

B9145

Reliable Statistical Learning

Fairness, Accountability, Transparency, Ethics

Failure modalities of AI

- Infrastructural view of data
- Recognize the embodiment of social, economic, and political interests in analytics / AI
- Discuss how to manage, communicate, and mitigate these limitations
- Learn how to prevent harmful adoptions of technology

AI and power

AI and power

- Like any engineering system, AI technology is borne out of **economic, social, and political forces**.
- They build on intangible and material infrastructure: human labor, computing servers, organization, natural resources
- How these systems are used is very much up to the people in power
- The omini-present nature of AI can rigidify existing socioeconomic structures

Potential of AI

- AI & analytics offer promise for people who wield power
- If you're (like me) used to technology being a positive force in your life, you can easily imagine AI models helping you

Smart Interfaces for Human-Centered AI

JAMES LANDAY March 12, 2019

You're in an AI-augmented office, hard at work:

"By observing cues like your posture, tone of voice, and breathing patterns, it can sense your mood and tailor the lighting and sound accordingly. Through gradual ambient shifts, the space around you can take the edge off when you're stressed, or boost your creativity when you hit a lull."

But for whom?

- They can be used against people who are already targeted and surveilled against
- Like any other technology, but with a wider reach and *omni-present*; codifies and automates existing structure

Anthropological/Artificial Intelligence & the HAI

Ali Alkhatib

You're in an AI-augmented office, hard at work:

Lights are carefully programmed by your employer to hack your body's natural production of melatonin through the use of blue light. The work day eke out every drop of energy, leaving you physically and emotionally drained at its end. Your eye movements are analyzed algorithms unknown to you determining your productivity levels. (Paraphrased)

But for whom?

- Surveillance controls the oppressed, but cannot enforce accountability on those in power
- 1000 police officers randomly assigned body cameras
- No evidence found between officers who knew they were being watched

Why filming police violence has done nothing to stop it

After years of police body cams and bystander cellphone video, it's clear that evidentiary images on their own don't bring about change. What's missing is power.

by **Ethan Zuckerman**

June 3, 2020

But for whom?

- Even though there is no deterrence in violent behavior, one might hope for accountability afterwards
- Officers legally justified in use of deadly force if they have an “objectively reasonable” fear that they are in danger
- Videos from body cameras and bystander cell phones have worked to bolster “reasonable fear” defense claims as much as they have demonstrated the culpability of police officers.

But for whom?

NYPD used facial recognition to track down Black Lives Matter activist

Mayor Bill de Blasio says standards need to be "reassessed"

By [James Vincent](#) | Aug 18, 2020, 5:26am EDT

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority

In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group.



By [Paul Mozur](#)

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.



By [Kashmir Hill](#)

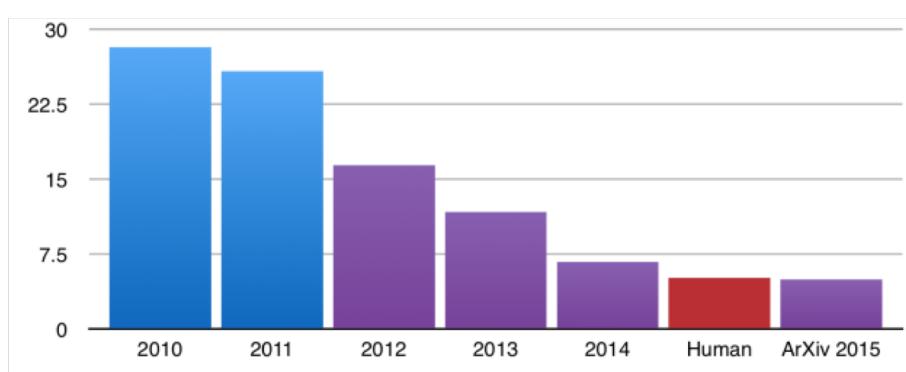
Published June 24, 2020 Updated Aug. 3, 2020

Progress for whom?

Progress in AI

Human-level average performance

Image recognition [Eckersley+ '17]



Face recognition [Harris+ '15]

TECH • GOOGLE
Google: Our new system for recognizing faces is the best one ever

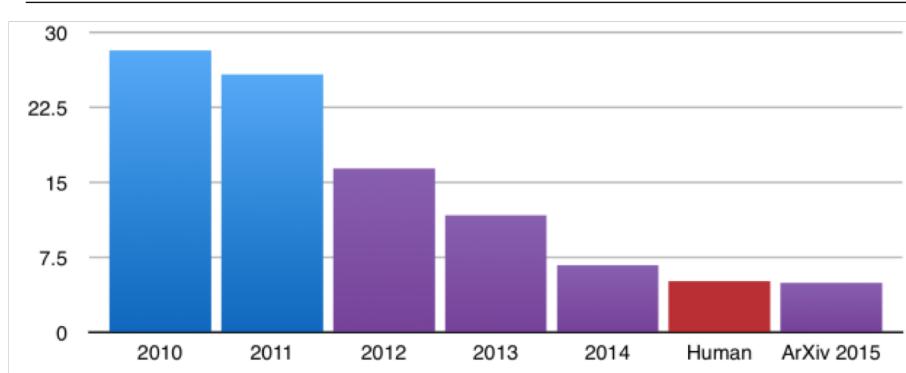
By DERRICK HARRIS March 17, 2015

FORTUNE

Progress in AI?

Human-level average performance

Image recognition [Eckersley+ '17]



Face recognition [Harris+ '15]

TECH • GOOGLE
Google: Our new system for recognizing faces is the best one ever

By DERRICK HARRIS March 17, 2015

FORTUNE

Poor performance on underrepresented examples

Amazon scraps secret AI recruiting tool that showed bias against women



Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

Feb. 9, 2018

The New York Times

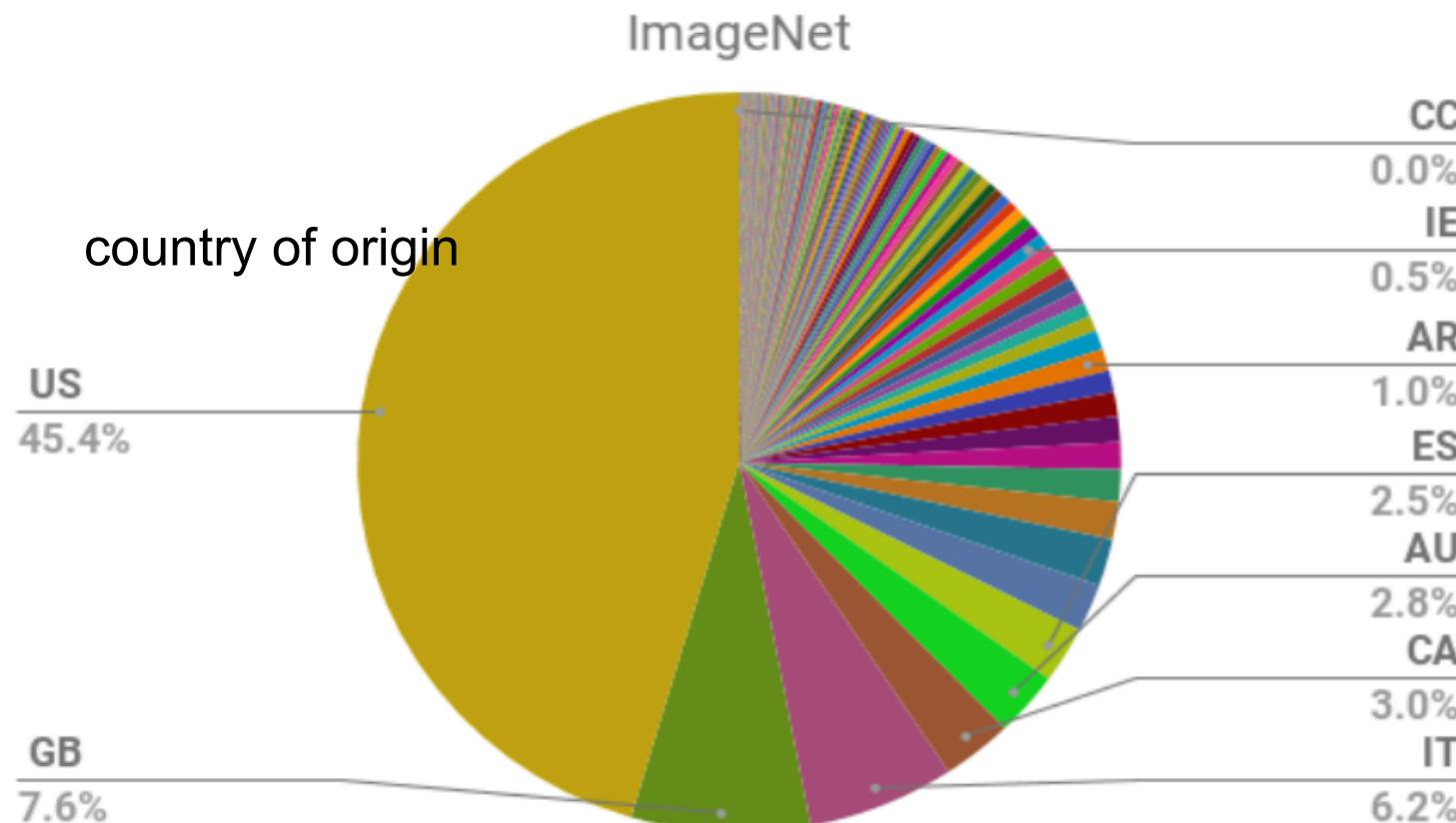
What are AI systems, really?

- They are optimizers. Models are explicitly trained to minimize prediction error on the training data.



Lack of diversity in the data

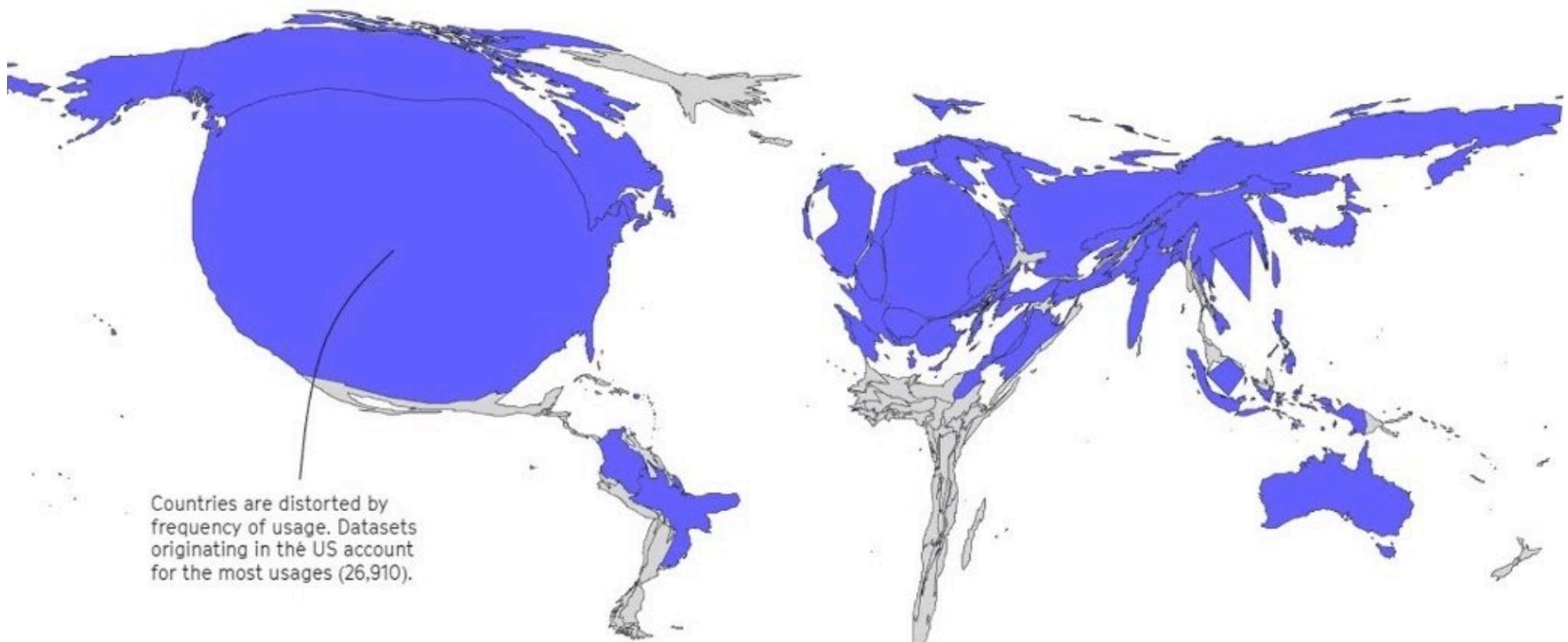
- “Clinical trials for new drugs skew heavily white”
- Less than 5% of cancer trial participants were non-white
- Majority of image data from US & Western Europe



World map as AI sees it

Frequency of dataset usage by country

● Usage of datasets from here ● No usage of datasets from here



Research by: [Koch, Denton, Hanna, and Foster \(2021\)](#)
Visual by: [The Mozilla Internet Health Report 2022](#)

Lack of diversity in the data



Facial recognition

- Labeled Faces in the Wild, a gold standard dataset for face recognition, is 78% male, and 84% White
- Commercial gender classification softwares had disparate performance on different subpopulations

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Gendered Shades:
Intersectional accuracy
disparity

An old problem

Kodak
“Shirley
cards”



- “Shirley cards” were used to calibrate colors when developing film
- Digital imaging still does not work well with dark skin tones

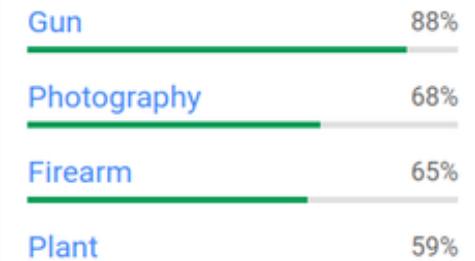
Object recognition



Screenshot from 2020-03-31 11-27-22.png



Screenshot from 2020-03-31 11-23-45.png



Machine translation



Alex Shams
@seyyedreza

Turkish is a gender neutral language. There is no "he" or "she" - everything is just "o". But look what happens when Google translates to English. Thread:

A screenshot of a Twitter thread by Alex Shams (@seyyedreza) comparing gender-neutral pronouns in Turkish and English. The interface shows 'Turkish - detected' and 'English' as the target languages. The thread lists various gender-neutral terms in Turkish followed by their English translations.

Turkish	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He/she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover

onu sevmiyor
onu seviyor

onu görüyor
onu göremiyor

o onu kucaklıyor
o onu kucaklamıyor

o evli
o bekar

o mutlu
o mutsuz

o çalışkan
o tembel

she does not like her
she loves him

she sees it
he can not see him

she is embracing her
he does not embrace it

she is married
he is single

he's happy
she is unhappy

he is hard working
she is lazy

6:36 PM · Nov 27, 2017 · Twitter Web Client

14.9K Retweets 2K Quote Tweets 27.2K Likes

Machine translation



Phoebe Tickell
@solarpunk_girl

In Hungarian, we don't use he/she there is only one gender pronoun “Ő”. But it's fascinating when this is fed through Google Translate, the algorithms highlight the biases that are there. Then imagine enacting any kind of change from those biases, encoded into computer code.

The screenshot shows the Google Translate mobile application. At the top, it says "Google Fordító". Below that, the source language is set to "MAGYAR" and the target language to "ANGOL". The input text in Hungarian is: "Ő szép. Ő okos. Ő érti a matematikát. Ő kedves. Ő egy orvos. Ő egy takarító. Ő egy politikus. Ő egy tanár. Ő erős. Ő okos. Ő sofőr. Ő bevásárol. Ő mosogat. Ő egy orvos. Ő horgászik. Ő sok pénzt keres. Ő szép. Ő okos. Ő még okosabb. Ő a legokosabb. Ő mosogat. Kapd be, Google." The output text in English is: "She is beautiful. He is clever. He understands math. She is kind. He is a doctor. She's a cleaner. He is a politician. She is a teacher. He is strong. He is clever. He's a driver. She's shopping. She washes the dishes. He is a doctor. He's fishing. He makes a lot of money. She is beautiful. He is clever. He's even smarter. He's the smartest. She washes the dishes. Get it, Google." There are icons for microphone, speaker, and a blue water drop at the bottom left. At the bottom right, it says "276 / 5000".

Speech recognition

MARCH 23, 2020

Stanford researchers find that automated speech recognition is more likely to misinterpret black speakers

The disparity likely occurs because such technologies are based on machine learning systems that rely heavily on databases of English as spoken by white Americans.



BY EDMUND L. ANDREWS

The technology that powers the nation's leading automated speech recognition systems makes twice as many errors when interpreting words spoken by African Americans as when interpreting the same words spoken by whites, according to a new study by researchers at Stanford Engineering.



Data as infrastructure

- Data provides the foundation on which we do knowledge work, and models reflect patterns in the data
- Once established, difficult to go beyond it
- Datasets are
 - **Contingent** on the social conditions of creation
 - **Constructed**: data is not objective
 - **Value-laden**: shaped by patterns of inclusion and exclusion

Data provenance

The Secretive Company That Might End Privacy as We Know It

By Kashmir Hill

A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.

“more than 600 law enforcement agencies have started using Clearview in the past year”

Data provenance

Google's DeepMind and UK hospitals made illegal deal for health data, says watchdog

The ruling concerns a 2015 agreement the AI subsidiary made with UK hospitals that has since been replaced

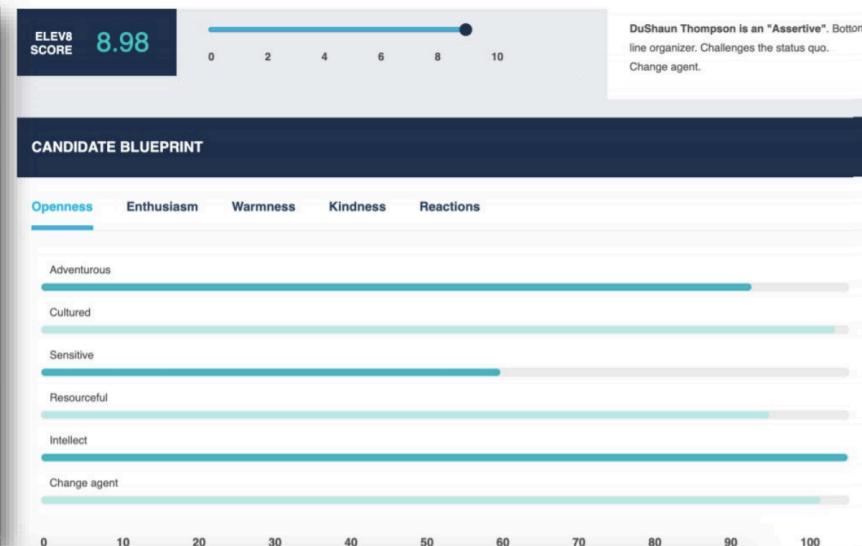
By [James Vincent](#) | Jul 3, 2017, 8:29am EDT

medical data including details of drug overdoses, abortions, and whether individuals were HIV positive, without explicit patient consent

AI snakeoil

- Inaccurate claims often lead to harmful adoptions
- Public view is often overly optimistic and inaccurate

The public predicts a 54% likelihood of high-level machine intelligence within 10 years



- If humans cannot assess competency based on a 30 second video, most likely neither can prediction models

Fixes are hard

Fixes are hard

Google using dubious tactics to target people with ‘darker skin’ in facial recognition project: sources

By GINGER ADAMS OTIS and NANCY DILLON
NEW YORK DAILY NEWS | OCT 02, 2019 AT 6:56 PM

Google hired temps go out to collect face scans from a variety of people on the street using \$5 gift cards as incentive. Homeless people and unsuspecting college students were targeted.

Employees were told to “go after people of color, conceal the fact that people’s faces were being recorded and even lie to maximize their data collections”

“not tell (people) that it was video, even though it would say on the screen that a video was taken”

Hotfixes

Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech **THE VERGE**

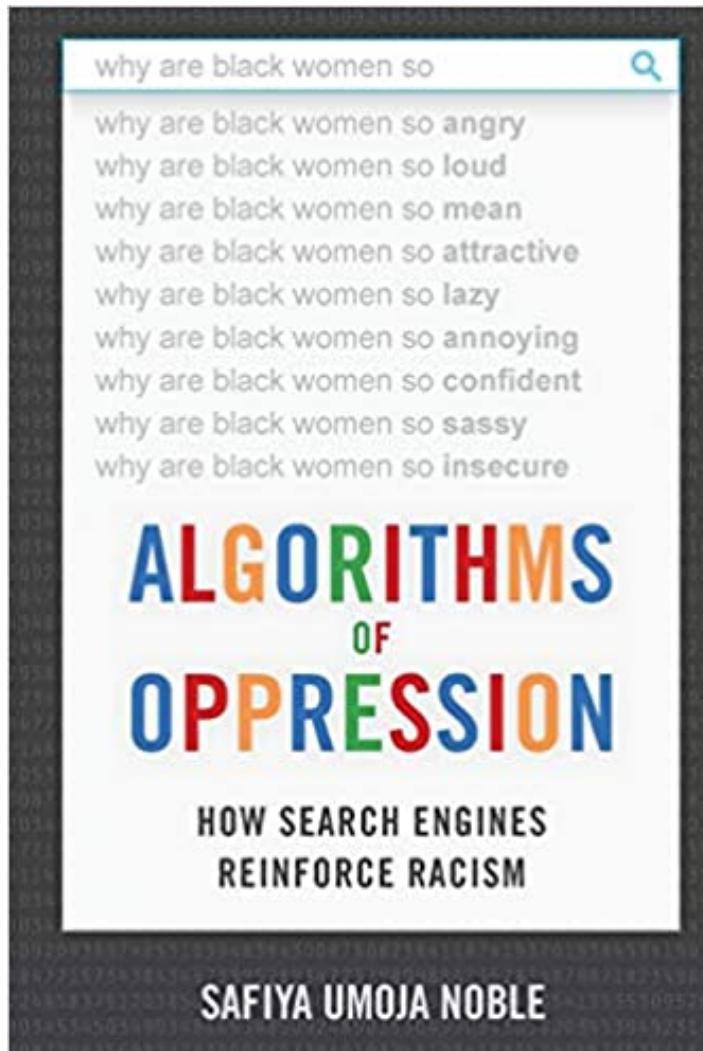
Nearly three years after the company was called out, it hasn’t gone beyond a quick workaround By [James Vincent](#) | Jan 12, 2018, 10:35am EST

Microsoft improves facial recognition technology to perform well across all skin tones, genders

June 26, 2018 | [John Roach](#)

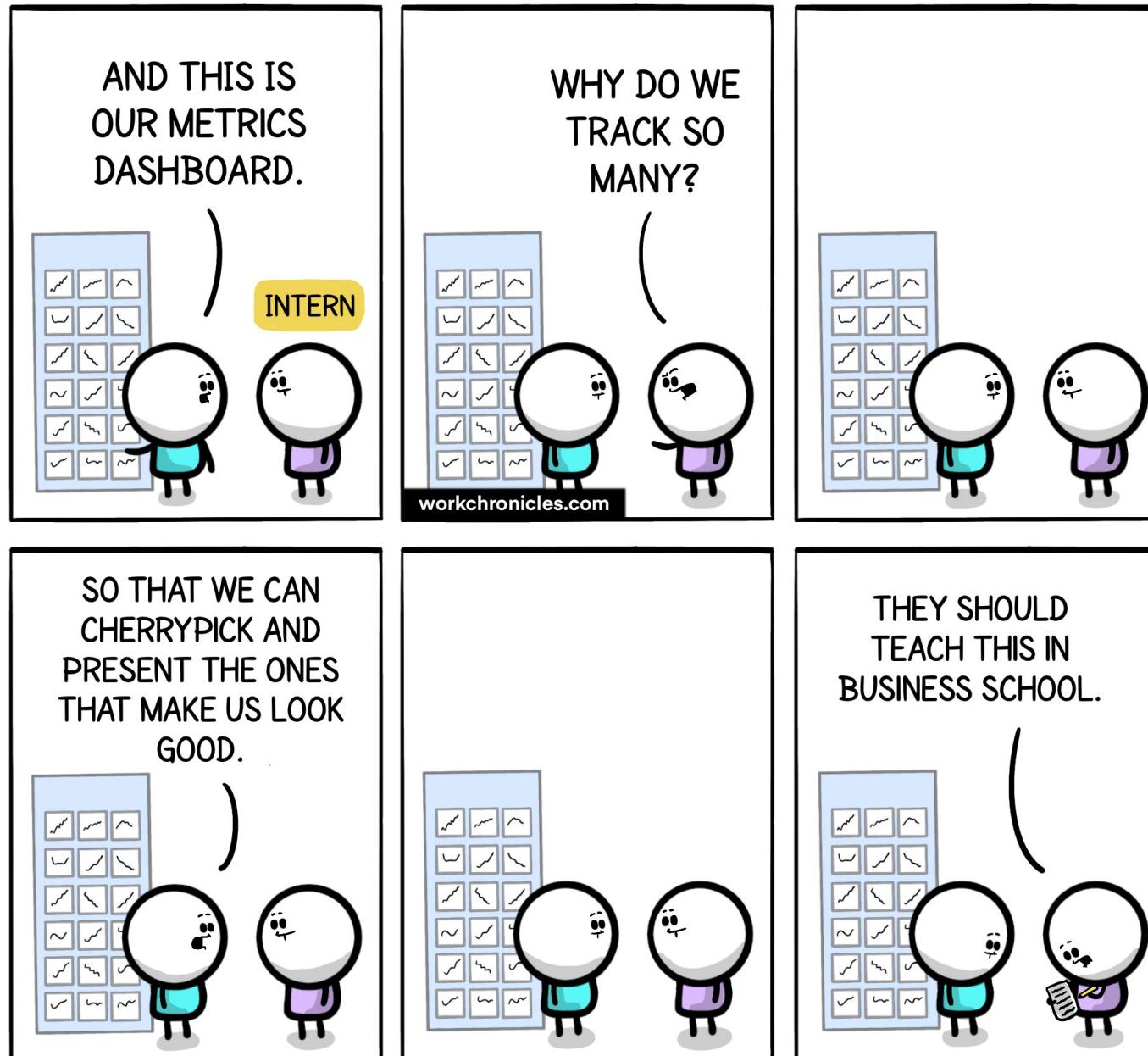
- Hotfix instead of major investments
- A “fair” gender classifier may not be the solution; prescribing gender without consent is inherently oppressive
- Using gender for prediction purposes may not be justified

Hotfixes



- Search algorithms perpetuate negative societal biases
- Noble dismantles the idea that search engines are inherently neutral: women of color and other marginalized populations are profiled and misrepresented
- Since the book, “Black girls” now give a carefully curated result, but “Asian girls” still gives inappropriate results

Metrics and incentives



Recap



4GIFs.com

Potential directions

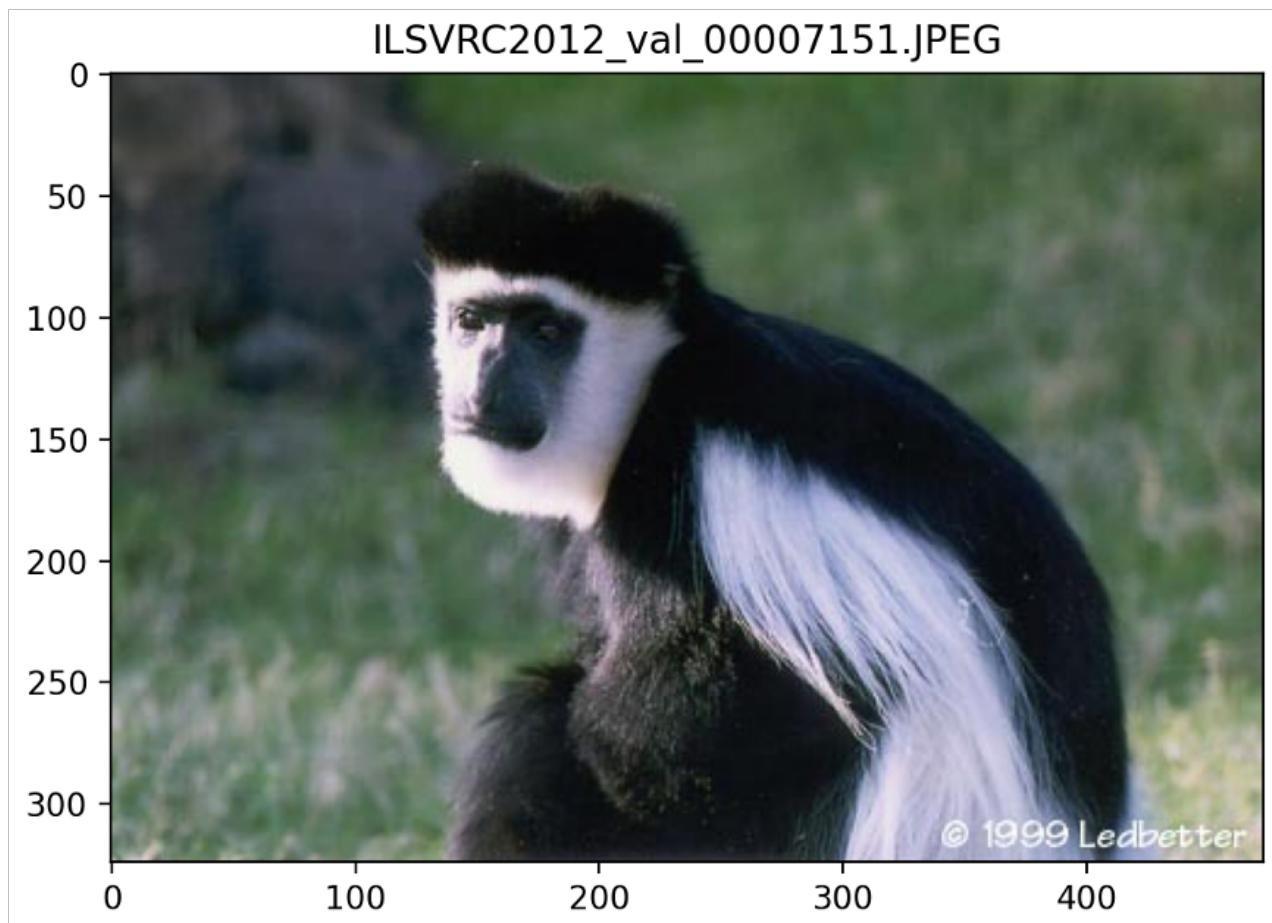
Lessons from archivists

Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning

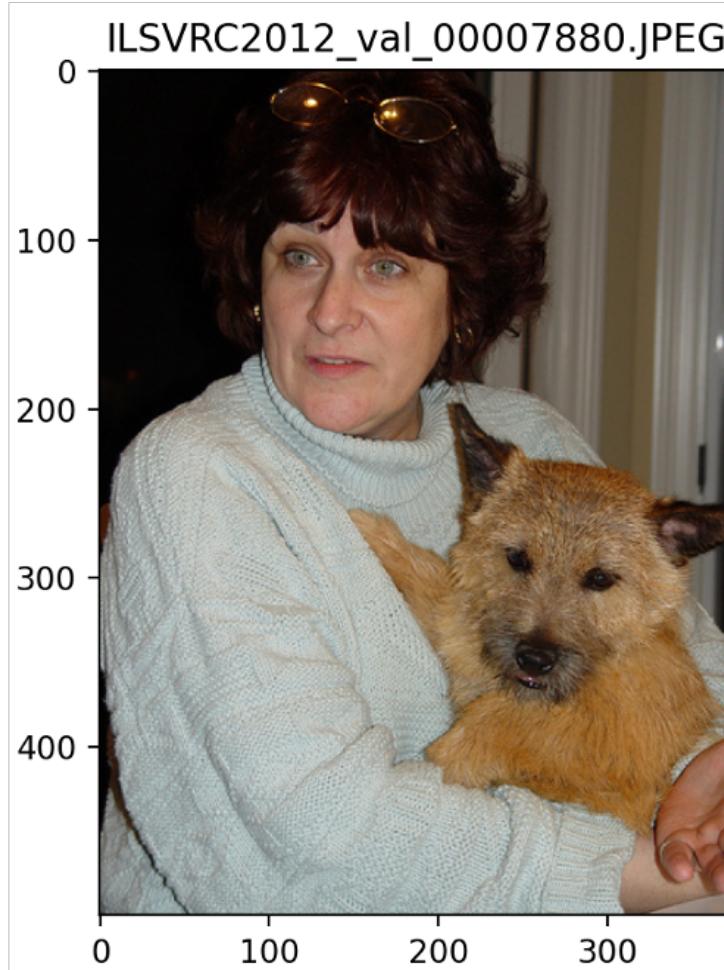
Jo and Gebru (2019)

Table 1: Lessons from Archives: summaries of approaches in archival and library sciences to some of the most important topics in data collection, and how they can be applied in the machine learning setting.

Consent	(1) Institute data gathering outreach programs to actively collect underrepresented data (2) Adopt crowdsourcing models that collect open-ended responses from participants and give them options to denote sensitivity and access
Inclusivity	(1) Complement datasets with “Mission Statements” that signal commitment to stated concepts/topics/groups (2) “Open” data sets to promote ongoing collection following mission statements
Power	(1) Form data consortia where data centers of various sizes can share resources and the cost burdens of data collection and management
Transparency	(1) Keep process records of materials added to or selected out of dataset. (2) Adopt a multi-layer, multi-person data supervision system.
Ethics & Privacy	(1) Promote data collection as a full-time, professional career. (2) Form or integrate existing global/national organizations in instituting standardized codes of ethics/conduct and procedures to review violations



n02488702 colobus, colobus monkey



n02094258 Norwich terrier

Consent

- (1) Institute data gathering outreach programs to actively collect underrepresented data
 - (2) Adopt crowdsourcing models that collect open-ended responses from participants and give them options to denote sensitivity and access
-

Power

- (1) Form data consortia where data centers of various sizes can share resources and the cost burdens of data collection and management
-

Publishers Prepare for Showdown With Microsoft, Google Over AI Tools

Media executives want compensation for use of their content in ChatGPT, Bing and Bard

By [Keach Hagey](#) [Follow](#), [Alexandra Bruell](#) [Follow](#), [Tom Dotan](#) [Follow](#) and [Miles Kruppa](#) [Follow](#), THE WALL STREET JOURNAL.

Data sheets

Gebru et al. (2020)

<https://arxiv.org/pdf/1803.09010.pdf>

- Documentation for datasets
- Akin to guidelines for archivists
- Primary audience: dataset creators & consumers
 - But also useful for policy makers, consumer advocates, study participants etc
- Motivation, composition, collection process, pre-processing, intended uses, maintenance etc

Model cards



Led by Dr. M. Mitchell and Dr. T. Gebru

- We should treat AI models as any other engineered product; document intended use cases and limitations
- Documentation for models providing benchmarked evaluations across race, geographic location, sex, skin type
- Factors for performance; instrumentation such as photo quality
- Usage contexts, evaluation & testing details

Back to power

- Timnit Gebru and Margaret Mitchell, leading researchers in fairness and ethics in AI, were fired from Google Research in 2020



Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.

Timnit Gebru, one of the few Black women in her field, had voiced exasperation over the company's response to efforts to increase minority hiring.

The New York Times



Google fires Margaret Mitchell, another top researcher on its AI ethics team

The Guardian

- Technical solutions aren't enough; power structure replicates and perpetuate in a myriad of ways

Structural representation

- Technology is political and value-laden
- The following does not free you from the social context you work in
 - I am a theoretical researcher
 - I work on basic research
 - I am a engineer
 - I do this out of technical interest
- You can play a unique role in recognizing, communicating, advocating, facilitating, and organizing around the varied powers and interests embedded in every AI system

Transition

Validation & Safety Testing

Safety testing

- AI-based products have an exceptional range of use cases
- This is one of their main appeal: AI models are versatile and aim to adapt to each situation (personalization)
- But this poses a substantial challenge to safety testing
- There are countless edge cases and models are often not used as intended / naively assumed by the engineers

Alexa

- It is difficult for voice assistants to account for social context



Alexa

- It is difficult for voice assistants to account for social context

Alexa tells 10-year-old girl to touch live plug with penny



⌚ 28 December 2021

A 10 yo asked Alexa for a “challenge to do”. Alexa responded with "Plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs"

Tesla

- Tesla's self-driving systems are notorious for only using visual information, rather than other sensors such as LiDAR
- This makes the entire system brittle to varied edge cases



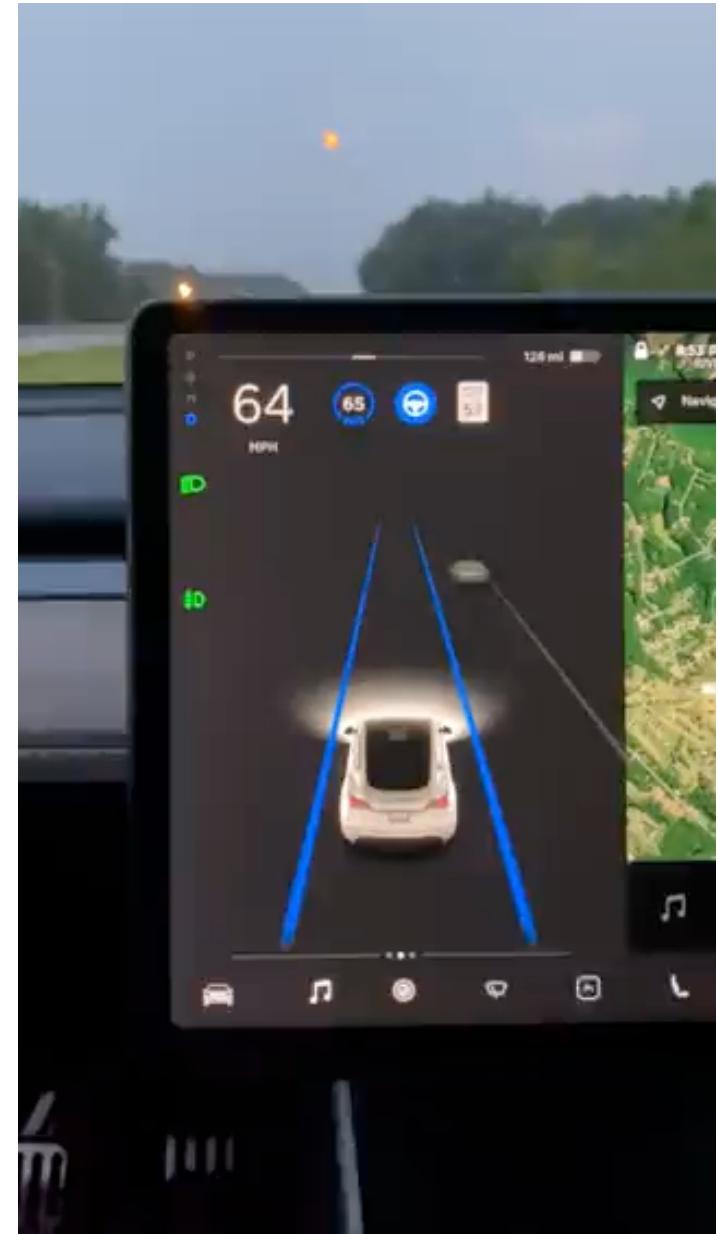
Owner: "Car kept jamming on the brakes thinking this was a person"

<https://twitter.com/TaylorOgan/status/1469404579439824899?s=20>

<https://twitter.com/besf0rt/status/1372205422426357766?s=20>

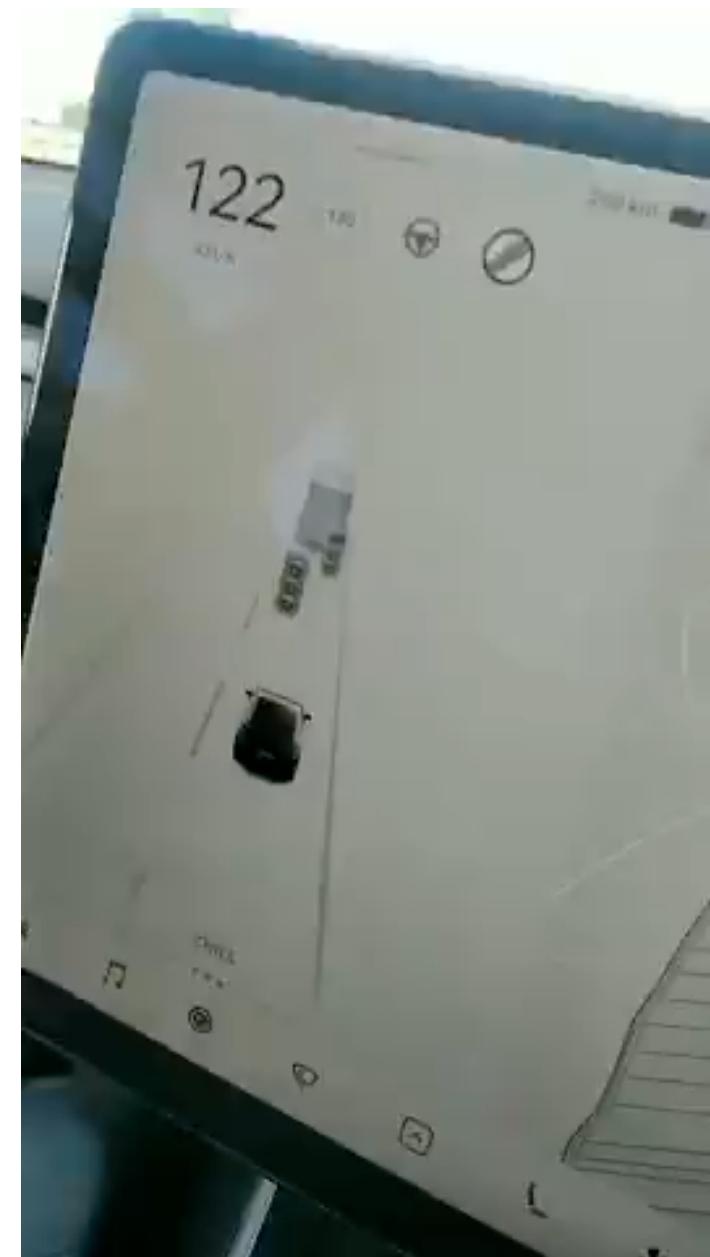
Tesla

- Tesla's self-driving systems are notorious for only using visual information, rather than other sensors such as LiDAR
- This makes the entire system brittle to varied edge cases

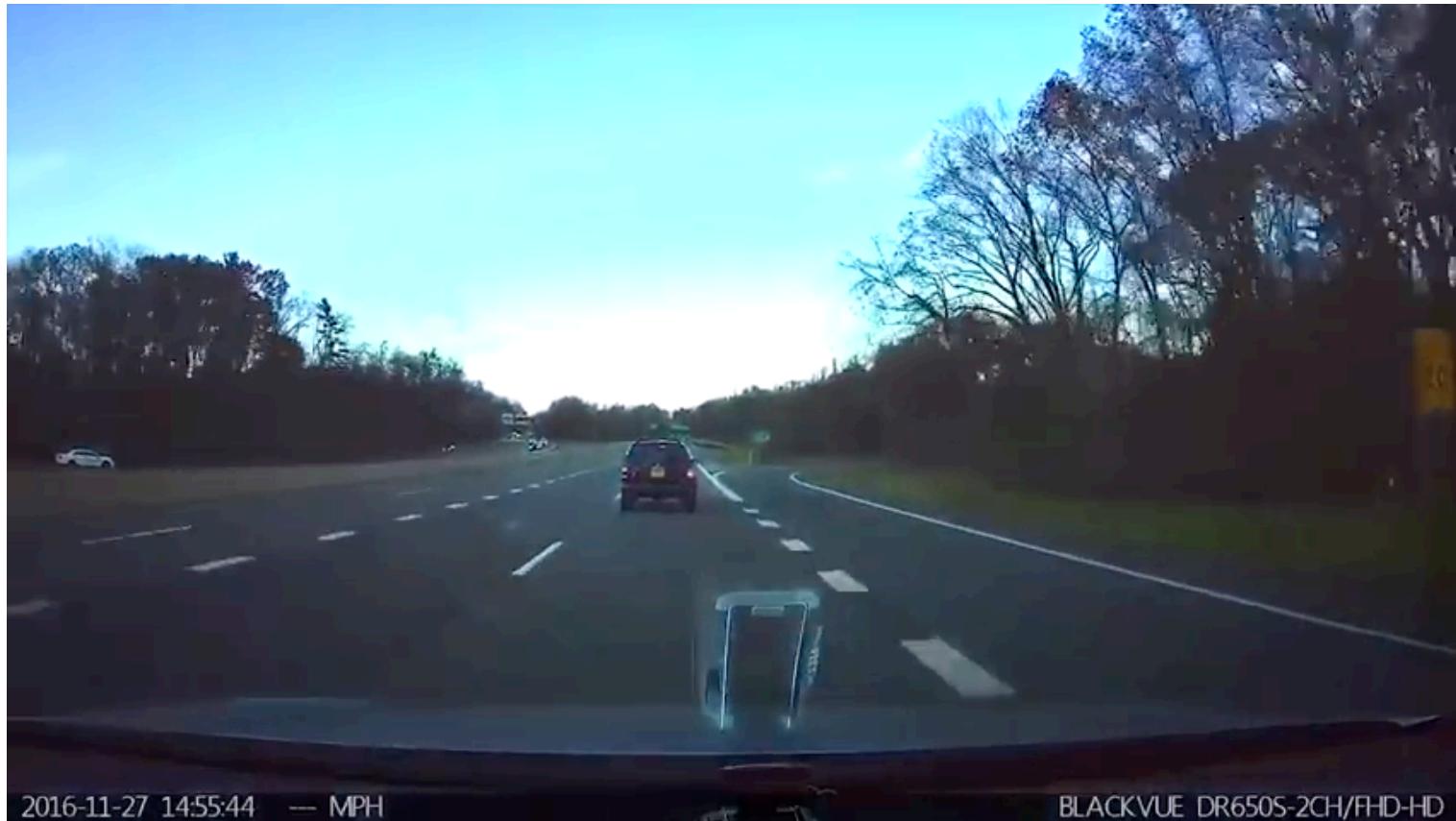


Tesla

- Tesla's self-driving systems are notorious for only using visual information, rather than other sensors such as LiDAR
- This makes the entire system brittle to varied edge cases



Tesla



Tesla

PREFECT



2016-05-19 07:03:59 --- km/h

DR650GW-2CH/FHD-HD

Tesla



Tesla

Catherine Guo



Federal Government Opens Safety Defect Investigation Into Tesla Autopilot Crashes

NHTSA is looking at whether the technology may be a contributing factor in multiple crashes with emergency vehicles

By Keith Barry

Published August 16, 2021 | Updated September 1, 2021

iRobot



Google Search

Kannada: Google apologises for 'ugliest Indian language' search result

BBC

4 June 2021



ugliest language in india



All

Videos

Images

News

Shopping

M

Kannada

What is the **ugliest language in India**? The answer is Kannada, a **language** spoken by around 40 million people in south **India**.

Soccer

AI Camera Ruins Soccer Game For Fans After Mistaking Referee's Bald Head For Ball



Garbage in, garbage out

Real World Example



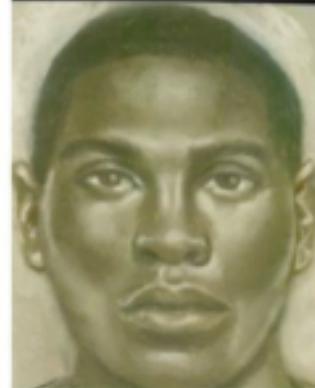
aws re:Invent
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



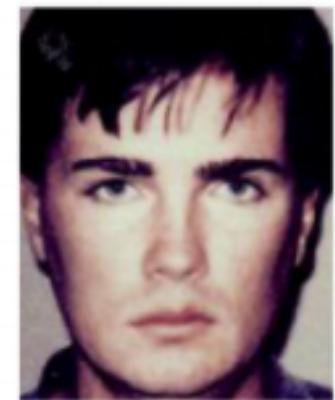
Figure 2: Slide from an AWS presentation titled "Washington County Sheriff's Office Rekognition Case Study."
(Source: Public records obtained by ACLU Oregon & Northern California.)



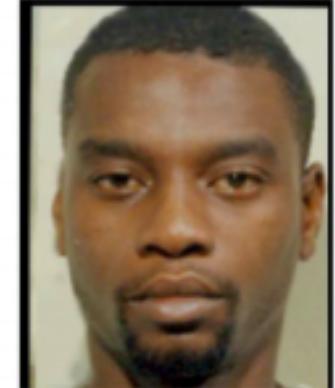
Probe Sketch



Top Retrieval



True Subject



<https://www.flawedfacedata.com/>

Spurious correlation

- Correlation is no substitute for **causal** evidence
- COVID prediction AIs were found to be “picking up on the text font that certain hospitals used to label the scans.”
- “As a result, fonts from hospitals with more serious caseloads became predictors of covid risk.”

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

**Technology
Review**

Spurious correlation

- Correlation is no substitute for **causal** evidence

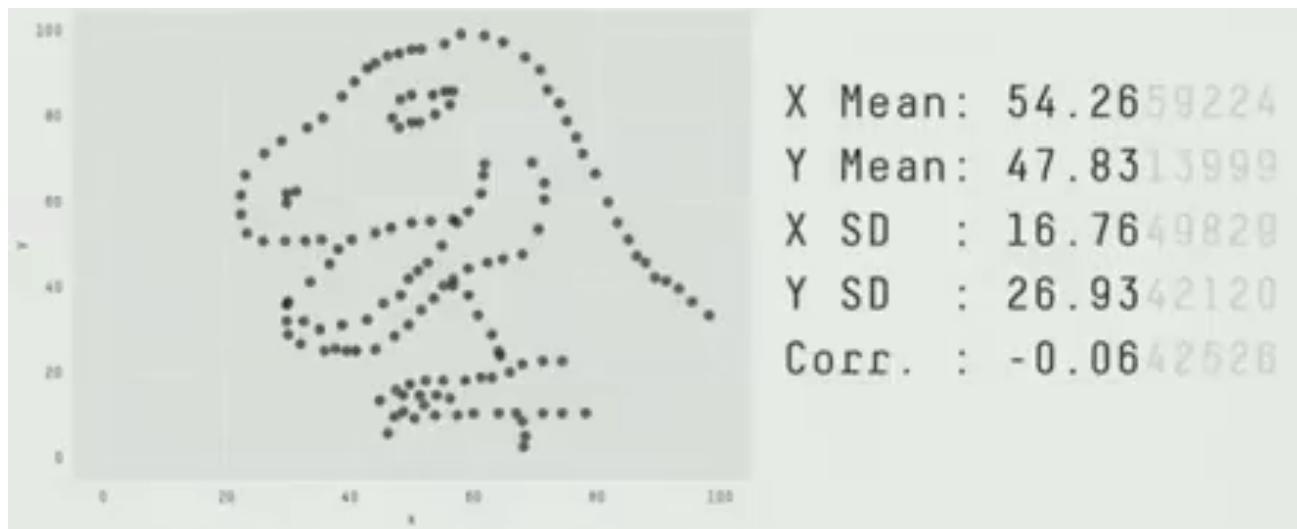


Amazon scraps secret AI recruiting tool that
showed bias against women

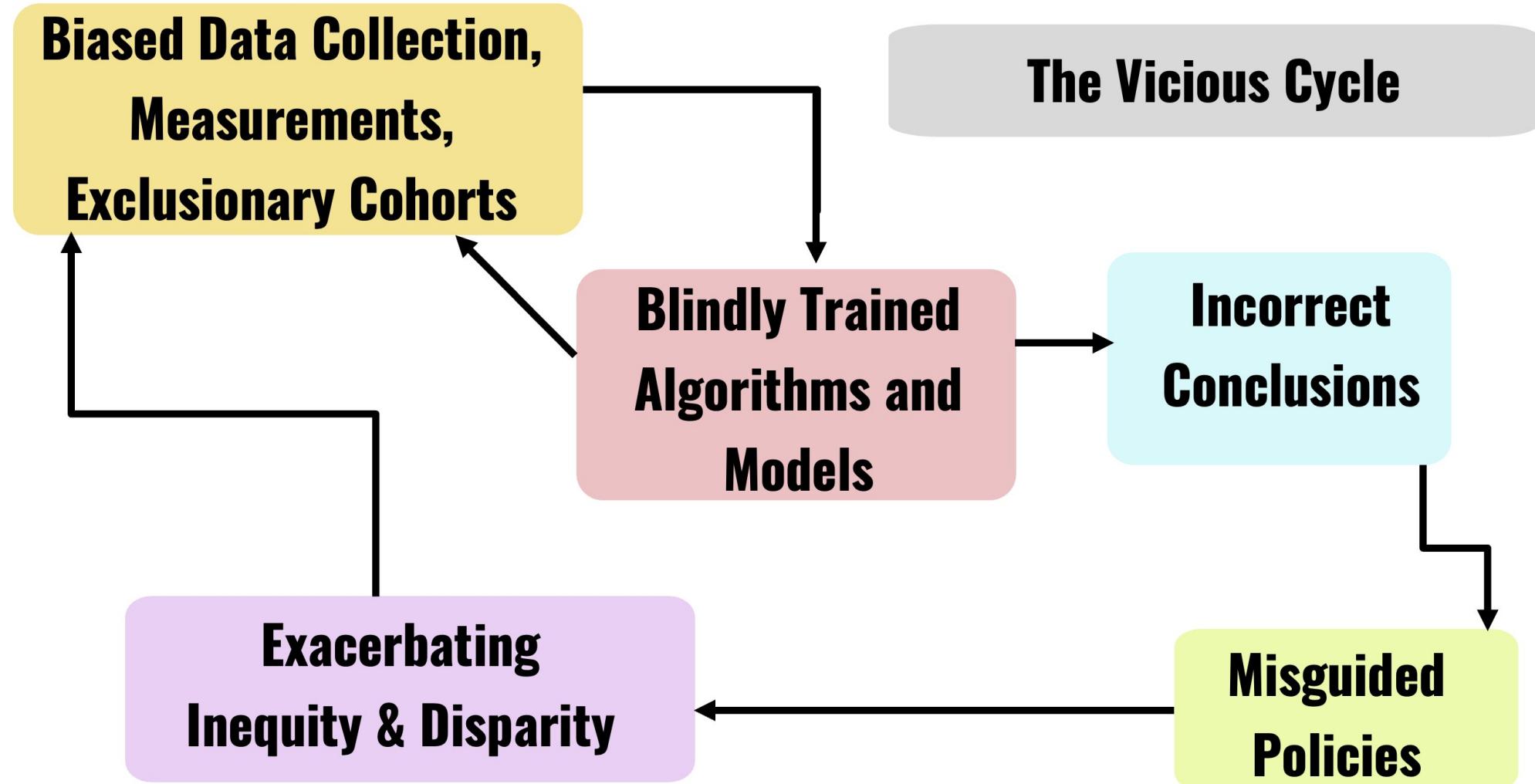
 REUTERS

Takeaways

- Don't let average-case metrics fool you!
- Often you will need to inspect inputs who suffer high prediction error for insights
- Summary statistics like mean prediction error is limiting, especially in light of systematic biases in data collection



Beware of vicious cycles



Summary

- Before you deploy a model, need to validate across a range of different datasets
- The more diverse that validation data, the better: across space, time, demographics, and labels
- You must continually monitor model performance after deployment; relationship between outcome and features may change
- It is crucial to design incentives around careful validation, monitoring, and maintenance

Fairness definitions and inherent trade-offs

Links

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<https://github.com/propublica/compas-analysis>

<https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html>

Chouldechova (2016) <https://arxiv.org/abs/1610.07524>

Kleinberg et al. (2016) <https://arxiv.org/abs/1609.05807>

https://www.youtube.com/watch?v=jIXluYdnyyk&ab_channel=ArvindNarayanan



COMPAS

Correctional Offender Management Profiling for Alternative Sanctions

- Used in prisons across US: AZ, CO, DL, KY, LA, OK, VA, WA, WI
 - Even used for sentencing in Wisconsin, California, New York
- Predicts recidivism = whether reoffend in two years
- Differential treatments across the judicial system based on risk score (likelihood of recidivism)
 - affects bail amount, waiting longer for parole, even sentencing
- Can't observe recidivism, so they use re-arrests as proxy
 - “Labels” are already heavily biased against Blacks

Risk scores attached to defendants unreliable

By Julia Angwin, Jeff Larson, Surya Mattu And Lauren Kirchner, ProPublica

"The first time Paul Zilly heard of his risk assessment score — and realized how much was riding on it — was during his sentencing hearing on Feb. 15, 2013, in a Barron County courtroom.

Zilly had been convicted of stealing a push lawn mower and some tools. The prosecutor recommended a year in county jail and follow-up supervision that could help Zilly with "staying on the right path." His lawyer agreed to a plea deal.

But Judge James Babler had seen Zilly's score.

The defendant was rated a high risk for future violent crime and a medium risk for general recidivism. "When I look at the risk assessment," Babler said in court, "it is about as bad as it could be."

Babler overturned the plea deal and imposed two years in state prison and three years of supervision."

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ProPublica: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

- Analyzed risk scores of 7,000+ people in 2013-2014

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%

- Among those who did not re-offend, Black defendants receive higher risk score than white counterparts
- Among those who re-offend, white defendants receive lower risk score than Black counterparts

Machine Bias

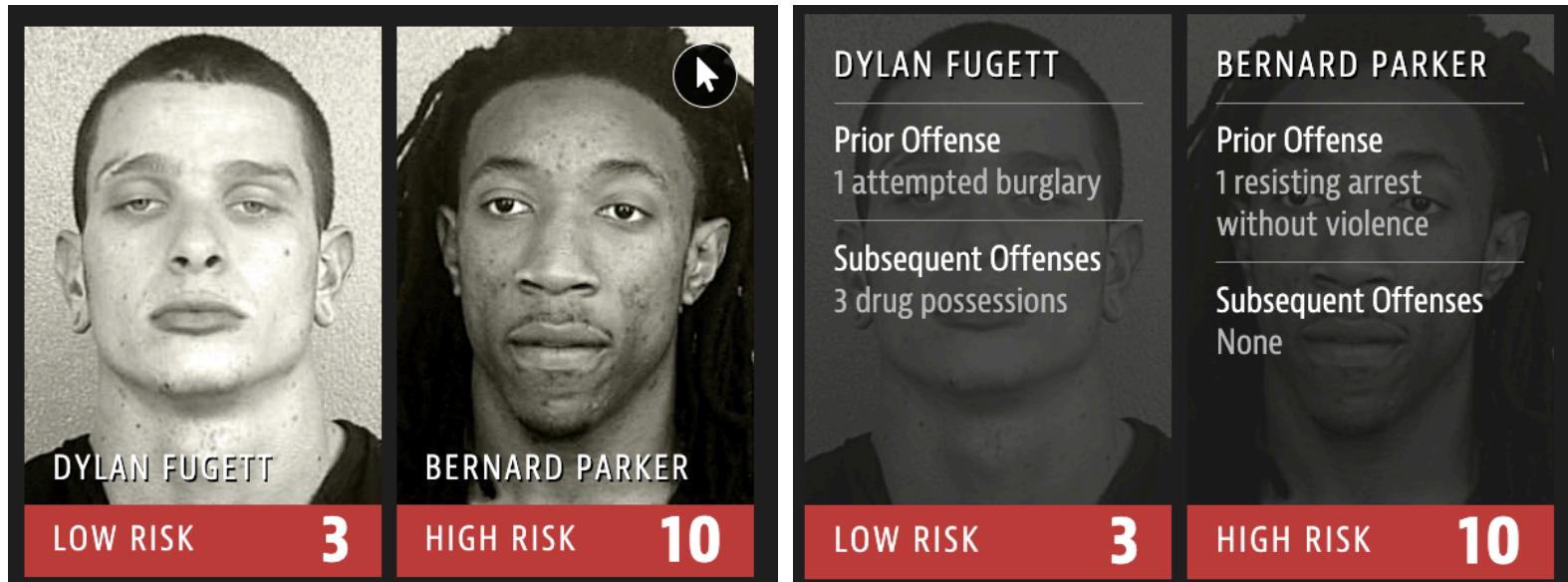
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Higher false positive rates (FPRs) and lower false negative rates (FNRs) for black defendants than for white defendant

- Hugely problematic, without even explicitly using race
- But prediction accuracy similar for both groups



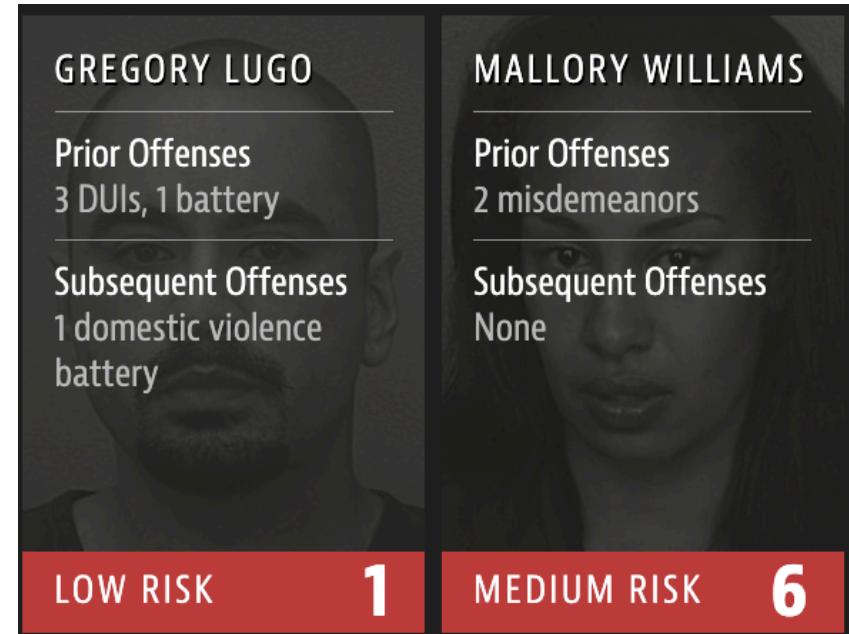
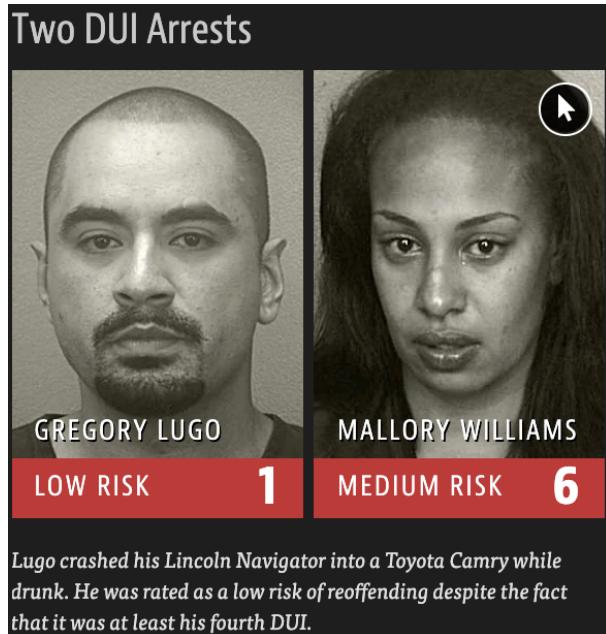
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Higher false positive rates (FPRs) and lower false negative rates (FNRs) for black defendants than for white defendant



Today: stylized perspective as a guide to various fairness definitions

Fairness

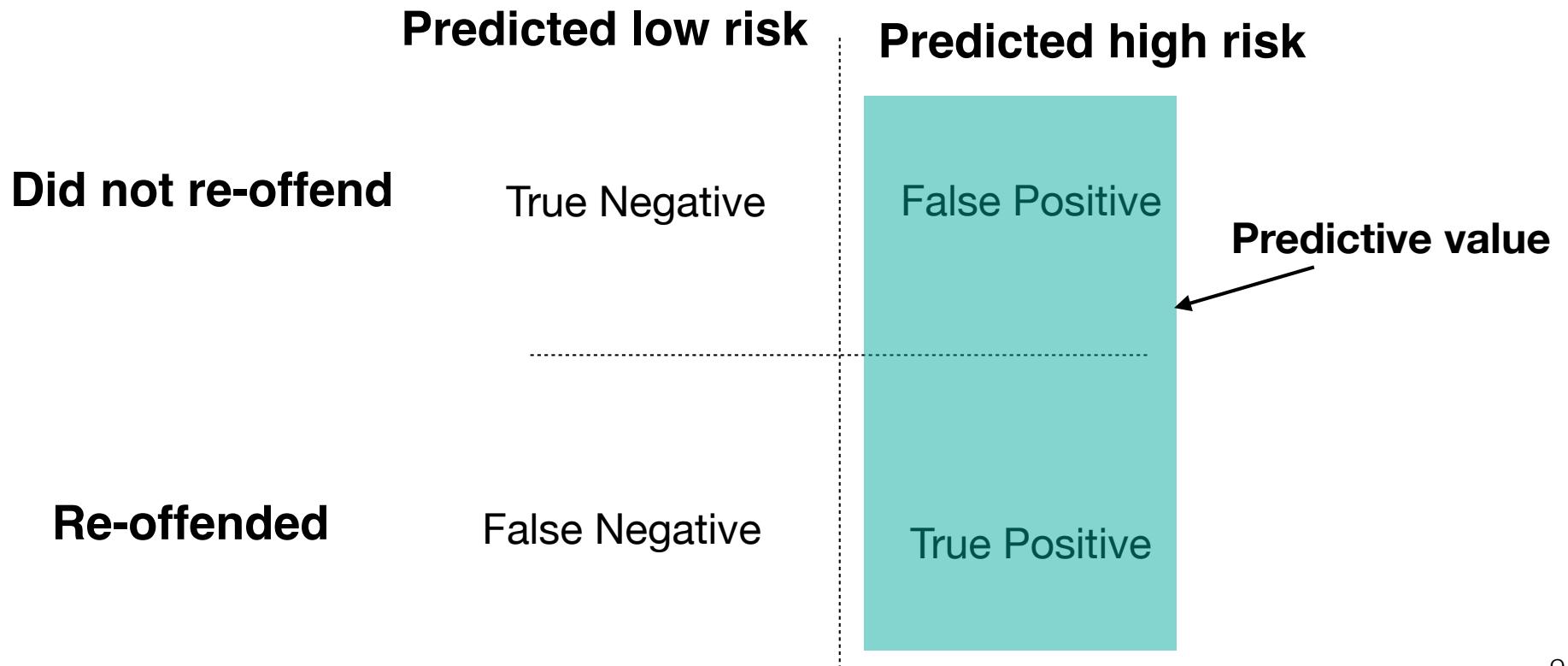
- Surprisingly common position among engineers: my model describes my data well, so my algo is faultless
- Make algorithmic systems support human values
 - Statistical bias is not enough
- Which values should it support?
- We consider a simple binary classification problem with pre-defined groups

Simple setup

	Predicted low risk	Predicted high risk
Did not re-offend	True Negative	False Positive
Re-offended	False Negative	True Positive

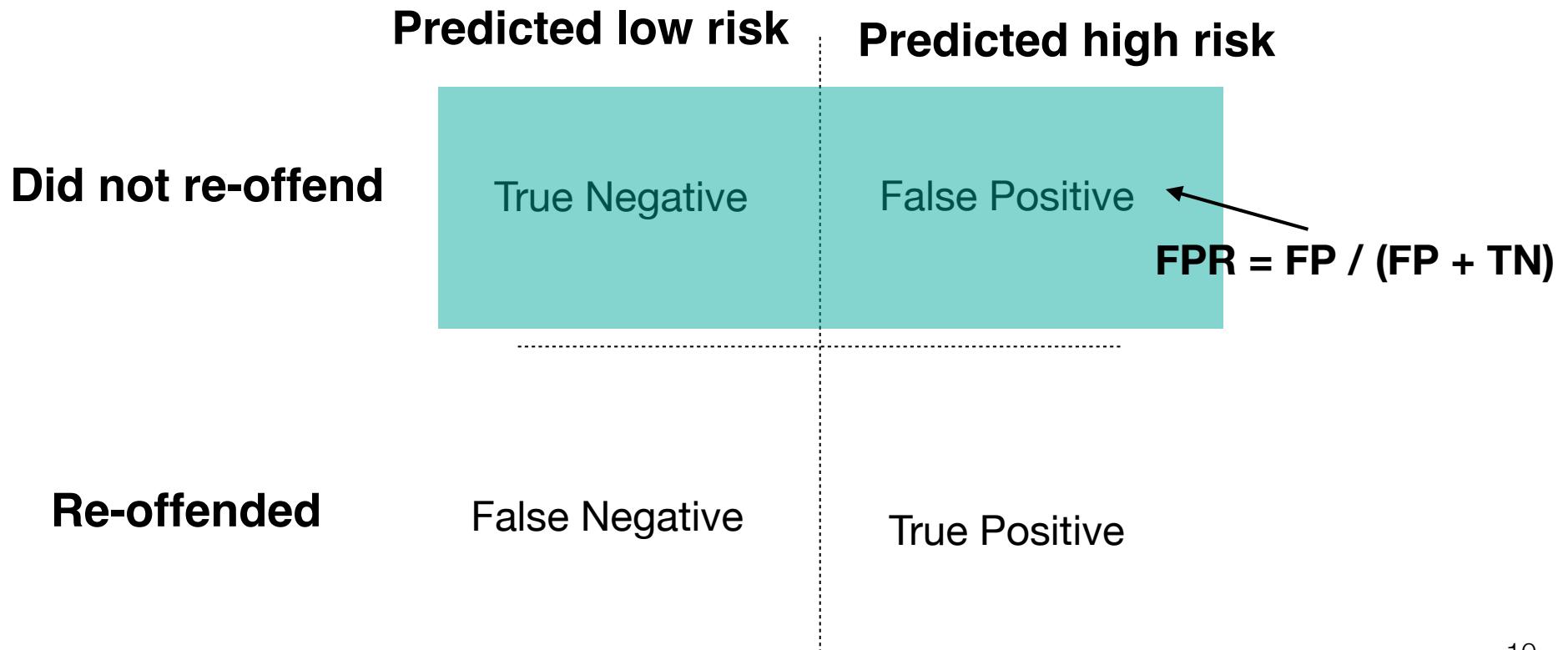
Perspectives matter

- Viewpoints vary substantially between stakeholders
- **Decision-maker / Northpointe:** of those I've predicted high-risk, what fraction will re-offend?



Perspectives matter

- Viewpoints vary substantially between stakeholders
- **Defendant:** what is the probability I'm wrongly labeled high-risk?



Perspectives matter

- Viewpoints vary substantially between stakeholders
- **Prosecution/law enforcement:** of those re-offend, how many did the system mark as high risk? (recall)

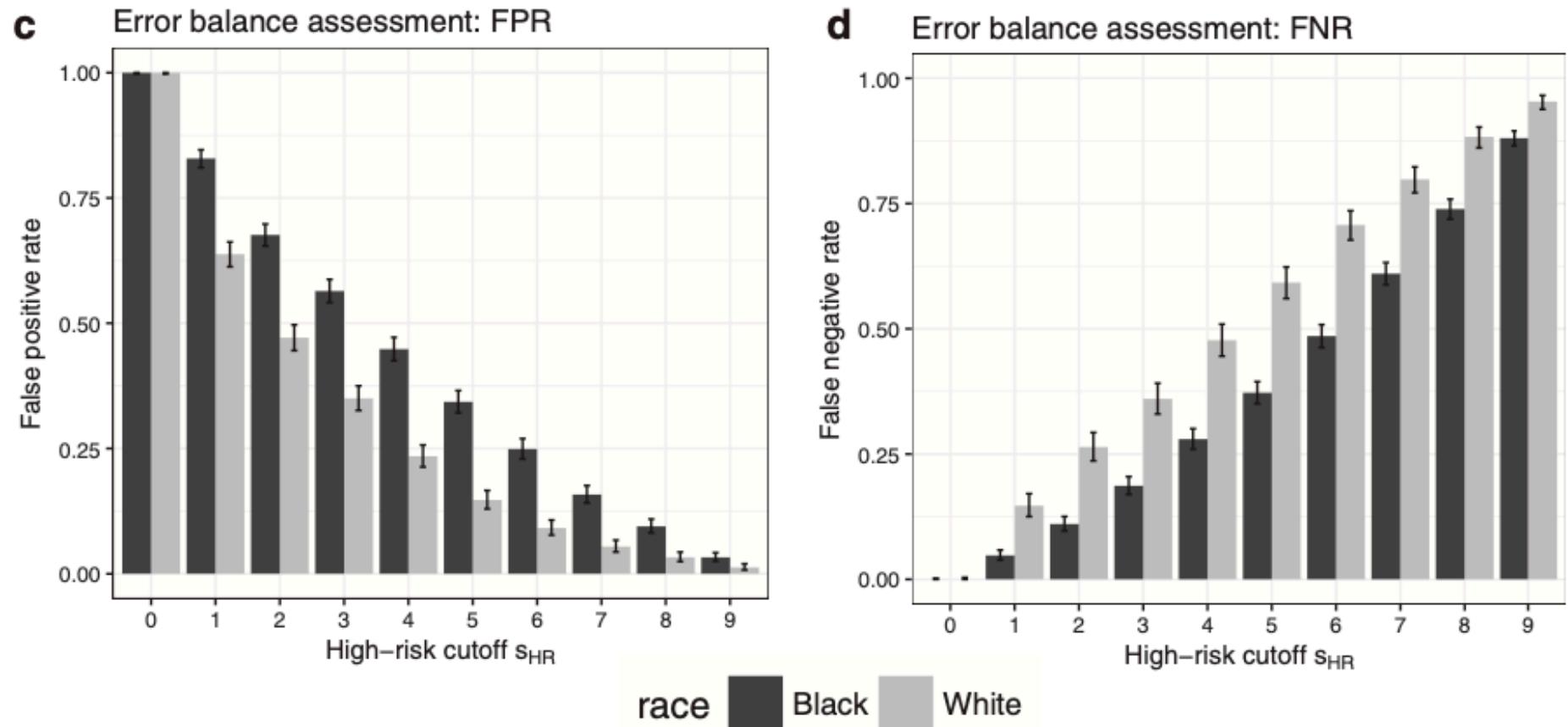
	Predicted low risk	Predicted high risk
Did not re-offend	True Negative	False Positive
Re-offended	False Negative	True Positive

Fairness definitions

- Consider fixed demographic groups
 - Let's consider Race = Black vs White
- Predictive parity
 - Equalize predictive value $P(Y = 1 | \text{predicted high risk}, R = *)$ across groups
- Error rate balance
 - Equalize FPR and FNR across groups, where $FPR = FP / (FP + TN)$, $FNR = FN / (FN + TP)$
 - Equalize $P(\text{predicted high risk} | Y = -1, R = *)$, $P(\text{predicted low risk} | Y = 1, R = *)$ across groups

COMPAS

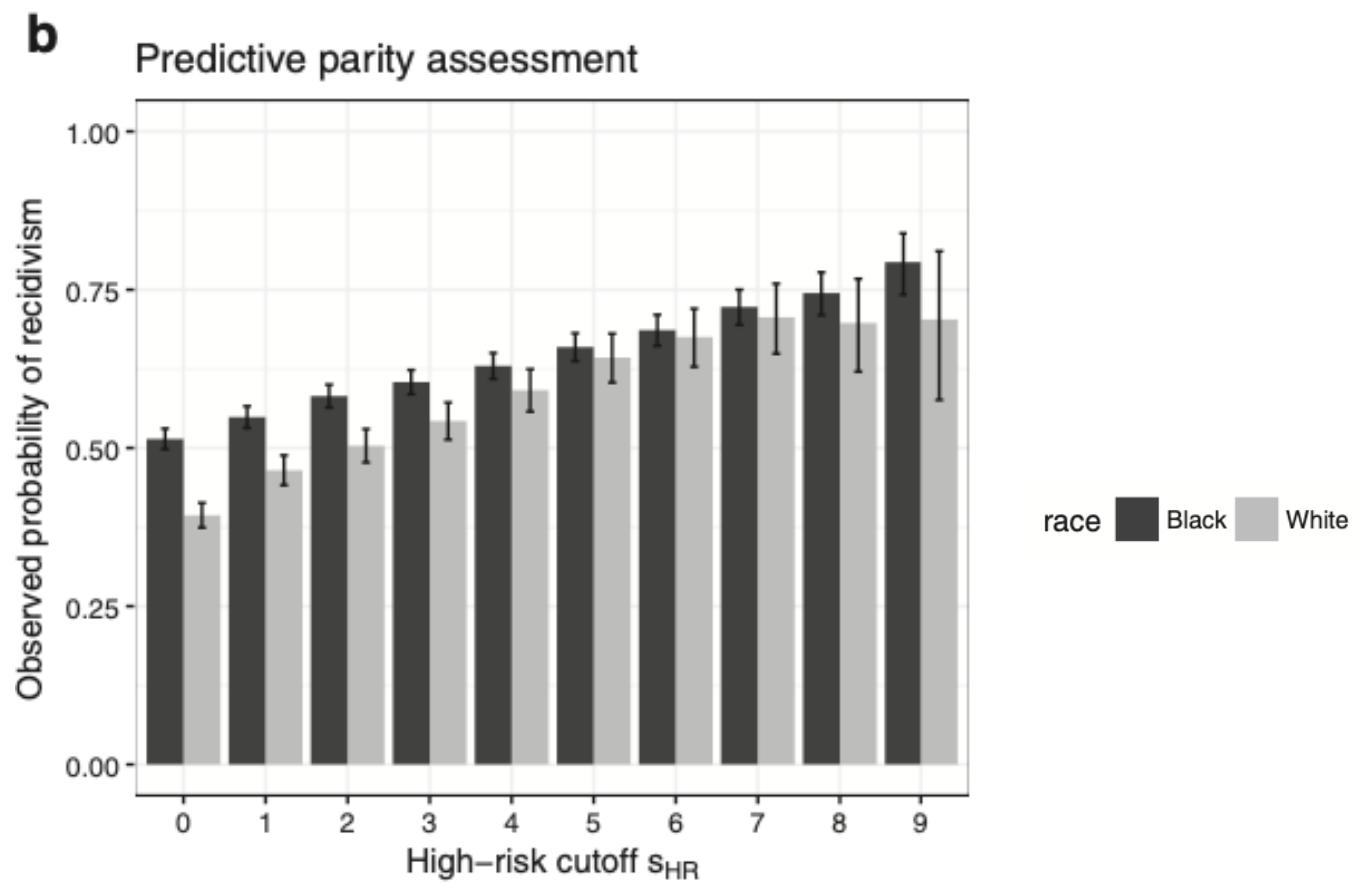
ProPublica: Higher false positive rates (FPRs) and lower false negative rates (FNRs) for black defendants than for white defendant (focused on cutoff ≥ 4)



Chouldechova (2016)

COMPAS

Northpointe: But predictive parity holds (well, kind of)



Chouldechova (2016)

Impossibility

Chouldechova (2016)

- Focus on relevant metrics in the COMPAS case: FPR, FNR, and predictive value
- Assume different prevalence across groups
 - Otherwise groups are identical from classification viewpoint
 - Race is only a proxy for determining prevalence. Determinants are often poverty and structural racism

Chouldechova (2016)

If a classifier satisfies predictive parity, i.e., identical $P(Y = 1 | \text{predicted high risk}, R = *)$ across groups, then it cannot jointly balance FPR and FNR

Proof

- Denote prevalence by p , (positive) predictive value by PV
- For each group

$$(1-p) PV * FPR = p * (1-PV) * (1- FNR)$$

$$\rightarrow FPR = p * (1-PV) * (1- FNR) / ((1-p) PV)$$

- So if p is different across groups, but PV is equalized, no way to equalize FPR *and* FNR across groups

Impossibility

- The result doesn't say anything about statistics nor computation
- Population level (non-)existence result
- Not limited to algorithmic decisions; impossibility applies to any decision mechanism including humans
- We can imagine showing similar results for other definitions

Managing trade-offs?

- How can we manage this trade-off?
 - Which one should we give up?
 - Equalize linear combination of multiple criteria?
- Very domain-dependent (previous caveats apply)
 - Balancing FPR makes sense from defendant's perspective
- Many papers equalize linear combination of two criteria, and train models over constraints / penalty terms
 - This is often not enough
- In the COMPAS context, Chouldechova recommends dispensing predictive parity, and equalizing FPR / FNR

More fairness definitions

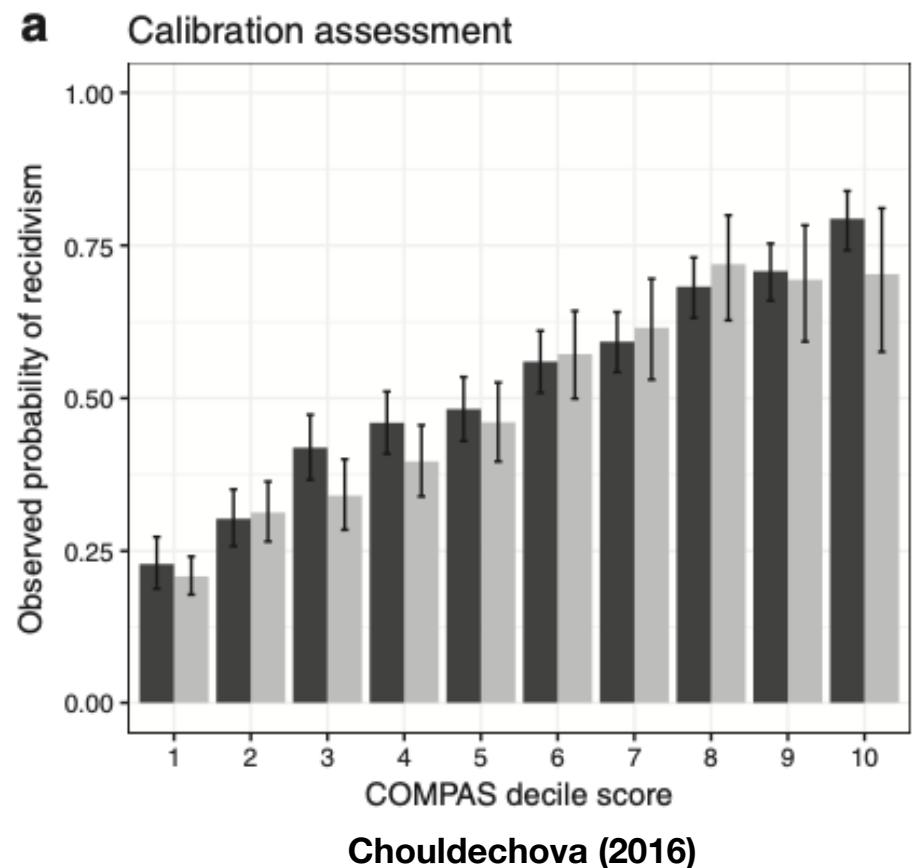
		True condition			
		Condition positive	Condition negative	Prevalence $= \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection, $\text{Power} = \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate $= \frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$		

Wikipedia: Evaluation of binary classifiers

Lots and lots of potential impossibility results

Calibration

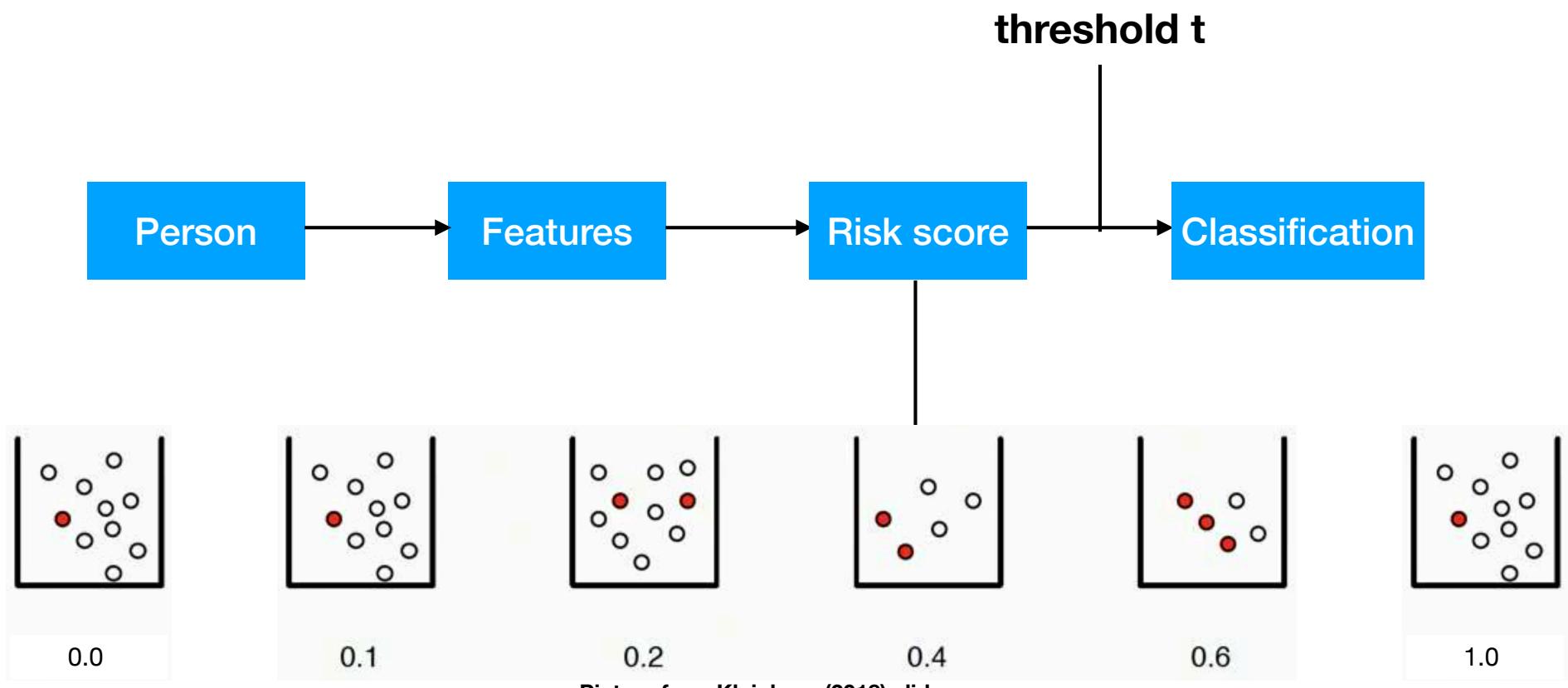
- Northpointe: COMPAS scores were well-calibrated within each group
- For all Black/white defendants with score s , (approximately) s fraction of them actually re-offends
 - It's nice that outputs actually mean what they claim
- But this is meaningless to Black defendants who won't re-offend but still receive high risk scores



Calibration

- Calibration can be useful in other contexts
 - Example: medical diagnoses
- Hospital uses uncalibrated scores w.r.t. gender to hire doctors
 - Candidate with highest score hired
 - Let's say female doctors with score s is likely to be good doctors with prob larger than that for males
 - Every patient now wants to be treated by female doctors

Setup



Discrete bins with risk scores

Goals

- Calibration within group
 - For each group, each bin with score s has $s\%$ positive people
- Balance in positive class
 - For every group, average score of positive people is same
- Balance in negative class
 - For every group, average score of negative people is same

ProPublica argued #2 and #3 does not hold for COMPAS

Impossibility

Kleinberg et al (2016)

- All three properties can be achieved in only the following two cases
 - Perfect prediction: every feature can be perfectly classified; risk score is always 0 or 1, with perfect accuracy
 - Groups are indistinguishable: every group have the same fraction of positive people
We can always predict this number for everyone
- Similar result for approximate fairness definitions

Proof sketch

Kleinberg (2018) slides

- In each group g , let N_g be the # people, k_g be expected # people in positive class
- By calibration, $k_g = \text{total score in group } g$
- Let x be the average score in negative class
- Let y be the average score in positive class
- **Since we've equalized averages, x and y are independent of group g**

Proof sketch

Kleinberg (2018) slides

- N_g = # people in group g, k_g = expected # group-g people in positive class
- By calibration, k_g = total score in group g
- x = average score in negative class, y = average score in positive class

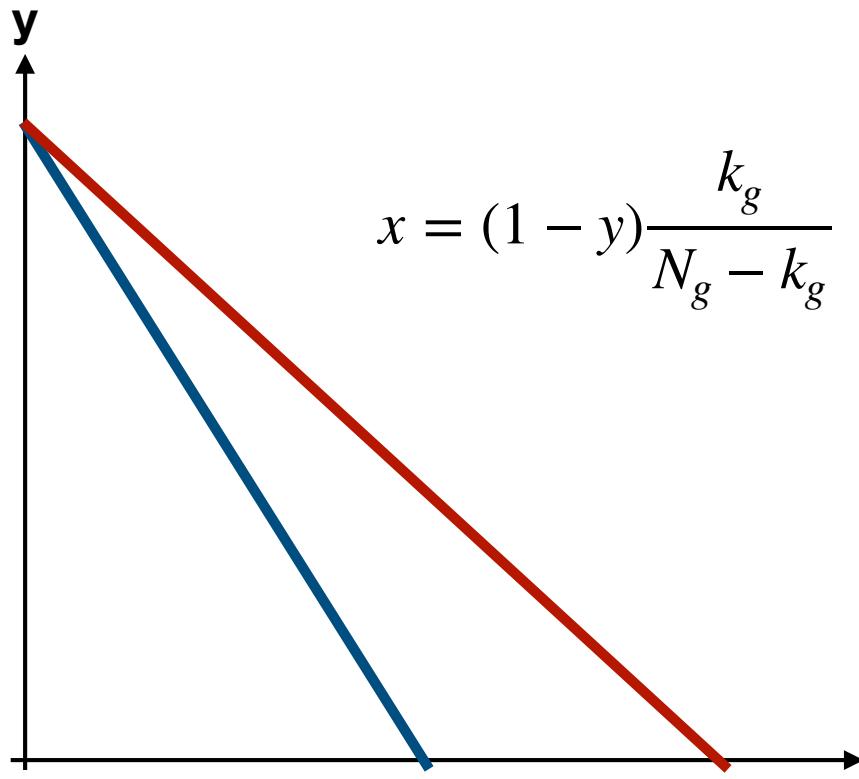
$$k_g = \text{total score in group g} = (N_g - k_g)x + k_g y$$

Imposes constraints on (x, y) space

$$x = (1 - y) \frac{k_g}{N_g - k_g}$$

Proof sketch

Kleinberg (2018) slides



- Case 1:** If slopes are different, feasible region = $(0, 1)$ => perfect classifier
Case 2: Slopes are identical across groups => identical prevalence

Representational harm

- So far, allocative harm, where system withholds resources and opportunity
- Representational harm is when system reinforces subordination of a group (e.g. stereotyping)
 - Harm may be more subtle, but has long-term effects
- Ex: Google image search on CEO used to show all white men

Kay et al. (2015)

Further questions

- Individualized notions of fairness?
- Causality
- How do we define groups?
 - Intersectionality is important
- Going beyond classification scenarios
 - utility, regression
 - complex interaction between prediction & decision
- Strategic behavior, dynamics across time and space
- Connections with mechanism design?