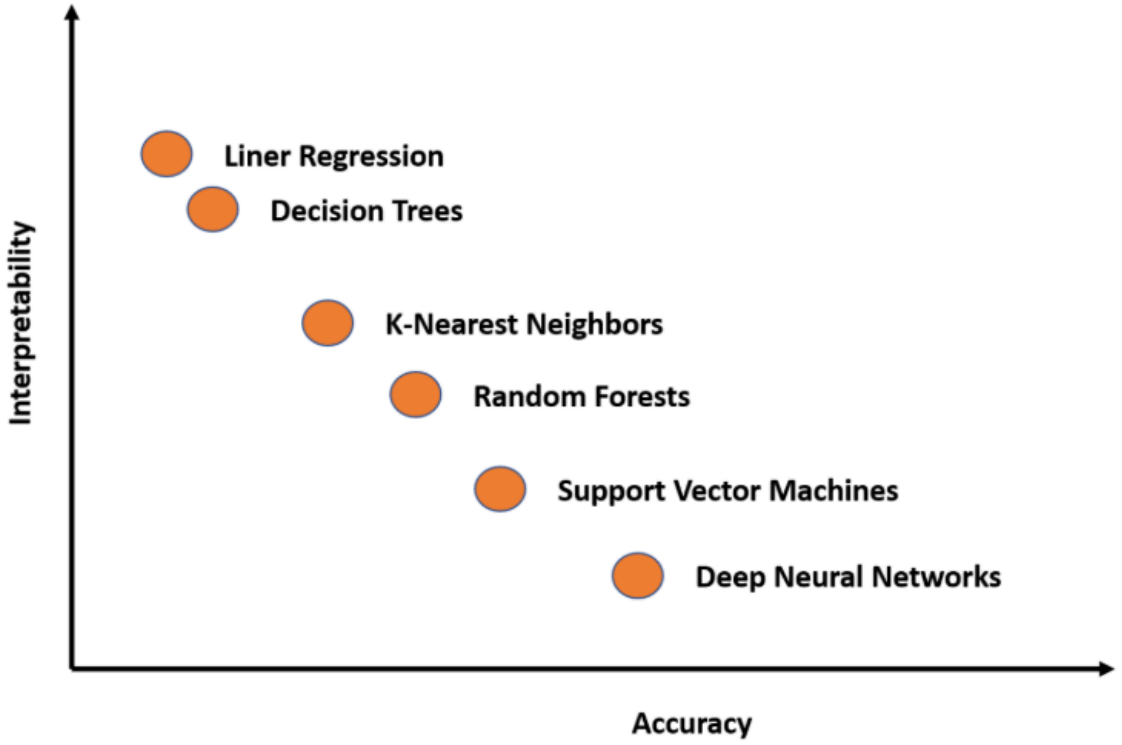
Interpretable Machine Learning

Part 1

Brief introduction

What is interpretability?

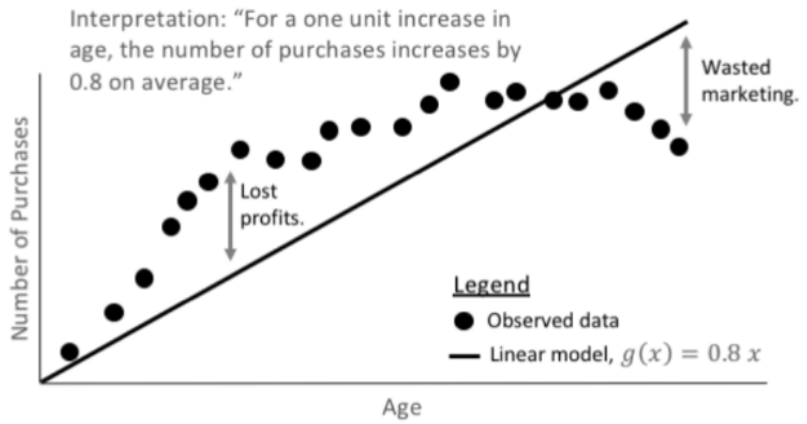


- Why do we need to interpret?
  - Fairness
  - Privacy
  - Reliability or Robustness
  - Causality
  - Trust

What is interpretability?

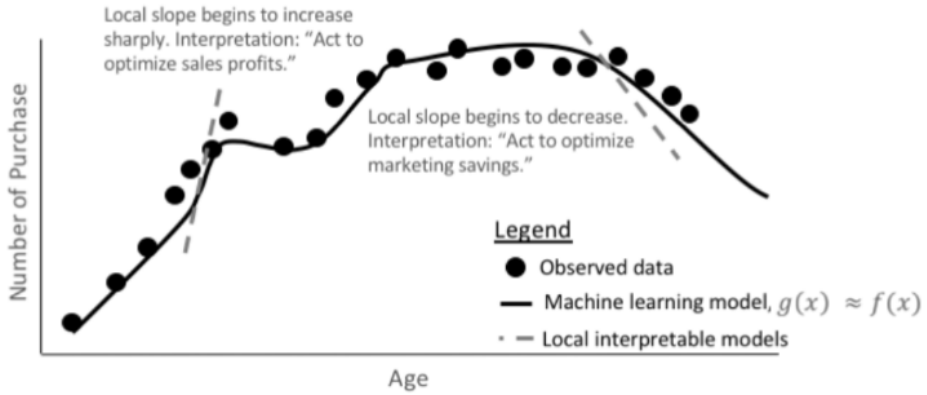
Linear Models

Exact explanations for ***approximate*** models.



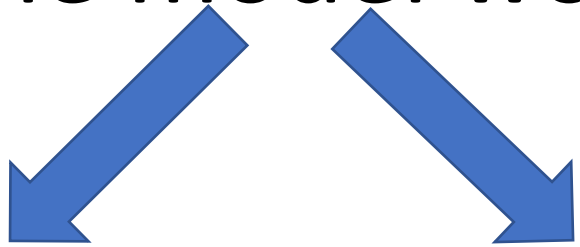
Machine Learning

Approximate explanations for ***exact*** models.



H$_2$O.ai

What if the model works well?



Trust the model

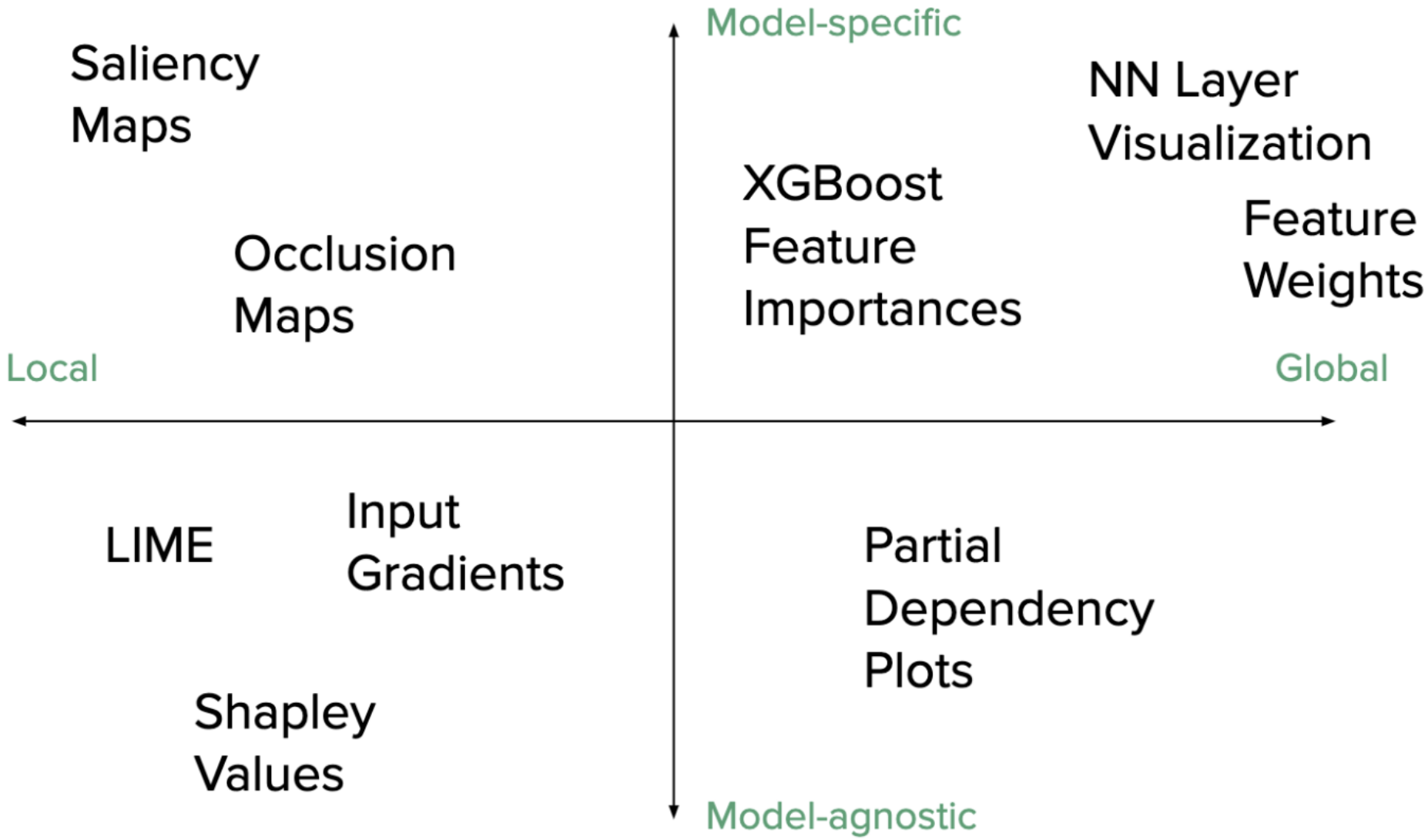
Interpret the result

When we do not need to interpret?

- No significant impact
- Self interpretable
- The problem is well studied for example:

  Optical character recognition

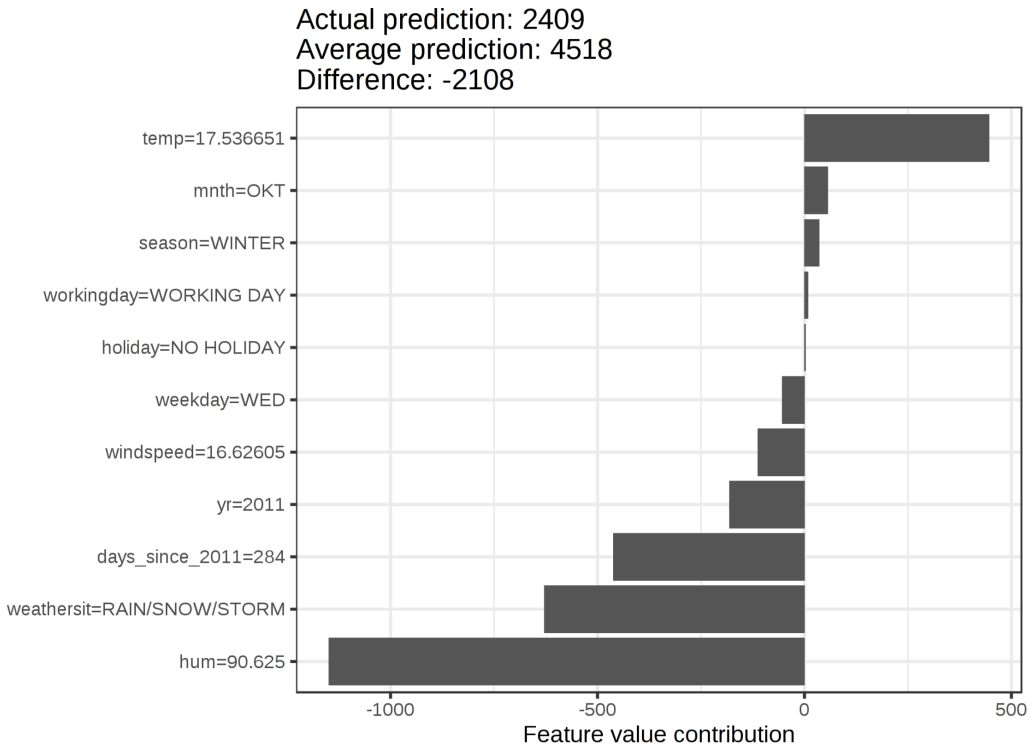
What if the model works well?



Outcomes

- Feature summary statistic
- Feature summary visualization
- Model internals (learned weights)
- Data point
- Intrinsically interpretable model

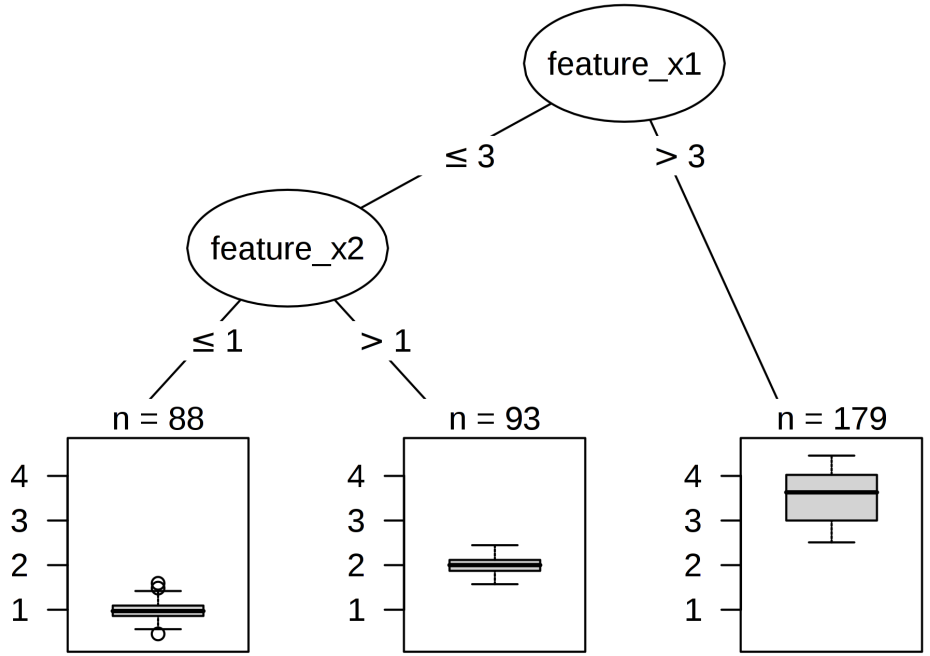
Feature summary statistic/visualization



- Be careful!
- The importance is only relative.

Model internals (decision tree)

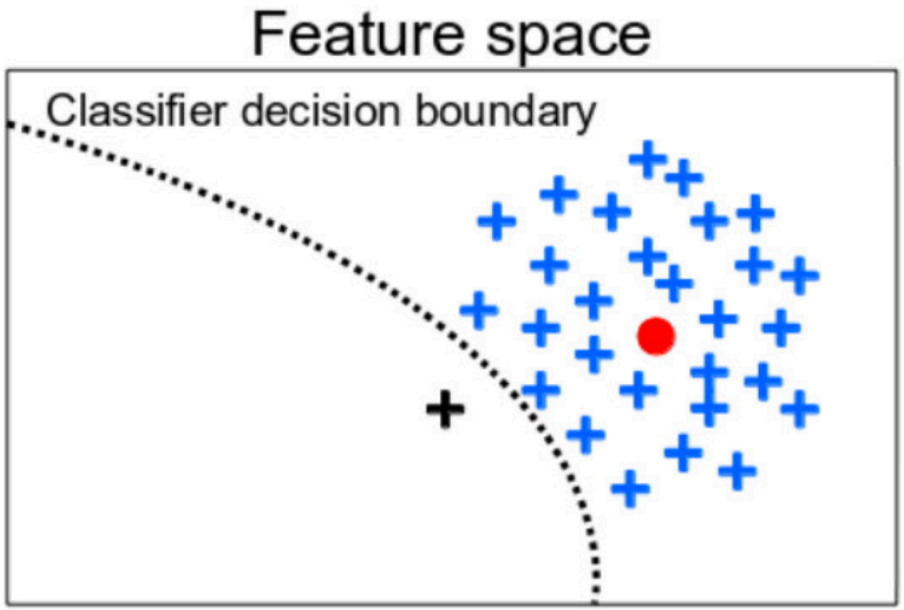
- Ideal for capturing interactions
- Natural visualisation



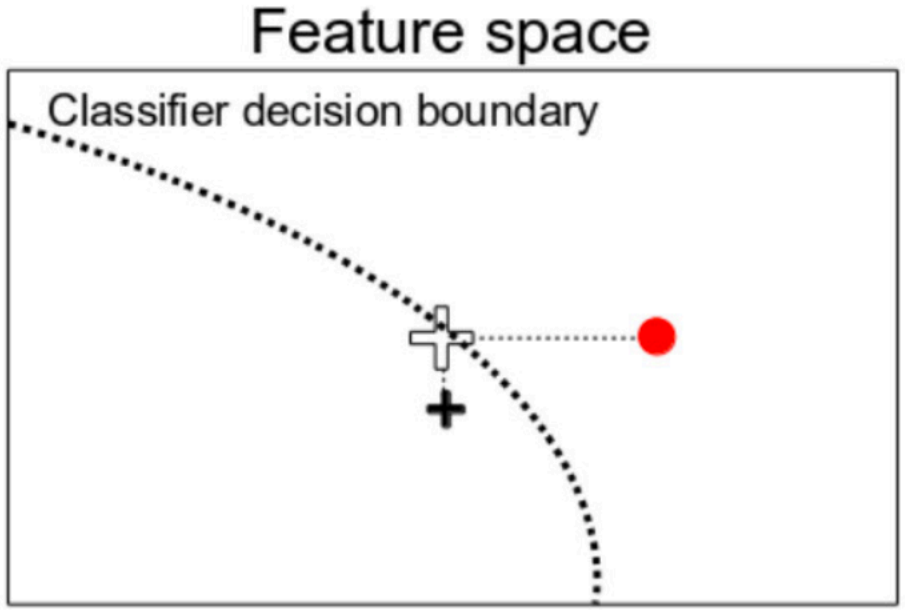
But…

- Lack of smoothness

Data point



Step 1: Generation



Step 2: Feature Selection

Properties of explanation

- Expressive Power
- Translucency
- Portability
- Algorithmic Complexity

Expressive power

- How understandable is your explanation?

What will you choose?

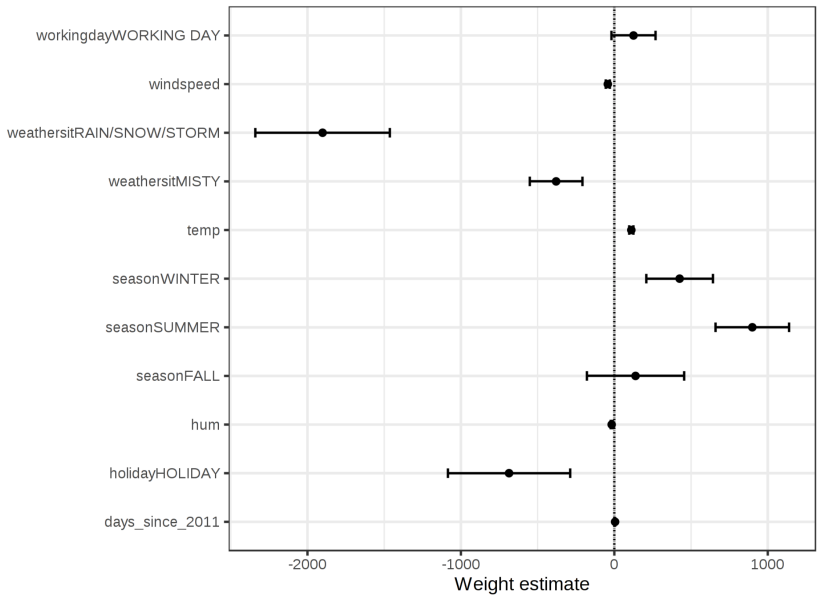
$$RS = \sum_{i=1}^{n} \ln(OR_i)\,\theta_i$$

Where θ_i is the vectors of parameters of a patient

Name of exposure	**Risk score(max 2)**
Risk factor 1	2
Risk factor 2	1
Risk factor 3	1
Risk factor 4	1

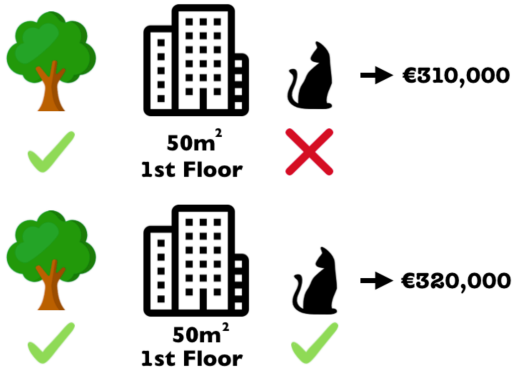
Translucency

- Do you use the structure of the model?



ML model

Counterfactual Explanations

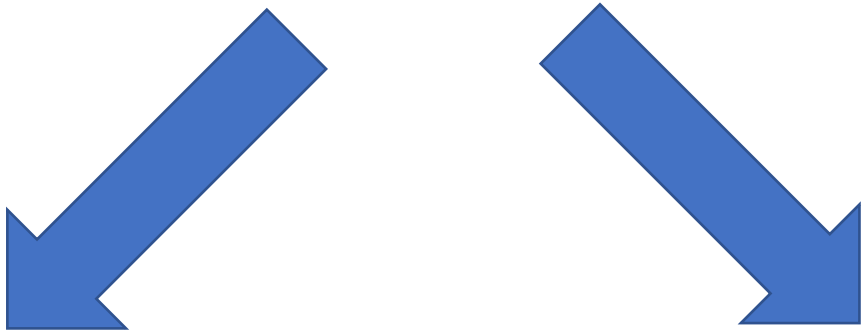


Properties of explanation (individual)

- Accuracy
- Fidelity
- Consistency
- Stability
- Comprehensibility
- Certainty
- Degree of Importance
- Novelty
- Representativeness

Accuracy and Fidelity

Does your explanation cover all the data?



Local

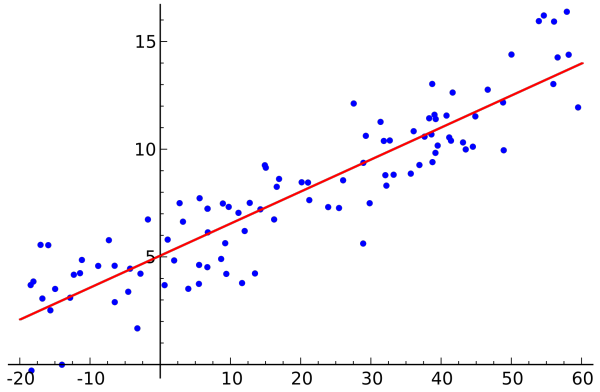
1) Local surrogate models
2) Shapley values

Global

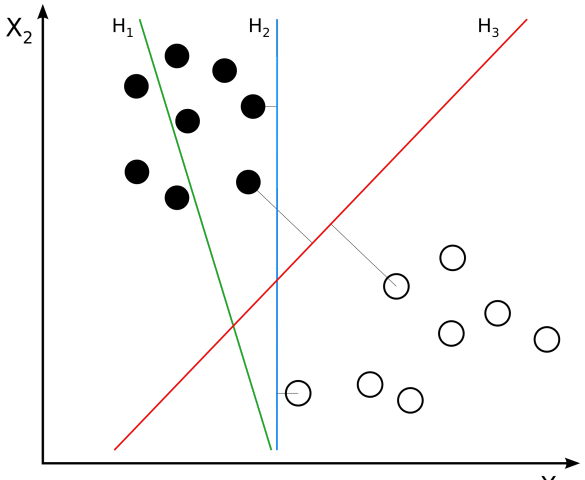
1) Decision Tree
2) Decision Rules

Consistency and Stability

Linear regression



Support vector machine

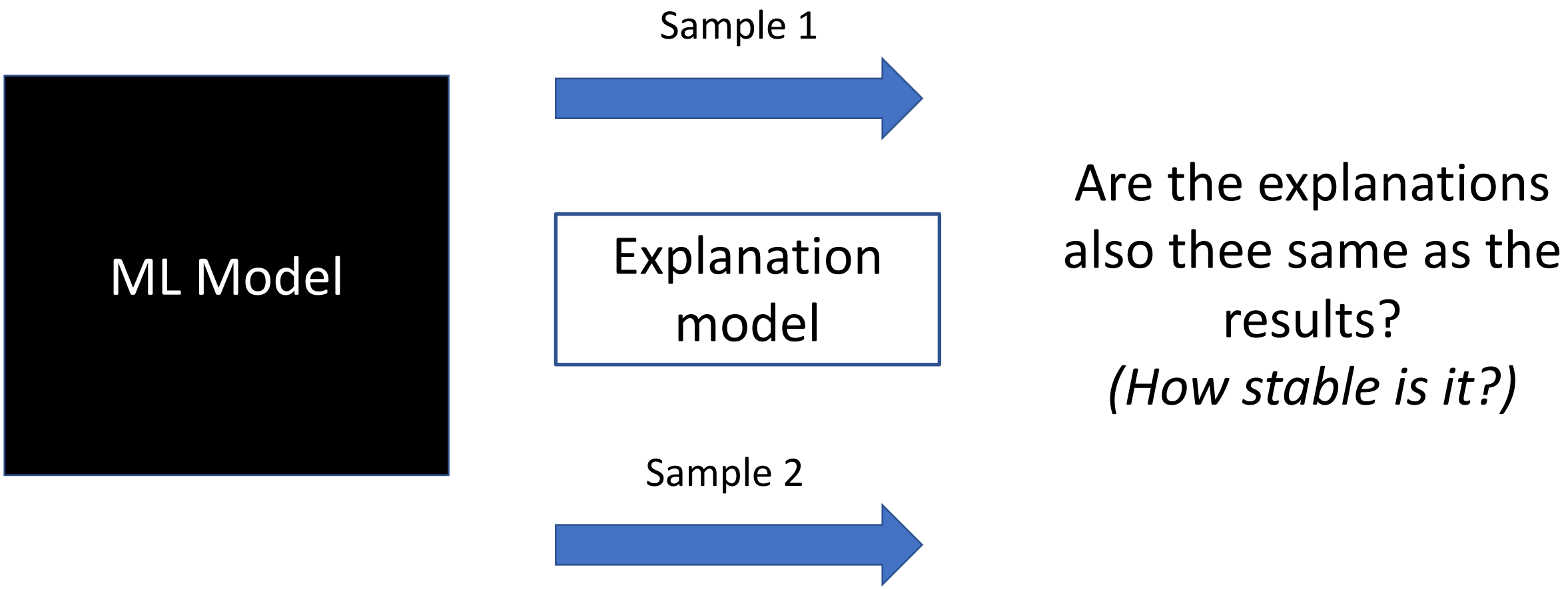


Explanation model

Are the explanations also thee same as the results?

(How consistent is it?)

Consistency and Stability



Are the explanations also thee same as the results?

(How stable is it?)

Certainty and Novelty

Does the prediction of value X have the same certainty in all cases?

Logistic regression

Risk score

Are the risk values the same?
Which data was used?

Sources / Additional links

- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
- “Guide to Interpretable Machine Learning”, Matthew Stewart, https://towardsdatascience.com/guide-to-interpretable-machine-learning-d40e8a64b6cf.
- ”Machine learning interpretability”, Patrick Hall, https://github.com/jphall663/GWU_data_mining/blob/master/10_model_interpretability/notes/MLI_good_bad_ugly.pdf.