system reform.

4.2 Promoting Reliability

We have an existing set of tools for reducing inconsistency in decision making, but precision-targeted statistical precedent is a powerful new addition. In this section, I briefly review some of the key existing approaches to limiting inconsistency.

4.2.1 The Rule of Precedent

In the United States, the rule of precedent has been the front-line defense against inconsistency in judicial decision making. By developing rules that bind future decisions, a rule of precedent can reduce inconsistency in outcomes insofar as some judges would have made decisions incompatible with the rule had it not existed (and do in fact make decisions compatible with the rule given its existence). While there have always been questions about how well precedent succeeds in constraining future decisions, the shortcomings of traditional precedent are stark in the modern era. The sheer number and factual complexity of contemporary claims have dramatically undermined the power of precedent. First, the explosion in the number of claims by individual systems (e.g., the 9th Circuit) has dramatically increased the amount of precedent, making it more difficult for systems to maintain coherent bodies of precedent that can reliably constrain future decisions. Second, government use of decentralized decision-making systems has aggressively expanded to more factually-intensive contexts where a rule of precedent is less valuable—adjudication of claims for social security disability benefits, parole, for example, frequently turn on differences between cases (e.g., credibility of parties) that are hard to usefully delineate with ex ante rules. In short, in many contexts, law cannot be moved further down the rule side of the rules versus standards spectrum without serious adverse consequences. The inability of traditional precedent to regulate decision-making in the contemporary world is perhaps best exemplified by the US Courts of Appeal. Once a bedrock for the rule of precedent, the federal appellate courts are increasingly abandoning it, designating less than 12% of opinions as precedent.[5]

4.2.2 Statistical Precedent

Though it does not have the pedigree of a traditional rule of precedent, statistical precedent is another tool for guarding against inconsistent decision making. Statistical precedent, rather than using stated rules in individual decisions, leverages historical statistical patterns in decision making to guide future decisions.

Experiments with statistical precedent have been limited so far, both in quantity and sophistication. At one point, the SSA targeted for review disability decisions by administrative law judges who had abnormally high approval rates (Krent and Morris 2013) and the Administrative Conference of the United States has newly recommended that the SSA consider reviewing "decisions from judges whose allowance rates are both well under and well above statistical average" [6]. In a classic work on inconsistency in social security disability decisions, (Mashaw 1985) suggested, though ultimately rejected, the idea that "disparities could be eliminated or sharply constrained by a quota system. State agencies or individual disability examiners could be given a grant rate (say 35 percent +/- 5 percent) for each time period (say a month). Awards would then be made on a comparative or relative basis and award rate disparities would virtually disappear." Though the SSA tried to implement such a quota, the ALJs aggressively fought for their "decisional independence," and the SSA eventually folded before the experiment could be completed (Hausman 2016).

The above types of statistical precedent are a clumsy response to inconsistency. Most importantly, while they might succeed in reducing disparities, they might fail to reduce inconsistency. As noted earlier, even if two judges grant claims in the same percentage of cases, they might grant claims in very different types of cases. The same problem plagues proposals to punish wayward judges, to encourage consistency by distributing information

[5] Table B-12, Annual Report of the Director: Judicial Business of the United States Courts 2016

[6] https://www.acus.gov/recommendation/improving-consistency-social-security-disability-adjudications

about peer grant rates, or targeting for review the cases that come from judges with deviant grant rates.

Despite a largely unsuccessful history, the use of statistical precedent is likely to grow precipitously. Some adjudication systems have begun exploring the potential of incorporating advanced analytics into their decision processes. The Administrative Conference of the United States, for example, has suggested that the Social Security Administration's Appeal's Council target for review decisions that "appear, based on statistical or predictive analysis of case characteristics, to have a likelihood of error or lack of policy compliance."[7] The Veteran's Benefits Administration has likewise explored the use of machine learning in its claims processing (Deutsch and Donohue 2009). And while details are a closely guarded secret, the Internal Revenue Service uses predictive technology to identify tax returns for auditing. Section IV.B sets the analytic foundations for the coming implementations of statistical precedent.

4.2.3 Other

Other tools for addressing inconsistency in decision making include centralized rulemaking, statistical rules, better judges, institutional design, and peer review. I briefly describe these tools below. Probably the most common antidote for inconsistency in adjudication is to move away from a standards-based system and more toward a rules-based system. The shortcomings of using rules to address inconsistency are well covered. While rules can reduce discretion and increase reliability, they can also introduce error by being over- and under-inclusive (Kaplow 1992). In short, because rules are constructed ex ante, it is difficult for them to account for the wide variety of events that actually occur, and they can thus generate suboptimal outcomes in cases.

Another option is decision matrices. The Federal Sentencing Guidelines stand as a prominent example, but other systems have experienced similar reforms. But like standard centralized rules, decision matrices are also often criticized for being too over- and under-inclusive. Critics of the federal sentencing guidelines, for example, have argued that they created excessive restrictions on judicial discretion and led to too many inapposite sentences (Alschuler 1991).

Another class of responses to inconsistency falls under the umbrella of what might be called "better judges." Either through improved hiring or improving existing adjudicators through professional development training, the notion is that better judges would be more consistent. But despite the intuitive appeal of the "better judges" approach to inconsistency, filling in the details is a difficult task. How do we identify and hire better judges? What should professional development programs develop, and how should they develop it? Despite frequent efforts to improve hiring or offer training to adjudicators, there is little to no evidence that they decrease inconsistency among adjudicators.

Institutional design can also be used to curtail inconsistency. A major design feature is the number of judges assigned to hear a given case. It is generally theorized that increasing

[7] ACUS Recommendation 2013-1, Improving Consistency in Social Security Disability Adjudications, 78 Fed. Reg. 41,352 (July 10, 2013)

the number of judges who participate in each decision is can increase reliability (Legomsky 2007). The notion is that larger decisional units will decrease the variance of decisions, both by mechanically limiting the power of extremist judges and by allowing for deliberation that can help prevent ill-considered decisions. But there are also countervailing pressures. First, research in psychology provides reason to doubt that group decision making limits extreme decisions. Groups can actually result in more decisions that are more extreme than the decisions that any of the individuals would have made (Moscovici and Zavalloni 1969). Second, while group decision making might help prevent sloppy decisions by adding eyes and minds, it's also possible that each adjudicator's effort and attention decreases as adjudicators are added because they can rely more on other judges (Halberstam 2015). Furthermore, increasing the size of decisional units can obviously be expensive if it requires hiring more judges, a necessity if we do not want to increase each decisional unit's caseload.

5 Machine Learning and Internalizing the External Perspective

In this section of the paper, I provide a non-technical account of how machine learning can be used to better measure inconsistency and control it through statistical precedent. For simplicity, the below discussion is restricted to binary legal decisions, but the key insights could be extended to legal contexts with non-binary decisions (e.g., criminal sentencing decisions).

5.1 Assessing Reliability with Targeted Disparity Studies

Recall the problems of measuring reliability in adjudication. Inter-rater reliability studies, because they involve observing different judges making decisions in identical cases, can provide highly accurately estimates the amount of disagreement between judges. In other words, they have high internal validity. But because judges do not duplicate effort on real cases as a matter of general practice, inter-rater reliability studies generally make use of artificial decision stimuli and environments, and those highly accurate estimates may be poor representations of disagreement over real world outcomes. In other words, inter-rater reliability studies have low external validity. In contrast, disparity studies, because they rely on decisions made in real cases, have high external validity. But because they rely on average differences in decision rates between judges rather than disagreement over outcomes in the exact same cases, they have low internal validity – they can fail to detect significant amounts of disagreement (e.g., two judges who both grant claims in 20% of cases may grant them in entirely different types of cases, meaning they could disagree in 40% of cases even though there is no disparity between them).

Machine learning can help us dramatically increase the internal validity of disparity studies, transforming them from studies of raw disparities to nuanced studies of disagreement. The key insight is simple: if we could perfectly predict the types of cases that Judge A grants and the type of cases that Judge B grants, we could also perfectly identify the types of cases in which they disagree – and thereby have a perfect measure of reliability. Of course, such perfect predictions will generally be impossible, but machine learning offers a powerful tool for getting as close as possible. By building a predictive model of Judge A's decisions and a separate predictive model of Judge B's decisions, we can make a best effort to identify the types of cases that Judge A is more likely to grant and the types of cases the Judge B is more likely to grant. This partition of the cases allows for our best estimate of disagreement.

The formal details of this approach to identifying disagreement are worked out and presented in Chapter 2, but the following should help crystallize the intuition. Suppose two judges, one generally viewed as liberal and another generally viewed as conservative, both reverse lower court employment discrimination decisions at a rate of 20%. There is a simple move that we imagine could reveal disagreement between the judges: rather than compare their rates of reversal, we can compare the rates at which they make "liberal" decisions. How might we do this? We could take divide the cases into two: those cases that, if reversed, would be liberal decisions (e.g., cases where the defendant won at the lower court) and those cases that, if reversed, would be conservative decisions (e.g., cases where the plaintiff won at the lower court). We then code the actual decisions as liberal or conservative based on whether they were actually reversed (e.g., code a decision as liberal if the defendant won in the lower court and the case is actually reversed). A comparison of the rate at which the two judges make these liberal decisions provides a new and—if our intuitions were correct—better estimate of disagreement.

But why stop there? Instead of (1) taking Judge A, who we think of as liberal, and Judge B, who we think of as conservative, (2) coding decisions as liberal or conservative, and (3) comparing the rates at which they make liberal decisions, why not instead (1) take Judge A and Judge B, (2) code decisions as either Judge-A-like or Judge-B-like, and (3) compare the rates at which the two judges make Judge-A-like decisions? The obvious objection is that we don't know what a Judge-A-like or Judge-B-like decision is. But we have a best answer from machine learning models of Judge and Judge B: a Judge-A-like decision is one where a Judge A is predicted as more likely to make the decision than Judge B is. Such a decision might be a reversal in cases where plaintiff won an employment discrimination at the lower court, but it might also be something much more complicated – something only a machine could pick up on. For example, Judge A might be more likely to reverse a case where the defendant won in lower court, the lower court judge was Judge Smith, and the case never went to discovery. And Judge A may not actually care about any of those factors! It may be that he's really just charmed by charming lawyers and that those factors happen to correlate with attorney charm. The success of a machine-learning approach to measuring reliability hinges only on the ability to predict decisions – not the much more difficult task of explaining them. Moreover, this analysis can be conducted separately for every pair of judges. While liberal/conservative my reasonably capture disagreement between Judge A and Judge B, it may do a very bad job of capturing the disagreement between Judge A and Judge C (or Judge B and Judge C). This flexible approach to disagreement, then, can avoid the trappings of unidimensional measures of disagreement, revealing significantly more inconsistency.

It is important to note that such measures of disagreement, despite being more accurate than measures based on raw disparities, are nonetheless only floors on the amount of inconsistency. Insofar as the machine learning models of the judges misclassify decisions (e.g., labeling a decision as a Judge-B-like decision when it is actually a Judge-A-like decision), it will still understate inconsistency (and overstate reliability). But machine learning can also help us solve this problem, allowing us to estimate a ceiling on disagreement in addition to a floor. The key here is the performance of a combined model of two judges that ignores any information about the two judges. By pooling their cases and testing to see how well their decisions can be collectively predicted, we can estimate the percentage of cases they would agree on. Formally, when the sample sizes of the two judges are equal, the estimate of agreement is one minus the correct classification rate ("CCR") of the pooled model, multiplied by two: (1-CCR)*2. This estimate is actually a floor on agreement, but it also serves as a ceiling on disagreement.

With all pairwise measures of inter-judge disagreement estimated, we can start to characterize the reliability of an entire adjudication system. For example, with the pairwise estimates in hand, it is fairly easy to provide estimates to questions like: what percentage of cases would be decided differently if, hypothetically, they had all been randomly reassigned to adjudicators? What percentage of cases could have been decided differently if they had be assigned to the most different judges? What are the case characteristics of the cases that judges disagree about most? Which judges' decisions patterns are the most extreme relative to other judges?

The accuracy of this approach to reliability is limited only by the ingenuity of our algorithms, the size or our datasets, and the detail with which we record case characteristics. If current trends continue, all of three of these factors will continue to increase, but they are already at a level such that we can start leveraging the benefits of machine learning. We can better detect whether inconsistency is at alarming levels. We can better understand what a legal decision means about the quality of a claim (e.g., is a denial of social security disability claim a strong signal that the person doesn't qualify for disability, or is it instead more a signal of which adjudicator was assigned to hear the case). We can better test the effects of legal reforms on reliability (e.g., how did the federal sentencing guidelines effect inter-judge disagreement?). In brief, the stage is set for the rigorous study of the reliability of our adjudication systems.

5.2 Promoting Reliability with Precision-Targeted Statistical Precedent

As discussed in section III.B.1, statistical precedent is not a new idea. Scholars and administrators have frequently proposed tethering future decisions to historical statistical patterns. But statistical precedent has so far been imprecisely targeted: because existing proposals are based on only unconditional averages in decision-making, their ability to reduce inconsistency is limited. With machine learning, we can generate precision-targeted statistical precedent. Return for a moment to our newly appointed parole board commissioner. Imagine that she has read about the limits of evidence based decision making and that she is now less inclined to rely on estimates of recidivism risk. But also imagine that she is hesitant to rely on her unaided judgment. What does she do about the 39-year-old white male who is currently serving year 20 for a rape and murder, has two prior robbery convictions, one previous assault conviction, a small facial tattoo, a 10th grade education, an acceptance letter from a half-way house, no offspring, two write-ups in prison for cell phone use, a history of moderate alcoholism, and a two-year sobriety token from Alcoholics Anonymous? Instead of relying on a machine learning algorithm for the best estimate of the inmate's probability of recidivating, she could leverage machine learning to help her answer a different question: what is the probability that my peers would release this inmate?

The question can best be answered with what I call a "systems model." A systems model excludes variables that are random with respect to the merits of cases. For example, a systems model would not include the identity of the judge randomly assigned to hear the case, the time of day it was randomly scheduled for, or whether it is decided on the day after a judge's football team won. By excluding these variables, the systems model smooths over the sources of inconsistent decision making. A perfect systems model is one whose predicted probabilities represent the probability that each individual case prevails – apart from sources of inconsistency. Thus, with a perfect systems model, a case with a predicted probability of .75 has a 75% chance of being granted – it is usually granted, but 25% of the time, perhaps because sometimes a particularly harsh judge in a harsh mood will decide it, the case will be denied. Such a perfect systems model has an intuitive interpretation: the predicted probabilities from such a model represent the percentage of votes that a case would receive if all judges cast multiple independent votes on the case. At the other extreme, a systems model may be perfectly imperfect: here, the predicted probabilities do not represent that probability that a case prevails. Instead, the predicted probability reflects a failure of the model to distinguish between different types of cases. Thus, a case with a .75 predicted probability does not have a 75% chance of being granted. Rather, 75% of cases with a .75 predicted probability have a 100% chance of being granted and 25% have a 0% of being granted. In reality, models will almost always be somewhere in between. And, as explained below, statistical precedent should take a different form depending on how perfect or imperfect a systems model is.

In addition to a systems model, sub-models of individual decision-making units (e.g., models of judges[8]) can be used to inform statistical precedent. Unlike a systems model, which is meant to smooth over the discordant patterns of a system, decision-unit models are specifically designed to capture inter- and intra-judge inconsistencies. Each unit's model is

[8] While individual judges will often be the most obvious and relevant boundary for decision-making units, units can be delineated in other ways. One might, for example, partition decision-making units into early-day decision makers and late-day decision makers, such that an individual judge's decisions—depending on whether she makes them early in the day or late in the day—are deemed to come from two different units. Because the decision-unit models are meant to capture disagreement, units should be delineated according to which boundaries maximize the detectable disagreement. But sample size is also a relevant consideration: more data allows an algorithm to trace more detailed patterns. Thus, one might also cluster together multiple judges to form a single decision-unit.

created by fitting the model on data generated just by that decision-making unit, and, as I explain in detail below, they can be used to better adapt statistical precedent to account for each unit's particular forms of inconsistency.

While predicted probabilities might provide useful information to a decision-maker, the full benefits of statistical precedent will often require that the they be converted into recommendations that can be consistently applied. A mere predicted probability could be interpreted and applied differently by different judges (or even by the same judge at different times), and, as I'll explain below, some of those interpretations will be better than others. The conversion process can vary along four key dimensions: (1) the grouping of cases; (2) the coarseness with which predicted probabilities are binned; (3) the extent to which targeted decision rates should be boosted so as to leverage collective wisdom; (4) whether recommendations are judge-neutral or judge-specific; (5) and whether the recommendations are for the initial decision or for review of the initial decision.

Importantly, statistical precedent can generally be used in a way that holds the overall leniency of a court system constant. Statistical precedent can thus avoid controversial issues about whether a court's mean decision rate is too harsh or too lenient. Again, I explain the details of holding such overall leniency constant below.

I discuss each of the five dimensions and the relevant considerations in detail below, but a few simple examples can help make the ideas more concrete. Suppose the Social Security Administration were to implement statistical precedent for their disability decisions. Imagine the SSA has built a system model that predicts whether an applicant will be granted disability benefits. Consider four possibilities:

(1) The SSA groups cases into two groups. In each group, predicted probabilities are binned into three bins: low quality (those cases with a predicted probability below .25), medium quality (cases with predicted probabilities between .25 and .75), and high quality (cases with predicted probabilities above .75). Those bins are assigned target rates that represent that are not boosted. For example, because 15% of cases in group one's low bin have been granted in recent history, the SSA targets a 15% grant rate for current cases in that bin. The recommendations are judge neutral, e.g., it is recommended that all ALJs target a 15% grant rate for low cases. Finally, the recommendations are directed to the ALJs making the initial decision rather than to the SSA Appeals Council.

(2) The SSA does not group cases and bins predicted probabilities into two bins: low and high. The low bin consists of all cases with predicted probabilities below .4. Those bins are assigned fully boosted target rates that represent: the predicted probabilities are interpreted as the indicating the merit of a case, and so it is recommended that ALJs grant 100% of the cases in the high bin and 0% of the cases in the low bin. The recommendations are judge neutral, as they do not alternate across judges. Finally, the recommendations are directed the SSA Appeals Council rather than the ALJ making the initial decision – if the ALJ's initial decision is contrary to the target rate, it is recommended that the Appeals Council review the decision.

(3) The SSA adopts a standard allowance-rate quota of 30%: all ALJs are instructed to target a grant rate of 30%, which is the percentage of cases that have been granted system-

wide in recent history. In the more formal terms of this paper, there is one group and one bin, and the recommendations are not boosted, are judge neutral, and are directed at the ALJs making the initial decision.

(4) The SSA chooses to use one group and not to bin. The target rates are a moderately boosted: based on estimates regarding the extent to which they represent case merit, each predicted probability is pushed toward 0 or 1 by varying amounts to set the target rate. The recommendations are judge specific, meaning that the predictions from judge-models are compared to target rates from the system models in order to form recommendations. It is recommended that the Appeals Council review cases where the judge predictions are far from the system predictions.

Below, I discuss in more detail each of the five key dimensions of statistical precedent.

5.2.1 Grouping Cases

Grouping cases can help statistical precedent target inconsistency with greater precision. To illustrate the core idea, consider a simple system with two judges and just two types of equally-prevalent cases, back pain and fibromyalgia cases. Judge A and Judge B both grant benefits in 30% of cases, but Judge A grants 60% of back-pain cases and 0% of fibromyalgia cases, while Judge B grants 0% of back-pain cases and 60% of fibromyalgia cases. A simple allowance-rate quota of 30% with one group would do nothing to promote agreement, as both judges could continue deciding cases in completely opposite ways. But by dividing the cases into two groups, we could begin to encourage cooperation. Not any division of cases would suffice. A random split, for example, would again do nothing to promote compromise: both judges could again continue deciding cases in a completely opposite way. But as the division of cases best reflected the division between the two judges, we could open room for agreement. In short, by creating two groups that track the disagreement between the two judges, one of back-pain cases and the other of fibromyalgia cases, we could maximize the possibility of compromise. In more formal terms, the optimal split is a group where Judge A's predicted probabilities are higher than Judge B's and another group where Judge B's are higher than Judge A's. By establishing an allowance-rate quota of 30% in fibromyalgia cases, Judge A would be encouraged to increase his grant rate from 0% to 30%, and Judge B would be encouraged to reduce her grant rate from 60% to 30%. It is conceivable that both judges continue to decide cases in the exact opposite ways even in the face of the two group quotas (e.g., Judge A grants fibromyalgia cases only for women and Judge B grants fibromyalgia cases only for men). But such continued high rates of disagreement would at least have to reflect new sources of disagreement, sources which could even be subsequently addressed with updated statistical precedent.[9]

[9] Updating statistical precedent to account for new forms disagreement can be complicated. Under the current scenario, statistical precedent means that Judge B is more likely to grant male fibromyalgia cases than Judge A, even though originally judge B was more likely to grant such cases. Three groups are needed to optimally target inconsistency. One group is cases that Judge A was more likely to grant originally. The second group of cases are those that Judge B was more likely to grant originally and continues to be more likely to grant. The third group of cases are those that Judge B was more likely to grant originally but is

The core idea extends to the more realistic and complex world of many judges with numerous dimensions of disagreement. Just as cases can be divided by the extent to which they best reflect disagreement between two judges, they can by divided by the extent to which they best reflect disagreement between many judges. While the full technical details are beyond the scope of this paper, the key intuition is simple: judges can be grouped together into decision-units by similarities in the decision-making patterns, and then groups of cases can be created according to variations in the ranking of predicted probabilities among decision units.[10]

While not all uses of statistical precedent are benefited, some forms can more precisely target inconsistency by grouping. But grouping is not costless. The primary cost of grouping is the taxation of judicial mental capacity. For meaningful application, a judge would have to become familiar with what a Group A case is: targeting a rate requires judgment about a case's merit relative to other cases in the group. As the number groups proliferates, the complexity of the judge's job increases as well.

5.2.2 Binning Predicted Probabilities

Within a group of cases, predicted probabilities may also be binned so as to help more precisely target inconsistency. For simplicity, assume one group and average grant rate of 50%, but that some cases have much higher predicted probabilities than others. Judges could be given access to each case's predicted probability, but such a raw, case-specific probability is only minimally useful. For example, how should a judge, armed with knowledge that the case she is currently deciding has a .30 predicted probability of being granted, alter her decision-making process so as to limit inconsistency? The answer is unclear. While the judge might intuitively use the predicted probability with good effect, an analytically coherent employment of predicted probabilities requires more guidance.

Binned predicted probabilities are a first step in providing that analytically coherent guidance. Predicted probabilities might, for example, be binned into low, medium, and high bins, with judges encouraged to grant cases in those bins at rates corresponding to the average predicted probability within the bin (e.g., 20% for cases in the low bin, 40% in the medium bin, and 80% in the high bin). But note that, standing alone, quotas within bins are a poor antidote to inconsistency. Generally, setting a target rate within bins is not as powerful as setting target rates in groups. Case groups are designed to target disagreement. And while bins can mitigate disagreement, they do so suboptimally. For example, it's possible that a bin of cases with an average predicted probability of .5 is created by half of judges granting all such cases and half granting none. In such case, a bin with a 50% target rate

less likely to grant under the first regime of statistical precedent.

[10] Inconsistency is optimally targeted by grouping cases into n! groups where n is the number of decision-units. For example, if there are three decision units, Judge A, Judge B, and Judge C, groups of case should be created such that predicted probabilities have the following relationship: $A > B > C, A > C > B, B > A > C, B > C > A, C > A > B, C > B > A$. If there are few or even no cases within any of those groups, it likely makes sense to discard them into a miscellaneous group. Similarly, it may make sense to choose only groups that show large deviations or to combine together groups where the deviations are similar in form (e.g., where $A >>>> B > C$ and $A >>>>> C > B$).

would have a high chance of reducing inconsistency, as both groups of judges would be encouraged to moderate their behavior and grant 50%. But it's also possible both groups of judges are granting 50% of cases in that bin already, just a different 50%. In that case, a bin quota would do nothing to encourage consistency. As such, binning of predicted probabilities is not an ideal solution to inconsistency. Instead, the main benefit of binning is the ability to attach analytically coherent recommendations to them when coupled with boosted predictions. Note that binning is generally not necessary when statistical precedent is used for targeted review rather than an initial recommendation.

5.2.3 Boosted Predictions

As briefly noted above, there are two basic ways that predicted probabilities from a systems model can be interpreted. First, we can view predicted probabilities as accurate with respect to groups of cases, but inaccurate with respect to individual cases. For example, under this interpretation, for a set of cases with predicted probabilities of approximately .40, 40% of those cases are indeed granted. But 40% of those cases would have been granted regardless of which judge decided them, when they decided them, or what mood they were in. And another 60% would have been denied regardless. But there is a second interpretation: predicted probabilities are accurate with respect to individual cases. Under this interpretation, a predicted probability of .4 means that if a case were hypothetically reentered into a judicial system, it would be granted 40% of the time and denied 60% of the time, with the decision differing only because it happened to be assigned to certain judges in certain moods.

Under the second interpretation, where predicted probabilities are accurate with respect to individual cases, the predicted probabilities have a normatively appealing elegance. A higher predicted probability means that more judges would more often grant that case than a case with a lower predicted probability. In brief, the predicted probabilities represent the percentage a votes a case would get in a world where each judge casts multiple independent votes in each case. On such a reading, according to basic democratic and crowdsourcing principles, the cases with more votes should be granted before those with fewer votes.

This provides an opportunity. Insofar as predicted probabilities reflect votes from the hypothetical world where all judges vote multiple times in every case, they can be used to provide additional guidance to judges. Rather than simply nudging judges to grant cases at a rate equal to their predicted probabilities, we should nudge judges to assess the quality of cases by their predicted probabilities. We should encourage judges to grant cases with high predicted probabilities (many votes) at a rate higher than their predicted probabilities (e.g., a case with 90% of votes should probably always be granted). Similarly, we should encourage judges to grant cases with low predicted probabilities (few votes) at a rate lower than their predicted probabilities (e.g., a case with only 10% of votes in favor of grant should probably never be granted). In short, we should seek to transfer grants from low quality cases to high quality cases, holding the overall grant rate constant while improving the overall quality of decisions. The transfer can be accomplished by boosting high predicted probabilities higher and lower predicted probabilities lower and calculating target rates accordingly.

The problem is that we will rarely know the extent to which a predicted probability is accurate with respect to an individual case, but we can make empirical progress. First, external benchmarks to help assess whether individual cases with higher predicted probabilities are indeed of higher quality than individual cases with lower predicted probabilities. For example, Chapter 3 leverages governor review of granted cases to assess the individual accuracy of predicted probabilities for California Parole decisions. In California, the governor reviews and has the ability to reverse decisions by the parole board that are in favor of release. If predicted probabilities are only accurate with respect to sets of cases, we have little reason to expect differential treatment by the governor: whether a particular case has a high or low predicted probability reveals nothing about its quality under the first interpretation – it only says something about the quality of cases with similar predicted probabilities. On the other hand, if the predicted probabilities are individually accurate, we would expect the governor to reverse low predicted probability cases at a higher rate than cases with a higher predicted probability. Indeed, we find that the governor is twice as likely to reverse cases in the lower half of predicted probabilities than cases in the higher half. We take this as evidence that we should at least moderately boost of predicted probabilities. While such appellate level results are often a convenient source of data, other external benchmarks can help assess the individual accuracy of predicted probabilities. A system might, for example, employ gold standard panels to assess the quality of cases to gather evidence about how closely connected case quality and predicted probabilities are. In other contexts, researchers may have access more objective proxies. For example, in the criminal justice context, recidivism stands as a prime candidate. A higher recidivism rate among those with lower predicted probabilities would provide evidence of individually accurate predicted probabilities, and researchers have recently demonstrated impressive levels of accuracy from a systems model of bail decisions(Lakkaraju et al. 2017).

A second class of techniques for assessing the extent to which predicted probabilities are accurate in individual cases involves building measures that proxy the extent to which noise variables explain the unexplained component of the systems model. Insofar as noise variables (e.g., the random assignment of judges and time slots) improve the predictive capacity of the model, the predicted probabilities of the systems model must reflect the individual quality of cases – if it is the noise variables that are limiting the predictive power of the systems model, then there is not a difference in quality between cases with similar predicted probabilities. Section IV.A outlines a method for aggressively detecting noise in a system. But it is also possible to generate more case-specific measures of noise, which can allow calibrating the boosting of particular predicted probabilities in accordance with an estimate of how much noise variables predict outcomes in that case.

There are three potential downsides to boosting predicted probabilities. First, predicted probabilities may be improperly boosted. If predicted probabilities are in fact not individually accurate, then boosting them could provide poor guidance to judges. For example, if the set of cases with predicted probabilities of approximately .7 are split between 70% of cases that are always granted and 30% of cases that are never granted, boosting those predicted probabilities to .9 would encourage judges to grant cases that none of them would have ever granted. It's thus important to assess the extent to which predicted probabilities are individually accurate. Second, boosting predicted probabilities can have complicated effects on systematic biases. Absent boosting, statistical precedent may capture and continue existing biases, but boosting could potentially exacerbate them. The effect stems from two sources. First, because boosting adds points to high predicted probabilities and deducts points from lower predicted probabilities, then cases whose predicted outcomes are deflated by systemic bias will be disfavored by boosting. Second, some methods of boosting specifically target those cases that are subject to the most inconsistency in decision making. Insofar as inconsistency is focused on cases that are subject to bias, such methods of boosting will be in even greater danger of increasing biases. Chapter 3 discusses methods for addressing bias in statistical precedent.

The final downside to boosting predicted probabilities is that it provides parties with more incentive to manipulate their variables in hopes of securing themselves a better recommendation. If parties, for example, know that an algorithm takes account of whether they have hired a lawyer or not, and hiring a lawyer is positively associated with a positive outcome, they could choose to hire a lawyer simply for the sake of the boost in their predicted probabilities. But the hiring of a lawyer may only be a signal of case quality, not a causal factor. Thus, the party who would not have otherwise hired a lawyer can gain advantage by doing so, masquerading as a higher quality case. Relatedly, it also increases the possibility that the mix of cases that enter a system is changed, as parties who would not have (or would have) entered the system now choose to do so (or not to) because of the boost given to them by statistical precedent. I discuss these problems in more detail in Section V.

5.2.4 Judge-Specific v. Judge-Neutral Recommendations

Statistical precedent can be judge neutral or judge specific. Judge-neutral guidance does not differ by judge – all judges are given the same nudge (e.g., a recommendation to target a grant rate of 30% for cases in a particular group). Judge-specific precedent, in contrast, adapts to the patterns of the particular judge making a decision or under review. It can do so by comparing the predicted probabilities from a systems model with the predicted probabilities from the decision-unit model of the judges whose decision is at issue. Cases where the difference between the two are large can be flagged, and a judge-specific recommendation can be generated.

Judge-specific precedent is more precisely targeted to inconsistency, and there is little to no reason not to use judge-specific precedent when it is employed for review of decisions. But there are strong benefits to judge-neutral precedent when it is being used to guide initial decisions. The downside to judge-specific precedent for guidance of the initial decision is the difficulty of converting it to analytically coherent recommendations. For example, we can alert a judge that she's 20% more likely to grant a case then her peers. She might even be able to make effective use of such information, perhaps by exercising extra caution or giving second thoughts to granting the case in front of her. But it is not possible to convert the predicted probability to a more specific recommendation. Judge-neutral precedent, in contrast, can easily be converted to target rates that can guide initial decisions.

5.2.5 Initial v. Review Recommendations

Statistical precedent can be used to guide either the initial decision or the to help target appellate review of outlier decisions. The benefits of using statistical precedent for targeted review are substantial. Importantly, the complications involved with grouping cases and binning predicted probabilities can be ignored. And, as noted above, statistical precedent for targeted review can more easily leverage the precision of judge-specific precedent. Moreover, employing statistical precedent for targeted review rather than guiding initial decisions mitigates the incentive that boosted probabilities provide for variable manipulation.

But the cost of reserving statistical precedent for only targeted review can also be significant. Because it necessitates duplication of judicial effort, appellate review is expensive. It is better that decisions are initially in conformity with statistical precedent rather than corrected to be so. The complexities and problems of initial review – grouping cases, binning predicted probabilities, increased risk of variable manipulation – may thus be worth taking on.

6 Problems and Barriers

Incorporating analytics into our adjudication systems will not be easy. There are both real barriers and problems to overcome. In this part of the paper, I briefly discuss some of those issues and suggest ways forward.

6.1 Data

Good data is a prerequisite for building good predictive models. The first step to good data is recording it. Unfortunately, many systems are still struggling to digitalize. But digitalizing standard data collection procedures won't be enough to fully unlock the potential of a machine learning approach to reliability. Because the tasks of the approach are predictive, information that is not ostensibly related to proceedings could still prove crucial. Collecting that type of information could be solved by an expansive ethic to data recording. But in the absence of such an ethic, or even in conjunction with it, the linking of externally created and aggregated datasets (the kind that private companies routinely now trade in so that they can better target advertisements) to court system datasets could dramatically improve the predictive power of models.

But even if the problem of recording data is solved, it is only the first step. The data will need to find its way into the hands of individuals and teams who can take full advantage of the data. While the internal staff of adjudication systems might be able to build and appropriately deploy predictive models by themselves, there is reason for skepticism. We might expect internal innovation to be slow for two reasons. First, the work of internal staff can be hampered by the very people whose decisions they are analyzing and, perhaps, seeking to regulate. Adjudicators are often in positions of prestige and power relative to administrators within a system. And, like most people, those adjudicators are unlikely to welcome scrutiny and oversight. The protracted battle between administrators and the SSA's administrative law judges stands as a prominent example.[11] Or consider the Federal Judicial Center's research on the federal courts. The Center does not conduct analysis that is dependent on judicial identity, going so far as to wipe datasets of judge-identifying information.

Second, internal staff are not sufficiently large or intellectually diverse enough to keep up with the state of the art. Instead, real advancements will likely require external assistance. But as someone who has spent much of the last five years trying to access data from adjudication systems, my hopes for extensive cooperation between administrators and external researchers are low. The fear, I think, is of bad publicity.

This unfortunate state of affairs is, in no small part, the fault of researchers and journalists. There is a certain "gotcha" style to much of the empirical work on adjudication, whether the work is journalistic or scholarly. The style makes sense in light of publishing incentives. Few readers are interested in balanced accounts of how a system handles intricate tradeoffs faced by adjudication systems. It's the charges of things like wide inter-judge disparities and racism in decisions that get researchers and journalists attention. Administrators thus face a high risk of bad publicity but little reward. I'm not sure how this trust problem can be solved.

The most likely path forward may be in researchers bypassing cooperation and instead collecting and linking datasets anew. Scholars studying adjudication systems are increasingly scraping the web for data and using text-parsing code to generate datasets, finding ways to study systems that don't seem to want studying. Perhaps once scholars can demonstrate to adjudicators, administrators, and the wider public that advanced analytics can be productively employed, cooperation becomes more common.

6.2 Status Quo Bias

Statistical precedent, because it leverages past decisions for guidance, can tether future decisions to the past in ways that may be harmful. For example, (Legomsky 2007) argues against quotas on asylum approval rates on the grounds that "rapid changes in human rights conditions would render the announced percentages continually obsolete." While that claim may be true in the asylum context, it is largely an empirical question: have decisional patterns—at the level captured by statistical models—actually changed dramatically over time? If not, we have less reason to think that they will in the future. Nonetheless, the danger may be real, and it is a good reason to hesitate in establishing strong controls on adjudicators (e.g., strict decision quotas). But the danger should not be overstated. If, on the one hand, there are rapid changes in the world that would dramatically affect the nature of the cases, we might also expect that those changes would be observable in such a way that administrators could alter or temporarily suspend regulations for affected cases.

[11] In one of the more recent installments of the battle, ALJs pushed back against administrators' setting of a goal that each ALJ decide 500-700 cases per year. *Association of Administrative Law Judges v. Colvin*, 777 F.3d 402 (7th Cir. 2015).

If, on the other hand, changes in the world are slow, then, as long as controls are not too tight, adjudicators can change their decision patterns to accommodate those changes, and subsequent algorithms can be updated to incorporate those accommodations.

But maybe concerns over status quo bias are less about changes in the world that lead to new mixes of cases and more about changes in the goals and values of an adjudicatory system. But again, as long as controls on adjudicators are not too restrictive, they can accommodate slow changes. Rapid changes in values pose a different problem. If, for example, statistical precedent is used to generate recommendations for decisions in the criminal justice system, those recommendations would be out of step with sharp shifts in views about the appropriate level of incarceration. It's possible, though, that such temporal flattening is, overall, a feature and not a bug. I imagine many readers would find a tethering to the status quo comforting in the context of a Trump administration's approach to immigration decisions. But whether it is a feature or a bug, statistical precedent's relationship to the status quo is probably overstated. In fact, the more serious concern may be that it too easily facilitates aggressive changes – an algorithm can easily be adjusted to better reflect new values by simple manipulation of the predicted probabilities, and those values can be easily transmitted to adjudicators (perhaps even in the form of quotas).

In the end, statistical precedent is probably neither inherently conservative nor progressive. It can be used in the service of different ideologies. The task will be remembering that and subjecting models to renewed scrutiny. While it may be tempting, we should not put automation on autopilot. Machine learning is merely a tool, and while a thorough understanding of how that tool works may be the province of technocrats, how we use that tool is not.

6.3 Legal Issues

The use of advanced analytics in adjudication implicates issues of equal protection and due process. Below, I highlight some of the issues that are likely to arise and offer preliminary thoughts on how to address them.

6.3.1 Equal Protection

When building predictive models of decision making, what should we do with suspect classifications? Should we—and does constitutional doctrine allow us to—include variables like race, gender, and national origin? Clearly the answer is yes where the model is only being used to estimate and describe inconsistency, but what about if it is being used to guide judicial decisions, target cases for appellate review, or otherwise influence outcomes?

What is clear is that ignoring suspect variables is not a satisfactory answer. If we're worried that algorithms may be biased toward members of protected classes, leaving suspect variables out of a model is not the way to address that bias. While our commitment to not discriminating on the basis of factors like race might be expressed by exclusion of those variables, it is poorly expressed: other variables will likely serve as proxies, soaking up the explanatory power that the suspect variables would otherwise carry. While the model would be different in content, it would be indistinguishable in consequences. This is a simple truth that the educated public will come to understand (if it doesn't already). If we all know that what we express by excluding suspect variables is void of consequence, it is difficult to see what expressive value it can have.

One alternative is to pursue the initial instinct more vigorously: if other variables are undoing the exclusion of suspect variables, then let's exclude them too. For example, if we're worried that zip code is serving as a proxy for race, exclude zip code from the algorithm too. Is crime type correlated with race or gender? Then drop crime type as well. The problem with this approach is that suspect variables can be related to other variables in a host of complex, counterintuitive, and hard-to-detect ways. To feel comfortable that we've purged a predictive model of troublesome variables, we would effectively have to abandon models altogether. There are more moderate and effective methods for addressing embedded biases.

The ideal option is to directly and explicitly address the possible underlying biases that render the suspect variables suspect in the first place. The technical complications of correcting for that bias are beyond the scope of this article, but the general ideas are simple enough to understand. We might, for example, try to account for any racial bias in an algorithm by artificially setting the race for all individuals to the same race (e.g., all individuals could be treated as white for the purposes of generating recommendations).[12] Alternatively, we could attempt to estimate the average bias against certain groups, and add back to the predicted probabilities any penalty that the group suffers (e.g., if, controlling for other variables, black inmates are 10% less likely to be paroled, we might add that 10% to black predicted probabilities). Or perhaps we could observe disparities in the way different judges make decisions about protected classes and make the assumption that the judges who treat protected classes more leniently are the judges that are making unbiased decisions.

But although a full accounting and correcting for bias is ideal, it can also be profoundly difficult to accomplish. In estimating bias, have we actually controlled for all of the other variables that might explain the disparities? Are the judges who treat protected classes more leniently actually doing so because an absence of bias, or are they perhaps simply less responsive to some of the other variables that happen to correlate with protected class? While we may be confident in our answers to these questions in some contexts, they will be nearly impossible to answer with any confidence in other contexts.

A much more technically feasible alternative is to guard against employing a predictive model of decision making that is worse than the decisions on which it is based. If we refrain from boosting predicted probabilities, the task is already complete. Because unboosted predicted probabilities only seek to spread idiosyncratic decision making more evenly through a system, it presents little danger of intensifying biases. But the story is different if statistical precedent is deployed boosting. The intuition is straightforward: if the majority is biased against a protected class, and the predicted probability for each case represents the results from a full vote, then the majority's bias is fully present in each case. Fortunately, we can

[12] While this approach is intuitively appealing, its shortcomings can be stark. Omitted variable bias. Even if there are no omitted variables, it's possible that although there is a moderate racial bias in decisions, a machine learning algorithm ends up not actually using race. The approach can also generate highly variable corrections that radically undermine the performance of the algorithm.

guard against such perverse results. Without having to estimate bias, we can simply add votes (i.e., add points to the predicted probabilities) until members of the protected class would prevail as frequently in the simulated world as they did in the actual world. Is "not making things worse" good enough? I think so. If we can prevent predictive models from increasing constitutionally troubling forms of bias, I find it difficult to see why we would forgo their benefits.[13]

6.3.2 Due Process

Statistical precedent, insofar as it is used to guide or otherwise regulate decisions, implicate issues of due process. Litigants obviously have due process interest in the ability to investigate and assure accuracy of the methods being used to decide their fates. The difficulty is in balancing that interest with competing interests. Evidence based decision making in criminal justice is already facing due process objections from scholars and litigants. In that context, the identified competing interest has, at least so far, been the proprietary nature of the algorithms. A private company, Northpointe, has developed one of the most widely used criminal risk assessment tools, the Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS"). The COMPAS algorithm was at the center of the Wisconsin Supreme Court's decision in *State v. Loomis.* Loomis, an individual who had been sentenced by a judge who had considered the COMPAS recommendation, challenged the constitutionality of using COMPAS to inform the sentencing decision, in part on the basis that "it violates a defendant's right to be sentenced based upon accurate information, in part because the proprietary nature of COMPAS prevents him from assessing its accuracy."[14] While the Wisconsin Supreme Court did not find a due process violation, it is almost certainly not the last word on the issue.

Regardless of how courts ultimately settle on the competition between due process interests and proprietary interests, there is another competing interest that may eventually prove to be even more important: the interest in maintaining an effective predictive model. The tension between transparency and an effective predictive model as an instance of Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[15]

Consider a systems model that is used generate recommendations for social security disability decisions. It is designed to guide administrative law judges' application of rules and standards—to help them distinguish the applicants that, according to the law, qualify for benefits from those that do not. But because the predictive model only predicts what judges decide—and not why they decide it—it is not well-suited to actually serve as the rules and standards. If the details of the predictive model were transparently accessible to the public, it might "distort and corrupt" the application process. For example, perhaps

[13] It is now popular to worry about the bias of algorithms, but we also need to worry about bias against algorithms (Dietvorst, Simmons, and Massey 2015).

[14] *State v. Loomis*, 881 N.W.2d 749, 770–71 (Wis. 2016).

[15] https://en.wikipedia.org/wiki/Campbell%27s_law

attorney representation is highly predictive of a successful claim but that it is not a causal relationship (Perhaps those who bother to hire an attorney are just those who are confident in their claims). Claimants who would not have otherwise hired an attorney, now aware that an algorithm gives claimants extra points for having an attorney, decide to hire an attorney in order to misrepresent themselves as the type of litigant that would hire an attorney, both wasting societal resources and undermining the ability of the model to distinguish those that qualify from those that don't.[16]

The tension between transparency and effectiveness is partially addressed by keeping statistical precedent flexible, allowing judges to distinguish the true high scores from the fake high scores, and therefore keeping the incentives for misrepresentation low. But insofar as misrepresentation is inexpensive and model recommendations exert influence on decision making, the tension remains. We might also imagine a central agency tasked with inspecting and assessing algorithms to assure their accuracy. Ultimately, there is significant uncertainty about how we will deal with the tension between the due process interest in transparency and the interests in model effectiveness and proprietary rights.

7 Conclusion

The relationship between the internal, normative perspective on law and the external, descriptive perspective on law is a dynamic one. While it's the results of the internal conversation that generate the data for the external perspective, we can and do refer back to that external perspective in our internal conversations. Judges point to what other judges do to justify themselves, even if they are not bound by precedent. Citizens point to court decisions as evidence of what the law required. Knowledge of what others think is, even if we don't know or understand the reasons for their thinking it, is relevant to our own thinking. Out of humility and support for democratic values, we give deference. But how much deference? That depends in no small part on the reliability of decision-making. Judging by coin flips is rarely deserving of our deference.

In this paper, I've argued that machine learning can advance both our understanding of reliability and our tools for promoting it. Comparing predictive models of judges to estimate inconsistency allows us to retain the external validity of disparity studies while improving internal validity. And the predictions from a model of a system, one that summarizes the votes of all judges, can be used to identify consensus views. The combination of the two—separate predictive models of judges and predictive models of the entire system—is particularly promising, allowing an aggressive search for idiosyncratic results. With this

[16] Northpointe, even though the details of its COMPAS algorithm remain a secret, already has to deal with the issue. It employs methods for detecting individuals who attempt to lower their risk score by answer questions dishonestly (Freeman 2016). Publicly revealing the inner workings of the predictive model could only exacerbate the problem. Even if the prospect of wide-scale study of the COMPAS algorithm by people charged with crime seems preposterous, more sophisticated businesses could conceivably publish study guides. In any case, the problem of transparency is clear in other contexts where litigants are more sophisticated.

capacity to better identify outliers comes the ability to better regulate decisions in an effort to make them more deserving of our deference.

Chapter 2: Detecting Inconsistency[17]

1 Introduction

Discretion to a potentially large set of semi-autonomous administrators is a fact of life for any government. Political principals at the top levels of government may articulate policies and priorities, but it is the administrators who make those policies and priorities into reality (Lipsky 1980; Daniel E. Ho 2017; Ting 2017). We refer to systems in which administrators operate, such as ICE, as *decentralized adjudication systems*. Decentralized adjudication systems are distinguished by their routine adjudicative, administrative or enforcement functions. This is in contrast to other parts of the government, where officials focus primarily on establishing rules to be applied across a wide variety of circumstances.[18] These two kinds of governance often interact. Political principals may adjust policy to constrain the behavior of those administrators, but even then, there are limits to their ability to do so. Sometimes, these administrators have different preferences than their political superiors and seek to selectively enforce policy, generating policy drift. But policy drift is not the sole problem confronting political principals; they are also concerned with the *quality* of administration.

One important element of quality is the degree to which administrators are making decisions consistently. In most areas of policy, political principals themselves view enforcement inconsistency as undermining their core policy goals. Thus, measurements of inconsistency are important for political principals who seek to advance their policy goals, as well as the public more broadly. For example, high-profile fatal incidents at nursing homes have raised awareness of the variation in enforcement across states, including variation in the size of fines levied for violations. As a result, the U.S. Centers for Medicare and Medicaid Services has worked to reduce inter-state variation in nursing homes inspections (Ornstein and Groeger 2012). Municipal governments often employ large cadres of inspectors and enforcement officers to monitor restaurants, to issue and verify building permits, to ensure sidewalks are clear of snow, to fine parking violators, etc. The degree of consistency is often a concern. Indeed, a 2011 grand jury oversight report on the Building Services Division in Oakland, California found (emphasis added):

> [C]ode enforcement inspectors have aggressively pursued blight and sub-standard properties throughout Oakland as determined by their *individual interpretations* of the applicable city code. This has led to an *inconsistent enforcement program* backed by inspectors' threats of filing large liens on the offending properties.
>
> Alameda County Grand Jury (2011)

In the context of immigration enforcement, consular and border officials all around the world process a large caseload of applications from non-citizens seeking to enter the United

[17] Coauthored with Ryan Hubert

[18] Many organizations perform both functions. For example, in this paper, we study the Ninth Circuit, which has both a rule-making function and an adjudicative function (Cameron and Kornhauser 2010).

States. High profile denials spark renewed attention to the seeming arbitrariness of visa or entry decisions (see for example, Turnbull 2013; Narea 2017; Rose 2017). For example, a recent Supreme Court decision permitted the Trump administration to deny entry to immigrants covered by Executive Order 13780 (the "Travel Ban") as long as those immigrants have no "bone fide relationship" with a U.S. person or entity (*Trump v. Int'l Refugee Assistance Project* 2017). Such determinations are made on a case by case basis by immigration officials. An policy maker overseeing immigration officials may observe that those officials deny entry to 80% of immigrants covered by the Travel Ban. On its face, this number provides some information about the overall degree of laxity of U.S. immigration officials. However, if a policy maker's goals are more sophisticated—for example, that the U.S. should deny entry to specific kinds of *high risk* immigrants—then this mean provides limited information. Indeed, if the policy maker *also* knew that, on average, immigration officers would make different decisions on 40% of of all applications, then the policy maker might conclude that the agents are not receiving sufficient guidance on how to identify high risk immigrants.

Accurate measures of inconsistency are rarely available to a policy maker or researcher. In fact, a version of the fundamental problem of causal inference often applies in this setting: because we may never observe different administrators working on the same case, estimating their disagreement on cases is difficult. Scholars and administrators have sought to overcome this difficulty in two main ways. First, they have surveyed judges with simulated case materials, allowing for observation of decisions on the same case (*e.g.*, Dhami 2005). While these inter-rater reliability studies are in high in internal validity, the use of simulated materials poses serious problems for external validity. A second method for estimating inconsistency is by estimating mean differences in the way different judges decide as-if randomly assigned cases (*e.g.*, Ramji-Nogales, Schoenholtz, and Phillip G. Schrag 2007; Nakosteen and Zimmer 2014b). Because these disparity studies rely on real decisions, they are high in external validity. But their reliance on simple differences in means poses serious problems for internal validity.

In this article, we show how machine learning can be used to optimize the estimation of inconsistency in decentralized adjudication systems using administrative data on actual decisions. The core dilemma is that mean differences between decision makers on outcomes—that is, disparity statistics—are not a good proxy for inconsistency. Those disparity statistics systematically *understate* the level of disagreement. As we describe in detail below, this downward bias can be mitigated through estimation of heterogeneous treatment effects (HTEs). However, estimating HTEs is not a trivial matter. Our contribution is to provide a method for optimally targeting the search for HTEs in order to minimize downward bias associated with estimating inconsistency. In particular, one only needs to subset the parameter space into a two-partition based on the "direction" of disagreement between decision makers.[19] Knowing the direction of disagreement on a decision requires knowing the counterfactual decisions made decision makers who were not assigned to a case. This latter step is feasible due to advances in machine learning, as well as increased data availability.

[19] For example, the set of cases where decision maker A is more lenient than decision maker B, and the set of cases where the opposite is true.

Using machine learning, we generate high quality predictions for these unobserved counterfactual quantities. Our method also solves a set of subsidiary technical problems that have plagued studies of inconsistency, including limitations imposed by small sample sizes and non-random assignment of cases to administrators.

We demonstrate our method by estimating inconsistency in the decision-making among the judicial panels of the Ninth Circuit Court of Appeals. To do so, we use an original dataset of all civil appeals heard between 1995 and 2013. Scholars, judges, and politicians have contended that the Ninth Circuit is chaotic, too inconsistent, and the home of "jackpot justice." These accusations have engendered routine calls and bills to split the Ninth Circuit. We find that at least 9% of appeals would be decided differently had they been (randomly) reassigned and that the two most dissimilar kinds of panels decide at least 40% of appeals differently. Because we lack data on other circuit courts that could allow us to make comparative assessments of the Ninth Circuit's level of inconsistency relative to those courts, we instead focus on the impact of inconsistency on the court's internal procedures. In particular, we examine whether unpublished opinions simply apply settled law, as required by the court's rules.[20] The evidence strongly suggests that they do not. As we detect more inconsistency over case outcomes, we also observe higher rates of non-publication. The pattern holds even after controlling for case area.

2 Quality and Inconsistency

At least since the dawn of the field of public administration, scholars have been concerned with the quality of governance (*e.g.*, Wilson 1887; Taylor 1911). While the literature on the topic is vast, there are two main ways to assess the quality of government: policy and implementation. The policy approach assesses policy outcomes against a predefined normative benchmark. For example, La Porta et al. (1999) consider good government to be that which promotes economic development. The authors take a set of variables, such as protection of property rights, tax rates and infrastructure quality, as evidence of good or bad government. However, one can also take policy goals as given and investigate how well they are implemented. Indeed, the vast majority of government is devoted to implementing policy. This takes a particular form: "administrators" make determinations on a set of "cases." Is company X liable for damages in a tort suit? Is person Y entitled to a social welfare benefit? Does restaurant Z comply with local health ordinances?

To fix ideas, consider a variant of the policy space adopted for use in models of judicial politics, the case space (Lax 2011). To emphasize that our argument generalizes beyond the context of the judicial system, we refer to this as *implementation space*. We assume the implementation space is unidimensional (specifically, a convex set $X \subseteq \mathbb{R}$), and thus each point in the space represents a possible constellation of facts—a case.[21] A decision maker

[20] In the U.S. federal courts, judges decide whether to officially publish their decisions. While the distinction between published and unpublished opinions has been blurred due to the advent of electronic databases, formally speaking, published opinions constitute legal precedent while unpublished opinions do not.

[21] The fact pattern is determinative in the sense that there are no other facts relevant for policy imple-

uses a rule, $\widehat{x}$ to decide what decision to make.[22] For example:

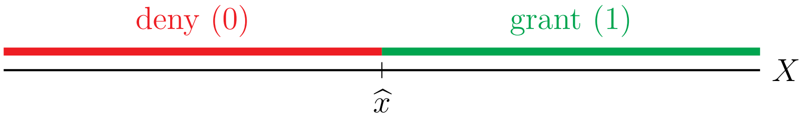


Figure 2: Implementation Space

A high level policy maker usually delegates enforcement to a large set of low level administrators in a decentralized adjudication system. For simplicity, suppose there are two administrators, 1 and 2. Those two administrators may—for whatever reason—use different criteria whenever they make their determinations. For some number of cases, the two administrators disagree, and their decision making is thus inconsistent. Typically, there have been two ways to study this phenomenon empirically. First are inter-rater reliability studies that measure whether two (or more) administrators would come to the same decision on the same case. These kinds of studies vary in their degree of external validity. Daniel E. Ho (2017) presents the results of a novel experiment with an intervention that enabled direct measurement of disagreement on identical cases. However, such experiments are expensive, and thus rare.

A second standard way to study this phenomenon is through disparity studies. Consider Figure 3. Administrator 1 has a grant rate of around 66% whereas Administrator 2 has a grant rate around 33%. When examining Figure 3, it is apparent that the two administrators disagree on $66 - 33 = 33\%$ of cases. This latter quantity is known as a "disparity." Many scholars have used disparites to study differences between decision makers across a wide array of institutions, including asylum applications (Ramji-Nogales, Schoenholtz, and Phillip G. Schrag 2007), social security disability appeals (Nakosteen and Zimmer 2014b), and judge sentencing (Anderson, Kling, and Stith 1999). The main advantage of disparity studies is that they allow us to study decentralized adjudication systems in as they exist in the real world.

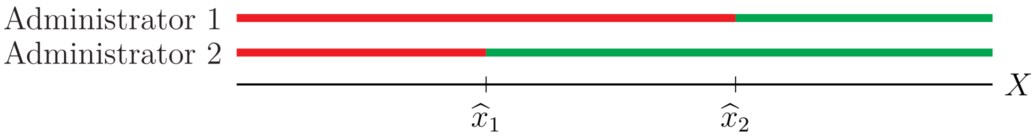


Figure 3: Decision Making with Multiple Administrators

But while disparity studies provide us with important information about variation among administrators, they understate how much disagreement there is among those administrators. This is because they rest on the unverified—and often incorrect—assumption that

mentation.

[22] Here, we assume rules to be monotonic, however, the discussion easily extends to non-monotonic rules. An example of a non-monotonic rule is a speed limit that specifies both a minimum and maximum speed.

disagreement between decision makers goes in the same direction across all cases. Roughly speaking, this means that two decision makers make their decisions in a "similar" manner, even if they sometimes disagree on marginal cases. To make this more concrete, suppose that the two administrators do not make their decisions in a similar manner. As depicted in Figure 4, Administrator 1 grants for *high* values in X while Administrator 2 grants for *low* values in X.

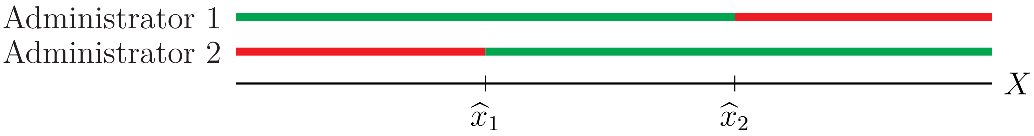


Figure 4: Decision Making with Multiple Administrators

The two administrators' grant rates remain unchanged, and so does the disparity statistic. But as Figure 4 illustrates, they disagree on many more than 33% of cases. Building on the work of Fischman (2014a), we show below that when scholars use disparity as a proxy for disagreement, they are always understating the degree to which the decision makers disagree. Our method minimizes that understatement.

3 Measuring Inconsistency

A decentralized adjudication system 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. Second, there is a finite set of cases about which decisions are made, which we label $\mathcal{N} = \{1, ..., N\}$, and 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.[23] For example, Y could be dichotomous, such as admit/deny or reverse/affirm, or continuous, such as the length of a criminal sentence or a fine.[24] As long as a single determination is made on each case, we allow decision makers to be groups, setting aside the micro-foundations of group decision making.[25] 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 border agents would specify how each agent would decide whether to admit each immigrant they process. Since we only observe actual decisions, we treat $\mathcal{C}$ as a black box and focus on measuring observable patterns in decision making.

[23] 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.

[24] We require that the elements of Y be ordinal. Our method does not allow for non-binary categorical outcomes.

[25] 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).

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 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.[26] 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:

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

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 (2014a), 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 randomly reassigned? We characterize average inconsistency differently than Fischman (2014a). Briefly, we allow that cases could be reassigned to *the same* decision maker, whereas Fischman (2014a) assumes that cases are reassigned to different decision makers.

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 Disparities

We have already noted the limitations of disparity studies, but here we show formally how they understate disagreement and inconsistency. As is apparent from equation (1) we express

[26] 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, $d(Y_i(j), Y_i(k)) = |Y_i(j) - Y_i(k)| \leq 10 = D$.

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 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 as-if random 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.

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).[27] 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 state court appellate judges. Suppose one judge is a Democratic and the other is a Republican. Moreover, suppose an analyst is interested in studying how this Democratic judge (D) and this Republican judge (R) decide civil rights cases differently. If she were to try to measure the extent of disagreement between these two judges, 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:

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

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

$$\phi(D,R) = \mathrm{E}[Y(D)] - \mathrm{E}[Y(R)]$$

Unfortunately, $\phi(D,R)$ would be a downward biased estimand for disagreement. In Propo-

[27] 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 Phillip G. Schrag (2007), Boyd, Epstein, and Martin (2010), and Kastellec (2013).

sition 1 in Appendix E, we show generally that[28]

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

Moreover, equation (5) holds strictly whenever there are "strong heterogeneous treatment effects" as defined by Definition 1 in Appendix E. 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 plaintiff is victorious in a civil rights case, then an ATE comparing D and R would understate the level of disagreement between them if D is more likely to reverse than R when the defendant won in the lower court but less likely to reverse than R when the plaintiff won in the lower court.

This problem is not solved by re-coding the outcome variable. For example, an analyst might re-code 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 judges, it still does not capture disagreement. Suppose, for example, that the two judges differ in their propensity to reverse lower court decisions: D always reverses the lower court decision, and R never does. 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 judges 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 appellate 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 dimensions characterizing the most disagreement.

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

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 the set of cases $\mathcal{N}$ that represents a partition of the case-level covariate space and where each $M \in \mathcal{M}$ is nonempty. Then, the estimand of interest is the conditional average treatment effect (CATE):[29]

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

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:

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

This approach helps researchers avoid making problematic assumptions on the joint distribution 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 $\mathcal{M}$. Proposition 2 in Appendix E shows that all possible partitions $\mathcal{M}$ generate estimands of disagreement that are smaller than the actual level of disagreement. Thus, our task is to pick $\mathcal{M}$ to maximize $\delta(j, k, \mathcal{M})$. Moreover, 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

[29] 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.

subsets:

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

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:

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

Estimation

We have shown that ATE-based estimands of disagreement, such as disparities, will always be biased downward. As a result, an ATE-based estimand constitutes a *lower bound* on the true level of disagreement among decision makers. It is important to emphasize once again that attempts to measure disagreement or inconsistency are always either lower or upper bounds on the true measure, including the method we introduce in this paper (see Fischman 2014a). However, while our method also yields a lower bound, our contribution is to substantially reduce the bias of more common disparity-type measures of inconsistency. 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 $Y(j) \geq Y(k)$ 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 recommend using 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 $\widehat{Y}_i(k)$. In our illustrative example, described in the next section, we use `Super Learner`, an ensemble method that uses a set of constituent algorithms to predict outcomes in the data (Laan, Polley, and Hubbard 2007a).

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 $\delta(B, C)$.

This poses a more serious problem where there are a small number of cases assigned to some of the treatments, as predictions would be extremely noisy and uninformative. Consider, for example, the Ninth Circuit. 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. It will therefore often be necessary to cluster judges into larger groupings to increase sample sizes used to build prediction models. Essentially, we must sometimes re-define the "treatment" and "control" to be panels of different *types* of judges, rather than specific judges. To be clear, clustering explicitly opts for increased bias in order to decrease variance, and the extent to which an analyst trades off bias for variance is context-dependent and discretionary. But we strongly recommend clustering decision makers by similarities in their decision patterns rather than shared demographic or political characteristics. Indeed, as we illustrate below, researchers can use a training set to build decision-predictive models for each decision maker, and decision makers can then be clustered by similarities in the outputs of those models.

Finally, our approach solves a problem identified by Fischman (2014a): finite sample bias artificially inflates estimates of inconsistency. Traditional estimates of disagreement overstate inconsistency because they treat all observed differences among decision-makers as reflecting true differences, ignoring the fact that differences are actually a combination of true differences and statistical noise. Moreover, the problem can become more severe as researchers increase variance by subsetting in the search of more inconsistency. Fischman (2014a) 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 when those expectations are not met, eliminating variance's contribution to bias.

4 Application: Ninth Circuit

In the previous section, we described a general approach for estimating inconsistency in an adjudication system. In this section, we demonstrate how to use machine learning to implement our method using a large and extensively coded original dataset of all civil cases filed in the Ninth Circuit and terminated on the merits over a period of nineteen years.[30]

[30] In Appendix B, we describe our data and discuss how it constitutes an improvement on other available datasets.

Our procedure generates estimates of extreme and average inconsistency that uncover a greater degree of disagreement among judges than traditional approaches can. In particular, we find levels of extreme and average inconsistency in the Ninth Circuit of 40% and 9%, respectively. That is, the two most dissimilar types of panels (which are endogenously determined by our machine-learning method) would decide 40% of cases differently, while two randomly selected panels would decide 9% of cases differently on average. As a benchmark, we can compare these estimates to two other disparity measures that cluster judges by party of appointing President, based on theoretical intuitions that partisanship is the dimension that captures the most disagreement among judges. Our method substantially outperforms both a naïve comparison of reversal rates and a comparison of pro-plaintiff decision rates. If we use reversal rates, we obtain estimates that uncover substantially less inconsistency: 12% and 4% for extreme and average inconsistency, respectively. If we use pro-plaintiff decision rates, we obtain estimates that uncover even less inconsistency: 2% and 1% for extreme and average inconsistency, respectively.

We now describe the procedure we used to obtain our machine-learning estimates of inconsistency. The procedure follows six basic steps.[31]

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 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

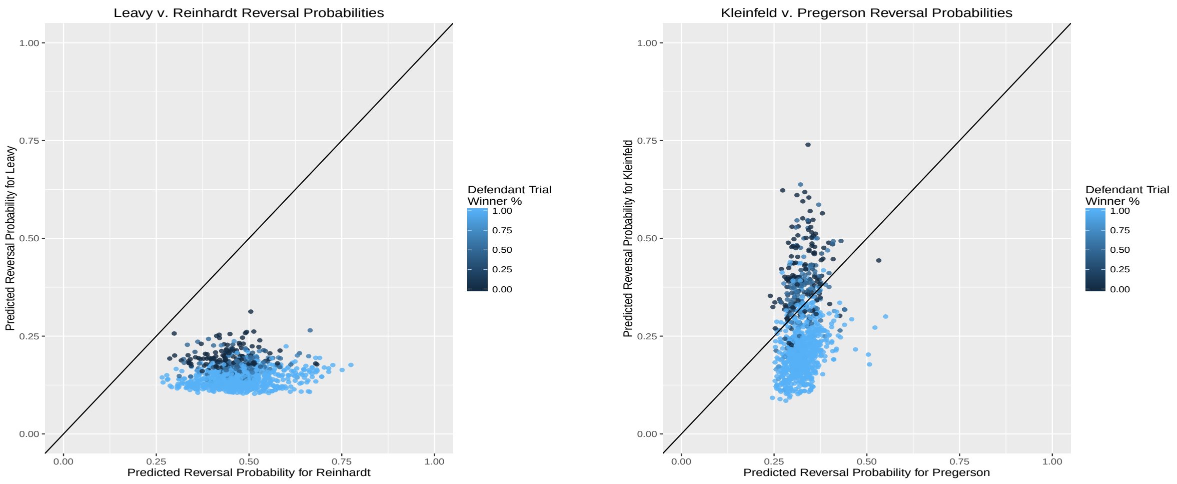
[31] We note that the six steps could be extended to include a seventh step for measuring uncertainty around the point estimates for inconsistency. Because we do not improve upon the sub-sampling method as described in Fischman (2014a), we restrict our focus here to point estimates. In principle, an analyst could use the sub-sampling technique in conjunction with our method to construct confidence intervals.

over all their cases leveraging information about the panels they sat on. We include six 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. The algorithm then constructs a weighted model of these constituent algorithms that generates the best predictions, as measured by mean squared error. Details are available in Appendix A.

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 5 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.

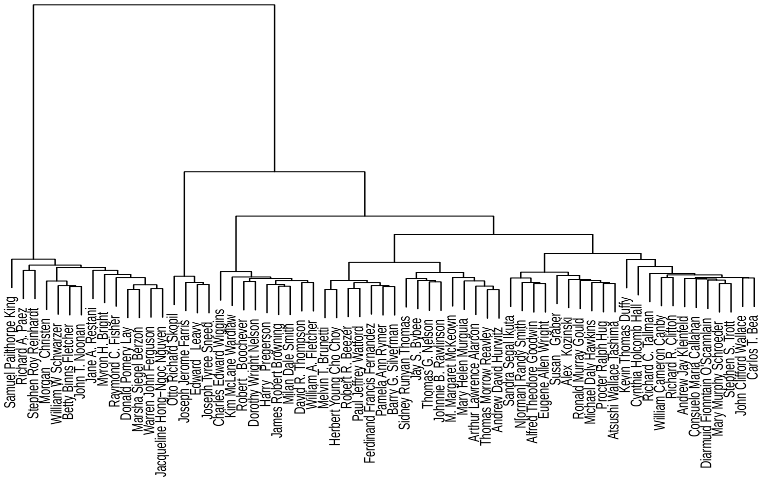
Figure 5: Comparing Judge Predictions



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.[32] Figure 6 shows the results of the cluster analysis.

Figure 6: Cluster Dendrogram



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.

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 demographic characteristics of the groups are available in Tables 3 to 8 in Appendix C. 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 an `O`-judge, and a panel of such judges as an `OOO`-panel.

[32] 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.

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.[33] As we show in Table 1, our `Super Learner` model outperforms each constituent algorithm.

Table 1: 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

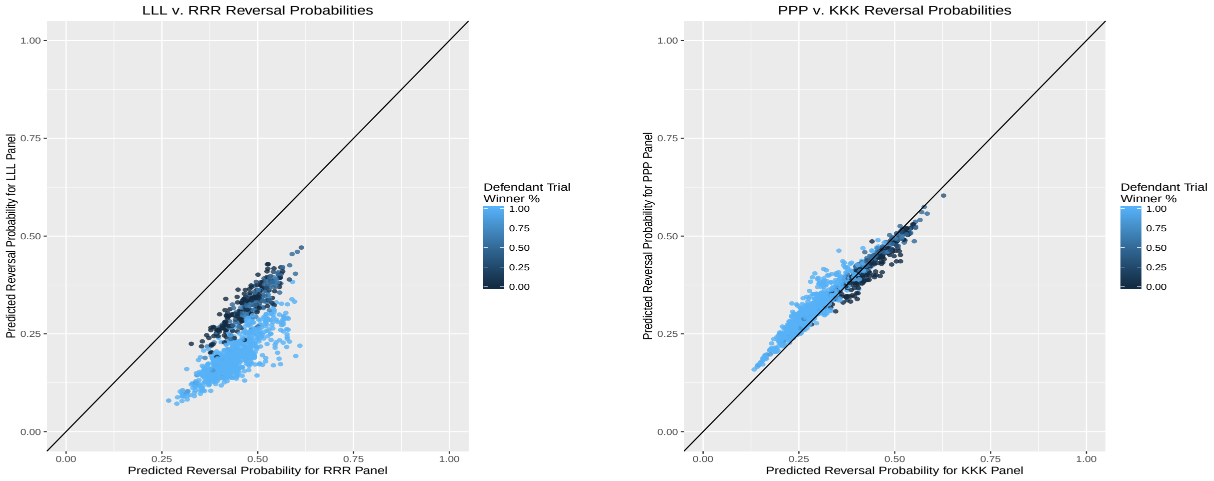
Figure 7 visualizes predictions for four of the panel types and highlights the potential for machine-learning approaches. The left panel suggests that the main source of disagreement between `LLL`-panels and `RRR`-panels is something related to judges' reversal proclivity. Even though `LLL`-panels appear 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 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 re-code decisions in the test set so that we can more accurately estimate the actual level of inconsistency in adjudication.

[33] 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.

Figure 7: Comparing Panel Predictions



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

As we discussed in Section 3, 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. 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:[34]

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

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.[35]

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 *unconfounded* 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. However, as we discuss in Appendix D, there are at least two possible threats to the randomization assumption in the context of the Ninth Circuit. In fact, proper randomization is rarely guaranteed in decision making systems, so this step provides a technique for correcting for potential selection bias.

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 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,

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

[35] 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.

which aims to make the confounding variable orthogonal to the outcome (Hansen 2008).[36] 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.[37] Our modified estimator, now resistant to 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

Now we estimate average inconsistency and extreme inconsistency. Our estimator for extreme inconsistency is straight forward:

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

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

[37] 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 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.

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 a 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 estimator for average inconsistency is therefore:

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

With these six steps, a researcher can generate estimates of inconsistency in adjudication systems, and, as we've detailed above, can offer substantial improvements to existing methods.

5 Discussion: Evaluating the Ninth Circuit

According to its critics, the Ninth Circuit is "in chaos," a system of "jackpot justice," and "nutty." One way of assessing these claims would be to compare our estimates of inconsistency in the Ninth Circuit with estimates from other circuits. Unfortunately, we lack data from other circuits needed to make these comparisons. Instead, we look to whether patterns in decision making comply with the Ninth Circuit's internal operating procedure: is the practice of designating opinions "not for publication," which renders them non binding on future cases, consistent with the court's appellate procedure and the policies that are used to defend that procedure?

According to Ninth Circuit policy, an opinion is to be published if it "[e]stablishes, alters, modifies or clarifies a rule of federal law" (Circuit Rule 36-2(a)). Consistent with this idea, the conventional wisdom among court observers is that cases with unpublished opinions are "easy" cases that are less controversial and more straightforward to resolve. Epstein, Landes, and Posner (2013), for example, argue that "most [unpublished opinions] are affirmances in uncontroversial cases" (p. 155) and that "... cases that arouse no disagreement among the judges tend not to be the cases that shape the law" (p. 55). Sunstein, Schkade, and Ellman (2004) have similarly claimed that "unpublished opinions are widely agreed to be simple and straightforward, and to involve no difficult or complex issues of law" (p. 313). And Judge Harry T. Edwards of the District of Columbia Circuit argued in a co-authored article with Michael Livermore that "judgments rendered in unpublished decisions ... typically involve more straightforward applications of law" (Edwards and Livermore 2009, p. 1923). Finally, in a 2002 hearing on unpublished opinions held in the U.S. House Subcommittee on Courts, the Internet, and Intellectual Property, Ninth Circuit Judge Alex Kozinski said

> But I can state with some confidence that the sinister suggestion that our unpublished dispositions conceal a multitude of injustices and inconsistencies is simply not borne out by the evidence. I feel so confident of this point, having participated in rendering thousands of these dispositions myself, that I would welcome an audit or evaluation by an independent source.
>
> *Unpublished Judicial Opinions* (2002)

Here, we lend our method for measuring inconsistency to the independent auditing effort. A logical implication—and a testable hypothesis—derived from this conventional wisdom is that cases featuring a low degree of inconsistency among judges (*i.e.*, a high degree of agreement) should be the ones that are more likely to be resolved with an unpublished opinion, and vice versa. In other words, there should be a (weakly) positive relationship between disagreement and opinion publication.

Testing this hypothesis empirically presents a practical problem: inconsistency is measured at the system-level, whereas opinion publication is measured at the case-level. In order to test the relationship between system-level inconsistency and case-level opinion publication, we therefore need to subset our sample into bins of cases that feature similar levels of disagreement. We do this by generating case-specific "disagreement scores" for each case.[38] By ordering cases by this disagreement score, we can plausibly partition the full data set of cases into smaller sets that will have increasing levels of inconsistency.

We separate cases into four quartiles by their case-level disagreement scores and measure systemic inconsistency among the cases in each of the four quartiles. This allows us to directly examine the relationship between our systemic measure of inconsistency and publication rates. Figure 8 plots those inconsistency estimates against publication rate. Disagreement between judges and opinion publication is *negatively* correlated.

[38] As a by-product of Step 3, there is a predicted probability of reversal for each case for each panel type. Disagreement scores are a measure of spread of each case's set of predicted probabilities. Here, we use mean absolute deviation (from the mean), but results using variance as the measure of spread are similar.

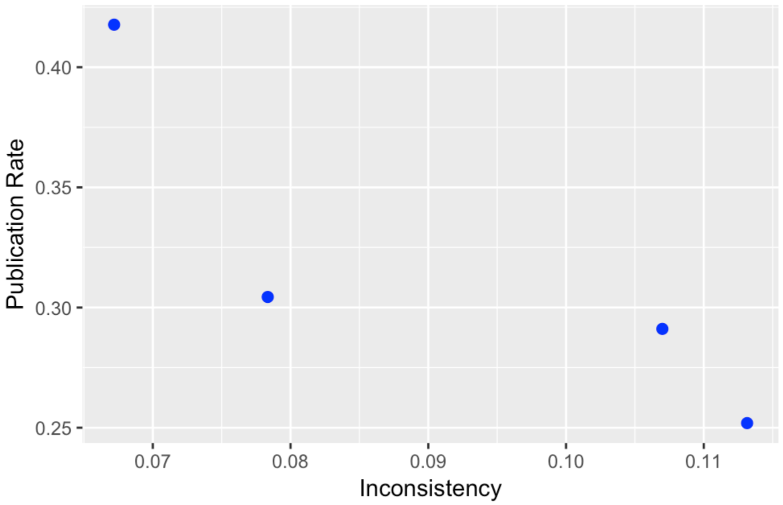


Figure 8: Inconsistency and Publication

We also regress publication directly on disagreement scores. There is again a strong negative relationship between these case-level disagreement scores and opinion publication: for a 1% increase in the disagreement score, there is 2.6% reduction in the probability that a case will be resolved with a published opinion (significant at the 0.001 level). This association holds even when we control for case type: 1% increase in the disagreement score is associated with 2.1% reduction in the probability that a case will be resolved with a published opinion.

In summary, our analysis strongly suggests that, contrary to conventional wisdom, unpublished opinions are not being used to dispose of the "easy" cases that judges agree on. In fact, it is where law does not determine the outcome that judges are most likely to issue unpublished decisions.

An astute observer of appellate courts might argue that although decision making may be inconsistent with standard justifications for the use of unpublished opinions, it may nonetheless be consistent with stated policy: while unpublished opinions do not resolve "easy" cases, they might be used to dispose of cases that are *legally* easy but *factually* difficult. In other words, judges may agree on what the law is and simply disagree over how to apply the law to complex fact patterns. Thus, judges could still be using published opinions to resolve a case that "[e]stablishes, alters, modifies or clarifies a rule of federal law" (Circuit Rule 36-2(a)) while using unpublished opinions to apply clear and settled law to difficult fact patterns.

We find little support for the argument that judges agree on law and disagree on application. To assess the argument, we run our method for estimating inconsistency on the publication decision instead of the substantive outcome. If judges are merely disagree on the application of clear and settled law, we should expect to find little disagreement over the decision to publish. This is not what we find. Figure 9 shows the distribution of estimated disagreement between all panel types over the decision to publish. Although the small part of the distribution below zero indicates that there would be some regression to the mean, a substantial number of panels disagree in a non-trivial percentage of cases. In short, judges do not only disagree on the application of law—they disagree as to whether law is clear and settled.

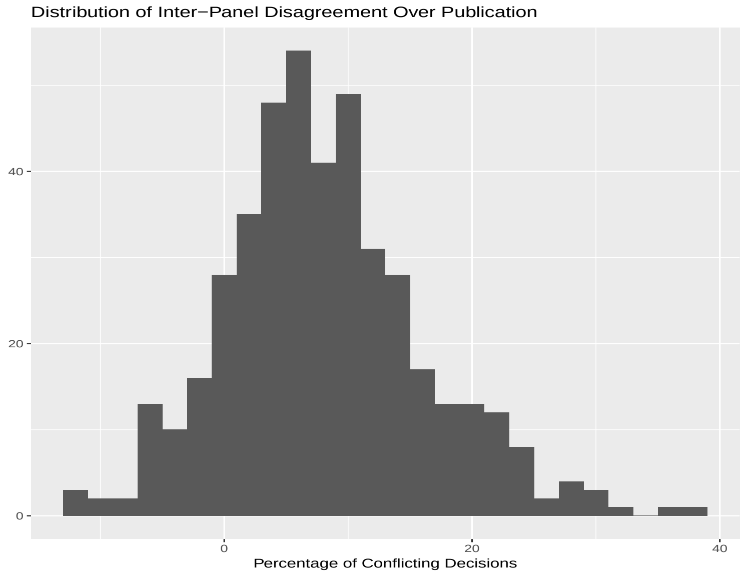


Figure 9: Disagreement about Whether to Publish

6 Conclusion

We have presented a method for estimating inconsistency in decentralized adjudication systems. Although it is still only a lower bound on inconsistency, it represents a vast improvement on the disparity studies that have so far been used. Particularly when coupled with inter-rater reliability studies that use simulated case materials, our observation-based method provides a way forward for the rigorous study of inconsistency. We applied our method to the decision making in the Ninth Circuit, showing how we could start building measures of performance. But lacking data on other circuits that could allow for meaningful comparisons, we instead focused on assessing Ninth Circuit compliance with internal circuit policy over the publication of judicial opinions. Evidence suggests that judges are not, as policy dictates and many judges and scholars have claimed, using unpublished opinions in cases that apply settled law. In fact, we find evidence that judges are more likely to use unpublished cases where inter-judge disagreement over the proper outcome is highest.

Chapter 3: Synthetic Crowdsourcing[39]

1 Introduction

Justice, the trope goes, is what the judge ate for breakfast. The problem of inconsistency in legal and administrative decisions is widespread and well documented. Recent research has demonstrated that case outcomes can vary significantly depending on the characteristics of the deciding judges: researchers have documented stark disparities between judges in domains including social security disability (Nakosteen and Zimmer 2014c), criminal sentencing (Abrams, Bertrand, and Mullainathan 2012) and asylum (Ramji-Nogales, Schoenholtz, and Philip G. Schrag 2007; Fischman 2014b). In addition to these disparities across judges, studies show that individual judges are themselves inconsistent from case to case. Circumstances such as the outcome of a football game (D. L. Chen and H. Spamann 2014) and the time of day (Danziger, Levav, and Avnaim-Pesso 2011) can substantially affect legal decisions.

The creation of general rules, whether imposed through centralized legislation or decentralized precedent, is a core strategy to protect against inconsistency and arbitrariness in decision making, but rules are often poorly suited for the fact-intensive contexts that make up a large portion of modern adjudication (Sunstein 1995). In contexts where small and varied deviations in fact patterns can substantially impact merits, rules will be crude and insensitive to the particulars of the case. For example, adjudication of claims for social security disability benefits, asylum, and parole frequently turn on issues that are hard to usefully delineate with ex ante rules.

Scholars and administrators have proposed a variety of methods to reduce inconsistency in fact-intensive decision-making settings. Proposed approaches include: increasing the number of decision makers responsible for making each panel (Legomsky 2007), matrix-based decision-making guides (e.g., Federal Sentencing Guidelines), peer review (Daniel E Ho 2017), statistical models that estimate outcomes that are highly relevant to the decision process (e.g., models of recidivism), more and better training for decision makers, and top-down review of deviant decision makers.[40]

In this article, we present a novel statistical tool that combines many of the benefits of existing approaches while avoiding many of their costs. ”Synthetic crowdsourcing” uses machine learning to predict future decisions using data on past decisions. The aim is to simulate a world in which all judges cast multiple independent votes in every case. By excluding variables that are statistically uncorrelated with the merits of a case (e.g., the identity of randomly assigned judge or whether a judge’s football team won the night before) and aggregating judgment across and within decision makers, such predictive models can cancel out arbitrary and contingent factors, smooth over sources of inconsistency, and capture

[39] Coauthored with Hannah Laqueur

[40] The Administrative Conference of the United States, for example, has newly recommended that the Social Security Administration review ”decisions from judges whose allowance rates are both well under and well above statistical average.” https://www.acus.gov/recommendation/improving-consistency-social-security-disability-adjudications.

the wisdom of the crowd. The results of the voting simulation can then be used to guide decision-making or to help administrators identify deviant decisions for further review.

The article proceeds as follows. Section 2 reviews the evidence of inconsistency in adjudication as well as existing approaches to reducing it. Section 3 offers a conceptual account of synthetic crowdsourcing and its potential uses. Section 4 provides proof of concept with an analysis of California Parole Board decisions. Section 5 addresses limitations. Section 6 concludes.

2 The Problem of Inconsistency

2.1 Evidence of Inconsistency

Researchers have documented stark disparities in the rates at which adjudicators grant asylum to refugees (Ramji-Nogales, Andrew I Schoenholtz, and Philip G Schrag 2007), provide social security disability benefits (Nakosteen and Zimmer 2014a), decide whether to remove children from parental custody (Nickerson 2007), and determine prison sentence lengths (Brantingham 1985). In U.S. asylum cases, for example, research suggests at least 27% of cases would be decided differently if they were randomly assigned to a different judge (Fischman 2014b). At the appellate level, it is estimated that roughly half of asylum appeals could have be decided differently had they been assigned to a different panel (Fischman 2014b).[41] Daniel Ho (2017) showed that Washington state food safety inspection unit disagreed in 60% of cases when assigned to evaluate the exact same establishments. Judges also show low levels of inter-rater reliability in hypothetical sentencing decisions, with standard deviations of 30% to 60% of the mean sentence length (Grunwald 2015).

Several recent studies have started to empirically explore inconsistencies *within* individual decision-makers. Research suggests, for example, that an inmate's chances of parole declines precipitously the longer a judge works without a break (Danziger, Levav, and Avnaim-Pesso 2011). Asylum officers are up to 3.3% more likely to reject asylum if they granted asylum in their previous case (Chen, Moskowitz, and Shue 2016) and 1.5% more likely to grant asylum after their city's NFL team won (Daniel Li Chen and Holger Spamann 2016).

These types of inter and intra-judge inconsistencies negatively affect the accuracy, predictability, fairness, and legitimacy of an adjudication system. The literature on the costs of inconsistency is extensive, and we do not attempt to summarize it here. Instead, we simply make what we think is the uncontroversial assumption that the outcome of a case should not depend on factors that are unrelated to the merits of the case. For example, outcomes should not depend on which judge is assigned to decide a case, whether that judge's football team won the night before, or what the judge ate for her proverbial breakfast. The synthetic crowdsourcing approach we propose serves to reduce decision inconsistencies that result from such factors.

[41] Generally, these disparity studies leverage the fact that adjudication systems frequently make use of random or as-if random assignment of cases to judges or administrators, allowing the researchers to attribute the cause of disparities to differences in adjudicators' preferences.

2.2 Existing Solutions to Inconsistency

Scholars and administrators have proposed a variety of approaches to reducing inconsistency in fact-intensive legal and administrative decision-making settings where the marginal benefit of more rule-making is minor; all are accompanied by serious costs. One set of responses aim to improve decision making by improving the decision-makers, either with professional development training or new hiring. The notion is that better, or better trained, judges and administrators will be more consistent (Legomsky 2007). Although the approach has intuitive appeal, filling in the details is a difficult task. How do we identify and hire "better" judges? What precisely should professional development programs work to develop, and how should they develop it? Despite frequent efforts to improve hiring or offer training to adjudicators, there is little to no evidence that such efforts in fact decrease inconsistency among adjudicators (Legomsky 2007).

Another approach to reducing inconsistency is to increase the number of judges who participate in each decision. The notion is that larger decision units will decrease the variance of any given decision, both by mechanically limiting the power of extremist judges and by allowing for deliberation that can help prevent ill-considered decisions (Legomsky 2007). But decision making in large groups can result in the amplification of errors, cascade effects, and group polarization (Sunstein 2005). Furthermore, increasing the size of decision units has substantial financial or labor costs: it requires either hiring more judges or increasing each decision unit's caseload.

Quotas have been proposed as a means to regulate decisions, but they represent a relatively clumsy response to inconsistency. Mashaw (1985) suggested, although ultimately rejected, the idea of reducing disparities in social security disability decisions with a quota system. "State agencies or individual disability examiners could be given a grant rate (say 35 percent +/- 5 percent) for each time period (say a month)," he wrote, "awards would then be made on a comparative or relative basis and award rate disparities would virtually disappear." More recently, the Administrative Conference of the United States has recommended that the Social Security Administration consider reviewing "decisions from judges whose allowance rates are both well under and well above statistical average" (Krent and Morris 2013). A similar principle underlies policies to punish decision-makers who consistently deviate from the average grant rate or policies that seek to encourage consistency by distributing information about peer grant rates (Legomsky 2007). While quotas might succeed in reducing the overall disparities in decisions, they might still fail to reduce inconsistency in decisions for comparable cases. This would arise if, for example, decision-makers grant the same percentage of cases but nonetheless grant claims in very different types of cases (Fischman 2014b).

Decision matrices, such as the Federal Sentencing Guidelines, explicitly attempt to generate consistent decisions for comparable cases, but in practice, such formulas can fail to account for the wide variety of events and circumstances that actually occur and thus result in sub-optimal outcomes (Kaplow 1992). For example, the Federal Sentencing Guidelines, developed in response to evidence of vast sentencing disparities, have since been criticized on a number of grounds, with arguments between advocates and critics echoing the fundamental rules vs. standards debates. The chief criticism is that the guidelines too crudely aggregate: limiting judicial discretion comes at the price of too many inapposite sentences.

Within the criminal justice system, statistical forecasts of future dangerousness—risk assessment instruments—are increasingly being used to help judges make less biased, more efficient, and more consistent decisions in contexts including parole, bail, and sentencing. Opponents argue they are opaque, unreliable, and unconstitutional (Starr 2014). Most importantly for the purposes of the current discussion, such tools are restricted to context in which an important outcome proxy for a correct or good decision is measurable (e.g., future offending). Thus, these statistical prediction instruments are, at a minimum, limited as approach. Many legal and administrative decision-making contexts do not have an identifiable proxy for whether a decision is correct or good.

Finally, peer review represents a recent and promising proposal to limit disparities and improve decisions. Ho (2017) recently published the results of the first and only randomized control trial of peer review in a government agency, randomizing food safety inspectors into weekly peer review inspections and meetings. The study found that the intervention increased the accuracy and decreased the variability of inspections. The greatest drawback to a peer review approach is its financial cost.

3 Synthetic Crowdsourcing

Synthetic crowdsourcing uses machine learning to generate predictions of case decisions. These predictions can be used to help decision-makers make better and more consistent decisions in the future or can be employed as a tool for administrative monitoring and review. By aggregating judgments across and within decision-makers, synthetic crowdsourcing cancels out arbitrary and contingent factors, leverages the wisdom of the crowd, and minimizes inter- and intra-judge inconsistency.[42] Synthetic crowdsourcing extends the core benefits of en banc decision making to the full population of cases while avoiding the dangers of group think. Similar to traditional matrix-based decision-making tools such as the Federal Sentencing Guidelines, synthetic crowdsourcing uses statistical patterns in historical decisions to guide future decisions. But unlike traditional approaches, it leverages machine learning to optimally tailor that guidance, allowing for substantial improvements in the consistency and overall quality of decision making.

Synthetic crowdsourcing can be understood as an effort to simulate a world in which each judge casts multiple independent votes in every case. In what follows, we explain how to best pursue that simulation effort and suggest how the results from the simulation can actually be implemented to guide and improve decision-making systems. In Section 4, we provide proof of concept using a predictive model of the California Board of Parole Hearings parole suitability decisions.

[42] We present the case for the guiding dichotomous outcomes, but most of the framework could be straightforwadly extended to continuous outcomes.

3.1 The Simulation Goal

In an effort to reduce inconsistency, our goal is to remove the influence of factors that are randomly or as-if randomly assigned by adjudication systems. A hypothetical but normatively appealing solution for mitigating the influence of (as-if) random factors is a world where, in each case, each judge independently casts multiple independent votes under a variety of conditions (e.g. after her football team won, after her football team lost, in the morning, in the afternoon etc.). Case outcomes would then be determined by the voting results. Such a world is normatively appealing for two reasons. First, it allows us to remain agnostic with respect to the value of different judges' decisions. We avoid potentially contentious debates and instead rely on the appeal of democratic principles. Second, Condorcet's Jury Theorem, the classic theorem of political science and antecedent to the "wisdom of the crowds," provides normative grounding for this approach: as long as decision-makers are, on average, making good decisions, then a world with more independent votes will generate better decisions (Austen-Smith and Banks 1996).

Of course, in reality, such a hypothetical world is unobtainable. Requiring all judges to participate in every case multiple times is prohibitively expensive and—if independence of votes is desired—pragmatically impossible. But we can statistically *simulate* such a world.

Before turning to the simulation technique we make a final note regarding the limitation of this simulation goal. It targets the elimination of factors that are, by practice of an adjudication system, assigned on a random or as-if random basis. But the simulation target does not directly address problems of systematic biases. To be sure, decisions should also not depend on race, class, gender or other morally arbitrary factors. The reason for our focus on randomly assigned factors is ultimately technical: their (as-if) random assignment renders them far easier to target. As we describe in Section 5, synthetic crowdsourcing can be potentially be supplemented to address problems of systemic bias. Nonetheless, its primary purpose is to eliminate the influence of random factors that generate inconsistencies within and across judges.[43]

3.2 The Simulation Technique

We recommend simulating the targeted world with a predictive model of decision making that excludes variables that are randomly or as-if randomly distributed among cases, includes all available variables that may be non-trivially related (statistically speaking) to the merits of a case, and is built with machine learning methods. Our recommendation is in conflict with the existing but limited literature on the use of decision predictive models to guide decision making. "Judgmental bootstrapping," as its proponents refer to it, "involves developing a model of an expert by regressing his forecasts against the information that he used" in order to "infer the rules that the expert is using" (Armstrong 2006). This approach is often effective. For example, bootstrapping models have been shown to be better at predicting loan defaults (Abdel-Khalik and El-Sheshai 1980) and forecasting the number of advertising

[43] Note also that the inconsistencies may be disparately distributed, so even without supplementation to directly address bias, synthetic crodwsourcing can ameliorate differential treatment.

pages a magazine will sell (Ashton, Ashton, and Davis 1994). But in contexts where decision-making requires nuanced judgments, its shortcomings can be stark. Most importantly, often we have no ability to measure much of the information that judges use in a complex decision task. Without that information, judgmental bootstrapping will, along with removing noise, remove much of the signal that it is designed to capture.

Machine learning offers several advantages over traditional regression-based methods. Machine learning algorithms can search over a rich set of variables and functional forms, thus capturing more signal in the data. And machine learning minimizes prediction error by using the data itself to decide how to make the bias-variance trade-off. Importantly, we dispense with any restriction that the model include only variables actually used by the decision-makers, and we instead exclude only the noise variables described above—those variables that we have good reason to believe are statistically unrelated to the actual merits of a case. These variables may include the judge to which one happens to be assigned, the results of the immediately preceding cases, the time of day, whether the judge's football team won the night before, the weather, and the judge's mood. By excluding these variables, predictions are averaged over the arbitrary circumstances that often influence case outcomes.

Our recommendation to include all variables that may be statistically related to the merits of a case is likely to be met with skepticism insofar as it implies the use of morally and constitutionally suspect variables like race and gender. It is now well understood, however, that the problem of algorithm-embedded bias cannot be resolved by simply excluding the variable of concern (e.g., race) from the predictive model. This is because other variables may serve as proxies for the bias-inducing variable, and the model may then simply capture biases through those proxies. Attempting to eliminate those proxies is also unlikely to be successful, as ostensibly benign variables might interact in complex ways that are difficult to identify. We discuss more promising approaches to ameliorating systematic bias in Section 5.

Building a model with machine learning rather than traditional linear regression methods not only allows for better signal capture, but it also has the benefit of separating the predictive task from controversial normative choices. The traditional regression approach puts modeling choices in the hands of the analyst. Different statistical models may generate different predictions and associated recommendations. An analyst may, intentionally or not, choose a model that generates recommendations in accordance with his or her own normative preferences. A machine learning approach lets the data determine which model or combination of models generates the best predictions.

3.3 Evaluating the Simulation Results

Responsibly employing the simulation results of a synthetic crowdsourcing model requires an assessment of the simulation's success. We identify three key dimensions along which the simulation results can be evaluated: calibration, discrimination, and residual system noise. The first two are standard performance metrics; the third is unique to synthetic crowdsourcing.

Like any classification model, a synthetic crowdsourcing model should be well calibrated. That is, a good model will demonstrate agreement between observed outcomes and predictions: X% of cases with a predicted probability of .X should have a positive outcome.

Unlike other classification models, a synthetic crowdsourcing model cannot be adequately assessed with standard measures of discrimination. Standard measures of discrimination (i.e., how well a model separates positive and negative outcomes) will systematically understate the success of a synthetic crowdsourcing model.[44] As applied to typical classification models, any failure to discriminate is a failure of the model: ideally, all predicted probabilities are either a 1 or a 0. The same is not true of synthetic crowdsourcing models. Consider, for example, a set of cases with predicted probabilities of .75. When generated from a synthetic crowdsourcing model, the correct interpretation of those probabilities is somewhere between two ends of a spectrum. On one end of the spectrum, each case has a 75% chance of a positive outcome, and the actual outcome is determined by factors excluded from the model, such as the judge randomly assigned to hear the case or the mood of the judge. On the other end of the spectrum, 75% of the cases would always end in a positive outcome and 25% would always end in a negative outcome regardless of random factors. Standard measures of discrimination reflect the latter interpretation, an interpretation that is only correct if the law is perfectly applied without judicial idiosyncrasy.

We propose two methods for assessing the extent to which standard metrics understate discriminatory power due to residual system noise. The first uses alternative markers of case merits such as the results of appellate review, recidivism in the context of criminal justice decisions, or subsequent employment in the case of social security disability decisions. If residual system noise is high, then there should be a strong positive relationship between predicted probabilities and case merits regardless of the actual decision outcome. In other words, a case with a high (low) predicted probability should reflect the fact that most judges, most of the time, would decide the case positively (negatively). Cases with high (low) predicted probabilities that nonetheless result in a negative (positive) outcome should be products of residual systemic noise, such as an extremist judge or a judge in an extreme mood. Thus, within the subsets of positively and negatively decided cases, one would expect positive correlations between predicted probabilities and markers of merits. On the other hand, if residual systemic noise is low, no such correlations would be expected: a high probability case that nonetheless results in a negative outcome is not an judge-induced aberration—it is just a consensus decision that the model failed to identify.[45]

[44] A common metric of discrimination is the area under the receiver operating curve (AUC), which provides the probability that a randomly selected case with a positive result (a hearing that ended in a grant) would have a higher predicted probability than a randomly selected case with a negative result (a hearing that ended in a denial). An area of 1 indicates a model that can perfectly discriminate between grants and denials; a model that is no better than random would score a 0.5.

[45] Our suggestion to use alternative markers of case merits raises the question: why not directly use those markers of case merits to build a model? The reason is two-fold. First, in most contexts, the ability to observe a marker of case merit is conditional on the outcome. For example, a researcher studying parole can only observe recidivism if an inmate is released from prison. It is precarious to apply a model to the entire population (e.g., paroled and not paroled inmates) if the model was built only on a subset of the population (e.g., paroled inmates) that may be observably different than the entire population. There have been recent advances against this problem, but it remains a substantial hurdle (Lakkaraju et al. 2017). Second, the alternative marker of merit may be a lower quality marker than judicial consensus (the concept targeted by

Absent an alternative marker of case merits, a second approach is to directly measure the residual system noise (i.e., the amount of intra and inter-judge disagreement) and to use those measurements to reduce the amount of variation that a synthetic crowdsourcing model is designed to explain. For example, if we know 20% of the decision-making is due to as-if randomly assigned factors excluded from the model, we should only expect a perfect synthetic crowdsourcing model to explain 80% of the decision variance. This approach is more difficult than using external markers of case merits, as there are technical impediments to the measurement of noise. But recent research shows how machine learning can improve measurement of noise in adjudication systems (Copus and Hubert 2017).

3.4 Employing the Simulation Results

Synthetic crowdsourcing can be used as a tool for targeting review of decisions or as a guide for making initial decisions. Implementation is simplest in the case of review. Synthetic crowdsourcing predictions can be used to target review resources towards cases that are most likely to be a product of deviations from judicial consensus. As resources permit, those cases with the highest predicted probabilities of a positive outcome that are decided negatively, and those cases with the lowest predicted probabilities that are decided positively, can be flagged for secondary or appellate review.

Employing the simulation results to help guide the primary decision demands more involved implementation choices. The predicted probabilities from a synthetic crowdsourcing model should generally be binned and converted to simple recommendations that judges can easily make sense of an apply; a raw predicted probability is likely to be used differently by different judges, doing little to address the problem of inconsistency that synthetic crowdsourcing is designed to address. For example, in the parole context, we might create three bins: cases with a low peer assessment (.0-.29 probability of parole release), moderate peer assessment (.30-.70 probability), and high peer assessment (.71-1) probability).[46] This binning parallels standard risk assessment instruments in the criminal justice system: predicted probabilities of re-offending are binned to indicate offender risk level (e.g. low, medium, high), and risk levels are associated with recommended judicial decisions (e.g., release, use discretion, detain).

4 Proof of Concept: California Board of Parole Hearings Release Decisions

In the following section, we build a decision predictive model of California Board of Parole Hearings decisions to demonstrate the potential of synthetic crowdsourcing to improve decision making.

synthetic crowdsourcing). Some scholars, for example, have argued that recidivism is a poor marker of merit in the criminal justice system because it only captures the incapacitation benefits of imprisonment (Starr 2014).

[46] Finer bins yield more precise guidance, but they also tax judicial competence to follow that guidance.

Through a public records request we obtained the population of California Board of Parole Hearings (the "Board") suitability hearing transcripts conducted between 2011 - 2014. The dataset was built using Python regular expressions to pull key information from each hearing transcript. The extracted variables used in the decision predictive algorithm include the commitment crime, the psychological risk assessment score as well as the identity of the evaluating psychologist, the minimum eligible parole date, the inmate's lawyer, the district attorney if present at the hearing, the number of victims present at the hearing, whether or not an interpreter was present at the hearing, the results of any previous suitability hearings, the inmate's date of entry into prison, and information concerning how many times the inmate has appeared before the Board, and the inmate's prison.[47]

As discussed in Section 3, in constructing the synthetic crowdsourcing model, we exclude variables that are, as a matter of system design, statistically unrelated to the merits of a case. That is, we exclude variables that are as-if randomly assigned. For example, we do not include the identity of the parole commissioners assigned to decide a case, the time of day a hearing is scheduled, or whether a judge's football team won the night before. Again, these exclusions from the model are an essential feature of synthetic crowdsourcing, which is designed to eliminate the arbitrary elements of the decision-making system.

We construct our predictive algorithm using *Super Learner*, a generalized stacking ensemble learning technique in which the output from a set of diverse base learning algorithms is combined via a meta-learning algorithm. The *Super Learner* has been theoretically proven to represent an asymptotically optimal system for combining the base-level predictions via cross-validated risk minimization (Laan, Polley, and Hubbard 2007b), with "risk" defined by a user-specified objective function, such as minimizing mean squared error or maximizing the area under the receiver operating characteristic curve.

Super Learner takes as input any number of user-supplied algorithms (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.

4.1 Evaluating The Simulation Results

We evaluate the simulation along the key dimensions identified in Section 3.3: calibration, discrimination, and residual systemic noise.

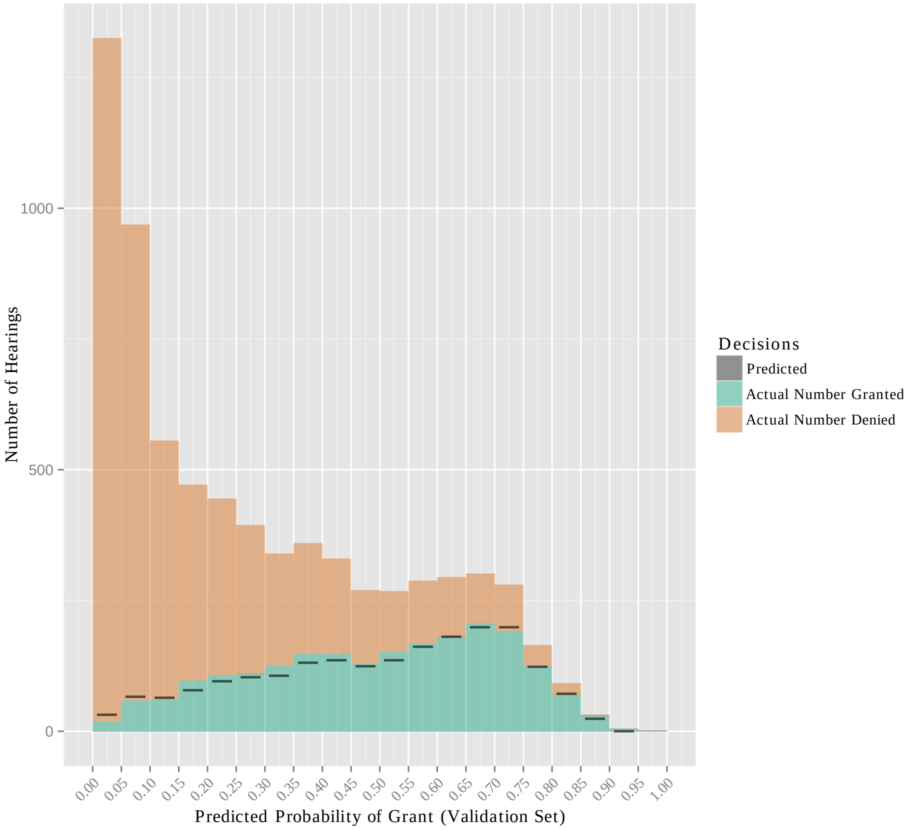
4.1.1 Calibration and Discrimination

Our model correctly predicts validation-set 2011-2014 suitability hearing decisions with 79% accuracy (the grant rate in this period was 28%). The cross-validated area under the ROC curve (AUC) is .80, meaning that, given any random pair of decisions, one a grant and one

[47] We also extracted information on a limited number of 'noise' variables - variables that are as-if randomly assigned - including the presiding and deputy commissioners, the date and time of the hearing, and the results of the immediately preceding hearings.

a denial, the model has an 80% success rate in ranking the grant higher than the denial in terms of predicted probabilities. Figure 10 graphically portrays the model's discriminatory power and calibration. The figure reveals the wide distribution of predicted probabilities provided by the model. It is particularly effective at identifying hearings that have a very low probability of resulting in a grant. The figure also shows that the model is well calibrated, indicating the actual proportion of hearings that resulted in a grant or denial closely tracks the model's (validation-set) predicted probabilities.

Figure 10: Validation Set Parole Predictions: 2011 - 2014



4.1.2 Evidence of Residual System Noise

Again, standard measures of discrimination understate the performance of synthetic crowdsourcing because the model is designed to exclude variation from as-if randomly assigned factors. We therefore want assurance that the distance between the predicted probabilities and zero or one reflect residual system noise rather than a mere lack of predictive power.

As described in section 3.3, if there is substantial residual system noise, one should expect a positive correlation between markers of case merit and a synthetic crowdsourcing model's predicted probabilities of a positive case outcome. In California, the Governor has the ability to review and reverse any parole granted to an inmate convicted of murder, and we use the results of the California Governor's review process as an external marker of case merit. Insofar as the model is merely failing to predict collective decision-making, we should expect no relationship between predicted probabilities and reversal rates. If, on the other hand, the model is excluding system noise, we would expect the Governor to reverse low probability grants more often than high probability grants as he corrects for commissioner deviations. Some readers might object that the Governor's assessment of a case is a poor marker of case merit. Perhaps, for example, the Governor is more concerned with political considerations than with assessing an inmate's chances of recidivating. But as long as the Governor's decisions are merely unrelated to merit rather than negatively correlated with it, the disconnect between actual merit and a marker of merit would only make it more difficult to find evidence of residual system noise.[48]

The relationship between the Governor's reversal rate and predicted probabilities demonstrates that the model is excluding and smoothing over a substantial amount of residual system noise. Figure 11 shows relationship between validation set predicted probabilities and the Governor's reversal rate.[49] Cases with low predicted probabilities are significantly more likely to be reversed by the Governor, which is the expected relationship when there is substantial residual system noise.

The calibration and discrimination metrics, along with evidence of residual system noise, strongly suggest that the model could be used to help the Board make better decisions. Importantly, our analysis is necessarily limited by the data we have access to: the variables that we could accurately extract from parole hearing transcripts. Were administrators of the parole system to actually implement a synthetic crowdsourcing model, it could be built with the more expansive set of variables maintained by the California Department of Corrections and Rehabilitation (CDCR) and the Board of Parole Hearings, thereby increasing the signal to noise ratio even further.

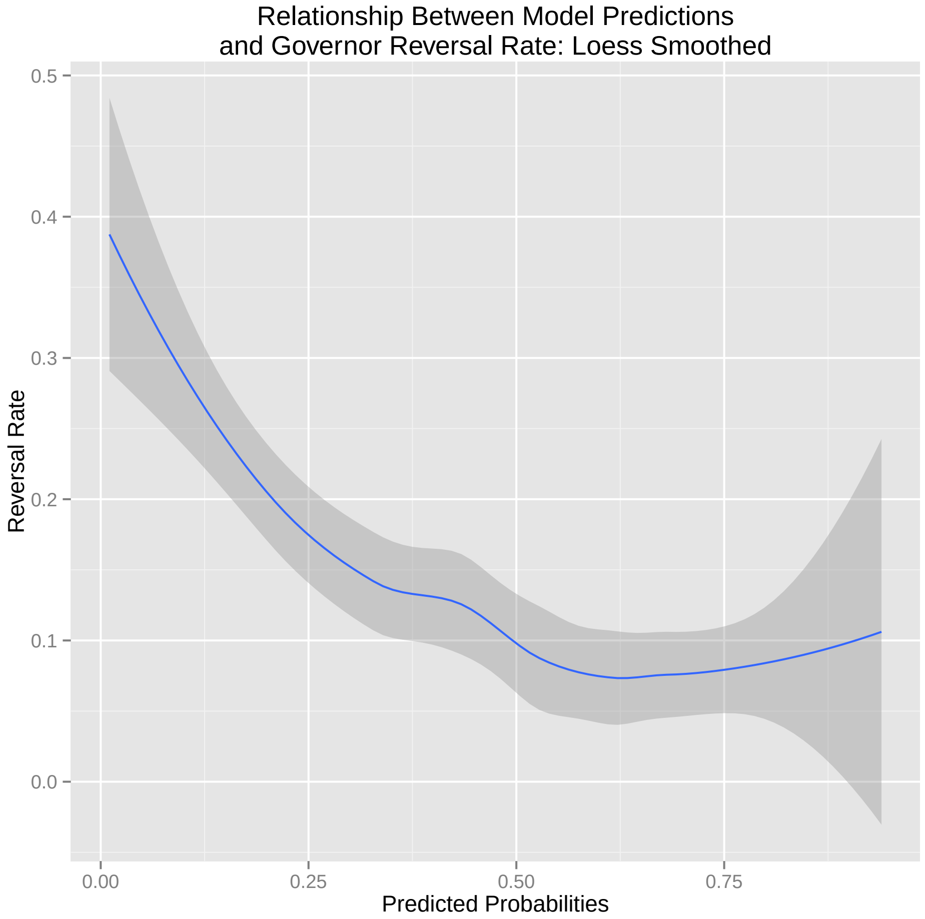
4.2 Implementation

A synthetic crowdsourcing model could improve both of the review processes currently employed in the parole system: the Governor reviews of all parole grants (for inmates convicted of murder), and the Board provides secondary review of cases when the assigned commissioners cannot come to a consensus. While Governor review may correct for unusual decisions to *grant* parole, it does nothing to correct for unusual decisions to *deny* parole. The governor's power to review could be extended to cases where parole was denied but the synthetic crowdsourcing model generated a high probability of a grant. The Board's internal review practices could also be improved upon with synthetic crowdsourcing. Relying merely on

[48] This is a simple extension of the well known fact that correlation coefficients are attenuated by random measurement error.

[49] As a proxy for governor reversals, we denote a case as reversed if an inmate's parole is granted but the inmate reappears in another hearing at a later date. In order to avoid problems of potentially biased missingness, we restrict the analysis to 2011 and 2012 hearings so that there is sufficient passage of time for inmates to show up in the dataset again if their parole is reversed.

Figure 11



explicit disagreement between commissioners to flag cases will miss the great majority of controversial decisions: judges explicitly disagree in less than 0.1% of cases. But a simple regression of the outcome on commissioner with prison-year fixed effects shows that implicit disagreement is significantly higher: among the 10 commissioners with the most decisions (> 600), grant rates differ by as much as 15% (p <.001). A synthetic crowdsourcing model can help regulate these sorts of controversial decisions by flagging low-probability grants and high-probability denials for secondary review.

Synthetic crowdsourcing could also help guide the primary parole decisions made by the Board of Parole Hearings commissioners. Such an implementation would require judgments along two important dimensions. First, an algorithmic recommendation could be more or less binding. For example, the algorithmic prediction could be offered as a mere recommendation to commissioners, serving as an anchor, but with commissioners free to ignore it. Alternatively, and more prescriptively, the prediction could be used as a third vote on a two-member panel, such that a single decision maker's agreement with the algorithm would be sufficient to decide the outcome. Second, the algorithmic recommendation could be more or less granular in its presentation of the model's predicted probabilities. It could, for example, be presented as a binary recommendation: if the predicted probability of an inmate's parole is above a specified threshold, the algorithm recommends parole. Or the predicted probabilities may be binned into three categories - low, medium, and high probability of release - with corresponding recommendations of denial for low, release for high, and discretion for medium. Regardless of the exact implementation, the notion is that the commissioner decisions would remain discretionary but would coincide more frequently with the algorithm—and the collective judgment of commissioners—than they would in the absence of the implementation.

5 Limitations

We now move from the parole example to discuss general limitations of synthetic crowdsourcing. We focus the discussion on synthetic crowdsourcing as tool for guiding primary decisions, where the issues are most pronounced. The concerns are substantially mitigated when synthetic crowdsourcing is simply used to target review of inconsistent decisions.

5.1 Biased Decision-Making

As noted in Section 3, synthetic crowdsourcing can help to ameliorate inter and intra-judge inconsistency, but, on its own, it does not address systematic biases. If judges, on average, discriminate against certain types of litigants, that bias can be embedded in the model. Racial bias is perhaps the most obvious concern, particularly in the criminal justice system.[50] If, for example, parole commissioners systematically release black inmates at a lower rate than their otherwise identical white counterparts, black inmates will, unfairly, have systematically lower predicted probabilities of release. Assuming synthetic crowdsourcing is used as a decision aid, with binned probabilities associated with release recommendations, this will impact inmates whose predicted probabilities of release are right above or below a decision threshold.

Embedded biases can be addressed in two ways. First, the problem can be assessed empirically to at least ensure that synthetic crowdsourcing is not inflating bias. That is, before implementation, recent decisions can be compared to what the synthetic crowdsourcing model would have recommended to assess whether relevant disparities are widened under the synthetically crowdsourced recommendations. If indeed synthetically crowdsourced predictions would increase disparities, this can be addressed by adjusting group-specific recommendation bins so as to maintain the status quo outcome distribution.[51]

[50] A substantial literature assessing discrimination in criminal justice decision making has generated mixed evidence (Mechoulan and Sahuguet 2015), but concern is still serious and widespread.

[51] One legal concern with this approach is that it has at least surface-level similarities to the type of mechanical race-based adjustments that have been found unconstitutional in the higher education context.

A second approach, if there is measurable bias in decisions, is to explicitly adjust the synthetically crowdsourced predictions to correct for the bias. That is, estimates of the causal impact of inmate race on decision-making can be used to adjust the algorithm's output. For example, if the estimate of disparate treatment of black offenders is a 10% reduction in parole chances, this method would add .10 to black offenders' predicted probabilities. The challenge of explicitly adjusting the algorithmic output is the challenge of estimating the causal effect of race on decision-making. While estimating discrimination is empirically challenging, recent advances in data collection and estimation techniques are allowing for more plausible causal inference with observational data. Methods such as propensity score matching (Hirano, Imbens, and Ridder 2003; Abadie and Imbens 2006) and those that combine treatment modeling with outcome modeling (Rubin 1979; Van der Laan and Robins 2003) provide extra traction in observational studies. These advances in research design are increasingly being coupled with non-parametric, data-adaptive estimation techniques, such as genetic matching (Diamond and Sekhon 2013) and targeted learning with *Super Learner*(Van der Laan and Rose 2011). New design and estimation techniques have the benefit of separating the research design from the estimation procedure so as to guard against specification searching. While data-adaptive estimation techniques cannot eliminate the need for assessing which variables must be controlled for, they can at least largely eliminate the debate over *how* variables are controlled for, allowing the data to determine which models are best.These advances in data, design and estimation are unlikely to ease all concerns of omitted variable bias, but the threshold of certainty required for the advancement of scientific knowledge should often be relaxed in policy world contexts that require action. Academics enjoy the luxury of waiting around or choosing questions based on the availability of a strong quasi-experimental research design, but the world sometimes demands a best answer. One might proceed even without robust causal estimates if the biases are particularly important to address.

5.2 Status Quo Bias

A second consideration is that synthetic crowdsourcing is inherently backward looking: using statistical predictions of past decisions to inform future decisions tethers future decisions to the status quo. This tethering to the past may be problematic if past decisions are or become incompatible with current values.[52] For example, in the California parole context, an algorithm built in the early 2000s would be modeled on an era in which fewer than 5% of inmates were released from prison (the current rate of release is over 30%). On the other hand, in some contexts, tying future decisions to past decisions may might offer an appealing temporal consistency, a version of legal precedent or institutional memory. Furthermore, synthetic crowdsourcing need not favor the status quo: it can be used quickly update a

[52] The tethering of past to future is not complete, as synthetic crowdsourcing models can automatically be updated as values shift. Insofar as judges express new values by deviating from the synthetic crowdsourcing recommendations, updated models will incorporate those deviations. The updating process can also be accelerated by selecting some random subset of cases to be decided without the aid of the synthetic crowdsourcing model. Departures from recommendations could alert administrators to a malfunctioning model and be used to update the model.

system in accordance with new values. For example, adjusting the release recommendation thresholds or making a racial bias correction to the algorithmic output could effect change that might otherwise come only with slow shifts in judicial attitudes.

5.3 Public Access and Litigant Adaptation

Recent litigation has highlighted the tension between proprietary rights to statistical algorithms used in the public arena and the public's ability to inspect those algorithms. In 2017, the US Supreme Court denied a petition for certiorari in *Loomis v. Wisconsin*, a case in which Loomis argued his due process rights were violated because the sentencing court relied on the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) criminal-risk instrument. The proprietary nature of COMPAS had prevented Loomis from challenging the algorithm's accuracy and scientific validity.

The focus on the tension between public access and proprietary rights has overshadowed what might be a more fundamental tension: given public access, parties can adapt to models in ways that undermine model accuracy. The problem of litigant adaptation comes in two forms. First, users of the legal system may be able to strategically alter their "variables" so as to obtain more favorable algorithmic recommendations. Consider, for example, an inmate's attorney in the California parole system. Some private attorneys are moderately associated with higher chances of parole. While that may be in part causal, some of the association is likely due to correlation with unobservables: those more eligible for release may also be more likely to hire a private attorney. If inmates know that an algorithm will give them a better recommendation if they hire certain attorneys, they may be more likely to do so, artificially boosting their release recommendations. Second, public access could result in parties entering adjudication systems that they would not have entered absent knowledge of the algorithm. This can create a disjunct between the population targeted by a model and the population to which it is applied. Consider an inmate who, absent an algorithm, would have chosen to defer his parole hearing because of his correct belief that parole commissioners would have judged him unsuitable for release. Despite the inmate's weak case for parole, he may share some characteristics—characteristic used by the algorithm—with inmates who have a high probability of release. Knowledge of his algorithmic recommendation might convince the inmate to proceed with his scheduled parole hearing. In such a case, the synthetic crowdsourcing model would incorrectly inform commissioners that their peers would generally recommend parole.

Technical solutions to the problem of litigant adaptation can obviate the need for restricting public access.[53] One obvious means of avoiding variable manipulation is to exclude variables that litigants can inexpensively manipulate. The drawback of this approach is that such variables may be highly predictive, so excluding them could substantially reduce

[53] Restricting public access is unlikely to be a long-term solution to the problem of litigant adaptation. Even if private companies or system administrators could be trusted to secretly develop accurate and well-functioning algorithms, maintaining that secrecy is likely to prove difficult. In a world of growing statistical and coding literacy, open source statistical software, and web-scraped data, efforts to reverse engineer models are likely to be plentiful.

the accuracy of models. An approach that helps to circumvent the strategic entry concern is to couple the decision recommendations with statistical models of entry. This strategy will only work, however, if data is available for both the population of actual litigants and potential litigants, and requires the ratio of potential litigants to actual litigants not be so high as to make identifying new types of entrants statistically unfeasible. A third solution, which addresses both strategic manipulation of variables and strategic entry into adjudication system, is to employ multiple output-equivalent models that use heterogeneous variables. Where data is aggressively recorded, it is likely that many variables and/or interactions of variables serve as reliable proxies for other variables/interactions. In this case, one set of variables and interactions can be substituted for another set without substantial changes to model predictions. By either randomly selecting one of the models or averaging over all models, this multiple-model strategy decreases litigant incentives to adapt. For example, if there are two models, the expected benefit from strategic manipulation of a variable is reduced by 50%. The incentive to strategically enter the system is also reduced: to obtain the full benefit of a excessively high recommendation with multiple models, a non-entrant must match entrants on the variables used in all of the models rather than in just one.

5.4 The Black Box Problem & Procedural Fairness

Even allowing public access to algorithms cannot entirely solve the problem of transparency. Machine learning as a statistical technique is inherently ”black box”—we can see the input and the output, but what happens in between is opaque. As machine learning is used in a growing number of real world applications from medicine to finance to public administration of criminal justice, health, and welfare, there is growing debate around the extent to which this black-box character represents a problem. We cannot resolve the general debate, and there are of course trade-offs between the superior predictive performance offered by machine learning and the relative opacity of the procedure. But for the purposes of considering synthetic crowdsourcing, we simply note that we are not advocating for automated decision making. Rather, we suggest machine learning predictions can *assist* human decisions. Synthetic crowdsourcing aims to simply nudge judges toward consensus to help mitigate the influence of mood and other judicial idiosyncrasies.

The black box nature of machine learning intensifies a related concern regarding the use of algorithms in administrative decision making: the idea that procedural fairness requires the ability to understand and engage with a decision. A large body of research in psychology shows that people care greatly about the process by which outcomes are reached, even if the result is an outcome they find unfavorable (Lind and Tyler 1988). Insofar as people think equitable and legitimate decisions are those in which each person is treated individually, an algorithm-assisted approach, particular one employing machine learning, may violate this sense of procedural fairness. Here too we underscore that employing synthetic crowdsourcing by no means implies replacing humans with machine. Further, it is not self-evident that fairness intuitions would, or at least *should*, be in greater conflict with an algorithm-assisted process than they are in a system where the idiosyncrasies or mood of a particular judge can dramatically affect the outcome of a case.

6 Conclusion

In this paper we have presented a powerful tool for reducing judge-to-judge and within-judge inconsistencies in decision making. By excluding variables that are statistically uncorrelated with the merits of a case (e.g., the identity of randomly assigned judges or whether a judge's football team won the night before) and aggregating judgments across and within decision makers, synthetic crowdsourcing decision models can cancel out arbitrary and contingent factors, smooth over sources of inconsistency, and capture the ”wisdom of the crowd.” The results of the voting simulation can then be used to guide future decision making or to flag deviant decisions for further review.

References

Abadie, Alberto, and Guido W. Imbens. 2006. “Large Sample Properties of Matching Estimators for Average Treatment Effects.” 01383, *econometrica* 74 (1): 235–267. Accessed May 22, 2016. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2006.00655.x/abstract.

Abdel-Khalik, A. Rashad, and Kamal M. El-Sheshai. 1980. “Information Choice and Utilization in an Experiment on Default Prediction.” 00120, *Journal of Accounting Research:* 325–342. Accessed October 7, 2015. http://www.jstor.org/stable/2490581.

Abrams, David, Marianne Bertrand, and Sendhil Mullainathan. 2012. “Do Judges Vary in Their Treatment of Race?” 00113, *Journal of Legal Studies* 41 (2): 347–383. Accessed October 7, 2015. http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1800840.

Alameda County Grand Jury. 2011. *2010-2011 Alameda County Grand Jury Final Report.* Technical report. http://www.acgov.org/grandjury/final2010-2011.pdf.

Alschuler, Albert W. 1991. “The failure of sentencing guidelines: A plea for less aggregation.” *The University of Chicago Law Review* 58 (3): 901–951.

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.

Armstrong, J. Scott. 2006. “Findings from Evidence-Based Forecasting: Methods for Reducing Forecast Error.” 00148, *International Journal of Forecasting* 22 (3): 583–598. Accessed October 5, 2015. http://www.sciencedirect.com/science/article/pii/S0169207006000537.

Ashton, Alison Hubbard, Robert H. Ashton, and Mary N. Davis. 1994. “White-Collar Robotics: Levering Managerial Decision Making.” 00009, *California Management Review* 37 (1): 83. Accessed October 7, 2015. http://search.proquest.com/openview/8590459ba0156326aa9c3ad830c53746/1?pq-origsite=gscholar.

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

Austen-Smith, David, and Jeffrey S Banks. 1996. “Information aggregation, rationality, and the Condorcet jury theorem.” *American Political Science Review* 90 (1): 34–45.

Austin, William, and Thomas A Williams III. 1977. “A survey of judges’ responses to simulated legal cases: Research note on sentencing disparity.” *J. Crim. L. & Criminology* 68:306.

Benitez-Silva, Hugo, Moshe Buchinsky, and John Rust. 2004. *How large are the classification errors in the social security disability award process?* Technical report. National Bureau of Economic Research.

Berk, Richard. 2012. *Criminal Justice Forecasts of Risk: A Machine Learning Approach.* 00021. Springer Science & Business Media. Accessed May 22, 2016. https://books.google.com/books?hl=en&lr=&id=Jrlb6Or8YisC&oi=fnd&pg=PR3&dq=Richard+Berk+(2012+machine+learning&ots=IuCe6dipsc&sig=CBHsKVxRWYVmFX4kxx8WOyFSyG0.

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. doi:10.1111/j.1540-5907.2010.00437.x.

Brantingham, Patricia L. 1985. "Sentencing disparity: An analysis of judicial consistency." *Journal of Quantitative Criminology* 1 (3): 281–305.

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.

Cameron, Charles M., and Lewis A. Kornhauser. 2010. "Modeling Collegial Courts (3): Adjudication Equilibria." http://ssrn.com/abstract=2153785.

Chen, D. L., and H. Spamann. 2014. *This Morning's Breakfast, Last Night's Game: Detecting Extraneous Factors in Judging.* Technical report. 00009. Working paper, ETH Zurich.

Chen, Daniel L, Tobias J Moskowitz, and Kelly Shue. 2016. "Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *The Quarterly Journal of Economics* 131 (3): 1181–1242.

Chen, Daniel Li, and Holger Spamann. 2016. "This Morning's Breakfast, Last Night's Game: Detecting Extraneous Influences on Judging." *Social Science Research Network Working Paper Series. The Impact of Value-Irrelevant Events on the Market Pricing of Earnings News. Contemporary Accounting Research* 33 (1): 172–203.

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.

Copus, Ryan, and Ryan Hubert. 2017. "Detecting Inconsistency in Governance." Accessed November 1, 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2812914.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." 00356, *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892. Accessed October 5, 2015. http://www.pnas.org/content/108/17/6889.short.

Deutsch, Emily Woodward, and Michael Donohue. 2009. "The Role of New Media in the Veterans Benefits Arena." *Veterans L. Rev* 1:183.

Dhami, Mandeep K. 2005. "From Discretion to Disagreement: Explaining Disparities in Judges' Pretrial Decisions." *Behavioral Sciences and the Law* 23 (3): 367–386.

Diamond, Alexis, and Jasjeet S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." 00458, *Review of Economics and Statistics* 95 (3): 932–945. Accessed May 22, 2016. `http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00318`.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey. 2015. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* 144 (1): 114.

Edwards, Harry T., and Michael A. Livermore. 2009. "Pitfalls of Empirical Studies that Attempt to Understand the Factors Affecting Appellate Decisionmaking." *Duke Law Journal* 58 (8): 1895–1989.

Epstein, Lee, William M. Landes, and Richard A. Posner. 2013. *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice.* Cambridge, MA: Harvard University Press.

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

———. 2014b. "Measuring Inconsistency, Indeterminacy, and Error in Adjudication." 00011, *American law and economics review* 16 (1): 40–85. Accessed October 5, 2015. `http://aler.oxfordjournals.org/content/16/1/40.short`.

Freeman, Katherine. 2016. "Algorithmic Injustice: How the Wisconsin Supreme Court Failed To Protect Due Process Rights in State v. Loomis." *North Carolina Journal of Law and Technology* 18:75–180.

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

Grunwald, Ben. 2015. "Questioning Blackmun's Thesis: Does Uniformity in Sentencing Entail Unfairness?" *Law & Society Review* 49 (2): 499–534.

Halberstam, Yosh. 2015. "Trial and Error: Decision Reversal and Panel Size in State Courts." *The Journal of Law, Economics, and Organization* 32 (1): 94–118.

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

Hausman, David. 2016. “Consistency and Administrative Review.”

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score.” 01273, *Econometrica* 71 (4): 1161–1189. Accessed May 22, 2016. `http://onlinelibrary.wiley.com/doi/10.1111/1468-0262.00442/abstract`.

Ho, Daniel E. 2017. “Does Peer Review Work? An Experiment of Experimentalism.” *Stanford Law Review* 69:1–119.

Ho, Daniel E. 2017. “Does Peer Review Work: An Experiment of Experimentalism.” *Stan. L. Rev.* 69:1.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 (396): 945–960. doi:`10.2307/2289064`.

Kaplow, Louis. 1992. “Rules versus standards: An economic analysis.” *Duke Lj* 42:557.

Kastellec, Jonathan P. 2013. “Racial Diversity and Judicial Influence on Appellate Courts.” *American Journal of Political Science* 57 (1): 167–183. doi:`10.1111/j.1540-5907.2012.00618.x`.

Krent, Harold J, and Scott Morris. 2013. “Achieving Greater Consistency in Social Security Disability Adjudication: An Empirical Study and Suggested Reforms.” In *Administrative Conference of the United States (April 3). Available at https://www. acus. gov/sites/default/files/documents/Achieving_Greater_Consistency_Fin al_Report_4-3-2013_clean. pdf.*

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 1999. “The Quality of Government.” *Journal of Law, Economics, and Organization* 15 (1): 222–279.

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

———. 2007b. “Super Learner.” 00267, *Statistical applications in genetics and molecular biology* 6 (1). Accessed October 5, 2015. `http://www.degruyter.com/view/j/sagmb.2007.6.1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml`.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. “The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables.” In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 275–284. ACM.

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

Lax, Jeffrey R. 2011. "The New Judicial Politics of Legal Doctrine." *Annual Review of Political Science* 14:131–157. doi:10.1146/annurev.polisci.042108.134842.

Legomsky, Stephen H. 2007. "Learning to Live with Unequal Justice: Asylum and the Limits to Consistency." *Stanford Law Review:* 413–474.

Lind, E Allan, and Tom R Tyler. 1988. *The social psychology of procedural justice.* Springer Science & Business Media.

Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services.* New York: Russell Sage Foundation.

Mashaw, Jerry L. 1985. *Bureaucratic justice: Managing social security disability claims.* Yale University Press.

Mechoulan, Stéphane, and Nicolas Sahuguet. 2015. "Assessing racial disparities in parole release." *The Journal of Legal Studies* 44 (1): 39–74.

Moscovici, Serge, and Marisa Zavalloni. 1969. "The group as a polarizer of attitudes." *Journal of personality and social psychology* 12 (2): 125.

Nakosteen, Robert, and Michael Zimmer. 2014a. "Approval of social security disability appeals: analysis of judges' decisions." *Applied Economics* 46 (23): 2783–2791.

———. 2014b. "Approval of Social Security Disability Appeals: Analysis of Judges' Decisions." *Applied Economics* 46 (23): 2783–2791.

———. 2014c. "Approval of Social Security Disability Appeals: Analysis of Judges' Decisions." 00001, *Applied Economics* 46 (23): 2783–2791. Accessed October 5, 2015. http://www.tandfonline.com/doi/abs/10.1080/00036846.2014.914147.

Narea, Nicole. 2017. *Iranian National Challenges USCIS Investor Visa Denial.* Accessed June 26, 2017.

Nickerson, Mike. 2007. "Child protection and child outcomes: Measuring the effects of foster care." *The American Economic Review* 97 (5): 1583–1610.

Ornstein, Charles, and Lena Groeger. 2012. *Two Deaths, Wildly Different Penalties: The Big Disparities in Nursing Home Oversight.*

Partridge, Anthony, and William Butler Eldridge. 1974. *The Second Circuit sentencing study: A report to the judges of the Second Circuit.* Federal Judicial Center.

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

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. 2007. "Refugee Roulette: Disparities in Asylum Adjudication." 00355, *Stanford Law Review:* 295–411. Accessed October 5, 2015. http://www.jstor.org/stable/40040412.

Ramji-Nogales, Jaya, Andrew I Schoenholtz, and Philip G Schrag. 2007. “Refugee roulette: Disparities in asylum adjudication.” *Stanford Law Review:* 295–411.

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.

Rose, Joel. 2017. *Canadians Report More Scrutiny And Rejection At U.S. Border Checkpoints.* `http://www.npr.org/2017/03/29/521920595/canadians-report-more-scrutiny-and-rejection-at-u-s-border-checkpoints`.

Rubin, Donald B. 1979. “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies.” 00671, *Journal of the American Statistical Association* 74 (366a): 318–328. Accessed May 22, 2016. `http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1979.10482513`.

Shapiro, Scott. 2006. “What Is the Internal Point of View?” *Fordham Law Review* 75:1157.

Starr, Sonja B. 2014. “Evidence-Based Sentencing and the Scientific Rationalization of Discrimination.” 00053, *Stan. L. Rev.* 66:803. Accessed October 5, 2015. `http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/stflr66&section=24`.

Sunstein, Cass R. 1995. “Problems with rules.” *California Law Review:* 953–1026.

———. 2005. “Group judgments: Statistical means, deliberation, and information markets.” *NYUL Rev.* 80:962.

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.

Taylor, Frederick Winslow. 1911. *The Principles of Scientific Management.* New York: Harper & Brothers Publishers.

Ting, Michael M. 2017. “Politics and Administration.” *American Journal of Political Science* 61 (2): 305–319.

Trump v. Int’l Refugee Assistance Project. 2017. 582 U. S. ______. LexisNexis Academic (June 27, 2017).

Turnbull, Lornet. 2013. *Suspicious Feds Turn Back Many Foreigners at Airport.* `http://www.seattletimes.com/seattle-news/suspicious-feds-turn-back-many-foreigners-at-airport/`.

Unpublished Judicial Opinions. 2002. Technical report. Washington, DC: U.S. House of Representatives Subcommittee on Courts, the Internet, and Intellectual Property.

Van der Laan, Mark J., and James M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality.* 00609. Springer Science & Business Media. Accessed May 22, 2016. https://books.google.com/books?hl=en&lr=&id=z4_-dXslTyYC&oi=fnd&pg=PR5&dq=Unified+Methods+for+Censored+Longitudinal+Data+and+Causality.&ots=uPP9VEFqeN&sig=_cQyRe35av79OveVZmJteVc0eAA.

Van der Laan, Mark J., and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data.* 00216. Springer Science & Business Media. Accessed October 5, 2015. https://books.google.com/books?hl=en&lr=&id=RGnSX5aCAgQC&oi=fnd&pg=PR3&dq=Targeted+Learning:+Causal+Inference+for+Observational+and+Experimental+Data&ots=FPb9cuAT7B&sig=P-RCh8efs-QtJYp1duNsq2W7Tbw.

Van Koppen, Peter J, and Jan Ten Kate. 1984. "Individual differences in judicial behavior: Personal characteristics and private law decision-making." *Law and Society Review:* 225–247.

Wilson, Woodrow. 1887. "The Study of Administration." *Political Science Quarterly* 2 (2): 197–222.

Appendices

A 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.

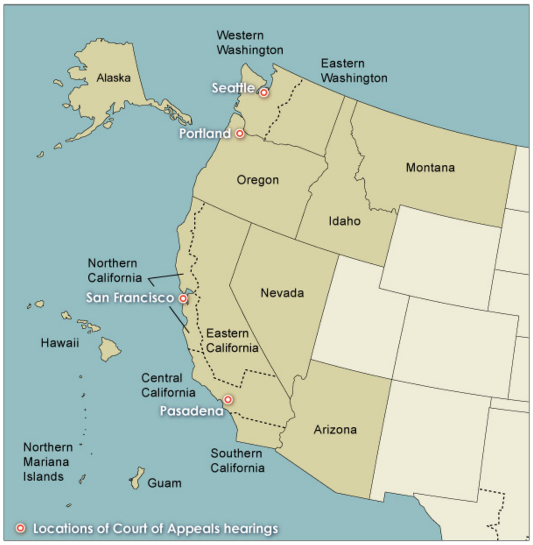
The first step generates validation set predictions for each data point for each underlying model. In the second step, `Super Learner` then leverages the cross-validation information on model performance to assign weights to each model according to how well their predictions match the true outcome. It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

B Data

Our main contribution is our methodological solutions to the problems of partitioning, clustering, and finite sample bias. However, the partitioning and clustering problems are more or less severe depending on the nature of the data one uses. 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.

To conduct our analysis, we constructed a large and extensively coded original dataset of *all* civil cases filed in the Ninth Circuit and terminated on the merits over a period of nineteen years. 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: The Ninth Circuit



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

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. 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`,[1] 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.[2]

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.

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.[3]

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 Appendix D), 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 1 and Table 2, we present some basic descriptive statistics for our sample.

[1] Code available upon request.

[2] As we discuss in Appendix D, there are potential deviations from random assignment in the Ninth Circuit. We account for this in two ways, which we describe in detail.

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

Table 1: Descriptive Statistics: Appeal Characteristics

	All Civil Cases		*Batched Cases*	
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 2: Descriptive Statistics: Trial Characteristics

	All Civil Cases		*Batched Cases*	
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	

C Judge Clusters

Table 3: 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 4: 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 5: 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 6: 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 7: 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 8: Visiting Cluster

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

D Threats to Identification

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.

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.[4] 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,[5] 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.

[4] 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.

[5] 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.

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

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

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

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

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