Biases are bugs

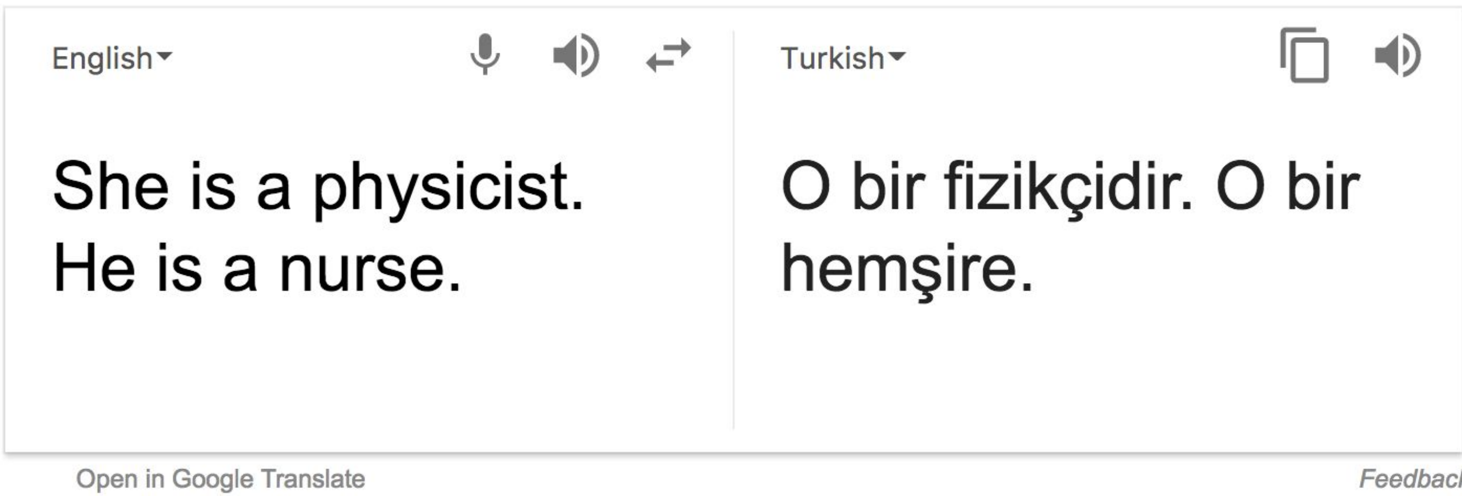
Algorithm fairness and machine learning ethics

Françoise Provencher, PhD
July 2nd, 2017 @ PyData Berlin

shopify

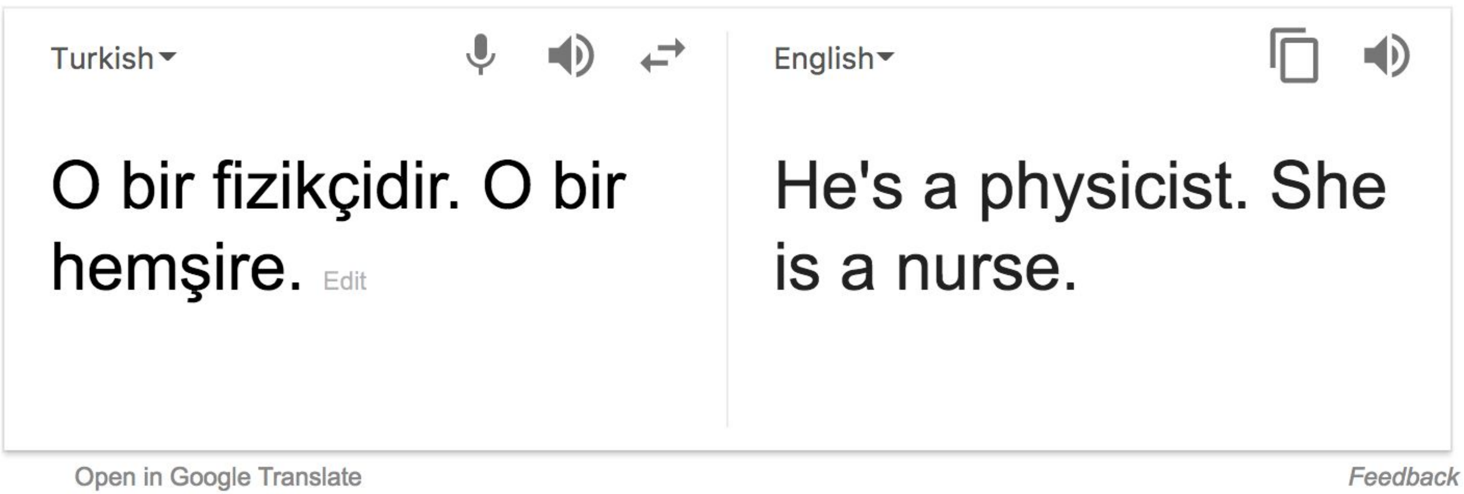
Algorithms learn biases from data.

Example : Google Translate



Turkish is a genderless language

Example : Google Translate



In this talk

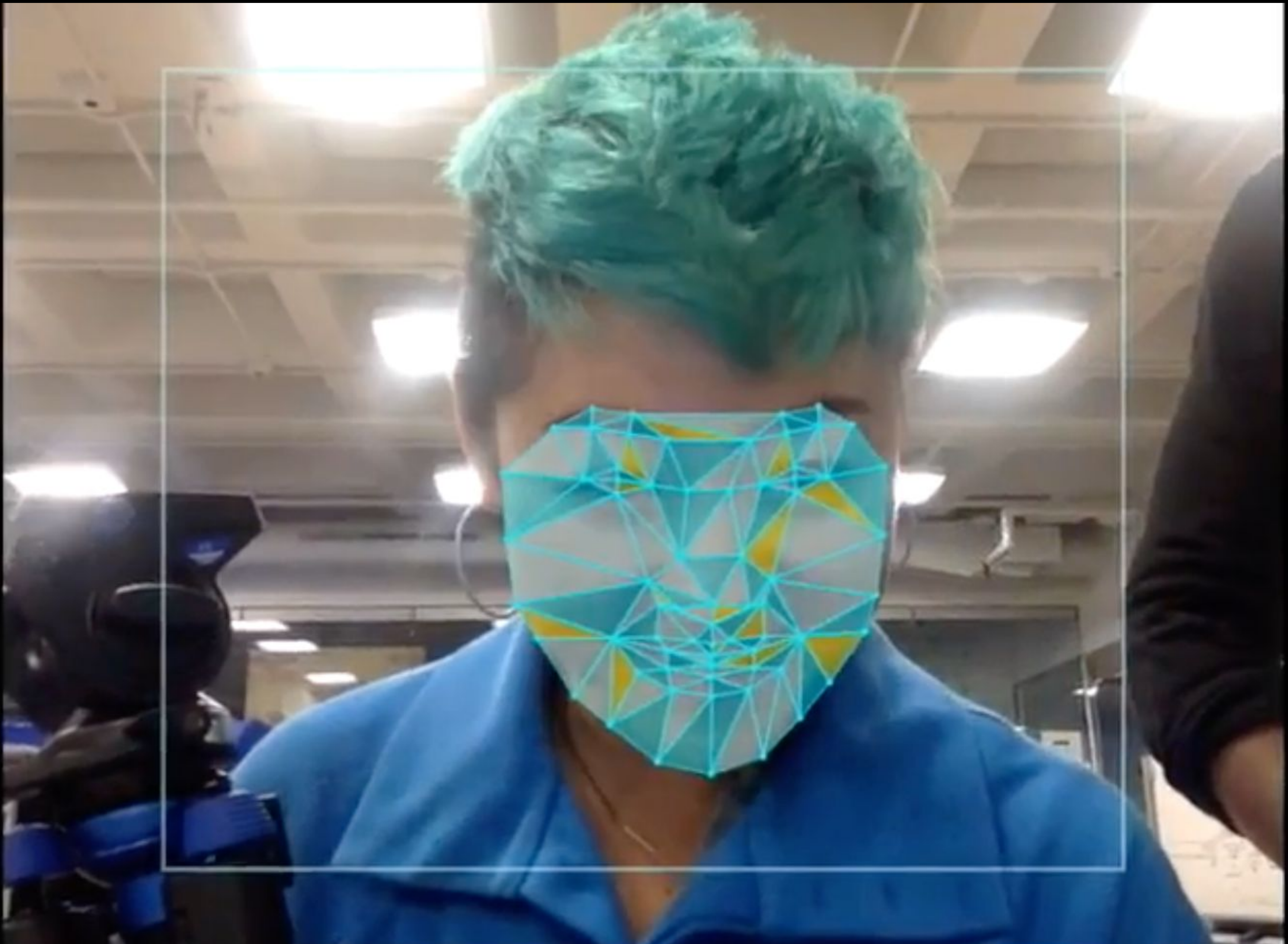
- Detecting biases in algorithms
- Removing biases from algorithms
- Making it work in your organisation

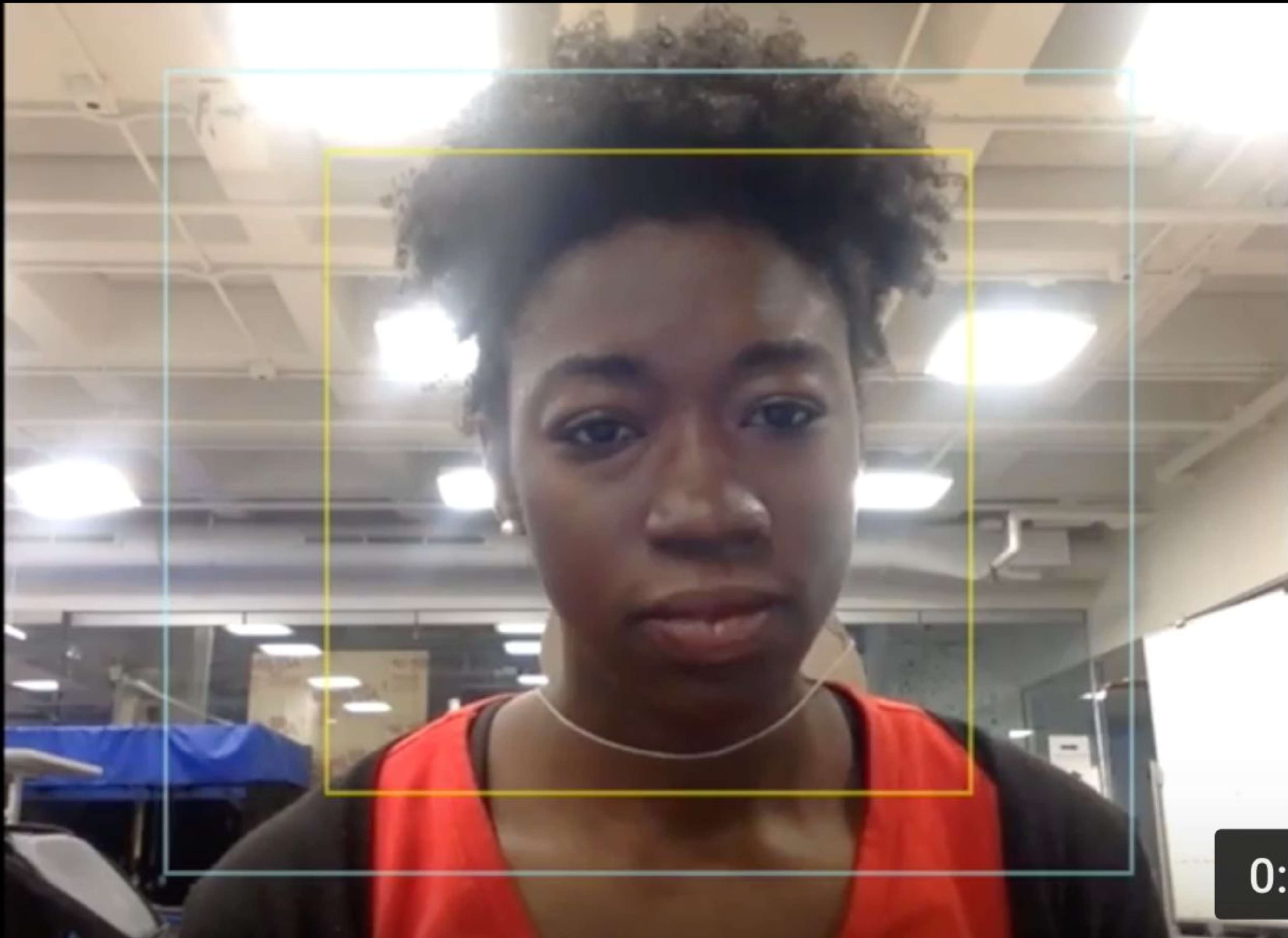
Strategies to detect biases in algorithms

Stress cases

Design for **worst case** scenario.

Eric Meyer & Sara Wachter-Boettcher, “Design for real life”, A book apart, 2016


Joy Buolamwini, Algorithmic Justice League, www.ajlunited.org



Joy Buolamwini, Algorithmic Justice League, www.ajlunited.org

Joy Buolamwini, Algorithmic Justice League, www.ajlunited.org

Stress cases

Pros

- Anticipate problems that are hard to quantify

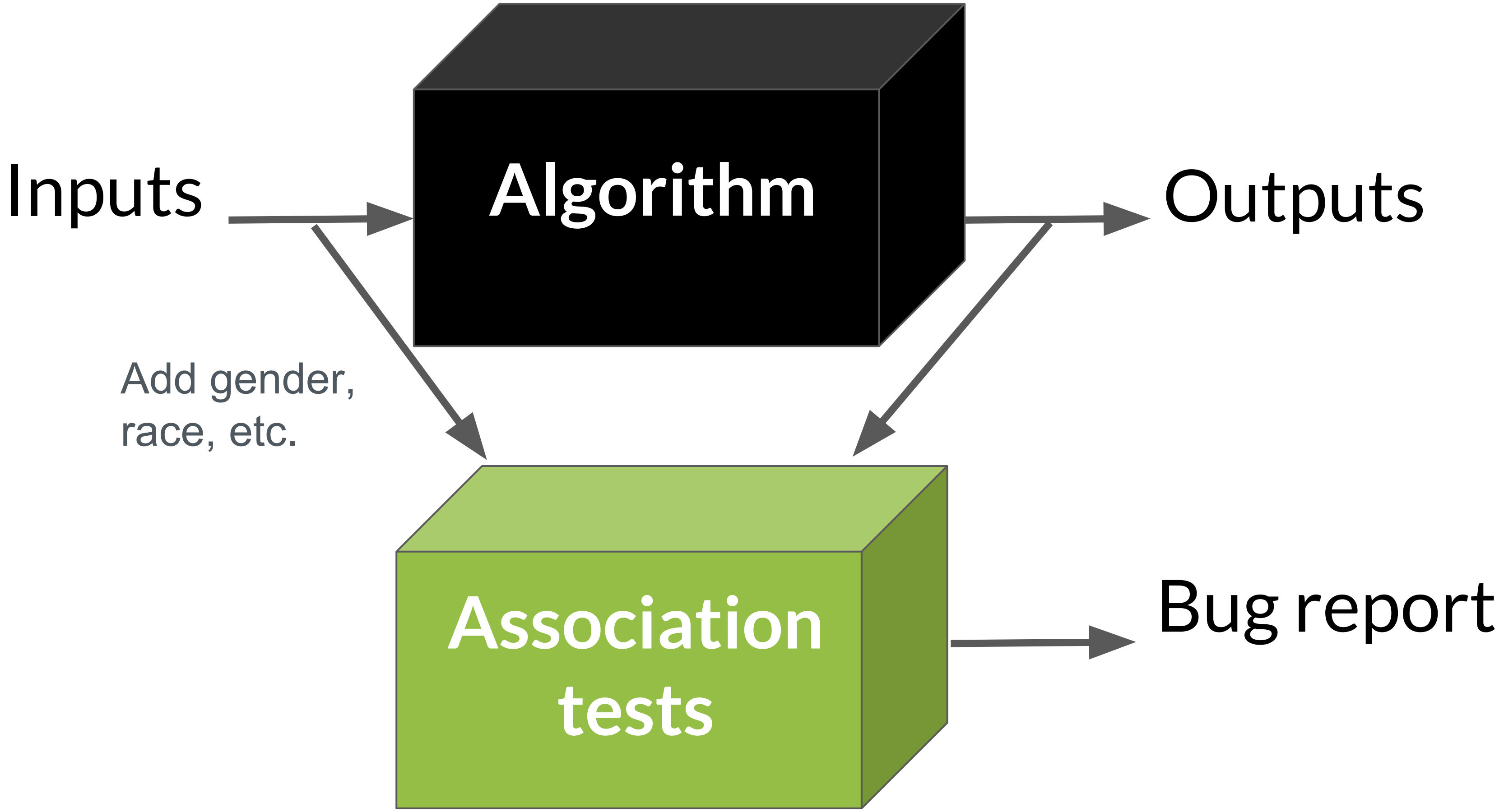
Cons

- Not systematic or scalable

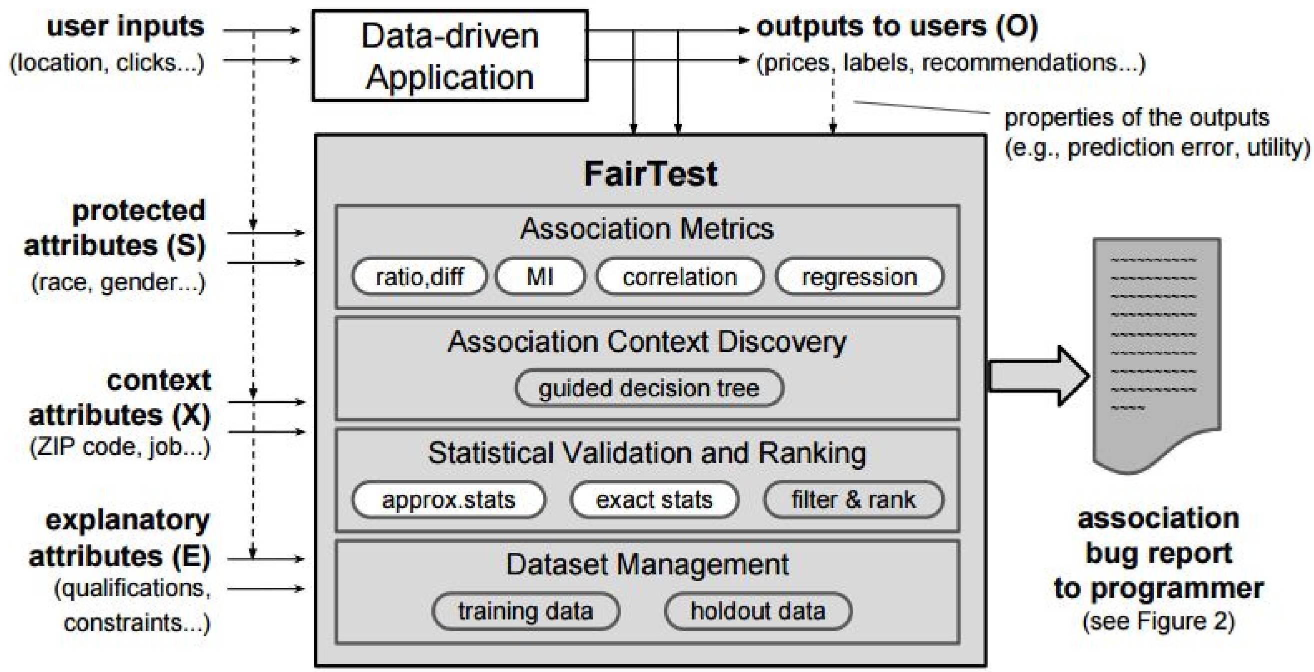
Association test package

Association test package

Association test package



FairTest



Tramers, F et al. (2015) “FairTest: Discovering Unwarranted Associations in Data-Driven Applications”, arXiv preprint arXiv:1510.02377

https://github.com/columbia/fairtest

Association test package

Pros

- Multiple association metrics
- Statistical significance
- Automated workflow

Cons

- Assumes we have the protected class attributes
- No agreed upon thresholds

Disparate impact test

In U.S. law, unintentional bias is encoded via **disparate impact**, which occurs when a selection process has widely different **outcomes** for different groups

Feldman, M. et al. (2015) “Certifying and removing disparate impact”, arXiv:1412.3756v3

Disparate impact test

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

Feldman, M. et al. (2015) “Certifying and removing disparate impact”,
arXiv:1412.3756v3

Disparate impact test

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

Probability of positive outcome should be about the same for members of the majority and the minority of a given protected class attribute.

Feldman, M. et al. (2015) “Certifying and removing disparate impact”, arXiv:1412.3756v3

Disparate impact test

Outcome	$X = 0$	$X = 1$
$C = \text{NO}$	a	b
$C = \text{YES}$	c	d

Tab. 1: A confusion matrix

The 80% rule can then be quantified as:

$$\frac{c/(a+c)}{d/(b+d)} \geq 0.8$$

Feldman, M. et al. (2015) “Certifying and removing disparate impact”,
arXiv:1412.3756v3

Disparate impact test

Pros

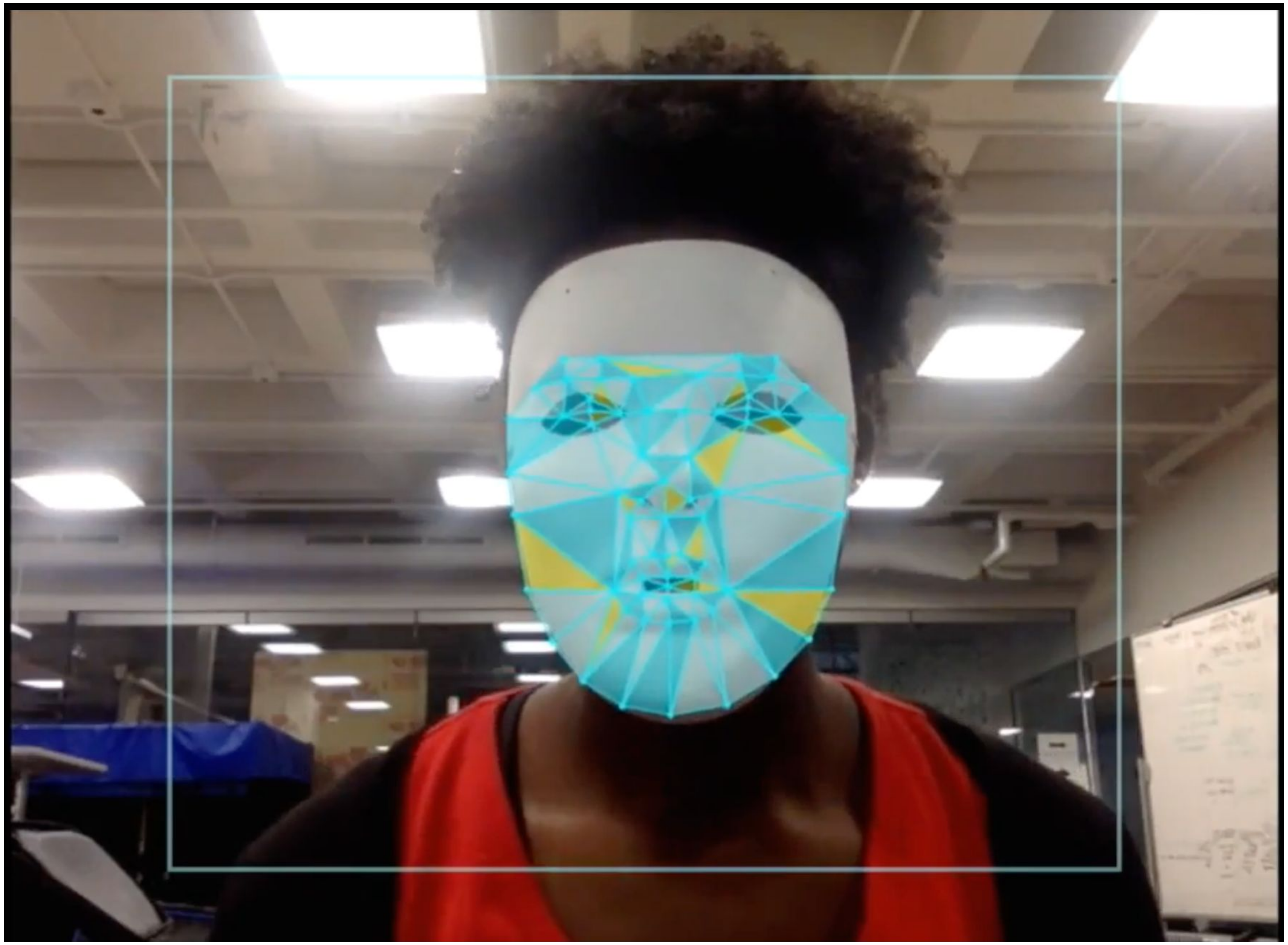
- Straightforward to compute
- Based on agreed upon concepts (law)

Cons

- Does not explicitly take into account statistical significance
- Assumes we have the protected class attributes

Strategies to remove biases from algorithms

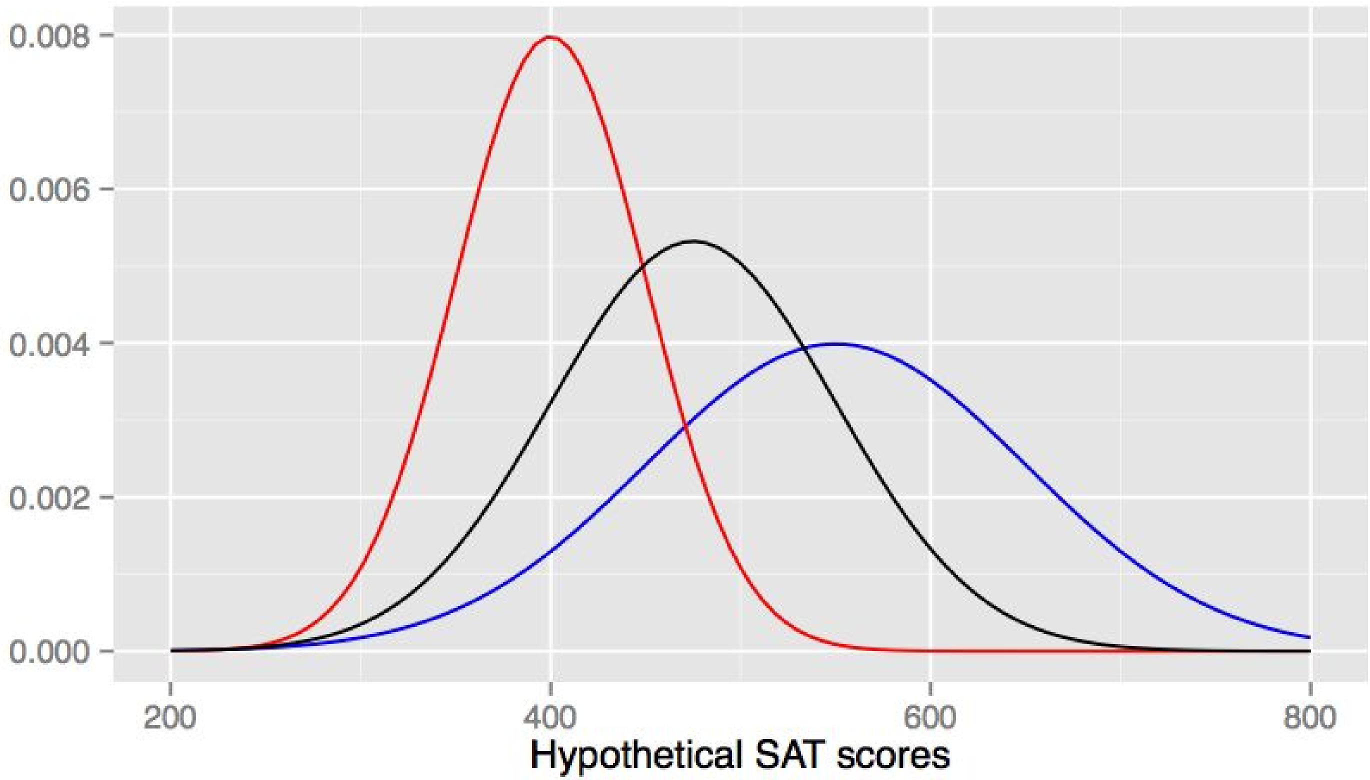
Get more data



Tweak the features

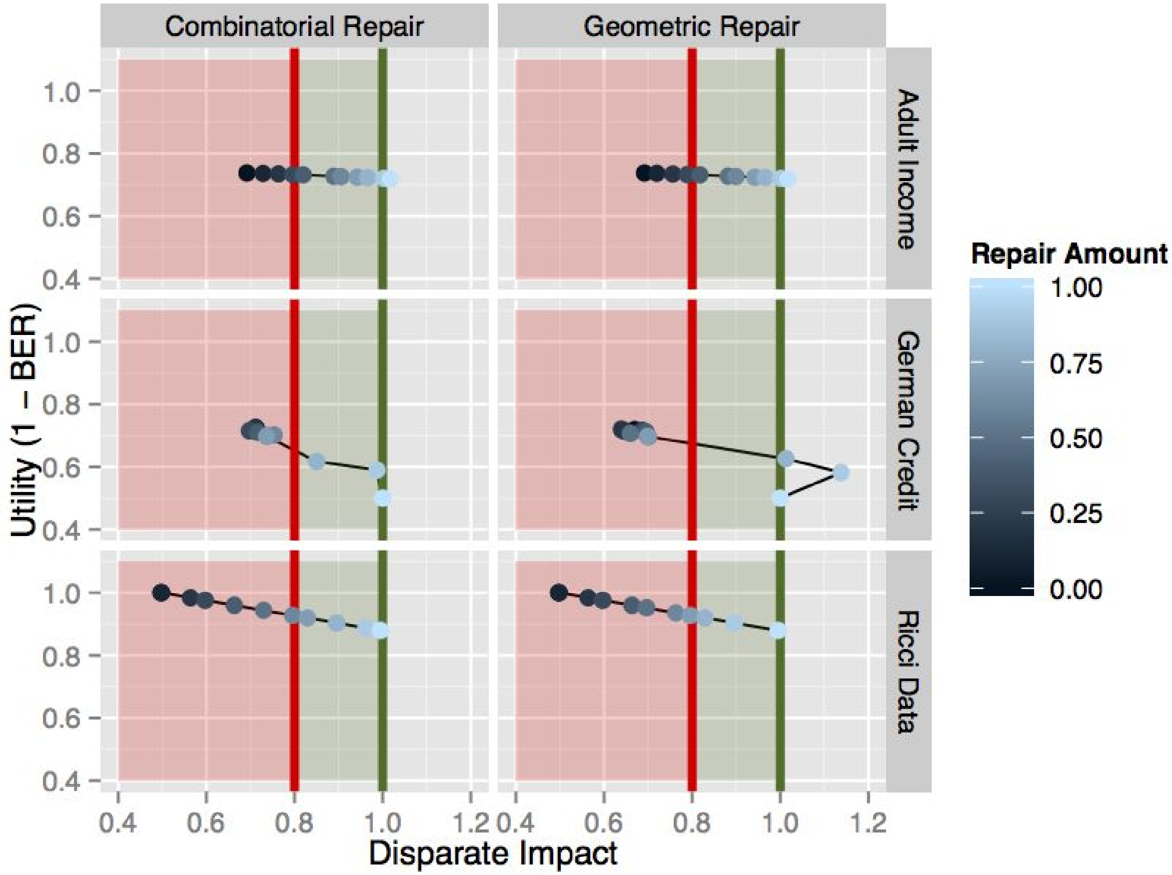
Include or exclude features until the ML algorithm behaves in the way you want.

Dataset repair:
Re-mapping distributions



Feldman, M. et al. (2015) “Certifying and removing disparate impact”, arXiv:1412.3756v3

Dataset repair: Tradeoff between DI and utility



Feldman, M. et al. (2015) “Certifying and removing disparate impact”, arXiv:1412.3756v3

Dataset repair: word2vec

Man	Woman
King	Queen
Lion	
Surgeon	
Computer Programmer	

Bolukbasi, T. et al. (2016) “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, arXiv:1607.06520v1

Dataset repair: word2vec

Man	Woman
King	Queen
Lion	Lioness
Surgeon	
Computer Programmer	

Bolukbasi, T. et al. (2016) “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, arXiv:1607.06520v1

Dataset repair: word2vec

Man	Woman
King	Queen
Lion	Lioness
Surgeon	Nurse
Computer Programmer	

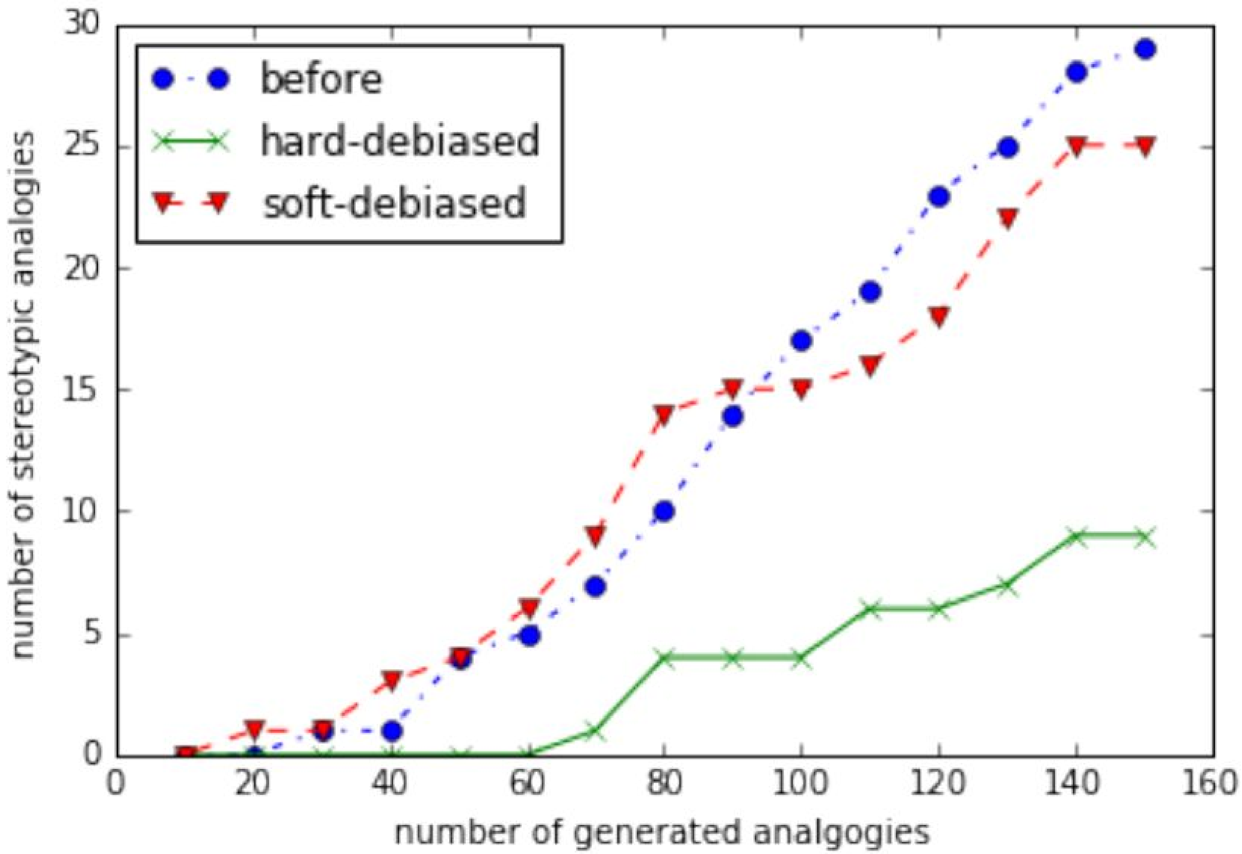
Bolukbasi, T. et al. (2016) “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, arXiv:1607.06520v1

Dataset repair: word2vec

Man	Woman
King	Queen
Lion	Lioness
Surgeon	Nurse
Computer Programmer	Homemaker

Bolukbasi, T. et al. (2016) “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, arXiv:1607.06520v1

Dataset repair: word2vec



Bolukbasi, T. et al. (2016) “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, arXiv:1607.06520v1


Strategies to make it
work in your organization

Leverage existing culture

- Unconscious bias (hiring)
- Code of conduct (legal)
- QA & fixing bugs (engineering)

Create a ML code-of-conduct

- Assemble a diverse team
- Craft a proposal and make it visible
- Use it to audit an algorithm
- Repeat

The Shopify data + ML CoC

- Is it fair?
- Is it legal?
- Would it be OK if it were publicized?
- What is the most horrible case than can happen? What am I doing to avoid this?
- Is the person I'm collecting data from directly benefiting from it?
- Am I mistaking correlation for causation?

The Shopify data + ML CoC

- Auditability
- Algorithm fairness
- Continuous improvements
- Interpretability

Get everyone on board

- Other data scientists
- Communicate to whole org
- Get execs buy-in

“If we can tweak a dial between the tradeoff of fairness and profit, turn it strongly towards fairness.”

- Tobias Lütke, CEO

Biases are bugs.
They need to be found,
fixed, and learnt from.