Surrogates

Surrogate models for explainability

A **surrogate** model is an intrinsically explainable machine learning model trained to predict the predictions of a black box model.

Surrogate models for explainability

A **surrogate** model is an intrinsically explainable machine learning model trained to predict the predictions of a black box model.

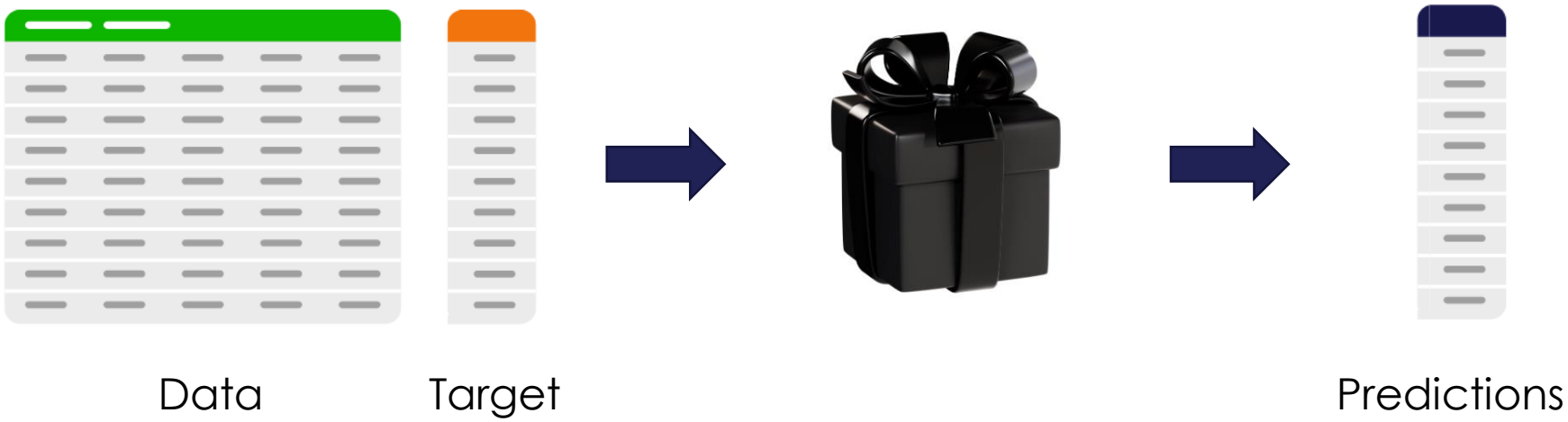
By interpreting the white box model, we try to gain insights into how the black box makes the predictions.

Surrogate models for explainability

The explanations are obtained by examining the components of the surrogate / intrinsically explainable model.

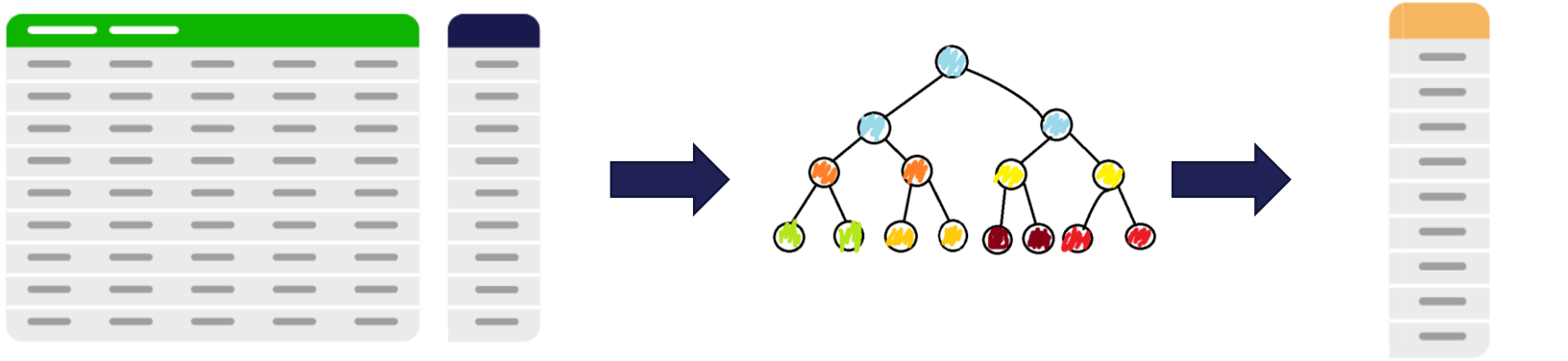
(First sections of the course).

Surrogate mechanism



First, we train a black box to predict some data.

Surrogate mechanism



Next, we train an intrinsically explainable model to **predict the predictions** of the black box.

Characteristics of an interpretability method

1) The explanations need to be easy to understand (for surrogates ✓)

2) The explanations need to correctly explain the original model (for surrogates ?)

Surrogate fidelity

For valid explanations, the surrogate must approximate the predictions of the black box very well.

We evaluate the surrogate's performance as we evaluate any machine learning model's performance.

Train In Data

THANK YOU

www.trainindata.com