Explanation Methods for Sequential Data Models

Francesco Spinnato

francesco.spinnato@di.unipi.it
University of Pisa

Tales on Data Science and Big Data

What is Sequential Data?

Type of data where the **order of information** matters.

TEXT

The **cat** *eats the* **mouse**

The **mouse** *eats the* **cat**

TIME SERIES

TRAJECTORIES

What are Sequential Data Models?

Models that take as input **sequential datasets**. Here we focus on **supervised learning**, in particular classification and regression.

CLASSIFICATION

"Is the patient healthy?"

"YES"

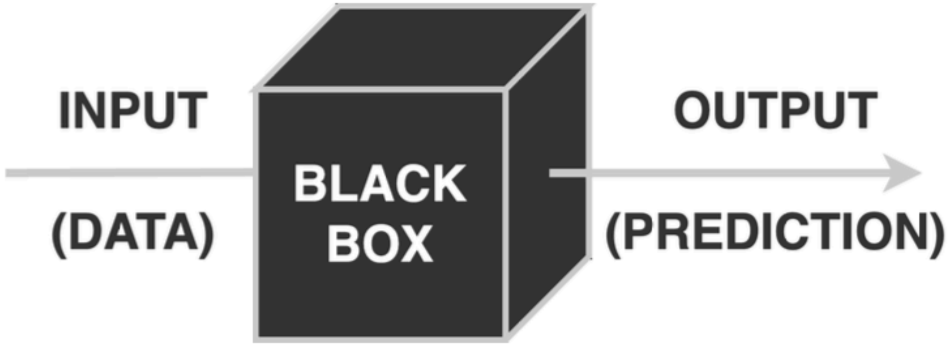
REGRESSION

"How old is the patient?"

"30"

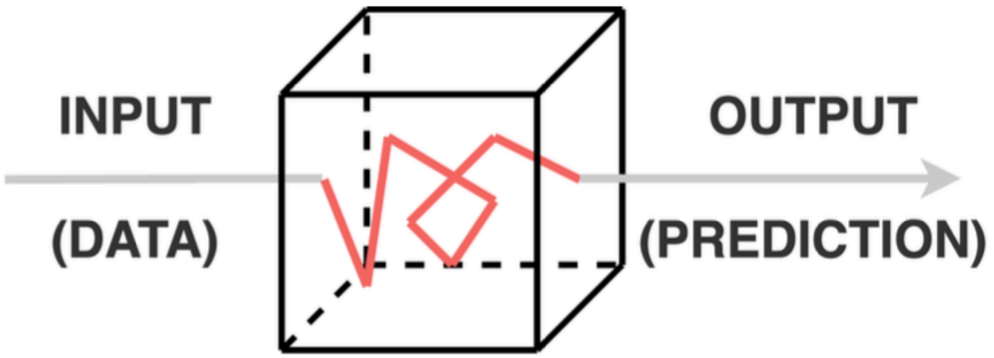
Why Explanations for Sequential Data Models?

The best machine learning models for sequential data are powerful but opaque: they are **black-boxes**!



Why Explanations for Sequential Data Models?

Explainable AI (XAI) is the branch of AI that tries to *open these black-boxes*, to understand the **relationship between input and output**.



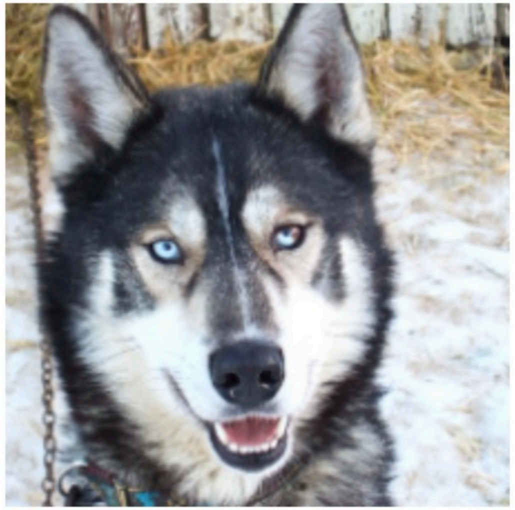
Why Explanations for Sequential Data Models?

- **Legally: GDPR, AI Act***
- **Trust and fairness in AI systems**
- Debugging and improving the model
- Understanding of the model's decision process

- European Parliament & Council of the EU. (2016). Regulation (EU) 2016/679 (GDPR).
- European Parliament & Council of the EU. (2024). Regulation (EU) 2024/1689 (AI Act).

Why Explanations for Sequential Data Models?

- **Debugging and improving the model**
- **Understanding of the model's decision process**



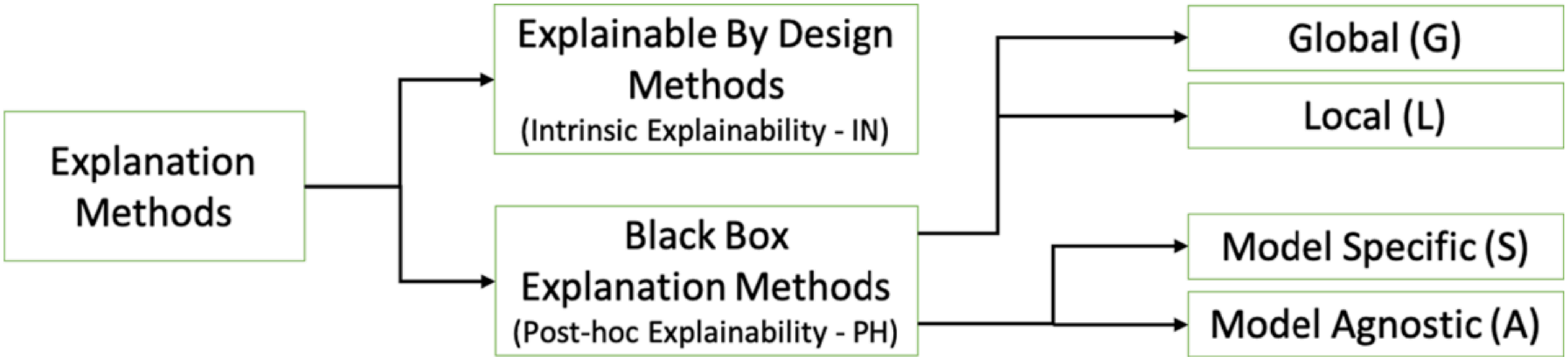
(a) Husky classified as wolf

(b) Explanation

Ribeiro, Marco Tulio et al. "Why should i trust you?" Explaining the predictions of any classifier." ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

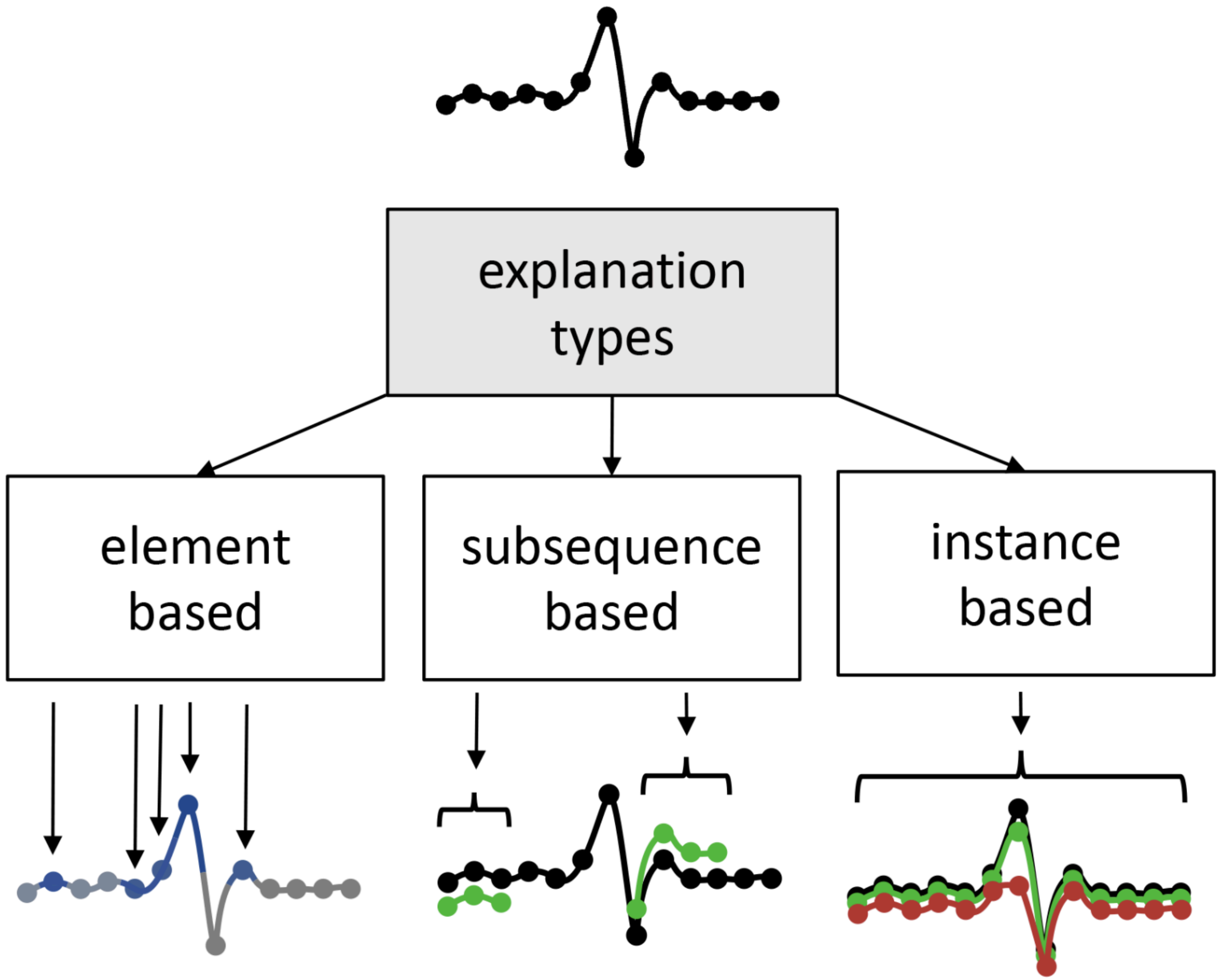
How can you obtain an explanation?

Explanations are obtained through **explainers**.



Bodria, Francesco, et al. "Benchmarking and survey of explanation methods for black box models." Data Mining and Knowledge Discovery 37.5 (2023): 1719-1778.

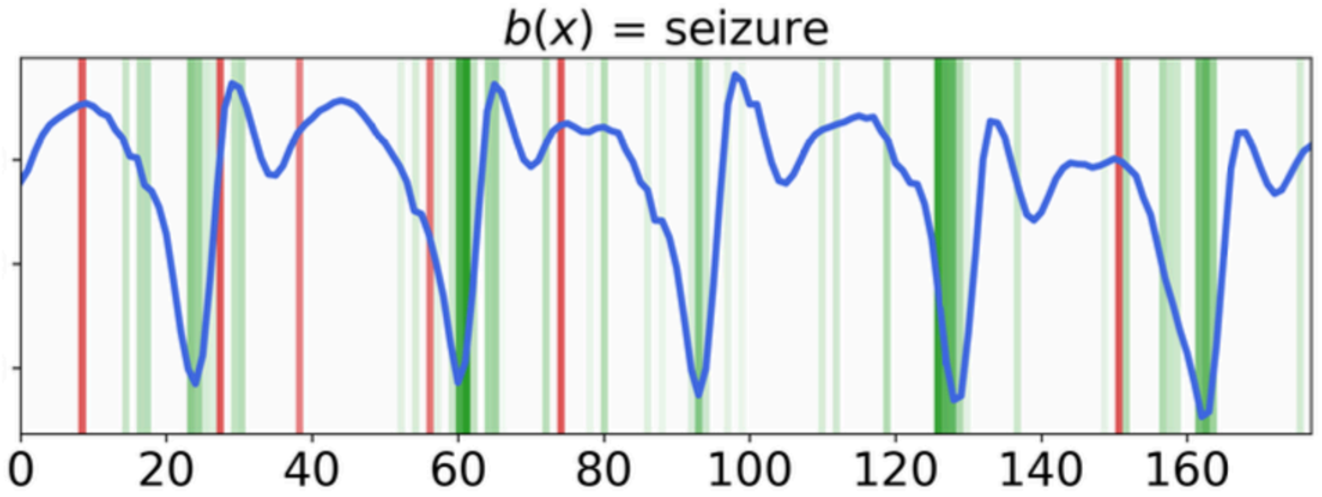
What does an explanation on sequential data look like?



Theissler, A., Spinnato, F., Schlegel, U., & Guidotti, R. (2022). Explainable AI for time series classification: a review, taxonomy and research directions. Ieee Access, 10, 100700-100724.

What does an explanation on sequential data look like?

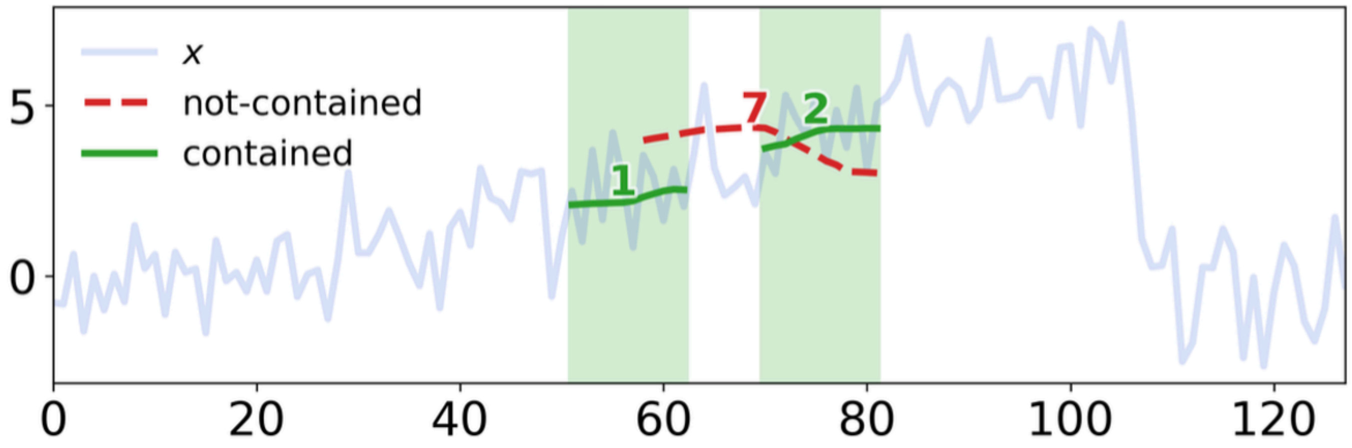
They can highlight the most important observations, using feature attribution methods, like **SHAP** (SHapley Additive exPlanations).



Scott M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NIPS 2017: 4765-4774

What does an explanation on sequential data look like?

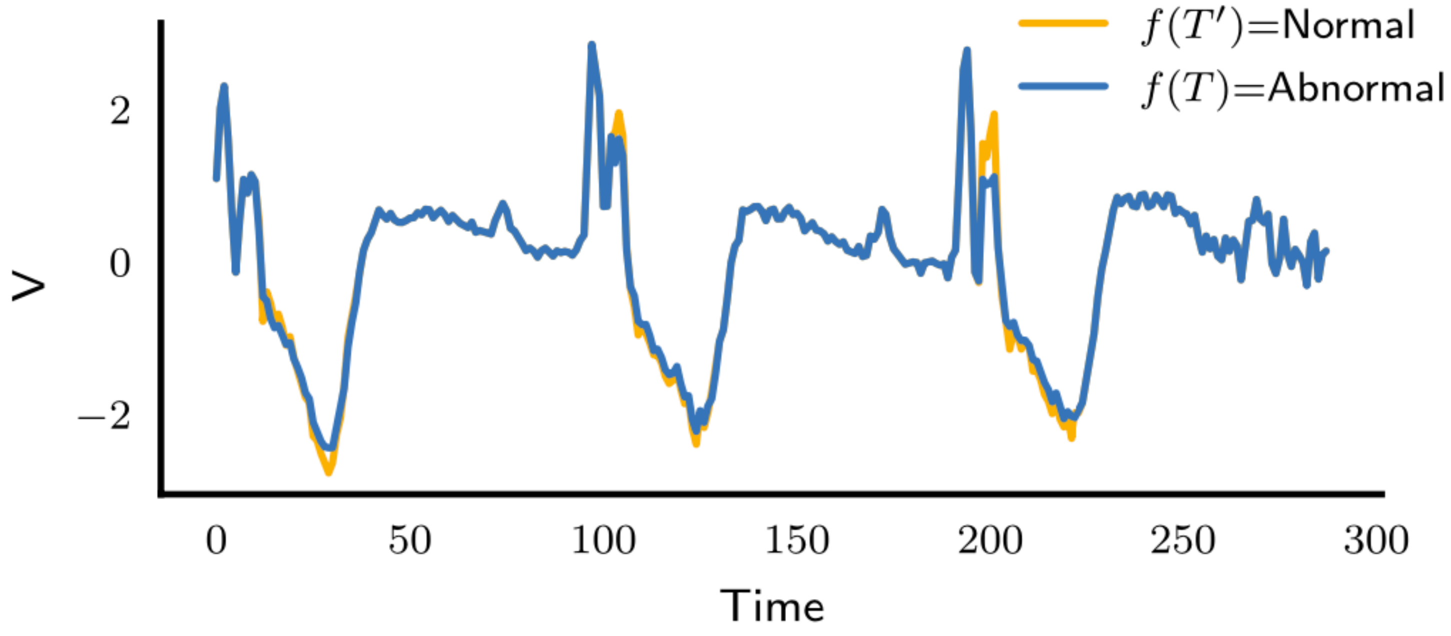
They can focus on subsets of observations, i.e., subsequences:



Guidotti, R., Monreale, A., Spinnato, F., Pedreschi, D., & Giannotti, F. (2020, October). Explaining any time series classifier. CogMI 2020 (pp. 167-176). IEEE.

What does an explanation on sequential data look like?

They can exploit entire instances, e.g., counterfactuals.



Karlsson, Isak, et al. "Locally and globally explainable time series tweaking." Knowledge and Information Systems 62.5 (2020): 1671-1700.

Challenges

Some challenges in XAI for sequential models are clear:

- most common explained task is **classification**;
- most common explained data are (univariate) **time series**;
- only **one kind of explanation** type;
- **lack of real applications** on complex datasets;
- implementation is **not standardized**.

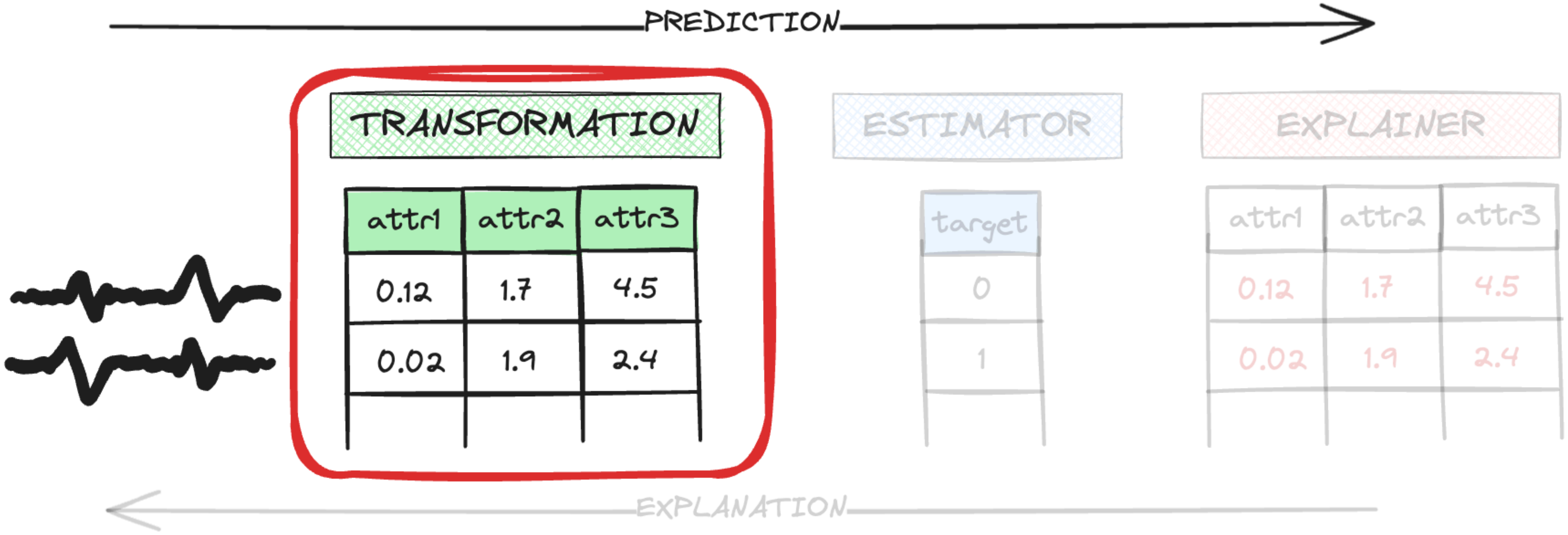
Background

Sequential Data Transformations

A Recipe for Interpretable Sequence Prediction (1)

To achieve interpretable sequence predictions you need:

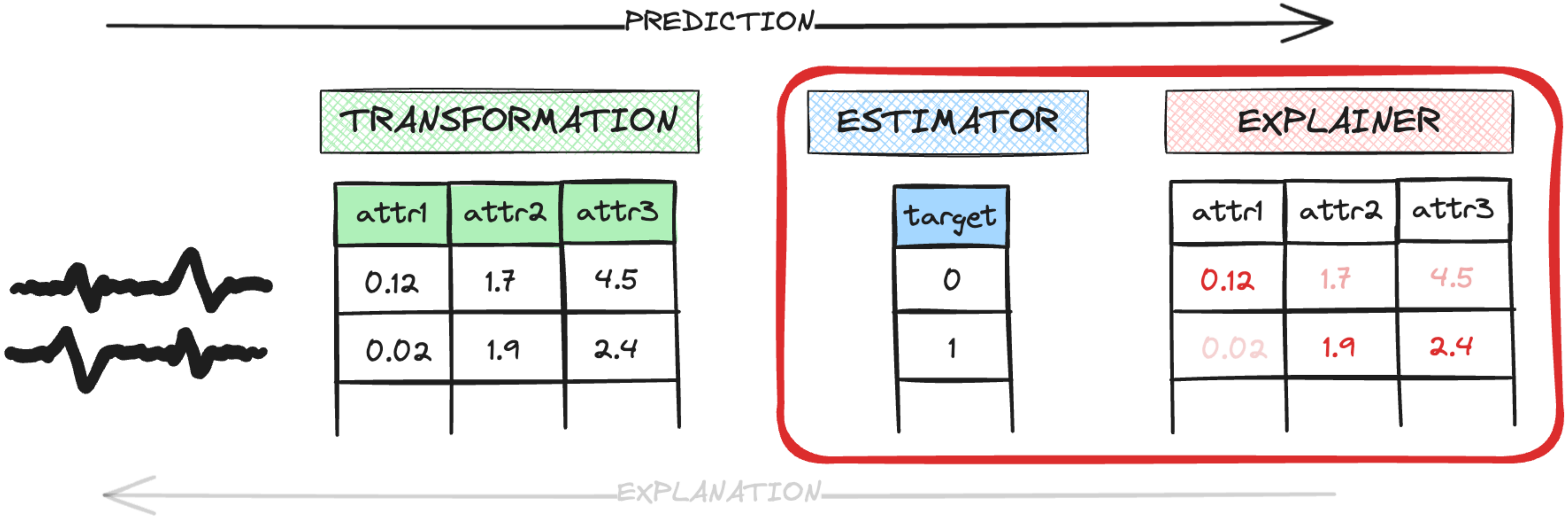
1. an **interpretable representation** for sequences;



A Recipe for Interpretable Sequence Prediction (2)

To achieve interpretable sequence predictions you need:

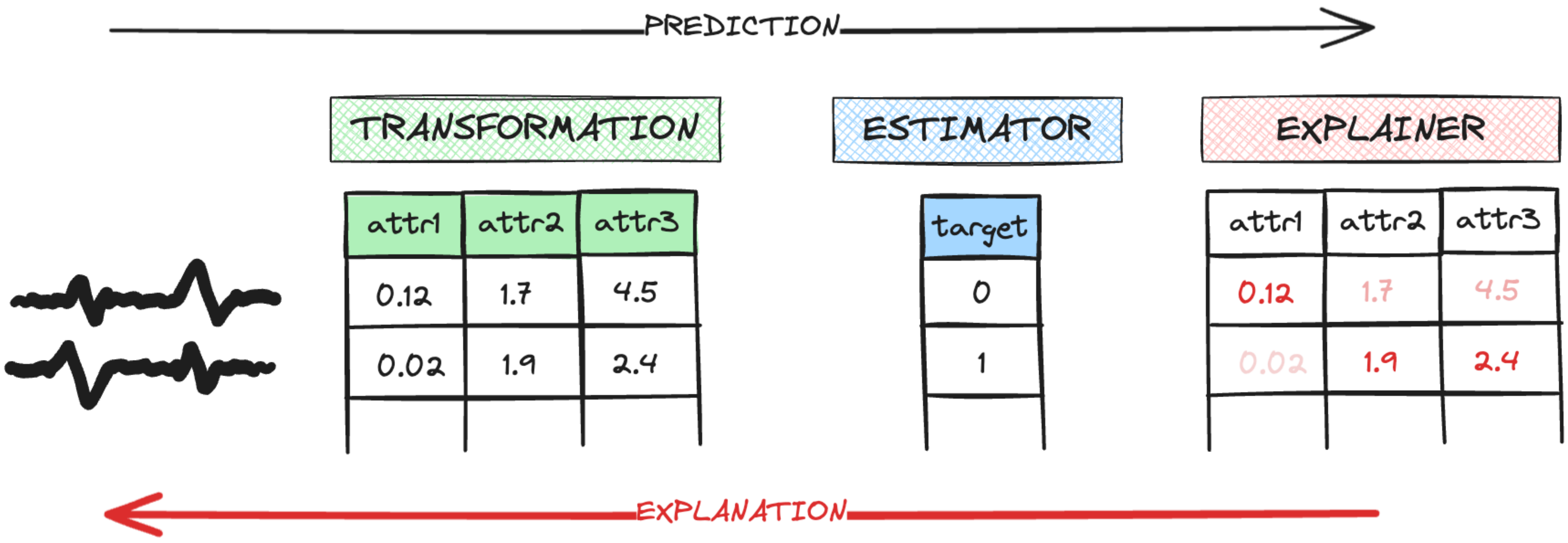
2. an **interpretable model** (or a post-hoc explainer);



A Recipe for Interpretable Sequence Prediction (3)

To achieve interpretable sequence predictions you need:

3. a way to **map** the explanation **back** to the sequence.



Bag-Of-Receptive-Fields

For Explaining Time Series Classification and Regression

Interpretable Transform ✅
Interpretable Model ✅
Mapping ✅

Spinnato, F., Guidotti, R., Monreale, A. and Nanni, M., 2024. Fast, Interpretable and Deterministic Time Series Classification with a Bag-Of-Receptive-Fields. IEEE Access.

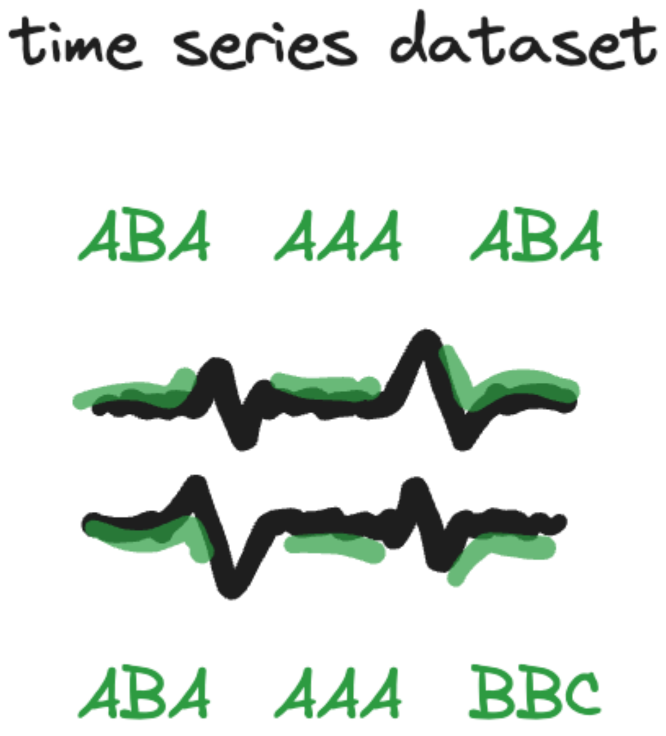
Bag-Of-Words (BOW)

The Bag-Of-Words represent a document as word counts.

	the	red	dog	cat	eats	food
1. the red dog	1	1	1	0	0	0
2. cat eats dog	0	0	1	1	1	0
3. dog eats food	0	0	1	0	1	1
4. red cat eats	0	1	0	1	1	0

Bag-Of-Patterns (BOP)

The Bag-Of-Patterns extracts "words" from time series using **SAX**.



Bag-Of-Patterns

AAA	BBC	ABA
1	0	2
1	1	1

Lin, Jessica, et al. "Experiencing SAX: a novel symbolic representation of time series." Data Mining and knowledge discovery 15 (2007): 107-144.

Bag-Of-Patterns (BOP)

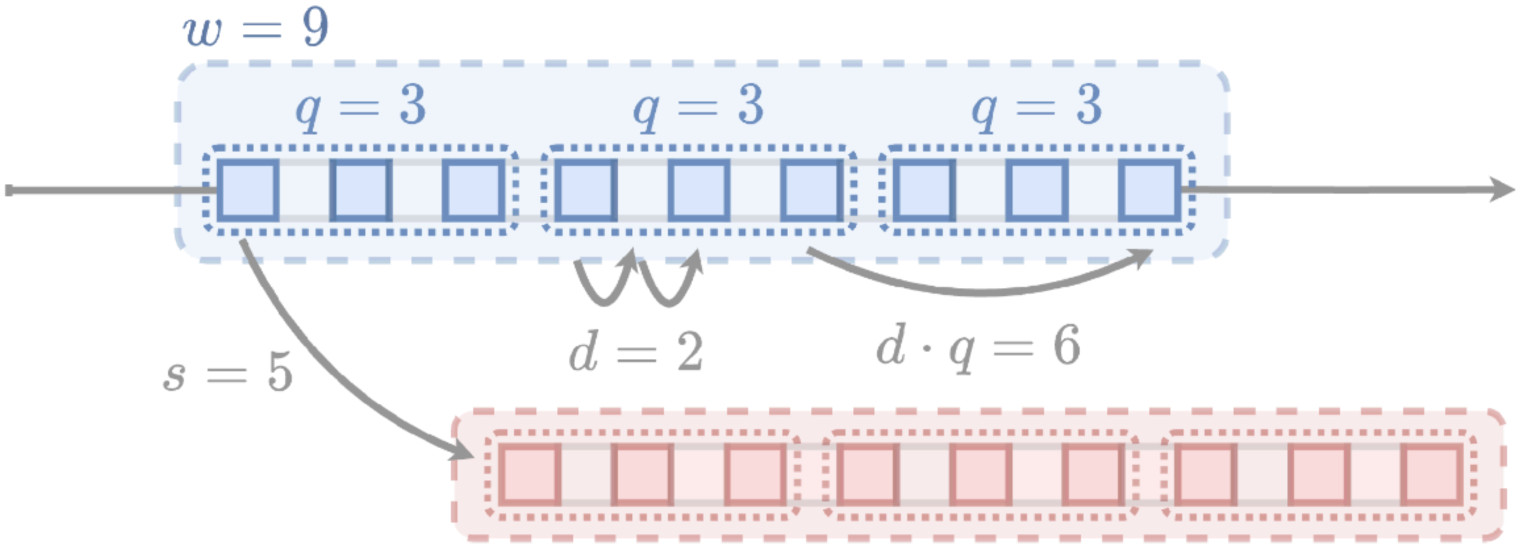
Limitations:

- it is **inefficient** (dense representation, quadratic complexity);
- has **terrible** downstream **performance**;
- works only on **regular**, **univariate** data.

Baydogan, M.G., Runger, G. and Tuv, E., 2013. A bag-of-features framework to classify time series. IEEE transactions on pattern analysis and machine intelligence, 35(11), pp.2796-2802.

Bag-Of-Receptive-Fields (BORF)

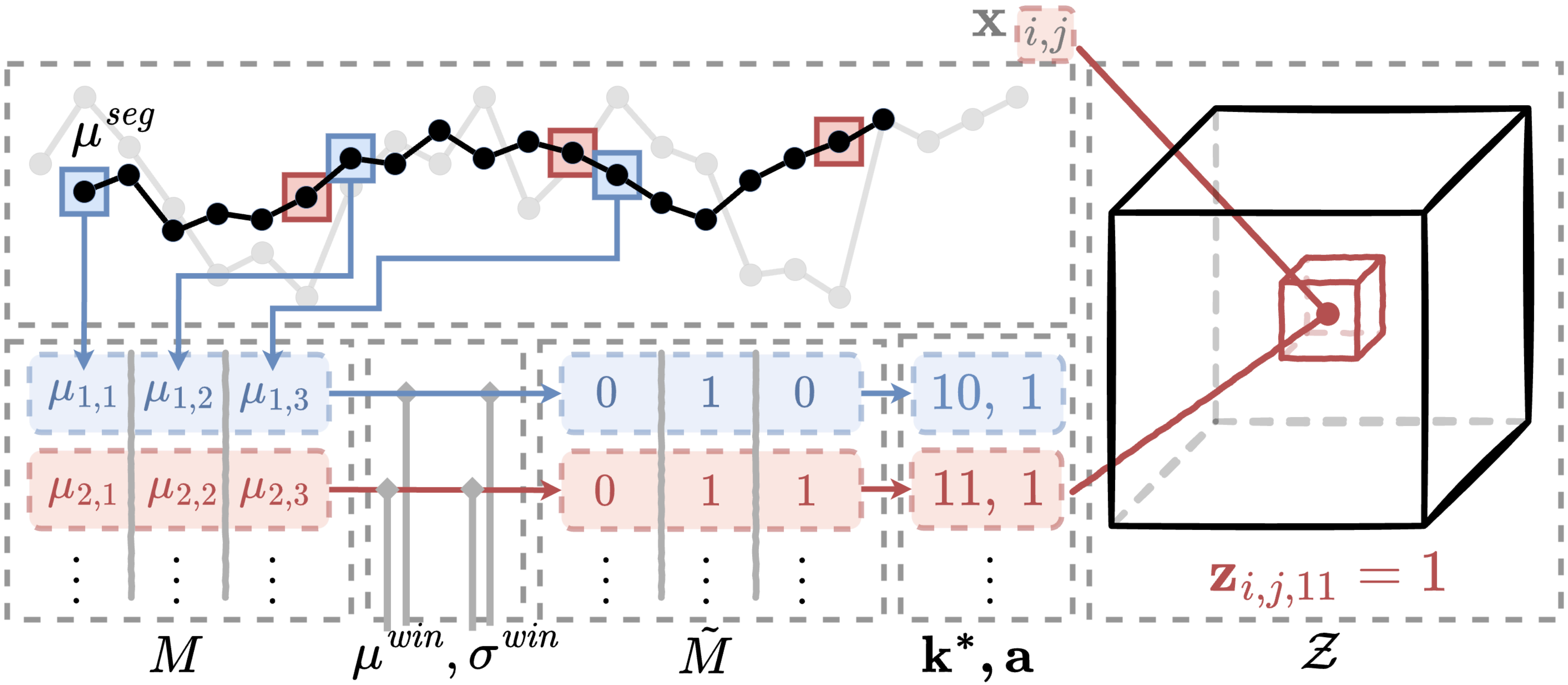
We generalize a subsequence to a **receptive field**...



Spinnato, F., Guidotti, R., Monreale, A. and Nanni, M., 2024. Fast, Interpretable and Deterministic Time Series Classification with a Bag-Of-Receptive-Fields. IEEE Access.

Bag-Of-Receptive-Fields (BORF)

... and speed up extraction, also using a **sparse representation.**



Spinnato, F., Guidotti, R., Monreale, A. and Nanni, M., 2024. Fast, Interpretable and Deterministic Time Series Classification with a Bag-Of-Receptive-Fields. IEEE Access.

Bag-Of-Receptive-Fields (BORF)

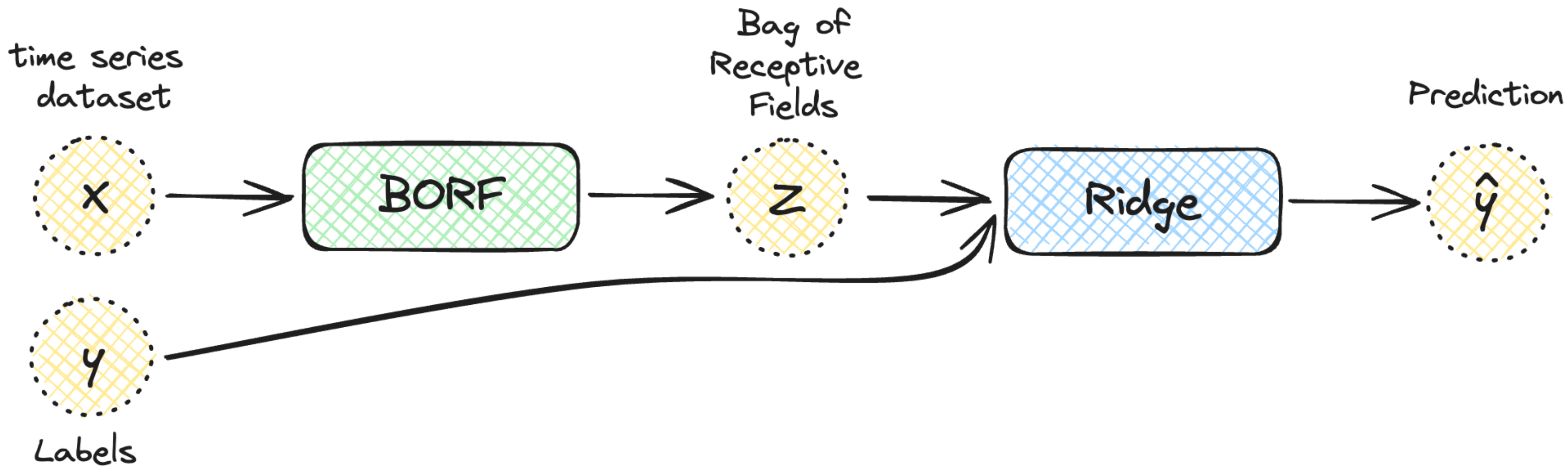
BORF:

- is **efficient** (sparse representation, linear complexity)
- has **good** downstream **performance**;
- works on **multivariate** data;
- can work on **irregular** time series.

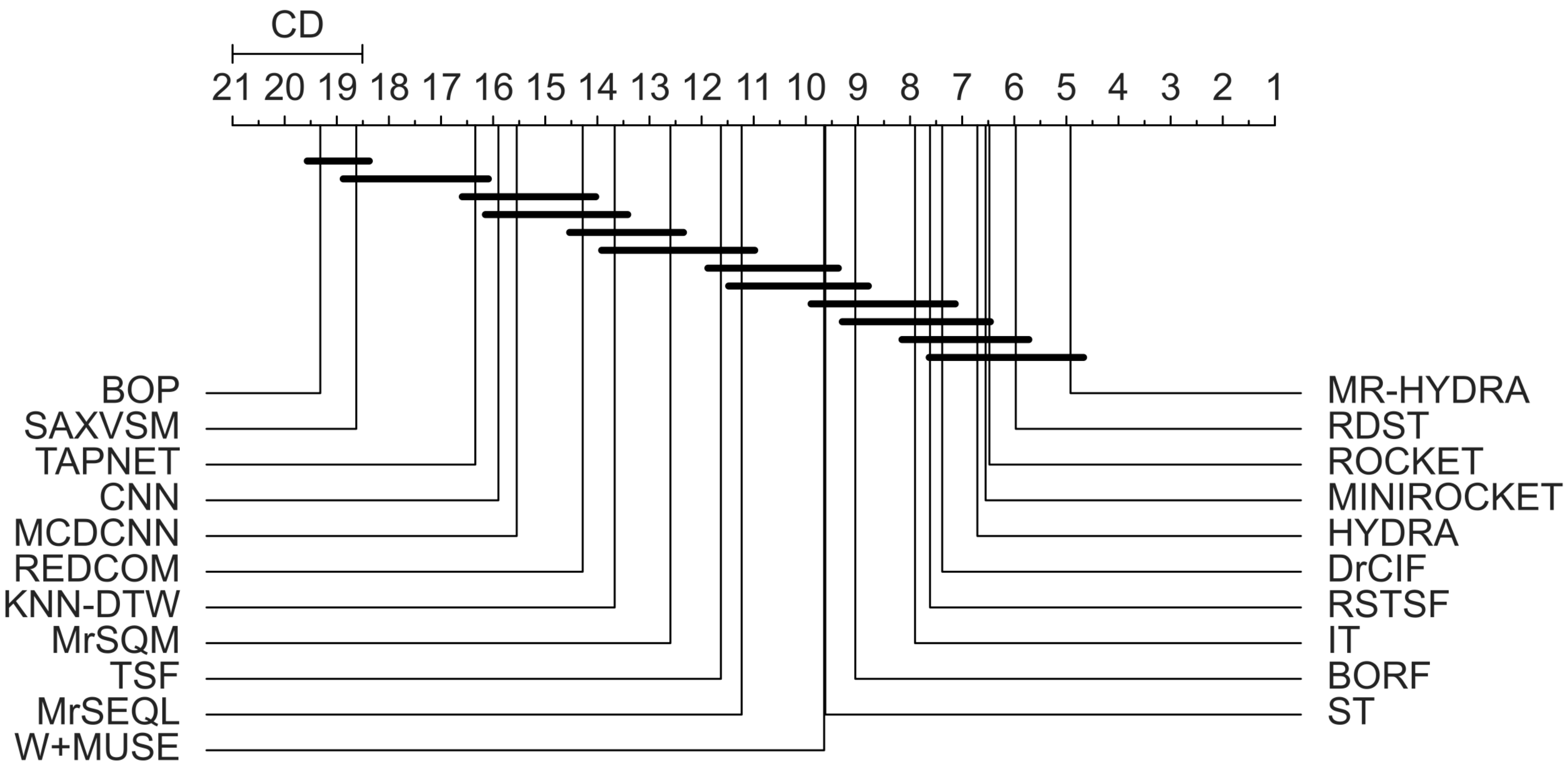
Spinnato, F., Guidotti, R., Monreale, A. and Nanni, M., 2024. Fast, Interpretable and Deterministic Time Series Classification with a Bag-Of-Receptive-Fields. IEEE Access.

How to use?

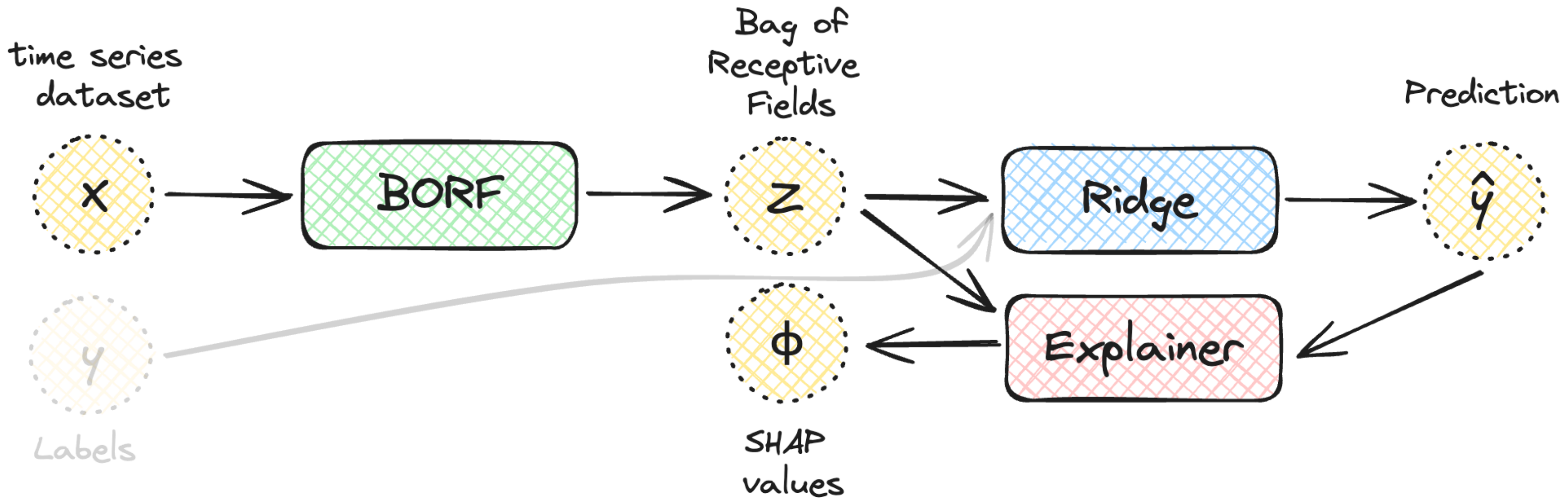
BORF is available in the *aeon* library and can be used in any pipeline



Predictive Performance

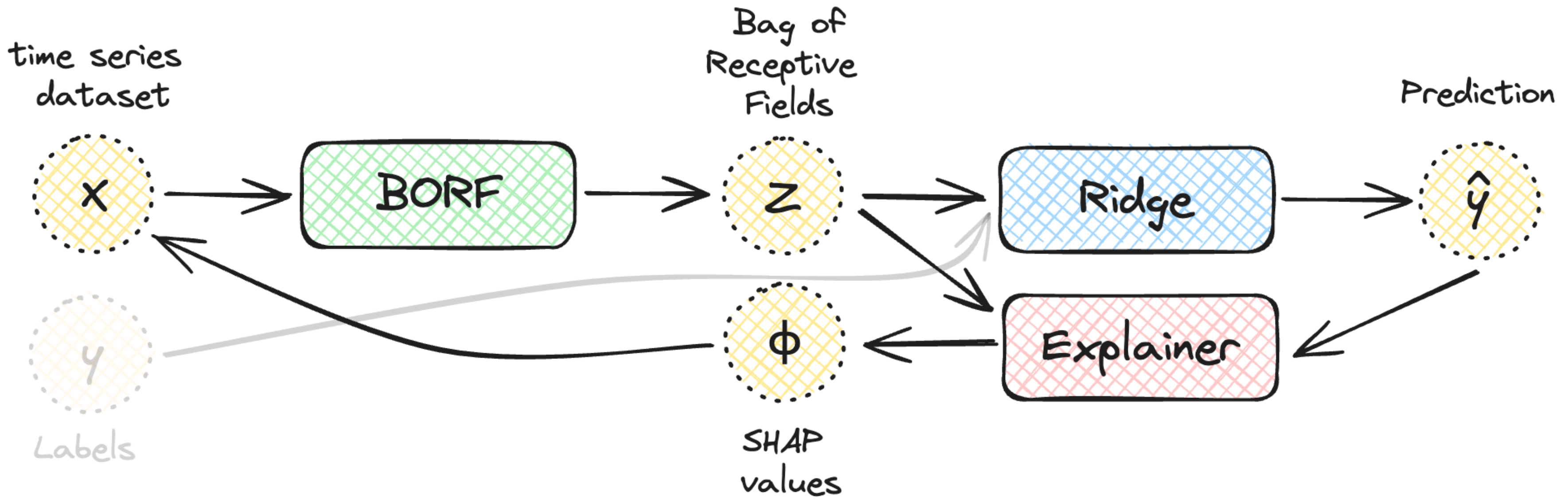


The output can be explained using any (tabular) explainer.

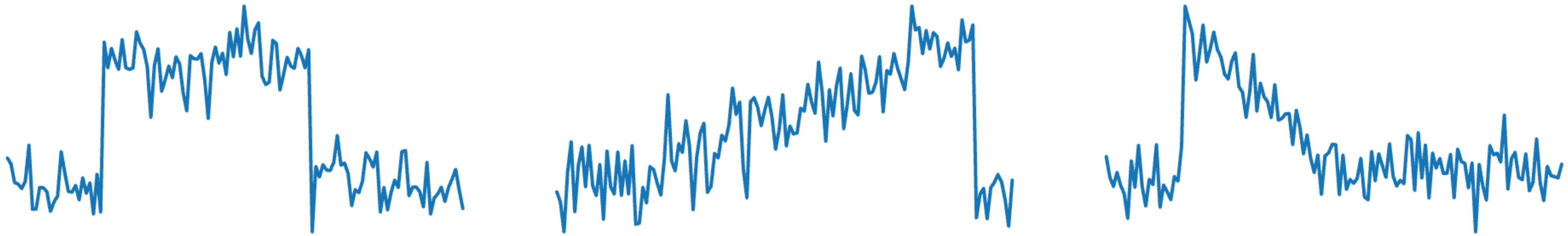


Mapping

The explanation is mapped back to the time series.



Example on a Toy Dataset (CBF)

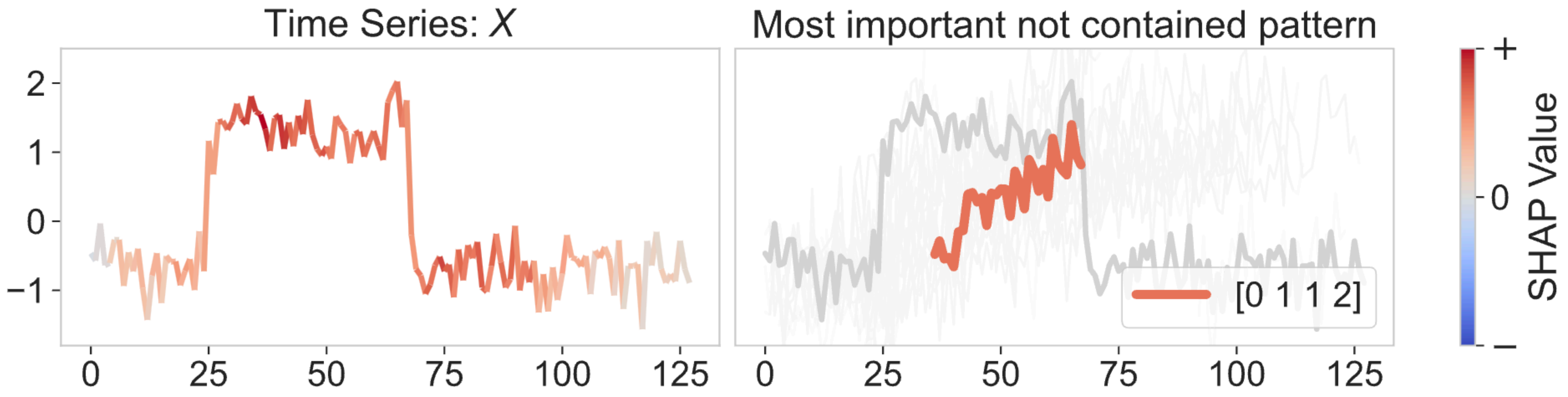


(a) Cylinder (b) Bell (c) Funnel

Scott, M., and Lee Su-In. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017): 4765-4774.

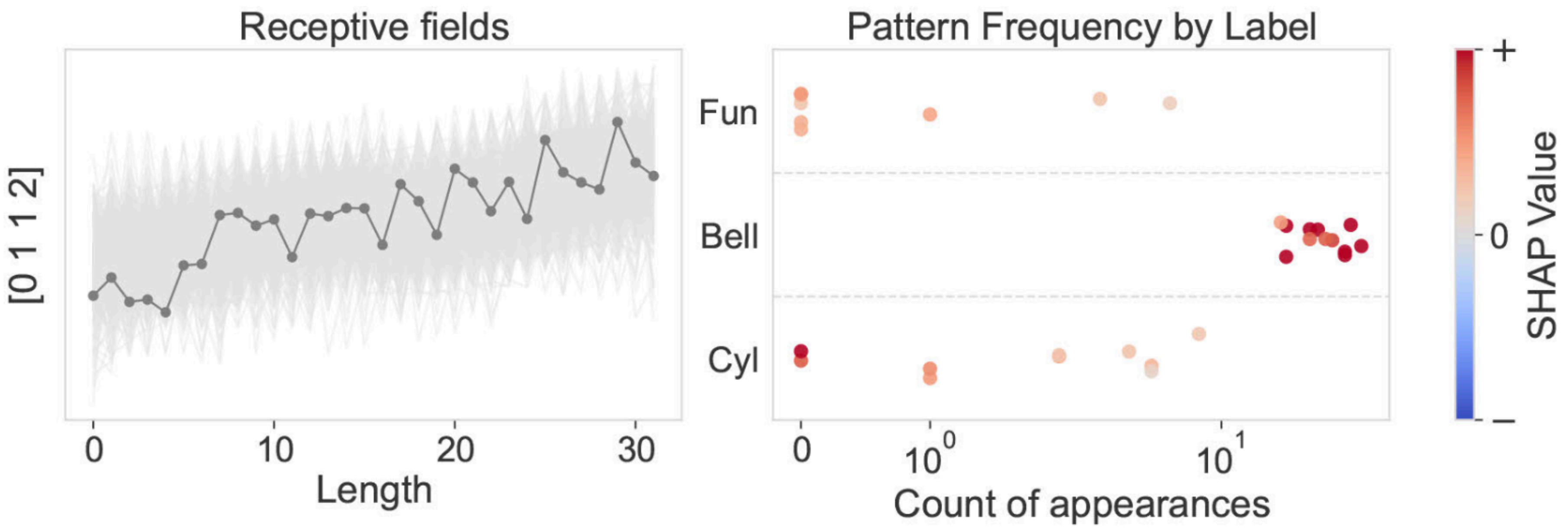
Example on a Toy Dataset (CBF)

The explanation can be *local*, i.e., on a single time series.



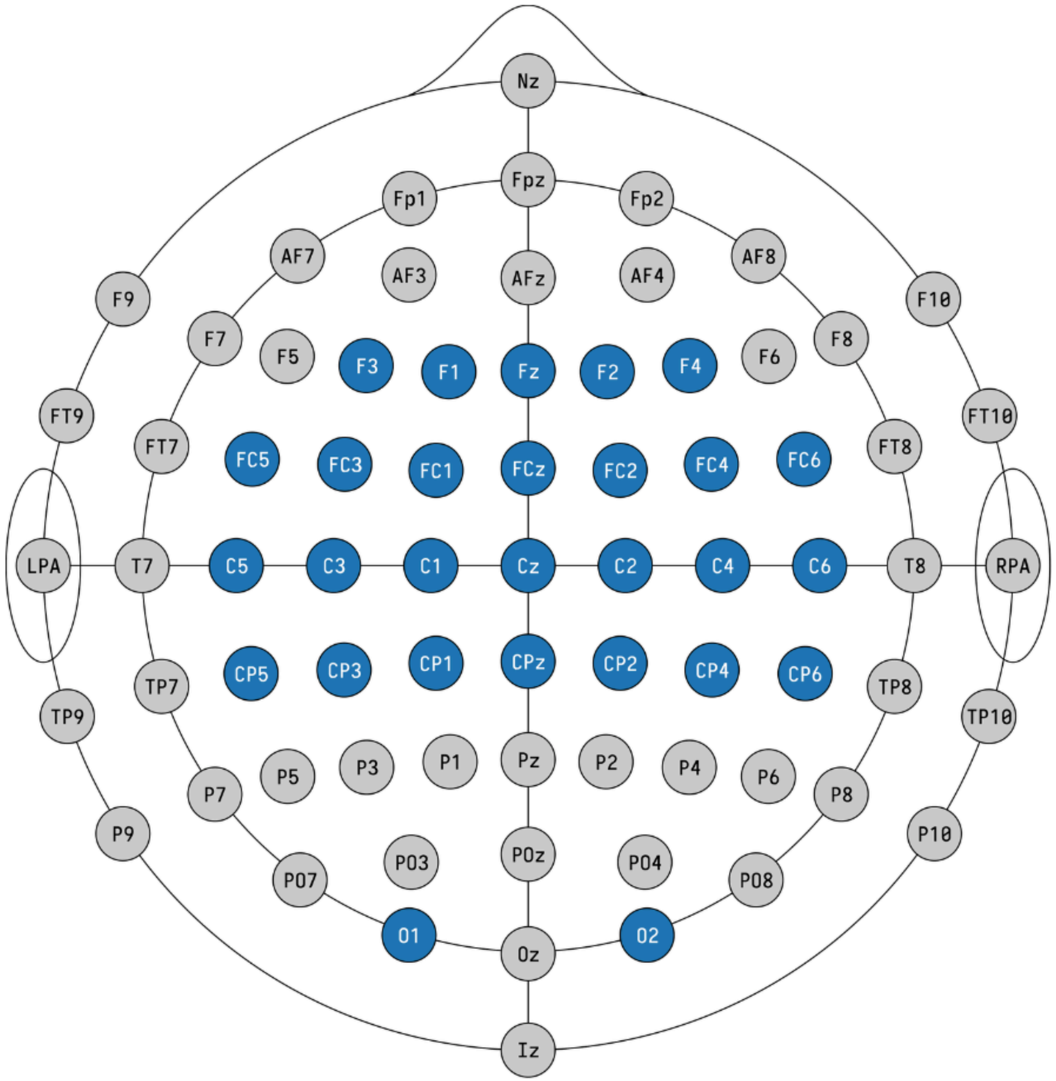
Example on a Toy Dataset (CBF)

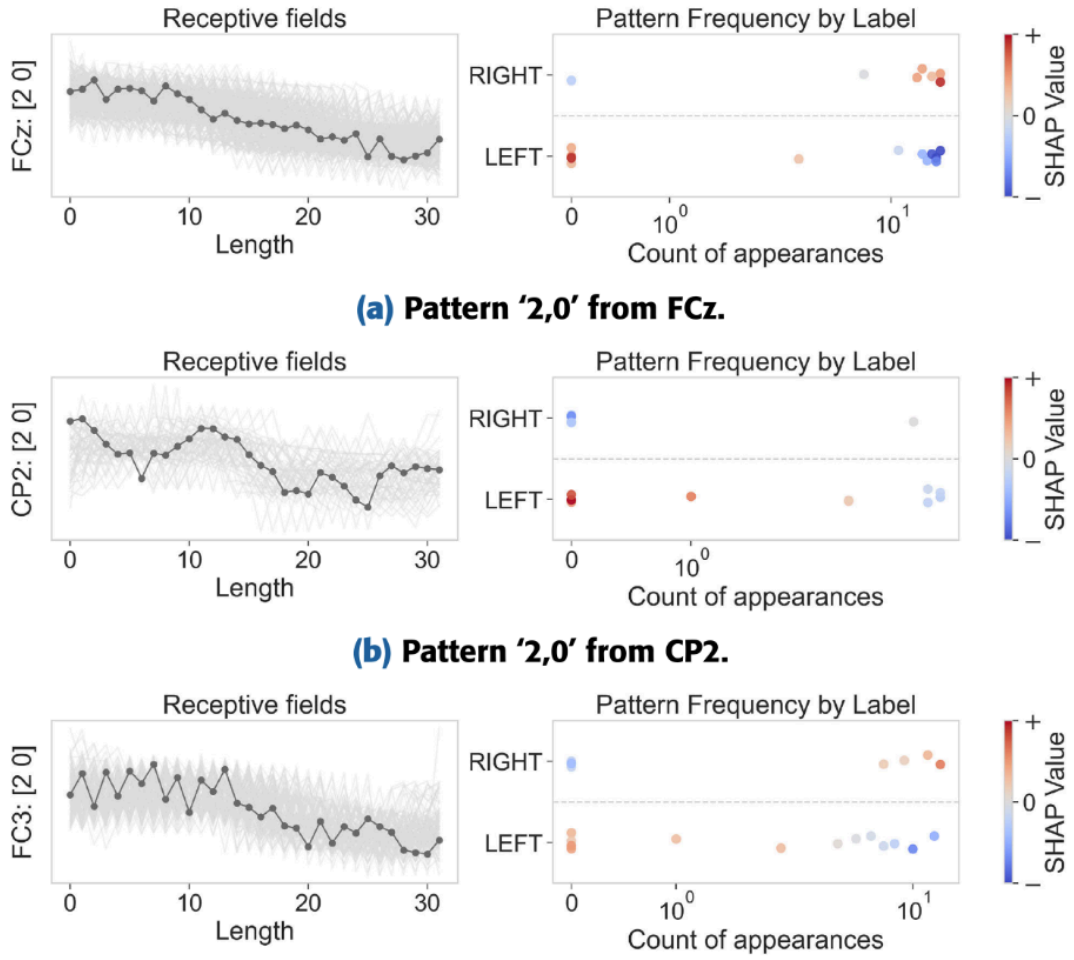
Or *global*, i.e., analyzing the pattern for a whole dataset.



(a) Pattern '0,1,1,2'.

Real Data - FingerMovements





(a) Pattern '2,0' from FCz.

(b) Pattern '2,0' from CP2.

(c) Pattern '2,0' from FC3.

Pros and Cons

✅

- fully interpretable
- deterministic
- good predictive performance
- fast and efficient
- global and local explanations
- element and subsequence-based explanations
- streamlined library

⛔

- ugly name

Shapelet Transform

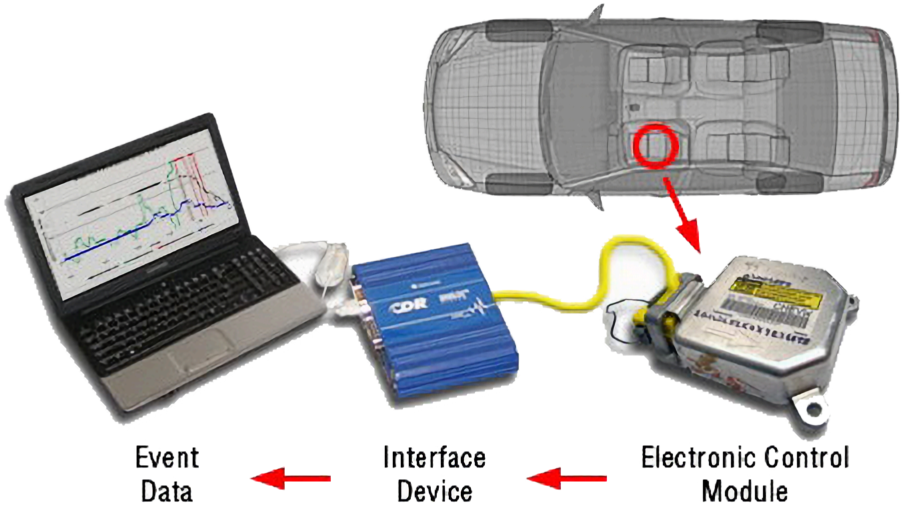
for Explaining Car Crash Predictions

Interpretable Transform ✅
Interpretable Model ✅
Mapping ✅

- Bianchi, M., Spinnato, F., Guidotti, R., Maccagnola, D., & Bencini Farina, A. (2024, October). Multivariate Asynchronous Shapelets for Imbalanced Car Crash Predictions. DS 2024
- Spinnato, Francesco, et al. "Explaining crash predictions on multivariate time series data." International Conference on Discovery Science. DS 2022.

Crash Data Recorders

Crash Data Recorders (CDRs) can retrieve car's event data which **monitor safety measures and record impact speeds.**

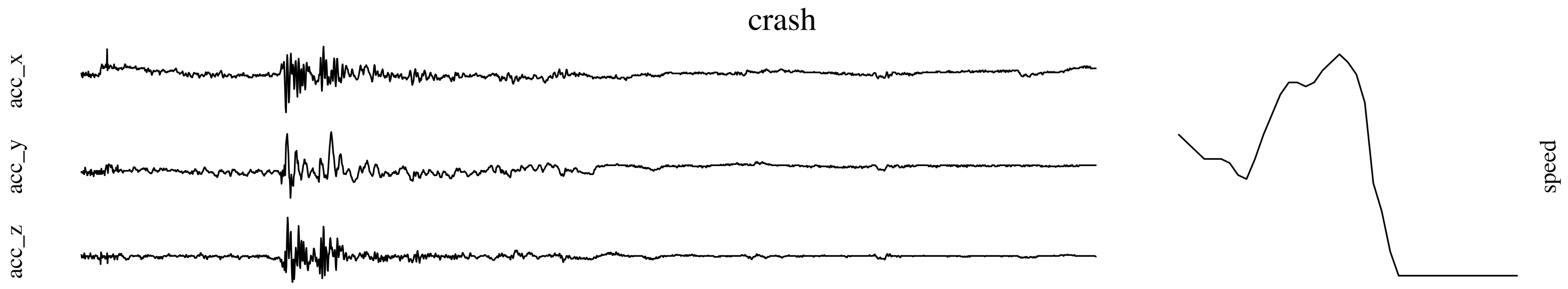


We collaborate with Assicurazioni Generali to **detect car crashes.**

---

https://www.generali.it/

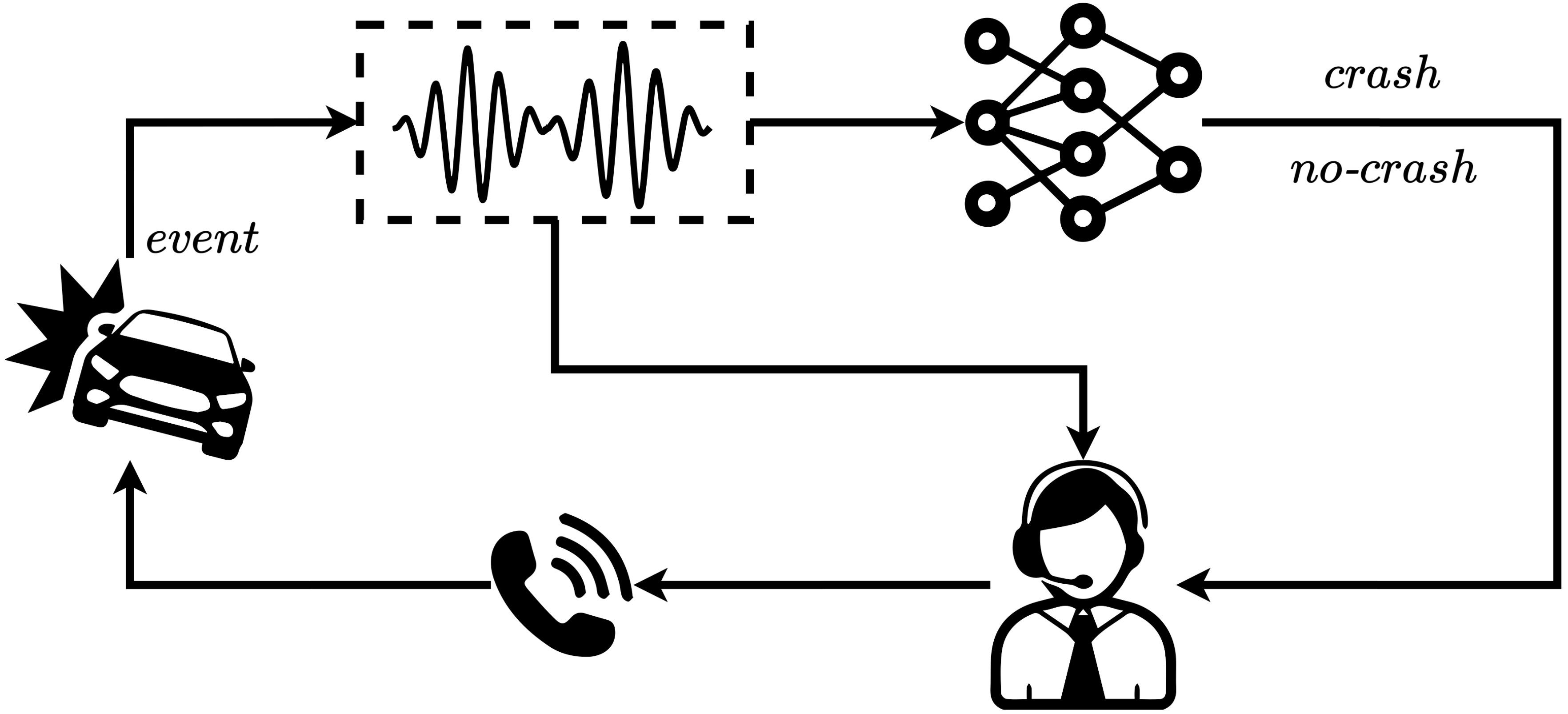
Generali's dataset is composed by **multivariate** time series containing 4 signals (acceleration on x,y,z axis and speed).



It is a challenging **binary classification dataset**, big, highly imbalanced, with classes: *Crash* ($\sim 1\%$) and *No-Crash* ($\sim 99\%$).

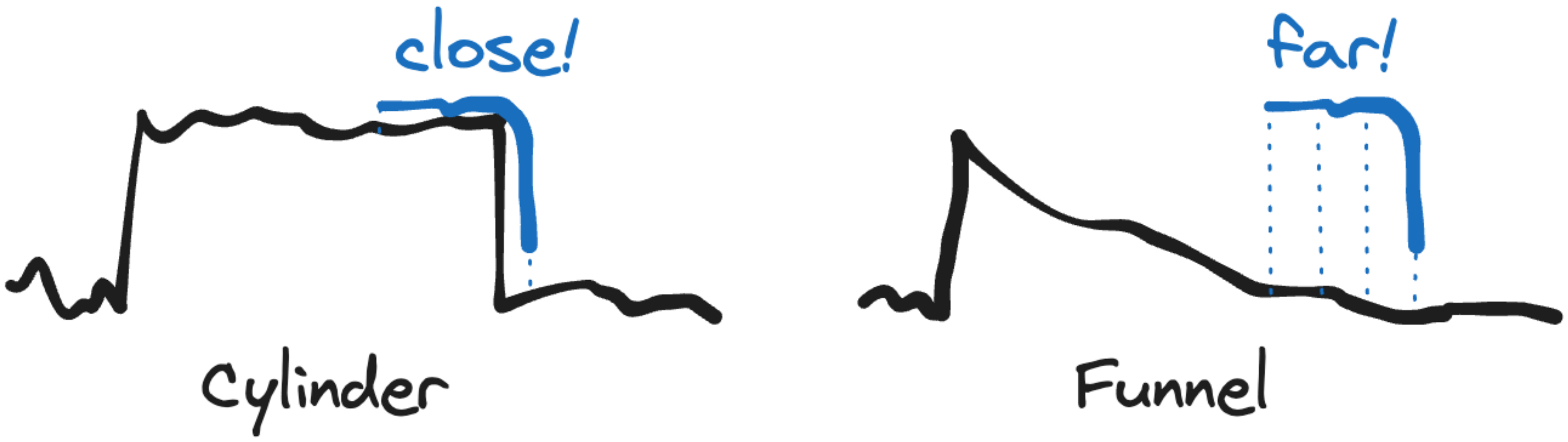
Generali's Workflow

Generali has a customer assistance workflow with a human-in-the-loop



Shapelets

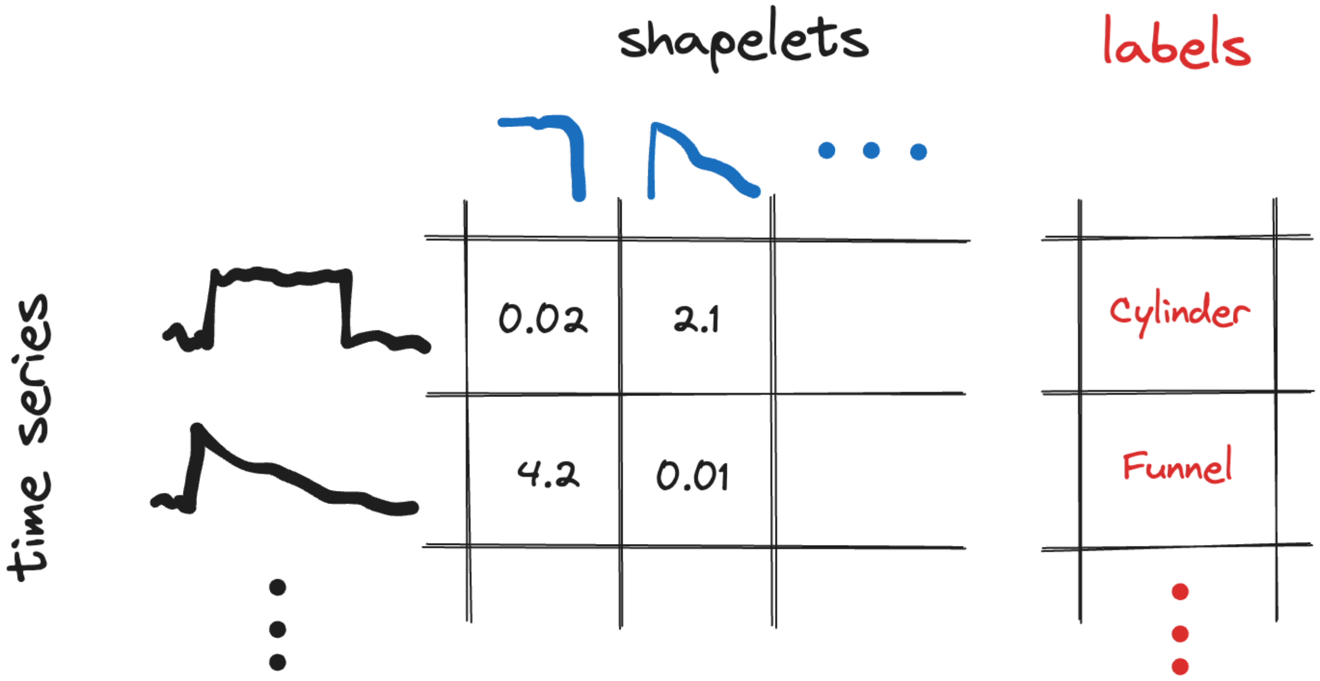
Highly representative and **discriminative** time series **subsequences** for a particular class in a time series dataset.



They can be extracted in many supervised and unsupervised ways.

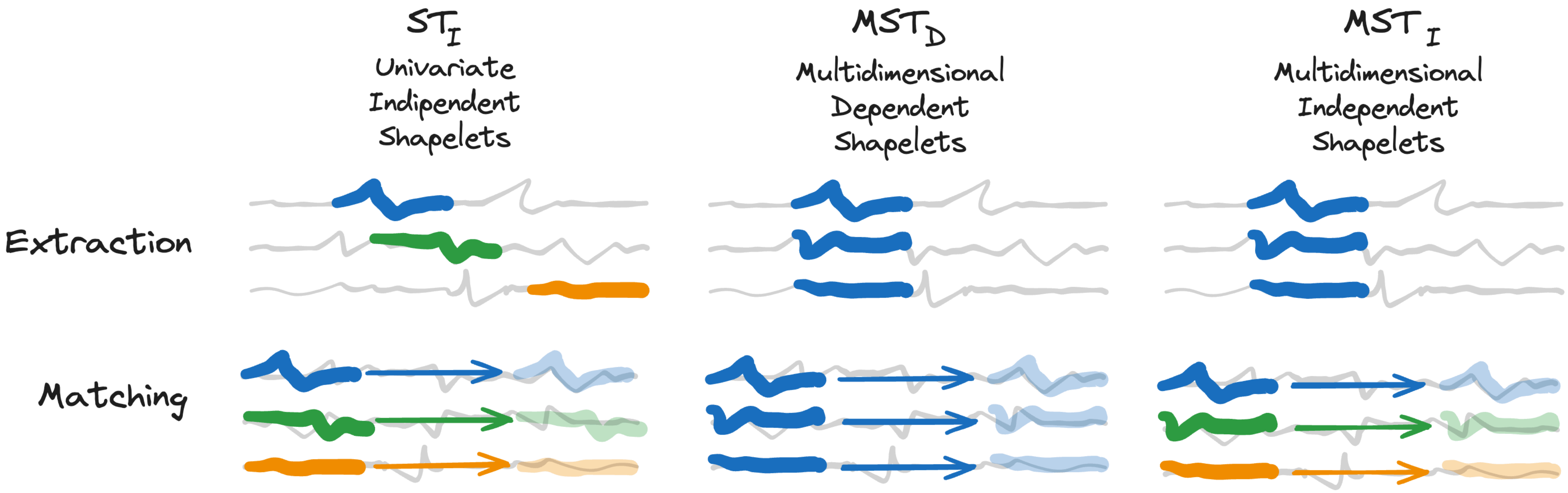
Shapelet Transform

Represent the time series dataset in a **tabular form**, as the **minimum** sliding-window **distance** between each shapelet and each time series.



Even for multivariate data, usually shapelets are **univariate**.

Multivariate shapelets are scarcely addressed in the literature.



Bostrom, Aaron. Shapelet transforms for univariate and multivariate time series classification. Diss. University of East Anglia, 2018.

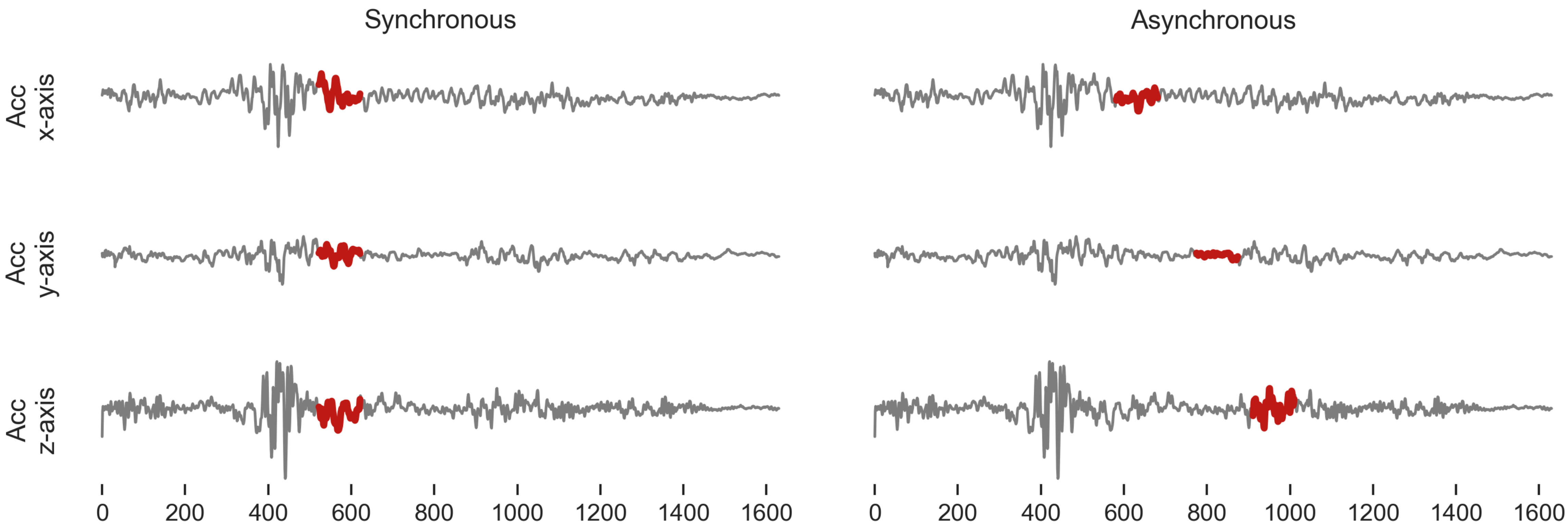
Limitations

There are several limitations to these approaches:

- ST_I shapelets are univariate;
- MST_D, MST_I only extract aligned/synchronous shapelets;
- poor performance on **imbalanced** datasets;
- **interpretability** is assumed but **rarely explored**;
- computational complexity.

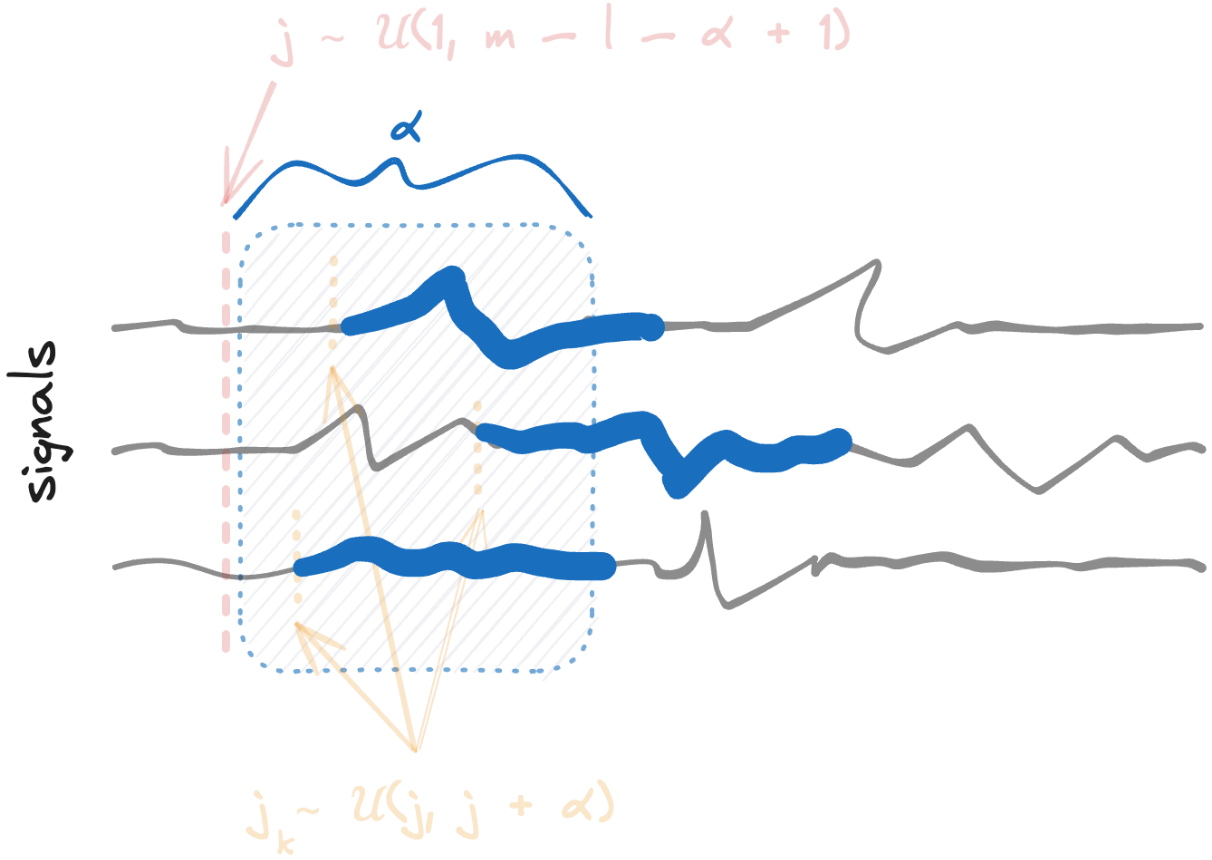
Multivariate Asynchronous Shapelets (MARS)

We want to control the **asynchronicity** of the shapelets.



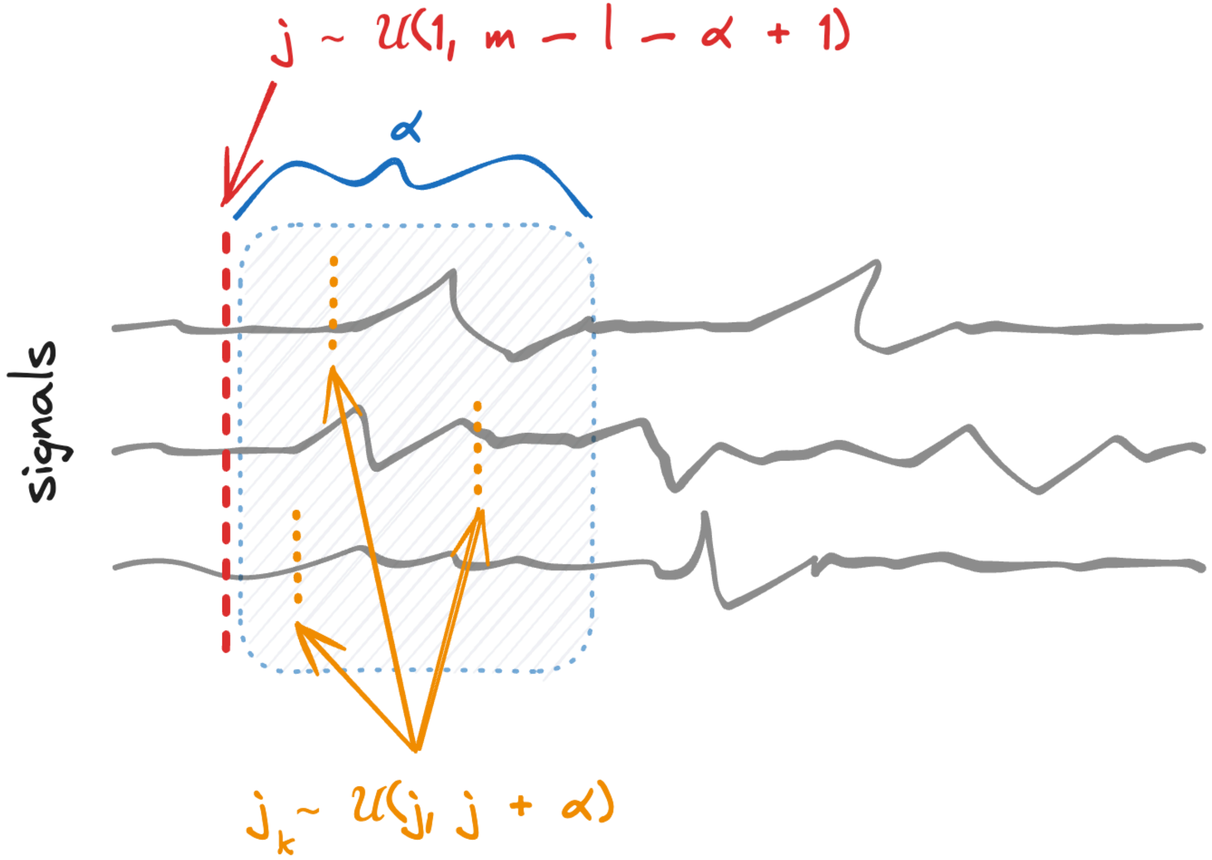
Shapelet Extraction

α defines how much the multivariate shapelet can be misaligned.



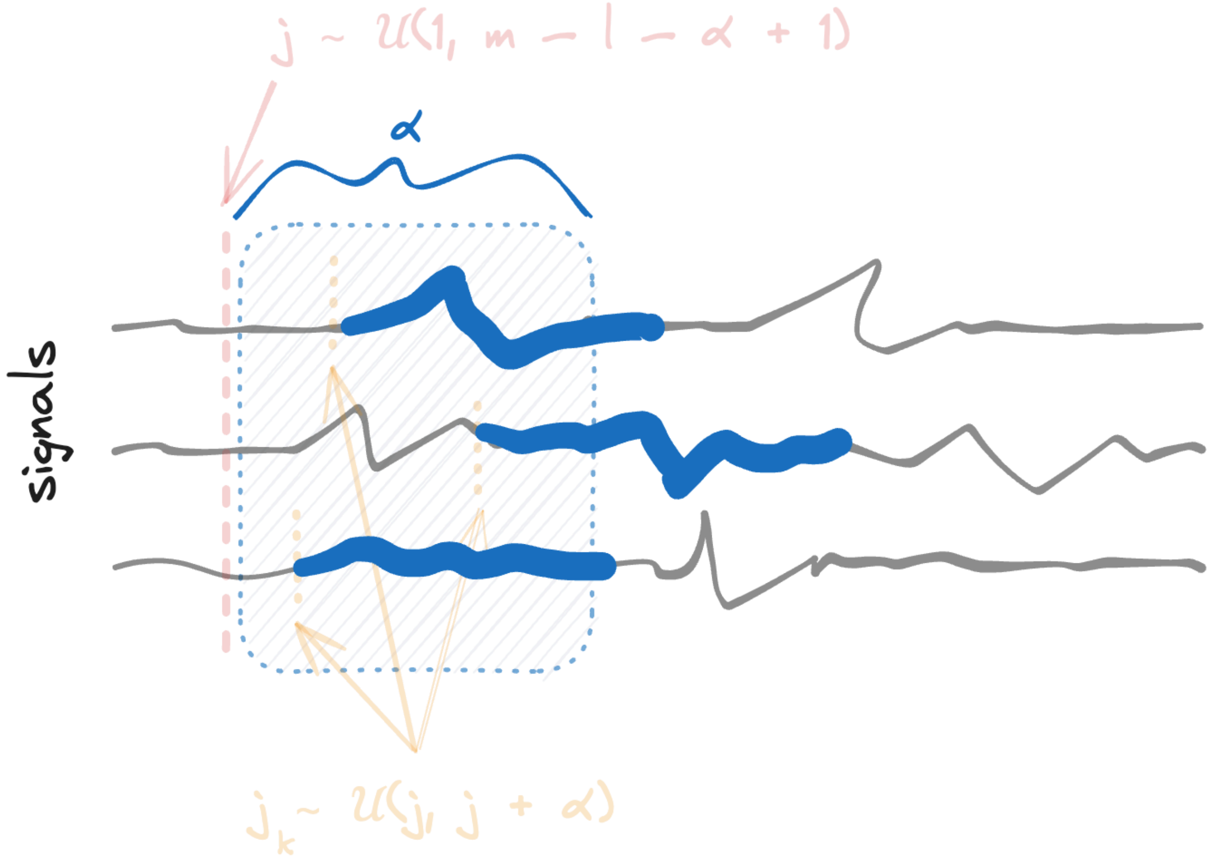
Shapelet Extraction

To enforce this we perform two index samplings: global and specific.



Shapelet Extraction

This ensures that the starting indexes are within α timesteps.



Shapelet Extraction for Imbalanced Datasets

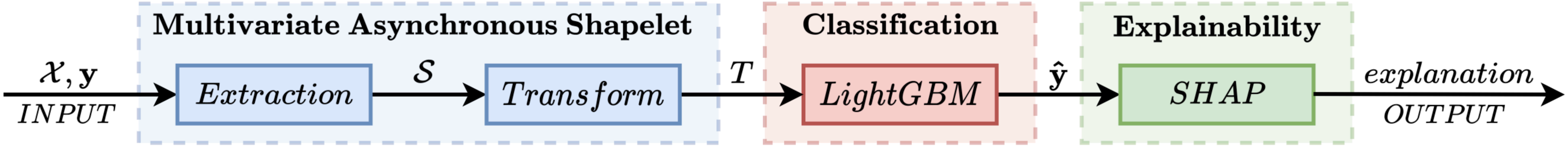
The **brute force** extraction of shapelets is **not feasible**, therefore, we randomly extract h shapelets from the TSC dataset.

Specifically we extract $\lfloor \frac{h}{2} \rfloor$ shapelets from time series belonging to **each of the 2 classes**.

This ensures that, if h is sufficiently high, both *Crash* and *No-Crash* instances will be represented by some shapelets.

MARS Pipeline

The Multivariate Asynchronous Shapelet extraction is used in a classification pipeline to **predict and explain** car crashes.



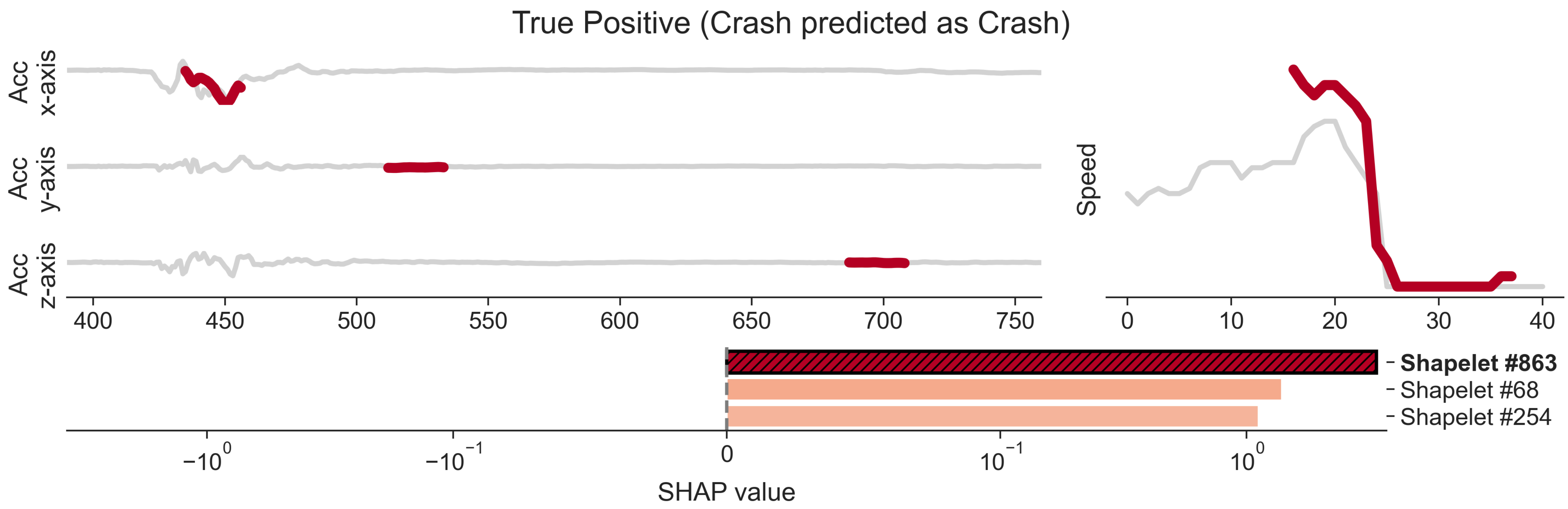
State-Of-The-Art Comparison

We compare MARS against the state-of-the-art **classifiers** and **anomaly detection** algorithms.

	MARS	$\mathrm{GEN_L}$	$\mathrm{GEN_H}$	TSF	ROCKET	ROCKAD	XGB	LGBM
f1	0.19	0.17	**0.31**	0.18	0.28	0.01	0.14	0.12
gini	**0.71**	0.66	0.47	0.64	0.53	0.28	0.58	0.66
gm	**0.84**	0.81	0.69	0.80	0.73	0.56	0.76	0.81
tp	**38**	35	25	34	28	18	31	35
fp	315	330	**84**	283	121	6122	349	514
fn	**15**	18	28	19	25	35	22	18
$time_{tr}$	17 h	-*	-*	3 h	1 h	5 h	5 min	**1 min**
$time_{in}$	0.82 s	0.09 s	0.09 s	0.14 s	0.07 s	0.01 s	35 µs	**30 µs**

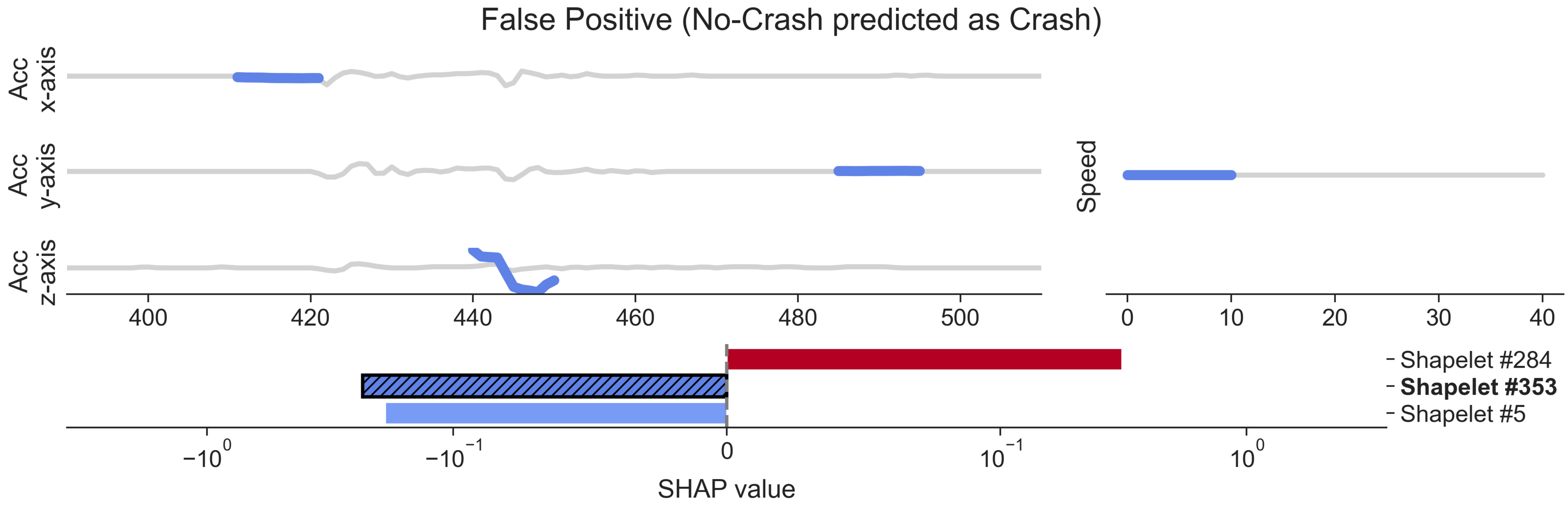
* Data unavailable due to training on Generali's system.

MARS returns an explanation in terms of **multivariate shapelet contribution** (SHAP values) toward either the class *Crash* or *No-Crash*.



Explaining Car Crashes (False Positive)

MARS returns an explanation in terms of **multivariate shapelet contribution** (SHAP values) toward either the class *Crash* or *No-Crash*.



Pros and Cons

✅

- fully interpretable
- good predictive performance on a specific task

⛔

- only subsequence-based explanations
- slow training time
- randomness

Shapelet Transform

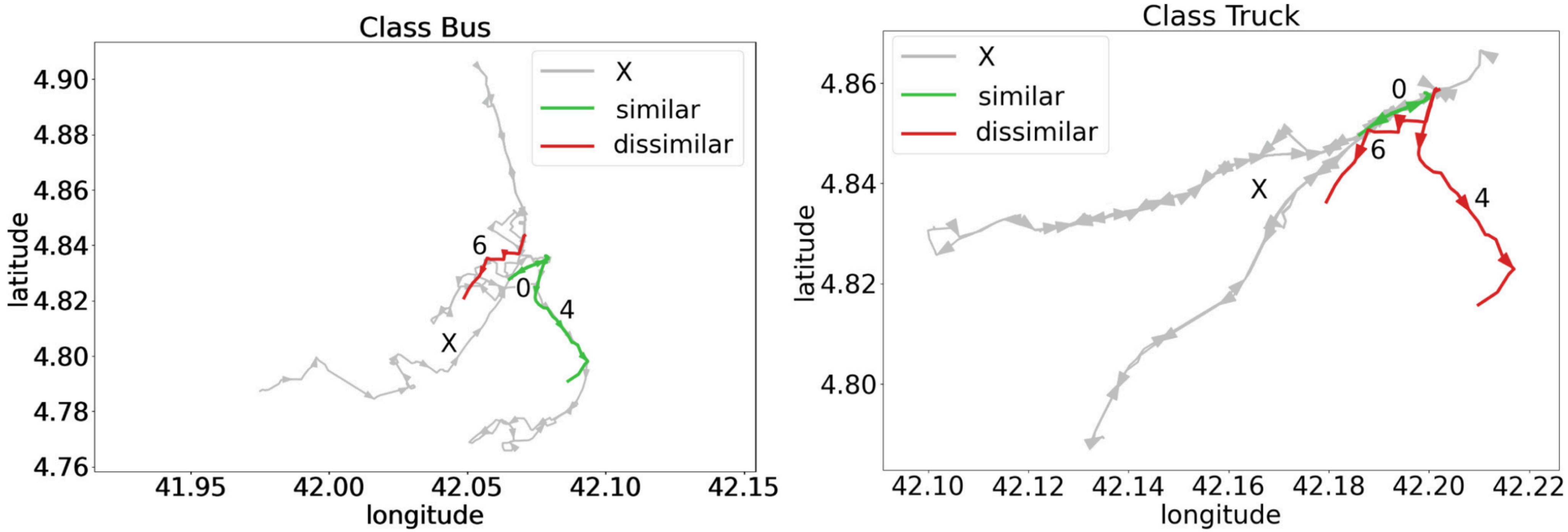
for Explaining Trajectory Classification

Interpretable Transform ✅
Interpretable Model ✅
Mapping ✅

Landi, C., Spinnato, F., Guidotti, R., Monreale, A., & Nanni, M. Geolet: An Interpretable Model for Trajectory Classification. IDA 2023

Geographic Shapelet Classifier (Geolet)

Multivariate shapelets can also be used with **trajectory** data. The main challenge is that trajectories are **irregular**.



Embeddings

for Explaining Time Series Classification

Interpretable Transform ⛔
Interpretable Model ⚠️
Mapping ✅

- Spinnato, Francesco, et al. "Understanding Any Time Series Classifier with a Subsequence-based Explainer." ACM Transactions on Knowledge Discovery from Data 18.2 (2023): 1-34.
- Guidotti, R., Monreale, A., Spinnato, F., Pedreschi, D., & Giannotti, F. (2020, October). Explaining any time series classifier. CogMI 2020 (pp. 167-176). IEEE.

Local Agnostic Subsequence-based Time Series explainer (LASTS)

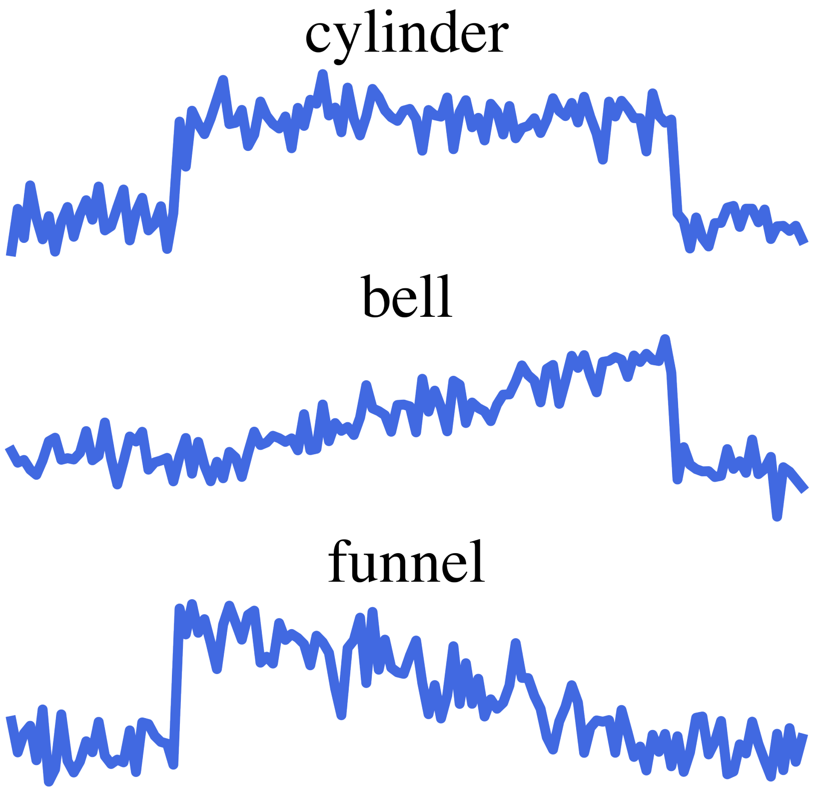
- **Local**: explains the prediction of the black-box for single univariate/multivariate instances (not the entire model);
- **Agnostic**: can explain any black-box.

LASTS can output three kinds of explanations:

- **Element-based**: saliency map;
- **Subsequence-based**: factual/counterfactual rules;
- **Instance-based**: prototypical/counterfactual instances.

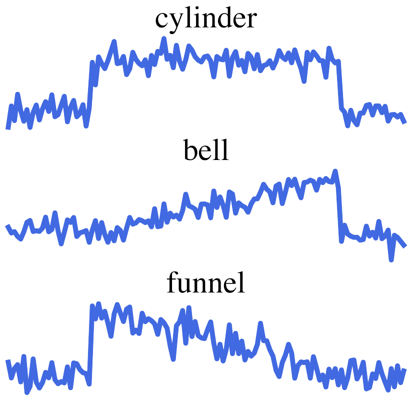
An Example of an Explanation

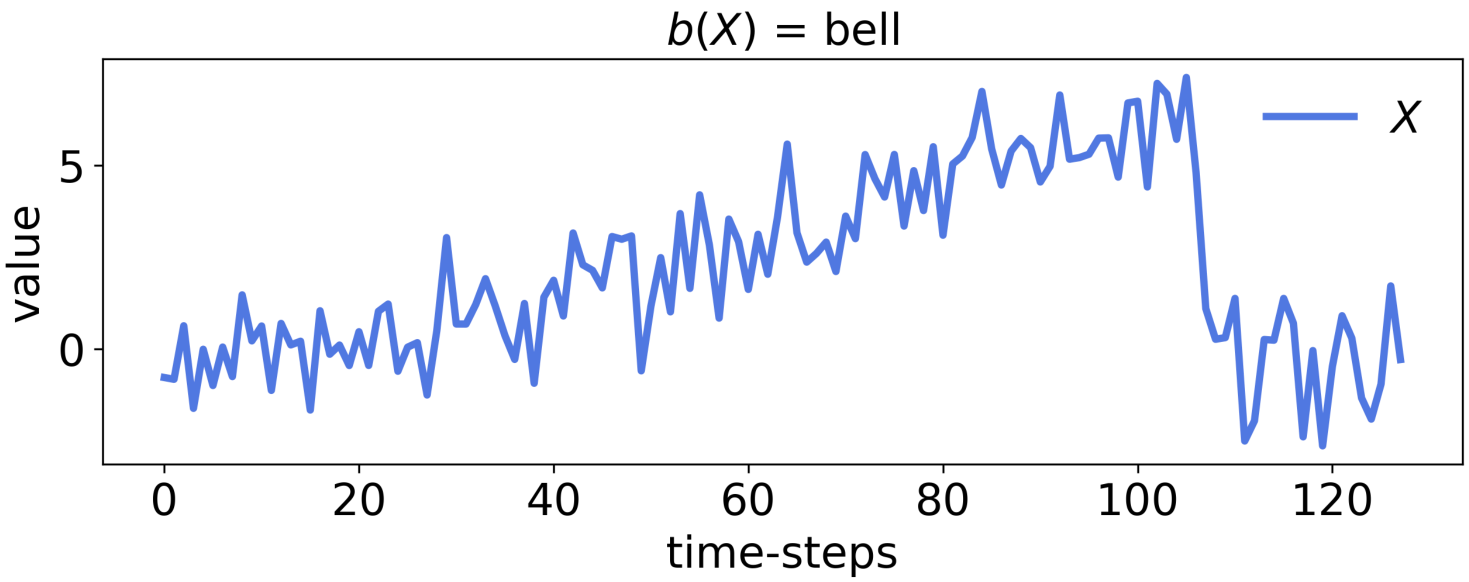
Let's take CBF, a simple dataset with **three classes.**



An Example of an Explanation

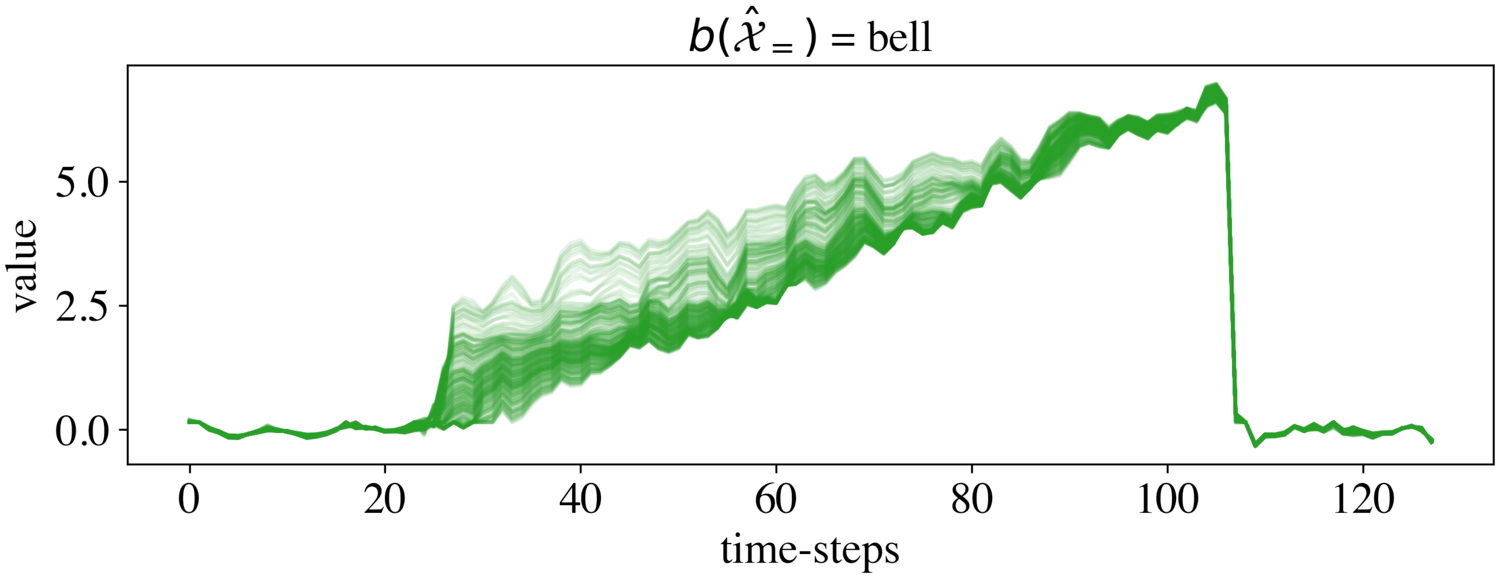
We want to **explain the prediction** of a black-box b for the time series X, i.e., why $b(X) = \text{bell}$?

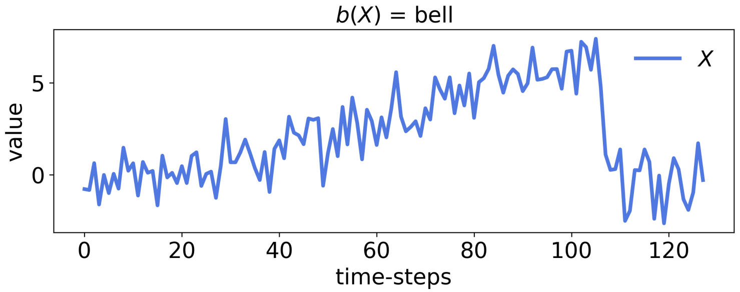




1. Instance-based Explanation: Prototypes

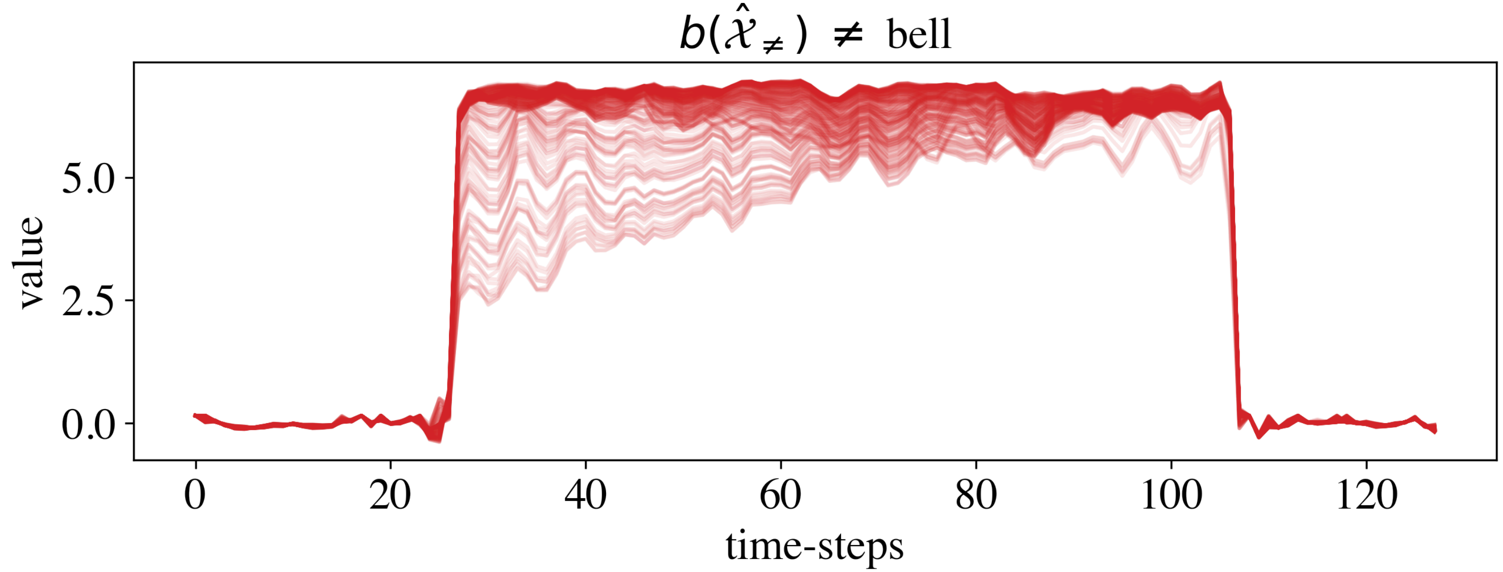
Prototypes are instances similar to X and with the same class.

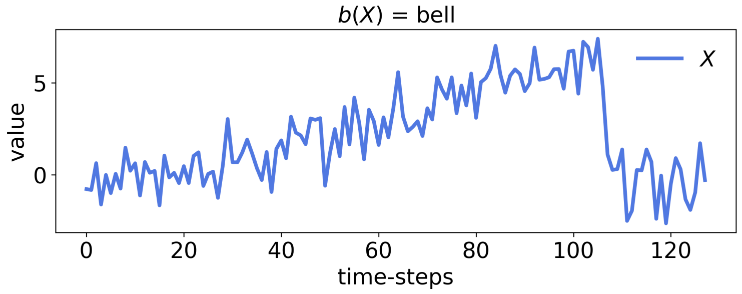




1. Instance-based Explanation: Counterfactuals

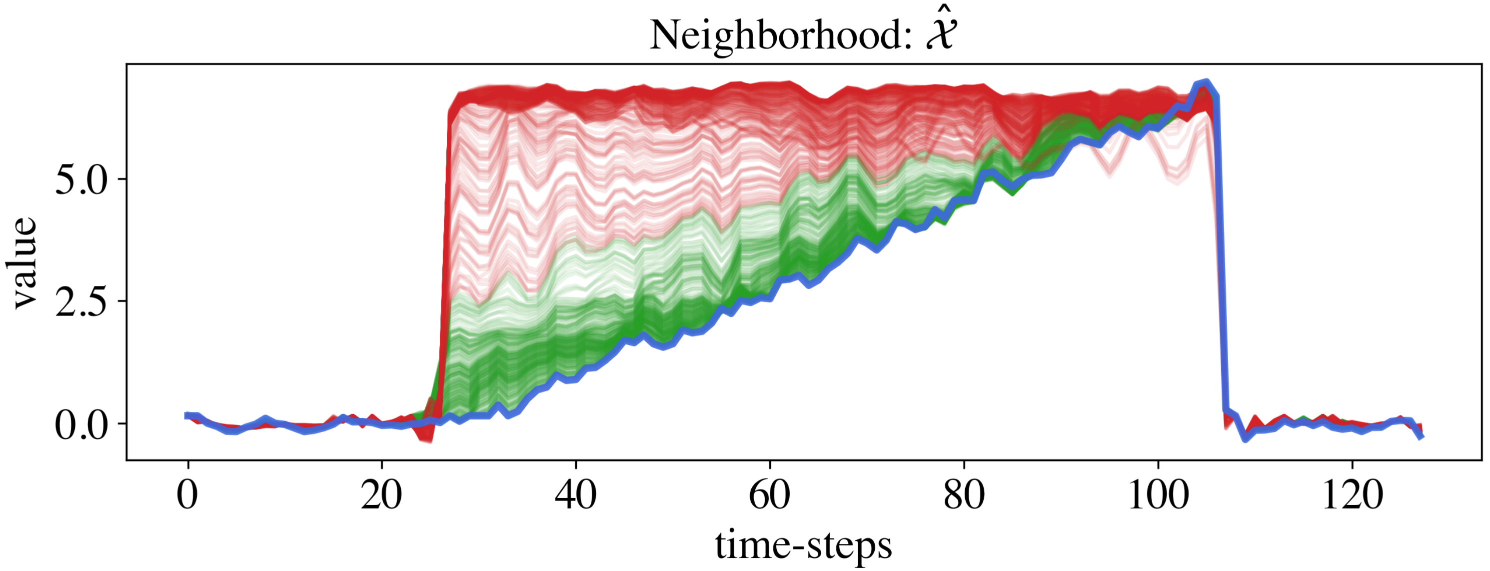
Counterfactuals are instances similar to X and with a different class.

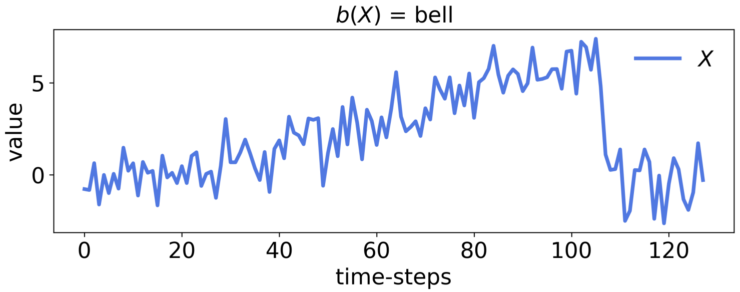




1. Instance-based Explanation

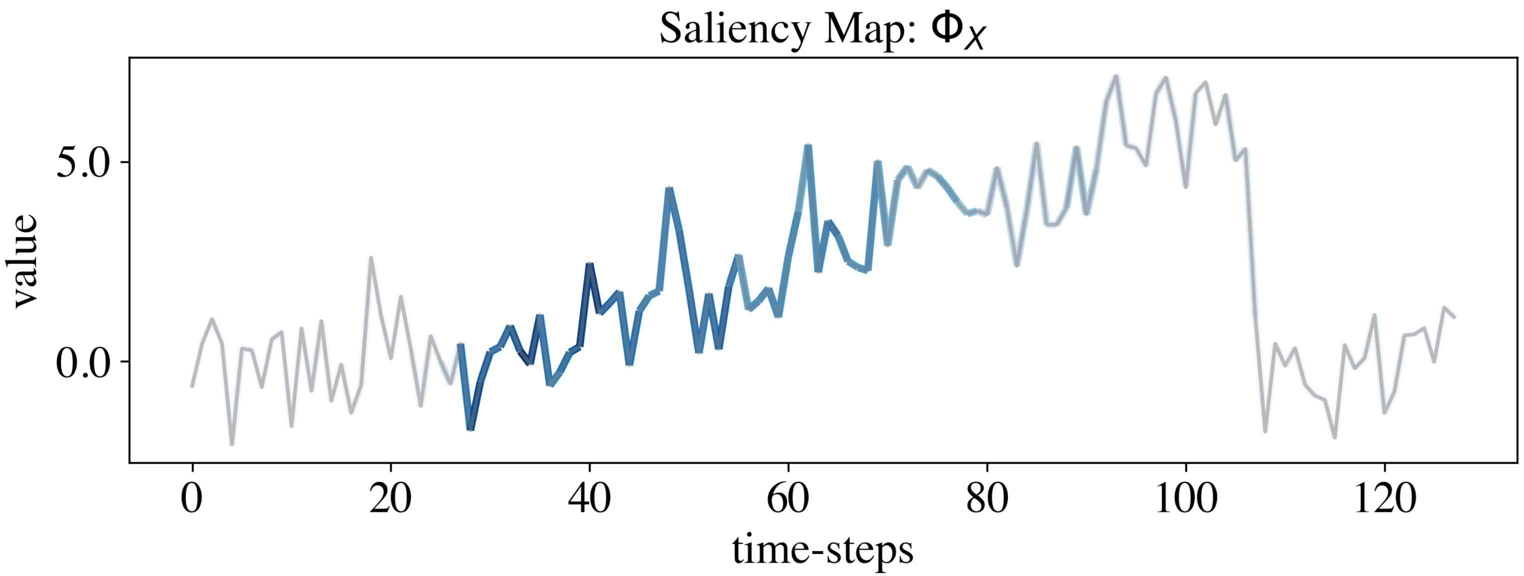
Together they form a local neighborhood.

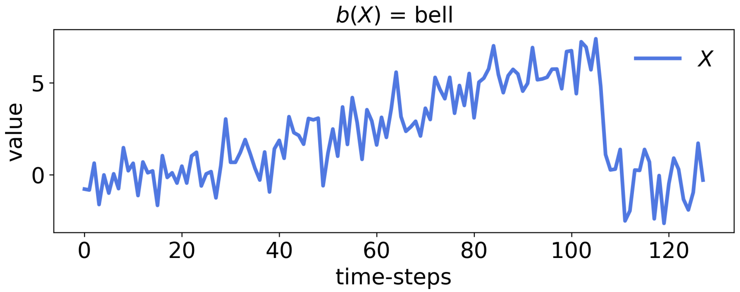




2. Element-based explanation

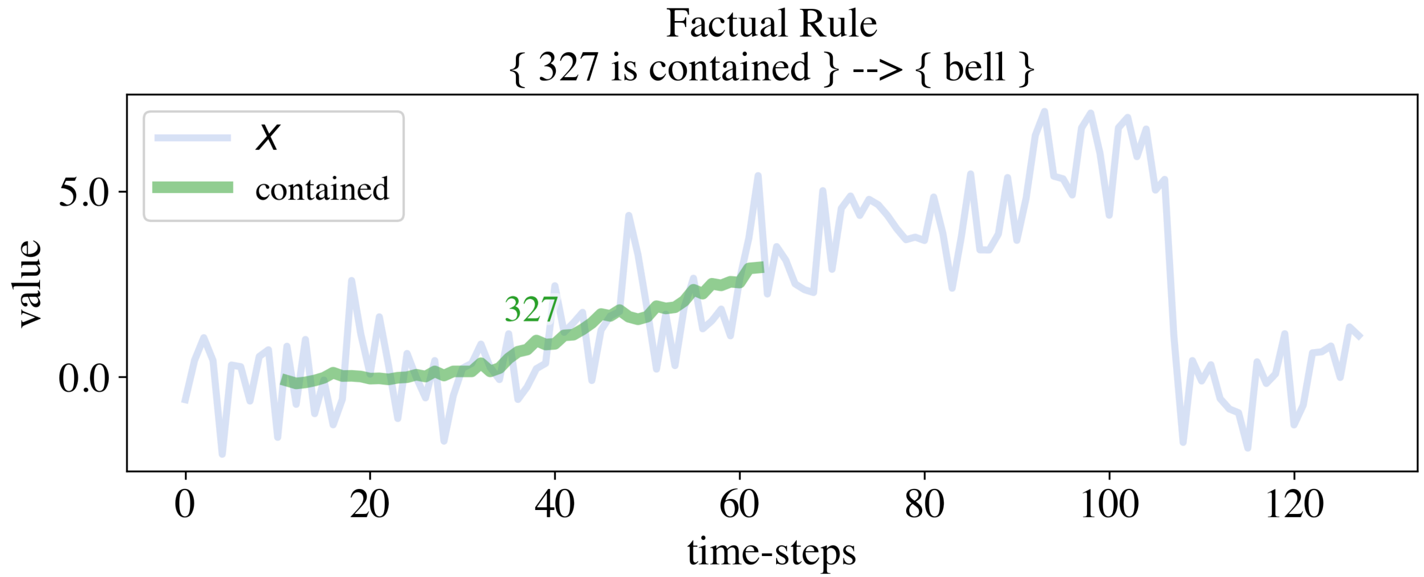
The most *sensitive* areas in the neighborhood form a saliency map.

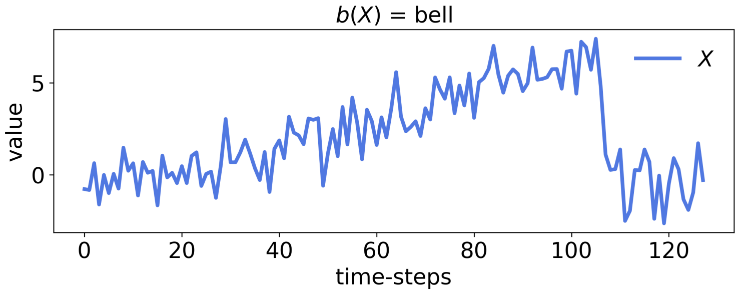




3. Subsequence-based explanation: Factual Rule

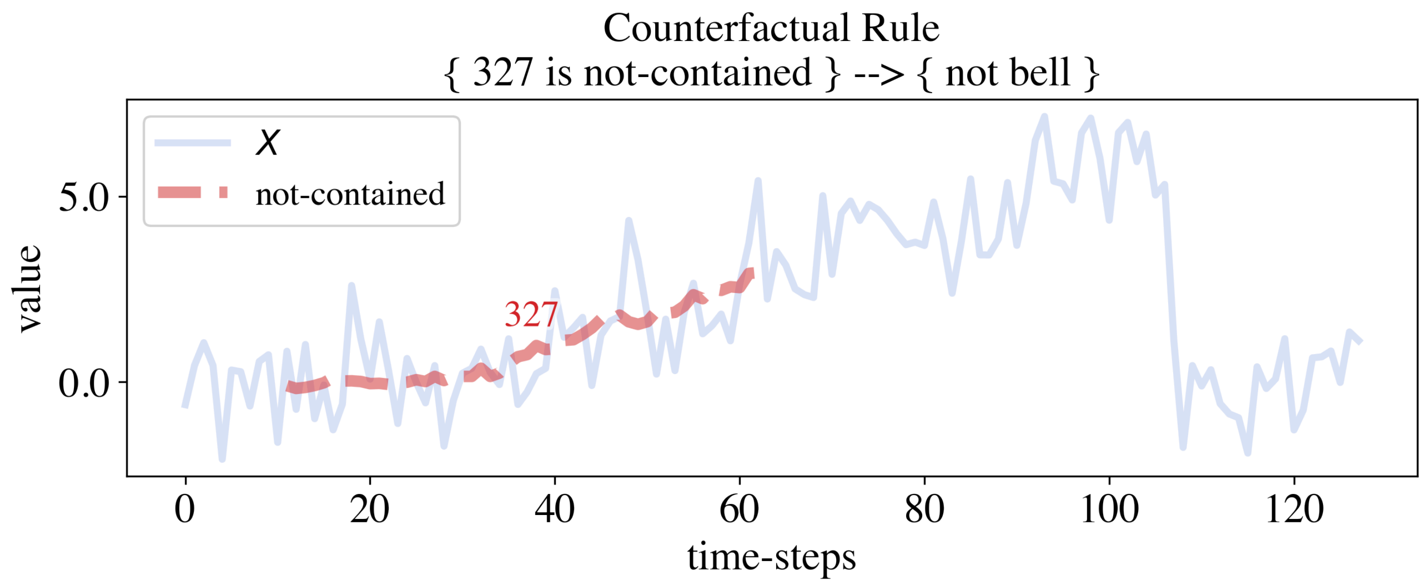
Explains the prediction of the black-box **directly**.

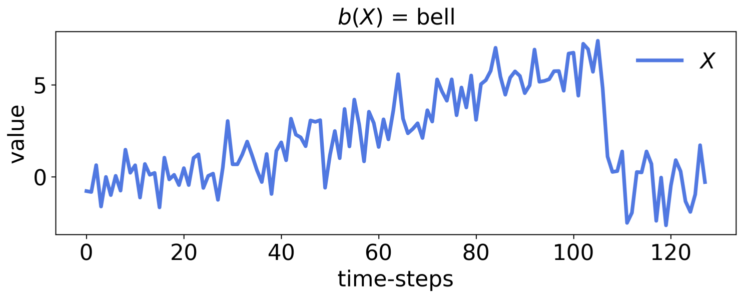




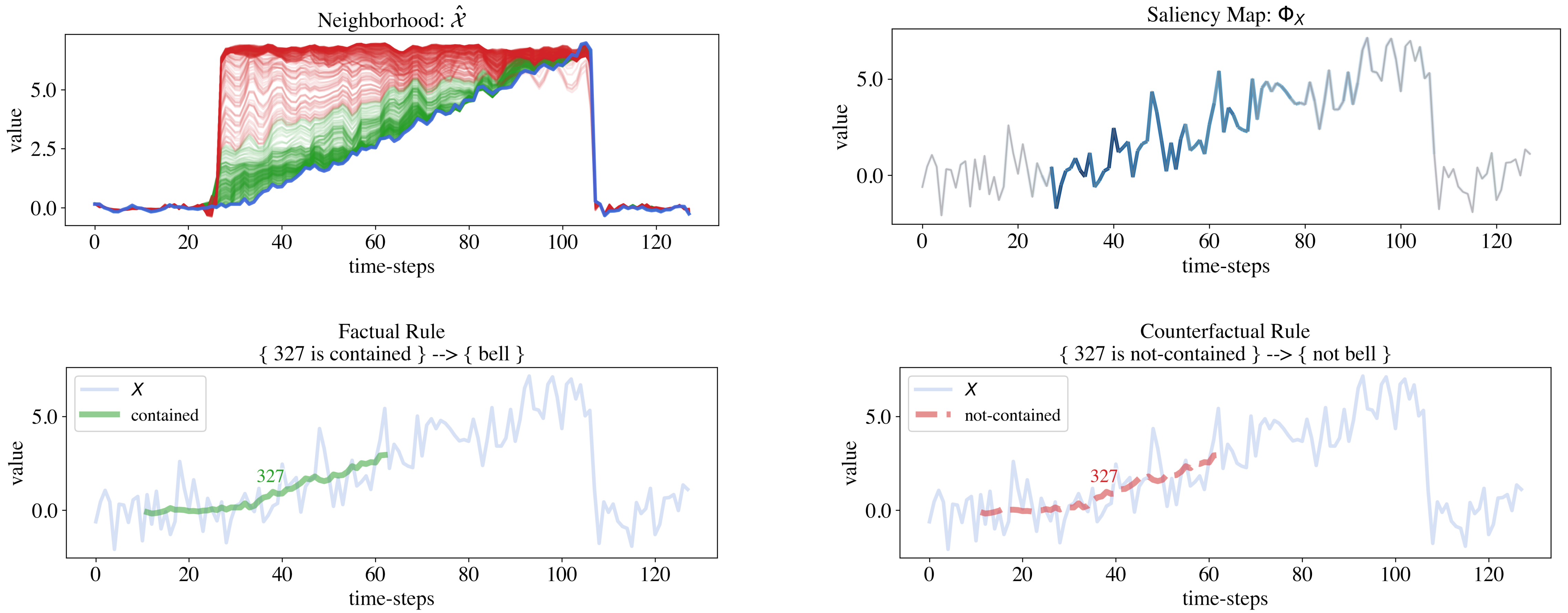
3. Subsequence-based explanation: Counterfactual Rule

Minimal rule variation resulting in a different black-box prediction.

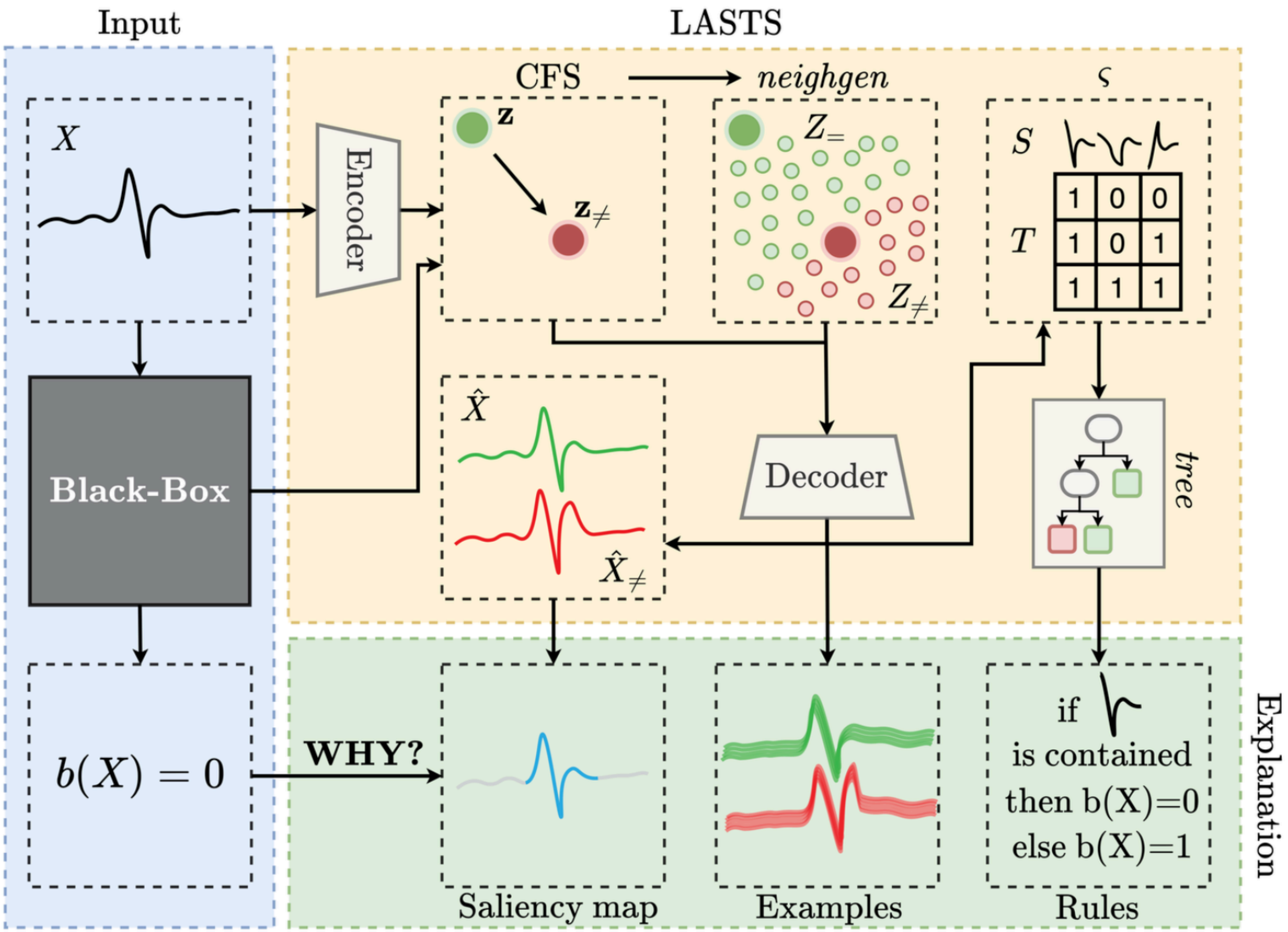




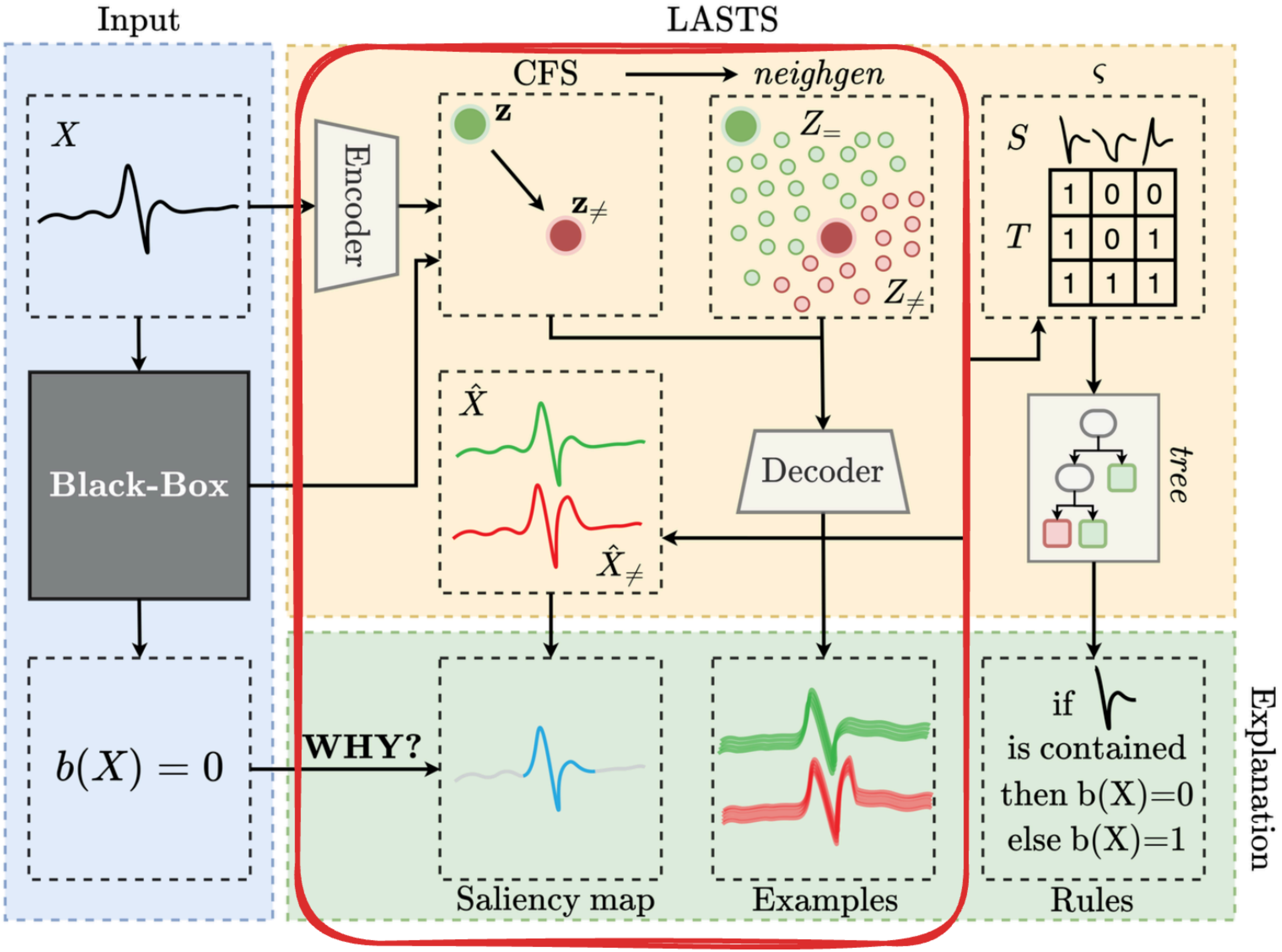
LASTS Explanation



Inside LASTS

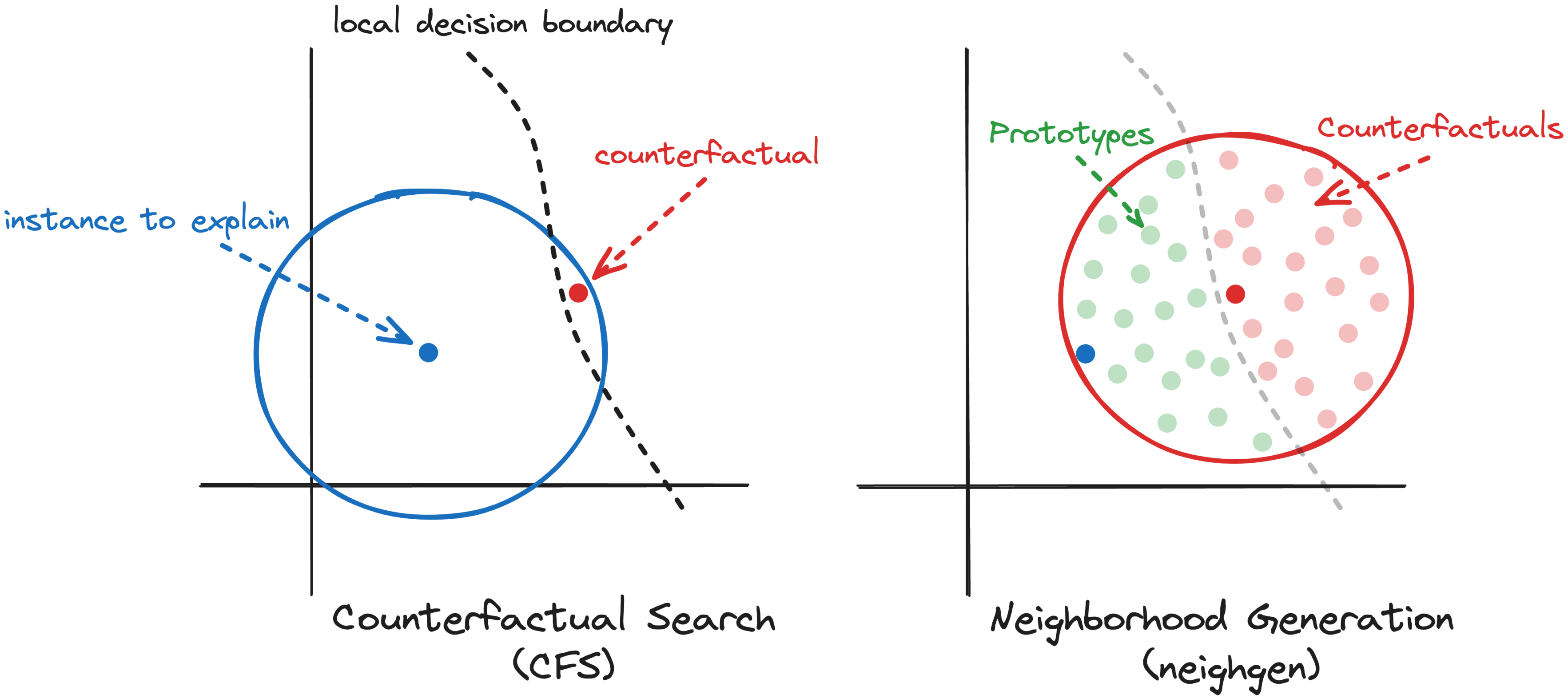


Inside LASTS



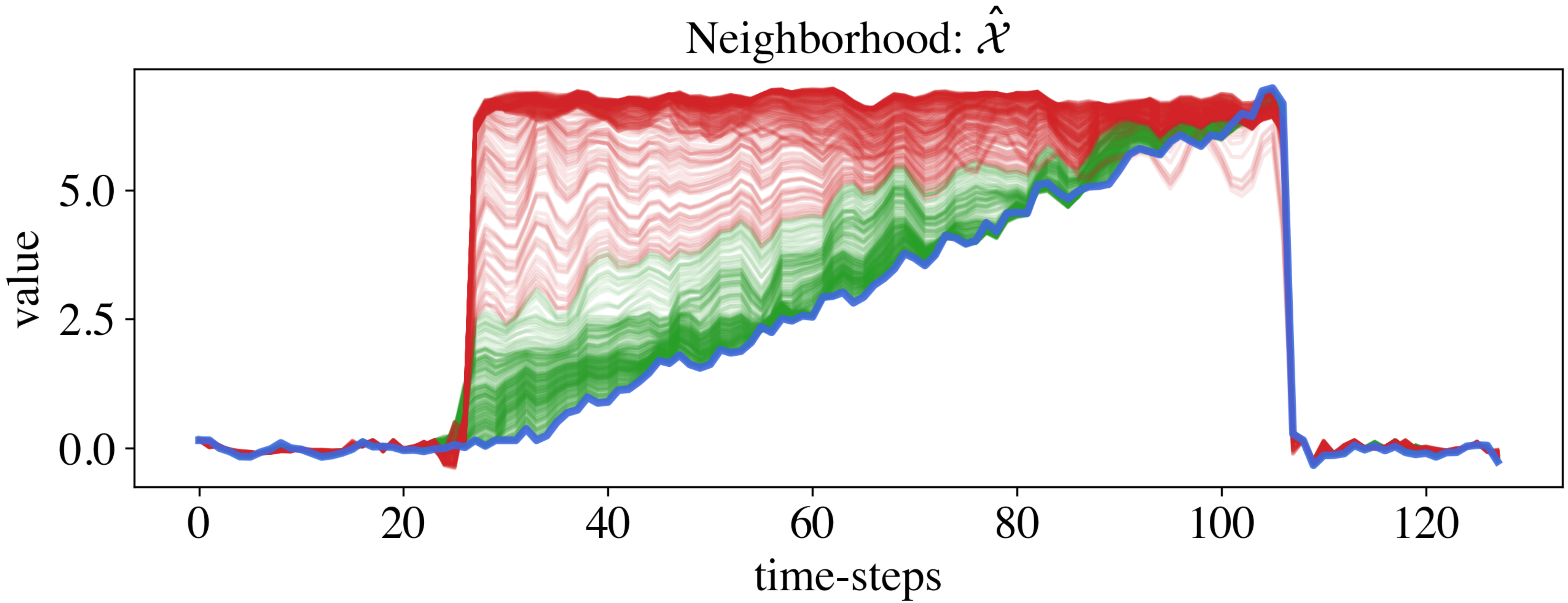
The Latent Space

LASTS uses the latent space of a Variational Autoencoder to **meaningfully** perturb the input time series.

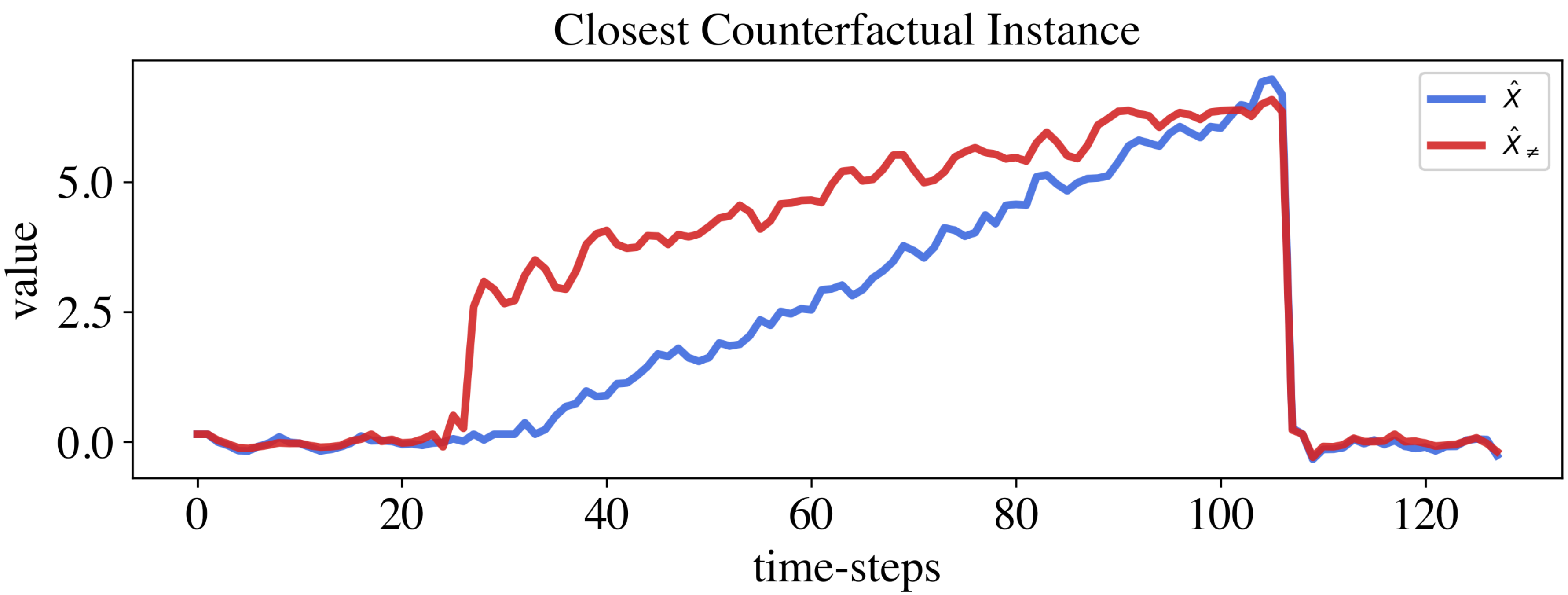


Back to the Time Series Space

LASTS uses the latent space of a Variational Autoencoder to **meaningfully** perturb the input time series.

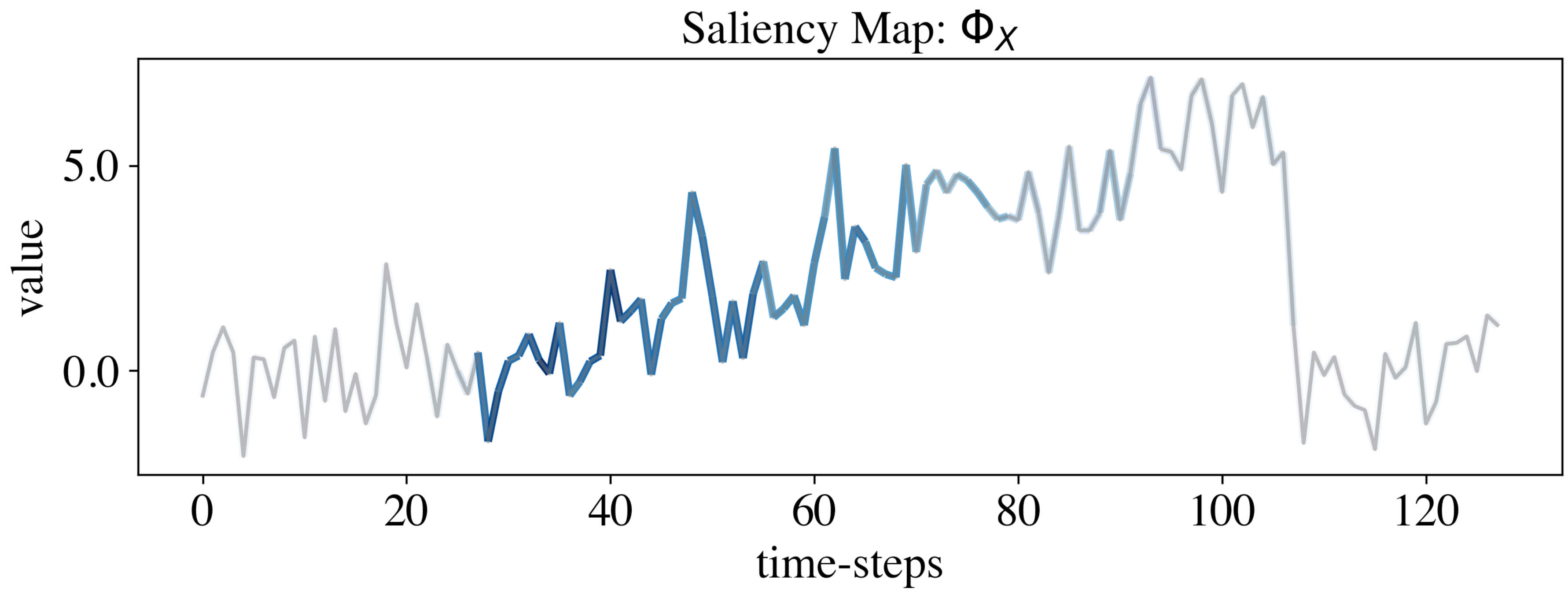


The saliency map is obtained by taking **the absolute difference between the instance to explain and the closest counterfactual.**

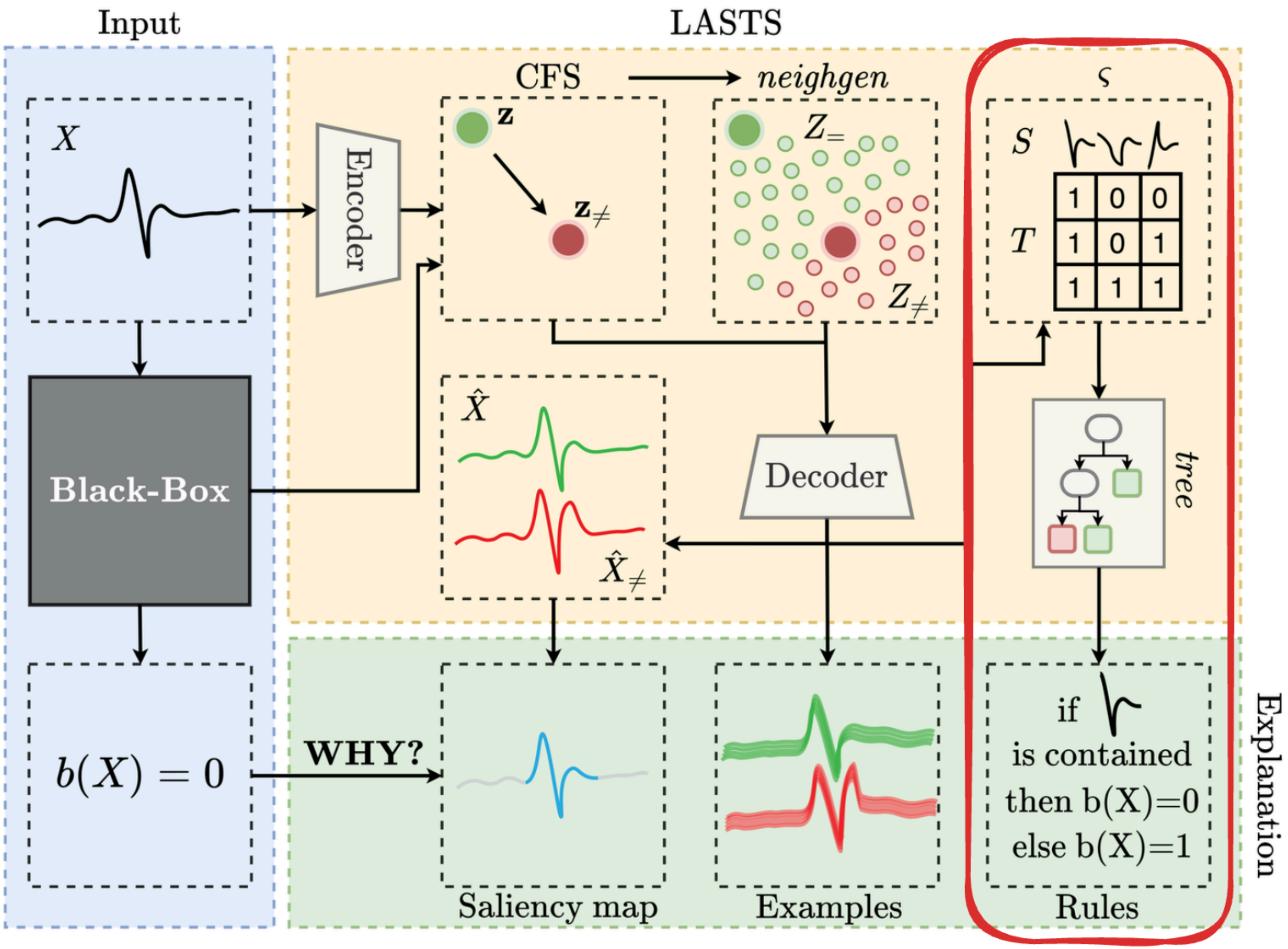


The Saliency Map

The saliency map is obtained by taking **the absolute difference between the instance to explain and the closest counterfactual.**

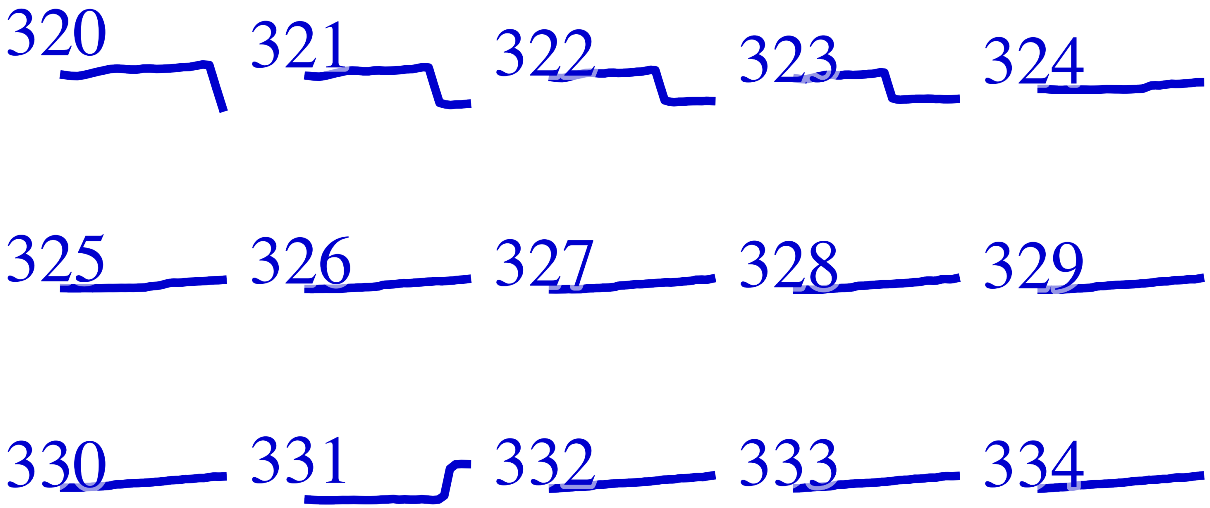


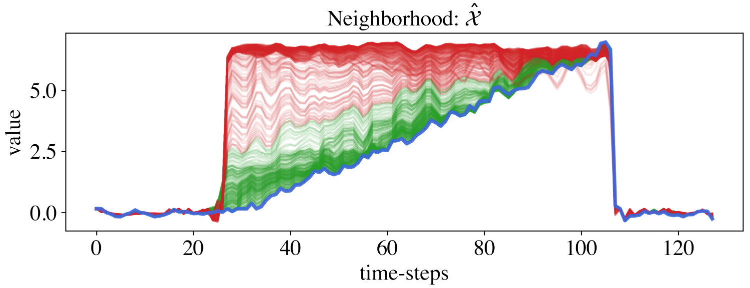
Inside LASTS



The Interpretable Subsequence-based Surrogate

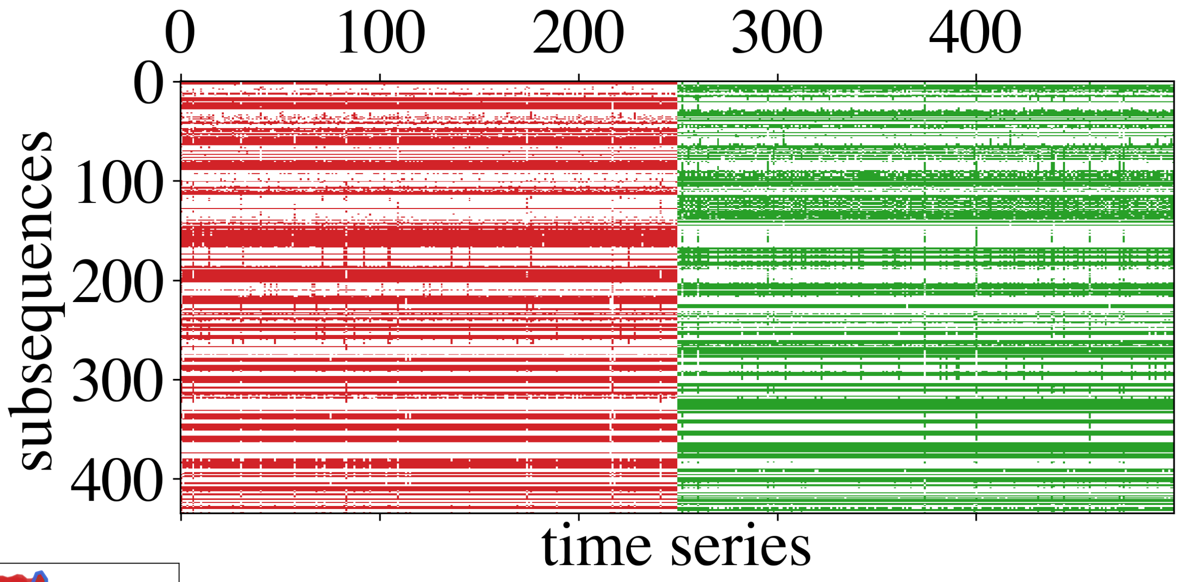
We convert the synthetic neighborhood in a **subsequence-based representation** with a Bag-Of-Patterns/Shapelet Transform.

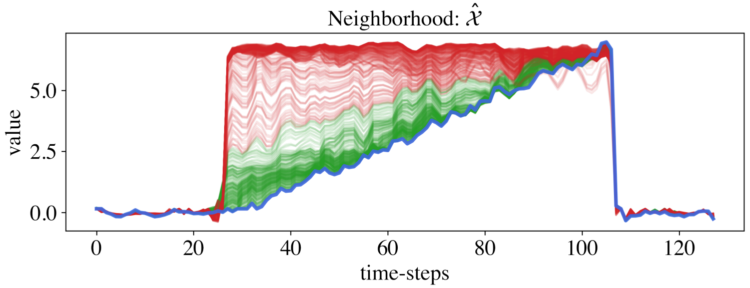




The Interpretable Subsequence-based Surrogate

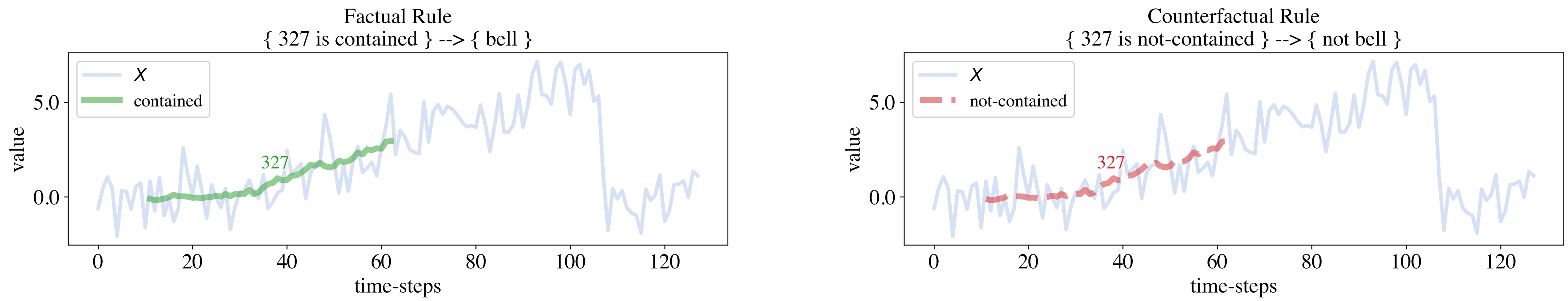
We convert the synthetic neighborhood in a **subsequence-based representation** with a Bag-Of-Patterns/Shapelet Transform.



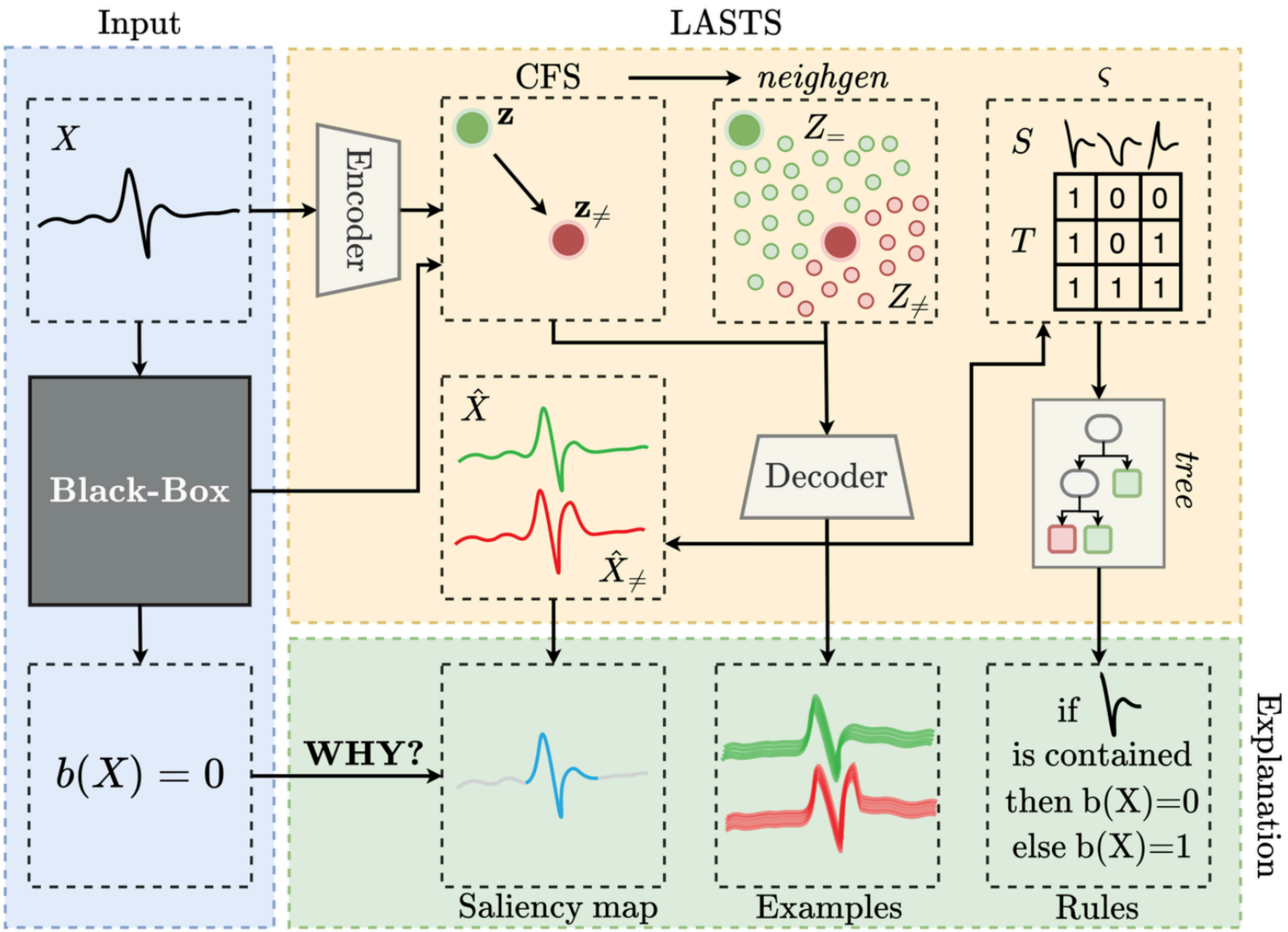


The Interpretable Subsequence-based Surrogate

The resulting dataset is passed to a **decision tree** from which we extract the factual and counterfactual rules.



LASTS



Input
LASTS
CFS
neighgen
ς
X
Encoder
z
z≠
Z=
Z≠
S
T
1 0 0
1 0 1
1 1 1
Black-Box
X̂
X̂≠
Decoder
tree
b(X) = 0
WHY?
if
is contained
then b(X)=0
else b(X)=1
Saliency map
Examples
Rules
Explanation

State-Of-The-Art Comparison

We benchmark the best version of LASTS on **15 datasets**, 10 univariate and 5 multivariate for the UEA/UCR repositories, using **ROCKET*** as black-box.

Each part of the explanation returned by LASTS is evaluated with different metrics and benchmarks.

---

* Dempster, Angus, François Petitjean, and Geoffrey I. Webb. "ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels."Data Mining and Knowledge Discovery 34.5 (2020): 1454-1495.

State-Of-The-Art Comparison

- instance-based $\rightarrow$ usefulness.
- saliency-based (against SHAP*) $\rightarrow$ stability, correctness and insertion/deletion benchmarks.
- rule-based (against a global surrogate and ANCHOR**) → fidelity, precision and coverage.

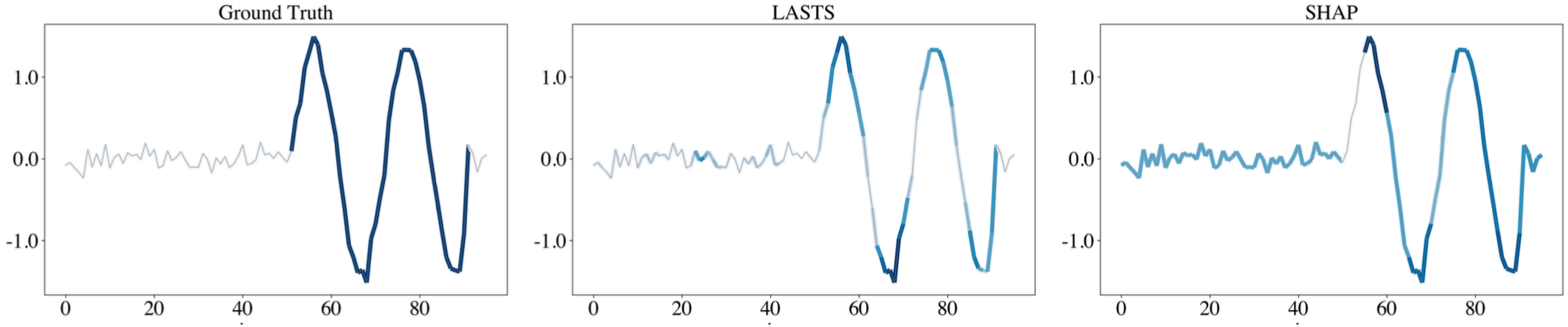
---

* Scott, M., and Lee Su-In. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017): 4765-4774.

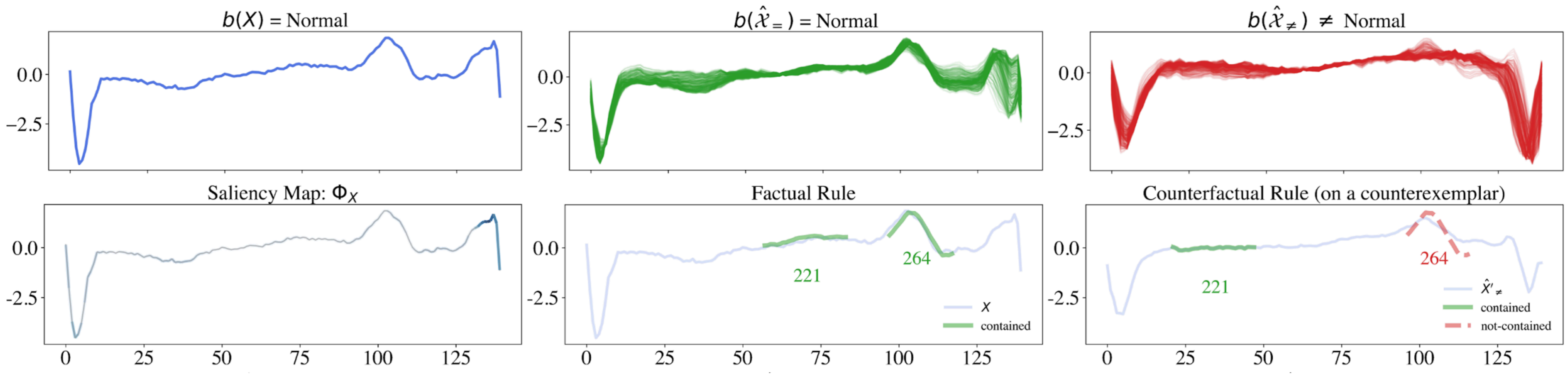
** Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

SHAP Comparison

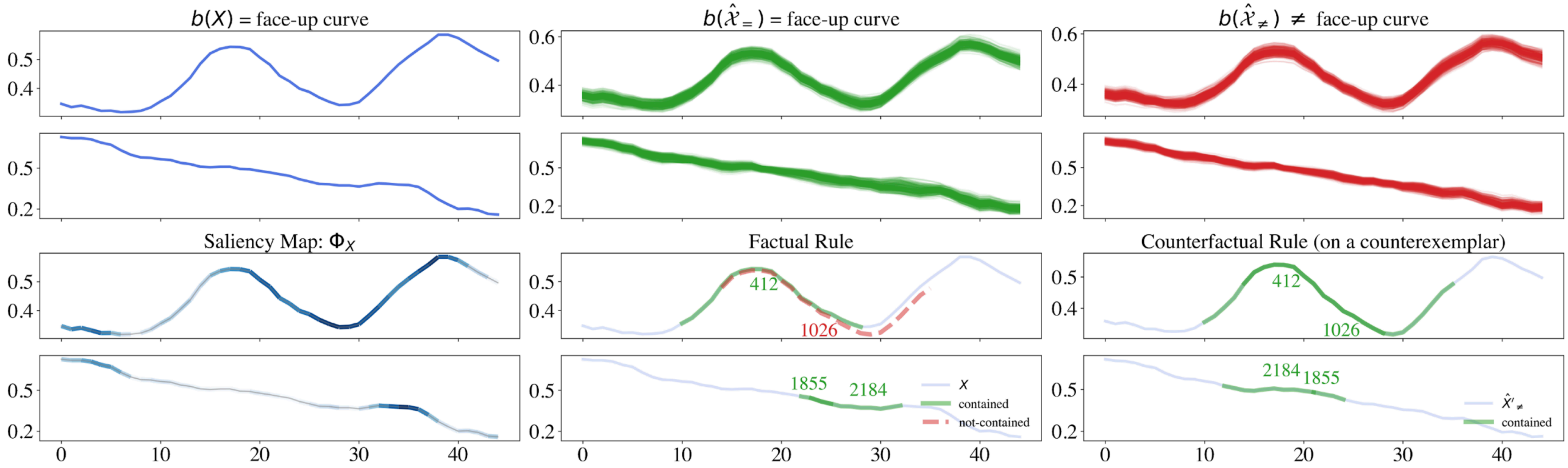
We use a **synthetic experiment** to check if the saliency maps obtained with agnostic explainers match with **custom-defined ground truths** (synthetic classifier on synthetic data).



Explaining ECG5000



Explaining Libras



Pros and Cons

✅

- the autoencoder is powerful...
- diverse explanations

⛔

- ...but cumbersome
- explanations can change depending ony many factors

Embeddings

for Explaining Text Classification

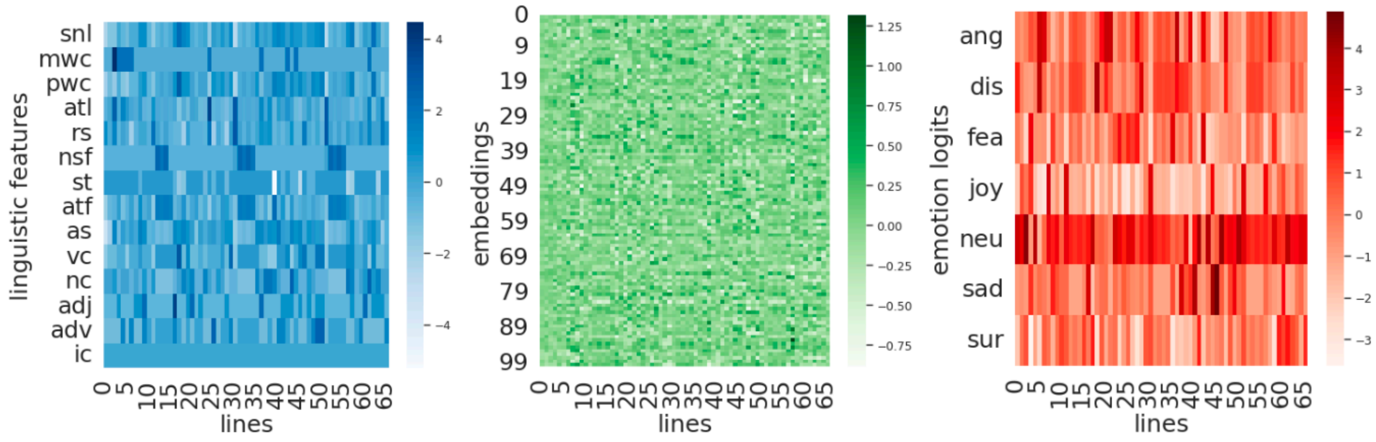
Interpretable Transform ⚠️
Interpretable Model ✅
Mapping ✅

Poggioli, Mattia, Francesco Spinnato, and Riccardo Guidotti. "Text to Time Series Representations: Towards Interpretable Predictive Models." DS 2023.

Text To Time Series

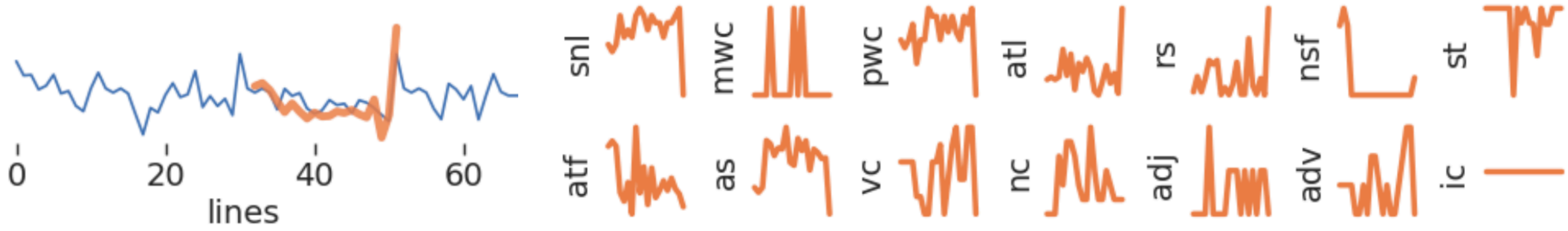
Use time series techniques to explain text classification.

Sometime we do bad, but we all in it / You gotta learn to dream, cause there's No Limit, ya heard me? / - singing / Y'all don't know what we goin through / Y'all don't know what they put us through / Y'all don't know what we goin through / Y'all don't know what they put us through / Don't treat me like a disease, cause my skin darker than yers / And my environment is hostile, nuttin like your suburbs / I'm from the ghetto, home of poverty - drugs and guns / Where hustlers night life for funds but, makin crumbs / in the slums in the street, in the cold in the heat / Rest in peace and then deceased but we still strugglin while you sleep / And the game never change it's still the same since you passed / We get beat and harassed, whenever them blue lights flash / To the little homies in the hood, claimin wards and wearin rags / Tryin to feel a part of a family he never had / And it's sad, I feel his pain, I feel his wants / To avoid bein locked up, there's do's and don'ts / Use your head little soldier, keep the coke out your system / that ? out your veins, that won't do away with the pain / Only prayers will get you through, ain't no use to bein foolish / Ain't got one life to live, so be careful how you use it / - singing / Y'all don't know what we goin through / Y'all don't know what they put us through / Y'all don't know what we goin through / Y'all don't know what they put us through /



Use time series techniques to explain text classification.

Sometime we do bad, but we all in it / You gotta learn to dream, cause there's No Limit, ya heard me? / - singing / Y'all don't know what we goin through / Y'all don't know what they put us through / Y'all don't know what we goin through / Y'all don't know what they put us through / Don't treat me like a disease, cause my skin darker than yers / And my environment is hostile, nuttin like your suburbs / I'm from the ghetto, home of poverty - drugs and guns / Where hustlers night life for funds but, makin crumbs / in the slums in the street, in the cold in the heat / Rest in peace and then deceased but we still strugglin while you sleep / And the game never change it's still the same since you passed / We get beat and harassed, whenever them blue lights flash / To the little homies in the hood, claimin wards and wearin rags / Tryin to feel a part of a family he never had / And it's sad, I feel his pain, I feel his wants / To avoid bein locked up, there's do's and don'ts / Use your head little soldier, keep the coke out your system / that ? out your veins, that won't do away with the pain / Only prayers will get you through, ain't no use to bein foolish / Ain't got one life to live, so be careful how you use it / - singing / Y'all don't know what we goin through / Y'all don't know what they put us through / Y'all don't know what we goin through / Y'all don't know what they put us through /



Conclusion

Open Challenges and Future Works

Data

Beyond simple datasets.

- bigger, more realistic datasets;
- irregular time series;
- missing values;
- multimodel data (sequences + tabular/images);

Models

Beyond single-label predictions towards multi-output models.

- forecasting;
- unsupervised learning:
  - clustering;
  - anomaly detection;
  - generation.

Explanations

Multi-faceted and domain-specific, with evaluation and benchmarks.

- interactive interfaces;
- more tailored explanations (LLMs + RAG?);
- standardization in libraries (`.fit`, `.predict`, `.explain`);
- explanation benchmarks.

THANK YOU FOR THE ATTENTION!

francesco.spinnato@di.unipi.it
https://github.com/fspinna