Explainability & Transparency

UNIVERSITÀ DI PISA

What is “Explainable AI” ?

- **Explainable-AI** explores and investigates methods to produce or complement **AI models** to make **accessible and interpretable** the internal logic and the outcome of the algorithms, making such process **understandable by humans**.

- **Explicability**, understood as incorporating both **intelligibility** *(“how does it work?”)* for non-experts, e.g., patients or business customers, and for experts, e.g., product designers or engineers) and **accountability** (*“who is responsible for”*).

Interpretability

- To ***interpret*** means to give or provide the meaning or to explain and present in understandable terms some concepts.

- In data mining and machine learning, interpretability is the ***ability to explain*** or to provide the meaning ***in understandable terms to a human***.

- https://www.merriam-webster.com/
- Finale Doshi-Velez and Been Kim. 2017. ***Towards a rigorous science of interpretable machine learning***. arXiv:1702.08608v2.

Motivating Examples

- Criminal Justice
  - People wrongly denied
  - Recidivism prediction
  - Unfair Police dispatch
- Finance:
  - Credit scoring, loan approval
  - Insurance quotes
- Healthcare
  - AI as 3rd-party actor in physician - patient relationship
  - Learning must be done with available data: cannot randomize cares given to patients!
  - Must validate models before use.

Opinion

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

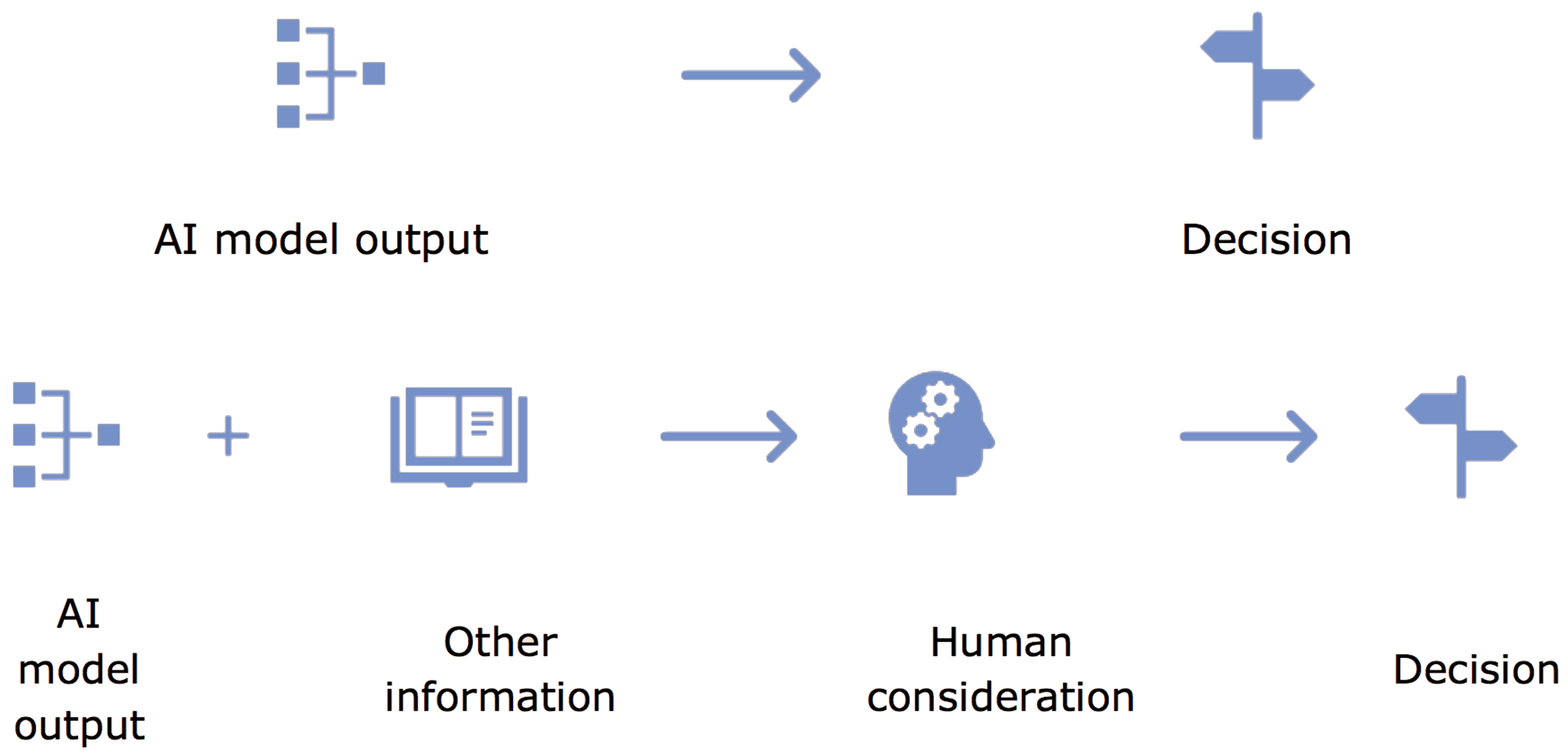
Stanford MEDICINE | News Center

Email Tweet

Researchers say use of artificial intelligence in medicine raises ethical questions

In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.

What is AI-assisted decision making?



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.

- 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.

Needs For Interpretable Models

Right of Explanation

Since 25 May 2018, GDPR establishes a right for all individuals to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making”, including profiling, takes place.

COMPAS recidivism black bias

Military tank classification depends on the background

Summarizing: the Need to Explain comes from …

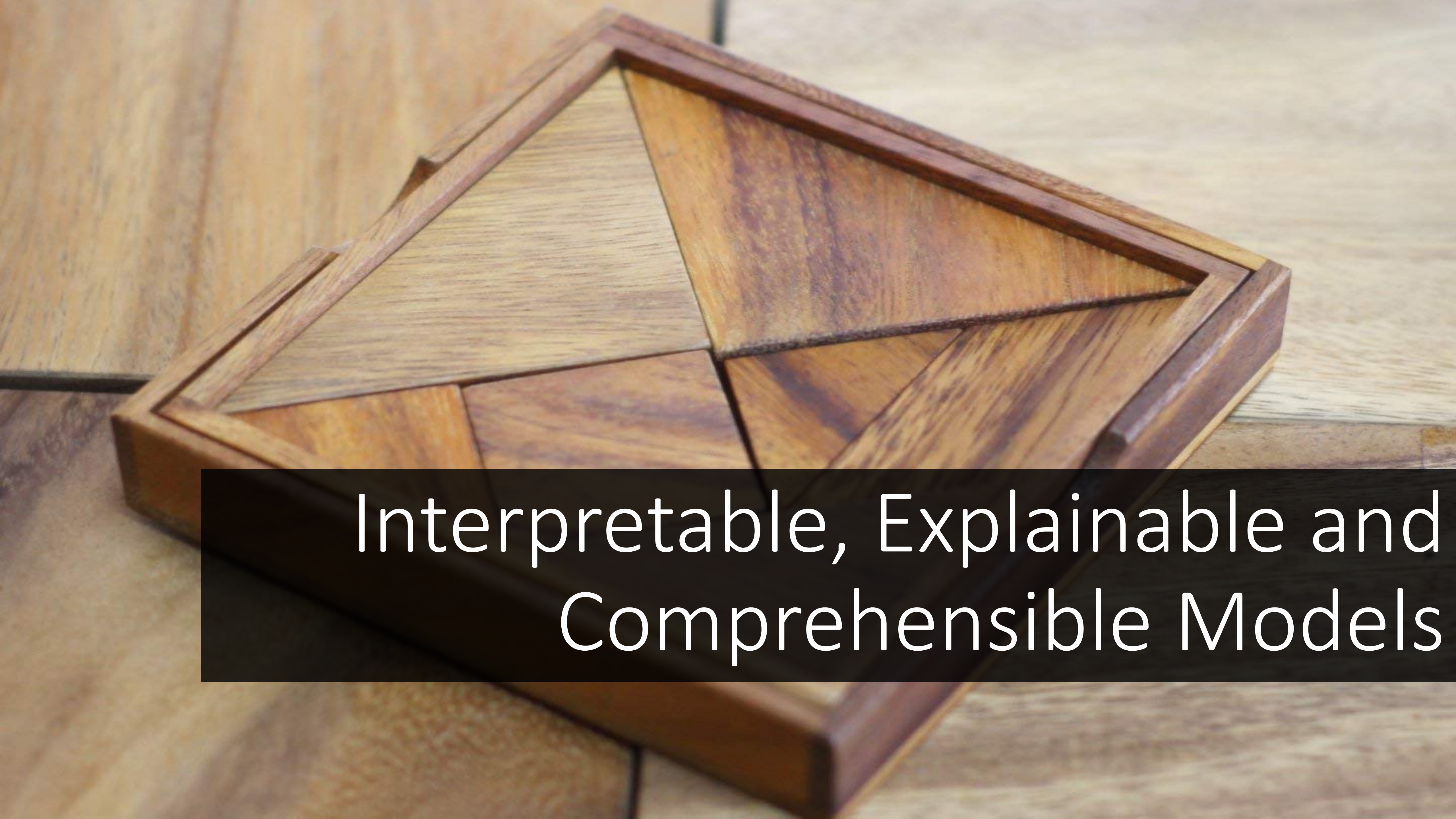
- User Acceptance & Trust [Lipton 2016, Ribeiro 2016, Weld and Bansal 2018]

- Legal
  - Conformance to ethical standards, fairness
  - *Right to be informed* [Goodman and Flaxman 2016, Wachter 2017]
  - Contestable decisions

- Explanatory Debugging [Kulesza et al. 2014, Weld and Bansal 2018]
  - Flawed performance metrics
  - Inadequate features
  - Distributional drift

Science and technology for the eXplanation of AI decision making

Explainable AI is the basic building brick
for **preserving and expanding human autonomy**,
and helping humans make better decisions



Interpretable, Explainable and Comprehensible Models

Dimensions of Interpretability

- ***Global and Local Interpretability***:
  - *Global*: understanding the whole logic of a model
  - *Local*: understanding only the reasons for a specific decision

- ***Time Limitation***: the time that the user can spend for understanding an explanation.

- ***Nature of User Expertise***: users of a predictive model may have different background knowledge and experience in the task. The nature of the user expertise is a key aspect for interpretability of a model.

Desiderata of an Interpretable Model

- ***Interpretability*** *(or* comprehensibility*)*: to which extent the model and/or its predictions are human understandable. Is measured with the *complexity* of the model.

- ***Fidelity***: to which extent the model imitate a black-box predictor.

- ***Accuracy***: to which extent the model predicts unseen instances.

- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

Desiderata of an Interpretable Model

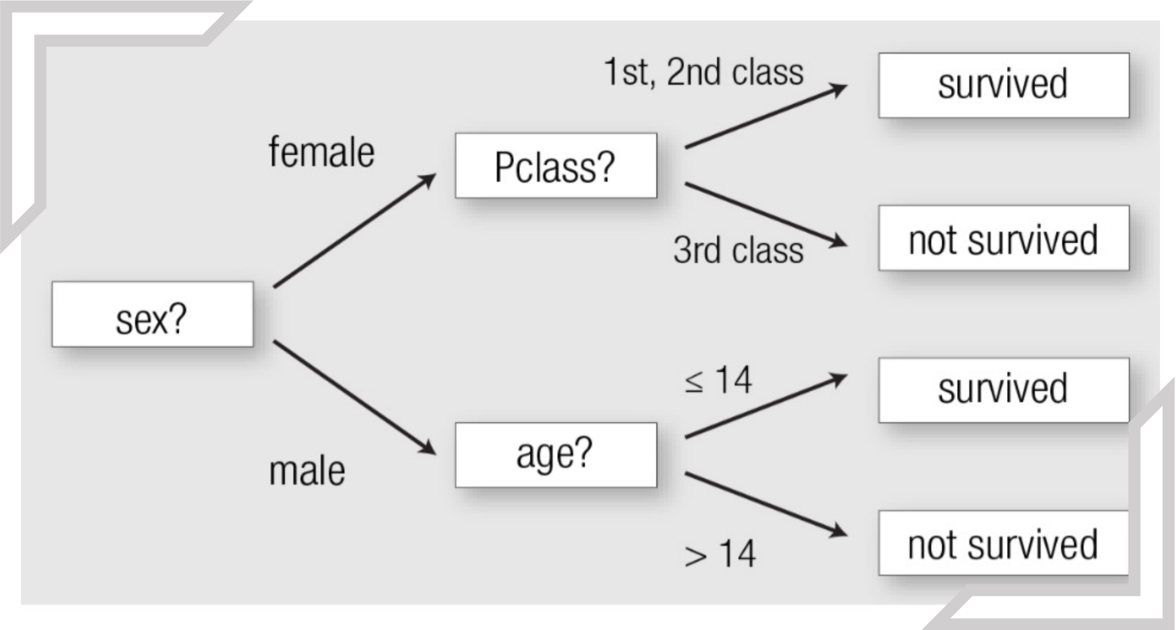
- ***Fairness***: the model guarantees the protection of groups against discrimination.
- ***Privacy***: the model does not reveal sensitive information about people.
- ***Respect Monotonicity***: the increase of the values of an attribute either increase or decrease in a monotonic way the probability of a record of being member of a class.
- ***Usability***: an interactive and queryable explanation is more usable than a textual and fixed explanation.

- Andrea Romei and Salvatore Ruggieri. 2014. ***A multidisciplinary survey on discrimination analysis***. Knowl. Eng.
- Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. 2015. ***A comprehensive review on privacy preserving data mining***. SpringerPlus .
- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

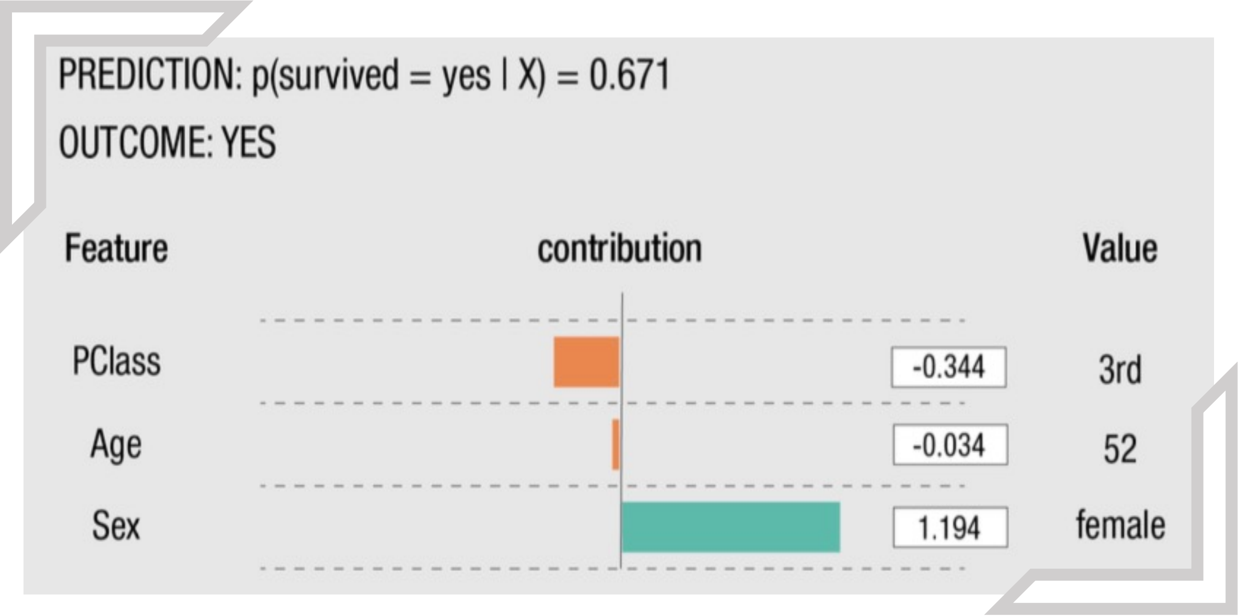
Desiderata of an Interpretable Model

- ***Reliability and Robustness***: the interpretable model should maintain high levels of performance independently from small variations of the parameters or of the input data.
- ***Causality:*** controlled changes in the input due to a perturbation should affect the model behavior.
- ***Scalability:*** the interpretable model should be able to scale to large input data with large input spaces.
- ***Generality:*** the model should not require special training or restrictions.

Recognized Interpretable Models



Decision Tree

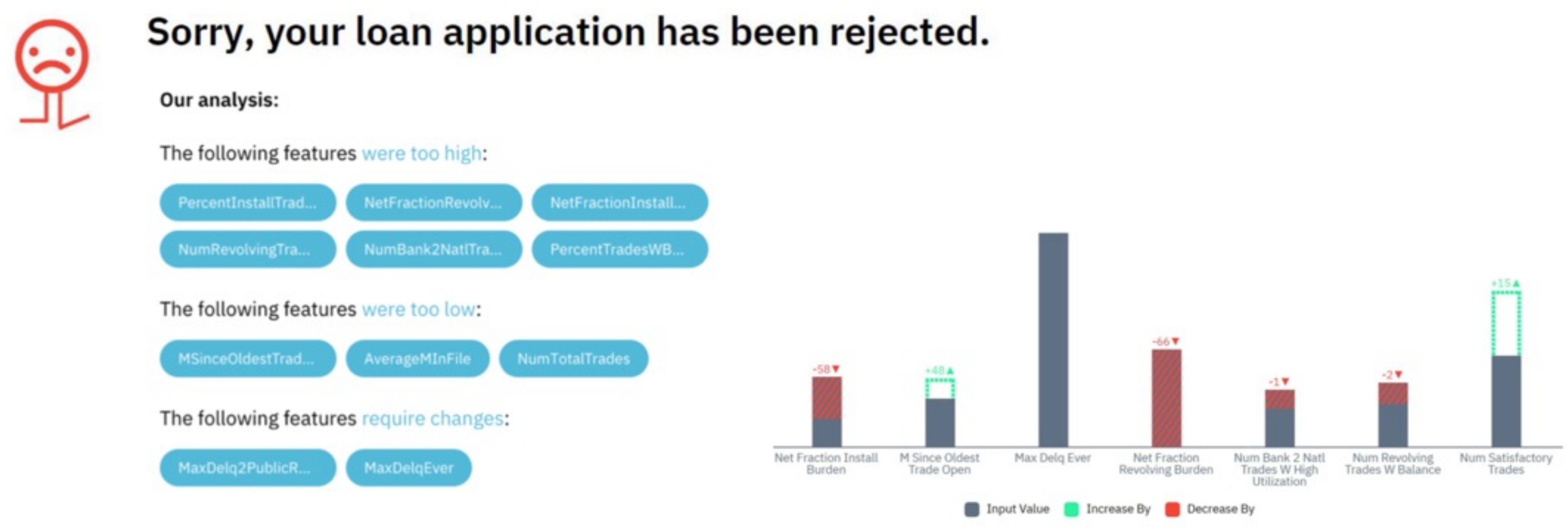


Linear Model

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

Rules

There are several kinds of explanations



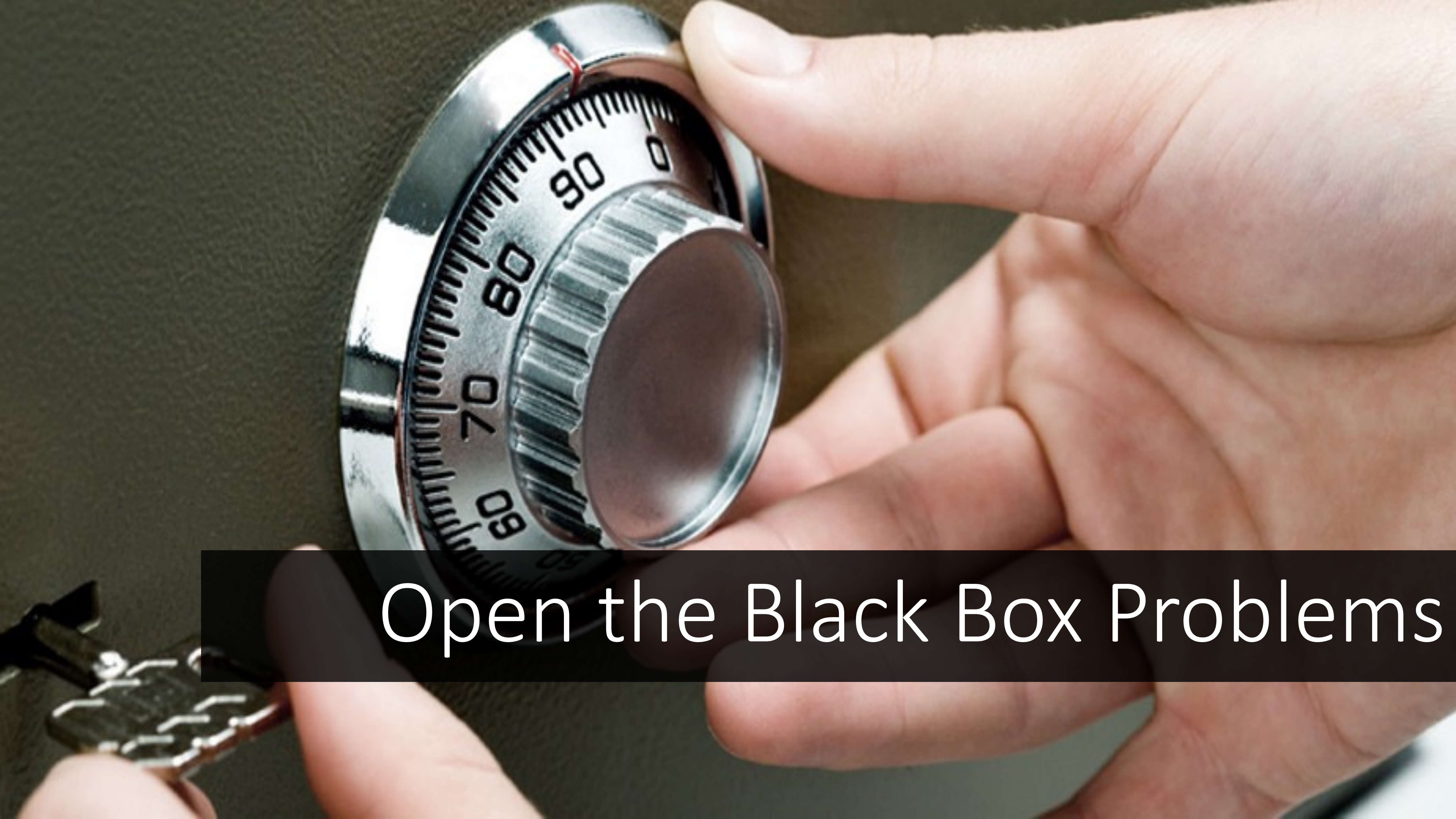
Counterfactuals suggest where to increase (green, dashed) or decrease (red, striped) each feature.

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Complexity

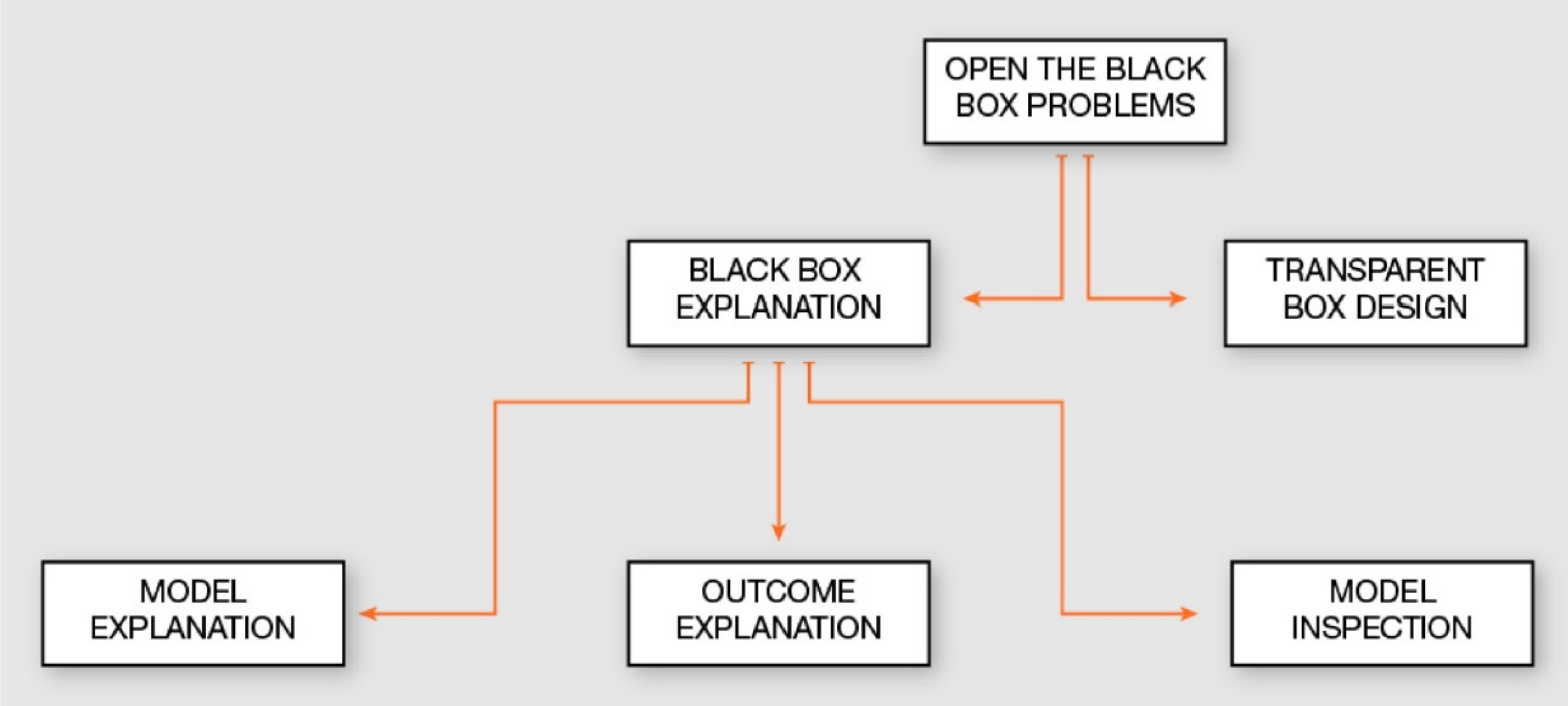
- Opposed to *interpretability*.
- Is only related to the model and not to the training data that is unknown.
- Generally estimated with a rough approximation related to the ***size*** of the interpretable model.
- Linear Model: number of non zero weights in the model.
- Rule: number of attribute-value pairs in condition.
- Decision Tree: estimating the complexity of a tree can be hard.

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ***Why should i trust you?: Explaining the predictions of any classifier***. KDD.
- Houtao Deng. 2014. ***Interpreting tree ensembles with intrees***. arXiv preprint arXiv:1408.5456.
- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

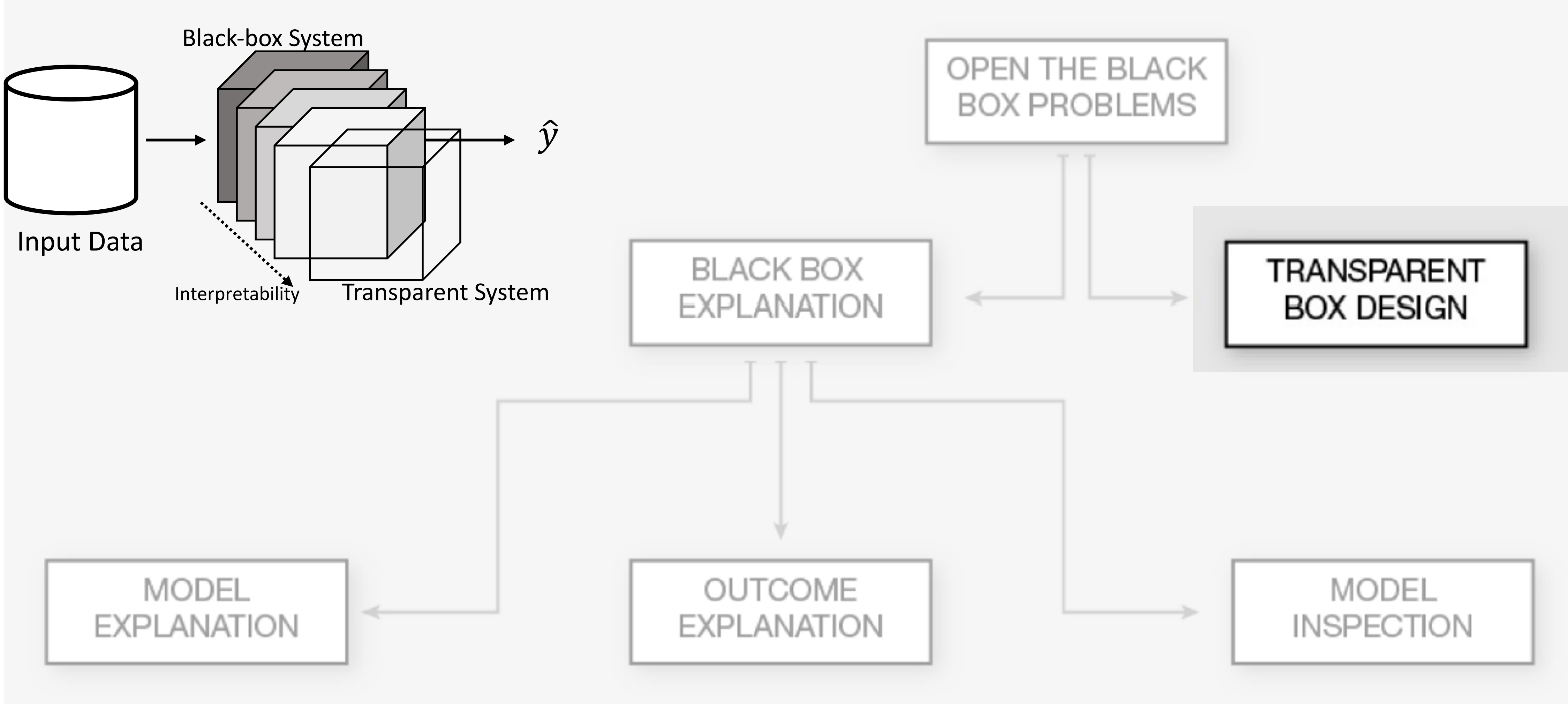


Open the Black Box Problems

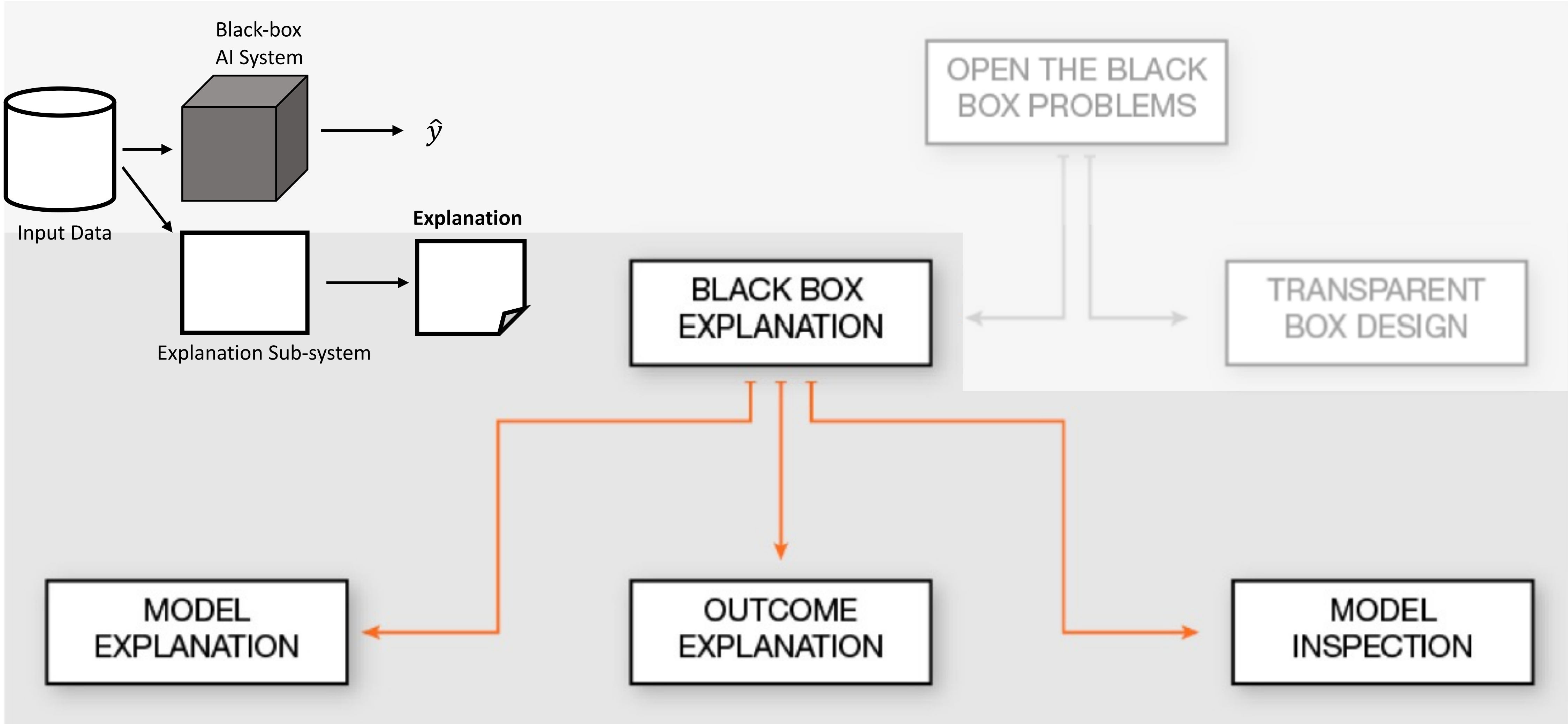
Problems Taxonomy



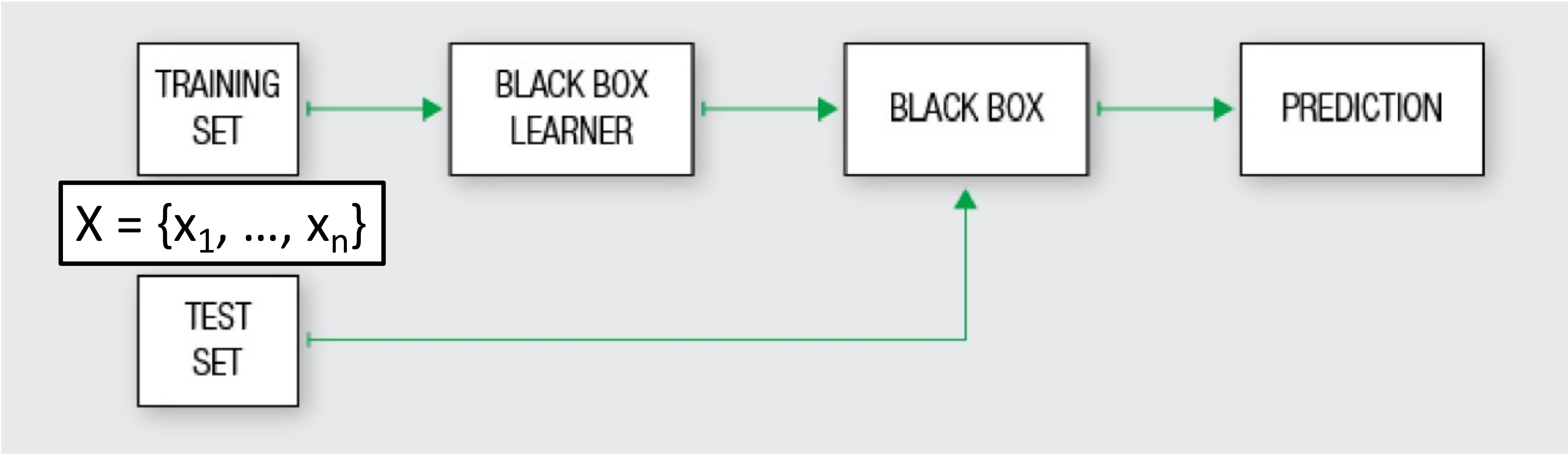
XbD – eXplanation by Design



BBX - Black Box eXplanation

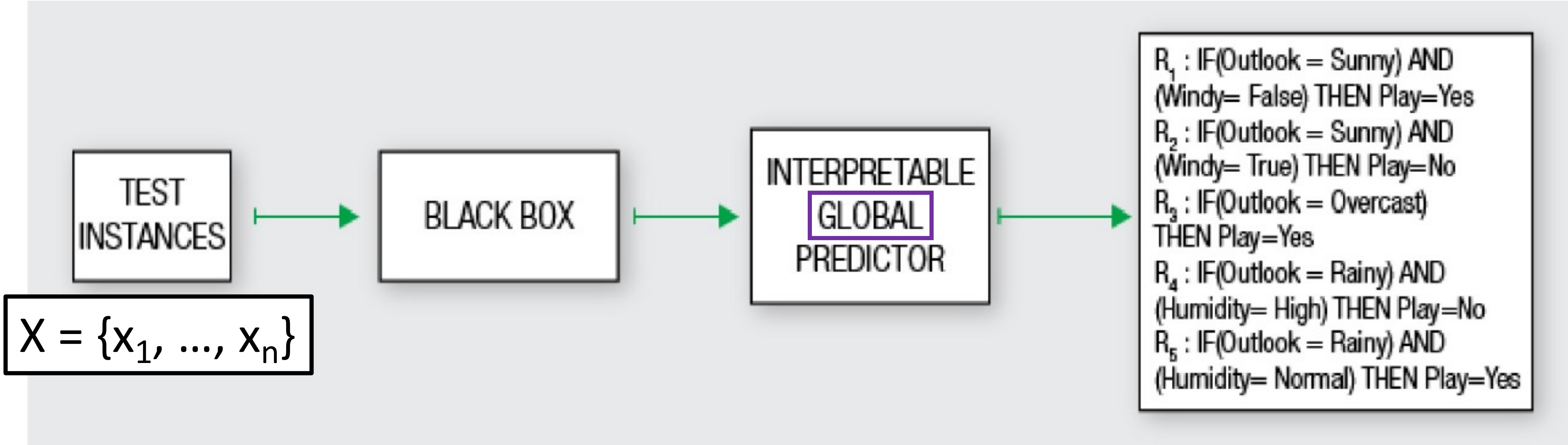


Classification Problem



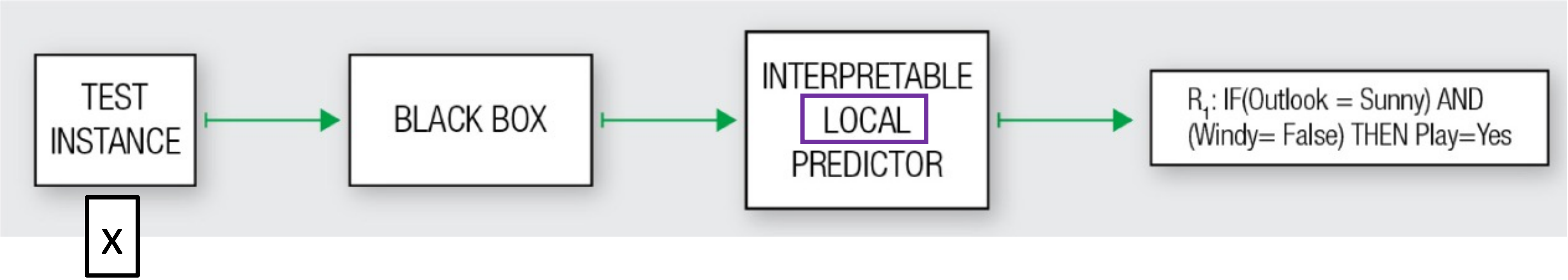
Model Explanation Problem

Provide an interpretable model able to mimic the ***overall logic/behavior*** of the black box and to explain its logic.



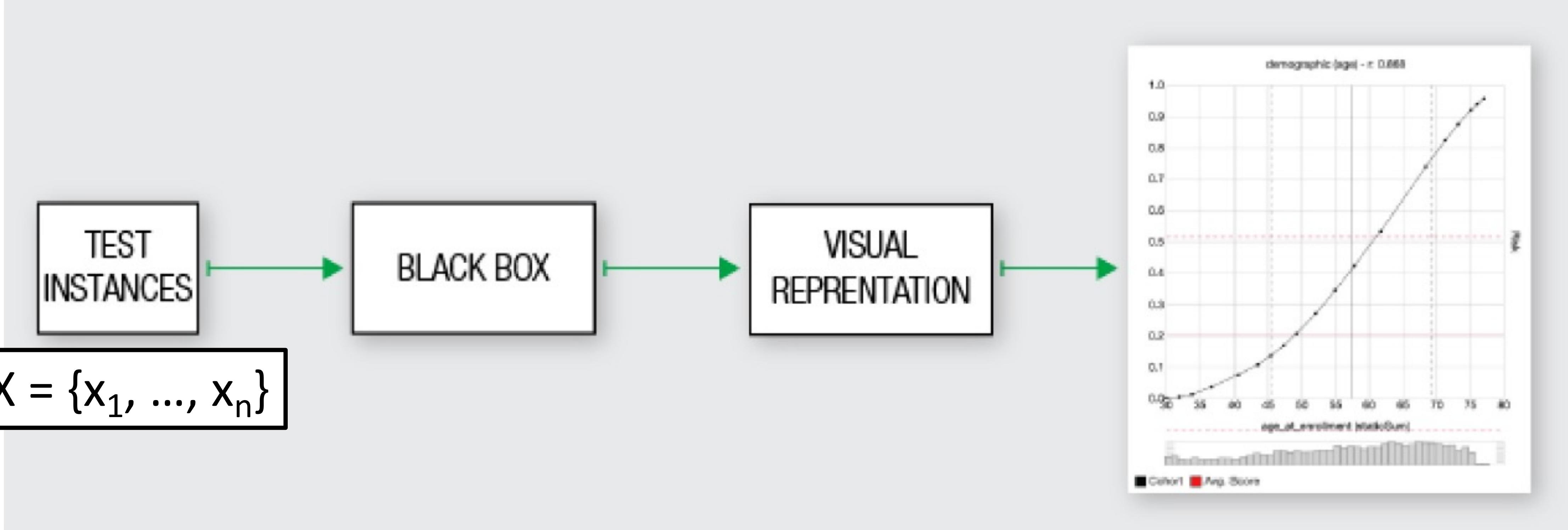
Outcome Explanation Problem

Provide an interpretable outcome, i.e., an ***explanation*** for the outcome of the black box for a ***single instance***.



Model Inspection Problem

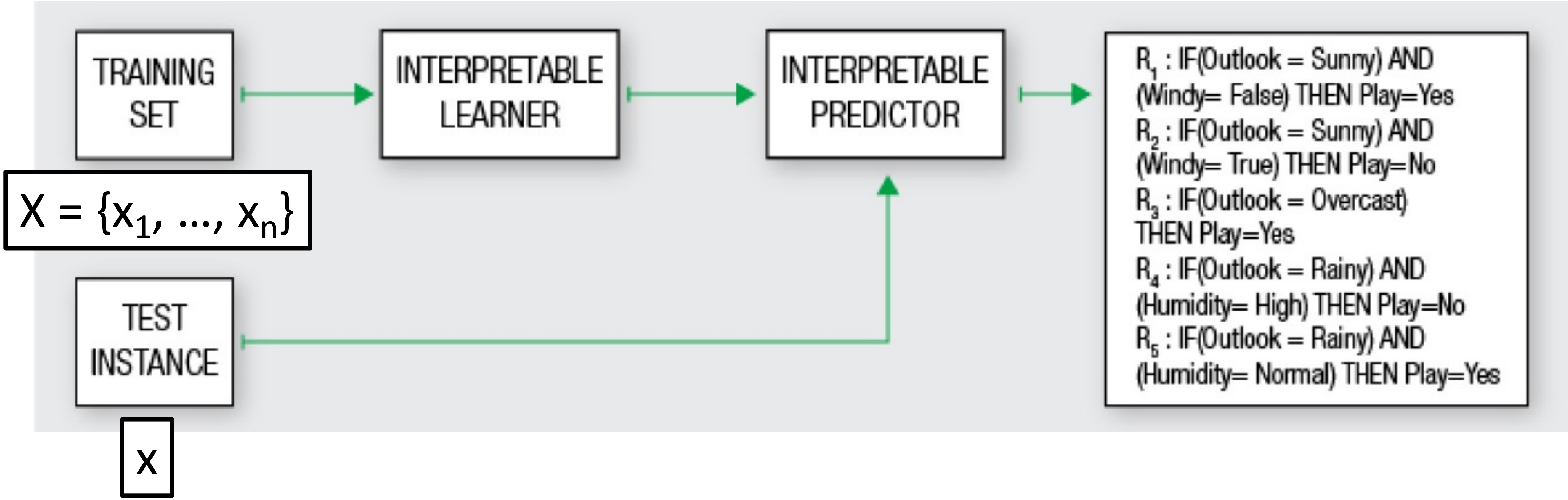
Provide a representation (visual or textual) for understanding either how the black box model works or why the black box returns certain predictions more likely than others.



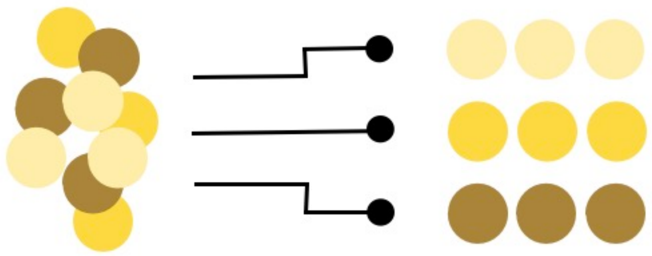
$X = \{x_1, \ldots, x_n\}$

Transparent Box Design Problem

Provide a model which is locally or globally interpretable on its own.



Categorization



- The type of ***problem***
- The type of ***black box model*** that the explanator is able to open
- The type of ***data*** used as input by the black box model
- The type of ***explanator*** adopted to open the black box

Black Boxes

- Neural Network (***NN***)
- Tree Ensemble (***TE***)
- Support Vector Machine (***SVM***)
- Deep Neural Network (***DNN***)

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)

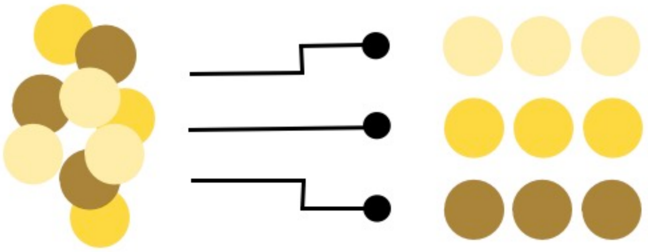
Images

(IMG)

Text

(TXT)

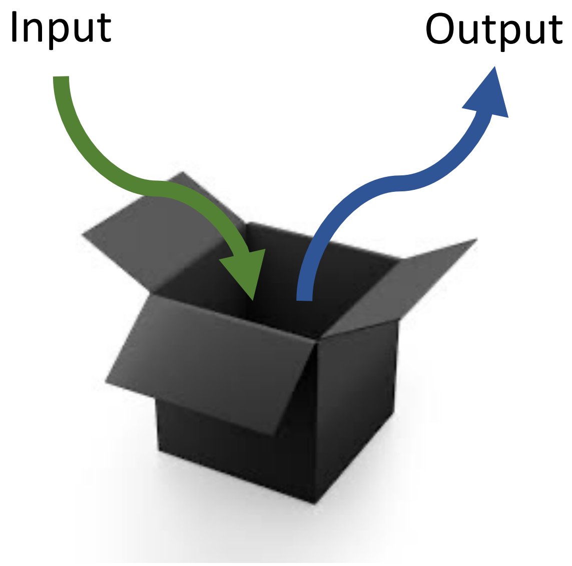
Explanators



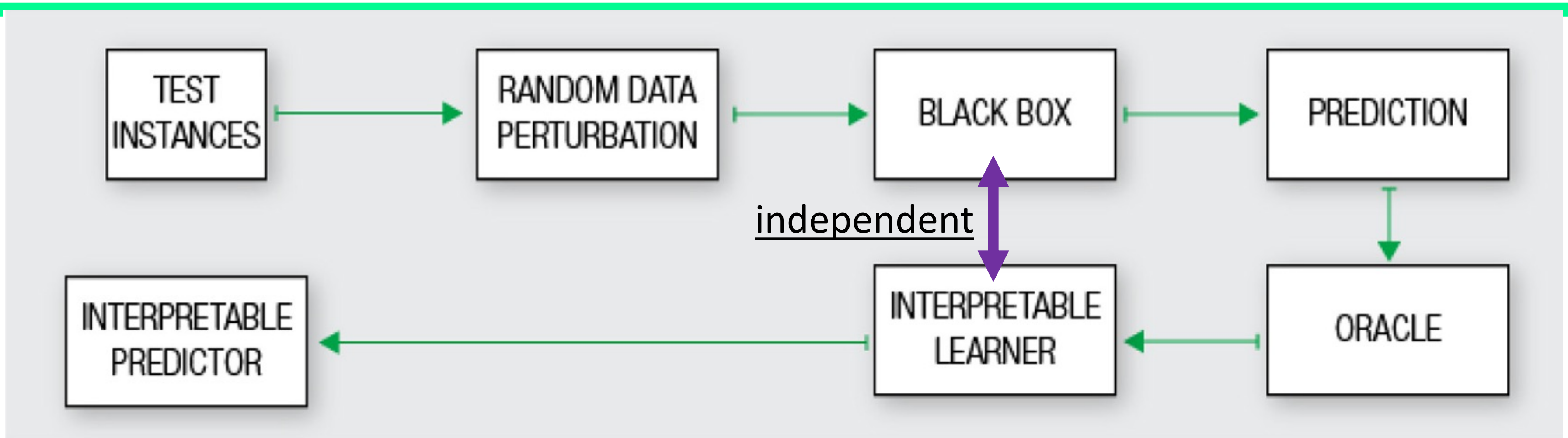
- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Maps (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)

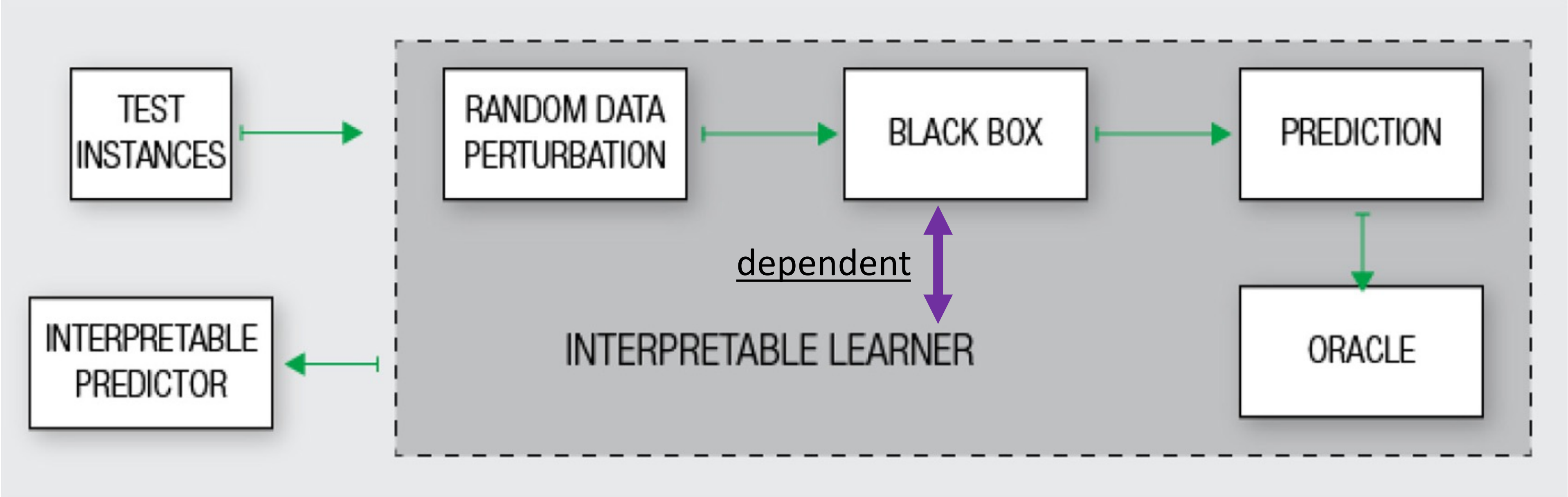
Reverse Engineering

- The name comes from the fact that we can only ***observe*** the ***input*** and ***output*** of the black box.
- Possible actions are:
  - ***choice*** of a particular comprehensible predictor
  - querying/auditing the black box with input records created in a controlled way using ***random perturbations*** w.r.t. a certain prior knowledge (e.g. train or test)
- It can be ***generalizable or not***:
  - Model-Agnostic
  - Model-Specific



Model-Agnostic vs Model-Specific





Solving The Model Explanation Problem

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
Trepan	[22]	Craven et al.	1996	DT	NN	TAB	✓				✓
–	[57]	Krishnan et al.	1999	DT	NN	TAB	✓		✓		✓
DecText	[12]	Boz	2002	DT	NN	TAB	✓	✓			✓
GPDT	[46]	Johansson et al.	2009	DT	NN	TAB	✓	✓	✓		✓
Tree Metrics	[17]	Chipman et al.	1998	DT	TE	TAB					✓
CCM	[26]	Domingos et al.	1998	DT	TE	TAB	✓	✓			✓
–	[34]	Gibbons et al.	2013	DT	TE	TAB	✓	✓			
STA	[140]	Zhou et al.	2016	DT	TE	TAB		✓			
CDT	[104]	Schetinin et al.	2007	DT	TE	TAB			✓		
–	[38]	Hara et al.	2016	DT	TE	TAB		✓	✓		✓
TSP	[117]	Tan et al.	2016	DT	TE	TAB	[illegible]	[illegible]	[illegible]	[illegible]	✓
Conj Rules	[21]	Craven et al.	[illegible]	DR	NN	TAB	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
G-REX	[44]	Johansson et al.	2003	DR	NN	TAB	✓	✓	✓	[illegible]	[illegible]
REFNE	[141]	Zhou et al.	2003	DR	NN	TAB	✓	✓	✓		✓
RxREN	[6]	Augasta et al.	2012	DR	NN	TAB		✓	✓		✓

Transparent methods

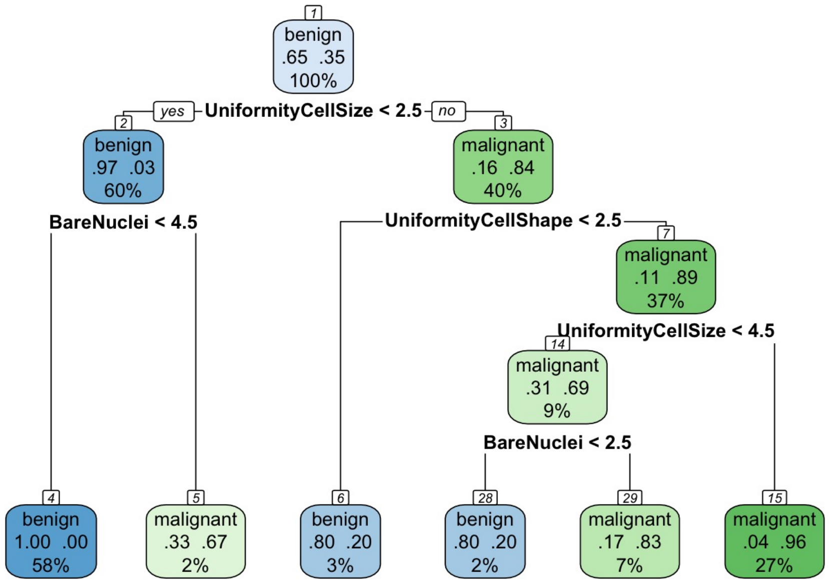
The explanation is ***embedded*** into the design of the AI system.

Most popular transparent methods:

- Decision tree (rules)
- Regressors (feature importance)

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

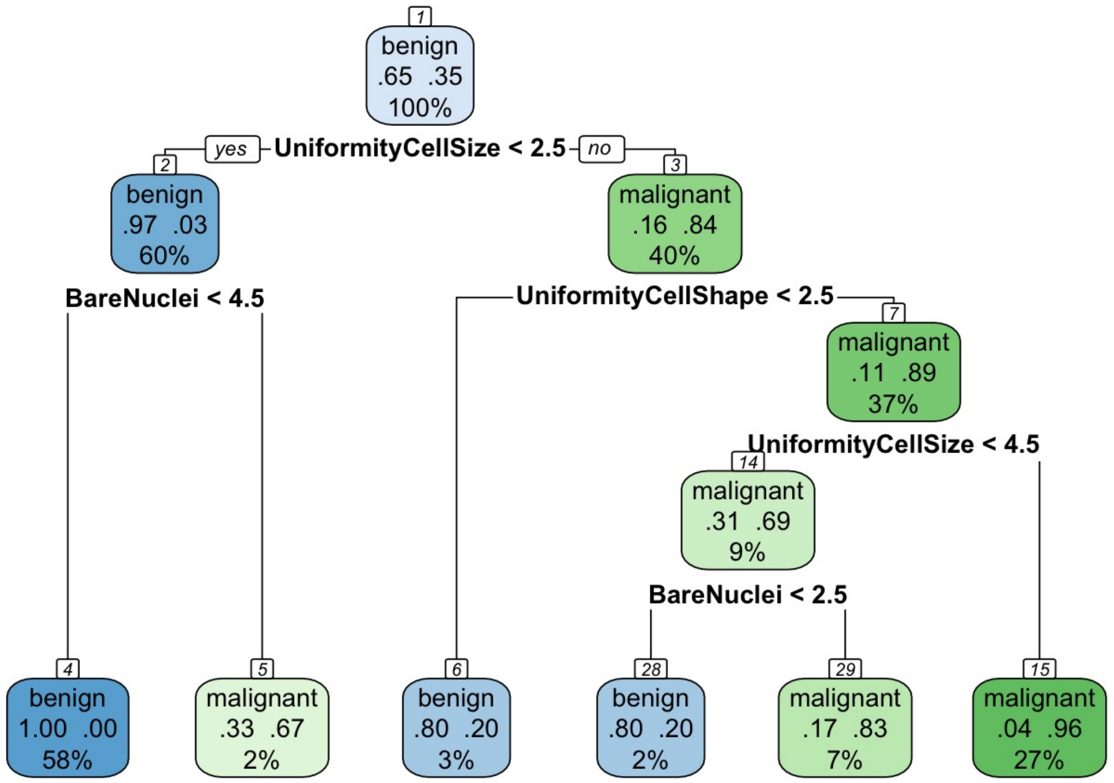
Φ = {({income > 900} -> grant),
({17 ≤ age < 25, job = other} -> grant)}



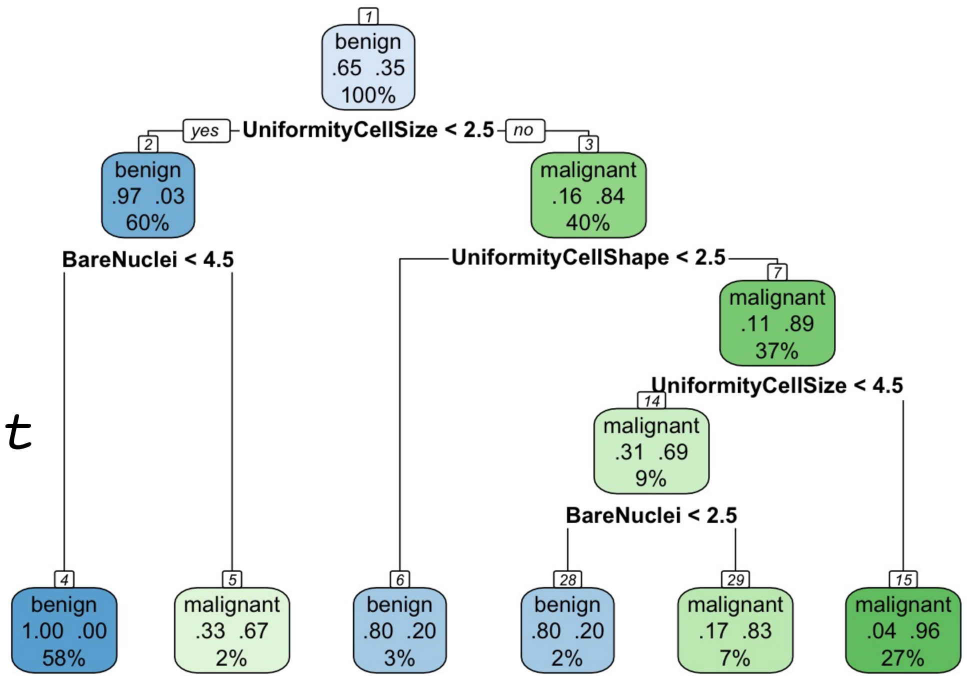
IF `SEX = female`
AND `Class = first`
THEN PREDICT `Survived = true`
WITH PRECISION 97%
AND COVERAGE 15%

Global Explainer: TREPAN

- Global explainer designed **to explain NN** but usable for any type of black box.
- It aims at **approximating a NN with a DT** classifier using best-m-of-n rules.
- At each node split the feature to split is selected on the original data extended with random samples respecting the current path.
- **It learns to predict the label returned by the black box**, not the original one.



Trepan – DT, NN, TAB



black box auditing

```
T = root_of_the_tree()
Q = <T, X, {}>
while Q not empty & size(T) < limit
      N, X_N, C_N  = pop(Q)
      Z_N = random(X_N, C_N)
      y_Z = b(Z), y = b(X_N)
      if same_class(y U y_Z)
            continue
      S = best_split(X_N U Z_N, y U y_Z)
      S'= best_m-of-n_split(S)
      N = update_with_split(N, S')
      for each condition c in S'
            C = new_child_of(N)
            C_C = C_N U {c}
            X_C = select_with_constraints(X_N, C_N)
            put(Q, <C, X_C, C_C>)
```

- Mark Craven and JudeW. Shavlik. 1996. ***Extracting tree-structured representations of trained networks***. NIPS.

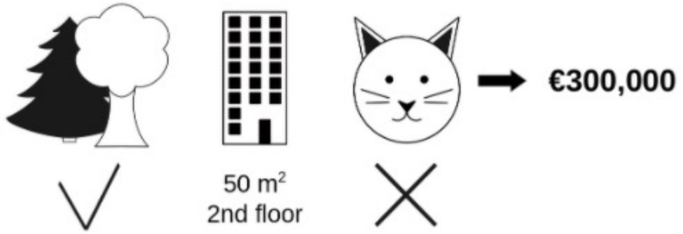
Solving The Outcome Explanation Problem

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
–	[134]	Xu et al.	2015	SM	DNN	IMG			✓	✓	✓
–	[30]	Fong et al.	2017	SM	DNN	IMG			✓		
CAM	[139]	Zhou et al.	2016	SM	DNN	IMG			✓	✓	✓
Grad-CAM	[106]	Selvaraju et al.	2016	SM	DNN	IMG			✓	✓	✓
–	[109]	Simonian et al.	2013	SM	DNN	IMG			✓		✓
PWD	[7]	Bach et al.	2015	SM	DNN	IMG			✓		✓
–	[113]	Sturm et al.	2016	SM	DNN	IMG			✓		✓
DTD	[78]	Montavon et al.	2017	SM	DNN	IMG			✓		✓
DeapLIFT	[107]	Shrikumar et al.	2017	FI	DNN	ANY			✓	✓	
CP	[64]	Landecker et al.	2013	SM	NN	IMG			✓		
–	[143]	Zintgraf et al.	2017	SM	DNN	IMG			✓	✓	✓
VBP	[11]	Bojarski et al.	2016	SM	DNN	IMG			✓		✓
–	[65]	Lei et al.	2016	SM	DNN	TXT			✓		✓
ExplainD	[89]	Poulin et al.	2006	FI	SVM	TAB		✓	✓		
–	[29]	Strumbelj et al.	2010	FI	AGN	TAB	✓	✓	✓		✓

SHAP

A prediction can be explained by assuming that **each feature value of the instance is a "player"** in a game where the **prediction is the payout**. Shapley values tells us how to fairly distribute the "payout" among the features.

Example



Prediction: You have trained a machine learning model to predict apartment prices. For a certain apartment it predicts €300,000 and you need to explain this prediction.

The apartment has an area of 50 m2, is located on the 2nd floor, has a park nearby and cats are banned.

The average prediction is €310,000.

How much has each feature value contributed to the prediction compared to the average prediction?

SHAP

The average prediction is **€310,000** while the prediction is **€300,000**

How much has each feature value contributed to the prediction compared to the average prediction?

The answer is simple for linear regression models. **The effect of each feature is the weight of the feature times the feature value.** This only works because of **the linearity of the model**.

For more complex models, we need a different solution!!!!

GOAL: explain the difference between the actual prediction (€300,000) and the average prediction (€310,000): a difference of -€10,000.

SHAP

GOAL: explain the difference between the actual prediction (€300,000) and the average prediction (€310,000): a difference of -€10,000.

Game theory:

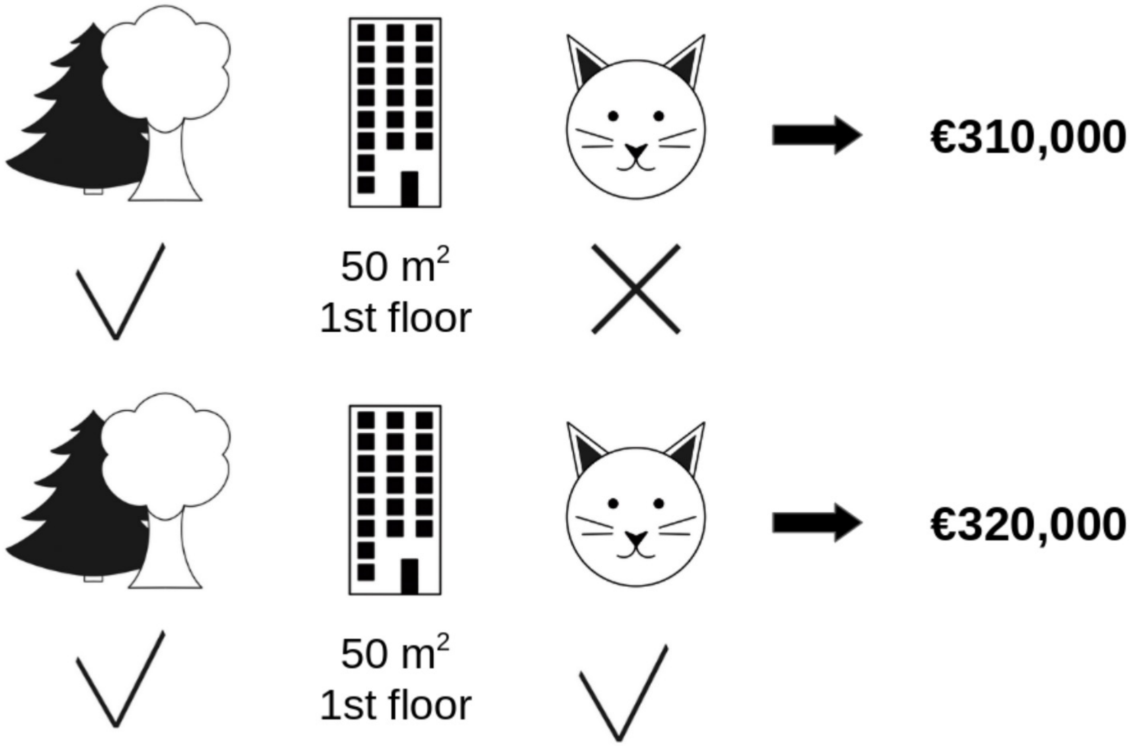
- The "game" is the prediction task for a single instance of the dataset.
- The "gain" is the **actual prediction** for this instance **minus** the **average prediction for all instances.**

The "players" are the feature values of the instance that collaborate to receive the gain (= predict a certain value).

The Shapley value is the average marginal contribution of a feature value **across all possible coalitions.**

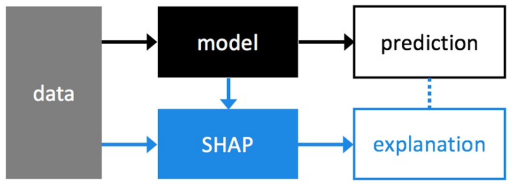
Shapely Values

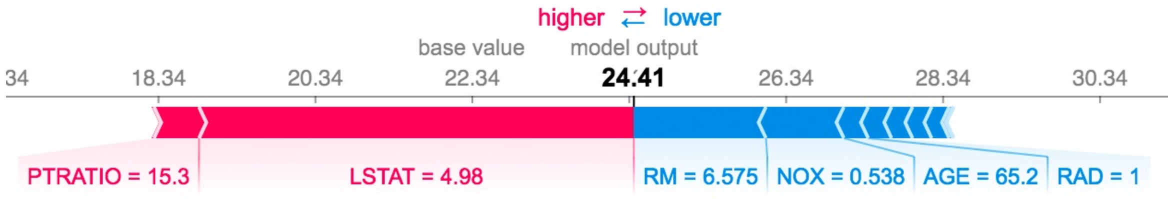
One sample repetition to estimate the contribution of cat-banned to the prediction when added to the coalition of *park-nearby* and *area-50.*

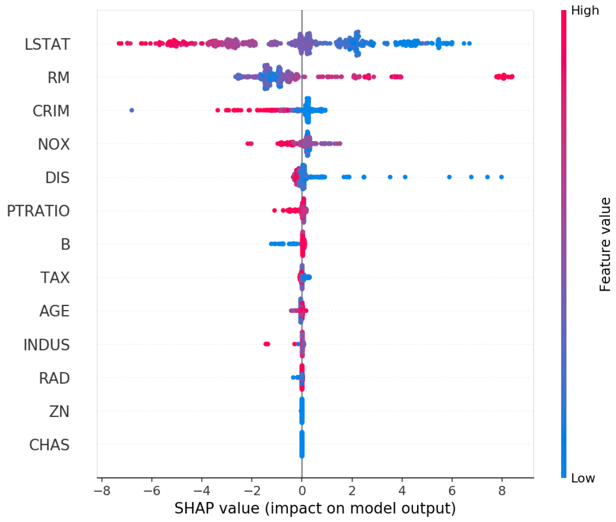


SHAP

- SHAP (SHapley Additive exPlanations) assigns each feature an importance value for a particular prediction by means of an additive feature attribution method.
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature

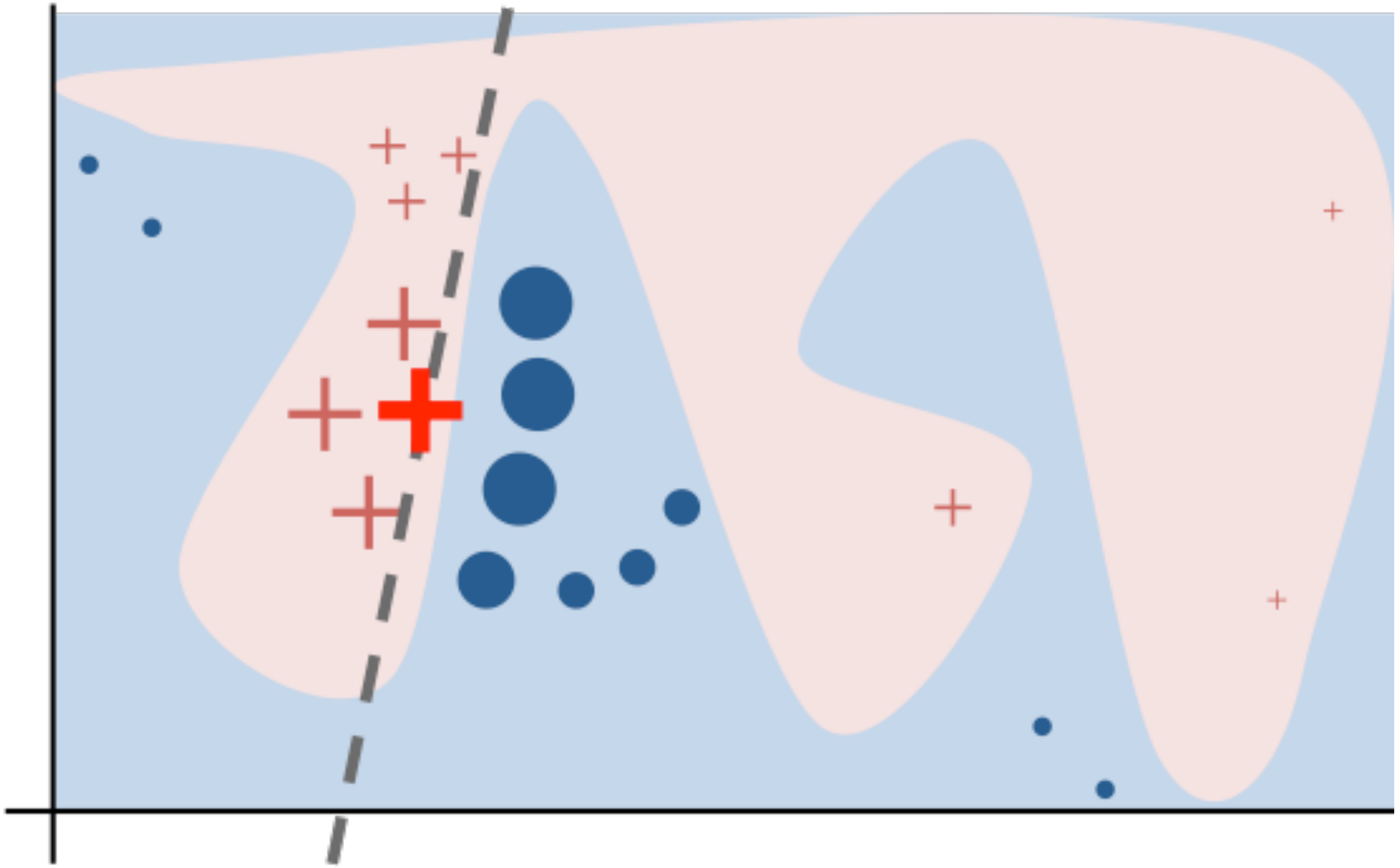






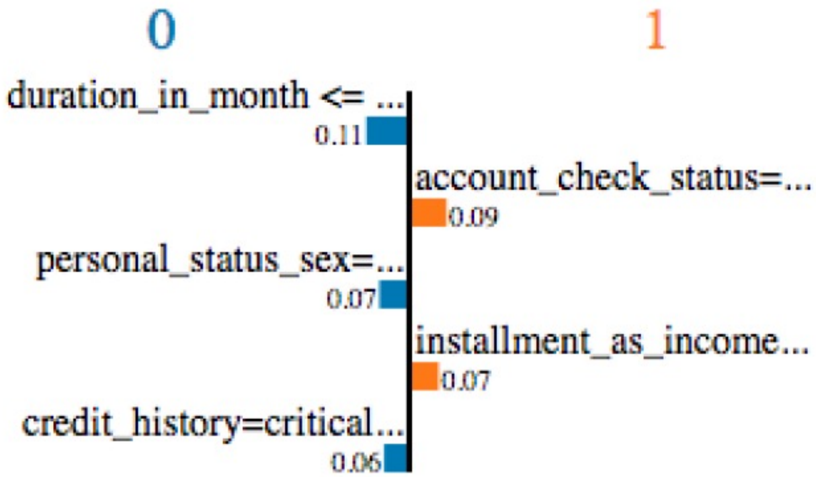
Local Explanation

- 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.



LIME

```
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
```



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

LIME

- 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.

LIME – tab data

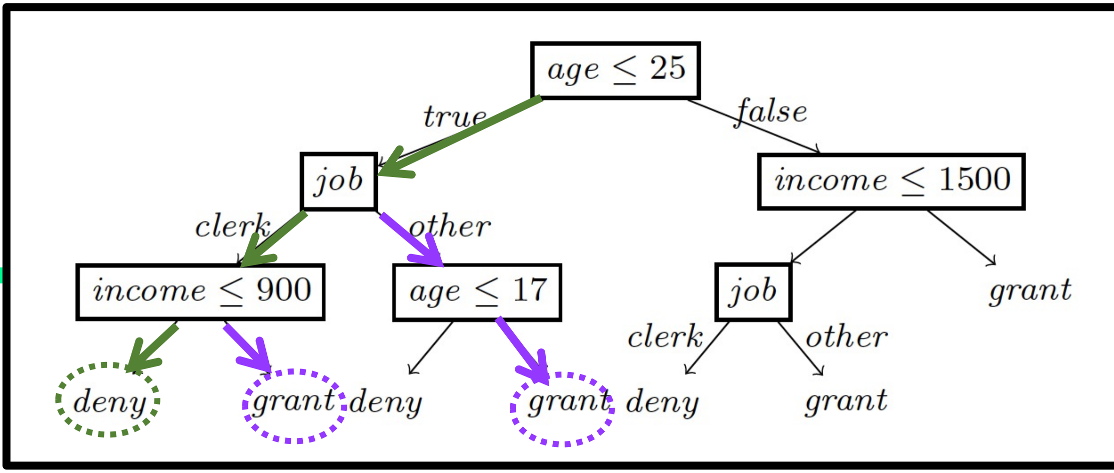
- LIME does not really generate images with different information: it randomly removes some superpixels, i.e. it suppresses the presence of an information rather than modifying it.

- 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 – DR, AGN, TAB



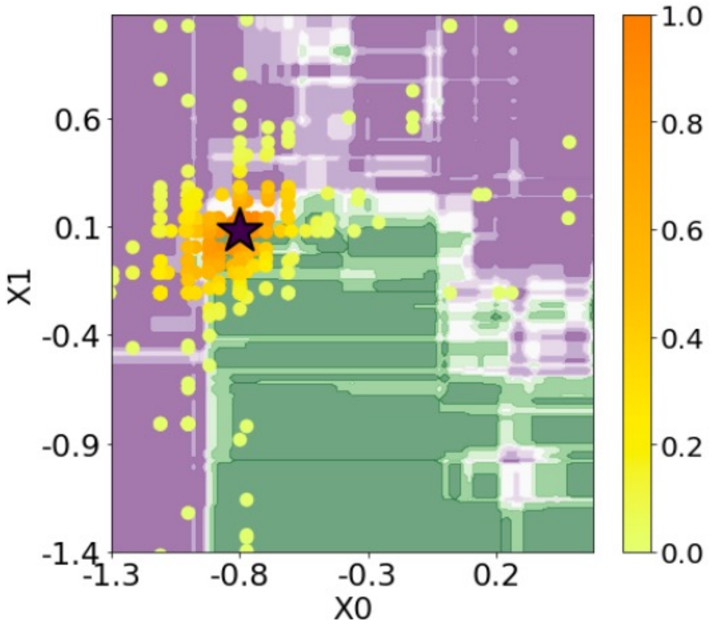

```
x instance to explain
Z_= = geneticNeighborhood(x, fitness_=, N/2)
Z_≠ = geneticNeighborhood(x, fitness_≠, N/2)
Z = Z_= U Z_≠
c = buildTree(Z, b(Z))
r = (p -> y) = extractRule(c, x)
ϕ = extractCounterfactual(c, r, x)
return e = <r, ϕ>
```

black box auditing

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

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

Pedreschi, Franco Turini,
f black box decision

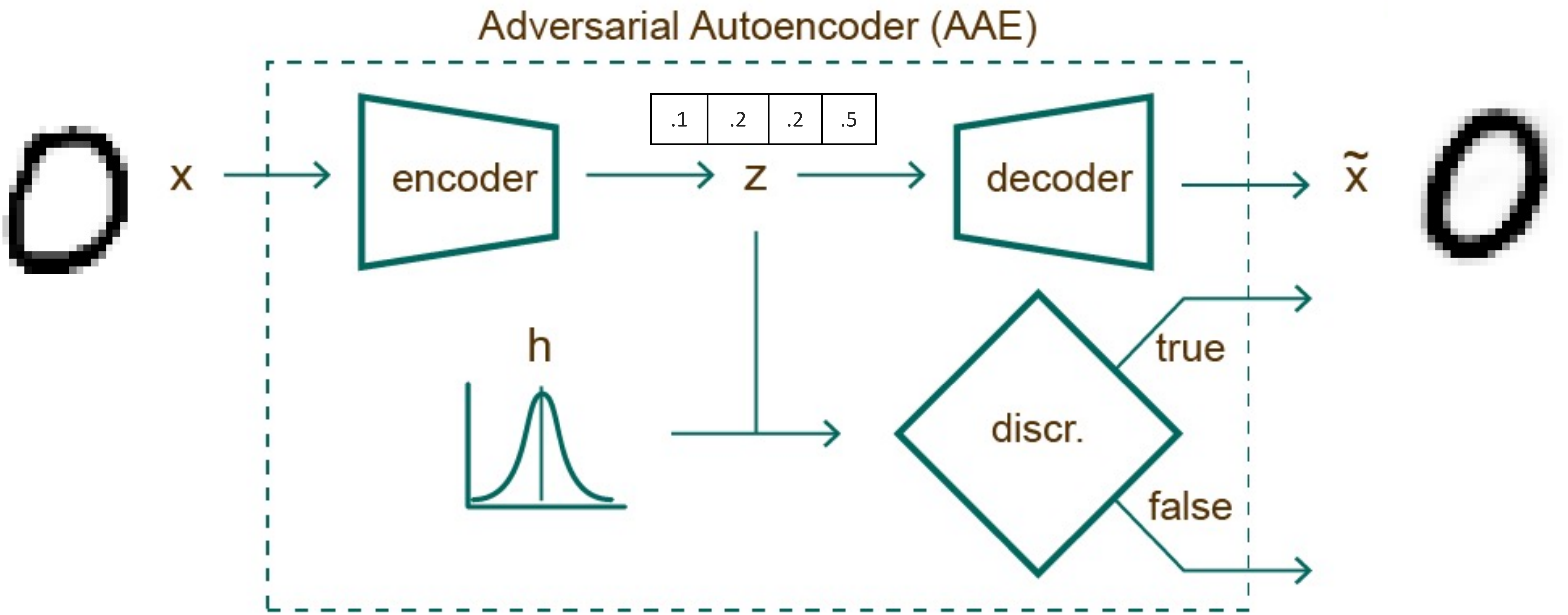


Adversarial Black box Explainer generating Latent Exemplars

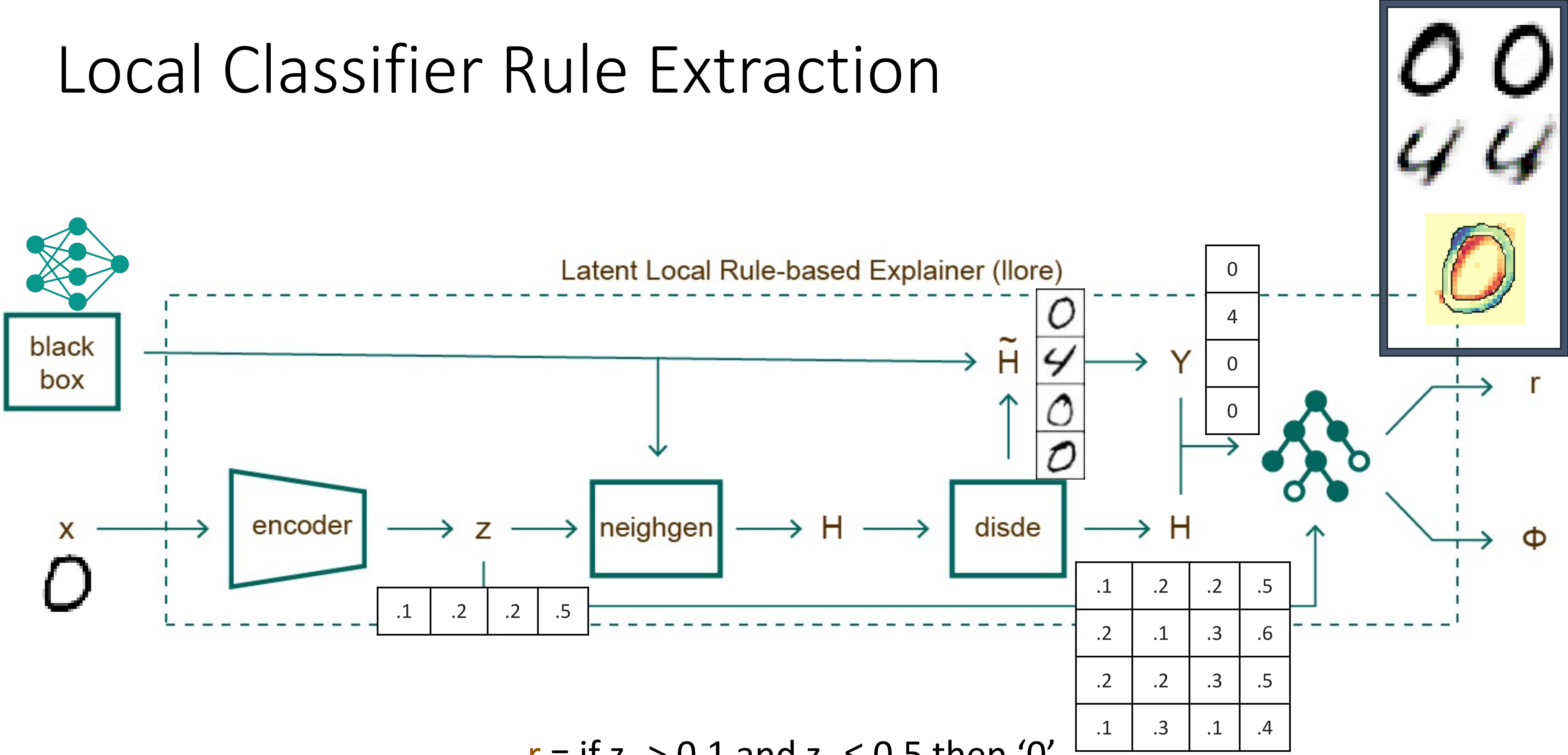
- Explaining image classification
- Solving the drawback of LIME
- Exploit adversarial autoencoders
- Providing explanations based on examplars and counter examplars

Explaining Image classifiers

Background - Adversarial Autoencoder



Local Classifier Rule Extraction



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

φ = {if $z_1 \leq 0.1$ then '4',

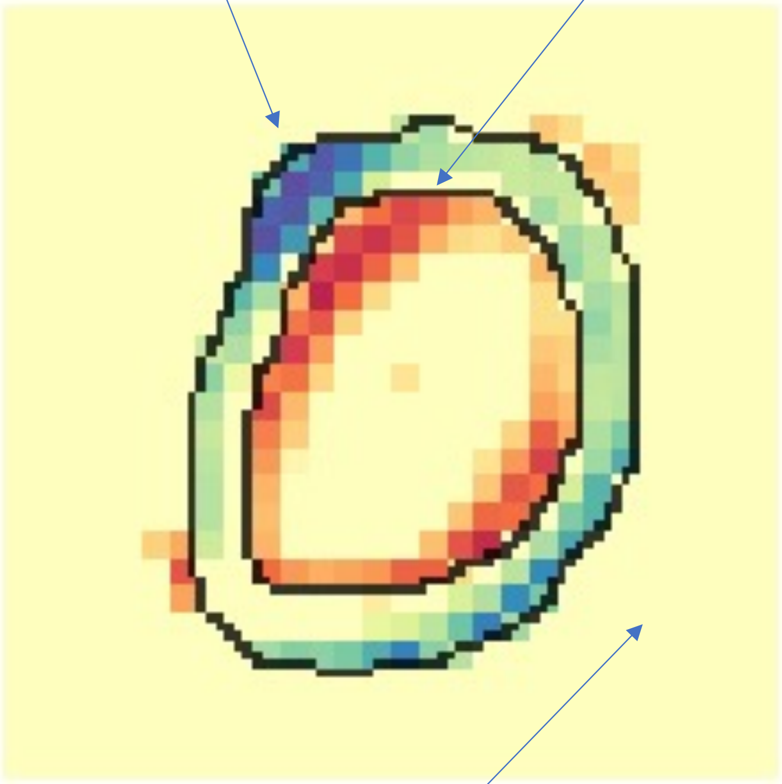
if $z_3 > 0.5$ then '8'}

- R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. Local rule-based explanations of black box decision systems. arXiv:1805.10820, 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 ***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”

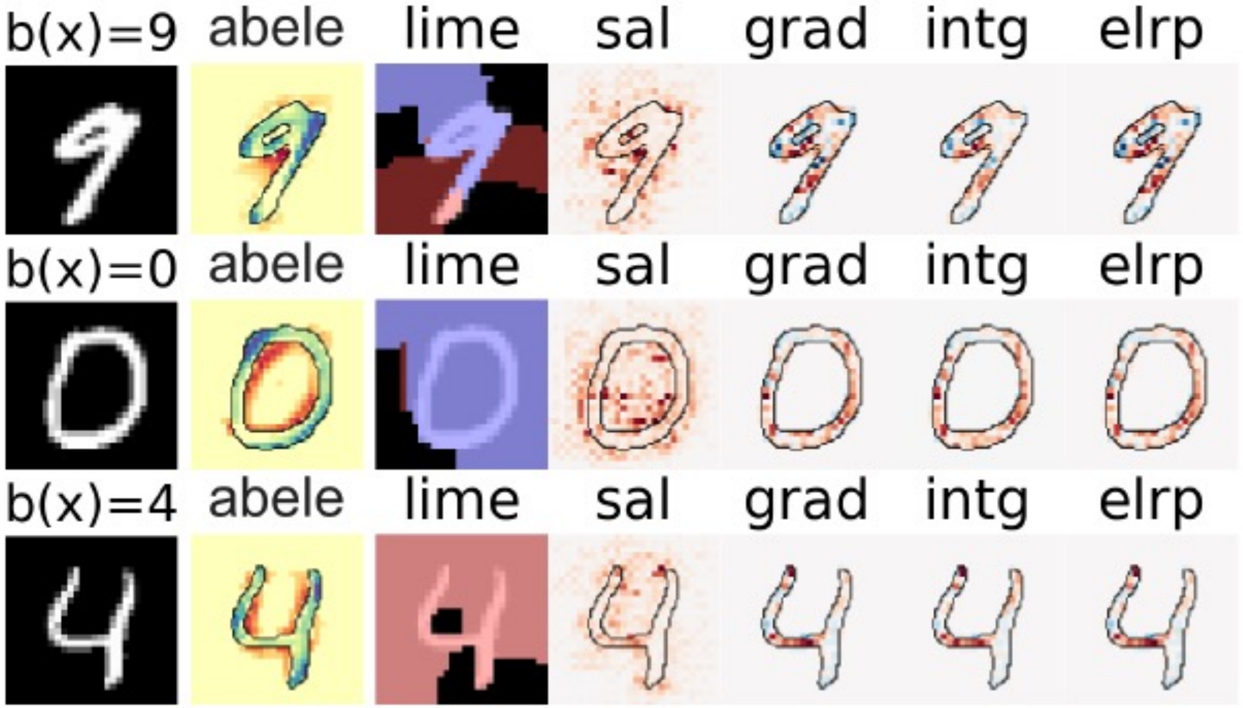
ABELE vs LIME Neighborhood

- ABELE

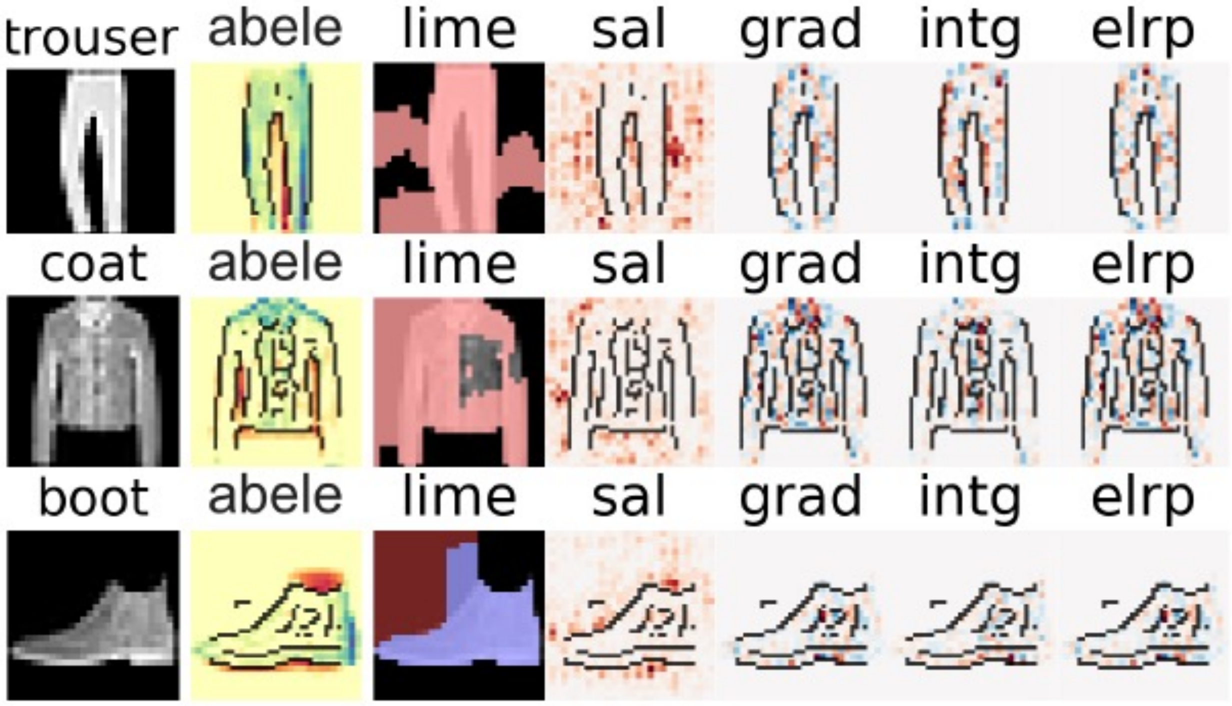
- LIME

Saliency Map Comparison

- **mnist**

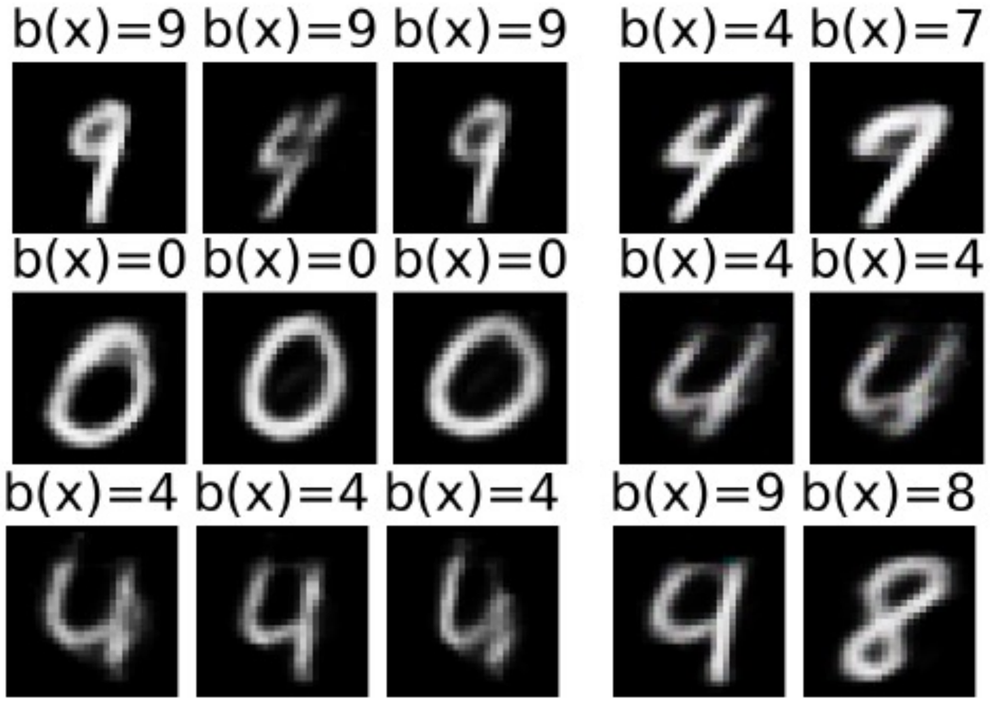


- **fashion**

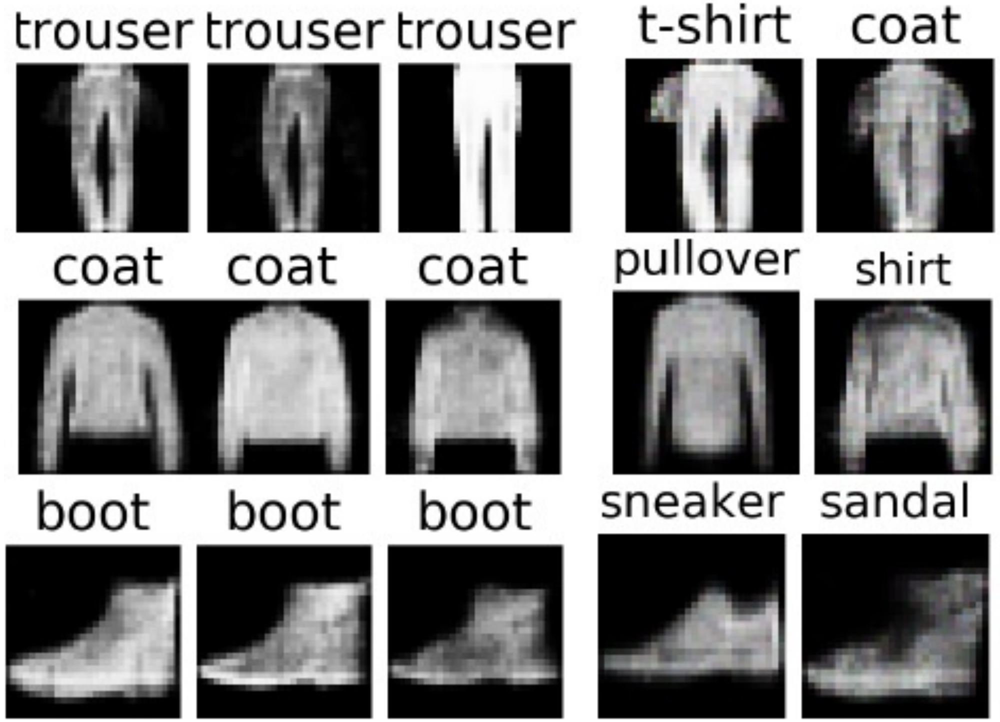


Exemplars and Counter-Exemplars

- **mnist**

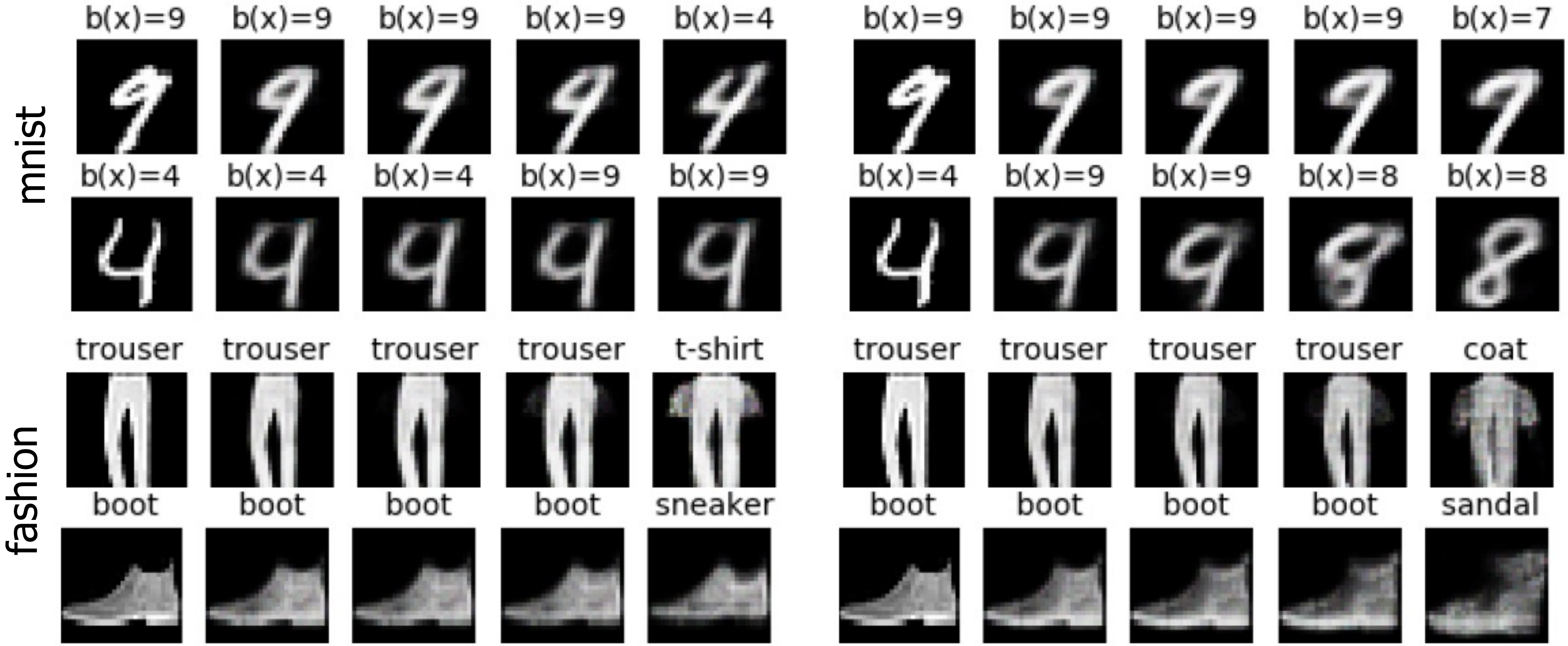


- **fashion**



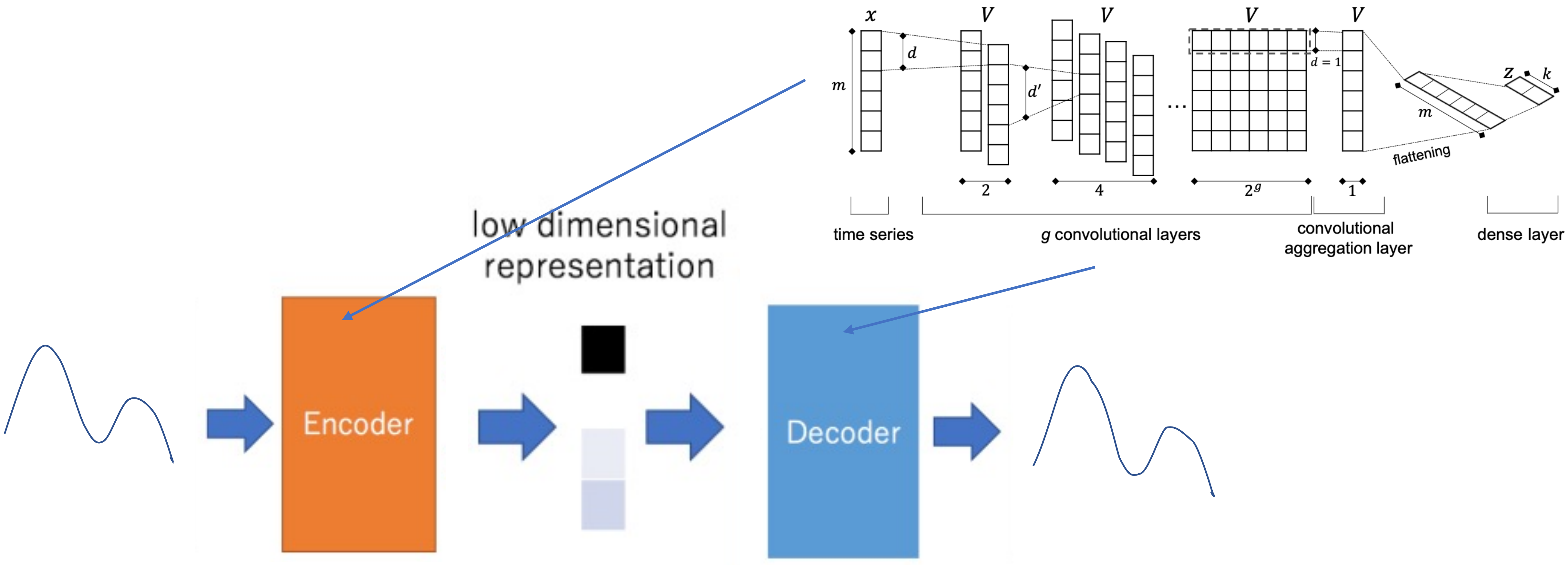
From Image to Counter-Exemplar

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

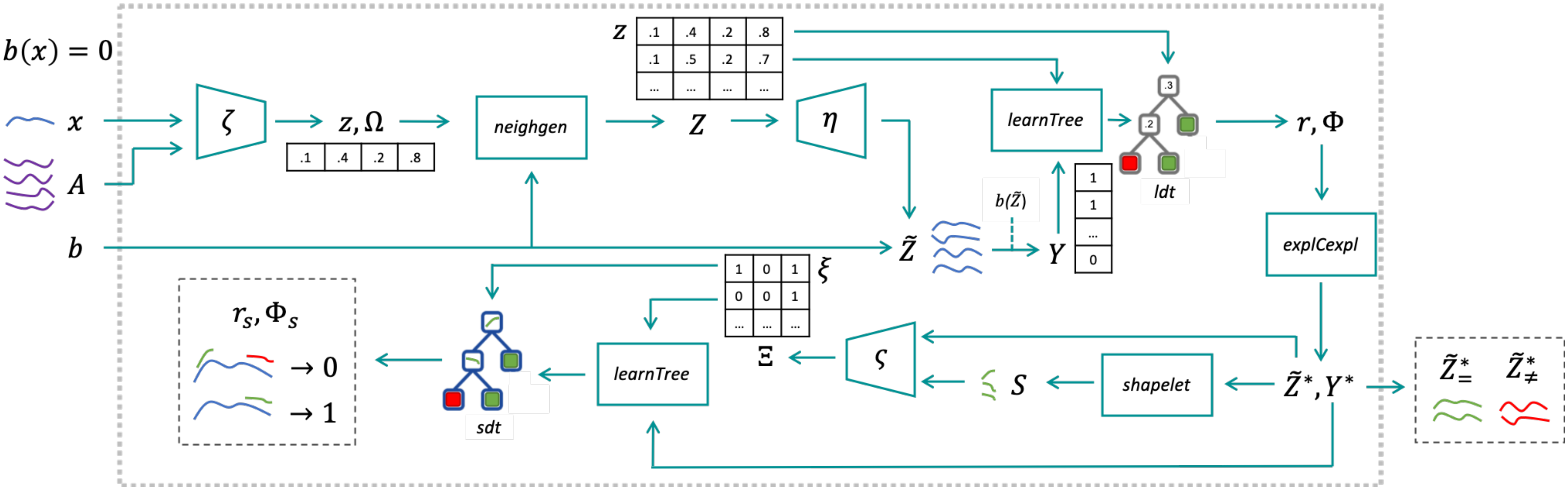


Explaining time series classifiers

Setting The Stage - Autoencoder

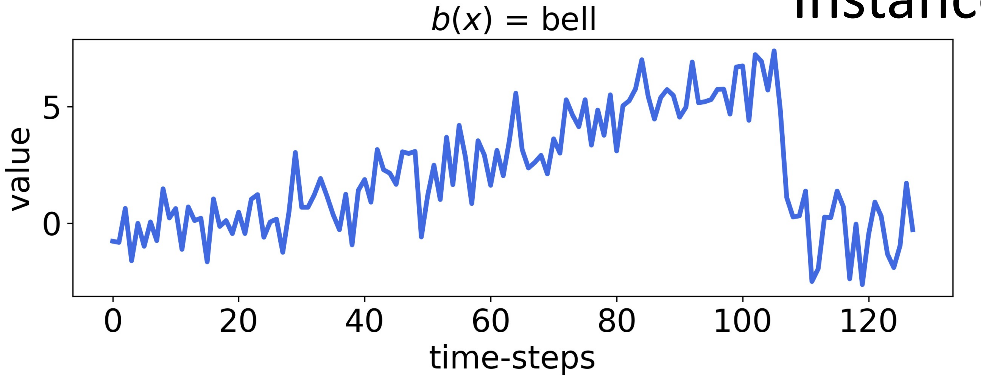


LASTS: Local Agnostic Subsequence-based Time Series explainer

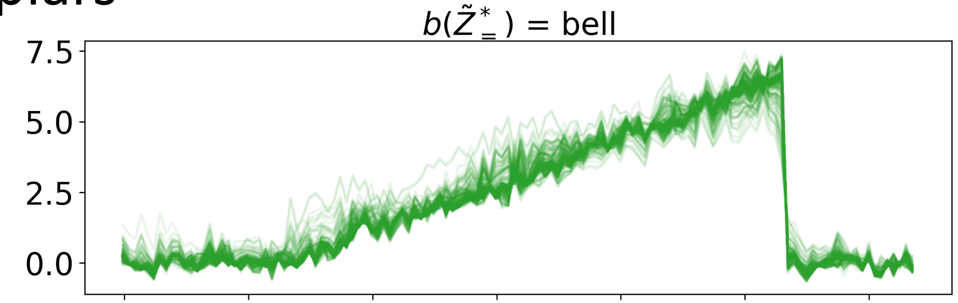


LASTS Explanation

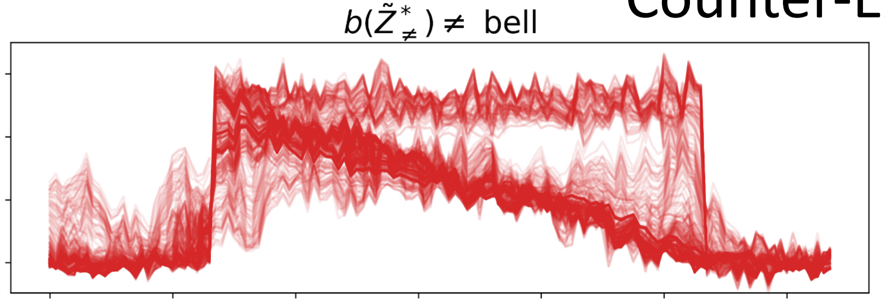
Instance to explain



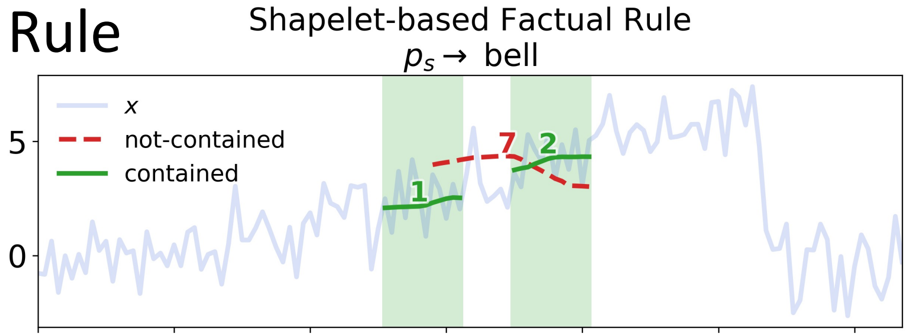
Exemplars



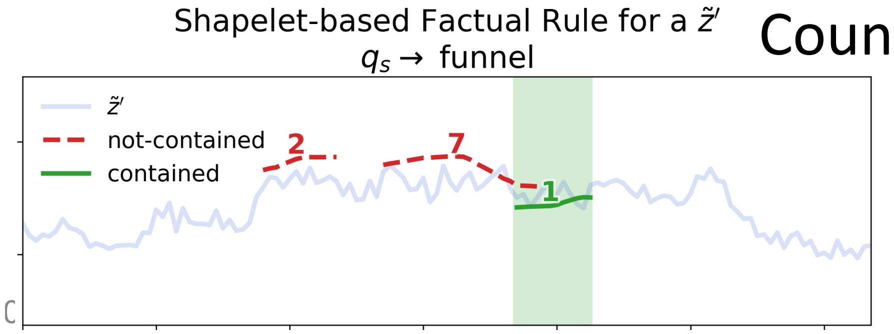
Counter-Exemplars



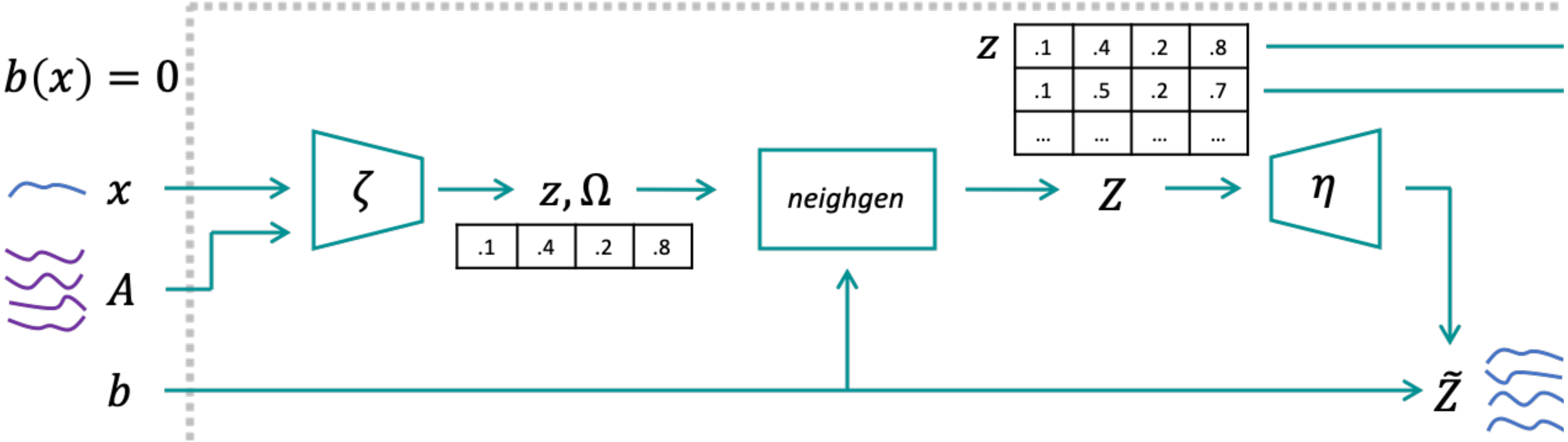
Factual Rule



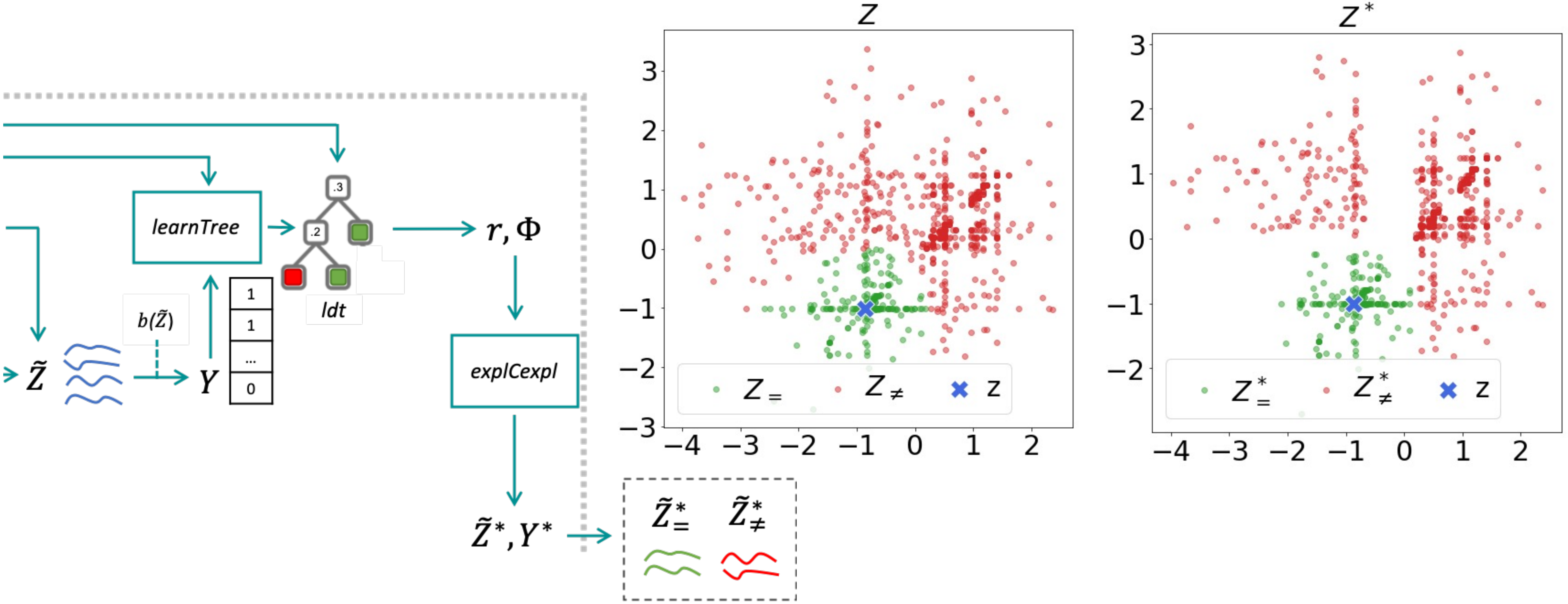
Counter-Factual Rule



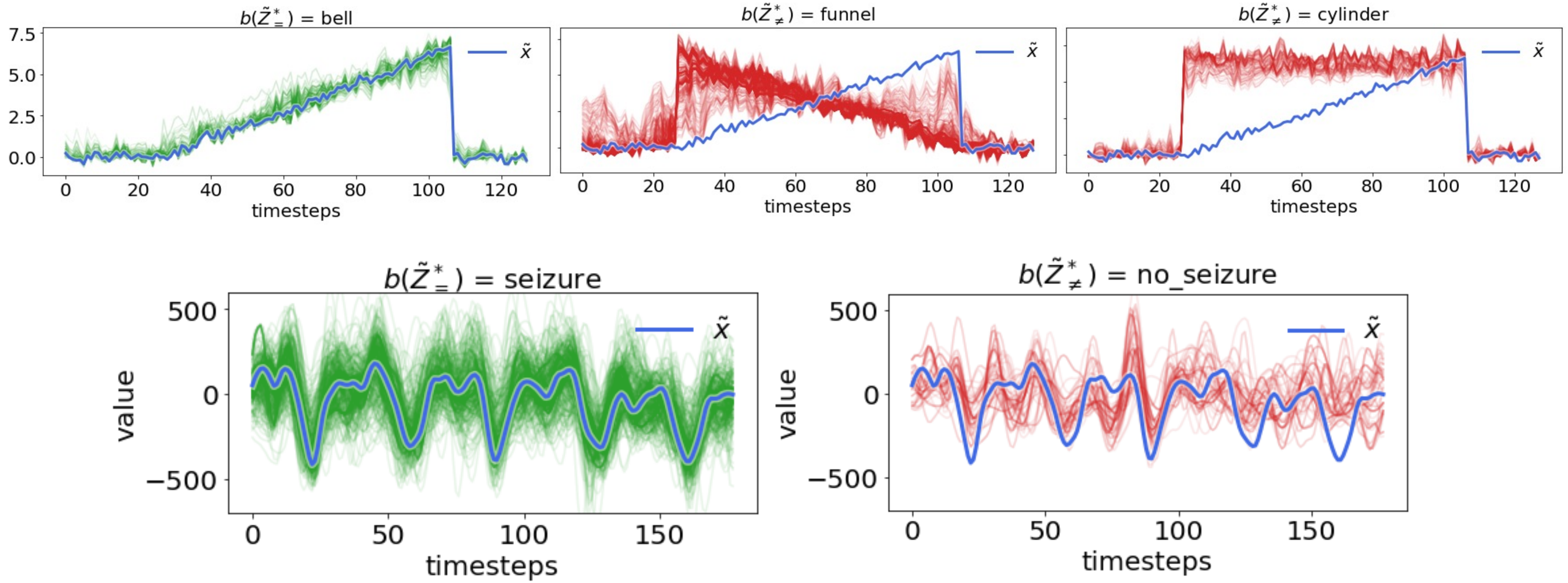
Latent Encoding and Neighborhood Generation



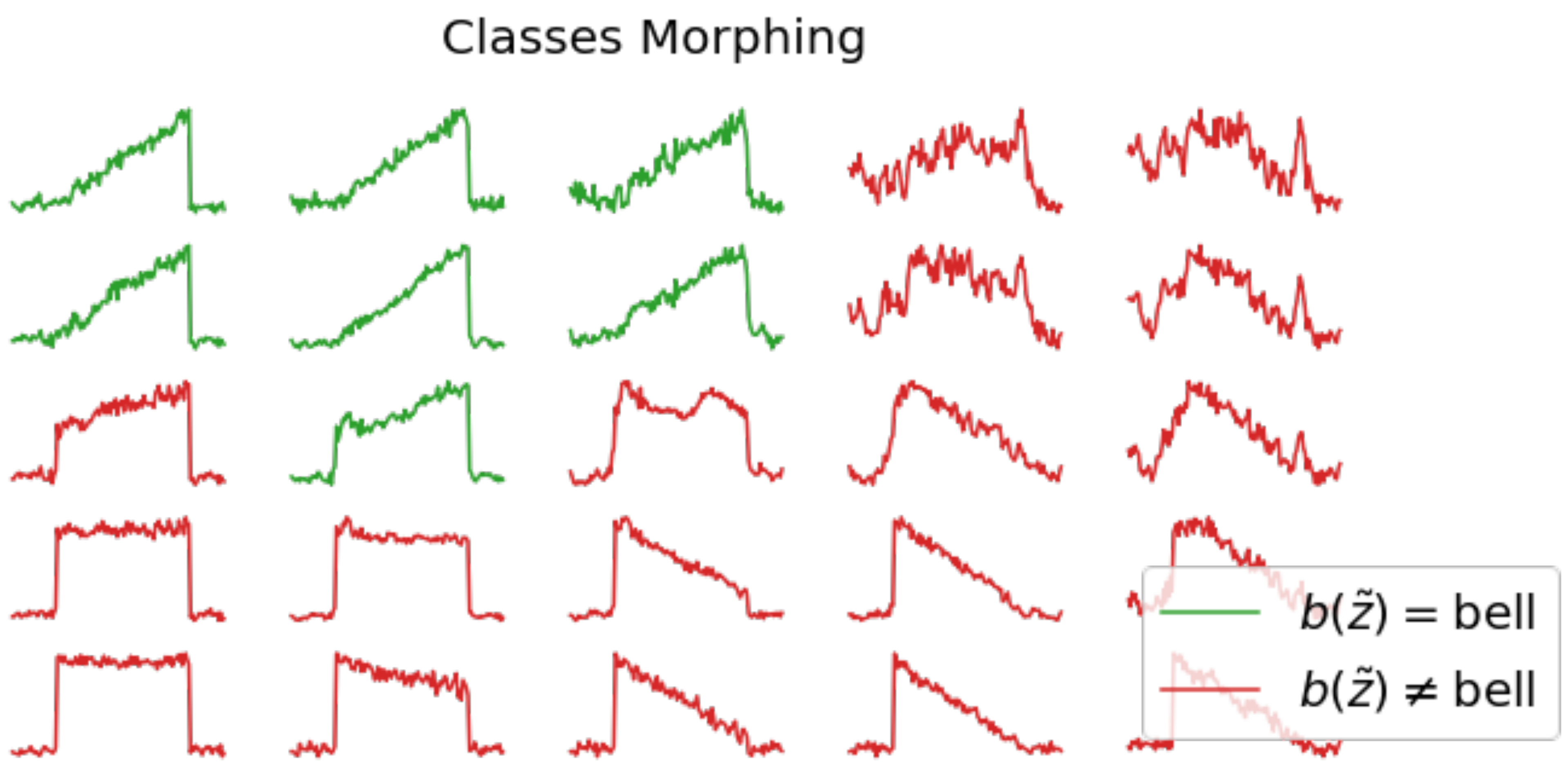
Local Latent Rules and (Counter-)Exemplars Selection



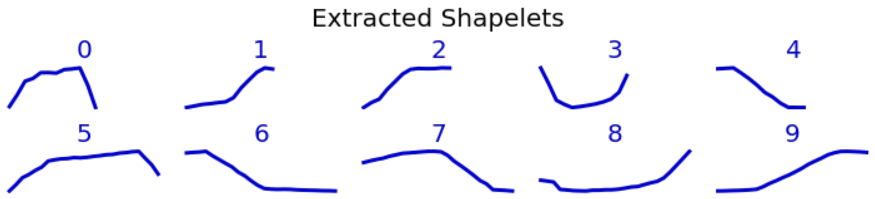
Exemplars and Counter Exemplars

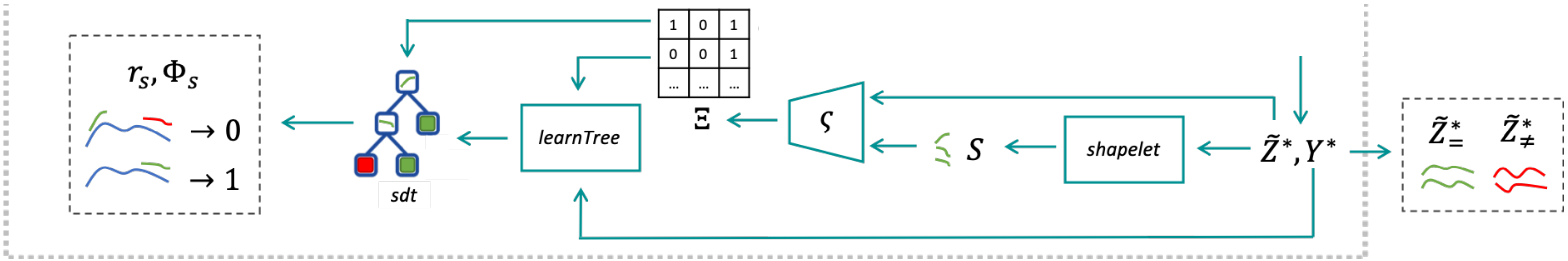


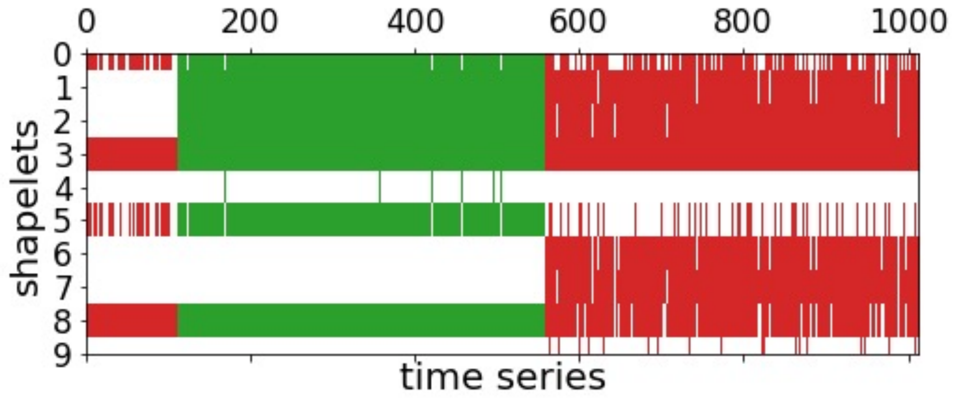
From Exemplars to Counter-Exemplars



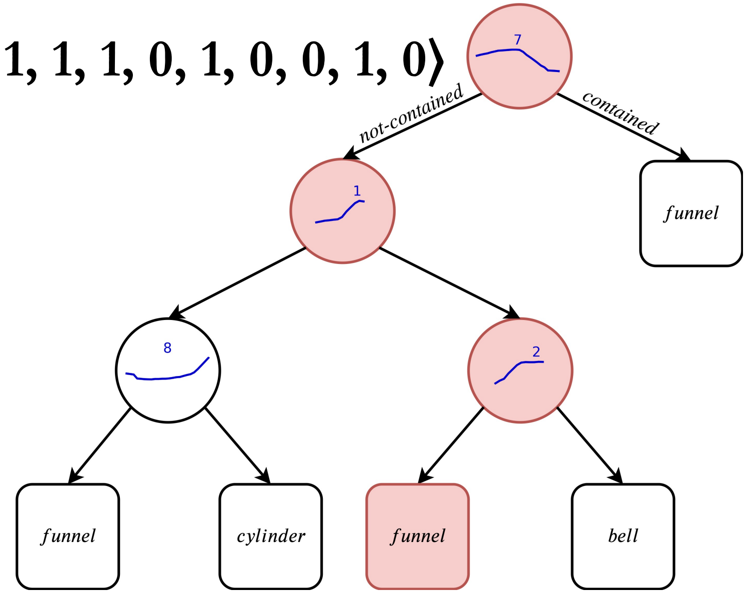
Shapelet-Based Rule Extraction







⟨1, 1, 1, 1, 0, 1, 0, 0, 1, 0⟩



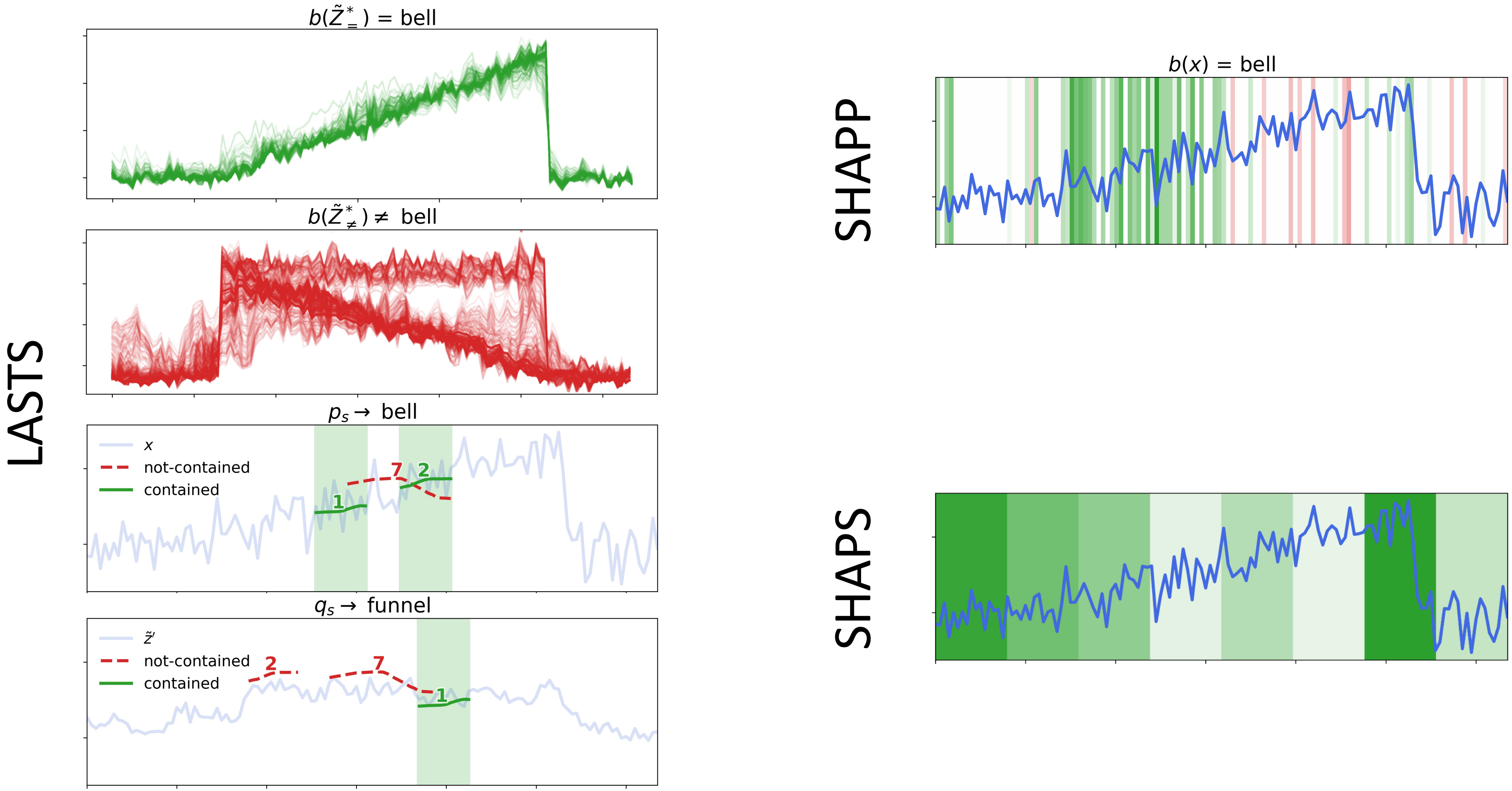
$\mathbf{r_s}$ IF S7 is NOT contained AND S1 AND s2 are contained

bell

$\mathbf{\Phi_s}$ IF S2 AND S7 are NOT contained AND S1 is contained

funnel

Comparing Time Series Explanations



Take Home Message

Take-Home Messages

- Explainable AI is motivated by real-world application of AI
- Not a new problem – a reformulation of past research challenges in AI
- Multi-disciplinary: multiple AI fields, HCI, social sciences (multiple definitions)
- In Machine Learning:
  - Transparent design or post-hoc explanation?
  - Background knowledge matters!
  - We can scale-up symbolic reasoning by coupling it with representation learning on graphs.
- In AI (in general): many interesting / complementary approaches

Open The Black Box!

- ***To empower*** individual against undesired effects of automated decision making
- ***To reveal*** and protect new vulnerabilities
- ***To implement*** the "right of explanation"
- ***To improve*** industrial standards for developing AI-powered products, increasing the trust of companies and consumers
- ***To help*** people make better decisions
- ***To align*** algorithms with human values
- ***To preserve*** (and expand) human autonomy

Open Research Questions

- There is ***no agreement*** on ***what an explanation is***
- There is ***not a formalism*** for ***explanations***
- There is ***no work*** that seriously addresses the problem of ***quantifying*** the grade of ***comprehensibility*** of an explanation for humans
- Is it possible to join ***local*** explanations to build a ***globally*** interpretable model?
- What happens when black box make decision in presence of ***latent features***?
- What if there is a ***cost*** for querying a black box?

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), 93
- Finale Doshi-Velez and Been Kim. 2017. ***Towards a rigorous science of interpretable machine learning***. arXiv:1702.08608v2
- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.
- Andrea Romei and Salvatore Ruggieri. 2014. ***A multidisciplinary survey on discrimination analysis***. Knowl. Eng.
- Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. 2015. ***A comprehensive review on privacy preserving data mining***. SpringerPlus
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ***Why should i trust you?: Explaining the predictions of any classifier***. KDD.
- Houtao Deng. 2014. ***Interpreting tree ensembles with intrees***. arXiv preprint arXiv:1408.5456.
- Mark Craven and JudeW. Shavlik. 1996. ***Extracting tree-structured representations of trained networks***. NIPS.

References

- M. Gethsiyal Augasta and T. Kathirvalavakumar. 2012. ***Reverse engineering the neural networks for rule extraction in classification problems***. NPL
- Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. ***Local rule-based explanations of black box decision systems***. arXiv preprint arXiv:1805.10820
- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).
- Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.

- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).

- Xiaoxin Yin and Jiawei Han. 2003. ***CPAR: Classification based on predictive association rules***. SIAM, 331–335
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. 2017. ***Learning certifiably optimal rule lists***. KDD.