Explanation, can I trust you?
A tutorial on Explainable AI

Anna Hedström, PhD candidate, TU Berlin

Northern Lights Deep Learning Conference (NLDL) 2023

@anna_hedstroem
@TUBerlin_UMI

Today's agenda

01 Fundamentals

02 Challenges

03 Evaluation

04 Summary + Q&A

Provide an **optimistic** and **critical** view

Fundamentals

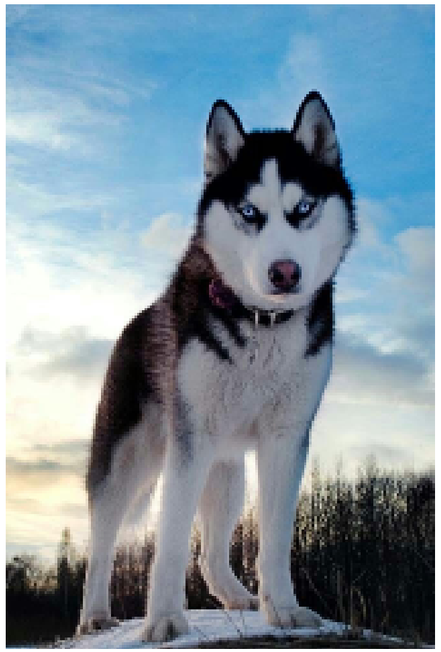
Why do we need Explainable AI?

Argument #1

Verification and trust

- Binary classification problem, train a classifier to distinguish between images

Wolfs



Huskies

Image credit: Prof. Dr. Marina Höhne

Argument #1

Verification and trust

- The model learns to discriminate between the classes to close to perfect accuracy

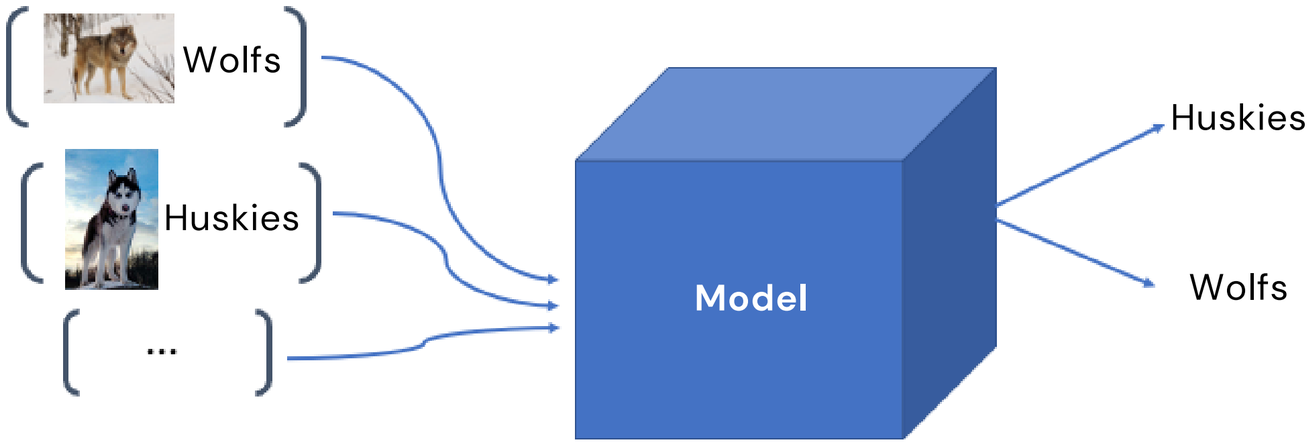


Image credit: Prof. Dr. Marina Höhne

Argument #1

Verification and trust

- To test how the model works in practice we run it on a test set

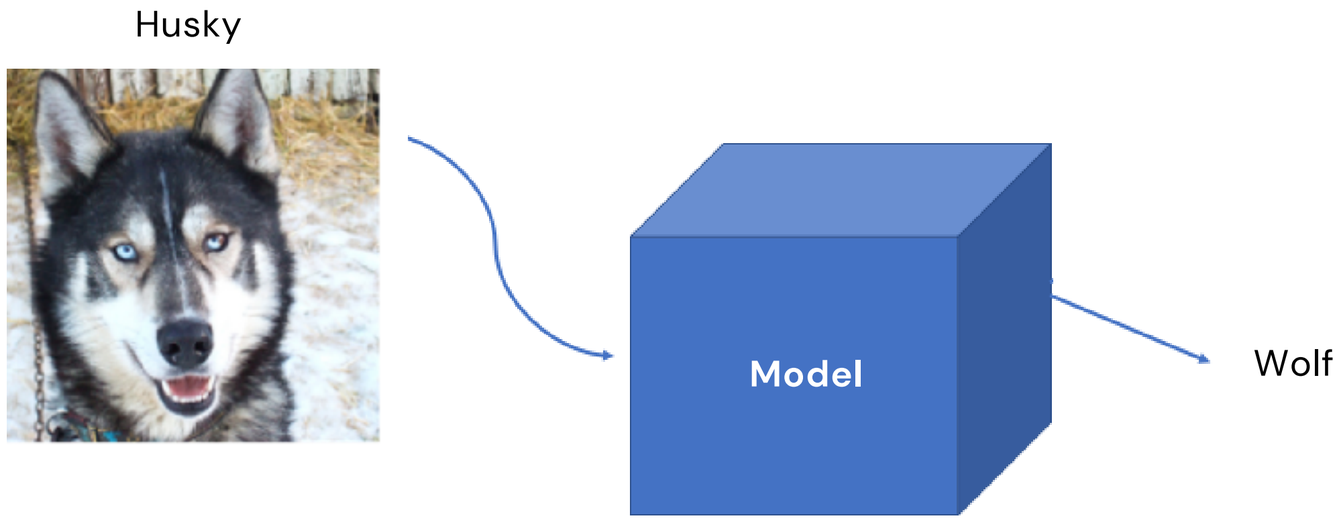


Image credit: Prof. Dr. Marina Höhne

Argument #1

Verification and trust

- The model appears as a black box

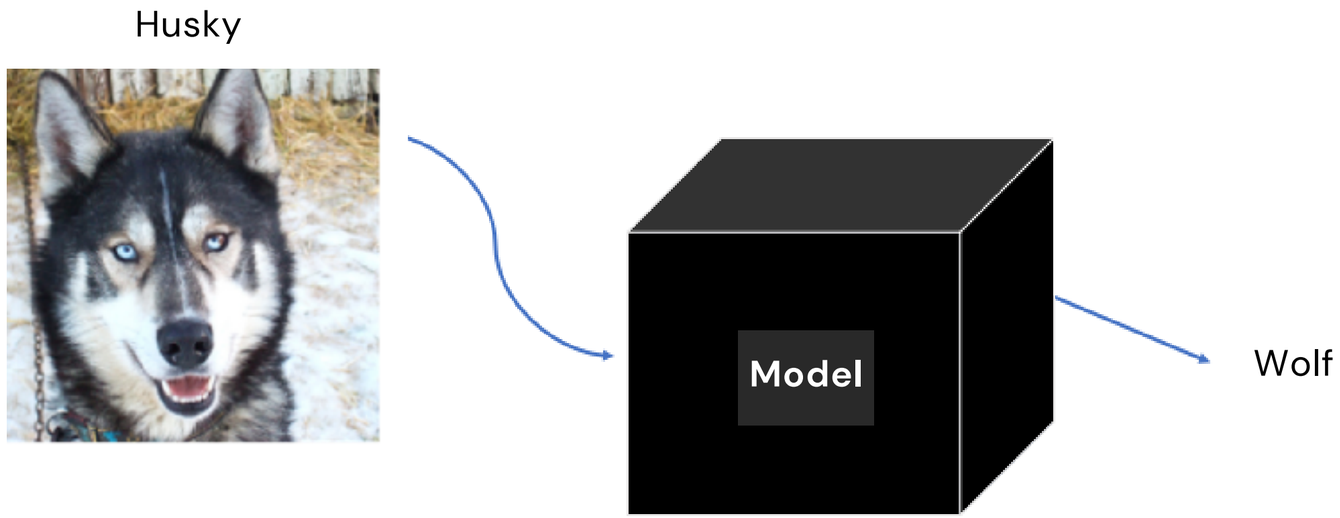


Image credit: Prof. Dr. Marina Höhne

Argument #1

Verification and trust

- We apply an explanation method showing the important features (i.e., pixels) for class wolf → the presence of snow

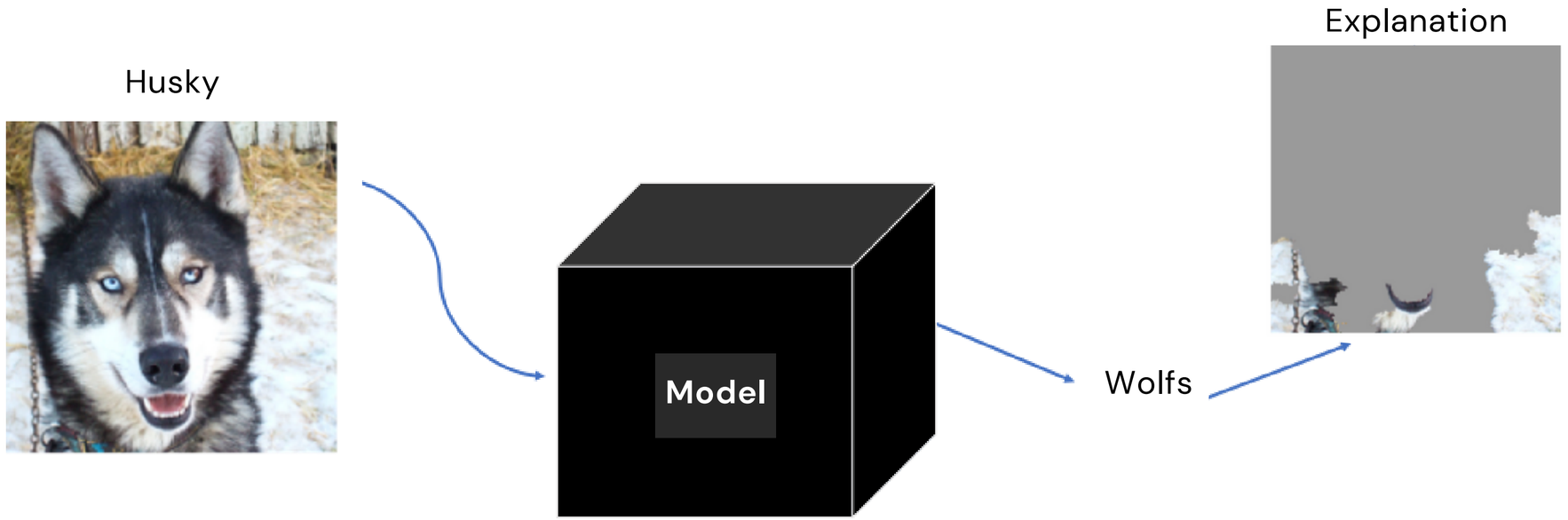
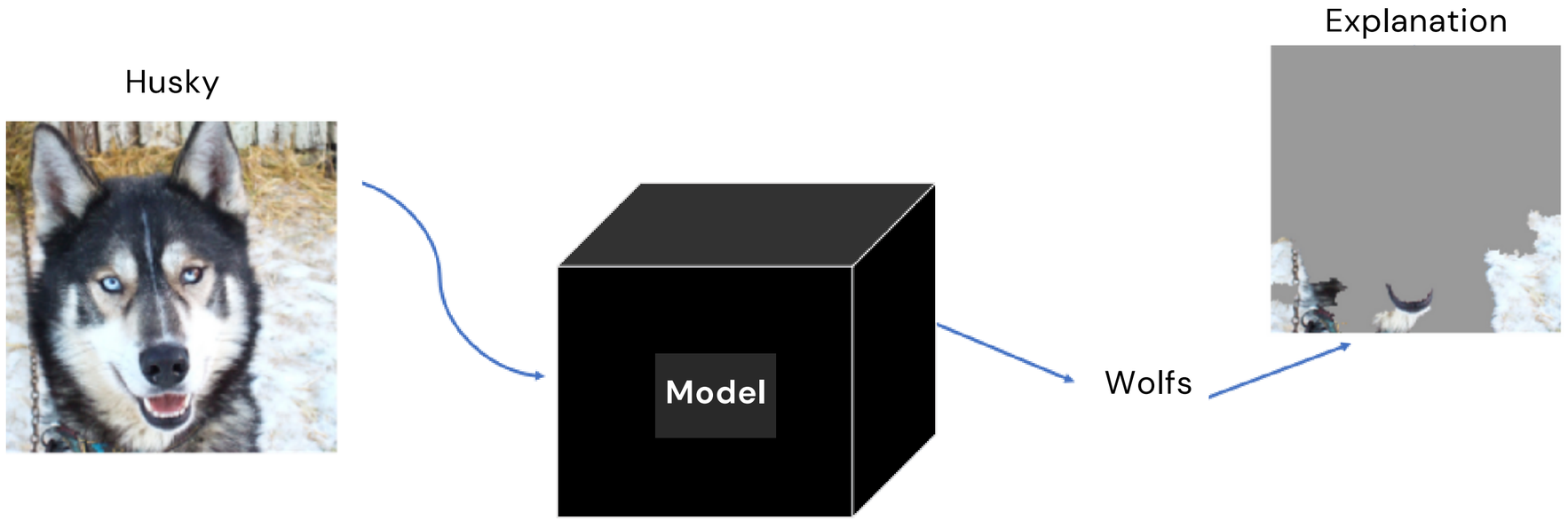


Image credit: Prof. Dr. Marina Höhne

Argument #1

Verification and trust

- We apply an explanation method showing the important features (i.e., pixels) for class wolf → the presence of snow



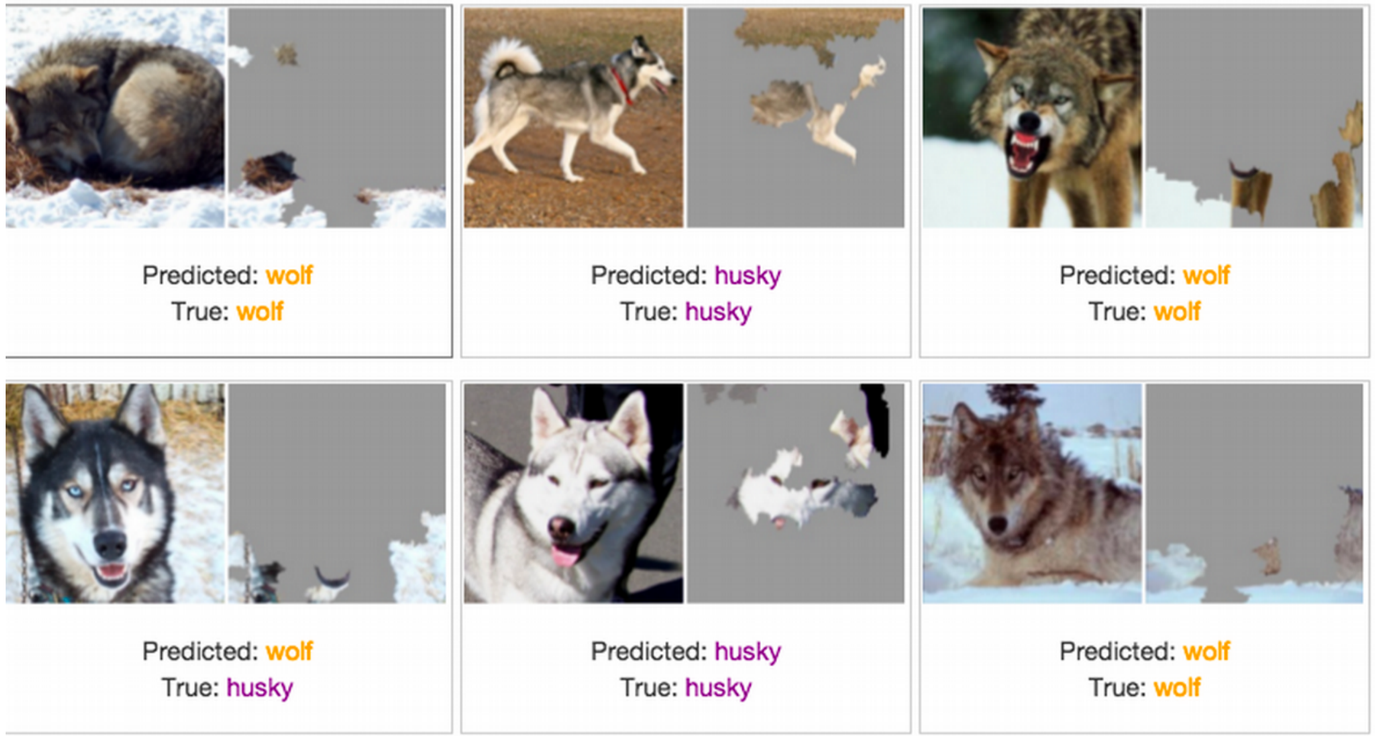
Is the goal of minimising error implying that the model works well in practice?

Image credit: Prof. Dr. Marina Höhne

Argument #1

Verification and trust

- **Not necessarily** — explanations can teach us if model features are representing a valid problem-solving behaviour for test environments



The model has learned to use "snow" when predicting "wolf"

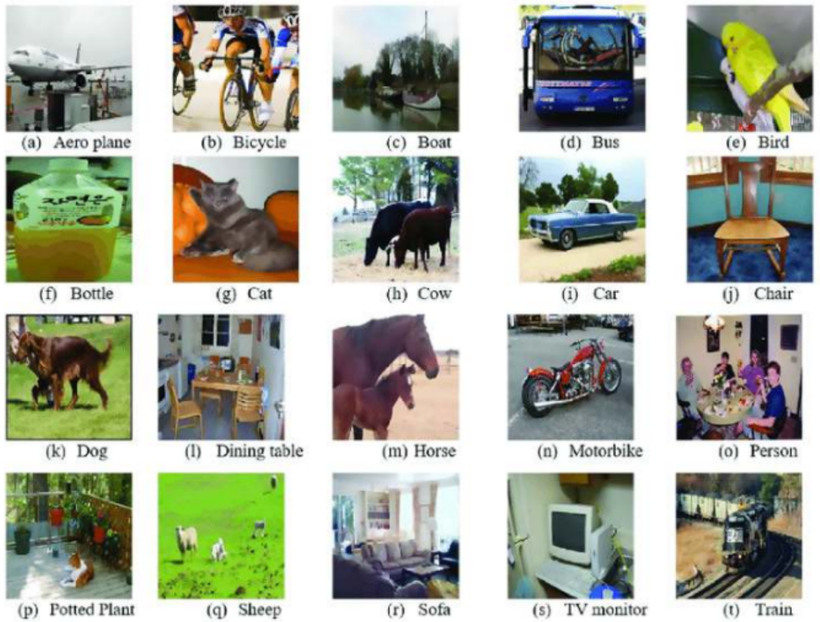
One wrong prediction

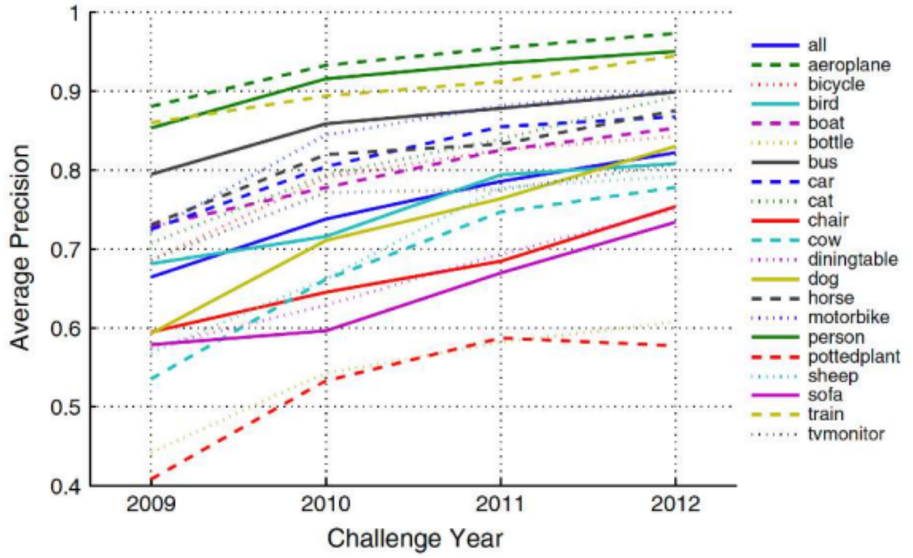
LIME explanation, https://theblue.ai/blog/lime-models-explanation/

Argument #2

Debugging purposes

- PASCAL-VOC Challenge (2005-2012) of multi-label classification for 20 classes
- Some of the best scientists participating in it, increasingly accurate predictions over the years





Argument #2

Debugging purposes

- When providing an input "horse", an expected prediction follows

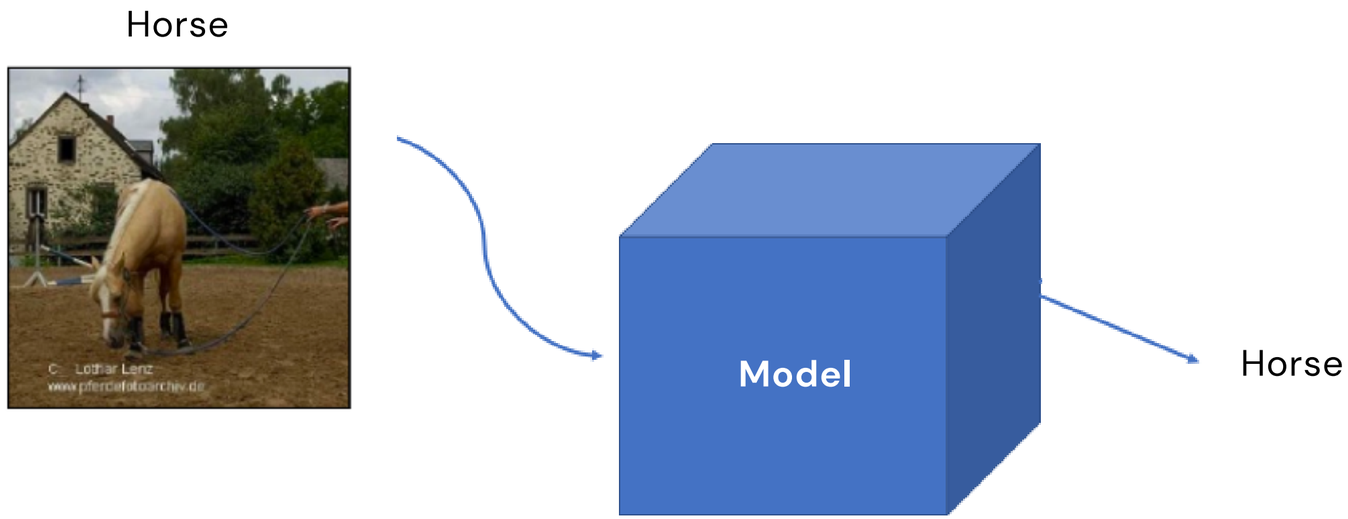


Image credit: Prof. Dr. Marina Höhne

Argument #2

Debugging purposes

- When providing an input "bird", an **unexpected prediction** follows

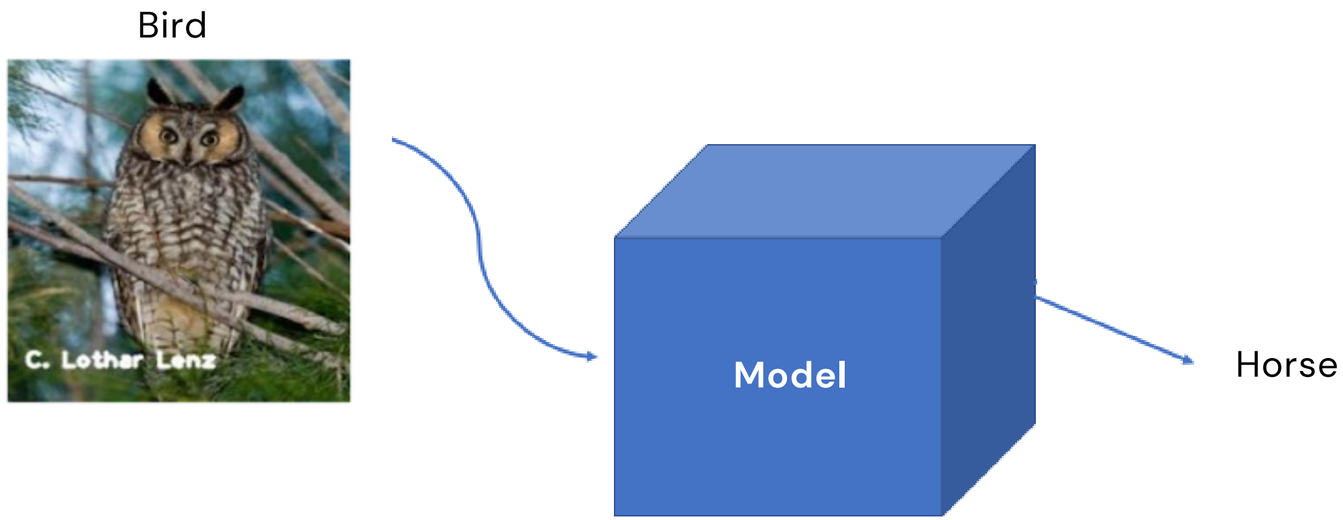
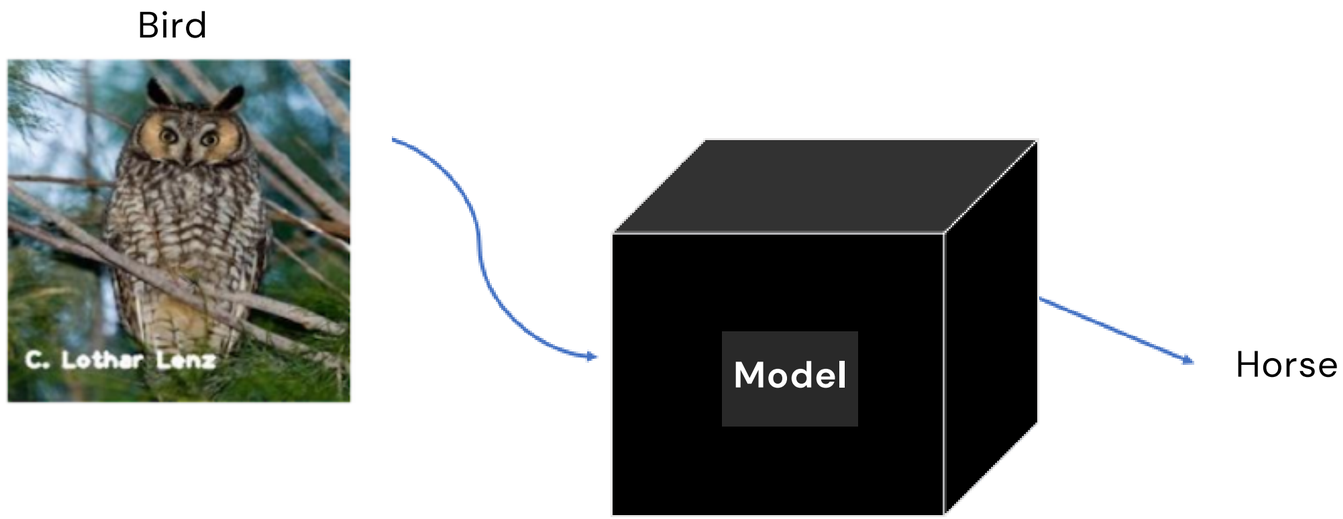


Image credit: Prof. Dr. Marina Höhne

Argument #2

Debugging purposes

- Our model behaves like a black box



Can we debug this?

Image credit: Prof. Dr. Marina Höhne

Argument #2

Debugging purposes

- By explaining the prediction "bird", we learn that a source tag was used as a feature

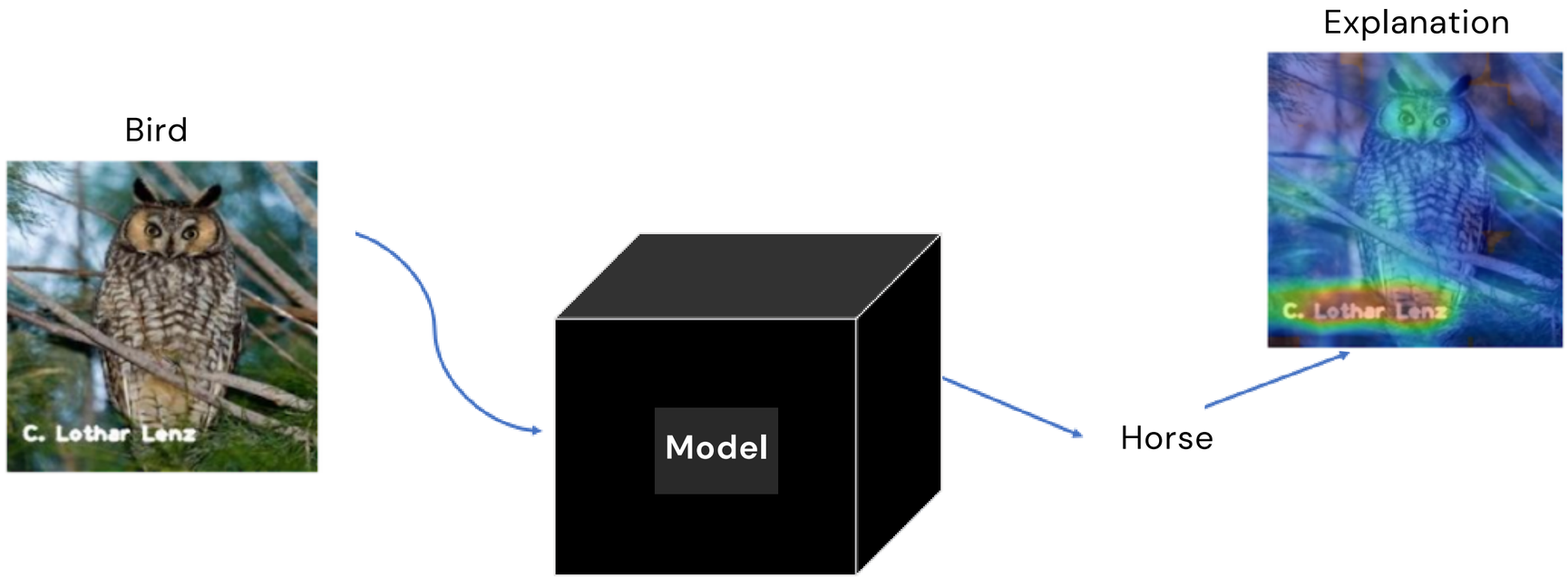
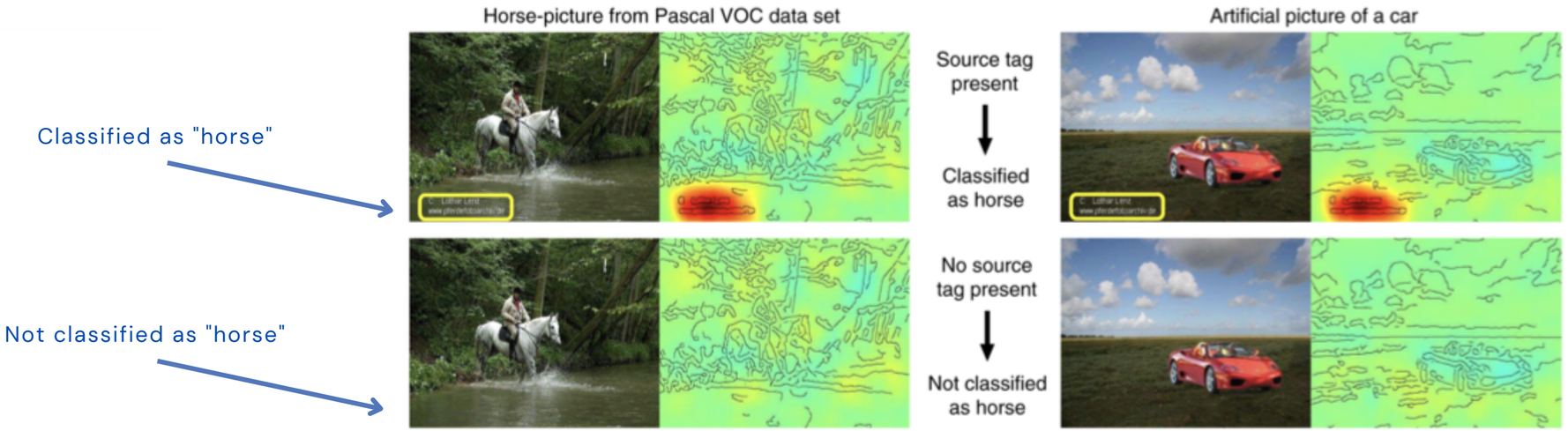


Image credit: Prof. Dr. Marina Höhne

Argument #2

Debugging purposes

- Lapuschkin et al., 2016 & 2019 showed with explanation methods that the model used non-object related features in its decision-making → "Clever Hans"



[4], Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking clever hans predictors and assessing what machines really learn. Nature communications, 10(1), 1-8.

What is "Clever Hans"?

Argument #2

Debugging purposes

- Clever Hans was a horse (1907) claimed to have performed arithmetic and other intellectual tasks

But

- The horse was responding directly to involuntary cues in the body language of the human trainer (who was entirely unaware)

Argument #3

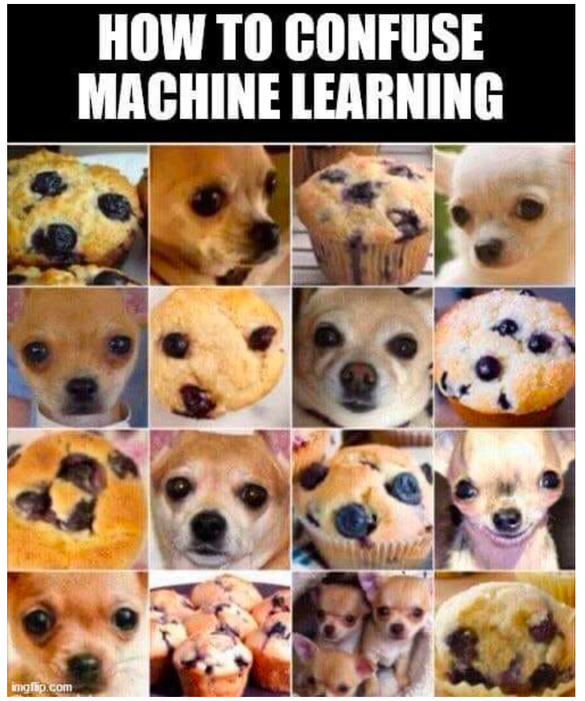
Legal aspects

- The General Data Protection Regulation laws in EU states all people have the right to “meaningful information about the logic behind automated decisions using their data”

"right to an explanation"

- Necessary in high-risk and in avoiding discrimination/ bias
  - e.g., for doctors, lawmakers, policymakers, banks

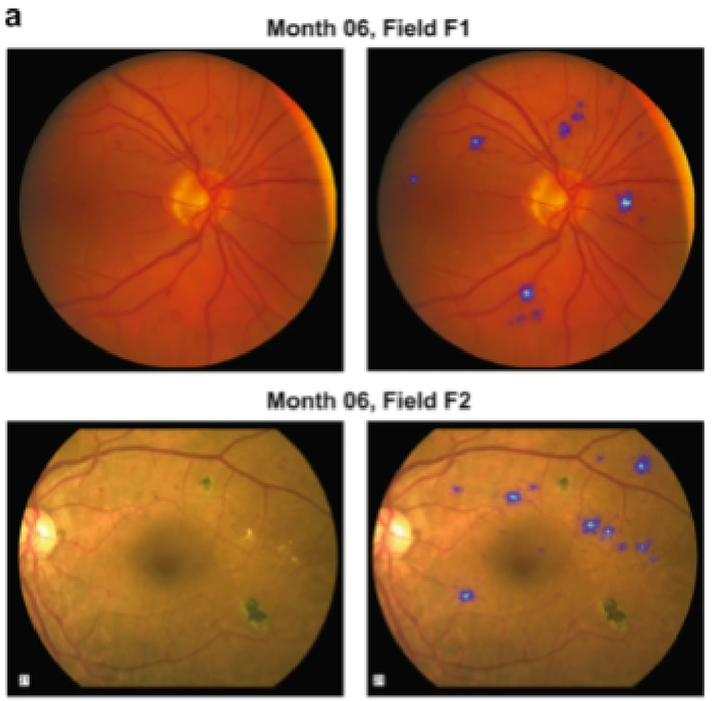
- Building explainable AI methods became more urgent
  - as Adversarial ML revealed the fragility of neural networks



Argument #4

Generating new insights

- Notable successes in using explanation AI methods to aid in the discovery of knowledge
- For example, Arcadu et al., 2019 used **explanations of a DNN to identify novel features of diabetic retinopathy progression** (a complication of diabetes, that damages the back of the eye)

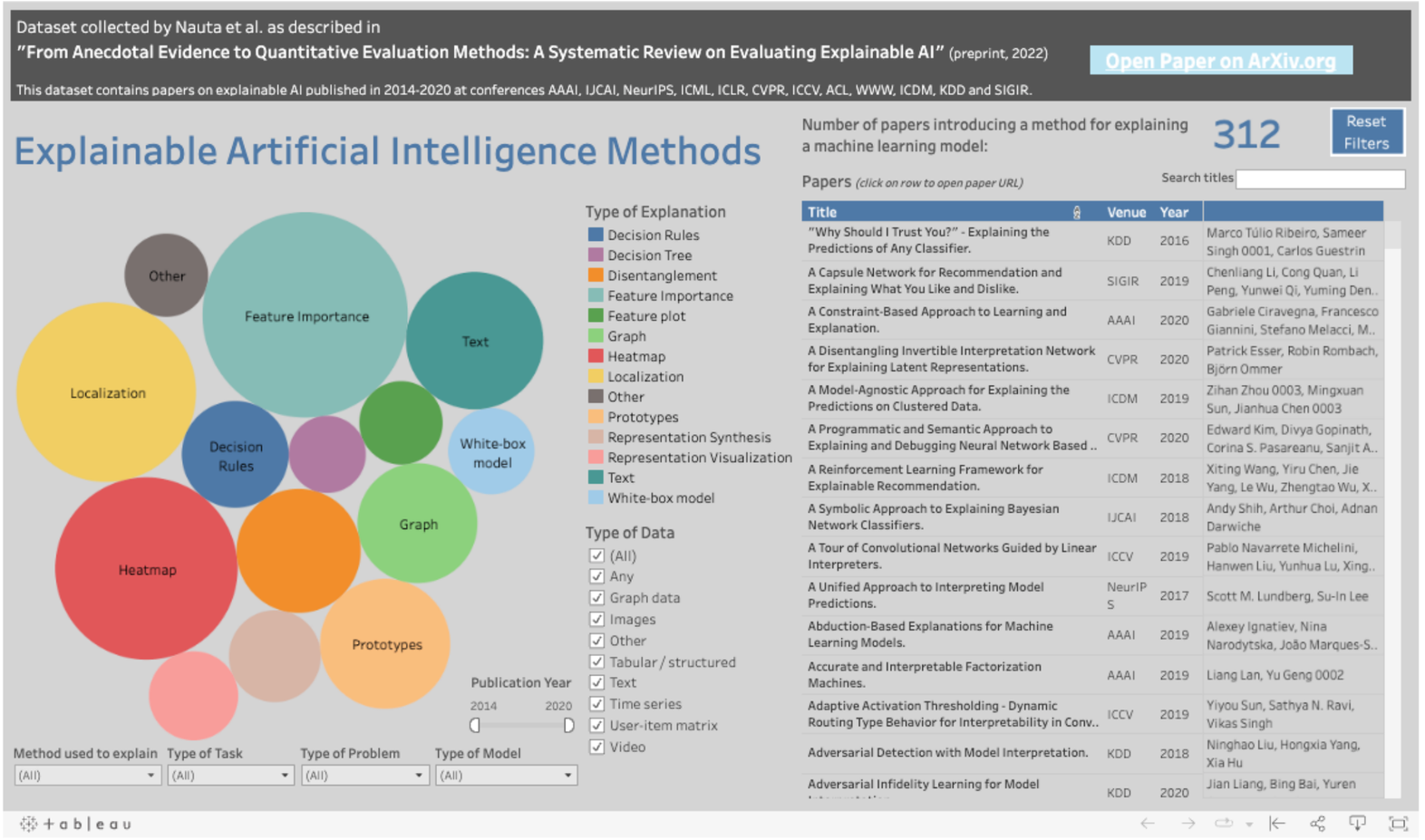


Left: original image, Right: attribution maps locating new areas "features" that the model used for predicting a month of progression

What did all these drivers result in?

Enormous activity

A diverse set of explanation methods for different tasks, models, data and users



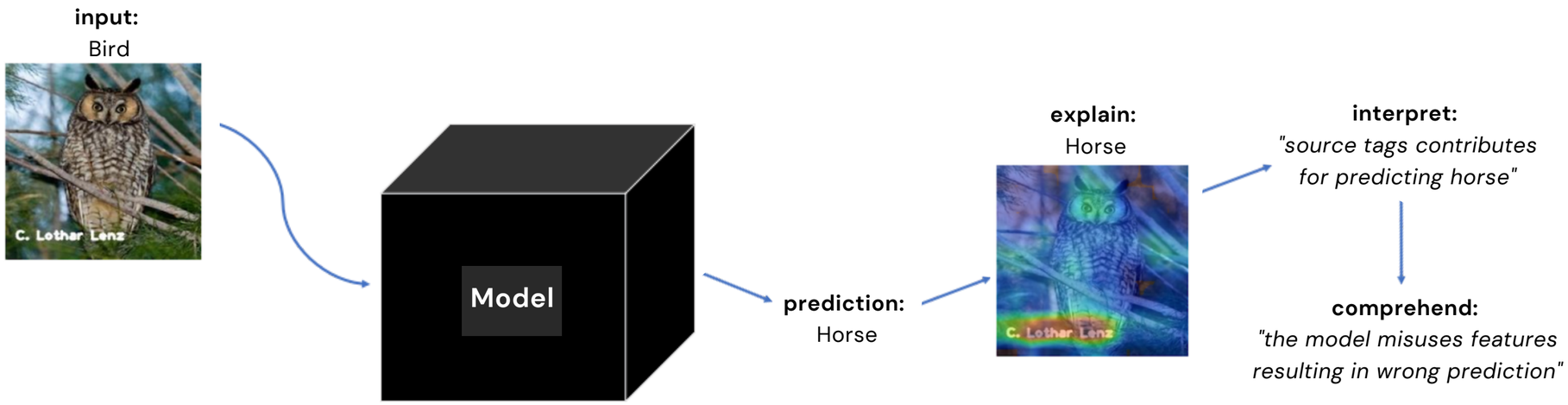
https://public.tableau.com/app/profile/m.nauta/viz/eval_xai/survey_data (et al., 2022)

Defining explainability

Defining explainability

The answer is "yes" and "no"

- **explain**: refers to the process of computation of the explanation (e.g., attribution map)
- **interpret**: refers to the process of assigning a meaning to the explanation
- **comprehend**: refers to a deeper functional insight (e.g., takeaway) of the model

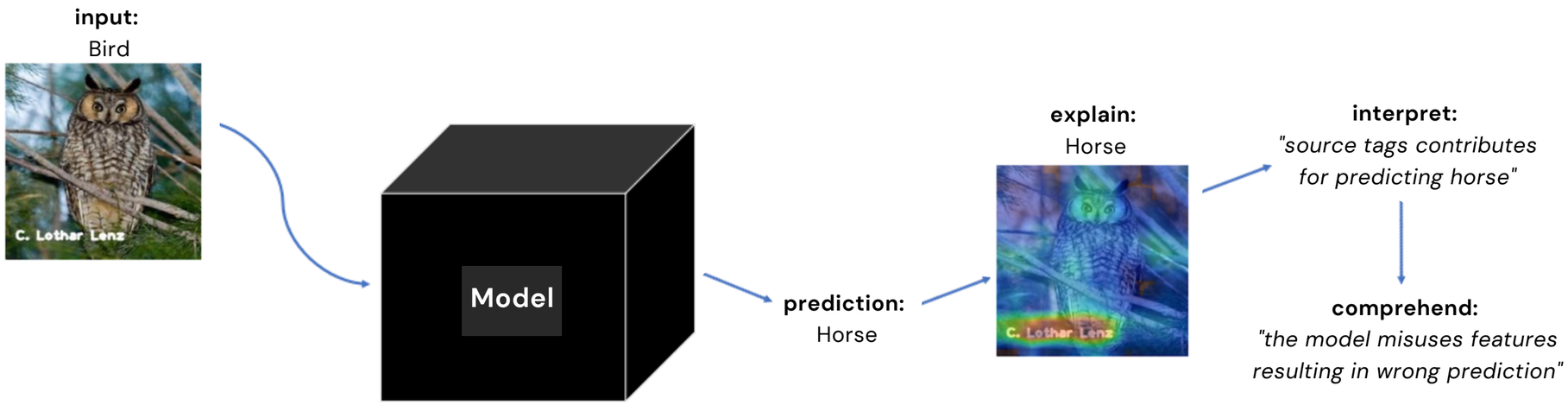


Definitions credit: Prof. Dr. Wojciech Samek

Defining explainability

The answer is "yes" and "no"

- **explain**: refers to the process of computation of the explanation (e.g., attribution map)
- **interpret**: refers to the process of assigning a meaning to the explanation
- **comprehend**: refers to a deeper functional insight (e.g., takeaway) of the model



To whom?
Developers, model owner, end-users

What to explain?
Local, global, combination

When to explain?
Before, during, after model training

Using what data?
Black-box, white-box, dataset

Medium to explain?
Heatmaps, dts, text

Methods

Overview of methods

Explainable AI can be taxonomised in infinitely many ways

- Local methods
  - Model-agnostic methods
  - Model-aware methods
- Global methods
- Combining/ enhancing methods

Overview of methods

Explainable AI can be taxonomised in infinitely many ways

- Local methods provide explanations for single predictions, decisions, in a post-hoc manner
  - Model-agnostic methods
  - Model-aware methods
- Global methods
- Combining/ enhancing methods

Overview of methods

Explainable AI can be taxonomised in infinitely many ways

- Local methods
  - Model-agnostic methods
  - Model-aware methods
- Global methods provide explanations for a model's representation, decision-making behaviour
- Combining/ enhancing methods

Overview of methods

Explainable AI can be taxonomised in infinitely many ways

- Local methods
  - Model-agnostic methods
  - Model-aware methods
- Global methods

- Combining/ enhancing methods provides explanation by means of aggregation or integration

Local methods

Local methods

Model-agnostic methods

Occlusion

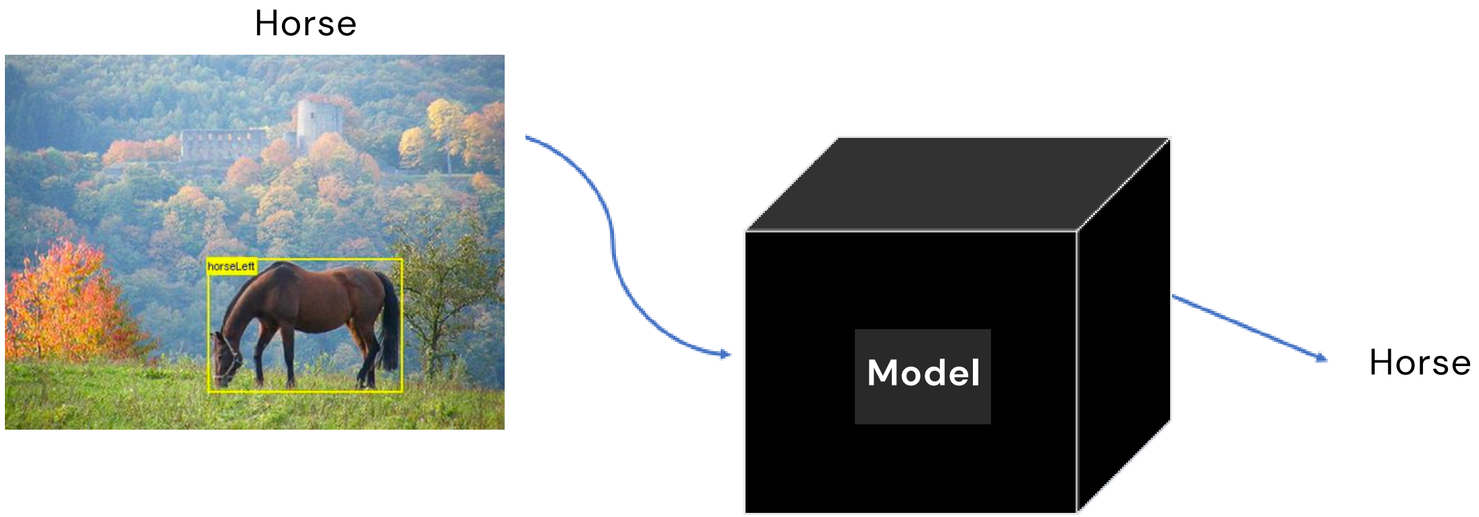
Model-agnostic method

- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat

Occlusion

Model-agnostic method

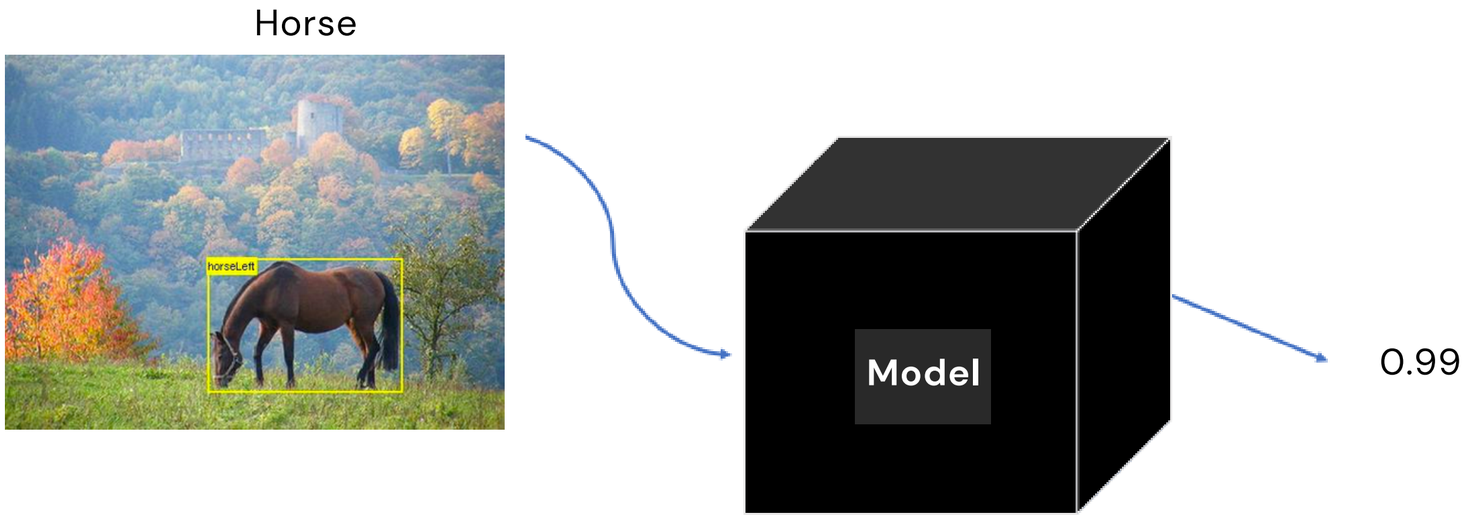
- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Occlusion

Model-agnostic method

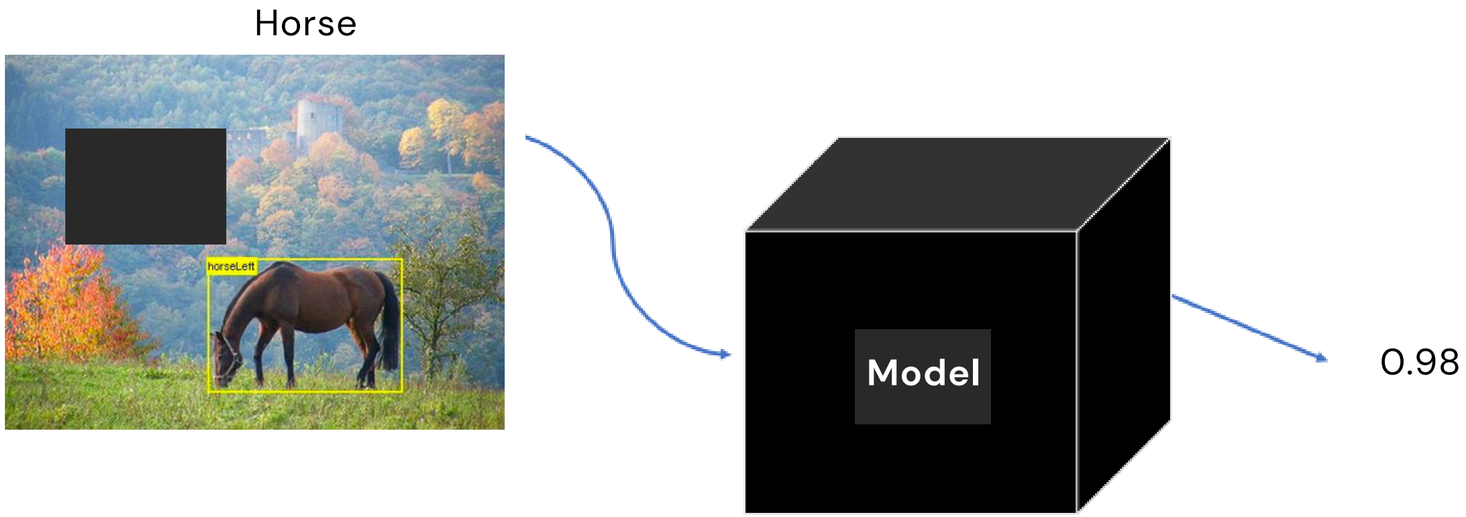
- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Occlusion

Model-agnostic method

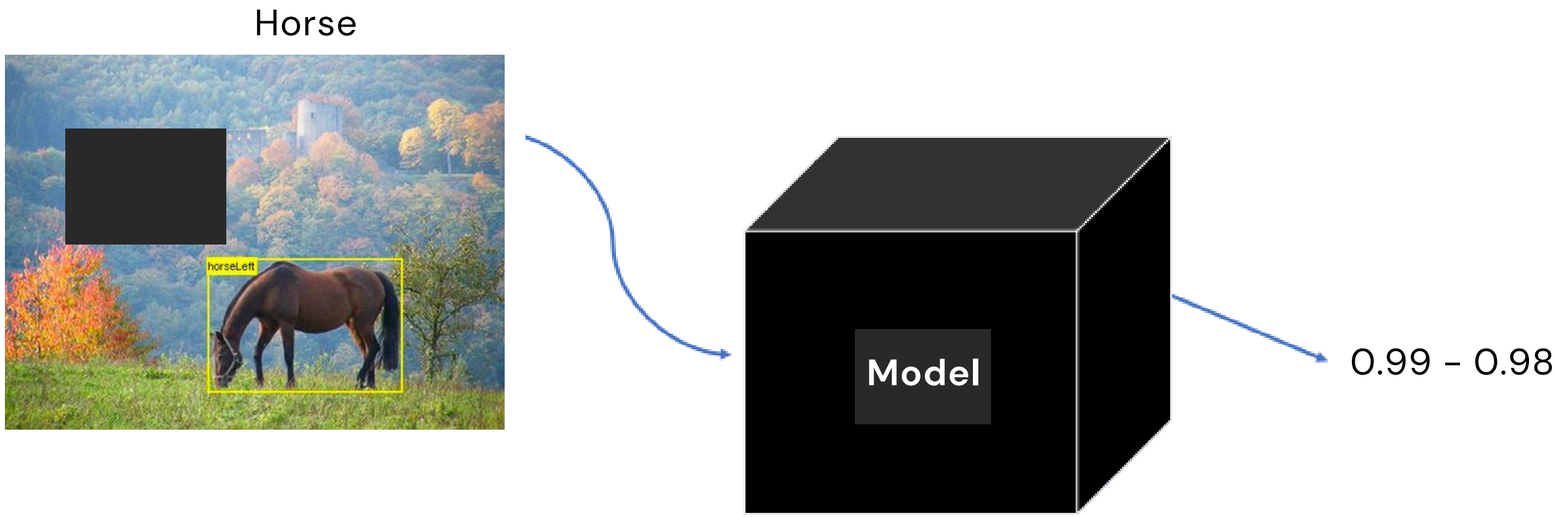
- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Occlusion

Model-agnostic method

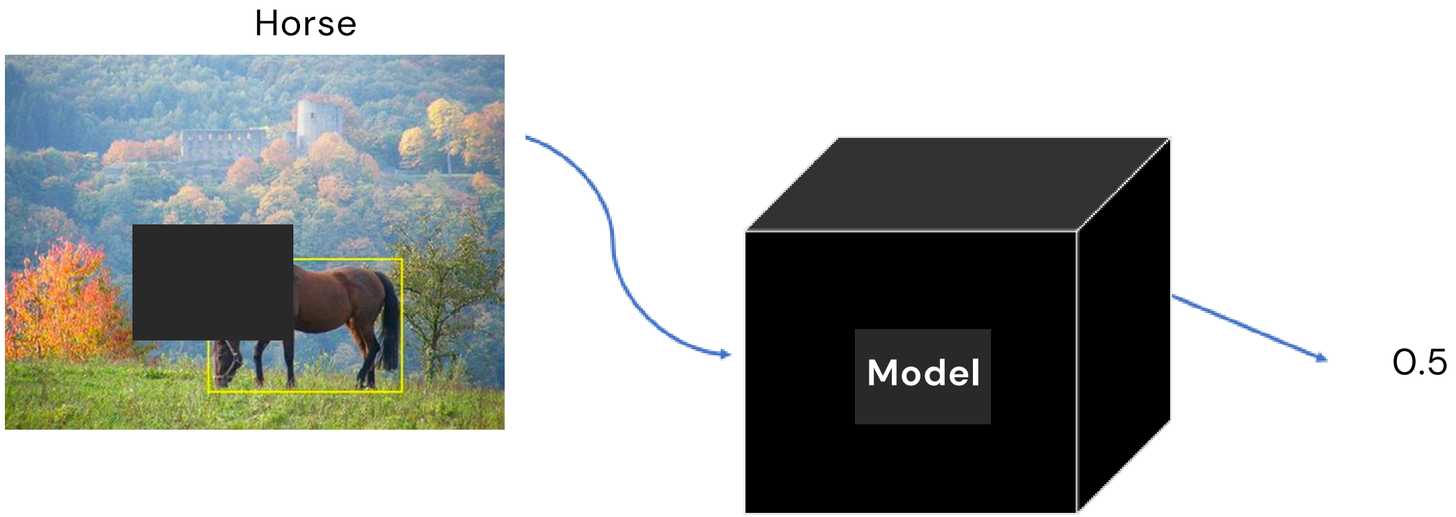
- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Occlusion

Model-agnostic method

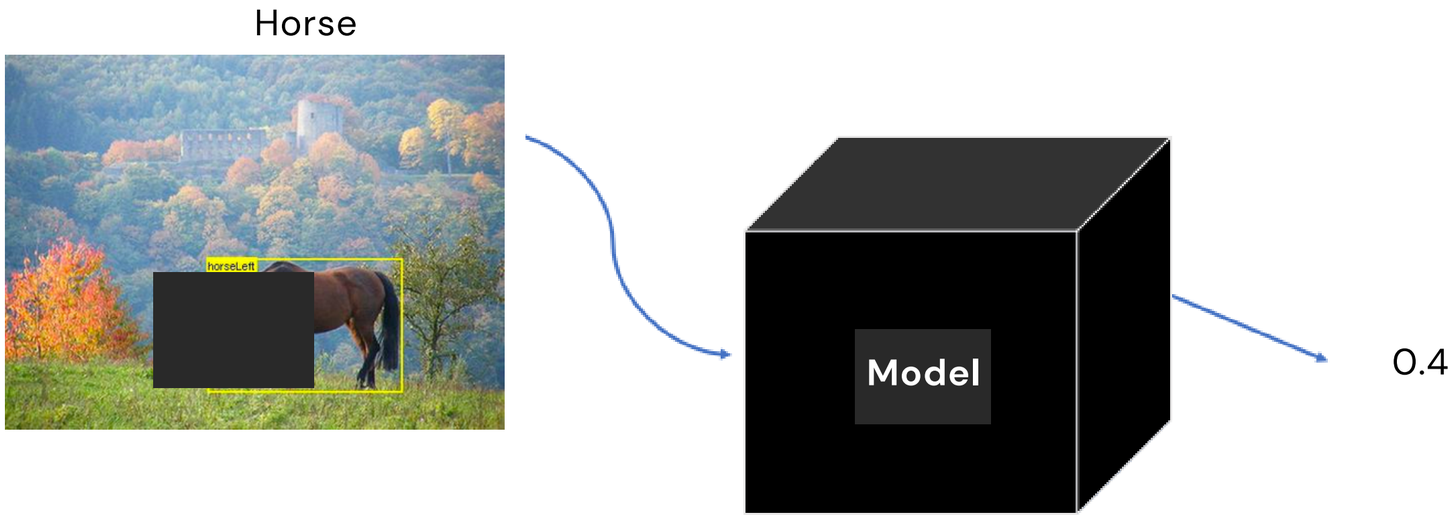
- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Occlusion

Model-agnostic method

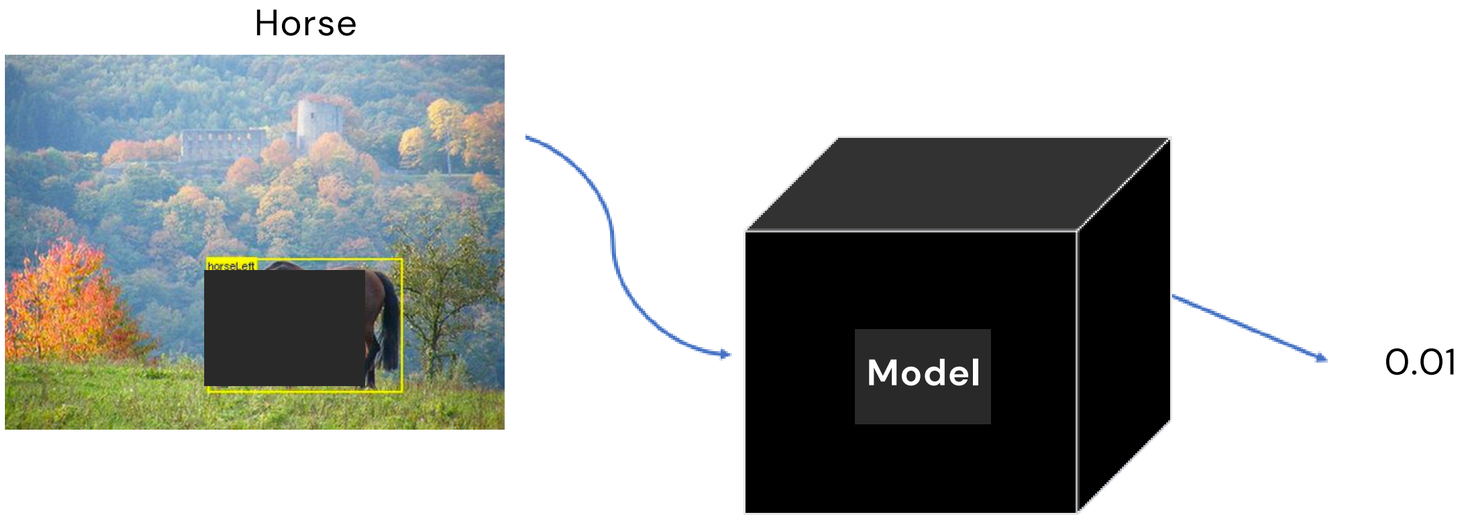
- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Occlusion

Model-agnostic method

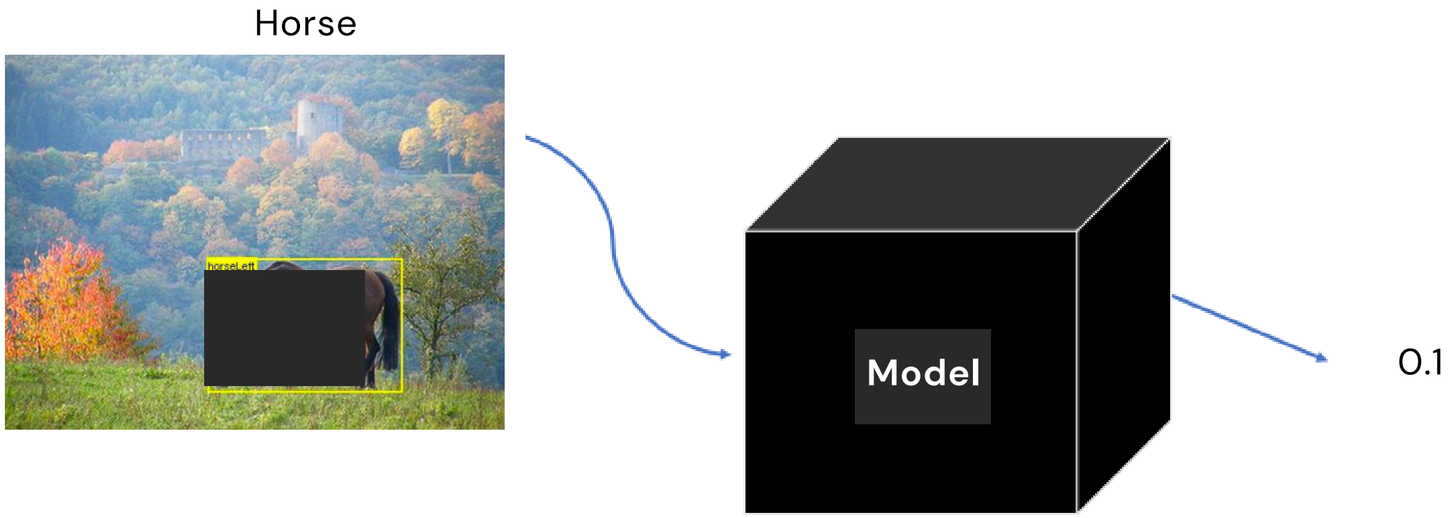
- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Occlusion

Model-agnostic method

- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat

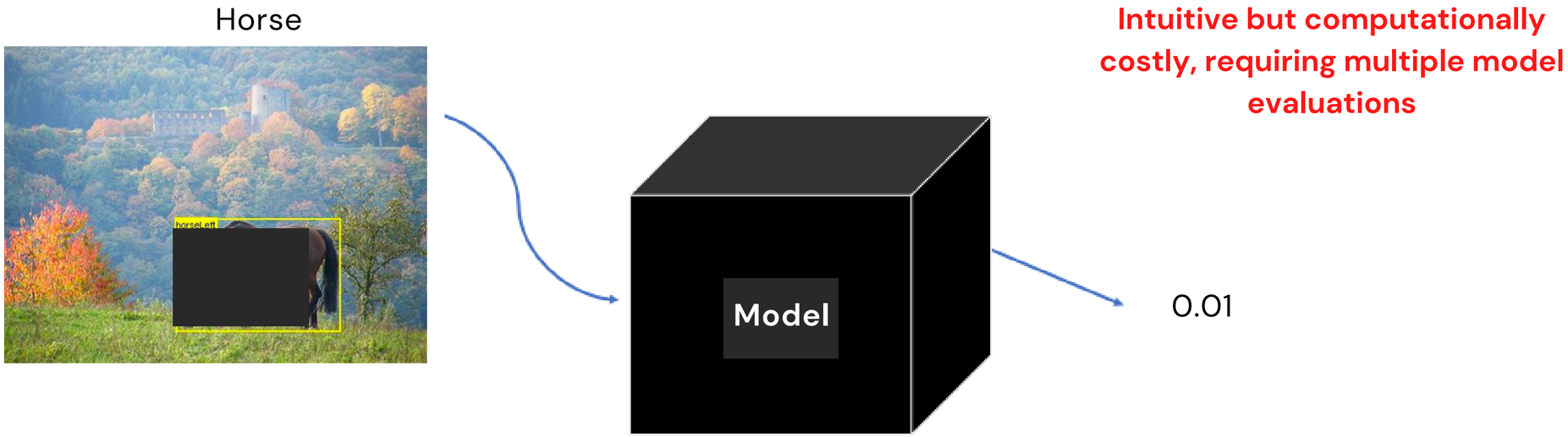


- Use this information to assign feature importance to individual pixels

Occlusion

Model-agnostic method

- For example, use Occlusion (Zeiler et al., 2014) to explain a prediction of a pre-trained model
  - Feed *x* through the neural network and record *y*
  - Occlude a region (e.g., using a black patch) and feed it and record *y'*
  - Compute the difference between *y* and *y'*
  - Repeat



Intuitive but computationally costly, requiring multiple model evaluations

Prototypes

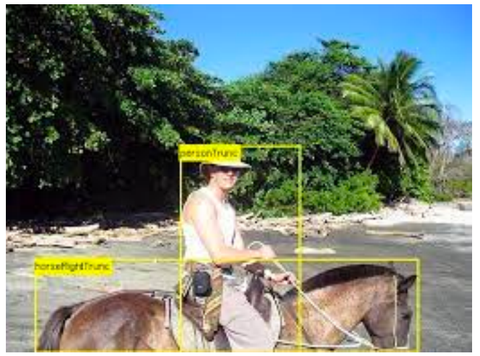
Model-agnostic method

- Prototype explanations are generated by finding a representative example or "prototype" from the training data that is similar to the sample being explained

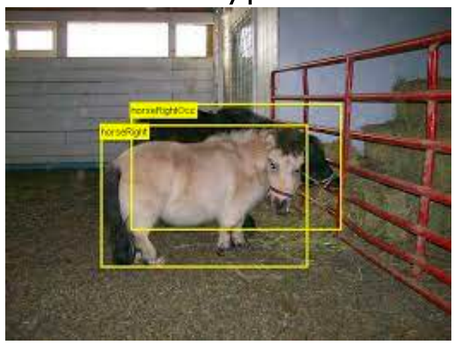
Explain "Horse"

Prototype 1

Prototype 2



Prototype 3



Prototypes

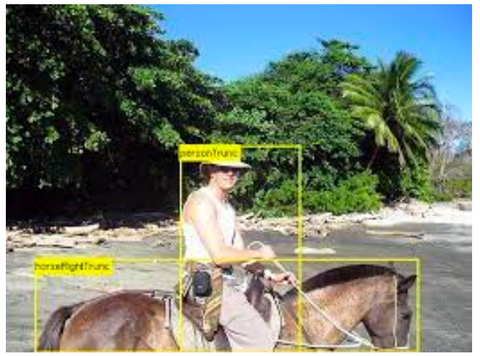
Model-agnostic method

- Prototype explanations are generated by finding a representative example or "prototype" from the training data that is similar to the sample being explained

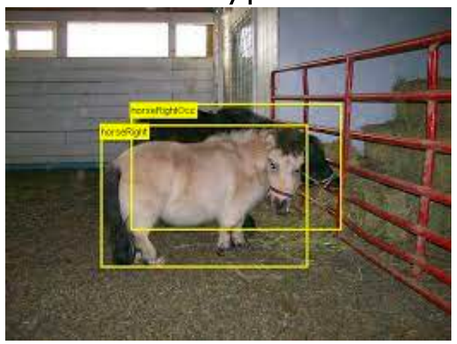
Explain "Horse"

Prototype 1

Prototype 2



Prototype 3

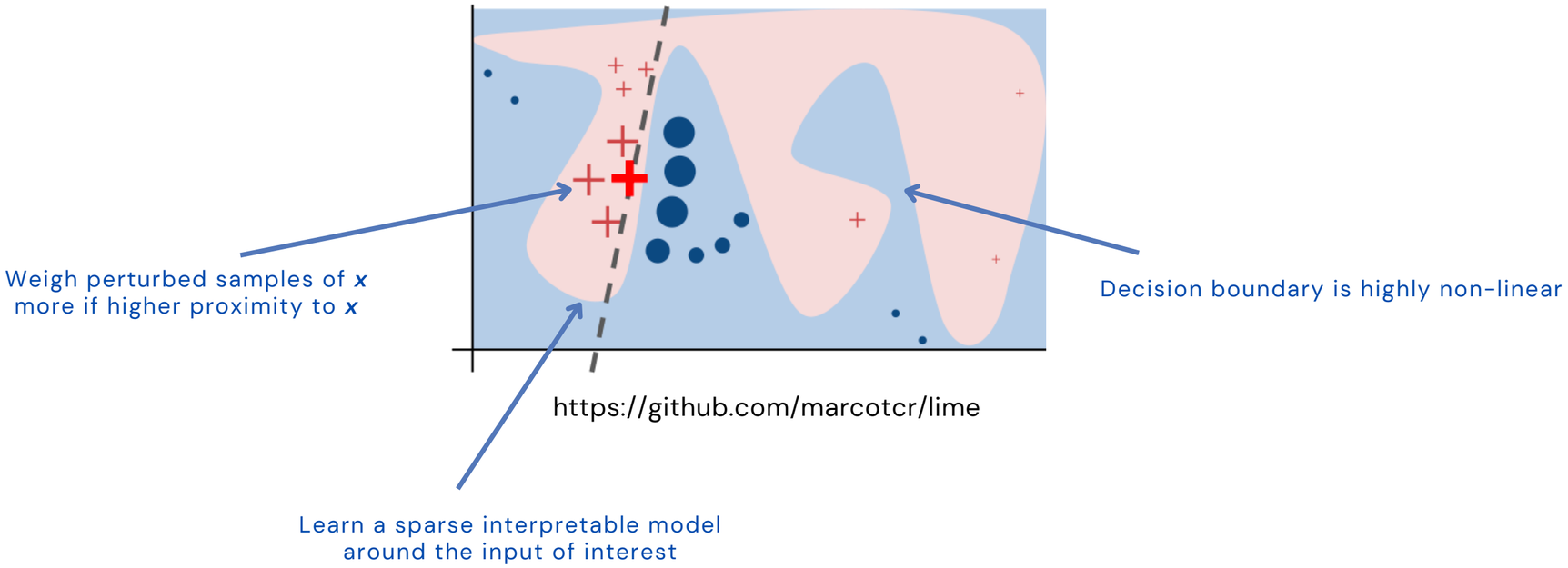


May not be representative of the model's decision boundary, especially in high-dimensional or complex data

LIME

Model-agnostic method

- A model might be complex globally, but locally it may be easier to understand it
- **Idea:** Explain a prediction with help of an interpretable surrogate model (Ribeiro et al., 2016)



LIME

Model-agnostic method

- A model might be complex globally, but locally it may be easier to understand it
- **Idea:** Explain a prediction with help of an interpretable surrogate model (Ribeiro et al., 2016)

$$\arg\min_{g \in G} L(f_c, g, \pi_x) + \Omega(g)$$

Learn a surrogate model *g* by minimising loss on ***x***, ***y*** pairs

Collect ***x***, ***y*** pairs by sampling in the neighbourhood of ***x*** with vicinity denoted ***π*** and then evaluate the network on these points ***f***

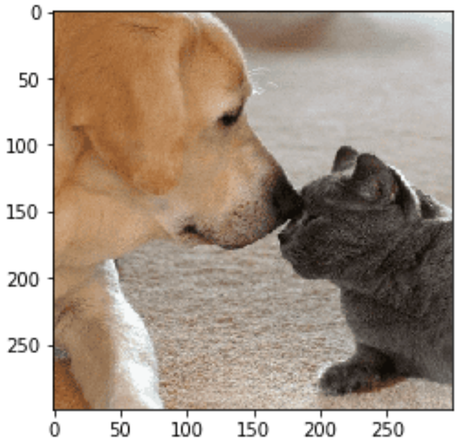
Keep model complexity low

- Weight vector of ***g*** is used as the basis for assigning the attributions to the input feature

LIME

Model-agnostic method

Explain:
"labrador"



Super-pixels as
features

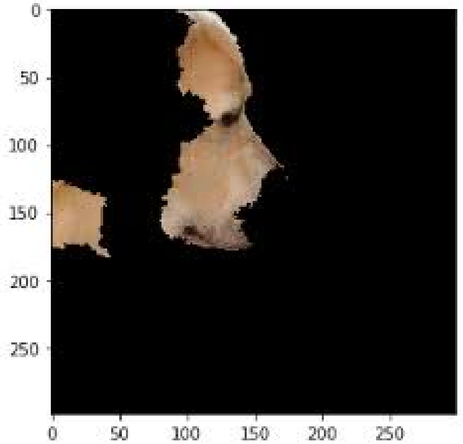
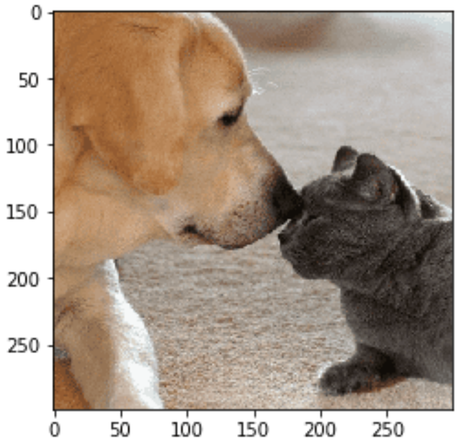


Image classification example

LIME

Model-agnostic method

Explain:
"labrador"



Super-pixels as
features

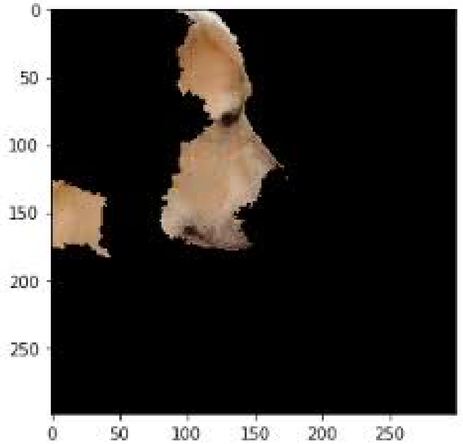


Image classification example

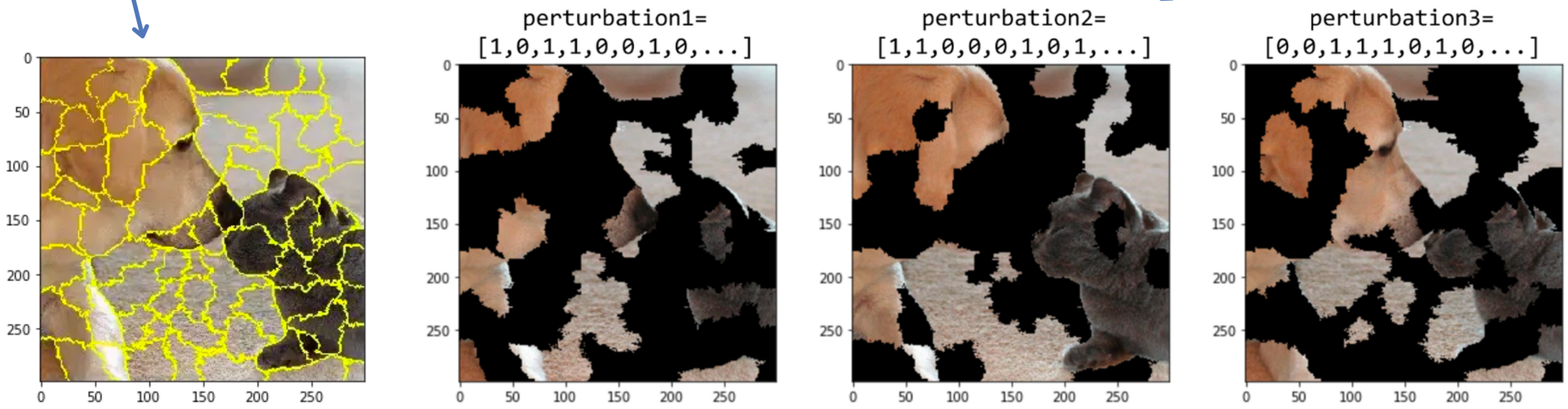
How were these "interpretable" features generated? How much should the input be"perturbed"?

LIME

Model-agnostic method

Segmentation mask of "interpretable" super-pixels, generated e.g., by SLIC algorithm

Different perturbations used to collect ***x, y*** pairs for surrogate model



Left: Segmentation of interpretable super-pixels, Right: Different perturbed samples

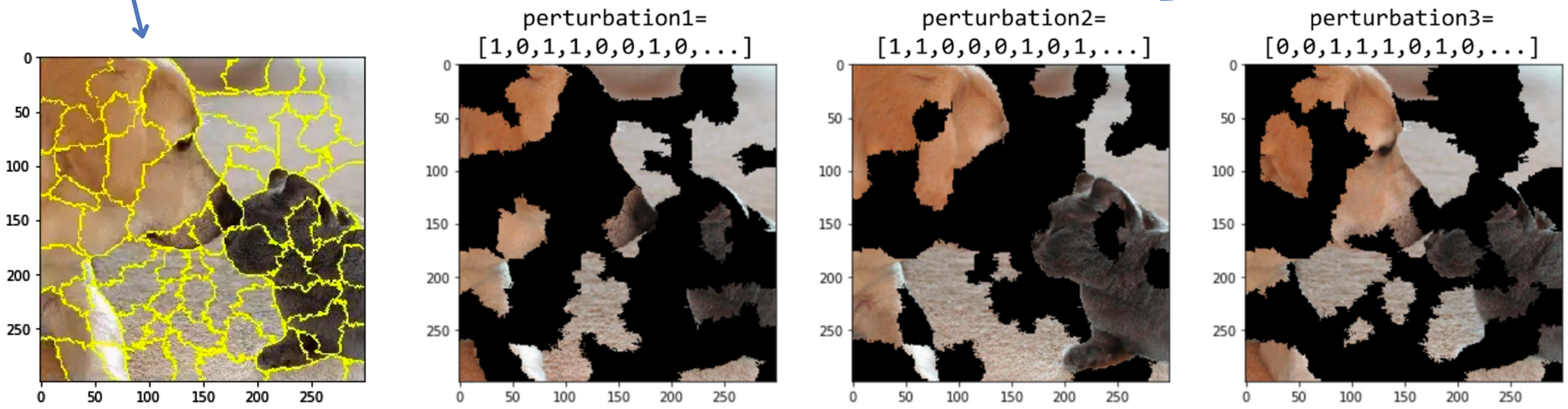
https://towardsdatascience.com/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13

LIME

Model-agnostic method

Segmentation mask of "interpretable" super-pixels, generated e.g., by SLIC algorithm

Different perturbations used to collect ***x, y*** pairs for surrogate model



Left: Segmentation of interpretable super-pixels, Right: Different perturbed samples
https://towardsdatascience.com/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13

LIME explanations can be sensitive to the choice of perturbation parameters

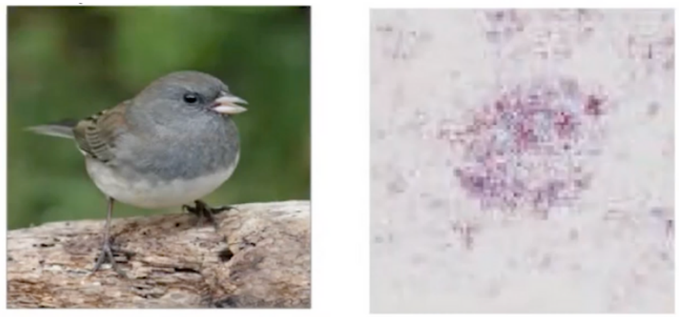
Local methods

Model-aware methods

Gradient-based methods

Model-aware method

- **Gradient-based methods** are a family of methods that explain a differentiable function ***f***
  - Compute the gradient of the output prediction w.r.t the input
  - The intensity of the colour indicates the importance of the feature
  - Encodes the local behaviour of the model



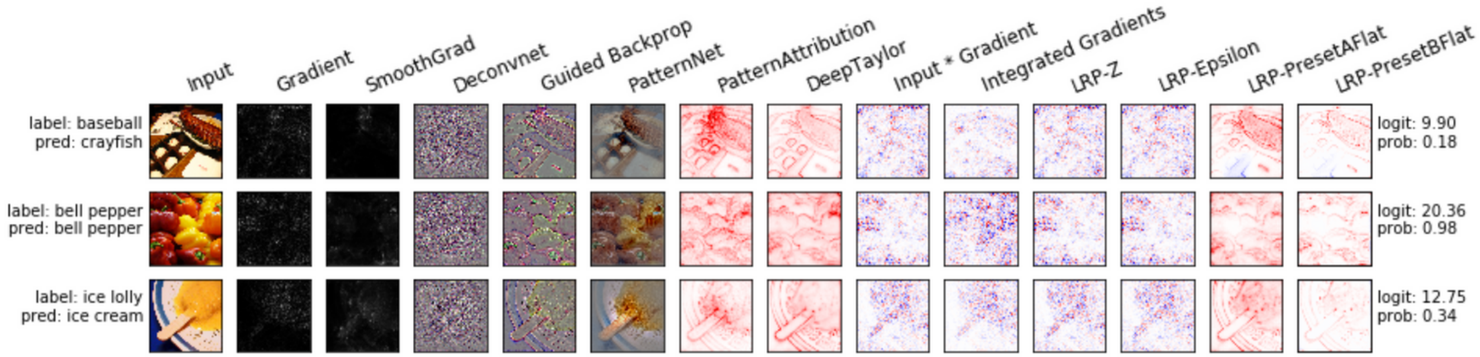
SmoothGrad: removing noise by adding noise (Smilkov et al., 2017)

- Many variations exist

Gradient-based methods

Model-aware method

- **Gradient-based methods** are a family of methods that explain a differentiable function ***f***
  - Compute the gradient of the output prediction w.r.t the input
  - The intensity of the colour indicates the importance of the feature
  - Encodes the local behaviour of the model



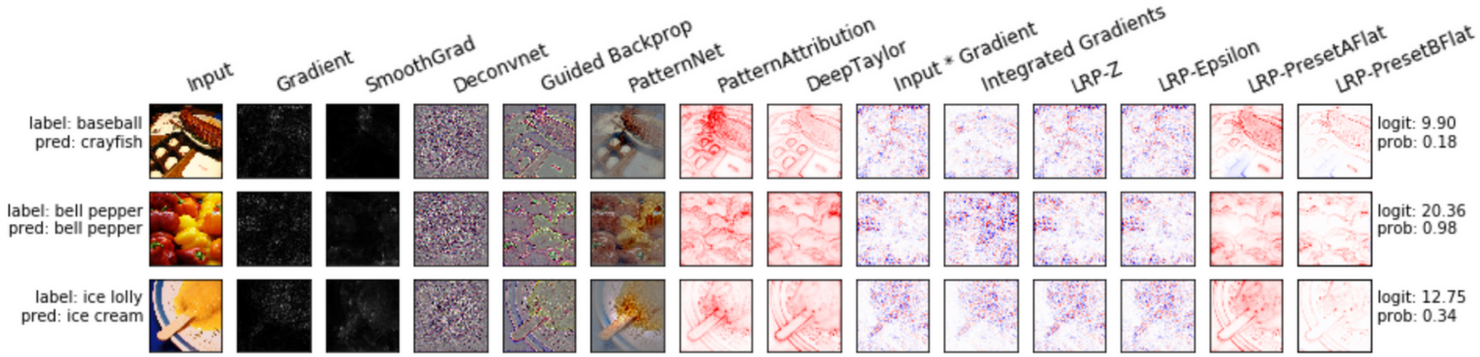
https://github.com/albermax/innvestigate

- Many variations exist

Gradient-based methods

Model-aware method

- **Gradient-based methods** are a family of methods that explain a differentiable function ***f***
  - Compute the gradient of the output prediction w.r.t the input
  - The intensity of the colour indicates the importance of the feature
  - Encodes the local behaviour of the model



https://github.com/albermax/innvestigate

- Many variations exist

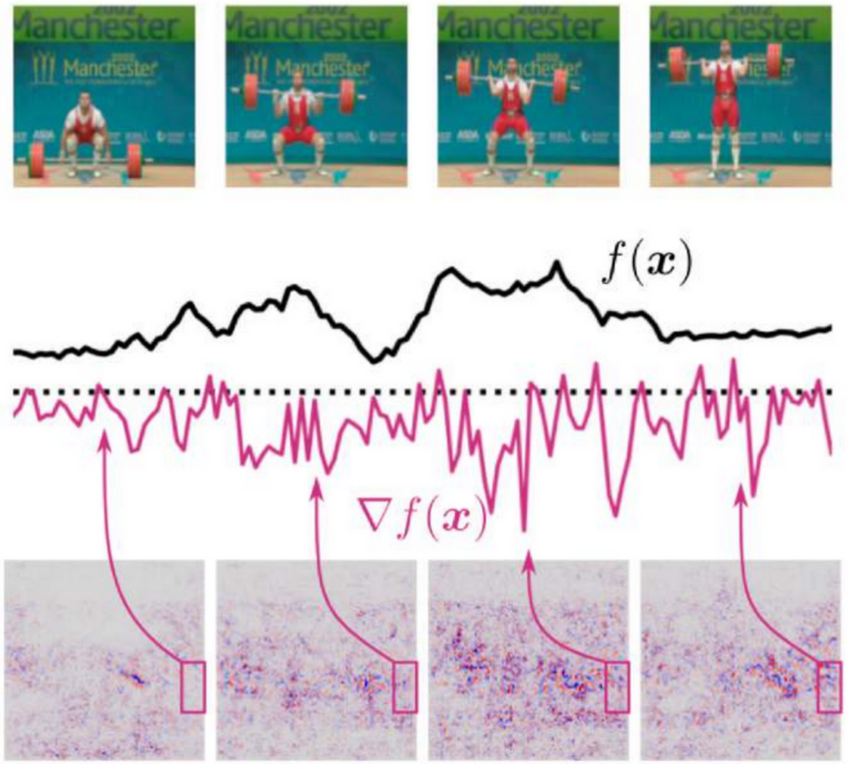
"I have no reason to believe the gradient holds anywhere other than very locally."

Gradients are "fragile"

How can gradient-based explanations be robust if the underlying model is not?

- Because of the depth of the function (Balduzzi et al., 2017), **the gradients may behave noisy while the function output may not**

- If following some trajectory along the input (e.g., an athlete lifting weights), there is a discrepancy between the output and gradient w.r.t. the input



http://www.heatmapping.org/

Image credit: Prof. Dr. Wojciech Samek

Propagation-based methods

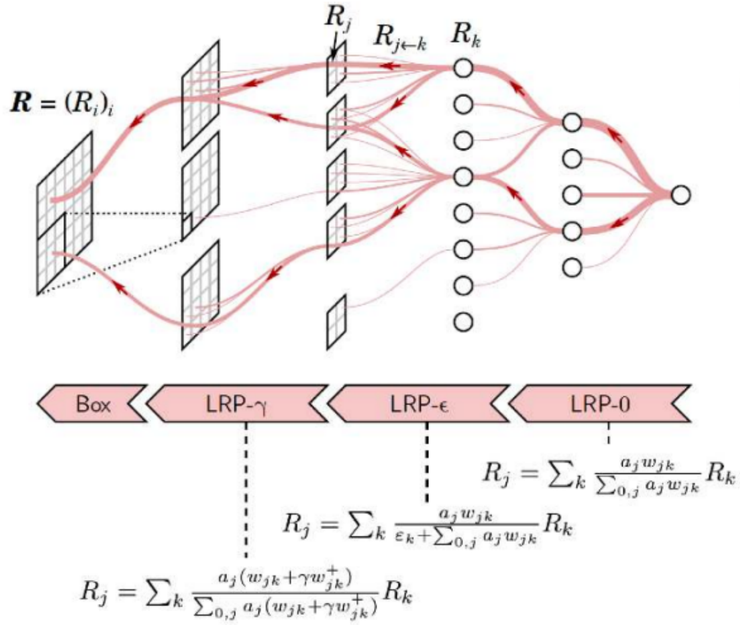
Model-aware method

- **Layer-wise Relevance Propagation (LRP)** is a white-box method grounded on the principles of flow conservation and proportional decomposition

- Works by performing a backward pass: distributing a model output quantity proportionally across the layers, according to the activations, back to input space

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

- Employs different propagation rules for different architectures and layer types



Propagation-based methods

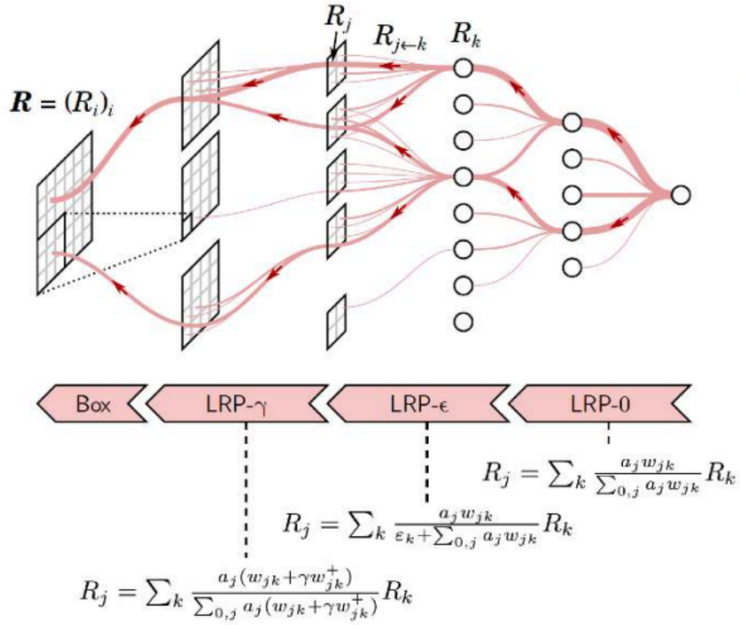
Model-aware method

- **Layer-wise Relevance Propagation (LRP)** is a white-box method grounded on the principles of flow conservation and proportional decomposition

- Works by performing a backward pass: distributing a model output quantity proportionally across the layers, according to the activations, back to input space

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

- Employs different propagation rules for different architectures and layer types



LRP has hyperparameters which may be non-trivial to optimise

Global methods

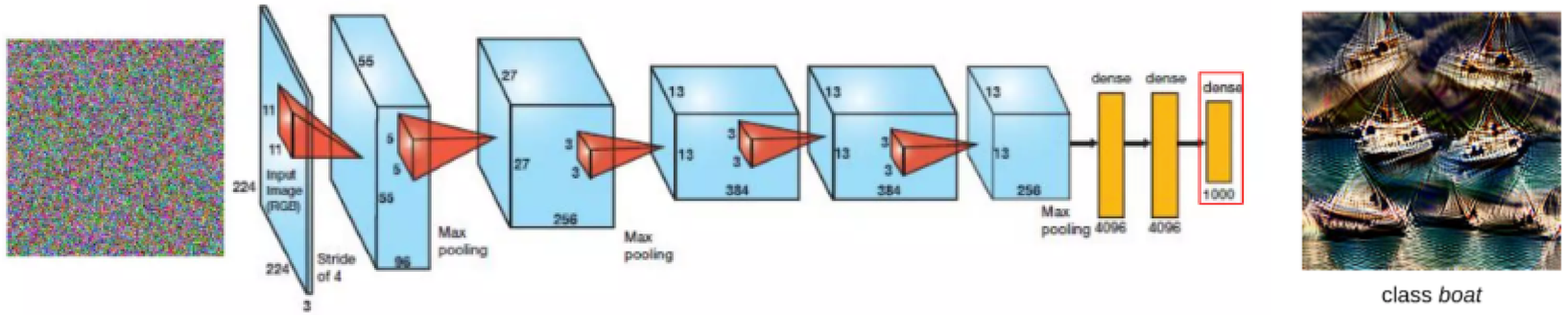
Activation Maximisation

Global method

- **Activation maximization** is an explanatory framework that generates an input that produces a maximum model response for a quantity of interest

How does a prototypical input look like that maximises a class "boat"?

1. Initialise the input image with random noise for the target class
2. Perform gradient descent to optimise the input image wrt target



class *boat*

"Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps" (Simonyan et al., 2014)

Activation Maximisation

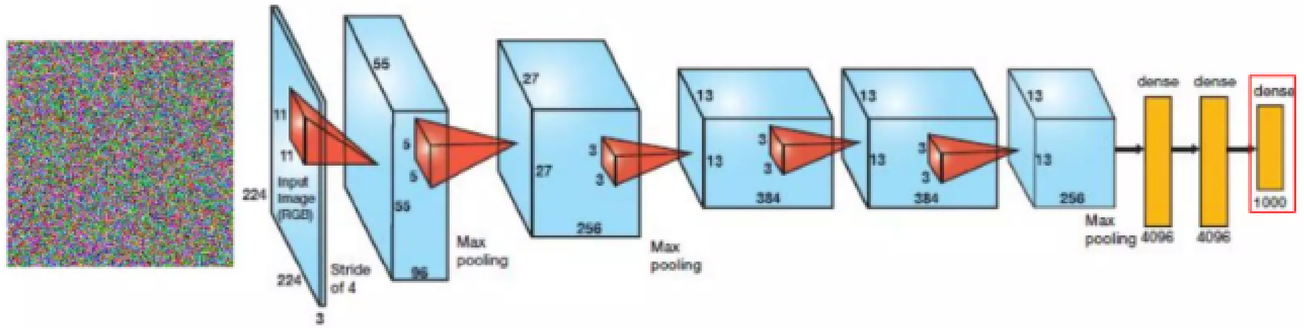
Global method

- **Activation maximization** is an explanatory framework that generates an input that produces a maximum model response for a quantity of interest

How does a prototypical input look like that maximises a class "boat"?

1. Initialise the input image with random noise for the target class
2. Perform gradient descent to optimise the input image wrt target

AM might be sensitive to its initialisation and difficult to scale



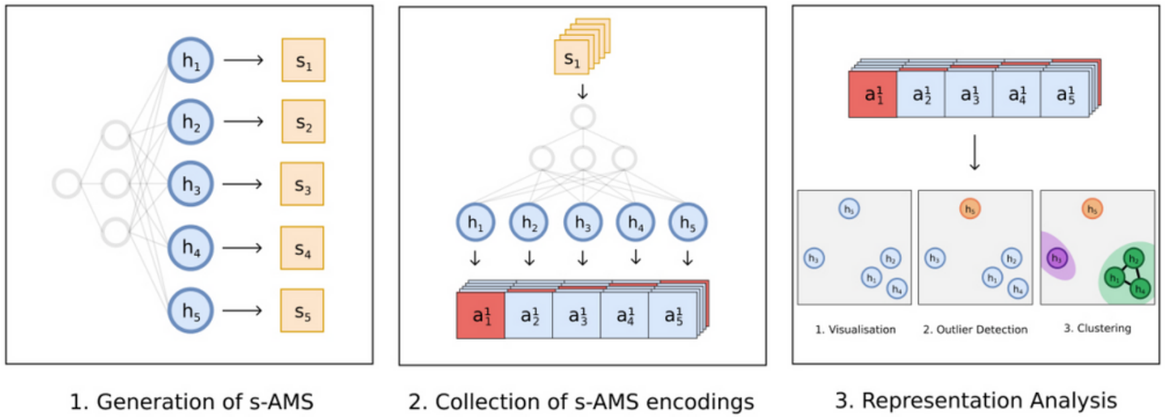
class *boat*

"Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps" (Simonyan et al., 2014)

DORA

Global method

- **DORA (Data-agnOstic Representation Analysis)** uses AM as a foundation, (1) collecting AMs from a layer and (2) feeding them through the network to produce embedding (activations) (3) analysing them e.g., with clustering



1. Generation of s-AMS
2. Collection of s-AMS encodings
3. Representation Analysis

"DORA: Exploring outlier representations in Deep Neural Networks" (Bykov et al., 2022)

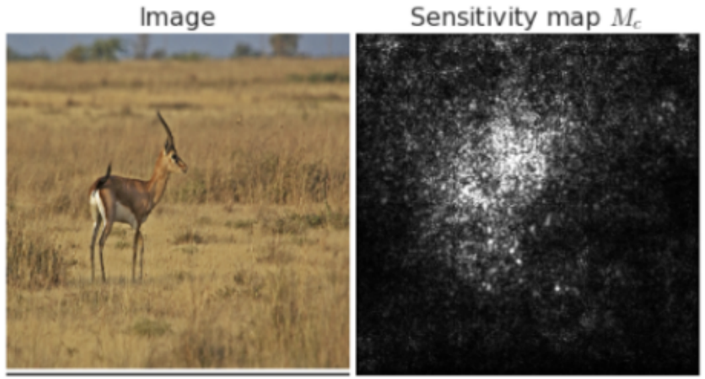
- **Idea:** Can be used to analyse the representations of the networks, e.g., detect outliers and controversial features

Combining/ enhancing methods

SmoothGrad

Enhancing explanations by adding noise to input

- A method called SmoothGrad was developed, aimed at reducing visual diffusion



A noisy sensitivity map, based on the gradient of the class score for gazelle for an image classification network. (Smilkov et al, 2017)

$$\frac{1}{N}\sum_{i=1}^{N} E\left(\boldsymbol{x}+\xi_i, f(\cdot,\hat{W})\right),$$

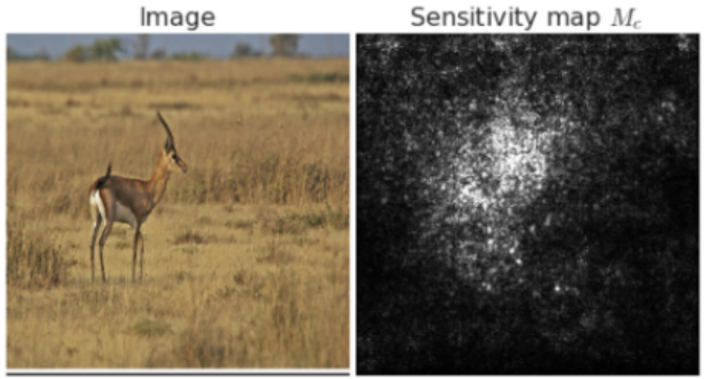
$$\xi_i \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathrm{SG}}^2\mathbf{I})$$

- Works by adding Gaussian noise to the input and takes the average over instances of noise

SmoothGrad

Enhancing explanations by adding noise to input

- A method called SmoothGrad was developed, aimed at reducing visual diffusion



A noisy sensitivity map, based on the gradient of the class score for gazelle for an image classification network. (Smilkov et al, 2017)

$$\frac{1}{N}\sum_{i=1}^{N} E\left(\boldsymbol{x}+\xi_i, f(\cdot,\hat{W})\right),$$

$$\xi_i \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathrm{SG}}^2\mathbf{I})$$

- Works by adding Gaussian noise to the input and takes the average over instances of noise

Instead of exploring the neighbourhood of the input, can we leverage the model space?

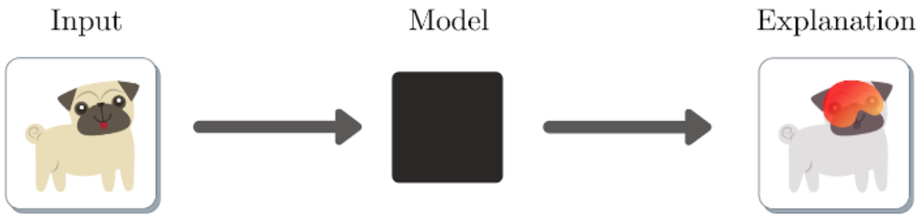
NoiseGrad

Enhancing explanations by introducing stochasticity to model weights

- We start with a baseline explanation

Baseline explanation

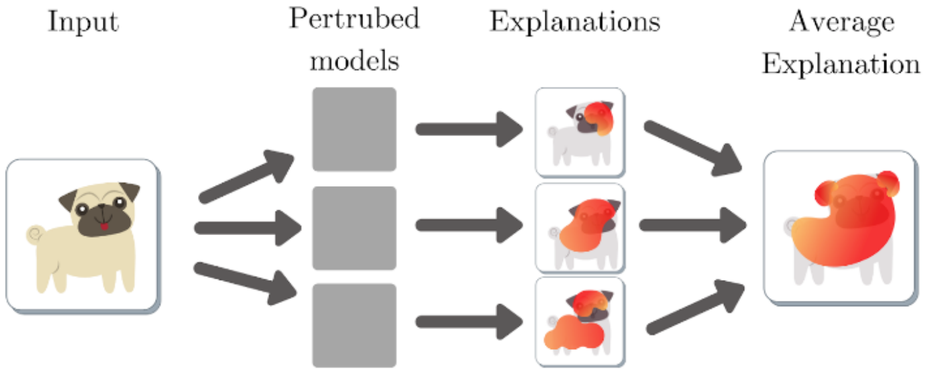
$$E\left(\boldsymbol{x}, f(\cdot, \hat{W})\right)$$



- Inspired by Bayesian learning, we approximate the posterior by drawing samples with multiplicative Gaussian noise added to the model's weights

NoiseGrad-enhanced explanation

$$\frac{1}{M}\sum_{i=1}^{M} E\left(\boldsymbol{x}, f(\cdot, \mathcal{W}_i)\right)$$

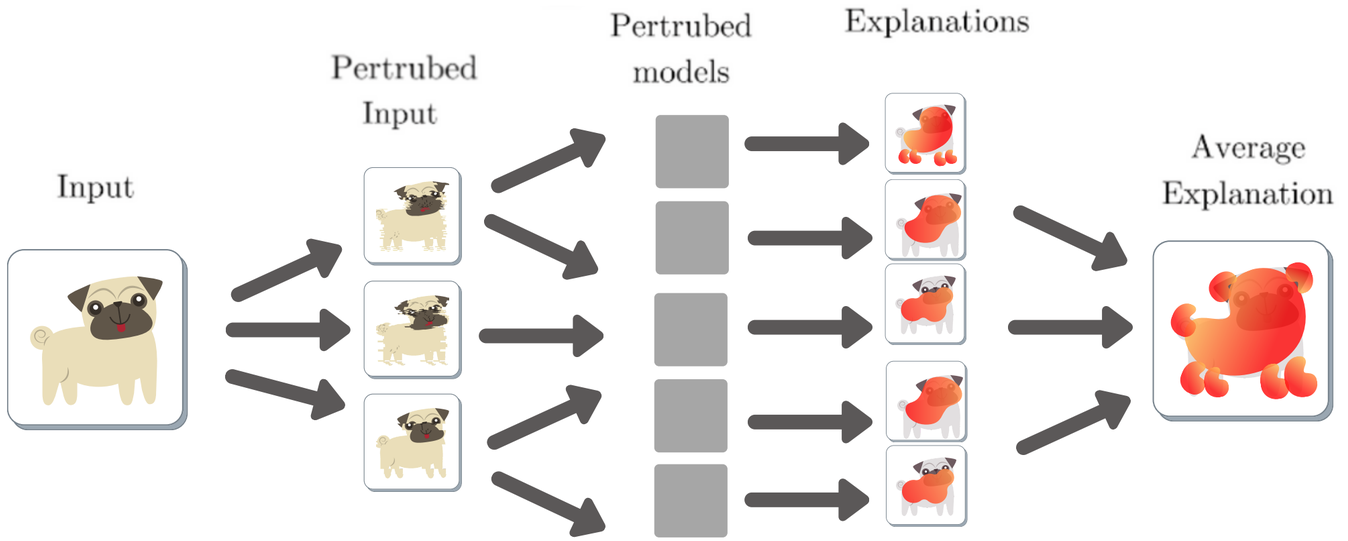


FusionGrad

Enhancing explanations by introducing stochasticity to model weights

- "Fused" NoiseGrad and SmoothGrad, creating FusionGrad

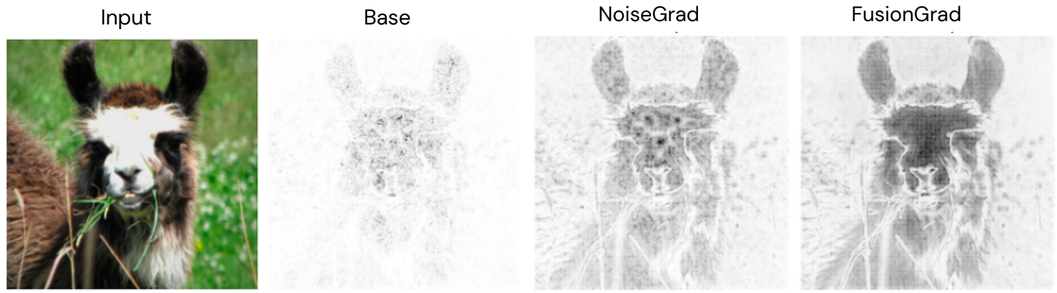
FusionGrad-enhanced explanation : $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M} E\left(\boldsymbol{x}+\xi_j, f(\cdot, \mathcal{W}_i)\right),$



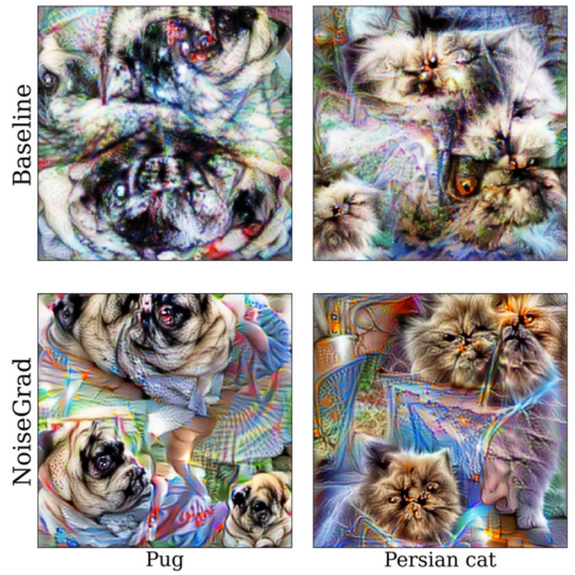
NoiseGrad and FusionGrad

Enhancing explanations by introducing stochasticity to model weights

- Explanations with NG and FG become significantly more localised, faithful and robust and:
  - Local (Saliency) explanations become more visually concise
  - Global (AM) explanations become more semantically meaningful



Kirill and Hedström et al„ *NoiseGrad — Enhancing Explanations by Introducing Stochasticity to Model Weights*, Proceedings of the AAAI Conference on Artificial Intelligence (2022)



AM explanations for ImageNet classes, ResNet-18 model

Self-explaining methods

Post-hoc explanations “must be wrong”; that they are by definition not completely faithful to the original model and must be less accurate with respect to the primary task.

"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" by Rudin (2019)

Post-hoc explanations “must be wrong”; that they are by definition not completely faithful to the original model and must be less accurate with respect to the primary task.

"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" by Rudin (2019)

But what do you do if you can't build your model from scratch?

Challenges

What are some challenges that the field is experiening?

Failure mode #1

Explanations can be manipulated

- Turns out that local methods are far from fault-proof

"Explanations can be manipulated and geometry is to blame" (Dombrowski, 2019)

Failure mode #1

Explanations can be manipulated

- **Idea:** similar to how adversarial attacks manipulate the model, manipulate explanations by optimising a customised loss function of the model using simple gradient descent (Dombrowski, 2019)

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

Create an adversarial example such that the norm of the input image is small << 1

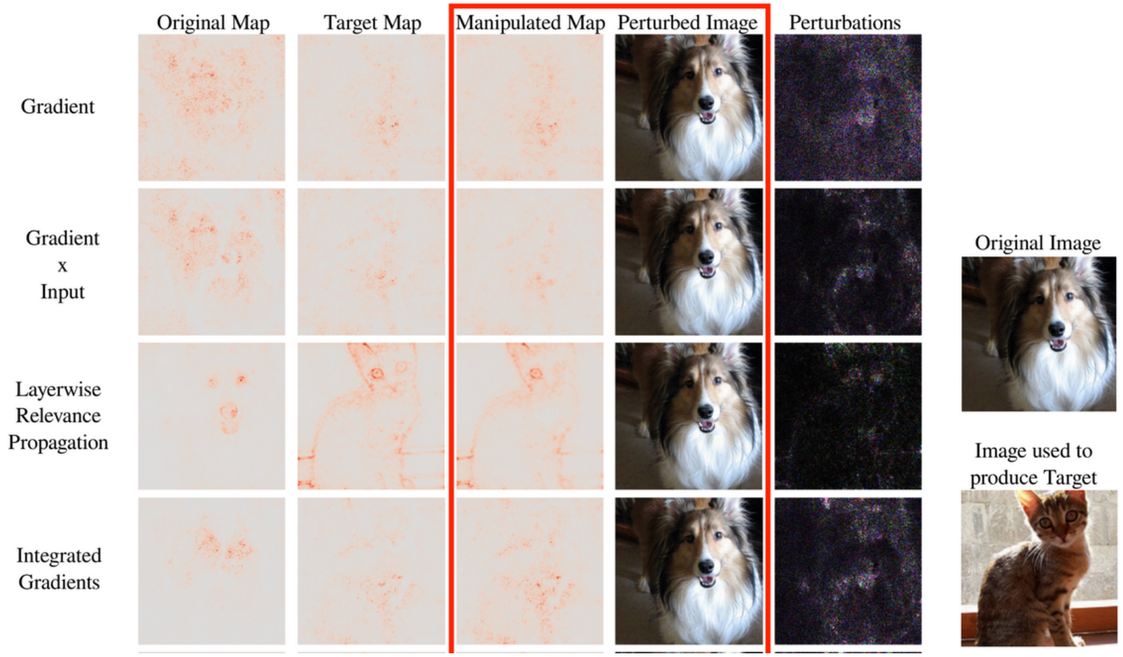
Ensure that the manipulated explanation map is close to the target

Encourage the network to have the same output

Failure mode #1

Explanations can be manipulated

- **Result:** While perturbations applied to the input can be visually indistinguishable, the result can be completely different explanations

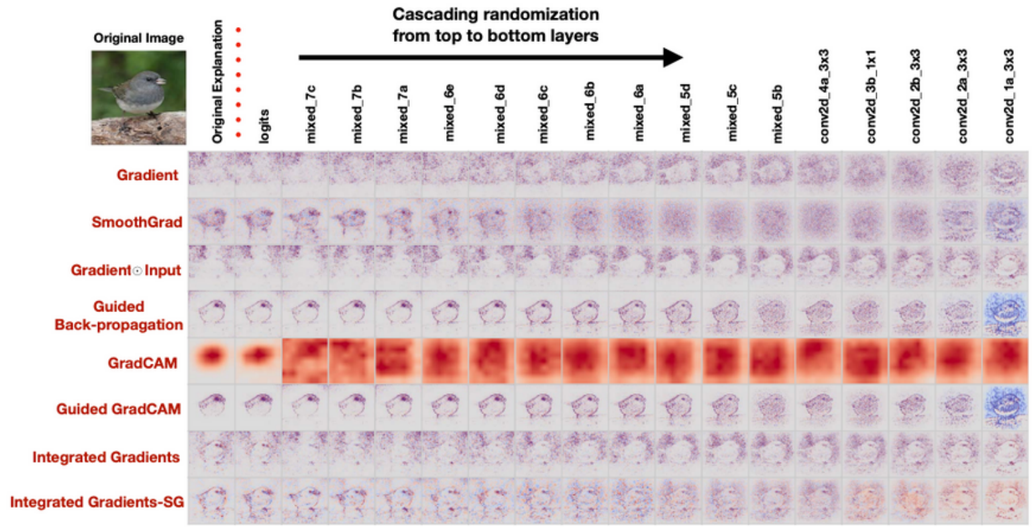


"Explanations can be manipulated and geometry is to blame" (Dombrowski, 2019)

Failure mode #2

Explanations are invariant to model parameters

- **Idea:** Randomise the parameters from top to bottom layers in a cascading way, and measure the distance of the resulting explanation to the original explanation (Adebayo et. al., 2018)



Seeking large difference in explanation outcomes between ***f*** and ***f'***

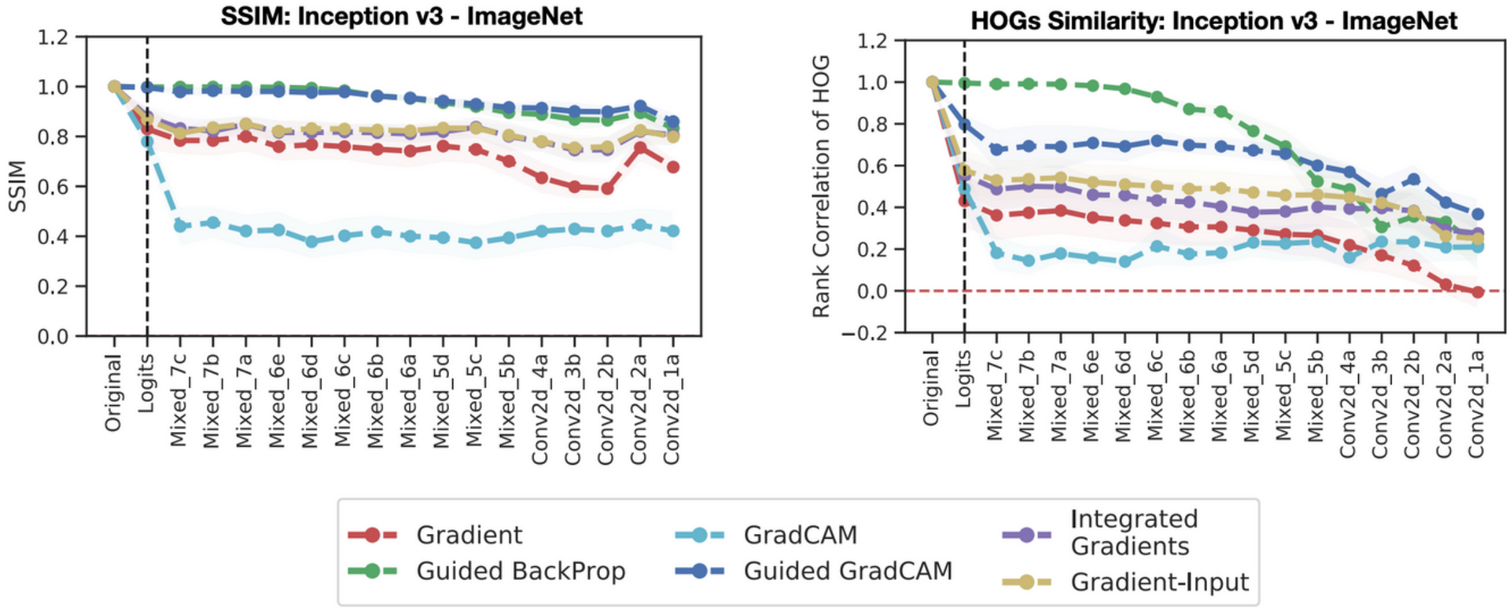
"Sanity Checks for Saliency Maps" (Adebayo et al., 2018)

Failure mode #2

Explanations are invariant to model parameters

- **Result:** Contrary to intuition, showing that explanations for accurate and random models are similar (high SSIM, high Rank Correlation)

Lower values are better meaning explanations are dis-similar



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

Failure mode #3

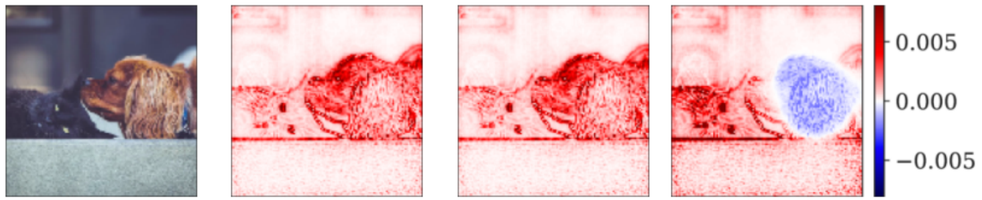
Explanations are invariant to class ("logit")

- **Idea:** Explanations are sensitive to its class: compute the distance between the original explanation and the explanation for a random logit (Sixt et. al., 2020)

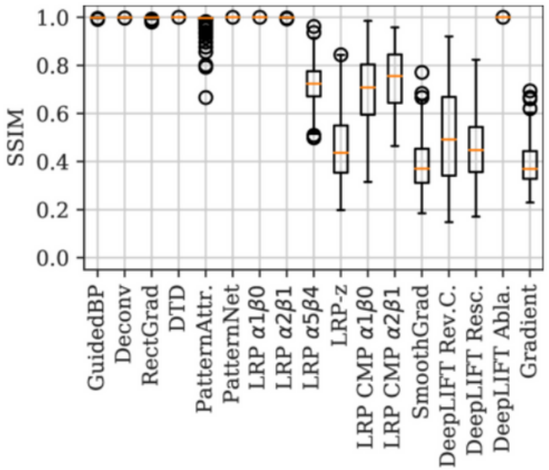
Failure mode #3

Explanations are invariant to class ("logit")

- **Result:** Right: LRP explanation insensitive where the difference between classes are small norm difference and Left: high SSIM



(b) Image (c) Expl. Cat (d) Expl. Dog (e) Diff. (c) - (d)



"When Explanations Lie: Why Many Modified BP Attributions Fail" (Sixt et al., 2020)

But are these findings stable enough to rule out explanation methods?

These studies paints a dark picture

But there is more to unpack

Rebuttal #1

Newly published work that sanity check existing sanity checks

- **Issue:** SSIM between any two attribution maps can be minimised by statistically uncorrelated and uninformative random processes

Presence of covariance term, with zero pixel-wise covariance easily would "pass" the tests

$$\frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1}\frac{2\sigma_{AB} + C_2}{\sigma_A^2 + \sigma_B^2 + C_2}.$$

- **Issue:** Due to the presence of skip connections in e.g., ResNets, explanations are only expected to differ to a limited extent since activations are maintained after randomisation

Rebuttal #1

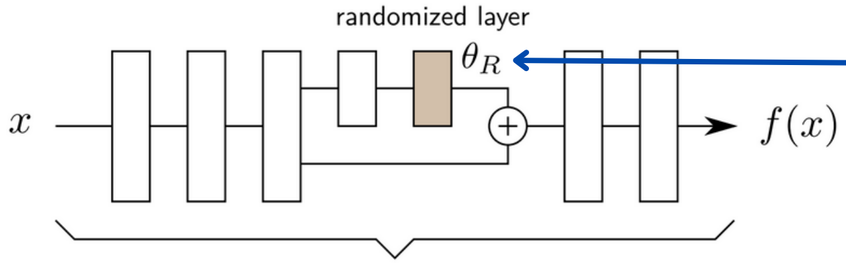
Newly published work that sanity check existing sanity checks

- **Issue:** SSIM between any two attribution maps can be minimised by statistically uncorrelated and uninformative random processes

Presence of covariance term, with zero pixel-wise covariance would easily "pass" the tests

$$\frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1}\frac{2\sigma_{AB} + C_2}{\sigma_A^2 + \sigma_B^2 + C_2}.$$

- **Issue:** Due to the presence of skip connections in e.g., ResNets, explanations are only expected to differ to a limited extent since activations are maintained after randomisation



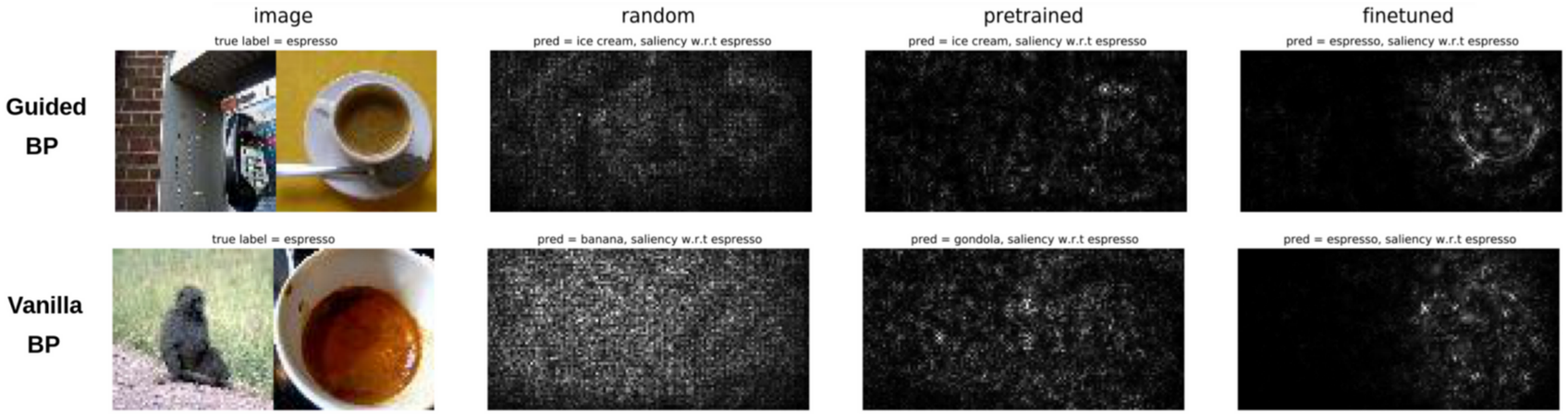
Simply replicating the activation and gradients from layer to layer

"Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations" (Binder et al., 2022)

Rebuttal #2

Newly published work that sanity check existing sanity checks

- **Issue:** The choice of task (and dataset) might have affected the results, authors claiming the presence of empirical confounds in Model Parameter Randomsiation test



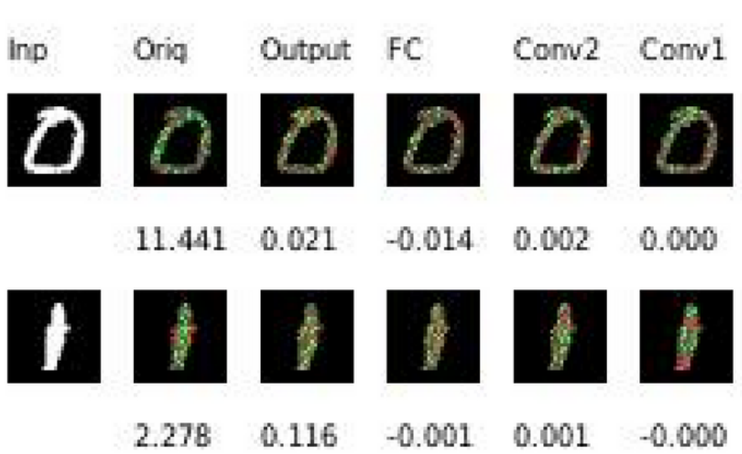
"Revisiting Sanity Checks for Saliency Maps" (Yona et al., 2021)

On the fine-tuned model,
attributional evidence is correctly absent
in the 'non-espresso' area

Rebuttal #3

Newly published work that sanity check existing sanity checks

- **Issue:** The choice of post-processing explanations may affect the results



Shows that if Integrated Gradient had signed attributions, then the MPR test would have passed and not failed (showing sum of IG attr)

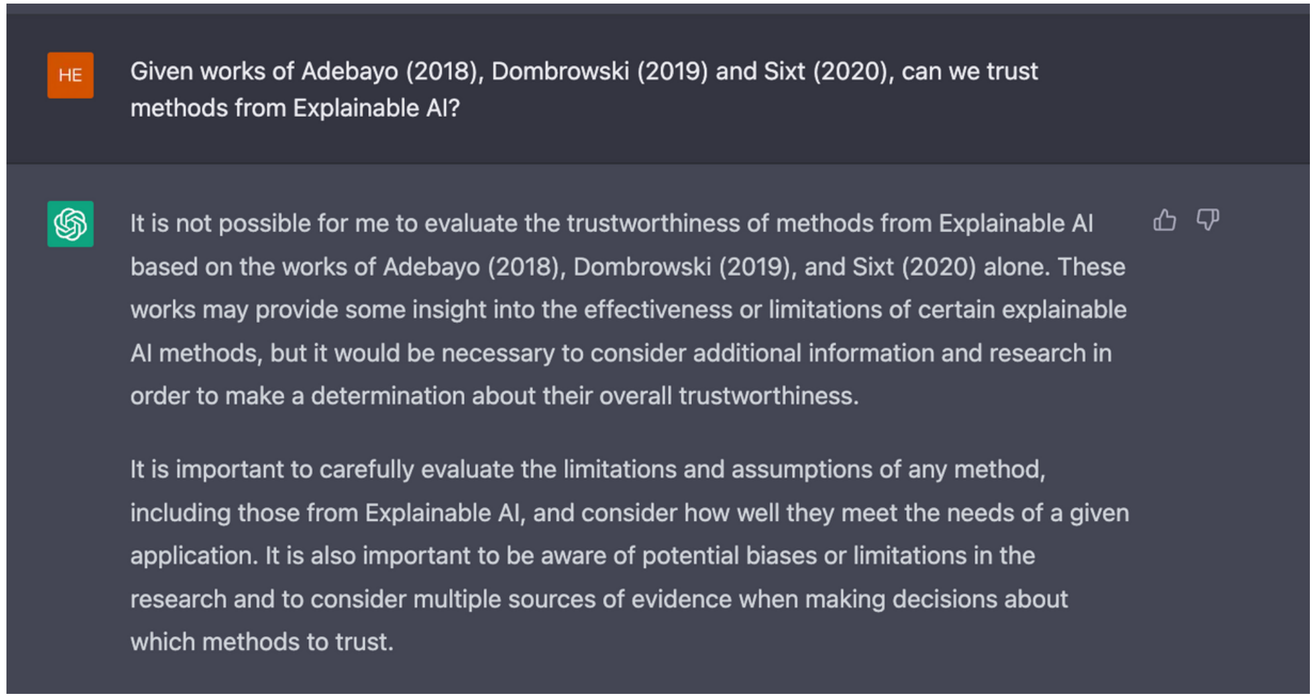
"A Note about: Local Explanation Methods for Deep Neural Networks lack Sensitivity to Parameter Values" (Sundararajan et al., 2018)

It looks worringly "easy to get it wrong"

Explanation, can I trust you?

We return back to the question

- Due to the presence of strong assertions about which methods work (and not), followed by rebuttals, this question remains difficult to answer



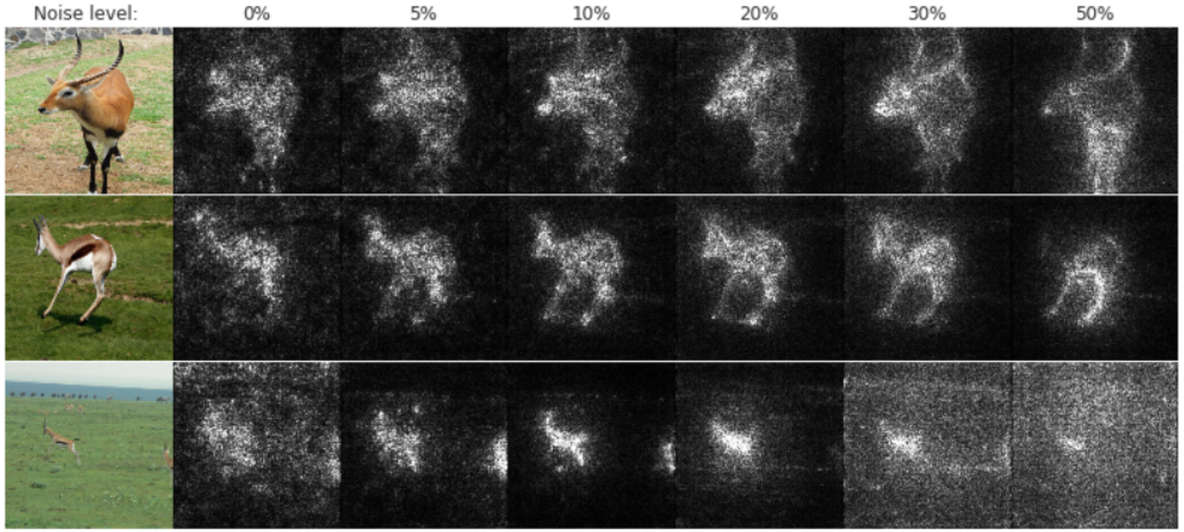
What can we do to tackle this?

Evaluation

Starting point for evaluation

An overreliance on qualitative evaluation methods

- **Idea:** Principal focus on qualitative evaluation, incl. influential papers (>1400)



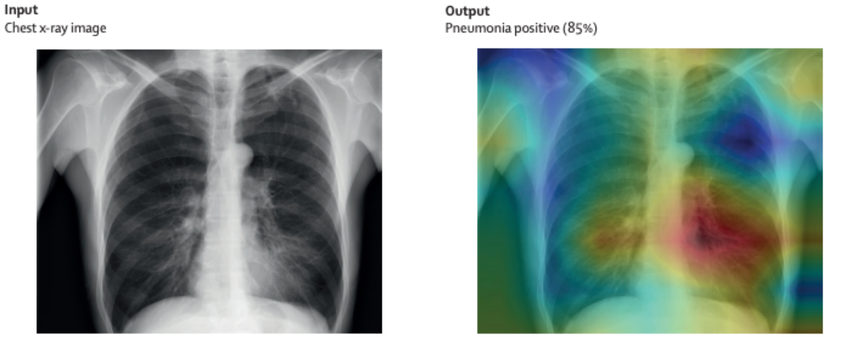
"SmoothGrad: removing noise by adding noise" (Smilkov et al, 2017)

- **Result:** Explanation methods were proposed with little to no performance guarantees, evaluating only "visual coherence" does not provide a satisfactory answer

Why qualitative methods fail

Because users misinterpret and over-trust AI explanations

- A great deal of evidence that humans tend to **ascribe a positive interpretation** to AI explanations (Ghassemi et al., 2021), and even **data scientists over-trust and misuse** interpretability tools (Kaur et al., 2020)



"CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning" (Rajpurkar et al, 2017)

- Issue of **normative vs descriptive explanations**: is its presence of (i) airspace opacity, (ii) the shape of the heart border or (iii) the left artery that contributed to the prediction?

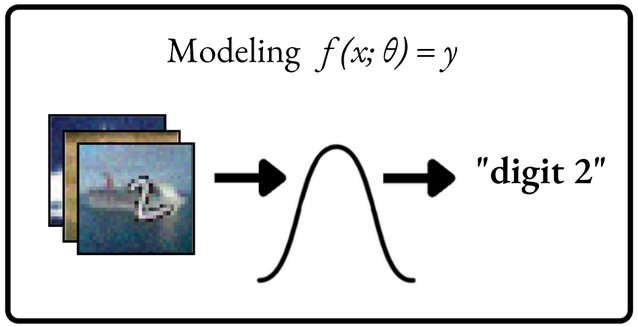
How do identify low- vs high-quality explanations?

Dislcaimer: it is not straight-forward

Step 1. Modeling

What makes evaluating XAI challenging?

- Consider a (simple) image classification task where a ResNet9 model is trained to classify MNIST digits from 0 to 9 pasted on CIFAR-10 backgrounds

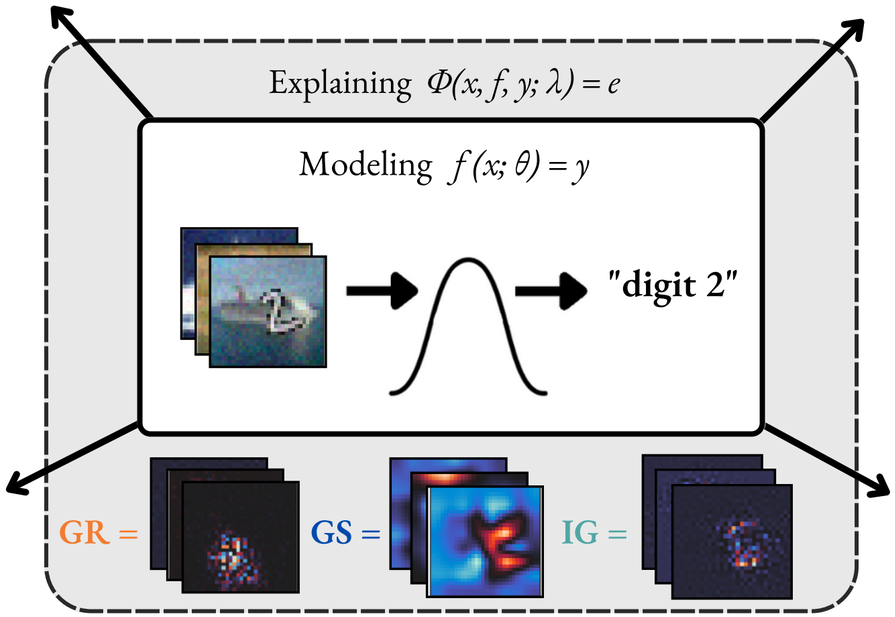


- We evaluate the "goodness" of the model by comparing labels $y = y_{true}$

Step 2. Explaining

What makes evaluating XAI challenging?

- To understand which parts of the input contributed to the prediction, we **apply several explanation methods**



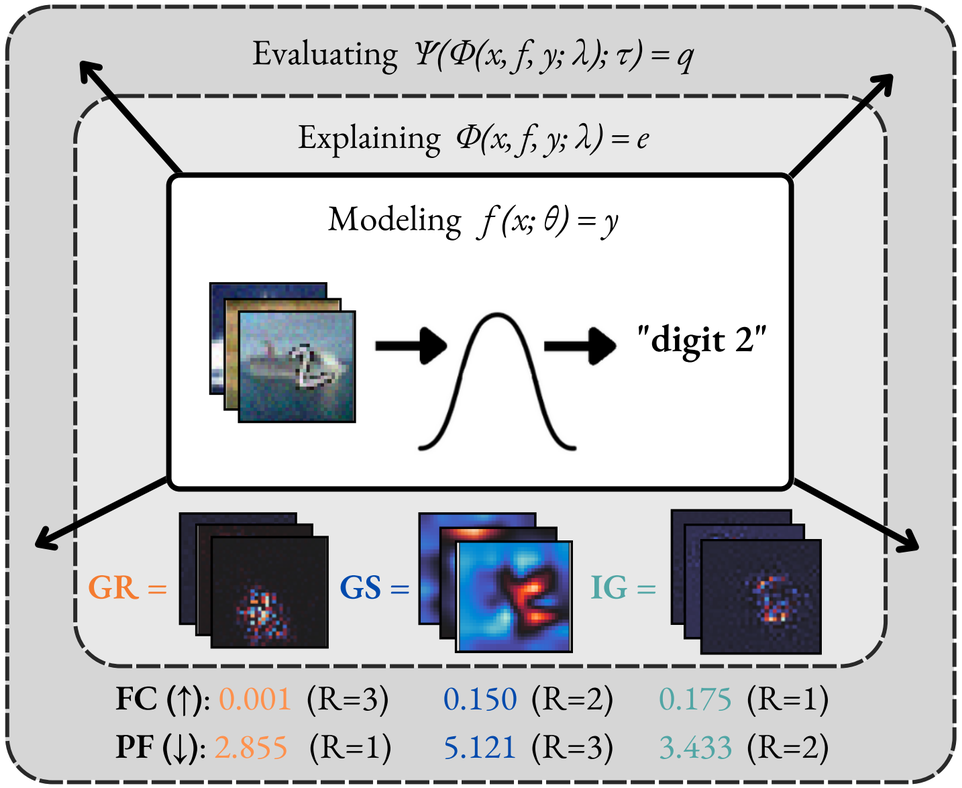
- Due to the black-box nature of the model (exceptions are linear models
- and shallow decision trees), explanations cannot be verified

$$e = e_{true}$$

Step 3. Evaluating

What makes evaluating XAI challenging?

- To evaluate the **quality of the explanation method**, we apply different metrics



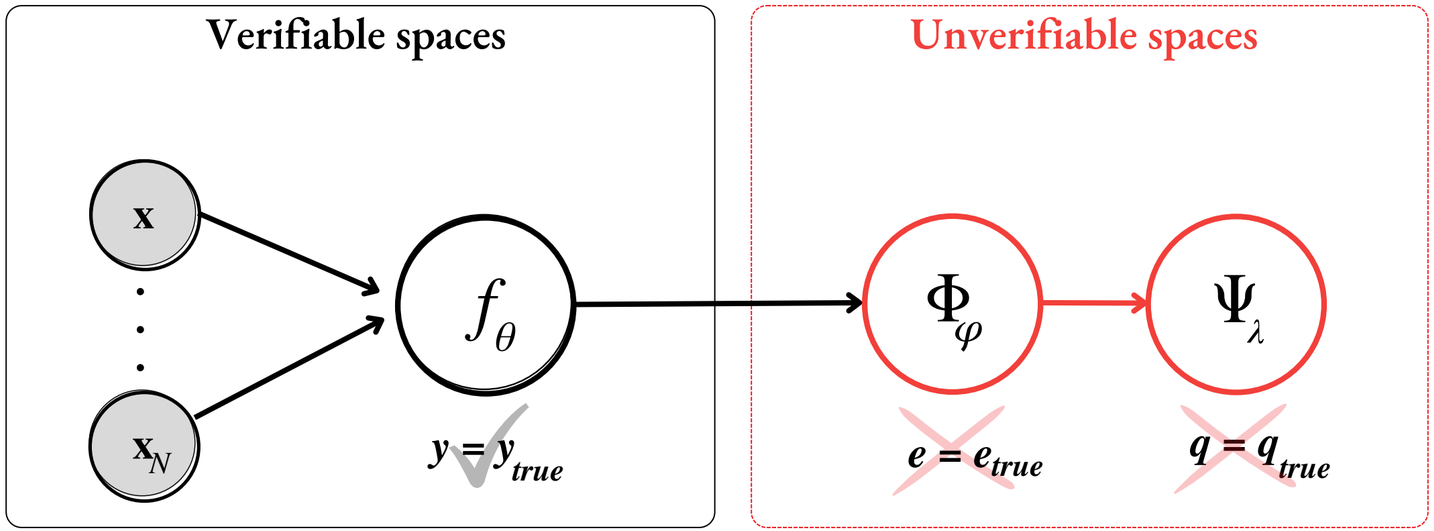
$$q = q_{true}$$

What makes evaluating XAI challenging?

The Challenge of Unverifiability

Evaluation is a conditional process from modeling to evaluating

- A key observation is that in XAI evaluation there exists a **conditional dependency** between the variables of modelling, explaining and evaluating



- **Uncertainty propagates through the Directed Acyclic Graph (DAG):** If a parent node is unverifiable, then the child node renders unverifiable

What are some different perspectives of explanation quality?

Defining explanation *quality*

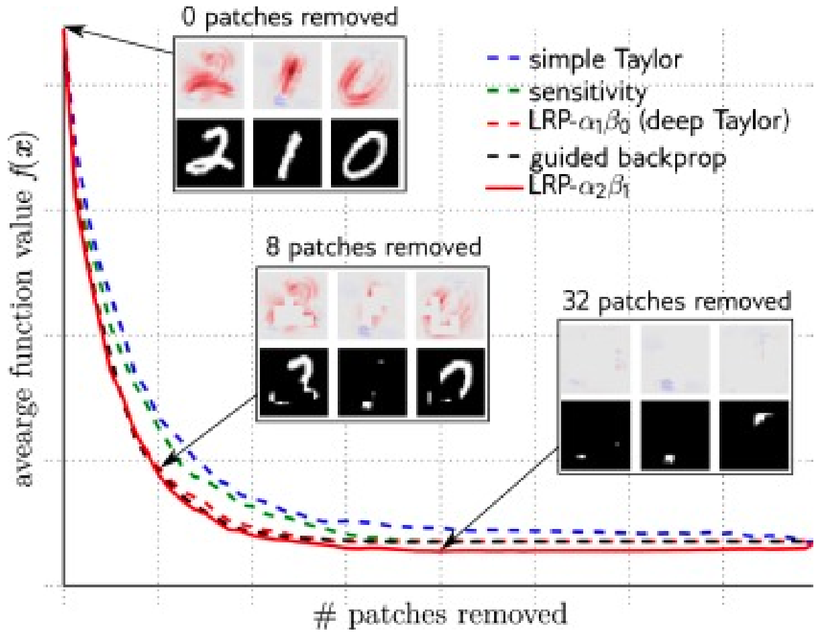
Test the relative fulfilment of human-desirable properties

- **Faithfulness** (↑) quantifies to what extent explanations follow the predictive behaviour of the model, asserting that more important features affect model decisions more strongly
- **Robustness** (↓) measures to what extent explanations are stable when subject to slight perturbations in the input, assuming that the model output approximately stayed the same
- **Randomisation** (↑) tests to what extent explanations deteriorate as the data labels or the model, e.g., its parameters are increasingly randomised
- **Localisation** (↑) tests if the explainable evidence is centred around a region of interest, which may be defined around an object by a bounding box, a segmentation mask or a cell within a grid
- **Complexity** (↓) captures to what extent explanations are concise, i.e., that few features are used to explain a model prediction
- **Axiomatic** (↑) measures if explanations fulfill certain axiomatic properties

Defining explanation quality

Test the relative fulfilment of human-desirable properties

- **Faithfulness** (↑) quantifies to what extent explanations follow the predictive behaviour of the model, asserting that more important features affect model decisions more strongly
- **Robustness** (↓) measures to what extent explanations are stable when subject to slight perturbations in the input, assuming that the model output approximately stayed the same
- **Randomisation** (↑) tests to what extent explanations deteriorate as the data labels or the model, e.g., its parameters are increasingly randomised



Pixel-Flipping (Bach et al., 2015): captures the impact of perturbing pixels in descending order according to the attributed value on the classification score
Image source: (Montavon, 2018)

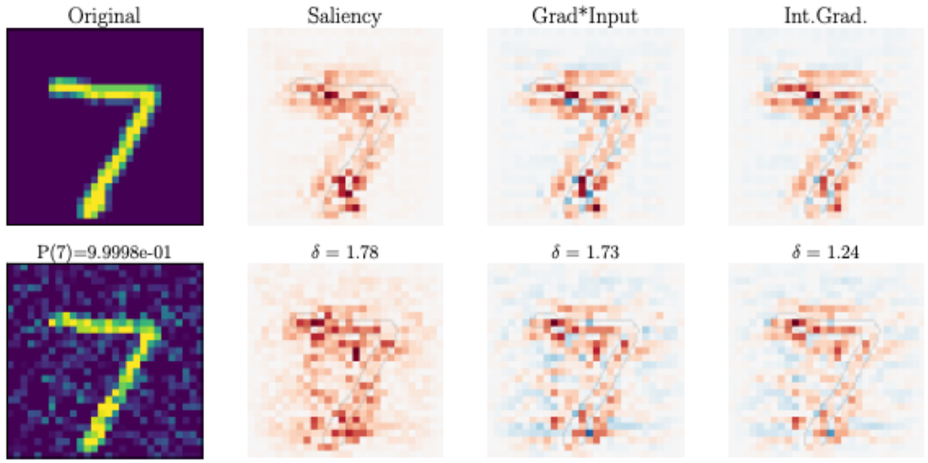
Defining explanation quality

Test the relative fulfilment of human-desirable properties

- **Faithfulness** (↑) quantifies to what extent explanations follow the predictive behaviour of the model, asserting that more important features affect model decisions more strongly

- **Robustness** (↓) measures to what extent explanations are stable when subject to slight perturbations in the input, assuming that the model output approximately stayed the same

- **Randomisation** (↑) tests to what extent explanations deteriorate as the data labels or the model, e.g., its parameters are increasingly randomised



Local Lipschitz Estimate (Alvarez-Melis et al., 2018): tests the consistency in the explanation between adjacent examples

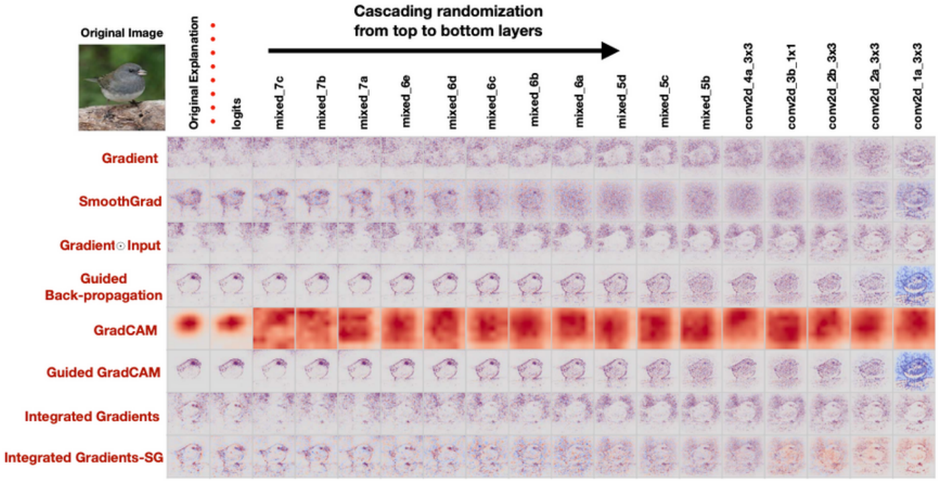
Defining explanation quality

Test the relative fulfilment of human-desirable properties

- **Faithfulness** (↑) quantifies to what extent explanations follow the predictive behaviour of the model, asserting that more important features affect model decisions more strongly

- **Robustness** (↓) measures to what extent explanations are stable when subject to slight perturbations in the input, assuming that the model output approximately stayed the same

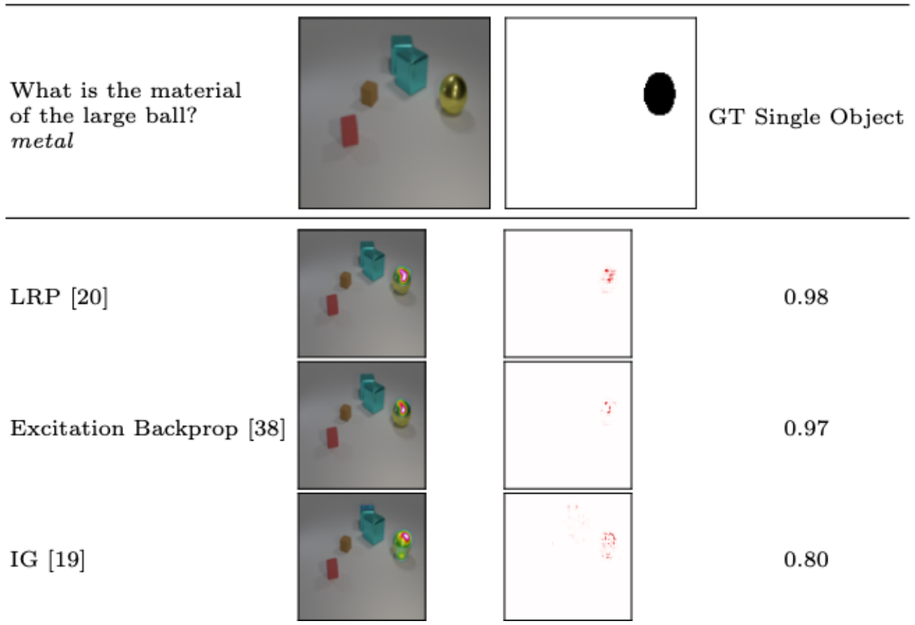
- **Randomisation** (↑) tests to what extent explanations deteriorate as the data labels or the model, e.g., its parameters are increasingly randomised



Model Parameter Randomisation (Adebayo et. al., 2018): randomises the parameters of single model layers in a cascading or independent way and measures the distance of the respective explanation to the original explanation

Defining explanation quality

Test the relative fulfilment of human-desirable properties



Relevance Mass Accuracy (Arras et al., 2021): measures the ratio of positively attributed attributions inside the ground-truth mask towards the overall positive attributions

- **Localisation** (↑) tests if the explainable evidence is centred around a region of interest, which may be defined around an object by a bounding box, a segmentation mask or a cell within a grid
- **Complexity** (↓) captures to what extent explanations are concise, i.e., that few features are used to explain a model prediction
- **Axiomatic** (↑) measures if explanations fulfill certain axiomatic properties

Defining explanation quality

Test the relative fulfilment of human-desirable properties

Definition 4 (Complexity)**.** Given a predictor $\boldsymbol{f}$, explanation function $\boldsymbol{g}$, and a point $\boldsymbol{x}$, the complexity of $\boldsymbol{g}$ at $\boldsymbol{x}$ is:

$$\mu_{\mathrm{C}}(\boldsymbol{f}, \boldsymbol{g}; \boldsymbol{x}) = \mathbb{E}_i\big[-\ln(\mathbb{P}_{\boldsymbol{g}})\big] = -\sum_{i=1}^{d} \mathbb{P}_{\boldsymbol{g}}(i)\ \ln(\mathbb{P}_{\boldsymbol{g}}(i))$$

Complexity (Bhatt et al., 2020): computes the entropy of the fractional contribution of all features to the total magnitude of the attribution individually

http://www.heatmapping.org/slides/2018_ICIP_3.pdf

- **Localisation** (↑) tests if the explainable evidence is centred around a region of interest, which may be defined around an object by a bounding box, a segmentation mask or a cell within a grid

- **Complexity** (↓) captures to what extent explanations are concise, i.e., that few features are used to explain a model prediction

- **Axiomatic** (↑) measures if explanations fulfill certain axiomatic properties

Defining explanation quality

Test the relative fulfilment of human-desirable properties

$$\sum_i R_i = f(\boldsymbol{x})$$

Completeness (Sundararajan et al., 2017): evaluates whether the sum of attributions (or "relevances") is equal to the function value

- **Localisation** (↑) tests if the explainable evidence is centred around a region of interest, which may be defined around an object by a bounding box, a segmentation mask or a cell within a grid
- **Complexity** (↓) captures to what extent explanations are concise, i.e., that few features are used to explain a model prediction
- **Axiomatic** (↑) measures if explanations fulfill certain axiomatic properties

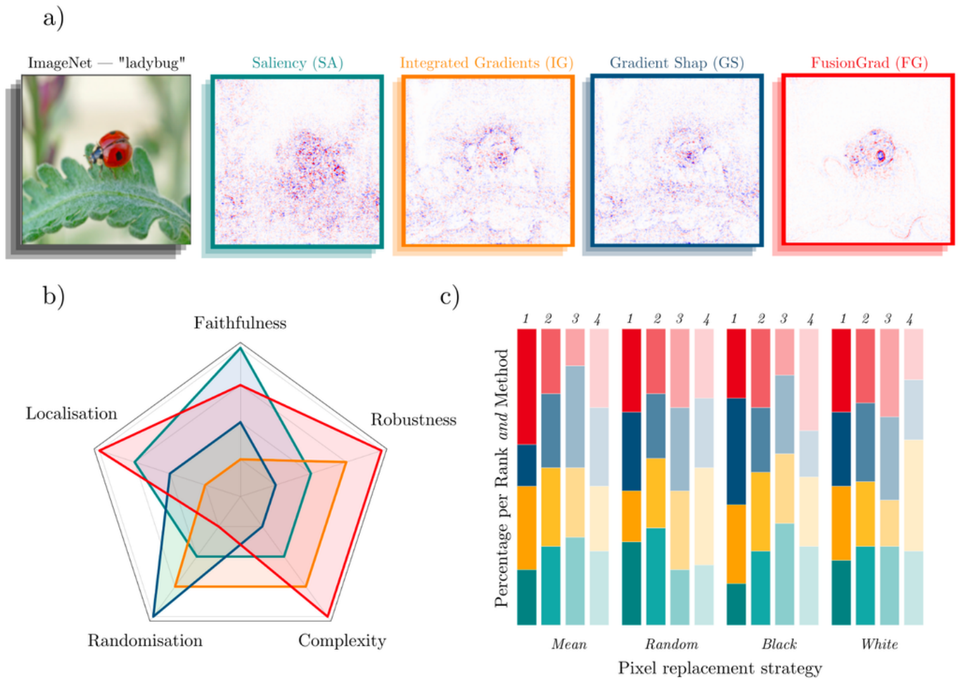
How can we automate evaluation of explainable methods?

Quantus, motivation

The goal of the project is to automate performance evaluation of XAI methods

- **For comparative analysis**
  a) visual inspection usually falls short

- **Quantus fills this gap, providing**
  b) a holistic snapshot of how different explanation methods rank
  c) ways to investigate the influence of metrics' parameterisation on ranking



"Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations" (Hedström et al., 2022)

Library content

A XAI quantification open-source XAI toolkit for ML practitioners and deep learners

- Provides more than **30+ metrics** in 6 categories for XAI evaluation
- Includes **tutorials and API reference** for developers and non-developers
- Implements an abstract layer for popular deep learning frameworks: **PyTorch** and **Tensorflow**
- Supports different **data types** (image, time-series, tabular), NLP next up!

Table 1: Comparison of four XAI libraries — (`AIX360` [2], `captum` [29], `TorchRay` [30] and `Quantus`) in terms of the number of XAI evaluation methods for six different evaluation categories, as implemented in each library.

Library	Faithfulness	Robustness	Localisation	Complexity	Axiomatic	Randomisation
`Captum` (2)	1	1	0	0	0	0
`AIX360` (2)	2	0	0	0	0	0
`TorchRay` (1)	0	0	1	0	0	0
`Quantus` (27)	**9**	**4**	**6**	**3**	**3**	**2**

"Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations" (Hedström et al., 2022)

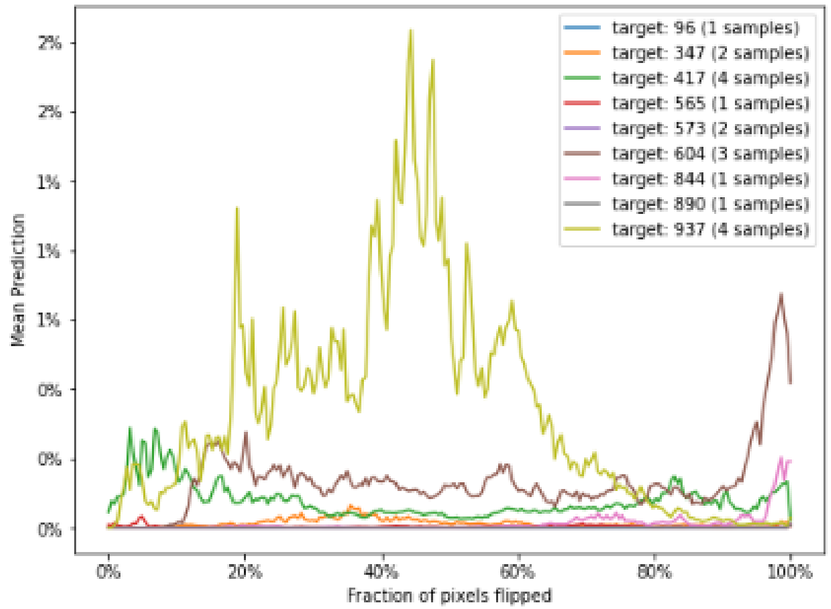
Syntax: simple but customisable

Evaluate XAI methods in a one-liner or compute scores with quantus.evaluate()

```
[ ]
# Create the pixel-flipping experiment.
pixel_flipping = quantus.PixelFlipping(
    features_in_step=224,
    perturb_baseline="black",
    perturb_func=quantus.baseline_replacement_by_indices,
)

# Call the metric instance to produce scores.
scores = pixel_flipping(model=model,
                        x_batch=x_batch,
                        y_batch=y_batch,
                        a_batch=a_batch,
                        device=device,)

# Plot example!
pixel_flipping.plot(y_batch=y_batch, scores=scores)
```



__init__ the metric in one go

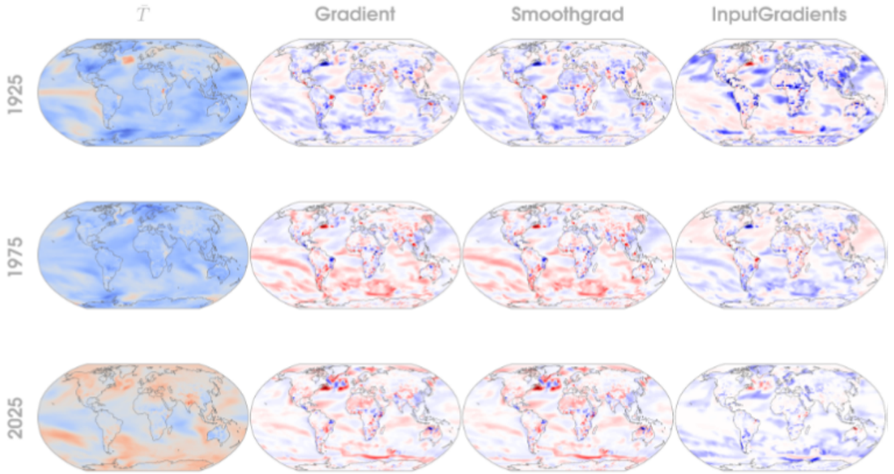
plot() to visualise some results

score xAI methods using **__call__**

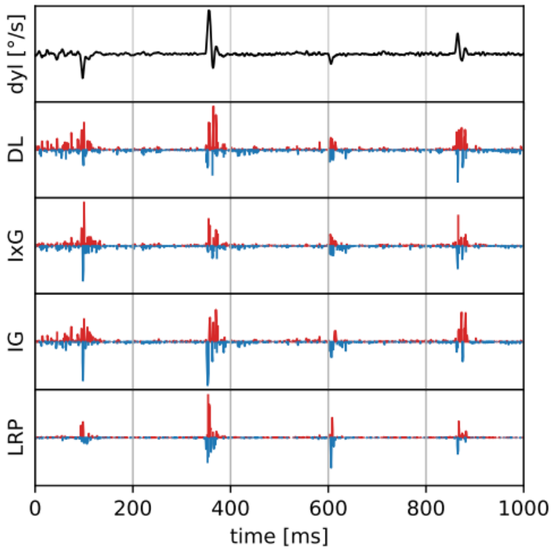
EVALUATION

Applications of Quantus

Evaluating explanations in climate and healthcare



Evaluate AI explanations for predictions of temperature over different decades 1920-2080 using MLPs and CNNs models (Bommer et al., unpublished)

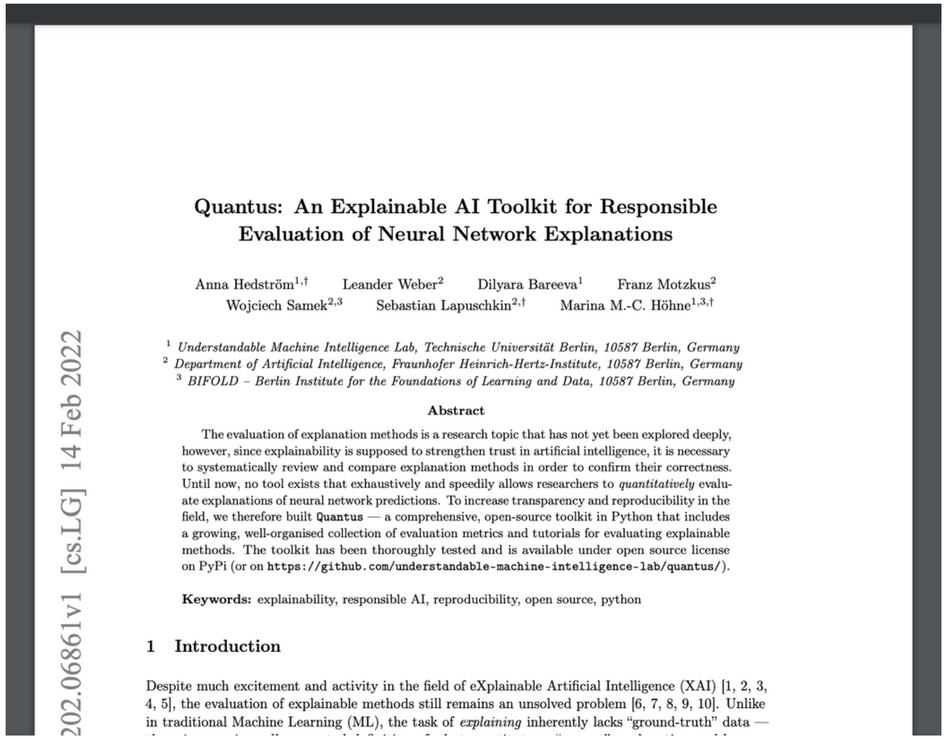


(c) pitch velocities of left eye

Evaluate explanation methods for eye movement-based biometric models, with time-series data (Krakowczyk et al., 2022)

Learn more

JMLR paper (preprint available on arxiv), code at Github and API documentation



Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations

Anna Hedström[1,†] Leander Weber[2] Dilyara Bareeva[1] Franz Motzkus[2]
Wojciech Samek[2,3] Sebastian Lapuschkin[2,†] Marina M.-C. Höhne[1,3,†]

[1] *Understandable Machine Intelligence Lab, Technische Universität Berlin, 10587 Berlin, Germany*
[2] *Department of Artificial Intelligence, Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany*
[3] *BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*

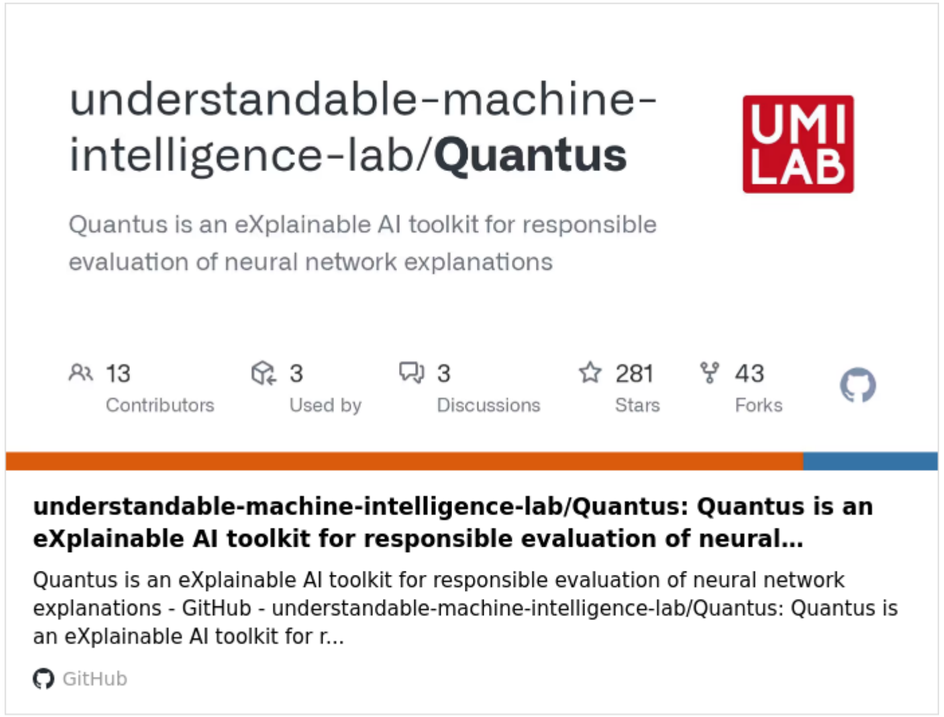
Abstract

The evaluation of explanation methods is a research topic that has not yet been explored deeply, however, since explainability is supposed to strengthen trust in artificial intelligence, it is necessary to systematically review and compare explanation methods in order to confirm their correctness. Until now, no tool exists that exhaustively and speedily allows researchers to *quantitatively* evaluate explanations of neural network predictions. To increase transparency and reproducibility in the field, we therefore built `Quantus` — a comprehensive, open-source toolkit in Python that includes a growing, well-organised collection of evaluation metrics and tutorials for evaluating explainable methods. The toolkit has been thoroughly tested and is available under open source license on PyPi (or on `https://github.com/understandable-machine-intelligence-lab/quantus/`).

Keywords: explainability, responsible AI, reproducibility, open source, python

1 Introduction

Despite much excitement and activity in the field of eXplainable Artificial Intelligence (XAI) [1, 2, 3, 4, 5], the evaluation of explainable methods still remains an unsolved problem [6, 7, 8, 9, 10]. Unlike in traditional Machine Learning (ML), the task of *explaining* inherently lacks "ground-truth" data —



What to contribute? Check out our issues!

Other research directions

Additional research directions

What are some suggestions to circumvent the missing ground truth?

- Various ideas have been proposed:
  - Simulate labels (Yang et al., 2019; Arras et al., 2020), known artefacts (Zhou et al, 2021) or actual truths with MRI data (Jin et al., 2022)
  - Measuring relative fulfilment against human-determined properties (Hedström et al., 2022; Nguyen et al., 2020; Montavon et al., 2018)
  - Axiomatic evaluation, free from empirical interpretations (Sundararajan et al., 2017; Kindermans et al., 2017)
  - Relying on self-explainable built-in mechanisms (Ghorbani et al., 2019; Rudin, 2019)

Summary

Summary

Navigating Explainable AI

- AI will have an extraordinary impact on society in the coming decades, and we should do all we can to ensure that it is implemented in a way that maximises our benefit

One route is explainability

- Explainability has the appeal of solving pressing problems for high-stake domains in the sciences, medicine, NLP and others

But in contrast to traditional ML, XAI is ambiguously defined

- Efforts are diverse, applications many, and researchers optimise against different targets, consequently, getting an overview of SOTA is difficult

Evaluation is a remaining challenge

Post-script

Thank you

Contact — hedstroem.anna@gmail.com

Twitter — https://twitter.com/anna_hedstroem

Understandable Machine Intelligence Lab — https://twitter.com/TUBerlin_UMI

Interpretability courses

- Intro XAI course '22 https://introinterpretableai.wordpress.com/
- NeuIPS'20 Tutorial https://explainml-tutorial.github.io/neurips20
- AAAI'21 Tutorial https://explainml-tutorial.github.io/aaai21

"Explanation separates us from astonishment,
which is the only gateway to the incomprehensible."

Eugene Ionesco, 1912-1994 French-Romanian playwright

References

References 1/4

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

Arras, Leila, Ahmed Osman, and Wojciech Samek. "CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations." Information Fusion 81 (2022): 14-40.

Arcadu, Filippo, et al. "Deep learning algorithm predicts diabetic retinopathy progression in individual patients." NPJ digital medicine 2.1 (2019): 1-9.

Alvarez-Melis, David, and Tommi S. Jaakkola. "On the robustness of interpretability methods." arXiv preprint arXiv:1806.08049 (2018).

Balduzzi, David, et al. "The shattered gradients problem: If resnets are the answer, then what is the question?." International Conference on Machine Learning. PMLR, 2017.

Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.

Bhatt, Umang, Adrian Weller, and José MF Moura. "Evaluating and aggregating feature-based model explanations." arXiv preprint arXiv:2005.00631 (2020).

References 2/4

Binder, Alexander, et al. "Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations." arXiv preprint arXiv:2211.12486 (2022).

Bykov, Kirill, et al. "DORA: Exploring outlier representations in Deep Neural Networks." arXiv preprint arXiv:2206.04530 (2022).

Bykov, Kirill, et al. "NoiseGrad—Enhancing Explanations by Introducing Stochasticity to Model Weights." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 6. 2022.

Dombrowski, Ann-Kathrin, et al. "Explanations can be manipulated and geometry is to blame." Advances in Neural Information Processing Systems 32 (2019).

Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. "The false hope of current approaches to explainable artificial intelligence in health care." The Lancet Digital Health 3.11 (2021): e745-e750.

Kaur, Harmanpreet, et al. "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning." Proceedings of the 2020 CHI conference on human factors in computing systems. 2020.

Krakowczyk, Daniel, et al. "Selection of XAI Methods Matters: Evaluation of Feature Attribution Methods for Oculomotoric Biometric Identification." NeuRIPS 2022 Workshop on Gaze Meets ML. 2022.

References 3/4

Lapuschkin, Sebastian, et al. "Analyzing classifiers: Fisher vectors and deep neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

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

Nauta, Meike, et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai." arXiv preprint arXiv:2201.08164 (2022).

Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225 (2017).

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.

Sixt, Leon, Maximilian Granz, and Tim Landgraf. "When explanations lie: Why many modified bp attributions fail." International Conference on Machine Learning. PMLR, 2020.

References 4/4

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).

Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." arXiv preprint arXiv:1706.03825 (2017).

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

Sundararajan, Mukund, and Ankur Taly. "A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values." arXiv preprint arXiv:1806.04205 (2018).

Yona, Gal, and Daniel Greenfeld. "Revisiting Sanity Checks for Saliency Maps." arXiv preprint arXiv:2110.14297 (2021).

Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.