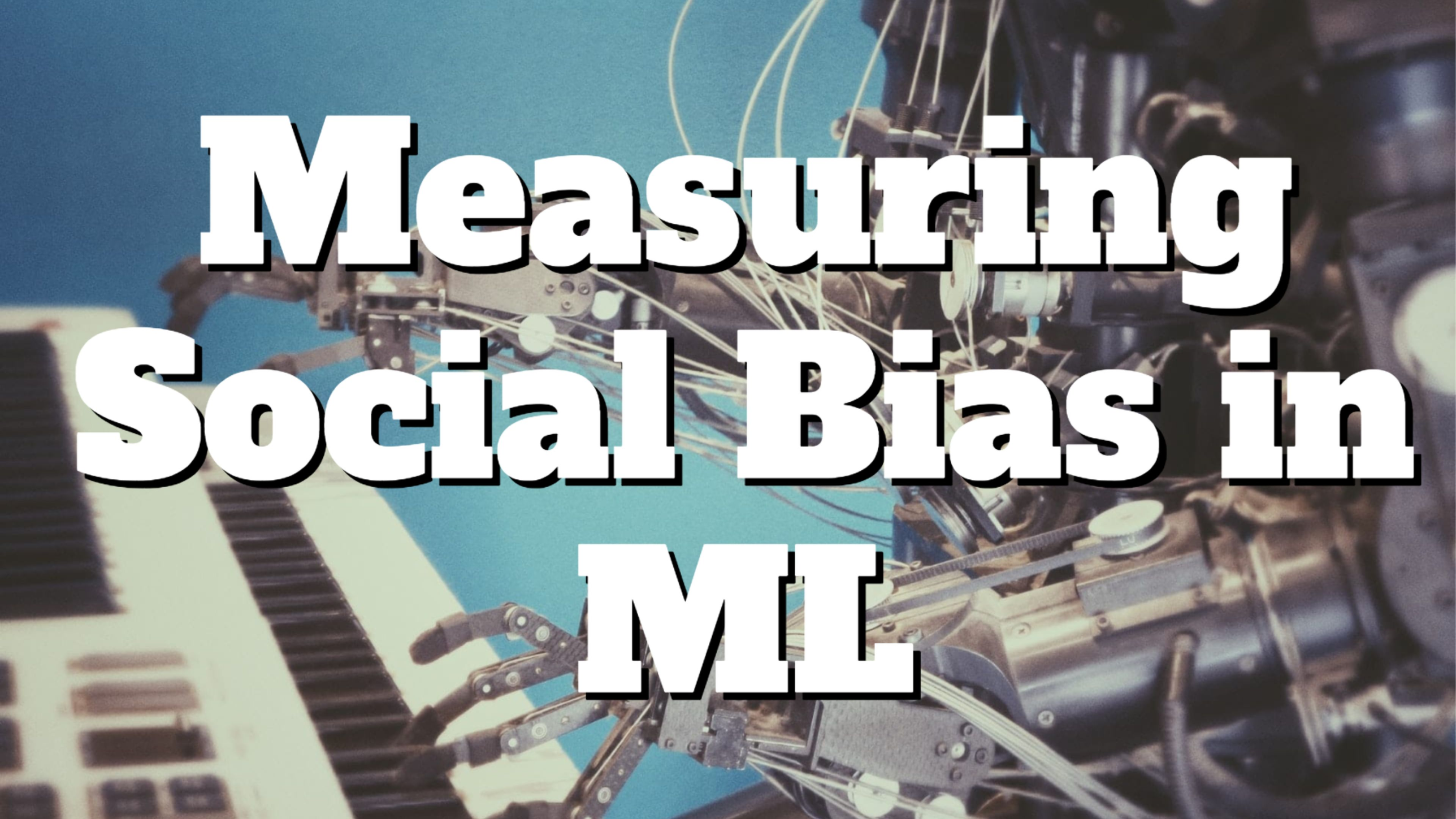
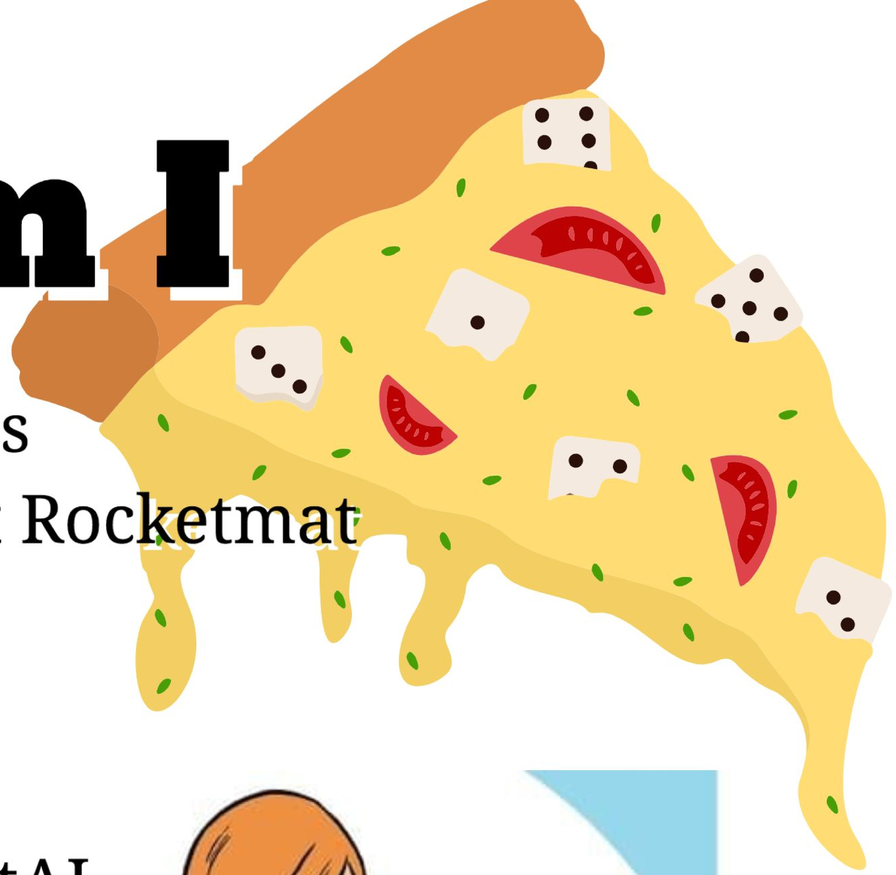


Measuring Social Bias in ML

SENORGIF.CO
Gustavo Coelho

Who am I



Undergrad in economics
Data Scientist Leader at Rocketmat
Pizza de Dados
Calango Hacker Clube
Pydata Brasília
@gusrabbit @rocketmatAI

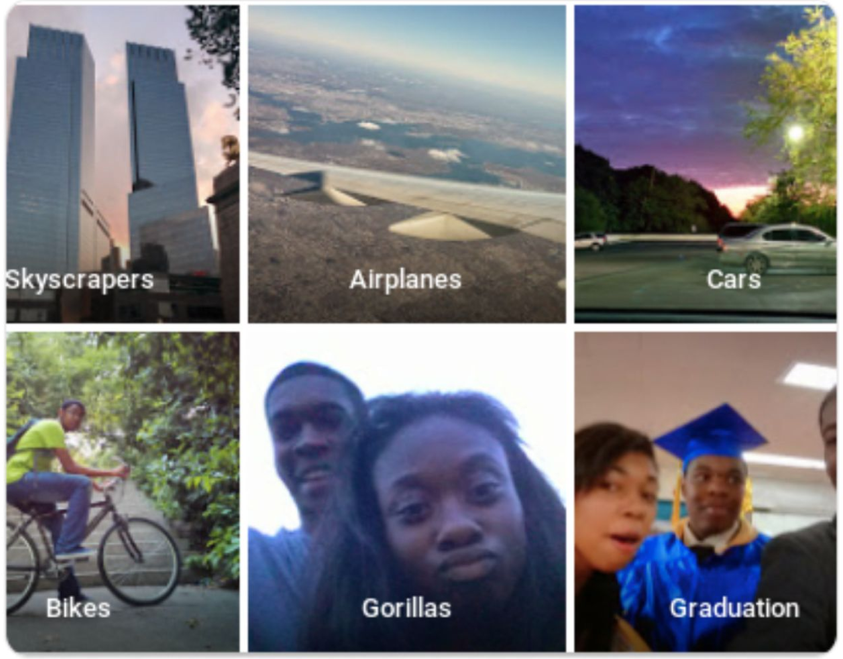
rocketmat[🚀]

we are reshaping the way HR selects candidates using artificial intelligence

Jacky lives on @jalcine@playviciou...
@jackyalcine

Follow

Google Photos, y'all f ed up. My friend's not a gorilla.



Social Bias in AI

Propublica vc Northpoint (Recindivism prediction)

Racial Discrimination in Auto Insurance Premiums

Algorithmic Justice League, Buolamwini, Code4rights

Weapons of Math Destruction, Cathy O’Neil

Trade-off between fairness and eficiency

Not enough to remove sensitive features

“It’s actually impossible for a risk score to satisfy both fairness criteria at the same time."

Sam Corbett-Davies ,Emma Pierson , Avi Feller and Sharad Goel

WHY NOT BOTH?
memegenerator.net

A Unified Approach to Quantifying Algorithmic Unfairness

Let's use inequality indices!
Can give different weights to different outcomes (FP, FN) depending on the case

Can consider sensitive features whether or not they are fed to algorithms

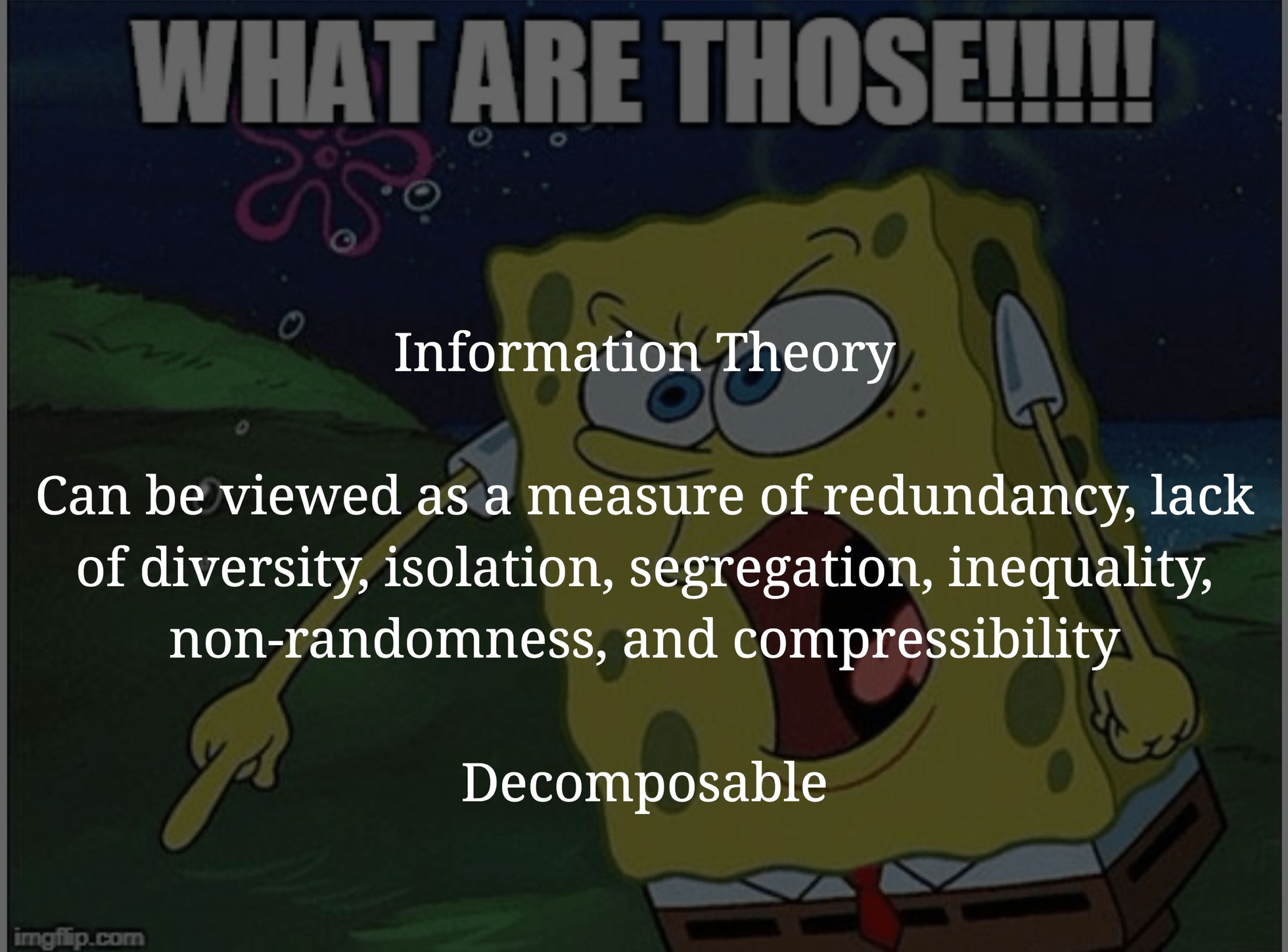
Previous fairness notions can be represented as especial cases

Till Speicher[1,*], Hoda Heidari[2,*], Nina Grgic-Hlaca[1],
Krishna P. Gummadi[1], Adish Singla[1], Adrian Weller[3,4], Muhammad Bilal Zafar[1]
[1]MPI-SWS, [2]ETH Zurich, [3]University of Cambridge, [4]The Alan Turing Institute

Inequality Indices

Economics has been using it for a long time

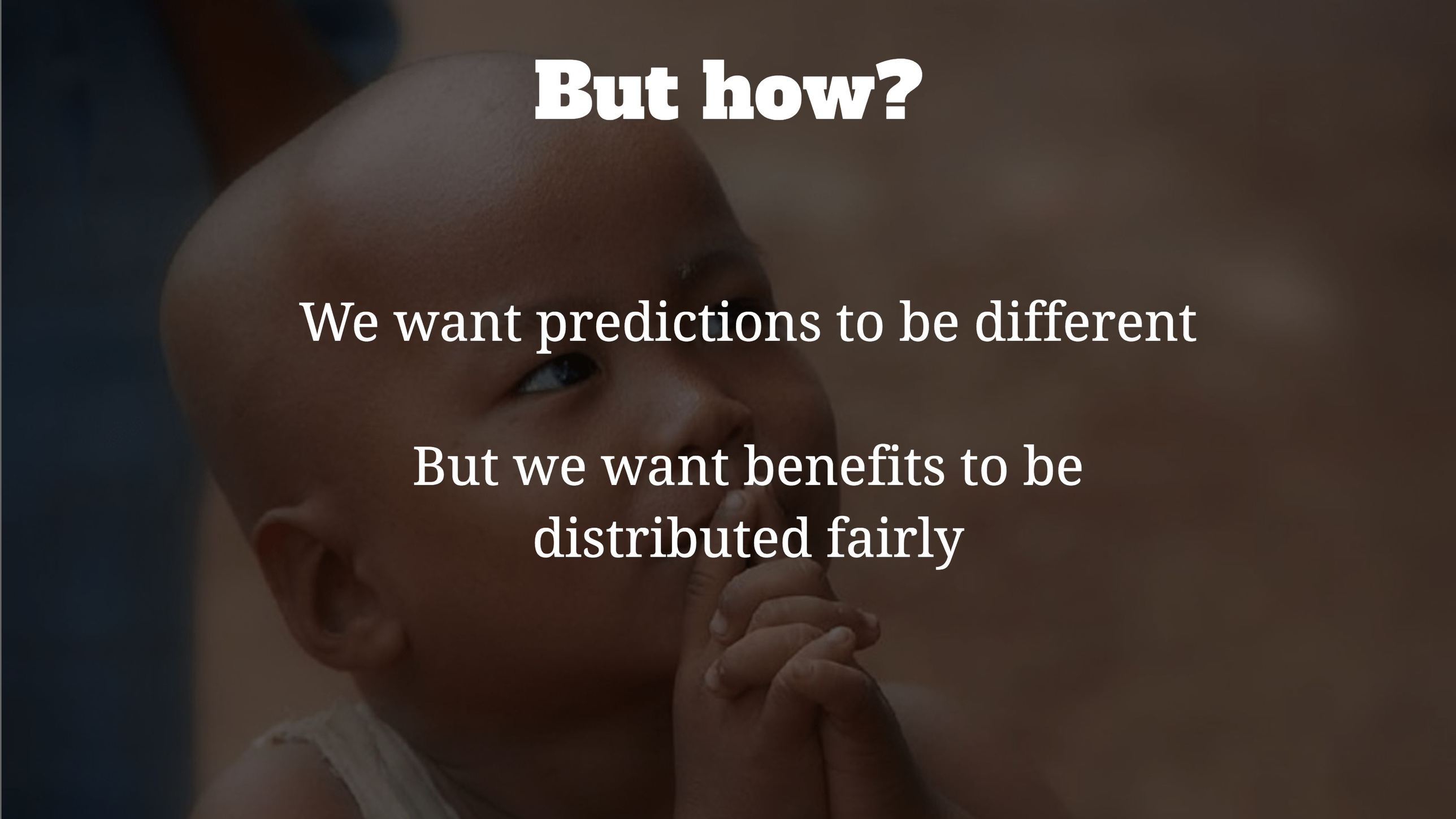
How and why to adapt it?



Information Theory

Can be viewed as a measure of redundancy, lack of diversity, isolation, segregation, inequality, non-randomness, and compressibility

Decomposable



But how?

We want predictions to be different

But we want benefits to be distributed fairly

Let's measure benefits

Benefit should capture the desirability of the predicted outcome for the individual

Basically, let's give weights to TP, FP, FN and TN

Predicting top performers

Difference between what the individual deserves and what they get

FP = 2

TP = 1

TN = 1

FN = 0

Overall Fairness

b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
1	1	1	2	0	0	1	1	1	2
Overall individual-level unfairness $= I(b_1, \ldots, b_{10})$									

We stop thinking about the ranking of predictions

And start measuring how well the benefits are distributed

Individual VS Group fairness

μ_{g_1}	μ_{g_1}	μ_{g_2}	μ_{g_2}	μ_{g_2}	μ_{g_2}	μ_{g_3}	μ_{g_3}	μ_{g_3}	μ_{g_3}
1	1	0.75	0.75	0.75	0.75	1.25	1.25	1.25	1.25
Between-group unfairness $= I(\mu_{g_1}, \mu_{g_1}, \mu_{g_2}, \mu_{g_2}, \mu_{g_2}, \mu_{g_2}, \mu_{g_3}, \mu_{g_3}, \mu_{g_3}, \mu_{g_3})$									

b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
0	2	1	0	1	1	1	1	1	2
Within-group unf. $=I(b_1, b_2)$		**Within-group unf.** $=I(b_3, b_4, b_5, b_6)$				**Within-group unf.** $=I(b_7, b_8, b_9, b_{10})$			

One more thing

Stop thinking in absolutes!

Inequality is never going to be zero or 100

Properties

ax·i·om

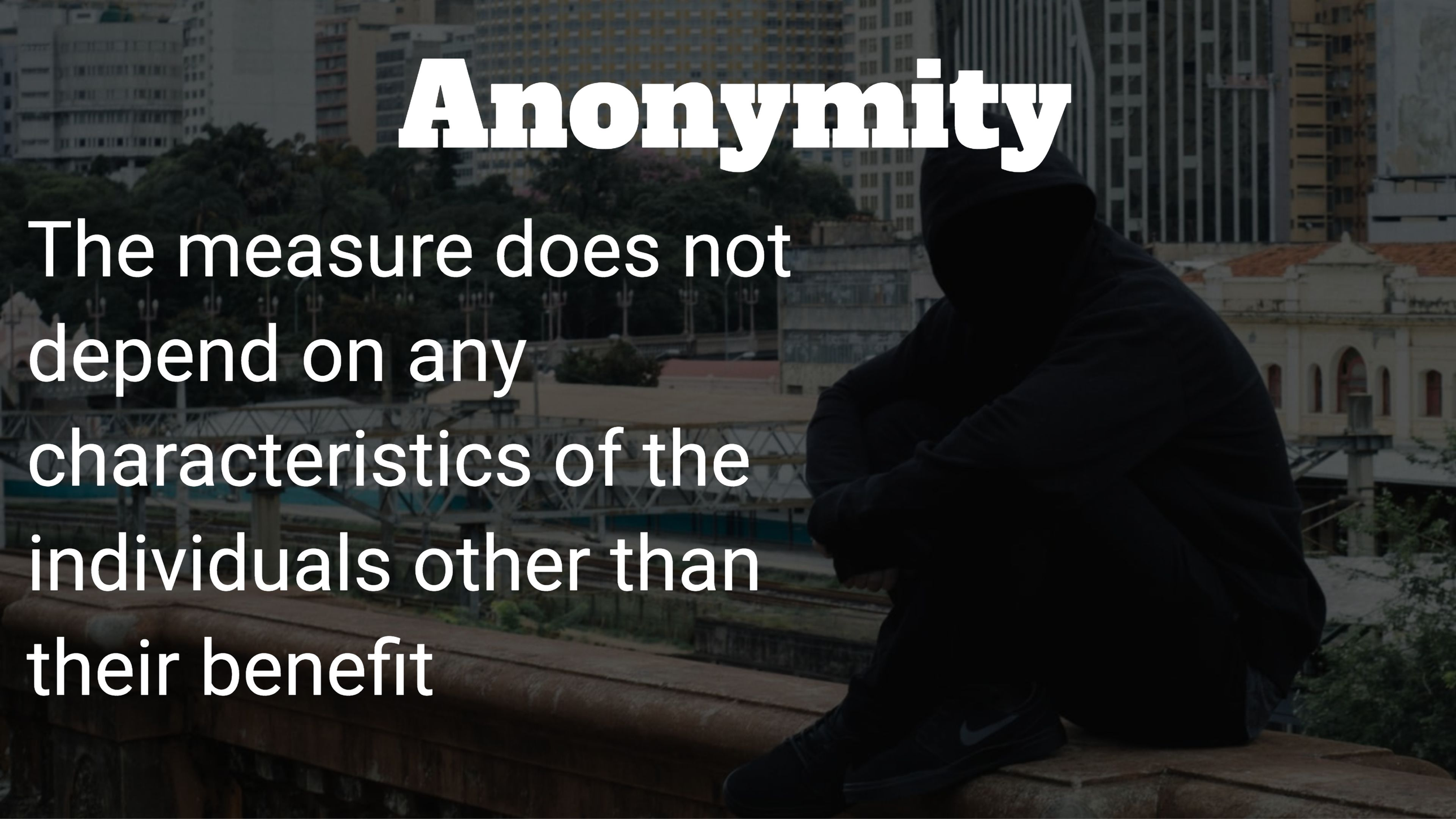
[**ak**-see-*uh*m]

–noun

1. a self-evident truth that requires no proof.
2. a universally accepted principle or rule.

Origin:

1475–85; < L *axiōma* < Gk: something worthy, equiv. to *axiō-*, var. s. of *axioûn* to reckon worthy + *-ma* resultative n. suffix



Anonymity

The measure does not depend on any characteristics of the individuals other than their benefit

Population invariance

The measure is independent of the size of the population under consideration

Transfer Principle

Transfers from a high-benefit individual to a low benefit individual must decrease inequality

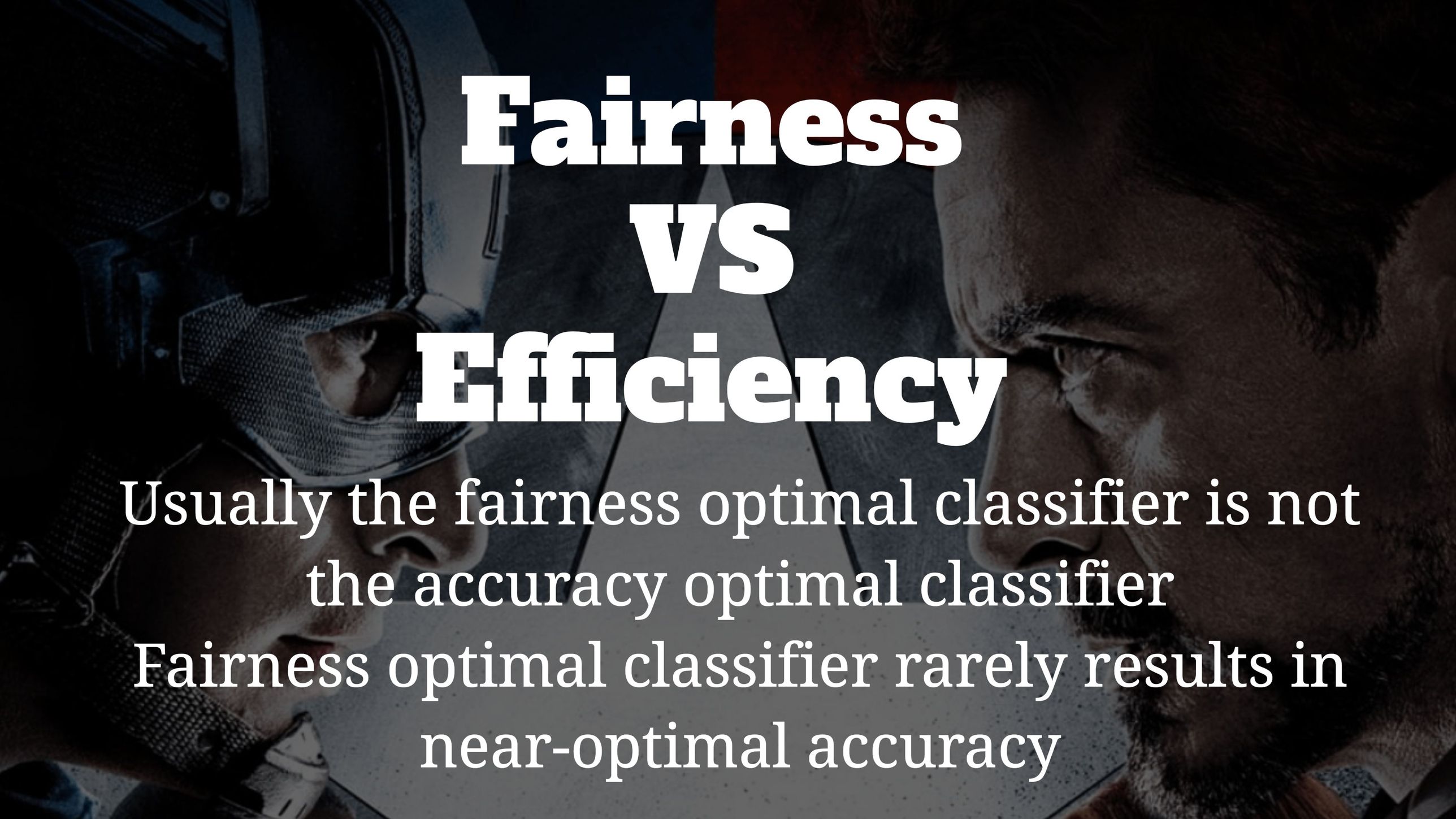
Zero-normalization

The measure is minimized when every individual receives the same level of benefit

Individual
VS
Group fairness

Minimizing between group may drive up within group

It may get an even bigger overall!



Fairness VS Efficiency

Usually the fairness optimal classifier is not the accuracy optimal classifier

Fairness optimal classifier rarely results in near-optimal accuracy

How to combat bias

Pre
During
Post

AI Fairness 360

References

http://code4rights.org/

https://weaponsofmathdestructionbook.com/

https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.8b72300511b7

https://github.com/IBM/AIF360

@gusrabbit

@gusrabbit

gusrabbit.com
pizzadedados.com.br