Past XAI work:
- GLocalX, RSS (tabular)
- TriplEx (text)

Current XAI work
- parametric explanations (tabular)
- sub-global explanations (tabular)
- language model beliefs (text)
- authorship attribution (text)
- neurosymbolic few-shot learning (images)

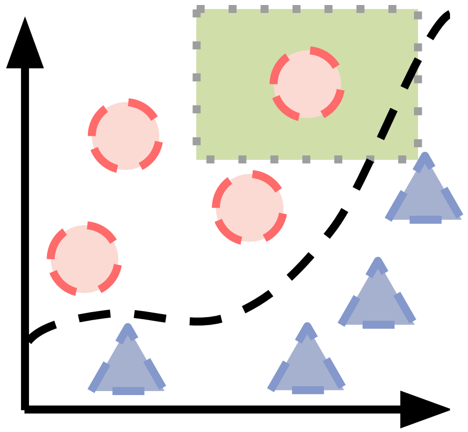
GLocalX

From local to global explanations

Mattia Setzu, Research fellow @UniPI

1/12/2022

Local explanations



Local explanations

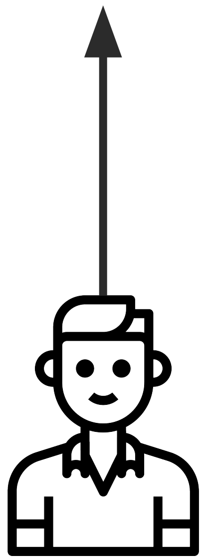
- explain one prediction on one record
- locally approximate the decision boundary

E.g. LIME[1], LORE[2], SHAP[3], etc.

[1] "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Ribeiro et al.
[2] Factual and Counterfactual Explanations for Black Box Decision Making, Guidotti et al.
[3] A Unified Approach to Interpreting Model Predictions, Lundberg & Lee

Local explanation

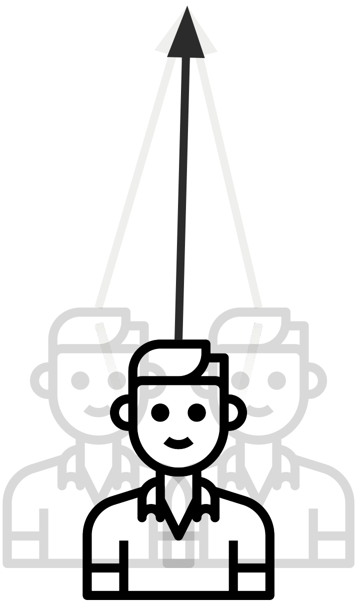
```
age <= 30, salary >= 2.2k,
status=”married” => grant loan
```

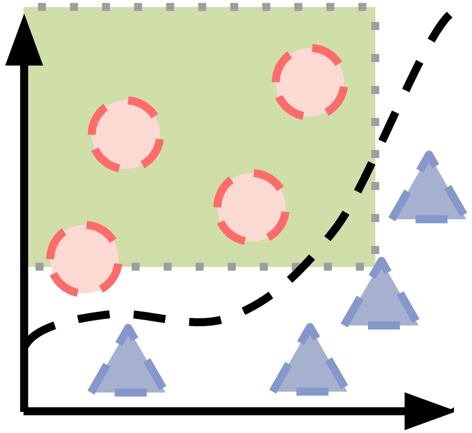


Global explanations

Global explanation

```
age <= 30,
salary >= 2.2k => grant loan
```



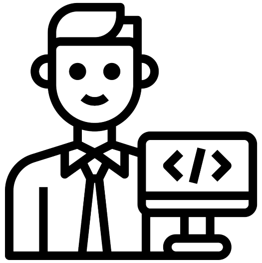


Global explanations
- explain all predictions on many records
- globally approximate the decision boundary

E.g. CART[4], CPAR[5], SBRL[6], etc.

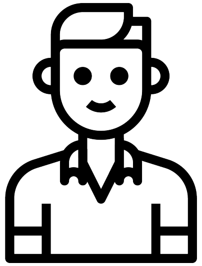
[4] Classification and Regression Trees, Breiman et al.
[5] CPAR: Classification based on Predictive Association Rules, Yin et al.
[6] Scalable Bayesian Rule Lists, Yang et al.

Who are the explanation users?



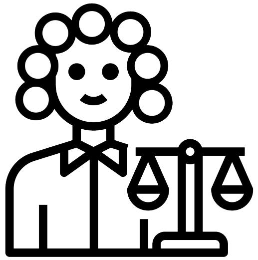
ML developer

Debug



End user

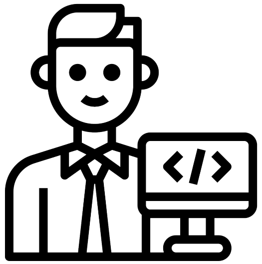
Act

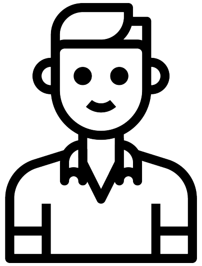


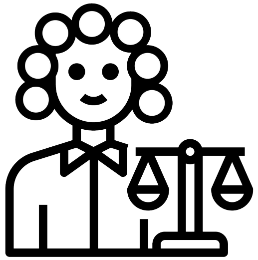
Auditor

Verify

Who are the explanation users?

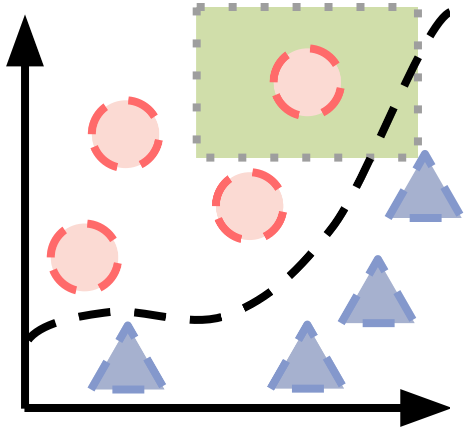






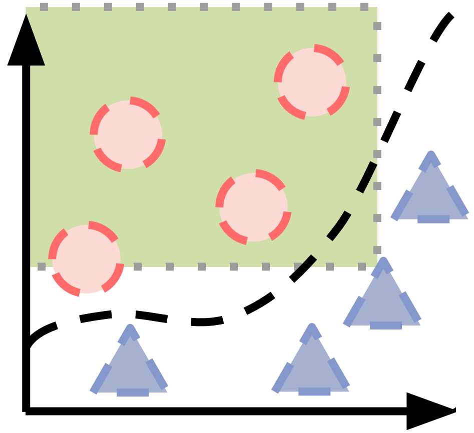
ML developer	End user	Auditor
• Has global access • Desires local and global understanding	• Little to no access • Desires local understanding	• Little to no access • Desires global understanding

Local **VS** Global



Local explanations
- require **only a fraction of the data**
- more **easily acquired**
- **precise** but potentially **complex**
- possibly diverse[7,8]

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

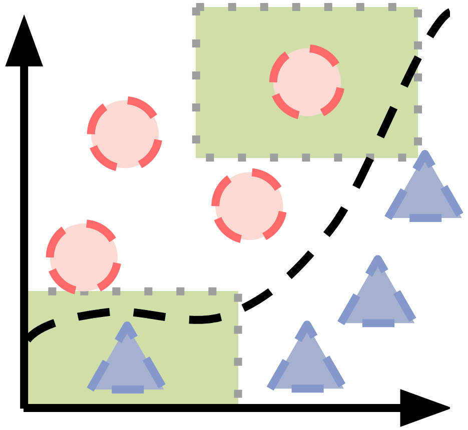


Global explanations
- require **data**
- more **cumbersome** to acquire
- **loose** but potentially **simple**

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

[7] Ensembles of locally independent prediction models, Ross et al.
[8] Learning qualitatively diverse and interpretable rules for classification, Ross et al.

Local **TO** Global



Local explanations

- require **only a fraction of the data**
- more **easily acquired**
- **precise** but potentially **complex**
- possibly diverse[1,2]

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

Global explanations

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

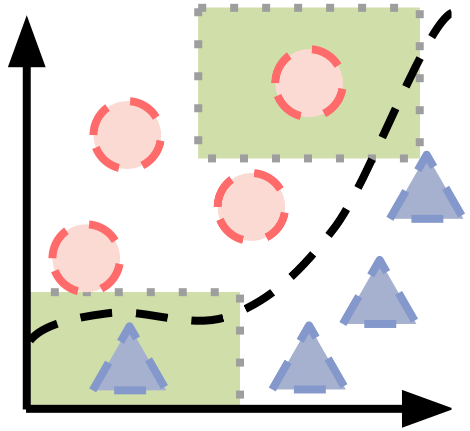
E.g. DT, CART, CPAR, SBRL, etc.

[7] Ensembles of locally independent prediction models, Ross et al.
[8] Learning qualitatively diverse and interpretable rules for classification, Ross et al.
[9] Meaningful explanations of black box ai decision systems, Pedreschi et al.

Explain globally by explaining locally!

- explanation-driven (decision rules)
- model-agnostic
- inferring instead of learning

GLocalX[10]: local-to-global decision rules as explanations

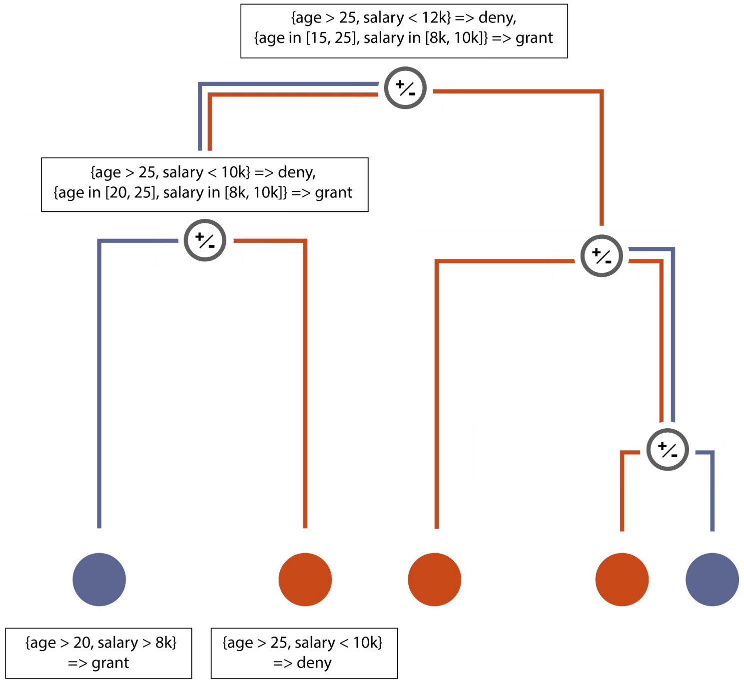


[10] GLocalX - From Local to Global Explanations of Black Box AI Models, Setzu et al.

Explain globally by explaining locally!

GLocalX[10]:

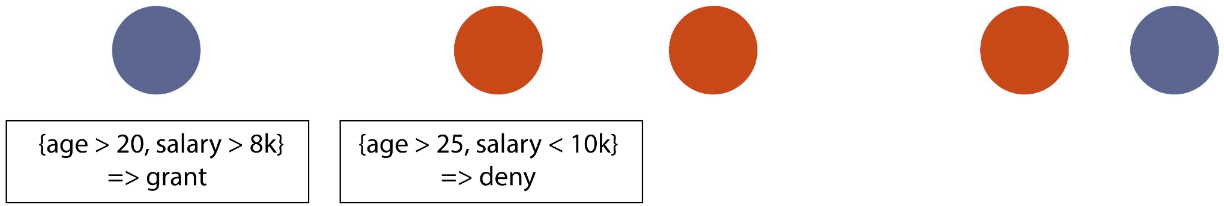
- input: local decision rules
- output: global decision rules
- inferring instead of learning
- model-agnostic



[10] GLocalX - From Local to Global Explanations of Black Box AI Models, Setzu et al.

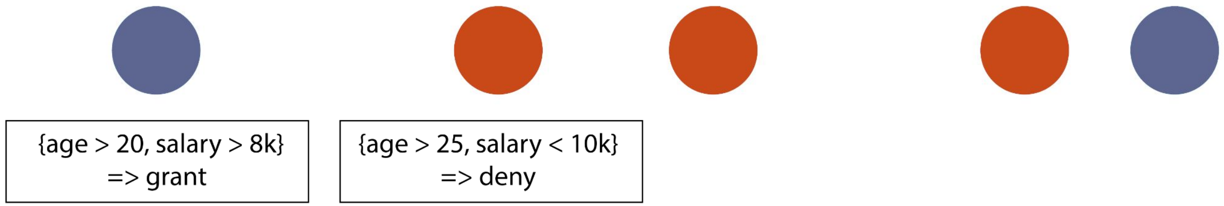
GLocalX

```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
```



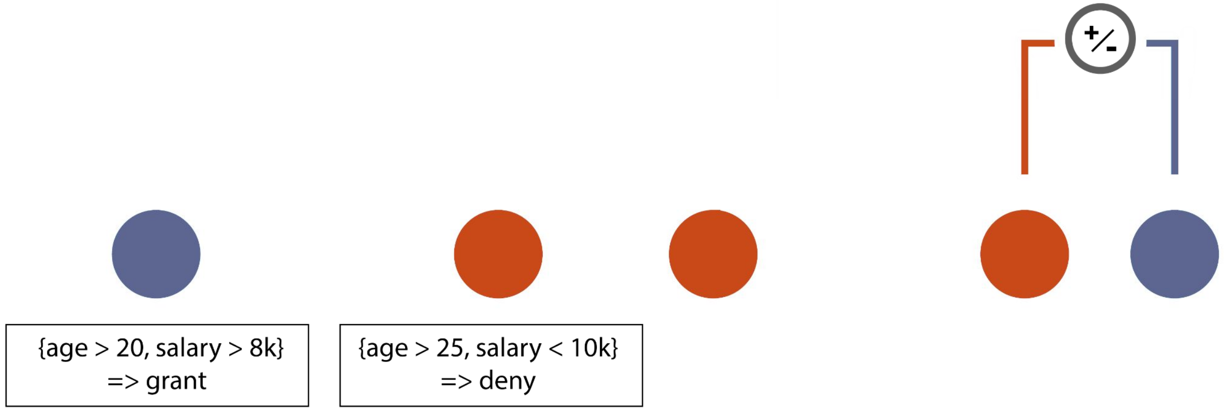
GLocalX

```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
```



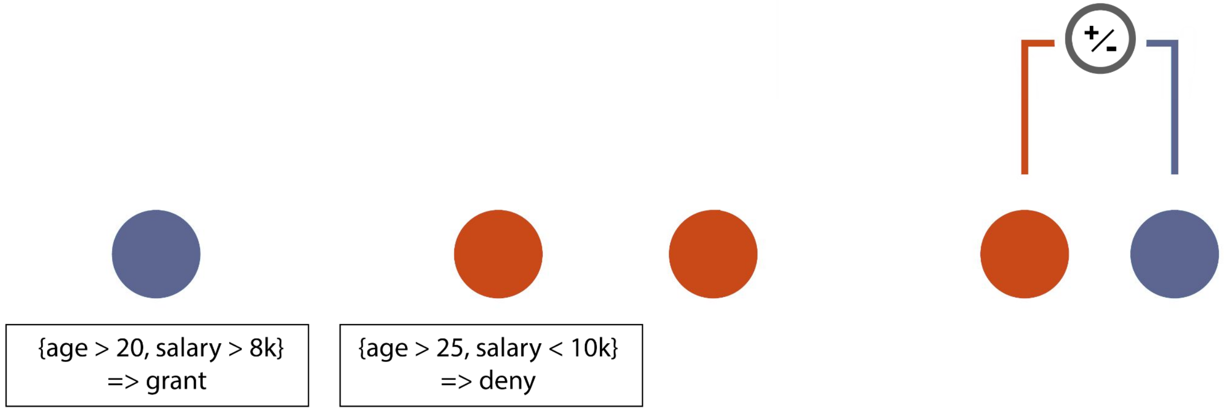
GLocalX

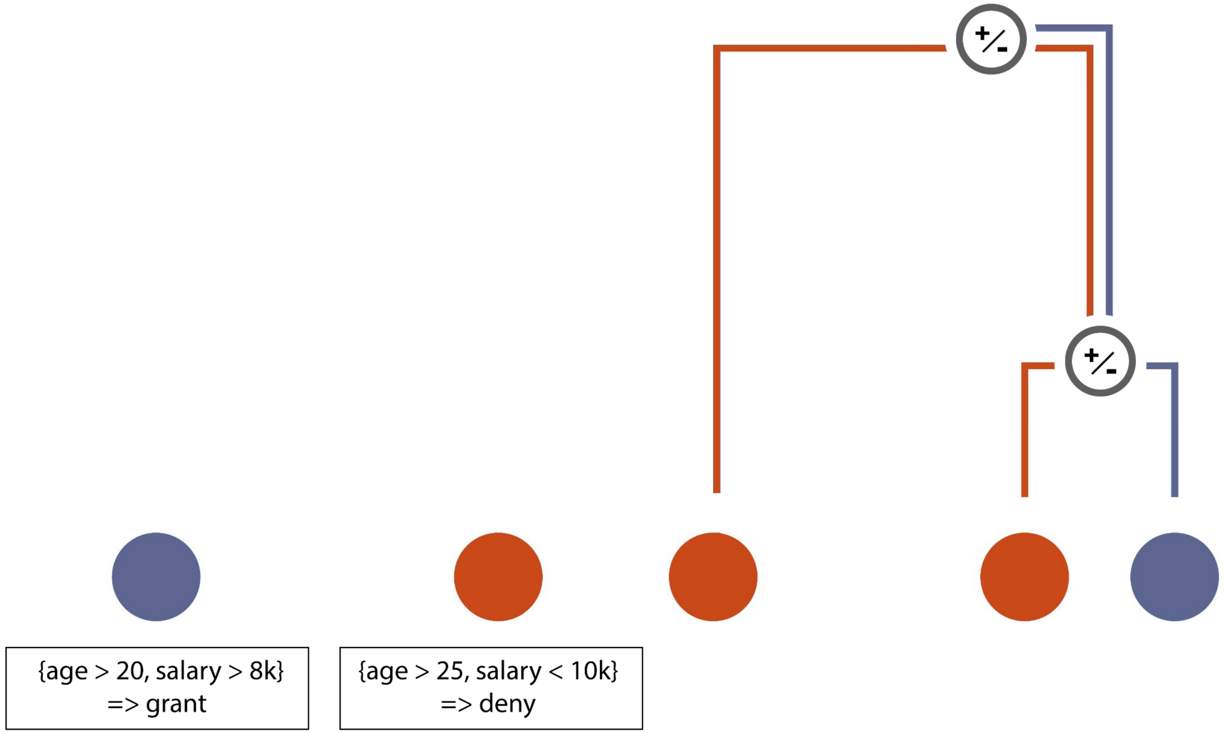
```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
    while len(q) > 1:
        e1, e2 = pop(q)
        M = merge(e1, e2, batch(X), f)
```



GLocalX

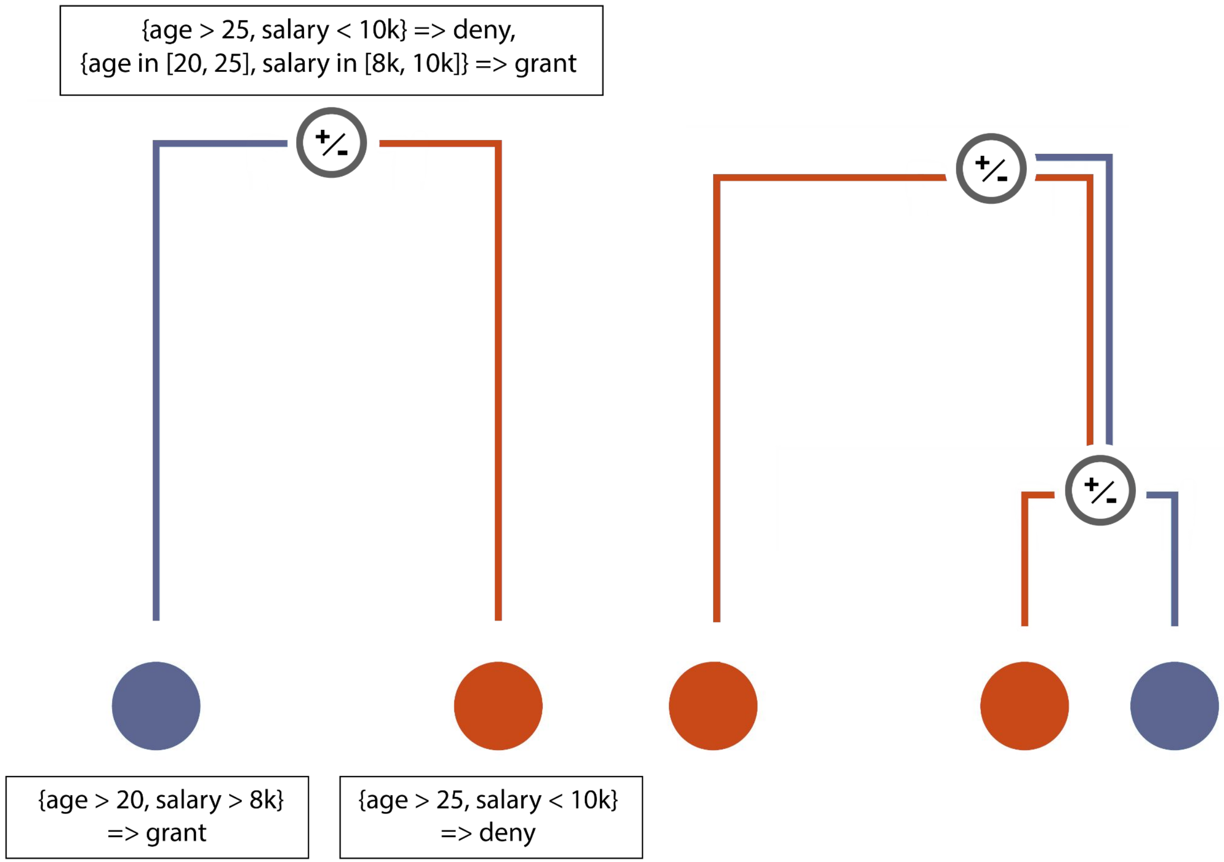
```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
    while len(q) > 1:
        e1, e2 = pop(q)
        M = merge(e1, e2, batch(X), f)
        if fitness(e1, e2, M, f, X):
            replace(boundary,
                            (e1, e2), M)
             q = sort(boundary, X)
            break
```



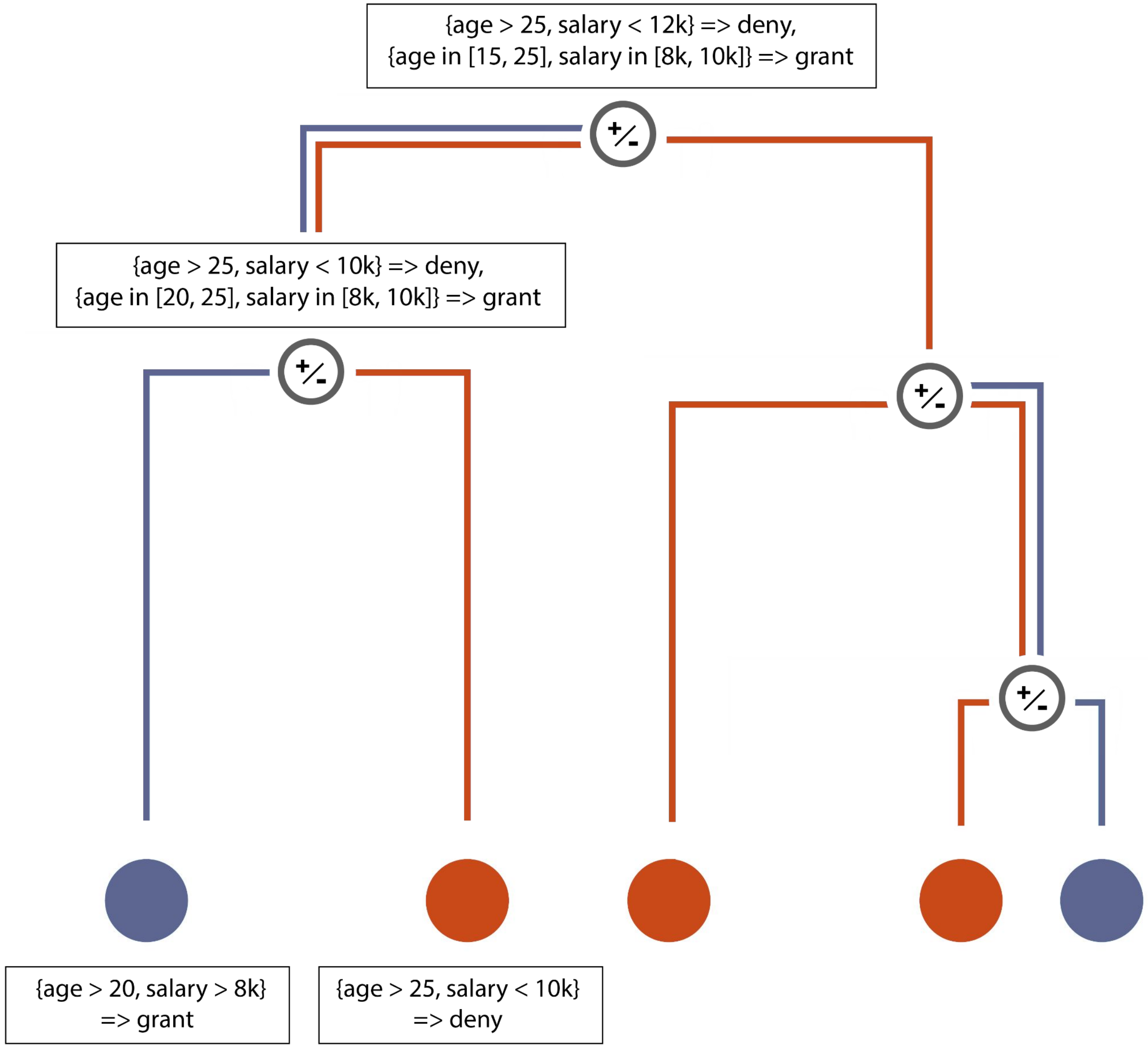



```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
    while len(q) > 1:
        e1, e2 = pop(q)
        M = merge(e1, e2, batch(X), f)
        if fitness(e1, e2, M, f, X):
            replace(boundary,
                            (e1, e2), M)
             q = sort(boundary, X)
            break
```

GLocalX

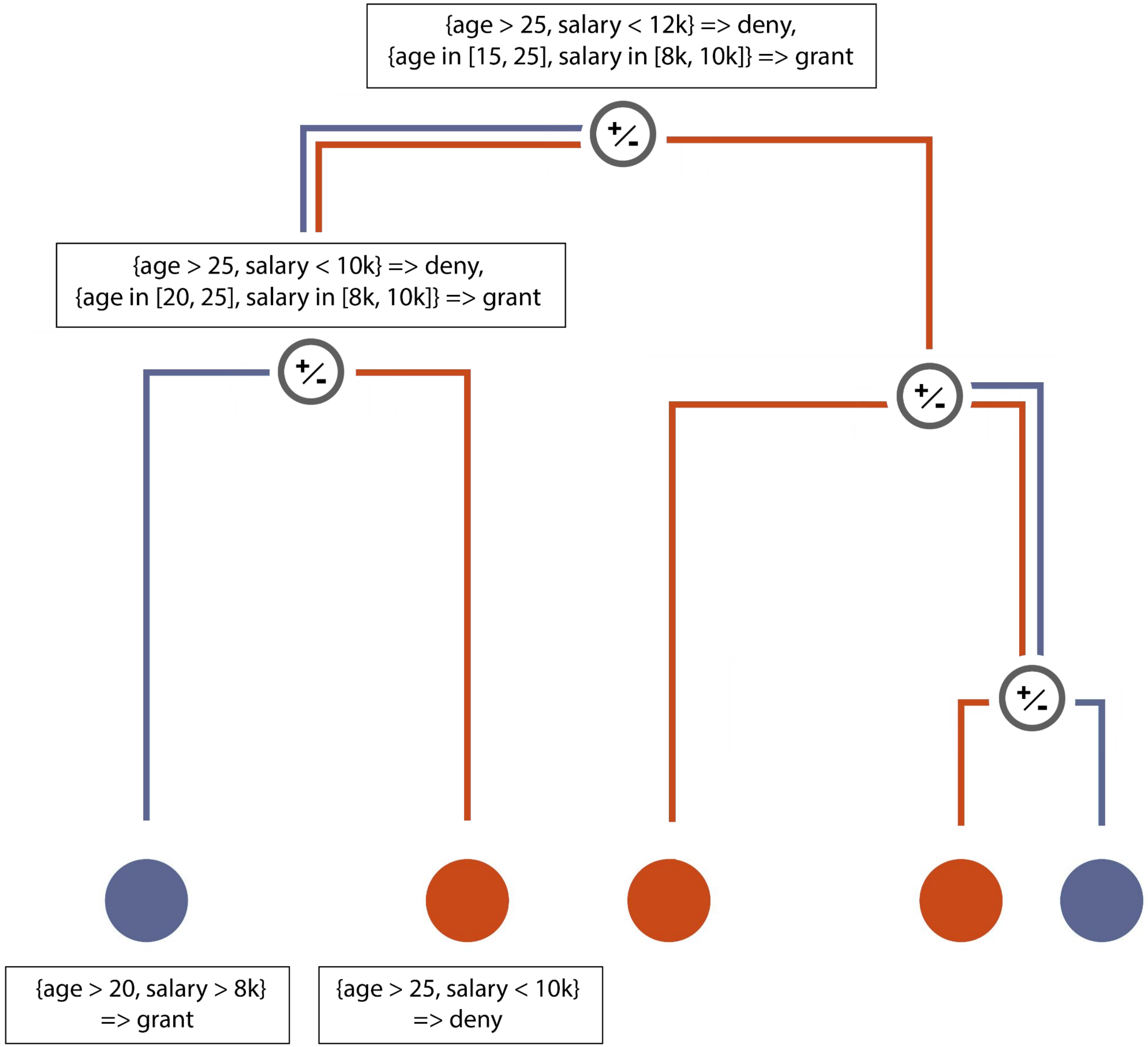



```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
    while len(q) > 1:
        e1, e2 = pop(q)
        M = merge(e1, e2, batch(X), f)
        if fitness(e1, e2, M, f, X):
            replace(boundary,
                            (e1, e2), M)
             q = sort(boundary, X)
            break
```



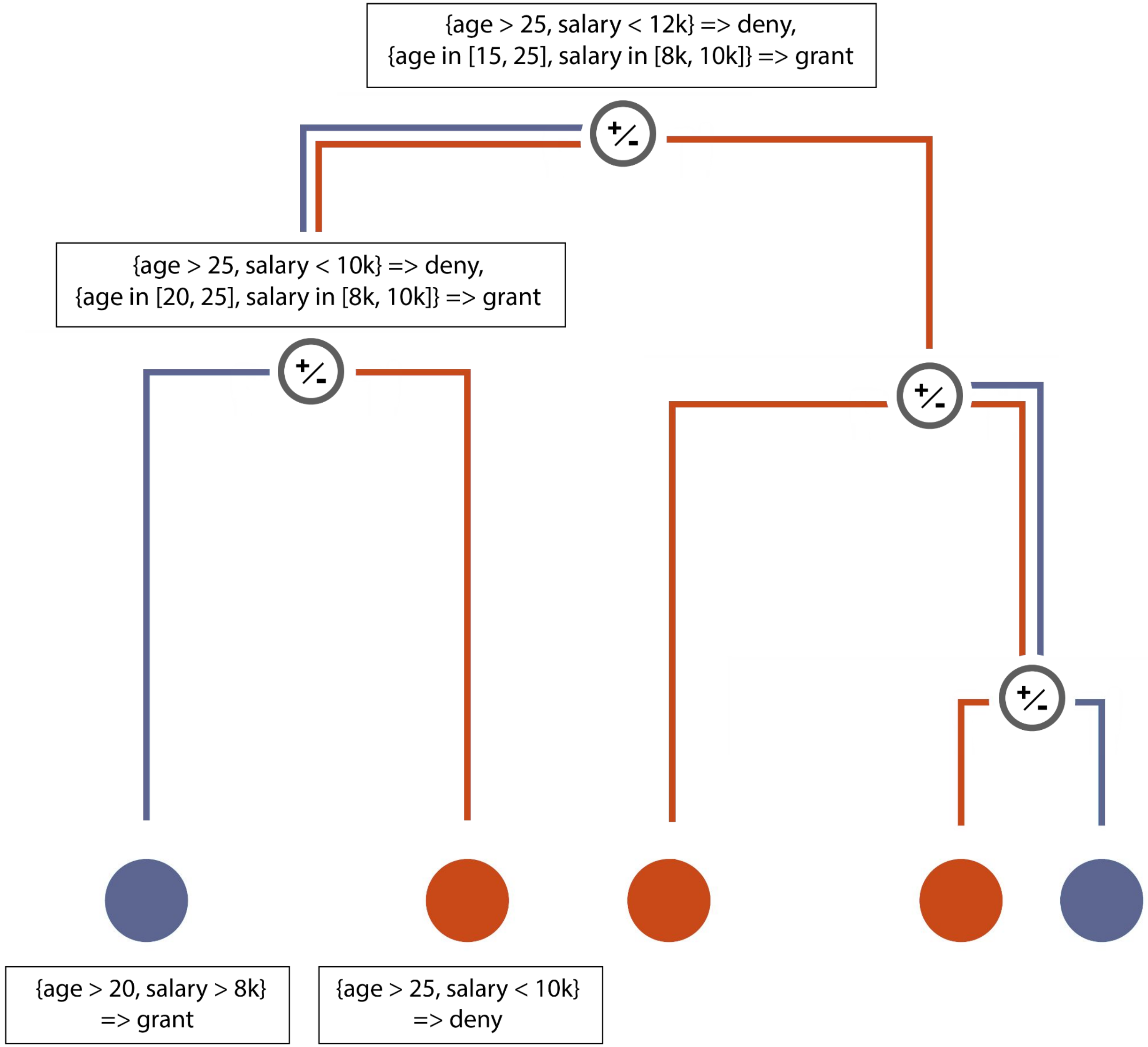

```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
    while len(q) > 1:
        e1, e2 = pop(q)
        M = merge(e1, e2, batch(X), f)
        if fitness(e1, e2, M, f, X):
            replace(boundary,
                            (e1, e2), M)
             q = sort(boundary, X)
            break
```

GLocalX




```python
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
    while len(q) > 1:
        e1, e2 = pop(q)
        M = merge(e1, e2, batch(X), f)
        if fitness(e1, e2, M, f, X):
            replace(boundary,
                            (e1, e2), M)
             q = sort(boundary, X)
            break
    return filter(boundary, a)
```

GLocalX




```
def glocalx(local_exp, X, f, a):
    boundary = copy(local_exp)
    q = sort(boundary, X)
  while len(q) > 1:
    e1, e2 = pop(q)
         M = merge(e1, e2, batch(X), f)
     if fitness(e1, e2, M, f, X):
                replace(boundary,
                                (e1, e2), M)
                 q = sort(boundary, X)
                break
    return filter(boundary, a)
```

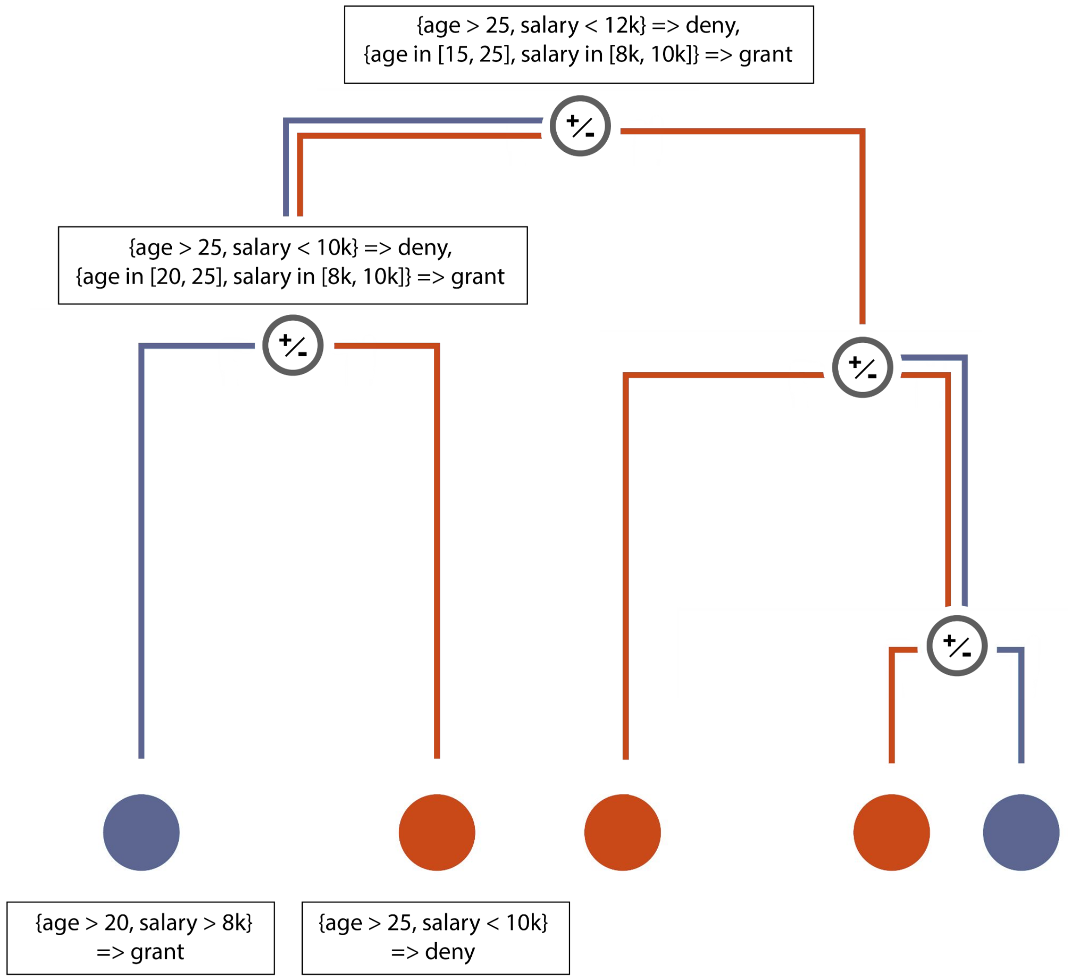
What to merge?

sort merge fitness

- Distance between explanations

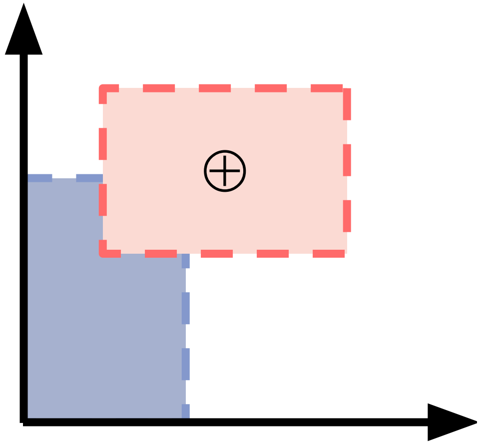
$$IoU(cov(e, X), cov(e', X))$$

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

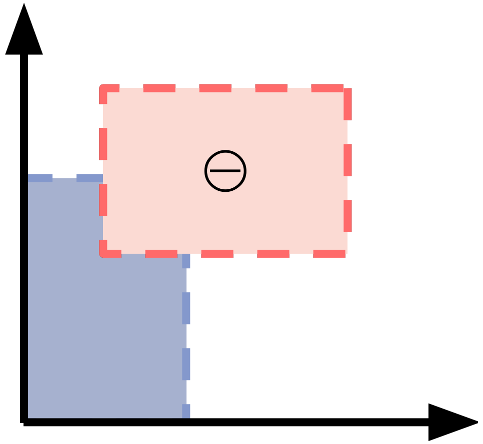


Twofold merge operator

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

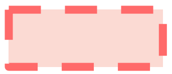


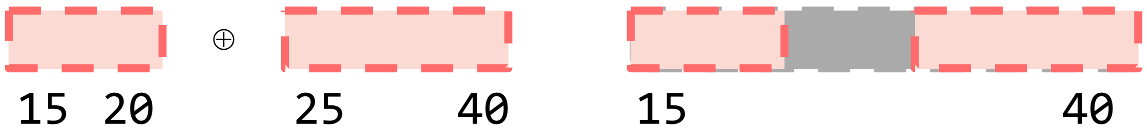
join



cut

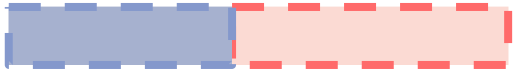
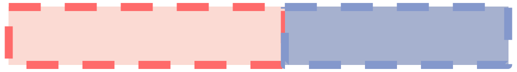
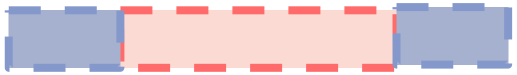
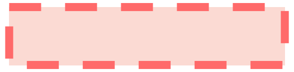
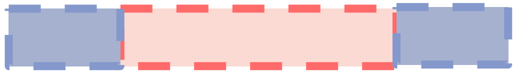
From local to global via premise relaxation.

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 $Q_i = \varnothing$		

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

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

From global to local via premise specification.

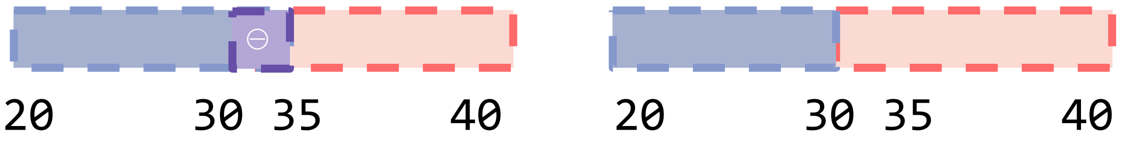
P_i: [a_P, b_p] - Q_i:[a_Q, b_Q]			
[left]	[a_P, a_Q]		
[right]	[b_P, b_Q]		
[in-between]	[a_Q, a_P], [b_P, b_Q]		
[everything]	[$a_<$, a_P], [b_P, $b_>$]		

cutting cut overlap

Cut

From global to local via premise specification.

age ∈ [30, 40) ⊖ age ∈ [20, 35) =



age ∈ [30, 40), age ∈ [20, 30)

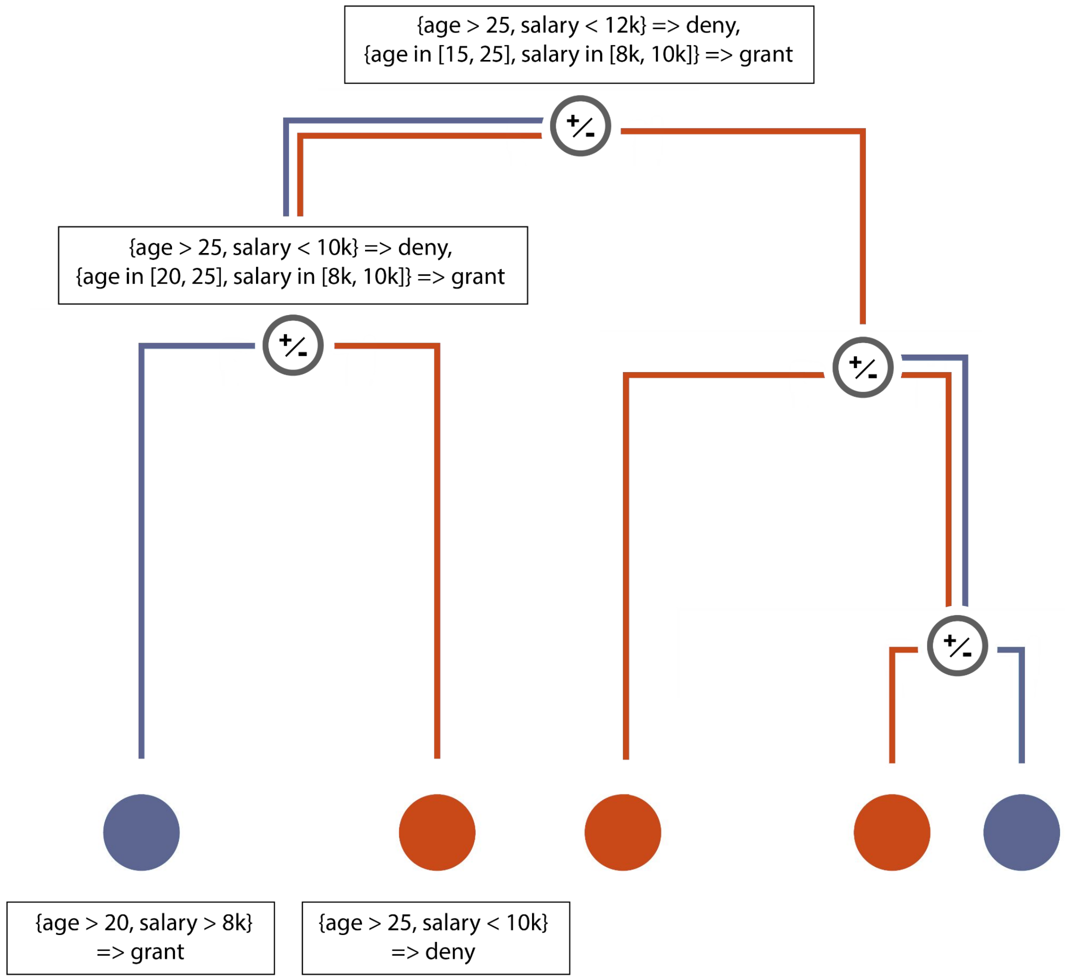
cutting cut overlap

Should we merge?

sort merge fitness

Not all merges are created equal!
- some are more global and less accurate
- some are less global and more accurate

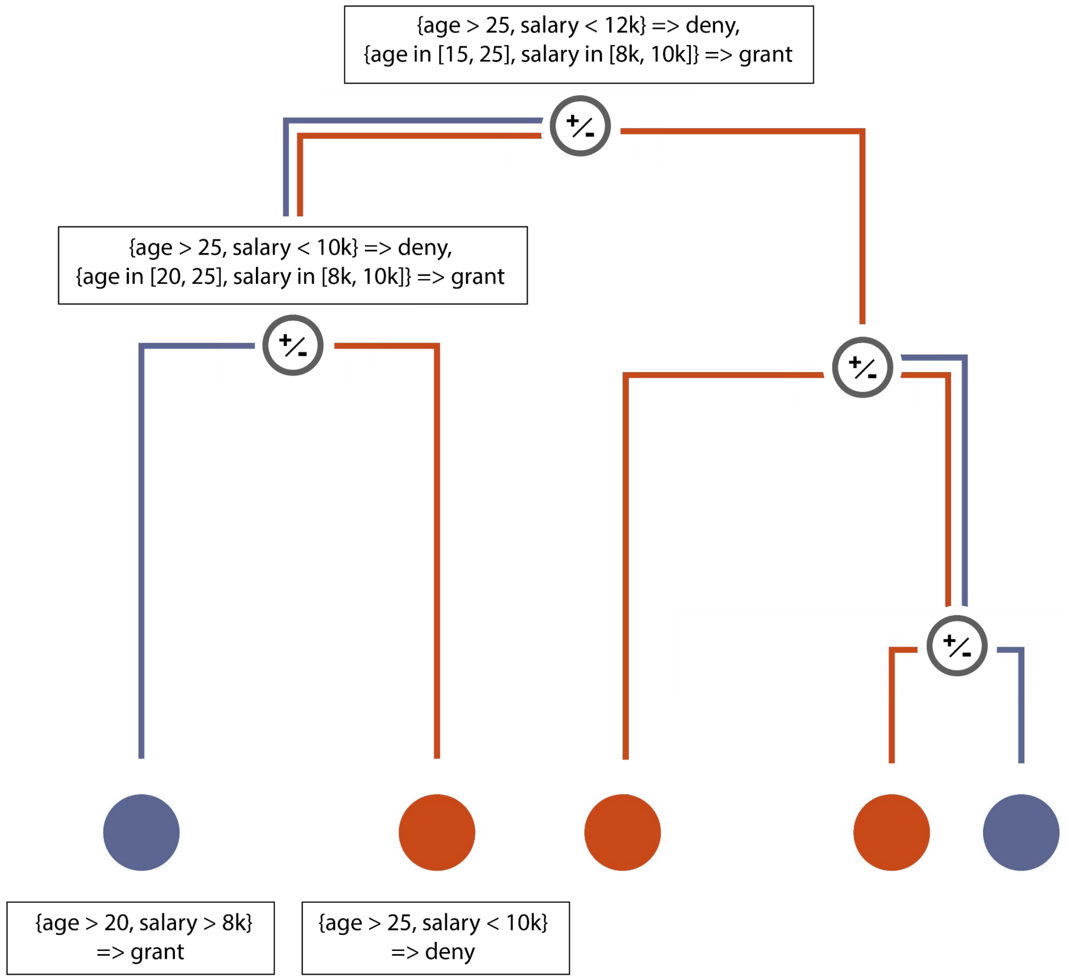
BIC(E)
- model likelihood as explanation fidelity
- complexity as avg. #rules and avg. length



404: data not found!

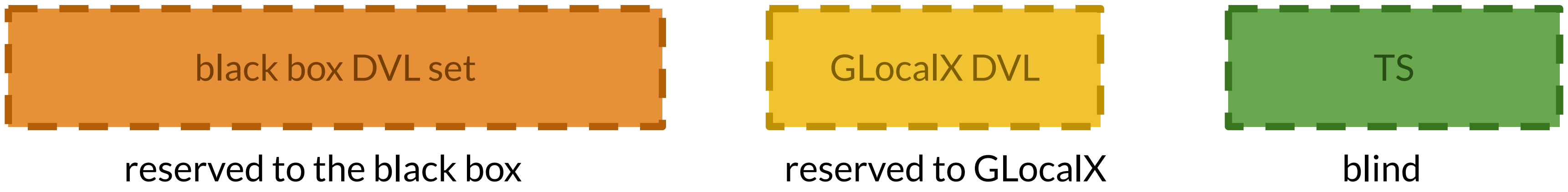
Data may be scarce for auditors and users

- density estimation of training data
- run GLocalX as is



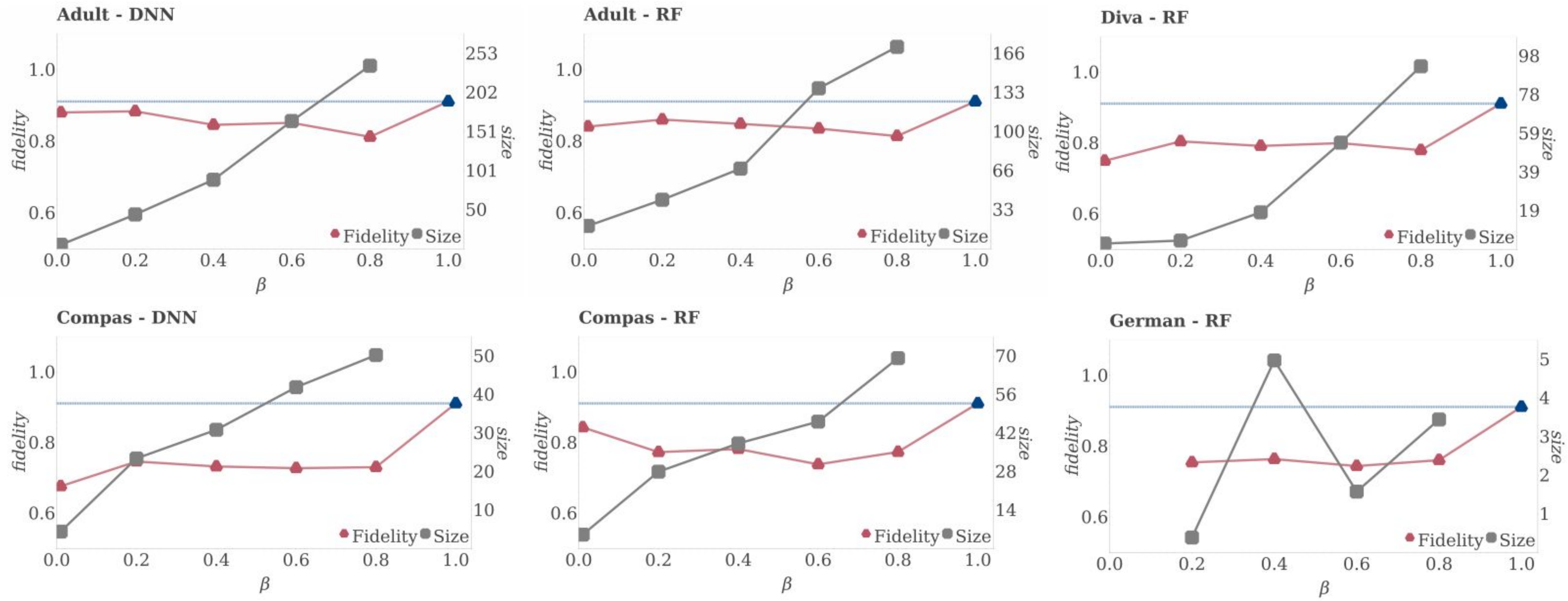
Validation setting

- 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)



How many local 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.

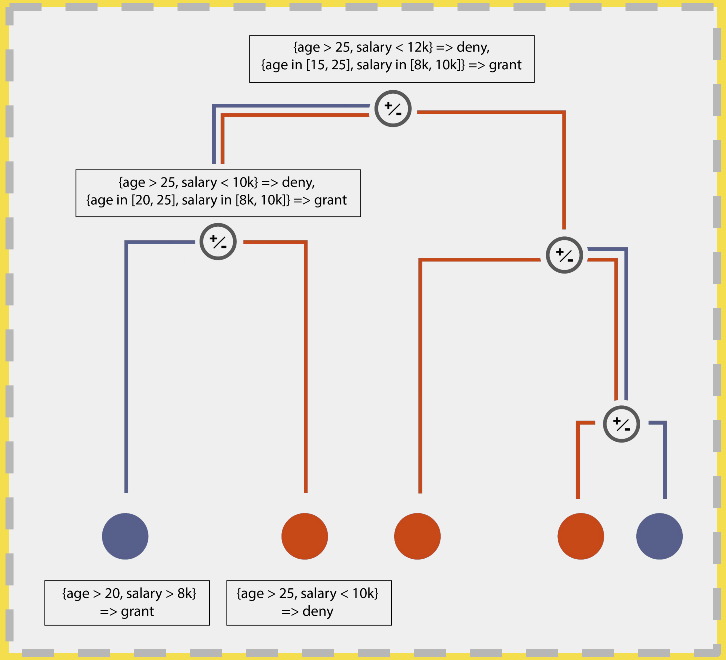
α-percentile	Fidelity	Size	Length
75	83.0 ± 3.6	31.0 ± 19.4	5.36 ± 2.41
90	84.7 ± 5.14	11.5 ± 6.4	5.43 ± 2.46
95	84.5 ± 5.48	6.625 ± 2.9	5.17 ± 2.59
99	84.0 ± 5.0	3.625 ± 2.6	5.97 ± 3.04

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

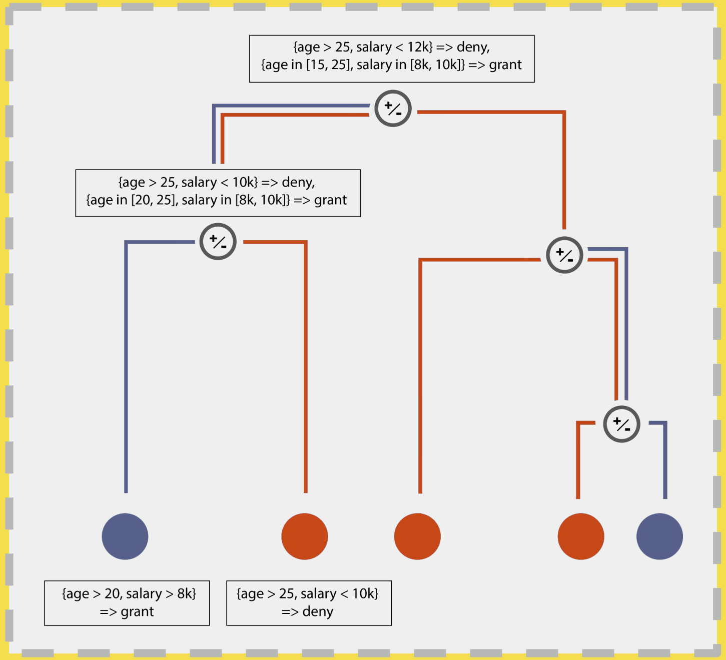
- Local to Global explanation paradigm
- Explaining globally by explaining locally
- Explanation cost: how many explanations do we really need?



github.com/msetzu/glocalx mattia.setzu@phd.unipi.it

Extensions and future work

- Local to Global in other domains
  - text
  - images
  - oblique rules



github.com/msetzu/glocalx mattia.setzu@phd.unipi.it

A plethora of challenges:

- [text] sparsity, merging tokens/text, few (if any) global families;
- [images] highly complex and entangled latent space.

Merge as subsumption?

May remind you of θ-subsumption in ILP[5]. In a LFE setting:

- [join] generalization as entailment (local entails global)
- [cut] specialization as inverse entailment (global entails local)

Why not apply classic LFE learning?

- lack of variables (what to substitute?);
- lattice already implicit in the polyhedral interpretation;
- practically: very few merges, less accurate models;

[5] Automatic Methods of Inductive Inference, Plotkin

Piggybacking again on ILP: background knowledge injection and predicate invention

- can generalize premises to domain-specific concepts
- can use more principled similarity measures
- invent symbols for common clauses (premises)

[5] Automatic Methods of Inductive Inference, Plotkin

Local to sub-global

Locality (globality) is a continuum!

Explain different (possibly related) groups/clusters, e.g.

- medical AI on white/black or young/old patients[7]
- AI judge on white/black defendants[8]

[6] Interpretable Decision Sets: A Joint Framework for Description and Prediction, Lakkaraju et al.
[7] FairLens: Auditing black-box clinical decision support systems, Panigutti et al.
[8] https://github.com/propublica/compas-analysis