Fairness in AI

What is it? Why is it a concern? How to ensure it?

Francisca Morgado
Data Scientist at NILG.AI

01

Why Fairness?

Types of bias and bias loop

02

What is Fairness?

Fairness definitions

03

How to ensure Fairness in ML?

Bias mitigation and fair algorithms

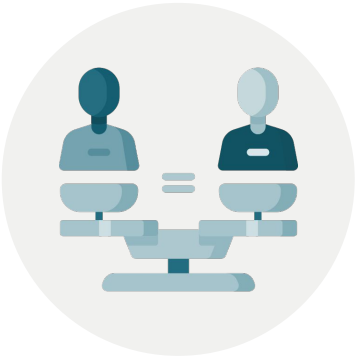
04

Where to apply it?

Use cases on Fairness

01 Why Fairness?

Why Fairness?



An algorithm used in USA courts, attributes a higher criminal risk on black defendants. [1]

Automatic sentiment analysis rates sentences as "I'm homosexual" with a negative score. [2]

Automatic translations have gender bias. When translating "She works in an Hospital, my friend is a doctor" from English to Portuguese, it assumes the doctor is a man. [3]

E-commerce website was displaying different prices depending on people localization and distance from a rival store. [4]

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] https://mashable.com/2017/10/25/google-machine-learning-bias/?europe=true
[3] https://translate.google.com/#view=home&op=translate&sl=en&tl=pt&text=She%20works%20in%20an%20Hospital%2C%20my%20friend%20is%20a%20nurse.
[4] https://www.wsj.com/articles/SB10001424127887323777204578189391813881534

Types of Bias

Historical Bias

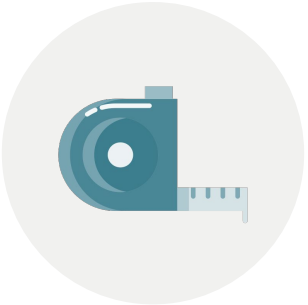
Bias present in our history that still affect the current systems. e.g., gender and racial bias.

Population Bias

This bias results when the statistics, demographics, representatives, and user characteristics represented in the training dataset differs from the population on the test set.

Observer Bias

It happens when the observer projects his/her expectations onto the problem, picking the solution that best suits the initial goal, instead of suiting the real data.



Measurement Bias

It occurs when unfair features are used as proxy labels. e.g., use family criminal records as a proxy to measure the level of “individual crime risk”.

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

Types of Bias

Funding

Aggregation

Social

Historical Bias

Bias present in our history that still affect the current systems. e.g., gender and racial bias.

Sampling

Simpson's Paradox

Population Bias

This bias results when the statistics, demographics, re[illegible]s, and user [illegible]presented in the [illegible]differs from the [illegible]est set.

User Interaction

Content Production

Self-Selection

Observer Bias

[illegible]hen the observer expectations onto [illegible]cking the solution [illegible] the initial goal, [illegible]ng the real data.

Evaluation

Emergent

[illegible] Bias

It o[illegible]en unfair features are used as proxy labels. e.g., use family criminal records as a proxy to measure the level of "individual crime risk".

Linking

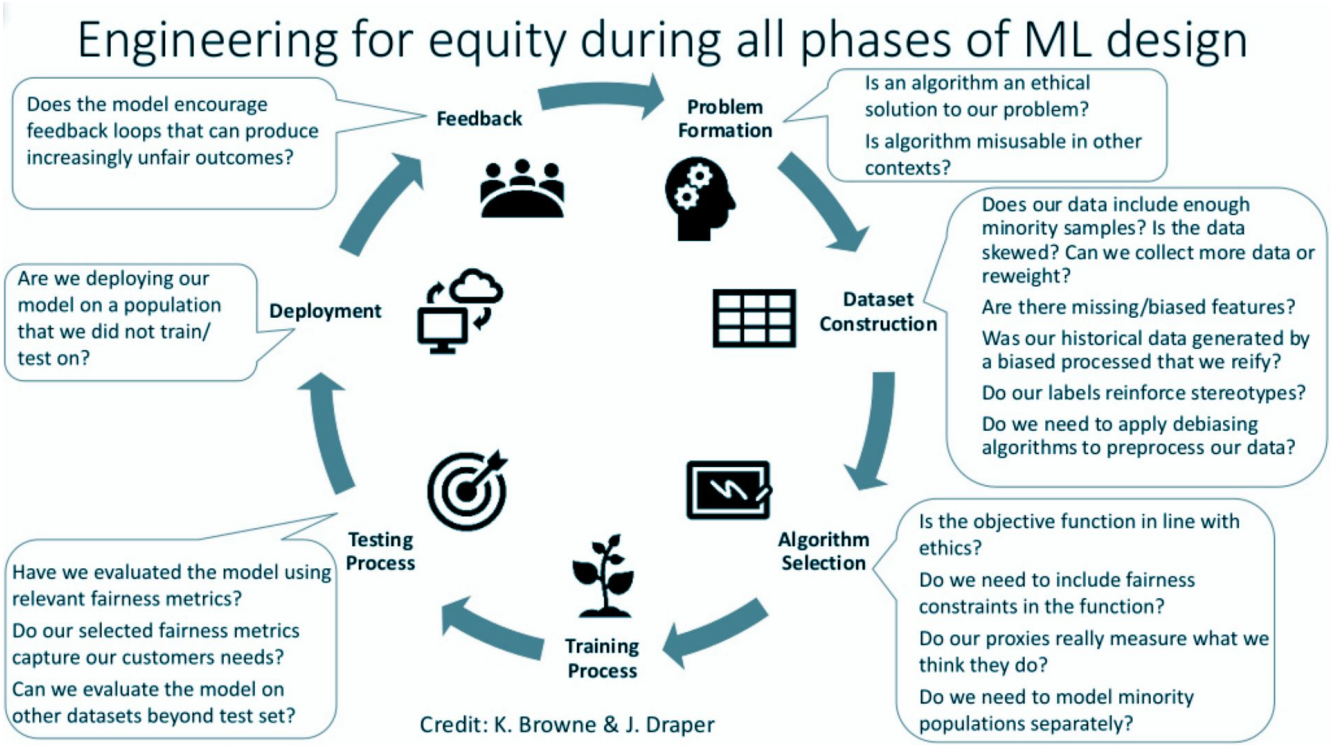
Omitted Variable

Popularity

Behavioral

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

Bias Feedback loop



02

What is Fairness?

What is Fairness?

Def 1. Equalized Odds

The protected and unprotected groups should have equal rates for true positives and false positives

Def 2. Equal Opportunity

The protected and unprotected groups should have equal true positive rates

Def 3. Fairness Through Awareness

Any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome

Def 4. Demographic Parity

Def 5. Fairness Through Unawareness

Def 6. Treatment Equality

Def 7. Test Fairness

Def 8. Counterfactual Fairness

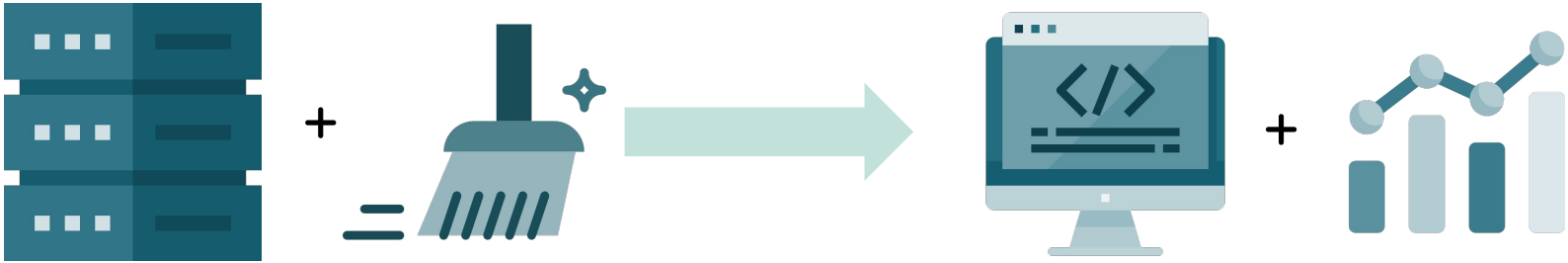
Def 9. Fairness in Relational Domains

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

03 How to ensure Fairness in ML?

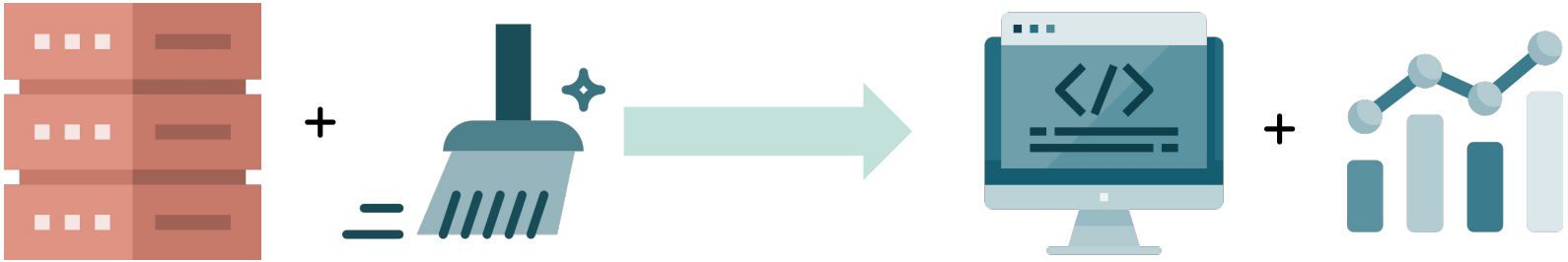
Fairness in AI

Mitigating Unfairness in the AI framework



Fairness in AI

Mitigating Unfairness in the AI framework



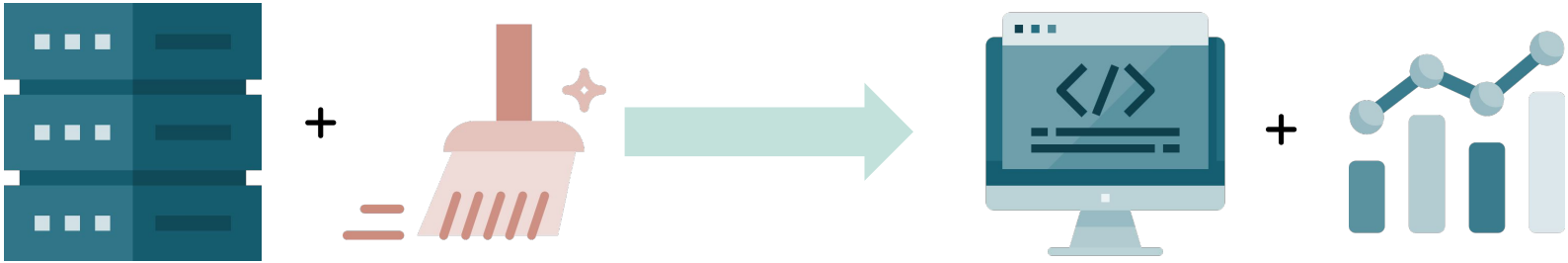
Data Collection

Rethink your data collection process, aiming representative datasets and avoiding unfair biased features.

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

Fairness in AI

Mitigating Unfairness in the AI framework



Data Collection

Rethink your data collection process, aiming representative datasets and avoiding unfair biased features.

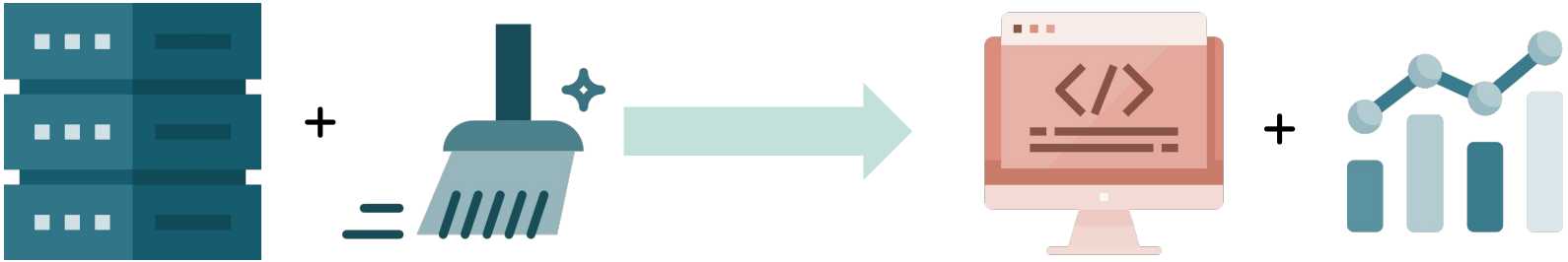
Pre-processing

Try to transform the data, removing underlying discrimination and unbalanced representations.

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

Fairness in AI

Mitigating Unfairness in the AI framework



Data Collection

Rethink your data collection process, aiming representative datasets and avoiding unfair biased features.

Pre-processing

Try to transform the data, removing underlying discrimination unbalanced representations.

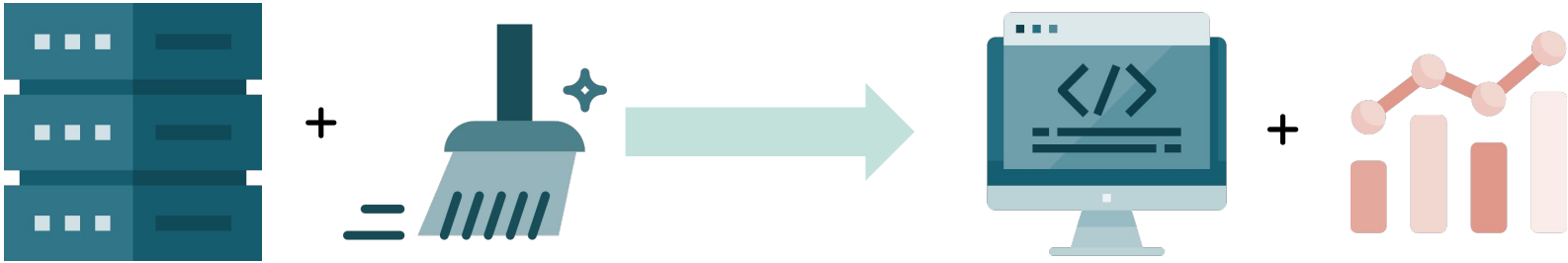
In-processing

Modify the algorithms in order to remove discrimination during the training process. E.g. Adding Fairness regularization.

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

Fairness in AI

Mitigating Unfairness in the AI framework



Data Collection

Rethink your data collection process, aiming representative datasets and avoiding unfair biased features.

Pre-processing

Try to transform the data, removing underlying discrimination unbalanced representations.

In-processing

Modify the algorithms in order to remove discrimination during the training process. E.g. Adding Fairness regularization.

Post-processing

Use Fair metrics to choose the best model; Use predictions from a "black-box" model to feed a Fair algorithm; Use the predictions from the model and apply fairness in the decision process.

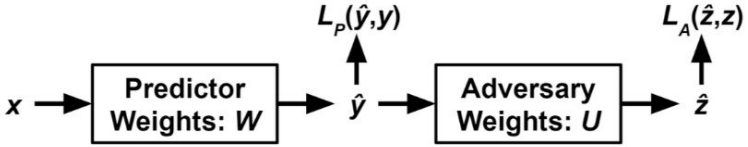
Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

Adversarial Learning

Adversarial Debiasing Network

Method:

1. Given X predicts Y
2. The dense layer is passed to an adversary network
3. Adversary network predicts Z (protected variable) given Y and $\hat{Y}$
4. Goal: minimizing the accuracy of the Adversary while maximizing Predictor accuracy



Results:

	Female		Male	
	Without	With	Without	With
FPR	0.0248	0.0647	0.0917	0.0701
FNR	0.4492	0.4458	0.3667	0.4349

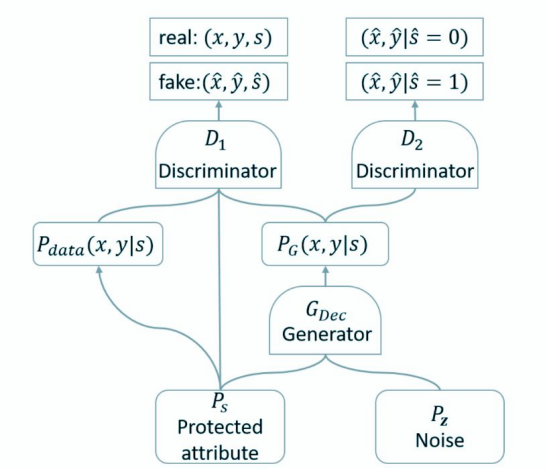
Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

Adversarial Learning

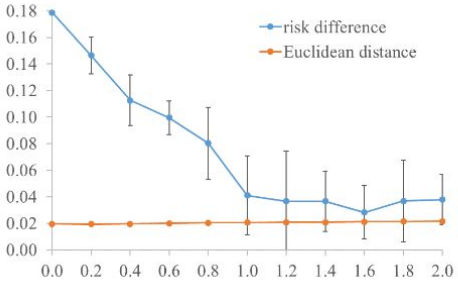
FairGAN

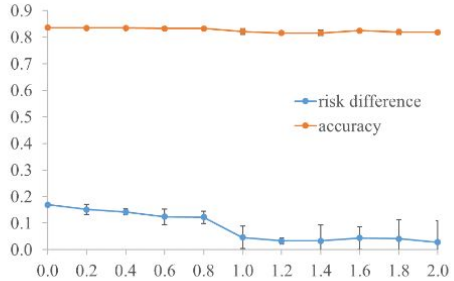
Method:

1. Use Generative Adversarial Networks (GAN) to generate artificial data, PG, adding noise to real data;
2. Use D1 Discriminator to predict if the data is real or fake;
3. Use D2 Discriminator to predict if the data is “protected” or not;
4. Train the GAN in order to fool the discriminators;
5. Use the generated debiased data to train your model



Results:





Xu, Depeng, et al. "Fairgan: Fairness-aware generative adversarial networks." *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018..

Adversarial Learning

Adversarial SHAP - Methods

1. Use SHapley Additive Explanations (SHAP) to measure the relevance of the "protected" features, Z, with respect to the label, Y;

2. Define "Fairness by Explicability" metrics:
   a. FE: difference in mean attribution of Z between the "protected" and "unprotected" groups
   b. SFE: total attribution of Z across the population

3. Train the model using Fairness regularization in the Loss function

$$\mathcal{L}_{\text{fair}} = (1-\lambda) * \mathcal{L}_o + \lambda * \mathcal{R}$$

SHAPSqueeze

$$\mathcal{R} = C \sum_i (\phi_Z^{i,g})^2$$

SHAPEnforce

$$\mathcal{P} = \begin{cases} -\phi_Z^{i,g}, \text{if } y_i = 1 \\ 0, \text{otherwise} \end{cases}$$

Hickey, James M., Pietro G. Di Stefano, and Vlasios Vasileiou. "Fairness by Explicability and Adversarial SHAP Learning." *arXiv preprint arXiv:2003.05330* (2020).

Adversarial Learning

Adversarial SHAP - Results

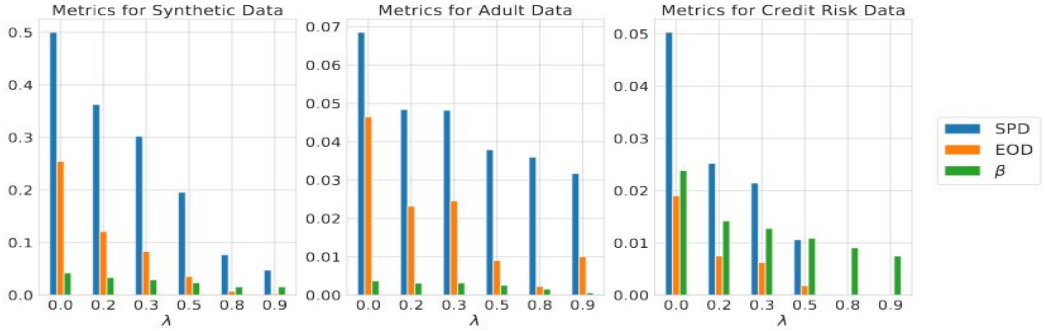
SHAPSqueeze

From λ=0 to λ=0.9, AUC and accuracy drop 0.04

SPD - Statistical Parity Difference
EOD - Equality of Opportunity Difference

SHAPEnforce

From λ=0 to λ=0.9, AUC drop 0.03 and accuracy drop 0.08



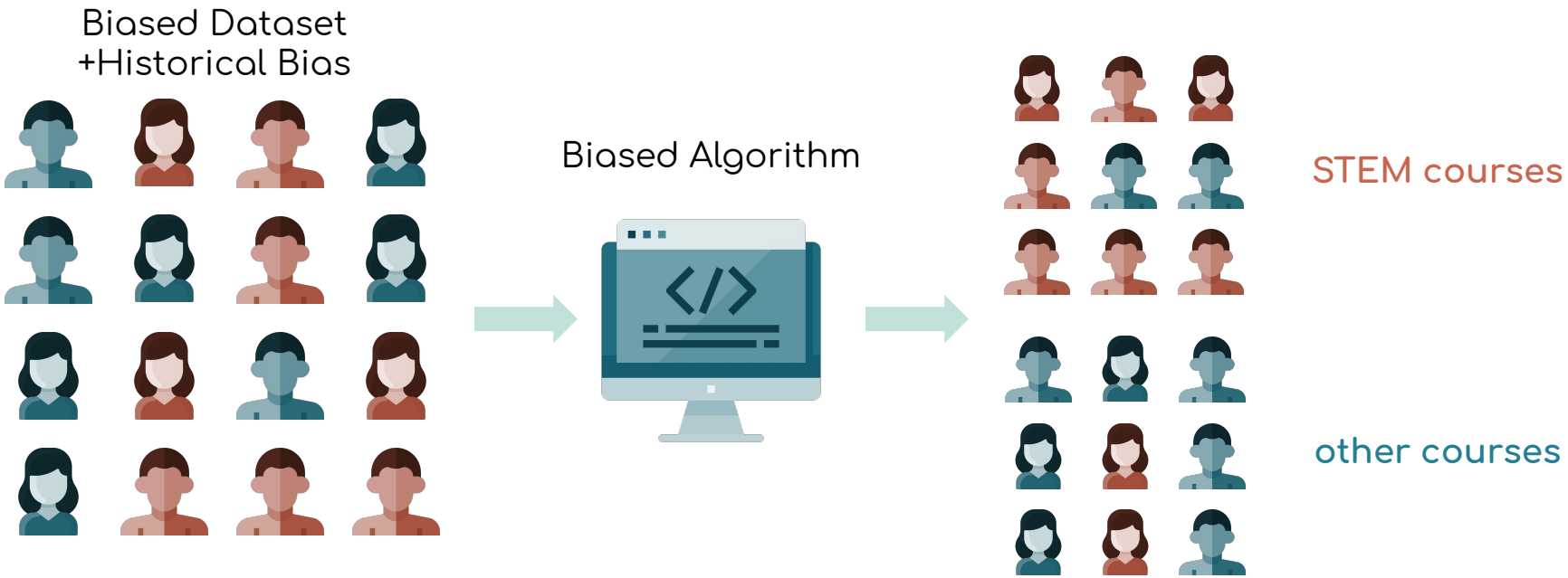
Hickey, James M., Pietro G. Di Stefano, and Vlasios Vasileiou. "Fairness by Explicability and Adversarial SHAP Learning." *arXiv preprint arXiv:2003.05330* (2020).

04

Where to apply it?

Literature Use Case

Gender Bias in Education Recommendation



Yao, Sirui, and Bert Huang. "Beyond parity: Fairness objectives for collaborative filtering." *Advances in Neural Information Processing Systems*. 2017..

Literature Use Case

Gender Bias in Education Recommendation

Value Unfairness

Measures inconsistency in signed prediction error across the user type.

Absolute Unfairness

Similar to Value Unfairness but does not consider the direction of the error

Underestimation Unfairness

Measures inconsistency in how much the predictions underestimate the true ratings.

Overestimation Unfairness

Measures inconsistency in how much the predictions overestimate the true ratings.

Non-parity unfairness

Measures the absolute difference between the overall average ratings of disadvantaged users and advantaged ones.

Yao, Sirui, and Bert Huang. "Beyond parity: Fairness objectives for collaborative filtering." *Advances in Neural Information Processing Systems*. 2017..

Literature Use Case

Gender Bias in Education Recommendation

Generated artificial datasets using uniform vs. unbalanced observations probability + unbalanced vs. balanced representations.

Trained matrix factorization algorithms with a regularization on fairness.

Tested the algorithm, evaluating the error and th metrics on fairness.

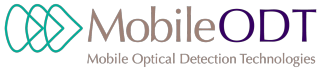
Conclusions:

- Regularizing on one of the metrics tends to decrease the other fairness metrics;
- Decreasing fairness metrics (decreasing unfairness) does minimized the prediction error;
- Regularization on the Value Unfairness was the most effective;

Yao, Sirui, and Bert Huang. "Beyond parity: Fairness objectives for collaborative filtering." *Advances in Neural Information Processing Systems*. 2017..

NILG.AI Use Cases

Fairness by Construction:



1. Project Goal:
   a. AI models that promote fair access to high-quality healthcare
   b. Data quality to prevent improper conclusions when decisions are provided by entry-level personnel
   c. See more at: https://tinyurl.com/modtsympbio

2. Data collection:
   a. Data collection that ensures representation of main protected groups:
      i. Gender, race
      ii. Others: staff skills, tatoos, room conditions
   b. See more on “Innovation in Medical Device Decontamination” at: https://tristelopenday.com/

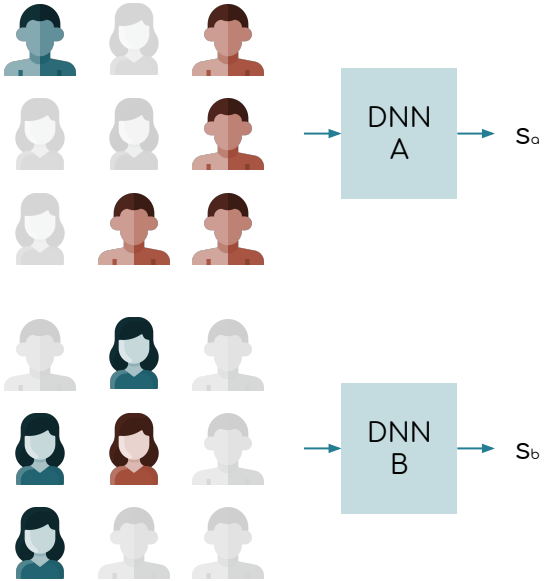
NILG.AI Use Cases

Group invariance:

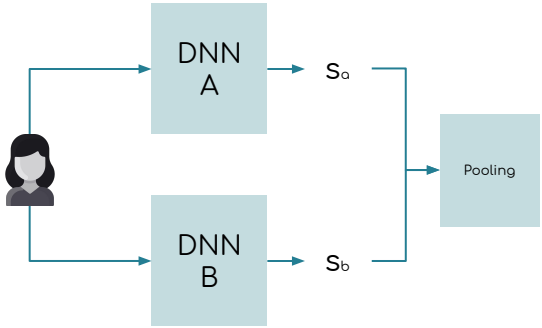
race, gender, country, deep-learning-framework-preference

Biased Dataset

Train a model per group



Avg. voting as prediction



nilg.ai

Thank you for your attention!