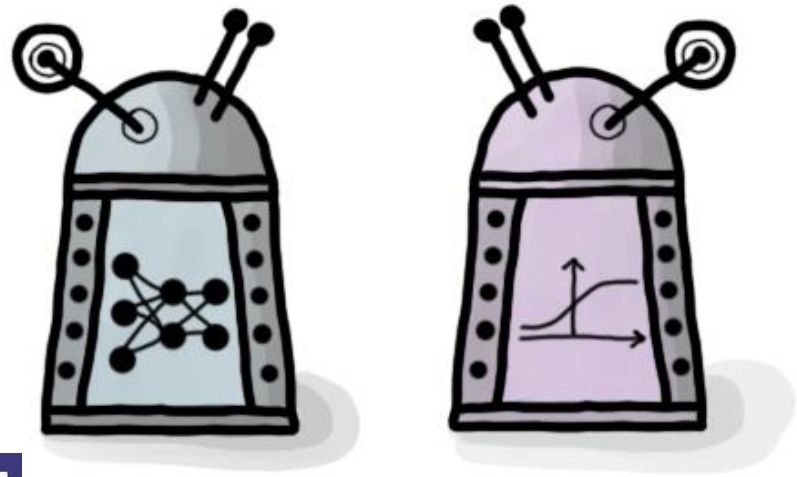
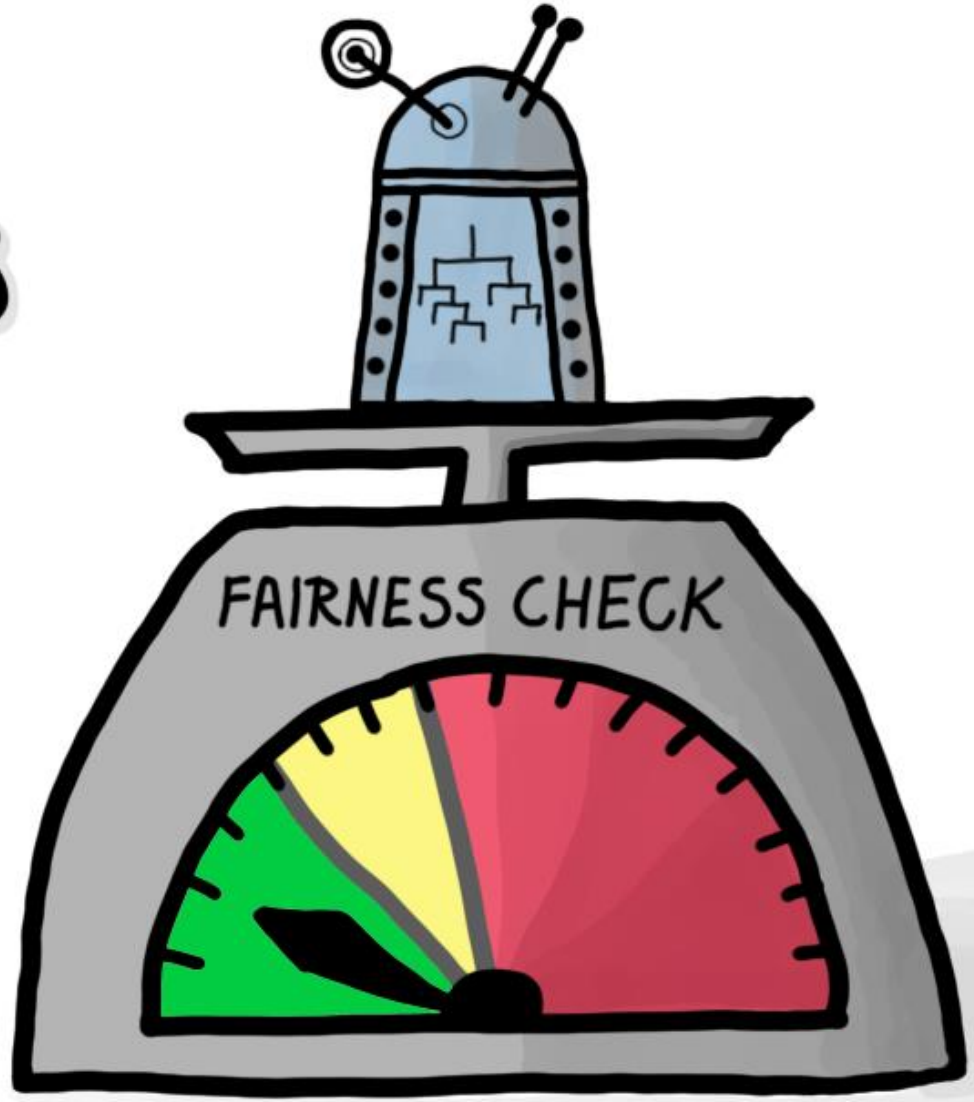
fairmodels

Jakub Wiśniewski
joint work with Przemysław Biecek





Faculty of Mathematics and Information Science
WARSAW UNIVERSITY OF TECHNOLOGY

Why is It important?

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them

propublica.org

When we analyze the results by intersectional subgroups - darker males, darker females, lighter males, lighter females - we see that all companies perform worst on darker females.

IBM and Microsoft perform best on lighter males. Face++ performs best on darker males.

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%

gendershades.org

GOOGLE \ TECH \ ARTIFICIAL INTELLIGENCE

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By James Vincent | Jan 12, 2018, 10:35am EST

SHARE

theverge.com

What is bias?

Ethnicity	Sex
African_American	Male
African_American	Male
African_American	Male
Other	Male
African_American	Female
Hispanic	Female

- Bias can have many sources
- Different treatment of some subgroups by model
  - Subgroups will be later called protected (vector)
  - One will be called privileged
- Can be described by non-discrimination criteria

Non-discrimination criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Fairness and Machine Learning, Barocas et al. (2019) – fairmlbook.org

Y – binary label
R – numerical response of a model
A – protected vector

How to measure it?

- With metrics
- From confusion matrix for each subgroup
- The metrics are either some form of relaxation or equivalents of Independence, Separation and Sufficency

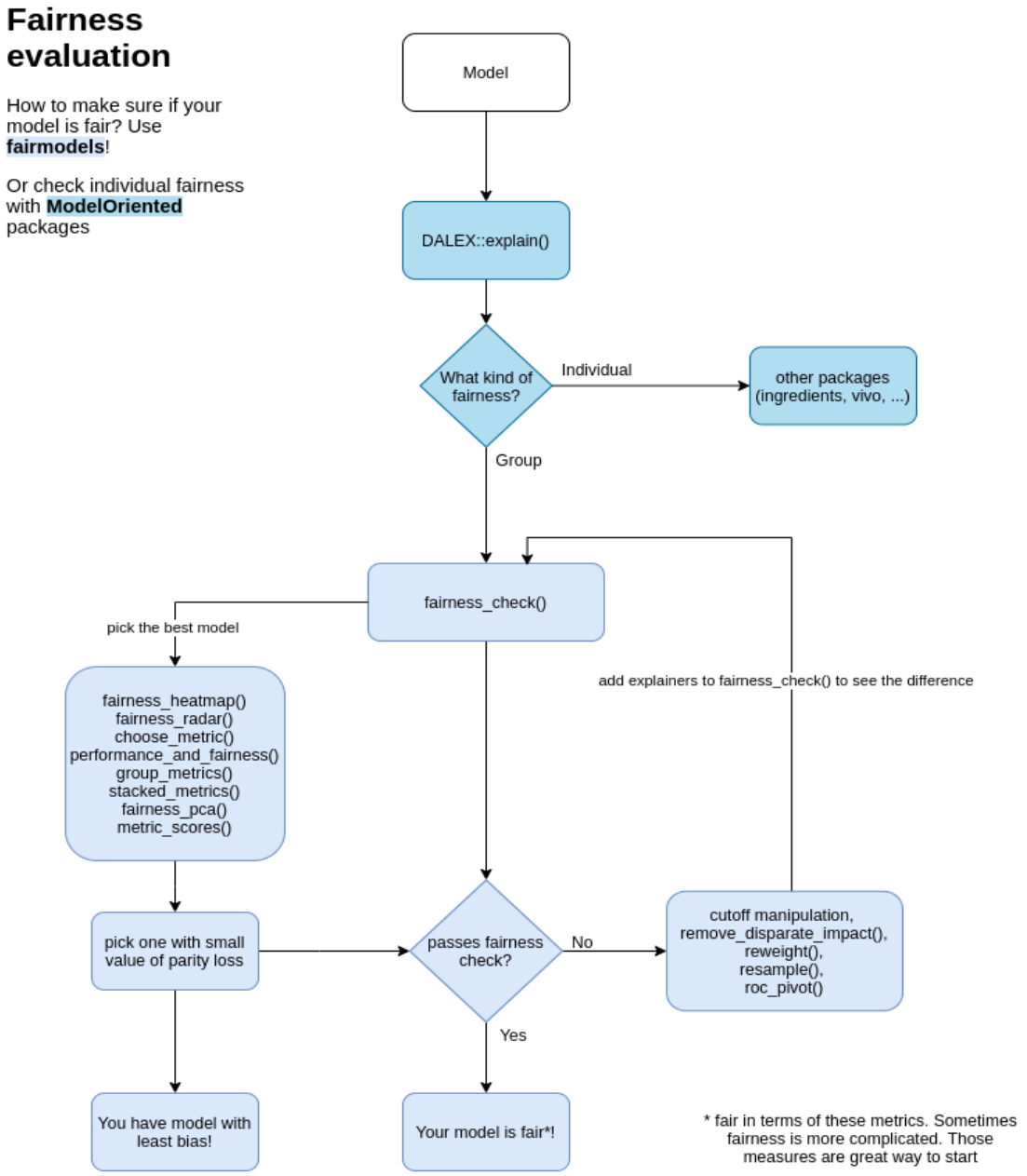
Metric	Formula	Name	Fairness criteria
TPR	$\frac{TP}{TP+FN}$	True positive rate	Equal opportunity (Hardt et al., 2016)
TNR	$\frac{TN}{TN+FP}$	True negative rate	
PPV	$\frac{TP}{TP+FP}$	Positive predictive value	Predictive parity (Chouldechova, 2016)
NPV	$\frac{TN}{TN+FN}$	Negative predictive value	
FNR	$\frac{FN}{FN+TP}$	False negative rate	
FPR	$\frac{FP}{FP+TN}$	False positive rate	Predictive equality (Corbett-Davies et al., 2017)
FDR	$\frac{FP}{FP+TP}$	False discovery rate	
FOR	$\frac{FN}{FN+TN}$	False omission rate	
TS	$\frac{TP}{TP+FN+FP}$	Threat score	
STP	$\frac{TP+FP}{TP+FP+TN+FN}$	Positive rate	Statistical parity (Dwork et al., 2012)
ACC	$\frac{TP+TN}{TP+TN+FP+FN}$	Accuracy	Overall accuracy equality (Berk et al., 2017)
F1	$\frac{2 \cdot PPV * TPR}{PPV+TPR}$	F1 score	

Intuition

- 2 subgroups
  - A - privileged
  - B – unprivileged
- Predicting credit rate
- Group A has acceptance rate of 80% and group B 50%
- From group A 90% good credit seekers got credit, meanwhile in group B it was only 60%
- In first case the used metric was STP and in second one TPR

How to do it easily? With fairmodels!

- With use of DALEX
- Group fairness metrics
- Any classification model works
- Iterative approach
  - *fairness_check() > add model > fairness_check()*
- Easy for testing and prototyping



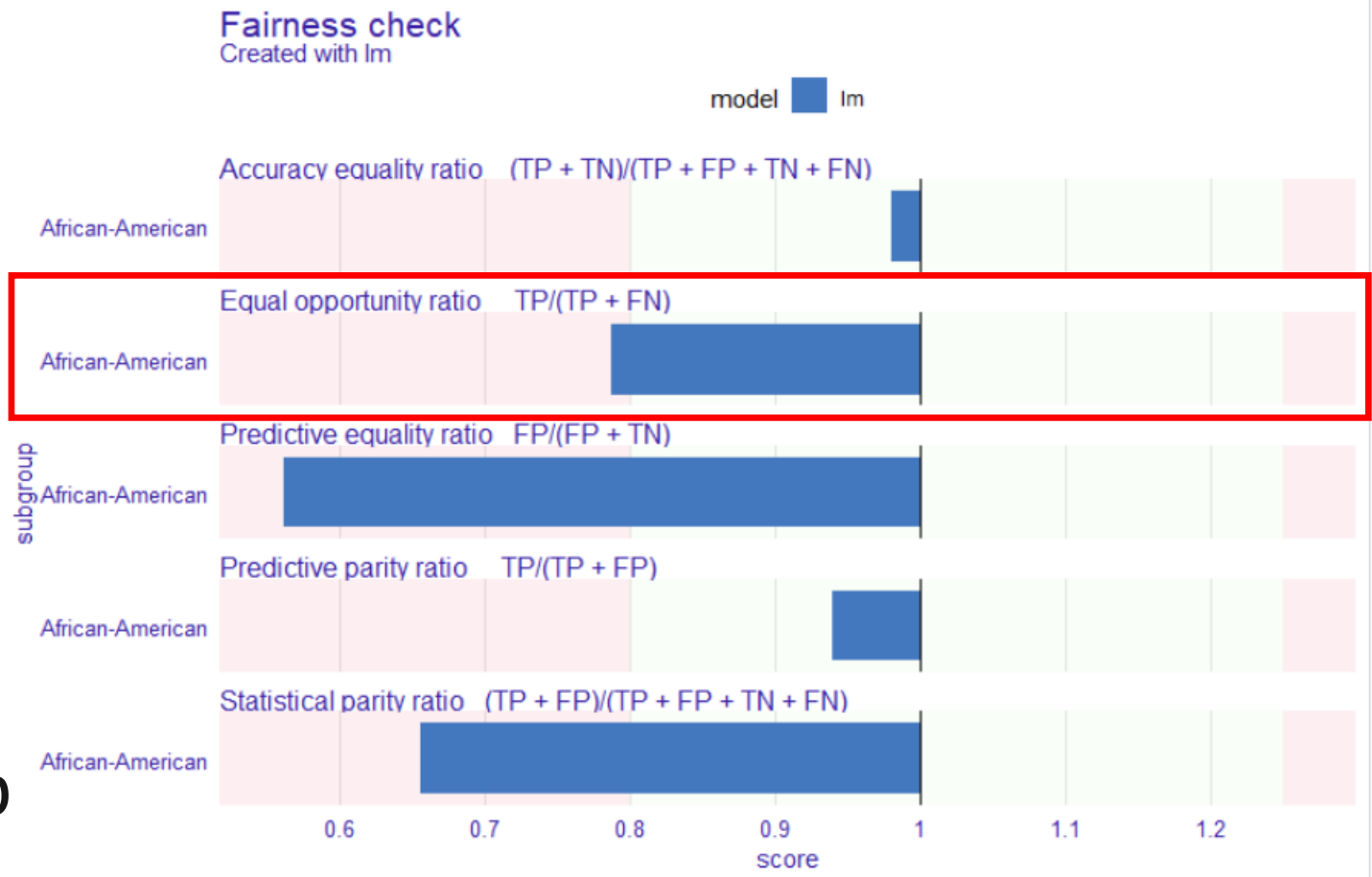
How? Let’s dive into details!

How to read it?

- The value of TPR bar is calculated by

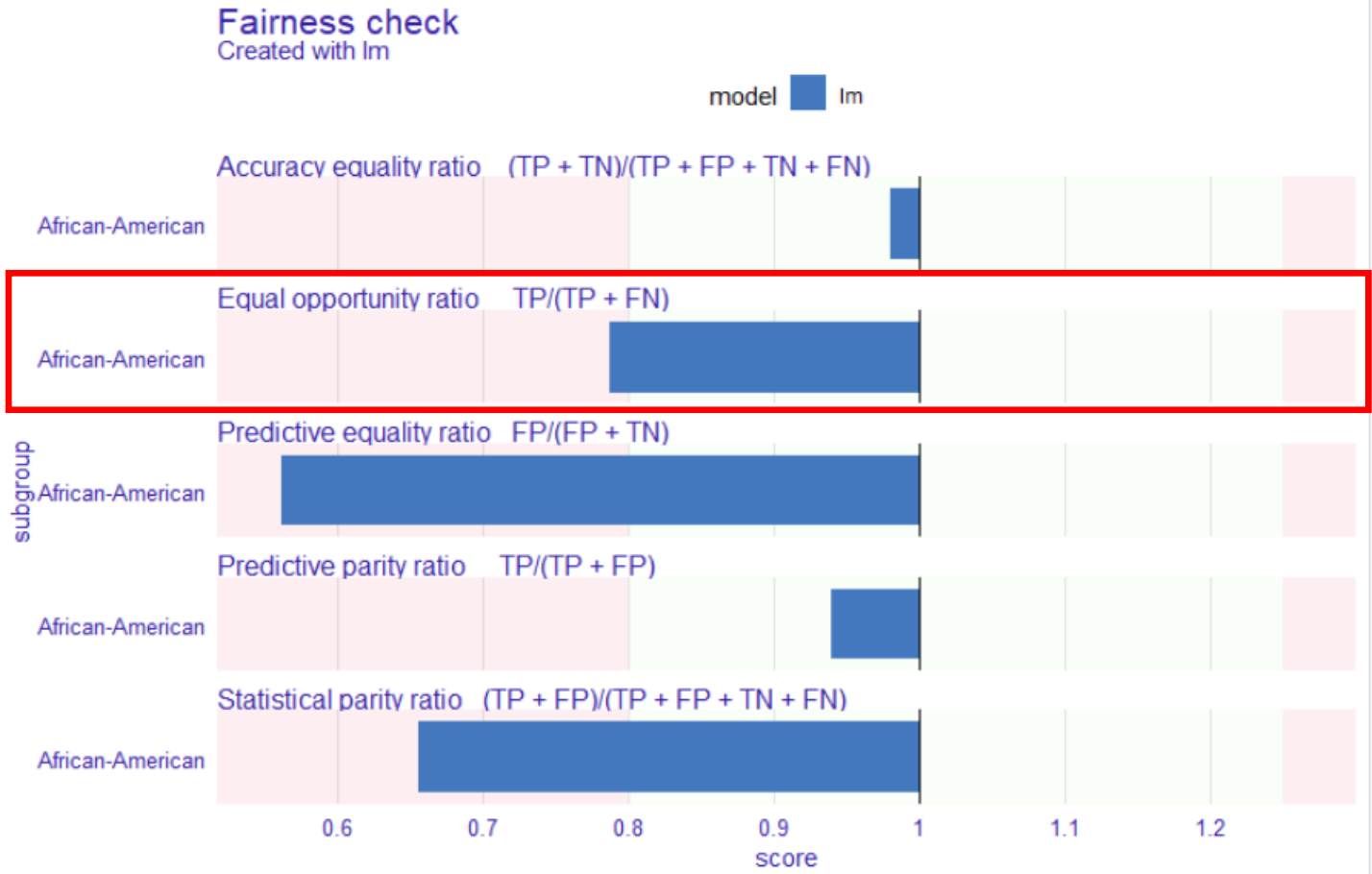
$$\frac{TPR_{African-American}}{TPR_{Caucasian}}$$

- Here *'Caucasian'* subgroup is considered privileged



How to read it?

- Epsilon as boundary
- The epsilon parameter is set to 0.8 due to EEOC four-fifths rule



$$\varepsilon < \frac{TPR_{African-American}}{TPR_{Caucasian}} < \frac{1}{\varepsilon}$$

How does it work?

Visualization tool

- Parity loss
- Example TPR parity loss:

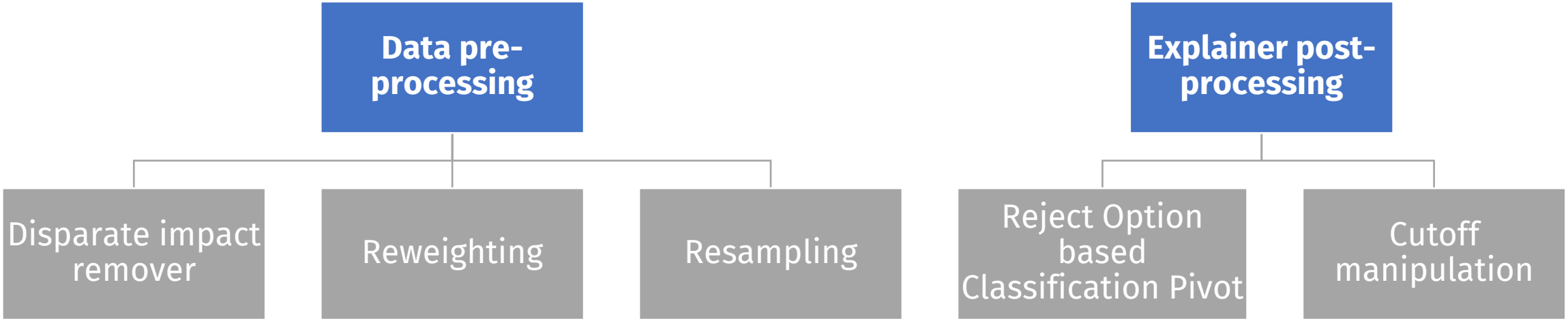
$$TPR_{parity\_loss} = \sum_{i \in \{a,b,\dots\}} \left| \ln\left(\frac{TPR_i}{TPR_a}\right) \right|$$

where a is privileged subgroup

- Intuition: the bigger the difference among subgroups the larger the parity loss

Let’s see how to make visualizations in code

Bias mitigation strategies

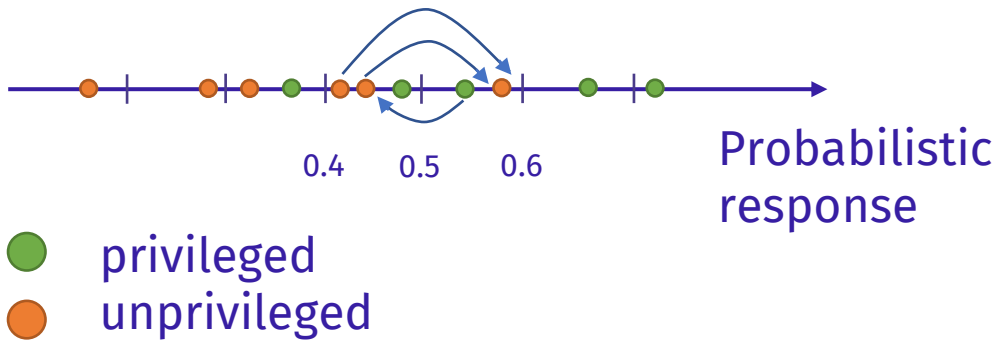


Mitigation in action

- Resample – focuses on mitigating STP
  - Duplicates underrepresented observations from unprivileged subgroups
  - Removes overrepresented observations from privileged subgroups

- Reweight – focuses on mitigating STP
  - Computes weights by dividing theoretical probability of assigning favorable label for subgroup by observed probability (based in data).

- Reject Option based Classification Pivot
  - Pivots the probabilities close to cutoff to its other side.



How to do it in fairmodels?

More fairness materials

- Landing page fairmodels.drwhy.ai
  - Article
  - Blogs
  - Documentation
  - Tutorials
  - GitHub
- fairmodels in Python as dalex module

Thank you for attention!

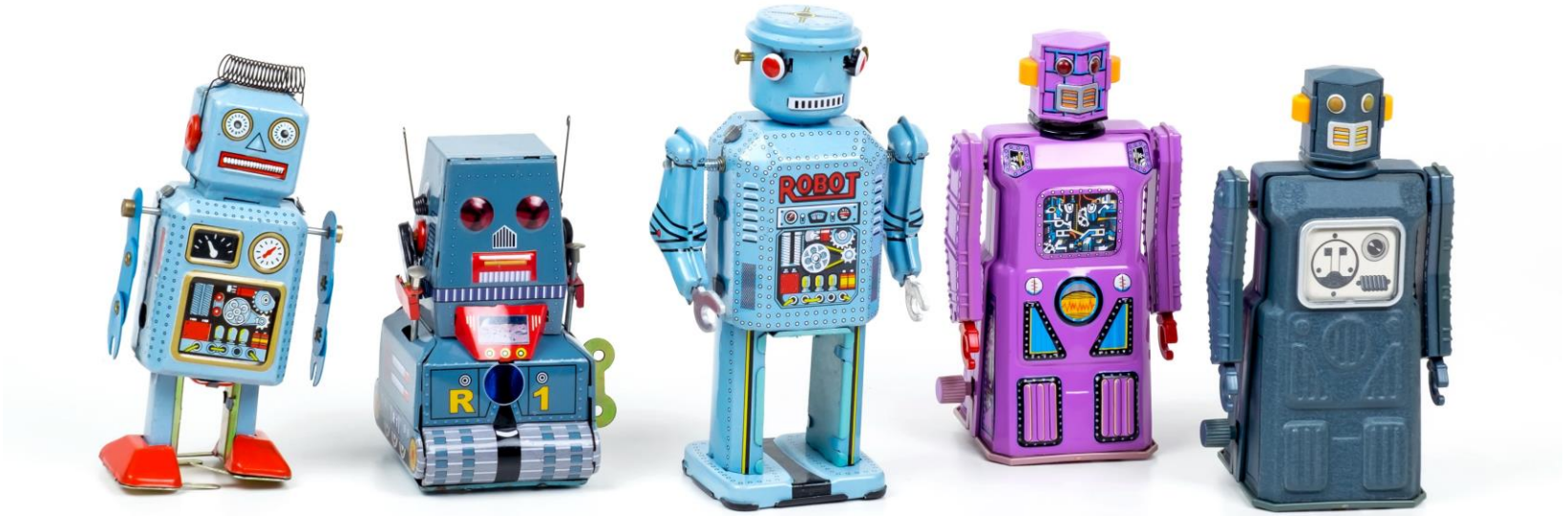


Photo by Eric Krull on Unsplash