# The Explanation Game
# Explaining ML models with Shapley Values

Joint work with Luke Merrick

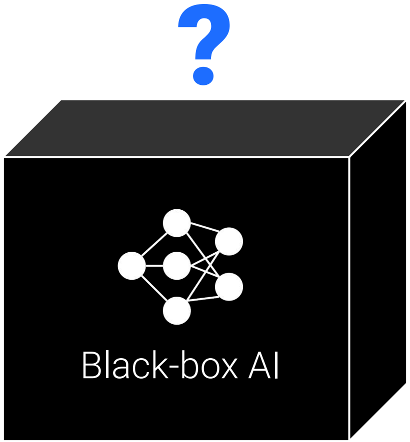
**Ankur Taly, Fiddler Labs**
ankur@fiddler.ai

fiddler

AI platform providing trust, visibility, and insights

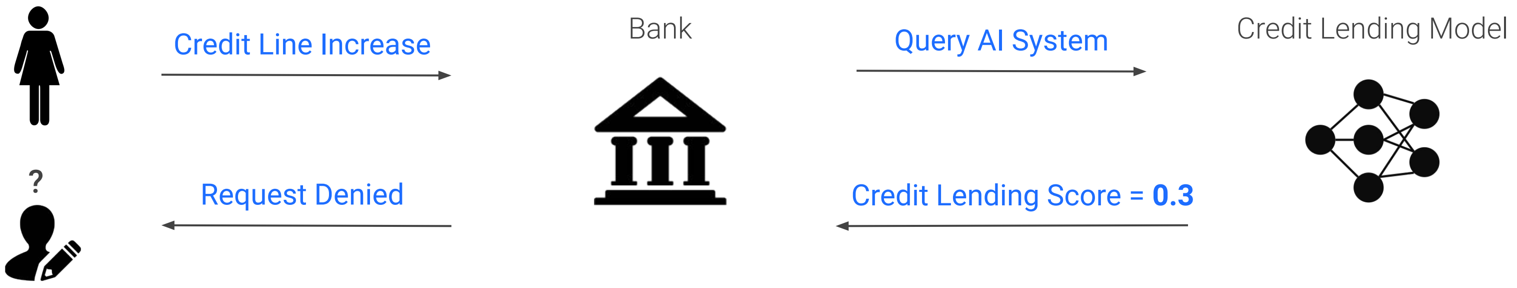
# Problem: Machine Learning is a Black box

**Output**
(Label, sentence, next word, game position)




**Input**
(Data, image, sentence, etc.)

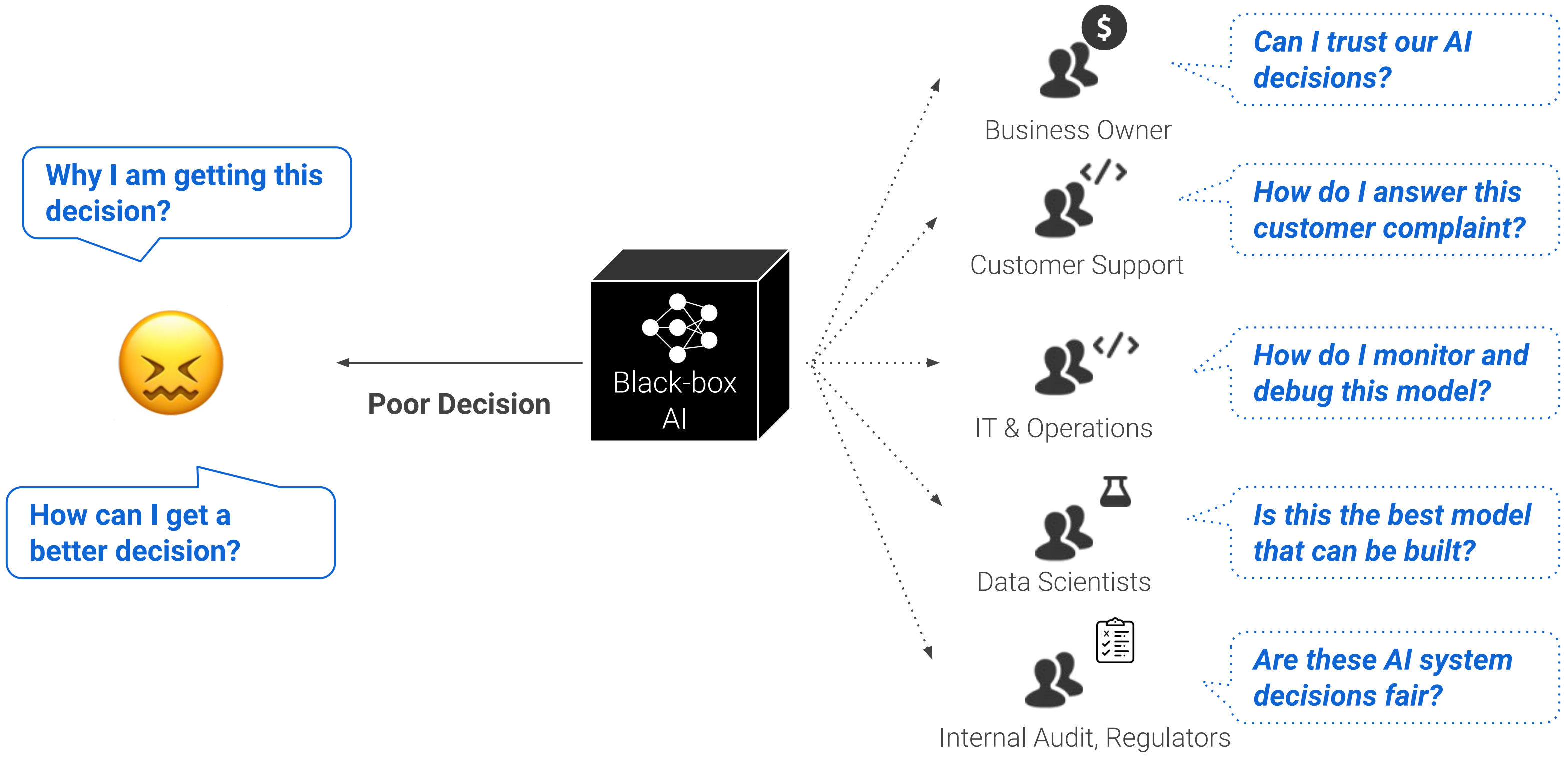
# Credit Lending in a black-box ML world




**Why? Why not? How?**

*Fair lending laws [ECOA, FCRA] require credit decisions to be explainable*

# Black-box AI creates confusion and doubt



# Why did the model make this prediction?

# The Attribution Problem

## Attribute a model's prediction on an input to features of the input

Examples:

- Attribute an object recognition network's prediction to its pixels
- Attribute a text sentiment network's prediction to individual words
- Attribute a lending model's prediction to its features

A reductive formulation of "why this prediction" but surprisingly useful :-)

# Applications of Attributions

- Debugging model predictions
- Generating an explanation for the end-user
- Analyzing model robustness
- Extracting rules from the model

# Gradient-Based Attribution Methods

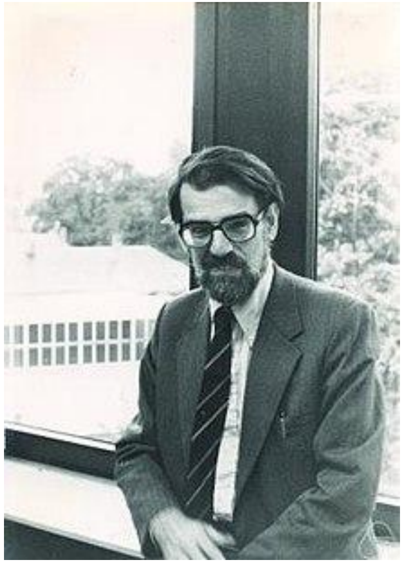
- **Feature*Gradient**
  - Paper: How to explain individual classification decisions, JMLR 2010
  - Inspired by linear models (where it amounts to feature*coefficient)
  - Does not work as well for highly non-linear models
- **Integrated Gradients**
  - Paper: Axiomatic Attribution for Deep Networks, ICML 2017
  - Integrate the gradients along a straight line path from the input at hand to a baseline
  - Inspired by **Aumann-Shapley values**
- Many more
  - GradCAM, SmoothGrad, Influence-Directed Explanations, …

But, what about non-differentiable models?

- Decision trees
- Boosted trees
- Random forests
- etc.

# Shapley Value

- Classic result in game theory on distributing the total gain from a **cooperative game**
- Introduced by **Lloyd Shapley** in **1953**[1], who later won the **Nobel Prize in Economics** in the 2012
- Popular tool in studying cost-sharing, market analytics, voting power, and most recently **explaining ML models**



Lloyd Shapley in 1980

[1] **"A Value for n-person Games"**. Contributions to the Theory of Games 2.28 (1953): 307-317

# Cooperative Game

- Players {1, …, M} collaborating to generate some **gain**
  - Think: Employees in a company creating some profit
  - Described by a **set function v(S)** specifying the gain for any subset $S \subseteq \{1, \ldots, M\}$

- **Shapley values** are a fair way to attribute the total gain to the players
  - Think: Bonus allocation to the employees
  - Shapley values are commensurate with the player's contribution

# Shapley Value Algorithm [Conceptual]

$$\phi_i(v) = \mathop{\mathbb{E}}_{\boldsymbol{O} \sim \pi(M)} [v(\text{pre}_i(\boldsymbol{O}) \cup \{i\}) - v(\text{pre}_i(\boldsymbol{O}))]$$

- Consider all possible permutations $\pi(M)$ of players (**M! possibilities**)
- In each permutation $\boldsymbol{O} \sim \pi(M)$
  - Add players to the coalition in that order
  - Note the marginal contribution of each player i to set of players before it in the permutation, i.e., $v(pre_i(\boldsymbol{O}) \cup \{i\}) - v(pre_i(\boldsymbol{O}))$
- The average marginal contribution across all permutations is the Shapley Value

# Example

A company with two employees **Alice** and **Bob**

- No employees, no profit [$v(\{\}) = 0$]
- Alice alone makes 20 units of profit [$v(\{Alice\}) = 20$]
- Bob alone makes 10 units of profit [$v(\{Bob\}) = 10$]
- Alice and Bob make 50 units of profit [$v(\{Alice, Bob\}) = 50$]

**What should the bonuses be?**

# Example

A company with two employees **Alice** and **Bob**

- No employees, no profit [v({}) = 0]
- Alice alone makes 20 units of profit [v({Alice}) = 20]
- Bob alone makes 10 units of profit [v({Bob}) = 10]
- Alice and Bob make 50 units of profit [v({Alice, Bob}) = 50]

**What should the bonuses be?**

| **Permutation** | **Marginal for Alice** | **Marginal for Bob** |
|---|---|---|
| Alice, Bob | 20 | 30 |
| Bob, Alice | 40 | 10 |
| **Shapley Value** | **30** | **20** |

# Axiomatic Justification

Shapley values are **unique under four simple axioms**

- **Dummy:** A player that doesn't contribute to any subset of players must receive zero attribution
- **Efficiency:** Attributions must add to the total gain
- **Symmetry:** Symmetric players must receive equal attribution
- **Linearity:** Attribution for the (weighted) sum of two games must be the same as the (weighted) sum of the attributions for each of the games

# Computing Shapley Values

**Exact computation**

- **Permutations-based approach** (Complexity: **O(M!)**)

$$\phi_i(v) = \mathop{\mathbb{E}}_{\boldsymbol{O} \sim \pi(M)} \left[ v(\mathrm{pre}_i(\boldsymbol{O}) \cup \{i\}) - v(\mathrm{pre}_i(\boldsymbol{O})) \right]$$

- **Subsets-based approach** (Complexity: **O($2^M$)**)

$$\phi_i(v) = \mathop{\mathbb{E}}_{S} \left[ \frac{2^{M-1}}{M} \binom{M-1}{|S|}^{-1} \left( v(S \cup \{i\}) - v(S) \right) \right]$$

# Computing Shapley Values

**Exact computation**

- **Permutations-based approach** (Complexity: **O(M!)**)

$$\phi_i(v) = \mathop{\mathbb{E}}_{\boldsymbol{O} \sim \pi(M)} \left[ v(\text{pre}_i(\boldsymbol{O}) \cup \{i\}) - v(\text{pre}_i(\boldsymbol{O})) \right]$$

- **Subsets-based approach** (Complexity: **O($2^M$)**)

$$\phi_i(v) = \mathop{\mathbb{E}}_{S} \left[ \frac{2^{M-1}}{M} \binom{M-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S)) \right]$$

- KernelSHAP: Solve a weighted least squares problem (Complexity: **O($2^M$)**)

$$\boldsymbol{\phi} = \arg\min_{\boldsymbol{\phi}} \sum_{S \subseteq \mathcal{M}} \frac{M-1}{\binom{M}{|S|} |S| (M - |S|)} \left( v(S) - \sum_{i=1}^{M} \phi_i \right)^2$$

# Computing Shapley Values

**Approximation computation**

- General idea: Express Shapley Values as an expectation over a distribution of marginals, and use sampling-based methods to estimate the expectation
- See: "Computational Aspects of Cooperative Game Theory", Chalkiadakis et al. 2011

# Shapley Values for Explaining ML Models

# Shapley Values for Explaining ML models

- Define a coalition game for each model input x to be explained
    - **Players are the features of the input**
    - **Gain is the model prediction F(x)**
- Feature attributions are the Shapley values of this game

We call the coalition game setup for computing Shapley Values as the ***"Explanation Game"***

# Setting up the Coalition Game

**Challenge**: Defining the prediction F(x) when only a subset of features  are present?

i.e., *what is* ***F(*** $x_1$***, <absent>,*** $x_3$***, <absent>, ..*** $x_m$***)?***

# Setting up the Coalition Game

**Challenge**: Defining the prediction F(x) when only a subset of features are present?
i.e., *what is* ***$F(x_1$, <absent>, $x_3$, <absent>, ..$x_m)$?***

**Idea 1: Model absent feature with an empty or zero value**

- Works well for image and text inputs
- Does not work well for structured inputs; what is the empty value for “income”?

# Setting up the Coalition Game

**Challenge**: Defining the prediction F(x) when only a subset of features are present?
i.e., *what is* ***F($x_1$, <absent>, $x_3$, <absent>, ..$x_m$)?***

**Idea 1: Model absent feature with an empty or zero value**

- Works well for image and text inputs
- Does not work well for structured inputs; what is the empty value for "income"?

**Idea 2: Sample values for the absent features and compute the expected prediction**

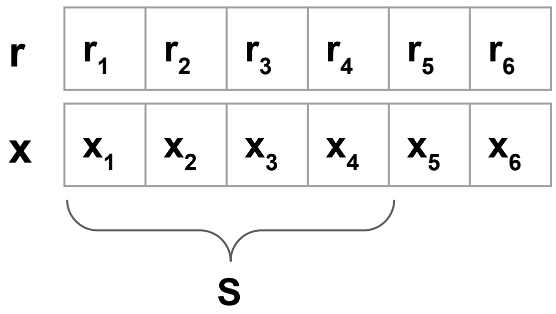
- This is the approach taken by most Shapley Value based explanation methods

# Notation for next few slides

- Model **F:** $\mathcal{X} \rightarrow \mathbf{R}$ where $\mathcal{X}$ is an M-dimensional input space
- Input distribution: $\mathbf{D}^{\mathbf{inp}}$
- Inputs to be explained: $\mathbf{x} \in \mathcal{X}$
- Reference inputs: $\mathbf{r}, \mathbf{r}_1, .. \in \mathcal{X}$

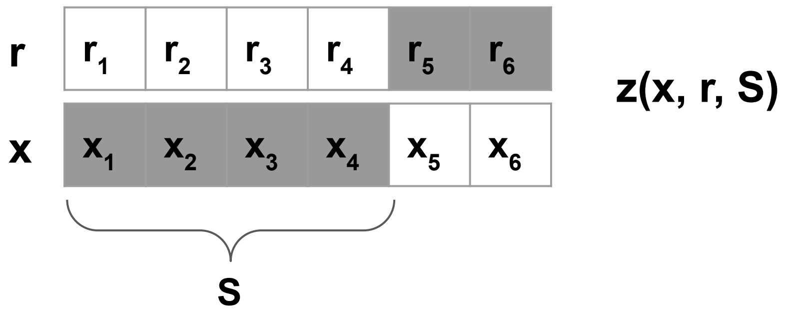
# Notation for next few slides

- Model **F:** $\mathcal{X} \rightarrow$ **R** where $\mathcal{X}$ is an M-dimensional input space
- Input distribution: $\mathbf{D}^{\mathbf{inp}}$
- Inputs to be explained: $\mathbf{x} \in \mathcal{X}$
- Reference inputs: $\mathbf{r}, \mathbf{r_1}, .. \in \mathcal{X}$
- A **composite input z(x, r, S)** is an input that agrees with x on features in S and with **r** on all the other features



# Notation for next few slides

- Model **F:** $\mathcal{X} \rightarrow \mathbf{R}$ where $\mathcal{X}$ is an M-dimensional input space
- Input distribution: $\mathbf{D}^{\mathbf{inp}}$
- Inputs to be explained: $\mathbf{x} \in \mathcal{X}$
- Reference inputs: $\mathbf{r}, \mathbf{r}_1, .. \in \mathcal{X}$
- A **composite input z(x, r, S)** is an input that agrees with x on features in S and with **r** on all the other features



# General game formulation

Given an input **x**, the payoff for a feature set S is the expected prediction over composite inputs z(x, r, S) where the references r are drawn from a distribution $\mathbf{D_{x,S}}$

$$v_x(S) ::= \mathop{\mathbb{E}}_{r \sim D_{x,S}}[F(z(x, r, S))] - \mathop{\mathbb{E}}_{r \sim D_{x,\phi}}[F(r)]$$

# General game formulation

Given an input **x**, the payoff for a feature set S is the expected prediction over composite inputs z(x, r, S) where the references r are drawn from a distribution $\mathbf{D}_{\mathbf{x,S}}$

$$v_x(S) ::= \mathop{\mathbb{E}}_{r\sim D_{x,S}}[F(z(x,r,S))] - \mathop{\mathbb{E}}_{r\sim D_{x,\phi}}[F(r)]$$

Features in S come from x while the remaining are sampled based on $D_{x,S}$

Offset term to ensure that the **gain for the empty set is zero**

# General game formulation

Given an input **x**, the payoff for a feature set S is the expected prediction over composite inputs z(x, r, S) where the references r are drawn from a distribution $\mathbf{D}_{\mathbf{x,S}}$

$$v_x(S) ::= \mathop{\mathbb{E}}_{r \sim D_{x,S}} [F(z(x, r, S))] - \mathop{\mathbb{E}}_{r \sim D_{x,\phi}} [F(r)]$$

Reference distribution $\mathbf{D}_{\mathbf{x,S}}$ varies across methods

- [SHAP, NIPS 2018] Uses conditional distribution, i.e., $D_{x,S} = \{r \sim D^{inp} \mid x_S = r_S\}$
- [KernelSHAP, NIPS 2018] Uses input distribution, i.e., $D_{x,S} = D^{inp}$
- [QII, S&P 2016] Uses joint-marginal distribution, i.e., $D_{x,S} = D^{J.M.}$
- [IME, JMLR 2010] Use uniform distribution, i.e., $D_{x,S} = \mathcal{U}$

# General game formulation

Given an input **x**, the payoff for a feature set S is the expected prediction over composite inputs z(x, r, S) where the references r are drawn from a distribution $\mathbf{D_{x,S}}$

$$v_x(S) ::= \mathop{\mathbb{E}}_{r \sim D_{x,S}} [F(z(x, r, S))] - \mathop{\mathbb{E}}_{r \sim D_{x,\phi}} [F(r)]$$

Reference distribution $\mathbf{D_{x,S}}$ varies across methods

- [SHAP, NIPS 2018] Uses conditional distribution, i.e., $D_{x,S} = \{r \sim D^{inp} \mid x_S = r_S\}$
- [KernelSHAP, NIPS 2018] Uses input distribution, i.e., $D_{x,S} = D^{inp}$
- [QII, S&P 2016] Uses joint-marginal distribution, i.e., $D_{x,S} = D^{J.M.}$
- [IME, JMLR 2010] Use uniform distribution, i.e., $D_{x,S} = \mathcal{U}$

**This is a critical choice that strongly impacts the resulting Shapley Values!!**

# Rest of the lecture

We will discuss the following preprint:

The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory,
Luke Merrick and Ankur Taly, 2019

- The many game formulations and the many Shapley values
- A decomposition of Shapley values in terms of single-reference games
- Confidence intervals for Shapley value approximations
- Ties to Norm Theory that enable contrastive explanations

# Mover Example 1 (from the QII paper)

**F(is_male, is_lifter) ::= is_male** (model only hires males)

Input to be explained: **is_male = 1, is_lifter = 1**

**Data and prediction distribution**

| **is_male** | **is_lifter** | **P[X=x]** | **F(x)** |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

# Mover Example 1 (from the QII paper)

**F(is_male, is_lifter) ::= is_male** (model only hires males)

Input to be explained: **is_male = 1, is_lifter = 1**

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

**Attributions for is_male=1, is_lifter = 1**

| Method | is_male | is_lifter |
|---|---|---|
| **SHAP** (conditional distribution) | 0.05 | 0.05 |
| **KernelSHAP** (input distribution) | 0.10 | 0.0 |
| **QII** (joint-marginal distribution) | 0.10 | 0.0 |
| **IME** (uniform distribution) | 0.50 | 0.0 |

# Mover Example 1 (from the QII paper)

**F(is_male, is_lifter) ::= is_male** (model only hires males)

Input to be explained: **is_male = 1, is_lifter = 1**

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

**Attributions for is_male=1, is_lifter = 1**

| Method | is_male | is_lifter |
|---|---|---|
| **SHAP** (conditional distribution) | 0.05 | 0.05 |
| **KernelSHAP** (input distribution) | 0.10 | 0.0 |
| **QII** (joint-marginal distribution) | [illegible] | 0.0 |
| **IME** (uni[illegible] | [illegible] | 0.0 |

Why does SHAP attribute to the **is_lifter** feature which plays no role in the model?

# Attributions under conditional distribution [SHAP]

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] ($D^{inp}$) | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

# Attributions under conditional distribution [SHAP]

### Data and prediction distribution

| is_male | is_lifter | P[X=x] ($D^{inp}$) | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

### Attributions for is_male=1, is_lifter = 1

| Permutation | Marginal for is_male | Marginal for is_lifter |
|---|---|---|
| is_male, is_lifter | 0.1 | 0.0 |
| is_lifter, is_male | 0.0 | 0.1 |
| **Average** | 0.05 | 0.05 |

# Attributions under conditional distribution [SHAP]

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] ($D^{inp}$) | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

**Attributions for is_male=1, is_lifter = 1**

| Permutation | Marginal for is_male | Marginal for is_lifter |
|---|---|---|
| is_male, is_lifter | 0.1 | 0.0 |
| is_lifter, is_male | 0.0 | 0.1 |
| **Average** | 0.05 | 0.05 |

$$
\begin{aligned}
v^{cond}(\{\}) &= 0.0 \\
v^{cond}(\{\text{is\_male}\}) &= \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}]) \mid \text{is\_male} = 1] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] \\
&= 1.0 - 0.9 = 0.1 \\
v^{cond}(\{\text{is\_lifter}\}) &= \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}]) \mid \text{is\_lifter} = 1] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] \\
&= 1.0 - 0.9 = 0.1 \\
v^{cond}(\{\text{is\_male}, \text{is\_lifter}\}) &= 1.0 - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] \\
&= 1.0 - 0.9 = 0.1
\end{aligned}
$$

# Attributions under conditional distribution [SHAP]

### Data and prediction distribution

| is_male | is_lifter | P[X=x] ($D^{inp}$) | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

### Attributions for is_male=1, is_lifter = 1

| Permutation | Marginal for is_male | Marginal for is_lifter |
|---|---|---|
| is_male, is_lifter | 0.1 | 0.0 |
| is_lifter, is_male | 0.0 | 0.1 |
| **Average** | 0.05 | 0.05 |

$$v^{cond}(\{\}) = 0.0$$

$$v^{cond}(\{\text{is\_male}\}) = \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}]) \mid \text{is\_male} = 1] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] = 1.0 - 0.9 = 0.1$$

$$v^{cond}(\{\text{is\_lifter}\}) = \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}]) \mid \text{is\_lifter} = 1] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] = 1.0 - 0.9 = 0.1$$

$$v^{cond}(\{\text{is\_male}, \text{is\_lifter}\}) = 1.0 - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] = 1.0 - 0.9 = 0.1$$

# Attributions under input distribution [KernelSHAP]

### Data and prediction distribution

| is_male | is_lifter | P[X=x] ($D^{inp}$) | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

### Attributions for is_male=1, is_lifter = 1

| Permutation | Marginal for is_male | Marginal for is_lifter |
|---|---|---|
| is_male, is_lifter | 0.1 | 0.0 |
| is_lifter, is_male | 0.1 | 0.0 |
| **Average** | 0.1 | 0.0 |

# Attributions under input distribution [KernelSHAP]

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] ($D^{inp}$) | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

**Attributions for is_male=1, is_lifter = 1**

| Permutation | Marginal for is_male | Marginal for is_lifter |
|---|---|---|
| is_male, is_lifter | 0.1 | 0.0 |
| is_lifter, is_male | 0.1 | 0.0 |
| **Average** | 0.1 | 0.0 |

$$
\begin{aligned}
v^{inp}(\{\}) &= 0.0 \\
v^{inp}(\{\text{is\_male}\}) &= \mathbb{E}[F([1, \text{is\_lifter}])] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] \\
&= 1.0 - 0.9 = 0.1 \\
v^{inp}(\{\text{is\_lifter}\}) &= \mathbb{E}[F([\text{is\_male}, 1])] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] \\
&= 0.9 - 0.9 = 0.0 \\
v^{inp}(\{\text{is\_male}, \text{is\_lifter}\}) &= 1.0 - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])] \\
&= 1.0 - 0.9 = 0.1
\end{aligned}
$$

# Attributions under input distribution [KernelSHAP]

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] ($D^{inp}$) | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

**Attributions for is_male=1, is_lifter = 1**

| Permutation | Marginal for is_male | Marginal for is_lifter |
|---|---|---|
| is_male, is_lifter | 0.1 | 0.0 |
| is_lifter, is_male | 0.1 | 0.0 |
| **Average** | 0.1 | 0.0 |

$$v^{inp}(\{\}) = 0.0$$

$$v^{inp}(\{\text{is\_male}\}) = \mathbb{E}[F([1, \text{is\_lifter}])] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])]$$
$$= 1.0 - 0.9 = 0.1$$

$$v^{inp}(\{\text{is\_lifter}\}) = \mathbb{E}[F([\text{is\_male}, 1])] - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])]$$
$$= 0.9 - 0.9 = 0.0$$

$$v^{inp}(\{\text{is\_male}, \text{is\_lifter}\}) = 1.0 - \mathbb{E}[F([\text{is\_male}, \text{is\_lifter}])]$$
$$= 1.0 - 0.9 = 0.1$$

# Mover Example 2

**F(is_male, is_lifter) ::= is_male AND is_lifter**   (model hires males who are lifters)

Input to be explained: **is_male = 1, is_lifter = 1**

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | 0.0 | 0 |
| 1 | 0 | 0.4 | 1 |
| 1 | 1 | 0.5 | 1 |

**Attributions for is_male=1, is_lifter = 1**

| Method | is_male | is_lifter |
|---|---|---|
| **SHAP** (conditional distribution) | 0.028 | 0.047 |
| **KernelSHAP** (input distribution) | 0.05 | 0.045 |
| **QII** (joint-marginal distribution) | 0.075 | 0.475 |
| **IME** (uniform distribution) | 0.375 | 0.375 |

# Mover Example 2

**F(is_male, is_lifter) ::= is_male AND is_lifter** (model hires males who are lifters)

Input to be explained: **is_male = 1, is_lifter = 1**

**Data and prediction distribution**

| is_male | is_lifter | P[X=x] | F(x) |
|---|---|---|---|
| 0 | 0 | 0.1 | 0 |
| 0 | 1 | [illegible] | [illegible] |
| 1 | 0 | [illegible] | [illegible] |
| 1 | 1 | [illegible] | [illegible] |

**Attributions for is_male=1, is_lifter = 1**

| Method | is_male | is_lifter |
|---|---|---|
| **SHAP** (conditional distribution) | 0.028 | 0.047 |
| **KernelSHAP** (input distribution) | 0.05 | 0.045 |
| **QII** (joint-marginal distribution) | 0.075 | 0.475 |
| **IME** (uniform distribution) | 0.375 | 0.375 |

Each method produces a different attribution!

How do we reconcile the differences between the various Shapley Values?

# The unconditional case

General game formulation

$$v_x(S) ::= \mathop{\mathbb{E}}_{r \sim D_{x,S}} [F(z(x, r, S))] - \mathop{\mathbb{E}}_{r \sim D_{x,\phi}} [F(r)]$$

# The unconditional case

General game formulation

$$v_x(S) ::= \mathop{\mathbb{E}}_{r \sim D_{x,S}} [F(z(x,r,S))] - \mathop{\mathbb{E}}_{r \sim D_{x,\phi}} [F(r)]$$

Consider the case where the reference distribution $\mathbf{D_{x,S} ::= D}$ is the same across all inputs x and subsets S

$$v_{x,D}(S) ::= \mathop{\mathbb{E}}_{r \sim D} [F(z(x,r,S))] - \mathop{\mathbb{E}}_{r \sim D} [r]$$

# The unconditional case

General game formulation

$$v_x(S) ::= \mathop{\mathbb{E}}_{r \sim D_{x,S}} [F(z(x, r, S))] - \mathop{\mathbb{E}}_{r \sim D_{x,\phi}} [F(r)]$$

Consider the case where the reference distribution $\mathbf{D_{x,S} ::= D}$ is the same across all inputs x and subsets S

$$v_{x,D}(S) ::= \mathop{\mathbb{E}}_{r \sim D} [F(z(x, r, S))] - \mathop{\mathbb{E}}_{r \sim D} [r]$$

Ensures that **irrelevant features get zero attribution (**see paper for proof)

**KernelSHAP**, **QII**, **IME** fall in this case (but choose different reference distributions)

# Single-reference Games

**Idea**: Model feature absence using a specific reference

Given an input **x and a specific reference r**,
the payoff for a feature set S is the prediction for the composite input z(x, r, S)

$$v_{x,r}(S) ::= F(z(x,r,S))] - F(r)$$

Side note: Integrated Gradients is a single-reference attribution method.

# Single-reference Games

**Idea**: Model feature absence using a specific reference

Given an input **x and a specific reference r**,
the payoff for a feature set S is the prediction for the composite input z(x, r, S)

$$v_{x,r}(S) ::= F(z(x, r, S))] - F(r)$$

Offset term to ensure that the **gain for the empty set is zero**

Side note: [Integrated Gradients]() is a single-reference attribution method.

# A decomposition in terms of single-reference games

Shapley values of $\mathbf{v}_{\mathbf{x,D}}$ can be expressed as an expectation over Shapley values from single-reference games $\mathbf{v}_{\mathbf{x,r}}$ where the references r are drawn from D.

**Lemma**: $\phi_i(v_{x,D}(S)) ::= \mathop{\mathbb{E}}_{r\sim D}[\phi_i(v_{x,r}(S))]$

# A decomposition in terms of single-reference games

Shapley values of $\mathbf{v}_{x,D}$ can be expressed as an expectation over Shapley values from single-reference games $\mathbf{v}_{x,r}$ where the references r are drawn from D.

**Lemma**: $\phi_i(v_{x,D}(S)) ::= \mathop{\mathbb{E}}_{r\sim D}[\phi_i(v_{x,r}(S))]$

Thus, the different Shapley Values across **KernelSHAP**, **QII**, **IME** are essentially differently weighted aggregations across a space of single-reference games

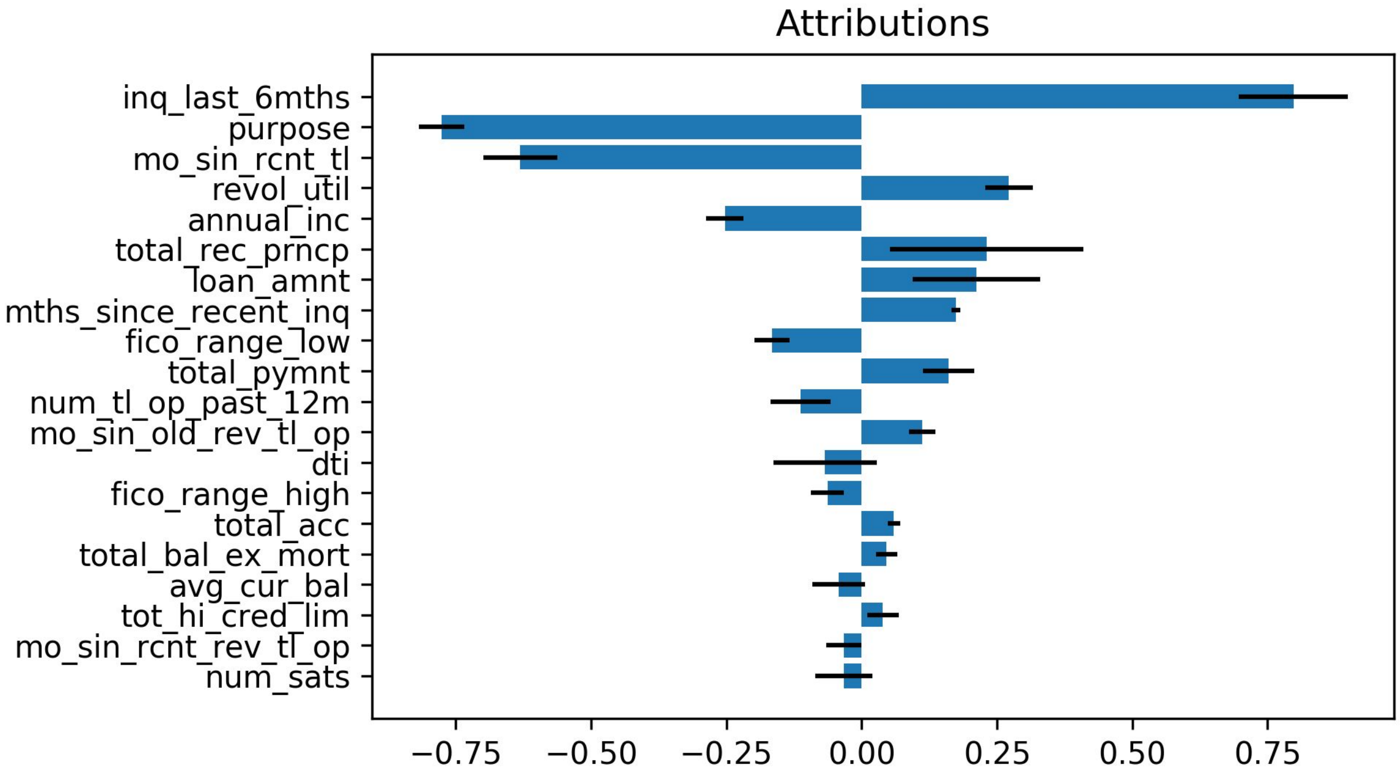
# Confidence Intervals

**Lemma**: $\phi_i(v_{x,D}(S)) ::= \mathop{\mathbb{E}}_{r\sim D}[\phi_i(v_{x,r}(S))]$

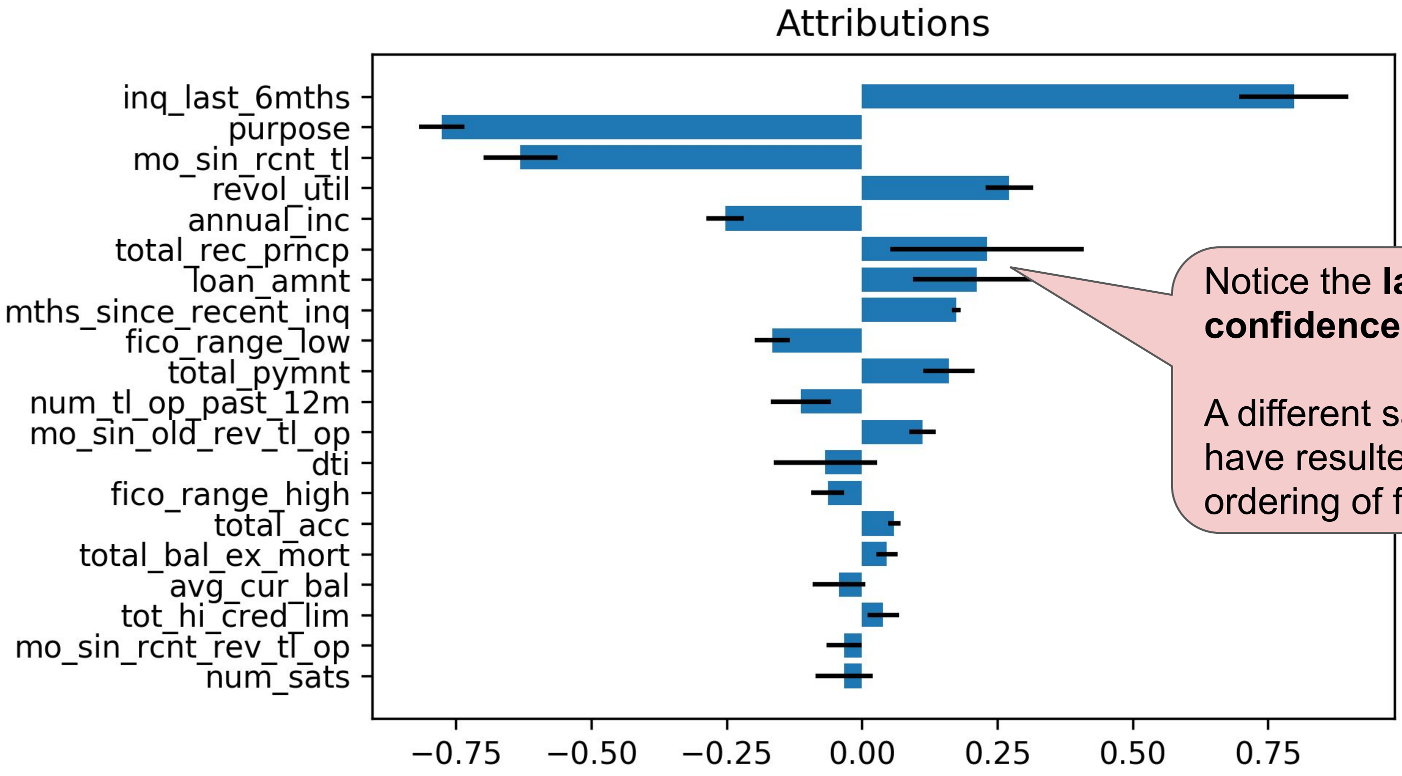
- Directly computing $\phi_i(v_{x,D}(S))$ involves estimating several expectations
- This makes it challenging to quantify the estimation uncertainty
- Our decomposition reduces the computation to estimating a single expectation
- Confidence intervals (CIs) can now easily be estimated from the sample standard deviation (SSD); courtesy central limit theorem.

$$\bar{\boldsymbol{\phi}} \pm \frac{1.96 \times \mathsf{SSD}(\{\boldsymbol{\phi}(v_{\boldsymbol{x},\boldsymbol{r}_i})\}_{i=1}^{N})}{\sqrt{N}}$$ [95% CIs]

# Showing Confidence Intervals is important!



# Showing Confidence Intervals is important!




Notice the **large confidence interval**.

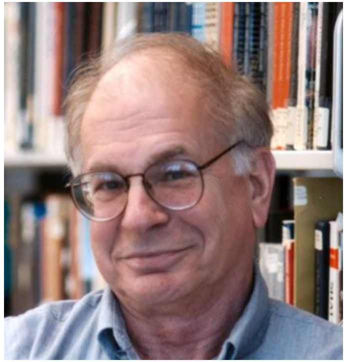
A different sampling may have resulted in a different ordering of features

# A new perspective on Shapley value attributions

# Norm Theory [Kahneman and Miller, 1986]

Classic work in cognitive psychology.

Describes a theory of psychological norms that shape the emotional responses, social judgments, and **explanations of humans**.



Daniel Kahneman

Dale T. Miller

# Three learnings from Norm Theory (and related work)

- **“Why” questions evoke counterfactual norms**
  - *“A why question indicates that a particular event is surprising and requests the explanation of an effect, denned as a contrast between an observation and a more normal alternative.”*
  - **Learning: Explanations are contrastive!**

# Three learnings from Norm Theory (and related work)

- **"Why" questions evoke counterfactual norms**
  - *"A why question indicates that a particular event is surprising and requests the explanation of an effect, denned as a contrast between an observation and a more normal alternative."*
  - **Learning: Explanations are contrastive!**
- **Norms vary depending on their context**
  - *"A man suffers from indigestion. Doctor blames it to a stomach ulcer. Wife blames it on eating turnips."* [Hart and Honoré., 1985]
  - **Learning**: **Different contrasts yield different explanations**

# Three learnings from Norm Theory (and related work)

- **"Why" questions evoke counterfactual norms**
  - *"A why question indicates that a particular event is surprising and requests the explanation of an effect, denned as a contrast between an observation and a more normal alternative."*
  - **Learning: Explanations are contrastive!**
- **Norms vary depending on their context**
  - *"A man suffers from indigestion. Doctor blames it to a stomach ulcer. Wife blames it on eating turnips."* [Hart and Honoré., 1985]
  - **Learning: Different contrasts yield different explanations**
- **Norms tend to be relevant to to the question at hand**
  - *"Our capacity for counterfactual reasoning seems to show a strong resistance to any consideration of irrelevant counterfactuals."* [Hitchcock and Knobecaus, 2009]
  - **Learning: Contrasts must be picked carefully**

# Shapley Values meet Norm Theory

**Lemma**: $\phi_i(v_{x,D}(S)) ::= \mathop{\mathbb{E}}_{r \sim D}[\phi_i(v_{x,r}(S))]$

- Shapley values **contrastively explain the prediction on an input against a distribution of references (norms)**

- Reference distribution can be varied to obtain different explanations.
  - E.g., Explain a loan application rejection by contrasting with:
    - All application who were accepted, or
    - All applications with the same income level as the application at hand

- Reference distribution must be relevant to the explanation being sought
  - E.g., Explain a B- grade by contrasting with B+ (next higher grade), not an A+

# Regulation may favor Contrastive Explanations

The Official Staff Interpretation to Regulation B of the Equal Credit Opportunity Act originally published in 1985[1] states:

> *"One method is to identify the factors for which the applicant's score fell furthest below the average score for each of those factors* ***achieved by applicants whose total score was at or slightly above the minimum passing score****. Another method is to identify the factors for which the applicant's score fell furthest below the average score for each of those factors achieved by all applicants."*

[1] 12 CFR Part 1002 - Equal Credit Opportunity Act (Regulation B), 1985

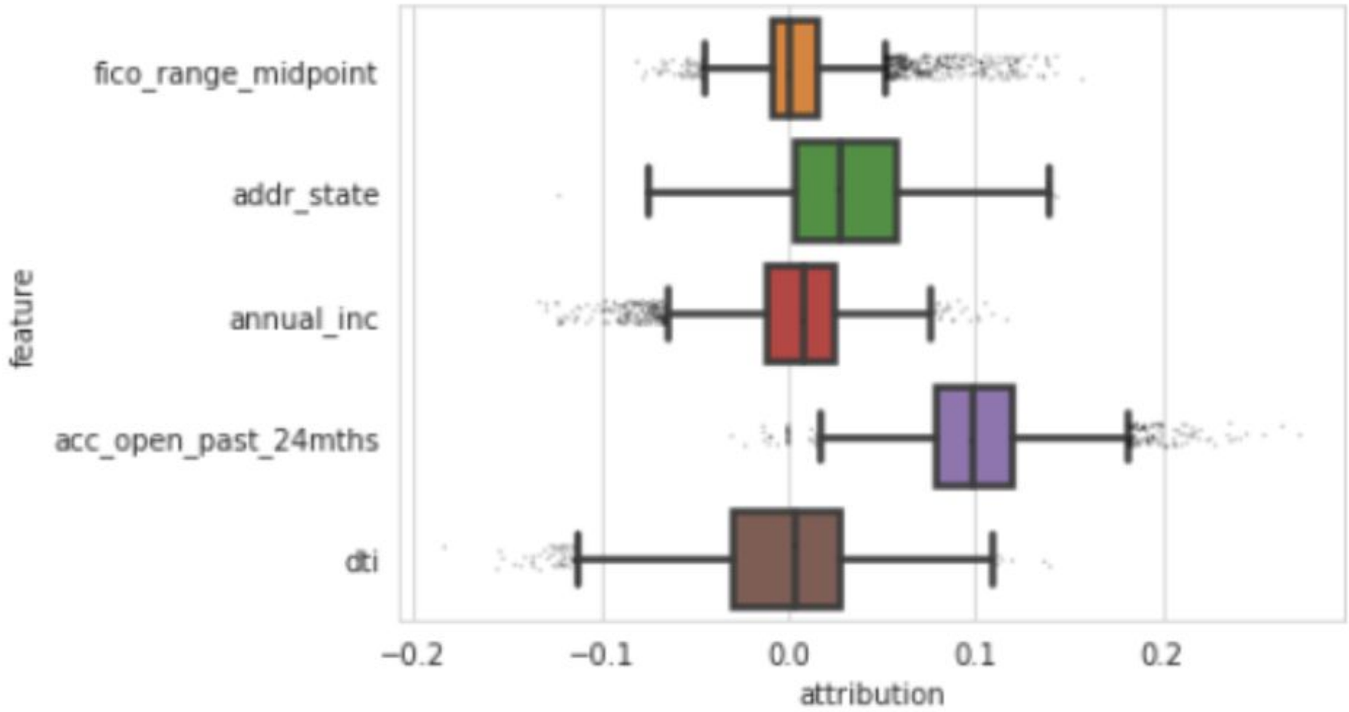
# Formulate-Approximate-Explain

Three step framework for explaining model predictions using Shapley values

- **Formulate** a contrastive explanation question by choosing an appropriate reference distribution D
- **Approximate** the attributions relative to the reference distribution D by sampling references $(r_i)_{i=1}^N \sim D$ and computing the single-reference game attributions $(\phi_i(v_{x,r_i}))_{i=1}^N$
- **Explain** the set of attributions $(\phi_i(v_{x,r_i}))_{i=1}^N$ by appropriate summarization
  - Existing approaches summarize attributions by computing a mean
  - But, **means could be misleading** when attributions have opposite signs

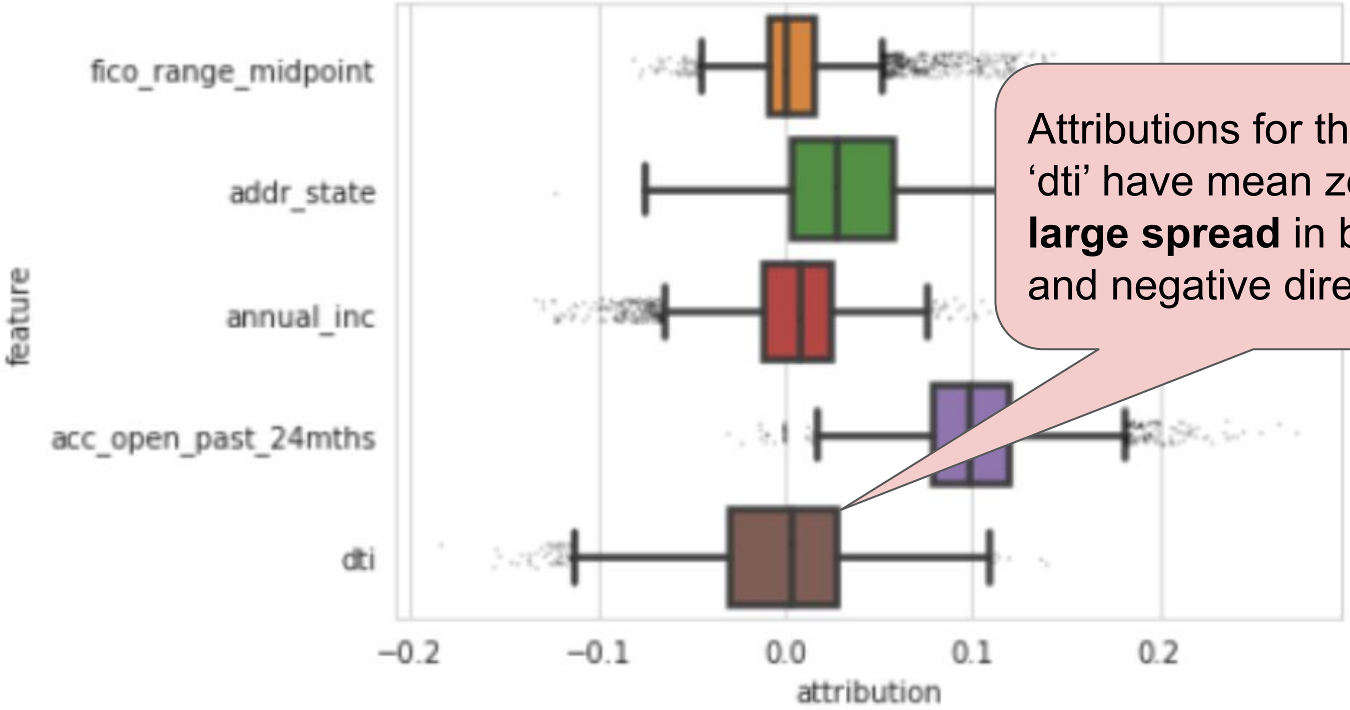
# Misleading Means

**Box plot of the attribution distribution $(\phi_i(v_{x,r_i}))_{i=1}^{N}$ for an input**



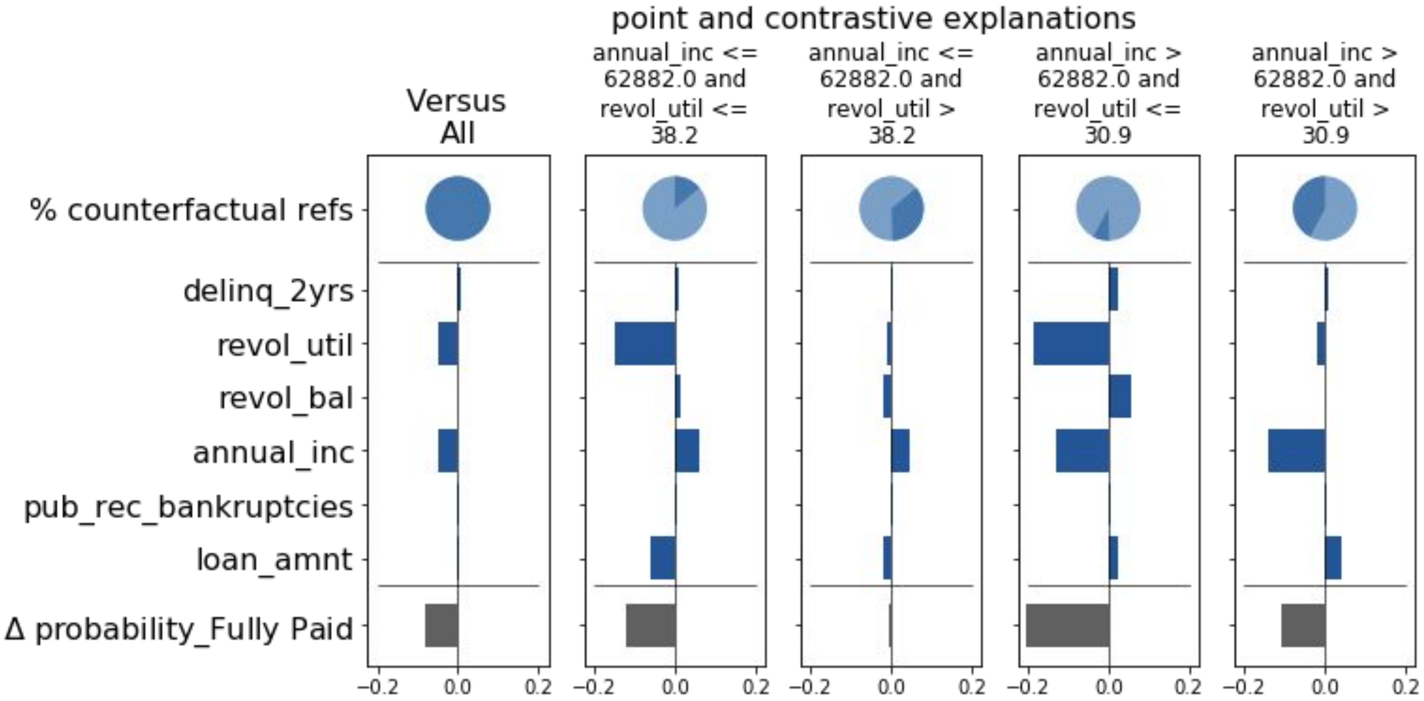
# Misleading Means

**Box plot of the attribution distribution** $(\phi_i(v_{x,r_i}))_{i=1}^{N}$ **for an input**

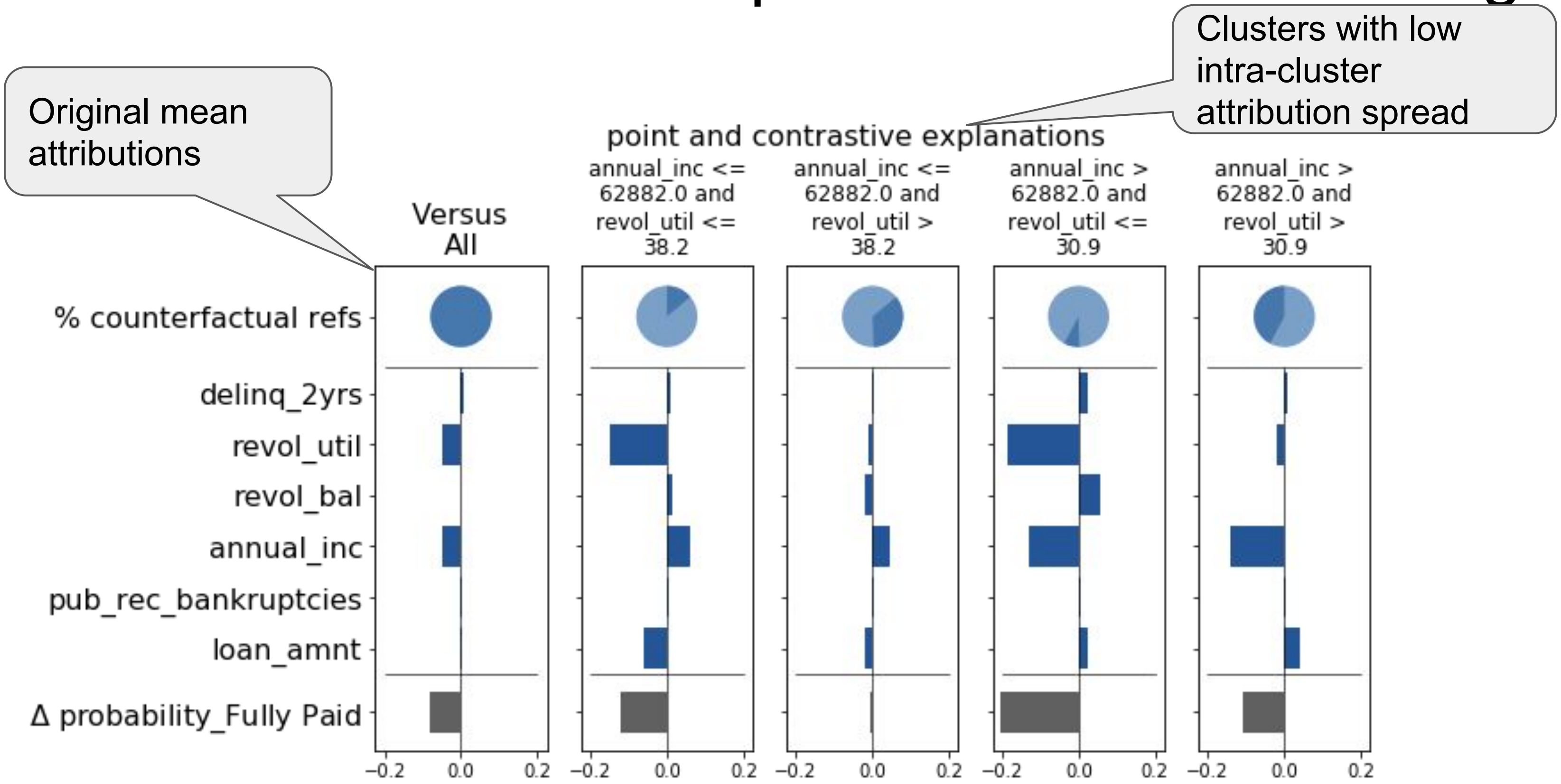



Attributions for the feature 'dti' have mean zero but a **large spread** in both positive and negative directions.

# Sneak Peak: Contrastive Explanations via Clustering



# Sneak Peak: Contrastive Explanations via Clustering



# Takeaways

- Shapley values is an **axiomatically unique method** for attributing the total gain from a cooperative game
- It has become popular tool for explaining predictions of machine learning models
- The key idea is to **formulate a cooperative game for each prediction** being explained
- There are many different game formulations in the literature, and hence **many different Shapley values**
    - See also: The many Shapley values for model explanation, arxiv 2019

# Takeaways

- **Shapley value explanations are contrastive**
  - The input at hand is contrasted with a distribution of references
  - This is well-aligned with how humans engage in explanations
- The choice of references (or norms) is an important knob for obtaining different types of explanations
- Shapley values must be interpreted in light of the references, along with rigorous quantification of any uncertainty introduced in approximating them

# References


- The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory
- A Unified Approach to Interpreting Model Predictions [SHAP and KernelSHAP]
- Algorithmic Transparency via Quantitative Input Influence [QII]
- An Efficient Explanation of Individual Classifications using Game Theory [IME]
- The many Shapley values for model explanation
- Norm Theory: Comparing Reality to Its Alternatives

# Questions?

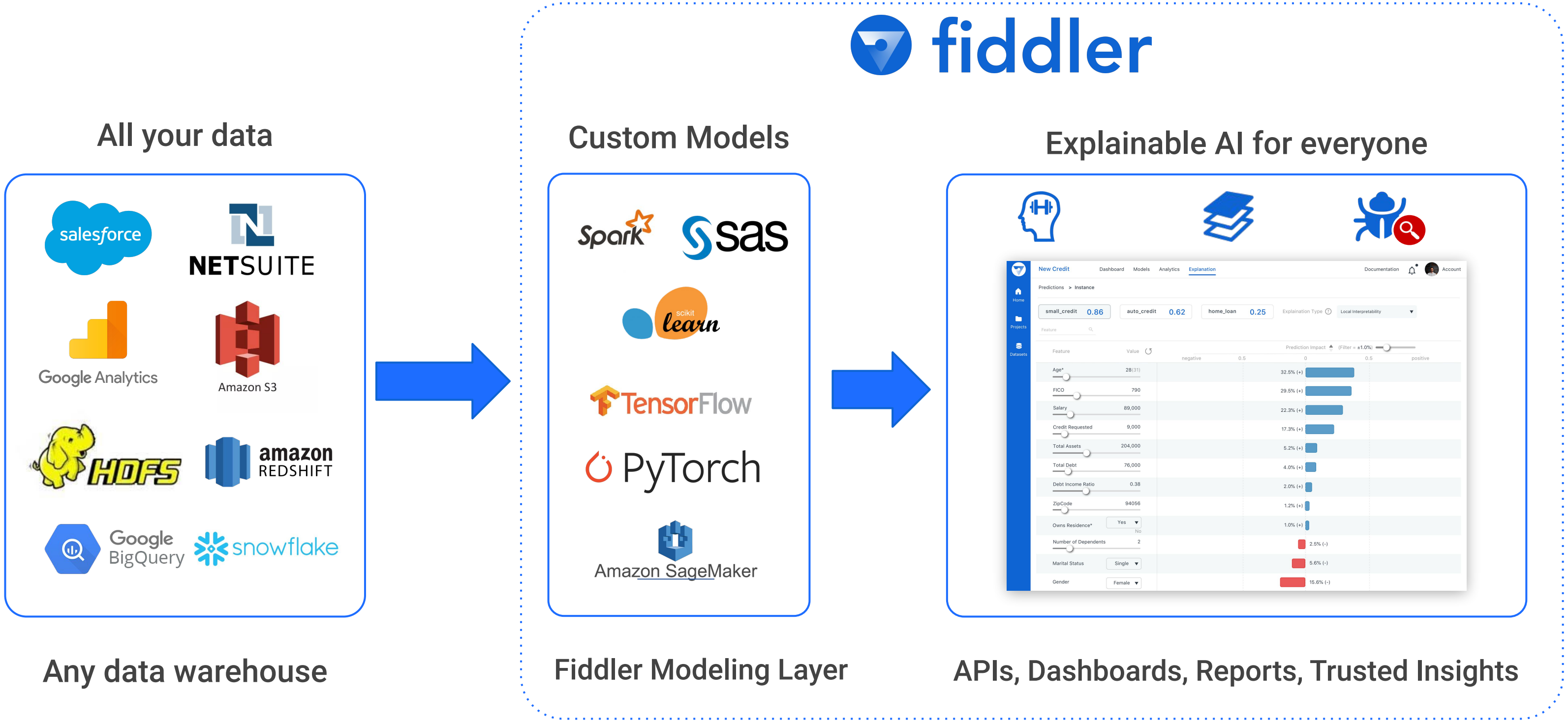
Please feel free to write to me at ankur@fiddler.ai

We are always looking for bright interns and data scientists :-)

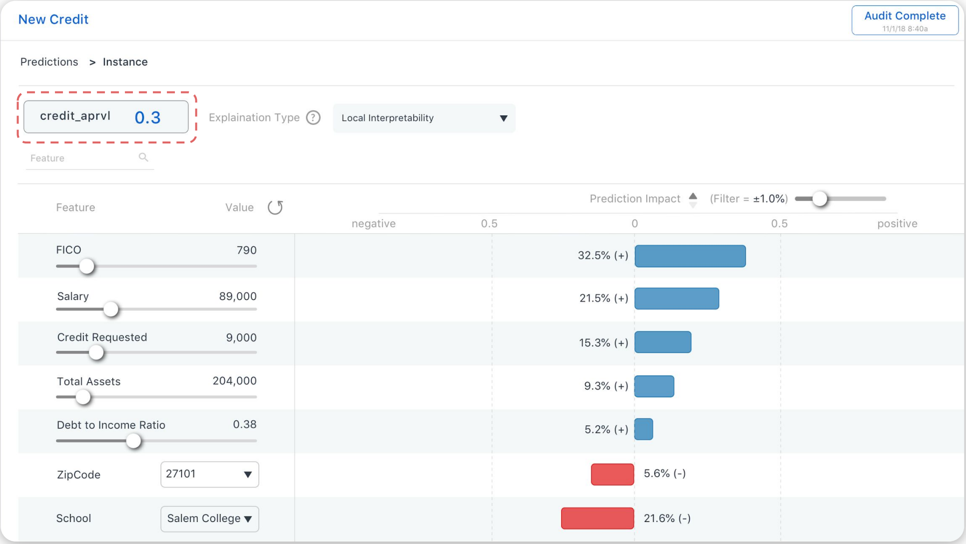
# Appendix

# Fiddler's Explainable AI Engine

**Mission**: Unlock **Trust, Visibility and Insights** by making **AI Explainable** in every enterprise



# Explain individual predictions (using Shapley Values)




How Can This Help…

**Customer Support**

Why was a customer loan rejected?

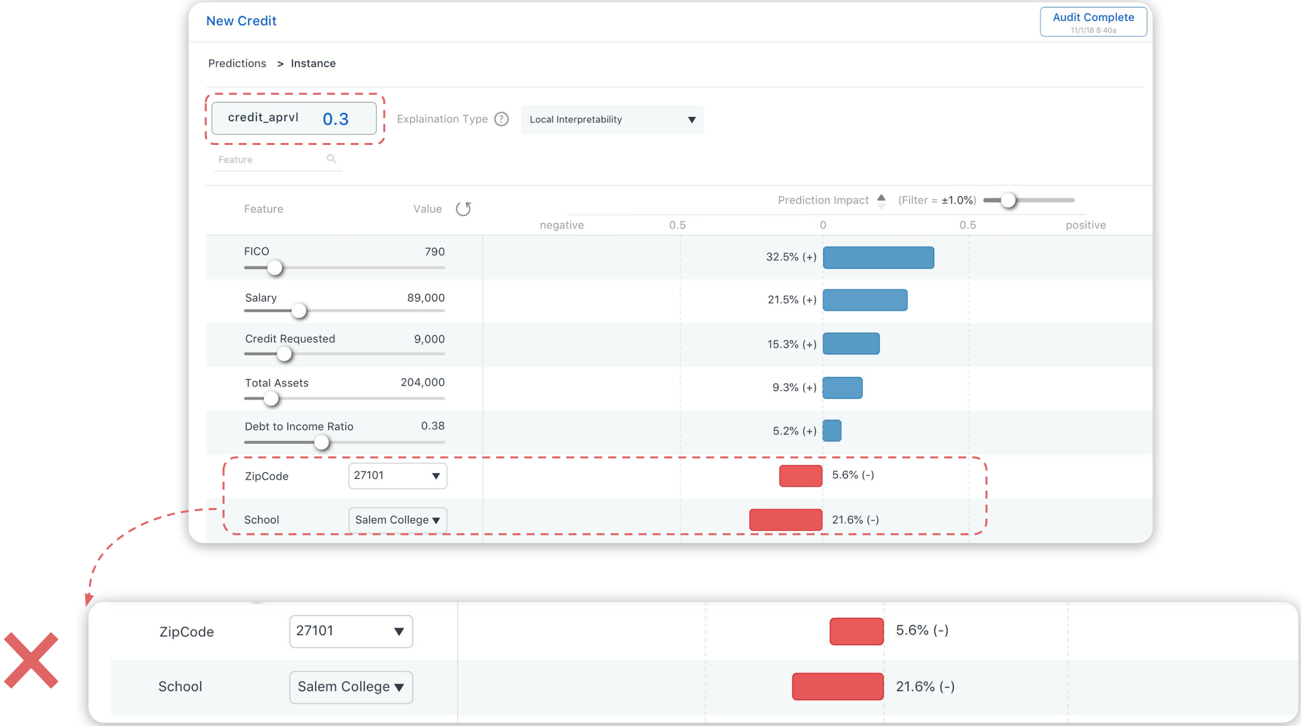
**Bias & Fairness**

How is my model doing across demographics?

**Lending LOB**

What variables should they validate with customers on “borderline” decisions?

# Explain individual predictions (using Shapley Values)




How Can This Help…

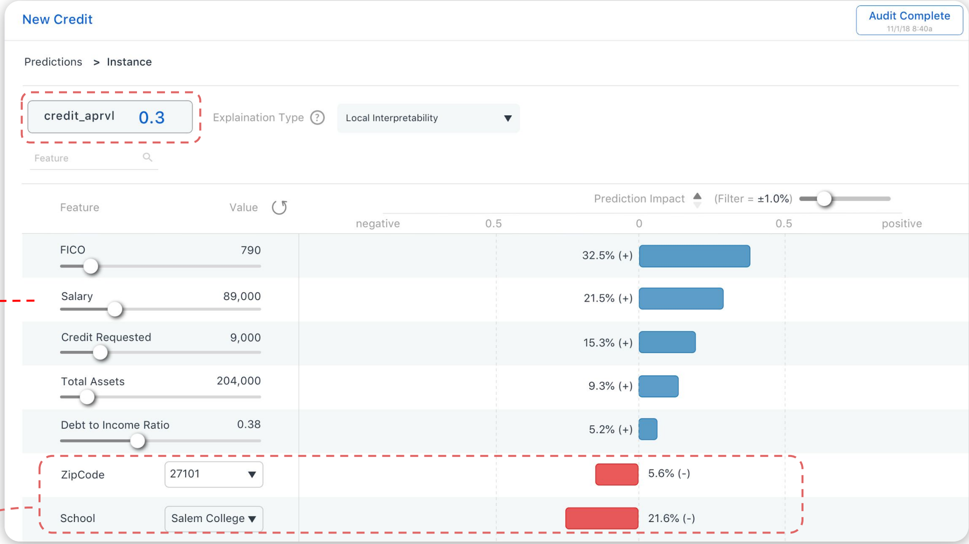
**Customer Support**
Why was a customer loan rejected?

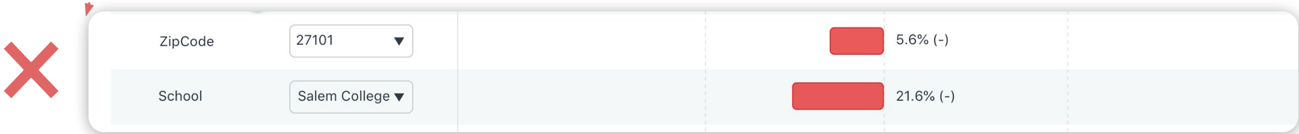
**Bias & Fairness**
How is my model doing across demographics?

**Lending LOB**
What variables should they validate with customers on "borderline" decisions?

# Explain individual predictions (using Shapley Values)

Probe the model on **counterfactuals**




How Can This Help…

**Customer Support**

Why was a customer loan rejected?

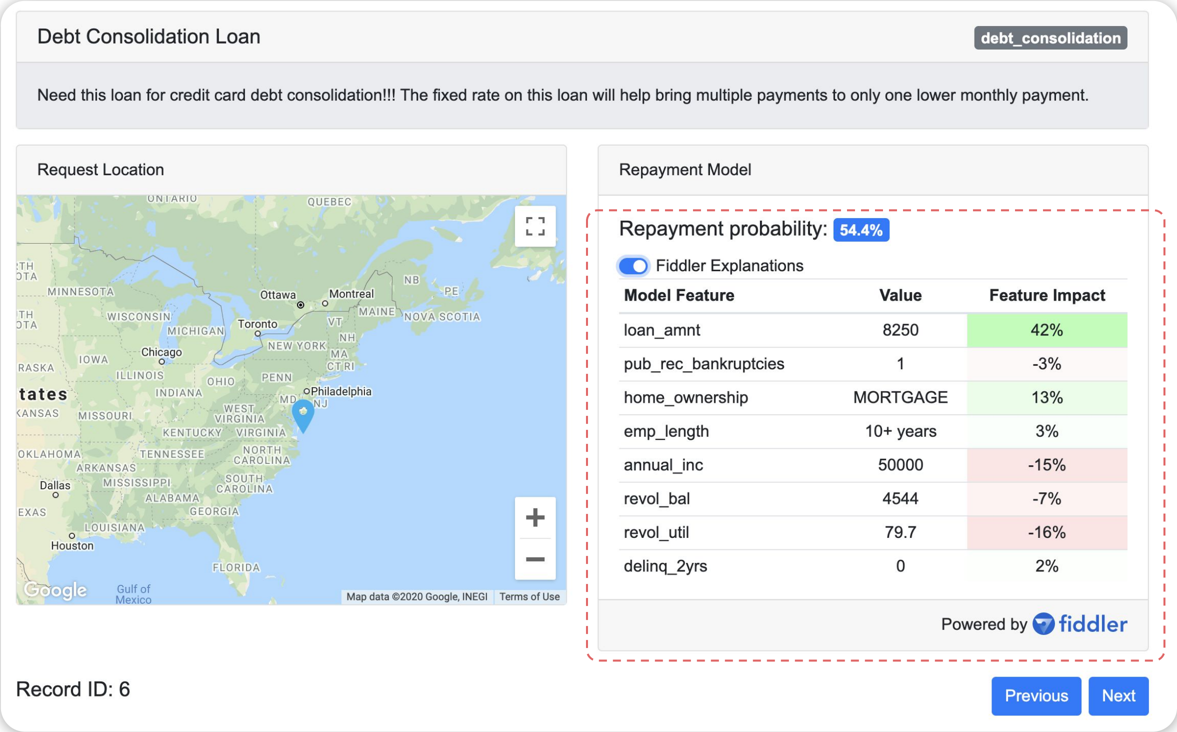
**Bias & Fairness**

How is my model doing across demographics?

**Lending LOB**

What variables should they validate with customers on “borderline” decisions?

# Integrating explanations




How Can This Help...

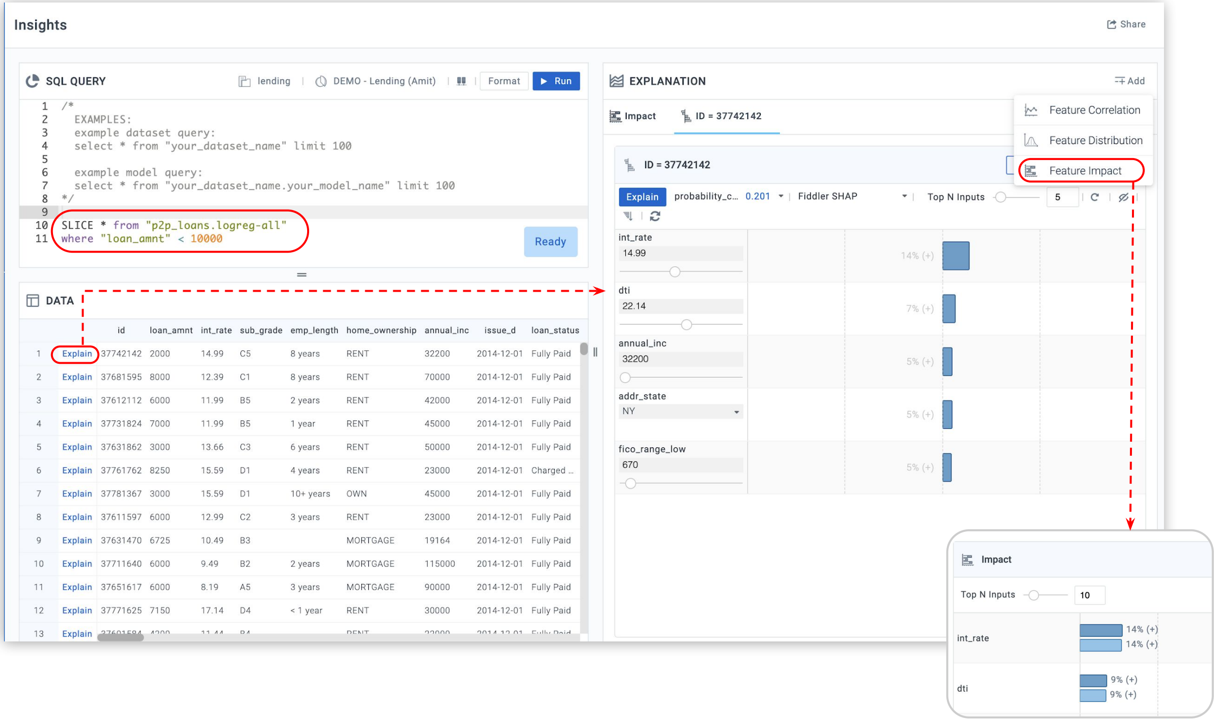
**Customer Support**

Why was a customer loan rejected?

Why was the credit card limit low?

Why was this transaction marked as fraud?

# Slice & Explain




## How Can This Help…

## Global Explanations

What are the primary feature drivers of the dataset on my model?

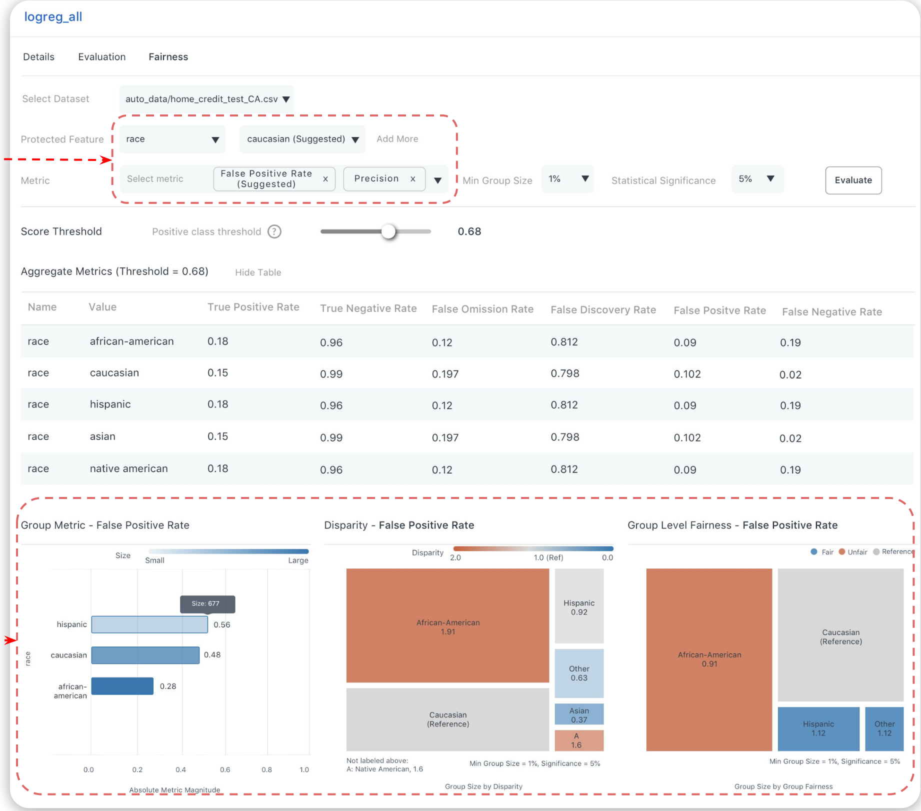
## Region Explanations

How does my model perform on a certain slice? Where does the model not perform well? Is my model uniformly fair across slices?

# Know Your Bias

Select protected feature and fairness metric

View fairness metric details




| Name | Value | True Positive Rate | True Negative Rate | False Omission Rate | False Discovery Rate | False Positve Rate | False Negative Rate |
|---|---|---|---|---|---|---|---|
| race | african-american | 0.18 | 0.96 | 0.12 | 0.812 | 0.09 | 0.19 |
| race | caucasian | 0.15 | 0.99 | 0.197 | 0.798 | 0.102 | 0.02 |
| race | hispanic | 0.18 | 0.96 | 0.12 | 0.812 | 0.09 | 0.19 |
| race | asian | 0.15 | 0.99 | 0.197 | 0.798 | 0.102 | 0.02 |
| race | native american | 0.18 | 0.96 | 0.12 | 0.812 | 0.09 | 0.19 |

How Can This Help...

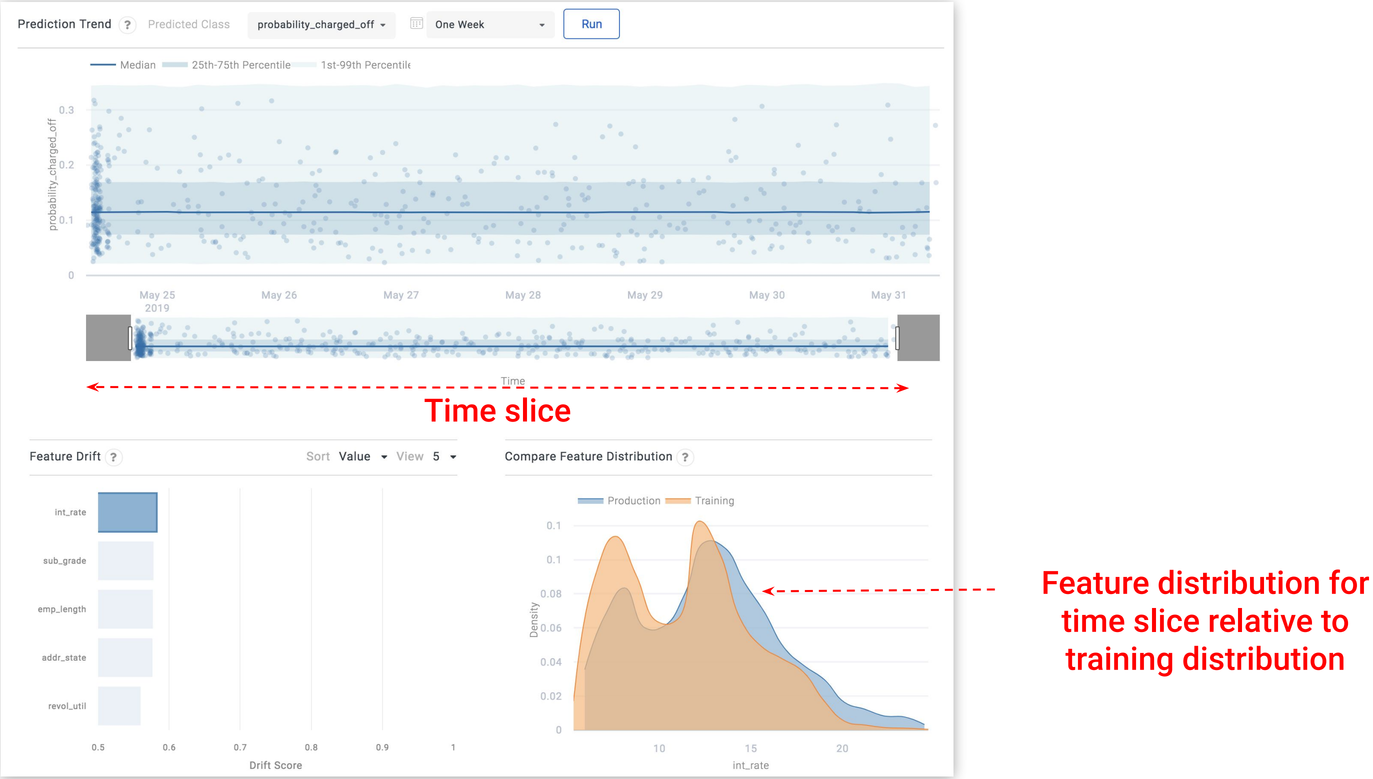
**Identify Bias**

How is my model doing across protected groups?

**Fairness Metric**

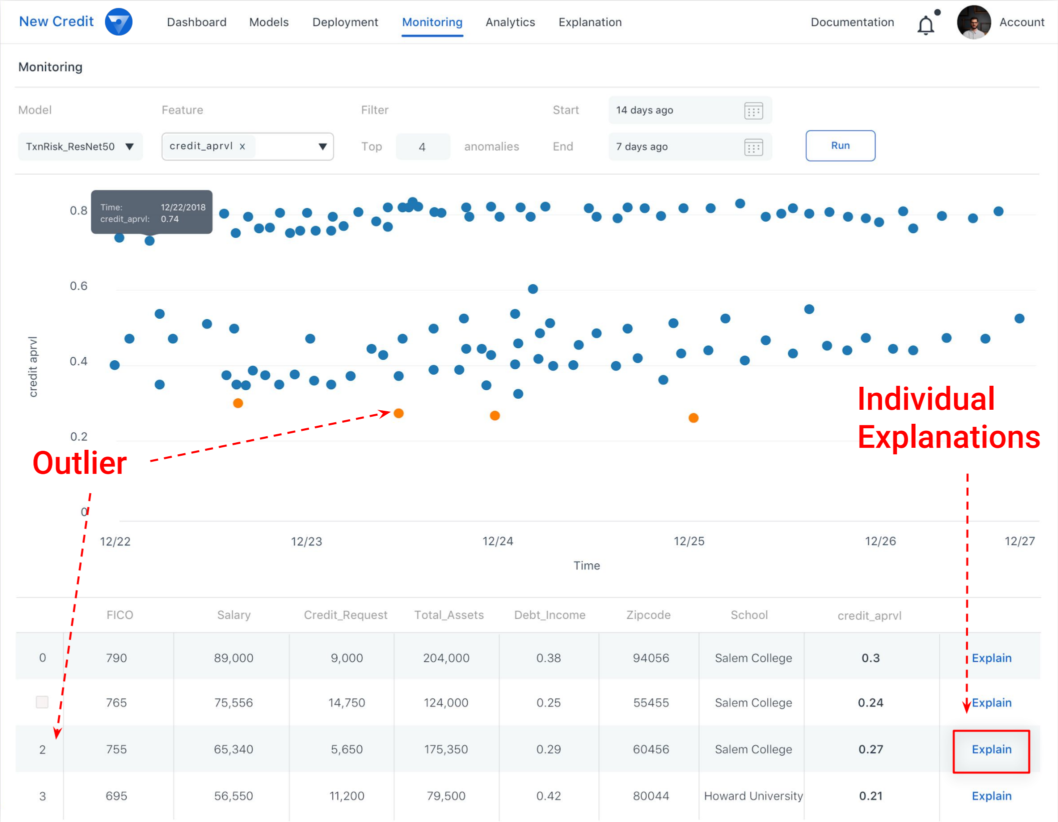
What baseline group and fairness metric is relevant?

# **Model Monitoring:** Feature Drift




*Investigate Data Drift Impacting Model Performance*

# **Model Monitoring:** Outliers with Explanations




| | FICO | Salary | Credit_Request | Total_Assets | Debt_Income | Zipcode | School | credit_aprvl | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 790 | 89,000 | 9,000 | 204,000 | 0.38 | 94056 | Salem College | 0.3 | Explain |
| | 765 | 75,556 | 14,750 | 124,000 | 0.25 | 55455 | Salem College | 0.24 | Explain |
| 2 | 755 | 65,340 | 5,650 | 175,350 | 0.29 | 60456 | Salem College | 0.27 | Explain |
| 3 | 695 | 56,550 | 11,200 | 79,500 | 0.42 | 80044 | Howard University | 0.21 | Explain |

## How Can This Help…

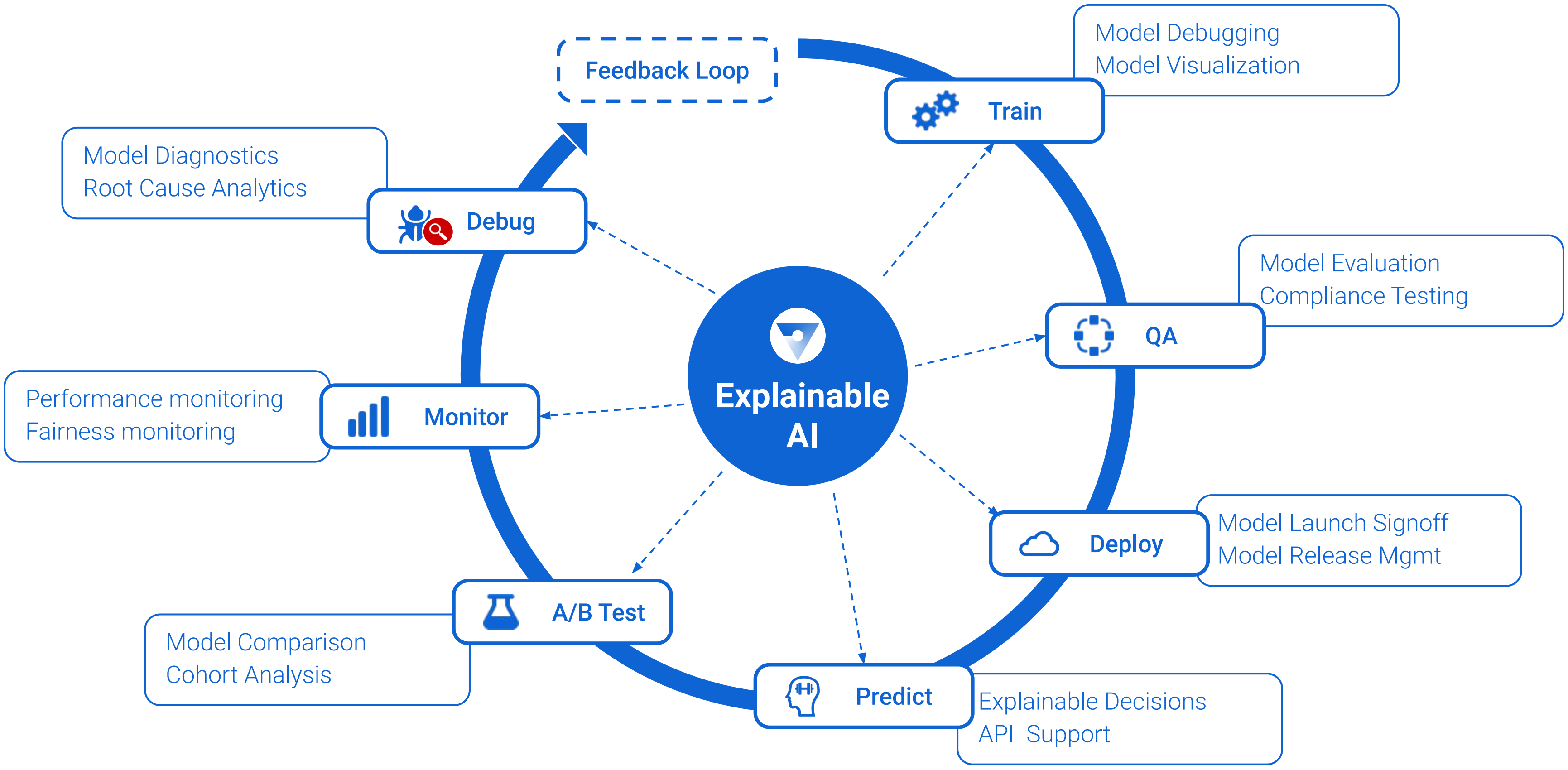
### Operations

Why are there outliers in model predictions? What caused model performance to go awry?

### Data Science

How can I improve my ML model? Where does it not do well?

# An Explainable Future



# Explainability Challenges & Tradeoffs

- Lack of standard interface for ML models makes pluggable explanations hard
- Explanation needs vary depending on the type of the user who needs it and also the problem at hand.
- The algorithm you employ for explanations might depend on the use-case, model type, data format, etc.
- There are trade-offs w.r.t. Explainability, Performance, Fairness, and Privacy.