# A Tutorial on Model Explainability

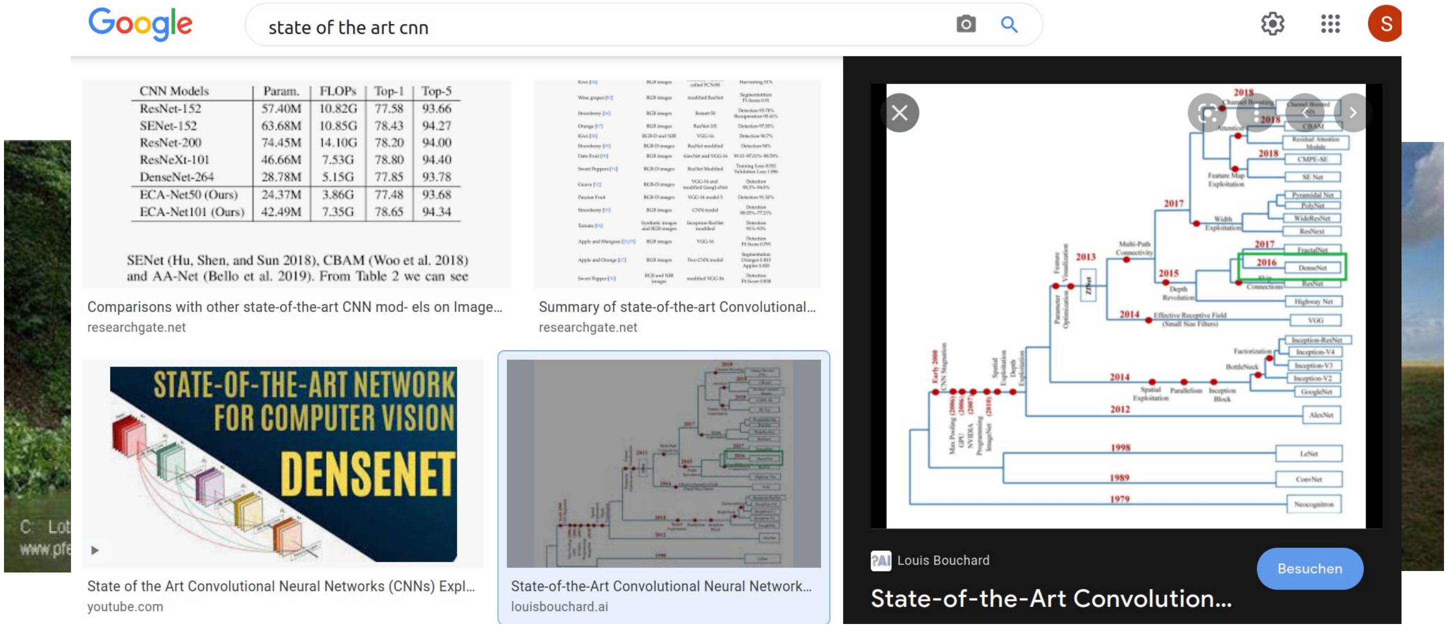
Sahra Ghalebikesabi

# Let us train a state of the art AI!

or

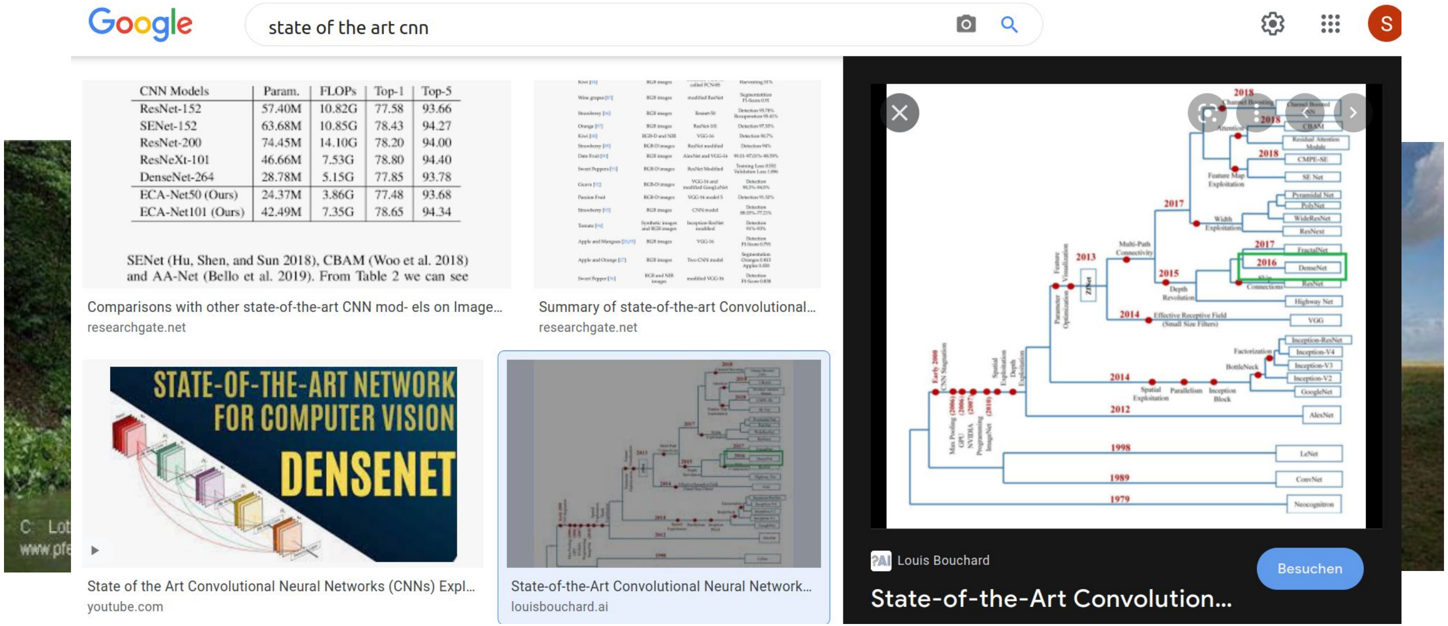

Images from the Pascal Voc 2007 data set

# Let us train a state of the art AI!

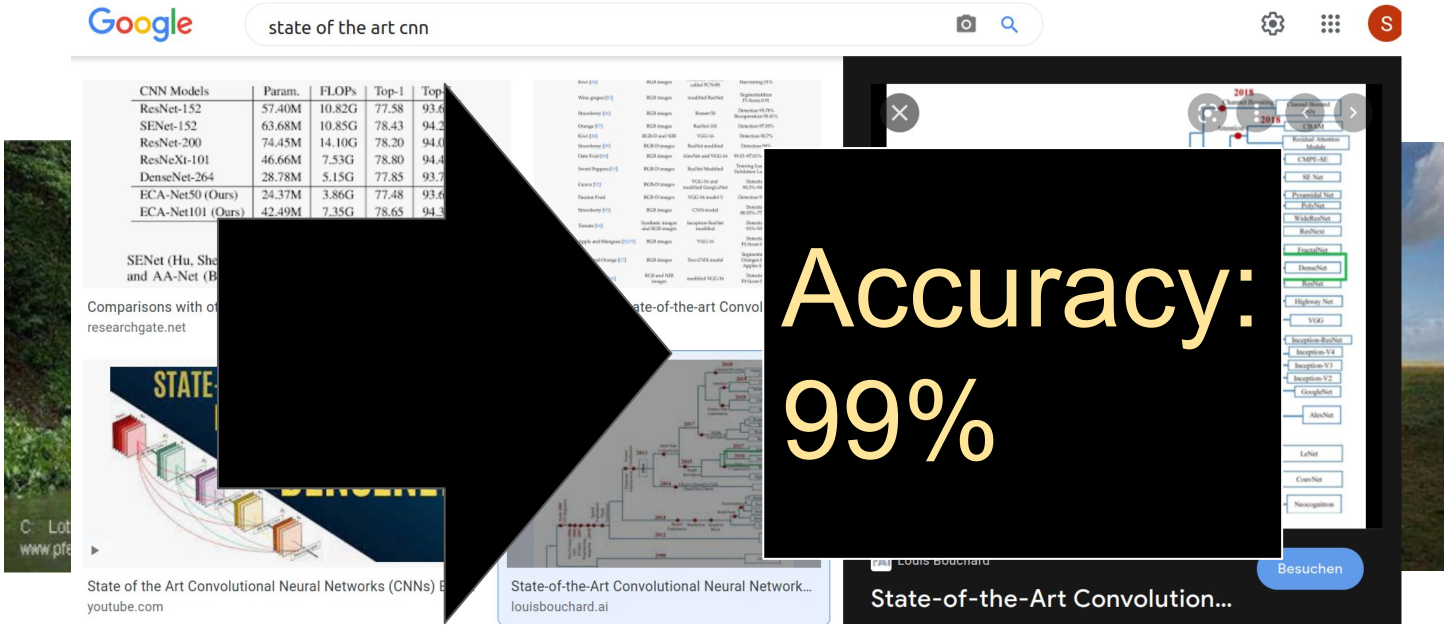


# Let us train a state of the art AI!

**My contribution:**



# Let us train a state of the art AI!



# Some people thought it might be cool to explain my model

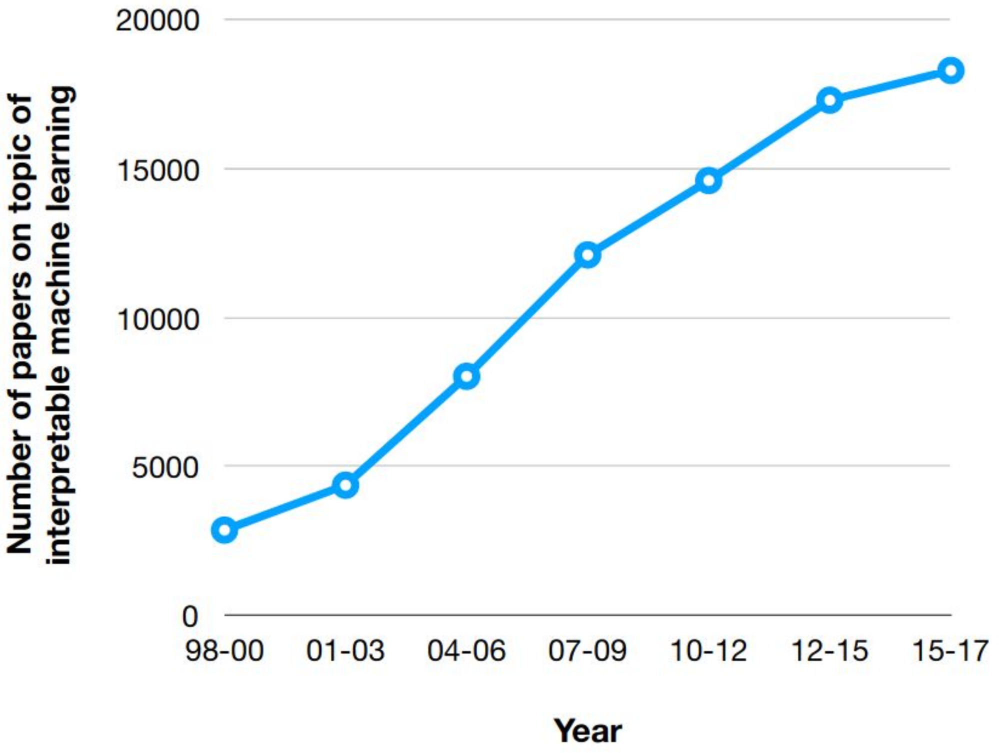



Image from Been Kim's DLSS 2018 talk

# Let us train a state of the art AI!

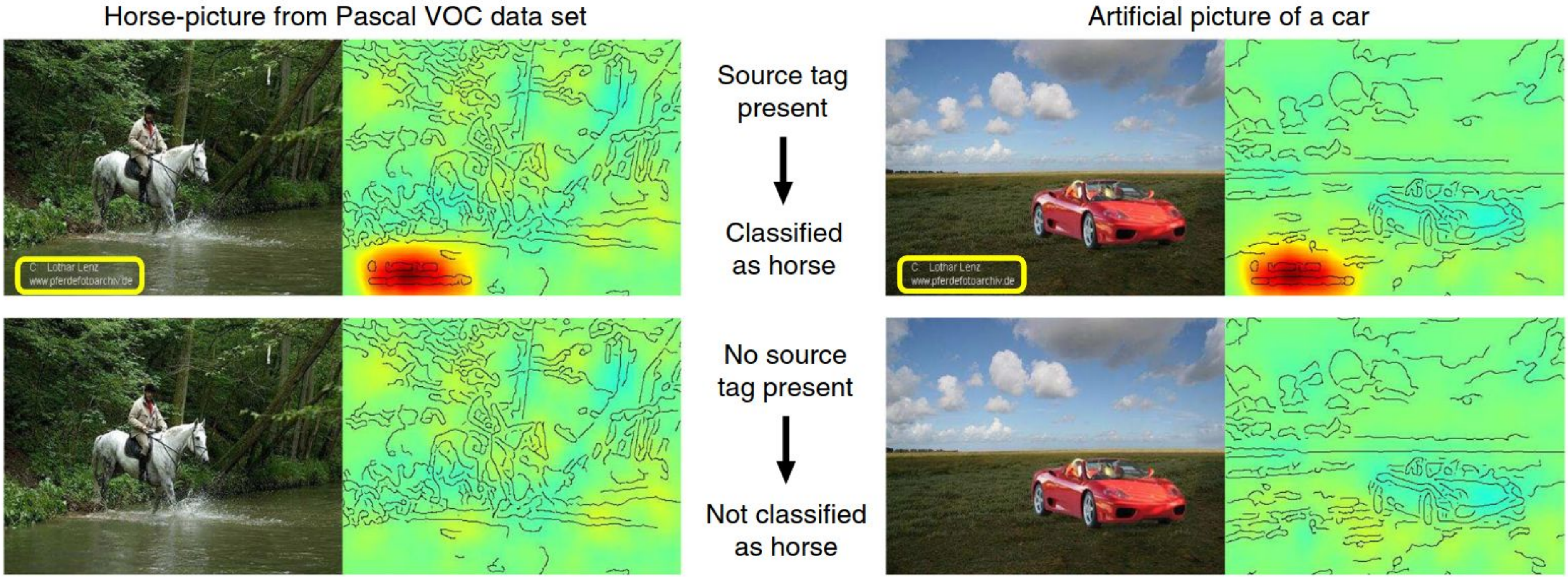



Image from Lapuschkin (2019)

# What is an Explanation Model?

# What is an Explanation Model?

- Model that explains a different model

# What is an Explanation Model?

- Model that explains a different model

→ ask a **better** question

# What is the **GOAL** of your explanation model?

- Understanding
- Trust
- Feature Selection
- Actionable Advice

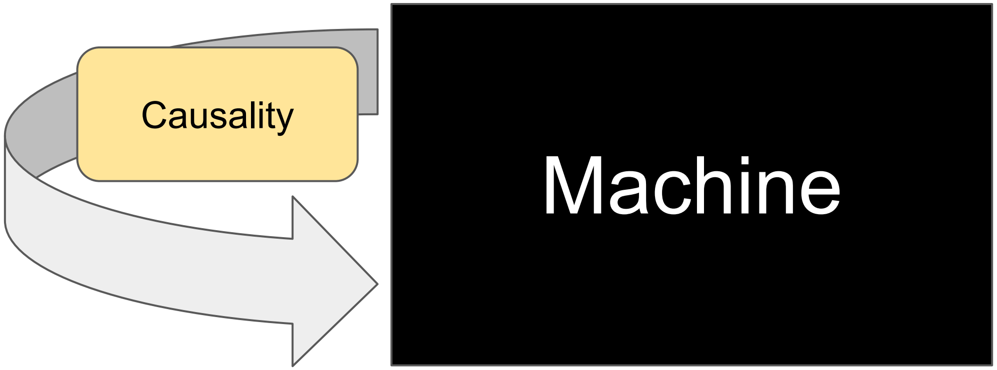
(a) Husky classified as wolf (b) Explanation

Image from Ribeiro et al (2017)

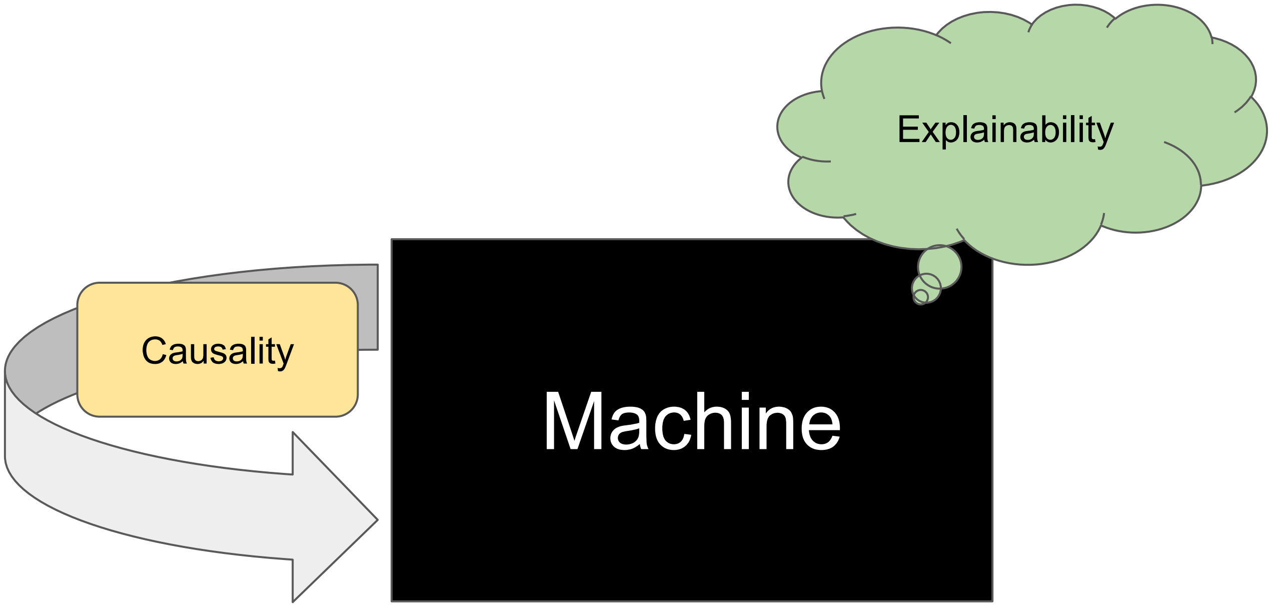
# Robust and verified Machine Learning

Machine

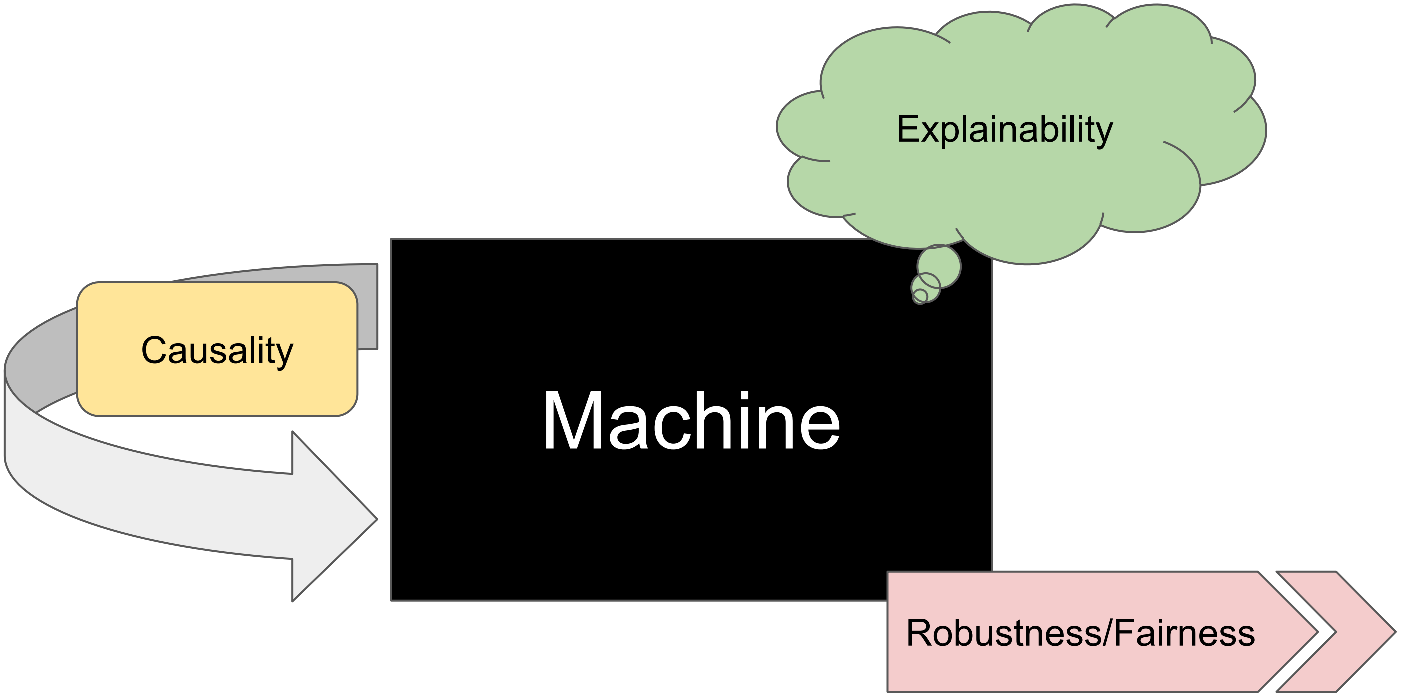
# Robust and verified Machine Learning



# Robust and verified Machine Learning



# Robust and verified Machine Learning



# We need Explainability!

# Different Camps of XAI

## Local interpretability

Why is **this image** labelled as a car?

Image from Lapuschkin (2019)

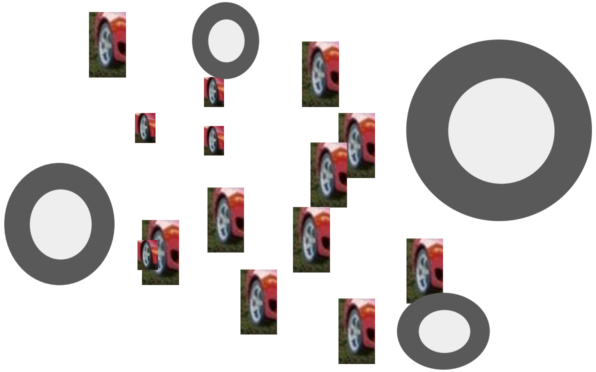


Image from Lapuschkin (2019)

## Global interpretability

What do **all car-labelled images** have in common?

# Different Camps of XAI

## Model-agnostic interpretability

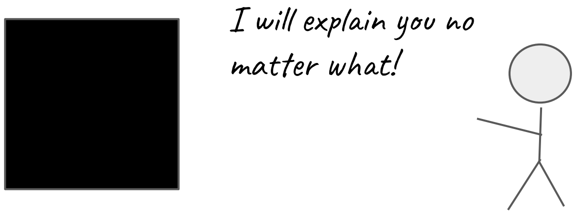
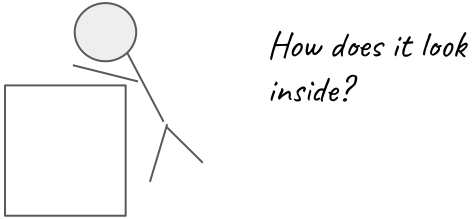



Image from https://subtleyoga.com/why-now-perhaps-more-than-ever-is-a-good-time-to-have-a-magic-word/wingardium-leviosa/

Image from https://saferoad.org/best-tool-chests-reviews/

## Model-specific interpretability




from her MLSS22 talk

# Different Camps of XAI

## Intrinsic interpretability

```
If age > 25 then predict favorite sport = tennis
```

https://corels.eecs.harvard.edu/corels/whatarerulelists.html

Image from https://users.cs.duke.edu/~cynthia/

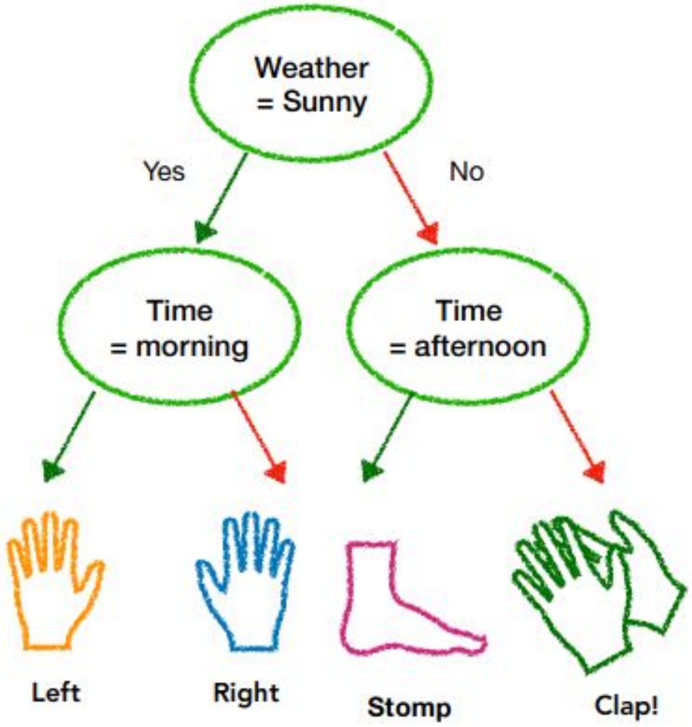
Image from https://beenkim.github.io/

## Post-hoc interpretability

"Probability distortion is that people generally do not look at the value of probability uniformly between 0 and 1. Lower probability is said to be over-weighted while medium to high probability is under-weighted" - Kahneman

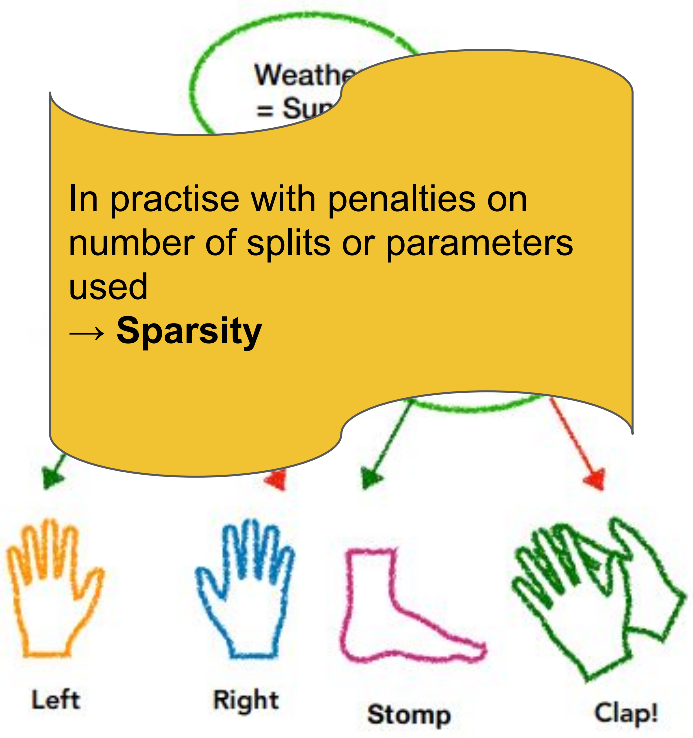
from her MLSS22 talk

# Decision trees / Rule Lists are explainable!




https://beenkim.github.io/slides/DLSS2018Vector_Been.pdf

# Decision trees / Rule Lists are explainable!




In practise with penalties on number of splits or parameters used
→ **Sparsity**

https://beenkim.github.io/slides/DLSS2018Vector_Been.pdf

# Risk-Calibrated Supersparse Linear Integer Models

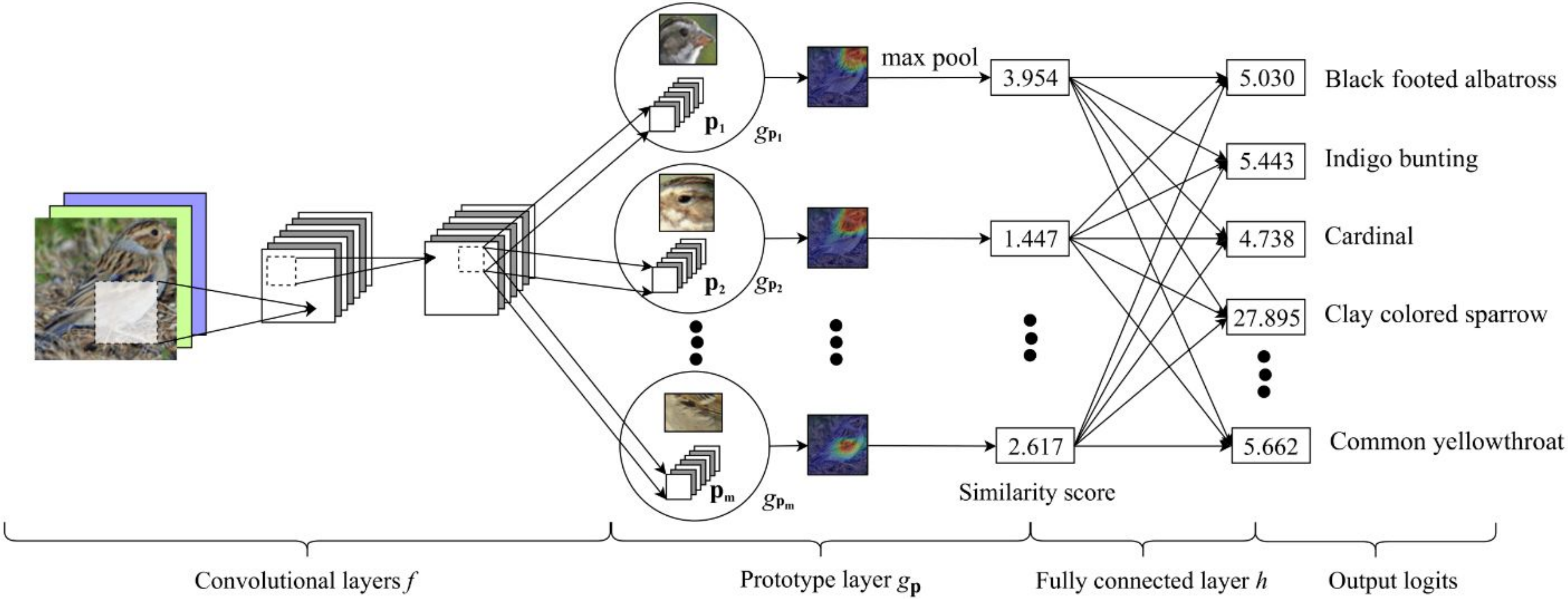
$$\min_{\lambda \in \mathcal{L}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \mathbf{x}_i \lambda}\right)}_{\text{Logistic Loss}} + \underbrace{C_0 \|\lambda\|_0}_{\text{Model Size}}$$

$$\lambda \in \mathcal{L} \quad \text{means that } \forall j, \;\; \lambda_j \in \underbrace{\{-10, -9, \ldots 0, \ldots 9, 10\}}_{\text{Small Integer Coefficients}}$$

RiskSLIM's Mixed-Integer Nonlinear Program

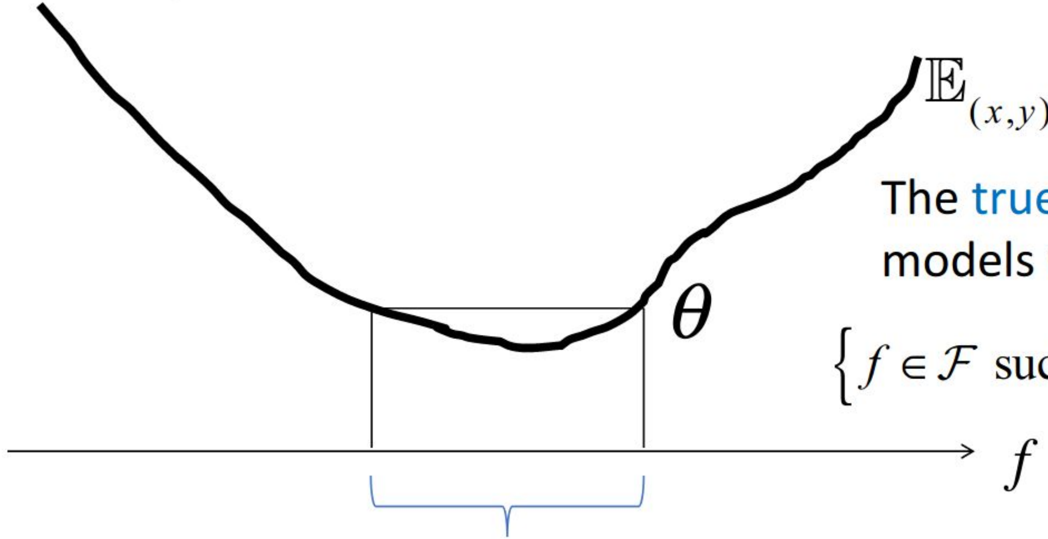
Ustun and Rudin. Learning Optimized Risk Scores from Large-Scale Datasets. KDD 2017, JMLR 2019 (accepted)
Code: https://github.com/ustunb/risk-slim

# This looks like that (Chen, 2019)



# This looks like that (Chen, 2019)

# True Rashomon Sets




$$\mathbb{E}_{(x,y)}\text{loss}\big(f(x),y\big)$$

The true Rashomon set is the set of models with low true loss.

$$\left\{f \in \mathcal{F} \text{ such that } \mathbb{E}_{(x,y)}\text{loss}(f(x),y) \le \theta\right\}$$

If the true Rashomon set is large, a simple-yet-accurate model is likely to exist.

https://users.cs.duke.edu/~cynthia/docs/KDDRashomonForPrint.pdf

# Rashomon Ratio

$$\text{Rashomon Ratio }(\mathcal{F}_2,\theta) := \frac{\#\left\{ f_2 \in \mathcal{F}_2 \text{ such that } L(f_2) \leq \theta \right\}}{|\mathcal{F}_2|}$$

https://users.cs.duke.edu/~cynthia/docs/KDDRashomonForPrint.pdf

# You can bound the loss in performance of simple models!

**Theorem.** For any $\epsilon > 0$, with probability at least $(1-\epsilon)p$, with respect to the random draw of functions from $\mathcal{F}_2$ to form $\mathcal{F}_1$, and with respect to the random draw of iid data:

$$\left|L\left(f_2^*\right) - \hat{L}\left(\hat{f}_1\right)\right| \leq \theta + 2b\sqrt{\frac{\log\left|\mathcal{F}_1\right| + \log 2/\epsilon}{2n}}, \text{ where } p = 1 - \frac{\binom{\left(1-\text{Rashomon Ratio}(\mathcal{F}_2,\theta)\right)|\mathcal{F}_2|}{|\mathcal{F}_1|}}{\binom{|\mathcal{F}_2|}{|\mathcal{F}_1|}}.$$

https://users.cs.duke.edu/~cynthia/docs/KDDRashomonForPrint.pdf

# You can bound the loss in performance of simple models!

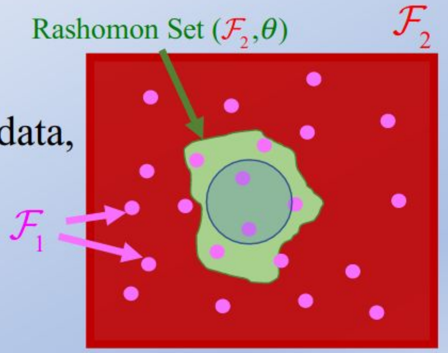
**Theorem**

For a $\mathcal{K}$-Lipschitz loss $l$ bounded by $b$, hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, $\mathcal{F}_1 \subset \mathcal{F}_2$, if for each $f_2 \in$ Rashomon set$(\mathcal{F}_2, \theta)$ there exists a model $f_1 \in \mathcal{F}_1$ such that $\| f_2 - f_1 \|_p \leq \delta$, and if the Rashomon set is large, in that it contains an $\ell_p$ ball of size at least $\delta$, then there exists $\bar{f}_1 \in$ Rashomon set$(\mathcal{F}_2, \theta)$ such that for a fixed parameter $\epsilon \in (0,1)$:

1. $\bar{f}_1$ is from simpler class $\mathcal{F}_1$.
2. With prob $\geq 1$-$\epsilon$ w.r.t. the random draw of training data,

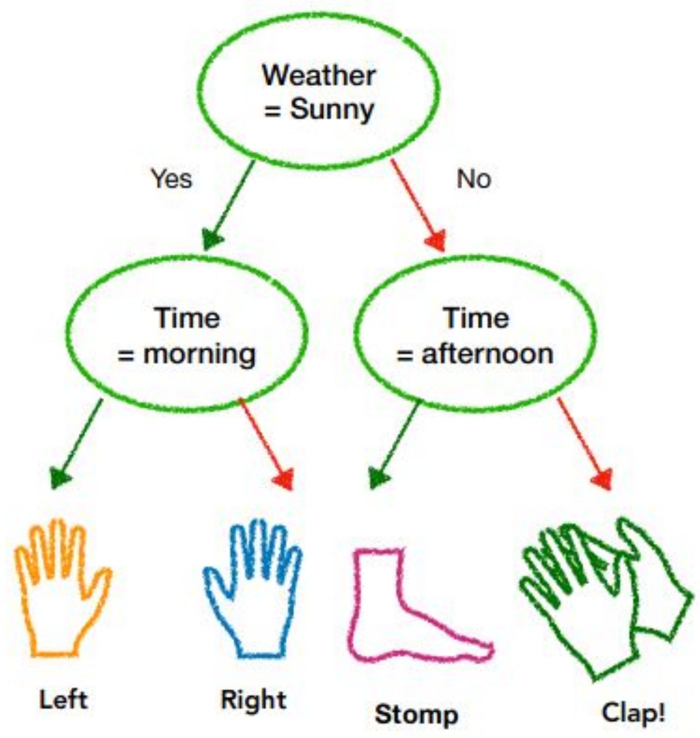
$$\left| L(\bar{f}_1) - \hat{L}(\bar{f}_1) \right| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}},$$

where $R_n(\mathcal{F}_1)$ is Rademacher complexity.



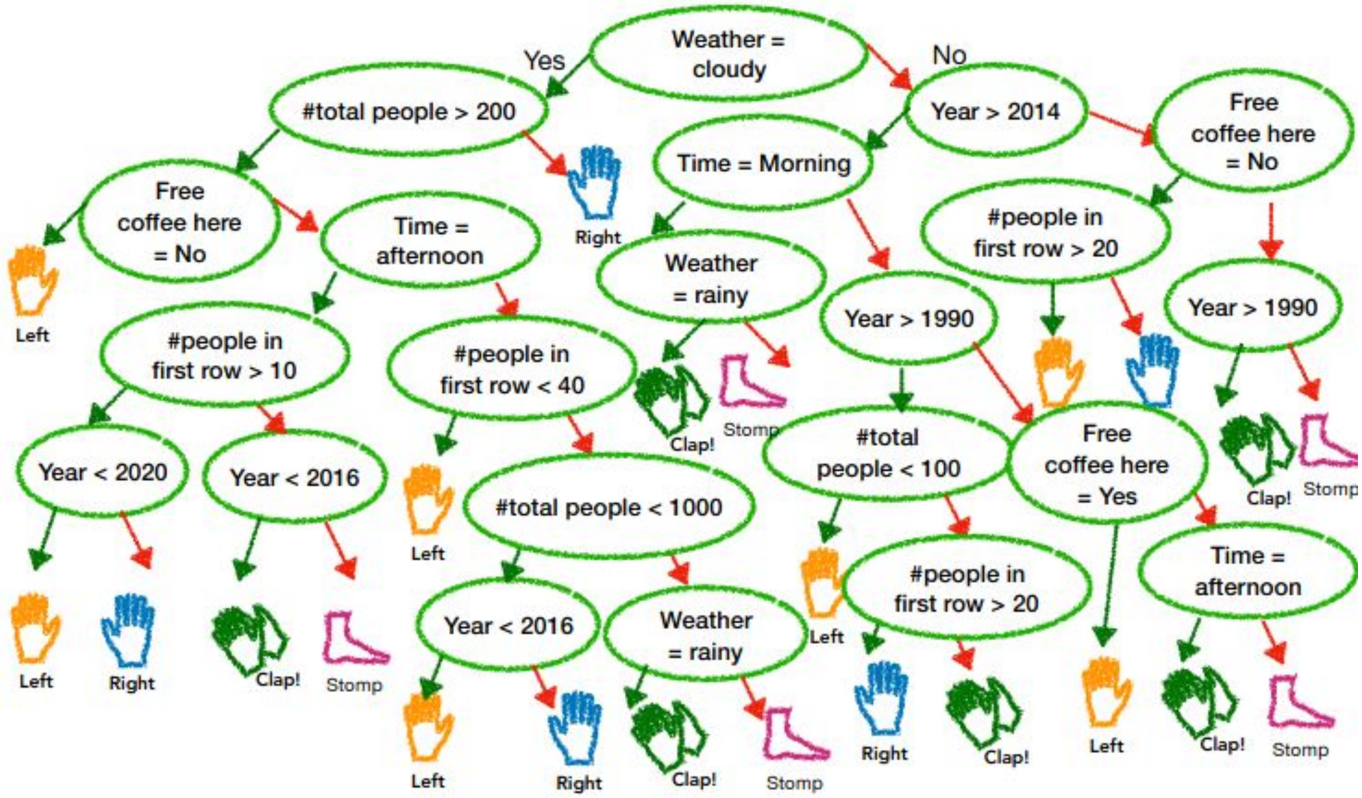

https://users.cs.duke.edu/~cynthia/docs/KDDRashomonForPrint.pdf

# Decision trees are explainable!




https://beenkim.github.io/slides/DLSS2018Vector_Been.pdf

# Decision trees are explainable?




https://beenkim.github.io/slides/DLSS2018Vector_Been.pdf

# Counterfactual Explanations (aka Algorithmic Recourse)

# Counterfactual Explanations (aka Algorithmic Recourse)

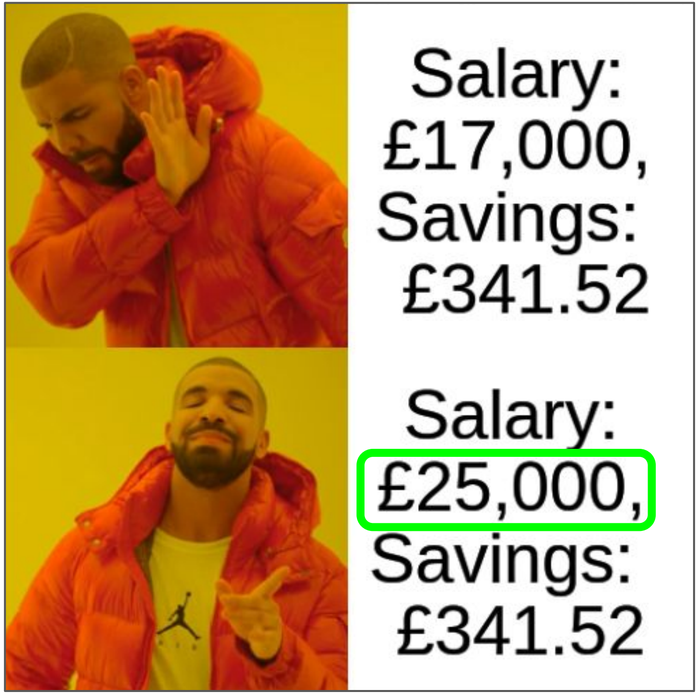

LOAN DENIED

# Counterfactual Explanations (aka Algorithmic Recourse)


**LOAN DENIED**

*How would **the numbers** need to change the least to flip the decision?*

# Counterfactual Explanations (aka Algorithmic Recourse)




**LOAN DENIED**

*How would **the numbers** need to change the least to flip the decision?*

# Counterfactual Explanations (aka Algorithmic Recourse)

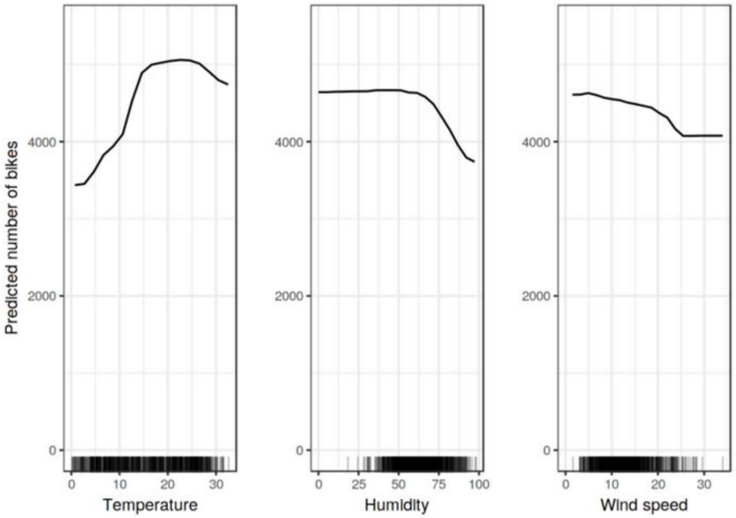

**LOAN DENIED**

*How would I need to* ***interfere*** *the least to flip the decision?*

**With causality**

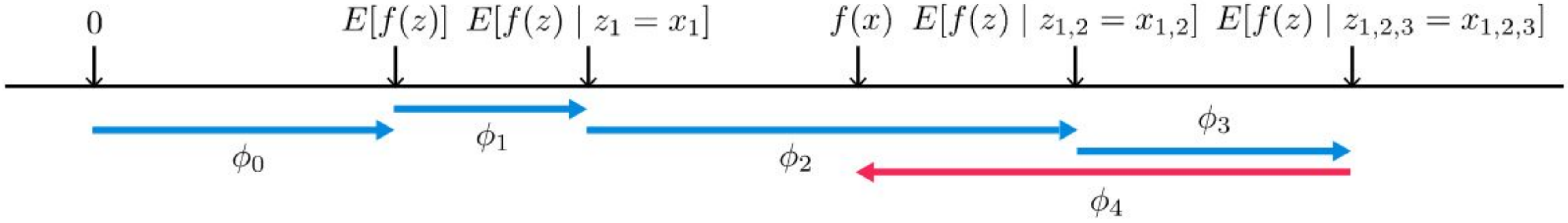
# Partial Dependence Plots

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)d\mathbb{P}(x_C)$$

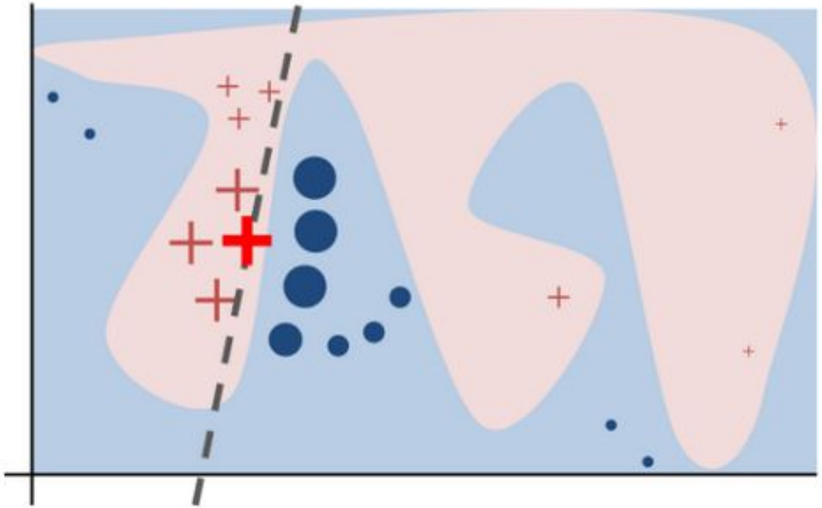



https://towardsdatascience.com/an-overview-of-model-explainability-in-modern-machine-learning-fc0f22c8c29a

# Shapley Values (Lundberg and Lee, 2017)



# LIME (Ribeiro et al., 2016)



Ribeiro et al. (2017)

- black box classification model *f*: pink and blue areas
- instance being explained: bold red cross
- instances sampled locally and weighted by their proximity: red crosses, and blue circles
- locally faithful explanation *g*: dashed line

# LIME (Ribeiro et al., 2016)

black box

neighbourhood kernel around x

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \underbrace{\mathcal{L}(f, g, \pi_x)}_{\text{fit } g \text{ to } f \text{ in a small neighbourhood around } x} + \underbrace{\Omega(g)}_{\text{ensure } g \text{ is simple}}$$

explanation model

# Problems with Tangent Line Approximations

- If a linear fit is good enough, why not just have a locally linear black box? (Rudin, 2019)
- Consider the black box

$$f(x) = \mathbb{I}(x_1 > 0)2x_2^2 - \mathbb{I}(x_1 \leq 0)x_2^2$$

# Problems with Tangent Line Approximations

- If a linear fit is good enough, why not just have a locally linear black box? (Rudin, 2019)
- Consider the black box

$$f(x) = \mathbb{I}(x_1 > 0)2x_2^2 - \mathbb{I}(x_1 \leq 0)x_2^2$$

for x1=-0.001, the feature attribution of Feature-2 is **negative**

# Can we just take the gradients?

No!

# Can we just take the gradients?

No!

Look at

$f(x)= \mathrm{ReLU}(1-x)$

we want to explain $x^*=2$

# Can we just take the gradients?

No!

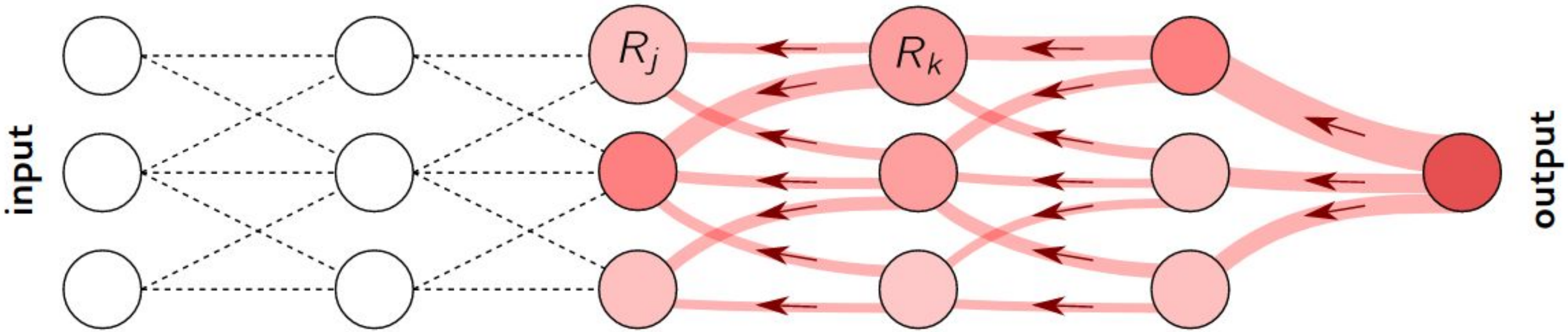
Look at

$f(x) = \mathrm{ReLU}(1-x)$

we want to explain $x^*=2$

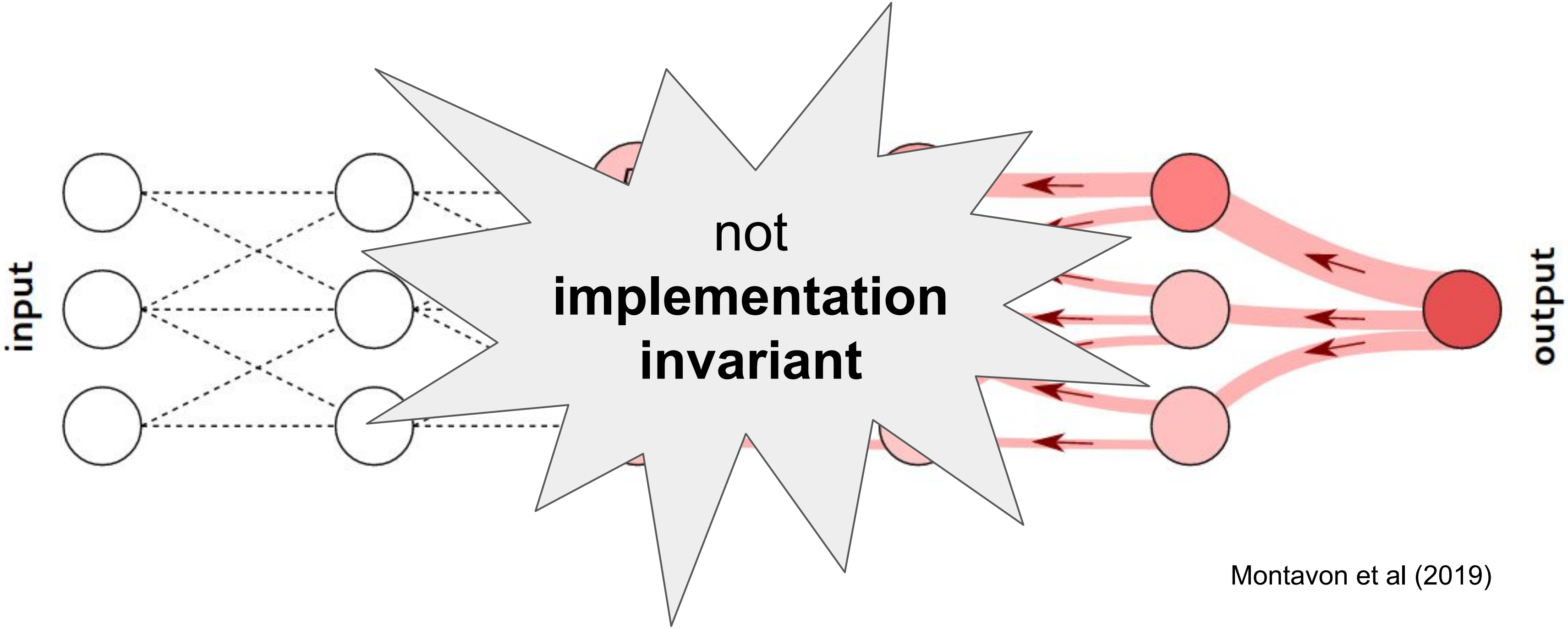
$\Rightarrow$ not **sensitive** wrt $x=0$

# Layer-wise relevance propagation (Bach et al, 2016)



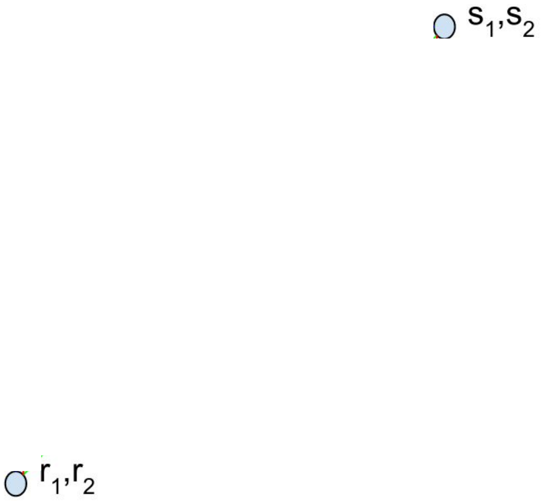

Montavon et al (2019)

# Layer-wise relevance propagation (Bach et al, 2016)



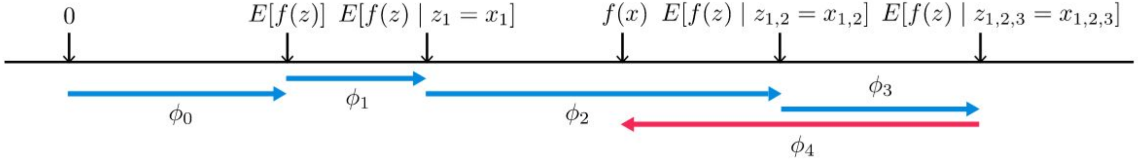

Montavon et al (2019)

# Integrated Gradients (Sundararajan et al, 2017)



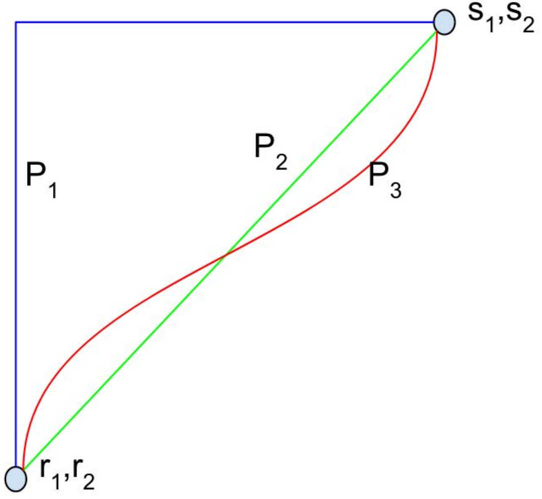
# Integrated Gradients (Sundararajan et al, 2017)

$s_1,s_2$

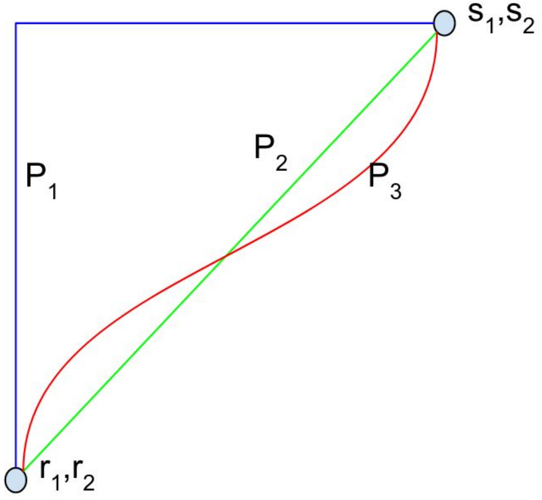



$r_1,r_2$

# Integrated Gradients (Sundararajan et al, 2017)

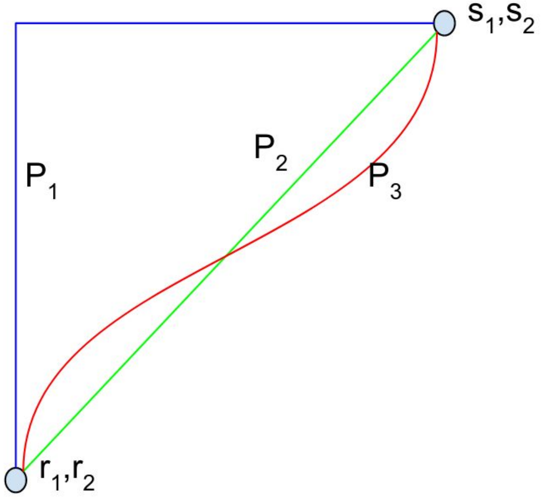


# Integrated Gradients (Sundararajan et al, 2017)




$$\mathsf{PathIntegratedGrads}_i^{\gamma}(x) ::= \int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \; \frac{\partial \gamma_i(\alpha)}{\partial \alpha} \; d\alpha$$

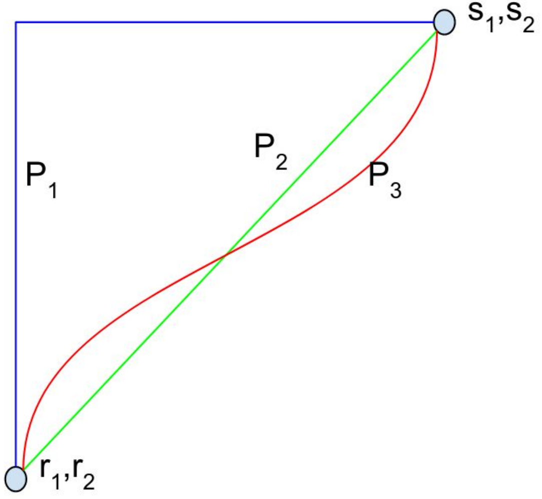
# Integrated Gradients (Sundararajan et al, 2017)




$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

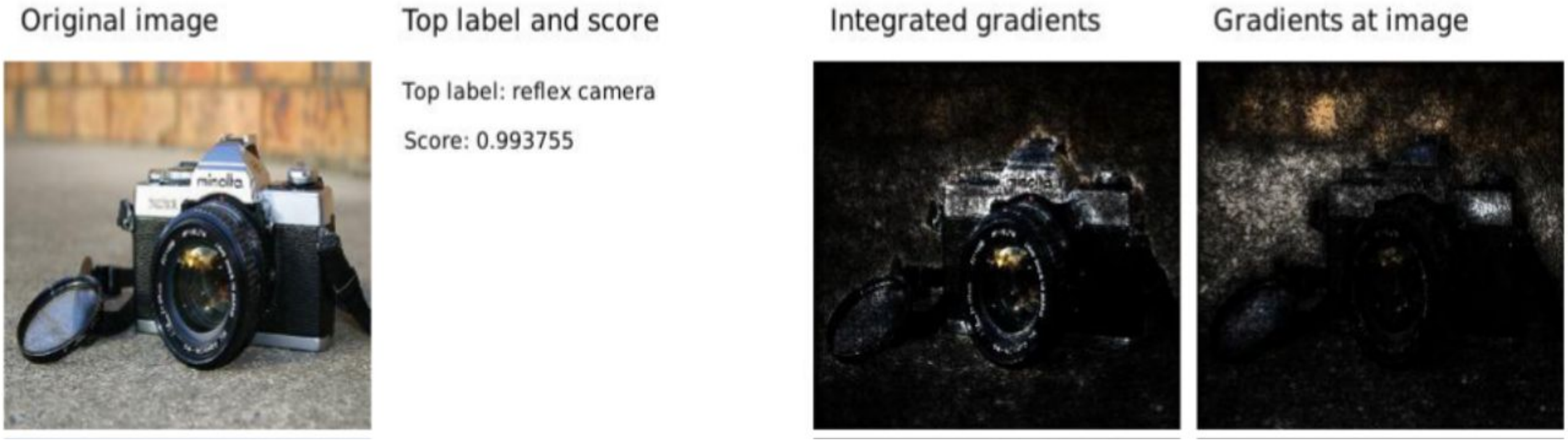
# Integrated Gradients (Sundararajan et al, 2017)

- Sensitivity
- Implementation Invariance
- Completeness
- Linearity

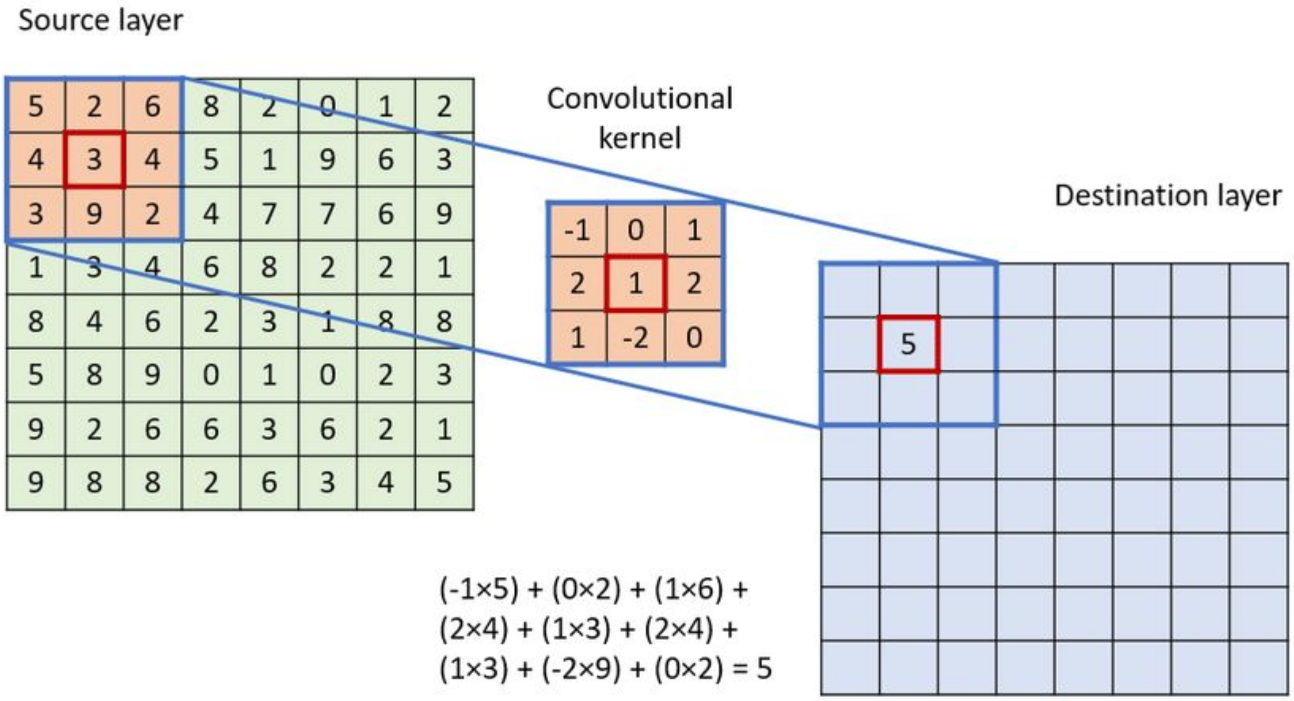



$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

# Integrated Gradients (Sundararajan et al, 2017)

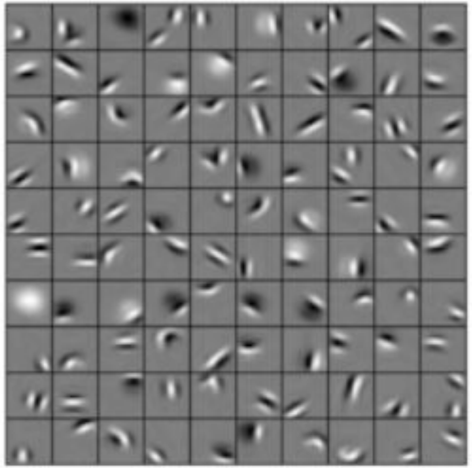


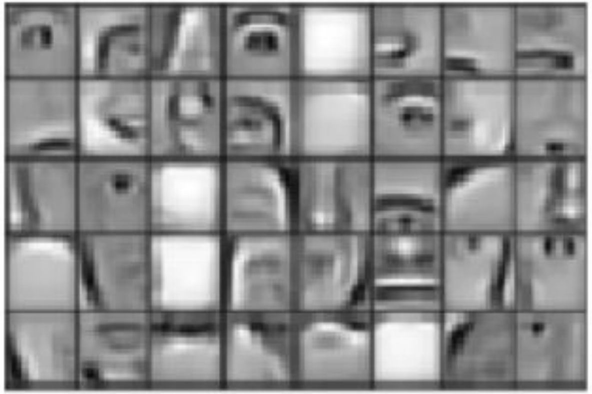
# Convolutional layers learn interpretable concepts




Best Practice Guide - Deep Learning - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Schematic-illustration-of-a-convolutional-operation-The-convolutional-kernel-shifts-over_fig2_332190148 [accessed 24 Mar, 2022]

# Convolutional layers learn interpretable concepts





https://developer.nvidia.com/blog/deep-learning-nutshell-core-concepts/

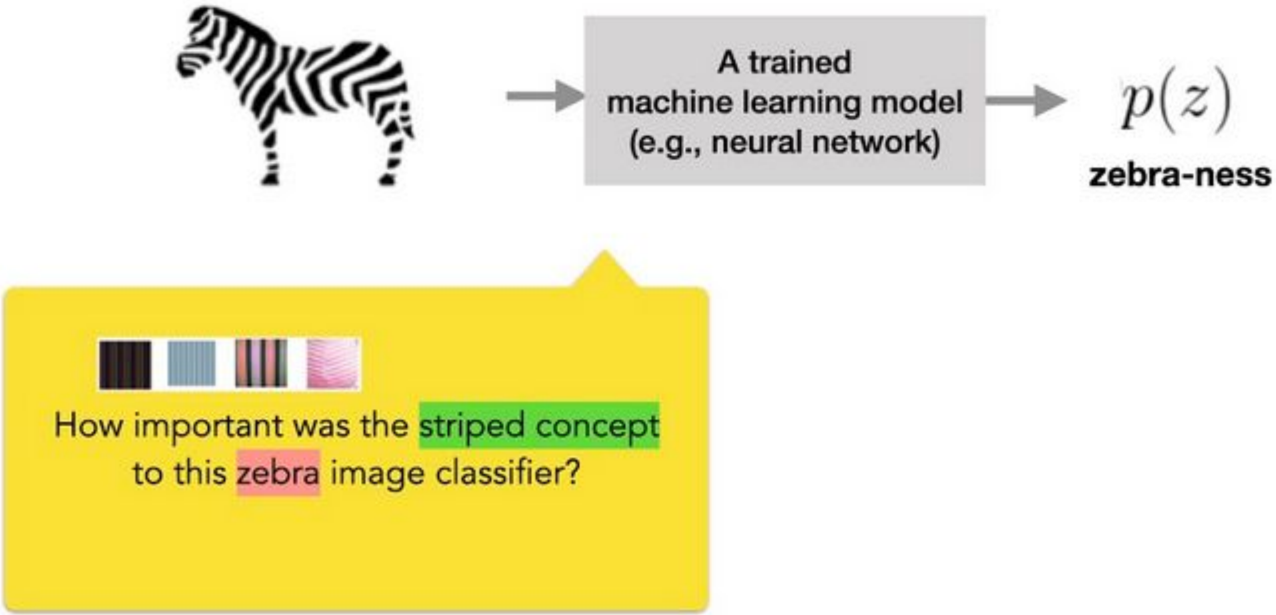
# Grad-CAM

$$L^c_{Grad-CAM} \in \mathbb{R}^{u \times v} = \underbrace{ReLU}_{\text{Pick positive values}} \left( \sum_k \alpha_k^c A^k \right)$$

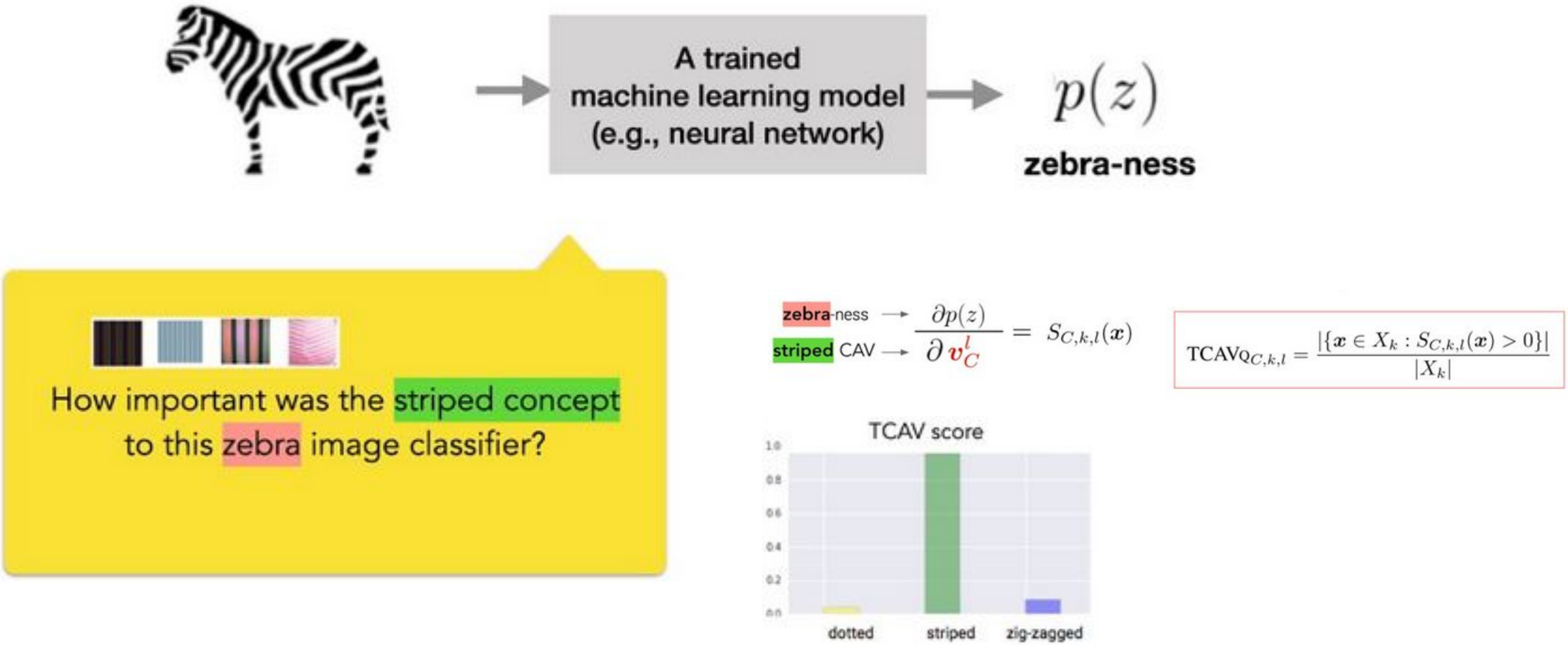
$$\alpha_k^c = \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\delta y^c}{\delta A_{ij}^k}}_{\text{gradients via backprop}}$$

# Attention Visualisation

# TCAV: Testing with Concept Activation Vectors (Kim et al, 2018)

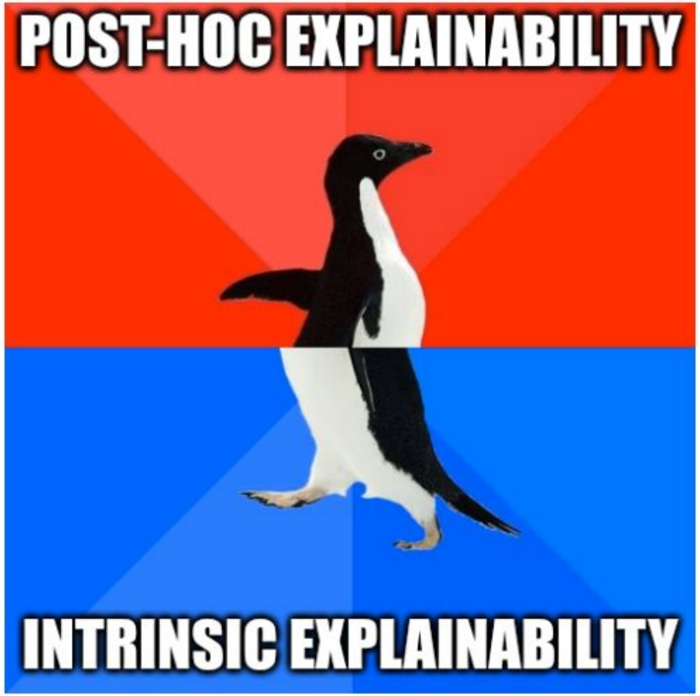


# TCAV: Testing with Concept Activation Vectors (Kim et al, 2018)



# Surrogate Modelling?!

- More Fidelity
- No “Double Trouble”




- Explainable models not necessarily explainable
- Loss in prediction accuracy
- Black box access

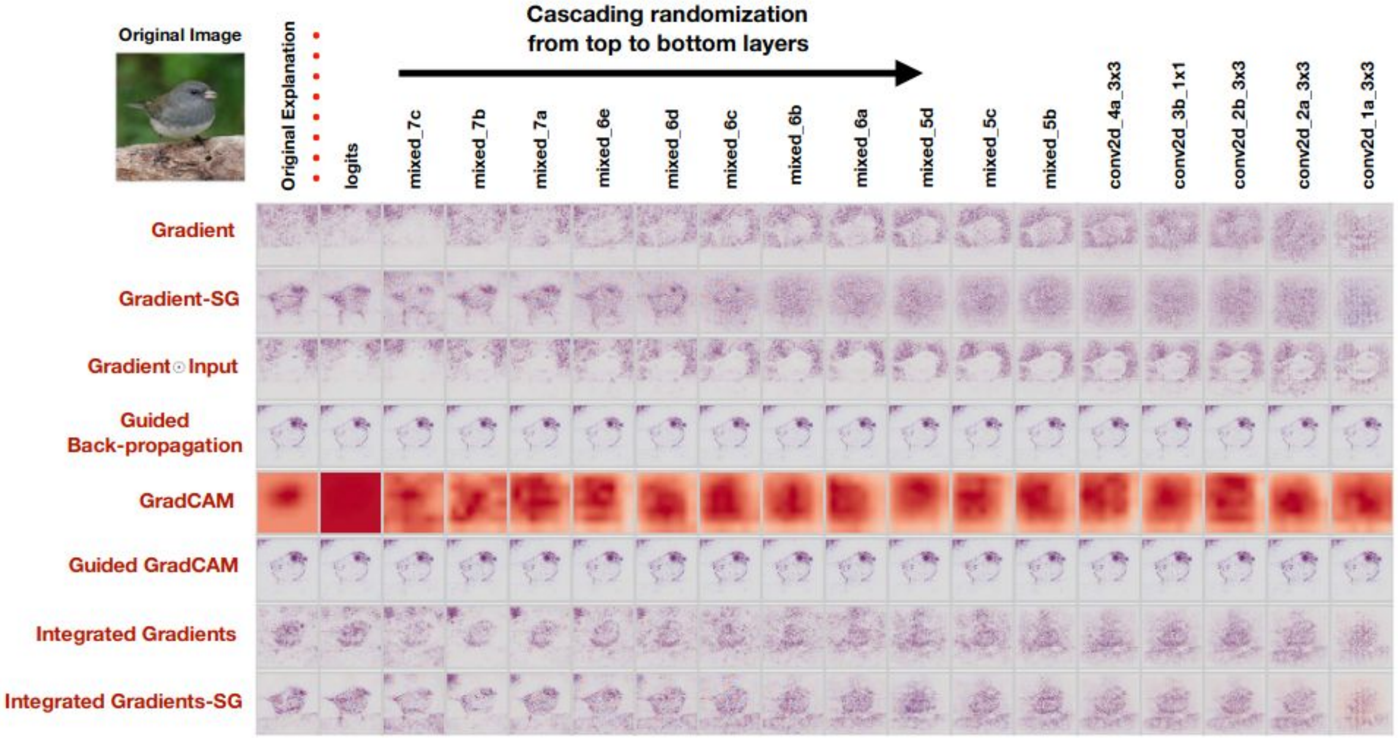
# Saliency maps are informative and elicit trust!


Image from Rudin (2019)

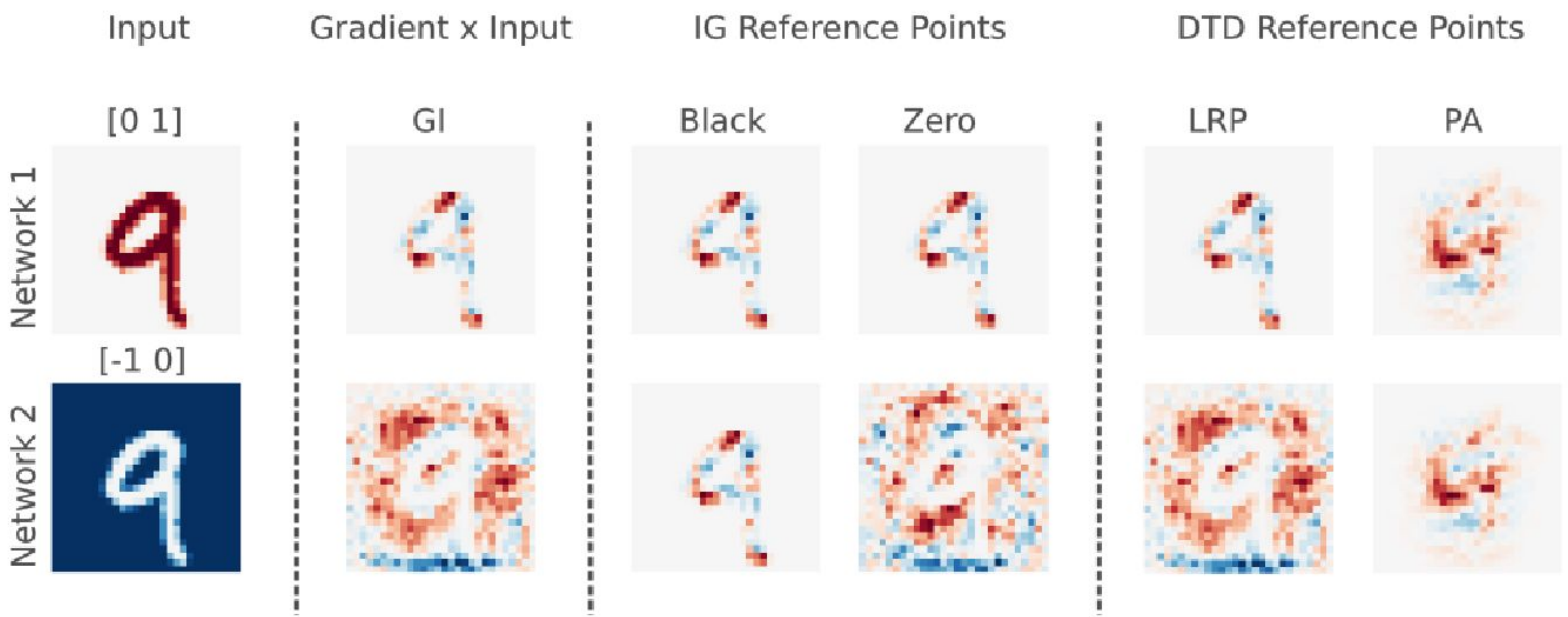
# Saliency maps are informative and elicit trust?


Image from Rudin (2019)

# Sanity checks of saliency maps (Adebayo, 2018)



# Saliency methods are unreliable (Kindermans et al., 2019)



# Fooling Explainability Tools (Slack et al., 2020)

**evil** model that does not hire black applicants

(*no one can know!*)

# Fooling Explainability Tools (Slack et al., 2020)

Imagine all **black applicants** live in **zip code 15235**

**evil** model that does not hire black applicants

(*no one can know*!)

# Fooling Explainability Tools (Slack et al., 2020)

**evil** model that does not hire black applicants

(*no one can know!*)

Imagine all **black applicants** live in **zip code 15235**

Strategy

⛔Reject black applicant from Zip code 1523

✅Accept black applicant from any other Zip code

# Fooling Explainability Tools (Slack et al., 2020)

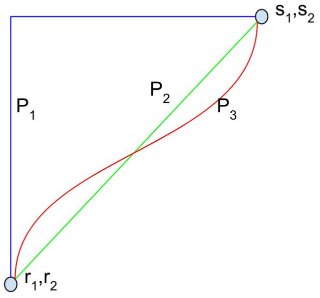
**evil** model that does not hire black applicants

(*no one can know!*)

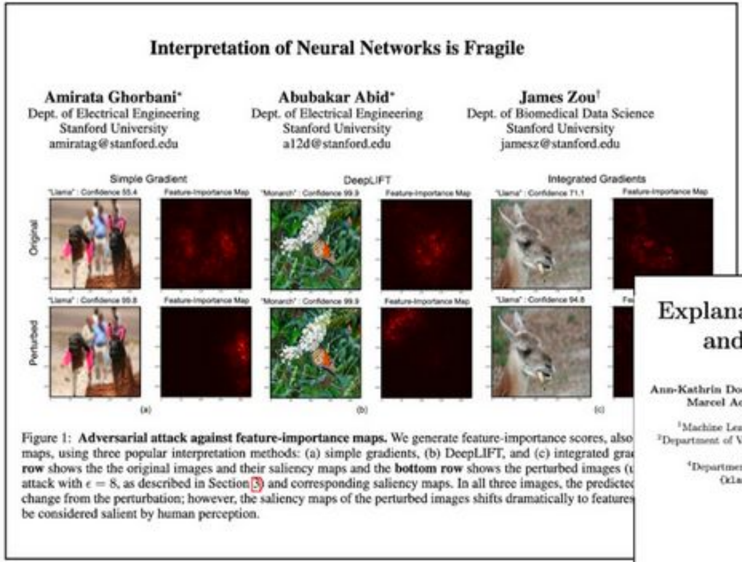
Imagine all **black applicants** live in **zip code 15235**

Strategy

⛔Reject black applicant from Zip code 1523
✅Accept black applicant from any other Zip code



# Adversarial Attacks on Explanations


Interpretation of Neural Networks is Fragile

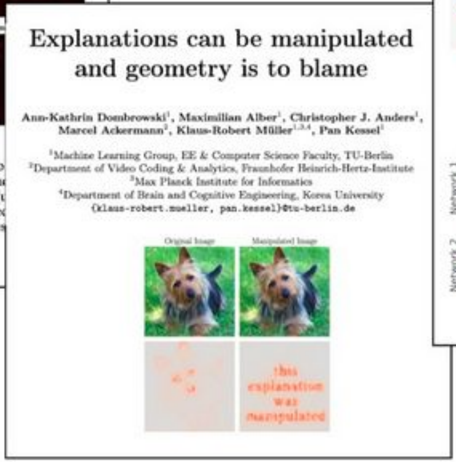
Amirata Ghorbani*
Dept. of Electrical Engineering
Stanford University
amiratag@stanford.edu

Abubakar Abid*
Dept. of Electrical Engineering
Stanford University
a12d@stanford.edu

James Zou†
Dept. of Biomedical Data Science
Stanford University
jamesz@stanford.edu



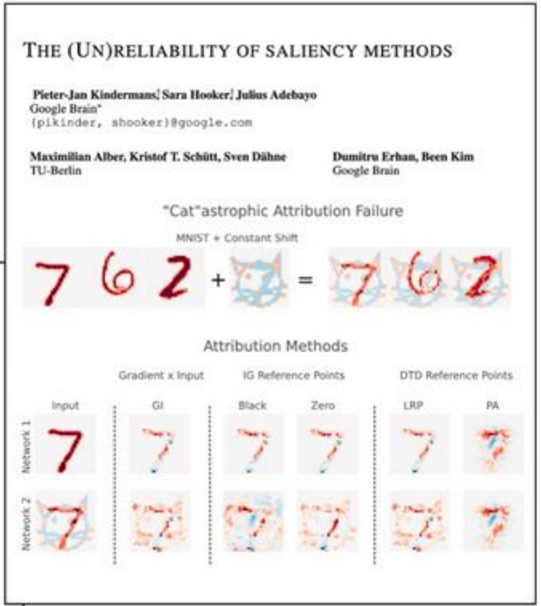
Figure 1: **Adversarial attack against feature-importance maps.** We generate feature-importance scores, also maps, using three popular interpretation methods: (a) simple gradients, (b) DeepLIFT, and (c) integrated gra **row** shows the the original images and their saliency maps and the **bottom row** shows the perturbed images (u attack with $\epsilon = 8$, as described in Section 3) and corresponding saliency maps. In all three images, the predicted change from the perturbation; however, the saliency maps of the perturbed images shifts dramatically to features be considered salient by human perception.


Explanations can be manipulated and geometry is to blame

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

[1]Machine Learning Group, EE & Computer Science Faculty, TU-Berlin
[2]Department of Video Coding & Analytics, Fraunhofer Heinrich-Hertz-Institute
[3]Max Planck Institute for Informatics
[4]Department of Brain and Cognitive Engineering, Korea University
{klaus-robert.mueller, pan.kessel}@tu-berlin.de




THE (UN)RELIABILITY OF SALIENCY METHODS

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo
Google Brain*
{pikinder, shooker}@google.com

Maximilian Alber, Kristof T. Schütt, Sven Dähne
TU-Berlin

Dumitru Erhan, Been Kim
Google Brain

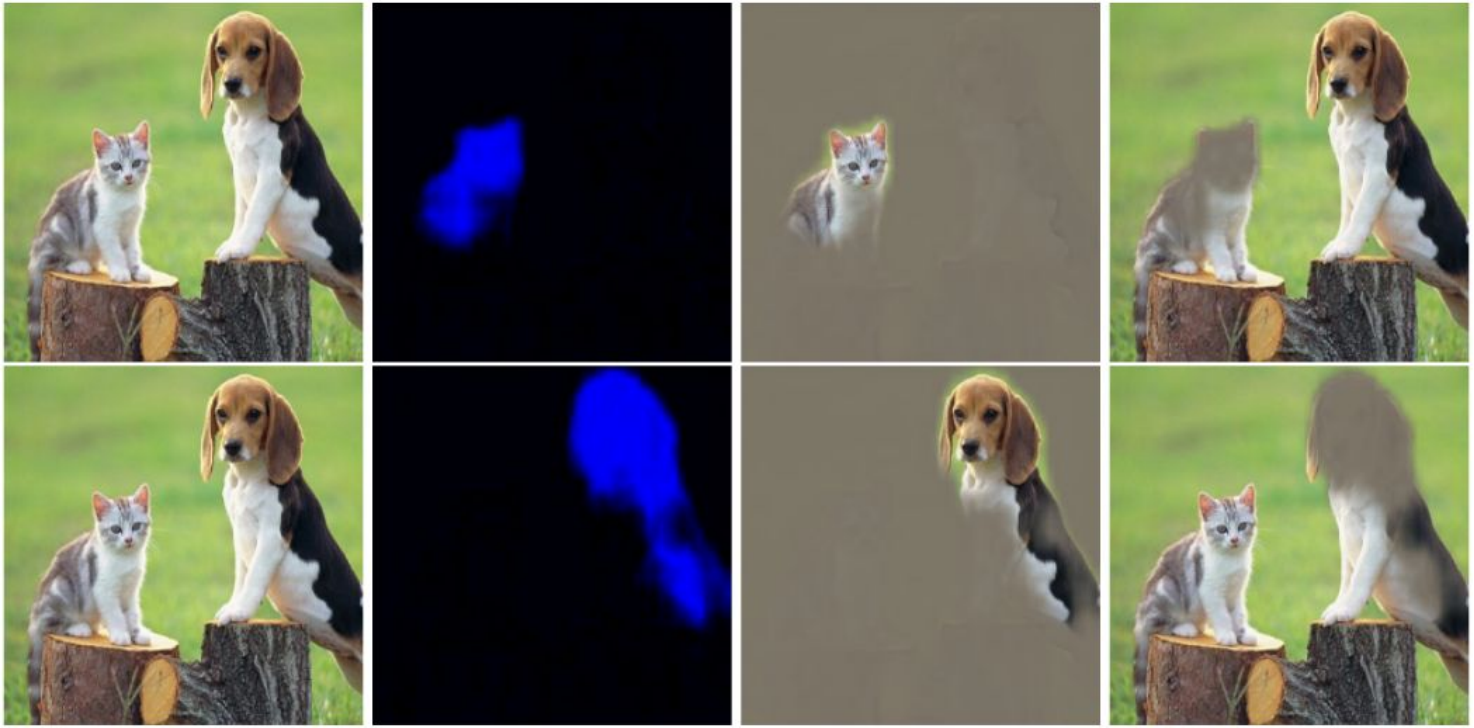


from Been Kim's MLSS 2021 talk

# But if XAI doesn't work, what can we do?

# But if XAI doesn’t work, what can we do?

# 1) Evaluate Exclusion and Inclusion Criteria (i.e. Dabkowski and Gal, 2017 or Hooker, 2019)

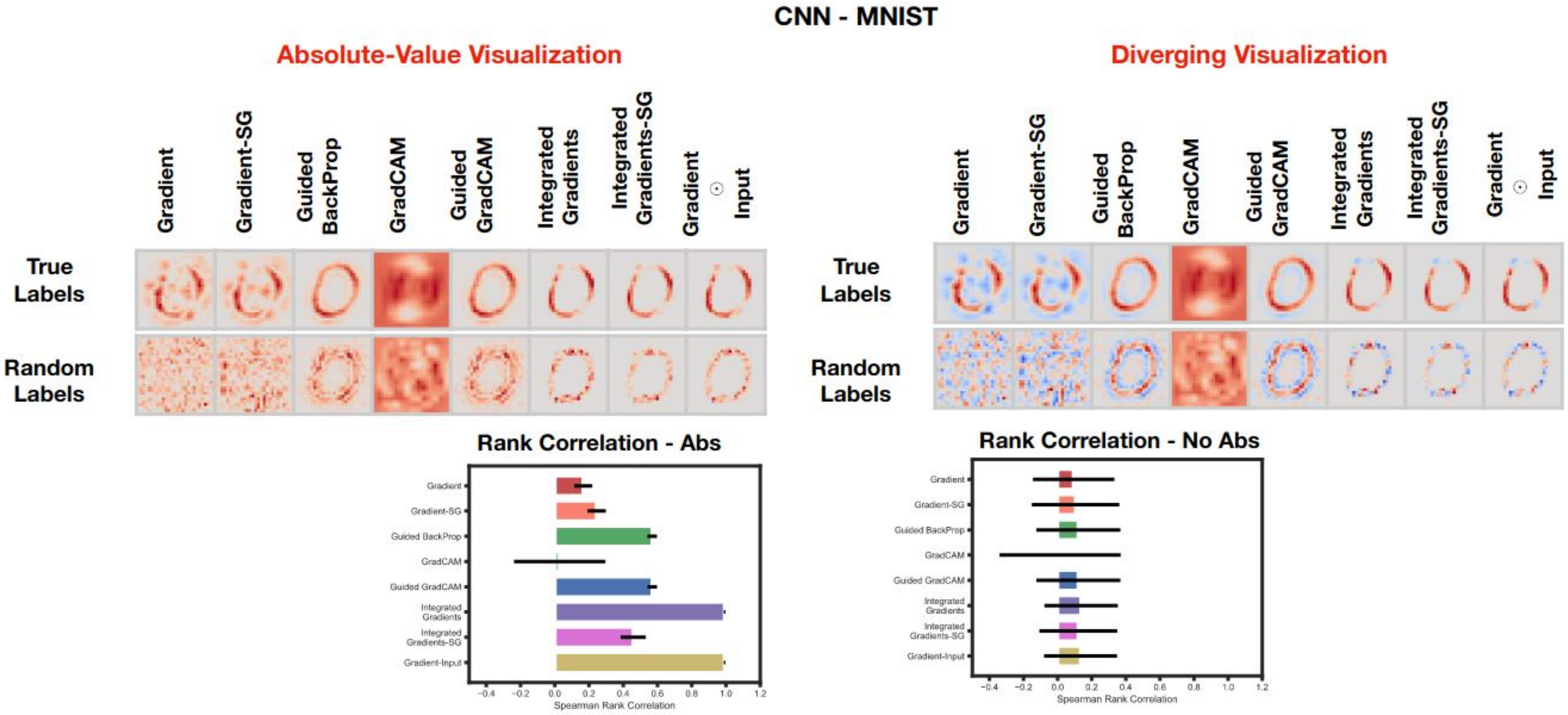


(a) Input Image (b) Generated saliency map (c) Image multiplied by the mask (d) Image multiplied by inverted mask

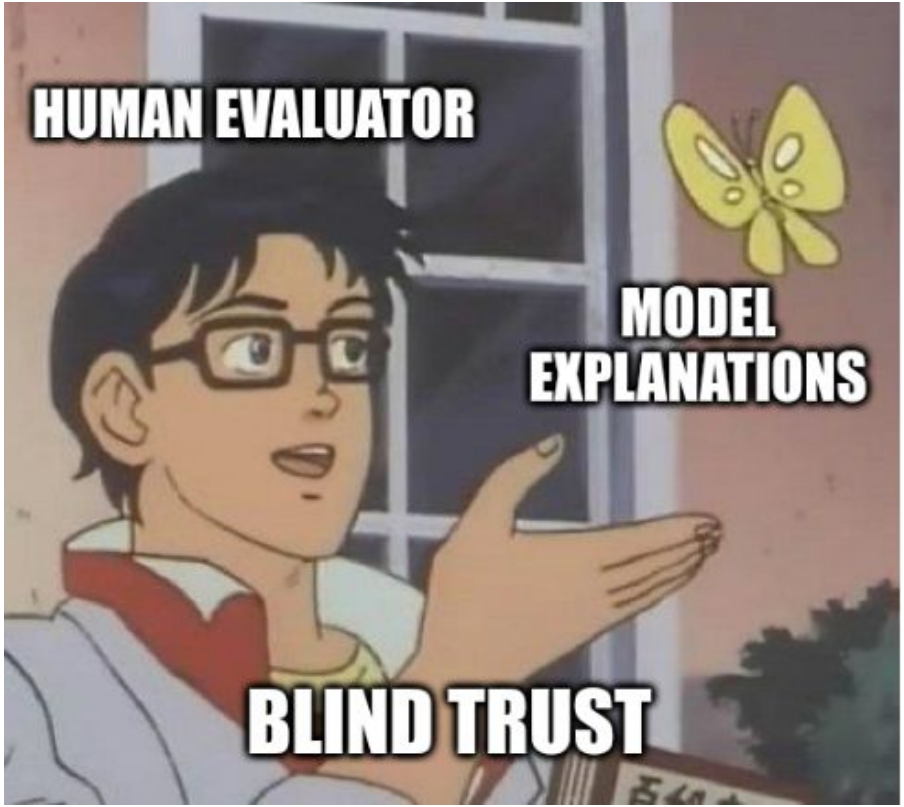
# 2) Handcraft sanity check data sets


https://github.com/google-research-datasets/bam

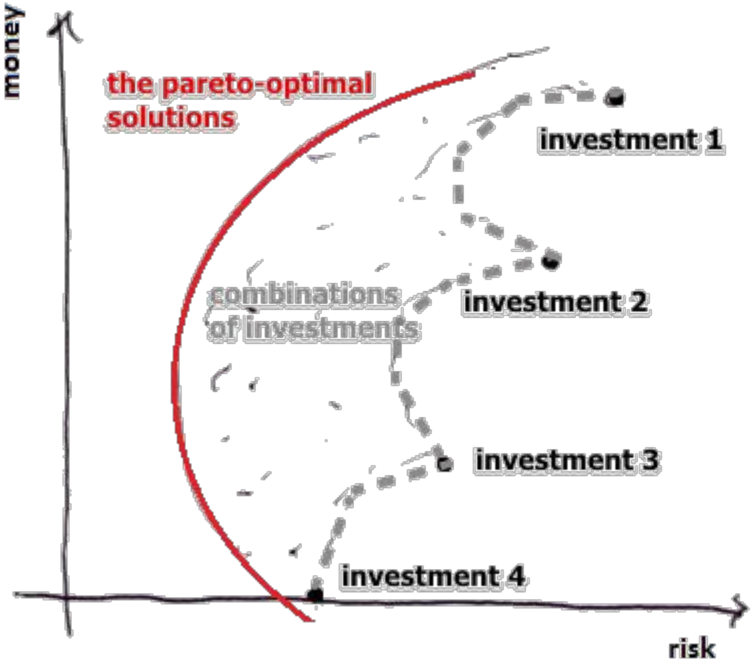
# 3) Look at classes that are not true, False positives, False negatives,... (Adebayo, 2018)



# 4) Human evaluations

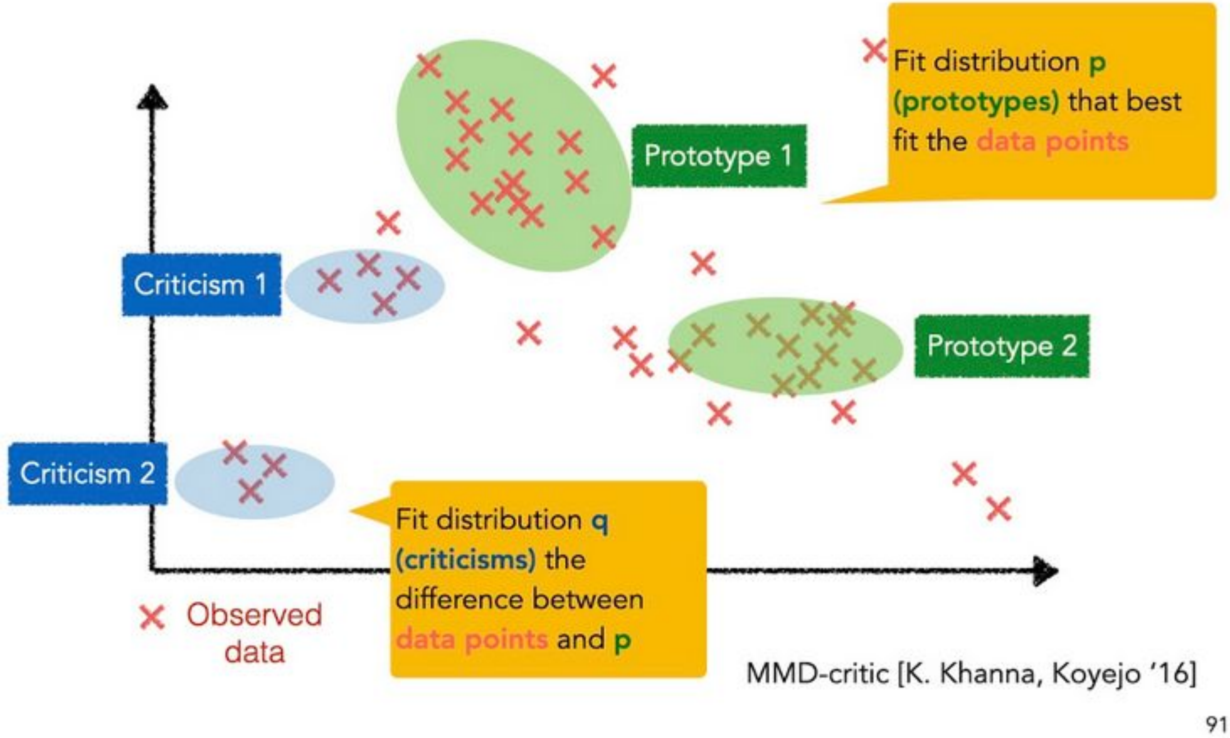


# 5) Risk diversification




http://mlwiki.org/index.php/Portfolio_Management

# 6) Exploratory data analysis

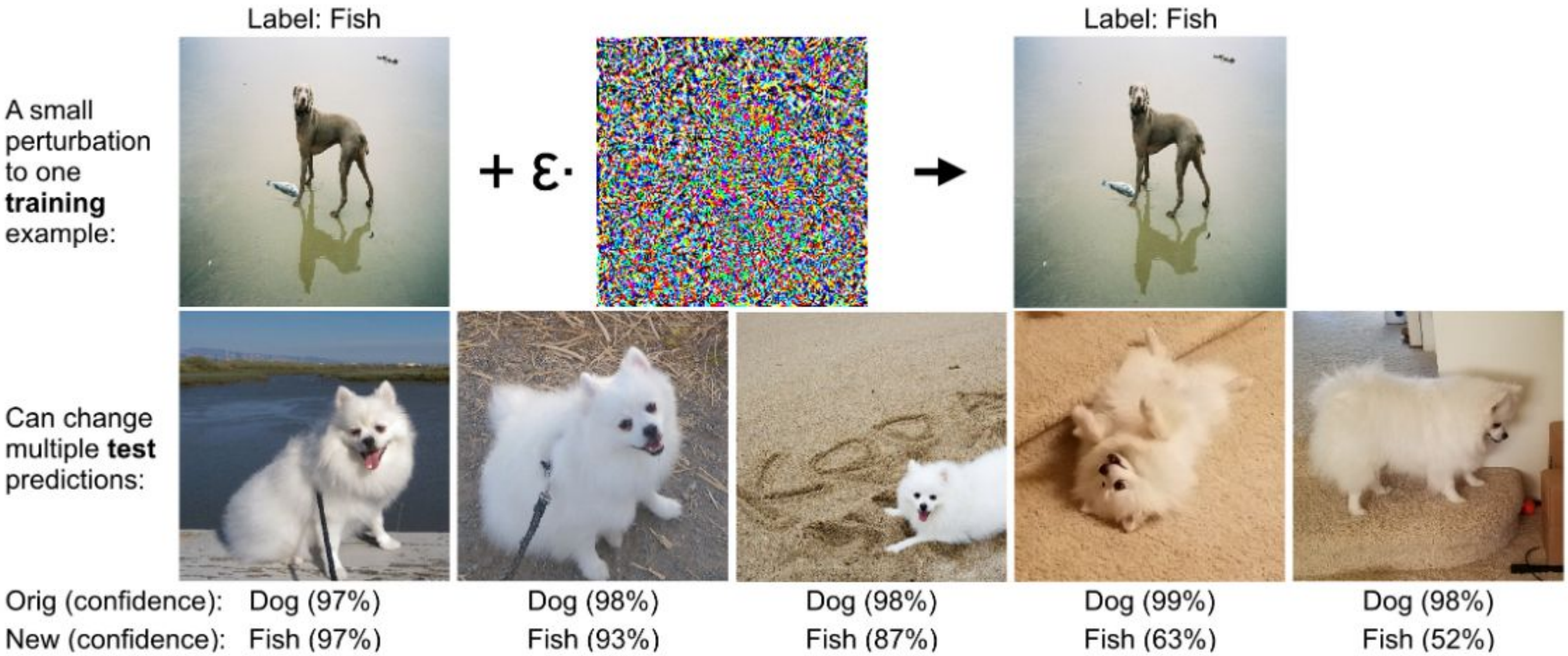



Been Kim's MLSS 2021 talk, https://beenkim.github.io/

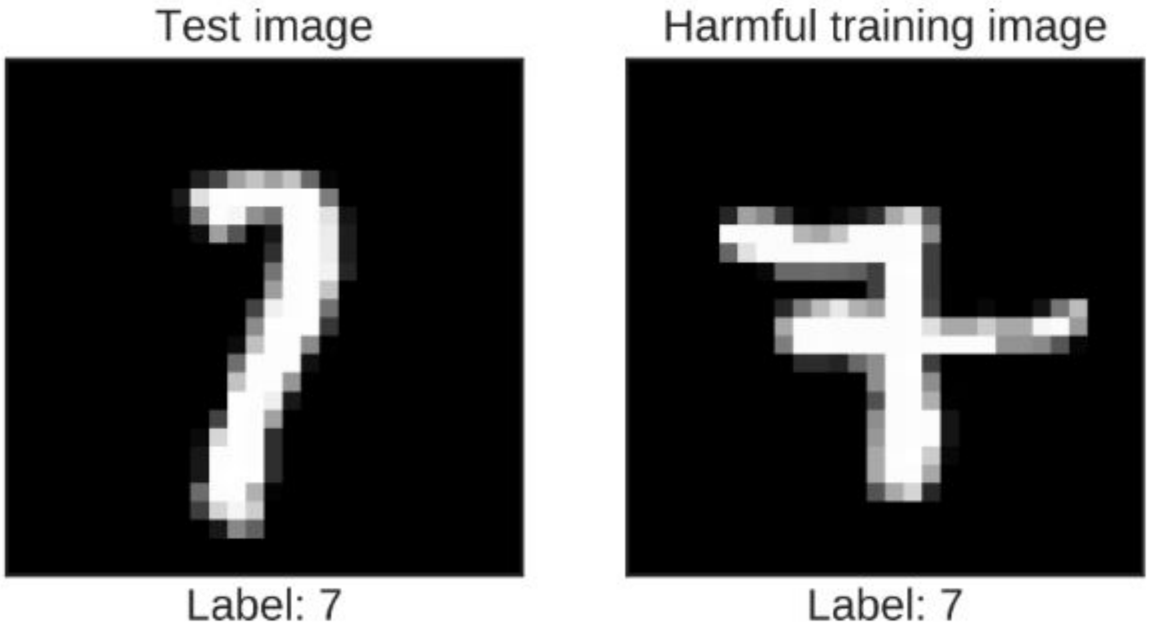
# 6) Exploratory data analysis


Been Kim's MLSS 2021 talk, https://beenkim.github.io/

# Influence Functions (Koh and Liang, 2017)



# Influence Functions (Koh and Liang, 2017)



# Influence Functions (Cook and Weisberg, 1982)

What happens if I upweight observation z by (1+$\epsilon$)

$$z = (x, y)$$

Find new optimal parameter

$$\hat{\theta}_{\epsilon,z} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z, \theta)$$

Influence of upweighting on parameters θ

$$\mathcal{I}_{\text{up,params}}(z) \stackrel{\text{def}}{=} \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta})$$

# Influence Functions (Koh and Liang, 2017)

$$z = (x, y)$$

perturb one training point

$$z_\delta \stackrel{\text{def}}{=} (x + \delta, y)$$

Find new optimal parameter

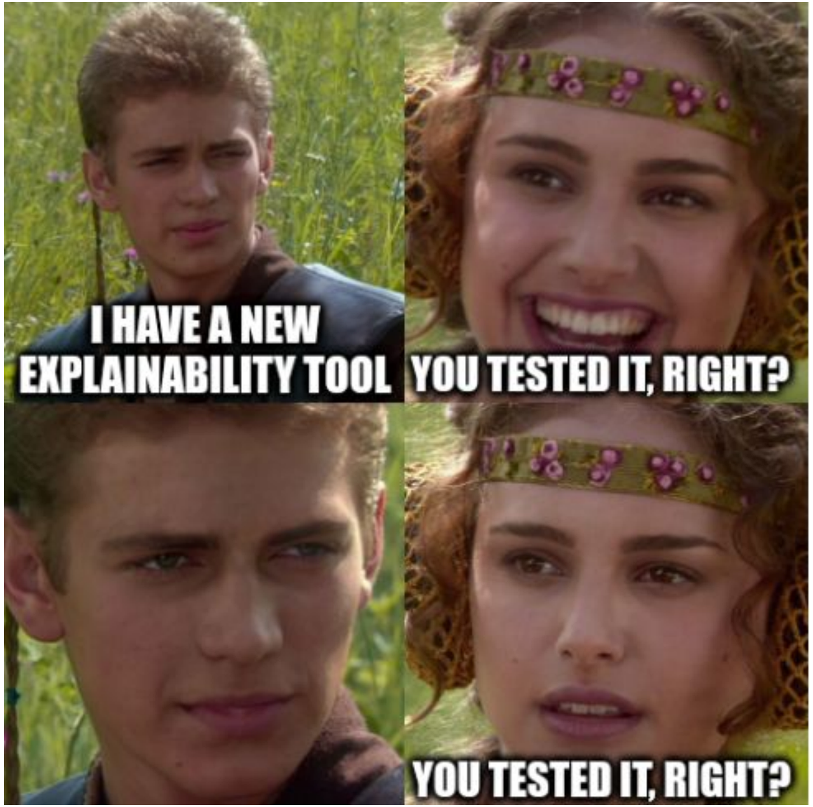
$$\hat{\theta}_{\epsilon, z_\delta, -z} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z_\delta, \theta) - \epsilon L(z, \theta)$$

Influence of perturbing z by δ

$$\begin{aligned} \left.\frac{d\hat{\theta}_{\epsilon, z_\delta, -z}}{d\epsilon}\right|_{\epsilon=0} &= \mathcal{I}_{\text{up,params}}(z_\delta) - \mathcal{I}_{\text{up,params}}(z) \\ &= -H_{\hat{\theta}}^{-1}\left(\nabla_\theta L(z_\delta, \hat{\theta}) - \nabla_\theta L(z, \hat{\theta})\right) \end{aligned}$$

Holds for arbitrary δ

# Take home message



# References


Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020.

Kindermans, Pieter-Jan, et al. "The (un) reliability of saliency methods." *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, Cham, 2019. 267-280.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

Hooker, Sara, et al. "A benchmark for interpretability methods in deep neural networks." *Advances in neural information processing systems* 32 (2019).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1.5 (2019): 206-215.

Adebayo, Julius, et al. "Sanity checks for saliency maps." *Advances in neural information processing systems* 31 (2018).

Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." *Nature communications* 10.1 (2019): 1-8.

Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *Advances in neural information processing systems* 32 (2019).

# References


Cook, R. Dennis, and Sanford Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

Ustun, Berk, and Cynthia Rudin. "Supersparse linear integer models for optimized medical scoring systems." *Machine Learning* 102.3 (2016): 349-391.

Montavon, Grégoire, et al. "Layer-wise relevance propagation: an overview." *Explainable AI: interpreting, explaining and visualizing deep learning* (2019): 193-209.

Bach, S., Binder, A., Montavon, G., Klauschen, F., M ̈uller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7), e0130140 (2015)

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." *International conference on machine learning*. PMLR, 2017.

Dabkowski, Piotr, and Yarin Gal. "Real time image saliency for black box classifiers." *Advances in neural information processing systems* 30 (2017).

Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." *International conference on machine learning*. PMLR, 2017.

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

# If you have any questions, please let me know

sghalebikesabi

sahra.ghalebikesabi@univ.ox.ac.uk

https://sghalebikesabi.github.io