Yale

Algorithmic Fairness in ML

CPSC680: Trustworthy Deep Learning

Rex Ying

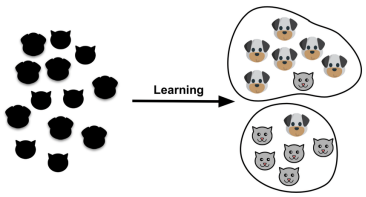
Content

- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

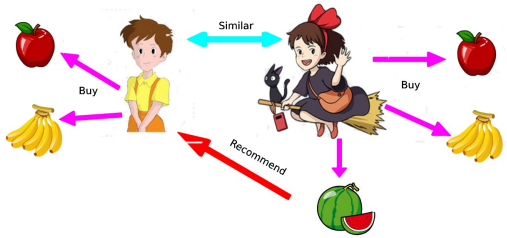
Machine Learning in Consequential Domains

- ML models are increasingly applied in **consequential domains** like healthcare, law enforcement, and employment to automate decision-making.
- The application of ML models in those domains, even though having many benefits (e.g., reducing labor costs), **raises many ethical concerns** as a decision can significantly influence people's lives.

_In_consequential domains



Dog-cat classification



Amazon recommendation

Consequential domains

Job hiring

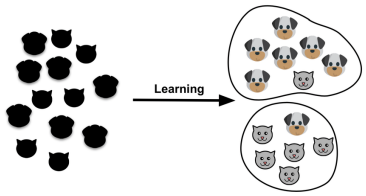
Criminal justice

Loan approval

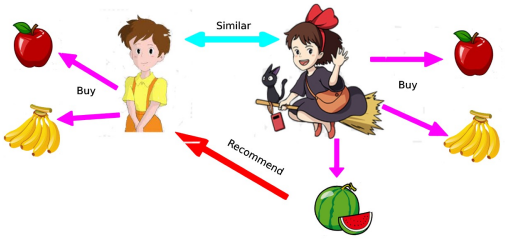
Machine Learning in Consequential Domains

- In high-stakes domains, we do need to care more than just performance

***In*consequential domains**



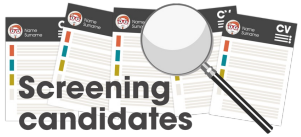
Dog-cat classification



Amazon Recommendation

Performance, performance, performance!!!

Consequential domains



Job hiring

Criminal justice

Loan approval

Hmm!
Is the algorithm accurate?

Is the algorithm private?

Is the algorithm fair?

This lecture

....

COMPAS Case Study

- COMPAS is an algorithm developed by Equivant to **predict the chance that a criminal will commit another crime** in the future (recidivism).
- Recidivism scores impact criminal sentences: if a person is likely to commit another crime, shouldn't they get a longer sentence?
- Real systems that **have been used** in New York, Wisconsin, California, Florida,…
- The system is **claimed to be *fair*** as it did correctly predict recidivism for Black and White defendants at roughly the same rate.

COMPAS Case Study

- The system is claimed to be *fair* as it did ***correctly*** predict recidivism for Black and White defendants at roughly the same rate.
- But, when it was wrong, it was wrong in different ways for Black and White

Black arrestees who would not be rearrested in a 2-year horizon scored as high risk at twice the rate of white arrestees not subsequently arrested

Prior Offense
1 attempted burglary
Subsequent Offenses
3 drug possessions

Prior Offense
1 resisting arrest without violence
Subsequent Offenses
None

This is correct

But

PRO PUBLICA

Hmm! This is ***not*** *fair*

It's not unfair; it satisfies a different notion of fairness!

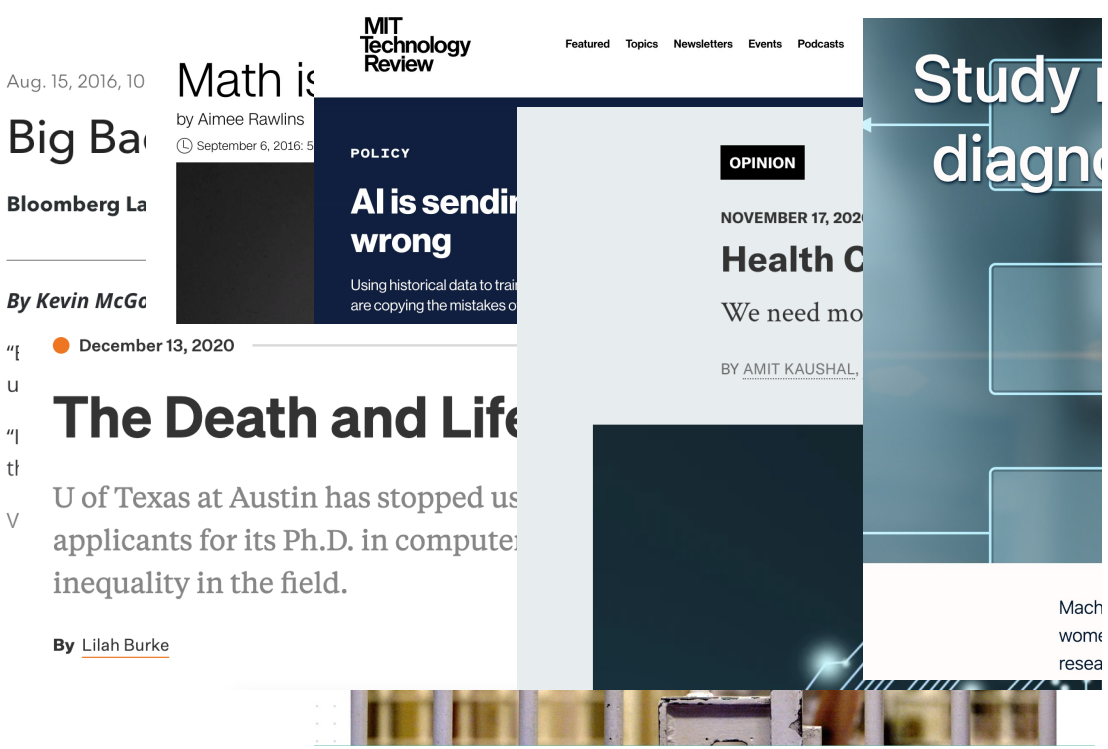
equivant

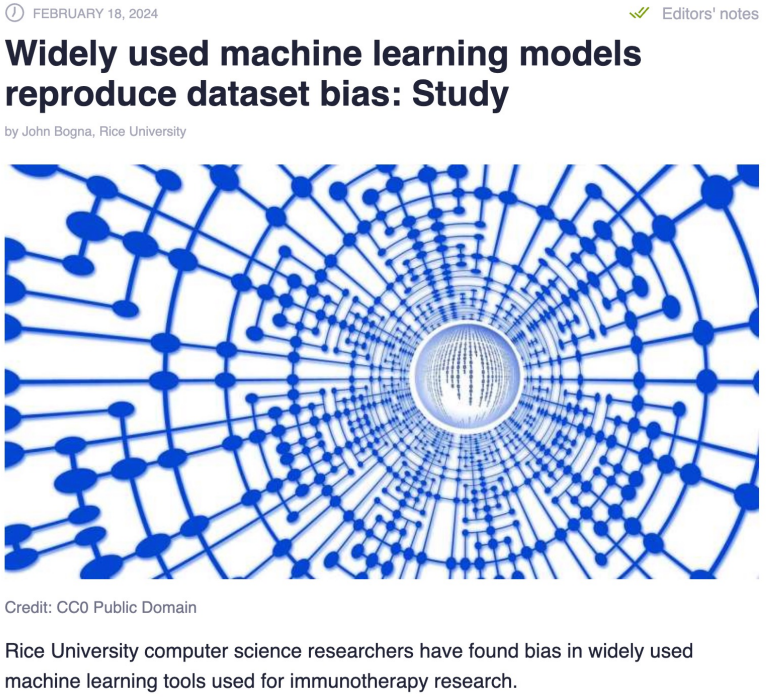
There is no clear "wrong" or "right".
It turns out that these were incompatible definitions.

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

Algorithmic Bias

It's more common than you thought!





Root Cause

- **Data, data, data!** If an ML model is **trained on data biased** by historical inequalities, human prejudices, or flawed collection practices, it will **likely learn and replicate these biases**.
- Biased data may come from
  - **Biased in data collection**: Data is gathered by humans who inherently possess biases.

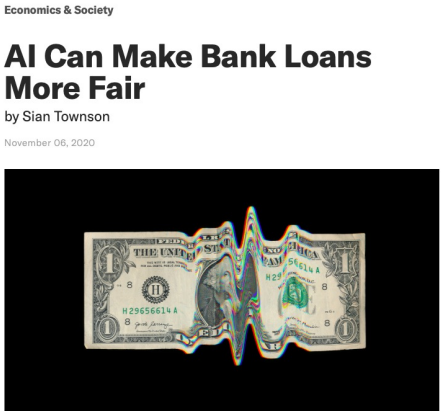
Study: People Associate 'Education' With Lighter Skin

Research participants remembered 'educated' black men as having a lighter skin tone.

By Brian Resnick and National Journal

Root Cause

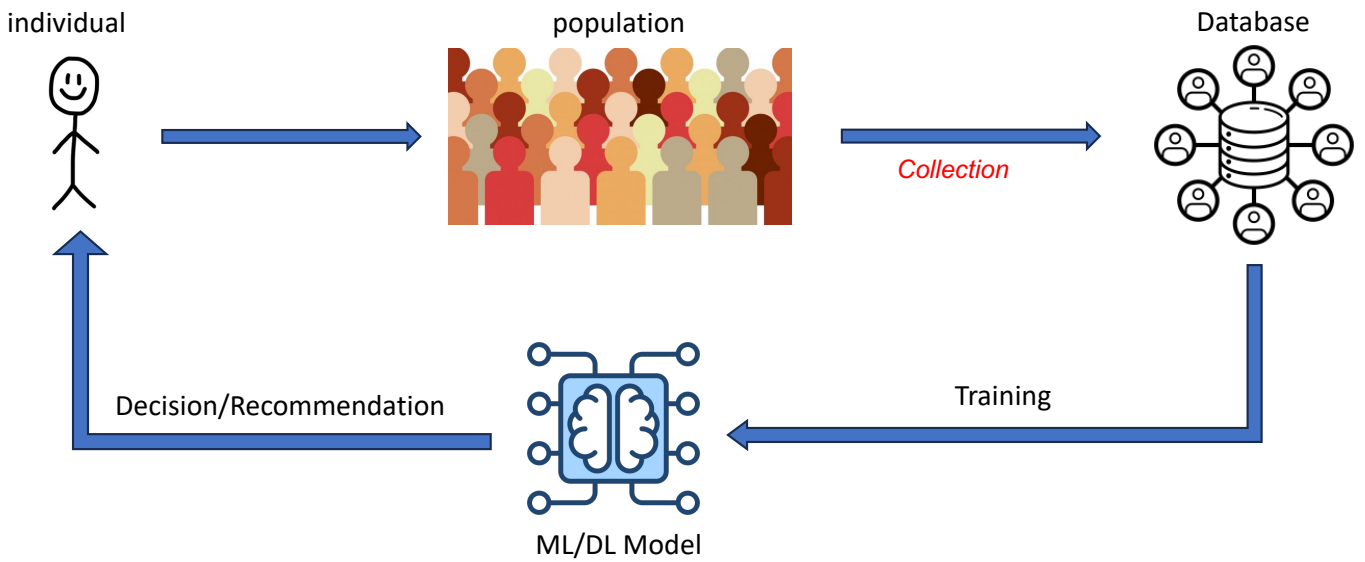
- If an ML model is **trained on data biased** by historical inequalities, human prejudices, or flawed collection practices, it will **likely learn and replicate these biases**.
- Biased data may come from
  - **Biased in data collection**: Data is gathered by humans who inherently possess biases.
  - **Imbalanced data**: ML models aim to optimize the performance on imbalanced data.



Dataset:	Adult
Field of application:	Credit loan
Goal:	Predict income
Covariate:	Age, Occupation, Education, Gender...
Problem:	~20% female, ~80% male

Feedback Loop

- **Amplifying feedback loop:** Systems that adapt based on their outputs can create a feedback loop that reinforces initial biases.



Emerging Legislation

- **Emerging Legislation:** The use of AI models is getting more attention from regulators.

NYC to Regulate Artificial Intelligence-Based Hiring Tools

Posted on December 15, 2021

POSTED IN U.S. STATE LAW, WORKPLACE PRIVACY

On November 10, 2021, the New York City Council passed a bill prohibiting employers and employment agencies from using automated employment decision tools to screen candidates or employees, unless a bias audit has been conducted prior to deploying the tool (the "Bill").

Main Questions

- The fairness of AI-based systems is **unquestionably crucial**.
- However, this is still an open research direction. There are generally **more problems than solutions**.

1. Measure of Fairness

Q: What is the correct measure of fairness?
A: It depends (maybe none)

2. Model Training

Q: How to train an ML model to satisfy a chosen fairness metric?
A: It depends (maybe impossible)

3. Verification

Q: How to verify that an ML model satisfies a chosen fairness metric?
A: It depends (maybe none)

Content

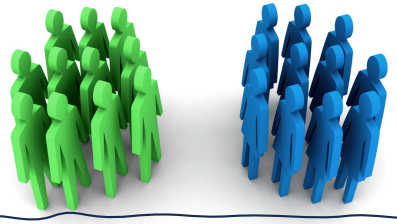
- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

Types of Fairness

- Our idea of fairness is clear, mainly based on two principles:

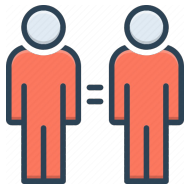
Group Fairness

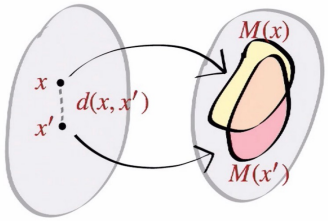
"Models should ***perform equally well*** across different subgroups of a population"



Individual Fairness

"***Similar people*** should be ***treated similarly***"





- But…

How to formally define these notions to measure fairness?

Group Measure of Fairness

- **Let's consider a binary classifier**
  - $Y = \{0, 1\}$ to be the ground truth label (e.g., recidivism)
  - $\hat{Y} = \{0, 1\}$ to be the model prediction
  - $G \in \{0, 1\}$ to be the sensitive attribute (e.g., race, gender)

		Predicted condition	
	Total population = P + N	Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P) [a]	True positive (TP), hit[b]	False negative (FN), miss, underestimation
	Negative (N)[d]	False positive (FP), false alarm, overestimation	True negative (TN), correct rejection[e]

- Error rate = $\frac{FP+FN}{TN+FP+FN+TP}$ — Accuracy

Defendants care →
- False Positive Rate = $\frac{FP}{FP+TN}$ — Probability at which non-offenders being predicted to re-offend

Judges care →
- False Negative Rate = $\frac{FN}{FN+TP}$ — Probability at which offenders were predicted to not re-offend?

Source

COMPAS Case Study

- We consider the COMPAS case study again…

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

- Error rate $\approx 36.2\%$ $\approx$ Error rate $\approx 36.2\%$
- False Positive Rate $\approx 44.9\%$ $>$ False Positive Rate $\approx 23.5\%$
- False Negative Rate $\approx 28.0\%$ $<$ False Negative Rate $\approx 47.7\%$

While having the the same error rate, Black defendants have 1.9x higher False Positive Rate!

Source

Group Measure of Fairness: Statistical Parity

- **Idea:** Different groups have the same probability of receiving favorable outcomes.

Definition #1 (Statistical parity): The classifier is said to satisfy statistical parity if the probability that the algorithm makes a positive prediction ($\hat{Y}$ = 1) is the same across different groups

$$P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1)$$

- **Cons**: Does not take the ground truth label Y into account (different groups have different underlying distributions for Y)

Group Measure of Fairness: Equal Opportunity

- **Idea:** Different groups have the same **true positive rate (TPR)** across different groups

Definition #2 (Equal Opportunity): The classifier is said to satisfy equal opportunity if the classifier has the same true positive rate across different groups

$$TP(G=0) = TP(G=1) \Leftrightarrow P\left(\hat{Y}=1 \mid G=0, \boldsymbol{Y=1}\right) = P\left(\hat{Y}=1 \mid G=1, \boldsymbol{Y=1}\right)$$

- **Cons**: Does not take error rate into account

		Predicted condition	
	Total population = P + N	Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P) [a]	True positive (TP), hit[b]	False negative (FN), miss, underestimation
	Negative (N)[d]	False positive (FP), false alarm, overestimation	True negative (TN), correct rejection[e]

Group Measure of Fairness: Equalized Odds

- **Idea:** Different groups have the same **true positive rate (TPR)** and **false positive rate (FPR)** across different groups

Definition #3 (Equalized Odds): The classifier is said to satisfy equal opportunity if the classifier has the same TPR and FPR across different groups

$$TP(G=0) = TP(G=1) \Leftrightarrow P\left(\hat{Y}=1 \mid G=0, \boldsymbol{Y=1}\right) = P\left(\hat{Y}=1 \mid G=1, \boldsymbol{Y=1}\right)$$

$$FP(G=0) = FP(G=1) \Leftrightarrow P\left(\hat{Y}=1 \mid G=0, \boldsymbol{Y=0}\right) = P\left(\hat{Y}=1 \mid G=1, \boldsymbol{Y=0}\right)$$

		Predicted condition	
	Total population = P + N	Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P) [a]	True positive (TP), hit[b]	False negative (FN), miss, underestimation
	Negative (N)[d]	False positive (FP), false alarm, overestimation	True negative (TN), correct rejection[e]

Group Measure of Fairness: Calibration

- **Let S** be the predicted score (e.g., logits) of the model for a given input
- **Idea:** when two people from different groups get the same predicted score, they should have the same probability of belonging to the favorable class.

Definition #4 (Calibration): The classifier is said to satisfy calibration if the probability of $Y = 1$ across different groups is the same given predicted score

$$P(Y = 1 \mid S = s, G = 0) = P(Y = 1 \mid S = s, G = 1)$$

i.e., people from different groups with the same predicted score should have the same probability of belonging to $y = 1$

- **Pros**: Wel-calibrated confidence score across different groups
- **Cons**: Does not compatible with equalized odds

Impossibility Theorem

- **Unfortunately**, we cannot consider all defined fairness criteria at the same time.

Theorem (Impossibility): No classifiers can satisfy three fairness (statistical parity, equalized odds, and calibration) criteria simultaneously.

- i.e., if an algorithm satisfies *statistical parity*, the algorithm ***cannot*** have *equal odds* and *calibration* at the same time.

Kleinberg et al. "Inherent trade-offs in the fair determination of risk scores." 2016

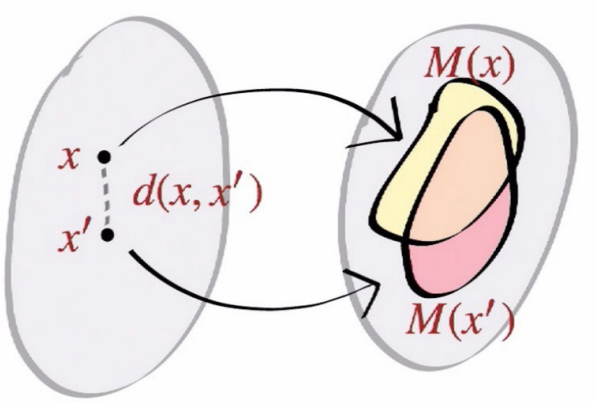
Individual Fairness

- **Idea**: *Similar* individuals should be treated *similarly*
- **Let** $M(x)$ to be the algorithm output for the individual x

Definition #5 (Individual fairness): The classifier is said to satisfy individual fairness, if for any x, x'

$$|M(x) - M(x')| \leq \delta \quad \text{if} \quad d(x, x') \leq \epsilon$$

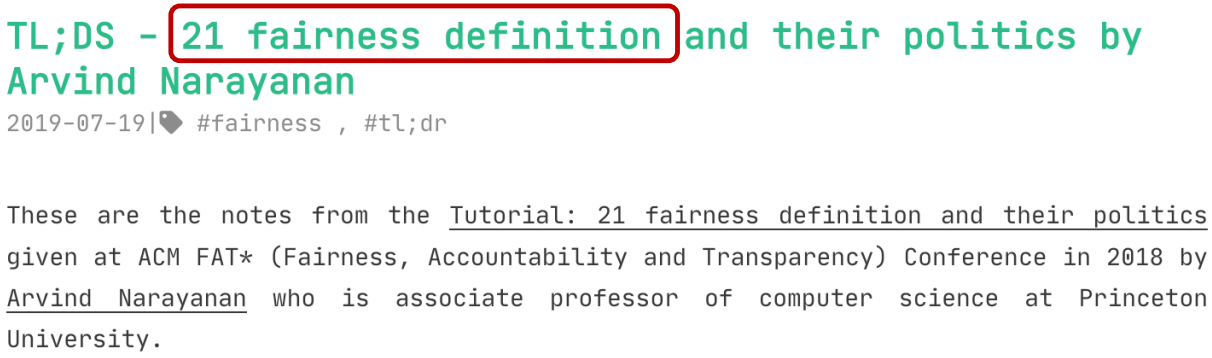
where d is the distance function measure similarity between two individuals.



- **Pros**: can model heterogeneity within each group
- **Cons**: Notion of "similar" is hard to define mathematically, especially in high-dimensions.

And many more…

- There are **many definitions** of fairness with different criteria.
  - Overall accuracy equality
  - Conditional use accuracy equality
  - Well-carlibration
  - Bayesian Fairness
  - Counterfactual Fairness
  - Generalized Entropy Index
  - Theil Index
  - …



TL;DS - 21 fairness definition and their politics by Arvind Narayanan

2019-07-19 | #fairness , #tl;dr

These are the notes from the Tutorial: 21 fairness definition and their politics given at ACM FAT* (Fairness, Accountability and Transparency) Conference in 2018 by Arvind Narayanan who is associate professor of computer science at Princeton University.

- Fairness is ***not a purely technical issue***. We need to think about context and stake holders. ***Different notions of fairness matter to different stakeholders***.

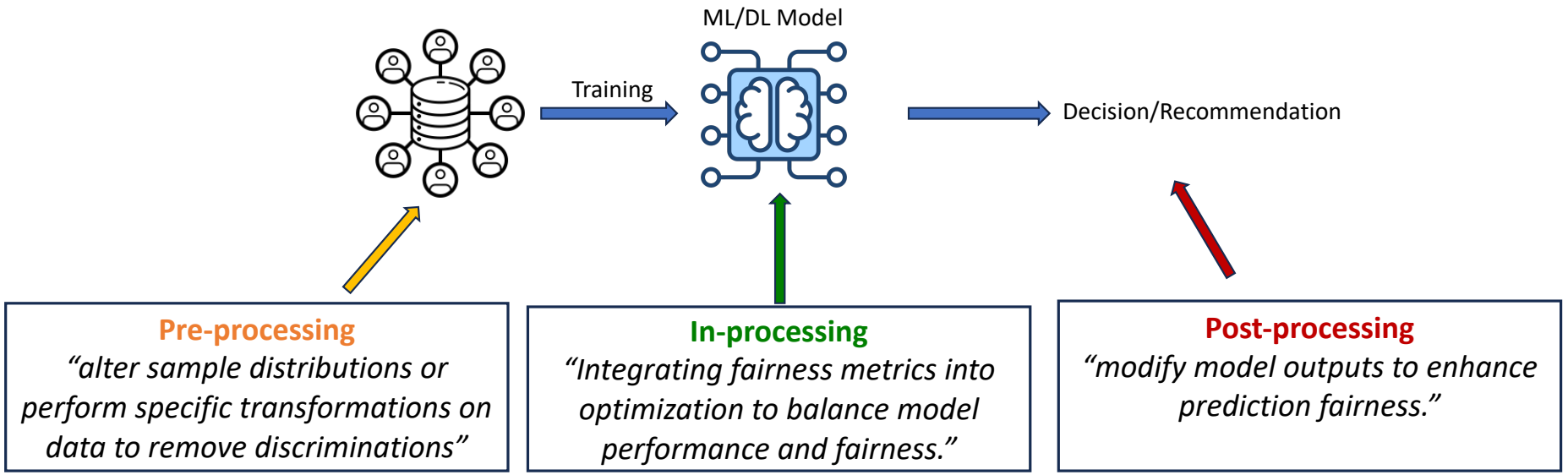
Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." ACM Computing Surveys (2020).

Content

- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

Mitigating Algorithmic Bias

- Technical fairness interventions operate in different location in ML pipeline



Pre-processing Approach

- **The goal** is to **make the classifier "immune" to one or more sensitive variables** (e.g., gender, race, etc.).

- **Naïve approach:** removing sensitive attribute?
  - COMPAS does not use race as an input to the algorithm but still gives very different outcomes for white vs black defendants!
  - **Reason**: other features (e.g. zip code) may correlate with the sensitive feature

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

- Error rate ≈ 36.2%
- False Positive Rate ≈ 44.9%
- False Negative Rate ≈ 28.0%

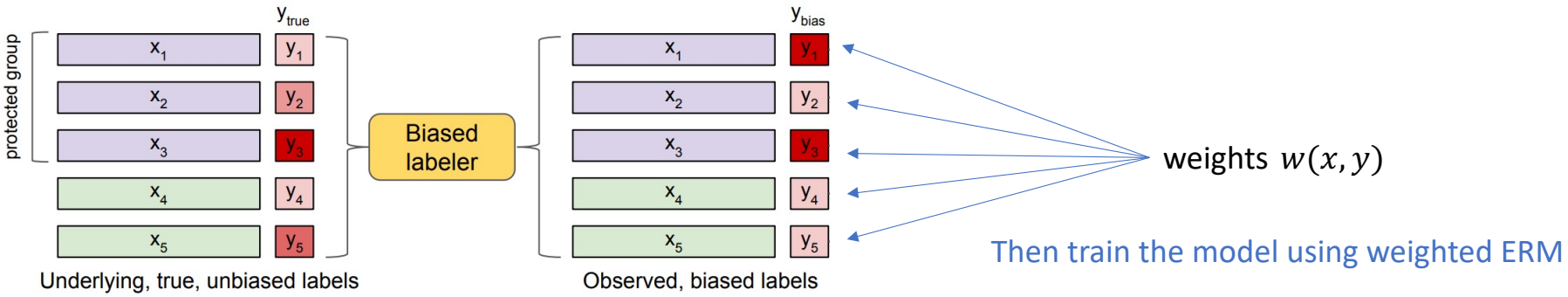
≈ / > / <

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

- Error rate ≈ 36.2%
- False Positive Rate ≈ 23.5%
- False Negative Rate ≈ 47.7%

Pre-processing: Reweighting Training Data

- **Goal:** reweighting the training data to reverse the bias in the training data.
- **Settings**:
  - There exists an *unbiased* labeling function y_{true}: $\mathcal{X} \to \{0, 1\}, P(y|x) = y_{true}(x)$
  - But we can only observe *biased labels* y_{bias}: $\mathcal{X} \to \{0, 1\}, P(y|x) = y_{bias}(x)$ produced by annotators or collectors who are biased against a certain group (labels for other groups remain accurate).



Jiang, Heinrich, and Ofir Nachum. "Identifying and correcting label bias in machine learning." ICML, 2020.

Pre-processing: Reweighting Training Data

- We can express notions of **fairness via linear constraints**. **Let**:
  - $h(y|x)$ be the probability of ML models labeling x by y.
  - $P_X = \mathbb{E}_{x\sim\mathcal{P}}\, y_{true}(x)$ be the proportion of input x having a positive label.
  - $\{G = b\}$ to be the protected group, probability of x in the protected group is $g(x)$
  - $P_b = \mathbb{E}_{x\sim\mathcal{P}}\, y_{true}(x, x \in \{G = b\})$ be proportion of positive input x in protected group.
- The classifier $h(\cdot)$ is said to satisfy ***equal opportunity*** if

$$\boldsymbol{P\left(\widehat{Y} = 1 \mid G = b, Y = 1\right) - P\left(\widehat{Y} = 1 \mid G = w, Y = 1\right) = 0}$$

$$\Leftrightarrow \mathbb{E}_{x\sim P}\left[\left(\frac{\boldsymbol{h(1|x)g(x)y_{true}(x)}}{\boldsymbol{P_b}} - \frac{\boldsymbol{h(1|x)y_{true}(x)}}{\boldsymbol{P_x}}\right)\right] = 0$$

$$\mathbb{E}_{x\sim P}[(\boldsymbol{h(1|x)\, c(x,y)})] = 0$$

We could do similarly for other notions of fairness

reference

Pre-processing: Reweighting Training Data

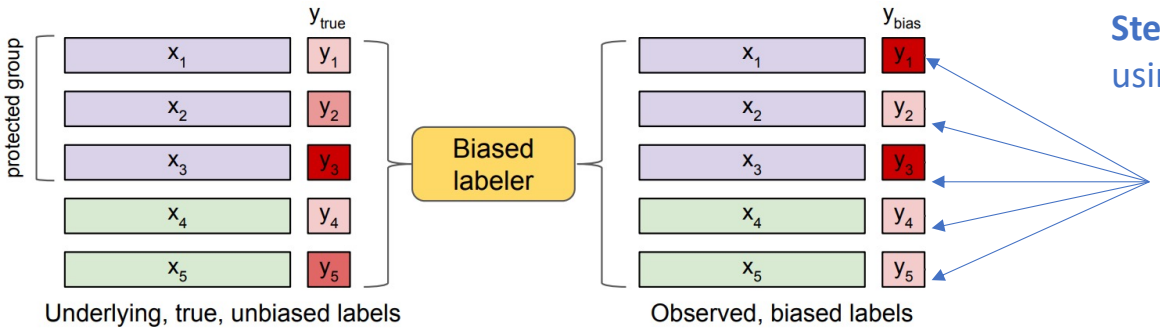
- Notice that y_{bias} can also be seen as an ML model having bias score

$$\mathbb{E}_{x\sim P}\left[\left(h(1|x)\, c(x,y)\right)\right] = \epsilon$$

- **Proposition**: Assume that y_{bias} **is the closest labeling function** (defined via KL divergence) to the true labeling function y_{true}, then, y_{bias} **has a closed form**

$$y_{bias}(y|x) \propto y_{true}(y|x)\exp\left(\lambda_\epsilon c(x,y)\right) \quad \text{for some param } \lambda_\epsilon$$

- **Weighting scheme is then defined as**



Step 1: optimize λ_ϵ to satisfy fairness constraint.
Step 2: The model is then trained on the dataset using these weights

$$w(x,y) = \frac{\exp\left(\lambda_\epsilon c(x,y)\right)}{\exp\left(\lambda_\epsilon c(x,0)\right) + \exp\left(\lambda_\epsilon c(x,1)\right)}$$

Guarantee?

Pre-processing: Reweighting Training Data

Theorem (statistical consistency): Let $h^\star$ minimize the re-weighted ERM with observed labels y_{bias} over all Lipschiz classifiers, then the mean square loss between $h^\star$ and y_{true} is bounded.

- **Empirical results:** smallest bias with acceptable utility trade-off

	Unconst.	Post-fix	Lagrange	Ours
Adult error	14.15	16.6	20.47	16.51
Adult bias	0.1173	0.0129	0.0198	**0.0037**
Bank error	9.41	9.7	10.46	9.63
Bank bias	0.0349	0.0068	0.0126	**0.0056**
COMPAS error	31.49	32.53	40.16	35.44
COMPAS bias	0.2045	0.0201	0.0495	**0.0155**
Crime error	11.62	32.06	28.46	30.06
Crime bias	0.4211	0.0653	0.1538	**0.0107**
German error	24.85	24.85	25.45	25.15
German bias	0.0766	0.0346	0.0410	**0.0137**