



**ThaiPy - Bangkok Python
Meetup**

THU, MAR 11, 6:45 PM GMT+7

Responsible AI toolkits

SeokJin Han
Microsoft

linkedin.com/in/seokjinhan
sehan@microsoft.com

Agenda

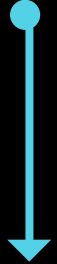
- What to consider to deal with Responsible AI
- Why
- What are the approaches
- Some demos and quizzes: Interpret ML, Error Analysis
- Resources
- Join the Community!

Get Hands Dirty

- <https://github.com/dem108/explain-ml-models>
 - Try git clone and follow README.md to get started

Progression of Responsible AI

Principles



Fairness

Reliability & Safety

Inclusiveness

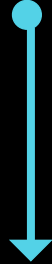
Privacy & Security

Transparency

Accountability



Practices



AETHER committee

The Partnership on AI

Guidelines for
Human-AI Design

Guidelines for
Conversational AI



Tools



Homomorphic Encryption

Interpret ML

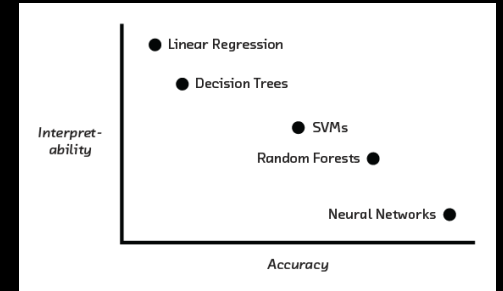
Differential Privacy

Data Drift

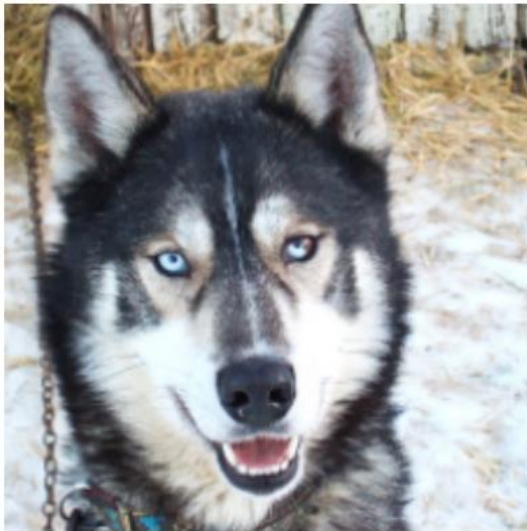
Secure MPC

Why this matters?

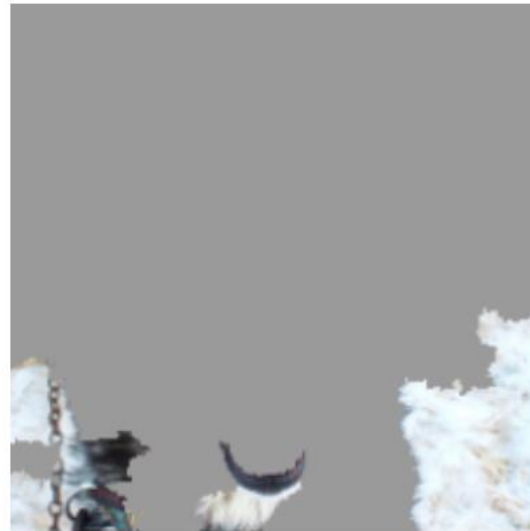
- “If you can't explain it simply, you don't understand it well enough.” — Albert Einstein
- Accuracy (as a whole) is not enough
- Is my model predicting well, or am I totally mistaken?



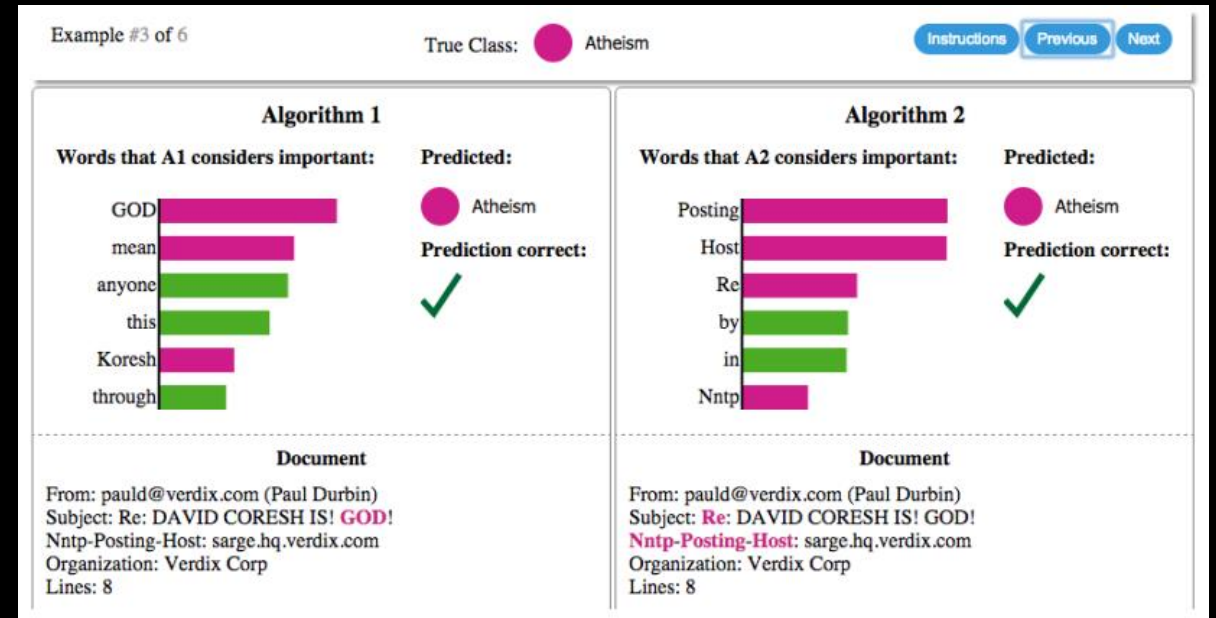
Dare to explain?



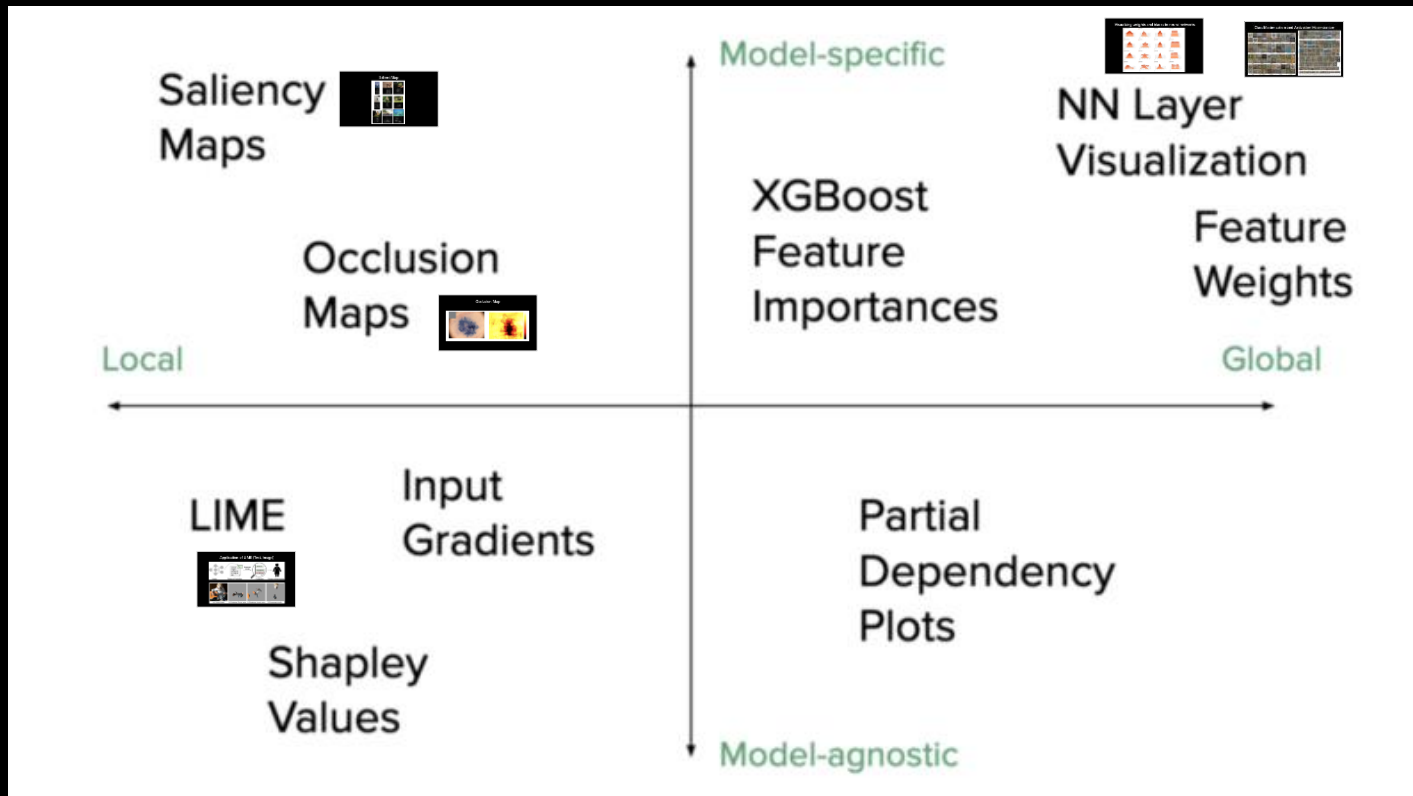
(a) Husky classified as wolf



(b) Explanation



Efforts to explain models



X-axis	Local	Explain locally (individual filters in NN etc)
	Global	Explain globally (visualize weights and biases in NN etc)
Y-axis	Model-agnostic	Can apply to many model types
	Model-specific	Works for specific model types

4 Interpretable Models

- 4.1 Linear Regression
- 4.2 Logistic Regression
- 4.3 GLM, GAM and more
- 4.4 Decision Tree
- 4.5 Decision Rules
- 4.6 RuleFit
- 4.7 Other Interpretable Models

5 Model-Agnostic Methods

- 5.1 Partial Dependence Plot (PDP)
- 5.2 Individual Conditional Expectation (ICE)
- 5.3 Accumulated Local Effects (ALE)
- 5.4 Feature Interaction
- 5.5 Permutation Feature Importance
- 5.6 Global Surrogate
- 5.7 Local Surrogate (LIME)
- 5.8 Scoped Rules (Anchors)
- 5.9 Shapley Values
- 5.10 SHAP (SHapley Additive exPlanation)

Explaining Black Box Models and Datasets

- Anchor** - An open-source bias audit toolkit for data scientists, machine learning researchers, and practitioners to audit for machine learning models for discrimination and bias to make informed and equitable decisions around developing and deploying predictive risk assessment tools.
- Alibi** - An open source Python library aimed at machine learning model inspection and interpretation. The initial focus of the library is on black-box, instance-based model explanations.
- anchors** - Code for the paper "High precision model agnostic explanations", a model-agnostic system that explains the behavior of complex models with high-precision rules called anchors.
- captum** - model interpretability and understanding library for PyTorch developed by Facebook. It contains general purpose implementations of integrated gradients, saliency maps, smoothgrad, vargrad and others for PyTorch models.
- class** - Example of using classifier-agnostic saliency map extraction on ImageNet presented in the paper "Classifier-agnostic saliency map extraction".
- ContrastiveExplanation** (old "Tree") - Python script for model agnostic contrastive/counterfactual explanations for machine learning. Accompanying code for the paper "Contrastive Explanations with Local Top Trees".
- DeepLIFT** - Codebase that contains the methods in the paper "Learning important features through propagating activation differences". Here is the slide and the video of the 15 minute talk given at ICML.
- Deepviz** - This is the code required to run the Deep Visualization ToolBox, as well as to generate the session-by-session visualizations using integrated optimization. The toolbox and methods are described casually here and more formally in this paper.
- eli5** - "Explain Like I'm 5" is a Python package which helps to debug machine learning classifiers and explains their predictions.
- FACETS** - Facets contains two robust visualizations to aid in understanding and analyzing machine learning datasets. Get a sense of the shape of each feature of your dataset using Facets Overview, or explain individual observations using Facets Dive.
- feature** - feature is a python toolkit to assess and mitigate unfairness in machine learning models.
- feature** - feature is a python toolbox auditing the machine learning models for bias.
- feature** - This repository is meant to facilitate the benchmarking of fairness aware machine learning algorithms based on this paper.
- GEU** - Global Explanations for Bias Identification - An attention-based constructed post-hoc explanations for detection and identification of bias in data. We propose a global explanation and introduce a step-by-step framework on how to detect and test bias. Python package for image data.
- IBM AI Explainability 360** - responsibility and explainability of data and machine learning models including a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics.
- IBM AI Fairness 360** - A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models.
- INtelligence** - An open-source library for analyzing kernel models visually by methods such as DeepSaliency-Disaggregation, FeatureSaliency Maps, and Integrated Gradients.
- Integrated-Gradients** - This repository provides code for implementing integrated gradients for networks with image inputs.
- interpret** - interpret is an open-source package for training interpretable models and explaining black-box systems.
- kernelviz** - kernelviz is a high-level toolkit for visualizing and debugging your trained kernel neural net models. Currently supported visualizations include Activation Maximization, Saliency Maps, Class Activation Maps.
- L3** - Code for replicating the experiments in the paper "Learning to Explain: An Information-Theoretic Perspective on Model Interpretation" at ICML 2016.
- Lightwood** - A Pytorch based framework that breaks down machine learning problems into smaller blocks that can be glued together iteratively with an objective to build predictive models with low line of code.
- lime** - Local Interpretable Model-agnostic Explanations for machine learning models.
- LM30 Importance** - LM30 (Local Model Feature Importance) calculates the importance of a set of features based on a metric of choice, for a model of choice, by iteratively removing each feature from the set, and evaluating the performance of the model, with a validation scheme of choice, based on the chosen metric.
- MindDB** - MindDB is an Explainable AI toolkit, framework for developers. With MindDB you can build, train and use state of the art ML models in as simple as one line of code.
- ml-explainability** - An Automated Machine Learning (AutoML) python package for tabular data. It can handle Binary Classification, Multi-Class Classification and Regression. It provides feature engineering, explanations and model metrics reports.
- NETRICH** - solves for neural network, deep learning and machine learning models.
- pyModelDash** - A model agnostic tool for decomposition of predictions from black boxes. Break Down Table shows contributions of every variable to a final prediction.
- robustness** - Code to implement learning robustness based predictions with code for paper "Robustness: Neural Predictions".
- responsibility** - Toolkit for auditing and mitigating bias and fairness of machine learning systems.
- SHAP** - SHapley Additive exPlanations is a unified approach to explain the output of any machine learning model.
- skater** - skater is a unified framework to enable Model Interpretation for all forms of model to help one build an interpretable machine learning system often needed for real world use-cases.
- TensorBoard's TensorBoard WhatIf** - TensorBoard screen to analyze the interactions between inference results and data inputs.
- TensorFlow's adversarial** - An adversarial example library for constructing attacks, building defenses, and benchmarking both. A python library to benchmark systems vulnerability to adversarial examples.
- tensorflow's back** - back is a collection of infrastructure and tools for research in neural network interpretability.
- tensorflow's Model Analysis** - TensorFlow Model Analysis (TMA) is a library for evaluating TensorFlow models. It allows users to evaluate their models on large amounts of data in a distributed manner, using the same metrics defined in their trainer.
- thorn** - thorn is a python library built on top of pandas and sklearn that implements fairness-aware machine learning algorithms.
- Thorn** - Thorn is a model-based approach for measuring discrimination in a software system.
- Treeinterpreter** - Package for interpreting scikit-learn's decision tree and random forest predictions. Allows decomposing each prediction into bias and feature contribution components as described in <http://blog.dominicam.io/interpreting-random-forests/>.
- vis** - vis is a python library for visualizing machine learning models.
- vis** - vis is a python library for visualizing machine learning models.

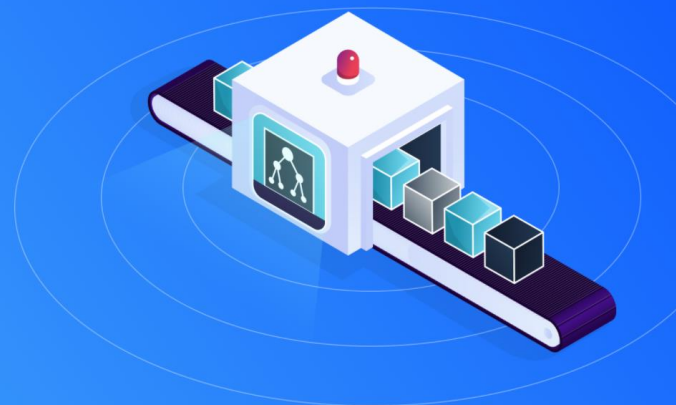
Source: <https://github.com/Harvard-IACS/2020-ComputeFest/>, <https://christophm.github.io/interpretable-ml-book/>, <https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets>

Understand Models. Build Responsibly.

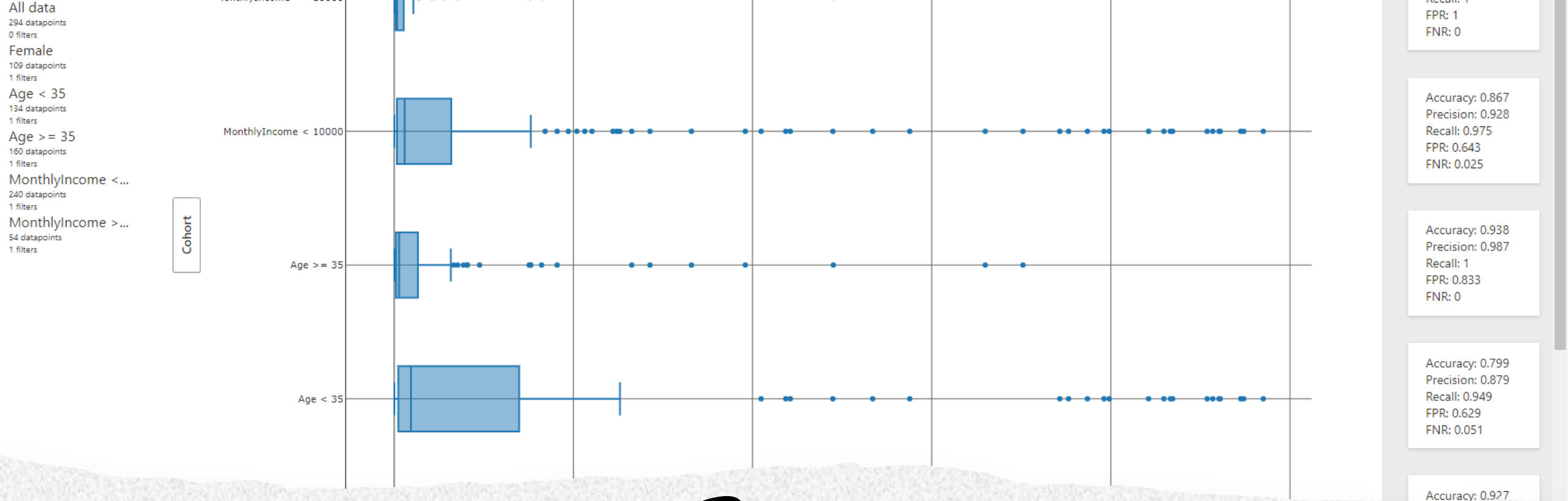
A toolkit to help understand models and enable responsible machine learning

[Get Started](#)

[Learn More](#)

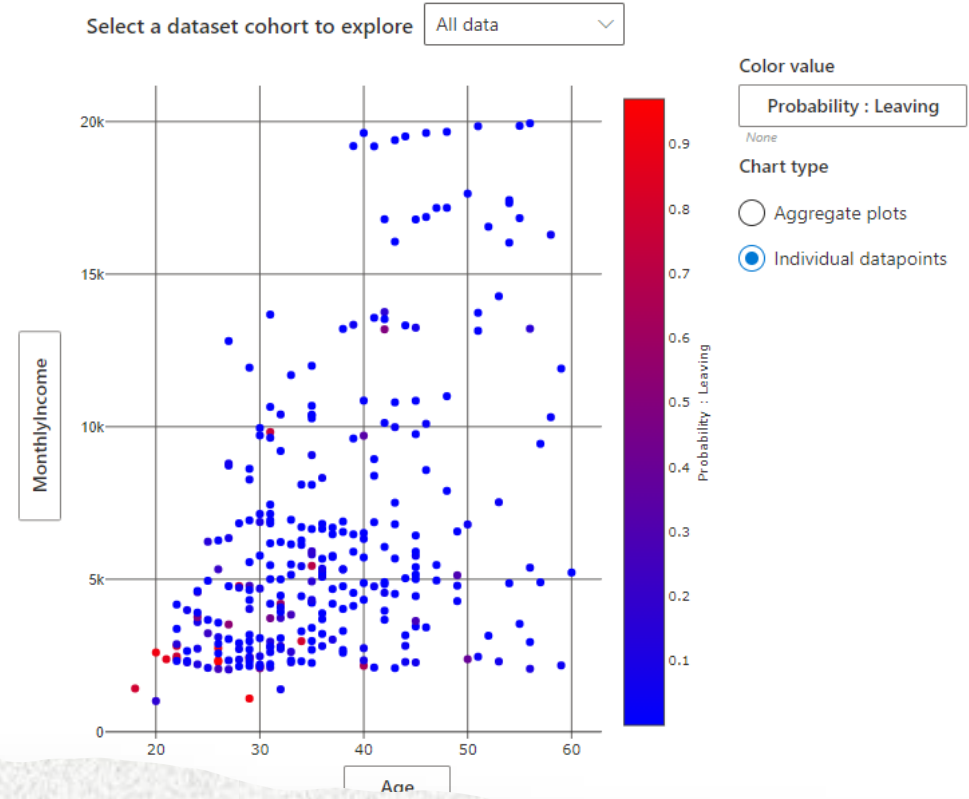
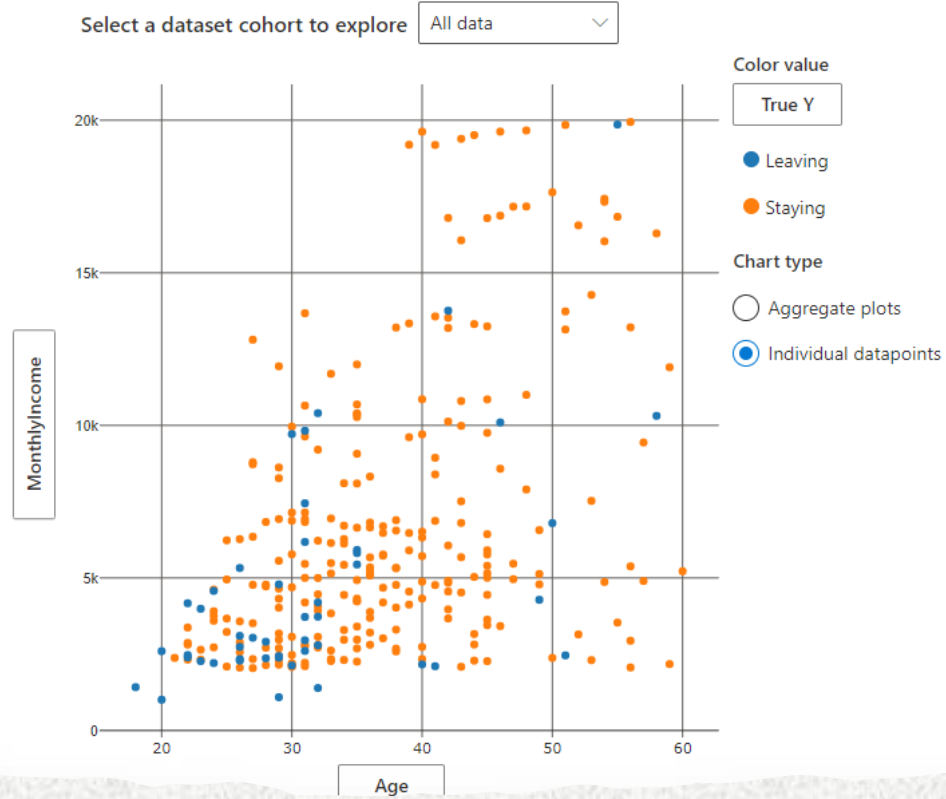


InterpretML



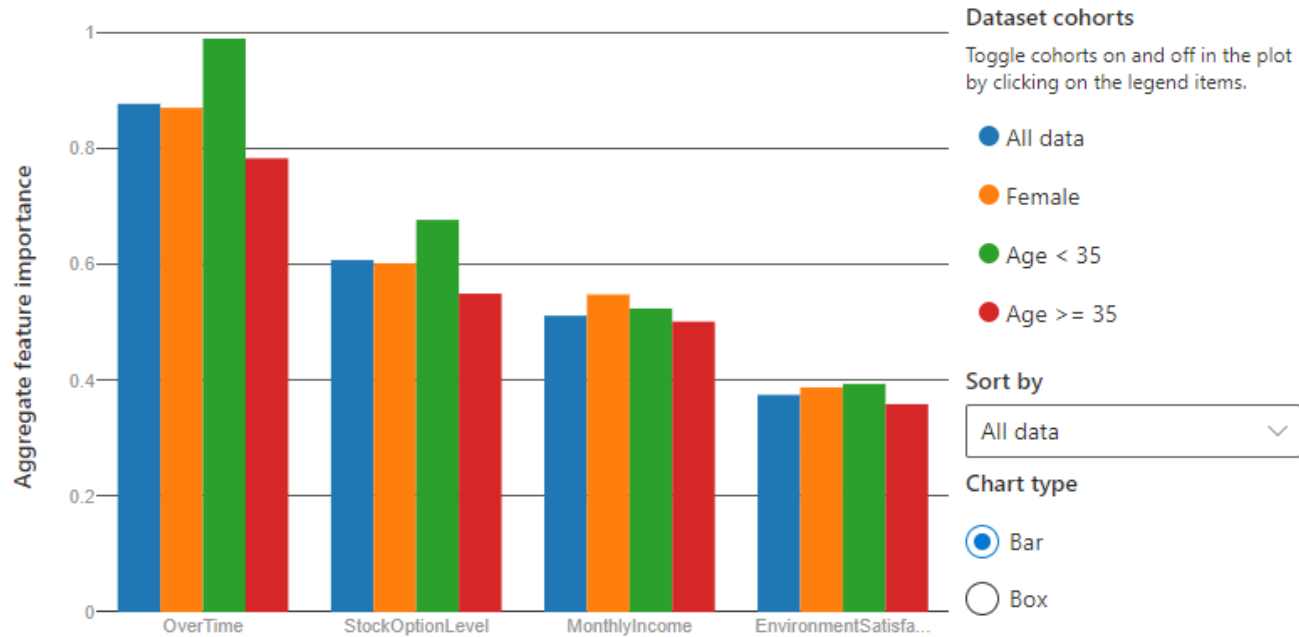
Does Model perform differently per Cohort?

** Disclaimer: Insights from these should be validated with different aspects*

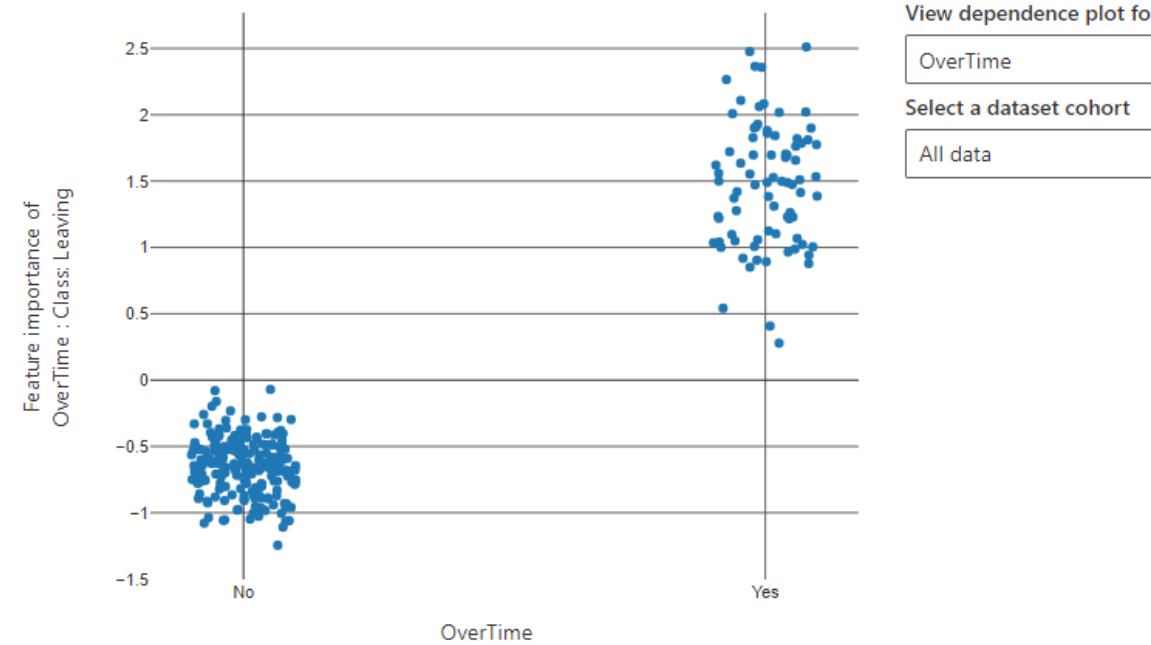


Are they making more money (on average) as they grow older?
When are they actually leaving?
When are they predicted to leave?
Does it change due to Gender?

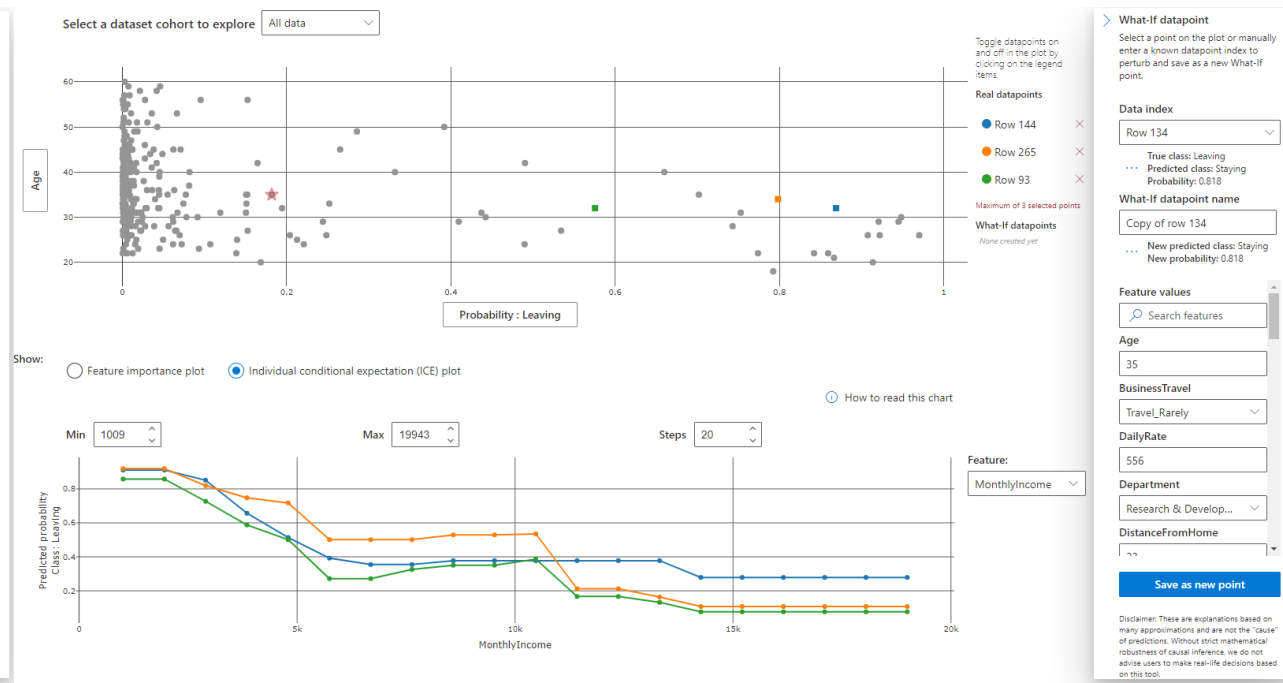
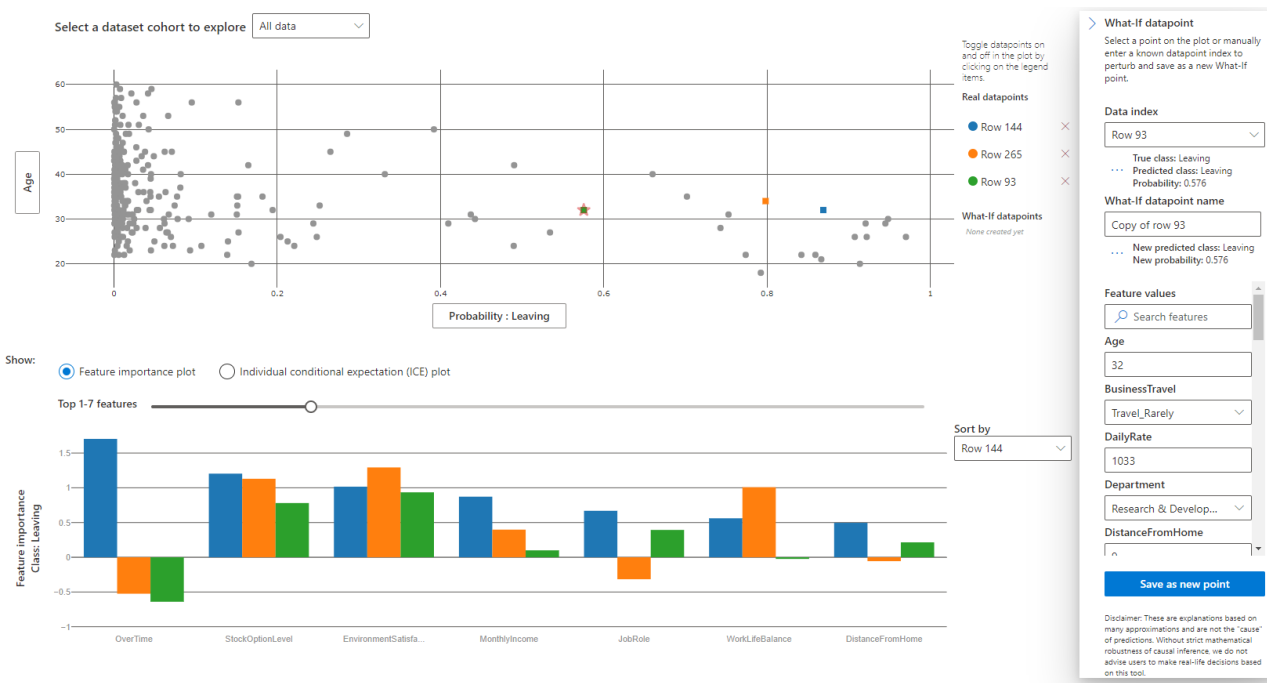
What do these explanations mean?



How to read this chart



What are the most important features?
How much does Over Time affect attrition?
What about Stock Option Level, Monthly Income?
Does Gender, Age make a difference?



Take 3 employees with similar age, can we tell what is the most importance feature for each person according to the model?

What is the impact of Age, Environment Satisfaction, Monthly Income on the probability of Leaving, and how does the impact differ per person?

Error Analysis

Identify & Diagnose Errors
Build Responsibly

A toolkit to help analyze and improve model accuracy.

Identify

Diagnose

Identify cohorts with high error rate versus benchmark and visualize how the error rate distributes.

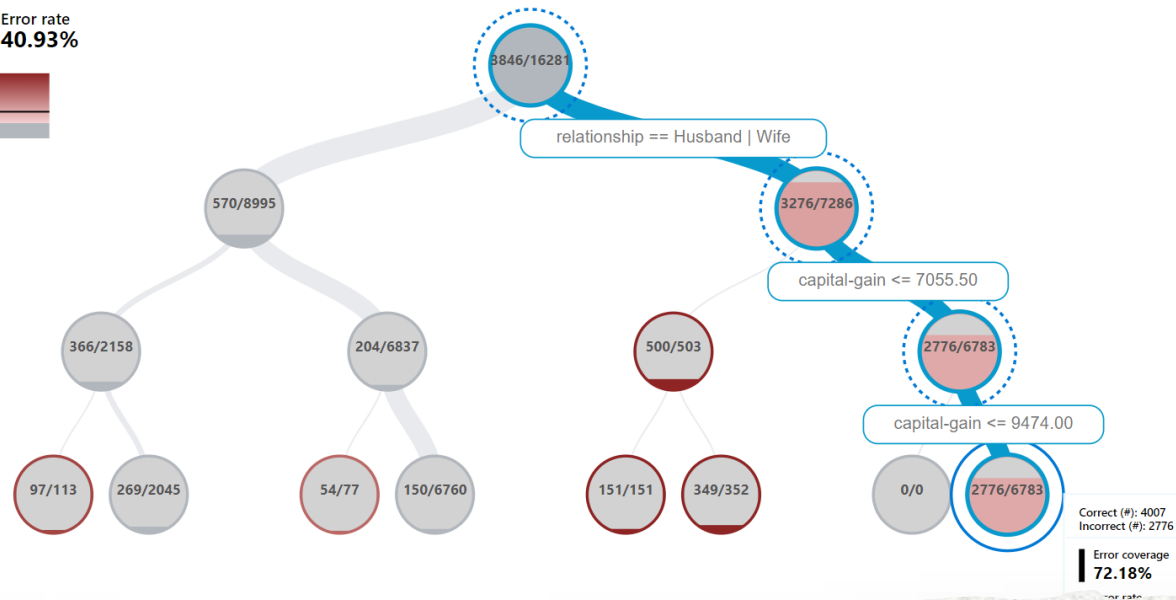
Error
Analysis

To find nodes in the tree map with the most errors, look for higher values and fuller circles.

Cohort: All data

Error coverage
72.18%

Error rate
40.93%



With the grid map you can focus on specific filters and combine error rates. Start with two dataset features to compare.

Cohort: All data

Cells
4

Error coverage
94.20%

Error rate
25.11%

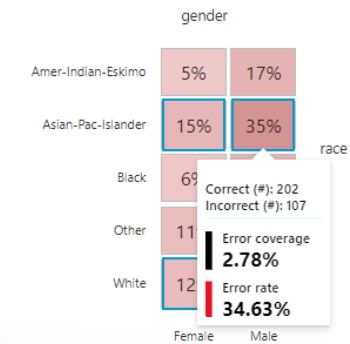


X-Axis: Feature 1

race

Y-Axis: Feature 2

gender



Which segments cover the most errors?

Which segments have the highest error rates?

Think fairness. Build for everyone.

A toolkit to assess and improve the fairness of machine learning models.

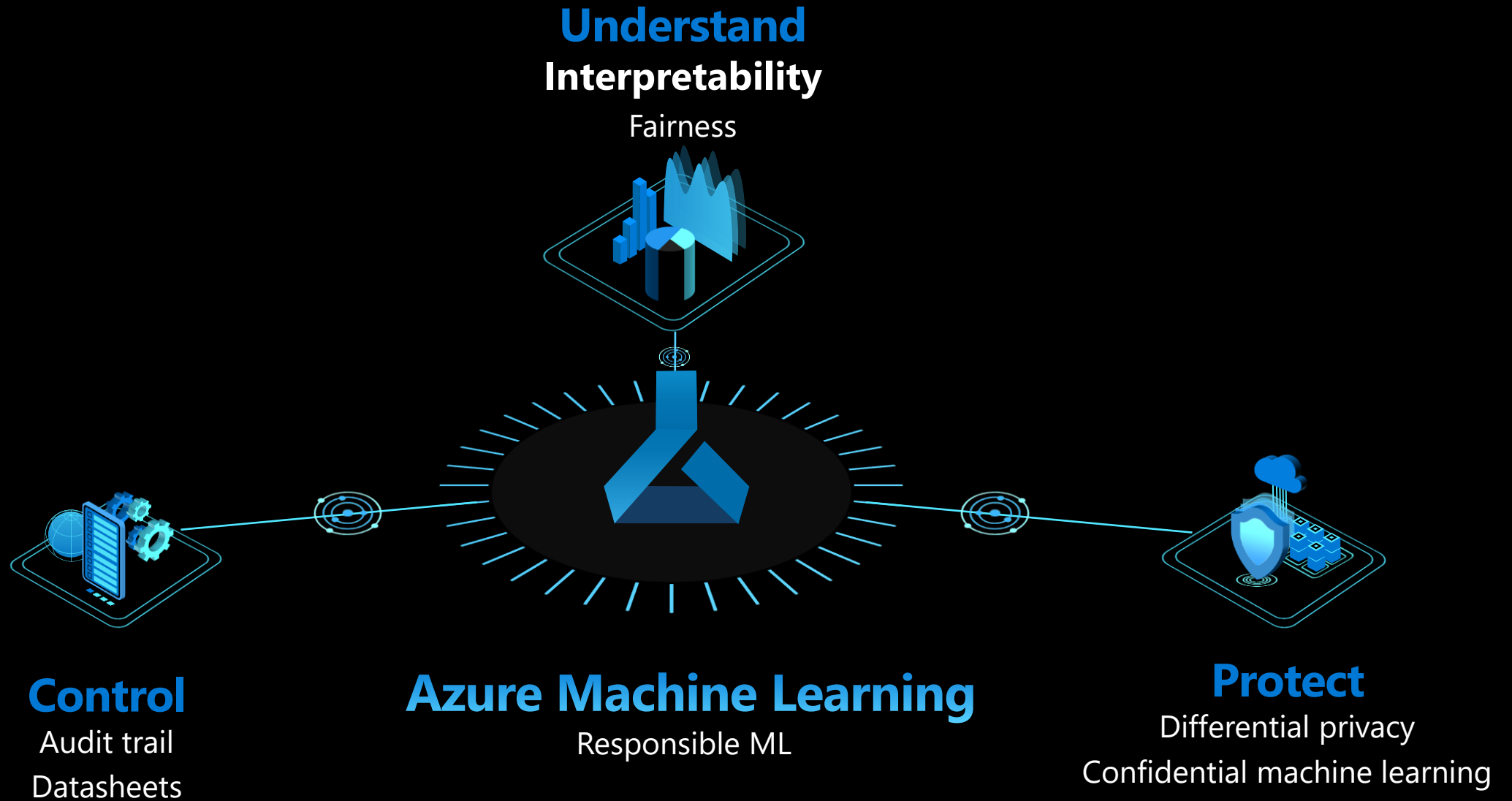
Assess

Mitigate

Use common **fairness metrics** and an **interactive dashboard** to assess which groups of people may be negatively impacted.

[Get Started](#)[API Docs](#)

Fairlearn



Codes, Blogs, Videos

- <https://github.com/dem108/explain-ml-models>
- [Enabling responsible AI development with new open-source capabilities](#)
- <https://github.com/microsoft/responsible-ai-widgets>
- [InterpretML](#)
- [Error Analysis](#)
- [Build Responsible AI using Error Analysis toolkit – YouTube](#)
- [Responsible Machine Learning with Error Analysis – Microsoft Blog](#)
- [Model interpretability in Azure Machine Learning](#)
- [Responsible AI at Microsoft](#)





Join our community!

Receive regular updates from our community experts and exclusive invites to our community events and workshops.



<https://aka.ms/MSSourceTH>