Algorithms in the Hands of Humans: Implications for Fairness

Jennifer L. Doleac

Texas A&M University

JUSTICE
TECH LAB

Real-world effects of algorithms depend on how humans use them

- Lots of attention paid to creating fair algorithms
  - Let's imagine we create an algorithm we're happy with — then what?
  - Most discussion is framed as human vs. machine
  - But machines' predictions rarely replace humans' predictions — the former aim to inform the latter

- How this new information affects real-world outcomes we care about will depend crucially on:
  - What humans might have done in the absence of that information
  - Human decision-makers' objective functions and the various incentives they face
  - **This is where social scientists are needed**

Case study: Effect of algorithmic risk scores on criminal sentencing

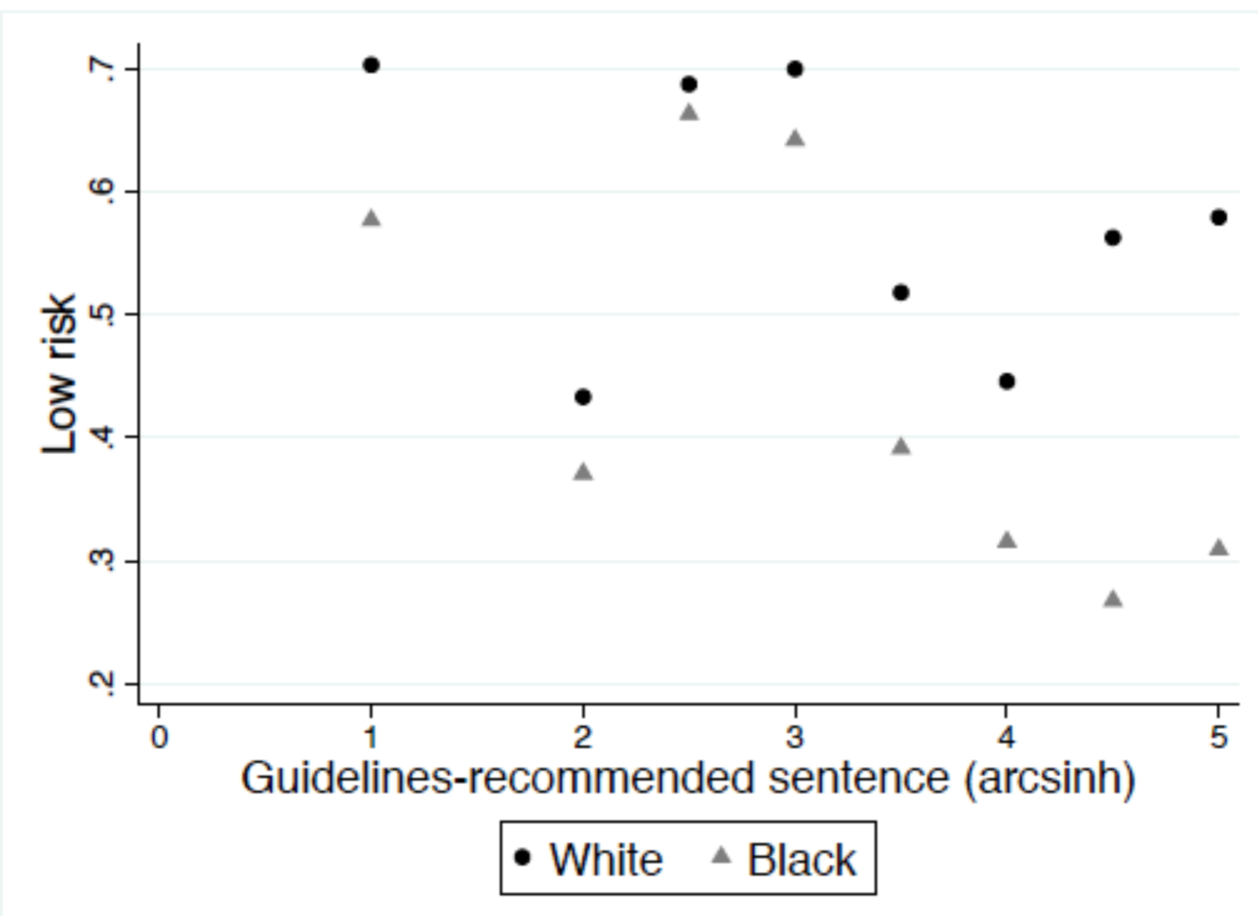
- Stevenson & Doleac (2019) considers effects on sentencing in Virginia
  - New risk assessment for non-violent offenders aimed to divert 25% lowest-risk offenders from incarceration
  - Risk assessment included controversial elements such as employment and marital status, in addition to less controversial variables like age and criminal history

- We find that judges pay attention to the risk scores: they change who they incarcerate
  - But that's the end of the good news
    - No net effect on incarceration rates
    - No efficiency gains (that is, no reduction in recidivism)
    - Judges appear to have responded as much to the absence of a diversion recommendation as to the recommendations themselves — this led to unintended consequences
  - What about fairness? We consider differential effects by:
    - Race (black vs. white)
    - Age (less than 23 vs. older)

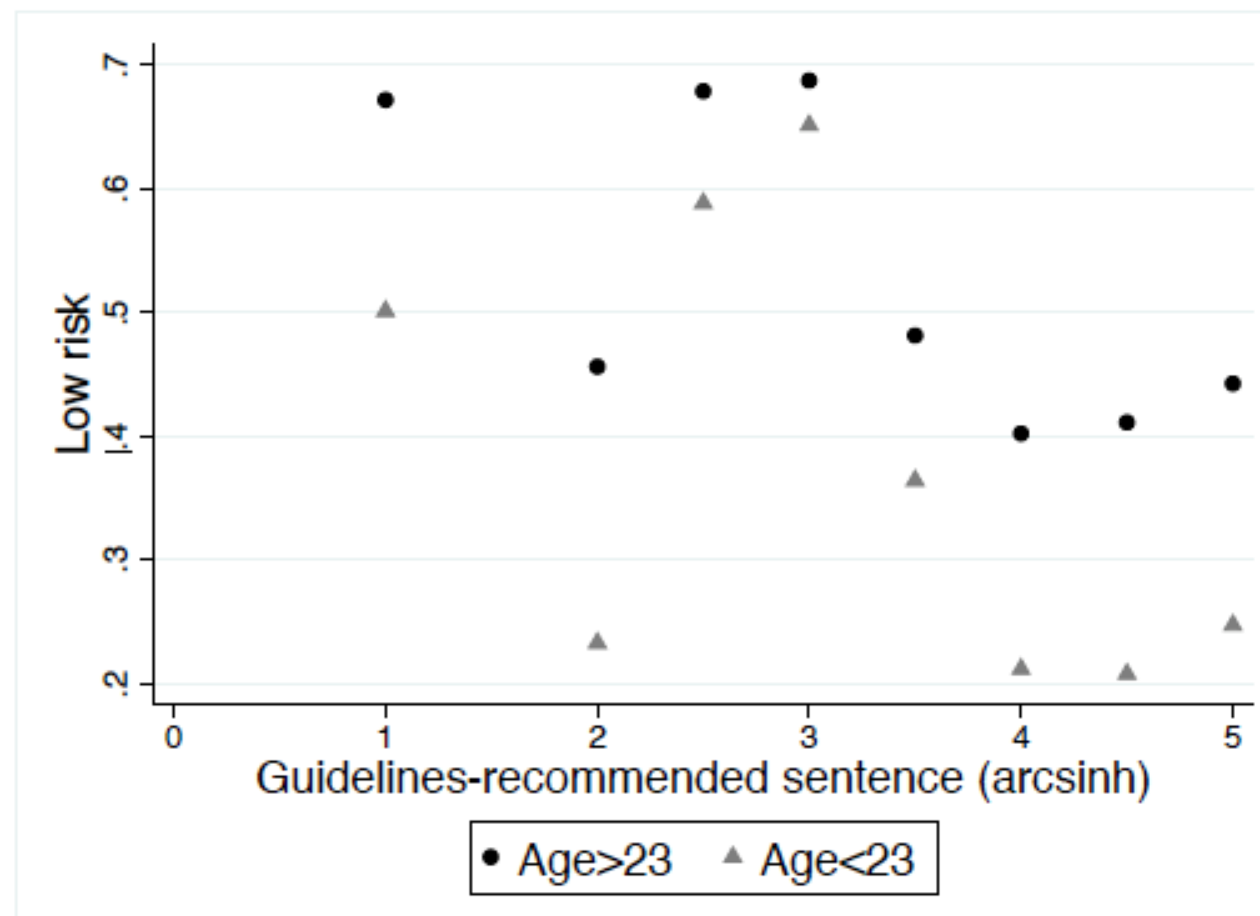
Risk scores are worse for black and young offenders

- Black (young) defendants have higher risk scores than white (older) defendants with the same guidelines-recommended sentence

Racial disparities in diversion recommendation



Age disparities in diversion recommendation



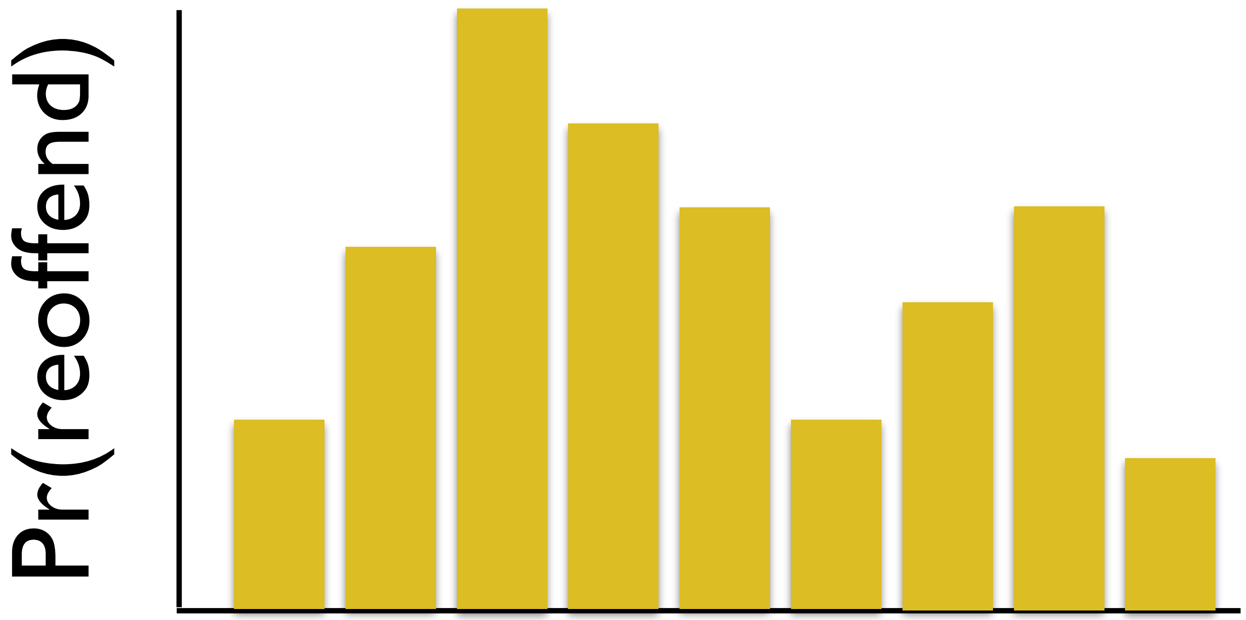
Adding information can help disadvantaged groups

- Does this mean risk assessments will increase disparities in sentencing?
  - Not necessarily!
  - Depends on judges' beliefs about group-level reoffending rates without the risk scores
  - Eliminating information that is unfavorable to a particular group does not necessarily help that group, due to statistical discrimination (see Ban the Box literature)

Statistical discrimination with threshold

- Imagine a set of offenders from a particular group (gender, race, crime type)
- Judges don't have enough info to distinguish between individuals within the group, so use group averages to predict what is likely true of the individual (**statistical discrimination**)

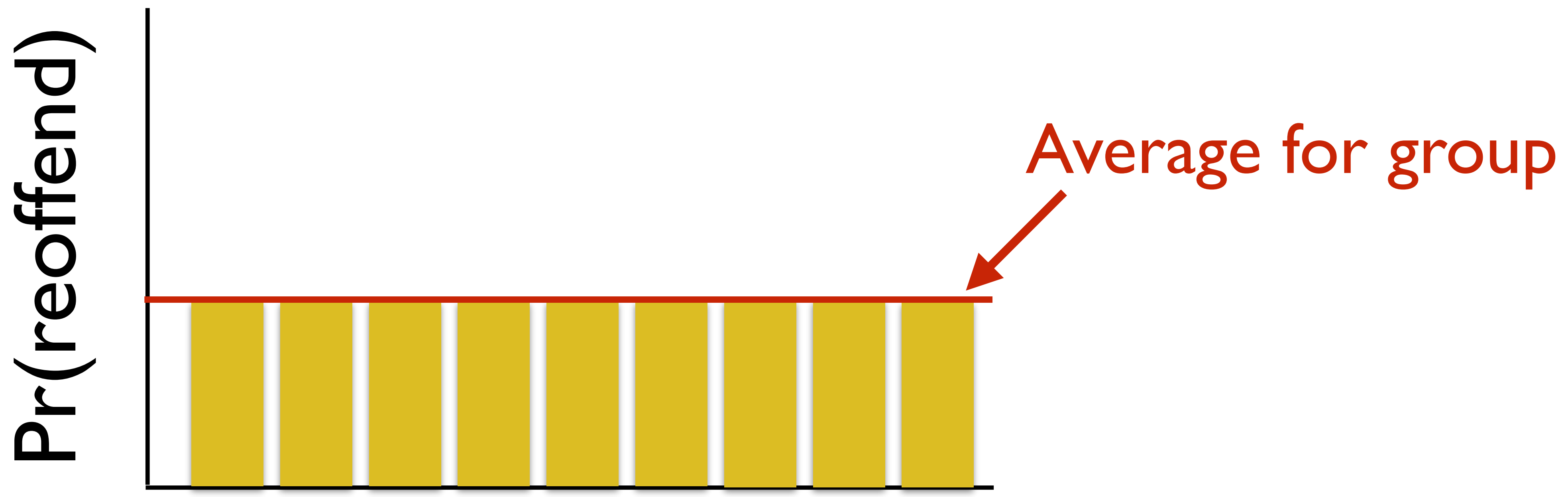
Actual distribution of risk:



Statistical discrimination with threshold

- Imagine a set of offenders from a particular group (gender, race, crime type)
- Judges don't have enough info to distinguish between individuals within the group, so use group averages to predict what is likely true of the individual (**statistical discrimination**)

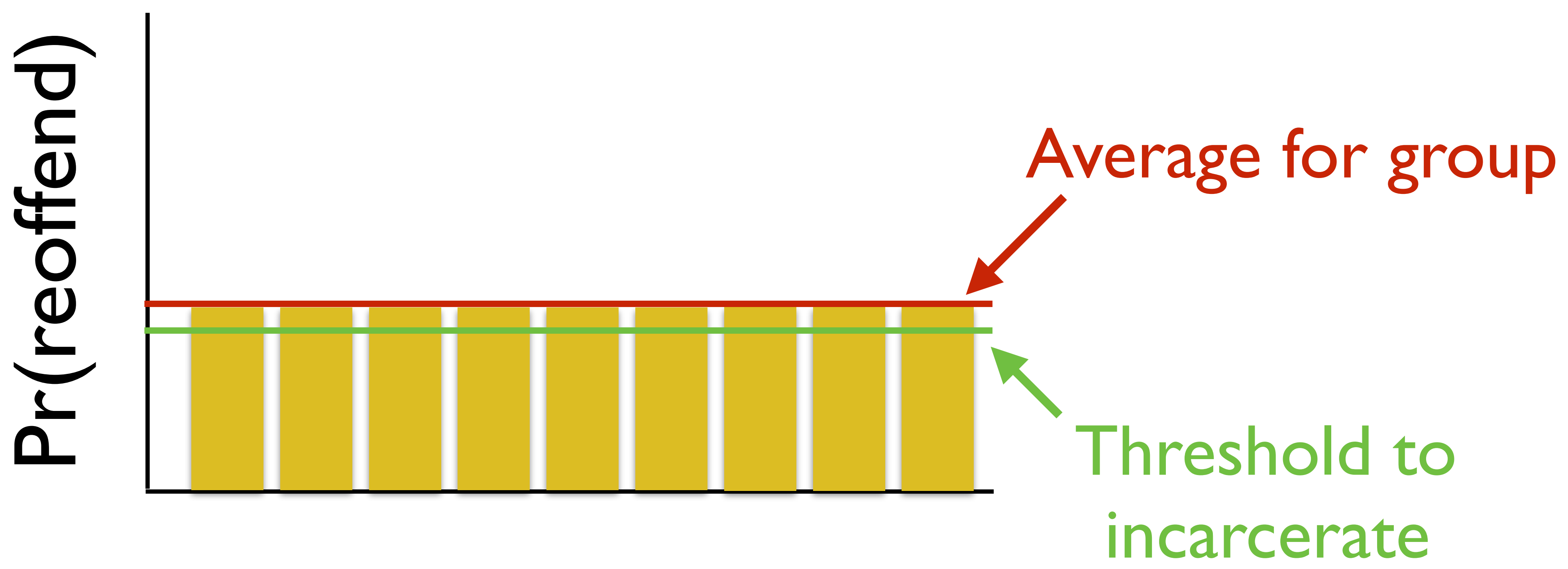
What the distribution looks like to the judge:



Statistical discrimination with threshold

- Imagine a set of offenders from a particular group (gender, race, crime type)
- Judges don't have enough info to distinguish between individuals within the group, so use group averages to predict what is likely true of the individual (**statistical discrimination**)

- If they incarcerate anyone above a certain risk threshold (green line below), then they'll incarcerate anyone in a group that has an average risk level above that threshold
  - In the example below, the **incarceration rate is 100%** when individual-level risk scores aren't available, because the group average is above the threshold



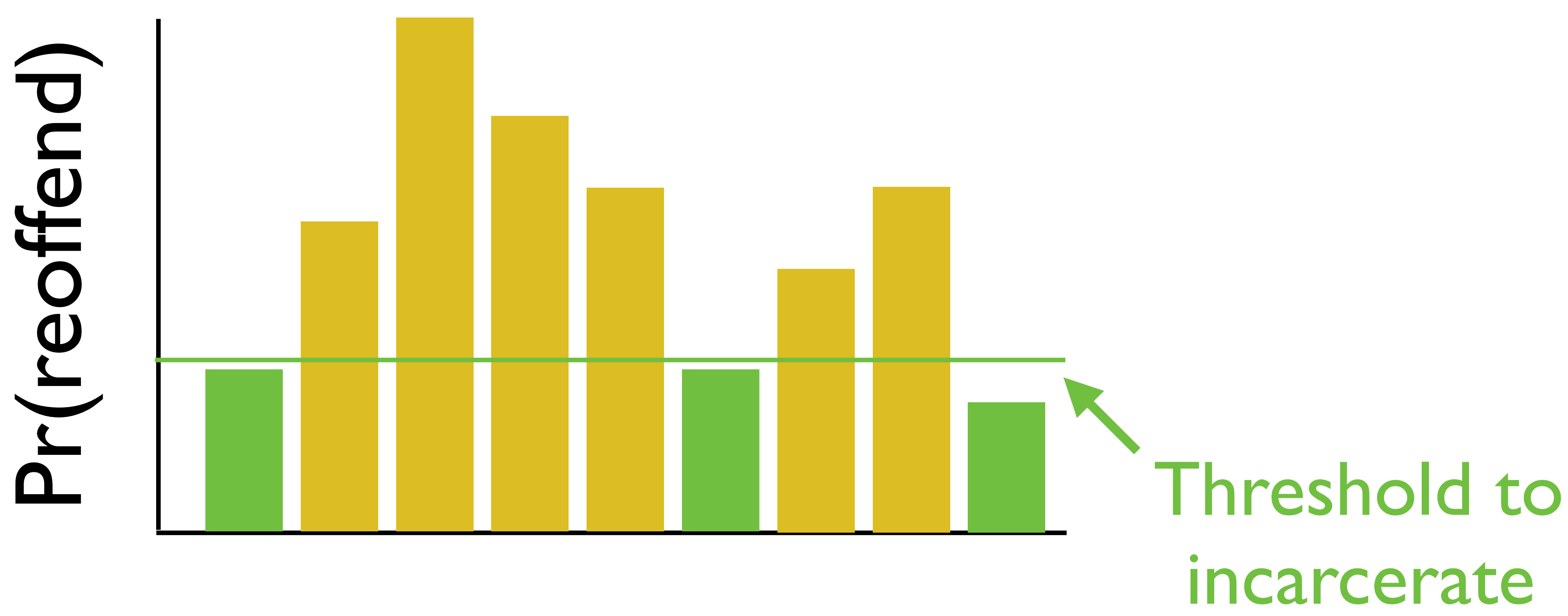
Adding information can help low-risk members of high-risk groups

- Now imagine that judges get risk score information that allows them to distinguish between individuals (or at least disaggregate the groups a bit)

Adding information can help low-risk members of high-risk groups

- Now imagine that judges get risk score information that allows them to distinguish between individuals (or at least disaggregate the groups a bit)

- Note that the information is still, on average, worse for everyone in the graph below — the underlying risk levels have not changed
  - But those who are lower-risk benefit from more detailed info being revealed
  - Judges are now able to distinguish between low- and high-risk defendants within the group
  - **Incarceration rate drops from 100% to 67%**



What does this mean for those concerned about fairness?

- Real-world effects can be tough to predict
- Average algorithmic risk scores for groups don't tell us whether those groups are helped or hurt by the use of algorithms — will depend on what human decision-maker assumes in the absence of those scores

- And judges may be considering lots of other factors, in addition to risk level:
  - Culpability of defendant
  - Victims' wishes
  - Political pressure to be tough on crime
  - Asymmetric cost of making the wrong decision

- Algorithmic risk scores may be better info about just one factor they're considering
  - Risk score info might also interact with some of the factors above (e.g. a low-risk score could reduce political pressure to incarcerate, and this interaction effect could vary across groups)

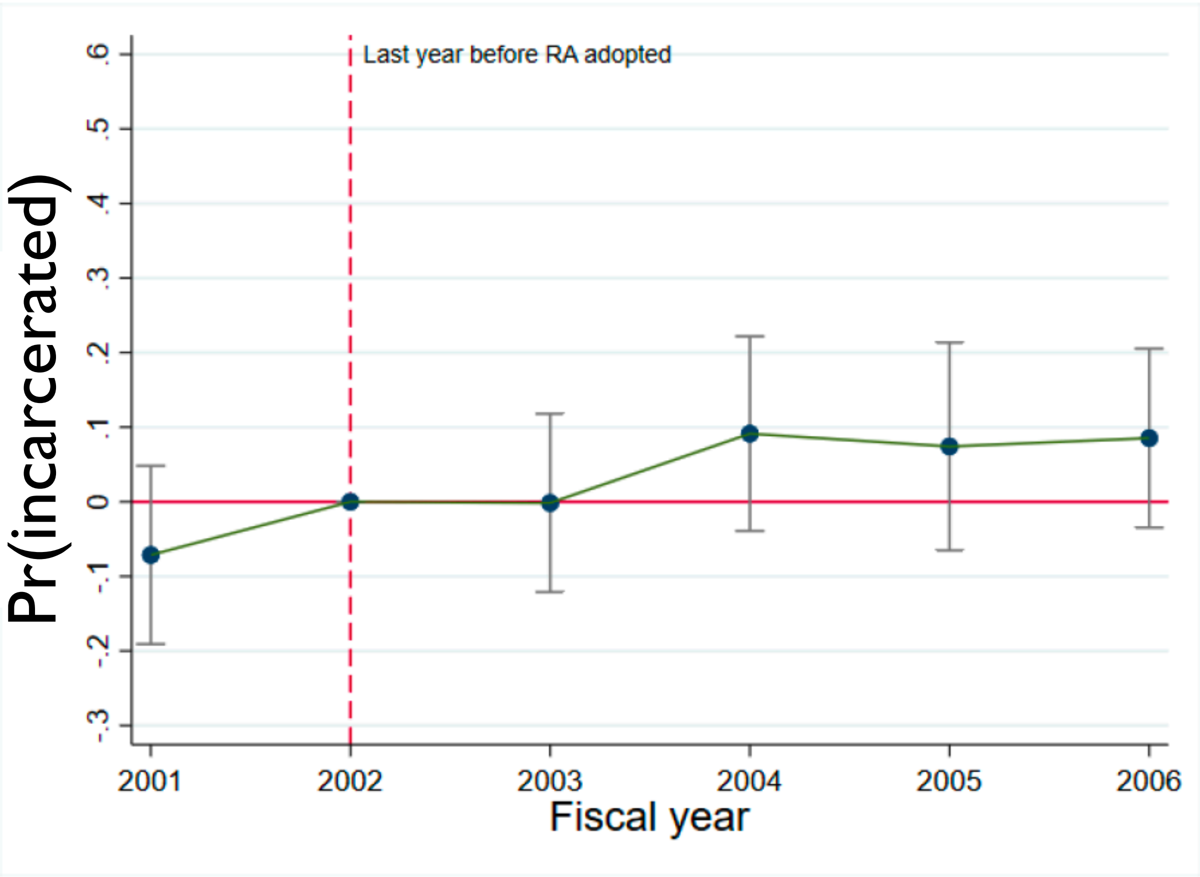
So what is the net effect of providing risk scores in the real world?

- This is where social scientists come in
- We need to measure causal effects on the outcomes we care about (e.g. sentencing disparities)
    - To do this, we need a randomized experiment or a natural experiment
- Let's turn back to Virginia, where risk scores were used to identify lowest-risk non-violent offenders…

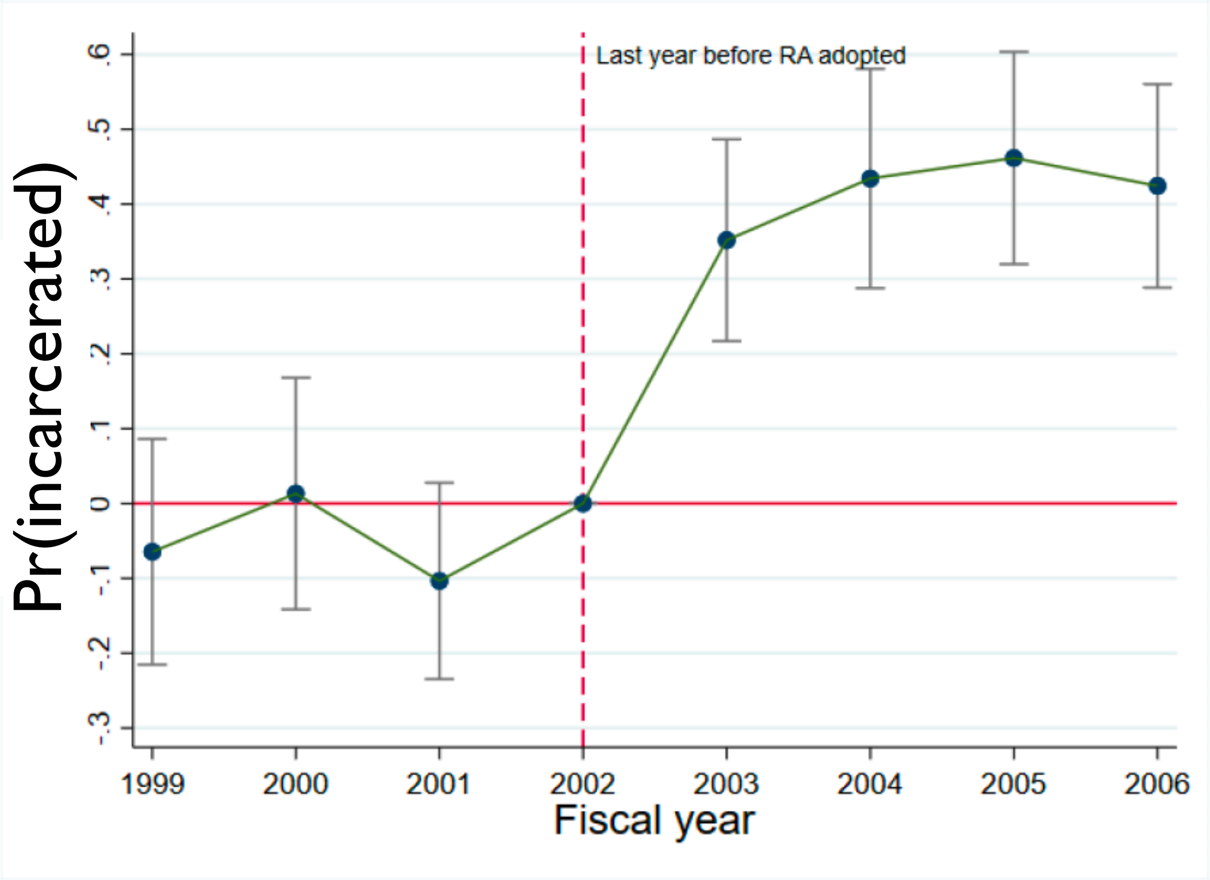
What should have happened?

- Simulation of what should have happened to sentencing for key groups if the risk assessment recommendations **replaced** judges' decisions:

Black defendants



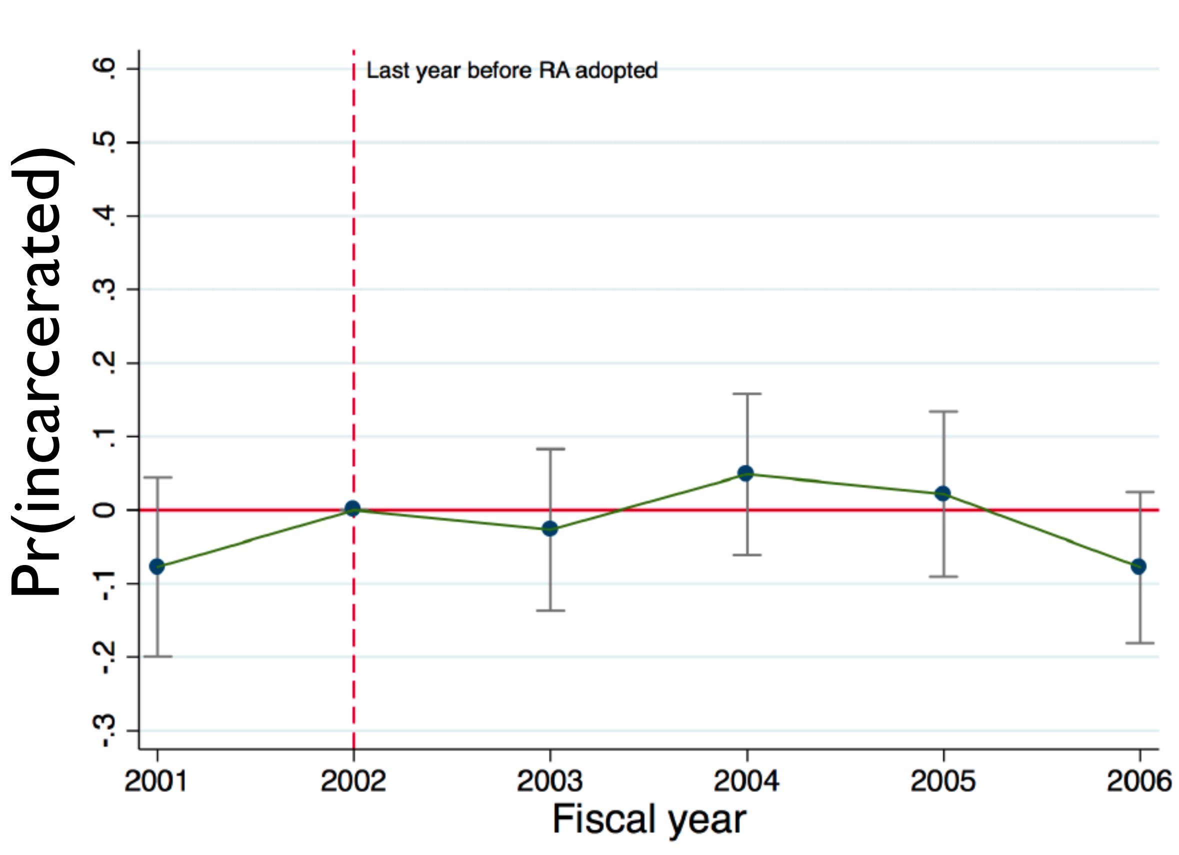
Young defendants



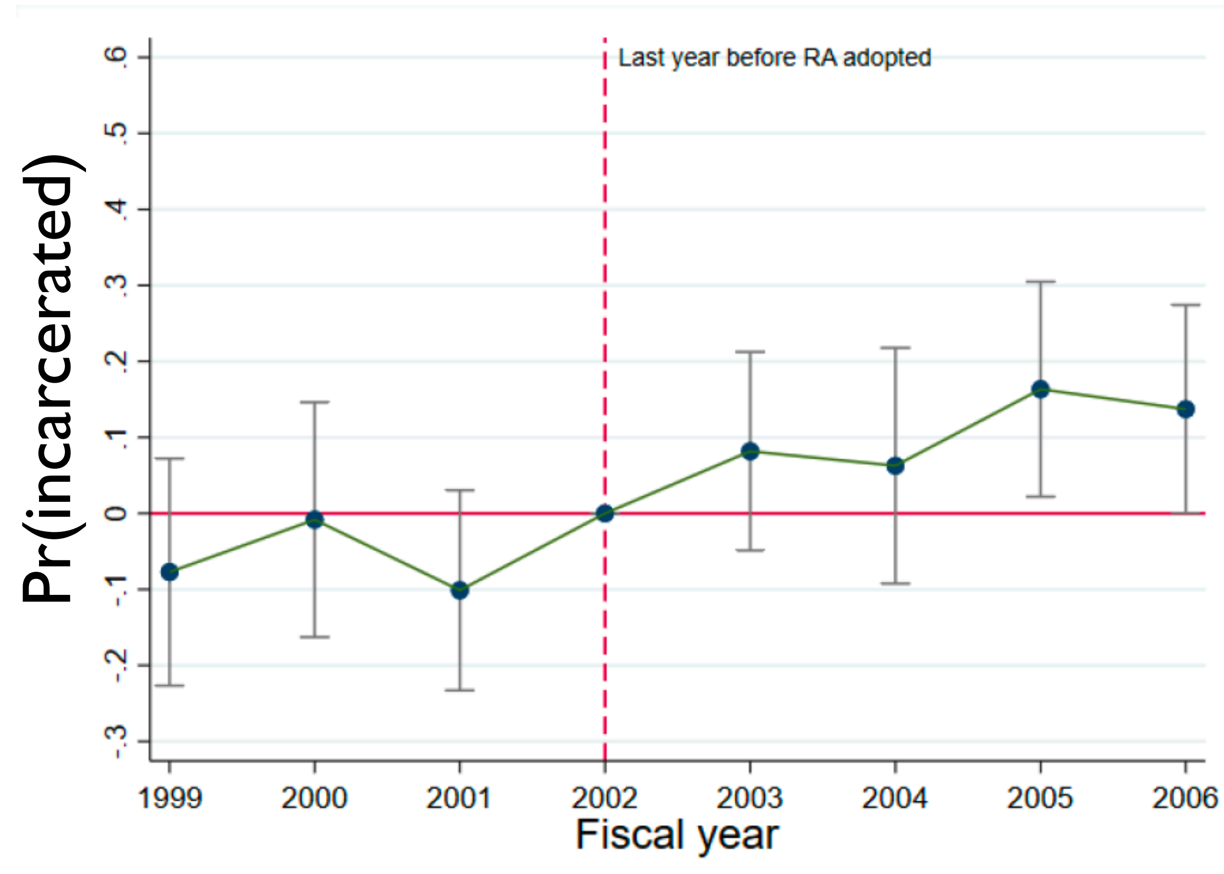
What actually happened?

- What actually happened to sentencing for key groups when the risk assessment recommendations **informed** judges' decisions:

Black defendants



Young defendants



Judges systematically deviate from the risk score recommendation

Panel A: Diverted | risk = low						
Alternative risk score	0.013					0.010
	(0.010)					(0.010)
Black		-0.015				-0.014
		(0.015)				(0.016)
Unemployed			0.025			0.009
			(0.017)			(0.018)
Female				0.040**		0.038**
				(0.016)		(0.017)
Age<23					0.069****	0.065***
					(0.020)	(0.020)
Observations	3943	3943	3943	3943	3943	3943
R^2	0.204	0.204	0.204	0.205	0.206	0.280
Mean DV	0.44	0.44	0.44	0.44	0.44	0.44

Panel B: Diverted | risk = high						
Alternative risk score	-0.004					-0.007
	(0.005)					(0.005)
Black		-0.029***				-0.045****
		(0.010)				(0.012)
Unemployed			0.043****			0.018
			(0.012)			(0.012)
Female				0.038***		0.040***
				(0.013)		(0.014)
Age<23					0.065****	0.058****
					(0.011)	(0.011)
Observations	7598	7598	7598	7598	7598	7598
R^2	0.142	0.143	0.144	0.143	0.146	0.197
Mean DV	0.16	0.16	0.16	0.16	0.16	0.16

Judges' bias affects when they pay attention to the risk scores

- When deciding whom to divert from incarceration:
  - They are more likely to follow the low-risk recommendation for female and younger defendants
  - They are more likely to deviate from the high-risk recommendation for white, female, and younger defendants

- **Punchline:** Even if the risk scores are perfectly fair, the way judges implement them may not be

Are judges actually making prediction errors?

- Simulated vs. actual results for young people raises the question of whether judges were actually making prediction errors in the absence of algorithmic risk scores

- Are they getting it wrong? Or do they simply have competing objectives?
    - Reluctance to incarcerate young defendants is in line with long-standing view that youth is a mitigating factor — young people are viewed as less culpable for their crimes
    - Most of the anticipated efficiency gains from the risk assessment would have come from locking up these young defendants
    - **Perhaps judges knew all along that young defendants were high-risk**, but they chose not to incarcerate them
        - Risk scores push them a bit in this direction, but is that what we want? Is this fair?

Where do we go from here?

- So far existing work finds little/no evidence of efficiency gains from algorithms, and some red flags with respect to how the use of the algorithms affects fairness (race/age disparities)
- **Important driver of real-world effects is how humans use the predictions**
    - We're hoping that algorithms will correct biases in human decision-making
    - But those biases (1) may be smaller than we think, and (2) affect when they defer to the algorithm's recommendation
    - Competing objectives (e.g. leniency toward young people, concern about public backlash, desire to be reelected/reappointed) will affect how judges and prosecutors use these tools
        - Real-world effects are difficult to predict

- Research frontier: How do we implement these tools in a manner that moves us closer to our societal goals?
    - To figure this out, it will be crucial to implement algorithms in a way that enables rigorous evaluation
    - Important area for social scientists and computer scientists to collaborate going forward!

Thank you!

Email: jdoleac@tamu.edu