NL4XAI

Interactive *Natural Language* Technology for eXplainable Artificial Intelligence

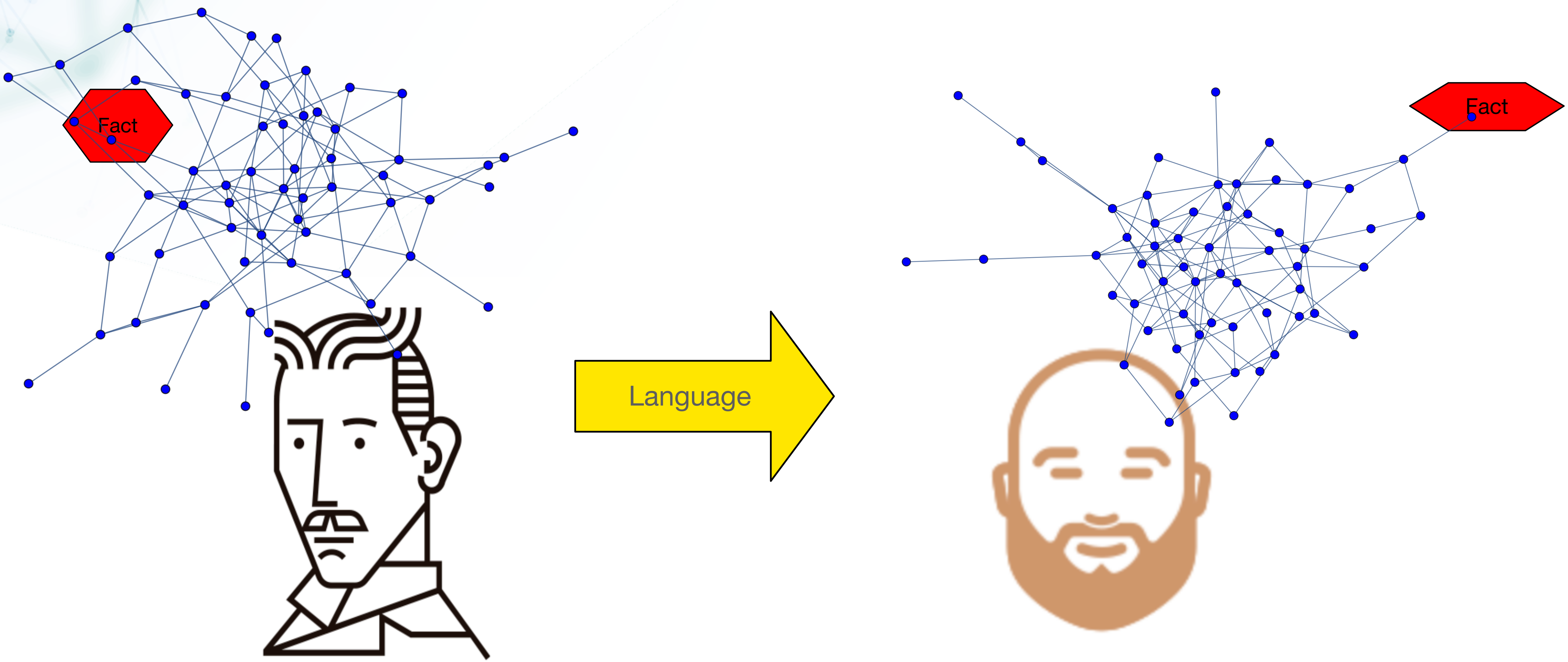
Navigating the Landscape of Explainable AI

Ettore Mariotti - 15 June 2023 - Barcelona Supercomputing Center

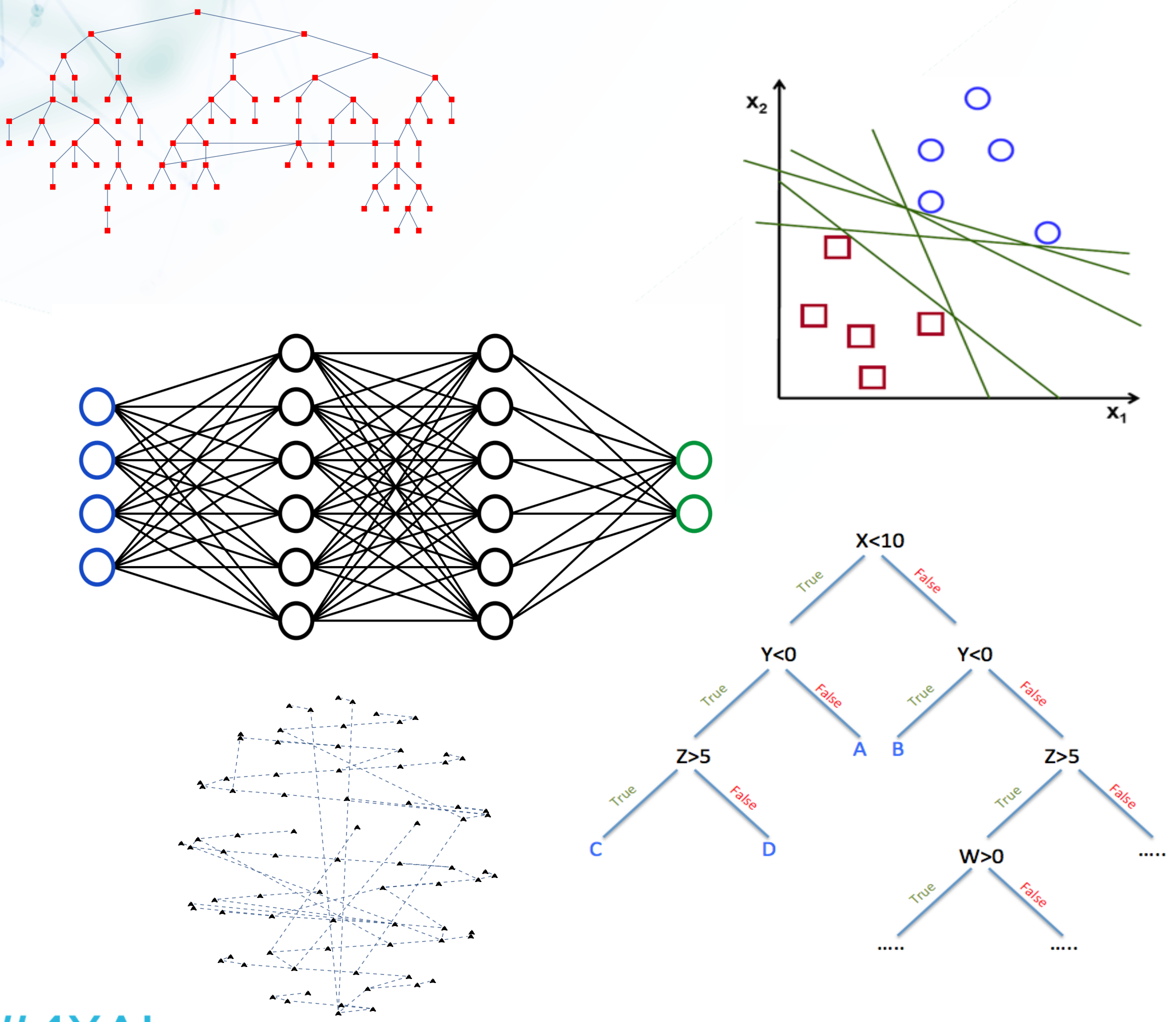
How can we understand AI?

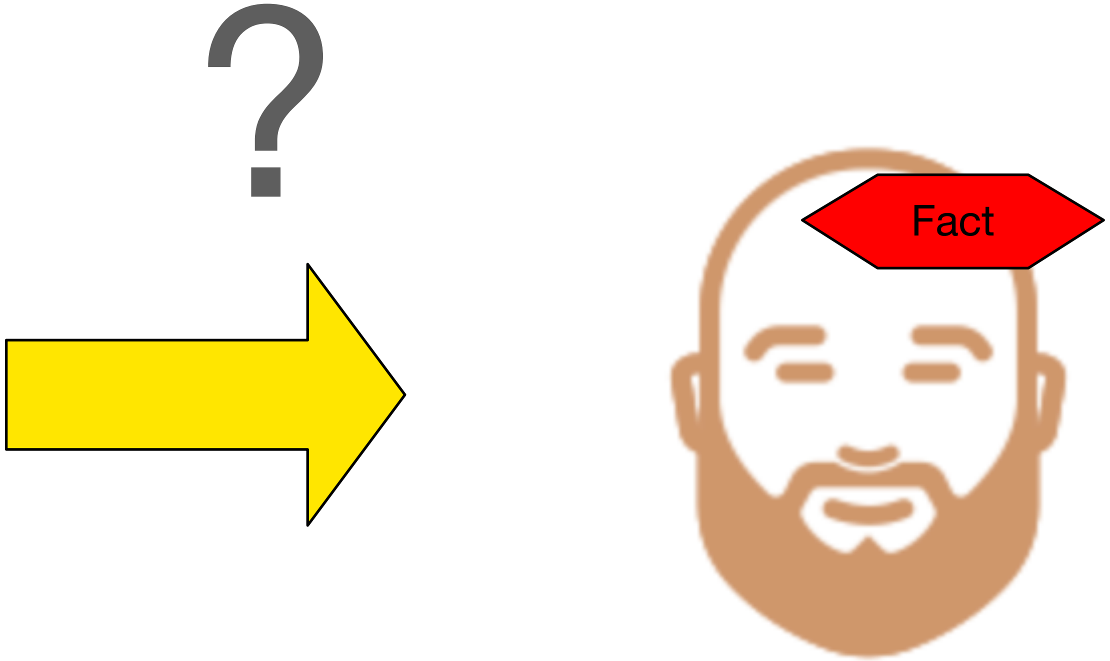
How can we understand AI?

What is an explanation



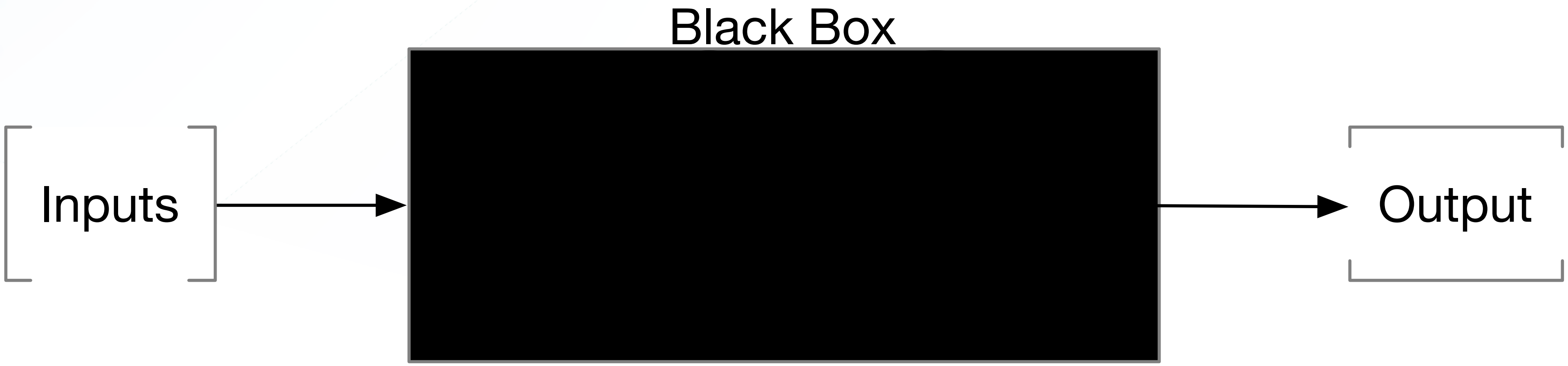
What is an explanation



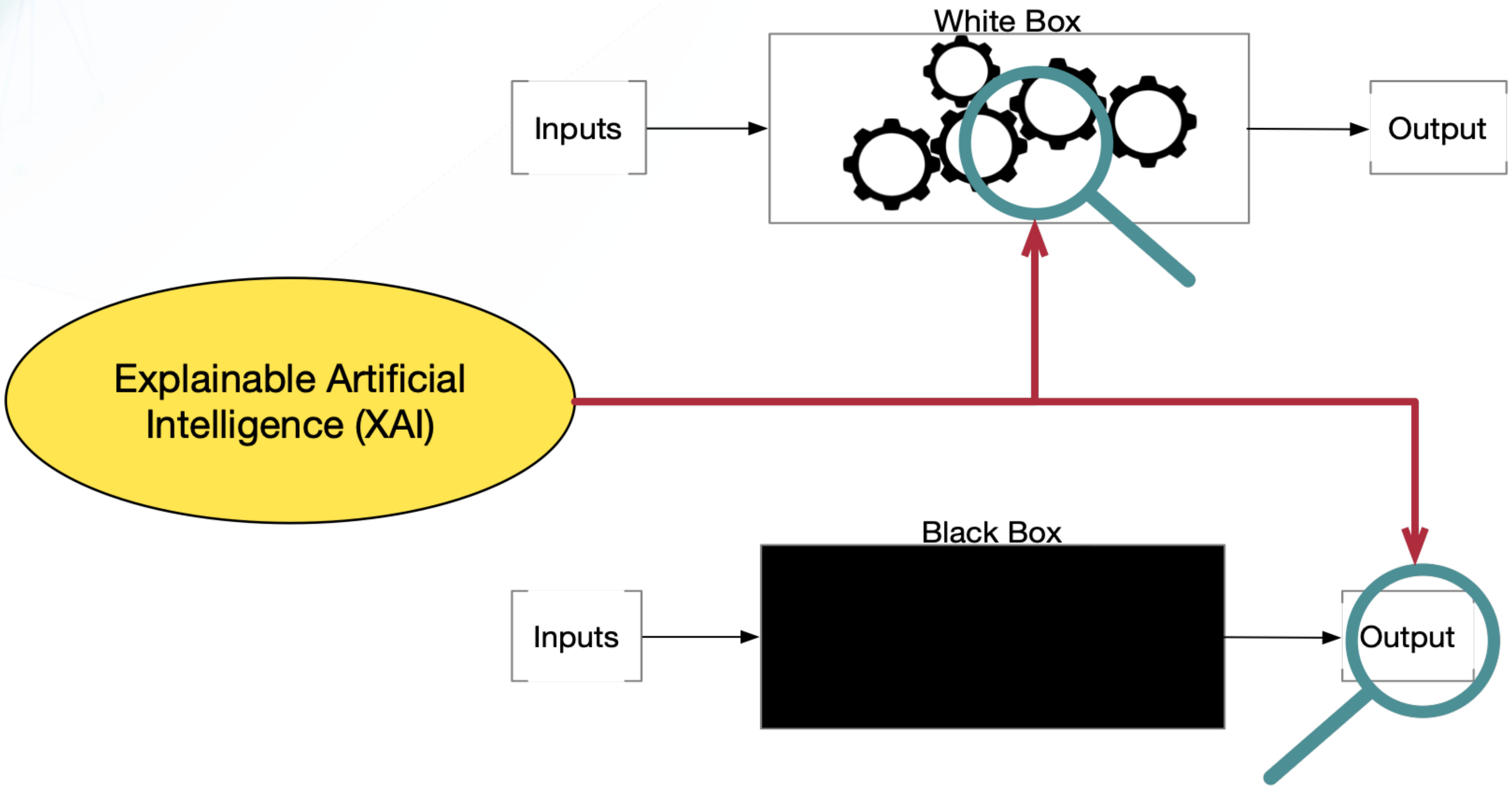


Explaining the black box

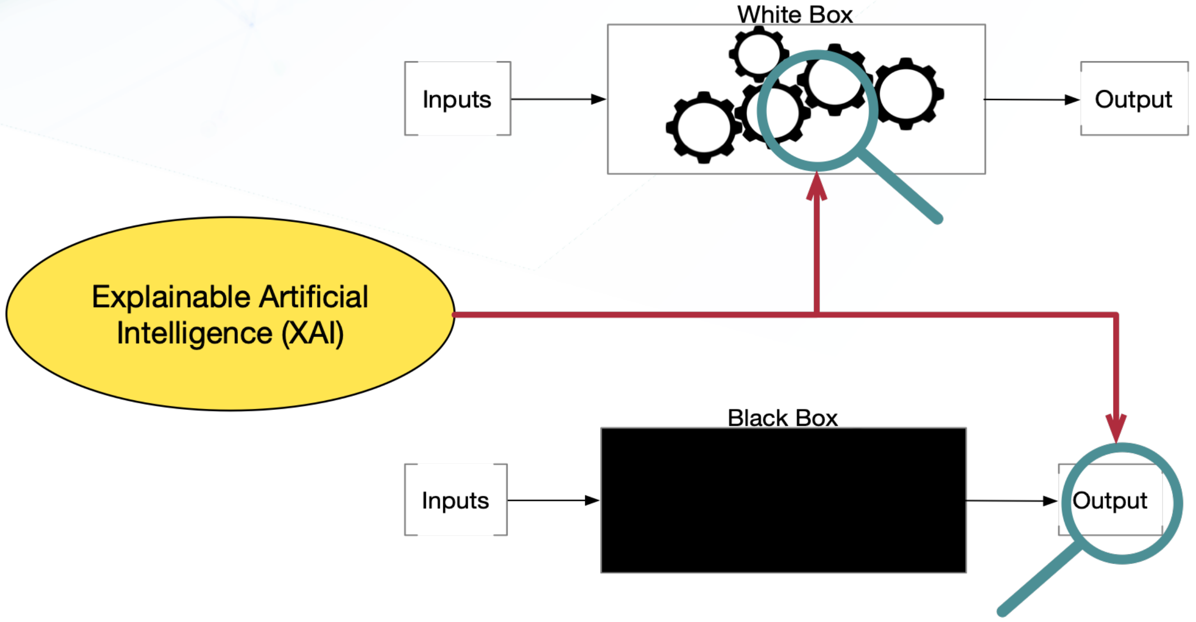
What is a black box

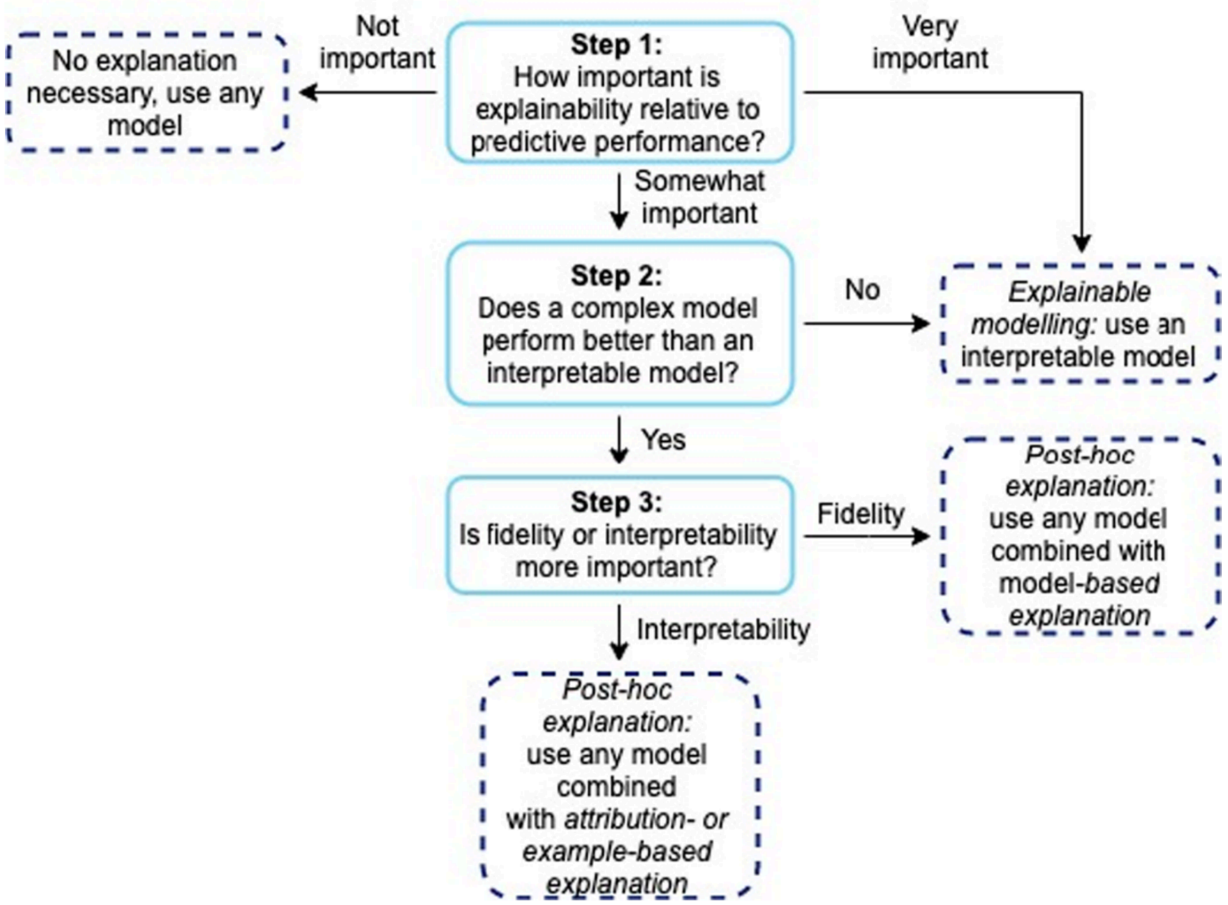


How/when should we use XAI?

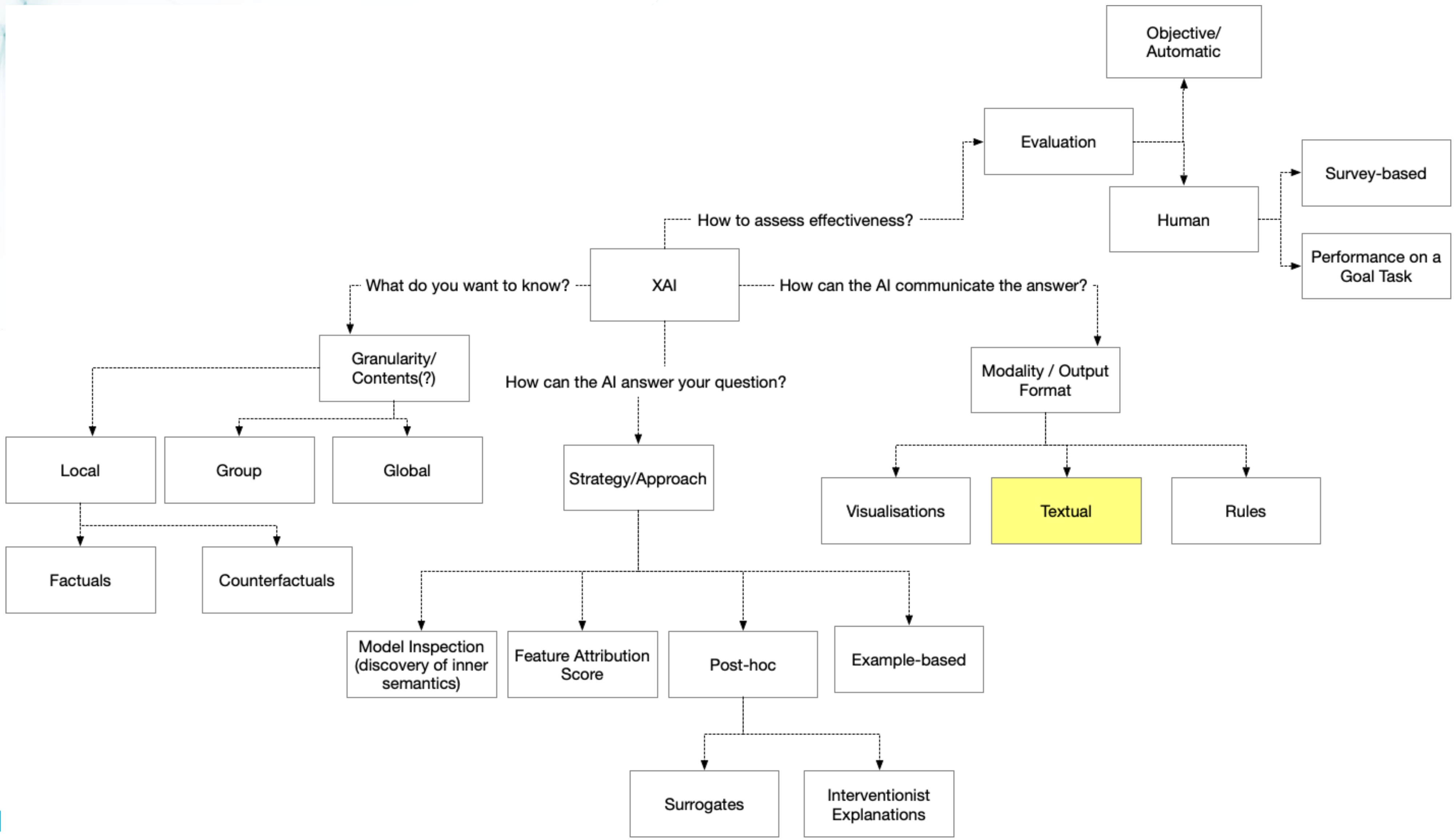


When/how should we use XAI?



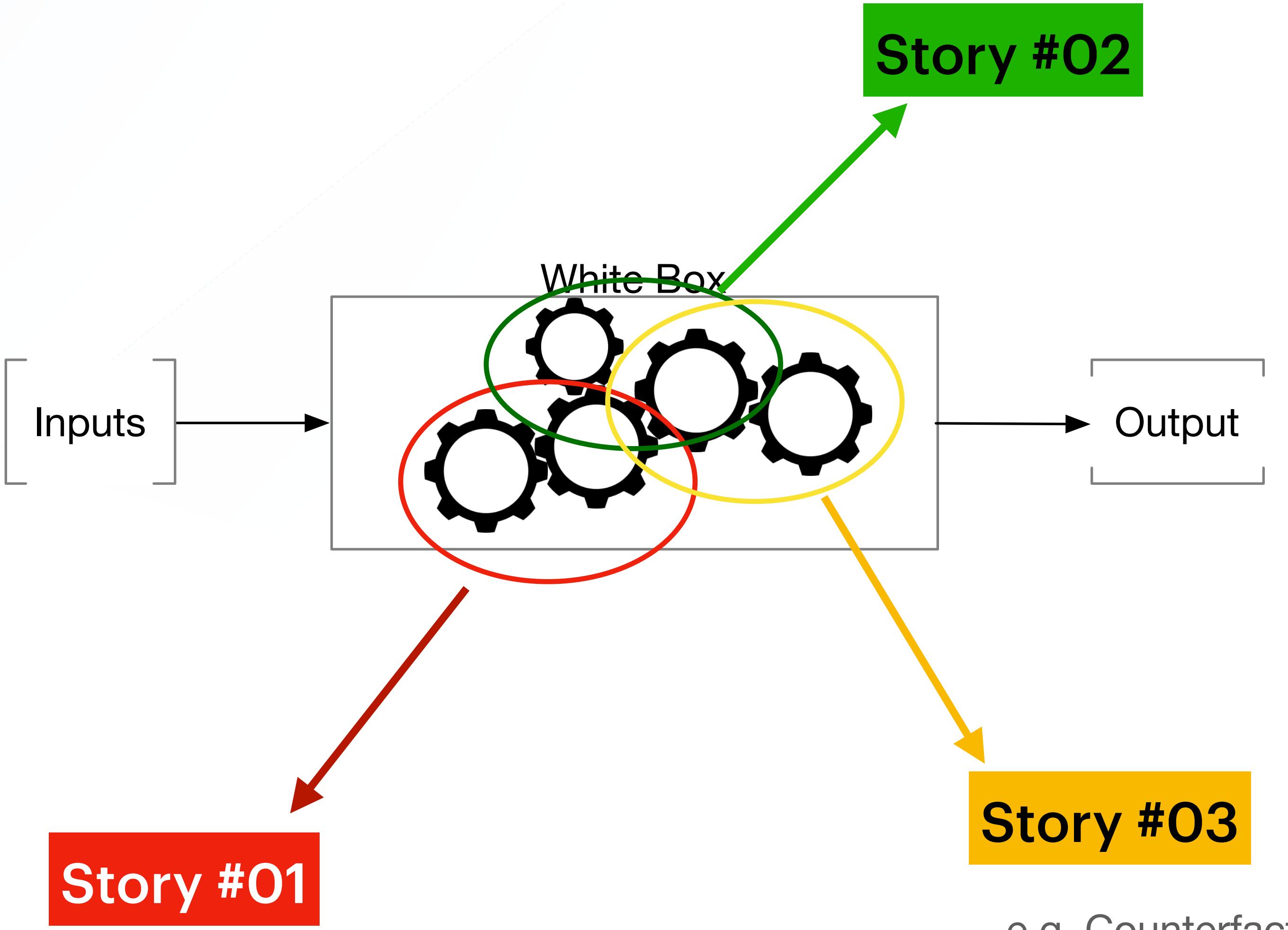


Bird-eye view of XAI field



Going Interpretable (white box)

e.g. Global Behaviour Explanation



White box examples

- Rule systems
- Linear models
- Trees
- GAMs

e.g. Counterfactual Explanation

e.g. Factual Explanation

Linear models and beyond

NL4XAI

Generic Formula $y = f(x_1, x_2, ..., x_p)$

Linear Model $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p$

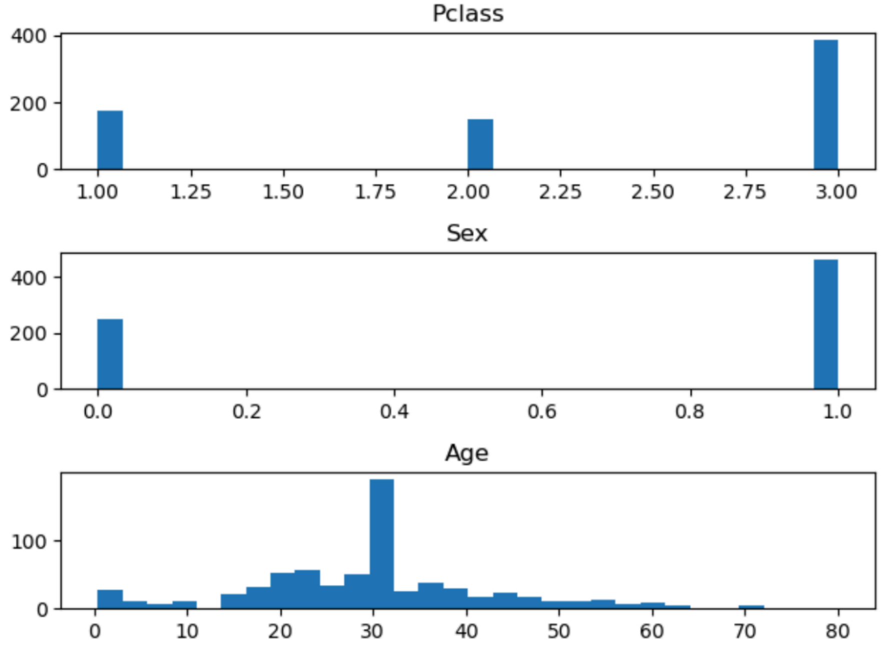
Generalised Additive Model $y = \alpha_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p)$

Example on a simple dataset

Titanic Dataset

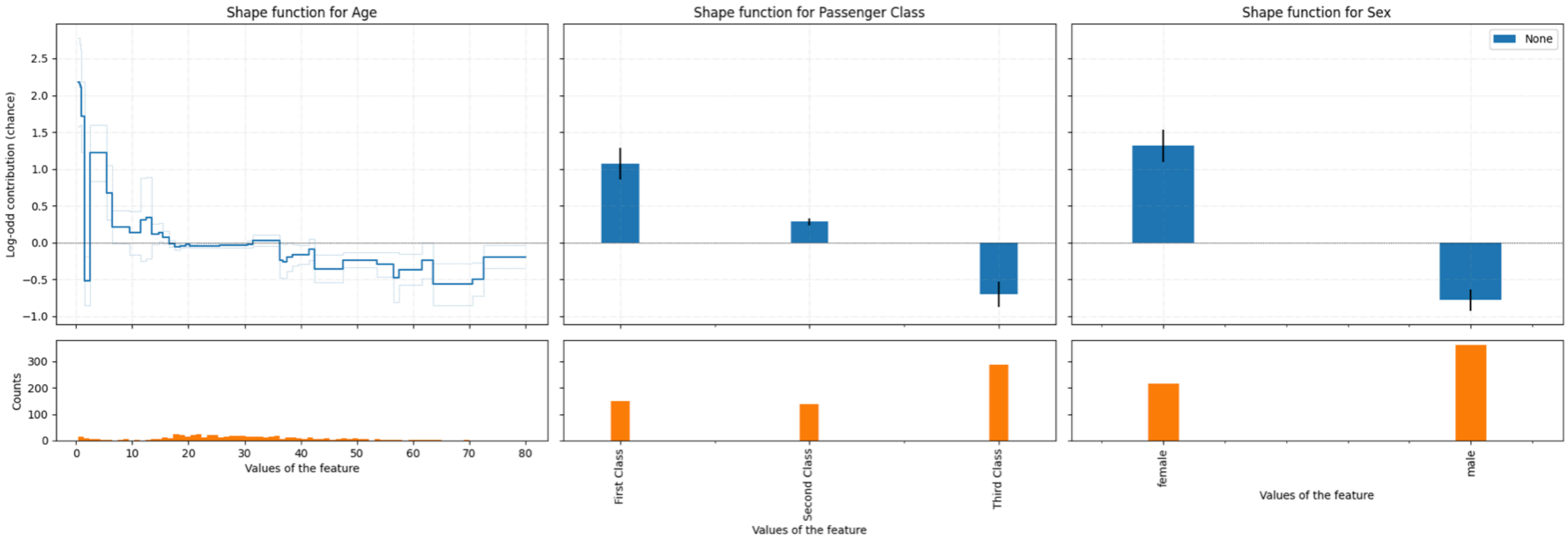
Target prediction: Will the passenger survive?

	Pclass	Sex	Age
0	2.00000	0.00000	29.00000
1	3.00000	1.00000	29.69591
2	1.00000	0.00000	35.00000
3	2.00000	0.00000	28.00000
4	3.00000	1.00000	34.00000
5	3.00000	1.00000	29.69591
6	2.00000	1.00000	29.00000
7	2.00000	1.00000	29.69591
8	2.00000	0.00000	40.00000
9	1.00000	0.00000	39.00000
10	1.00000	0.00000	18.00000
11	1.00000	0.00000	29.69591
12	3.00000	1.00000	29.69591
13	3.00000	1.00000	29.69591
14	3.00000	0.00000	28.00000
15	3.00000	1.00000	9.00000
16	1.00000	1.00000	45.00000
17	1.00000	1.00000	29.69591



Global behaviour

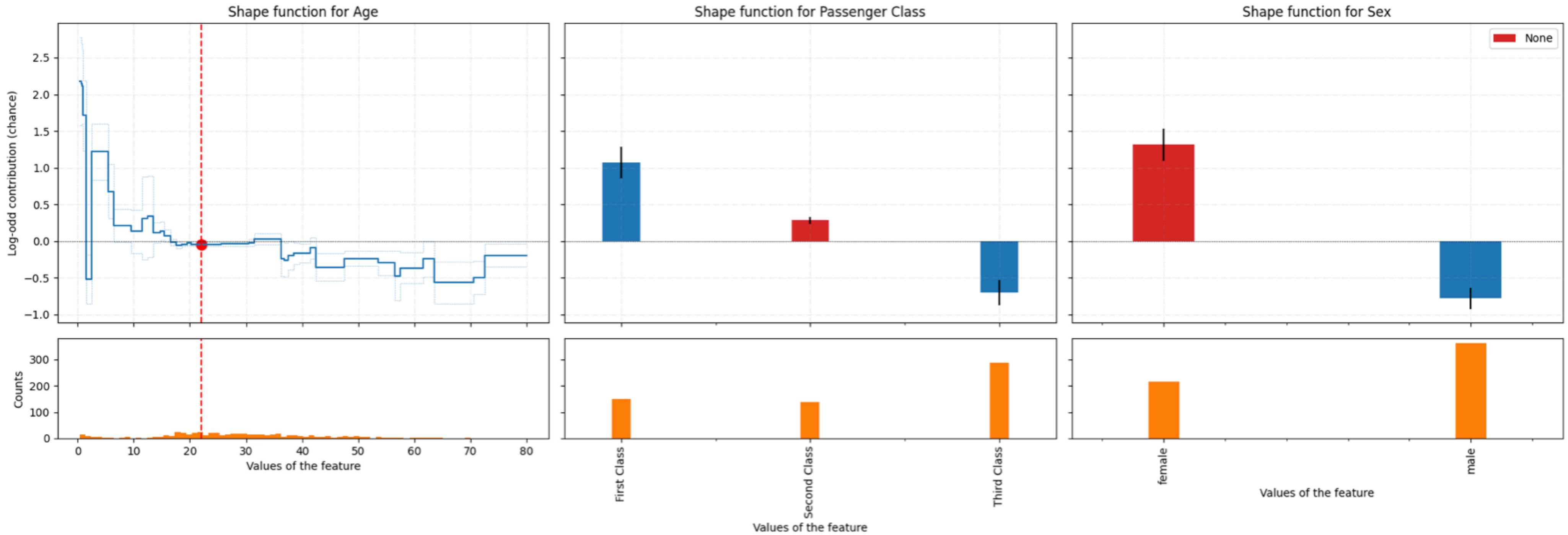
$$g(y) = -0.47 + f_{Age}(x_{Age}) + f_{PClass}(x_{PClass}) + f_{Sex}(x_{Sex})$$



Factual Local Explanation (single instance)

24 years old — Second class — Woman

$$g(y) = -0.47 + f_{Age}(x_{Age}) + f_{PClass}(x_{PClass}) + f_{Sex}(x_{Sex})$$

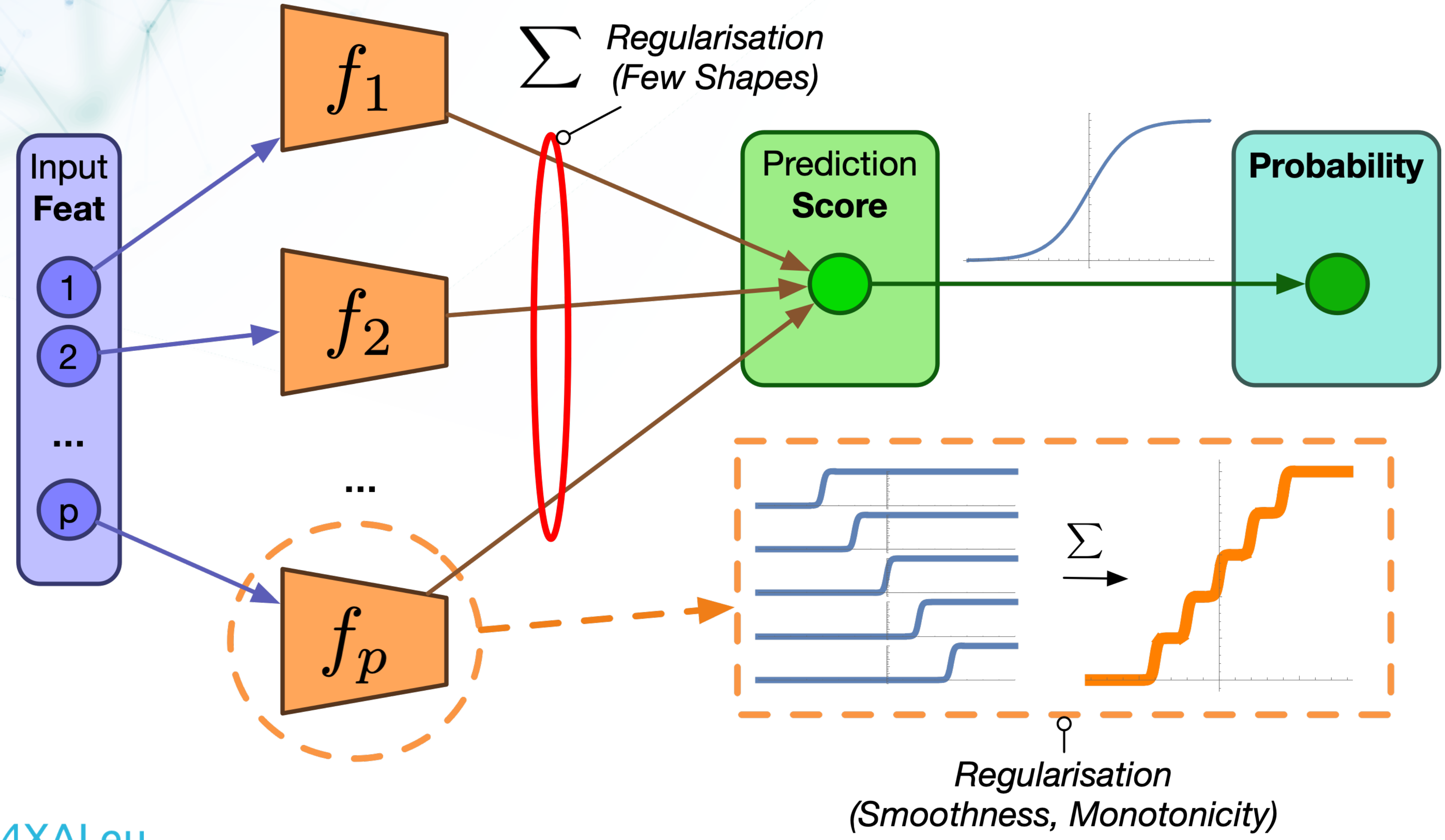


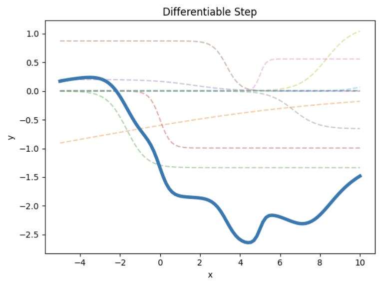
This instance is classified as Survived with probability 74.72%. (logodds 1.1)

Exploring the balance between interpretability and performance with carefully designed constrainable neural additive models, Mariotti et al, 2023

https://doi.org/10.1016/j.inffus.2023.101882

GAM and NAM and CNAM



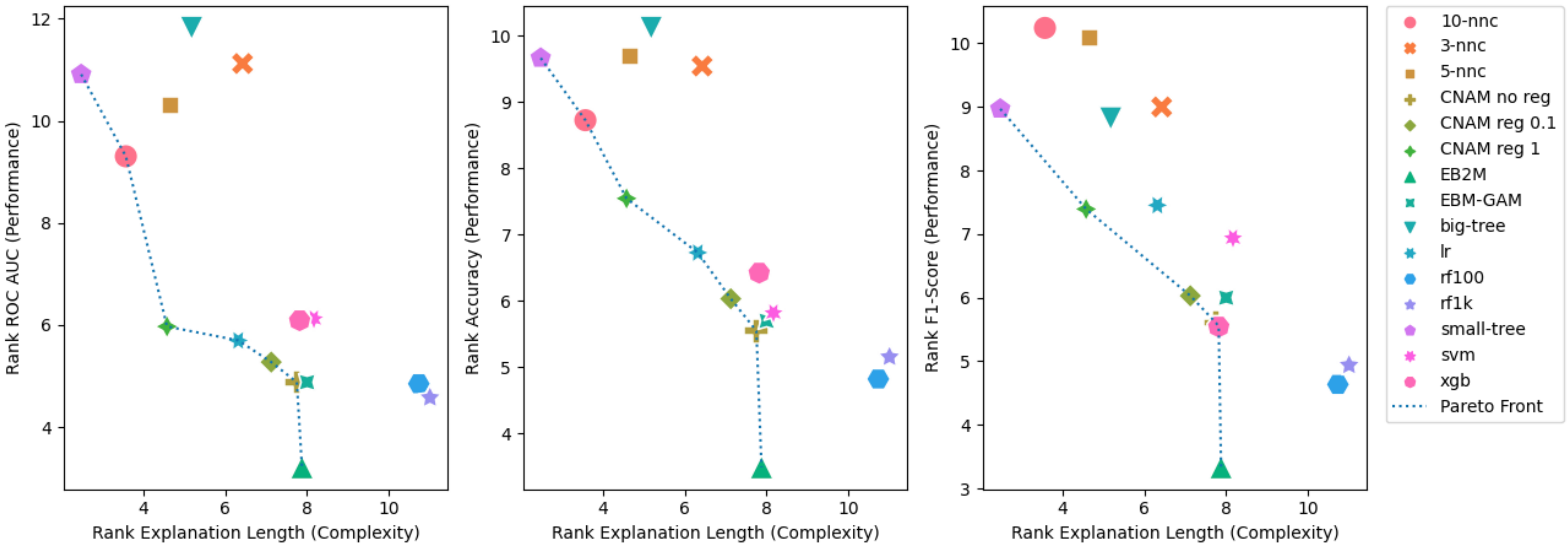


Exploring the balance between interpretability and performance with carefully designed constrainable neural additive models, Mariotti et al, 2023

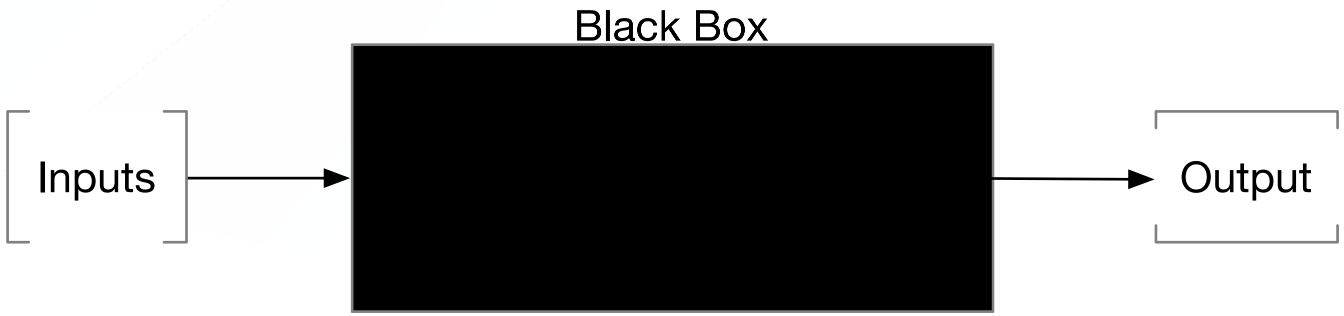
https://doi.org/10.1016/j.inffus.2023.101882

CNAM Performances

Performances/Complexity Tradeoff averaged over 33 datasets

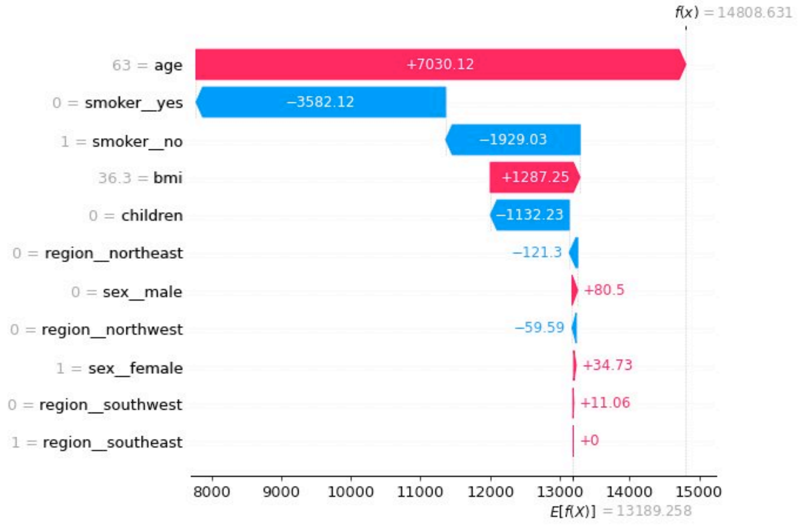


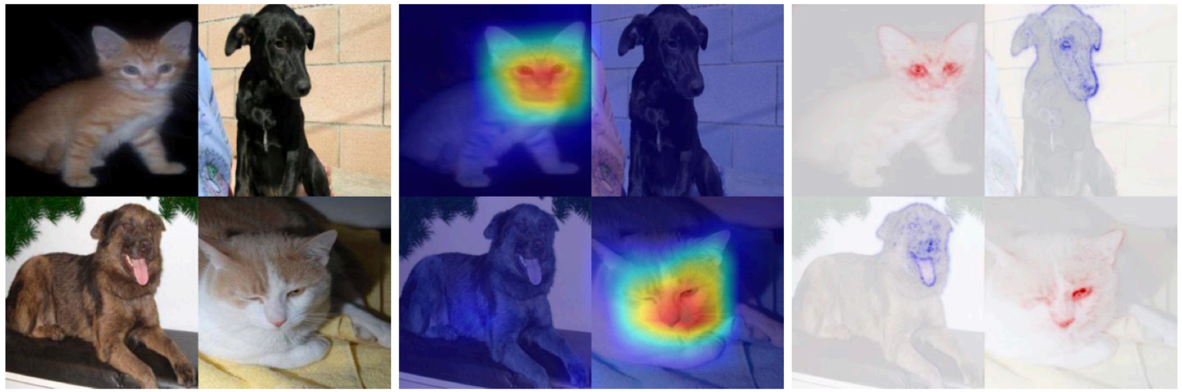
Going post-hoc for explaining black box



Feature Attribution Techniques

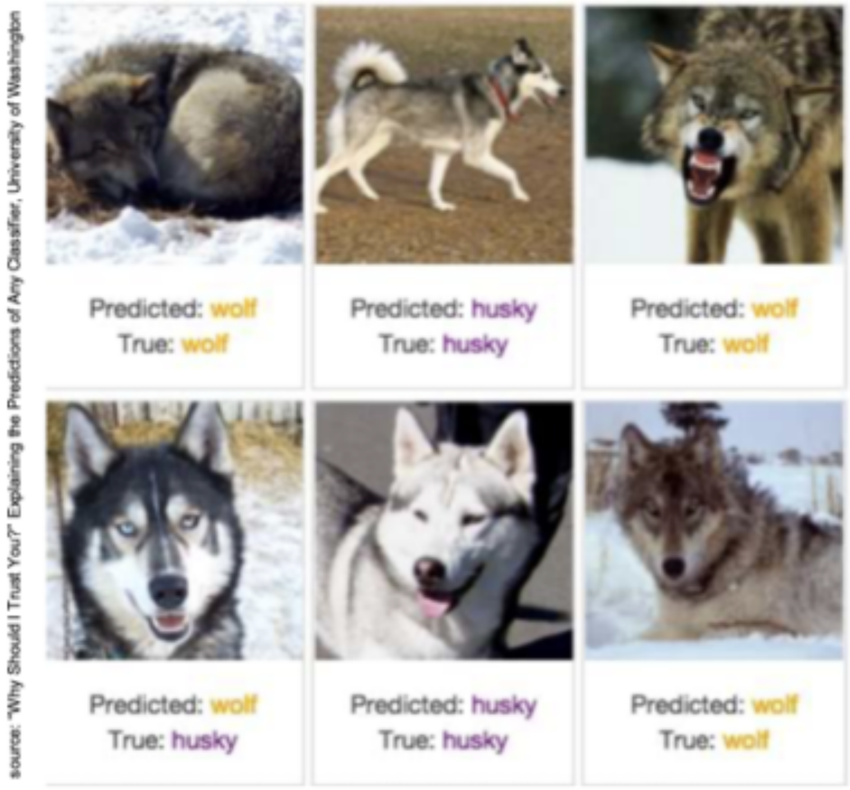
- Answering the question:
- *What was the **impact** that the features had on the prediction?*





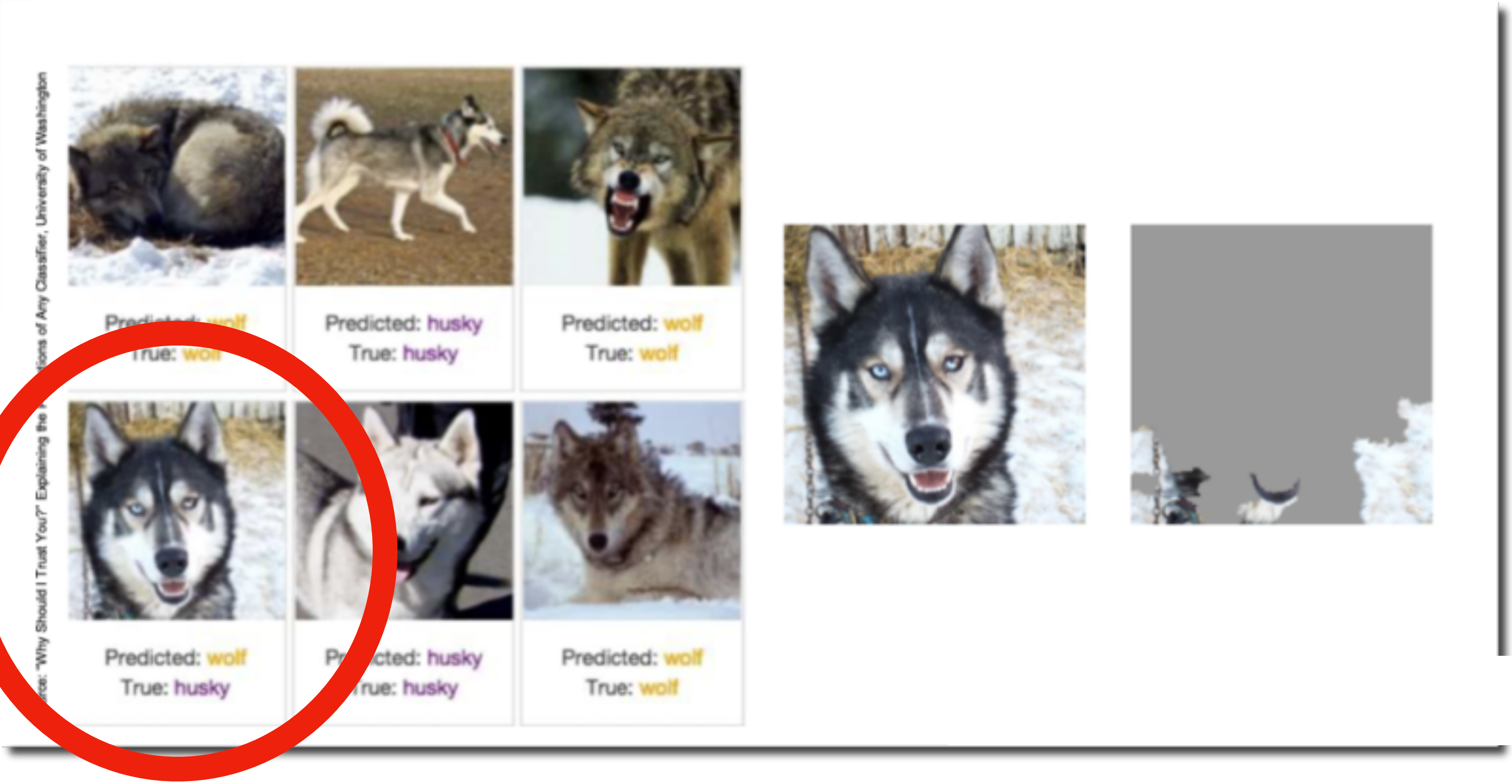
(d) Dogs vs. Cats (e) GradCAM (f) LRP

Husky vs wolf



Husky vs wolf

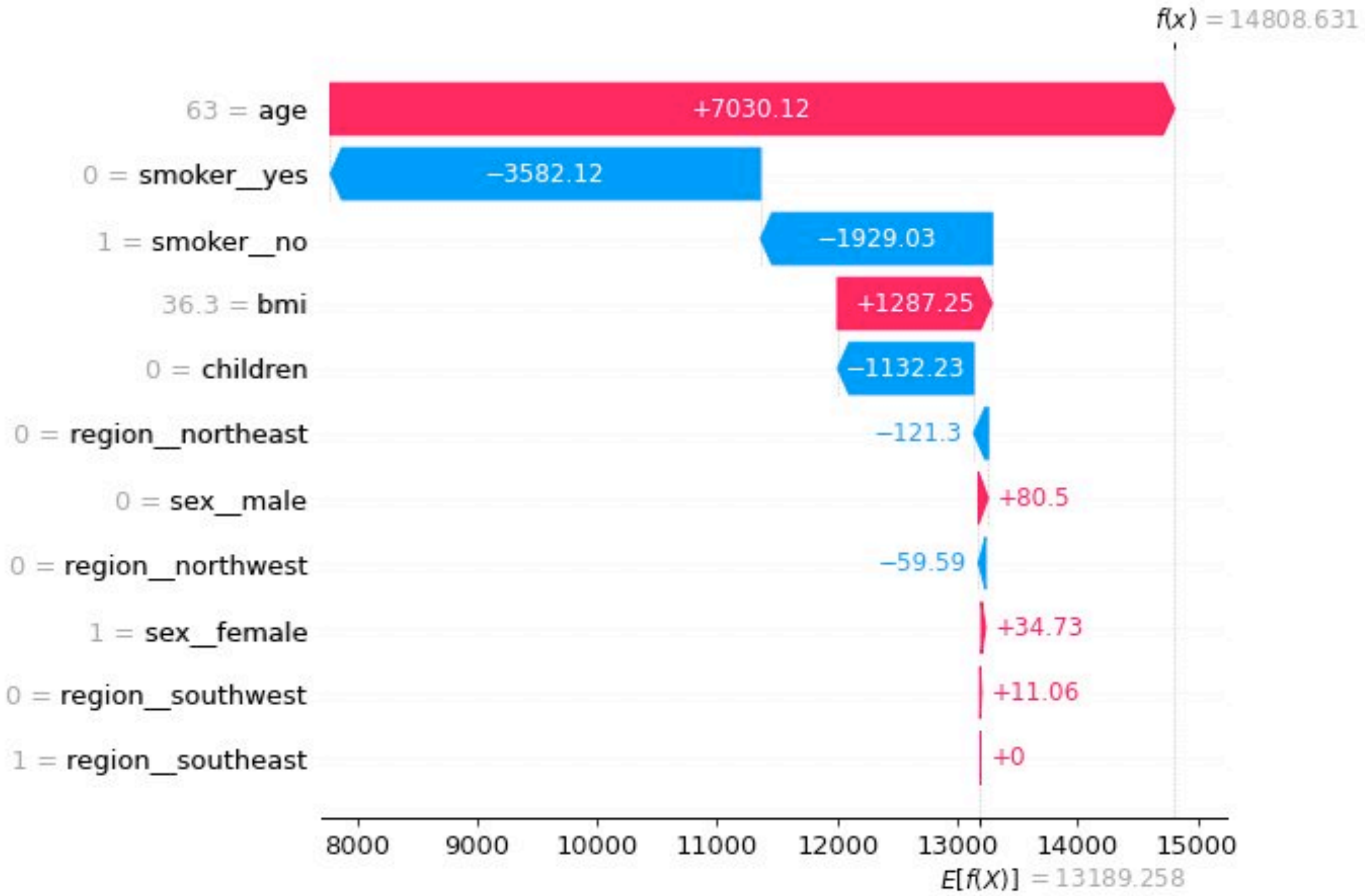
Exposing “Artificial Stupidity”



Feature Attribution with Shapley Values (Tabular data)

NL4XAI

- Answering the question:
- *What is the **impact** that the features had on moving the model away from its baseline?*



What are Shapley Values?

- Game-theoretic way of dividing a payoff to the players of a coalition game *v*
  - Null players have attribution 0 $\phi_i = 0$
  - Attributions sums up to the payoff $\{\phi_i \text{ such that} \sum_i \phi_i = f[x] - E[f[x]]\}$
  - The attribution is the **average** Marginal contribution of a player *i* to every possible sub coalitions S of N players that does not contain *i*

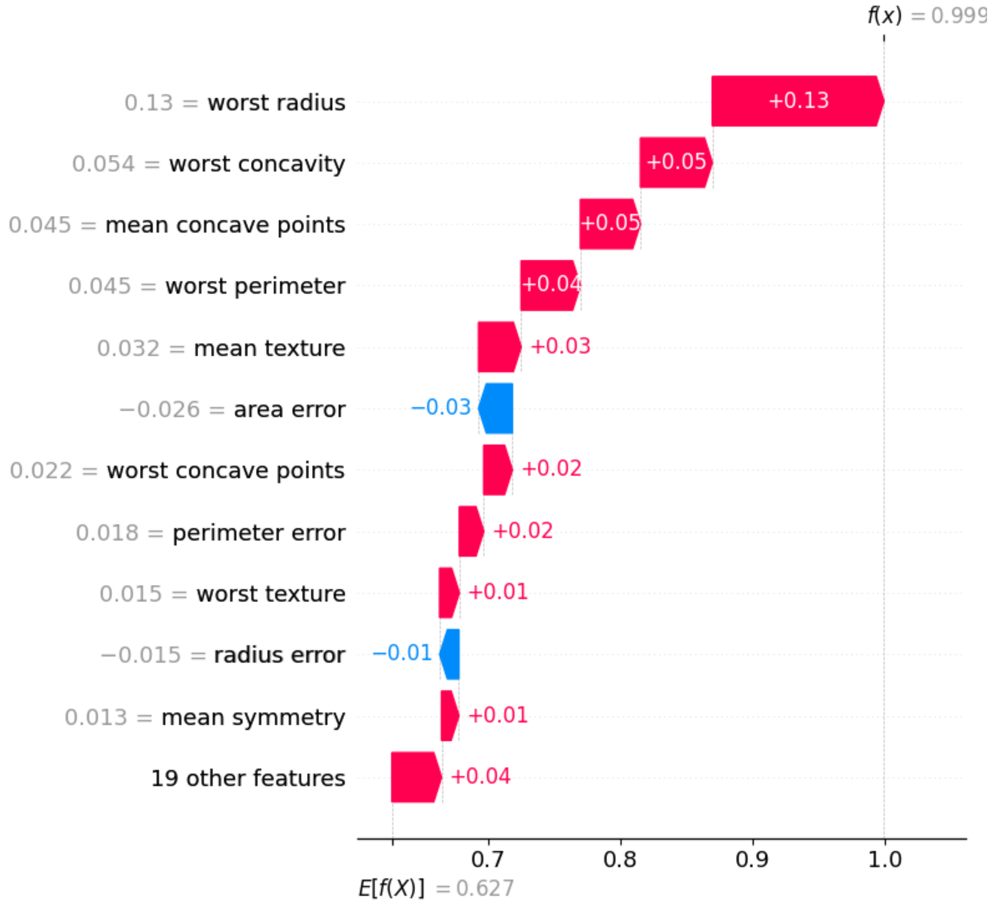
$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \underbrace{\frac{|S|!\,(n - |S| - 1)!}{n!}}_{\text{Appropriate normalisation}} \underbrace{(v(S \cup \{i\}) - v(S))}_{\text{Marginal contribution}}$$

https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882773

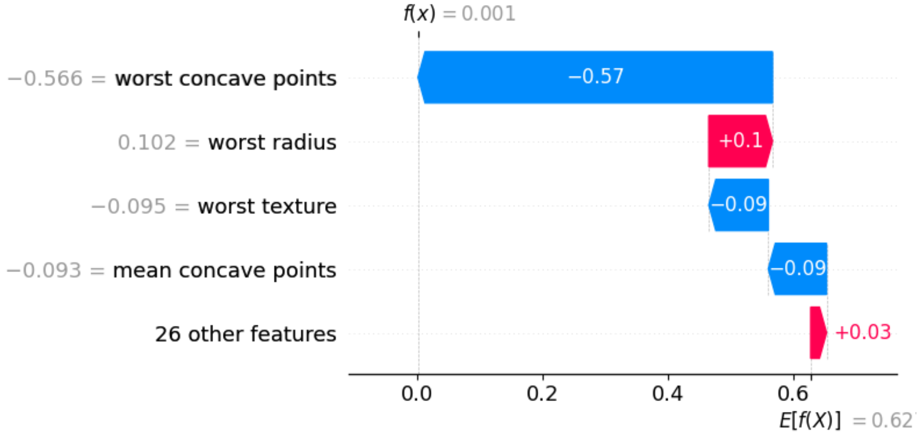
Measuring Model Understandability by means of Shapley Additive Explanations, Mariotti et al, 2022

An heuristic for complexity

Explanation of A



Explanation of B



Formal Definition of Shap Length

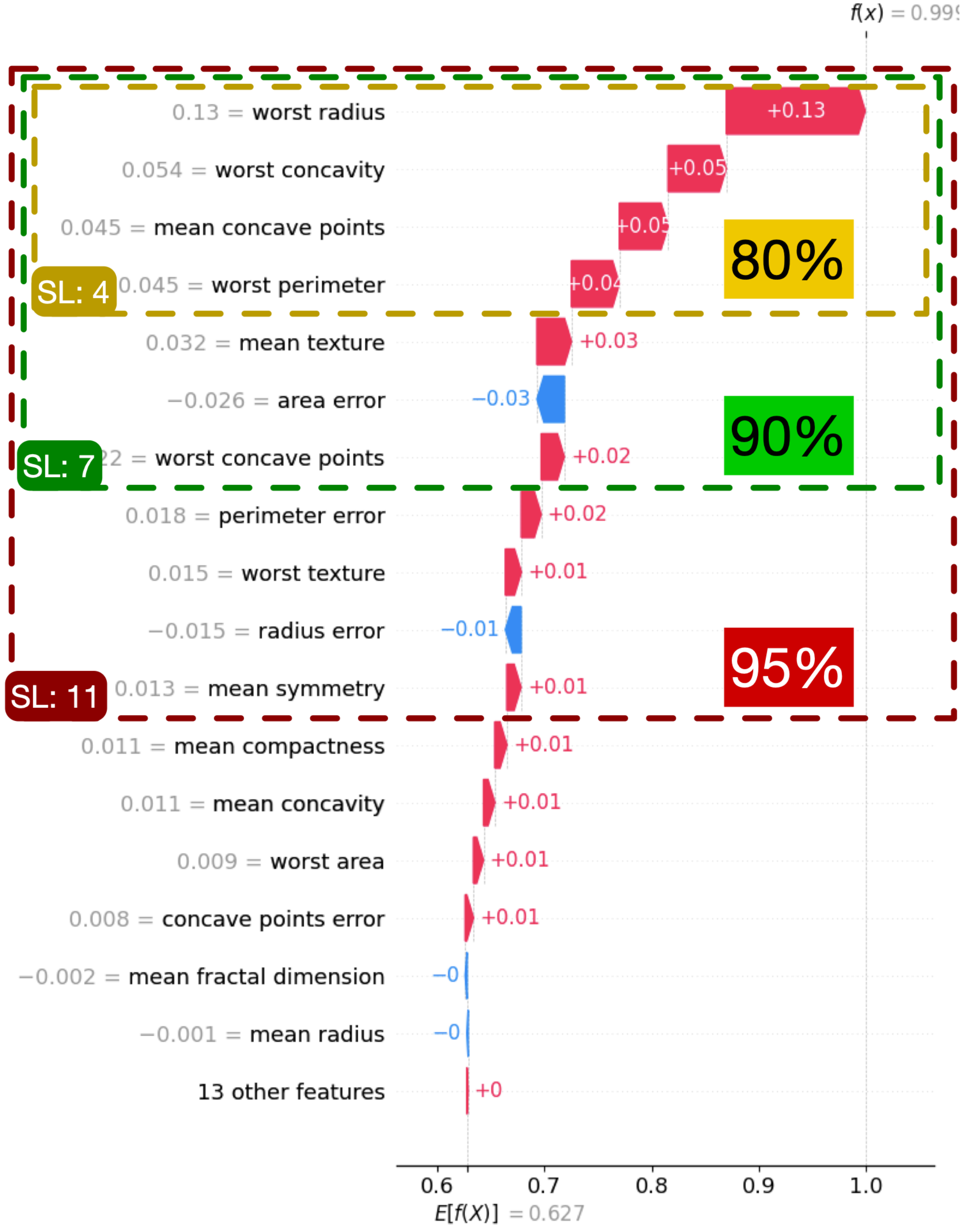
- *Explanation Mass of* ϕ_i *:* $|\phi_i|$
- *Explanation Completeness:*

$$\Gamma(\Phi) := \frac{\sum_{i \in \Phi_{subset}} |\phi_i|}{\sum_{i \in \Phi} |\phi_i|}$$

- *p%-complete explanation:*

$$\Phi_s \text{ such that } \Gamma(\Phi_s) \geq p$$

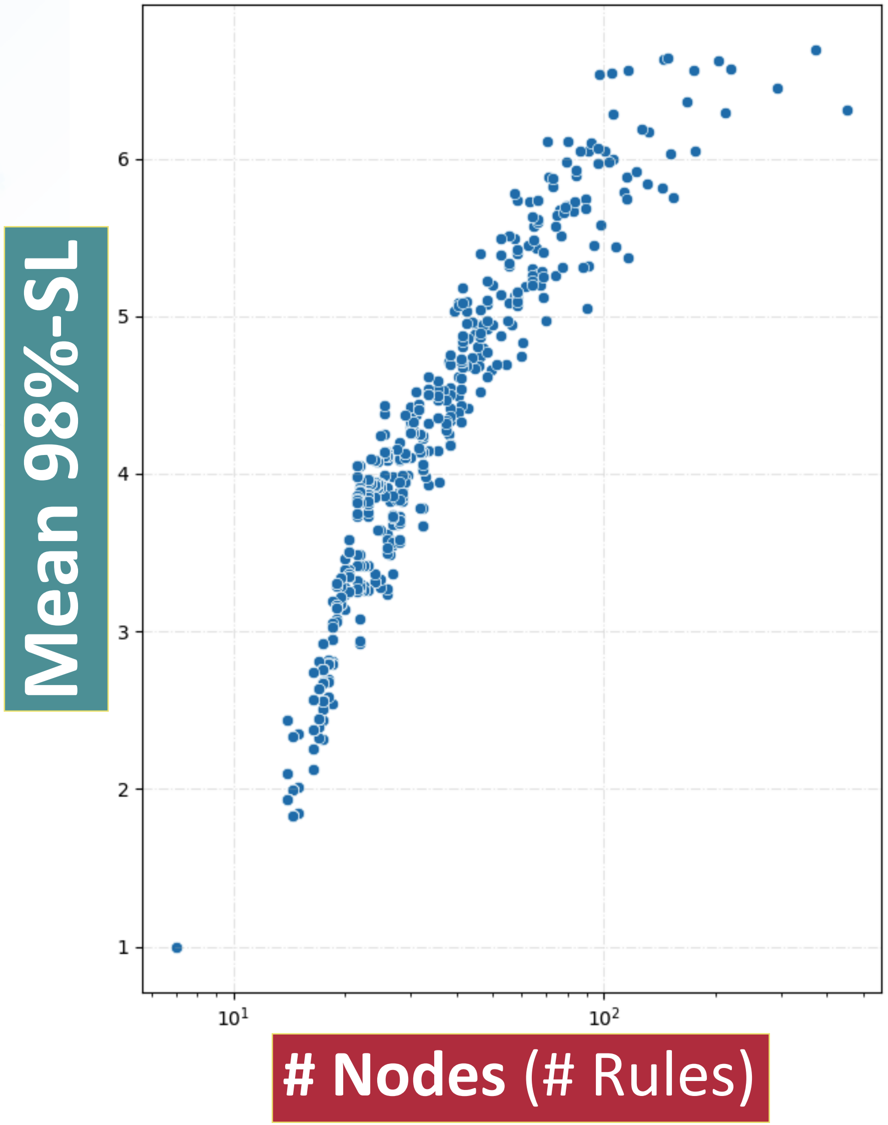
- *Shap Length* $SL_{p\%} := ||\Phi_p||$

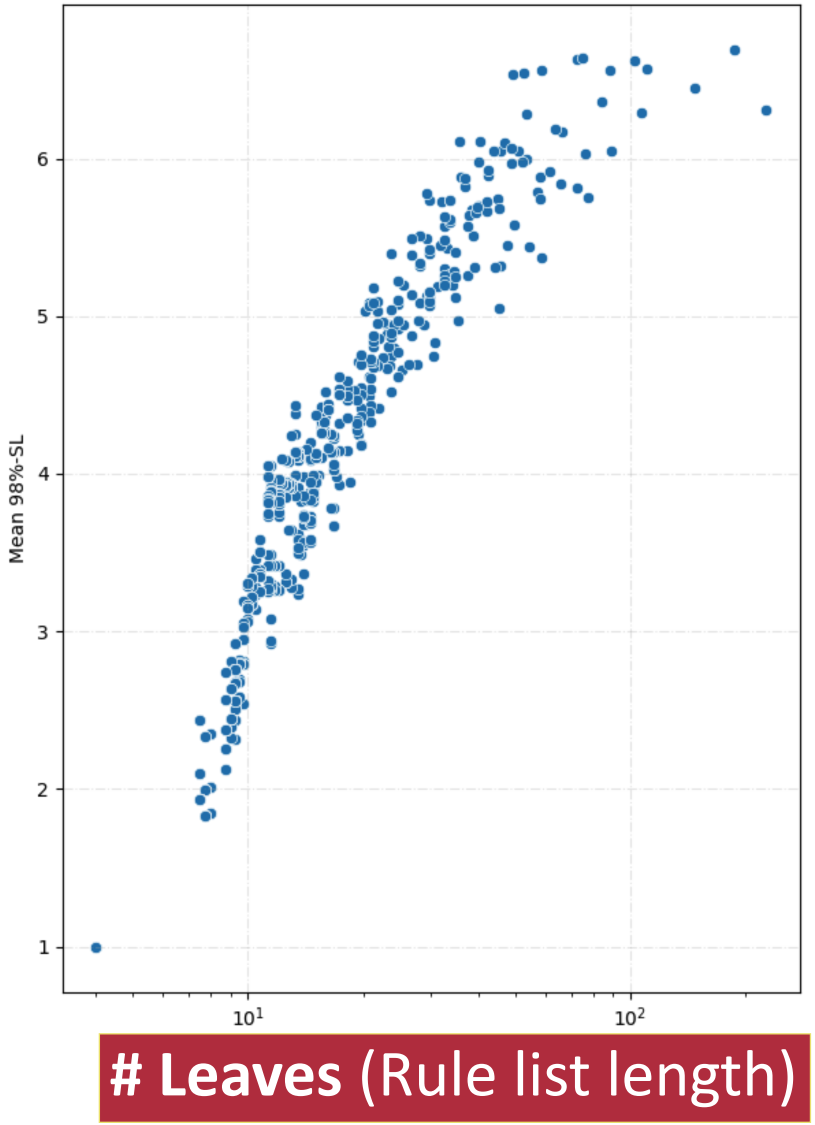


https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882773

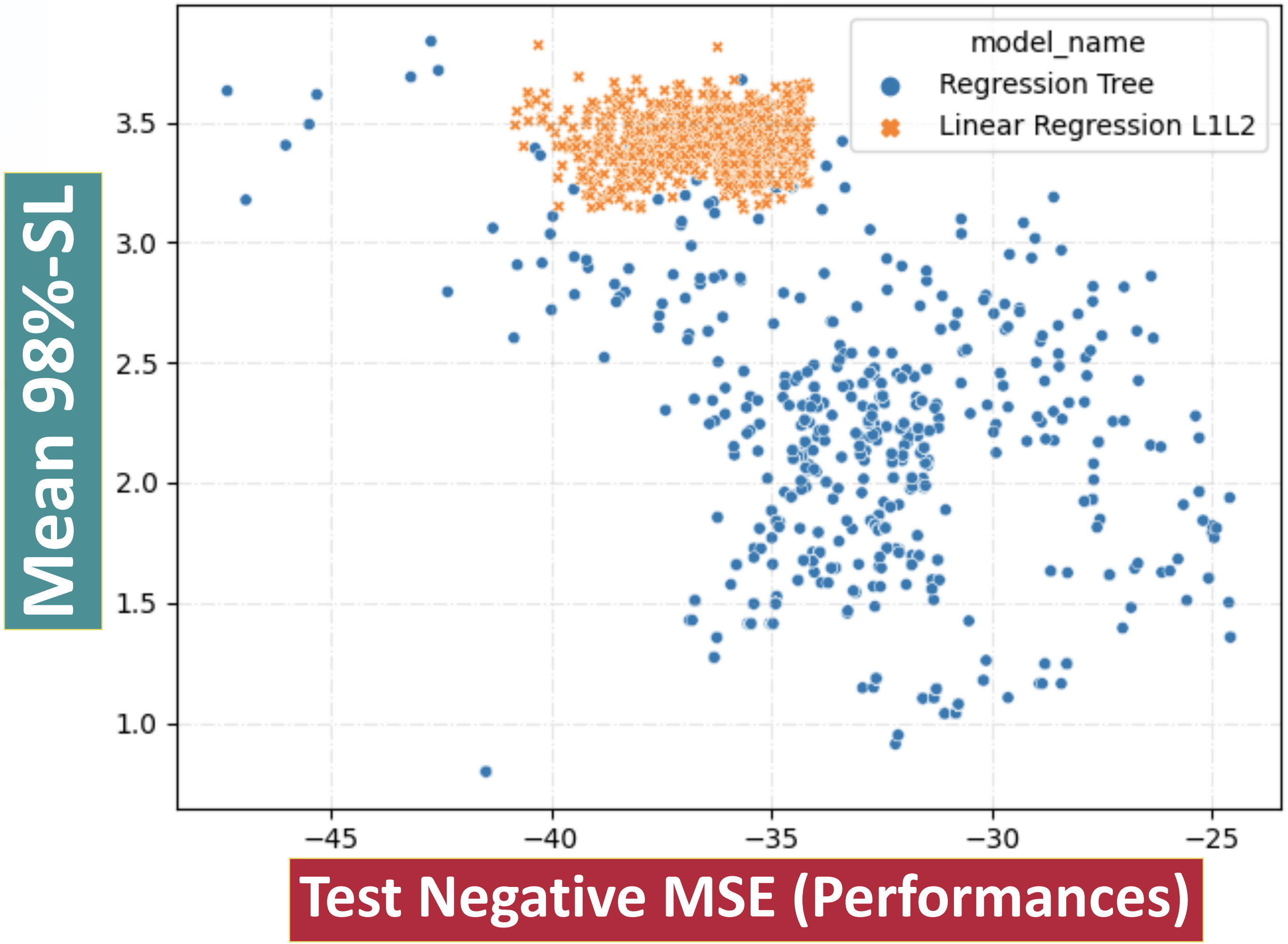
Measuring Model Understandability by means of Shapley Additive Explanations,
Mariotti et al, 2022

Proportionality with tree-related complexity metrics



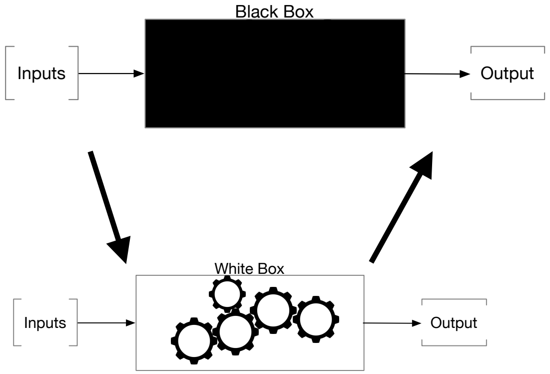


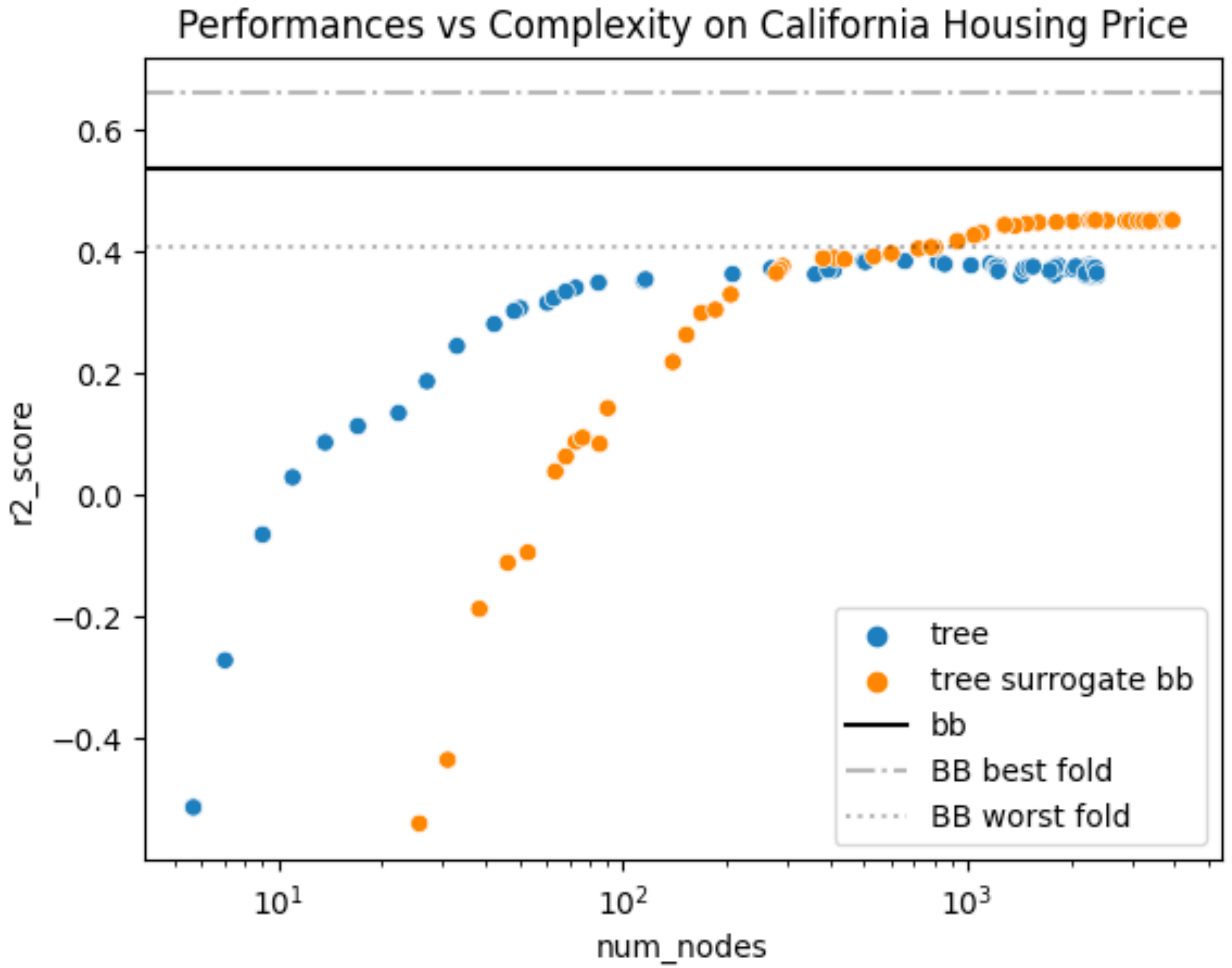
Comparison of Tree vs Linear models on Boston-house dataset



Surrogation of black box with interpretable white box

- Have white-boxes behave like black-boxes
- Train on Black-Box Labels
- Exploring the different tradeoffs of this approach

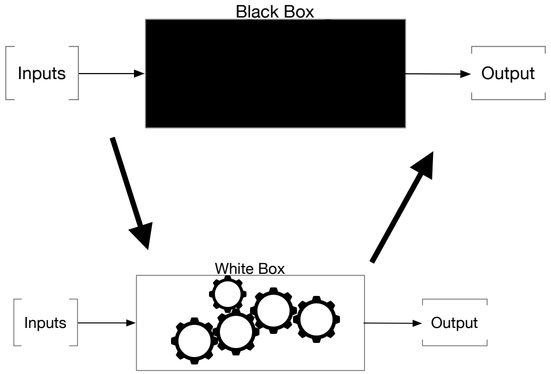




How to measure faithfulness?

- What does it mean “behave like the black box”?
  - Fidelity Accuracy: How often they predict the same label
    - A task-related measure
  - **SHAP-Gap**: How similar are their reasoning (in SHAP approximation)
    - Distance of SHAP Explanations (L2 or Cos)
    - Measures wether predictions are not only similar, but also if their rationale is similar

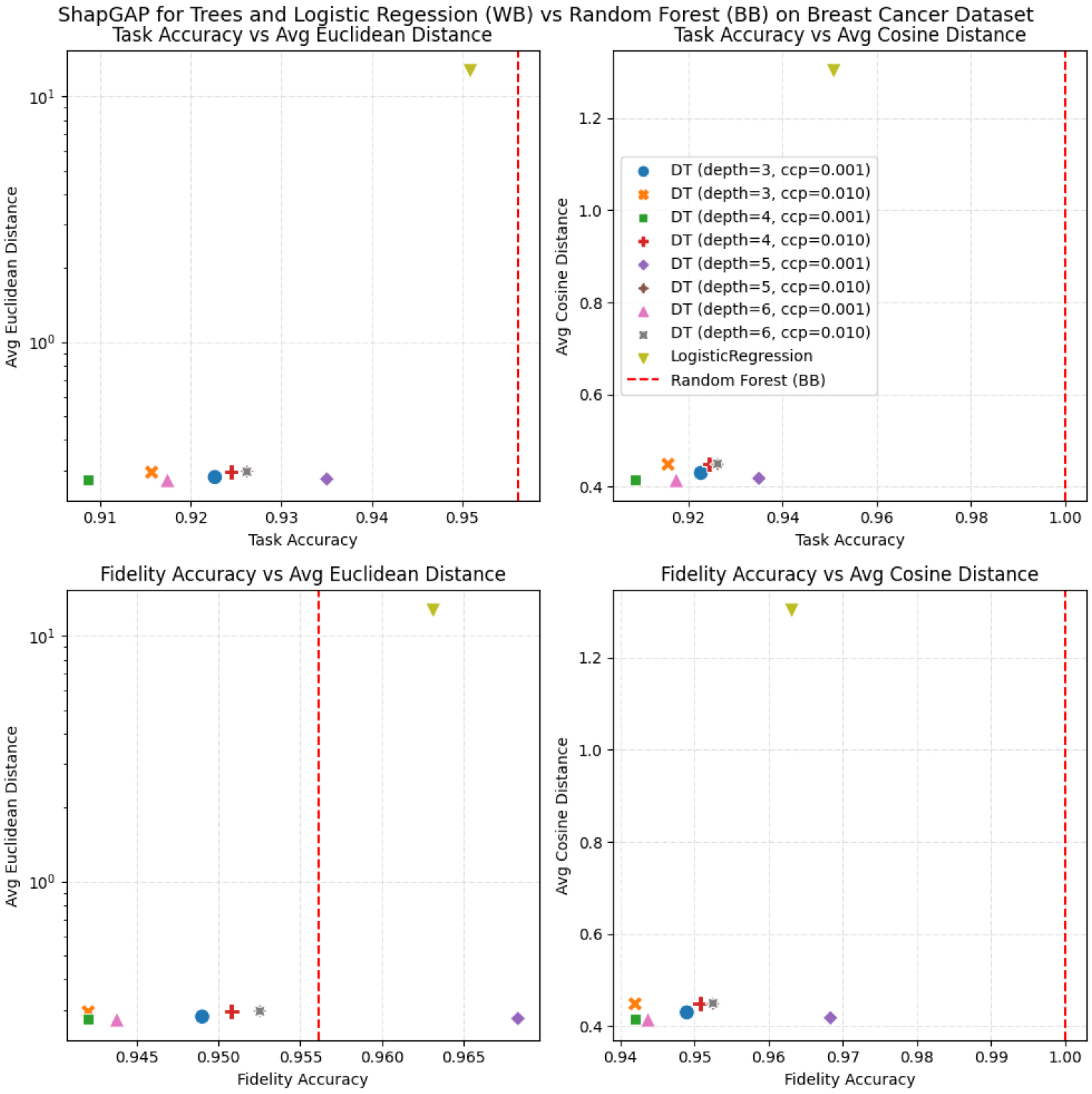
$$ShapGAP(D,d) = \frac{1}{n}\sum_{i=1}^{n} d(S_{bb}(x_i), S_{wb}(x_i))$$



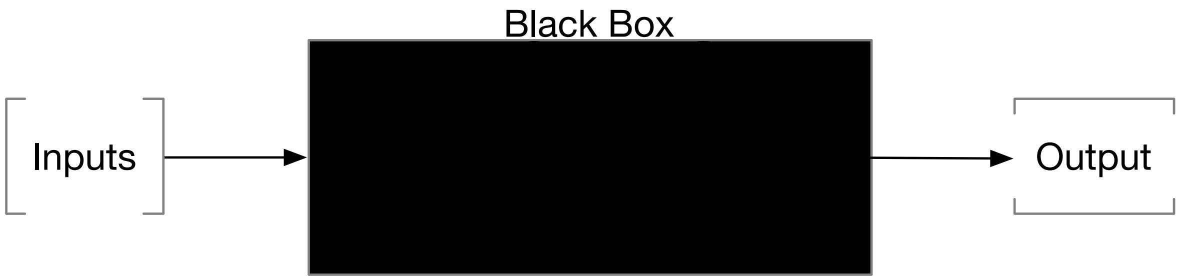
Illustrative Example

$$ShapGAP_{L2}(D) = \frac{1}{n}\sum_{i=1}^{n} ||S_{bb}(x_i) - S_{wb}(x_i)||_2$$

$$ShapGAP_{Cos}(D) = \frac{1}{n}\sum_{i=1}^{n}(1 - \frac{S_{bb}(x_i)\cdot S_{wb}(x_i)}{||S_{bb}(x_i)||_2||S_{wb}(x_i)||_2})$$

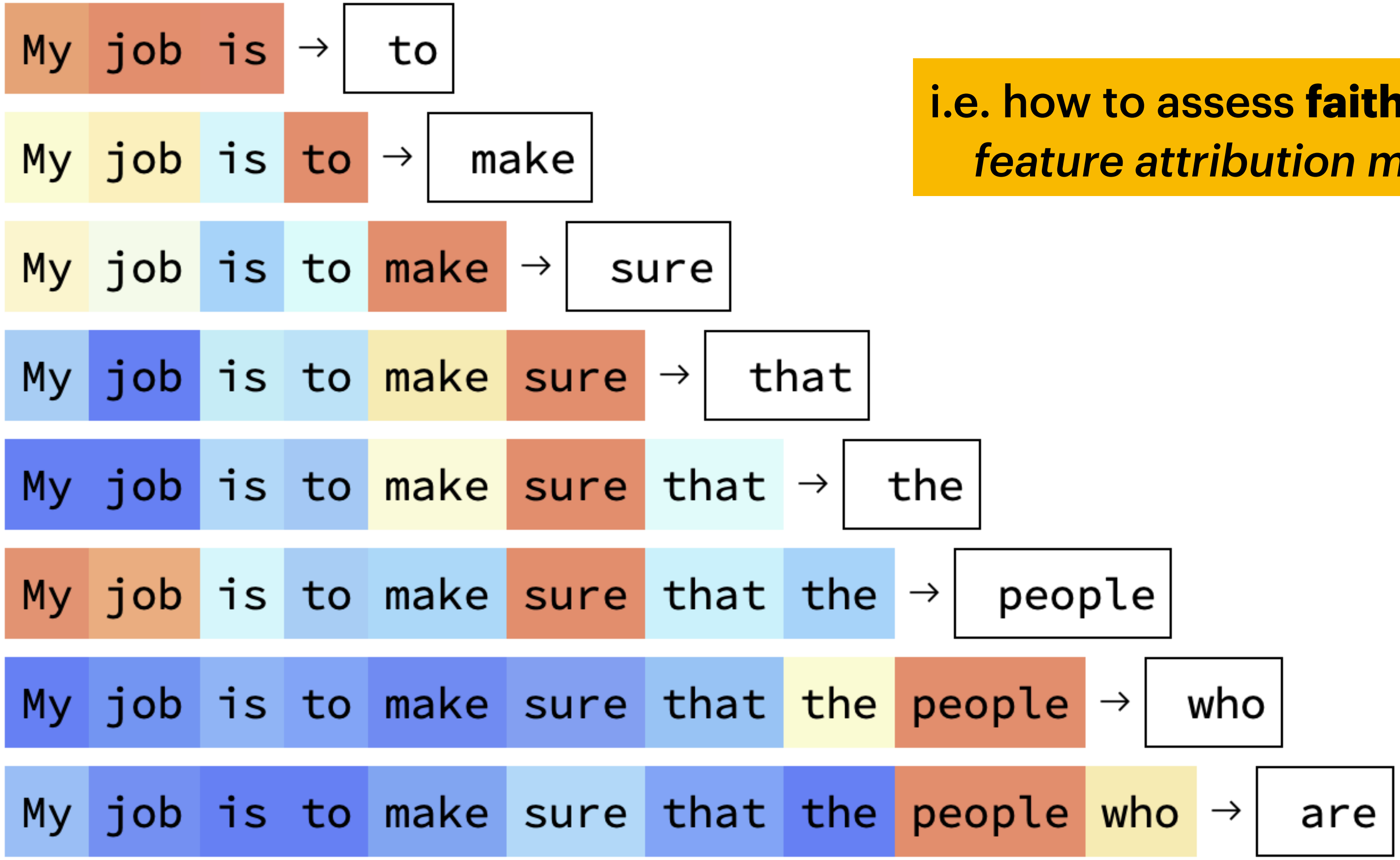


Examples of Feature Attributions for Text (Language Generation)



Examples of Feature Attributions for Text (Language Generation)

NL4XAI

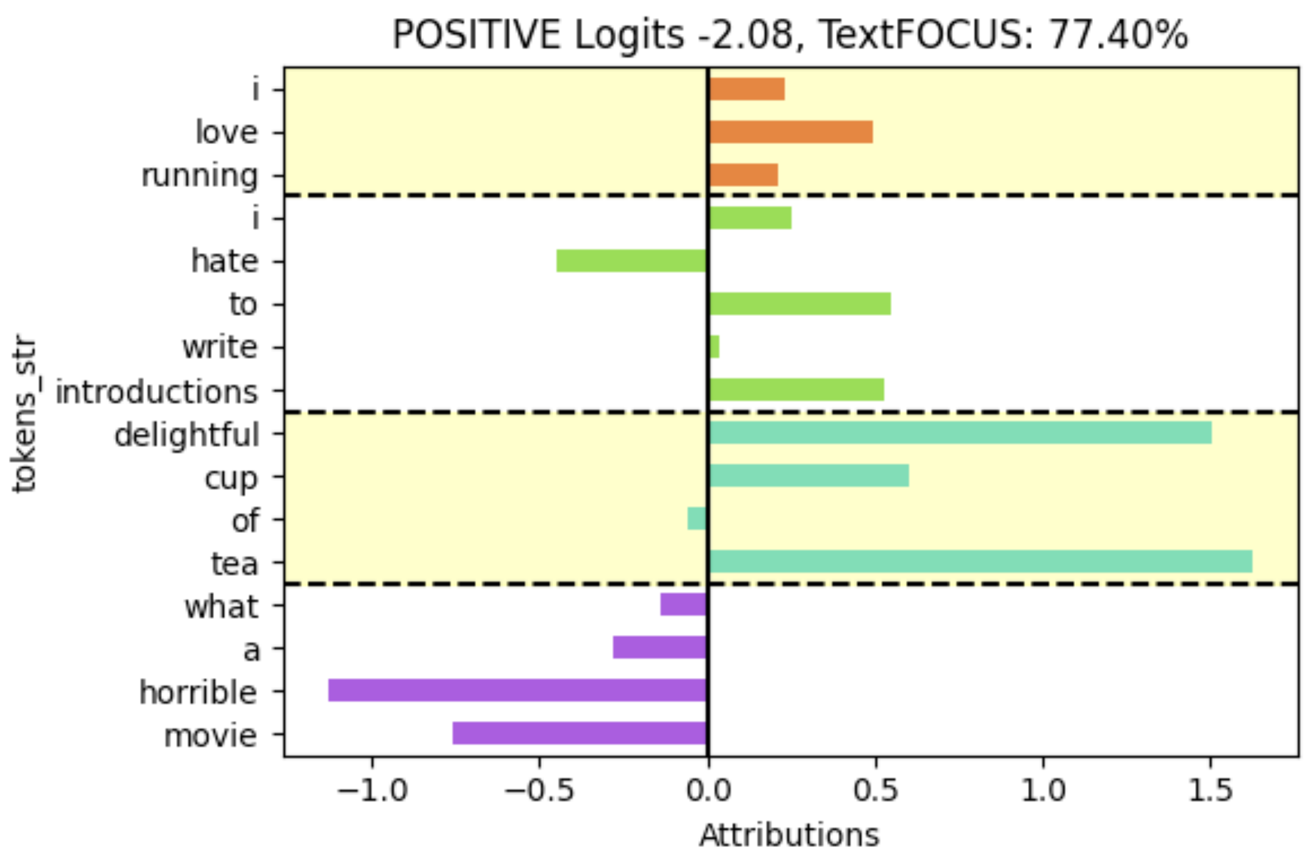


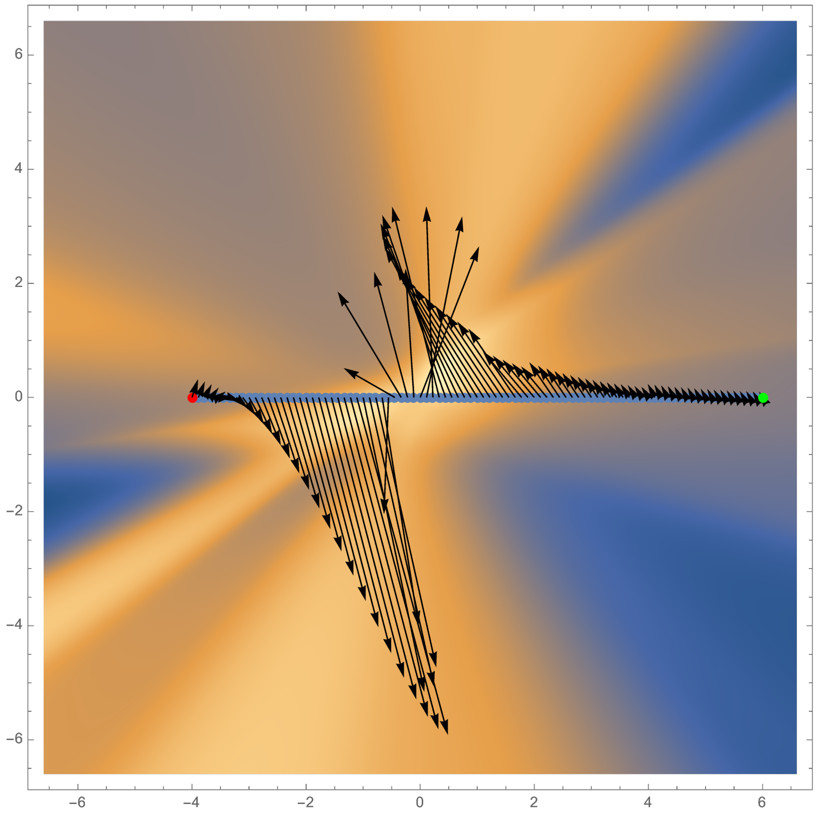
i.e. how to assess **faithfulness** of *feature attribution methods?*

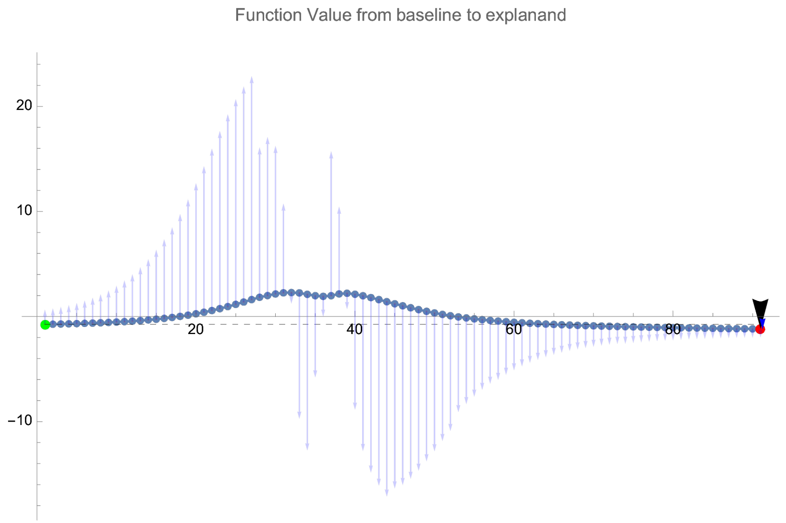
Problem of evaluating faithfulness

[CLS]	sentence 1	[SEP]	sentence 2	[SEP]					N = 2
[CLS]	sentence 1	[SEP]	sentence 2	[SEP]	sentence 3	[SEP]	sentence 4	[SEP]	N = 4

Feature attributions in NLP tasks (Integrated Gradients)

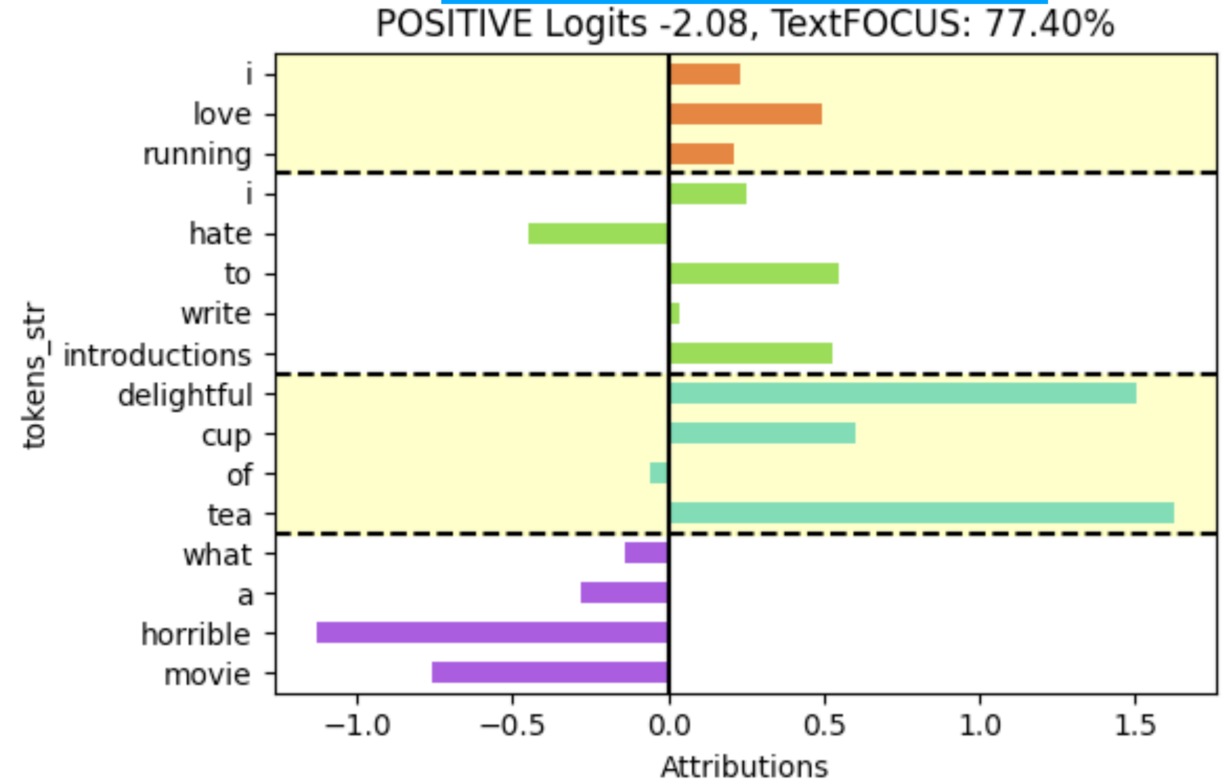




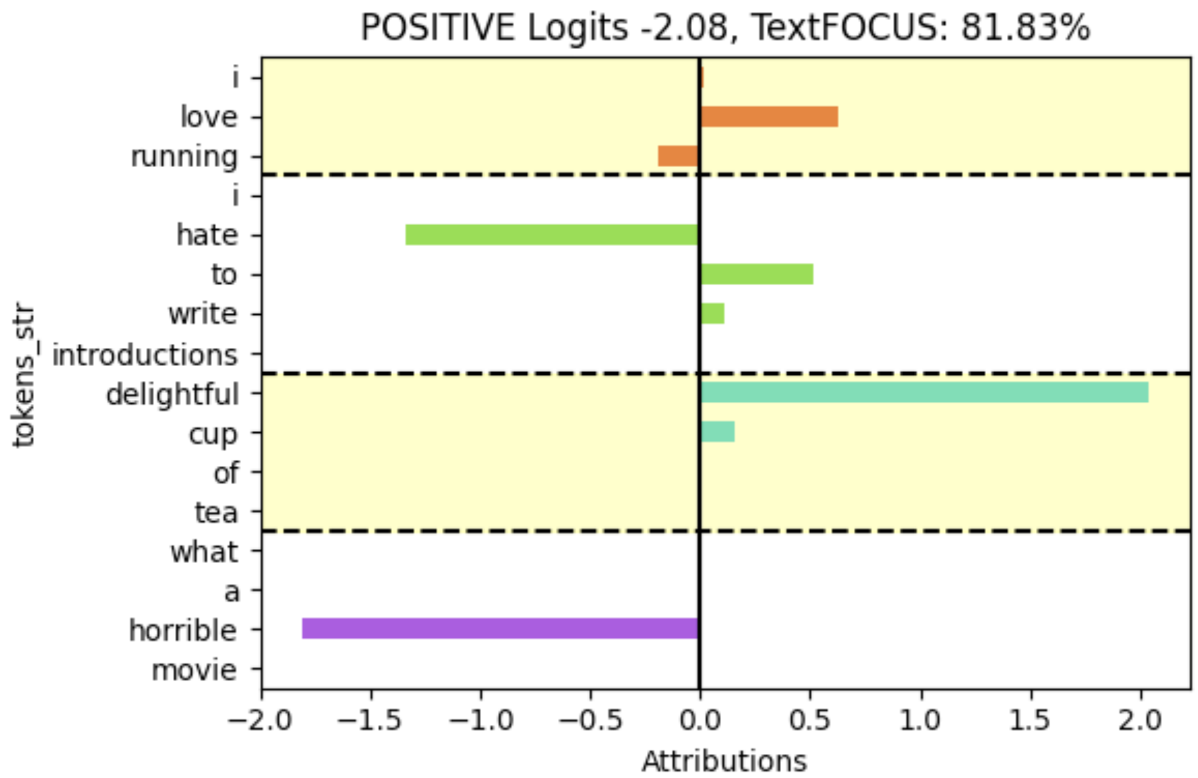


Problem of evaluating faithfulness

Integrated Gradients



LIME

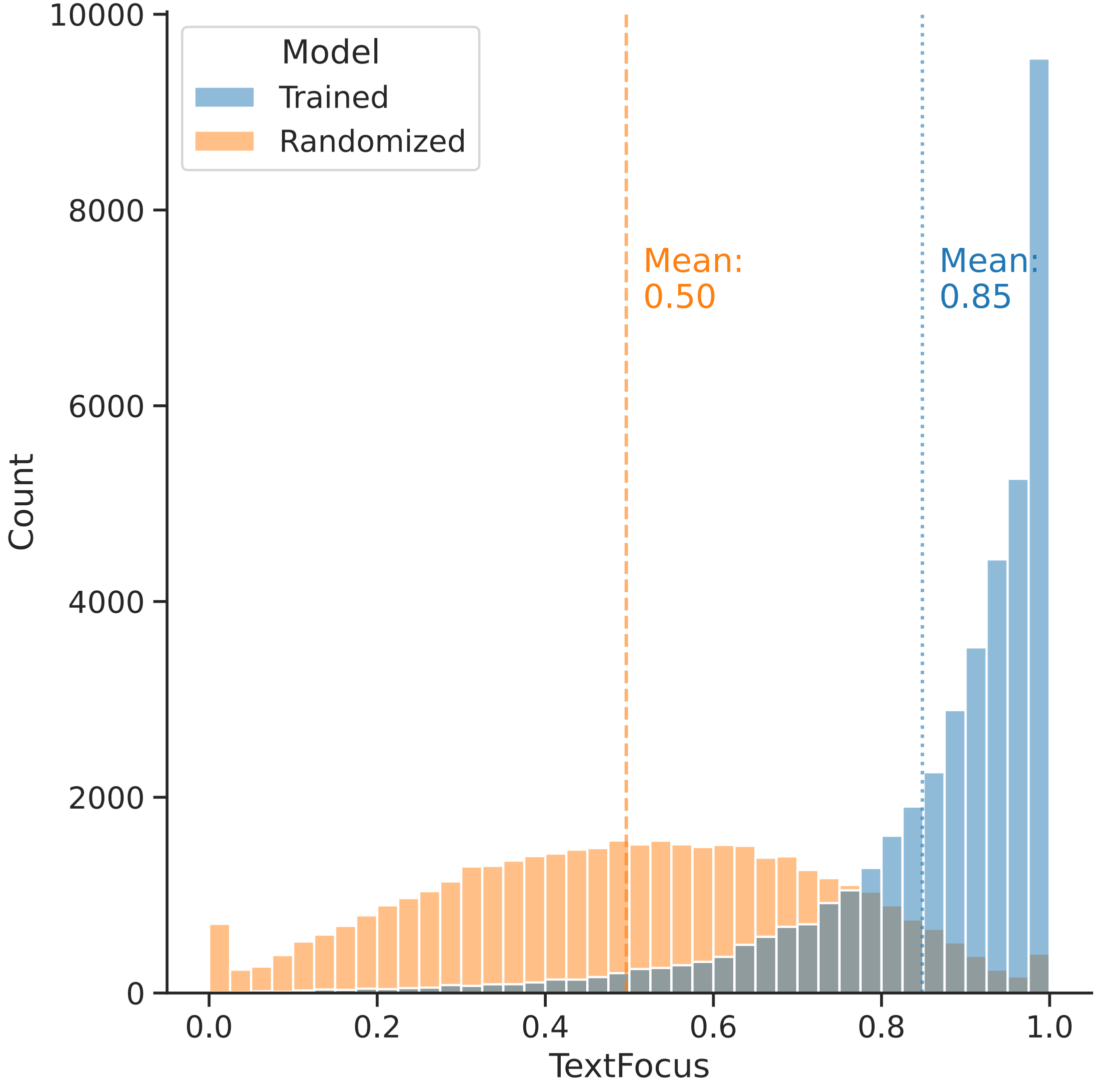


- True Positive evidence (TP) $= \sum_{i \in T} |\max(0, \alpha_i)|$
- False Positive evidence (FP) $= \sum_{i \in N} |\max(0, \alpha_i)|$
- True Negative evidence (TN) $= \sum_{i \in N} |\min(0, \alpha_i)|$
- False Negative evidence (FN) $= \sum_{i \in T} |\min(0, \alpha_i)|$

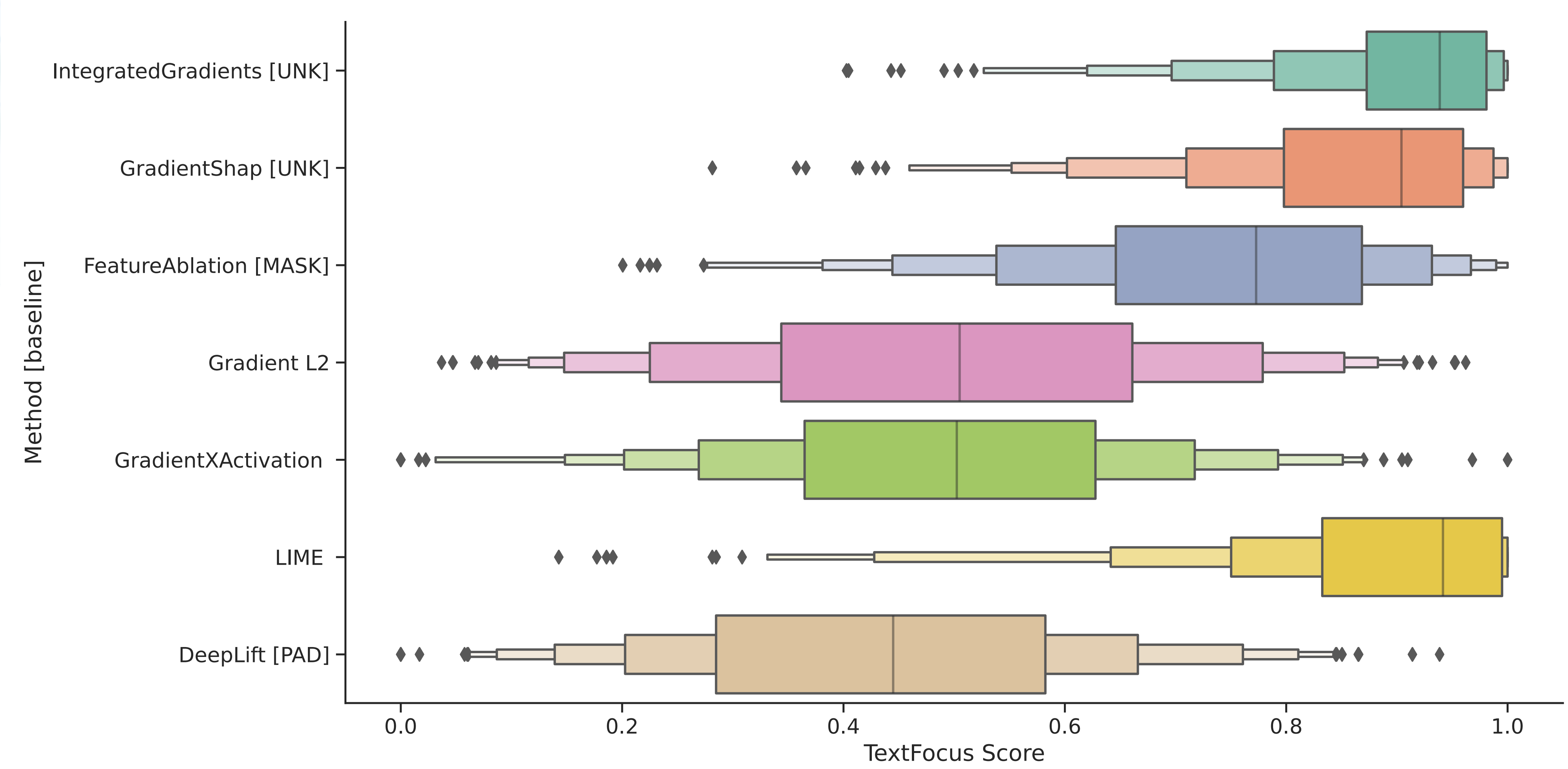
N = 2

N = 4

Problem of evaluating faithfulness



Problem of evaluating faithfulness



Supporting *the right to explanation by* AI POWERED SYSTEMS