

# Fooling Partial Dependence via Data Poisoning



**Hubert Baniecki, Wojciech Kretowicz, Przemysław Biecek**

MI<sup>2</sup>.AI, Warsaw University of Technology, Poland

**ECML PKDD, Grenoble, France**

September 21, 2022



# Outline

1. Adversary in explainable machine learning – Why should I care?
2. Fooling PD via Data Poisoning – What and how?
3. Experimental results
4. Discussion & future work

🕒 This article was published more than 3 years ago

BUSINESS

2019

# Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has a higher credit score



By Taylor Telford

November 11, 2019 at 10:44 a.m. EST

*Many articles have been published in 2020 describing new machine learning-based models for [detection and prognostication of COVID-19], but it is unclear which are of potential clinical utility. [...] Our review finds that **none of the models identified are of potential clinical use** due to methodological flaws and/or underlying biases.*



## Explainable machine learning: from credit scoring to precision diagnostics in bio-medicine

### nature machine intelligence

Explore content ▾

About the journal ▾

Publish with us ▾

2021

[nature](#) > [nature machine intelligence](#) > [analyses](#) > article

Analysis | [Open Access](#) | [Published: 15 March 2021](#)

## Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

[Michael Roberts](#) , [Derek Driggs](#), [Matthew Thorpe](#), [Julian Gilbey](#), [Michael Yeung](#), [Stephan Ursprung](#), [Angelica I. Aviles-Rivero](#), [Christian Etmann](#), [Cathal McCague](#), [Lucian Beer](#), [Jonathan R. Weir-McCall](#), [Zhongzhao Teng](#), [Effrossyni Gkrania-Klotsas](#), [AIX-COVNET](#), [James H. F. Rudd](#), [Evis Sala](#) & [Carola-Bibiane Schönlieb](#)

[Nature Machine Intelligence](#) **3**, 199–217 (2021) | [Cite this article](#)

**74k** Accesses | **237** Citations | **1159** Altmetric | [Metrics](#)

# We **ex**plain black-box machine learning for...

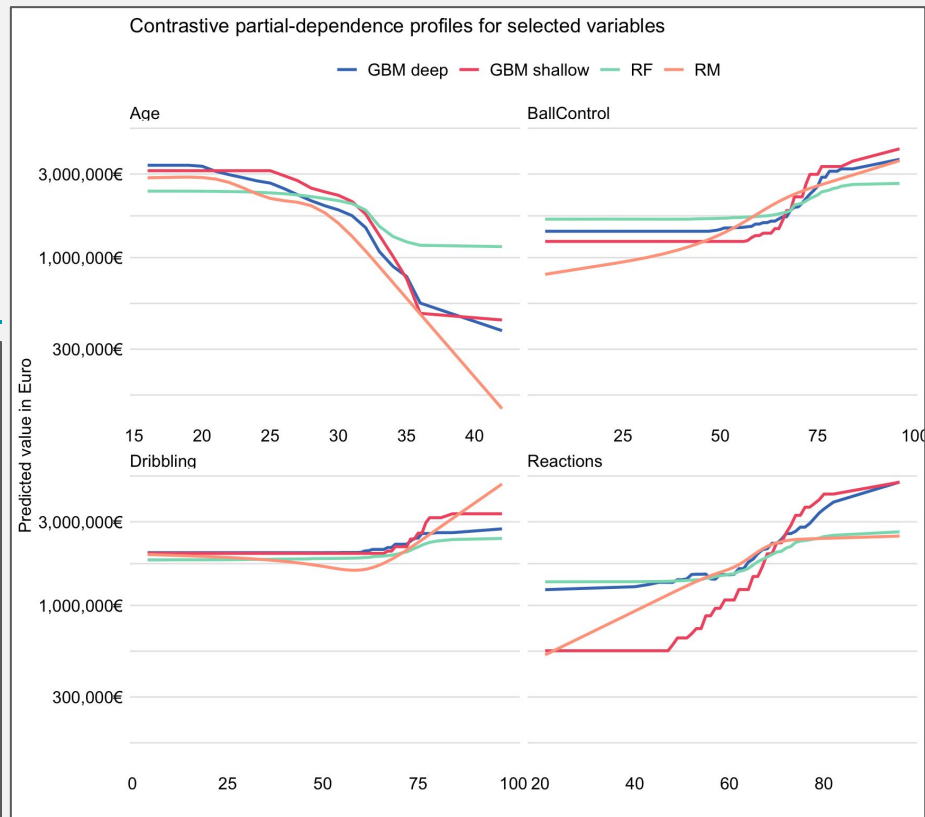
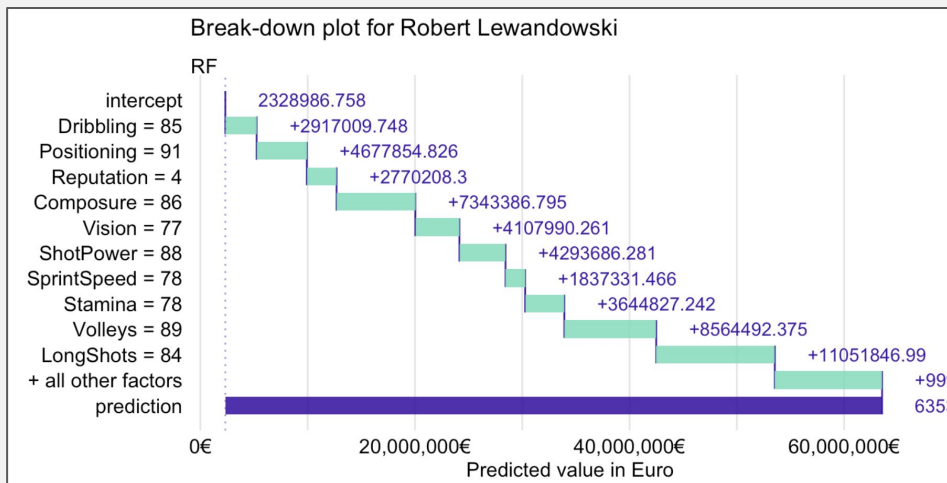
W. Samek (Monday, 4th XKDD Workshop @ECML PKDD 2022)

(1) Validation & debugging

(2) Scientific insights

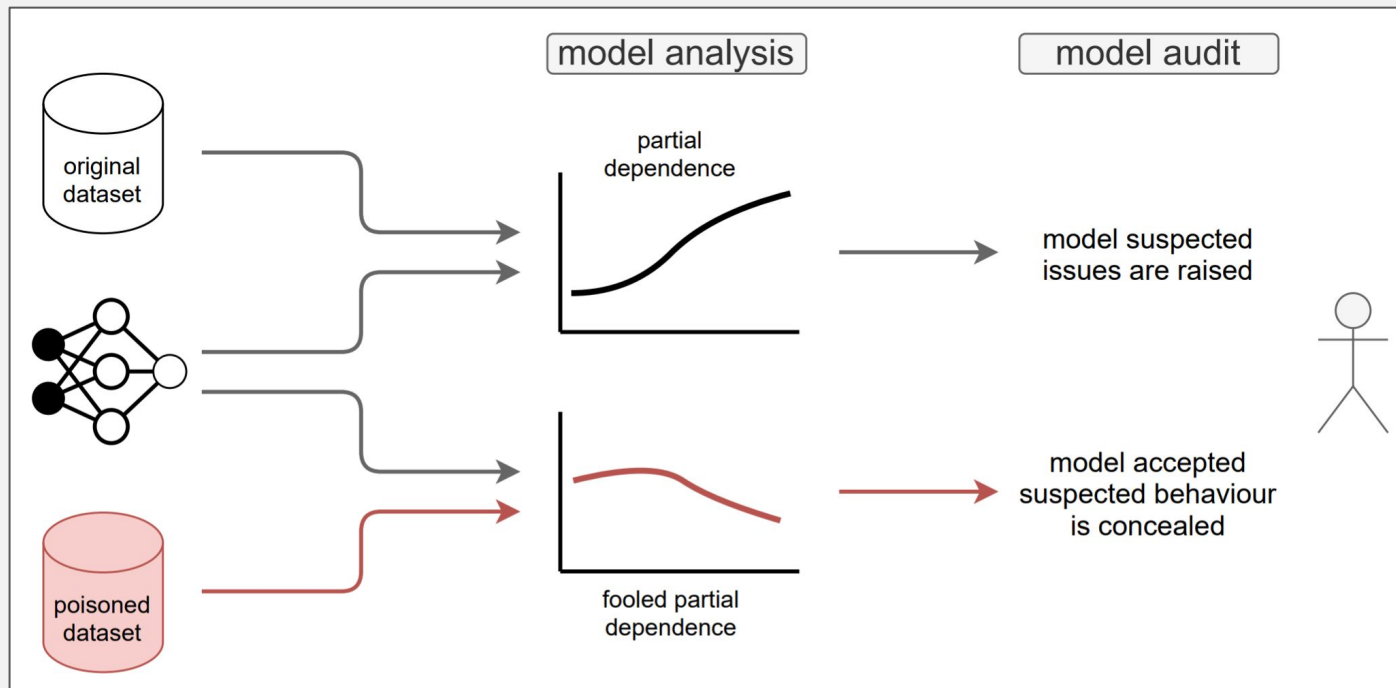
(3) Model improvement

[ema.drwhy.ai](http://ema.drwhy.ai)

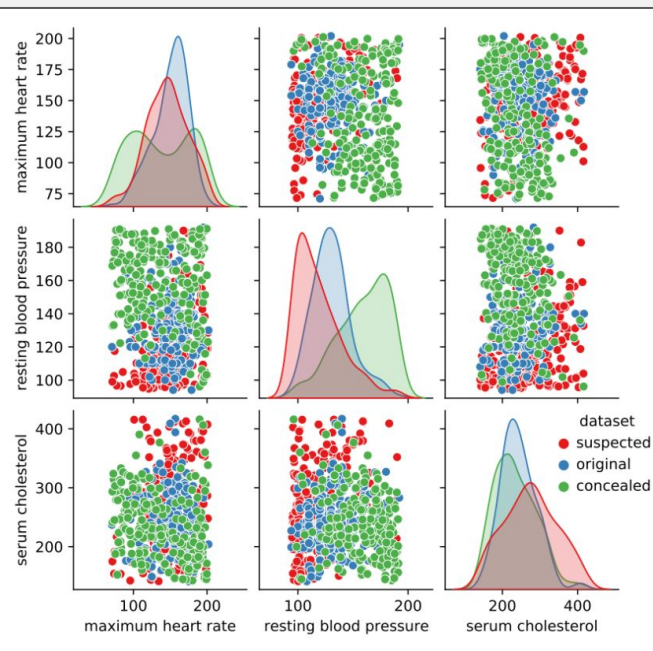
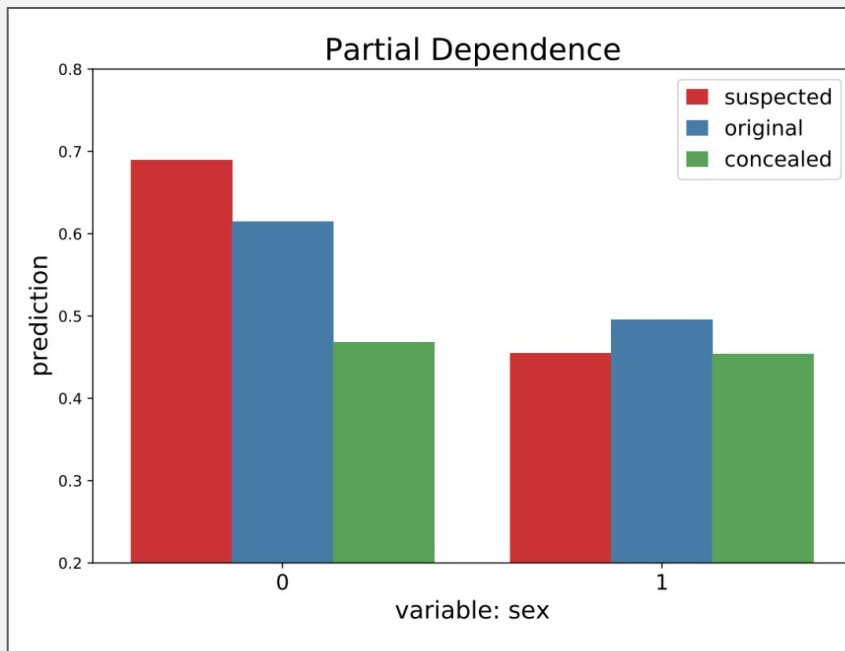


# Attacking model-agnostic explanations, e.g. PDP

Why?

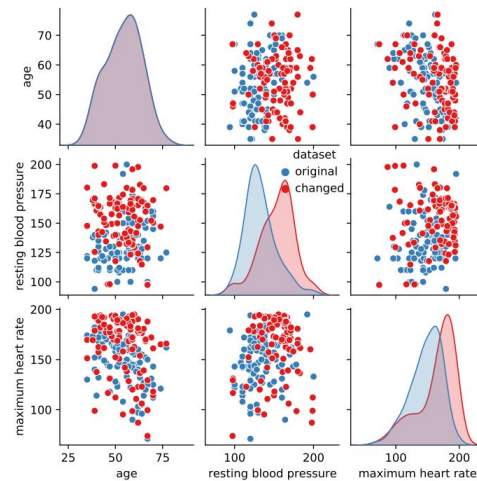
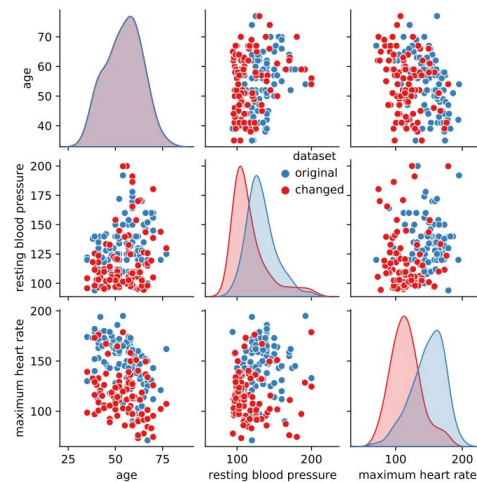
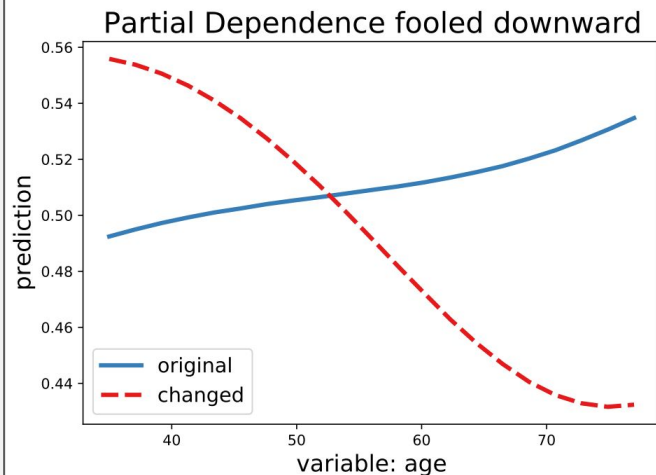
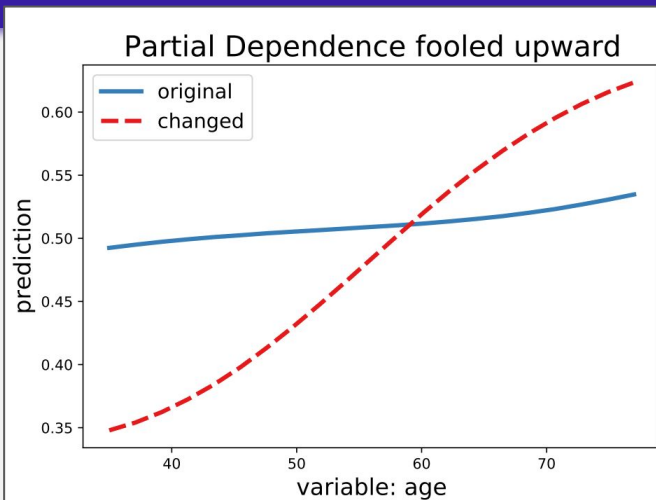


# Motivational example (Heart disease classification)



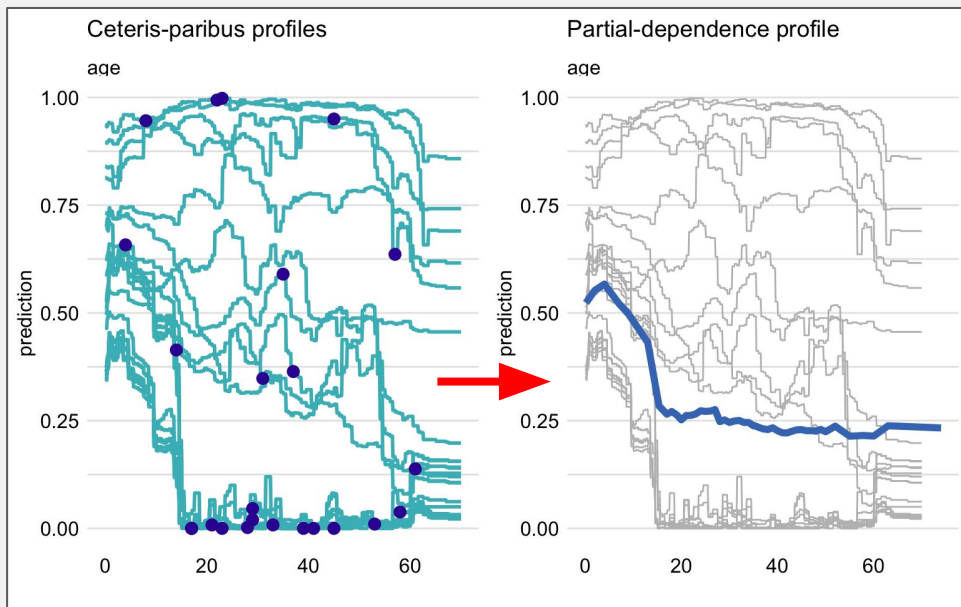
We explain the same model - it is unchanged!

*Faking  
explanations to  
confirm a  
hypothesis?*

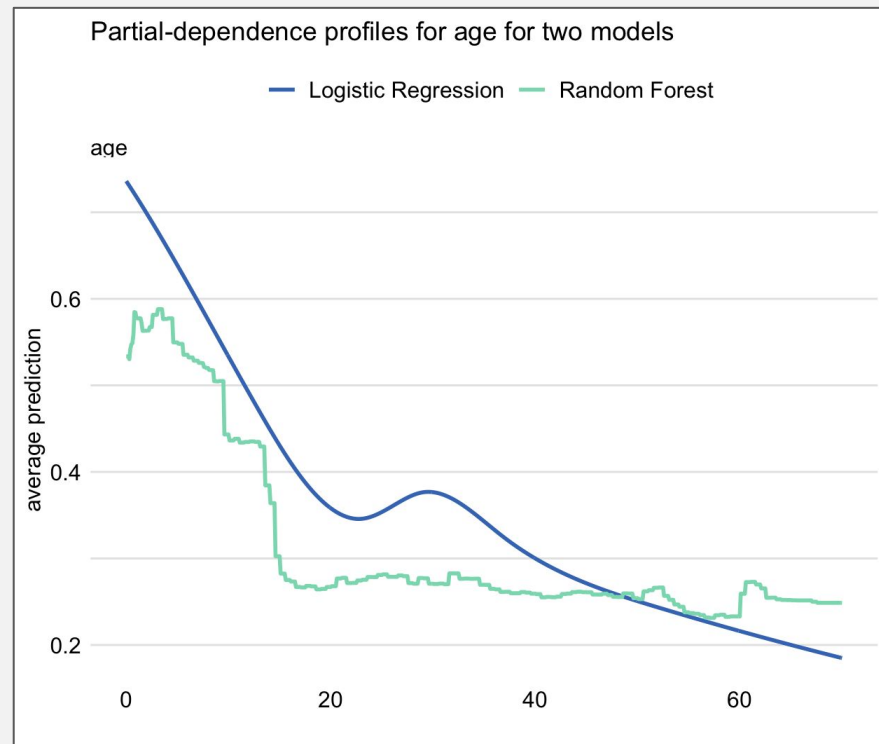


*Can we trust  
explanations?*

# Partial Dependence (plot, profile, PDP)



[ema.drwhy.ai](http://ema.drwhy.ai)





# Data Poisoning (1/2)

We want to **optimize the distance** between the original and changed explanation by **iteratively changing** the dataset. We propose:

- (1) A flexible model-agnostic **genetic-based algorithm**, which does not assume any structure about the model or explanation. (Wright, 1991)
- (2) An efficient **gradient-based algorithm**, which is specific to differentiable models, e.g. neural networks. **Analytical derivation + automatic differentiation.**

## Data Poisoning (2/2)

We want to **optimize the distance** between the original and changed explanation by **iteratively changing** the dataset. Possible strategies:

(1) **Targeted attack** changes the dataset to achieve the closest explanation result to the predefined desired function **T**.

(2) **Robustness check** aims for the most distant model explanation from the original one.

# Experiments (1/2)

## Fooling Partial Dependence via Data Poisoning

9

**Table 1.** Attack loss values of the robustness checks for Partial Dependence of various machine learning models (**top**), and complexity levels of tree-ensembles (**bottom**). Each value corresponds to the scaled distance between the original explanation and the changed one. We perform the fooling 6 times and report the mean  $\pm$  sd. We observe that the explanations' vulnerability increases with GBM complexity.

Task \ Model	LM	RF	GBM	DT	KNN	NN	SVM
<b>friedman</b>	0 $\pm$ 0	152 $\pm$ 76	127 $\pm$ 71	332 $\pm$ 172	164 $\pm$ 61	269 $\pm$ 189	576 $\pm$ 580
<b>heart</b>	2 $\pm$ 3	20 $\pm$ 5	77 $\pm$ 28	798 $\pm$ 192	133 $\pm$ 21	501 $\pm$ 52	451 $\pm$ 25

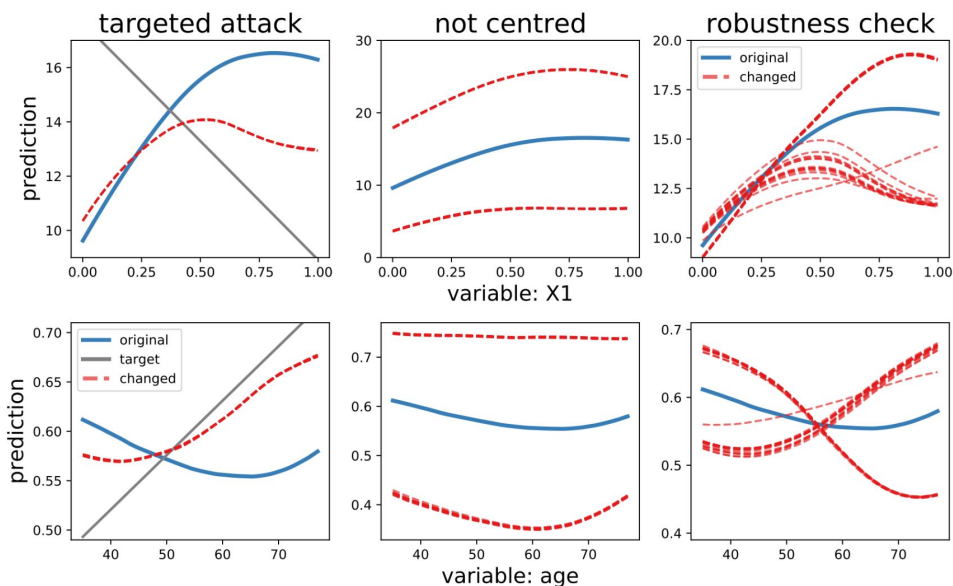
  

Task \ Model	Trees		10	20	40	80	160	320
<b>friedman</b>	GBM		57 $\pm$ 12	114 $\pm$ 20	157 $\pm$ 37	176 $\pm$ 20	189 $\pm$ 8	210 $\pm$ 9
	RF		233 $\pm$ 22	219 $\pm$ 25	219 $\pm$ 9	201 $\pm$ 23	216 $\pm$ 13	209 $\pm$ 15
<b>heart</b>	GBM		1 $\pm$ 0	3 $\pm$ 1	29 $\pm$ 4	70 $\pm$ 24	152 $\pm$ 56	321 $\pm$ 95
	RF		62 $\pm$ 7	55 $\pm$ 3	29 $\pm$ 9	21 $\pm$ 6	14 $\pm$ 5	13 $\pm$ 2

## Experiments (2/2)

10

H. Baniecki et al.



**Fig. 3.** Fooling Partial Dependence of a  $3 \times 32$  neural network fitted to the **friedman** (top row) and **heart** (bottom row) datasets. We performed multiple randomly initiated gradient-based fooling algorithms on the explanations of variables  $X_1$  and **age**

# Main contributions

- (1) We highlight that Partial Dependence can be **manipulated** with adversarial data perturbations.
- (2) We introduce a novel concept of using a **genetic algorithm** for attacking explanations of **any black-box**. We propose a gradient algorithm for neural networks.
- (3) Experiments on various models and their sizes shows the **hidden debt of model complexity** related to explainable machine learning.

## Remarks

- **Claim:** Partial Dependence can be fooled, but not necessarily always
- **Assumption:** an auditor has no access to the original (unknown) data
- **Takeaway:** interpret model explanation in the context of data distribution

## Future? work

- Sanity checks for other explanations:
  - PDP -> **Accumulated Local Effects** (Apley & Zhu, J. R. Stat. Soc. 2020)
  - SHAP (Lundberg & Lee, NeurIPS 2017) -> (Baniecki & Biecek, AAAI 2022)
- Remove the **assumption** and analyze the **detectability** of data poisoning by measuring the distance between data distributions.

# Details? Algorithms, benchmarks, and related work!

Paper ID **176**

Poster:

**When?** Tomorrow, 18:30 (Thursday)

**Where?** A1, Robust & Adv. ML (1)

**Paper:** arXiv:2105.12837

**Contact:** hbaniecki.com



Call for  
postdocs ;-)  
[www.mi2.ai](http://www.mi2.ai)

This work was financially supported by the NCN OPUS grant no. 2017/27/B/ST6/0130  
and SONATA BIS grant no. 2019/34/E/ST6/00052.