

# Responsible Machine Learning\*

## Lecture 3: Discrimination Testing and Remediation

Patrick Hall

The George Washington University

July 30, 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.

# Contents

Introduction

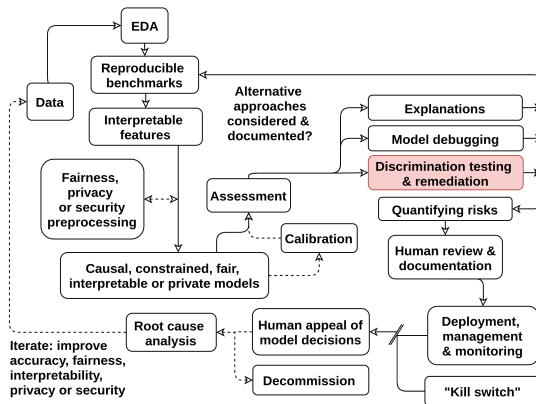
Bias and Discrimination

Testing for Discrimination in ML

Remediation

Acknowledgements

# A Responsible Machine Learning Workflow<sup>†</sup>



## † A Responsible Machine Learning Workflow

## Why Care About Discrimination in Machine Learning?

- **Responsible practice of machine learning (ML):** ML can affect millions of people! [7]
- **Discrimination is often illegal (in the U.S.):** Non-compliance fines and litigation costs.
- **Reputational risk:** Upon encountering a perceived unethical ML system, 34% of consumers are likely to, “stop interacting with the company.”<sup>‡</sup>

---

<sup>‡</sup>See: [Why addressing ethical questions in AI will benefit organizations.](#)

# What Is Bias?

- Almost *all* data, statistical models, and ML models encode different types of *bias*, i.e., **systematic misrepresentations of reality**.
- Sometimes, bias is *helpful*.
  - Shrunk and robust  $\beta_j$  coefficients in penalized linear models.
- Other types of bias can be unwanted, unhelpful, discriminatory, or illegal.
- Many instances of discrimination in ML arise from sociologically biased data collection, labeling, or storage processes.

## What is Discrimination in ML?

In many 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 or individual discrimination*).[\[3\]](#)

---

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

## What Kinds of Discrimination Occur in ML?

Several forms of discrimination may manifest in ML, including:

- Group disparities:
  - Overt discrimination against groups, i.e., *disparate treatment*.
  - Unintentional discrimination against groups, i.e., *disparate impact* (DI).
- Local or individual discrimination.

# How Does Discrimination Arise in ML?

Discrimination originates from training data:

- Incomplete or inaccurate data, e.g., under-representation of minorities. See [Gender Shades \[2\]](#).
- 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](#).



# How Does Discrimination Arise 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., *DI*.
- Exhibiting different accuracies across demographic groups, i.e., *differential validity*.<sup>¶</sup>

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

- Local response function or decision boundary form.
- Capacity to form local complex demographic proxies on a row-by-row basis.

## Common Metrics of Discrimination in ML

Common metrics for DI and **group** disparities:

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

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

There are many other, sometimes conflicting, mathematical definitions of discrimination. See [21 Definitions of Fairness and Their Politics](#).

## Additional Considerations for Discrimination Testing

- Local discrimination, i.e., the model treats a small number of similar people differently.
  - Search around probability thresholds.
  - Adversarial models.
- Post-hoc explanation to understand drivers of discrimination:
  - To be conducted after discrimination is confirmed by standard tests.
  - Be aware of:
    - No demographic features in model.
    - Fairwashing [1] and scaffolding [8].

## How to Fix Discrimination in ML?

**Fix organizational processes:** Lecture 6

**Fix the data:**

- Collect demographically representative training data.
- Label and annotate data carefully.
- Select features judiciously.
- Sample and reweigh training data to minimize discrimination.[4]

# How to Fix Discrimination in ML?

## Fix the model:

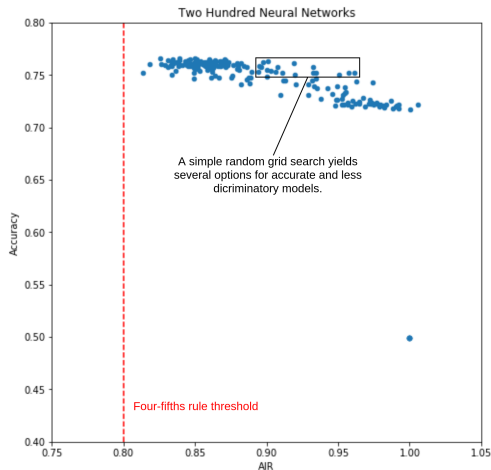
- Consider fairness metrics when selecting hyperparameters and cutoff thresholds.
- Train fair models directly:
  - Learning fair representations (LFR) and adversarial de-biasing.[9], [10]
  - Use dual objective functions that consider both accuracy and fairness metrics.
- Edit model mechanisms to ensure less biased predictions, e.g., with GA2M/EBM models.

## Fix the predictions:

- Balance model predictions, e.g., reject-option classification.[5]
- Correct or override predictions with model assertions or appeal mechanisms.[3], [6]

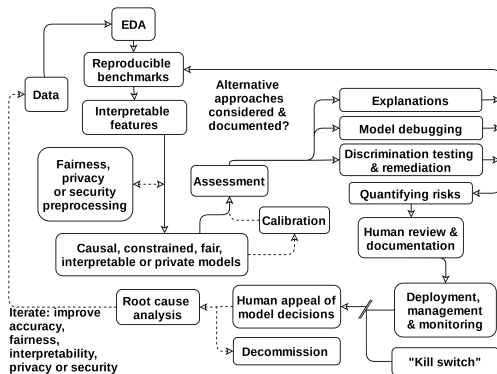
# How to Fix Discrimination in ML?

Consider discrimination measures during model selection.



# 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.

Thanks to Lisa Song for her continued assistance in developing these course materials.

Some materials ©Patrick Hall and the H2O.ai team 2017-2020.



## References

Ulrich Aïvodji et al. “Fairwashing: the Risk of Rationalization.” In: *arXiv preprint arXiv:1901.09749* (2019). URL: <https://arxiv.org/pdf/1901.09749.pdf>.

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/2lH95lQ>, 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://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf>. IEEE. 2012, pp. 924–929.

## References

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.

Ziad Obermeyer et al. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” In: *Science* 366.6464 (2019). URL: <https://bit.ly/36XK6yk>, pp. 447–453.

Dylan Slack et al. “How Can We Fool LIME and SHAP? Adversarial Attacks on Post-hoc Explanation Methods.” In: *arXiv preprint arXiv:1911.02508* (2019). URL: <https://arxiv.org/pdf/1911.02508.pdf>.

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.