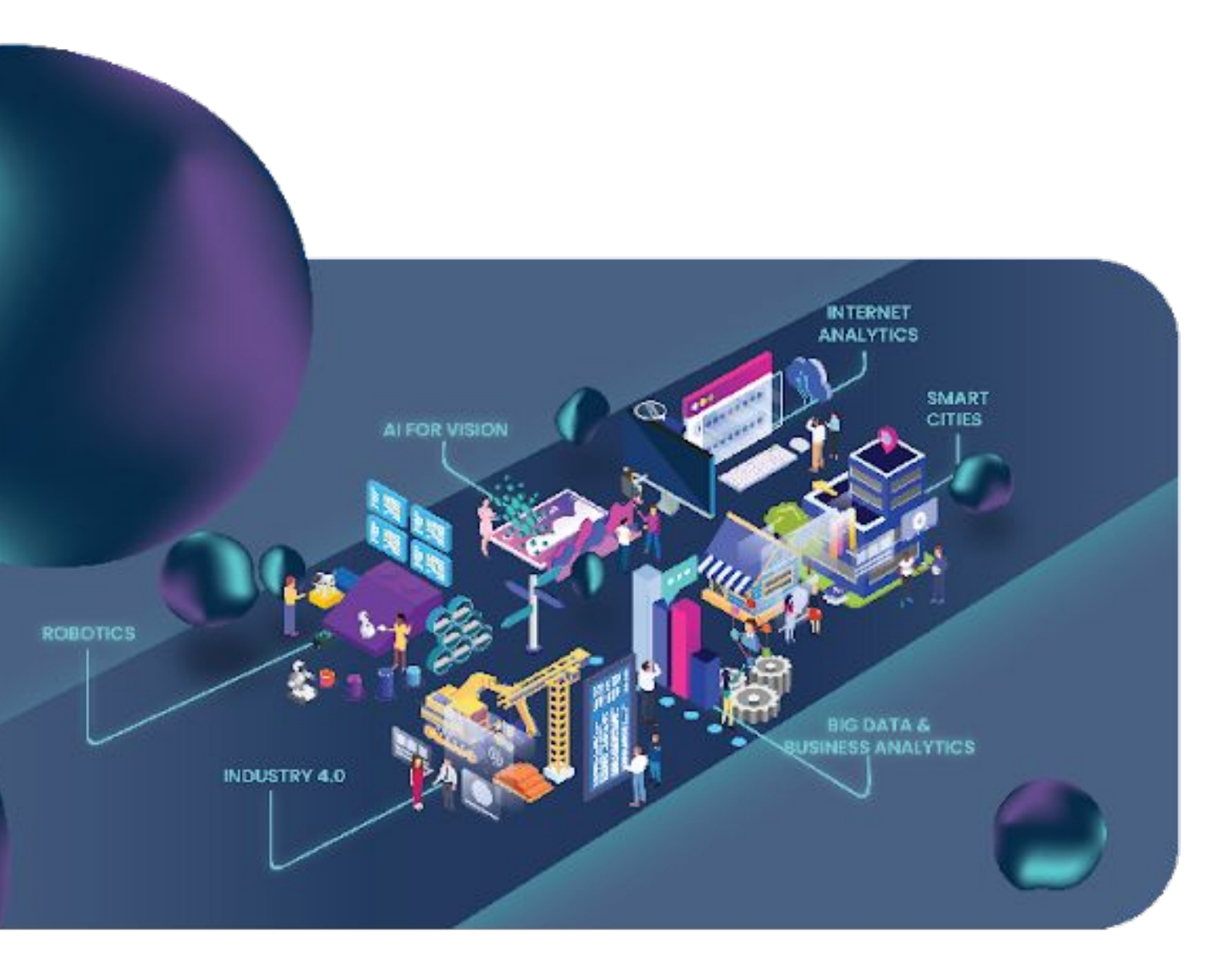


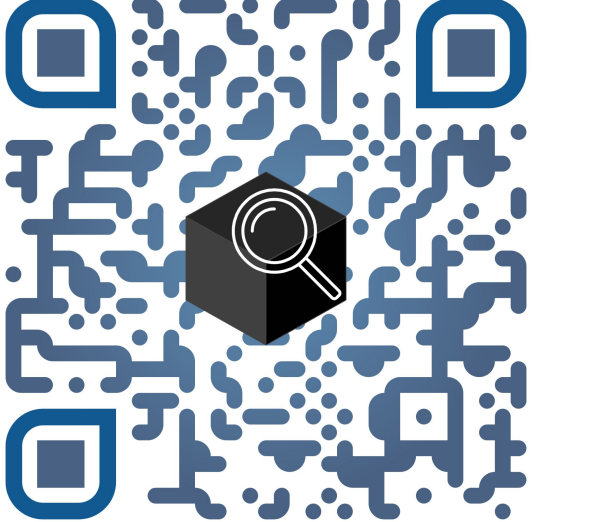
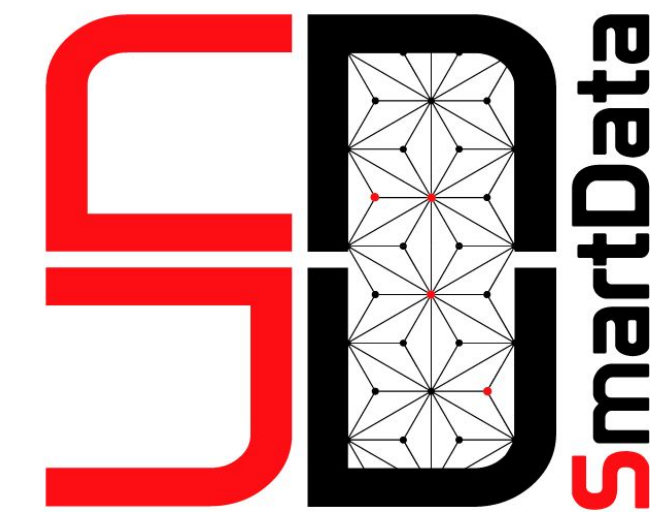
DATI, AI E ROBOTICA @POLITO

RICERCA, TRASFERIMENTO TECNOLOGICO E SUPPORTO ALLE AZIENDE SUI TEMI FONDAMENTALI DEI BIG DATA, INTELLIGENZA ARTIFICIALE, ROBOTICA E RIVOLUZIONE DIGITALE

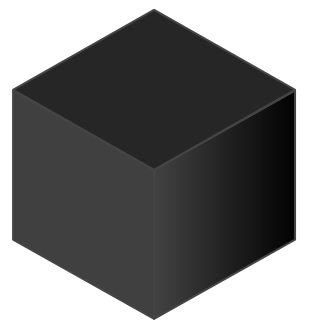


EXPLAINABLE AI AND FAIRNESS

Eliana Pastor, Elena Baralis, Luca de Alfaro
POLITO POLITO UCSC



ON THE NEED OF XAI

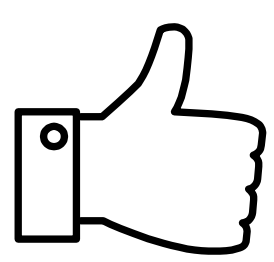


Black box models → hide from users the reasons behind predictions

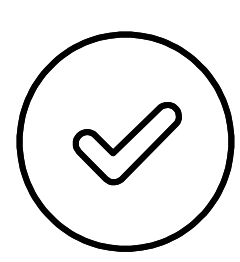
Why Explainable AI?



FAIRNESS



TRUST



MODEL
VALIDATION



DEBUG &
ERROR ANALYSIS

EXPLAINING MODEL BEHAVIOR

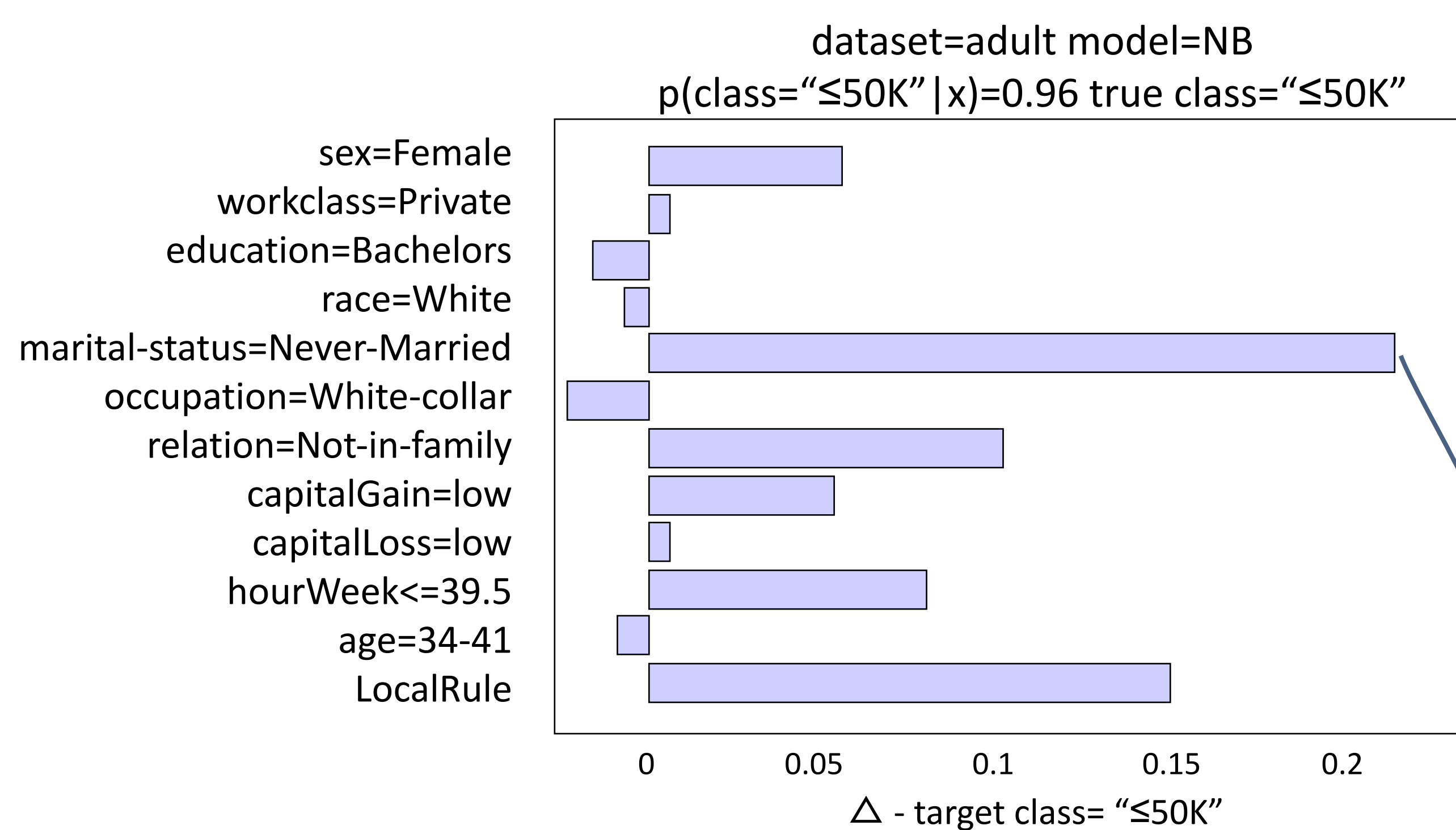
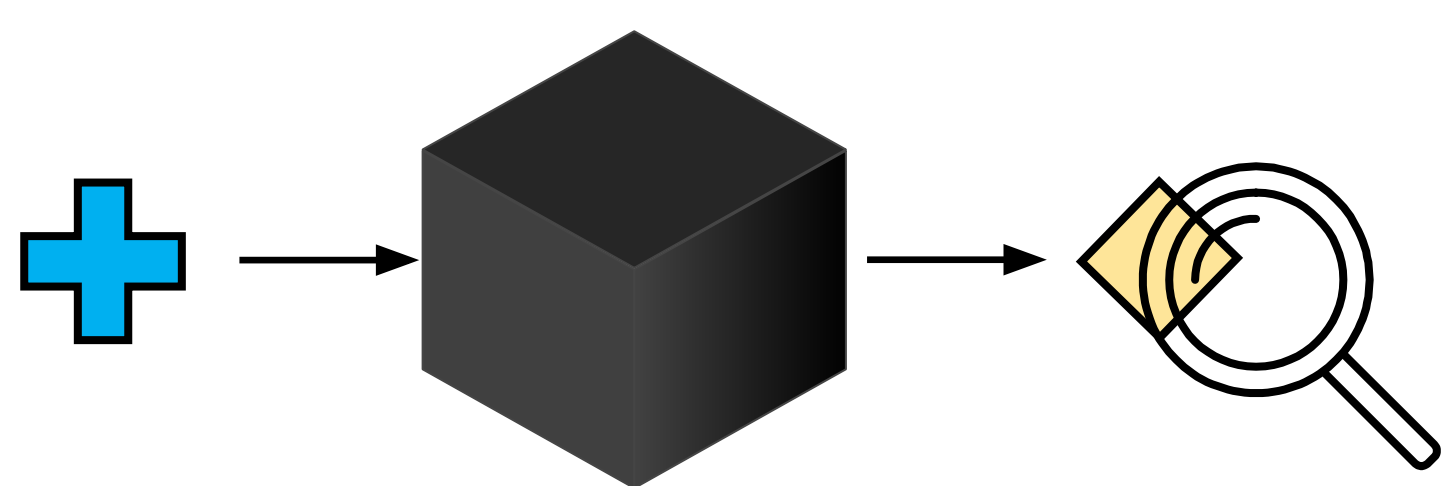
Explaining classifier behavior

- LACE - **Individual** prediction
- DivExplorer - Data **subgroup**

Model agnostic → applicable for a generic classifier

LACE - EXPLAINING INDIVIDUAL PREDICTIONS

Given a single prediction, we explain the reasons behind it



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

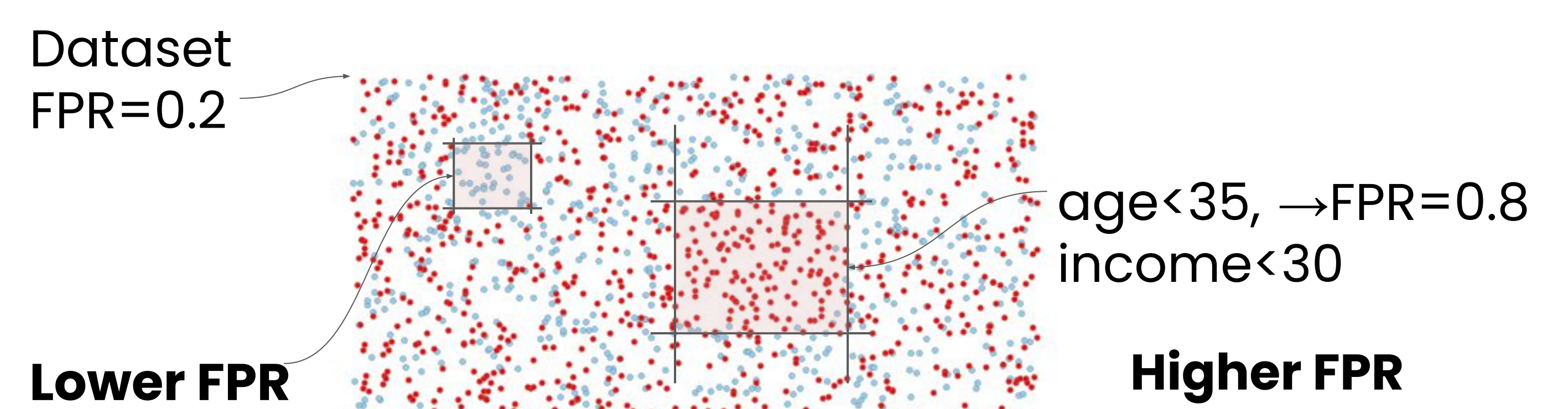
- Quantitative insight → prediction difference
- Qualitative insight → local rules

It can reveal if model predictions are based on **protected** attributes

DIVEXPLORER - EXPLAINING SUBGROUPS

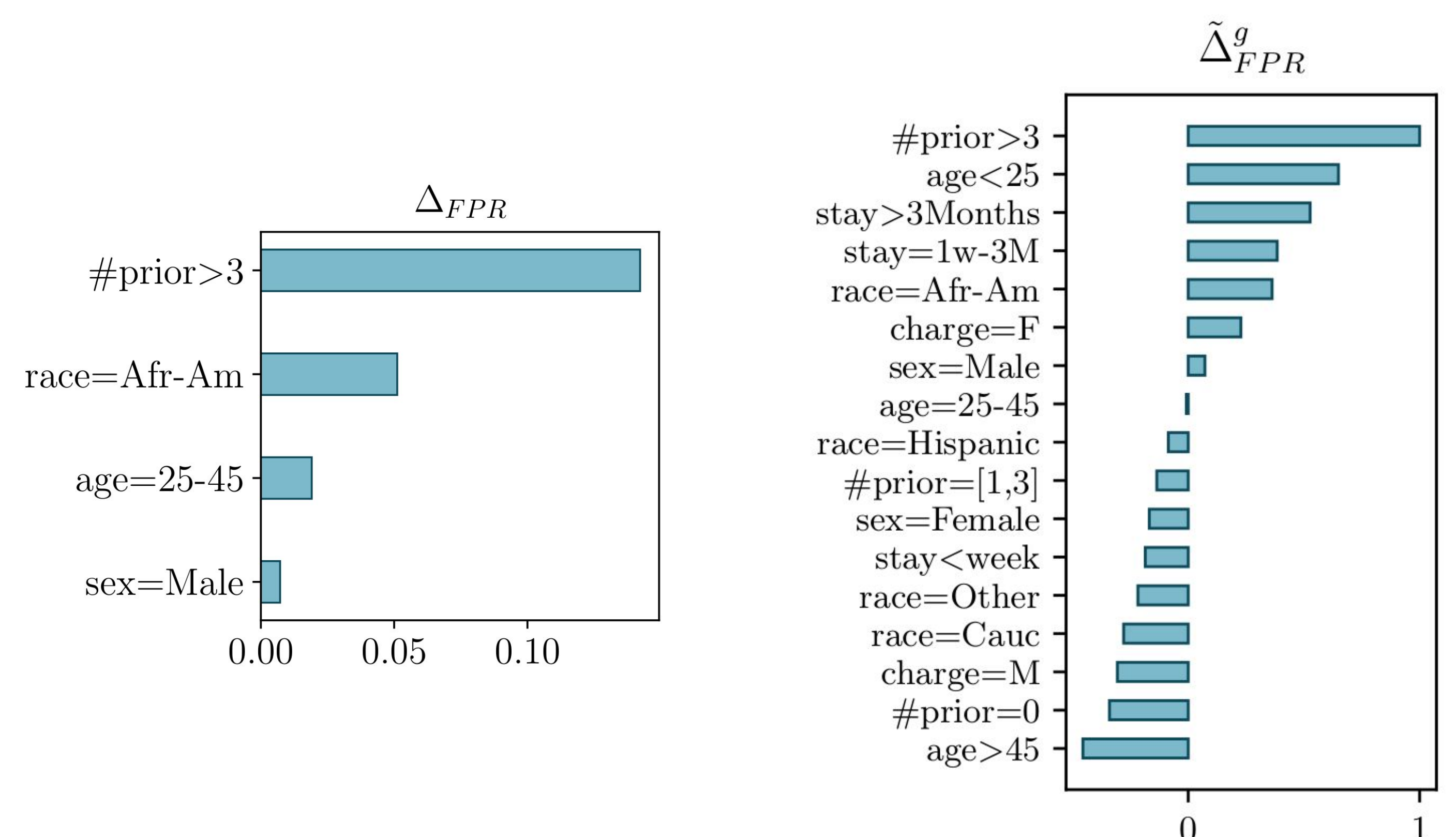
Evaluation typically considers overall performance

However, the model **behavior** can **differ across** data **subgroups**



DivExplorer

- **Automatic identification** of subgroups with **divergent behavior**
- **Local contribution** to divergence
- **Global contribution** to divergence



TRY IT AT DIVEXPLORER.ORG!

Adjustments:

Prune Redundancy by

0

Prune

Show Corrective Values

Search for specific records here

Clear

Search

Edit Columns

Support	Itemset	Δ_{fpr}
0.13	(age=25-45, #prior=>3, race=Afr-Am, sex=Male)	0.22
0.1	(#prior=>3, age=25-45, sex=Male, charge=F, race=Afr-Am)	0.217
0.06	(#prior=>3, sex=Male, stay=1w-3M)	0.216
0.15	(age=25-45, #prior=>3, race=Afr-Am)	0.211
0.07	(#prior=>3, stay=1w-3M)	0.207

Globals Computation:

Compute Global Values

<< 1 >>