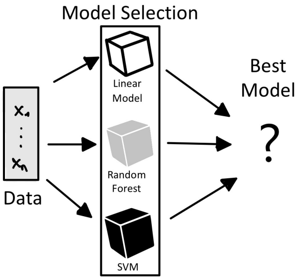
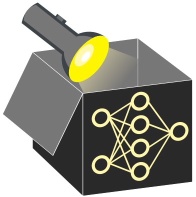
Interpretable Machine Learning

Dimensions of Interpretability

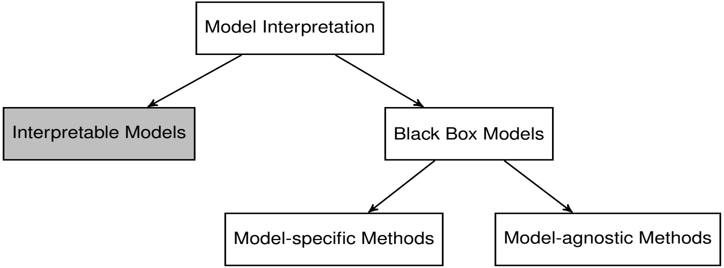


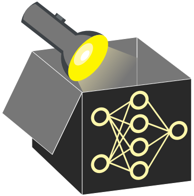
Learning goals

- Difference between intrinsic, model-specific, and model-agnostic interpretability
- Different types of explanations
- Local, global, and regional explanations
- Model/learner explanation (without/with refits)
- Levels of interpretability



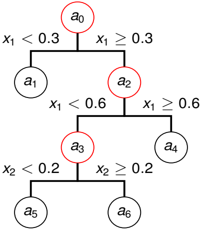
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



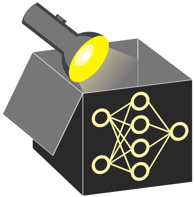


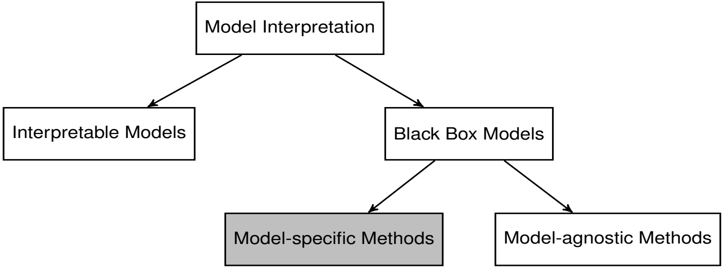
Intrinsically Interpretable Models:

- Simple model structure (e.g., weighted sum or tree)
- Examples: GLMs, decision trees
- Pro: Additional IML methods not necessarily required
- Con: Limited model complexity can reduce performance; can still be hard to interpret with many features /interactions



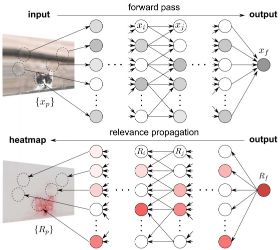
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



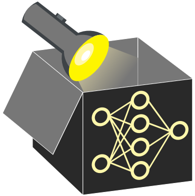


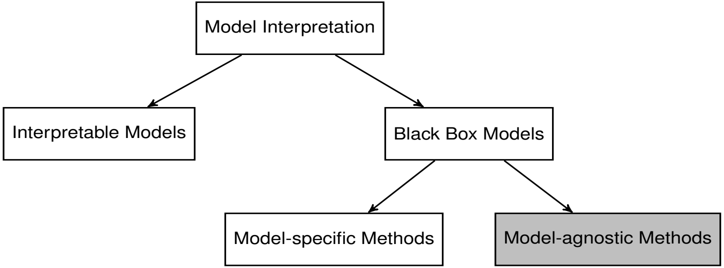
Model-specific Methods:

- Designed for specific model types (e.g., NNs)
- Examples: Gini importance of tree-based models, Layer-wise relevance propagation (LRP)
- Pro: Exploit model structure
- Con: Restricted to specific model class



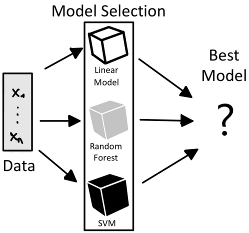
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



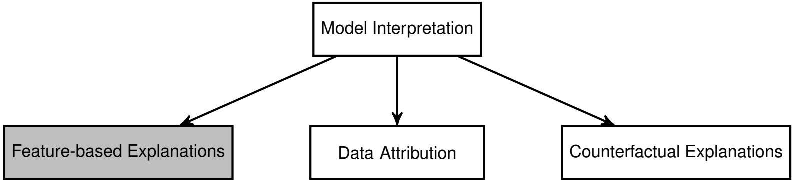


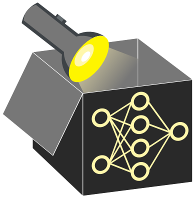
Model-agnostic Methods:

- In ML: Tune over many model classes
  ⇝ Unknown which model is best / deployed
  ⇝ Need for IML methods that work for any model
- Applied after training (post-hoc)
- Applicable to intrinsically interpretable models
  ⇝ provides insights into explanations



TYPES OF EXPLANATIONS

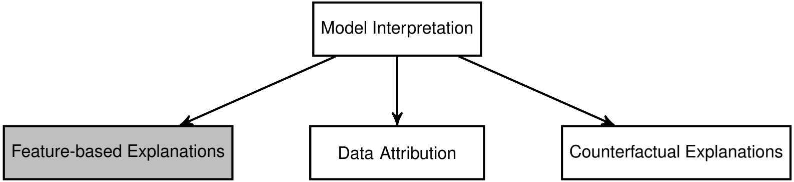


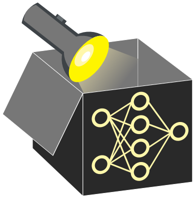


Feature-based Explanations:

- Analyze the role of individual features in model behavior.
- Types of feature-based explanations:
  - Feature Importance
  - Feature Effects
  - Feature Interactions
- Common principle: Vary or perturb feature values and observe changes in predictions, variance, or performance.

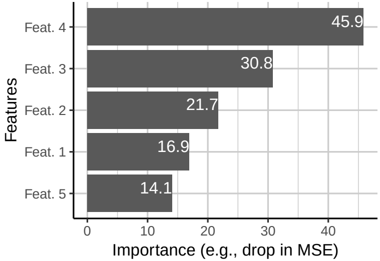
TYPES OF EXPLANATIONS



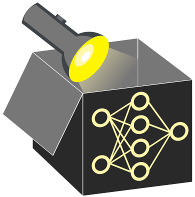


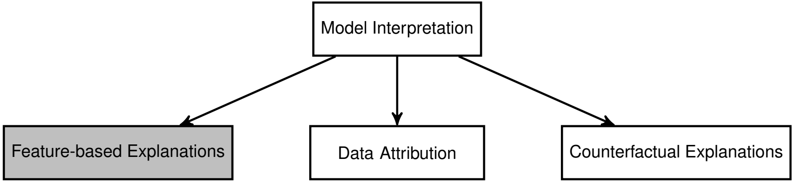
Feature Importance quantifies relevance of features, e.g., their contribution to model prediction, predictive performance, or prediction variance.

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



TYPES OF EXPLANATIONS



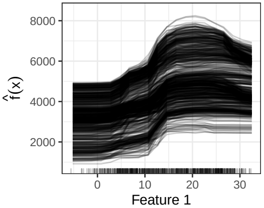


Feature Effects indicate changes (direction and magnitude) in model prediction due to changes in feature values.

- Model-agnostic methods: ICE curves, PD plots . . .
- Pendant in linear models: Weights / coefficients θ_j
- Further examples: ALE, SHAP, and LIME

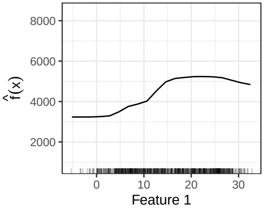
Individual Effects (curves)

ICE plot

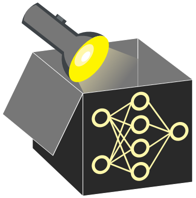


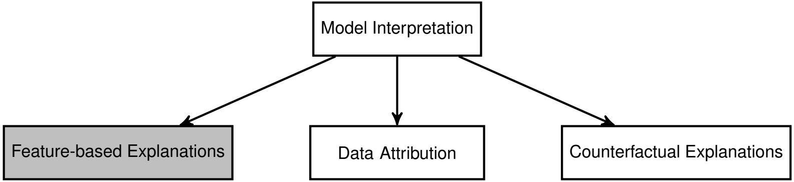
Global Effect (curve)

Partial Dependence (PD) plot



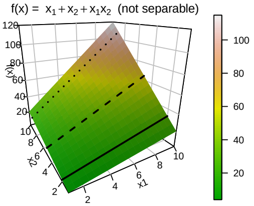
TYPES OF EXPLANATIONS

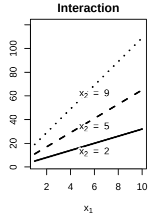




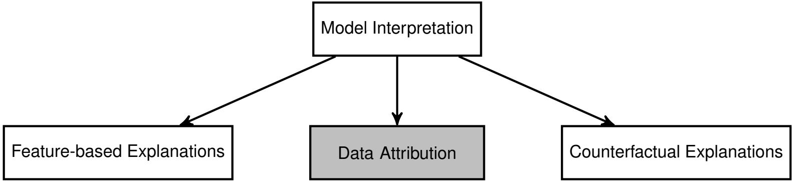
Feature Interaction: How combinations of features jointly affect predictions.

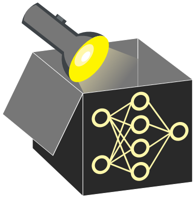
- Model-agnostic methods: Friedman's H-statistic
- Pendant in linear models: Coefficients of interaction terms θ_{jk}





TYPES OF EXPLANATIONS

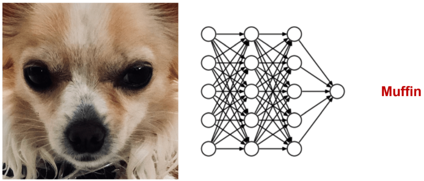




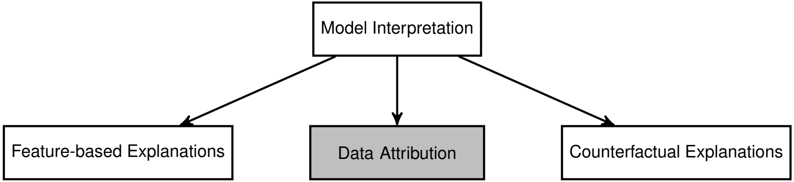
Data Attribution: Identify which training instances most influenced a prediction.

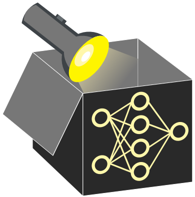
Example: A model should distinguish muffins and dogs.

Question: Why does it misclassify this dog image (test point) as a muffin?



TYPES OF EXPLANATIONS





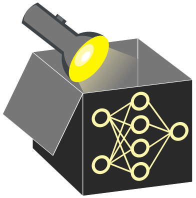
Data Attribution: Identify which training instances most influenced a prediction.

Example: A model should distinguish muffins and dogs.

Approach: Measure how perturbations to training instances affect prediction or loss.

- ⇝ Influential training instances drive prediction of test points.
- ⇝ If these resemble muffins, the model may predict muffin instead of dog.

TYPES OF EXPLANATIONS

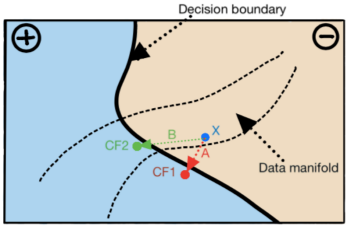


Model Interpretation

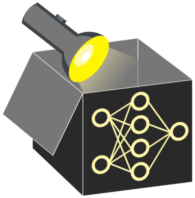
- Feature-based Explanations
- Data Attribution
- Counterfactual Explanations

Counterfactual Explanations:

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



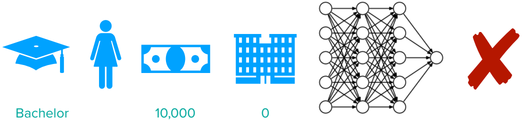
TYPES OF EXPLANATIONS



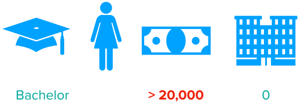
Model Interpretation

Feature-based Explanations | Data Attribution | Counterfactual Explanations

Example (loan application):



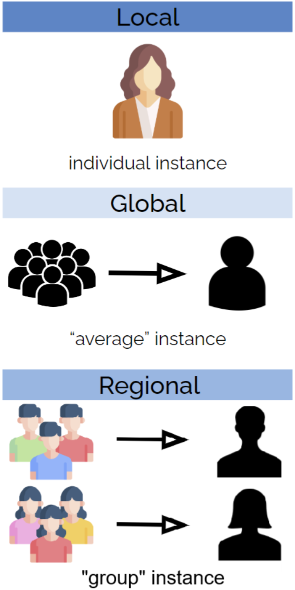
What can a person do to obtain a favorable prediction from a given model ?

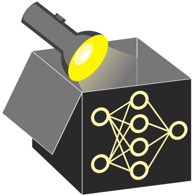


LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

Local – Explain model behavior for **single instances**:

- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE





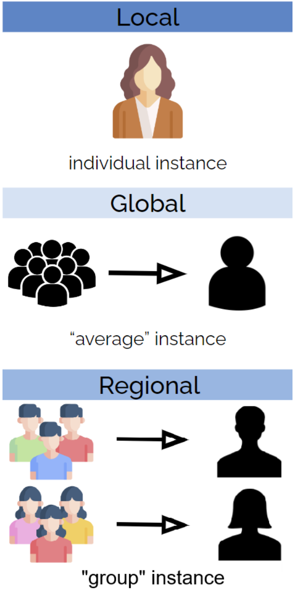
LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

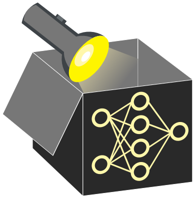
Local – Explain model behavior for **single instances**:

- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE

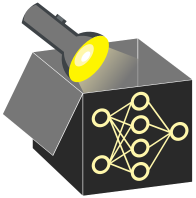
Global – Explain model behavior for **entire input space**:

- Provide high-level insights into model behavior, often by aggregating local explanations
- Easier to communicate but loss of detail & over-simplification (hides differences)
- Examples: PD plots, ALE plots, PFI





LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS



Local – Explain model behavior for **single instances**:

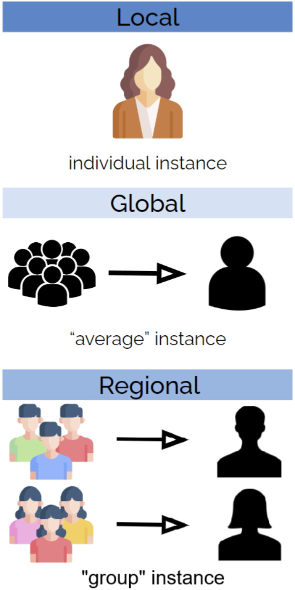
- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE

Global – Explain model behavior for **entire input space**:

- Provide high-level insights into model behavior, often by aggregating local explanations
- Easier to communicate but loss of detail & over-simplification (hides differences)
- Examples: PD plots, ALE plots, PFI

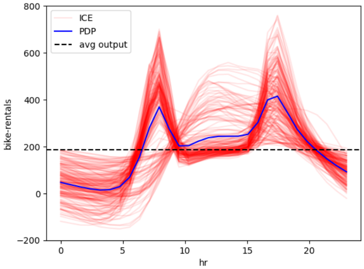
Regional explanations – for **subspaces** / **regions**:

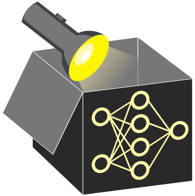
- Compromise between nuanced & high-level insights
- Useful when local explanations group well without losing much detail



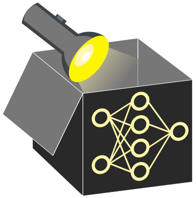
LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

- **Local** (red): ICE curves for one instance
  $\rightsquigarrow$ Detailed but cluttered/obscure pattern
- **Global** (blue): PDP averaged over *all* days
  $\rightsquigarrow$ Averaged curve hides heterogeneity



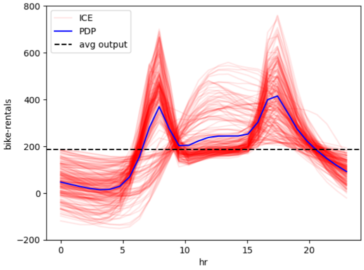


LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

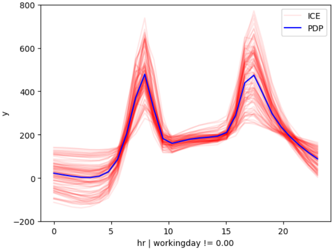


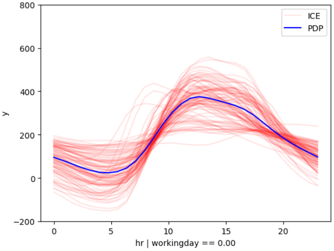
- **Local** (red): ICE curves for one instance
  $\rightsquigarrow$ Detailed but cluttered/obscure pattern
- **Global** (blue): PDP averaged over *all* days
  $\rightsquigarrow$ Averaged curve hides heterogeneity
- **Regional**: Split data on `workingday`
  - Region 1: morning and evening peak
  - Region 2: late-morning leisure peak

  $\rightsquigarrow$ Preserves detail without overload (challenge: find regions automatically)



Region 1 (working day)

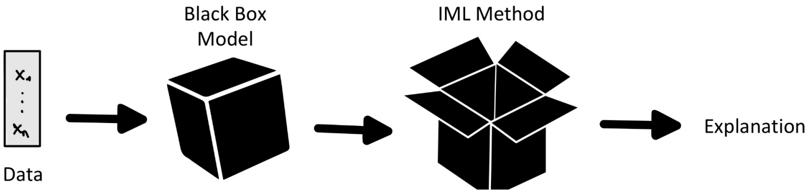




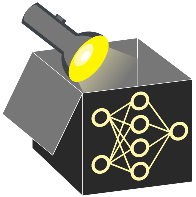
Region 2 (non-working day)

FIXED MODEL VS. REFITS

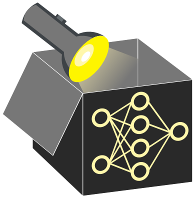
- Input of global interpretation methods: model + data, output: explanations
  ⇝ Explanations can be viewed as statistical estimators



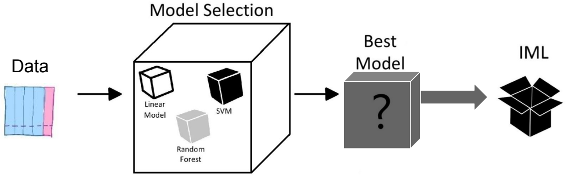
- Situation in ML: Deployed model is trained on all available data
  ⇝ No unseen test data left to, e.g., reliably estimate performance
  ⇝ IML method could use same data model was trained on
  ⇝ But: Some IML methods rely on measuring loss requiring unseen test data
- Alternative: Explain the inducer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



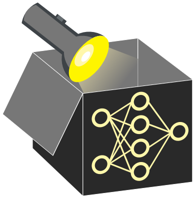
LEVELS OF INTERPRETABILITY



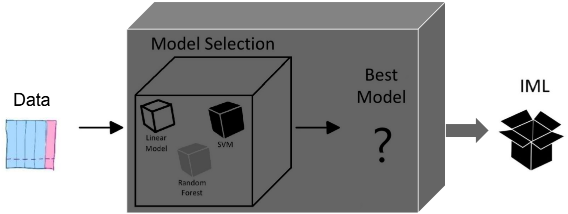
	Research Question	Objects of analysis
1^{st} level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$



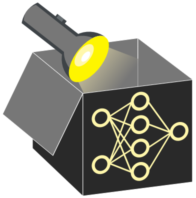
LEVELS OF INTERPRETABILITY



	Research Question	Objects of analysis
1^{st} level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$
2^{nd} 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^{st} level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$
2^{nd} 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^{rd} level view	How do data properties relate to performance of a learner and its hyperparameters?	properties of ML algorithms in general (benchmark)