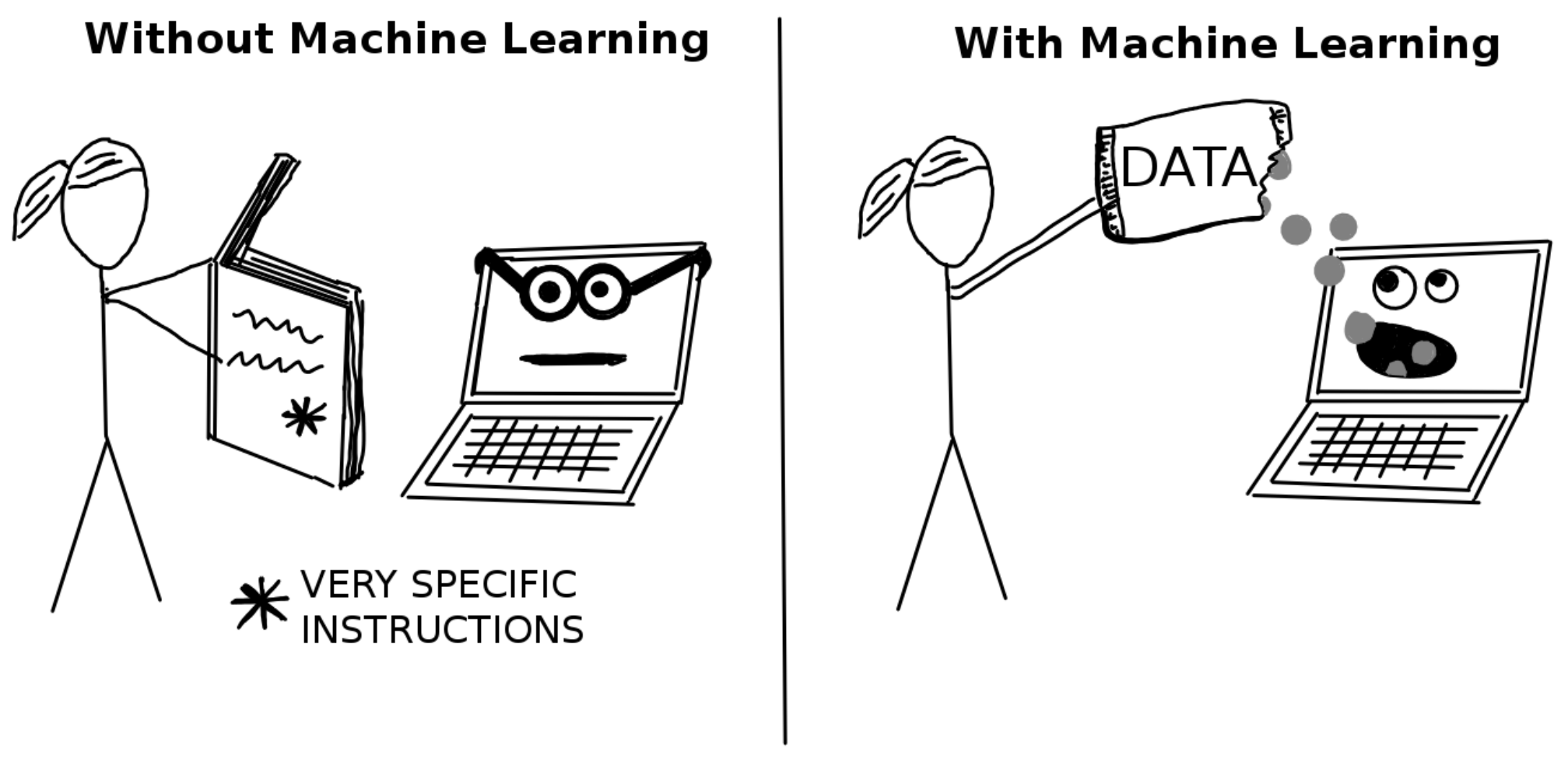
Interpretable Machine Learning

Christoph Molnar `christoph.molnar@stat.uni-muenchen.de`

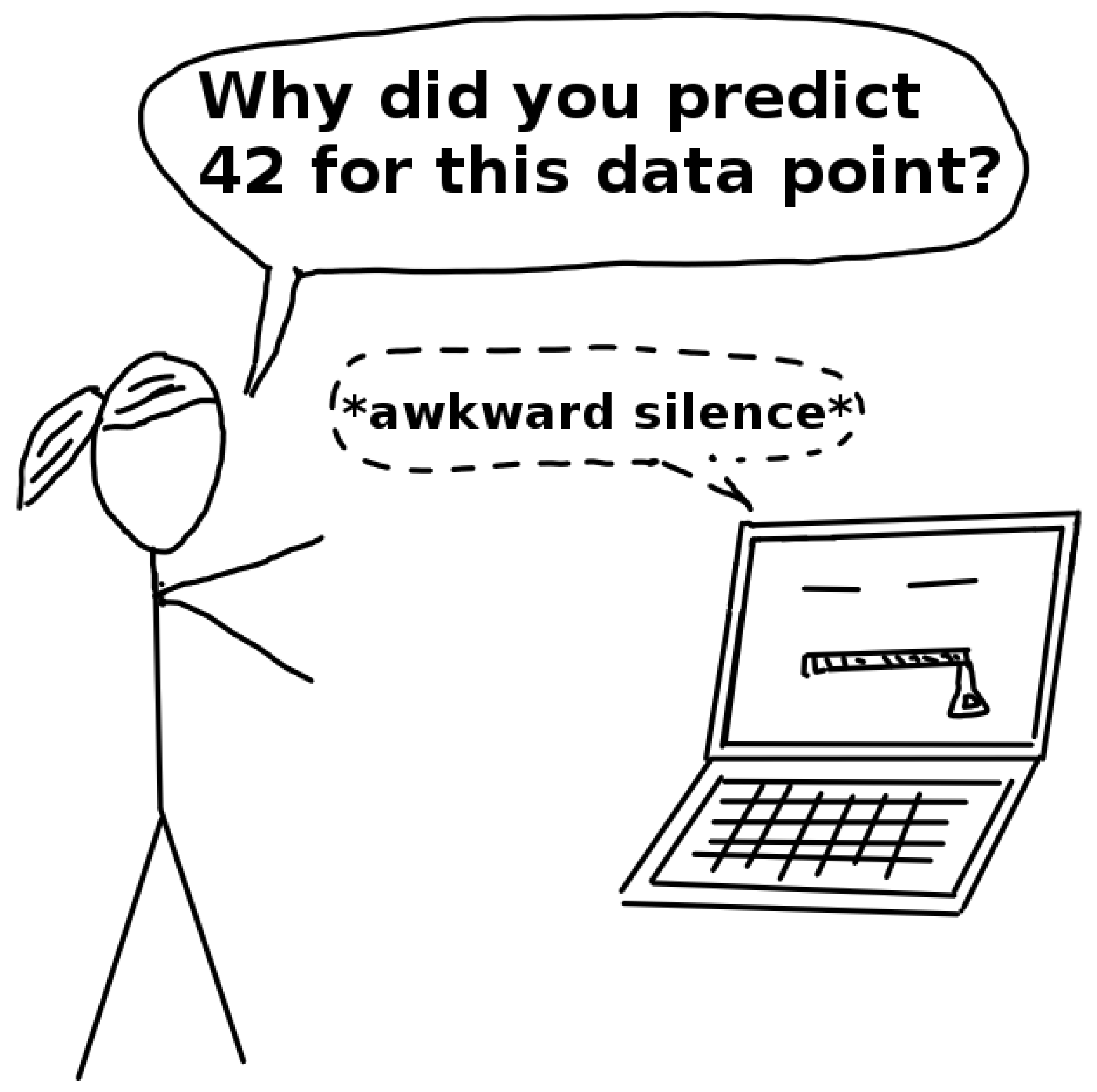
LMU Munich / Department of Statistics / Prof. Dr. Bernd Bischl

Machine Learning

Machine learning is a set of methods that allow computers to learn from data to make and improve predictions (e.g. cancer, weekly sales, credit default).



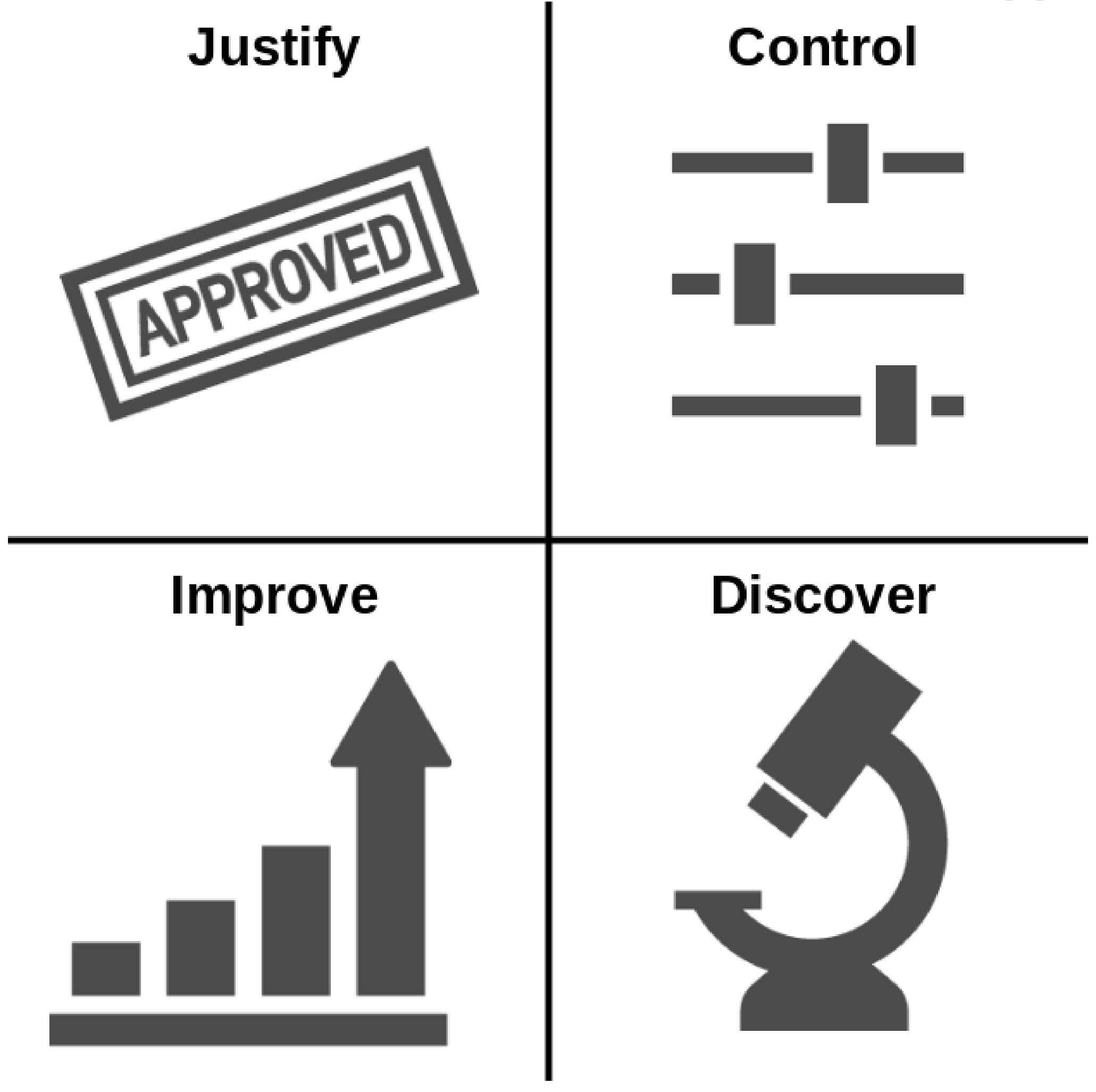
Black Box Problem



Solution: **Interpretable Machine Learning** is a set of methods and models that make it possible to explain the behaviour and predictions of machine learning systems.[4]

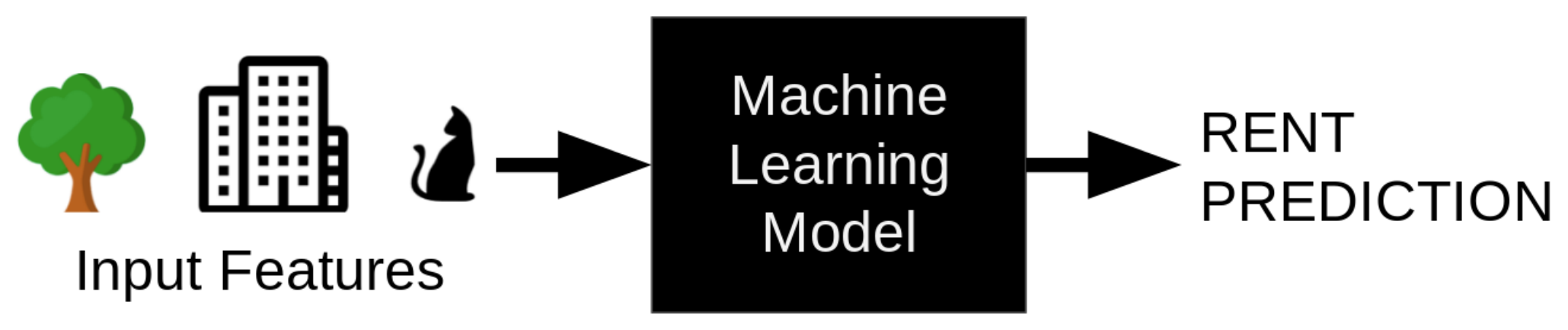
Why Explain?

We need to explain machine learning models to ... [1]:



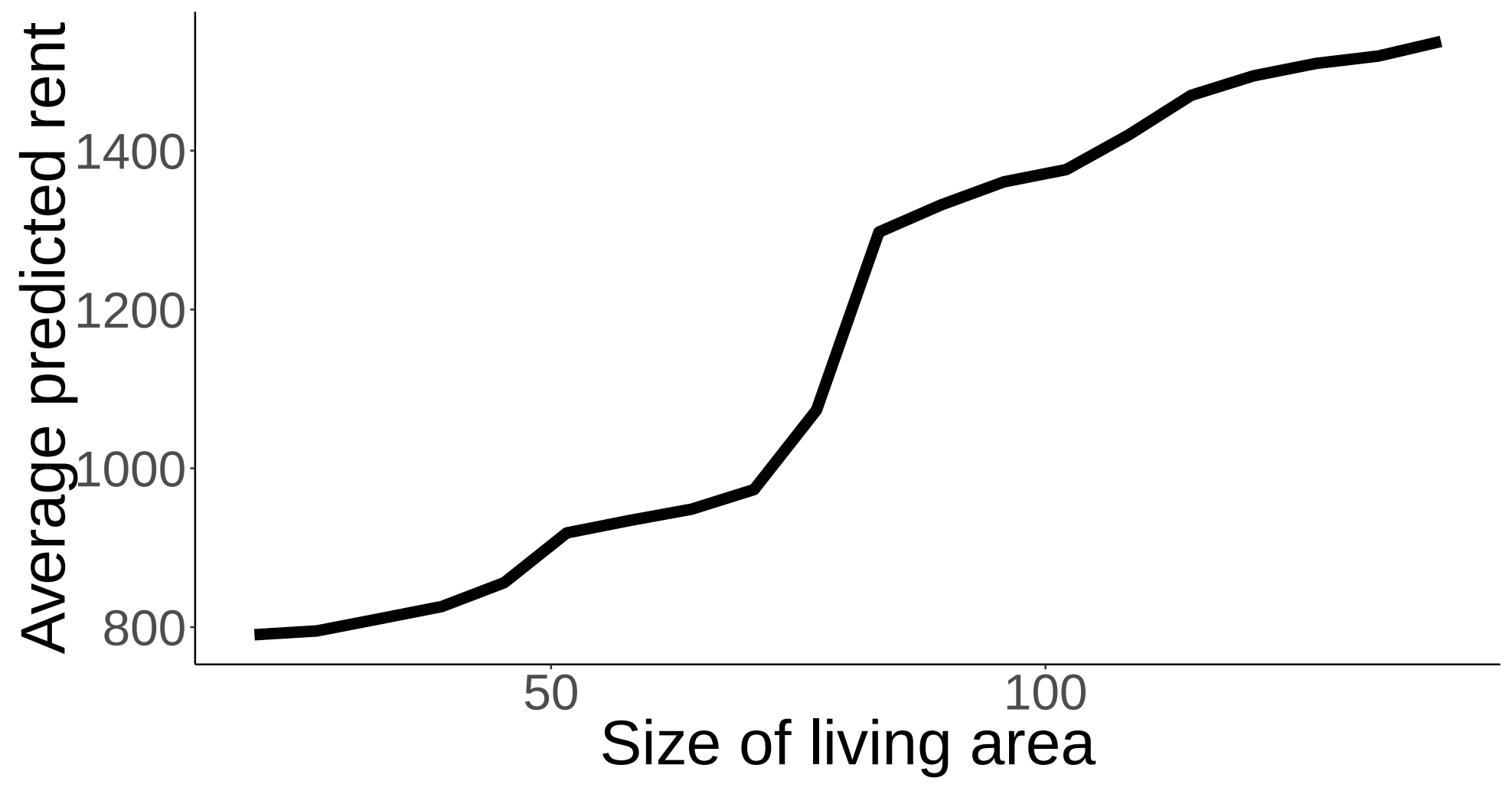
Example: Predict Rent

As an application example we train a machine learning model (here a random forest), which predicts the rent of an apartment based on the size of the living space, the location ("good" or "bad") and whether cats are allowed.



Feature Effect

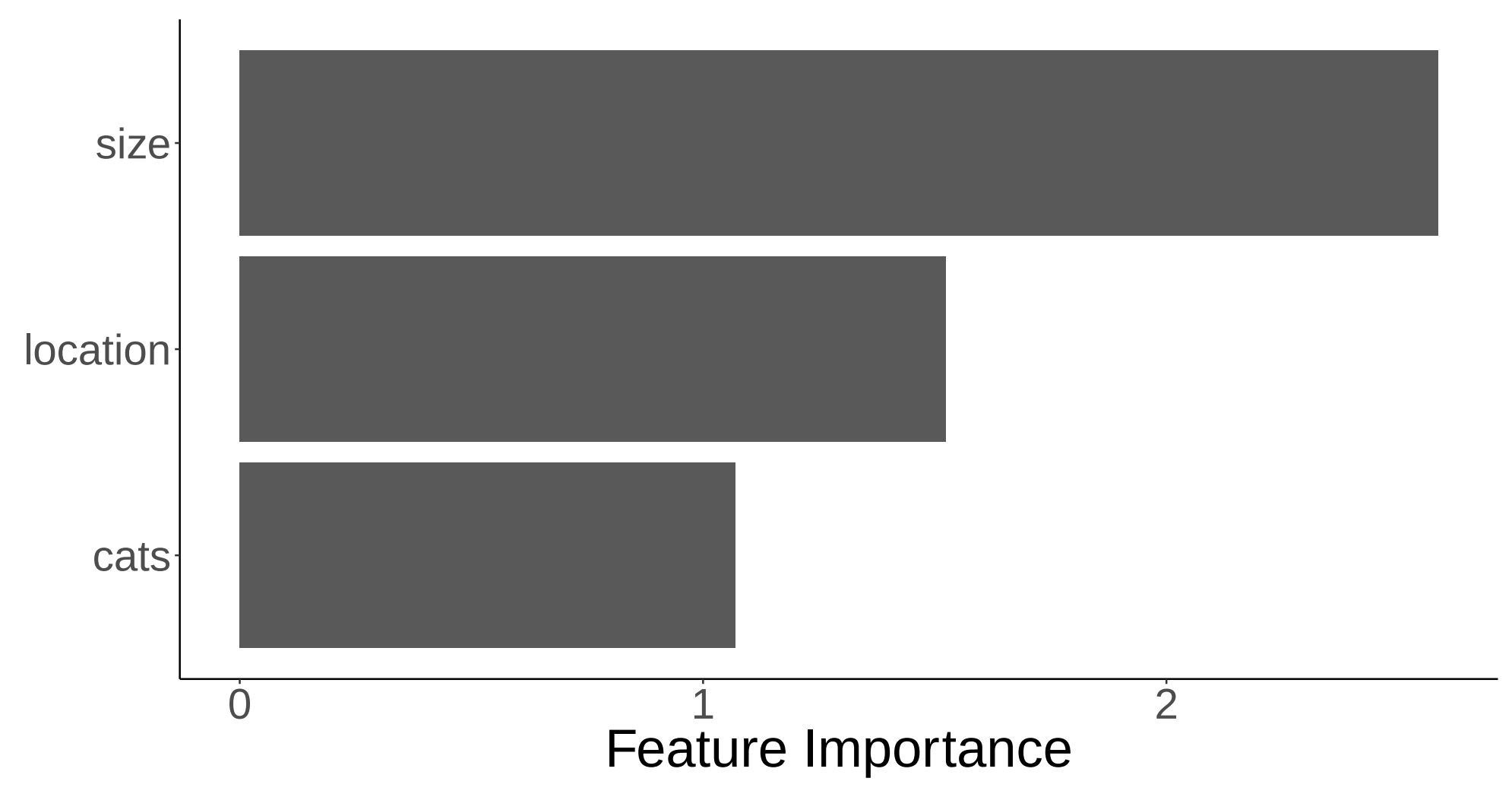
The feature effect describes how changing a feature changes on average the prediction of all data points. [3]



If we vary the size of the living area, we observe how the predicted rent increases with increasing size.

Feature Importance

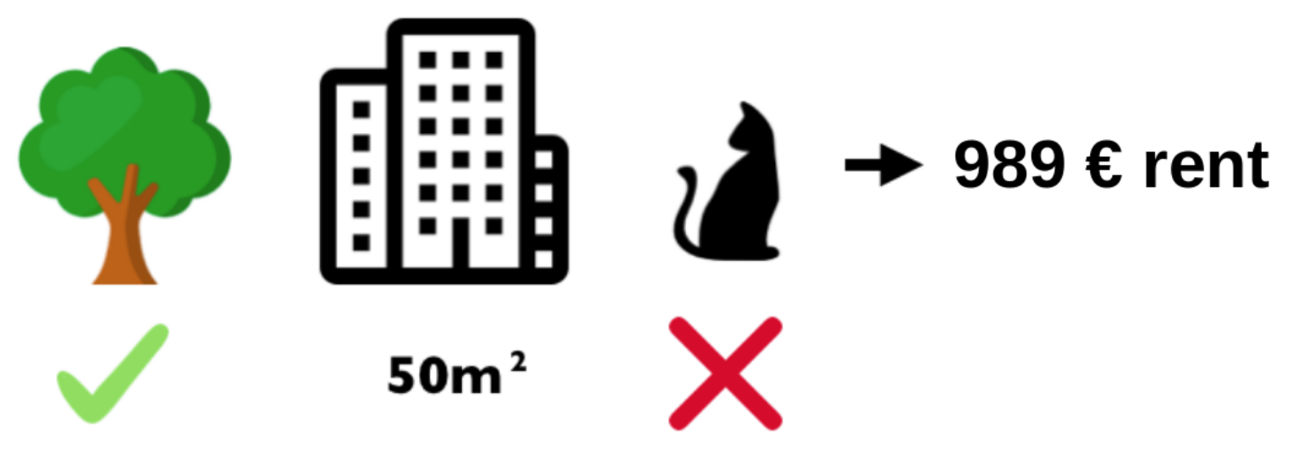
Feature importance [2] tells us how much the model error increases when we shuffle the values of a feature in the data (= ”destroying” the relationship between the feature and the outcome). The greater the increase in error, the more important the feature is.



In the rent example the most important feature was the size of the living area. Shuffling the size feature increases the model error by a factor of 2.5.

Individual Predictions

Sometimes we want to explain why a certain prediction was made by a machine learning model. *Suppose we want to explain the 989Euro rent prediction for a* $50m^2$ *apartment in a good location and where cats are forbidden.*

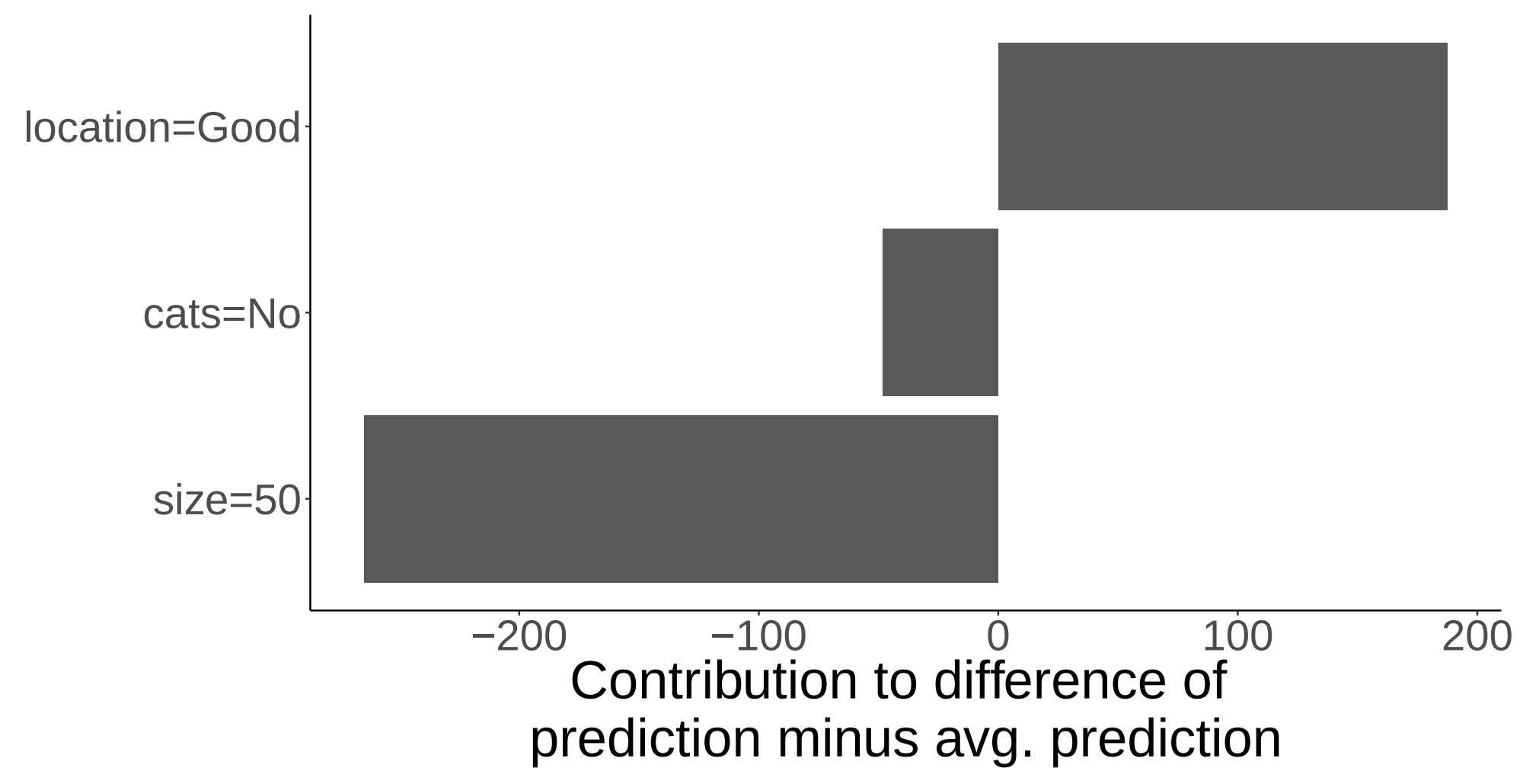


Counterfactual Explanation

A counterfactual explanation describes the smallest changes to a data points feature values (inputs) that change the prediction to a predefined output. *How do the inputs for the 989 Euro apartment have to change so that the predicted rent is over 1100 Euros? Answer (one of many possible ones): If cats were allowed, the predicted rent would be 1111 Euros.*

Shapley Values

From a game theory point of view, the feature values are players in a cooperative game who receive the value of the prediction as a payout. The Shapley Value method [6] splits the difference between the prediction and the average prediction fairly among the feature values.

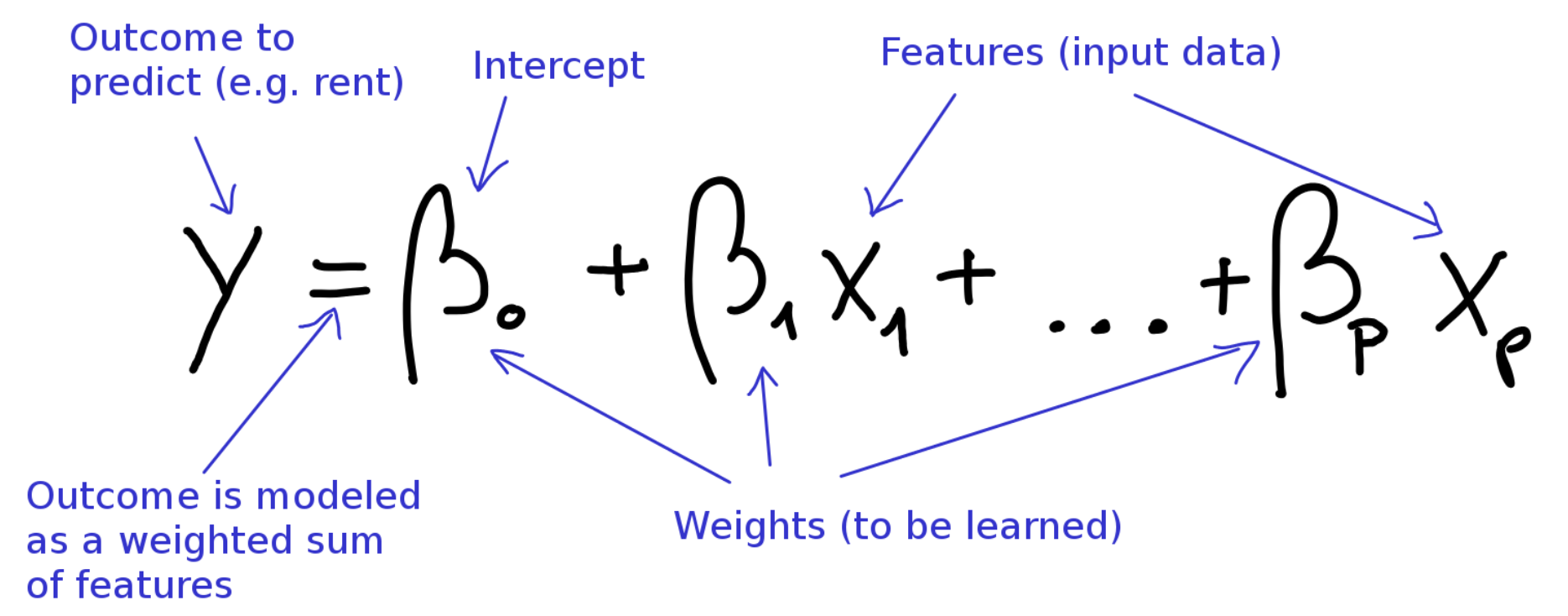


The predicted value of the apartment is 989Euros, the average prediction of all apartments is 1158. The negative difference is explained by the apartments small size and the ban on cats. The good location has a positive effect on the predicted rent.

Interpretable Models

Intrinsically interpretable models learn simple relationships between input and output.

Linear Regression Models


Outcome to predict (e.g. rent)
Intercept
Features (input data)
Y = β₀ + β₁X₁ + ... + βₚXₚ
Outcome is modeled as a weighted sum of features
Weights (to be learned)

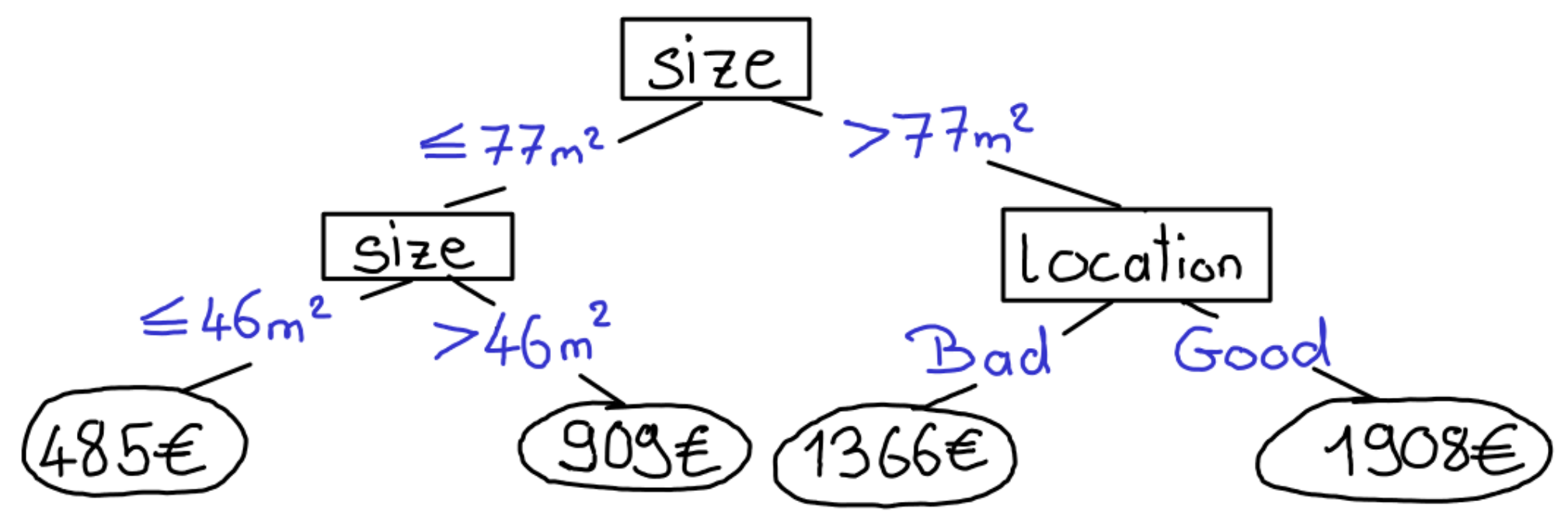
In the rent example, the following weights were learned:

$$\text{rent} = -147 + 14 \cdot \text{size} + 427 \cdot I_{\text{good location}} + 116 \cdot I_{\text{cats allowed}}$$

The predicted rent increases by 14 Euros per each m^2.

Decision Trees

Decision trees divide the data into smaller subsets based on ’decisions’ made on input features.



The tree predicts a rent of 1908 Euros for an apartment larger than 77 m^2 *in a good location.*

Decision Rules

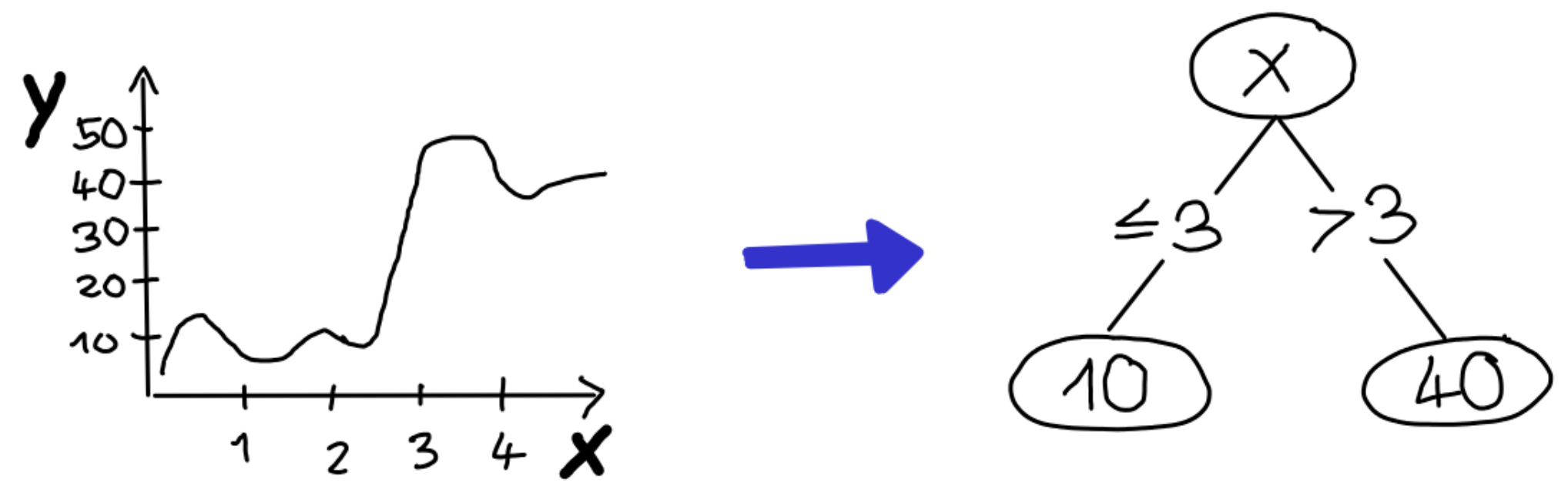
Decision rules are IF-THEN statements that consist of a condition and a prediction. One or more rules can be used to make predictions. For example:

IF $90m^2 \leq$ *size* $< 110m^2$ *AND location = ”good” THEN rent is between 1540 and 1890 EUR*

Surrogate Models

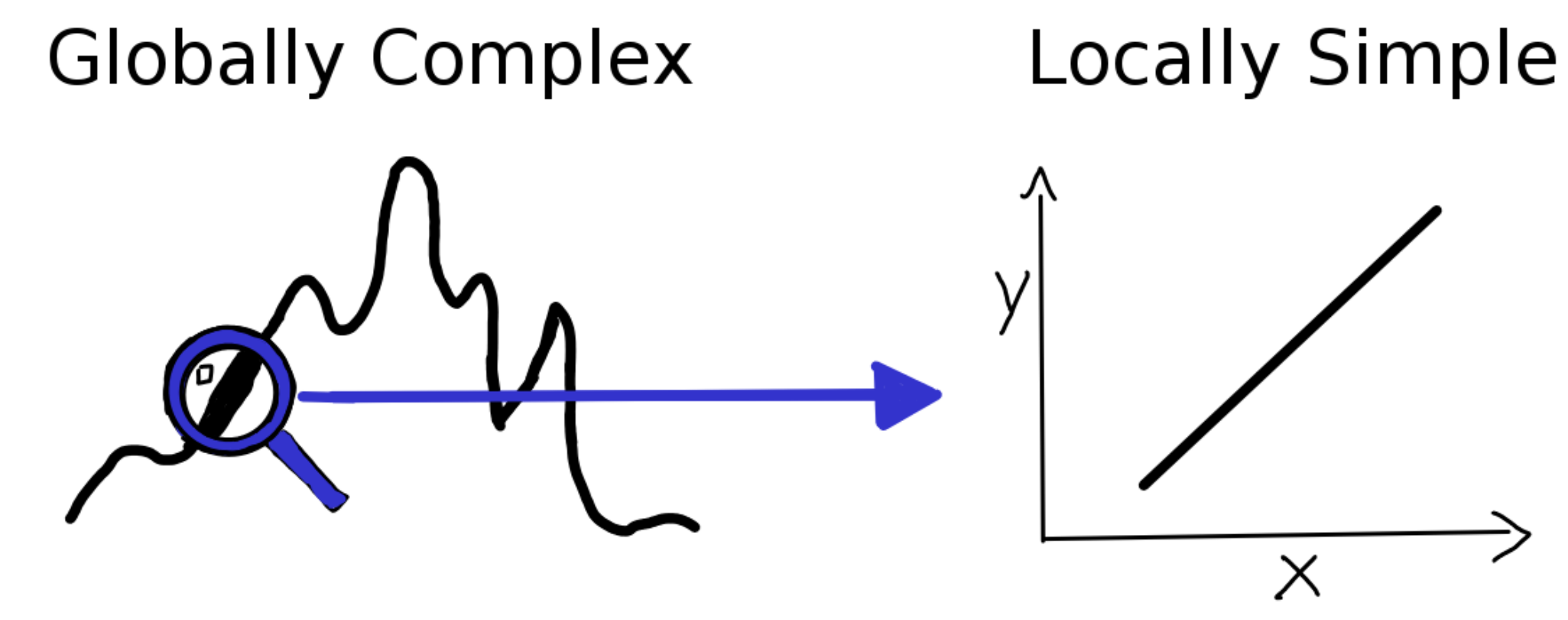
Global Surrogate Models

Intrinsically interpretable models (e.g. a tree) can be used to increase interpretability of a black box model (e.g. a neural network) by approximating its predictions by acting as a surrogate model that can be interpreted.



Local Surrogate Models[5]

To explain an individual prediction, replace a complex model with a locally weighted, interpretable model (e.g. decision tree). Data points that are close to the data point of interest get a high weight.



References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 2018.

[2] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. Model class reliance: Variable importance measures for any machine learning model class, from the ”Rashomon” perspective. *arXiv preprint arXiv:1801.01489*, 2018.

[3] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[4] Christoph Molnar. *Interpretable Machine Learning*. 2018. `https://christophm.github.io/interpretable-ml-book/`.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144. ACM, 2016.

[6] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

LaTeX TikZposter