Demonstration of InXAI framework on ensemble classifier ML model

Maciej Mozolewski, Jagiellonian University in Kraków, Edrone, Kraków, Poland

Szymon Bobek, Jagiellonian University, Kraków, Poland

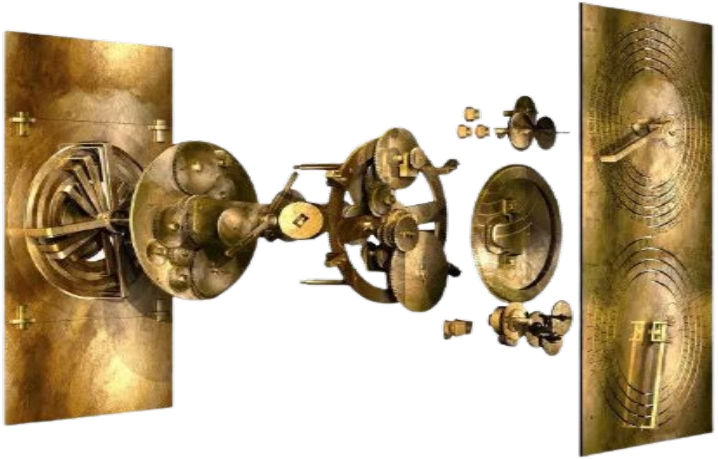
Grzegorz J. Nalepa, Jagiellonian University, Kraków, Poland

https://geist.re

Artificial Intelligence in Research and Applications Seminar (AIRA)
10.03.2022

1. e**X**plainable **A**rtificial **I**ntelligence
2. InXAI
3. First research paper
4. Current work
5. Q&A

eXplainable Artificial Intelligence



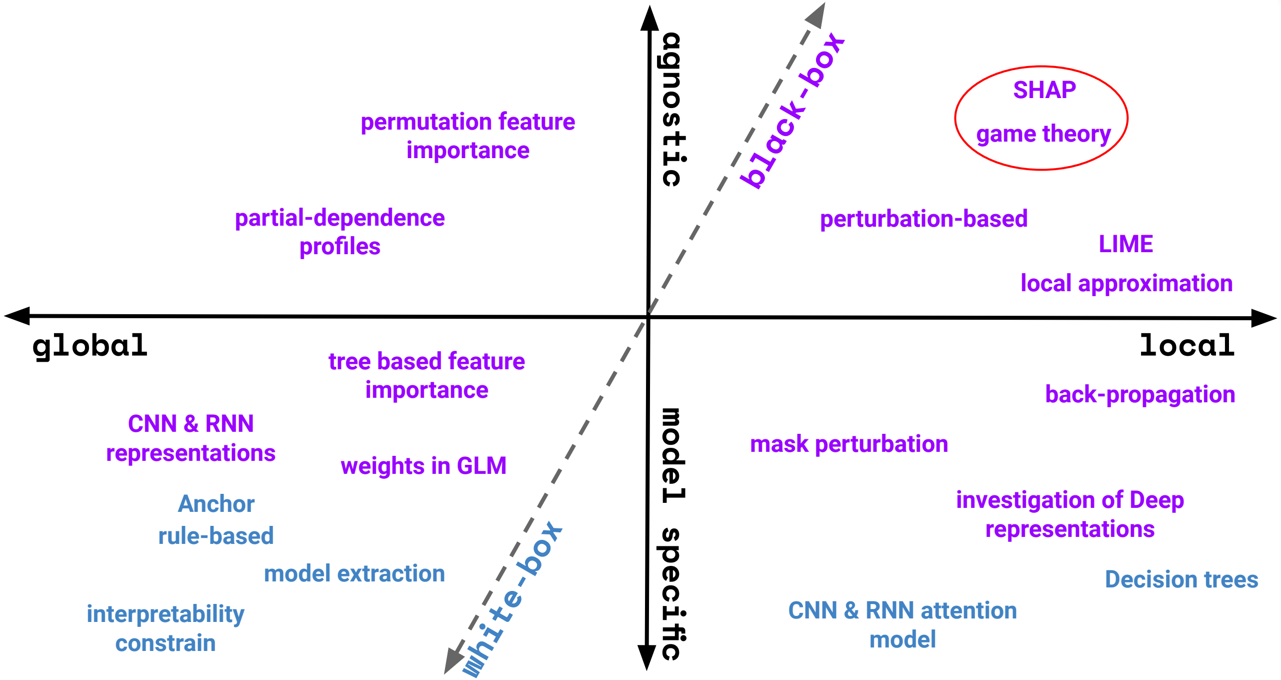
Scientists unlock the 'Cosmos' on the Antikythera Mechanism, the world's first computer, [Livescience]

Mainstream approach to XAI

What variables have contributed to a given output of a model?

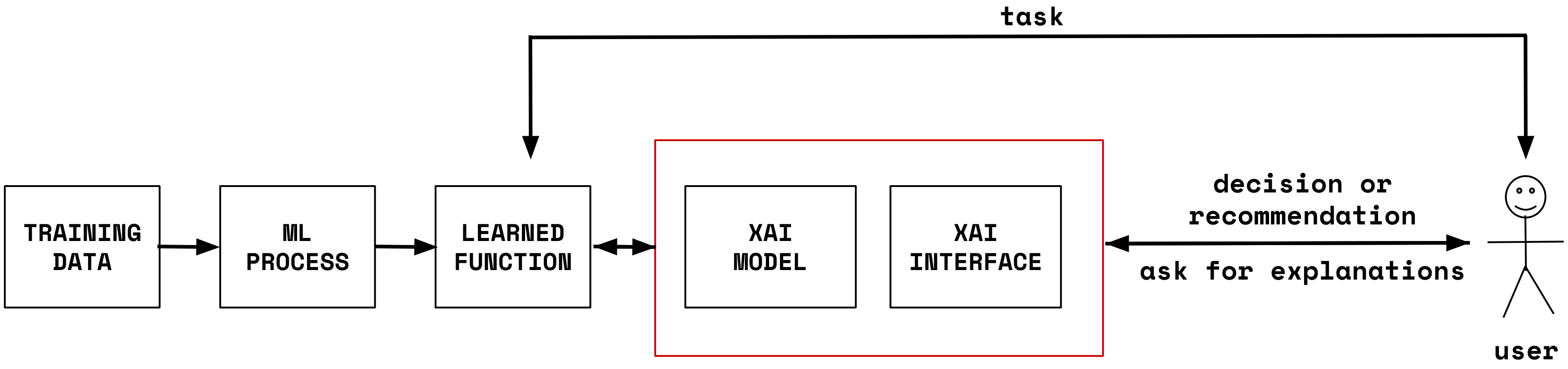
Biecek P., Burzykowski T. (2020). Explanatory Model Analysis. online

The map of XAI



Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable ML. Communications of the ACM, 63(1), 68-77.

Human-in-the-Loop

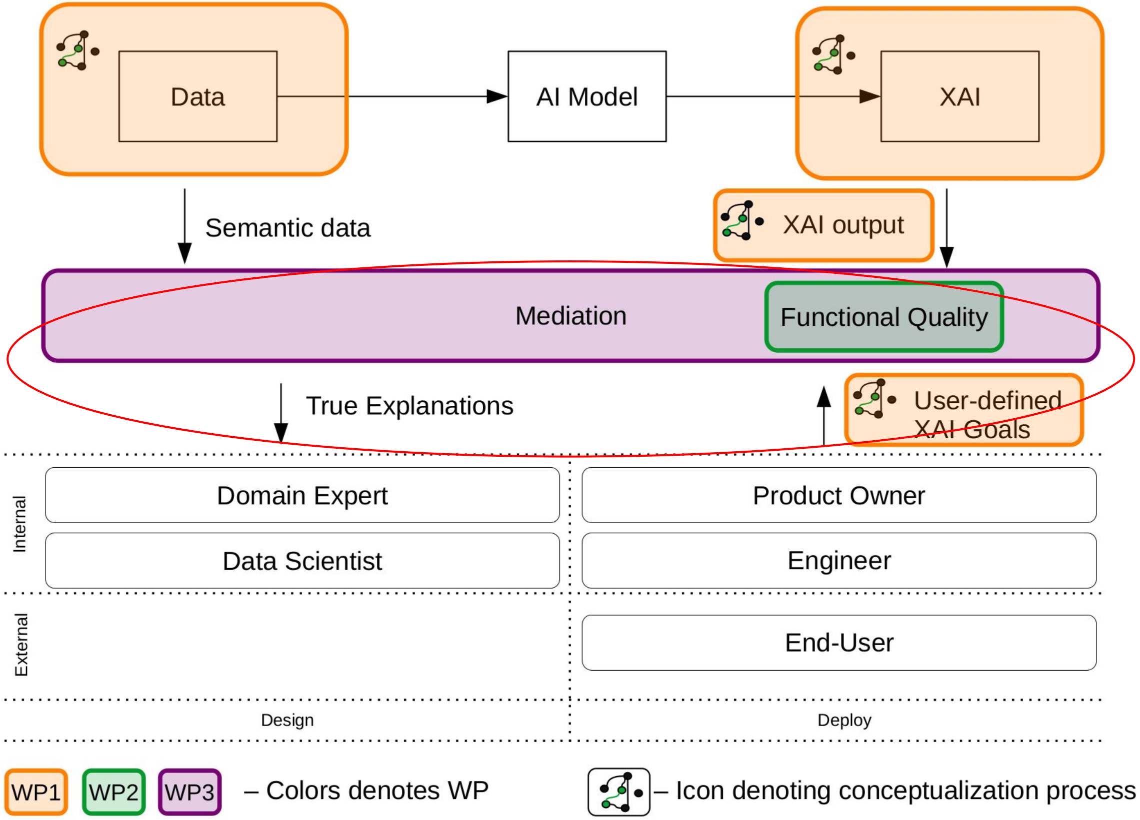


Understand *Why* **_Why Not_** **Know *when to trust AI***

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable ML. Communications of the ACM, 63(1), 68-77.

1. e**X**plainable **A**rtificial **I**ntelligence
2. InXAI
3. First research paper
4. Current work
5. Q&A

InXAI

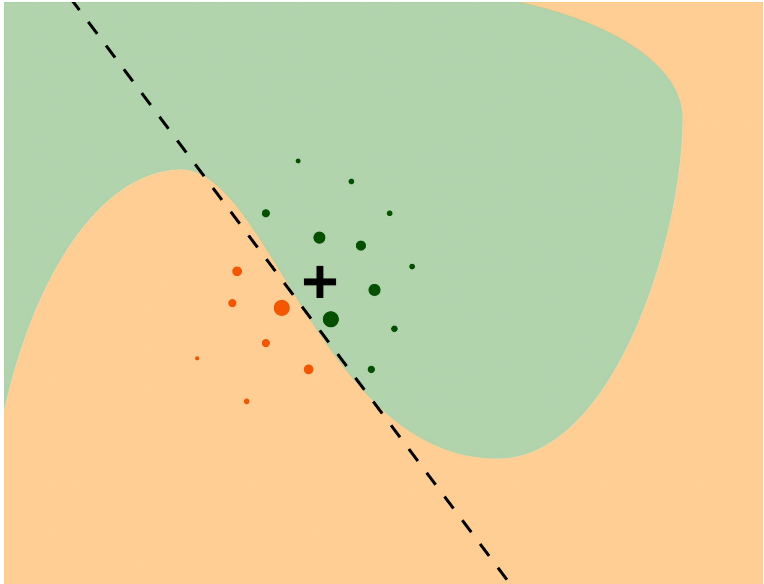


github.com/sbobek/inxai

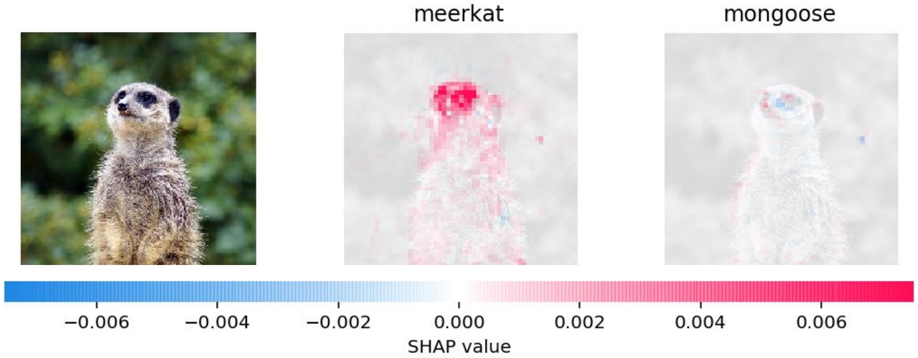
metric	what it measure?	when to use?
Consistency ↑ better	To what extent **different explainers** for predictions of ML model(s) are similar to each other	At the beginning of the pipeline: model/explainer selection Given **model with satisfactory explainer, choose similar one**
Stability robustness ↑ better	For **given explainer**, are explanations similar for similar input, measured with local Lipschitz continuity in the fixed neighborhood of any datapoint	Towards the end of the pipeline: provide end user with model/explainer with **predictable explanations**
AUC Perturbational Accuracy Loss ↓ better	For **given explainer**, how accuracy deteriorates as the data get progressively perturbed, according to their inverse importance in explanation	Compare performance of different explainers Assert **ensemble with good aggregate performance**

Local explainers

LIME



SHAP



$$\varphi(\underline{x}_*, j) = \frac{1}{p!} \sum_J \Delta^{j|\pi(J,j)}(\underline{x}_*)$$

Biecek P., Burzykowski T. (2020). Explanatory Model Analysis. Online.
Lundberg S.M., Lee S. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st NIPS'17. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

1. eXplainable Artificial Intelligence
2. InXAI
3. First research paper
4. Current work
5. Q&A

ICCS 2021:
“Explanation-driven model stacking”

How to have better explanations with InXAI framework?

Bobek S., Mozolewski M., Nalepa G.J. (2021) Explanation-Driven Model Stacking. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J.J., Sloot P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12747. Springer, Cham.

Ensemble model

Weighted sum of several **classifiers**

$$\mathbb{P}_{mm}(Q|x^{(i)}) = \frac{\sum_k \mathbb{P}_k(Q|x^{(i)})w_k}{\sum_k w_k}$$

$$\sum_k w_k > 0; \; w_k \geq 0$$

Optimise w_k for the selected InXAI metric, while keeping "standard" metrics for ML models at a decent level

Metrics for ensemble model

Ensemble Inner Consistency

$$C_{mm} = C\left(\frac{w_1}{\sum_k w_k}\Phi^{m_1}, \frac{w_2}{\sum_k w_k}\Phi^{m_2}, \dots, \frac{w_1}{\sum_k w_k}\Phi^{m_k}\right)$$

Φ - explainer

Consistency

$$C(\Phi^{m_1}, \Phi^{m_2}, \dots, \Phi^{m_k}) = \frac{1}{\max\limits_{a,b \in 1,2,\dots,k} ||\Phi^{m_a} - \Phi^{m_b}||_2 + 1}$$

Optimization of weights of ensemble

$$L_{mm} = \frac{AUCx_{mm}^{\gamma_{auc}}}{\overline{S_{mm}}^{\gamma_s} \cdot \overline{C_{mm}}^{\gamma_c}}$$

SMAC

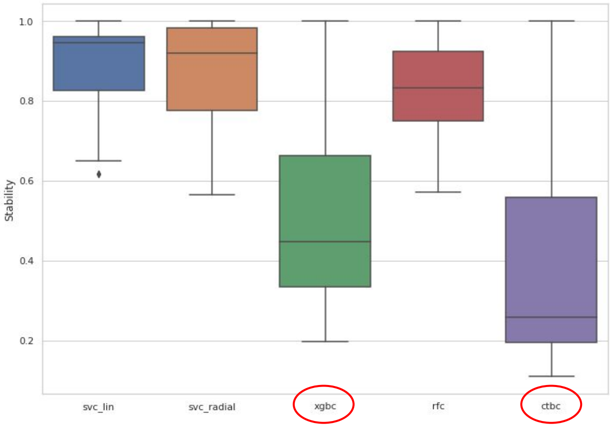
$$\left.\begin{aligned} \overline{S_{mm}} &= \frac{\sum_i^N {S_{mm}}^i}{N} \\ \overline{C_{mm}} &= \frac{\sum_i^N {C_{mm}}^i}{N} \end{aligned}\right\}$$ mean value across all observations

$$\left.\begin{aligned} AUCx_{approx} &= \frac{\sum_k AUCx_k\, w_k}{\sum_k w_k} \\ S_{approx} &= \frac{\sum_k S_k\, w_k}{\sum_k w_k} \end{aligned}\right\}$$ runtime approx.

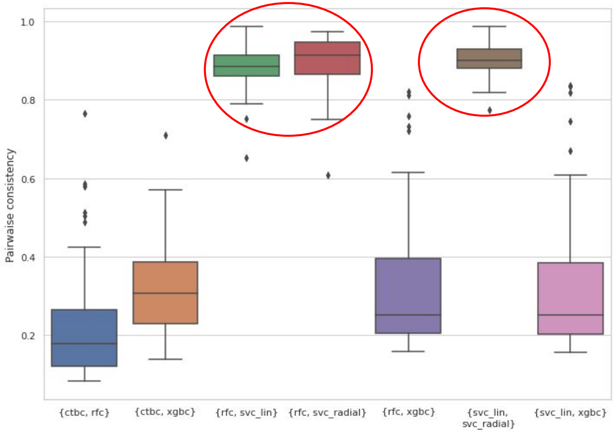
“Toy” example: binary classifier

model	model abbreviation	accuracy score	F1-score class ”0”	F1-score class ”1”
SVMClassifier with RBF kernel	svc_radial	0.76	0.78	0.73
SVMClassifier with linear kernel	svc_lin	0.82	0.83	0.80
XGBClassifier	xgbc	0.74	0.76	0.72
RandomForestClassifier	rfc	0.74	0.77	0.70
CatBoostClassifier	ctbc	0.65	0.66	0.65

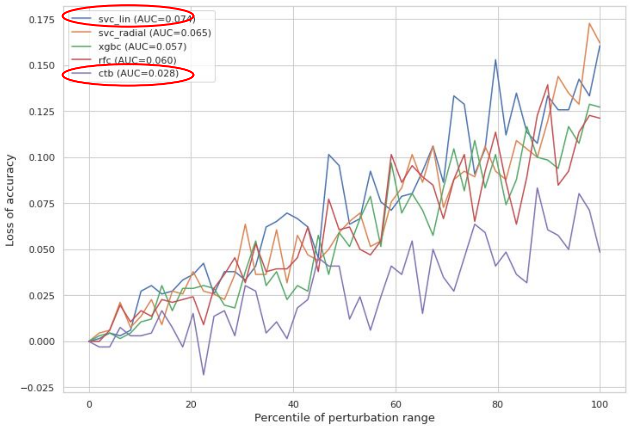
InXAI metrics for SHAP explainer



Stability per unit model



Pairwise consistency



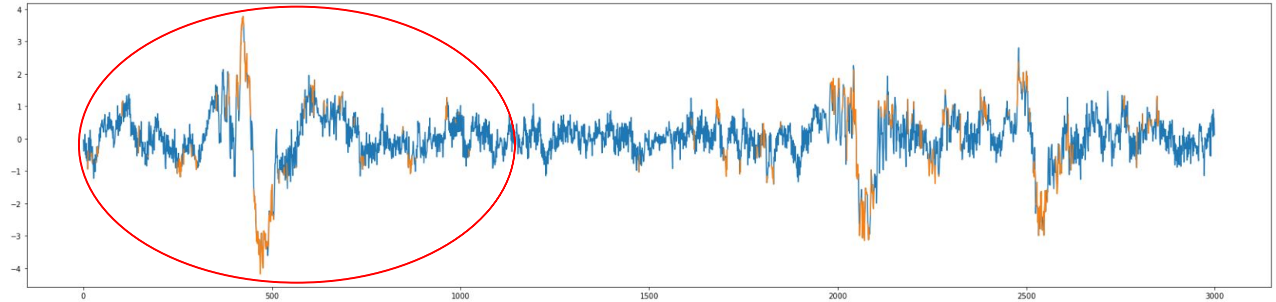
AUC Perturb. Acc. Loss

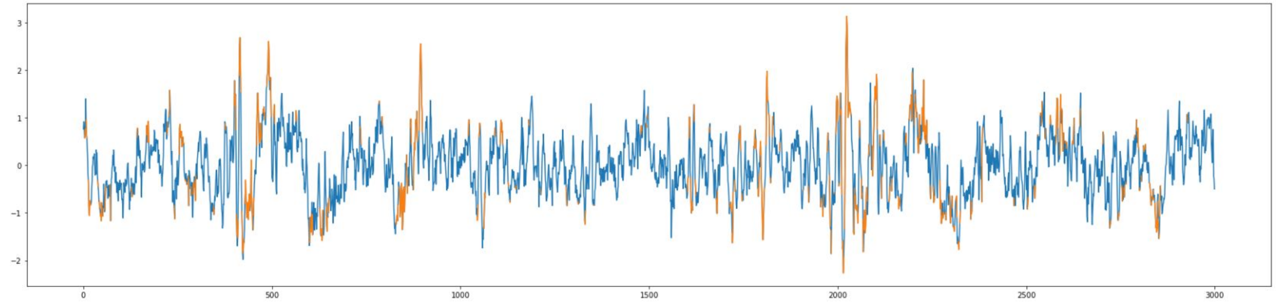
Results and conclusions

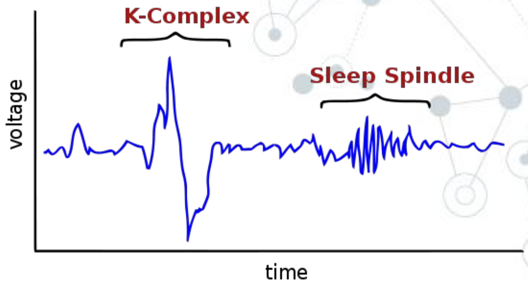
#	meta-parameter			weights for models after optimization					metrics			
	AUC acc. loss	Stabi- -lity	Consis- -tency	xgbc	rfc	ctbc	svc_lin	svc_radial	model acc.	AUC acc. loss	Stabi- -lity	Consis- -tency
1	1.0	1.0	1.0	.000087	.363524	.000031	.272740	.363619	0.76	0.060	0.872	0.895
2	**3.0**	1.0	1.0	.000042	**.499893**	.000025	**.000004**	**.500035**	0.73	**0.048**	0.858	0.862
3a	1.0	**3.0**	1.0	.000007	.315697	.000021	.312844	.371430	0.77	0.062	**0.874**	0.899
3b	1.0	**5.0**	1.0	.000021	**.000013**	.000020	**.499952**	**.499993**	0.77	0.059	**0.887**	0.871
4a	1.0	1.0	**3.0**	.000062	.318573	.000074	.310580	.370711	0.77	0.064	0.874	**0.899**
4b	1.0	1.0	**5.0**	.000026	.293124	.000037	**.350562**	**.356252**	0.77	0.067	0.876	**0.902**

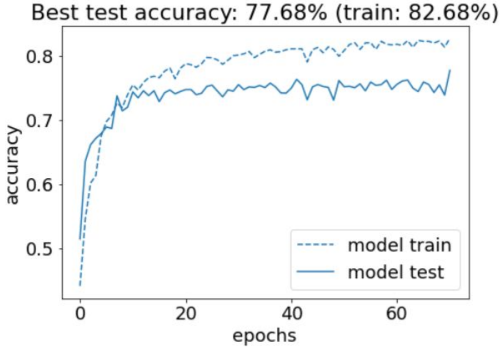
1. e**X**plainable **A**rtificial **I**ntelligence
2. InXAI
3. First research paper
4. Current work
5. Q&A

InXAI: Time Series









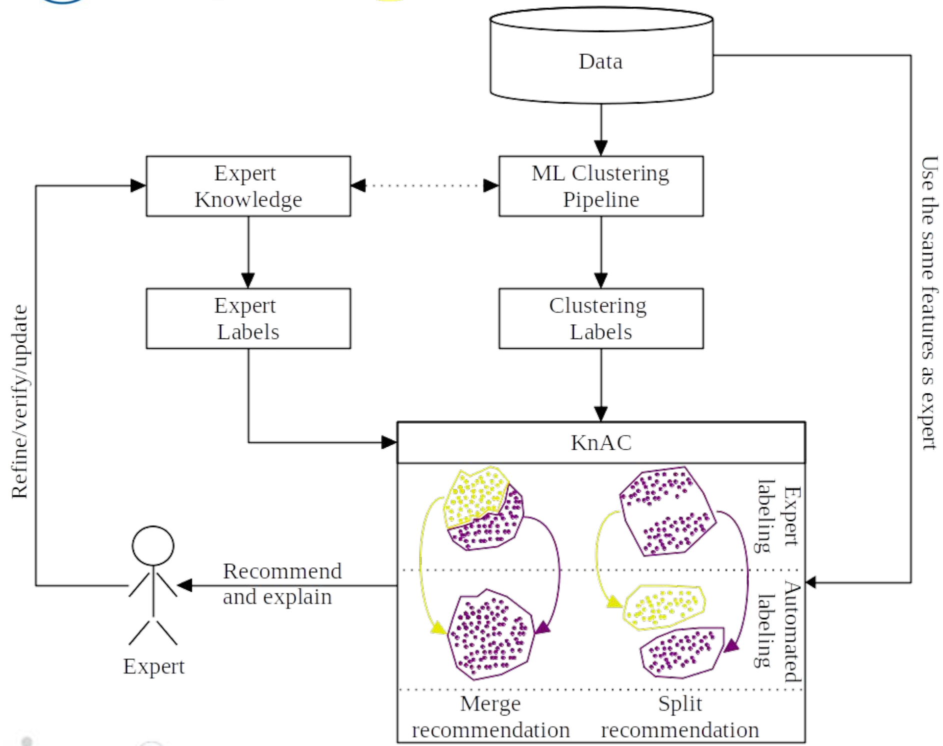
Supratak A., Guo Y. (2020), **TinySleepNet**: An Efficient Deep Learning Model for Sleep Stage Scoring based on Raw Single-Channel EEG.

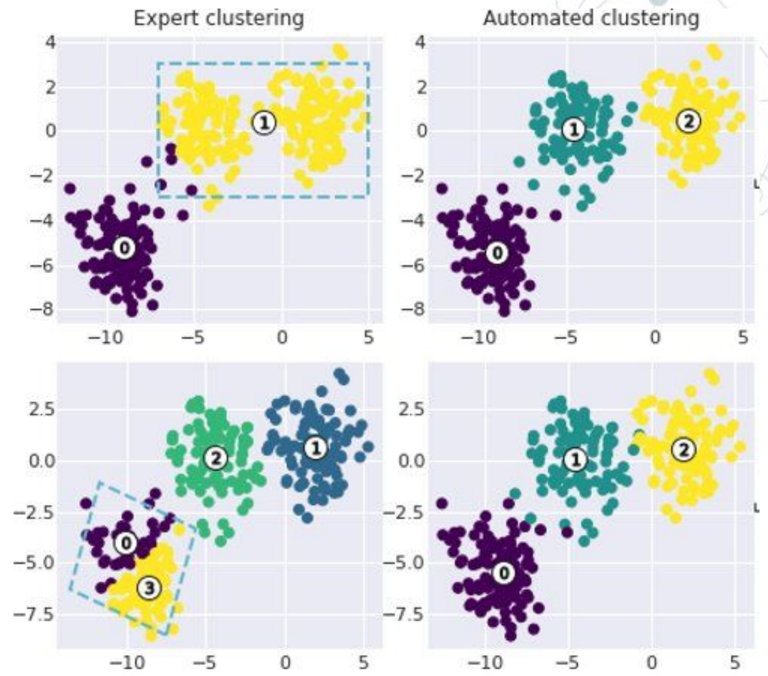
Berry R. B. (2011), Fundamentals of Sleep Medicine, 1st Edition, Elsevier.

github.com/flower-kyo/Tinysleepnet-pytorch

en.wikipedia.org/wiki/K-complex

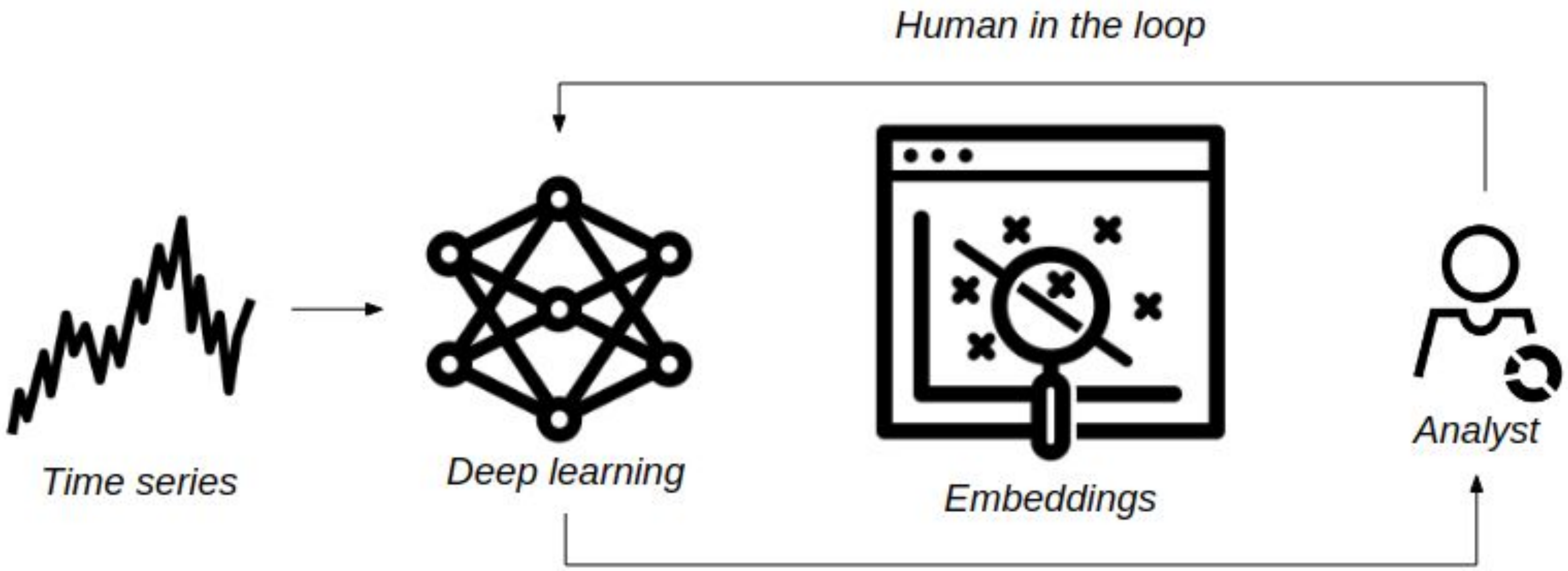
KnAC





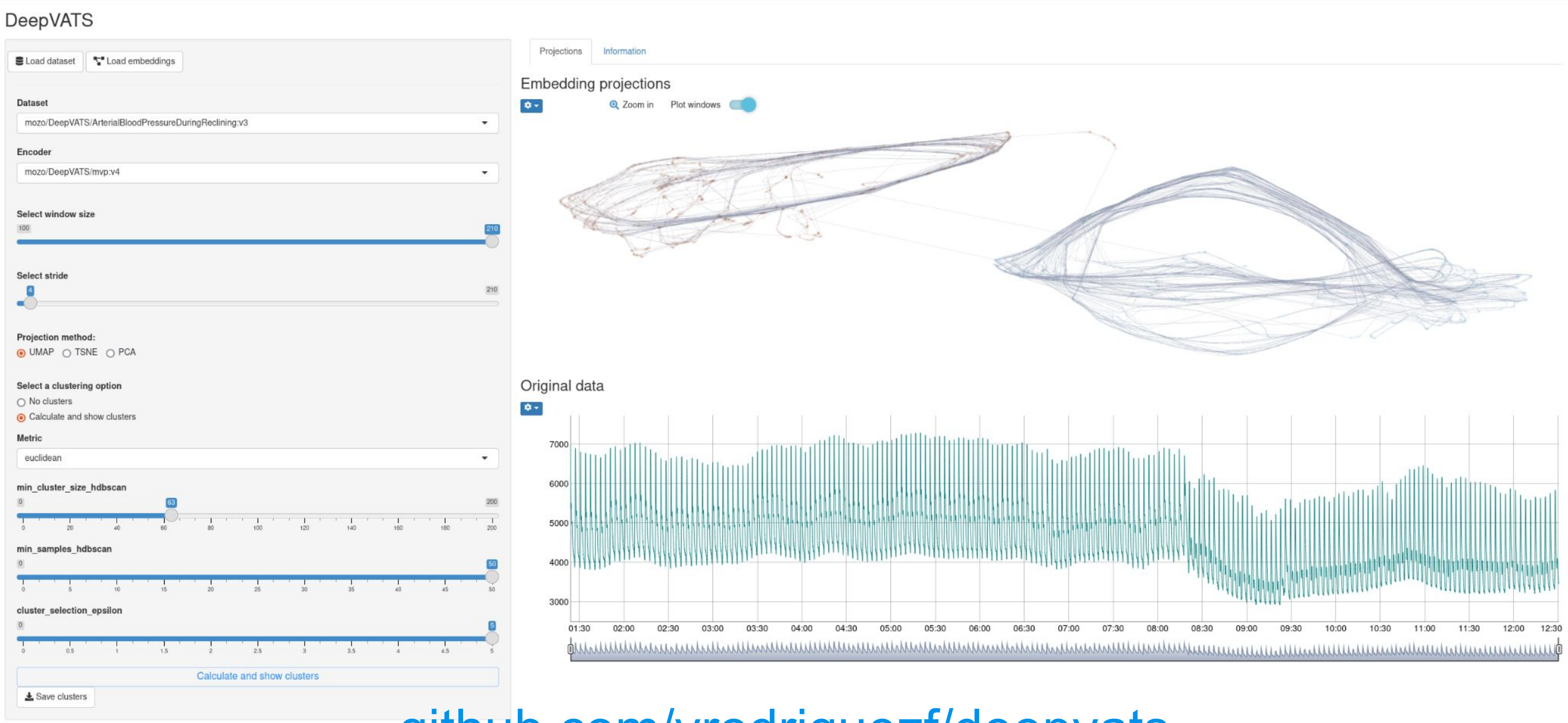
github.com/sbobek/knac

DeepVATS for Time Series



github.com/vrodriguezf/deepvats

DeepVATS for Time Series



github.com/vrodriguezf/deepvats

DeepVATS for Time Series

1. e**X**plainable **A**rtificial **I**ntelligence
2. InXAI
3. First research paper
4. Current work
5. Q&A

Summary

1. **InXAI** / Beyond feature importance
2. **KnAC** / Human-in-the-Loop
3. **DeepVATS** / Time Series clustering

Bibliography

1. Bobek S., Mozolewski M., Nalepa G.J. (2021) Explanation-Driven Model Stacking. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J.J., Sloot P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12747. Springer, Cham. Download.
2. Bobek S., Bałaga P., Grzegorz N. (2021). Towards Model-Agnostic Ensemble Explanations. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J.J., Sloot P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12747. Springer, Cham. Download.
3. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.
4. Shapley, Lloyd S. 1953. “A Value for n-Person Games.” Contributions to the Theory of Games II,(Harold W. Kuhn and Albert W. Tucker), 307–17. Princeton: Princeton University Press.
5. Štrumbelj, Erik, and Igor Kononenko. 2010. “An Efficient Explanation of Individual Classifications Using Game Theory.” Journal of Machine Learning Research 11 (March): 1–18.
6. Arrieta B. A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, 2020, Pages 82-115, ISSN 1566-2535.
7. Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Communications of the ACM, 63(1), 68–77.

Questions & Answers

Thank you!