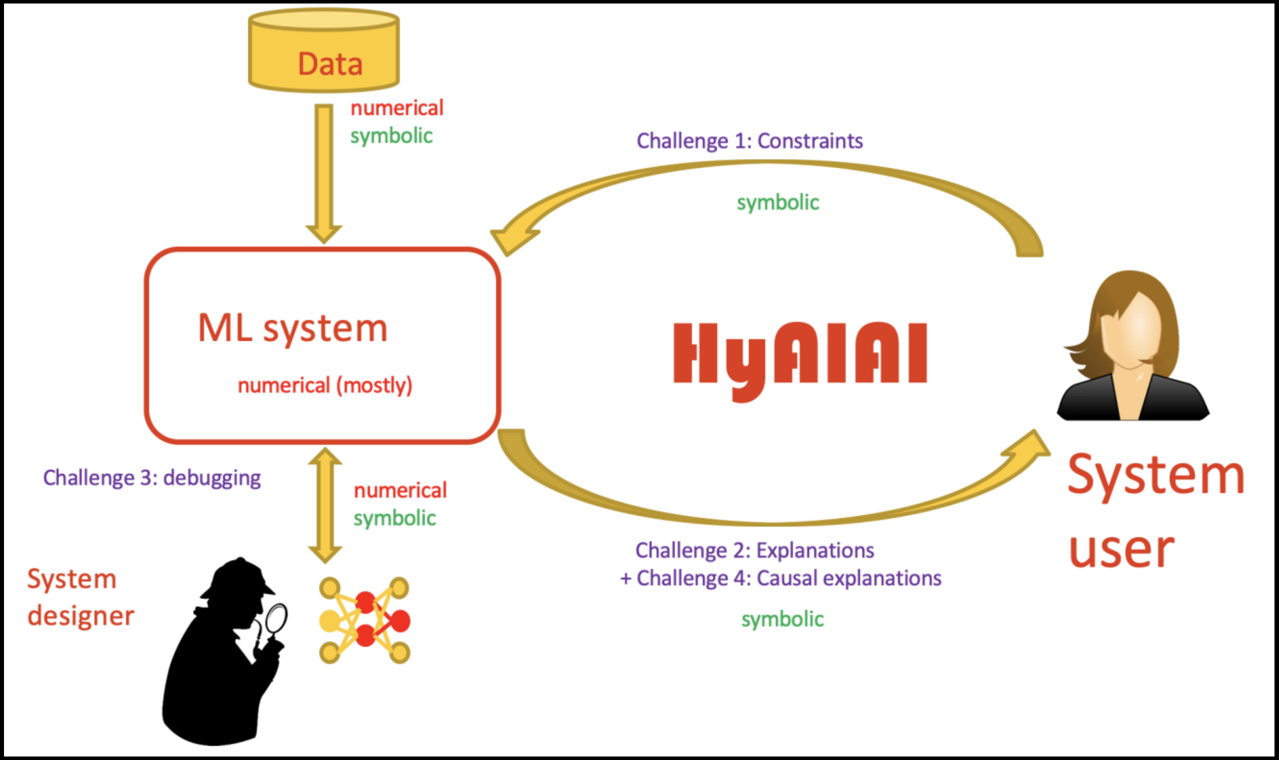
Explainable Time Series Classification

Elisa Fromont
Inria LACODAM team

SIFED 2022
(Symposium International Francophone sur l'Ecrit et le Document)
14/10/2022

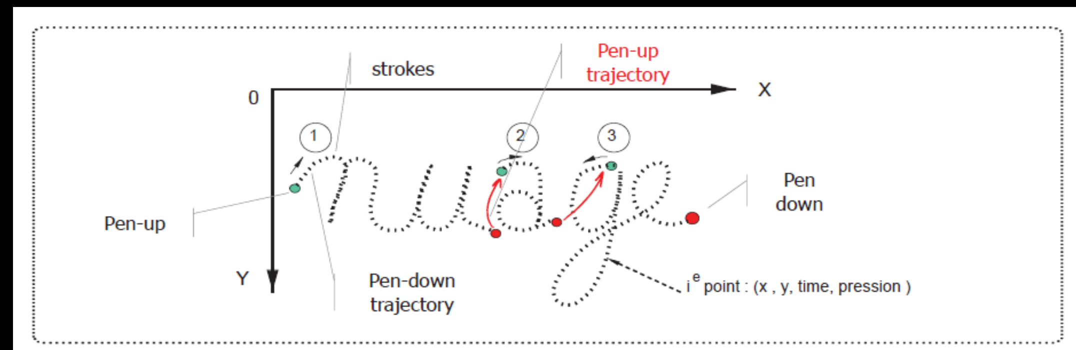
Resources

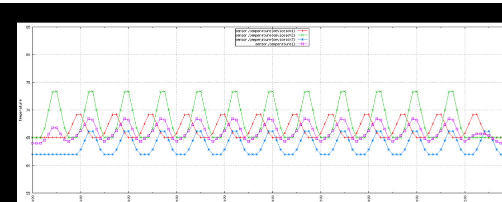
- Somes slides are borrowed from Alessandro Leite & Marc Schoenauer https://project.inria.fr/hyaiai/files/2021/04/slides_explainable_ai_review.pdf
- Hybrid Approaches for Interpretable AI: https://project.inria.fr/hyaiai/
  - Interesting links, Softwares, Publications, …



Why explaining TSC?

- TS are ubiquitous
  - E.g. handwritten documents (online or offline)
- The best TS classifiers are often black boxes
- They can be used for critical decisions: law, medicine, *education*, insurances, ecology, self-driving cars,…





moman → Black box ML model → You wrote « *moman* » 0.734 → Explanation → « You wrote « *moman* » instead of « *maman* », this is because you did not put a long enough downward stroke attached to the circle and there is an ambiguous upward-ish stroke after the letter circle that makes this ressemble a « o ».

TS classifiers?

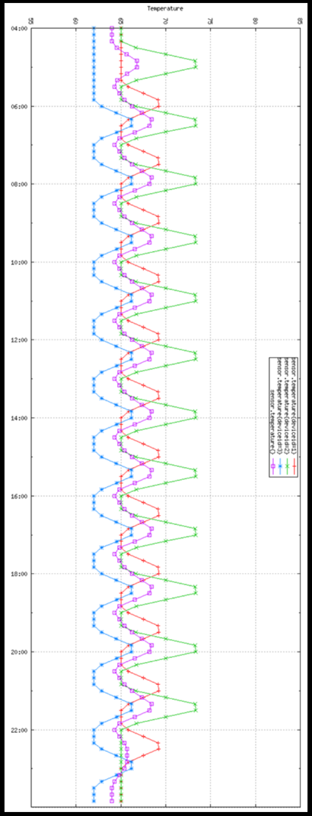
Oct 2021

- For general TS, a common benchmark (univariate and multivariate TS): www.timeseriesclassification.com
  - ROCKET (simple linear classifiers using random convolutional kernels)
    Exceptionally fast and accurate time series classification using random convolutional kernels. Dempster et. al. DAMI 2020
  - INCEPTION TIME (a neural network dedicated to TS)
    InceptionTime: Finding AlexNet for Time Series Classification. Fawaz et. al. DAMI 2019
    TS-CHIEF (a diverse tree ensemble with handcrafted features on random intervals)
    TS-CHIEF: A Scalable and Accurate Forest Algorithm for Time Series Classification. Shifaz et. al. DAMI 2020

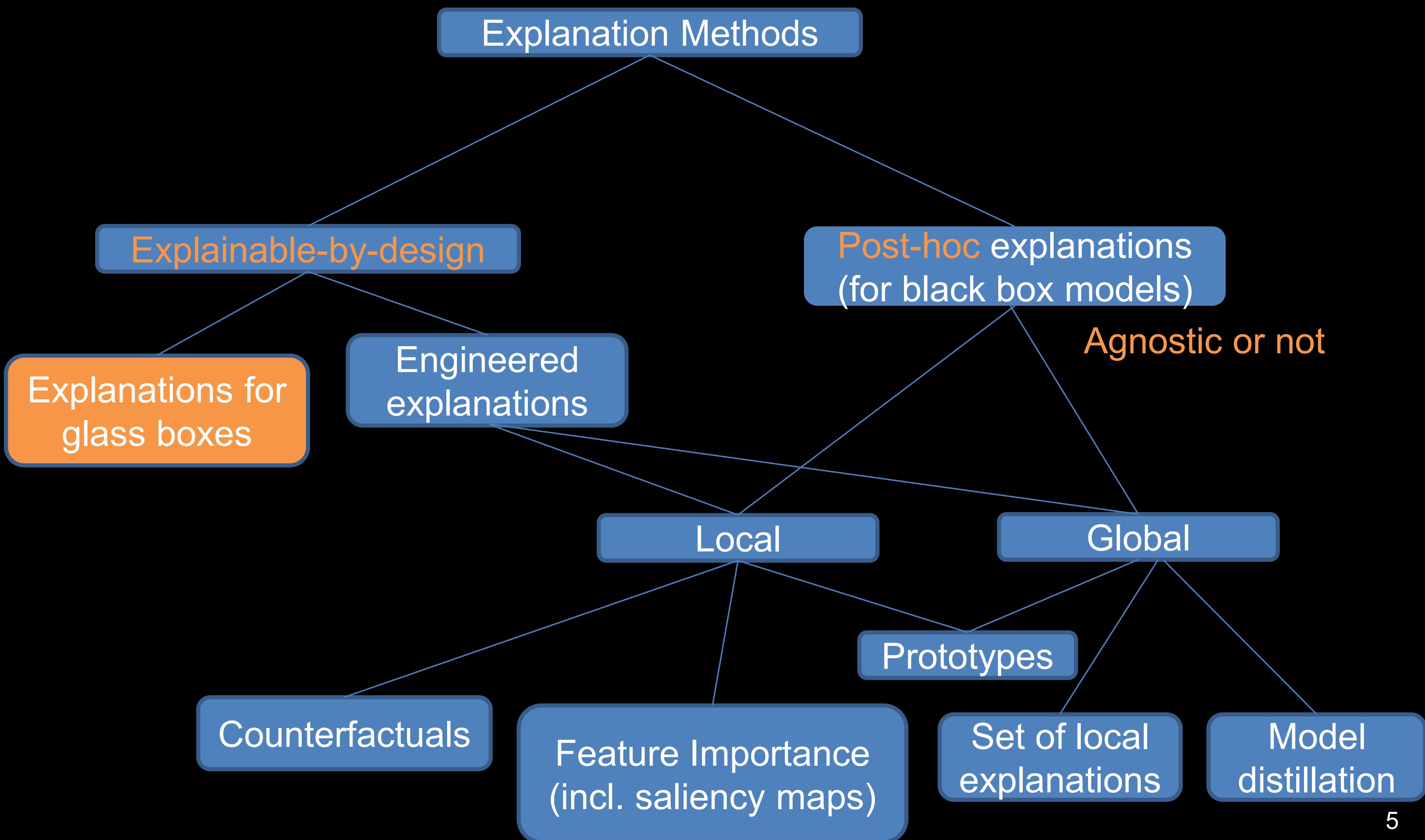
– VAN (encoder/decoder FCN + LSTM) *Denis Coquenet, Clément Chatelain, and Thierry Paquet. “End-to-end HandwrittenParagraph Text Recognition Using a Vertical Attention Network”. In:Transactionson Pattern Analysis and Machine Intelligence (TPAMI)(2022).*

– Vision Transformer: *Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, XiaohuaZhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “An Image is Worth 16x16 Words:Transformers for Image Recognition at Scale”. In:9th International Conference onLearning Representations (ICLR). 2021.*

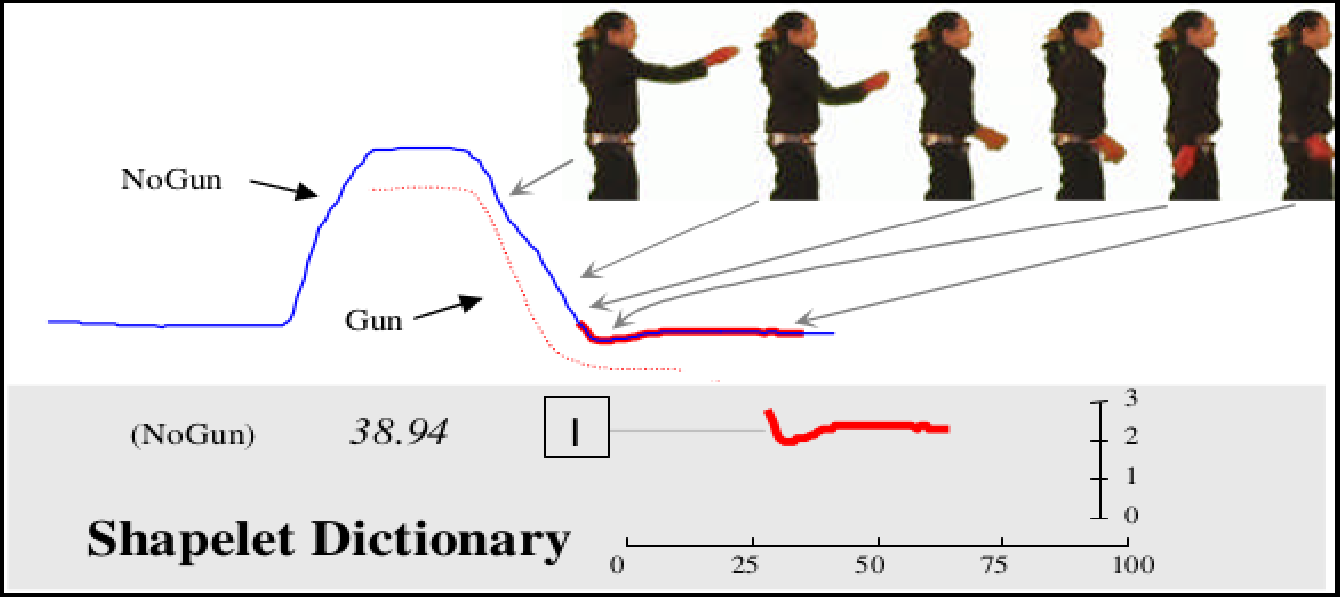
Black boxes



Taxonomy of explanation methods



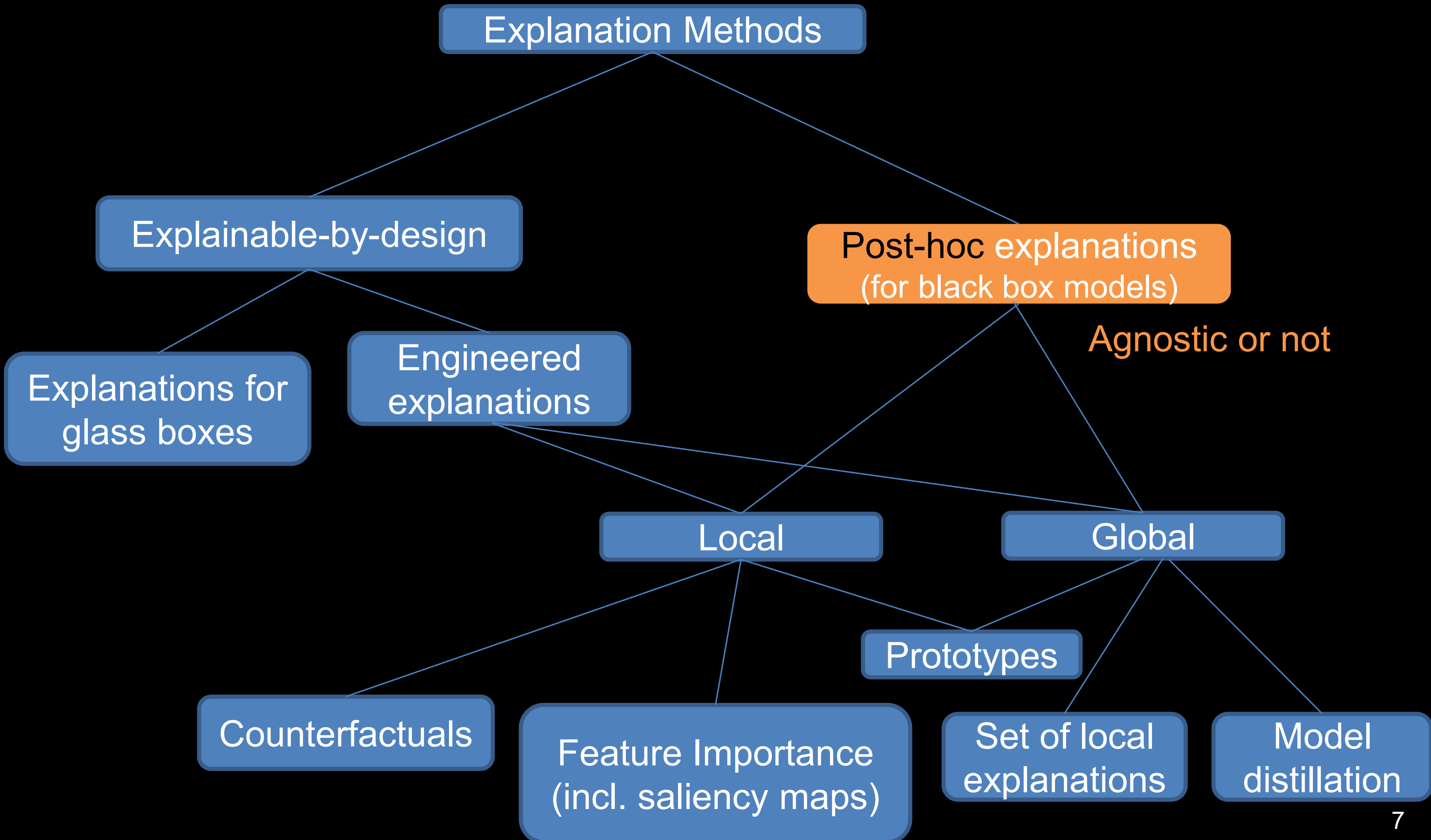
A glass box for TS classification



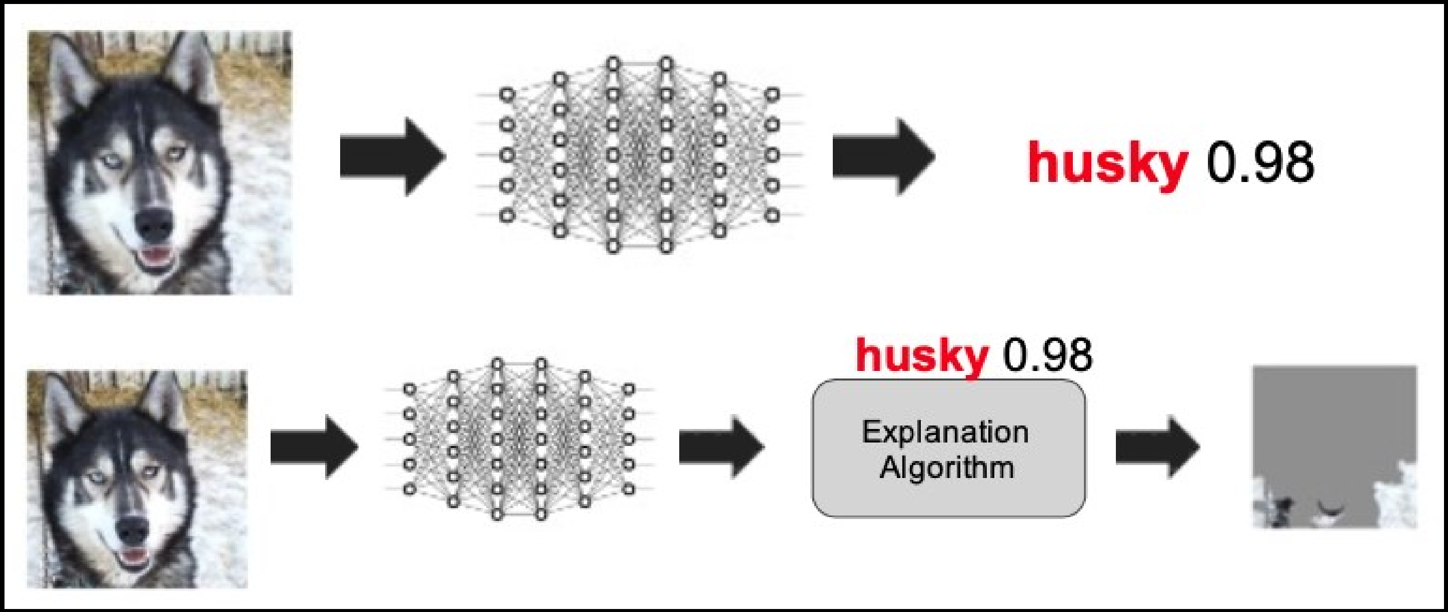
Both features and models are interpretable
… but inefficient (shapelet enumeration) and inaccurate (too simple ?)

L. Ye, E. Keogh. **Time Series Shapelets: A New Primitive for Data Mining**, KDD 2009

Taxonomy of explanation methods



Post-hoc explanations



Local vs Global

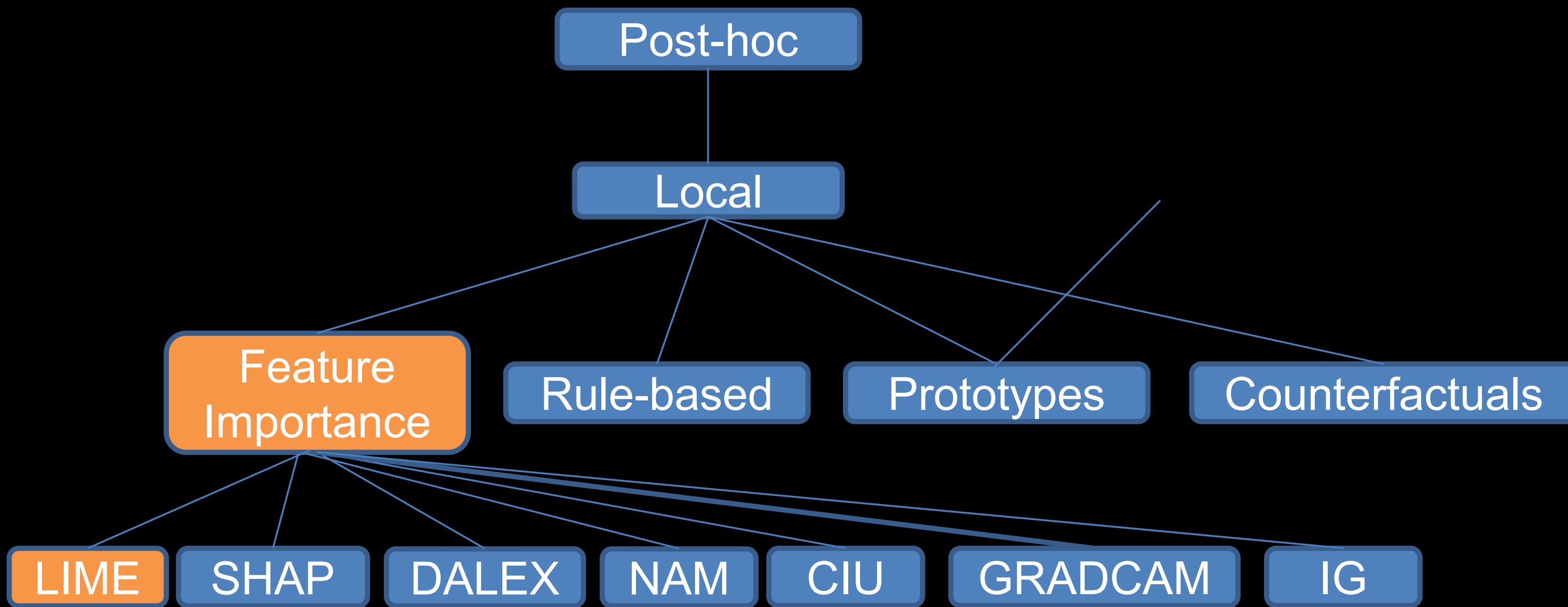
LOCAL

- Explain individual predictions
  - Help in unearthing biases in the neighbourhood of a given sample
  - Help in checking out if individual predictions are correctly being made

GLOBAL

- Explain the behaviour of a model (e.g. for each class)
  - Highlight biases affecting larger subgroups
  - Help in determining if the model is “ready” for deployment

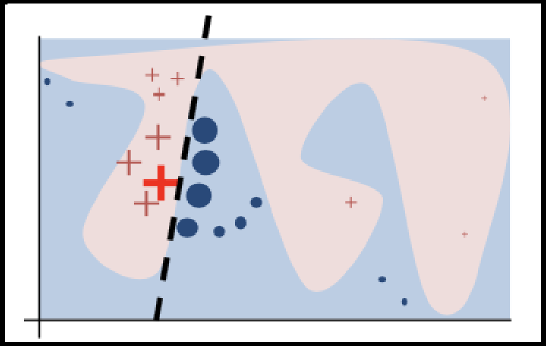
(local) Post-hoc explainability Feature importance



[LIME] Tulio et. al. "Why should i trust you?" Explaining the predictions of any classifier". In: ACM SIGKDD 2016
[DALEX] P. Biecek and T. Burzykowski. ***Explanatory Model Analysis***. Chapman and Hall/CRC, New York, 2021.
[NAM] Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, Geoffrey E. Hinton. ***Neural Additive Models: Interpretable Machine Learning with Neural Nets***. NeurIPS 2021
[CIU] S. Anjomshoae, K. Främling and A. Najjar . ***Explanations of Black-Box Model Predictions by Contextual Importance and Utility*** International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems 2019

Local Interpretable Model-Agnostic Explanations (LIME)

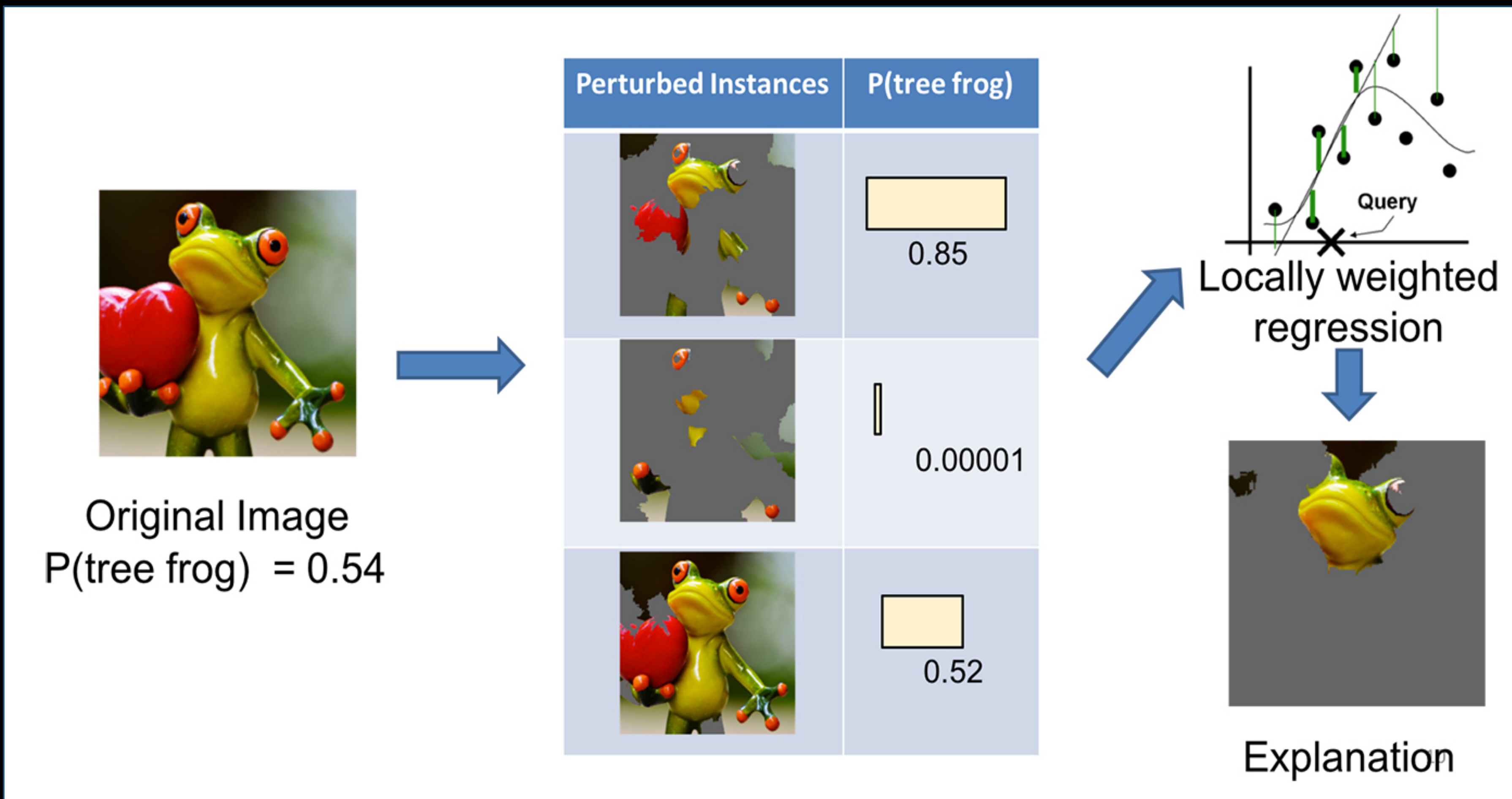
- Model agnostic explanation method based on (some) feature importance
- Draw a perturbed sample of weighted instances $\{z \in R^d\}$ around a point x_i by exploiting a proximity measure π_x
- Fed them to the black-box model b(z) to predict the output for each sample
- Weights the samples according to the distance to x_i
- Train an explanation model g(·): sparse linear model on the weighted samples
- Use g(·) to explain. The explanations are the weights of the linear model



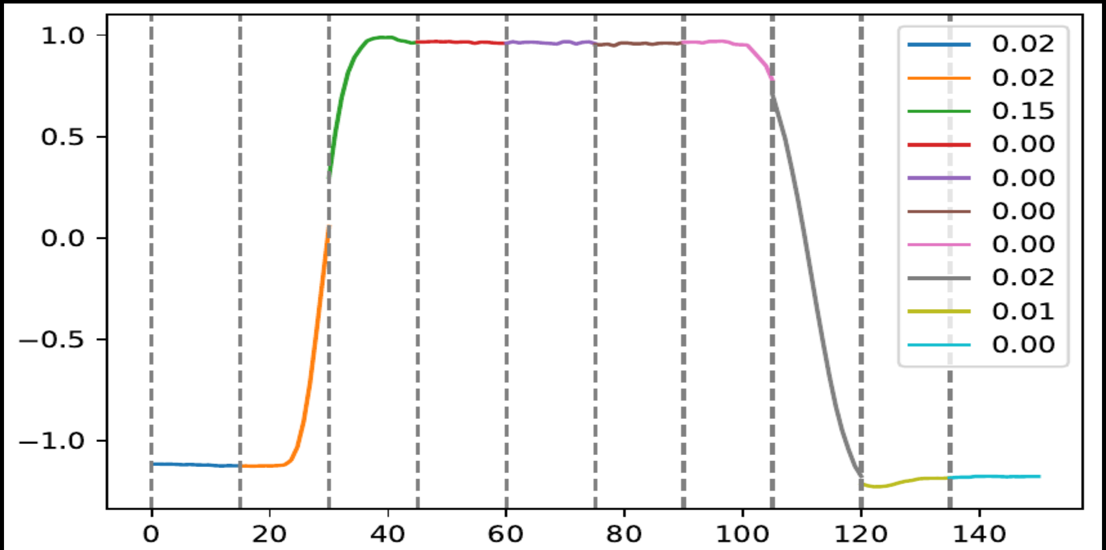
There are (a lot of) variants to overcome LIME's limitations: KL-LIME, s-LIME, DLIME, ILIME, ALIME, ….

[LIME] Tulio et. al. "Why should i trust you?" Explaining the predictions of any classifier". In: ACM SIGKDD 2016

Example with LIME



LIME for TS: LEFTIST

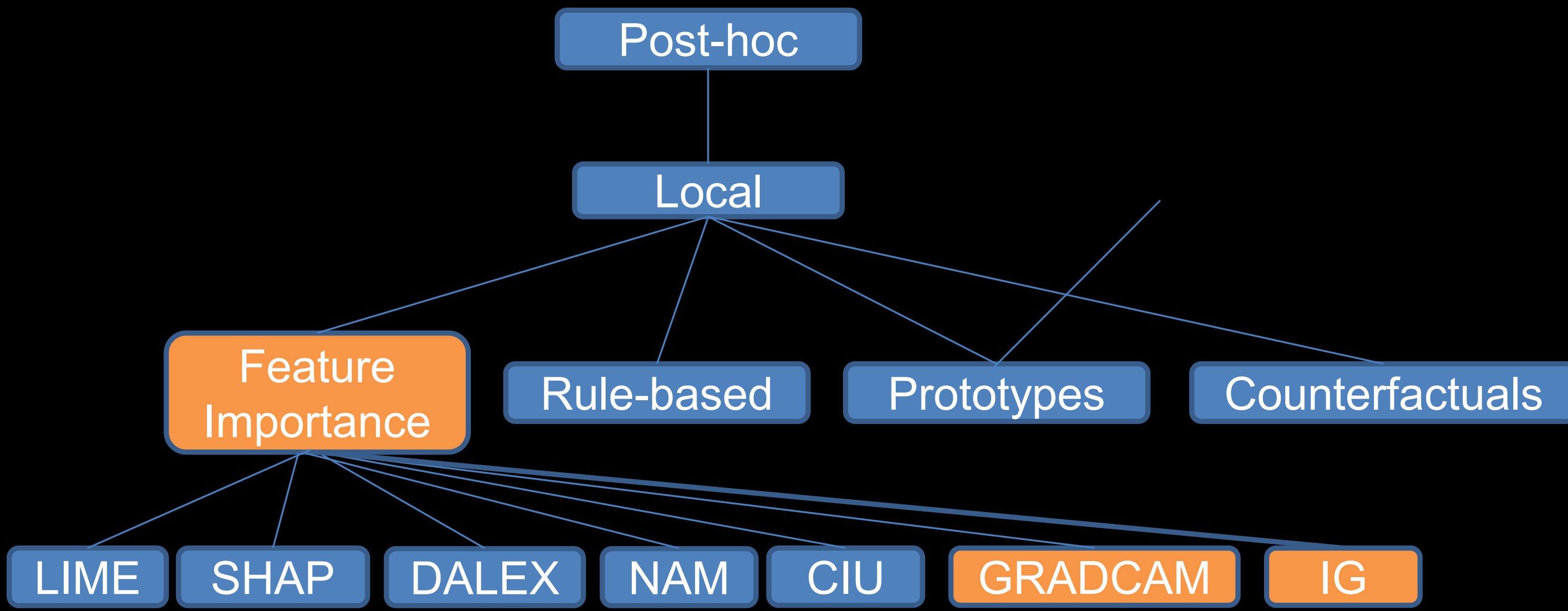


LEFTIST explication for an SVM classifier for a time series belongs to one class (GunPoint dataset)

Language = segment of a TS (similar to shapelets)

[LEFTIST] M. Guillemé; V. Masson; L. Rozé; Al Termier. ***Local Explainer For TIme Series classification***, ICTAI 2019

(local) Post-hoc explainability
Feature importance



[GRADCAM] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in ICCV, 2017, pp. 618–626.

[Integrated Gradient] M. Sundararajan, A. Taly, Q/Yan: **Axiomatic Attribution for Deep Networks**. ICML 2017: 3319-3328

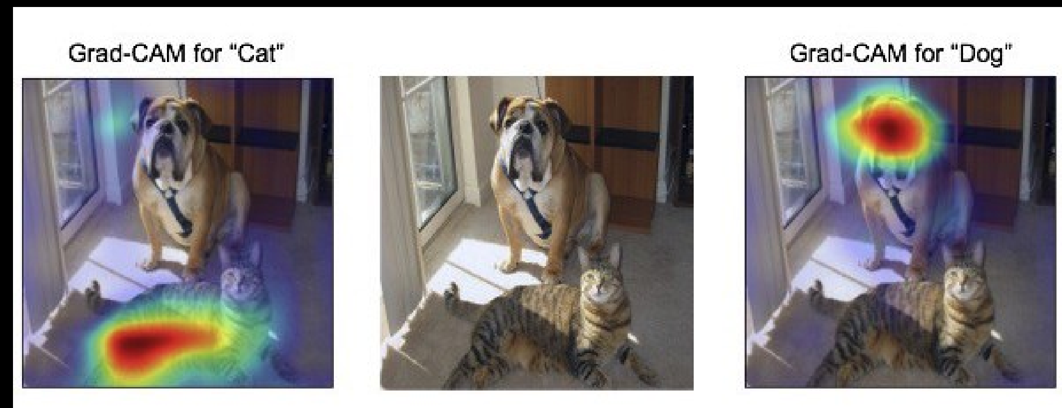
[DALEX] P. Biecek and T. Burzykowski. ***Explanatory Model Analysis***. Chapman and Hall/CRC, New York, 2021.

[NAM] Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, Geoffrey E. Hinton. ***Neural Additive Models: Interpretable Machine Learning with Neural Nets***. NeurIPS 2021

[CIU] S. Anjomshoae, K. Främling and A. Najjar . ***Explanations of Black-Box Model Predictions by Contextual Importance and Utility*** International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems 2019

Saliency Maps

Heat map on the original input of the neural network (not agnostic) to highlight the regions (features) important for the neural network to take a decision.

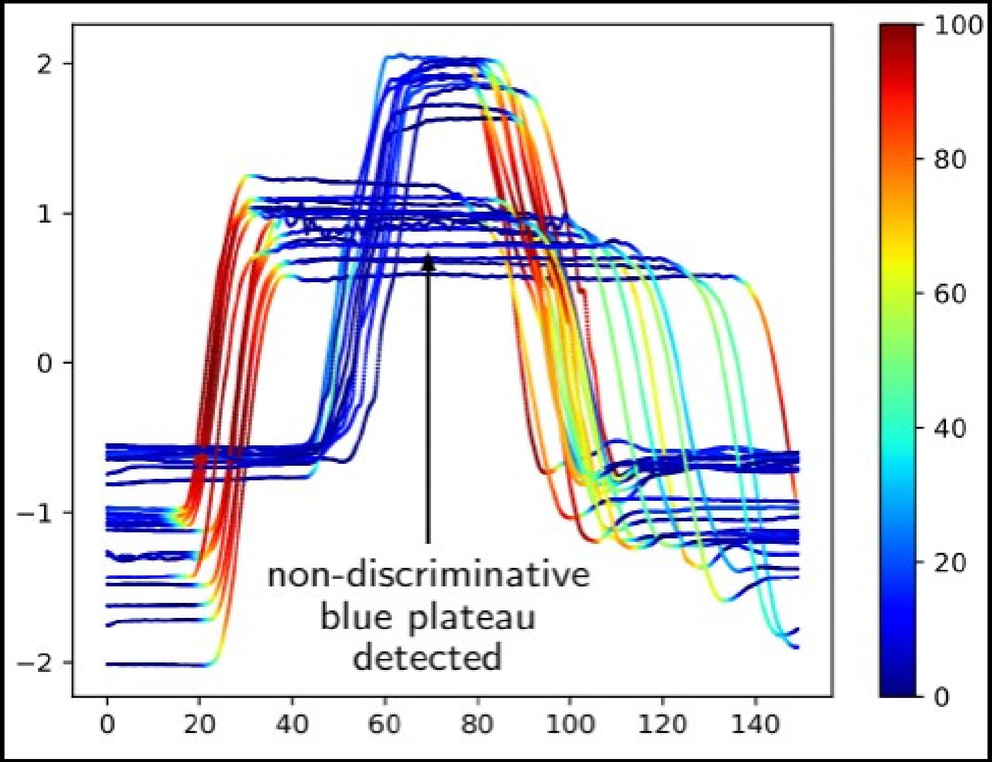


$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

[GRADCAM] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in ICCV, 2017, pp. 618–626.
[Integrated Gradient] Mukund Sundararajan, Ankur Taly, Qiqi Yan: Axiomatic Attribution for Deep Networks. ICML 2017: 3319-3328

GradCAM on univariate TS

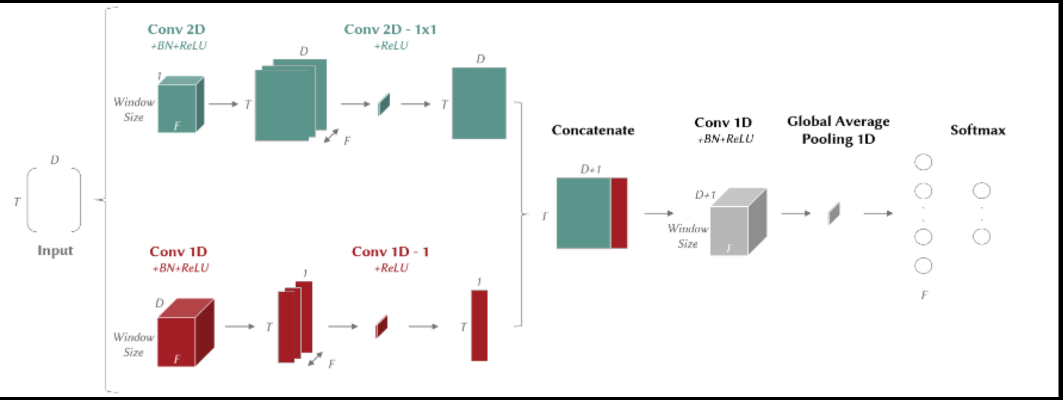


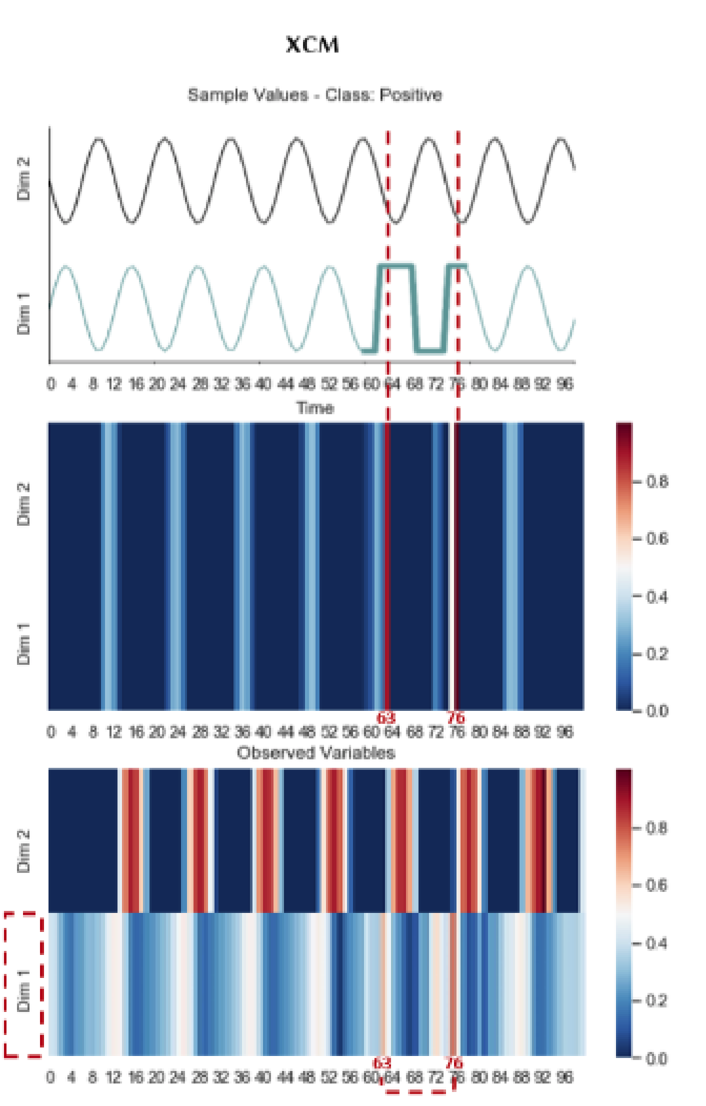
GRADCAM on a NN to classify examples of the Gunpoint dataset

H.Fawaz, G. Forestier, J. Weber, L. Idoumghar, P. Muller. ***Deep learning for time series classification: a review.*** DAMI 2019

GradCam on Multivariate TS

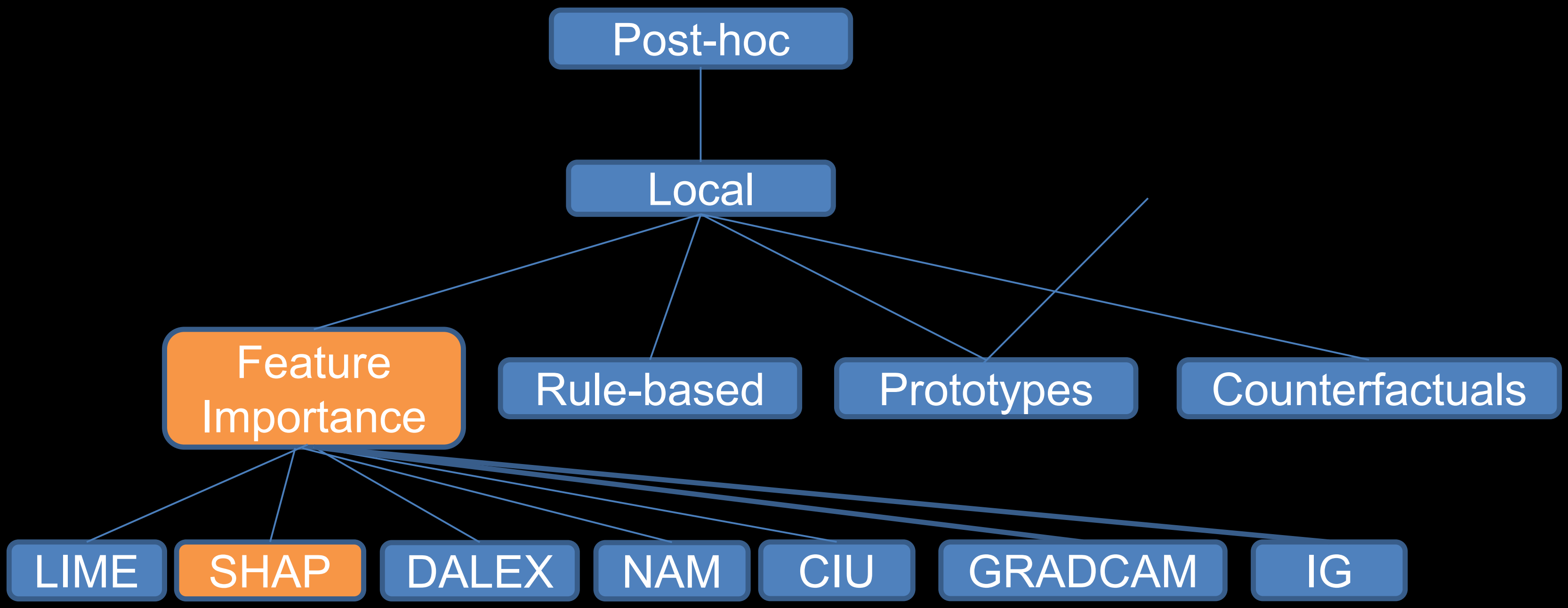
Design a multivariate TS classifier which separate explicitly 1D convolutions (time) and 2D convolutions (variables)





K. Fauvel, T. Lin, V. Masson, E. Fromont, and A. Termier. Mathematics 2021. **XCM: An Explainable Convolutional Neural Network for Multivariate Time Series Classification**

(local) Post-hoc explainability
Feature importance



[SHAP]S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

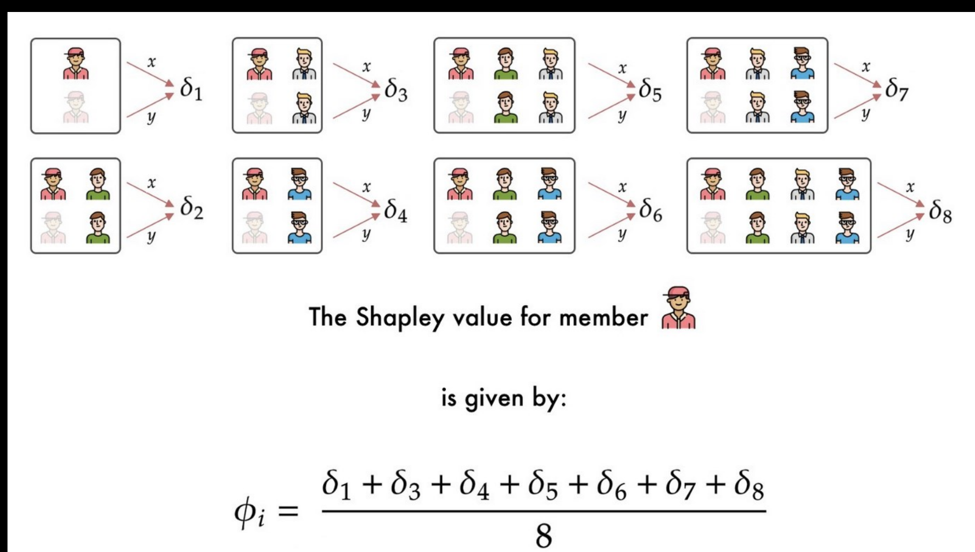
[DALEX] P. Biecek and T. Burzykowski. ***Explanatory Model Analysis***. Chapman and Hall/CRC, New York, 2021.

[NAM] Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, Geoffrey E. Hinton. ***Neural Additive Models: Interpretable Machine Learning with Neural Nets***. NeurIPS 2021

[CIU] S. Anjomshoae, K. Främling and A. Najjar . ***Explanations of Black-Box Model Predictions by Contextual Importance and Utility*** International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems 2019

Shapley Values [L. Shapley 1953]

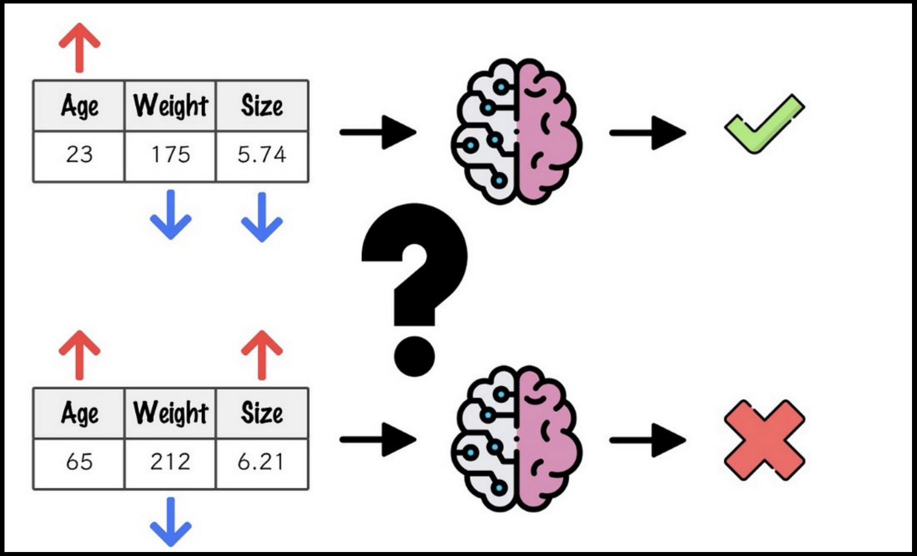
- Shapley values are a concept of the *cooperative game theory* field, whose objective is to measure each player's contribution to the game
- Context where "*n*" players participate collectively obtaining a reward "*p*" which is intended to be fairly distributed at each one of the "*n*" players according to the individual contribution, such a contribution is a Shapley value = average marginal contribution of an instance of a feature among all possible coalitions.



δi = marginal Contribution of member "A" to the coalition of members B, C, D.

https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

SHapley Additive exPlanations (SHAP)

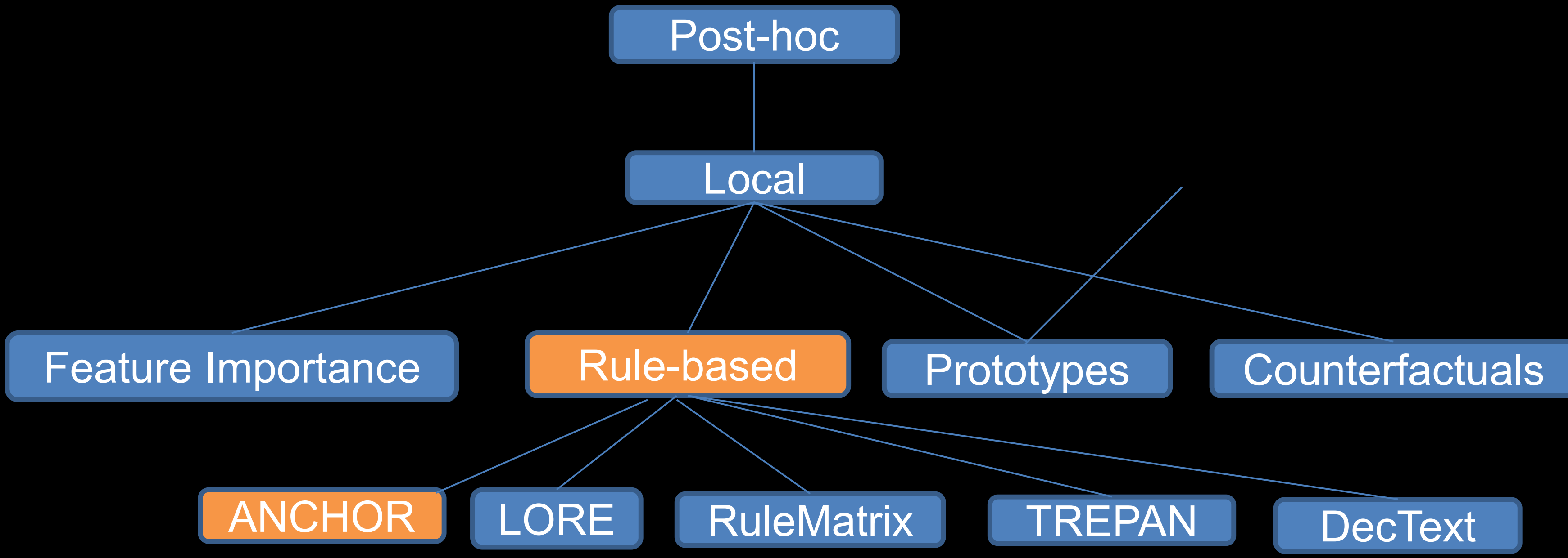


Local and global model-agnostic explanation method

[SHAP]S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

- Calculating Shapley values for each instance of each feature is NP-hard ➔ Kernel Shap (much fewer coalition samples)
- Kernel Shap is based on a weighted linear regression where the coefficients of the solution are the Shapley values
- SHAP is directly usable for time series provided that the time series has already been "represented" with interpretable features

Post-hoc explainability
Rule-based methods



[ANCHORS] Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin . **Anchors: High-Precision Model-Agnostic Explanations**. AAAI 2018: 1527-1535

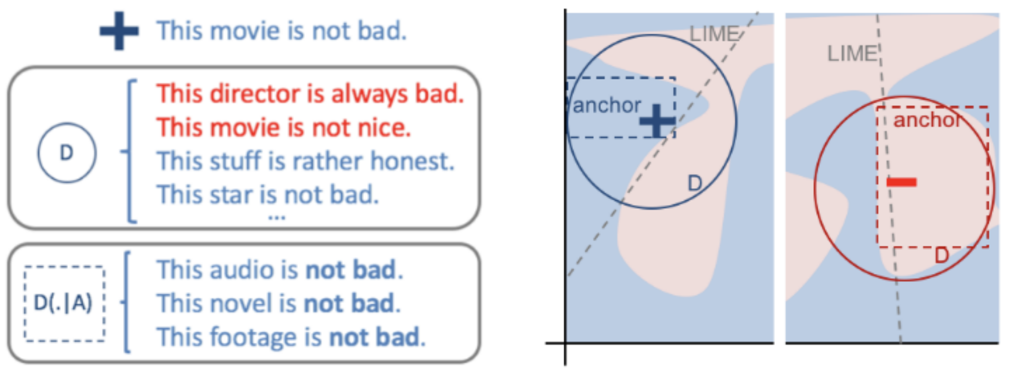
[LORE] Guidotti et. al.. **Factual and Counterfactual Explanations for Black Box Decision Making** ”. **IEEE Intelligent Systems Vol: 34(6), 2019**

[RuleMatrix] Ming. Et. al **RuleMatrix: Visualizing and Understanding Classifiers with Rules**, IEEE Transactions on Visualization and Computer Graphics 2018

[TREPAN] Craven et. al..**Extracting tree-structured representations of trained networks. NIPS 1996**

[DecText]Boz, **Extracting decision tree from trained neural networks**, **KDD 2002**

Anchors: High-Precision Model-Agnostic Explanations

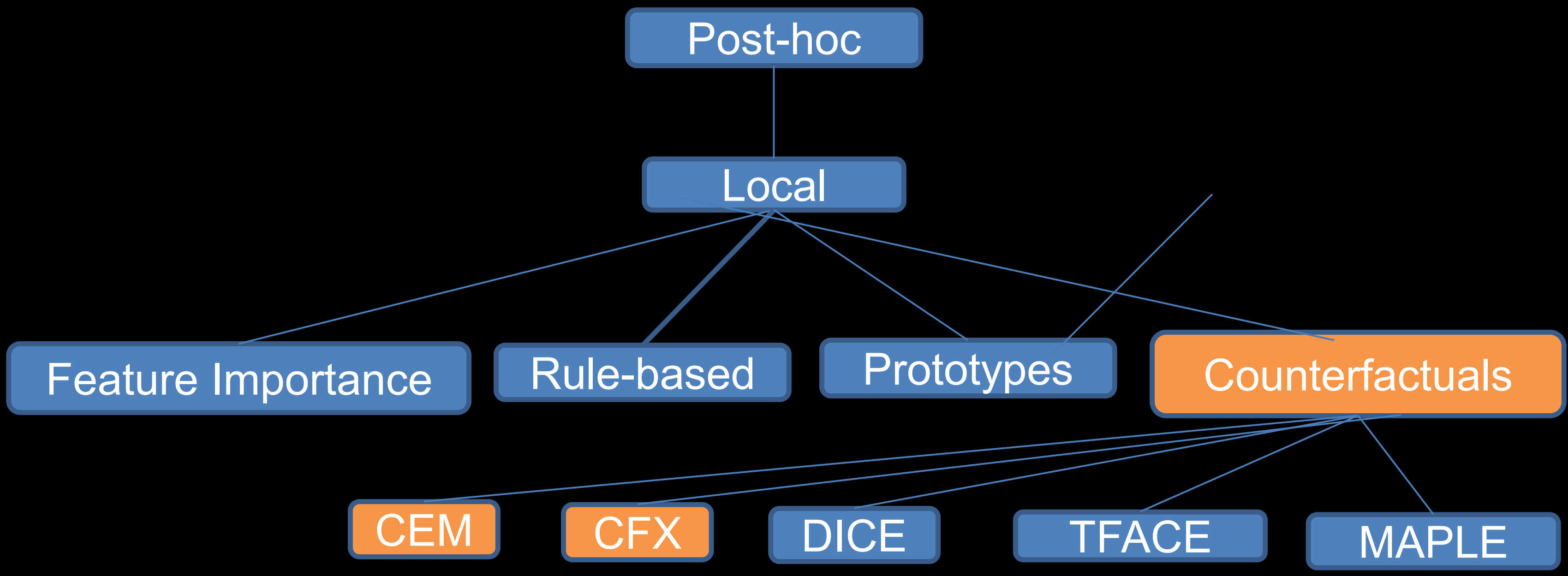


(a) $\mathcal{D}$ and $\mathcal{D}(.|A)$ (b) Two toy visualizations

- Model agnostic rule-based explanation method
- Anchor explanation = rule that sufficiently "anchors" the prediction locally – such that changes to the rest of the feature values of the instance do not matter
- Given a sample x_i, r is an anchor if $r(x_i) = b(x)$
- Build a perturbed sample from x_i
- Extract all anchors with precision greater than a defined threshold
- Employs a multi-armed bandit algorithm
- Uses a bottom-up and beam search to explore the anchors

Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018: 1527-1535

Post-hoc explainability
Counterfactuals methods

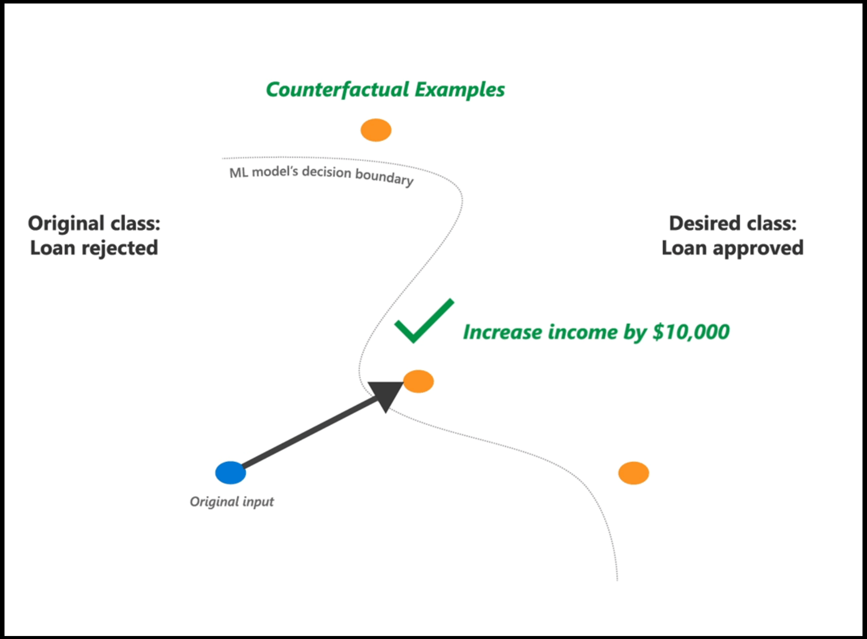


[CEM] Dhurandhar et .al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives », NeurIPS. 2018
[CFX] Albini et. al. "Relation-based counterfactual explanations for Bayesian network classifiers". IJCAI 2020.
[DICE] Mothilal et. al. Explaining machine learning classifiers through diverse counterfactual explanations, FAT 2020
[FACE] Poyiadzi et. al. FACE: Feasible and Actionable Counterfactual Explanations AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society
[MAPLE] Plumb et. al. Model Agnostic Supervised Local Explanations, **NeurIPS 2018**

Counterfactuals

Contrastive explanation method (CEM)

- Local explanation method for neural network
- Given x to explain, CEM considers $x1 = x + \delta$
- Separate positive (δ^p) and negative (δ^n) perturbations w.r.t. label
- Use an autoencoder to explore the boundary between both regions

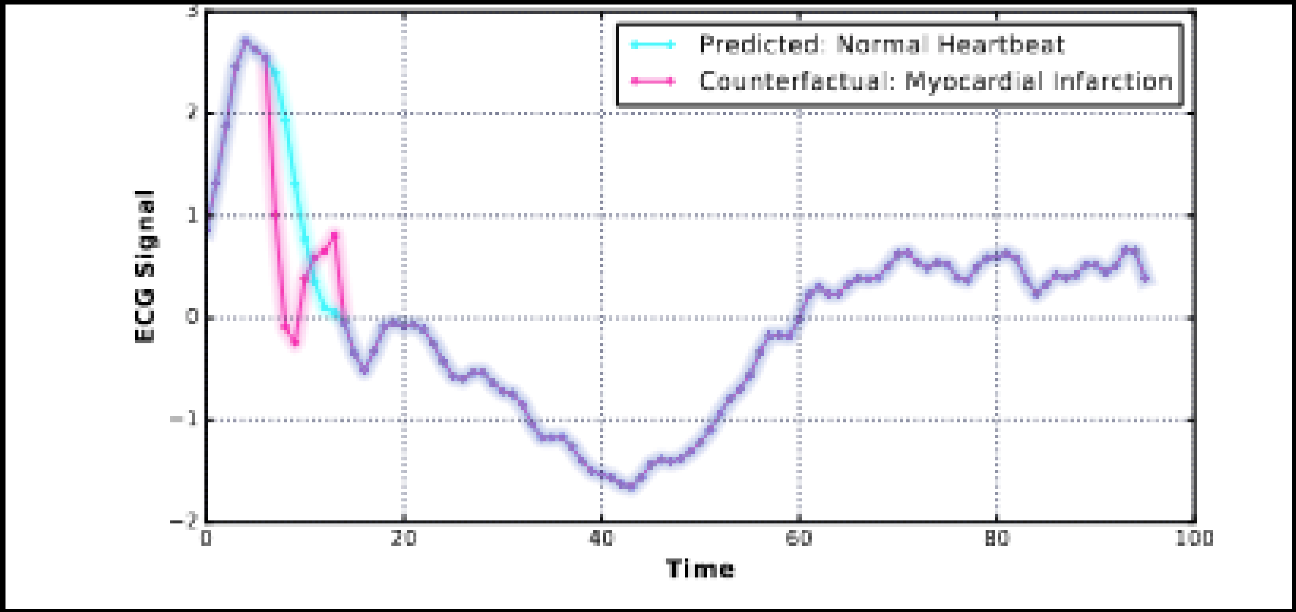


CFX

- Local explanation method for Bayesian network classifiers
- Explanations are built from relations of influence between variables, indicating the reasons for the classification

[CEM] Dhurandhar et .al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives », NeurIPS. 2018
[CFX] Albini et. al. "Relation-based counterfactual explanations for Bayesian network classifiers". IJCAI 2020.

Counterfactuals for TS



« Adapts existing counterfactual instances in the case-base by highlighting and modifying discriminative areas of the time series that underlie the classification »

E Delaney, D. Greene, M. T. Keane. ***Instance-based Counterfactual Explanations for Time Series Classification.*** International Conference on Case-Based Reasoning ICCBR 2021

Factual and counterfactual shapelet-based rules for TS data

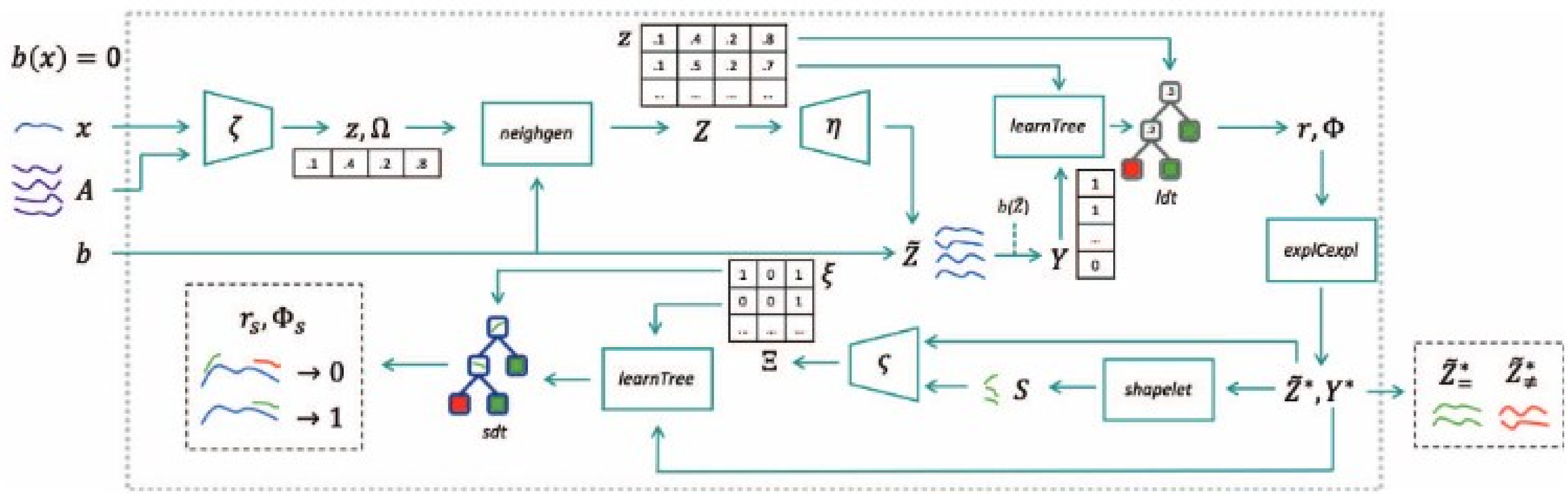
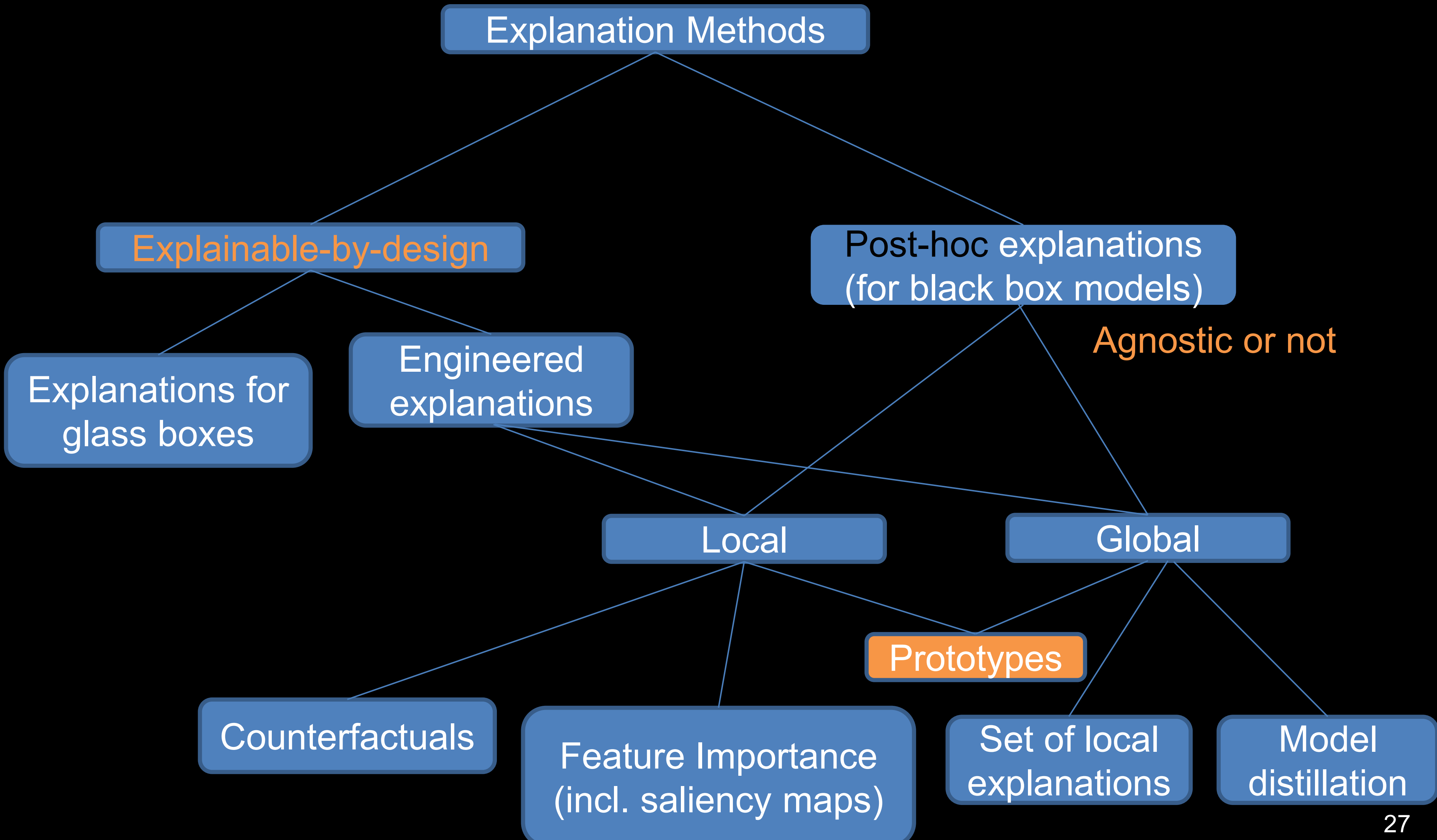


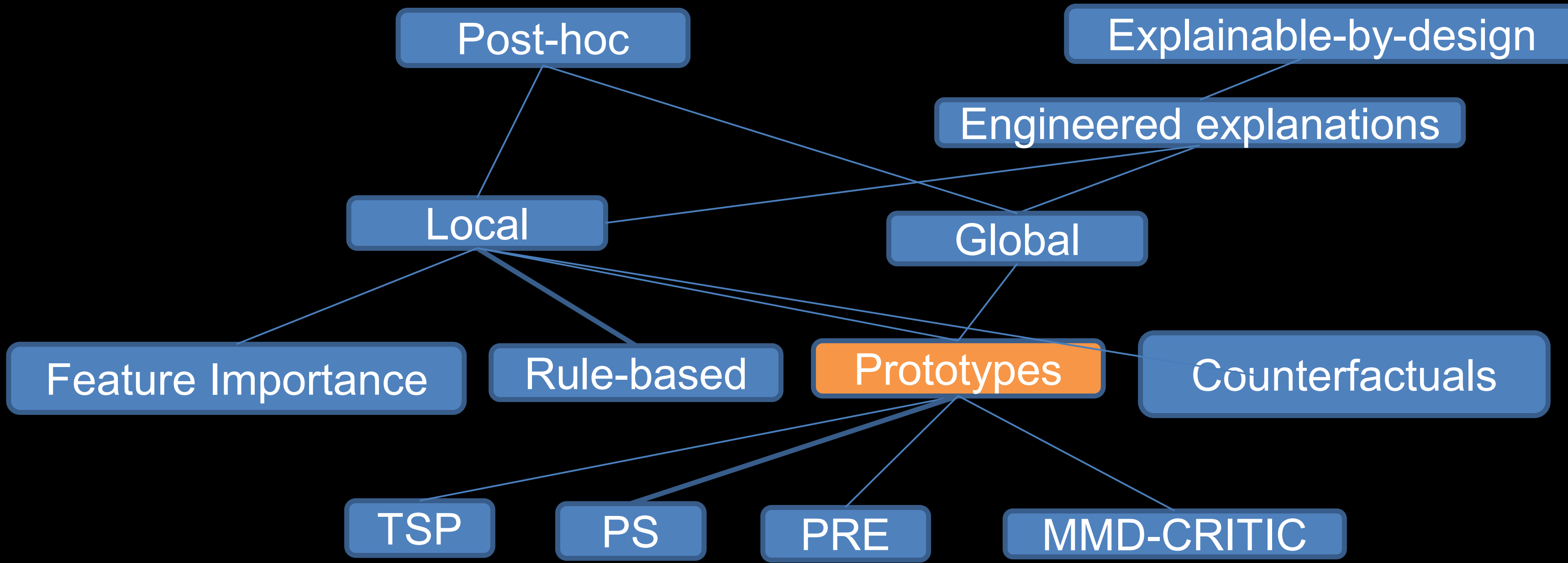
Fig. 1. LASTS architecture. LASTS takes as input the time series x, the black box b and some knowledge on time series A. It uses the AE ζ and η for generating Z and for selecting exemplars and counter-exemplars $\tilde{Z}^*$. From $\tilde{Z}^*$ it extracts the shapelets S and retrieves the shapelet-based rules r_s, Φ_s. The output explanation $e = \langle r_s, \Phi_s, \tilde{Z}^* \rangle$ is contained in the black dashed rectangles.

Riccardo Guidotti; Anna Monreale; Francesco Spinnato; Dino Pedreschi; Fosca Giannotti. **Explaining Any Time Series Classifier**. IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), 2020

Taxonomy of explanation methods



(Post-hoc/By-design) explainability prototypes methods

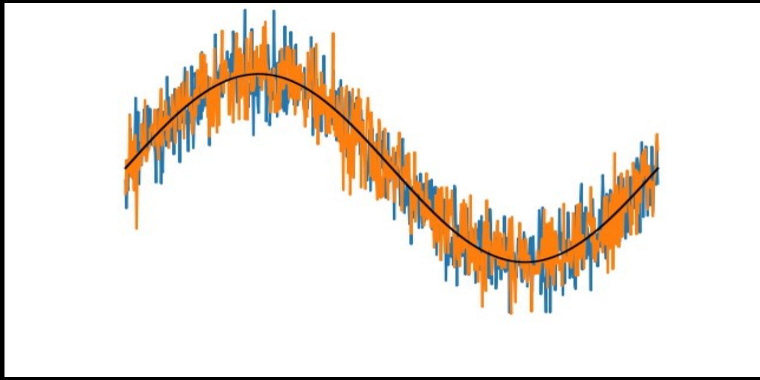


[TSP] Koh et. al. Understanding black-box predictions via influence functions . ICML 2017
[PS] Chen et. al. This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS 2019
[PRE] Li et. al. : "Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions". AAAI 2018
[MMD-CRITIC] Kim et. al. "Examples are not enough, learn to criticize! Criticism for interpretability." NeurIPS 2016

Prototypes

- Explain a model using a synthetic or natural example:
  - from the training set close to the a sample xi
  - a centroid of a cluster for which xi belongs to
  - generated by some ad-hoc process

- Humans observers usually understand a model's reasoning by looking at similar cases



Engineered prototype

- Use white boxes (DT, symbolic rules)
- Learn the explanations WITHIN the model

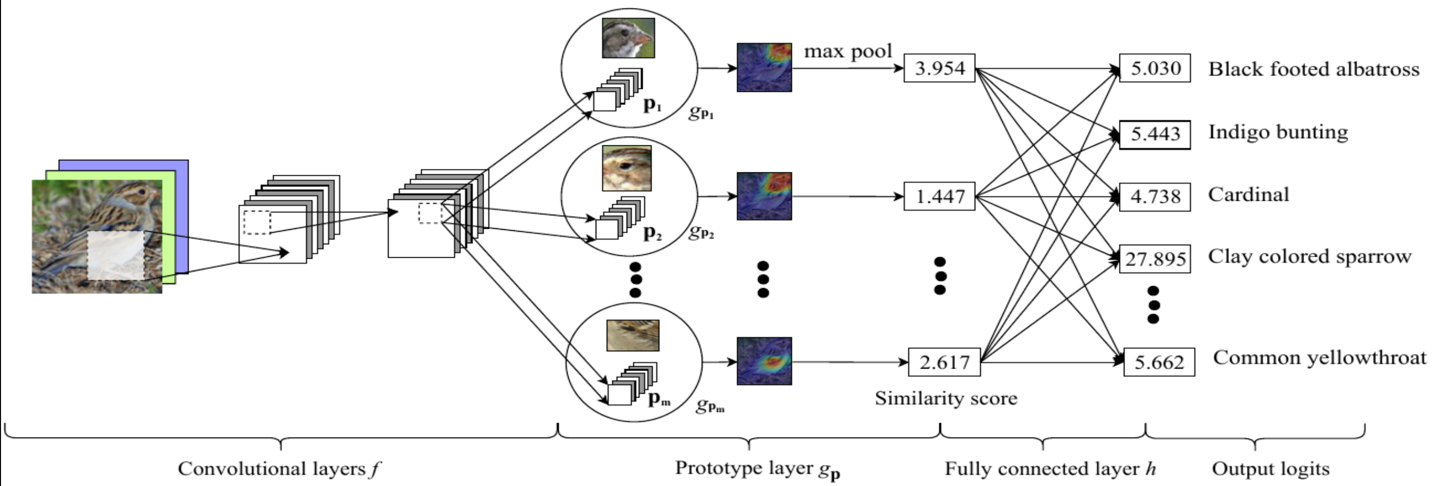
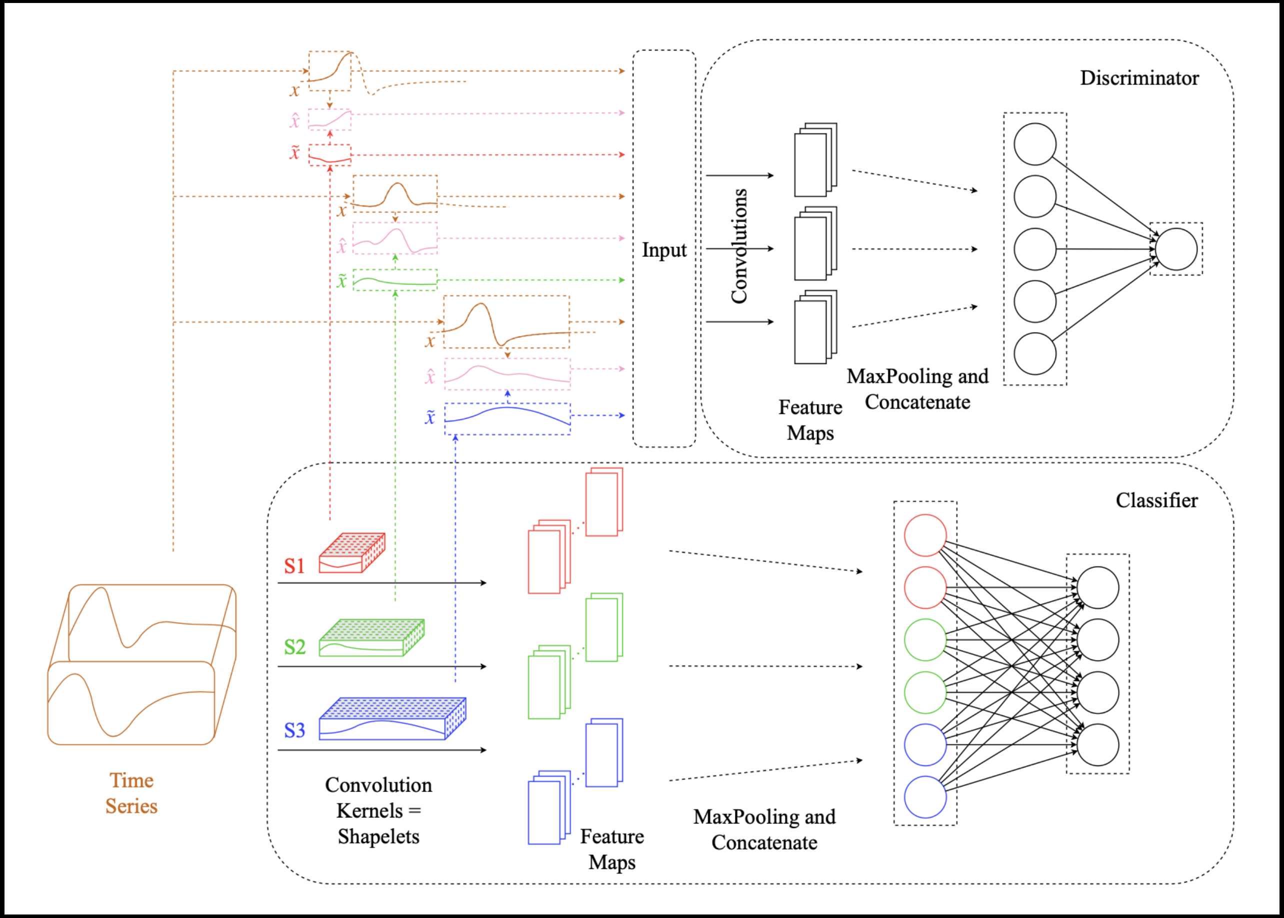


Figure 2: ProtoPNet architecture.

[Prototype Explanations] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, Jonathan Su: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS 2019: 8928-8939

Another example of engineered prototype: XCNN



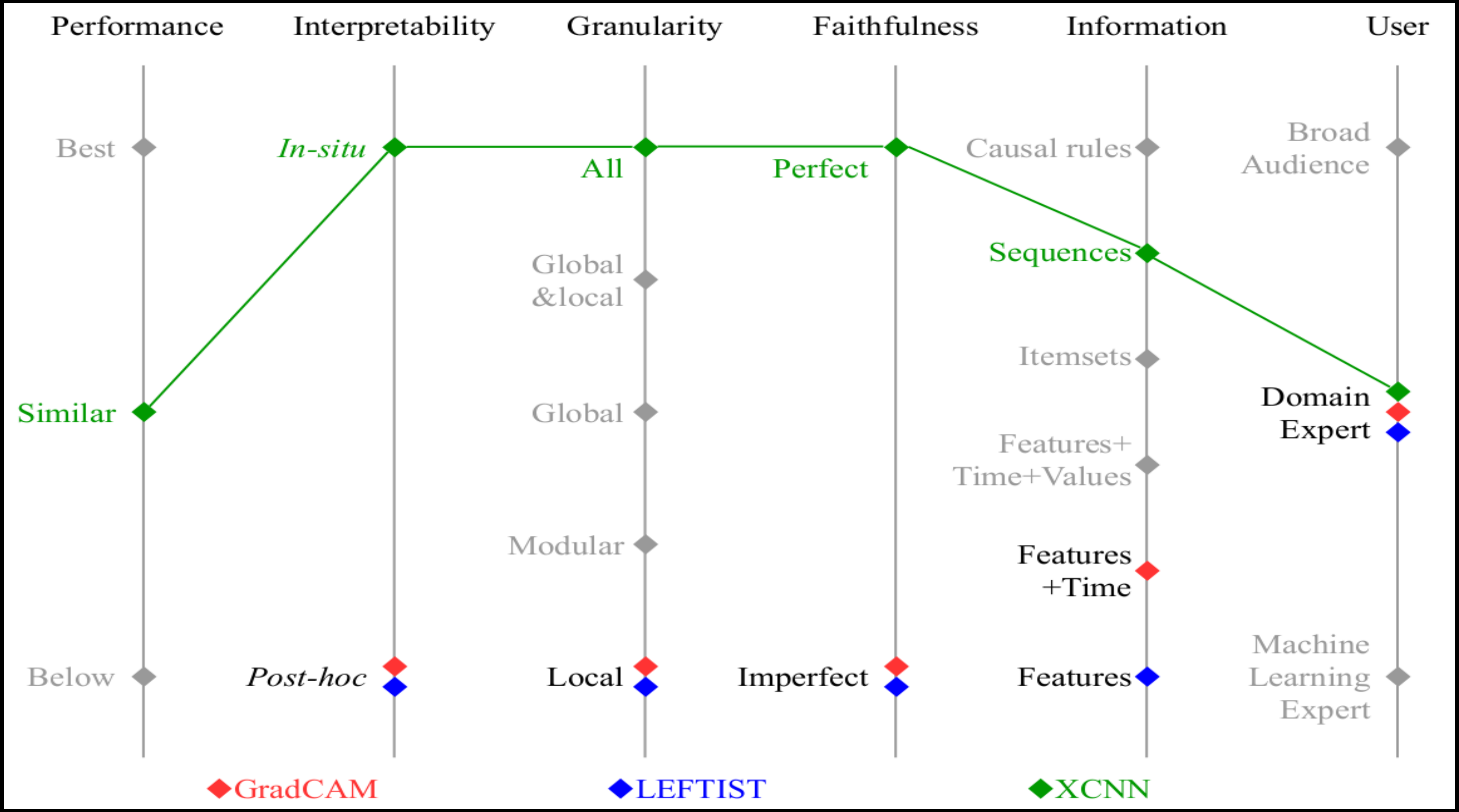
[XCNN] Y. Wang, R. Emonet, E. Fromont, S. Malinowski, R. Tavenard **Adversarial Regularization for Explainable-by-Design Time Series Classification** ICTAI 2020 – 32th International Conference on Tools with Artificial Intelligence, Nov 2020.

XAI Evaluation

An open problem…

- **Fidelity**: how good is f(·) at mimicking b(·)?
- **Stability**: how consistent are the explanations for similar samples?
- **Faithfulness**: how are the relevance scores indicating the *true* important features?
- **Monoticicity**: how is the accuracy of b(·) improved when new a new important feature is added?
- …

A comparison framework?



K. Fauvel, V. Masson, E. Fromont **A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers** IJCAI-PRICAI 2020 – Workshop on Explainable Artificial Intelligence

Conclusion

- Various explainable AI methods have been developed over the last years. They are ALL adapted to TS classification
  - Feature importance is the most widely adopted strategy
  - Rule-based explanations are gaining attention due to the logical formalization strategy
  - Explainable-by-design helps with faithfulness and **stability** of explanations

- There are still space to make explainable AI stable, as well as understandable by different human observers (➔ causal explanations ?)

"Explaining black boxes, rather than replace them with interpretable models, can make the problem worse by providing misleading or false characterizations to the black box. – Rudin (2019)"

Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

Questions ?