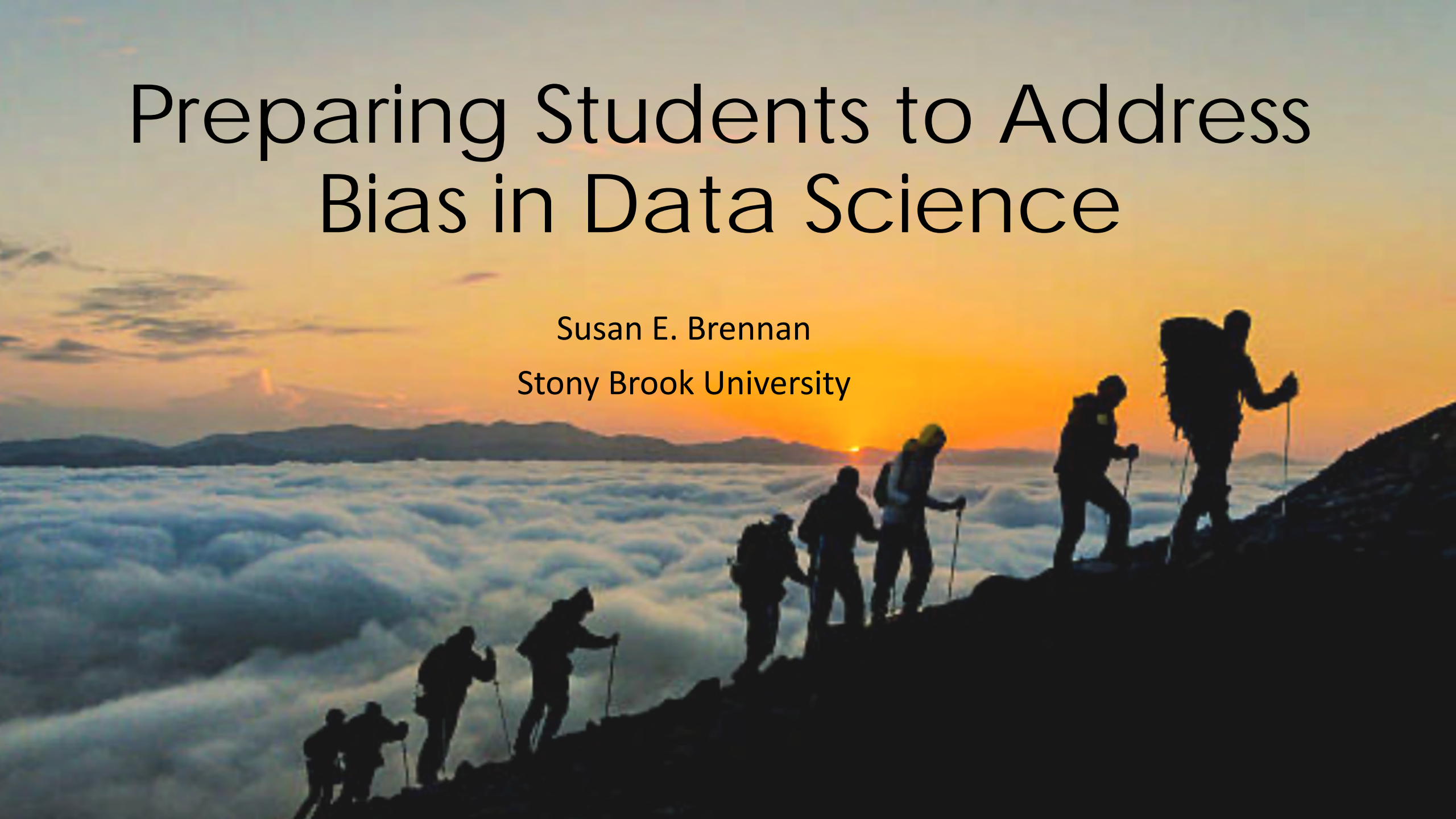


Preparing Students to Address Bias in Data Science

Susan E. Brennan

Stony Brook University



What is a model?

“An opinion, embedded in math”
(Cathy O’Neil, 2016, p. 21)

CGS COUNCIL OF GRADUATE SCHOOLS

PROGRAM

Friday, December 7, continued

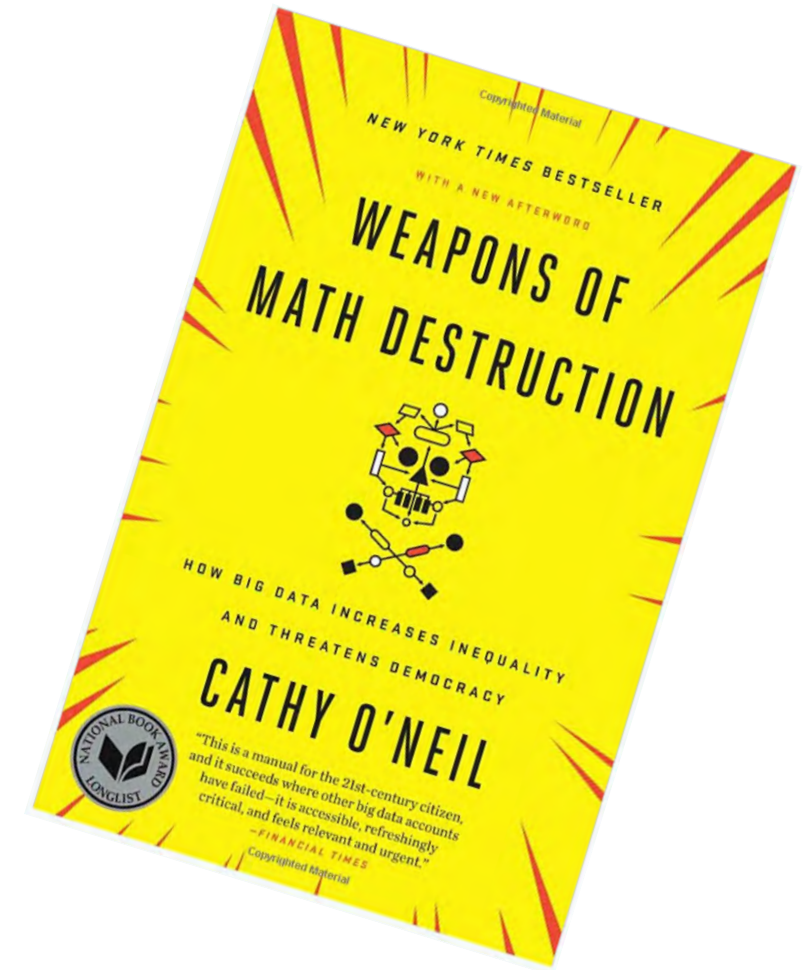
8:00 AM - 12:00 PM EXHIBITS OPEN

9:00 - 10:30 AM **PLENARY III**
**WEAPONS OF MATH DESTRUCTION AND THE ETHICAL MATRIX:
HOW TO EMBED ETHICAL CONSIDERATION IN BIG DATA CONTEXTS**

In *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Cathy O’Neil argues that algorithms used in big data analysis are deciding society’s winners and losers. After explaining the biases inherent in algorithms, which determine which data are collected and analyzed, Dr. O’Neil will address implications for the ethical training of graduate students across all disciplines.




Presider: Phillip Harper, Dean, Graduate School of Arts and Science, New York University
Speaker: Cathy O’Neil, Author, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*

AMBASSADOR
REGENCY WEST



When people use technology for decision-making, where might bias come into play?

The AI Technology Stack (*Moore, 2018*)

| | |
|--|---|
|  DECIDE | Weigh the output, search, plan, predict |
|  LEARN | Deep learning, ML (a black box) or something more transparent |
|  PERCEIVE | Data input (historical data), "Big Data," hardware, the cloud |

Example: Bias in Automated Decision Systems

Machine Bias in the Northpointe COMPAS Algorithm:

What is the likelihood that someone arrested for an \$80 misdemeanor) will commit another crime?

by Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner

ProPublica, May 23, 2016

Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1
attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK

3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK

8

ProPublica's Method

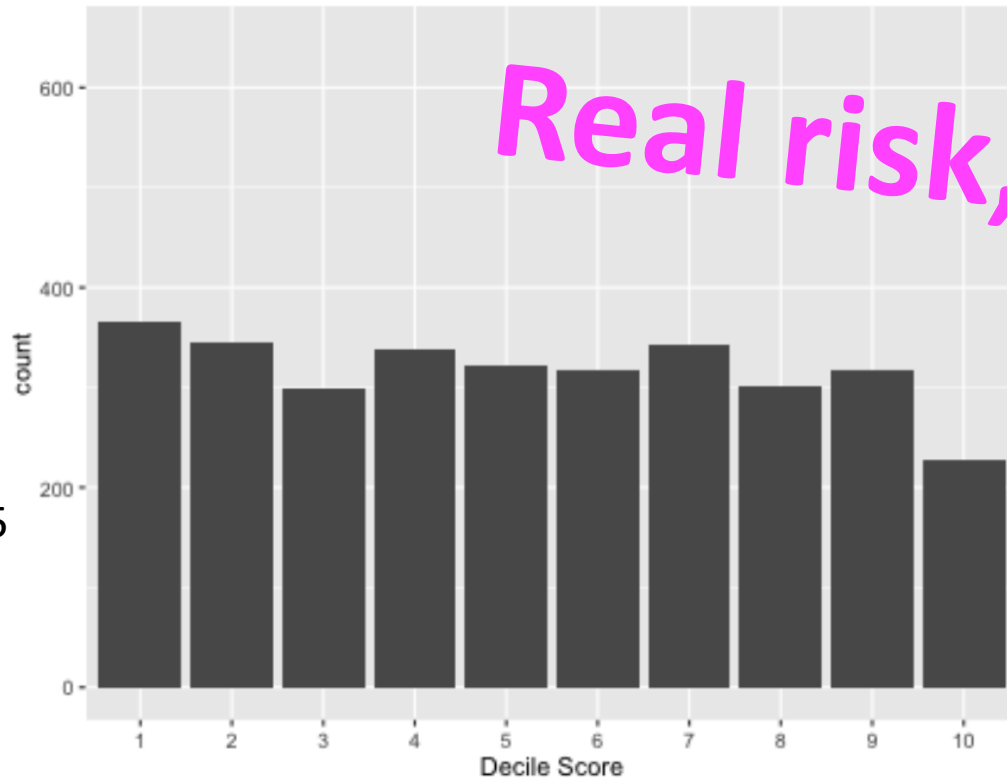
- Obtained records for 18,000 Broward Cty arrests (2013 & 2014)
- Input data: [137-Q COMPAS questionnaire](#) upon being booked into jail
 - Joined these data with individuals' race
 - Recidivism algorithms do not use race in their ML models (illegal), but *do* use many variables correlated with race or neighborhood
 - ❖ When was the first time you were ever involved with the police?
 - ❖ Do any of your relatives have criminal records?
 - ❖ Are you employed?
 - ❖ How long have you lived in your home? How often have you moved?

Proxies!

ProPublica's Method

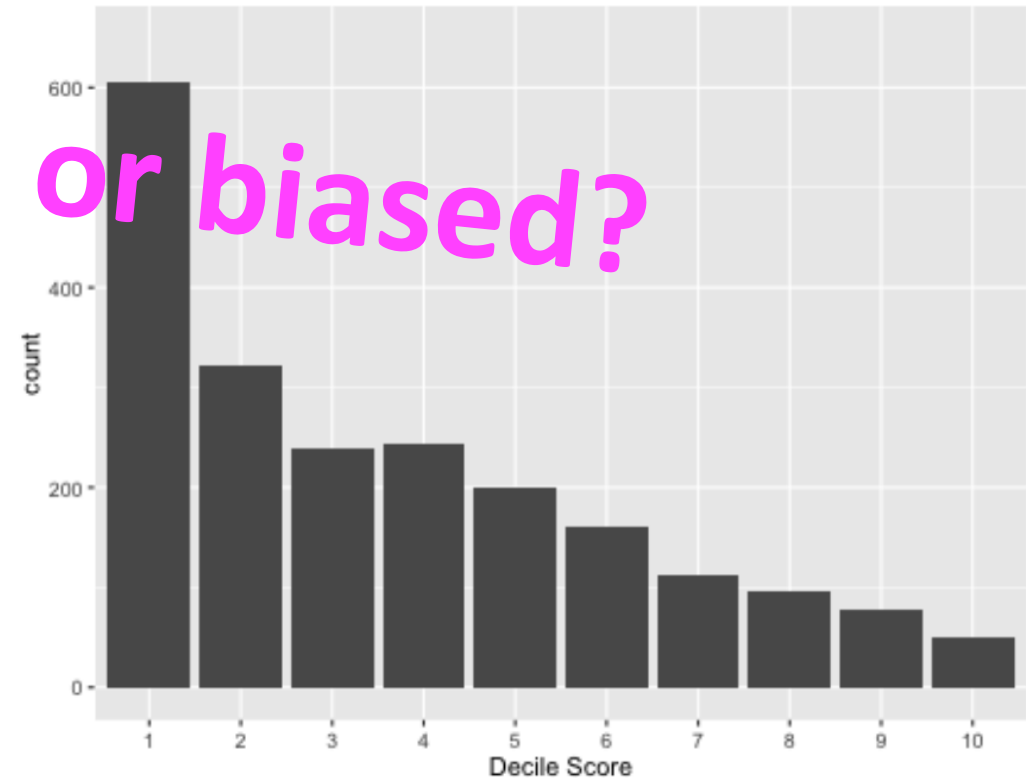
- Obtained records for 18,000 Broward Cty arrests (2013 & 2014)
- Input data: [137-Q COMPAS questionnaire](#) upon being booked into jail

Black Defendant's Decile Scores



N=3,175

White Defendant's Decile Scores



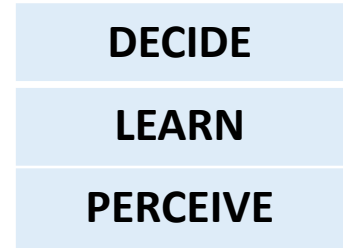
N=2,101

ProPublica's Method

- Obtained records for 18,000 Broward Cty arrests (2013 & 2014)
- Input data: [137-Q COMPAS questionnaire](#) upon being booked into jail
- Compared predictions with outcomes (ground truth) over 2 yrs
 - COMPASS was 61% correct in predicting nonviolent recidivism
 - For 7,000 inmates, error rate was roughly equal for Blacks & Whites (39%), so were “well-calibrated” —a common way of defining “fairness”
 - **BUT Whites were misclassified as low risk** twice as often as Blacks
 - and **Blacks were misclassified as high risk** twice as often as Whites

| | Black Defendants | White Defendants |
|--|-------------------------|-------------------------|
| False Positive Rate (no +crime) | 44.85% | 23.45% |
| False Negative Rate (+crime) | 27.99% | 47.72% |

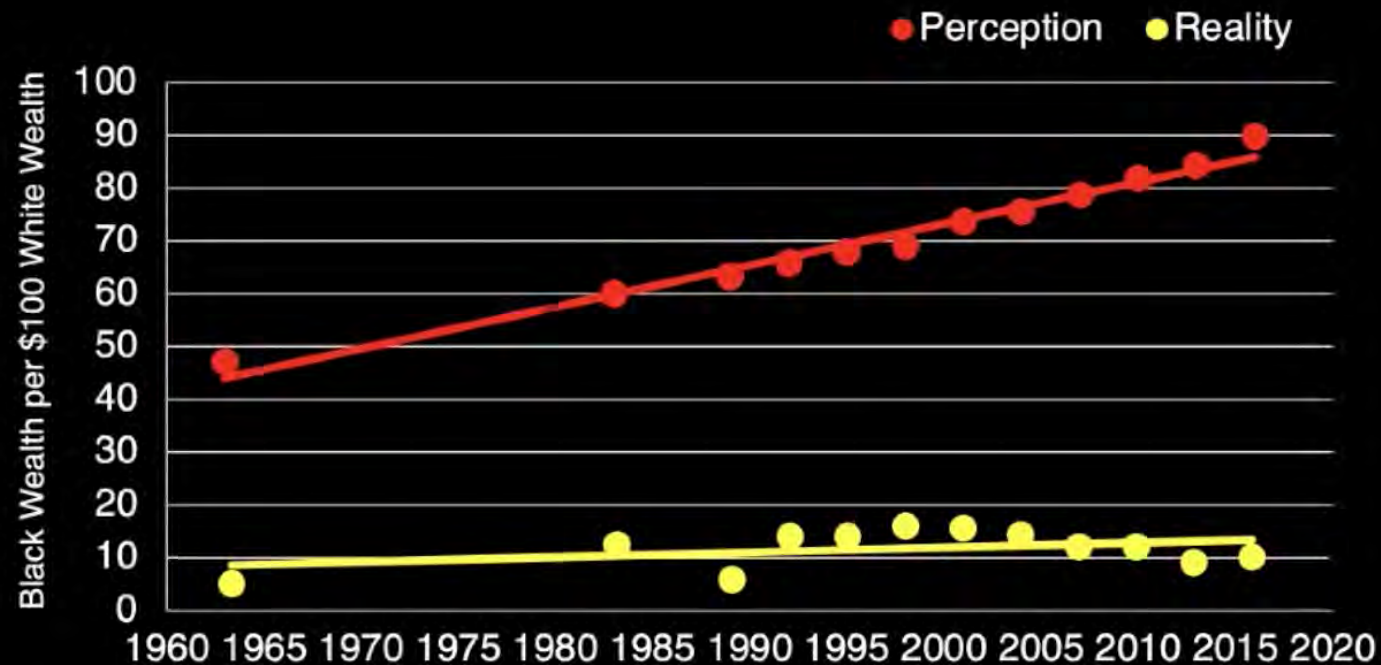
Where is the bias in COMPAS?



- In the **input data** (*PERCEIVE*) – with the use of proxies. *Ignoring* race does not factor it out, but can amount to colorblind racism.
- In the **algorithm** (*LEARN*) – “well-calibrated” does not mean “fair.”
- In its **use for decisions** – COMPAS was designed to predict recidivism, but has been used for sentencing (even Northpointe agrees this is misguided).
- Policy issue: The biases in COMPAS could be deployed differently:
 - For EVIL: Risk scales filter individuals by a flawed notion of risk to society (unfair).
 - For GOOD: Needs scales can measure the needs of stakeholders, inform case plans, and identify opportunities for intervention and anti-recidivism support

Bias exists simultaneously on multiple levels: *Data vs. What People Think about Data*

Perceptions of Racial Wealth Equality 1963-2016



(← The Ground Truth)
Jennifer Richeson &
colleagues (2019)

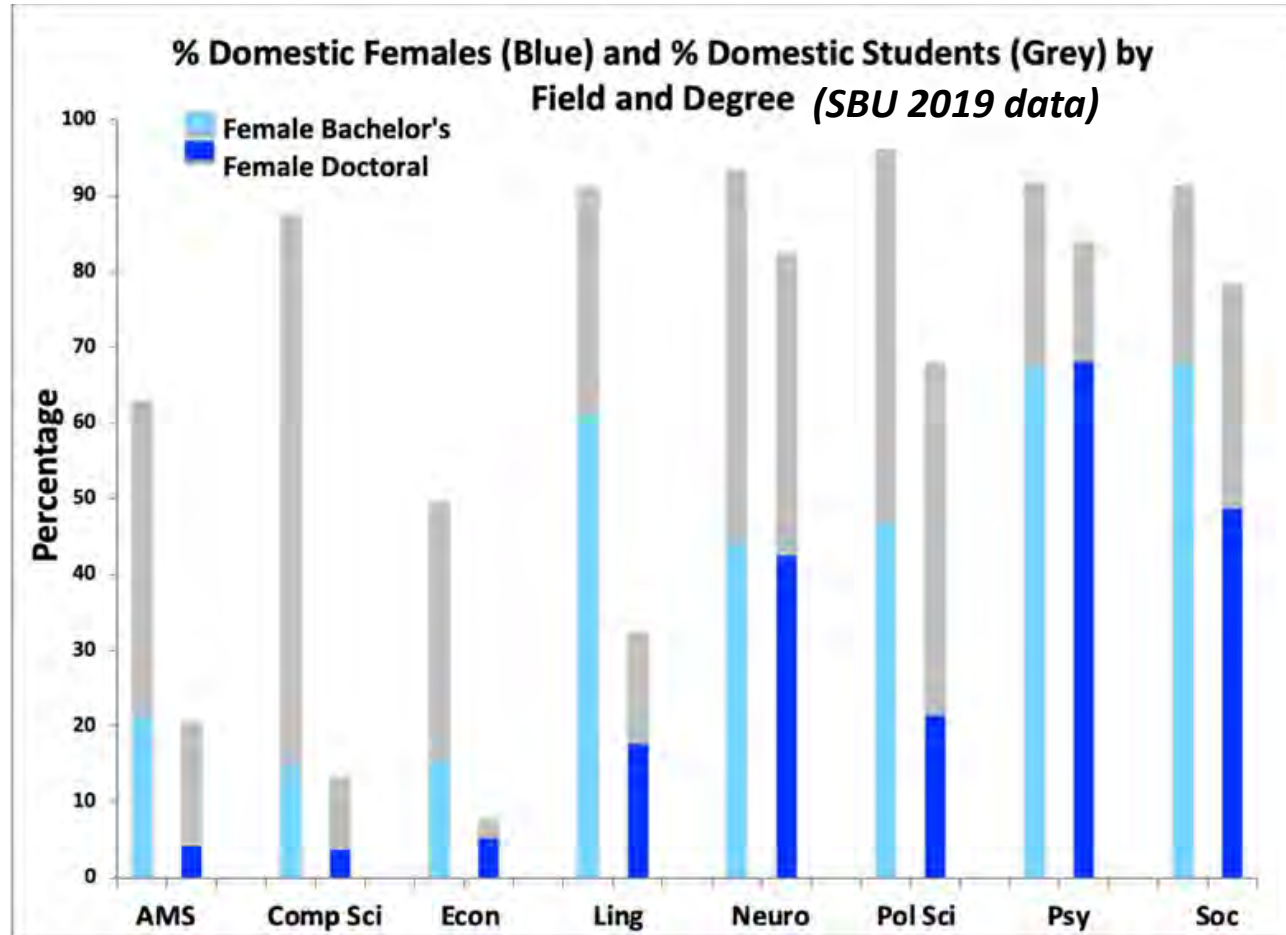
This is why **history** – another word for data –
is so very important!

Now, back to the story of our traineeship....

Why are there so few women in our computer science classes?

Where *are* the data scientists?

One program's challenge can be another program's solution.





Data-Centered Sciences

Human-Centered Sciences

AI Certificate
CS Bridge Courses

Our NSF Research Traineeship – A timeline

- Feb 6, 2019 – First proposal submitted. Good pedagogical model, but weak theme.
- July 2019 – Declined (encouraging but critical reviews; but proposal was not even discussed)
- Early 2020 competition – Sat this one out; we missed SBU's internal competition deadline

Detecting and Addressing

B I A S

in Data, Humans, and Institutions

Our NSF Research Traineeship – A timeline

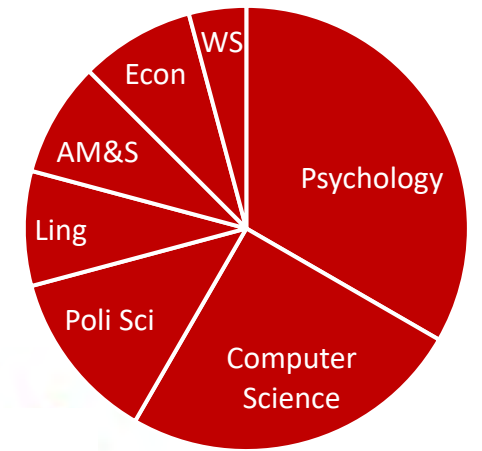
- **Feb 6, 2019** – First proposal submitted. Good pedagogical model, but weak theme
- **July 2019** – Declined (encouraging but critical reviews; but proposal was not even discussed)
- **Early 2020 competition** – Sat this one out; we missed SBU’s internal competition deadline
- **Oct 2020** – With our new (bias) theme, we were selected in SBU’s internal competition
- **Feb 25, 2021** – Second proposal submitted
- **May 27, 2021 (Memorial Day holiday weekend)** – We made NSF’s short list. With only a week to respond to a set of questions, we were able to flesh out our model and add a Human-Centered Data Science certificate to our curriculum.
- **July 12, 2021** – Funding recommended by NSF!
- **Sept 1, 2021** – Project start date (unexpectedly, a month early)
- **August, 2022** – First cohort of 11 NRT trainees admitted (from 7 graduate programs, 5 depts)
- **July, 2023** – Second cohort of 13 NRT trainees admitted (from 9 programs, 7 depts)

Detecting and Addressing

BIAS

in Data, Humans, and Institutions

Features of the Traineeship Model



Convergent Research Practica

Data Science Internships

Cohort-based Training

Bridge Courses

2 Certificates



34 Funded Trainees

34 Non-Funded Trainees

Alan Alda Center Courses

Cross-Disciplinary Mentoring

Travel funds



This project is based
on work supported by
NSF Grant #2125295

CONVERGENT RESEARCH PROJECTS

COHORT 1

- *Post-Conviction Project* – with data from the National Registry of Exonerations and advisors from the Innocence Project’s Network
- *Bias in Large Language Models* (especially gender bias; several projects)
- *Bias in Facial Emotion Recognition* (clinical applications)


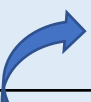



Post-Conviction Project

COHORT 1

- Partners: Domain experts (including 4 Innocence Organizations)
- The National Registry of Exonerations – Public facing database with ~3,200 exonerations since 2012; continuously updated
 - Explore factors underlying wrongful conviction and successful exoneration; categorize cases (latent classes) and make predictions
 - Ask: How can complex data analyses be made more transparent? (decision-trees)
 - **Recently submitted our first paper to the *Just Data* Conference**
- Next steps:
 - Observe intake staff making decisions under uncertainty and with incomplete data – (1) whether an applicant can be shown to be factually innocent, and (2) whether the organization has the resources to help
 - Create and test data-intensive tools to support transparent decision-making and communication for Innocence Project staff
 - Identify biases at different stages of the conviction and exoneration processes

Where is the bias? *In the context of use.*
Let's extend the AI technology stack.

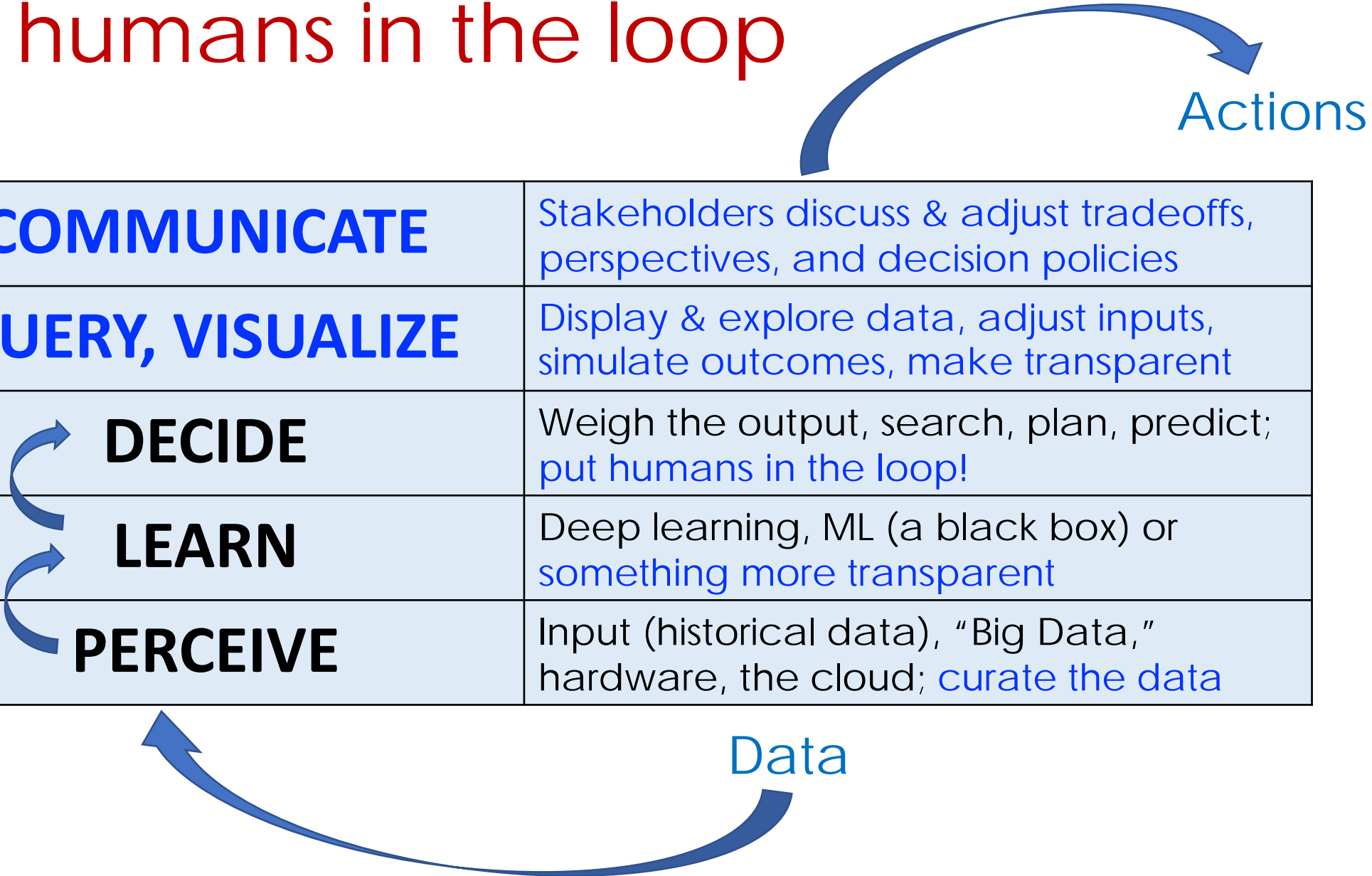
| | |
|---|---|
|  DECIDE | Weigh the output, search, plan, predict |
|  LEARN | Deep learning, ML (a black box) or something more transparent |
|  PERCEIVE | Data input (historical data), "Big Data," hardware, the cloud |

Put humans in the loop

Actions

| | |
|-------------------------|--|
| COMMUNICATE | Stakeholders discuss & adjust tradeoffs, perspectives, and decision policies |
| QUERY, VISUALIZE | Display & explore data, adjust inputs, simulate outcomes, make transparent |
| DECIDE | Weigh the output, search, plan, predict; put humans in the loop! |
| LEARN | Deep learning, ML (a black box) or something more transparent |
| PERCEIVE | Input (historical data), "Big Data," hardware, the cloud; curate the data |

Data

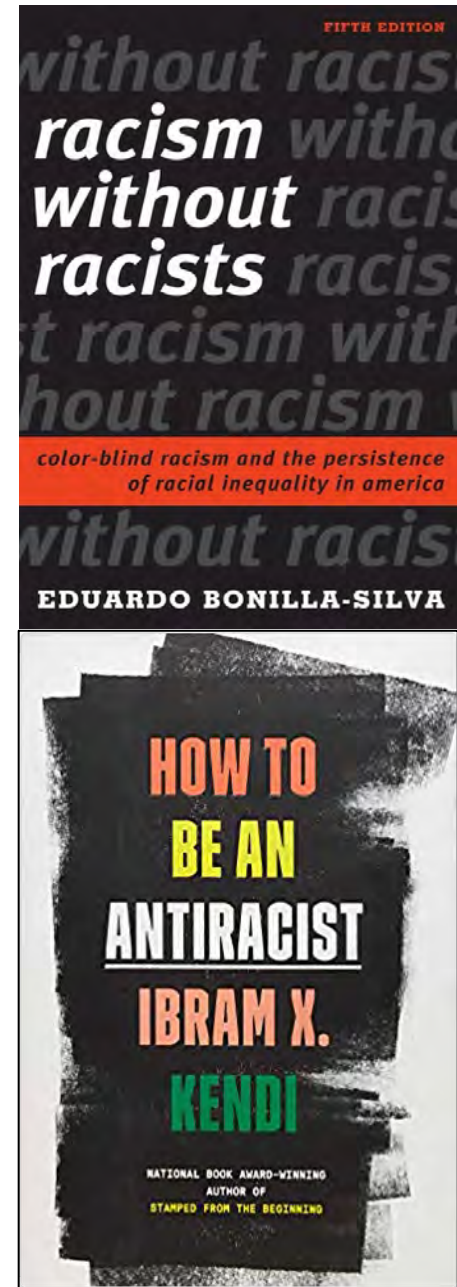


We can't address bias by ignoring it

- Bias-blindness ignores the strong historical impacts of bias
- Historical impacts are embedded in our institutions and infrastructure (redlining; property tax school funding; healthcare algorithms; and even the LI highway system)
- Color- (or gender-) blindness makes Whiteness (or maleness) the norm, and everyone else, the exception.

Addressing bias is even harder than detecting it.

- As scientists, we're equipped to *detect* bias w.r.t. a particular context.
- It's much more difficult to *address* bias; that depends on values, goals, policies, and the ability to effect change.



We can't address bias by ignoring it

“The only way out of this morass—for all of us—is to stare at racial disparity unblinkingly, and then do what evidence and experts tell us is required to level the playing field and march forward together, collectively striving to achieve true equality for all Americans.”

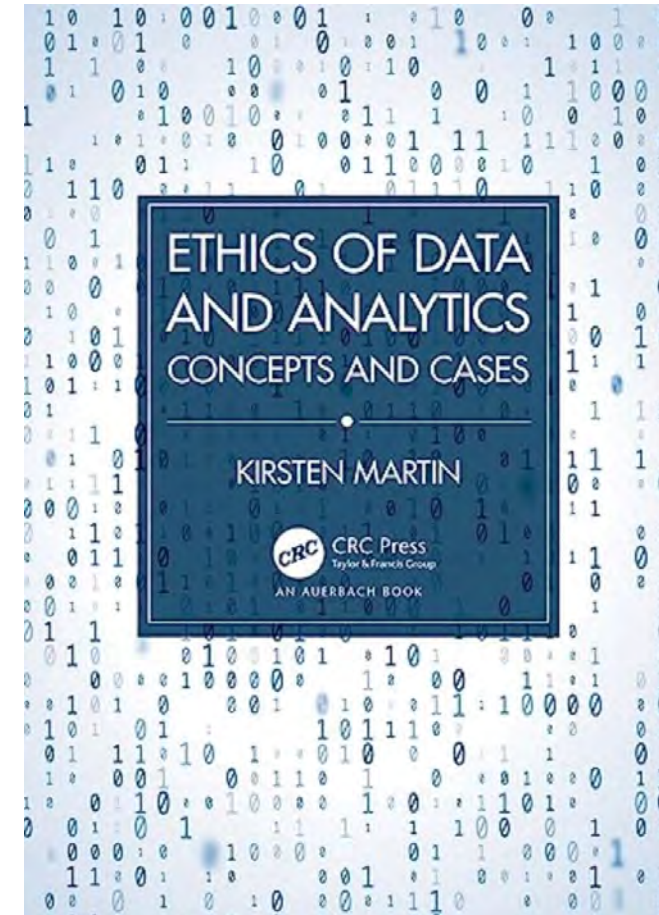
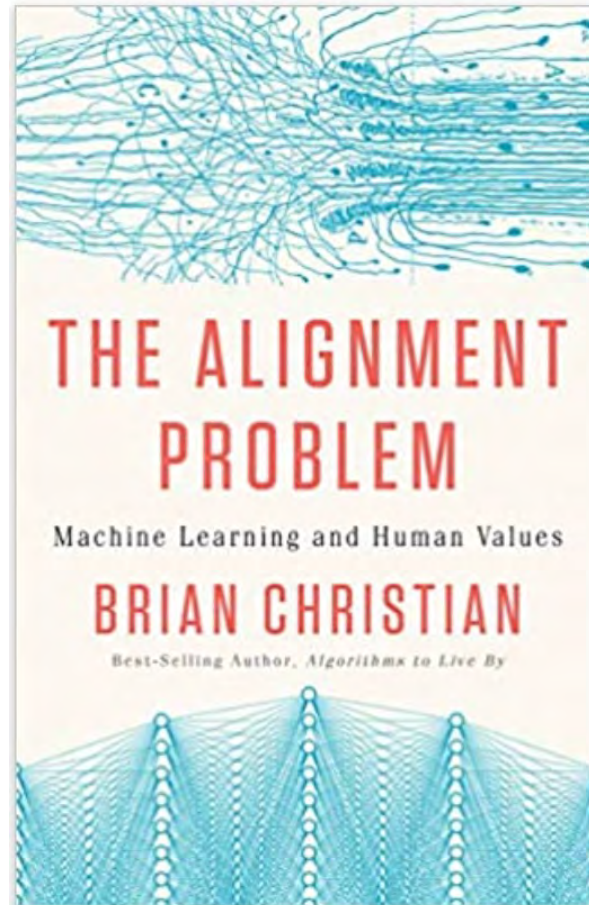
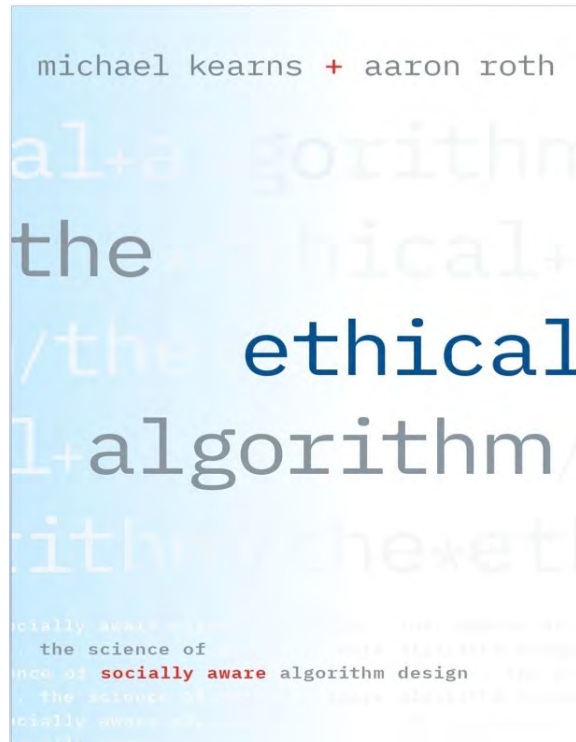
Judge K. B. Jackson's dissent in 21–707, SFFA Inc. v. UNC, p. 26

How can we get to accountability for AI?

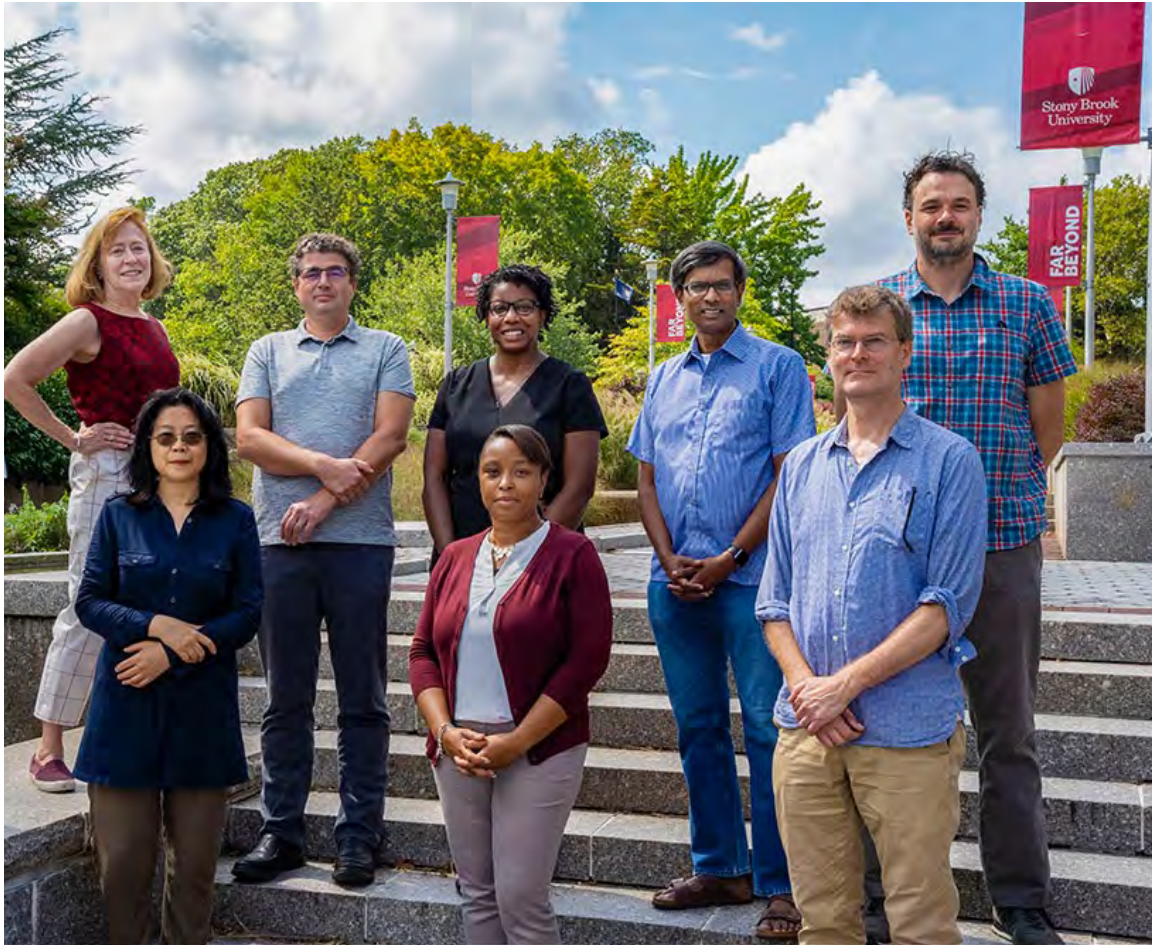
Training in data science should include auditing datasets and algorithms for bias

- Auditing for bias should consider the algorithm's broad contexts of use.
- It should involve stakeholders—not just developers and those with power.
- Establish clear and transparent standards, without COI (tech can't monitor itself).
- Don't ask if AI is good or fair—ask how it shifts power (Kalluri, 2020, *Nature*)
- See also Timnit Gebru et al.'s work (incl. *Stochastic Parrot* and *Datasheets* papers)
- Auditing adaptive algorithms is itself quite a research problem! Policy and science need to work together on this.

Resources and Inspiration:



Thanks!



This project is based on work supported by NSF Grant #2125295