Challenges to interpretability

Challenges to interpretabiliy

Understanding what the model has learned and how they produce their predictions is challenging, even for intrinsically explainable models.

Challenges to interpretabiliy

Challenges related to:

◒ **Input data**

Correlation

Bias

◒ **Model**

Complexity

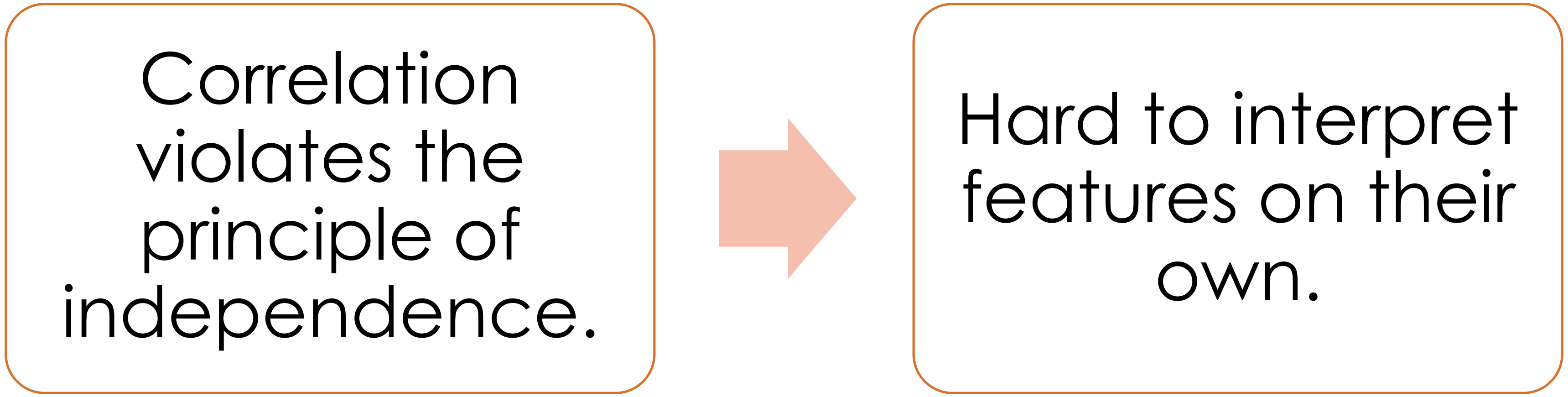
Inscrutability

Performance

● **Us**

Bias

Correlation



Correlation violates the principle of independence.

Hard to interpret features on their own.

Correlation

Coefficients in linear models and importance in decision trees are affected by correlation.

If we remove / perturb a feature, a correlated feature takes its place in the model, masking the effect of the removal / perturbation.

Permutation importance

Feature elimination methods

Biased data

When the data used to train a model does not represent the population that will be scrutinized by the model:

➔ the interpretations are not useful.

Coded bias ➔ algorithm for facial recognition trained on mostly white men used also (also) to identify people of colour.

Confirmation bias

When we think a feature is important, we may be susceptible to selecting the model / post-hoc method that shows that feature as important.

Bad model performance

Stating the obvious here:

If a model does not fit the data well, the interpretations that we can get from it, are meaningless.

Black box models

Models that are inscrutable by design ➔ Hard to know what is really driving the predictions.

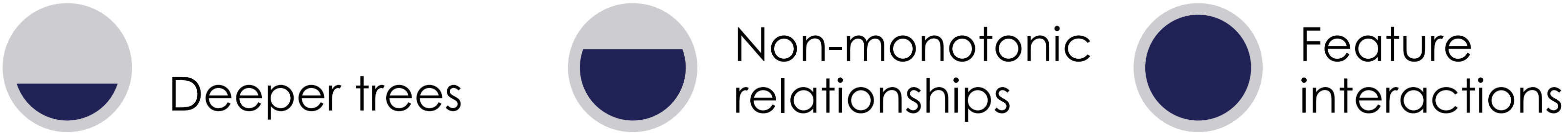
Explaining black boxes (with post-hoc methods) may lead to wrong or misleading interpretations, because there is no way to confirm that a post-hoc method truly reflects the inner working of the black box.

White box models limitations

The main challenge is to create / use models that are simple enough to understand, yet complex enough to properly fit the data.

White box models limitations

- Models trained using too many features are hard to interpret.
- Complex models, that is, model with too many parameters are harder to explain.

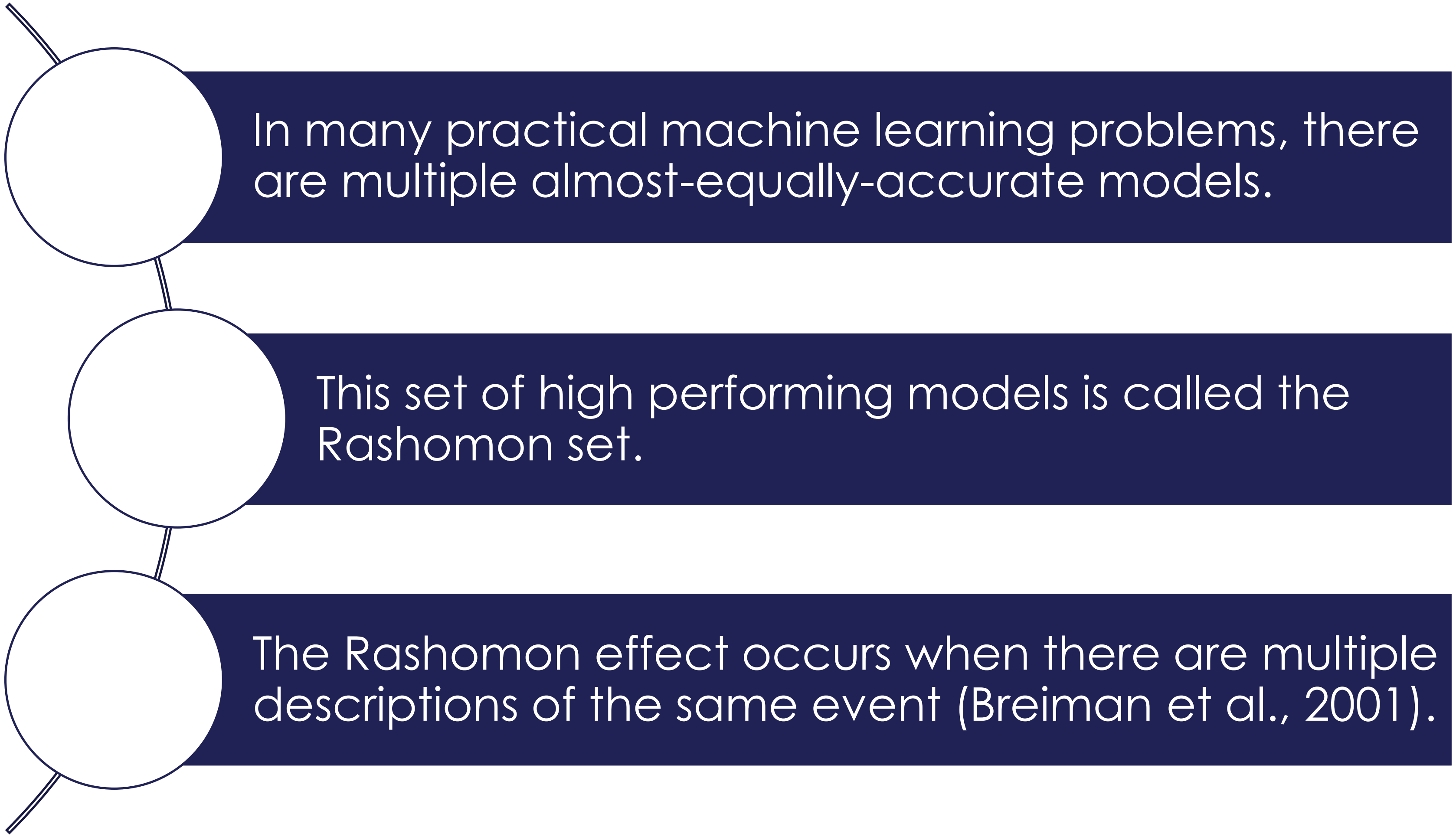


Deeper trees

Non-monotonic relationships

Feature interactions

Rashomon sets



In many practical machine learning problems, there are multiple almost-equally-accurate models.

This set of high performing models is called the Rashomon set.

The Rashomon effect occurs when there are multiple descriptions of the same event (Breiman et al., 2001).

Train In Data

THANK YOU

www.trainindata.com