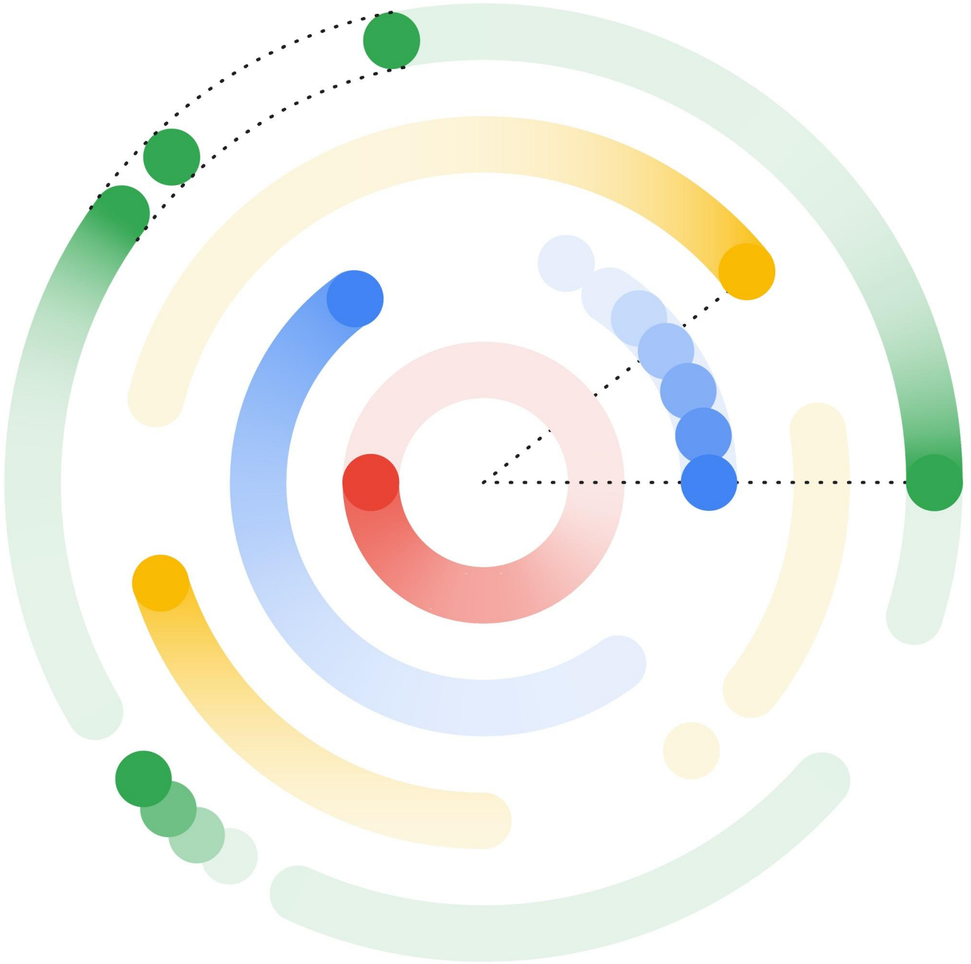
Google Cloud

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

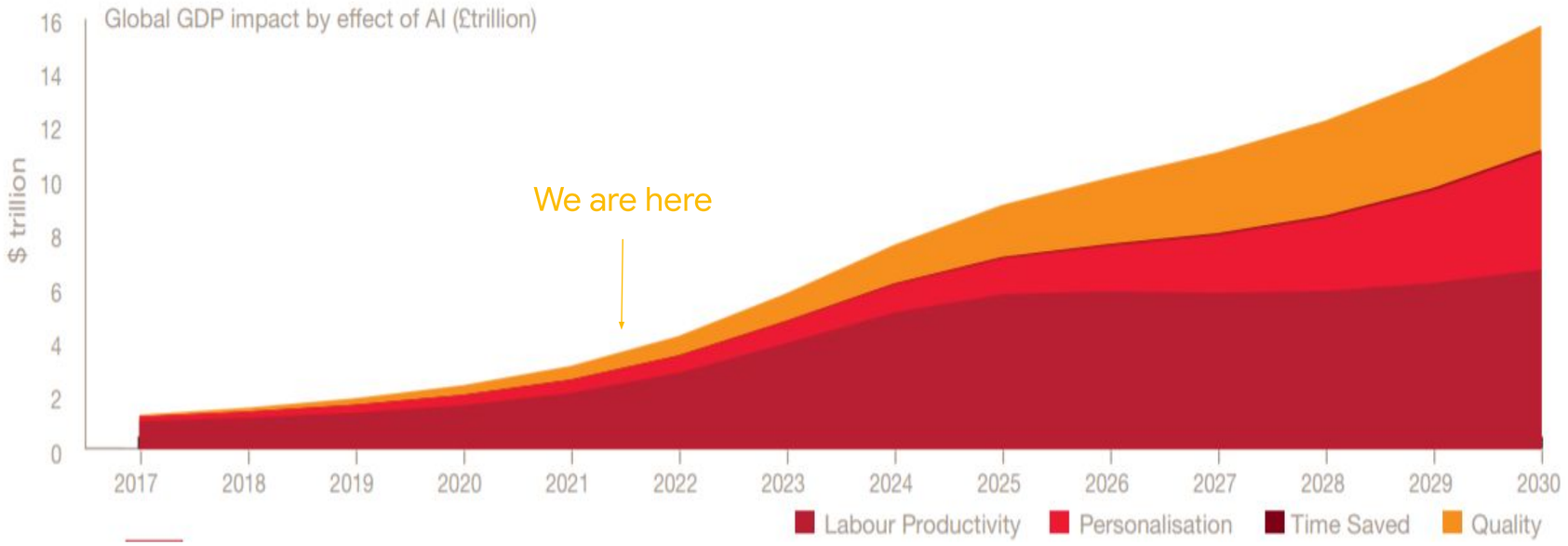
Agenda

01

The case for responsible AI

By 2030, AI will be the
#1 driver
of global GDP growth.

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

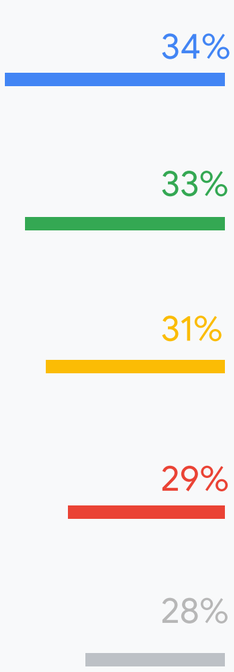


Organizations that achieve AI absorption will be the leaders of the global economy

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

Adopting Responsible AI practices: the research

What fuels ethical issues



34% Pressure to urgently implement AI without adequately addressing ethical issues

33% Ethical issues were not considered while constructing AI systems

31% Lack of resources (funds, people, technology) dedicated to ethical AI systems

29% Lack of a diverse team with respect to race, gender, etc. developing AI systems

28% Lack of ethical AI code of conduct or the ability to assess deviation from it

Ethics at the center

90% Organizations (in one survey) globally have encountered ethical issues resulting from the use of AI

40% AI systems (in one survey) were abandoned when ethical issues are raised

1.7 x Companies who have achieved scale in AI are more likely to understand & implement Responsible AI

Source: Capgemini “Organizations must address ethics in AI to gain public's trust and loyalty Accenture-Built-to-Scale-PDF-Report

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)

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?

AI Principles & Practices

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?

Governance Programs & Reviews

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?

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:

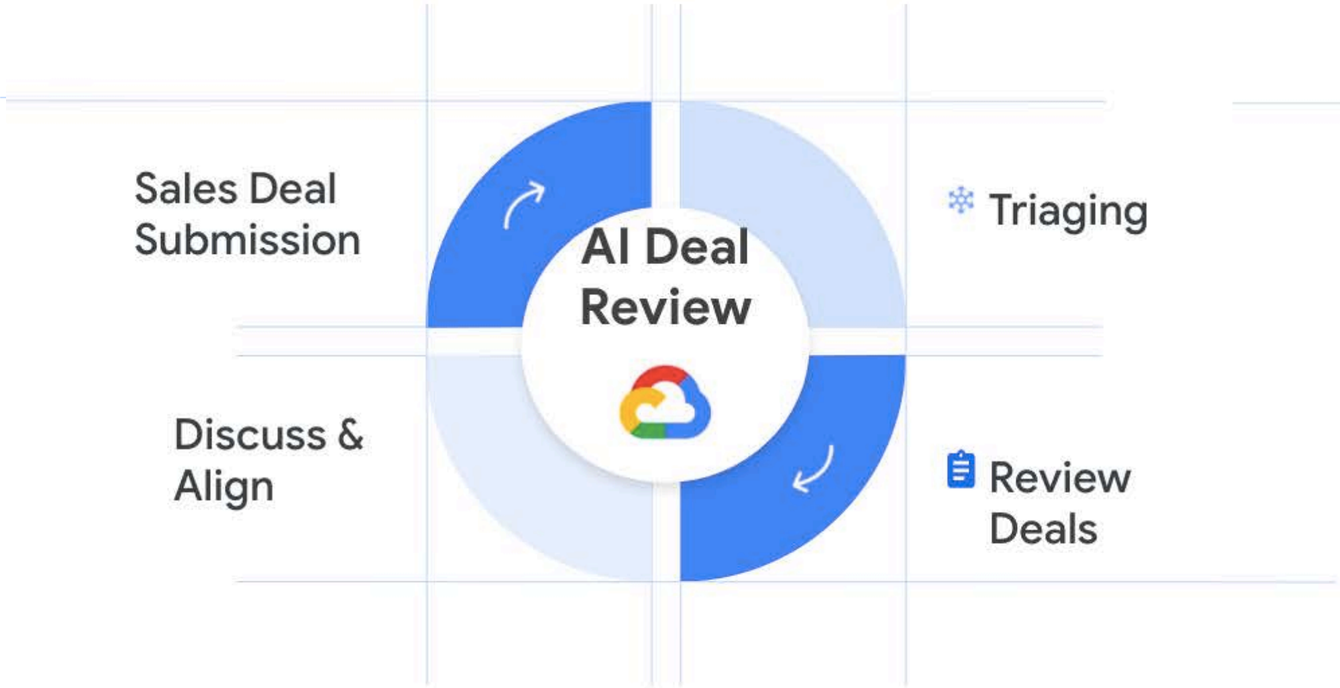
1. likely to cause overall harm
2. principal purpose to direct injury
3. surveillance violating internationally accepted norms
4. purpose contravenes international law and human rights

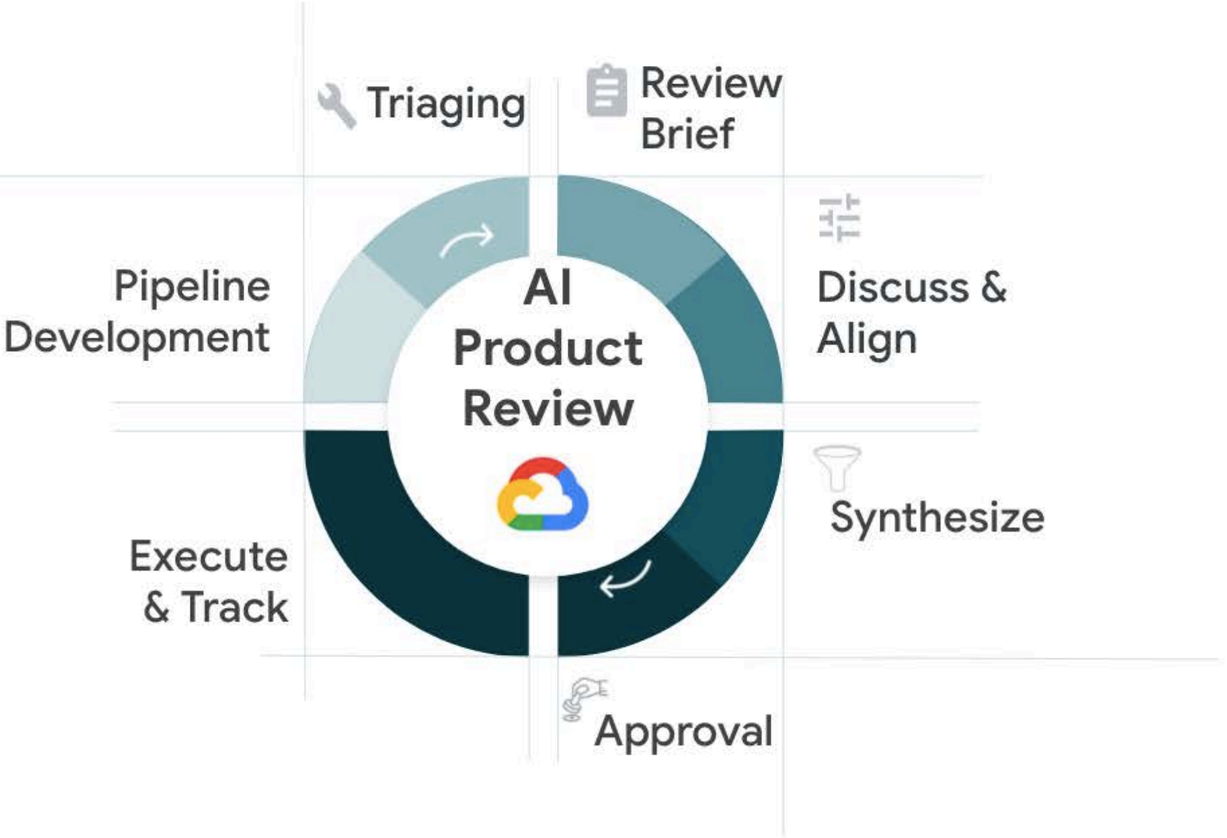
Principles that remain on paper

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

1. **Articulate Your Values & Principles***
   What are your shared commitments, guidelines, guardrails & lines you will not cross?
2. **Operationalize Your Commitments**
   What processes will you put in place to ensure your actions align with your values.
3. **Evaluate Every Use Case***
   Who will this benefit? Who is impacted? Who might it harm or benefit?
   What is the societal context in which this will live?
4. **Define Fairness***
   What does fair mean for this specific use case?
   How will you assess & measure it?
5. **Review & Discuss**
   Create a safe space for complex & sensitive conversations.
   Ensure a full range of voices can input.
6. **Craft Your Alignment Plan**
   How specifically will your product, system, tool, etc. align with your values?
7. **NOW Build, Integrate, Monitor***
   This is when tooling comes into the picture.

* 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

- 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 and GCP resources and tools for ML developers

- **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

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

Fairness

Send feedback

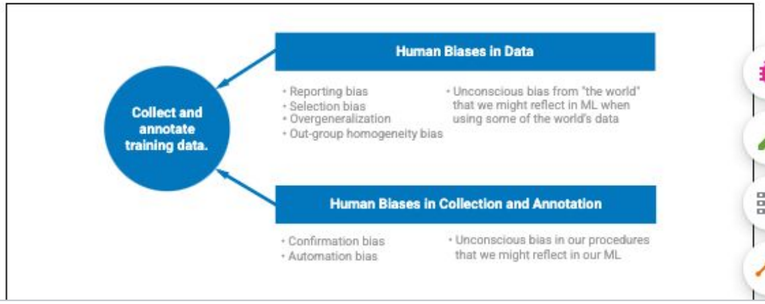
Estimated Time: 5 minutes

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.

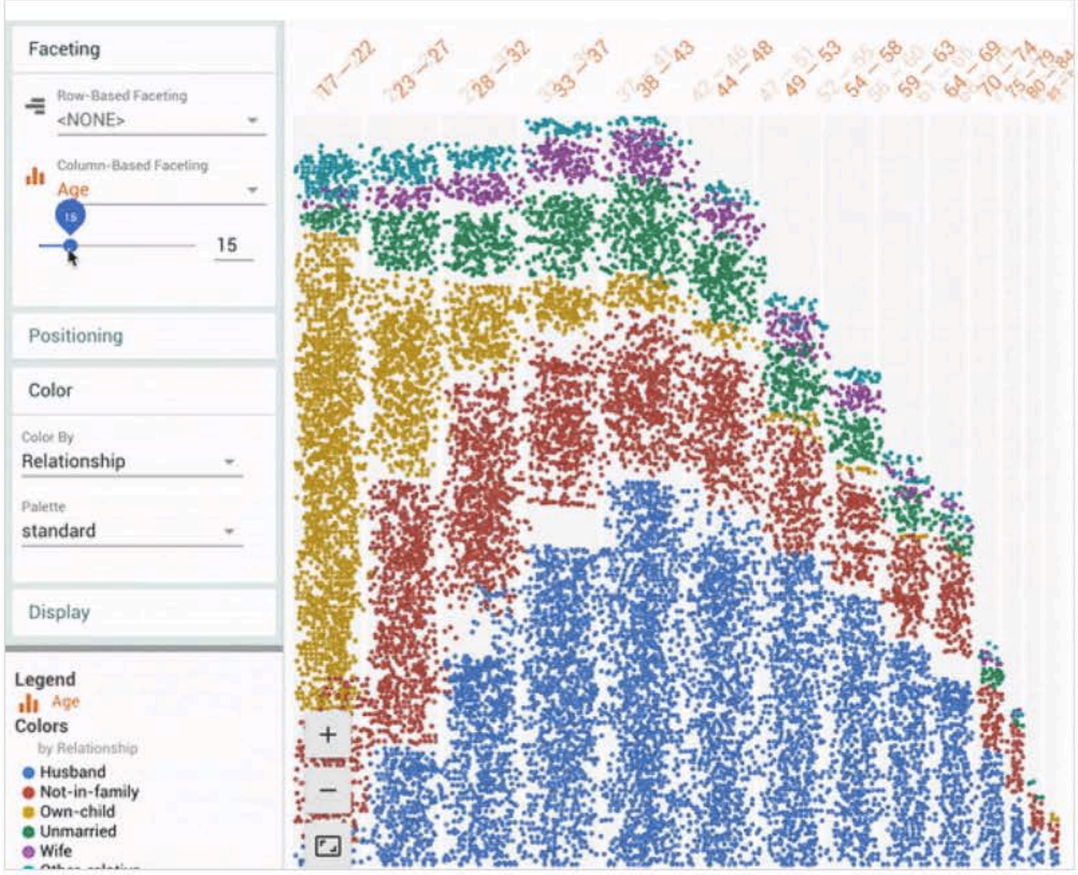
This module looks at different types of human biases that can manifest in training data. It then provides strategies to identify them and evaluate their effects.

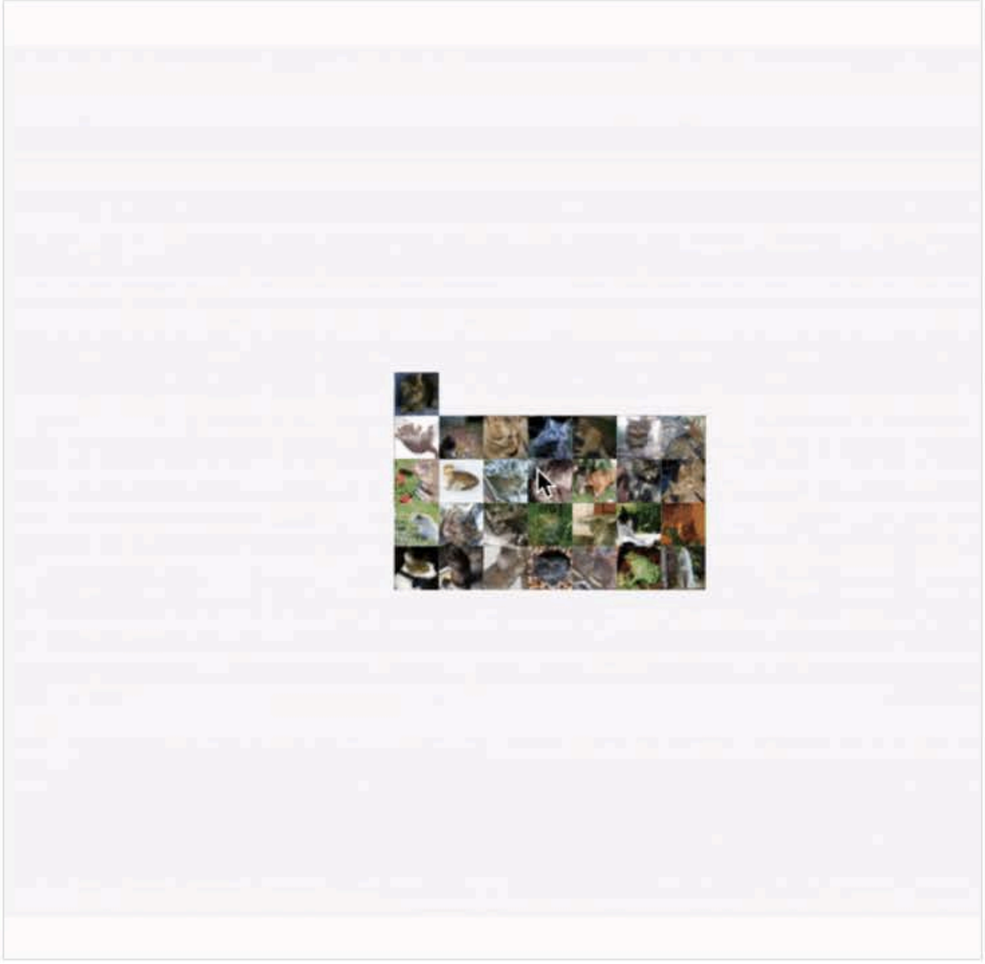


Construct and prepare your data

Facets

An open-source tool for visualizing and inspecting your datasets



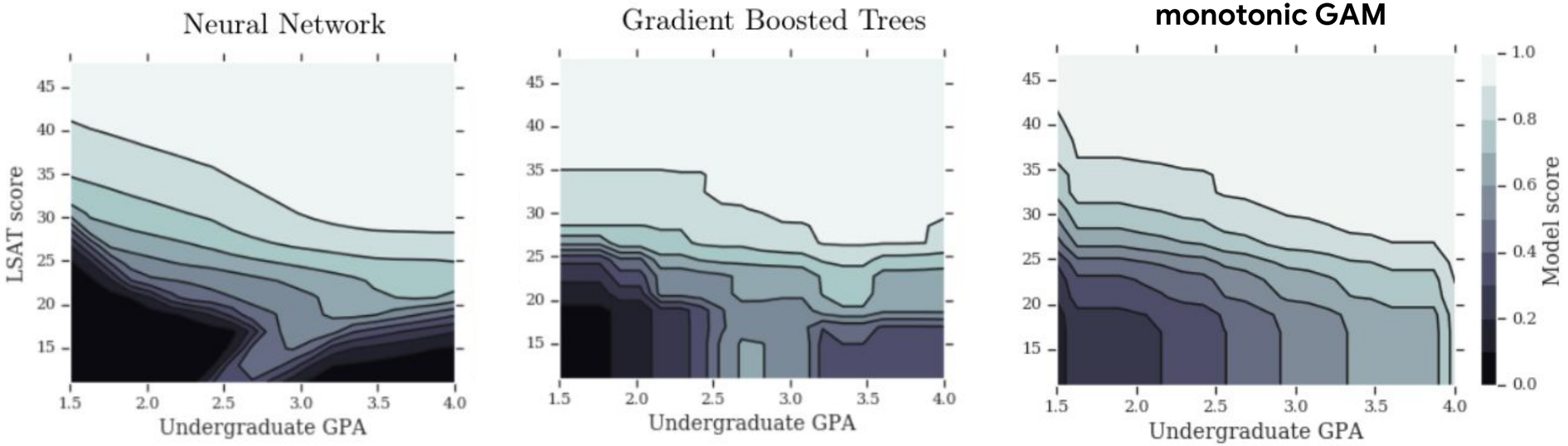


pair-code.github.io/facets/

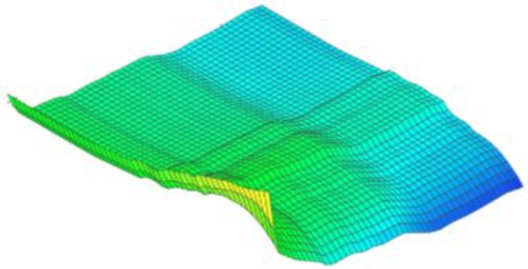
Build and Train Model

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 $\{x_1, x_2, x_3, ...\}$
- Ensembles of 1-D piecewise-linear functions (GAMs) or multi-D functions (lattices)



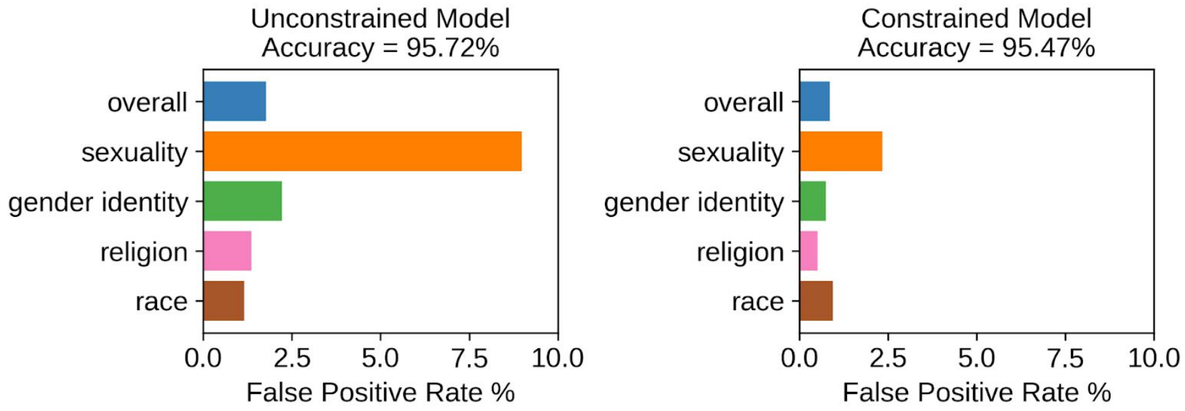
Partially monotonic lattice



tensorflow.org/lattice/overview

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 ≤ 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**

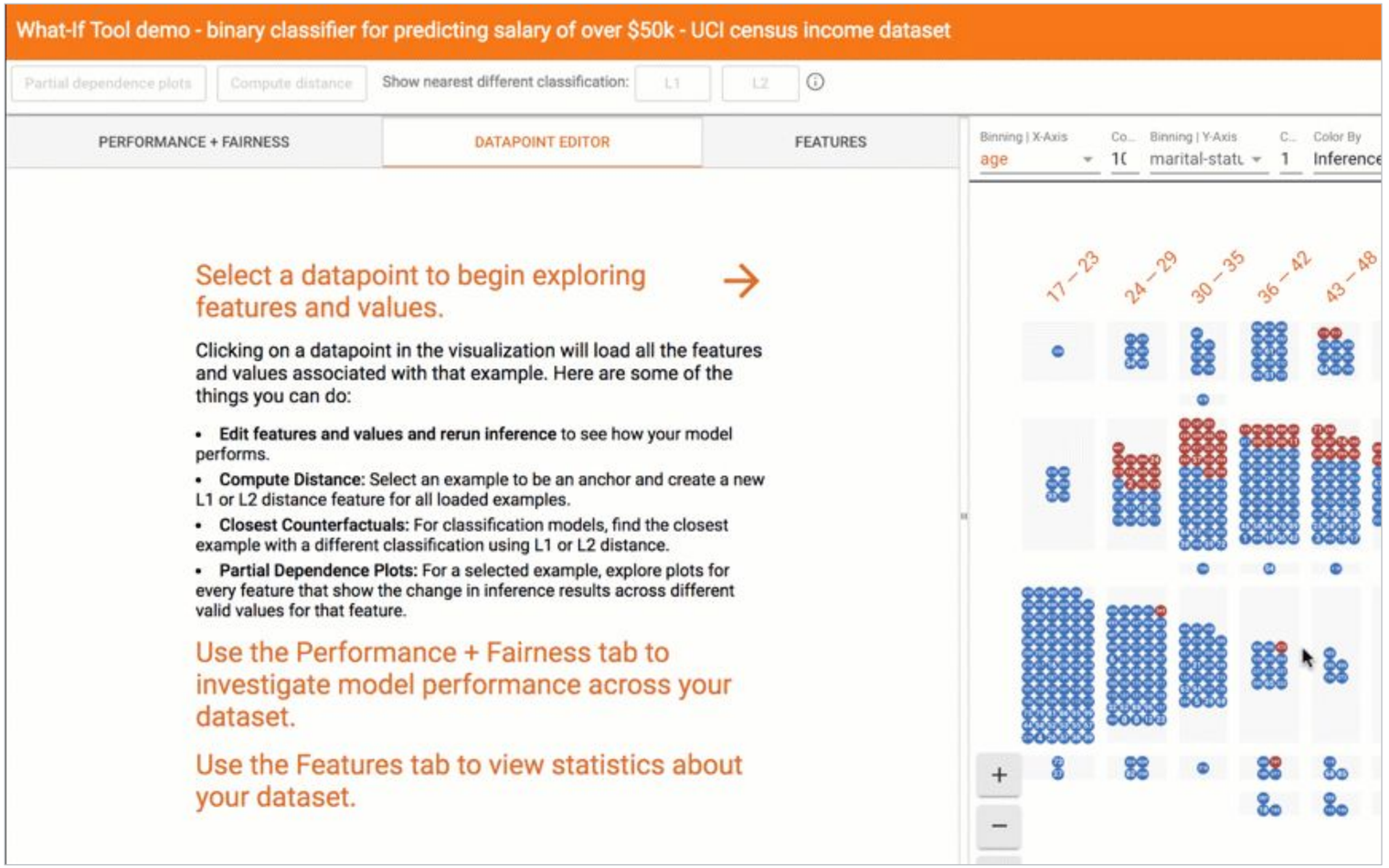


github.com/google-research/tensorflow_constrained_optimization/

Evaluate your model

What-If Tool

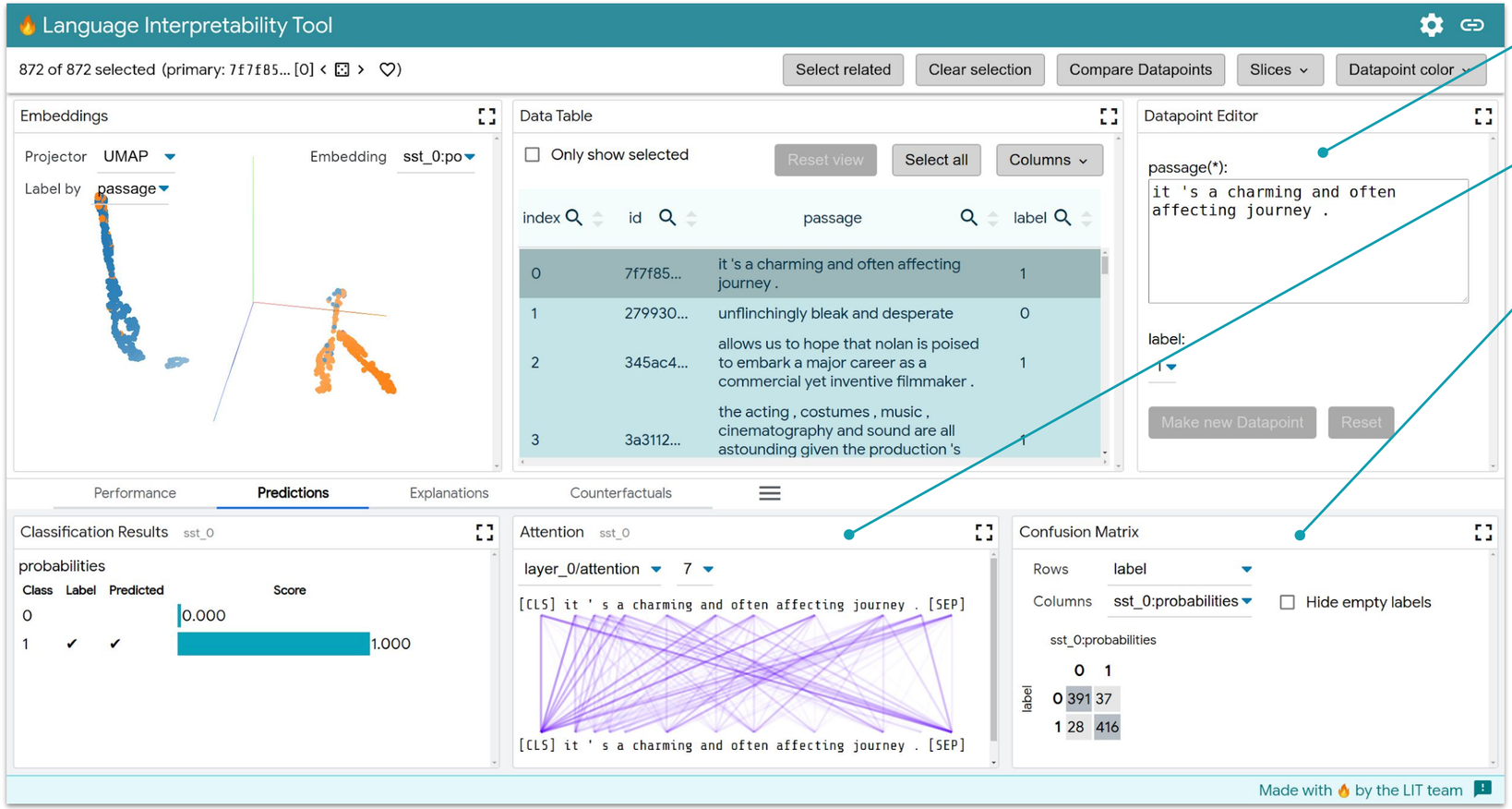
Visually probe the behavior of **tabular ML models**, with minimal coding



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

Language Interpretability Tool

A visual, interactive model-understanding tool for NLP

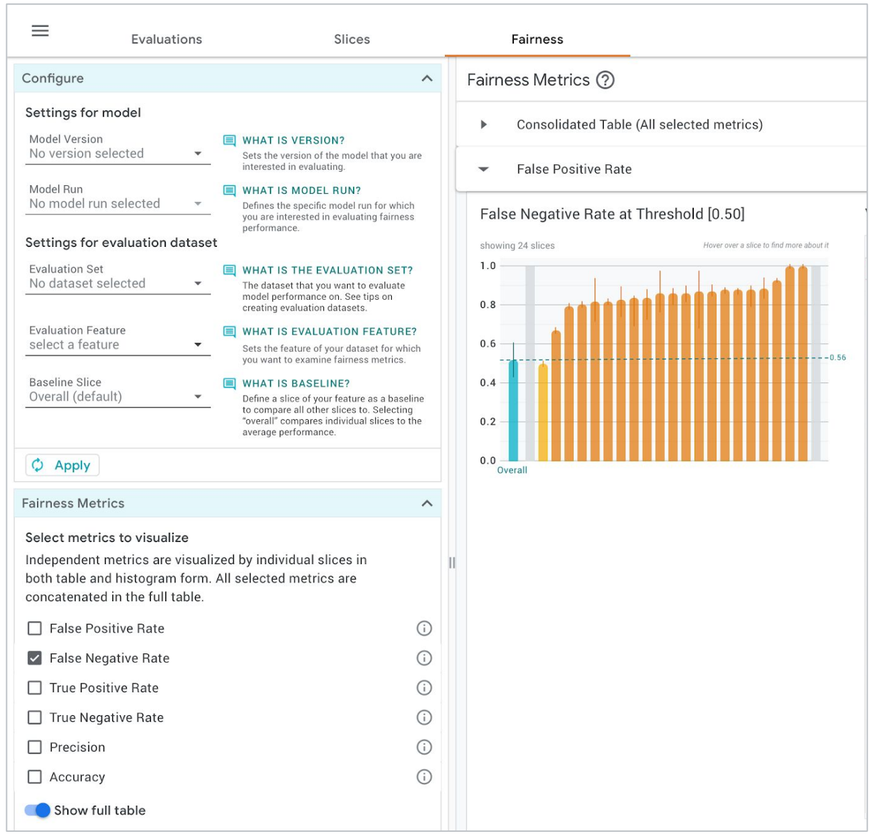


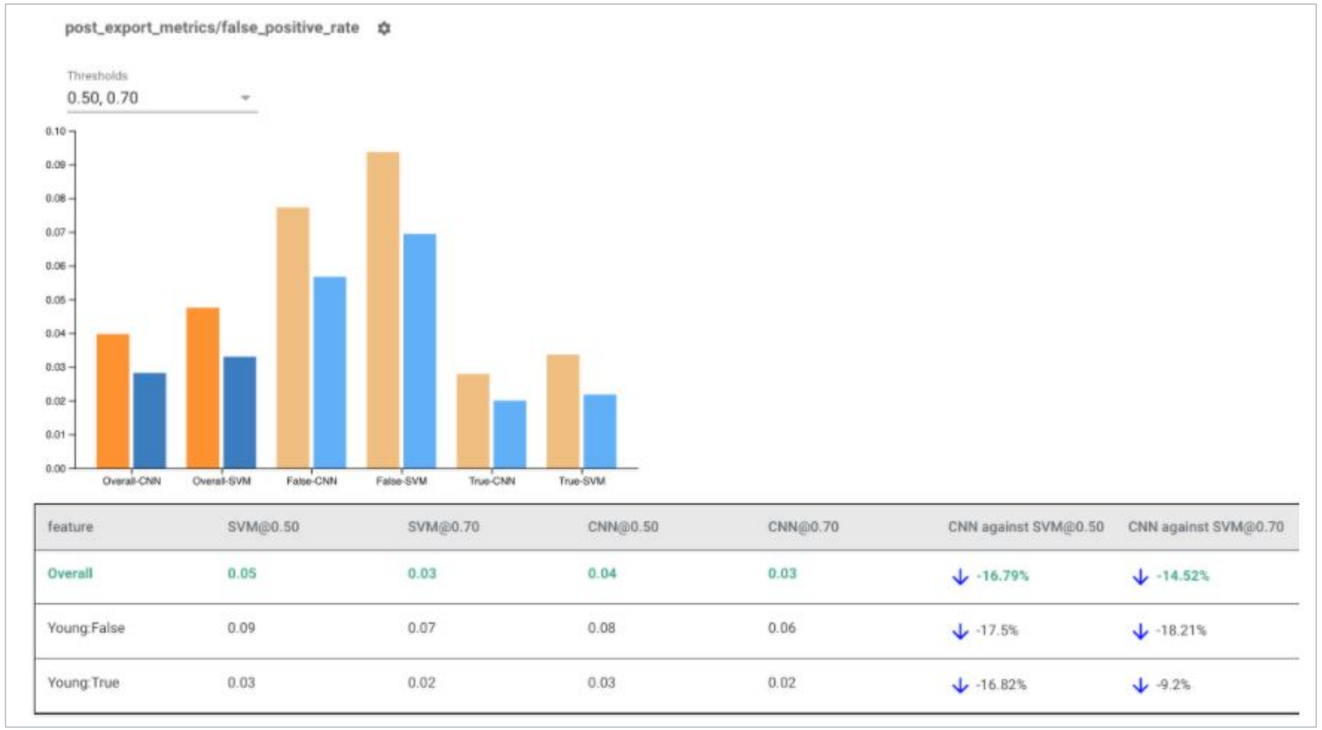
- **Counterfactual generation** via manual edits or algorithmic generator plug-ins
- **Local explanations** via salience maps, attention, and rich visualizations
- **Aggregate analysis** with custom metrics, slice-by-feature, and embedding clusters
- **Side-by-side mode** to compare two models, or a 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

pair-code.github.io/lit/

ML Fairness Indicators

Compute & visualize ML fairness metrics in TFX pipelines





feature	SVM@0.50	SVM@0.70	CNN@0.50	CNN@0.70	CNN against SVM@0.50	CNN against SVM@0.70
Overall	0.05	0.03	0.04	0.03	-16.79%	-14.52%
Young:False	0.09	0.07	0.08	0.06	-17.5%	-18.21%
Young:True	0.03	0.02	0.03	0.02	-16.82%	-9.2%

tensorflow.org/tfx/guide/fairness_indicators

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.

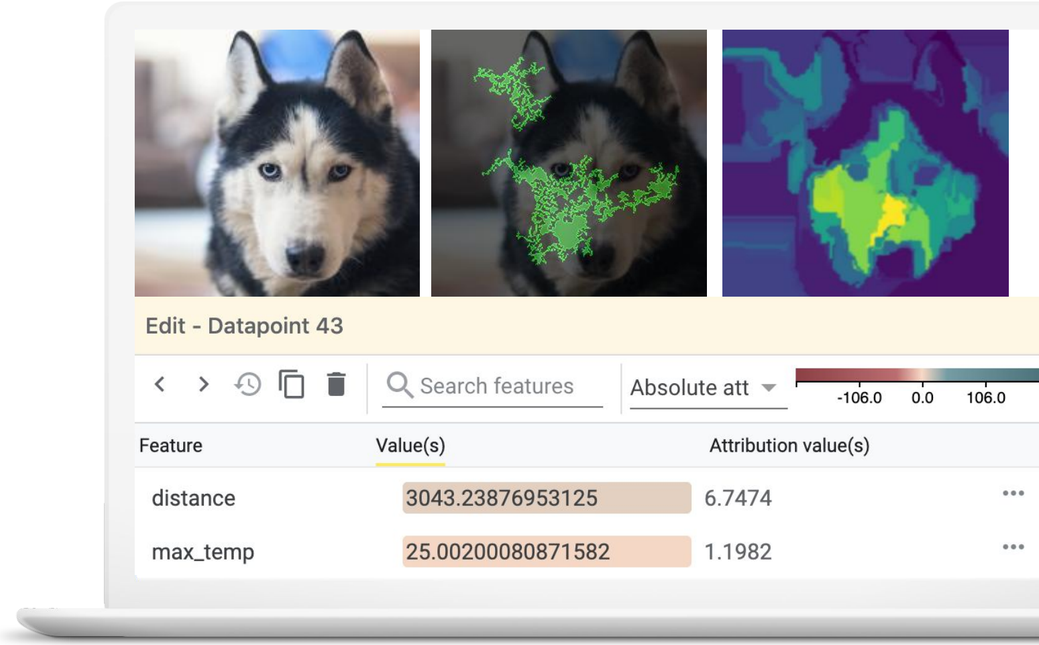
Built into multiple AI Platform services

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 significant faster than open-source alternatives.



cloud.google.com/explainable-ai

Integrate and Monitor

Vertex Model Monitoring

Easy and proactive monitoring of model performance

Monitor and alert

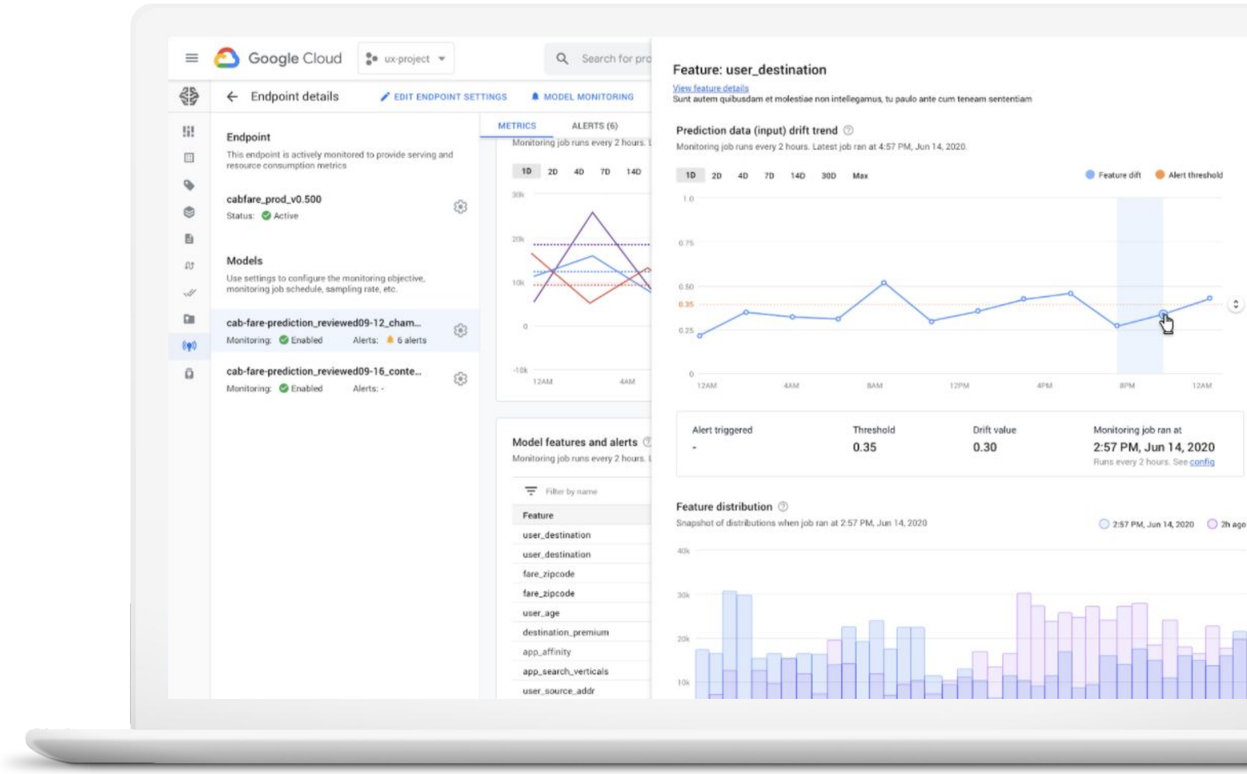
Monitor signals for model's predictive performance, and alert when those signals deviate.

Diagnose

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

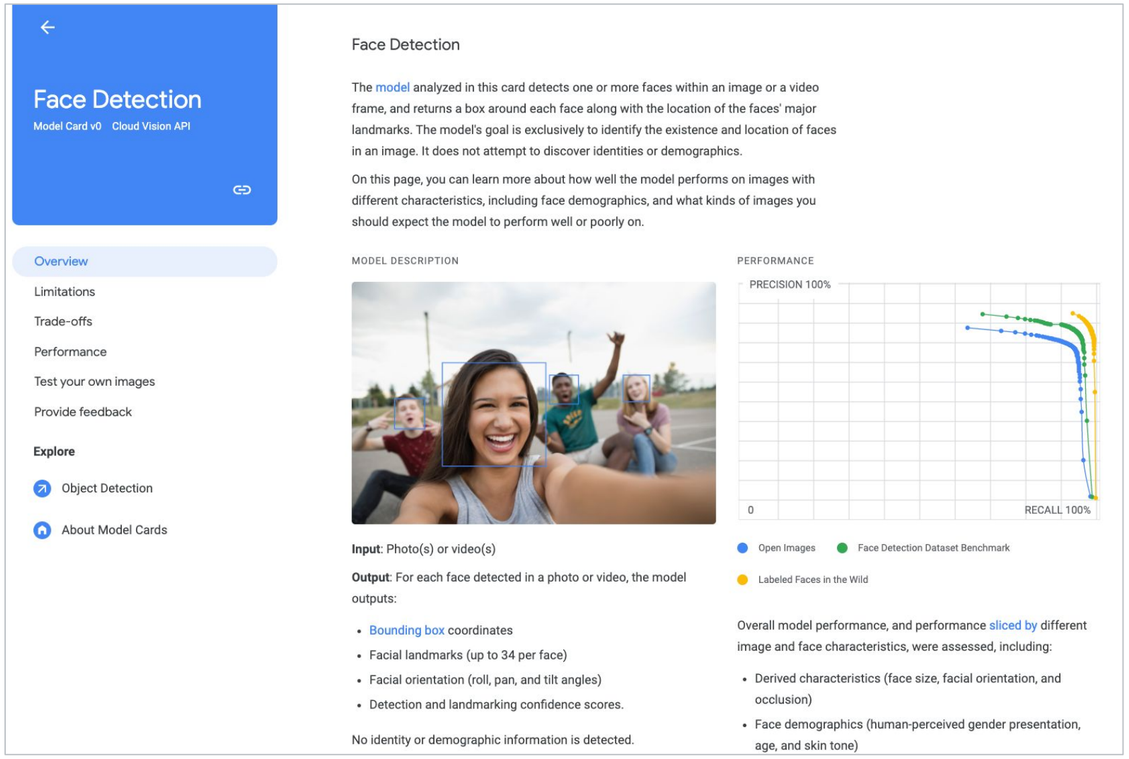
Update Model

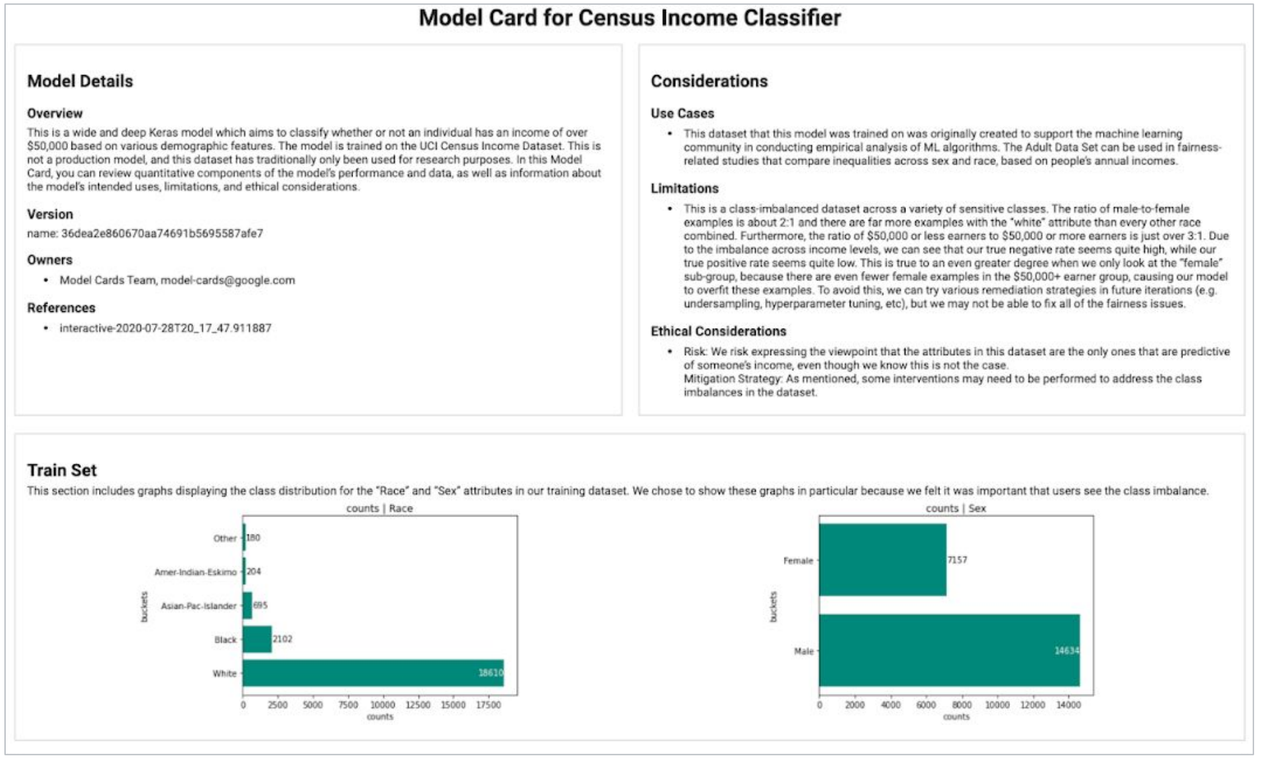
Trigger model re-training pipeline or collect relevant training data to address performance degradation.



Model Cards & Model Card Toolkit

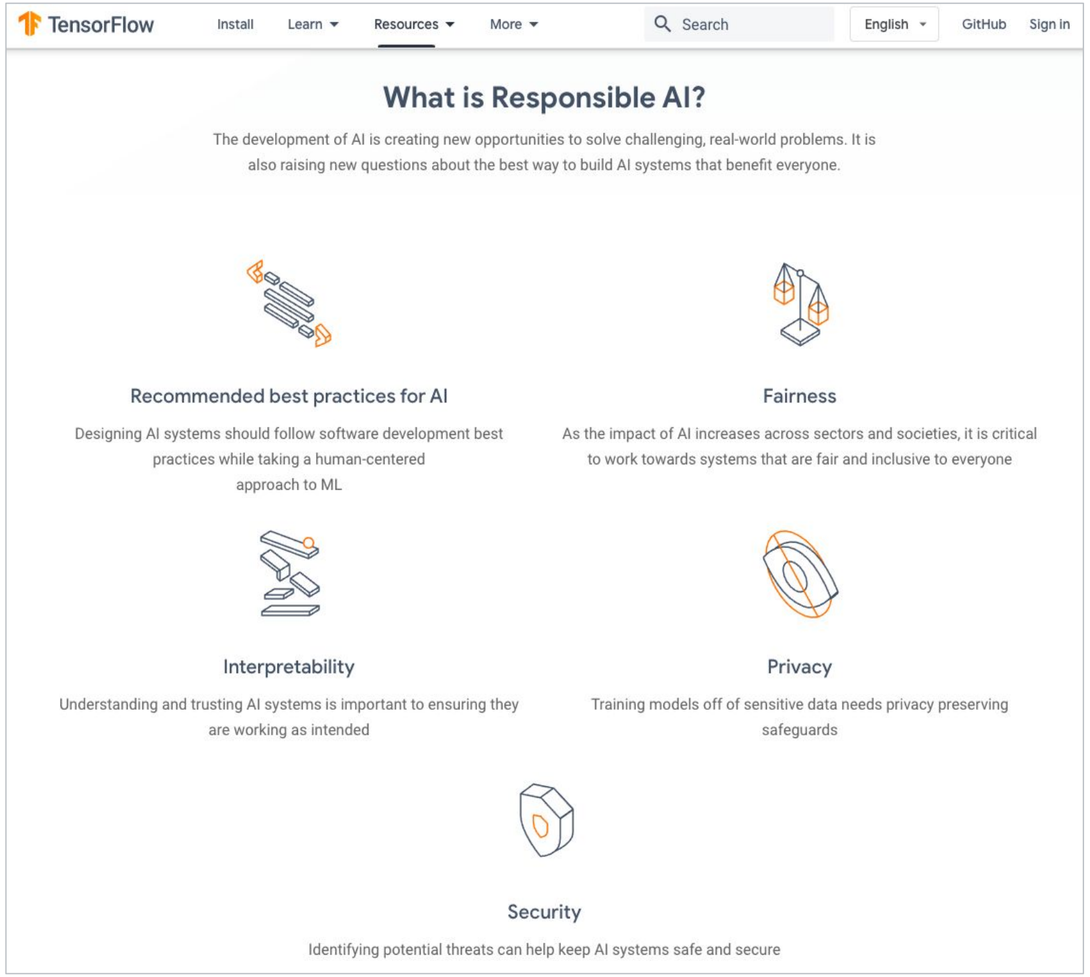
Organize and communicate the essential facts of your models in a structured way



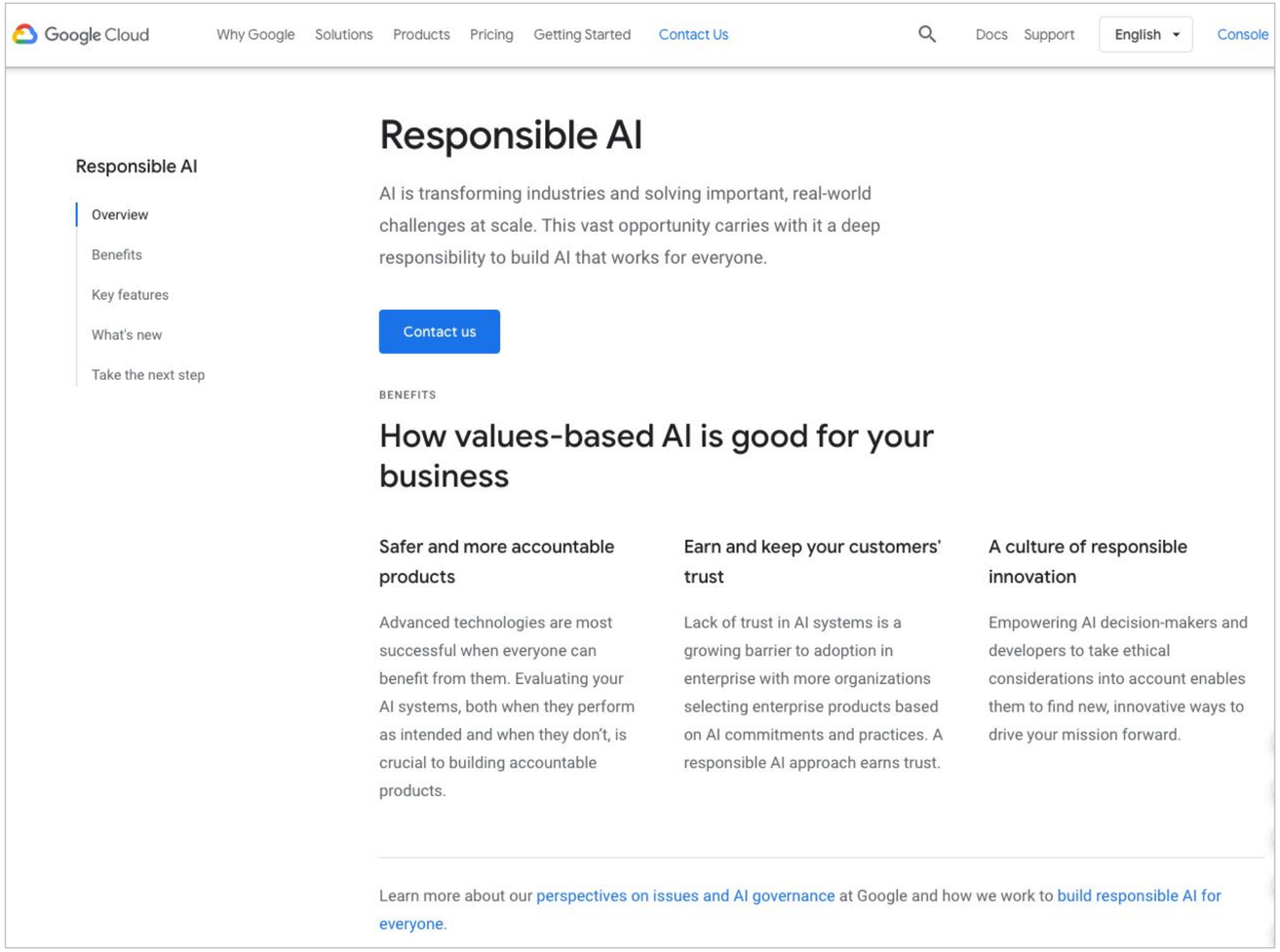


github.com/tensorflow/model-card-toolkit

Learn more about these and other tools at...



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**

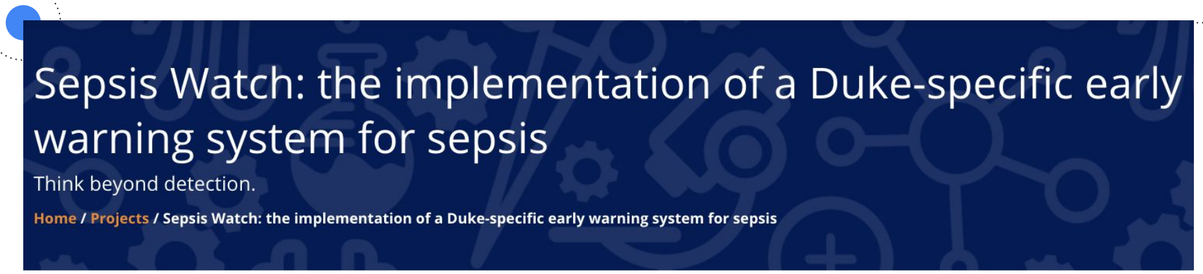
02 **Recognized by WHO as a major cause of preventable disease and death globally**

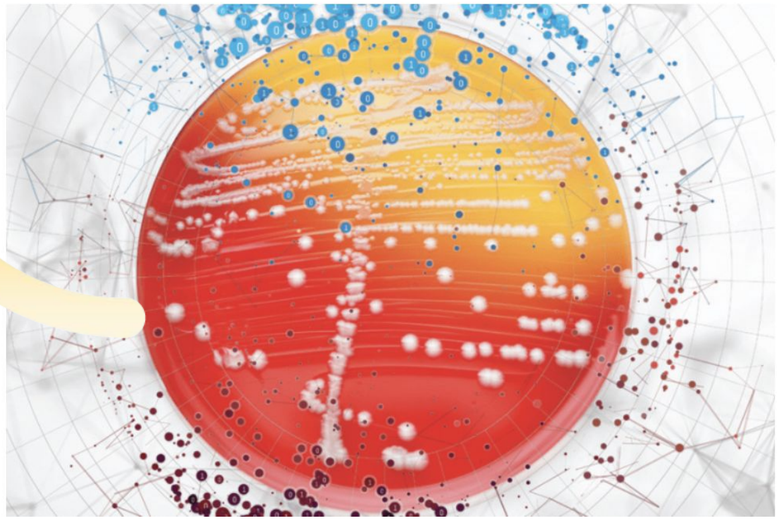
03 **Difficult to diagnose; no universally agreed upon definition**

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.





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”

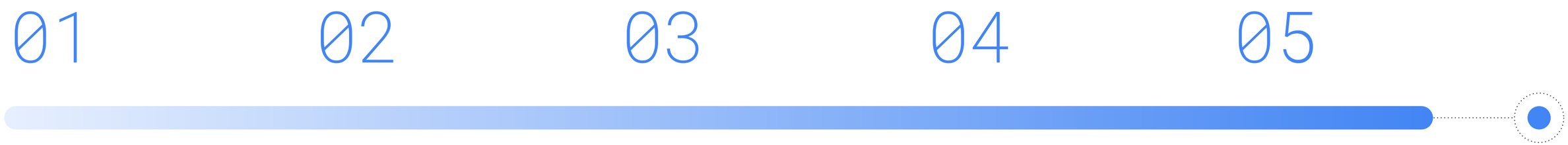
IEEE Spectrum

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

Source: Accenture-Built-to-Scale-PDF-Report

~~Deploy~~ Integrate

The responsible AI workflow



01	02	03	04	05
Define Problem	Construct and Prepare Data	Build and Train Model	Evaluate	Integrate and Monitor

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

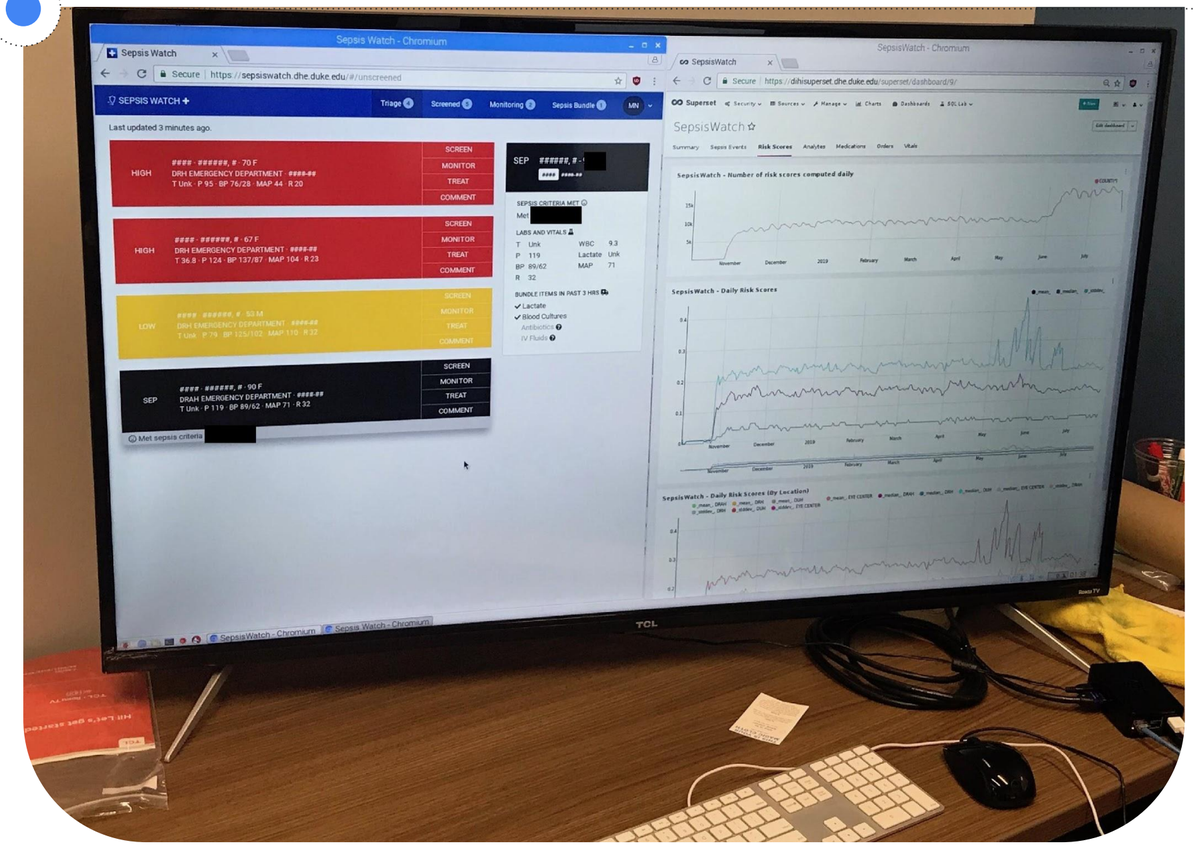
Clinicians helped lead institution-specific validation and evaluation.

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

- **Outcome** .....sepsis within the next 4 hours, see (1) for sepsis criteria
- **Output** .....0% - 100% probability of sepsis occurring in the next 4 hours
- **Patient population** .....all adult patients >18 y.o. presenting to DUH ED and admitted
- **Time of prediction** .....every hour of a patient's encounter
- **Input data source**.....electronic health record (EHR)
- **Input data type** .....demographics, analytes, vitals, medication administrations
- **Training data location and time-period** .....DUH, 10/2014 – 12/2015
- **Model type**..... Recurrent Neural Network

Validation and performance

	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

- **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

Google Cloud

Thank you.