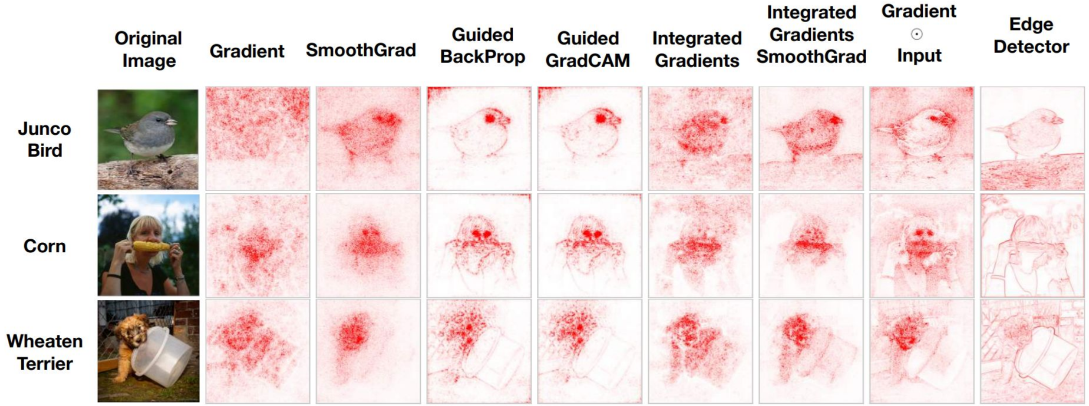
Evaluating explanations

Hubert Baniecki

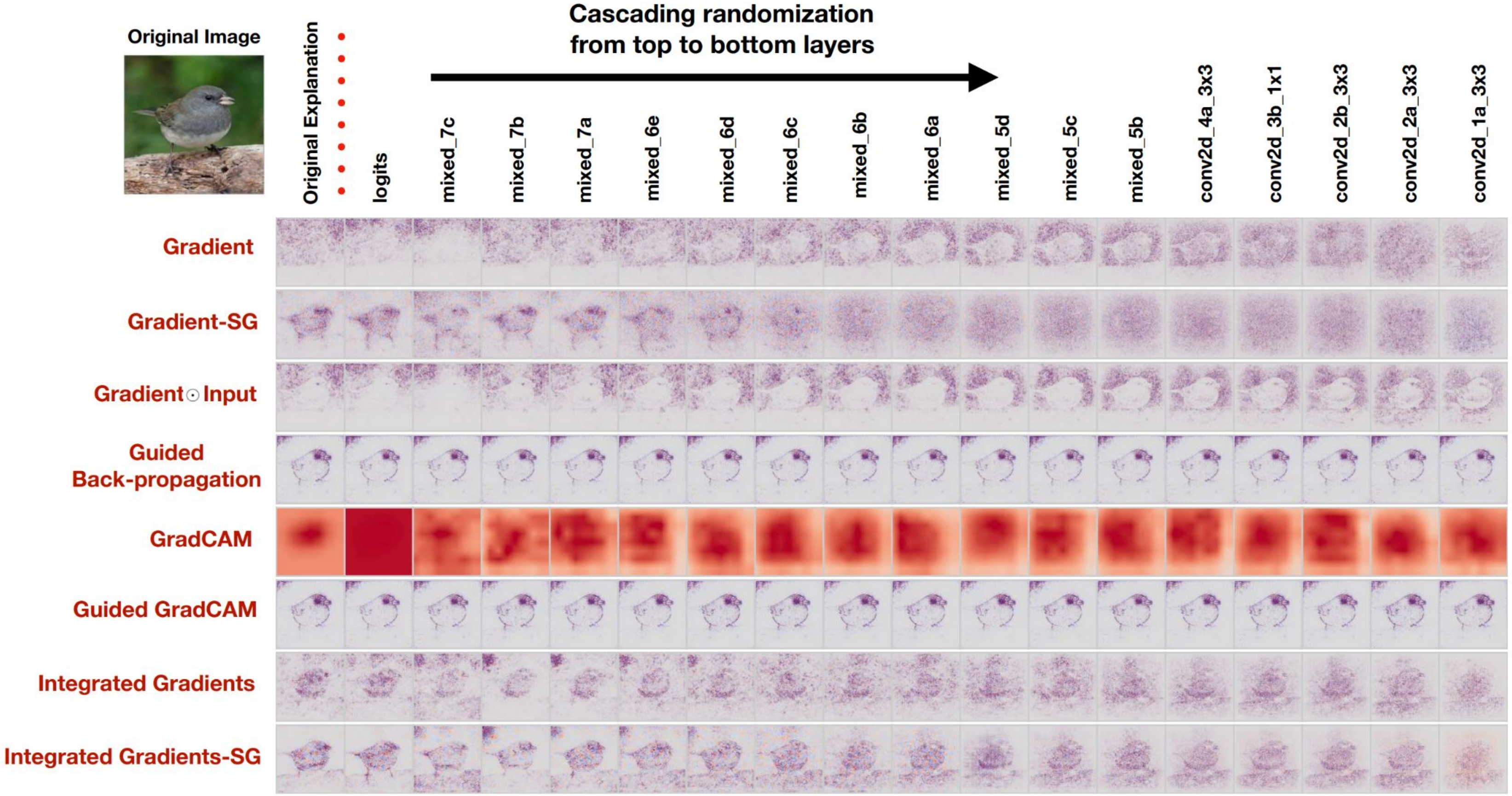
January 10th, 2022

Motivation testing explanations

“Through extensive experiments we show that some existing saliency methods are independent both of the **model** and of the **data** generating process.”

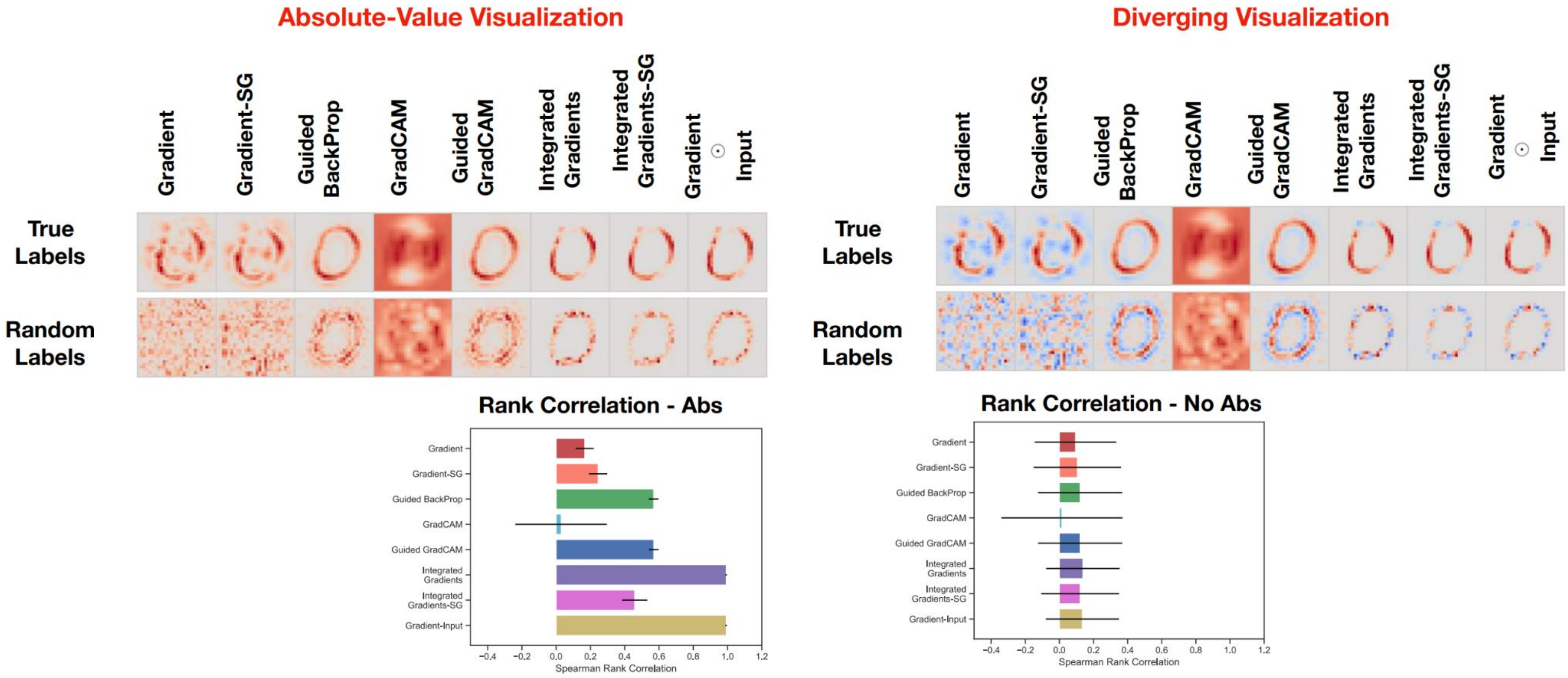


Adebayo et al. Sanity Checks for Saliency Maps (NeurIPS 2018)



Adebayo et al. Sanity Checks for Saliency Maps (NeurIPS 2018)

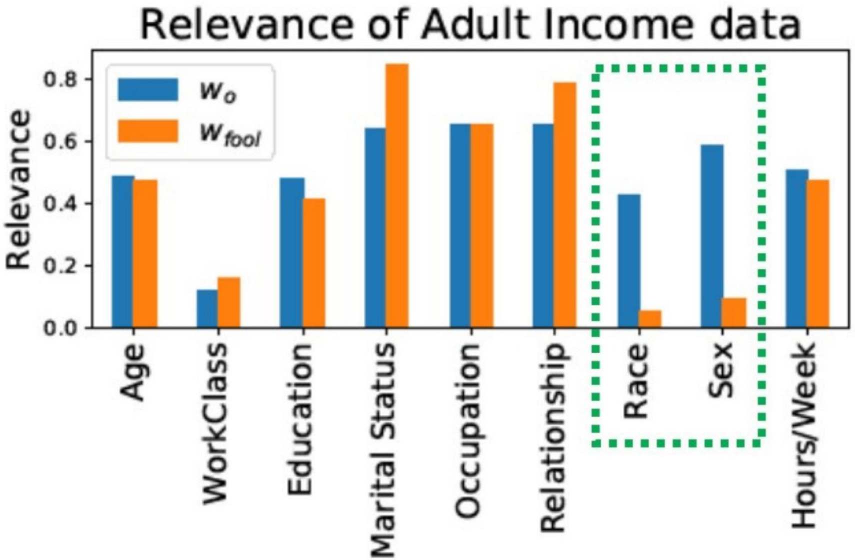
CNN - MNIST



Adebayo et al. Sanity Checks for Saliency Maps (NeurIPS 2018)

Motivation manipulating explanations

Dombrowski et al. Explanations can be manipulated and geometry is to blame (NeurIPS 2019)



Heo et al. Fooling Neural Network Interpretations via Adversarial Model Manipulation (NeurIPS 2019)

Agenda discuss three ways of evaluating explanations

1. **User** study: Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)
2. **Theoretical** benchmark: Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)
3. **Practical** experiment: Zhou et al. Do Feature Attribution Methods Correctly Attribute Features? (AAAI 2022)

User study

Poursabzi-Sangdeh et al.
*Microsoft Research
Manipulating and Measuring Model Interpretability
(CHI 2021)

Idea. Compare usefulness of four models in the experiments:
white-box and black-box with 2 or 8 features used.

(1) **How well can people simulate a model's predictions?**
(2) **To what extent do people follow a model's predictions when it is beneficial for them to do so?**
(3) **How well can people detect when a model has made a mistake and correct for it?**

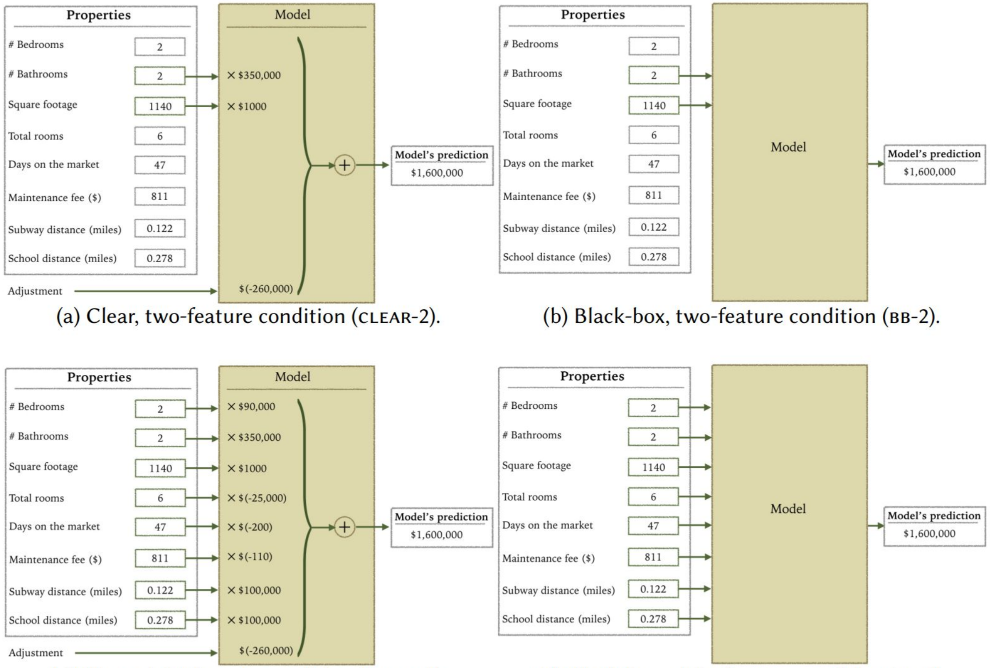
Methods. Four sequential user-studies on Amazon Technical Turk.

EXPERIMENT 1: PREDICTING APARTMENT SELLING PRICES (N=1250)

EXPERIMENT 4: OUTLIER FOCUS AND DETECTION OF MISTAKES (N=800)

Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

EXPERIMENT 1: PREDICTING APARTMENT SELLING PRICES (N=1250)



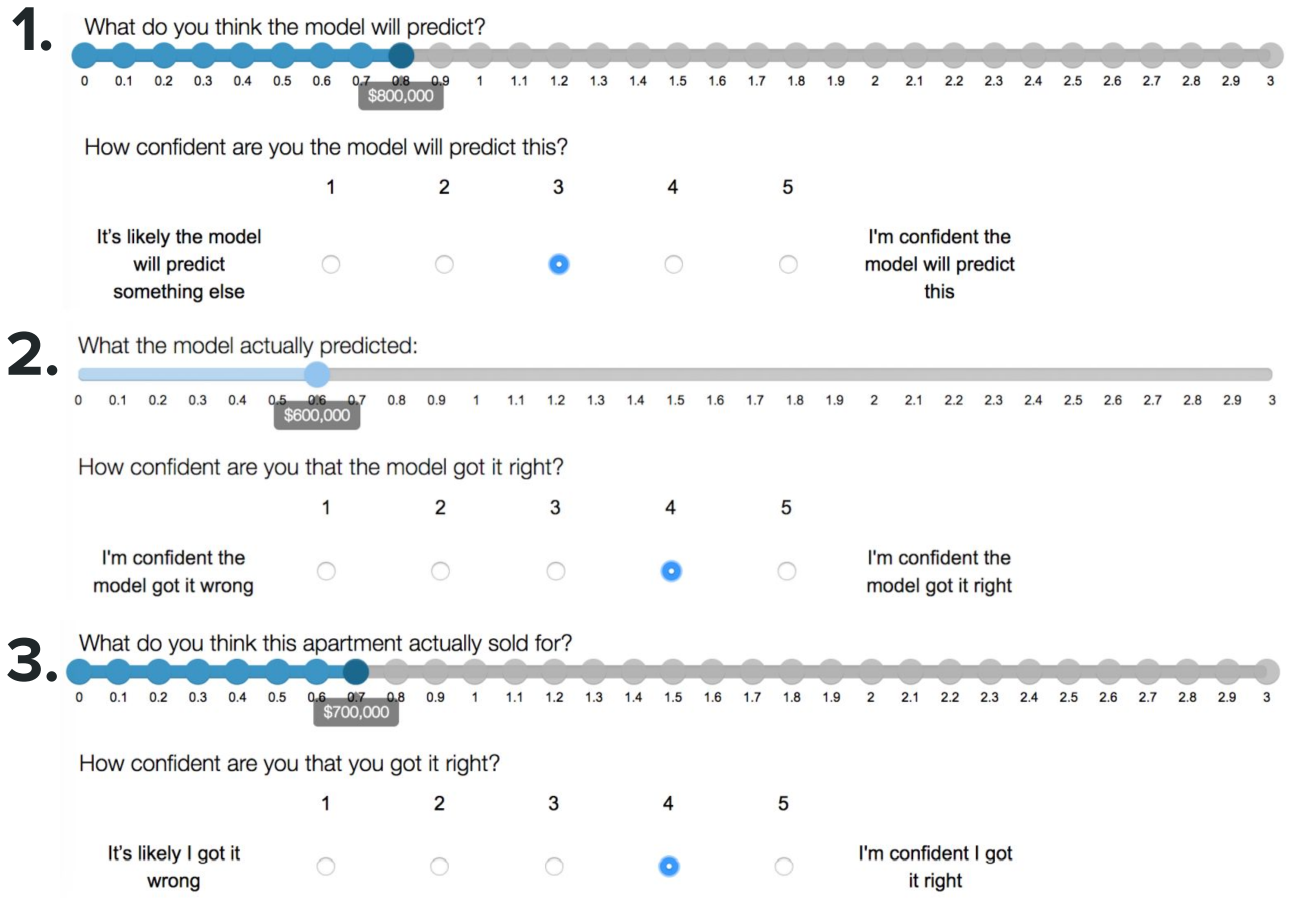
(a) Clear, two-feature condition (CLEAR-2).

(b) Black-box, two-feature condition (BB-2).

(c) Clear, eight-feature condition (CLEAR-8).

(d) Black-box, eight-feature condition (BB-8).

Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)



Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

Hypotheses and measures

1. **Simulation:** users will better simulate predictions of a simple model

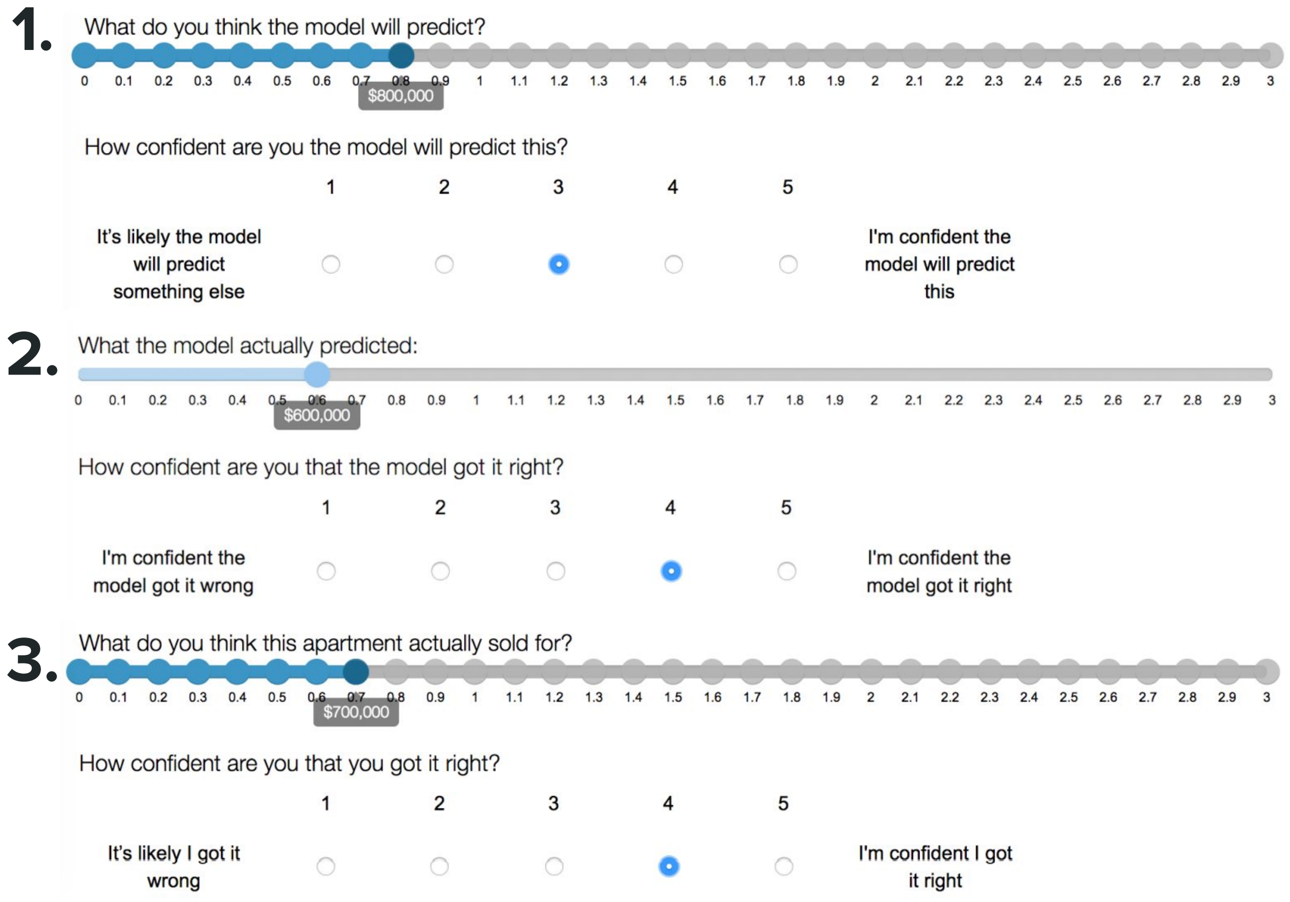
   error: model prediction - user prediction of the prediction

2. **Deviation:** users will follow the predictions of a simple model

   error: model prediction - user prediction of the price

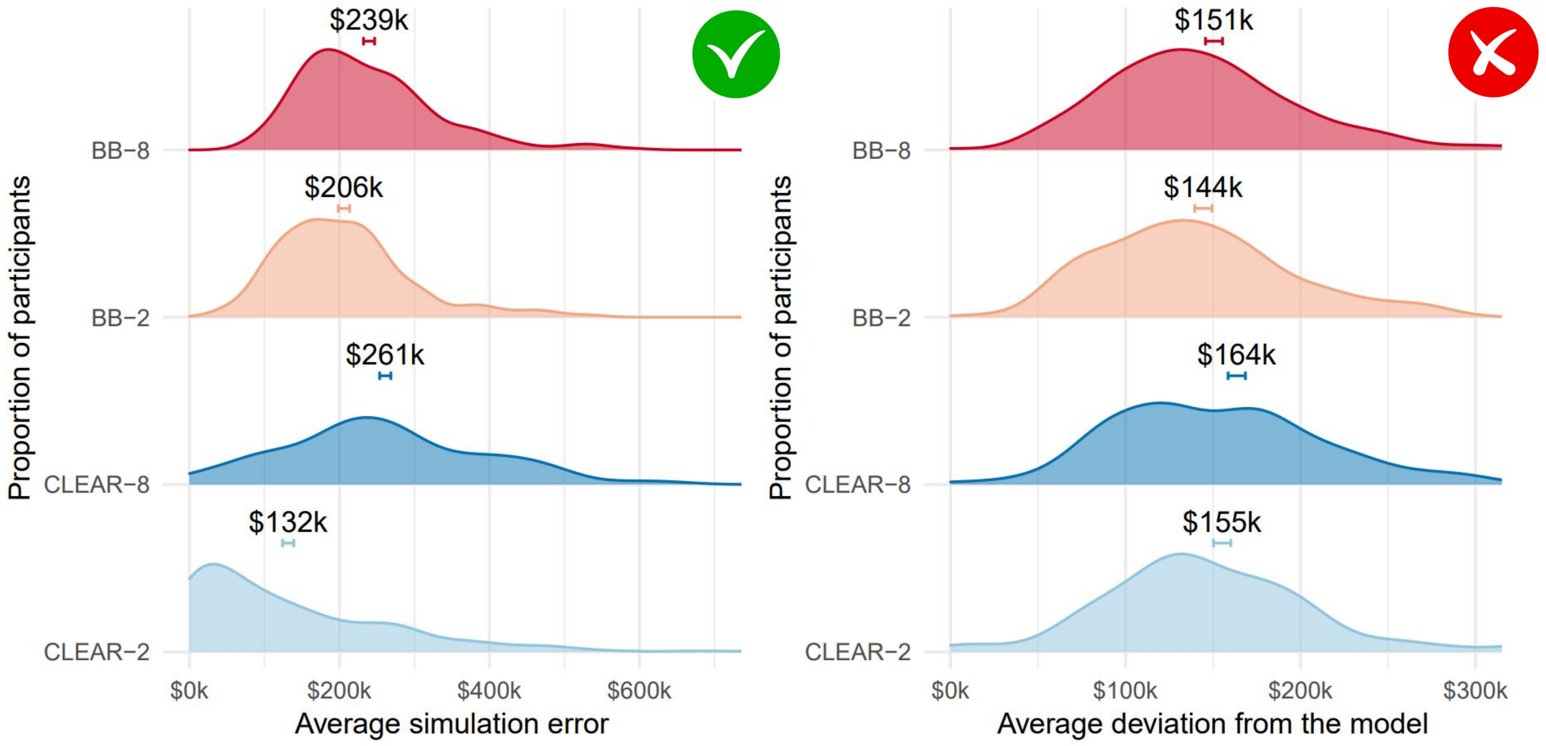
3. **Detection of mistakes:** users will differently find the models' mistakes

   same as above but only for 2 apartments

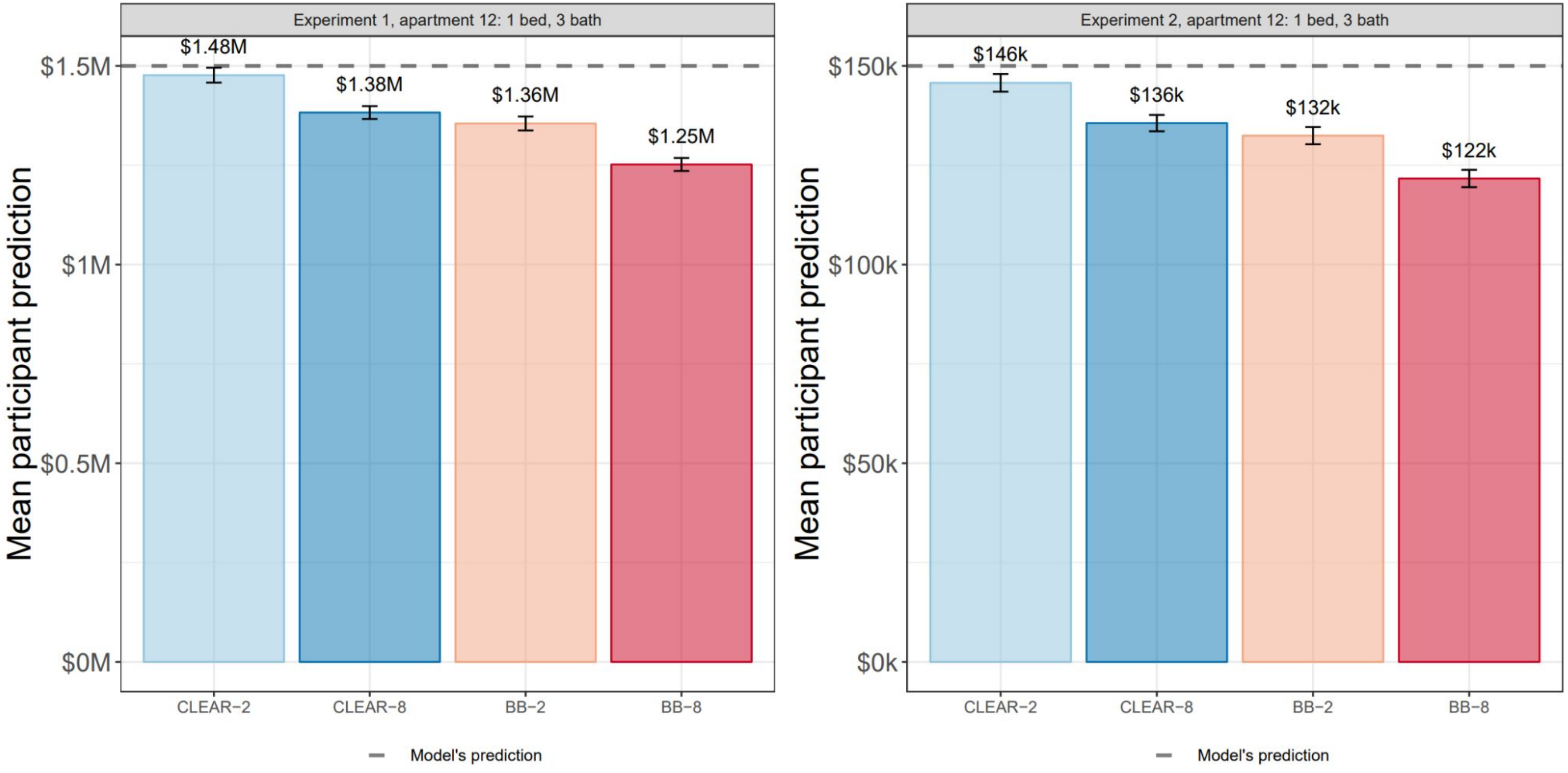


Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

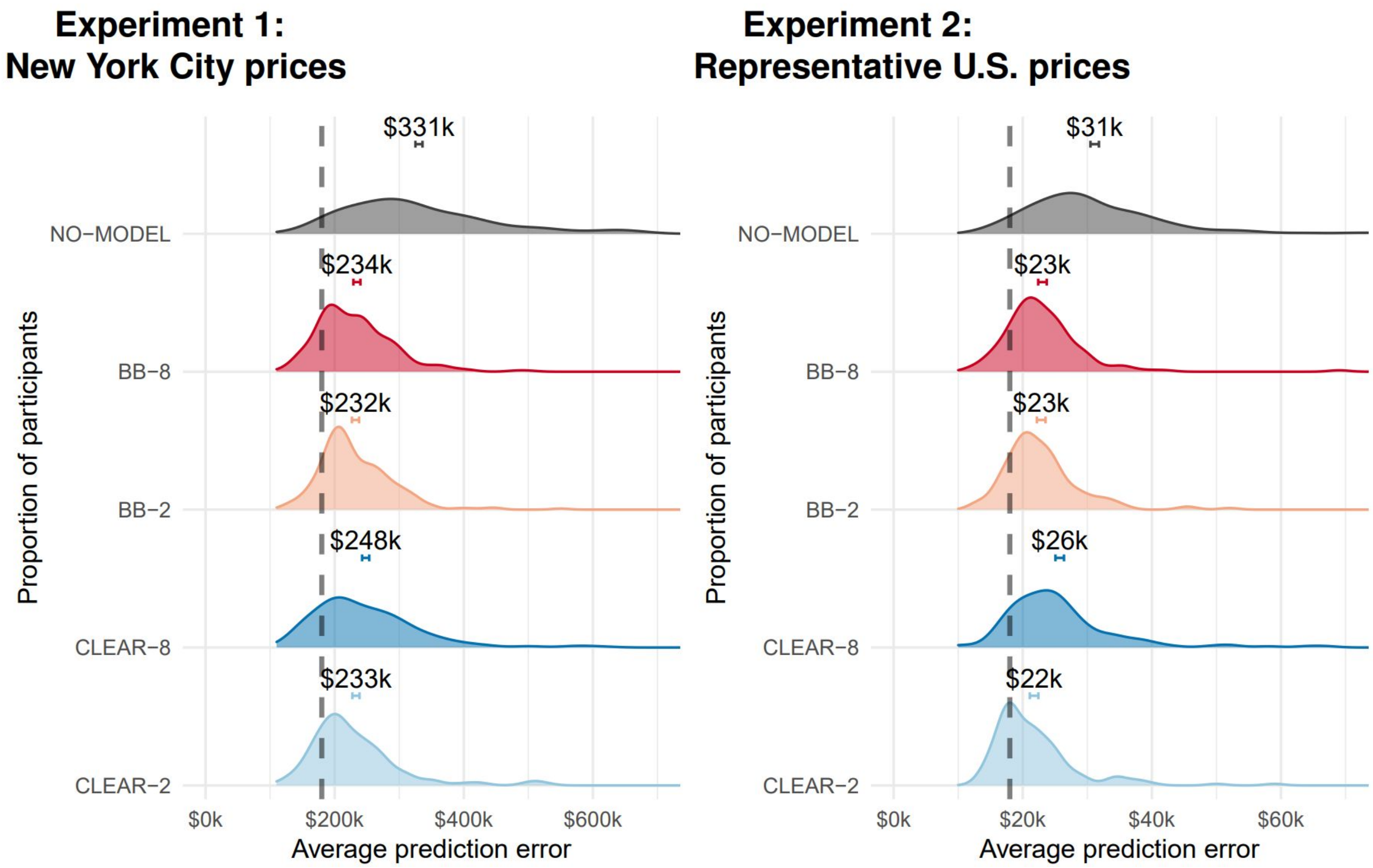
EXPERIMENT 1: PREDICTING APARTMENT SELLING PRICES (N=1250)



Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

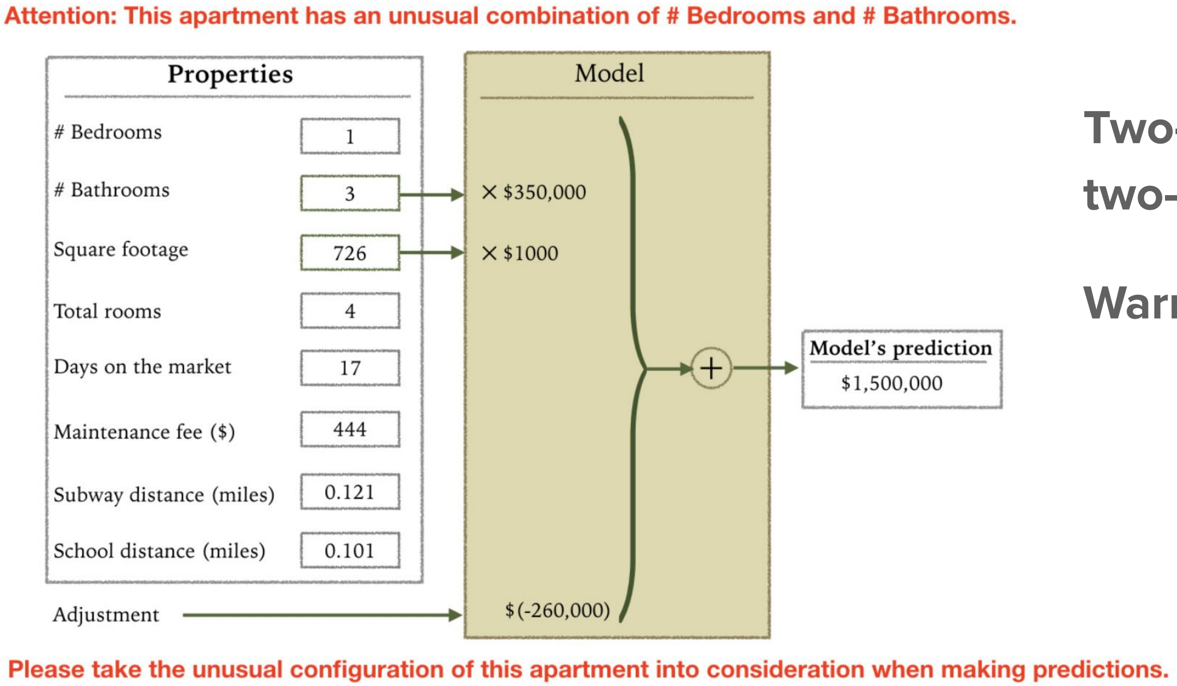


Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)



Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

EXPERIMENT 4: OUTLIER FOCUS AND DETECTION OF MISTAKES (N=800)

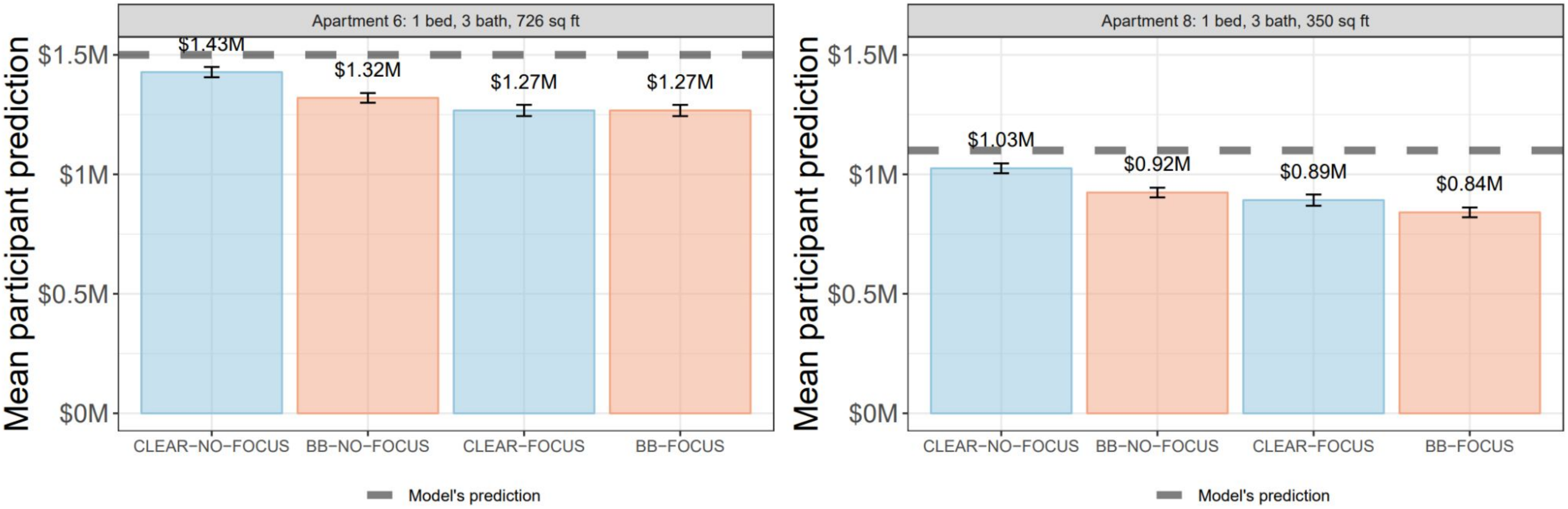


Two-feature white-box vs two-feature black-box

Warning vs no warning

Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

EXPERIMENT 4: OUTLIER FOCUS AND DETECTION OF MISTAKES (N=800)



Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

Hypotheses and outcomes

1. **Outlier focus.** Showing the warning improves performance.
2. **Transparency (clear vs. black box) and no outlier focus.** Showing the white-box improves performance.
3. **Transparency (clear vs. black box) and outlier focus.** No differences in white-box and black-box.

Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)

1. No significant differences (in the users' performance) between a simple white-box and complex black-box
2. **Information overload**
3. Set goals with respect to interpretability: use testing, not intuition

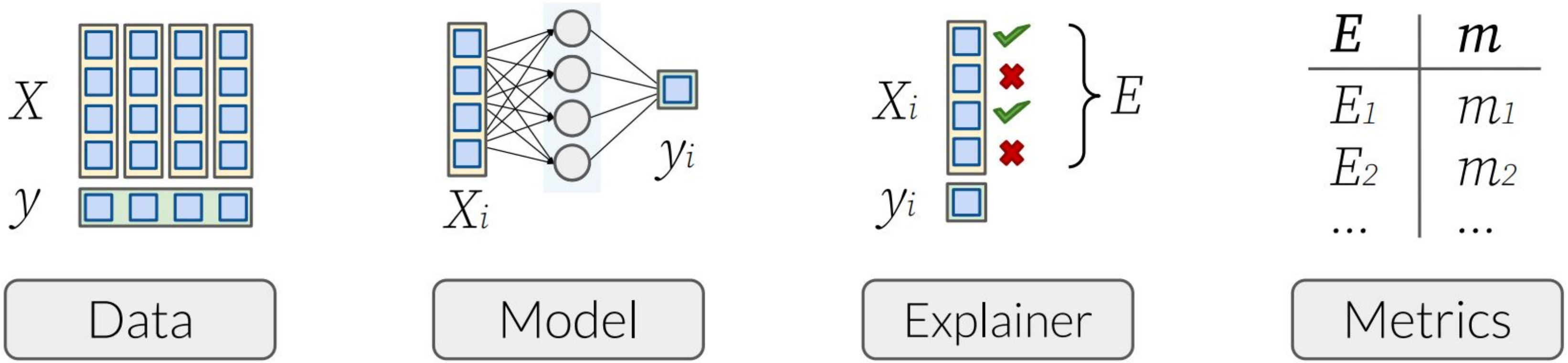
Theoretical benchmark

Liu et al.

*Abacus.AI

Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)

XAI-Bench https://github.com/abacusai/xai-bench



Examples: real data, synthetic data families

Examples: multilayer perceptron, decision tree, linear regression

Examples: SHAP, SHAPR, BF-SHAP, MAPLE, LIME, L2X, breakDown, Random

Examples: GT-shapley, ROAR faithfulness, monotonicity

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)

Data

1. **Generate features X.** They have known distributions, e.g. multivariate Gaussian, mixture of Gaussians, and multinomial. Moreover, we known the conditional distributions.

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

2. **Generate labels Y from X.** For example: piecewise linear, additive targets, nonlinear cosine, exponent. We can scale regression tasks to have the target of mean 0/sd 1.

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)

Metrics

$$\text{faithfulness} = \text{Pearson}\left(\left|\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{D}(\boldsymbol{x}_{F\setminus i})}[f(\boldsymbol{x}')] - f(\boldsymbol{x})\right|_{1\le i\le D}, [w_i]_{1\le i\le D}\right)$$

Comparing which feature would have the most impact on the model output when individually changed.

$$\text{monotonicity} = \frac{1}{D-1}\sum_{i=0}^{D-2}\mathbb{I}_{|\delta_i^+|\le|\delta_{i+1}^+|},$$

$$\text{where } \delta_i^+ = \mathbb{E}_{\boldsymbol{x}'\sim\mathcal{D}(\boldsymbol{x}_{S+(\boldsymbol{w},i+1)})}[f(\boldsymbol{x}')] - \mathbb{E}_{\boldsymbol{x}'\sim\mathcal{D}(\boldsymbol{x}_{S+(\boldsymbol{w},i)})}[f(\boldsymbol{x}')]$$

Fraction of indices *i* such that the marginal improvement for feature *i* is greater than the marginal improvement for feature *i+1*.

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)

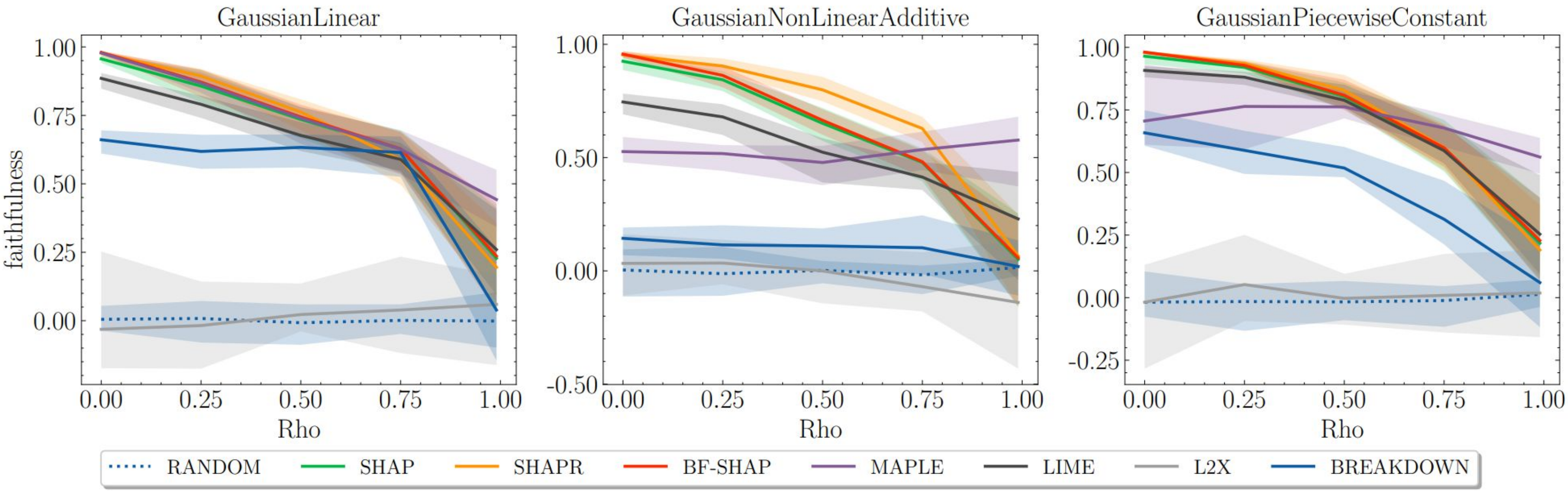
Metrics

1. **Remove-and-retrain (ROAR).** Hooker et al. A Benchmark for Interpretability Methods in Deep Neural Networks (NeurIPS 2019)
   — The model is retrained using a new dataset with the features removed.

2. **GT-Shapley.**
   — Pearson between an explanation and ground-truth Shapley values

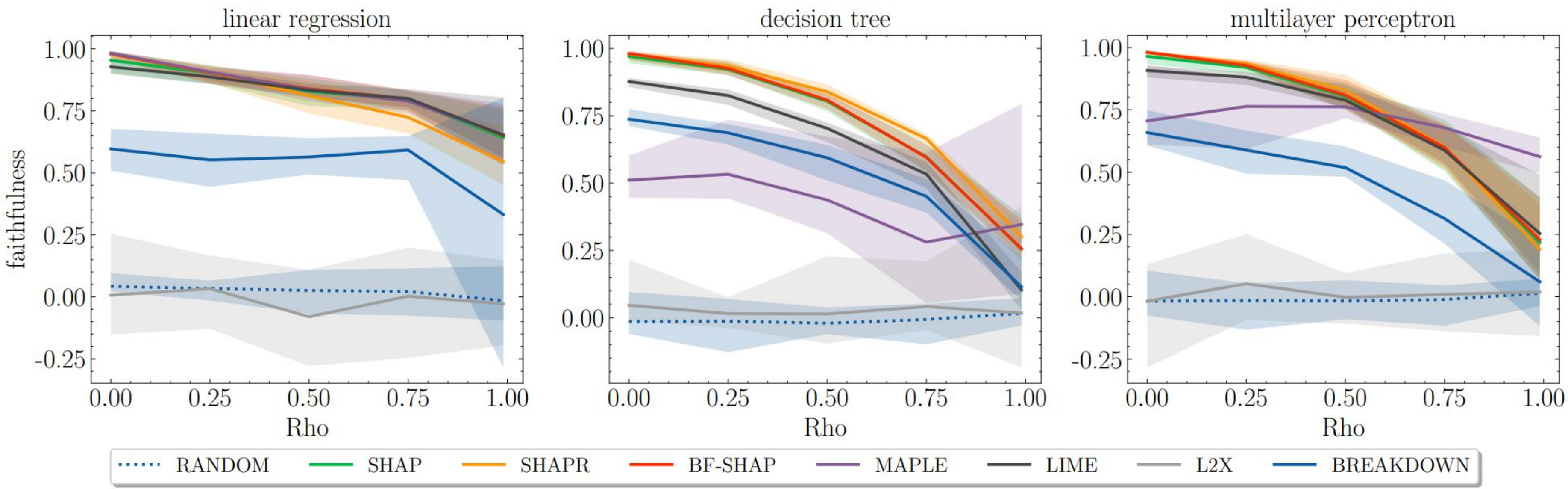
3. **Infidelity.** Yeh et al. On the (in) fidelity and sensitivity of explanations (NeurIPS 2019)
   — Fidelity but with noisy baseline conditional expectation

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)



Rho: scale of feature dependence

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)



Rho: scale of feature dependence

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)

We can approximate the known datasets!

Measure similarities between data distribution with Jensen-Shannon Divergence

Table 2: Explainer performance on the simulated wine dataset across metrics. All performance numbers are from explainers for a decision tree.

	RANDOM	SHAP	SHAPR	LIME	MAPLE	L2X	BREAKDOWN
faithfulness (↑)	$-0.007_{\pm 0.005}$	$\mathbf{0.534}_{\pm 0.045}$	$0.528_{\pm 0.032}$	$0.368_{\pm 0.031}$	$0.034_{\pm 0.033}$	$-0.030_{\pm 0.018}$	$-0.042_{\pm 0.011}$
monotonicity (↑)	$0.529_{\pm 0.008}$	$0.549_{\pm 0.009}$	$\mathbf{0.551}_{\pm 0.009}$	$0.547_{\pm 0.007}$	$0.520_{\pm 0.014}$	$0.522_{\pm 0.005}$	$0.493_{\pm 0.014}$
ROAR (↑)	$0.698_{\pm 0.031}$	$0.780_{\pm 0.016}$	$0.549_{\pm 0.031}$	$0.738_{\pm 0.026}$	$\mathbf{0.818}_{\pm 0.022}$	$0.664_{\pm 0.02}$	$0.625_{\pm 0.002}$
GT-Shapley (↑)	$0.004_{\pm 0.013}$	$0.825_{\pm 0.006}$	$\mathbf{0.945}_{\pm 0.002}$	$0.745_{\pm 0.015}$	$0.685_{\pm 0.008}$	$-0.108_{\pm 0.029}$	$-0.064_{\pm 0.02}$
infidelity (↓)	$0.353_{\pm 0.174}$	$0.234_{\pm 0.124}$	$\mathbf{0.212}_{\pm 0.146}$	$0.234_{\pm 0.126}$	$0.234_{\pm 0.132}$	$0.285_{\pm 0.115}$	$0.365_{\pm 0.133}$

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)

Table 3: Explainer performance on the simulated forest fires dataset across metrics. All performance numbers are from explainers for a decision tree.

	RANDOM	SHAP	LIME	MAPLE	L2X	BREAKDOWN
faithfulness (↑)	$0.022_{\pm 0.034}$	$\mathbf{0.571}_{\pm 0.023}$	$0.449_{\pm 0.007}$	$0.080_{\pm 0.056}$	$0.001_{\pm 0.008}$	$0.158_{\pm 0.032}$
monotonicity (↑)	$0.537_{\pm 0.02}$	$0.591_{\pm 0.007}$	$\mathbf{0.598}_{\pm 0.002}$	$0.561_{\pm 0.002}$	$0.527_{\pm 0.01}$	$0.575_{\pm 0.012}$
ROAR (↑)	$0.575_{\pm 0.002}$	$0.615_{\pm 0.011}$	$0.616_{\pm 0.008}$	$\mathbf{0.696}_{\pm 0.024}$	$0.534_{\pm 0.018}$	$0.604_{\pm 0.019}$
GT-Shapley (↑)	$0.012_{\pm 0.06}$	$\mathbf{0.870}_{\pm 0.005}$	$0.779_{\pm 0.027}$	$0.804_{\pm 0.011}$	$0.031_{\pm 0.12}$	$0.105_{\pm 0.013}$
infidelity (↓)	$0.207_{\pm 0.125}$	$\mathbf{0.075}_{\pm 0.074}$	$0.077_{\pm 0.075}$	$0.077_{\pm 0.079}$	$0.091_{\pm 0.07}$	$0.117_{\pm 0.076}$

Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)

We should always be able to create artificial data that allows evaluating our algorithms.

Practical experiment

Zhou & Ribeiro et al.
*MIT
Do Feature Attribution Methods Correctly Attribute Features?
(AAAI 2022)

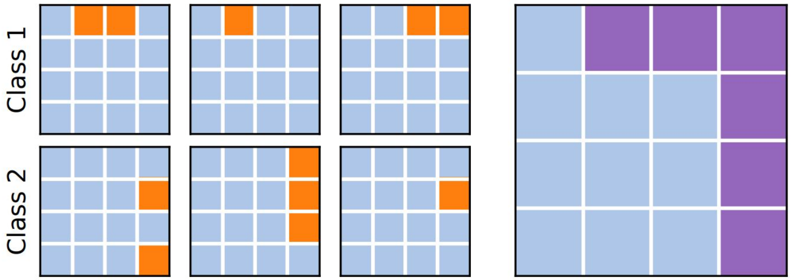


Figure 1: The intuition behind our feature attribution ground truth: if we know that for every input, only specific features (orange) are informative to the label, then across the dataset, a high-performing model has to focus on them and not get "distracted" by other irrelevant features. Thus, feature attributions should highlight the union *union* of these features (purple), and any attribution outside this area is misleading.

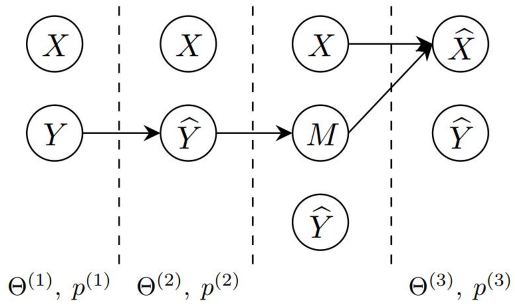
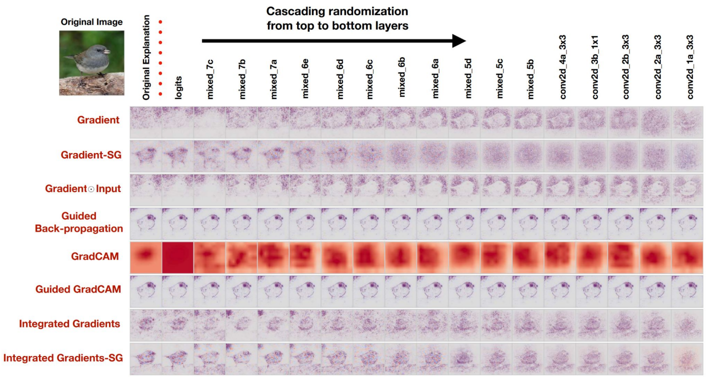


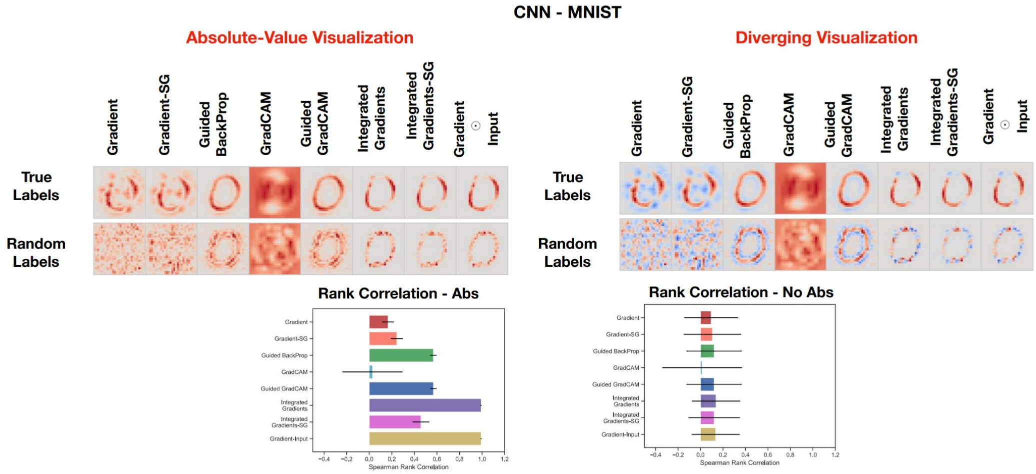
Figure 2: The graphical model for our dataset modification.

Zhou et al. Do Feature Attribution Methods Correctly Attribute Features? (AAAI 2022)

We can evaluate explanations against the crafted ground-truth.

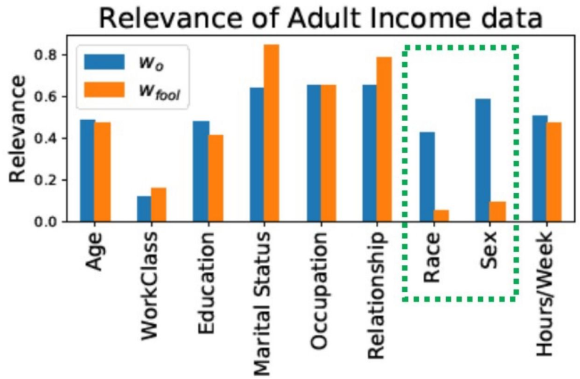
Recap





Adebayo et al. Sanity Checks for Saliency Maps (NeurIPS 2018)

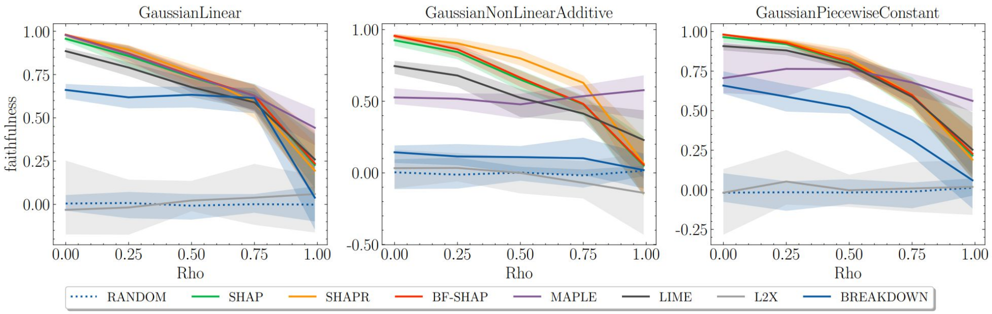
Dombrowski et al. Explanations can be manipulated and geometry is to blame (NeurIPS 2019)



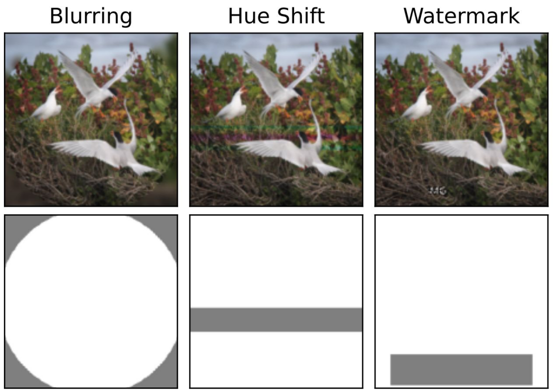
Heo et al. Fooling Neural Network Interpretations via Adversarial Model Manipulation (NeurIPS 2019)

Recap

Poursabzi-Sangdeh et al. Manipulating and Measuring Model Interpretability (CHI 2021)



Liu et al. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (NeurIPS Dataset Track 2021)



Zhou et al. Do Feature Attribution Methods Correctly Attribute Features? (AAAI 2022)