Tutorial on

eXplainable Knowledge Discovery in Data Mining

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

UNIVERSITÀ DI PISA

Definitions

Oxford Dictionary of English

explanation | ɛkspləˈneɪʃ(ə)n |

noun

a statement or account that makes something clear: *the birth rate is central to any explanation of population trends*.

interpret | ɪnˈtəːprɪt |

verb **(interprets, interpreting, interpreted)** *[with object]*

1 explain the meaning of (information or actions): *the evidence is difficult to interpret*.

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"*).
- 5 core principles for ethical AI:
  - beneficence, non-maleficence, autonomy, and justice
  - a new principle is needed in addition: explicability

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

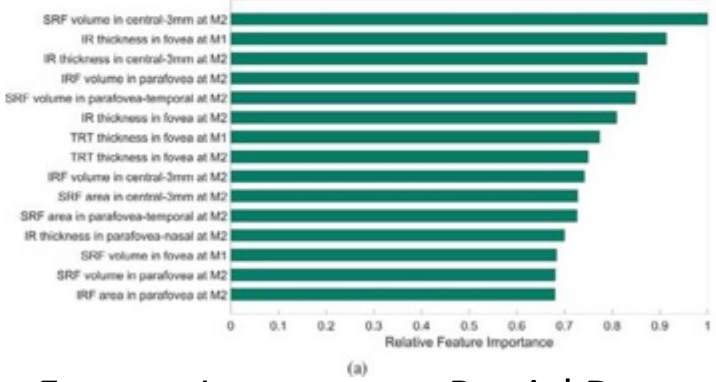
In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.

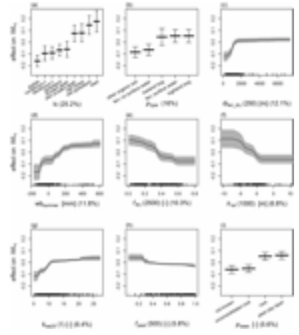
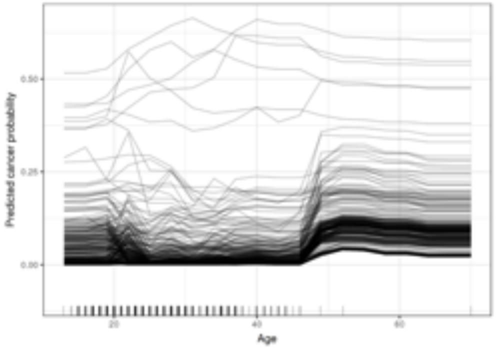
Right of Explanation

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.

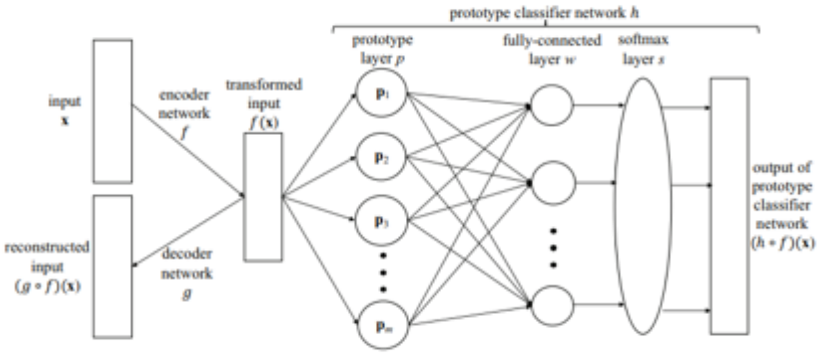
Explanation in different AI fields

- Machine Learning



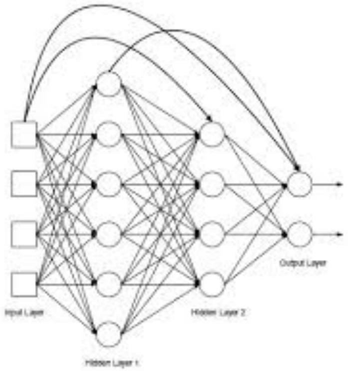
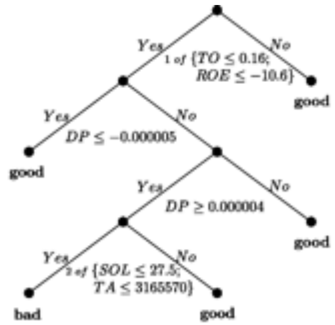



Feature Importance, Partial Dependence Plot, Individual Conditional Expectation



Auto-encoder

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537

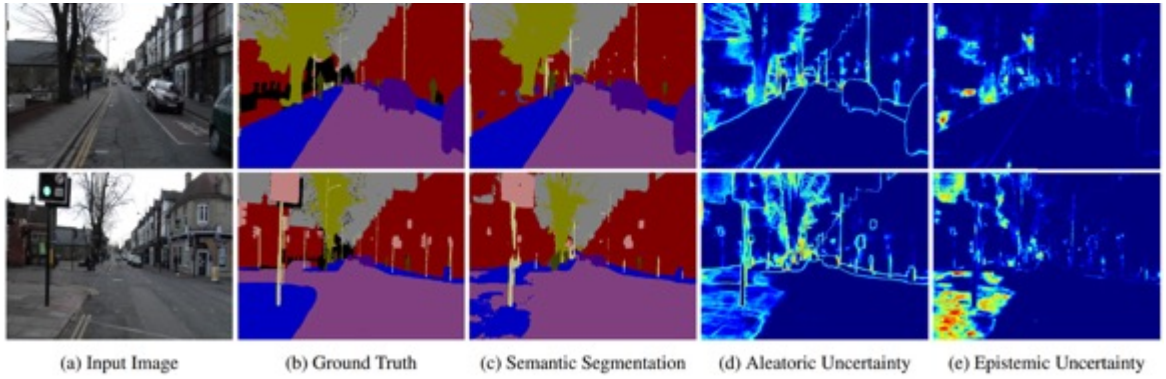



Surogate Model

Mark Craven, Jude W. Shavlik: Extracting Tree-Structured Representations of Trained Networks. NIPS 1995: 24-30

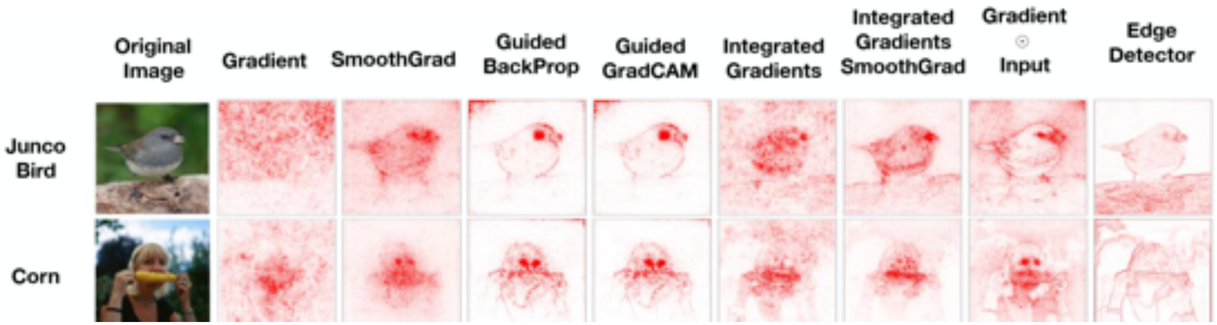
Explanation in different AI fields

- Machine Learning
- Computer Vision



Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

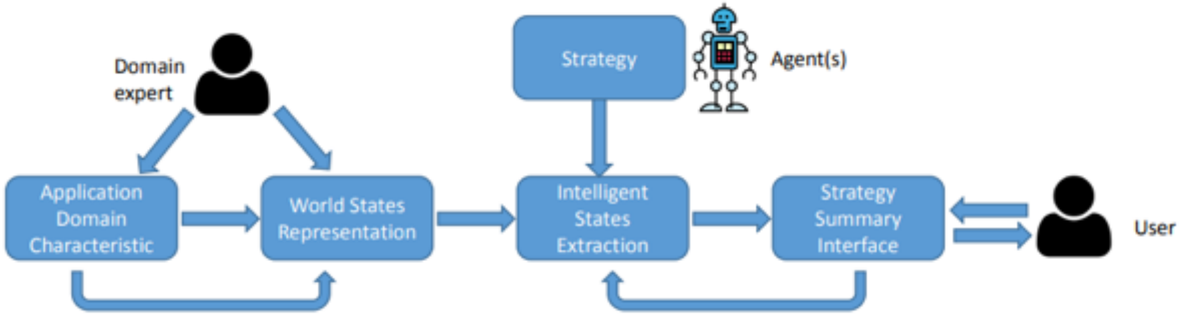


Saliency Map

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems



Agent Strategy Summarization

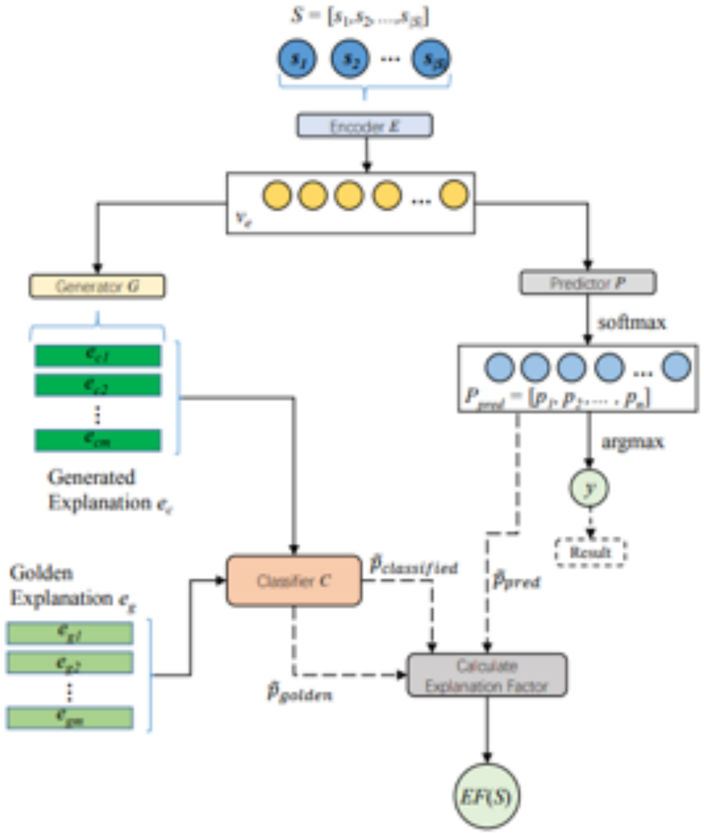
Ofra Amir, Finale Doshi-Velez, David Sarne: Agent Strategy Summarization. AAMAS 2018: 1203-1207

Explainable Agents

Joost Broekens, Maaike Harbers, Koen V. Hindriks, Karel van den Bosch, Catholijn M. Jonker, John-Jules Ch. Meyer: Do You Get It? User-Evaluated Explainable BDI Agents. MATES 2010: 28-39

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP

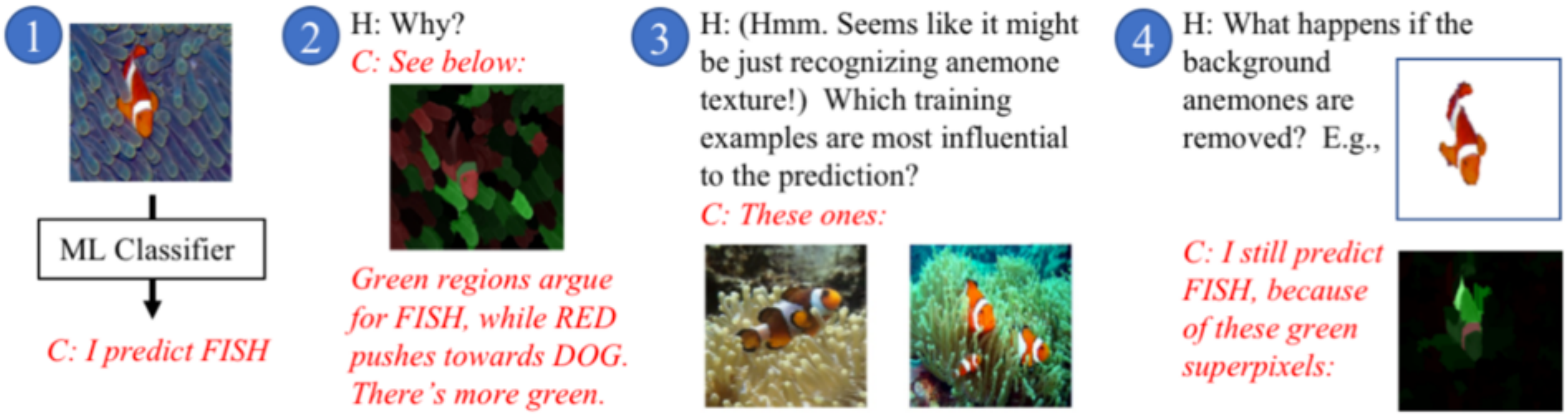


Explainable NLP

Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)

Explanation as *Machine-Human Conversation*

[Weld and Bansal 2018]



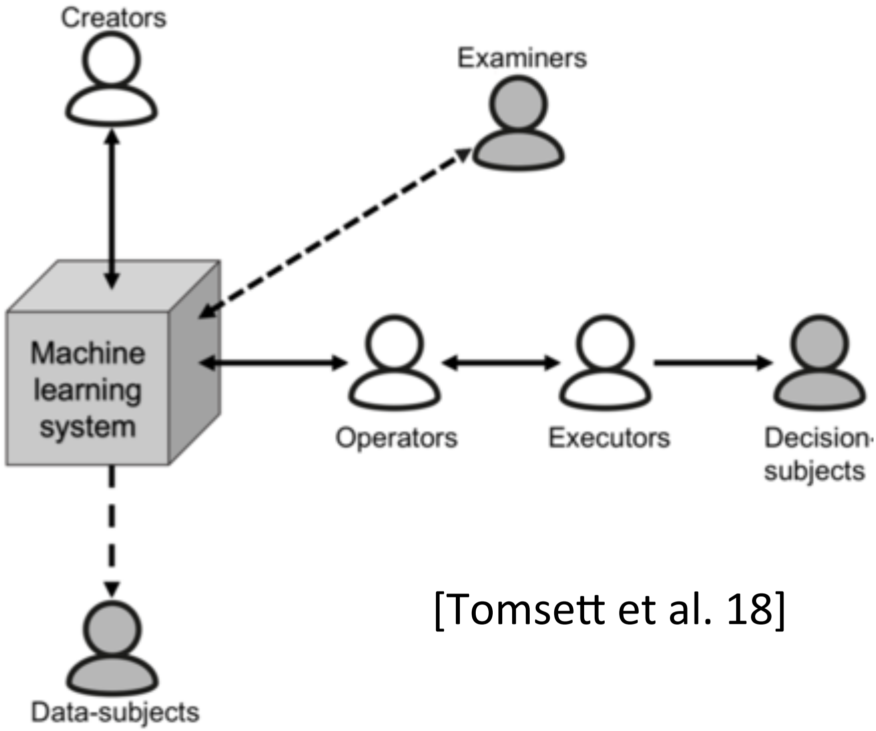
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

Role-based Interpretability

"~~Is the explanation interpretable~~?" → "*To whom* is the explanation interpretable?"

No Universally Interpretable Explanations!

- **End users** "Am I being treated fairly?"
  "Can I contest the decision?"
  "What could I do differently to get a positive outcome?"
- **Engineers, data scientists**: "Is my system working as designed?"
- **Regulators** " Is it compliant?"



[Tomsett et al. 18]

An ideal explainer should model the *user background.*

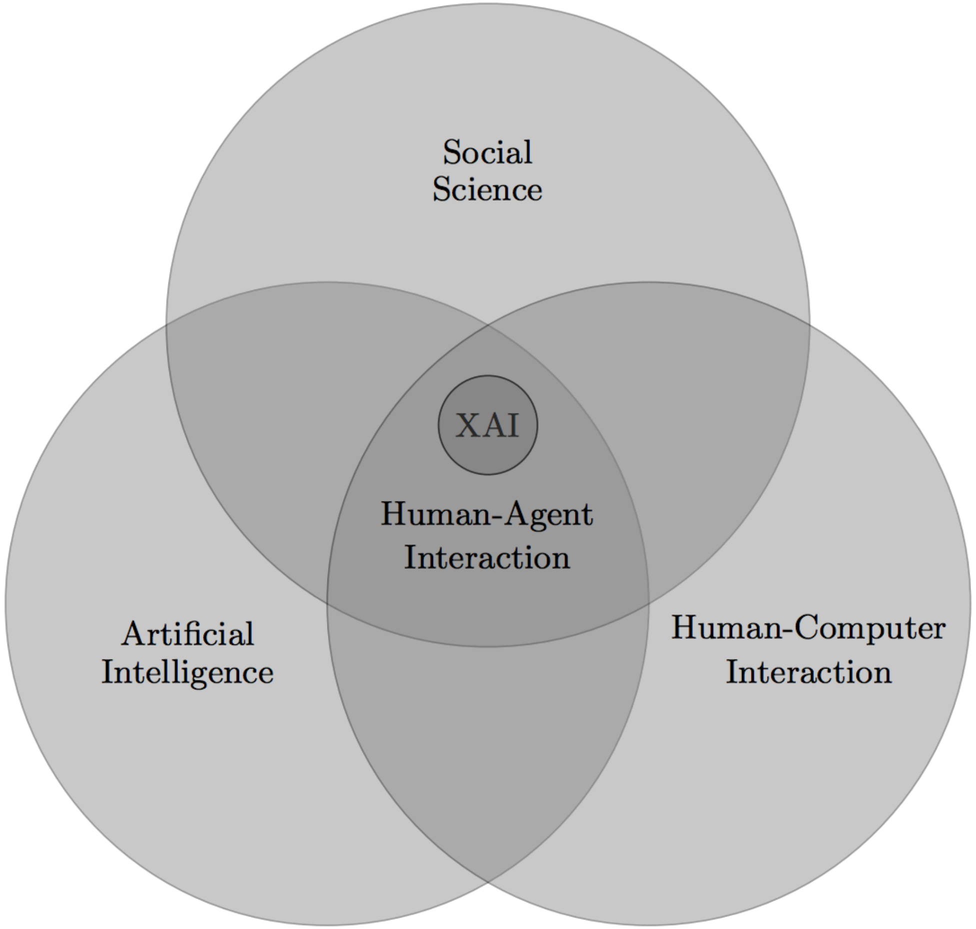
[Tomsett et al. 2018, Weld and Bansal 2018, Poursabzi-Sangdeh 2018, Mittelstadt et al. 2019]

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

XAI is Interdisciplinary

- For millennia, philosophers have asked the questions about what constitutes an explanation, what is the function of explanations, and what are their structure
- **[Tim Miller 2018]**



References

- [Tim Miller 2018] Tim Miller Explanaition in Artificial Intelligence: Insight from Social Science
- [Alvarez-Melis and Jaakkola 2018] Alvarez-Melis, David, and Tommi S. Jaakkola. "On the Robustness of Interpretability Methods." arXiv preprint arXiv:1806.08049 (2018).
- [Chen and Rudin 2018]: Chaofan Chen and Cynthia Rudin. An optimization approach to learning falling rule lists. In Artificial Intelligence and Statistics (AISTATS), 2018.
- [Doshi-Velez and Kim 2017] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
- [Goodman and Flaxman 2016] Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a" right to explanation"." arXiv preprint arXiv:1606.08813 (2016).
- [Freitas 2014] Freitas, Alex A. "Comprehensible classification models: a position paper." ACM SIGKDD explorations newsletter 15.1 (2014): 1-10.

- [Goodman and Flaxman 2016] Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a" right to explanation"." arXiv preprint arXiv:1606.08813 (2016).

- [Gunning 2017] Gunning, David. "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency (DARPA), nd Web (2017).
- [Hind et al. 2018] Hind, Michael, et al. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." arXiv preprint arXiv:1808.07261 (2018).

References

- [Kulesza et al. 2014] Kulesza, Todd, et al. "Principles of explanatory debugging to personalize interactive machine learning." Proceedings of the 20th international conference on intelligent user interfaces. ACM, 2015.
- [Lipton 2016] Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.
- [Mittelstatd et al. 2019] Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." arXiv preprint arXiv:1811.01439 (2018).
- [Poursabzi-Sangdeh 2018] Poursabzi-Sangdeh, Forough, et al. "Manipulating and measuring model interpretability." arXiv preprint arXiv:1802.07810 (2018).
- [Rudin 2018] Rudin, Cynthia. "Please Stop Explaining Black Box Models for High Stakes Decisions." arXiv preprint arXiv: 1811.10154 (2018).
- [Wachter et al. 2017] Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation." International Data Privacy Law 7.2 (2017): 76-99.
- [Weld and Bansal 2018] Weld, D., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of ACM (2018).
- [Yin 2012] Lou, Yin, Rich Caruana, and Johannes Gehrke. "Intelligible models for classification and regression." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2012).

Explaining Explanation Methods

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

COMPAS recidivism black bias

The background bias

(a) Husky classified as wolf

(b) Explanation

FAIRNESS

Is the model fair to every group and/or individual?

ACCOUNTABILITY

What can we attribute the decision to?

TRANSPARENCY

How did the model come up with the decision?

FAIRNESS

Is the model fair to every group and/or individual?

ACCOUNTABILITY

What can we attribute the decision to?

TRANSPARENCY

How did the model come up with the decision?

FAIRNESS

Is the model fair to every group and/or individual?

ACCOUNTABILITY

What can we attribute the decision to?

TRANSPARENCY

How did the model come up with the decision?

FAIRNESS

Is the model fair to every group and/or individual?

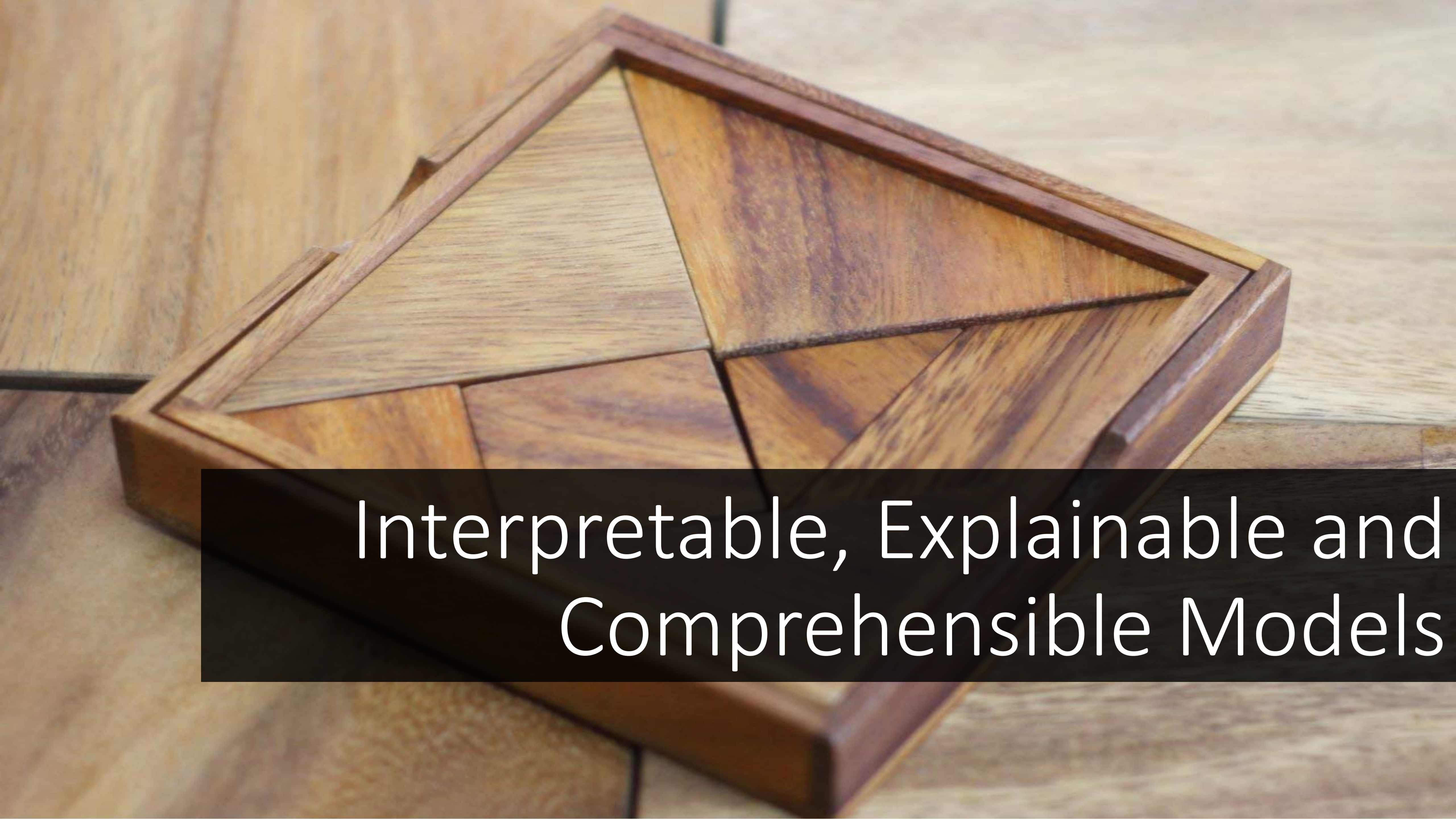
ACCOUNTABILITY

What can we attribute the decision to?

TRANSPARENCY

How did the model come up with the decision?

Interpretable, Explainable and Comprehensible Models



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.

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.

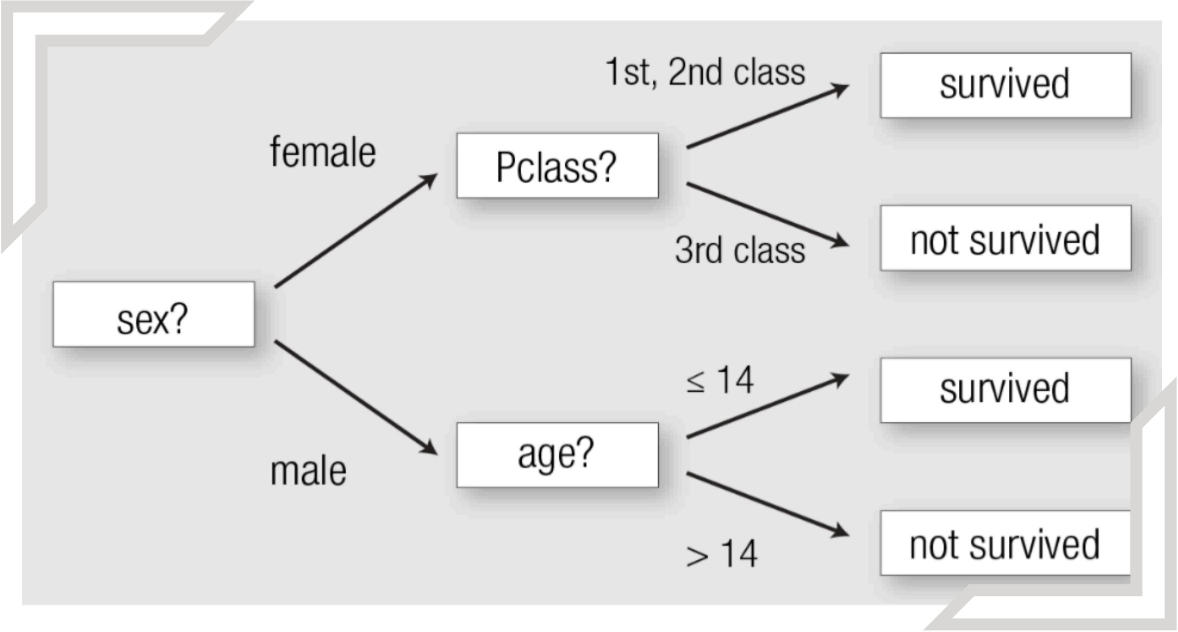
- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

Desiderata of an Interpretable Model

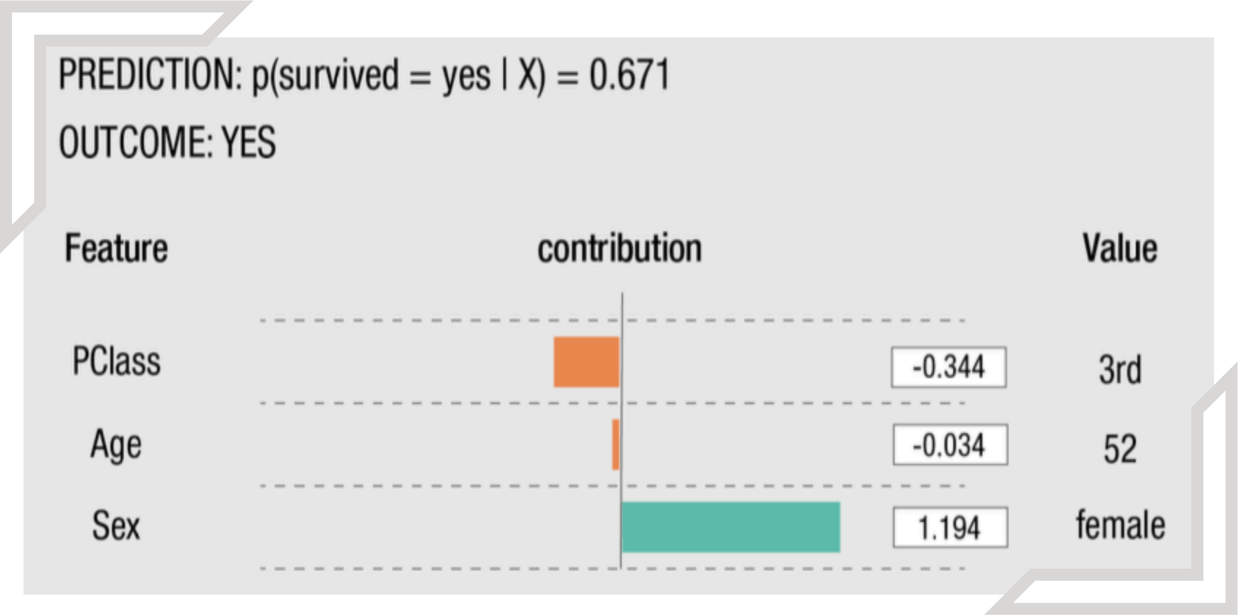
- ***Fairness***: the model guarantees the protection of groups against discrimination.
- ***Privacy***: the model does not reveal sensitive information about people.
- ***Respect Monotonicity***: the increase of the values of an attribute either increase or decrease in a monotonic way the probability 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.

Recognized Interpretable Models



Decision Tree

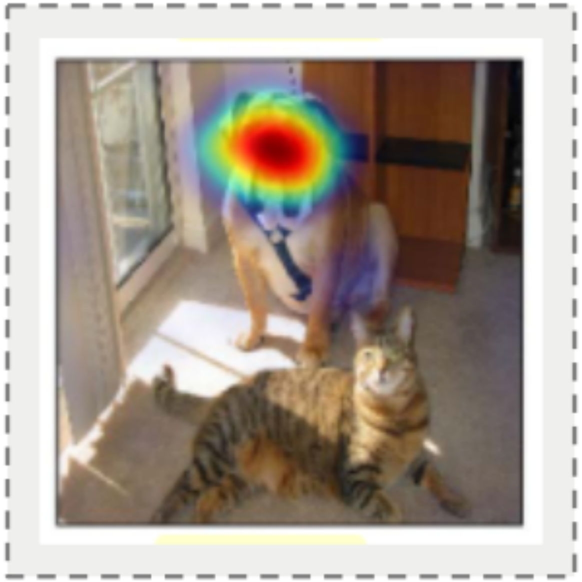


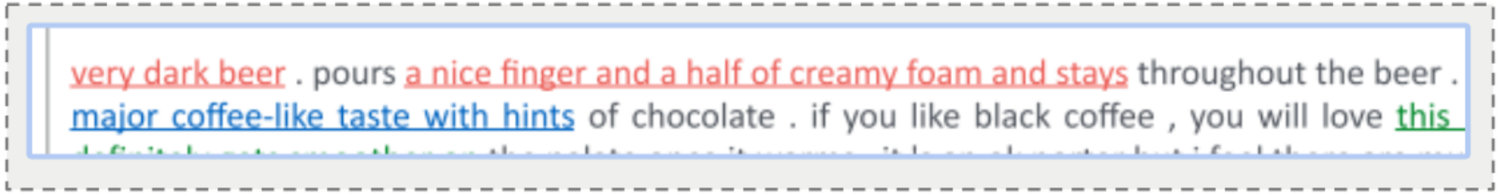
Linear Model

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

Rules

Explainations: Saliency Maps

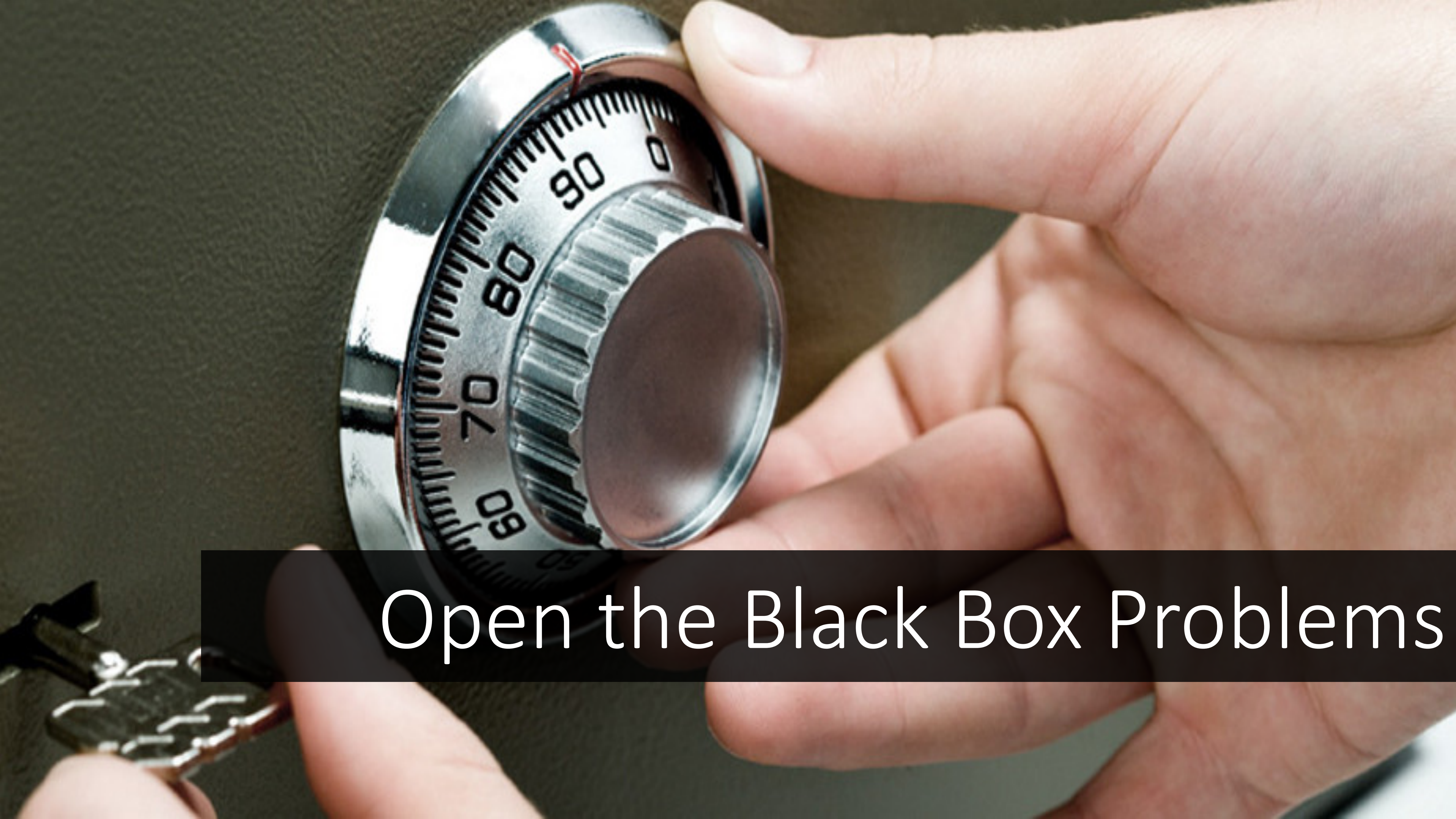




Complexity

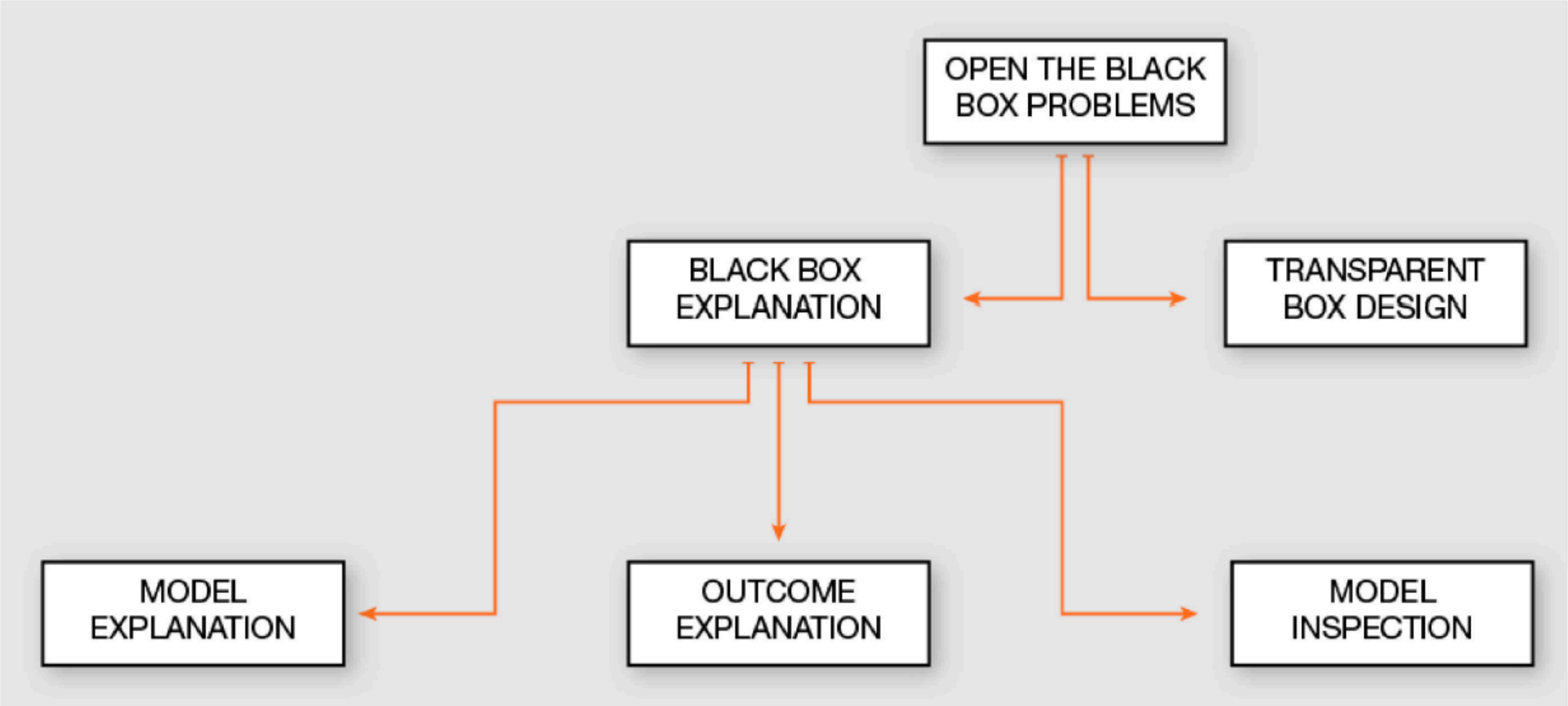
- Opposed to *interpretability*.
- Is only related to the model and not to the training data that is unknown.
- Generally estimated with a rough approximation related to the ***size*** of the interpretable model.
- Linear Model: number of non zero weights in the model.
- Rule: number of attribute-value 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.

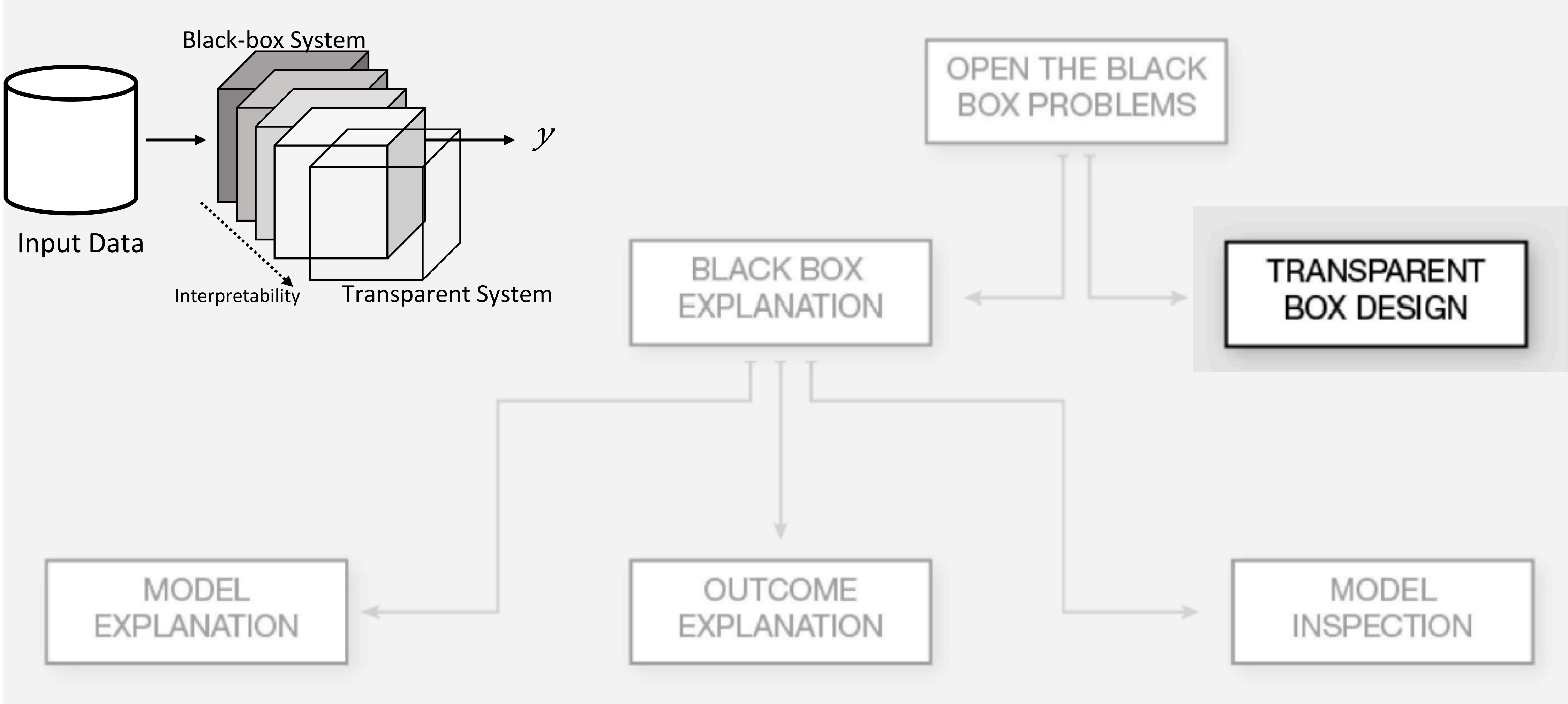


Open the Black Box Problems

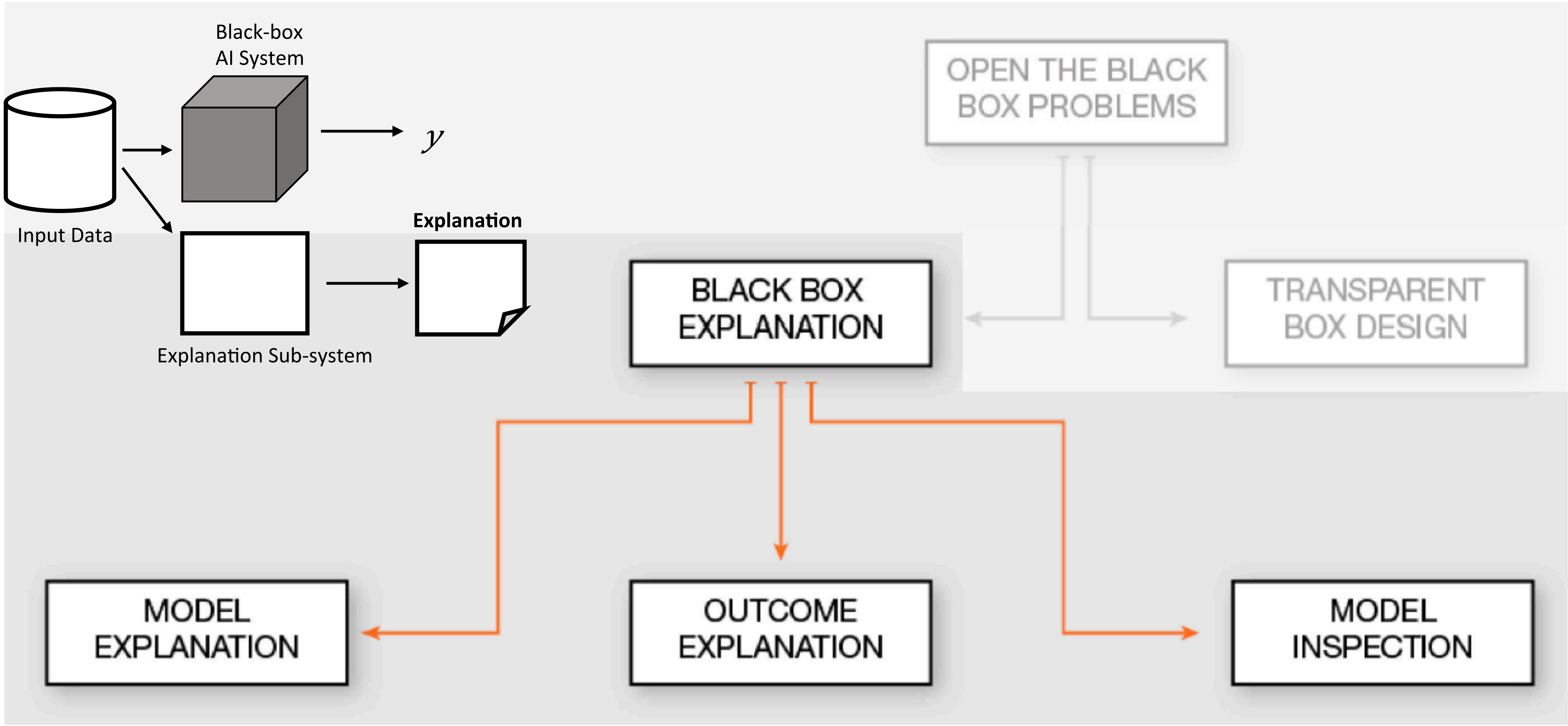
Problems Taxonomy



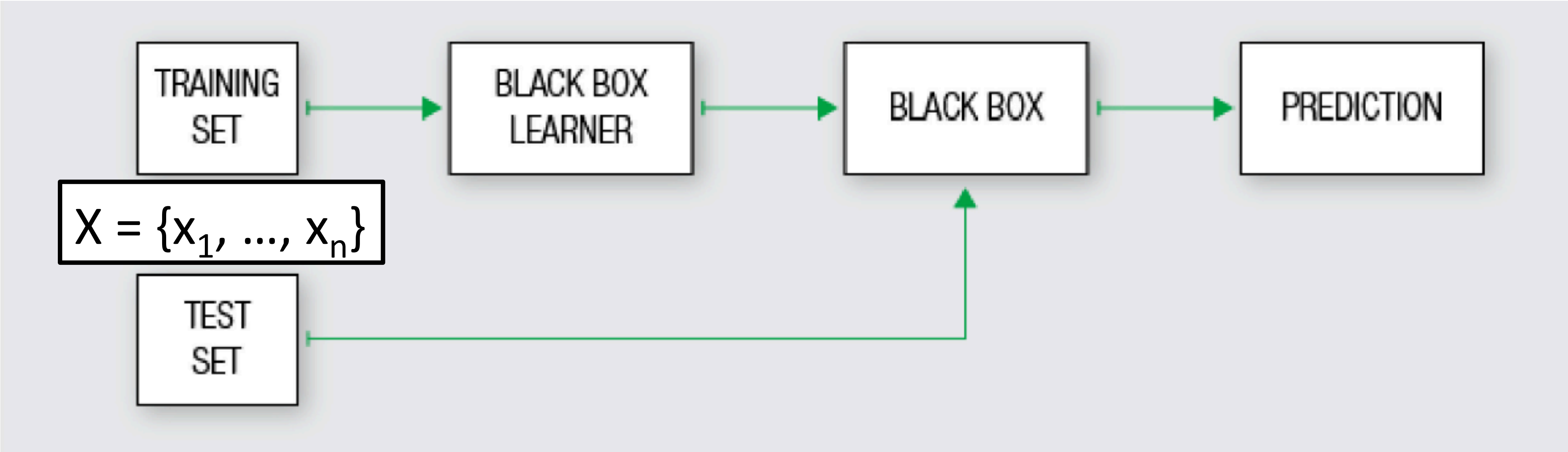
XbD – eXplanation by Design



BBX - Black Box eXplanation

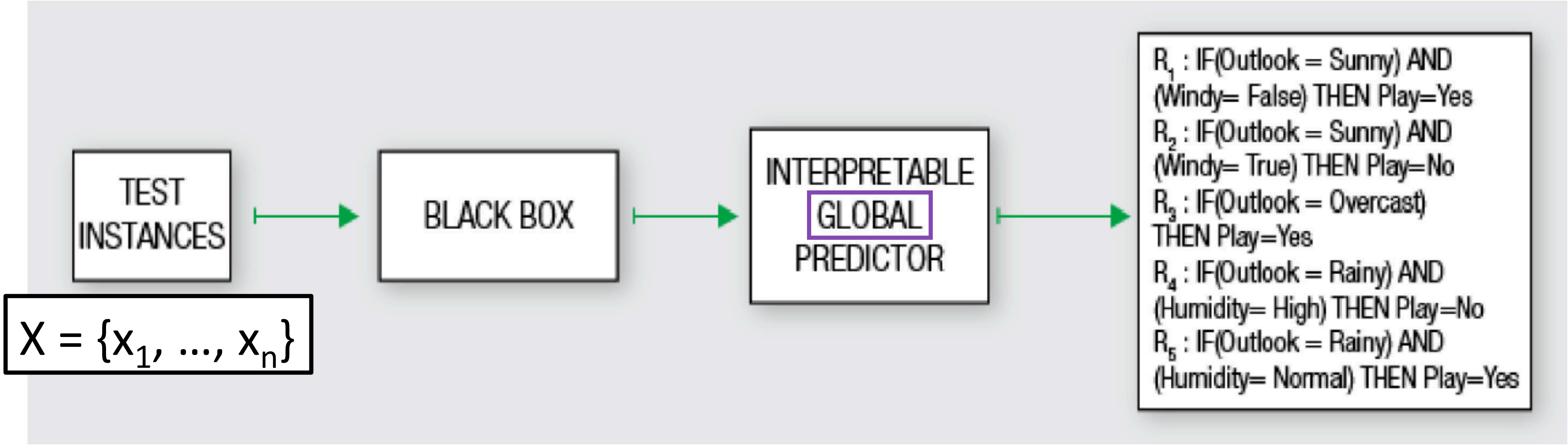


Classification Problem



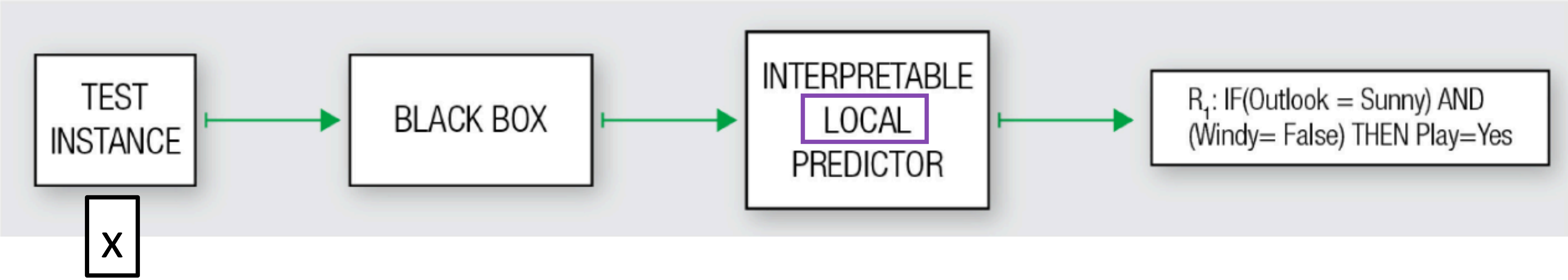
Model Explanation Problem

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



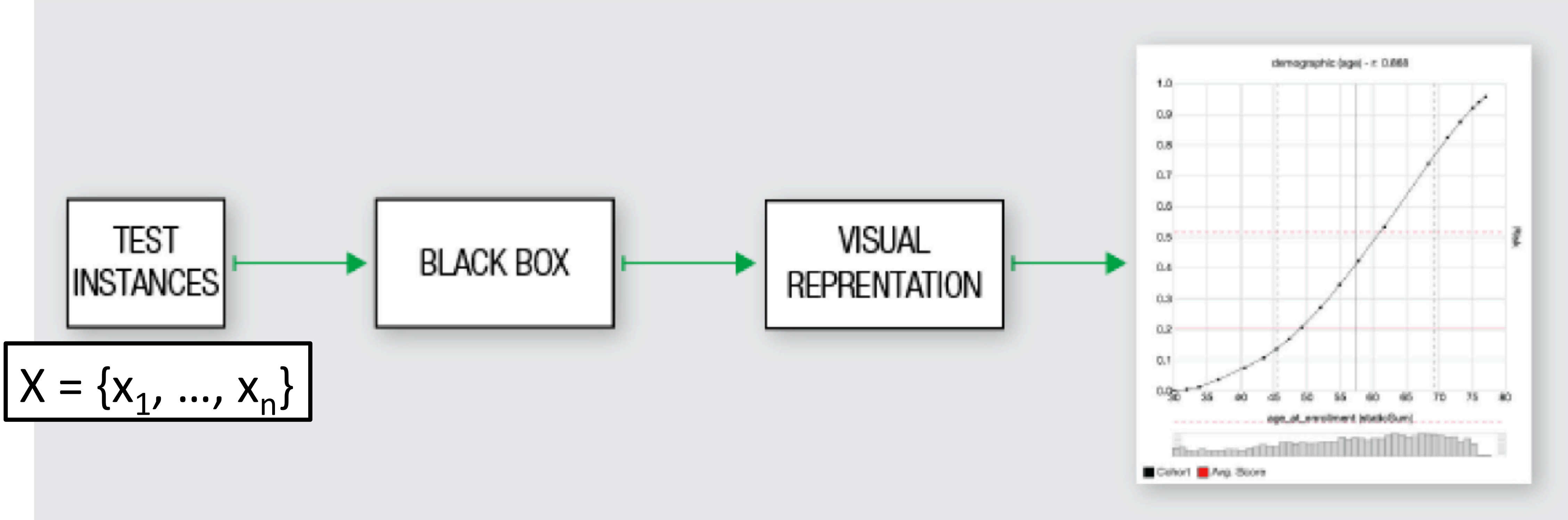
Outcome Explanation Problem

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



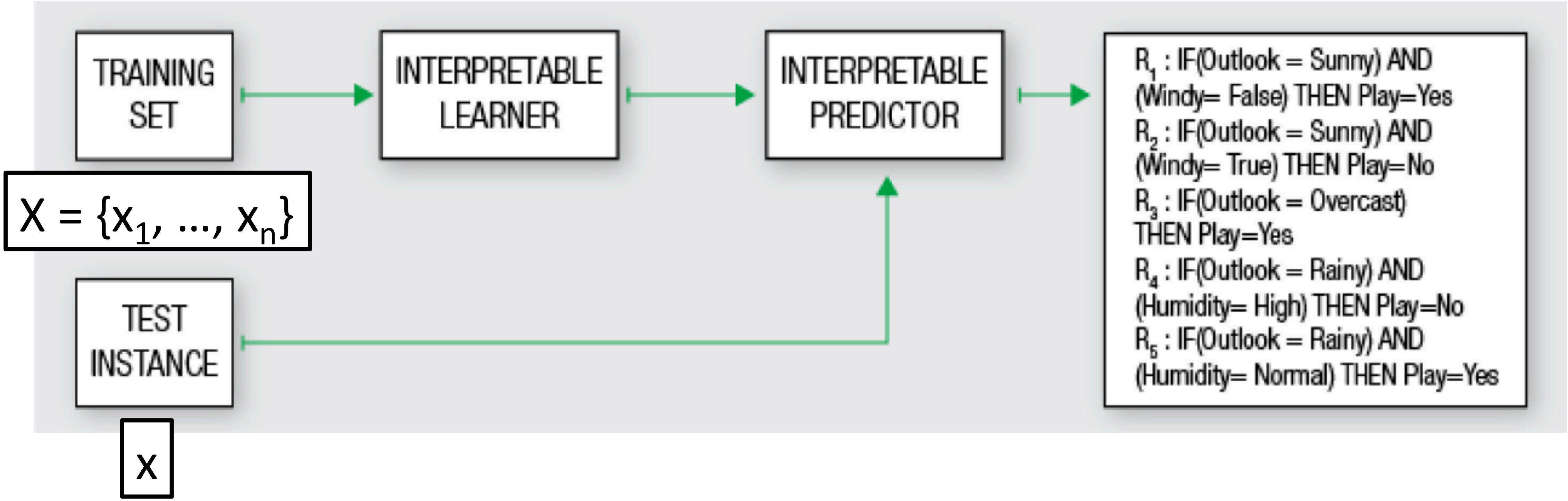
Model Inspection Problem

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.

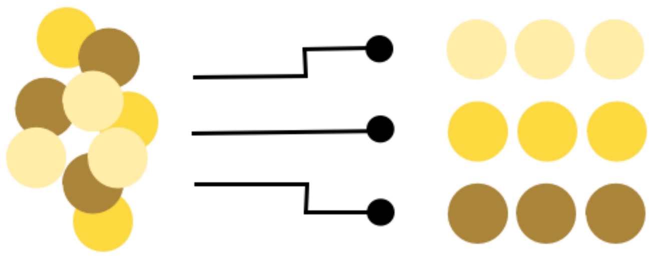


Transparent Box Design Problem

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)

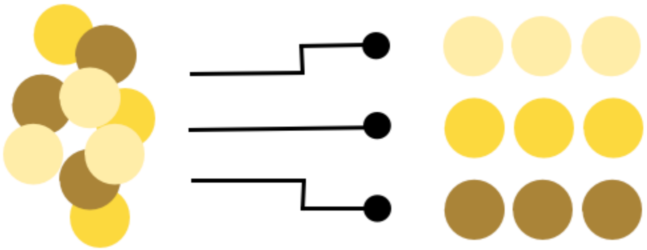
2000 rows all told

Tabular
(TAB)

Images
(IMG)

Text
(TXT)

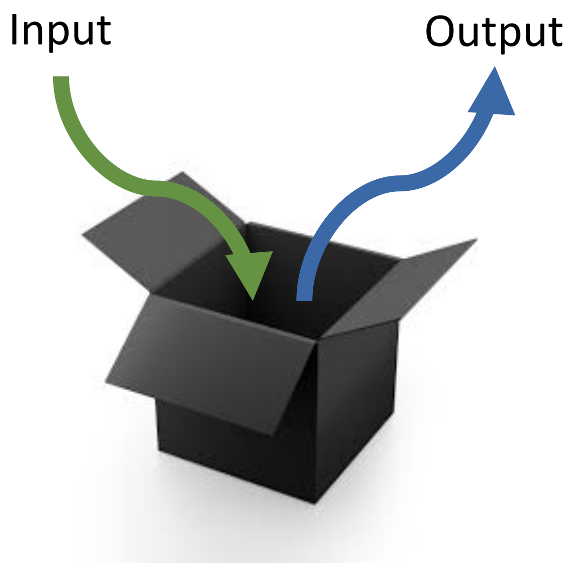
Explanators



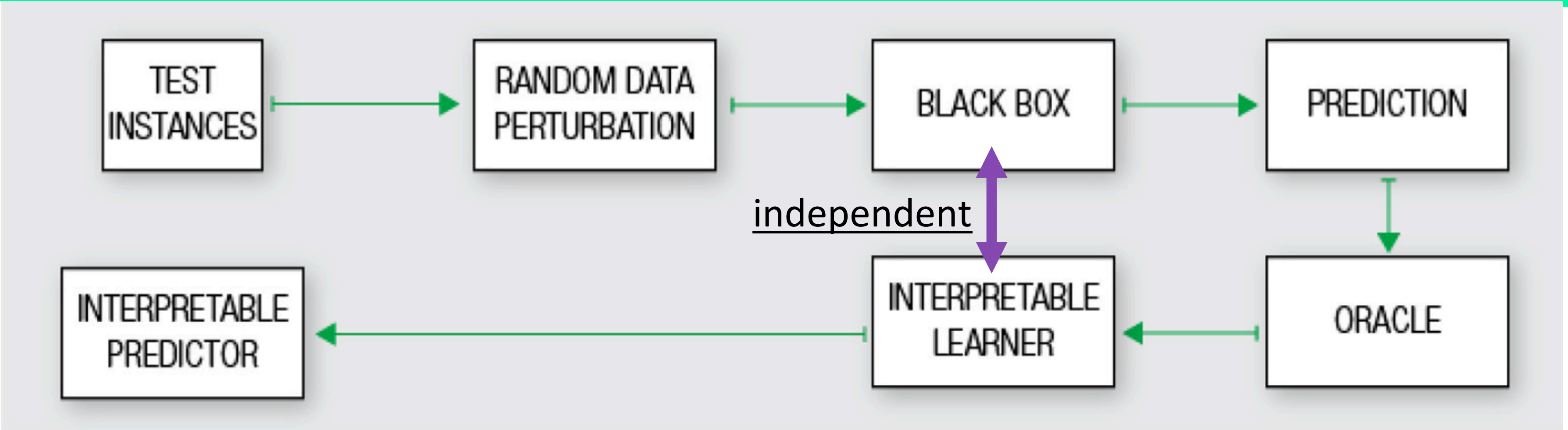
- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Maps (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)
- Activation Maximization (***AM***)

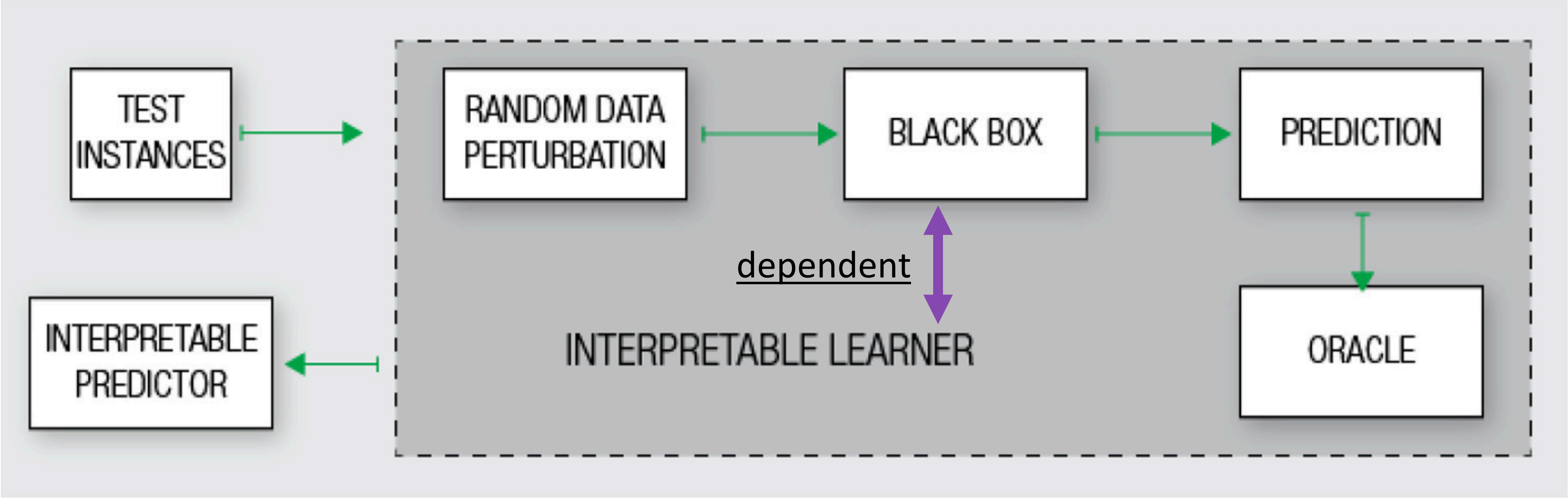
Reverse Engineering

- 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					✓
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]
REFNE	[141]	Zhou et al.	2003	DR	NN	TAB	✓	✓	✓		✓
RxREN	[6]	Augasta et al.	2012	DR	NN	TAB		✓	✓		✓

Global Model Explainers

- Explanator: DT
  - Black Box: NN, TE
  - Data Type: TAB

- Explanator: DR
  - Black Box: NN, SVM, TE
  - Data Type: TAB

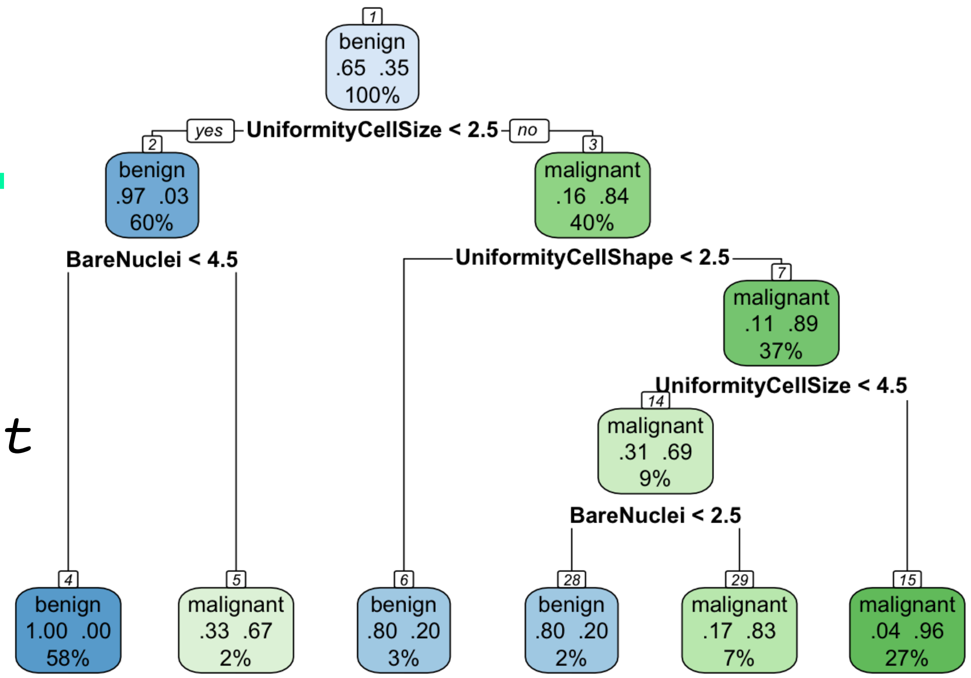
- Explanator: FI
  - Black Box: AGN
  - Data Type: TAB

R_1 : IF(Outlook = Sunny) AND (Windy= False) THEN Play=Yes
R_2 : IF(Outlook = Sunny) AND (Windy= True) THEN Play=No
R_3 : IF(Outlook = Overcast) THEN Play=Yes
R_4 : IF(Outlook = Rainy) AND (Humidity= High) THEN Play=No
R_5 : IF(Outlook = Rainy) AND (Humidity= Normal) THEN Play=Yes

Trepan – DT, NN, TAB

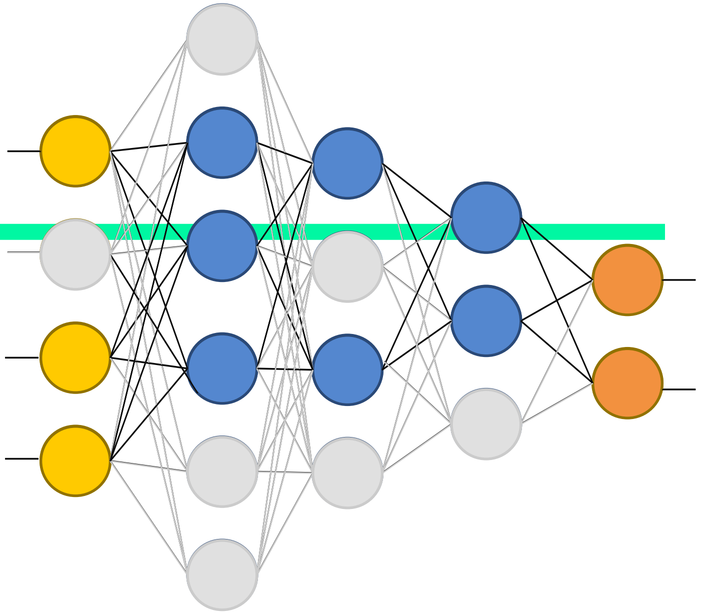
black box auditing

```
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 ∪ y_Z)
        continue
    S = best_split(X_N ∪ Z_N, y ∪ 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 ∪ {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.

RxREN – DR, NN, TAB



```
prune insignificant neurons
for each significant neuron
  for each outcome
    compute mandatory data ranges
for each outcome
  build rules using data ranges of each neuron
prune insignificant rules
update data ranges in rule conditions analyzing error
```

black box auditing

if $((data(I_1) \geq L_{13} \wedge data(I_1) \leq U_{13}) \wedge (data(I_2) \geq L_{23} \wedge data(I_2) \leq U_{23}) \wedge$
$(data(I_3) \geq L_{33} \wedge data(I_3) \leq U_{33}))$ then class $=C_3$
else
if $((data(I_1) \geq L_{11} \wedge data(I_1) \leq U_{11}) \wedge (data(I_3) \geq L_{31} \wedge data(I_3) \leq U_{31}))$
then class $=C_1$
else
class $= C_2$

- M. Gethsiyal Augasta and T. Kathirvalavakumar. 2012. ***Reverse engineering the neural networks for rule extraction in classification problems***. NPL.

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	✓	✓	✓		✓

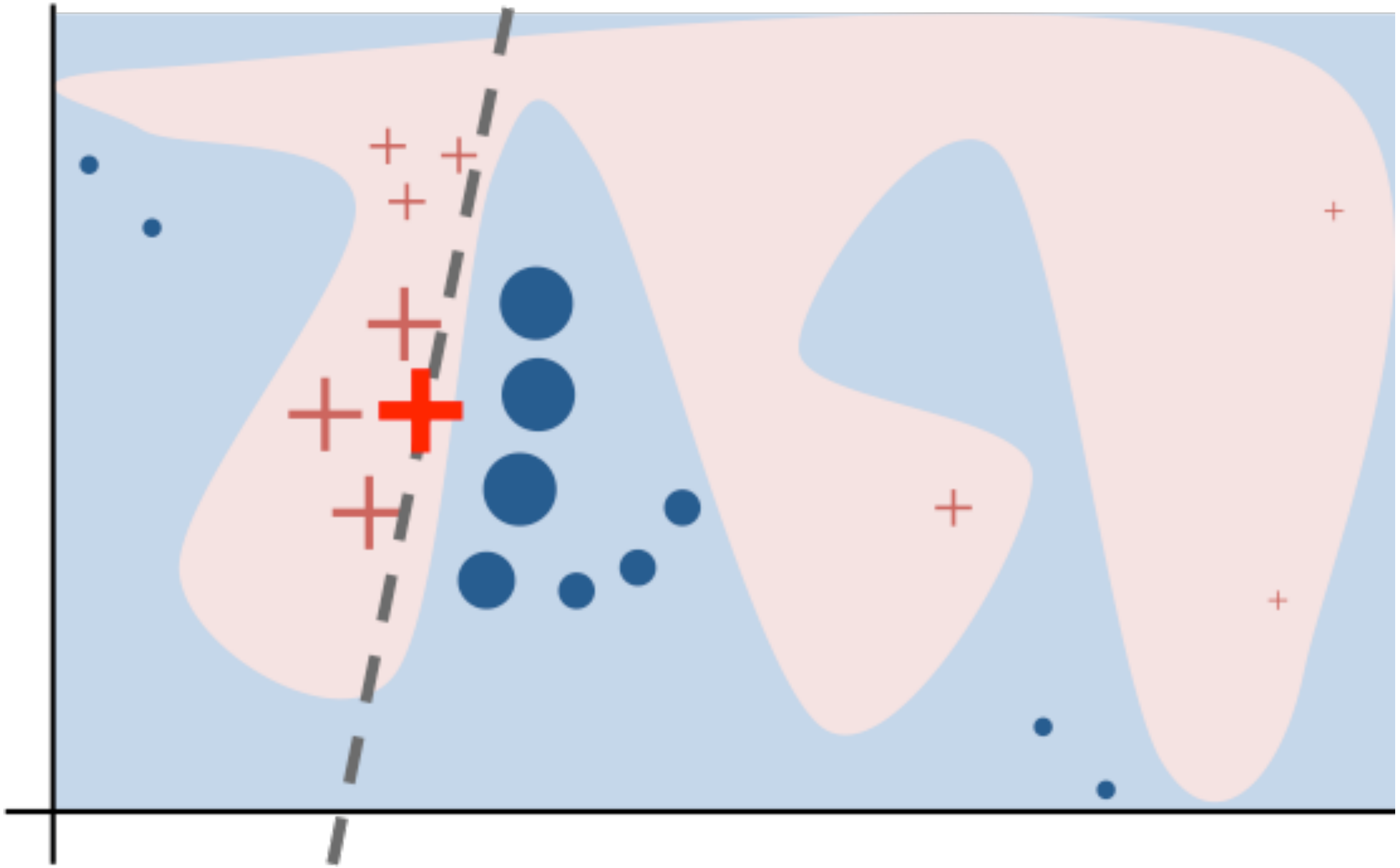
Local Model Explainers

- Explanator: SM
  - Black Box: DNN, NN
  - Data Type: IMG
- Explanator: FI
  - Black Box: DNN, SVM
  - Data Type: ANY
- Explanator: DT
  - Black Box: ANY
  - Data Type: TAB

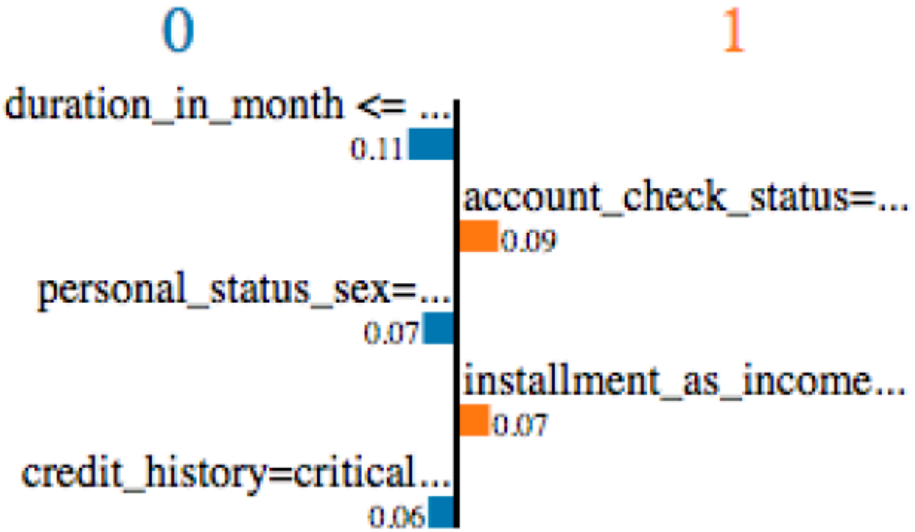
R_1: IF(Outlook = Sunny) AND (Windy= False) THEN Play=Yes

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.

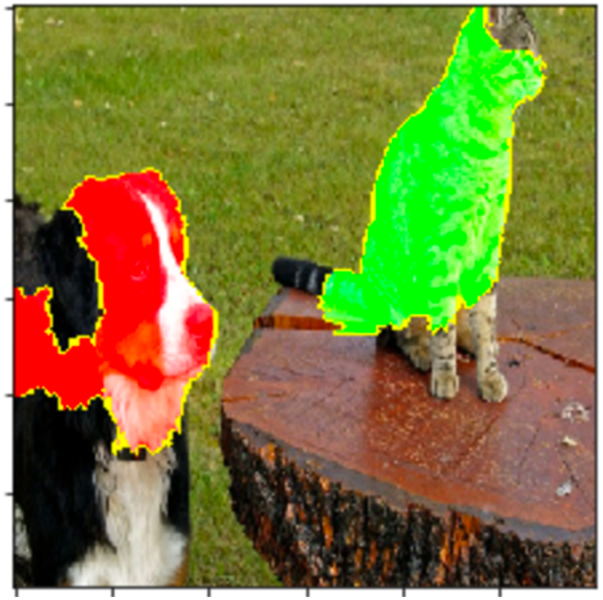


LIME – FI, AGN, ANY



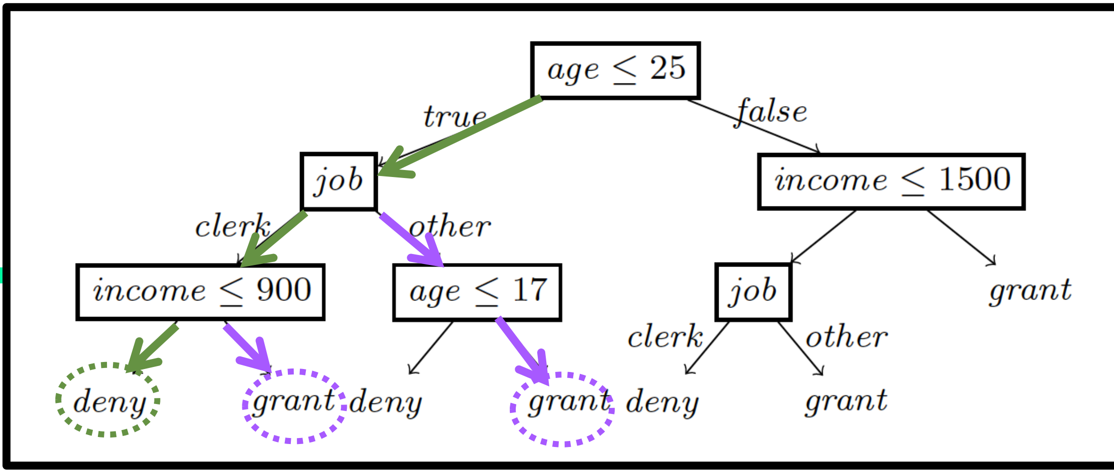

```
Z = {}
x instance to explain
x' = real2interpretable(x)
for i in {1, 2, …, N}
      zi= sample_around(x')
      z = interpretabel2real(z')
      Z = Z ∪ {<zi, b(zi), d(x, z)>}
w = solve_Lasso(Z, k)
return w
```

black box auditing



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

LORE – DR, AGN, TAB



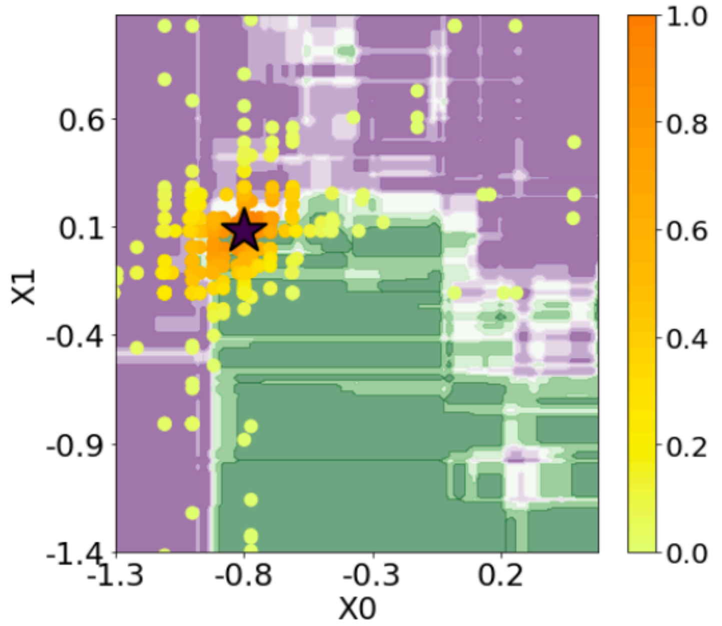

```
x instance to explain
Z= = geneticNeighborhood(x, fitness=, N/2)
Z≠ = geneticNeighborhood(x, fitness≠, N/2)
Z = Z= ∪ Z≠
c = buildTree(Z, b(Z))
r = (p -> y) = extractRule(c, x)
φ = extractCounterfactual(c, r, x)
return e = <r, φ>
```

black box auditing

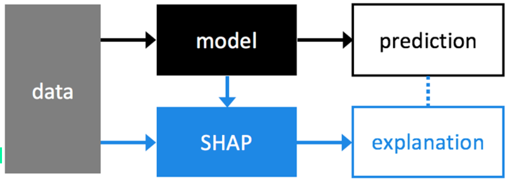
r = {age ≤ 25, job = clerk, income ≤ 900} -> deny

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

Pedreschi, Franco Turini,
f black box decision



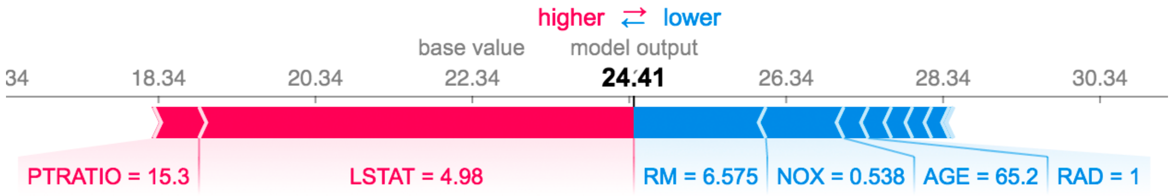
SHAP (SHapley Additive exPlanations)

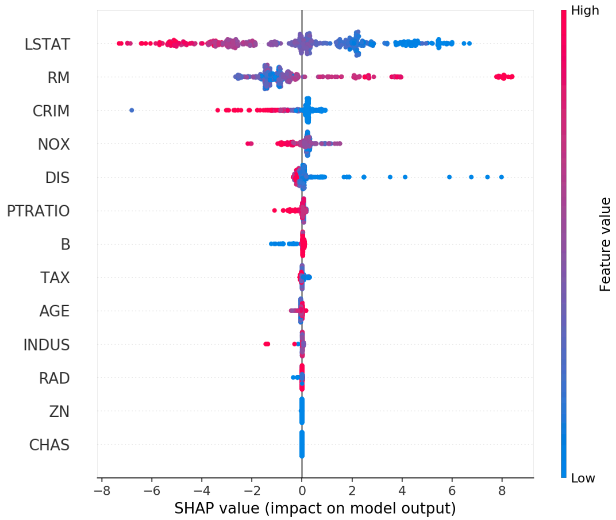


- SHAP assigns each feature an importance value for a particular prediction by means of an additive feature attribution method.
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$



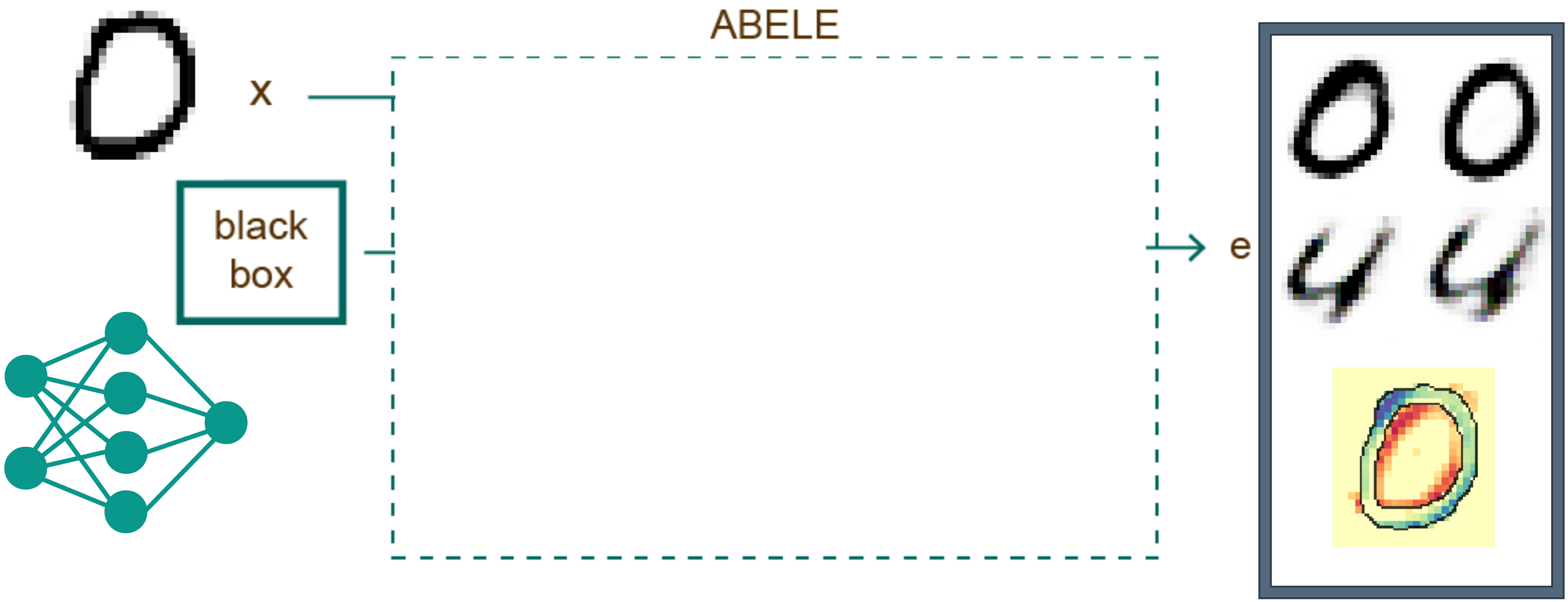


- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017.

Black Box Explanation by **Learning Image Exemplars** in the Latent Feature Space

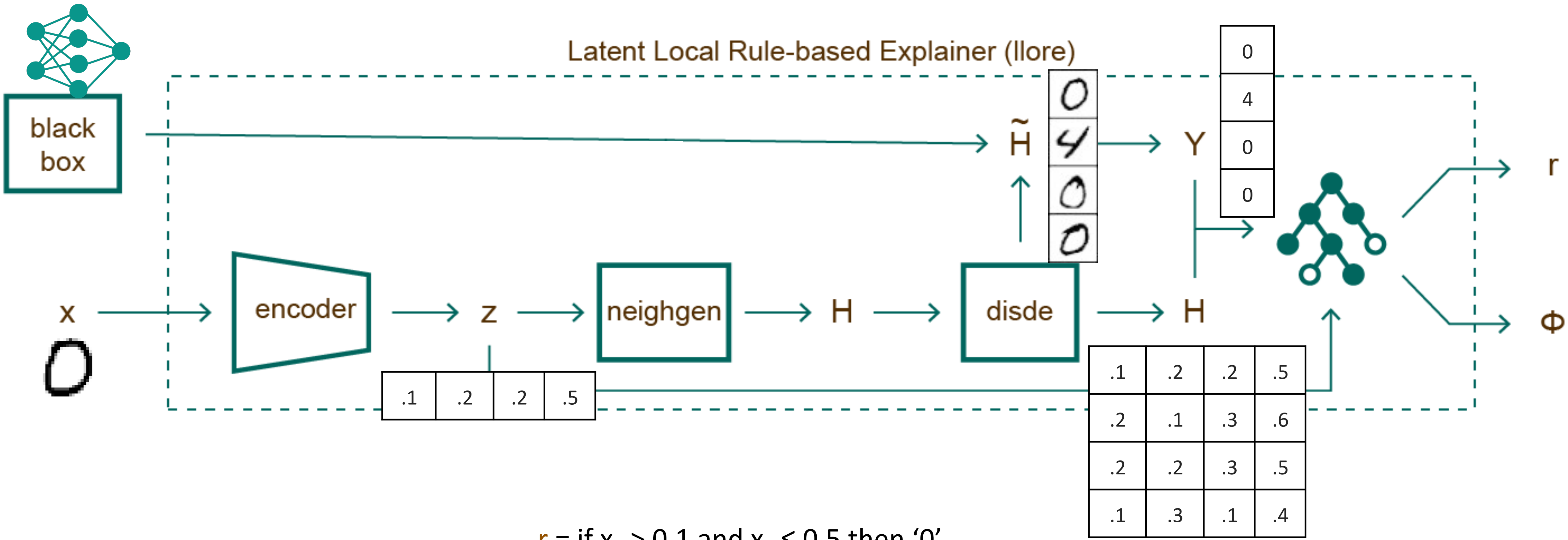
DALHOUSIE UNIVERSITY

Adversarial Black box Explainer generating Latent Exemplars



https://github.com/riccotti/ABELE

Latent Local Rule Extraction



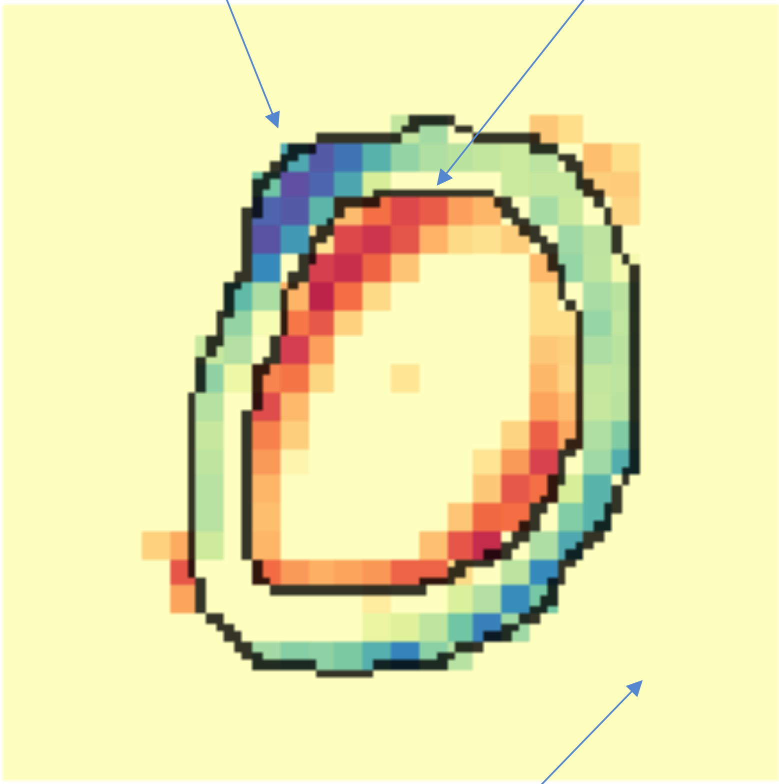
r = if $x_1 > 0.1$ and $x_3 \leq 0.5$ then '0'

φ = {if $x_1 \leq 0.1$ then '4',
if $x_3 > 0.5$ 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.

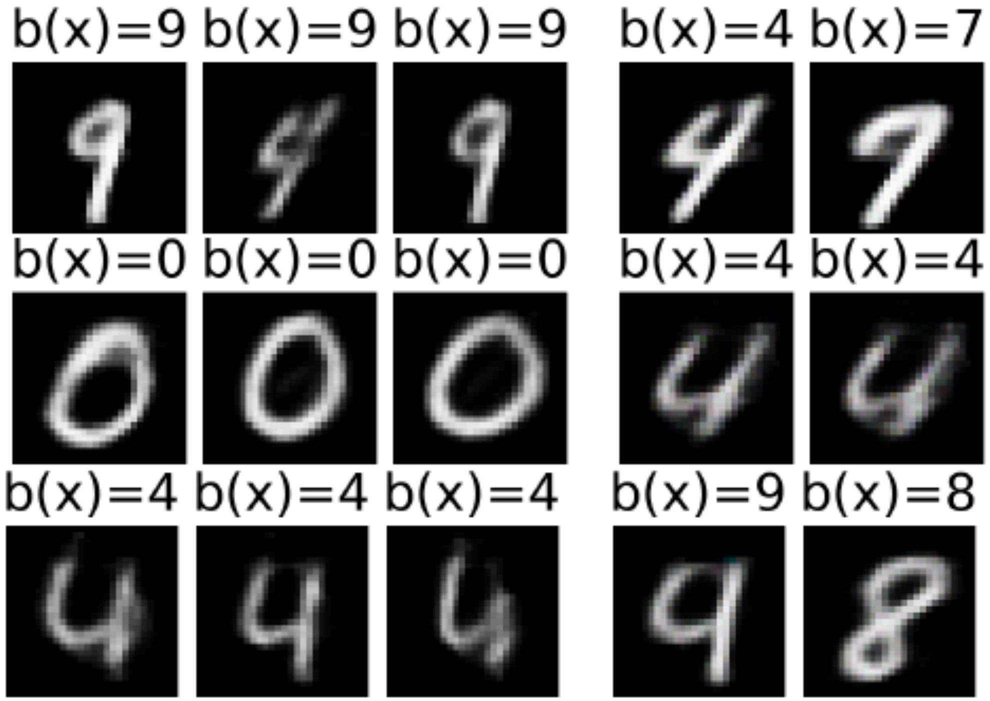
Saliency Map from Exemplars

- 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.

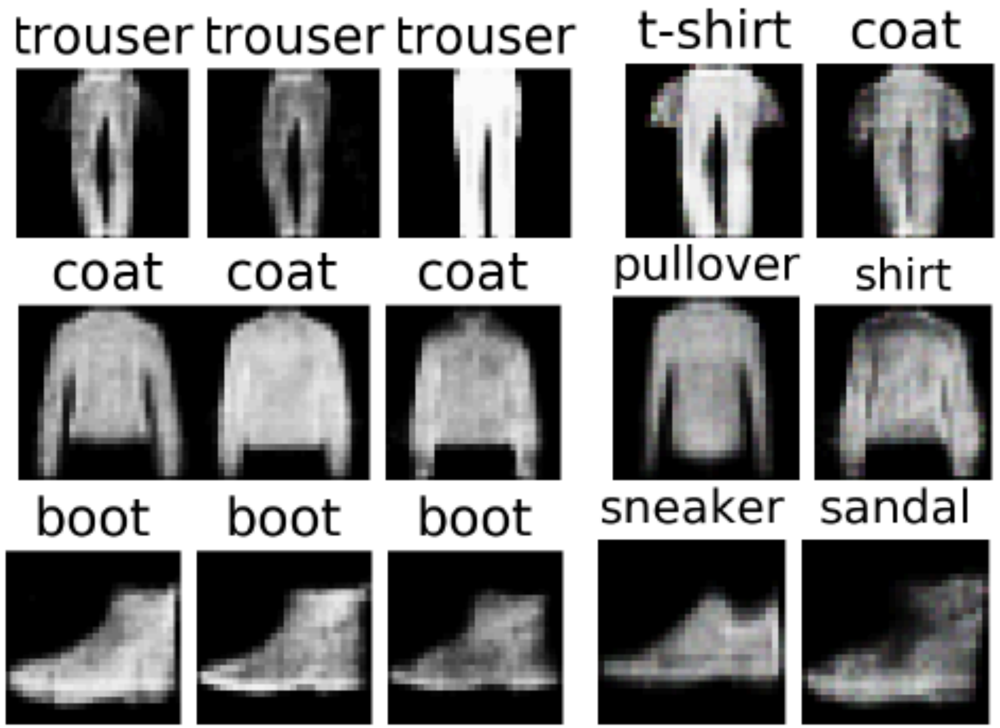


Exemplars and Counter-Exemplars

- **mnist**

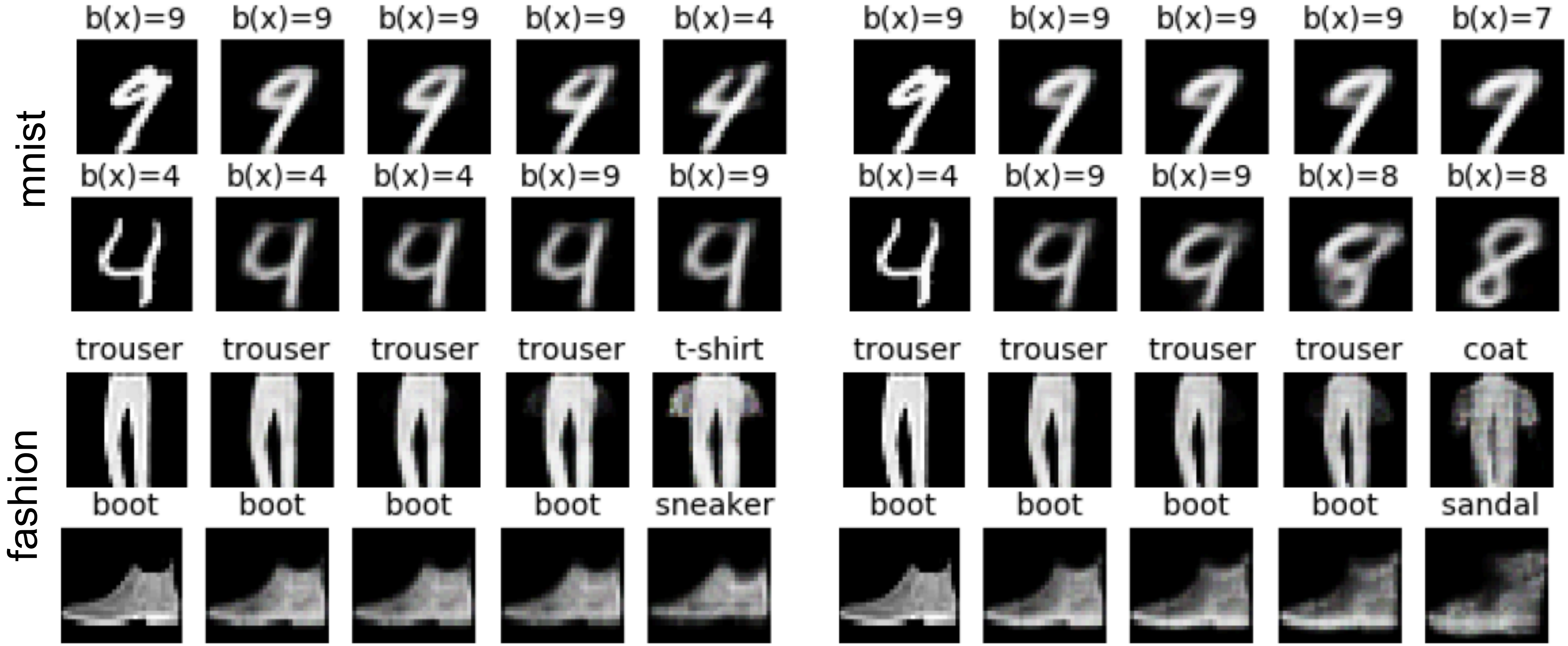


- **fashion**



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.

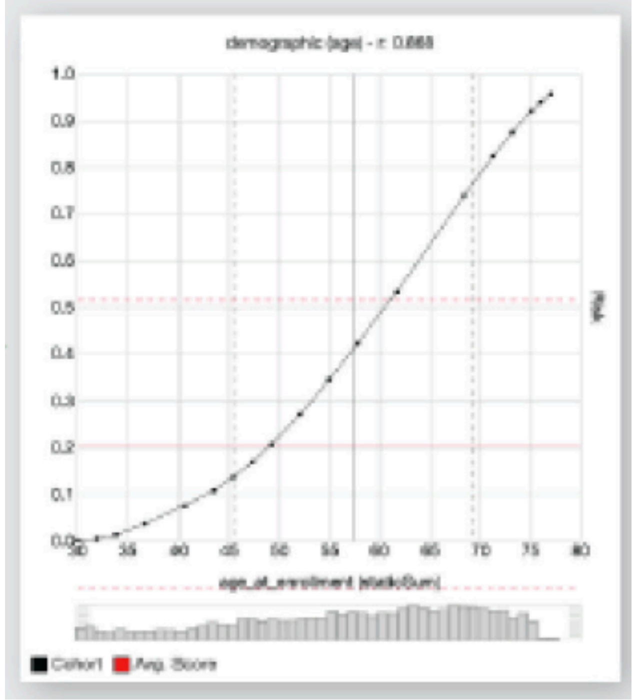


Solving The Model Inspection Problem

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
NID	[83]	Olden et al.	2002	SA	NN	TAB			✓		
GDP	[8]	Baehrens	2010	SA	AGN	TAB	✓		✓		✓
QII	[24]	Datta et al	2016	SA	AGN	TAB	✓		✓		✓
IG	[115]	Sundararajan	2017	SA	DNN	ANY			✓		✓
VEC	[18]	Cortez et al.	2011	SA	AGN	TAB	✓		✓		✓
VIN	[42]	Hooker	2004	PDP	AGN	TAB	✓		✓		✓
ICE	[35]	Goldstein et al.	2015	PDP	AGN	TAB	✓		✓	✓	✓
Prospector	[55]	Krause et al.	2016	PDP	AGN	TAB	✓		✓		✓
Auditing	[2]	Adler et al.	2016	PDP	AGN	TAB	✓		✓	✓	✓
OPIA	[1]	Adebayo et al.	2016	PDP	AGN	TAB	✓		✓		
–	[136]	Yosinski et al.	2015	AM	DNN	IMG			✓		✓
IP	[108]	Shwartz et al.	2017	AM	DNN	TAB			✓		
–	[137]	Zeiler et al.	2014	AM	DNN	IMG		✓		✓	
–	[112]	Springenberg et al.	2014	AM	DNN	IMG			✓		✓
DGN-AM	[80]	Nguyen et al.	2016	AM	DNN	IMG			✓	✓	✓

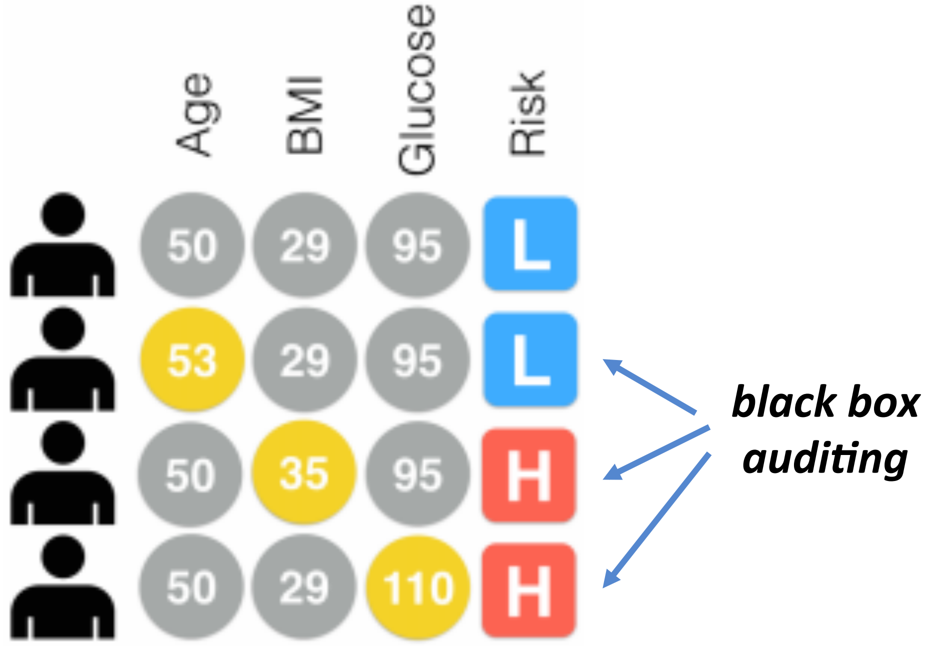
Inspection Model Explainers

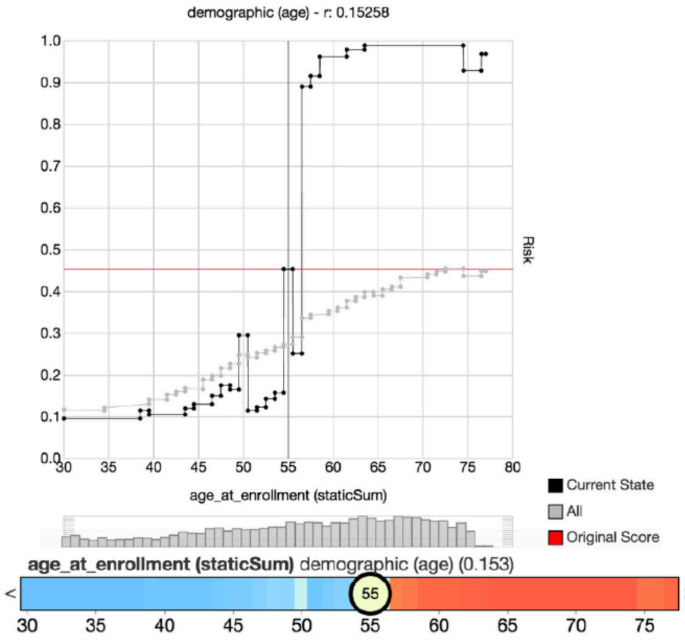
- Explanator: SA
  - Black Box: NN, DNN, AGN
  - Data Type: TAB
- Explanator: PDP
  - Black Box: AGN
  - Data Type: TAB
- Explanator: AM
  - Black Box: DNN
  - Data Type: IMG, TXT



Prospector – PDP, AGN, TAB

- Introduce ***random perturbations*** on input values to understand to which extent every feature impact the prediction using PDPs.
- The input is changed ***one variable at a time***.





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

Conclusions

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?

Future Challenges

- Creating awareness! Success stories!
- Foster multi-disciplinary collaborations in XAI research.
- Help shaping industry standards, legislation.
- More work on transparent design.
- Investigate symbolic and sub-symbolic reasoning.
- Evaluation:
  - We need benchmark - Shall we start a task force?
  - We need an XAI challenge - Anyone interested?
  - Rigorous, agreed upon, human-based evaluation protocols

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

Thank you!

Anna Monreale
University of Pisa

Dino Pedreschi
University of Pisa

Riccardo Guidotti
University of Pisa