Life Sciences and
Facility Management

Institute of
Computational Life Sciences

Explainable AI

Machine Learning

Norman Juchler

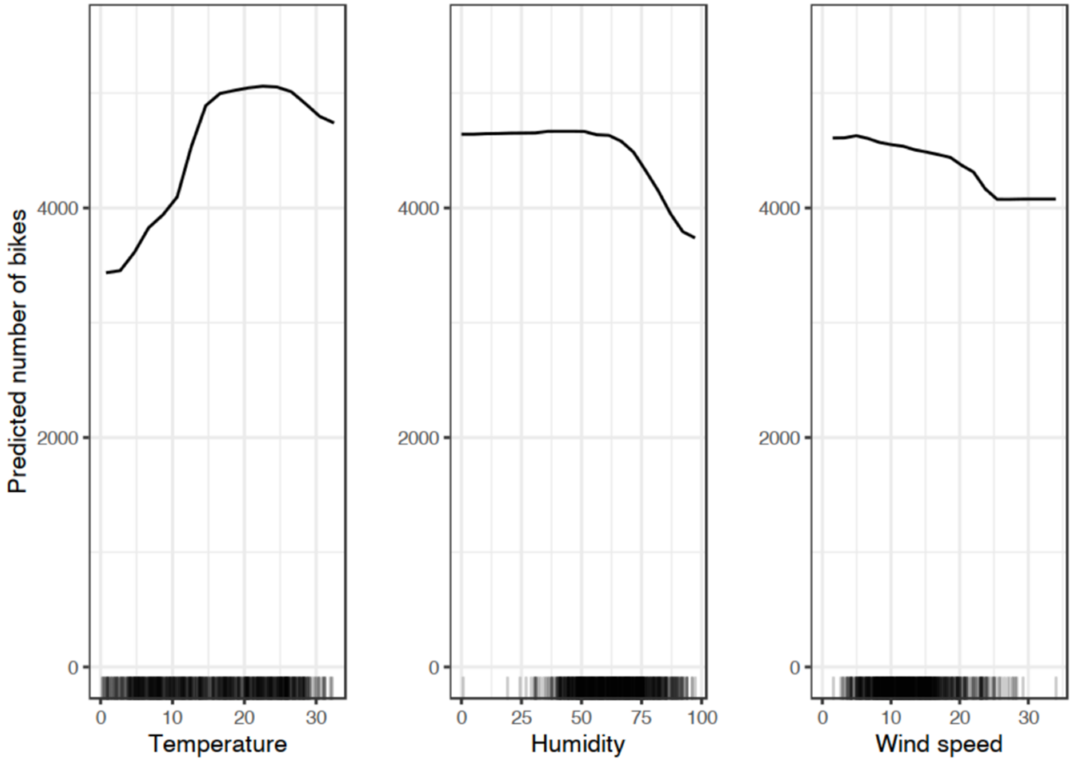
Motivation

- In many real-world applications of machine learning, it is often more important to understand the reasons for the predictions than to achieve the highest accuracy.

Data understanding through exploration

- Example: Marginal distributions examples from bike rental data

Question: What can you infer from the plots?



Black and white boxes

- White boxes:
  - Interpretable models where the internal workings (equations, parameters) are transparent and easily understood
  - Examples: linear regression or decision trees
- Black boxes:
  - Are typically complex and opaque, making it difficult to directly interpret how the methods arrive at predictions.
  - Examples: Neural network, ensemble methods.

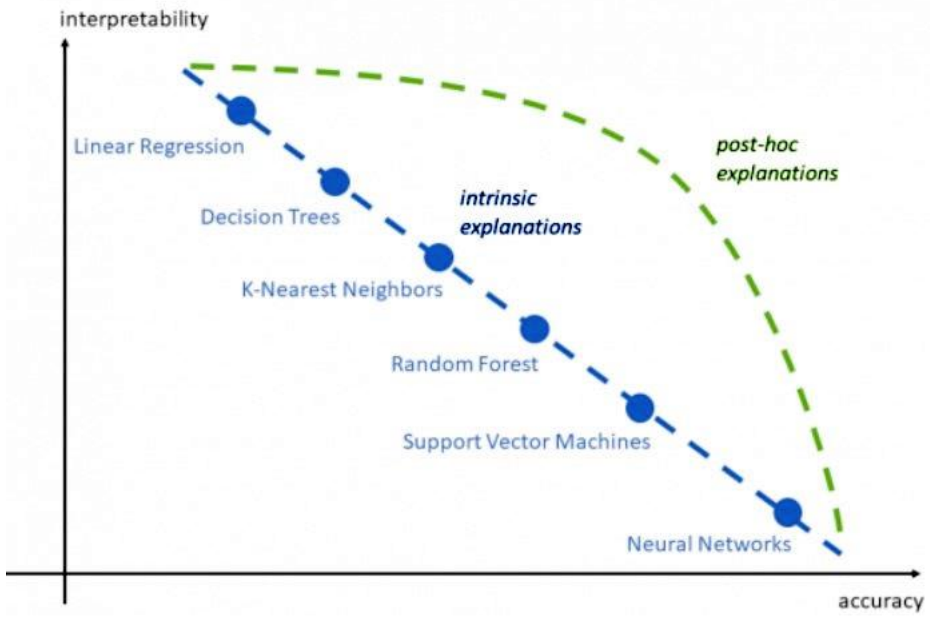
- Observation: Many standard ML models behave like black boxes – it's often difficult to interpret the relationship between in- and output.

Interpretability: Relevance

- **Definition**: Model interpretability refers to the ability to understand and explain how a machine learning model makes its predictions or decisions.
- Knowing *why* our model made a certain prediction allows us to learn more about the problem, the data and the reasons why the model made a wrong prediction: Interpretability enables us to **extract the knowledge** learned from the model.
- Some models run in high-risk environments that require **safety** and prediction **guarantees** (e.g., specificity > 0.99 and sensitivity > 0.95)
- Models can be **biased** without us being aware of it. Model interpretability can help to reveal unfair or unintended behavior.
- Interpretability increases **human/social acceptance** of machine learning models
- Interpretability facilitates **debugging** and auditing.

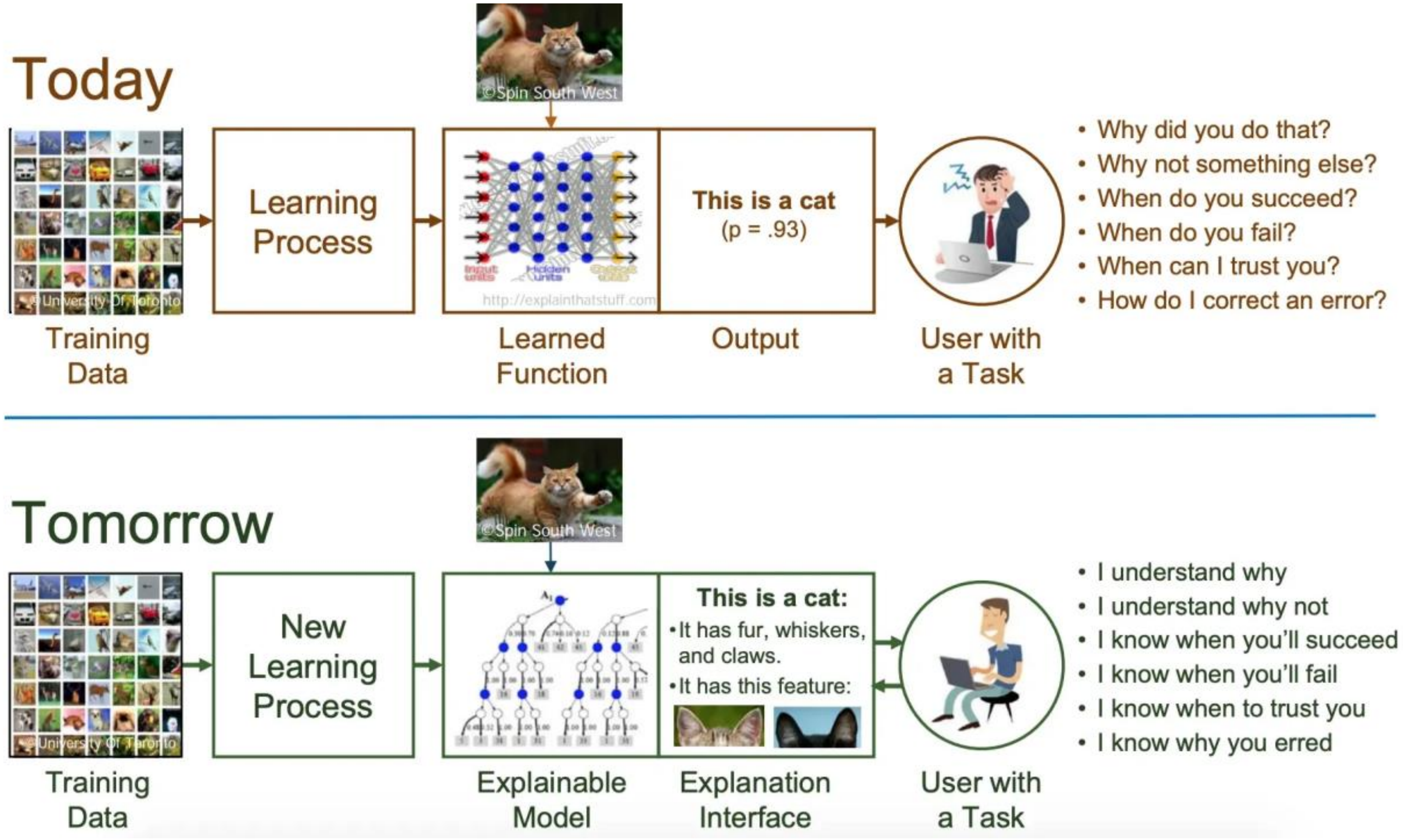
Types of interpretable models

- **Intrinsic explanation**: simple models
  - Refers to the interpretability of ML models that are inherently transparent and understandable by design
  - E.g., linear regression, decision trees, kNN
- **Post hoc explanation**: additional calculations
  - Refers to interpreting or explaining the predictions of a machine learning model *after* it has been trained.
  - E.g., feature importance, SHAP or LIME



- **Global explanation**: Explains model behavior
  - E.g., feature importance
- **Local explanation**: Explains individual prediction
  - E.g., LIME, SHAP

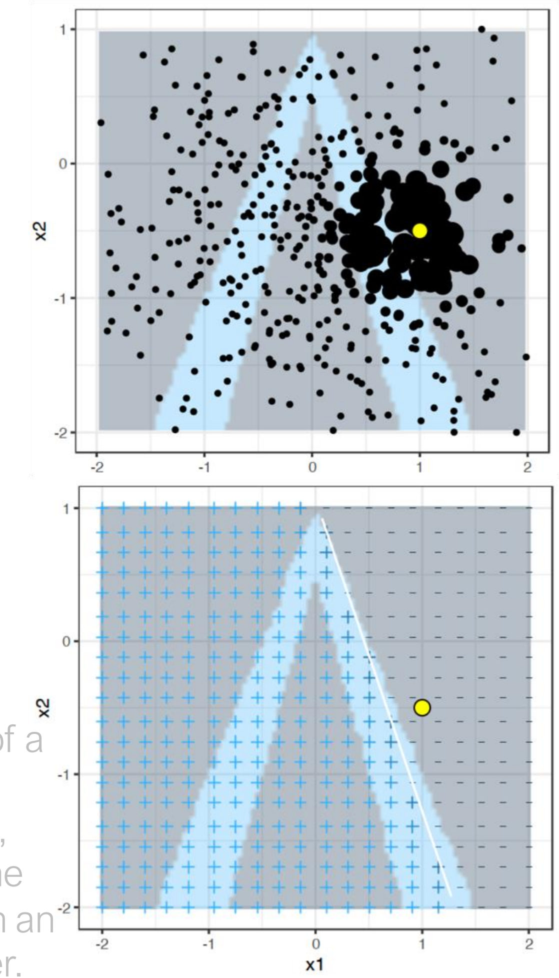
The vision of explainable ML



Source: Could not identify the original source

LIME: Local interpretable model-agnostic explanations

- LIME explains a prediction by replacing the complex (black-box) model with a locally interpretable substitute model.
- Algorithm:
  - Select a sample of interest for which you want to explain the prediction.
  - Randomly sample new data points and get their predictions from the black-box model
  - Weight new samples according to their proximity to the sample of interest.
  - Train a weighted, interpretable model on the dataset with the variations.
  - Explain the prediction by interpreting the local model.



Top: Decision regions (gray and light blue areas) of a black box model for binary classification. We can apply LIME to the sample of interest (yellow). First, we sample new points (or perturbed versions of the original data). Bottom: For these samples, we train an interpretable model. For example: a linear classifier. This local model can be used to interpret the prediction for the sample of interest.

- Sources:
  - Section on LIME in Molnar, 2024
  - Python package implementing LIME
  - Original paper, Ribeiro et al., 2016

Shapley Additive exPlanations (SHAP)

- **Shapley values** aim to fairly assign a prediction to individual features.
- The Shapley value is the average marginal contribution of a feature value across all possible feature sets.
- We can approximate it by Monte Carlo sampling:

$$\hat{\phi}_j = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{f}(x^m_{+j}) - \hat{f}(x^m_{-j})\right)$$

x: Sample whose prediction we want to explain

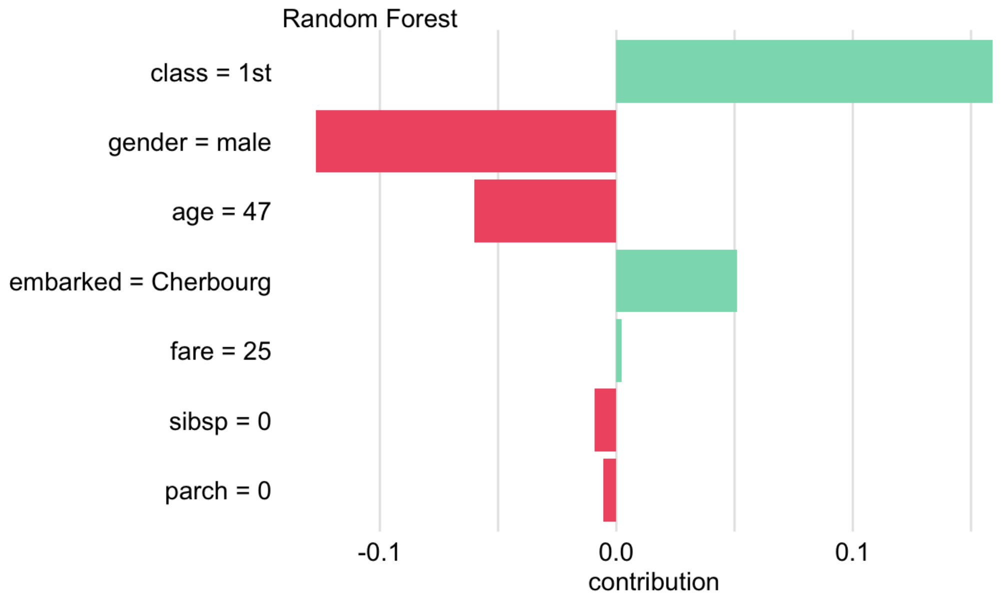
same as x^m_{+j}, but without **feature j**

x with a random number of feature values replaced by feature values from a random data point z

- The concept of Shapley values is borrowed from **game theory**:
  - Idea: Shapley values fairly distribute the "payout" (here: the prediction) among the "players" (here: input features) based on their contribution to the outcome.
  - The Shapley value is the only attribution method that satisfies the properties *Efficiency*, *Symmetry*, *Dummy* and *Additivity*, which together can be considered a definition of a fair payout (here: the prediction).

Shapley Additive exPlanations (SHAP)

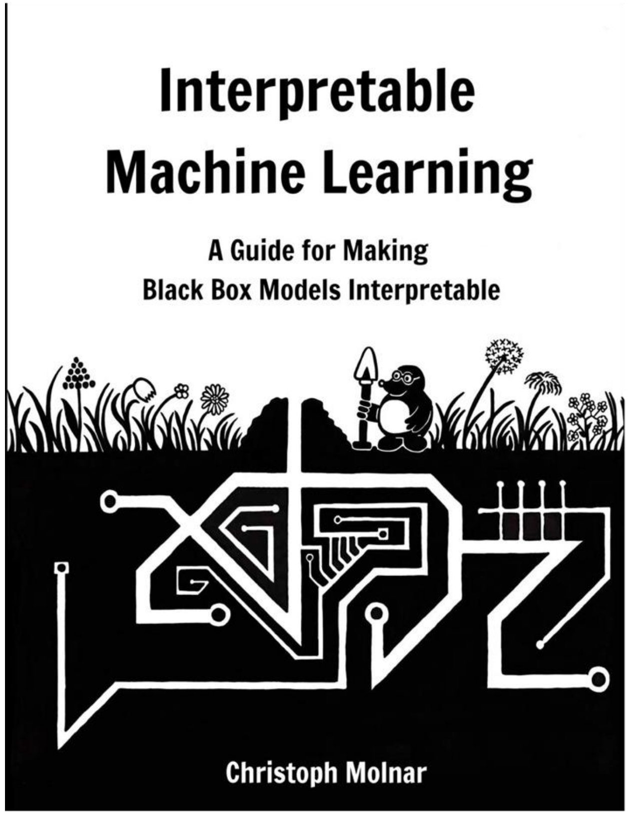
- Properties of SHAP:
  - Local and global interpretability: Explains individual predictions and summarizes feature importance globally.
  - Model-agnostic: Works with any machine learning model (e.g., tree-based, neural networks, SVMs).
  - Fairness: Ensures consistent and fair attribution of feature contributions.

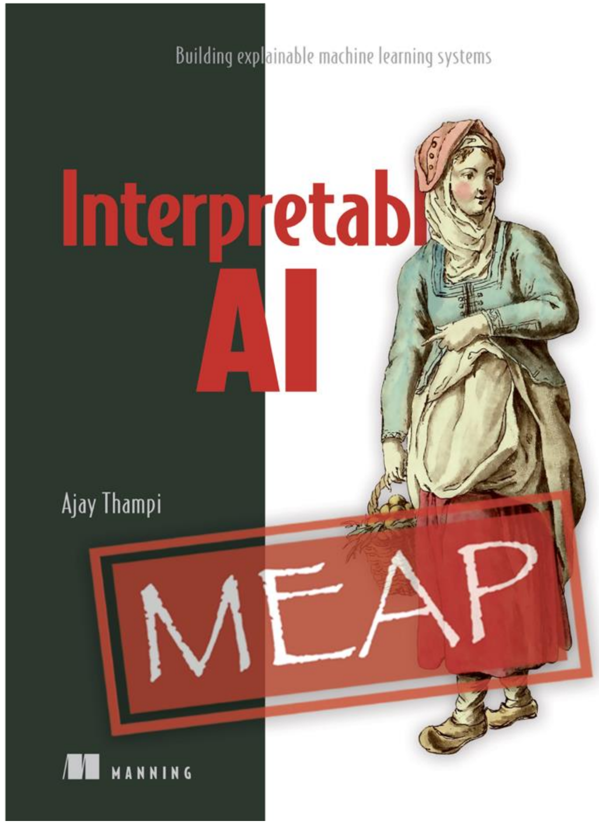


A plot of Shapley values for a model of the Titanic dataset, evaluated for a specific passenger. Source: Link

- Further reading:
  - SHAP: Python package
  - Introduction to explainable AI with SHAP: Link
  - About Shapley values and the SHAP method
  - Another SHAP tutorial (with examples in R): Link

References





- More at: https://christophm.github.io/interpretable-ml-book/index.html

Summary

- Local, post-hoc approaches promise to deliver explanations for individual predictions of high-performance models.
- We have learned about two popular approaches of this kind: LIME and SHAP.