LIME
Motivation

Surrogates for explainability

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

Limitations of surrogates

- Black box models have complex separation boundaries.
- White box models have simpler separation boundaries.

- Hard for a white box to approximate well a black box throughout the entire data space.

Train a surrogate locally!

The idea is that the separation boundaries in the proximity of a data point are less complex than those for the entire dataset.

LIME

LIME = **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations.

LIME

LIME trains a surrogate model in the proximity of the data point we want to explain, to model (and hence interpret) the prediction of the black box.

To explain multiple data points, we train multiple surrogates.

Train In Data

THANK YOU

www.trainindata.com