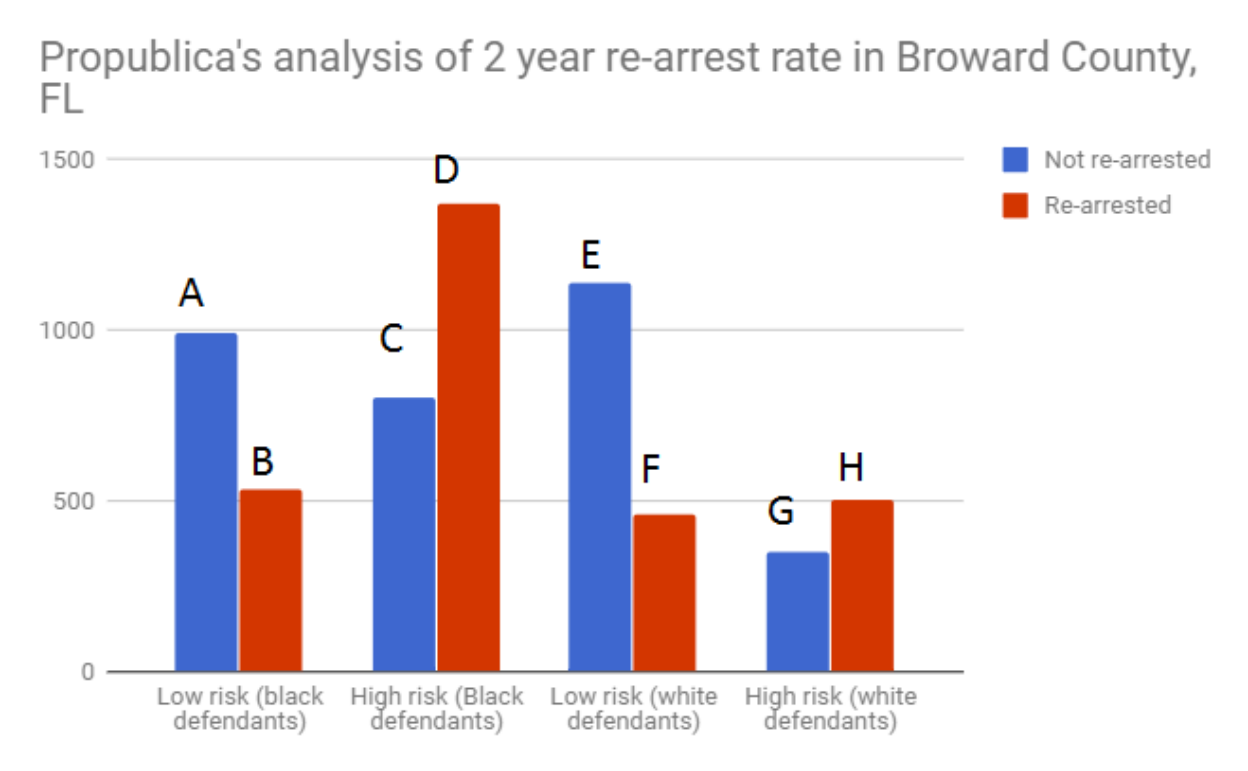
Algorithms week 5-2:
Algorithmic Accountability

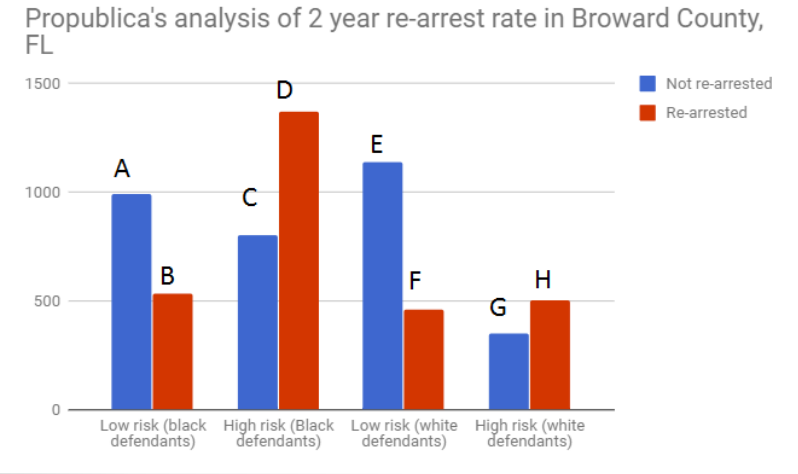
Jonathan Stray
Columbia Lede Program
August 15, 2018

	Low risk (black d	High risk (Black (	Low risk (white d	High risk (white defendants)
Not re-arrested	990	805	1139	349
Re-arrested	532	1369	461	505



Stephanie Wykstra, personal communication

	Low risk (black d	High risk (Black (	Low risk (white d	High risk (white defendants)
Not re-arrested	990	805	1139	349
Re-arrested	532	1369	461	505



ProPublica argument

False positive rate

P(high risk | black, no arrest) = C/(C+A) = 0.45

P(high risk | white, no arrest) = G/(G+E) = 0.23

False negative rate

P(low risk | black, arrested) = B/(B+D) = 0.28

P(low risk | white, arrested) = F/(F+H) = 0.48

Northpointe response

Positive predictive value

P(arrest | black, high risk) = D/(C+D) = 0.63

P(arrest | white, high risk) = H/(G+H) = 0.59

Quantitative Definitions of Fairness

. . .

Three criteria commonly proposed as statistical definition of "fairness." They are mutually exclusive!

"Independence" or "demographic parity". The classifier predicts the same number of people in each group.

"Separation" or "equal error rates". The classifier has the same false positive rate / true positive rate for each group.

"Sufficiency" or "calibration." When classifier predicts true, both groups have the same probability of having a true outcome.

The Impossibility result: With different base rates, only one of these criteria at a time is achievable (ref: Barocas and Hardt, NIPS 2017 tutorial)

“Independence” or “demographic parity”

The idea: the prediction should not depend on the group.

Same percentage of black and white defendants scored as high risk.
Same percentage of men and women hired.
Same percentage of rich and poor students admitted.

Mathematically: Equal rate of “true” prediction for all groups.
A classifier with this property: choose the 10 best scoring applicants in each group.
Drawbacks: Doesn’t measure who we accept, as long as we accept equal numbers in each group. The “perfect” predictor, which always guesses correctly, is considered unfair if the base rates are different.
Legal principle: disparate impact
Moral principle: equality of outcome

“Sufficiency” or “Calibration”

The idea: a prediction means the same thing for each group.

Same percentage of re-arrest among black and white defendants who were scored as high risk.
Same percentage of equally qualified men and women hired.
Whether you will get a loan depends only on your probability of repayment.

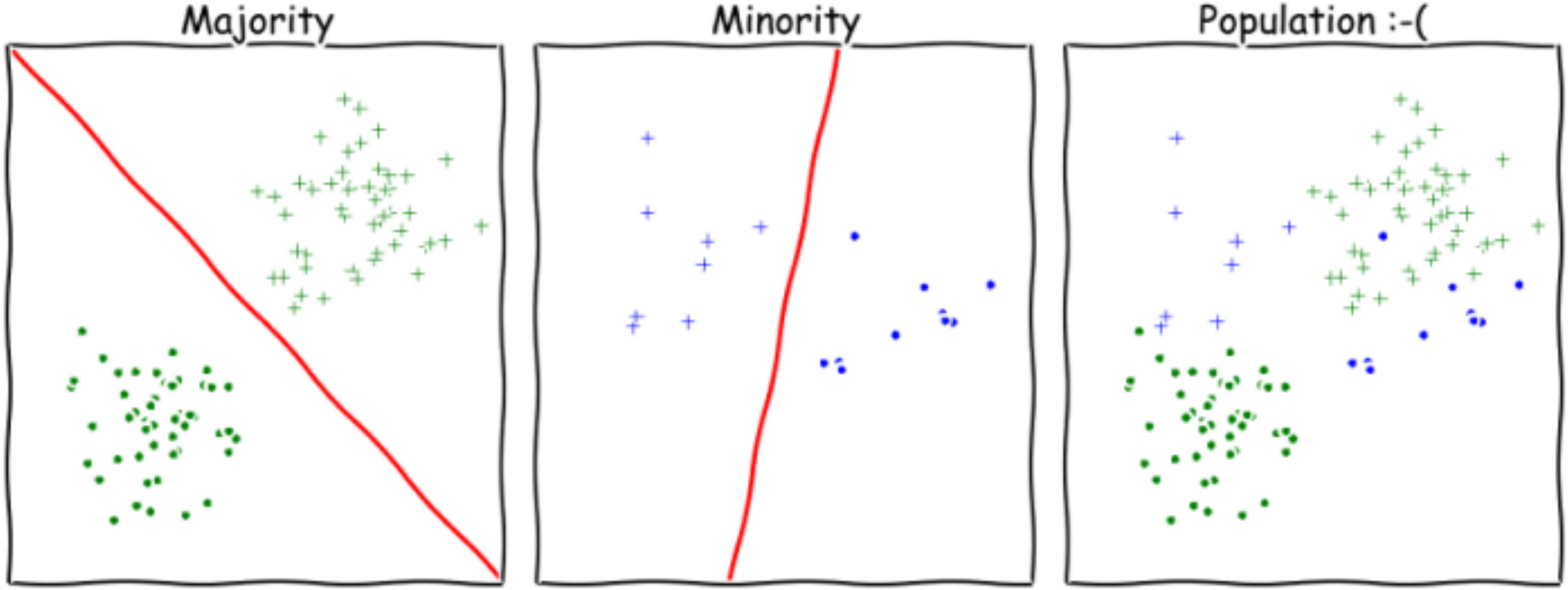
Mathematically: Equal PPV (Precision) for each group.
A classifier with this property: any standard machine learning algorithm.
Drawbacks: Disparate impacts may exacerbate existing disparities. Error rates may differ between groups in unfair ways.
Legal principle: disparate treatment
Moral principle: equality of opportunity

“Separation” or “Equal error rates”

The idea: Don’t let a classifier make most of its mistakes on one group.

Same percentage of black and white defendants who are not re-arrested are scored as high risk.
Same percentage of qualified men and women mistakenly turned down.
If you would have repaid a loan, you will be turned down at the same rate regardless of your income.

Mathematically: Equal false positive and true positive rates.
A classifier with this property: use different thresholds for each group.
Drawbacks: Classifier must use group membership explicitly. Predictive accuracy will differ between groups.
Legal principle: disparate treatment
Moral principle: equality of opportunity



Even if two groups of the population admit simple classifiers, the whole population may not

How Big Data is Unfair, Moritz Hardt

Algorithms as part of a system

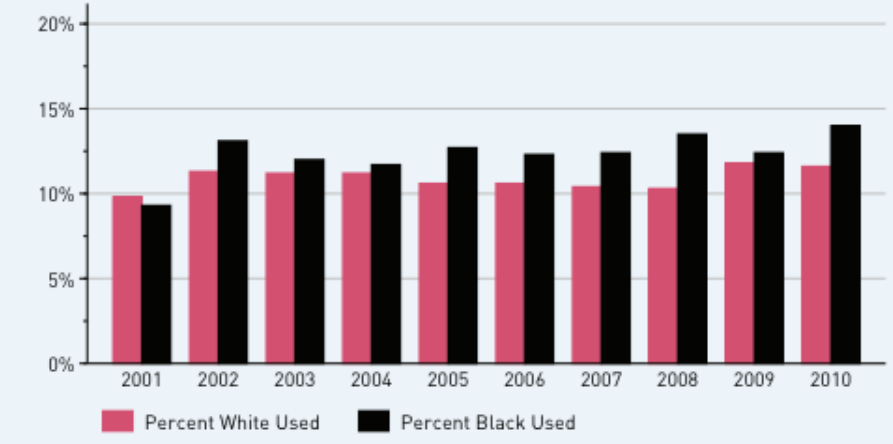
. . .

Data Quality

. . .

FIGURE 21

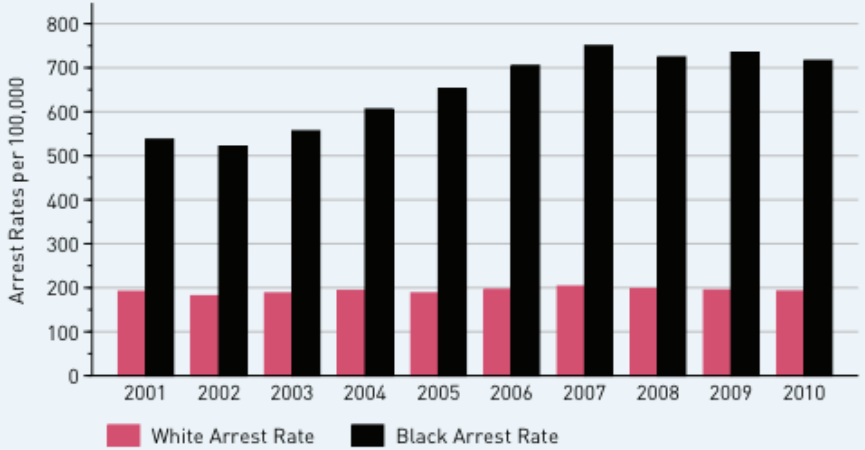
Marijuana Use by Race: Used Marijuana in Past 12 Months (2001-2010)



Source: National Household Survey on Drug Abuse and Health, 2001-2010

FIGURE 10

Arrest Rates for Marijuana Possession by Race (2001-2010)



Source: FBI/Uniform Crime Reporting Program Data and U.S. Census Data

The black/white marijuana arrest gap, in nine charts,
Dylan Matthews, Washington Post, 6/4/2013

A senior police official recently testified to the City Council that there was a simple justification — he said more people call 911 and 311 to complain about marijuana smoke in black and Hispanic neighborhoods

...

Robert Gebeloff, a data journalist at The Times, transposed Census Bureau information about race, poverty levels and homeownership onto a precinct map. Then he dropped the police data into four buckets based on the percentage of a precinct's residents who were black or Hispanic.

What we found roughly aligned with the police explanation. In precincts that were more heavily black and Hispanic, the rate at which people called to complain about marijuana was generally higher.

…

What we discovered was that when two precincts had the same rate of marijuana calls, the one with a higher arrest rate was almost always home to more black people. The police said that had to do with violent crime rates being higher in those precincts, which commanders often react to by deploying more officers.

Using Data to Make Sense of a Racial Disparity in NYC Marijuana Arrests, New York Times, 5/13/2018

The proportion of racial disparities in crime explained by differential participation versus differential selection is hotly debated

…

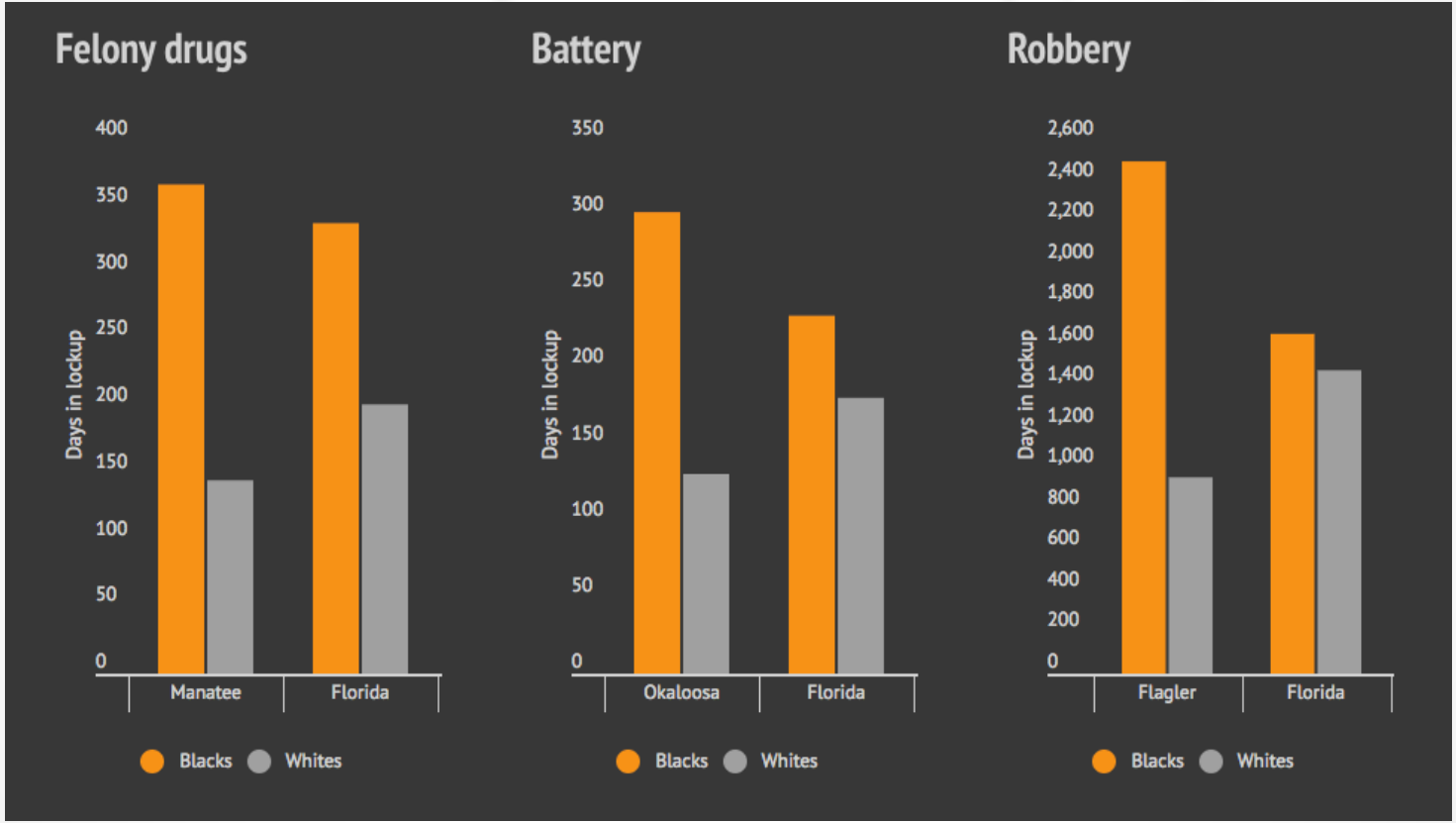
In our view, official records of arrest—particularly for violent offenses—are a valid criterion. First, surveys of victimization yield "essentially the same racial differentials as do official statistics. For example, about 60 percent of robbery victims describe their assailants as black, and about 60 percent of victimization data also consistently show that they fit the official arrest data" (Walsh, 2004: 29). Second, self-reported offending data reveal similar race differentials, particularly for serious and violent crimes (see Piquero, 2015).

Risk, Race, and Recidivism: Predictive Bias and Disparate Impact
Jennifer L. Skeem, Christopher T Lowenkamp, Criminology 54 (4) 2016

Transparency & Oversight

. . .

How are “points” used by judges?



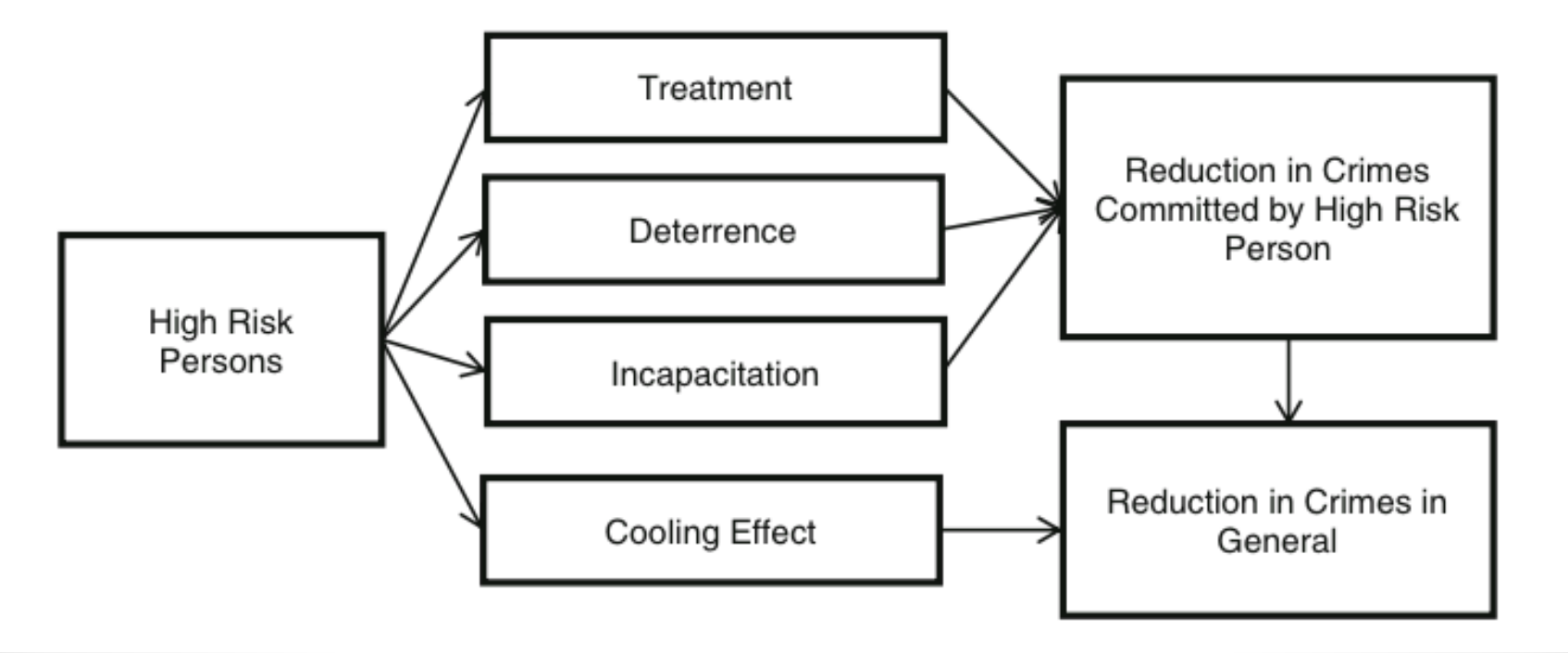
Bias on the Bench, Michael Braga, Herald Tribune

There are a number of interventions that can be directed at individual-focused predictions of gun crime because intervening with high-risk individuals is not a new concept. There is research evidence that targeting individuals who are the most criminally active can result in significant reductions in crime

…

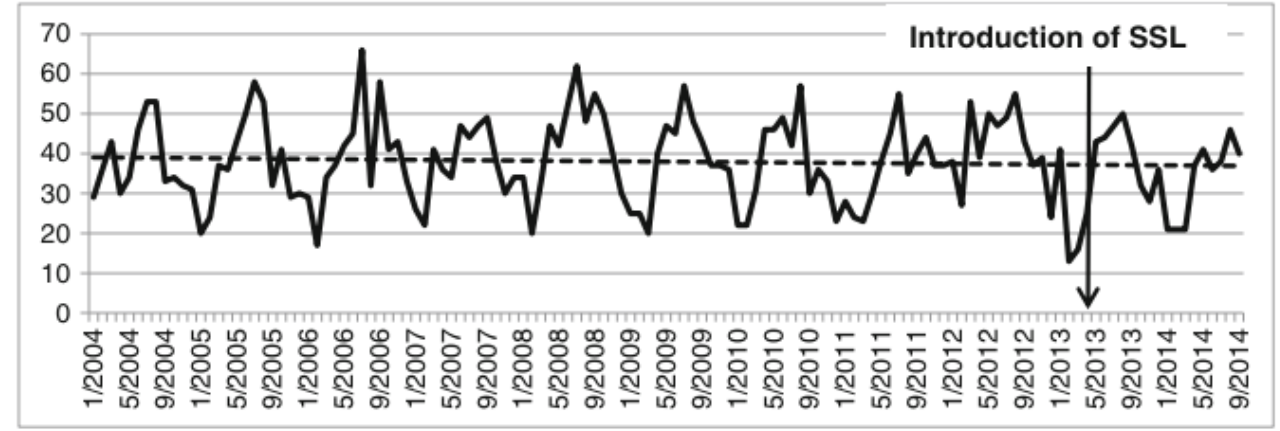
Conversely, some research shows that interventions targeting individuals can sometimes backfire. As an example, some previous proactive interventions, including increased arrest of individuals perceived to be at high risk (selective apprehension) and longer incarceration periods (selective incapacitation), have led to negative social and economic unintended consequences. Auerhahn (1999) found that a selective incapacitation model generated a large number of persons falsely predicted to be high-risk offenders, although it did reasonably well at identifying those who were low risk.

Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot, Saunders, Hunt, Hollywood, RAND, 2016



Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot, Saunders, Hunt, Hollywood, RAND, 2016

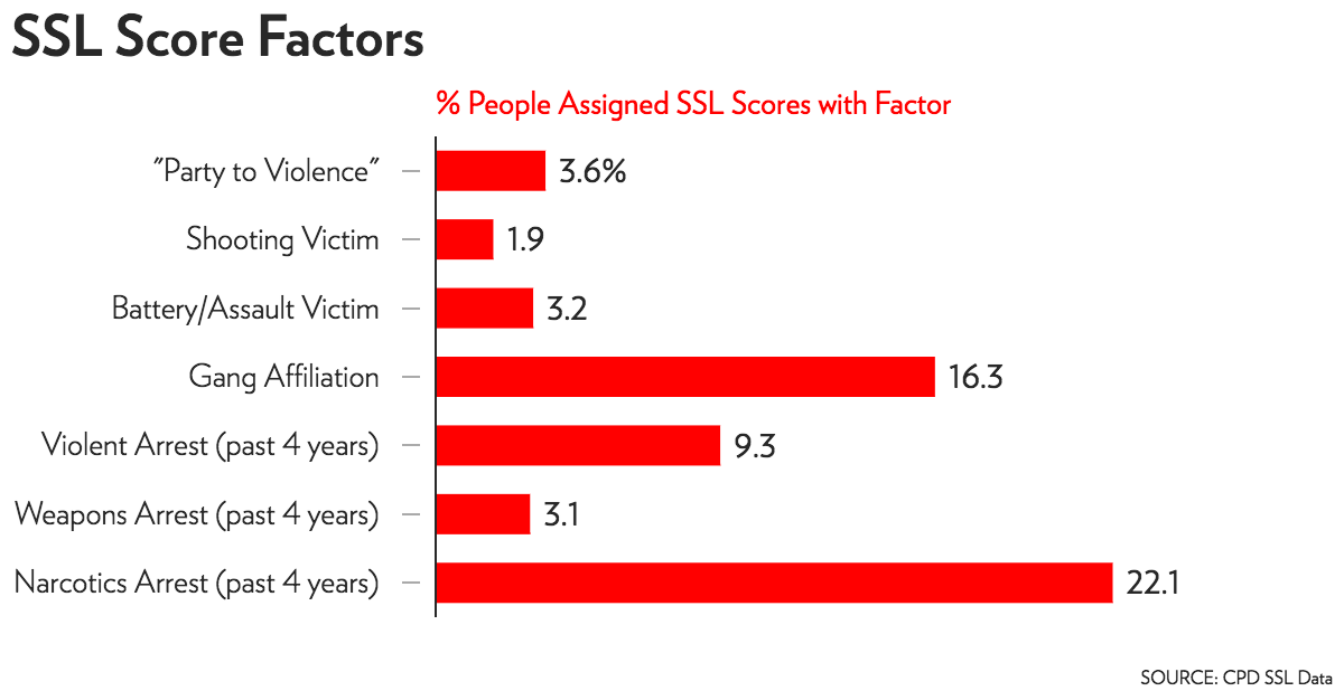
Once other demographics, criminal history variables, and social network risk have been controlled for using propensity score weighting and doubly-robust regression modeling, being on the SSL did not significantly reduce the likelihood of being a murder or shooting victim, or being arrested for murder. Results indicate those placed on the SSL were 2.88 times more likely to be arrested for a shooting



Monthly homicides in Chicago from January 2004 to September 2014

Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot, Saunders, Hunt, Hollywood, RAND, 2016

Reverse-engineering the SSL score



The contradictions of Chicago Police's secret list,
Kunichoff and Sier, Chicago Magazine 2017

Chicago Police Cut Crime with Major Upgrades to Analytics and Field Technology

Since deploying six Strategic Decision Support Centers across the city, Chicago saw a 21 percent drop in shootings last year.

The Chicago Police Department (CPD) is deploying predictive and analytic tools after seeing initial results and delivering on a commitment from Mayor Rahm Emanuel, a bureau chief said recently.

Last year, CPD created six Strategic Decision Support Centers (SDSCs) at police stations, essentially local nerve centers for its high-tech approach to fighting crime in areas where incidents are most prevalent.

…

Connecting features like predictive mapping and policing, gunshot detection, surveillance cameras and citizen tips lets police identify "areas of risk, and ties all these things together into a very consumable, very easy to use, very understandable platform," said Lewin.

"The predictive policing component … the intelligence analyst and that daily intelligence cycle, is really important along with the room itself, which I didn't talk about," Lewin said in an interview.

Theo Douglas, Government Technology, 2018

Why algorithmic decisions?

. . .

TABLE 1: Studies Included in Meta-Analysis

Citation	Prediction	Accuracy Statistic Reported	Accuracy		
			Clinical	Statistical	d^+
Adams (1974)[1]	Brain impairment	Hit rate	53	52	.02
Adams (1974)[2a]	Brain impairment	Hit rate	53	56	–.06
Alexakos (1966)[1]	Academic performance	Hit rate	39	51	–.24
Alexakos (1966)[2]	Academic performance	Hit rate	39	52	–.27
Astrup (1975)[b]	Psychiatric diagnosis	Hit rate	78	74	.09
Barron (1953)	Psychotherapy outcome	Hit rate	62	73	–.23
Blumetti (1972)	Length of psychotherapy	Hit rate	61	54	.15
Bolton et al. (1968)	Prognosis	Correlation	.35	.48	–.16
Carlin and Hewitt (1990)	Real vs. random MMPI profile	Hit rate	63	95	–.73
Conrad and Satter (1954)	Academic performance	Correlation	.36	.46	–.12
Cooke (1967a)	Psychiatric diagnosis	Hit rate	77	76	.02
Cooke (1967b)[a]	Psychiatric diagnosis	Correlation	.42	.51	–.11
Danet (1965)	Prognosis	Hit rate	64	70	–.13
Devries and Shneidman (1967)[a]	Matching MMPI profiles to persons	Hit rate	75	100	–.81
Dickerson (1958)	Compliance with counseling plan	Hit rate	57	52	.10
Dunham and Meltzer (1946)[b]	Length of hospital stay	Hit rate	38	58	–.40
Evenson, Altman, Sletten, and Cho (1975)	Length of hospital stay	Hit rate	64	71	–.14
Fero (1975)[1]	Prognosis	Correlation	.35	.57	–.29
Fero (1975)[2b]	Prognosis	Correlation	.35	.73	–.57
Gardner et al. (1996)[1b]	Offense / violence	Hit rate	62	74	–.25
Gardner et al. (1996)[2]	Offense or violence	Hit rate	62	71	–.20
Gardner et al. (1996)[3]	Offense or violence	Hit rate	62	70	–.17
Gardner et al. (1996)[3]	Offense or violence	Hit rate	62	70	–.17

From The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction, Ægisdóttir et al.

Abstract

Can machine learning improve human decision making? Bail decisions provide a good test case. Millions of times each year, judges make jail-or-release decisions that hinge on a prediction of what a defendant would do if released. The concreteness of the prediction task combined with the volume of data available makes this a promising machine-learning application. Yet comparing the algorithm to judges proves complicated. First, the available data are generated by prior judge decisions. We only observe crime outcomes for released defendants, not for those judges detained. This makes it hard to evaluate counterfactual decision rules based on algorithmic predictions. Second, judges may have a broader set of preferences than the variable the algorithm predicts; for instance, judges may care specifically about violent crimes or about racial inequities. We deal with these problems using different econometric strategies, such as quasi-random assignment of cases to judges. Even accounting for these concerns, our results suggest potentially large welfare gains: one policy simulation shows crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates. Moreover, all categories of crime, including violent crimes, show reductions; and these gains can be achieved while simultaneously reducing racial disparities. These results suggest that while machine learning can be valuable, realizing this value requires integrating these tools into an economic framework: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals. *JEL* Codes: C10 (Econometric and statistical methods and methodology), C55 (Large datasets: Modeling and analysis), K40 (Legal procedure, the legal system, and illegal behavior)

Human Decisions and Machine Predictions, Kleinberg et. al. 2017

Machine learning in lending

. . .

None of the new start-ups are consumer banks in the full-service sense of taking deposits. Instead, they are focused on transforming the economics of underwriting and the experience of consumer borrowing — and hope to make more loans available at lower cost for millions of Americans.

…

They all envision consumer finance fueled by abundant information and clever software — the tools of data science, or big data — as opposed to the traditional math of creditworthiness, which relies mainly on a person's credit history.

…

The data-driven lending start-ups see opportunity. As many as 70 million Americans either have no credit score or a slender paper trail of credit history that depresses their score, according to estimates from the National Consumer Reporting Association, a trade organization. Two groups that typically have thin credit files are immigrants and recent college graduates.

Banking startups adopt new tools for lending,
Steve Lohr, New York Times

The essential insight of our paper is that a more sophisticated statistical technology (in the sense of reducing predictive mean squared error) will, by definition, produce predictions with greater variance. Put differently, improvements in predictive technology act as mean-preserving spreads for predicted outcomes—in our application, predicted default propensities on loans.[4] This means that there will always be some borrowers considered less risky by the new technology ("winners"), while other borrowers will be deemed riskier ("losers"), relative to their position in equilibrium under the pre-existing technology. The key question is then how these winners and losers are distributed across societally important categories such as race, age, income, or gender.

Predictably Unequal? The Effects of Machine Learning on Credit Markets, Fuster et al

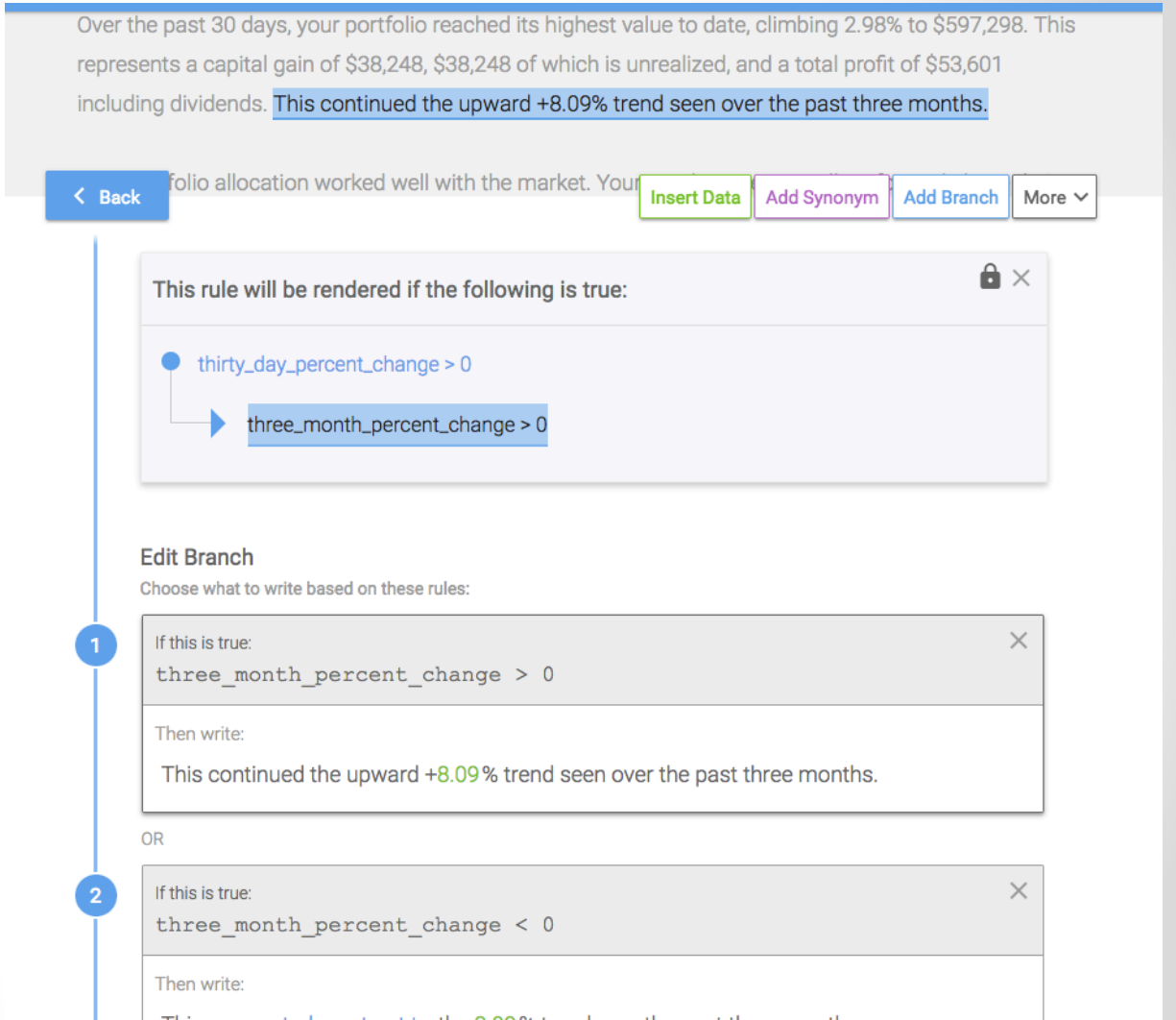
We compute counterfactual equilibria associated with each statistical technology on a subset of our data (loans originated in 2011, in this version of the paper), and then compare the resulting equilibrium outcomes with one another to evaluate comparative statics on outcomes across groups. We find that the machine learning model appears to provide a slightly larger number of borrowers access to credit, and marginally reduces disparity in acceptance rates (i.e., the extensive margin) across race and ethnic groups in the borrower population. However, the story is different on the intensive margin. Here, the cross-group disparity of equilibrium rates increases significantly (by 23%) under the machine learning model relative to the less sophisticated logistic regression models. This is also accompanied by a substantial increase in within-group dispersion in equilibrium interest rates as technology improves—it rises significantly more for Black and Hispanic borrowers under the machine learning model than it does for White non-Hispanic borrowers, i.e., Black and Hispanic borrowers get very different rates from one another under the machine learning technology.

Predictably Unequal? The Effects of Machine Learning on Credit Markets, Fuster et al

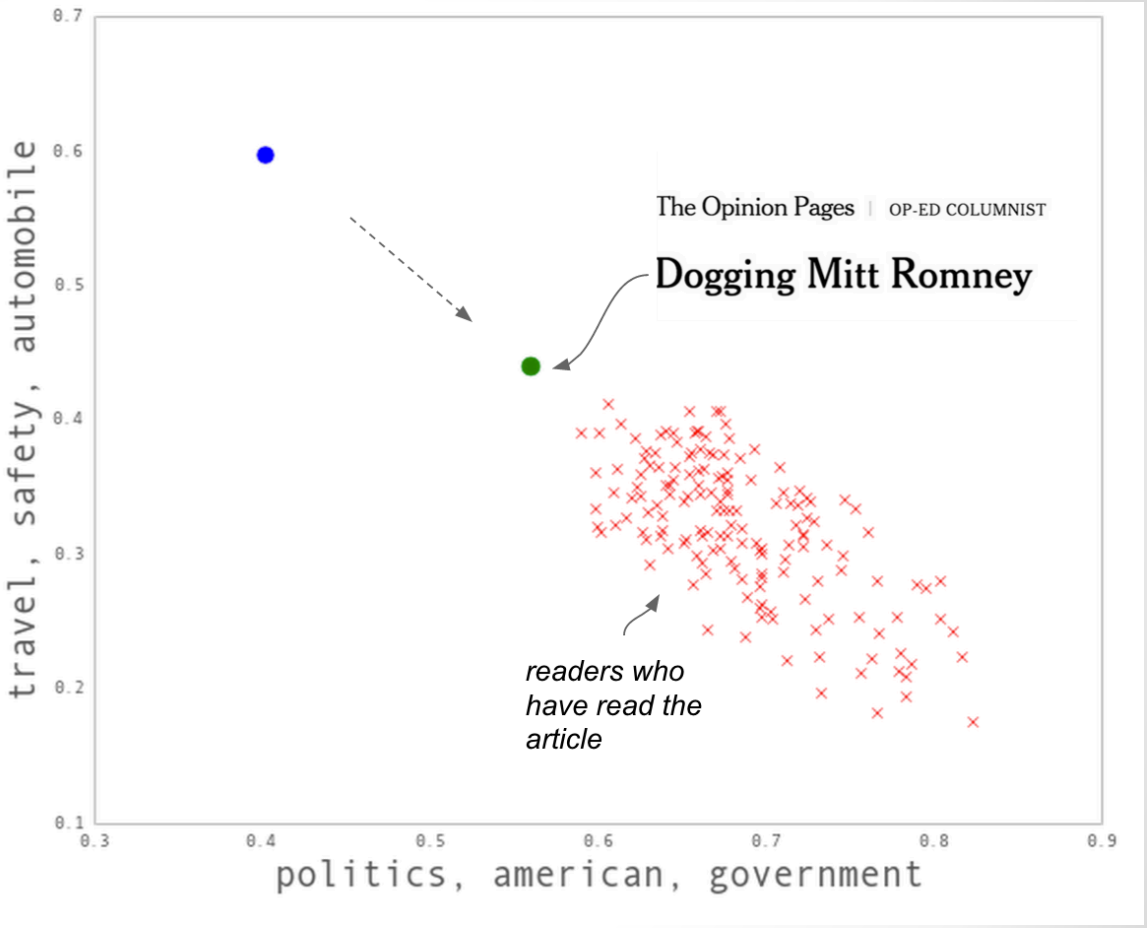
Transparency of News Algorithms

. . .

AP's use of Automated Insights to generate earnings stories.



Functioning of New York Times' article recommendation system



content only content + social

News Algorithms

Algorithms used at all stages of journalism.

- **Data mining** or analysis during reporting
- **Automated story writing**
- **Recommendation systems** both publisher and platform
- **Ad targeting**,
- **Subscriber retention…**

Three major stakeholders: technologists, readers, and journalists.

Stages of news algorithm transparency

- **Disclosure.** An algorithm was involved.
  "AP created this story using an automated system"

- **Justification.** *Some* reason for the algorithmic result in this case.
  "You're seeing this article because you said you liked vegetables"

- **Explanation.** More detailed algorithmic analysis of this case.
  "This is how our election prediction model works…"

- **Reproduction.** Enough information to allow independent replication.
  "This is our newsroom's github repo" or "Here's a link to Workbench"

How can we make algorithmic news more transparent?
Stuart Myles, AP, 2018