Building a Robot Judge

13. Algorithmic Fairness

Elliott Ash

April 28, 2019

Humans vs. Machines

Kleinberg et al (2019)

- Given the same data/features X, machines tend to beat humans:
  - Games: Chess, AlphaGo, Poker
  - Image classification
  - Question answering (IBM Watson)
- But humans see more than machines do. Humans make decisions based on (X, Z)

Bail Decision: Detain or Relase

Kleinberg et al (2019)

- Costs of detention (avg. 2-3 months):
  - Consequential for jobs, families
  - Costs to taxpayers of jails
- Costs of release:
  - failure to appear at trial
  - commit more crimes
- Judge is implicitly making an assessment/prediction about these outcomes.

Data: Kentucky & Federal

Kleinberg et al (2019)

Jurisdiction	Number of cases	Fraction released people	Fraction of Crime	Failure to Appear at Trial	Non-violent Crime	Violent Crime
Kentucky	362k	73%	17%	10%	4.2%	2.8%
Federal Pretrial System	1.1m	78%	19%	12%	5.4%	1.9%

Machine Learning

Kleinberg et al (2019)

- Use labeled dataset (released defendants), to predict whether they fail to appear or commit more crimes. Assess accuracy in test set.
- Issue: Judge sees factors the machine does not
  - Machine makes decisions based on $P(Y|X)$
  - Judge makes decisions based on $P(Y|X,Z)$
    - X, prior crime history
    - Z, other factors not seen by the machine

Data: Defendant Features

Kleinberg et al (2019)

`Age at first arrest, Times sentenced residential correction, Level of charge, Number of active warrants, Number of misdemeanor cases, Number of past revocations, Current charge domestic violence, Is first arrest, Prior jail sentence, Prior prison sentence, Employed at first arrest, Currently on supervision, Had previous revocation, Arrest for new offense while on supervision or bond, Has active warrant, Has active misdemeanor warrant, Has other pending charge, Had previous adult conviction, Had previous adult misdemeanor conviction, Had previous adult felony conviction, Had previous Failure to Appear, Prior supervision within 10 years`

- excludes race, gender, and religion
  - not legal to include – will come back to this issue

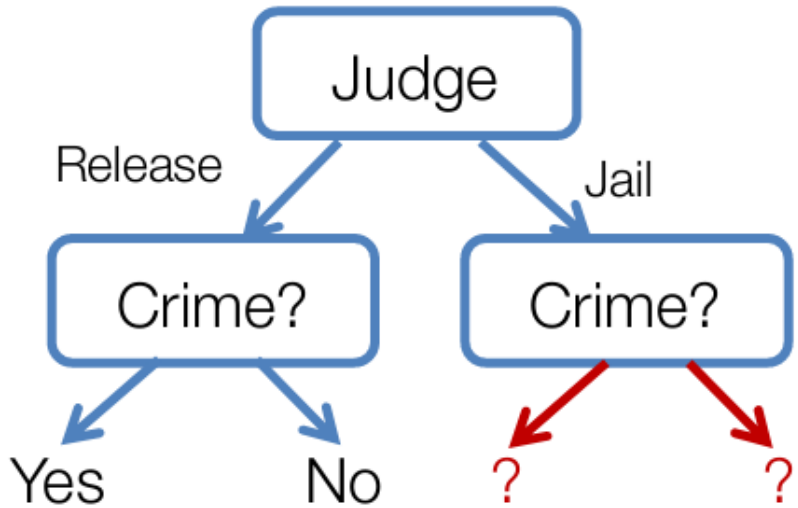
Prediction→Release Rule

Kleinberg et al (2019)

- Predictions create a new release rule:
  - For every defendant predict P(crime)
  - Sort by increasing P(crime)
  - Release bottom k
- What is the fraction released vs. crime rate tradeoff?

Judge is selectively labeling the dataset

Kleinberg et al (2019)



- We can only train on released people:
  - By jailing, judge is selectively hiding labels!

Selection on unobservables

Kleinberg et al (2019)

- Selective labels introduce bias:
  - Say young people with no tattoos have no risk for crime. Judge releases them.
  - Machine observes age, but does not observe tattoos.
  - So, the machine would falsely conclude that all young people do no crime.
  - It would falsely presume that by releasing all young people, it does better than judge!

Keys to Solution

Kleinberg et al (2019)

- Selection problem is one-sided:
  - we observe the counterfactual (crime rate) for released defendants, but not jailed defendants.
- Cases are randomly assigned:
  - this means that on average all judges have the same cases
- Natural variability between judges in leniency.

- $\rightarrow$ Analyze most lenient judges, where released population is minimally selected.

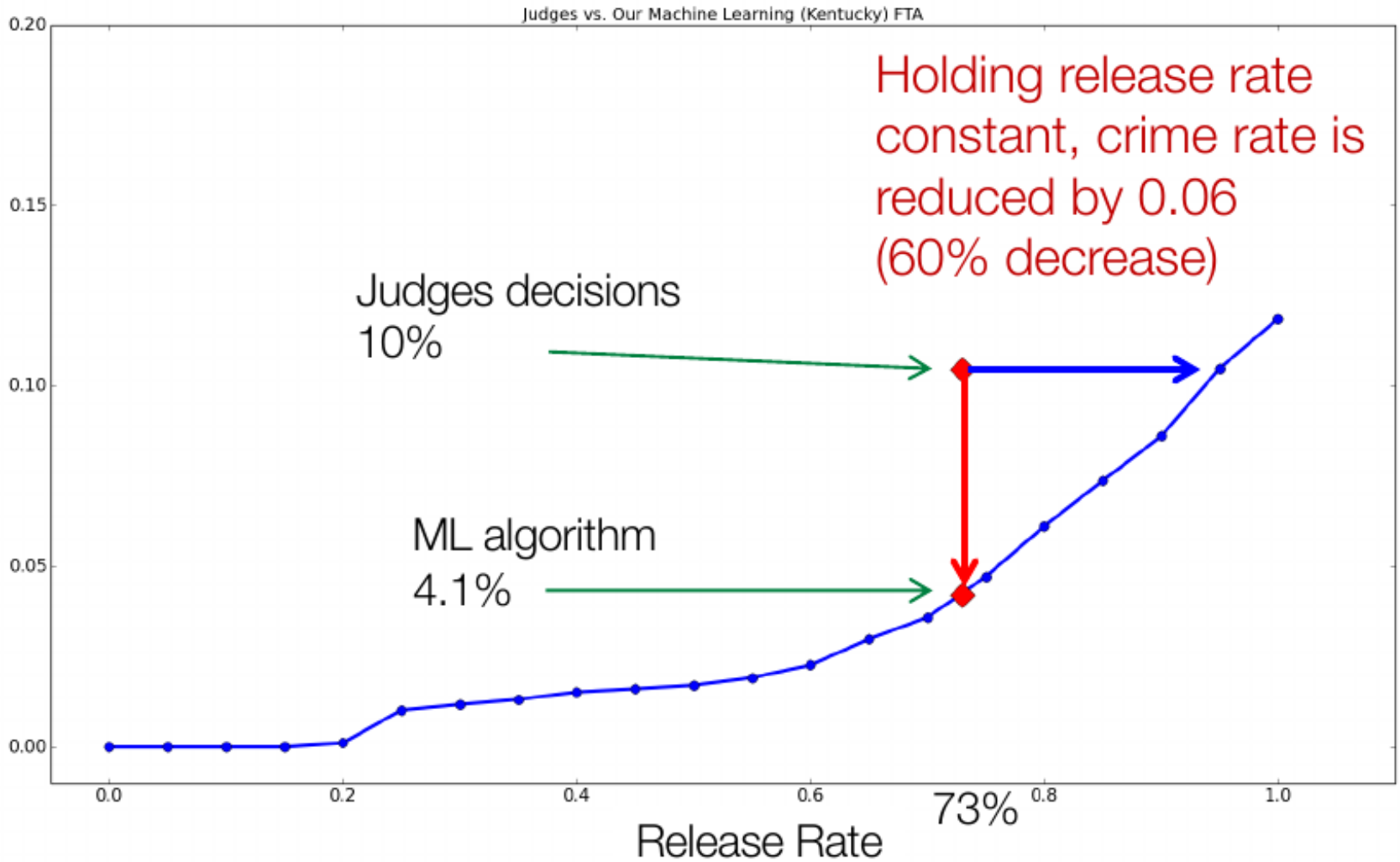
Solution: Contraction

Kleinberg et al (2019)

- Take released population of a lenient judge:
  - Then ask which additional defendant we would we jail to minimize crime rate.
  - Compare change in crime rate to a strict judge

Compare Judge to ML in predicted crime rate

Kleinberg et al (2019)



Algorithm decision doesn't depend on race/ethnicity

Kleinberg et al (2019)

Release Rule	Crime Rate	Drop Relative to Judge	Percentage of Jail Population Black	Hispanic	Minority
Distribution of Defendants (Base Rate)			.4877	.3318	.8195
Judge	.1134 (.0010)	0%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)
Algorithm					
Usual Ranking	.0854 (.0008)	-24.68%	.5984 (.0029)	.3023 (.0027)	.9007 (.0017)
Match Judge on Race	.0855 (.0008)	-24.64%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)
Equal Release Rates for all Races	.0873 (.0008)	-23.02%	.4877 (.0029)	.3318 (.0028)	.8195 (.0023)

Analyzing Judicial mistakes

Kleinberg et al (2019)

- Odds are against algorithm. Judge sees many things the algorithm does not:
  - e.g. defendent "demeanor"
- Two possible reasons why judges are making mistakes:
  - Misuse of observable features
  - Misuse of unobservable features

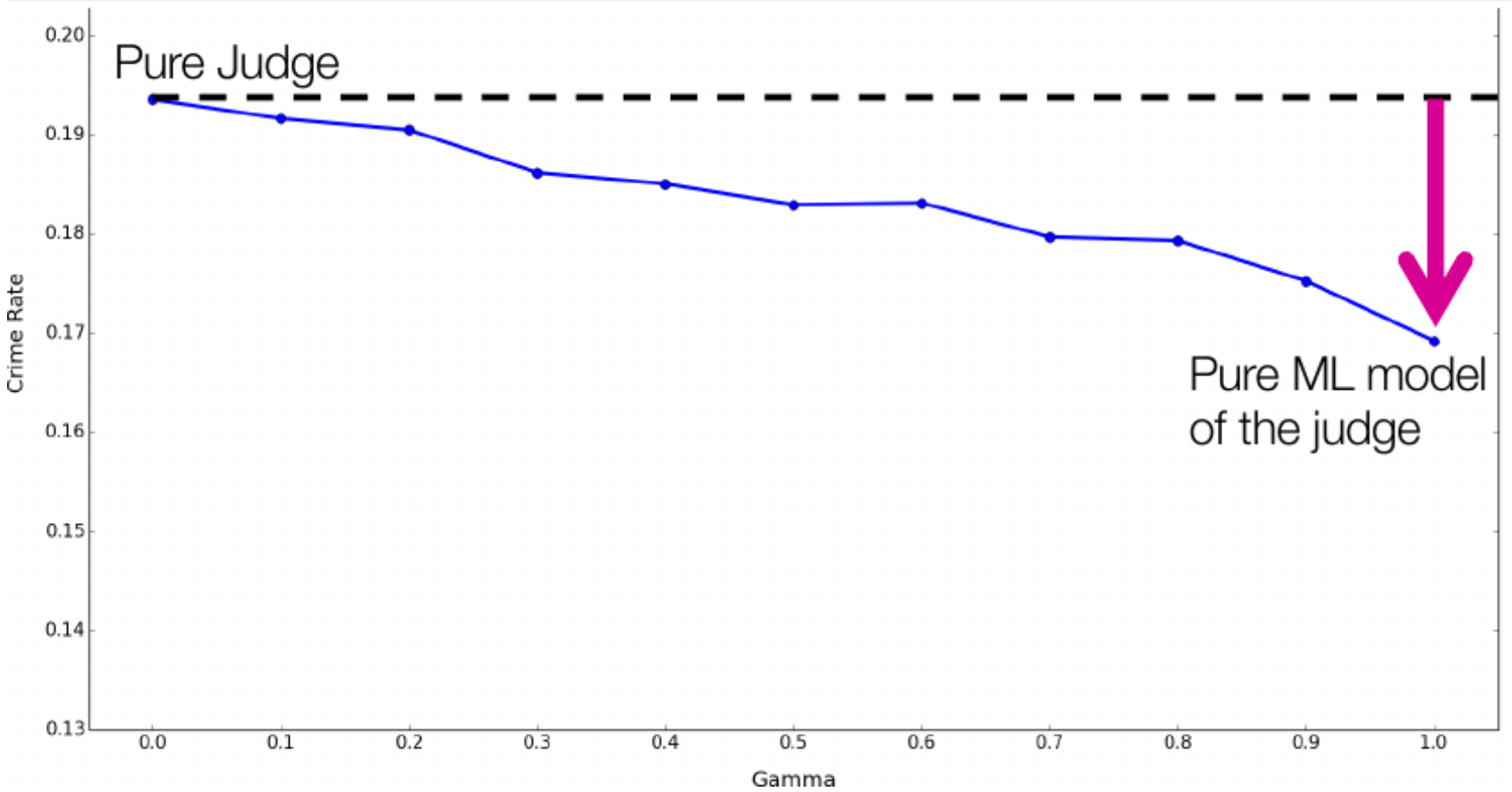
Solution: Predict Judge Decision

Kleinberg et al (2019)

- Predict whether judge gives bail or not:
  - This weights features the way judge does, but does not use unobservables.
- Compute a mix of the real decision and predicted decision:
  - Does this mixture beat the judge in terms of reducing crime rates?
  - If so, the judge is mis-weighing unobservable features.

ML Model of judge decision reduced predicted crime rate

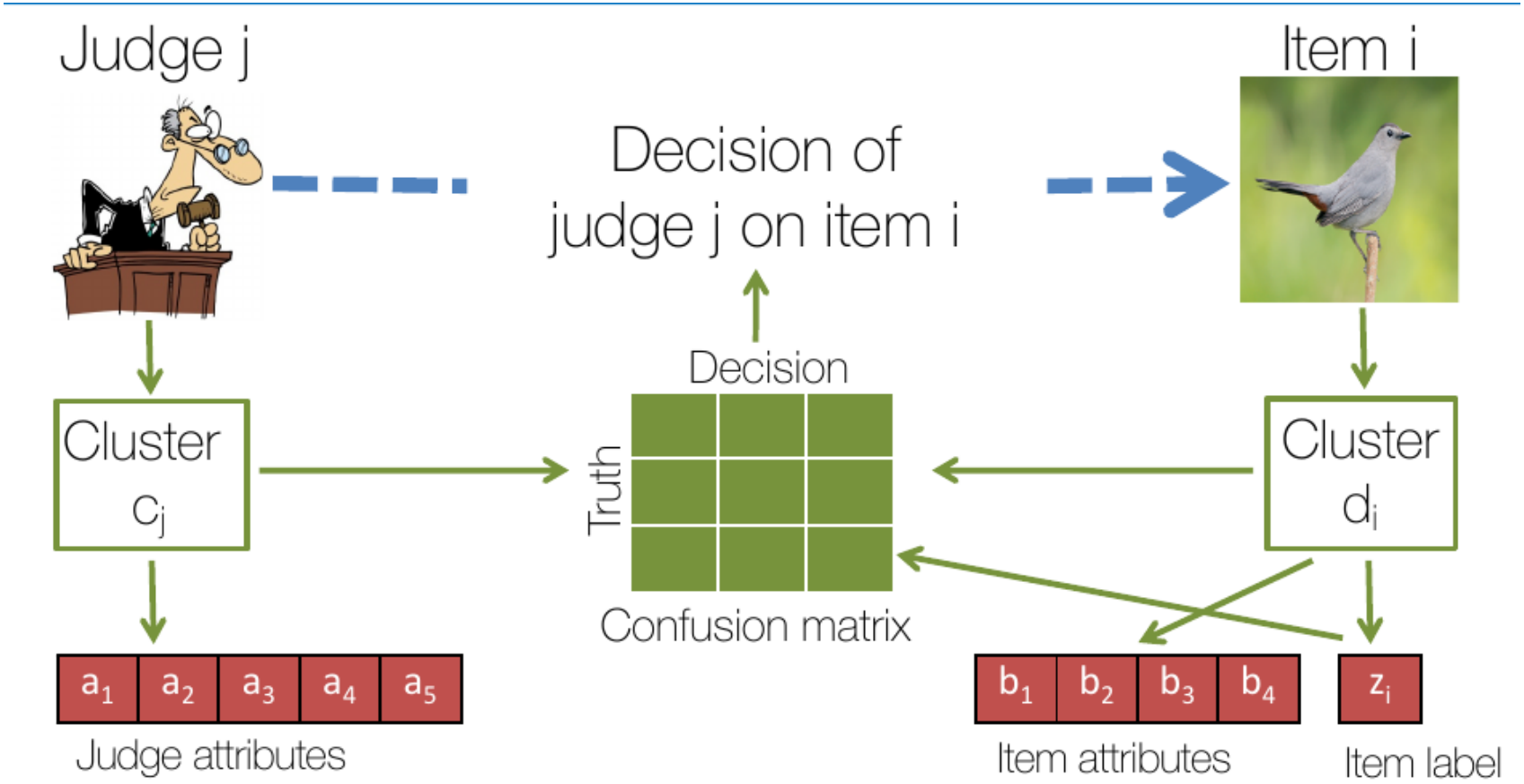
Kleinberg et al (2019)



Algorithm that replicates judge reduces crime rate by 35% (relative to true judge).

Understanding Judicial Errors

Kleinberg et al (2019)



Cluster judges, defendants, and judge-defendant pairs by similarity of confusion matrix.

Judge and defendant attributes

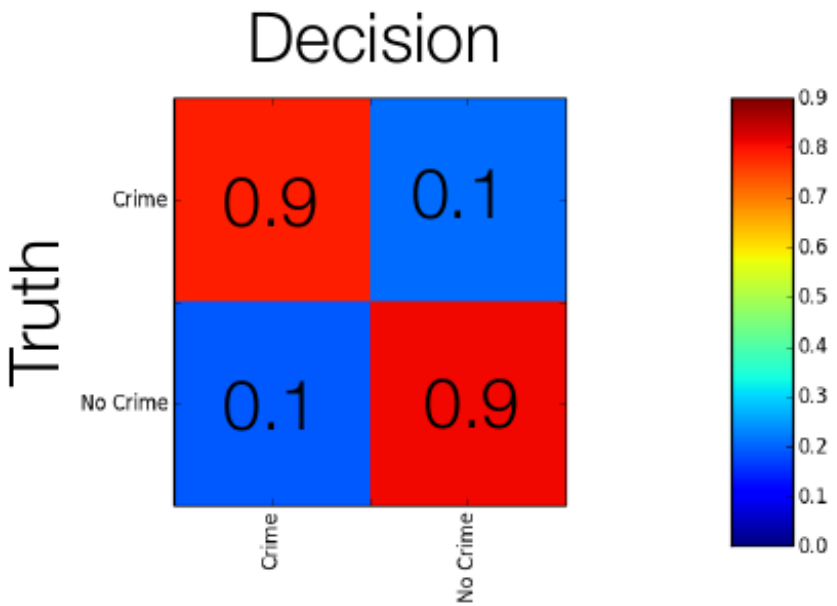
Kleinberg et al (2019)

- Judge attributes:
  - Year, County, Number of Previous Cases, Number of Previous Felony Cases, Number of Previous Misd. Cases, Number of Previous Minor Cases
- Defendant attributes:
  - Previous Crime History, Personal Attributes (Children/Married etc.), Social Status (Education/House Own/Moved a lot in past 10 years), Current Offense Details

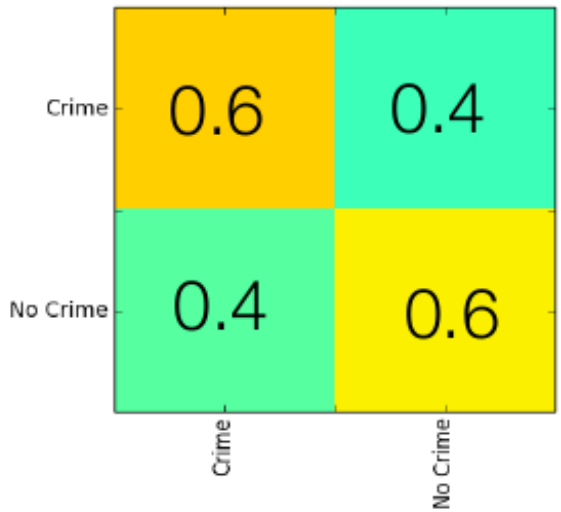
Analyzing judge mistakes

Kleinberg et al (2019)

Experienced Judges

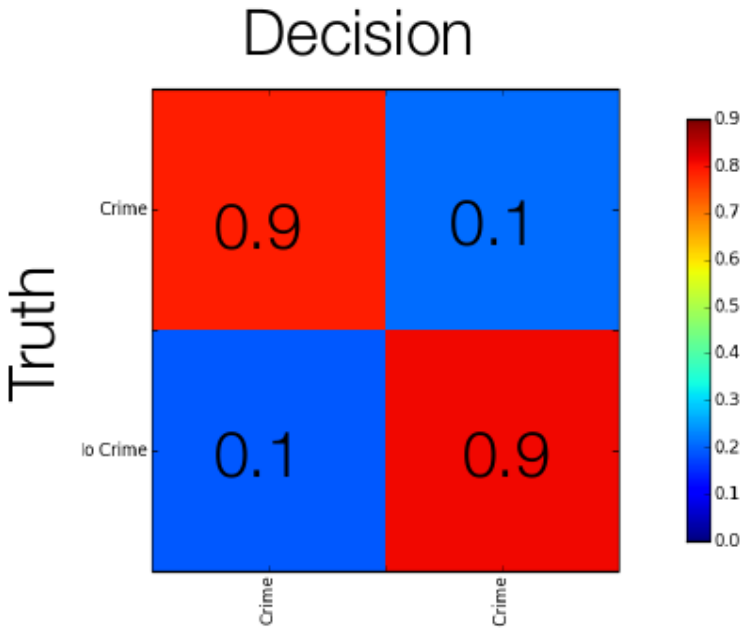


Inexperienced Judges

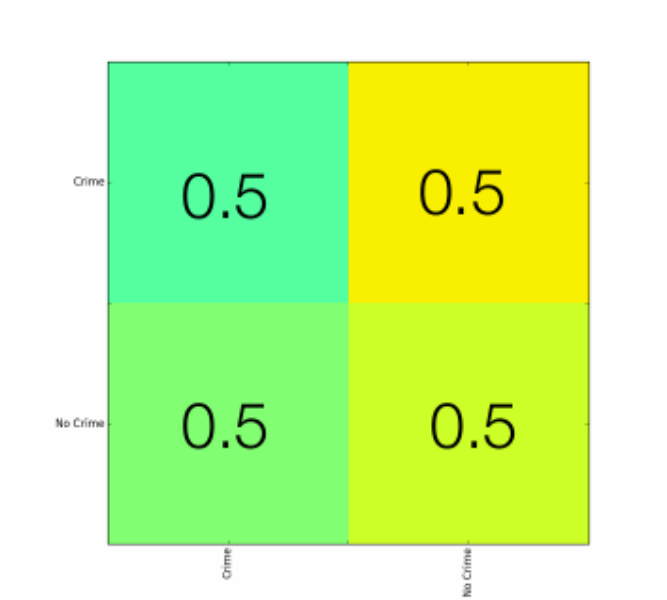


Analyzing judge mistakes

Kleinberg et al (2019)



Defendants who are single, did felonies, and moved a lot are accurately judged



Defendants who have kids are confusing to judges

- Or are judges balancing crime risk against kids' welfare?

Algorithmic Bias

- Algorithms can help us understand if human judges are biased, and diagnose reasons for bias.
- Not just about prediction. Key is starting with decision:
  - Performance benchmark: Current "human" decisions
  - Not ROC but "human ROC"
- Question: What are we really optimizing?

Labels are Driven by Decisions

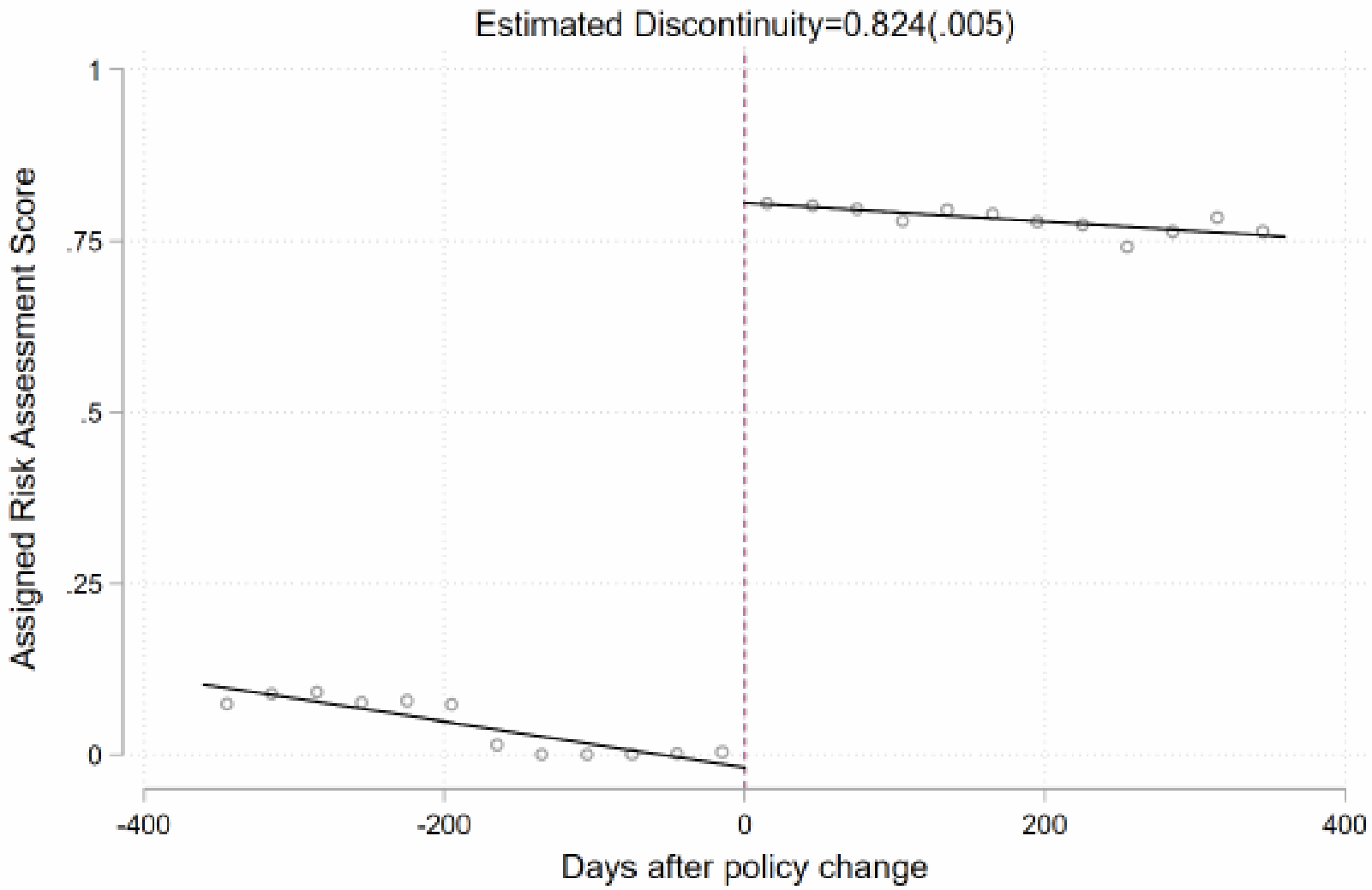
- We don't see labels of people that are jailed
- This is a broader problem in policymaking systems:
  - Prediction -> Decision -> Outcome
- Which outcomes we see depends on our decisions

Behavioral responses to decisions

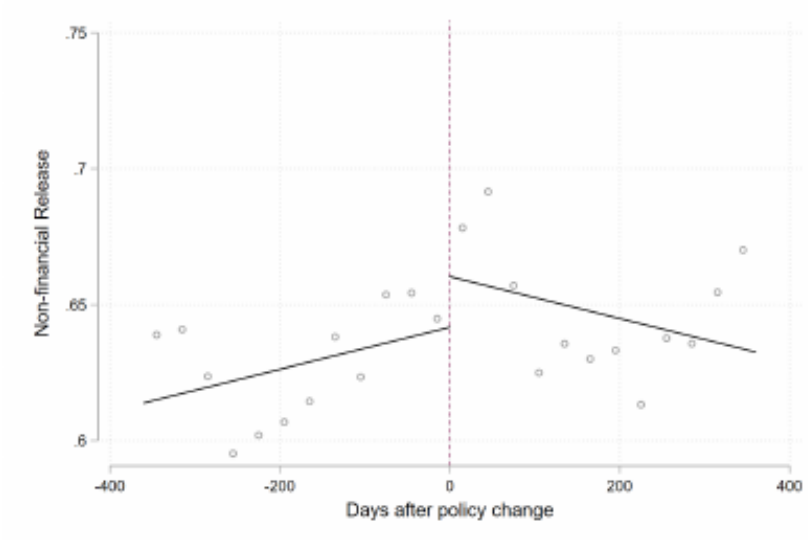
- Judges and criminals will change their behavior in response to adopting machine decision supports.
  - Could have unintended consequences, or create a self-reinforcing feedback loop.

Sloan et al: Fuzzy RD for risk scoring

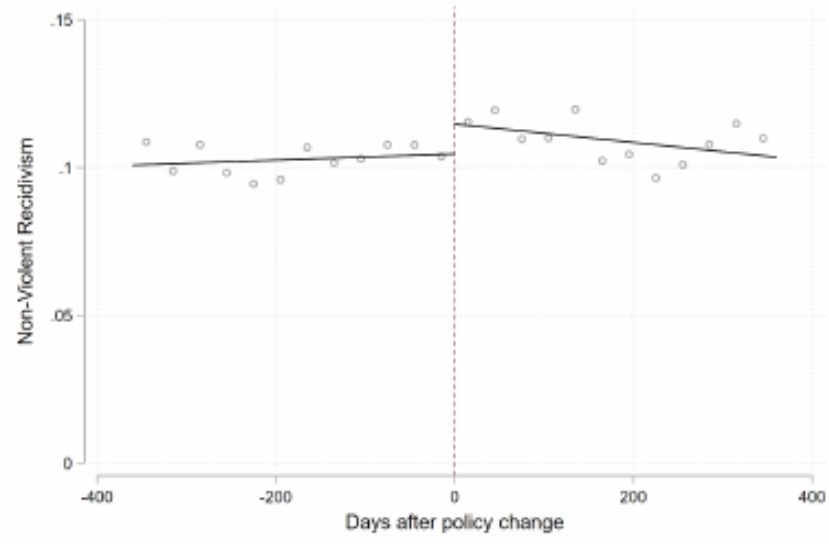
Figure 4: Regression Discontinuity Results for the Probability of Receiving a Risk Assessment Score



Sloan et al: Risk scoring increases release rates and recidivism



(a) Non-financial Bond



(a) Probability of Non-Violent Recidivism

- In response to risk scoring, judges release more poor defendants.

Constraints on input features

- For example, race would be illegal to include.
  - But many other characteristics correlate with race.
- Equalizing metrics across races/groups will result in other distortions.

Focusing on re-arrest rates is limited

- Is minimizing the crime rate really the right goal?
- There are other important factors
  - Consequences of jailing on the family
  - Jobs and the workplace
  - Future behavior of the defendant
- How could we measure/model these?

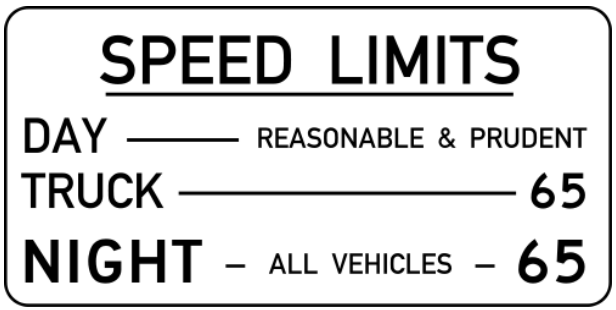
Algorithmic Bias

- Algorithm generates consistent decisions for same evidence, correcting individual-level biases across judges.
- But *systematic* biases across all judges will *not* be corrected:
  - These could be reproduced or even *amplified* in the automated decisions.
- Skeem and Lovenkamp (2016) analyze popular criminal risk metric:
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories.
  - **Pre-existing criminal-justice biases are reproduced in decisions guided by the metric.**

Other limitations

- Transparency:
  - Closed-source algorithms result in "black box justice" and could be abused by insiders.
  - But open-source algorithms are prone to gaming: savvy attorneys could "trick" the algorithm.
- Algorithm can only use evidence that appears in a lot of cases; it ignores special/mitigating circumstances.
- Would not work on new types of cases.
  - In particular, would not account for new laws/legislation.
- Teaching the algorithm to understand rare evidence, and to understand new laws, would require something much closer to **legal artificial intelligence**.

Legal Vagueness and Value Judgments



- Even if the AI could read new laws, there is the problem of legal vagueness:
  - How will the AI decide in this circumstance?

- Making choices in the presence of vagueness or indeterminacy requires value judgements.
  - What counts as a "good" outcome? Is it even measurable?