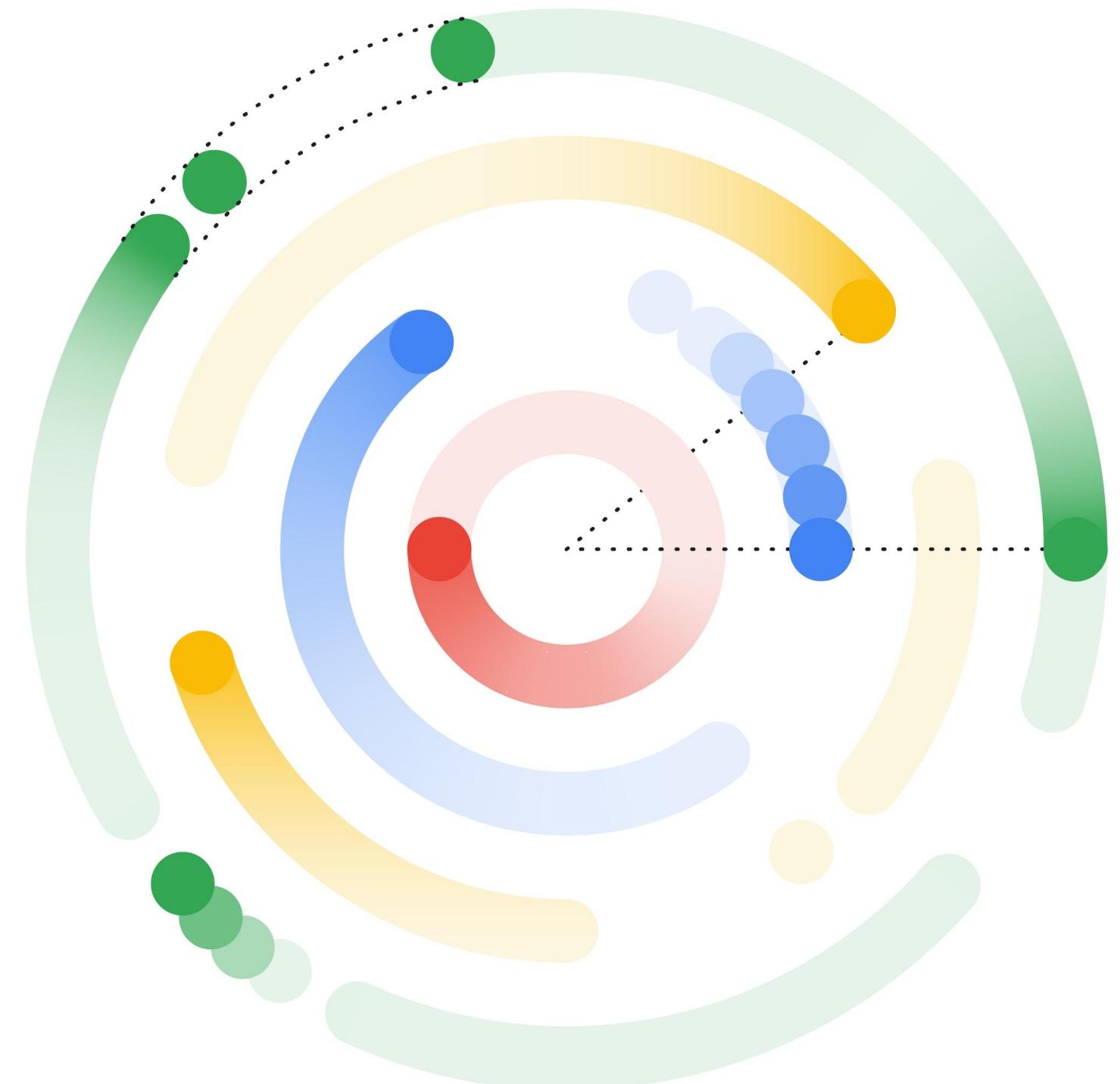


Designing Inclusive ML Systems

Google Cloud Applied ML Summit
Solving for the future.

06/10/21

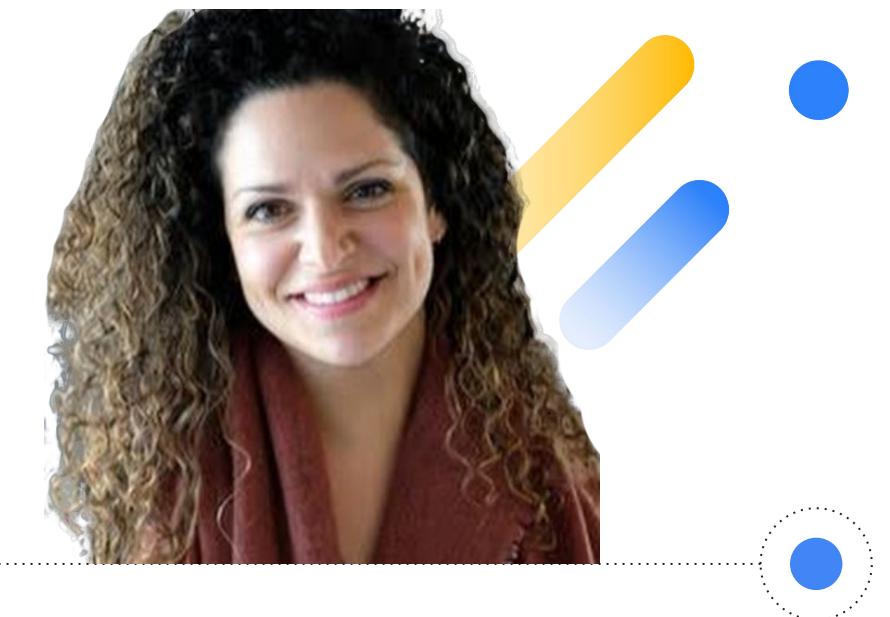




Tracy Frey
Managing Director, Outbound
Product Management &
Responsible AI
Google Cloud



Parker Barnes
Product Manager, Cloud AI
and Industry Solutions
Google Cloud



Madeleine Elish
Senior Research Scientist,
Responsible AI
Google

Agenda

The case for Responsible AI

01

Responsible AI Tooling

02

Responsible AI Case Study

03

Resources

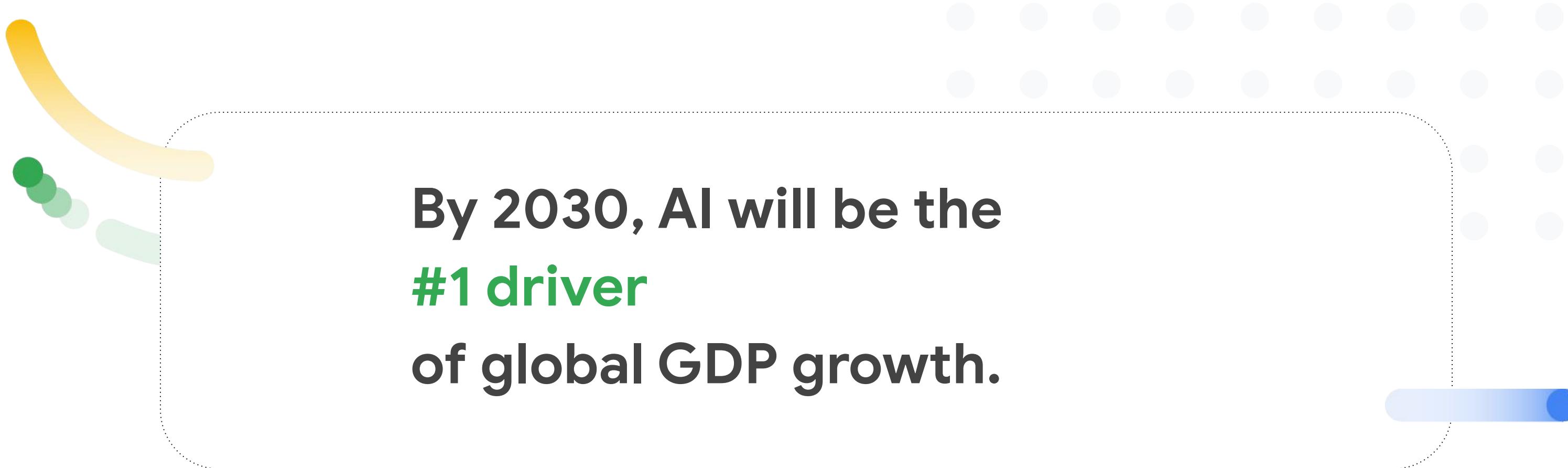
04



01

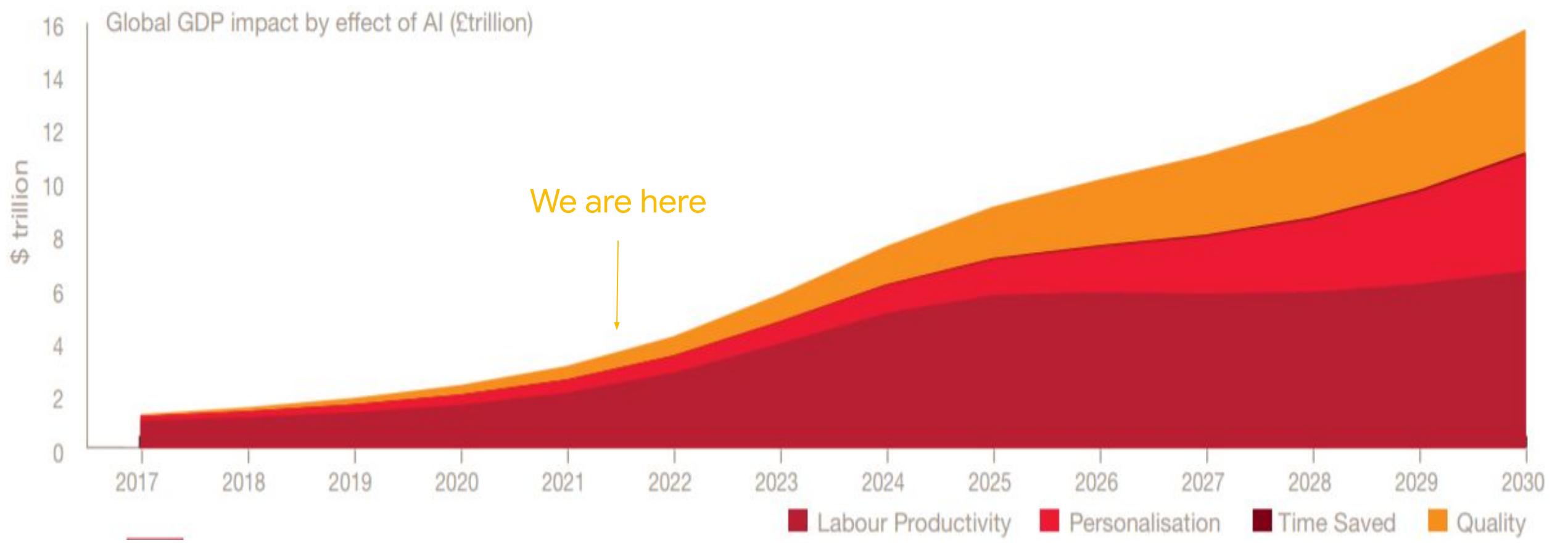


The case for responsible AI



PWC 2019 (chart, data), McKinsey (data)

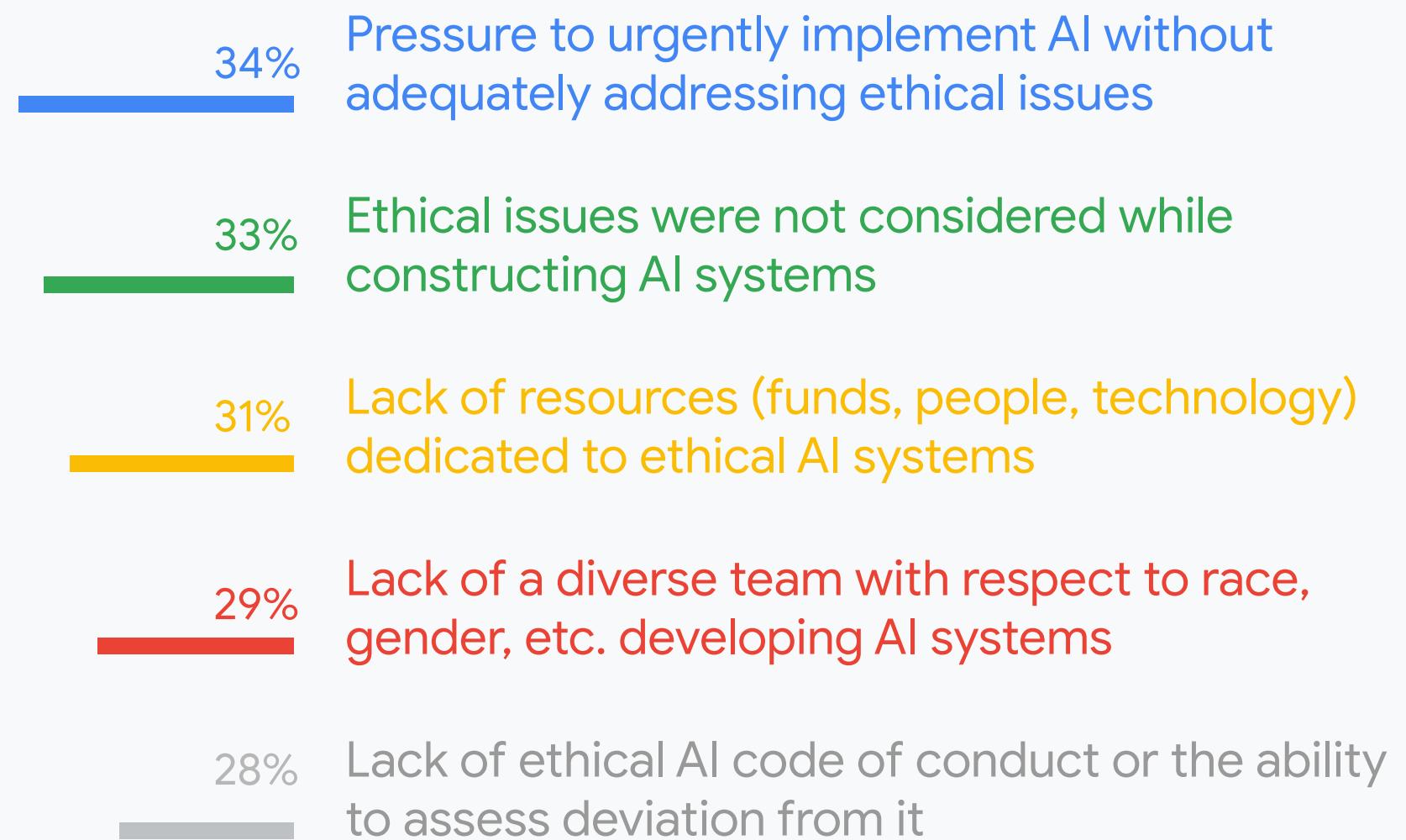
Google Cloud



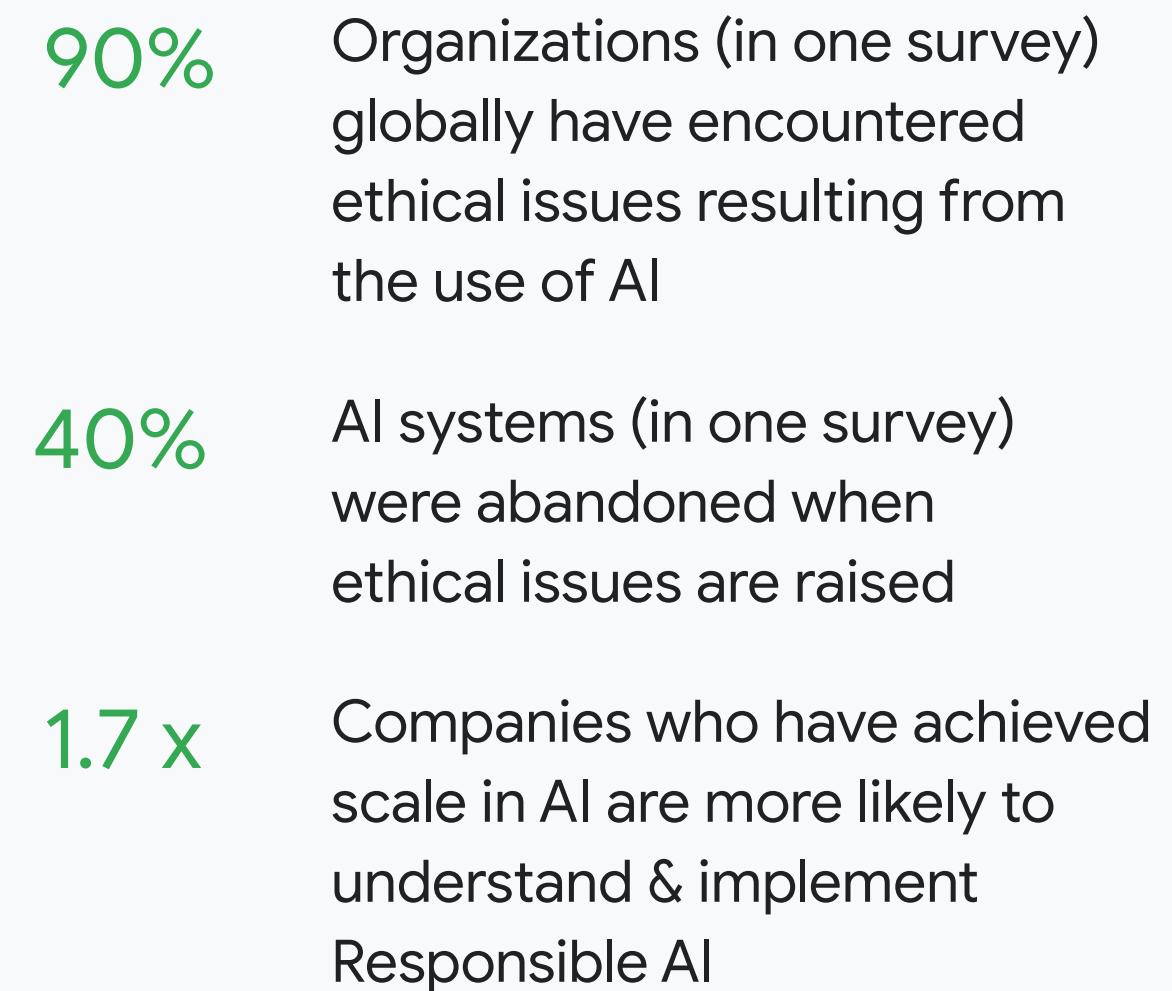
Organizations that achieve AI absorption will be the leaders of the global economy

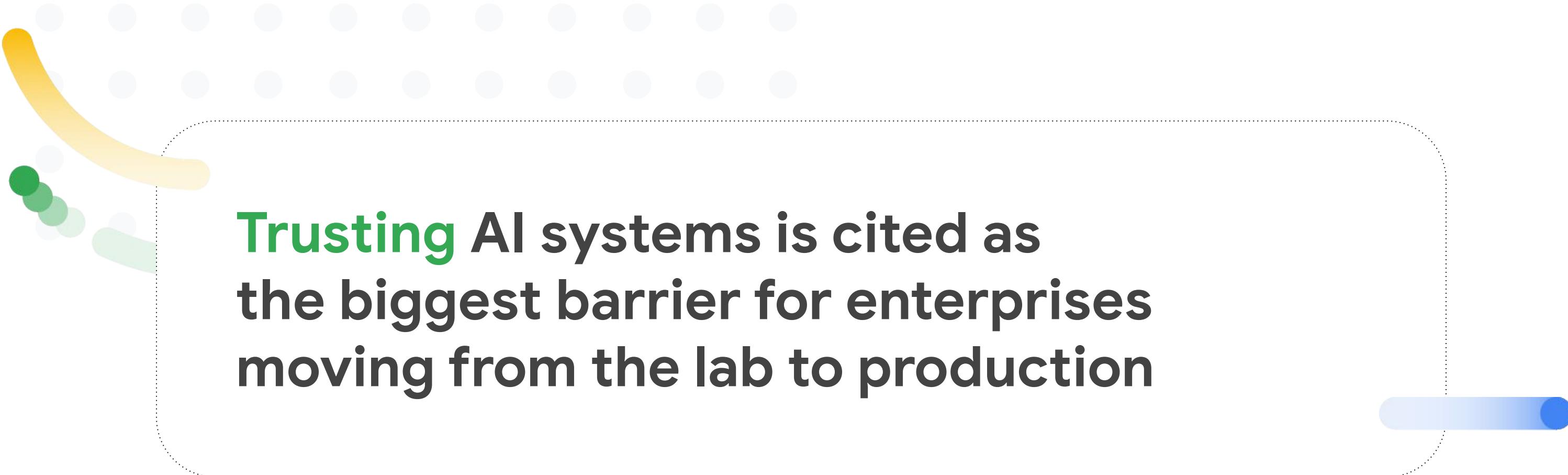
Adopting Responsible AI practices: the research

What fuels ethical issues



Ethics at the center

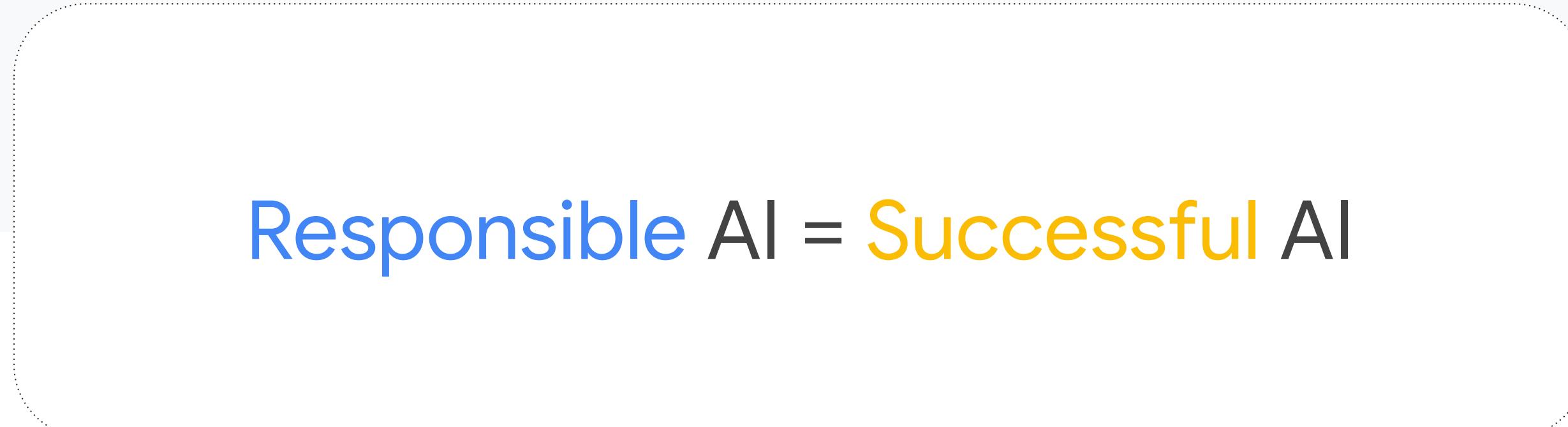




**Trusting AI systems is cited as
the biggest barrier for enterprises
moving from the lab to production**

Sources: CCS Insights IT Decision-Maker Workplace Technology Survey 2019 (pdf)

Google Cloud



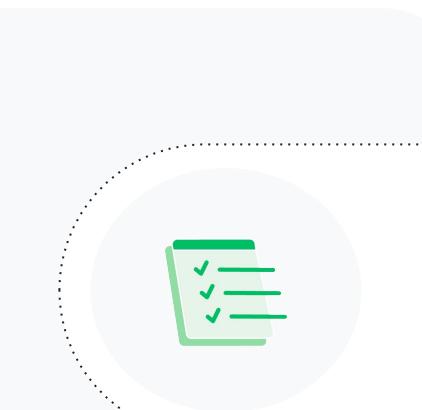
Responsible AI = Successful AI

Adopting Responsible AI practices: our approach

How can I begin to understand and apply the sometimes opposing ethical frameworks in my organization?

How do I manage and govern AI in my organization?
We are excited about AI and development is taking off, but how do we identify the pitfalls before they happen?

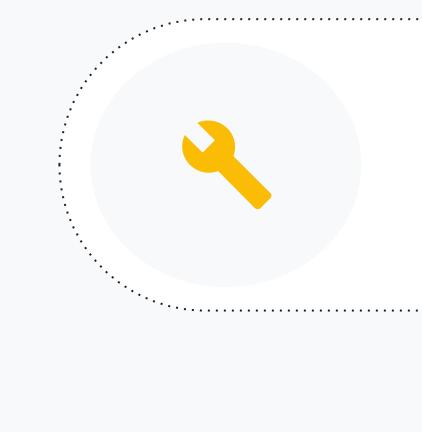
How do I understand the predictions made by AI?
How do I know if my data/model/outcomes perpetuate unfair bias?
How do I know what is influencing my model's predictions?



AI Principles & Practices



Governance Programs & Reviews



Responsible AI Tools



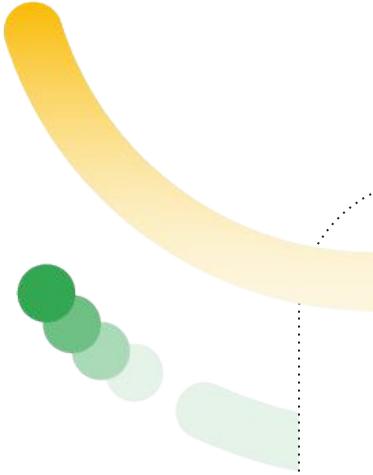
Google AI Principles

AI should:

- 1 be socially beneficial
- 2 avoid creating or reinforcing unfair bias
- 3 be built and tested for safety
- 4 be accountable to people
- 5 incorporate privacy design principles
- 6 uphold high standards of scientific excellence
- 7 be made available for uses that accord with these principles
 - Primary purpose and use
 - Nature and uniqueness
 - Scale
 - Nature of Google's Involvement

Applications we will not pursue:

- 1 likely to cause overall harm
- 2 principal purpose to direct injury
- 3 surveillance violating internationally accepted norms
- 4 purpose contravenes international law and human rights



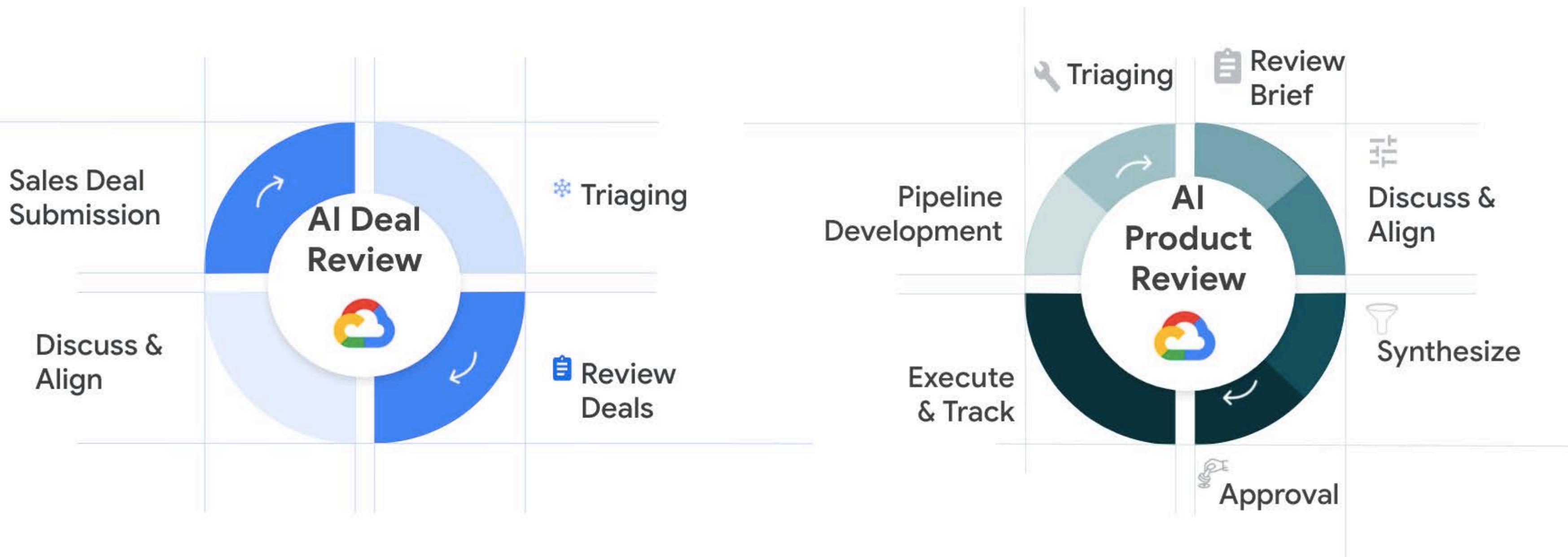
“

Principles that remain on paper
are **meaningless.**

-Sundar Pichai



How we approach this in Cloud



How might you go about this in your organizations?

AKA: Lessons we learned the hard way so you don't have to



* A laundry list or decision tree is not possible.

* There is no “Ethics Checklist.”

* This is critical, and different every time.

* Don’t do this first! Tooling is an aid, not a solution.

02

Responsible AI tooling



Building AI responsibly requires answering hard questions across the ML lifecycle

01

Define Problem

What problem(s) will the model solve?

Who's the intended user?

What are the risks associated with the use-case?

What will 'success' look like?

02

Collect and prepare data

How was the training data collected, sampled, labeled?

Is it representative of the real-world?

Is the training data skewed?

Is the data privacy-protecting?

03

Train model

How was the model trained?

Who trained it? When?

How was the model debugged/improved?

What are the models' limitations?

04

Evaluate

How was the model tested?

What test datasets were used?

Is the model stable, high-performing and safe?

Is the model trustworthy?

05

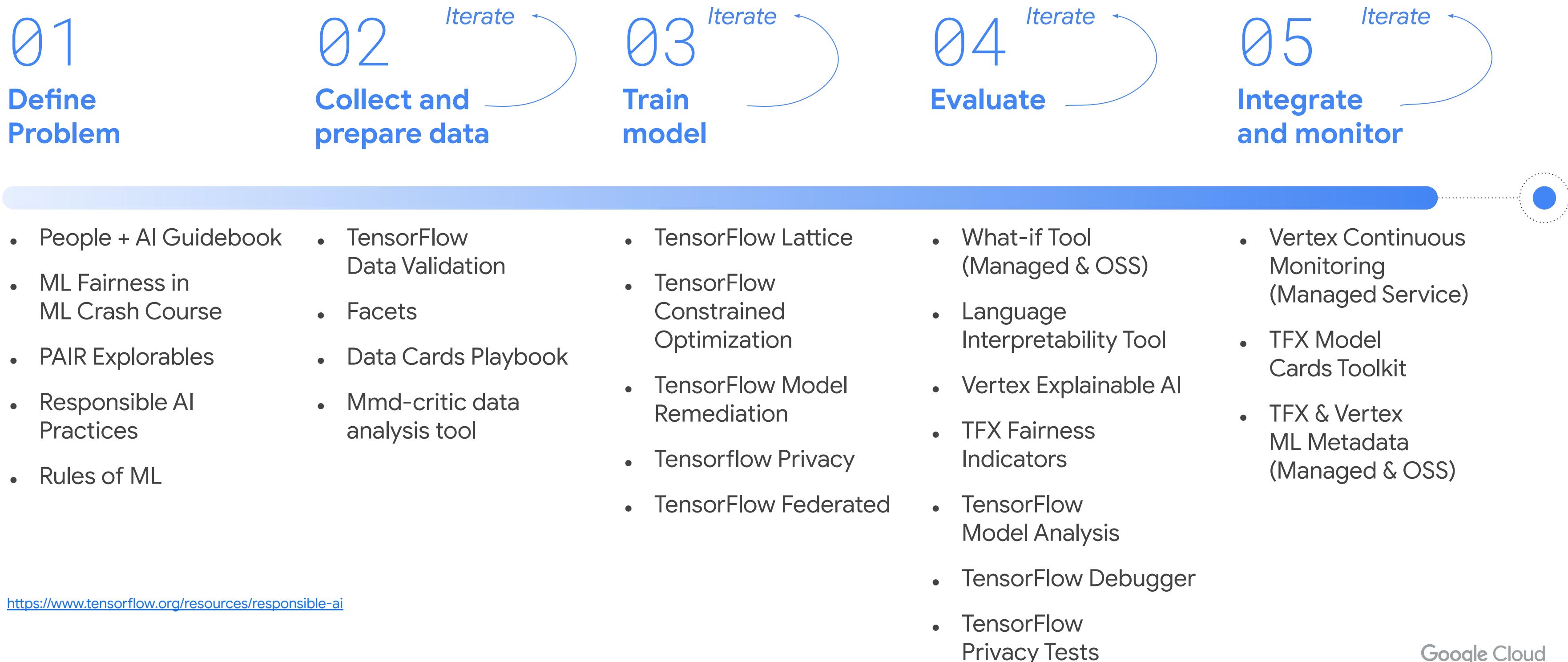
Integrate and monitor

Is the model behaving as expected?

Why did the model fail in this case?



Google and GCP resources and tools for ML developers



Google and GCP resources and tools for ML developers

01

Define Problem

02

Collect and prepare data

03

Train model

04

Evaluate

05

Integrate and monitor

Iterate

Iterate

Iterate

Iterate

- [People + AI Guidebook](#)
- [ML Fairness in ML Crash Course](#)
- PAIR Explorables
- Responsible AI Practices
- Rules of ML

- TensorFlow Data Validation
- Facets
- Data Cards Playbook
- Mmd-critic data analysis tool

- [TensorFlow Lattice](#)
- [TensorFlow Constrained Optimization](#)
- TensorFlow Model Remediation
- Tensorflow Privacy
- TensorFlow Federated

- [What-if Tool \(Managed & OSS\)](#)
- [Language Interpretability Tool](#)
- [Vertex Explainable AI](#)
- [TFX Fairness Indicators](#)
- TensorFlow Model Analysis
- TensorFlow Debugger
- TensorFlow Privacy Tests

- [Vertex Continuous Monitoring \(Managed Service\)](#)
- [TFX Model Cards Toolkit](#)
- TFX & Vertex ML Metadata (Managed & OSS)

<https://www.tensorflow.org/resources/responsible-ai>

Google Cloud



Define the ML problem

People + AI Guidebook

The People + AI Guidebook was written to help user experience (UX) professionals and product managers follow a human-centered approach to AI.



Getting Started

The Guidebook's recommendations are based on data and insights from over a hundred individuals across Google product teams, industry experts, and academic research.

These six chapters follow the product development flow, and each one has a related worksheet to help turn guidance into action.

User Needs + Defining Success

Identify user needs, find AI opportunities, and design your reward function.

↗ Read Chapter ↘ Get Worksheet

Data Collection + Evaluation

Decide what data are required to meet your user needs, source data, and tune your AI.

↗ Read Chapter ↘ Get Worksheet

Mental Models

Introduce users to the AI system and set expectations for system-change over time.

↗ Read Chapter ↘ Get Worksheet

Explainability + Trust

Explain the AI system and determine if, when, and how to show model confidence.

↗ Read Chapter ↘ Get Worksheet

Machine Learning Crash Course

Courses

Search

English



Crash Course

Problem Framing

Data Prep

Clustering

Recommendation

Testing and Debugging

More

Quick Links

- ☰ Overview
- ☰ Prerequisites and Prework
- ☰ Exercises

ML Concepts

- ▶ Introduction to ML (3 min)
- Framing (15 min)
- Descending into ML (20 min)
- Reducing Loss (60 min)
- First Steps with TF (65 min)
- Generalization (15 min)
- Training and Test Sets (25 min)
- Validation Set (35 min)
- Representation (35 min)
- Feature Crosses (70 min)
- Regularization: Simplicity (40 min)
- Logistic Regression (20 min)
- Classification (90 min)
- Regularization: Sparsity (20 min)
- Neural Networks (65 min)
- Training Neural Nets (10 min)
- Multi-Class Neural Nets (45 min)
- Embeddings (50 min)

ML Engineering

- ▶ Production ML Systems (3 min)
- Static vs. Dynamic Training (7 min)
- Static vs. Dynamic Inference (7 min)
- Data Dependencies (14 min)
- Fairness (70 min)

▶ Video Lecture

☰ Types of Bias

Home > Products > Machine Learning > Courses

Rate and review



Send feedback

Fairness

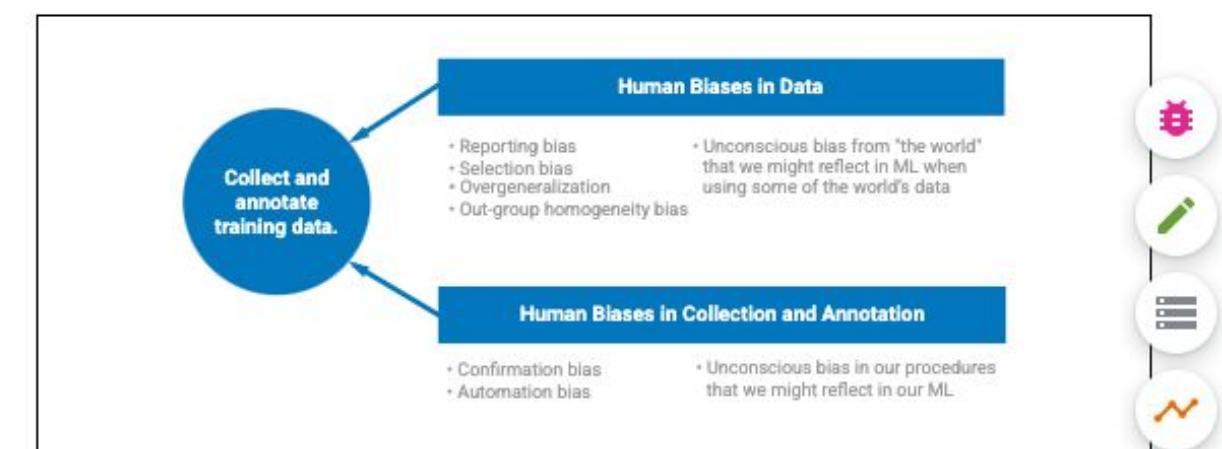
⌚ Estimated Time: 5 minutes

Learning Objectives

- Become aware of common human biases that can inadvertently be reproduced by ML algorithms.
- Proactively explore data to identify sources of bias before training a model
- Evaluate model predictions for bias

Evaluating a machine learning model responsibly requires doing more than just calculating loss metrics. Before putting a model into production, it's critical to audit training data and evaluate predictions for bias.

This module looks at different types of human biases that can manifest in training data. It then provides strategies to identify them and evaluate their effects.

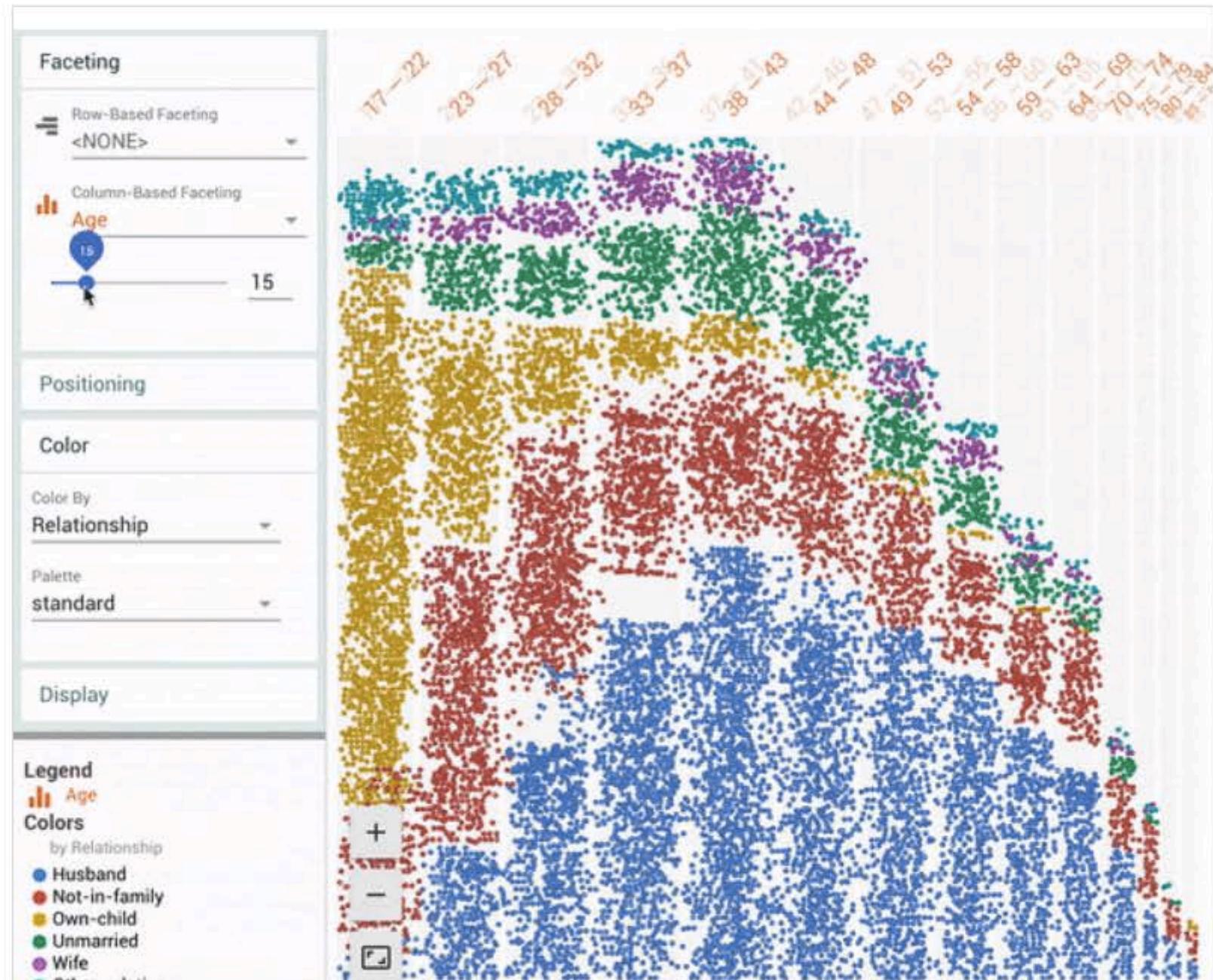




Construct and prepare your data

Facets

An open-source tool for visualizing and inspecting your datasets



pair-code.github.io/facets/



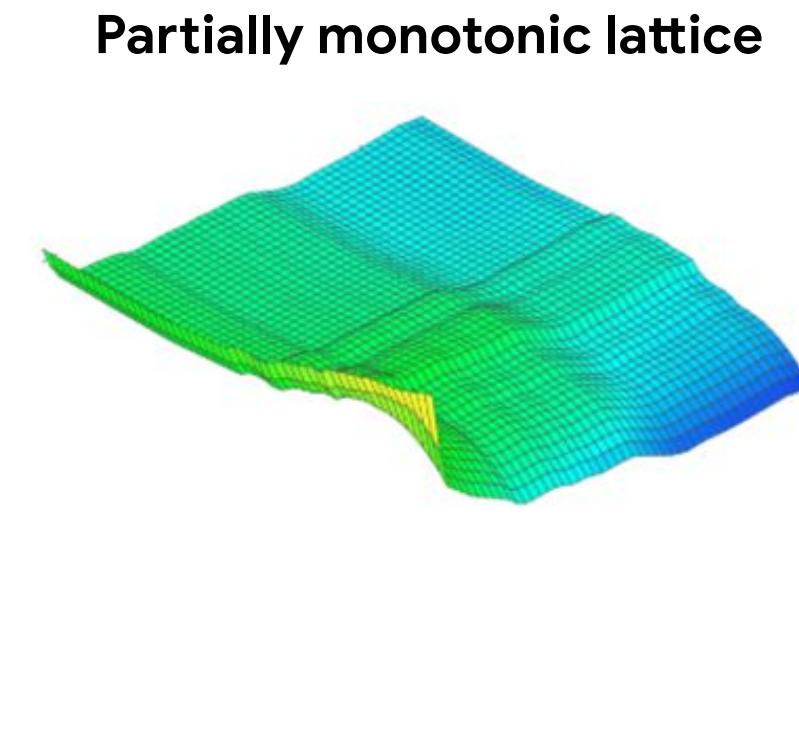
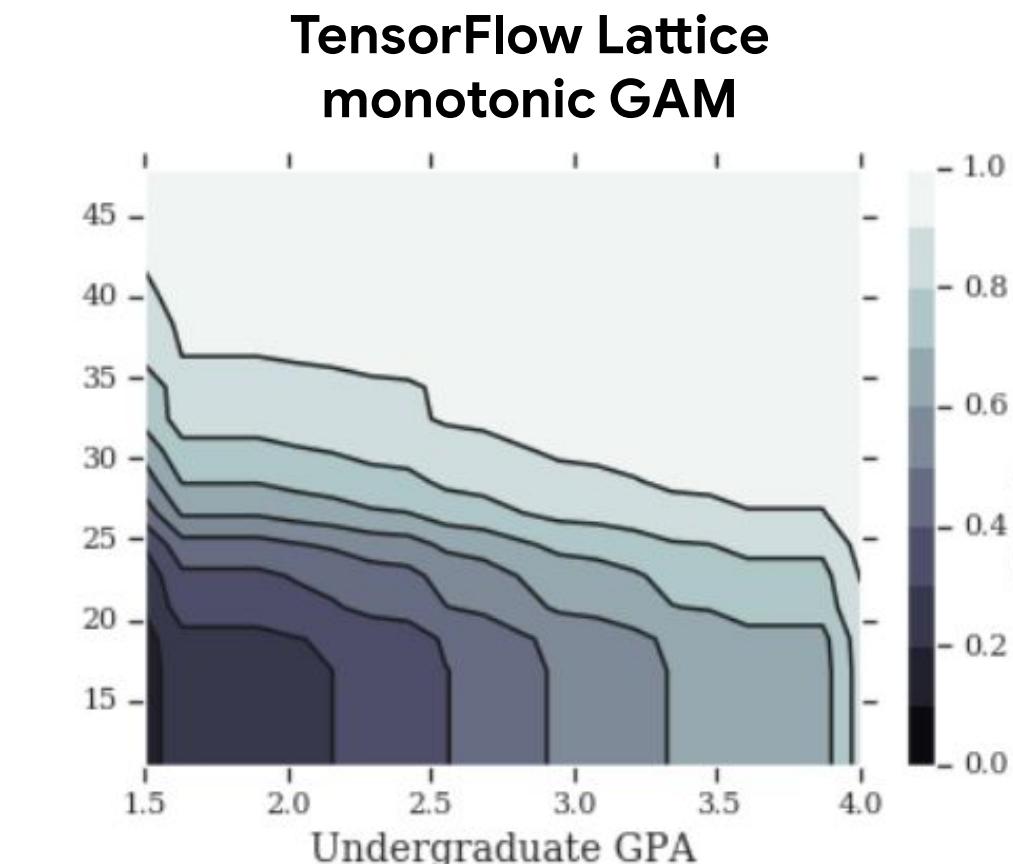
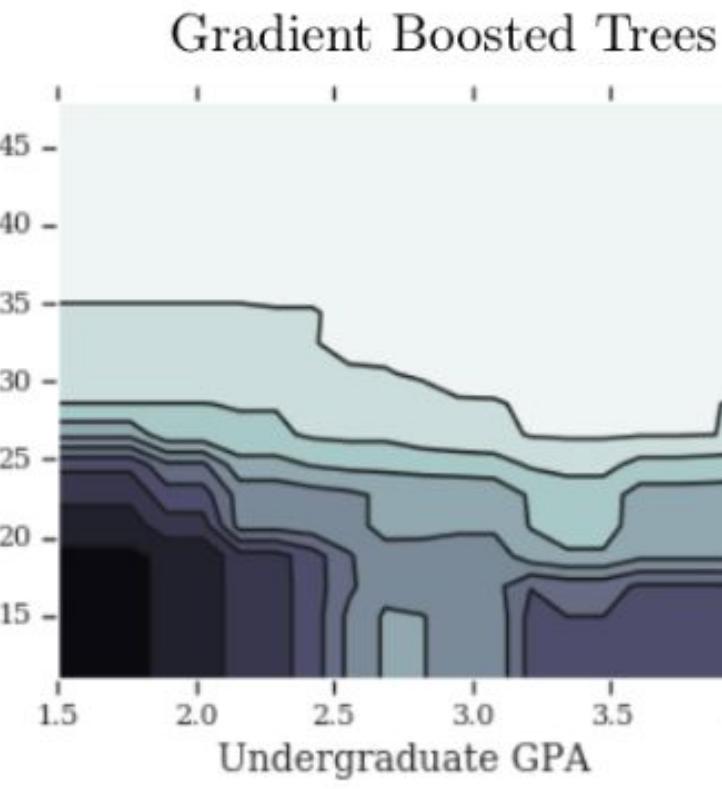
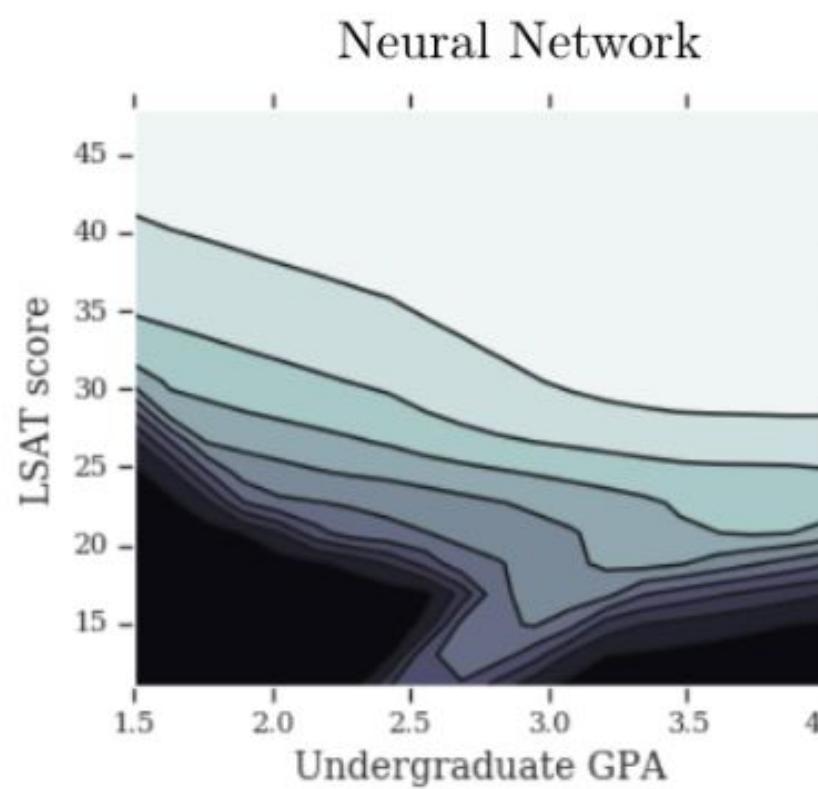
Google Cloud



Build and Train Model

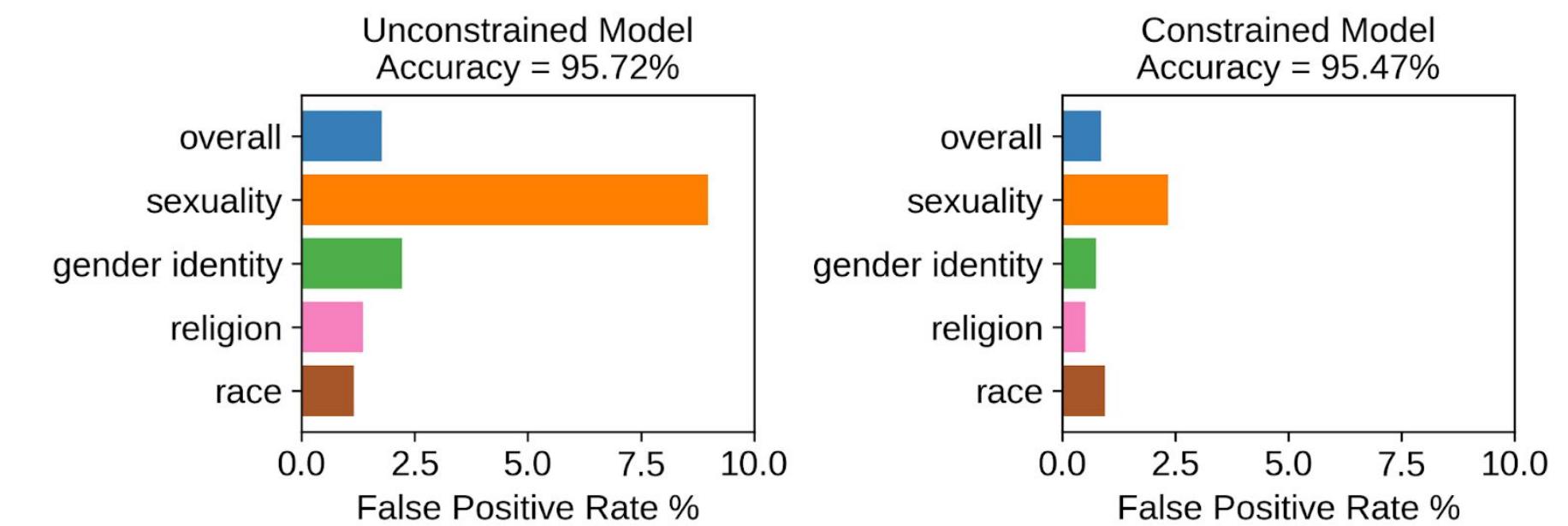
TensorFlow Lattice (TFL)

- Set of modeling layers that enforce semantically meaningful **shape constraints**
 - **Monotonicity:** If x_1 increases, prediction stays the same or increases
 - **Convexity:** If x_1 increases, slope of prediction stays the same or increases
 - **Trust:** If x_1 increases, prediction becomes more sensitive to x_2
 - **Dominance:** Prediction is always more sensitive to x_1 than x_2
 - **Unimodality:** Prediction always has exactly one local maximum in $\{x_1, x_2, x_3, \dots\}$
- Ensembles of 1-D piecewise-linear functions (GAMs) or multi-D functions (lattices)



TensorFlow Constrained Optimization (TFCO)

- Library that enforces semantically meaningful **rate constraints** on models
 - Example rates: Accuracy, precision, recall, fairness, churn, ROC or P/R AUC
 - Optimize/constrain rates: E.g., minimize false positive rate (FPR) s.t. FNR $\leq 5\%$
- Constrains the **actual rates** instead of relaxations (like hinge or sigmoid)
- Constrains model on **any dataset** (or slices of any dataset), not just training set
- User specifies *what they want to accomplish*; TFCO handles *how to do it*
- Works with any gradient-based TensorFlow model, including **TensorFlow Lattice**





Evaluate your model

What-If Tool

Visually probe the behavior of **tabular ML models**, with minimal coding

What-If Tool demo - binary classifier for predicting salary of over \$50k - UCI census income dataset

Partial dependence plots Compute distance Show nearest different classification: L1 L2 ⓘ

PERFORMANCE + FAIRNESS DATAPoint EDITOR FEATURES

Binning | X-Axis Co... Binning | Y-Axis C... Color By
age 10 marital-stat... 1 Inference

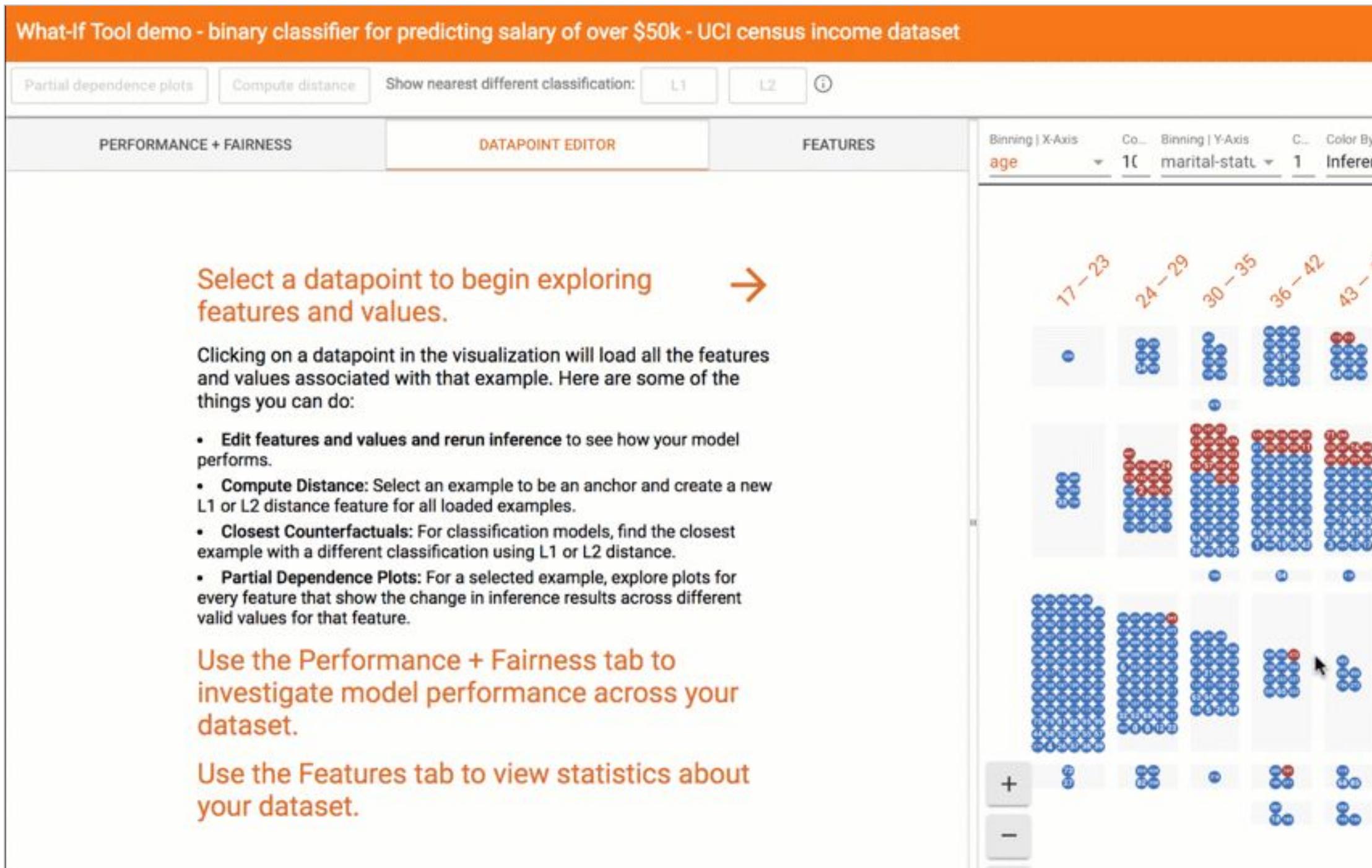
Select a datapoint to begin exploring →

Clicking on a datapoint in the visualization will load all the features and values associated with that example. Here are some of the things you can do:

- Edit features and values and rerun inference to see how your model performs.
- Compute Distance: Select an example to be an anchor and create a new L1 or L2 distance feature for all loaded examples.
- Closest Counterfactuals: For classification models, find the closest example with a different classification using L1 or L2 distance.
- Partial Dependence Plots: For a selected example, explore plots for every feature that show the change in inference results across different valid values for that feature.

Use the Performance + Fairness tab to investigate model performance across your dataset.

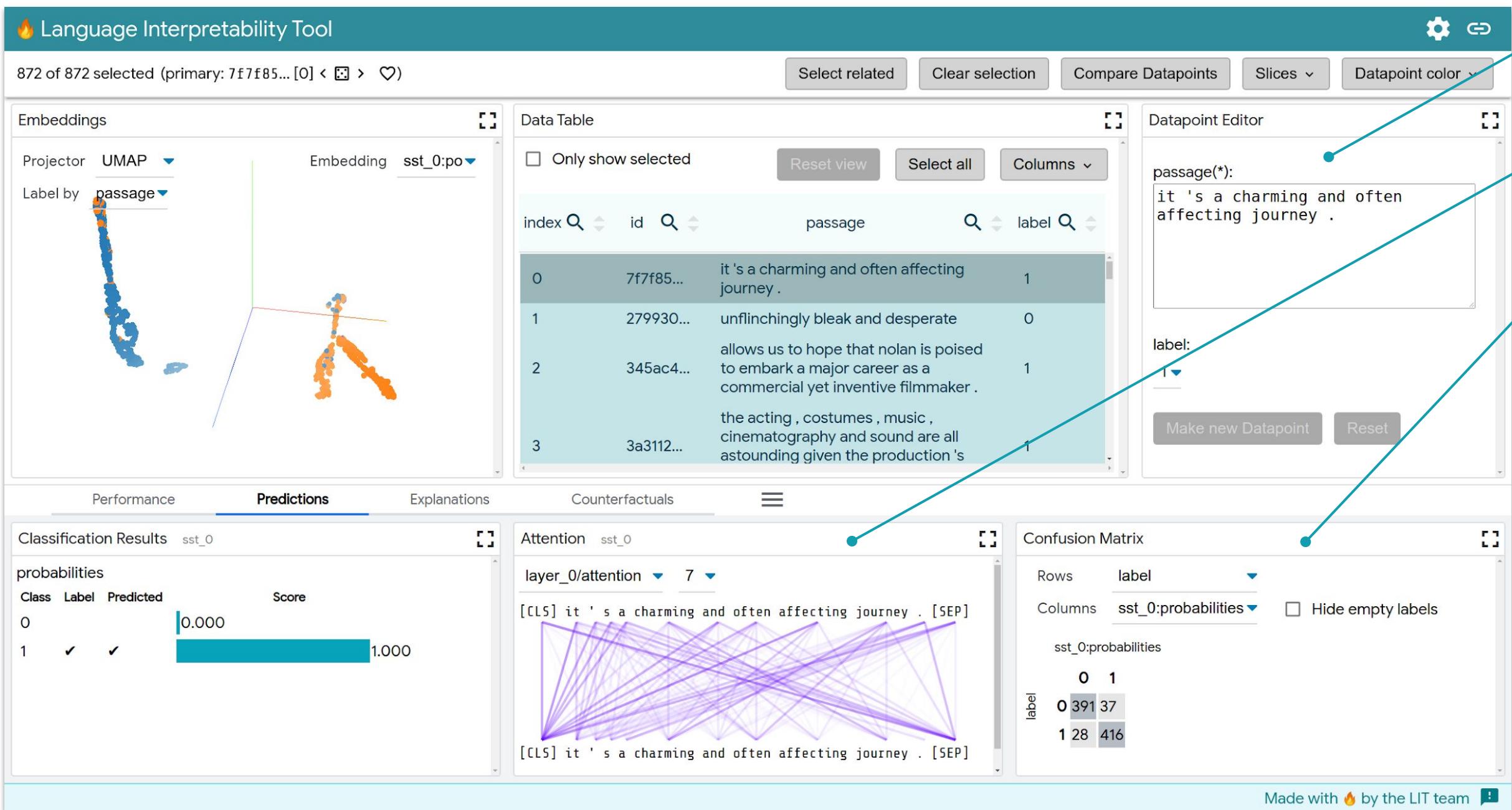
Use the Features tab to view statistics about your dataset.



<https://pair-code.github.io/what-if-tool/>

Language Interpretability Tool

A visual, interactive model-understanding tool for NLP



- ▶ **Counterfactual generation** via manual edits or algorithmic generator plug-ins
- ▶ **Local explanations** via salience maps, attention, and rich visualizations
- ▶ **Aggregate analysis** with custom metrics, slice-by-feature, and embedding clusters
- ▶ **Side-by-side mode** to compare two models, or a pair of examples
- ▶ **Highly extensible** to new model types, including classification, regression, structured prediction, and seq2seq
- ▶ **Framework agnostic** and compatible with TensorFlow, PyTorch, and more

ML Fairness Indicators

Compute & visualize ML fairness metrics in TFX pipelines

Evaluations Slices Fairness

Configure

Settings for model

Model Version: No version selected **WHAT IS VERSION?** Sets the version of the model that you are interested in evaluating.

Model Run: No model run selected **WHAT IS MODEL RUN?** Defines the specific model run for which you are interested in evaluating fairness performance.

Settings for evaluation dataset

Evaluation Set: No dataset selected **WHAT IS THE EVALUATION SET?** The dataset that you want to evaluate model performance on. See tips on creating evaluation datasets.

Evaluation Feature: select a feature **WHAT IS EVALUATION FEATURE?** Sets the feature of your dataset for which you want to examine fairness metrics.

Baseline Slice: Overall (default) **WHAT IS BASELINE?** Define a slice of your feature as a baseline to compare all other slices to. Selecting "overall" compares individual slices to the average performance.

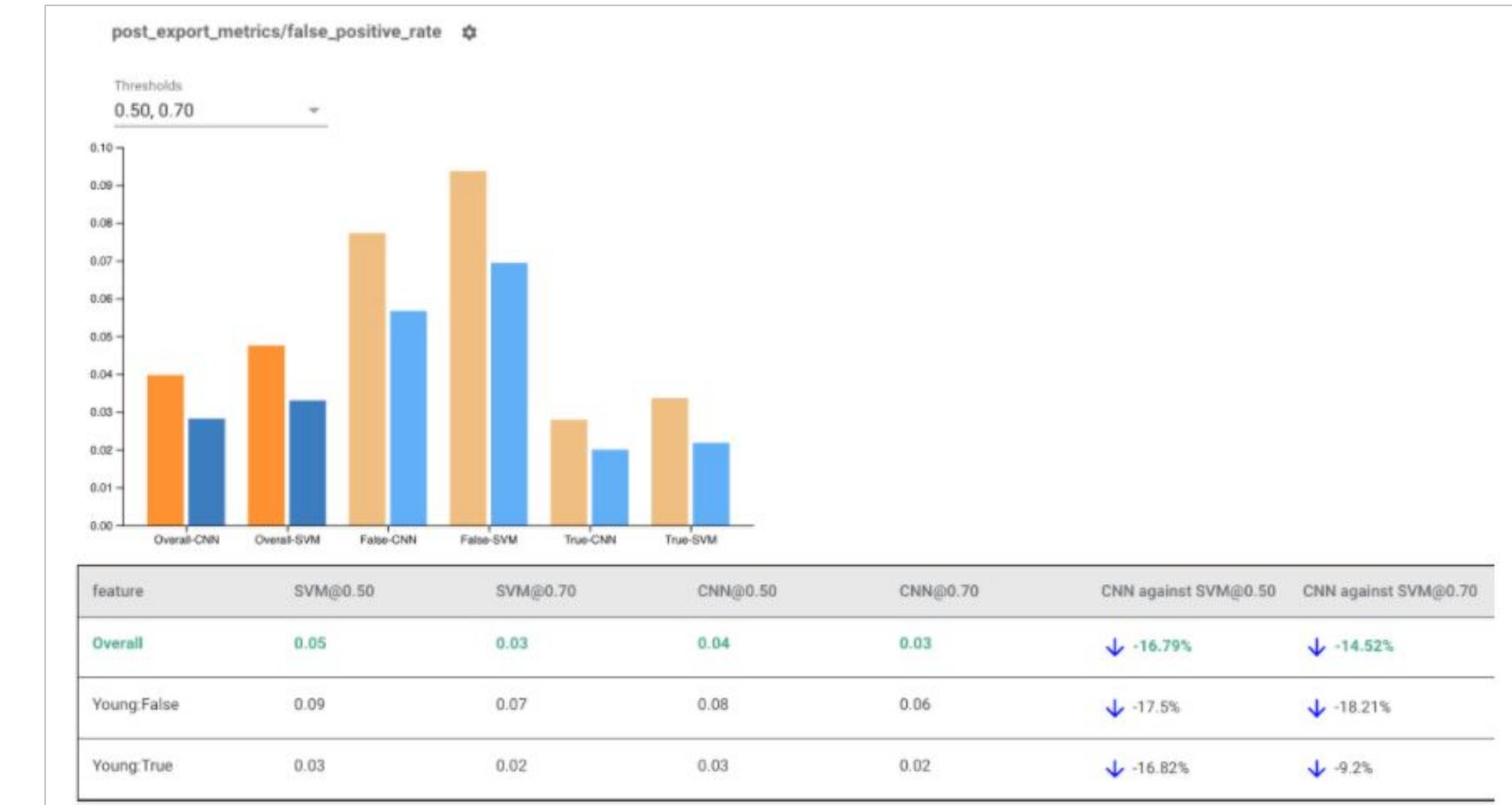
Apply

Fairness Metrics

Select metrics to visualize

Independent metrics are visualized by individual slices in both table and histogram form. All selected metrics are concatenated in the full table.

False Positive Rate *i*
 False Negative Rate *i*
 True Positive Rate *i*
 True Negative Rate *i*
 Precision *i*
 Accuracy *i*
 Show full table



tensorflow.org/tfx/guide/fairness_indicators

Google Cloud

Vertex Explainable AI managed service

Understand the ‘why’ behind ML models & predictions



Robust, actionable explanations

“Feature attributions” show you which input features are most important to your model overall and for specific predictions.



Built into multiple AI Platform services

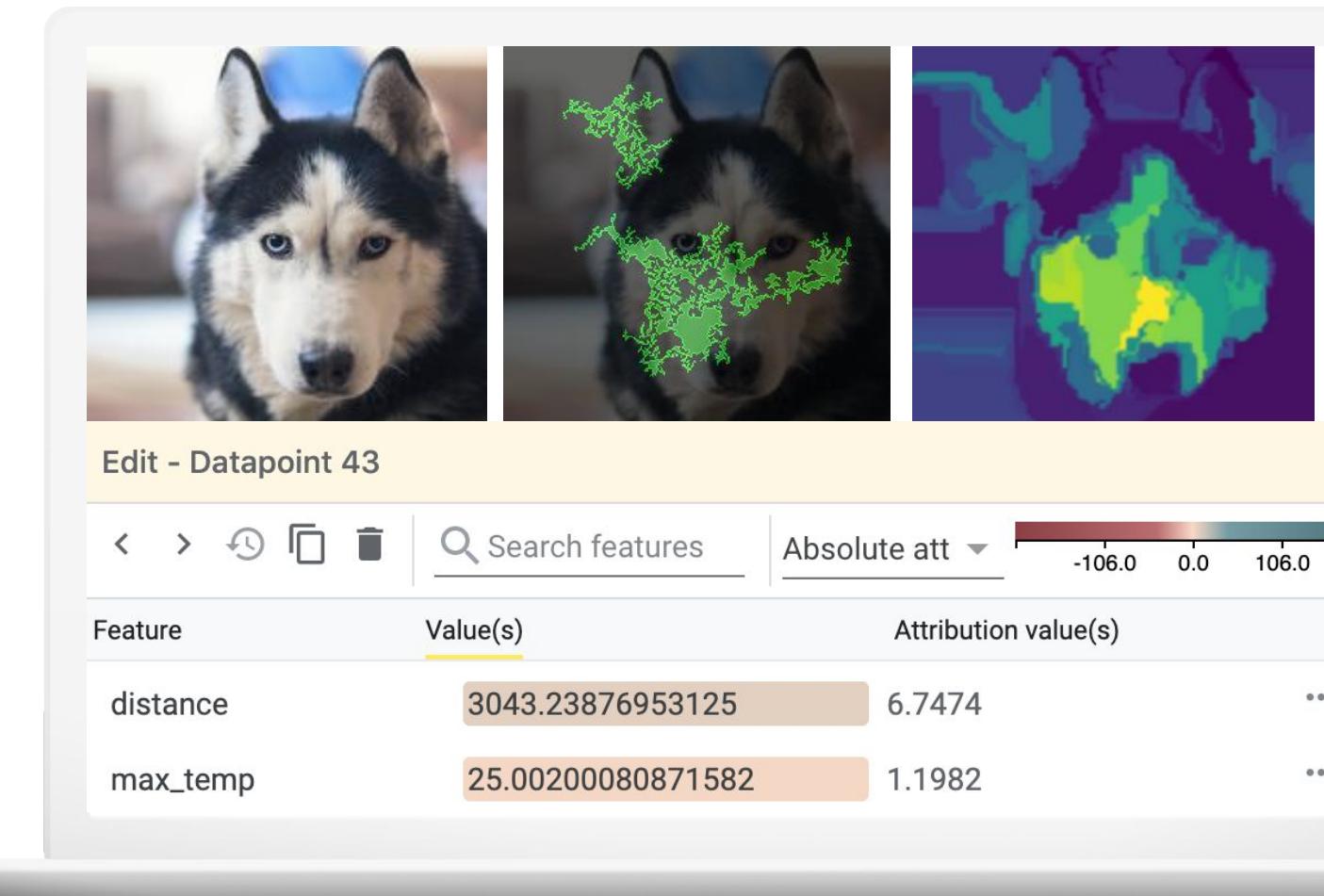
Get explanations easily through Vertex Prediction, AutoML Tables, and Notebooks (+ more coming!)



Flexible, fast & scalable

Supports tabular, image & text models from any ML framework, Online & Batch processing use-cases.

Fully managed, serverless, and significantly faster than open-source alternatives.

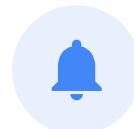




Integrate and Monitor

Vertex Model Monitoring

Easy and proactive monitoring of model performance



Monitor and alert

Monitor signals for model's predictive performance, and alert when those signals deviate.



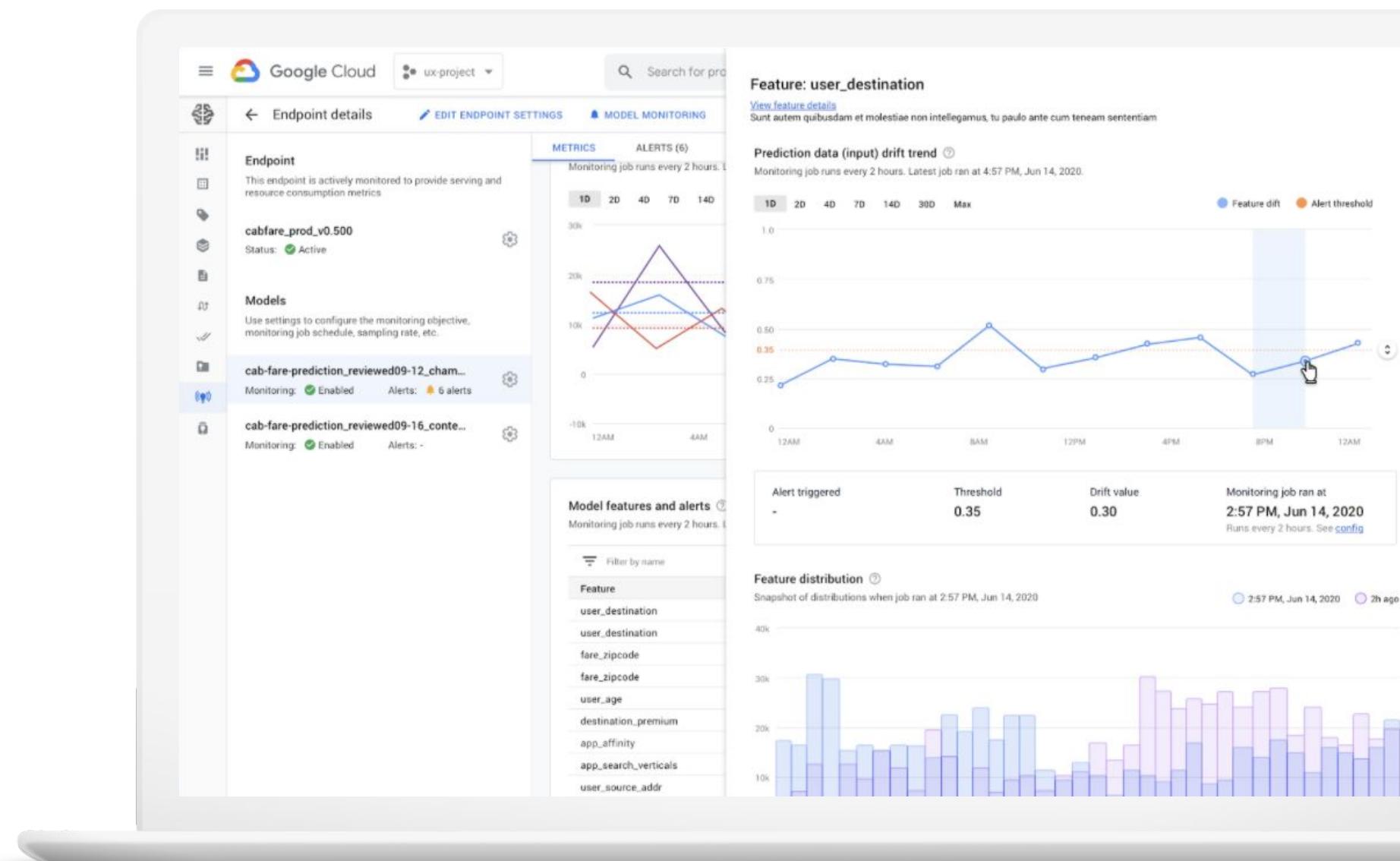
Diagnose

Help identify the cause for the deviation i.e. what changed, how and how much?



Update Model

Trigger model re-training pipeline or collect relevant training data to address performance degradation.



Model Cards & Model Card Toolkit

Organize and communicate the essential facts of your models in a structured way

[←](#)

Face Detection

Model Card v0 - Cloud Vision API

[Overview](#) [Limitations](#) [Trade-offs](#) [Performance](#) [Test your own images](#) [Provide feedback](#)

[Explore](#)

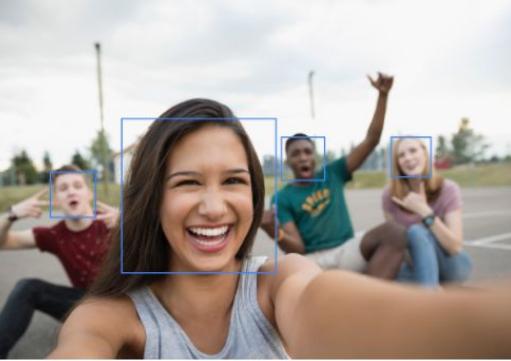
- [Object Detection](#)
- [About Model Cards](#)

Face Detection

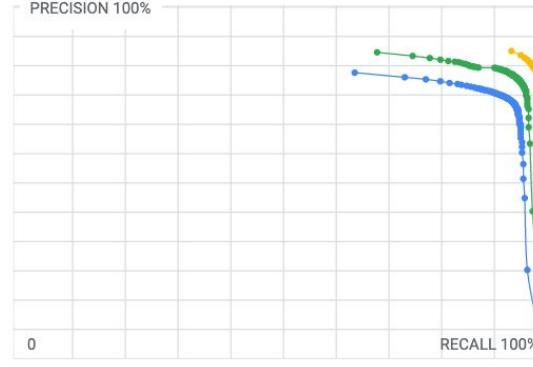
The model analyzed in this card detects one or more faces within an image or a video frame, and returns a box around each face along with the location of the faces' major landmarks. The model's goal is exclusively to identify the existence and location of faces in an image. It does not attempt to discover identities or demographics.

On this page, you can learn more about how well the model performs on images with different characteristics, including face demographics, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION



PERFORMANCE



Precision 100%
RECALL 100%

Legend: ● Open Images ● Face Detection Dataset Benchmark
● Labeled Faces in the Wild

Overall model performance, and performance sliced by different image and face characteristics, were assessed, including:

- Derived characteristics (face size, facial orientation, and occlusion)
- Face demographics (human-perceived gender presentation, age, and skin tone)

No identity or demographic information is detected.

Model Card for Census Income Classifier

Model Details

Overview

This is a wide and deep Keras model which aims to classify whether or not an individual has an income of over \$50,000 based on various demographic features. The model is trained on the UCI Census Income Dataset. This is not a production model, and this dataset has traditionally only been used for research purposes. In this Model Card, you can review quantitative components of the model's performance and data, as well as information about the model's intended uses, limitations, and ethical considerations.

Version
name: 36dea2e860670aa74691b5695587afe7

Owners

- Model Cards Team, model-cards@google.com

References

- interactive-2020-07-28T20_17_47.911887

Considerations

Use Cases

- This dataset that this model was trained on was originally created to support the machine learning community in conducting empirical analysis of ML algorithms. The Adult Data Set can be used in fairness-related studies that compare inequalities across sex and race, based on people's annual incomes.

Limitations

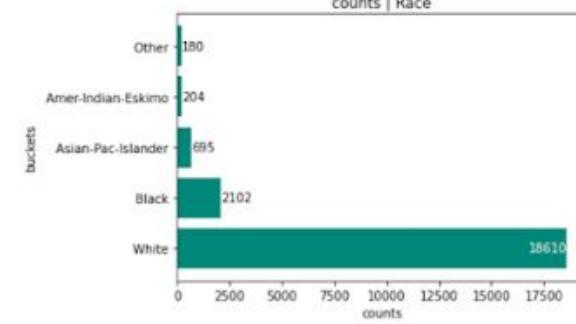
- This is a class-imbalanced dataset across a variety of sensitive classes. The ratio of male-to-female examples is about 2:1 and there are far more examples with the "white" attribute than every other race combined. Furthermore, the ratio of \$50,000 or less earners to \$50,000 or more earners is just over 3:1. Due to the imbalance across income levels, we can see that our true negative rate seems quite high, while our true positive rate seems quite low. This is true to an even greater degree when we only look at the "female" sub-group, because there are even fewer female examples in the \$50,000+ earner group, causing our model to overfit these examples. To avoid this, we can try various remediation strategies in future iterations (e.g. undersampling, hyperparameter tuning, etc), but we may not be able to fix all of the fairness issues.

Ethical Considerations

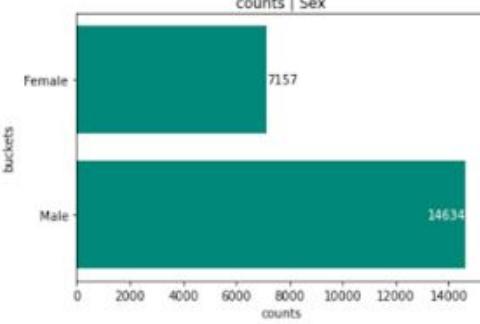
- Risk: We risk expressing the viewpoint that the attributes in this dataset are the only ones that are predictive of someone's income, even though we know this is not the case.
Mitigation Strategy: As mentioned, some interventions may need to be performed to address the class imbalances in the dataset.

Train Set

This section includes graphs displaying the class distribution for the "Race" and "Sex" attributes in our training dataset. We chose to show these graphs in particular because we felt it was important that users see the class imbalance.



buckets	counts Race
Other	180
Amer-Indian-Eskimo	204
Asian-Pac-Islander	695
Black	2102
White	18610



buckets	counts Sex
Female	7157
Male	14634

Learn more about these and other tools at...

 [TensorFlow](#) [Install](#) [Learn](#) ▾ [Resources](#) ▾ [More](#) ▾ [!\[\]\(fac3df7ac48ec2707f62dd77e89888f8_img.jpg\) Search](#) [English](#) ▾ [GitHub](#) [Sign in](#)

What is Responsible AI?

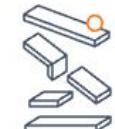
The development of AI is creating new opportunities to solve challenging, real-world problems. It is also raising new questions about the best way to build AI systems that benefit everyone.


Recommended best practices for AI

Designing AI systems should follow software development best practices while taking a human-centered approach to ML


Fairness

As the impact of AI increases across sectors and societies, it is critical to work towards systems that are fair and inclusive to everyone


Interpretability

Understanding and trusting AI systems is important to ensuring they are working as intended


Privacy

Training models off of sensitive data needs privacy preserving safeguards


Security

Identifying potential threats can help keep AI systems safe and secure

 [Google Cloud](#) [Why Google](#) [Solutions](#) [Products](#) [Pricing](#) [Getting Started](#) [Contact Us](#) [!\[\]\(8e2b6fa99613fbe158bbb17338f0c874_img.jpg\) Search](#) [Docs](#) [Support](#) [English](#) ▾ [Console](#)

Responsible AI

Responsible AI

- [Overview](#)
- [Benefits](#)
- [Key features](#)
- [What's new](#)
- [Take the next step](#)

[Contact us](#)

BENEFITS

How values-based AI is good for your business

<p>Safer and more accountable products</p> <p>Advanced technologies are most successful when everyone can benefit from them. Evaluating your AI systems, both when they perform as intended and when they don't, is crucial to building accountable products.</p>	<p>Earn and keep your customers' trust</p> <p>Lack of trust in AI systems is a growing barrier to adoption in enterprise with more organizations selecting enterprise products based on AI commitments and practices. A responsible AI approach earns trust.</p>	<p>A culture of responsible innovation</p> <p>Empowering AI decision-makers and developers to take ethical considerations into account enables them to find new, innovative ways to drive your mission forward.</p>
--	---	--

Learn more about our [perspectives on issues and AI governance](#) at Google and how we work to [build responsible AI for everyone](#).

tensorflow.org/responsible_ai

cloud.google.com/responsible-ai

03



Case Study: Duke Sepsis Watch

Sepsis

- 
- 01 **Leading cause of inpatient mortality within hospitals in the US**
 - 02 **Recognized by WHO as a major cause of preventable disease and death globally**
 - 03 **Difficult to diagnose; no universally agreed upon definition**
 - 04 **Treatable if diagnosed in time**

Duke Institute for Health Innovation Sepsis Watch

A system combining a deep learning model that predicts the risk of a patient developing sepsis with new hospital protocols to raise the quality of treatment.

Sepsis Watch: the implementation of a Duke-specific early warning system for sepsis

Think beyond detection.

Home / Projects / Sepsis Watch: the implementation of a Duke-specific early warning system for sepsis

Related Results & Outcomes

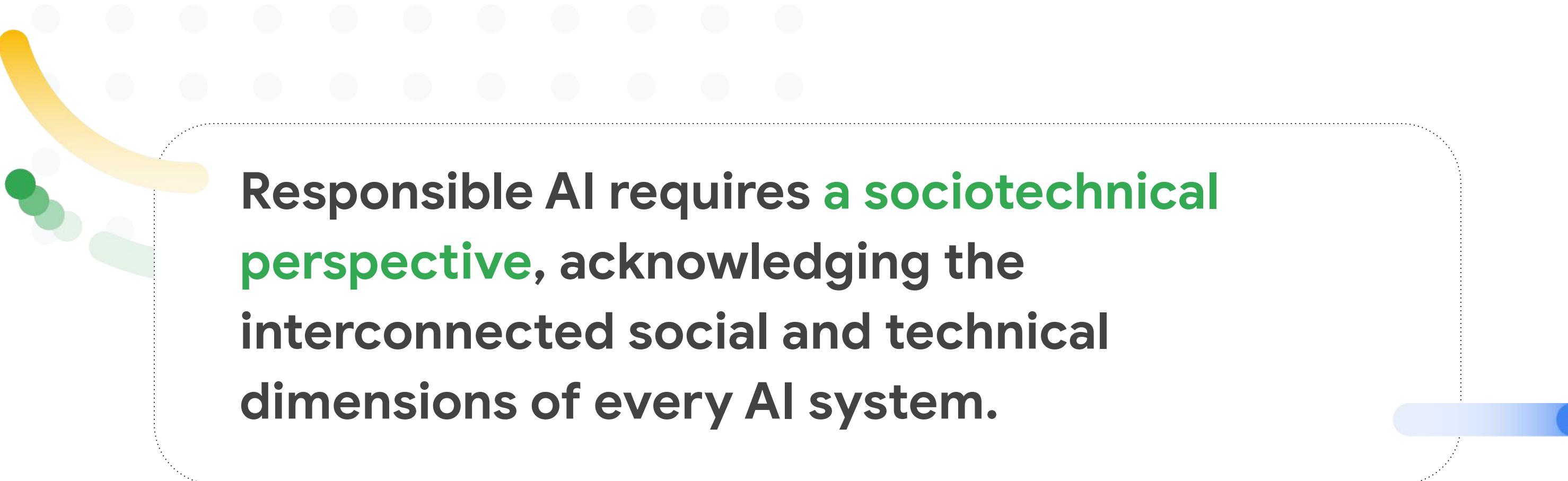
- An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection
- 'The Human Body Is a Black Box': Supporting Clinical Decision-Making with Deep Learning
- Deep Sepsis: a look "under the hood" at the model powering Sepsis Watch

“

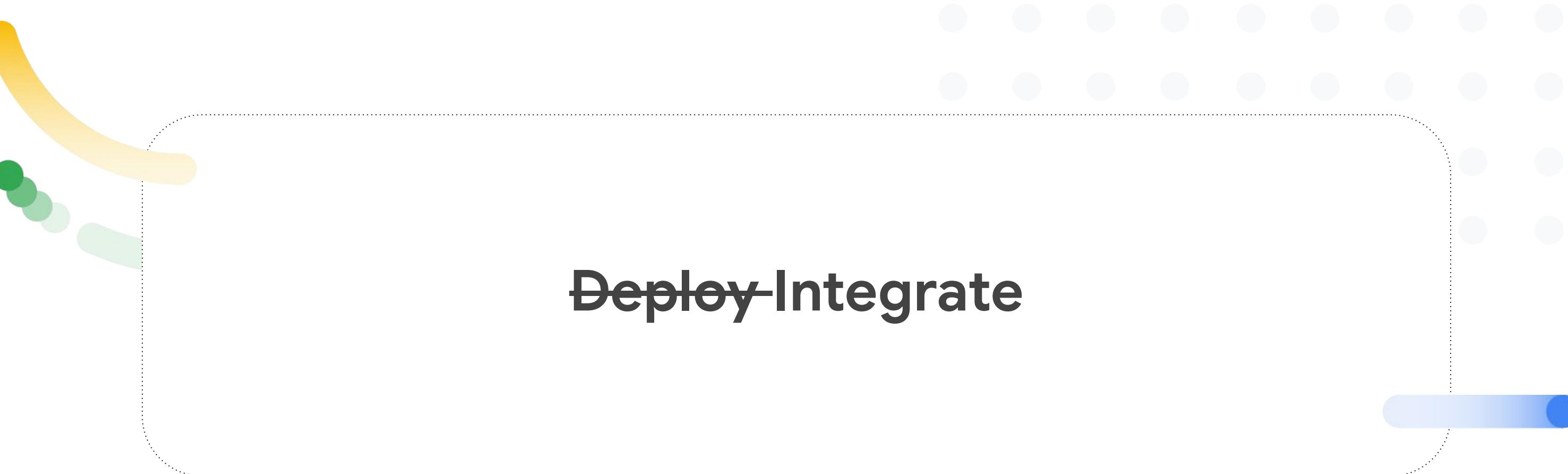
Hospitals Fight Sepsis with AI”

IEEE Spectrum



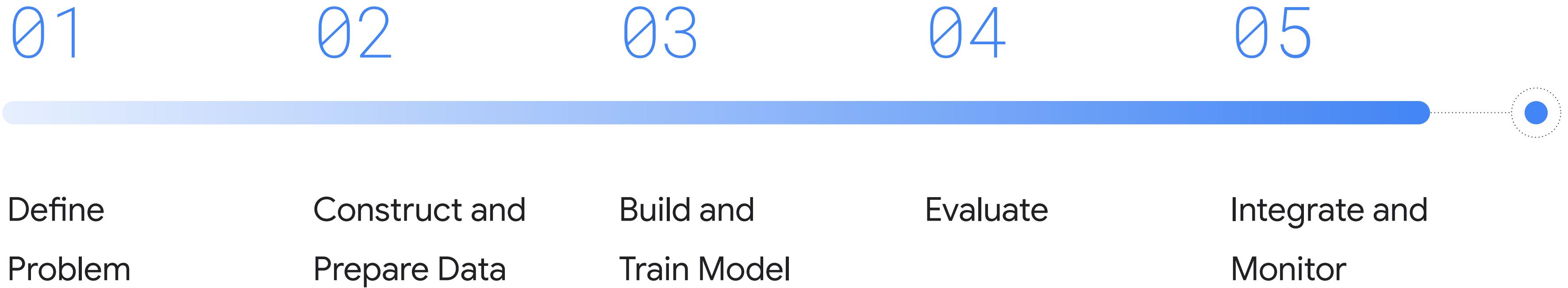


Responsible AI requires a sociotechnical perspective, acknowledging the interconnected social and technical dimensions of every AI system.



Deploy Integrate

The responsible AI workflow





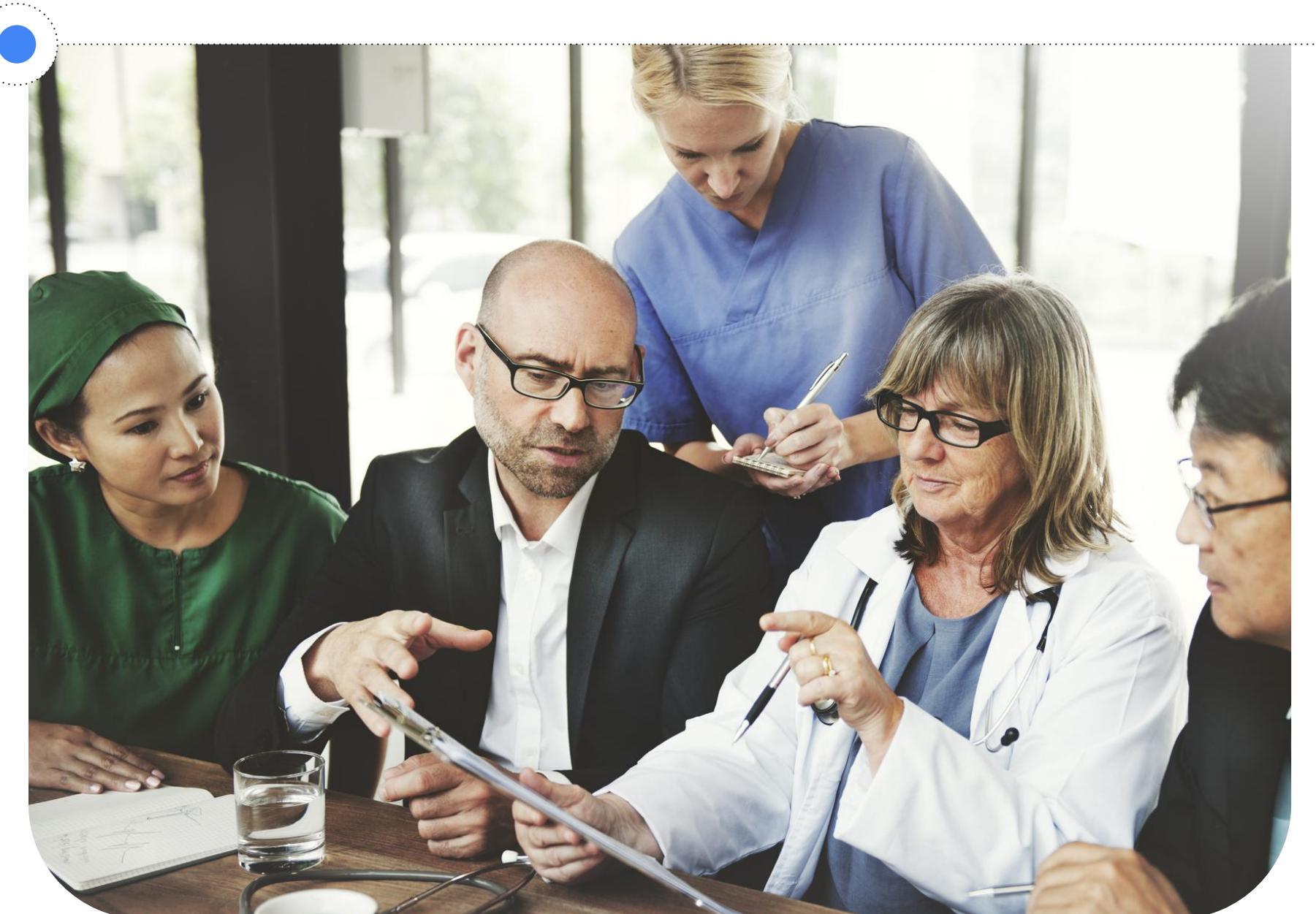
Understand and **Define the Problem**

Is this problem a good one for AI/ML?

What risks are associated with my use case?

Local expertise

Stakeholder engagement was prioritized in problem understanding phase and throughout the design process.





02

Construct and Prepare Data

03

Build and Train Model

04

Evaluate model

Local context shaped data, model training, and evaluation

Clinicians helped lead institution-specific validation and evaluation.



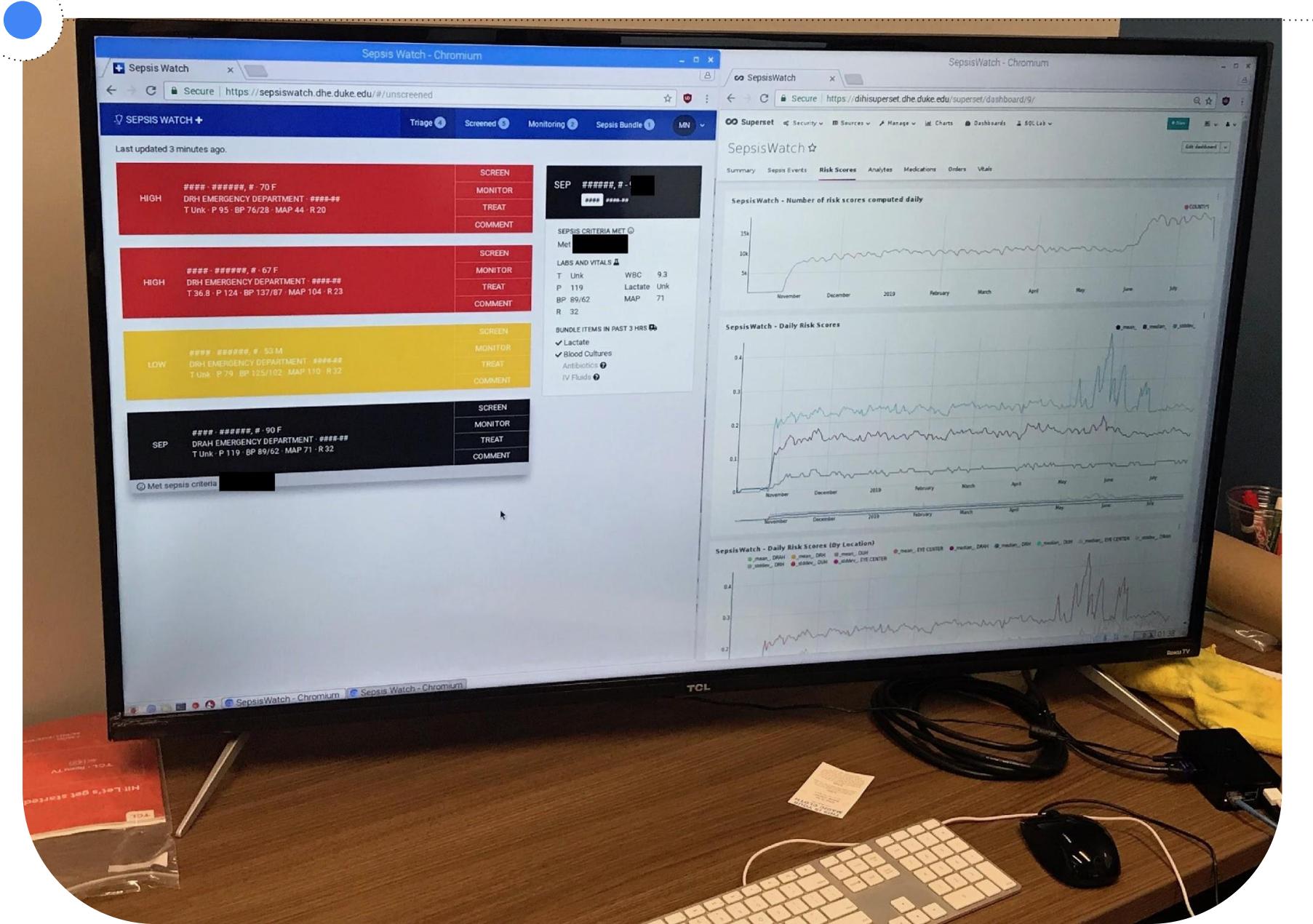


Integrate and Monitor

What should others know about my model and its implications for use?

Ongoing monitoring

A 55' monitor on a desk at the office of the innovation team.



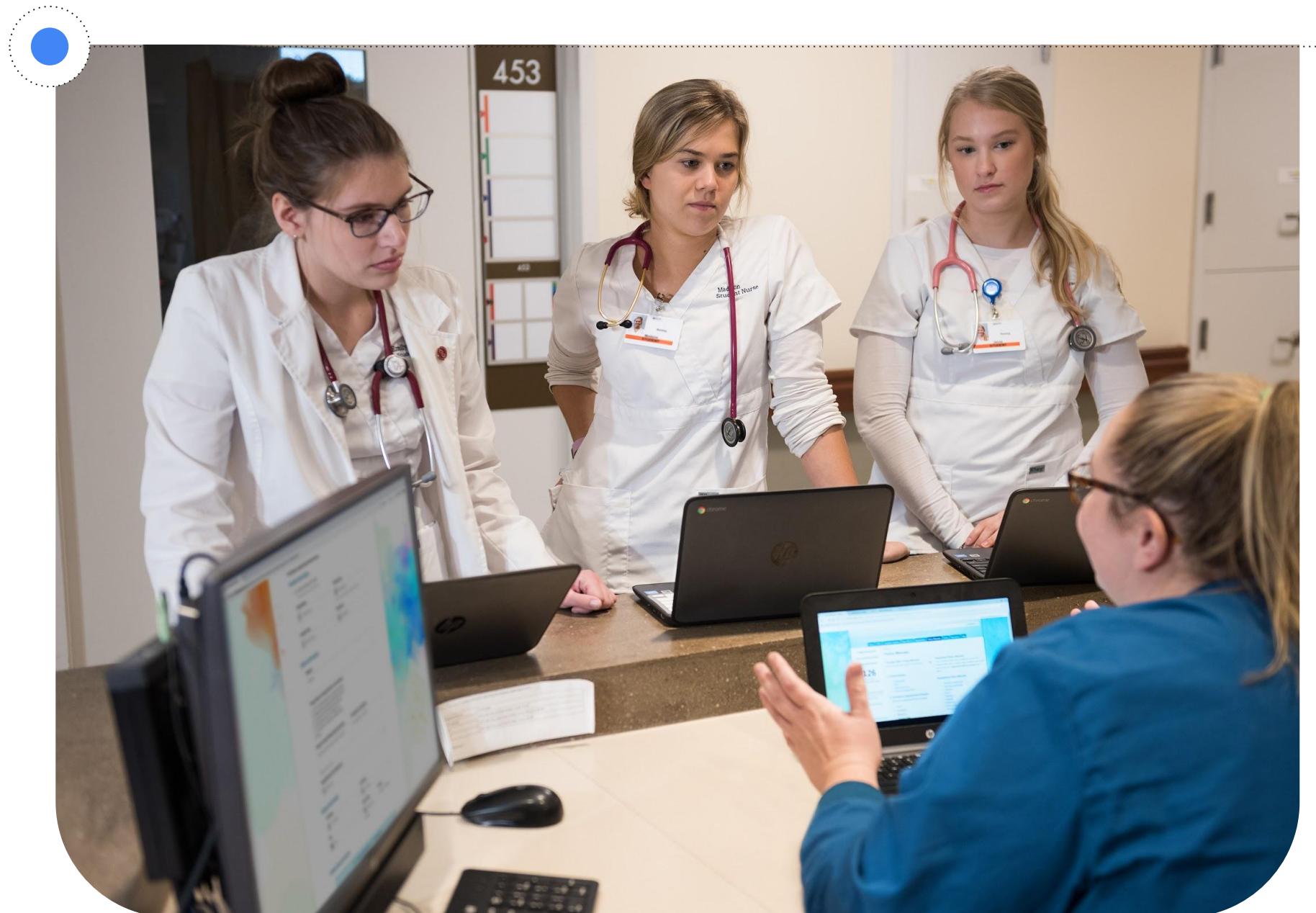
Model Card

Stakeholder engagement was prioritized in problem definition phase and throughout the design process.

Model Facts	Model name: Deep Sepsis	Locale: Duke University Hospital																									
Approval Date: 09/22/2019	Last Update: 09/24/2019.	Version: 1.0																									
Summary This model uses EHR input data collected from a patient's current inpatient encounter to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.																											
Mechanism <ul style="list-style-type: none">▪ Outcomesepsis within the next 4 hours, see (1) for sepsis criteria▪ Output0% - 100% probability of sepsis occurring in the next 4 hours▪ Patient populationall adult patients >18 y.o. presenting to DUH ED and admitted▪ Time of predictionevery hour of a patient's encounter▪ Input data source.....electronic health record (EHR)▪ Input data typedemographics, analytes, vitals, medication administrations▪ Training data location and time-periodDUH, 10/2014 – 12/2015▪ Model type.....Recurrent Neural Network																											
Validation and performance <table border="1"><thead><tr><th></th><th>Prevalence</th><th>AUC</th><th>PPV @ Sensitivity of 60%</th><th>Sensitivity @ PPV of 20%</th></tr></thead><tbody><tr><td>Local Retrospective</td><td>18.9%</td><td>0.88</td><td>0.14</td><td>0.50</td></tr><tr><td>Local Temporal</td><td>6.4%</td><td>0.94</td><td>0.20</td><td>0.66</td></tr><tr><td>Local Prospective</td><td>TBD</td><td>TBD</td><td>TBD</td><td>TBD</td></tr><tr><td>External</td><td>TBD</td><td>TBD</td><td>TBD</td><td>TBD</td></tr></tbody></table>				Prevalence	AUC	PPV @ Sensitivity of 60%	Sensitivity @ PPV of 20%	Local Retrospective	18.9%	0.88	0.14	0.50	Local Temporal	6.4%	0.94	0.20	0.66	Local Prospective	TBD	TBD	TBD	TBD	External	TBD	TBD	TBD	TBD
	Prevalence	AUC	PPV @ Sensitivity of 60%	Sensitivity @ PPV of 20%																							
Local Retrospective	18.9%	0.88	0.14	0.50																							
Local Temporal	6.4%	0.94	0.20	0.66																							
Local Prospective	TBD	TBD	TBD	TBD																							
External	TBD	TBD	TBD	TBD																							
Uses and directions <ul style="list-style-type: none">▪ Operational use case(s): Every hour, data is pulled from the EHR to calculate risk of sepsis for every patient at the DUH ED. A rapid response team nurse reviews every high-risk patient with a physician in the ED to confirm whether or not to initiate treatment for sepsis.▪ General use: This model is intended to be used to by clinicians to identify patients for further assessment for sepsis. The model is not a diagnostic for sepsis and is not meant to guide or drive clinical care. This model is intended to complement other pieces of patient information related to sepsis as well as a physical evaluation to determine the need for sepsis treatment.▪ Examples of appropriate decisions to support: Patient X has a high risk of sepsis according to the model. A rapid response team nurse discusses the patient with the ED physician caring for the patient and they agree the patient does not require treatment for sepsis.▪ Before using this model: Test the model retrospectively and prospectively on local data to confirm generalizability of the model to the local setting.▪ Safety and efficacy evaluation: Analysis of data from clinical trial (NCT03655626) underway. Preliminary data shows rapid response team, nurse-driven workflow was effective at improving sepsis treatment bundle compliance.																											

Clinical expertise

Nurses played an unanticipated but essential role in effectively integrating Sepsis Watch into clinical care.



Every responsible AI story is different.



What's yours?

inclusive-ml-feedback@google.com

cloud.google.com/responsible-ai



Thank you.