Big Data for Public Policy

9. Policy Implications of AI

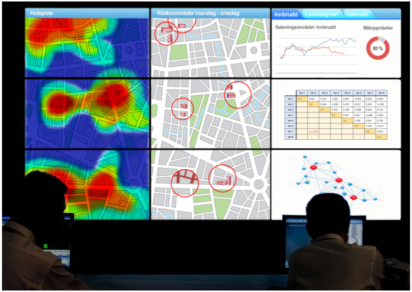
Elliott Ash & Malka Guillot

(Bigger) Data can help solve (bigger) policy problems.

(Bigger) Data can help solve (bigger) policy problems.

(Bigger) Data can cause its own (bigger) problems.

Predictive Policing

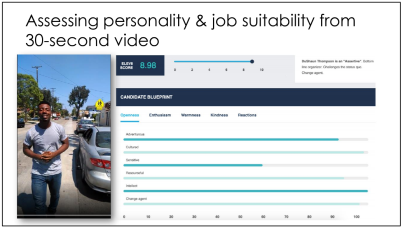


Predictive policing poses discrimination risk, thinktank warns

Machine-learning algorithms could replicate or amplify bias on race, sexuality and age

▲ One officer said human biases including more stop and searches of black men were likely to be introduced into algorithm data sets. Photograph: Carl Court/Getty Images

Algorithmic Hiring Systems Don't Work!



Source: Raghavan et al, 2019.

(1) Genuine, rapid technological progress

Narayanan Slides

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

These are *perception tasks.*
Ethical concerns stem from *high accuracy.*

(2) Imperfect but improving steadily

- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

These are *human judgment tasks.*
Ethical concerns stem from *subjectivity* → some *error is inevitable.*

(3) Fundamentally suspect
Narayanan Slides

- Predicting criminal recidivism
- Predicting job performance
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These are *social outcome prediction tasks.*
Ethical concerns are fundamental, amplified by inaccuracy due to the difficulty of predicting these outcomes.

Accuracy of recidivism prediction

COMPAS tool (137 features): 65% ± 1% (slightly better than random)

Logistic regression (2 features): 67% ± 2%

Age and number of priors

Dressel & Farid. *The accuracy, fairness, and limits of predicting recidivism*. Science Advances 2018.

- algorithm predicts re-arrest (not recidivism), so some of the predictive performance comes from being able to predict the biases of policing.

Harms of using AI for predicting social outcomes

Narayanan slides

- Hunger for personal data
- Transfer of power from domain experts & workers to unaccountable tech companies
- Lack of explainability
- Distraction from interventions
- Veneer of objectivity
- ...

Outline

Machine Predictions and Human Decisions

How Judges Respond to Decision Support

Algorithmic Bias in the Courts

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)

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

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%

Source: Jure Leskovec slides.

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.

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

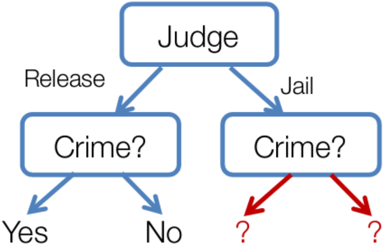
Prediction→Release Rule

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.
  - could be any predictive characteristic that human judge sees, but not recorded in dataset.
- Machine observes age, but does not observe tattoos.

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.
  - could be any predictive characteristic that human judge sees, but not recorded in dataset.
- 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.

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.

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.

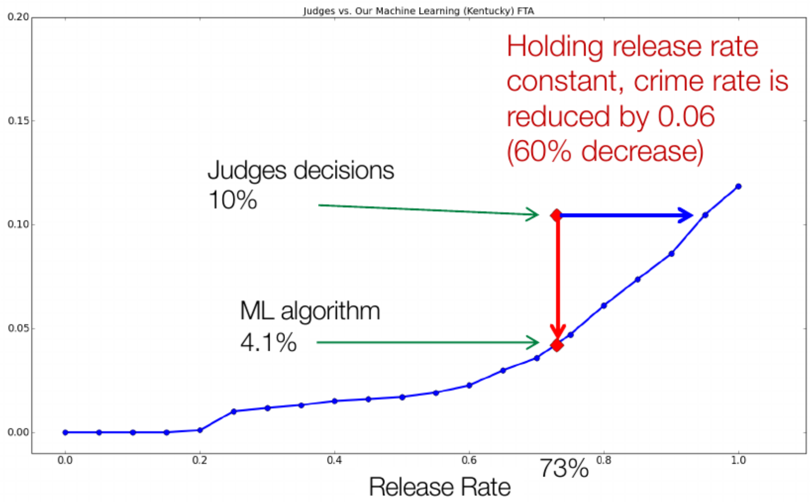
Solution: Contraction Approach

Kleinberg et al (2019)

- Take released population of a lenient judge:
  - Then ask which additional defendant we would 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's decisions don't depend on race/ethnicity

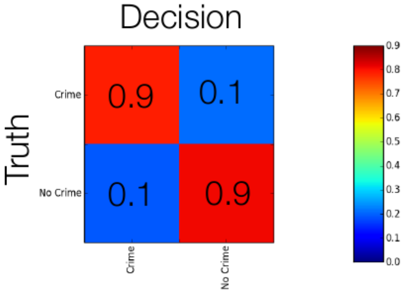
Kleinberg et al (2019)

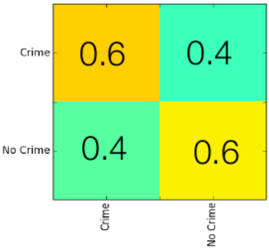
Release Rule	Crime Rate	Drop Relative to Judge	Percentage of Jail Population Black	Hispanic	Minority
Distribution of Defendants (Base Rate)			.4877	.3318	.8195
Judge	.1134 (.0010)	0%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)
Algorithm					
Usual Ranking	.0854 (.0008)	-24.68%	.5984 (.0029)	.3023 (.0027)	.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 judge mistakes
Kleinberg et al (2019)

Analyzing judge mistakes
Kleinberg et al (2019)

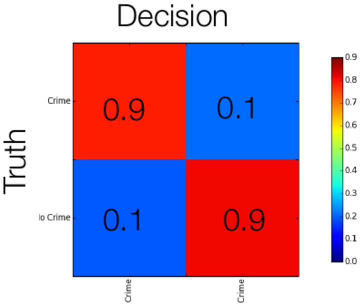
Experienced Judges	Inexperienced Judges

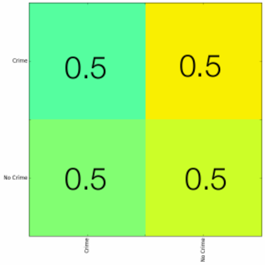




Analyzing judge mistakes

Kleinberg et al (2019)



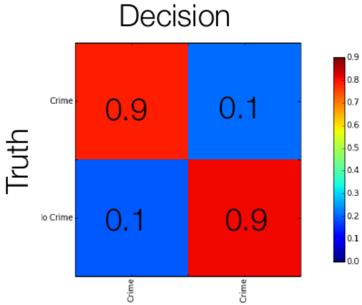


Defendants who are single, did felonies, and moved a lot are accurately judged

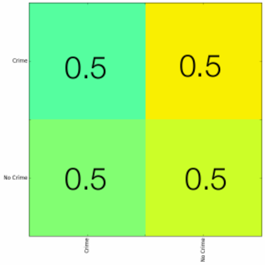
Defendants who have kids are confusing to judges

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?

Evaluating Machine Decision Support

- Not just about prediction. Key is starting with decision:
  - Performance benchmark: Current “human” decisions

Evaluating Machine Decision Support

- Not just about prediction. Key is starting with decision:
  - Performance benchmark: Current "human" decisions
- 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:

$$\text{Prediction} \rightarrow \text{Decision} \rightarrow \text{Outcome}$$

- Observed outcomes depend on decisions.

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?

Outline

Machine Predictions and Human Decisions

How Judges Respond to Decision Support

Algorithmic Bias in the Courts

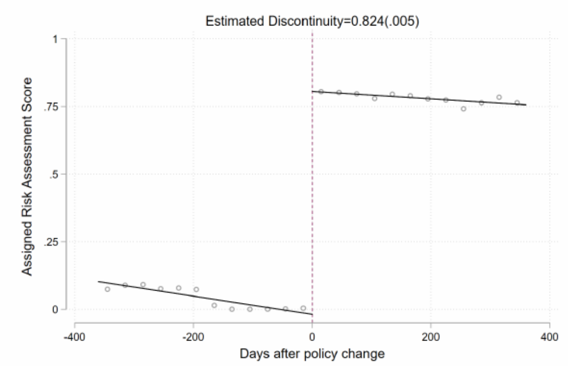
Behavioral responses to decisions

- 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.

How do judges respond to risk scoring?

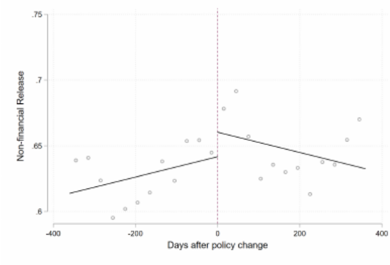
Sloan et al (2018)

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

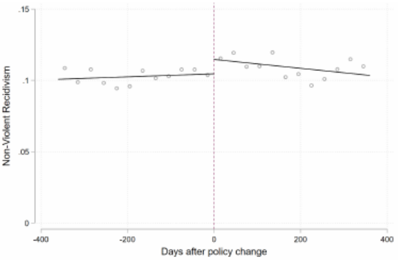


Fuzzy RD, comparing outcome before/after assessment score introduced.

Risk scoring increases release rates and recidivism
Sloan et al (2018)



(a) Non-financial Bond



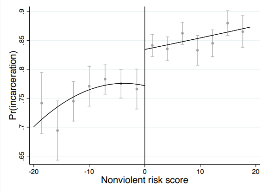
(a) Probability of Non-Violent Recidivism

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

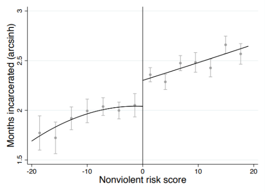
Similar Evidence from Florida

Stevenson and Doleac 2020

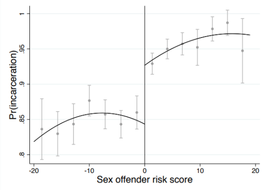
(a) Nonviolent risk score and probability of incarceration



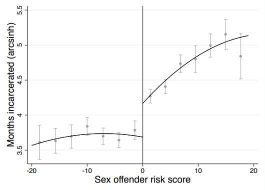
(b) Nonviolent risk score and the sentence length



(c) Sex offender risk score and probability of incarceration



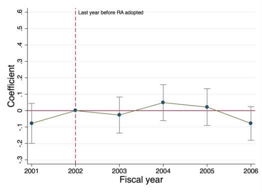
(d) Sex offender risk score and the sentence length



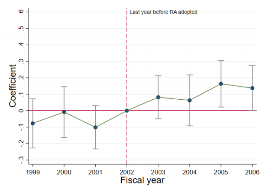
Judge Response is Much Lower than Predicted

Stevenson and Doleac 2020

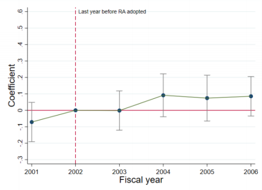
(a) Risk assesment's **actual** impact on sentencing for black defendants



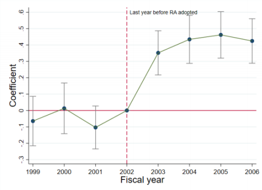
(b) Risk assesment's **actual** impact on sentencing for young defendants



(c) Risk assesment's **simulated** impact on sentencing for black defendants



(d) Risk assesment's **simulated** impact on sentencing for young defendants



Outline

Constraints on input features

- How to prevent algorithms from being biased toward different social groups?

Constraints on input features

- How to prevent algorithms from being biased toward different social groups?
- For example, race would be illegal to include.
    - But many other characteristics correlate with race.

Constraints on input features

- How to prevent algorithms from being biased toward different social groups?
- For example, race would be illegal to include.
  - But many other characteristics correlate with race.
- Equalizing metrics (e.g. risk scores, or accuracy) across races/groups will result in other distortions.

Algorithmic Bias

- Algorithm generates consistent decisions for same evidence, correcting individual-level biases across judges.

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.

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.

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 of algorithmic decisions

- 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.

Other limitations of algorithmic decisions

- 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.

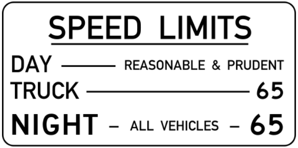
Other limitations of algorithmic decisions

- 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.

Other limitations of algorithmic decisions

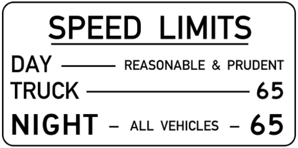
- Transparency:
  - Closed-source algorithms result in "black box justice" and could be abused by insiders.
  - But open-source algorithms are prone to gaming: savvy attorneys could "trick" the algorithm.
- Algorithm can only use evidence that appears in a lot of cases; it ignores special/mitigating 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?

Legal Vagueness and Value Judgments

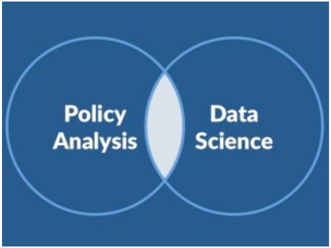


- Even if the AI could read new laws, there is the problem of legal vagueness:
  - How will the AI decide in this circumstance?

- Making choices in the presence of vagueness or indeterminacy requires value judgements.

What counts as a "good" outcome? Is it even measurable?

What are some other issues in big data and public policy?



1. **What is the policy problem or research question?**
2. Data:
    - What is interesting about the data? Is it the right dataset to solve this problem? Were sufficient visuals and descriptive statistics provided to trust the data and its usefulness for the stated purpose?
3. Machine learning:
    - What are we trying to measure or predict? Is the right model being used for that purpose?
    - Were hyperparameters properly tuned without seeing test data? Were informative test-set metrics reported and/or visualized?
    - Were the model predictions effectively validated, for example through model explanation methods?
4. Policy analysis:
    - Did the resulting statistics or predictions provide some evidence or solutions ot the stated problem or research question?
    - **Highlight limitations and open questions.**