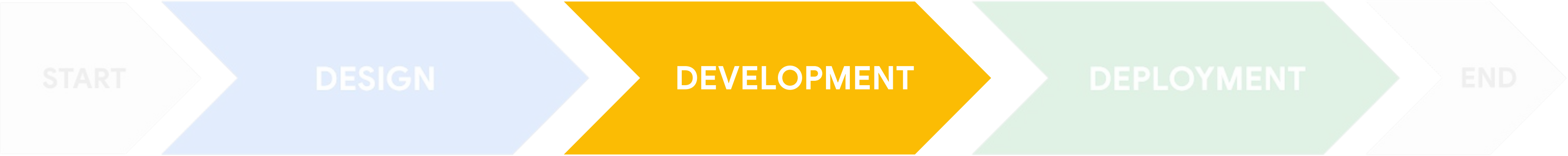
How can we ensure the model is fair?

Responsible AI: Human-Centered Design



START

DESIGN

DEVELOPMENT

DEPLOYMENT

END

Course 1

Fundamentals of TinyML

- **What** am I building?
- **Who** am I building this for?
- What are the **consequences** for the user if it ***fails***?

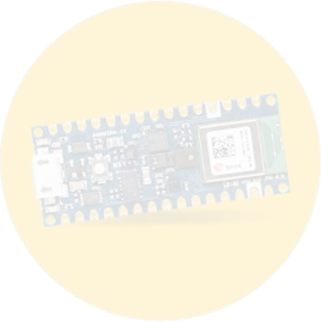
Course 2

Applications of TinyML

- What data will be collected to train the model?
- Is the dataset biased?
- **How can we ensure the model is fair?**

Course 3

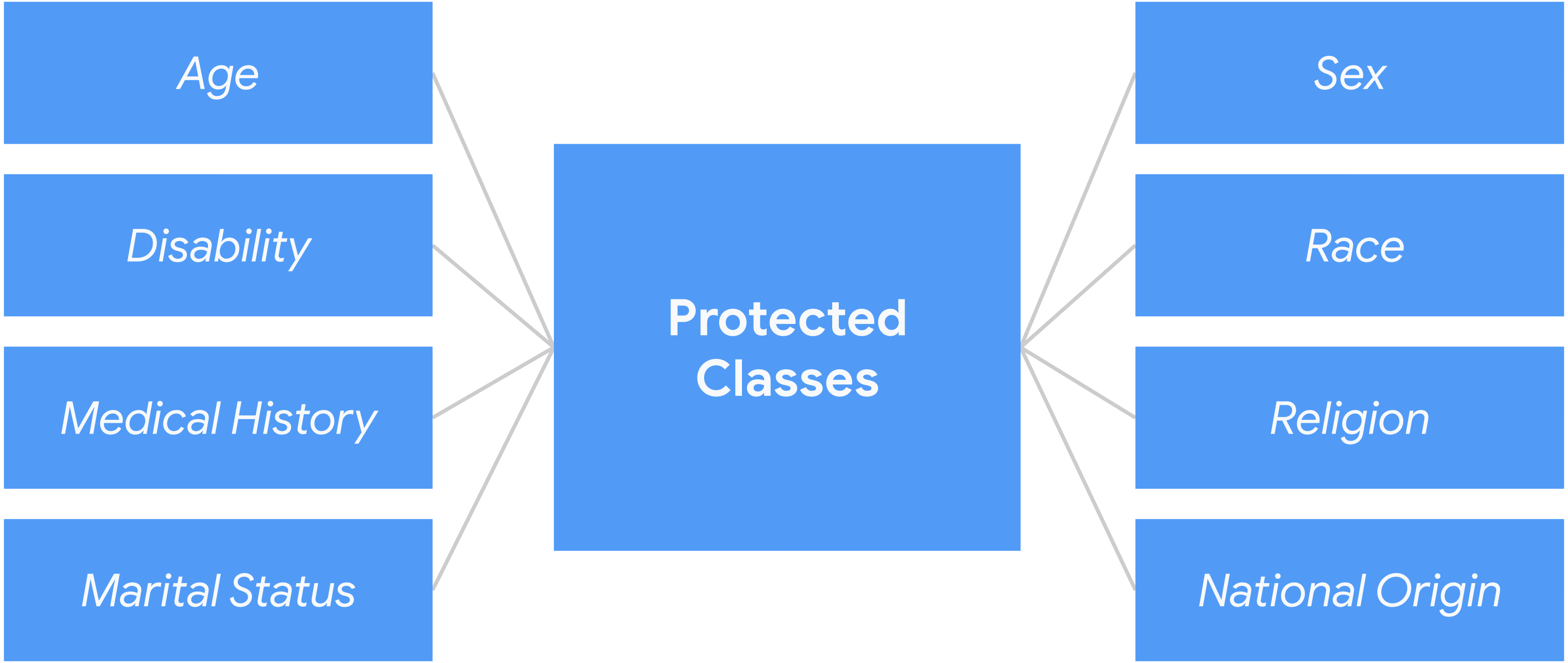
Deploying TinyML



Unfairness in ML

Model exhibits **discriminatory biases**, perpetuates **inequality** or performs less well for historically **disadvantaged groups**

- ***All ML discriminates*** (it just means to recognize a distinction, differentiate)
- Fairness is concerned with **wrongful** discrimination



Protected Classes

Age

Disability

Medical History

Marital Status

Sex

Race

Religion

National Origin

Discrimination

Disparate **Treatment**:

Membership in a protected class is used as an input to the model, decisions are differentiated on that basis in a way that disadvantages members of a protected class

Disparate **Impact**:

Outcomes of the model disproportionately disadvantage members of a protected class

1. Group Unawareness

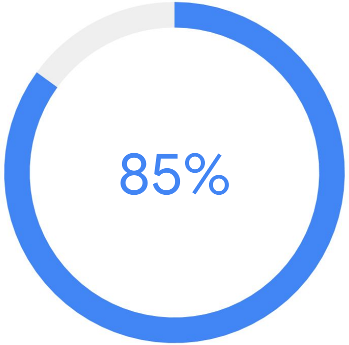
Sensitive attributes are **not** included as features of the data (e.g. race, gender)

Pro: Avoids disparate treatment

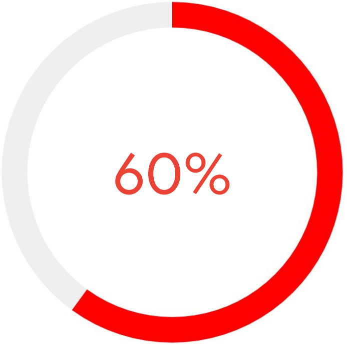
Con: Possibility of highly correlated features that are proxies of the sensitive attribute

2. Group Threshold

Counteract historical biases in data by **adjusting** confidence thresholds ***independently*** for each group



Group A



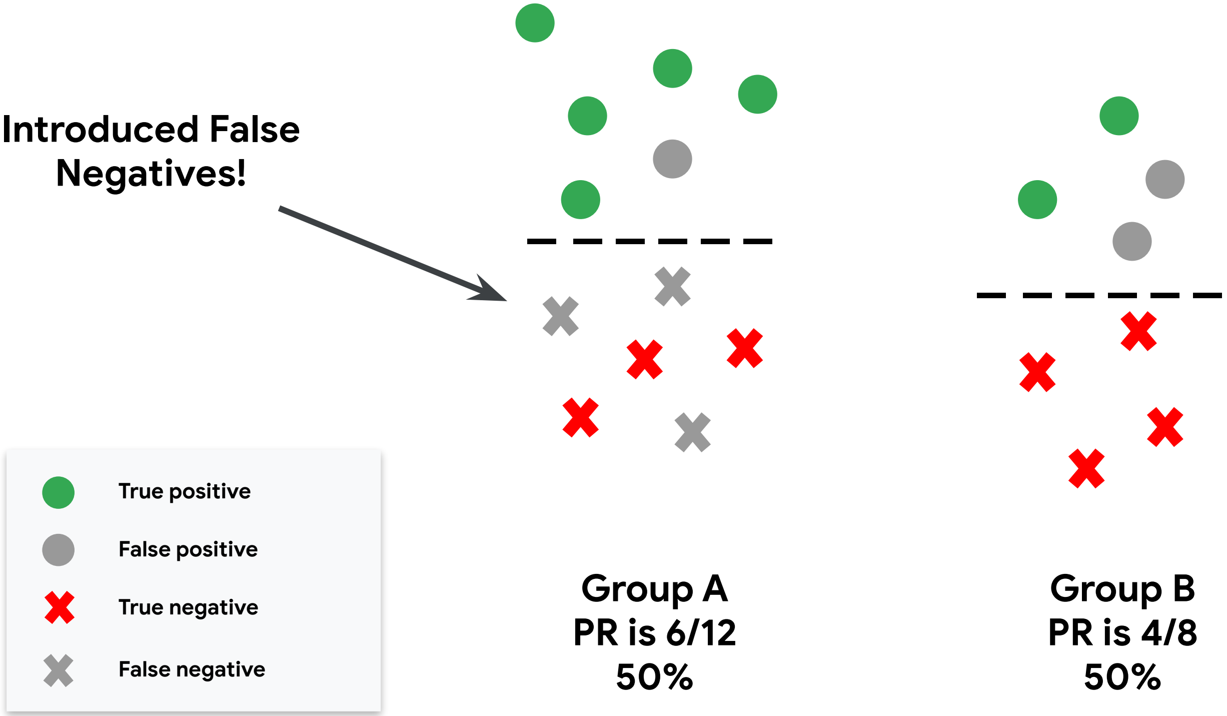
Group B

3. Demographic Parity

	Actually Healthy = Yes	Actually Healthy = No
Predicted Healthy = Yes	***True Positive***	***False Positive***
Predicted Healthy = No	False Negative	True Negative

The positive rate is the same across groups

Problem with Demographic Parity



Introduced False Negatives!

Group A
PR is 6/12
50%

Group B
PR is 4/8
50%

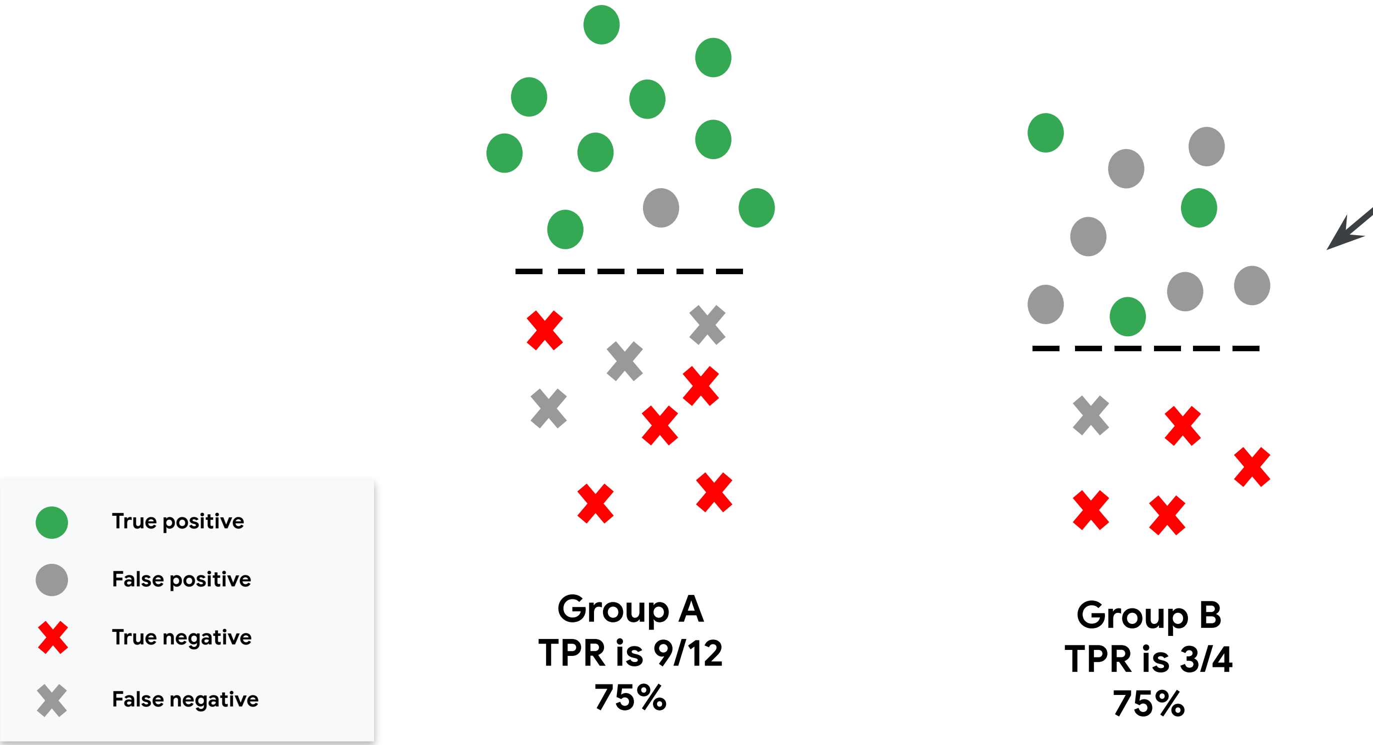
- ● True positive
- ● False positive
- ✖ True negative
- ✖ False negative

4. Equal Opportunity

	Actually Healthy = Yes	Actually Healthy = No
Predicted Healthy = Yes	***True Positive***	False Positive
Predicted Healthy = No	***False Negative***	True Negative

Qualified individuals should have an equal chance of being correctly classified for a desirable outcome.

Problem with Equality of Opportunity



Introduced False Positives!

● True positive

● False positive

✖ True negative

✖ False negative

Group A
TPR is 9/12
75%

Group B
TPR is 3/4
75%

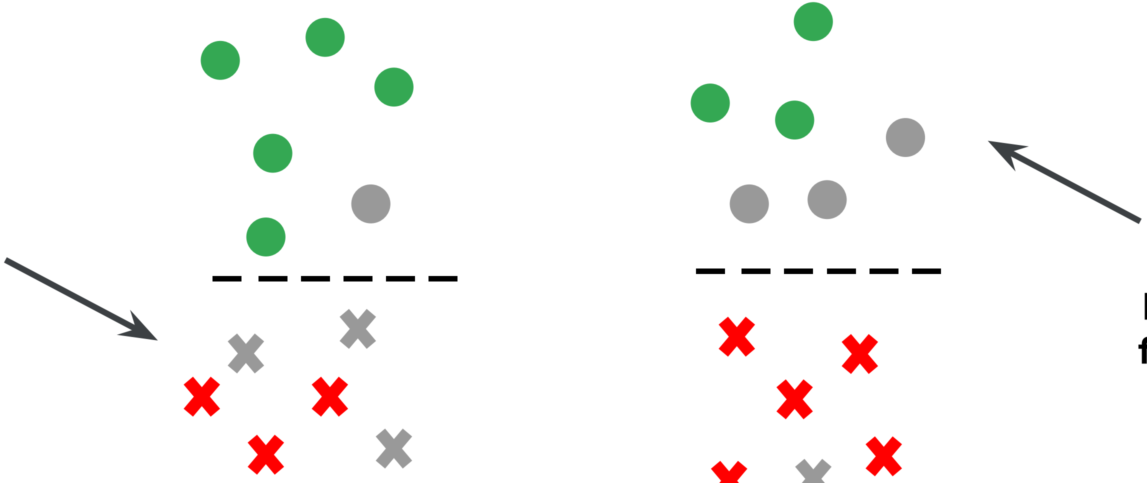
4. Equal Accuracy

	Actually Disease = Yes	Actually Disease = No
Predicted Disease = Yes	***True Positive***	False Positive
Predicted Disease = No	False Negative	***True Negative***

The percentage of correct classifications should be the same for all individuals

Problem with Equal Accuracy

Higher rate of false negatives



Higher rate of false positives

- 🟢 True positive
- ● False positive
- ✗ True negative
- False negative

Group A
Accuracy is
75%

Group B
Accuracy is
75%

Impossibility Theorem

We cannot satisfy all fairness metrics

at the same time!

For example:

- **Group Unawareness** is incompatible with **Group Threshold**
- **Equal Opportunity** is incompatible with **Equal Accuracy**

How can we mitigate ***unfairness*** in ML?

The Framing Trap

Algorithmic Frame

Do properties of the output match the input? Does the algorithm provide good accuracy on unseen data?

Data Frame

Has bias been removed from the training data? Does the demographic information of the data require optimization of the model?

Sociotechnical Frame

How does the model operate when considered as part of a system of humans and social institutions?

The Framing Trap

Algorithmic Frame

Do properties of the output match the input? Does the algorithm provide good accuracy on unseen data?

Data Frame

Has bias been removed from the training data? Does the demographic information of the data require optimization of the model?

Sociotechnical Frame

How does the model operate when considered as part of a system of humans and social institutions?

The Framing Trap

Algorithmic Frame

Do properties of the output match the input? Does the algorithm provide good accuracy on unseen data?

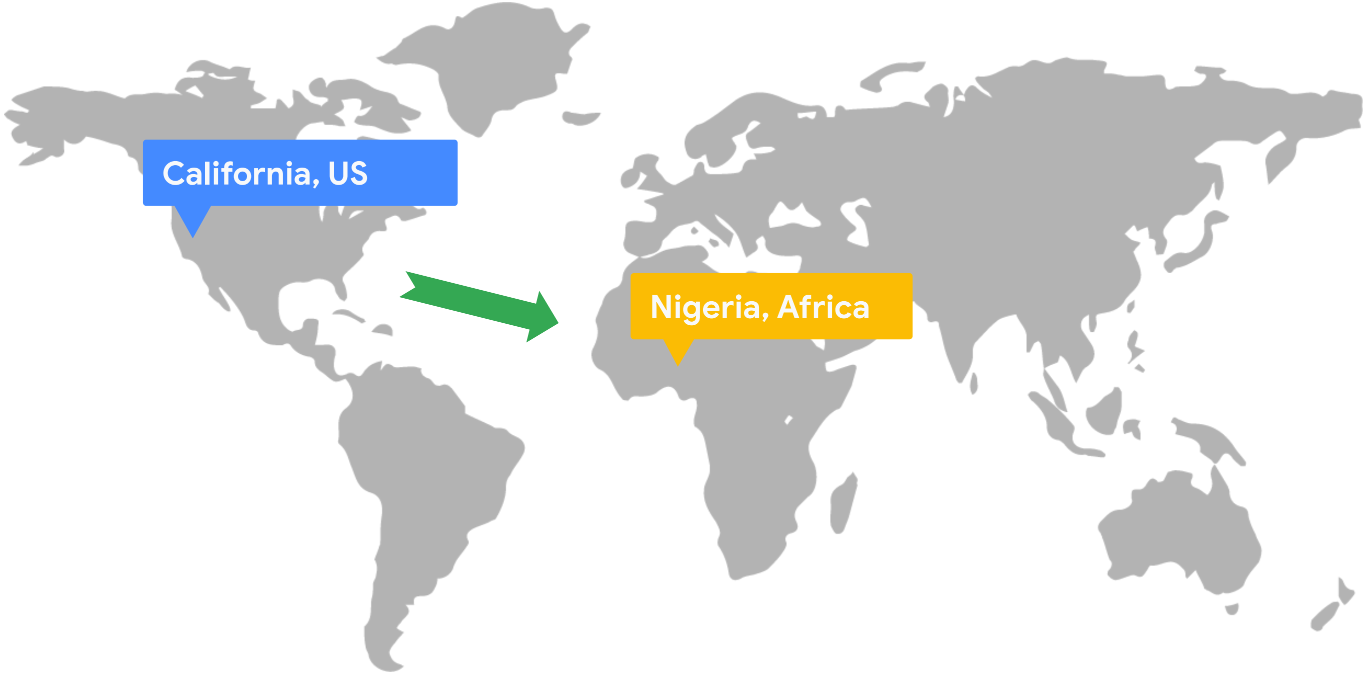
Data Frame

Has bias been removed from the training data? Does the demographic information of the data require optimization of the model?

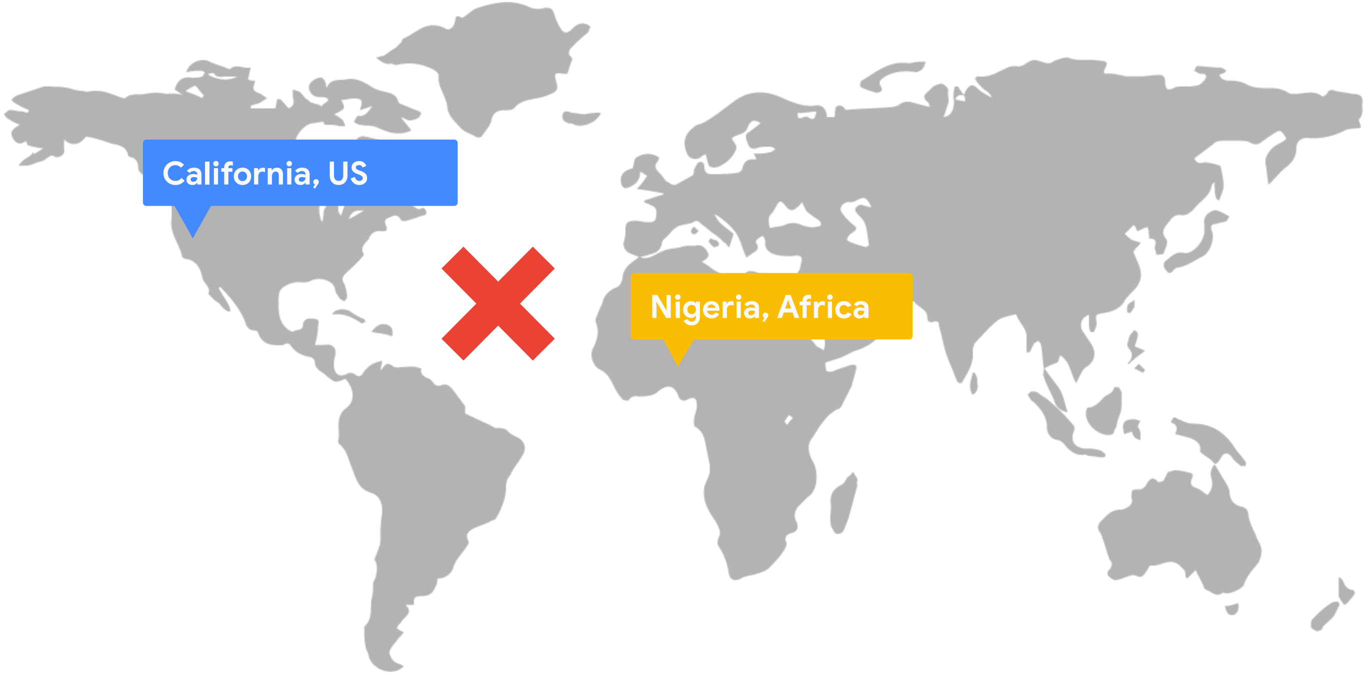
Sociotechnical Frame

How does the model operate when considered as part of a system of humans and social institutions?

The Portability Trap



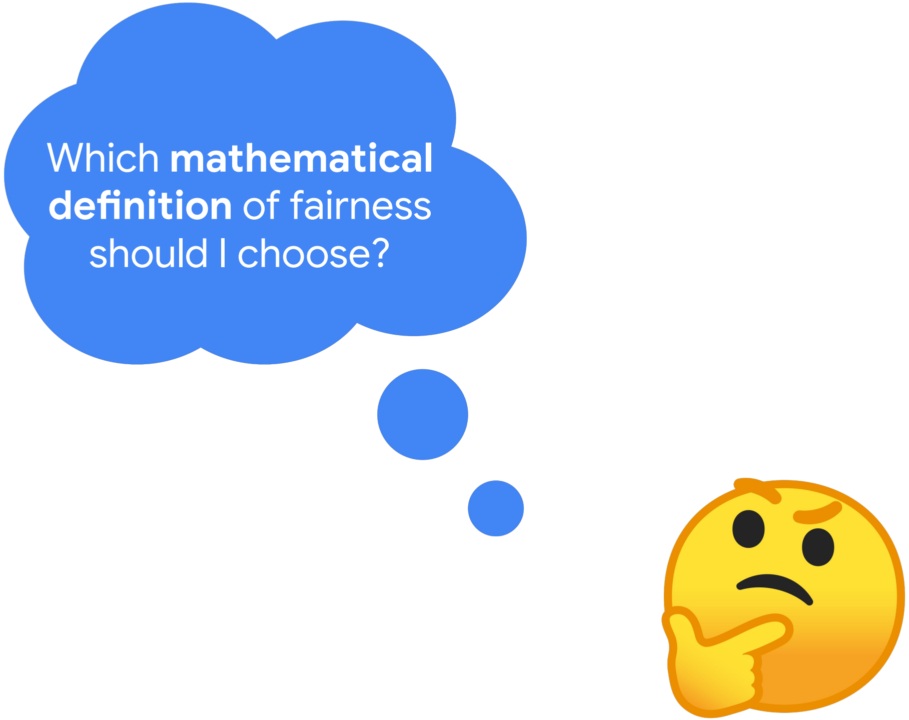
The Portability Trap



Context Matters!

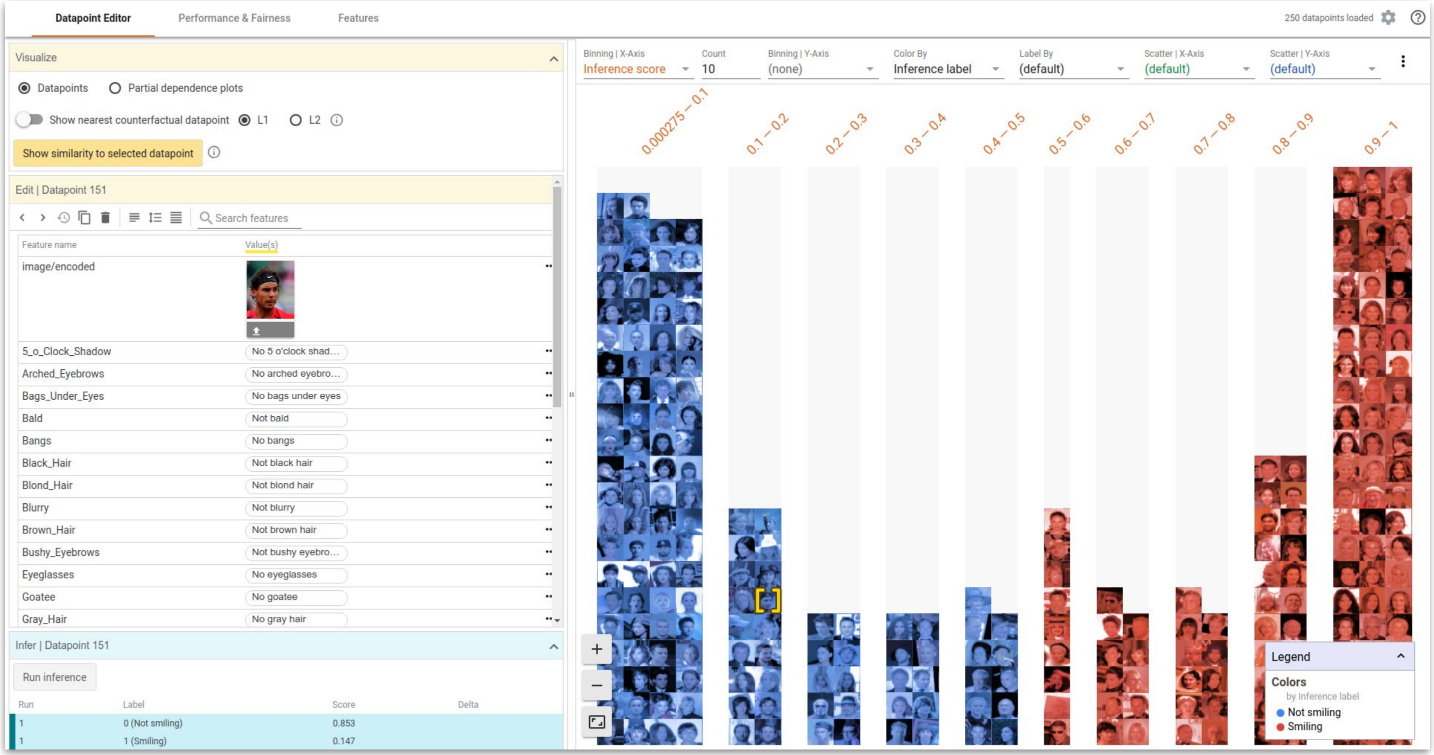
Repurposing algorithmic solutions may not preserve fair outcomes.

The **Formalism** Trap



Which **mathematical definition** of fairness should I choose?

Google's **What-If Tool**



Visualize

● Datapoints ○ Partial dependence plots

Show nearest counterfactual datapoint ● L1 ○ L2

Show similarity to selected datapoint

Edit | Datapoint 151

Search features

Feature name	Value(s)
image/encoded	
5_o_Clock_Shadow	No 5 o'clock shad…
Arched_Eyebrows	No arched eyebro…
Bags_Under_Eyes	No bags under eyes
Bald	Not bald
Bangs	No bangs
Black_Hair	Not black hair
Blond_Hair	Not blond hair
Blurry	Not blurry
Brown_Hair	Not brown hair
Bushy_Eyebrows	Not bushy eyebro…
Eyeglasses	No eyeglasses
Goatee	No goatee
Gray_Hair	No gray hair

Infer | Datapoint 151

Run inference

Run	Label	Score	Delta
1	0 (Not smiling)	0.853	
1	1 (Smiling)	0.147	

Binning | X-Axis Inference score

Count 10

Binning | Y-Axis (none)

Color By Inference label

Label By (default)

Scatter | X-Axis (default)

Scatter | Y-Axis (default)