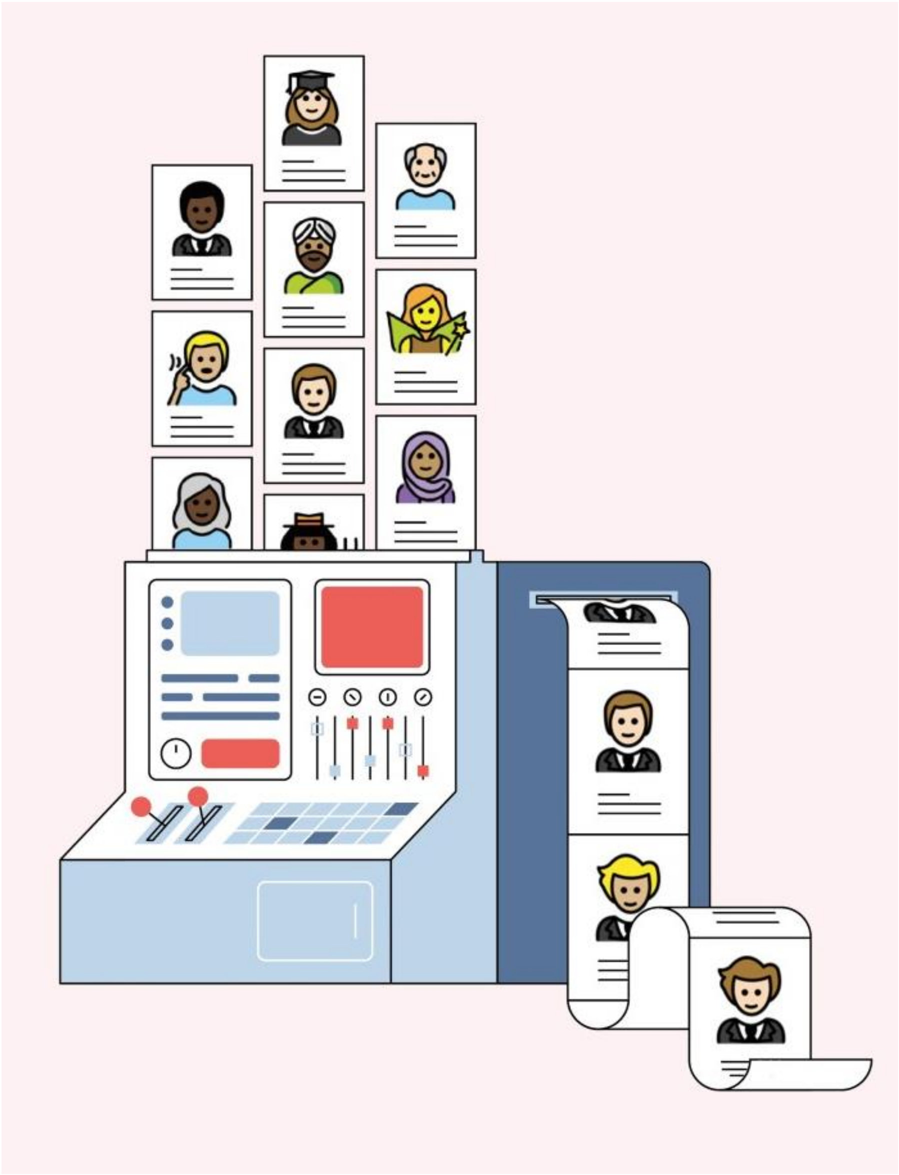
How to be fair?
A study of label and selection bias

Marco Favier, Sam Pinxteren, Jonathan Meyer and Toon Calders

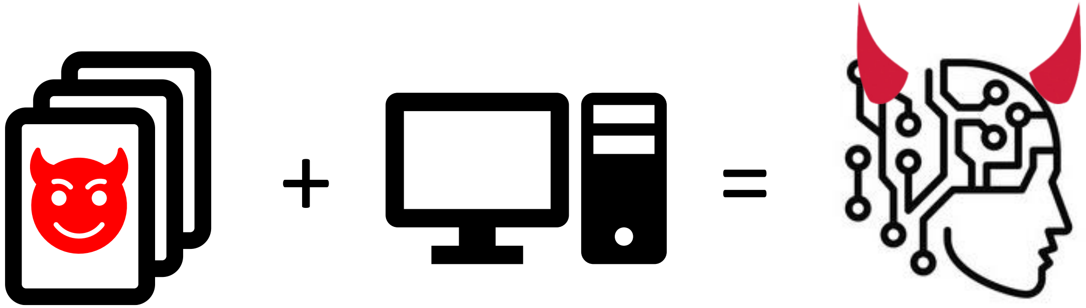
University
of Antwerp



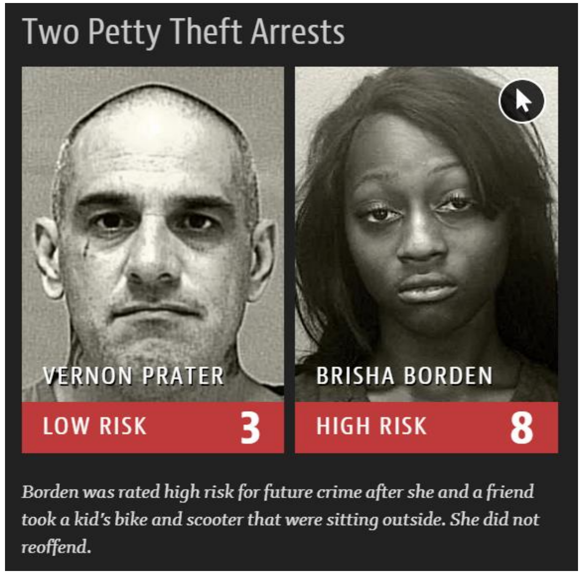
Data Bias

What happens when data do not represent reality or are biased?

Biased data lead to biased models



COMPAS



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

- **Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool to predict risk of recidivism**
  - Label: was there a new arrest within two years?
  - Data: pending charges, prior arrest history, previous pretrial failure, residential stability, substance abuse, …

ProPublica Study (2016)

- ProPublica study showed that the errors made by the model are highly biased:

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.)

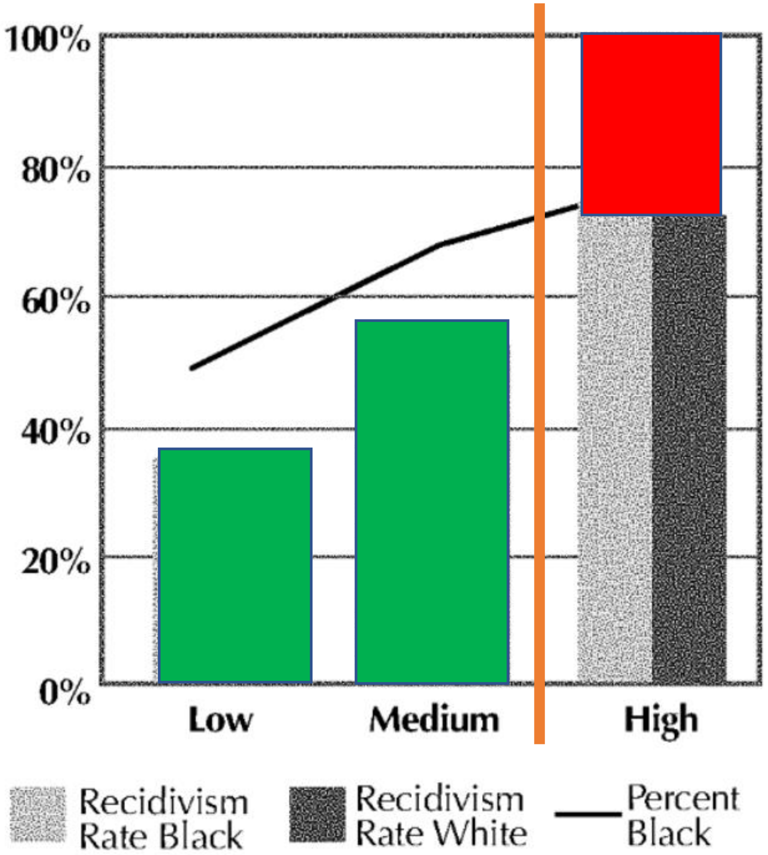
Fair or Unfair?

- Northpointe's defence:
  - scores are *calibrated*

- All false positives are in High risk
- All false negatives in other groups

- Black is relatively more frequent in High than in Low and Medium

FIGURE 3.
Recidivism Rates by Race and Percent Black in Each Risk Category—Any Arrest



Anthony W. Flores et al., False Positives, False Negatives, and False Analyses., 80 Fed. Probation 38 (2016)

Definition of Fairness

- Different definitions concentrate on other aspects of fairness
- We will use the following naming conventions:
  - **X** independent variables
  - A Sensitive attribute (Age, gender, ethnicity, …)
    0 = “protected” group ; 1 = “privileged” group
  - Y dependent variable; target to predict
    0 = undesirable label ; 1 = desirable label
  - $\hat{Y}$ predicted label

Definition of Fairness

- Different definitions concentrate on other aspects of fairness
  - Disparate impact: P($\hat{Y}$=1 |A=0) vs P($\hat{Y}$=1 |A=1)
  - Equal opportunity: P($\hat{Y}$=1 | Y=1, A=0) vs P($\hat{Y}$=1 | Y=1, A=1)
  - Calibration: P(Y=1 | $\hat{Y}$=1, A=0) vs P(Y=1 | $\hat{Y}$=1, A=1)
- And they all make sense

- Which one should I pick ? Situation dependent!
- So let’s look at a concrete situation

Pro Publica: *Equal opportunity*

- If you deserve to stay in prison, it shouldn't matter whether you're black or white *P(High | reoffender)*

- If you deserve to be released, it shouldn't matter whether you're black or white *P(Low | no reoffender)*

$$\mathrm{P}[\, \hat{Y} = 1 \mid Y = 1, A = 0 \,] = \mathrm{P}[\, \hat{Y} = 1 \mid Y = 1, A = 1 \,]$$
$$\mathrm{P}[\, \hat{Y} = 1 \mid Y = 0, A = 0 \,] = \mathrm{P}[\, \hat{Y} = 1 \mid Y = 0, A = 1 \,]$$

Northpointe: *Calibration*

- What it means to be a high/low risk should not depend on your ethnicity *P(reoffend | High), P(reoffend | Low)*

$$P[\,Y = 1 \mid \hat{Y} = 1\,, A = 0\,] = P[\,Y = 1 \mid \hat{Y} = 1\,, A = 1\,]$$
$$P[\,Y = 1 \mid \hat{Y} = 0\,, A = 0\,] = P[\,Y = 1 \mid \hat{Y} = 0\,, A = 1\,]$$

Comparison calibration – equal odds

- Equal odds:
  - Errors should affect all groups in the same proportion
- Calibration:
  - Assigned labels should have the same interpretation for all groups

- These criteria seem quite similar in nature; maybe they can be combined?

Illustration: Calibrated

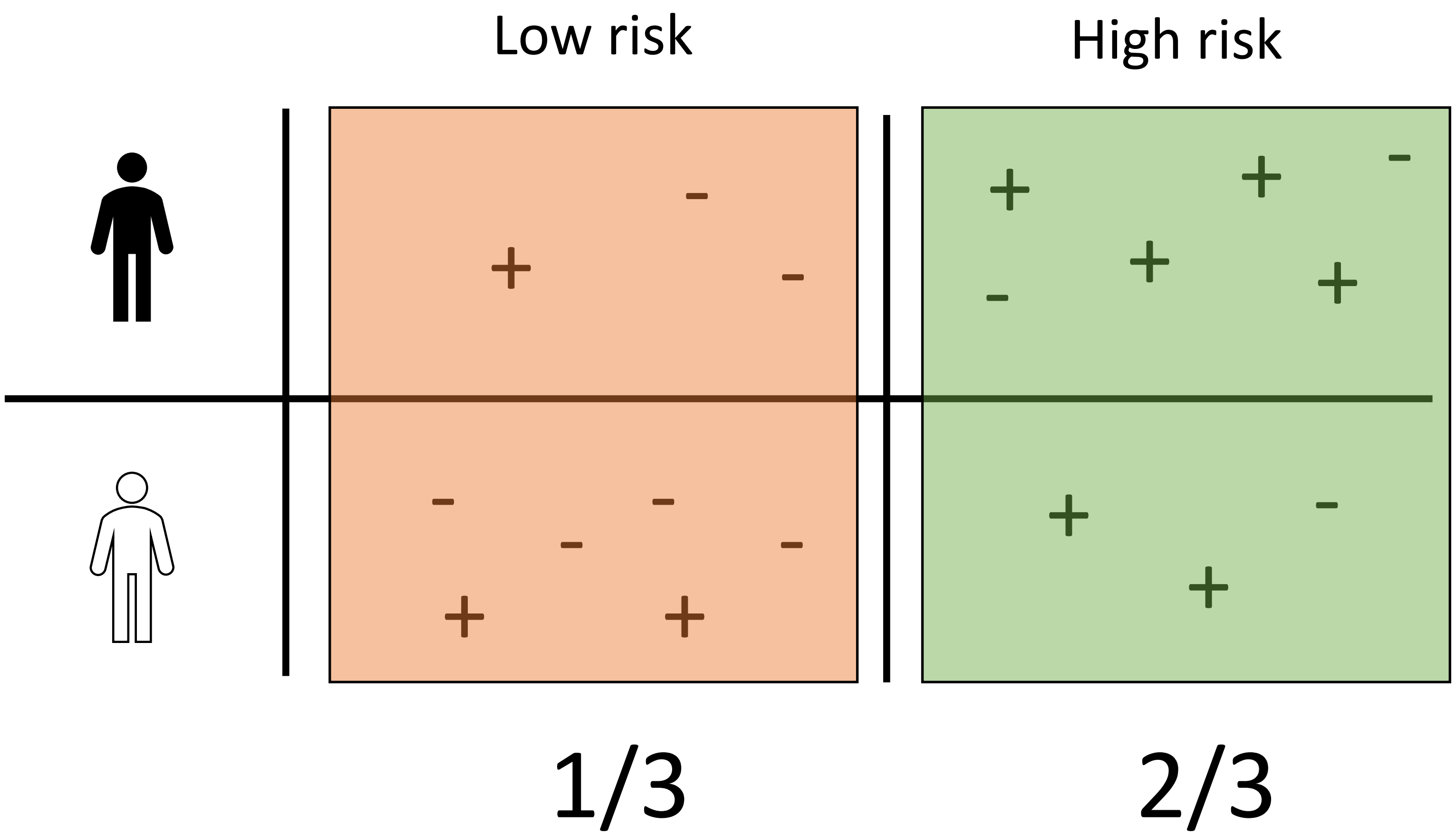
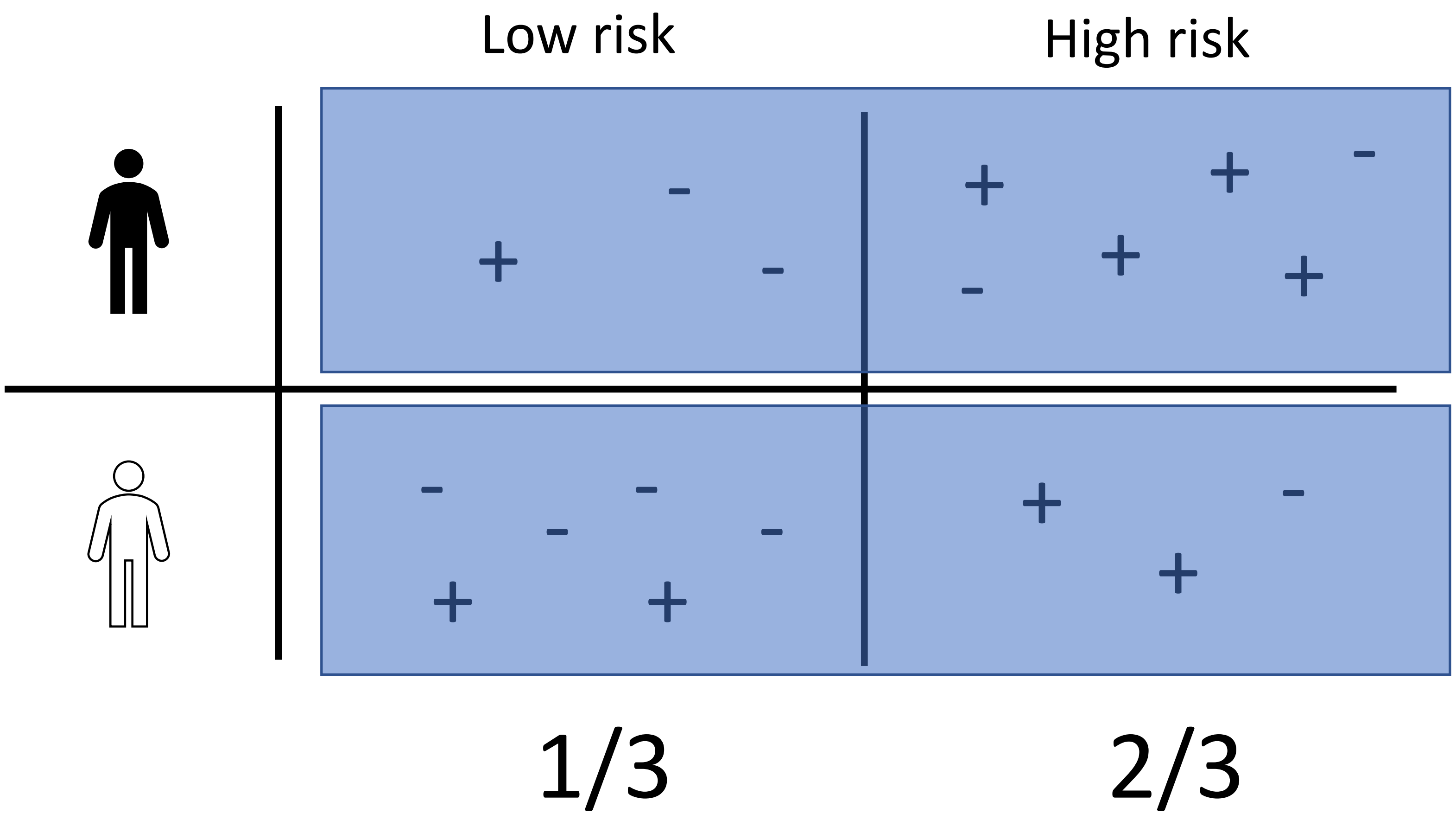


Illustration: But Not Equal Opportunity



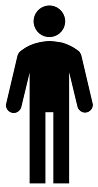
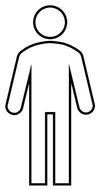
$P(H|+) = 4/5$

$P(H|-) = 1/2$

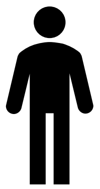
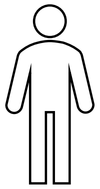
$P(H|+) = 1/2$

$P(H|-) = 1/5$

Let’s Satisfy Equal Odds

	Low risk	High risk	
	$P_1(1\text{-}x)$ + $N_1(1\text{-}y)$ -	$P_1\ x$ + $N_1\ y$ -	$P(H \mid +) = x$ $P(H \mid -) = y$
	$P_2(1\text{-}x)$ + $N_2(1\text{-}y)$ -	$P_2\ x$ + $N_2\ y$ -	$P(H \mid +) = x$ $P(H \mid -) = y$

What About Calibration ?

	Low risk	High risk	
	$\frac{P_1(1-x)+}{N_1(1-y)-}$	$\frac{P_1\,x+}{N_1\,y-}$	$P(H\|+) = x$ $P(H\|-) = y$
	$\frac{P_2(1-x)+}{N_2(1-y)-}$	$\frac{P_2\,x+}{N_2\,y-}$	$P(H\|+) = x$ $P(H\|-) = y$

(P1/N1=P2/N2) or (x=1 and y=0) or (x=0 and y=1)

Theorem (Kleinberg et al. 2016)

Let S be a score function mapping instances $\mathbf{X}$,A to the interval [0,1]

If S satisfies:

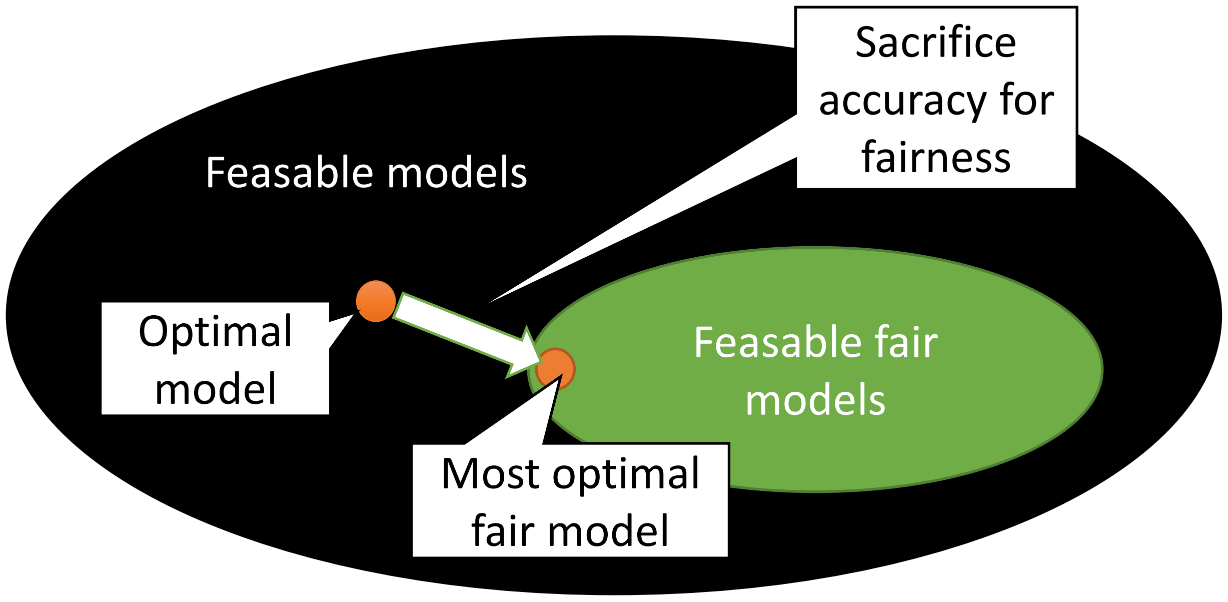
- *Calibration:* $E[\, Y \mid A=a, S(\mathbf{X}) = s \,] = s$ for all a, s
- *Equal odds:* $E[\, S(\mathbf{X}) \mid Y=y, A=0 \,] = E[\, S(\mathbf{X}) \mid Y=y, A=1 \,]$ for all y

Then, one of the following two holds:

- *Perfect predictability:* For all $\mathbf{X}$, $S(\mathbf{X}) \in \{ 0, 1 \}$
- *Equal base rates:* $E[\, Y=1 \mid A=0 \,] = E[\, Y=1 \mid A=1 \,]$

Common Approach to Fairness

- Select a fairness measure (how?)
- Optimize the fairness measure while keeping accuracy as high as possible



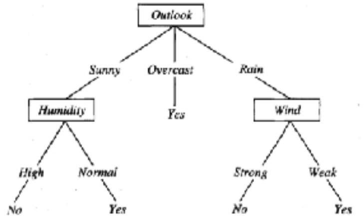
Fairness in ML

Different metrics to measure if the model is biased.	E.G. Statistical Parity, Equal Opportunity, Calibration etc...
Different techniques to minimize those measures	E.G. pre/in/post-processing techniques.
How to be fair?	Select a fairness metric Minimize the metric using a suitable technique Claim the model is now unbiased

The “Old Way”: Fairness by design

Learning + fairness intervention;
e.g., reweighing

Observed data



Bias detected by measure;
e.g., demographic parity (DP)

Model that is *“fair by design”*;
e.g., predictions satisfy DP

Accuracy – Fairness Trade-Off

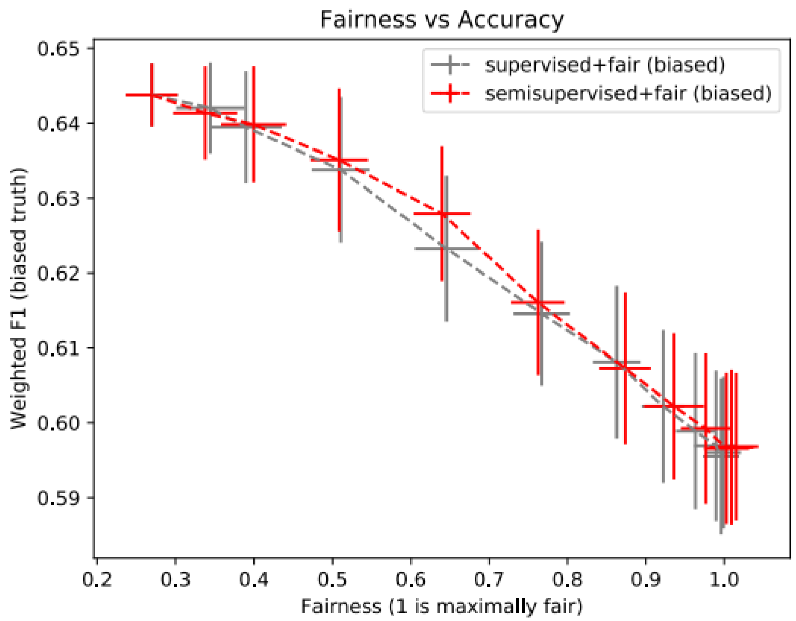
- Common assumption:
  Most accurate *fair* model is less accurate than most accurate model overall

- However: in case of label-bias, fair models may be *more accurate* than unfair models
  - Do not optimize accuracy on *tainted training data*, but optimize expected accuracy for *fair test data*

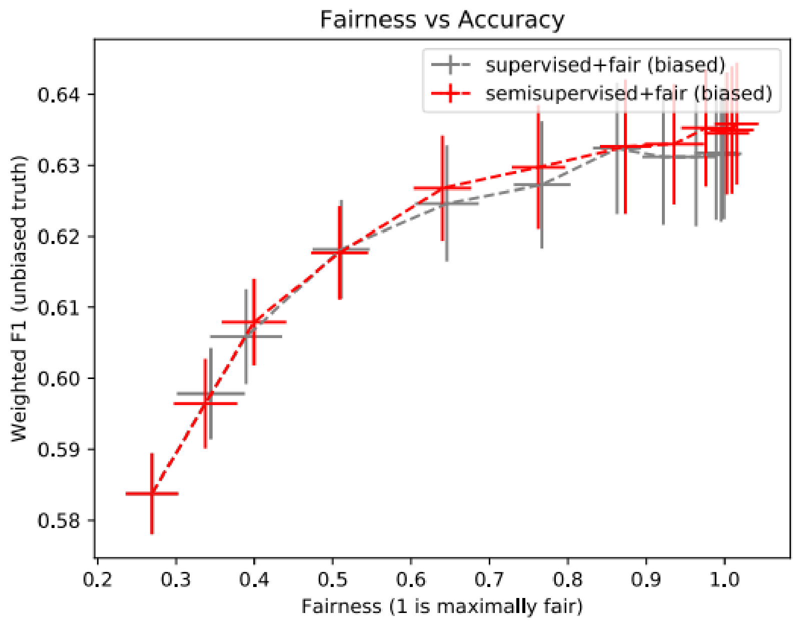
Wick, Michael, Swetasudha Panda, and Jean-Baptiste Tristan (2019) "Unlocking fairness: a trade-off revisited.“ Advances in Neural Information Processing Systems 32

Unlocking fairness: a trade-off revisited

- Synthetically generate data with label bias
- Test on ground truth

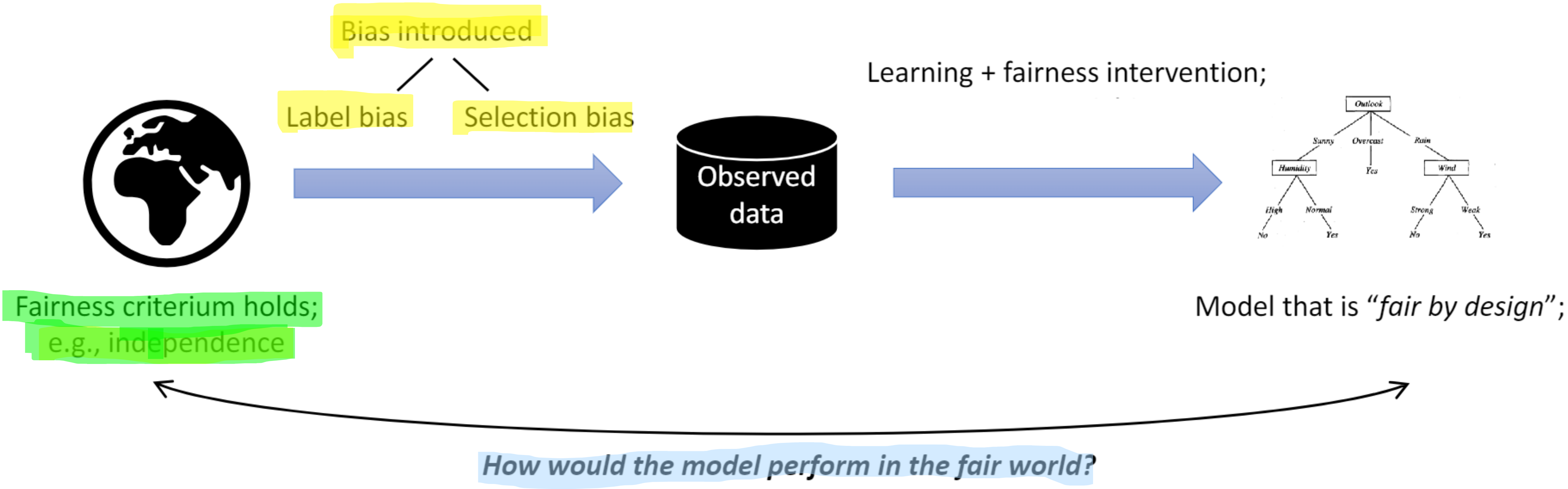


(a) COMPAS (biased ground truth)



(b) COMPAS (unbiased ground truth)

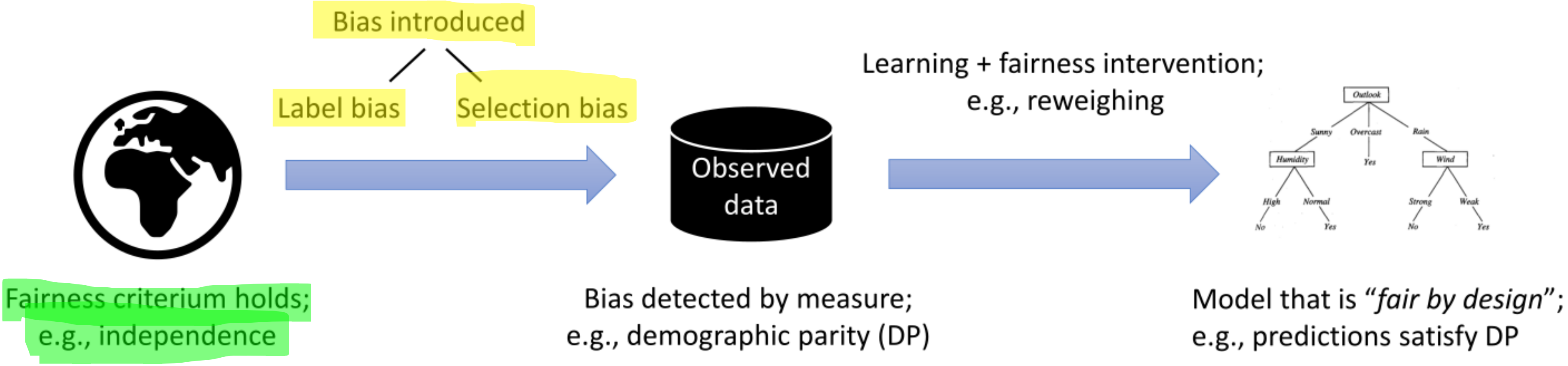
A New Way to be Fair



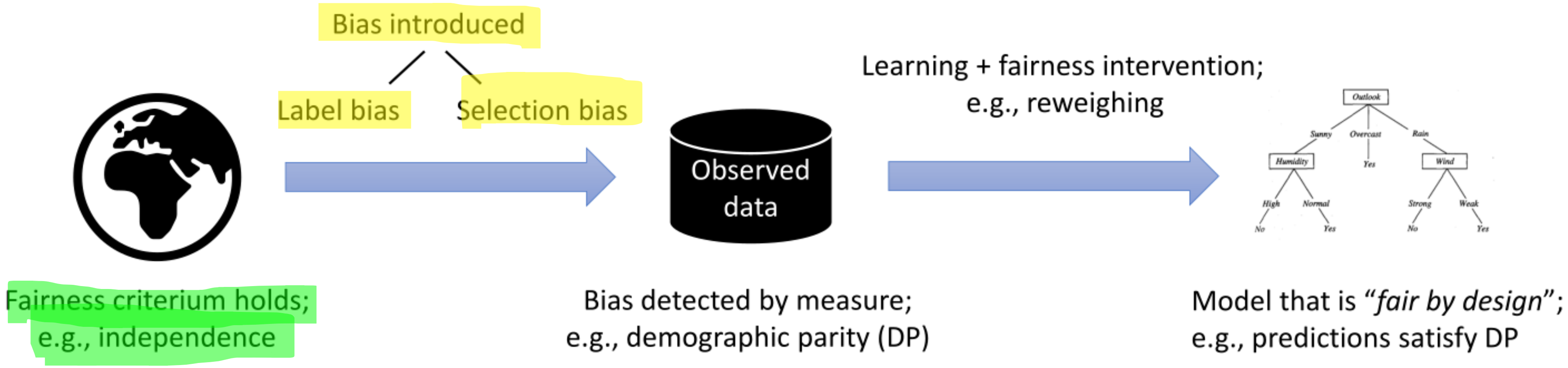
Favier, M., Calders, T., Pinxteren, S., & Meyer, J. (2023). How to be fair? A study of label and selection bias. Machine Learning, 1-24.

Fairness Framework

- Different types of bias introduction
- Different fairness assumptions

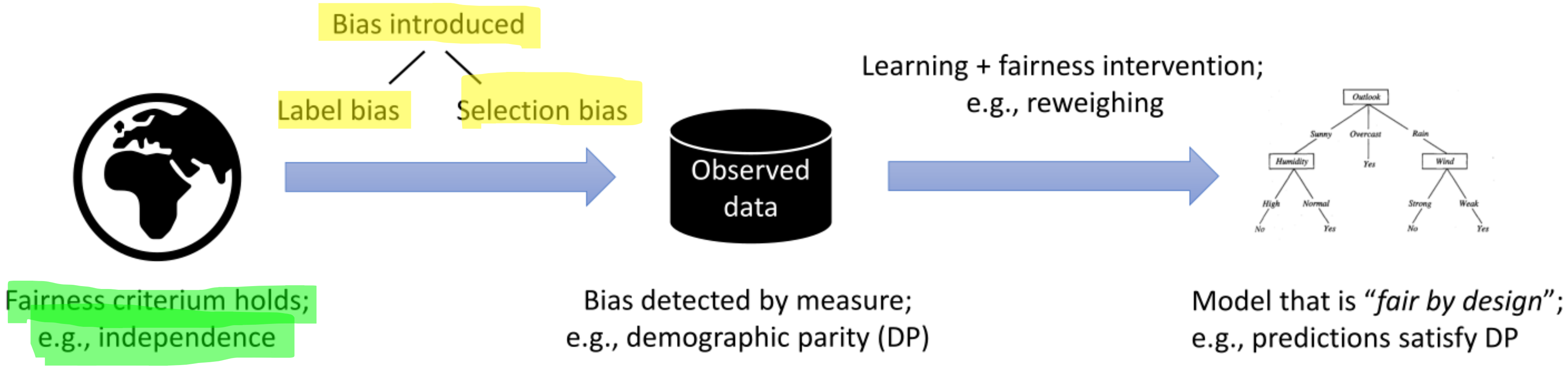


Research Questions



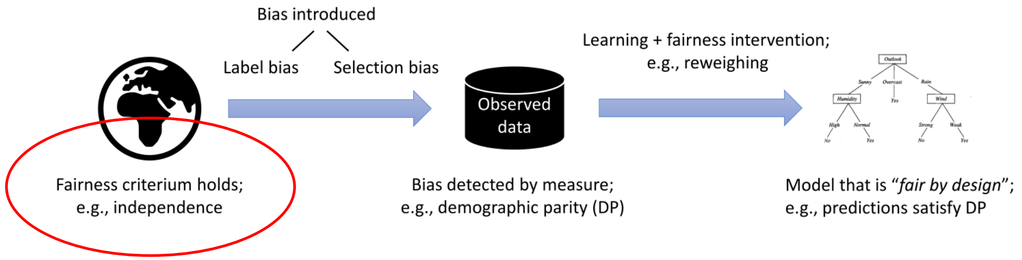
- How to formalize the sources of bias?
- Is the *observed data* consistent with the bias assumptions?
- What should we optimize w.r.t. the biased data?

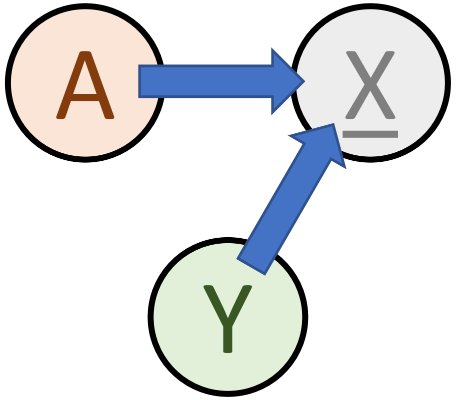
Research Questions



- **How to formalize the sources of bias?**
- Is the *observed data* consistent with the bias assumptions?
- What should we optimize w.r.t. the biased data?

Two definitions of Fair Data

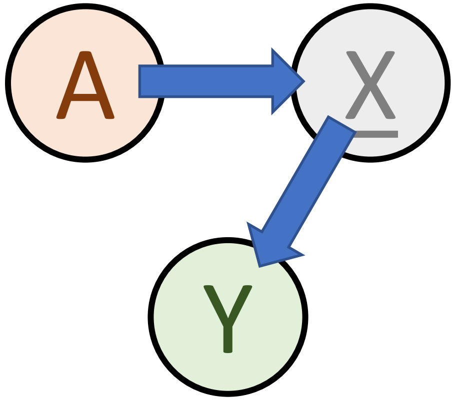




Statistical Parity

$$Y \perp\!\!\!\perp A$$

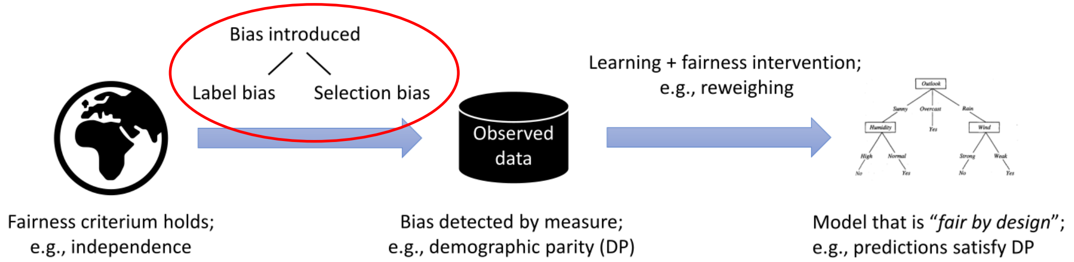
"The Label is equally distributed between sensitive groups"



We're All Equal

$$Y \perp\!\!\!\perp A | X$$

"Identical individuals with different sensitive attributes are treated equally"

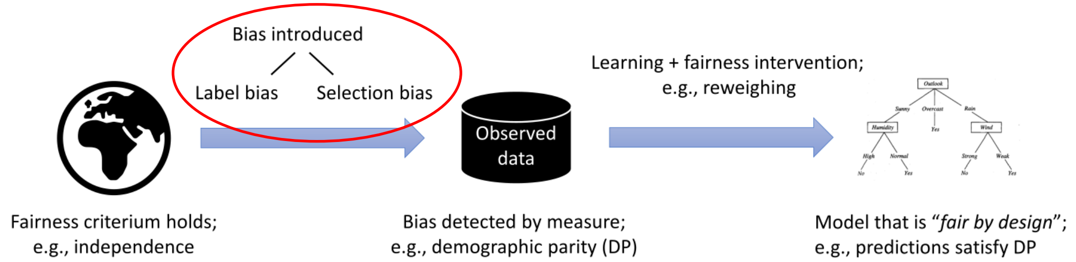


Label Bias

The label that some individuals received does not represent the label they deserved.

Example: Job Hiring
If a racist person is responsible for hiring people, he will deny the positive label to valuable workers from the discriminated group. Those people did not receive the label they deserved.

X_1	…	X_n	A	Y
0	…	1	♂	✓
1	…	1	♀	✗
0	…	0	♂	✓
0	…	1	♀	✓
1	…	0	♂	✗
1	…	0	♂	✓
0	…	0	♀	✗
1	…	1	♀	✗



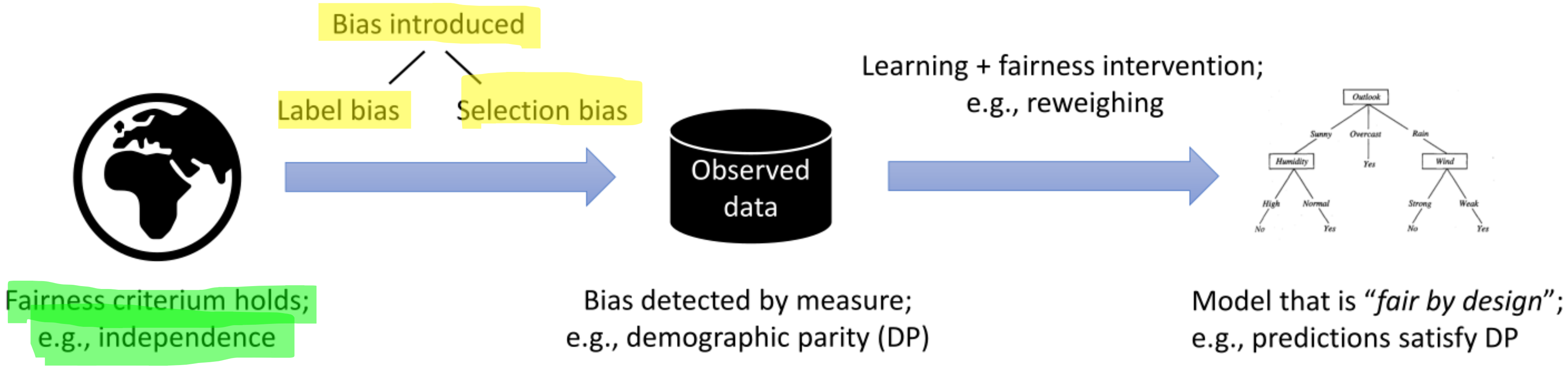
Selection Bias

the selection of the individuals in the dataset is not independent from the individuals' features.

Example: the *toeslagenaffaire*
The data collected relied on anonymous tips, which may lead discriminated groups to be over-represented.

X_1	...	X_n	A	Y
0	...	1	♂	✓
1	...	1	♀	✗
0	...	0	♂	✓
0	...	1	♀	✓
1	...	0	♂	✗
1	...	0	♂	✓
0	...	0	♀	✗
1	...	1	♀	✗

Research Questions

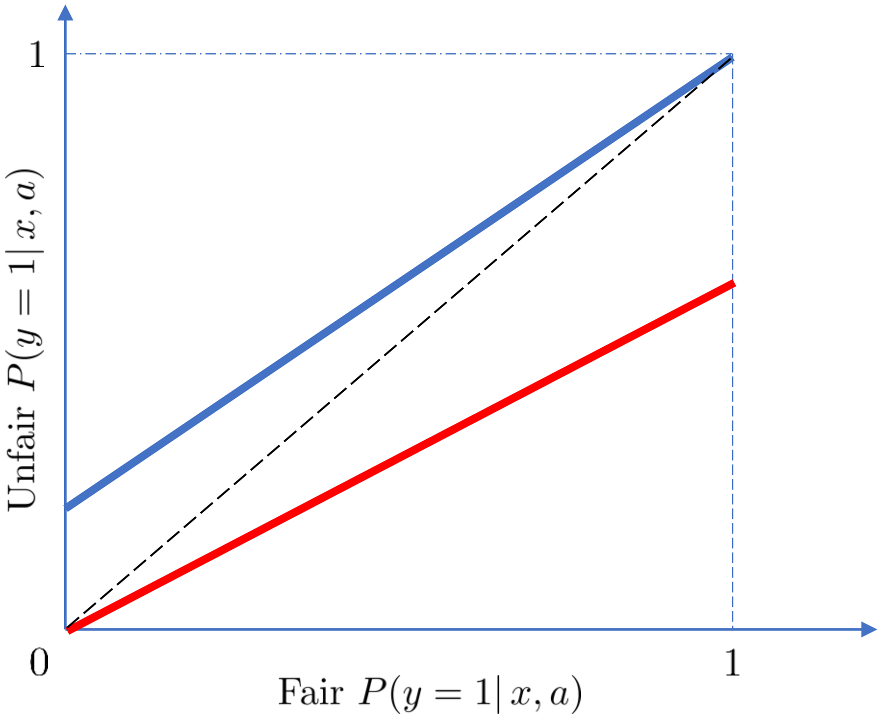


- How to formalize the sources of bias?
- **Is the *observed data* consistent with the bias assumptions?**
- What should we optimize w.r.t. the biased data?

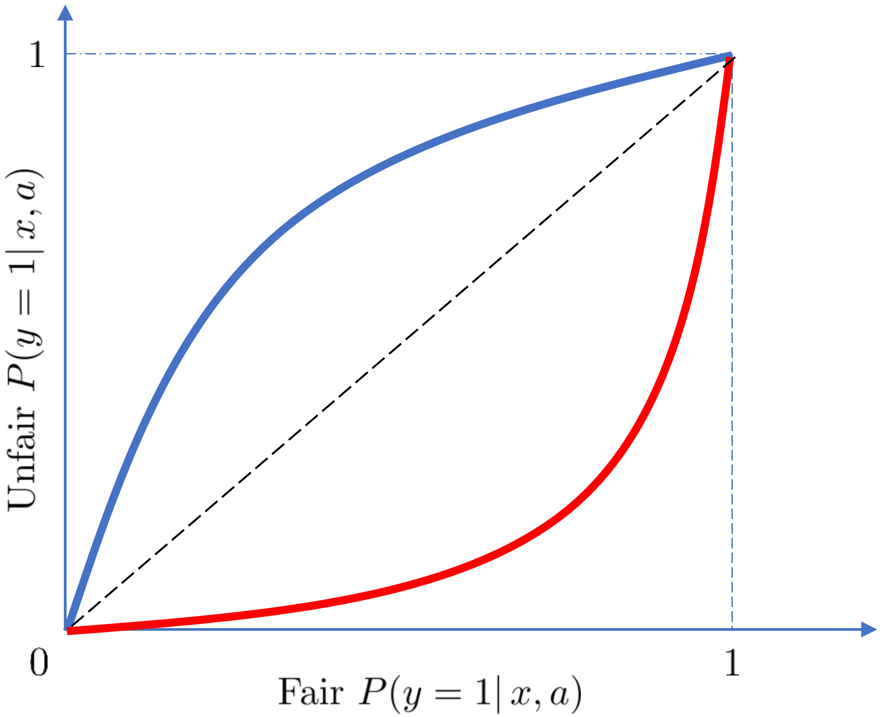
How Bias Changes Probabilities

Privileged Group

Unprivileged Group



Label Bias



Selection Bias

	Statistical Parity	We're All Equal
	$Y \perp\!\!\!\perp A$	$Y \perp\!\!\!\perp A \mid X$
Label Bias	$\left[1 - \frac{P_U(y_0 \mid a_1)}{\max_{x \in X} P_U(y_0 \mid x, a_1)}, \frac{P_U(y_1 \mid a_1)}{\max_{x \in X} P_U(y_1 \mid x, a_1)}\right] \bigcap \left[1 - \frac{P_U(y_0 \mid a_0)}{\max_{x \in X} P_U(y_0 \mid x, a_0)}, \frac{P_U(y_1 \mid a_0)}{\max_{x \in X} P_U(y_1 \mid x, a_0)}\right] \neq \emptyset$	$P_U(y_1 \mid x, a_1) = mP_U(y_1 \mid x, a_0) + c$
Selection Bias	$P_M(y = 1 \mid a = 1) \neq P_M(y = 1 \mid a = 0)$	$\frac{P_U(y_1 \mid x, a_1)}{P_U(y_0 \mid x, a_1)} = k\frac{P_U(y_1 \mid x, a_0)}{P_U(y_0 \mid x, a_0)}$

Statistical Parity + Label Bias

- What does it mean?
  - Necessary condition that allows us to check if label bias has occurred.
  - The amount of bias is tied with how difficult it is to make predictions.
  - In general, the process that generates the biased data might be not unique.

Theorem:

The following set must be non-empty:

$$\left[1 - \frac{P_U(y_0|\, a_1)}{\max_{x\in X} P_U(y_0|\, x, a_1)}, \frac{P_U(y_1|\, a_1)}{\max_{x\in X} P_U(y_1|\, x, a_1)}\right] \bigcap \left[1 - \frac{P_U(y_0|\, a_0)}{\max_{x\in X} P_U(y_0|\, x, a_0)}, \frac{P_U(y_1|\, a_0)}{\max_{x\in X} P_U(y_1|\, x, a_0)}\right]$$

We're All Equal +

Label Bias

$$P_U(y_1|x, a_1) = mP_U(y_1|x, a_0) + c$$

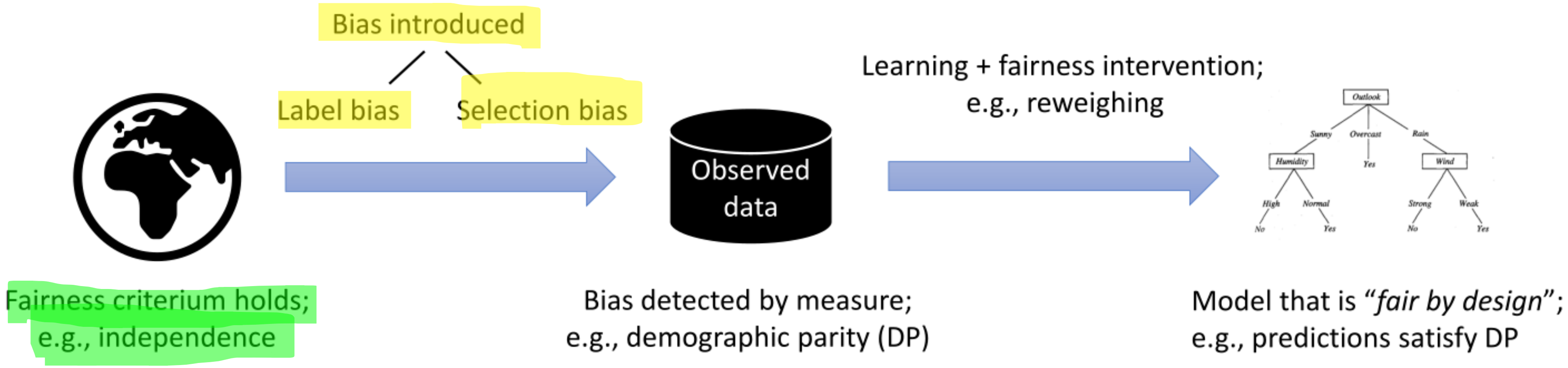
Selection Bias

$$\frac{P_U(y_1|x, a_1)}{P_U(y_0|x, a_1)} = k\frac{P_U(y_1|x, a_0)}{P_U(y_0|x, a_0)}$$

What does it mean?

- Necessary condition that allows us to check if label/selection bias has occurred.
- Very limiting equation to solve, which may explain why fairness by unawareness rarely works.
- In general, the process that generates the biased data might be not unique.

Research Questions



- How to formalize the sources of bias?
- Is the *observed data* consistent with the bias assumptions?
- **What should we optimize w.r.t. the biased data?**

	Statistical Parity $Y \perp\!\!\!\perp A$	We're All Equal $Y \perp\!\!\!\perp A \vert X$
Label Bias	$\left[1 - \frac{P_U(y_0 \vert a_1)}{\max_{x \in X} P_U(y_0 \vert x, a_1)}, \frac{P_U(y_1 \vert a_1)}{\max_{x \in X} P_U(y_1 \vert x, a_1)}\right] \bigcap \left[1 - \frac{P_U(y_0 \vert a_0)}{\max_{x \in X} P_U(y_0 \vert x, a_0)}, \frac{P_U(y_1 \vert a_0)}{\max_{x \in X} P_U(y_1 \vert x, a_0)}\right] \neq \emptyset$	$P_U(y_1 \vert x, a_1) = mP_U(y_1 \vert x, a_0) + c$
Selection Bias	$P_M(y = 1 \vert a = 1) \neq P_M(y = 1 \vert a = 0)$	$\frac{P_U(y_1 \vert x, a_1)}{P_U(y_0 \vert x, a_1)} = k \frac{P_U(y_1 \vert x, a_0)}{P_U(y_0 \vert x, a_0)}$

Statistical Parity + Sampling Bias

What does it mean?

- The fairness measure of the fair model is NOT zero.
- Multiple fairness techniques FAIL to find the fair model.
- Different interventions are needed for different biases, even if the fairness measure is the same.

Theorem:

Let $P_M(y=1|\,x,a)$ be the fair model, then

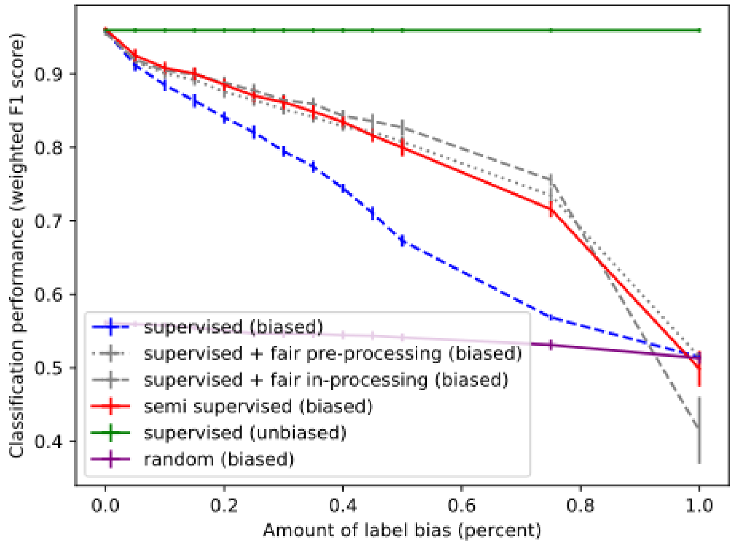
$$P_M(y=1|a=1) \neq P_M(y=1|a=0)$$

Conclusion and Future Work

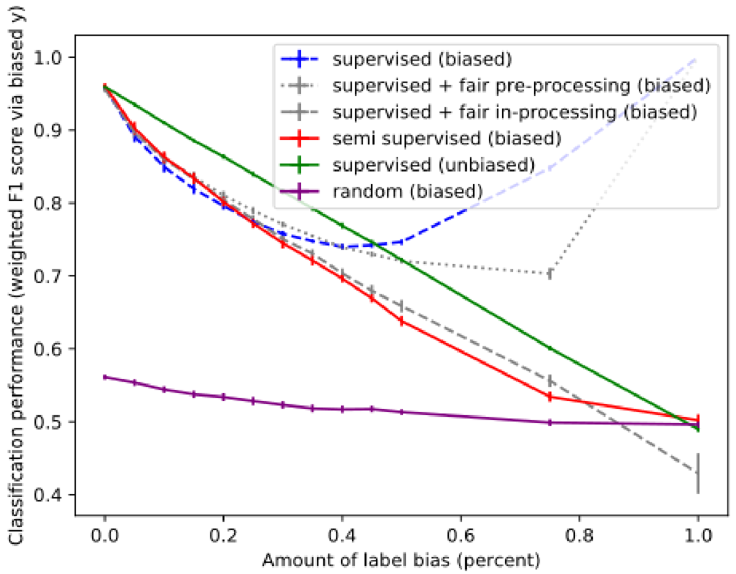
	Statistical Parity	We're all equal
Label Bias	• Can be detected? ✔ • Still satisfied? ✔	• Can be detected? ✔ • Still satisfied? ✔
Sampling Bias	• Can be detected? ✘ • Still satisfied? ✘	• Can be detected? ✔ • Still satisfied? ✔

- From theory to practice: we need methods able to describe the bias that happened.
- Evil is not banal: bias can depend on some other features.
- Connect intervention to bias.

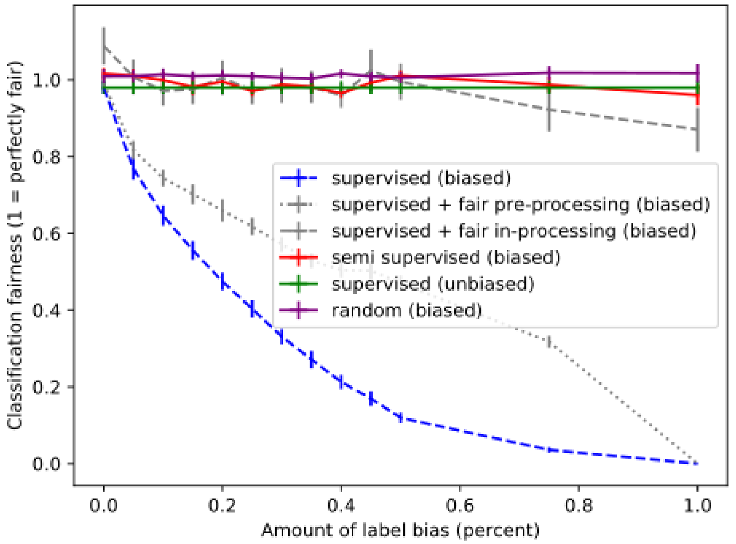
Ethical machine learning is maximizing accuracy in a fair world



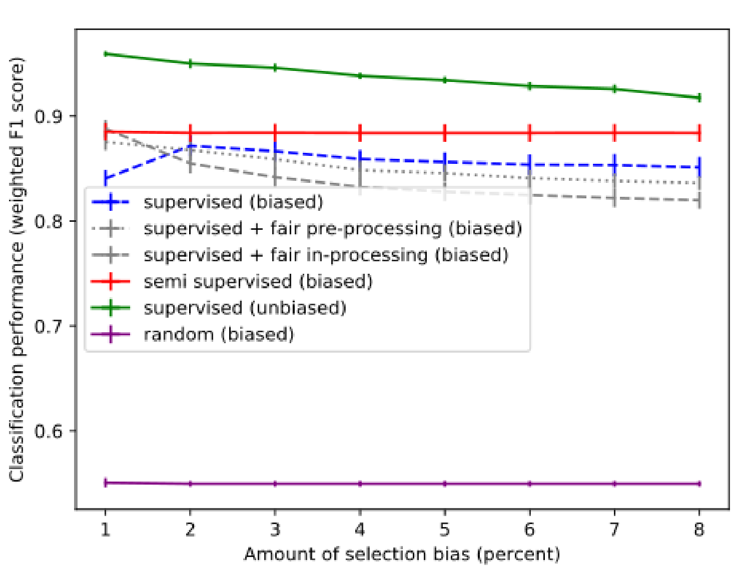
(a) F1 (unbiased truth)



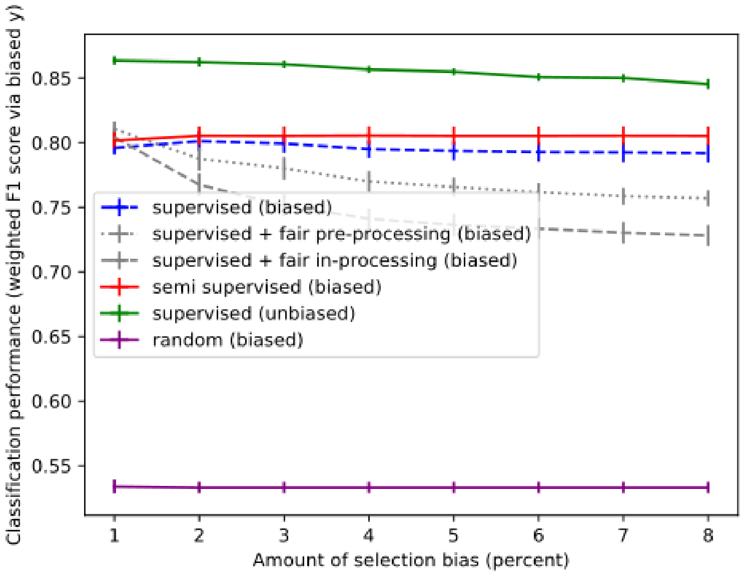
(b) F1 (biased truth)



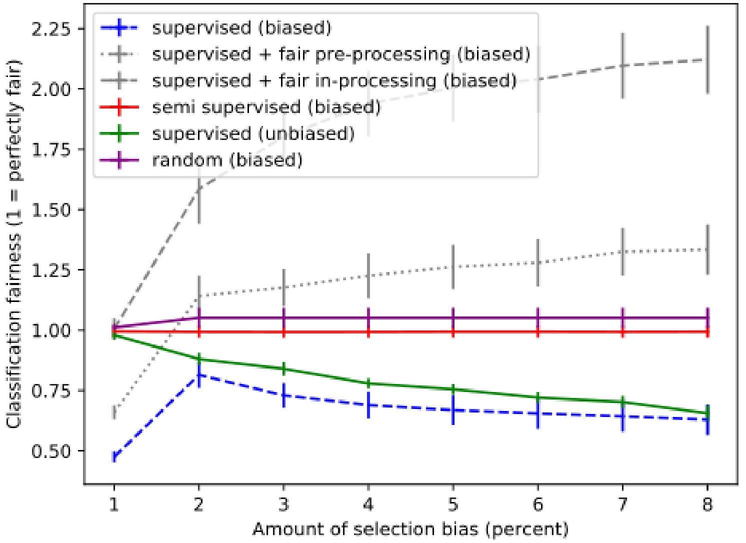
(c) Fairness



(a) F1 (unbiased truth)



(b) F1 (biased truth)



(c) Fairness