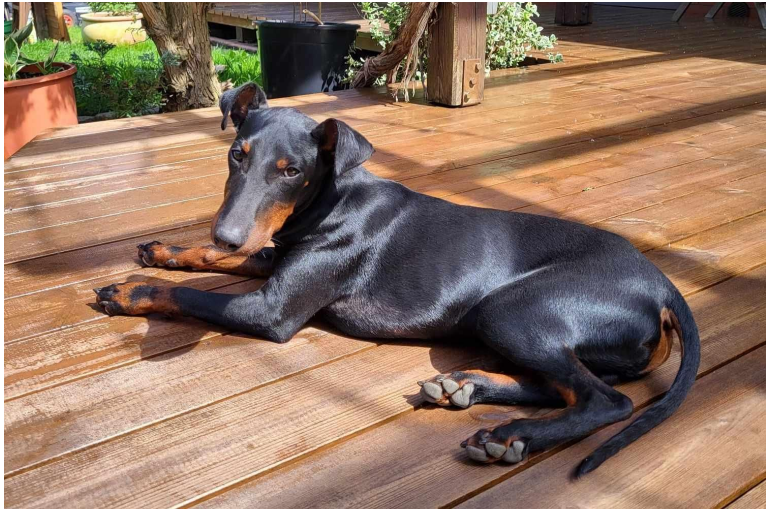
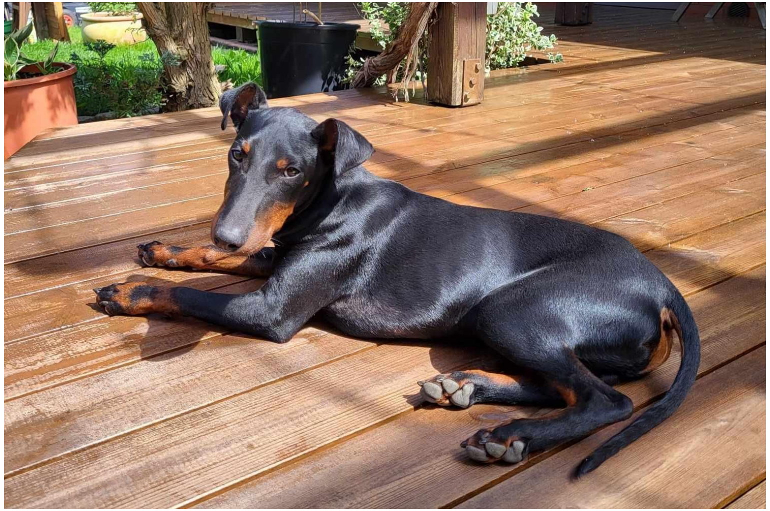
+ 4 months

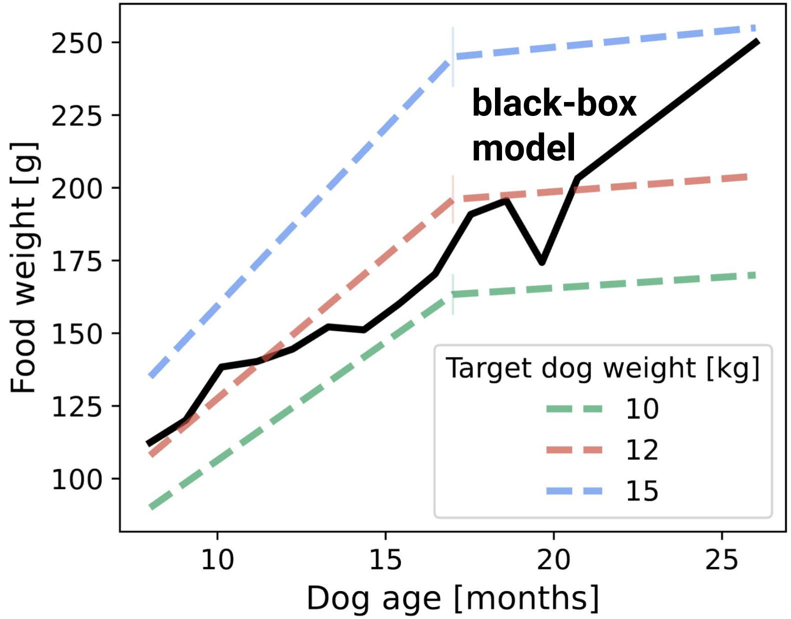


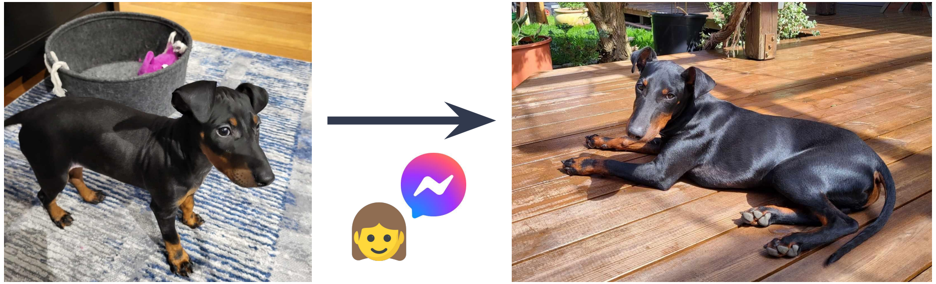
+ 4 months



A dog aged 2–4 months with a target weight of 10–15 kg should eat 90–235 g daily.
aged 4–6 months with a target weight of 10–15 kg should eat 170–255 g

How much should the dog (with a target weight of 12 kg) eat every week?





Global Feature Effect Explanation

e.g. PDP, ALE, SHAP dependence

(x) explained feature

(y) effect on model prediction

(color, **invisible**) marginalised features – dog weight, **breed**, ...

On the Robustness of Global Feature Effect Explanations

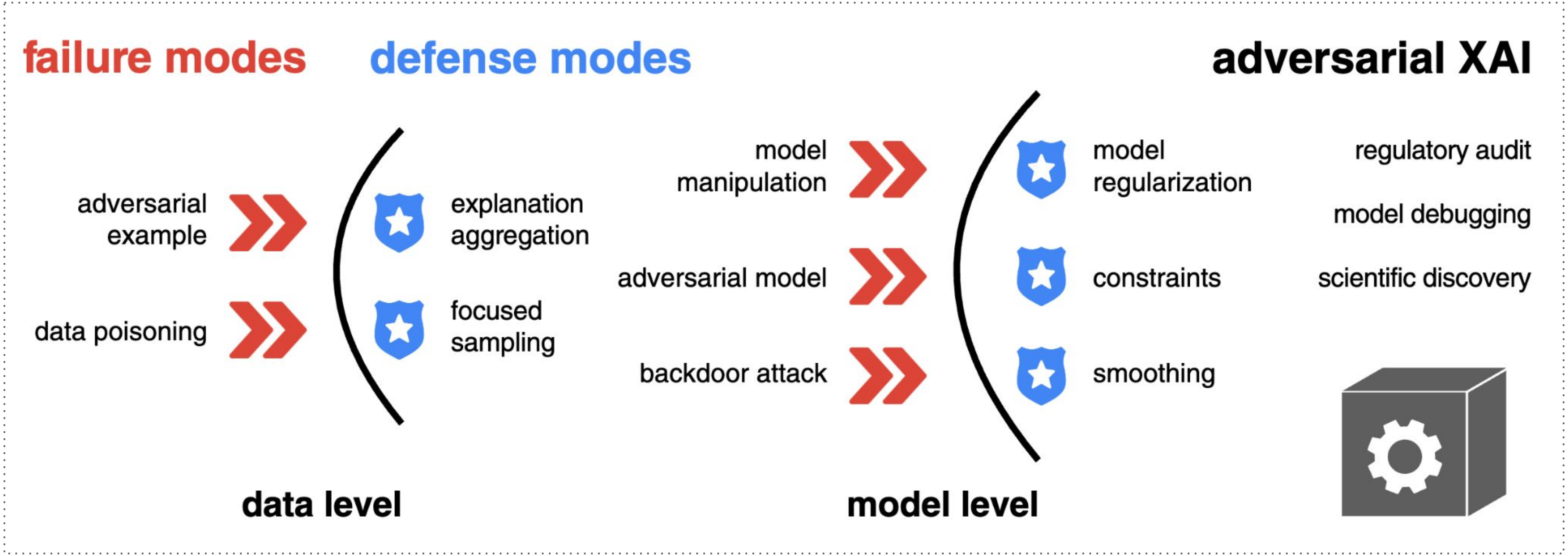
Hubert Baniecki, Giuseppe Casalicchio, Bernd Bischl, Przemyslaw Biecek

UNIVERSITY OF WARSAW
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Munich Center for Machine Learning
Warsaw University of Technology

ECML PKDD 2024

Adversarial robustness of explanations

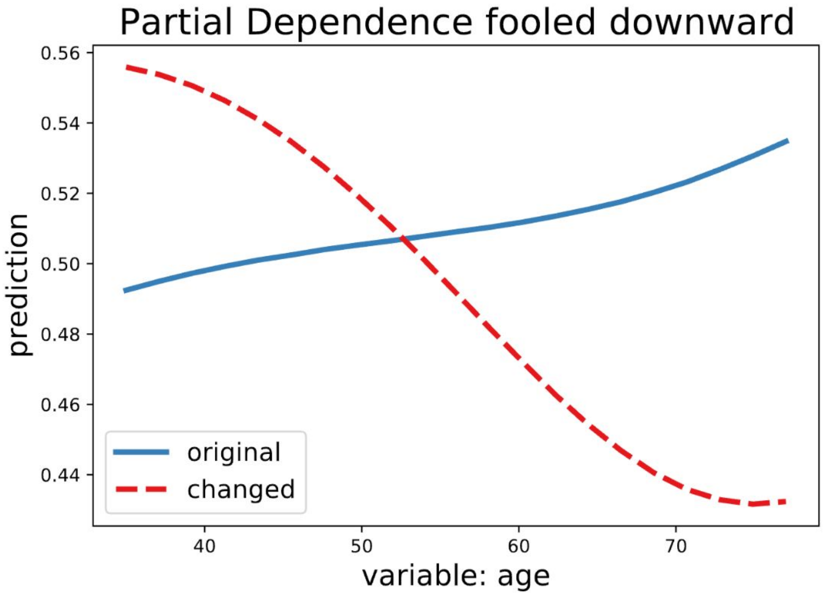
(Baniecki & Biecek, *Information Fusion* 2024)

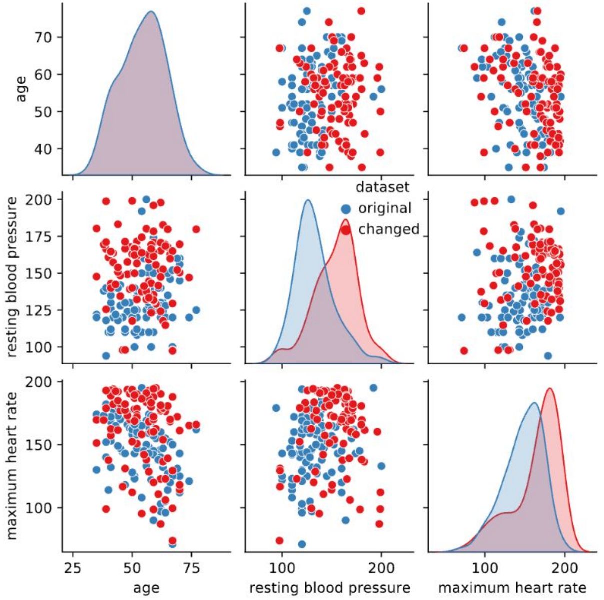


Attack on partial dependence $\in$ global feature effects

$$f_S(x_S) := \mathbb{E}_{q(\boldsymbol{x}_{\bar{S}})}\left[f(x_S, \boldsymbol{x}_{\bar{S}})\right] = \int f(x_S, x_{\bar{S}}) q(x_{\bar{S}}) dx_{\bar{S}}$$

(Baniecki et al., *ECML PKDD* 2022)





TL;DR:

1. Theoretical guarantees, **bounds** – how badly can it be manipulated?
2. Analyse **three** feature effect methods incl. accumulated local effects.
3. Robustness w.r.t marginal vs. **conditional** distribution.
4. What about **model** perturbation?
5. Empirical analysis of a **gradient-based** estimator of feature effects.

Theoretical analysis

$$\text{PD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) := \mathbb{E}_{\mathbf{X}_{\bar{s}} \sim p_{\mathbf{X}_{\bar{s}}}} [f(\mathbf{x}_s, \mathbf{X}_{\bar{s}})] := \int f(\mathbf{x}_s, \mathbf{x}_{\bar{s}}) p_{\mathbf{X}_{\bar{s}}}(\mathbf{x}_{\bar{s}}) d\mathbf{x}_{\bar{s}}$$

$$\text{CD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) := \mathbb{E}_{\mathbf{X}_{\bar{s}} \sim p_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}} [f(\mathbf{x}_s, \mathbf{X}_{\bar{s}})] := \int f(\mathbf{x}_s, \mathbf{x}_{\bar{s}}) p_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}(\mathbf{x}_{\bar{s}} | \mathbf{x}_s) d\mathbf{x}_s$$

Assumption 1. *We assume that the model f has bounded predictions, i.e., there exists a constant B such that $|f(\mathbf{x})| \leq B$ for all $\mathbf{x} \in \mathbb{R}^p$.*

Theorem 2. *The robustness of partial dependence and conditional dependence to* ***data perturbations*** *is given by the following formulas*

$$\left| \text{PD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \text{PD}_s(\mathbf{x}_s; f, p'_{\mathbf{X}}) \right| \leq 2B \cdot d_{\text{TV}}\left(p_{\mathbf{X}_{\bar{s}}}, p'_{\mathbf{X}_{\bar{s}}}\right),$$

$$\left| \text{CD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \text{CD}_s(\mathbf{x}_s; f, p'_{\mathbf{X}}) \right| \leq 2B \cdot d_{\text{TV}}\left(p_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}, p'_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}\right),$$

where the total variation distance d_{TV} is defined via the l_1 functional distance.

Perturb the dataset

Theorem 2. *The robustness of partial dependence and conditional dependence to* ***data perturbations*** *is given by the following formulas*

$$\left|\mathrm{PD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \mathrm{PD}_s(\mathbf{x}_s; f, p'_{\mathbf{X}})\right| \leq 2B \cdot d_{\mathrm{TV}}\left(p_{\mathbf{X}_{\bar{s}}}, p'_{\mathbf{X}_{\bar{s}}}\right),$$

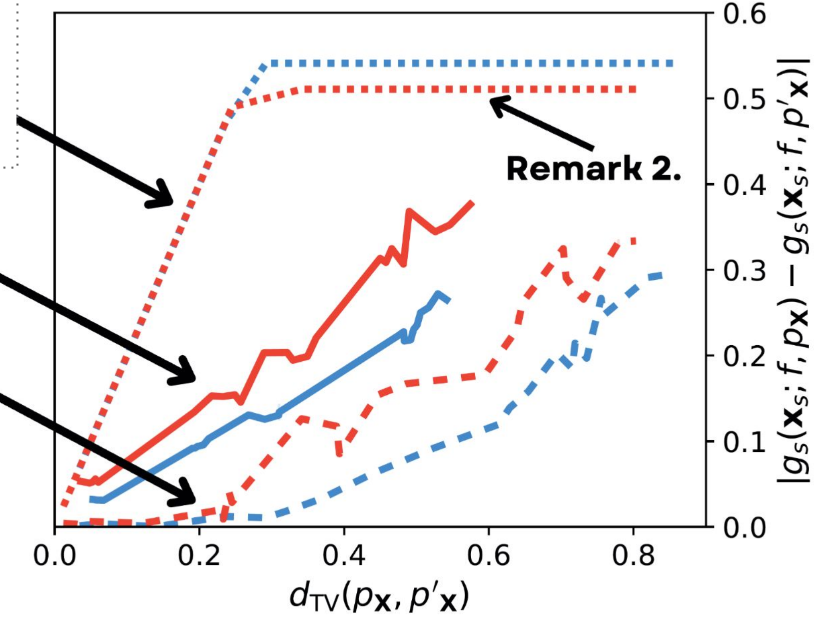
$$\left|\mathrm{CD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \mathrm{CD}_s(\mathbf{x}_s; f, p'_{\mathbf{X}})\right| \leq 2B \cdot d_{\mathrm{TV}}\left(p_{\mathbf{X}_{\bar{s}}|\mathbf{X}_s=\mathbf{x}_s}, p'_{\mathbf{X}_{\bar{s}}|\mathbf{X}_s=\mathbf{x}_s}\right),$$

where the total variation distance d_{TV} is defined via the l_1 functional distance.

Adversarial attack

(Baniecki et al., *ECML PKDD* 2022)

Gaussian noise (baseline)



More experiments in the paper.

Robustness to model perturbations

Lemma 2. *The robustness of partial dependence and conditional dependence to* ***model perturbations*** *is given by the following formulas*

$$\left|\mathrm{PD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \mathrm{PD}_s(\mathbf{x}_s; f', p_{\mathbf{X}})\right| \leq \|f - f'\|_\infty,$$

$$\left|\mathrm{CD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \mathrm{CD}_s(\mathbf{x}_s; f', p_{\mathbf{X}})\right| \leq \|f - f'\|_{\infty,\mathcal{X}},$$

where $\|f\|_\infty := \sup_{\mathbf{x}\in\mathbb{R}^p} |f(\mathbf{x})|$ *denotes an infinity norm for a function and* $\|f\|_{\infty,\mathcal{X}} := \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$ *is the same norm taken over the domain* $\mathcal{X} \subseteq \mathbb{R}^p$.

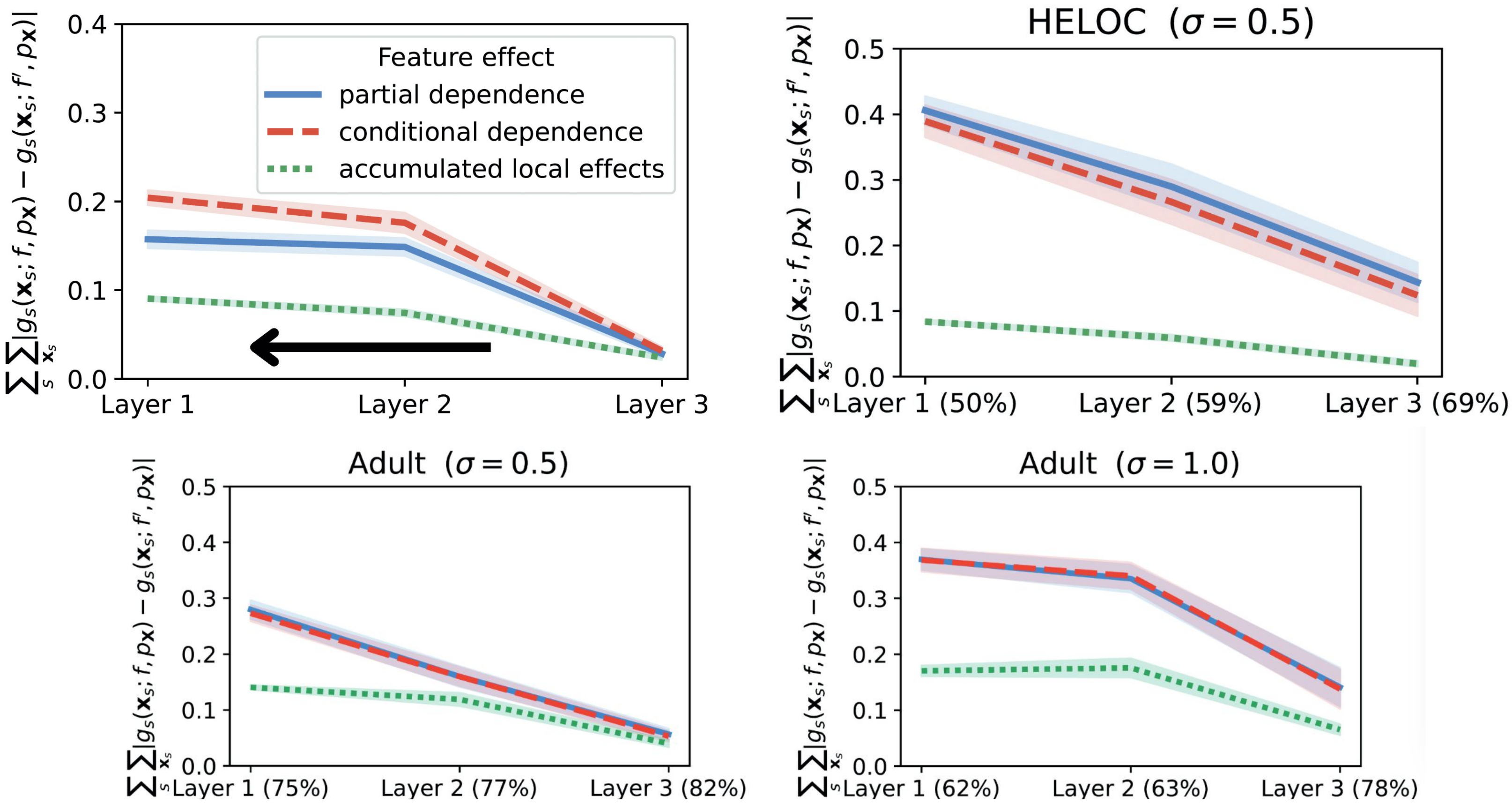
Proof. Follows directly from [26, lemmas 5 & 6]. **(Lin et al., *NeurIPS* 2023)**

Theorem 5. *The robustness of accumulated local effects to* ***model perturbations*** *is given by the following formula*

$$\left|\mathrm{ALE}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \mathrm{ALE}_s(\mathbf{x}_s; f', p_{\mathbf{X}})\right| \leq (\mathbf{x}_s - \mathbf{x}_{\min,s}) \cdot \|h - h'\|_{\infty,\mathcal{X}},$$

where $h := \frac{\partial f}{\partial \mathbf{x}_s}$ *and* $h' := \frac{\partial f'}{\partial \mathbf{x}_s}$ *denote partial derivatives of* f *and* f' *respectively.*

Perturb network weights



Takeaway & future work directions

1. (on average in our setting) Partial dependence **is more robust** to data perturbation than conditional dependence.
2. (Differential) ALE **does not pass** the model randomization test.
3. **Future:** tighten the bound, improve the attack.
4. **Feature dependence**, e.g. correlation, interactions.

On the Robustness of Global Feature Effect Explanations

Hubert Baniecki, Giuseppe Casalicchio, Bernd Bischl, Przemyslaw Biecek

UNIVERSITY OF WARSAW

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Munich Center for Machine Learning

Warsaw University of Technology

ECML PKDD 2024