Ethical Concerns in Data Science

Richard De Veaux

Williams College

Statistically Speaking

March 31, 2022

All Models are Wrong. What about Data?

- Why do we analyze data?
- To learn about the world,

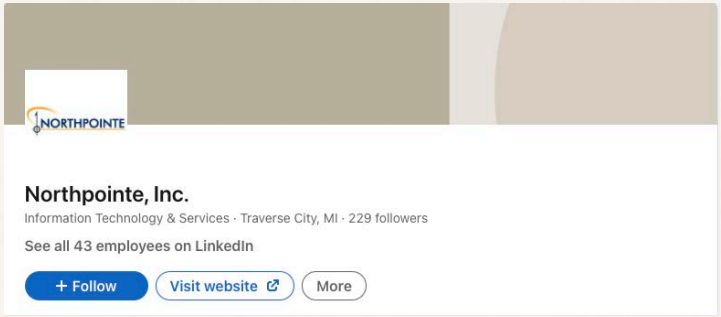
- Or just to prove a point?

Three examples —is There a Theme?

- Transparency and bad data
  - Proprietary algorithms (black box)
- Jumping to your own conclusions
  - Don't be fooled by your own assumptions
- Bad data
  - Statisticians must take responsibility

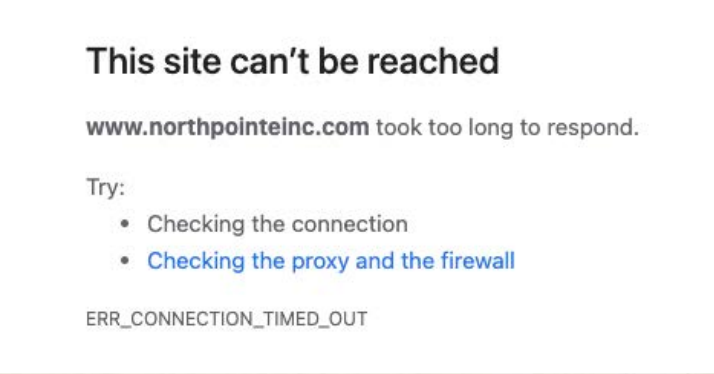
COMPAS Recidivism Algorithm — Details

- Proprietary algorithm
  - 137 variables
  - Based on questionnaire and criminal justice database
  - Incentivized by profit
- The score and methodology
  - An integer between 1 and 10
  - Compute *subscales* from the questionnaire
  - Weights w



Violent Recidivism Risk Score

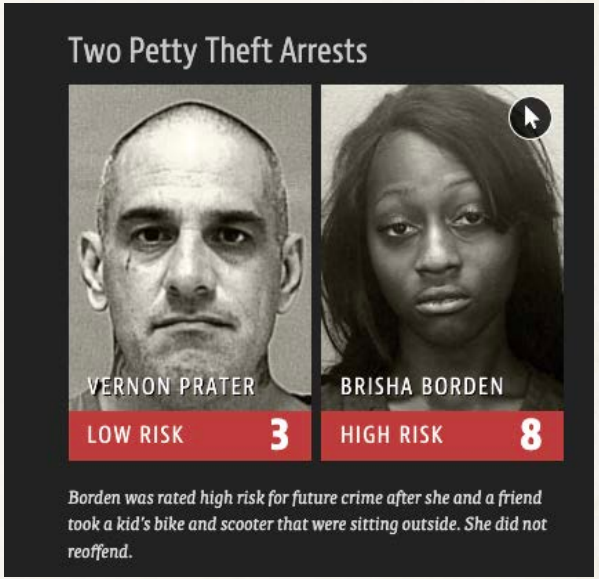
$$= (\text{age} * -\text{w}) + (\text{age-at-first-arrest} * -\text{w}) + (\text{history of violence} * \text{w}) + (\text{vocation education} * \text{w}) + (\text{history of noncompliance} * \text{w}),$$



ProPublica — Machine Bias?

PROPUBLICA
Investigative Journalism in the Public Interest

When he was U.S. Attorney General, Eric Holder asked the U.S. Sentencing Commission to study potential bias in the tests used at sentencing. “Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice,” he said, adding, “they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.” The sentencing commission says it is not currently conducting an analysis of bias in risk assessments.



ProPublica Summary

- Critique
  - Huge false positive rate
  - Differential rate by race
- Examples
  - Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45% vs. 23%).
  - White defendants who re-offended (within 2 yrs) were labeled low risk almost twice as often as black re-offenders (48% vs. 28%)
  - Even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.
  - The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.

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%

Two Drug Possession Arrests

DYLAN FUGETT

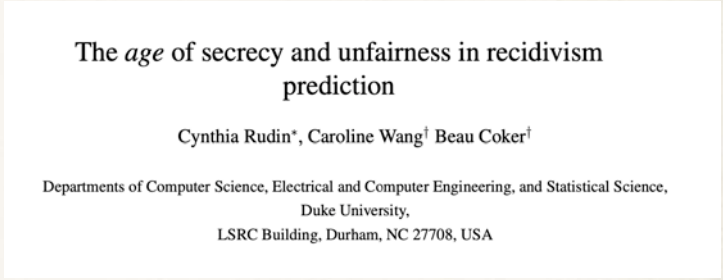
RISK: 3

BERNARD PARKER

RISK: 10

Rudin *et al* Analysis

- Reproduced the analysis using a transparent model
- Algorithm bias?
  - Or bad data? Or incorrect explanation of COMPAS by ProPublica?
  - Proprietary algorithm makes it hard to tell.
- ProPublica's analysis assumes age linear in COMPAS model
  - As Northpointe claims
  - ProPublica's conclusion about race does not follow if age non-linear as data suggests
- ML methods for predicting COMPAS scores performed equally well with or without direct knowledge of race.
  - So, what other variables does it depend on?
- Should freedom be determined by proprietary models anyway?
  - Interpretable models are equally accurate.



The *age* of secrecy and unfairness in recidivism prediction

Cynthia Rudin*, Caroline Wang† Beau Coker†

Departments of Computer Science, Electrical and Computer Engineering, and Statistical Science,
Duke University,
LSRC Building, Durham, NC 27708, USA

Criminal Justice Paper

- In this paper, we answer two primary questions:
  - (1) Which federal judges give the harshest sentence length penalty to Black defendants?
  - (2) Which federal judges give the harshest sentence length penalty to Hispanic defendants

The Most Discriminatory Federal Judges Give Black and Hispanic Defendants At Least Double the Sentences of White Defendants

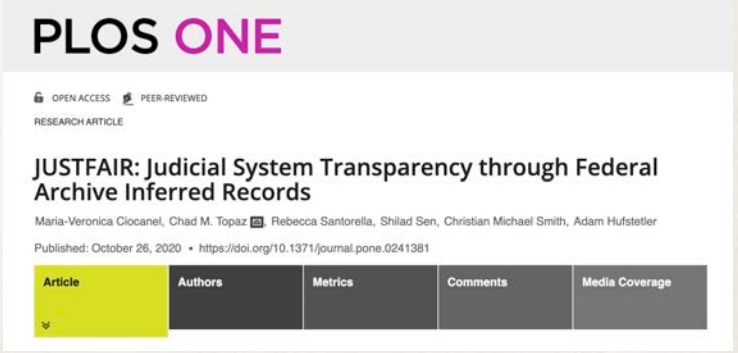
Christian Michael Smith[1], Nicholas Goldrosen[2], Maria-Veronica Ciocanel[3], Rebecca Santorella[4], Chad M. Topaz[5,6], Shilad Sen[7]

Criminal Justice Paper

- In this paper, we answer two primary questions:
  - (1) Which federal judges give the harshest sentence length penalty to Black defendants?
  - (2) Which federal judges give the harshest sentence length penalty to Hispanic defendants
- Use JUSTFAIR database
  - Across all individuals i, we define judge j's degree of racial discrimination against Black defendants as. $E[Y_{ijb} - Y_{ijw}]$
    - Y_{ijb} is the potential sentence that judge j gives to defendant i if defendant i is Black
    - Y_{ijw} is the potential sentence that judge j gives to defendant i if the same defendant i is White.
- Identifies two judges in Philadelphia as discriminating
  - "Random chance is not a viable explanation for the large disparities…"

The Most Discriminatory Federal Judges Give Black and Hispanic Defendants At Least Double the Sentences of White Defendants

Christian Michael Smith[1], Nicholas Goldrosen[2], Maria-Veronica Ciocanel[3], Rebecca Santorella[4], Chad M. Topaz[5,6], Shilad Sen[7]



PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records

Maria-Veronica Ciocanel, Chad M. Topaz, Rebecca Santorella, Shilad Sen, Christian Michael Smith, Adam Hufstetler

Published: October 26, 2020 • https://doi.org/10.1371/journal.pone.0241381

Article | Authors | Metrics | Comments | Media Coverage

Reactions — Jonah Gelbach

- Economist and law professor, UC Berkeley
  - Data are incomplete. Missing data rampant — low coverage.
    - Match rate is less than 50% — judge to sentence
  - Model does not account for district differential plea offers
    - Matched for severity of case load?
  - Two judges called out for the worst "violation"
    - Hon. C. Darnell Jones II (< 200 cases total)
    - Hon. Timothy J. Savage (< 300 cases total)
  - 760 fair coins flipped 150 times showed two coins with extremely low number of heads (the two judges?)

- Twitter reaction
  - Twitter account is deleted

Jonah B. Gelbach

Professor of Law

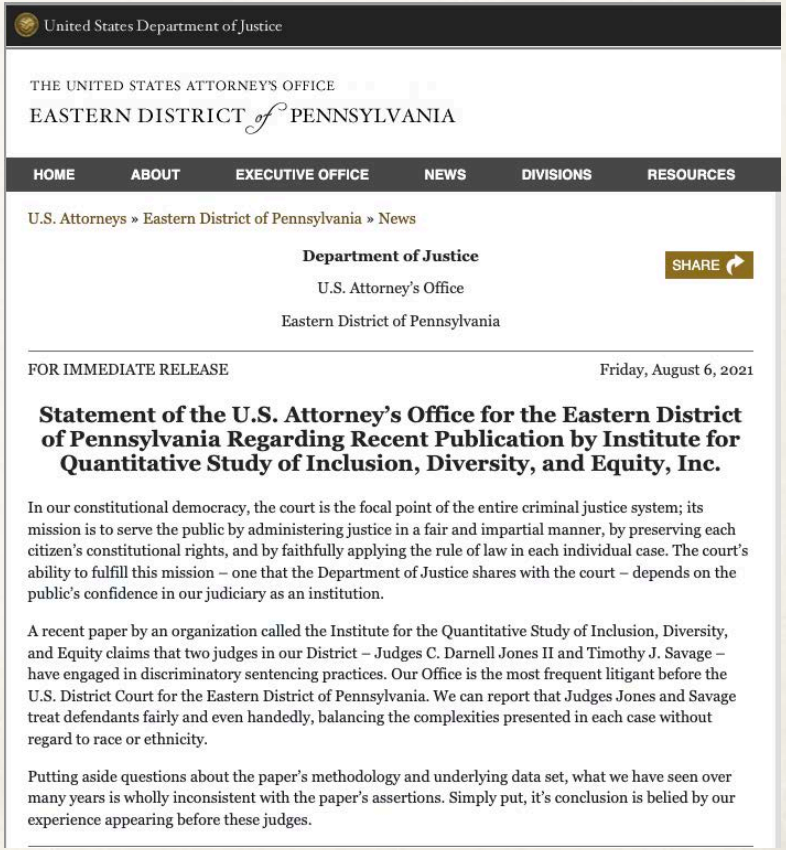
gelbach@berkeley.edu

Faculty Support Contact: Aman Dosanjh

Areas of Expertise: Civil Procedure and Litigation | Evidence | Law and Economics | Securities Law

Reactions — US Attorney's Office

- Eastern District of Pennsylvania
  - Our office is the most frequent litigant before the U.S. District Court for the Eastern District of Pennsylvania.
  - We report that Judges Jones and Savage treat defendants fairly and even handedly, balancing the complexities presented in each case without regard to race or ethnicity.
  - Putting aside questions about the paper's methodology and underlying data set, what we have seen over many years is wholly inconsistent with the paper's assertions.
  - Simply put, it's [sic] conclusion is belied by our experience appearing before these judges



United States Department of Justice

THE UNITED STATES ATTORNEY'S OFFICE
EASTERN DISTRICT of PENNSYLVANIA

HOME ABOUT EXECUTIVE OFFICE NEWS DIVISIONS RESOURCES

U.S. Attorneys » Eastern District of Pennsylvania » News

Department of Justice
U.S. Attorney's Office
Eastern District of Pennsylvania

SHARE

FOR IMMEDIATE RELEASE Friday, August 6, 2021

Statement of the U.S. Attorney's Office for the Eastern District of Pennsylvania Regarding Recent Publication by Institute for Quantitative Study of Inclusion, Diversity, and Equity, Inc.

In our constitutional democracy, the court is the focal point of the entire criminal justice system; its mission is to serve the public by administering justice in a fair and impartial manner, by preserving each citizen's constitutional rights, and by faithfully applying the rule of law in each individual case. The court's ability to fulfill this mission – one that the Department of Justice shares with the court – depends on the public's confidence in our judiciary as an institution.

A recent paper by an organization called the Institute for the Quantitative Study of Inclusion, Diversity, and Equity claims that two judges in our District – Judges C. Darnell Jones II and Timothy J. Savage – have engaged in discriminatory sentencing practices. Our Office is the most frequent litigant before the U.S. District Court for the Eastern District of Pennsylvania. We can report that Judges Jones and Savage treat defendants fairly and even handedly, balancing the complexities presented in each case without regard to race or ethnicity.

Putting aside questions about the paper's methodology and underlying data set, what we have seen over many years is wholly inconsistent with the paper's assertions. Simply put, it's conclusion is belied by our experience appearing before these judges.

Fall out

- The court's ability to fullfill its mission depends on the public confidence in the judiciary.

- Second "version" of paper 1 month later
  - No individual judges are named — only averages
  - Note: A previous version of this work included estimates on individually identified judges. Thanks to helpful feedback, we no longer place enough credence in judge-specific estimates to make sufficiently confident statements on any individual judge. We encourage others not to rely upon results from earlier versions of this work.

- What if the named judges had been more vulnerable?

Racial Disparities in Criminal Sentencing Vary Considerably across Federal Judges

Christian Michael Smith[1], Nicholas Goldrosen[2], Maria-Veronica Ciocanel[3], Rebecca Santorella[4], Chad M. Topaz[5,6], Shilad Sen[7]

Republican Undercount?

- "Williams College Professor Steven Miller, a Yale and Princeton trained math expert, said he analyzed Pennsylvania ballot data collected by former Trump campaign data chief Matt Braynard as well as 2,684 voter interviews conducted by a phone bank and found two concerning patterns.
    - One involved possible votes that were not counted
    - The other ballots that appeared to be requested by someone other than a registered voter."

- "I estimate that the number of ballots that were either requested by someone other than the registered Republican or requested and returned but not counted range from 89,397 to 98,801," Miller said in the sworn statement provided to Just the News

Justification

- All three professionals … took serious issue with Miller's assumption that the data that he analyzed was accurate and representative. They pointed to the fact that the survey's response rates were low, that the survey only reached out to registered Republicans and that voters often tell pollsters untrue statements on their voting history.
- "I agree completely that this could have been done better and welcome the opportunity to help others not make these mistakes," Miller told the Record in response to the criticism he has received.
- Miller also apologized via email to community members on Monday, saying that his dependence on the accuracy and reliability of the data he analyzed should have been spelled out more clearly.

Professor Steven Miller issues legal statement suggesting PA ballot irregularities; conclusions repudiated by statisticians, political scientists

Samuel Wolf
November 25, 2020

Professor of Mathematics Steven Miller issued a sworn legal statement suggesting ballot irregularities in Pennsylvania. (Samuel Wolf/ *The Williams Record*)

Ethical Standards

The ethical statistician:

1. **Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis. When reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms.**
2. **Reports the limitations of statistical inference and possible sources of error.**
3. In publications, reports, or testimony, identifies who is responsible for the statistical work if it would not otherwise be apparent.
4. **Reports the sources and assessed adequacy of the data, accounts for all data considered in a study, and explains the sample(s) actually used.**
5. **Clearly and fully reports the steps taken to preserve data integrity and valid results.**
6. **Where appropriate, addresses potential confounding variables not included in the study.**
7. In publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics.
8. **In publications or testimony, identifies the ultimate financial sponsor of the study, the stated purpose, and the intended use of the study results.**
9. **When reporting analyses of volunteer data or other data that may not be representative of a defined population, includes appropriate disclaimers and, if used, appropriate weighting.**
10. To aid peer review and replication, shares the data used in the analyses whenever possible/allowable and exercises due caution to protect proprietary and confidential data, including all data that might inappropriately reveal respondent identities.
11. Strives to promptly correct any errors discovered while producing the final report or after publication. As appropriate, disseminates the correction publicly or to others relying on the results.

Disavowal and Apology

- Miller told The Eagle that he made a mistake separating his analysis of the data from questions about the reliability of the data itself.
- "Especially as the consequences are so important, I should have made a greater effort to go deeply into and share how the data was collected and not treat this solely as an independent calculation," he wrote in a statement Monday night.
- The data Miller used came from Braynard's "Voter Integrity Fund," a group led by former Trump campaign staffers and government employees, Miller confirmed to The Eagle

Williams prof disavows own finding of mishandled GOP ballots

By Francesca Paris, The Berkshire Eagle
Nov 24, 2020

Steven Miller, a professor of mathematics at Williams College, issued a statement Monday in which he apologized for a "lack of clarity and due diligence" after his statistical analysis of Pennsylvania mail-in votes was used by conservative lawmakers to push unsubstantiated claims of voter fraud.

Summary

- We all need to be aware of our biases in our analysis
- Proprietary and/or black box algorithms?
  - Have no place making decisions about justice, finance, health etc.
  - Interpretable models are equally accurate.
  - Reproducibility
- Public shaming is dangerous
  - Make sure your analysis is iron clad and reproducible
- Data pedigree
  - Is essential
  - Is the responsibility of the analyst
  - Limits to analysis

Thank you!!!