RESPONSIBLE AI

Week 6: Replication Project 2 – Fairness Assessments

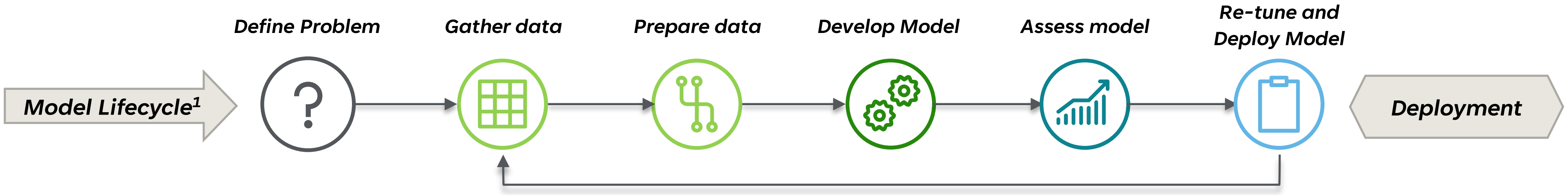
Meira Gilbert and Nandita Rahman

TODAY'S OBJECTIVES

- How to build a model?
- How can we assess algorithmic fairness? What metrics are commonly used?
- What can't be captured in data, and what are the limitations of fairness metrics?

REPLICATION PROJECT: RESPONSIBLE AI IN ACTION

OPERATIONALIZING RESPONSIBLE AI ISN'T JUST USING CODE.
IT ALSO REQUIRES HUMANS IN THE LOOP FOR EACH STAGE OF THE MODEL LIFECYCLE



Start: Week 03 (10/14/22)

Part 0: Overview of Data, Scope, and Preparation

- Dataset
- Scope Definition

Week 04 (10/21/22)

Part 1: Exploratory Data Analysis

- Review of protected classes
- Analysis insights and explanation
- Data imbalances

Week 06 (11/4/22)

Part 2: Model Development and Fairness Evaluation (Bias Detection)

- Model Development
- Bias detection metrics
- Fairness metrics, assessment and visualizations

Week 07 (TBD*)

Part 3: Bias Mitigation Techniques

- Pre, in, post-processing bias mitigation methods
- Discussion

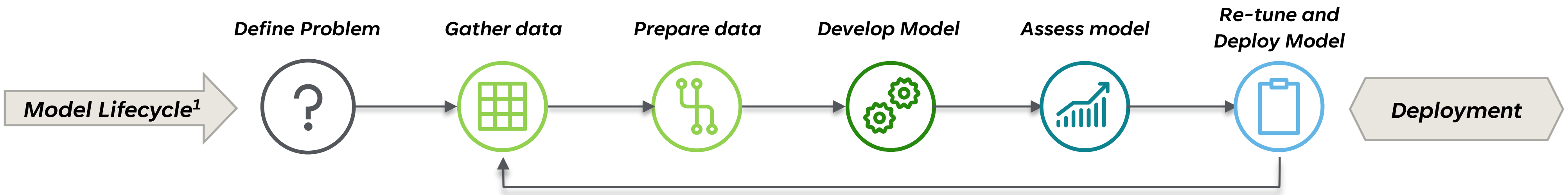
End: Week 09 (12/02/22)

Part 4: Presentation, Conclusions and Explainability

- Local, Global Explainability Results
- Explainability integration into Model Lifecycles
- Future perspectives on explainability

REPLICATION PROJECT PART 02

MODEL DEVELOPMENT & FAIRNESS EVALUATION



Model Lifecycle[1]

Define Problem

Gather data

Prepare data

Develop Model

Assess model

Re-tune and Deploy Model

Deployment

- **Model Development and Fairness Evaluation:**
  After your Exploratory Data Analysis (Part 01), you'll now train your data and assess fairness metrics without de-biasing.

- There will be **TWO PARTS** to this portion of the replication project.
  (1) Training models without de-biasing, using IBM's tutorial
  (2) Training models without de-biasing, using any insights gained from your EDA step in Part 01. In addition, apply any model development techniques including (1) Feature Selection, (2) Encoding, and others that you may have learned in your methodology portion of this course.

Part 1

Use IBM AIF360's Tutorial to run training and testing, and capture fairness metrics

Evaluate whether the model performs sufficiently for production.

Does the model answer the question with sufficient confidence given the test data?

Part 2

Based on your EDA results should you collect additional data, do feature engineering, or experiment with other algorithms?

Use visualizations to understand your model

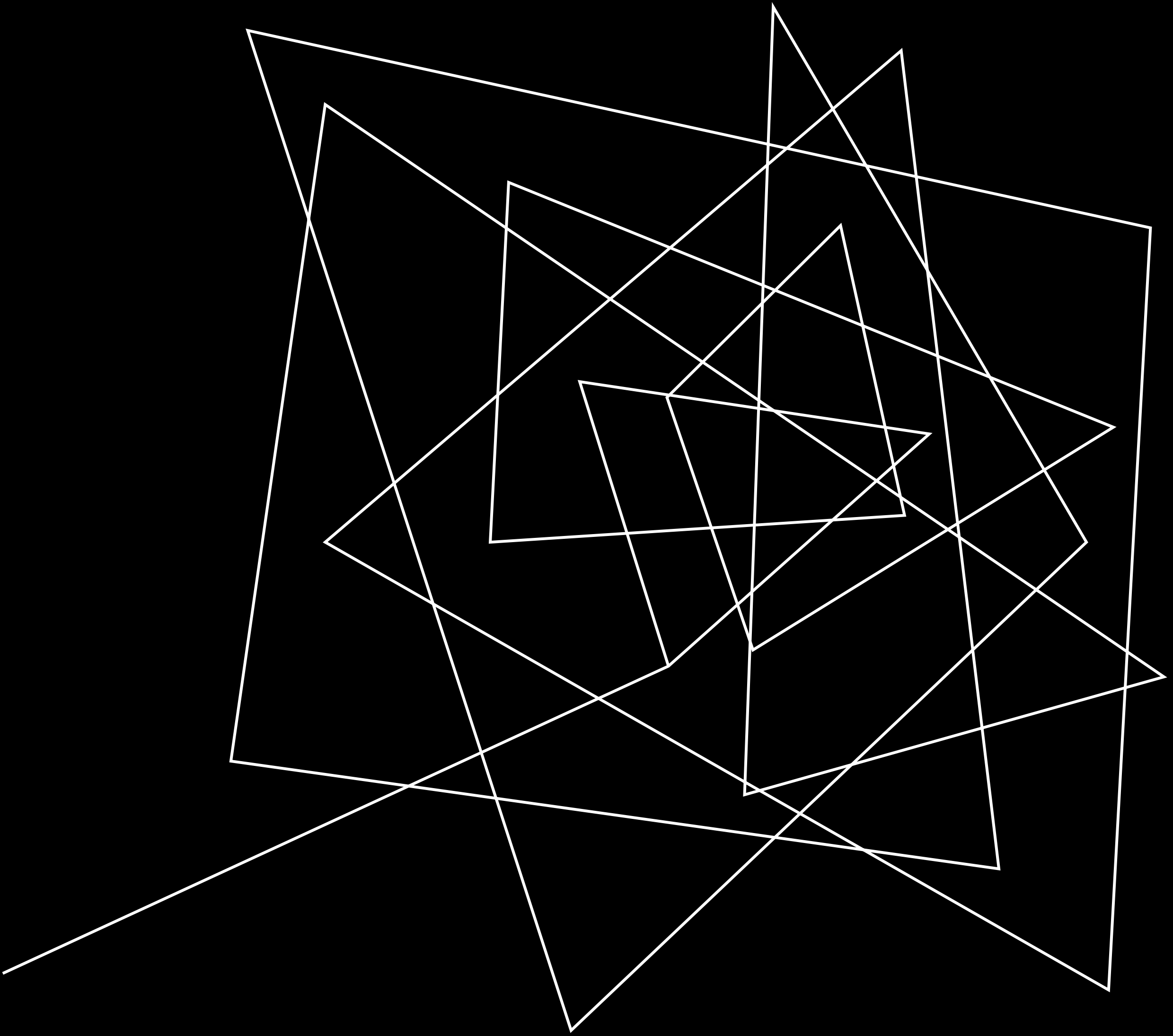
Rerun the same analysis you completed in Part 1 to capture fairness metrics.

Assessing Fairness

Explain the fairness of your model predictions using your fairness metric results.

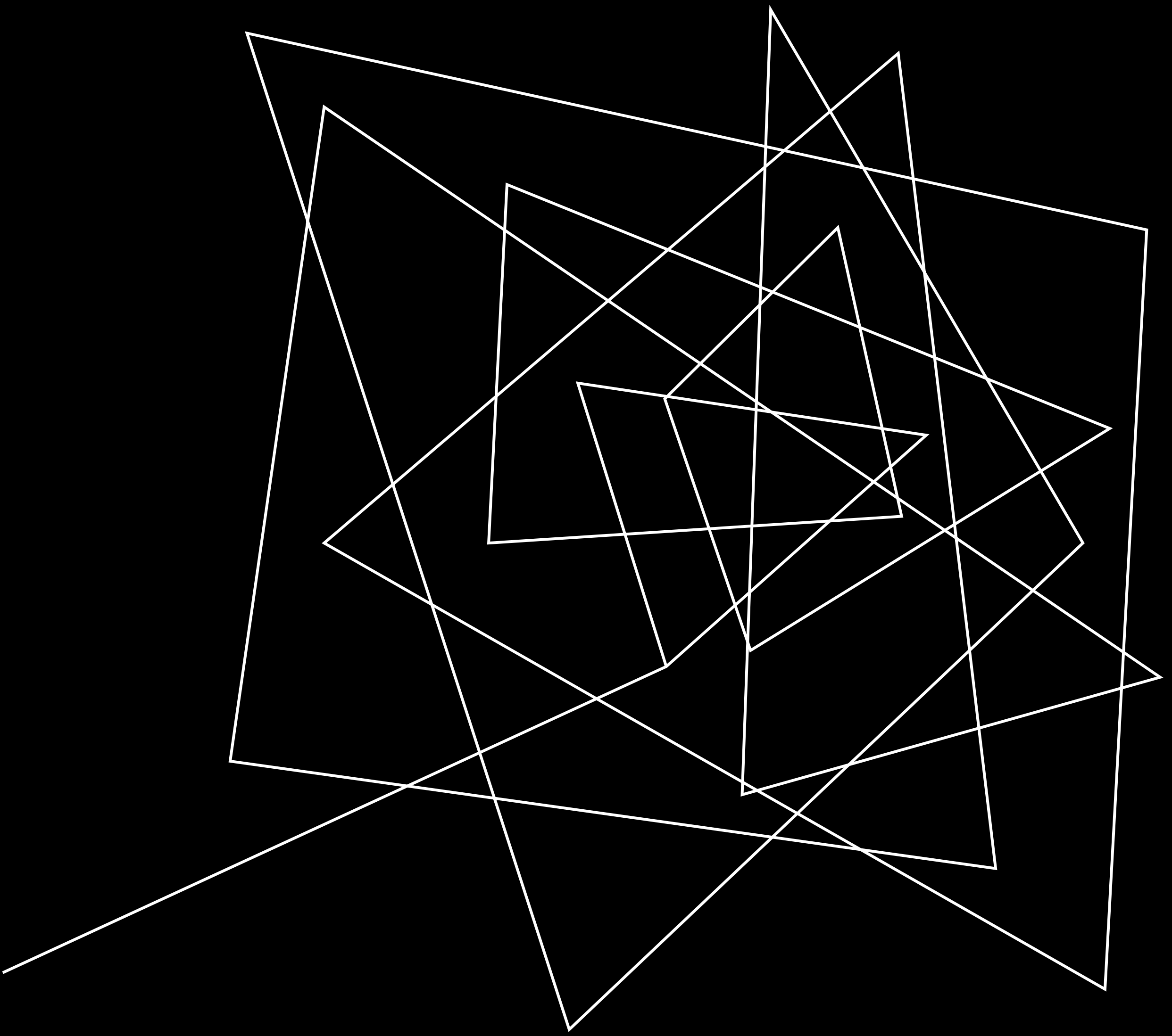
Compare and contrast the differences between the fairness metrics results for Part 1 and Part 2.

Based on your results, provide your initial judgement for which type of **model** and **fairness metrics** seem appropriate to use.



READING PRESENTATION #1

Ethical Machine Learning in Health Care (Chen et al. 2021)



READING PRESENTATION #2

Millions of black people affected by racial bias in health-care algorithms (Nature)

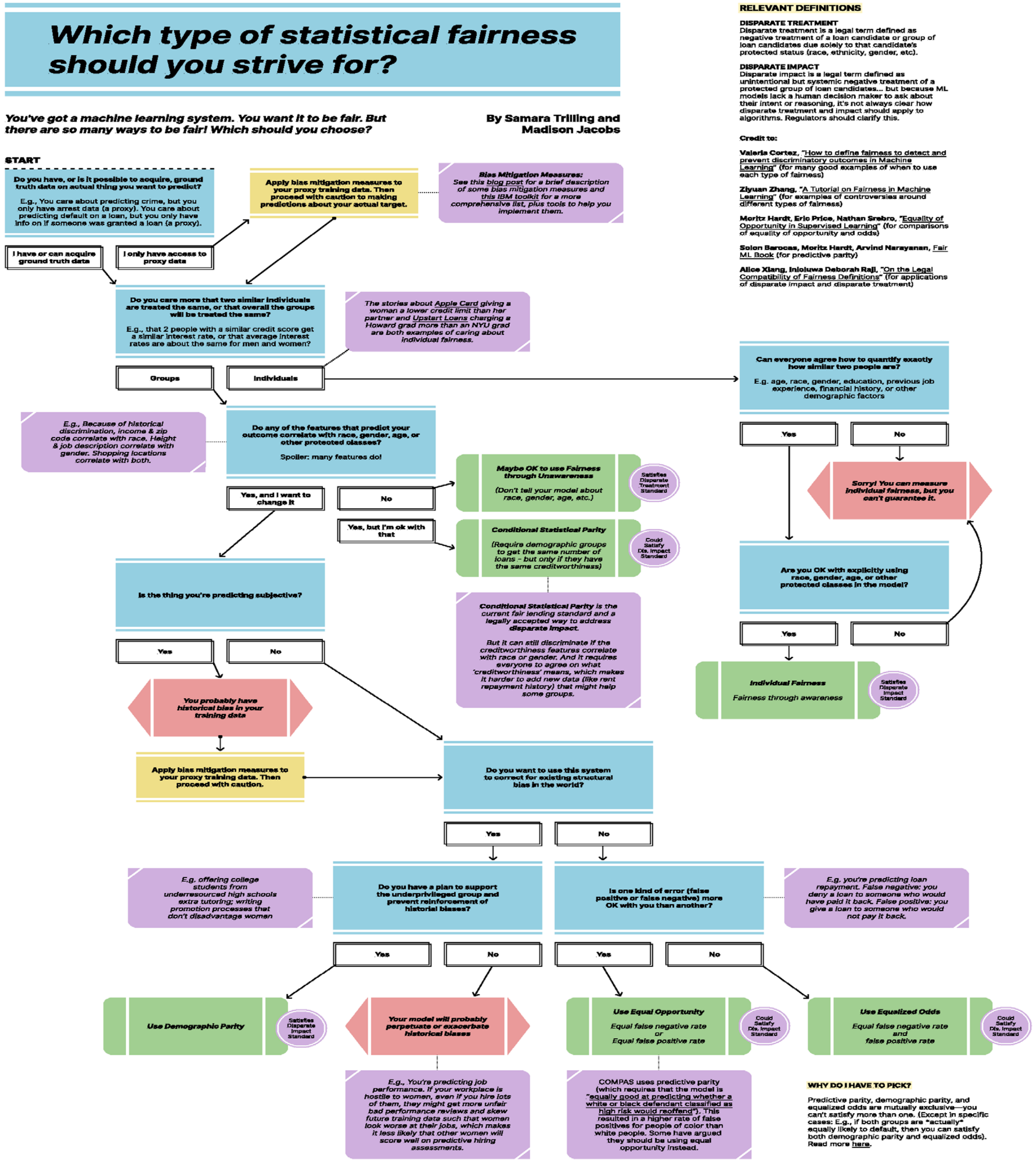
FAIRNESS METRICS

Statistical Parity / Demographic Parity	Equal proportion of outcomes between groups regardless of other factors
Statistical Parity Difference	Computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group
Equalized Odds	Equal false negative rates and false positive rates
Equal Opportunity	Equal false negative or false positive rates between groups
Equal Opportunity Difference	This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.
Average Odds Difference	Computed as average difference of false positive rate (false positives / negatives) and true positive rate (true positives/positives) between unprivileged and privileged groups.
Disparate Impact	Computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.
Theil Index	Computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1. It measures the inequality in benefit allocation for individuals.

How can we determine which metrics to use, given our data and use case?

When you have competing fairness metrics, how to pick which to prioritize?

What do you do when you encounter different definitions for similar metrics?



Which type of statistical fairness should you strive for?

You've got a machine learning system. You want it to be fair. But there are so many ways to be fair! Which should you choose?

By Samara Trilling and Madison Jacobs

START

Do you have, or is it possible to acquire, ground truth data on actual thing you want to predict?

E.g., You care about predicting crime, but you only have arrest data (a proxy). You care about predicting default on a loan, but you only have info on if someone was granted a loan (a proxy).

I have or can acquire ground truth data

I only have access to proxy data

Apply bias mitigation measures to your proxy training data. Then proceed with caution to making predictions about your actual target.

Bias Mitigation Measures: *See this blog post for a brief description of some bias mitigation measures and this IBM toolkit for a more comprehensive list, plus tools to help you implement them.*

Do you care more that two similar individuals are treated the same, or that overall the groups will be treated the same?

E.g., that 2 people with a similar credit score get a similar interest rate, or that average interest rates are about the same for men and women?

The stories about Apple Card giving a woman a lower credit limit than her partner and Upstart Loans charging a Howard grad more than an NYU grad are both examples of caring about individual fairness.

Groups

Individuals

Do any of the features that predict your outcome correlate with race, gender, age, or other protected classes?

Spoiler: many features do!

E.g., Because of historical discrimination, income & zip code correlate with race. Height & job description correlate with gender. Shopping locations correlate with both.

Yes, and I want to change it

No

Yes, but I'm ok with that

Maybe OK to use Fairness through Unawareness

(Don't tell your model about race, gender, age, etc.)

Satisfies Disparate Treatment Standard

Conditional Statistical Parity

(Require demographic groups to get the same number of loans – but only if they have the same creditworthiness)

Could Satisfy Dis. Impact Standard

***Conditional Statistical Parity** is the current fair lending standard and a legally accepted way to address **disparate impact**.*

But it can still discriminate if the creditworthiness features correlate with race or gender. And it requires everyone to agree on what 'creditworthiness' means, which makes it harder to add new data (like rent repayment history) that might help some groups.

Is the thing you're predicting subjective?

Yes

No

You probably have historical bias in your training data

Apply bias mitigation measures to your proxy training data. Then proceed with caution.

Do you want to use this system to correct for existing structural bias in the world?

Yes

No

Do you have a plan to support the underprivileged group and prevent reinforcement of historial biases?

E.g. offering college students from underresourced high schools extra tutoring; writing promotion processes that don't disadvantage women

Yes

No

Use Demographic Parity

Satisfies Disparate Impact Standard

Your model will probably perpetuate or exacerbate historical biases

E.g., You're predicting job performance. If your workplace is hostile to women, even if you hire lots of them, they might get more unfair bad performance reviews and skew future training data such that women look worse at their jobs, which makes it less likely that other women will score well on predictive hiring assessments.

Is one kind of error (false positive or false negative) more OK with you than another?

E.g. you're predicting loan repayment. False negative: you deny a loan to someone who would have paid it back. False positive: you give a loan to someone who would not pay it back.

Yes

No

Use Equal Opportunity

Equal false negative rate
or
Equal false positive rate

Could Satisfy Dis. Impact Standard

COMPAS uses predictive parity (which requires that the model is "equally good at predicting whether a white or black defendant classified as high risk would reoffend"). This resulted in a higher rate of false positives for people of color than white people. Some have argued they should be using equal opportunity instead.

Use Equalized Odds

Equal false negative rate
and
false positive rate

Could Satisfy Dis. Impact Standard

Can everyone agree how to quantify exactly how similar two people are?

E.g. age, race, gender, education, previous job experience, financial history, or other demographic factors

Yes

No

Sorry! You can measure individual fairness, but you can't guarantee it.

Are you OK with explicitly using race, gender, age, or other protected classes in the model?

Yes

No

Individual Fairness

Fairness through awareness

Satisfies Disparate Impact Standard

RELEVANT DEFINITIONS

DISPARATE TREATMENT
Disparate treatment is a legal term defined as negative treatment of a loan candidate or group of loan candidates due solely to that candidate's protected status (race, ethnicity, gender, etc).

DISPARATE IMPACT
Disparate impact is a legal term defined as unintentional but systemic negative treatment of a protected group of loan candidates... but because ML models lack a human decision maker to ask about their intent or reasoning, it's not always clear how disparate treatment and impact should apply to algorithms. Regulators should clarify this.

Credit to:

Valeria Cortez, "How to define fairness to detect and prevent discriminatory outcomes in Machine Learning" (for many good examples of when to use each type of fairness)

Ziyuan Zhang, "A Tutorial on Fairness in Machine Learning" (for examples of controversies around different types of fairness)

Moritz Hardt, Eric Price, Nathan Srebro, "Equality of Opportunity in Supervised Learning" (for comparisons of equality of opportunity and odds)

Solon Barocas, Moritz Hardt, Arvind Narayanan, Fair ML Book (for predictive parity)

Alice Xiang, Inioluwa Deborah Raji, "On the Legal Compatibility of Fairness Definitions" (for applications of disparate impact and disparate treatment)

WHY DO I HAVE TO PICK?

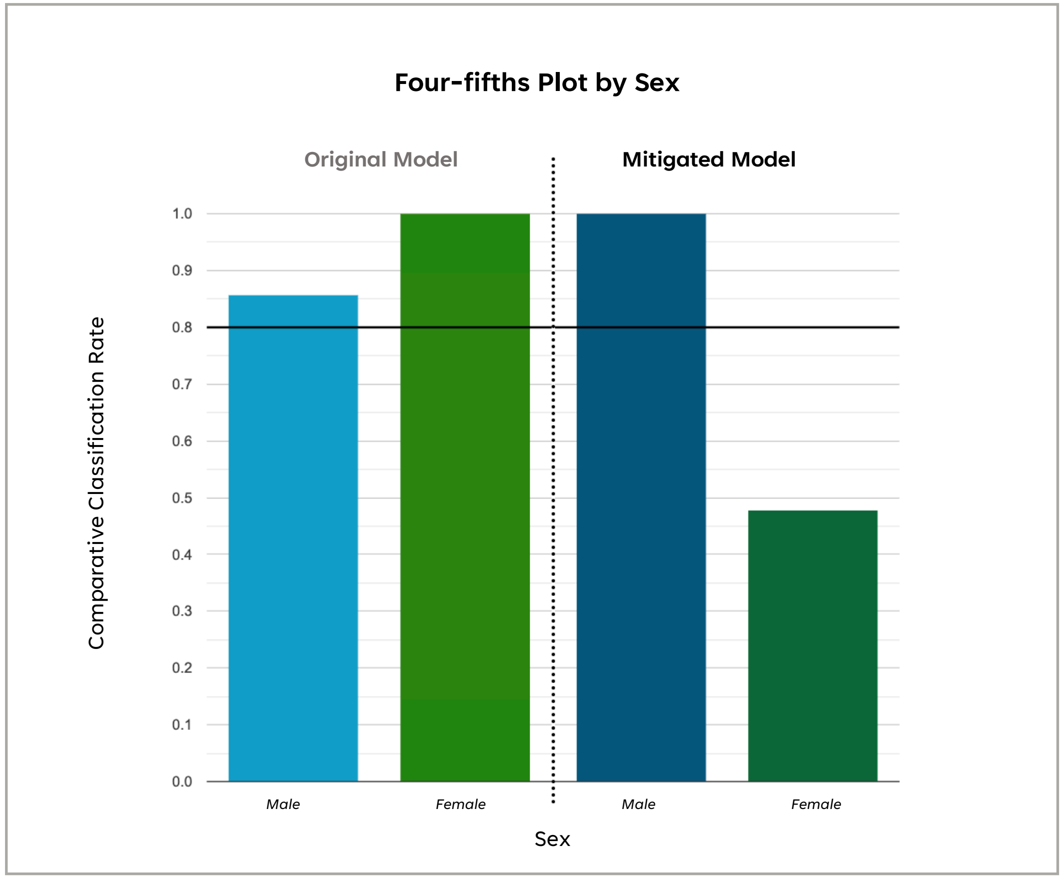
Predictive parity, demographic parity, and equalized odds are mutually exclusive—you can't satisfy more than one. (Except in specific cases: E.g., if both groups are *actually* equally likely to default, then you can satisfy both demographic parity and equalized odds). Read more here.

https://www.aspentechpolicyhub.org/wp-content/uploads/2020/07/FAHL-Tree.pdf

FAIRNESS PLOT EXAMPLES

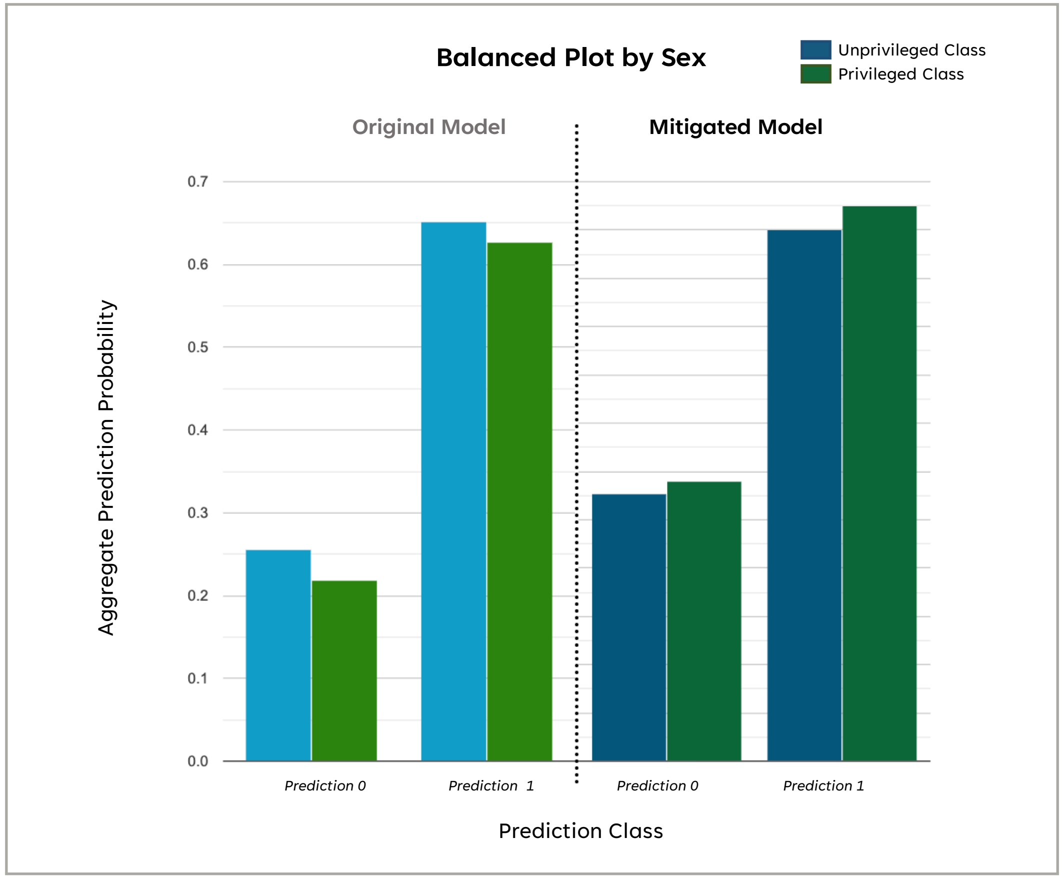
Four Fifths Plot

Identifies if there is adverse impact for unprivileged groups in comparison to the group with the highest selection rate.



Balance Plot

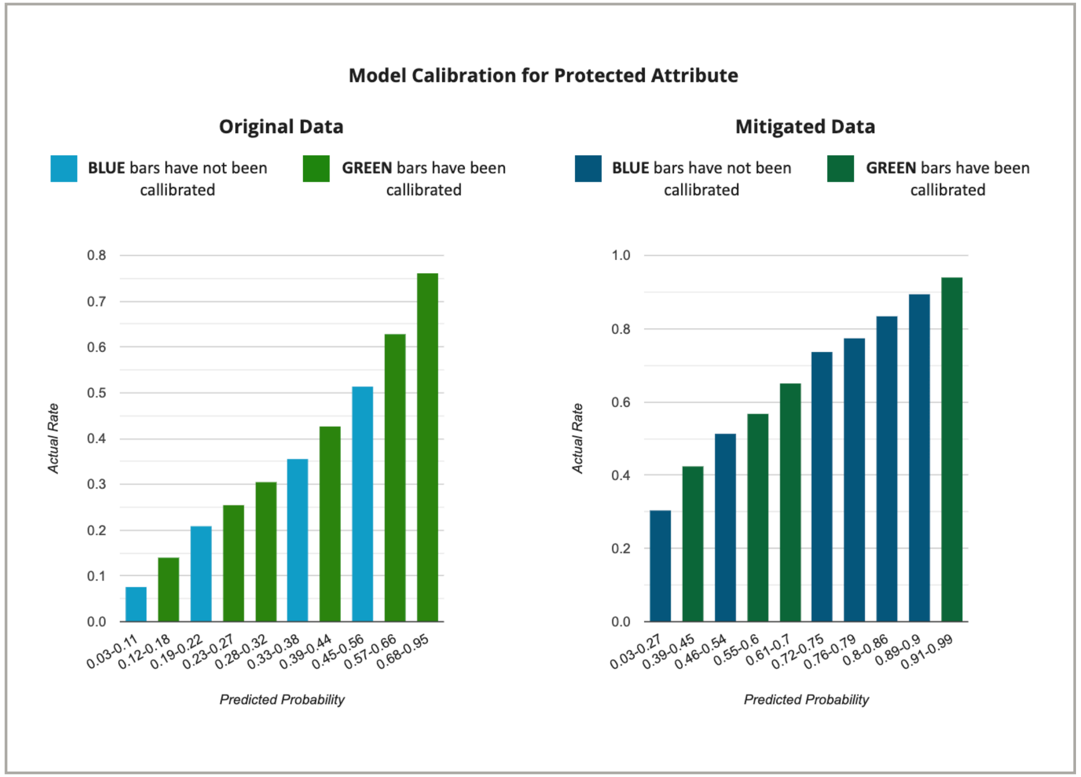
Examines whether average score received by individuals in positive and negative instances are similar regardless of sensitive attributes.



OTHER FAIRNESS VISUALIZATIONS (*IMPUTED*)

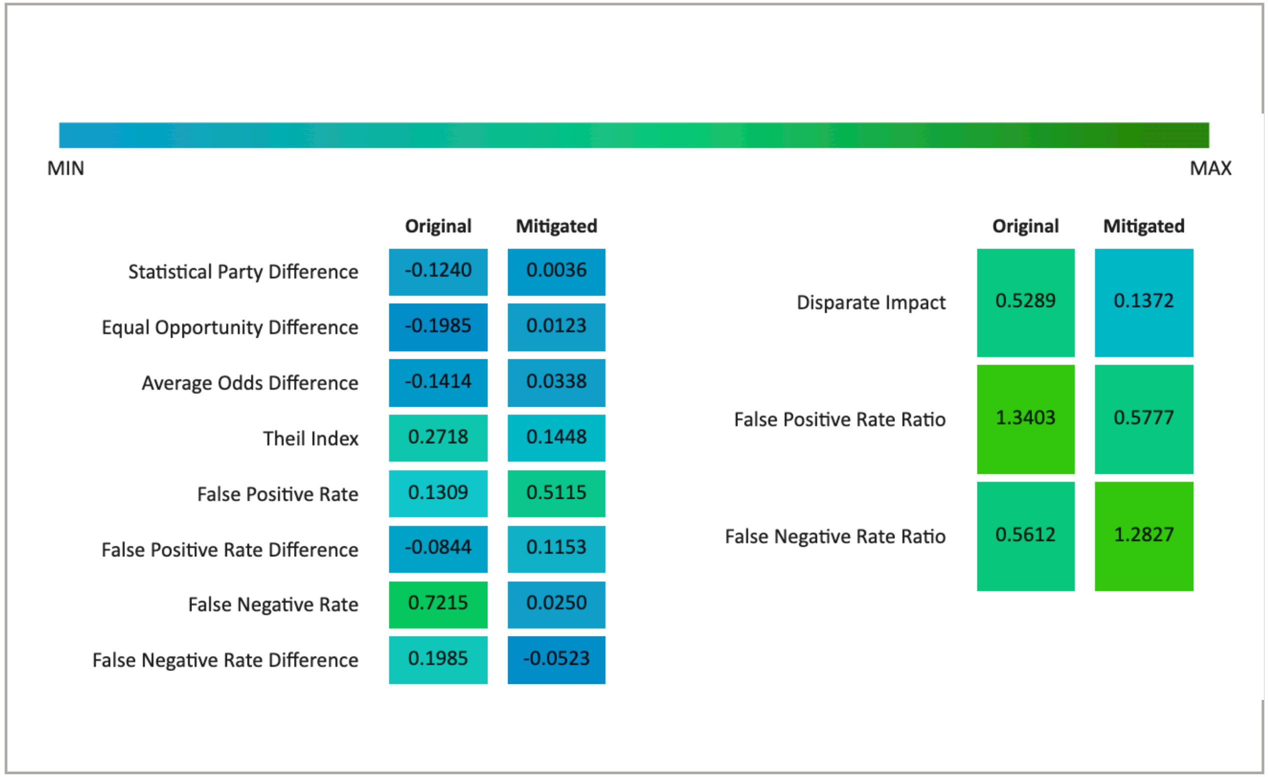
Calibration Plots

Checks if model makes accurate predictions in aggregate for members of each class.



Fairness Ratios

A selection of metrics is designed to help data scientists detect and evaluate bias within AI models.



MIN MAX

	Original	Mitigated
Statistical Party Difference	-0.1240	0.0036
Equal Opportunity Difference	-0.1985	0.0123
Average Odds Difference	-0.1414	0.0338
Theil Index	0.2718	0.1448
False Positive Rate	0.1309	0.5115
False Positive Rate Difference	-0.0844	0.1153
False Negative Rate	0.7215	0.0250
False Negative Rate Difference	0.1985	-0.0523

	Original	Mitigated
Disparate Impact	0.5289	0.1372
False Positive Rate Ratio	1.3403	0.5777
False Negative Rate Ratio	0.5612	1.2827

DISCUSSION QUESTIONS

- Which fairness metrics do you think would be most valuable for our use case (predicting utilization)?
- Which fairness metrics are you still confused or concerned about? Why?
- What visualizations are you interested in creating, based off your EDA and understanding of the case?

“ZOOMING OUT” - *Health Expenditure Use Case*

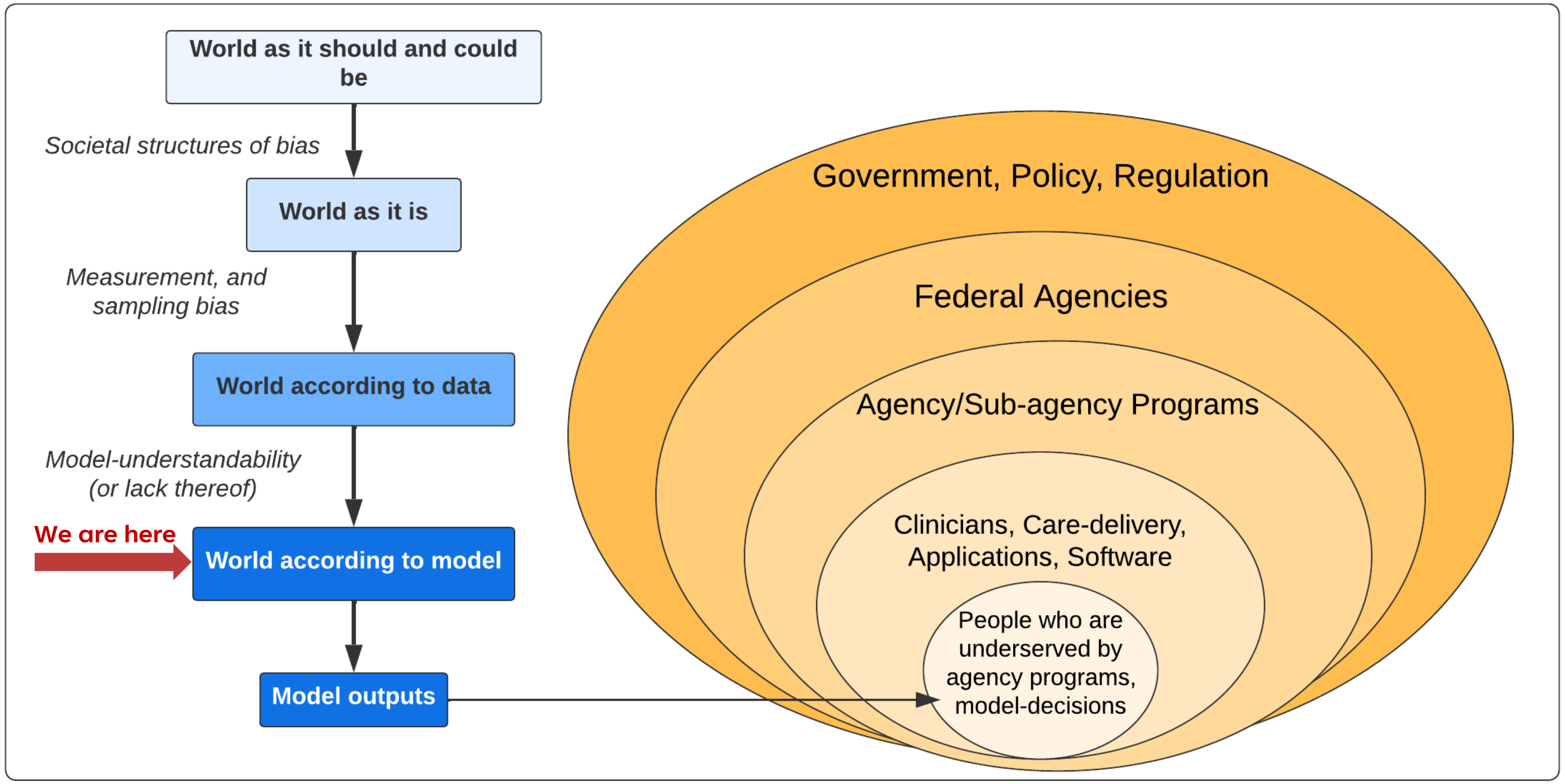

World as it should and could be
Societal structures of bias
World as it is
Measurement, and sampling bias
World according to data
Model-understandability (or lack thereof)
We are here
World according to model
Model outputs
Government, Policy, Regulation
Federal Agencies
Agency/Sub-agency Programs
Clinicians, Care-delivery, Applications, Software
People who are underserved by agency programs, model-decisions

Figure above: Illustration of the types of bias that can enter a model’s decision-space (stages of data and algorithm use), resulting in model-outputs that affect individuals downstream. *(Adapted from: Mhasawadade 2020, Mitchell 2020)*

Bias and Algorithmic Fairness

Use Cases	• Why are we implementing this? Who are affected parties?
Build	• What are the environmental costs of AI? Labor? Human Rights?
Deployment	• Once used, are the outcomes of the AI fair, equitable, and impartial? How can we measure this?

FOR NEXT WEEK

- Next week's class will be held **on THURSDAY, November 10th 12:30 PST!**
- Complete next week's **readings**
  - If you signed up to present **Model Cards for Model Reporting (Mitchell et al.)** or **Fairness Through Awareness (Dwork et al.)** come prepared to present next week and submit your presentation to Gradescope by 10 AM PT, Thursday, November 10th

- Submit your answers to next week's participation questions on Gradescope by 10 PM PT, Thursday, November 10th

- **Replication Part #2:** Notebook and writeup on model development and fairness metric assessments
  - Primary contact for replication project: Nandita Rahman (nanrahman@deloitte.com)
  - Office hours: Mondays 1-2pm PST
  - Notebook link here

- Office Hours for Replication Project
  - Mondays from 1-2pm PT (4-5pm ET)
  - Nandita hosting virtually - Zoom link