

# Discrimination in Machine Learning\*†

© Patrick Hall

H<sub>2</sub>O.ai

May 28, 2020

---

\* This material is shared under a [CC By 4.0 license](#) which allows for editing and redistribution, even for commercial purposes. However, any derivative work should attribute the author and H2O.ai.

† This presentation is not, and should not be construed as, legal advice or requirements for regulatory compliance.

Why?

- 
- 

What?

- 
- 

How?

- 
- 

Acknowledgements

- 

References

# Contents

Why?

What?

How?

Acknowledgements

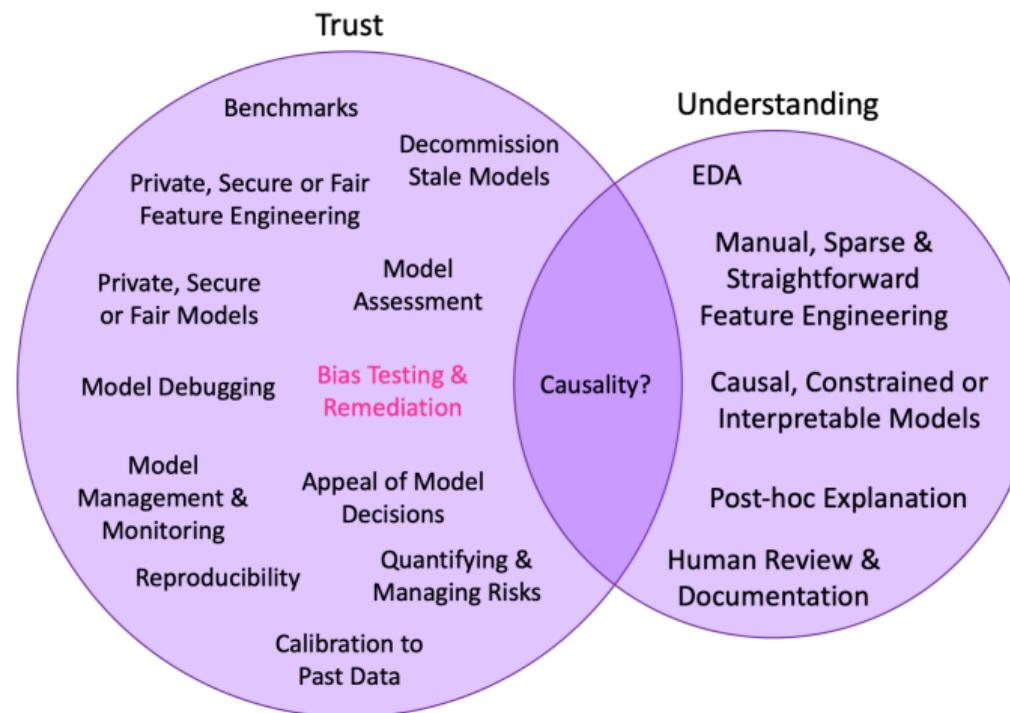
## Why Care About Discrimination in ML?

- **Reputational risk:** Upon encountering a perceived unethical ML system, 34% of consumers are likely to, "stop interacting with the company."<sup>‡</sup>
- Non-compliance fines and litigation costs.
- Responsible practice of ML.

---

<sup>‡</sup>See: Why addressing ethical questions in AI will benefit organizations.

# Elements of Responsible Practice of ML



## What Is Bias?

- Almost all data, statistical models, and machine learning (ML) models encode different types of *bias*, i.e., systematic misrepresentations of reality.
- Sometimes, bias is helpful, e.g. shrunken, robust  $\beta_j$  coefficients in penalized linear models.
- Other types of bias might be unwanted, unhelpful, or illegal discrimination.

# What is Discrimination in ML?

In some applications<sup>§</sup>, model predictions should *ideally* be independent of demographic group membership.

In these applications, a model exhibits discrimination if:

1. Demographic group membership is not independent of the likelihood of receiving a favorable or accurate model prediction.
2. Membership in a *subset* of a demographic group is not independent of the likelihood of receiving a favorable or accurate model prediction (i.e., *local bias*).[2]

Several forms of discrimination may manifest in ML, including:

- Overt discrimination, i.e. *disparate treatment*.
- Unintentional discrimination, i.e. *disparate impact* (DI).
- Discrimination in ML may or may not be illegal, depending on how it arises and applicable discrimination laws.[2]

---

<sup>§</sup>e.g., Under the Equal Credit Opportunity Act (ECOA), as implemented by Regulation B, and the Fair Credit Reporting Act (FCRA)

# What is Discrimination in ML?

Discrimination originates from training data:

- Incomplete or inaccurate data, particularly under-representation of minorities, e.g. [Gender Shades](#)[1].
- Accurate but differing patterns of causation, correlation, or dependency between demographic groups and past outcomes, e.g. traditional FICO credit scores.<sup>¶</sup>
- Explicit encoding of historical social biases into training data, e.g. criminal records.<sup>¶</sup>

---

<sup>¶</sup>See: [Responsible Data Science: Identifying and Fixing Biased AI](#).

# What is Discrimination in ML?

ML models can perpetuate or exacerbate discrimination.

**Group disparities**, i.e. different or inaccurate treatment of entire demographic groups:

- Learning different correlations between demographic groups and favorable model outcomes, i.e. *adverse impact*.
- Exhibiting different accuracies across demographic groups, i.e. *differential validity*.¶

**Locally**, i.e. different or inaccurate treatment of similar individuals:

- Local response function or decision boundary form.
- Capacity to form local complex or latent demographic proxies.

# What is Discrimination in ML?

Many kinds of group disparities can be measured, e.g.:

- Accuracy disparity:  $\frac{\text{accuracy}_p}{\text{accuracy}_r}$
- Adverse impact ratio:  $\frac{\% \text{ accepted}_p}{\% \text{ accepted}_r}$
- Marginal effect:  $\% \text{ accepted}_p - \% \text{ accepted}_r$
- Standardized mean difference:  $\frac{\bar{y}_p - \bar{y}_r}{\sigma_{\bar{y}}}$

where,  $p \equiv$  protected group and  $r \equiv$  reference group (often white males),

$$\% \text{ accepted}_{\text{group}} = 100 \cdot \frac{\text{tn}_{\text{group}} + \text{fn}_{\text{group}}}{N_{\text{group}}}, \text{ and } \text{accuracy}_{\text{group}} = \frac{\text{tp}_{\text{group}} + \text{tn}_{\text{group}}}{N_{\text{group}}}.$$

Local bias is much trickier to measure ... and often an unmitigated risk for consumer-facing ML systems.

# How to Fix Discrimination in ML?

**Fix the process:** ensure diversity of experience in design, training, and review of ML systems.

**Fix the data:**

- Collect demographically representative training data.
- Select features judiciously, e.g. using `time_on_file` as an input variable as opposed to `bankruptcy_flag`.<sup>¶</sup>
- Sample and reweigh training data to minimize discrimination.<sup>[3]</sup>

**Fix the model:**

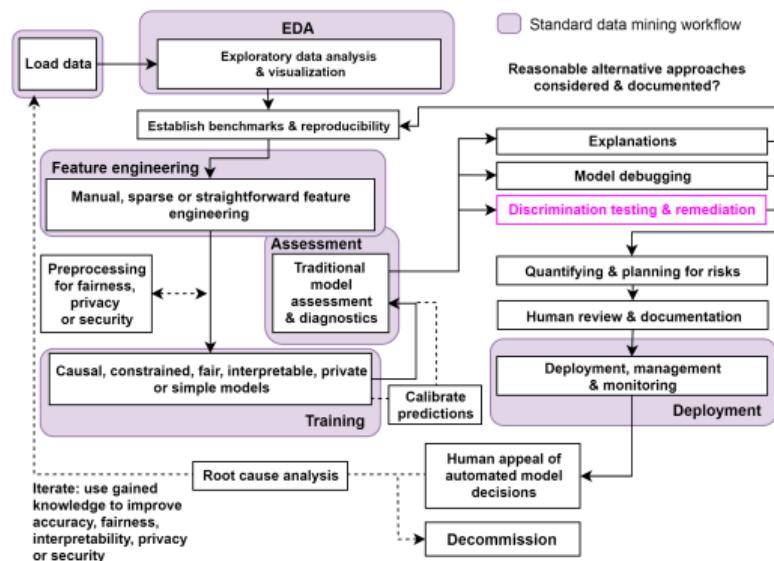
- Consider fairness metrics when selecting hyperparameters and cutoff thresholds.
- Train fair models directly:
  - Learning fair representations (LFR) and adversarial de-biasing.<sup>[6], [7]</sup>
  - Use dual objective functions that consider both accuracy and fairness metrics.
- Edit model mechanisms to ensure less biased predictions, e.g. with **GA2M** models.

**Fix the predictions:**

- Balance model predictions, e.g. reject-option classification.<sup>[4]</sup>
- Correct or override predictions with model assertions or appeal mechanisms.<sup>[2], [5]</sup>

# How to Fix Discrimination in ML?

As part of a responsible ML workflow.



# Acknowledgements

This presentation borrows heavily from the expertise of Nicholas Schmidt and Bryce Stephens of **BLDS, LLC**, a leading fair lending compliance firm.

## References

Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In: *Conference on Fairness, Accountability and Transparency*. URL: <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>. 2018, pp. 77–91.

Patrick Hall, Navdeep Gill, and Nicholas Schmidt. “Guidelines for the Responsible Use of Explainable Machine Learning.” In: *arXiv preprint arXiv:1906.03533* (2019). URL: <https://arxiv.org/pdf/1906.03533.pdf>.

Faisal Kamiran and Toon Calders. “Data Preprocessing Techniques for Classification Without Discrimination.” In: *Knowledge and Information Systems* 33.1 (2012). URL: <https://bit.ly/21H951Q>, pp. 1–33.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. “Decision Theory for Discrimination-aware Classification.” In: *2012 IEEE 12th International Conference on Data Mining*. URL: <http://citeserx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf>. IEEE. 2012, pp. 924–929.

Daniel Kang et al. *Debugging Machine Learning Models via Model Assertions*. URL: [https://www-cs.stanford.edu/~matei/papers/2018/mlsys\\_model\\_assertions.pdf](https://www-cs.stanford.edu/~matei/papers/2018/mlsys_model_assertions.pdf). 2019.

## References

Rich Zemel et al. "Learning Fair Representations." In: *International Conference on Machine Learning*. URL: <http://proceedings.mlr.press/v28/zemel13.pdf>. 2013, pp. 325–333.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. URL: <https://arxiv.org/pdf/1801.07593.pdf>. ACM. 2018, pp. 335–340.