Microsoft

Responsible AI: Building trustworthy, secure and transparent machine learning

Mehrnoosh Sameki
Senior Program Manager, Azure AI
mesameki@microsoft.com

Talk Outline

Introduction to Responsible AI

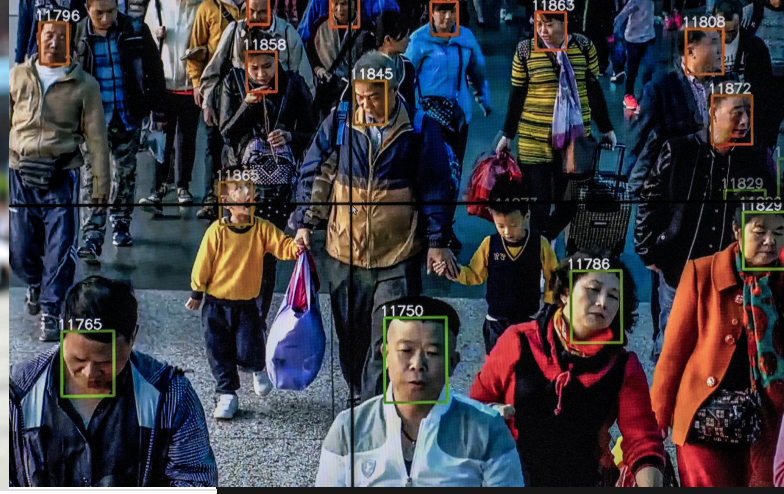
Responsible AI at Microsoft

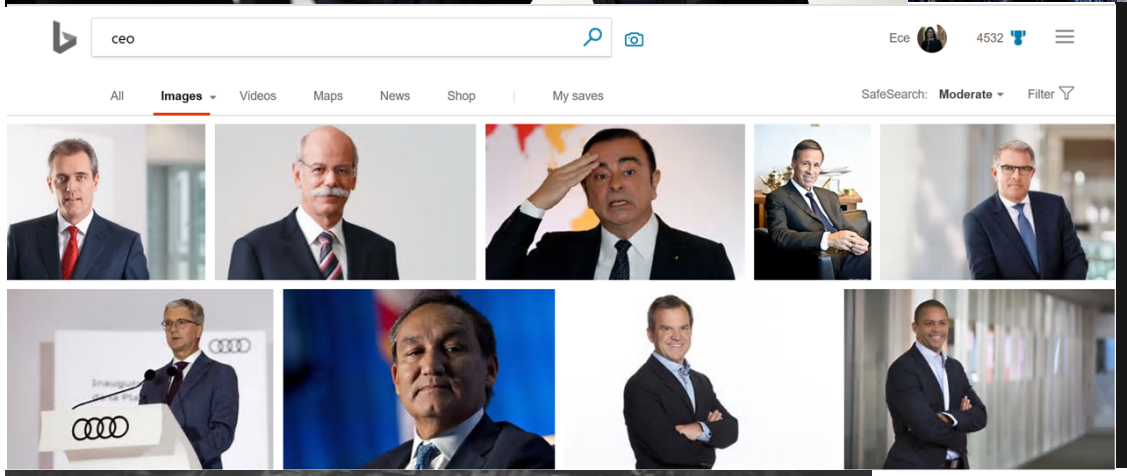
Tools to aid you

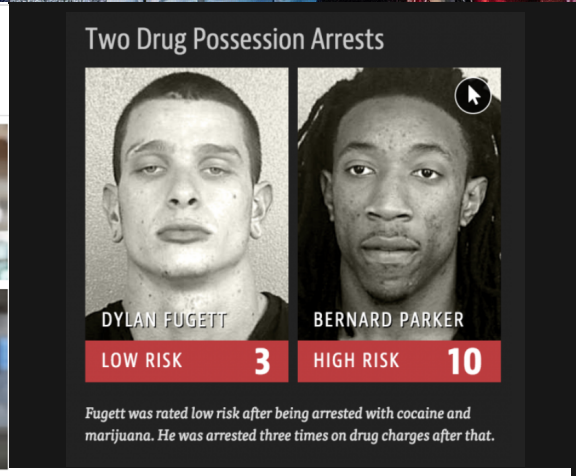
Future Directions

AI will have a considerable impact on business and society as a whole

But this impact raises a host of complex and challenging questions

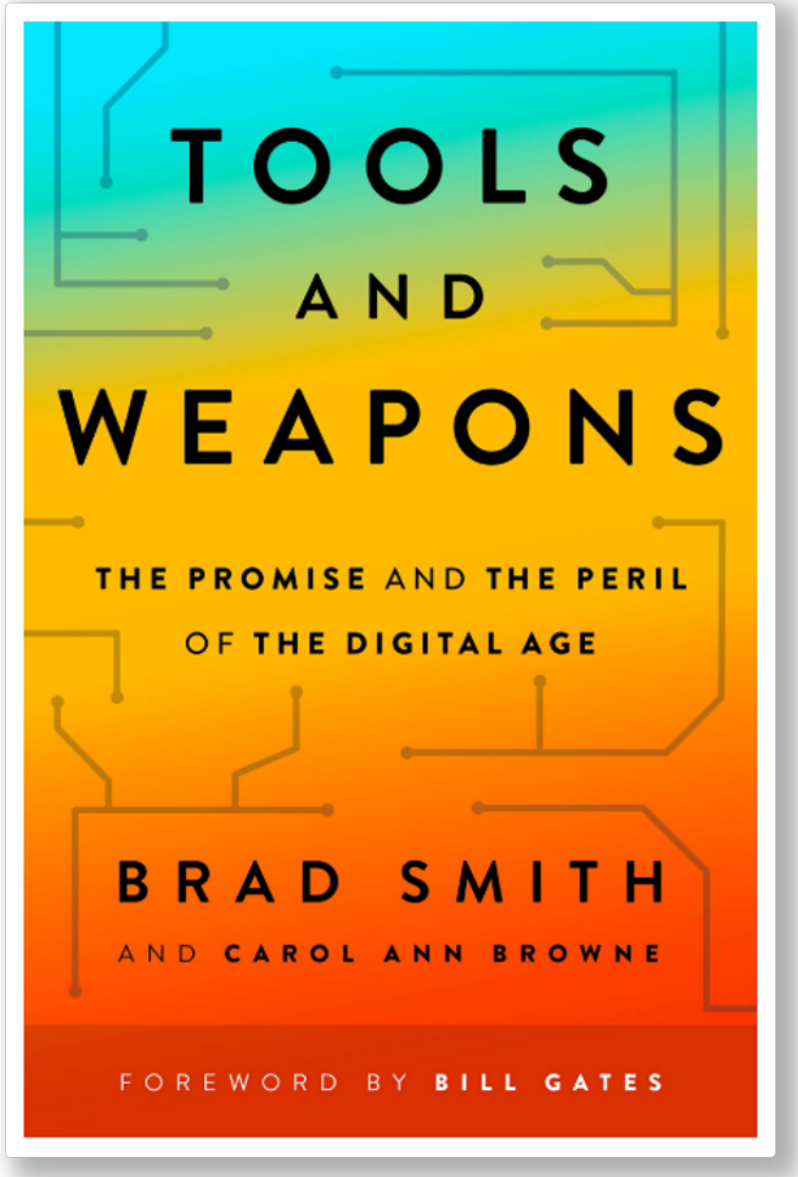






Original Video for Input Audio

“When your technology changes the world, you bear a responsibility to help address the world you have helped create.”



Why Responsible AI?

7.5.19 Report from Capgemini

Consider People and the Real-World Implications

Bring in real-world context
domain expertise,
and diversity

Focus on envisioning uses and impact at the beginning

Building Trusted Machine Learning

Platform

Build on a platform that is secure and trustworthy

Process

Focus on reliable, reproducible processes that incorporate human oversight and the state-of-art best practices

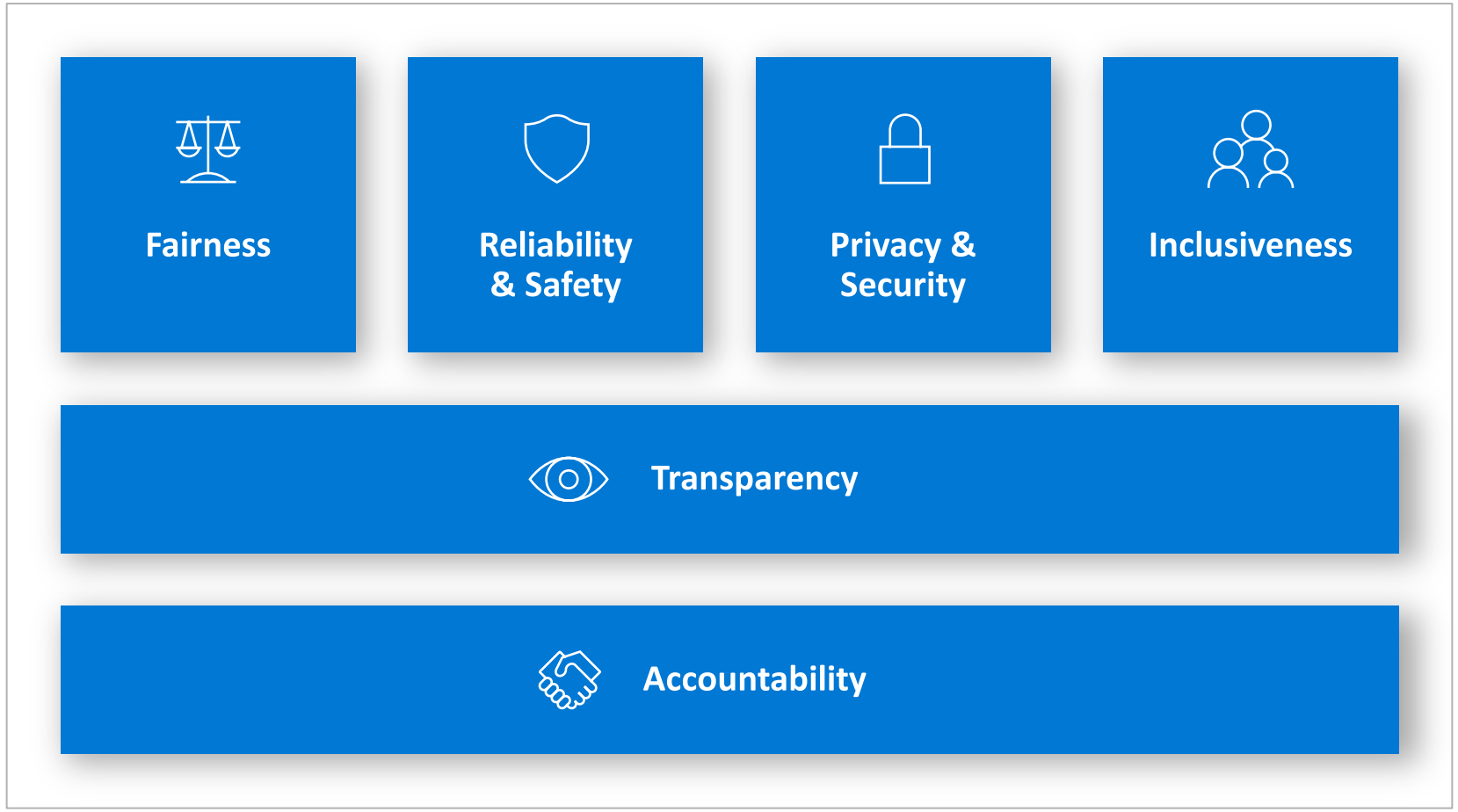
Models

Emphasize deeper understanding of model behavior, analysis, and testing

“Ultimately the question is not only what computers can do. It’s what computers should do.”

—The Future Computed

Microsoft’s AI Principles



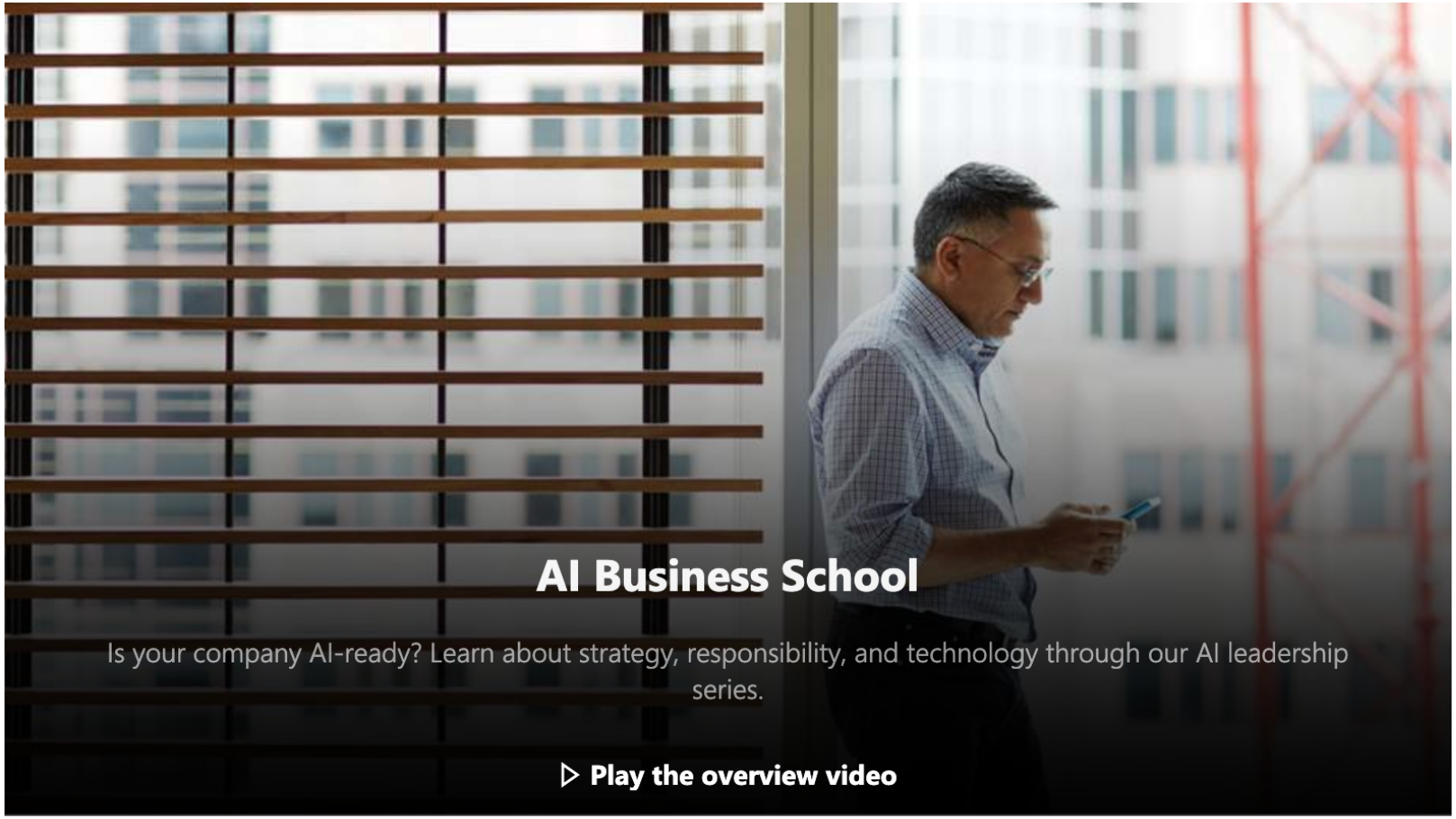
Putting our ethical principles into action

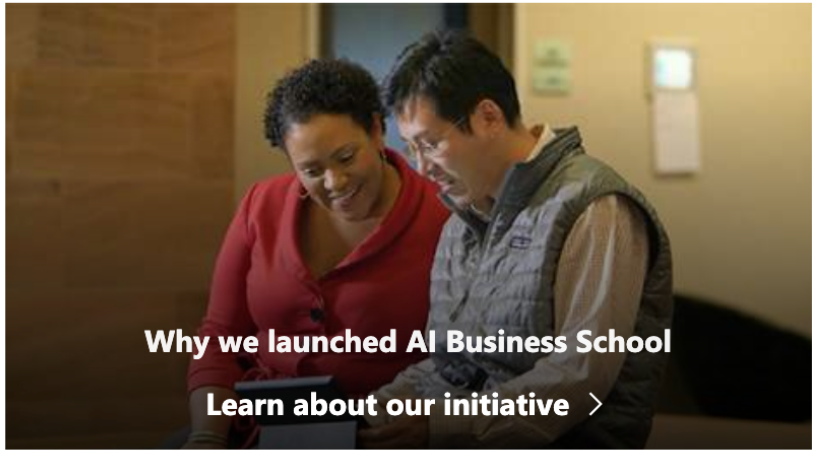
Aether Working Groups

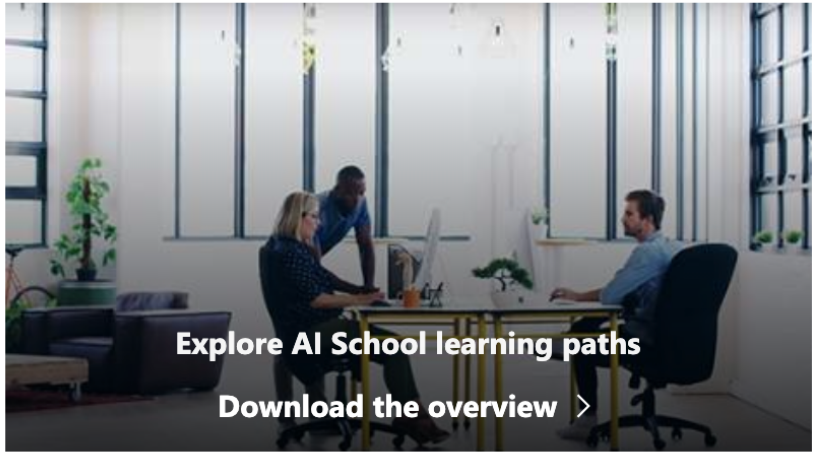
AI, Ethics, and Effects in Engineering and Research

"By working arm-in-arm with multiple stakeholders, we can address the important topics rising at the intersection of AI, people, and society."

ERIC HORVITZ, TECHNICAL FELLOW AND DIRECTOR OF MICROSOFT RESEARCH, MICROSOFT



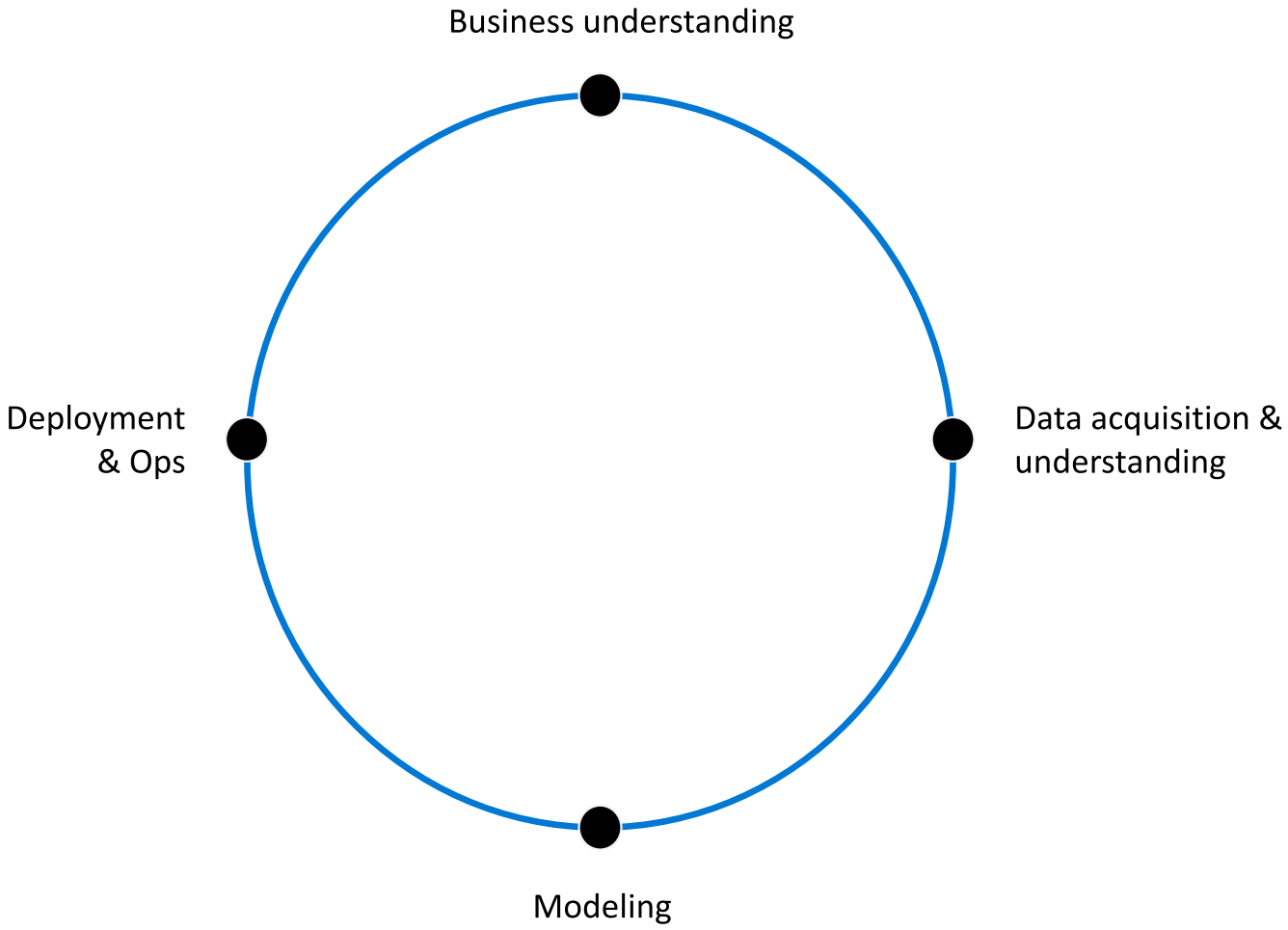




Lead with confidence in the age of AI

AI is changing how business works across all industries. We created AI Business School to share insights and practical guidance from top executives on how to strategically apply AI in your organization.

Responsible AI is an End-to-End Process



Types of harm

- Allocation: extends or withholds opportunities, resources, or information.
- Quality of service: whether a system works as well for one person as it does for another
- Stereotyping: reinforce existing societal stereotypes
- Denigration: actively derogatory or offensive
- Over or under representation: over-represent, under-represent, or even erase particular groups of people

Crawford et al. 2017

Data: Source

- Think critically before collecting any data
- Try to identify societal biases present in data source
- Check for biases in cultural context of data
- Check the data source matches deployment context

Data: Collection Process

- Check for biases in technology used to collect data
- Check for biases in humans involved in collecting data
- Check for biases in strategy used for sampling
- Ensure sufficient representation of subpopulations
- Check the collection process itself is fair and ethical

Data: Labeling and preprocessing

- Check whether discarding data introduces biases
- Check whether bucketing introduces bias
- Check preprocessing software for bias
- Check labeling/annotation software for biases
- Check that human labelers do not introduce biases

To Do:

- Better data-related documentation
  - Datasheets for datasets: every dataset, model, or pre-trained API should be accompanied by a data sheet that documents its
    - Creation
    - Intended uses
    - Limitations
    - Maintenance
    - Legal and ethical considerations
    - Etc.

https://arxiv.org/abs/1803.09010

Model Definition

- Clearly define all assumptions about model
- Try to avoid biases present in assumptions
  - Using number of arrests as a proxy for amount of crime
- Check whether model structure introduces biases
- Check objective function for unintended effects
- Consider including “fairness” in objective function

Testing

- Check that test data matches deployment context
- Ensure test data has sufficient representation
- Continue to involve diverse stakeholders
- Revisit all fairness requirements
- Use metrics to check that requirements are met

Deployment

- Continually monitor
  - Match between training data, test data, and instances you encounter in deployment
  - Fairness metrics
  - User reports and user complaints
- Invite diverse stakeholders to audit systems for biases
- Monitor users' interactions with systems

Tools you can use

Overview of Transparency and Fairness Tools

Transparency

Unfairness Detection

InterpretML

Google What-if

IBM Fairness 360

Fairlearn

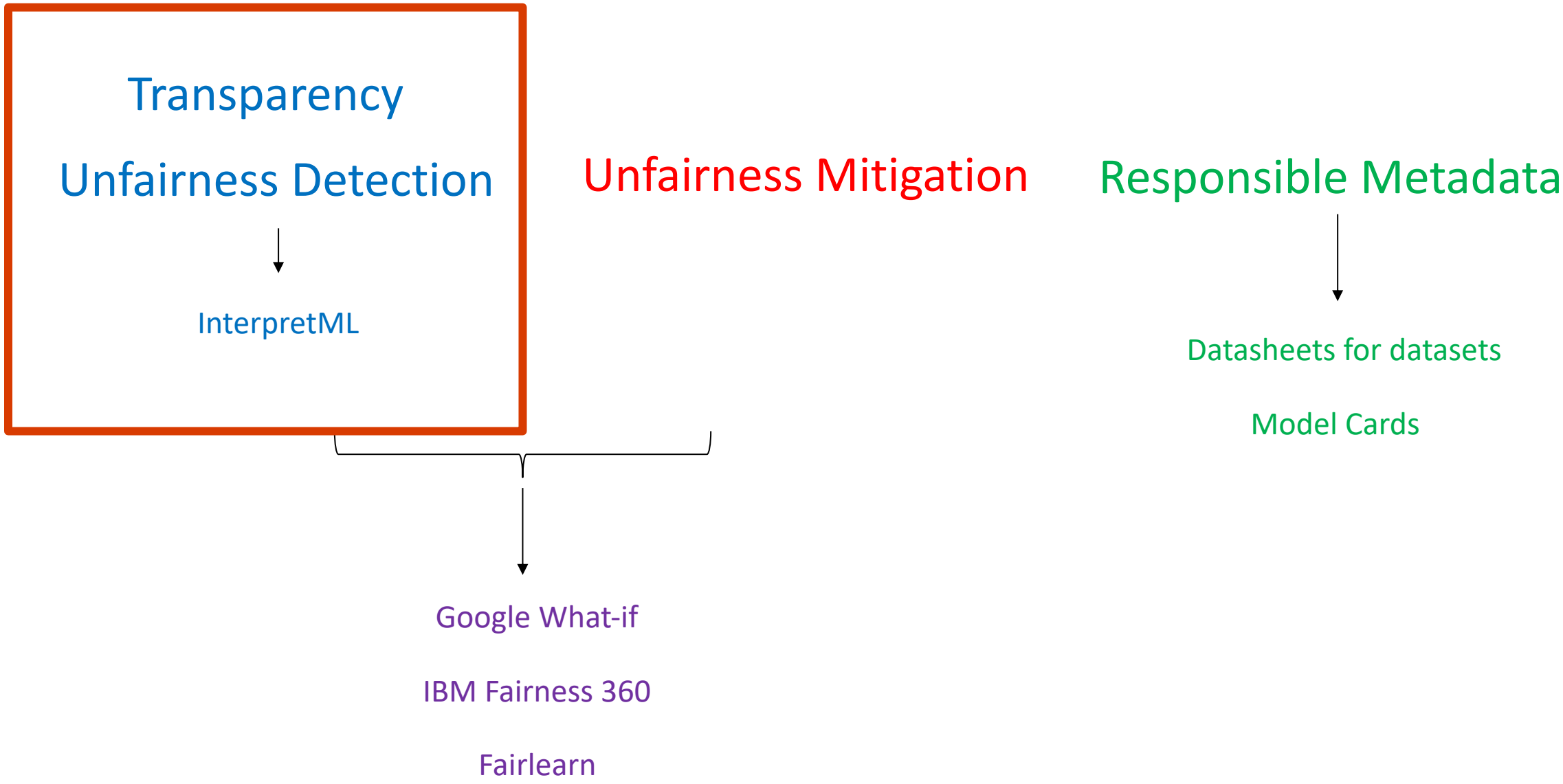
Unfairness Mitigation

Responsible Metadata

Datasheets for datasets

Model Cards

Overview of Transparency and Fairness Tools



Machine Learning Transparency and Fairness

Model Designers/Evaluators
Training Time

End users or providers of solutions to end users
Inferencing Time

- Data scientists need to **explain the output of a model to stakeholders** (business, users, clients) to build trust
- Data scientists need tools to **debug their models** and make informed decision on how to improve them
- Data scientists need tools to verify if model's behavior **matches pre-declared objectives**

- AI predictions need to be explained at the inferencing time:
  - e.g., **health care**: Why the model classified Fabio at risk for colon cancer?
  - e.g., **finance**: Why Rosine was denied a mortgage loan or why his investment portfolio carries a higher risk?

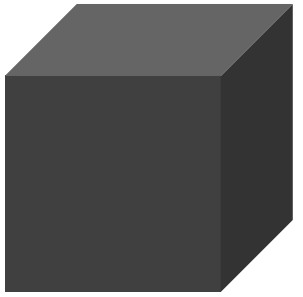
Interpretability: InterpretML

github.com/interpretML
pip install –U interpret

Tools to **understand how the system is working**

Glassbox Models

Explainable Boosting
Linear Models
Decision Tree
Rule Systems
…



Blackbox Explainers

LIME
SHAP
Partial Dependence
Sensitivity Analysis
…

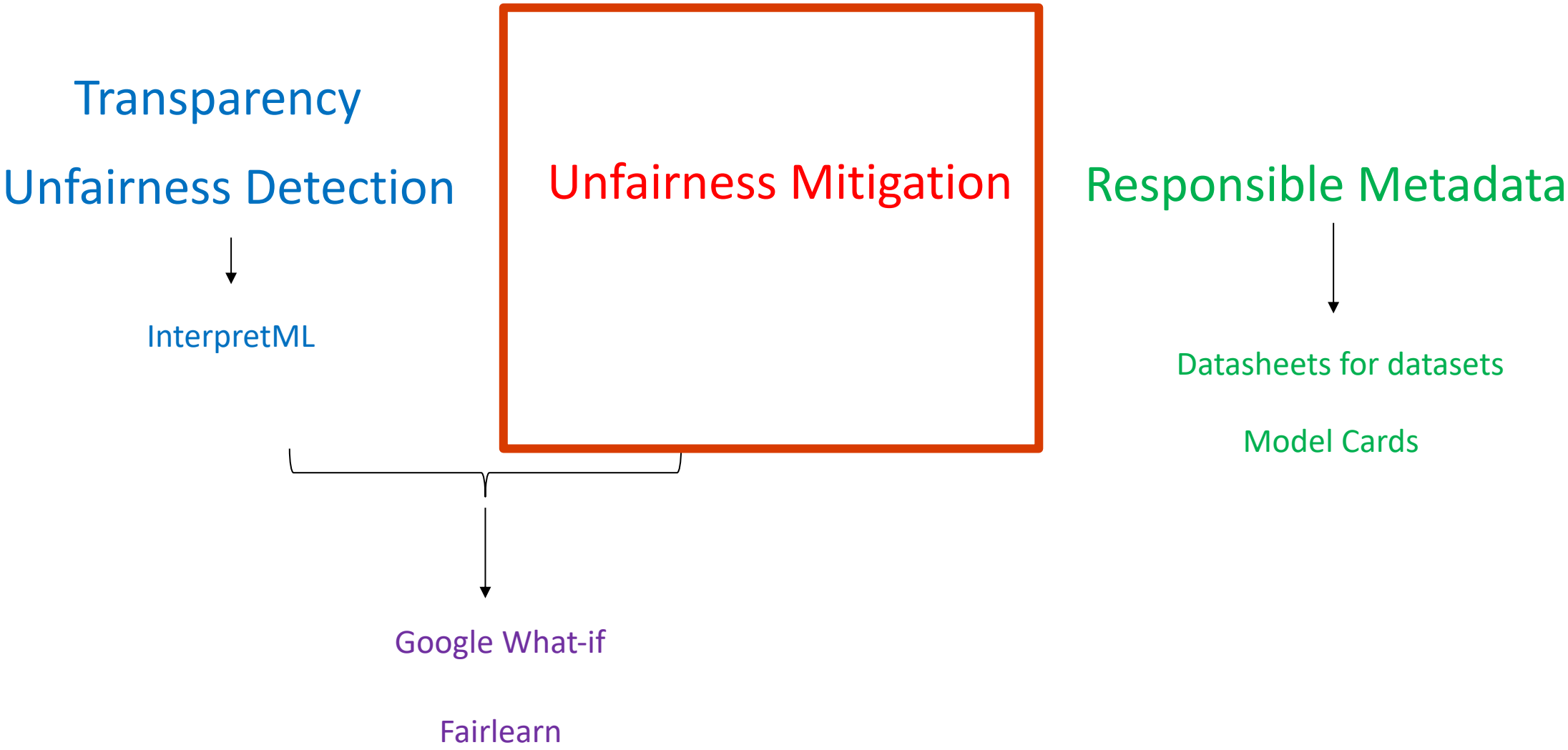
InterpretML

github.com/interpretML
pip install –U interpret

Goal: to provide researchers and AI developers with a toolkit that allows for:

- Explaining machine learning models globally on all data, or locally on a specific data point using the state-of-art technologies
- Easily adding new explainers and compare them to the state-of-the-art explainers
- A common API and data structure across the integrated libraries

Overview of Transparency and Fairness Tools

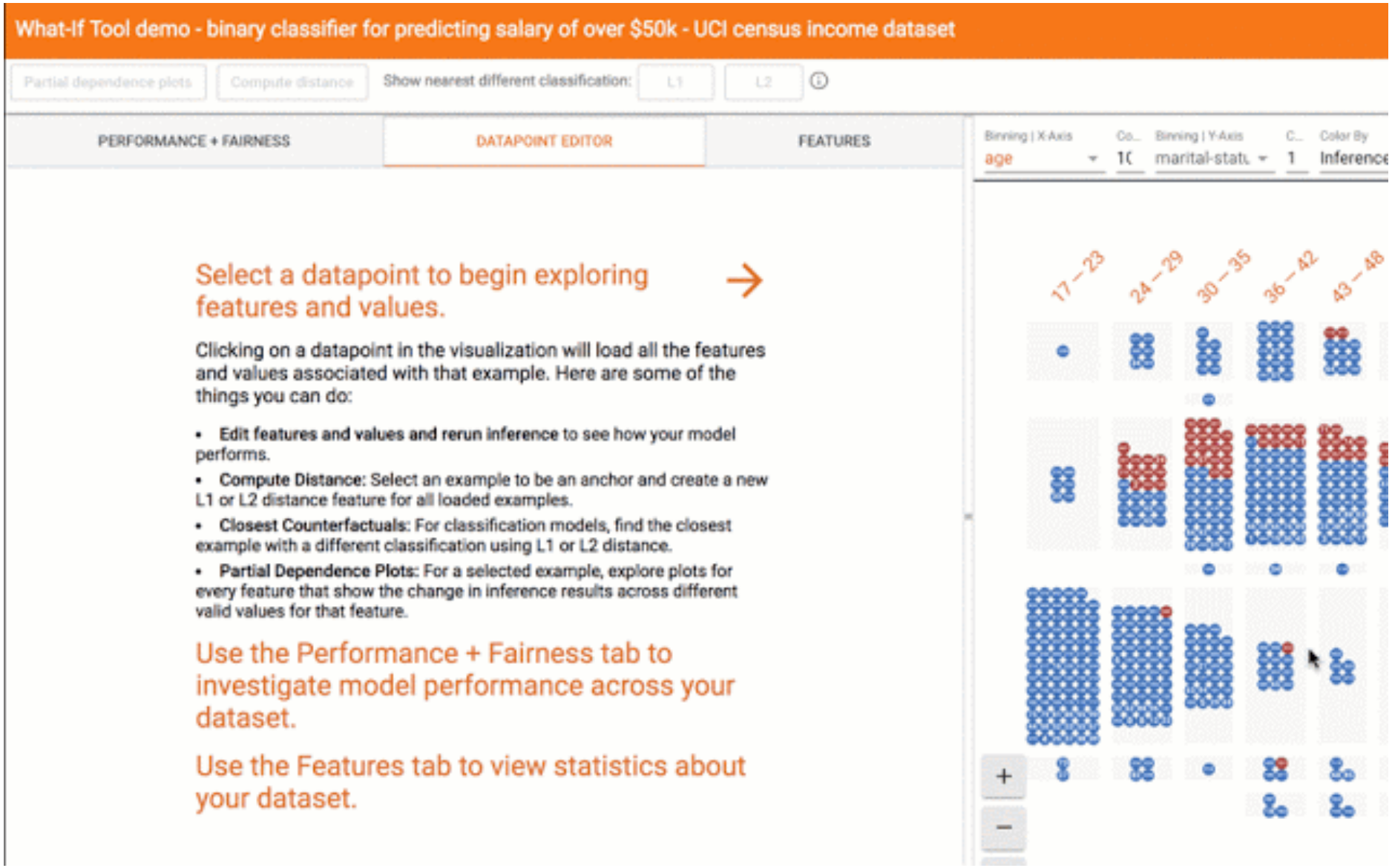


What If Tool

Goal: Code-free probing of machine learning models

- Feature perturbations (what if scenarios)
- Counterfactual example analysis
- [Classification] Explore the effects of different classification thresholds, taking into account constraints such as different numerical fairness metrics.

What If Tool


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
Select a datapoint to begin exploring features and values.
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.
Binning | X-Axis
age
Binning | Y-Axis
marital-stat
Color By
Inference
17 – 23
24 – 29
30 – 35
36 – 42
43 – 48

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

Datapoint editor | Performance & Fairness | Features

500 datapoints loaded

Visualize

Datapoints | Partial dependence plots

Show nearest counterfactual datapoint | L1 | L2 | Model: 1

Show similarity to selected datapoint

Edit

Search features

Select a datapoint to begin exploring model behavior for your selection.

Edit and Infer: Edit your datapoint here and run inference in the Infer table to see differences in model behavior.

Visualize: Switch between visualizing datapoints and exploring partial dependence plots to gain insights into your model's behavior. Explore counterfactuals or see how similar (or different) the rest of your dataset is from your selection.

Infer

Run inference

Binning | X-Axis: (none)

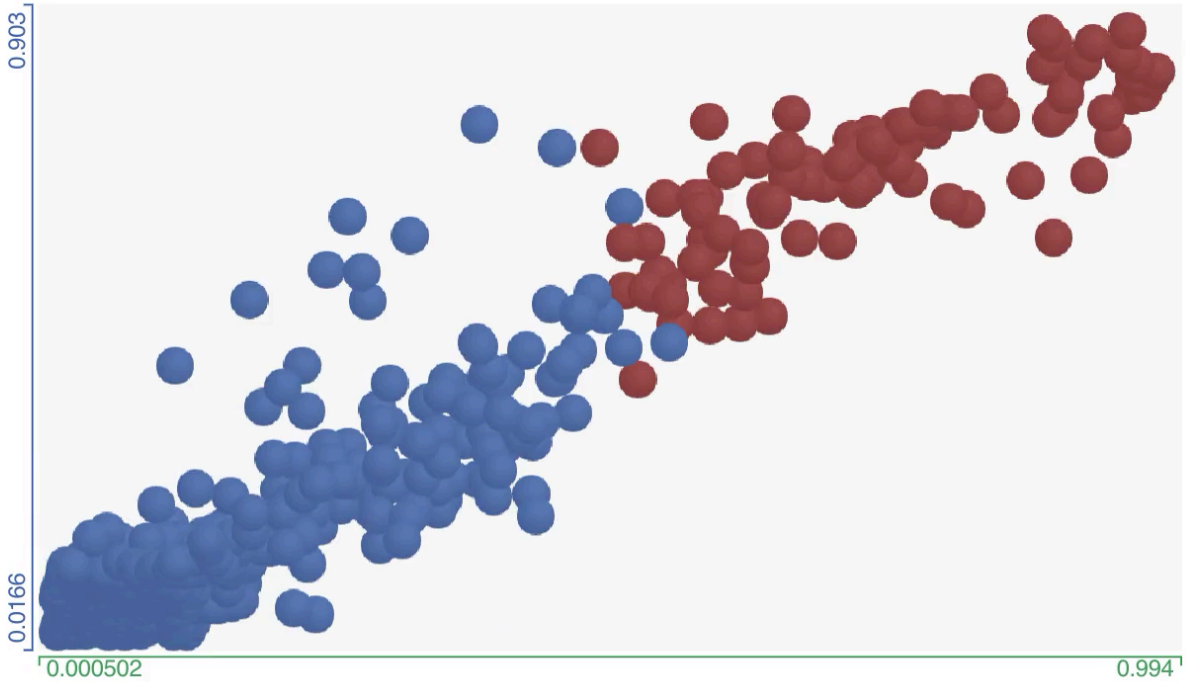
Binning | Y-Axis: (none)

Color By: Inference lab…

Label By: (default)

Scatter | X-Axis: Inference score

Scatter | Y-Axis: Inference score 2



Legend

Colors
by Inference label 1

<=50k

>50k

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

Datapoint editor | Performance & Fairness | Features

500 datapoints loaded

Configure

Ground Truth Feature: over_50k

WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: <none>

WHAT DOES SLICING DO?
Shows performance for each value of the selected feature.

Fairness

Apply an optimization strategy

Select a strategy to set classification thresholds based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will default back to custom thresholds.

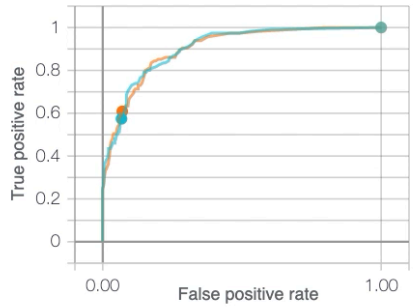
- (•) Custom thresholds
- ( ) Single threshold
- ( ) Demographic parity
- ( ) Equal opportunity
- ( ) Equal accuracy
- ( ) Group thresholds

Explore overall performance

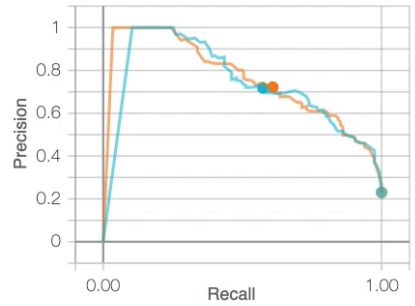
Sort by: Count

Feature Value	Count	Model	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	500	1	0.5	5.2	9.8	85.0	0.64
		2	0.5	5.4	9.0	85.6	0.66

ROC curve



PR curve



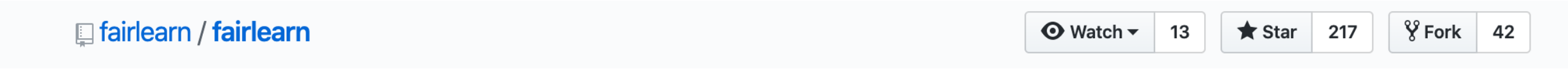
Confusion matrix

1	Predicted Yes		Predicted No		Total	
Actual Yes	13.2%	(66)	9.8%	(49)	23.0%	(115)
Actual No	5.2%	(26)	71.8%	(359)	77.0%	(385)
Total	18.4%	(92)	81.6%	(408)		

2	Predicted Yes		Predicted No		Total	
Actual Yes	14.0%	(70)	9.0%	(45)	23.0%	(115)
Actual No	5.4%	(27)	71.6%	(358)	77.0%	(385)
Total	19.4%	(97)	80.6%	(403)		

Fairlearn Toolkit

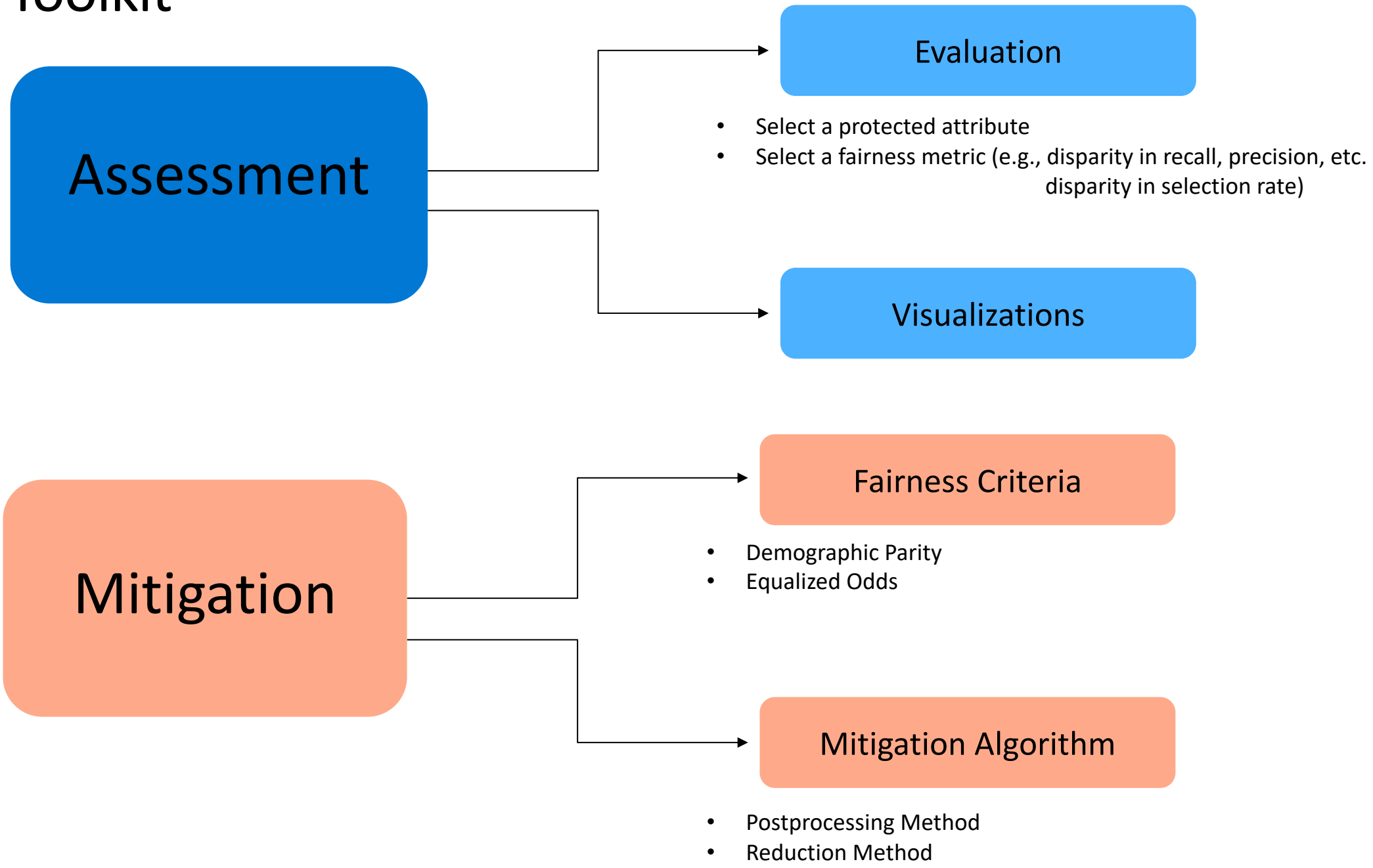
- Released at Ignite 2019 (Nov 2019)



- Empowers developers of artificial intelligence systems to **assess their systems' fairness** and **mitigate any observed fairness issues**.

- Focuses on **negative impacts for groups of people**, such as those defined in terms of race, gender, age, or disability status.

Fairlearn Toolkit



Fairness Criteria

Demographic parity

Applicants of each race (gender, …) have the
same odds of getting approval on their loan applications

Loan approval decision is independent of protected attribute

Equalized odds

Qualified applicants have the same odds of getting approval on their loan applications regardless of race (gender, …)

Unqualified applicants have the same odds of getting approval on their loan applications regardless of race (gender, …)

Overview of Transparency and Fairness Tools

Bias Detection

H2O

Microsoft InterpretML

Microsoft Azure Interpretability Toolkit

IBM Open Scale

Google What-if

IBM Fairness 360

Microsoft Fairlearn

Bias Mitigation

Responsible Metadata

Datasheets for datasets

Model Cards

Fact Sheets

Datasheets for Datasets [Gebru et al., 2018]

- Better data-related documentation
  - Datasheets for datasets: every dataset, model, or pre-trained API should be accompanied by a data sheet that documents its
    - Creation
    - Intended uses
    - Limitations
    - Maintenance
    - Legal and ethical considerations
    - Etc.

Model Cards for Model Reporting [Mitchell et al., 2018]

Intended use

Human-assisted moderation

Make moderation easier with an ML assisted tool that helps prioritize comments for human moderation, and create custom tasks for automated actions. See our moderator tool as an example.

Author feedback

Assist authors in real-time when their comments might violate your community guidelines or be may be perceived as "Toxic" to the conversation. Use simple feedback tools when the assistant gets it wrong. See our authorship demo as an example.

Read better comments

Organize comments on topics that are often difficult to discuss online. Build new tools that help people explore the conversation.

Uses to avoid

Fully automated moderation

Perspective is not intended to be used for fully automated moderation. Machine learning models will always make some mistakes, so it is essential to build in systems for humans to catch and correct those mistakes.

Character judgement

In order to maintain user privacy, the TOXICITY model only helps detect toxicity in an individual statement, and is not intended to detect anything about the individual who said it. In addition, Perspective does not use prior information about an individual to inform toxicity predictions.

Model details

Training data

Proprietary from Perspective API, which includes comments from online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic", defined as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion".

Model architecture

The model is a Convolutional Neural Network (CNN) trained with GloVe word embeddings, which are fine-tuned during training. You can also train your own deep CNN for text classification on our public toxicity dataset, and explore our open-source model training tools to train your own models.

Values

Community, Transparency, Inclusivity, Privacy, and Topic neutrality. These values guide our product and research decisions.

Future Directions

Future Directions

Additional Open Source Capabilities

Deeper Platform Integration

New Security and Privacy Technology