PONTIFICIA
UNIVERSIDAD
CATÓLICA DE
VALPARAÍSO

pucv.cl

Explainable Artificial Intelligence

Escuela de Ingeniería Informática

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Explainable Artificial Intelligence

Introduction

Sources

- Ali et al.: „Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence.“ In: Elsevier Information Fusion, 2023

Material auxiliar

Papers

- Ali et al.: „Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence.“ In: Elsevier Information Fusion, 2023

Deep Dive

- Anders et al.: “Finding and removing Clever Hans: Using explanation methods to debug and improve deep models”. In: Elsevier Information Fusion, 2022.
- Ali et al.: ”Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence“. In: Elsevier Information Fusion, 2023. https://doi.org/10.1016/j.inffus.2023.101805
- Hedström et al.: ”Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond“. In: Journal of Machine Learning Research, 2023.
- Rojat et al.: “Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey“. arxiv preprint, 2021.

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Explainable Artificial Intelligence (XAI)

In Introduction to XAI

Agenda

Explainable Artificial Intelligence

1. Motivation
    1. Clever Hans
    2. Watermark on image classification
    3. Cow-on-beach
2. Definitions
    1. White, gray, black
    2. Balance between accuracy and interpretability
3. Types of explainers
    1. Scoop-based
        - LIME, SHAP/Shapely values
    2. Complexity-based
        - Decision Trees, TREPAN
    3. Methodology-based
        - Saliency Maps, LIME, LRP
4. Summary

Motivation: Clever Hans

Clever Hans The performing horse Clever Hans with his trainer, Wilhelm von Osten, 1904.

Source https://www.britannica.com/topic/Clever-Hans

Motivation: Clever Hans

In exhibitions beginning in **1891** and led by his trainer, Wilhelm von Osten, **Hans would demonstrate almost "human" intelligence** by responding to questions with a variety of **hoof taps** or other actions.

After a series of **carefully designed experiments** and close behavioral observations, Oskar Pfungst—a student at the **Psychological Institute** at the University of Berlin—concluded that **Clever Hans** was, in fact, **simply responding** to very **subtle**, probably involuntary**, cues from von Osten**.

"Clever Hans Effect" explained as simple behavioral responses to subtle cues provided (perhaps unintentionally) by his handler.

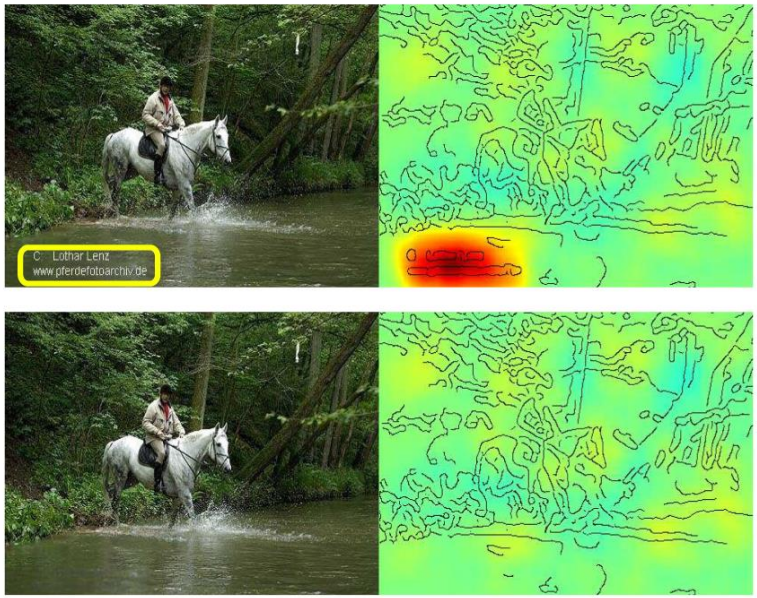
Motivation: Clever Hans

"Clever Hans effect" and Artificial Neural Networks.

"Detect whether the learned strategy [of the neural network] is **valid and generalizable** or whether the model has based its decision on a **spurious correlation in the training data**"

Classifier trained on the PASCAL VOC 2007 data set focuses on a source tag present in about one-fifth of the horse figures. Removing the tag also removes the ability to classify the picture as a horse.

Horse-picture from Pascal VOC data set



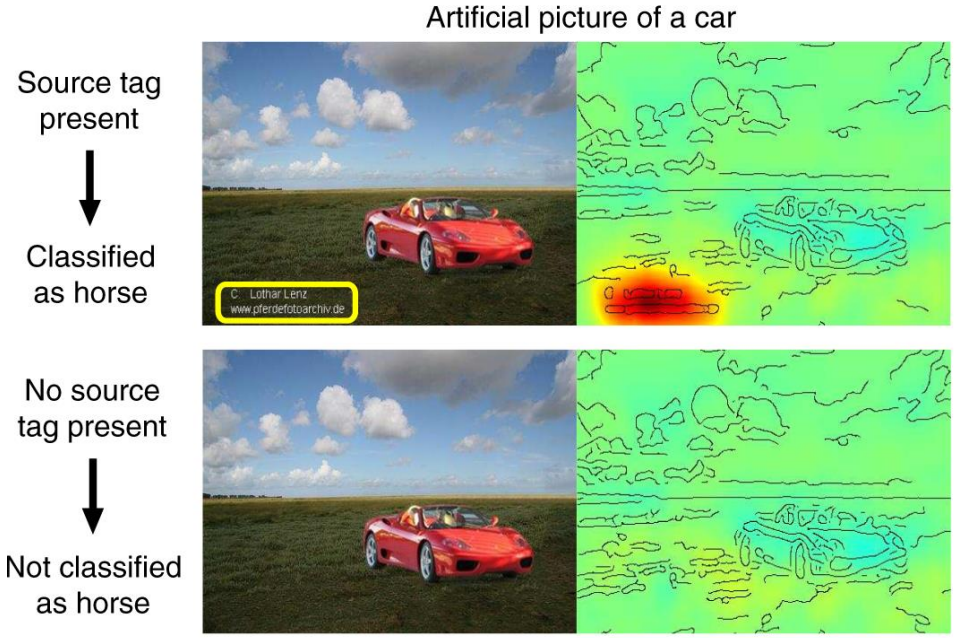
Source: Lapuschkin, S., Wäldchen, S., Binder, A. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* **10**, 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4

Motivation: Clever Hans

"Clever Hans effect" and Artificial Neural Networks.

"Detect whether the learned strategy [of the neural network] is **valid and generalizable** or whether the model has based its decision on a **spurious correlation in the training data**"

Inserting the tag on a car image changes the classification from car to horse.



Source: Lapuschkin, S., Wäldchen, S., Binder, A. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* **10**, 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4

Motivation: Clever Hans

Spurious correlations: In psychology the reliance on such spurious correlations is typically referred to as the "**Clever Hans phenomenon**".

A model implementing a 'Clever Hans'-type decision strategy **will likely fail** to provide correct classification and thereby usefulness once it is deployed in the **real world**, where **spurious** or artifactual **correlations may not be present**.

Source: Lapuschkin, S., Wäldchen, S., Binder, A. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* **10**, 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4

Motivation: Invisible cow on beach

Spurious correlations limit generalization of ANNs in the real world.

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

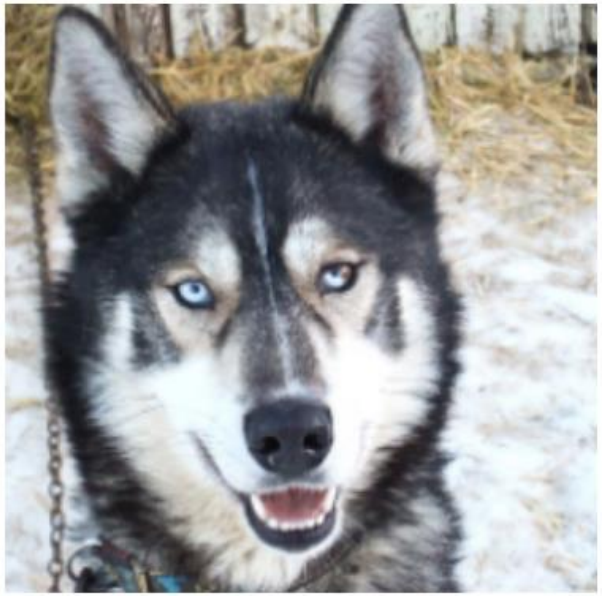
(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Fig. 1. Recognition algorithms generalize poorly to new environments. Cows in 'common' contexts (e.g. Alpine pastures) are detected and classified correctly (A), while cows in uncommon contexts (beach, waves and boat) are not detected (B) or classified poorly (C). Top five labels and confidence produced by ClarifAI.com shown.

Source: Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Motivation: what features were relevant?

Spurious correlations limit generalization of ANNs in the real world.



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

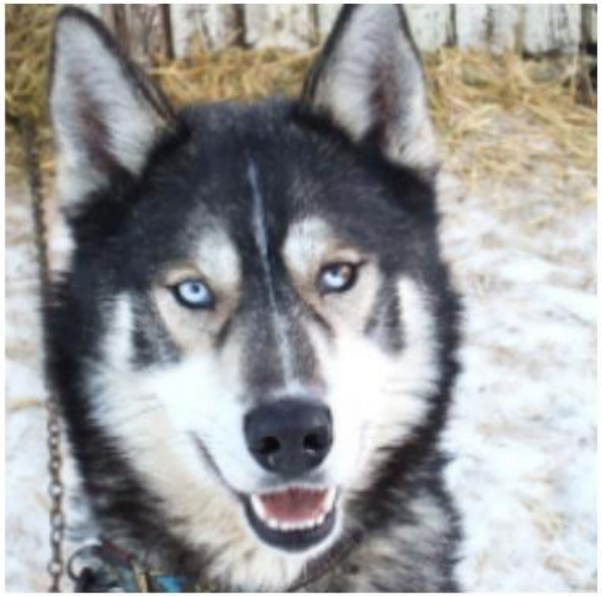
Source: LIME Ribeiro et al. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier"

Motivation: what features were releva

Spurious correlations limit generalization of ANNs i

2nd XAI class: can we trust the explanations?

1st XAI class: how can we obtain such explanations?



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Source: LIME Ribeiro et al. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier"

What does XAI answer?

Research Questions for explaining models

1. Data Explainability (Exploratory Data Analysis)

   What sort of information do we have in the database?

   What are the most important portions of the data?

   How is the information distributed?

What else is it good for?

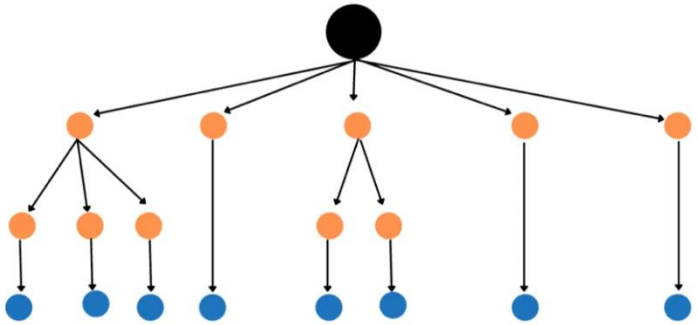
Research Questions for explaining models

2. Model Explainability

   What makes a parameter, objective, or action important to the system?

   What are the **consequences** of making a different decision or **adjusting a parameter**?

   How does the system carry out a certain action?



What else is it good for?

Research Questions for explaining models

3. Post-hoc Explainability

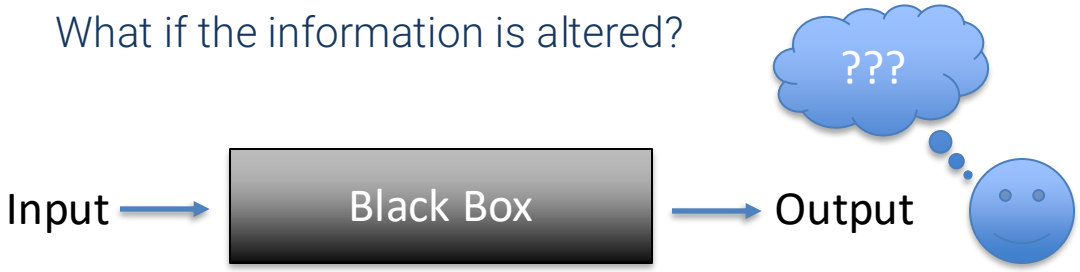
   What is the reason behind the model's prediction?

   What was the reason for occurrence X?

   What would happen **if** Y was the cause of occurrence X?

   What variables have the most influence on the user's decision?

   What if the information is altered?



Agenda

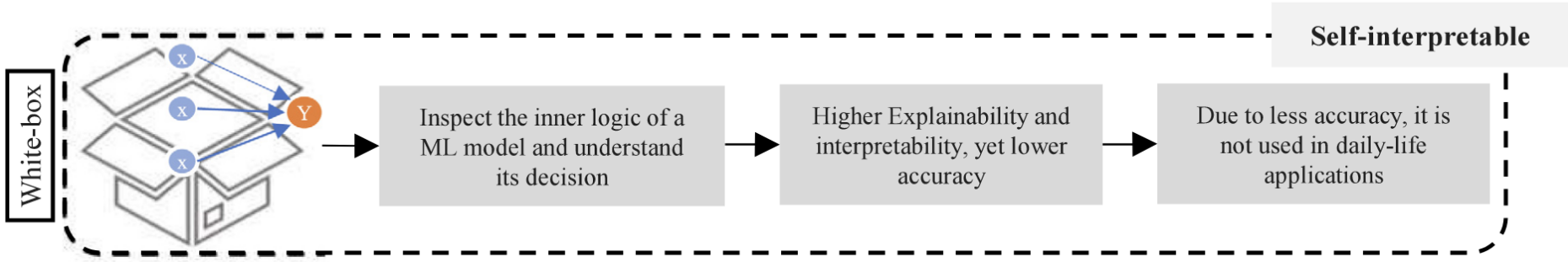
1. Motivation
    1. Clever Hans
    2. Water-mark on image classification
    3. Cow-on-beach
2. Definitions
    1. White, gray, black
    2. Balance between accuracy and interpretability
3. Types of explainers, with examples
    1. Scoop-based
        - LIME, SHAP/Shapely values
    2. Complexity-based
        - Decision Trees, TREPAN
    3. Methodology-based
        - Saliency Maps, LIME, LRP
4. Summary

Definition: *white-box models*

How do we explain AI's reasoning to uncover spurious correlations?

Definition: White-box (or glass box) models. The internal workings of white box models are fully visible and understandable. This means that every step of the model's decision-making process can be examined and interpreted by humans.

Examples: Decision Trees, Linear Regression, Rule-based systems.



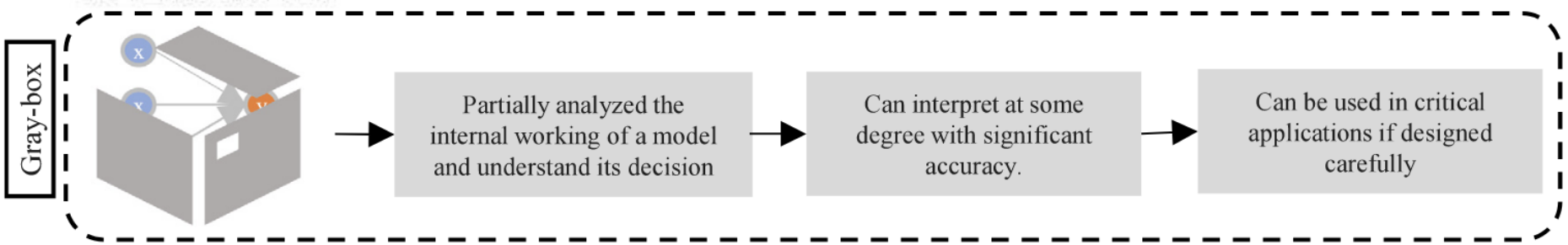
Source: Ali et al.: „Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence." In: Elsevier Information Fusion, 2023

Definition: *grey-box models*

How do we explain AI's reasoning to uncover spurious correlations?

Definition: Grey-box models are only partially observable. They are interpretable to some degree and still achieve a higher accuracy.

Examples: Fuzzy Systems, Bayesian Networks.



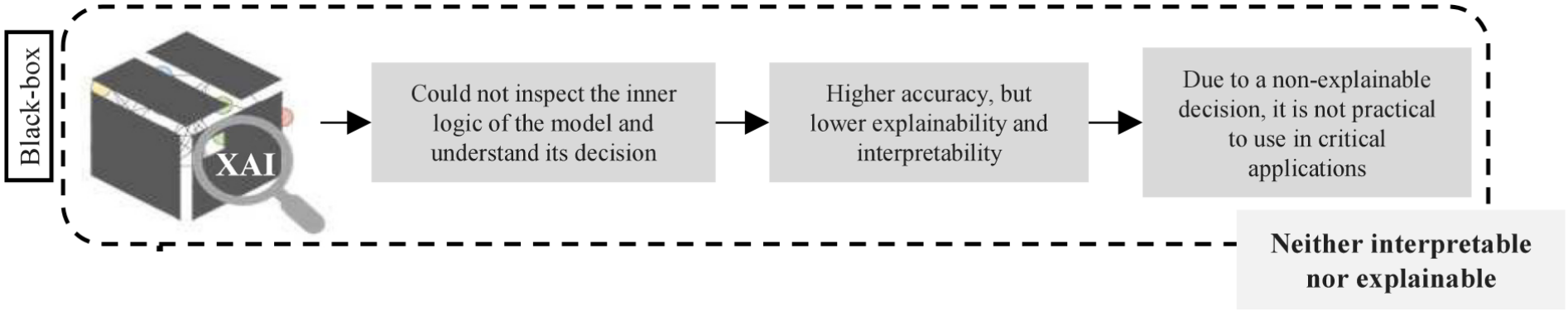
Source: Ali et al.: „Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence." In: Elsevier Information Fusion, 2023

Definition : *black-box models*

How do we explain AI's reasoning to uncover spurious correlations?

Definition: Black-box models. The internal mechanisms of black box models are not easily interpretable. The model's predictions are clear, but the way they are derived is not transparent.

Examples: Deep Neural Networks, Ensemble methods (Random Forest), Gradient Boosting.



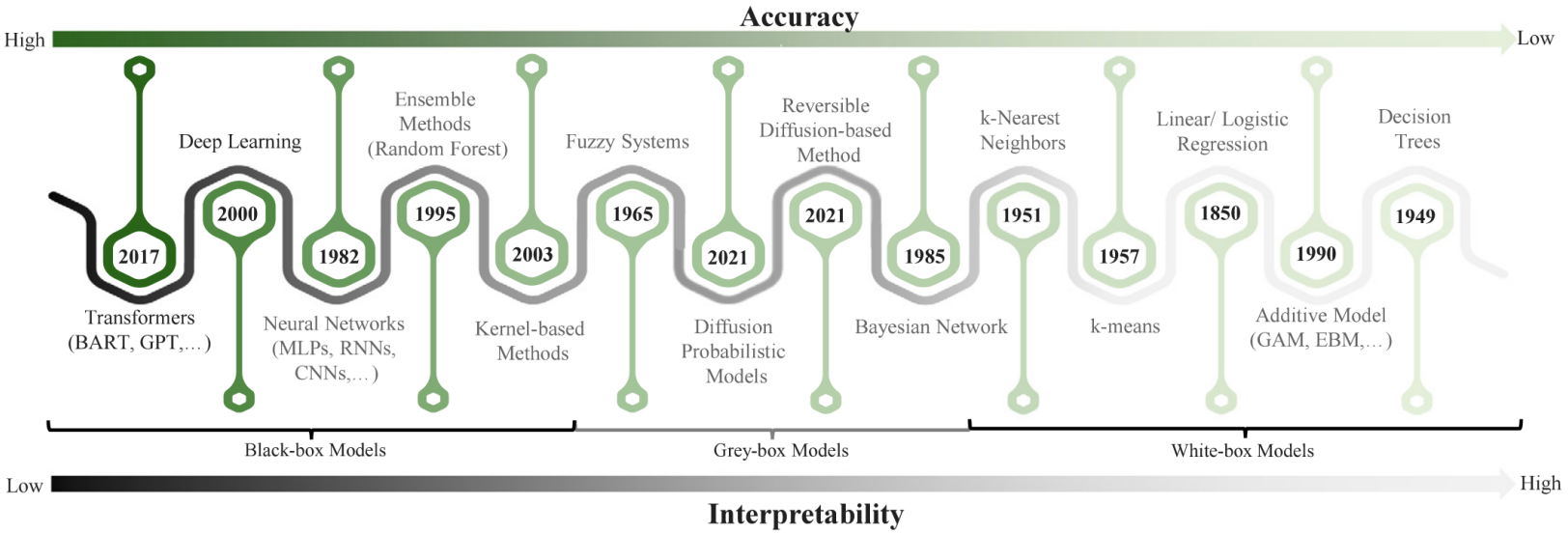
Trade-off accuracy/explainability

Apparent **balance** between ML model's **Performance** and **Interpretability**.

Example: CNNs are harder to understand than Decision Trees.

→ The models become harder to understand if the number of parameters increases

However, "there is no scientific evidence for a general tradeoff between accuracy and interpretability" (Rudin et al., 2022)



Agenda

1. Motivation
    1. Clever Hans
    2. Water-mark on image classification
    3. Cow-on-beach
2. Definitions
    1. White, gray, black
    2. Balance between accuracy and interpretability
3. Types of explainers, with examples
    1. Scoop-based
        - LIME, SHAP/Shapely values
    2. Complexity-based
        - Decision Trees, TREPAN
    3. Methodology-based
        - Saliency Maps, LIME, LRP
4. Summary

Types of Explainers

We use the following taxonomy for explainability methods:

1. Scoop-based
2. Complexity-based
3. Methodology-based

Scoop-based

These explainers are used to analyze feature importance:
→ How to input(s) relate to model output(s)?

What *scoop* of input data are we looking at?

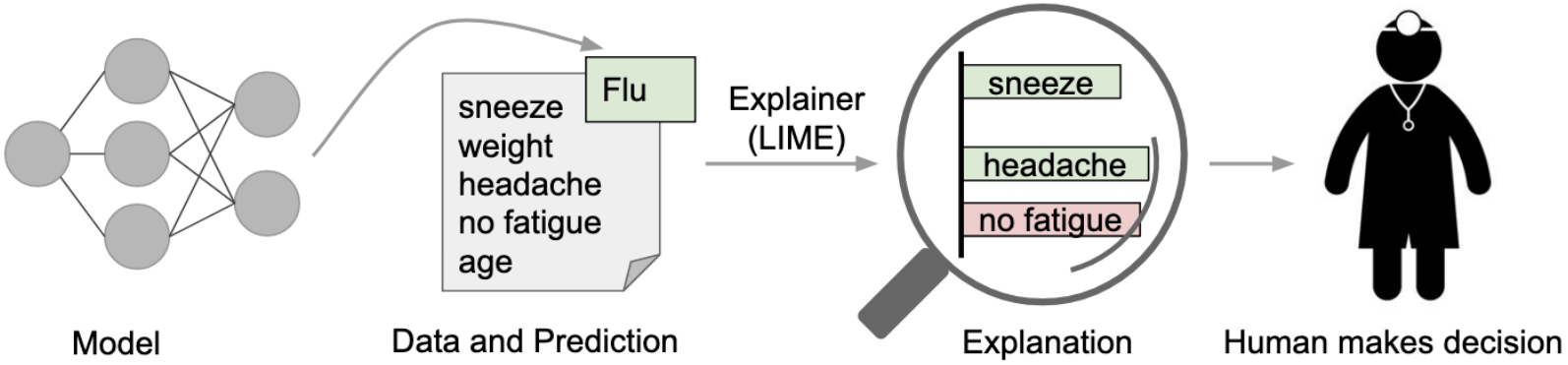
Image: Getty Images

Scoop-based

Local explanation only looks at a specific input sample.

- Explanation can only be for one prediction/output/decision.

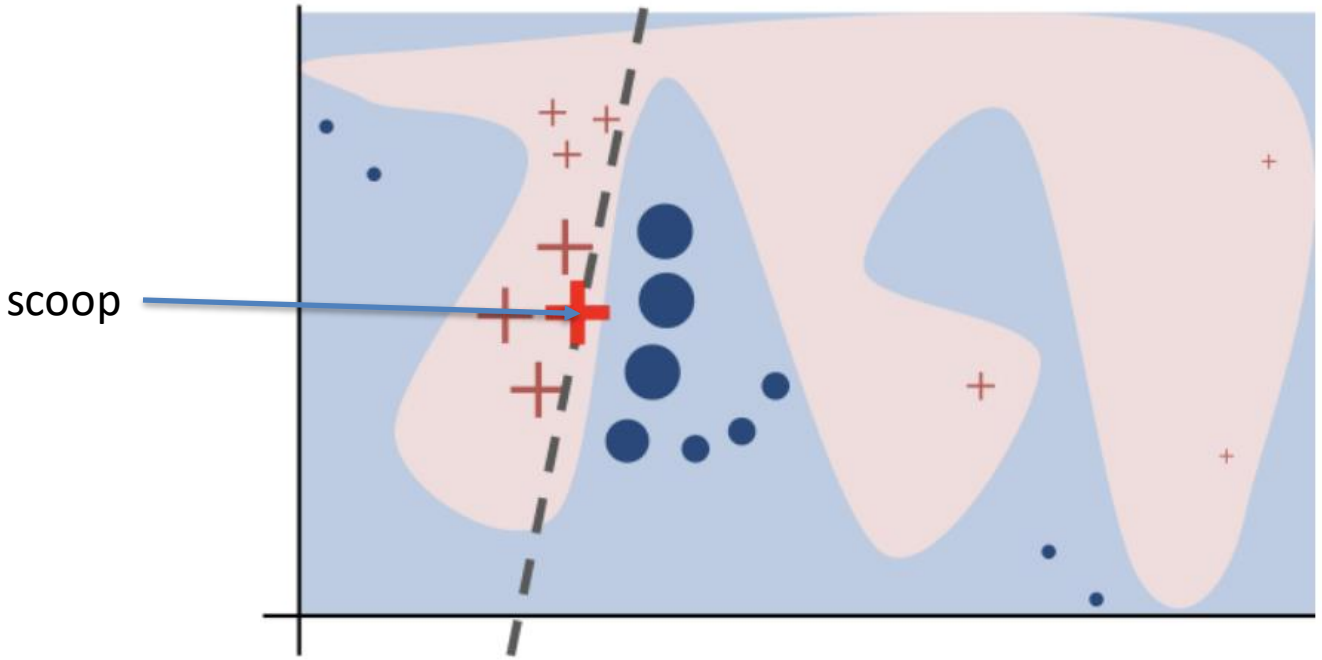
Example: **Locally Interpretable Model-Agnostic Explainer** (LIME)



Source LIME: M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

Locally Interpretable Model-Agnostic Explainer (LIME)

The overall goal of LIME is to identify an interpretable model over the interpretable representation that is **locally faithful** to the classifier. (=scoop)



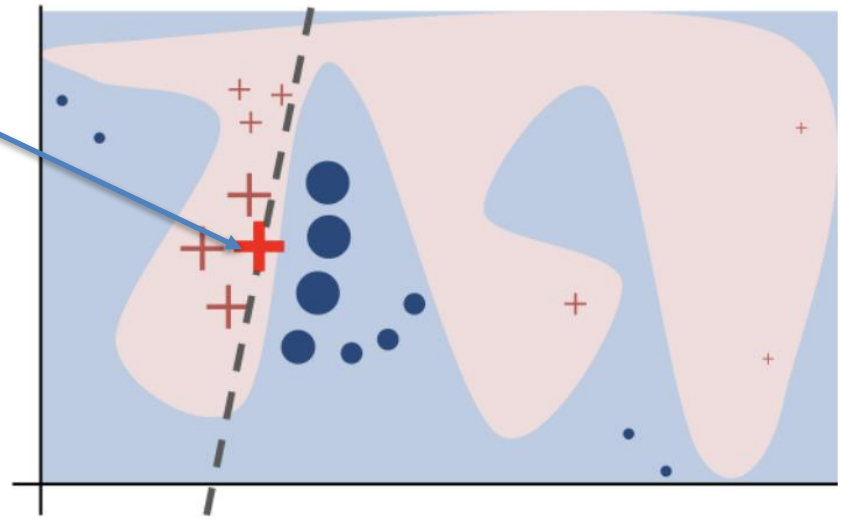
Source LIME: M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

LIME

Blue region shows a complex Neural Network
LIME creates interpretable linear model

Only bold-red cross is being explained

LIME samples instances, gets predictions using ANN, and weighs them by the proximity to the instance being explained (represented here by size).

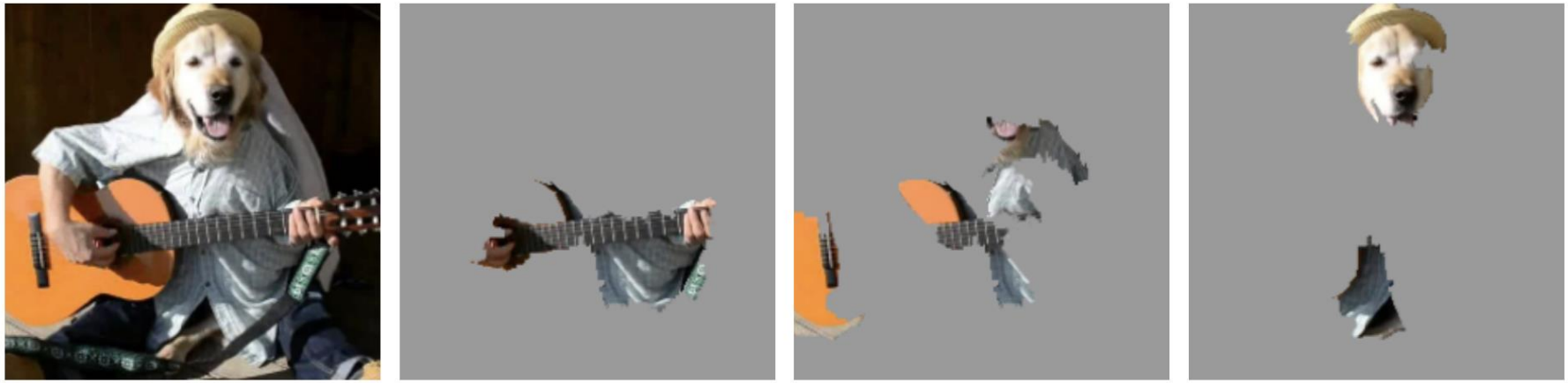


Linear model is only locally faithful!

Source LIME: M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

LIME

Example: Explaining what pixels contribute to a ANNs prediction. LIME builds a simple surrogate model for this picture and for the classes.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Source LIME: M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

Global Explanations?

How can we get a **Global explanation that** looks at all data?

- Explanation that provides rationale for whole dataset.

Example: **SHapley Additive exPlanation** (SHAP) **Values**

How much did a single feature affect the prediction?

Source SHAP: S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4768–4777.

SHAP Values

Example: SHapley Additive exPlanation (SHAP) Values
How much did a single feature affect the prediction?

y = Predict House Price in USD

Features X = [
Longitude,
Latitude,
Median Income,
Number of Rooms,
Age,
Population
…
]

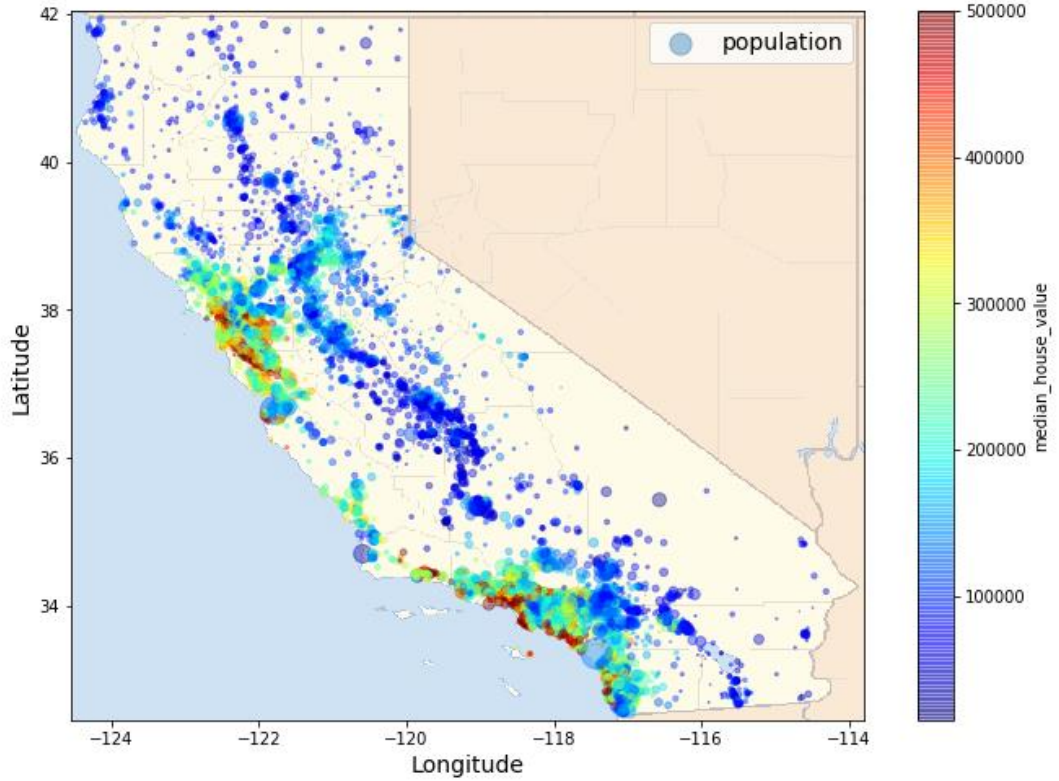


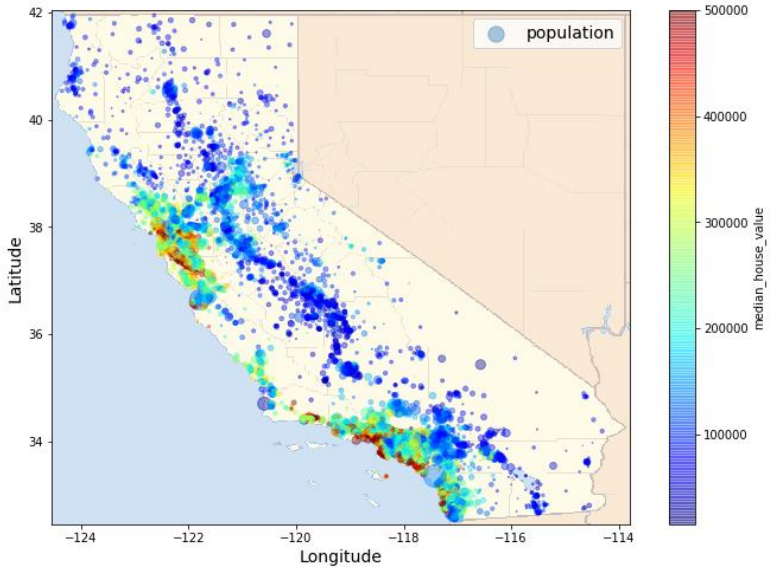
Image source: https://github.com/amansingh9097/California-housing-price-prediction

SHAP Values

Example: SHapley Additive exPlanation (SHAP) Values
How much did a single feature affect the prediction?

y Predict Price in USD
X Longitude, Latitude, Median Income, Number of Rooms, Age, Population

House prices in California



SHAP values of feature importance for house prices

Before revealing the results, what do you think?

What feature is most important

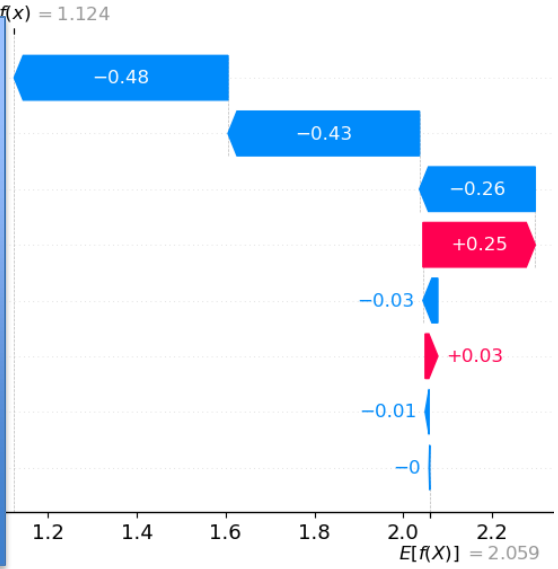


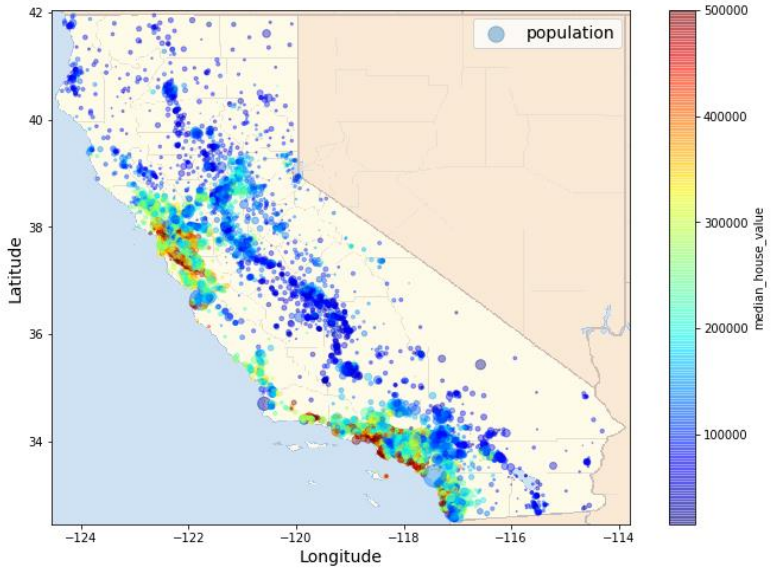
Image source: https://github.com/amansingh9097/California-housing-price-prediction

SHAP Values

Example: SHapley Additive exPlanation (SHAP) Values
How much did a single feature affect the prediction?

y Predict Price in USD
X Longitude, Latitude, Median Income, Number of Rooms, Age, Population

House prices in California



SHAP values of feature importance for house prices

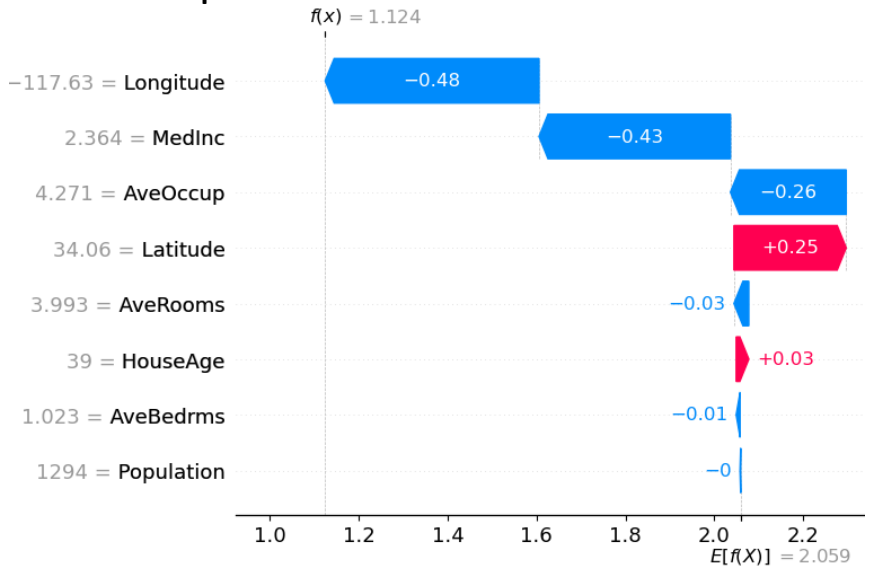


Image source: https://github.com/amansingh9097/California-housing-price-prediction

SHAP Values

Algorithm

- build power set of models
- Over features (A, B, C)
- Calculate marginal contributions
  - Weight are connections per layer, e.g., 1/3

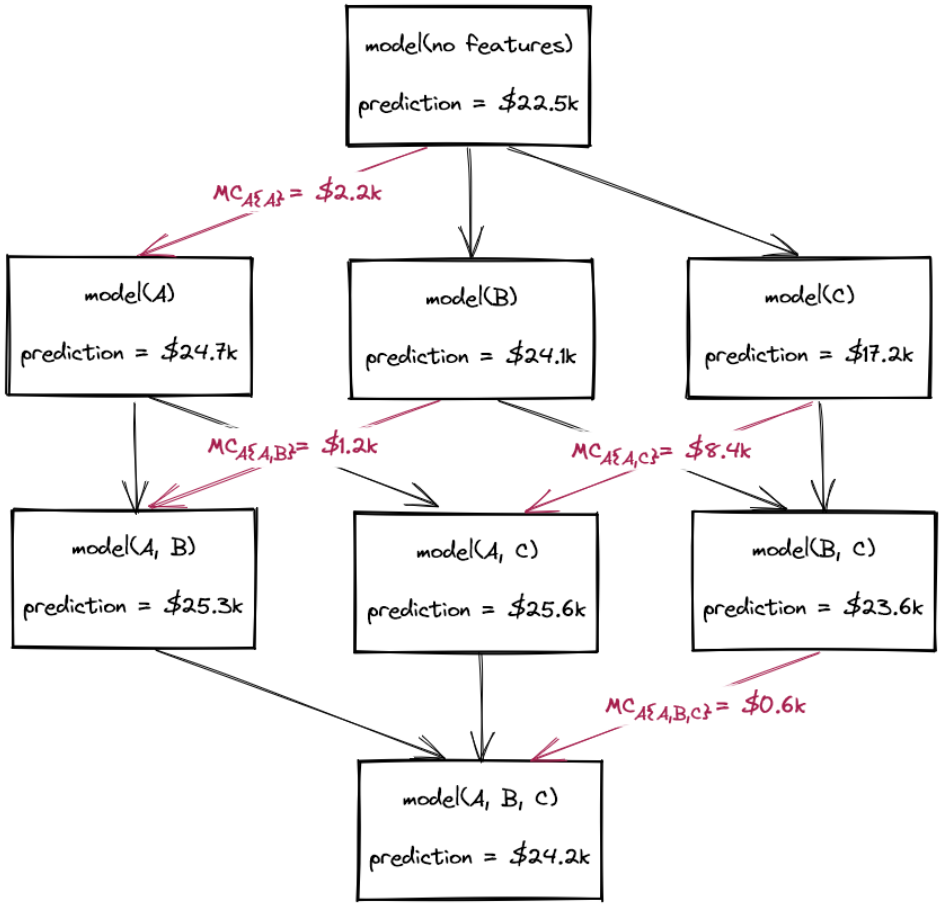
$shapley_A(house1)$=

1/3×2200+
1/6×1200+ 1/6×8400+
1/3×600 = $2,550 USD

Downside: Train model 2^F times where F is #features

SHAP vs Shapley values?

SHAP value use cheap approximated model instead of full model (LIME ☺)



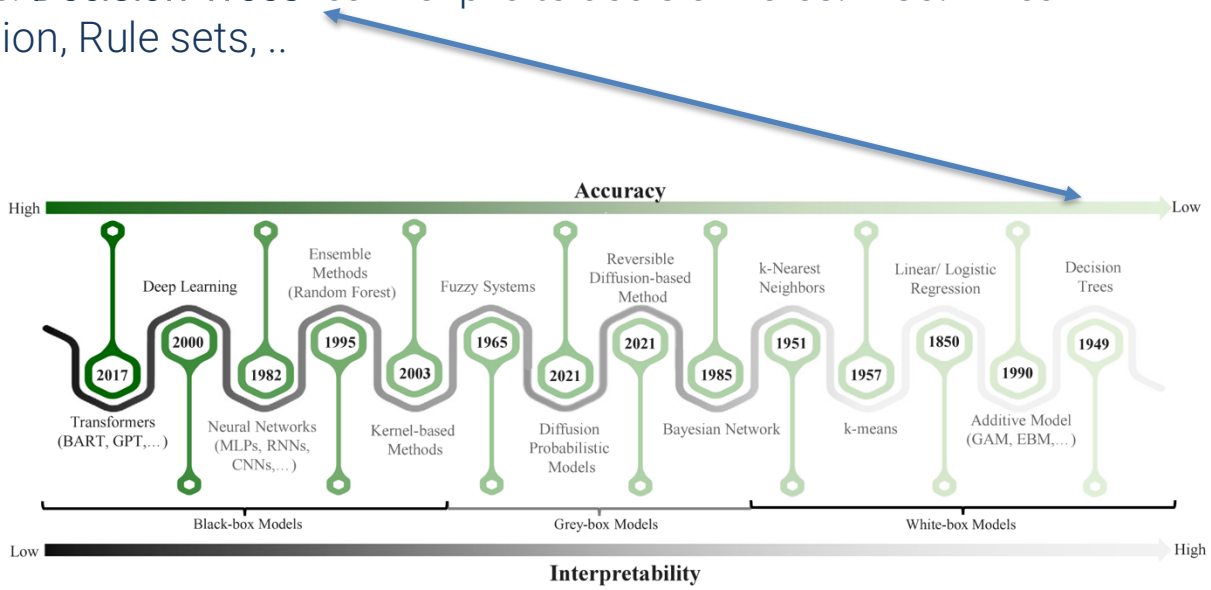
Types of Explainers: Complexity-based

Interpretabiliy is proportional to the model complexity.
→ More complex models are more difficult to explain.

We classify interpretation methods into **intrinsic** or **post-hoc** explainers.
Intrinsic interpretability is accomplished using simple models

- Self-explanatory models with interpretation built-in

Example: **Decision Trees** learn explicite decision rules. Also: Linear Regression, Rule sets, ..



Decision Trees

Example: **Decision Trees** learn explicite decision rules.
Intrinsically interpretable, as the following tree shows:

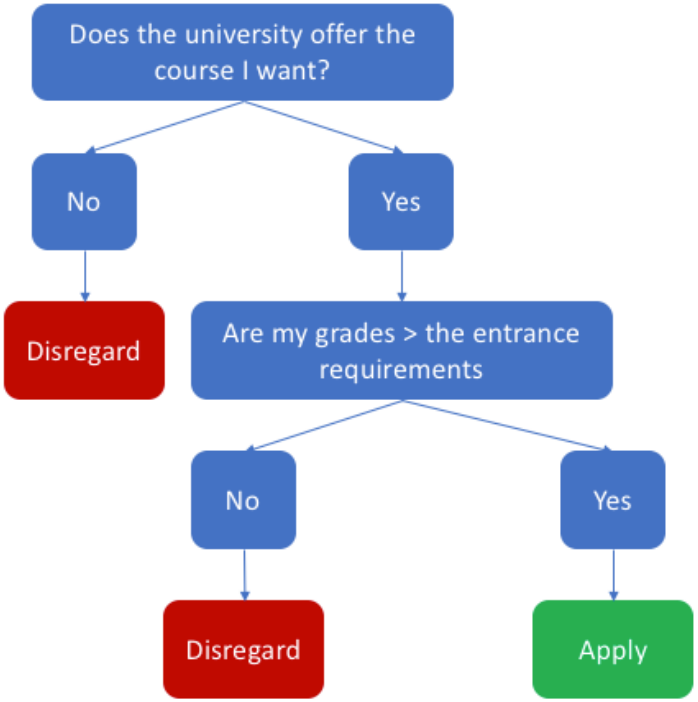


Image source: https://medium.com/data-science/decision-trees-understanding-explainable-ai-620fc37e598d

Types of Explainers: Complexity-based

Post-hoc interpretation may be accomplished using a second, simpler model

- The surrogate of the original model is intrinsically explainable.
- (Remember LIME? *Linear* surrogate model)

Example: TREPAN fits a decision tree model to the behaviour of an ANN.

Source TREPAN: 144] M.W. Craven, J.W. Shavlik, Extracting tree-structured representations of trained networks, in: Proceedings of NIPS, 1995, pp. 24–30.

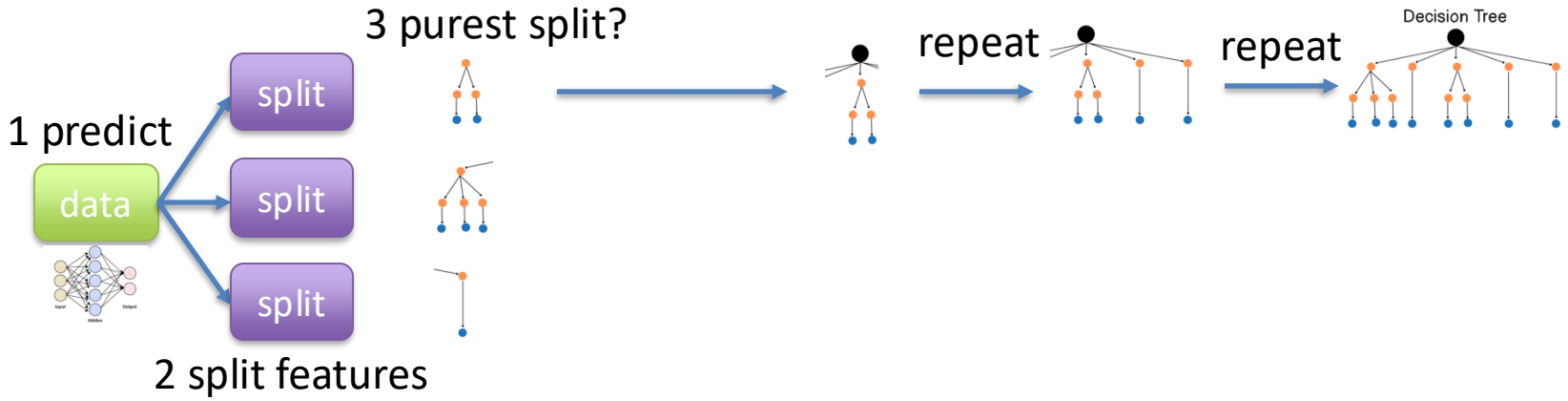
TREPAN

Example: TREPAN fits a decision tree model to the behaviour of an ANN.

How do you build a surrogate Decision Tree for an ANN?

Loop:

1. Draw samples of dataset
2. Use trained network (ANN) to label the samples
3. Generate many splits using different features ("*bagging*")
4. Select "**best split**" (like, purest split), for the node of the Decision Tree
5. For each possible split make new leaf



Source TREPAN: 144] M.W. Craven, J.W. Shavlik, Extracting tree-structured representations of trained networks, in: Proceedings of NIPS, 1995, pp. 24–30.

Methodology-based: core XAI algorithms

Methdology-based methods are the are the **core algorithms** of e**X**plainable **A**rtificial **I**ntelligence.

They are classified by their **method of calculating interpretations.**

1. Backpropagation-based
2. Perturbation-based

Backpropagation-based

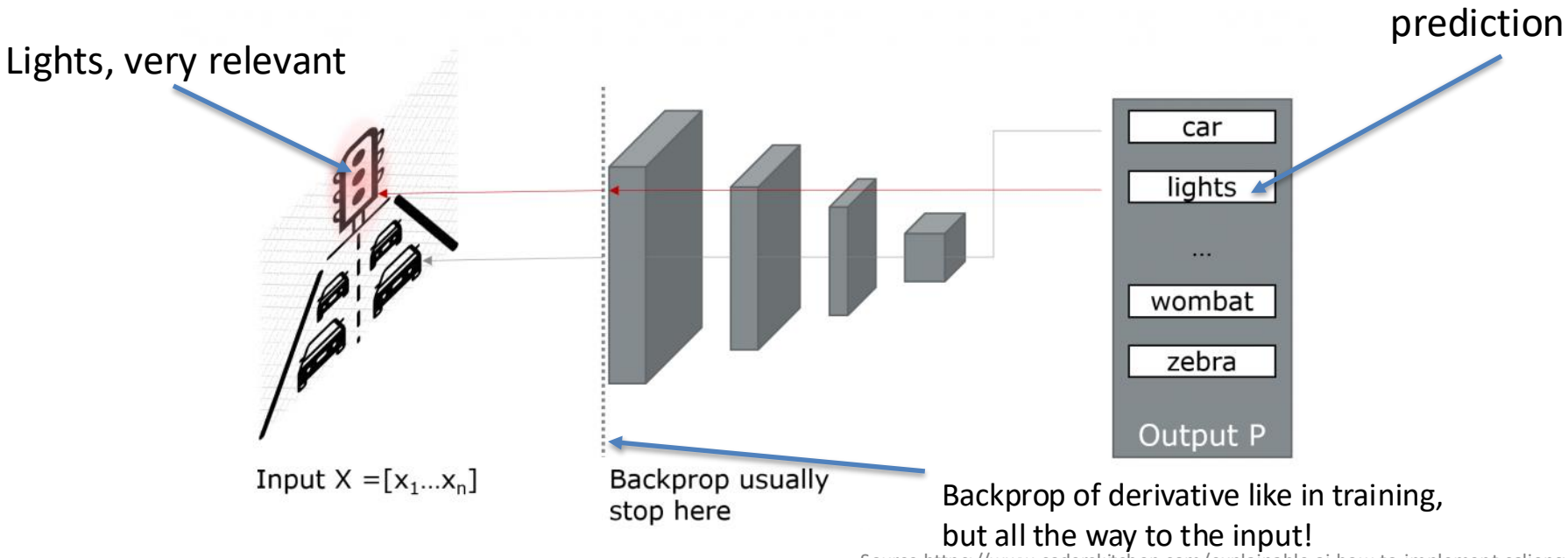
Backpropagation-based (aka gradient-based) methods use the back-propagation algorithm

- Backprop a significance signal backwards from output to input
- Backprop adds “weight” to each activation from forward-pass
  - gradient signal of each parameter

Examples: Salience Maps. Intuition: Where do people focus on first when they see an image? This is the degree of importance of pixels in the image.

Saliency Maps

Algorithm: use same mechanism from BackPropagation of **Gradients** (partial derivatives) from training of neural networks. But use it as if you propagate back a "sort-of" **relevancy for the prediction of the class P** (e.g., lights) from **output back to the input** picture (here, photo of cars waiting at lights)



Source https://www.coderskitchen.com/explainable-ai-how-to-implement-saliency-maps/

Types of Explainers: Methodology-based

Examples: Salience Maps)

Input images

backprop the gradient to input to generate saliency

Source https://www.coderskitchen.com/explainable-ai-how-to-implement-saliency-maps/

Perturbation-based

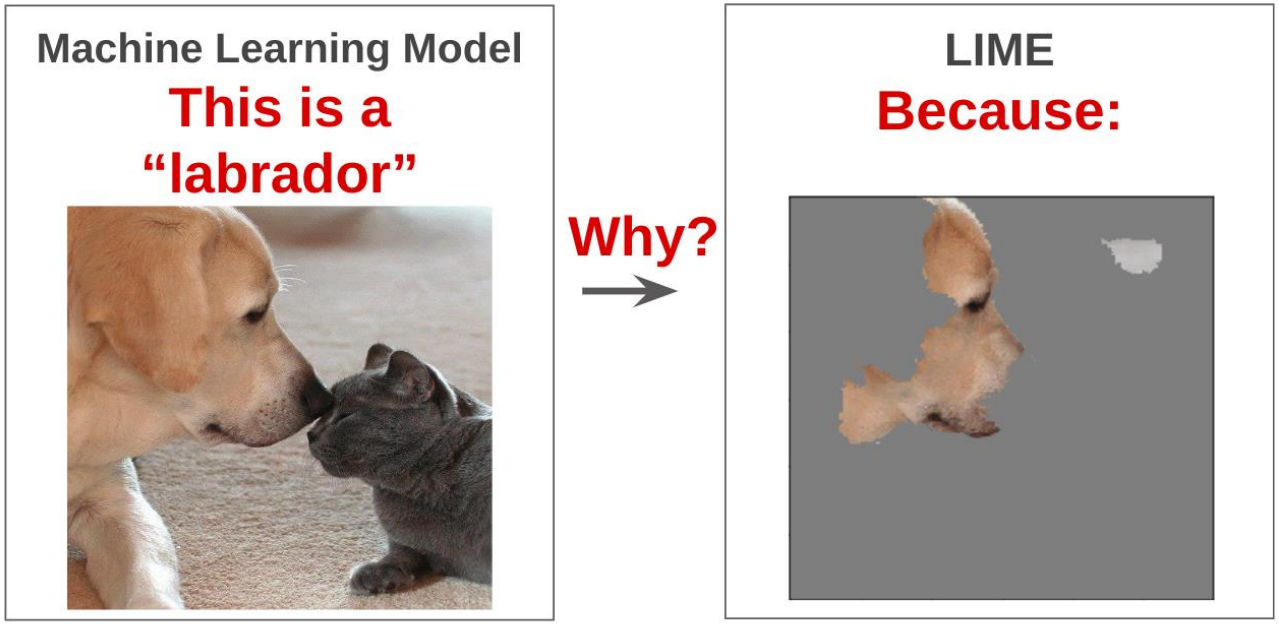
Perturbation-based methods change inputs and measure changes in outputs

- Change input features: occlusions of inputs, partly replacing features, masking, conditional sampling, etc.
- Evaluate impact on model outputs
- One Forward-pass is enough (no gradients needed)

Example: Locally Interpretable Model-Agnostic Explainer (LIME)

LIME revisited

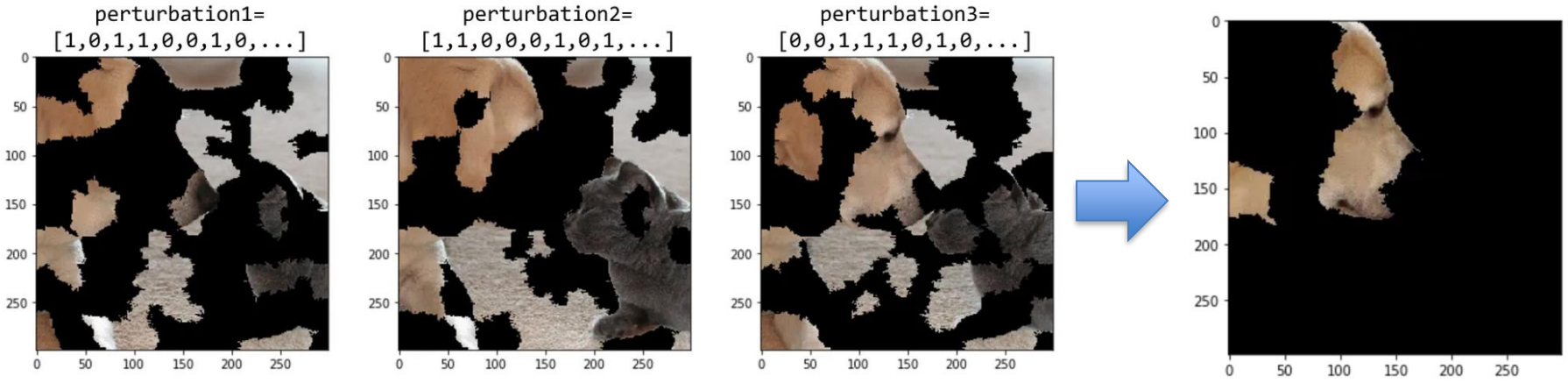
Perturbation-based method
Example: Locally Interpretable Model-Agnostic Explainer (LIME)



Source image: https://medium.com/data-science/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13

LIME with Perturbations

Perturbation-based method
Example: **Locally Interpretable Model-Agnostic Explainer** (LIME)



1. Create many perturbations of the image
2. Predict the class for each perturbation (via ANN)
3. Fit linear model to dataset of (perturbation, label)
4. Rank coefficients to find most important region of image

Source image: https://medium.com/data-science/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13

Explaining Neural Networks

Influence methods alter the input or the parameters of the ANN to see what alters the model performance

Today: Feature Importance and Layer-wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation

Example: **Layer-wise Relevance Propagation** (LRP)

LRP propagates the model's prediction backwards in the neural network, by means of purposely designed **local propagation rules:** $R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k$

Similar to Saliency Maps.

Source LRP: Montavon et al.: "Layer-Wise Relevance Propagation: An Overview". In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 2019

Layer-wise Relevance Propagation

Assumptions

- ANN can be broken down into **layers of computations**
- **Relevance is conserved** from *layer k* to *layer j*: It is redistributed to lower layers
  (Analogously to Kirchoff's conservation laws in electrical circuits)

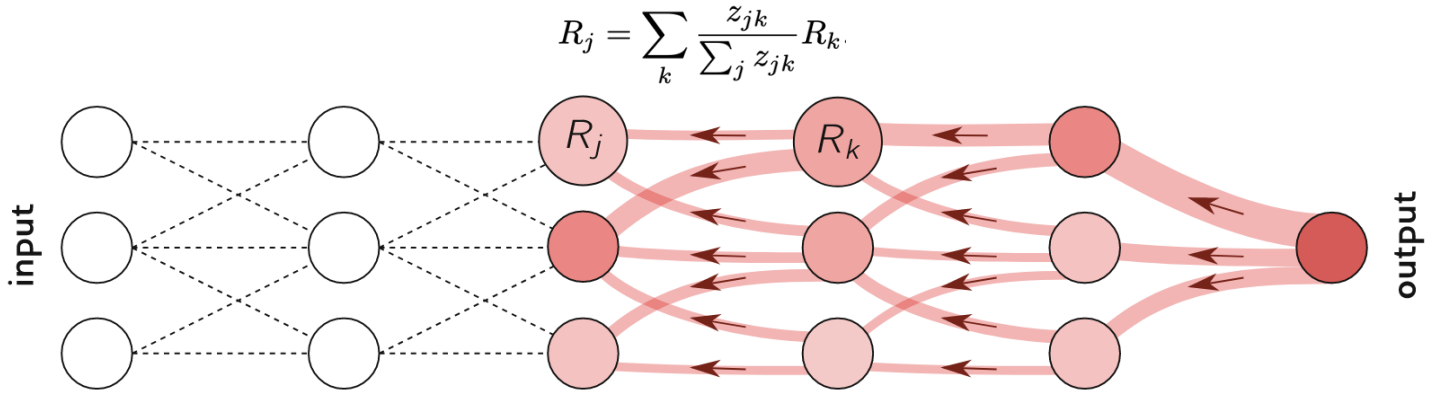


Fig. 10.2. Illustration of the LRP procedure. Each neuron redistributes to the lower layer as much as it has received from the higher layer.

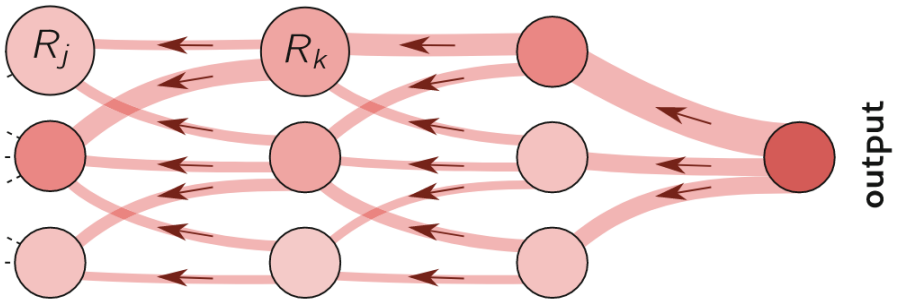
Source LRP: Montavon et al.: "Layer-Wise Relevance Propagation: An Overview". In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 2019

Types of Explainers:
Neural network-based - influence methods

Local propagation rules are at the core of LRP
They are tailored for different ANN types.
Here, we see rules for ANNs with ReLU activations.

Basic Rule (LRP-0). Compute relevance R of lower layer j, where a are the activation outputs and w weights from layer j to layer k.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$



The gradient of a deep neural network is typically noisy, therefore one needs to design more robust propagation rules.

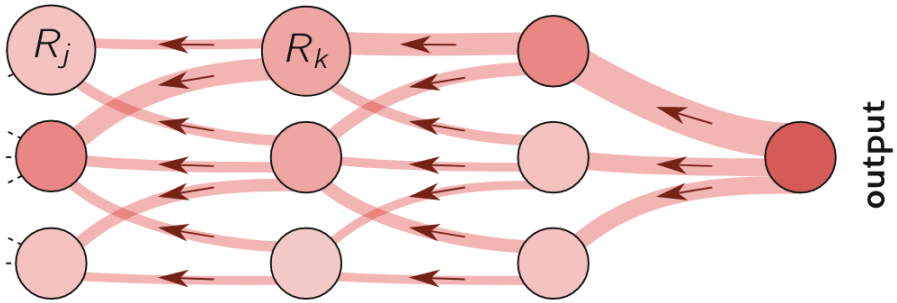
Source LRP: Montavon et al.: "Layer-Wise Relevance Propagation: An Overview". In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 2019

Types of Explainers:
Neural network-based - influence methods

Local propagation rules are at the core of LRP
They are tailored for different ANN types.
Here, we see rules for ANNs with ReLU activations.

Epsilon Rule (LRP-ε): with larger ε only most salient explanation factors survive.

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$$



This typically leads to explanations that are sparser in terms of input features and less noisy.

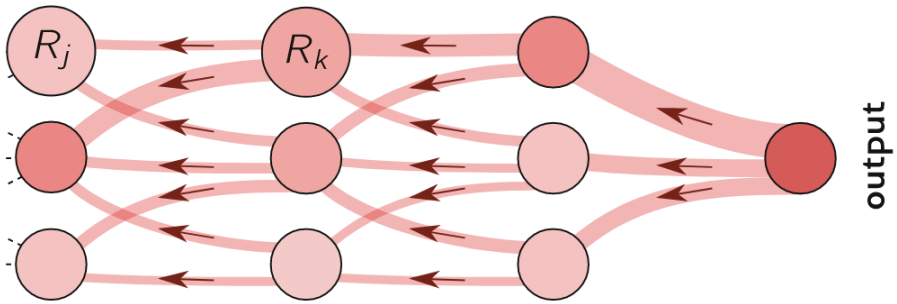
Source LRP: Montavon et al.: "Layer-Wise Relevance Propagation: An Overview". In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 2019

Types of Explainers:
Neural network-based - influence methods

Local propagation rules are at the core of LRP
They are tailored for different ANN types.
Here, we see rules for ANNs with ReLU activations.

Gamma Rule (LRP-γ): where γ favors positive weights. Larger γ make negative contributions disappear

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j \cdot (w_{jk} + \gamma w_{jk}^+)} R_k$$



This helps to deliver more stable explanations

Source LRP: Montavon et al.: "Layer-Wise Relevance Propagation: An Overview". In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 2019

Layer-wise Relevance Propagation: visualized

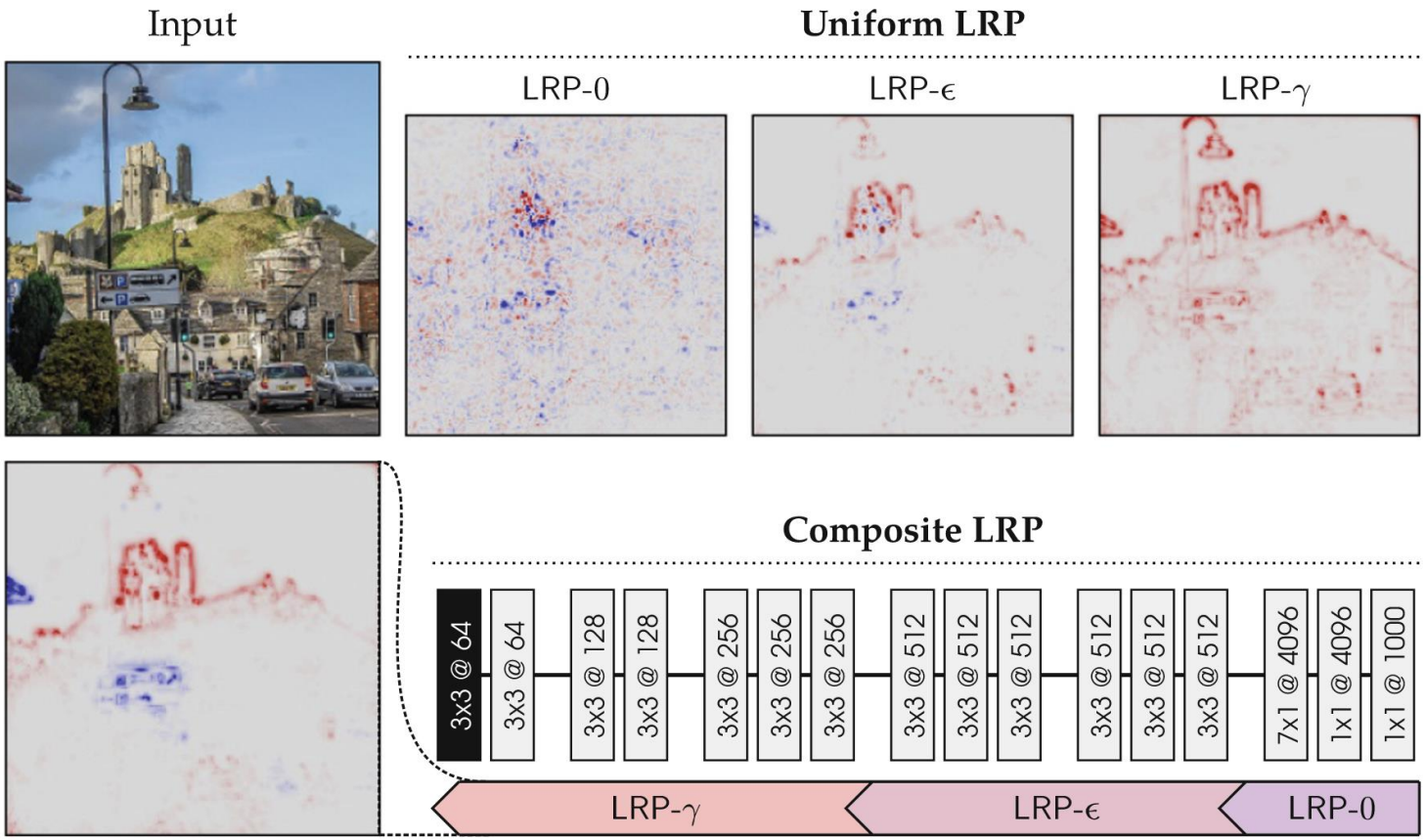


Fig. 10.4. Input image and pixel-wise explanations of the output neuron 'castle' obtained with various LRP procedures. Parameters are $\epsilon = 0.25\,\text{std}$ and $\gamma = 0.25$.

Source LRP: Montavon et al.: "Layer-Wise Relevance Propagation: An Overview". In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 2019

Summary

1. Motivation
    1. Clever Hans
    2. Water-mark on image classification
    3. Cow-on-beach
2. Definitions
    1. White, gray, black
    2. Balance between accuracy and interpretability
3. Types of explainers, with examples
    1. Scoop-based
        - LIME, SHAP/Shapely values
    2. Complexity-based
        - Decision Trees, TREPAN
    3. Methodology-based
        - Saliency Maps, LIME, LRP
4. Summary

Types of Explainers

Summary of types of explainers

- Scoop-based
  Local vs Global explanations, e.g., LIME, SHAP
- Complexity-based
  Intrinsic or post-hoc, e.g., Decision Tree, TREPAN
- Methodology-based
  - Backpropagation vs Perturbation-based, e.g., Saliency Maps, LIME
  - Neural network-based, e.g., Layer-wise Relevance Propagation (LRP)