

11 - Prediction Policy Problems

ml4econ, HUJI 2020

Itamar Caspi

June 14, 2020 (updated: 2020-06-14)

Outline

- Prediction Policy Problems
- Algorithmic Fairness

Prediction Policy Problems

Prediction policy problems

What we've covered in the previous two lectures focused on evaluating policy based on causal inference and treatments effects.

Nevertheless, as Kleinberg, Ludwig, Mullainathan, and Obermeyer (AER 2015) point out, some policy decisions solely depend on prediction:

- Which teacher is best? (Hiring, promotion)
- Unemployment spell length? (Savings)
- Risk of violation of regulation (Health inspections)
- Riskiest youth (Targeting interventions)
- Creditworthiness (Granting loans)

Illustration

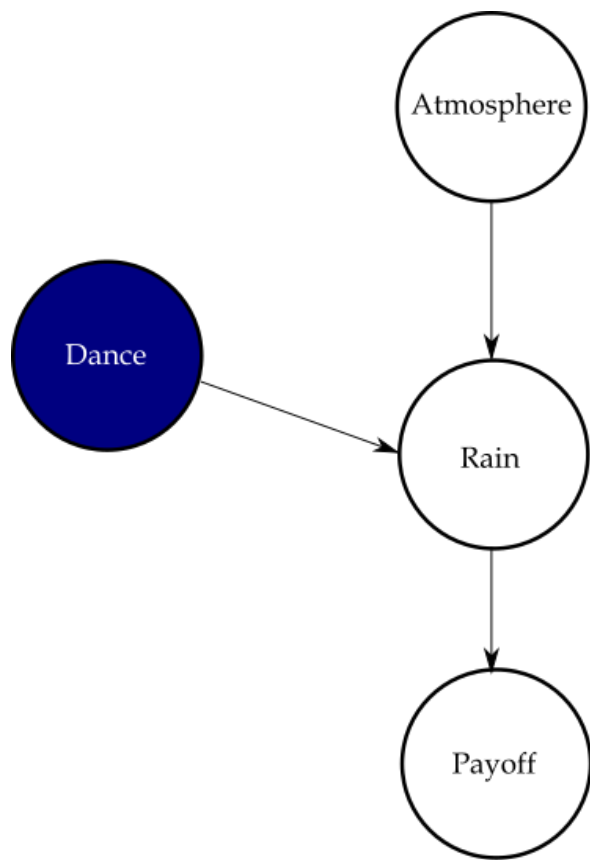
Consider the following toy example from Kleinberg, Ludwig, Mullainathan, and Obermeyer (AER 2015):

- $Y = \{\text{rain, no rain}\}$
- X atmospheric conditions
- D is a binary policy decision
- $\Pi(Y, D)$ payoff (utility)

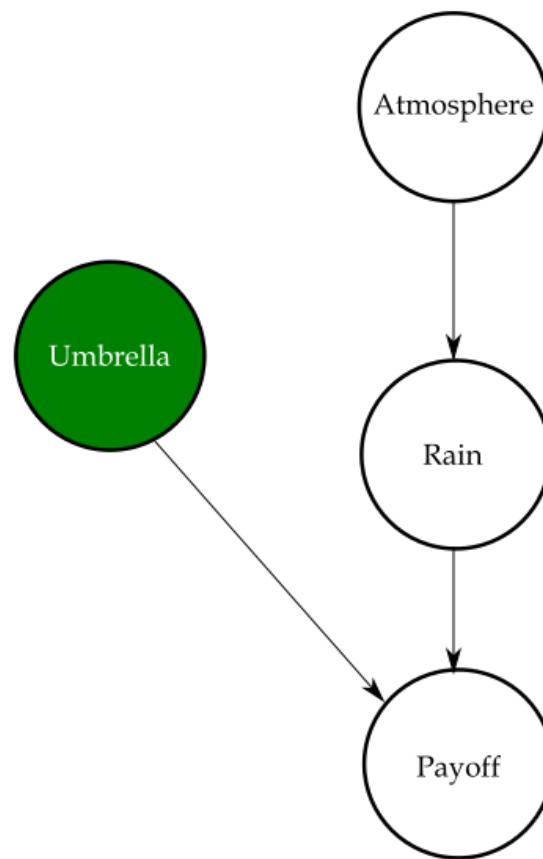
The change in payoff resulting from a policy decision is given by

$$\frac{\partial \Pi}{\partial D} = \underbrace{\frac{\partial \Pi}{\partial D} (Y)}_{\text{prediction}} + \frac{\partial \Pi}{\partial Y} \underbrace{\frac{\partial Y}{\partial D}}_{\text{causation}}$$

Rain dance vs. umbrella



CAUSATION



PREDICTION

Real life prediction policy problem: Joint replacement

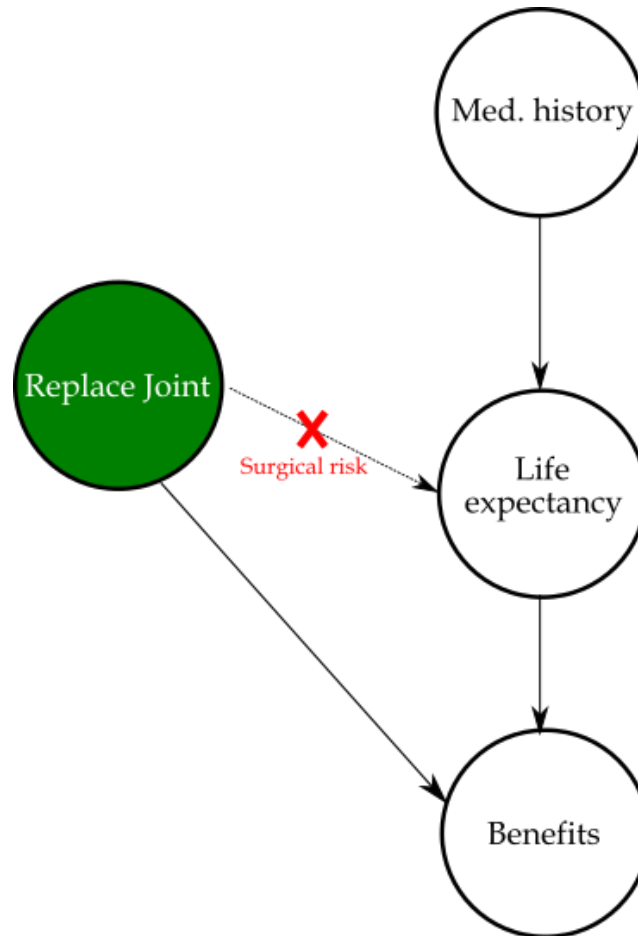
Over 750,000 joint replacements every year in the US.

- Benefits: Improved mobility and reduced pain.
- Costs: \sim \$15, 000 + painful recovery from surgery.

Working assumption: Benefits Π depend on longevity Y .

QUESTION: Can we improve resource allocation by predicting which surgeries will be futile using data available at the time of the surgery?

Joint replacement DAG



Note: Kleinberg et al. (2015) abstract from surgical risk.

The data

- A 20% sample of 7.4 million Medicare beneficiaries, 98,090 (1.3%) of which had a claim for joint replacement in 2010.
- 1.4 percent of this sample die in the month after surgery, potentially from complications of the surgery itself, and 4.2 percent die in the 1–12 months after surgery.
- Average mortality rate $\sim 5\%$ - *on average*, surgeries are not futile.
- This is perhaps misleading. A more appropriate question is whether surgeries on the predictably riskiest patients were futile.

Predicting mortality risk

Kleinberg et al. setup:

- *Outcome*: mortality in 1-12 months
- *Features*: Medicare claims dated prior to joint replacement, including patient demographics (age, sex, geography); co-morbidities, symptoms, injuries, acute conditions, and their evolution over time; and health-care utilization.
- *Sample*: training sample 65K observations / test sample, 33K observations
- *ML algorithm*: Lasso

The play book:

- Put beneficiaries from the test-set into percentiles by model predicted mortality risk.
- Attache to each percentile its corresponding share of surgeries.
- Show that an algorithm can do better then physicians.

The riskiest people receiving joint replacement

Predicted Mortality Percentile	Observed Mortality Rate	Total Number Annually
1	0.562 (.027)	4905
2	0.530 (.02)	9810
5	0.456 (.012)	24525
10	0.345 (.008)	49045
20	0.228 (.005)	98090
30	0.165 (.004)	147135
100	0.057 (.001)	490450

How to read this table

- column 1: Percentile by model predicted mortality risk
- column 2: Actual 1-12 months mortality risk
- column 3: Annual number of surgeries

For example, The riskiest beneficiaries go through 4,905 surgeries even though 56.2% of them are expected to die within 1-12 months, i.e., more than half of these surgeries are expected to be futile.

Source: Kleinberg et al. (2015).

Can an algorithm do better than physicians?

The first econometric challenge in prediction policy:

■ **Selective labels:** We only observe those who got surgery.

How to construct a counterfactual?

- Look at patients eligible for replacement who didn't get it.
- Working assumption: Physicians allocate replacements to the least risky first.

So, can the lasso beat doctors?

Predicted Mortality Percentile	Observed Mortality Rate	Total Number Annually	Substitute with 50th percentile Eligibles	
			Futile Procedures Averted	Annual Savings (in millions)
1	0.562 (.027)	4905	2403	36
2	0.530 (.02)	9810	4485	67
5	0.456 (.012)	24525	9398	141
10	0.345 (.008)	49045	13350	200
20	0.228 (.005)	98090	15219	228
30	0.165 (.004)	147135	13548	203
100	0.057 (.001)	490450		

Source: Kleinberg et al. (2015).

Simulation: substitute riskiest recipients with those who might have benefited from joint replacement procedures (drawn from *median* predicted risk) under Medicare eligibility guidelines, but did not receive treatment.

For example: *"Replacing the riskiest 10 percent with lower-risk eligibles would avert 13,350 futile surgeries and reallocate the 200 million per year to people who benefit from the surgery, at the [...] cost of postponing joint replacement for 35,695 of the riskiest beneficiaries who would not have died."* (Kleinberg et al., 2015)

What can go wrong?

The second econometric issue with prediction policy problems:

Omitted payoff bias: What if the unobserved objective (payoff) physicians see include other variables?

EXAMPLE: Patients with high mortality benefit most in terms of pain reduction.

We can assess omitted payoff bias based on observable post-replacement outcomes, e.g., number of visits to physicians for osteoarthritis and multiple claims for physical therapy or therapeutic joint injections.

Higher mortality-risk shows no sign of higher benefits.

Predicted Mortality Percentile	Observed Mortality Rate	Total Number Annually	PT + Joint Injections	Physician Visits for Osteo.
1	0.562 (.027)	4905	4.4 (.356)	1.4 (.173)
2	0.530 (.02)	9810	4.0 (.316)	1.8 (.13)
5	0.456 (.012)	24525	3.9 (.208)	2.0 (.092)
10	0.345 (.008)	49045	3.8 (.143)	2.1 (.066)
20	0.228 (.005)	98090	3.9 (.091)	1.8 (.042)
30	0.165 (.004)	147135	3.8 (.076)	1.9 (.035)
100	0.057 (.001)	490450	3.9 (.046)	2.1 (.023)

Leveling the playing field

Before we can compare decision making made by humans and ML algorithms we need to make sure we've leveled the playing field.

In particular, a fair comparison human decision makers and algorithms necessitates that both "competitors" have:

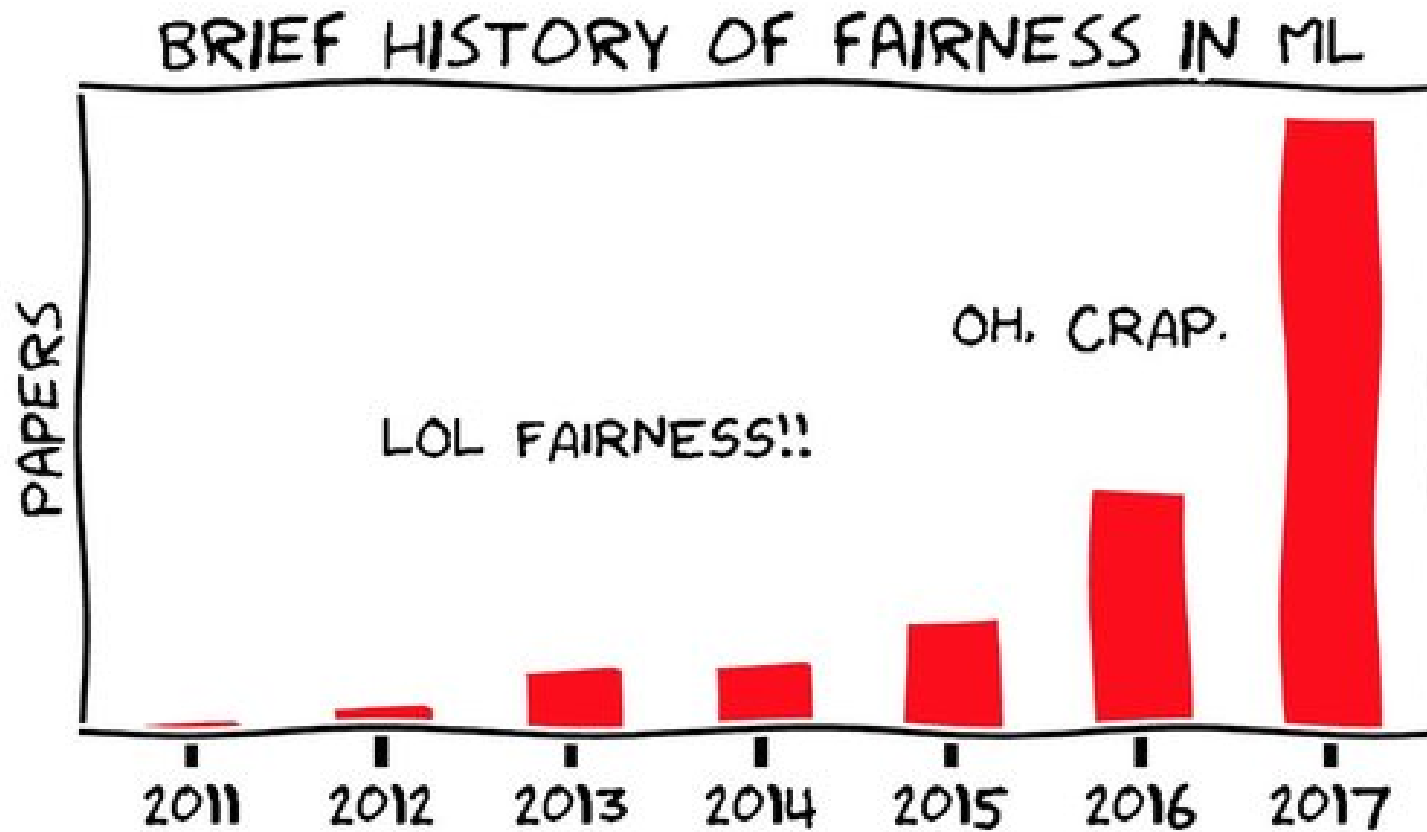
- Access to the entire pre-decision information set.
 - The same objective function.
-

Main takeaways from prediction policy problems

- Better ML models do not necessarily imply better decision making
- Insights from the social sciences can help analyze and improve algorithmic decision making.

Algorithmic Fairness

Fairness??






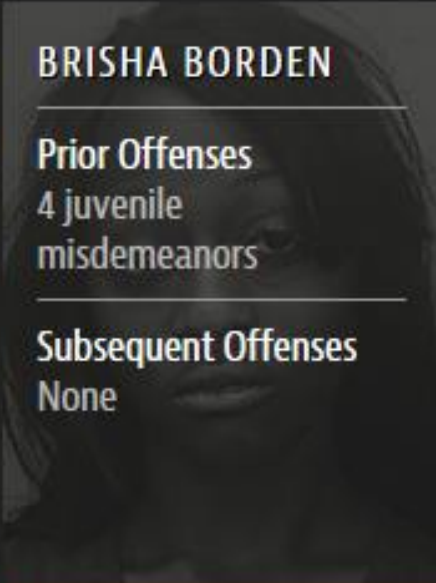
Source: <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

ProPublica and COMPASS

"**Machine Bias**", by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica May 23, 2016

COMPAS, Correctional Offender Management Profiling for Alternative Sanction, is a software, developed by Northpointe which includes algorithm used to assess potential *recidivism* risk.

Bias in recidivism risk assessment?

Two Petty Theft Arrests		Two Petty Theft Arrests	
			
VERNON PRATER	BRISHA BORDEN	VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors	Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None	Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8	LOW RISK 3	HIGH RISK 8
<i>Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.</i>		<i>Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.</i>	

COMPAS accuracy rates

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Desired fairness properties

Fairness Properties for Risk Assignments (Kleinberg, Mullainathan, and Raghavan ,2017):

- (A) *Calibration within groups*, e.g., the score returned from COMPAS for a defendant should reflect the probability of re-offending.
- (B) *Balance for the negative class*: false positive rates must be equal for all groups, i.e., the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.
- (C) *Balance for the positive class*, e.g., predicted recidivism risk assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.























An Impossibility theorem of fairness

Kleinberg, Mullainathan, and Raghavan (2017):

Theorem 1.1: *Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions (A), (B), and (C). Then the instance must either allow for perfect prediction [...] or have equal base rates.*

Illustration











Perfect prediction for group **a**, 1 false positive for group **b**

Group	a							b				Unequal base rates
Outcome												
Predictor												

Source: <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

Illustration

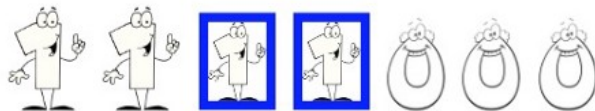
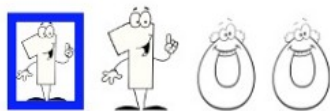
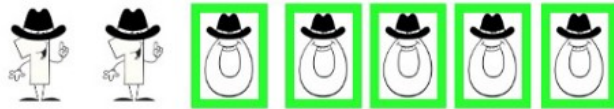
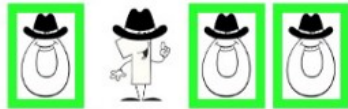
Now, the true positive rates as well as true negative rates are equal for both groups (both have 1/2 and 1):

Group	a	b	
Outcome	 	 	Unequal base rates
Predictor	  	  	

Source: <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

Illustration

Now, balance for the negative class is violated with this setting.

Group	a	b
Outcome		 Unequal base rates
Predictor		
NPV	$\frac{2}{5}$	$\frac{1}{3}$

Source: <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

Back to COMPAS

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Northpointe's main defense is that risk scores reflect observed probability of re-offending.

Blind Algorithms

Algorithmic Fairness (Kleinberg, Ludwig, Mullainathan, and Rambachan, AER 2018):

Can we increase algorithmic fairness by ignoring variables that induce such bias such as race, age, sex, etc.?

Short answer: Not necessarily.

The basic setup

The context: Student admission to college.

Data: $\{Y_i, X_i, R_i\}_{i=1}^N$, where

- Y_i is performance
- X_i is a set of features
- R_i is a binary race indicator where $R_i = 1$ for individuals that belong to the minority group and $R_i = 0$ otherwise.

Predictors:

- **"Aware"**: $\hat{f}(X_i, R_i)$
- **"Blind"**: $\hat{f}(X_i)$
- **"Orthogonality"**: $\hat{f}(\widetilde{X}_i)$, where $\widetilde{X}_i \perp R_i$.

Definitions

Let S denote the set of admitted students and $\phi(S)$ denote a function that depends only on the predicted performance, measured by \hat{f} , of the students in S .

Compatibility condition: If S and S' are two sets of students of the same size, sorted in descending order of predicted performance $\hat{f}(X, R)$, and the predicted performance of the i^{th} student in S is at least as large as the predicted performance of the i^{th} in S' for all i , then $\phi(S) \geq \phi(S')$.

- The *efficient* planner maximizes $\phi(S)$ where $\phi(S)$ is compatible with \hat{f} .
- The *equitable* planner seeks to maximize $\phi(S) + \gamma(S)$, where $\phi(S)$ is compatible with \hat{f} , and $\gamma(S)$ is monotonically increasing in the number of students in S who have $R = 1$.

Main result: Keep R in

Kleinberg et al. (2018) main result:

THEOREM 1: *For some choice of K_0 and K_1 with $K_0 + K_1 = K$, the equitable planner's problem can be optimized by choosing the K_0 applicants in the $R = 0$ group with the highest $\hat{f}(X, R)$, and the K_1 applicants in the $R = 1$ group with the highest $\hat{f}(X, R)$.*

(See Kleinberg et al., 2018 for a sketch of the proof.)

Intuition

- Good ranking of applicants is desired for both types of planners.
 - Equitable planners still care about ranking *within* groups.
 - Achieving a more balanced acceptance rate is a *post* prediction step. Can be adjusted by changing the group-wise threshold.
-

Illustration of the result

Say that we have 10 open slots, 100 admissions from the majority group ($R = 0$) and 20 from the minority group ($R = 1$). In addition, assume that the acceptance rate for the minority group is set to 30%.

An equitable planner should:

1. Rank candidates within each group according to $\hat{f}(X_i, R_i)$.
2. Accept the top 7 from the $R = 0$ group, and top 3 from the $R = 1$ group.

Empirical application

DATA: Panel data on This representative sample of students who entered eighth grade in the fall of 1988, and who were then followed up in 1990, 1992, 1994, and mid-20s).

OUTCOME: $\text{GPA} \geq 2.75$.

FEATURES: High school grades, course taking patterns, extracurricular activities, standardized test scores, etc.

RACE: White ($N_0 = 4,274$) and black ($N_1 = 469$).

PREDICTORS: OLS (random forest for robustness)

RESULT: The "aware" predictor dominates for both efficient planner and equitable planner.

Sources of disagreement

- On the right: The distribution of black students in the sample across predicted-outcome deciles according to the race-blind or race-aware predictors.
- How to read this: In the case of agreement between race-blind and race-aware, the values would be aligned on the main diagonal. By contrast, disagreement is characterized by off-diagonal non-zero values.
- Bottom line (again): Adding race to the equation improves within group ranking.

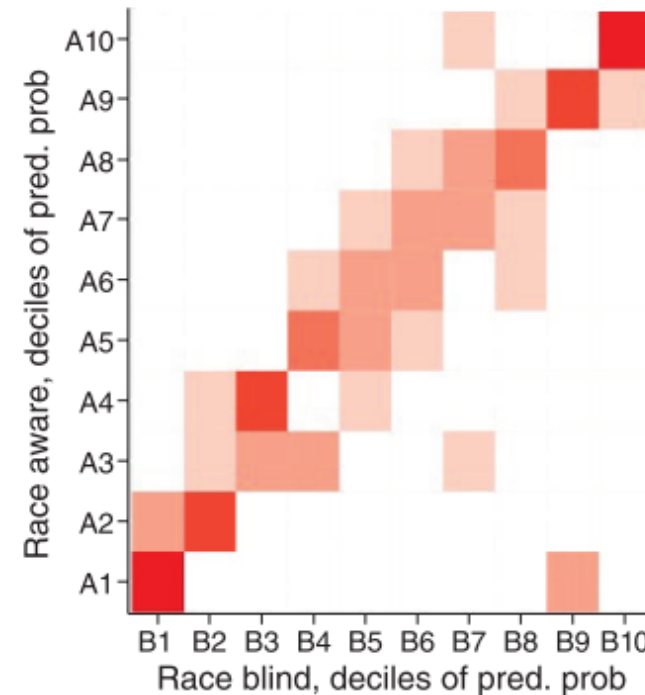


FIGURE 2. HEATMAP OF RANKINGS OF BLACK APPLICANTS BY PREDICTED PROBABILITY OF GPA < 2.75, USING RACE-AWARE VERSUS RACE-BLIND ALGORITHMS

Main takeaways

- Turning algorithms blind might actually do harm.
 - What actually matters is the rankings within groups.
 - Caveat: This is a very specific setup and source of bias.
-

```
slides %>% end()
```

 [Source code](#)

Selected references

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine Bias.” *ProPublica*, May 23. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Athey, S. (2018). The Impact of Machine Learning on Economics.

Athey, S., & Wager, S. (2018). Efficient Policy Learning.

Kleinberg, B. J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review: Papers & Proceedings*, 105(5), 491–495.

Kleinberg, B. J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic Fairness. *American Economic Review: Papers & Proceedings*, 108, 22–27.

Selected references

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics*, 133(1), 237–293.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of the 8th Conference on Innovation in Theoretical Computer Science*, 43, 1–23.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87–106.