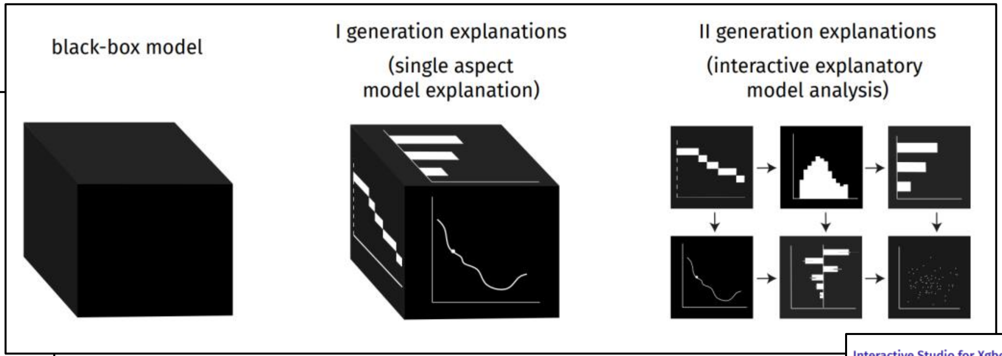
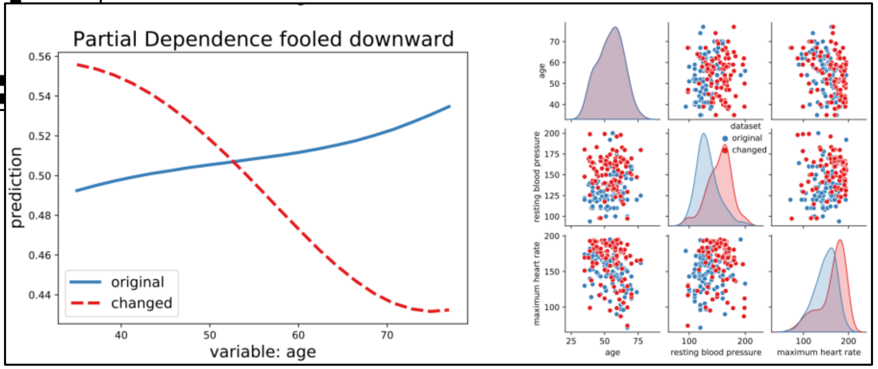
Manipulating explainability and fairness in machine learning

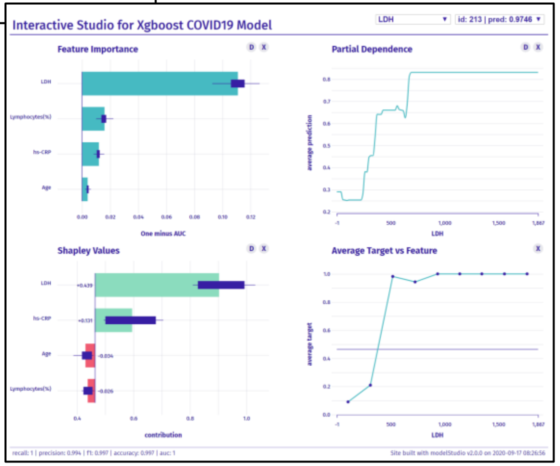
Hubert Baniecki
@ MI2DataLab, Warsaw University of Technology

ML in PL Conference, Poland
November 2021

```
hbaniecki:~$ whoami
- Reseracher and Data Science student at Warsaw University of Technology
- Interested in explainable machine learning, developing open source software
- Specifically joint with adversarial machine learning, and evaluation
```







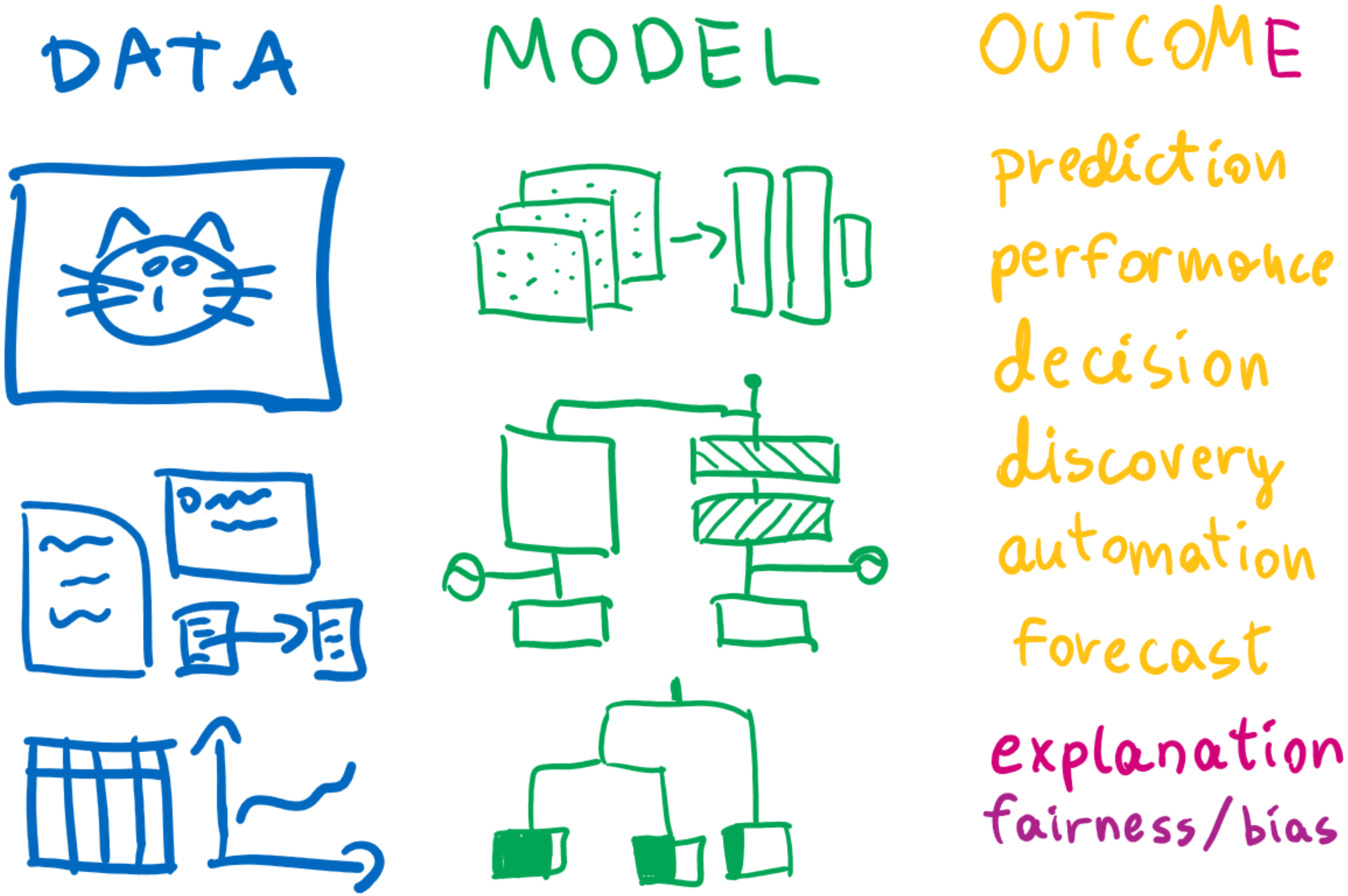
DISCLAIMER

1. Presenting interesting work of others (fair use)

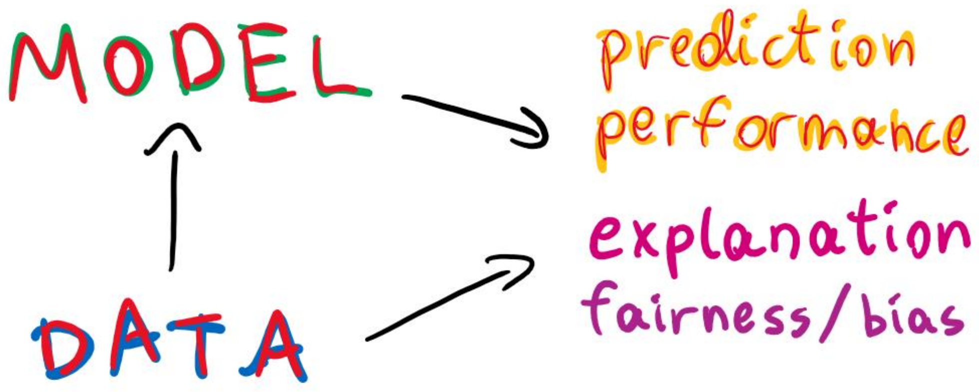
2. Selective survey based on a live list of related work:
   https://github.com/hbaniecki/adversarial-explainable-ai (contributions are welcomed)

3. Omitting technical details and math

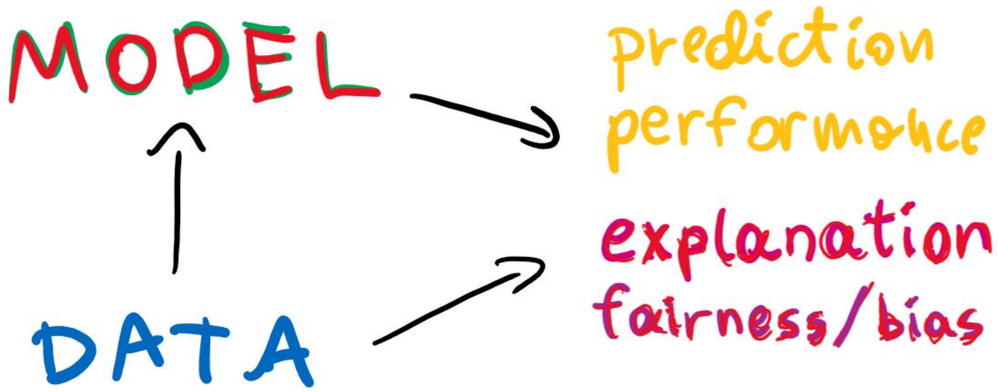
explainability and fairness



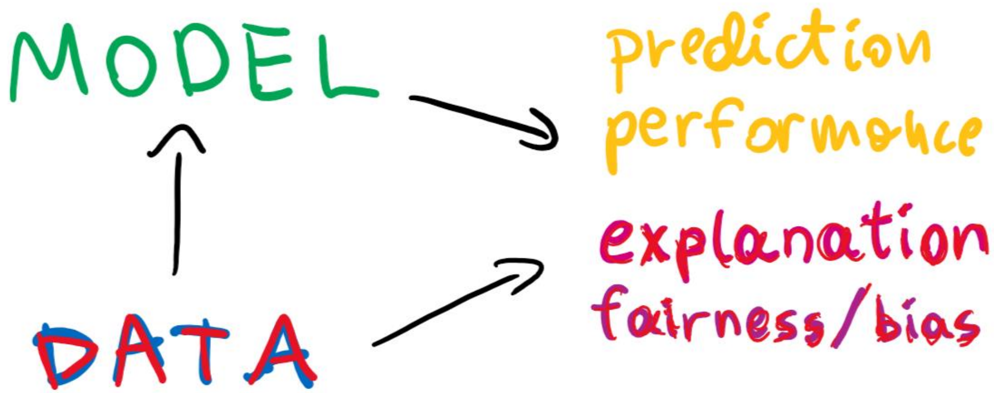
Manipulating explainability and fairness



1. change the model



2. change the data



Manipulating explainability via data change

Target: change explanation/saliency maps
for deep neural network image classification

Method: perturb an image with gradient optimization;
aim for an arbitrary target map

Loss ~
distance(manipulated explanation, target explanation) +
*γ * distance(manipulated prediction, original prediction)*

Result: It is possible to manipulate explanations
(ImageNet + VGG, ResNet, DenseNet)

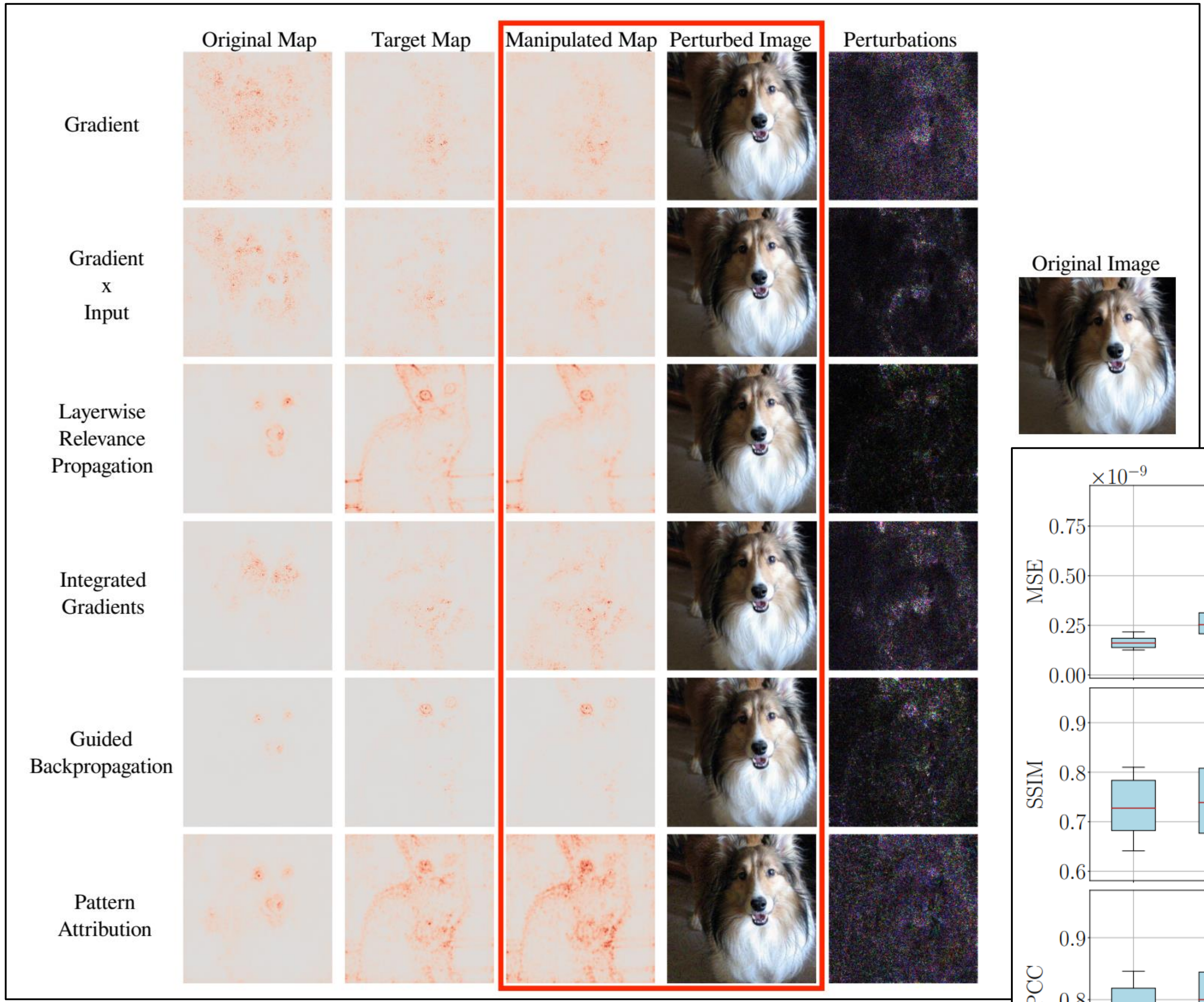
Explanations can be manipulated and geometry is to blame

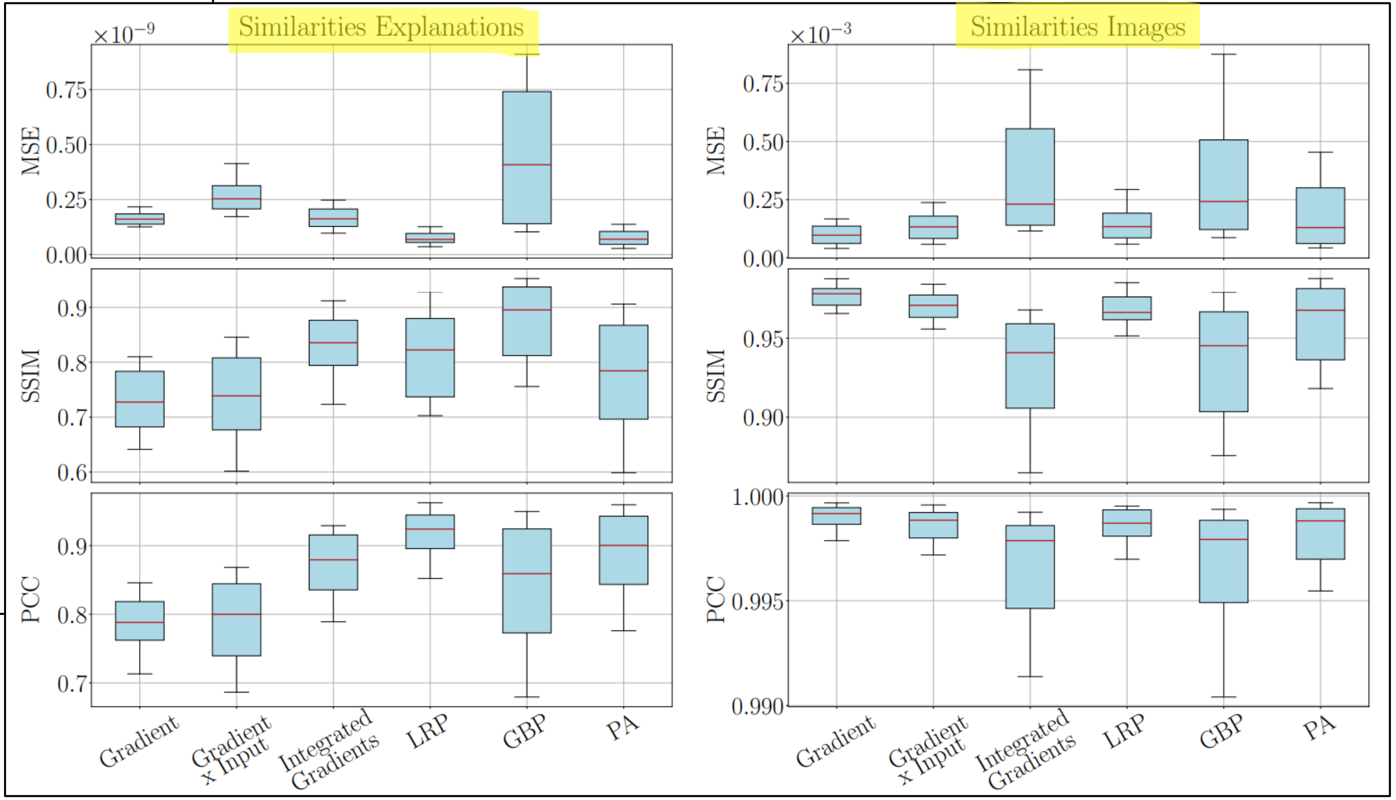
Ann-Kathrin Dombrowski[1], Maximilian Alber[5], Christopher J. Anders[1], Marcel Ackermann[2], Klaus-Robert Müller[1,3,4], Pan Kessel[1]

[1]Machine Learning Group, Technische Universität Berlin, Germany
[2]Department of Video Coding & Analytics, Fraunhofer Heinrich-Hertz-Institute, Berlin, Germany
[3]Max-Planck-Institut für Informatik, Saarbrücken, Germany
[4]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
[5]Charité Berlin, Berlin, Germany

Original Image Manipulated Image

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.





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

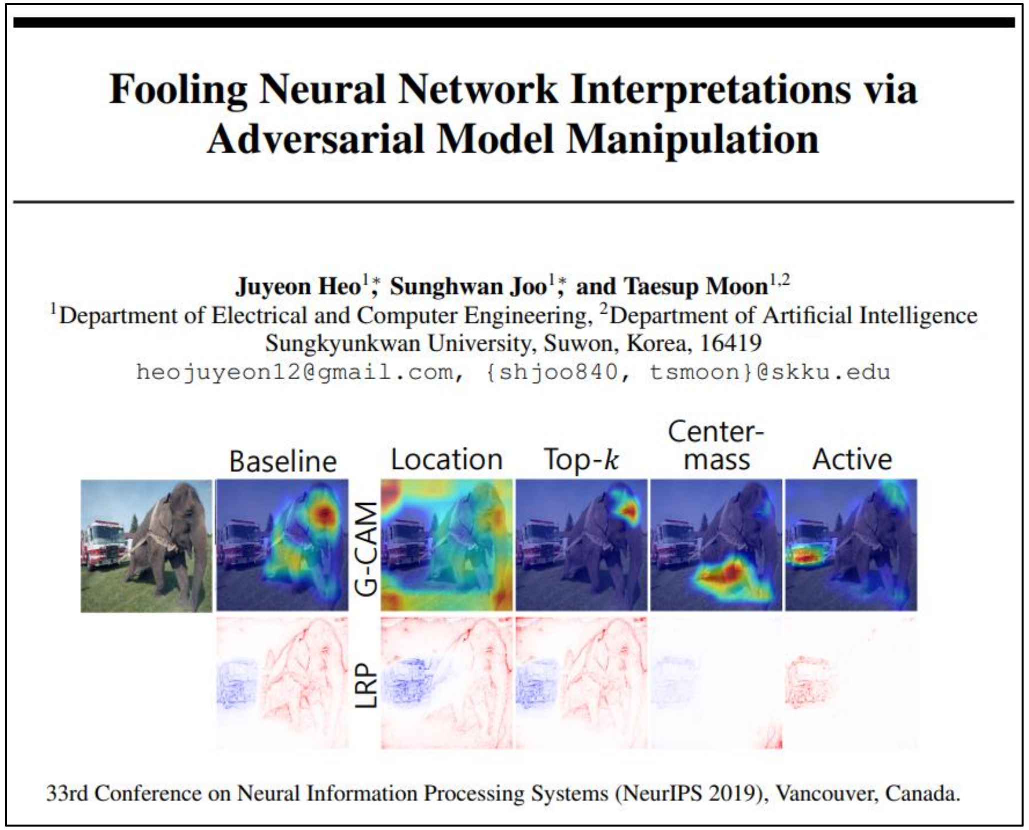
Manipulating explainability via model change

Target: change explanation/saliency maps
for deep neural network image classification

Method: fine-tune a model; change parameters
to generate uninformative or false explanations

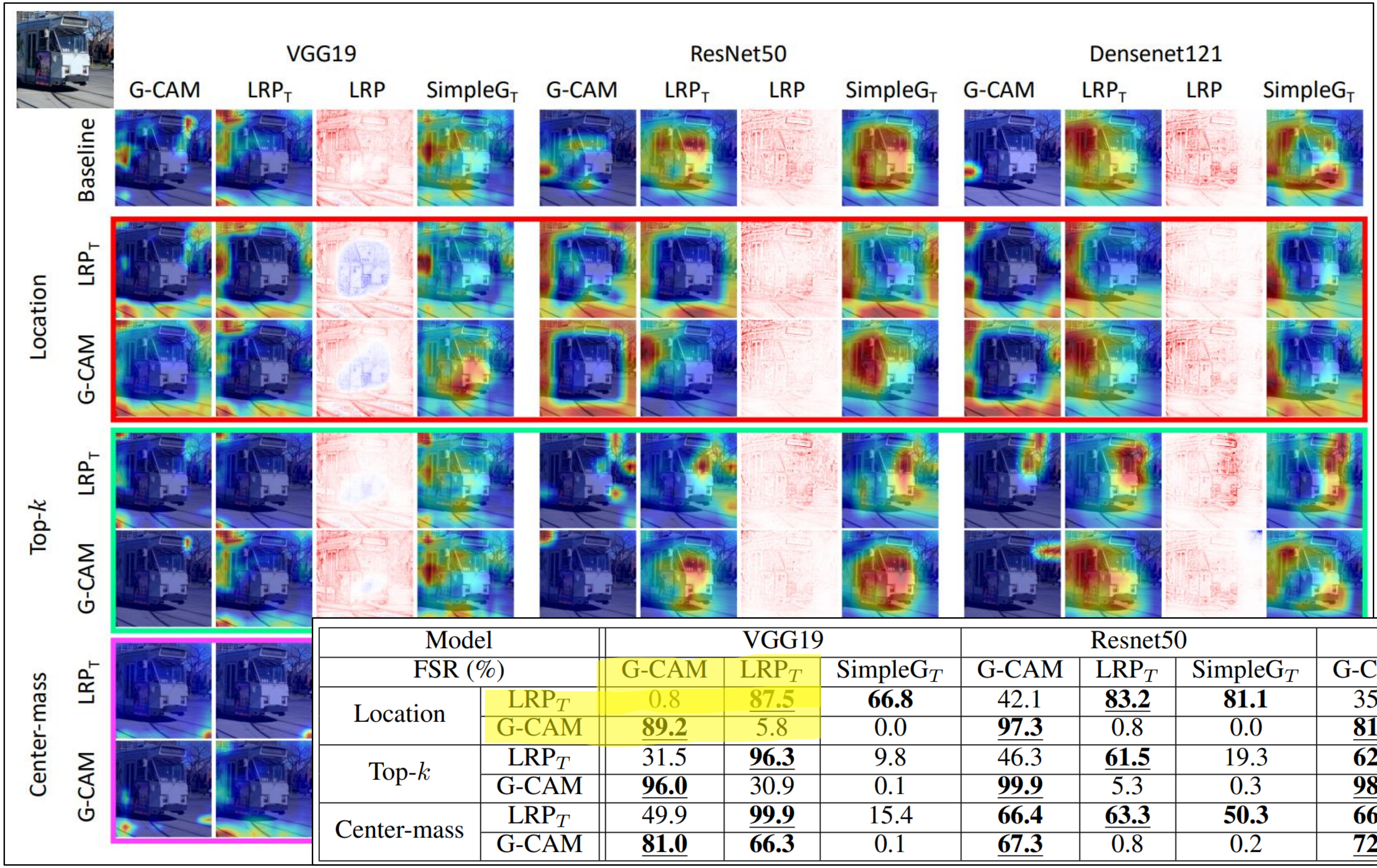
Loss ~
performance of the manipulated model +
*λ*distance(manipulated explanations, target explanations)*

Result: It is possible to manipulate explanations - globally
(ImageNet + VGG, ResNet, DenseNet)


Fooling Neural Network Interpretations via Adversarial Model Manipulation

Juyeon Heo[1]*, **Sunghwan Joo**[1]*, **and Taesup Moon**[1,2]
[1]Department of Electrical and Computer Engineering, [2]Department of Artificial Intelligence
Sungkyunkwan University, Suwon, Korea, 16419
heojuyeon12@gmail.com, {shjoo840, tsmoon}@skku.edu

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



Fooling Success Rate

Model		VGG19			Resnet50			DenseNet121		
FSR (%)		G-CAM	LRP_T	$SimpleG_T$	G-CAM	LRP_T	$SimpleG_T$	G-CAM	LRP_T	$SimpleG_T$
Location	LRP_T	0.8	**87.5**	**66.8**	42.1	**83.2**	**81.1**	35.7	26.6	**88.2**
	G-CAM	**89.2**	5.8	0.0	**97.3**	0.8	0.0	**81.8**	0.4	**92.1**
Top-k	LRP_T	31.5	**96.3**	9.8	46.3	**61.5**	19.3	**62.3**	**53.8**	**66.7**
	G-CAM	**96.0**	30.9	0.1	**99.9**	5.3	0.3	**98.3**	1.9	3.7
Center-mass	LRP_T	49.9	**99.9**	15.4	**66.4**	**63.3**	**50.3**	**66.8**	**51.9**	28.8
	G-CAM	**81.0**	**66.3**	0.1	**67.3**	0.8	0.2	**72.7**	21.8	29.2

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

2018– evaluate explanation maps

Ancona et al. **Towards better understanding of gradient-based attribution methods for Deep Neural Networks**.
International Conference on Learning Representations (ICLR). 2018.

Theoretical unification of Grad*Input, IG, LRP & DeepLIFT methods and evaluating them with an introduced *sensitivity-n* property.

Alvarez-Melis & Jaakkola. **Towards Robust Interpretability with Self-Explaining Neural Networks**.
Neural Information Processing Systems (NeurIPS). 2018.

Introduce a *self-explaining neural network* with native concept-based interpretability and evaluating its faithfulness and stability against the explanation maps.

Adebayo et al. **Sanity Checks for Saliency Maps**.
Neural Information Processing Systems (NeurIPS). 2018.

Develop model randomization and data *randomization tests* to evaluate explanations.

2018~ evaluate explanation maps

Kindermans et al. **The (Un)reliability of Saliency Methods**.
Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer. 2019.

Analyze *input invariance* of saliency maps and the choice of the reference point.

Ghorbani et al. **Interpretation of Neural Networks Is Fragile**.
AAAI Conference on Artificial Intelligence (AAAI). 2019.

Adversarial attack on explanations via gradient-based data perturbations.
(See Domrowski et al. 2019 for differences)

Manipulating explainability and fairness via model change

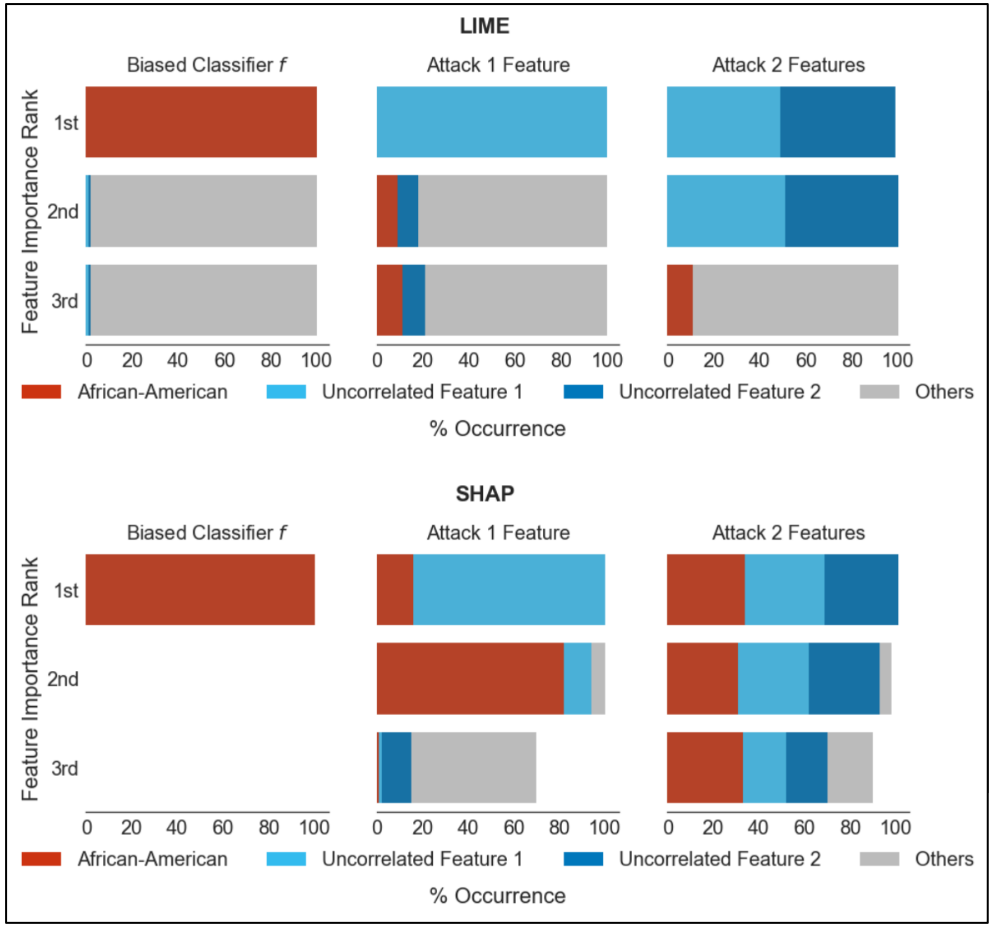
Slack et al. **Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods**. *AAAI/ACM Conference on AI, Ethics, and Society (AIES).* 2020.

Target: change feature attributions of a (biased) black-box model

Framework:

1. build a surrogate model that predicts the same as black-box in-distribution, but arbitrarily out-of-distribution (as needed)
2. support it with another classifier of out-of-distribution samples
3. black-box predictions in-distribution don't change, but LIME and SHAP attributions change as they use perturbed data

Result: Fooled LIME and SHAP attributions



Manipulating fairness via model change

Aivodji et al. **Fairwashing: the risk of rationalization**.
International Conference on Machine Learning (ICML). 2019.

Target: change the value of fairness parity measure of a black-box

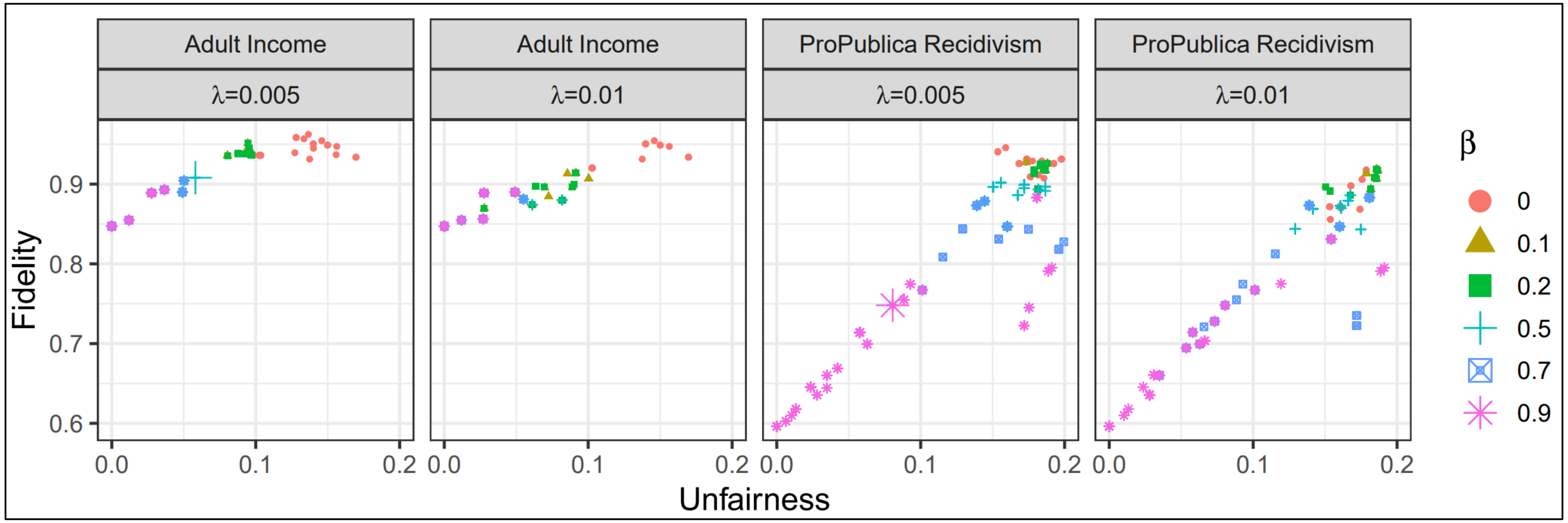
Method: build a fair surrogate model that predicts the same as black-box

$$Loss \sim (1 - \beta) * (1 - fidelity) + \beta * unfairness + \lambda * complexity$$

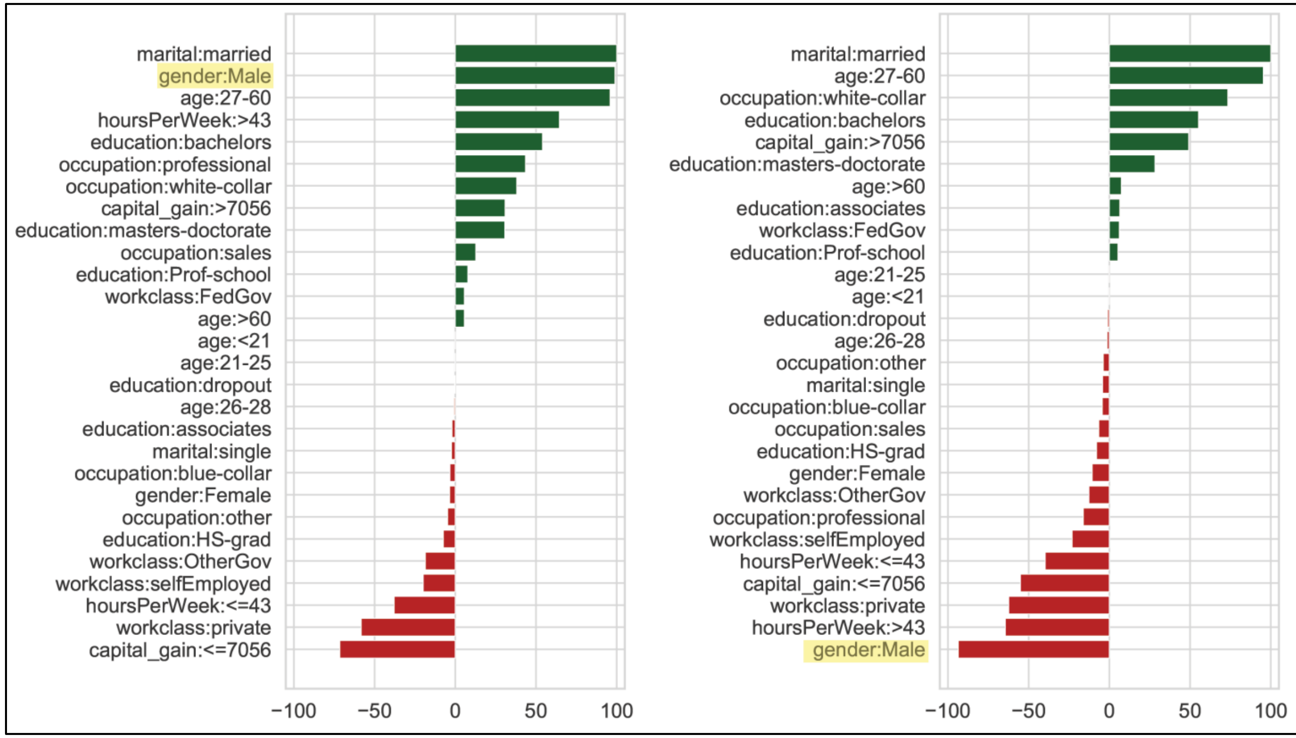
1. fidelity**:** accuracy of predicting the same outcomes as the black-box
2. unfairness: demographic parity measure of bias with respect to the sensitive attribute, e.g. gender or race
3. complexity: a number of rules in a list (serves as a regularization)

In practice: a classification rule list approximates a random forest model

Result: A fairer surrogate model of high fidelity



Relative feature dependence



if capital_gain:>7056 **then**
  income$\geq 50k$
else if marital:single **then**
  income$<50k$
else if education:HS-grad **then**
  income$<50k$
else if occupation:other **then**
  income$<50k$
else if occupation:white-collar **then**
  income$\geq 50k$
else
  income:$<50k$

Aivodji et al. **Fairwashing: the risk of rationalization**. *ICML*. 2019.

Manipulating fairness via data change

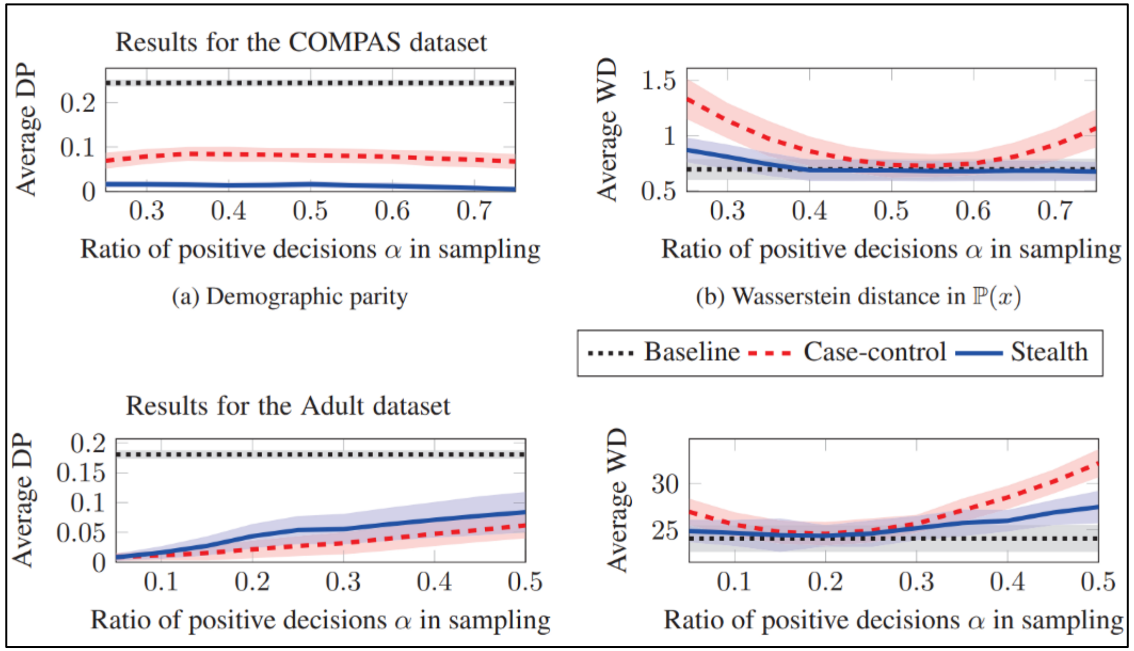
Fukuchi et al. **Faking Fairness via Stealthily Biased Sampling**.
AAAI Conference on Artificial Intelligence (AAAI). 2020.

Target: change the value of fairness parity measure of a black-box

Method: subset a sample of dataset on which a model *appears to* be fair; minimize:

1. similarity of data distributions, specifically Wasserstein distance
2. unfairness: demographic parity measure of bias with respect to the sensitive attribute, e.g. gender or race

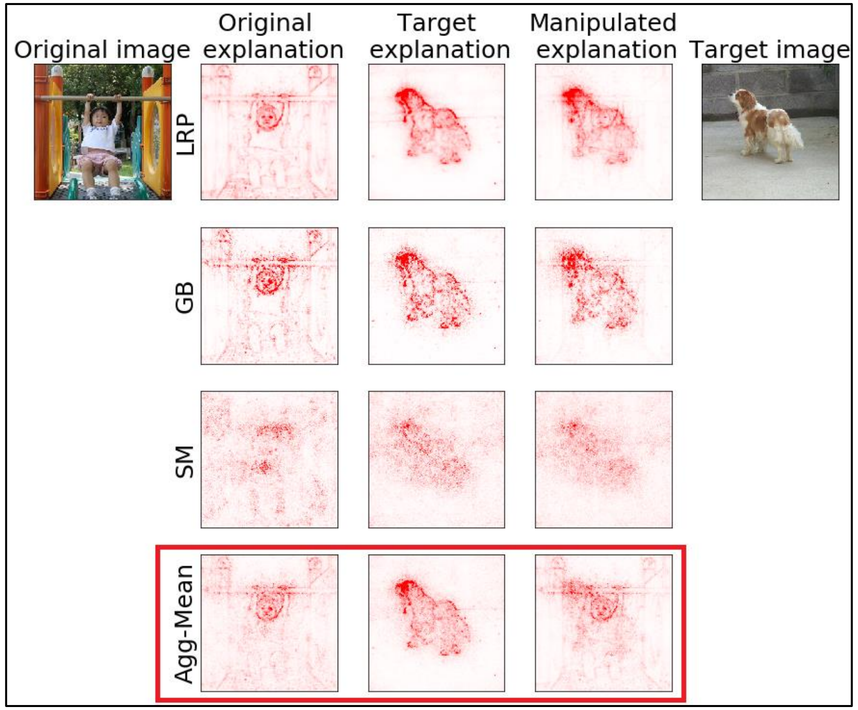
Result: A benchmark dataset, which proves model fairness



Summary & future work

- It is <u>possible</u> to manipulate explanations and fairness measures
- Evaluation approaches are required for a trustworthy adoption of these methods
- Develop defense mechanisms

Vilone & Longo. **Notions of explainability and evaluation approaches for explainable artificial intelligence**. *Information Fusion*. 2021.



Rieger & Hansen. **A simple defense against adversarial attacks on heatmap explanations**. *Workshop on Human Interpretability in Machine Learning (ICML WHI)*. 2020.

Summary & future work (2022)

Investigate critical vulnerabilities in novel explanations

Brittle interpretations: The Vulnerability of TCAV and Other Concept-based Explainability Tools to Adversarial Attack

Anonymous

29 Sept 2021 (modified: 06 Oct 2021) ICLR 2022 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisions

Keywords: interpretability, adversarial attack

One-sentence Summary: We identify a novel vulnerability in the deep learning interpretability pipeline, and design attacks that mislead model explanations for two well known interpretability tools.

Develop robust explanations

Dombrowski et al. **Explanations can be manipulated and geometry is to blame**. *NeurIPS*. 2019.
Dombrowski et al. **Towards robust explanations for deep neural networks**. *Pattern Recognition*. 2022.

Feedback appreciated!

Contact: https://www.linkedin.com/in/hbaniecki

Resources: https://github.com/hbaniecki/adversarial-explainable-ai