



# Responsible AI Toolbox

Besmira Nushi, Rahee Ghosh Peshawaria, Wenxin Wei  
Minsoo Thigpen, Mehrnoosh Sameki  
Microsoft

Contact: [rai-toolbox@microsoft.com](mailto:rai-toolbox@microsoft.com)

---

# Agenda

Introduction to the Responsible AI Toolbox



Responsible AI Dashboard - Demo



Vision and Language Tasks – Demo & QA

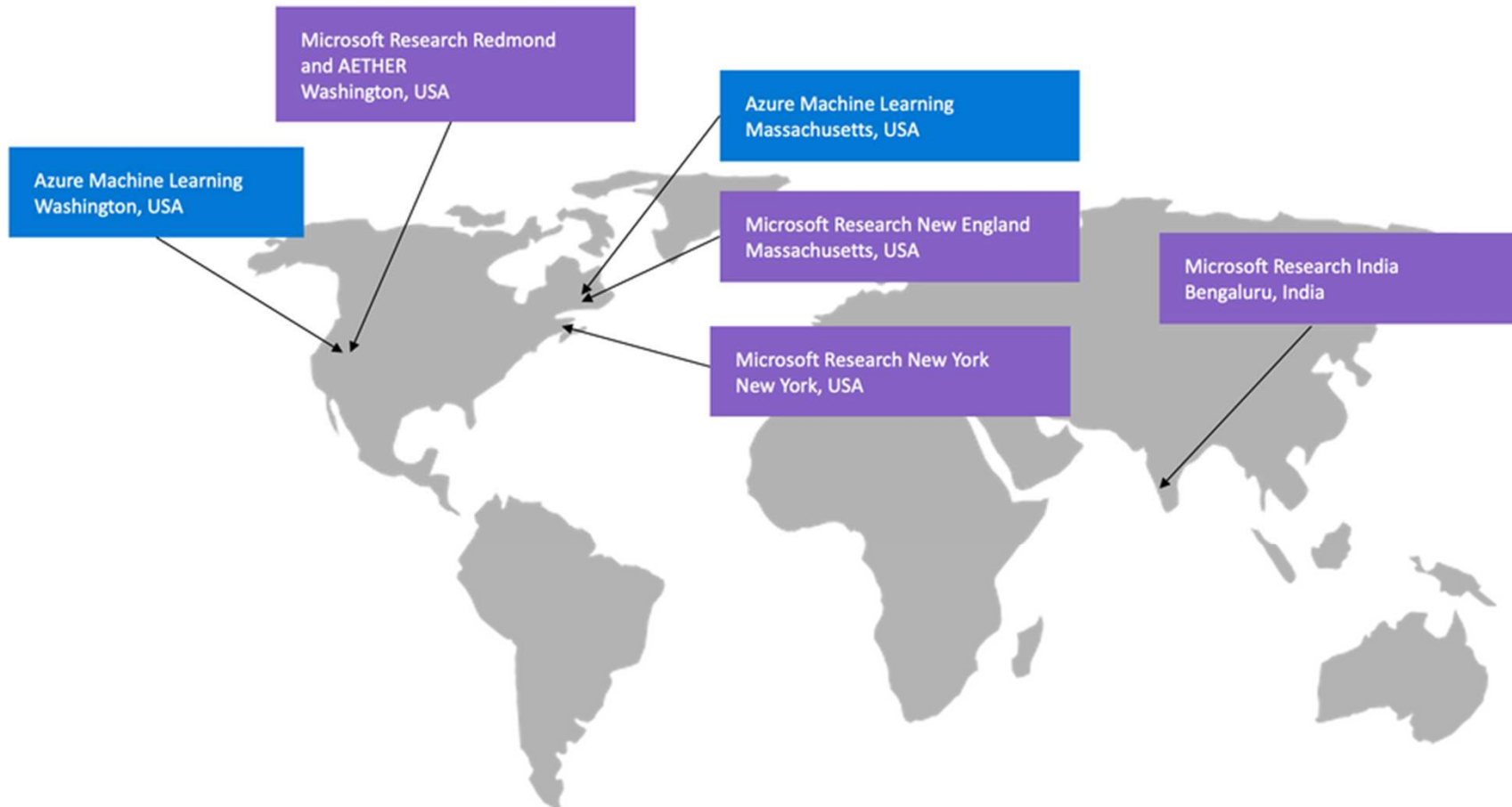


Responsible AI Mitigations and Tracker - Demo

User Insights, Challenges, and Opportunities

QA

# Thanks to our v-team!



# Microsoft's Responsible AI Principles

<https://microsoft.sharepoint.com/sites/ResponsibleAI>



Fairness



Reliability  
& Safety



Privacy &  
Security



Inclusiveness



Transparency



Accountability

Common need: Evaluation of system performance across demographic groups, key use cases, and operational factors.



**The path to  
deploying reliable  
machine learning  
systems is still  
unpaved.**

Software Engineering for ML: A Case Study  
ICSE 2019



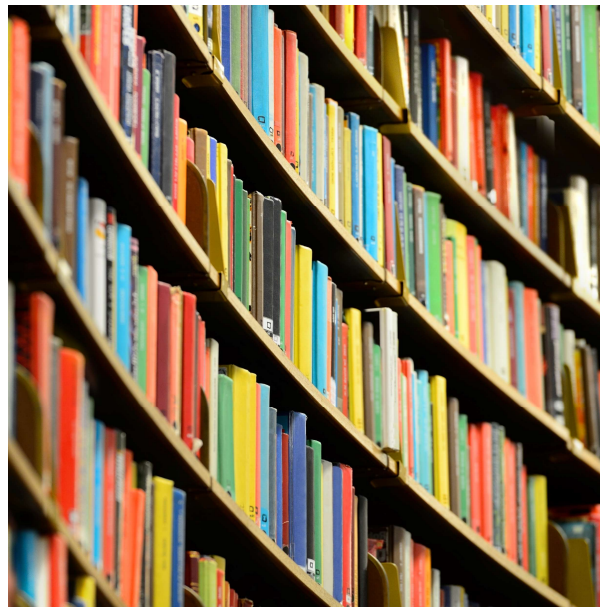


# Key Challenge: Tool Fragmentation

## Desiderata for Tool Integration



**Learnability**



**Discoverability**



**Sharing Insights & Data**

# Current Tools: Open-source Building Blocks

- InterpretML – [interpret.ml](https://interpret.ml)
- Error Analysis – [erroranalysis.ai](https://erroranalysis.ai)
- Fairlearn – [fairlearn.github.io](https://fairlearn.github.io)
- DiCE – [github.com/interpretml/dice](https://github.com/interpretml/dice)
- EconML – [aka.ms/econml](https://aka.ms/econml)
- DoWhy – [github.com/microsoft/dowhy](https://github.com/microsoft/dowhy)
- BackwardCompatibilityML – [github.com/microsoft/BackwardCompatibilityML](https://github.com/microsoft/BackwardCompatibilityML)

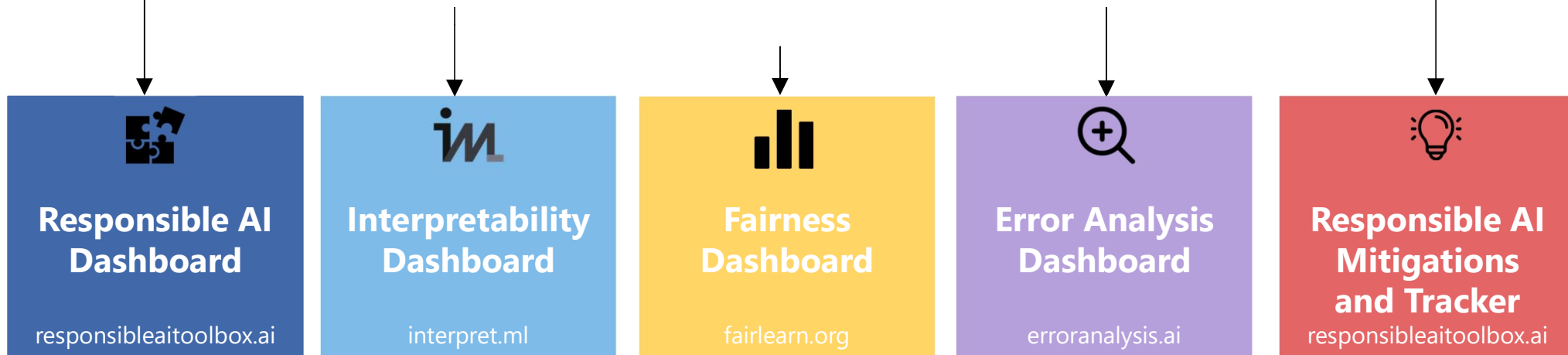


# Introducing: Responsible AI Toolbox

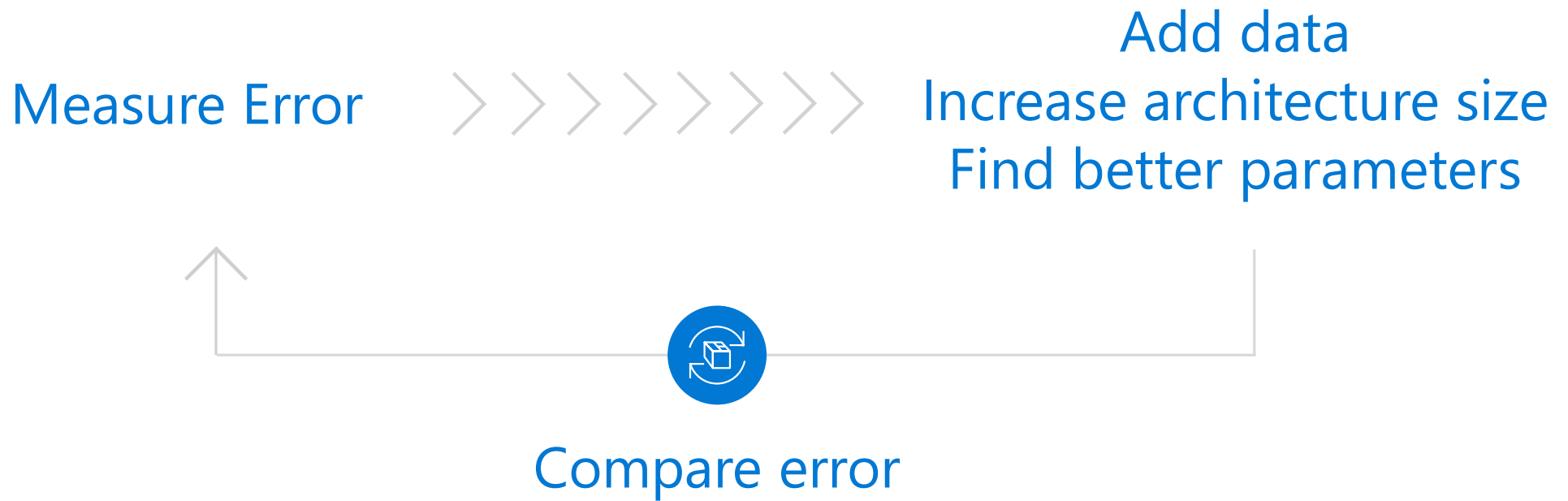
# Responsible AI Toolbox

[responsibleaitoolbox.ai](https://responsibleaitoolbox.ai)

An open-source framework for **accelerating** and **operationalizing Responsible AI** via a set of **interoperable** tools, libraries, and customizable dashboards.



# Current Model Debugging & Improvement Approaches



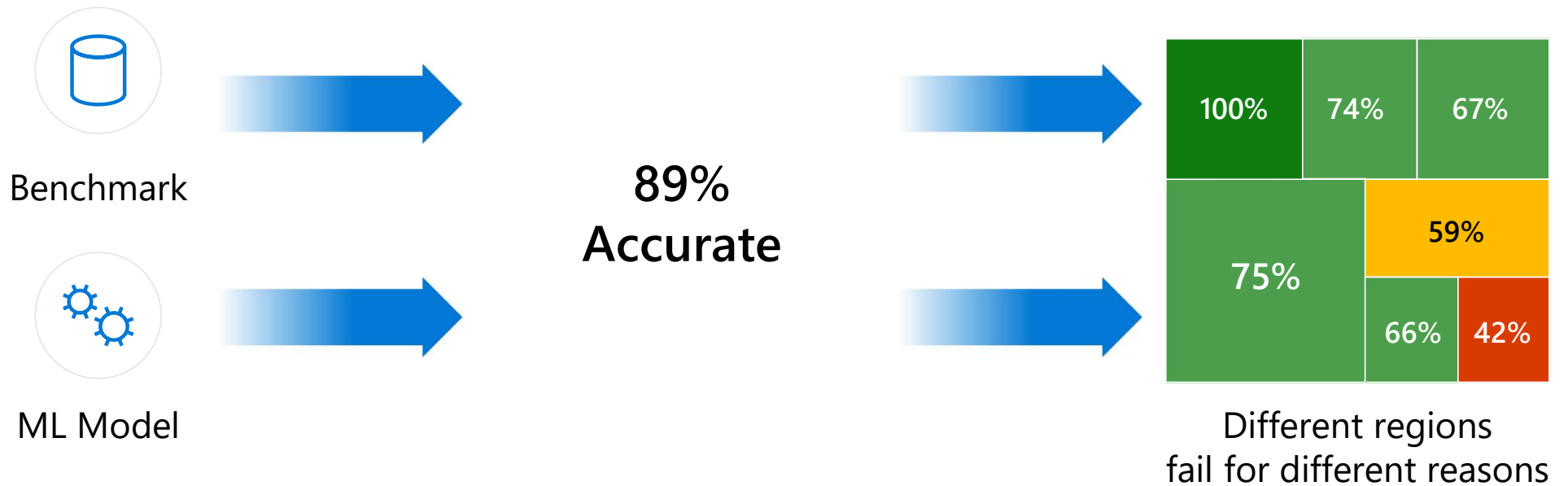
# Evaluating machine learning models aka the problem with aggregated metrics

## AI-powered scans can identify people at risk of a fatal heart attack almost a **DECADE** in advance 'by looking at the entire iceberg and not just the tip'

- The AI predicted heart risk with **90% accuracy**, according to data
- Current medical scans are only able to see 'the tip of the iceberg'
- It could benefit around 350,000 in Britain, cardiologists believe
- Government funding will fast track the tech into the NHS in two years

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	PFDet		+3	0.62882	49	1y
2	Avengers			0.62161	48	1y
3	kivajok			0.61707	102	1y
4	XJTU			0.61559	22	1y
5	ikciting		+5	0.59472	39	1y
6	Sogou_MM			0.57936	105	1y
7	QLearning			0.56688	20	1y
8	[RingUkraine] CloudResearch			0.53742	50	1y
9	Res101+SoftNMS			0.53413	29	1y
10	Kyle L.			0.51464	53	1y

# Why isn't this sufficient?





# Emotion Recognition

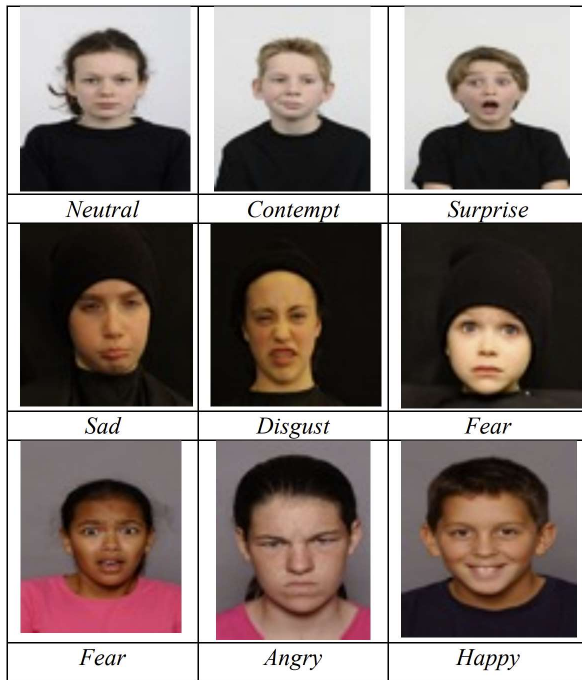







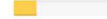












TABLE I. DEEP LEARNING RECOGNITION RATES ACROSS THE DIFFERENT STIMULI SETS (IN %): (FE)AR, (AN)GRY, (HA)PPY, (SA)D, (NE)UTRAL, (SU)RPRISED, (DI)SGUST, (CO)NTEMPT

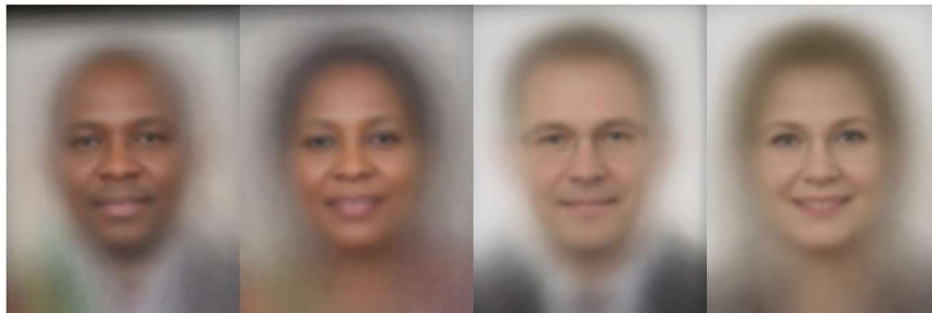
	Fe	An	Di	Ha	Ne	Sa	Su	Co
<b>NIMH-ChEFS</b>	13	43		100	100	48		
<b>Dartmouth</b>	25	35	55	100	99	64	91	
<b>Radboud</b>	33	54	100	100	100	95	100	50
<b>CEPS</b>	5	50	10	95	92	52	81	

[Howard et al., ARSO 2017]

Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems

# GenderShades Study

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 



[Buolamwini and Gebru, FAccT 2018]

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

# Follow up case study

All data

→ 5.5 % error rate

Women, No makeup,  
Short/Tied hair, Not smiling

→ 35.7 % error rate

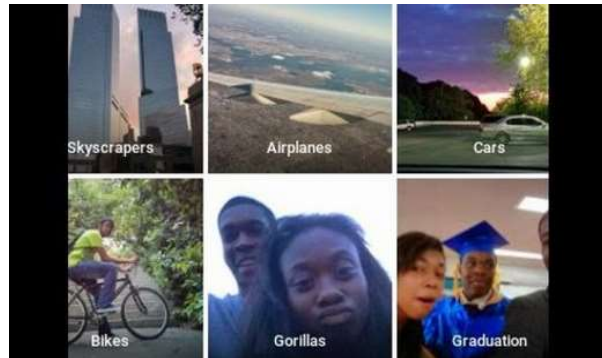
[Nushi et al., ICLR DebugML 2018]

Error terrain analysis for machine learning:  
Tool and visualizations

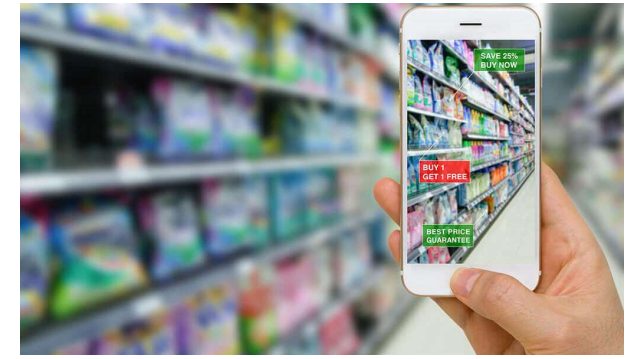
# Performance discrepancies in the real world



Safety



Fairness



Trust

# Concepts of disaggregated evaluation

## **Cohort (aka data slices, regions, subgroups, clusters):**

Subsets of data created by adding filters to the overall test or train datasets.

Examples:

```
"age > 40 and residency= 'Florida'"
```

```
"gender=female and 'diabetes' in pre_existing_conditions"
```

## **Performance discrepancy (ratio or difference):**

- Discrepancy between all data vs. cohort of interest
- Discrepancy between two cohorts of interest  
Example: WA residents vs NY residents
- The best and worst performance across combinations of features.  
Example: the best and worst performance for combinations of gender and age
- Discrepancy between cohorts with the best and worst performance

# Cohort design considerations

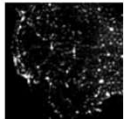
1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery

Filters can be created based on feature values for tabular data.

## Census Income Data Set

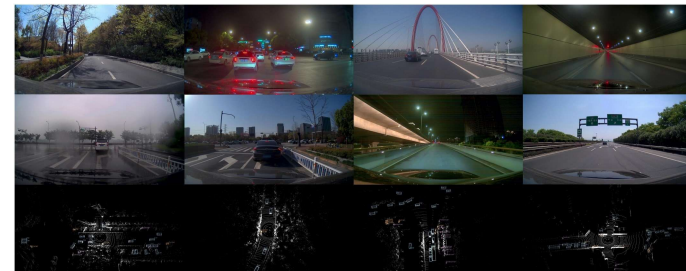
Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** Predict whether income exceeds \$50K/yr based on census data. Also known as "Adult" dataset.



<b>Data Set Characteristics:</b>	Multivariate	<b>Number of Instances:</b>	48842	<b>Area:</b>	Social
<b>Attribute Characteristics:</b>	Categorical, Integer	<b>Number of Attributes:</b>	14	<b>Date Donated</b>	1996-05-01
<b>Associated Tasks:</b>	Classification	<b>Missing Values?</b>	Yes	<b>Number of Web Hits:</b>	739048

Filters can be (softly) inferred using image/text processing or auxiliary models.



# Cohort design considerations

1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery

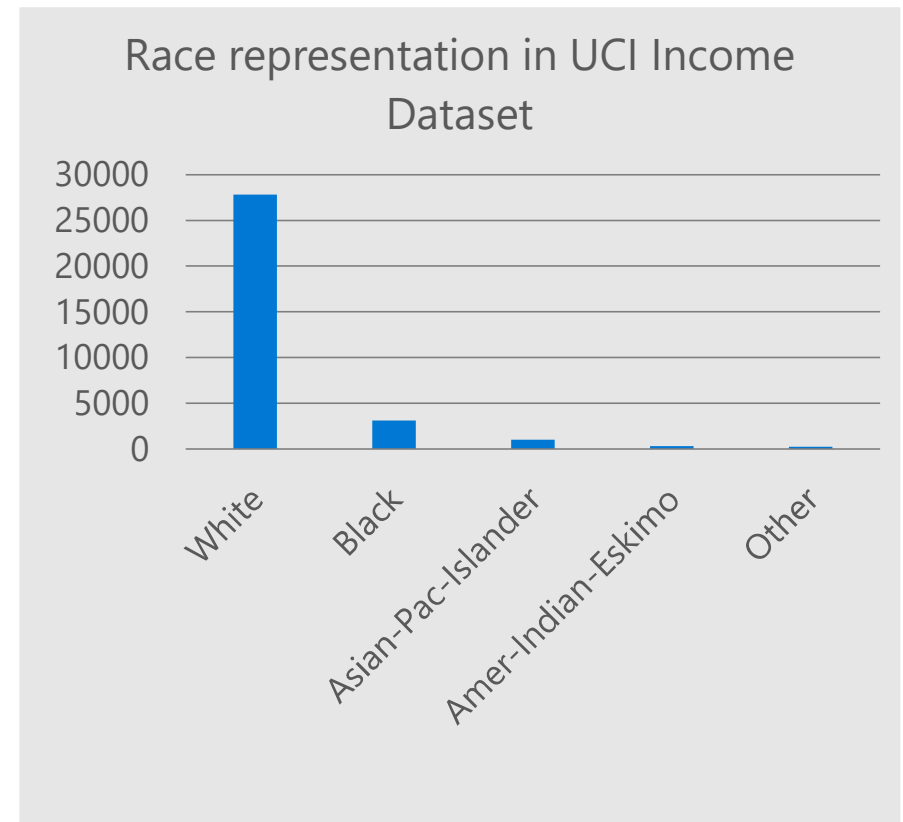


## **Credit risk assignment example**

- 20% false positives for small loans  
(e.g. < \$5000)
- 5% false positives for larger loans  
(e.g. > \$20,000)

# Cohort design considerations

1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery

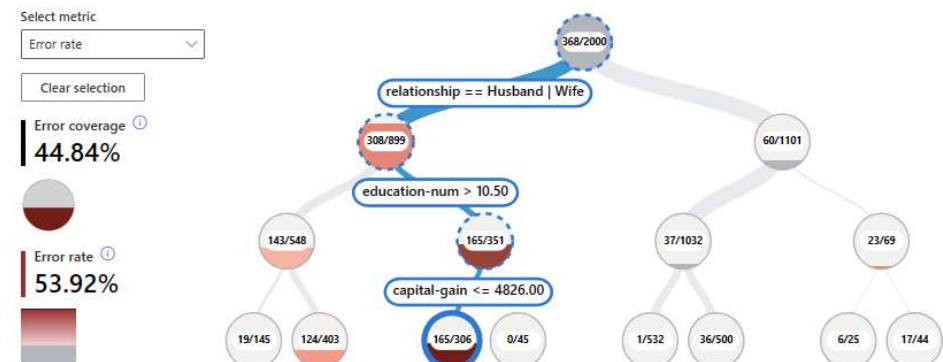


# Cohort design considerations

1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery

## Automated discovery

Useful for quick discovery of cohorts with significantly higher error rates



Visualization based on Responsible AI Dashboard:  
<https://github.com/microsoft/responsible-ai-toolbox>



# Cohort design considerations

1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery

## Manual discovery

Useful for exploring errors on known important cohort definitions.

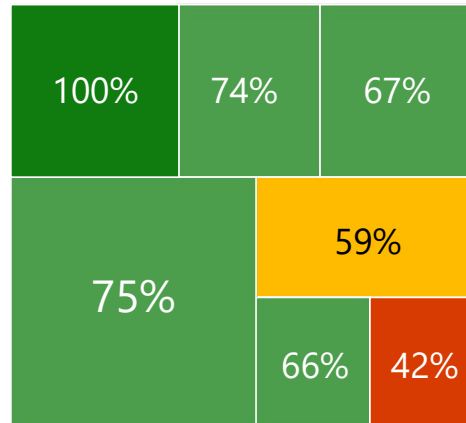


Visualization based on Responsible AI Dashboard:  
<https://github.com/microsoft/responsible-ai-toolbox>

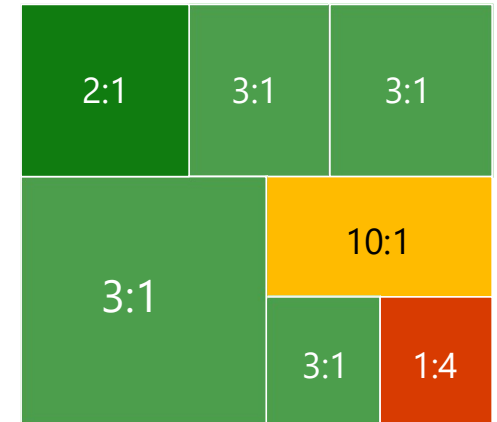
# Disaggregation: from evaluation to debugging

## ! Data and model debugging

- Imbalance
- Noise
- Missing values
- Distribution shifts
- Spurious correlations
- Wrong labels



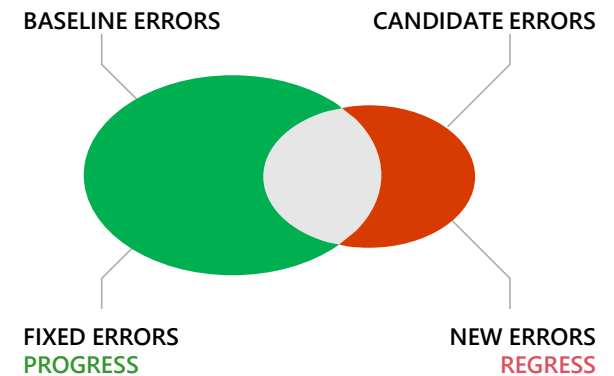
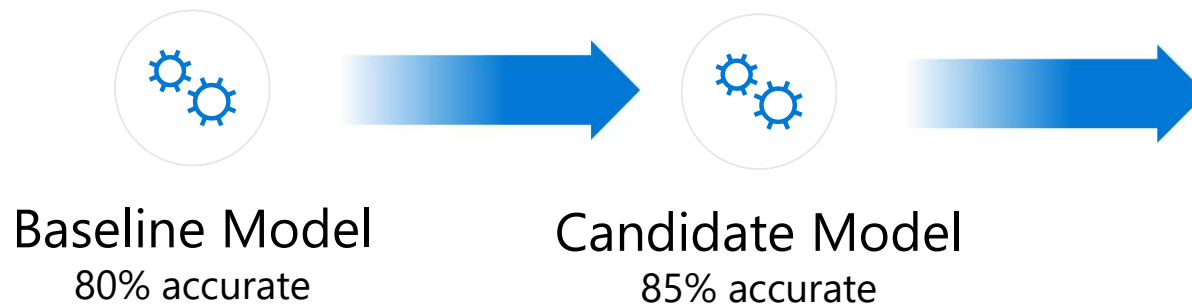
Disaggregated evaluation  
Discrepancy metrics



Disaggregated training data metrics  
e.g. class imbalance etc.

Different cohorts may have very different class imbalances which may or may not align with the overall class balance ratios in the training data.

# Disaggregated model comparison



Model Updates may lead to new mistakes and lost trust.

# Incompatibility Sources

## **Optimization Stochasticity**

Stochastic batches in gradient descent

Model initialization

Random data augmentation

Distributed training

## **Label Noise**

Semi-supervised learning with noisy data

Human labeling error

## **Distributional Shifts**

Training data is not a representation of the real world

Bias in data collection

The concept definition changes

Domain transfer

## **Model Class**

Fundamental architectural changes

# Compatibility is not built-in

## Updates in Practice

[Bansal et al., AAI 2019]  
Updates in Human-AI Teams:  
Understanding and Addressing  
the Performance/Compatibility  
Tradeoff

[Srivastava et al., KDD 2020]  
An empirical analysis of backward  
compatibility in machine learning  
systems

Classifier	Dataset	Perf. v1	Perf. v2	Compatibility
Logistic Regression	Recidivism	0.68	0.72	72%
	Credit Risk	0.72	0.77	66%
	Mortality	0.68	0.77	40%
Multi-layered Perceptron	Recidivism	0.59	0.73	53%
	Credit Risk	0.70	0.80	63%
	Mortality	0.71	0.84	76%

High-stake decision-making

Backward compatibility scores available at:  
<https://github.com/microsoft/BackwardCompatibilityML>

Low compatibility

Percentage of  
predictions that  
remain correct.

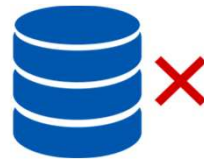
# Targeted Debugging for Machine Learning

Responsible AI  
Dashboard



Identify

Responsible AI  
Dashboard & Mitigations

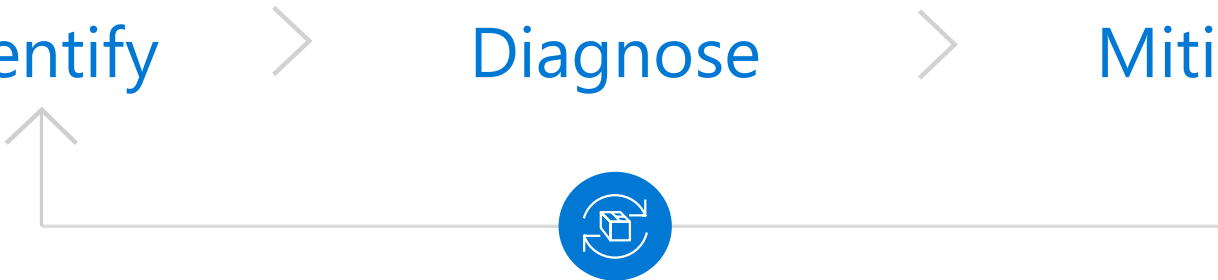


Diagnose

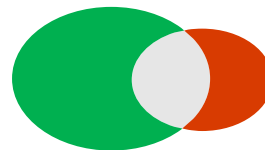
Responsible AI  
Mitigations



Mitigate



Track, Compare, Validate



Responsible AI Tracker

# Debugging Machine Learning Models

Identify



Diagnose



Mitigate



**Fairlearn**

Fairness Assessment



**InterpretML**

Interpret and Debug Models  
Perform Feature Perturbations



**Fairlearn**

Unfairness Mitigation Algorithms



**Error-Analysis**

Error Analysis



**Counterfactual**

Diverse Counterfactual Explanations  
for Debugging



**Responsible AI  
Mitigations**

Enhance your dataset and  
retrain model



**Exploratory-Data-Analysis**

Understand Dataset Characteristics



Responsible AI Tracker  
Model  
Comparison

**Compare & Validate**

Backward  
Compatibility

# The future of data science productivity and tools



code



data



model



visualizations

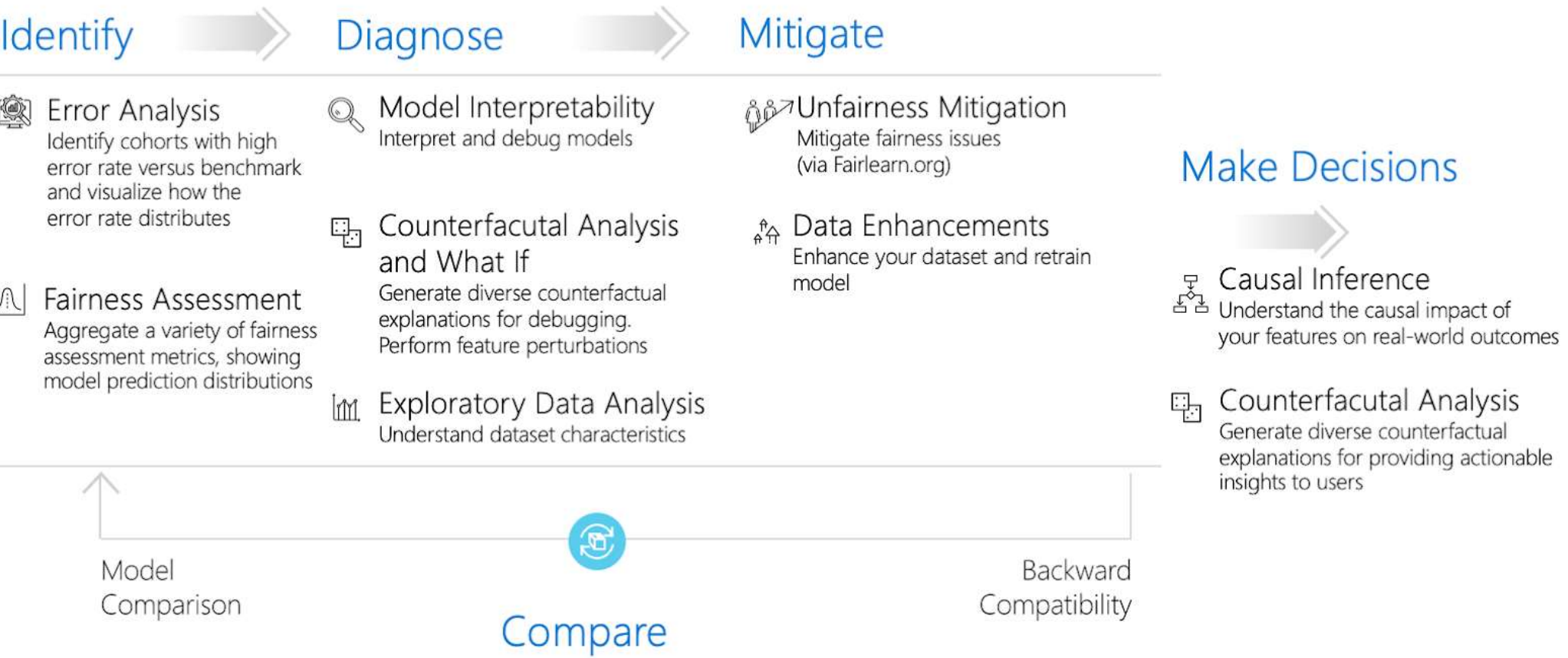


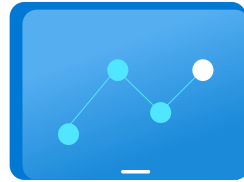


# Responsible AI Dashboard ML Debugging and Causal Decision-Making

# Responsible AI Dashboard

An open-source framework for accelerating and operationalizing Responsible AI via a set of interoperable tools, libraries, and customizable dashboards.



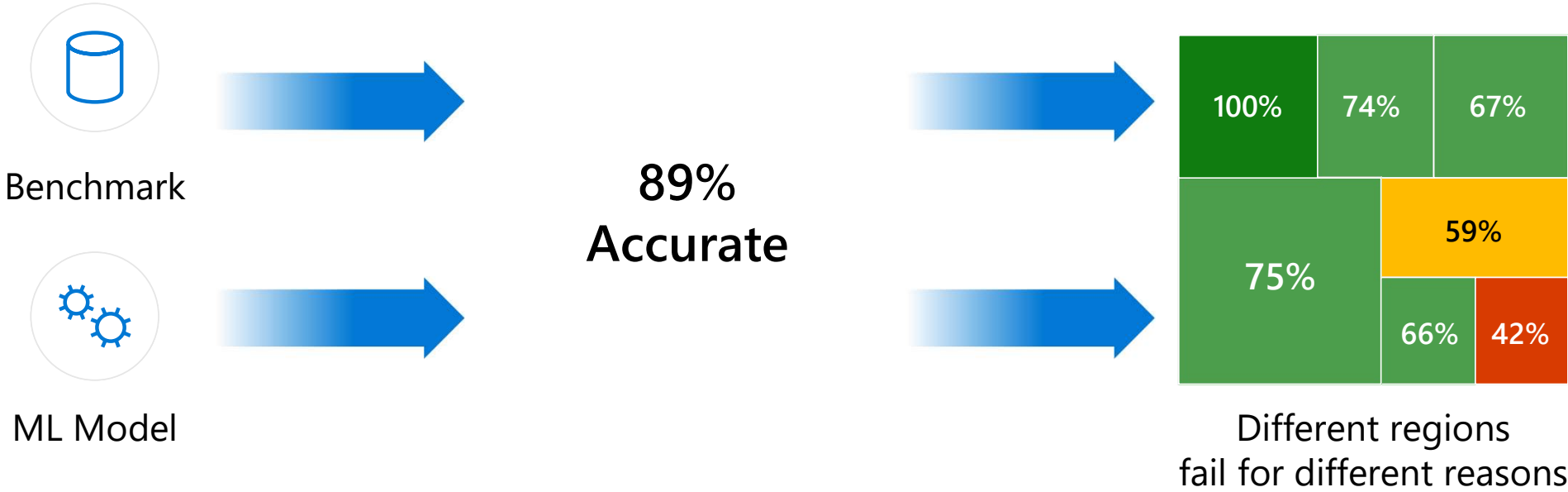


## Identify

Error Analysis  
Fairness Assessment

# Error Analysis

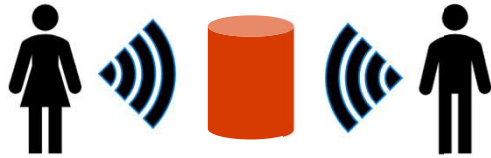
Rigorous performance evaluation and testing is often needed to deploy models in production.



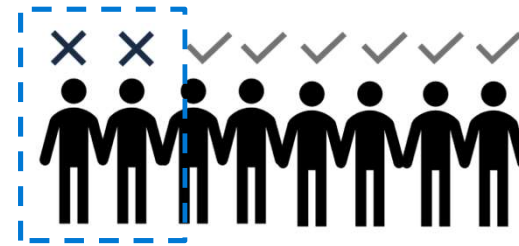
Analyze and debug model errors

# Fairness in AI

There are many ways that an AI system can behave unfairly.



A voice recognition system might fail to work as well for women as it does for men.



A model for screening job application might be much better at picking good candidates among white men than among other groups.

Avoiding negative outcomes of AI systems for different groups of people

Learn more

<https://github.com/microsoft/responsible-ai-toolbox> and <https://github.com/fairlearn>



## Diagnose

Interpretability  
Counterfactuals  
Data Exploration

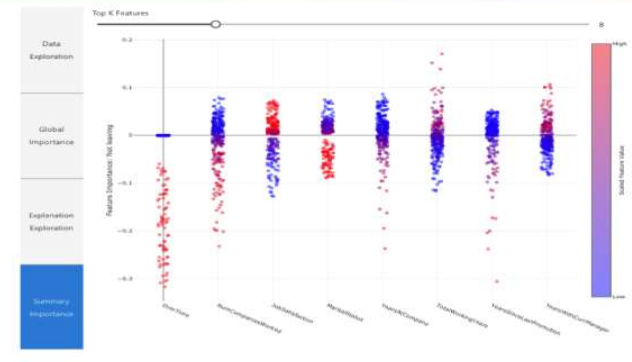
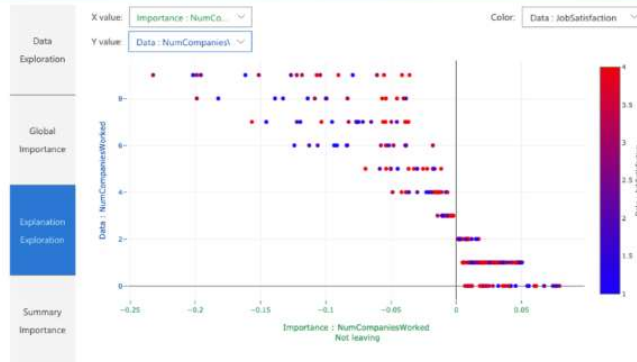
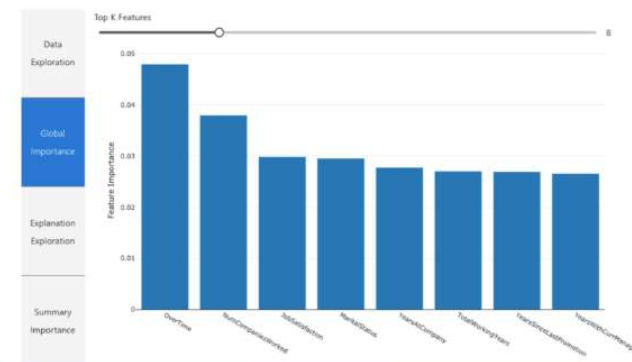
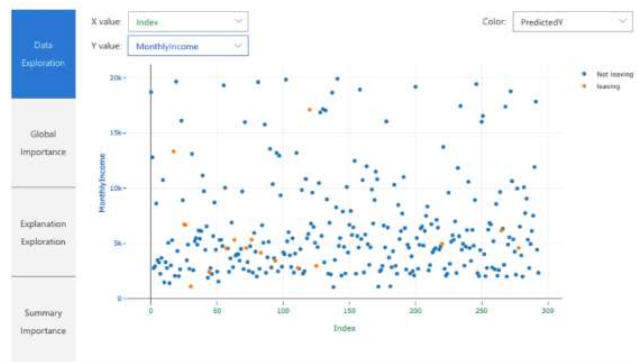
# Interpretability

## Understand overall model predictions

What are the top K important factors impacting your overall model predictions?

## Understand individual model predictions

What are the top K important factors impacting your model predictions for a single sample?



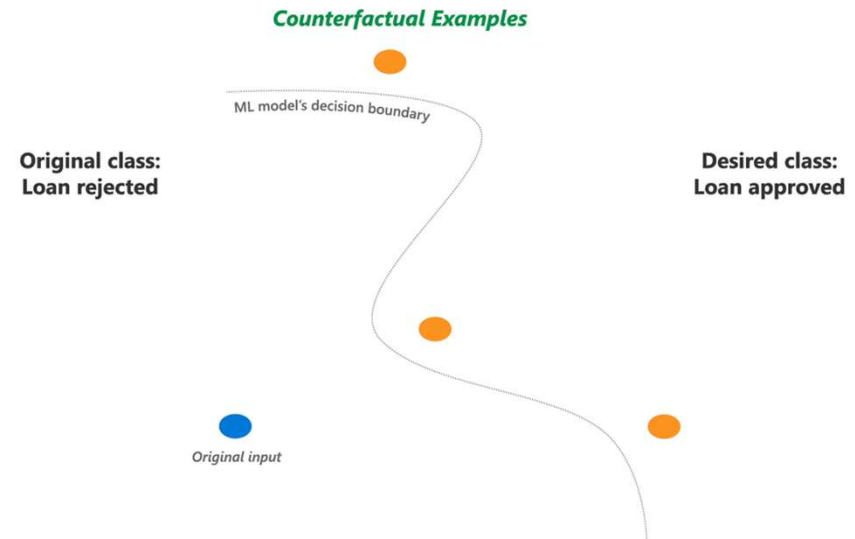
# Counterfactuals

## Debug model predictions

Enable data scientists and model evaluators to debug models by understanding the closest datapoints with different prediction outcomes

## Make responsible model-driven decisions

Answer end-users' questions such as "what can I do to get a different outcome from the AI model?"







## Mitigate

Model Fairness mitigations

Data mitigations



## Take Action

Causal Inference  
Counterfactual Analysis

# Causal Inference

## Understand overall causal effects

Answer real-world "what if" questions about how an outcome would have changed under different policy choices.

## Explore individual causal effects

Inform personalized interventions, such as a targeted promotion to customers or an individualized treatment plan. Learn about how an individual with a particular set of features respond to a change in a causal feature, or treatment.

## Extract a treatment policy

Build policies for future interventions. Identify what parts of your sample experience the largest responses to changes in causal features, or treatments, and construct rules to define which future populations should be targeted for particular interventions.

Aggregate causal effects Individual causal what-if **Treatment policy**

These tools help build policies for future interventions. You can identify what parts of your sample experience the largest responses to changes in causal features, or treatments, and construct rules to define which future populations should be targeted for particular interventions.

Set treatment feature  
ScreenPorch

Interpretable recommended global treatment policy for sample size (n) = 730

	EnclosedPorch <= 15	EnclosedPorch > 15
OverallCond <= 6.5	n = 508 Recommended treatment = increase	n = 73 Recommended treatment = decrease
OverallCond > 6.5	OpenPorchSF <= 151 n = 138 Recommended treatment = decrease	OpenPorchSF > 151 n = 11 Recommended treatment = increase

This table shows a recommended treatment policy that can be applied to the current data sample or other populations. The table provides a simple rule to segment observations into data cohorts based on the features with the largest impact on whether the individual will respond to the selected treatment. The table also specifies number of datapoints in the current data sample assigned to each segment. The table can be read by taking a row and then taking a column of that specific row.

**Causal analysis**

Global cohorts are not currently supported for causal analysis. All causal analyses will be shown for all data.

Aggregate causal effects Individual causal what-if **Treatment policy**

Individual causal effects can inform personalized interventions, such as a targeted promotion to customers or an individualized treatment plan. How would an individual with a particular set of features respond to a change in a causal feature, or treatment? The causal what-if tool calculates marginal changes in real-world outcomes for a particular individual if you change their level of a treatment. This analysis enables you to understand how real-world outcomes would have changed under different policy choices, such as a different pricing strategy for a product or an alternative treatment for a patient. Specify the treatment of interest and observe how the real-world outcome would change.

Datapoint index  
Index 31

Select treatment  
OverallQual

Current treatment value: 6

Set new treatment value: 6

Current outcome: 196 (196, 196)

New outcome: 196 (196, 196)

Confidence interval (upper): 93.468  
Causal effect point: 38.596  
Confidence interval (lower): -14.276

GarageCars(num) OverallQual(num) ScreenPorch(num) Fireplaces(num) OverallCond(num)

Continuous treatments: On average in this sample, increasing this feature by 1 unit will cause the probability of class=label 1 to increase by X units.  
Binary treatments: On average in this sample, turning on this feature will cause the probability of class=label 1 to increase by X units.

A lasso (or logistic regression if y is binary) was fit to predict y from X[-i], and a lasso (or logistic regression if X[i] is categorical) was fit to predict X[i] from X[-i]. The causal effect can be viewed as the average correlation of the residuals/surprise variation of the two prediction tasks. Learn more about Double Machine Learning [here](#)

# Responsible AI Dashboard in Azure Machine Learning

*Generally Available in Azure Machine Learning and Open Source*

Microsoft Azure Machine Learning Studio

Microsoft > RAIPM2 > Models

Model List

+ Register model Refresh Delete Deploy Edit columns Reset view Show latest versions only

Search

Created on Created by Tags All filters Clear all

Showing 1-25 of 28 models Page size: 25

Name	☆	Version	Experiment	Run ID	Created on	Tags	Properties	Created by
component_registered_jr_01_16...		1		9f749efd-d0fc-43f3-9e89-5027...	Mar 24, 2022 6:37 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		a2773b64-ba51-413a-87de-01f...	Mar 24, 2022 6:37 PM		flavors.python_function	John Wu
component_registered_boston_...		1		fa76dcd0-a4d1-4e19-ba87-87ff...	Mar 22, 2022 12:57 PM		flavors.python_function	Minsoo Thigpen (SHE/HER)
component_registered_jr_01_16...		1		92bfe2d7-e940-4b34-bf1f-9563...	Feb 14, 2022 6:49 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		a89c54b-8380-4567-a796-078...	Feb 14, 2022 6:23 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		7bfa937d-cb39-4c0f-95af-7501...	Feb 14, 2022 4:42 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		c713933b-a4eb-468f-8af0-4c55...	Feb 14, 2022 4:40 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		c4d9334-0d20-4371-88cc-c3aa...	Feb 14, 2022 4:31 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		2759751a-b8f9-4cc0-9879-569...	Feb 14, 2022 11:51 AM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		a716ca86-b02c-432c-ba00-943...	Feb 13, 2022 2:45 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		5feb4b2b-6b36-4701-8c38-793...	Feb 13, 2022 2:19 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		b527ed9f-3bbc-41cd-a242-1db...	Feb 12, 2022 5:51 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		0798733-bcf6-46d0-9ac5-bf73...	Feb 12, 2022 5:36 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		940c7214-8f9c-4ce2-69bc-54e9...	Feb 12, 2022 5:32 PM		flavors.python_function	John Wu
component_registered_jr_01_16...		1		76a25d57-d166-4f95-a9e6-0fa...	Feb 12, 2022 5:29 PM		flavors.python_function	John Wu
component_registered_boston_...		1		e85a9394-86ab-4796-8090-a00...	Feb 8, 2022 4:57 PM		flavors.python_function	Rachel Kallam
RAI-Classification-Adult_v2_164...		1		0bb17c59-fa2e-47cf-852f-0beb...	Jan 13, 2022 2:02 PM		flavors.python_function	Rachel Kallam
my_trained_nb_model_1642098...		1		ae4b0bfe-9993-4b13-9891-a30...	Jan 13, 2022 1:37 PM		flavors.python_function	Rachel Kallam
my_trained_nb_model_1641918...		1		3f13e0d6-7348-41dc-a9e2-d09...	Jan 11, 2022 11:43 AM		flavors.python_function	Rachel Kallam

<< Page 1 of 2 >>

*A comprehensive single-pane-of-glass experience with a variety of model and data exploration capabilities such as **Error Analysis, Model Explanations, Fairness metrics, and Data Exploration.***



# Responsible AI Scorecard in Azure Machine Learning

Public Previewed in Azure Machine Learning

## Fairness Assessment

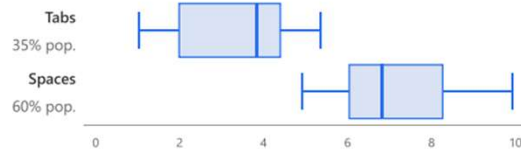
Observe performance differences between identified demographic groups, paying particular attention to the subgroups whose differences exceed the target maximum or minimum.

Tabs has the highest MSE: 3.78

Spaces has the lowest MSE: 3.55

Difference in MSE: 0.23

Prediction distribution chart

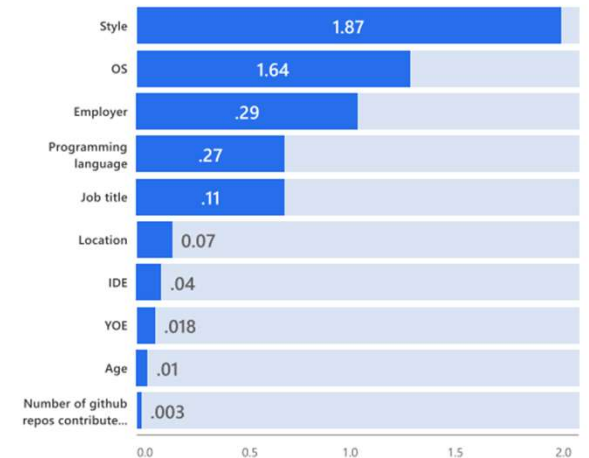


Analysis across cohorts:

	Average Prediction	Average Ground Truth	Mean Squared Error	Mean Absolute Error
Tabs	3.4	3	3.78	658
Spaces	7.1	8	3.55	543
Difference	3.7	5	.23	115
Ratio	.47	.37	.93	.82

## Feature relevance (explainability)

Understand factors that have impacted your model predictions the most. These are factors that may account for performance levels and differences.



Generate key summaries of Responsible AI insights by exporting to PDF. Share with technical and non-technical stakeholders to aid in compliance review.

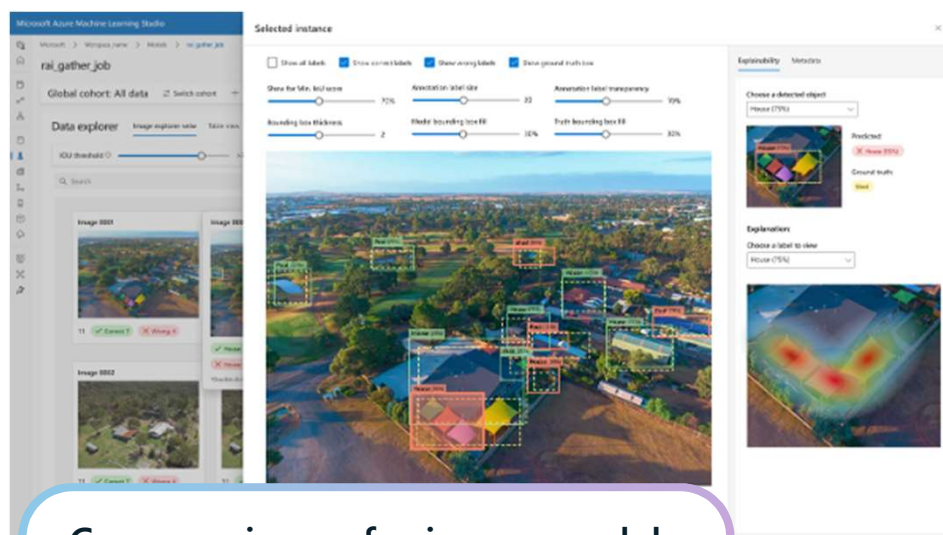


# Responsible AI Dashboard for Vision and Language



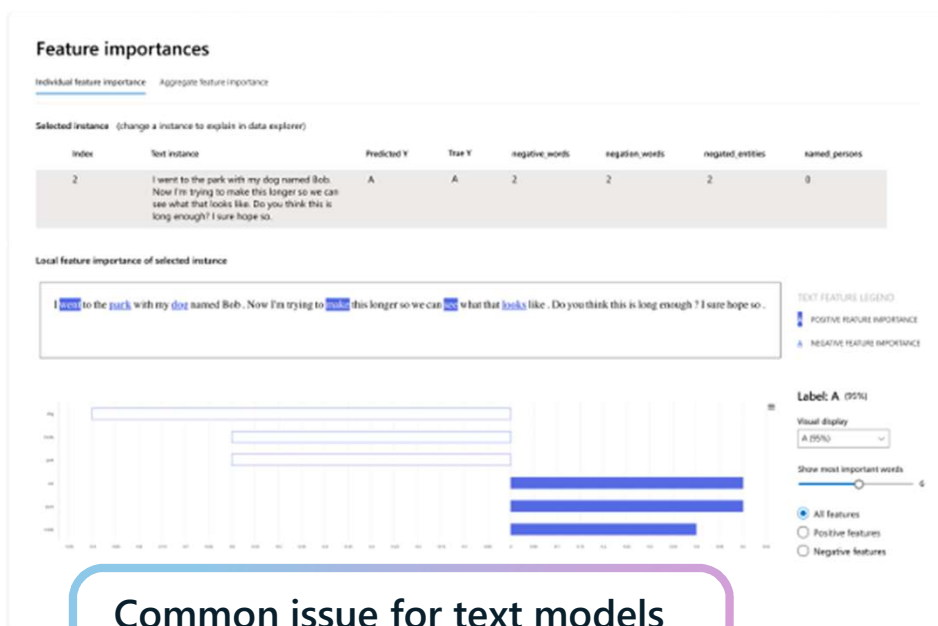
# Responsible AI dashboard support for image and text data

Newly available:



## Common issues for image models

- Misalignment of bounding boxes
- Object overlap
- Spurious correlations
- Labeling errors



## Common issue for text models

- Problems with grounding
- Linguistic shortcuts



# What is new?



Rich visualizations for vision and text



Meta-data support and ingestion

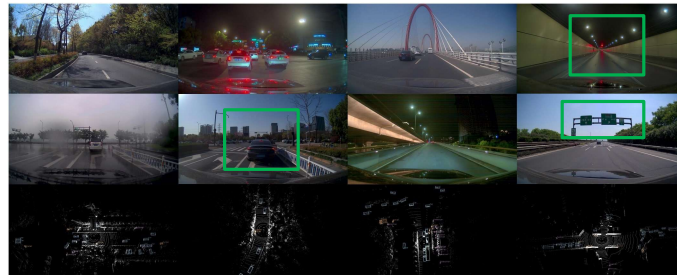


Interpretability for vision and text



Consistent design with customizable debugging workflows

# Meta-data for cohort design in Vision



Ground truth

Demographics  
Synthetics  
Bounding Box Info



System data

Camera Settings  
Time of day  
Location



Inferred attributes

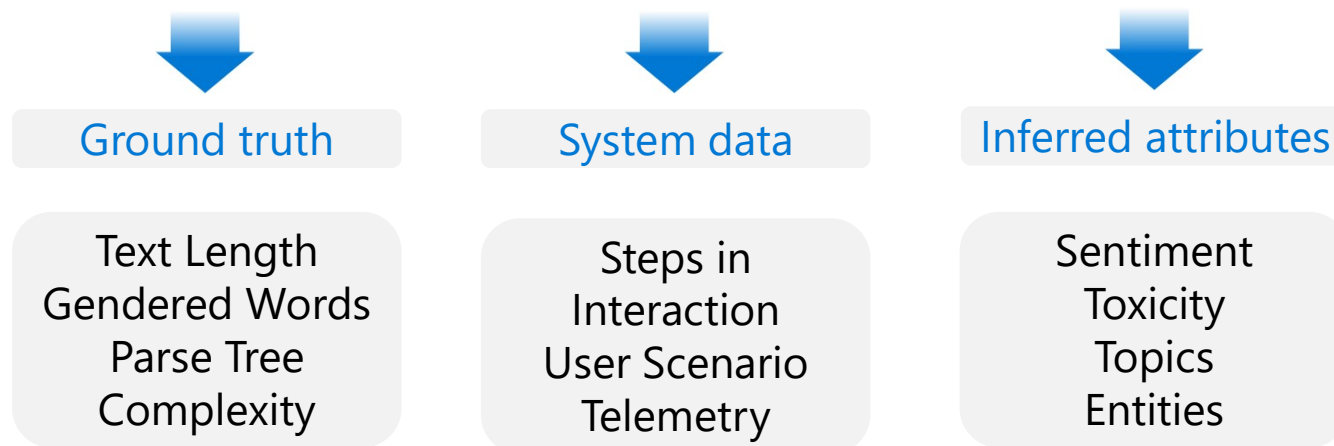
Brightness, Noise  
Objects  
Auto Captions

# Meta-data for cohort design in Language

**ACM Conference on Fairness, Accountability, and Transparency** (ACM FAccT, formerly known as ACM FAT\*) is a peer-reviewed academic conference series about ethics and computing systems.<sup>[1]</sup>

Sponsored by the [Association for Computing Machinery](#), this conference focuses on issues such as [algorithmic transparency](#), [fairness in machine learning](#), [bias](#), and [ethics](#) from a multi-disciplinary perspective. The conference community includes computer scientists, statisticians, social scientists, scholars of law, and others.<sup>[2]</sup>

The conference is sponsored by [Big Tech](#) companies such as [Facebook](#), [Twitter](#), and [Google](#), and large foundations such as the [Rockefeller Foundation](#), [Ford Foundation](#), [MacArthur Foundation](#), and [Luminate](#).<sup>[3]</sup> Sponsors contribute to a general fund (no "earmarked" contributions are allowed) and have no say in the selection, substance, or structure of the conference.<sup>[4]</sup>

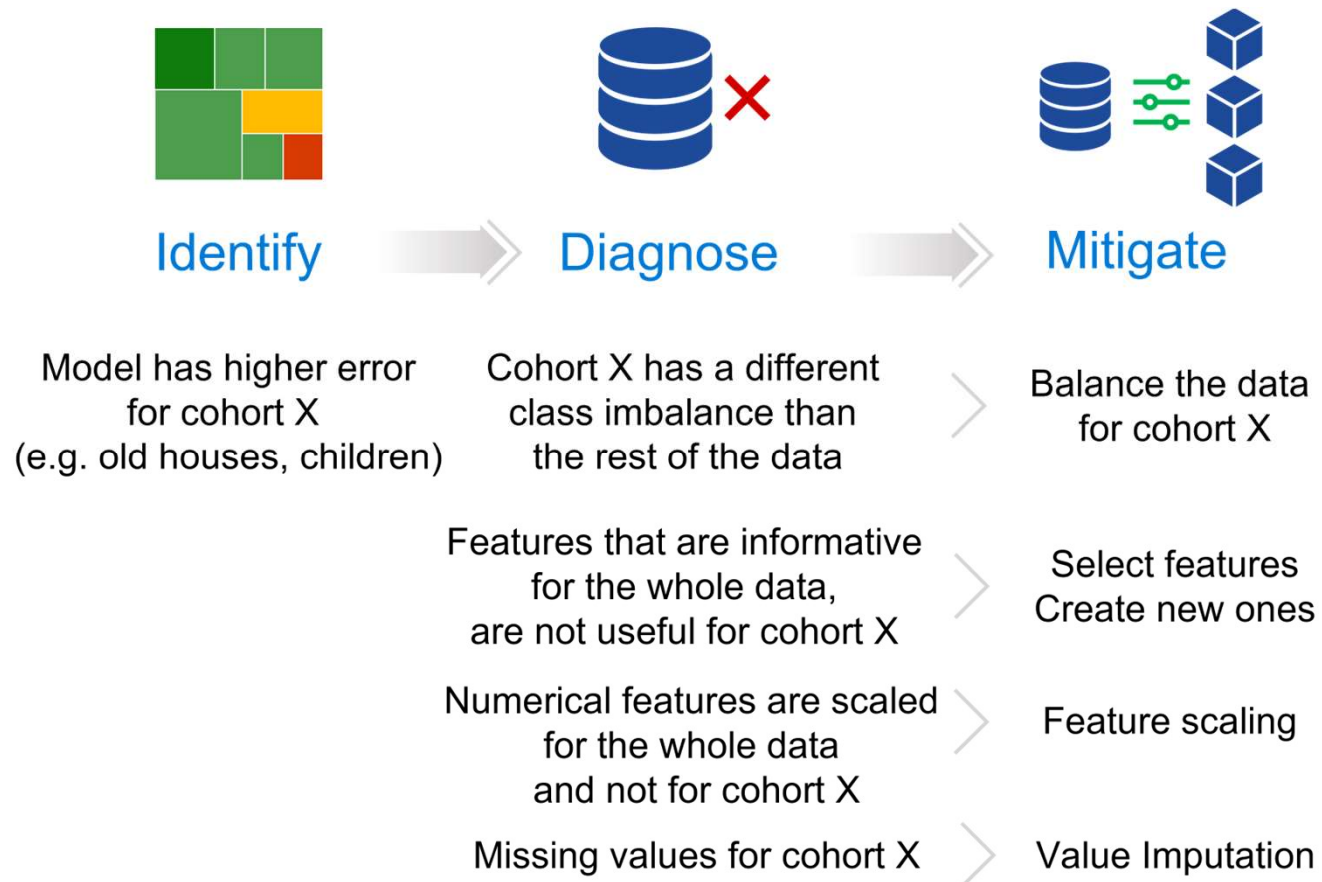




# Responsible AI Mitigations and Tracker

# Responsible AI Mitigations

`pip install raimitigations`



# An overview

```
pip install raimitigations
```

---

A rich set of mitigations focusing on **data quality** as it relates to the quality of ML models.

---

A simple interface for mitigation steps that follows the **.fit()** and **.transform()** convention.

---

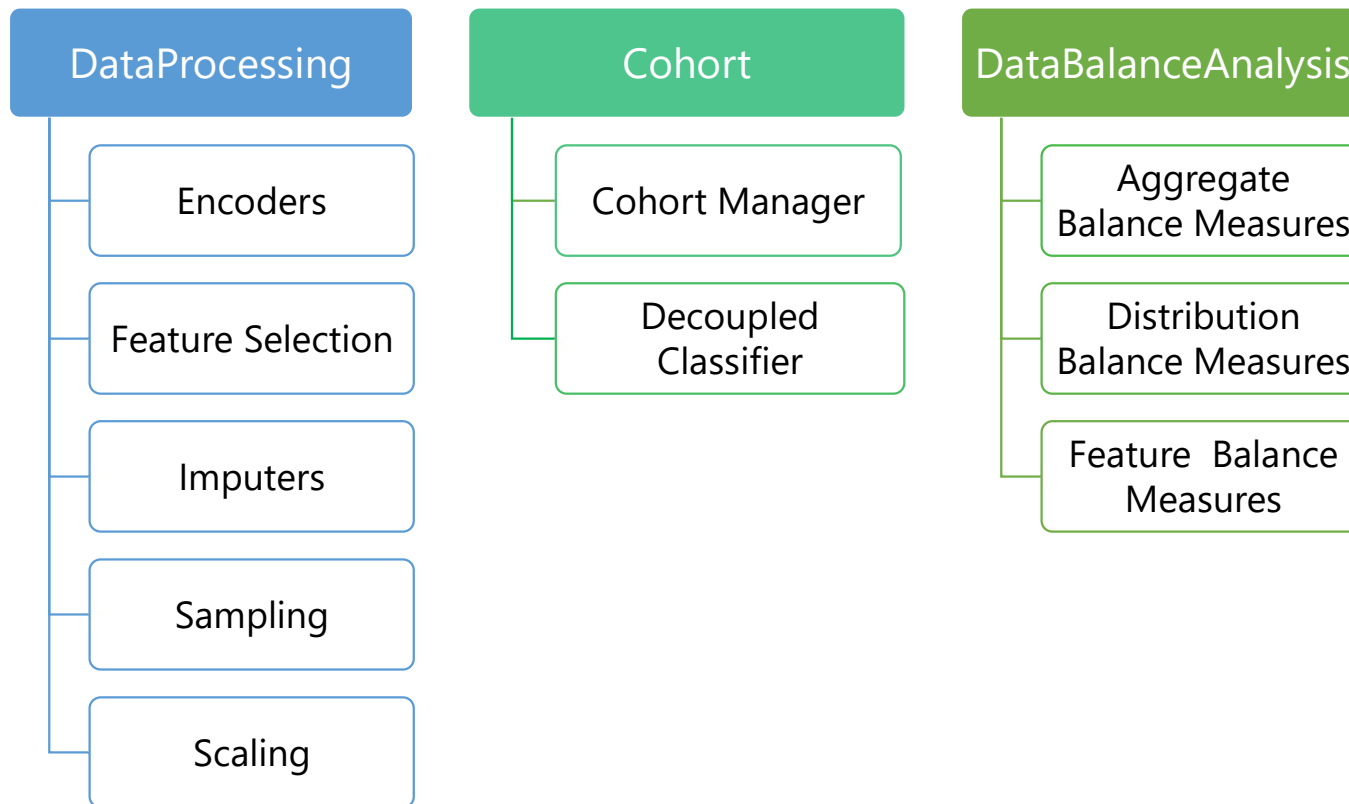
Function calls adapted for responsible AI by extending existing calls either with **target features or cohorts**.

---

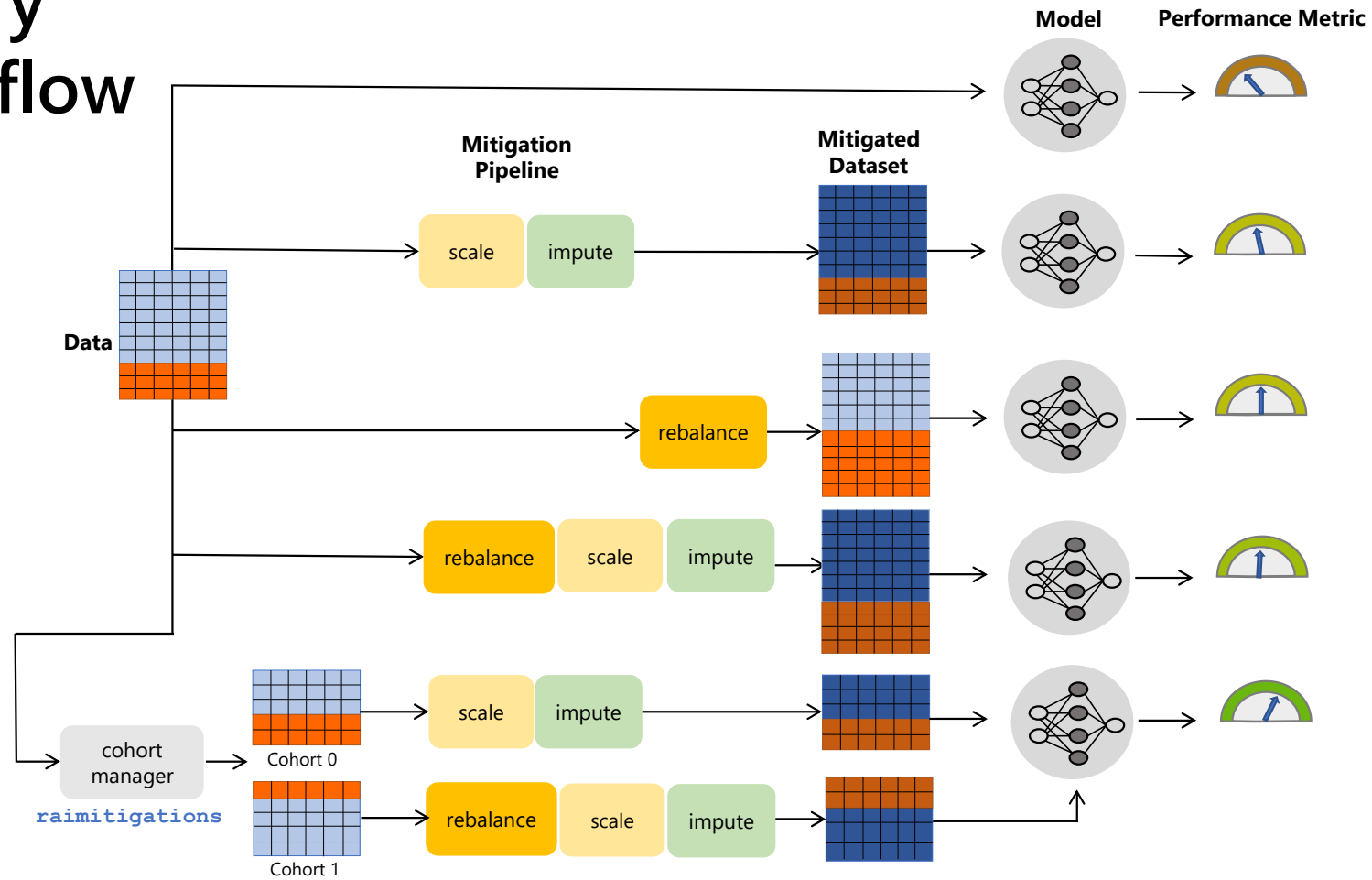
Possible to create **different models for different cohorts**, or **post process** predictions for improving predictions in a cohort..

# Library Components

<https://github.com/microsoft/responsible-ai-toolbox-mitigations>



# Library Workflow

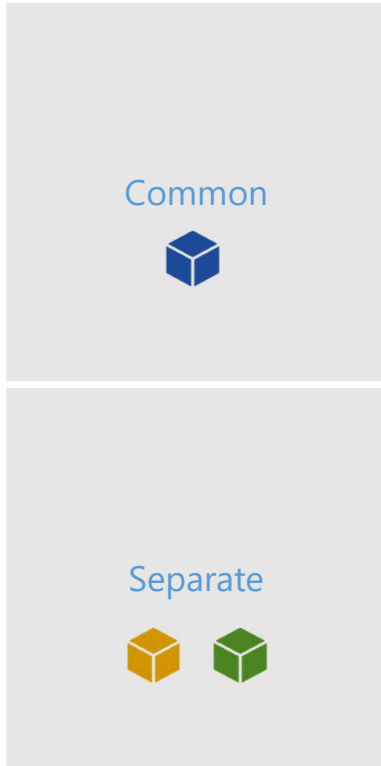




# Targeted Mitigations



Model training



Data mitigation strategy		
Common	Separate Same type	Separate Different types
<p><b>Blanket mitigation</b></p> <p>Applies the same mitigation type to all cohorts and uses all data as context.</p> <p>Trains a single model for all cohorts.</p>	<p><b>Targeted mitigation</b></p> <p>Applies the same mitigation type to all cohorts but uses only the cohort data as context.</p> <p>Trains a single model for all cohorts.</p>	<p><b>Targeted mitigation</b></p> <p>Applies different mitigation types to different cohorts and uses only the cohort data as context.</p> <p>Trains a single model for all cohorts.</p>
<p><b>Targeted mitigation</b></p> <p>Applies the same mitigation type to all cohorts and uses all data as context.</p> <p>Trains different models for different cohorts</p>	<p><b>Targeted mitigation</b></p> <p>Applies the same mitigation type to all cohorts but uses only the cohort data as context.</p> <p>Trains different models for different cohorts.</p>	<p><b>Targeted mitigation</b></p> <p>Applies different mitigation types to different cohorts and uses only the cohort data as context.</p> <p>Trains different models for different cohorts.</p>

# Responsible AI Tracker

<https://github.com/microsoft/responsible-ai-toolbox-tracker>



---

Managing and linking model improvement artefacts for cleaner data-science practices: **code, models, visualizations, data.**

---

Disaggregated model evaluation and comparison, for tracking both **performance improvements and declines.**

---

Initial integration with the **Responsible AI Mitigations library.**  
More to be done for e2e model improvement.

---

Initial integration with **mlflow.**

### Responsible AI Tracker

Adult Census Income

Notebook	Model	Accuracy
5 estimators.ipynb	✓	0.789
balance all data.ipynb	✓	0.827
target balance per ...ipynb	✓	0.849
balance per cohort ...ipynb	✓	0.793

Code

Models

Compare models

Compare Models
balance all data.ipynb
target balance per cohort.ipynb
balance per cohort both.ipynb

Notebooks

Metrics

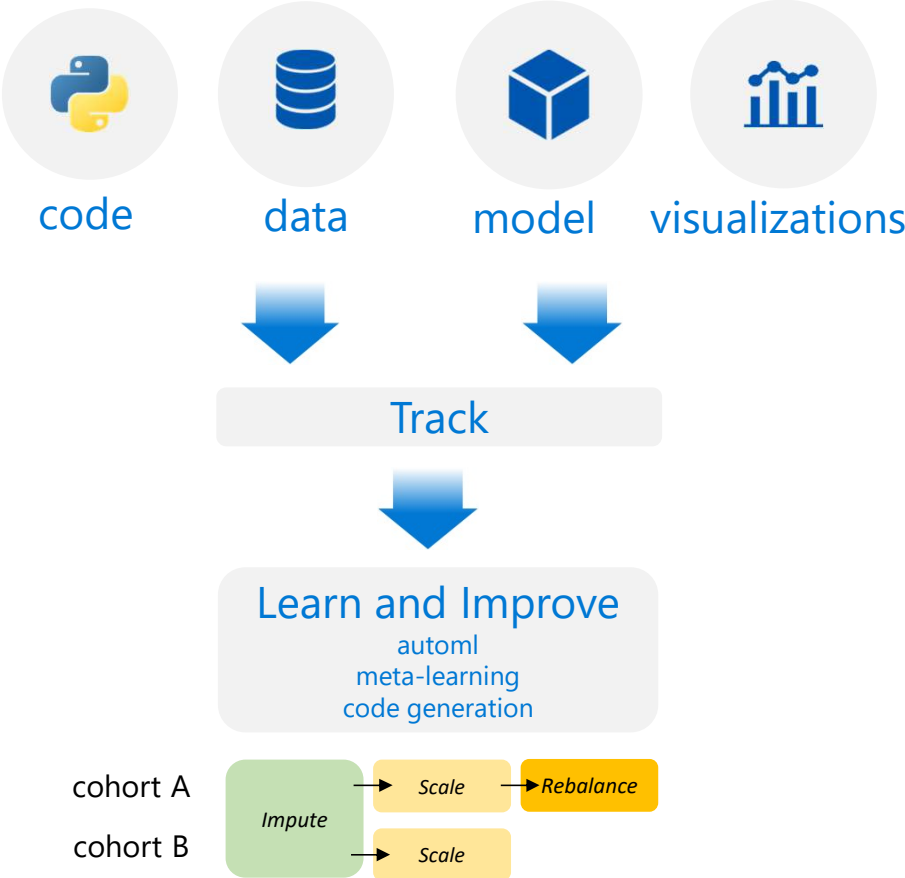
Cohorts

Visualization reports

Visual display
  Absolute
  Comparative

Notebook	Cohort	Accuracy	Precision	Recall	F1 Score	Log Loss	ROC AUC
		-0.119 0.119	-0.73 0.73	-0.683 0.683	-0.443 0.443	-0.234 0.234	-0.046 0.046
5 estimators.ipynb <span style="font-size: small;">baseline</span>	All adult-test-sampl...csv	0.789	1	0.134	0.237	0.433	0.892
	Married	0.611	1	0.149	0.259	0.635	0.797
	Not married	0.936	1	0.054	0.102	0.266	0.836
balance all da...ipynb	All adult-test-sampl...csv	0.827 (0.038) ↑	0.626 (0.374) ↓	0.727 (0.593) ↑	0.673 (0.436) ↑	0.511 (0.078) ↑	0.889 (0.003) ↓
	Married	0.683 (0.072) ↑	0.616 (0.384) ↓	0.815 (0.666) ↑	0.702 (0.443) ↑	0.62 (0.015) ↓	0.784 (0.013) ↓
	Not married	0.947 (0.011) ↑	0.936 (0.064) ↓	0.237 (0.183) ↑	0.378 (0.276) ↑	0.421 (0.155) ↑	0.837 (0.001) ↑
target balance...ipynb	All adult-test-sampl...csv	0.849 (0.06) ↑	0.813 (0.187) ↓	0.495 (0.361) ↑	0.615 (0.378) ↑	0.455 (0.022) ↑	0.888 (0.004) ↓
	Married	0.73 (0.119) ↑	0.803 (0.197) ↓	0.542 (0.393) ↑	0.647 (0.388) ↑	0.594 (0.041) ↓	0.792 (0.005) ↓
	Not married	0.947 (0.011) ↑	0.977 (0.023) ↓	0.231 (0.177) ↑	0.374 (0.272) ↑	0.341 (0.075) ↑	0.81 (0.026) ↓
balance per co...ipynb	All adult-test-sampl...csv	0.793 (0.004) ↑	0.563 (0.437) ↓	0.674 (0.54) ↑	0.614 (0.377) ↑	0.544 (0.111) ↑	0.846 (0.046) ↓
	Married	0.728 (0.117) ↑	0.72 (0.28) ↓	0.663 (0.514) ↑	0.69 (0.431) ↑	0.596 (0.039) ↓	0.79 (0.007) ↓
	Not married	0.846 (0.09) ↓	0.27 (0.73) ↓	0.737 (0.683) ↑	0.395 (0.293) ↑	0.5 (0.234) ↑	0.876 (0.04) ↑

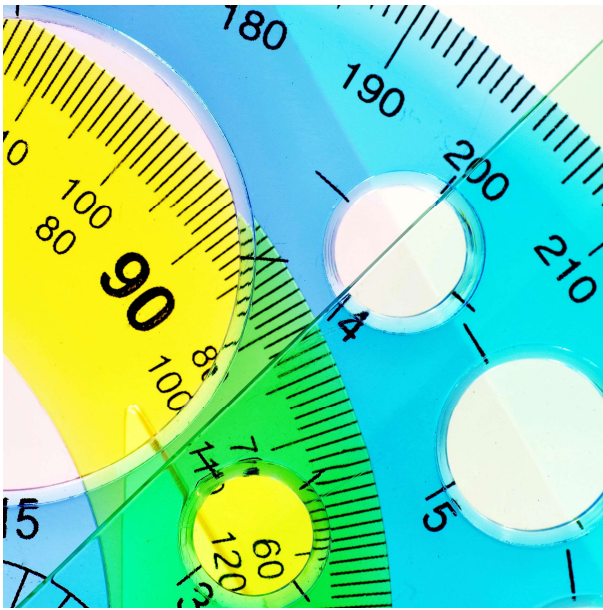
# Learning to mitigate for Responsible AI





# User Insights, Challenges, and Opportunities

# Responsible AI as an open-source opportunity



Transparency



Research & Education



Integration with OSS frameworks

# Adoption Challenges



Disaggregated evaluation, reliability and ML criticality



Choosing the right metrics (domain expertise)



Integration of RAI, ML tools with other tools in the ML Lifecycle



Wide range of ML expertise and problem domains



Responsible AI pre- and post-production



# Insights - What works?



Co-design with users/customers



Vertical solutions (e.g. Responsible AI for Healthcare)



Customization and flexibility (metrics, components)



Transparency, reproducibility, reusability of evaluation pipelines



Processes, Culture, Education – beyond tools





# RESPONSIBLE AI MATURITY MODEL

MAPPING YOUR ORGANIZATION'S GOALS ON THE PATH TO RESPONSIBLE AI



MIHAELA VORVOREANU ■ AMY HEGER ■ SAMIR PASSI ■ SHIPI DHANORKAR ■ ZOE KAHN ■ RUOTONG WANG

AETHER CENTRAL UX RESEARCH & EDUCATION ■ MICROSOFT

V1 ■ MAY 17, 2023

<https://aka.ms/raimm>

# Stay tuned

-  <https://github.com/microsoft/responsible-ai-toolbox>
- <https://github.com/microsoft/responsible-ai-toolbox-mitigations>
- <https://github.com/microsoft/responsible-ai-toolbox-tracker>



Extend the Responsible AI Dashboard for Generative AI



More functionality around model comparison and monitoring



Scalability investments and distributed mitigations



Learning to mitigate for Responsible AI

# Useful links

## **Responsible AI Toolbox**

<https://github.com/microsoft/responsible-ai-toolbox>

## **Responsible AI Tracker**

<https://github.com/microsoft/responsible-ai-toolbox-tracker>

## **Responsible AI Mitigations**

<https://github.com/microsoft/responsible-ai-toolbox-mitigations>

## **Responsible AI: The research collaboration behind new open-source tools offered by Microsoft**

<https://www.microsoft.com/en-us/research/blog/responsible-ai-the-research-collaboration-behind-new-open-source-tools-offered-by-microsoft/>

## **Responsible AI Dashboard Deep Dive Blogs**

Responsible AI dashboard: A one-stop shop for operationalizing Responsible AI in practice: [Tech Community blog](#)

Responsible AI Dashboard in Azure Machine Learning: [Tech Community blog](#)

Debug Object Detection Models with the Responsible AI Dashboard: [Tech Community blog](#)

## **Responsible AI Mitigations and Tracker: New open-source tools for guiding mitigations in Responsible AI**

<aka.ms/rai-mitigationstracker-blog>



Questions?

[rai-toolbox@microsoft.com](mailto:rai-toolbox@microsoft.com)