RSAConference2022
San Francisco & Digital | February 7 – 10

TRANSFORM

SESSION ID: MLAI-M06

It's Not Fair! Detecting Algorithmic Bias with Open Source Tools

Mike Kiser
Director of Strategy and Standards
SailPoint
@_MikeKiser

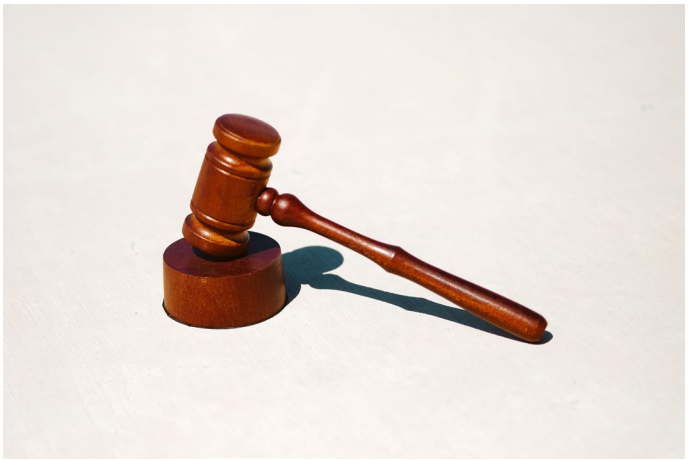
Mo Badawy
Principal Data Scientist
SailPoint

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.


TEMPLE OF JUSTICE



Transparency

Fairness

Transparency

Fairness

Why are bubbles round?

Consider a smooth surface in $\mathbb{R}^{n+1}$ representing the graph of a function $x_{n+1} = u(x_1, \ldots, x_n)$ defined on a bounded open set Ω in $\mathbb{R}^n$. Assuming that u is sufficiently smooth, the area of the surface is given by the nonlinear functional

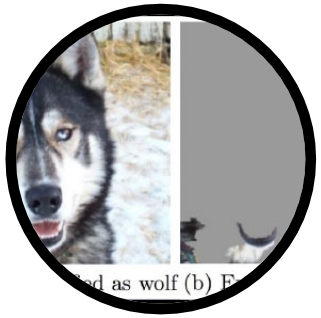
$$\mathcal{A}(u) = \int_\Omega \left(1 + |\nabla u|^2\right)^{1/2} \mathrm{d}x_1 \ldots \mathrm{d}x_n, \tag{1}$$

where ∇u is the gradient vector $(\partial u/\partial x_1, \ldots, \partial u/\partial x_n)$ and $|\nabla u|^2 = (\nabla u) \cdot (\nabla u)$.

Impact of Transparency

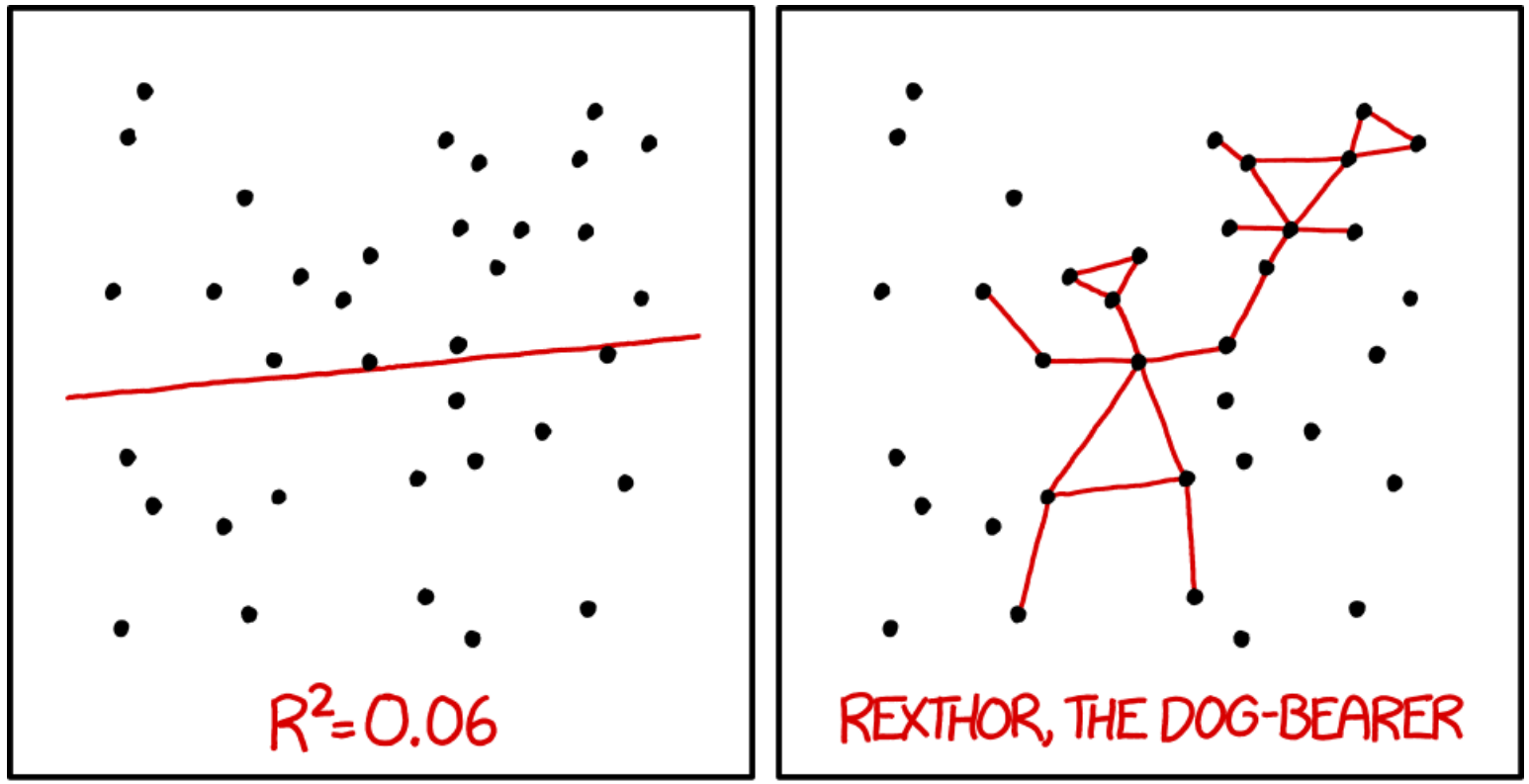
Safety

Model Adjustment



Objective Assessment

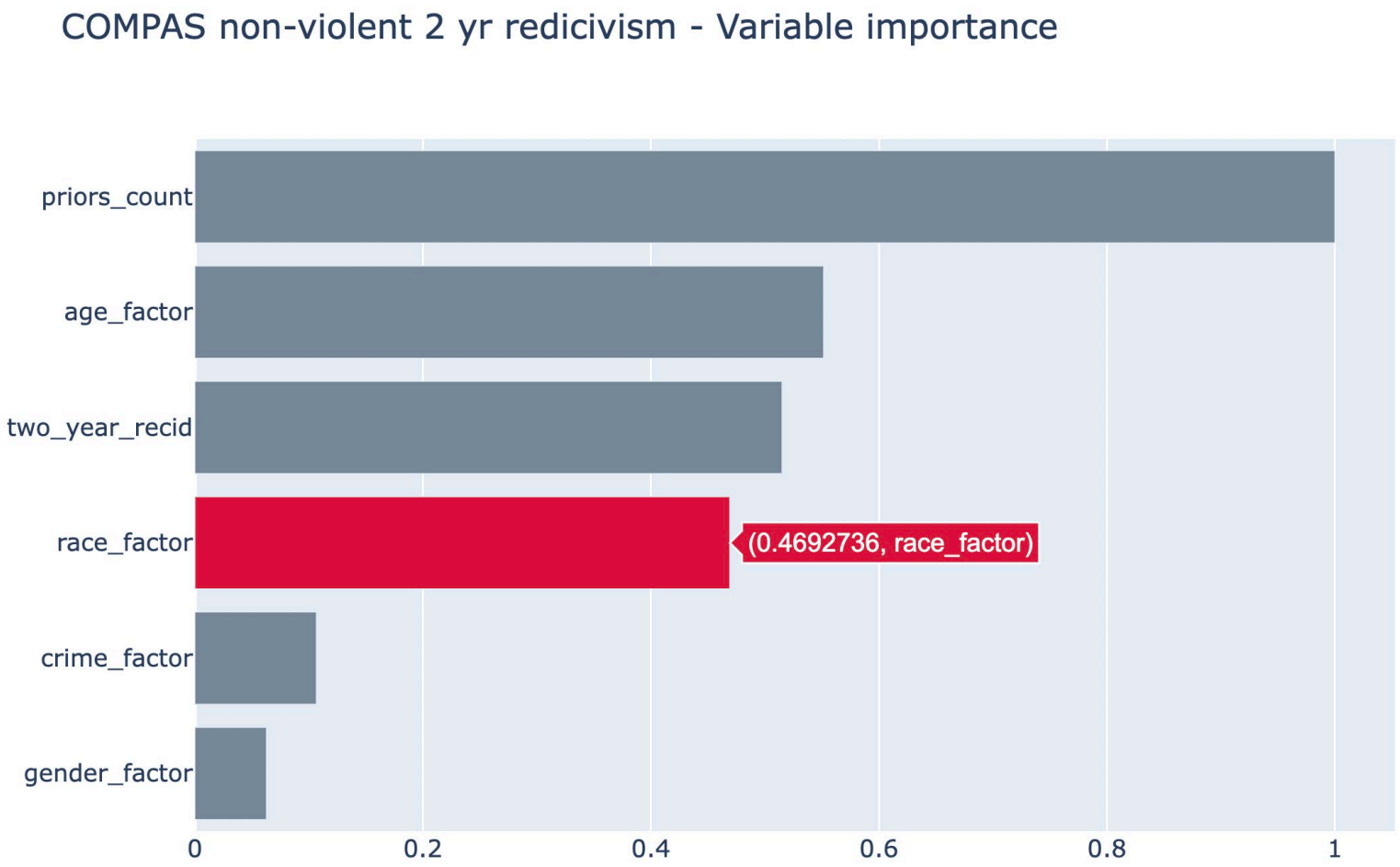
Model Interpretability



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

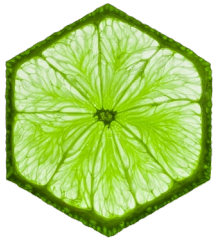
https://xkcd.com/1725/

COMPAS Model Interpretability



Model Agnostic Methods

- **Global vs Local**



- **LIME** - **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations

- **SHAP** - **SH**apley **A**dditive ex**P**lanations

Comparison of Transparency Tools

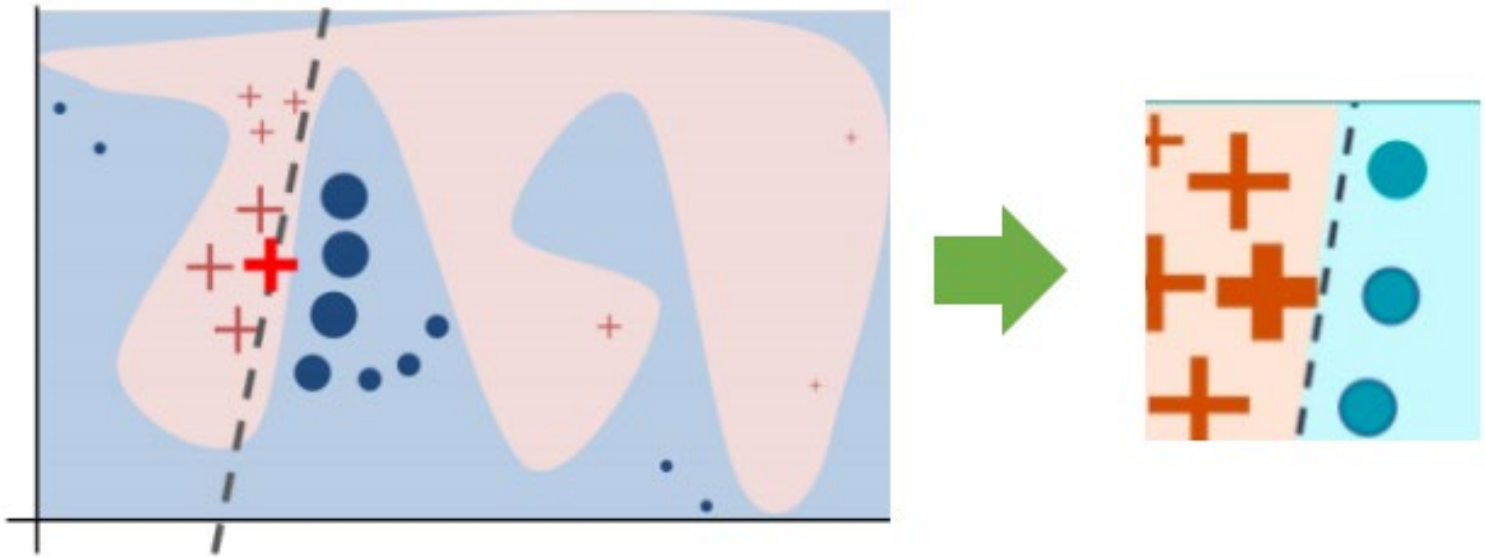
Local Interpretable Model-agnostic Explanations (LIME)

- Local Model Approximation
- Less Intuitive to Human Reasoning
- Relatively Fast

SHAP

- Game Theory
- More Intuitive to Human Reasoning
- Relatively Slow
  (Faster on tree structures, slow on k-nearest neighbor)

Global vs Local Interpretation



- **Global interpretation:** general landscape of features/attributes

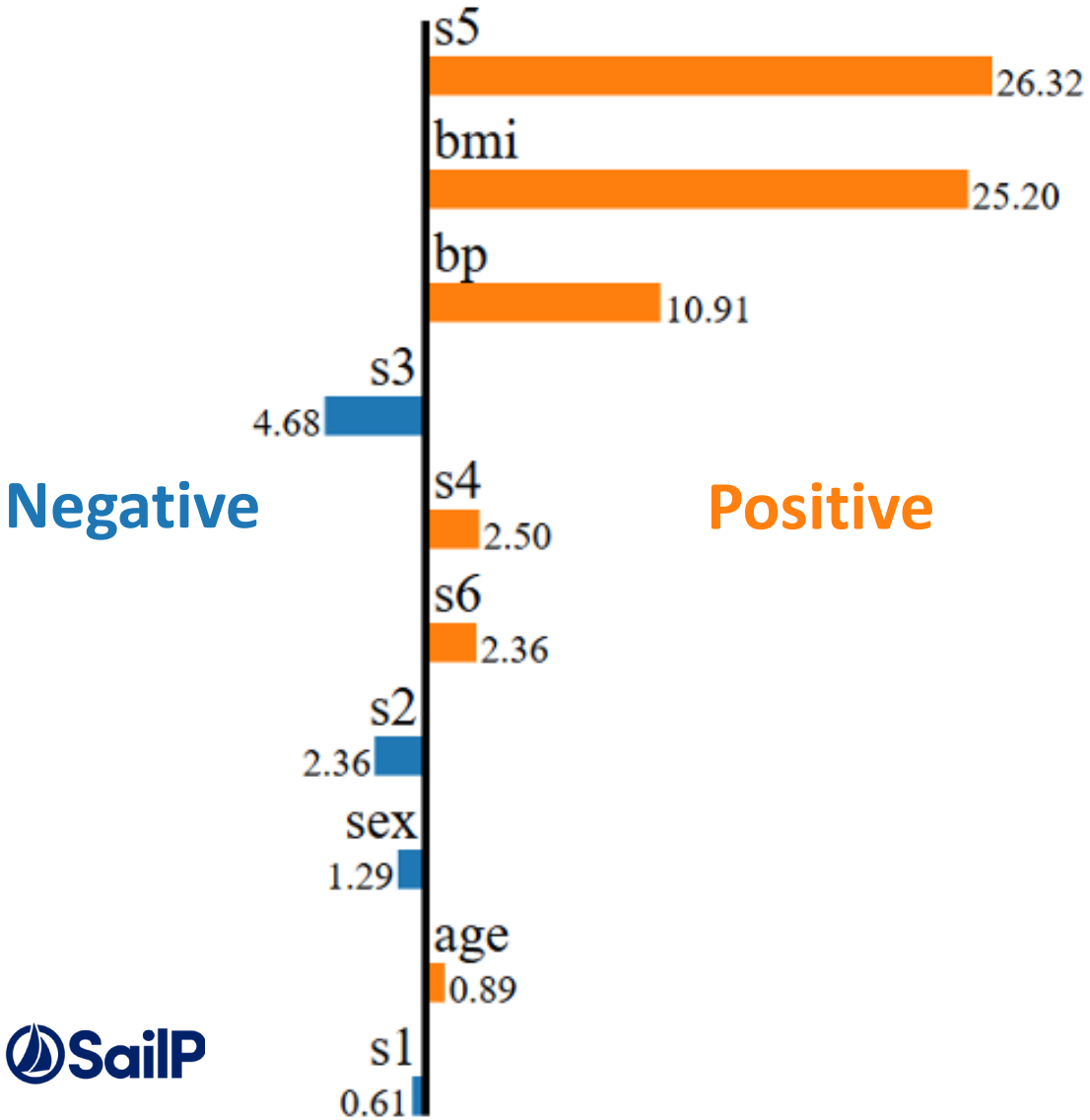
- **Local interpretation:** a particular instance or a small set of instances.

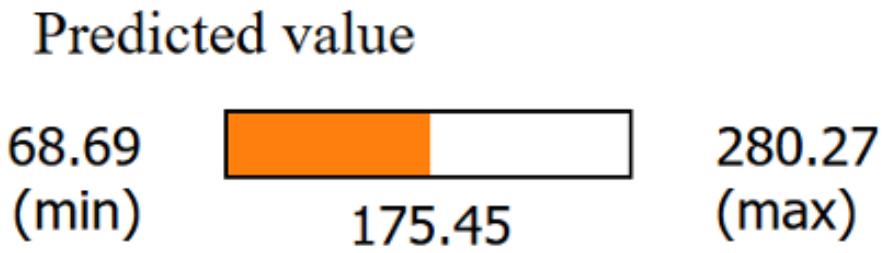
Diabetes Progression Dataset

Prediction Target:	Quantitative measure of disease progress
Instances:	442
Attributes:	Age Sex Body Mass Index (BMI) Blood Pressure (BP) LDL HDL …

https://web.stanford.edu/~hastie/StatLearnSparsity_files/DATA/diabetes.html

LIME Results: Diabetes Dataset



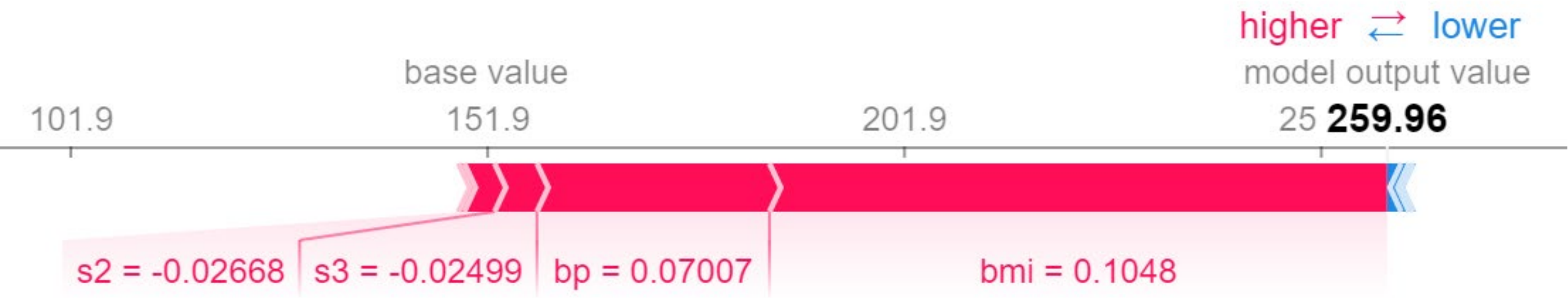


- Predicted value is a representation of diabetes progression
- Features are identified by their relative impact

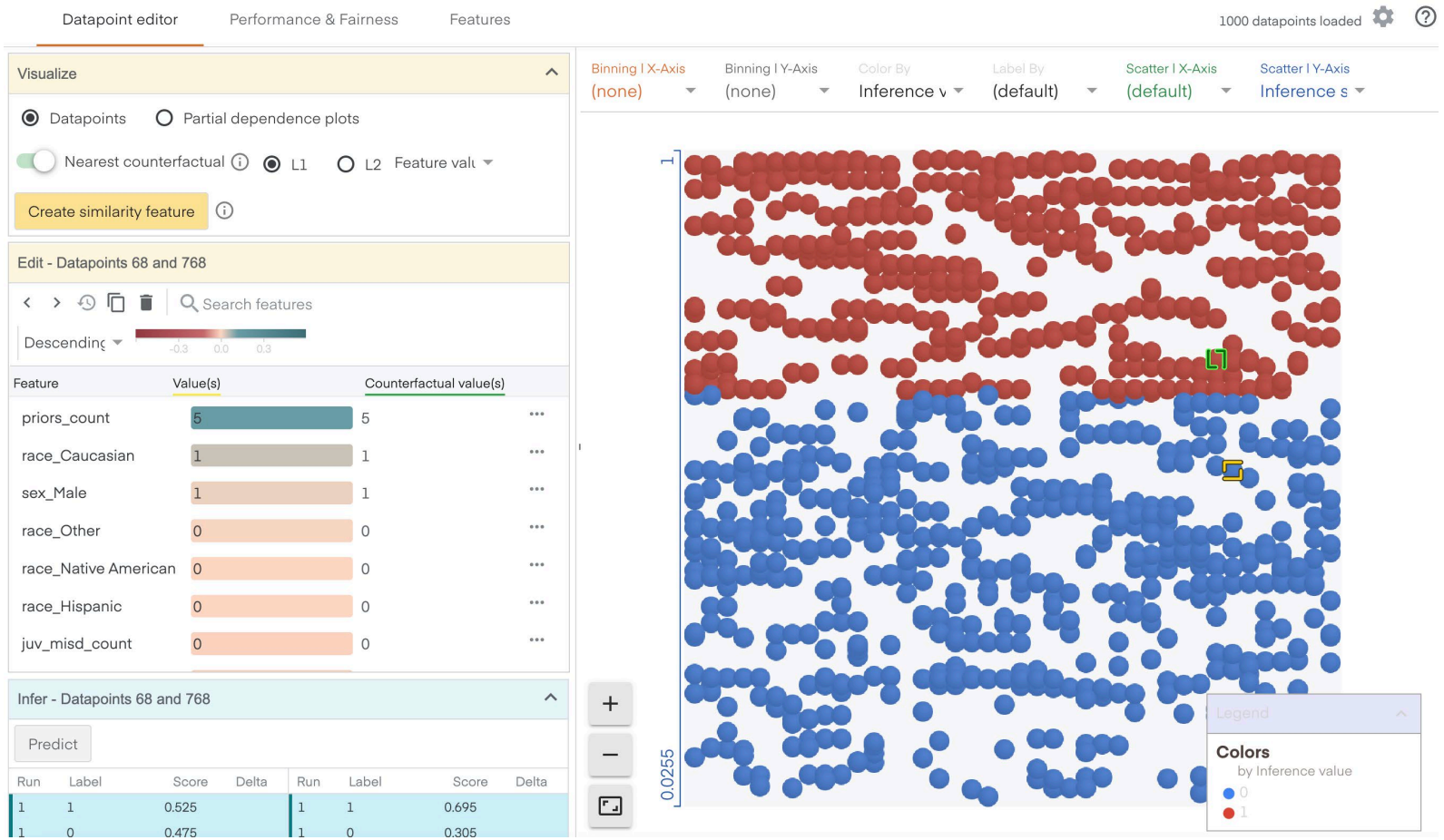
SailP

SHAP Results: Diabetes Dataset

- Shapley value in game theory
- Shows change in prediction value by feature



Google's What-If Tool



Data Exploration

"Counter Factuals"

Fairness

Supported Models:

Binary classification
Mutli-class classification
Regression tasks

Transparency

Fairness

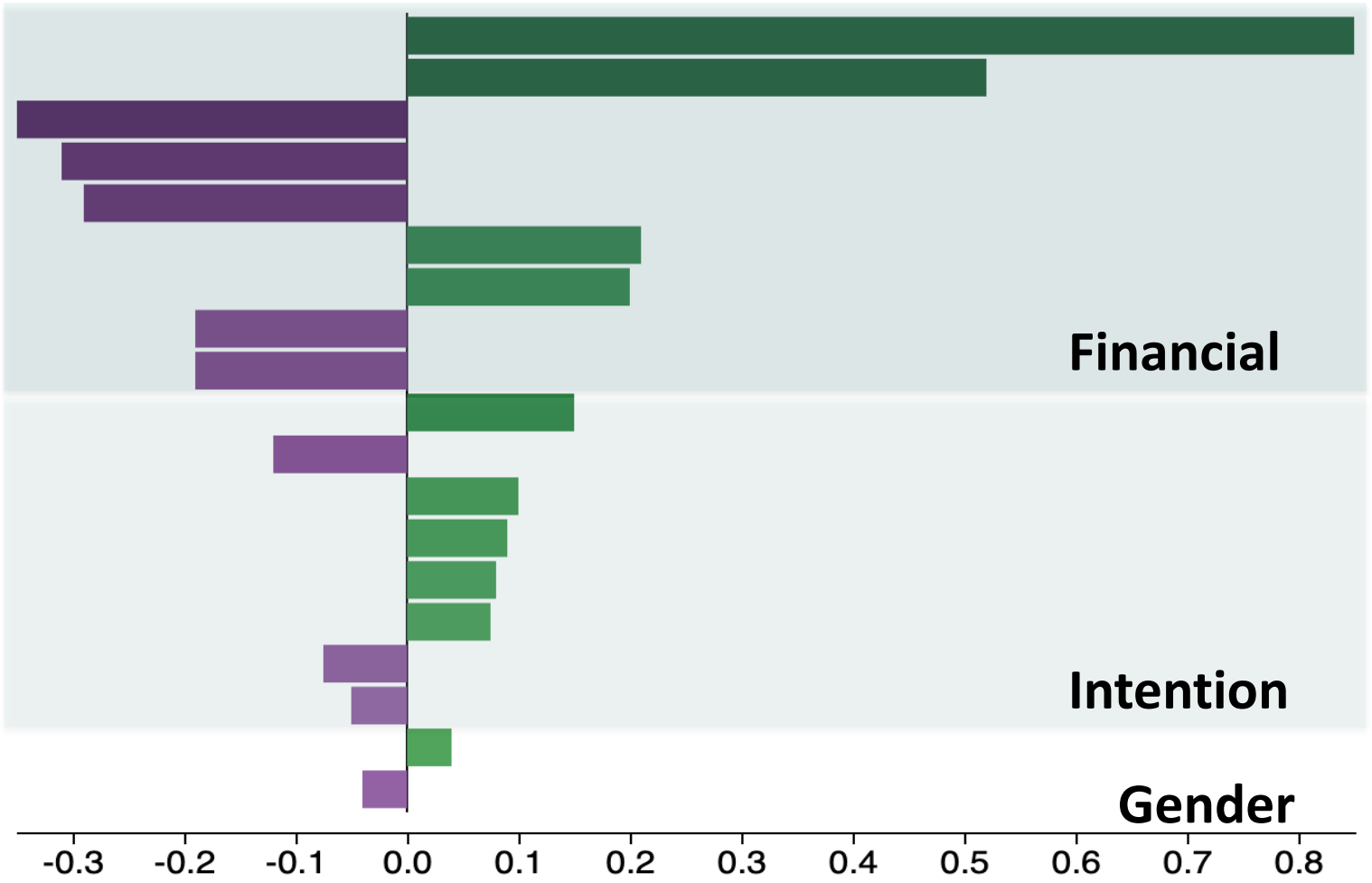
Bias

Shortcut

Shortcut | Impartiality

Shortcut | Impartiality | Self-Interest

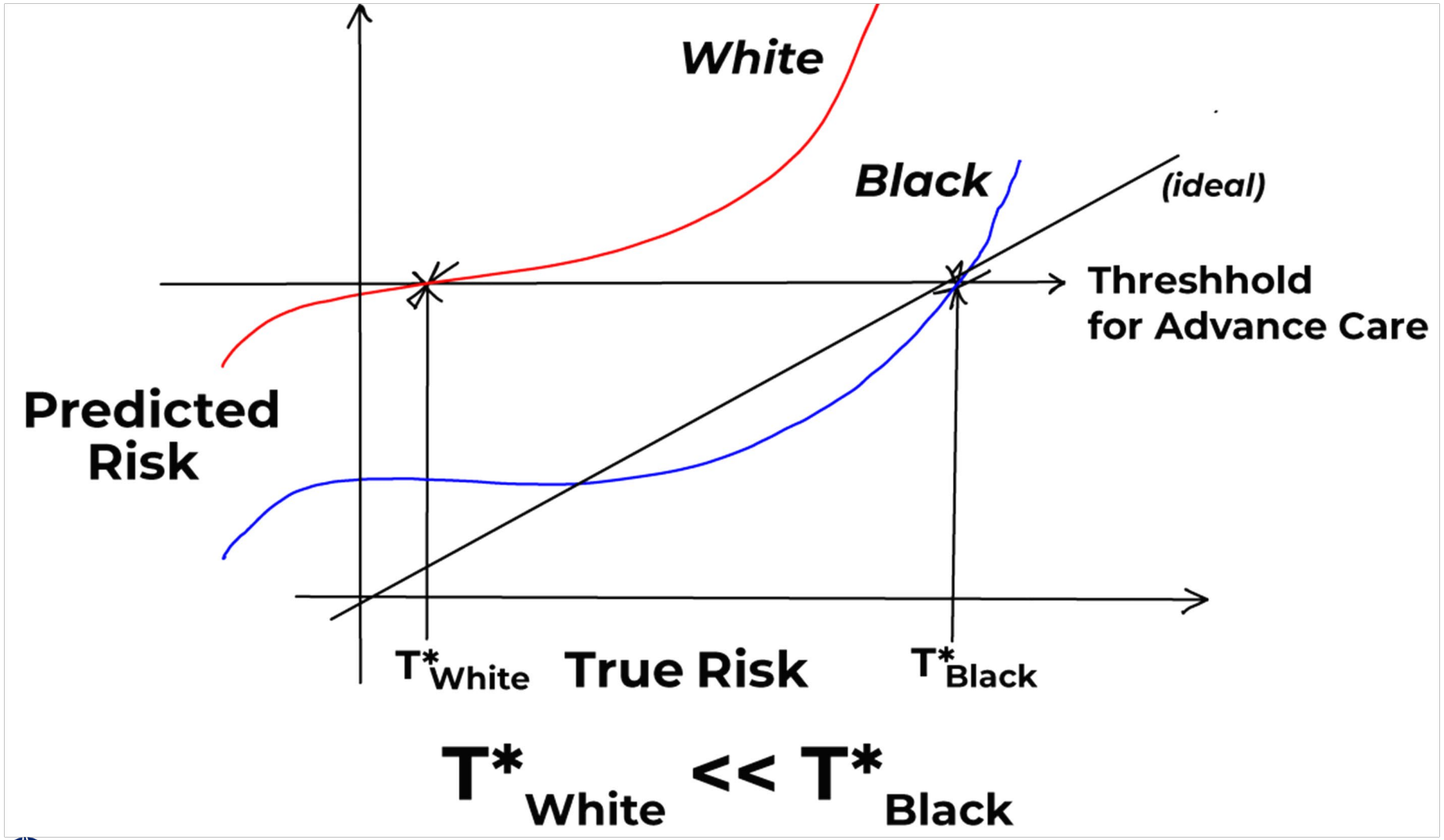
Gender Bias in German Credit Dataset



Relative Impact of Attributes on Credit

- Gender part of original data
- Checking for bias is relatively straightforward
- Gender is relatively low impact

Hidden Bias: Renal Failure Dataset

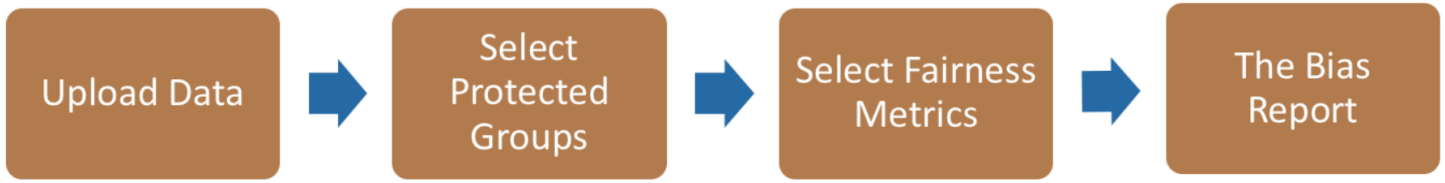


- Racial bias not immediately apparent
- Algorithm was optimized for cost

Aequitas Fairness Toolkit

Bias and Fairness Audit Toolkit

The Bias Report is powered by Aequitas, an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.



See an example report on COMPAS risk assessment scores.

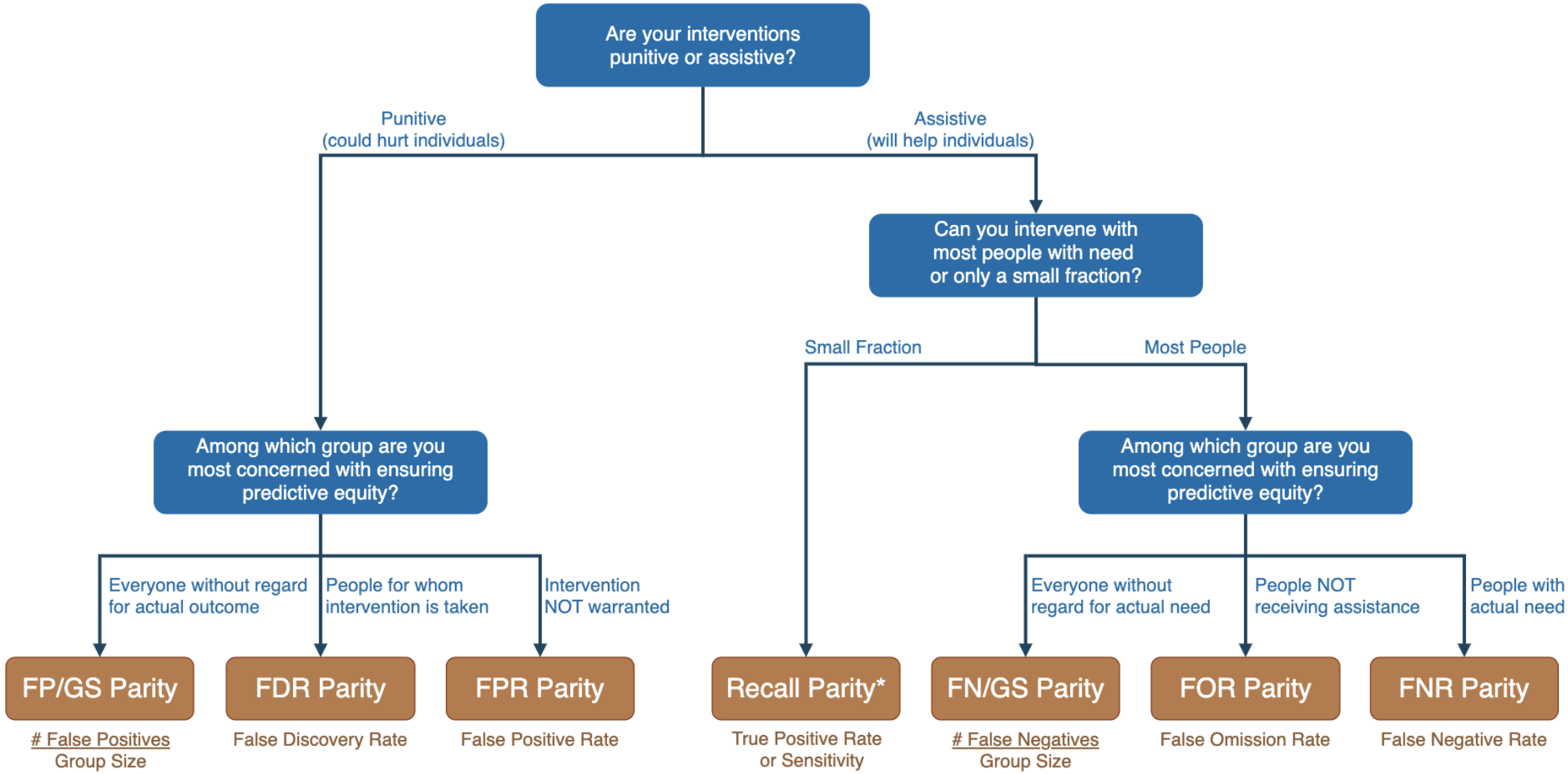
Or try out the audit tool using your own data or one of our sample data sets.

Get Started!

http://aequitas.dssg.io/

Aequitas Result Set

Aequitas Fairness Decision Tree



Aequitas: Formatting the Data Set

score
The predicted outcome

label_value
Ground truth (the actual result)

attribute_n
Attribute within the model

score	label_value	attribute_1	…	attribute_n
0	1	A		40
1	0	B		30
1	1	A		32
0	1	B		50
↑	↑	↑		↑
binary	**binary**	**categorical or continuous**		**categorical or continuous** (will be discretized into quartiles)

Comparison of Fairness Tools

Fairness Comparison	Benchmarks fairness of algorithms
AI Fairness 360	Embeddable code for bias/mitigation Developer-focused Input cleansing
Audit-AI	Equal Employment Opportunity Commission Standards Fairness in hiring driven
Thesis-ML	Fairness-aware ML algorithms
Aequitas	Toolkit (libraries, command line, web interface) Usable by policymakers and auditors
FairML	Calculates attribute significance for black-box models
FairTest	Subgroup / Decision Tree created from user population Reports on subgroup / feature association for potential bias
Themis	Python module for statistical analysis of bias
Fairness Measures	Top K ranking algorithms

So, What Now?

- Next week:
  - Inventory and consider the use cases of machine learning within your organization (both internal and outsourced)
  - Discuss with IRB, etc about potential issues that might need to be addressed
- In three months:
  - Investigate the tools that might be applicable for your datasets and machine learning usage:
    - Transparency: LIME / SHAP
    - Fairness: (Various tools)
- Within six months:
  - Begin to implement these tools and seek to rectify issues that may present themselves


TEMPLE OF JUSTICE

Links to Various Tools

Transparency	**Lime**	https://github.com/marcotcr/lime
	SHAP	https://github.com/slundberg/shap
	Google What-if?	https://pair-code.github.io/what-if-tool/
Fairness	**Aequitas**	https://github.com/dssg/aequitas
	Fairness Comparison	https://github.com/algofairness/fairness-comparison
	AI Fairness 360	https://github.com/IBM/aif360
	Audit-AI	https://github.com/pymetrics/audit-ai
	Thesis-ML	https://github.com/cosmicBboy/themis-ml
	FairML	https://github.com/adebayoj/fairml
	FairTest	https://github.com/columbia/fairtest
	Themis	https://themis-ml.readthedocs.io/en/latest/index.html
	Fairness Measures	https://fairnessmeasures.github.io/

Explore, Explain and Examine Predictive Models: https://pbiecek.github.io/ema/