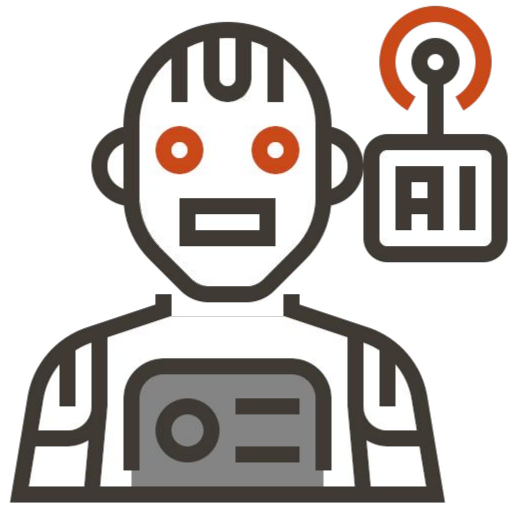
# GLocalX

## From Local to Global Explanations of Black Box AI Models

Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, Fosca Giannotti

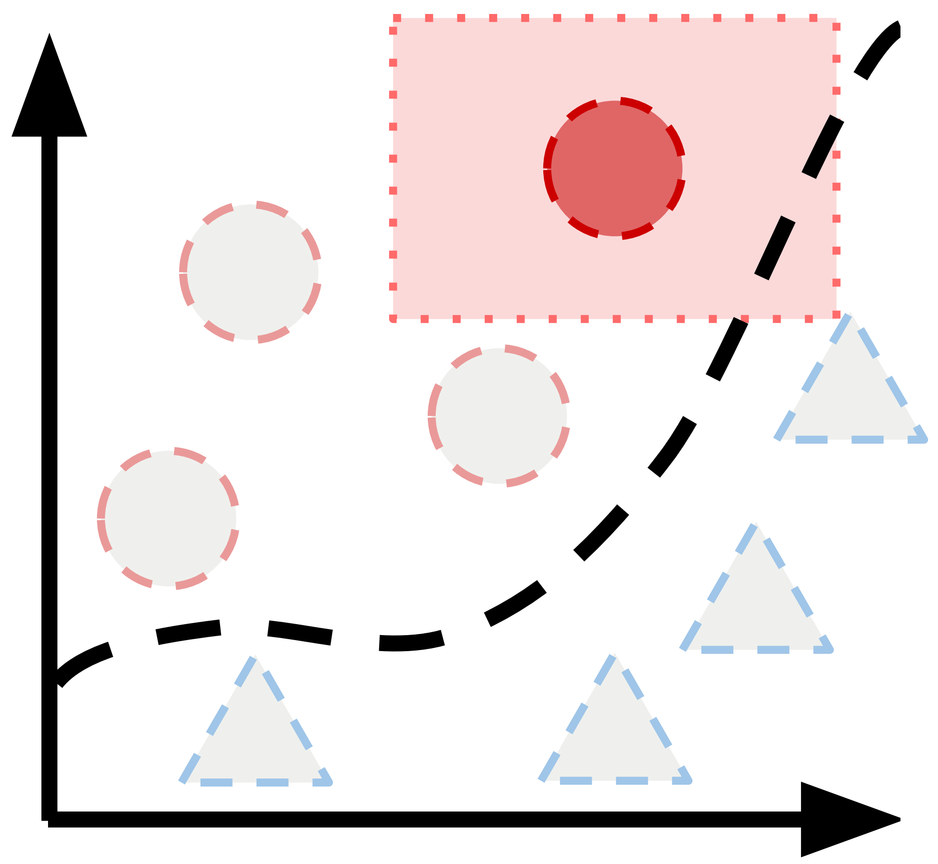
Consiglio Nazionale delle Ricerche

# Why explainability?




| ML developers | Users | Auditors |
|---|---|---|
| Debug | Act | Verify |

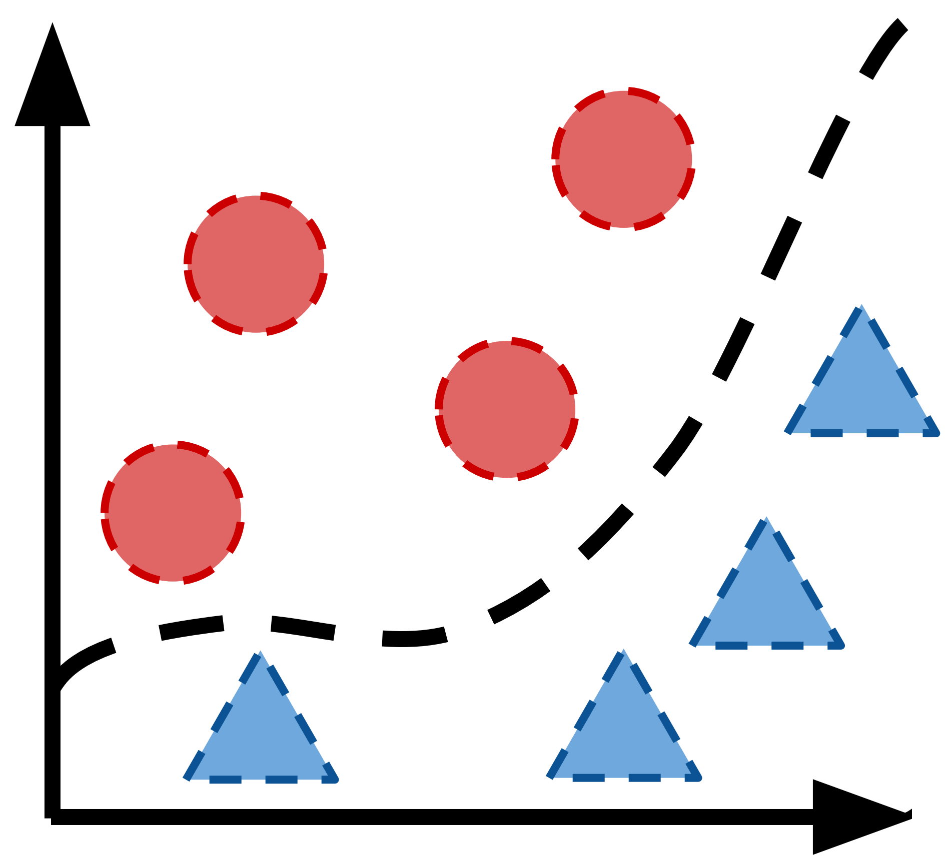
# Local VS Global explanations



Local explanations

- require only a fraction of the data
- more easily acquired
- precise but potentially complex

E.g. LIME, LORE, SHAP, etc.

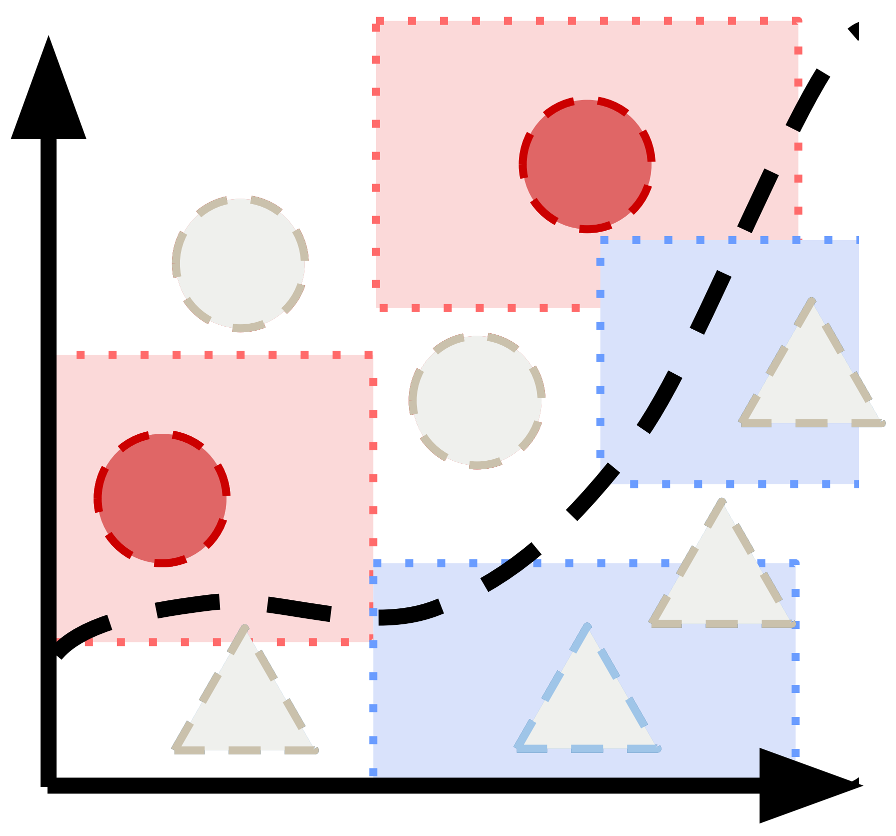


Global explanations

- require data
- more cumbersome to acquire
- loose but potentially simple

E.g. CART, CPAR, SBRL, etc.

# A third way: Local to Global



Local explanations

- require **only a fraction of the data**
- more **easily acquired**
- **precise** but potentially **complex**

E.g. LIME, LORE, SHAP, etc.

Global explanations

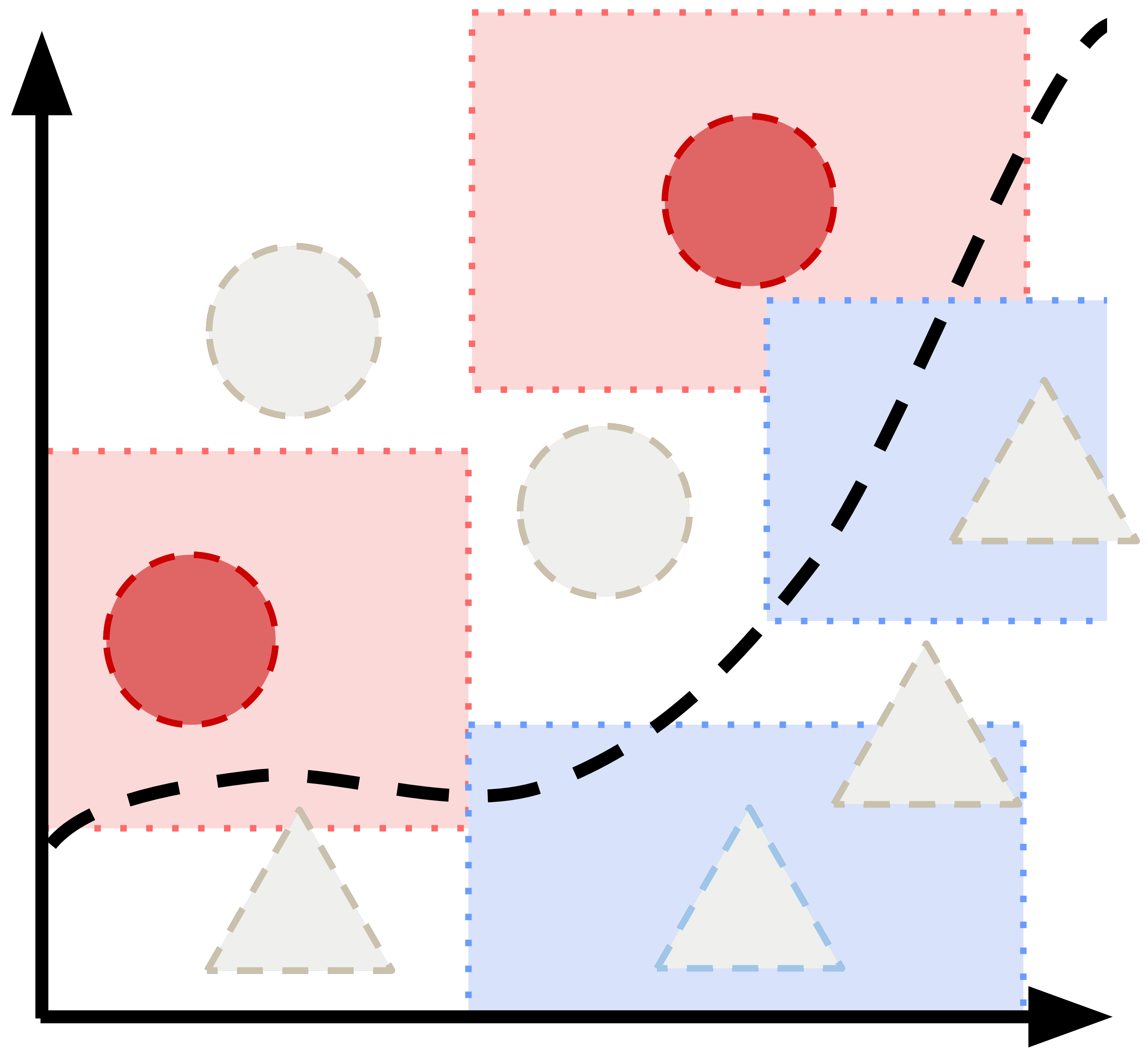
- require **data**
- more **cumbersome** to acquire
- **loose** but potentially **simple**

E.g. CART, CPAR, SBRL, etc.

# The Local to Global setting

Explain globally by explaining locally!

- explanation-driven
- inferring instead of learning
- black-box model as oracle



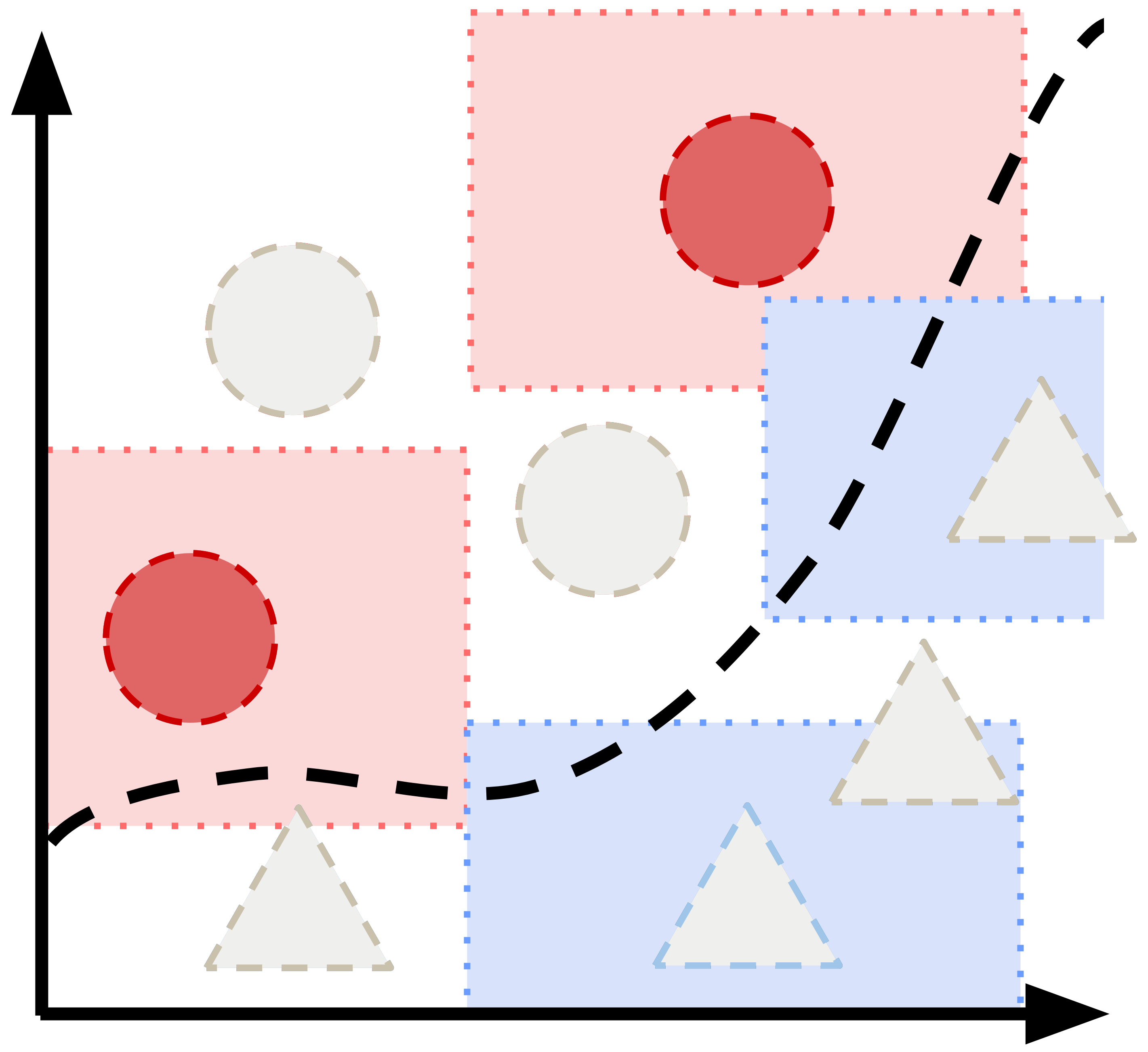
# The Local to Global setting

Explain globally by explaining locally!

- explanation-driven
- inferring instead of learning
- black-box model as oracle

Our proposal, **GLocalX**

- iterative and hierarchical inference
- axis-parallel decision rules as explanations

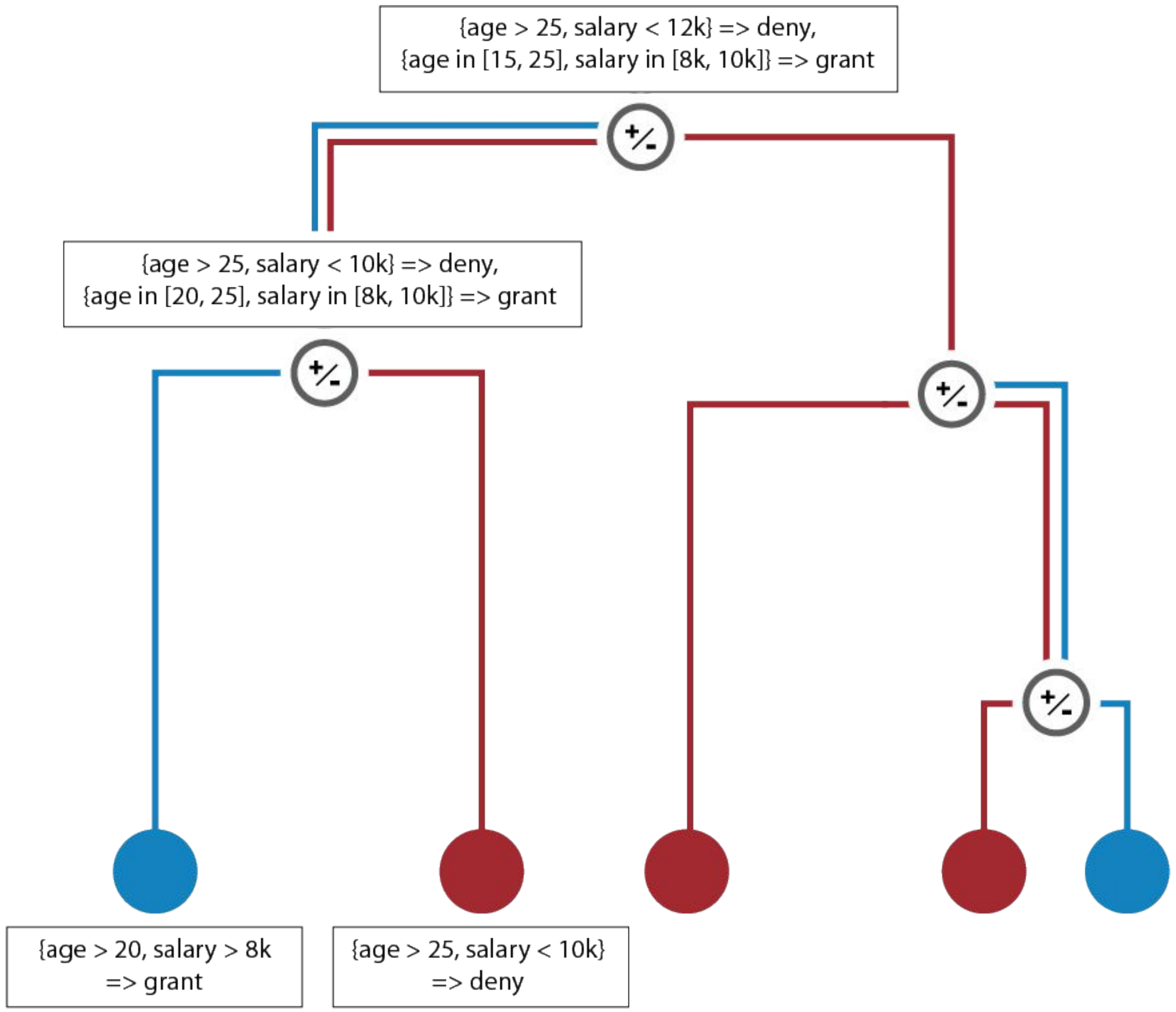


# What to merge?

- Distance between explanations

$$IoU(cov(e, X), cov(e', X))$$

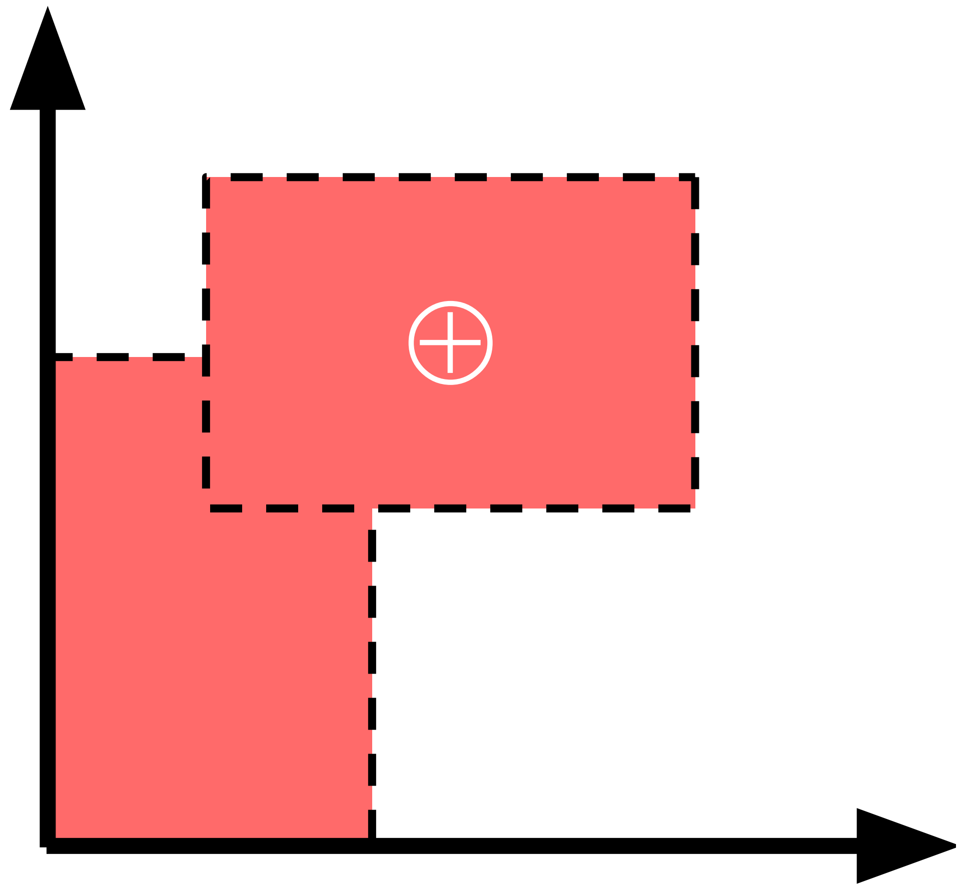
- Linkage for sets of explanations
  - `min`
  - `max`
  - `full`



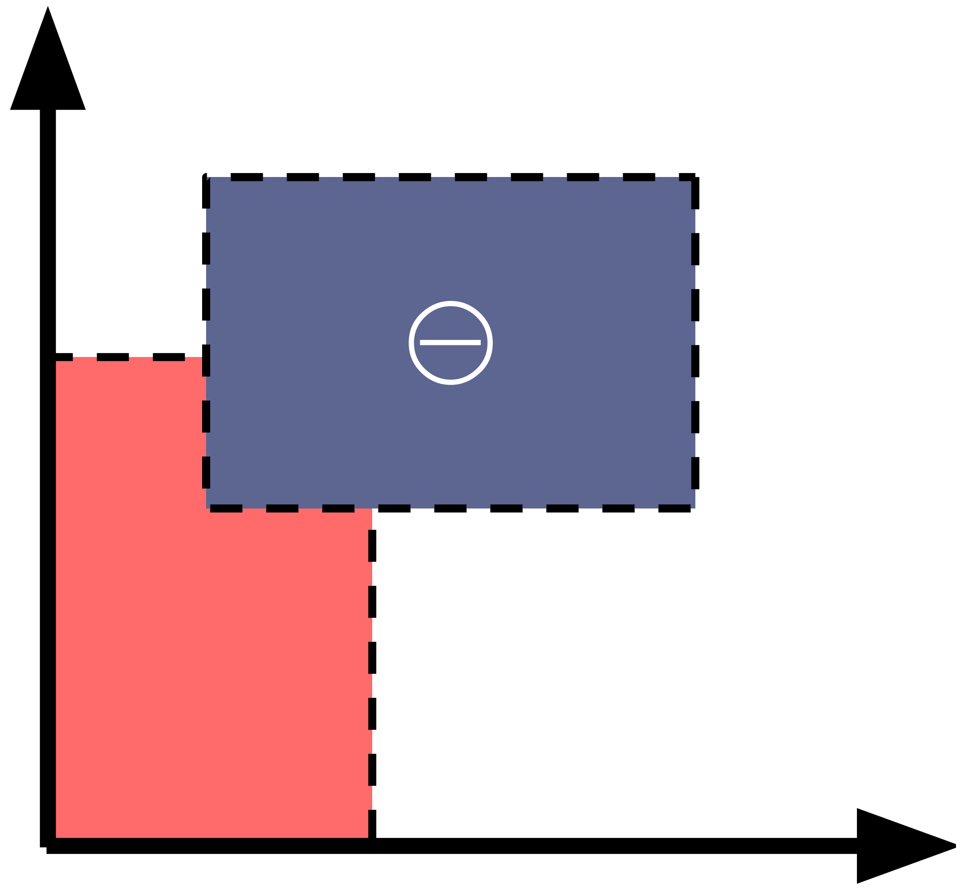
# How to merge?

Twofold merge operator

- approximate union (⊕) for concordance, approximate difference (⊖) for discordance
- each premise is an axis-parallel polyhedron, e.g. premise `age > 20` is polyhedron $P_{age}$`: [20, +∞)`




join




cut

# Join

From local to global via premise relaxation.

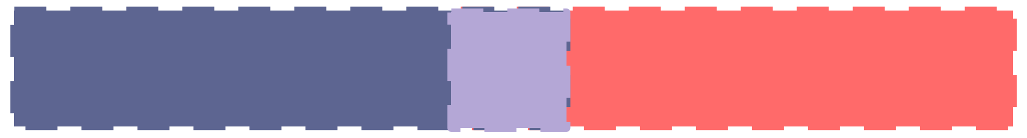
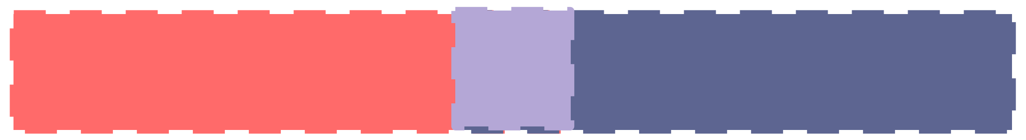
| $P_i: [a_P, b_p] + Q_i: [a_Q, b_Q]$ | | | |
|---|---|---|---|
| [non-empty] | $P_i, Q_i \neq \varnothing$ |  |  |
| [empty] | $P_i = \varnothing$ XOR $Q_i = \varnothing$ |  |  |

$\text{age} \in [15, 20) \oplus \text{age} \in [25, 40) =$


$\text{age} \in [15, 40)$

# Cut

From global to local via premise specification.

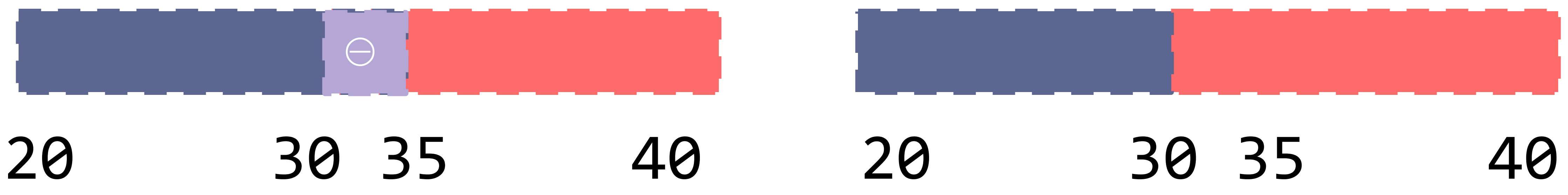
| $P_i$: $[a_P, b_p]$ - $Q_i$: $[a_Q, b_Q]$ | | | |
|---|---|---|---|
| [left] | $[a_P, a_Q]$ |  |  |
| [right] | $[b_P, b_Q]$ |  |  |
| [in-between] | $[a_Q, a_P]$, $[b_P, b_Q]$ |  |  |
| [everything] | $[a_<, a_P]$, $[b_P, b_>]$ |  |  |

cutting cut overlap

# Cut

From global to local via premise specification.

$\text{age} \in [30, 40) \ominus \text{age} \in [20, 35) =$




$\text{age} \in [30, 40), \text{age} \in [20, 30)$

cutting cut overlap

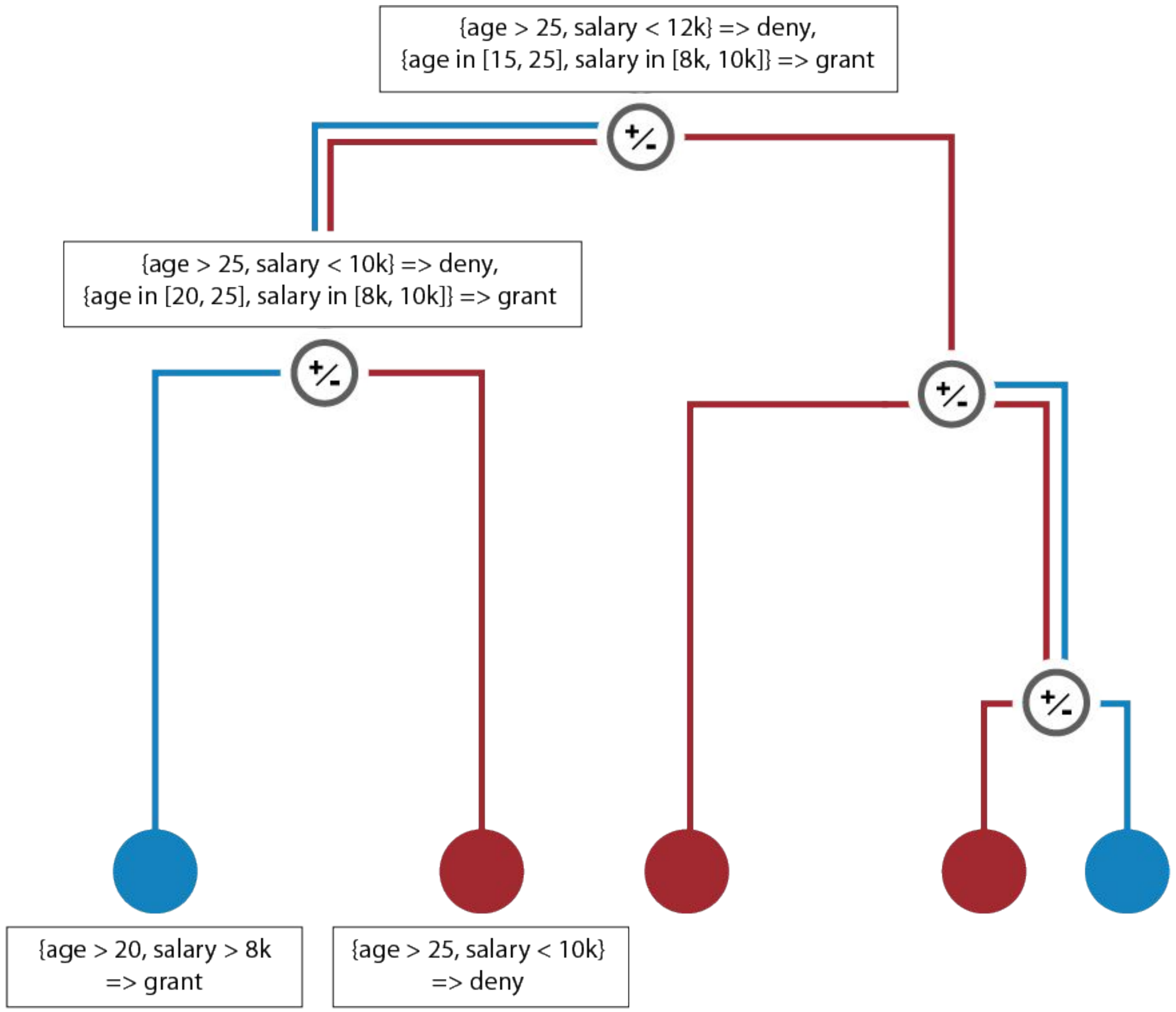
# Should we merge?

Not all merges are created equal!

- some are more global and less accurate
- some are less global and more accurate

`BIC(E)`

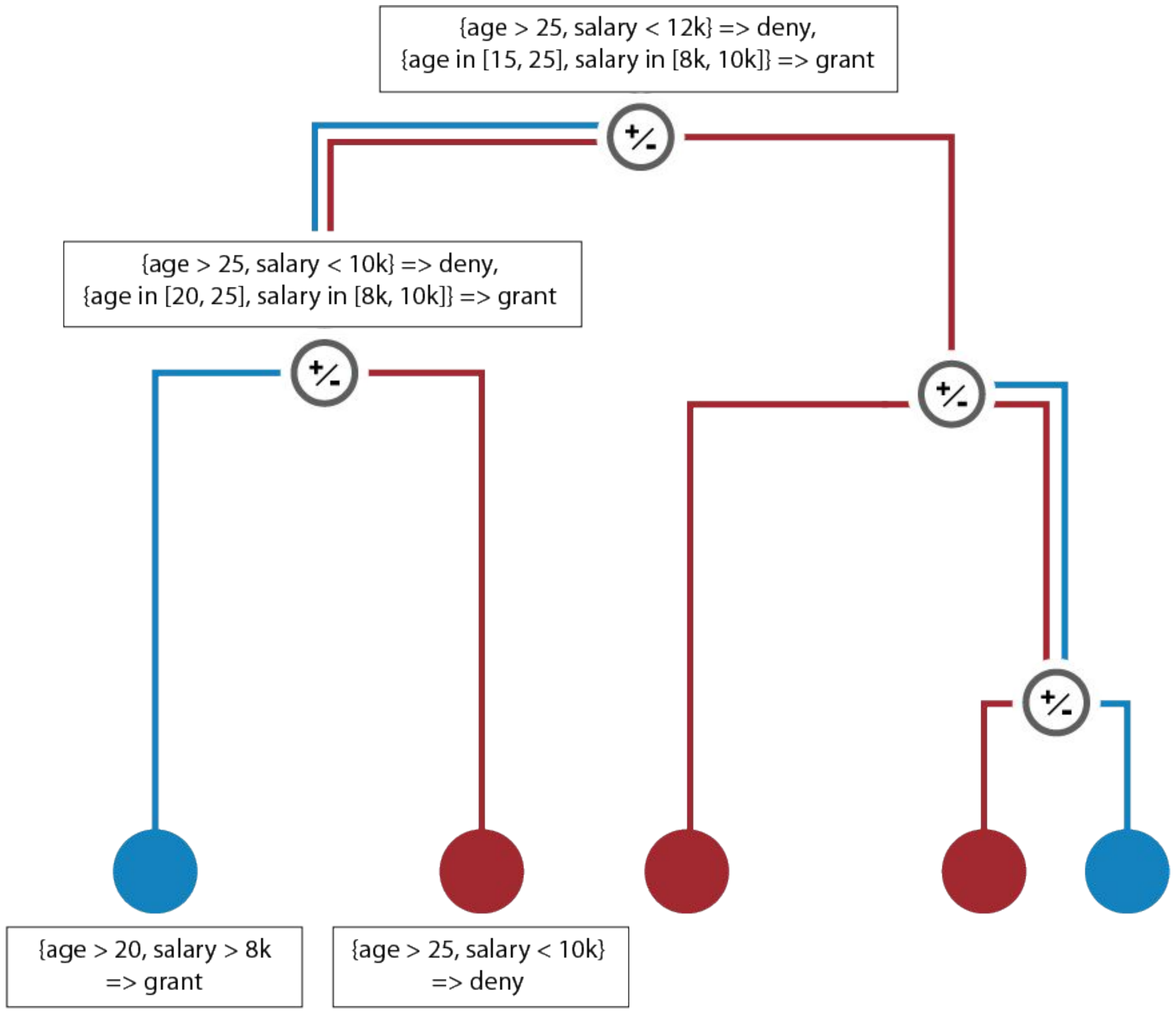
- model likelihood as explanation fidelity
- complexity as avg. #rules and avg. length



# 404: data not found

Data may be scarce for auditors and users

- density estimation of training data
- run GLocalX as is



# Full algorithm

**Algorithm 1** GLOCALX($\mathbb{E}, \alpha$)

**Input:** $\mathbb{E}$ explanation theories, $\alpha$ filter threshold
**Output:** $E$ explanation theory

**repeat**
  $\mathbb{Q} \leftarrow \textsc{sort}(\mathbb{E})$ ▷ `sort pairs of theories by similarity`
  $\mathit{merged} \leftarrow$ `False`
  $X' \leftarrow \mathrm{batch}(\mathrm{X})$
  **while** $\neg$ $\mathit{merged} \wedge \mathbb{Q} \neq \emptyset$ **do**
    $E_i, E_j \leftarrow \textsc{pop}(\mathbb{Q})$ ▷ `select most similar theories`
    $E_{i+j} \leftarrow \textsc{merge}(E_i,\ E_j,\ X')$ ▷ `merge theories`
    **if** $\textsc{bic}(E_{i+j}) \leq \textsc{bic}(E_i \cup E_j)$ **then** ▷ `verify improvement`
      $\mathit{merged} \leftarrow$ `True`
      **break**
  **if** $\mathit{merged}$ **then** ▷ `merge occurred`
    $\mathbb{E} \leftarrow \textsc{update}(E_i,\ E_j,\ E_{i+j})$ ▷ `update hierarchy`
**until** $\mid \mathrm{E} \mid > 1 \wedge \mathit{merged}$ ▷ `until the merge is successful`
$E \leftarrow \textsc{filter}(E, \alpha)$ ▷ `Filter final theory`
**return** $E$

# Full algorithm: filtering

**Algorithm 1** GLocalX($\mathbb{E}, \alpha$)

**Input:** $\mathbb{E}$ explanation theories, $\alpha$ filter threshold
**Output:** $E$ explanation theory

**repeat**
  $\mathbb{Q} \leftarrow \text{SORT}(\mathbb{E})$ ▷ sort pairs of theories by similarity
  $merged \leftarrow$ `False`
  $X' \leftarrow \text{batch}(X)$
  **while** $\neg\, merged \wedge \mathbb{Q} \neq \emptyset$ **do**
    $E_i, E_j \leftarrow \text{POP}(\mathbb{Q})$ ▷ select most similar theories
    $E_{i+j} \leftarrow \text{MERGE}(E_i, E_j, X')$ ▷ merge theories
    **if** $\text{BIC}(E_{i+j}) \leq \text{BIC}(E_i \cup E_j)$ **then** ▷ verify improvement
      $merged \leftarrow$ `True`
      **break**
  **if** $merged$ **then** ▷ merge occurred
    $\mathbb{E} \leftarrow \text{UPDATE}(E_i, E_j, E_{i+j})$ ▷ update hierarchy
**until** $\mid \text{E} \mid > 1 \wedge merged$ ▷ until the merge is successful
$\boxed{E \leftarrow \text{FILTER}(E, \alpha)}$ ▷ Filter final theory
**return** $E$

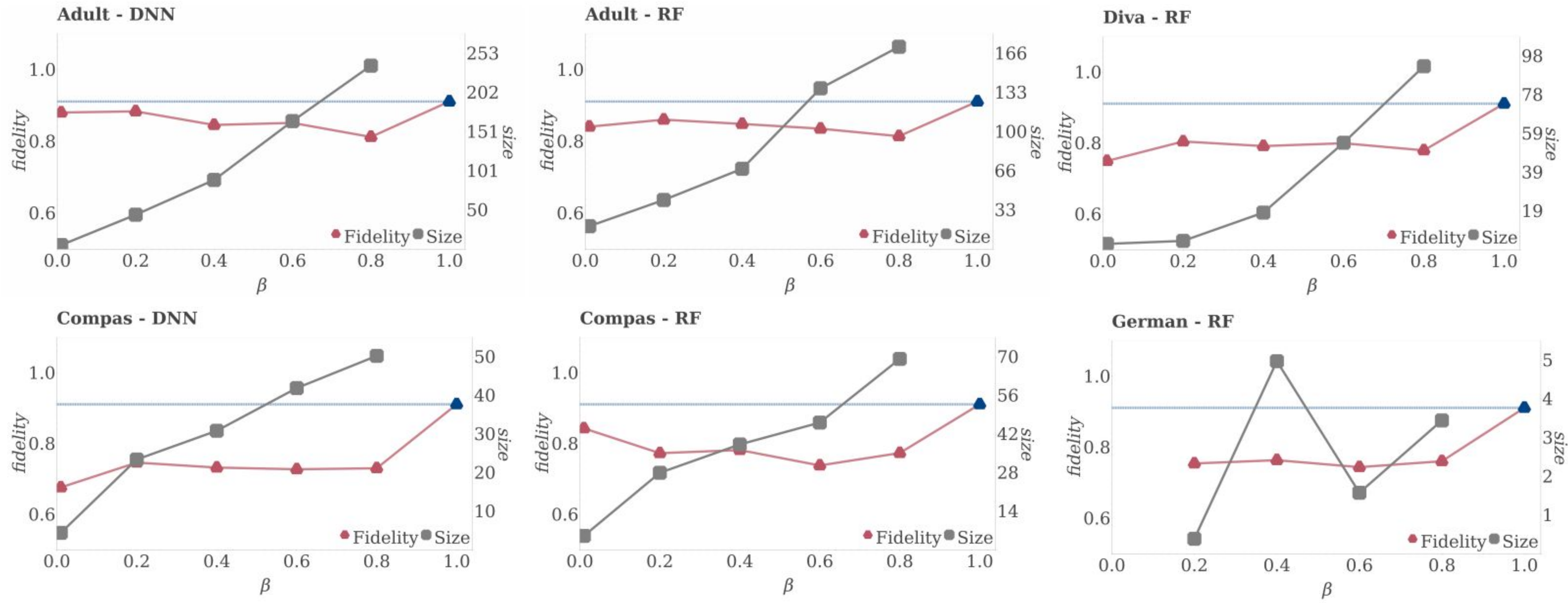
# Validation

# Setting

- 3 UCI datasets (~1k to ~50k records) , 8 black boxes (DNN, RF, SVM)
- 1 real-world fraud detection dataset (from the Italian Ministry of Economics)
- Natively global models:
  - rule-based models (CPAR)
  - decision tree (pruned/not pruned)

# Input size: how many rules do we need?

Acquiring local explanation can be costly, can we get away with using fewer local explanations?



# How simple can we make our explanations?

The higher the filter, the less rules we output.

| *α-percentile* | **Fidelity** | **Size** | **Length** |
|---|---|---|---|
| *75* | 83.0 ± 3.6 | 31.0 ± 19.4 | 5.36 ± 2.41 |
| *90* | 84.7 ± 5.14 | 11.5 ± 6.4 | 5.43 ± 2.46 |
| *95* | 84.5 ± 5.48 | 6.625 ± 2.9 | 5.17 ± 2.59 |
| *99* | 84.0 ± 5.0 | 3.625 ± 2.6 | 5.97 ± 3.04 |

# GLocalX VS natively global models

| | Fidelity | Size | Length |
|---|---|---|---|
| *GLocalX* | 85.1 | **8.5** | 4.28 ± 1.42 |
| *GLocalX** | 83.5 | 9.5 | 4.79 ± 1.67 |
| *CPAR* | 86.6 | 91.6 | 3.06 ± 1.66 |
| *Decision Tree* | **87.5** | 1036.5 | 6.60 ± 1.86 |
| *Pruned Decision Tree* | 85.5 | 29.1 | **2.64 ± 0.73** |
| *Union* | 76.8 | 2660.2 | 4.14 ± 1.63 |

# GLocalX

## From Local to Global Explanations of Black Box AI Models

- Explaining globally by explaining locally
- Explanation cost: how many explanations do we really need?
- Local to Global *vs* Global explanation paradigm

github.com/msetzu/glocalx

mattia.setzu@phd.unipi.it