One Tree to Explain Them All: Importance-Aware Rule Models for Black-Box Systems

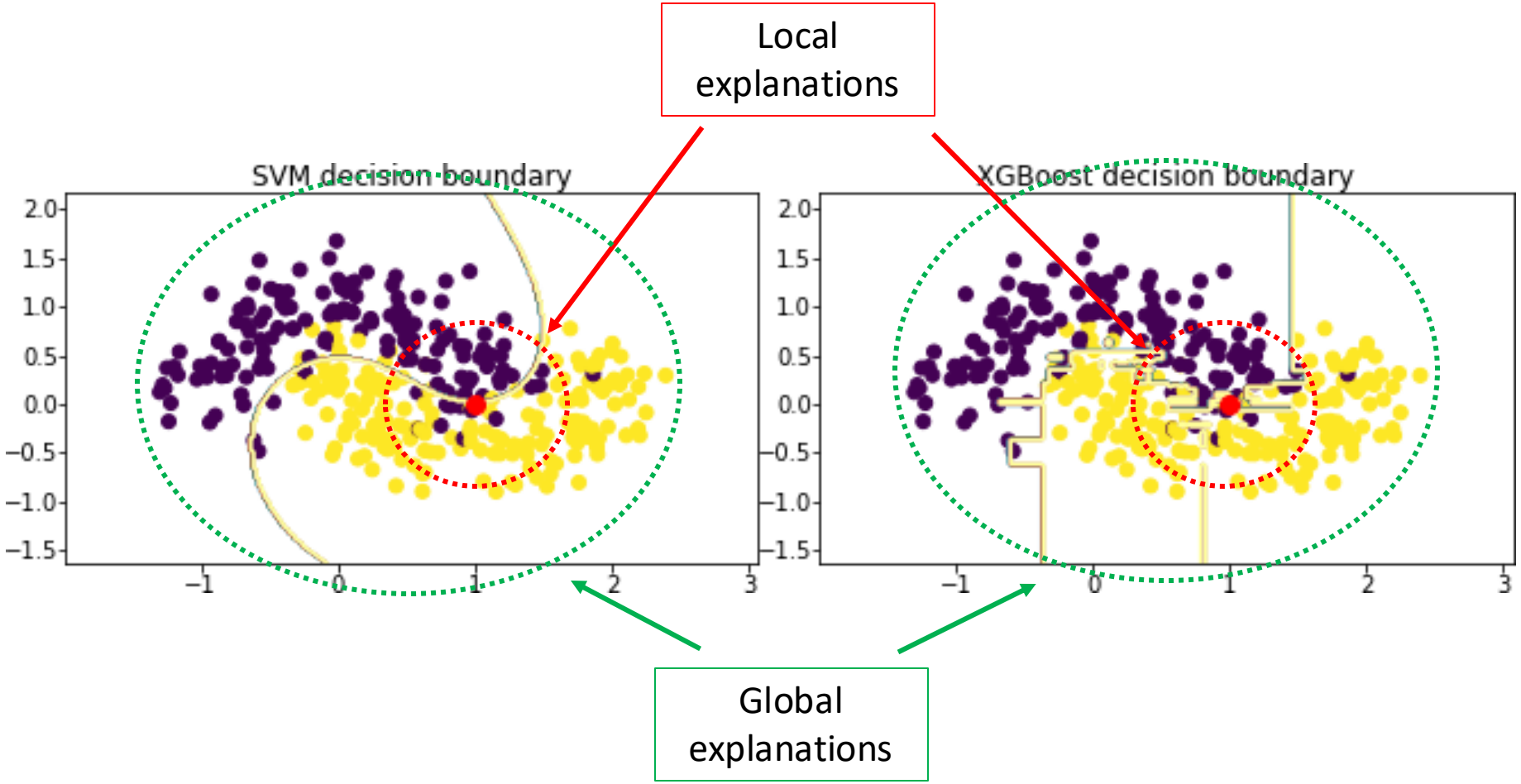
Szymon Bobek

RuleML Webinar, November 2025

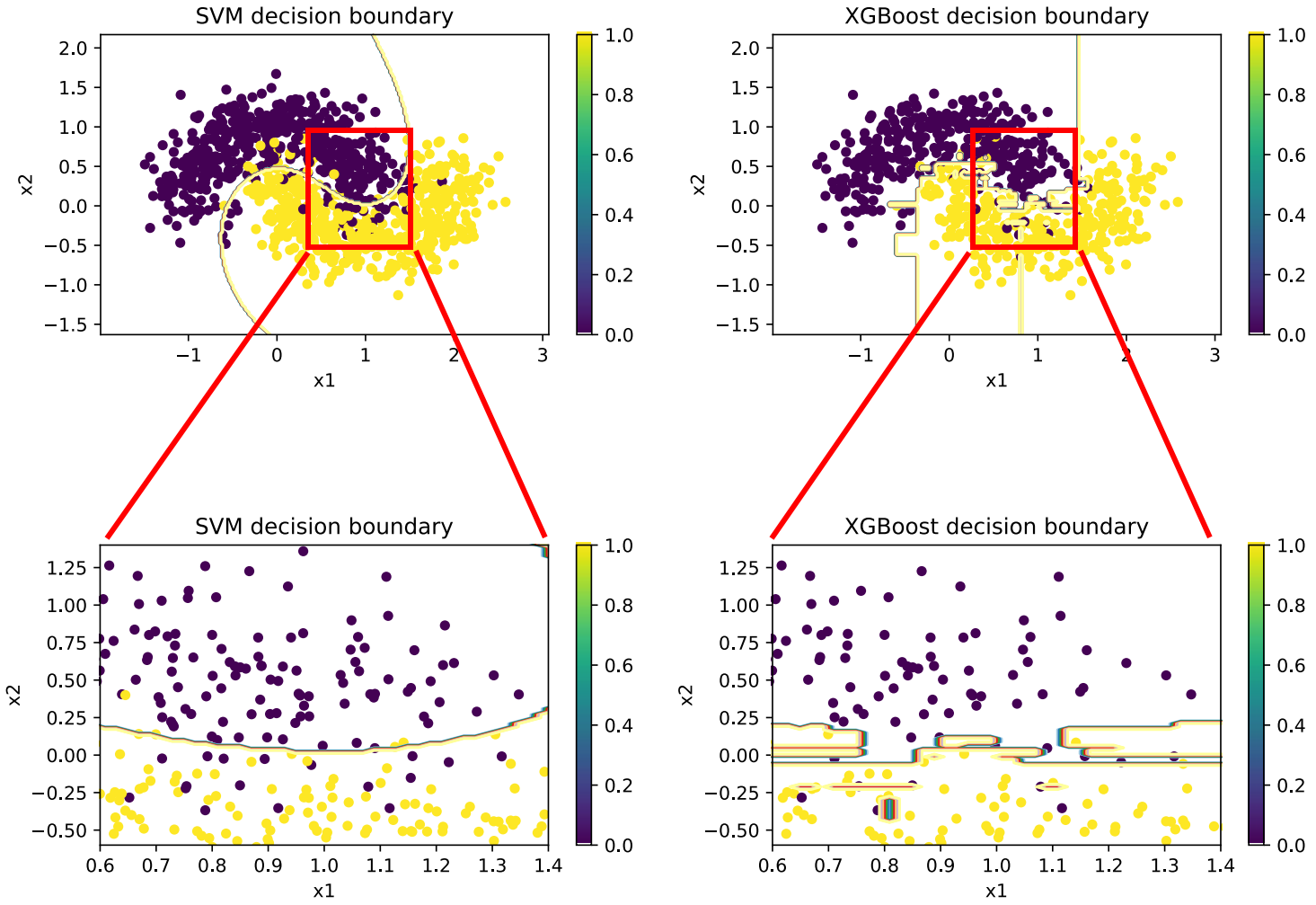
https://geist.re

The presentation covers several different works founded from the XPM (NCN UMO-2020/02/Y/ST6/00070) project funded by the National Science Centre, Poland under CHIST-ERA program .

Local vs Global explanations

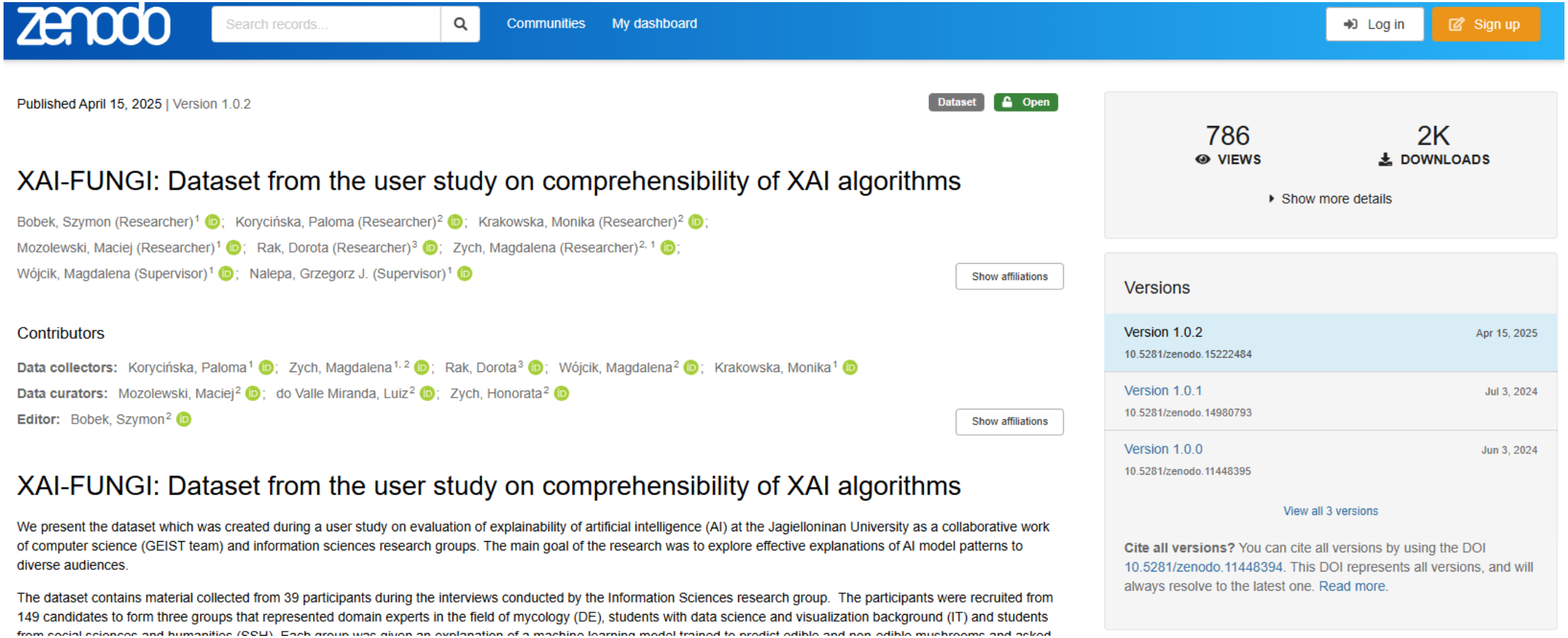


Locally, the decision boundary is simpler



- In this approach we focus on explaining an instance
- "Zooming in" we can fit inherently interpretable model that will approximate the decision of the blackbox one
- The assumption is not always valid. There are models which have complex decision boundary even locally
- Term "Locally" is vague. The locality is subjective
- When zooming in, we are limiting the number of samples that can be used for training
- **Every ML model can be local surrogate**

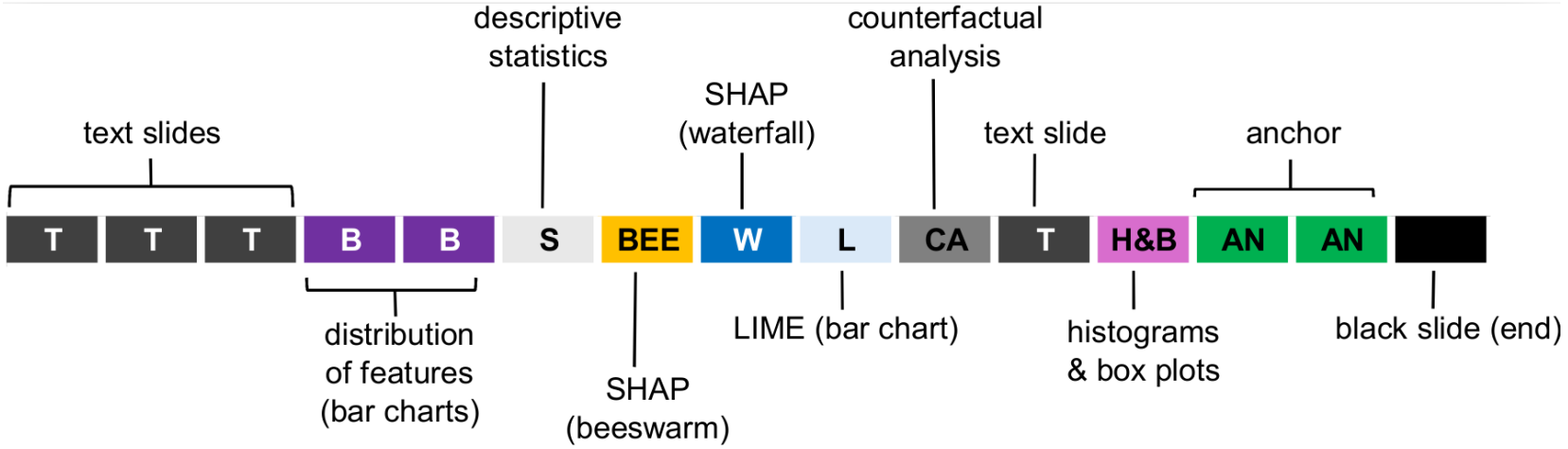
XAI-FUNGI: Dataset from the user study on comprehensibility of XAI algorithms

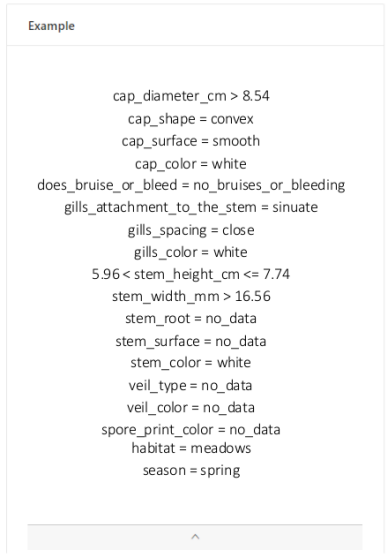


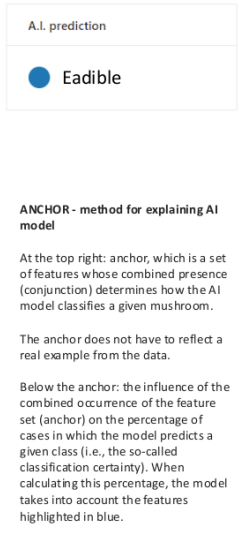
https://zenodo.org/records/15222484

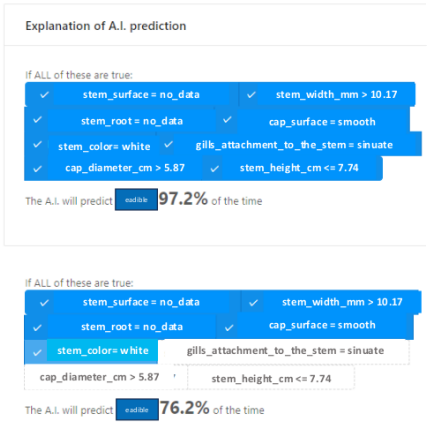
Rules to rule them all

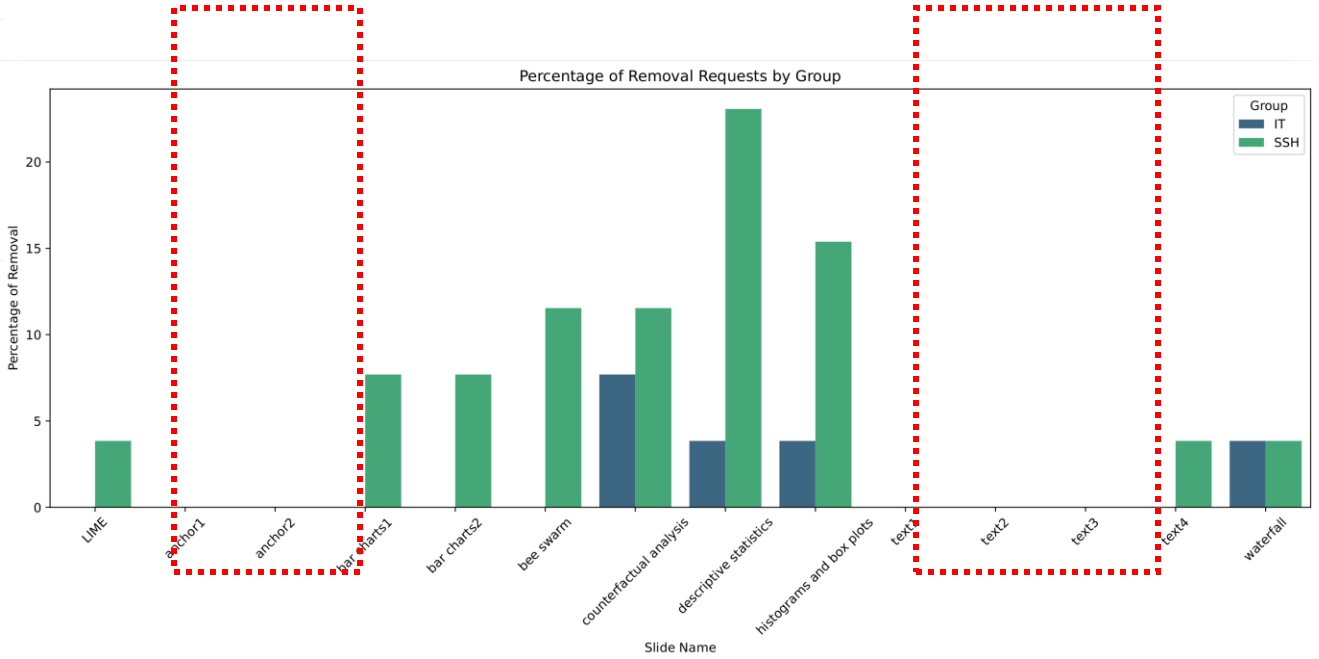
DOI 10.5281/zenodo.15222484





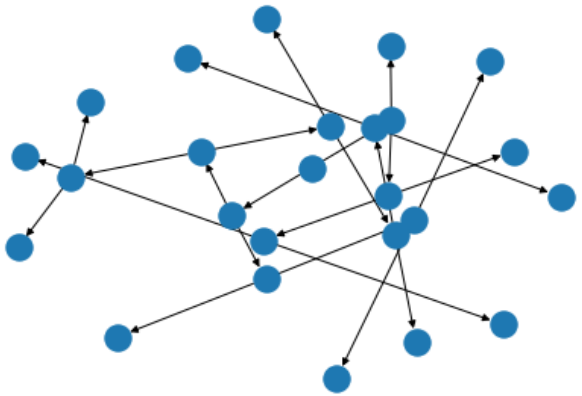


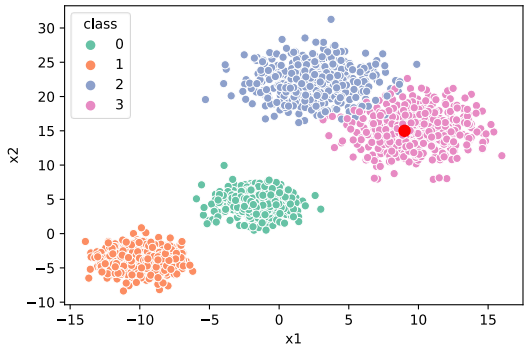


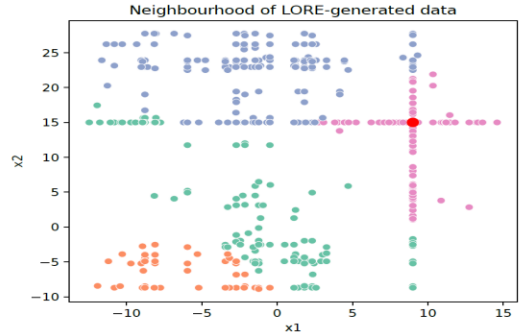


LORE: Dataset generation and explanation creation

```
([
  {'class': 3},
  { 'x1': '>6.532643',
    'x2': '-0.82784< x2 <=20.426761'}
 ],
 [
  {'x2': '<=-0.82784'}
 ])
```







- Generate samples by maximizing following fitness functions:

$$fitness^{x}_{=}(z) = I_{b(x)=b(z)} + (1 - d(x,z)) - I_{x=z}$$

$$fitness^{x}_{\neq}(z) = I_{b(x)\neq b(z)} + (1 - d(x,z)) - I_{x=z}$$

- Fitness funcitons determine survivors and generaiton performs (2-point) crossover and mutation scheme:

parent 1	25	clerk	10k	yes
parent 2	30	other	5k	no
children 1	25	other	5k	yes
children 2	30	clerk	10k	no

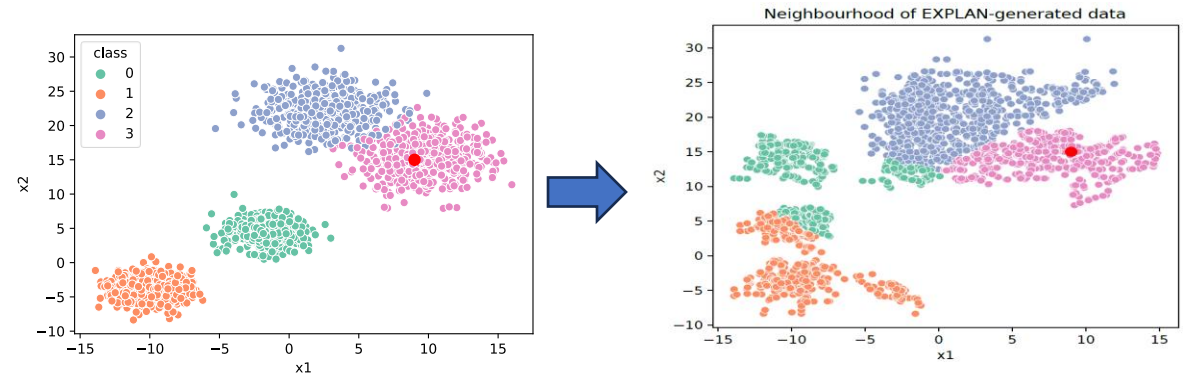
parent	25	clerk	10k	yes
children	27	clerk	7k	yes

Replacing (mutation) according to empirical distribution of a feature

- The resulting dataset is balanced and completely artificial

EXPLAN: Explanation creation

```
[
  {'class': 3},
  {'x1': '>5.384601', 'x2': '<=18.633979'}
]
```




```
1: {'f_6': ['<=0.051216'], 'f_1': ['>-0.118408']}
2: {'f_6': ['<=0.034339'], 'f_3': ['<=0.231442'],
    'f_1': ['>-0.387023']}
3: {'f_6': ['<=-0.156712'], 'f_3': ['<=0.457593'],
    'f_1': ['>-0.619056']}
4: {'f_1': ['>-0.052240'], 'f_6': ['<=-0.101768']}
5: {'f_6': ['<=0.075657'], 'f_7': ['<=1.201282']}
```

EXPLAN

```
1: {'f_6': ['<=-1.20169']}
2: {'f_6': ['<=0.152367'], 'f_3': ['<=1.206294']}
3: {'f_6': ['<=0.358263'], 'f_1': ['>-0.691120']}
4: {'f_6': ['<=-0.852889']}
5: {'f_6': ['<=-1.322858']
```

LORE

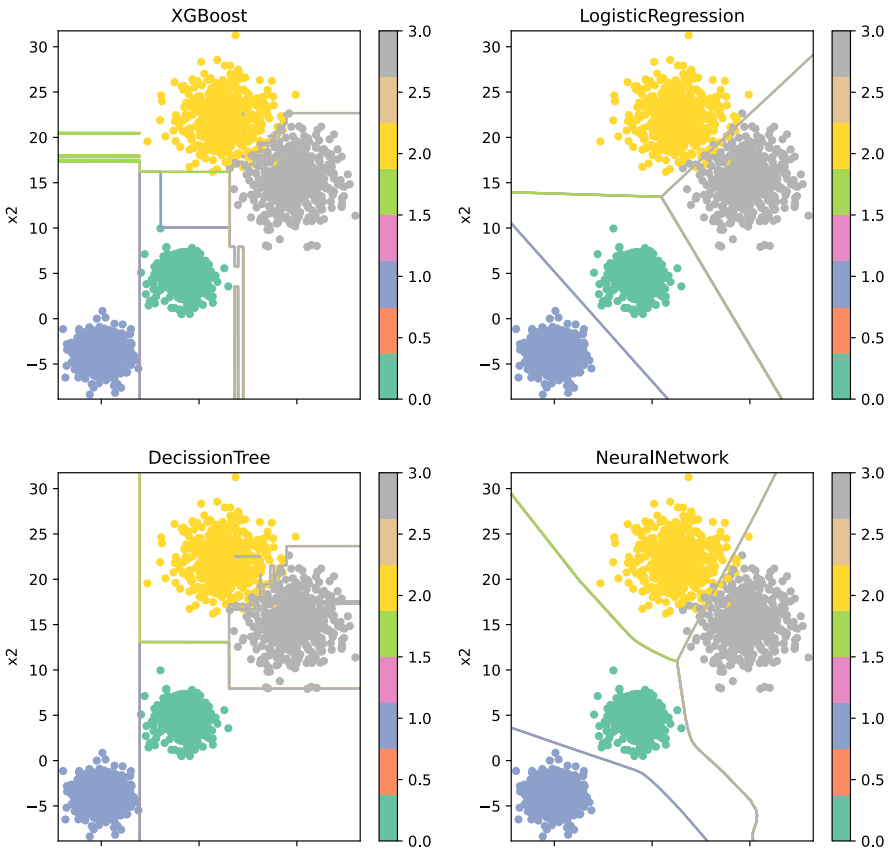
- The explanation is created using the same decision tree algorithm as LORE
- Generation with Importance-similarity perturbation, aglomerative clustering and SMOTE rebalancing
- Suprisingly the implementaiton of EXPLAN does not provide counterfactuals, which could be extracted
- The generation process may result in different explanations for the same instance

State of the art methods

	Factual	Counterfactual	Visual	Example-based	Deterministic	Synthetic data
LORE	Yes	Yes	No*	No*	No	Only
EXPLAN	Yes	No*	No*	No*	No	Only
Anchor	Yes	No	No	Yes	No	Mainly

- They focus on one aspect of explanation, not the whole spectrum
- They use synthetic data, which is perfectly fine for ML engineer, but might be confusing for domain knowledge/user
- Consecutive runs yield different results
- Lack compelling visual representation

Surrogate models are prone to Rashomon effect

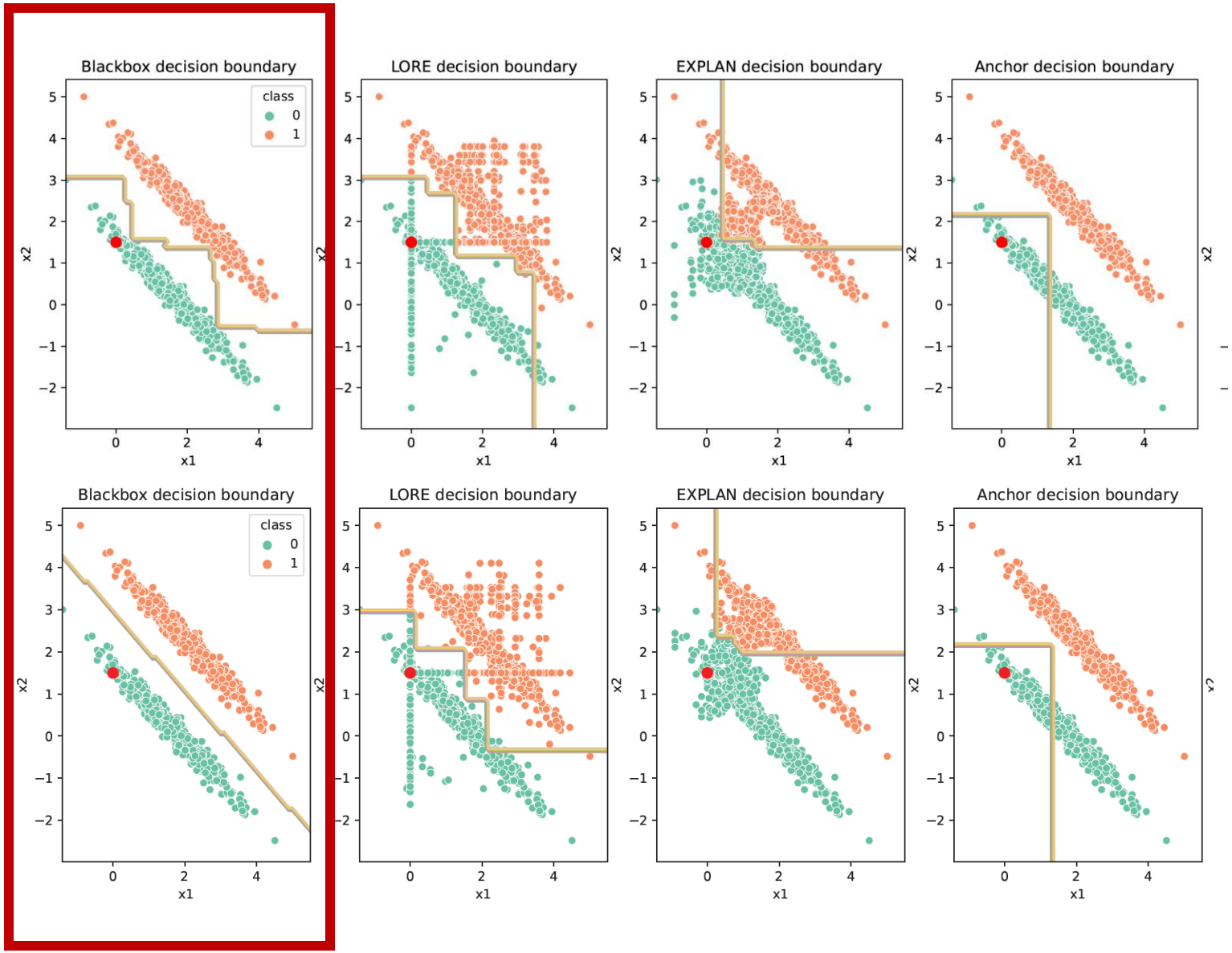


- Many models may be "right" but use very different methods to derive the "right"
- In Explainability we care about how the "right" is derived
- In such a case the more Rashomon effect the more doomed we are

Rashomon (1950) by Akira Kurosawa

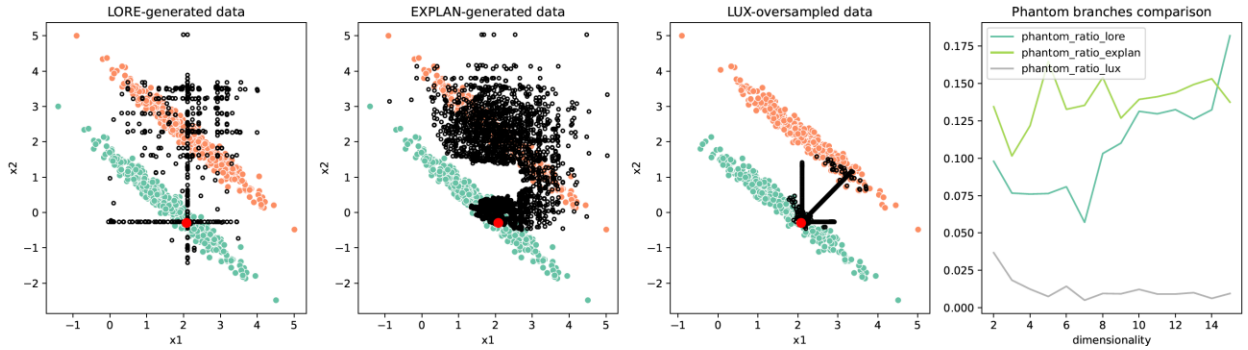
Surrogate models are prone to Rashomon effect

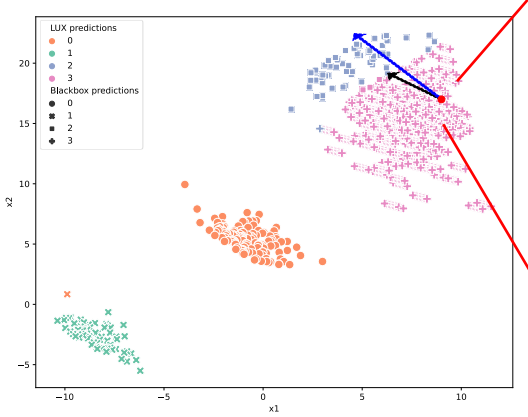
Rashomon (1950) by Akira Kurosawa



LUX: Local Universal Rule-based Explainer

Local Universal eXplainer

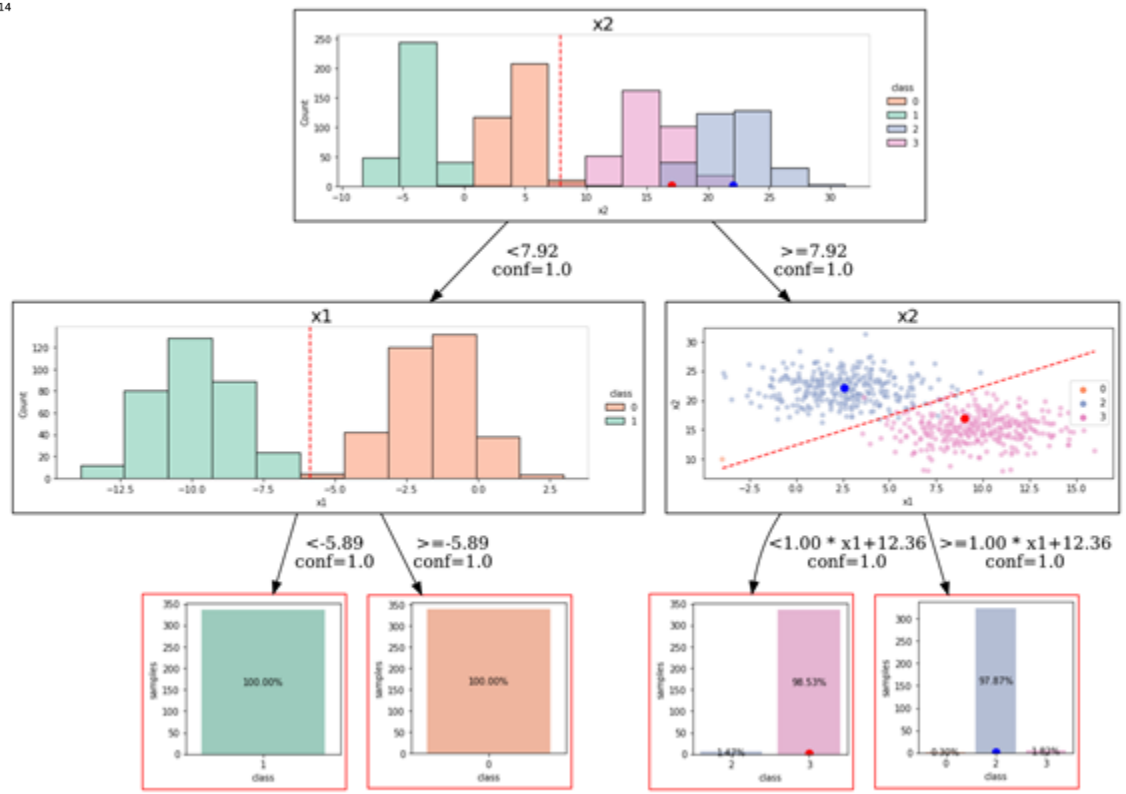




Gradients of SHAP points towards decision boundary in "important" directions.

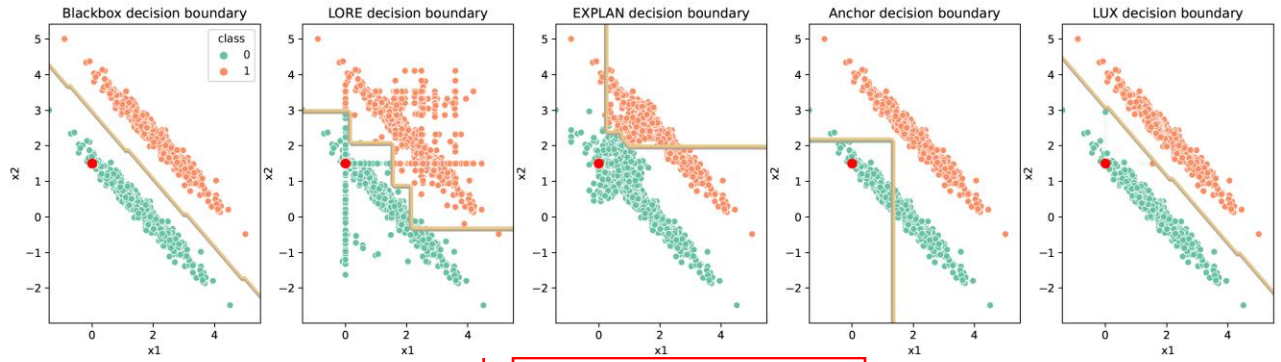
Countercastuals can be extracted by traversing a decision tree in local neighbourhood of an instance

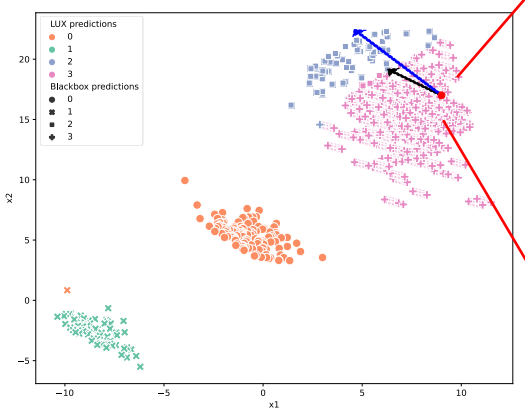
- Neighbourhood creation from original data
- Minimal artificial data generation (along SHAP-based gradients)
- Counterfactuals extraction
- Visualization
- Oblique splits



Bobek, S., & Nalepa, G. J. (2023). **Local Universal Explainer (LUX) -- a rule-based explainer with factual, counterfactual and visual explanations**. *arXiv [Cs.AI]*. Retrieved from http://arxiv.org/abs/2310.14894

Local Universal eXplainer

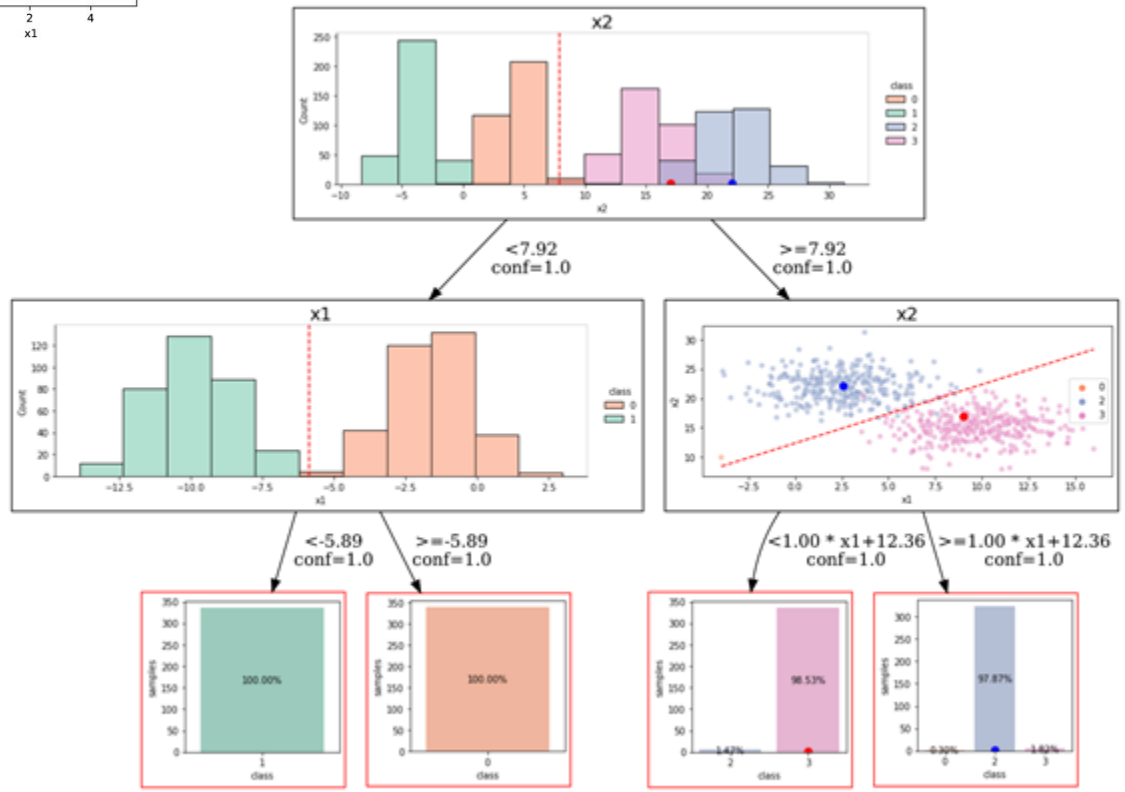




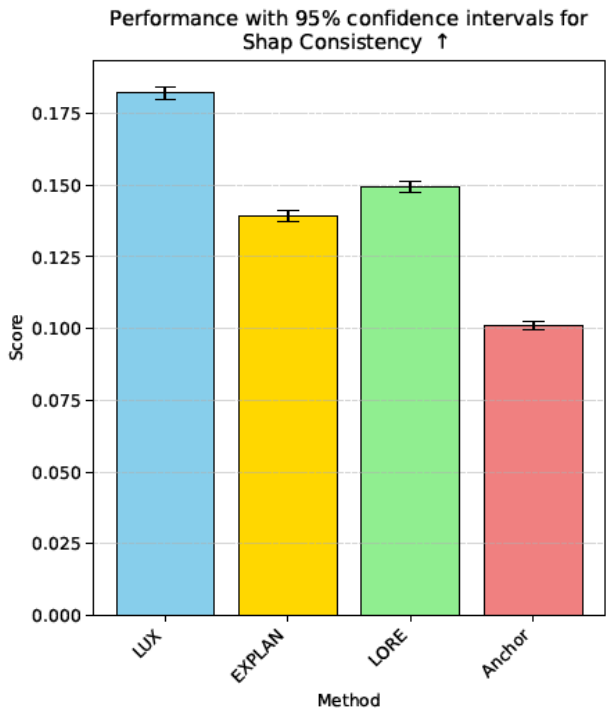
Gradients of SHAP points towards decision boundary in "important" directions.

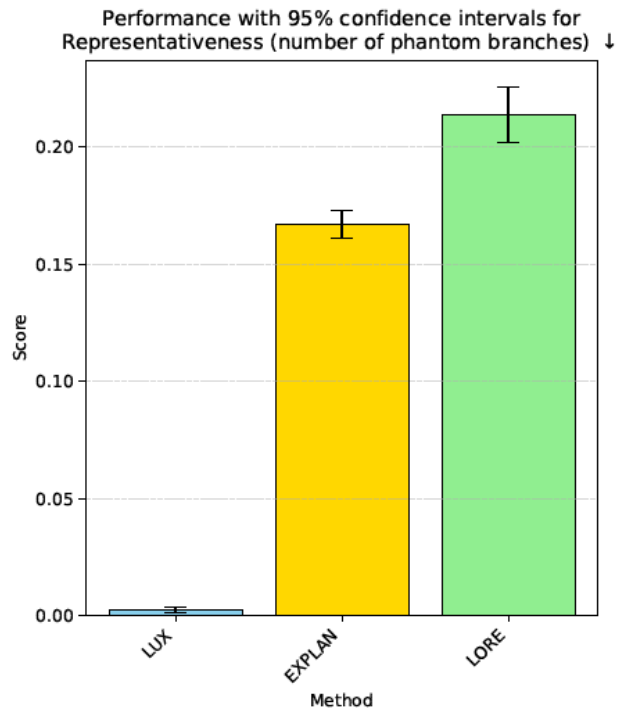
Countercastuals can be extracted by traversing a decision tree in local neighbourhood of an instance

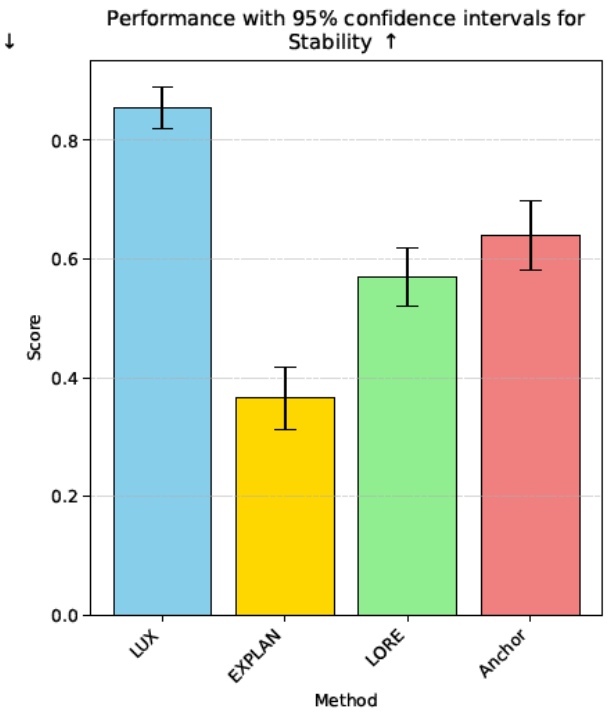
- Neighbourhood creation from original data
- Minimal artificial data generation (along SHAP-based gradients)
- Counterfactuals extraction
- Visualization
- Oblique splits

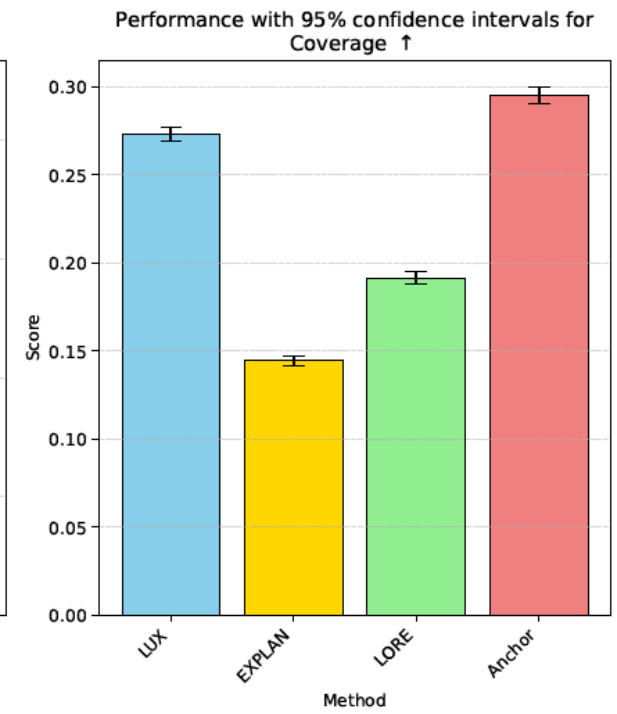


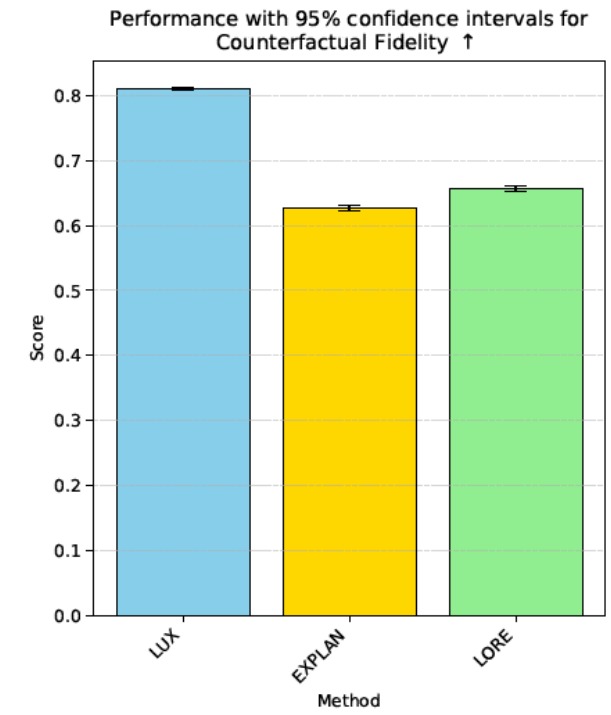
Bobek, S., & Nalepa, G. J. (2023). **Local Universal Explainer (LUX) -- a rule-based explainer with factual, counterfactual and visual explanations**. *arXiv [Cs.AI]*. Retrieved from http://arxiv.org/abs/2310.14894

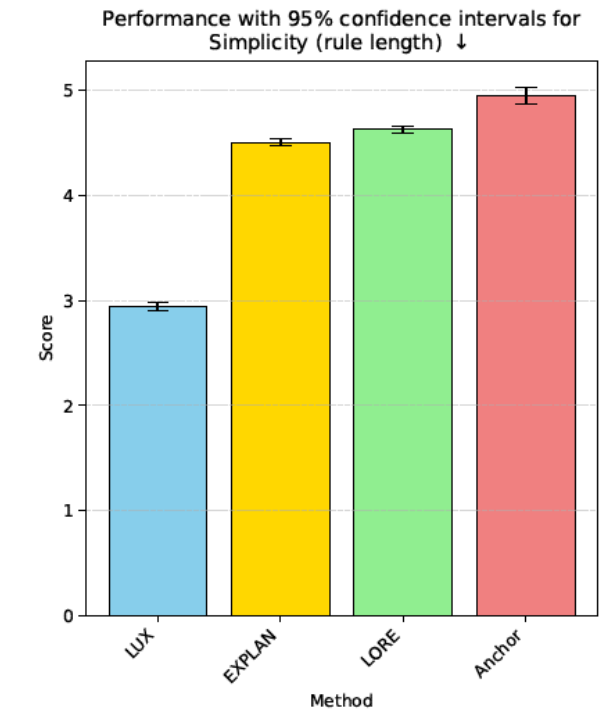

Performance with 95% confidence intervals for
Shap Consistency ↑
Score
Method
LUX
EXPLAN
LORE
Anchor

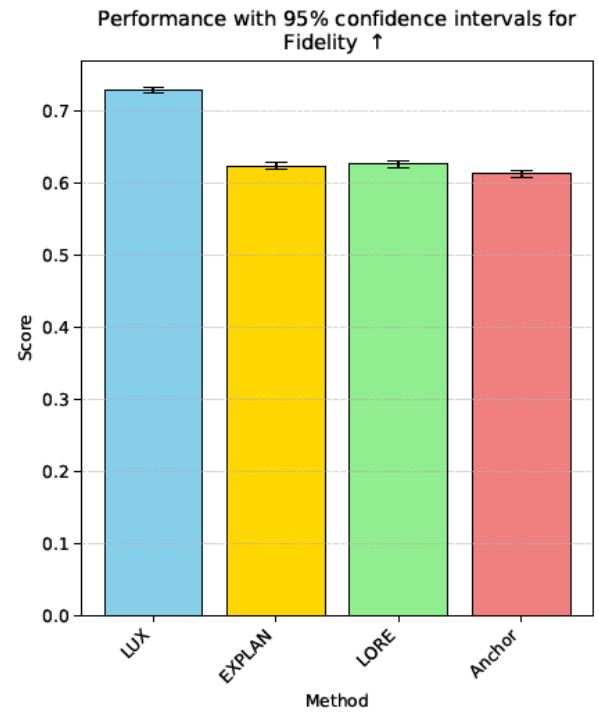

Performance with 95% confidence intervals for
Representativeness (number of phantom branches) ↓
Score
Method
LUX
EXPLAN
LORE

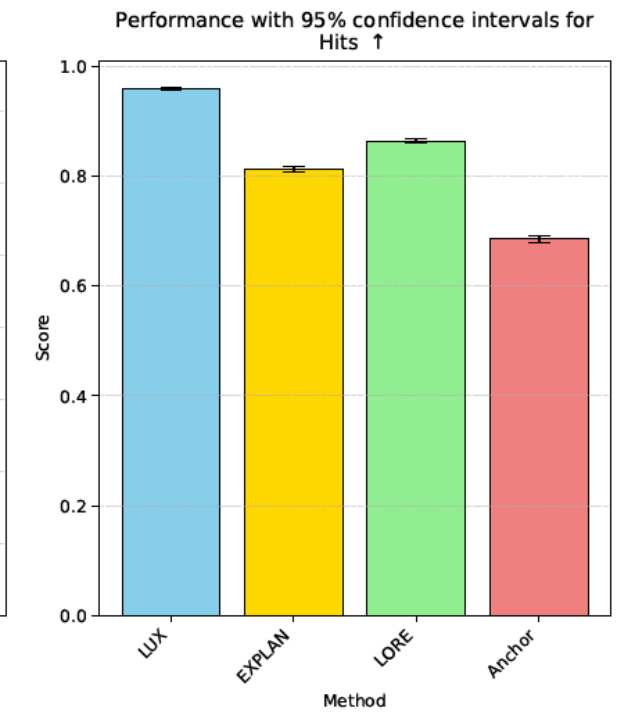

Performance with 95% confidence intervals for
Stability ↑
Score
Method
LUX
EXPLAN
LORE
Anchor


Performance with 95% confidence intervals for
Coverage ↑
Score
Method
LUX
EXPLAN
LORE
Anchor


Performance with 95% confidence intervals for
Counterfactual Fidelity ↑
Score
Method
LUX
EXPLAN
LORE


Performance with 95% confidence intervals for
Simplicity (rule length) ↓
Score
Method
LUX
EXPLAN
LORE
Anchor


Performance with 95% confidence intervals for
Fidelity ↑
Score
Method
LUX
EXPLAN
LORE
Anchor


Performance with 95% confidence intervals for
Hits ↑
Score
Method
LUX
EXPLAN
LORE
Anchor

LUX is OpenSource

Bobek, S.; Nalepa, G. **J. Local Universal Rule-Based eXplainer (LUX)**. SoftwareX 2025, 30, 102102. https://doi.org/10.1016/j.softx.2025.102102.

Welcome to the LUX documentation

LUX (Local Universal Rule-Based Explainer) is an XAI algorithm that produces explanations for any type of machine-learning model. It provides local explanations in a form of human-readable (and executable) rules, but also provide counterfactual explanations as well as visualization of the explanations.

Install

LUX can be installed from either PyPI or directly from source code GitHub

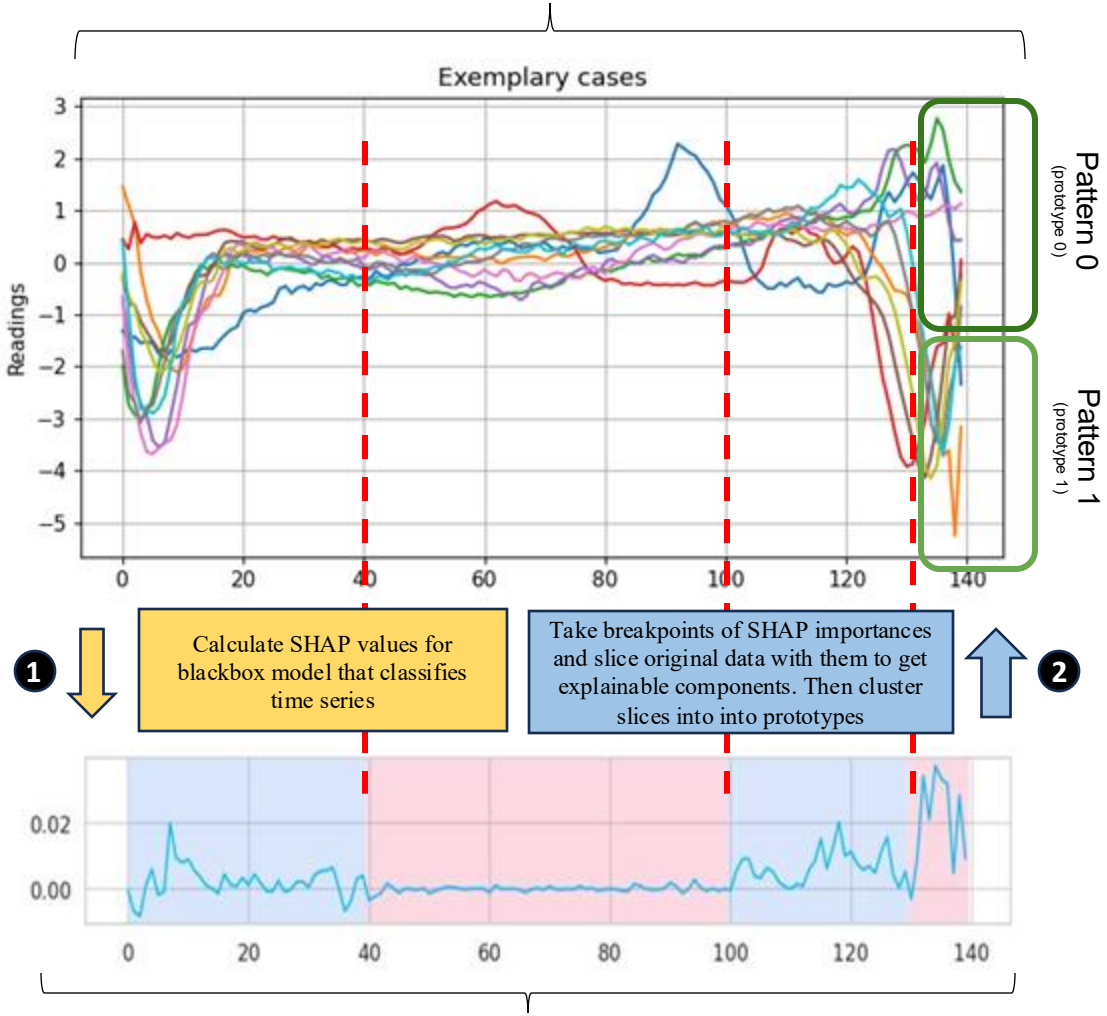
To install form PyPI:

```
pip install lux-explainer
```

TSProto: Rule-based explanations for time-series

Explainable components and prototypes discovery

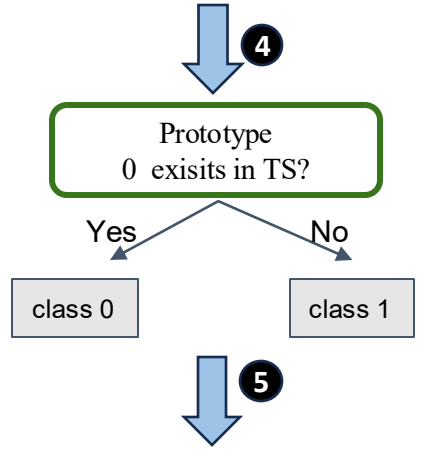
The input is a dataset and a blackbox model that classifies time series



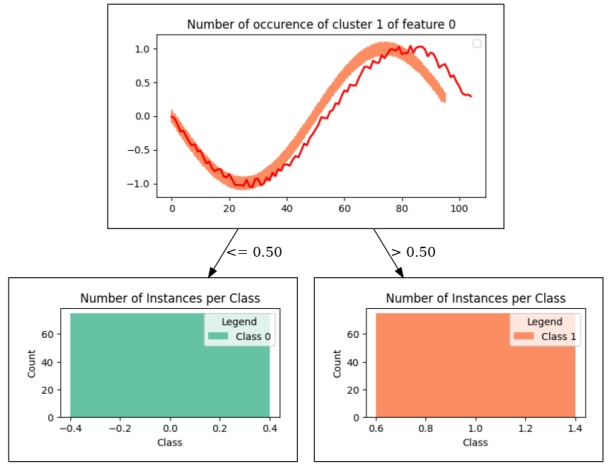
The ruptures changepoint detection is used to find regions of similar importance in time series. Similar importance form dense regions that are considered interpretable components.

Context aggregation

OHE streams with respect to presence/absence of particular pattern (prototype), enrich with statistics & build interpretable global model



Encrich explanations with visual representation of prototypes and interpretable components

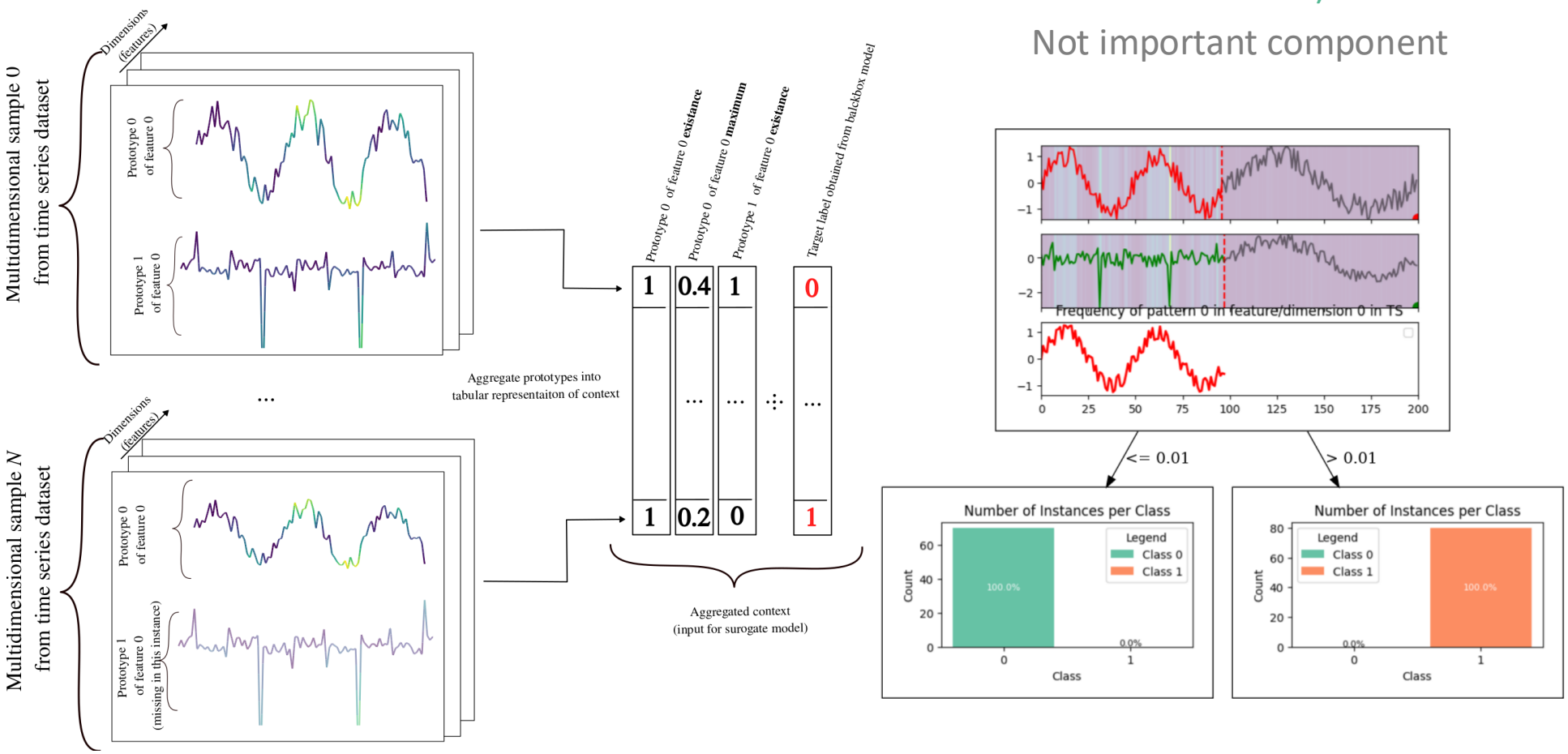


Context aggregation

Prototype visualization/Important part similar to prototype (barycenter of cluster)

Contrastive exmaple (most similar to prototype, but different class)

Not important component



Szymon Bobek and Grzegorz J. Nalepa. 2025. **TSProto: Fusing deep feature extraction with interpretable glass-box surrogate model for explainable time-series classification**. Inf. Fusion 124, C (Dec 2025). https://doi.org/10.1016/j.inffus.2025.103357

TSProto is OpenSource

tsproto

Search docs

EXAMPLES

Interpreting TSProto

Basic Usage examples

REFERENCE

API reference

DEVELOPMENT

Release notes

Contributing guide

Welcome to the TSProto documentation

View page source

Welcome to the TSProto documentation

TSProto (Post-host prototype-based explanations with rules for time-series classifiers) is an XAI algorithm that produces explanations for any type of machine-learning model. It provides local explanations in a form of human-readable (and executable) rules, but also provide counterfactual explanations as well as visualization of the explanations.

Install

TSProto can be installed from either PyPI or directly from source code GitHub

To install form PyPI:

```
pip install tsproto
```

People

GEIST
Research Group

Prof. Grzegorz
J. Nalepa, PhD. Eng.
GEIST Leader













Thank you for your attention!

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re

References

- Bobek, S.; Nalepa, G. J. **Local Universal Explainer (LUX) -- a Rule-Based Explainer with Factual, Counterfactual and Visual Explanations**. arXiv February 9, 2024. https://doi.org/10.48550/arXiv.2310.14894.
- Bobek, S.; Nalepa, G. J. **Local Universal Rule-Based eXplainer (LUX)**. SoftwareX 2025, 30, 102102. https://doi.org/10.1016/j.softx.2025.102102.
- Bobek, S.; Nalepa, G. J. **TSProto: Fusing deep feature extraction with interpretable glass-box surrogate model for explainable time-series classification.** Inf. Fusion 124, C (Dec 2025). https://doi.org/10.1016/j.inffus.2025.103357
- Bobek, S.; Korycińska, P.; Krakowska, M.; Mozolewski, M.; Rak, D.; Zych, M.; Wójcik, M.; Nalepa, G. J. **User-Centric Evaluation of Explainability of AI with and for Humans: A Comprehensive Empirical Study**. International Journal of Human-Computer Studies 2025, 205, 103625. https://doi.org/10.1016/j.ijhcs.2025.103625.
- Bobek, S., Korycińska, P., Krakowska, M. et al. **Dataset resulting from the user study on comprehensibility of explainable AI algorithms**. Sci Data 12, 1000 (2025). https://doi.org/10.1038/s41597-025-05167-6