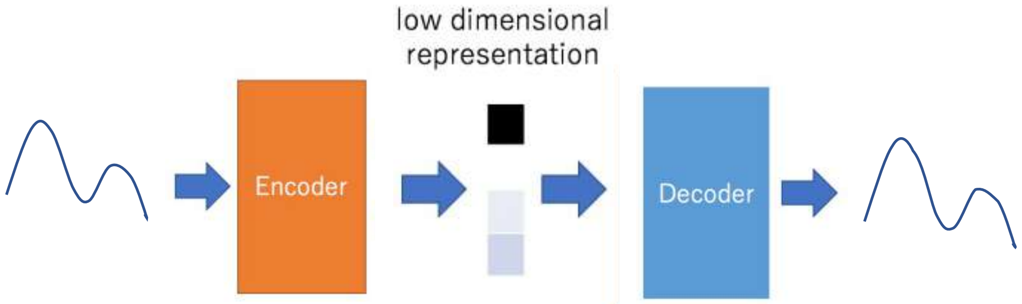
Exploiting Auto-Encoders for Explaining Black Box Classifiers

Riccardo Guidotti, Anna Monreale



What is a Black Box Model?

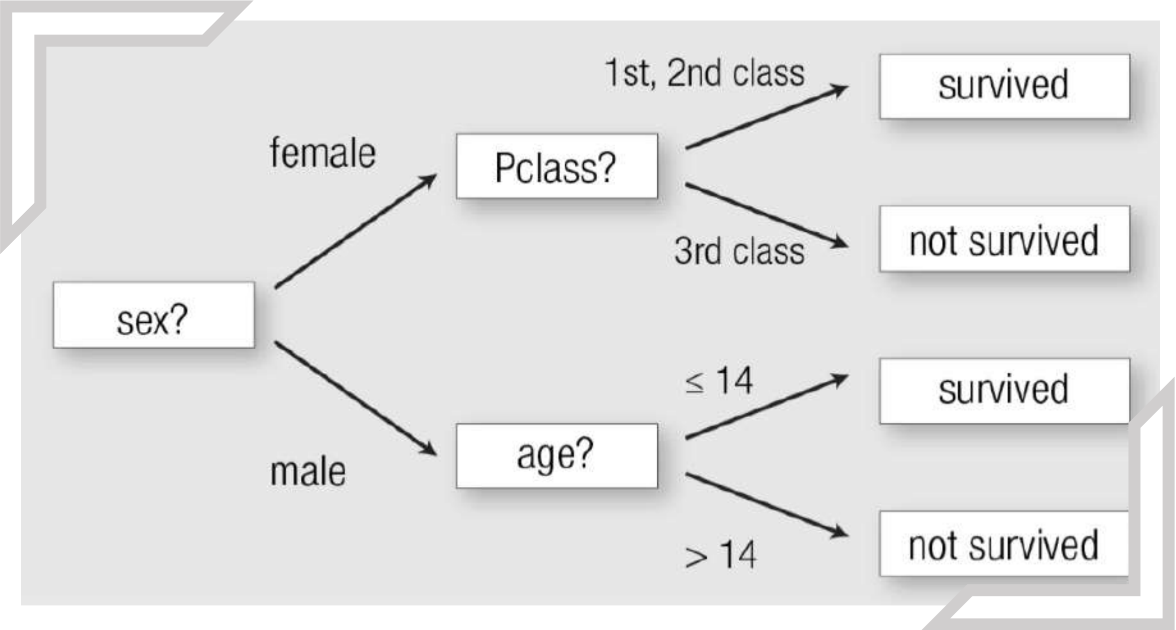
A ***black box*** is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

Example:

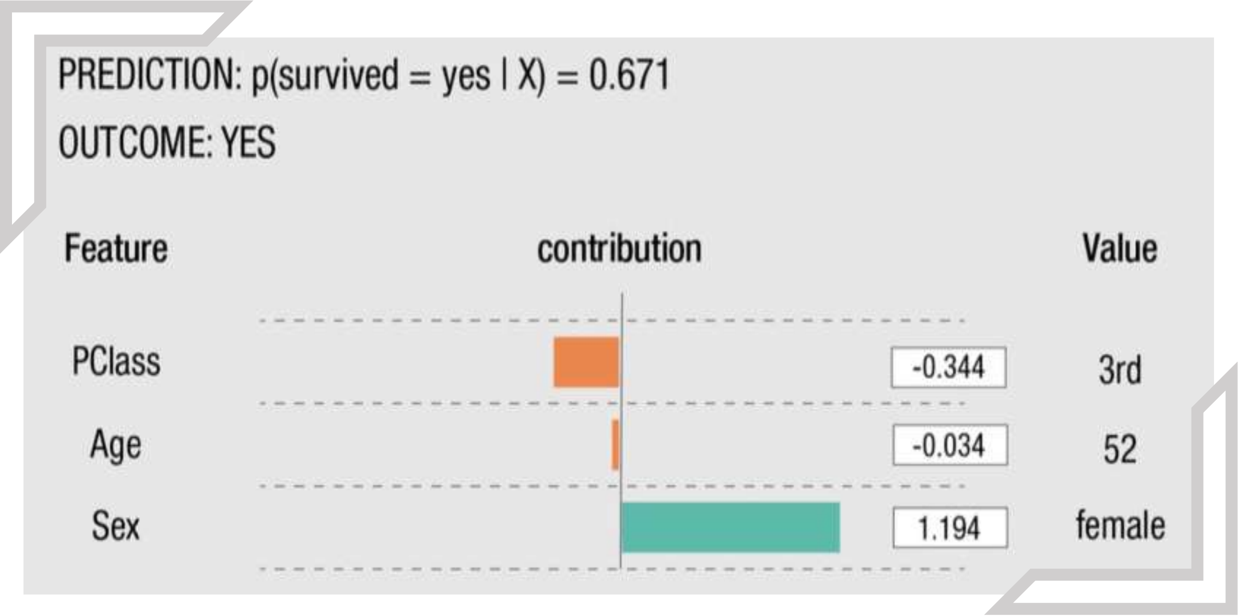
- DNN
- SVM
- Ensemble

- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). ***A survey of methods for explaining black box models***. *ACM Computing Surveys (CSUR)*, *51*(5), 93.

Interpretable Models



Decision Tree



Linear Model

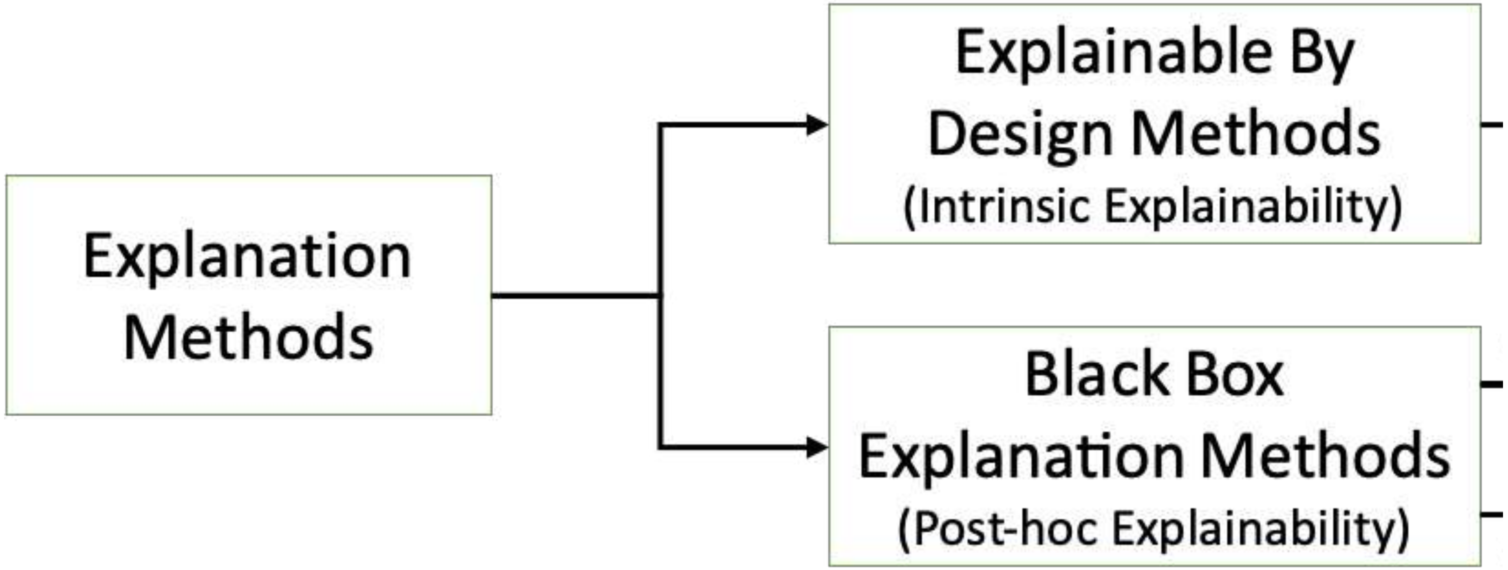
$$\text{if } condition_1 \wedge condition_2 \wedge condition_3 \text{ then } outcome$$

Rules

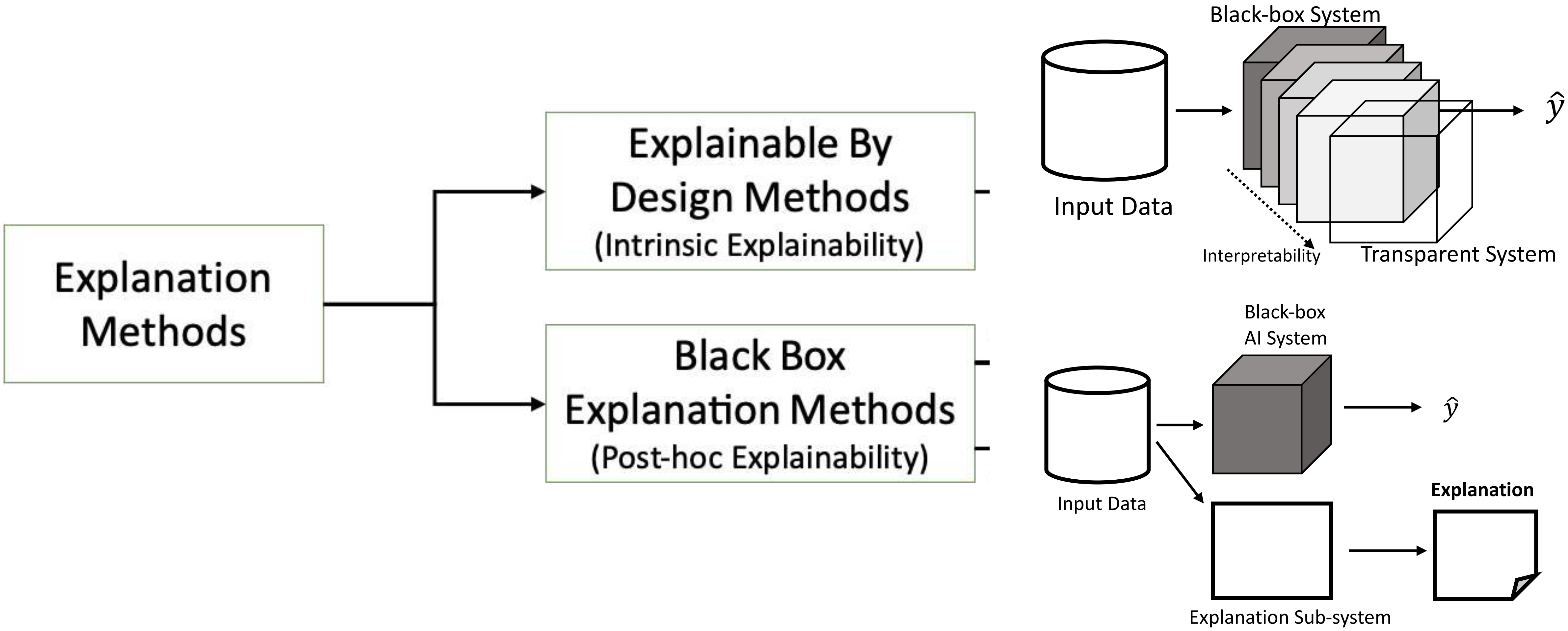
How to Open the Black Box

XAI Taxonomy of Explanation Methods

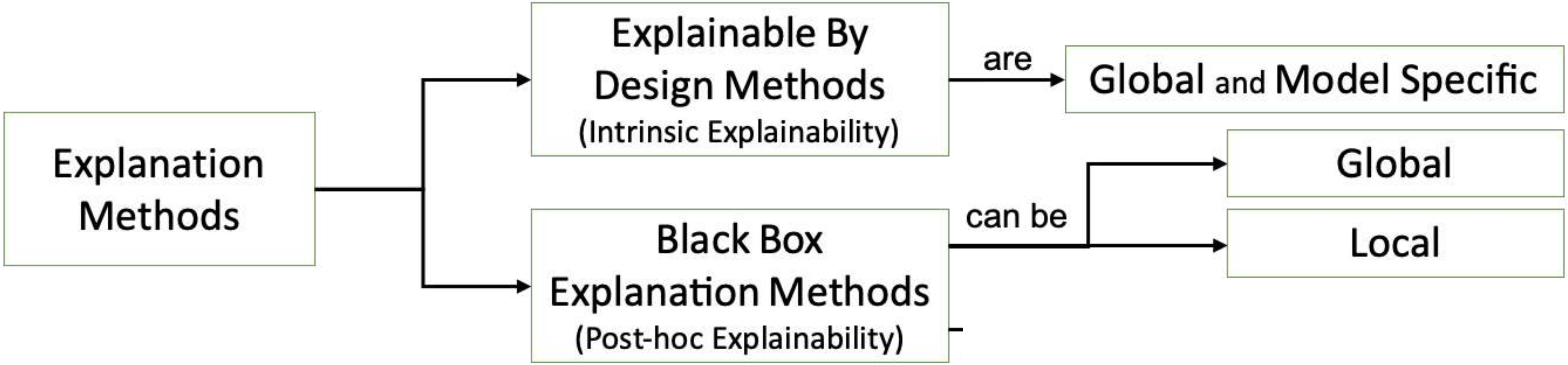
XAI Taxonomy of Explanation Methods



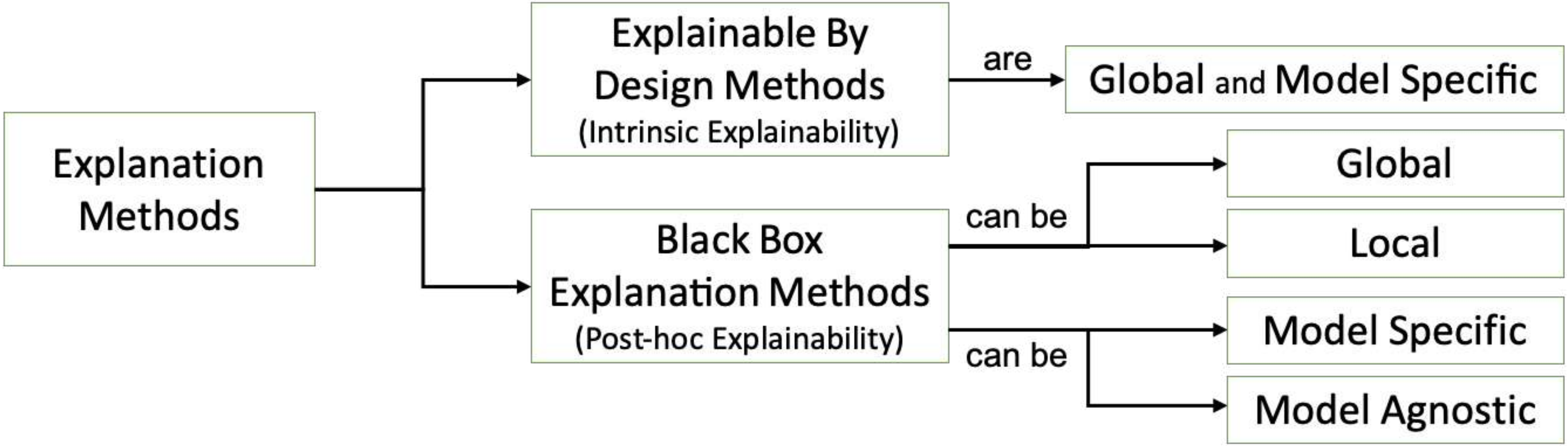
XAI Taxonomy of Explanation Methods



XAI Taxonomy of Explanation Methods



XAI Taxonomy of Explanation Methods



Types of Data

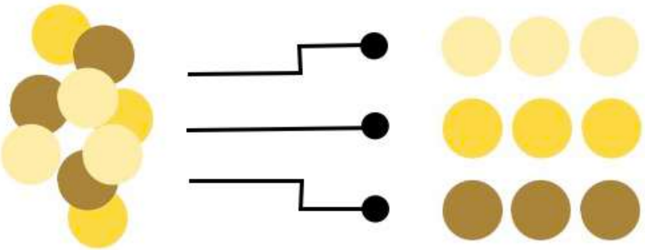


Table of baby-name data
(baby-2010.csv)

name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

Field names

One row (4 fields)

2000 rows all told

Tabular

(**TAB**)

Images

(**IMG**)

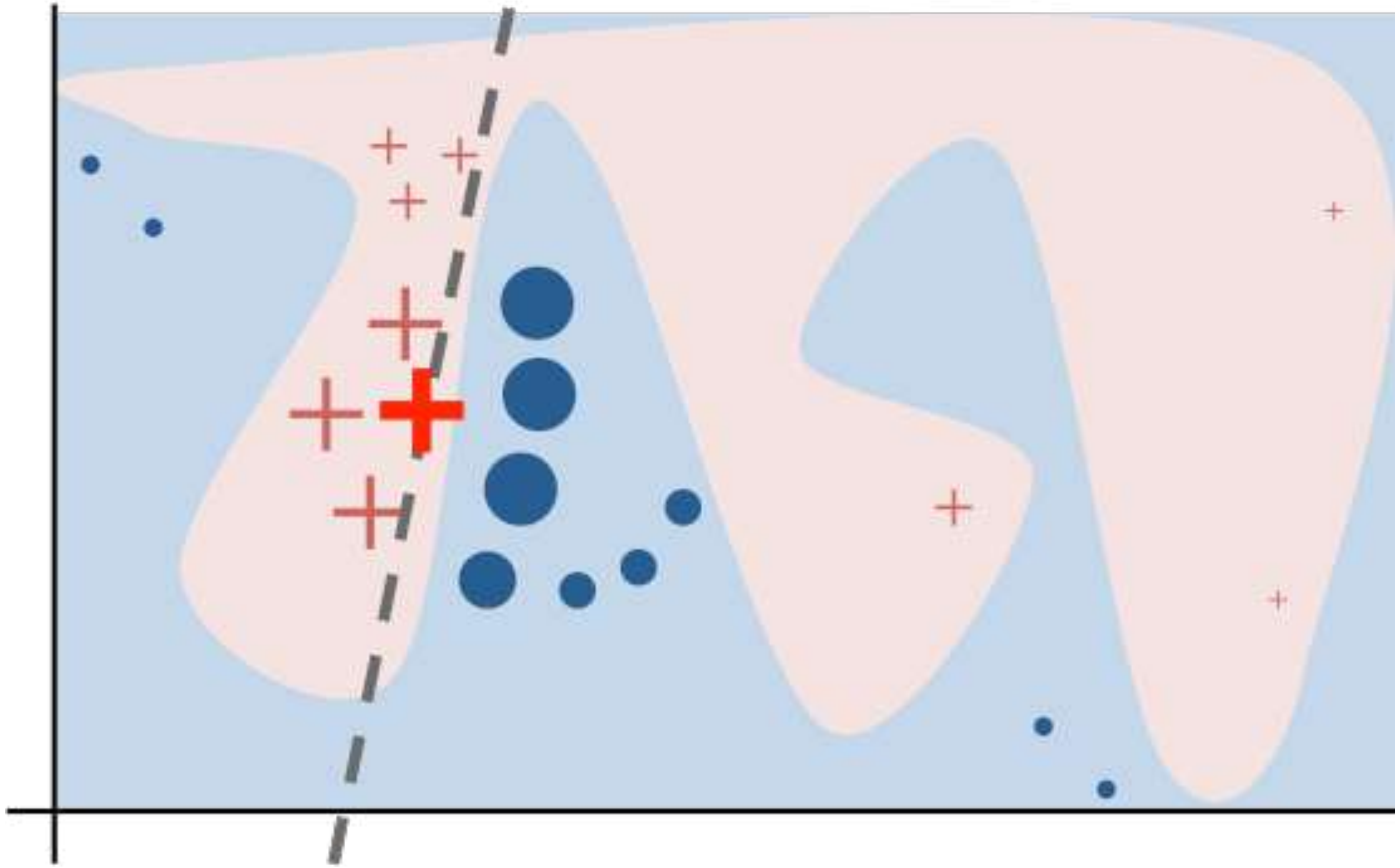
Text

(**TXT**)

Background and Motivations

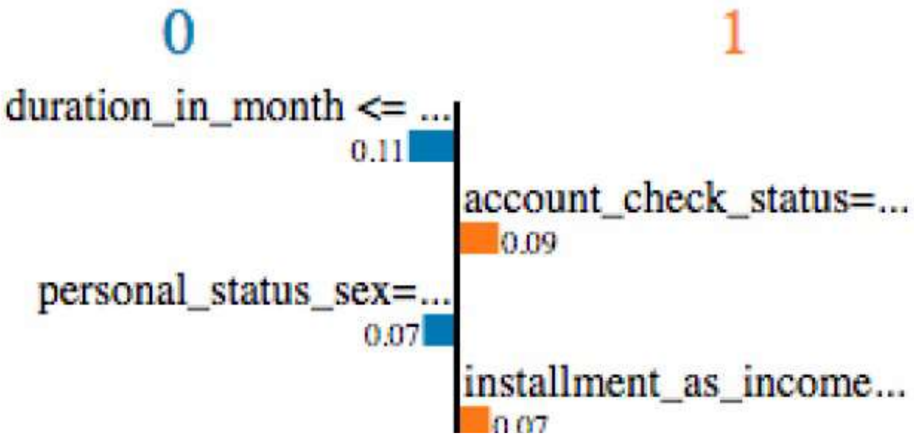
Local Explanation

- The overall decision boundary is complex
- In the neighborhood of a single decision, the boundary is simple
- A single decision can be explained by auditing the black box around the given instance and learning a ***local*** decision.



Local Interpretable Model-agnostic Explanations

- Local model-agnostic explainer that reveals the black box decisions through features importance/saliency maps.
- It locally approximates the behavior of a black box with a local surrogate expressed as a logistic regressor (with Lasso or Ridge penalization).
- Synthetic neighbors are weighted w.r.t. the distance with the instance to explain.



LIME on Tabular Data

Sepal length	Sepal width	Petal length	Petal width	b(setosa)	b(versic)	b(virgi)
3	4	3	6	0.1	0.7	0.2

Train a Logistic Regressor

Returns the coefficients as Explanation

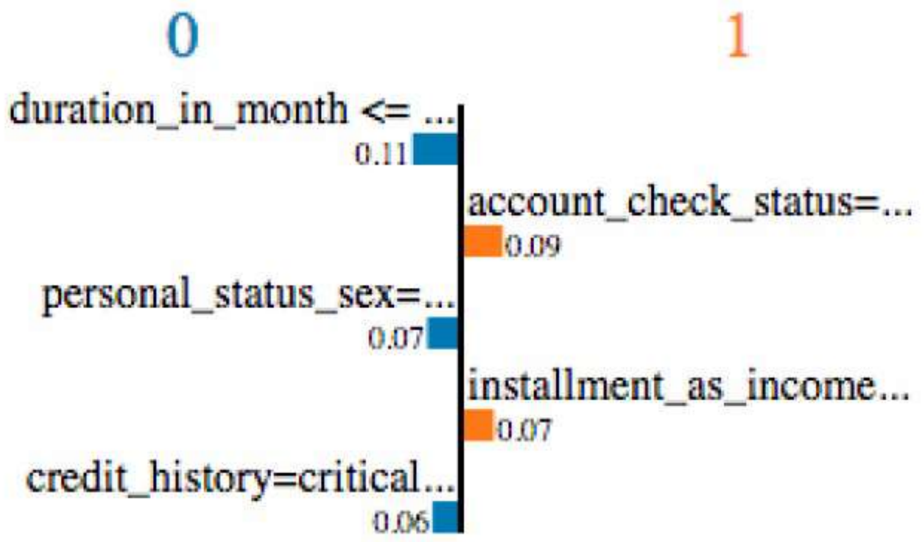
LIME Pseudo - Code

```
Z = {}
x instance to explain
x' = real2interpretable(x)
for i in {1, 2, …, N}
    z_i= sample_around(x')
    z = interpretabel2real(z')
    Z = Z U {<z_i, b(z_i), d(x, z)>}
w = solve_Lasso(Z, k)
return w
```

black box auditing

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. KDD.

Features Importance

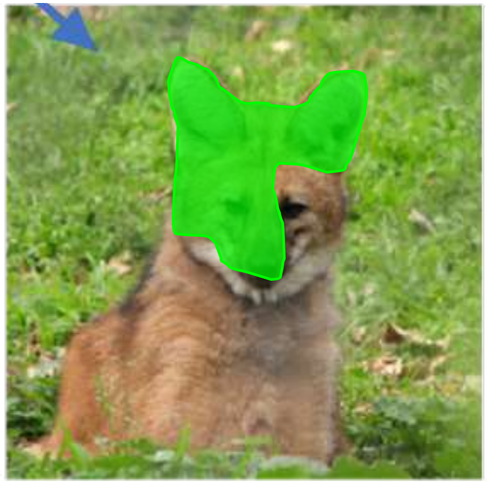


Saliency Map

LIME on Images

- LIME ***turns*** an image x to a vector x’ of interpretable superpixels expressing presence/absence.
- It ***generates*** a synthetic neighborhood Z by randomly perturbing x’ and labels them with the black box.
- It ***trains*** a linear regression model (interpretable and locally faithful) and assigns a weight to each superpixel.

x

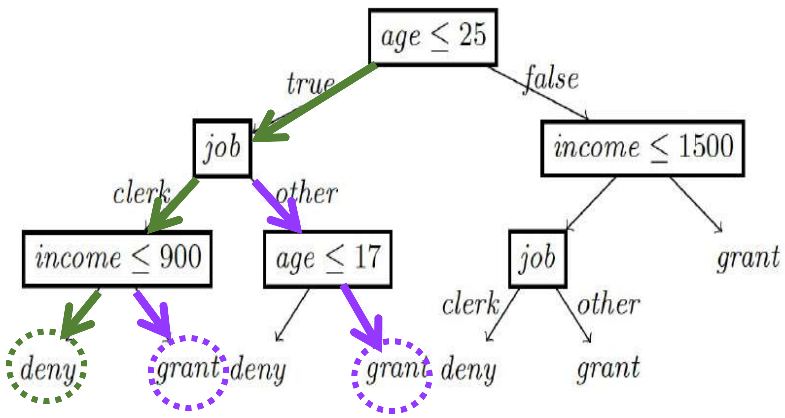


LIME Issues

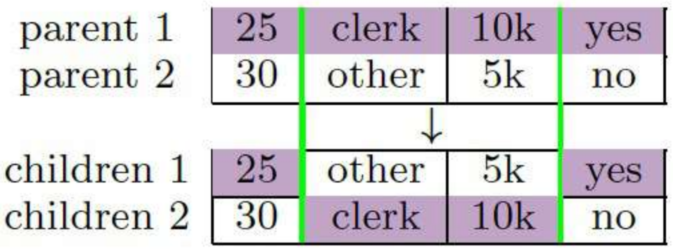
- A set of values with weights (both for images and tabular data) is not necessarily a good and human comprehensible explanation
- LIME does not really generate images with different information: it randomly removes some superpixels, i.e. it suppresses the presence of an information rather than modifying it.
- On tabular data LIME generates the neighborhood by changing the feature values with other values of the domain.
  - x = {age=24, sex=male, income=1000} (x = x’)
  - z = {age=30 , sex=male, income=800} (z = z’)

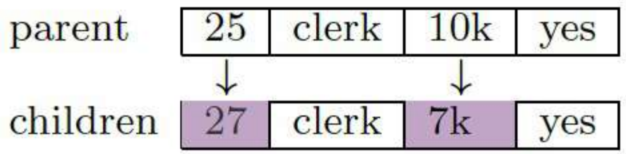
LORE: LOcal Rule-based Explainer

- LORE extends LIME adopting as local surrogate a decision tree classifier and by generating synthetic instances through a genetic procedure that accounts for both instances with the same labels and different ones.
- It can be generalized to work on images and text using the same data representation adopted by LIME.



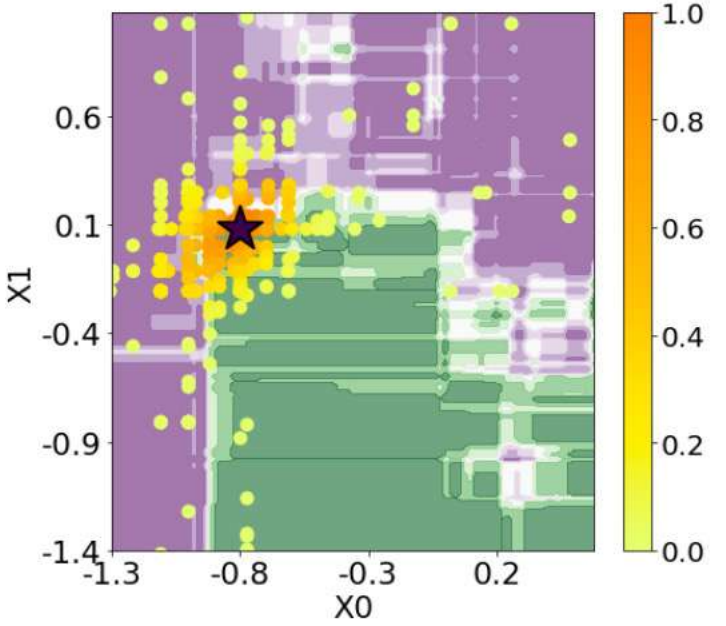
LORE Pseudo - Code

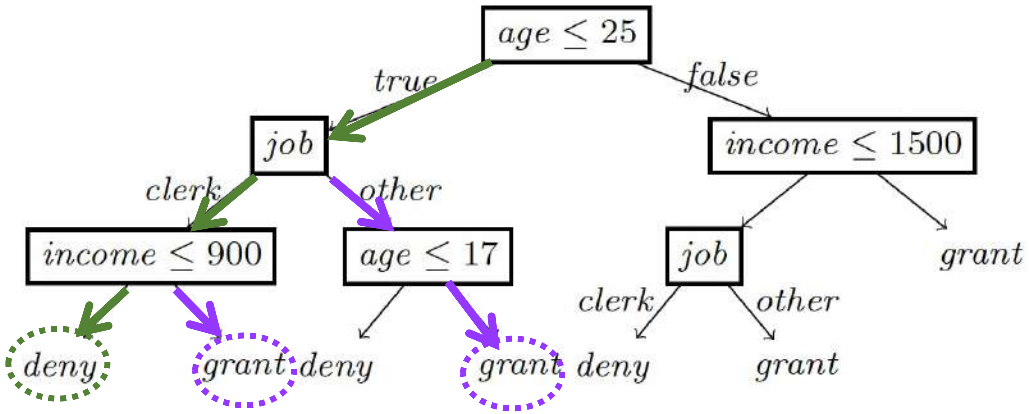





```
x instance to explain
Z= = geneticNeighborhood(x, fitness=, N/2)
Z≠ = geneticNeighborhood(x, fitness≠, N/2)
Z = Z= ∪ Z≠
c = buildTree(Z, b(Z))
r = (p -> y) = extractRule(c, x)
ϕ = extractCounterfactual(c, r, x)
return e = <r, ϕ>
```

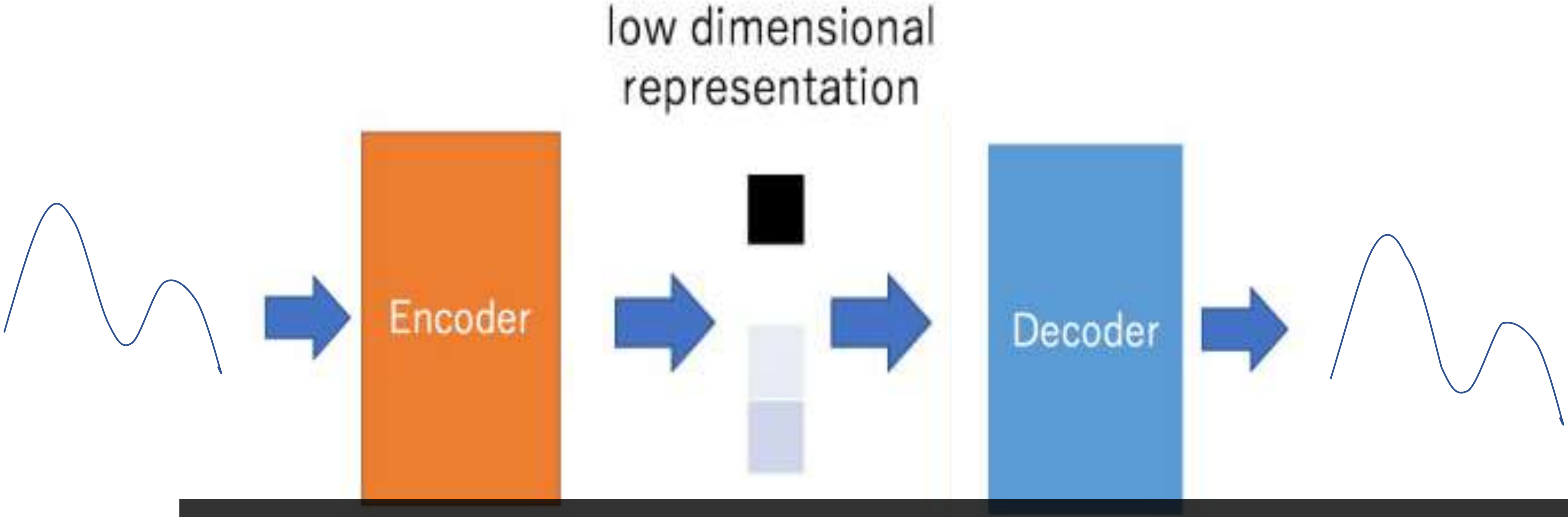
black box auditing





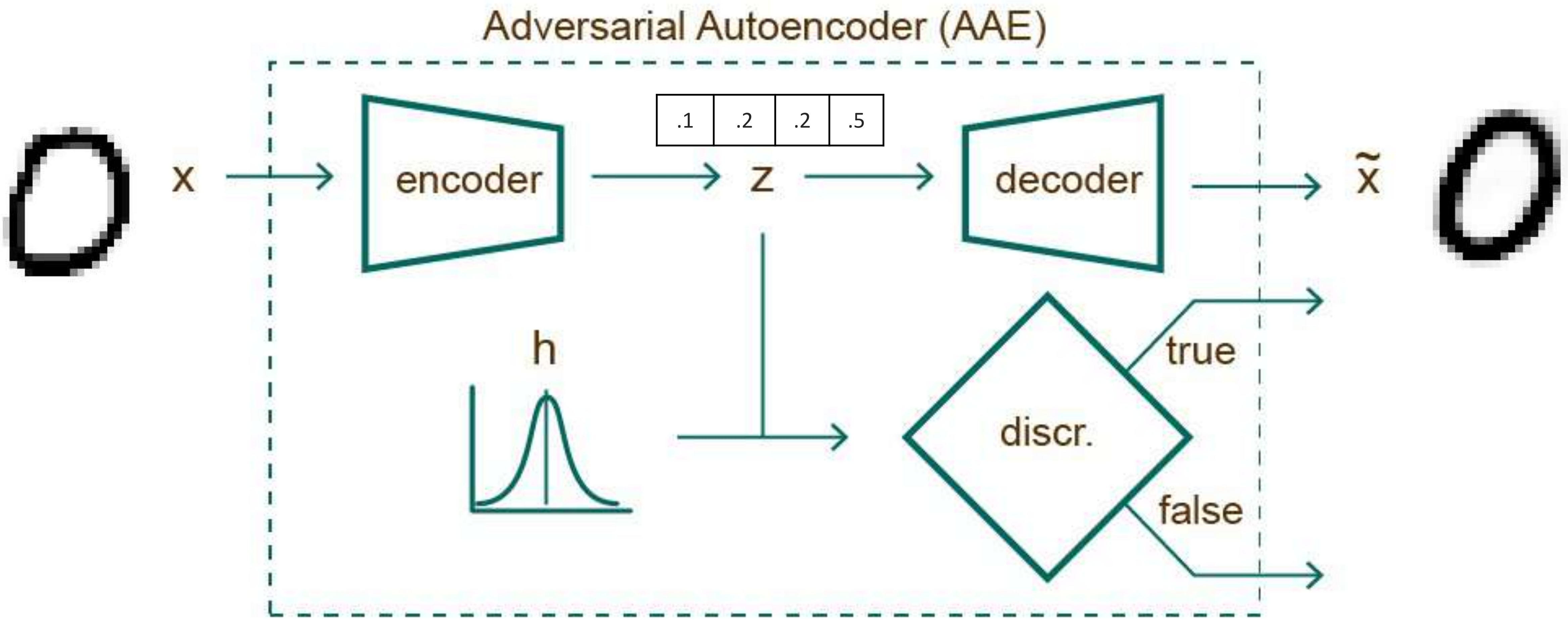
r = {age ≤ 25, job = clerk, income ≤ 900} -> deny

Φ = {({income > 900} -> grant),
({17 ≤ age < 25, job = other} -> grant)}

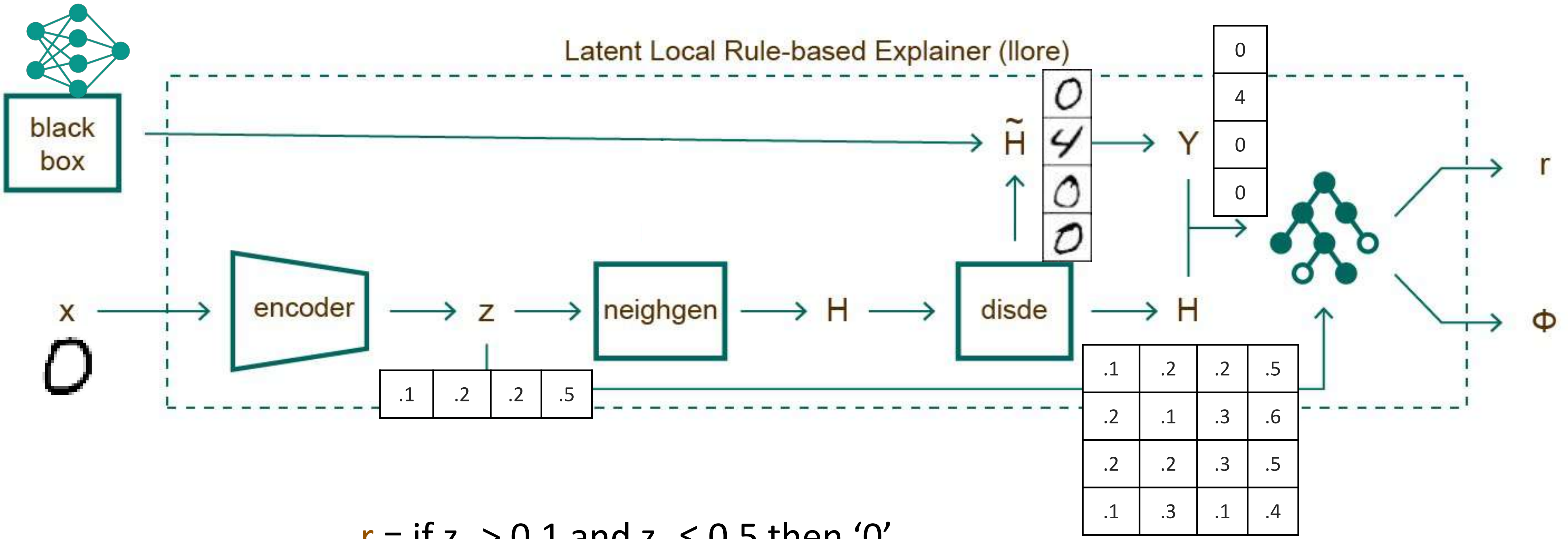


Explaining with Autoencoders

Adversarial Autoencoder



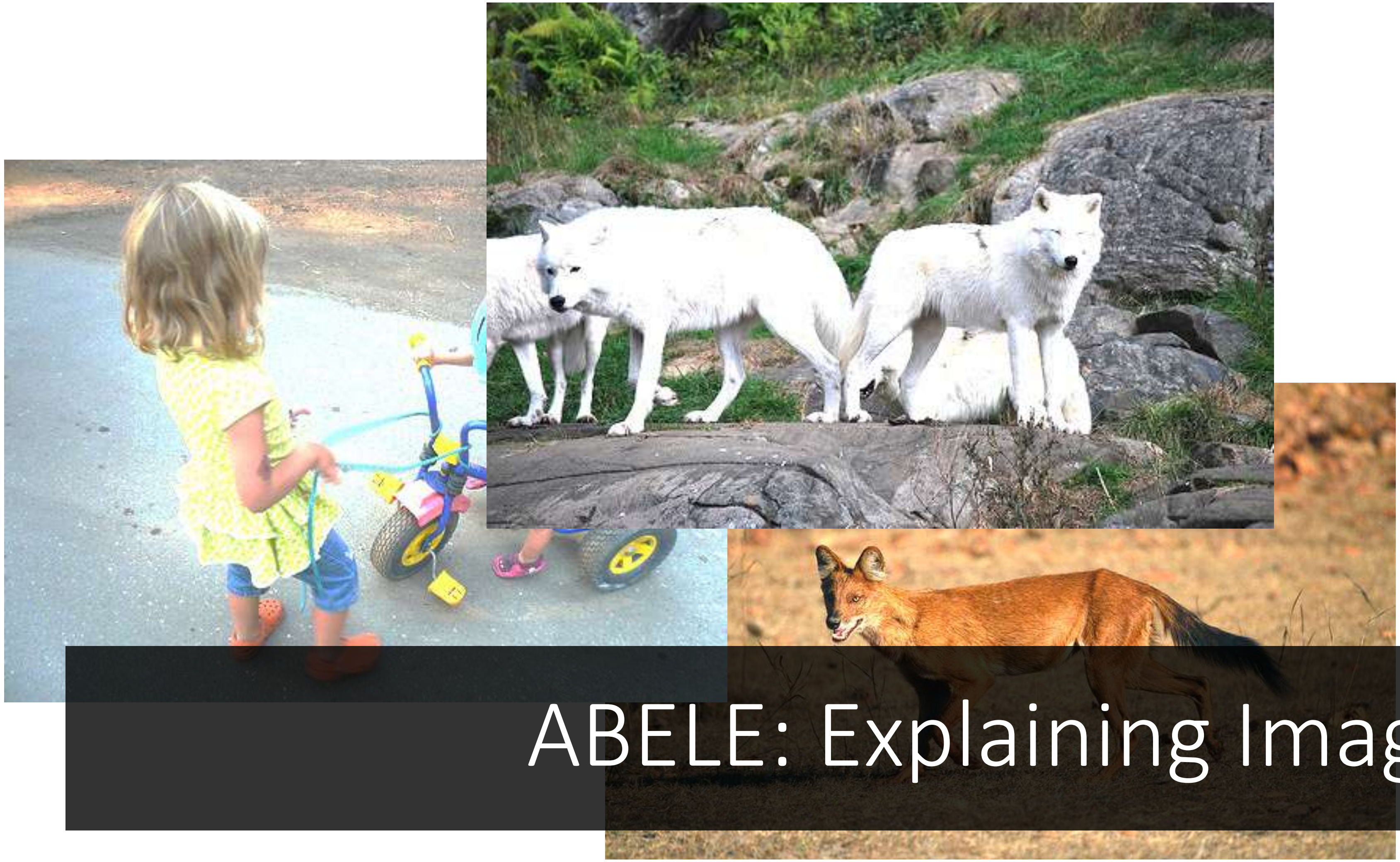
Latent Local Rule Extraction



r = if $z_1 > 0.1$ and $z_3 \leq 0.5$ then '0'

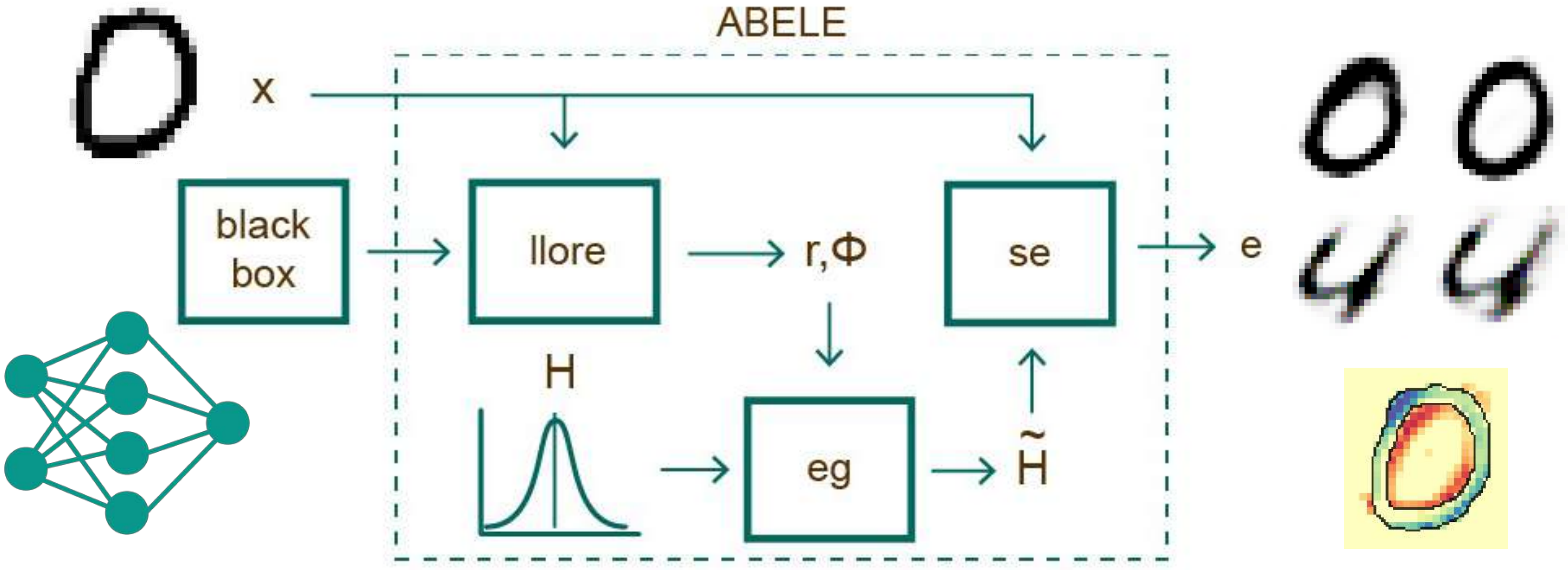
Φ = {({$z_1 \leq 0.1$} --> '4'), ({$z_3 > 0.5$} --> '8')}

- Guidotti, Riccardo, et al. ***Black Box Explanation by Learning Image Exemplars in the Latent Feature Space.*** ECML-PKDD, 2019.



ABELE: Explaining Image Classifiers

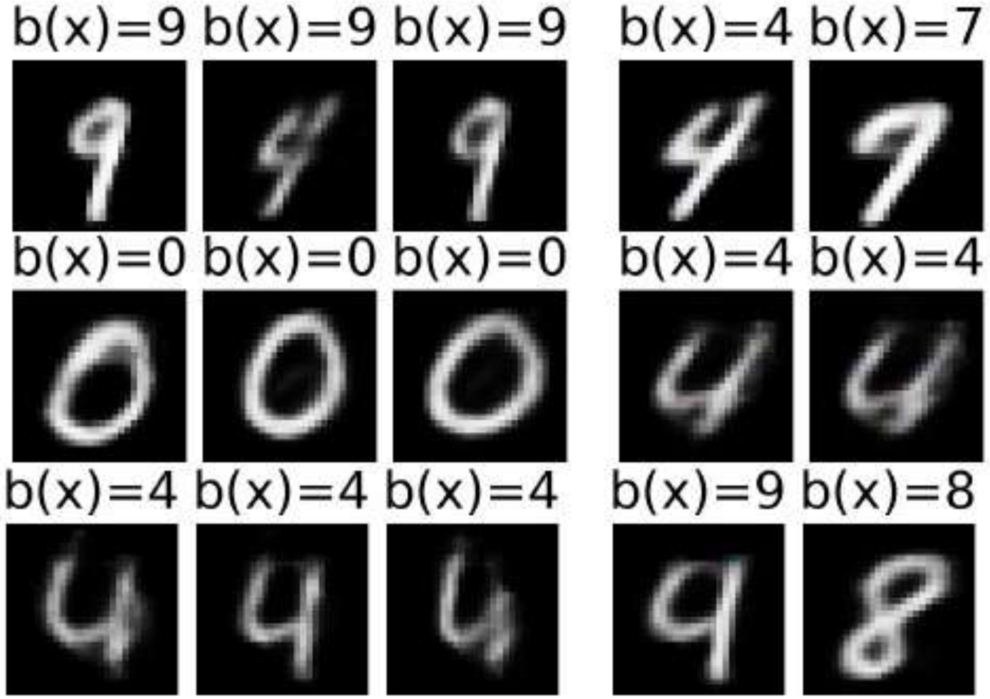
ABELE: Adversarial Black box Explainer generating Latent Exemplars



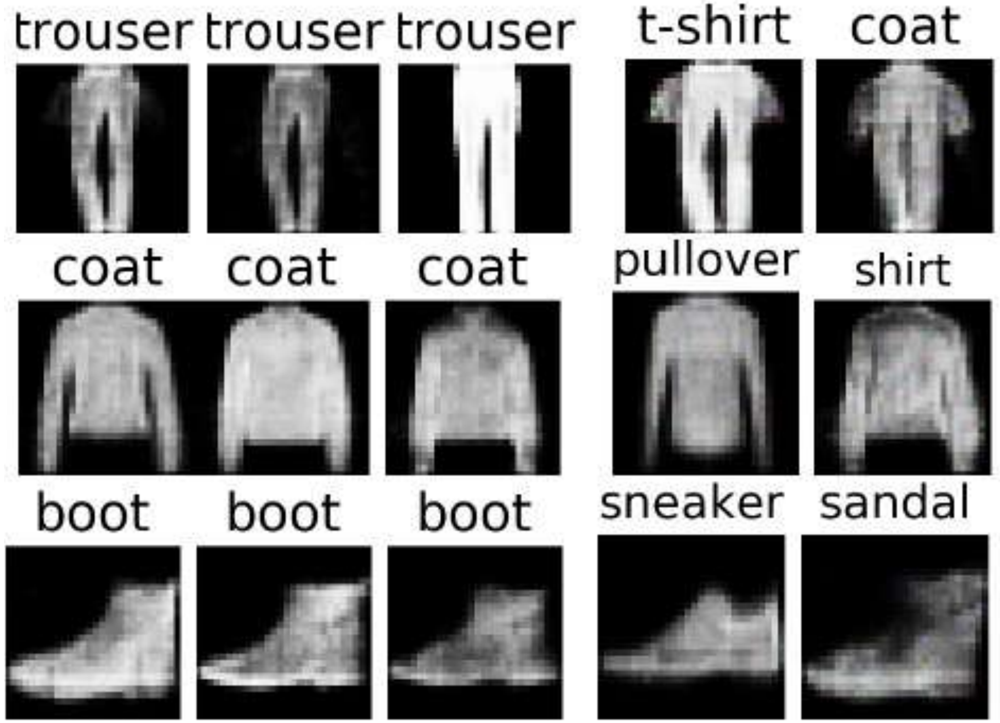
- Guidotti, Riccardo, et al. ***Black Box Explanation by Learning Image Exemplars in the Latent Feature Space.*** ECML-PKDD, 2019.

Exemplars and Counter-Exemplars

- **mnist**

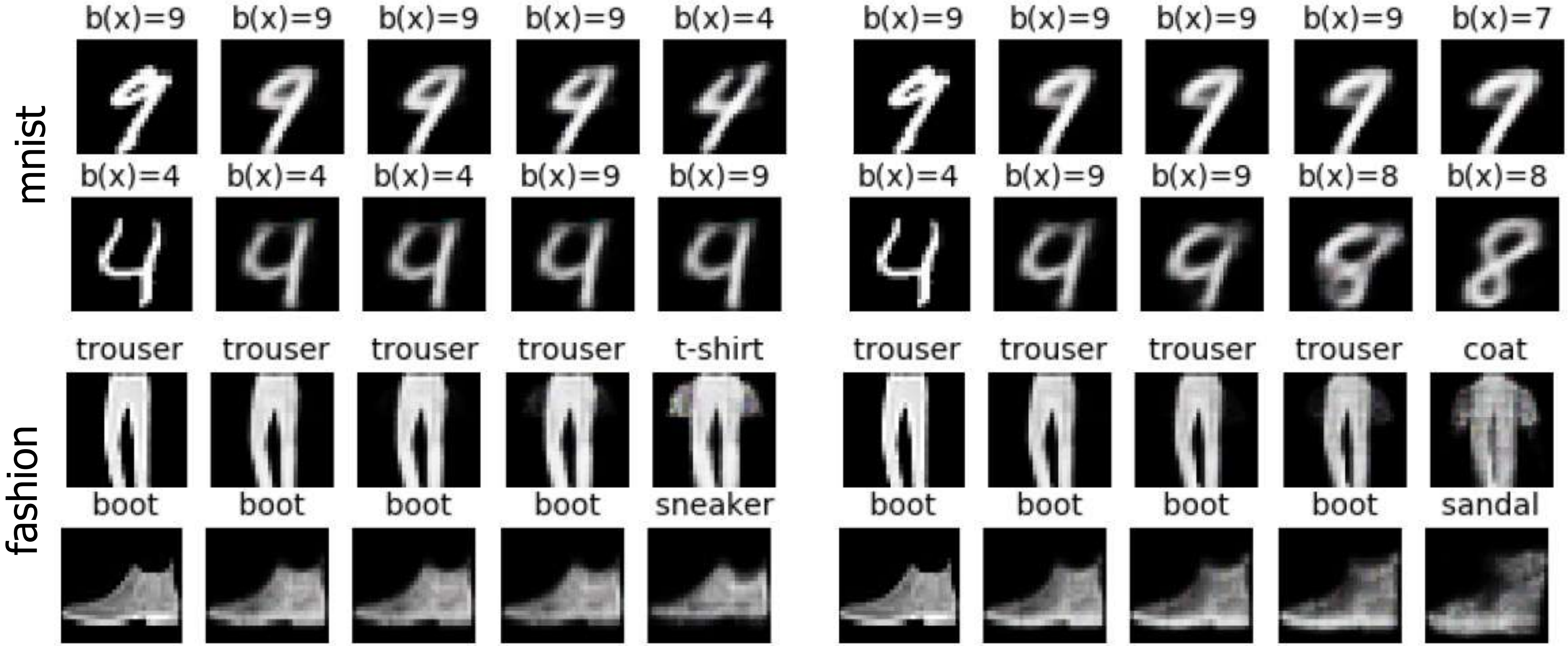


- **fashion**



From Image to Counter-Exemplar

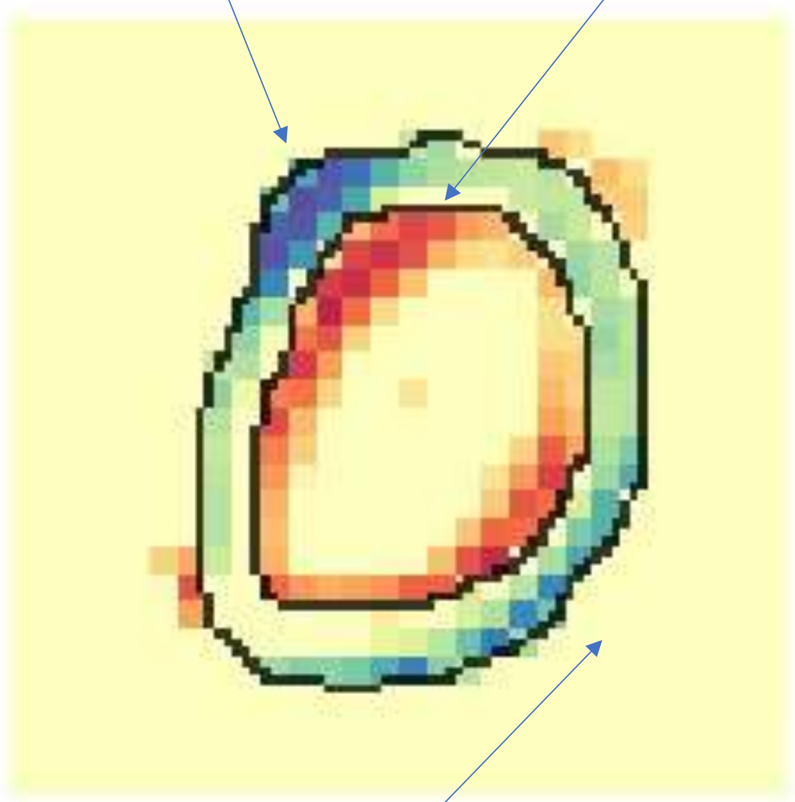
- T. Spinner et al. Towards an interpretable latent space: an intuitive comparison of autoencoders with variational autoencoders. In IEEE VIS 2018, 2018.



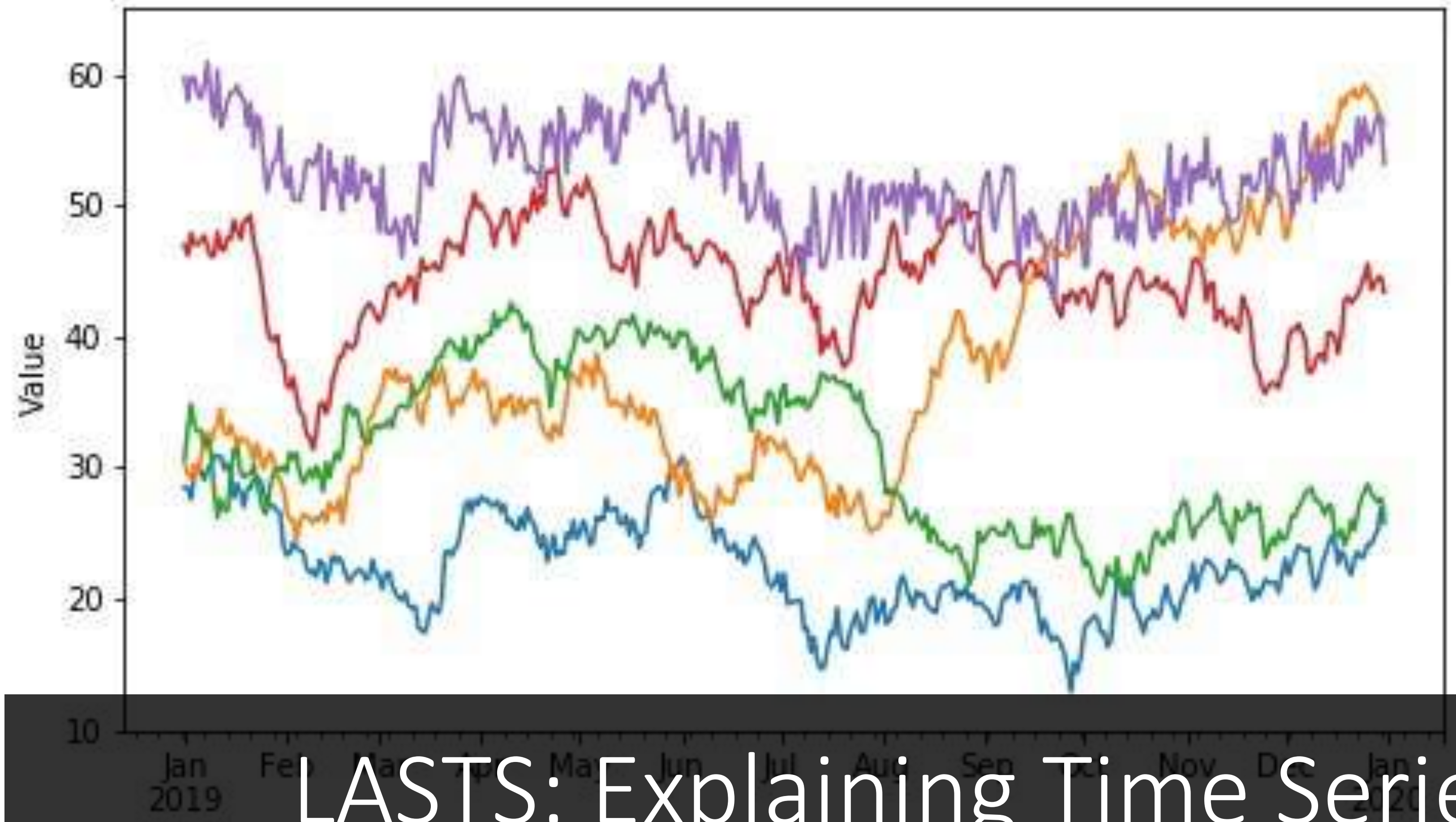
Saliency Map from Exemplars

- The saliency map ***s*** highlights areas of **x** that contribute to ***b(x)*** and that push it to ≠ ***b(x)***.
- It is obtained as follows:
  - pixel-to-pixel-difference between ***x*** and each exemplar in $\widetilde{\boldsymbol{H}}$
  - each pixel of ***s*** is the median value of the differences calculated for that pixel.

Red/Blue means consistent difference “variable area”

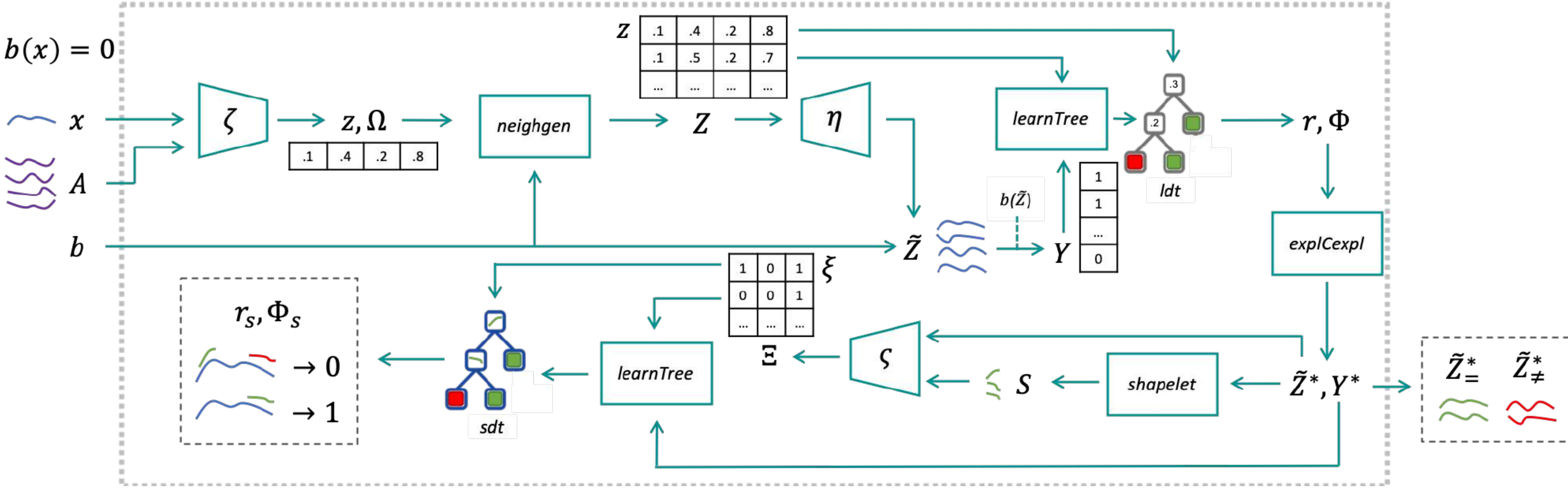


Yellow means no difference “no change area”



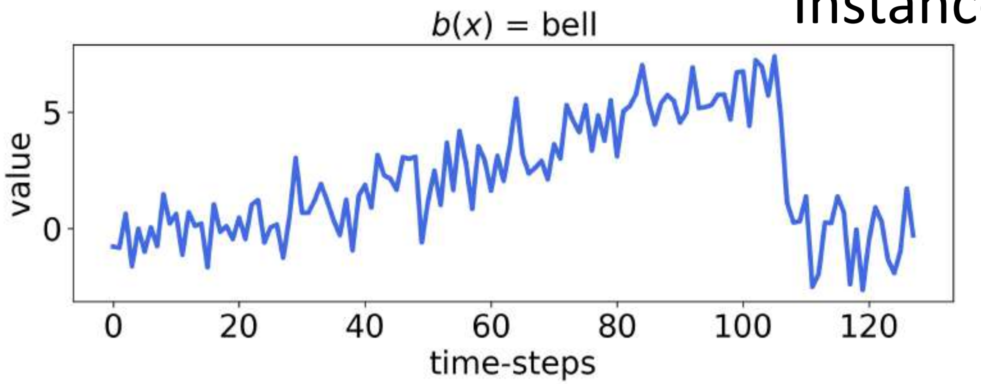
LASTS: Explaining Time Series Classifiers

LASTS: Local Agnostic Shapelet-based Time Series explainer

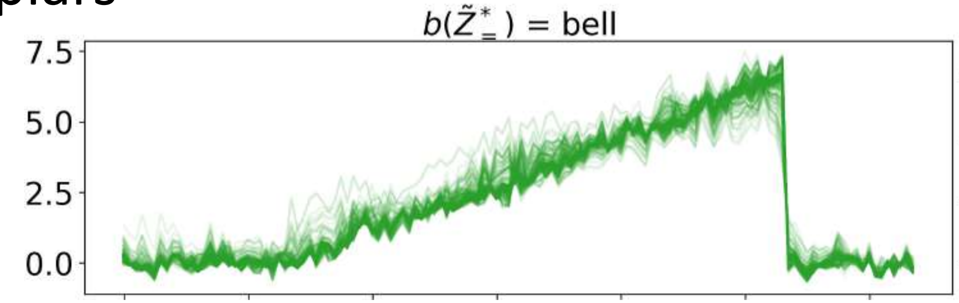


LASTS Explanation

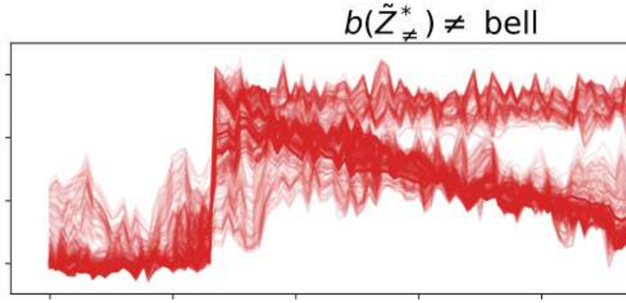
Instance to explain



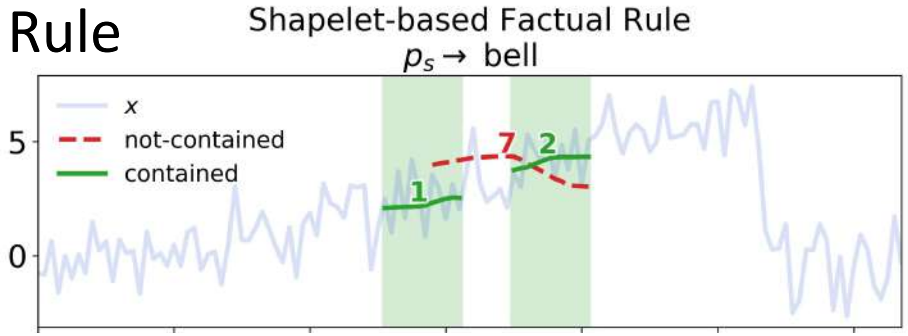
Exemplars



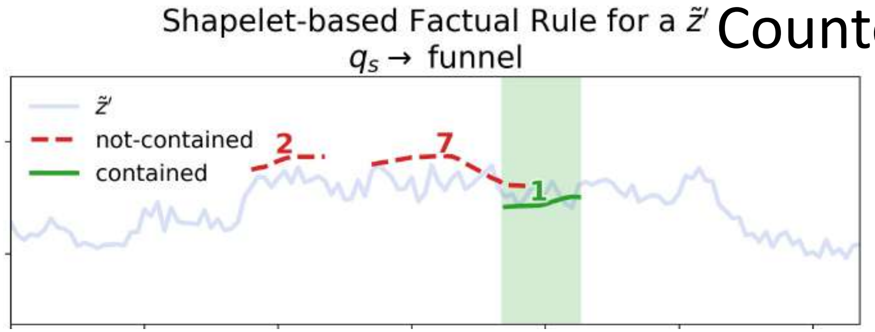
Counter-Exemplars



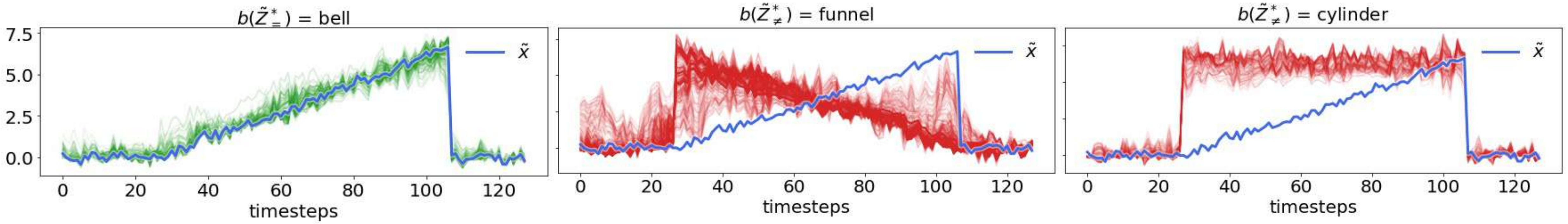
Factual Rule

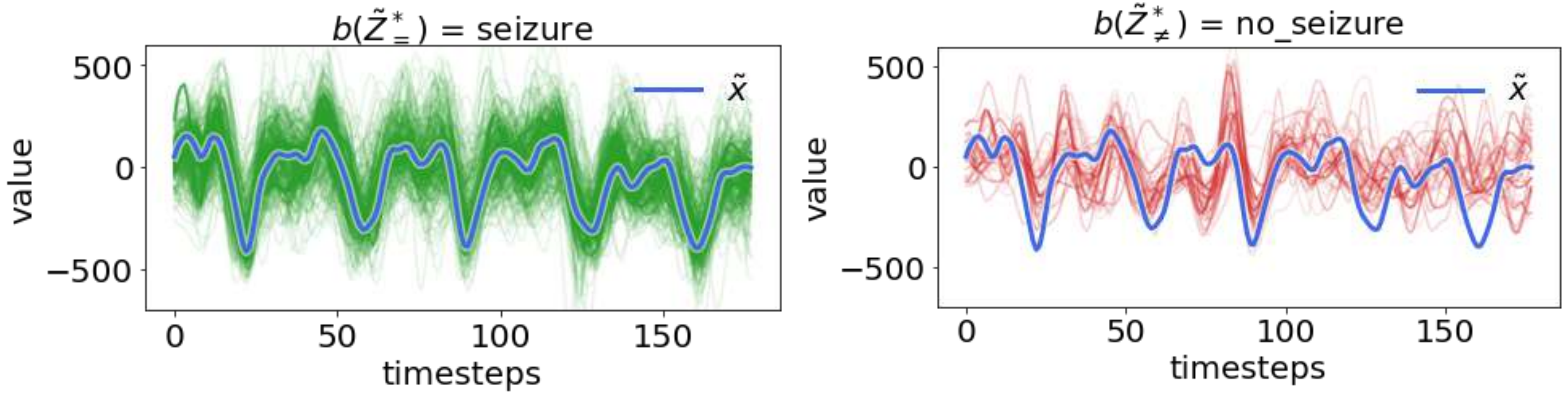


Counter-Factual Rule

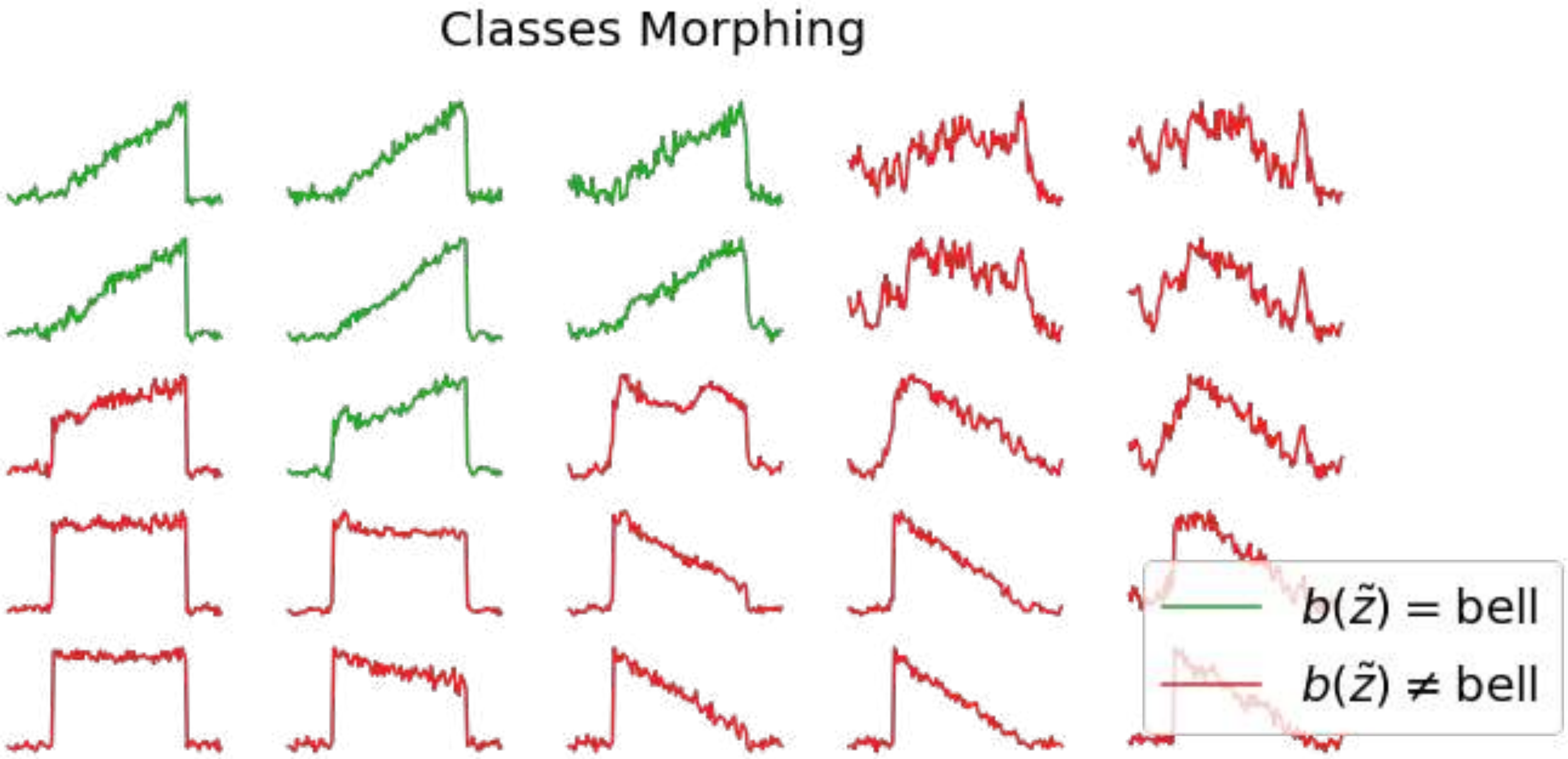


Exemplars and Counter Exemplars

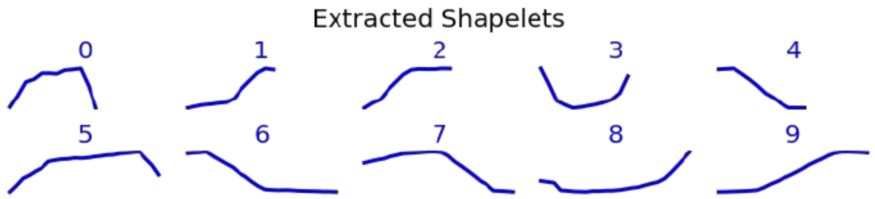


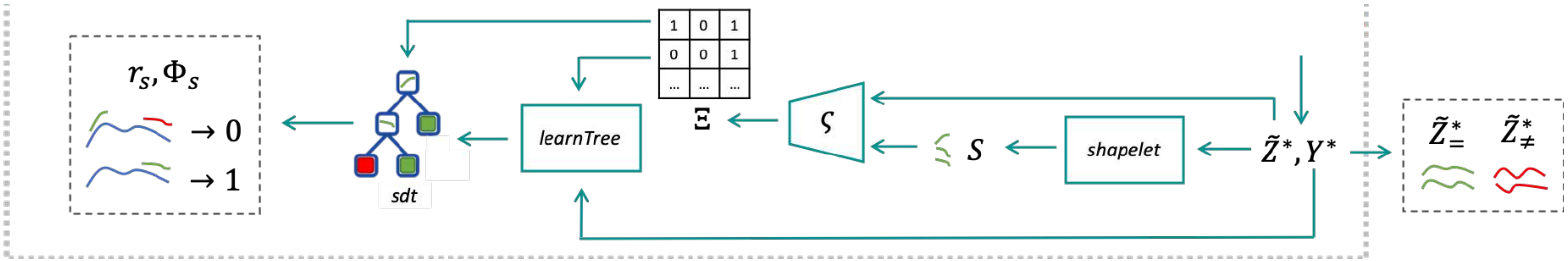


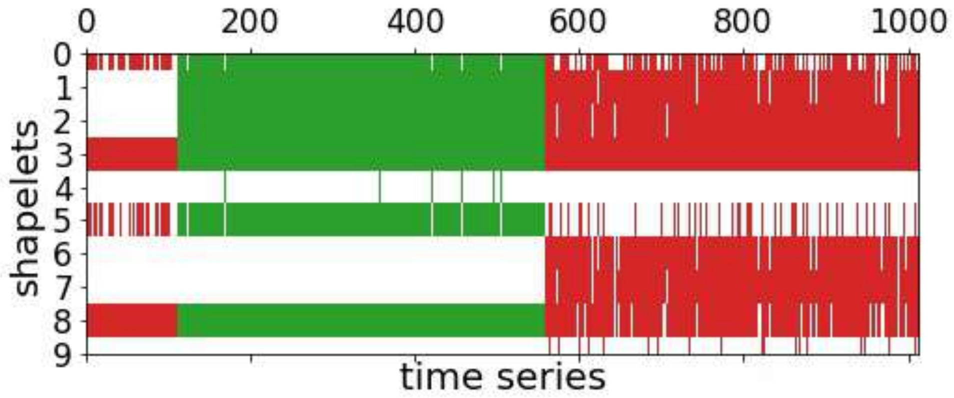
From Exemplars to Counter-Exemplars

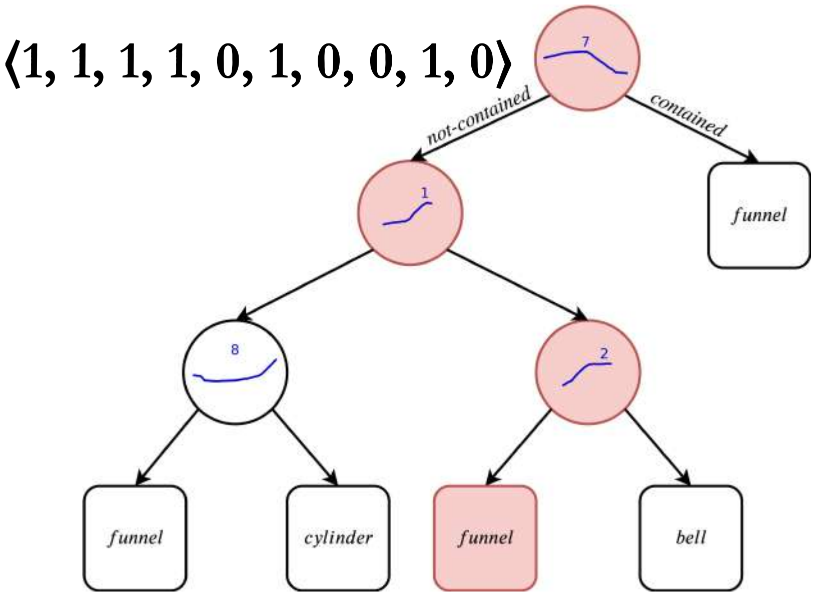


Shapelet-Based Rule Extraction









$\mathbf{r}_s$ IF *s7* is NOT contained AND *s1* AND *s2* are contained

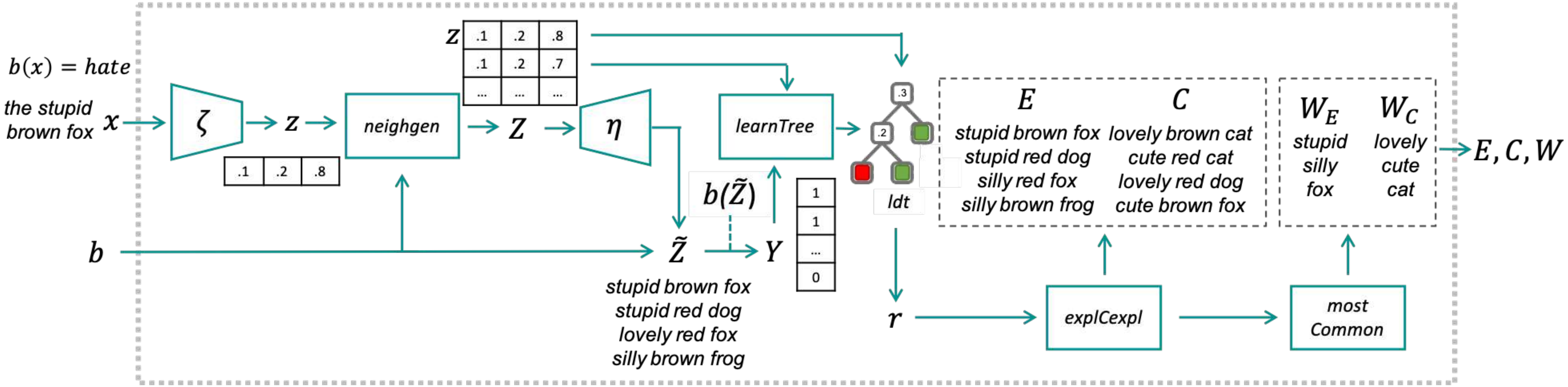
bell

$\mathbf{\Phi}_s$ IF *s2* AND *s7* are NOT contained AND *s1* is contained

funnel

XSPELLS: Explaining Text Classifiers

X-SPELLS: eXplaining Sentiment Prediction generating ExempLars in the Latent Space



Take Home Message

- Idea: Enable model and data agnostic local explanations through autoencoders using simple and effective methods.

- In turns, different and complementary types of explanations become available.

Thank you

riccardo.guidotti@di.unipi.it

XAI

European Research Council
Established by the European Commission

ERC-AdG-2019 “Science & technology for the eXplanation of AI decision making”

References

- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
- Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2021). Benchmarking and Survey of Explanation Methods for Black Box Models. arXiv preprint arXiv:2102.13076.
- Molnar, C. (2020). Interpretable machine learning. Lulu. com.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEWEE. 2018.
- Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations--A survey. arXiv preprint arXiv:1911.07749.

- Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations--A survey. arXiv preprint arXiv:1911.07749.

- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080.
- Zhang, Y., & Chen, X. (2018). Explainable recommendation: A survey and new perspectives. arXiv preprint arXiv:1804.11192.

References

- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.
- Ribeiro, M. T., et al. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD. 2016
- Lundberg, S., & Lee, S. I. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. 2017
- Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14-23.
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019, July). Meaningful explanations of Black Box AI decision systems. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9780-9784).
- Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2019, September). Black box explanation by learning image exemplars in the latent feature space. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 189-205). Springer, Cham.
- Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX-From Local to Global Explanations of Black Box AI Models. Artificial Intelligence, 294, 103457.
- Guidotti, R. (2021). Evaluating local explanation methods on ground truth. Artificial Intelligence, 291, 103428.

References

- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.
- Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279-288).
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- Freitas, A. A. (2014). Comprehensible classification models: a position paper. ACM SIGKDD explorations newsletter, 15(1), 1-10.
- Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(5), 582-638.
- Craven, M. W., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. Advances in neural information processing systems, 24-30.
- Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse engineering the neural networks for rule extraction in classification problems. Neural processing letters, 35(2), 131-150.
- Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3429-3437).
- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020, February). FACE: feasible and actionable counterfactual explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 344-350).

References

- Cortez, P., & Embrechts, M. J. (2011, April). Opening black box data mining models using sensitivity analysis. In 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (pp. 341-348). IEEE.
- Kim, B., Gilmer, J., Wattenberg, M., & Viégas, F. (2018). Tcav: Relative concept importance testing with linear concept activation
- Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617). vectors.
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. arXiv preprint arXiv:1704.01701.
- Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328). PMLR.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.
- Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018, July). Learning to explain: An information-theoretic perspective on model interpretation. In International Conference on Machine Learning (pp. 883-892). PMLR.
- Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. arXiv preprint arXiv:1802.07623.

Explanation Toolboxes and Repositories

- https://github.com/jphall663/awesome-machine-learning-interpretability
- https://github.com/pbiecek/xai_resources
- https://github.com/ModelOriented/DrWhy
- https://fat-forensics.org/
- https://github.com/Trusted-AI/AIX360
- https://captum.ai/
- https://github.com/interpretml/interpret
- https://github.com/SeldonIO/alibi
- https://github.com/pair-code/what-if-tool