Designing Against Bias in Machine Learning and AI

David J Corliss, PhD is managing director at Grafham Analytics, a data science consulting company. His work in best practices for ethical machine learning and AI includes chairing the 2022 Conference of Statistical Practice from the American Statistical Association (ASA), writing a column on Data for Good in the ASA's monthly member magazine, and serves on the steering committee of the Statistics section of the American Association for the Advancement of Science. Dr. Corliss is the founder of Peace-Work, a volunteer cooperative of statisticians, data scientists and other researchers applying analytics in issue-driven advocacy.

PEACE-WORK

Designing Against Bias: Identifying and Mitigating Bias in Machine Learning and AI

David J Corliss, Peace-Work

Modern Modeling Methods
University of Connecticut
June 27-28, 2023

PEACE-WORK

OUTLINE

Overview of Bias in ML and AI

Root Causes of Bias

Measuring Bias

Designing Against Bias

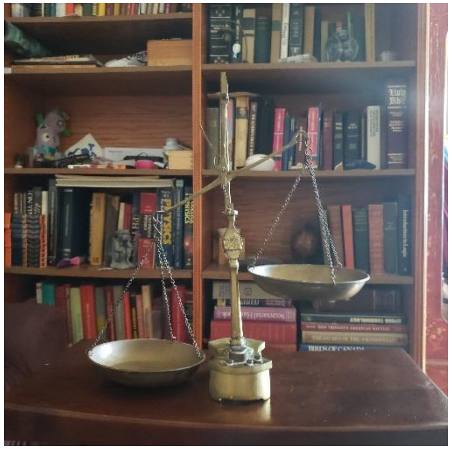
Conclusions

PEACE-WORK

OVERVIEW OF BIAS IN ML AND AI

Bias in Machine Learning Algorithms

Taking human decisions out of the process was supposed to make things more fair…



…but often it hasn’t

=> What went wrong??

PEACE-WORK

Racial Bias: Bail and Parole Algorithms

The “Solution”: ML says who gets bail or parole

COMPAS Algorithm:

RISK = AGE * Weight 1
+ AGE AT FIRST ARREST * Weight 2
+ HISTORY OF VIOLENCE * Weight 3
+ EDUCATION LEVEL * Weight 4
+ HISTORY OF NONCOMPLIANCE * Weight 5

The Problem: using the algorithm results in the exact same bias

PEACE-WORK

Gender Bias: Amazon Resume Screening

The "Solution": ML picks top resumes

Amazon Algorithm:

Resume Quality = ? + ? + ? + ? + ? ...

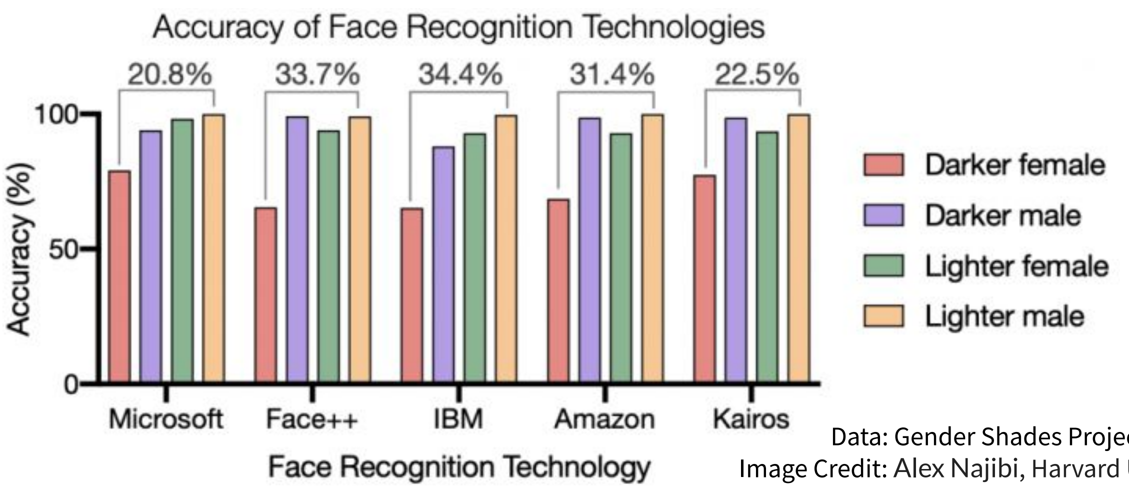
Image Credit: flazingo_photos -

The Problem: the algorithm is biased against women applicants

PEACE-WORK

ROOT CAUSES OF BIAS

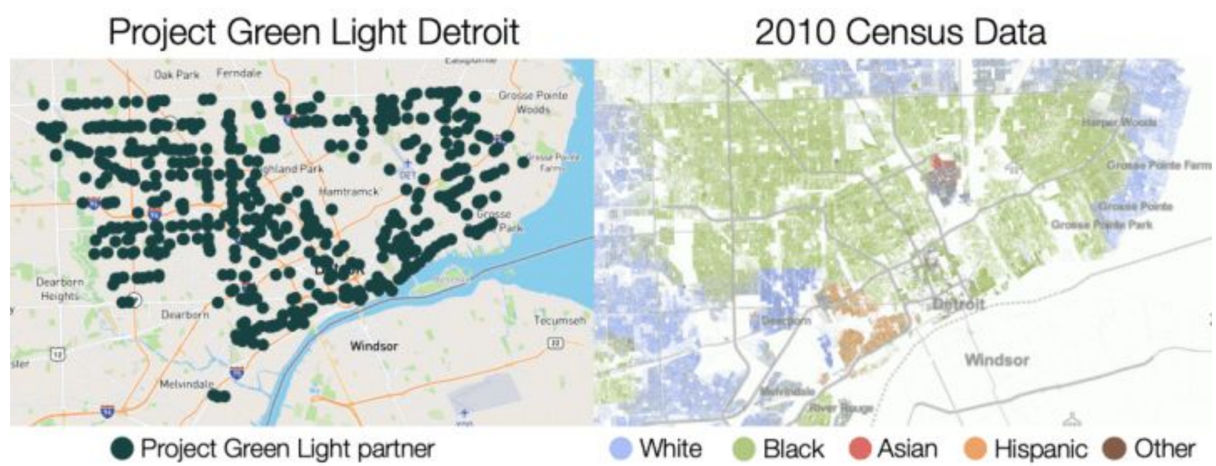
Root Causes of Bias: Selection Bias



Data: Gender Shades Project
Image Credit: Alex Najibi, Harvard University

Algorithm trained using biased subset

Usage results in disparate impact



=> Biased Training Population = Biased Results

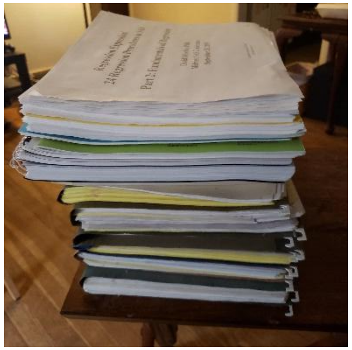
Root Causes of Bias: The History Problem

ML replaces human decision making

Image Credit: David Davies -CC BY-SA 2.0

```
library(tensorflow)
library(keras)
model <- keras_model_sequential() %>%
  layer_conv_2d(filters = 32,
  kernel_size = c(3,3), activation = "relu",
```

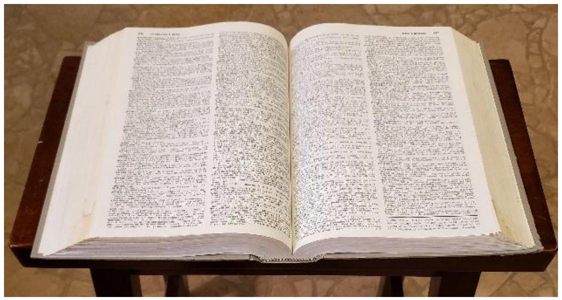
The algorithm is trained using earlier, biased human decisions



=> Bias In = Bias Out

PEACE-WORK

Root Causes of Bias: Spaghetti Problem



DATA DICTIONARY

Hundreds or even thousands of potential predictors

Algorithm trained uncritically using “anything that sticks”

Image Credit: snackdinner.com

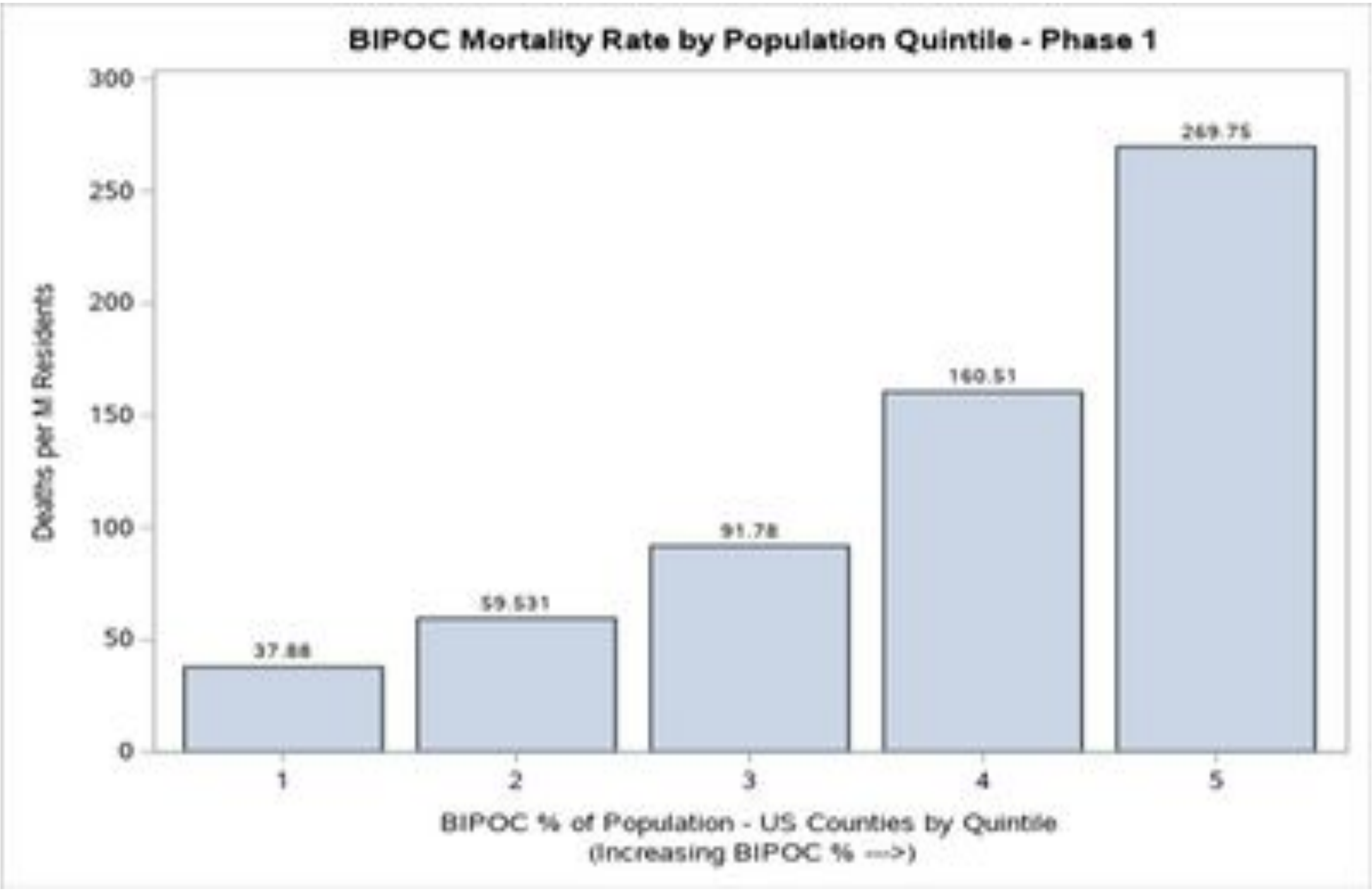
=> Biased Predictors = Biased Outcome

PEACE-WORK

STATISTICAL MEASURES OF BIAS

Measuring Bias: Disparate Impact

Example: COVID-19 Initial Mortality



PEACE-WORK

Measuring Bias: Disparate Impact

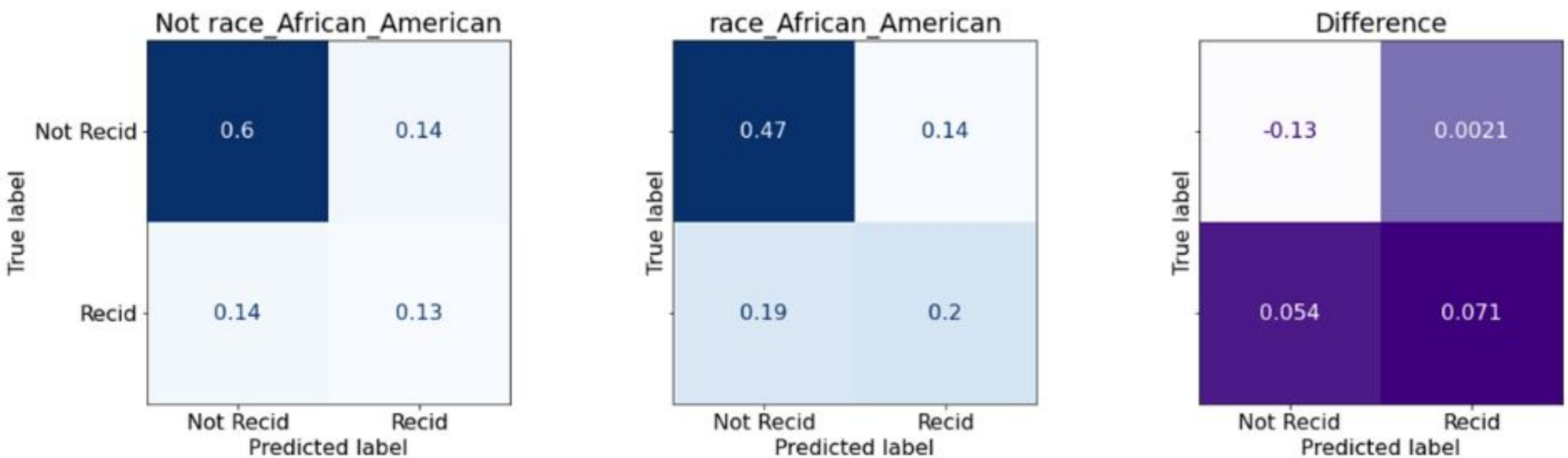
Odds Ratios for demographic factors compare highest % prevalence (60%+) vs. lowest (<5%)

Black / African American	10.1
Cardiovascular Disease	9.3
Chronic Lung Disease	5.9
Prison Populations	5.5
Indigenous	3.3
Poverty (High % Below Poverty Line)	2.9
High Population Density	1.9

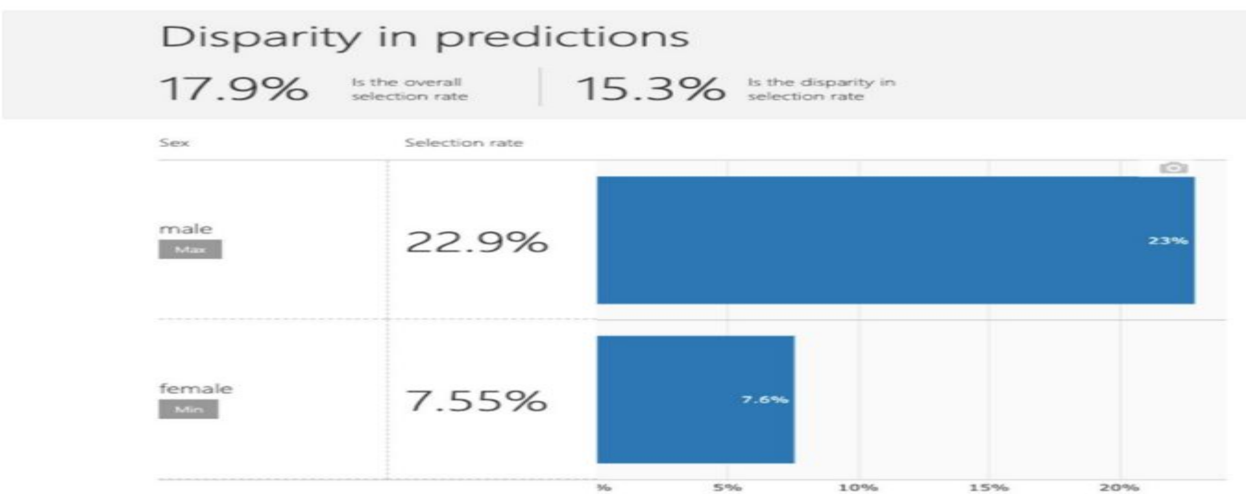
Prison numbers compared to overall US population. Reported by Saloner et al, COVID-19 Cases and Deaths in Federal and State Prisons, JAMA, August 11, 2020

PEACE-WORK

Measuring Bias: Fairlearn Algorithm

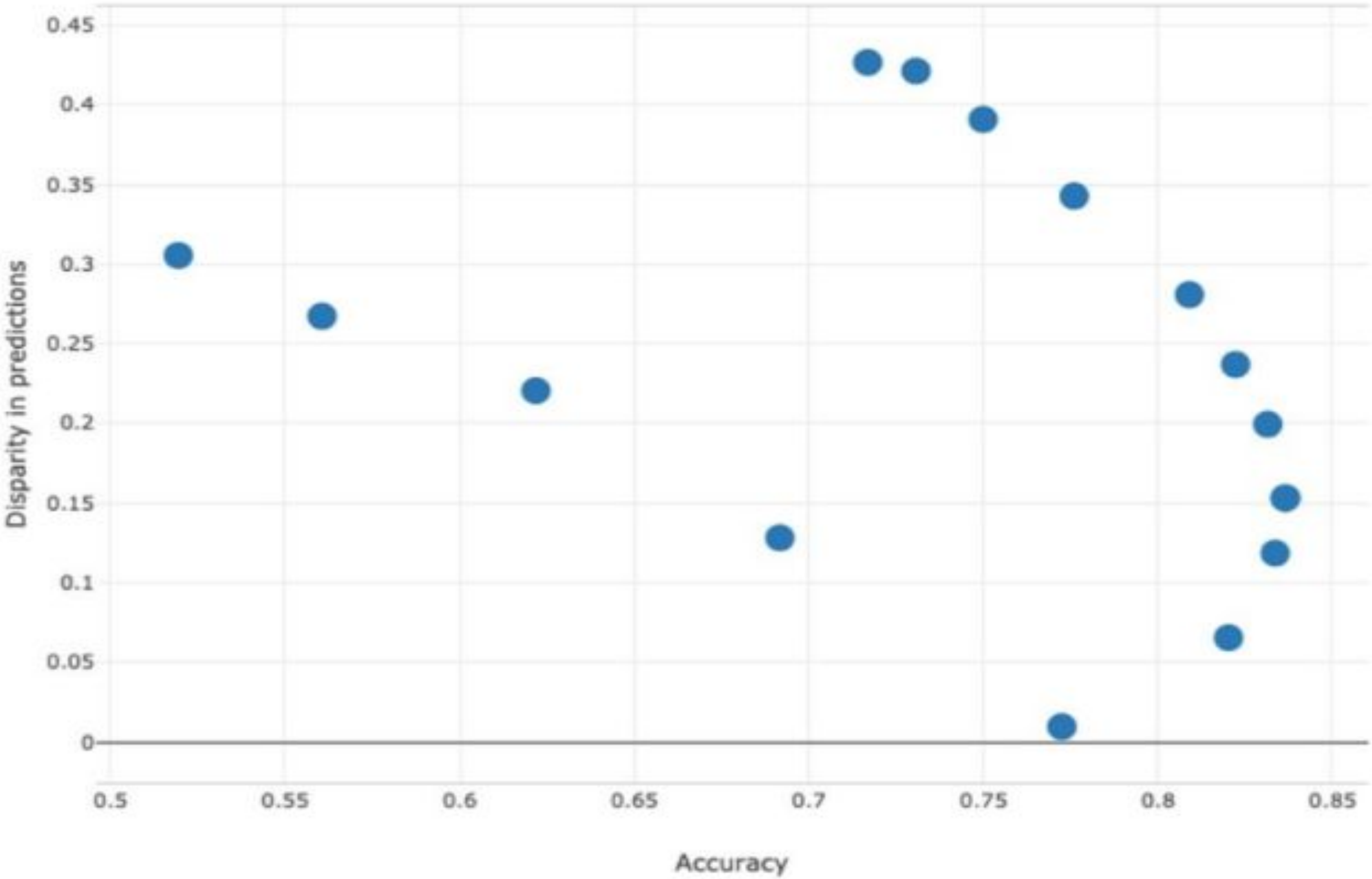


Confusion matrices for African-American defendants vs rest, and difference, for Fairlearn-adjusted model



DESIGNING AGAINST BIAS IN ML AND AI

Designing Against Bias: Fairlearn



Designing Against Bias: Bias-Minimized Comparison Algorithm

1. Develop a new predictive algorithm
2. Create a second model - the BMCA - by removing predictors that might confer bias
3. Test the new model against the BMCA to estimate the amount of bias in any variables causing concern

PEACE-WORK

CONCLUSIONS

Best Practices for Design to Minimize Bias

1. Parsimonious Models
2. Screen all predictors for bias
3. Transparent Methods, not Black Box
4. Develop the model using new outcomes screened for bias - not past human decisions
5. Test for bias w/ FairLearn, BCMA, etc.
6. Present using Odds Ratios or Relative Risk
7. Open Source the data and algorithm

PEACE-WORK

References

Corliss, D. (2021), *Disproportional Impact of COVID-19 on Marginalized Communities*, Proc. SAS Global Forum 2021

Dastin, J., (2018), Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, October 2018
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Hielbrun, L., Comment on testing using relative risk (personal communication), February 2023

Larson J., Mattu S., Kirchner L., Angwin J. (2016), *How We Analyzed the COMPAS Recidivism Algorithm*, ProPublica

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Larson J., Mattu S., Kirchner L., Angwin J. (2016), *COMPAS Recidivism Risk Score Data and Analysis*, ProPublica

https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

Najibi, A. (2020), *Racial Discrimination in Face Recognition Technology*, Gender Shades Project, Harvard
https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

New York Times, Corona Virus County Data, New York Times, 2021

New York Times COVID-19 Data, accessed 10/8/2021: https://github.com/nytimes/covid-19-data

Owen, S. (2022), Mitigating Bias in Machine Learning With SHAP and Fairlearn, Databricks
https://www.databricks.com/blog/2022/09/16/mitigating-bias-machine-learning-shap-and-fairlearn.html

US Census Bureau Demographic Data

https://www.census.gov/programs-surveys/ces/data/restricted-use-data/demographic-data.html

PEACE-WORK

Questions?

David J Corliss, PhD
Peace-Work
E-mail: davidjcorliss@peace-work.org

PEACE-WORK