Attribution-based Explanations that Provide Recourse Cannot be Robust

Joint work with

Rianne de Heide and Tim van Erven

2022-11-21

Programme of today

- Brief introduction to Explainable Artificial Intelligence
- Attribution Methods
- Recourse and Robustness
- Impossibility result
- When Recourse is possible

Explainable Artificial intelligence (XAI)

Interpretable Machine Learning (IML)

Call for XAI

Some Reasons

- Fairness: Biases can be detected earlier
- Trustworthiness
- Increases reliability
- Regulation

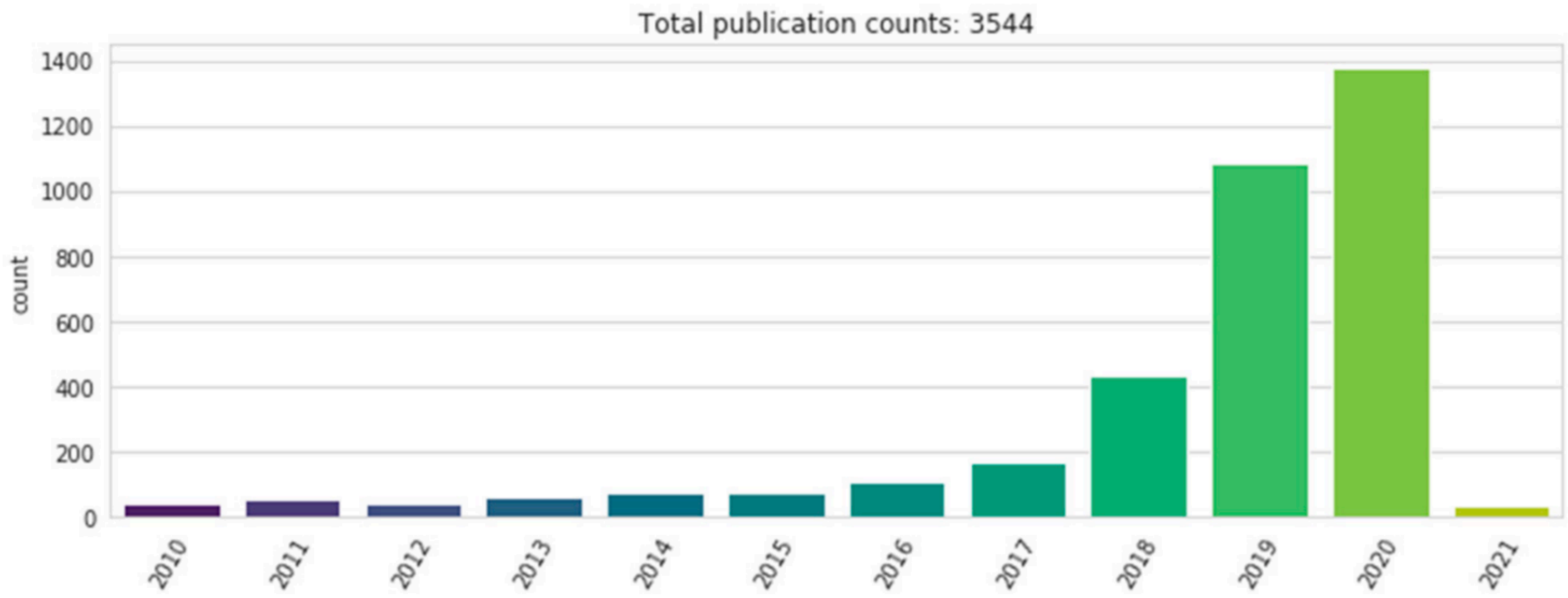
Explanations

Explosion

Methods
CAM with global average pooling [42], [91]
+ Grad-CAM [43] generalizes CAM, utilizing gradient
+ Guided Grad-CAM and Feature Occlusion [68]
+ Respond CAM [44]
+ Multi-layer CAM [92]
LRP (Layer-wise Relevance Propagation) [13], [53]
+ Image classifications. PASCAL VOC 2009 etc [45]
+ Audio classification. AudioMNIST [47]
+ LRP on DeepLight. fMRI data from Human Connectome Project [48]
+ LRP on CNN and on BoW(bag of words)/SVM [49]
+ LRP on compressed domain action recognition algorithm [50]
+ LRP on video deep learning, *selective relevance method* [52]
+ BiLRP [51]
DeepLIFT [57]
Prediction Difference Analysis [58]
Slot Activation Vectors [41]
PRM (Peak Response Mapping) [59]
LIME (Local Interpretable Model-agnostic Explanations) [14]
+ MUSE with LIME [85]
+ Guidelinebased Additive eXplanation optimizes complexity, similar to LIME [93]
# Also listed elsewhere: [56], [69], [71], [94]
Others. Also listed elsewhere: [95]
+ Direct output labels. Training NN via multiple instance learning [65]
+ Image corruption and testing Region of Interest statistically [66]
+ Attention map with autofocus convolutional layer [67]
DeconvNet [72]
Inverting representation with natural image prior [73]
Inversion using CNN [74]
Guided backpropagation [75], [91]
Activation maximization/optimization [38]
+ Activation maximization on DBN (Deep Belief Network) [76]
+ Activation maximization, multifaceted feature visualization [77]
Visualization via regularized optimization [78]
Semantic dictionary [39]
Network dissection [36], [37]
Decision trees
Propositional logic, rule-based [82]
Sparse decision list [83]
Decision sets, rule sets [84], [85]
Encoder-generator framework [86]
Filter Attribute Probability Density Function [87]
MUSE (Model Understanding through Subspace Explanations) [85]

(2014.03)	SEDC [129]
(2015.08)	OAE [51]
(2016.05)	HCLS [110, 112]
(2017.06)	Feature Tweaking [186]
(2017.11)	CF Expl. [196]
(2017.12)	Growing Spheres [114]
(2018.02)	CEM [55]
(2018.02)	POLARIS [209]
(2018.05)	LORE [80]
(2018.06)	Local Foil Trees [190]
(2018.09)	Actionable Recourse [189]
(2018.11)	Weighted CFs [77]
(2019.01)	Efficient Search [175]
(2019.04)	CF Visual Expl. [76]
(2019.05)	MACE [99]
(2019.05)	DiCE [145]
(2019.05)	CERTIFAI [179]
(2019.06)	MACEM [56]
(2019.06)	Expl. using SHAP [165]
(2019.07)	Nearest Observable [201]
(2019.07)	Guided Prototypes [191]
(2019.07)	REVISE [95]
(2019.08)	CLEAR [202]
(2019.08)	MC-BRP [123]
(2019.09)	FACE [162]
(2019.09)	Equalizing Recourse [83]
(2019.10)	Action Sequences [163]
(2019.10)	C-CHVAE [156]
(2019.11)	FOCUS [124]
(2019.12)	Model-based CFs [127]
(2019.12)	LIME-C/SHAP-C [164]
(2019.12)	EMAP [41]
(2019.12)	PRINCE [71]
(2019.12)	LowProFool [18]
(2020.01)	ABELE [79]
(2020.01)	SHAP-based CFs [66]
(2020.02)	CEML [11–13]
(2020.02)	MINT [100]
(2020.03)	ViCE [74]
(2020.03)	Plausible CFs [22]
(2020.04)	SEDC-T [193]
(2020.04)	MOC [52]
(2020.04)	SCOUT [199]
(2020.04)	ASP-based CFs [28]
(2020.05)	CBR-based CFs [103]
(2020.06)	Survival Model CFs [106]
(2020.06)	Probabilistic Recourse [101]
(2020.06)	C-CHVAE [155]
(2020.07)	FRACE [210]
(2020.07)	DACE [96]
(2020.07)	CRUDS [60]
(2020.07)	Gradient Boosted CFs [5]
(2020.08)	Gradual Construction [97]
(2020.08)	DECE [44]
(2020.08)	Time Series CFs [16]
(2020.08)	PermuteAttack [87]
(2020.10)	Fair Causal Recourse [195]
(2020.10)	Recourse Summaries [167]
(2020.10)	Strategic Recourse [43]
(2020.11)	PARE [172]

Methods	H
Linear probe [101]	
Regression based on CNN [106]	
Backwards model for interpretability of linear models [107]	
GDM (Generative Discriminative Models): ridge regression + least square [100]	
GAM, GA^{2}M (Generative Additive Model) [82], [102], [103]	
ProtoAttend [105]	
Other content-subject-specific models:	N
+ Kinetic model for CBF (cerebral blood flow) [131]	N
+ CNN for PK (Pharmacokinetic) modelling [132]	N
+ CNN for brain midline shift detection [133]	N
+ Group-driven RL (reinforcement learning) on personalized healthcare [134]	N
+ Also see [108]–[112]	N
PCA (Principal Components Analysis), SVD (Singular Value Decomposition)	N
CCA (Canonical Correlation Analysis) [113]	
SVCCA (Singular Vector Canonical Correlation Analysis) [97] = CCA+SVD	
F-SVD (Frame Singular Value Decomposition) [114] on electromyography data	
DWT (Discrete Wavelet Transform) + Neural Network [135]	
MODWPT (Maximal Overlap Discrete Wavelet Package Transform) [136]	
GAN-based Multi-stage PCA [118]	
Estimating probability density with deep feature embedding [119]	
t-SNE (t-Distributed Stochastic Neighbour Embedding) [77]	
+ t-SNE on CNN [120]	
+ t-SNE, activation atlas on GoogleNet [121]	
+ t-SNE on latent space in meta-material design [122]	
+ t-SNE on genetic data [137]	
+ mm-t-SNE on phenotype grouping [138]	
Laplacian Eigenmaps visualization for Deep Generative Model [124]	
KNN (k-nearest neighbour) on multi-center low-rank rep. learning (MCLRR) [125]	
KNN with triplet loss and *query-result activation map pair* [139]	
Group-based Interpretable NN with RW-based Graph Convolutional Layer [123]	
TCAV (Testing with Concept Activation Vectors) [96]	
+ RCV (Regression Concept Vectors) uses TCAV with Br score [140]	
+ Concept Vectors with UBS [141]	
+ ACE (Automatic Concept-based Explanations) [56] uses TCAV	
Influence function [129] helps understand adversarial training points	
Representer theorem [130]	
SocRat (Structured-output Causal Rationalizer) [127]	
Meta-predictors [126]	
Explanation vector [128]	
# Also listed elsewhere: [14], [43], [85], [94]	N
# Also listed elsewhere: [14], [60], [85] etc	N
CNN with separable model [142]	
Information theoretic: Information Bottleneck [98], [99]	
Database of methods v.s. interpretability [10]	N
Case-Based Reasoning [143]	



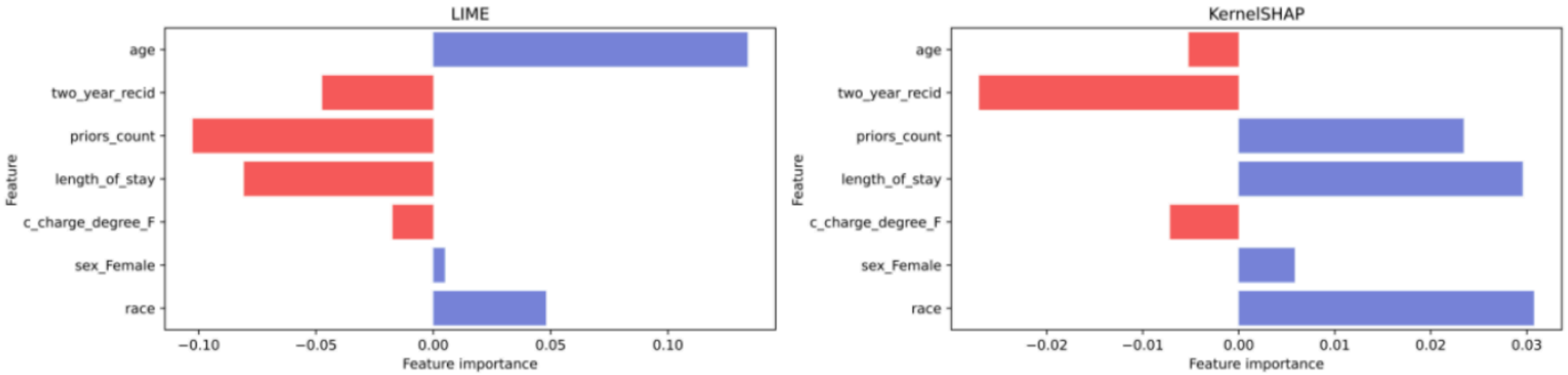
Explanations

Some issues

- Easily manipulated
- Disagreement problem

- Can be Unreliable

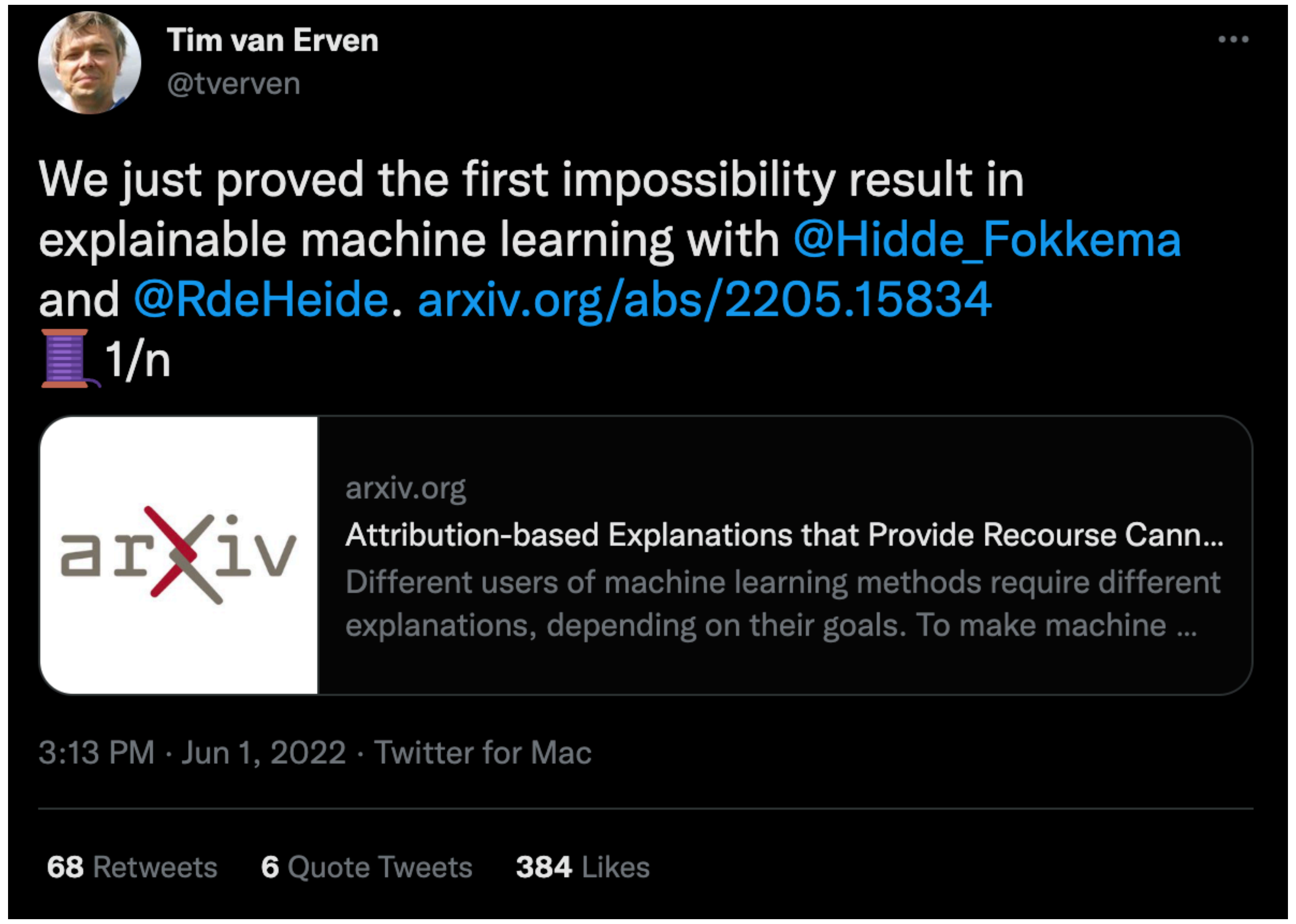
Below, you see a data point, as well as its explanation using methods **LIME** and **KernelSHAP**.



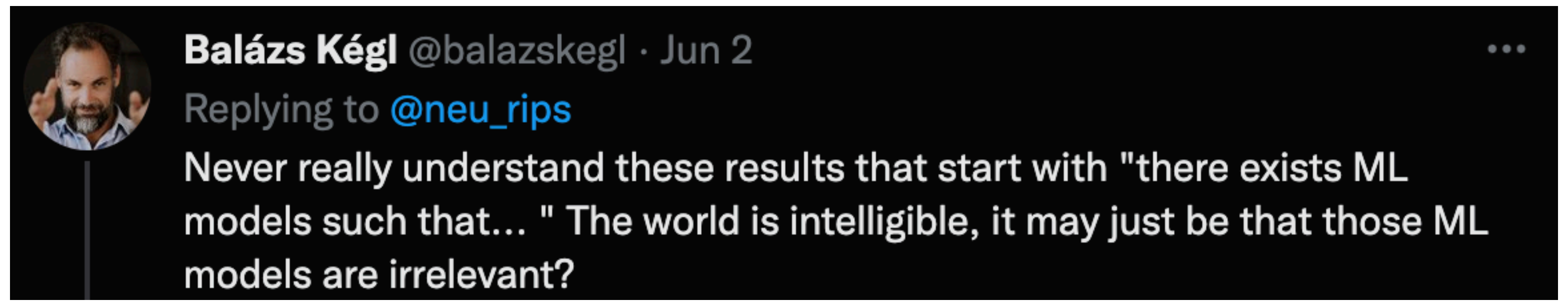
Explanations

Call for Rigor

- Mythos of model interpretability, [Zachary Lipton, 2017]
- Towards a rigorous science of interpretable machine learning, [Doshi-Velez, Kim, 2017]
- “Interpretability research suffers from an over-reliance on intuition-based approaches that risk—and in some cases have caused—illusory progress and misleading conclusions”, [Leavitt, Morcos, 2020]



Some people liked this

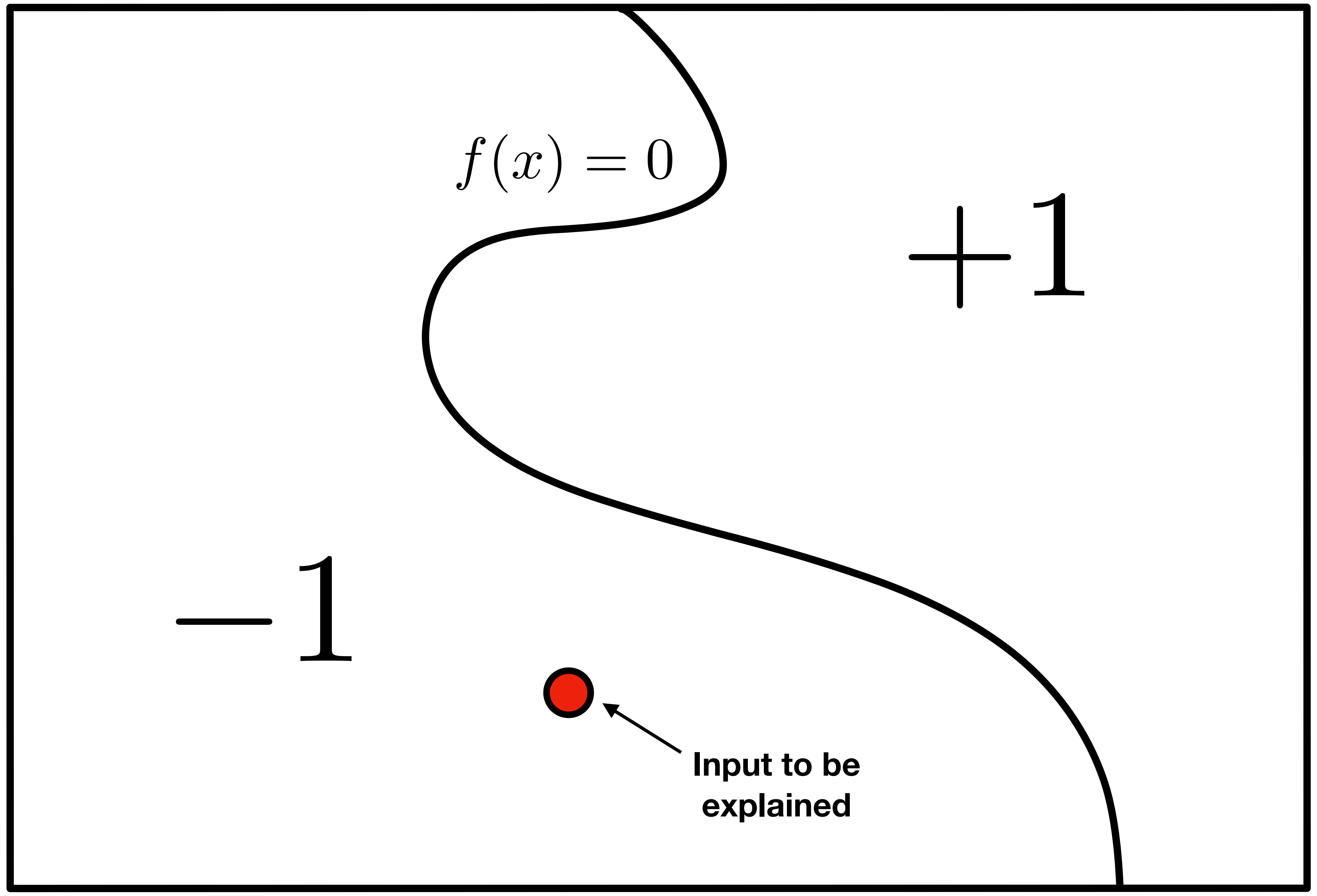


Some people didn’t

Attribution methods

Setting

Post-Hoc and local explanations



Machine learning model, e.g. a classifier:

$$f : \mathcal{X} \subseteq \mathbb{R}^d \to [0,1], \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \mapsto y$$

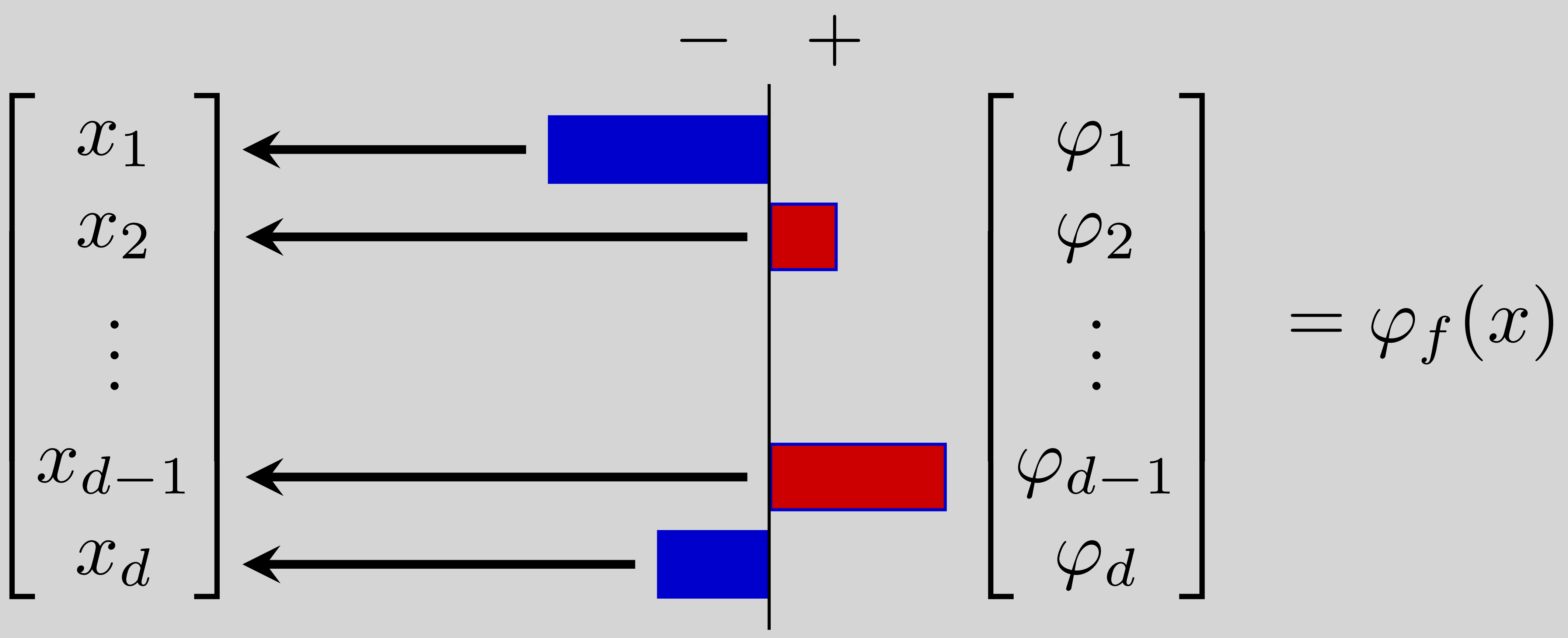
- **Local:** Only explain the part of f that is relevant for x
- **Post-Hoc:** The function f is given and fixed

Setting

Attribution methods

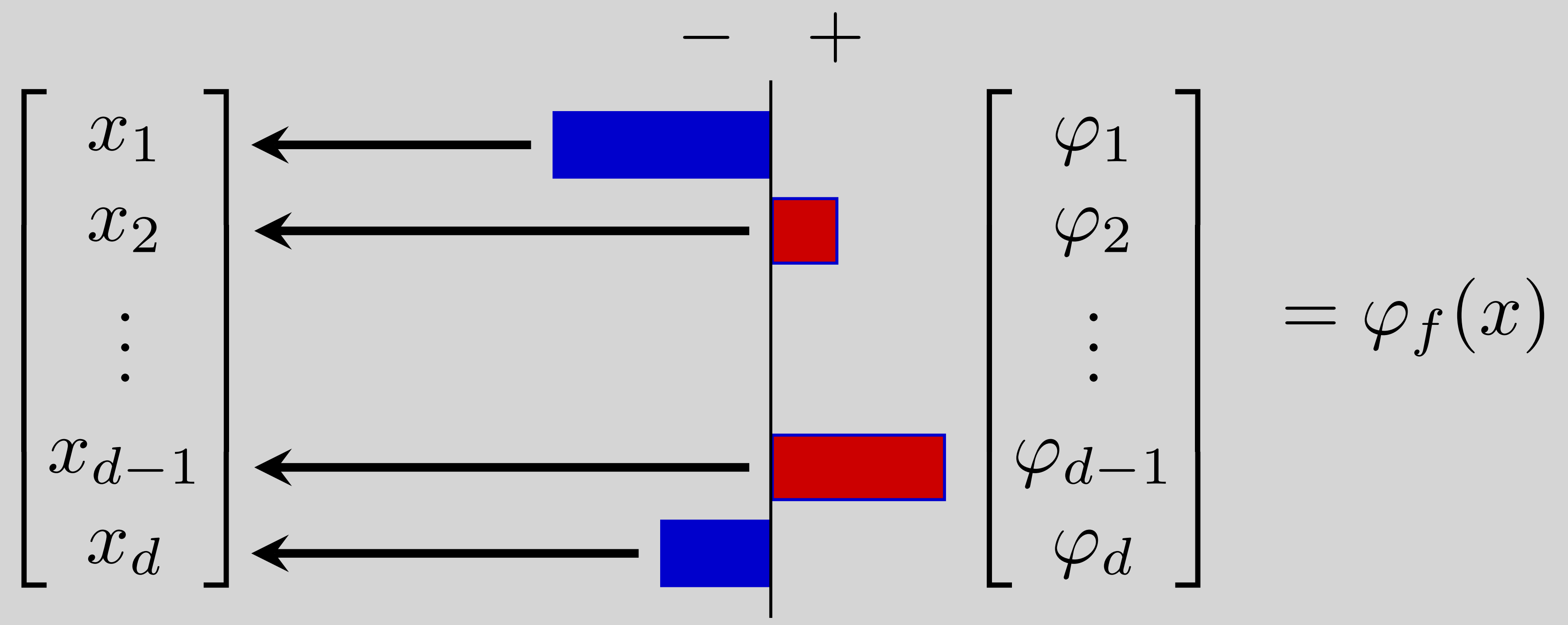
Machine learning model, e.g. a classifier:

$$f : \mathcal{X} \subseteq \mathbb{R}^d \to [0,1], \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \mapsto y$$



$\varphi_f(x) \in \mathbb{R}^d$ attributes a weight to each feature which explains how important the feature was for the classification of x of f

Example
Attribution methods



f linear, low dimension d

$$f(x) = \theta_0 + \sum_{i=1}^{d} x_i \theta_i$$

$$\varphi_f(x)_i = \theta_i$$

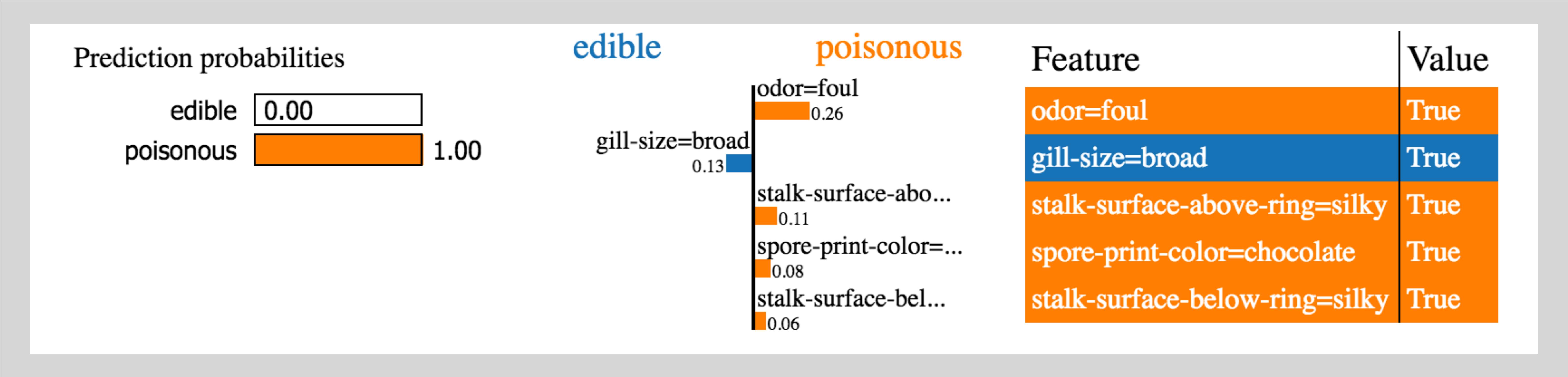
- In general φ_f will depend on x
- Most methods follow this example, by locally linearising around x

Examples (LIME)

Lime: Extract local linear approximations of f near x and report coefficients

- Optionally: Apply some dimension reduction before linearising.

Lime for tabular data[1]



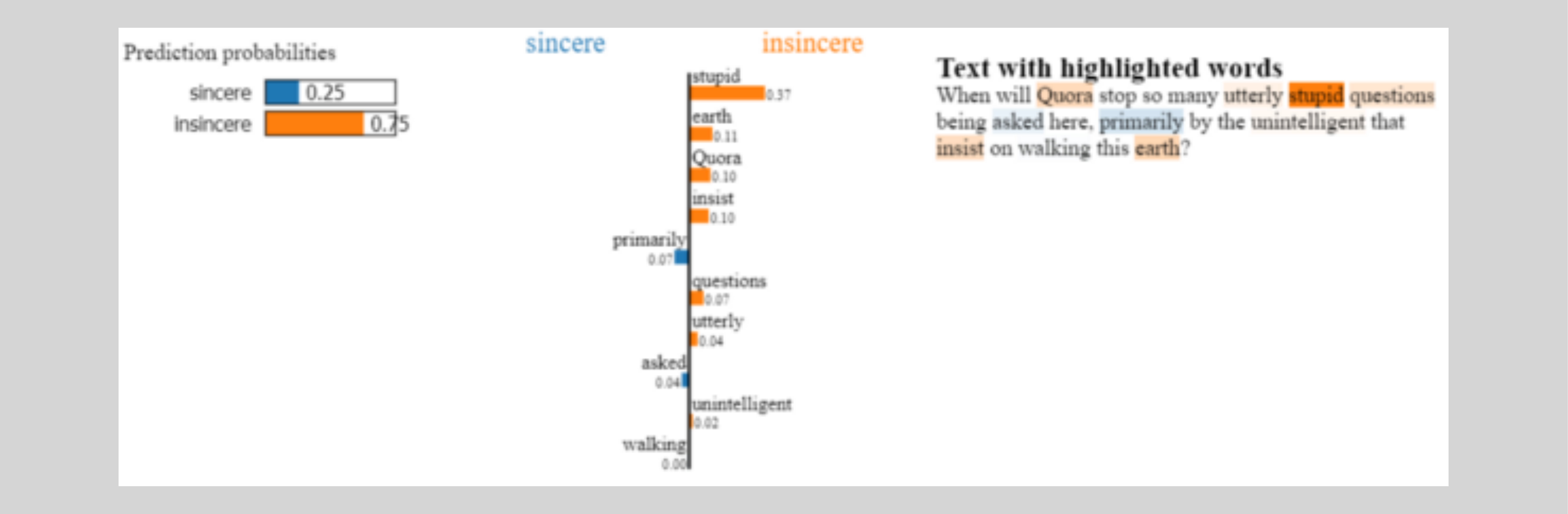
[1]Image source: https://github.com/marcoctr/lime

Examples (LIME for Text)

Lime: Extract local linear approximations of f near x and report coefficients

- Optionally: Apply some dimension reduction before linearising.

Lime for text data[2]



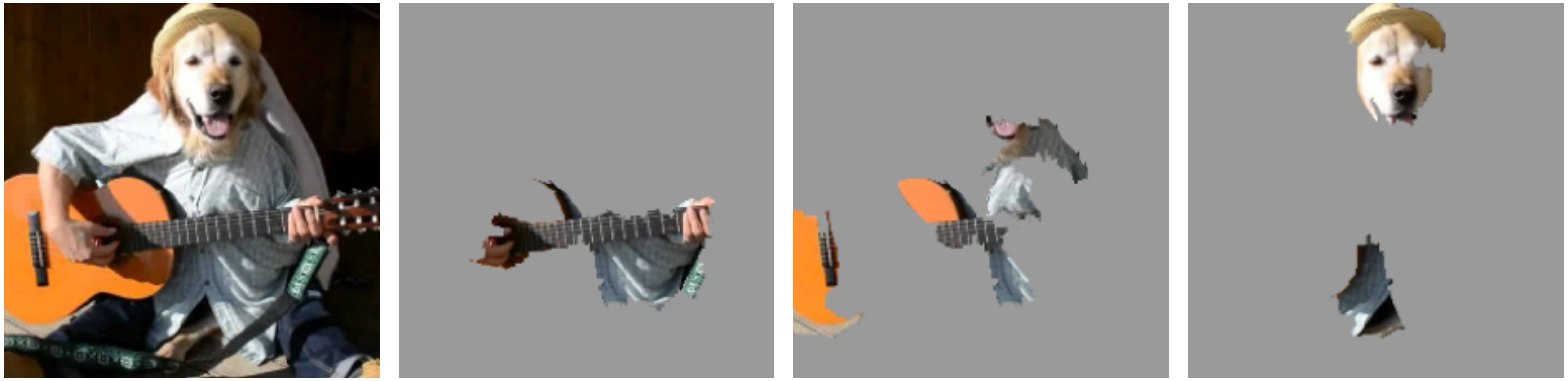
[2]Image source: https://towardsdatascience.com/what-makes-your-question-insincire-in-quora-26ee7658b010.

Examples (LIME for Images)

<u>Lime</u>: Extract local linear approximations of f near x and report coefficients

- Optionally: Apply some dimension reduction before linearising.

Lime for image data[3]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

[3] Image source: [Ribeiro et al., 2016]

Examples (Gradient)

Gradient methods

- Vanilla gradient:

$$\varphi_f(x) = \nabla f(x)$$

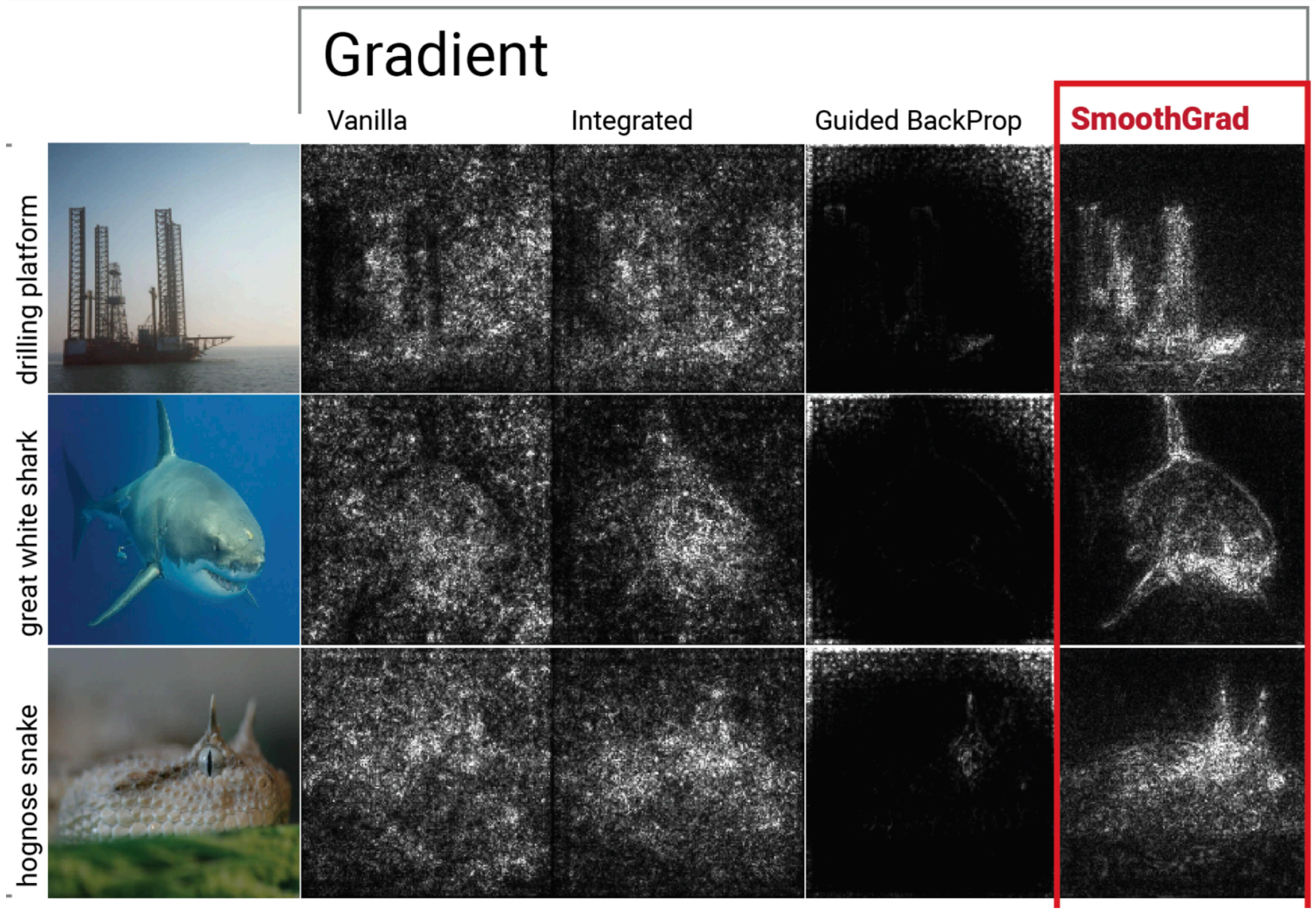
- SmoothGrad:

$$\varphi_f(x) = \mathbb{E}_{Z\sim\mathcal{N}(x,\Sigma)}[\nabla f(Z)]$$

- Integrated Gradients:

$$\varphi_f(x) = (x - x_0)\int_0^1 \nabla f(x_0 + t(x - x_0))\mathrm{d}t$$

- …

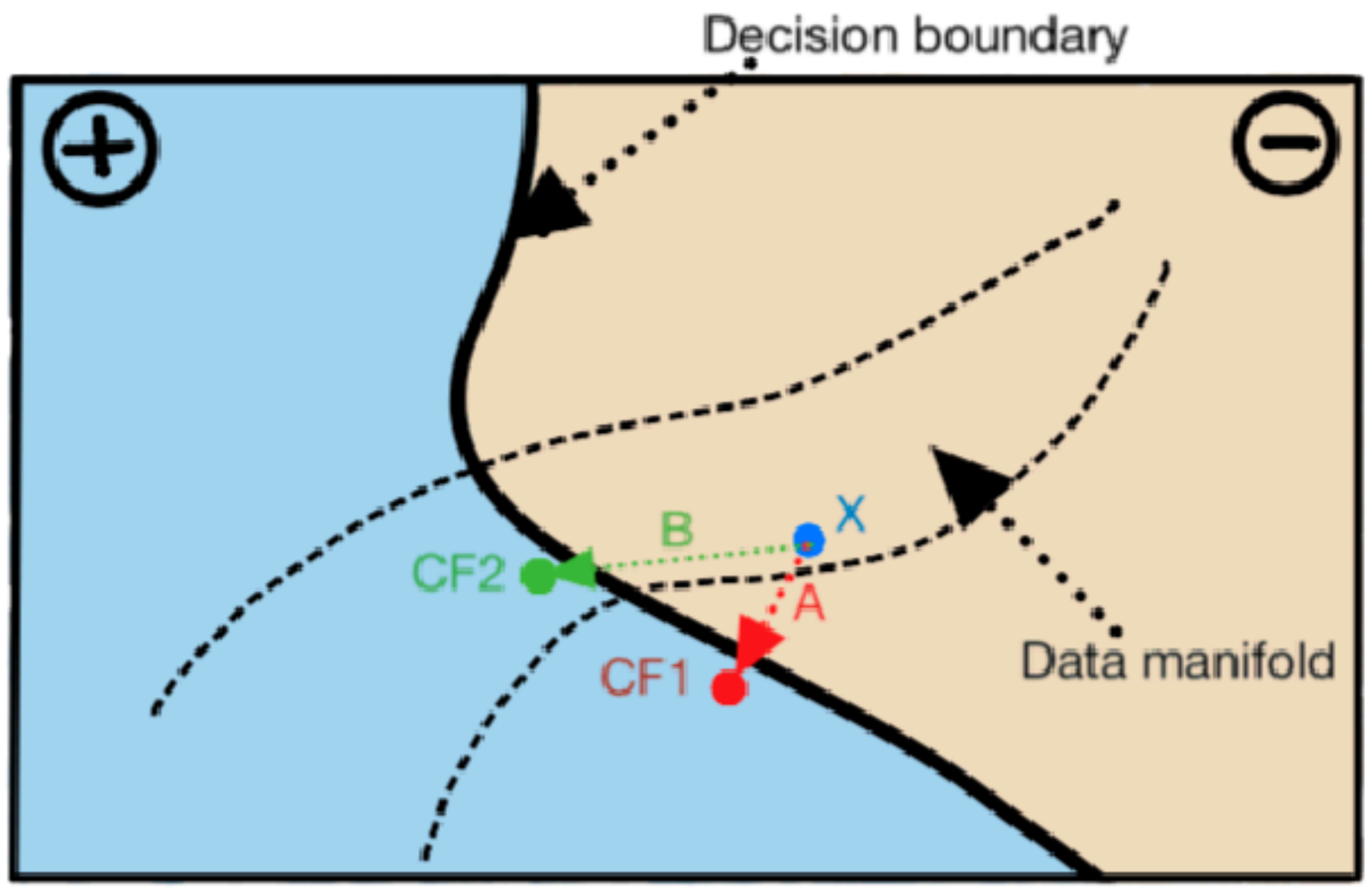


[4]Image source: [Smilkov et al., 2017]

Counterfactuals

Example

"If your Income would have been € 40.000,- instead of € 35.000,-, your loan application would have been accepted"



[5]Image source: [Verma et al., 2020]

Counterfactuals as attributions

Definition

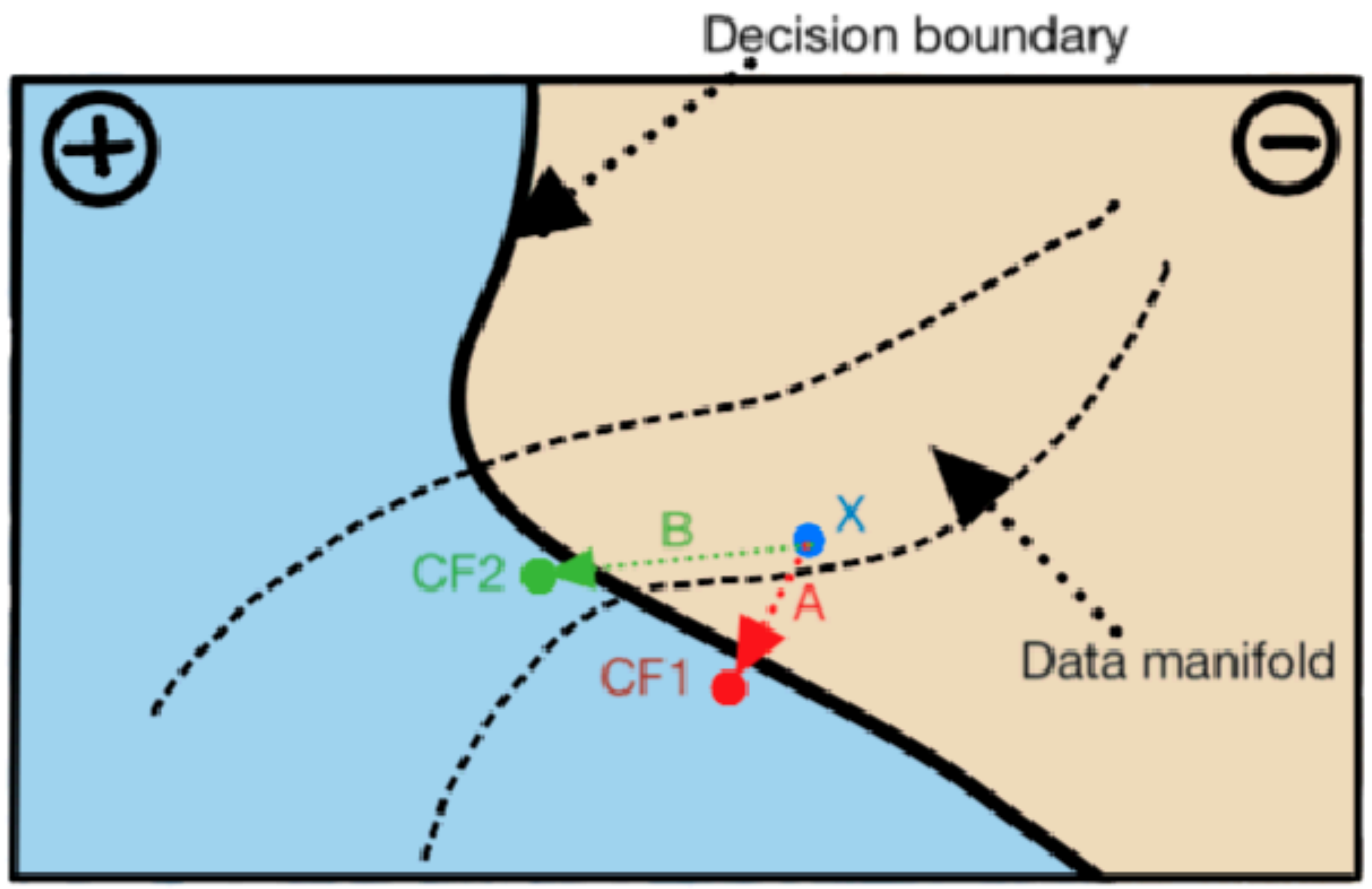
Consider Binary classification $f\colon \mathcal{X} \to \{-1,1\}$ and
let $x \in \mathcal{X}$.

A ***counterfactual*** x^{CF} for x is

$$x^{\mathrm{CF}} \in \arg\min_{y \in C} \|x - y\| \quad \mathrm{s.t.} \quad f(x^{\mathrm{CF}}) \neq f(x)$$

Counterfactuals can be seen as Attributions.
Write

$$\varphi_f(x) = x^{\mathrm{cf}} - x$$



[5]Image source: [Verma et al., 2020]

What are Good Explanations/Attributions?

- How to say some explanations are better than others?
- What is the (implicit) goal of the explanations?

[3]Image source: [Verma et al., 2020]

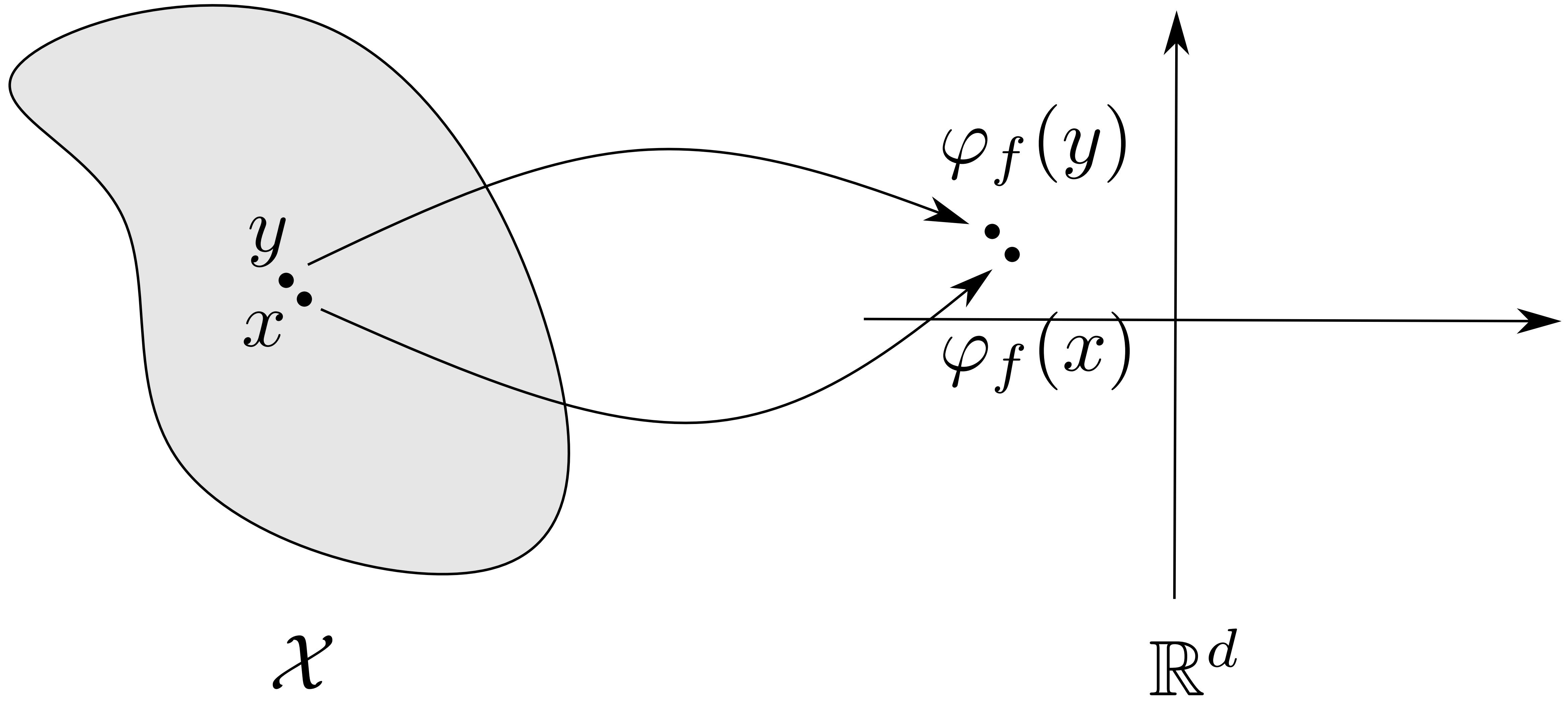
Robustness & Recourse sensitivity

Robustness

Definition

An attribution method φ_f for f is called ***Robust*** if it is continuous

Similar users require similar explanations



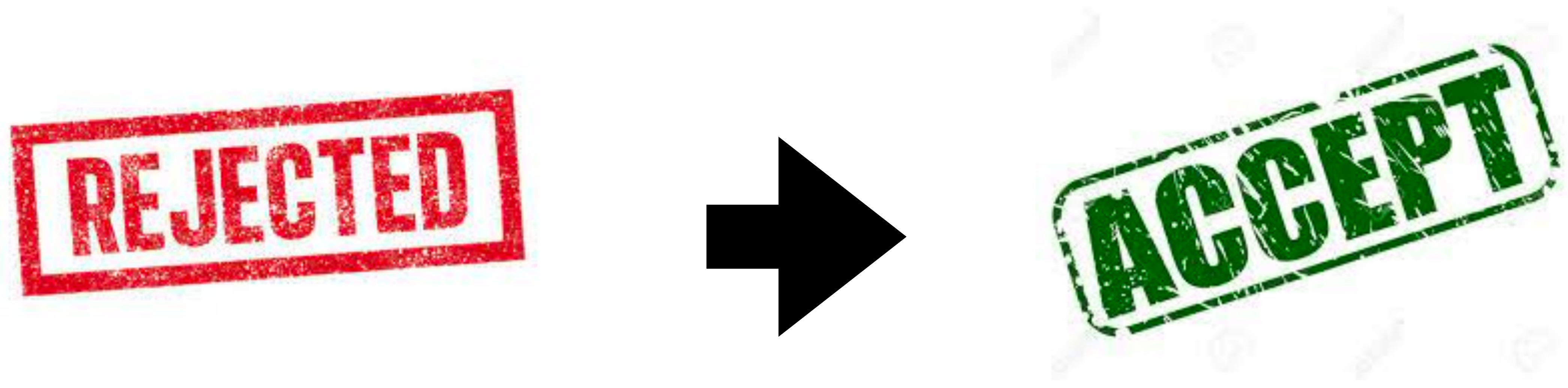
Recourse sensitivity

Motivation

User has some goal in mind:

- Wants to get a loan
- Increase their credit score
- Increase a probability
- Wants to upload a profile picture to get an OV card.

The explanation should allow the user to reach this goal

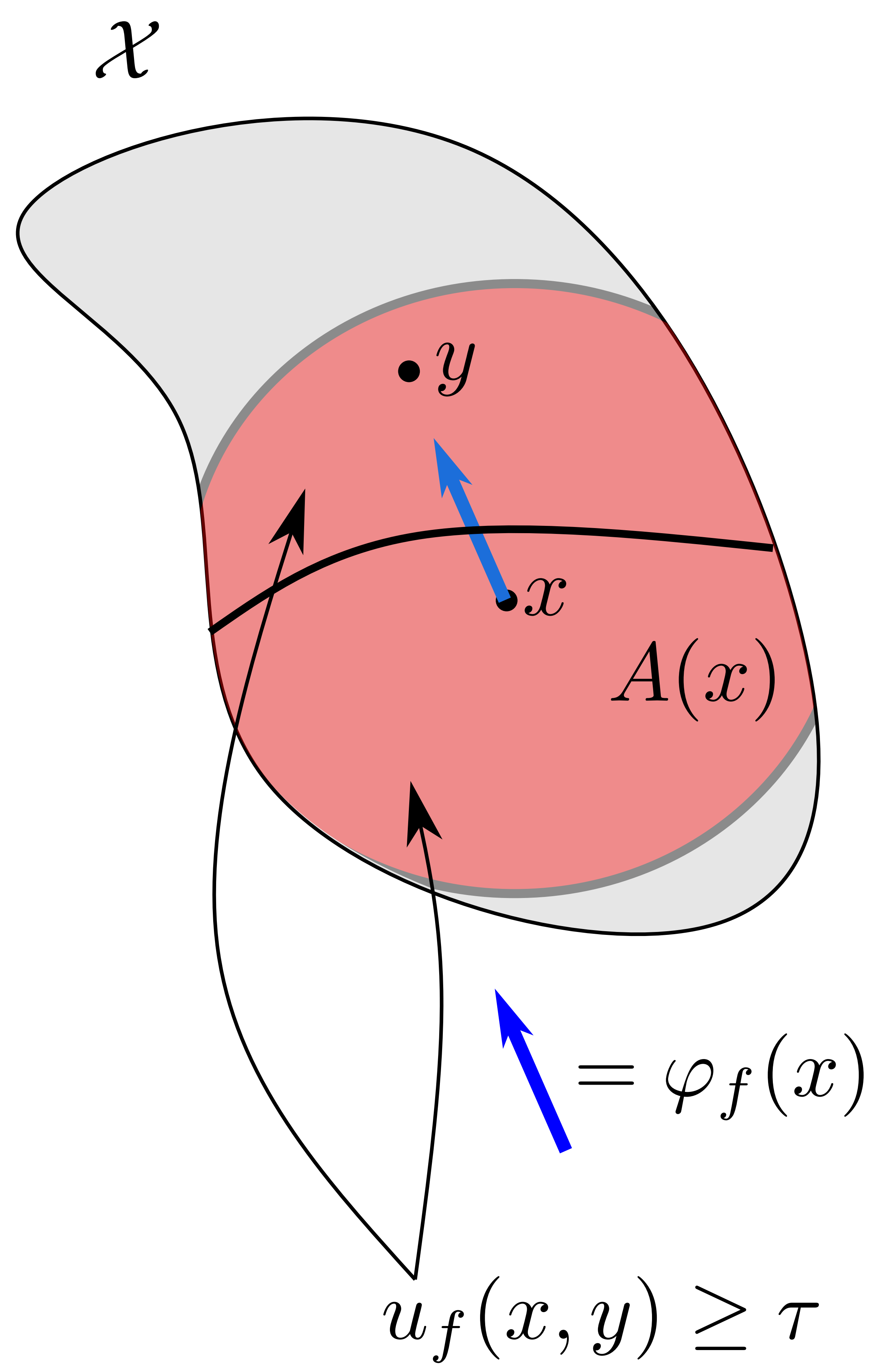


Recourse sensitivity

Informal definition

An Attribution method is called ***Recourse Sensitive*** if the user can achieve a sufficient utility increase when moving in the direction of $\varphi_f(x)$

This is very weak form of Recourse!



Recourse sensitivity

Utility & Attainable points

Measure if some utility $u_f \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exceeds some threshold $u_f(x, y) \geq \tau$:

- Undesirable classification: $u_f(x, y) = f(y) \geq 0$
- Increase score: $u_f(x, y) = f(y) - f(x) \geq \tau$
- Decrease a probability: $u_f(x, y) = \dfrac{f(x)}{f(y)} \geq \dfrac{1}{1-p} = \tau$

Define set of attainable points from x

$$A(x) = \{y \in \mathcal{X} \mid \|x - y\| \leq \delta, y \in C(x)\}$$

The set $C(x)$ will impose some constraints. Examples:

- $C(x) = \mathcal{X}$, Unrestricted case,
- $C(x) = \{y \in \mathcal{X} \mid \|x - u\|_0 \leq k\}$, Sparse change,
- $C(x) = \{u \in \mathcal{X} \mid y = x + \alpha z, \alpha \geq 0, z \in D\}$, Certain Directions D.

Recourse sensitivity

Definition

Definition

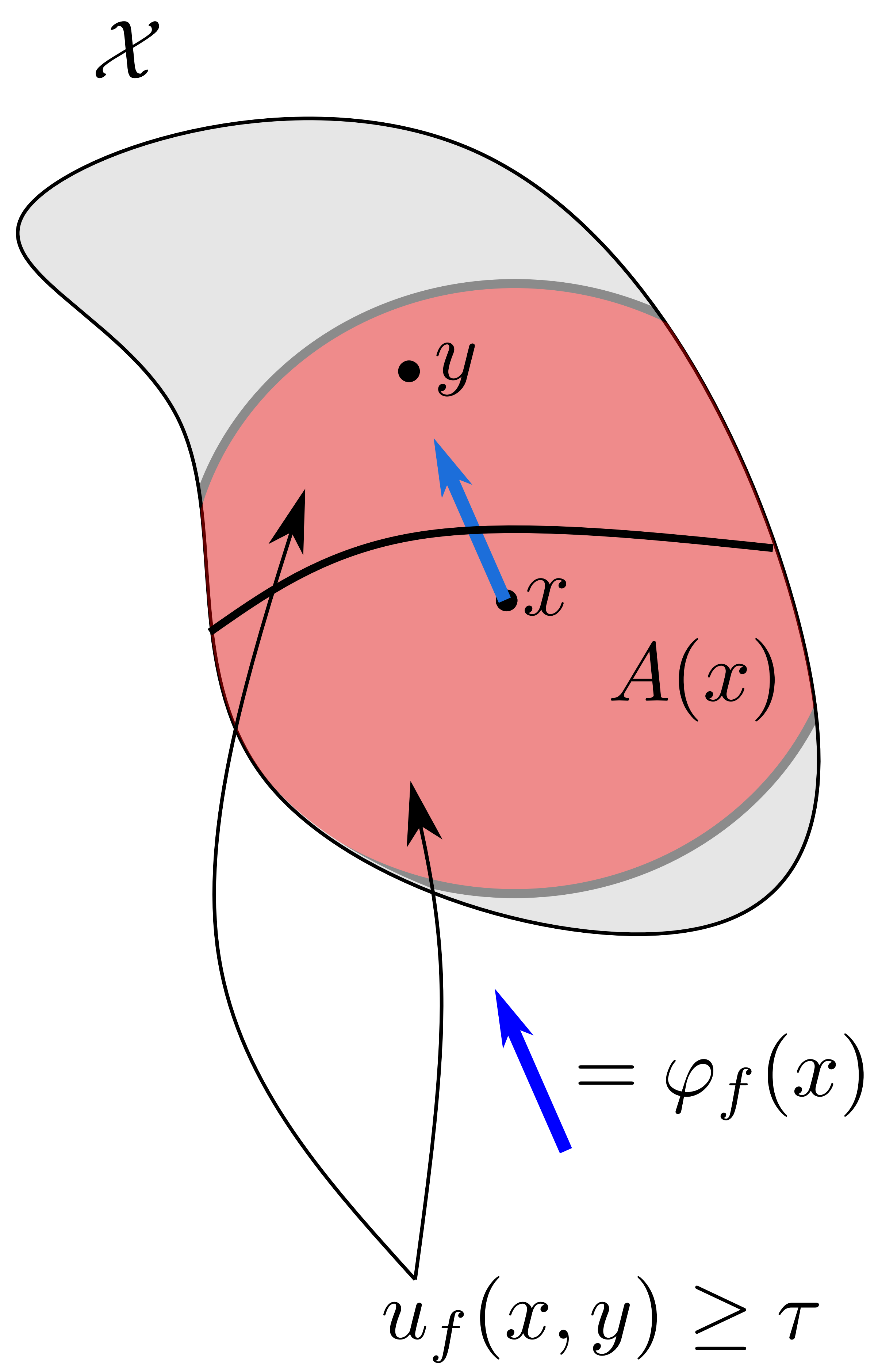
Consider the points close to x that achieve sufficient utility

$$U(x) = \{y \in \mathscr{X} \mid u_f(x, y) \geq \tau,\ \|x - y\| \leq \delta\}$$

An Attribution function φ_f is called ***Recourse Sensitive*** if

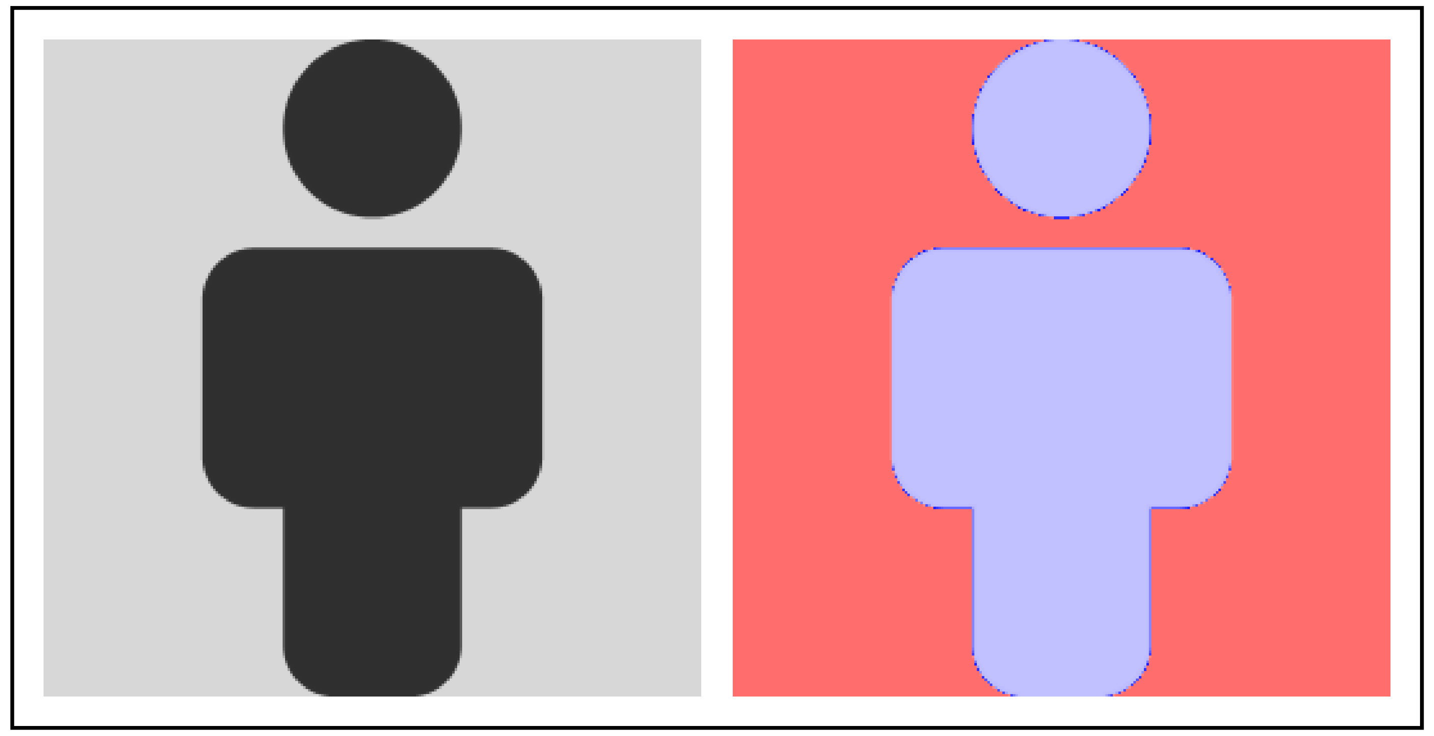
$$\varphi_f(x) = \alpha(y - x), \qquad \alpha > 0 \text{ and } y \in U(x),$$

for all $x \in \mathscr{X}$ for which $U(x) = \emptyset$.

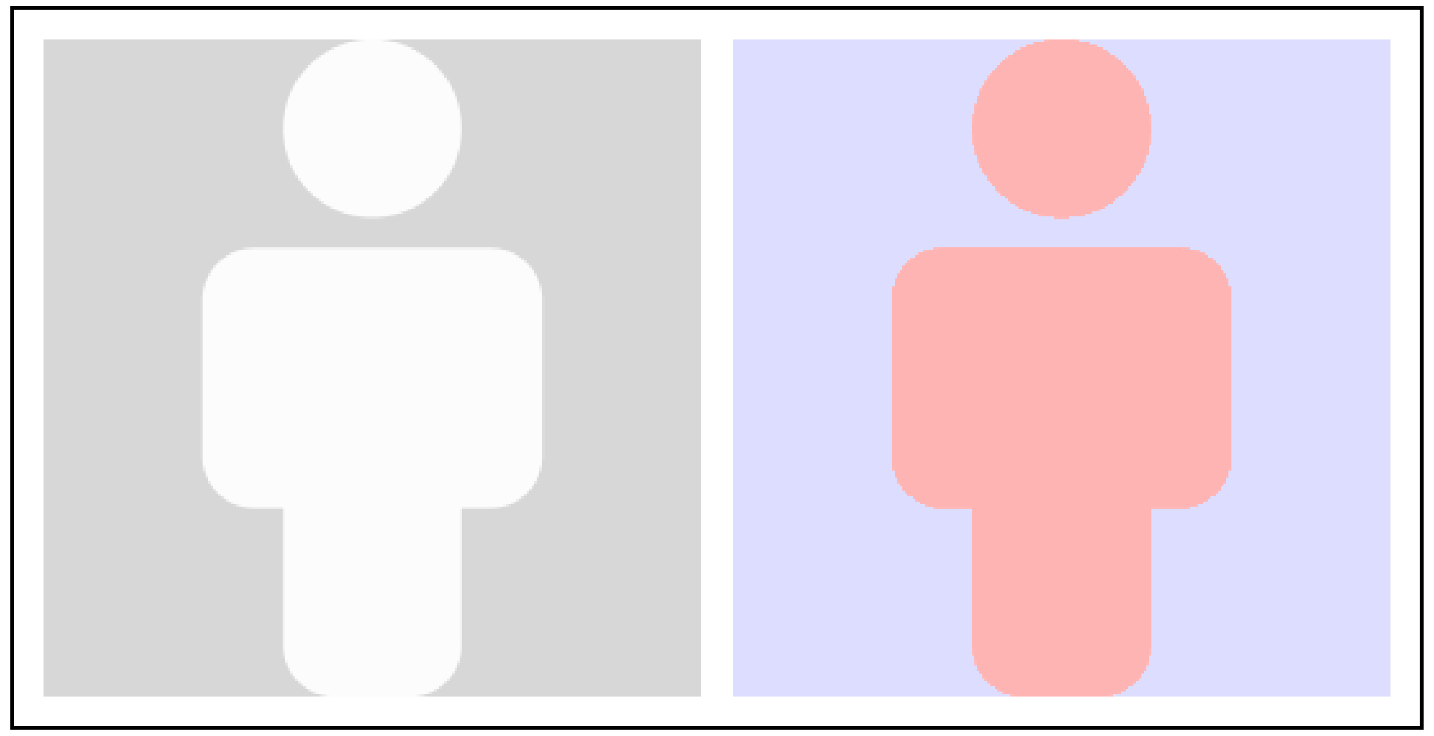


Recourse sensitivity

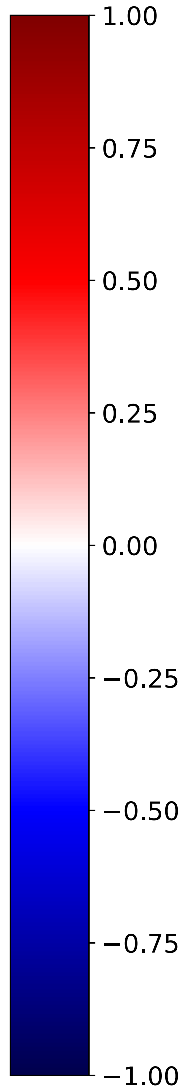
Example



(a) Accepted profile picture



(b) Rejected profile picture



Impossibility

Impossibility result

Attribution methods cannot always

- Provide Recourse
- Be Robust

Impossibility result

Specific case (Binary classification)

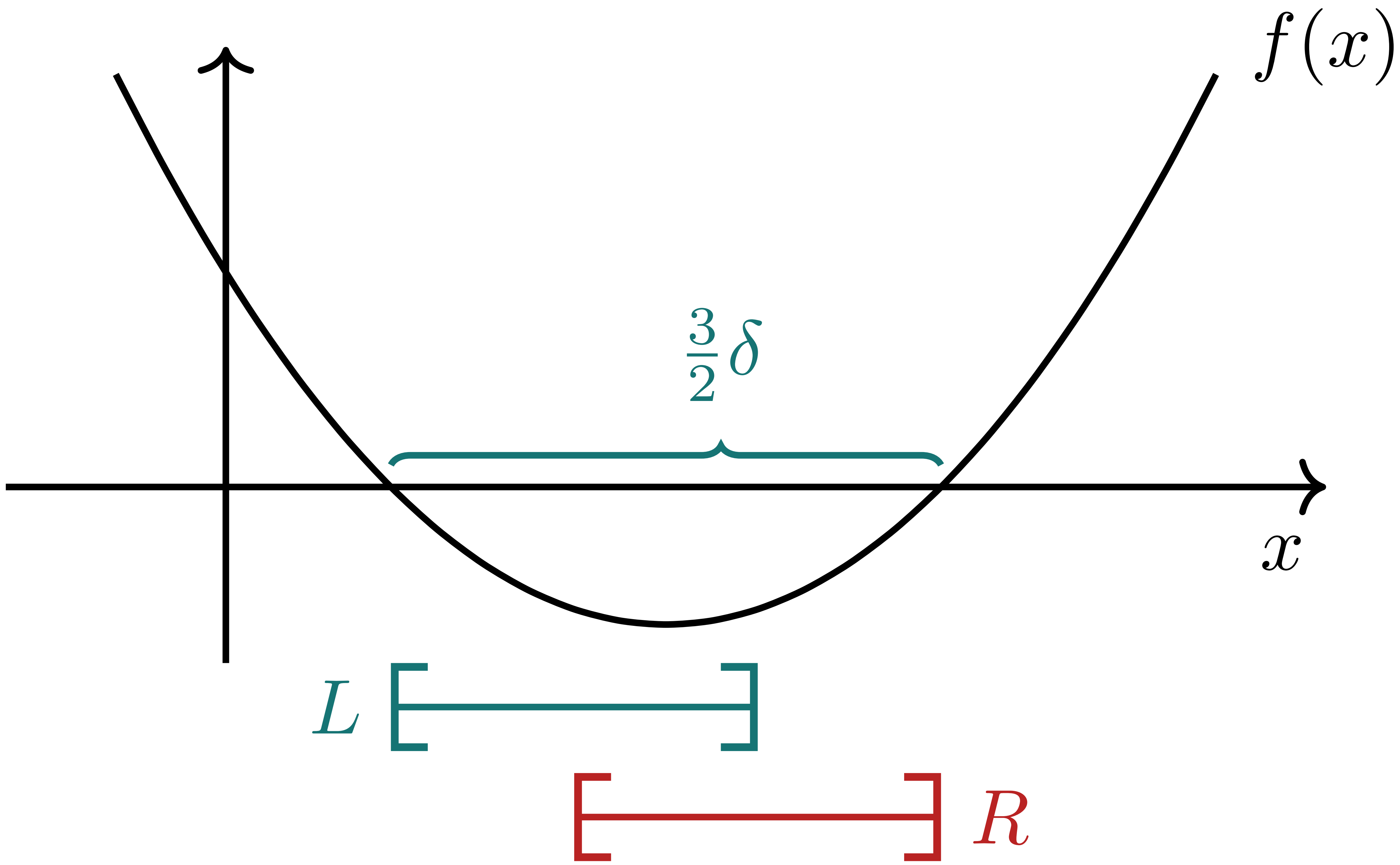
Setting

- $\mathscr{X} = \mathbb{R}^d$,
- $u_f(x, y) = f(y)$,
- $\tau = 0, \delta > 0.$

Theorem

There exists a continuous function f such that no attribution method φ_f can be both recourse sensitive and continuous

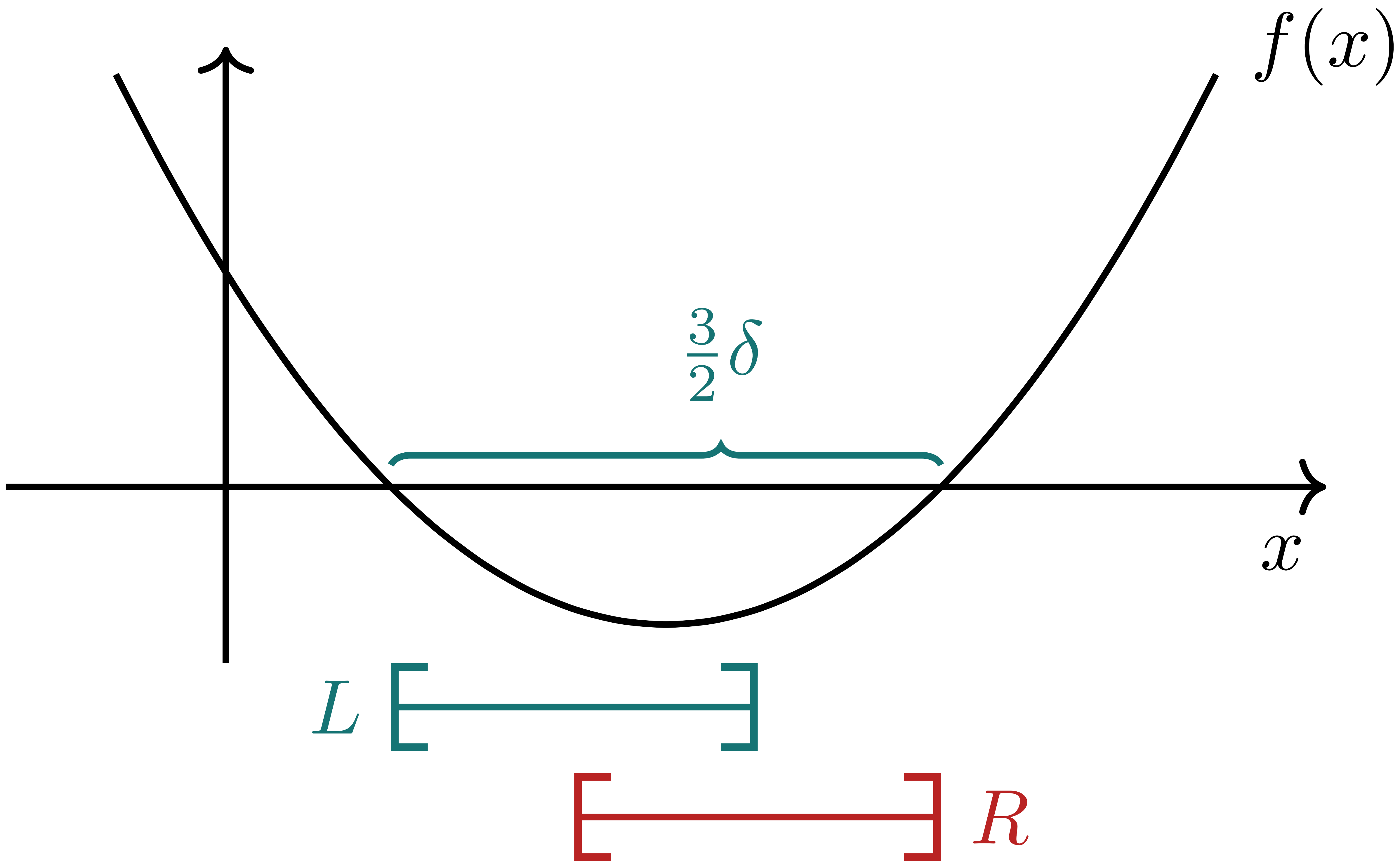
Proof sketch



$R = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

$L = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

Proof sketch

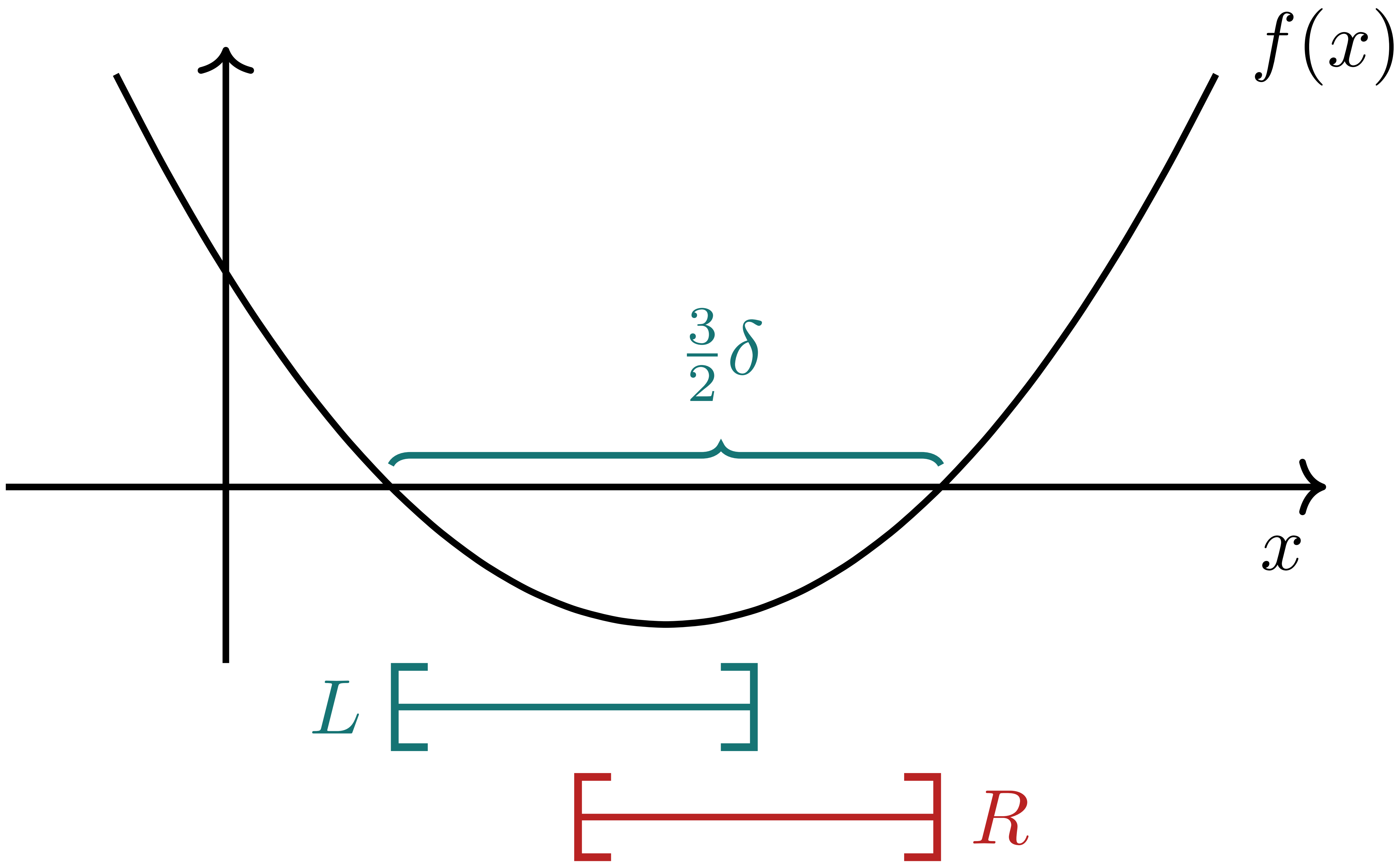


$R = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

$L = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

$$\varphi_f(x) = \begin{cases} < 0 & \text{for } x \in L \backslash R \\ > 0 & \text{for } x \in R \backslash L \\ \neq 0 & \text{for } x \in L \cap R \end{cases}$$

Proof sketch



$R = \{x \mid$ recourse is possible by moving at most δ left$\}$
$L = \{x \mid$ recourse is possible by moving at most δ left$\}$

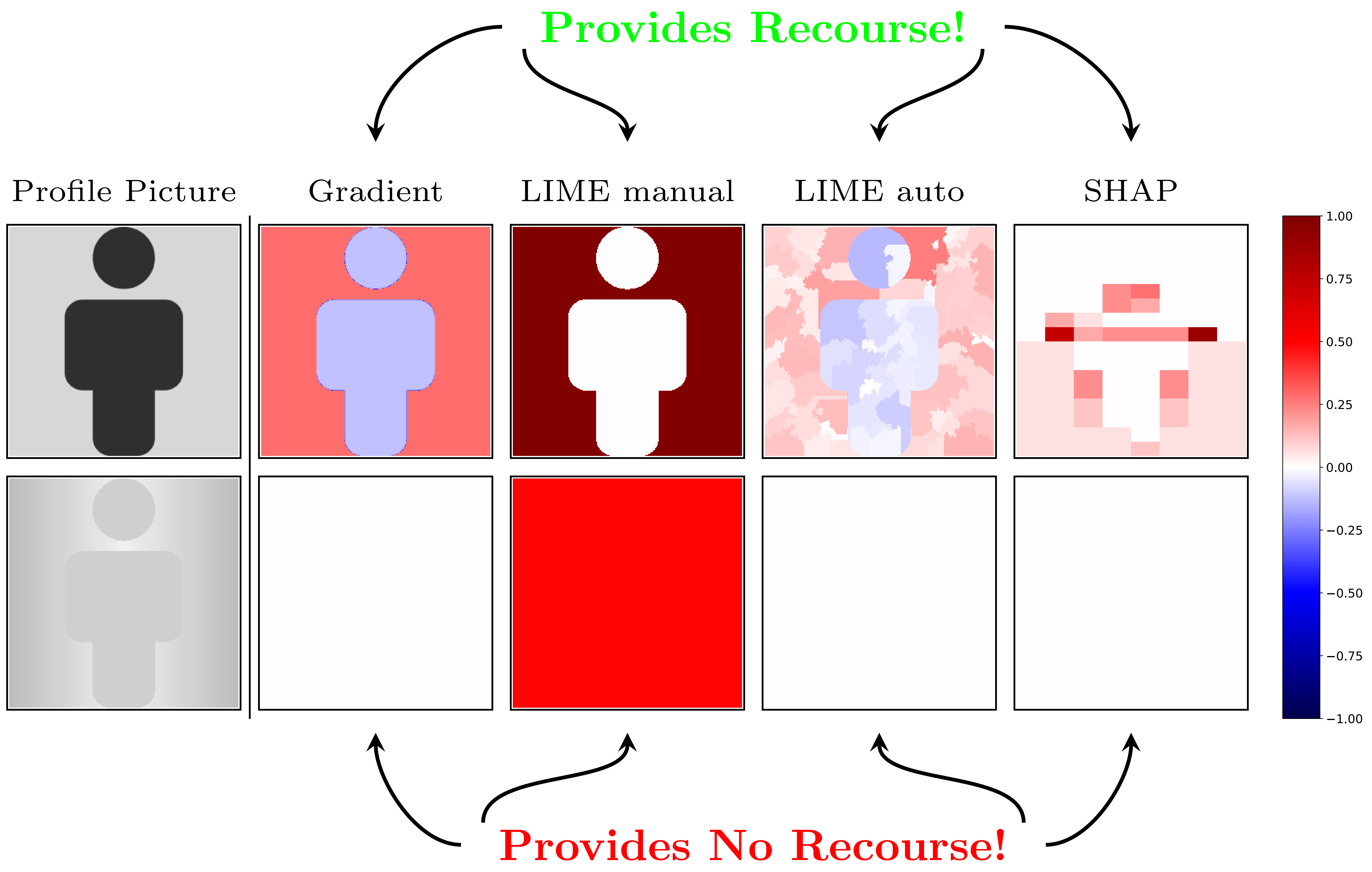
$$\varphi_f(x) = \begin{cases} < 0 & \text{for } x \in L \setminus R \\ > 0 & \text{for } x \in R \setminus L \\ \neq 0 & \text{for } x \in L \cap R \end{cases}$$

But this contradicts continuity!
(By the intermediate-value theorem)

This example can be embedded into higher dimensions

Recourse sensitivity

Example



Impossibility result

General case

Theorem

If u_f is of the form $u_f(x, y) = \widetilde{u}(f(x), f(y))$ and if there exist $z_1, z_2 \in \mathbb{R}^d$ such that $\widetilde{u}(z_1, z_2) \geq \tau$ and $\widetilde{u}(z_1, z_1) < \tau$.

Then, there exists a continuous $f \colon \mathscr{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and robust.

Attribution methods cannot always

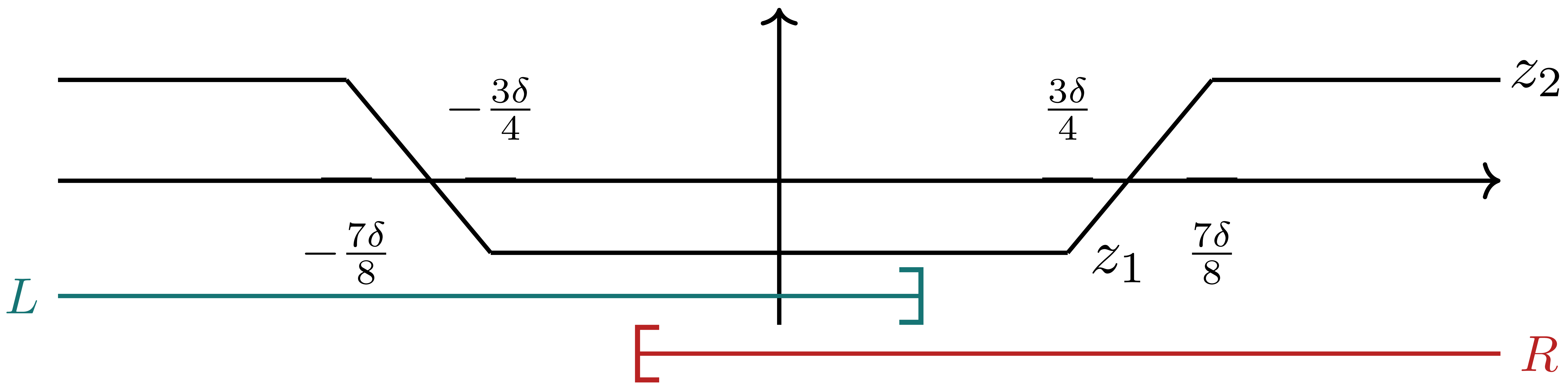
- Provide recourse
- Be continuous

Impossibility result

Proof sketch

Define

- Take z_1, z_2 such that $\widetilde{u}(z_1, z_2) \geq \tau$.
- $L = \{x \in \mathcal{X} \mid$ there exists some $y \in [x - \delta, x]$ with $u_f(x, y) \geq \tau\}$,
- $R = \{x \in \mathcal{X} \mid$ there exists some $y \in [x + \delta, x]$ with $u_f(x, y) \geq \tau\}$.

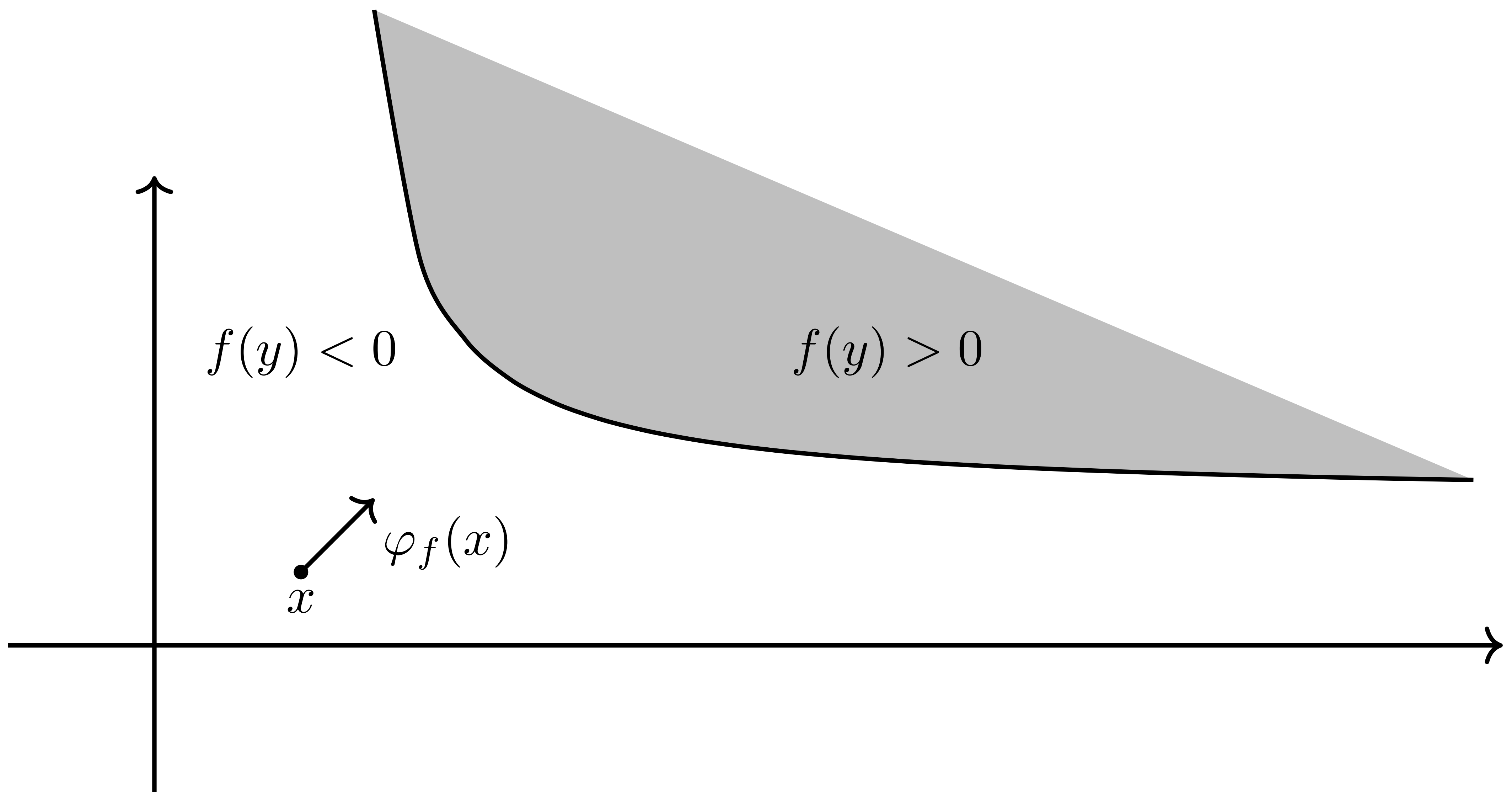


When is Recourse and Robustness possible?

Recourse and Robustness is possible sometimes

Binary classification

- Preferred class ($u_f(x, y) = f(y) \geq 0$)
- Let $U = \{x \in \mathscr{X} \mid f(x) > 0\}$ be convex
- Then Recourse and Robustness is possible!

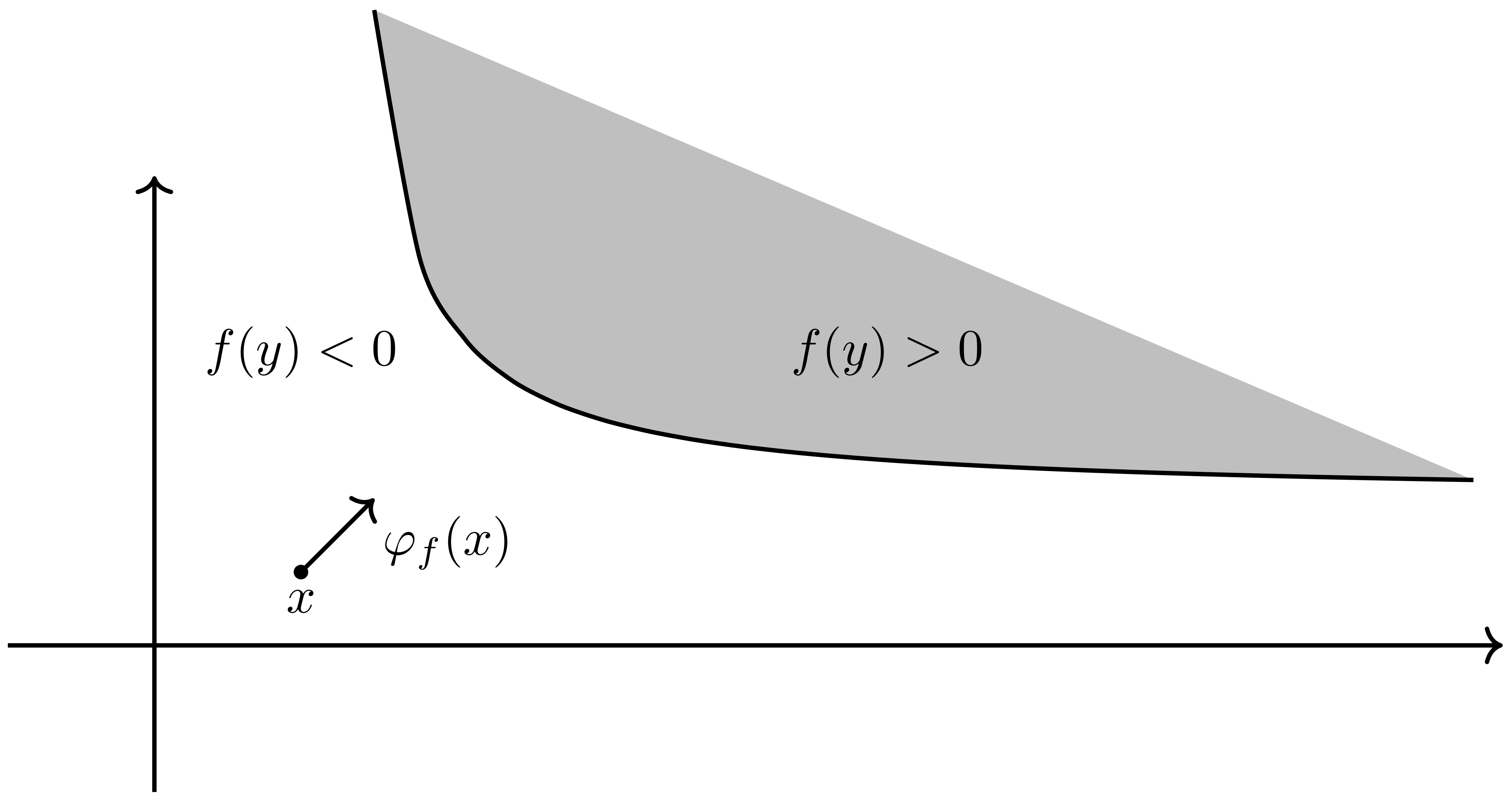


$$\varphi_f(x) = P_U(x) - x$$

Recourse and Robustness is possible sometimes

General case

- General Utility $u_f(x, y)$
- $U(x) = \{y \,|\, u_f(x, y) \geq \tau\}$ become x dependent
- We need:
  - "Continuity of $U(x)$"
  - Projections should exist and be unique



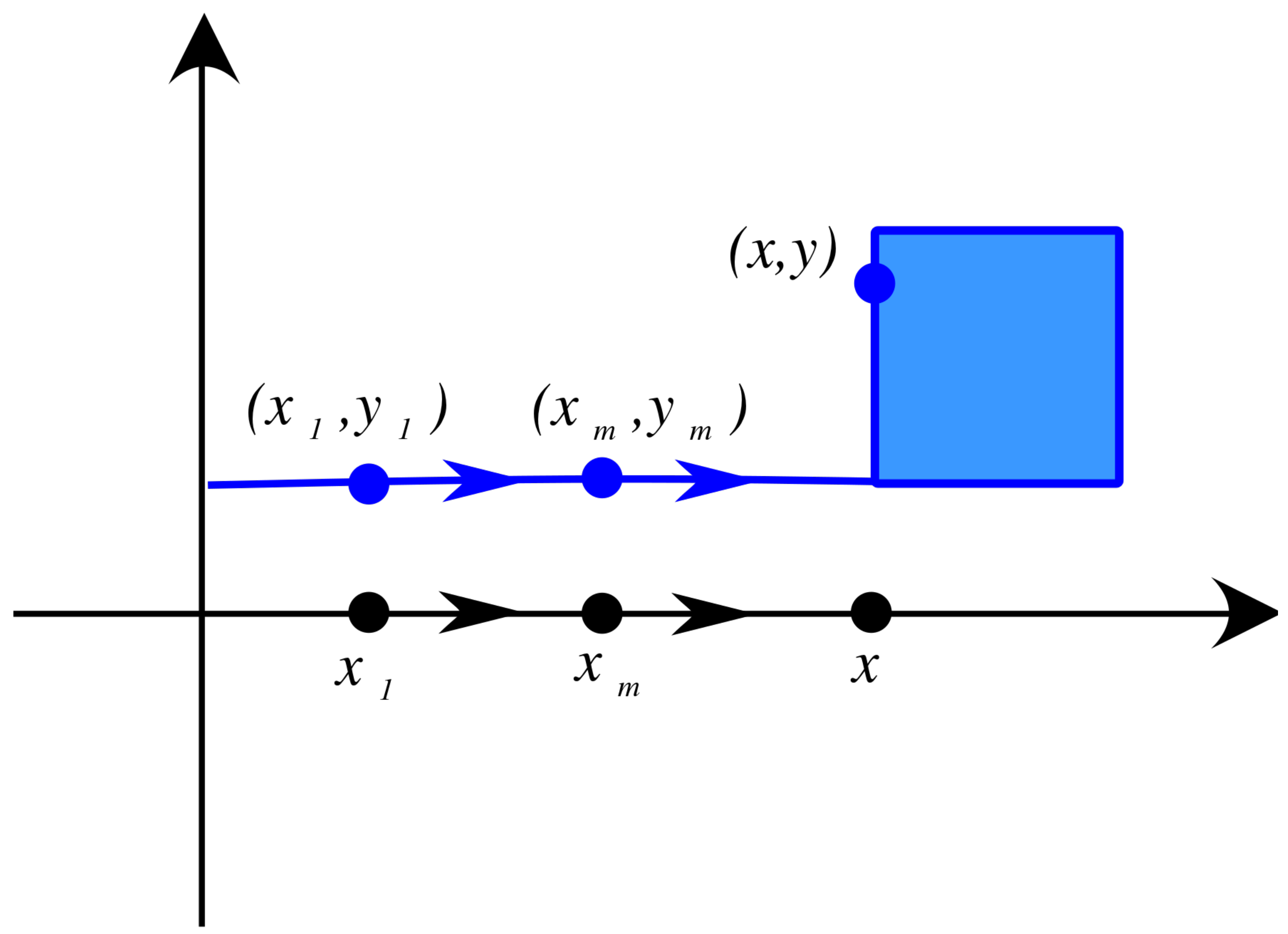
$$\varphi_f(x) = P_U(x) - x$$

Hemi-continuity

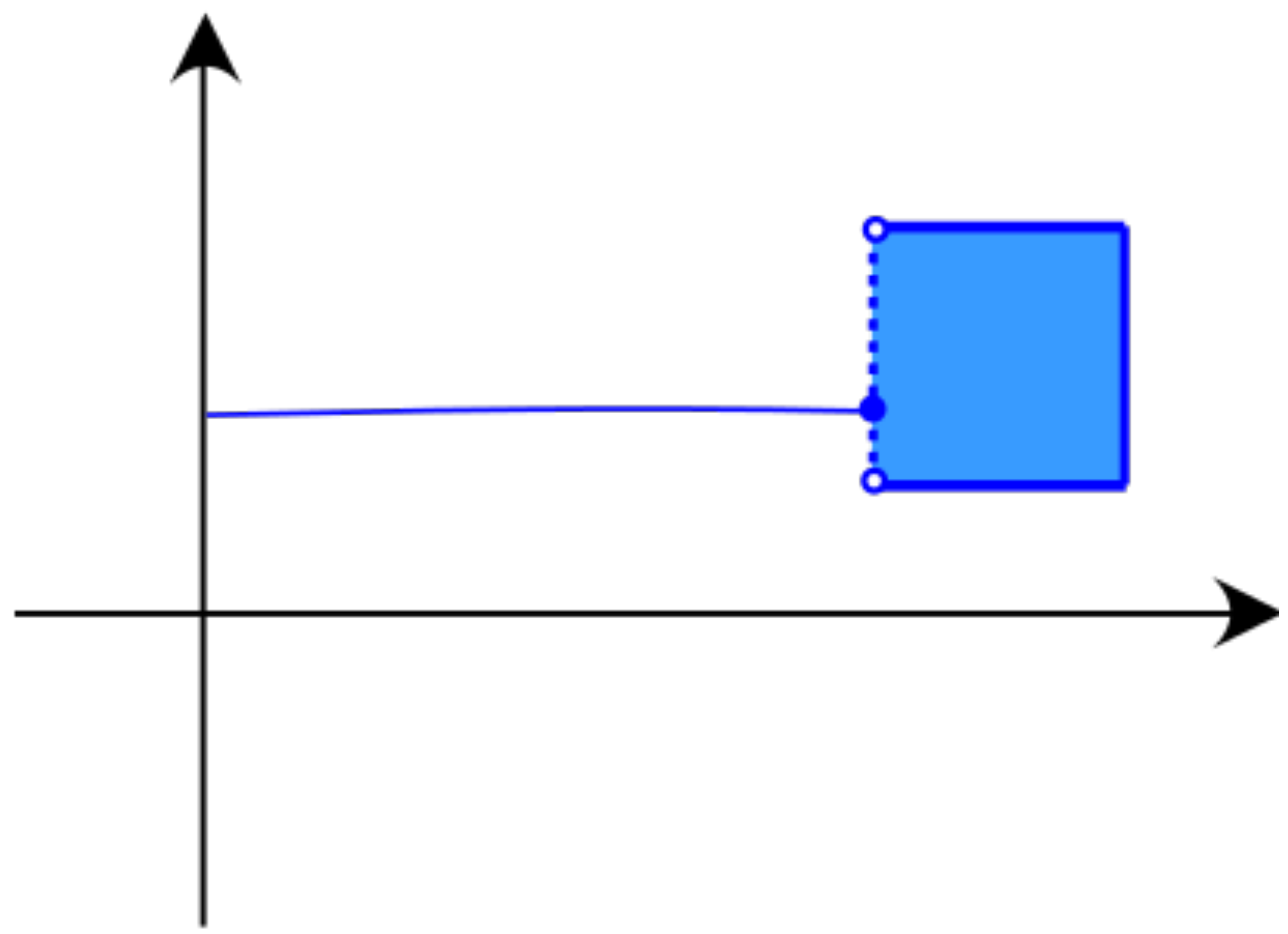
Set-valued function $U \colon \mathscr{X} \to 2^{\mathscr{Y}}$:

- Upper Hemi-continuity: U(x) cannot suddenly explode
- Lower Hemi-continuity: U(x) cannot suddenly implode

UHC, but not LHC[4]



LHC, but not UHC[4]



[5]Image source: https://en.wikipedia.org/wiki/Hemicontinuity

Recourse and Robustness is possible sometimes

General case

Theorem

Let $\delta > 0$, $\tau \geq 0$, $f\colon \mathscr{X} \to \mathbb{R}$ be a continuous function and $u_f(x, y)$ a utility function with the following properties:

1. For every $x \in \mathscr{X}$, the projection onto $U(x)$ exists and is unique;
2. The set-valued function $U(x)$ is Hemi-continuous and closed.

Then the function given by:

$$\varphi_f(x) = \underset{y \in U(x)}{\text{arg min}} \, \|x - y\| - x = P_{U(x)}(x) - x$$

Is a recourse sensitive and robust attribution map.

Recourse and Robustness is possible sometimes

General case

Berge's Maximum Theorem

Let $\mathscr{X}, \mathscr{Y} \subseteq \mathbb{R}^d$, assume that:

1. The function $v : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ is a continuous function;
2. The set-valued $U \colon \mathscr{X} \to 2^{\mathscr{Y}}$ is semi-continuous, never empty, and attains compact sets.

Then, the parametrized optimization problem $v^*(x) := \inf_{yinU(x)} v(x, y)$ is continuous and the set-valued solution function $U^*(x) = \{y \in U(x) \mid v^*(x) = v(x, y)\}$ is UHC and compact-valued

Recourse and Robustness is possible sometimes

General case

Theorem

Let $\delta > 0$, $\tau \geq 0$, $f \colon \mathcal{X} \to \mathbb{R}$ be a continuous function and $u_f(x, y)$ a utility function with the following properties:

1. For every $x \in \mathcal{X}$, the projection onto $U(x)$ exists and is unique;
2. The set-valued function $U(x)$ is Hemi-continuous and closed.

Then the function given by:

$$\varphi_f(x) = \underset{y \in U(x)}{\text{arg min}} \, \|x - y\| - x = P_{U(x)}(x) - x$$

Is a recourse sensitive and robust attribution map.

Proof idea:

- Berge's Maximum Theorem gives Continuity almost immediately
  - $v(x, y) = \|x - y\|$
  - $U(x)$ is not compact, but there are general versions of Berge that deal with this case
- Check that Recourse sensitivity is satisfied

Full Characterisation in the Single-Feature Case

Attribution methods can

- Provide Recourse
- Be Robust

In the case of ***Single Feature Recourse Sensitivity*** and L and R are sufficiently disjoint.

Characterisation

Warm up,
One dimensional case

Define

- $L = \{x \in \mathcal{X} \mid$ there exists some $y \in [x - \delta, x]$ with $u_f(x, y) \geq \tau\}$,
- $R = \{x \in \mathcal{X} \mid$ there exists some $y \in [x + \delta, x]$ with $u_f(x, y) \geq \tau\}$.

Theorem

Let $\delta, \tau > 0, f : \mathcal{X} \to \mathbb{R}$ and let u_f be a utility function with $u_f(x, x) < \tau$ for all $x \in \mathcal{X}$. Then, there exists a continuous recourse sensitive attribution method φ_f for f if and only if there exists $\widetilde{L} \subseteq L$ and $\widetilde{R} \subseteq R$ such that $\widetilde{L} \cup \widetilde{R} = L \cup R$
and $\widetilde{L}$ and $\widetilde{R}$ are separated.

Definition

Two sets $A, B \subseteq \mathcal{X}$ are called *separated* if $\mathrm{cl}(A) \cap B = A \cap \mathrm{cl}(B) = \emptyset$

Characterisation

Warm up,
One dimensional case

L
R

Allowed	*Not Allowed*
[ ) ( ]	[( )]
[( )]	[ ]( ]
[ )( ]	[ )[ ]

Conclusion

Summary:

- There exist f for which recourse sensitivity + robustness is impossible, for several machine learning tasks
- There are cases for which it is possible, but they require strong conditions
- Full Characterisation for Single-Feature case
- Further extensions in the paper:
  - Sufficient Conditions for when Recourse and Robustness is possible
  - Discussion on possible ways around this Impossibility result
  - Constraints on user actions

Thank you for your attention!

References

- Fokkema, Hidde, Rianne de Heide, and Tim van Erven. "Attribution-based Explanations that Provide Recourse Cannot be Robust." arXiv preprint arXiv:2205.15834 (2022).
- M. T. Ribeiro, S. Singh, and C. Guestrin. ”Why should I trust you?” explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.
- D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. ArXiv:1706.03825, 2017.
- Verma, Sahil, John Dickerson, and Keegan Hines. "Counterfactual explanations for machine learning: A review." arXiv preprint arXiv:2010.10596 (2020).
- Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.
- Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
- Leavitt, Matthew L., and Ari Morcos. "Towards falsifiable interpretability research." arXiv preprint arXiv:2010.12016 (2020).