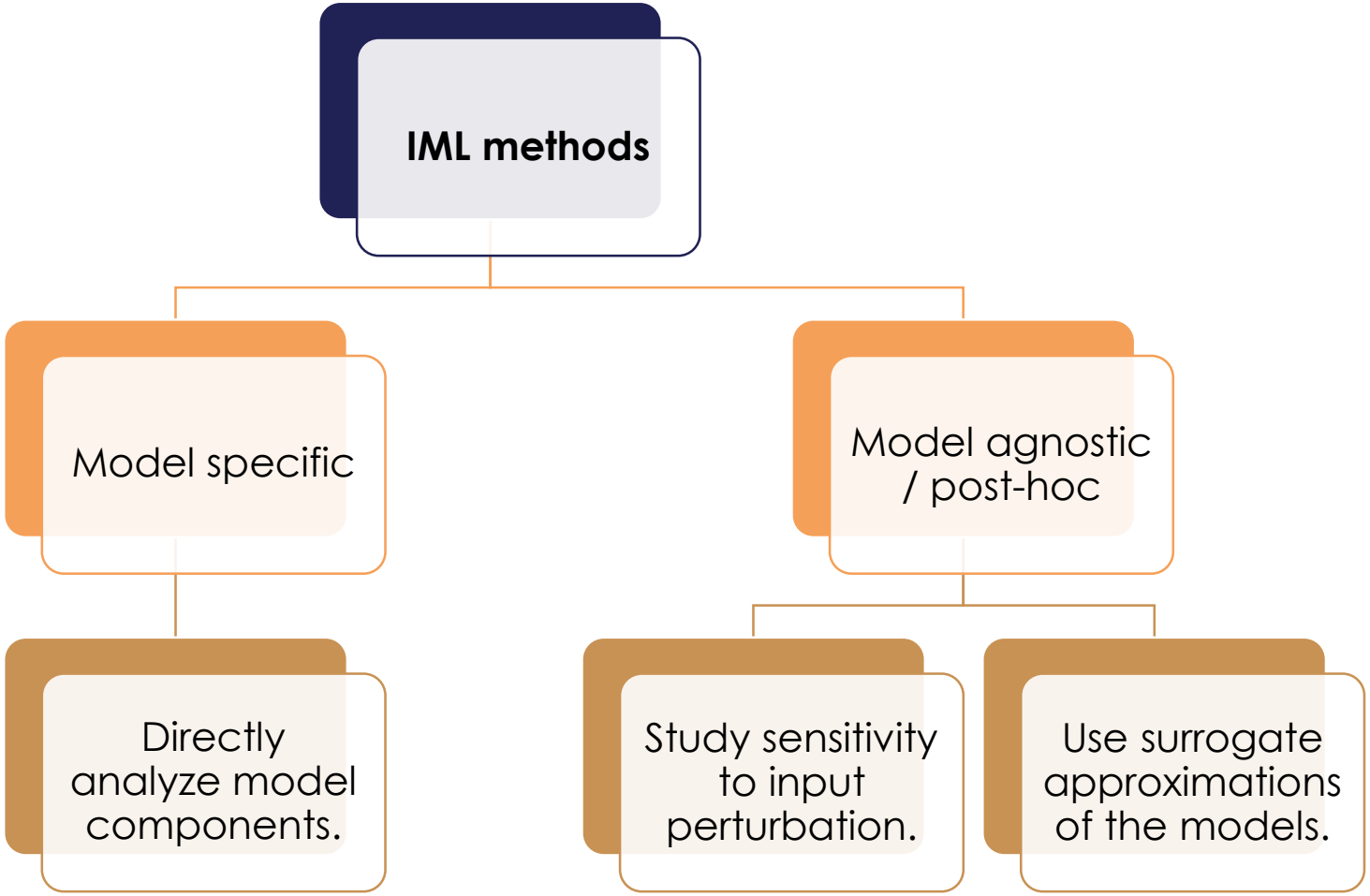
# Interpretability methods

# IML methods

Methods or algorithms that allow us to extract information from ML models.

IML methods allow us to "*explain*" the outputs of machine learning models.

# IML methods


IML methods
Model specific
Model agnostic / post-hoc
Directly analyze model components.
Study sensitivity to input perturbation.
Use surrogate approximations of the models.

# Intrinsically explainable models

Intepreting model components

# Interpreting model components

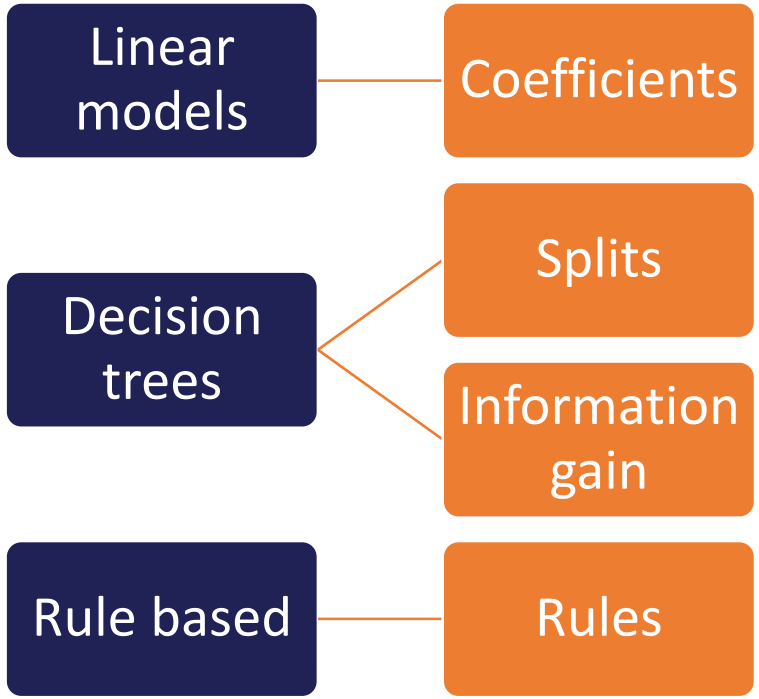
Intrinsically explainable models

- Linear models
- GAMS
- Decision Trees
- Constrained GBMS
- Rule based models
- KNNs

Some machine learning models are interpretable by design.

# Interpreting model components

By analyzing the model parameters (components), we can understand how they produce the predictions.

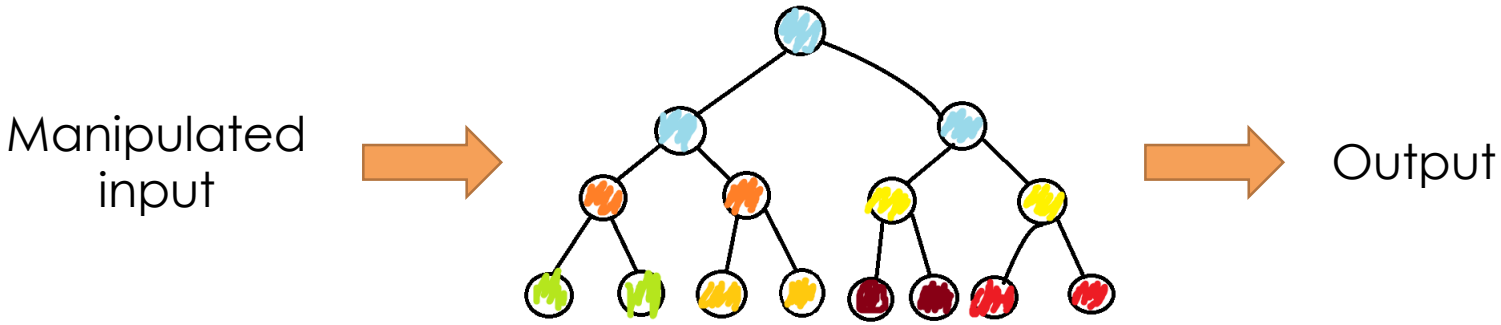


# Sensitivity methods

Altering the input data

# Sensitivity methods

Manipulate the input data and analyze the effect on the model prediction.




The greater the change in the output ➔ the more relevant the manipulation.

# Sensitivity methods

- ✓Permutation feature importance
- ✓Partial Dependency Plots
- ✓Feature hiding
- ✓SHAP

# Surrogates

Predicting the model's predictions

# Surrogates

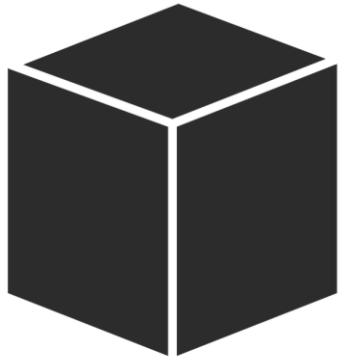
Interpretable models are trained to predict the predictions of the black box models.

- Linear models
- Decision trees

Analyze the components of the surrogates to try and understand what is driving the black box predictions.

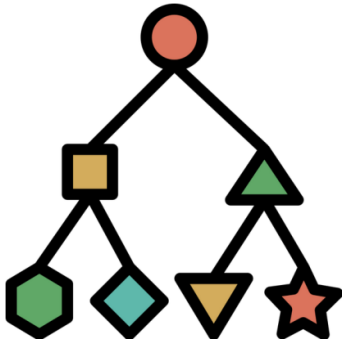
# Surrogates

X , y



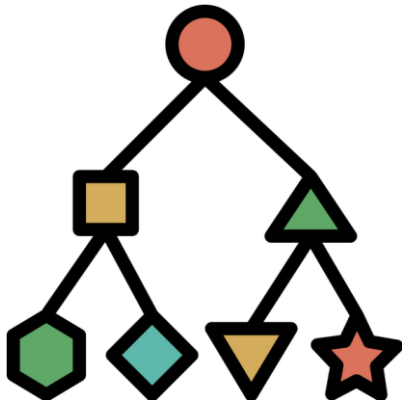
Predictions

X , predictions



Output

# Surrogates



Analyze the components of the intrinsically explainable model.

Proxy to explain the black-box.

If it is important for the white box, it must also be important for the black box… right?

# Surrogates

No guarantee that they faithfully represent the black box model.

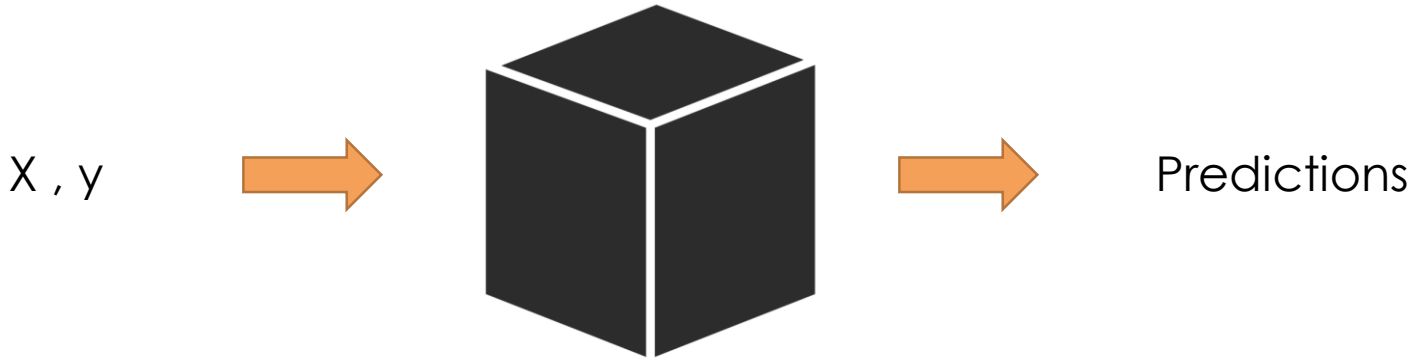
Don’t use them on their own.

# Mixed

LIME: combines a surrogate with a perturbation method.

# A word of caution

# Challenges

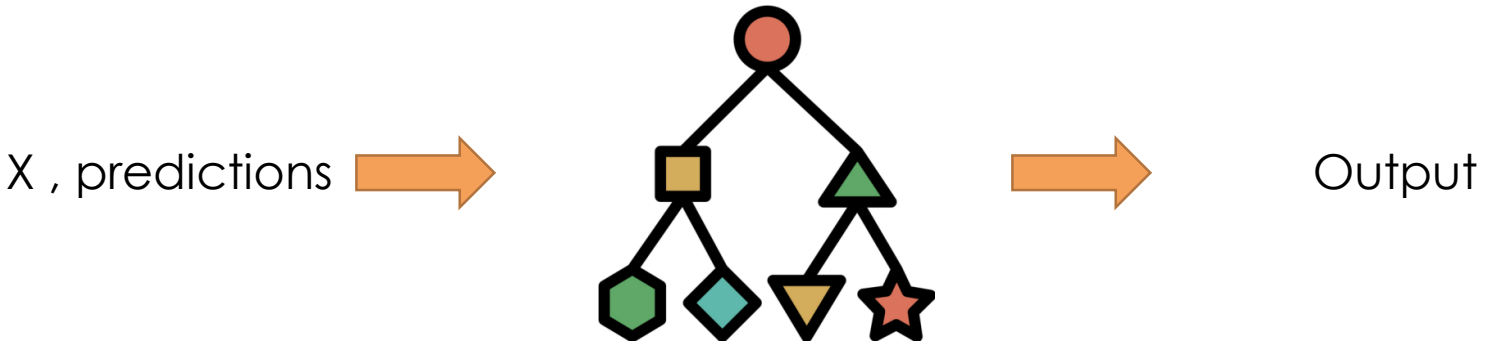



**Model performance**: we compare the *predictions* to *y* ➔ we know how well the model is doing.

# Challenges

**Quality of the explanations**: we don't know what the real explanation is.

- There is no way to assess if the surrogate model represents the behaviour of the black box accurately.



# IML methods

**Whenever possible, use intrinsically interpretable machine learning models.**

Train In Data

# THANK YOU

www.trainindata.com