(Un)fairness in Machine Learning

A (un)fairly short overview

Christopher Akiki, Alina Mailach, Fabian Dannegger

Overview

1. Context: Pervasiveness of Machine Learning
2. Fairness: Definitions and Considerations.
3. Unfairness: Origin and Quantification.
4. Mitigation: Where to intervene?
5. Practical Example: Deciding who gets a vaccine

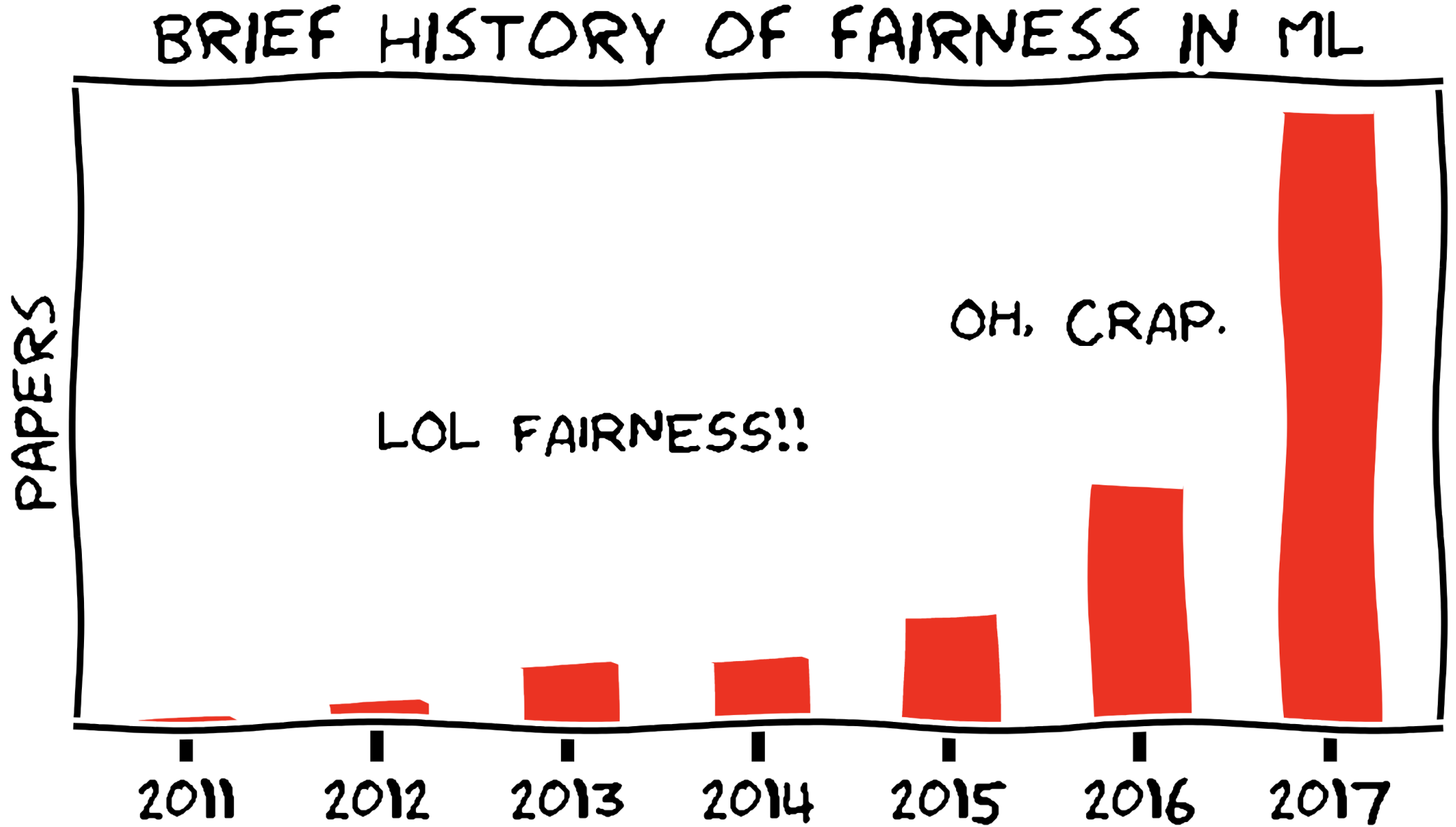


Figure by Moritz Hardt

Unfairness in ML: Hey Siri, find me the nearest criminal

"After all, as the former CPD [Chicago Police Department] computer experts point out, the algorithms in themselves are neutral. 'This program had absolutely nothing to do with race... but multi-variable equations,' argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound." — **Dr. Gillian Tett**

https://mathbabe.org/2014/08/25/gilian-tett-gets-it-very-wrong-on-racial-profiling/

Fairness in ML: A Story in Two ~~Acts~~ Tweets

Timnit Gebru
@timnitGebru

I'm sick of this framing. Tired of it. Many people have tried to explain, many scholars. Listen to us. You can't just reduce harms caused by ML to dataset bias.

> **Yann LeCun** @ylecun · Jun 21
>
> ML systems are biased when data is biased.
> This face upsampling system makes everyone look white because the network was pretrained on FlickFaceHQ, which mainly contains white people pics.
> Train the *exact* same system on a dataset from Senegal, and everyone will look African. twitter.com/bradpwyble/sta...

11:00 PM · Jun 21, 2020 · Twitter Web App

(Un)fairness: A few (very) high-level definitions

- **Context:**

  A classification task where the data points are people

- **Protected attribute:**

  "An attribute that partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion. Protected attributes are not universal, but are application specific."

- **Bias:**

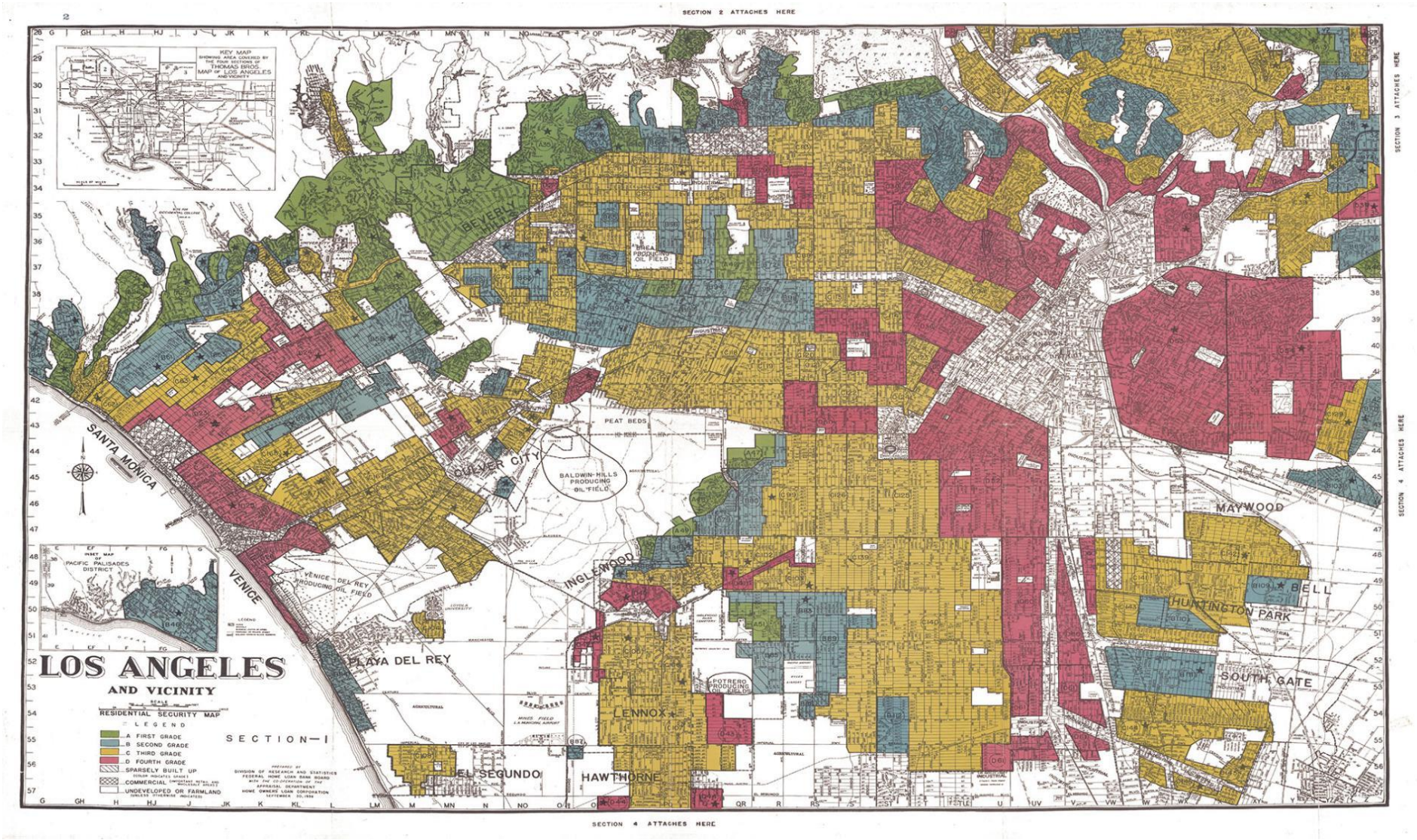
  "A systematic error. In the context of fairness, we are concerned with unwanted bias that places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage."

- **Problem:**

  Formally defining unwanted bias in an actionable way is extremely contextual, contentious, and sometimes even conflicting.

https://aif360.mybluemix.net/resources#glossary

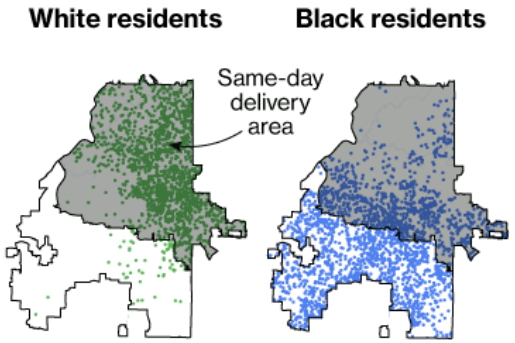
1939: Just remove the sensitive attributes?



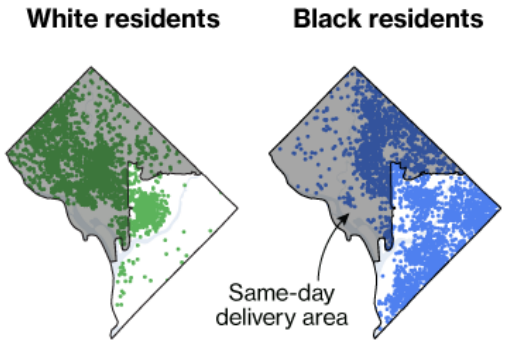
https://www.nationalgeographic.com/news/2016/10/housing-discrimination-redlining-maps/

2016: Just remove the sensitive attributes?

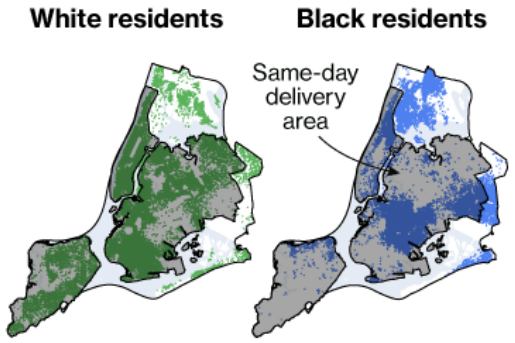
The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.



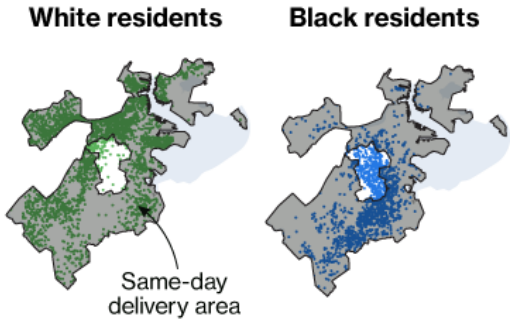
Same-day service is unavailable in southeast Washington, D.C. including neighborhoods located blocks from the Capitol building and all areas across the Anacostia River.



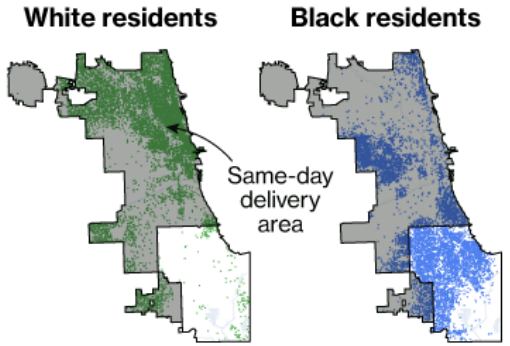
More than 6.5 million New York City residents live in ZIP codes with same-day delivery. But the Bronx is the only city borough completely excluded from the service.



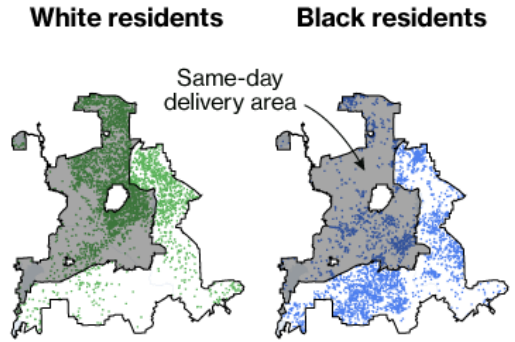
Three ZIP codes in the center of Boston, including the Roxbury neighborhood, are excluded from same-day coverage.



About half of Chicago's black residents live in the southern half of the city where they do not have access to Amazon's same-day delivery service.



Dallas has the lowest overall coverage rate among cities with same-day delivery. White residents are more than twice as likely as black residents to have access to the service.



https://www.bloomberg.com/graphics/2016-amazon-same-day/

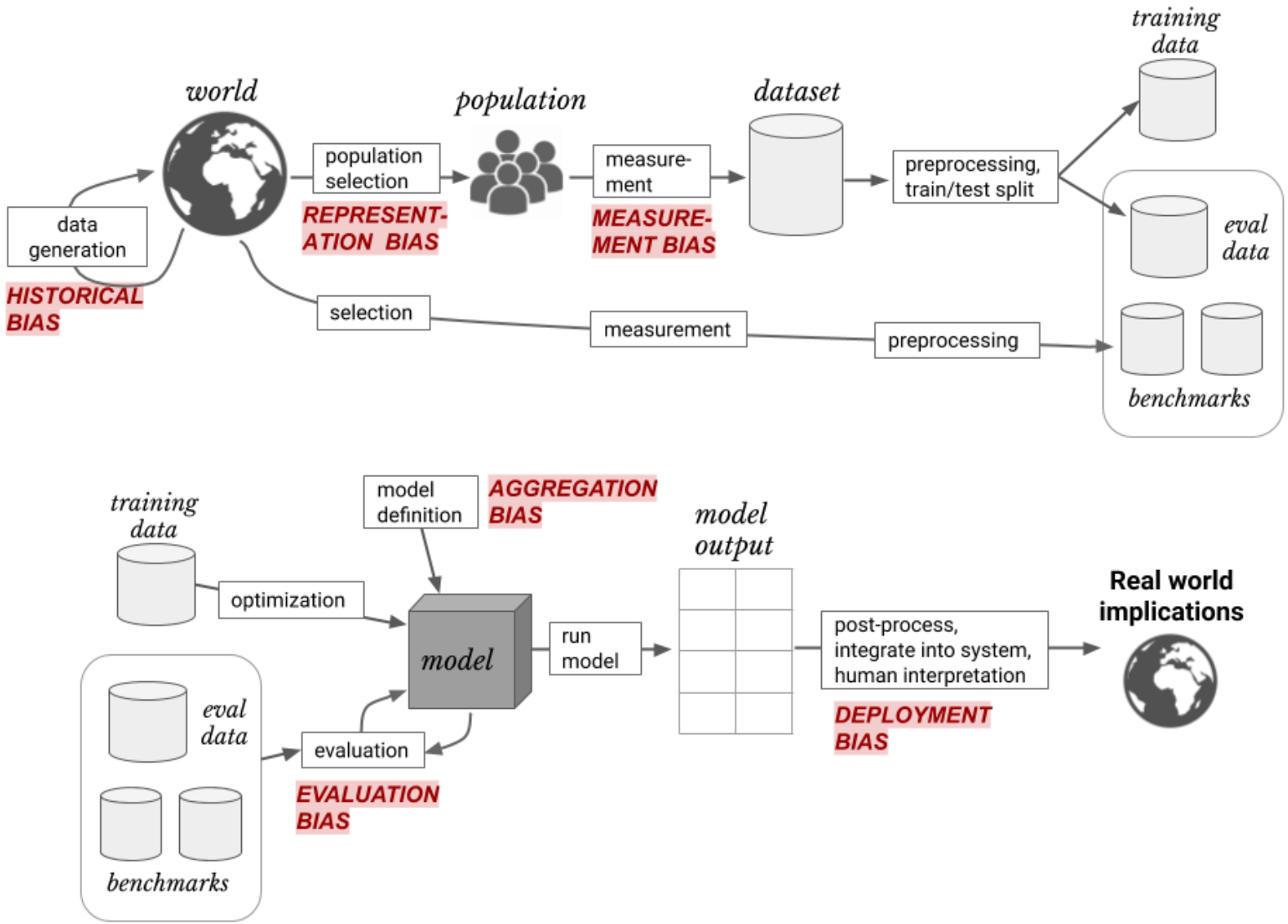
2016: A big year for ML Fairness

Context: COMPAS algorithm, Predicting future recidivism in practice

	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%

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

Bias: Where does it come from?



Suresh, Harini, and John V. Guttag. "A framework for understanding unintended consequences of machine learning." arXiv preprint arXiv:1901.10002 (2019).

Many ways to rate a binary classifier

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

https://en.wikipedia.org/wiki/Confusion_matrix

Many ways to rate a binary classifier

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

https://en.wikipedia.org/wiki/Confusion_matrix

Many ways to rate a binary classifier

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Problem: No imperfect classifier can guarantee all three unless the base rates are equal

(Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." Big Data 5.2 (2017): 153–163. Crossref. Web.)

https://en.wikipedia.org/wiki/Confusion_matrix

Really many ways to rate a binary classifier

		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) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		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−) = $\frac{\text{FNR}}{\text{TNR}}$		

https://en.wikipedia.org/wiki/Confusion_matrix

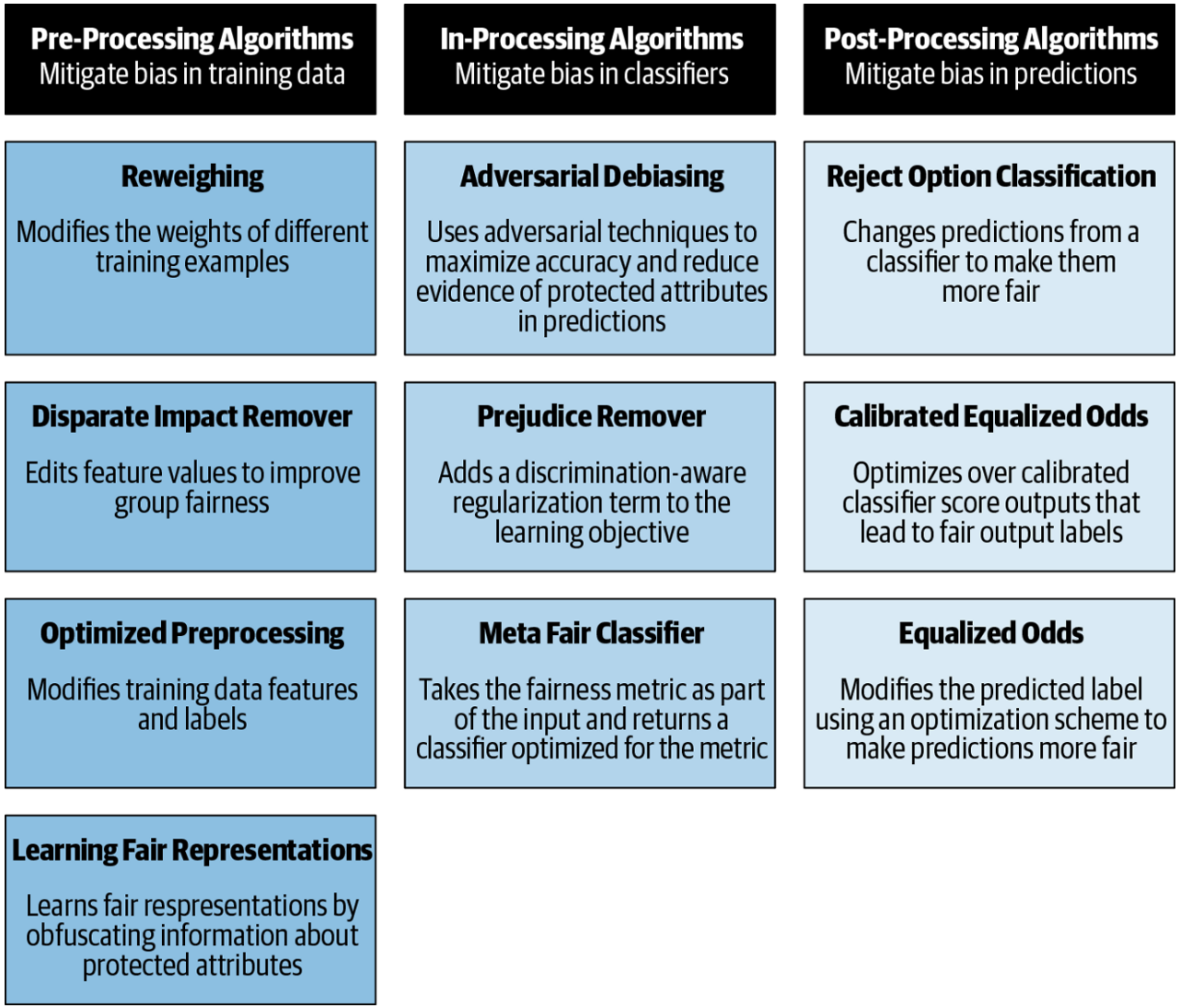
Really many metrics of fairness

List of demographic fairness criteria

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)

https://fairmlbook.org/classification.html

Mitigating unfairness: How to intervene?



. AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning. Trisha Mahoney, Kush R. Varshney, and Michael Hind. O'Reilly Media, Sebastopol, CA, USA, 2020.

Practical example:

- Use the IBM fairness toolkit on data from the Medical Expenditure Panel Survey (MEPS) to try and distribute fictional COVID-19 vaccines.
- https://nbviewer.jupyter.org/github/cakiki/ml-fairness-validity/blob/master/fairness.ipynb