Using patterns in judicial data to identify bias in decision making

Daniel L. Chen

Judicial Analytics and Law *J of Artificial Intelligence & Law 2018*

Justice: **equal treatment before the law** $(y = f(X) + \varepsilon, a \rightarrow X)$
equality based on recognition of difference
$(y \perp W, var(\varepsilon) \perp W, a \nrightarrow W)$

control principle and **merit** principle: individuals responsible only for events that are under their control
W: race, gender, masculinity, name, football, weather, judge's lunchtime, preceding case, ...

Judicial Analytics and Law *J of Artificial Intelligence & Law 2018*

Justice: **equal treatment before the law** $(y = f(X) + \varepsilon, a \to X)$
equality based on recognition of difference
$(y \perp W, var(\varepsilon) \perp W, a \nrightarrow W)$

control principle and **merit** principle: individuals responsible only for events that are under their control
W: race, gender, masculinity, name, football, weather, judge's lunchtime, preceding case, ...

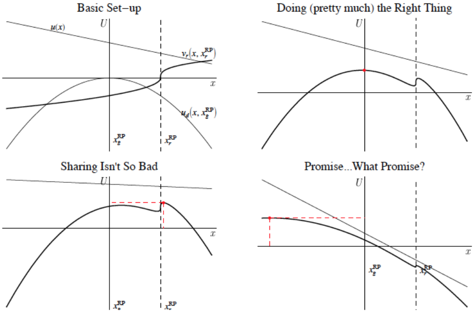
Judicial Analytics and Law *J of Artificial Intelligence & Law 2018*

Justice: **equal treatment before the law** $(y = f(X) + \varepsilon, a \rightarrow X)$
equality based on recognition of difference
$(y \perp W, var(\varepsilon) \perp W, a \nrightarrow W)$

control principle and **merit** principle: individuals responsible only for events that are under their control
W: race, gender, masculinity, name, football, weather, judge's lunchtime, preceding case, ...

Machine Learning and Rule of Law *Computational Analysis of Law* 2018

- Behavioral anomalies offer intuitive understanding of feature relevance
- "settings where people are closer to *indifference* among options are more likely to lead to detectable effects [of behavioral biases] outside of it." (Simonsohn, JPSP 2011)



A model of recognition-respect and

revealed preference indifference

Research in Experimental Economics 2017

Natural Laboratory to Study Normative Judgments

U.S. Circuit Courts

- All 380K cases, 1M judge votes, from 1891-
- 2B 8-grams, 5M citation edges across cases

U.S. District Courts

- 1M criminal sentencing decisions
- 2.5M opinions from 1923-

U.S. Supreme Court

- Speech patterns in oral arguments from 1955-
- Identical introductory sentences

U.S. Immigration Courts
Prosecutors
WW1 Courts martials
Chile, India, Kenya

Natural Laboratory to Study Normative Judgments

U.S. Circuit Courts

- All 380K cases, 1M judge votes, from 1891-
- 2B 8-grams, 5M citation edges across cases

U.S. District Courts

- 1M criminal sentencing decisions
- 2.5M opinions from 1923-

U.S. Supreme Court

- Speech patterns in oral arguments from 1955-
- Identical introductory sentences

U.S. Immigration Courts

Prosecutors

WW1 Courts martials

Chile, India, Kenya

Natural Laboratory to Study Normative Judgments

U.S. Circuit Courts

- All 380K cases, 1M judge votes, from 1891-
- 2B 8-grams, 5M citation edges across cases

U.S. District Courts

- 1M criminal sentencing decisions
- 2.5M opinions from 1923-

U.S. Supreme Court

- Speech patterns in oral arguments from 1955-
- Identical introductory sentences

U.S. Immigration Courts
Prosecutors
WW1 Courts martials
Chile, India, Kenya

Natural Laboratory to Study Normative Judgments

U.S. Circuit Courts

- All 380K cases, 1M judge votes, from 1891-
- 2B 8-grams, 5M citation edges across cases

U.S. District Courts

- 1M criminal sentencing decisions
- 2.5M opinions from 1923-

U.S. Supreme Court

- Speech patterns in oral arguments from 1955-
- Identical introductory sentences

U.S. Immigration Courts

Prosecutors

WW1 Courts martials

Chile, India, Kenya

Natural Laboratory to Study Normative Judgments

U.S. Circuit Courts

- All 380K cases, 1M judge votes, from 1891-
- 2B 8-grams, 5M citation edges across cases

U.S. District Courts

- 1M criminal sentencing decisions
- 2.5M opinions from 1923-

U.S. Supreme Court

- Speech patterns in oral arguments from 1955-
- Identical introductory sentences

U.S. Immigration Courts

Prosecutors

WW1 Courts martials

Chile, India, Kenya

Natural Laboratory to Study Normative Judgments

U.S. Circuit Courts

- All 380K cases, 1M judge votes, from 1891-
- 2B 8-grams, 5M citation edges across cases

U.S. District Courts

- 1M criminal sentencing decisions
- 2.5M opinions from 1923-

U.S. Supreme Court

- Speech patterns in oral arguments from 1955-
- Identical introductory sentences

U.S. Immigration Courts

Prosecutors

WW1 Courts martials

Chile, India, Kenya

Natural Laboratory to Study Normative Judgments

U.S. Circuit Courts

- All 380K cases, 1M judge votes, from 1891-
- 2B 8-grams, 5M citation edges across cases

U.S. District Courts

- 1M criminal sentencing decisions
- 2.5M opinions from 1923-

U.S. Supreme Court

- Speech patterns in oral arguments from 1955-
- Identical introductory sentences

U.S. Immigration Courts

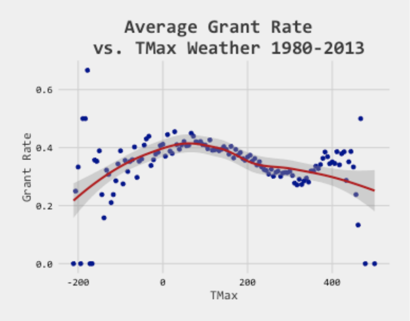
Prosecutors

WW1 Courts martials

Chile, India, Kenya

The weather

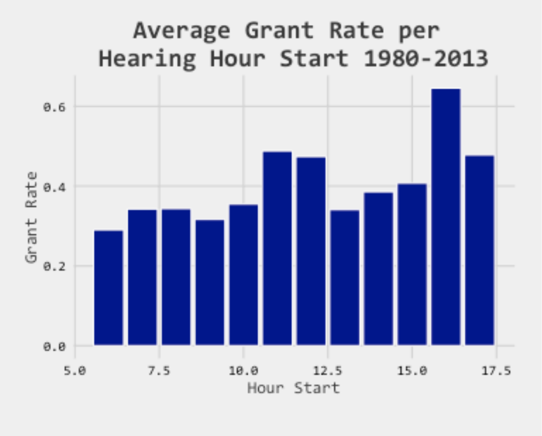
Judges deny refugees asylum when the weather is too hot or too cold



Chen and Eagel, ICAIL 2017

Time of Day

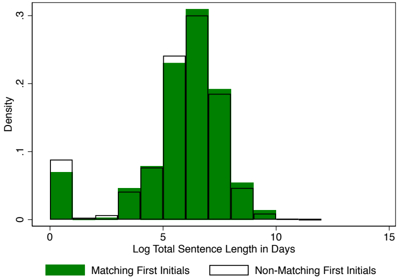
They grant asylum more before lunch and less after.



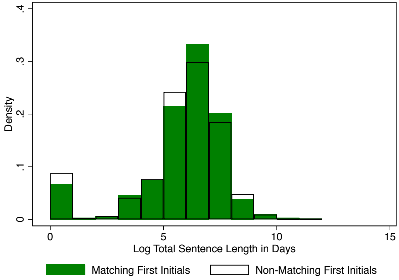
1M decisions

The defendant's name

They assign longer sentence lengths to defendants whose first initial matches their own.



First Letter of First Name



First Letter of Last Name

The defendant's birthday

When they do the opposite and give the gift of leniency

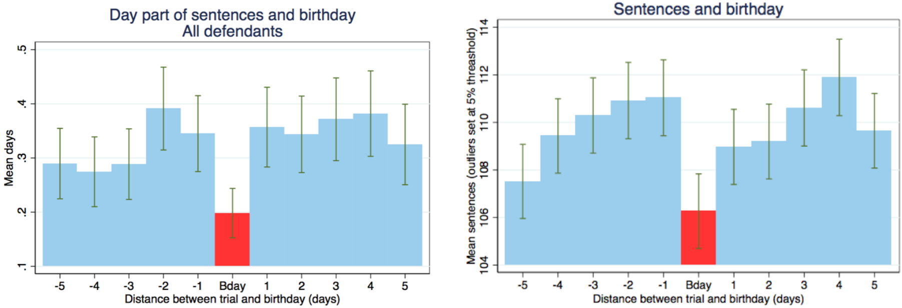
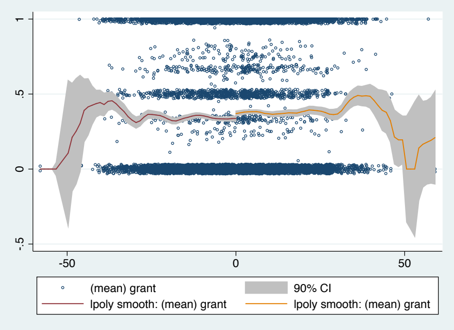


Figure: US and French judicial leniency on defendant birthdays

Chen and Philippe, JEBO R&R

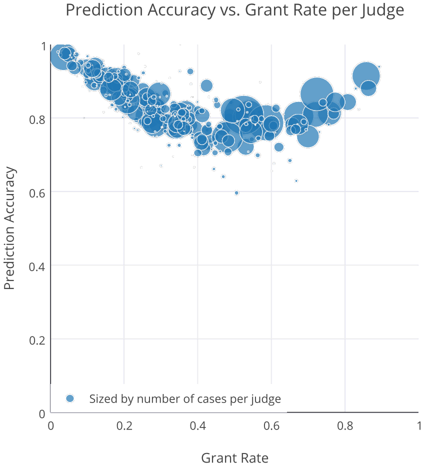
NFL Football

Judges are more lenient the day after their team wins, rather than loses.



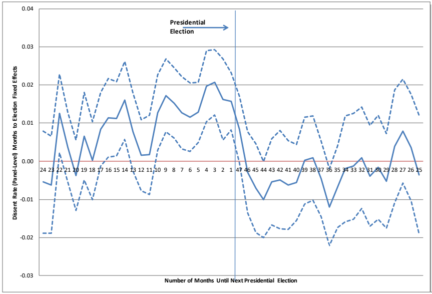
Snap judgments

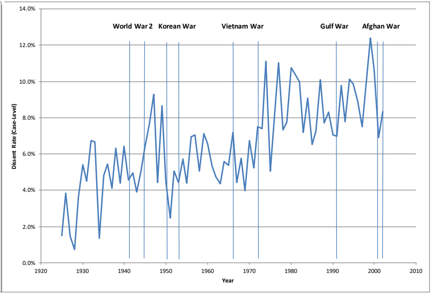
We can use machine learning to predict asylum decisions with 80% accuracy the date the case opens.. and when it closes.



Dunn, Sagun, Sirin, and Chen, ICAIL **2017**

Elections and wartime also affect decisions





Berdejo and Chen, JLE 2017

Gambler's Fallacy

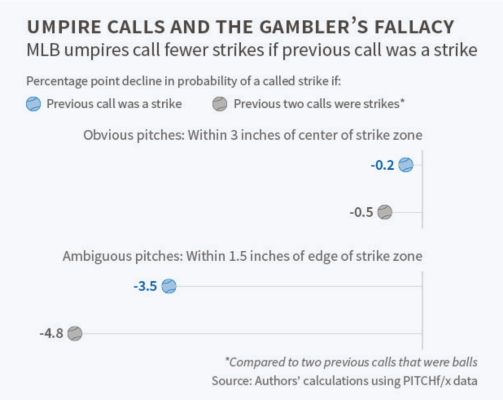
How people often imagine a sequence of coin flips:

0101001011001010100110100

A real sequence of coin flips:

0101011111011000001001101

Up to 5% of decisions reversed due to the gambler's fallacy



Chen, Moskowitz, and Shue, QJE 2016

In the US Supreme Court, the first sentence of the lawyers oral arguments are identical

Recording 1 of 66

0:00

1. Please provide your impression of the voice recording in the matrix below:

Very Attractive	○	○	○	○	○	○	○	Very Unattractive
Very Masculine	○	○	○	○	○	○	○	Not At All Masculine
Not Intelligent	○	○	○	○	○	○	○	Intelligent
Very Unaggressive	○	○	○	○	○	○	○	Very Aggressive
Not Trustworthy	○	○	○	○	○	○	○	Trustworthy
Very Confident	○	○	○	○	○	○	○	Very Timid

2. Assuming that this is a lawyer arguing a case in front of a panel of judges, how likely do you think this lawyer will win the case?

Will Definitely Lose ○--○--○--○--○--○--○ Will Definitely Win

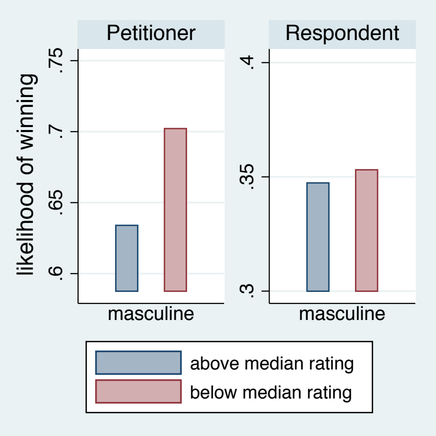
3. How good is the quality of the recording?

Very Bad ○--○--○--○--○--○--○ Very Good

Next

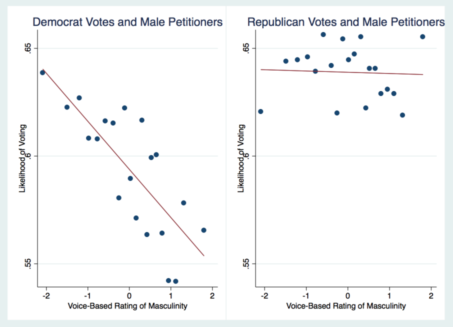
"Mr. Chief Justice, (and) may it please the Court?"

Male petitioners below median in masculinity rating are 7 percentage points more likely to win



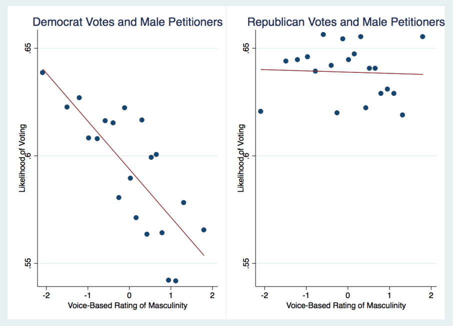
Chen, Halberstam, and Yu, Plos-ONE 2016

Democrats vote against masculine-sounding lawyers



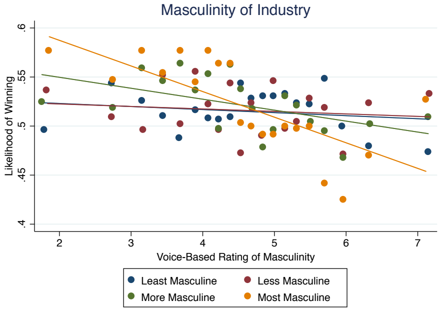
Profit-maximizing firms would tend to erode this correlation

Democrats vote against masculine-sounding lawyers



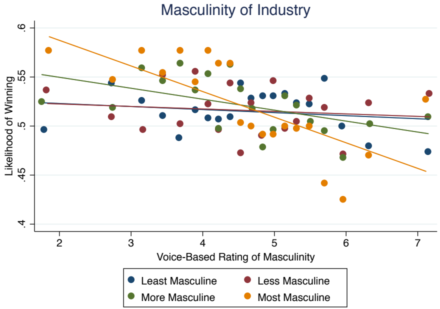
Profit-maximizing firms would tend to erode this correlation

Negative correlation is stronger in more masculine industries



consistent with their perceiving masculine-sounding lawyers as winners

Negative correlation is stronger in more masculine industries



consistent with their perceiving masculine-sounding lawyers as winners

De-Biasing Experiment Reduces Misbeliefs

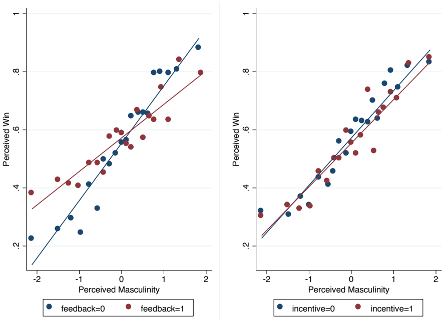


Figure: Feedback ($p < 0.01$), Incentives

Incentives Further Erodes Misbeliefs

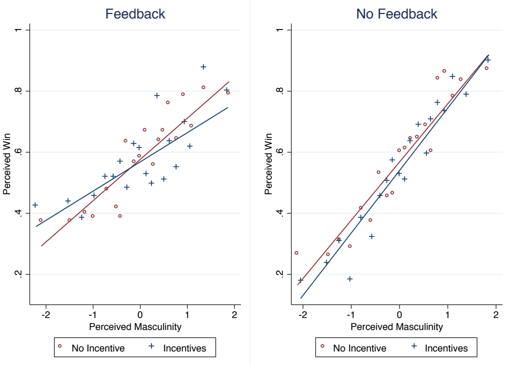


Figure: Incentives ($p < 0.05$) with Feedback

identifying a taste for masculine-sounding lawyers

Incentives Further Erodes Misbeliefs

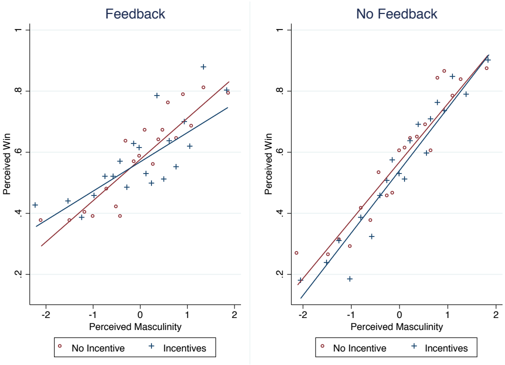


Figure: Incentives ($p < 0.05$) with Feedback

identifying a taste for masculine-sounding lawyers

Gender

- Female lawyers are also coached to be more masculine (Starecheski 2014)
  - Are our findings restricted to male advocates alone or do they extend?

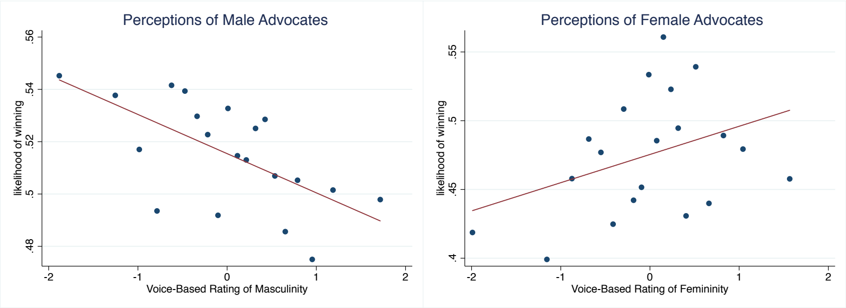


Figure: Extends: Less masculine males and more feminine females ↑win

- If masculine = - feminine, pooled results would be stronger

Robust to Lawyer Heterogeneity and the Best ML Prediction of the Supreme Court

	Judge Votes for Lawyer					
Predicted Vote	0.257***		0.258***	0.250***		0.248***
from Random Forest	(0.0486)		(0.0487)	(0.0485)		(0.0489)
Masculine		-0.0223**	-0.0207**		-0.0852**	-0.0780**
		(0.0101)	(0.0101)		(0.0359)	(0.0361)
Cluster	Lawyer and Judge					
Collapsed	No	No	No	Yes	Yes	Yes
Observations	26447	26391	26391	1229	1229	1229
R-squared	0.061	0.002	0.063	0.058	0.008	0.064
Sample: Male Petitioners, Democrat Judges						

Figure: Best Prediction and Perceived Masculinity

- Random forest also selects perceptions

Besides voice, there is text

- Females: Migraine, hysterical, morbid, obese, terrified, unemancipated, battered
- Males: Reserve, industrial, honorable, commanding, conscientious, duty

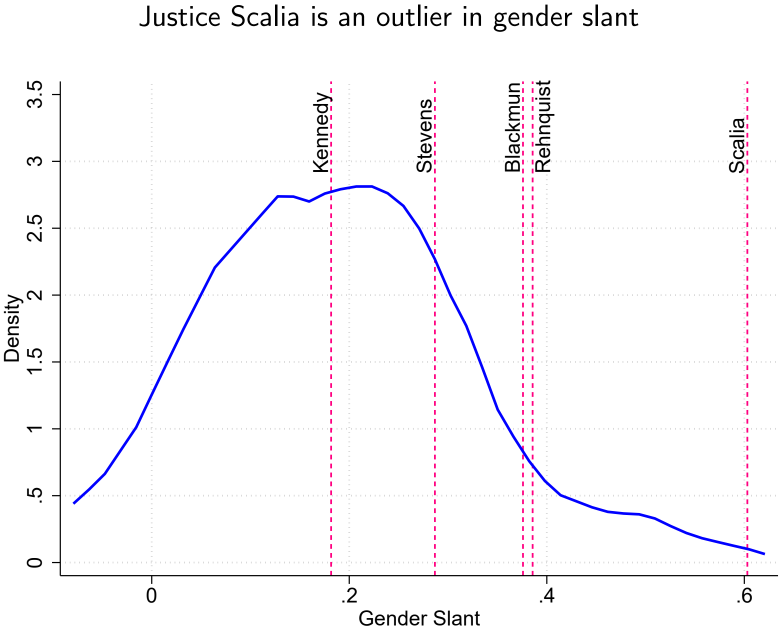
Besides voice, there is text

- Females: Migraine, hysterical, morbid, obese, terrified, unemancipated, battered
- Males: Reserve, industrial, honorable, commanding, conscientious, duty

Besides voice, there is text

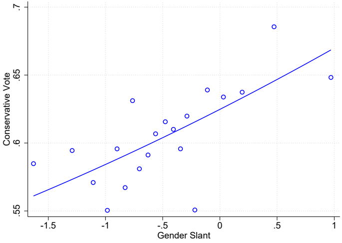
- Females: Migraine, hysterical, morbid, obese, terrified, unemancipated, battered
- Males: Reserve, industrial, honorable, commanding, conscientious, duty

We can do this judge by judge

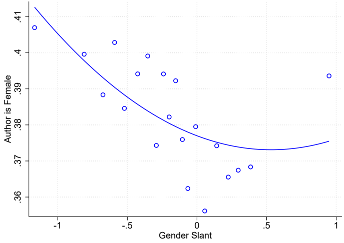


In the Circuit Courts, judges with more gender slant..

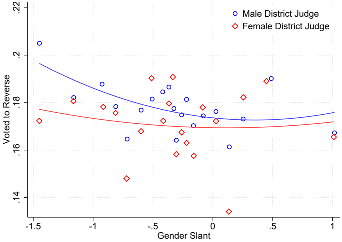
Vote against women's rights issues



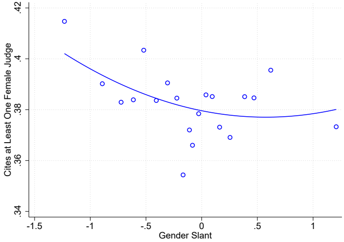
Assign fewer opinions for females to author



Reverse male judges less often



Cite female judges less often



Ash, Chen, and Ornaghi, AEJ: Applied 2022

Five Ways for ML to Diagnose Judicial Inattention

1. Early predictability
2. Behavioral anomalies
3. Inattentiveness to appellate reversals
4. Implicit risk rankings of asylees closer to random
5. Is indifference greater for some refugees (e.g., from Global South)?

Five Ways for ML to Diagnose Judicial Inattention

1. Early predictability
2. Behavioral anomalies
3. Inattentiveness to appellate reversals
4. Implicit risk rankings of asylees closer to random
5. Is indifference greater for some refugees (e.g., from Global South)?

Five Ways for ML to Diagnose Judicial Inattention

1. Early predictability
2. Behavioral anomalies
3. Inattentiveness to appellate reversals
4. Implicit risk rankings of asylees closer to random
5. Is indifference greater for some refugees (e.g., from Global South)?

Five Ways for ML to Diagnose Judicial Inattention

1. Early predictability
2. Behavioral anomalies
3. Inattentiveness to appellate reversals
4. Implicit risk rankings of asylees closer to random
5. Is indifference greater for some refugees (e.g., from Global South)?

Five Ways for ML to Diagnose Judicial Inattention

1. Early predictability
2. Behavioral anomalies
3. Inattentiveness to appellate reversals
4. Implicit risk rankings of asylees closer to random
5. Is indifference greater for some refugees (e.g., from Global South)?

Five Ways for ML to Diagnose Judicial Inattention

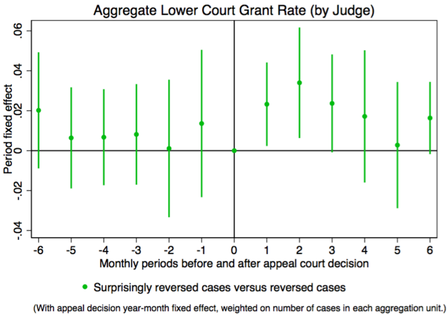
1. Early predictability
2. Behavioral anomalies
3. Inattentiveness to appellate reversals
4. Implicit risk rankings of asylees closer to random
5. Is indifference greater for some refugees (e.g., from Global South)?

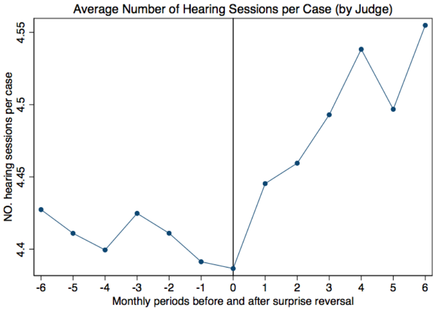
Five Ways for ML to Diagnose Judicial Inattention

1. Early predictability
2. Behavioral anomalies
3. Inattentiveness to appellate reversals
4. Implicit risk rankings of asylees closer to random
5. Is indifference greater for some refugees (e.g., from Global South)?

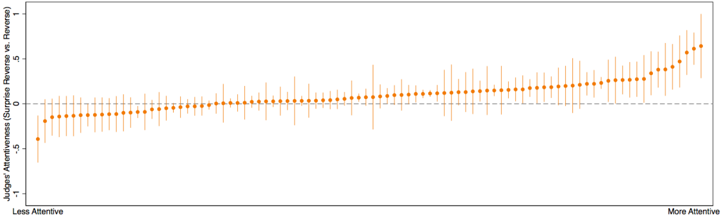
After "Surprise" Reversals, Judges Grant More Asylum and Hold More Hearing Sessions

Surprise Reversal is a reversal of a decision that was predicted to be "Affirm"





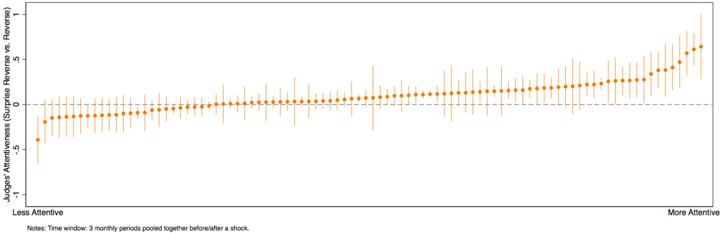
Judges Vary in Responsiveness to Reversal



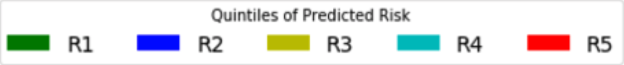
Notes: Time window: 3 monthly periods pooled together before/after a shock.

Do less attentive judges have implicit risk rankings closer to random?

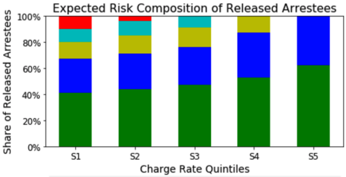
Judges Vary in Responsiveness to Reversal



Do less attentive judges have implicit risk rankings closer to random?

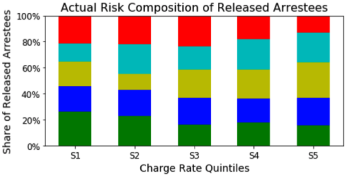


Robot Prosecutors



- If defendants released based only on risk score, the harshest prosecutors would only be releasing low-risk defendants.

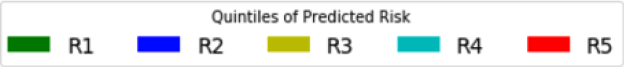
Human Prosecutors



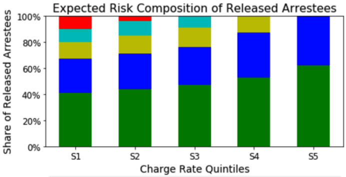
- Distribution of risk scores for released defendants is similar for most lenient and least lenient prosecutors.

- Are the lenient asylum judges, only denying the 'riskiest' applicants
  - i.e., seeing the lowest reversal rates (of their asylum denials)?

See also *Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan, QJE 2017*

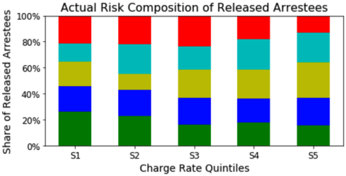


Robot Prosecutors



- If defendants released based only on risk score, the harshest prosecutors would only be releasing low-risk defendants.

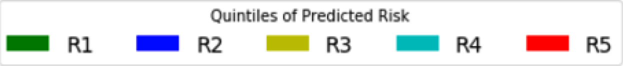
Human Prosecutors



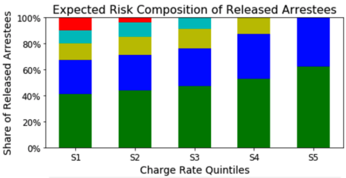
- Distribution of risk scores for released defendants is similar for most lenient and least lenient prosecutors.

- Are the lenient asylum judges, only denying the 'riskiest' applicants
  - i.e., seeing the lowest reversal rates (of their asylum denials)?

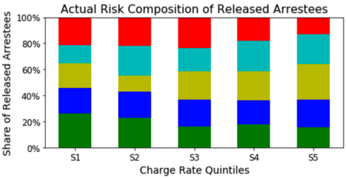
See also *Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan, QJE 2017*



Robot Prosecutors



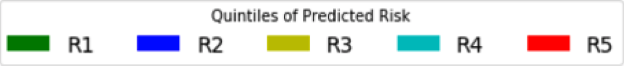
Human Prosecutors



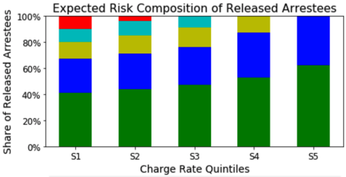
- If defendants released based only on risk score, the harshest prosecutors would only be releasing low-risk defendants.
- Distribution of risk scores for released defendants is similar for most lenient and least lenient prosecutors.

- Are the lenient asylum judges, only denying the 'riskiest' applicants
  - i.e., seeing the lowest reversal rates (of their asylum denials)?

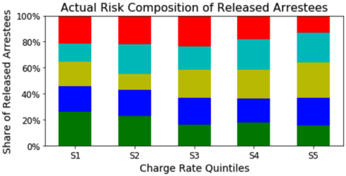
See also *Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan, QJE 2017*



Robot Prosecutors



Human Prosecutors

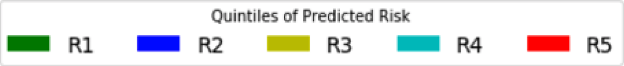


- If defendants released based only on risk score, the harshest prosecutors would only be releasing low-risk defendants.

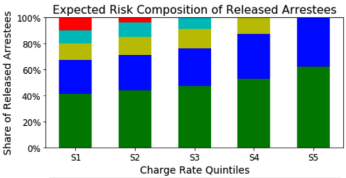
- Distribution of risk scores for released defendants is similar for most lenient and least lenient prosecutors.

- Are the lenient asylum judges, only denying the 'riskiest' applicants
  - i.e., seeing the lowest reversal rates (of their asylum denials)?

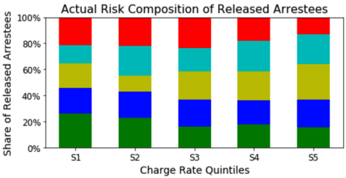
See also *Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan, QJE 2017*



Robot Prosecutors



Human Prosecutors



- If defendants released based only on risk score, the harshest prosecutors would only be releasing low-risk defendants.

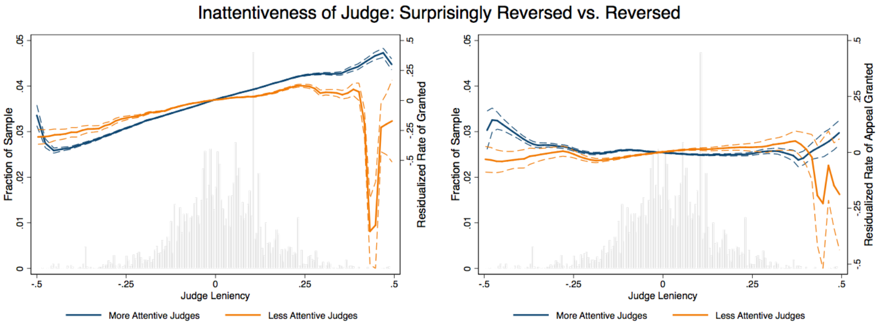
- Distribution of risk scores for released defendants is similar for most lenient and least lenient prosecutors.

- Are the lenient asylum judges, only denying the 'riskiest' applicants
  - i.e., seeing the lowest reversal rates (of their asylum denials)?

See also *Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan, QJE 2017*

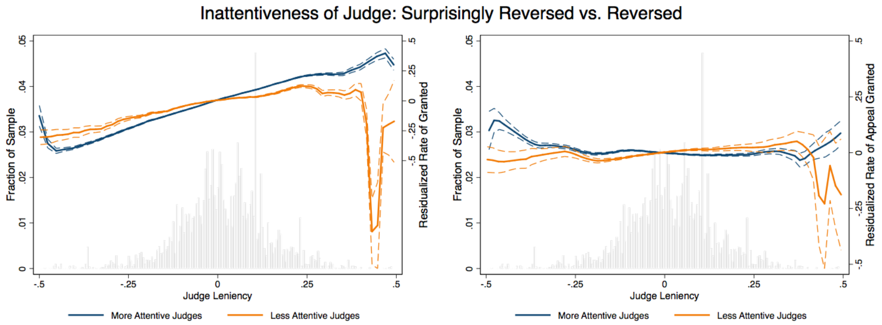
Left Figure: Judges have strong habits

A judge who is generally lenient in other cases is likely to be lenient in a given case



(Time window: 3 monthly periods pooled together before/after shock. More attentiveness: the coefficient of interaction of surprisingly reversed dummy and time-period dummy is bigger)

Right Figure: Assess implicit risk ranking

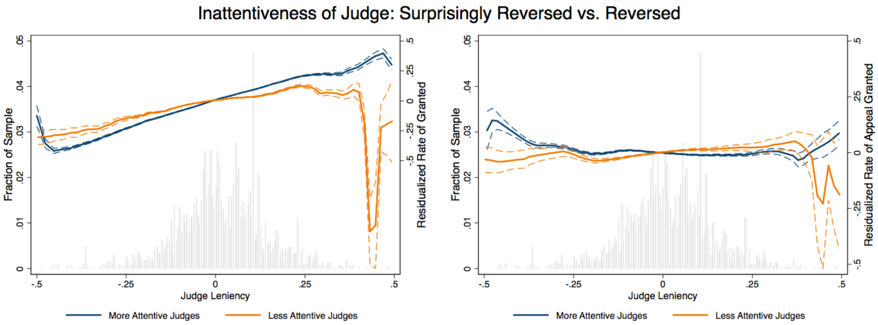


(Time window: 3 monthly periods pooled together before/after shock. More attentiveness: the coefficient of interaction of surprisingly reversed dummy and time-period dummy is bigger)

If judges are 'ordering' their asylees, the most lenient judge letting in the most applicants should be rejecting only the "least safe" applicants

Their appeal success should be lower, which we see among more attentive judges

Right Figure: Assess implicit risk ranking

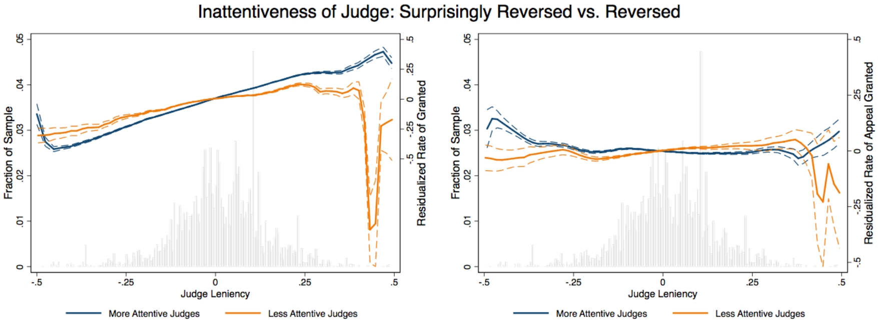


(Time window: 3 monthly periods pooled together before/after shock. More attentiveness: the coefficient of interaction of surprisingly reversed dummy and time-period dummy is bigger)

If judges are 'ordering' their asylees, the most lenient judge letting in the most applicants should be rejecting only the "least safe" applicants

Their appeal success should be lower, which we see among more attentive judges

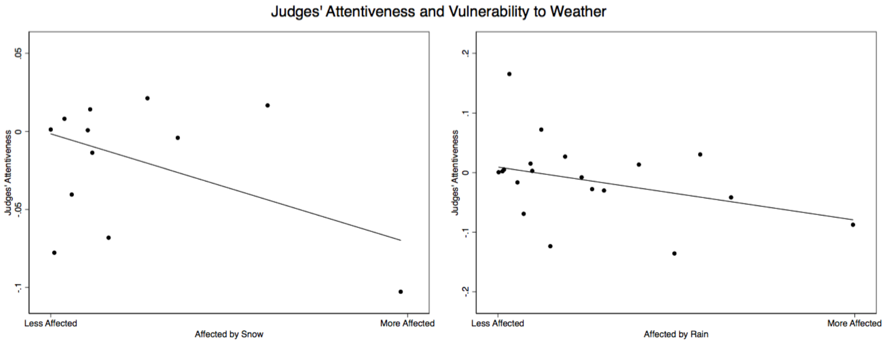
.. but not less attentive judges



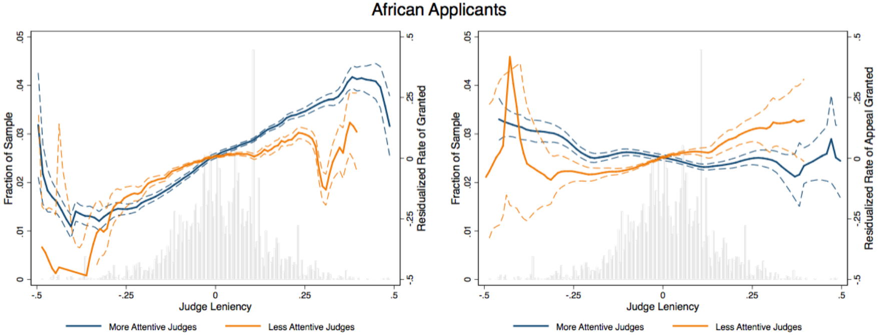
(Time window: 3 monthly periods pooled together before/after shock. More attentiveness: the coefficient of interaction of surprisingly reversed dummy and time-period dummy is bigger)

.. who may be more prone to other extraneous factors

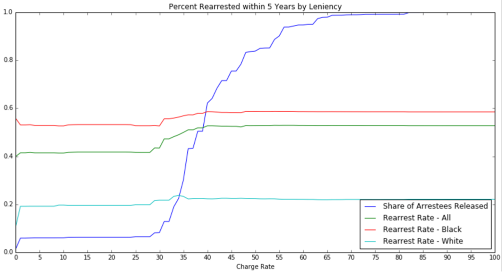
.. such as weather

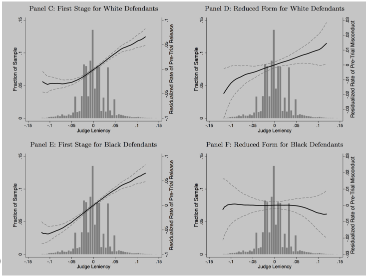


Difference in Indifference for asylees from the Global South



Using ML to Understand how Screeners Screen





Actually, flat for Whites, *upward* slope for Blacks (left)

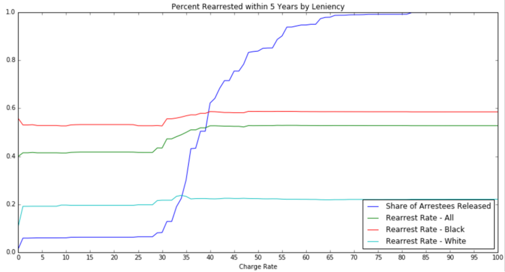
Algorithms as Prosecutors: Identifying Characteristics Noisy to Human Prosecutors

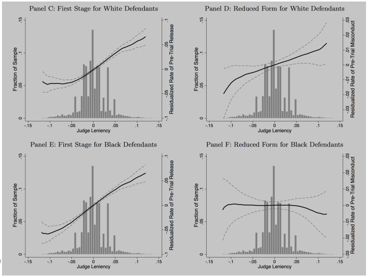
- Judges released along "right" diagonal for Whites but not Blacks (right)

in *Arnold, Dobbie, Yang, QJE 2017*

WHY "WRONG DIAGONAL" FOR BLACK DEFENDANTS?

Using ML to Understand how Screeners Screen





Actually, flat for Whites, *upward* slope for Blacks (left)

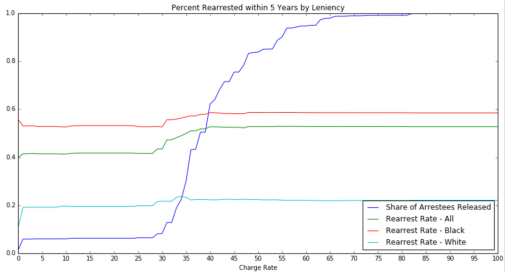
Algorithms as Prosecutors: Identifying Characteristics Noisy to Human Prosecutors

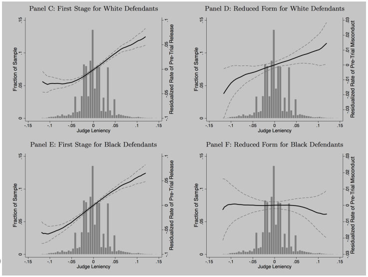
- Judges released along "right" diagonal for Whites but not Blacks (right)

in ***Arnold, Dobbie, Yang, QJE 2017***

Why "wrong diagonal" for Black defendants?

Using ML to Understand how Screeners Screen





Actually, flat for Whites, *upward* slope for Blacks (left)

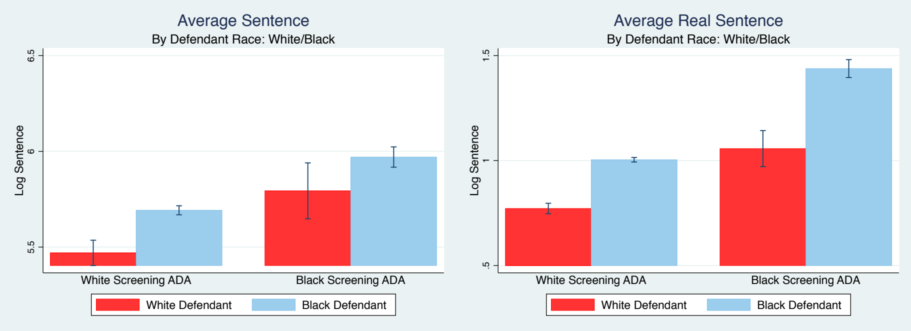
Algorithms as Prosecutors: Identifying Characteristics Noisy to Human Prosecutors

- Judges released along "right" diagonal for Whites but not Blacks (right)

in ***Arnold, Dobbie, Yang, QJE 2017***

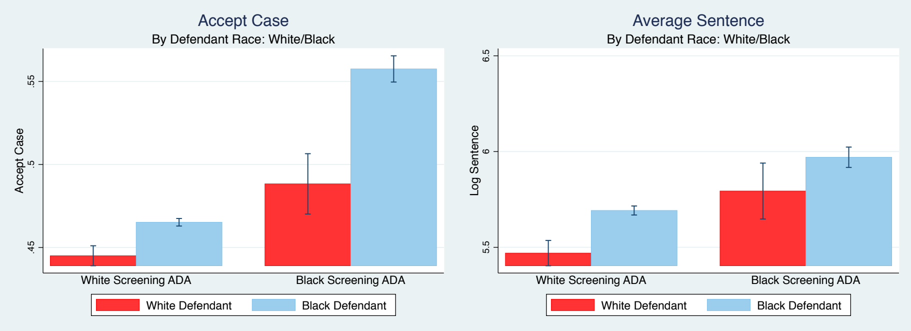
Why "wrong diagonal" for Black defendants?

1. Screening Increases Racial Sentencing Gap



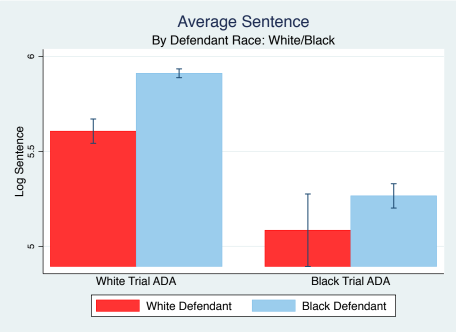
- Black-white sentence differences (on left)
- Since black defendants are less likely to be declined, "real" racial disparity magnifies (on right)
  - Effects are quite large in log scale
  - Is statistical discrimination the reason for disparate screening?

2. White Prosecutors Screen-In Fewer Cases that result in Lower Sentences



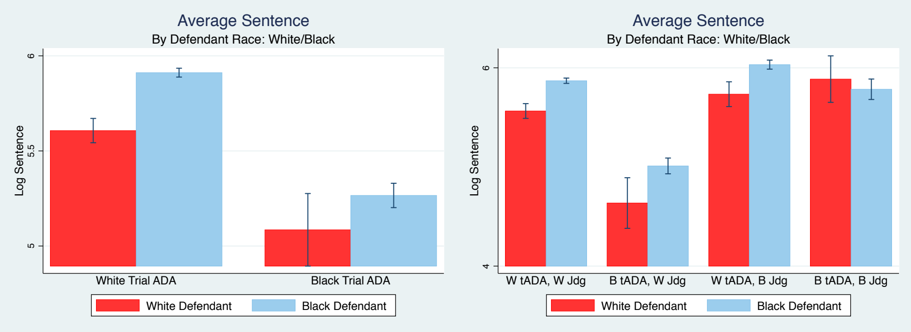
- Black defendants are screened in more (on left)
- White and black screeners let in different cases (on right)
  - If targeting the most severe ones, white screener cases should have *longer* sentences
  - Suggests not about statistical discrimination

3. White Trial Prosecutors Obtain Longer Sentences



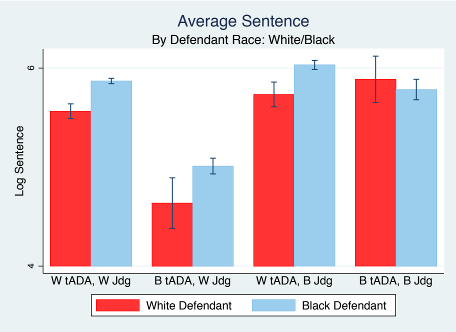
- Most District Attorneys are elected; want to appear tough-on-crime (Pfaff 2016)
- Why are White trial prosecutors more effective in this goal?

4. Black Trial Prosecutors + White Judges Render Shorter Sentences



- The difference seems attributeable to the interaction of hierarchy and race
  - Black trial prosecutors + Black judges render similar average sentences as White trial prosecutors do
  - Effects are quite large in log scale (on right)

5. Black Trial Prosecutors + Black Judges Eliminate or Reverse Racial Sentencing Gap



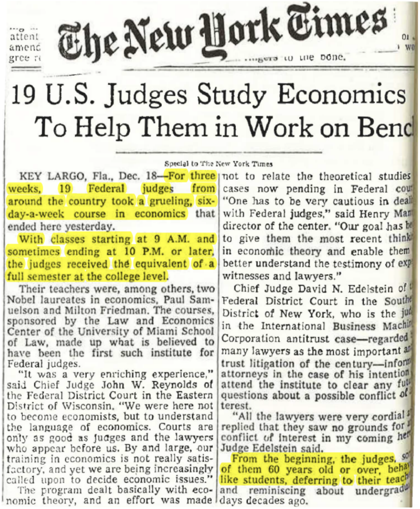
- Hard to explain as statistical discrimination rather than ingroup bias

Revealed Preference Indifference

	Log of Total Sentence in Days	
	(1)	(2)
First Letter Match × Negro	0.174	0.168
	(0.0687)	(0.0686)
N	41793	40011
adj. R-sq	0.475	0.442
First Letter Match × Judge FE	X	X
First Letter Match × Month × Year FE	X	X
First Letter Match × Case Type FE	X	X
First Letter Match × Skin Color FE		X
First Letter Match × Hair Color FE		X
First Letter Match × Eye Color FE		X

- Name letter effects appear only for African Americans labeled “Negro” and not for “Black”
  - robust to controls for skin, hair, eye color

By 1990, 40% of federal judges had attended an economics-training program.



The New York Times

19 U.S. Judges Study Economics To Help Them in Work on Benc

Special to The New York Times

KEY LARGO, Fla., Dec. 18—For three weeks, 19 Federal judges from around the country took a grueling, six-day-a-week course in economics that ended here yesterday.

With classes starting at 9 A.M. and sometimes ending at 10 P.M. or later, the judges received the equivalent of a full semester at the college level.

Their teachers were, among others, two Nobel laureates in economics, Paul Samuelson and Milton Friedman. The courses, sponsored by the Law and Economics Center of the University of Miami School of Law, made up what is believed to have been the first such institute for Federal judges.

"It was a very enriching experience," said Chief Judge John W. Reynolds of the Federal District Court in the Eastern District of Wisconsin. "We were here not to become economists, but to understand the language of economics. Courts are only as good as judges and the lawyers who appear before us. By and large, our training in economics is not really satisfactory, and yet we are being increasingly called upon to decide economic issues."

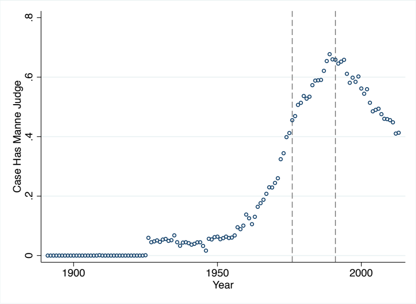
The program dealt basically with economic theory, and an effort was made not to relate the theoretical studies cases now pending in Federal cour

"One has to be very cautious in deal with Federal judges," said Henry Man director of the center. "Our goal has be to give them the most recent think in economic theory and enable them better understand the testimony of exp witnesses and lawyers."

Chief Judge David N. Edelstein of t Federal District Court in the South District of New York, who is the jud in the International Business Machin Corporation antitrust case—regarded many lawyers as the most important an trust litigation of the century—inform attorneys in the case of his intention attend the institute to clear any fut questions about a possible conflict of terest.

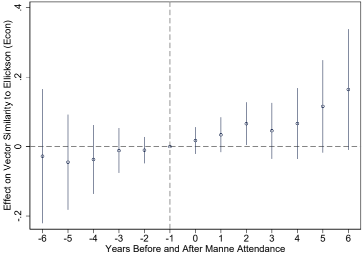
"All the lawyers were very cordial a replied that they saw no grounds for conflict of interest in my coming her Judge Edelstein said.

From the beginning, the judges, so of them 60 years old or over, behav like students, deferring to their teach and reminiscing about undergradu days decades ago.



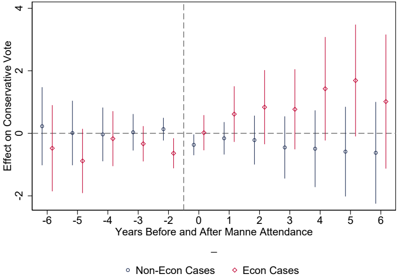
The results of these seminars were dramatic

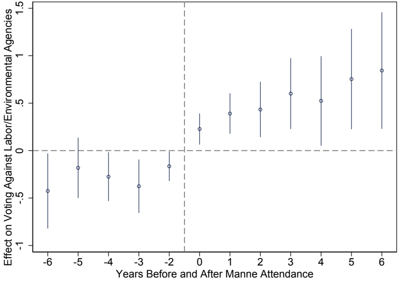
We can see economics language used in academic articles became prevalent in opinions.



The results of these seminars were dramatic

We can see economics trained judges changing how they decided



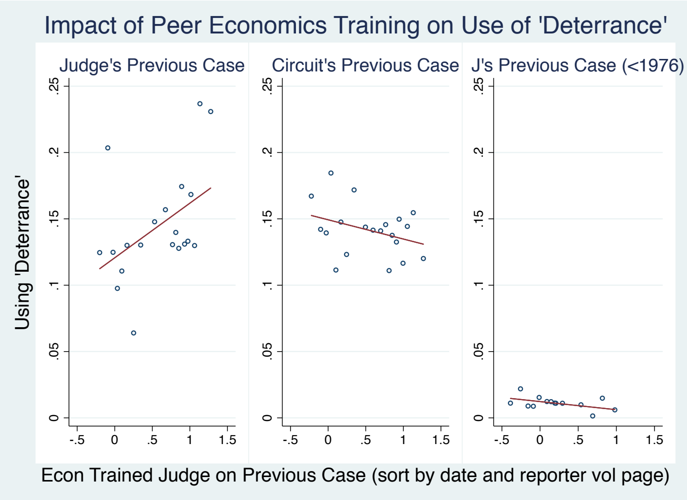


Econ vs Non-Economics Cases

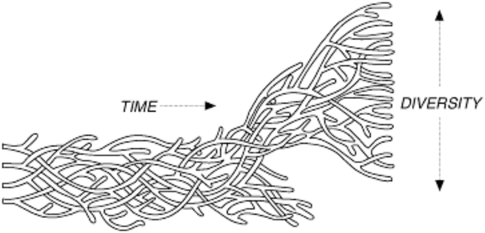
on Labor/Environmental Cases

Impacting their peers

We can see economic language traveling from one judge to another and across legal areas.



The Geneology of Ideology



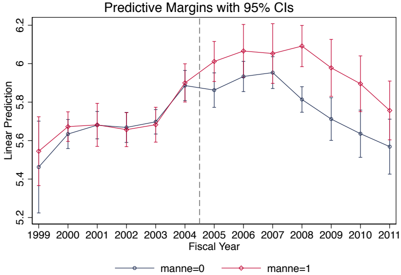
$$P_m = \frac{d_{m\to m}}{d_{\to m} + \delta} \Big/ \frac{d_{m\to \cancel{m}} + \delta}{d_{\to \cancel{m}} + \delta}$$

Scoring Memetic Phrases

Varma, Parthasarathy, and Chen, ICAIL 2017

When judges were given discretion in sentencing

economics trained judges immediately rendered 20% longer sentences relative to the non-economics counterparts.



Ideas Have Consequences: The Impact of Law and Economics on American Justice

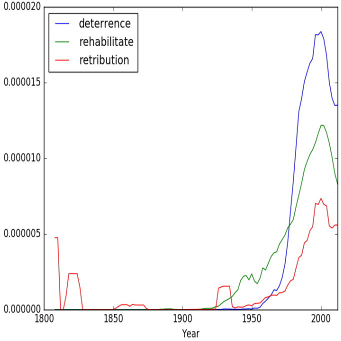
Impact of Economics Judges on Racial Gaps

Economics trained judges are harsher to minorities

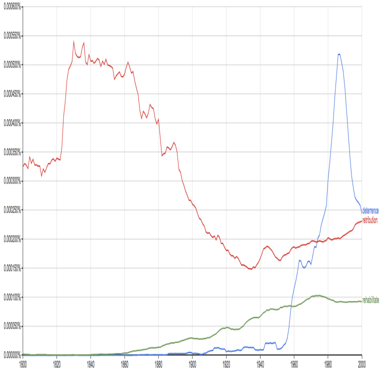
	Life	Months	Life	Months
	(1)	(2)	(3)	(4)
Minority	*0.00395****	*20.84****	*0.00388****	*20.34****
	(0.000770)	*(1.979)*	*(0.00102)*	*(2.170)*
*** Economics**	**0.00401****	**5.413*****	**0.00379****	**3.180***
	(0.00157)	**(2.044)**	**(0.00170)**	**(1.910)**
*** Republican**			0.000641	**4.096****
			(0.00103)	**(1.723)**
*** Minority J**			-0.00119	**-7.451****
			(0.00135)	**(3.167)**
N	156650	155977	154920	154253
adj. R-sq	0.015	0.102	0.015	0.102
Judge FE	Y	Y	Y	Y
Sample	All	All	All	All

Ingroup bias coefficient reduces gradient by one-third

The Great Transformation mentalities changed to be more economical (*Polyani 1944*)



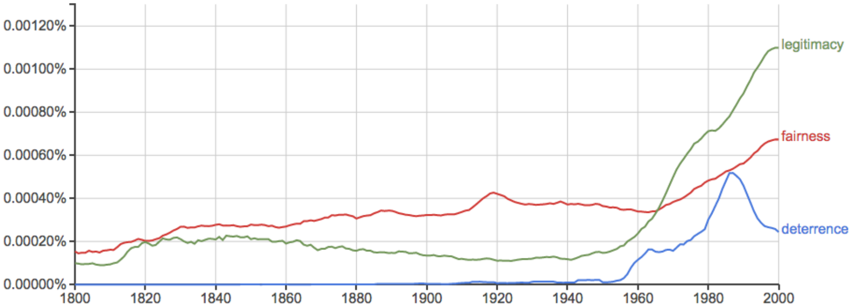
Word Frequency in State Court Opinions



Word Frequency in Google Books

Massive build-up of prisons

AI and the Next Transformation of Law?



Word Frequency in Google Books

- ~~retribution, rehabilitation,~~ deterrence, legitimacy, fairness