Practical Approaches to Explainable Machine Learning

Umang Bhatt
PhD Candidate, University of Cambridge
Fellow, Mozilla Foundation
Student Fellow, Leverhulme Center for the Future of Intelligence

@umangsbhatt
usb20@cam.ac.uk

Our Starting Point

How are existing approaches to explainability used in practice?

Explainable Machine Learning in Deployment

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yuhuan Jia, Joydeep Ghosh, Ruchir Puri, José Moura, and Peter Eckersley

Appeared at the ACM Conference on Fairness, Accountability, and Transparency 2020
https://arxiv.org/abs/1909.06342

Growth of Transparency Literature

Many algorithms proposed to “explain” machine learning model output

We study how organizations use these algorithms, if at all

Our Approach

30 minute to 2 hour semi-structured interviews

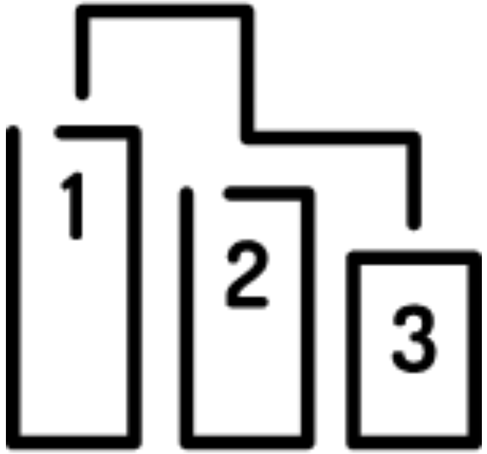
50 individuals from 30 organizations interviewed

Shared Language

- **Transparency**: Providing stakeholders with relevant information about how the model works: this includes documentation of the training procedure, analysis of training data distribution, code releases, feature-level explanations, etc.

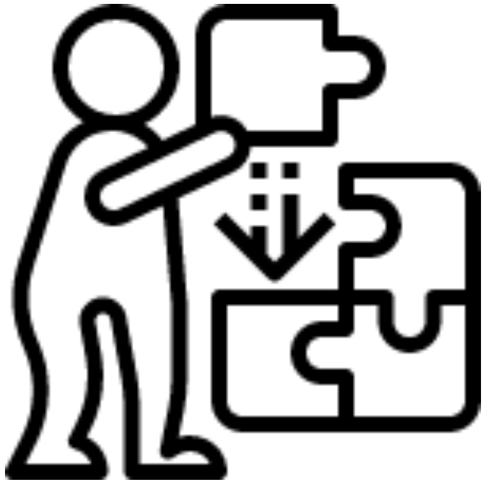
- **Explainability**: Providing insights into a model's behavior for specific datapoint(s)

Types of Explanations



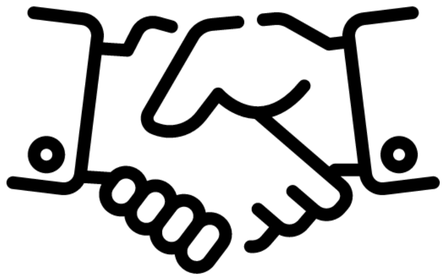
Feature Importance

Sample Importance

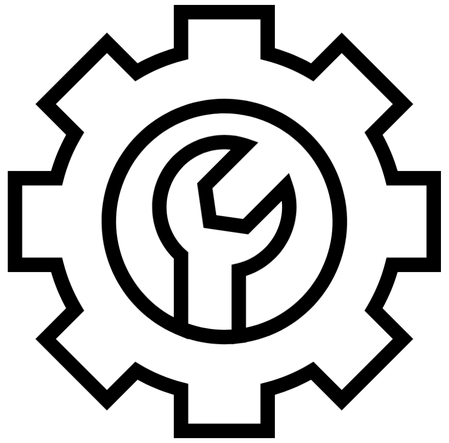


Counterfactuals

Stakeholders



Executives



Engineers

End Users

Regulators

Findings

1. Explainability is used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

Machine Learning Explainability for External Stakeholders

Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang

Convening hosted by CFI, PAI, and IBM
Appeared at the ICML 2020 Workshop on Extending Explainable AI: Beyond Deep Models and Classifiers
https://arxiv.org/abs/2007.05408

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

PARTNERSHIP ON AI

IBM

Overview

- 33 participants from 5 countries
  - 15 ML experts, 3 designers, 6 legal experts, 9 policymakers
  - Domain expertise: Finance, Healthcare, Media, and Social Services
- Goal: ***facilitate an inter-stakeholder conversation around explainable machine learning***

Community Engagement

1. *In which **context** will this explanation be used? Does the context change the properties of the explanations we expose?*

2. *How should the explanation be **evaluated**? Both quantitatively and qualitatively…*

3. *Can we prevent data misuse and preferential treatment by involving **affected groups** in the development process?*

4. *Can we **educate** external stakeholders (and data scientists) regarding the functionalities and limitations of explainable machine learning?*

Deploying Explainability

1. *How does **uncertainty** in the model and introduced by the (potentially approximate) explanation technique affect the resulting explanations?*

2. *How can stakeholders **interact** with the resulting explanations? Can explanations be a conduit for interacting with the model?*

3. *How, if at all, will stakeholder **behavior** change as a result of the explanation shown?*

4. *Over **time**, how will the explanation technique adapt to changes in stakeholder behavior?*

Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

Evaluating and Aggregating Feature-based Model Explanations

Umang Bhatt, Adrian Weller, and José Moura

Appeared at the International Joint Conference on Artificial Intelligence 2020
https://arxiv.org/abs/2005.00631

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \{C_1 \ldots C_j\}$

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \{C_1 \ldots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \{C_1 \ldots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)
  - Which feature(s) x_i contributes to classifying $\boldsymbol{x}$ with $\boldsymbol{f}$

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \{C_1 \ldots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)
  - Which feature(s) x_i contributes to classifying $\boldsymbol{x}$ with $\boldsymbol{f}$
  - Let $\boldsymbol{g}$ be our explanation technique and let $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})$ be the explanation for $\boldsymbol{f}(\boldsymbol{x})$

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \to \{C_1 \dots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)
  - Which feature(s) x_i contributes to classifying $\boldsymbol{x}$ with $\boldsymbol{f}$
  - Let $\boldsymbol{g}$ be our explanation technique and let $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})$ be the explanation for $\boldsymbol{f}(\boldsymbol{x})$
- Feature Importance: $\boldsymbol{g}_m = |\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})| \in \mathbb{R}^d_{\geq 0}$

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \{C_1 \ldots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)
  - Which feature(s) x_i contributes to classifying $\boldsymbol{x}$ with $\boldsymbol{f}$
  - Let $\boldsymbol{g}$ be our explanation technique and let $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})$ be the explanation for $\boldsymbol{f}(\boldsymbol{x})$
- Feature Importance: $\boldsymbol{g}_m = |\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})| \in \mathbb{R}^d_{\geq 0}$
  - Normalize $\boldsymbol{g}_m$ to induce a probability distribution over the features in the input space

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \to \{C_1 \dots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)
  - Which feature(s) x_i contributes to classifying $\boldsymbol{x}$ with $\boldsymbol{f}$
  - Let $\boldsymbol{g}$ be our explanation technique and let $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})$ be the explanation for $\boldsymbol{f}(\boldsymbol{x})$
- Feature Importance: $\boldsymbol{g}_m = |\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})| \in \mathbb{R}^d_{\geq 0}$
  - Normalize $\boldsymbol{g}_m$ to induce a probability distribution over the features in the input space
- Feature Contribution: $\boldsymbol{g}_c = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}) \in \mathbb{R}^d$ and $|\boldsymbol{g}_c| = \boldsymbol{g}_m$

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \to \{C_1 \dots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)
  - Which feature(s) x_i contributes to classifying $\boldsymbol{x}$ with $\boldsymbol{f}$
  - Let $\boldsymbol{g}$ be our explanation technique and let $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})$ be the explanation for $\boldsymbol{f}(\boldsymbol{x})$
- Feature Importance: $\boldsymbol{g}_m = |\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})| \in \mathbb{R}^d_{\geq 0}$
  - Normalize $\boldsymbol{g}_m$ to induce a probability distribution over the features in the input space
- Feature Contribution: $\boldsymbol{g}_c = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}) \in \mathbb{R}^d$ and $|\boldsymbol{g}_c| = \boldsymbol{g}_m$
  - Presence of rain increases the chance of precipitation $(+)$

Feature Importance Desiderata

Explain $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \to \{C_1 \dots C_j\}$

- Restrict ourselves to feature-based explanations in the classification setting (i.e. attributions, saliency maps, etc.)
  - Which feature(s) x_i contributes to classifying $\boldsymbol{x}$ with $\boldsymbol{f}$
  - Let $\boldsymbol{g}$ be our explanation technique and let $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})$ be the explanation for $\boldsymbol{f}(\boldsymbol{x})$
- Feature Importance: $\boldsymbol{g}_m = |\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})| \in \mathbb{R}^d_{\geq 0}$
  - Normalize $\boldsymbol{g}_m$ to induce a probability distribution over the features in the input space
- Feature Contribution: $\boldsymbol{g}_c = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}) \in \mathbb{R}^d$ and $|\boldsymbol{g}_c| = \boldsymbol{g}_m$
  - Presence of rain increases the chance of precipitation $(+)$
  - Presence of clouds detracts from the chance of sunny weather $(-)$

Common feature-based explanations

LIME (Ribeiro et al. KDD 2016)

$$\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \underset{\boldsymbol{g} \in \mathcal{G}}{\arg\min}\ \mathcal{L}(\boldsymbol{f}, \boldsymbol{g}, \pi_x) + \Omega(\boldsymbol{g})$$

Common feature-based explanations

LIME (Ribeiro et al. KDD 2016)

$$\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \ \mathcal{L}(\boldsymbol{f}, \boldsymbol{g}, \pi_x) + \Omega(\boldsymbol{g})$$

Local surrogate model, $\boldsymbol{g}$, to approximate f in some kernelized region π_x, and encourages **sparsity** by keeping model complexity, $\Omega(\boldsymbol{g})$, low

SHAP (Lundberg and Lee. NeurIPS 2017)

$$\boldsymbol{g}_{ci} = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \phi_i = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{i\}} \binom{F-1}{S}^{-1} \left(\boldsymbol{f}(\boldsymbol{x}_{S \cup \{i\}}) - \boldsymbol{f}(\boldsymbol{x}_S) \right)$$

Considers contribution over all features, as Shapley values satisfy **efficiency**, symmetry, additivity, and dummy (zero).

Common feature-based explanations

LIME (Ribeiro et al. KDD 2016)

$$\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \ \mathcal{L}(\boldsymbol{f}, \boldsymbol{g}, \pi_x) + \Omega(\boldsymbol{g})$$

Local surrogate model, $\boldsymbol{g}$, to approximate f in some kernelized region π_x, and encourages **sparsity** by keeping model complexity, $\Omega(\boldsymbol{g})$, low

SHAP (Lundberg and Lee. NeurIPS 2017)

$$\boldsymbol{g}_{ci} = \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \phi_i = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{i\}} \binom{F-1}{S}^{-1} \left(\boldsymbol{f}(\boldsymbol{x}_{S \cup \{i\}}) - \boldsymbol{f}(\boldsymbol{x}_S) \right)$$

Considers contribution over all features, as Shapley values satisfy **efficiency**, symmetry, additivity, and dummy (zero).

Integrated Gradients (Sundarajan et al. ICML 2017)

Accumulates the gradients along a straight line path between $\boldsymbol{x}$ and $\bar{x}$, where $\boldsymbol{f}(\bar{x}) \approx 0$, and satisfies **completeness**, $\sum_{i=1}^{d} \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(x)$.

Evaluating explanations

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z})\leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z})d\boldsymbol{z}$$

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z})) \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z}) d\boldsymbol{z}$$

$$\mu_{\text{MAX}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \max_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))$$

Let D be the distance between explanations and ρ be the distance between inputs

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z})) \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z}) d\boldsymbol{z}$$

$$\mu_{\text{MAX}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \max_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))$$

Let D be the distance between explanations and ρ be the distance between inputs

Faithfulness

Does the explanation capture features important to the prediction?

Evaluating explanations

Sensitivity

Do similar inputs have similar explanations?

$$\mu_{\text{AVG}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \int_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z})d\boldsymbol{z}$$

$$\mu_{\text{MAX}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, r) = \max_{\rho(\boldsymbol{x},\boldsymbol{z}) \leq r} D(\boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{z}))$$

Let D be the distance between explanations and ρ be the distance between inputs

Faithfulness

Does the explanation capture features important to the prediction?

$$\mu_{\text{F}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}, S) = \text{corr}(\tfrac{1}{|S|}\textstyle\sum_{i \in S} \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i, \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}))$$

Fix a subset size and randomly sample subsets of that size from $\boldsymbol{x}$ to estimate the Pearson Correlation Coefficient

Evaluating explanations (cont.)

Evaluating explanations (cont.)

Complexity

Is the explanation digestible?

Evaluating explanations (cont.)

Complexity

Is the explanation digestible? We define an attribution contribution distribution:

$$\mathbb{P}_A = \left\{ \frac{|\boldsymbol{g}(\boldsymbol{x})_1|}{\sum_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \frac{|\boldsymbol{g}(\boldsymbol{x})_2|}{\sum_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \ldots, \frac{|\boldsymbol{g}(\boldsymbol{x})_d|}{\sum_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|} \right\}$$

$$\mu_{\mathsf{C}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}) = H(\boldsymbol{x}) = \mathbb{E}_i\big[-\ln(\mathbb{P}_A(i))\big] = -\sum_{i=1}^{d} \mathbb{P}_A(i) \ln(\mathbb{P}_A(i))$$

Evaluating explanations (cont.)

Complexity

Is the explanation digestible? We define an attribution contribution distribution:

$$\mathbb{P}_A = \left\{ \frac{|\boldsymbol{g}(\boldsymbol{x})_1|}{\sum_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \frac{|\boldsymbol{g}(\boldsymbol{x})_2|}{\sum_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|}, \ldots, \frac{|\boldsymbol{g}(\boldsymbol{x})_d|}{\sum_{j\in[d]} |\boldsymbol{g}(\boldsymbol{x})_j|} \right\}$$

$$\mu_{\mathsf{C}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{x}) = H(\boldsymbol{x}) = \mathbb{E}_i\big[-\ln(\mathbb{P}_A(i))\big] = -\sum_{i=1}^{d} \mathbb{P}_A(i) \ln(\mathbb{P}_A(i))$$

The least complex explanation is one where $\boldsymbol{g}(\boldsymbol{x})_i = 1$ and the most complex explanation is one where $\boldsymbol{g}(\boldsymbol{x})_i = \frac{1}{d}$.

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Three candidate methods for $h(\cdot)$.

- Convex Combination: $\boldsymbol{g}_{agg} = w\boldsymbol{g}_1 + (1 - w)\boldsymbol{g}_2$

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Three candidate methods for $h(\cdot)$.

- Convex Combination: $\boldsymbol{g}_{agg} = w\boldsymbol{g}_1 + (1 - w)\boldsymbol{g}_2$
- Centroid Aggregation: $\boldsymbol{g}_{\text{agg}} \in \arg\min_{\boldsymbol{g} \in \mathcal{G}} \sum_{i=1}^{m} d(\boldsymbol{g}, \boldsymbol{g}_i)$

Aggregating Existing Techniques

Can we learn an aggregate explanation of existing techniques that does better with respect to a criterion of interest? An approach to study $\boldsymbol{g}_{\text{agg}}$ can be to set the problem up as follows:

$$\boldsymbol{g}_{\text{agg}} = \arg\max_{\boldsymbol{g} \in \mathcal{G}} \mu(\boldsymbol{f}, \boldsymbol{g}), \text{ s.t. } \boldsymbol{g} = h(\mathcal{G}_m)$$

Three candidate methods for $h(\cdot)$.

- Convex Combination: $\boldsymbol{g}_{agg} = w\boldsymbol{g}_1 + (1-w)\boldsymbol{g}_2$
- Centroid Aggregation: $\boldsymbol{g}_{\text{agg}} \in \arg\min_{\boldsymbol{g} \in \mathcal{G}} \sum_{i=1}^{m} d(\boldsymbol{g}, \boldsymbol{g}_i)$
- Bayesian Optimization: $\max_{\boldsymbol{g}_{agg} \in \mathcal{G}} \mu(\boldsymbol{g}_{agg})$ where

$$k(\boldsymbol{g}_i, \boldsymbol{g}_j) = \mathbb{E}_{x \sim \mathcal{D}_x}\left[k(\boldsymbol{g}_i(\boldsymbol{x}), \boldsymbol{g}_j(\boldsymbol{x}))\right] = \mathbb{E}_{x \sim \mathcal{D}_x}\left[e^{-\frac{1}{2}||\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_j(\boldsymbol{x})||^2}\right]$$

Generalized Aggregation

Convex Combination

$$\boldsymbol{g}_{agg} = w^T G$$

$$G^T = \begin{pmatrix} | & | & & | \\ \text{SHAP}_1 & \text{LIME}_2 & \cdots & \text{IG}_m \\ | & | & & | \end{pmatrix}$$

Generalized Aggregation

Convex Combination

$$\boldsymbol{g}_{agg} = w^T G$$

$$G^T = \begin{pmatrix} | & | & & | \\ \text{SHAP}_1 & \text{LIME}_2 & \cdots & \text{IG}_m \\ | & | & & | \end{pmatrix}$$

$$w_{agg} \in \arg\max_{w} \sum_{i=1}^{m} \mu(w^T G_i)$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \ldots IG_m\}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \ldots IG_m\}$$

$$\boldsymbol{g}_c^S = \begin{bmatrix} 1 & -2 & 7 \end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix} .1 & .2 & .7 \end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix} C & B & A \end{bmatrix}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \ldots IG_m\}$$

$$\boldsymbol{g}_c^S = \begin{bmatrix} 1 & -2 & 7 \end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix} .1 & .2 & .7 \end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix} C & B & A \end{bmatrix}$$

$$\text{rank}^{S_1} = \begin{bmatrix} C & B & A \end{bmatrix} \quad \text{rank}^{S_2} = \begin{bmatrix} C & A & B \end{bmatrix} \quad \text{rank}^{S_3} = \begin{bmatrix} A & B & C \end{bmatrix}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \ldots IG_m\}$$

$$\boldsymbol{g}_c^S = \begin{bmatrix} 1 & -2 & 7 \end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix} .1 & .2 & .7 \end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix} C & B & A \end{bmatrix}$$

$$\text{rank}^{S_1} = \begin{bmatrix} C & B & A \end{bmatrix} \quad \text{rank}^{S_2} = \begin{bmatrix} C & A & B \end{bmatrix} \quad \text{rank}^{S_3} = \begin{bmatrix} A & B & C \end{bmatrix}$$

$$\text{Borda Count: } \boldsymbol{g}_{agg} = \text{rank}^{agg} = \begin{bmatrix} C & A & B \end{bmatrix}$$

Generalized Aggregation

Classical Rank Aggregation

$$\mathcal{G}_m = \{SHAP_1, LIME_2, \ldots IG_m\}$$

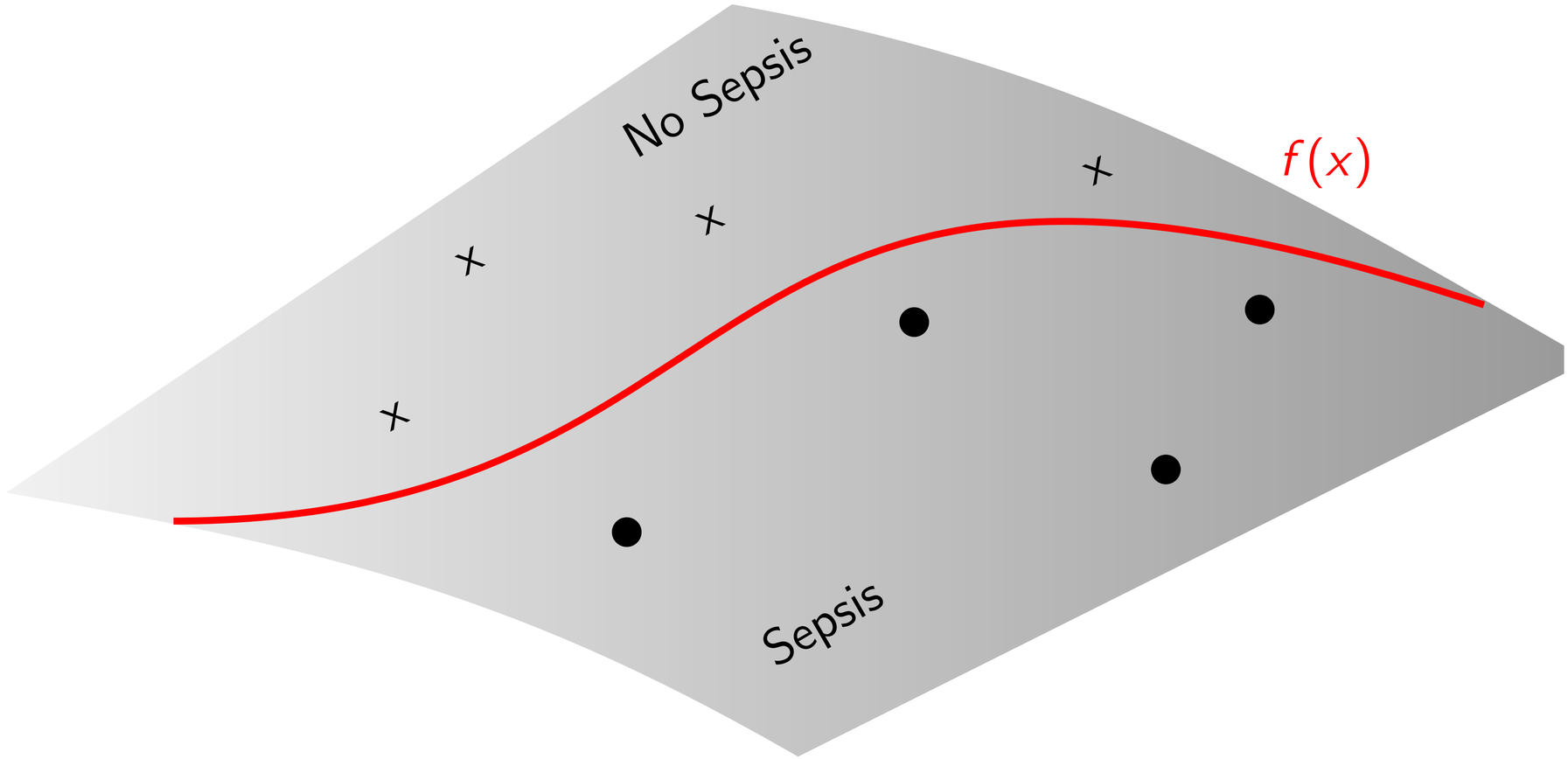
$$\boldsymbol{g}_c^S = \begin{bmatrix} 1 & -2 & 7 \end{bmatrix} \rightarrow \boldsymbol{g}_m^S = \begin{bmatrix} .1 & .2 & .7 \end{bmatrix} \rightarrow \text{rank}^S = \begin{bmatrix} C & B & A \end{bmatrix}$$

$$\text{rank}^{S_1} = \begin{bmatrix} C & B & A \end{bmatrix} \quad \text{rank}^{S_2} = \begin{bmatrix} C & A & B \end{bmatrix} \quad \text{rank}^{S_3} = \begin{bmatrix} A & B & C \end{bmatrix}$$

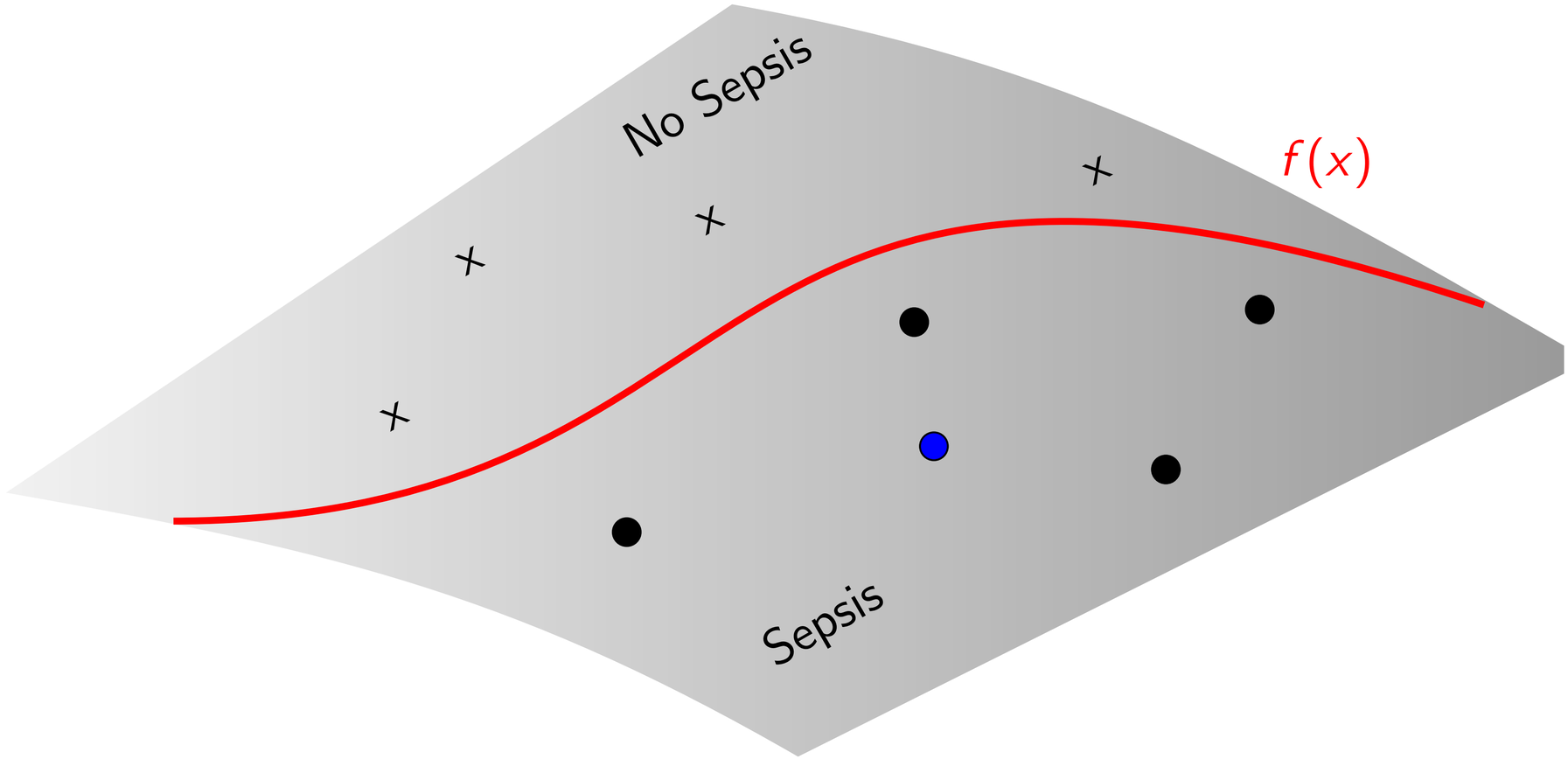
$$\text{Borda Count: } \boldsymbol{g}_{agg} = \text{rank}^{agg} = \begin{bmatrix} C & A & B \end{bmatrix}$$

$$\boldsymbol{g}_{agg} \in \arg\min_{\boldsymbol{g}} \sum_{\boldsymbol{g}_i \in \mathcal{G}_m} d(\boldsymbol{g}, \boldsymbol{g}_i)$$

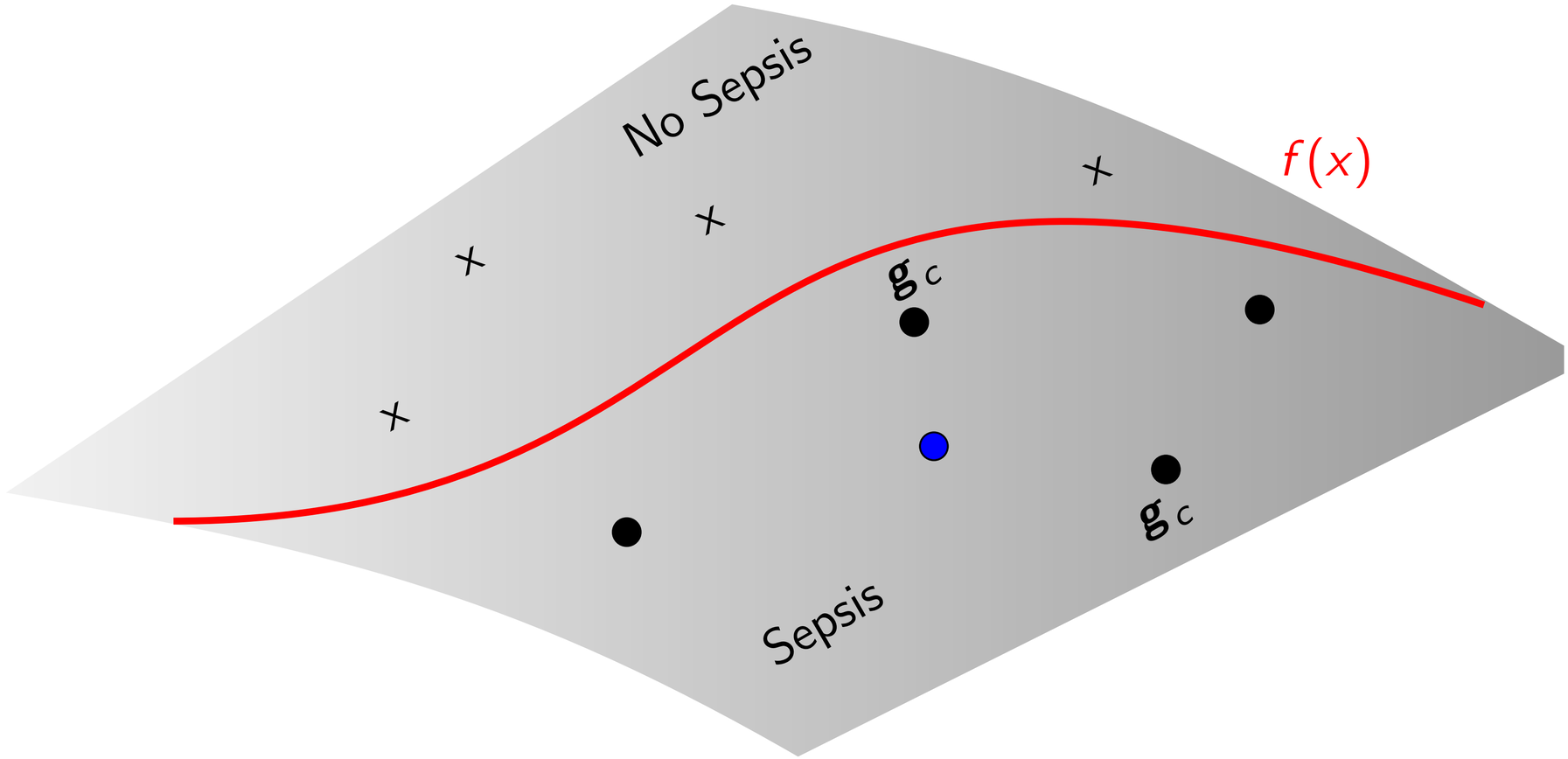
Aggregating Local Explanations



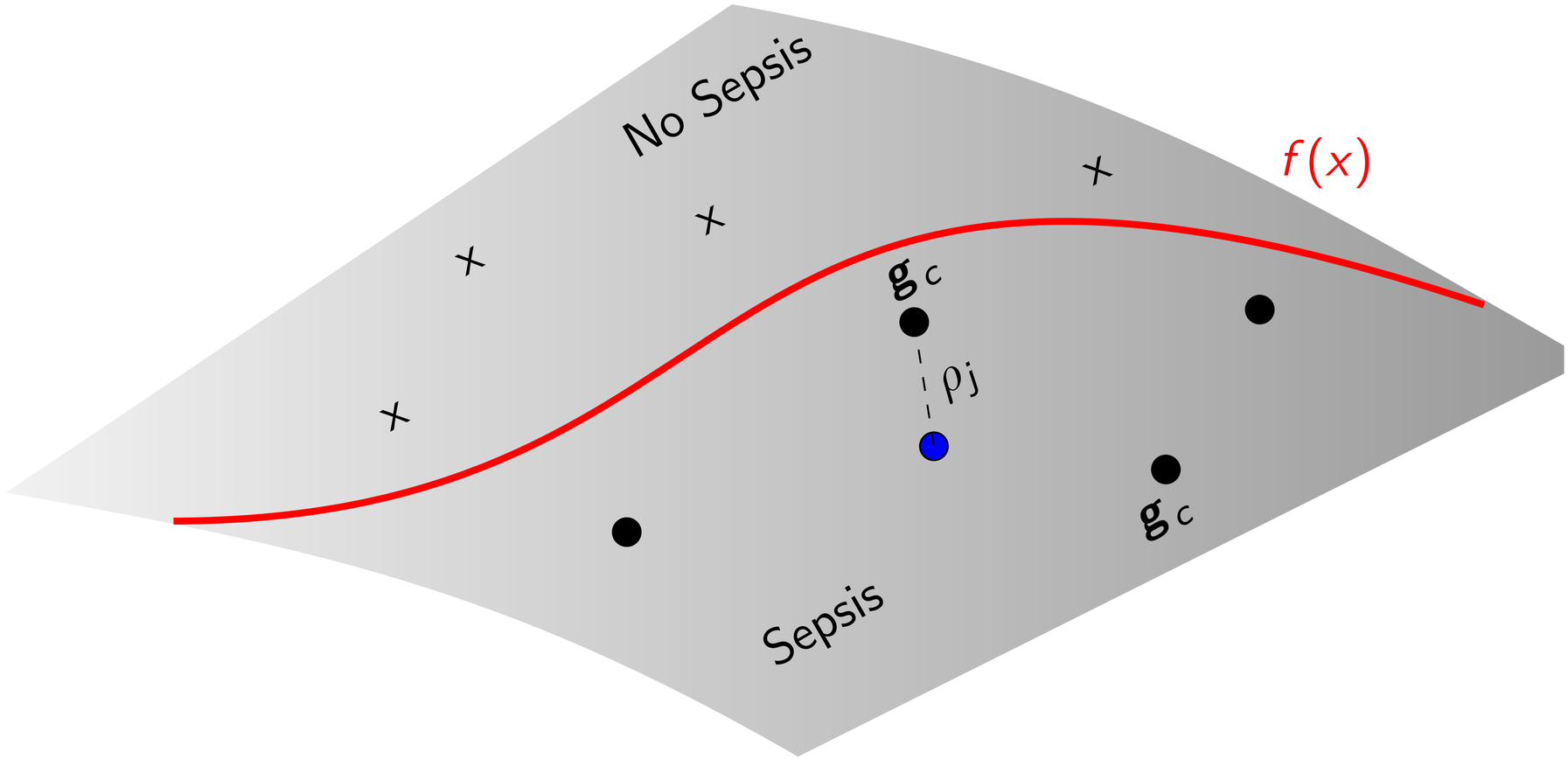
Aggregating Local Explanations



Aggregating Local Explanations



Aggregating Local Explanations



AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

1. Find k nearest neighbors, $\mathcal{N}_k$, of x_{test} and their weights, ρ_j

$$\rho_j = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon,x^{(j)}}, x_{\text{test}})\Big|_{\epsilon=0}$$

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N}\subset\mathcal{D},|\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)}\in\mathcal{N}} \rho_j$$

AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

1. Find k nearest neighbors, $\mathcal{N}_k$, of x_{test} and their weights, ρ_j

$$\rho_j = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon,x^{(j)}}, x_{\text{test}})\big|_{\epsilon=0}$$

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N}\subset\mathcal{D},|\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)}\in\mathcal{N}} \rho_j$$

2. Calculate the attributions, $\boldsymbol{g}_c$, for all points in $\mathcal{N}_k$

$$\boldsymbol{g}_{ci} = \phi_i = \frac{1}{|F|} \sum_{S\subseteq F\setminus\{i\}} \binom{F-1}{S}^{-1} \left(f(x_{S\cup\{i\}}) - f(x_S)\right)$$

AVA: Aggregate Valuation of Antecedents

Can we use weighted Shapley values (Kalai et al. Journal of Game Theory 1987) to aggregate feature-based explanations with lower sensitivity?

1. Find k nearest neighbors, $\mathcal{N}_k$, of x_{test} and their weights, ρ_j

$$\rho_j = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon,x^{(j)}}, x_{\text{test}})\big|_{\epsilon=0}$$

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N}\subset\mathcal{D},|\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)}\in\mathcal{N}} \rho_j$$

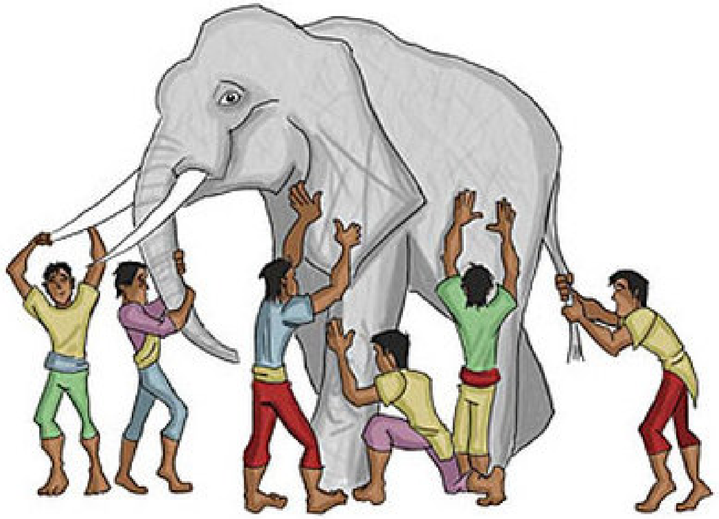
2. Calculate the attributions, $\boldsymbol{g}_c$, for all points in $\mathcal{N}_k$

$$\boldsymbol{g}_{ci} = \phi_i = \frac{1}{|F|} \sum_{S\subseteq F\setminus\{i\}} \binom{F-1}{S}^{-1} \left(f(x_{S\cup\{i\}}) - f(x_S)\right)$$

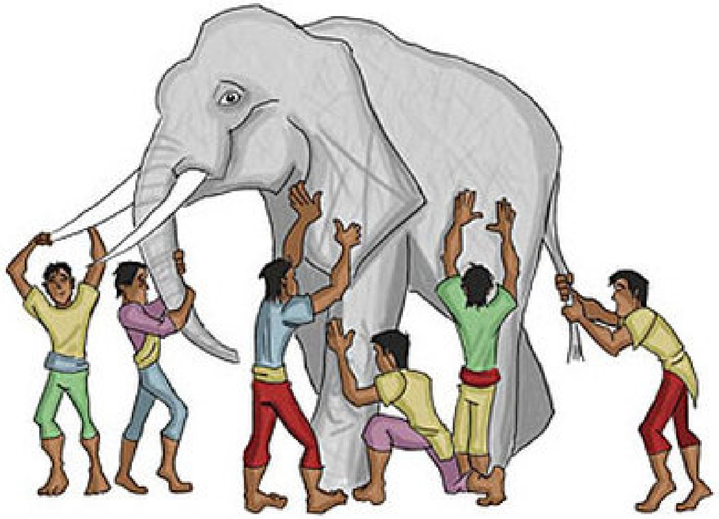
3. Aggregate the k explanations into a consensus, $\boldsymbol{g}_{agg}$

$$\boldsymbol{g}_{agg} = \sum_{x^{(j)}\in\mathcal{N}_k} \frac{\rho_j}{\rho} \boldsymbol{g}_c^j$$

Why aggregate?

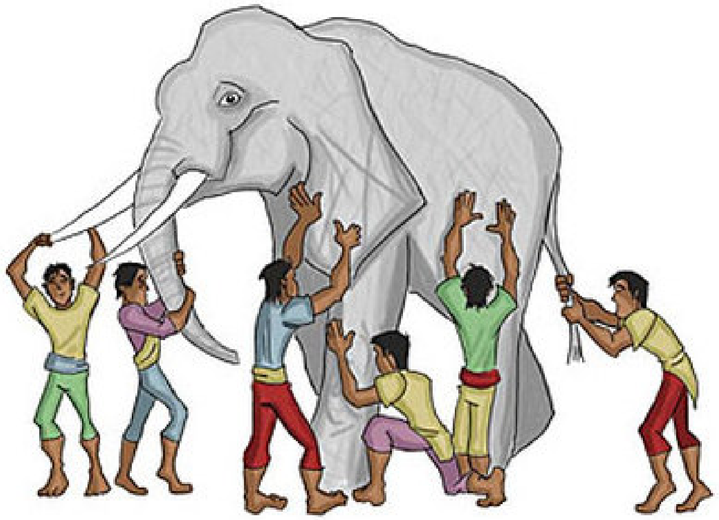


Why aggregate?



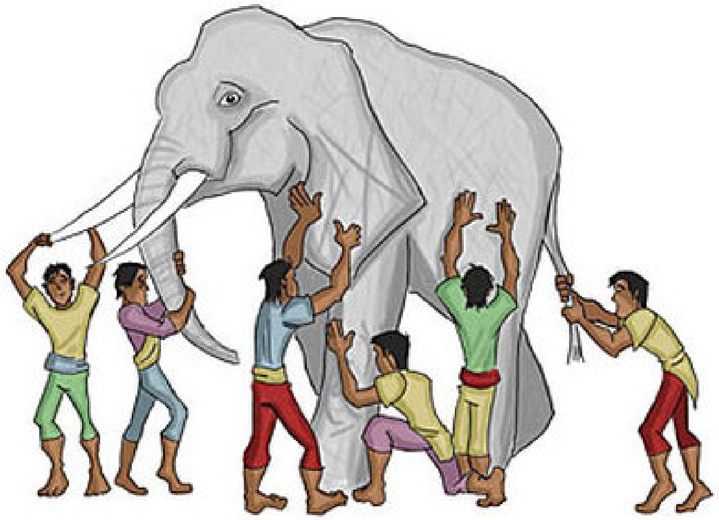
- Each training point has its own “learned” attribution

Why aggregate?



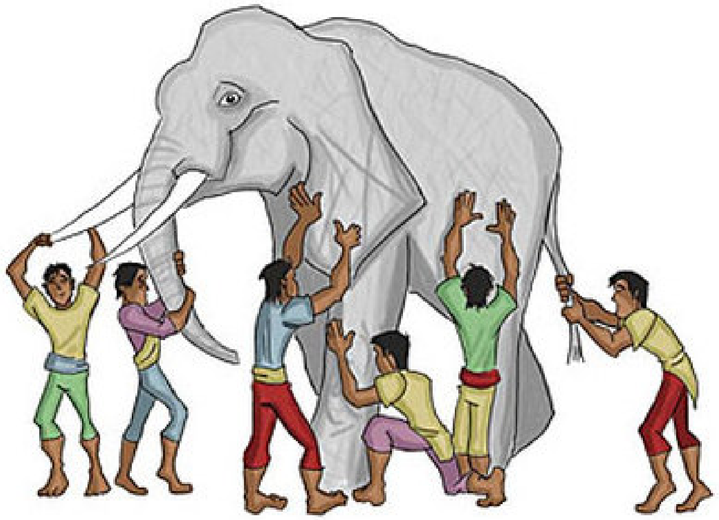
- Each training point has its own “learned” attribution
- Aggregate explanation now has lower sensitivity

Why aggregate?



- Each training point has its own “learned” attribution
- Aggregate explanation now has lower sensitivity
- Resulting attribution uses motivating reasoning of a doctor

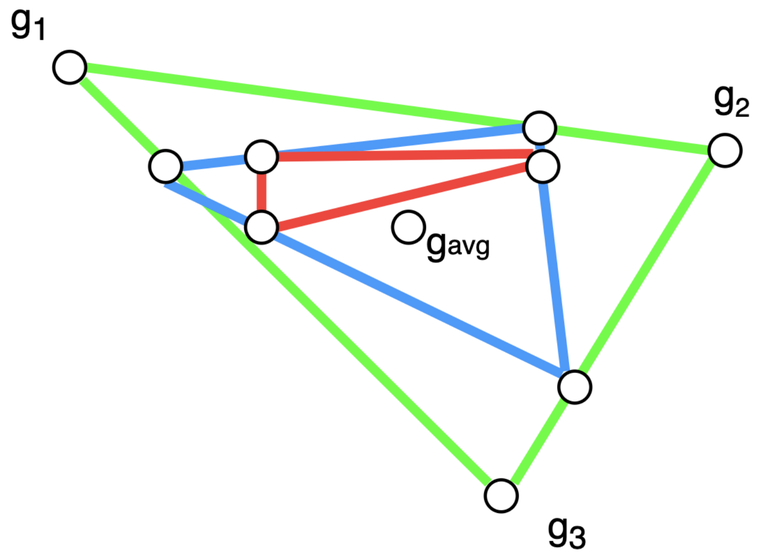
Why aggregate?



- Each training point has its own “learned” attribution
- Aggregate explanation now has lower sensitivity
- Resulting attribution uses motivating reasoning of a doctor
- **SUPER SUPER** cheap to compute

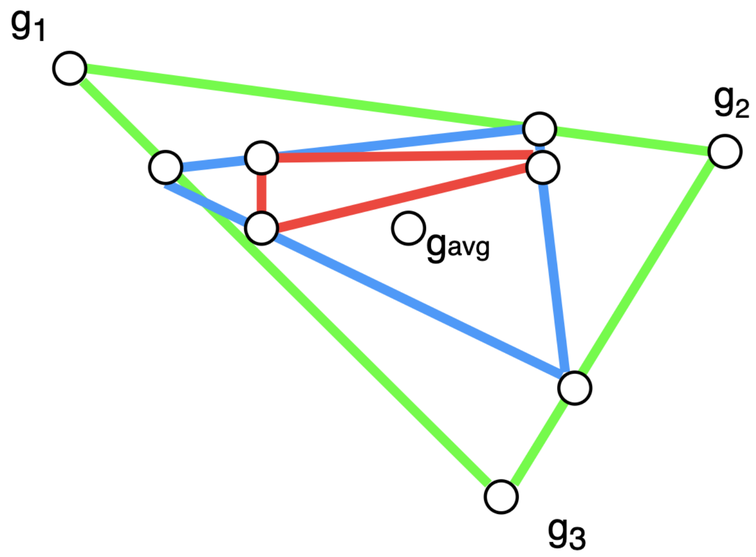
Minimizing Complexity

Region Shrinking Method

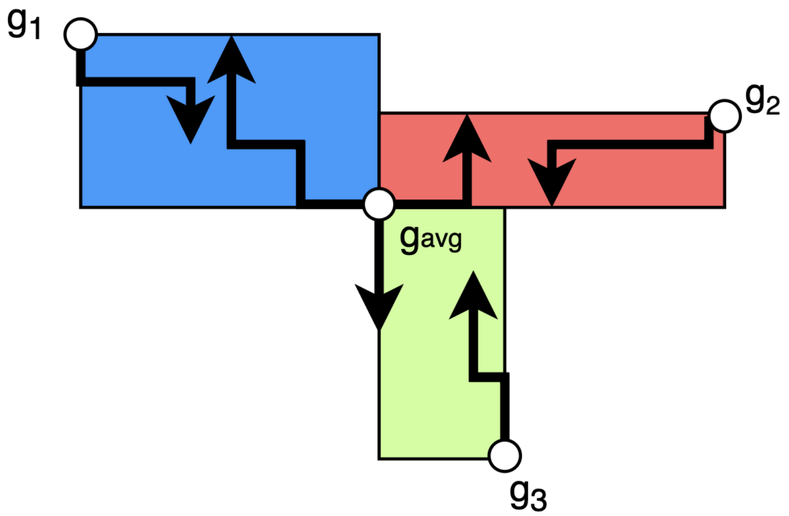


Minimizing Complexity

Region Shrinking Method



Gradient-Descent Style Method



Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
   - Consider counterfactuals, natural language explanations, etc.

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
   - Consider counterfactuals, natural language explanations, etc.
3. Working with experts to find a $\boldsymbol{g}^*$

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
   - Consider counterfactuals, natural language explanations, etc.
3. Working with experts to find a $\boldsymbol{g}^*$
4. Multi-Objective optimization

Summary

This Work

1. Aggregate local explanations with classical rank aggregation or via convex combination can be useful
2. Aggregating Shapley value explanations results in a Shapley value
3. We can learn aggregate explanations to lower sensitivity and complexity

Future Work

1. Axiomatic Aggregation
   - If $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ satisfy Axiom R, then $\boldsymbol{g}_{\mathcal{A}}$ satisfies R.
2. Are feature-based explanations even useful?
   - Consider counterfactuals, natural language explanations, etc.
3. Working with experts to find a $\boldsymbol{g}^*$
4. Multi-Objective optimization
   - Resulting Setup

$$max\ \text{faithfulness}(\boldsymbol{g}_{agg}) + \text{sensitivity}(\boldsymbol{g}_{agg})$$

Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

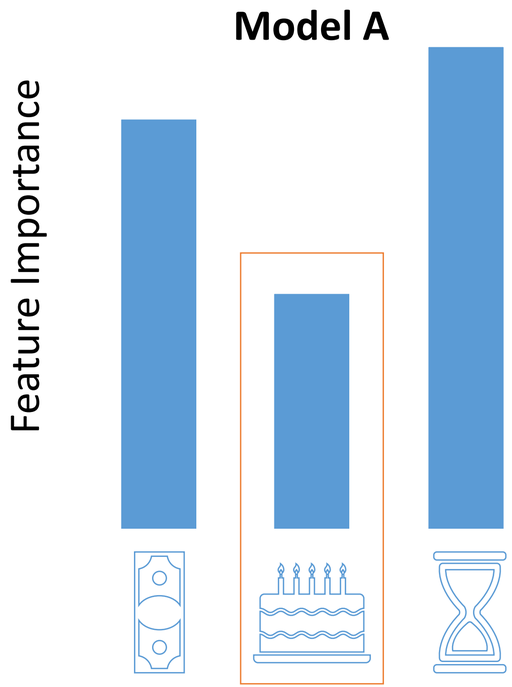
You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods

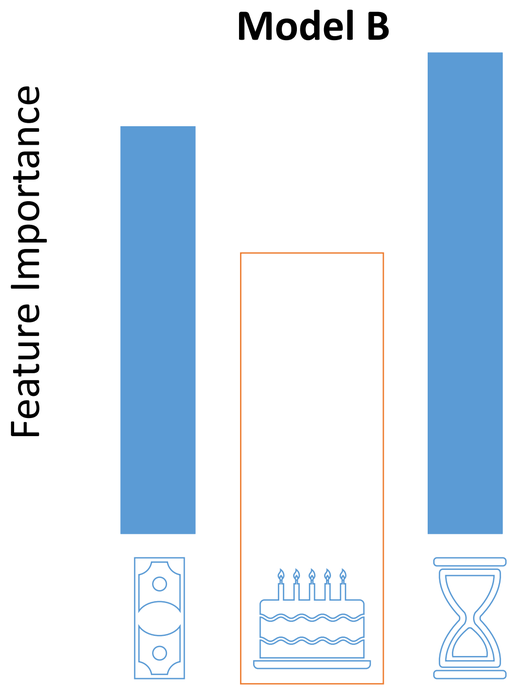
Botty Dimanov, **Umang Bhatt**, Mateja Jamnik, and Adrian Weller

Appeared at the European Conference on Artificial Intelligence 2020
http://ecai2020.eu/papers/72_paper.pdf

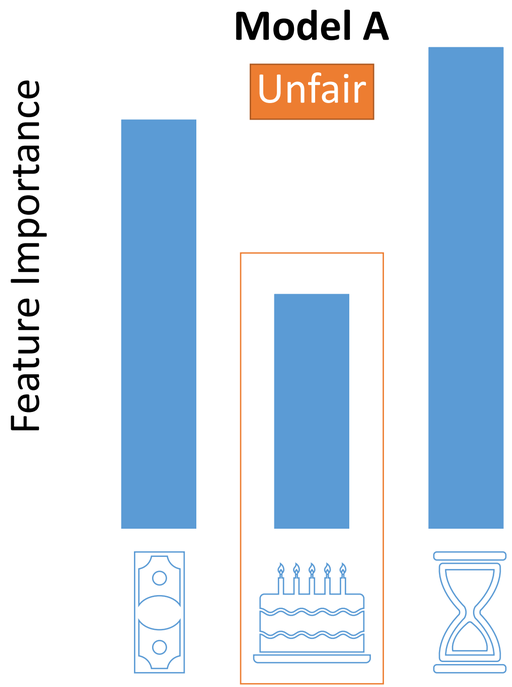
The Alan Turing Institute

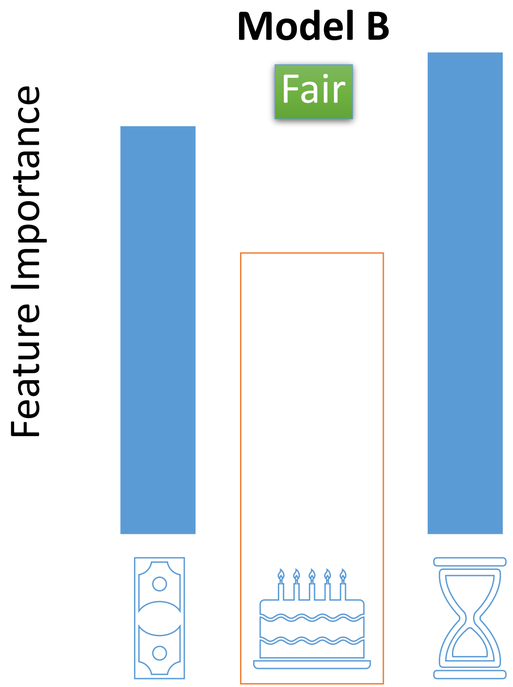
Why do we care?





Why do we care?

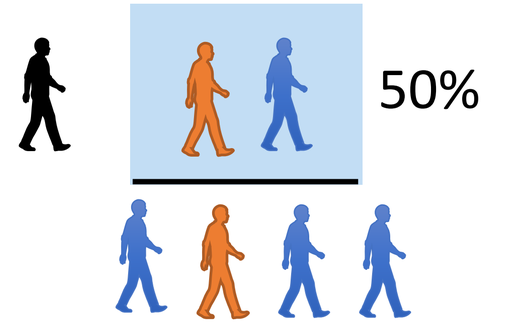


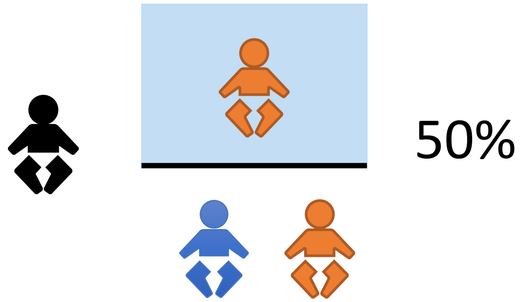


What is fairness?

Demographic Parity

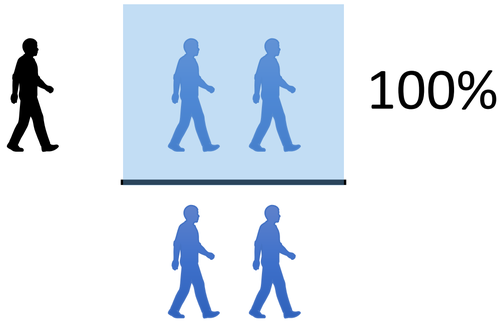
positive rate

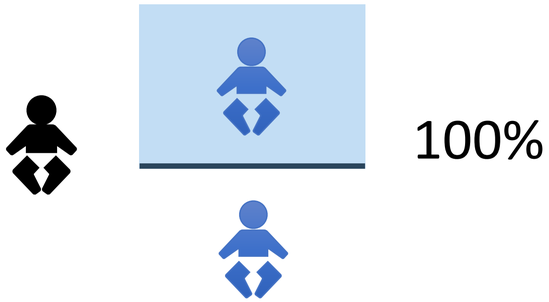




Equal Opportunity

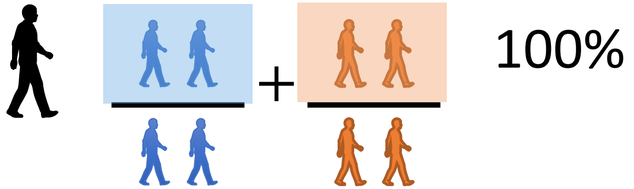
true positive rate

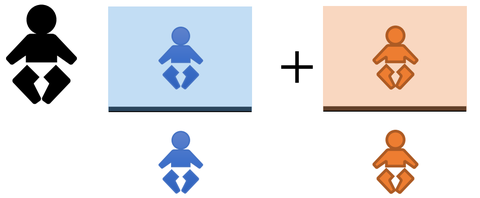




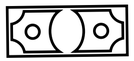
Equal Accuracy

true positive +
true negative





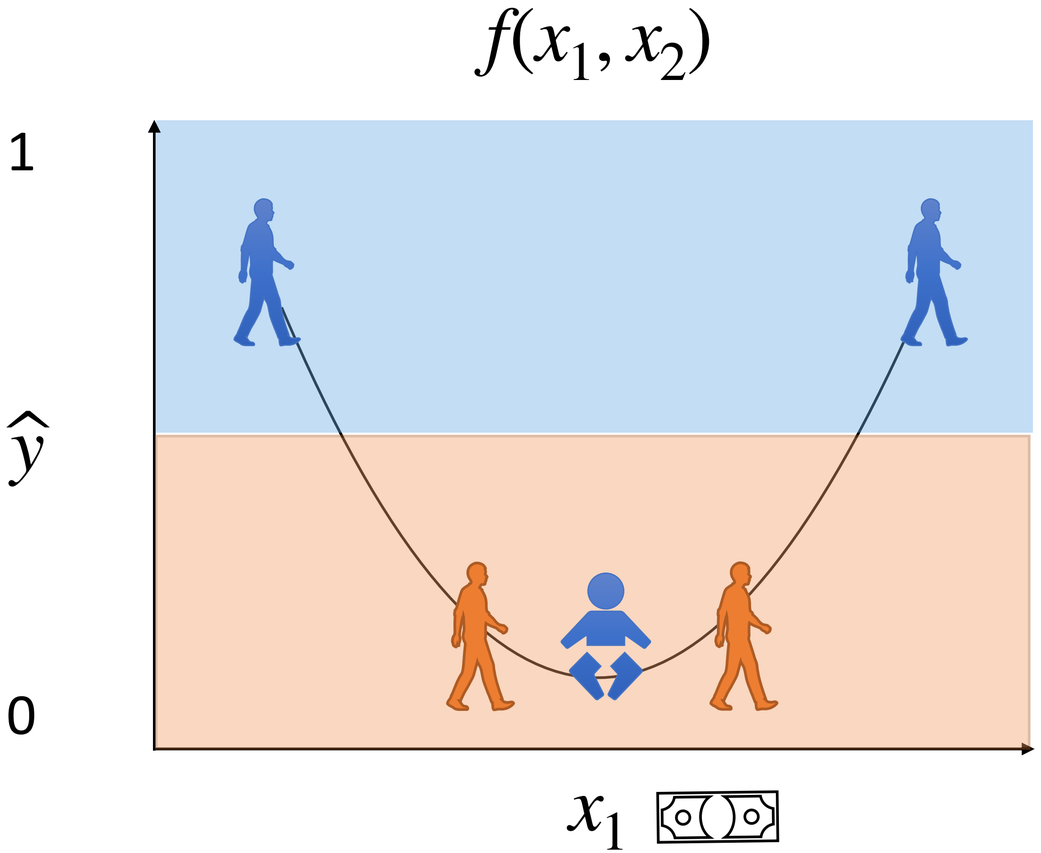
Toy Example

$x_1 \in R$ 

$x_2 \in \{0,1\}$ 

$y \in \{0,1\}$ 

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = 0$$



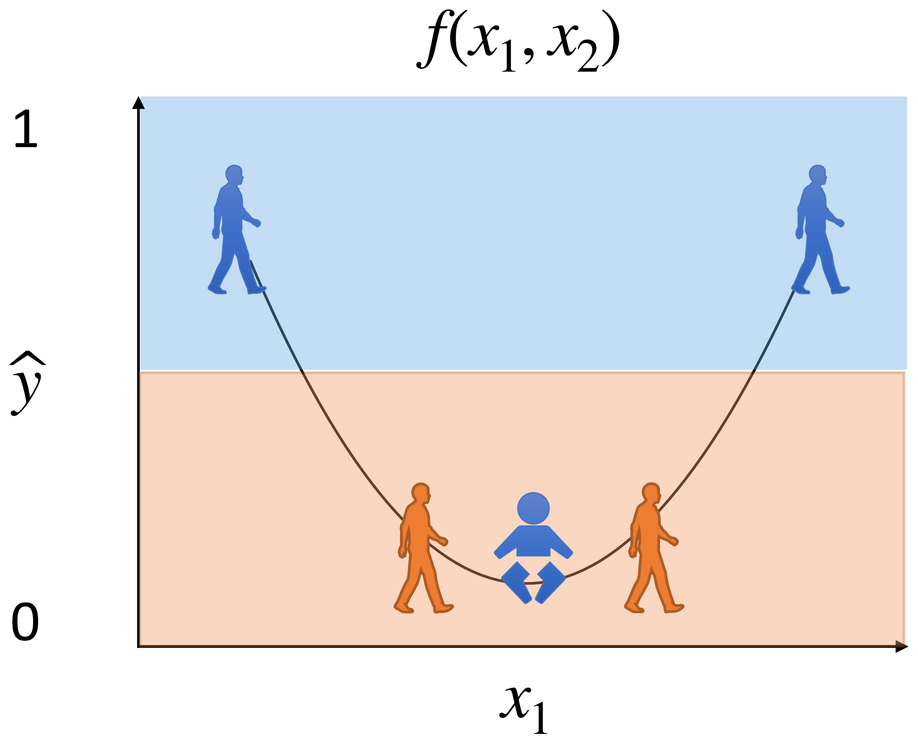
Toy Example

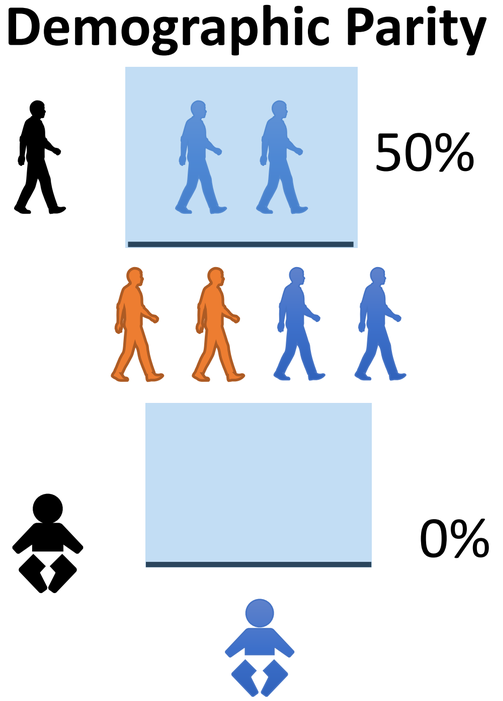
$x_1 \in R$

$x_2 \in \{0,1\}$

$y \in \{0,1\}$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = 0$$





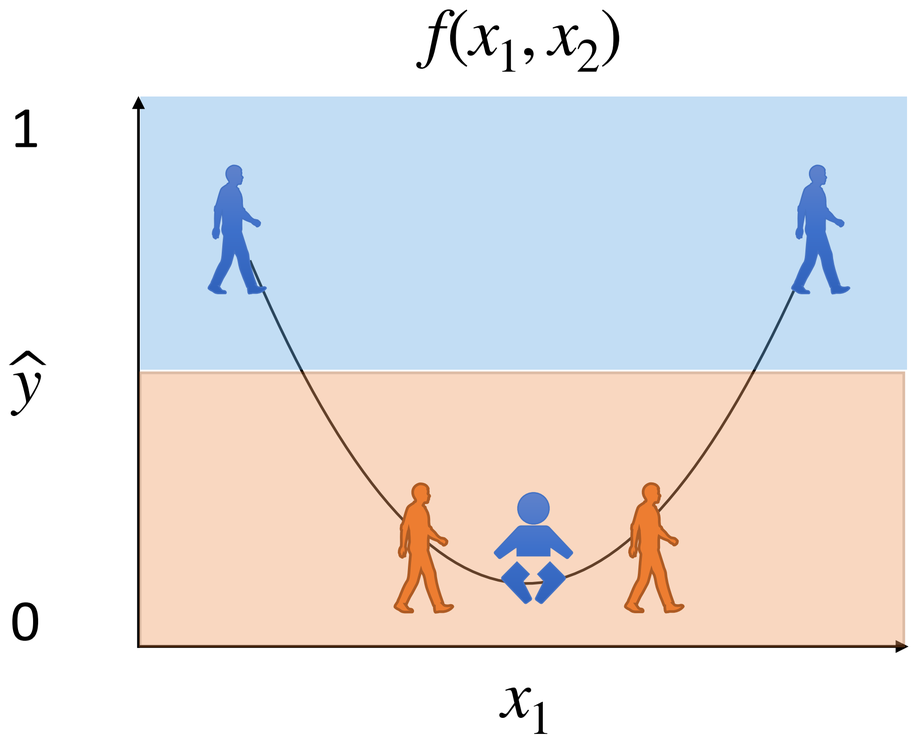
Toy Example

$x_1 \in R$

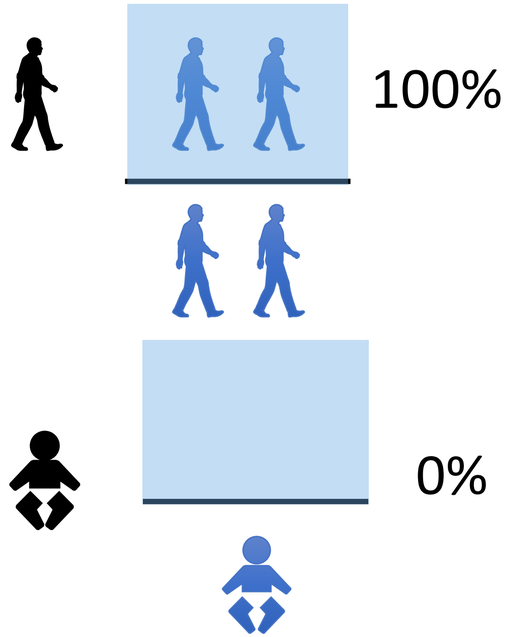
$x_2 \in \{0,1\}$,

$y \in \{0,1\}$,

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = 0$$



Equal Opportunity



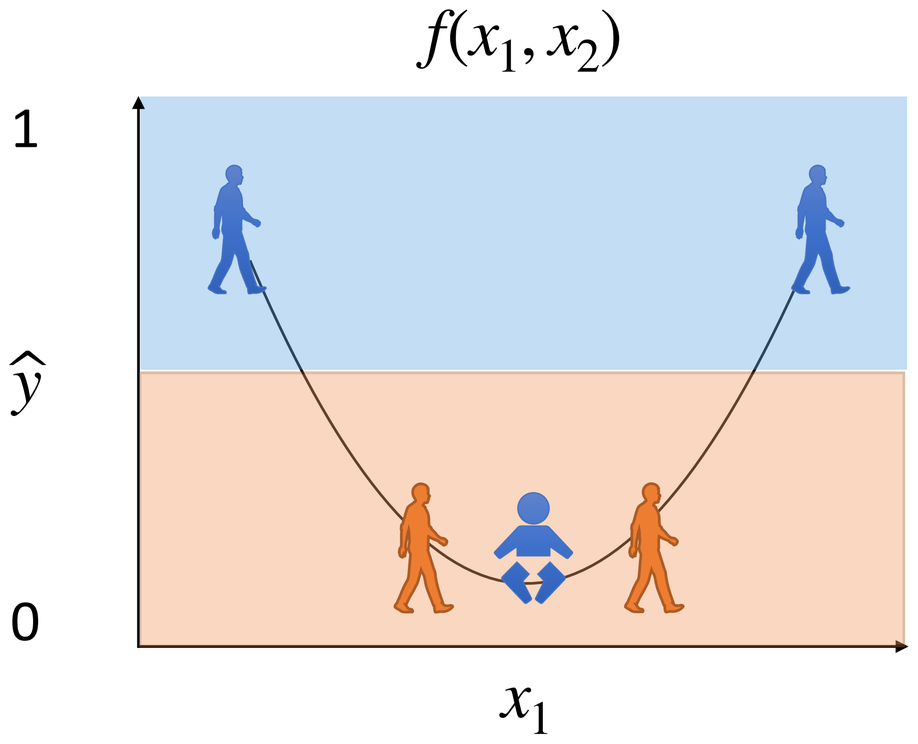
Toy Example

$x_1 \in R$

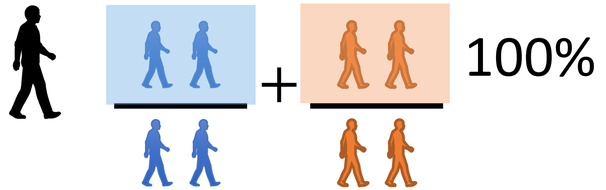
$x_2 \in \{0,1\}$

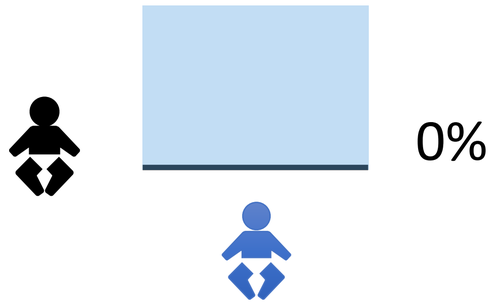
$y \in \{0,1\}$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = 0$$



Equal Accuracy





Can we manipulate explanations?

Modified explanations via adversarial perturbations of **inputs**

- Ghorbani, Abid, and Zou. AAAI 2019
- Dombrowski et al. NeurIPS 2019
- Slack et al. AIES 2019

Control visual explanations via adversarial perturbations of **parameters**

- Heo, Joo, and Moon. NeurIPS 2019

Downgrade explanations via adversarial perturbations of **parameters to hide unfairness**

Our Setup

Classifier $f : \mathbf{X} \mapsto \mathbf{Y}$

Explanation $g(f, \mathbf{x})_j$

Our Goal

$f_\theta \longmapsto f_{\theta+\delta}$

Desirable Properties

Model Similarity $\forall i,\ f_{\theta+\delta}(\mathbf{x}^{(i)}) \approx f_\theta(\mathbf{x}^{(i)})$

Low Target Feature Attribution $\forall i,\ |g(f_{\theta+\delta}, \mathbf{x}^{(i)})_j| \ll |g(f_\theta, \mathbf{x}^{(i)})_j|$

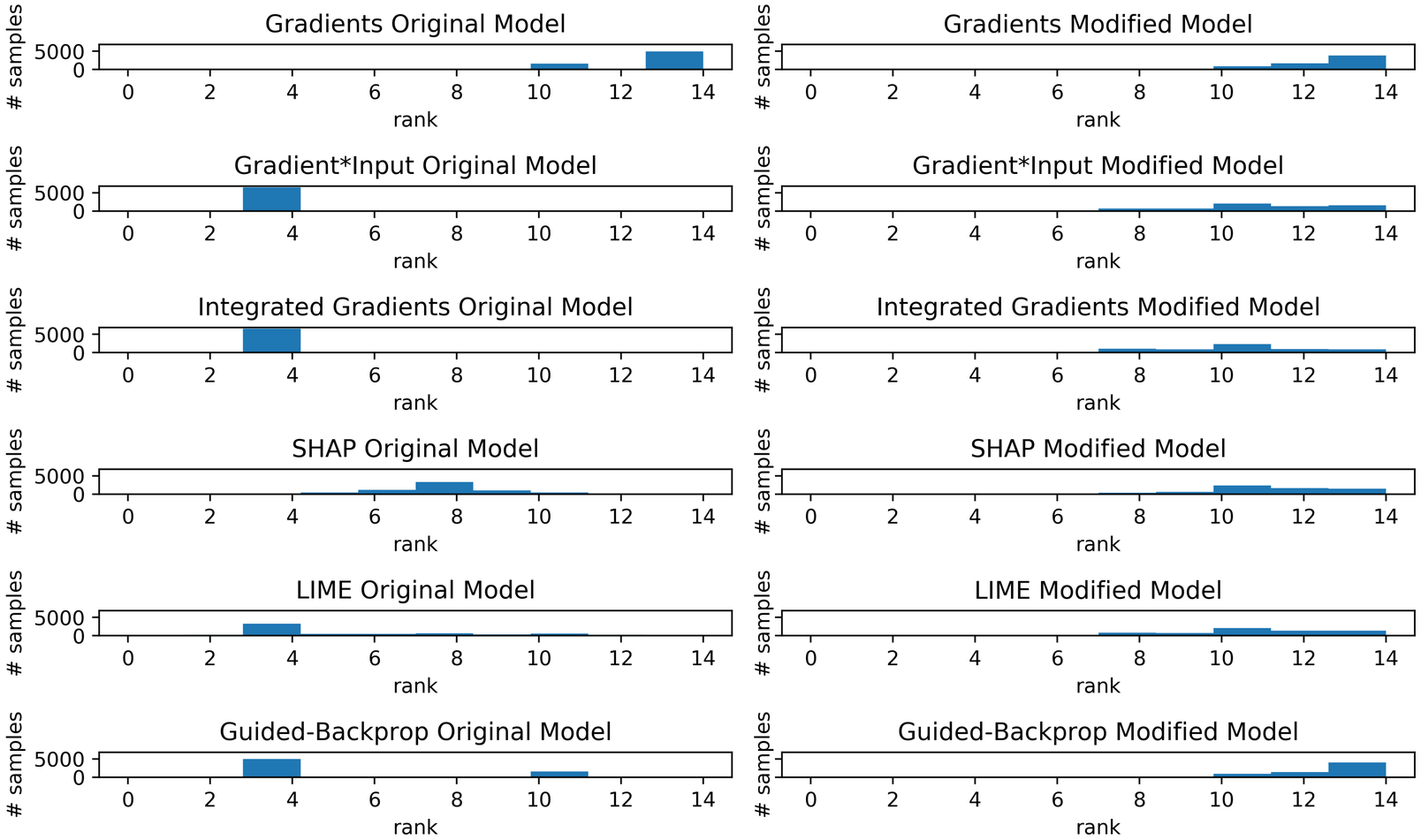
Our Method

Adversarial Explanation Attack

$$\text{argmin}_{\delta} \;\; L' = L(f_{\theta+\delta}, \mathbf{x}, \mathbf{y}) + \frac{\alpha}{n} \left|\left| \nabla_{\mathbf{X}_{:,j}} L(f_{\theta+\delta}, \mathbf{x}, \mathbf{y}) \right|\right|_p$$

Results (Importance Ranking)

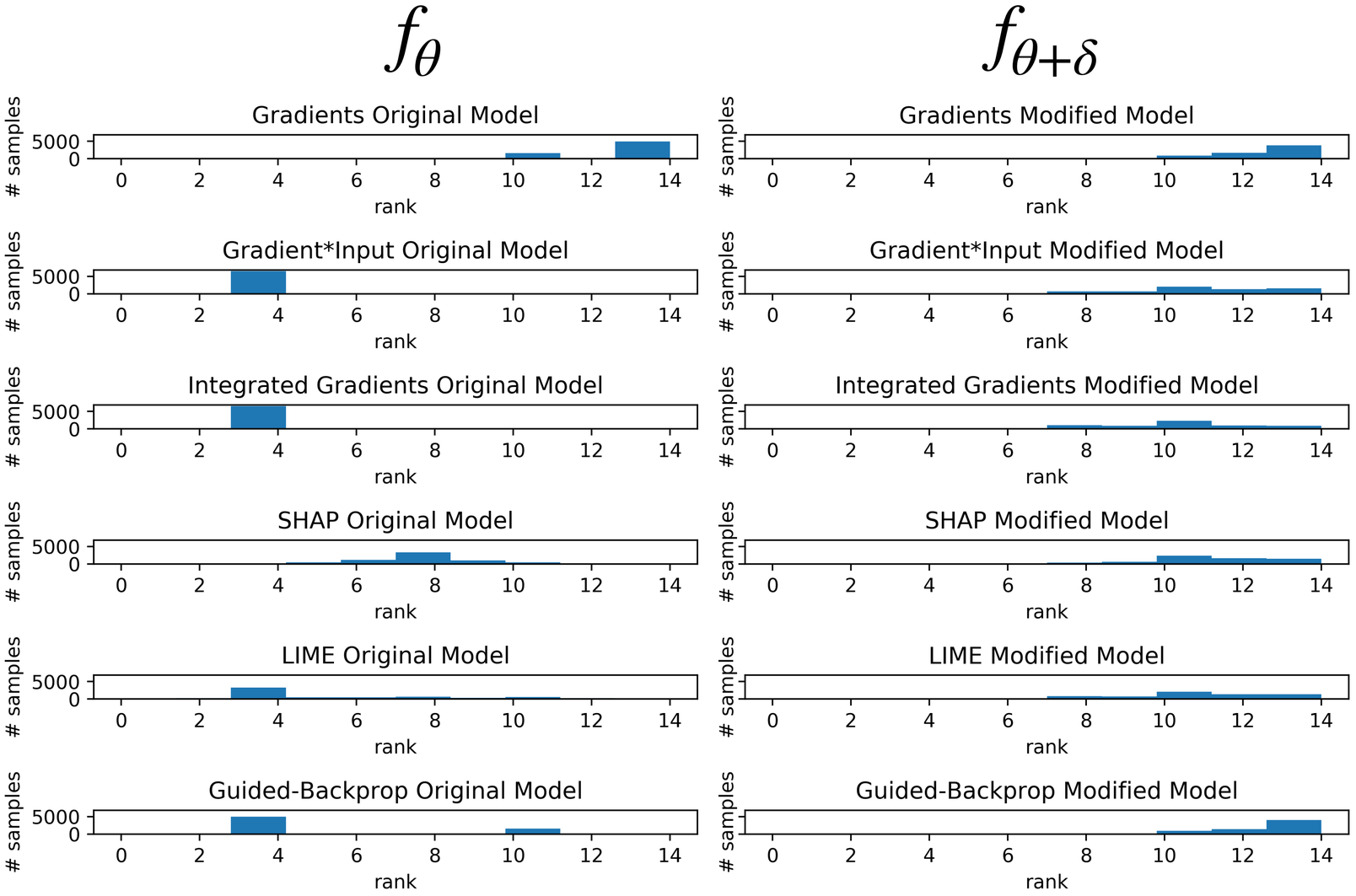
f_θ $f_{\theta+\delta}$



Results (Importance Ranking)

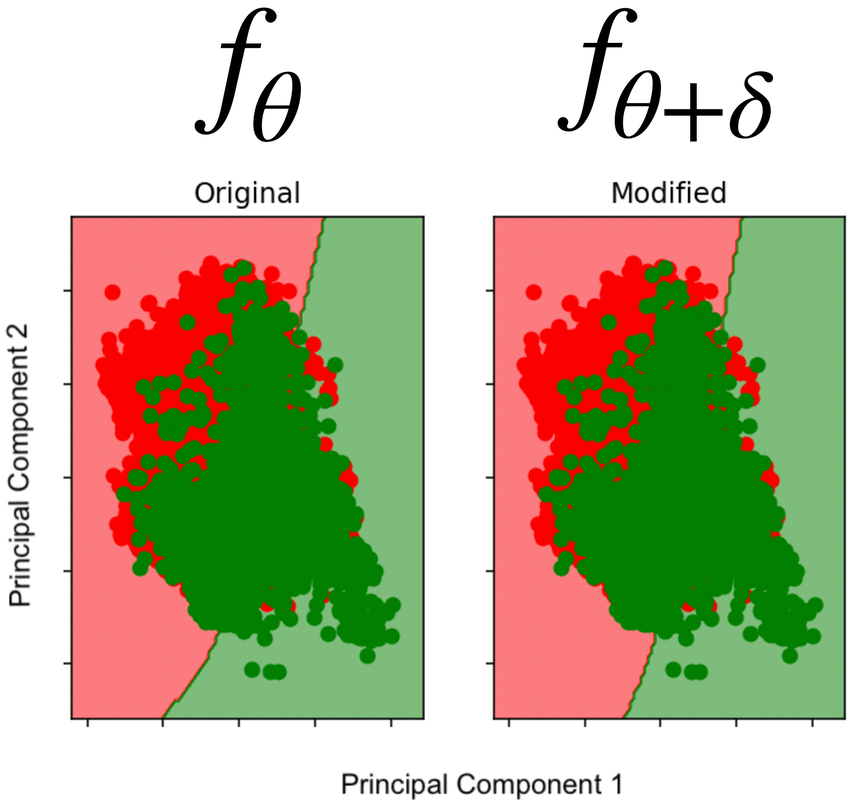
Our *adversarial explanation attack:*

1. Significantly decreases relative importance.
2. Generalizes to test points.
3. Transfers across explanation methods.



Findings

- Little change in accuracy, but difference in outputs is detectable
- Low attribution achieved with respect to **multiple** explanation methods
- High unfairness across multiple fairness metrics (compared to holding feature constant)



Takeaway

Feature importance reveals nothing reliable about model fairness.

Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

Effects of Uncertainty on the Quality of Feature Importance Explanations

Torgyn Shaikhina,* **Umang Bhatt**,* Roxanne Zhang, Konstantinos Georgatzis, Alice Xiang, and Adrian Weller

Appeared at the AAAI Workshop on Explainable Agency in Artificial Intelligence 2021

Motivation

A. Vulnerability of feature-based model explanations

B. Practitioners may use such explanations for model selection, which leads to confirmation bias, over-reliance, and potential manipulation

Our Assumptions

- [Assumption] **Fixed model assumption is restrictive**: Existing methods for generating explanations tend to explain a point estimate, which assumes that only one suitable model exists.

- [Remark] Obtain uncertainty estimates by **subsampling** from the posterior of models (from the same model class and with the same model specification) given training samples and initial states.

Shapley Value

$$\phi_i(v, x, f) = \frac{1}{|T|} \sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1} (v_x(S \cup \{i\}, f) - v_x(S, f))$$

$$v_x(S, f) = \mathbb{E}\left[f(z) \,|\, z = \bar{x}_{[\bar{x}_s = x_s]}\right]$$

Our Approach

$$\phi_i(v, x)^* = \int_{f \in \mathcal{F}} P(f \mid \mathcal{D}) \phi_i(v, x, f) df$$

$$\tilde{\phi}_i(v, x) = \sum_{f \in \mathcal{B}} w_f \phi_i(v, x, f) = \mathbb{E}_{f \in \mathcal{B}}[\phi_i(v, x, f)]$$

Our Approach

$$\phi_i(v, x)^* = \int_{f \in \mathcal{F}} P(f \mid \mathcal{D}) \phi_i(v, x, f) df$$

$$\tilde{\phi}_i(v, x) = \sum_{f \in \mathcal{B}} w_f \phi_i(v, x, f) = \mathbb{E}_{f \in \mathcal{B}}[\phi_i(v, x, f)]$$

Variance

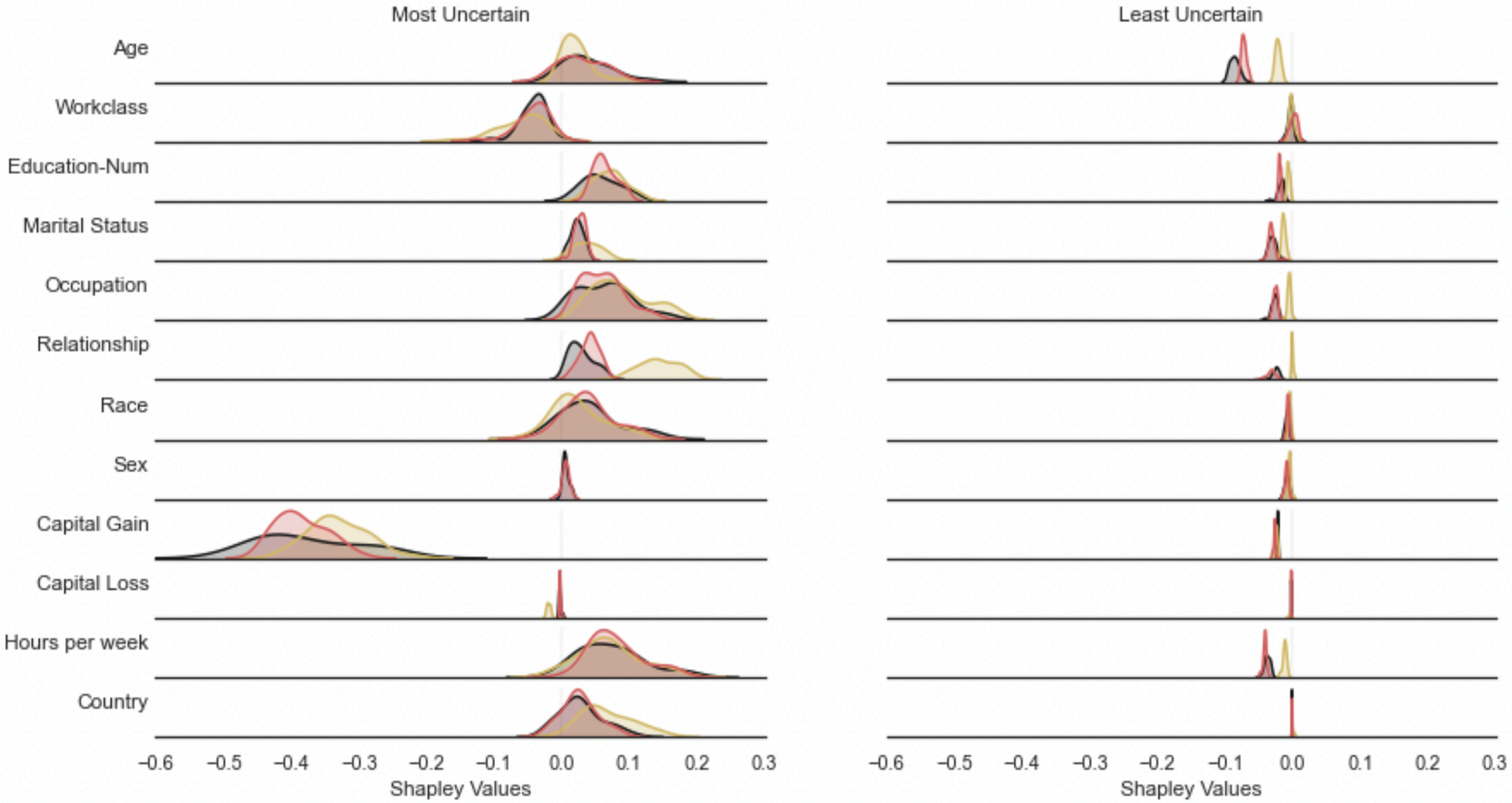
$$s_i(\mathcal{B}, x) = \mathbb{V}_{f \in \mathcal{B}}[\phi_i(v, x, f)]$$

Experimental Setup

- Model is uncertainty on a point if it's prediction lies 1.5 times the IQR
- Set of models found by (i) changing initialization or (ii) subsampling from the training data
- Multiple Evaluation Criteria: Complexity, Monotonicity, Efficiency, Faithfulness…

Finding #1

Predictive uncertainty leads to high variance in explanations



Feature importance scores of an OOD, most uncertain (left) and least uncertain (right) samples of the Adult dataset computed using Exact (red), Tree (black) and Kernel (yellow) SHAP. Each feature is presented as a distribution of its Shapley values associated with the ensemble of 30 Gradient Boosted Trees.

*Similar empirical evidence for other tree-based models and neural nets

Finding #2

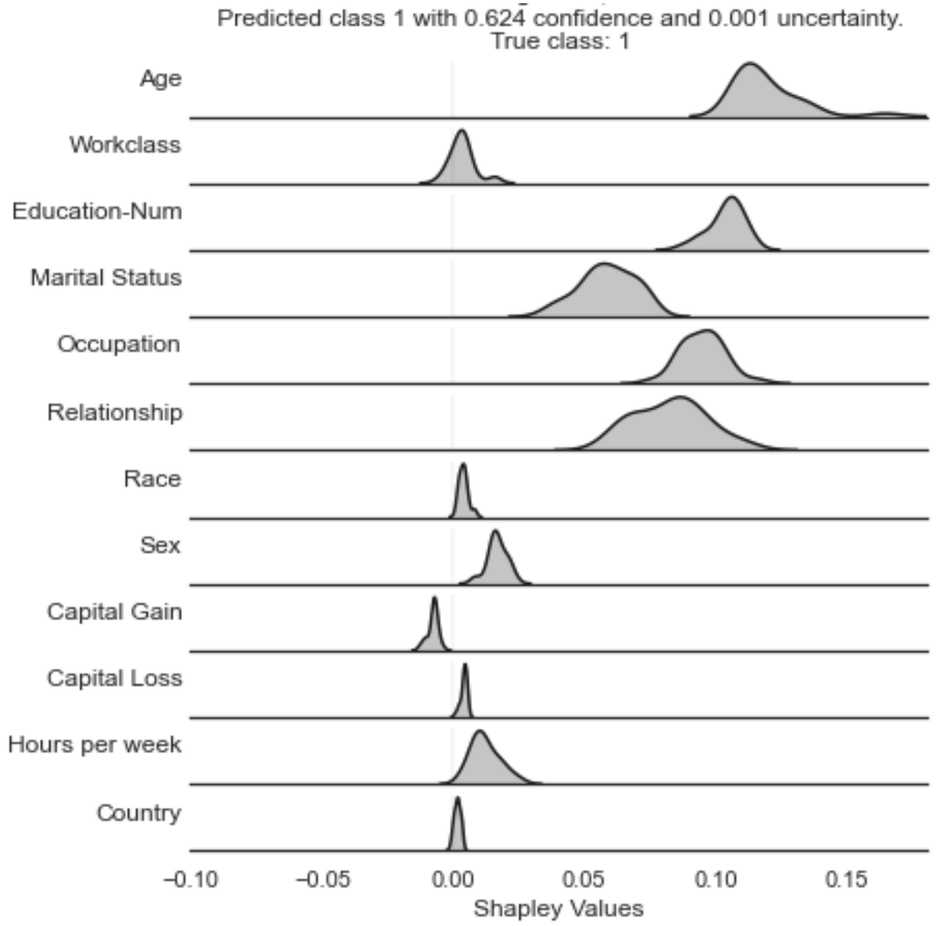
Uncertainty degrades quality of explanations across w.r.t. all metrics:

		Variance	Complexity	Monotonicity	Efficiency	Faithfulness
Exact Shapley	non-OOD	0.07 ± 0.00	1.88 ± 0.22	0.88 ± 0.04	1.00 ± 0.00	0.00 ± 0.24
	OOD	$\mathbf{0.39 \pm 0.00}$	$\mathbf{1.96 \pm 0.14}$	$\mathbf{0.85 \pm 0.05}$	1.00 ± 0.00	$\mathbf{-0.01 \pm 0.23}$
TreeSHAP	non-OOD	0.08 ± 0.00	1.90 ± 0.20	0.88 ± 0.04	0.99 ± 0.09	0.00 ± 0.24
	OOD	$\mathbf{0.44 \pm 0.00}$	$\mathbf{1.96 \pm 0.13}$	$\mathbf{0.85 \pm 0.04}$	$\mathbf{0.97 \pm 0.18}$	$\mathbf{-0.01 \pm 0.23}$
KernelSHAP	non-OOD	0.19 ± 0.00	1.81 ± 0.23	0.89 ± 0.05	1.00 ± 0.00	0.00 ± 0.24
	OOD	$\mathbf{0.809 \pm 0.00}$	$\mathbf{1.85 \pm 0.16}$	$\mathbf{0.82 \pm 0.05}$	1.00 ± 0.00	0.00 ± 0.23
Integrated Gradients	non-OOD	20.00 ± 5.34	1.53 ± 0.22	0.96 ± 0.03	$\mathbf{0.99 \pm 0.10}$	0.00 ± 0.16
	OOD	$\mathbf{46.48 \pm 18.61}$	$\mathbf{1.55 \pm 0.18}$	0.96 ± 0.03	1.00 ± 0.03	0.00 ± 0.13
Smoothgrad	non-OOD	20.35 ± 0.52	1.92 ± 0.11	0.92 ± 0.04	1.00 ± 0.10	0.00 ± 0.13
	OOD	19.97 ± 0.43	$\mathbf{1.98 \pm 0.06}$	0.92 ± 0.03	1.00 ± 0.09	$\mathbf{-0.02 \pm 0.10}$

Comparison of the explanation quality on OOD and in-distribution samples of the Adult dataset.
Presented are mean and standard deviation of quality scores averaged across the ensemble of models.
Values in bold indicate where the degradation in quality was statistically significant ($p<0.05$).

Implication #1

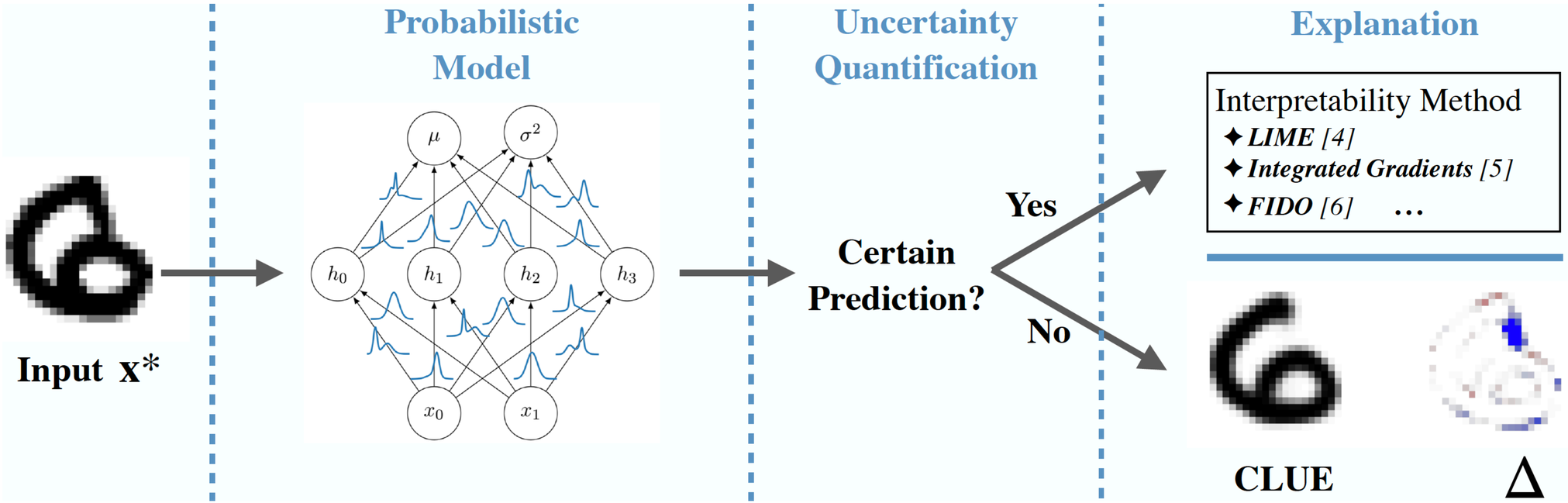
When interpreting feature importance scores, data practitioners must consider the entire distribution (not just point estimates).



Implication #2

Don’t use feature importance on uncertain data — it’s unreliable…

Use counterfactual latent uncertainty explanations (CLUEs) instead!

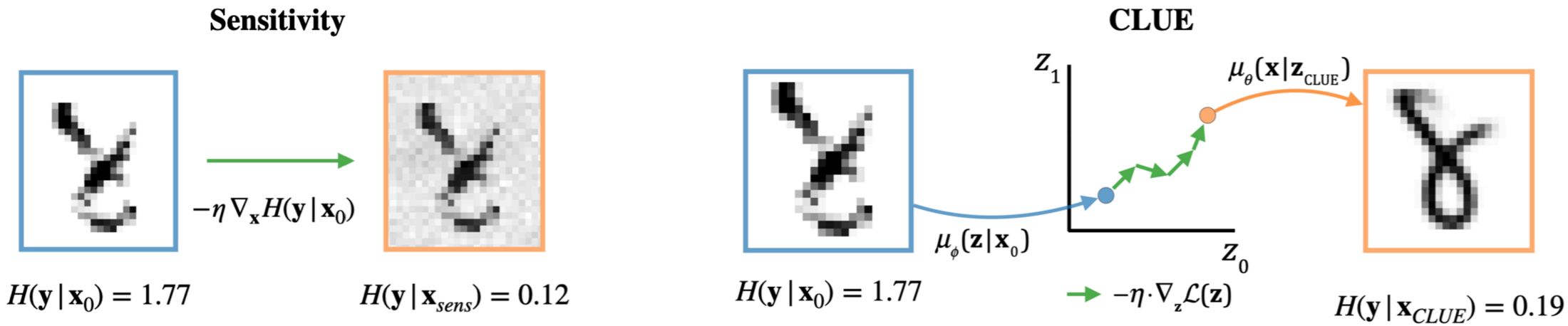


Getting a CLUE: A Method for Explaining Uncertainty Estimates

Javier Antoran, **Umang Bhatt**, Tameem Adel, Adrian Weller, and Jose Miguel Hernandez-Lobato

To appear at the International Conference on Learning Representations 2021
https://arxiv.org/abs/2006.06848

Optimization in the Latent Space of a Deep Generative Model



Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

Practitioner-Driven Questions

1. Can we provide methodology for practitioners to evaluate explanations?
2. Can existing explainability tools be used to identify model unfairness?
3. Can we quantify how much uncertainty is associated with given explanations?

Practical Approaches to Explainable Machine Learning

Thanks for listening!

Umang Bhatt
usb20@cam.ac.uk
@umangsbhatt