XAI - Science & technology for the eXplanation of AI decision making

Riccardo Guidotti

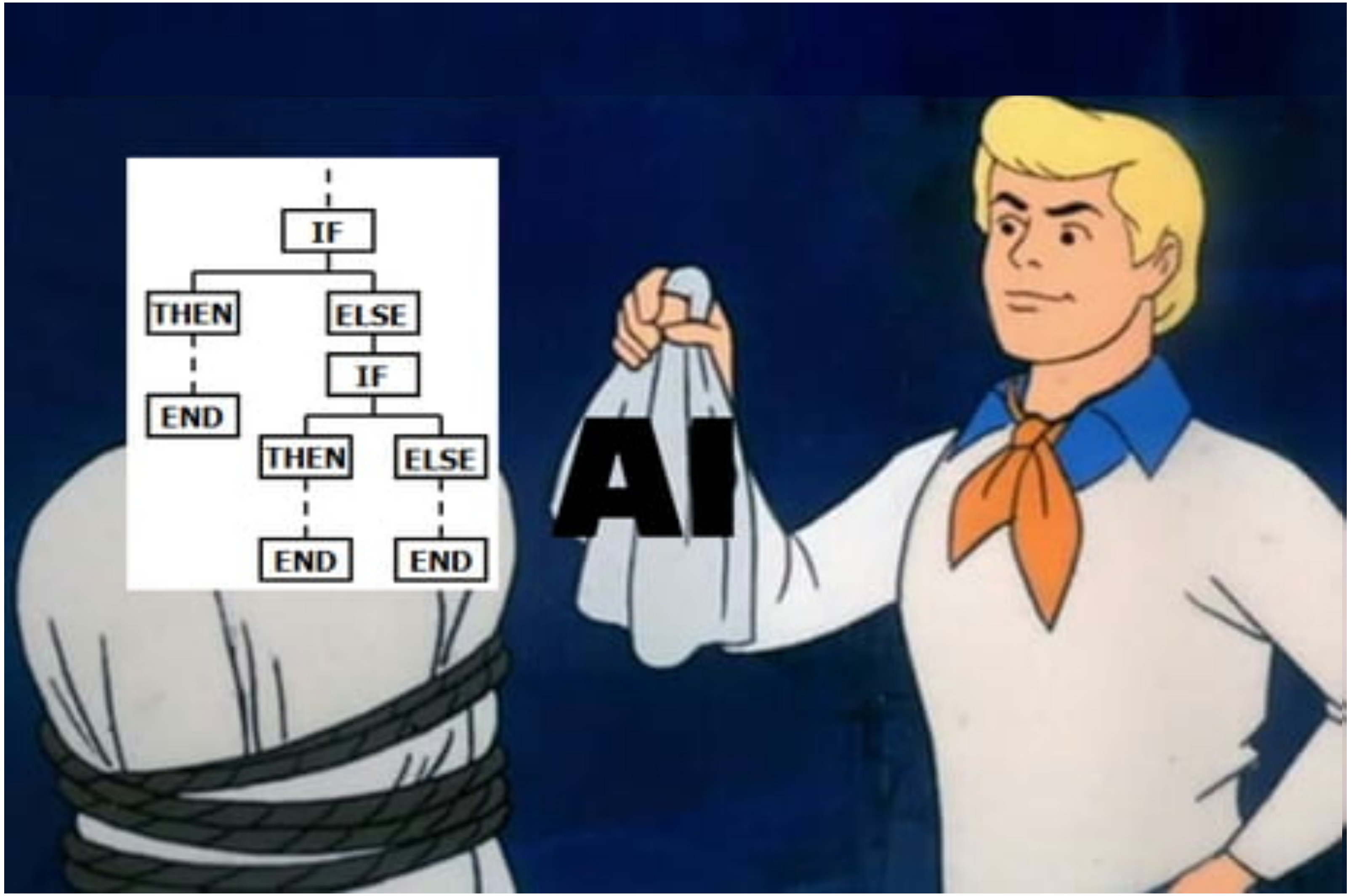
KDD Lab

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

UNIVERSITÀ DI PISA

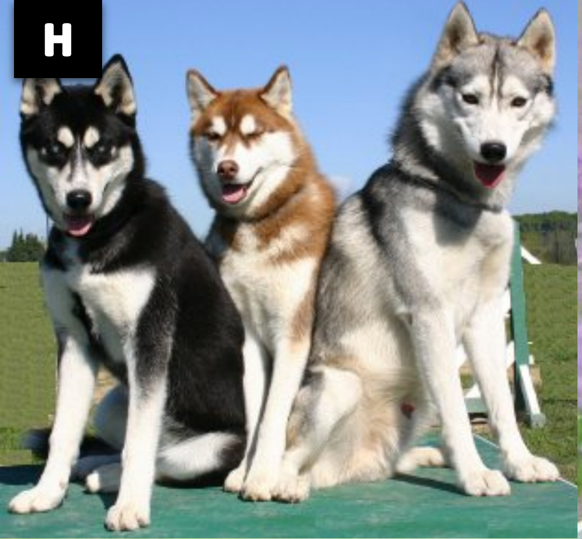
What is “Explainable AI” ?

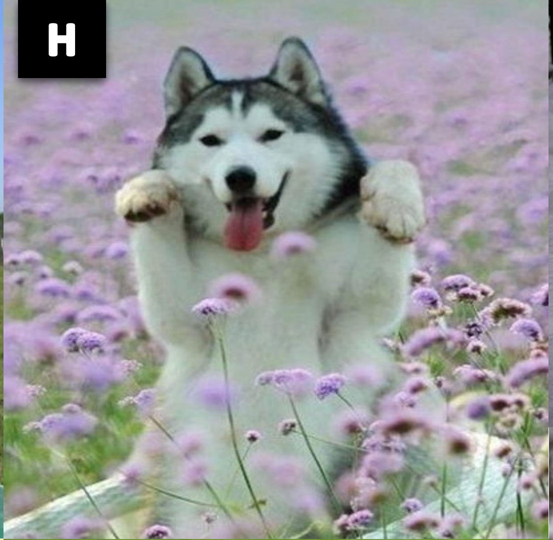
What is “Explainable AI” ?

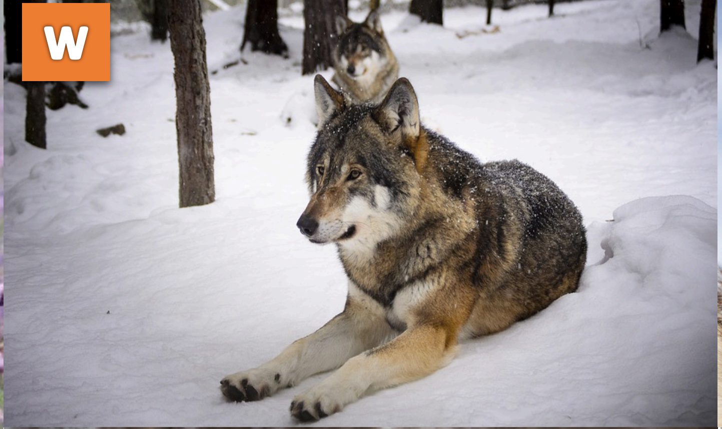


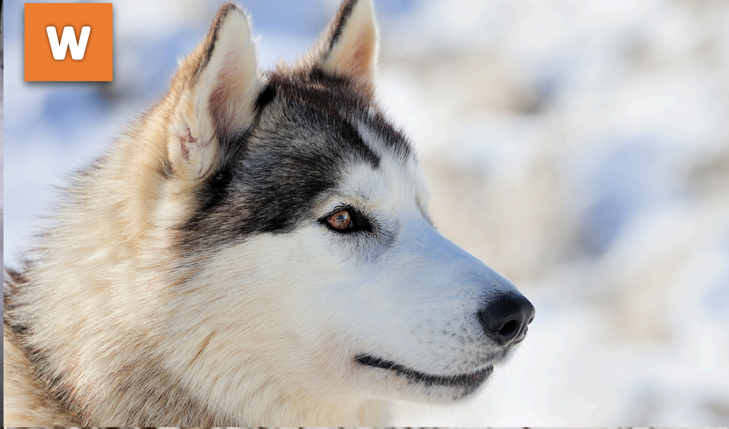
Needs For Interpretable Models

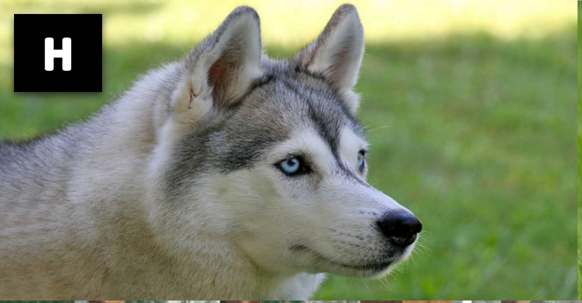
bamahuskies.com

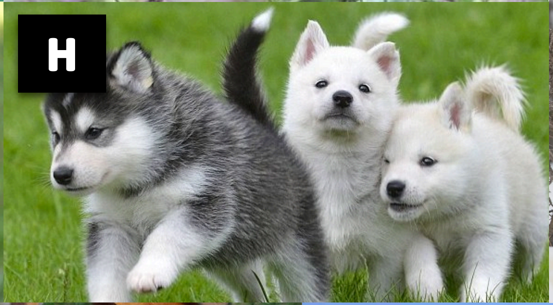

H

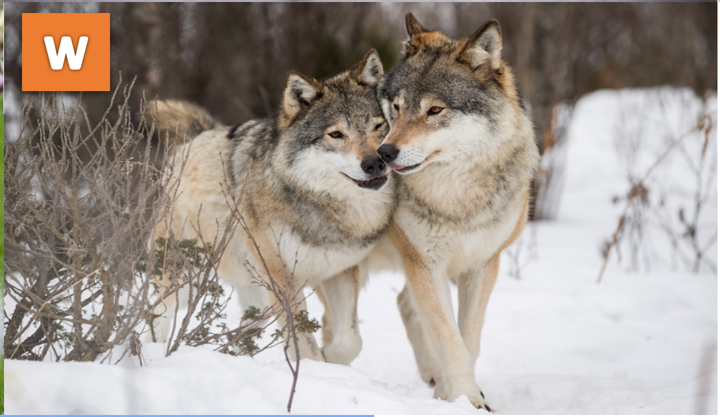

H

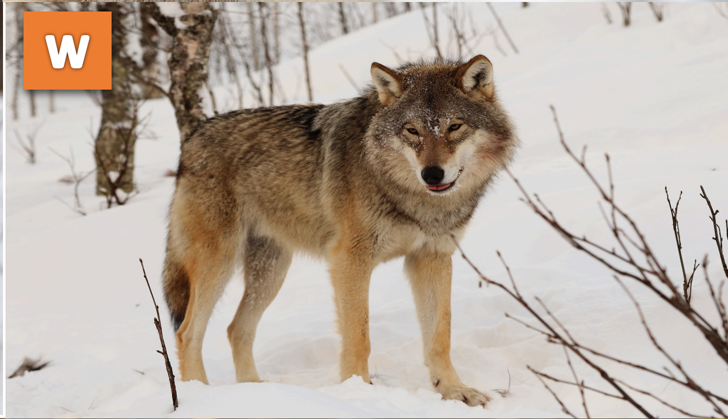

W

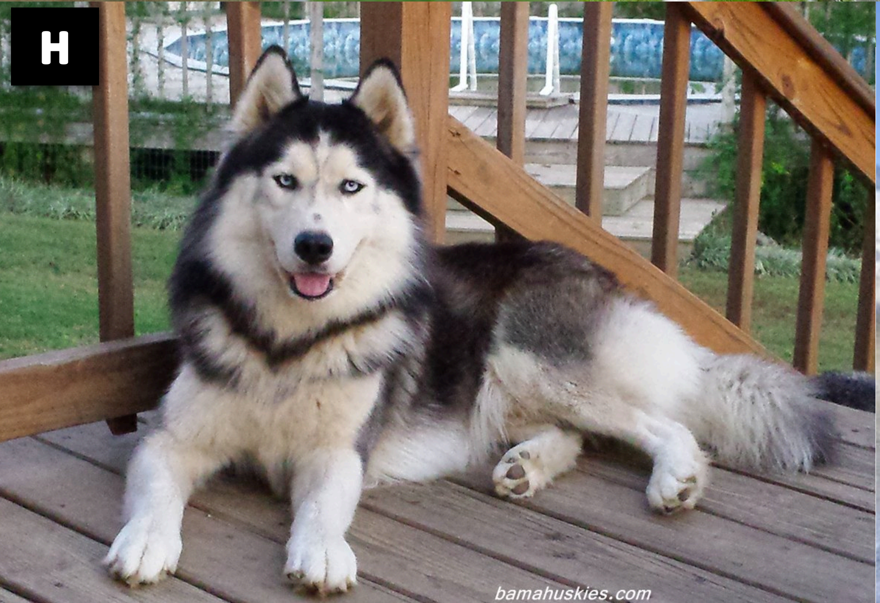

W

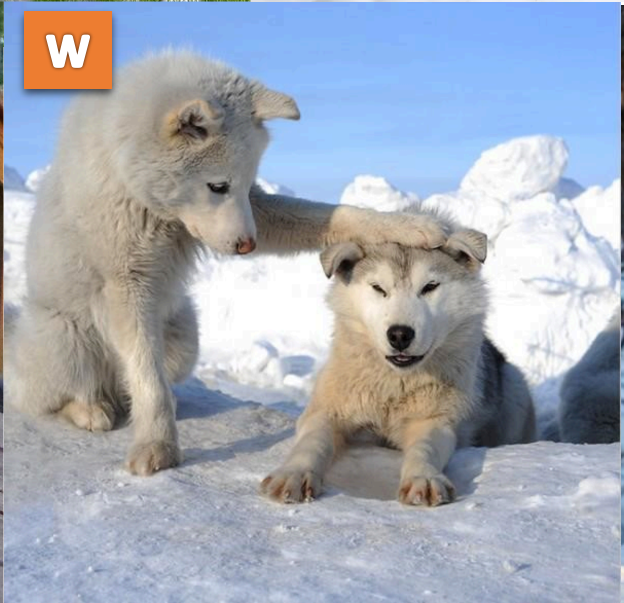

H

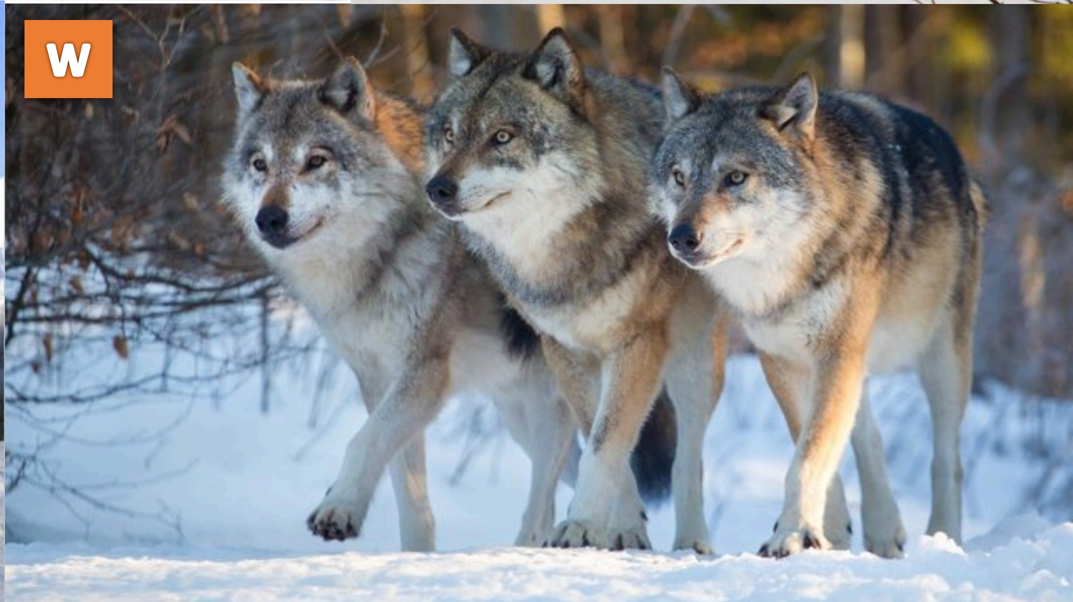

H


W

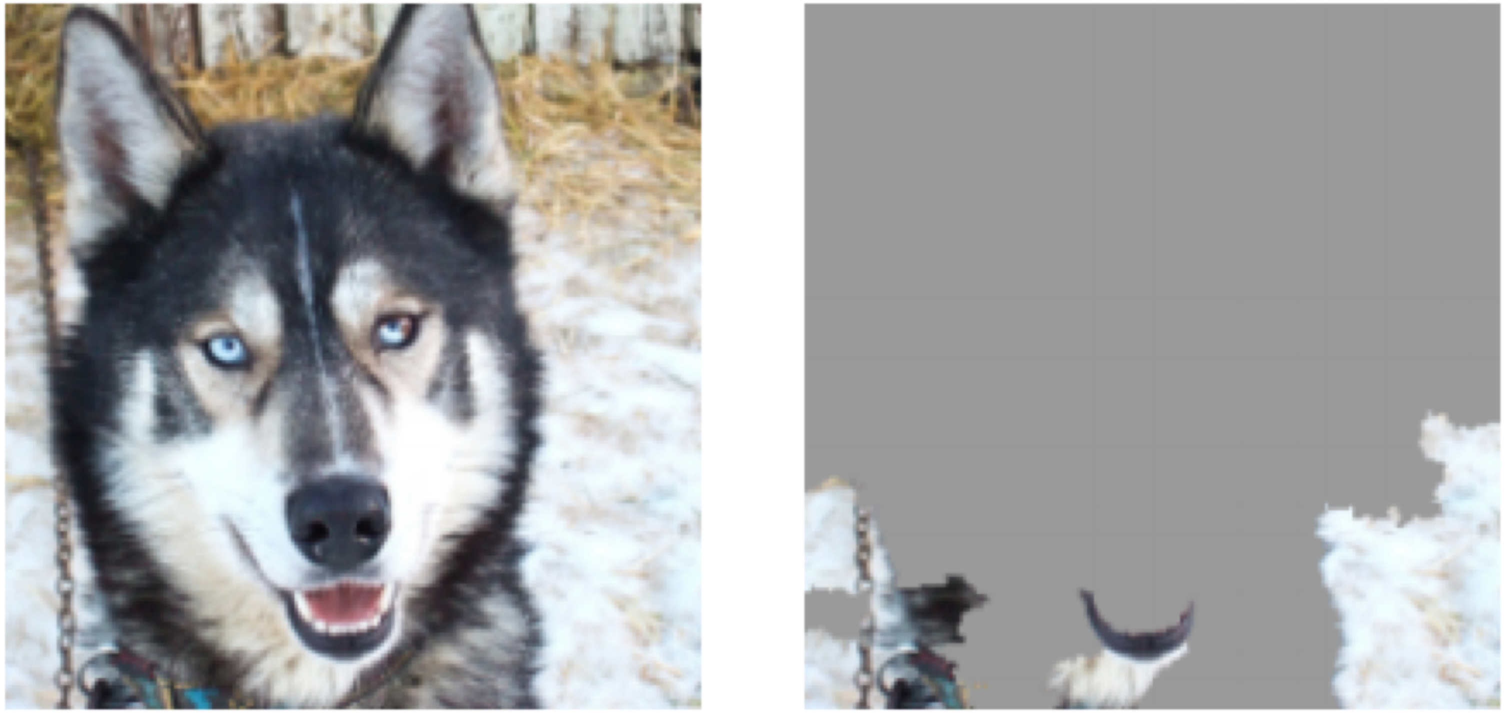

W


H
bamahuskies.com


W


W

The background bias



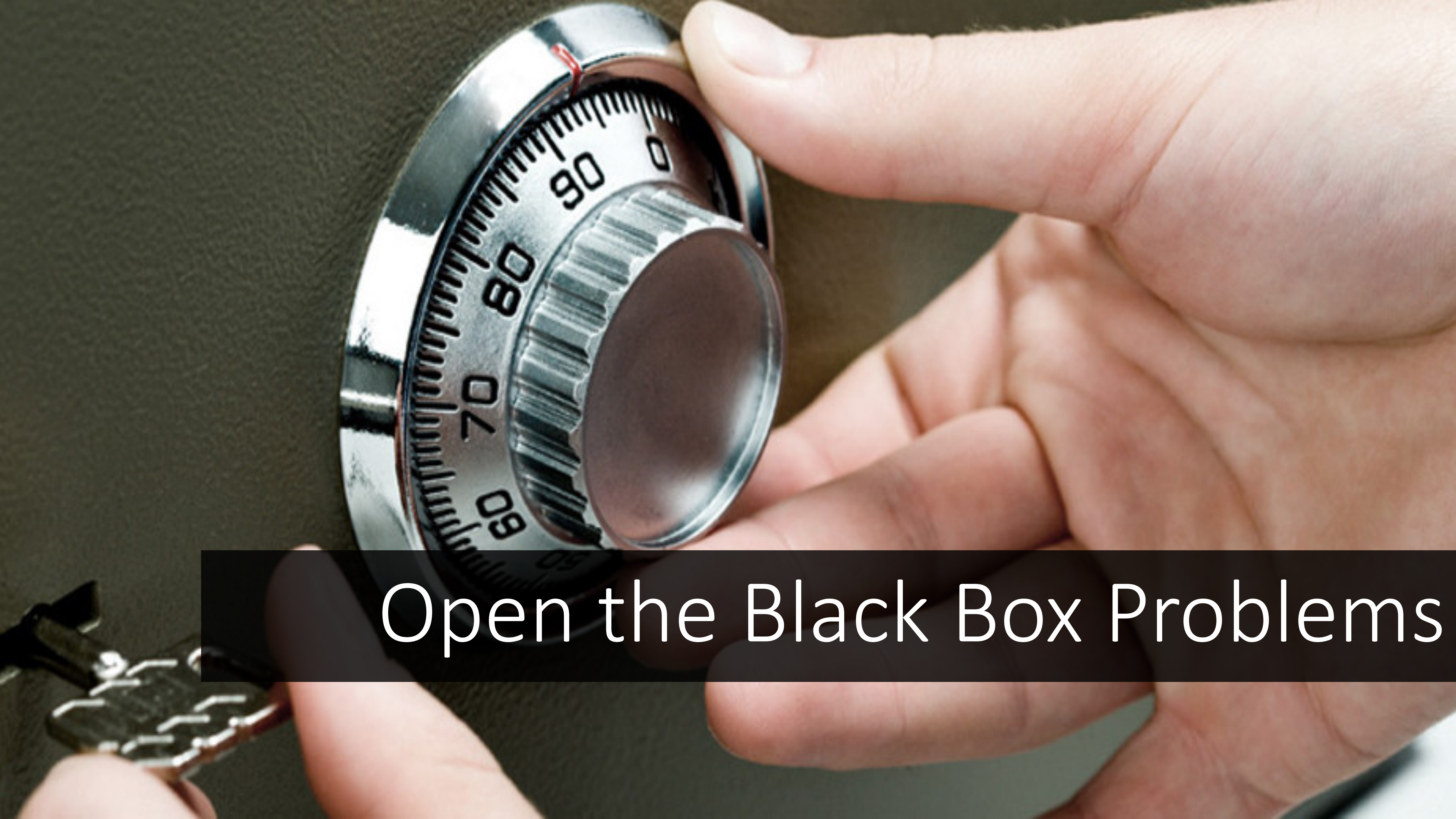
(a) Husky classified as wolf

(b) Explanation

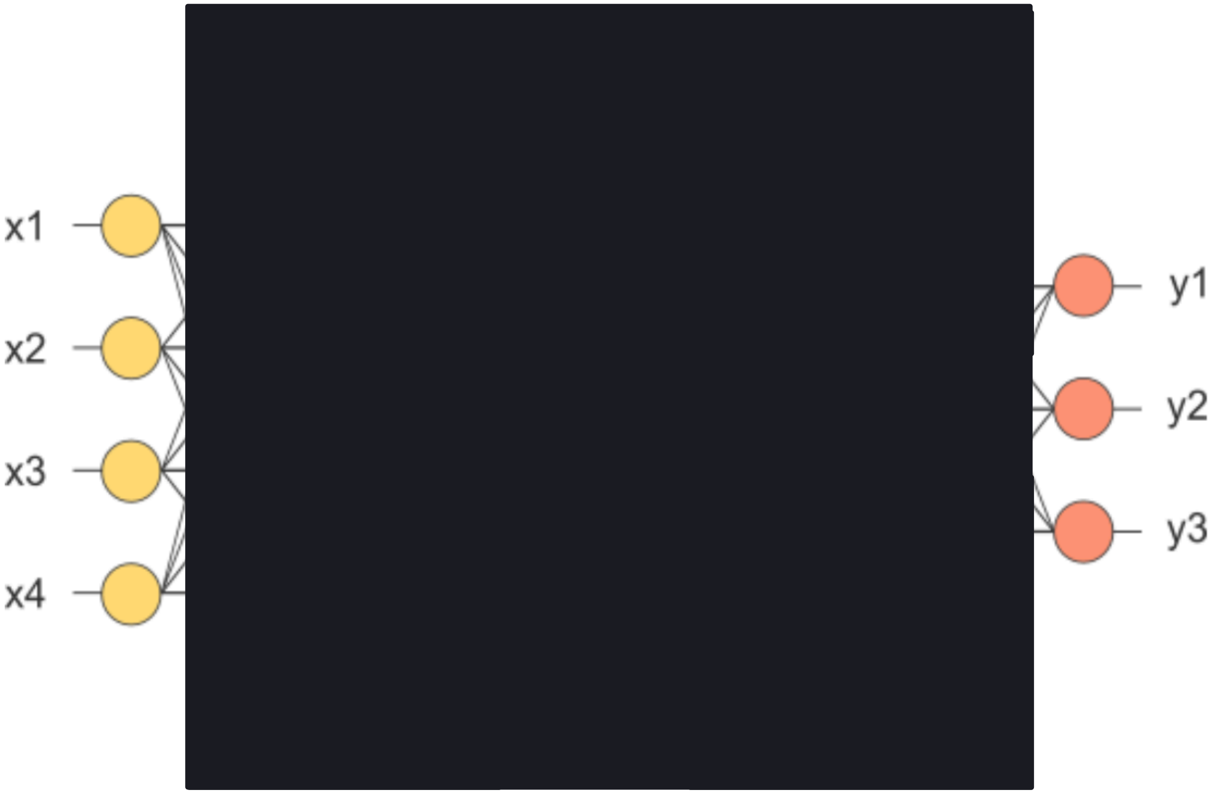
Right of Explanation

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.

Open the Black Box Problems



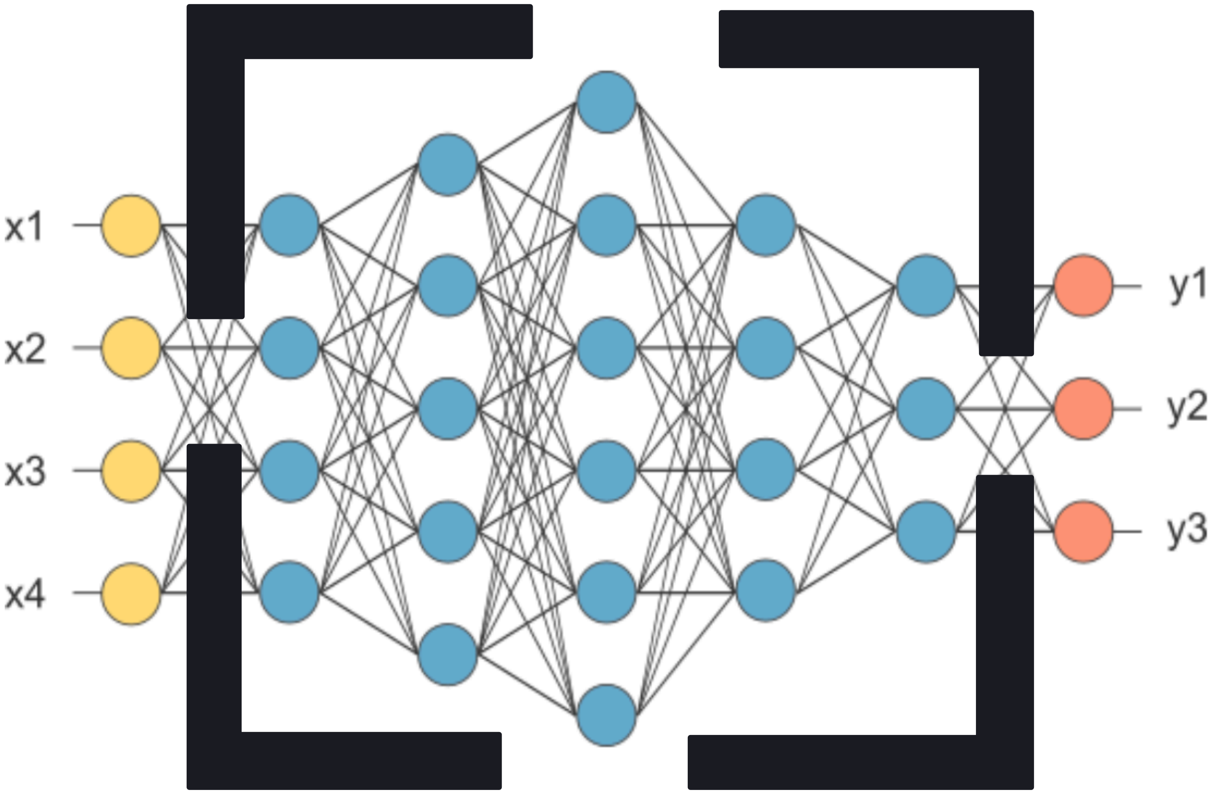
What is a Black Box Model?



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.

What is a Black Box Model?

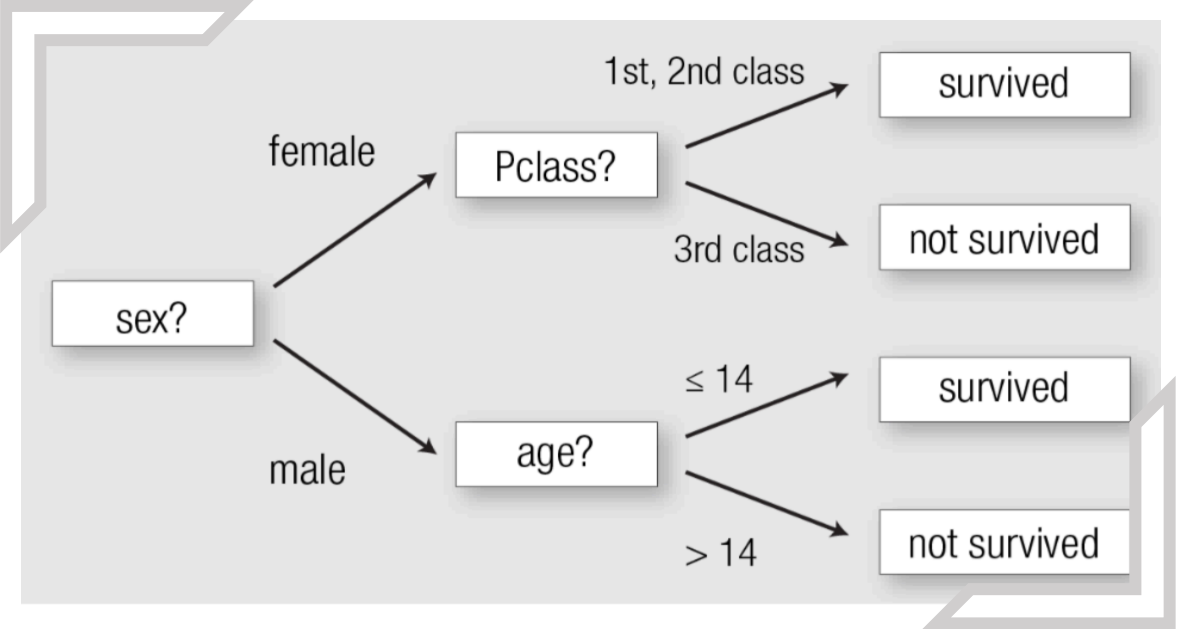


A ***black box*** is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

- 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.

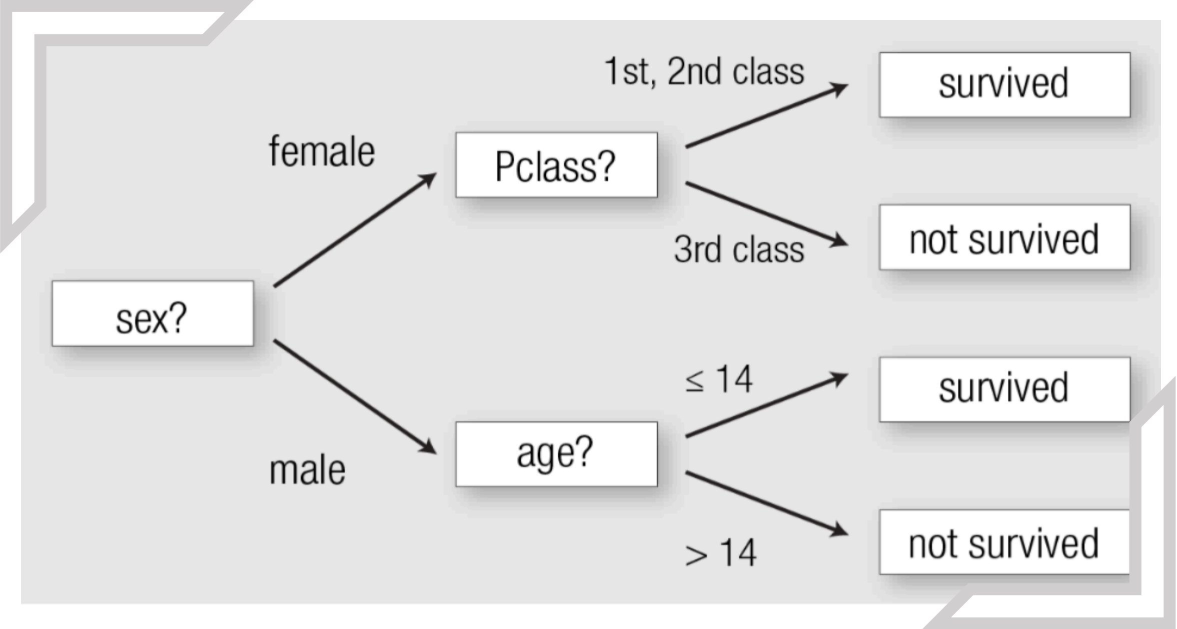
Recognized Interpretable Models

Recognized Interpretable Models

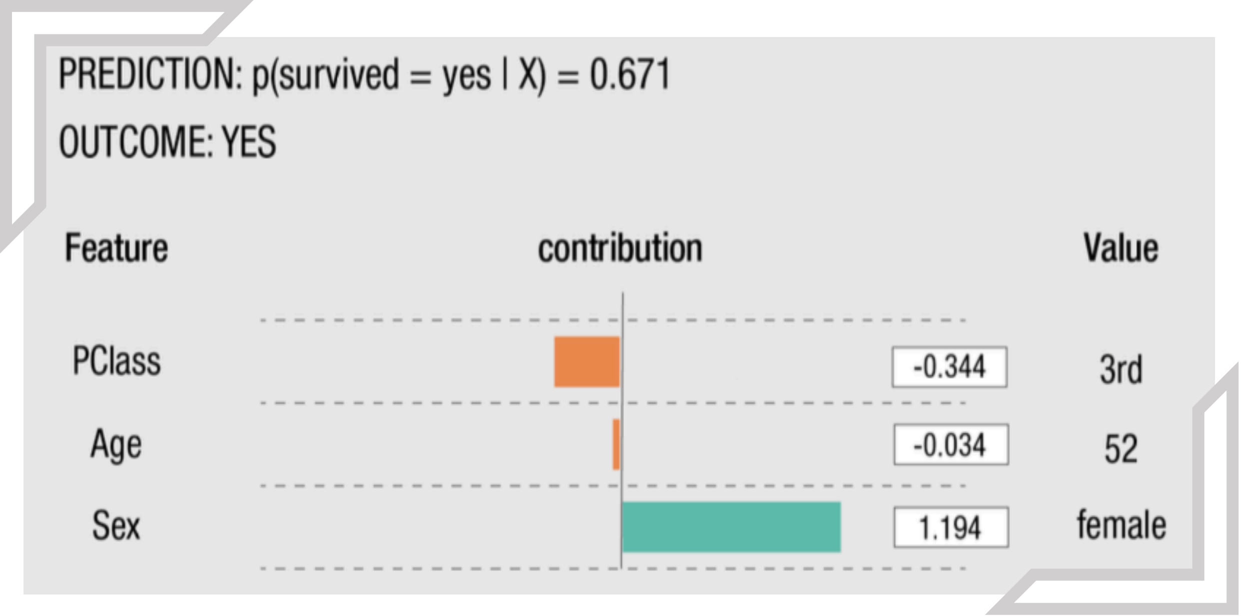


Decision Tree

Recognized Interpretable Models

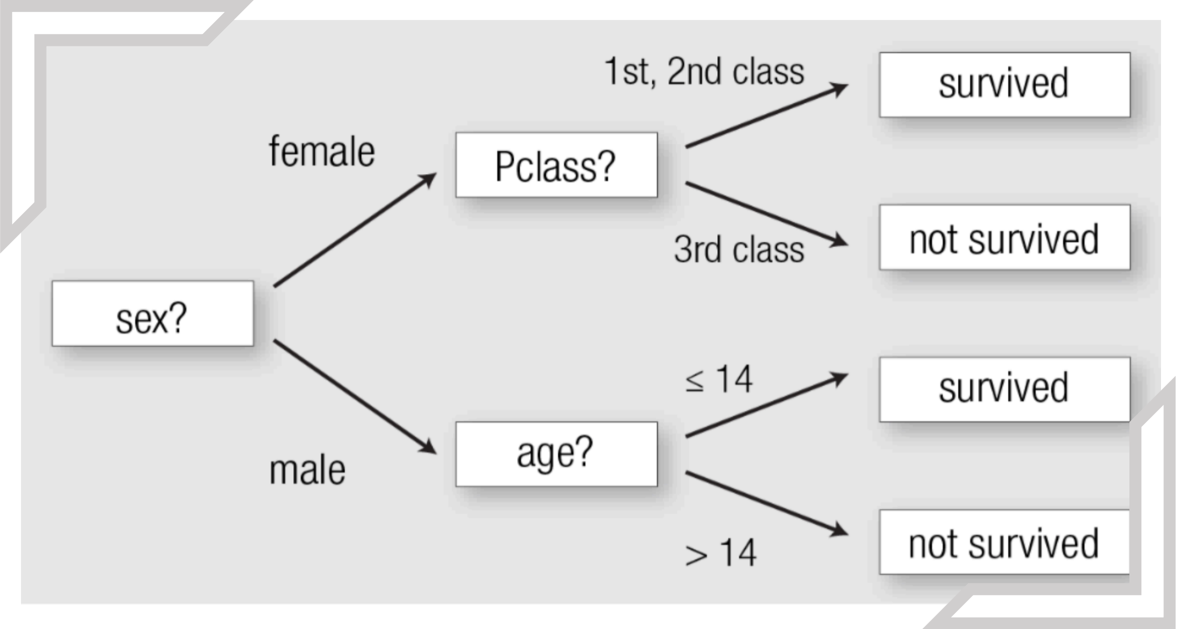


Decision Tree



Linear Model

Recognized Interpretable Models



sex?

female

male

Pclass?

age?

1st, 2nd class

3rd class

≤ 14

> 14

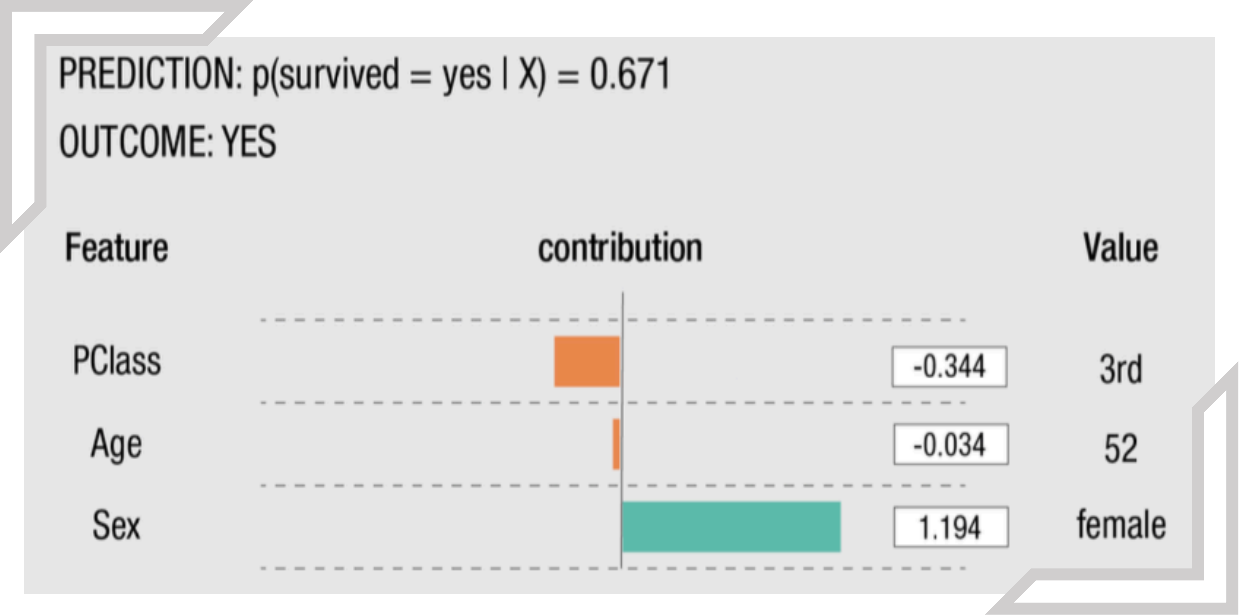
survived

not survived

survived

not survived

Decision Tree

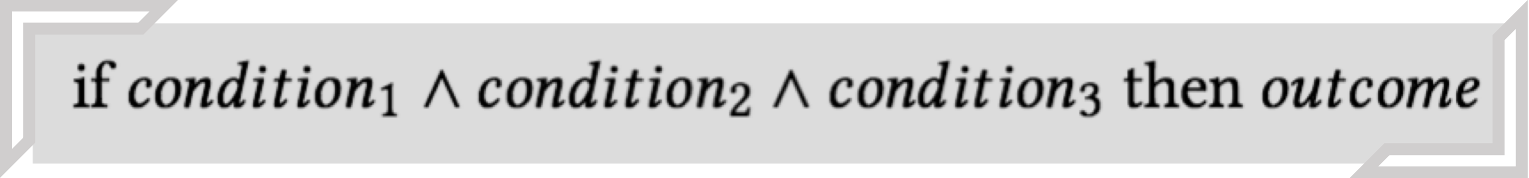


PREDICTION: $p(\text{survived} = \text{yes} \mid X) = 0.671$

OUTCOME: YES

Feature	contribution	Value
PClass	-0.344	3rd
Age	-0.034	52
Sex	1.194	female

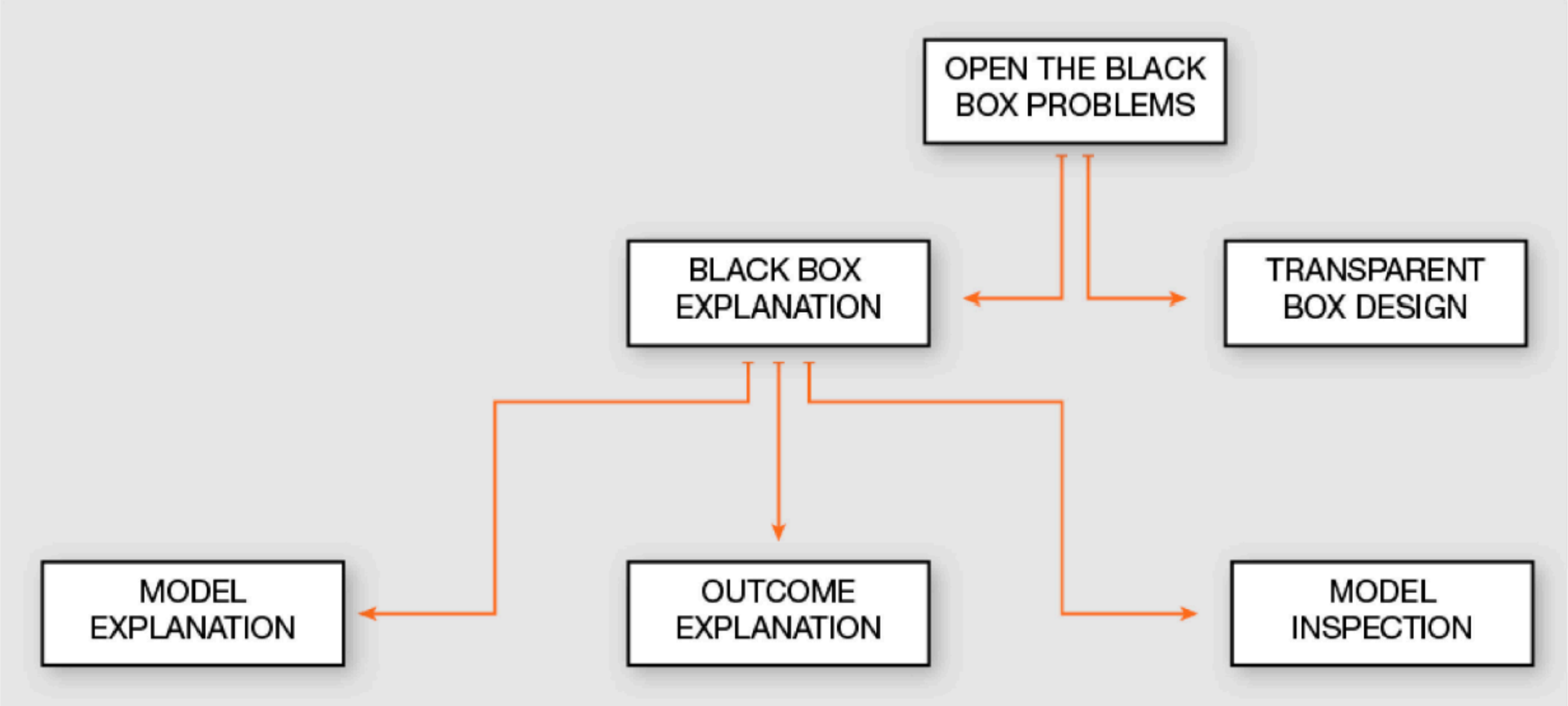
Linear Model



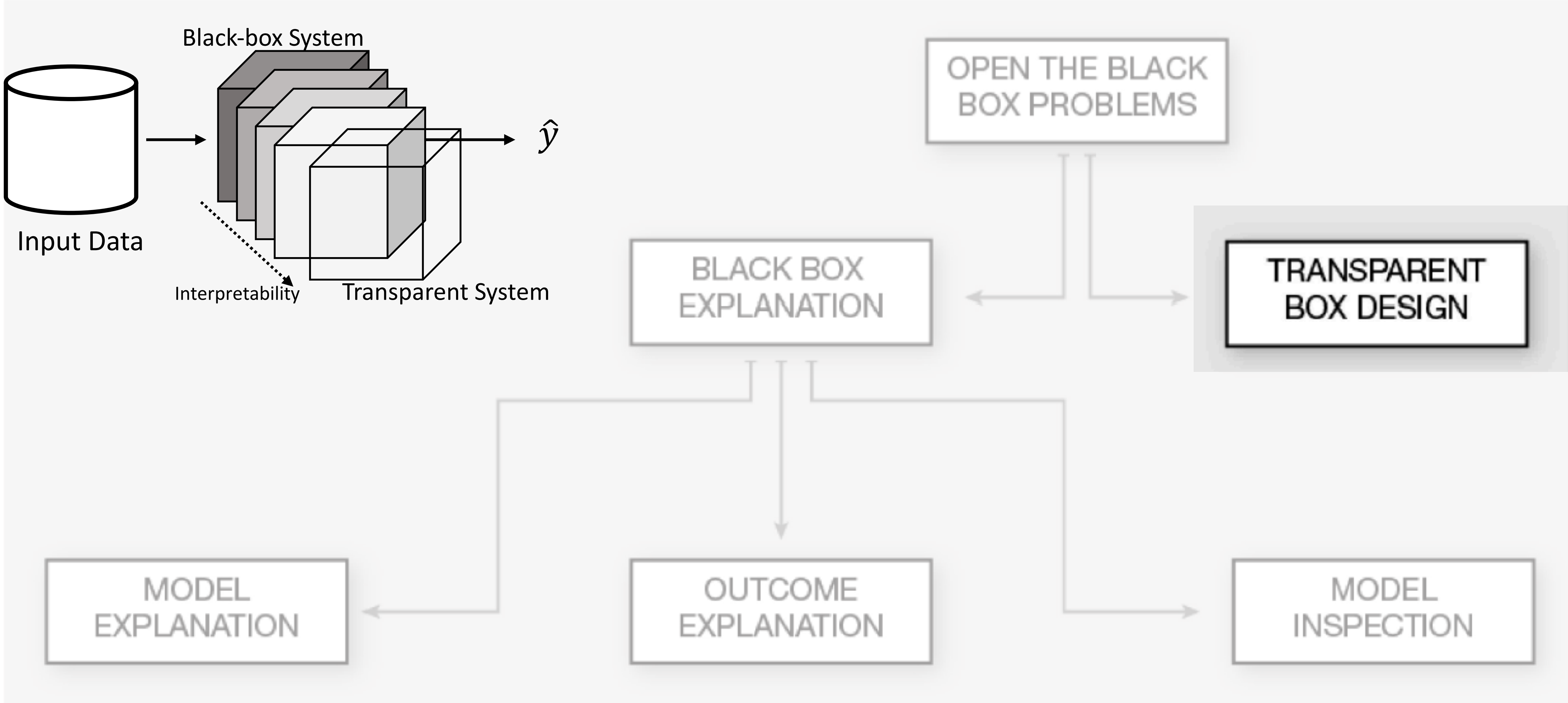
if $\mathit{condition}_1 \land \mathit{condition}_2 \land \mathit{condition}_3$ then $\mathit{outcome}$

Rules

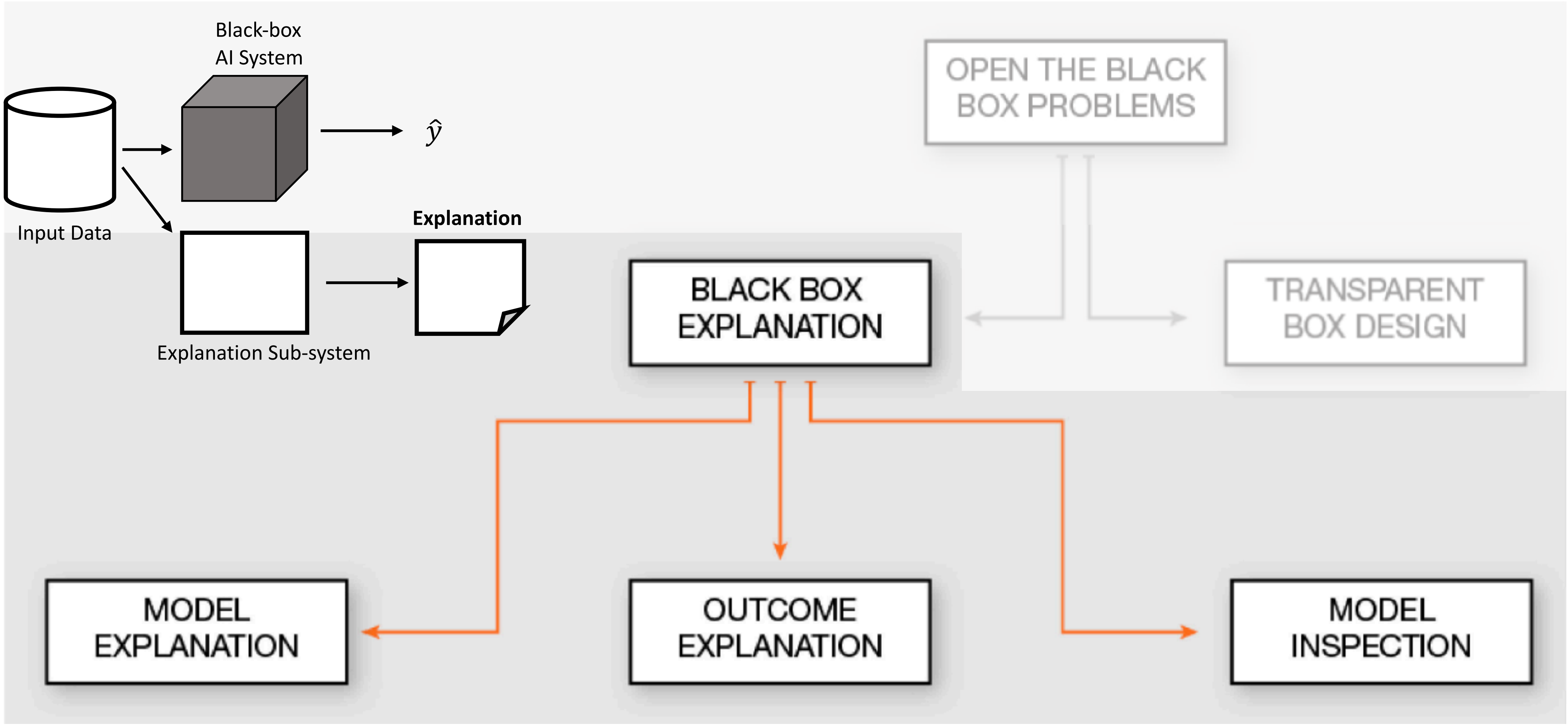
Problems Taxonomy



XbD – eXplanation by Design

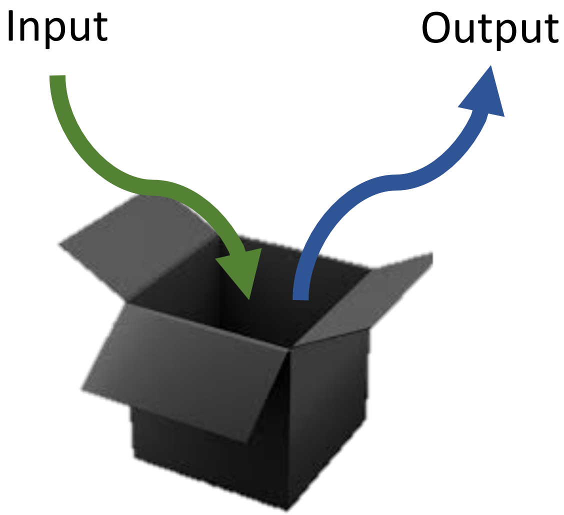


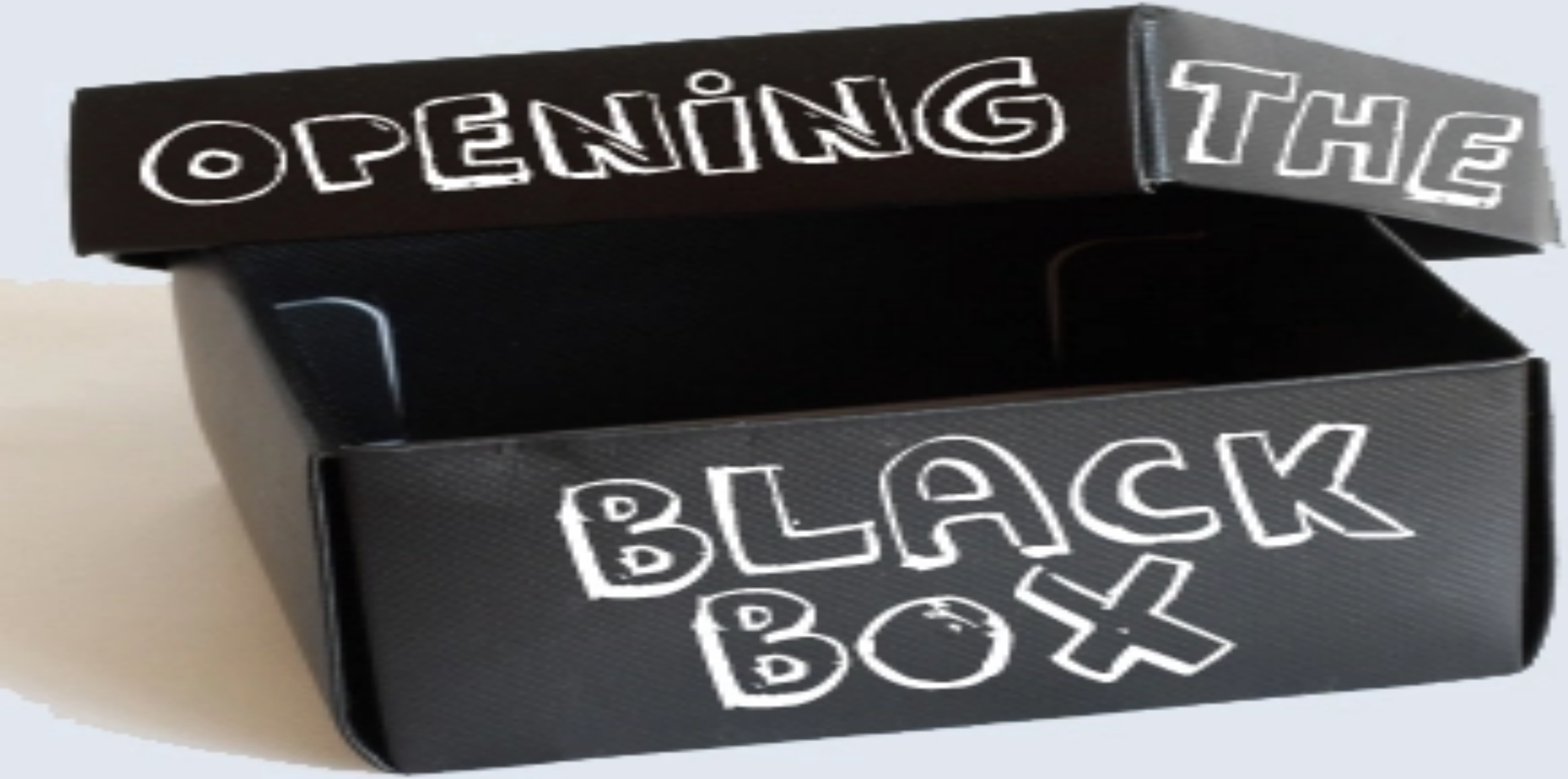
BBX - Black Box eXplanation



How Can We Explain?

- We adopt **reverse engineering**: we can only ***observe*** the ***input*** and ***output*** of the black box.
- Possible actions are:
  - querying/auditing the black box with input records created in a controlled way using ***random perturbations***
  - ***choice*** of a particular interpretable model
- The explanation process can be ***generalizable or not***:
  - Model-Agnostic
  - Model-Specific




OPENING THE
BLACK
BOX

Research Proposals

- ***Local Explanation***
  - for different type of data
  - for **pairwise learning**
  - with **causal reasoning**
  - with **inductive logic programming**
- ***Transparent Design***
  - Data-driven merge of decision trees
  - Prototype-based decision trees for interpretability also in latent space
  - Evolving decision trees in real or latent space with a genetic algorithm
- ***Defining Explanations***
  - What is an explanation? For whom is an explanation?
  - Design of **languages for explanation** context-dependent
  - Design explanation as **human-machine conversation**
- ***Explanation Evaluation*** & design of a benchmarking platform

Thank you!

riccardo.guidotti@isti.cnr.it

SoBigData

AI4EU

HUMANE AI

Track & Know

PRO-RES

XAI

erc

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. (2019). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 93.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).
- Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2019) Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. In Proceedings of ECML-PKDD.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.