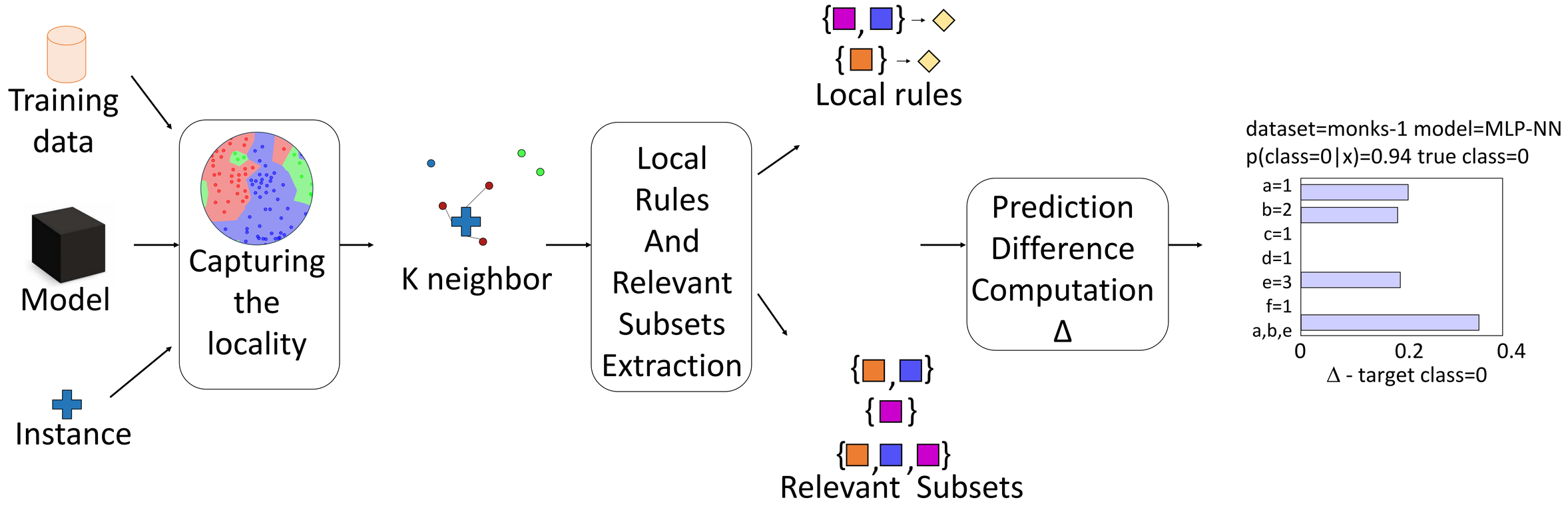
Enhancing Interpretability of Black Box Models by means of Local Rules

Eliana Pastor and Elena Baralis
Politecnico di Torino, Italy

eliana.pastor@polito.it

Paper
Scan me!

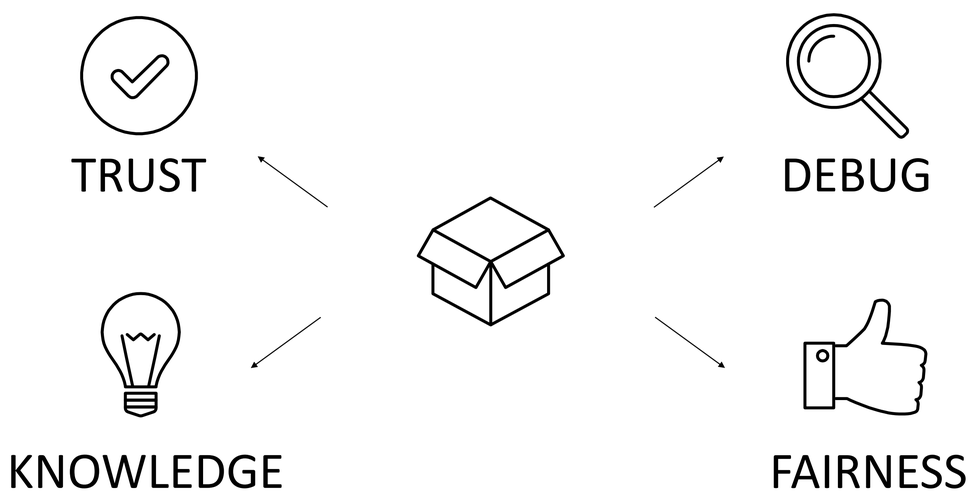
LACE explanation method



Enhancing interpretability

Black box models → hide from users the reasons behind predictions

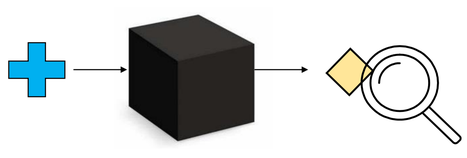
Why explainable AI?



Demand from institutions ⟶ GDPR

- *"meaningful information about the logic involved"*
- *"to ensure fair and transparent processing"*

LACE Explanations



- Qualitative insight ⟶ **local rules**

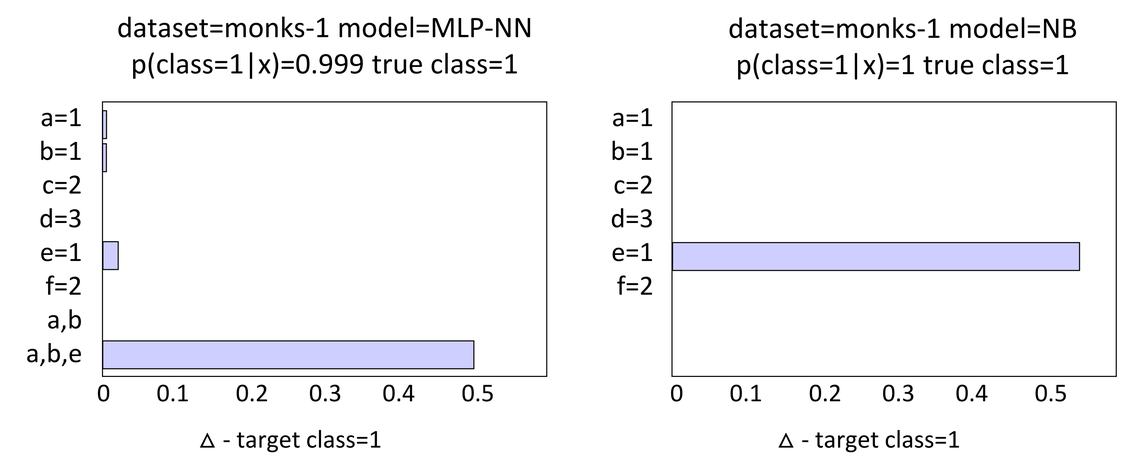
Rules learned in the locality of the prediction

- Quantitative insight ⟶ **prediction difference**

Prediction change when one or more attribute values are omitted.

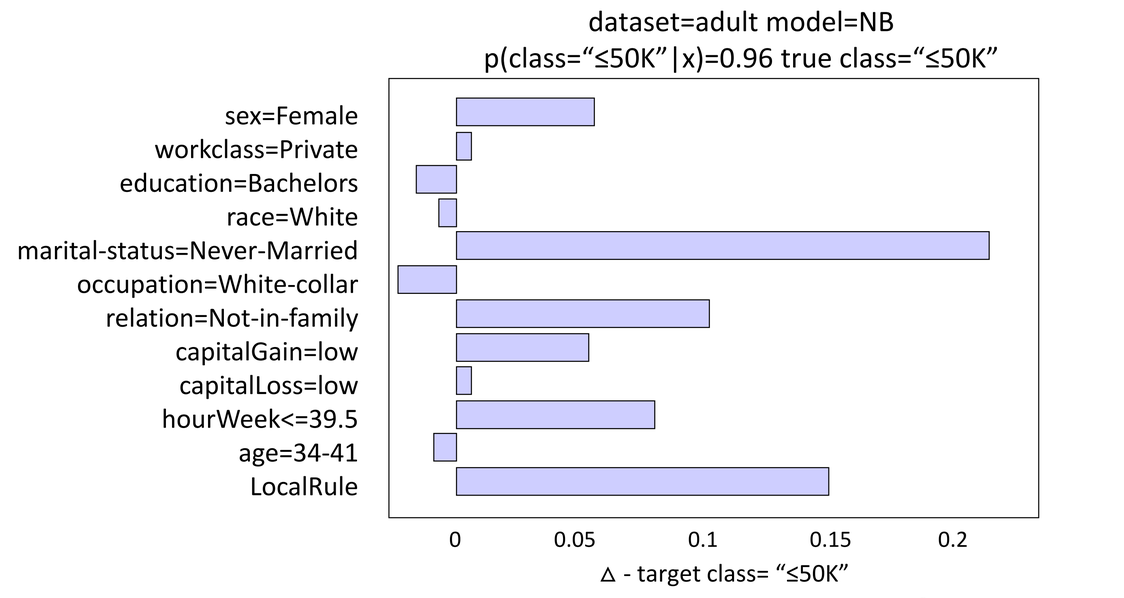
LACE Explanation Results

- Explanation comparison



Local rules:
{e=1} → class= 1
{a=1, b=1} → class=1

- Insight of reasons behind predictions



Local Rule: {sex=Female, workclass=Private, capital-gain=low} → class= "≤50K"

Future work

- User studies ⟶ to assess LACE ability to provide actionable insights on model behavior
- Explanation of Big Data model predictions

Eliana Pastor and Elena Baralis, Explaining black box models by means of local rules. ACM SAC 2019