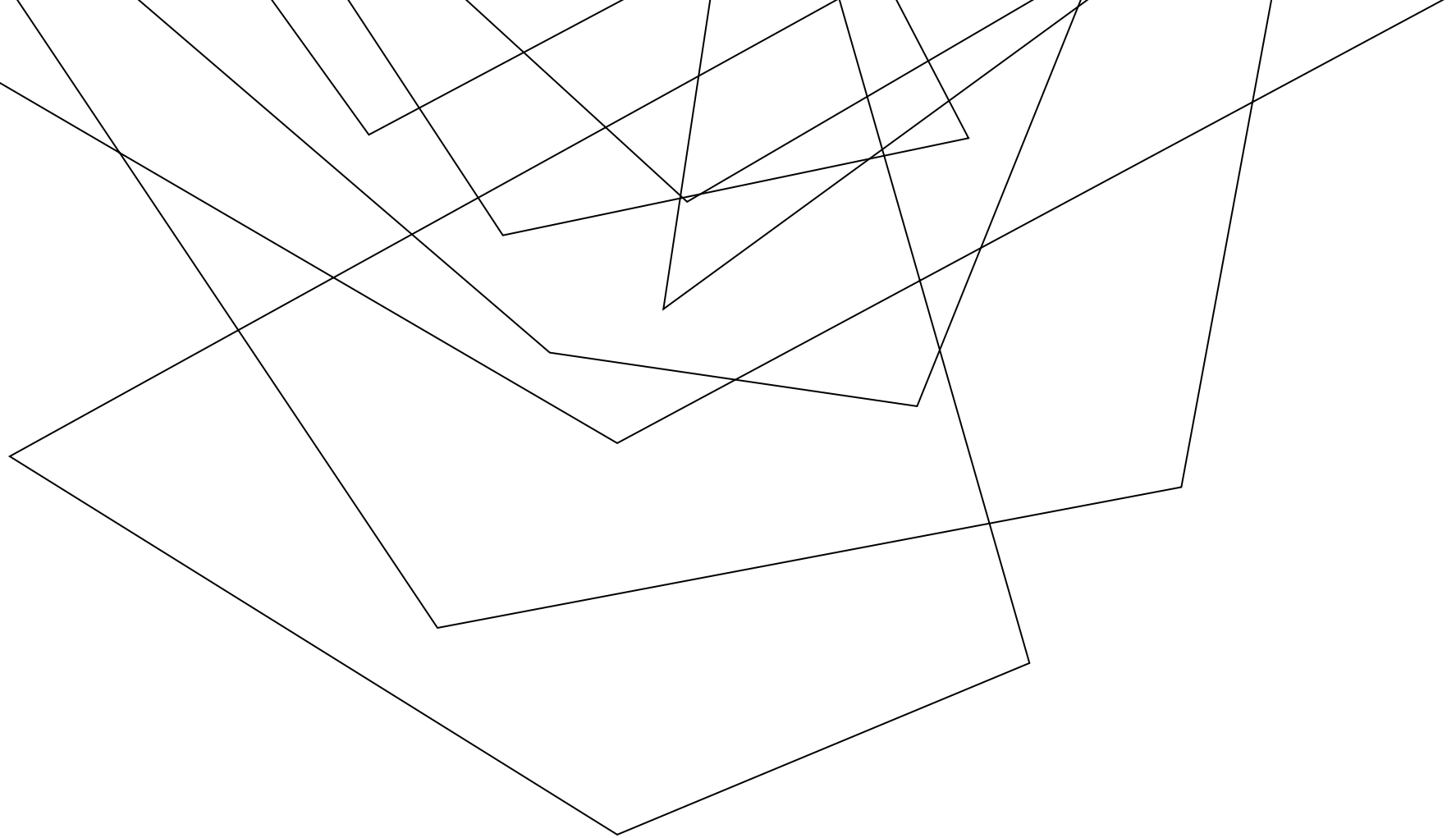


RESPONSIBLE AI

Week 6: Replication Project 2 – Fairness Assessments

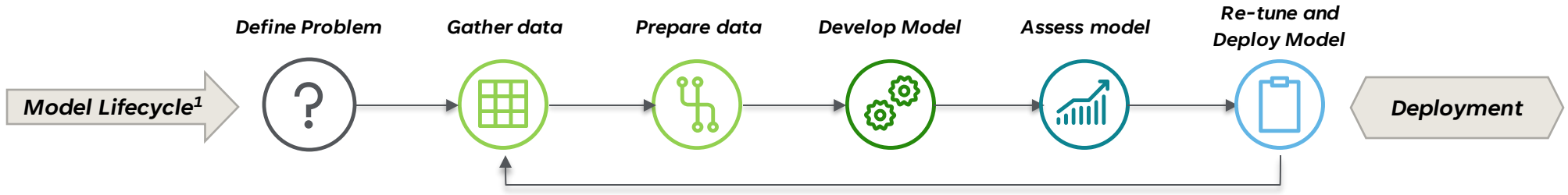
Credit: Meira Gilbert and Nandita Rahman

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

Part 0: Overview of Data, Scope, and Preparation

- Dataset
- Scope Definition

Week 04

Part 1: Exploratory Data Analysis

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

Week 06

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

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

Week 07

Part 3: Bias Mitigation Techniques

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

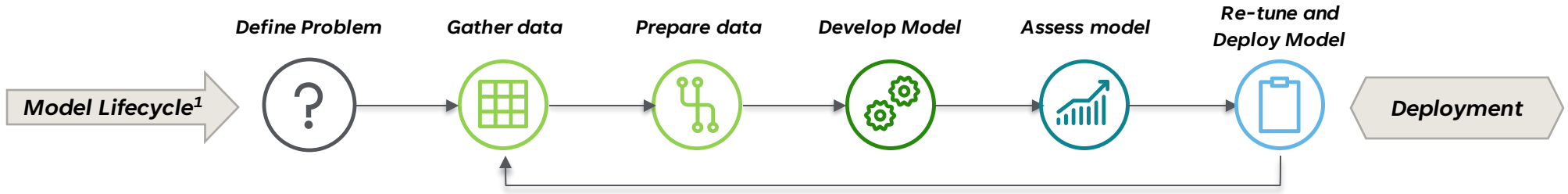
End: Week 09

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 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

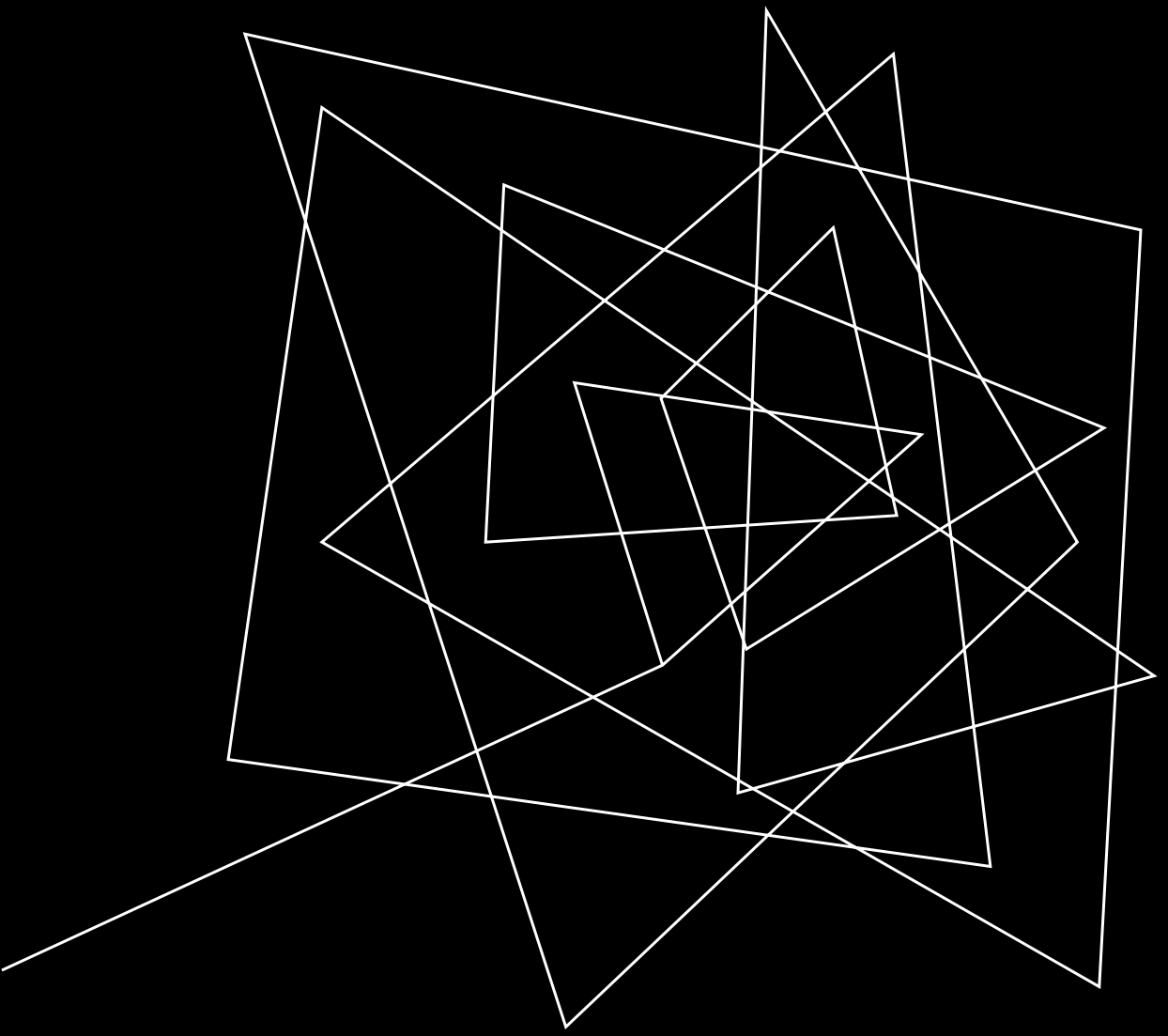
- Use IBM AIF360's Tutorial to run training and testing, and capture fairness metrics
- Evaluate whether the model performs sufficiently for production.
- Does the model answer the question with sufficient confidence given the test data?

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)

FAIRNESS METRICS

FAIRNESS CRITERIA DESCRIBES THE CONNECTION BETWEEN SENSITIVE ATTRIBUTES AND TRUE/PREDICTED LABELS MATHEMATICALLY.

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?

DISPARATE IMPACT

Disparate Impact:

$$\frac{= P(\text{favorable} = 1) \mid C = \text{unprivileged})}{P(\text{favorable} = 1) \mid C = \text{privileged})}$$

In other words:

$$\frac{\frac{\#\text{ favorable from underprivileged}}{\#\text{ unprivileged}}}{\frac{\#\text{ favorable from privileged}}{\#\text{ privileged}}}$$

General Consensus (industry standard):

If your disparate impact is LESS THAN 80% - your model is unfair and should be revised.

STATISTICAL PARITY

Something is "fair" if there isn't a difference on whether or not you have a given attribute (A):

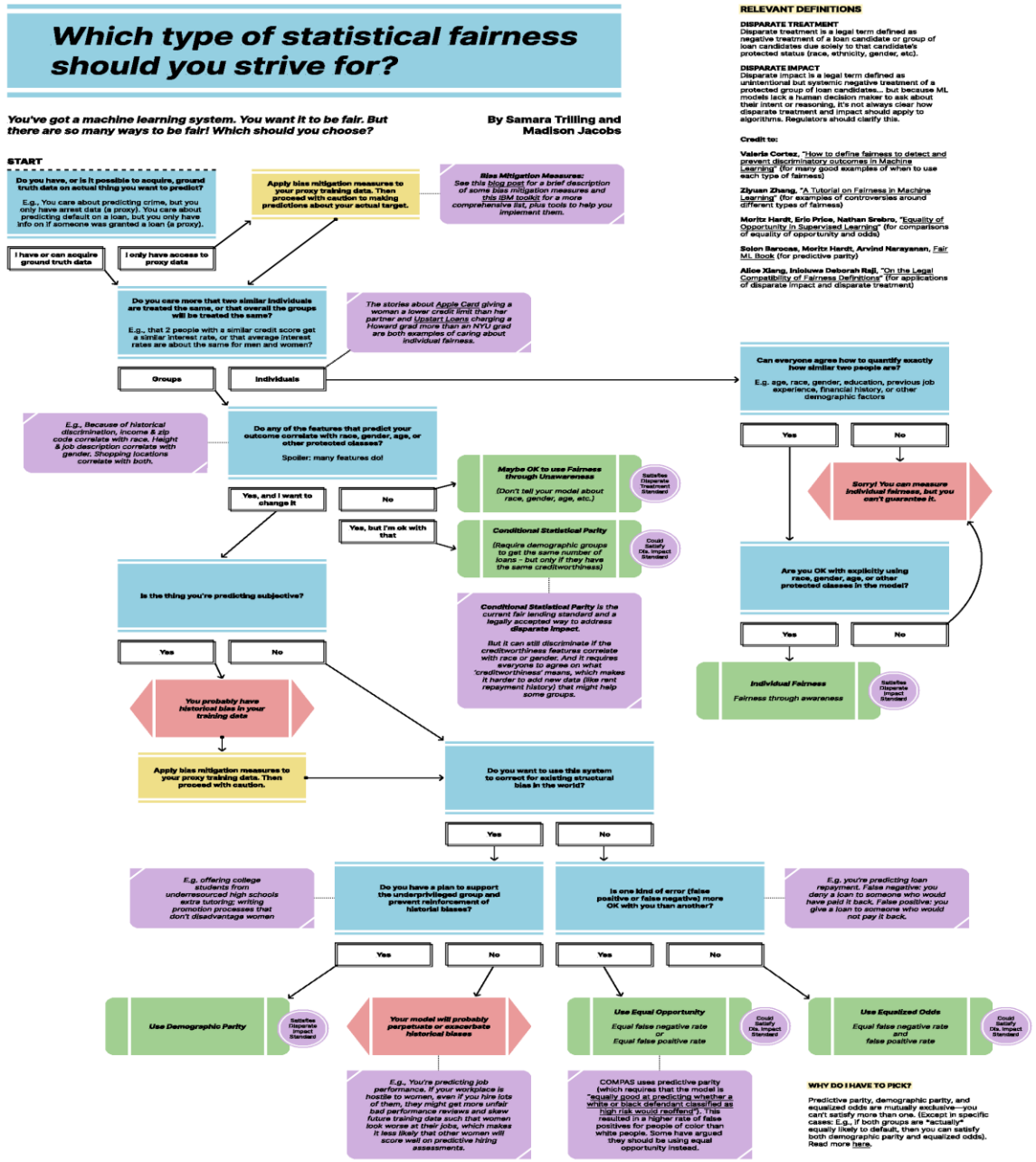
$$disparity = Pr(Y = y \mid A = 0) - Pr(Y = y \mid A = 1)$$

```
                                num_positives(privileged=False)     (num_positives(privileged=True)
Statistical parity difference =  ------------------------------- -  -------------------------------
                                num_instances(privileged=False)     num_instances(privileged=True)
```

- **Under 0**: Higher benefit for the monitored group.
- **At 0**: Both groups have equal benefit.
- **Over 0** Implies higher benefit for the reference group.

General Consensus (industry standard):

There is no disparity if disparity = 0. This (obviously) will not occur – so we introduce a threshold.

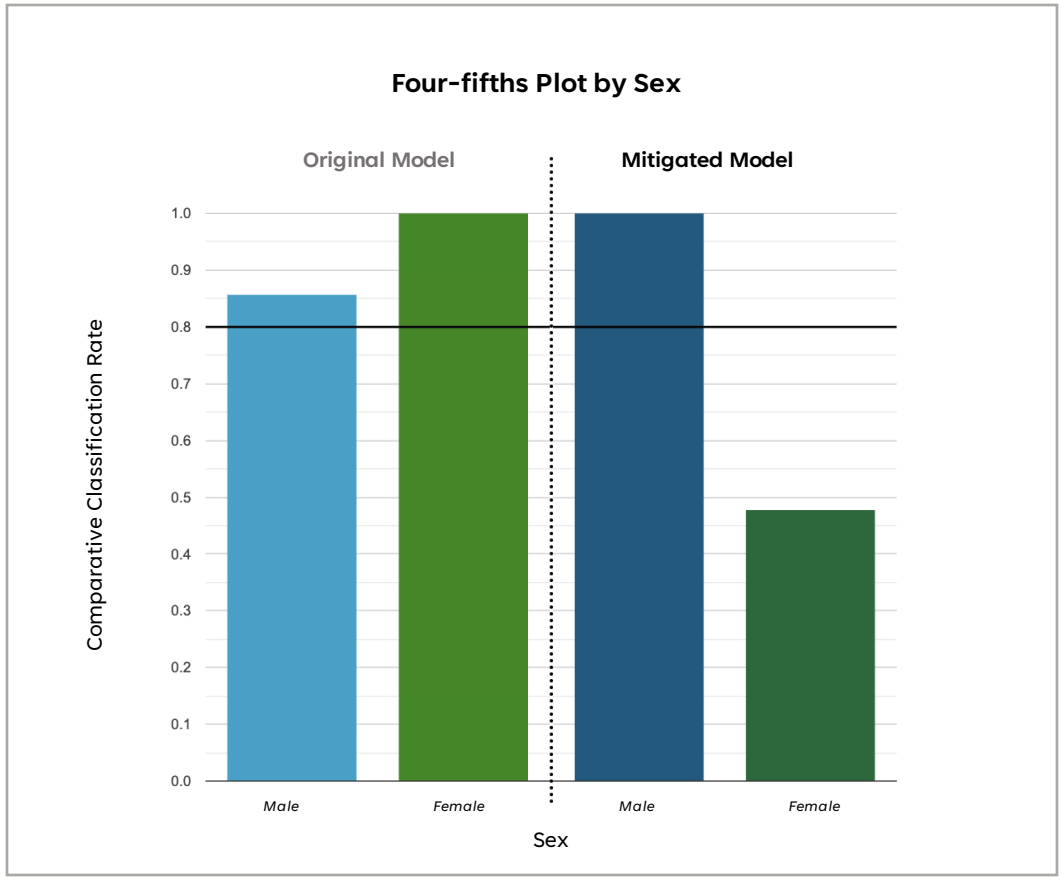


https://www.aspentechpolicyhub.org/wp-content/uploads/2020/07/FAHL-Tree.pdf

FAIRNESS PLOT EXAMPLES

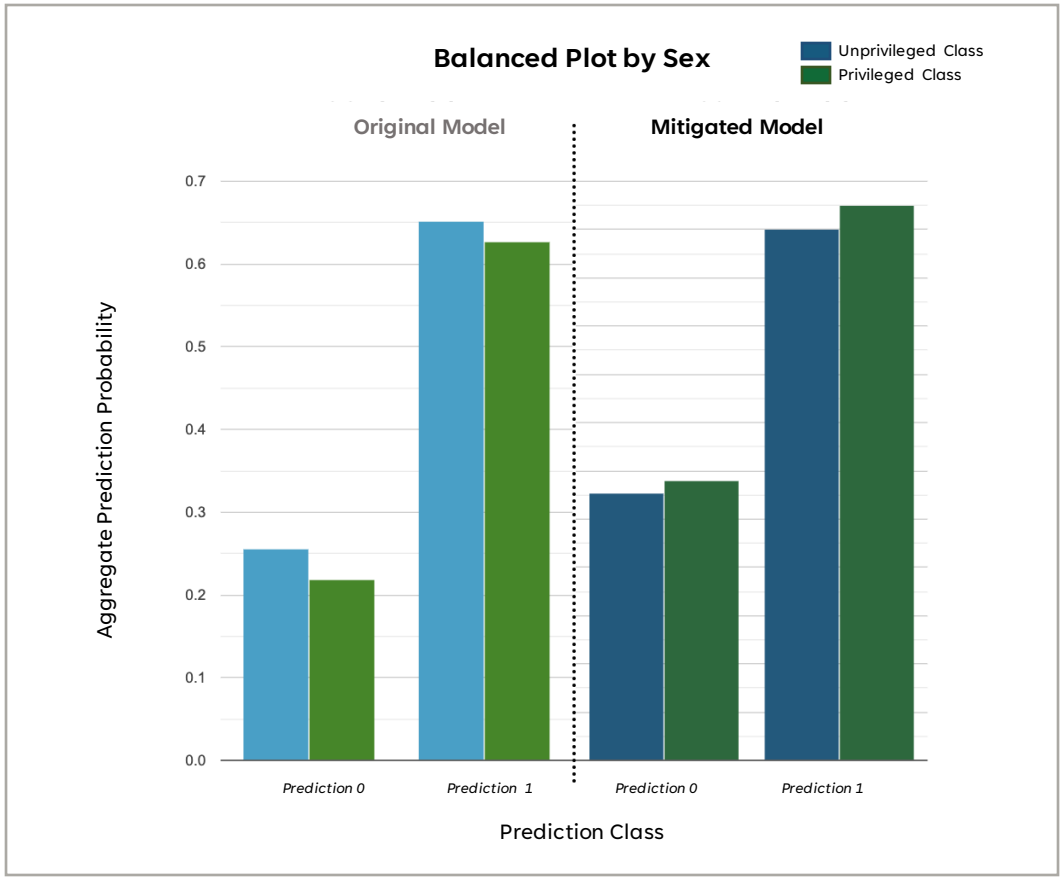
Four Fifths Plot

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



Balance Plot

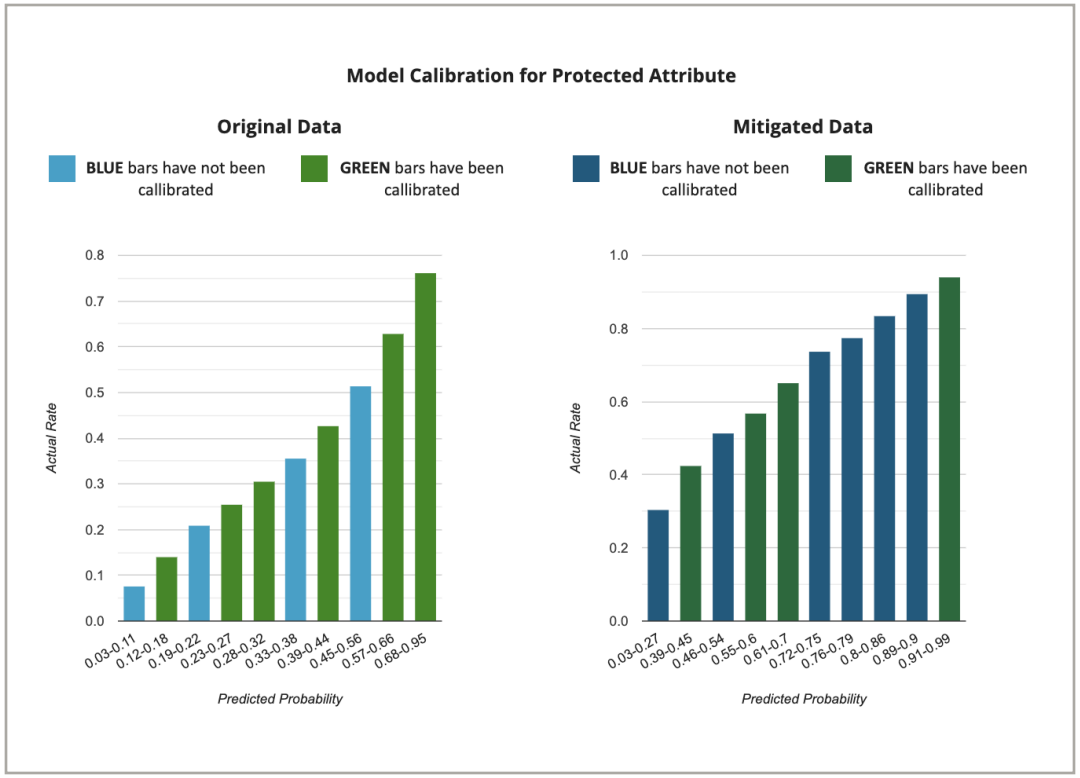
Examines whether average score received by individuals in positive and negative instances are similar regardless of sensitive attributes.



OTHER FAIRNESS VISUALIZATIONS (*IMPUTED*)

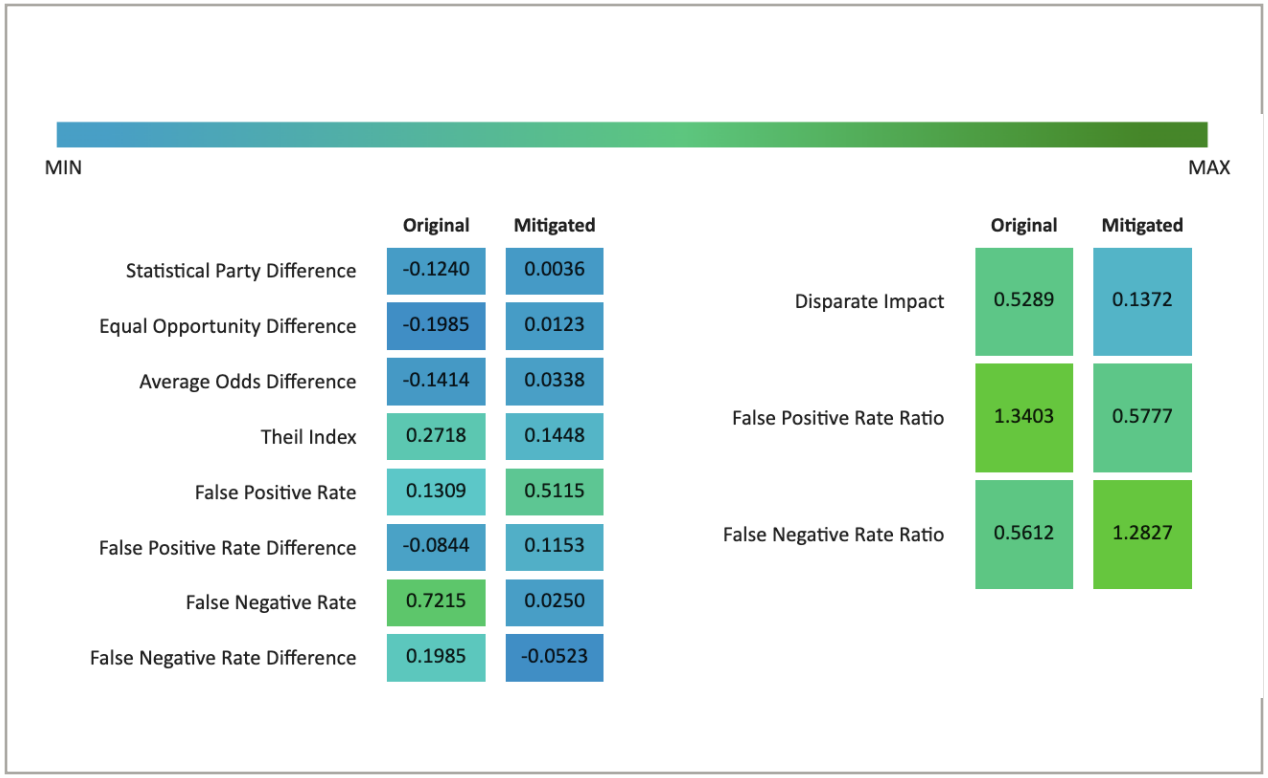
Calibration Plots

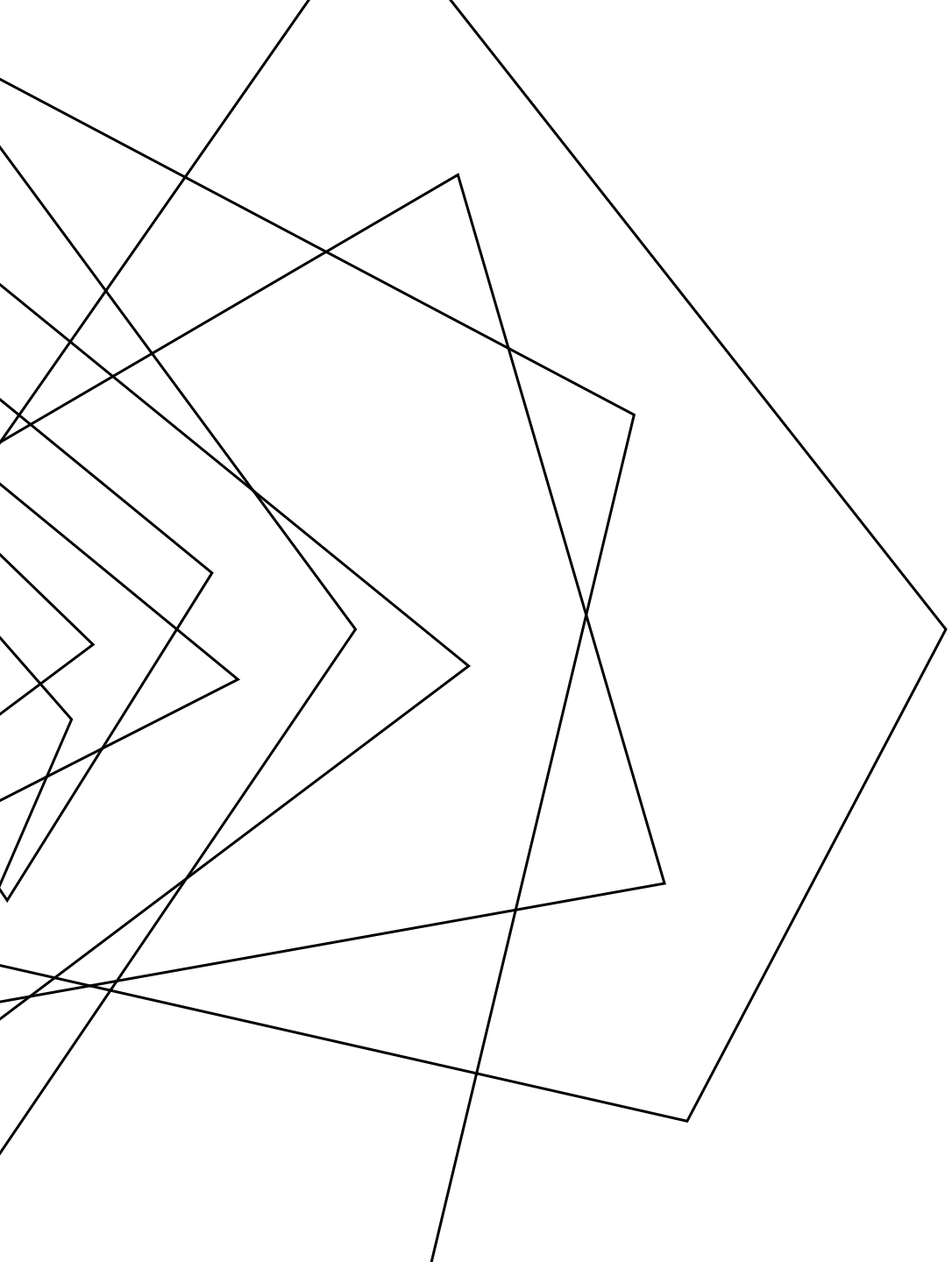
Checks if model makes accurate predictions in aggregate for members of each class.



Fairness Ratios

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





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*

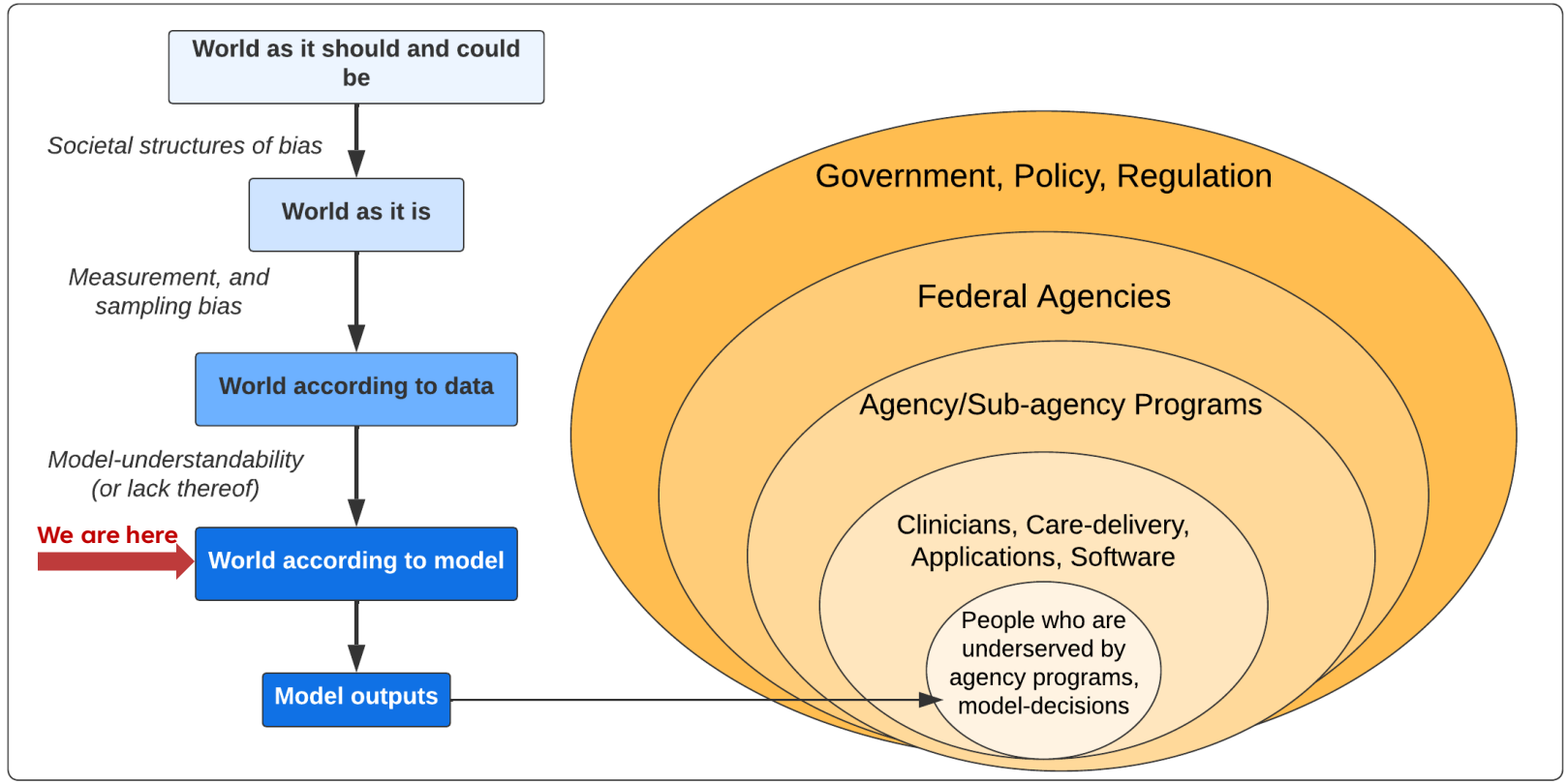


Figure above: Illustration of the types of bias that can enter a model's decision-space (stages of data and algorithm use), resulting in model-outputs that affect individuals downstream. *(Adapted from: Mhasawadade 2020, Mitchell 2020)*

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 as normal
- 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 by 11 AM PT, Thursday, November 16th

- Submit your answers to next week's participation questions on Gradescope by 11 AM PT, Thursday, November 16th

- **Replication Part #2:** Notebook and writeup on model development and fairness metric assessments
  - Primary contact for replication project: Emily Ramond & Parker Addison
  - Office hours: Tuesdays 3:30-4:30pm
  - Start early! You'll need to learn/research fairness metrics.