Design Patterns for Responsible AI

Sara Robinson, Developer Advocate

@SRobTweets

sararobinson.dev

Agenda

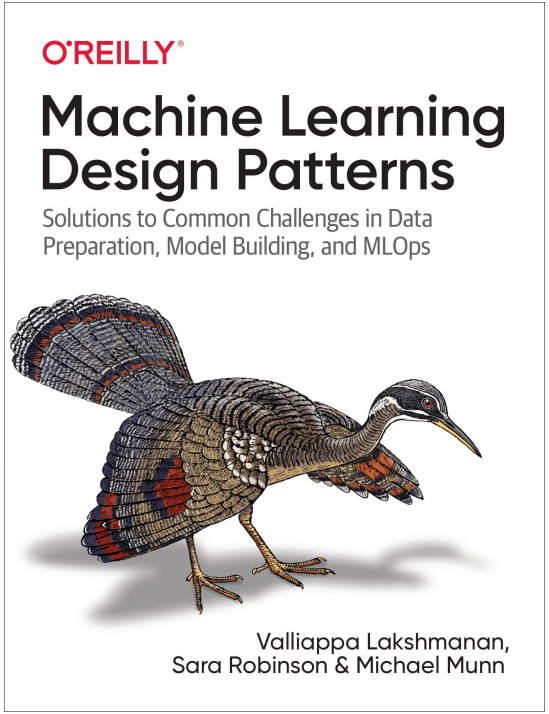
1. What are design patterns?
2. Defining Responsible AI
3. Three patterns + some live demos

What are design patterns?

@SRobTweets

Design patterns are formalized best practices to solve common problems when designing a software system.

We wrote a book!



Pre-order **bit.ly/ml-design-patterns**

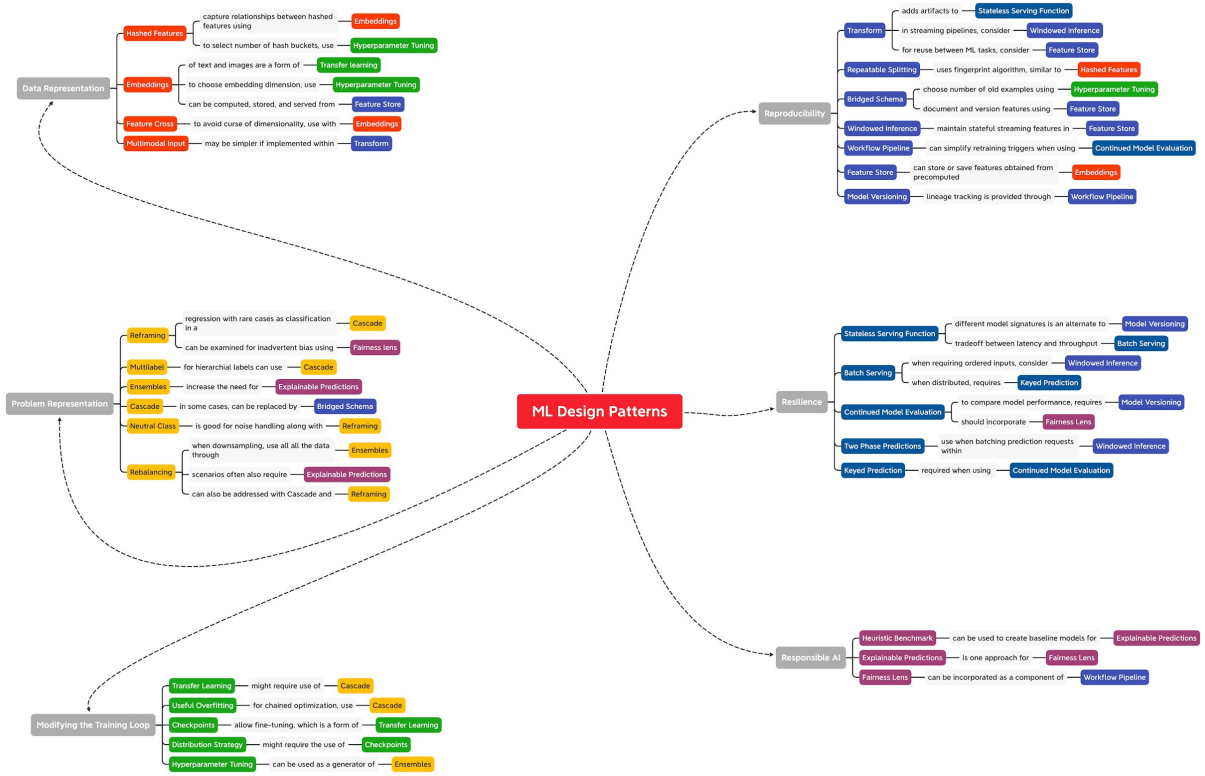
Launching **November 2020**

Follow us on Twitter:

@lak_gcp

@SRobTweets

ML Design Patterns: quick preview



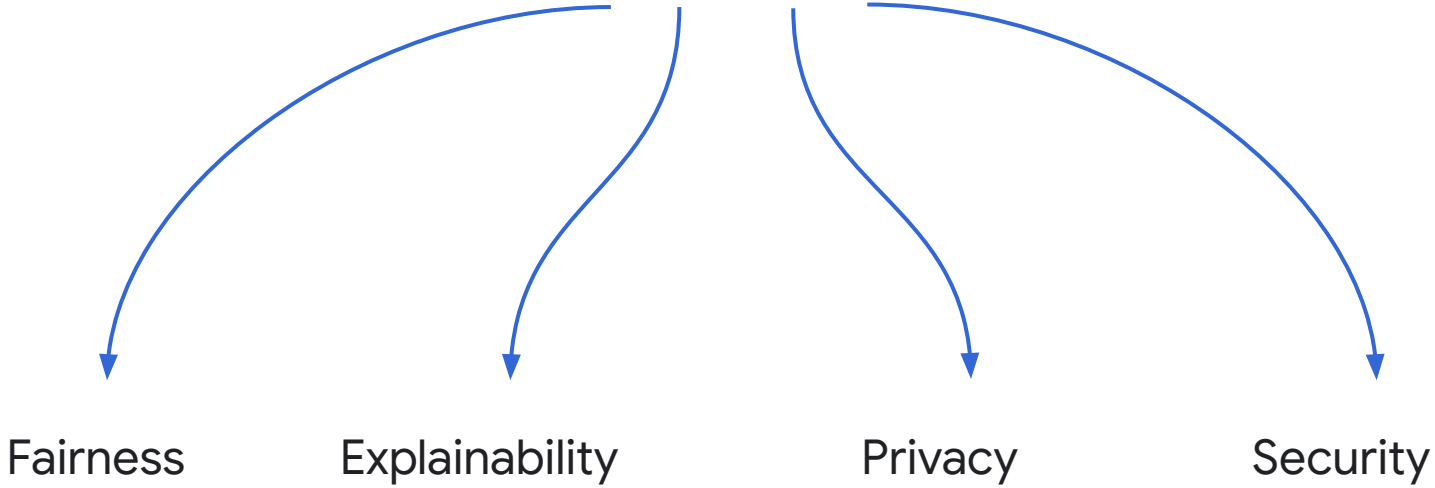
Defining Responsible AI

@SRobTweets

What is **Responsible AI**?

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education. It is also raising new questions about the best way to build **fairness**, **explainability**, **privacy**, and **security** into these systems.

Responsible AI



3 patterns for Responsible AI

1. Heuristic Benchmark — Developing a starting point for summarizing and evaluating a model
2. Explainable Predictions — Understanding the features influencing model behavior
3. Fairness Lens — Ensuring models are fair and equitable for different groups of users

Fairness

Understanding the reasons behind a model's predictions can help ensure models are **treating all users fairly**

Explainability

The process of understanding **how** and **why** a machine learning model is making predictions.

Pattern #1: Heuristic Benchmark

@SRobTweets

Let’s start with an example

🚴 You’re building a model to predict bike rental duration

📊 *The model’s mean absolute error (MAE) is 1,200 seconds. Great!*

🤔 But is that good or bad??

Heuristic benchmark = simple point of comparison

- Good benchmarks:
  - Constant
  - Rule of thumb
  - Mean / median / mode
  - Human experts

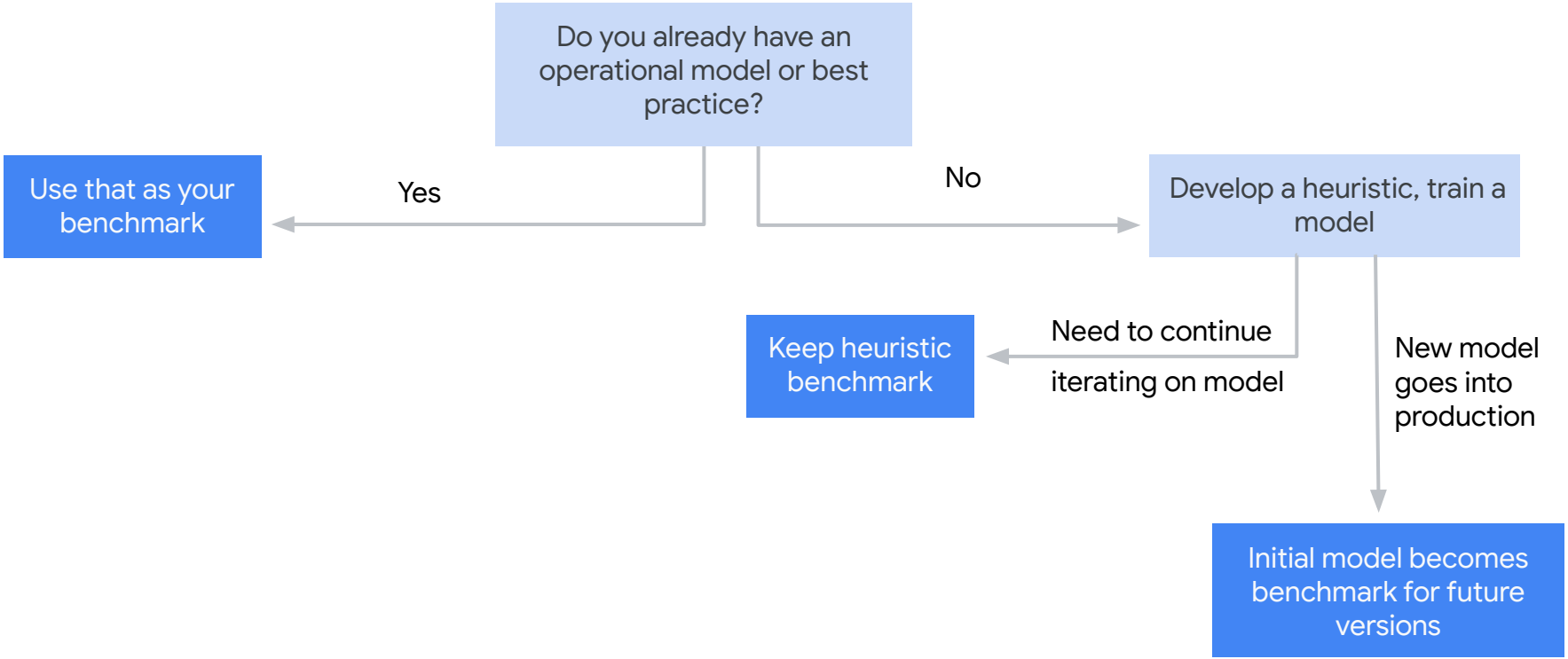
- Not necessarily determined by ML: comparing to a linear regression model isn't always best

Returning to our bike example

- In our training dataset, what is the average rental duration given the station name and whether or not it is a peak commute hour?

- How does our model performance compare to this benchmark?

Should you use a heuristic benchmark?



Pattern #2: Explainable Predictions

@SRobTweets

“Explainability” evokes a variety of related concepts

Human-in-the-loop

Data provenance

Auditability

Transparency

Model Lineage

Explainability

Accountability

Interpretability

Traceability

Contestability

Human-on-the-loop

Who are model stakeholders?

	Example questions
Model builders & ML Ops	• Why is my model not performing? • How can I improve it?
‘End users’ of ML systems	• Should I trust the model’s output? • How should I respond to the prediction?
Public stakeholders	• Is the model safe and fit-for-purpose? • Does it comply with regulations?

How can users **take action** from explanations?

Model builders & ML Ops

- Improve training data, refine features
- Update model architecture
- Involve stakeholders and domain experts

‘End users’ of ML systems

- Make informed decisions
- Identify new areas for model refinement
- Take recourse on contentious predictions

Public stakeholders

- Audit a model’s behavior to see if it complies with regulations/standards
- Use explanations to inform future policy

It depends on the **data type**

Images

Text

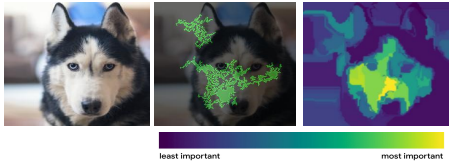
How could you not love cake?!

Tabular

Feature name	Feature value
start_hr	18
weekday	1
distance	1395.51
temp	16.168
dew_point	7.83396
wdsp	0
max_temp	20.7239
prcp	0.03
rain_drizzle	0
duration	11

It depends on the **data type**

Images



Text

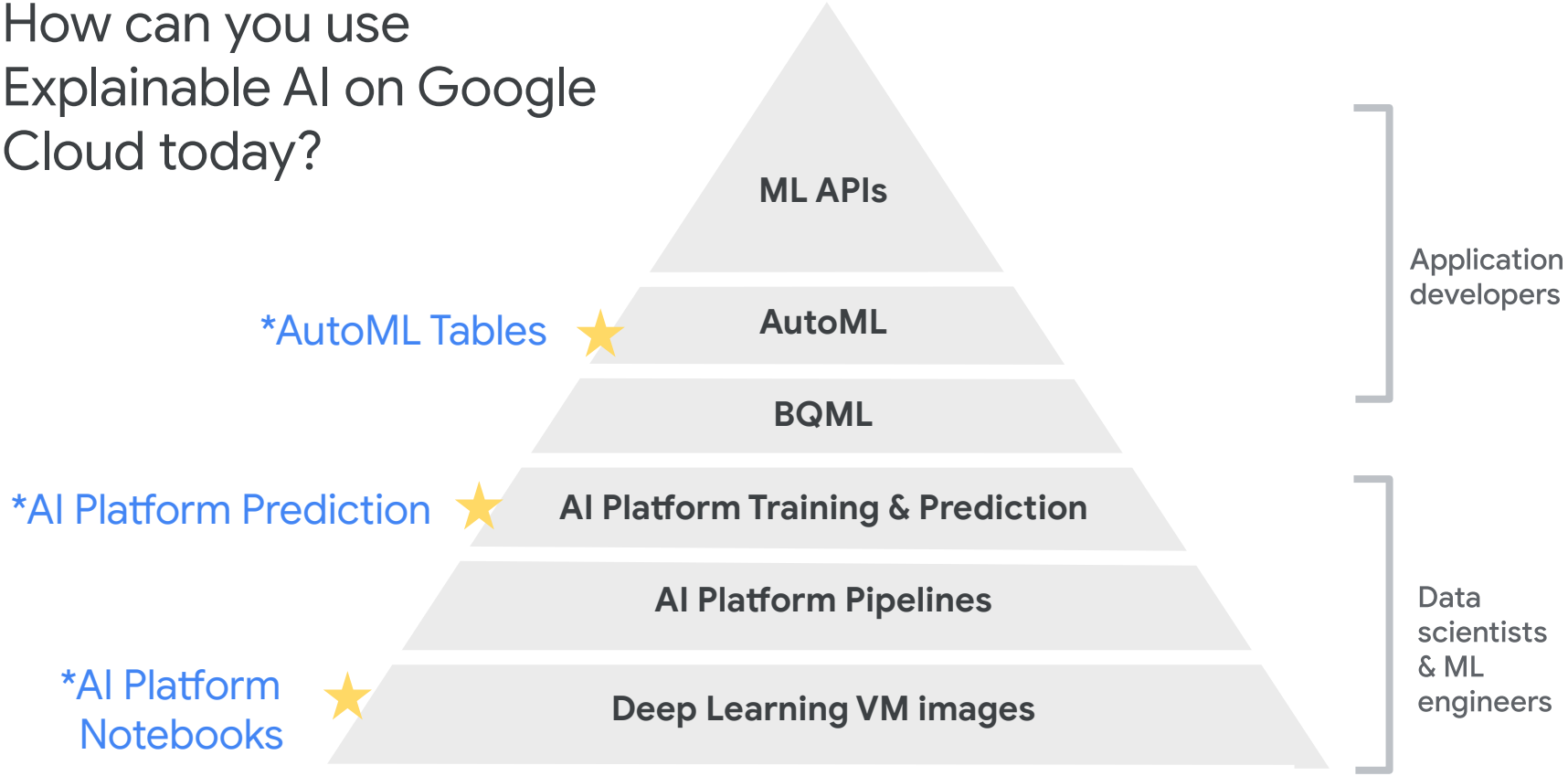
How could you not love cake?!

Sentiment score: 0.9

Tabular

Name	Feature value	Attribution value
distance	1395.51	-2.44478
start_hr	18	-1.29039
max_temp	20.7239	0.690506
temp	16.168	0.12629
dew_point	7.83396	0.0110318
prcp	0.03	-0.00134132

How can you use Explainable AI on Google Cloud today?



Google Cloud

We offer **feature attributions** today

AI explainability

Post-hoc explainers
Explain pre-trained models

- **Feature-based**
- Data-based
- Confidence-based
- Concept-based
- Counterfactuals
- Others...

Interpretable models
Architectures explainable-by-design

- Simple models (linear, GAMs, GLMs, trees, etc.)
- Interpretable neural nets (e.g. TabNet, TFT)
- Models with constraints (e.g. monotonicity)
- RL-based Data Valuation
- Associative/logical decision rule learning
- Others...

Demo time!

@SRobTweets

AutoML Tables

Building a **fraud detection model** using public data from BigQuery

TensorFlow on AI Platform

Building an **image classification** for medical images

Pattern #3: Fairness Lens

@SRobTweets

A shoe example

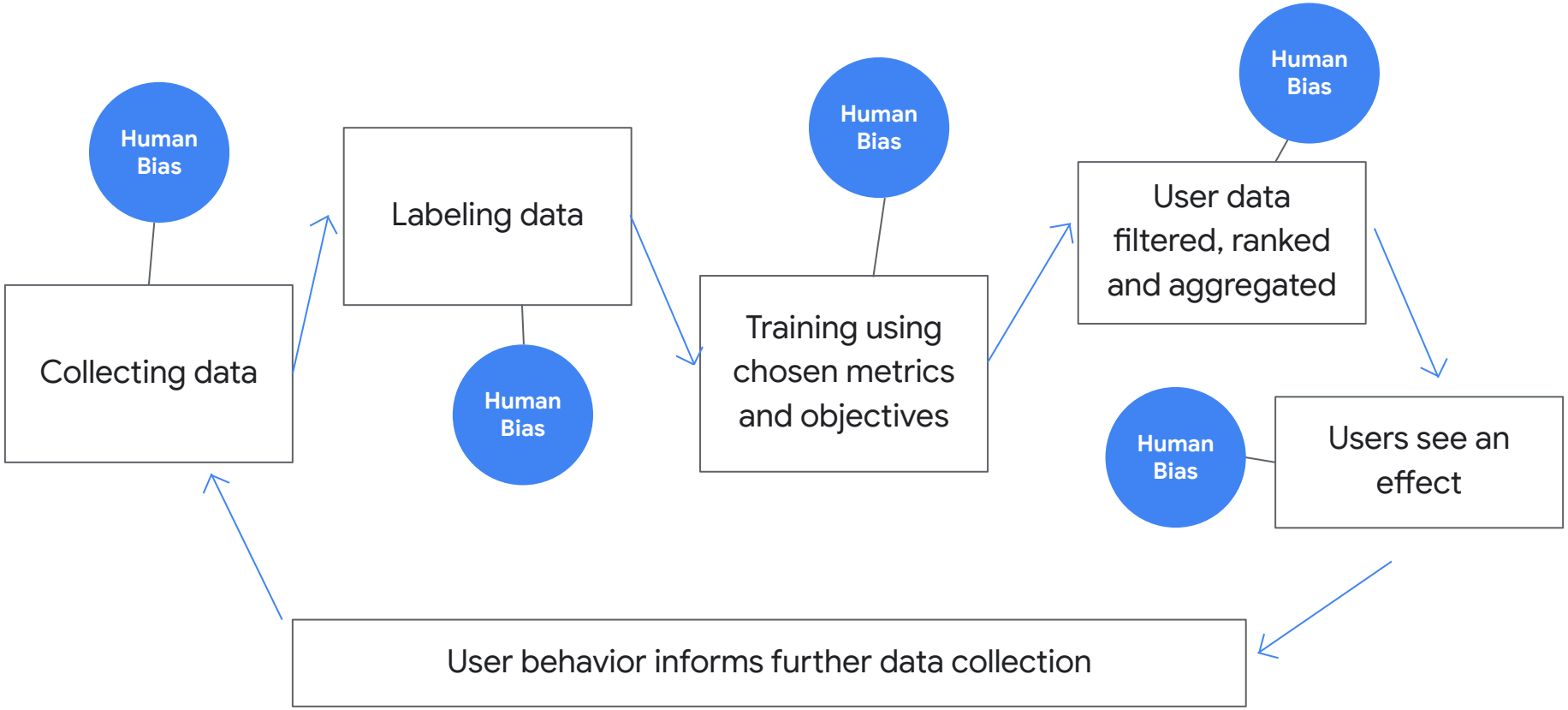
- You're in charge of collecting all of the shoe images for a fashion classification model

- Which types of shoes come to mind?

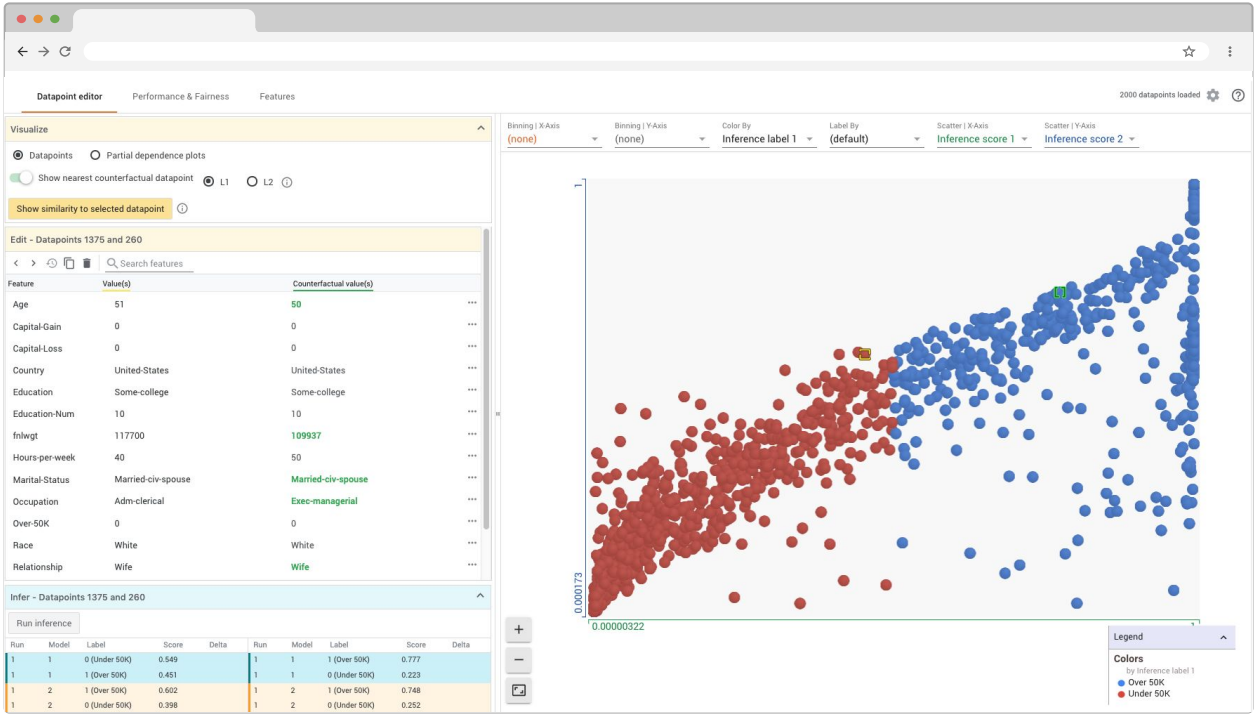
Types of data bias

- Bias isn’t always bad: naturally occuring vs. harmful
- Data distribution bias
- Data representation bias
- Experimenter bias

How can bias affect your ML system?

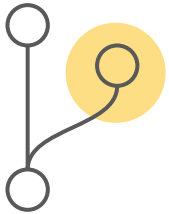


Solving for fairness with the What-If Tool



Google Cloud

How can you use it?



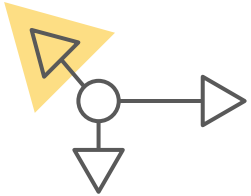
Available on many platforms

TensorBoard dashboard

Google Colaboratory

Jupyter Notebook

Cloud AI Platform Notebooks



Supports What-If Analysis

Explore counterfactuals

Fairness measures

Partial dependence plots

Visualizes Model Performance

Threshold simulation

Up to 2 model comparisons

Dataset summary statistics

Integration with Explainable AI

Edit - Datapoint 43

Search features | Absolute att | -106.0 0.0 106.0

Feature	Value(s)	Attribution value(s)
distance	3043.23876953125	6.7474
max_temp	25.00200080871582	1.1982
temp	17.223600387573242	0.1657
dew_point	6.5005202293396	0.1352
prcp	0	0
rain_drizzle	0	0
start_hr	14	0
wdsp	0	0
weekday	1	0
baseline_score	13.613849639892578	
duration	19	

Google Cloud

Demo time!

@SRobTweets

Resources

- Pre-order the book: bit.ly/ml-design-patterns
- Explainability whitepaper: bit.ly/xai-whitepaper
- Explainability sample code: bit.ly/xai-sample-code
- What-If Tool: pair-code.github.io/what-if-tool
- Codelabs: codelabs.developers.google.com

Thank you

Sara Robinson
@SRobTweets