Designing Against Bias in Machine Learning and AI

David J Corliss, PhD
AVP – Data Science, OnStar Insurance

Great Lakes Data and Analytics Summit
April 27, 2023

Designing Against Bias in Machine Learning and AI

David J Corliss, PhD is an AVP Technical Expert in Data Science at GM OnStar Insurance. 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, serving on the Data User Advisory Committee for the US Bureau of Labor Statistics, and was recently named to the steering committee of the Statistics section of the American Association for the Advancement of Science. Outside of work, Dr. Corliss is the founder of Peace-Work, a volunteer cooperative of statisticians, data scientists and other researchers applying analytics in issue-driven advocacy.

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??

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

Gender Bias: Amazon Resume Screening

The "Solution": ML picks top resumes

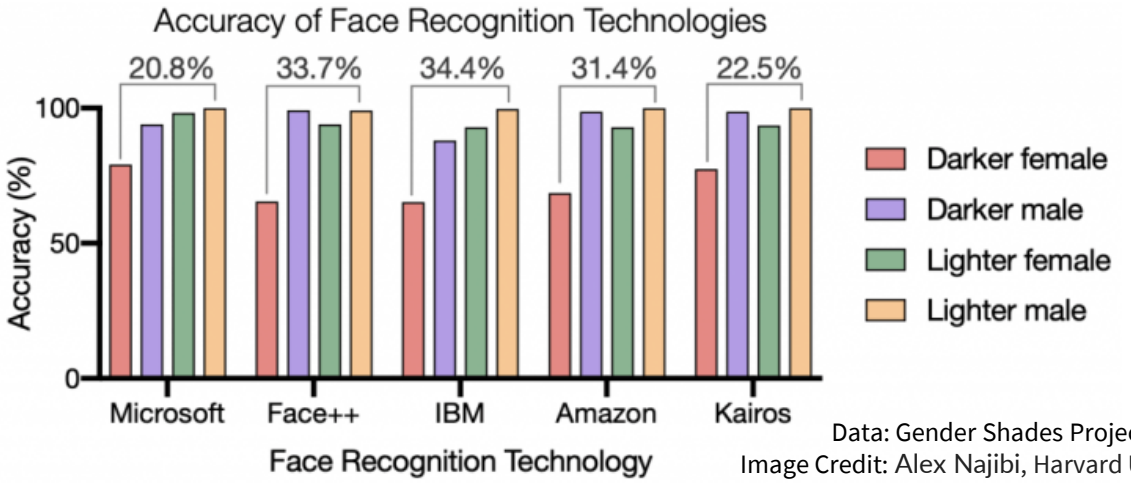
Amazon Algorithm:

Resume Quality = ? + ? + ? + ? + ? …

Image Credit: flazingo_photos - CC BY-SA 2.0

The Problem: the algorithm is biased against women applicants

Root Causes of Bias: Selection Bias



Algorithm trained using biased subset

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

Usage results in disparate impact

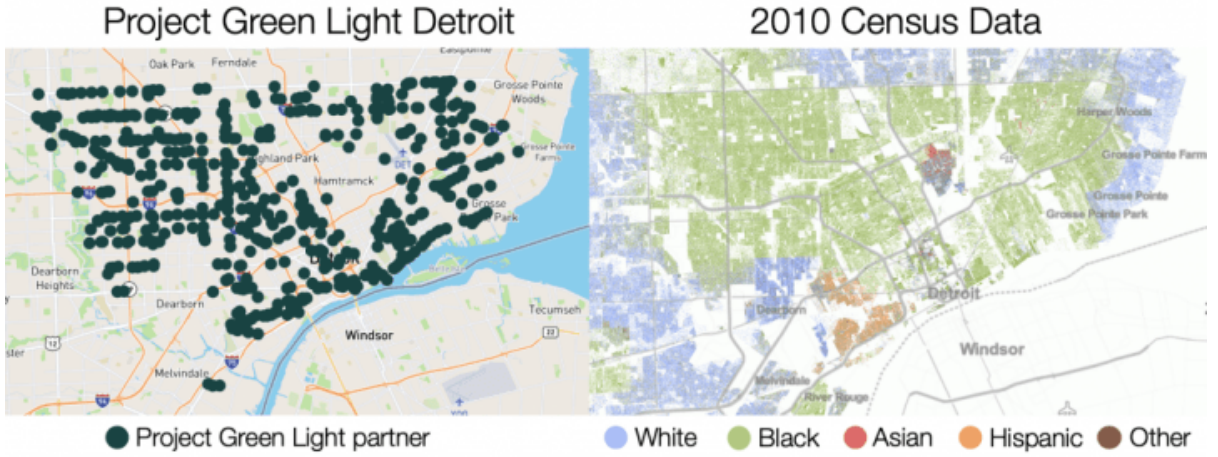


Image Credit: Alex Najibi, Harvard University

=> Biased Training Population = Biased Results

Root Causes of Bias: The History Problem

ML replaces human decision making

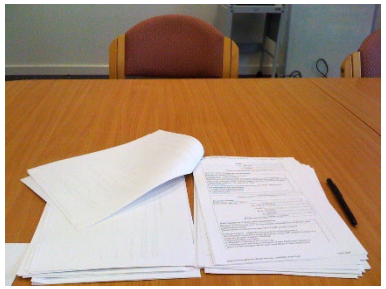
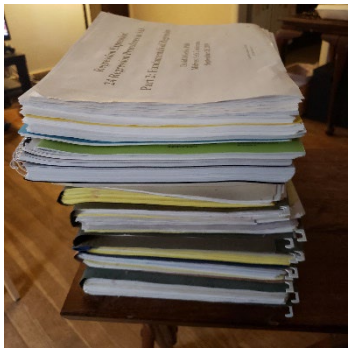


Image Credit: David Davies -CC BY-SA 2.0

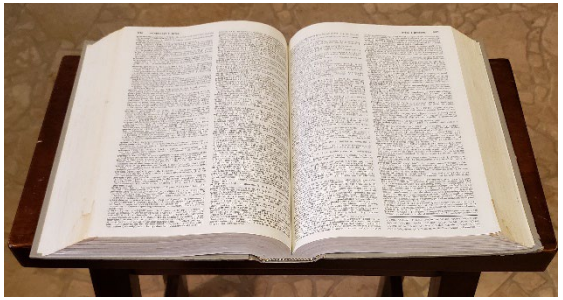
```
library(tensorflow)
library(keras)
model <- keras_model_sequential() %>%
  layer_conv_2d(filters = 32,
  kernel_size = c(3,3), activation = "relu",
```

The algorithm is trained using earlier, biased human decisions



=> Bias In = Bias Out

Root Causes of Bias: Spaghetti Problem



DATA DICTIONARY

Hundreds or even thousands of potential predictors

Algorithm trained uncritically using “anything that sticks”

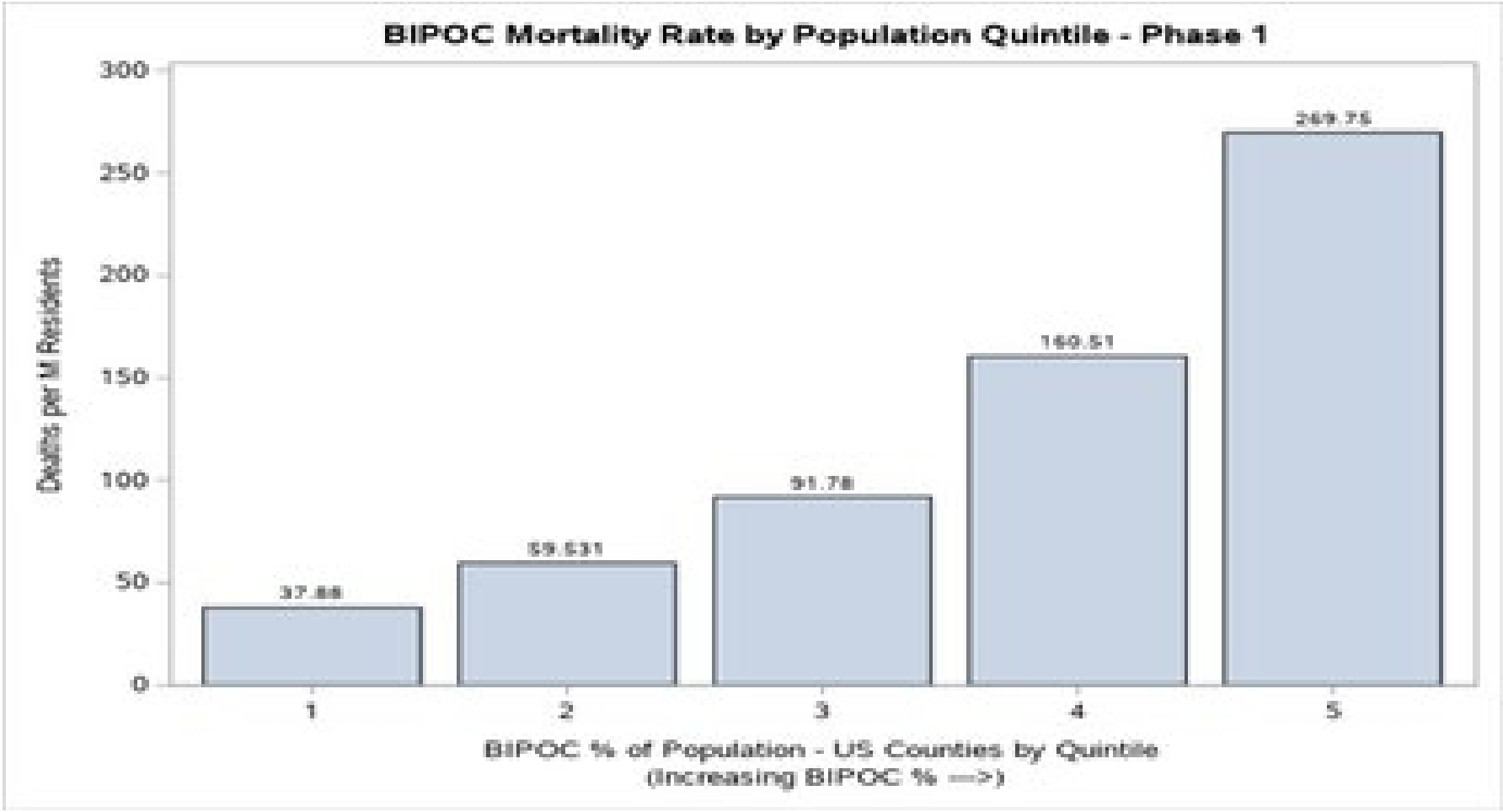
Image Credit: snackdinner.com

=> Biased Predictors = Biased Outcome

GREAT LAKES DATA & ANALYTICS SUMMIT

Measuring Bias: Disparate Impact

Example: COVID-19 Initial Mortality



Measuring Bias: Disparate Impact

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

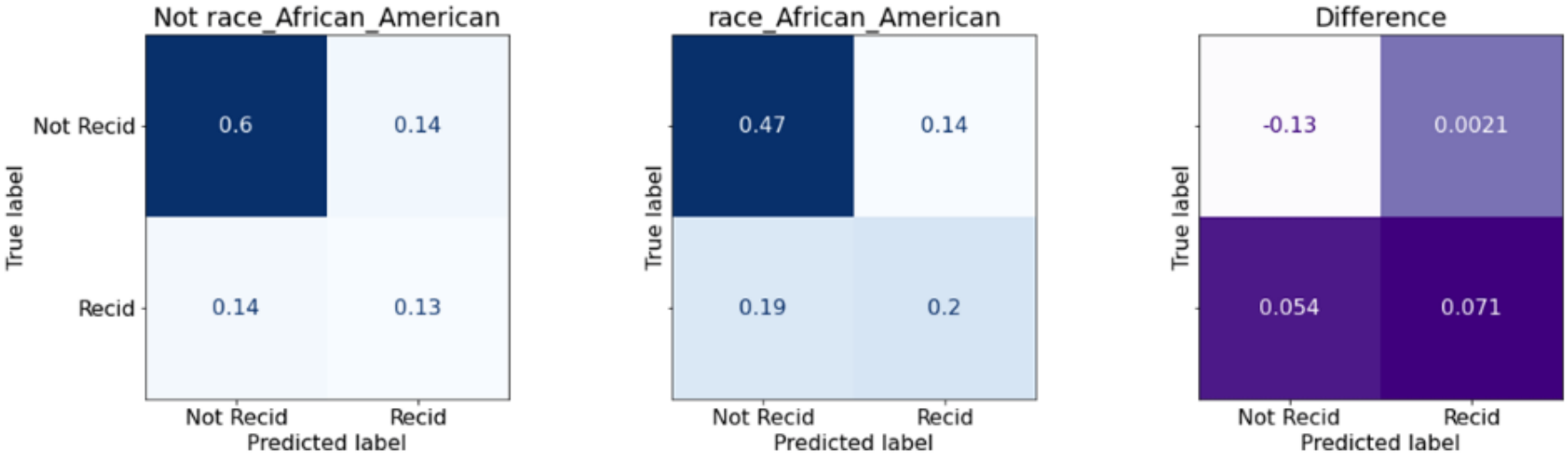
Black / African American	10.1
Cardiovascular Disease	9.3
Chronic Lung Disease	5.9
Prison Populations	5.5
Indigenous	3.3
Poverty (High % Below Poverty Line)	2.9
High Population Density	1.9

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

Designing Against Bias: Bias-Minimized Comparison Algorithm

1. Create a second predictive model (BMCA)
2. Screen input variables for minimum bias
3. Transparent Algorithm– Regression, Decision Trees, etc. – not Black Box
4. Develop BMCA against new outcomes, not past decisions and tune for minimum bias
5. Test models vs. BMCA using Odds Ratios

Measuring Bias: Fairlearn Algorithm



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

Bias Mitigation with Fairlearn

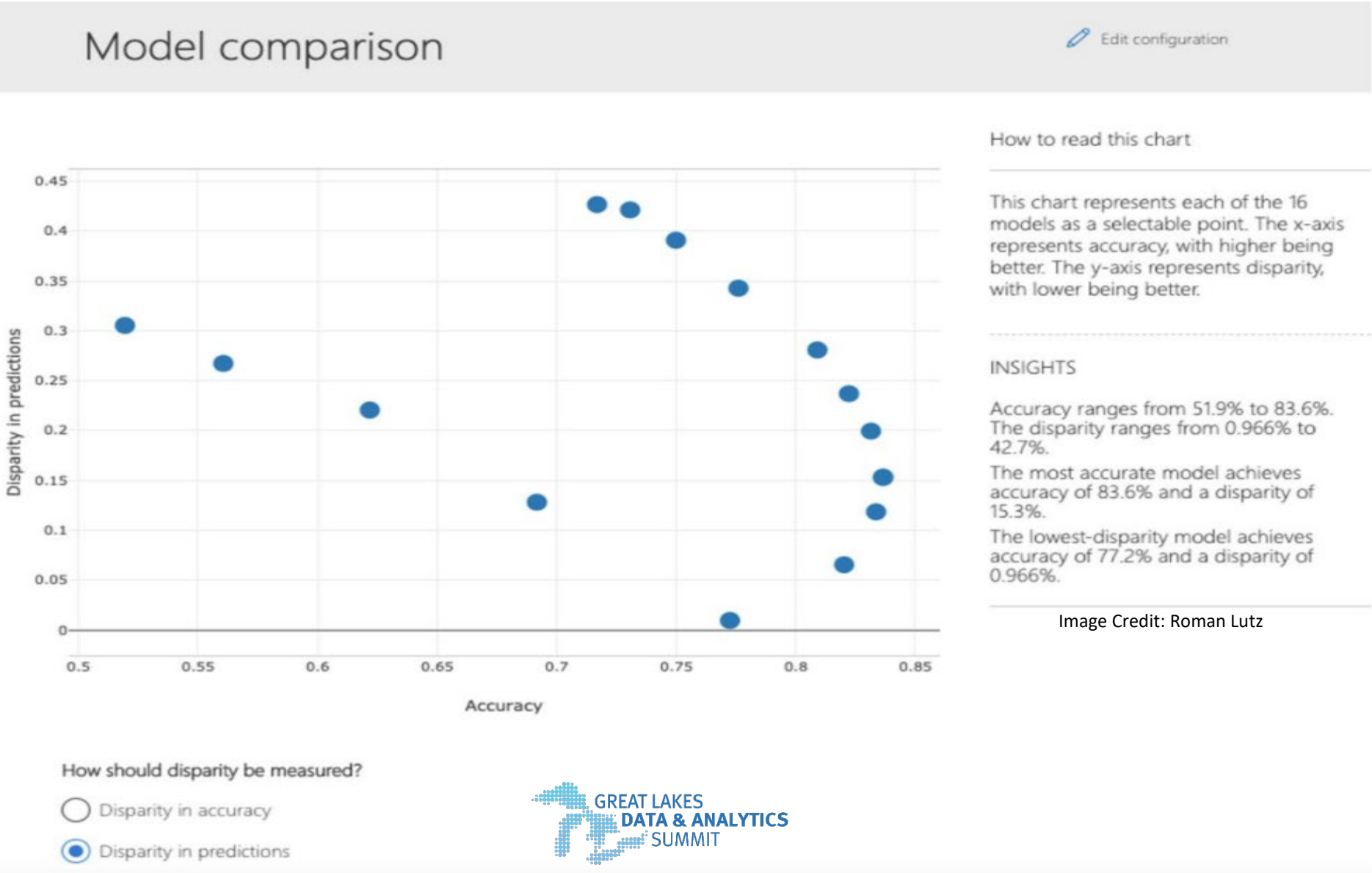


Image Credit: Roman Lutz

Summary of Best Practices 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 decisions
5. Test using Fairlearn and/or a BMCA
6. Present results using Odds Ratios
7. Open Source the data and algorithm

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

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 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

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker., K. (2020), Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

US Census Bureau Demographic Data

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

Questions?

David J Corliss
AVP – Data Science
OnStar Insurance
E-mail: david.corliss@gmfinancial.com

GREAT LAKES DATA & ANALYTICS SUMMIT