Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Exercise 10

Lars Sipos

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

28.06.2022

Introductory Questions

» Local vs. Global explanations?

» Intrinsic vs. Post-hoc Interpretability Techniques?

Intrinsic vs. Post-hoc Explanation Methods?

Intrinsic Methods

» Decision Trees

» Linear Models (e.g. Linear Regression)

» Scalable Bayesian Rule Lists (SBRLs)

» Generalized Additive Models with pairwise interactions (GA^2M)

Post-hoc Methods

» Local Interpretable Model-agnostic Explanations (LIME)

» SHapley Additive exPlanations (SHAP)

» Explainable Neural Networks (XNNs)

» Gradient-weighted Class Activation Mapping (Grad-CAM)

Introductory Questions

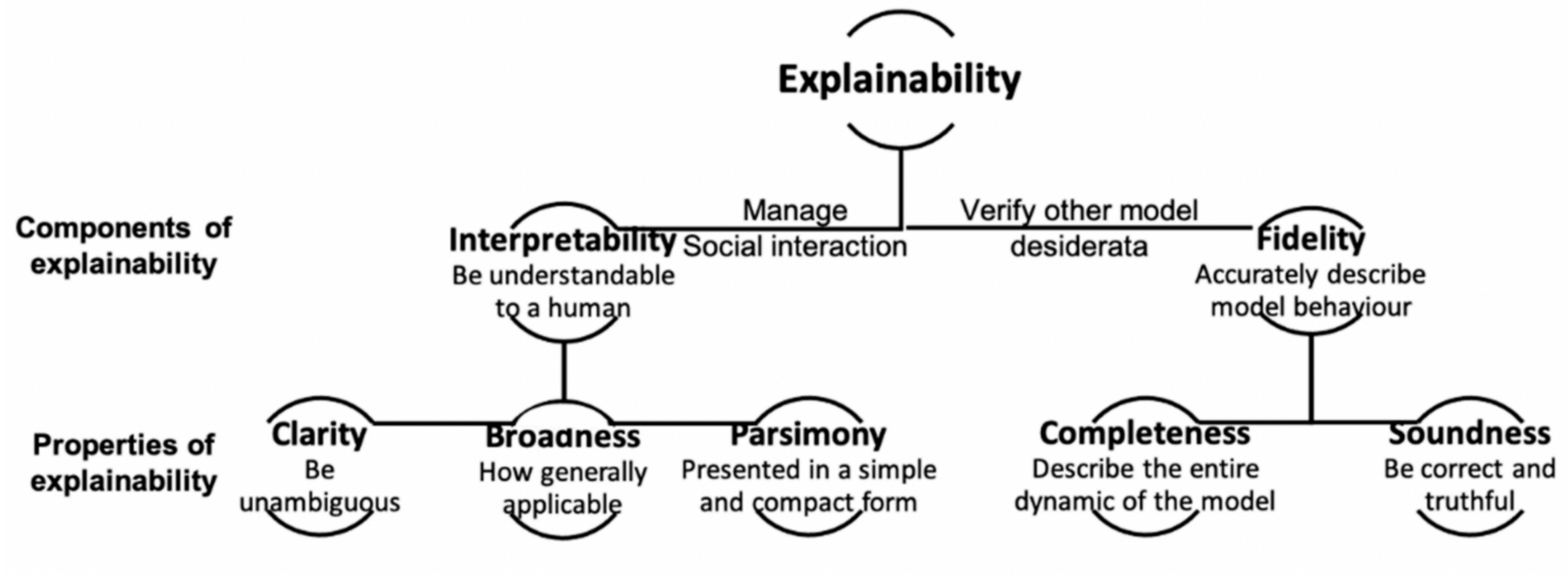
» Local vs. Global explanations?

» Intrinsic vs. Post-hoc Interpretability Techniques?

» Interpretability vs. Explainability?

Explainability and Related Properties

In the data science community the terms explainability and interpretability have often been used synonymously, but there have been attempts to make a distinction:



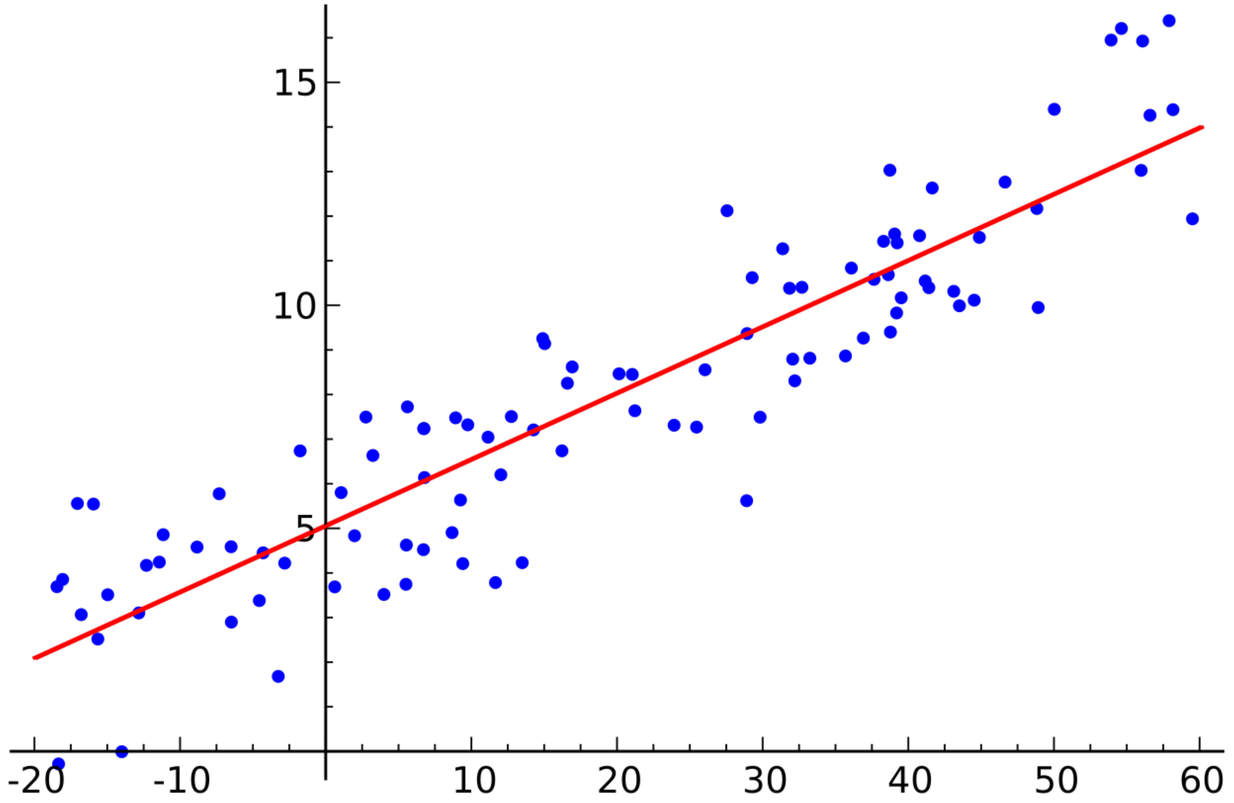
J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics," Electronics, vol. 10, no. 5, p. 593, Mar. 2021, doi: 10.3390/electronics10050593.

SHAP Demonstration

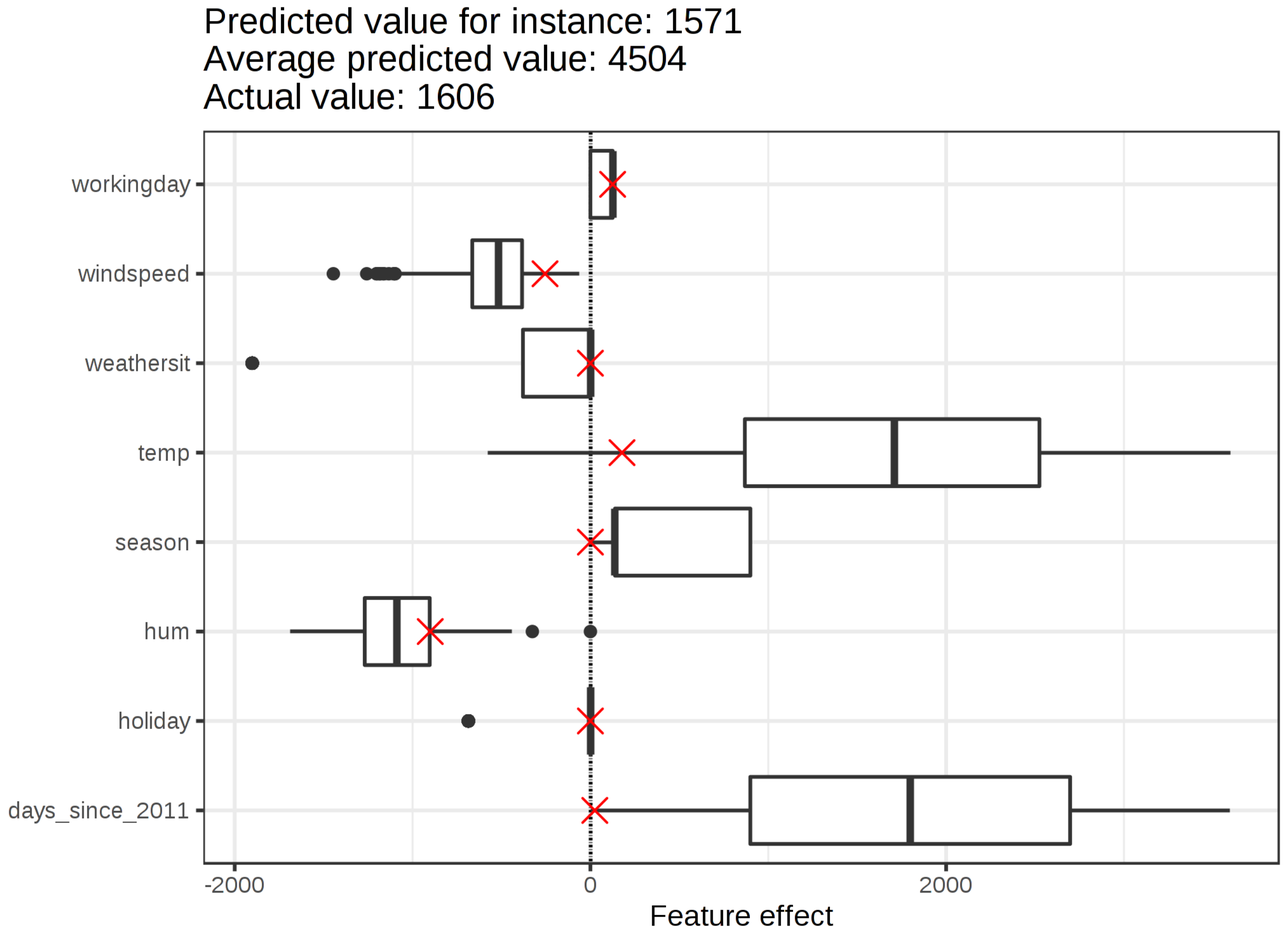
You can follow along here:

https://github.com/FUB-HCC/hcds-summer-2022/tree/main/exercise/tasks/ex10_explanations/shap_demonstration.ipynb

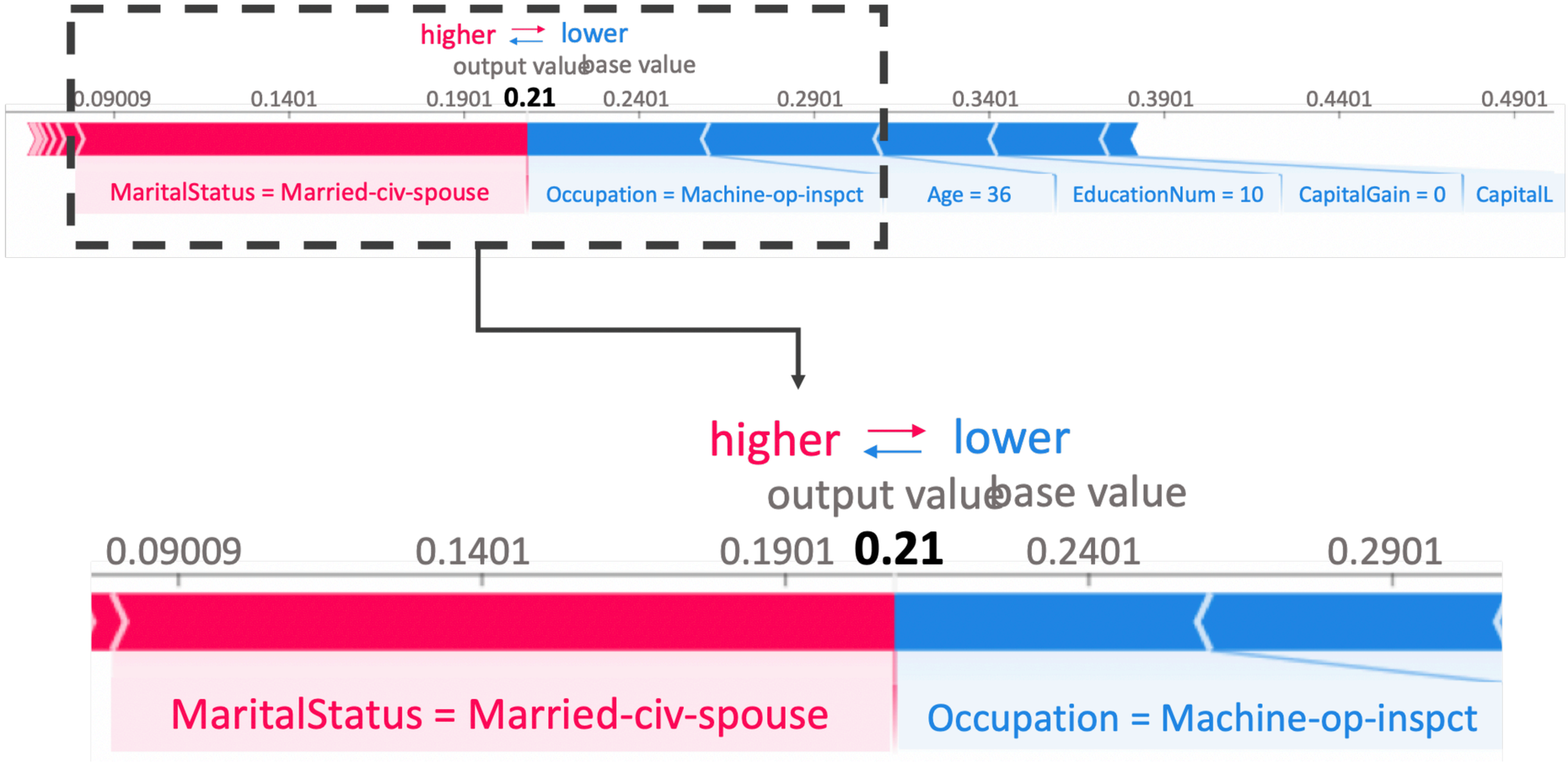
Interpretable to Whom?



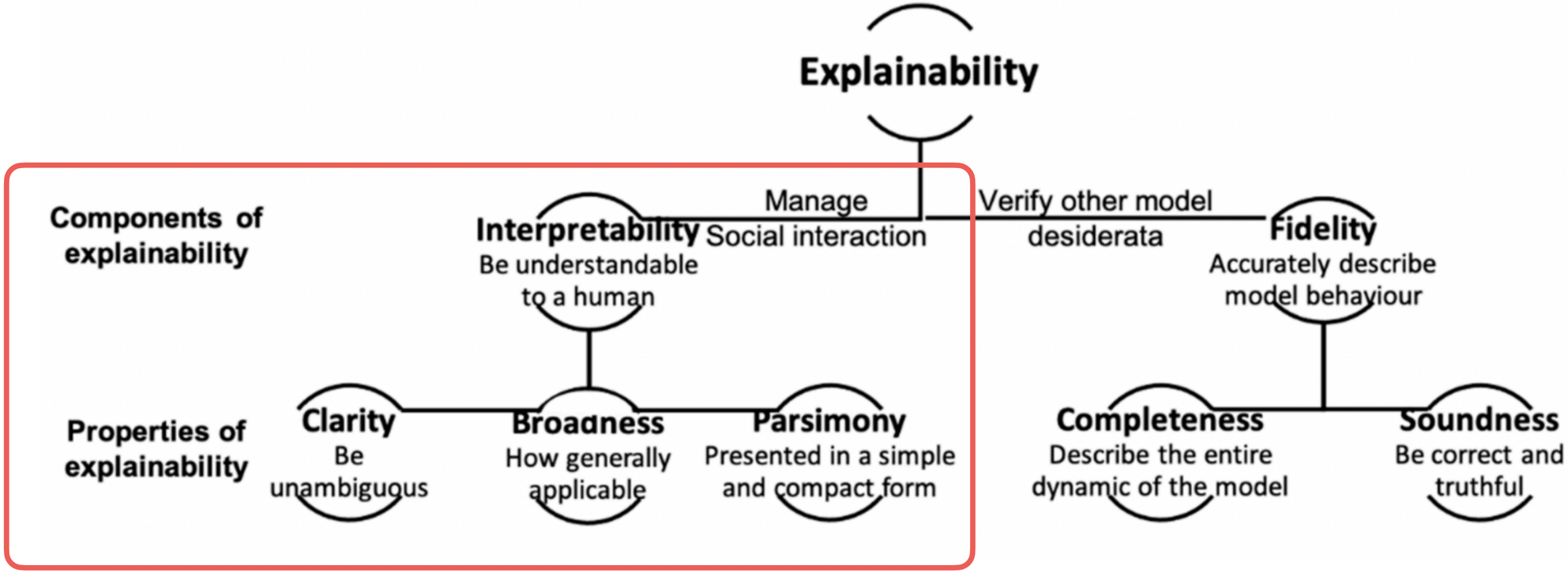
Interpretable to Whom?



Interpretable to Whom?



Explainability and Related Properties



J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics," Electronics, vol. 10, no. 5, p. 593, Mar. 2021, doi: 10.3390/electronics10050593.

Types of Explanations

» **Rationale explanation**: The reason that led to a ML decision, delivered in an accessible, understandable and non-technical way, especially for lay users.

» **Responsibility explanation**: Who is involved in the development, management and implementation of an artificial intelligence system, and who to contact for a human review of a decision.

» **Data explanation**: What data has been used in a particular decision and how; what data has been used to train and test the ML model and how.

M. E. Webb et al., "Machine learning for human learners: opportunities, issues, tensions and threats," Education Tech Research Dev, vol. 69, no. 4, pp. 2109–2130, Aug. 2021, doi: 10.1007/s11423-020-09858-2.

Types of Explanations (ctd.)

» **Safety and performance explanation**: Steps taken across the design and implementation of an ML system to maximise the accuracy, reliability, security and robustness of its decisions and behaviours.

» **Fairness explanation**: Steps taken across the design and implementation of an ML system to ensure that the decisions it supports are generally unbiased and fair, and whether or not an individual has been treated equitably.

» **Impact explanation**: The impact that the use of an artificial intelligence system and its decisions has or may have on an individual, and on wider society.

M. E. Webb et al., “Machine learning for human learners: opportunities, issues, tensions and threats,” Education Tech Research Dev, vol. 69, no. 4, pp. 2109–2130, Aug. 2021, doi: 10.1007/s11423-020-09858-2.

Interpretable Explanations Task

https://github.com/FUB-HCC/hcds-summer-2022/wiki/10_exercise

Challenges in Current Explanation Methods

» **Statistical uncertainty and inferences**: Not only the ML model, but also its explanations are statistically computed from data and are subject to uncertainty. However, many explanation methods provide explanations without quantifying the uncertainty of the explanation.

» **Causal explanations**: Most statistical learning procedures reflect correlation structures between features instead of true, inherent, causal structure of the underlying phenomena.

» **Evaluation**: The ground truth explanation is not known, and any straightforward way is not available to quantify how interpretable a model is, or how correct an explanation is

» **Feature dependence**: Feature dependence introduces problems with attribution and extrapolation. Extrapolation and correlated features can cause misleading explanations.

C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges,” vol. 1323, 2020, pp. 417–431. doi: 10.1007/978-3-030-65965-3_28.,
J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, “Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics,” Electronics, doi: 10.3390/electronics10050593.

Next Time

you will have …

1. actively participated in the lecture
2. submitted the fifth programming assignment

Have fun!