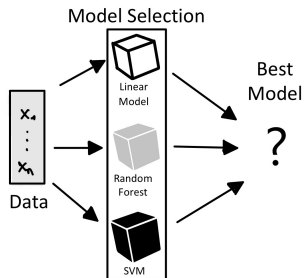


# Interpretable Machine Learning

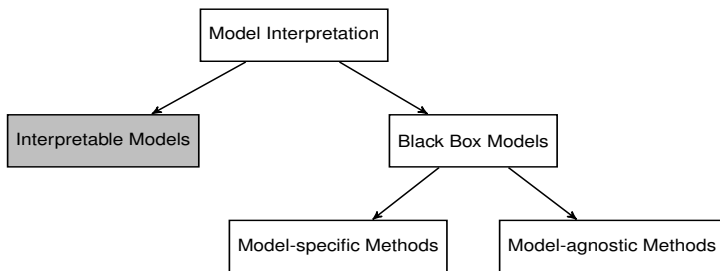
## Dimensions of Interpretability



### Learning goals

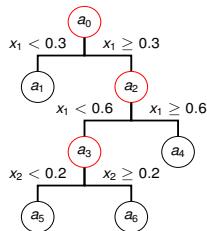
- Intrinsic vs. model-agnostic methods
- Different types of explanations
- Local vs. global methods
- Model or learner explanations – with or without refits
- Levels of interpretability

# INTRINSIC VS. MODEL-AGNOSTIC

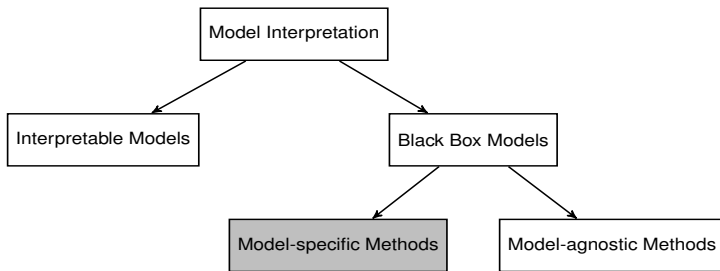


Intrinsically interpretable models:

- Examples: linear model, decision tree, decision rule, GLMs
- Interpretable because of simple model structure, e.g., weighted combination of feature values or tree structure
- Difficult to interpret with many features / complex interactions

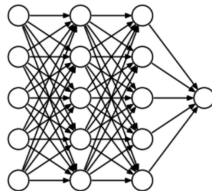


# INTRINSIC VS. MODEL-AGNOSTIC

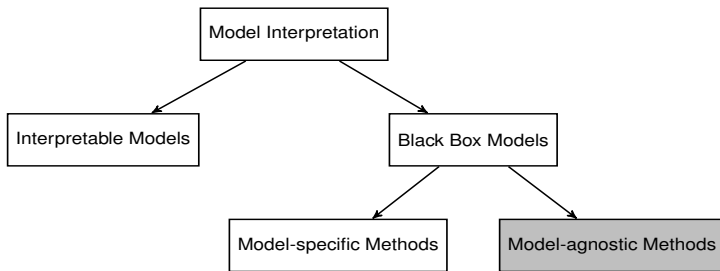


Model-specific methods:

- Interpretation method applicable to a specific ML model
- Example: implicitly integrated feature interpretation methods in tree based models, e.g., Gini Importance
- Advantage: Can exploit model structure
- Visualize activations of NNs

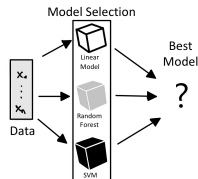


# INTRINSIC VS. MODEL-AGNOSTIC

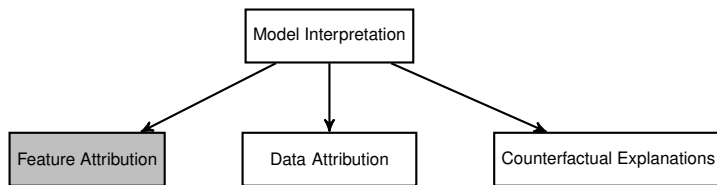


Model-agnostic methods:

- In ML: Tune over many model classes
  - ↪ Unknown which model is best / deployed
  - ↪ Need for interpretation methods applicable to any model
- Applied after training (post-hoc)
- Applicable to intrinsically interpretable models
  - ↪ provides insights into other types of explanations



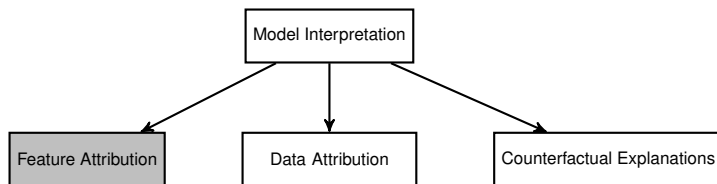
# TYPES OF EXPLANATIONS



## Feature Attribution:

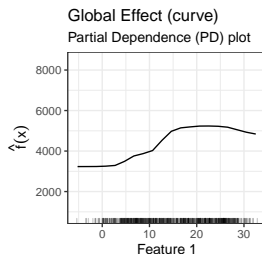
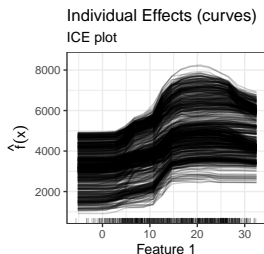
- Produce explanations on a per-feature level, e.g., feature effects or feature importance
- Vary feature values, inspect change of model prediction, model variance or model error

# TYPES OF EXPLANATIONS

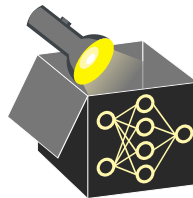
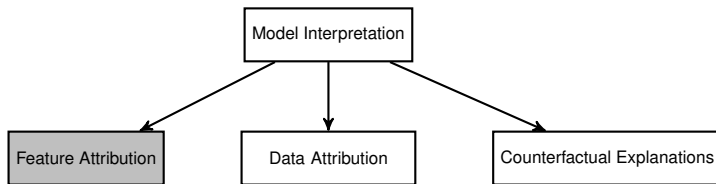


**Feature Effects** indicate the change in prediction due to changes in feature values.

- Model-agnostic methods:  
ICE curves, PD plots . . .
- Pendant in linear models:  
Regression coefficient  $\theta_j$
- Further examples: Saliency Maps, model-agnostic methods such as SHAP and LIME

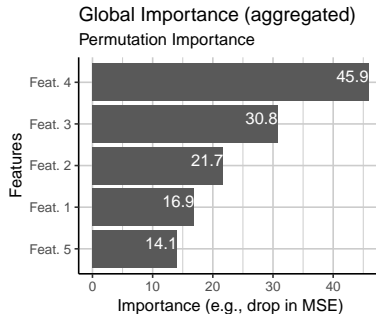


# TYPES OF EXPLANATIONS

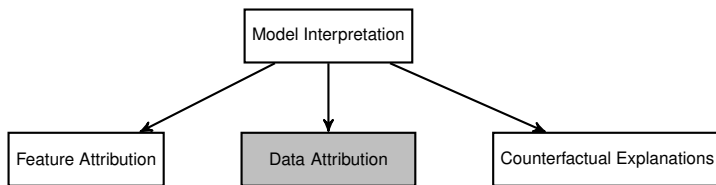


**Feature importance** methods rank features by how much they contribute to the predictive performance or prediction variance of the model.

- Model-agnostic methods: PFI, ...
- Pendant in linear models: t-statistic, p-value (significant effect)



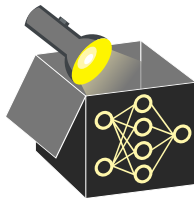
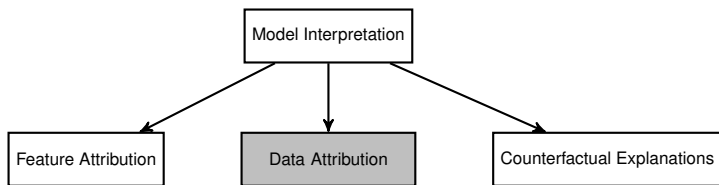
# TYPES OF EXPLANATIONS



Data Attribution: Identify training instances most responsible for a decision (e.g. Influence Functions)

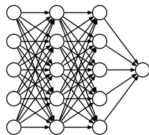


# TYPES OF EXPLANATIONS



**Data Attribution:** Identify training instances most responsible for a decision (e.g. Influence Functions)

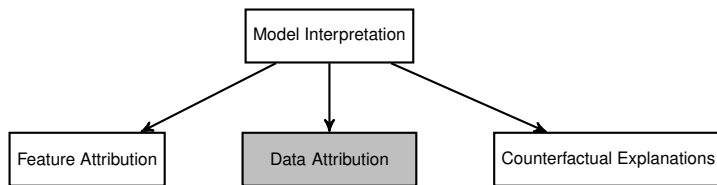
**Example:** Consider a model which should distinguish muffins and dogs



Muffin

How does this incorrect prediction come about?

# TYPES OF EXPLANATIONS



**Data Attribution:** Identify training instances most responsible for a decision (e.g. Influence Functions)

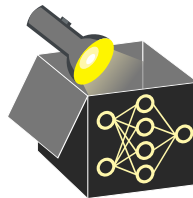
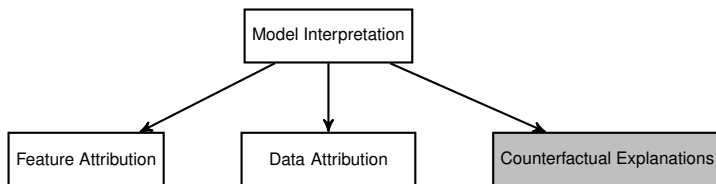
Look at training data: Which data points caused the model prediction?



Method searches for the most similar images and bases the decision on them

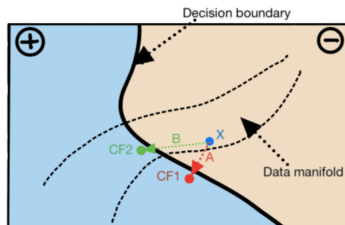
- ~ Training images looking most like new input show a muffin
- ~ Wrong output (muffin instead of dog)

# TYPES OF EXPLANATIONS

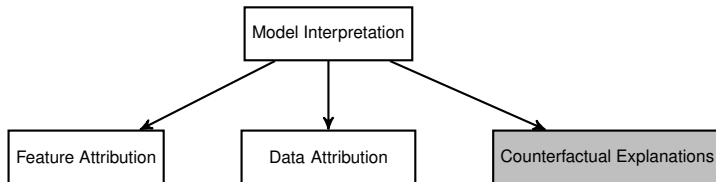


## Counterfactual Explanations:

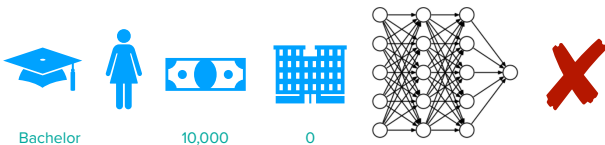
- Identify smallest necessary change in feature values so that a desired outcome is predicted
- Contrastive explanations
- Diverse counterfactuals
- Feasible & actionable explanations



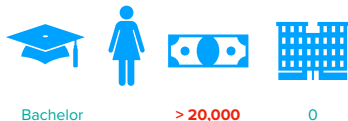
# TYPES OF EXPLANATIONS



**Example** (loan application):



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



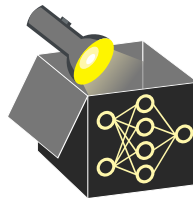
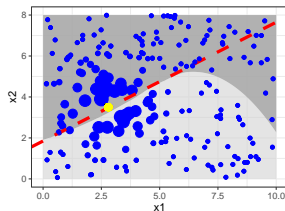
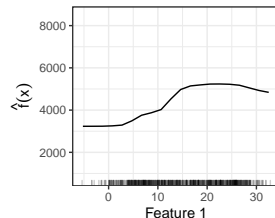
# GLOBAL VS. LOCAL

Global interpretation methods explain the model behavior for the entire input space by considering all available observations:

- Permutation Feature Importance (PFI)
- Partial Dependence (PD) plots
- Accumulated Local Effect (ALE) plots
- ...

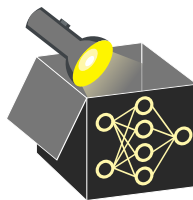
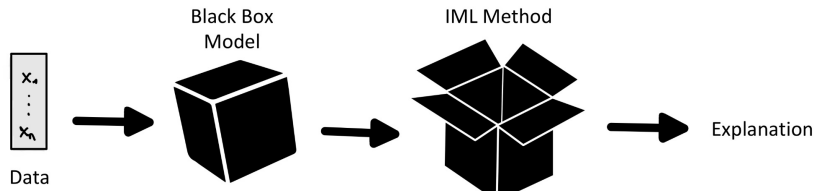
Local interpretation methods explain the model behavior for single data instances:

- Individual Conditional Expectation (ICE) curves
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley values, SHAP
- ...



# FIXED MODEL VS. REFITS

- Input of global interpretation methods: model + data, output: explanations  
~> Explanations can be viewed as statistical estimators



- Situation in ML: Deployed model is trained on all available data  
~> No unseen test data left to, e.g., reliably estimate performance  
~> IML method could use same data model was trained on  
~> But: Some IML methods rely on measuring loss requiring unseen test data
- Alternative: Explain the inducer that created the model (instead of a fixed model)  
~> Idea: Use resample strategies (e.g., 4-fold CV) as in performance estimation  
~> Requires refitting

# LEVELS OF INTERPRETABILITY

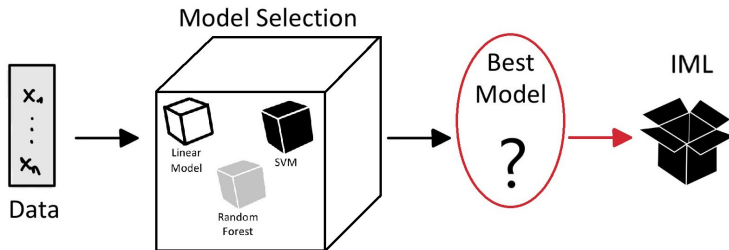
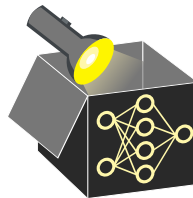
1<sup>st</sup>  
level  
view

## Research Question

How to explain a given model  
fitted on a data set?

## Objects of analysis

(deployed) model  
 $\theta \mapsto \hat{f}(\theta)$



# LEVELS OF INTERPRETABILITY

1<sup>st</sup>  
level  
view

## Research Question

How to explain a given model  
fitted on a data set?

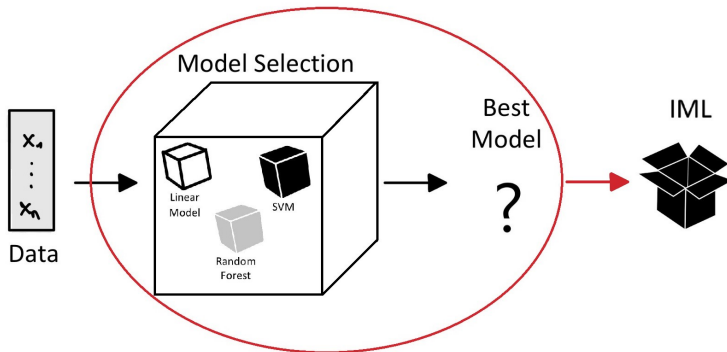
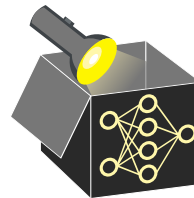
## Objects of analysis

(deployed) model  
 $\theta \mapsto \hat{f}(\theta)$

2<sup>nd</sup>  
level  
view

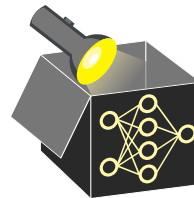
How does an optimizer choose  
a model based on a data set?

Model selection process (e.g.,  
decisions made by AutoML  
systems or HPO process)





# LEVELS OF INTERPRETABILITY



	Research Question	Objects of analysis
1 <sup>st</sup> level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$
2 <sup>nd</sup> level view	How does an optimizer choose a model based on a data set?	Model selection process (e.g., decisions made by AutoML systems or HPO process)
3 <sup>rd</sup> level view	How do data properties relate to performance of a learner and its hyperparameters?	properties of ML algorithms in general (benchmark)

