

# Measuring Model Fairness

J. Henry Hinefeld



hhinnefeld@civisanalytics.com



hinnefe2.github.io



DrJSomeday

# Outline

1. Motivation
2. Subtleties of measuring fairness
3. Case Study
4. Python tools
5. Conclusion



Models determine whether you can buy a home...

# Credit Score

☒ Excellent

☐ Good

☐ Fair



and what advertisements you see ...



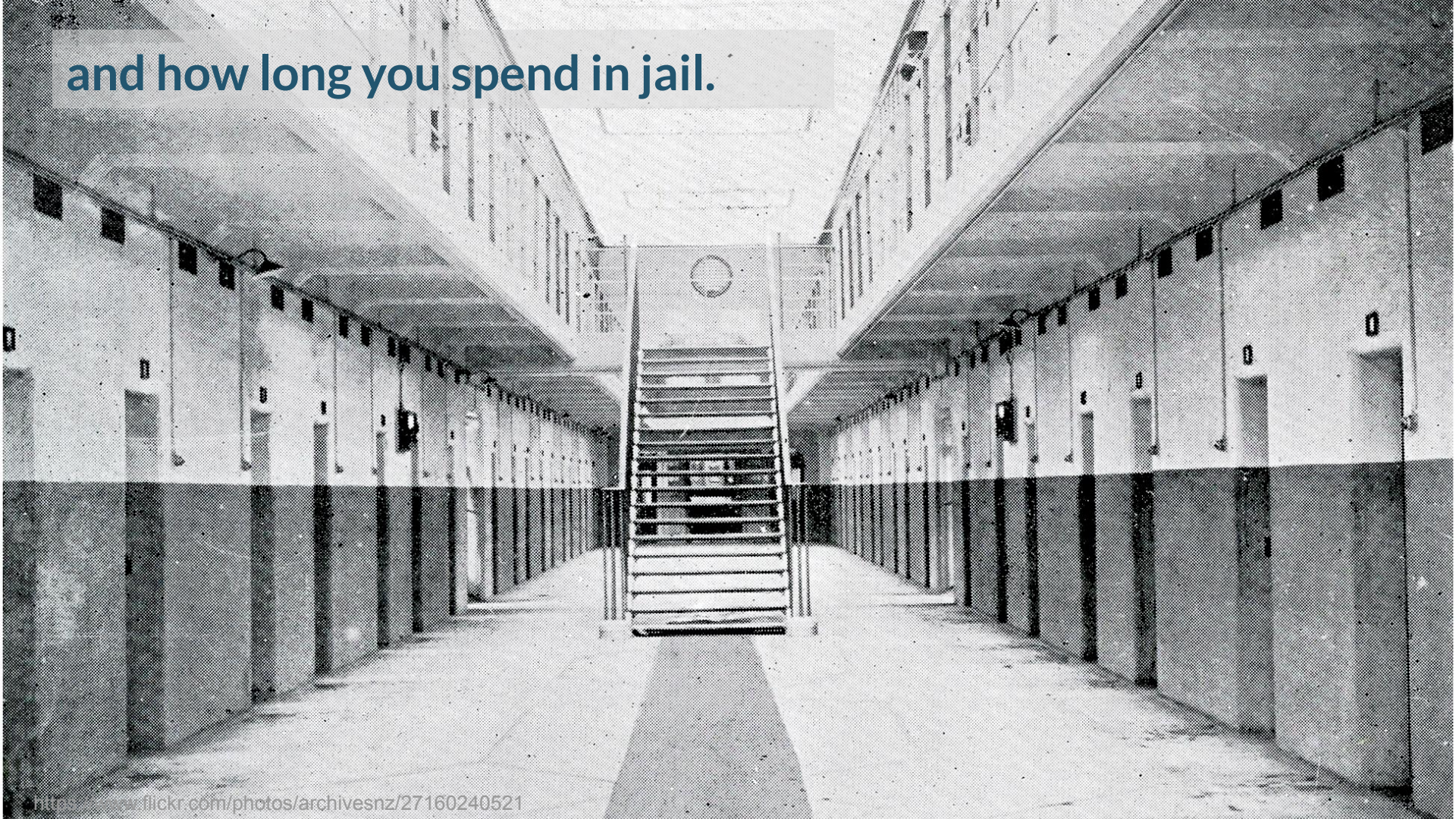
# Jobs

## 100 Help Wanted

Volunteer-Admin Manager (P/T)  
Foundation and Family Office  
high intelligence and ability  
without supervision, involves  
investment management  
person office



and how long you spend in jail.





# How do you measure if your model is fair?



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

# How do you measure if your model is fair?

## 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%

*Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)*

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

# How do you measure if your model is fair?

**TABLE 7: Area Under the Curve Values for COMPAS Risk Models Predicting Any Offense, Offenses Against Persons, and Felony Offenses for White and African American Men**

<i>Model</i>	<i>White Men (n = 1,412)</i>			<i>African American Men (n = 296)</i>		
	<i>Any</i>	<i>Person</i>	<i>Felony</i>	<i>Any</i>	<i>Person</i>	<i>Felony</i>
COMPAS I	.69	.74	.73	.64	.69	.69
COMPAS II	.71	.75	.75	.66	.71	.72
Recidivism Risk III	.69	.71	.71	.67	.72	.73

<http://www.northpointeinc.com/files/publications/Criminal-Justice-Behavior-COMPAS.pdf>

The predictive accuracy of the COMPAS recidivism score was consistent between races in our study - 62.5 percent for white defendants vs. 62.3 percent for black defendants. The authors of the Northpointe study found a small difference in the concordance scores by race: 69 percent for white defendants and 67 percent for black defendants.

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



# How do you decide which measure of fairness is appropriate?



## Inherent Trade-Offs in the Fair Determination of Risk Scores


Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan. 2016. <https://arxiv.org/abs/1609.05807>

<https://pixabay.com/en/legal-scales-of-justice-judge-450202/>

## Subtlety #1:

Different groups  
can have different  
ground truth  
positive rates

## U.S. Breast Cancer Statistics

 [Save as Favorite](#)

 Sign in to receive recommendations ([Learn more](#))



- About 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer over the course of her lifetime.
- In 2018, an estimated 266,120 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 63,960 new cases of non-invasive (in situ) breast cancer.
- About 2,550 new cases of invasive breast cancer are expected to be diagnosed in men in 2018. A man's lifetime risk of breast cancer is about 1 in 1,000.



## Certain fairness metrics make assumptions about the balance of ground truth positive rates

**Disparate Impact** is a popular metric which assumes that the ground truth positive rates for both groups are the same

**Definition 1.1** (Disparate Impact (“80% rule”)). *Given data set  $D = (X, Y, C)$ , with protected attribute  $X$  (e.g., race, sex, religion, etc.), remaining attributes  $Y$ , and binary class to be predicted  $C$  (e.g., “will hire”), we will say that  $D$  has disparate impact if*

$$\frac{\Pr(C = \text{YES} | X = 0)}{\Pr(C = \text{YES} | X = 1)} \leq \tau = 0.8$$

## Subtlety #2: Your data is a biased representation of ground truth

Datasets can contain **label bias** when a protected attribute affects the way individuals are assigned labels.

In addition, the results indicate that students from African American and Latino families are more likely than their White peers to receive expulsion or out of school suspension as consequences for the same or similar problem behavior.

"Race is not neutral: A national investigation of African American and Latino disproportionality in school discipline." Skiba et al.

A dataset for predicting “student problem behavior” that used “has been suspended” for its label could contain label bias.



## Certain fairness metrics are based on agreement with possibly biased labels

**Equal Opportunity** is a popular metric which compares the True Positive rates between protected groups

**Definition 2.2** (Equal opportunity). We say that a binary predictor  $\widehat{Y}$  satisfies *equal opportunity* with respect to  $A$  and  $Y$  if  $\Pr\{\widehat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\widehat{Y} = 1 \mid A = 1, Y = 1\}$ .

Equality of Opportunity in Supervised Learning, Hardt et al. (<https://arxiv.org/pdf/1610.02413.pdf>)

## Subtlety #2: Your data is a biased representation of ground truth

Datasets can contain **sample bias** when a protected attribute affects the sampling process that generated your data.

We find that persons of African and Hispanic descent were stopped more frequently than whites, even after controlling for precinct variability and race-specific estimates of crime participation.

"An analysis of the NYPD's stop-and-frisk policy in the context of claims of racial bias" Gelman et al.

A dataset for predicting contraband possession that used stop-and-frisk data could contain sample bias.



## Certain fairness metrics compare classification ratios between groups

**Disparate Impact** is a popular metric which compares the ratio of positive classifications between groups

**Definition 1.1** (Disparate Impact (“80% rule”)). *Given data set  $D = (X, Y, C)$ , with protected attribute  $X$  (e.g., race, sex, religion, etc.), remaining attributes  $Y$ , and binary class to be predicted  $C$  (e.g., “will hire”), we will say that  $D$  has disparate impact if*

$$\frac{\Pr(C = \text{YES} | X = 0)}{\Pr(C = \text{YES} | X = 1)} \leq \tau = 0.8$$

## Subtlety #3:

**It matters whether the modeled decision's consequences are positive or negative**

When a model is **punitive** you might care more about False Positives.

When a model is **assistive** you might care more about False Negatives.

The point is you have to think about these questions.

# We can't math our way out of thinking about fairness

You still need a person to think about the ethical implications of your model

Originally people thought “Models are just math, so they must be fair”

**definitely not true**

Now there's a temptation to say ‘Adding this constraint will make my model fair’

**still not automatically true**

# Can we detect real bias in real data?

Spoiler: it can be tough!

- Start with real data from Civis's work
  - Features are demographics, outcome is a probability
  - Consider racial bias; white versus African American



# Can we detect real bias in real data?

Create artificial datasets with known bias; then we'll see if we can detect it.

- Start with real data from Civis's work
  - Features are demographics, outcome is a probability
  - Consider racial bias; white versus African American
- Two datasets:
  - Artificially balanced: select white subset and randomly re-assign race
  - Unmodified (imbalanced) dataset

Next introduce known sample and label bias

Sample bias: protected class affects whether you're in the sample at all

- Create a modified dataset with labels taken from the original data

$$\tilde{P}(x \in X) = \begin{cases} 0.8 & \text{if race = white and score} \geq 0.5 \\ 0.2 & \text{if race = white and score} < 0.5 \\ 1 & \text{if race = black} \end{cases}$$

Next introduce known sample and label bias

Label bias: you're in the dataset, but protected class affects your label

- Use the original dataset but modify the labels

$$\tilde{Y} = \begin{cases} 1 \text{ if score} \geq 0.3, \text{ else } 0 & \text{if race} = \text{white} \\ 1 \text{ if score} \geq 0.7, \text{ else } 0 & \text{if race} = \text{black} \end{cases}$$

There are many possible metrics for model fairness

These are two popular ones

Disparate Impact

$$\frac{\Pr \left\{ \hat{Y} = 1 \mid S = 1 \right\}}{\Pr \left\{ \hat{Y} = 1 \mid S = 0 \right\}}$$

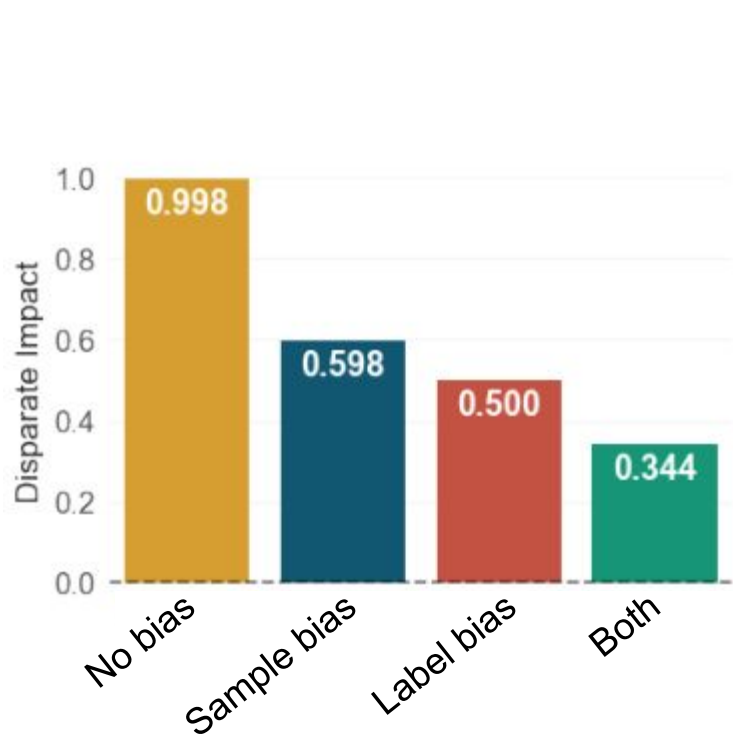
Equal Opportunity

$$\Pr \left\{ \hat{Y} = 1 \mid S = 1, Y = 1 \right\} - \Pr \left\{ \hat{Y} = 1 \mid S = 0, Y = 1 \right\}$$

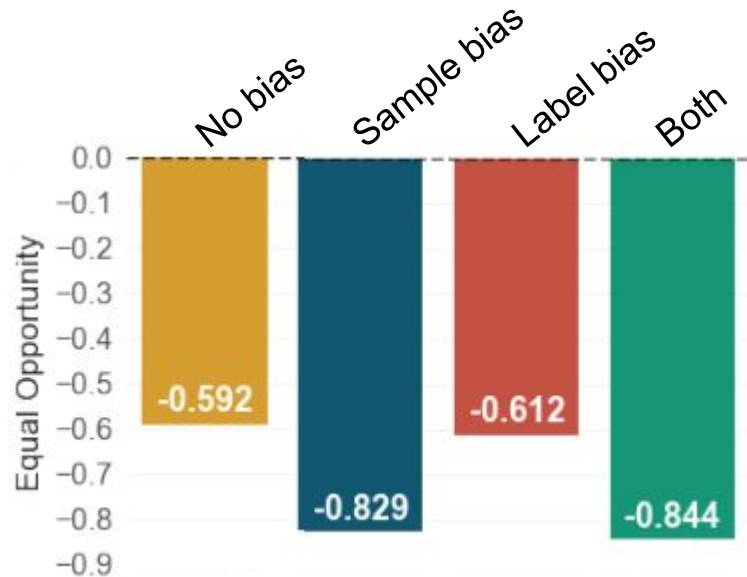
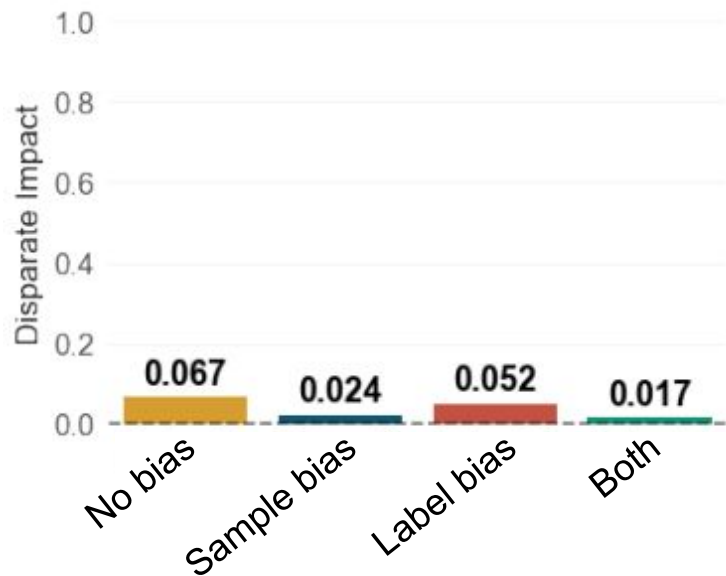


# With balanced ground truth, both metrics detect bias

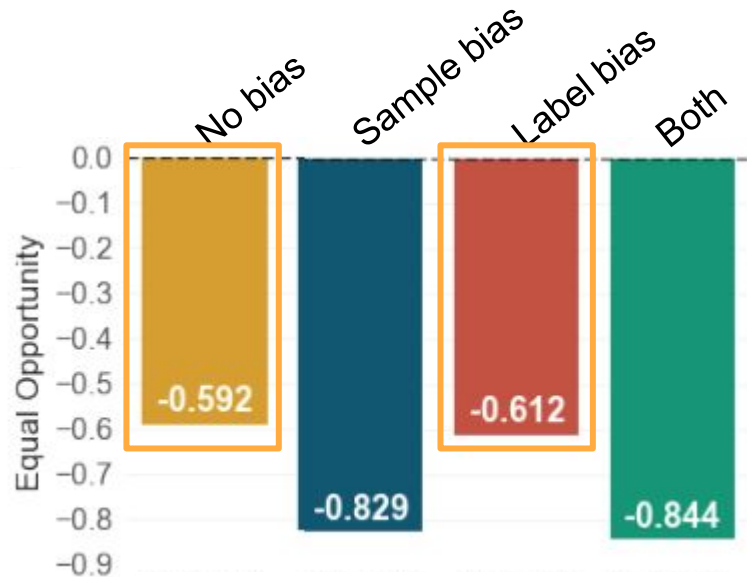
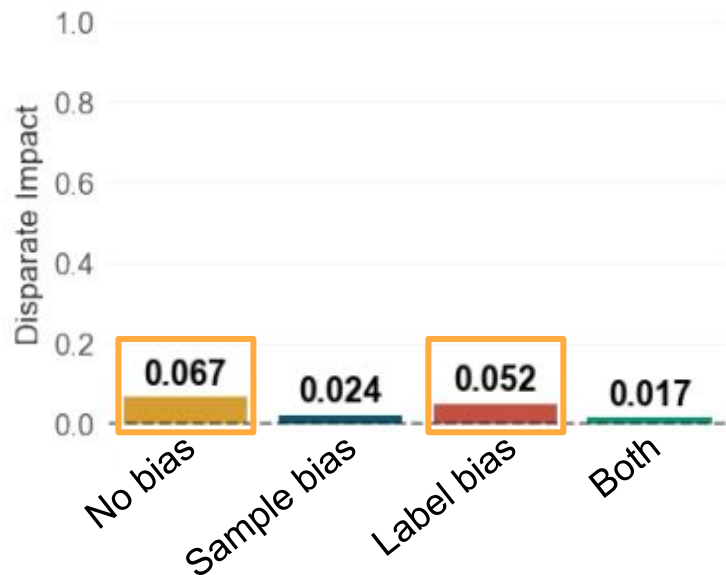
Good news!



With imbalanced ground truth, both metrics still detect bias...  
...even when there isn't any bias in the "truth".



Label bias is particularly hard to detect when the ground truth is imbalanced



There are open source python tools for measuring model fairness  
once you've decided what definition of 'fair' makes sense for your problem

# Aequitas

Bias & Fairness Audit

[aequitas.dssg.io](https://aequitas.dssg.io)

- **Pro:** easy to use
- **Con:** non-standard license





There are open source python tools for measuring model fairness  
once you've decided what definition of 'fair' makes sense for your problem



AI Fairness 360 Open Source Toolkit

[github.com/IBM/AIF360](https://github.com/IBM/AIF360)

- Pro: comprehensive, lots of documentation + tutorials
- Con: more comprehensive than you need, lots of dependencies

There are open source python tools for measuring model fairness  
once you've decided what definition of 'fair' makes sense for your problem

Model interpretation tools: LIME and SHAP

[github.com/marcotcr/lime](https://github.com/marcotcr/lime)

[github.com/slundberg/shap](https://github.com/slundberg/shap)

- Pro: offer a deeper understanding of your model's behavior
- Con: harder to explain, existing code is research quality

There's no one-size fits all solution

Except for "think hard about your inputs and your outputs"

- These metrics (and others) can help but you have to use them carefully

$$\frac{\Pr \left\{ \widehat{Y} = 1 \mid S = 1 \right\}}{\Pr \left\{ \widehat{Y} = 1 \mid S = 0 \right\}}$$

$$\Pr \left\{ \widehat{Y} = 1 \mid S = 1, Y = 1 \right\} - \Pr \left\{ \widehat{Y} = 1 \mid S = 0, Y = 1 \right\}$$

# There's no one-size fits all solution

Except for "think hard about your inputs and your outputs"

- These metrics (and others) can help but you have to use them carefully
- Use a diverse team to create the models and think about these questions!



<https://imgur.com/gallery/hem9m>



# There's no one-size fits all solution

Except for "think hard about your inputs and your outputs"

- These metrics (and others) can help but you have to use them carefully
- Use a diverse team to create the models and think about these questions!
- Know your data and think about your consequences



“Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that’s something only humans can provide. We have to explicitly embed better values into our algorithms, creating Big Data models that follow our ethical lead. Sometimes that will mean putting fairness ahead of profit.”

— **Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy**