Using Machine Learning to Detect Human Rights Abuses

Daniel L. Chen*

Abstract

Predictive judicial analytics holds the promise of increasing efficiency and fairness of law. While much empirical work evaluates judges to observe inconsistencies in their behavior, the advent of machine learning tools offers an approach to automate the detection of inconsistencies and, concomittantly, indifference of judges to the law and to litigants. This article presents a theoretical framework to understand a large set of behavioral findings on judicial decision-making. First, settings where judges are closer to indifference among options are more likely to lead to detectable effects of behavioral biases outside the lab. Second, inter-judge disparities in prediction accuracy can reveal indifference of judges to the circumstances of cases. Third, algorithms can identify and reduce difference in judicial indifference. Fourth, implicit bias in judicial opinions and oral arguments can be predictive of disparities. Applications are illustrated with asylum judges, criminal justice, and textual and audio data. Machine learning may help detect and remedy due process violations and human rights abuses.

Keywords: Judicial Analytics, Due Process, Behavioral Judging

*Daniel L. Chen, daniel.chen@iast.fr, Toulouse School of Economics, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, Toulouse, France; dchen@law.harvard.edu, LWP, Harvard Law School. First draft: April 2018. Current draft: April 2018. Latest version at `nber.org/~dlchen/papers/Using_Machine_Learning_to_Detect_Human_Rights_Abuses.pdf` Work on this project was conducted while Chen received financial support from the European Research Council (Grant No. 614708), Swiss National Science Foundation (Grant Nos. 100018-152678 and 106014-150820), and Agence Nationale de la Recherche.

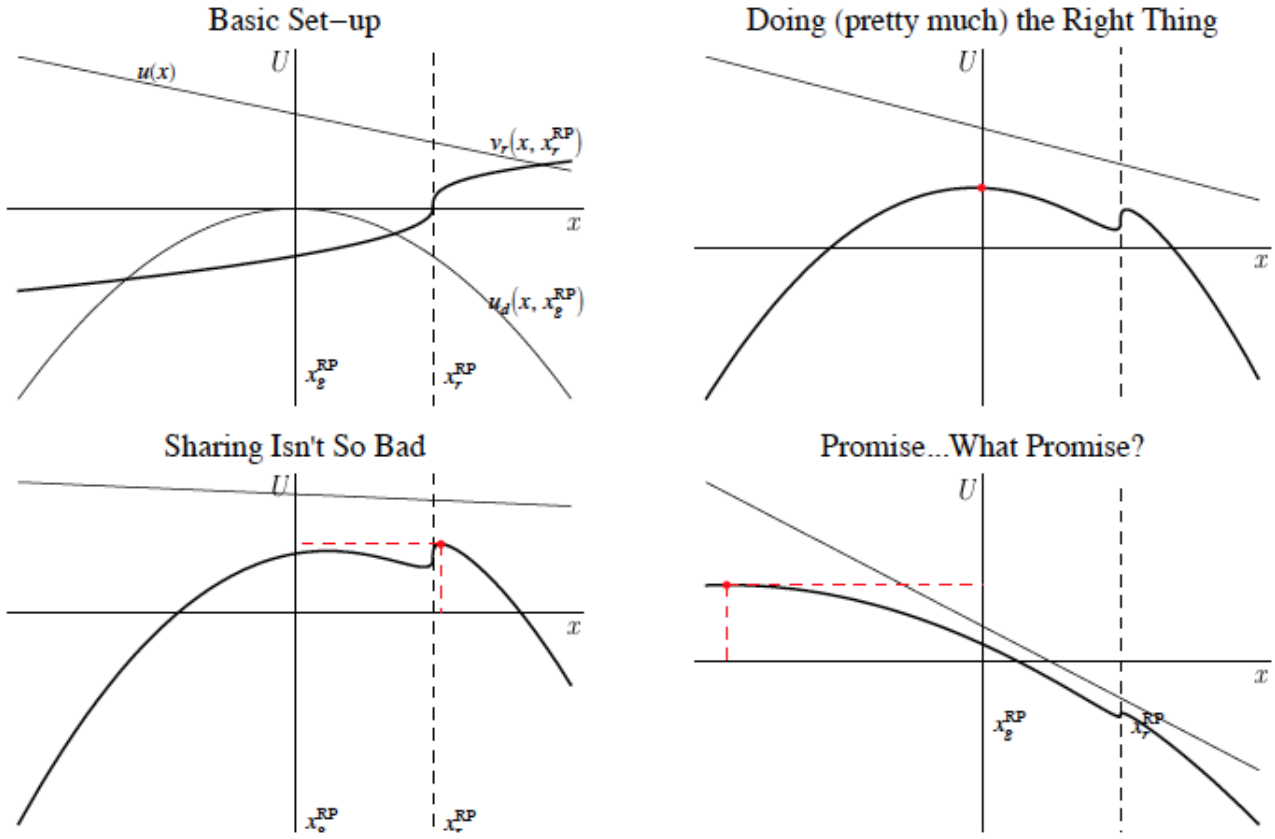
This article presents a set of ideas related to legitimacy in law, how to formalize recognition-respect theory, and what it means for legal institutions, judges -- to be indifferent, such that it violates our notion of justice. Consider a definition of justice as equal treatment before the law and equality based on recognition of difference. We can imagine a set of covariates X that should lead to the same prediction or predictability of outcomes $Y = f(X) + \varepsilon$; the X's should improve predictions. And, we can think of a set of W's that should not ($y \perp W, var(\varepsilon) \perp W$). We tend to think of X's as mutable—as consequences of choices ($a \rightarrow X, a \nrightarrow W$), and the W's as immutable, unrelated to one's actions, though this view is changing, the W's can also be mutable if they are expressions of one's identity.

Recent research has identified a collection of W's that appear to affect decisions. U.S. federal appeals court judges become more politicized before elections and more unified during war (Berdejo et al. 2016; Chen 2016a). Refugee asylum judges are 2 percentage points more likely to deny asylum to refugees if their previous decision granted asylum (Chen et al. 2015). Politics and race also appear to be influential (Schanzenbach 2005, Bushway and Piehl 2001, Mustard 2001, Steffensmeier and Demuth 2000, Albonetti 1997, Klein et al. 1978, Humphrey and Fogarty 1987, Thomson and Zingraff 1981, Abrams et al. 2012, Boyd et al. 2010; Shayo and Zussman 2011) as does masculinity (Chen et al. 2017a, 2016a), birthdays (Chen and Phlippe 2017), football (Chen 2017a; Eren and Mocan 2016), time of day (Chen and Eagel 2016, Danziger et al. 2011), weather (Barry et al. 2016), name (Chen 2016b), and shared biographies (Chen et al. 2016b) or dialects (Chen et al. 2016c). There are also various papers showing clear judicial biases in laboratory environments (Guthrie et al. 2000, 2007, Rachlinski et al. 2009, 2013, Simon 2012). In particular, these experiments identify racial bias (Rachlinski et al. 2009).

Information acquisition on the X's can be endogenous to judicial preferences. Let's say the parabola in Figure 1 captures what the judge thinks the sentence should be. Small deviations cost little, as the judge is not sent to jail. The defendant has a reference point on what is fair and just. Loss aversion with respect to this reference can lead to anger. Sympathy might be social preferences, but empathy is recognizing and respecting the reference points. Recognition-respect may be empathy, separate from sympathy.

One way to know if a judge is indifferent, is by observing where behavioral biases arise. Psychologists find many effects of moderate sizes in the lab, so settings where

Figure 1: Sympathy and Empathy



Source: Chen [2017b]

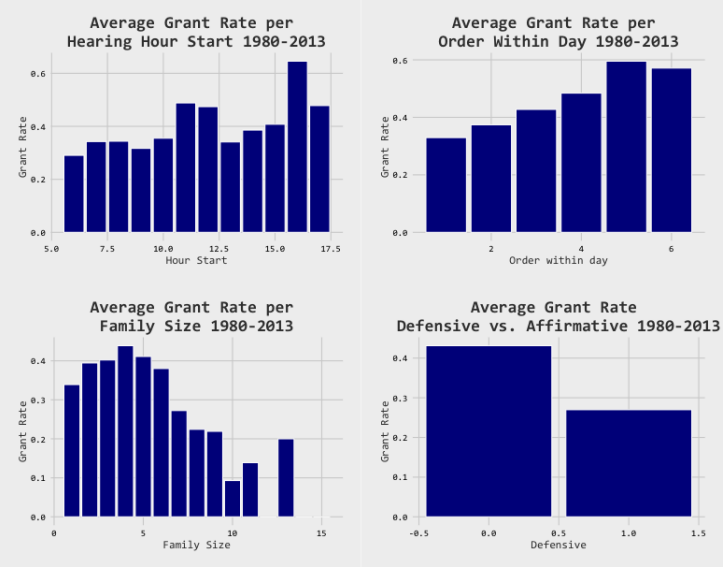
people are closer to indifference among options are more likely to lead to detectable effects outside of it (Simonsohn 2011). For example, if a judge is indifferent, the parabola becomes flatter, and the decisions *less* predictable, holding *fixed* the prediction.

1 Using machine learning to automate the detection of judicial inconsistencies (Application 1)

Let me demonstrate this idea with the asylum courts where I have the administrative universe since 1981. This data comprise half a million asylum decisions across 336 hearing locations and 441 judges. The applicant for asylum reasonably fears imprisonment, torture, or death if forced to return to their home country. The average grant rate is about 35%. Applicants are randomly assigned. Chen et al. [2017b] shows that using data only available up to the decision date, you can achieve 80% predictive accuracy. It is predominately driven by trend features and judge characteristics, W's beyond the applicant control, that might raise questions of due process violations. About one-third is driven by case information, news events, and court information. Then we use only the data available to the case at the opening date and we show that you can achieve 78% accuracy, which raises questions about snap judgments, heuristics, or pre-determined judgments playing a role in how judges decide. In other words, information on the X's may not be factoring into the final outcome of the case, even though one might think that they should.

Figure 2 shows some descriptive statistics. Judges are more lenient before lunch and towards the end of the day. The lower left of Figure 2 shows that there is a U-shape relationship with family size, and the lower right shows that defensive cases are less likely to be granted – defensive cases are those where the applicant has been caught, rather than applying for an extension to stay. Figure 3 shows that judges are more lenient with good weather rather than extreme weather and more lenient with a genocide news indicator. The bottom part shows strong trend factors both within the court on the left and over time on the right. While the literature typically studies one behavioral feature at a time, Chen and Eagel [2017] demonstrates the possibility for machine learning to automate the detection of judicial inconsistencies due to W.

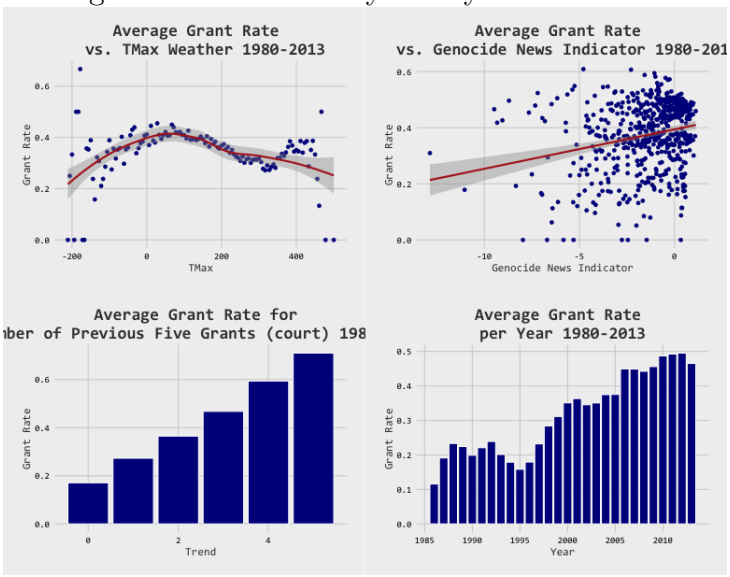
Figure 2: Predictability of Asylum Decisions



More lenient before lunch & towards end of day & for affirmative asylum, U-shape with family size

Source: Chen and Eagel [2017]

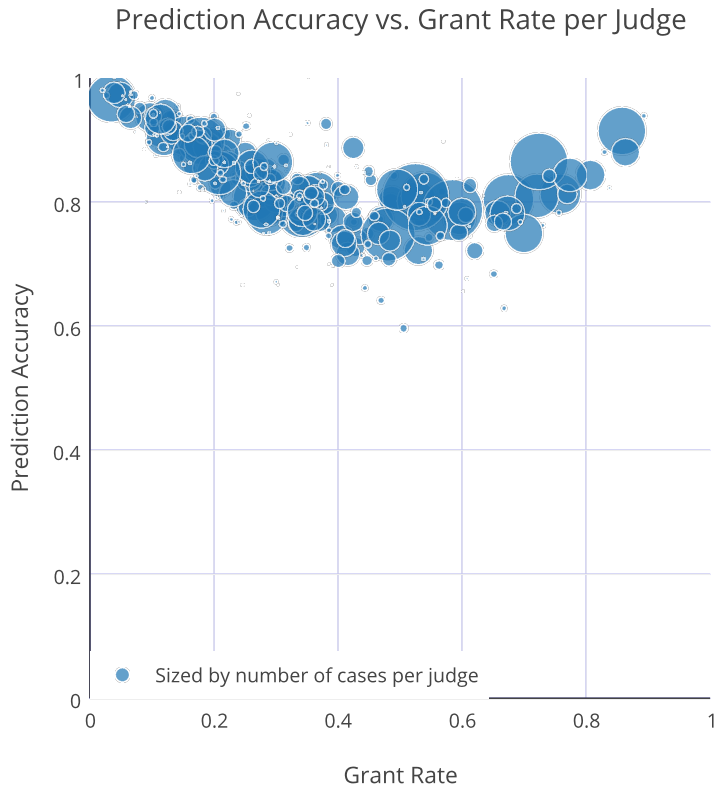
Figure 3: Predictability of Asylum Decisions



More lenient with good weather & genocide news indicator & over time & female judges. Strong trend factors within-court & within-judge.

Source: Chen and Eagel [2017]

Figure 4: Early Predictability of Asylum Decisions

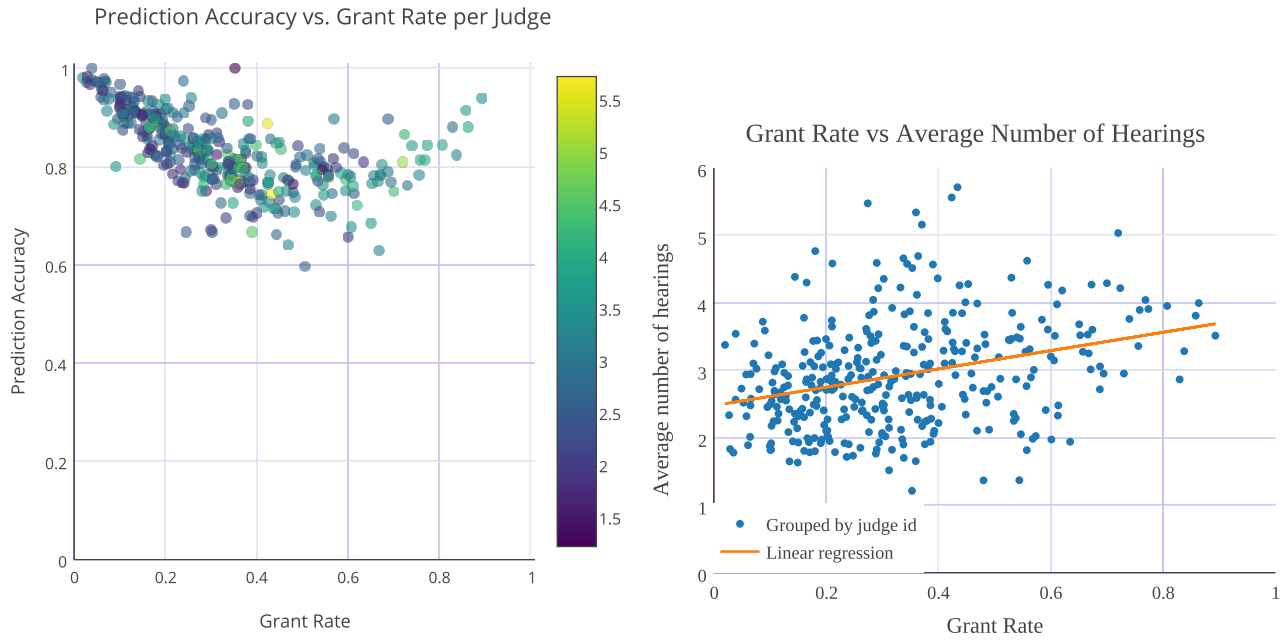


Judges with high and low grant rates are more predictable
Source: Chen et al. [2017b]

2 Using machine learning to detect judicial indifference to case features (Application 2)

Chen et al. [2017b] conceptualizes early predictability, the possibility to use machine learning to automate the detection of judicial indifference, where the judges appear to ignore the circumstances of the case, X. In asylum courts, judges can be predicted with the same level of accuracy at the time the case opens --- and at the time the case closes. To be sure, there may be external circumstances like country war that should dictate the outcome of the case. But significant inter-judge disparities in predictability suggest not all the judges interpret that information in the same way, raising questions of snap or predetermined judgment. Information acquisition can be endogenous to preferences, blurring the boundary between Becker's statistical and taste-based discrimination (Becker 1957).

Figure 5: Early Predictability of Asylum Decisions



Less predictable judges are not simply flipping a coin: hearing sessions are greater for less predictable judges and for judges with higher grant rates
Source: Chen et al. [2017b]

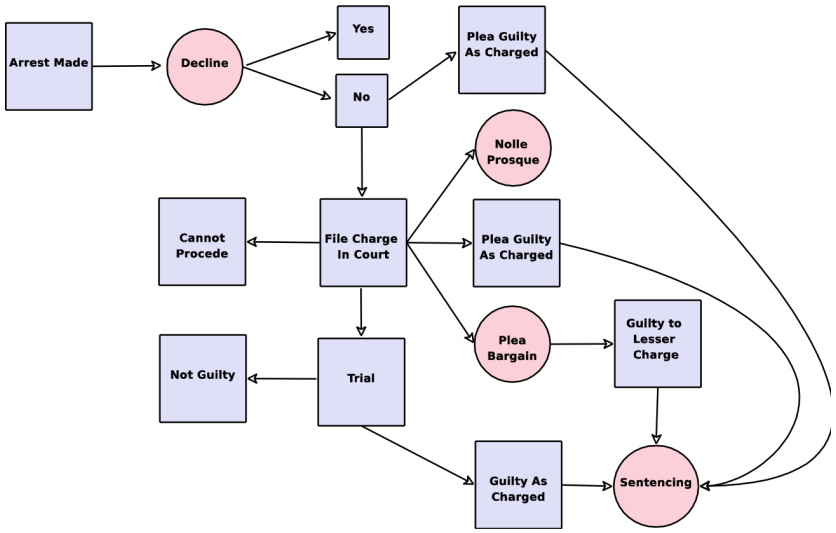
Figure 4 shows that judges with low and high grant rates are more predictable. We might wonder, maybe the judges with a middle-grant rate are simply flipping a coin, but that is not the case. Figure 5 shows that they have more hearing sessions than the judges who rarely grant asylum.

We may also wonder about the judges that are highly predictable with low or high grant rates–maybe both sides are using heuristics equally. But we see that the judges with higher grant rates are having more hearing sessions on average. It seems that these judges are collecting more information to potentially justify their decisions.

3 Using machine learning to detect difference-in-indifference (Application 3)

We see that machine learning can detect W's that affect judicial decisions, though they should not; and we see machine learning detecting judicial indifference, ignoring X's relevant to the case. Next, we see how machine learning can document information

Figure 6: Detailed Information on Arrests Linked to Final Resolution

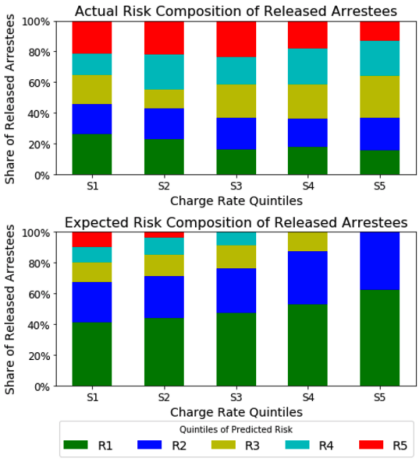


acquisition that varies by a protected class W, like race. Amaranto et al. [2017] use data from New Orleans for 12 years. The data is incredibly detailed - 430,000 charges, 145,000 defendants - and a 594 page codebook including the name, social security number, victims, witnesses, police officers, and so on–a perfect setting for machine learning, with more columns than rows. The entire data collection process begins at the moment of the arrest to the final sentence, if any (Figure 6).

To put this in perspective, other datasets are not linked: they study victimization, or police reports, or arrests, and the many random judge assignment studies examine (and can only examine) the final node only. This data reveals the screening (rejections and dismissals) decision. Why is that important? Amaranto et al. [2017] addresses the broader disparities in criminal justice, an issue of recurring interest in light of Ferguson and Baltimore, and perhaps also Paris and Brussels, where motivations like the perceived legitimacy of the lawmaker have been hotly debated alongside racial differences in the police use of force.

An unexamined issue is what happens after an arrest and before the trial. Information about cases dropped by the prosecutor has been-to date-unavailable---and they are largely unaccountable. The prosecutor is said to decide the fate of 15 cases for every case presided over in trial (Wilson and Petersilia 2010). From 1990-2010, roughly 50% of increase in felony filings comes from misdemeanors being charged as felonies (Pfaff 2017). A recent study argues that the prosecutor charge type can essentially make the racial gap in sentences disappear (Rehavi and Starr 2014). Prosecutors are there-

Figure 7: Risk Distribution of Released Arrestees



Source: Amaranto et al. [2017]

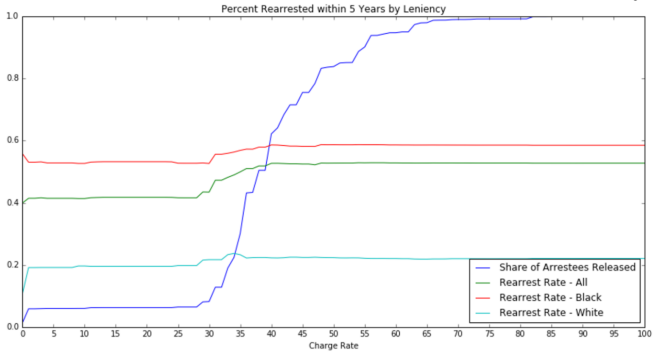
fore very powerful. The data shows the racial gap reappears once prosecutor screening is taken into account. Namely, if Black defendants are less likely to have their case screened out, that mechanically reintroduces the sentencing gap in a very real manner. Cases are randomly or rotationally assigned to screeners. The screening decision is interpretive and discretionary, and recently policy discussions have called for the decision to be made without observation of race.

How the screeners rank the risk of the arrestees is unobserved. However, we can assess their implicit risk ranking by comparing the distribution of predicted risk of the arrestees charged by the "strict" and the "lenient" screeners. That is, we can use machine learning to understand how screeners screen. The actual risk distribution amongst strict and lenient screeners differ from what we would expect to see if the screeners were releasing based on predicted risk.

Figure 7 presents two histograms. In the lower panel, we see what would happen if screeners are targeting whom to release based on risk (in this case, recidivism). The far right bar indicates the strictest screeners who are releasing the fewest arrestees. One might expect the released arrestees to be very "safe" and unlikely to be recidivate. As one moves to the left, the screeners are letting out more and more arrestees. We would expect to see arrestees released to be more likely to recidivate. However, if the screeners were to release defendants at random, then we should expect to see an even distribution of predicted risk for each of the sets of screeners. The top panel shows what actually happens in the data, which looks closer to random.

Figure 8 shows that the slope is essentially flat for Whites and *upwards* sloping for

Figure 8: Risk Distribution of Released Arrestees by Race



Source: Amaranto et al. [2017]

Blacks. Another way to observe judicial indifference would be to create a prediction model of judicial decisions and compare the predicted judge with the actual judge. If the actual judge does better than the predicted judge, we might infer the judge is paying attention to important features of the case not collected in the data. If the actual judge does worse than the predicted judge, we might infer the judge is more susceptible to irrelevant features of the case.

4 Using machine learning to detect implicit bias (Application 4)

We have seen machine learning help identify violations of due process. The final section describes how machine learning can uncover implicit bias in judicial texts or oral arguments (Caliskan et al. 2017), which in turn can be surprisingly predictive and impacts decisions. For this and related work, I have digitized all 380,000 cases and a million judge votes from 1891 in the Circuit Courts. I have engineered 2 billion N-grams of up to length 8 and 5 million citation edges across cases, collated 250 biographical features on 268 judges, and linked this to the 5% random sample[1] of over 400 hand-coded features and 6000 cases hand-coded for meaning in 25 legal areas. I also utilize a data set on millions of criminal sentencing decisions in U.S. District Courts since 1992, linked to judge identity via FOIA-request, and a digital corpus of their opinions since 1923. These data are linked to publicly available Supreme Court datasets, US District docket datasets, geocoded judge seats, biographies of judicial clerks, and administrative data

[1] U.S. Courts of Appeals Database Project (http://www.cas.sc.edu/poli/juri/auburndata.htm)

from Administrative Office of the US Courts (date of key milestones, e.g., oral arguments, when was the last brief filed, etc.). I have also digitized the speech patterns in US Supreme Court oral arguments since 1955 - longitudinal data on speech intonation (linguistic turns) are rare. The data are linked to oral advocates' biographies, faces, clipped identical introductory sentences, and ratings of their traits.

What we can do is to look at the judge's own past writings and see if that correlates with their biographies, and when the judges are randomly assigned, does it impact the decisions? Does it predict sentencing harshness and disparities? Two broad themes emerge. First, textual features broadly predict decisions more than political factors alone, suggesting the importance of text. Second, implicit associations (bias) in judicial texts or oral arguments can be surprisingly predictive.

Ash et al. [2017a] uses the universe of U.S. Circuit Court cases appealed to the Supreme Court since 1946. We show that case similarity among Circuit Court opinions achieves better prediction accuracy of Supreme Court decisions relative to the current best prediction model, which is based on ideology of judges and trends of how they vote. Relative to the benchmark prediction accuracy of 59%, textual measures of case similarity achieve prediction accuracy of 64%. Combining case similarity with ideological features further improves accuracy to 72%, suggesting that ideology affects interpretation of precedent.

Next, Ash et al. [2017b] buttress the importance of text in the development of law in U.S. federal courts. We employ a supervised learning approach to measure the polarization of U.S. Circuit Court judges using the text of court opinions for the years 1930-2013. Our results show persistent but low partisanship of court opinions in the past century, with little evidence of an increase or decrease in partisanship. We visualize voting networks of Circuit Court judges, and we find that judges are not polarized. Third, the mechanism appears partly due to career concerns. We study the behavior of Circuit Court judges during Supreme Court vacancies; judges who were candidates of nomination write fewer discretionary opinions when the Senate is controlled by the opposing political party.

Agrawal et al. [2017] predicts higher court reversals of lower court decisions using the text. Every year more than 300,000 civil and criminal cases are heard in the district courts all over the U.S. Less than 5% of these cases are appealed and heard in circuit courts. For most of the cases, the circuit court either affirms the decision of the district court or reverses it. Out of the cases heard in circuit courts, only about 2% are heard in the Supreme Court. The Supreme Court again can again either affirm or reverse

the circuit court decision. We build a model to predict the higher court decision using the lower court's opinion. Comparing a wide variety of classification, dimensionality reduction, and oversampling techniques, we are able to achieve an accuracy of 79% in Circuit Courts and 68% accuracy in the Supreme Court.

Ash and Chen [2017] show that judges' writings can predict average harshness and racial and sex disparities in sentencing decisions. We document significant reductions in mean square error relative to a naive prediction (the mean of the training data) on the test dataset by approximately 24 percent in predicting punitiveness. In ongoing work, we show that male judges write opinions whose semantics reflect attitudinal disparities in favor of males. We show that judges whose semantics reflect positive attitudes towards government are more likely to vote in favor of government regulation. We compare our empirical approach to orthogonalized machine learning to measure the causal impact of judge's implicit biases on outcomes. We then ask whether judicial attitudes, reflected in court opinions, impact population attitudes. For example, we assess the sentiment of each paragraph on thermometer (sentiment) questions in the American National Election Survey. We calculate cosine similarity of each paragraph to each thermometer target (e.g., Republicans, Democrats, woman, feminists, etc.), and use a sentiment analyzer to compute the average sentiment (positive or negative) of each opinion towards each target. Using LASSO in two-stage least squares for causal inference (Belloni et al. 2012), we identify biographical characteristics predictive of judicial attitudes and show that sentiment reflected in court opinions impact population attitudes.

Implicit bias may also be observed in the oral arguments of the courts. In ongoing work, we show that, even in the Supreme Court, vocal intonation of gendered words (e.g., actor vs. actress) classify vocal intonations of neutral words into stereotypically male (e.g., logical, ability, think) and female (e.g., looking, cook, goodwill), surprisingly suggesting the relevance of people's perceptions of gender being revealed in how people speak. Furthermore, the vocal intonations of judges' speaking these words are predictive of their decisions. These results complement other work indicating that perceptions of gender improve predictions of Supreme Court outcomes and continue to play a role in a manner more complex and nuanced than conventionally perceived (Chen et al. 2017a, 2016a).

Next, we examine whether speech variation beyond word choice, that is, fluctuations in the way one speaks holding the words fixed is predictive of ideology in the U.S. Supreme Court. We use lawyers' campaign donations as a commonly-used measure

of political ideology. We find that audio significantly improves prediction accuracy of ideology relative to using the text alone. AUC increases from 0.55 to 0.61, even in a setting as solemn as the Supreme Court.

5 Conclusion

The legal profession is undergoing a great transformation. The tools of machine learning and causal inference can be used to increase efficiency and fairness of the law. In this article, I discuss how these tools can be used to detect human rights violations.

We have found behavioral factors associated with judge's verdicts. But another element appears very important. By 1990, 40% of federal judges had attended an economics-training program. This law and economics program was founded in 1976 as a two-week training course with lectures by Nobel Prize economists Milton Friedman, Paul Samuelson, and other luminaries. We obtained a list of all the attendees from an article written by a former director. This article also has a number of letters written by judges expressing appreciation. "I can't believe how much I have learned, but I'm glad I didn't have to take this course in college," said one judge. Another wrote, "As a result of what I have learned, I have become a much better judge." Justice Ginsberg wrote, "The instruction is far more intense than the Florida sun."

The results of these seminars were dramatic. We can see economics language used in academic articles become rapidly prevalent in judicial opinions (Ash et al. 2017c). We can see economics trained judges changing how they decided and impacting their peers. We can see economic language traveling from one judge to another and across legal areas. Economics changed how they perceived the consequence of their decisions. Judges shifted their votes by 10% in economics cases. If you teach judges markets work, then they will make decisions that deregulate government. If you teach judges deterrence works, then they will become harsher to criminal defendants. In the district courts, when judges were given discretion in sentencing, economics trained judges immediately rendered 20% longer sentences relative to the non-economics counterparts. Economics judges exacerbate racial and gender disparities in sentencing.

Part of what made the economics training program successful is likely because theory provided structure for judges to understand the patterns they saw. But maybe we can go further. If judges are shown the behavioral findings, will they become less prone to behavioral biases? If judges are taught theoretical structure that drive the behavioral

bias, will they become better judges? Would better justice increase cooperation, trust, recognition and respect?

References

David S. Abrams, Marianne Bertrand, and Sendhil Mullainathan. Do judges vary in their treatment of race? *The Journal of Legal Studies*, 41(2):347 – 383, 2012. URL http://ideas.repec.org/a/ucp/jlstud/doi10.1086-666006.html.

Sharan Agrawal, Elliott Ash, Daniel Chen, Simranjyot Singh Gill, Amanpreet Singh, and Karthik Venkatesan. Affirm or reverse? using machine learning to help judges write opinions. 2017.

Celesta A Albonetti. Sentencing under the federal sentencing guidelines: Effects of defendant characteristics, guilty pleas, and departures on sentence outcomes for drug offenses, 1991-1992. *Law and Society Review*, pages 789–822, 1997.

Daniel Amaranto, Elliott Ash, Daniel L Chen, Lisa Ren, and Caroline Roper. Algorithms as prosecutors: Lowering rearrest rates without disparate impacts and identifying defendant characteristics ânoisyâto human decision-makers. 2017.

Elliott Ash and Daniel Chen. Predicting punitiveness from judicial corpora. 2017.

Elliott Ash, Daniel Chen, Shivendra Panicker, and Akshay Trivedi. Precedent vs. politics? case similarity predicts supreme court decisions better than ideology. 2017a.

Elliott Ash, Daniel L Chen, and Wei Lu. The (non-) polarization of us circuit court judges, 1930-2013. 2017b.

Elliott Ash, Daniel L Chen, and Suresh Naidu. Ideas have consequences: The impact of law and economics on american justice. Technical report, working paper, 2017c.

Nora Barry, Laura Buchanan, Evelina Bakhturina, and Daniel L. Chen. Events Unrelated to Crime Predict Criminal Sentence Length. Technical report, 2016.

Gary S Becker. *Economics of Discrimination*. University of Chicago Press, 1957. URL http://press.uchicago.edu/ucp/books/book/chicago/E/bo3630686.html.

Alex Belloni, Daniel L. Chen, Victor Chernozhukov, and Chris Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, November 2012. ISSN 00129682. doi: 10.3982/ECTA9626. URL http://www.jstor.org/stable/23357231.

Carlos Berdejo and Daniel L. Chen. Electoral Cycles Among U.S. Courts of Appeals Judges. Technical report, 2016.

Christina Boyd, Lee Epstein, and Andrew D. Martin. Untangling the causal effects of sex on judging. *American Journal of Political Science*, 54(2):389–411, 2010. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract\_id=1001748`.

Shawn D Bushway and Anne Morrison Piehl. Judging judicial discretion: Legal factors and racial discrimination in sentencing. *Law and Society Review*, pages 733–764, 2001.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Daniel Chen. Mood and the malleability of moral reasoning. 2017a.

Daniel Chen and Arnaud Philippe. Reference points, mental accounting, and social preferences: Sentencing leniency on birthdays. Technical report, mimeo, 2017.

Daniel Chen, Yosh Halberstam, and CL Alan. Perceived masculinity predicts us supreme court outcomes. *PloS one*, 11(10):e0164324, 2016a.

Daniel Chen, Yosh Halberstam, and Alan Yu. Covering: Mutable Characteristics and Perceptions of Voice in the U.S. Supreme Court. *Review of Economic Studies*, 2017a. URL `http://nber.org/~dlchen/papers/Covering.pdf`. invited to resubmit, TSE Working Paper No. 16-680.

Daniel L. Chen. Priming Ideology: Why Presidential Elections Affect U.S. Courts of Appeals Judges. Technical report, 2016a.

Daniel L Chen. Implicit egoism in sentencing decisions: First letter name effects with randomly assigned defendants. 2016b.

Daniel L Chen. Tastes for desert and placation: A reference point-dependent model of social preferences. 2017b.

Daniel L. Chen and Jess Eagel. Can Machine Learning Help Predict the Outcome of Asylum Adjudications? Technical report, 2016.

Daniel L. Chen and Jess Eagel. Can Machine Learning Help Predict the Outcome of Asylum Adjudications? *Artificial Intelligence and the Law*, March 2017. URL https://www.tse-fr.eu/publications/can-machine-learning-help-predict-outcome-asylum-adjudications. Accepted at ICAIL, TSE Working Paper No. 17-782.

Daniel L. Chen, Tobias J. Moskowitz, and Kelly Shue. Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires. Working paper, ETH Zurich, March 2015.

Daniel L. Chen, Xing Cui, Lanyu Shang, and Jing Zhang. What Matters: Agreement Between U.S. Courts of Appeals Judges. Technical report, 2016b.

Daniel L. Chen, Damian Kozbur, and Alan Yu. Mimicry: Phonetic Accommodation Predicts U.S. Supreme Court Votes. Working paper, ETH Zurich, 2016c.

Daniel L. Chen, Matt Dunn, Levent Sagun, and Hale Sirin. Early Predictability of Asylum Court Decisions. *Artificial Intelligence and the Law*, March 2017b. URL https://www.tse-fr.eu/publications/early-predictability-asylum-court-decisions. Accepted at ICAIL, TSE Working Paper No. 17-781.

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

Ozkan Eren and Naci Mocan. Emotional judges and unlucky juveniles. Working paper, 2016.

Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. Inside the judicial mind. *Cornell Law Review*, 86(4):777–830, 2000. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=257634.

Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. Blinking on the bench: How judges decide cases. *Cornell Law Review*, 93(1):1–44, 2007. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1026414.

John A Humphrey and Timothy J Fogarty. Race and plea bargained outcomes: A research note. *Social Forces*, 66(1):176–182, 1987.

Benjamin Klein, Robert G. Crawford, and Armen A. Alchian. Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21(2):297–326, October 1978. ISSN 0022-2186. URL http://www.jstor.org/stable/725234.

David B Mustard. Racial, ethnic, and gender disparities in sentencing: Evidence from the us federal courts. *Journal of Law and Economics*, 44(1):285–314, 2001.

John F Pfaff. *Locked In.* Basic Books, 2017.

Jeffrey J. Rachlinski, Sheri Lynn Johnson, Andrew J. Wistrich, and Chris Guthrie. Does Unconscious Racial Bias Affect Trial Judges? *Notre Dame Law Review*, 84: 1195–1246, 2009. URL http://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=1691&context=facpub.

Jeffrey J. Rachlinski, Andrew J. Wistrich, and Chris Guthrie. Altering Attention in Adjudication. *UCLA Law Review*, 60:1586–1618, 2013. URL http://www.uclalawreview.org/pdf/60-6-6.pdf.

M Marit Rehavi and Sonja B Starr. Racial disparity in federal criminal sentences. *Journal of Political Economy*, 122(6):1320–1354, 2014.

Max Schanzenbach. Racial and sex disparities in prison sentences: the effect of district-level judicial demographics. *The Journal of Legal Studies*, 34(1):57–92, 2005.

Moses Shayo and Asaf Zussman. Judicial ingroup bias in the shadow of terrorism. *The Quarterly Journal of Economics*, 126(3):1447–1484, 2011. ISSN 00335533. URL http://www.jstor.org/stable/23015705.

Dan Simon. *In Doubt: The Psychology of the Criminal Justice Process.* Harvard University Press, Cambridge, MA, 2012. URL http://books.google.com/books?id=0OgsbJwZ5eEC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

Uri Simonsohn. Spurious? name similarity effects (implicit egotism) in marriage, job, and moving decisions. *Journal of personality and social psychology*, 101(1):1, 2011.

Darrell Steffensmeier and Stephen Demuth. Ethnicity and sentencing outcomes in us federal courts: Who is punished more harshly? *American sociological review*, pages 705–729, 2000.

Randall J Thomson and Matthew T Zingraff. Detecting sentencing disparity: Some problems and evidence. *American Journal of Sociology*, pages 869–880, 1981.

James Q Wilson and Joan Petersilia. *Crime and public policy*. Oxford University Press, 2010.