

A critical perspective to **Fairness, Accountability, Transparency, Ethics in ML**

Other links:

CVPR Tutorial by Timnit Gebru and Emily Denton

<https://sites.google.com/view/fatecv-tutorial/schedule?authuser=0>

ICLR Talk by Ruha Benjamin

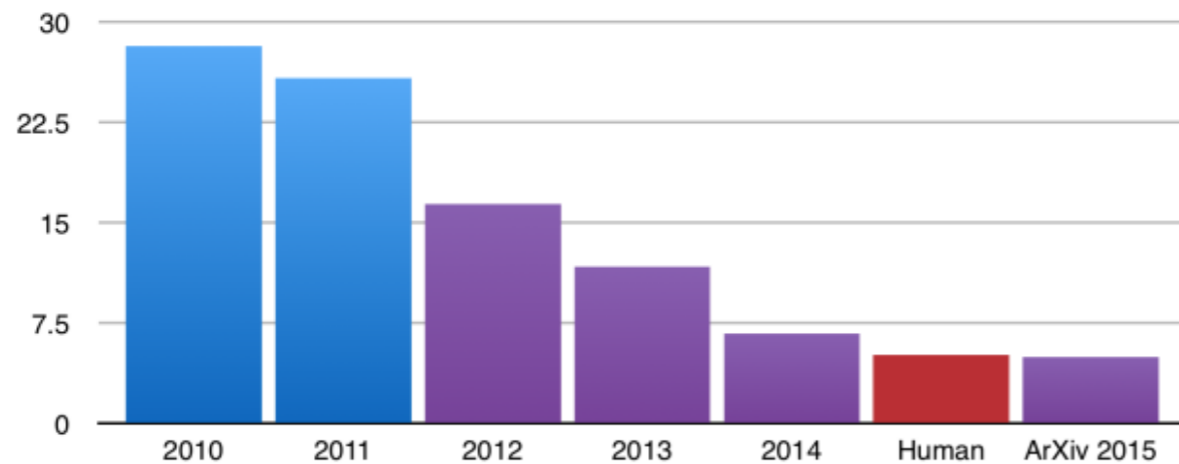
https://iclr.cc/virtual_2020/speaker_3.html

Neurips tutorial by Kate Crawford

https://www.youtube.com/watch?v=fMym_BKWQzk&ab_channel=TheArtificialIntelligenceChannel

Progress in machine learning

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

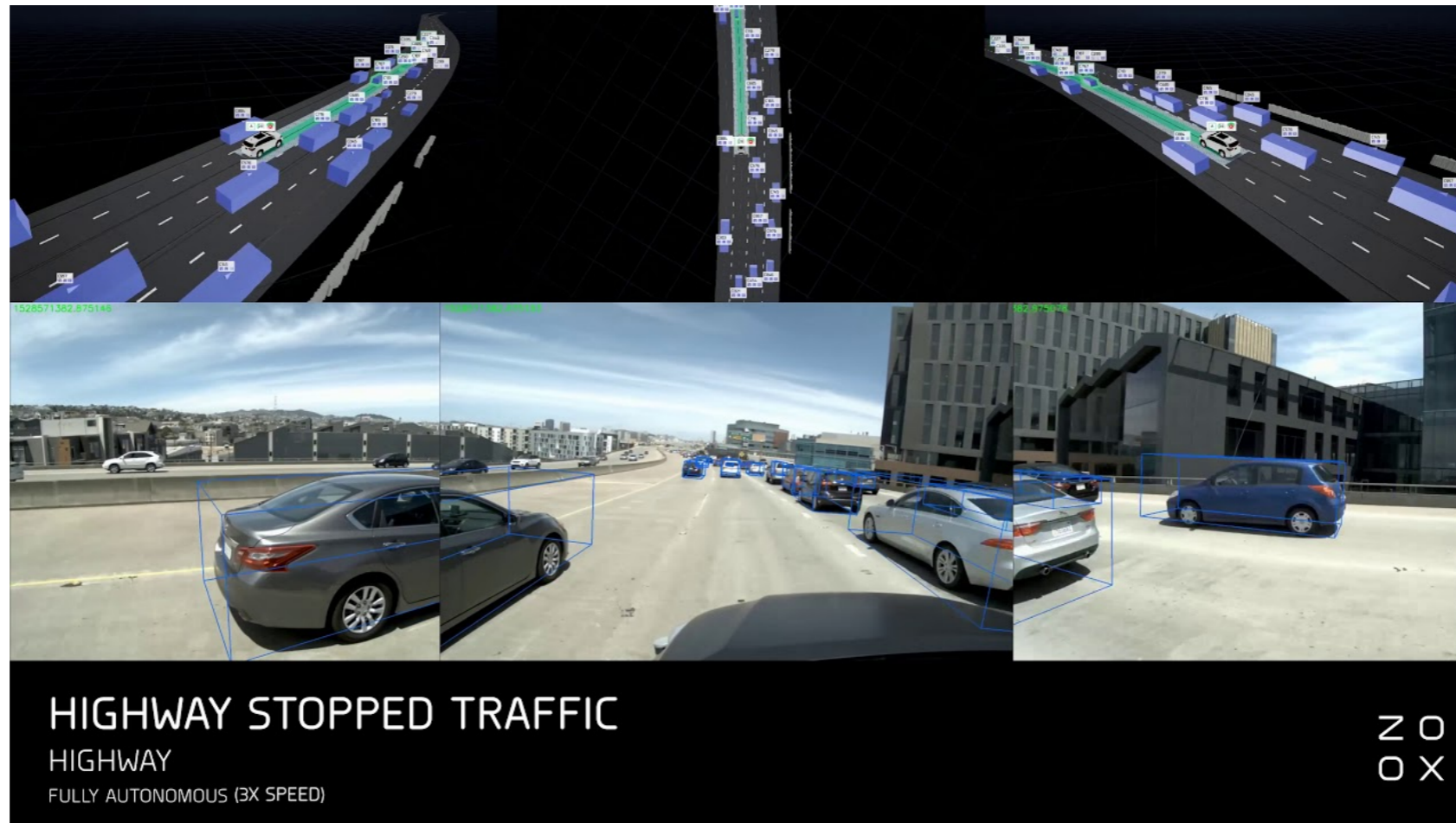
Training time

Rank	Time to 93% Accuracy	Model	Hardware	Framework
1 Mar 2020	0:02:38	ResNet50-v1.5 <i>Apsara AI Acceleration(AIACC) team in Alibaba Cloud source</i>	16 ecs.gn6e-c12g1.24xlarge (AlibabaCloud)	AIACC-Training 1.3 + Tensorflow 2.1

Training cost

Rank	Cost (USD)	Model	Hardware	Framework
1 Mar 2020	\$7.43	ResNet50-v1.5 <i>Apsara AI Acceleration(AIACC) team in Alibaba Cloud source</i>	1 ecs.gn6e-c12g1.24xlarge (AlibabaCloud)	AIACC-Training 1.3 + Tensorflow 2.1

Autonomous driving



Introducing the Voyage G3 Robotaxi

A better, safer shared vehicle for a COVID-19 world

 [Oliver Cameron](#) Aug 26 · 8 min read

Waymo's robo-taxi service opens to the public in Phoenix

GPT-3


Latest language model from Open AI

Describe a layout.

Just describe any layout you want, and it'll try to render below!

A div that contains 3 buttons each with a random color.

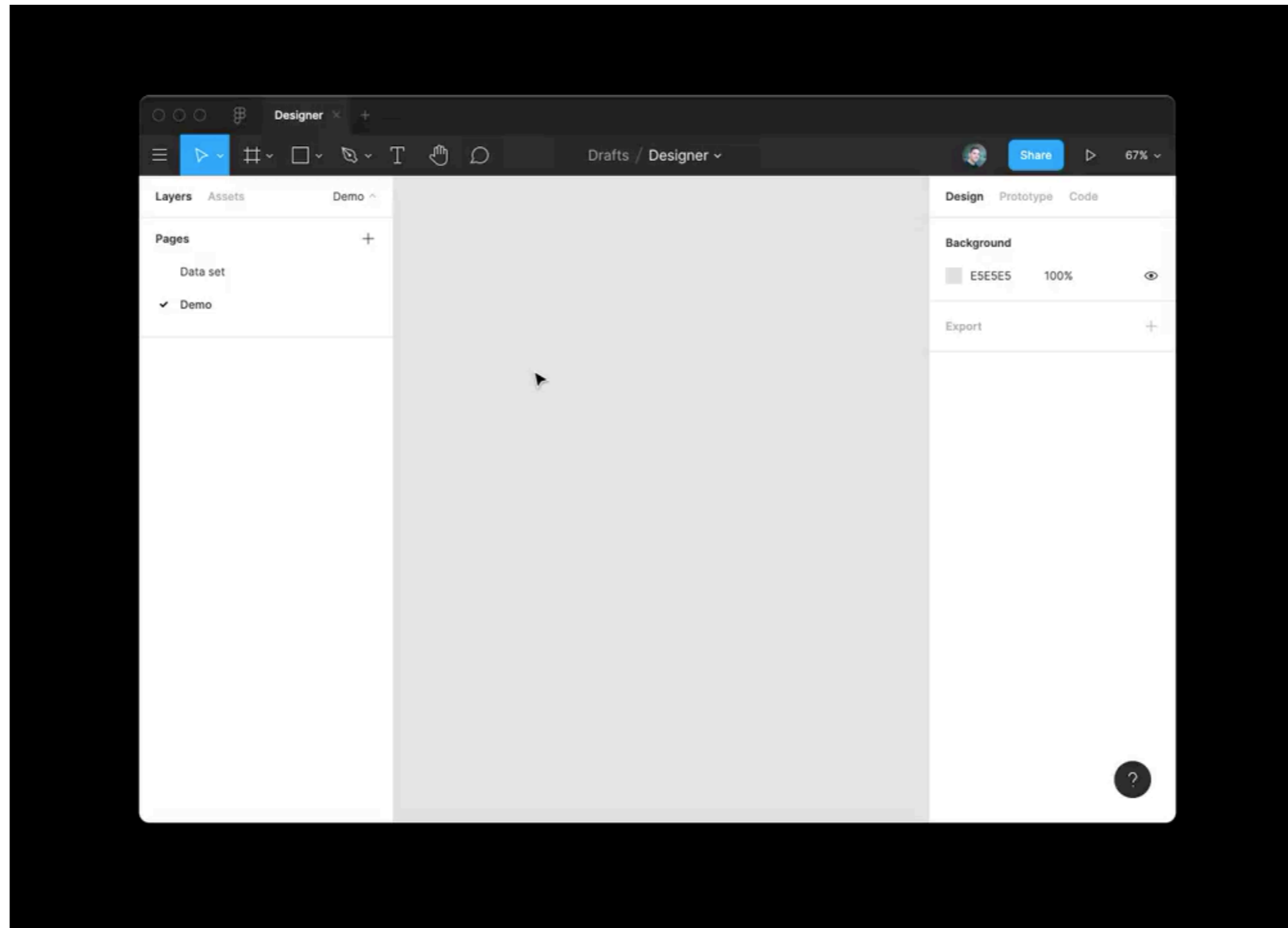
Generate



<https://twitter.com/sharifshameem/status/1282676454690451457>

GPT-3

Latest language model from Open AI



<https://twitter.com/jsngr/status/1284511080715362304>

Potential of AI

- AI technology offers lots of 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 ML 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 constantly targeted and surveilled against
- Like any other technology, but with a wider reach, and omni-present

Anthropological/Artificial Intelligence & the HAI (Paraphrased)

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.

But for whom?

- Surveillance controls the oppressed, but cannot enforce accountability on those in power

How China Uses High-Tech Surveillance to Subdue Minorities

By Chris Buckley and Paul Mozur

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

Facial recognition

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

AI Interviews

Hire★Vue

**Video interview software
and platform that makes
hiring simple**

JOBFLEX

좋은 기업과 좋은 인재의 Seamless한 연결

Data provenance

The Secretive Company That Might End Privacy as We Know It

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.

By [Kashmir Hill](#)

Published Jan. 18, 2020 Updated Feb. 10, 2020

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

IBM Research Releases ‘Diversity in Faces’ Dataset to Advance Study of Fairness in Facial Recognition Systems

IBM will no longer offer, develop, or research facial recognition technology

IBM's CEO says we should reevaluate selling the technology to law enforcement

By [Jay Peters](#) | [@jaypeters](#) | Jun 8, 2020, 8:49pm EDT

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

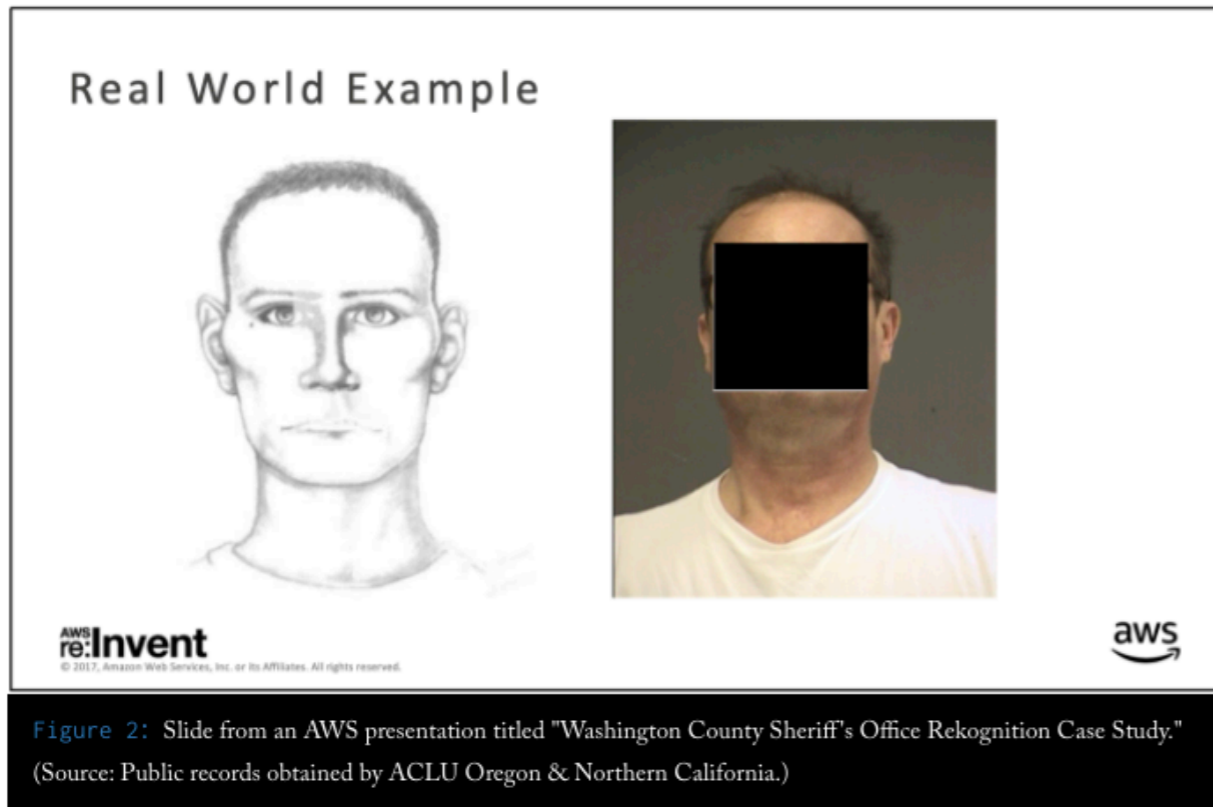
Not used as intended

GARBAGE IN, GARBAGE OUT

FACE RECOGNITION ON FLAWED DATA



Clare Garvie
ClareAngelyn

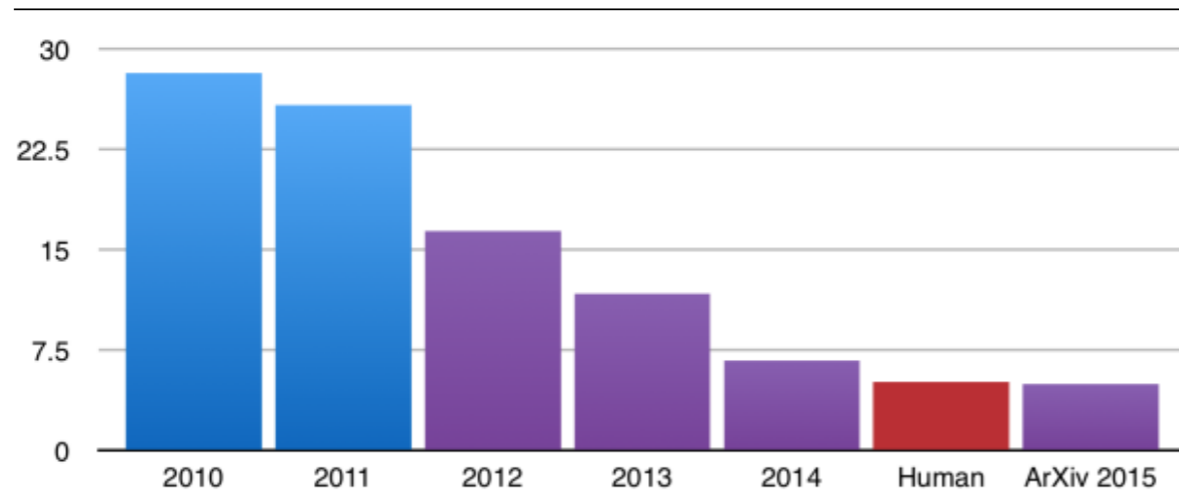


<https://www.flawedfacedata.com/>

Progress in machine learning?

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  REUTERS

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

By Steve Lohr

Feb. 9, 2018

The New York Times



















Limits of abstraction

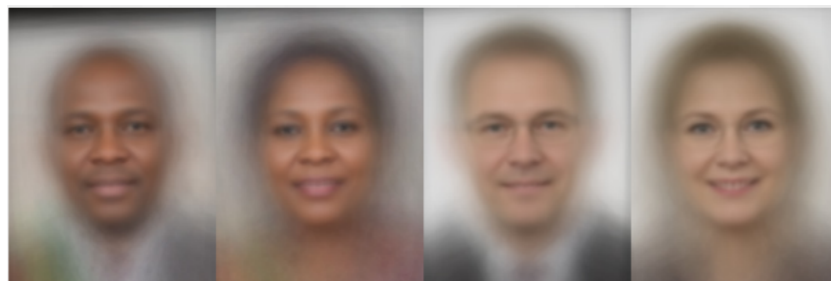
$$\text{minimize}_{\theta \in \Theta} \mathbb{E}_P[\ell(\theta; Z)]$$

- Optimize performance under data-collection system P
- But data collection is always biased
- Structured racism, sexism, underrepresentation

Facial recognition

- Labeled Faces in the Wild, a gold standard dataset for face recognition, is **77.5% male**, and **83.5% White** [Han and Jain '14]
- Commercial gender classification softwares have **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 [Buolamwini and Gebru '18]

Not a new 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

Lack of diversity in data

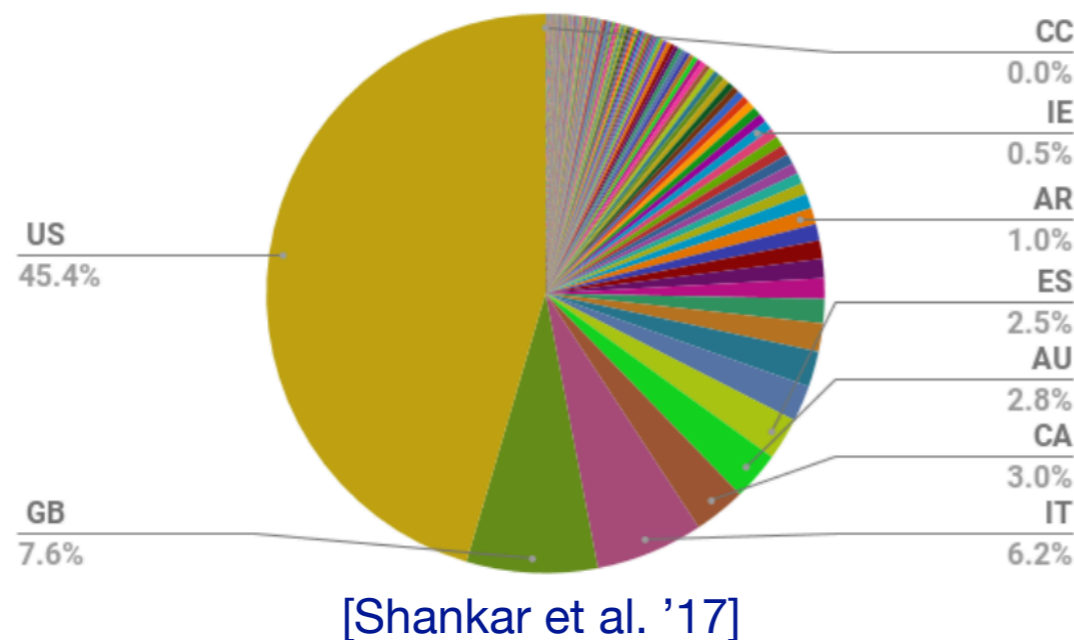
- “Clinical trials for new drugs **skew heavily white**”

- Less than 5% of cancer trial participants were non-white

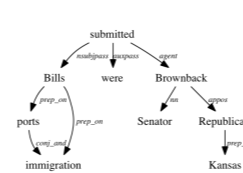
[Oh et al. '15, Burchard et al. '15, Chen et al., '14, SA Editors '18]

- Majority of image data from **US & Western Europe**

ImageNet: country of origin



Other examples



Dependency parsing

[Blodgett+ 16]



Captioning

[Tatman+ 17]



Recommender systems

[Ekstrand+ 17,18]



Face recognition

[Grother+ 11]



Language identification

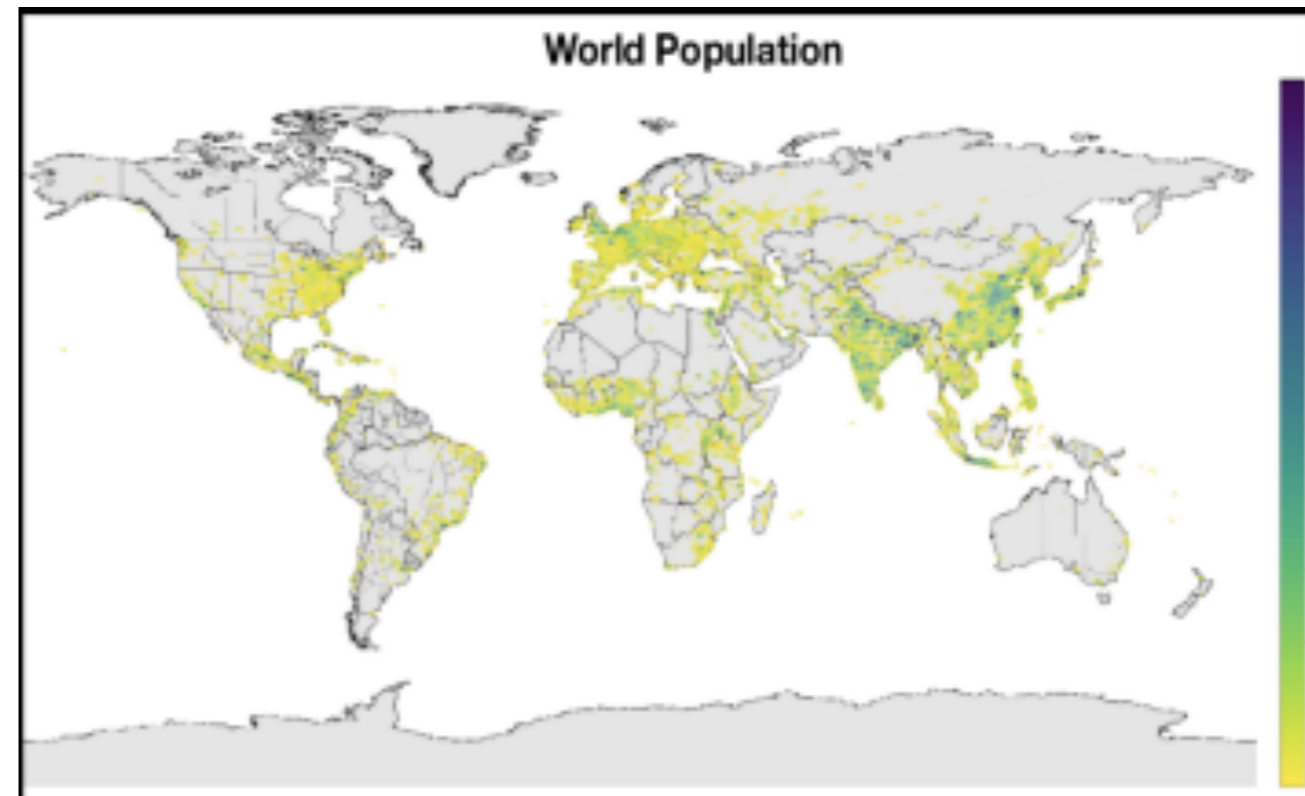
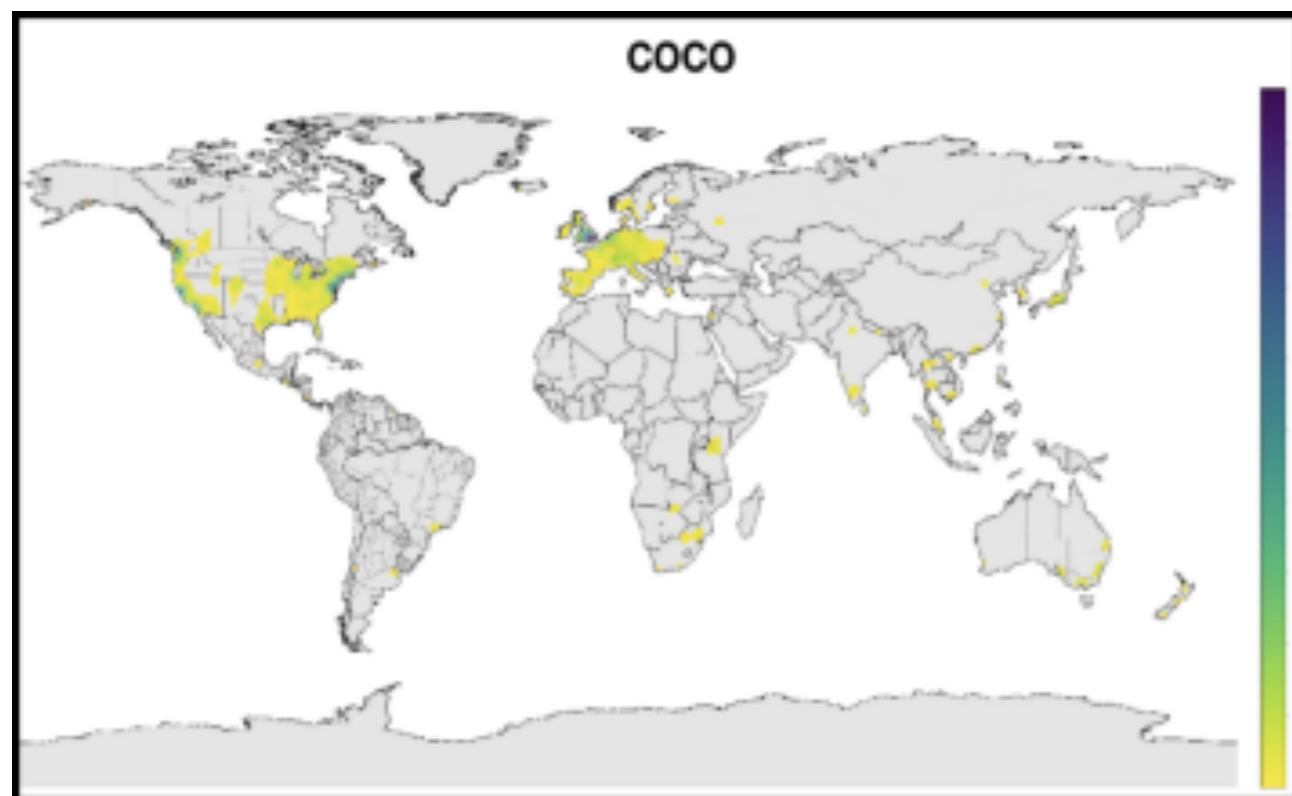
[Blodgett+ 16, Jurgens +17]



Part-of-speech tagging

[Hovy+ 15]

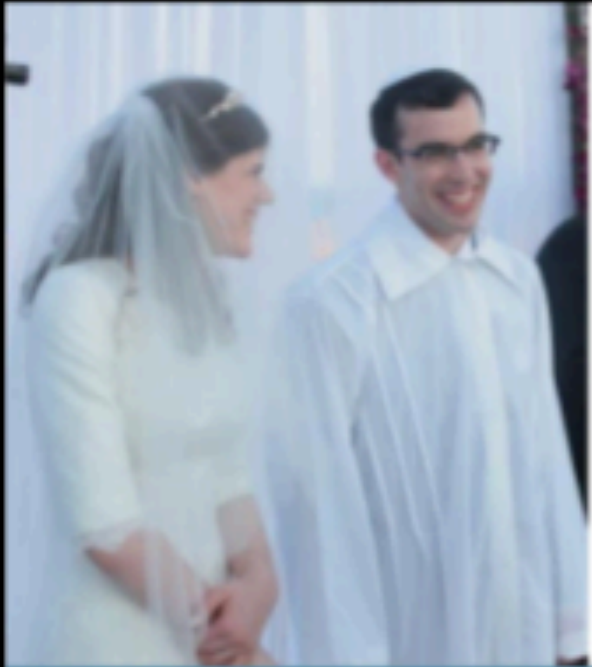
Lack of diversity in data



[DeVries et al. 2019, Does object recognition work for everyone?]



Who is seen? How are they seen?



*ceremony,
wedding, bride,
man, groom,
woman, dress*



*bride,
ceremony,
wedding, dress,
woman*



*ceremony,
bride, wedding,
man, groom,
woman, dress*



person, people

[Shankar et al. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World]



Slide from Timnit Gebru & Emily Denton's CVPR2020 tutorial

Who is seen? How are they seen?

Training data: 33% of cooking images have man in the agent role
Model predictions: 16% cooking images have man in the agent role

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

[Zhao et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints]
 [Hendricks et al. Women also snowboard: Overcoming bias in captioning models.]



Gender bias in 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:

Turkish - detected 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

she does not like her
she loves him

onu görüyor
onu göremiyor

she sees it
he can not see him

o onu kucaklıyor
o onu kucaklamıyor

she is embracing her
he does not embrace it

o evli
o bekar

she is married
he is single

o mutlu
o mutsuz

he's happy
she is unhappy

o çalışkan
o tembel

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

Racial bias in 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.



Datasets and humans

“Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings.”

Mimi Onuoha

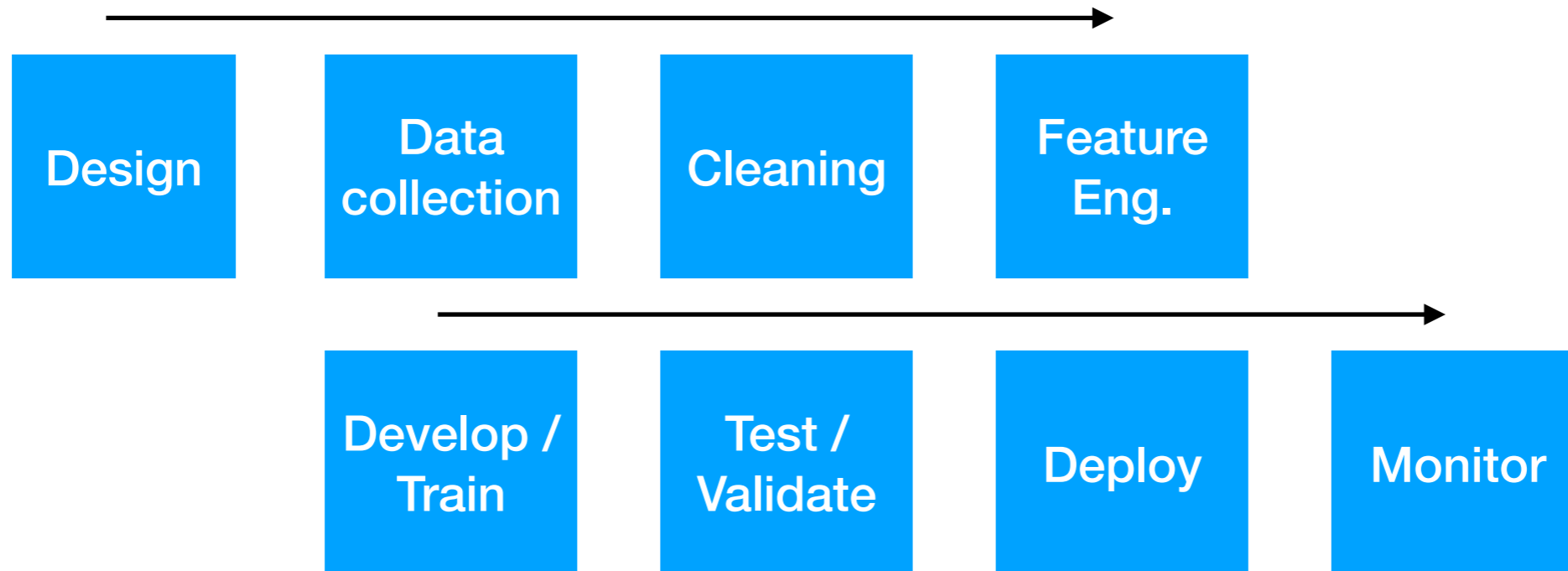
<https://datasociety.net/people/onuoha-mimi/>

- We always need to remember the humans in our mathematical abstractions

Seeta Peña Gangadharan: https://trustworthyiclr20.github.io/Gangadharan_ICLR_trustworthy_ML_slides.pdf

- Our research (even theoretical research) is always conducted under social conditions

ML Pipeline



Every aspect of the pipelines gets affected by structural inequities of the world we live in

Data as infrastructure

Denton et al. (2020+)

- Data provides the foundation on which we do knowledge work, just like electricity, sewage, roads etc
- Once established, difficult to go beyond it (more on this next week)

Contingent → Datasets are contingent on the social conditions of creation

Constructed → Data is not objective; 'Ground truth' isn't truth

Value-laden → Datasets are shaped by patterns of inclusion and exclusion

Slide from Timnit Gebru & Emily Denton's CVPR 2020 tutorial on FATE

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 wants to avoid that pitfall — so much so it paid to have hired temps go out to collect face scans from a variety of people on the street using \$5 gift cards as incentive....teams were dispatched to target homeless people in Atlanta, unsuspecting students on college campuses around the U.S.”


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”

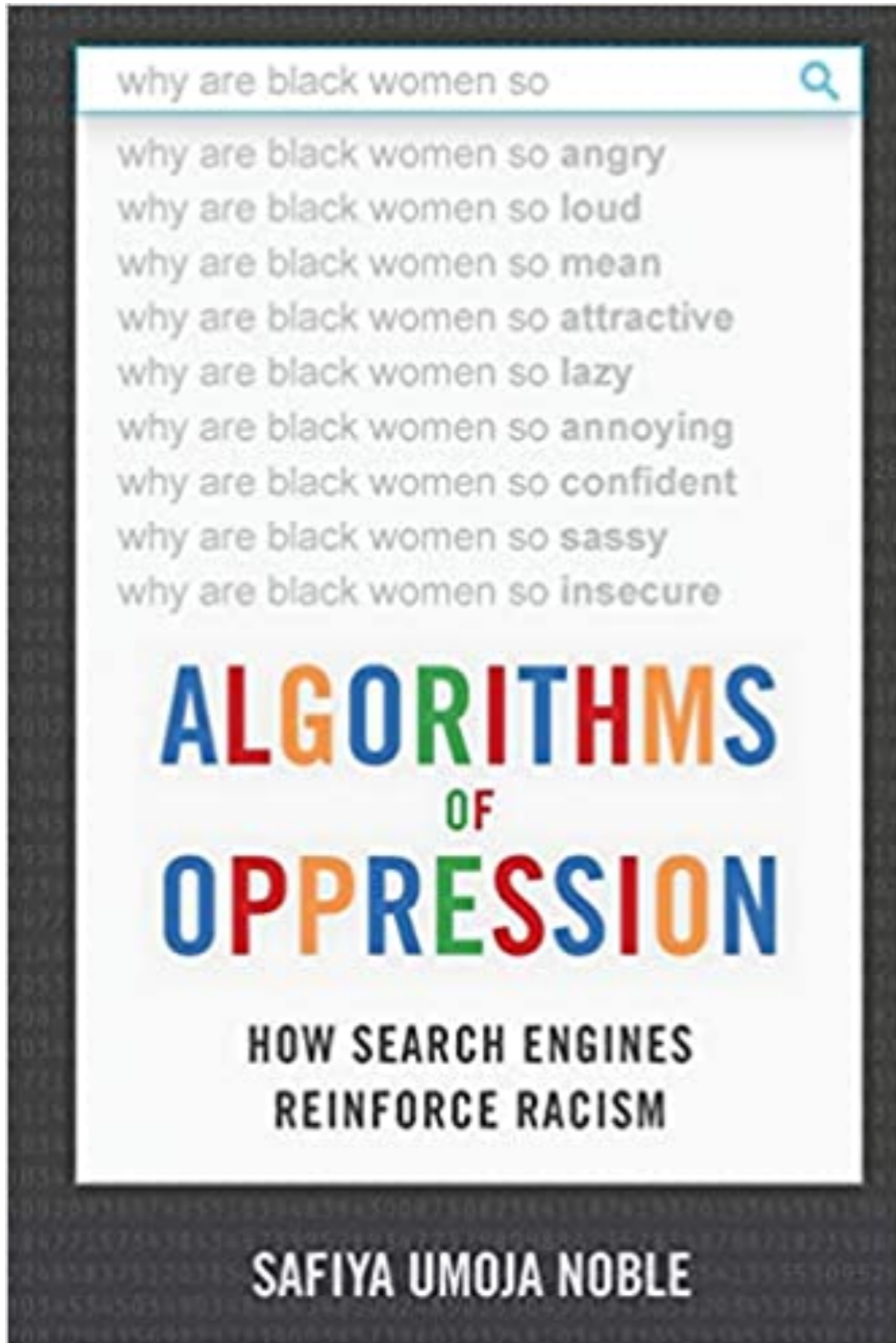
“If the person were to look at that screen after the task had been completed, and say, ‘Oh, was it taking a video?’... we were instructed to say, ‘Oh it’s not really,’”

Fixes are hard

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

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

- A “fair” gender classifier is not the solution
- Prescribing gender without consent is inherently oppressive, especially for trans people
- Gender is a social construct, differing across space and time
- E.g. Maybe a gender classifier to be used for predicting conversion-rates in an ads auction should not exist



Deb Raji
@rajiinio



I often wonder about this as well.

Google responded to Black people being tagged "gorillas" by blocking the "gorilla" tag altogether. And I notice that although popularized examples of "doctor" & "CEO" now show diverse Image search results, something like "foot" or "hair" won't

Alison Gerber @alisonkgerber

is there good writing on the tech side of Big Search's responses to scholars'/activists' demands? can look more like whack-a-mole than anything else - e.g., after @safiyanoble's work, "black girls" results appear to be carefully tidied up, but "asian girls" is still a horrorshow



asian girls



All Images Videos News Shopping More Settings Tools

About 4 470 000 000 results (0,69 seconds)

Ad · www.filipinocupid.com/

Want an Asian Girlfriend? - Browse 1000s of Asian Women

Discover beautiful **Asian** women seeking dating and relationships today. Join now! Anti-Fraud Screening. Over 3 Million Filipinas. Mobile Friendly. Review Matches for Free. Featured on 90 Day Fiancé. Join in 30 Seconds. As Seen on TV. #1 Filipino Dating Site.

Success Stories

Read Success Stories From Happy Customers Who've Found Love.

Meet Filipino Women

Browse Stunning Filipino Women Aged Between 18 and 30

Ad · www.internationalcupid.com/

Browse Single Women From Asia - Meet Beautiful African Singles

4 Million+ Single Women on the #1 International Dating Site. Advanced Matching. Start Now. Meet Foreign Singles on the #1 International Dating Site. 4 Million+ Singles. Join Now. Join in 30 Seconds. Date Safely from Home. Backed by Cupid Media. Meet Singles from Home. Real Foreign Ladies · Read Success Stories · International Women 18-24

www.pinterest.com > flowbacg > sexy-asian-girls

Sexy asian girls - Pinterest

Oct 18, 2020 - Explore Flowbacg's board "Sexy asian girls" on Pinterest. See more ideas about Sexy asian girls, Sexy asian, Asian girl.

www.pinterest.com > irvadelman > beautiful-asian-girls

Beautiful asian girls - Pinterest

Sep 17, 2020 - Explore Irwin Adelman's board "Beautiful asian girls" on Pinterest. See more ideas about Beautiful asian girls, Asian girl, Asian beauty.

thehive.com > category > sexy-girls > asian

Hot Asian Girls | Hot Korean & Chinese Girls - theCHIVE



Fixes are hard

- How do we define a fix?
 - We can infer race without asking an explicit question
- Consent, privacy etc
- Although our technical language is very limited in scope, vast majority of fairness papers
 - Posit a notion of fairness in an abstract learning setting
 - Formulate a constrained loss minimization problem over (approximate) fairness, or relaxations thereof
 - Propose solution approaches

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 an engineer
 - I do this out of technical interest
- Our research communities, and the corporations it supports need dramatic shifts towards better representation

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

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

Mitchell et al. (2019)

<https://arxiv.org/pdf/1810.03993.pdf>

- Documents for “trained ML models that provide benchmarked evaluation in a variety of conditions”
 - “different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, skin type)”
 - “intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains.”
- Factors for performance (e.g. instrumentation)
- Usage contexts, evaluation & testing details

<https://modelcards.withgoogle.com/model-reports>

Industry efforts

- IBM Open Scale <https://www.ibm.com/cloud/watson-openscale>
 - Monitoring and (re)evaluation of deployed models
- IBM AI Fairness 360 <https://aif360.mybluemix.net/>
 - API for examining and mitigating potential discrimination
- Microsoft Fairlearn <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>
 - API (wrapper) for imposing fairness constraints on a learner, as well as visualization tools
- Google responsible AI practices <https://ai.google/responsibilities/responsible-ai-practices/>
 - Guidelines and recommendations on best practices

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=jlXluYdnyyk&ab_channel=ArvindNarayanan

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

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

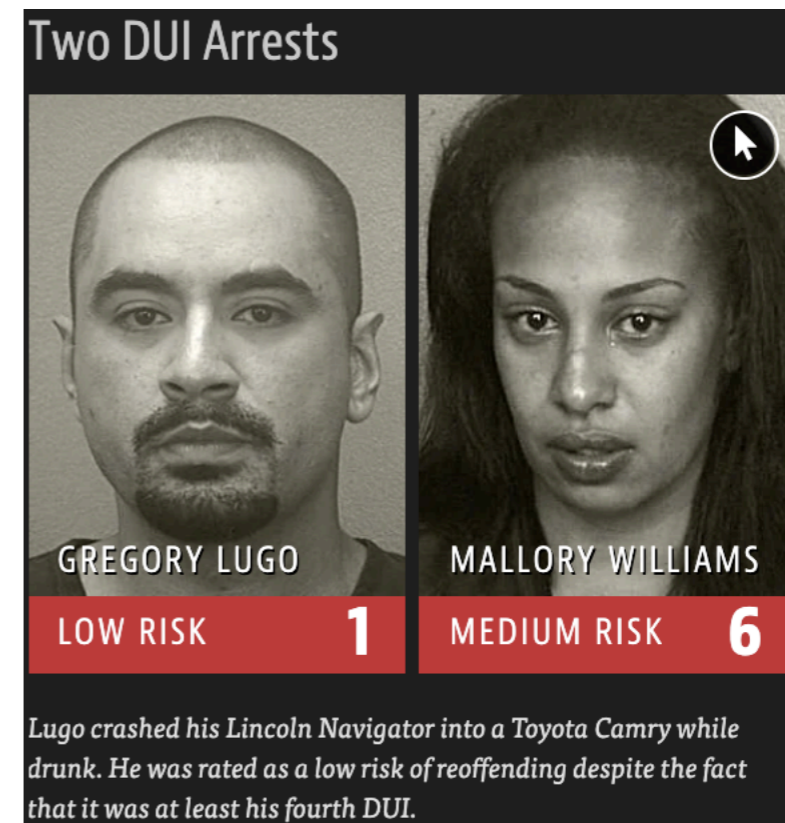
There's software used across the country to predict future criminals. And it's biased against blacks.

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

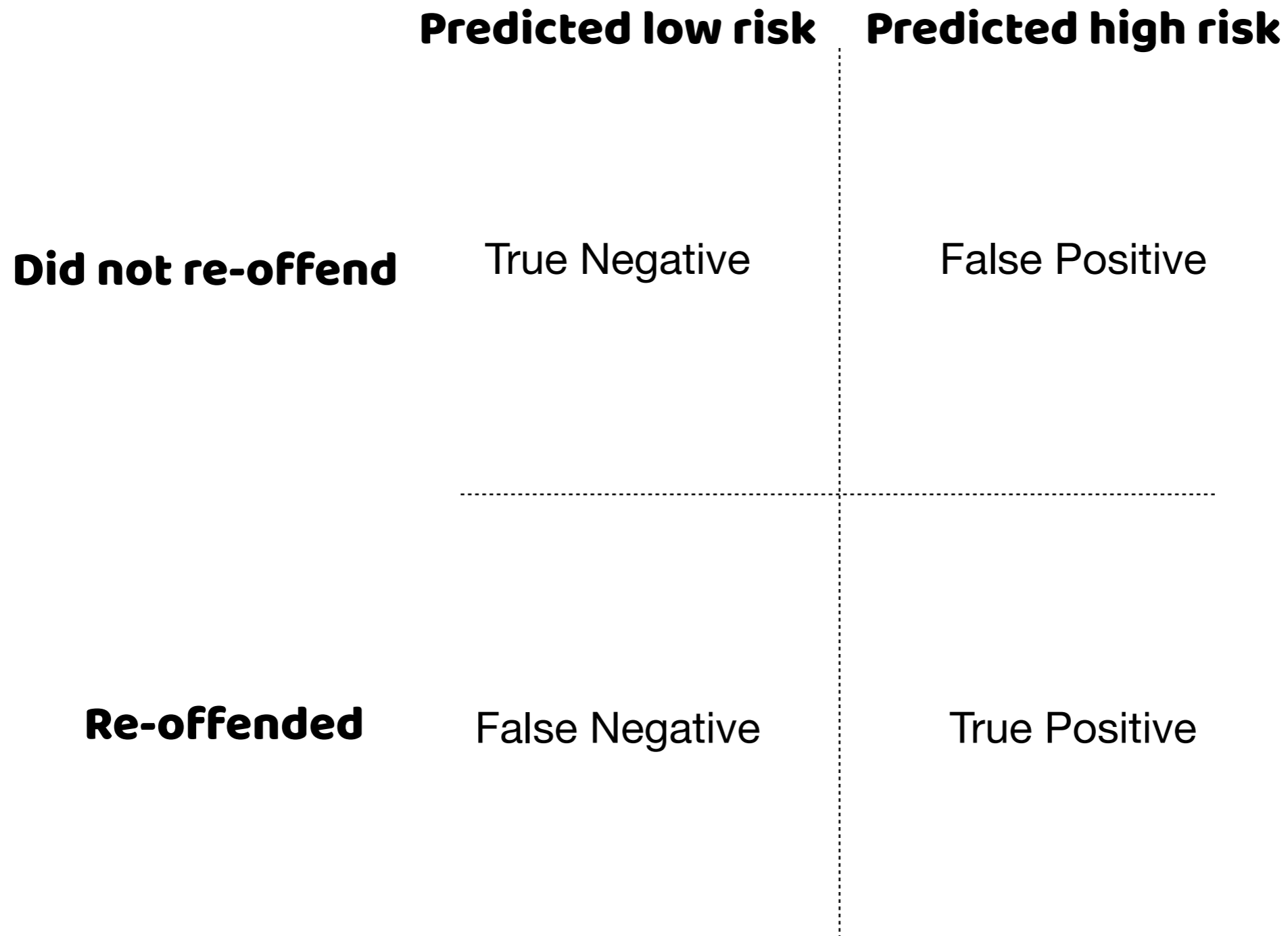


Today: **stylized** perspective as a guide to various fairness definitions

Fairness

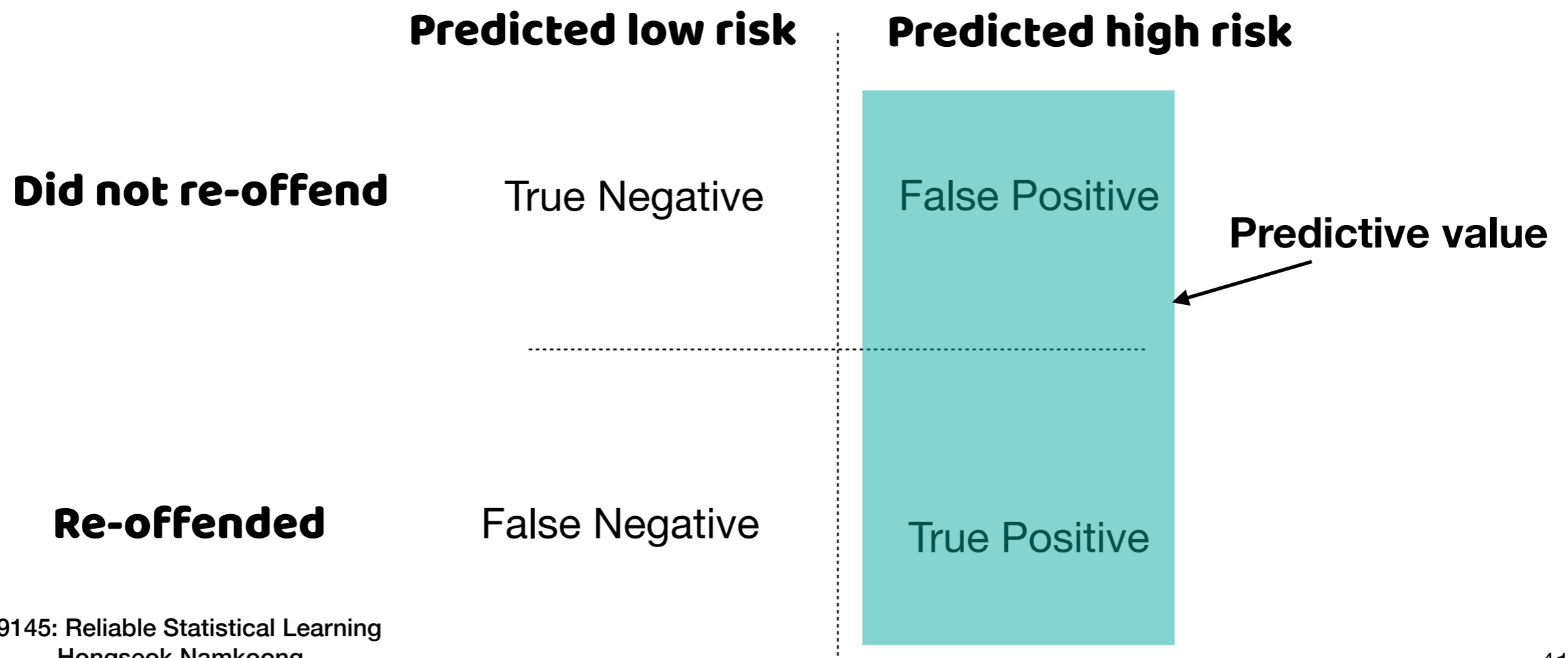
- Surprisingly common position among engineers: my model describes my data well, so my algo is faultless
- Make algorithmic systems support human values
 - Statistical bias is not enough
- Which values should it support?
- We consider a simple binary classification problem with pre-defined groups

Simple setup



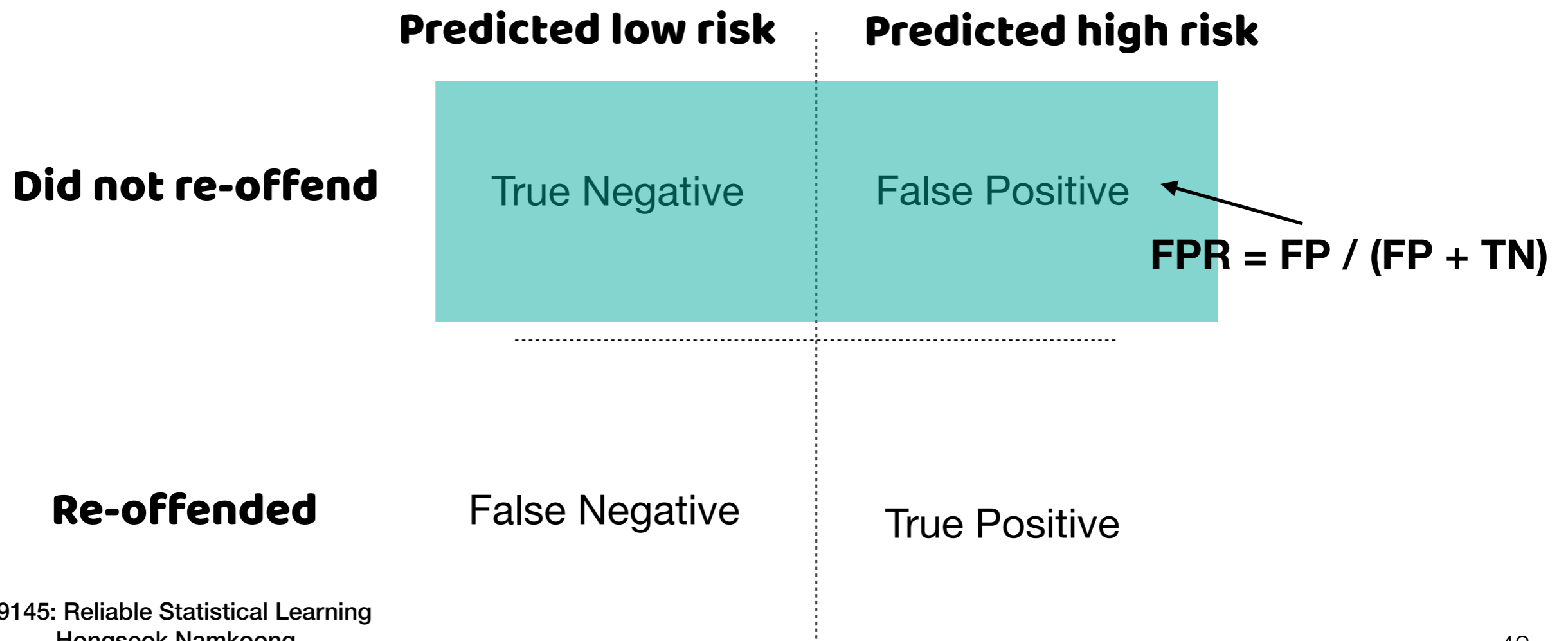
Perspectives matter

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



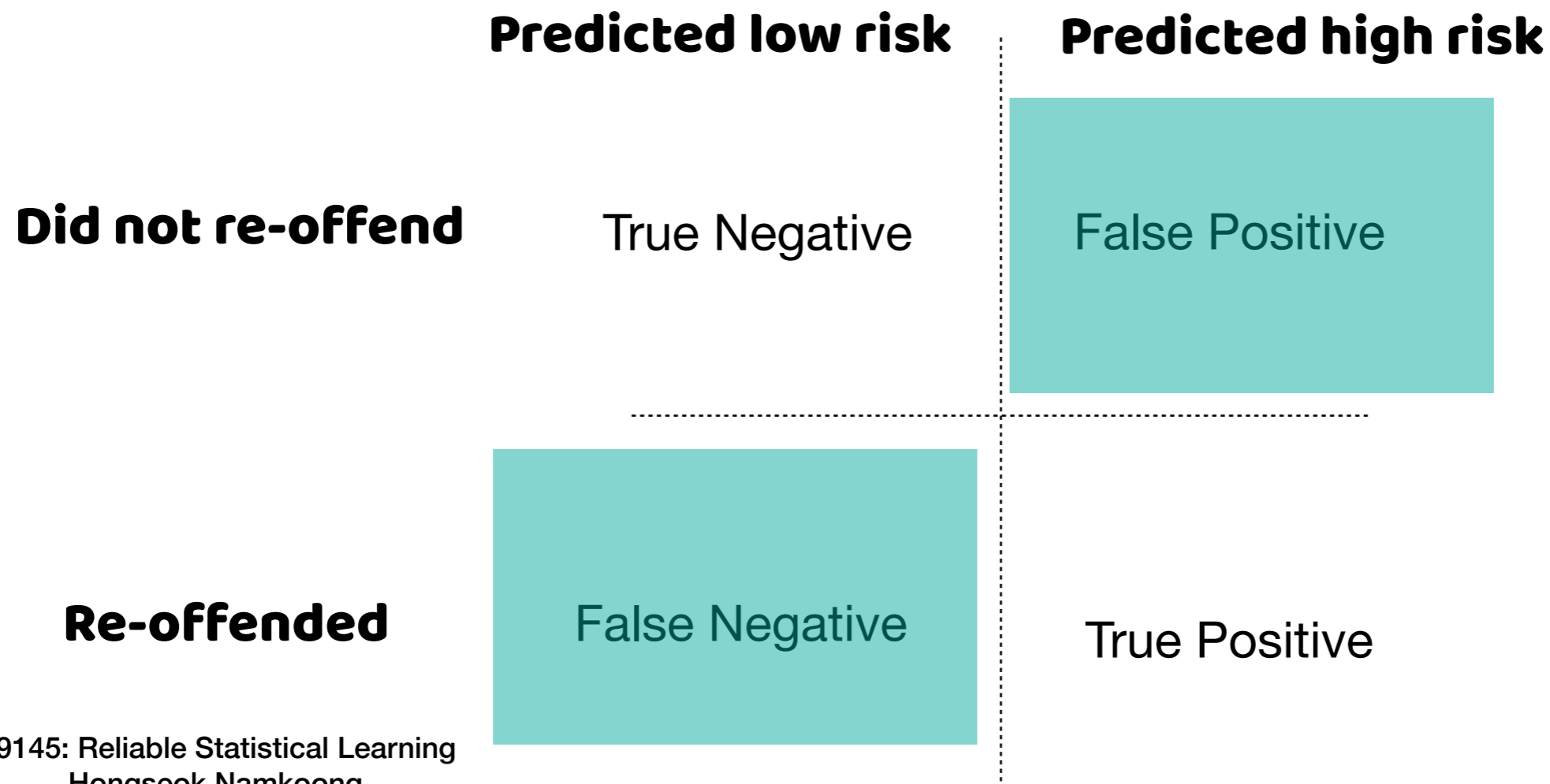
Perspectives matter

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



Perspectives matter

- Let's take college admission or hiring
- **Society:** how do we maximally benefit from diversity by ensuring selected set is demographically balanced.

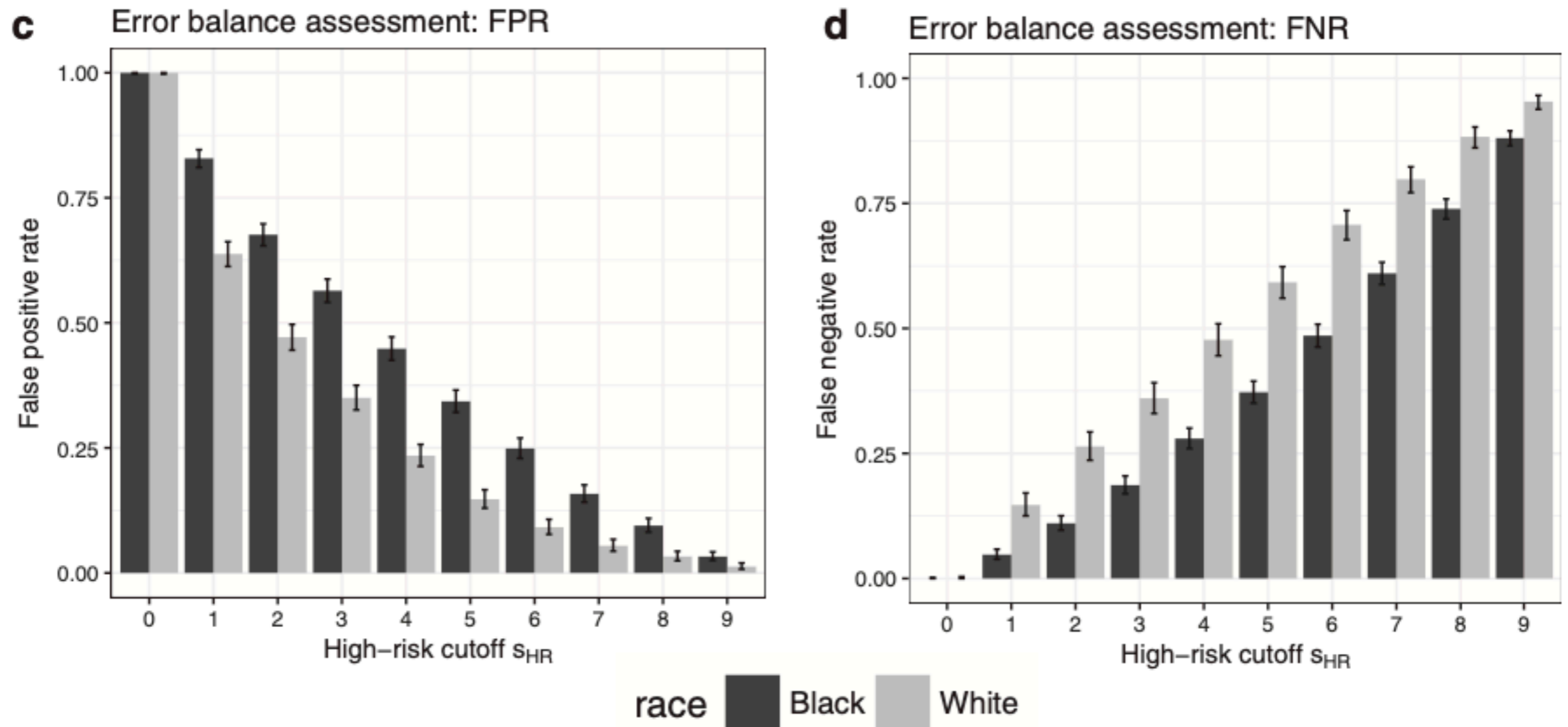


Fairness definitions

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

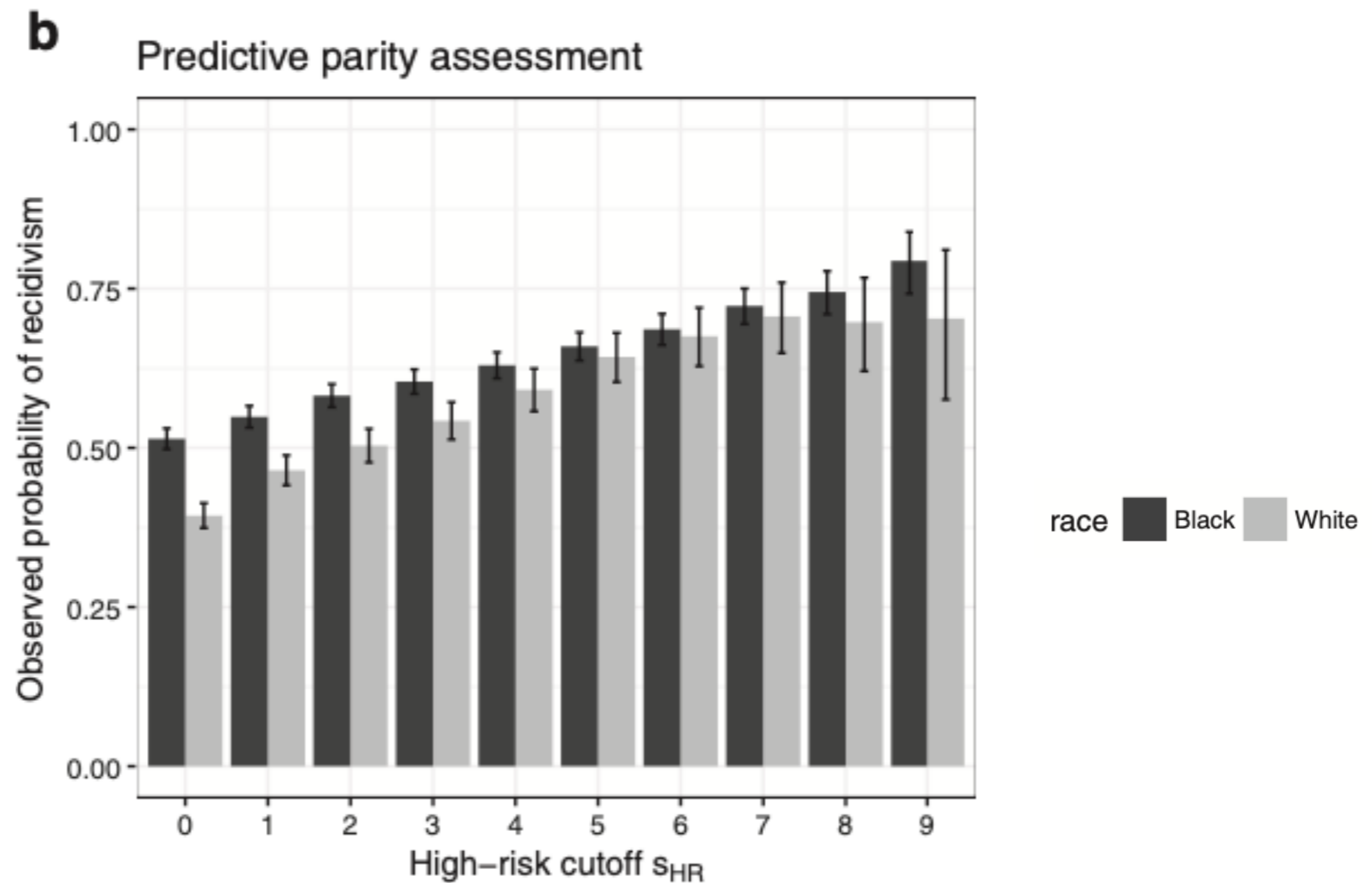
COMPAS

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



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, then it cannot jointly balance FPR and FNR

Proof

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

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

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

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

Proof

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 not enough

Managing trade-offs?

- Chouldechova recommends dispensing predictive parity, and equalizing FPR / FNR
 - Makes sense in COMPAS context
 - See also Hardt et al (2016) <https://arxiv.org/pdf/1610.02413.pdf>
- But even this may not be possible if you consider a single threshold rule
- Also trade-offs between fairness and utility (e.g. safety, # defendants released)

More fairness definitions

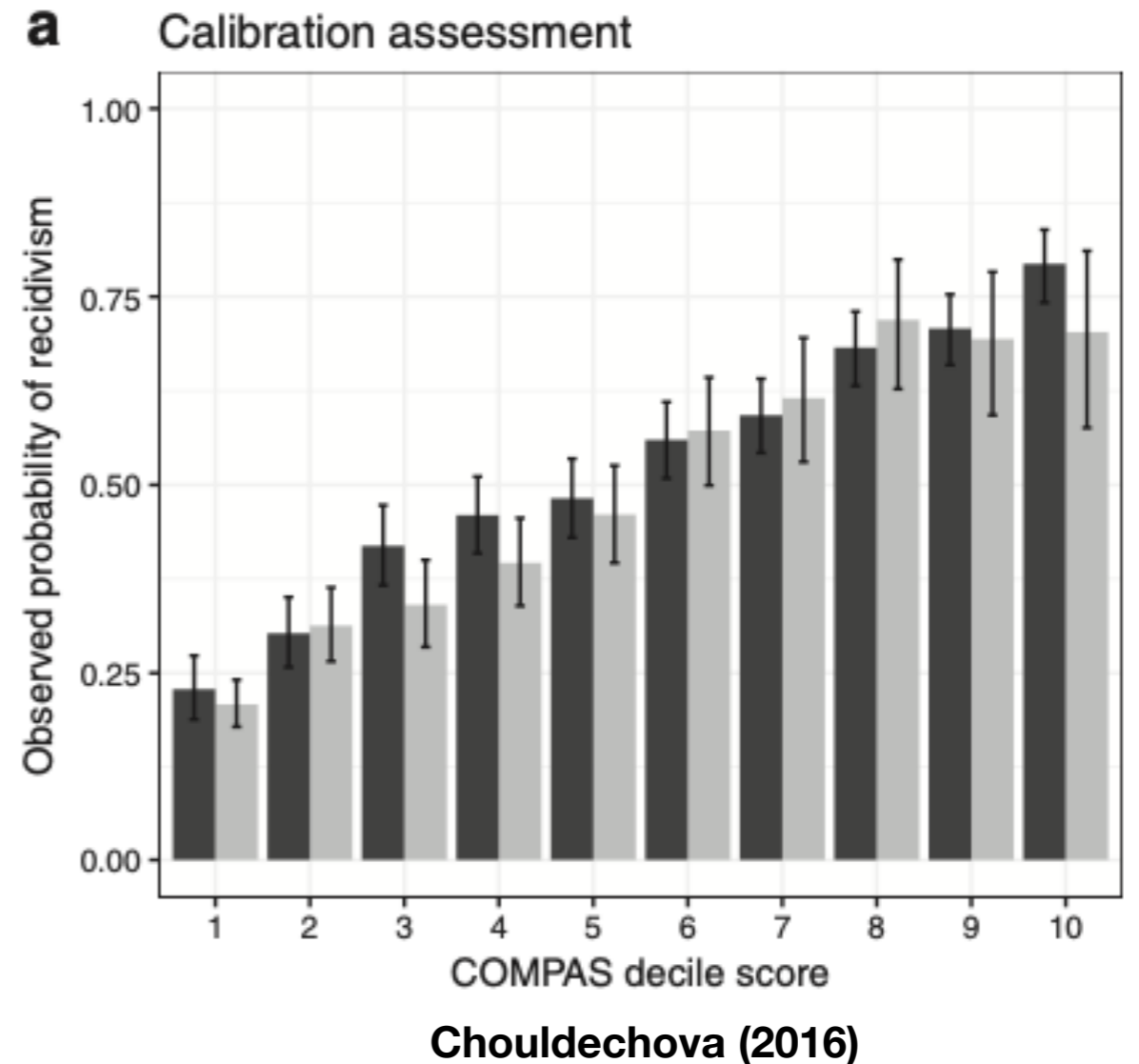
		True condition			
Total population		Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$
Predicted condition	Predicted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

Wikipedia: Evaluation of binary classifiers

Lots and lots of potential impossibility results

Calibration

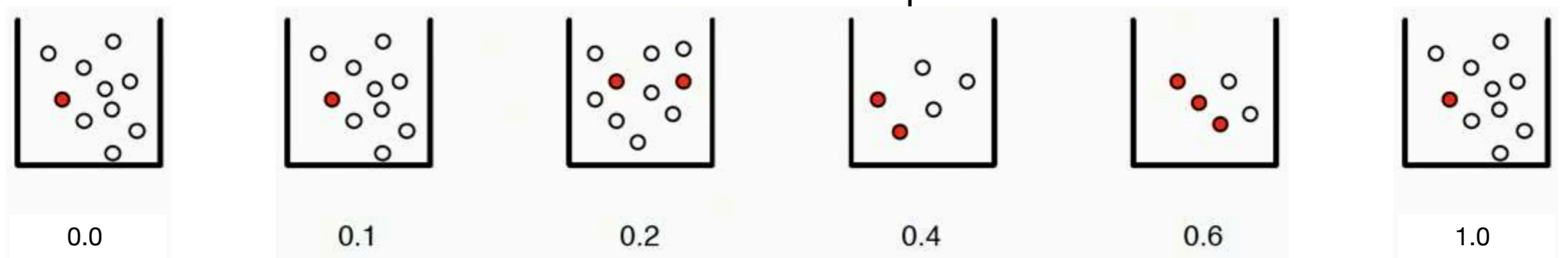
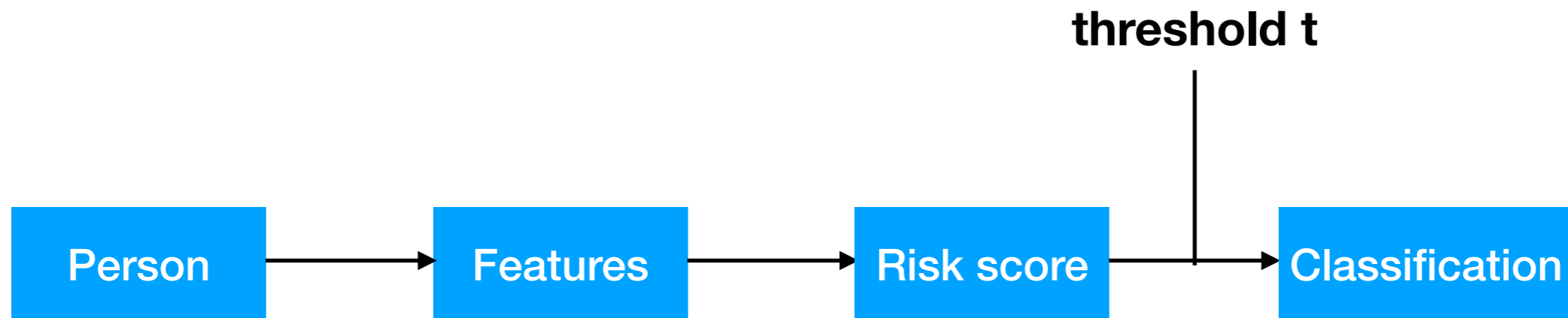
- Northpointe: COMPAS scores were well-calibrated within each group
- For all Black/white defendants with score s , (approximately) s fraction of them actually re-offends
 - It's nice that outputs actually mean what they claim
- But this is meaningless to Black defendants who won't re-offend but still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



Picture from Kleinberg (2018) slides

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, $k_g =$ 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

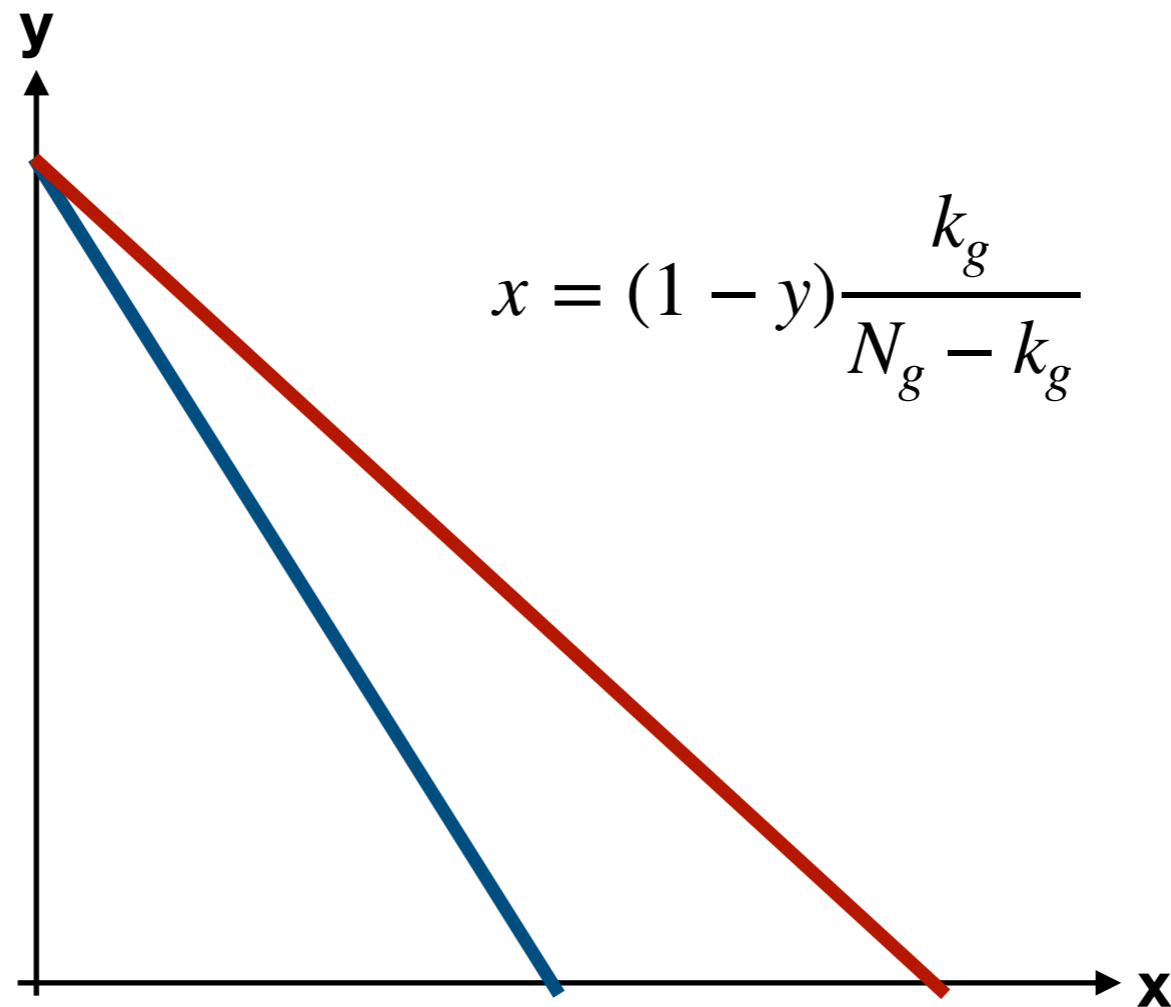
$$k_g = \text{total score in group } g = (N_g - k_g)x + k_g y$$

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?