

# Interpretable Machine Learning

## Fundamental Terms and Concepts



### Learning goals

- What is interpretable machine learning (IML) and Explainable Artificial Intelligence (XAI)?
- What is interpretability?
- What is the purpose of IML?
- What are the fundamental terms and concepts of IML?

# INTERPRETABLE MACHINE LEARNING

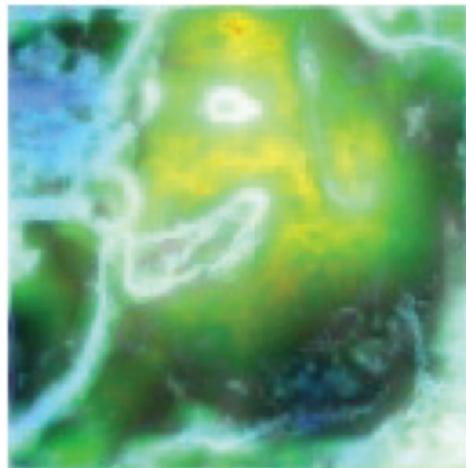
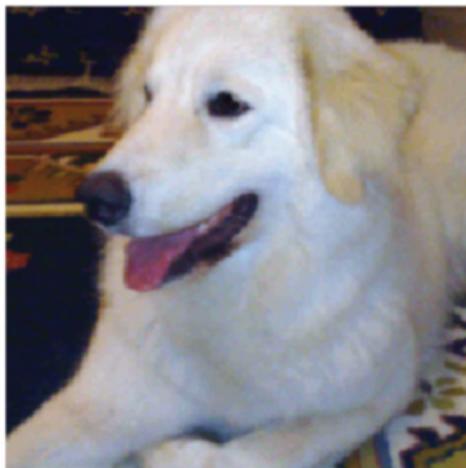
- Machine learning (ML) algorithms algorithmically train predictive models with no or little pre-specifications and assumptions about the data.
- Several algorithms such as decision tree learning create interpretable models. However, most algorithms create models which can be considered a black box.
- We use the term black box, although the internal workings of the model are in fact accessible, but too complex for the human mind to comprehend.
- Interpretable machine learning (IML) is an umbrella term for all models and methods that allow for some kind of interpretation.

# EXPLAINABLE AI

- IML is often used synonymously with Explainable AI (XAI). There is no unified standard for these terminologies. We find that XAI often is specifically concerned with the interpretation of neural networks, whereas IML is used as an encompassing term for everything related to model interpretability, i.e., interpretable models such as generalized additive models, model-agnostic techniques, as well as interpretations of neural networks.
- The nature of neural networks allows for powerful model-specific interpretation techniques, e.g., layer-wise relevance propagation (LRP) and saliency maps. They have in common that influence on the output layer is backpropagated through the entire network layer by layer up to the input layer.

# XAI - SALIENCY MAPS

- For visual data, i.e., pixels being represented as a matrix, deep learning has proven to deliver remarkable outperformance over other model types.
- A saliency map is a heatmap indicating pixel influence on the prediction (e.g., a classification of an image):



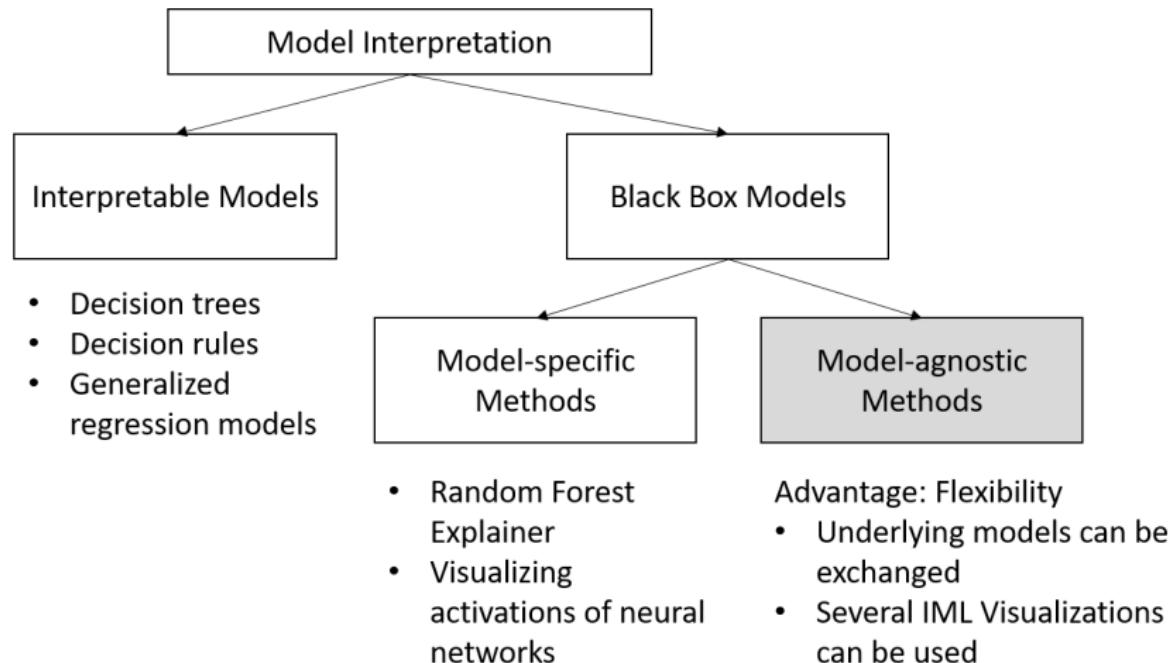
# WHAT IS INTERPRETABILITY?

- There is no scientific consensus on the definition of interpretability.
- We need to differentiate between interpretations of a model or reality. The latter is distorted by all modeling fallacies involved in predictive modeling, e.g., data quality, under- and overfitting, or model extrapolations.
- For practical purposes, we decide to define interpretability such that it is beneficial to our modeling tasks. Think back to the foundations of statistical modeling: the linear regression model (LM). The LM, with its known equation of beta coefficients, represents a paradigm for statistical interpretability.
- It follows that it would be beneficial to create techniques that give us an interpretation similar to the one of an LM.

# DIMENSIONS OF INTERPRETABILITY

- Intrinsic vs Model-agnostic Interpretability.
- Interpretability based on the **type or style** of explanations.
- Local vs Global interpretability.

# INTRINSIC VS MODEL-AGNOSTIC INTERPRETABILITY



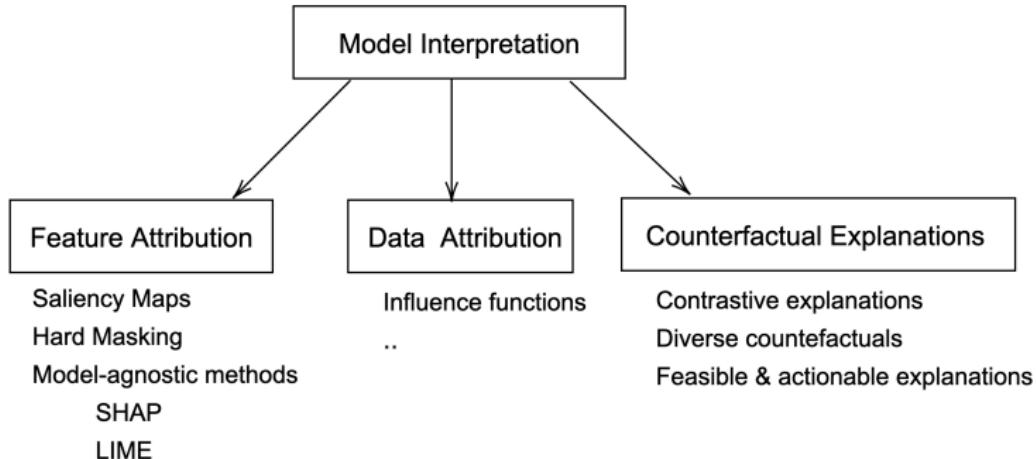
# INTRINSIC AND MODEL-AGNOSTIC INTERPRETATION

- Intrinsicly interpretable models:
  - Examples are linear models and decision trees.
  - interpretable because of their simple structures,  
e.g. weighted combination of feature values or tree structure.
  - difficult to interpret with many features or complex interaction terms.
- Model-agnostic interpretation methods:
  - applied after training (post-hoc).
  - also work for more complex black box models.
  - can also be applied to intrinsicly interpretable models,  
e.g. feature importance for decision trees.

# MODEL-AGNOSTIC INTERPRETABILITY

- Model-agnostic interpretability methods work for **any** kind of machine learning model.
- Explanation type is not tied to the underlying model type.
- Often, only access to data and fitted predictor is required.  
No further knowledge about the model itself is necessary.
- We usually distinguish between **feature effect** and **feature importance** methods.

# TYPES OF EXPLANATIONS



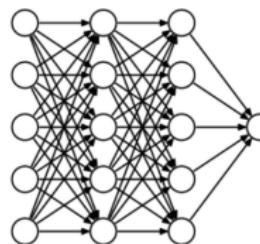
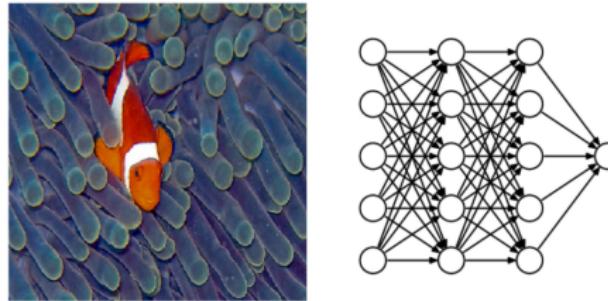
# TYPES OF EXPLANATIONS

The output of an interpretability procedure is an **explanation**. We can differentiate the types of approaches based on the kinds of output

- **Feature attribution:** Feature attribution identifies the subset of features instances that are most responsible for a decision.
- **Data attribution:** Data attribution identifies training instances that are most responsible for a decision.
- **Counterfactual Explanation:** Counterfactual explanations are obtained by identifying the smallest change made to an input to change a prediction made by a model.

# EXPLANATION USING TRAINING INSTANCES [HTTPS://ARXIV.ORG/PDF/1703.04730.PDF](https://arxiv.org/pdf/1703.04730.pdf)

**Data attribution:** Which training instances results in the decision for the instance  $x$  of the model ?



Fish

# EXPLANATION USING TRAINING INSTANCES [HTTPS://ARXIV.ORG/PDF/1703.04730.PDF](https://arxiv.org/pdf/1703.04730.pdf)

**Data attribution:** Which training instances results in the decision for the instance  $x$  of the model ?



- Methods: Influence functions, prototype generation.

# EXPLANATION USING COUNTERFACTUALS

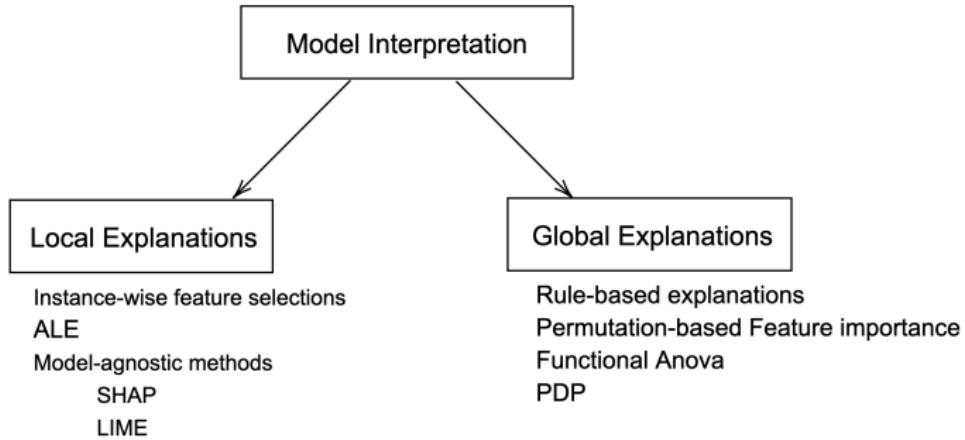
A counterfactual is small “imperceptible” change in  $x$ . What if a small difference  $|x - x'| \leq \epsilon$  to  $x$  causes a large change in the model output ?



What can a person do to obtain a favorable prediction from a given model ?



# LOCAL VS GLOBAL EXPLANATIONS



# GLOBAL AND LOCAL INTERPRETABILITY

Global interpretability methods explain the expected model behavior for the entire input space by considering all available observations (or representative subsets). For example:

- Permutation Feature Importance
- Partial Dependence Plot
- Functional Anova
- ...

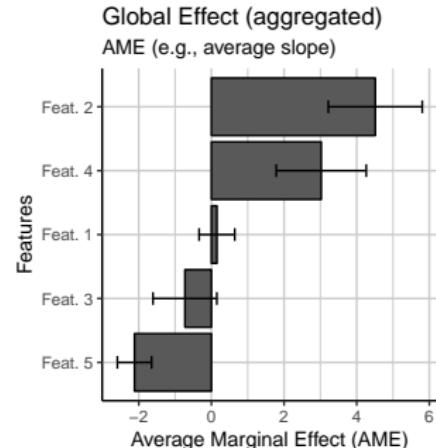
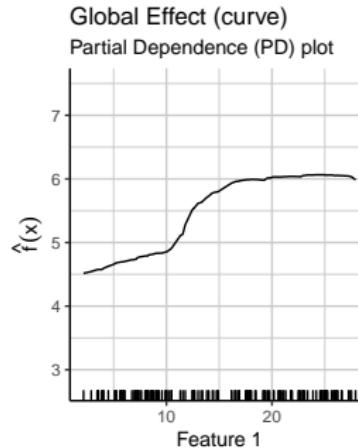
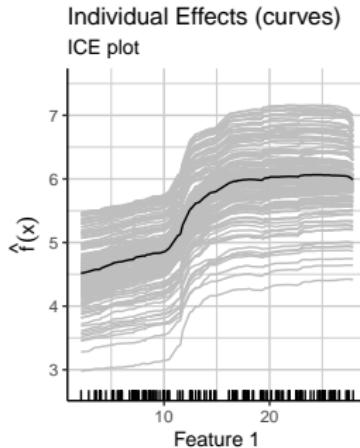
Local interpretability methods explain single predictions or a group of similar observations.

For example:

- Individual Conditional Expectation (ICE) Plots
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley Values
- ...

# FEATURE EFFECTS VS. FEATURE IMPORTANCE

**Feature Effects** visualize or quantify the (average) relationship or contribution of a feature to the model prediction.

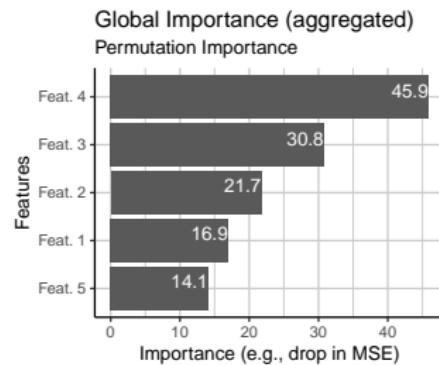


- Methods: Partial Dependence Plots, Individual Conditional Expectation, Accumulated Local Effects (ALE)
- Pendant in linear models: Regression coefficient  $\hat{w}_j$

# FEATURE EFFECTS VS. FEATURE IMPORTANCE (CONT'D)

**Feature importance** methods rank features by how much they contribute to the predictive performance or prediction variance of the model.

- Methods: Permutation Feature Importance, Functional Anova
- Analog in linear models: Absolute t-statistic



## FIXED MODEL VS. REFITS

- Most methods, we will discuss, analyze a fixed, trained model (e.g., permutation feature importance).
- Some methods require refitting the model (e.g., PIMP).
- Trained model  $\Rightarrow$  Model is the object of analysis.
- Refitting  $\Rightarrow$  Learning process is the object of analysis.
- The advantage of refitting is  
that it includes information about the variability in the learning process.

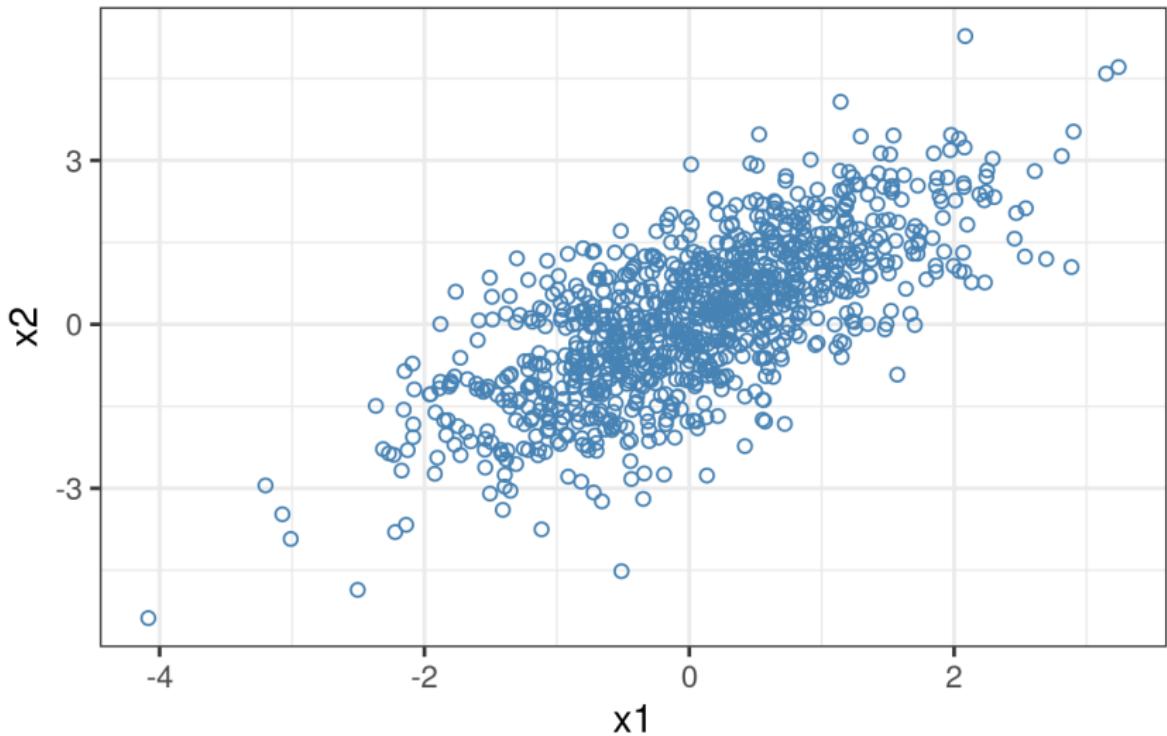
# CORRELATION VERSUS INTERACTION

- A lot of problems in IML arise due to correlated features and feature interactions. One needs to be careful not to confuse them.
- Correlated features imply that the joint probability density function (PDF) of a set of  $p$  features is not a product of their marginal distributions, i.e.,  $\mathcal{F}(x_1, \dots, x_p) \neq \mathcal{F}(x_1) \cdot \dots \cdot \mathcal{F}(x_p)$ . This often becomes a problem as many interpretation methods rely on model evaluations with recombinations of feature values. If the training data is correlated, these recombinations are located in low training density areas of the feature space, which results in model extrapolations.
- An interaction is a product term between features inside the prediction function. As such, interactions are detached from the constitution of the data, i.e., regardless of the degree of correlation, the effect of a feature on the target will depend on the values of one or multiple other features.

# CORRELATION

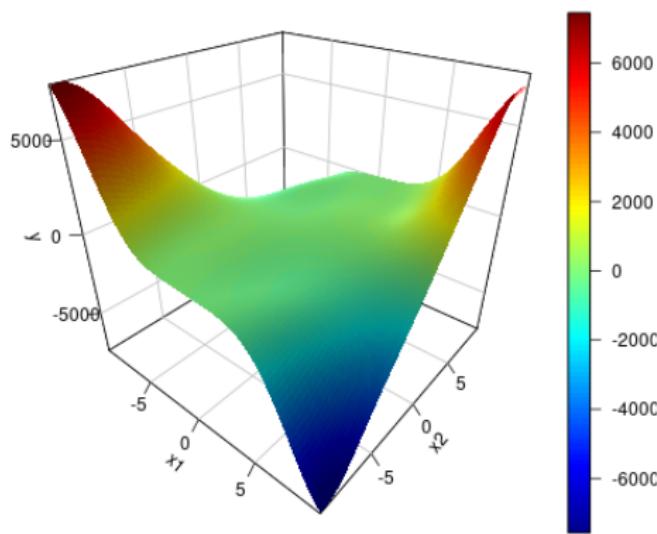
1000 randomly sampled observations with correlation between both features:

# CORRELATION



# INTERACTION

Shape of the prediction function for a two-way interaction:



# EXTRAPOLATION

- As we do not assume any data generating process, ML models cannot extrapolate reliably in areas where the model was not trained with a sufficient amount of training data. The resulting predictions are subject to a high variance.
- If we solely wish to interpret the model, this is not an issue. However, in most applications we wish to use model interpretations to evaluate some kind of real-world phenomenon that generated the data, i.e., conduct inference. In such a case, interpretations based on model extrapolations can be misleading.
- In IML, extrapolation issues often arise due to recombining values of correlated features.

# EXTRAPOLATION

- Model extrapolation essentially corresponds to prediction uncertainty, i.e., the less the training data on a feature subspace, the higher the prediction uncertainty and thus model extrapolation. Unfortunately, virtually no ML models are capable of quantifying prediction uncertainty.
- There is no consensus definition of when a model extrapolates and to what degree. Furthermore, the severity of the problem depends on the model itself. Some models might extrapolate more reliably than others.
- Theoretically, we could use the training density as a proxy. However, density estimation in high dimensions is extremely infeasible.