

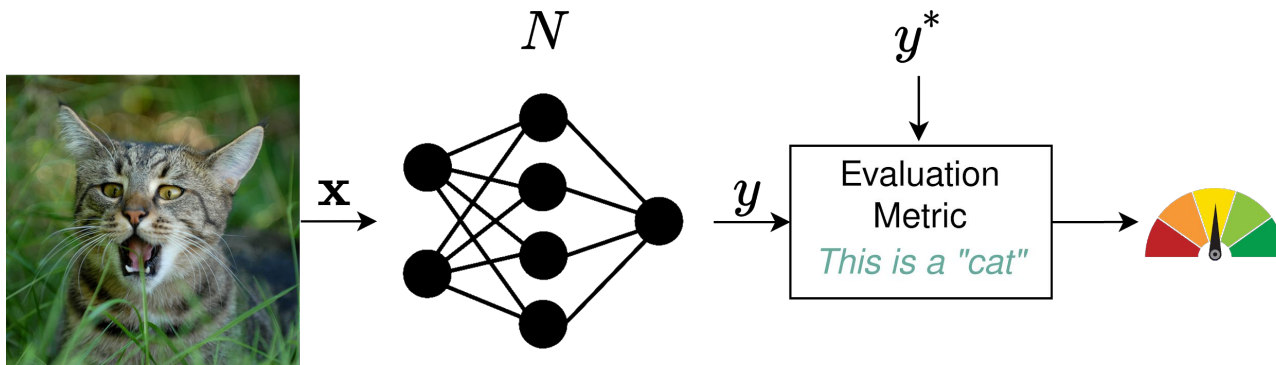
# Why you shouldn't trust me: A survey on Adversarial Model Interpretation Manipulations.

Verena Heusser  
Seminar Explainable Machine Learning

# Motivation

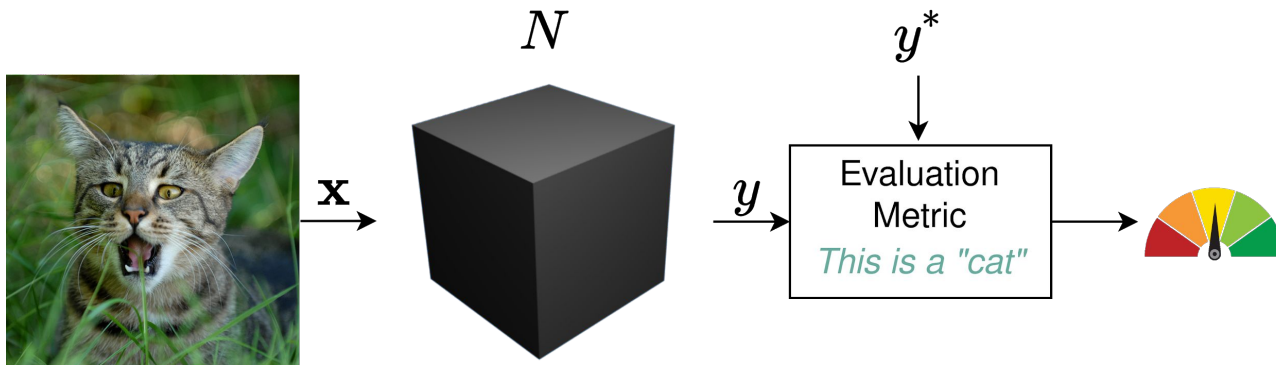
# Motivation: Omnipresent ML

- Machine learning algorithms are moving out of the lab into the real world
- Performance comes at the cost of complexity



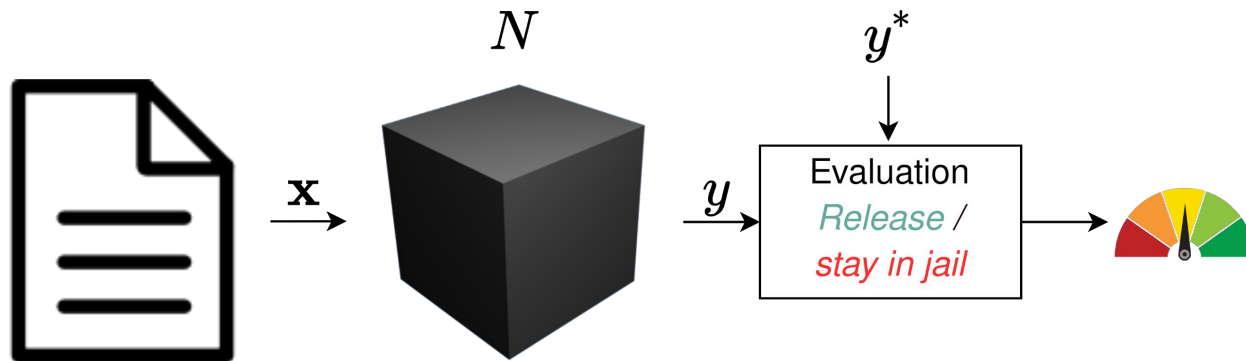
# Motivation: Omnipresent ML

- Machine learning algorithms are moving out of the lab into the real world
- Performance comes at the cost of complexity → black box



# Motivation: Omnipresent ML

- Machine learning algorithms are moving out of the lab into the real world
- Performance comes at the cost of complexity → black box

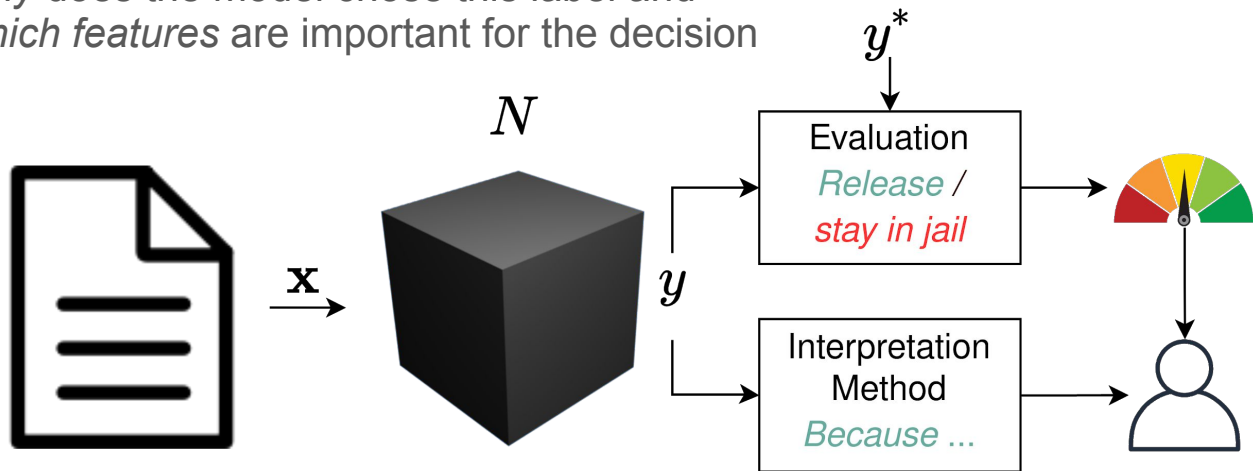


# Motivation: Interpretable ML

- so far: *what* is the most likely label → accuracy
- now also: *why* does the model chose this label and  
*which features* are important for the decision

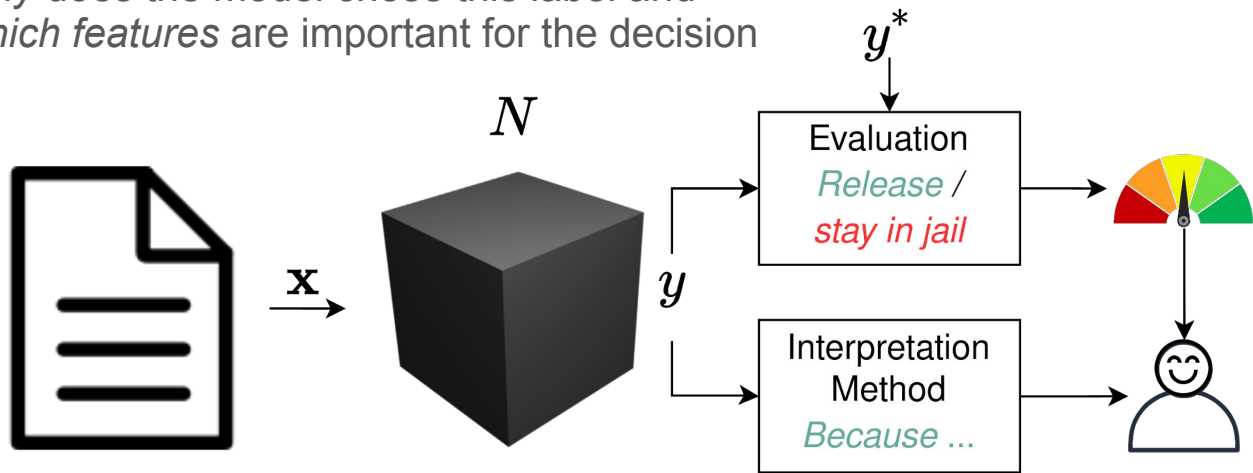
# Motivation: Interpretable ML

- so far: *what* is the most likely label  $\rightarrow$  accuracy
- now also: *why* does the model chose this label and *which features* are important for the decision

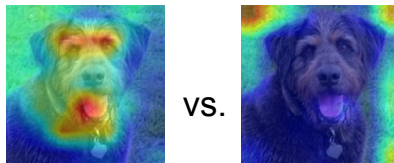


# Motivation: Interpretable ML

- so far: *what* is the most likely label  $\rightarrow$  accuracy
- now also: *why* does the model chose this label and *which features* are important for the decision



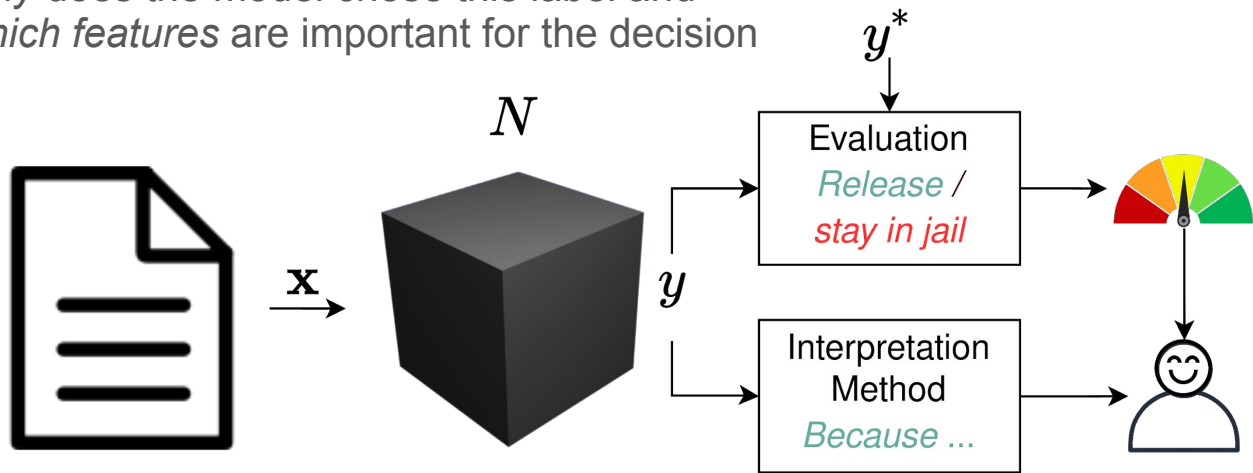
- $\rightarrow$  Reliability, robustness, fairness, trust



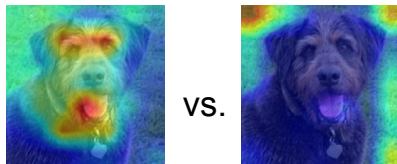


# Motivation: Interpretable ML

- so far: *what* is the most likely label  $\rightarrow$  accuracy
- now also: *why* does the model chose this label and *which features* are important for the decision

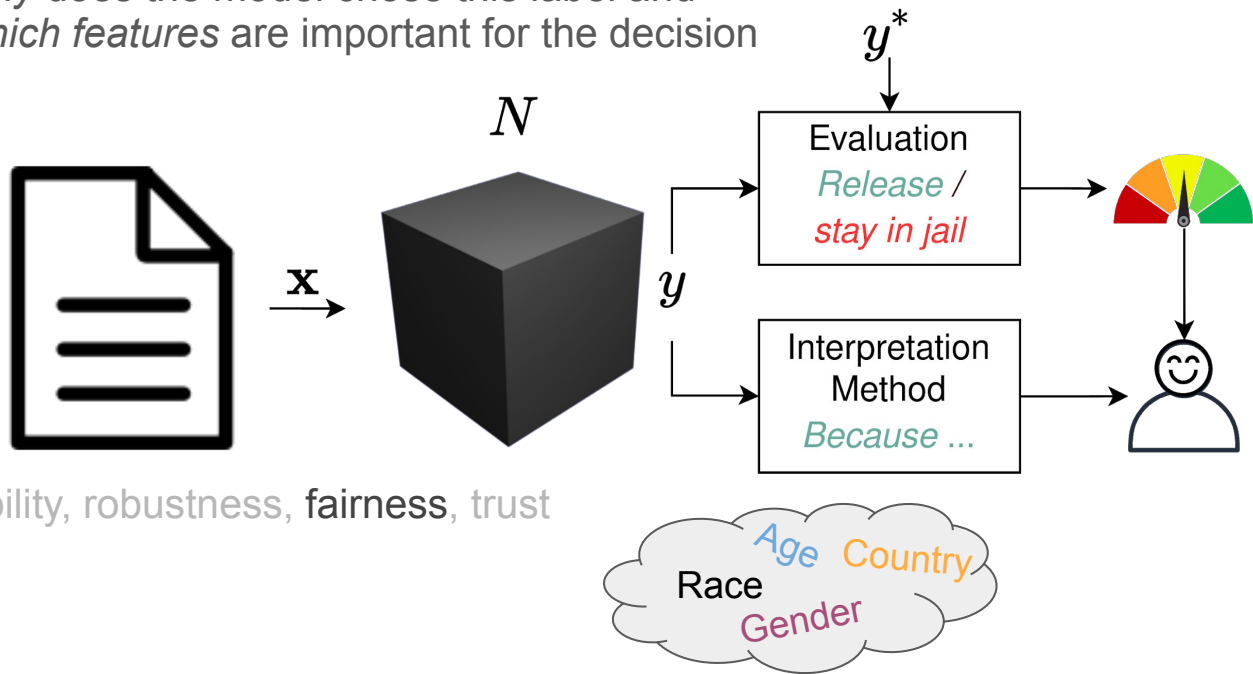


- $\rightarrow$  Reliability, robustness, fairness, trust



# Motivation: Interpretable ML

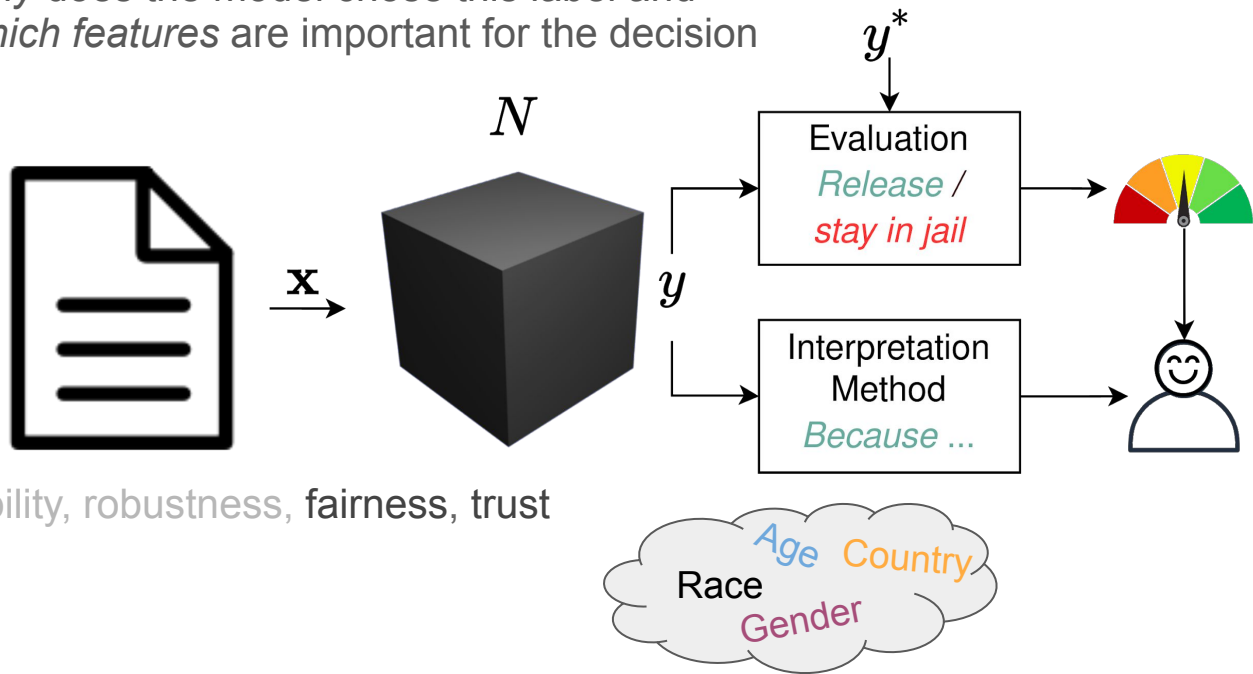
- so far: *what* is the most likely label  $\rightarrow$  accuracy
- now also: *why* does the model chose this label and *which features* are important for the decision



- $\rightarrow$  Reliability, robustness, fairness, trust

# Motivation: Interpretable ML

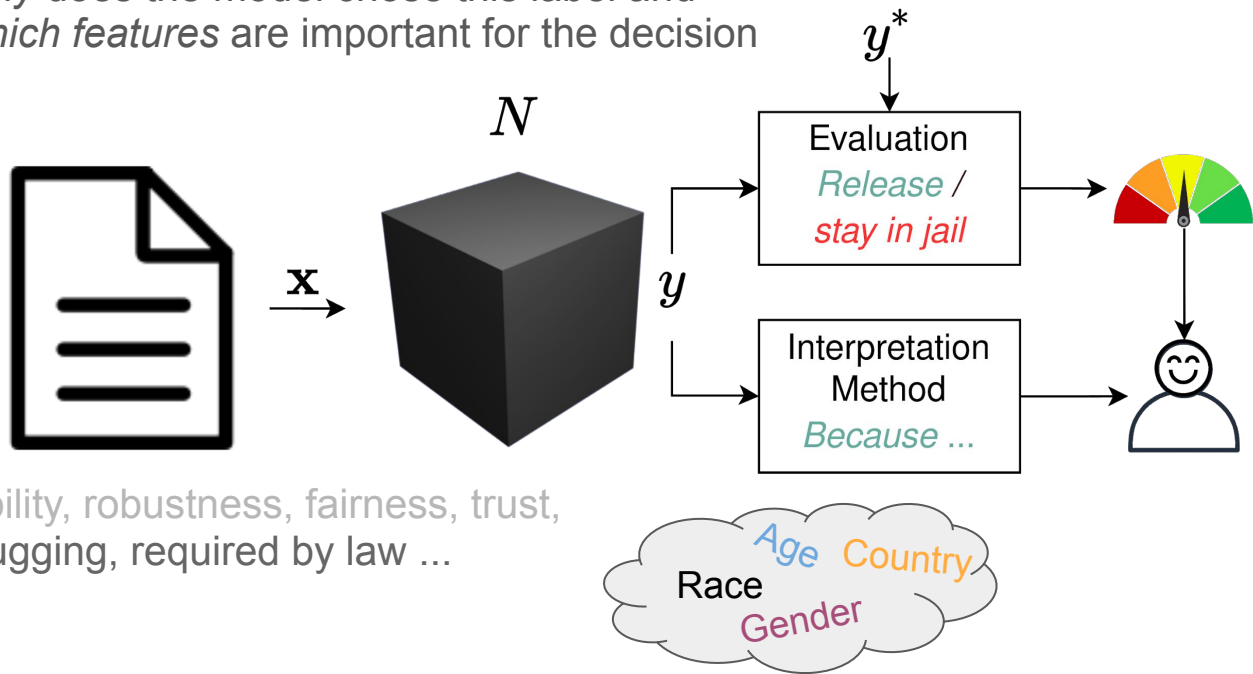
- so far: *what* is the most likely label  $\rightarrow$  accuracy
- now also: *why* does the model chose this label and *which features* are important for the decision



- $\rightarrow$  Reliability, robustness, fairness, trust

# Motivation: Interpretable ML

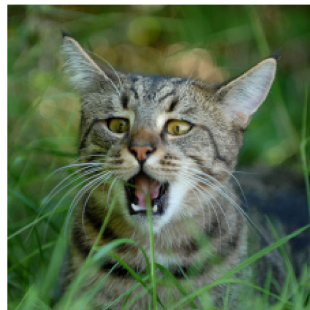
- so far: *what* is the most likely label → accuracy
- now also: *why* does the model chose this label and *which features* are important for the decision



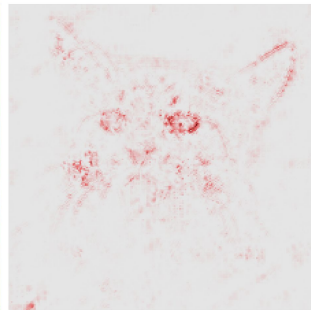
- → Reliability, robustness, fairness, trust,  
+ debugging, required by law ...

# Motivation: Interpretable ML

- Local vs. global
  - Local: Explain the decision  
→ Why is this image a cat?
  - Global: explain the whole model  
→ What does a cat look like?
- White box vs. black box
  - White box: use the model itself to compute interpretations
  - Black box: use an interpretable model to mimic an uninterpretable model



(a) Original.



(b) Map.

LRP [Bach et al., 2015]



(a) Original



(b) Mask.



(c) Saliency Map.

LIME [Ribeiro et al., 2016]

# Motivation: Interpretable ML

- Problem solved? → Not quite ...
- Interpretation methods are already used in many domains for model validation
- However
  - Humans do not benefit from interpretation methods
    - they cannot build better models [Hase et al., 2020]
    - improve their performance [Hase et al., 2020]
    - and are not better at detecting false model decisions [Poursabzi-Sangdeh et al., 2018]
  - Methodological difficulties: it is unclear
    - how to evaluate
    - how to compare different interpreters

# Motivation: Adversarial ML

- Adversarial model fooling
  - attacks on the **model**
  - altered input [\[Szegedy et al., 2013\]](#)  
→ model makes false predictions



correctly labeled  
image

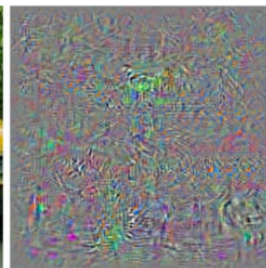


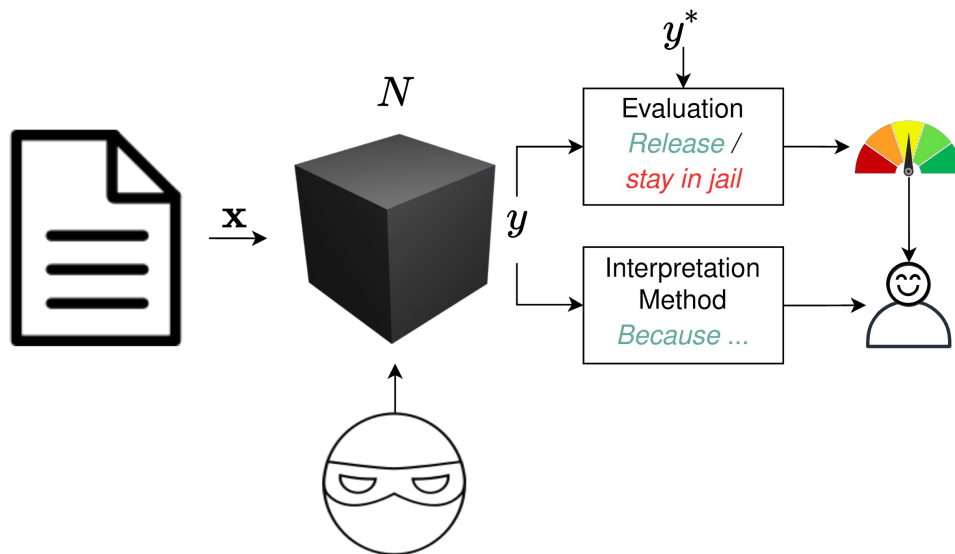
image difference



incorrectly  
labeled image

# Motivation: Adversarial ML

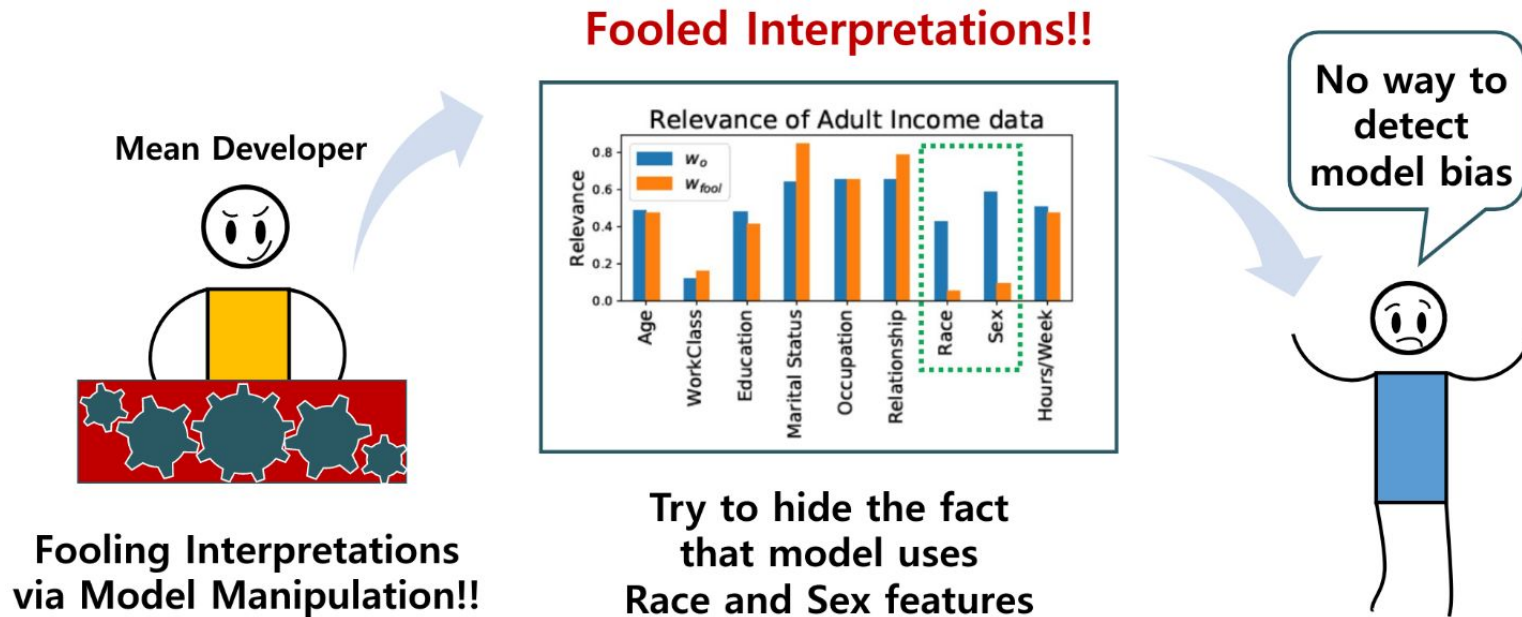
- Adversarial model fooling
  - attacks on the **model**
  - altered input [\[Szegedy et al., 2013\]](#)  
→ model makes false predictions
- Adversarial interpreter fooling
  - attacks on the **interpreter**
  - → interpreter makes false interpretations





# Motivation: Adversarial ML

- Adversarial Interpreter Fooling

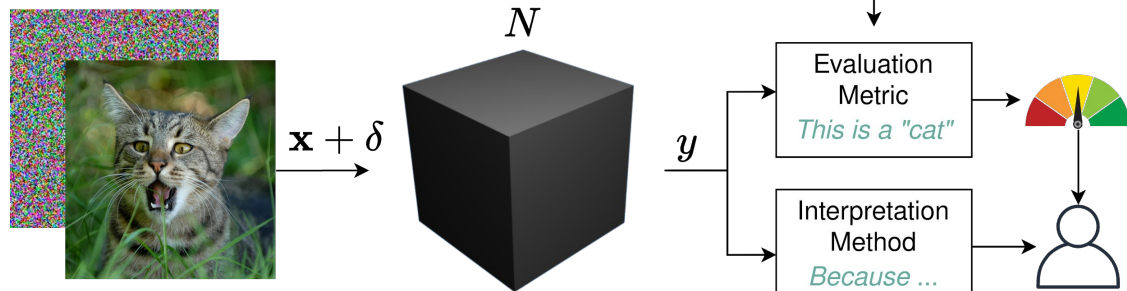


# Manipulation Methods

# Manipulation Types

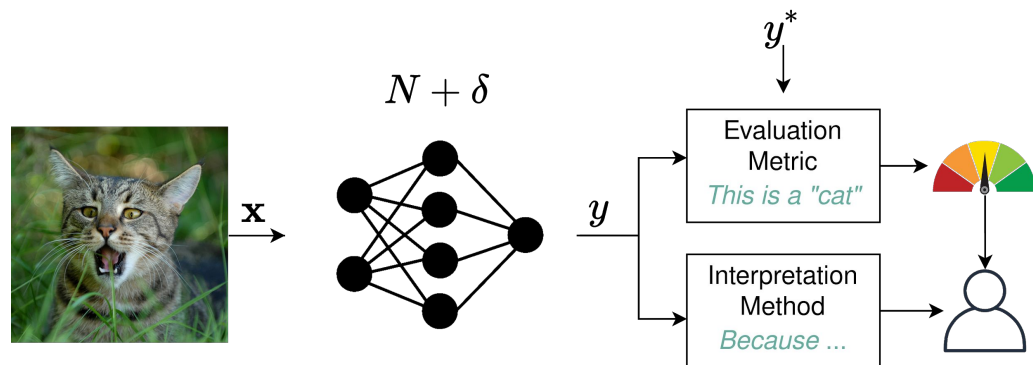
- **Input Level Manipulations**

- [Subramanya et al., 2019]
- [Dombrowski et al., 2019]
- [Ghorbani et al., 2019]



- **Model Level Manipulations**

- [Heo et al., 2019]
- [Dimanov et al., 2020]
- [Slack et al., 2020]

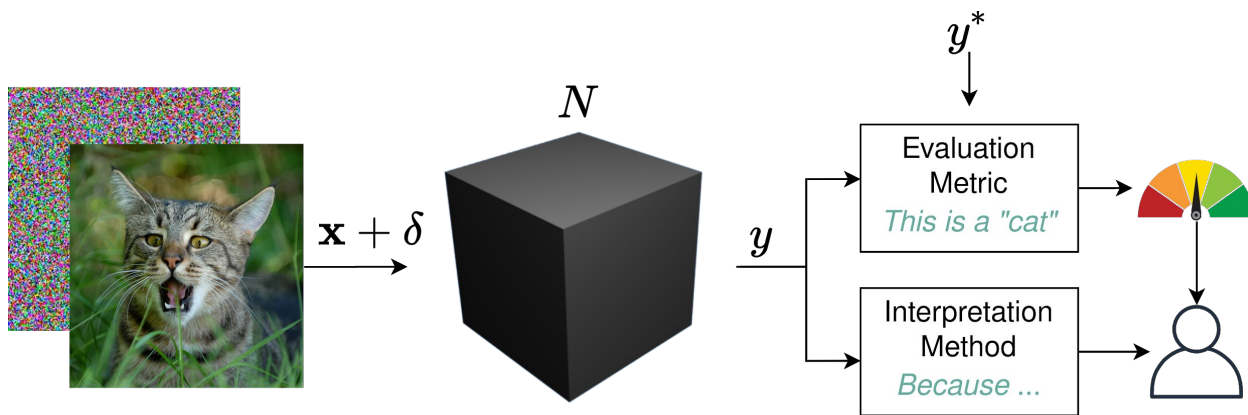


# Evaluation Criteria: Is the fooling successful?

- Fooling successful if [\[Dimanov et al., 2020\]](#)
  - (Model prediction similarity)
  - Interpretation dissimilarity
- Other criteria
  - Effectiveness: no computational overhead
  - Transferability: manipulation does not only affect one type of interpretation
- Evaluation → which is the best interpreter?
  - Qualitative Evaluation: Inspection and random sampling
  - Quantitative Evaluation → similarity scores

# Interpreter Manipulation Examples

## Input Level



$$\mathbf{x} + \delta$$

---

## **Explanations can be manipulated and geometry is to blame**

---

**Ann-Kathrin Dombrowski<sup>1</sup>, Maximilian Alber<sup>5</sup>, Christopher J. Anders<sup>1</sup>,  
Marcel Ackermann<sup>2</sup>, Klaus-Robert Müller<sup>1,3,4</sup>, Pan Kessel<sup>1</sup>**

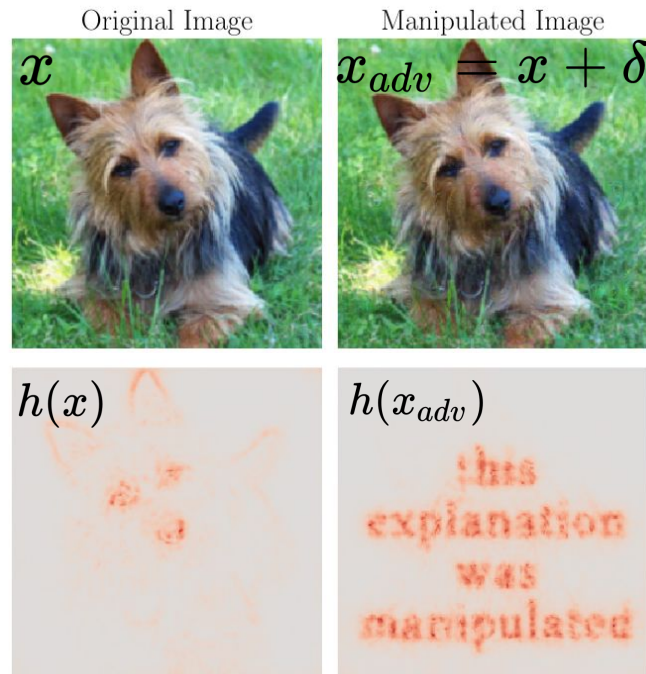
# Explanations can be manipulated and geometry is to blame $\mathbf{x} + \delta$

[Dombrowski et al., 2019]

- **manipulate an image** with a hardly perceptible perturbation such that the explanation map matches an arbitrary target map

$$\mathcal{L} = \|h(x_{adv}) - h^t\|^2 + \gamma \|g(x_{adv}) - g(x)\|^2$$

- Practical implication:
  - adversary can imperceptibly change the input to a model  
→ arbitrary + drastic manipulation of the interpreter



## Further Studies

$$\mathbf{x} + \delta$$

- Learned adversarial patches can cause both model and interpreter to fail [Subramanya et al., 2019]



(a) Original.  
Pred: 'French bulldog'.



(b) Original with  
Patch.



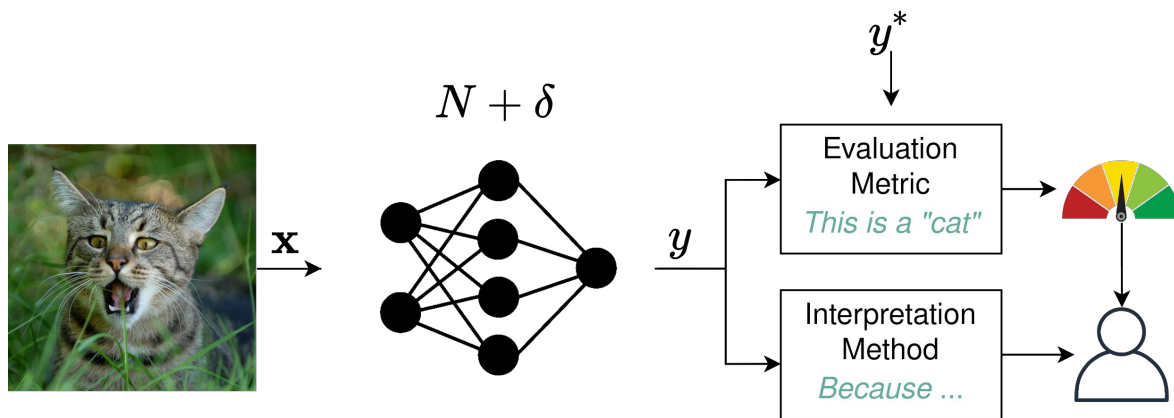
(c) Fooled Map, fooled  
model. Pred: 'Soccer ball'.

- Interpreters are susceptible even to infinitesimal perturbations [Ghorbani et al., 2019]



# Interpreter Manipulation Examples

## Model Level



$$N + \delta$$

---

# Fooling Neural Network Interpretations via Adversarial Model Manipulation

---

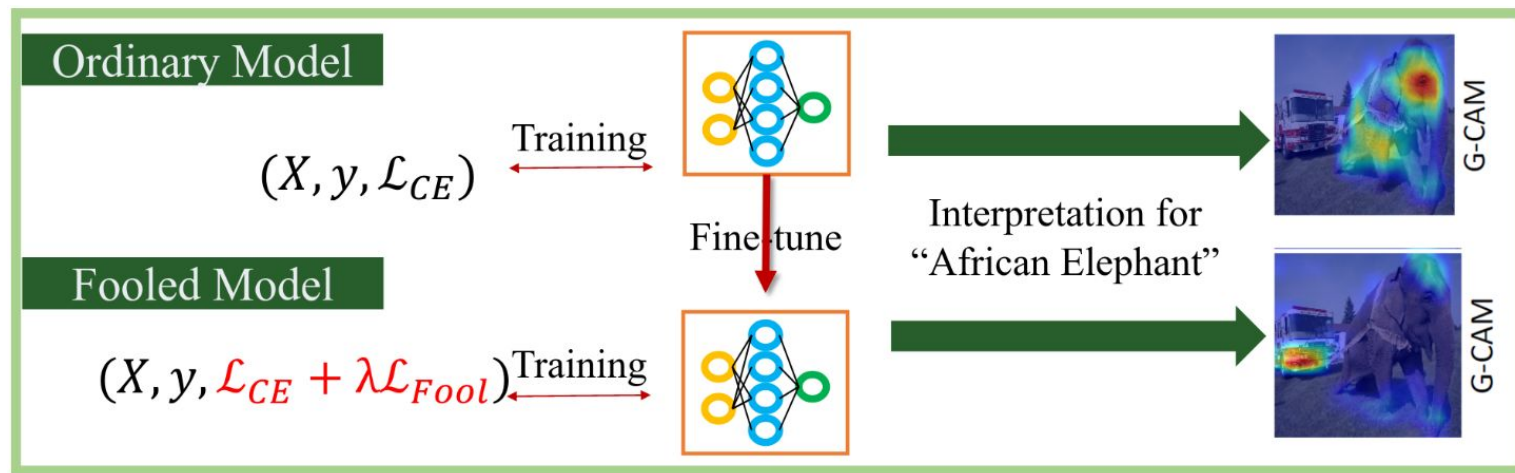
**Juyeon Heo<sup>1,\*</sup>, Sunghwan Joo<sup>1,\*</sup>, and Taesup Moon<sup>1,2</sup>**

<sup>1</sup>Department of Electrical and Computer Engineering, <sup>2</sup>Department of Artificial Intelligence  
Sungkyunkwan University, Suwon, Korea, 16419

heojuyeon12@gmail.com, {shjoo840, tsmoon}@skku.edu

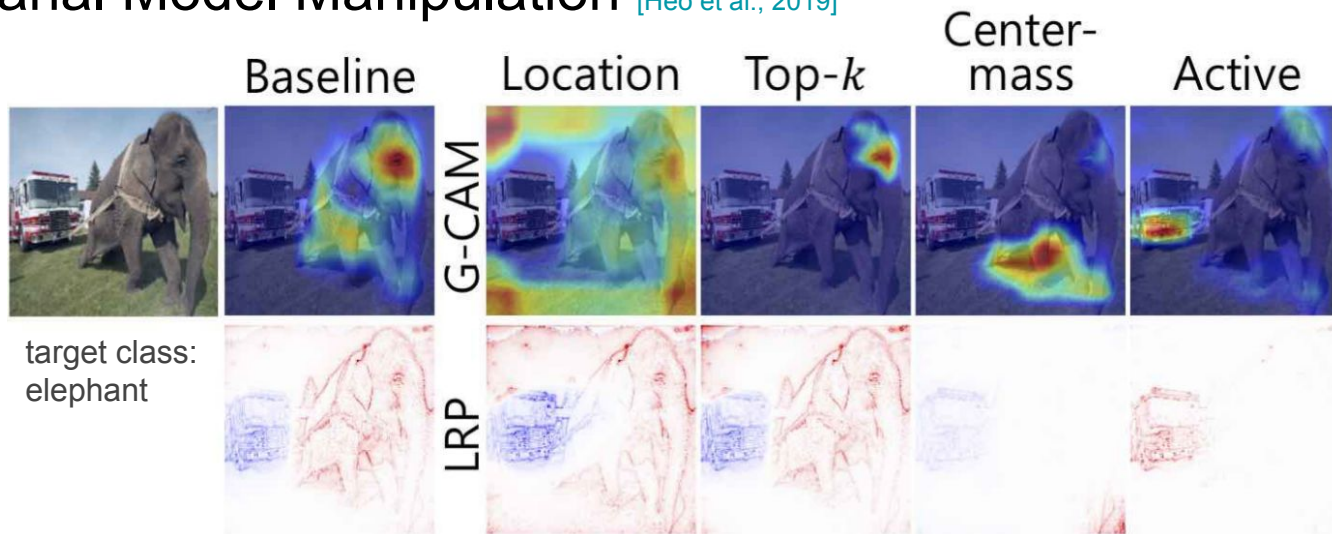
# Fooling Network Interpretations via Adversarial Model Manipulation [Heo et al., 2019]

$$N + \delta$$



# Fooling Network Interpretations via Adversarial Model Manipulation [Heo et al., 2019]

$$N + \delta$$



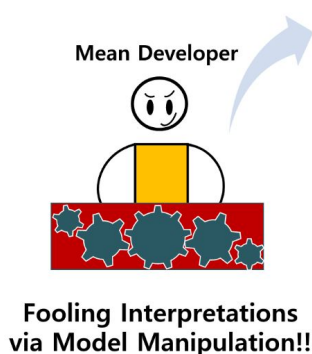
(b) Examples of different kinds of foolings

# Fooling Network Interpretations via Adversarial Model Manipulation [Heo et al., 2019]

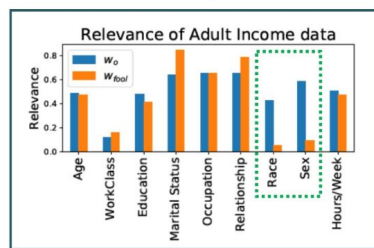
$$N + \delta$$

- Results:
  - generalization to unseen test samples
  - different types of interpreters are be fooled
  - while the model performance stays approx. the same
  - → the model is robust but the interpreter is not

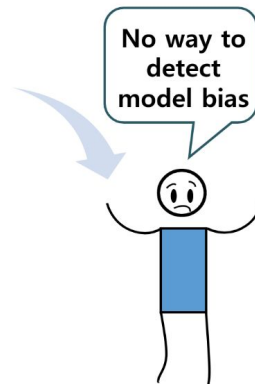
- Practical implication:
  - No way to detect the model inherent bias
  - Interpreters can be systematically manipulated to contain unfair biases



## Fooled Interpretations!!



Try to hide the fact  
that model uses  
Race and Sex features



$$N + \delta$$

## Further Studies

- Interpreters fail to decide if a model is fair [\[Dimanov et al., 2020\]](#)
  - create adversarial models that focus only on sensitive features
  - → model interpreters fail to incorporate fairness and fail to detect model biases

→ use real-world datasets

⇒ Core motivational concern of Interpretable ML

# Conclusion

# Summary

→ Adversarial setting for fooling model interpreters

- Interpretation methods can be tricked by applying input and model perturbations
  - interpreters can be fooled with simple input perturbations
  - Biases can be encoded into the model
  - and there might be no way to uncover the hidden biases

$$\mathbf{x} + \delta$$

$$N + \delta$$



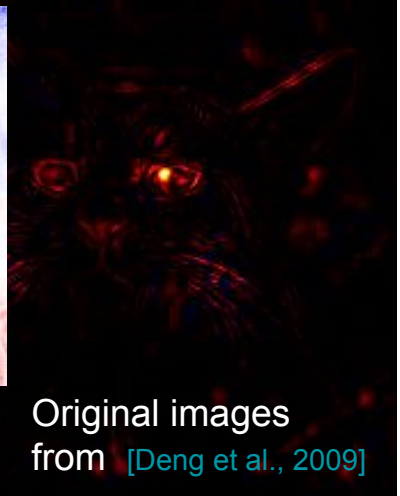
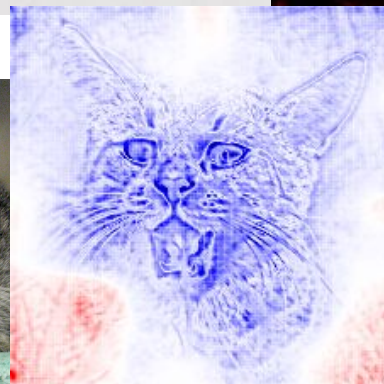
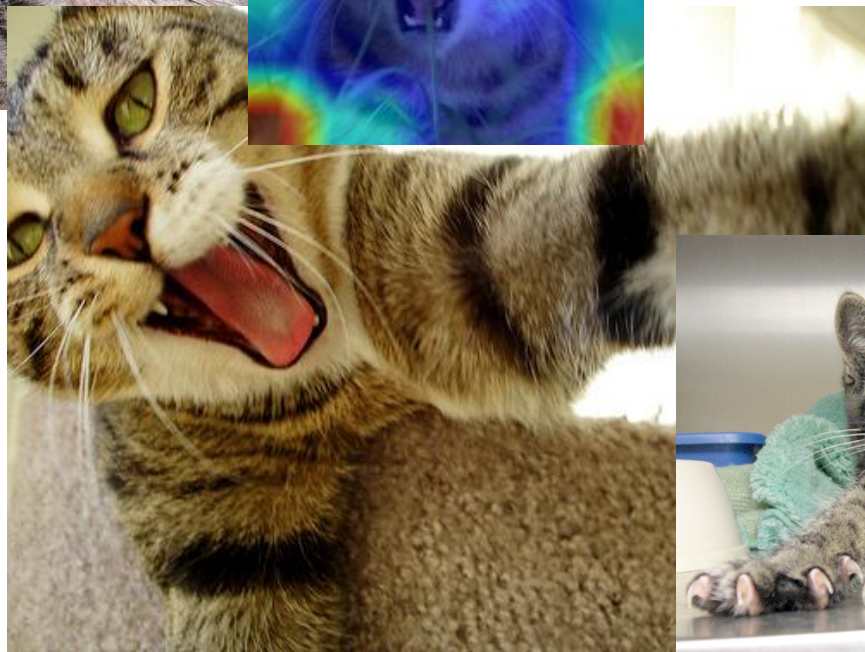
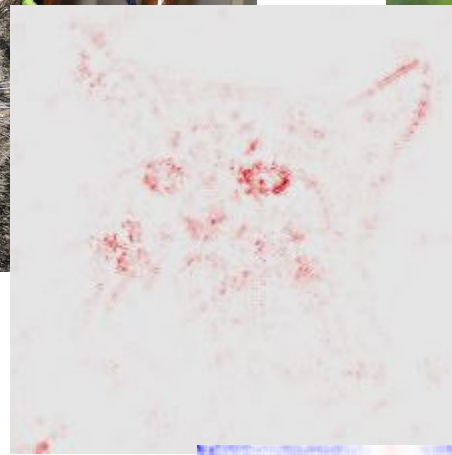
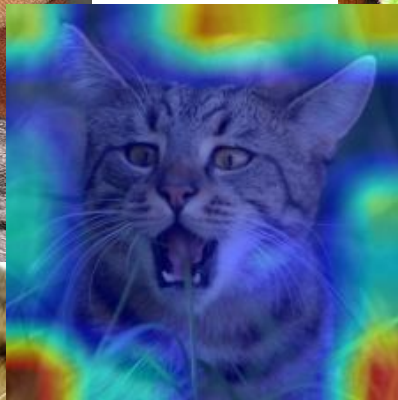
# Conclusion

- Models and interpreters can be misled in a large and systematic manner
- However, this does not mean that interpreters are useless

⇒ **Caution** when using interpretation techniques

⇒ Future work:

- Benchmarking
- Robustness
- Theoretical understanding
- Extension to other task domains



Original images  
from [Deng et al., 2009]

# References

# References

- [Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.
- [Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.
- [Dimanov et al., 2020] Dimanov, B., Bhatt, U., Jamnik, M., & Weller, A. (2020, February). You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods. In *SafeAI@ AAAI* (pp. 63-73).
- [Dombrowski et al., 2019] Dombrowski, A. K., Alber, M., Anders, C., Ackermann, M., Müller, K. R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems* (pp. 13589-13600).
- [Ghorbani et al., 2019] Ghorbani, A., Abid, A., & Zou, J. (2019, July). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 3681-3688).
- [Hase et al., 2020] Hase, P., & Bansal, M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. *arXiv preprint arXiv:2005.01831*.
- [Heo et al., 2019] Heo, J., Joo, S., & Moon, T. (2019). Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems* (pp. 2925-2936).
- [Subramanya et al., 2019] Subramanya, A., Pillai, V., & Pirsaviash, H. (2019). Fooling network interpretation in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2020-2029).
- [Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- [Poursabzi-Sangdeh et al., 2018] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.
- [Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "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* (pp. 1135-1144).