Explainable Artificial Intelligence (XAI)

Rémi Kazmierczak
Unité d'informatique de d'ingénierie des systèmes
Ensta Paris

26/04/2023

Context

An autonomous car makes a decision causing an accident and resulting in human and material damage

The company responsible for the car is asked to justify the decision-making of the autonomous vehicle

“Tesla hit with another lawsuit over a fatal Autopilot crash”

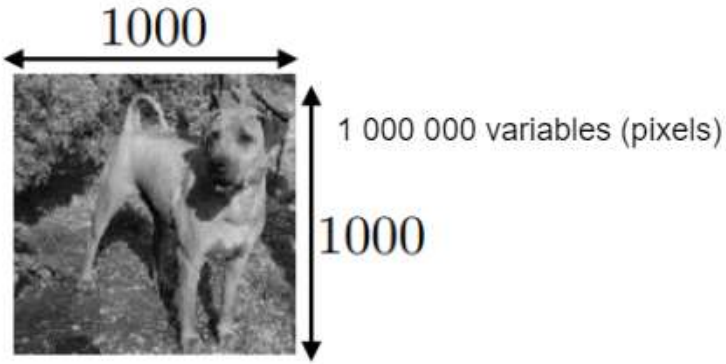
Plan

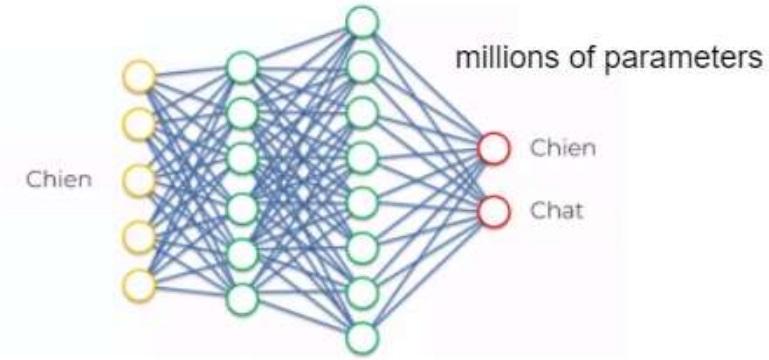
1. What is XAI
2. Why do we need it
3. Levels of Transparency
4. Trade-off between interpretability and performance
5. Post-hoc methods to achieve XAI
6. Challenges to achieve Responsible AI
7. Bibliography

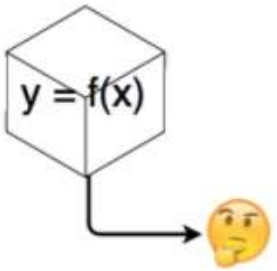
What is explainability in machine/deep Learning

Deep Learning models are powerful but opaque

- Gap between high dimensional mathematical optimization and human abilities



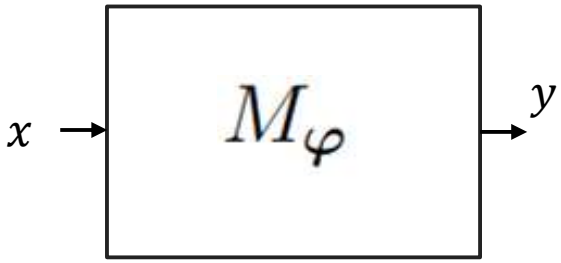




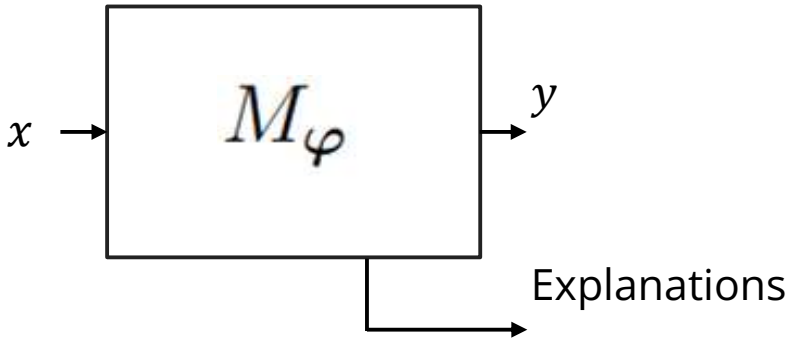
Definition of Explainability

Given an audience, an **explainable Artificial Intelligence** is one that produces details or reasons to make its functioning clear or easy to understand.

Classical Model :

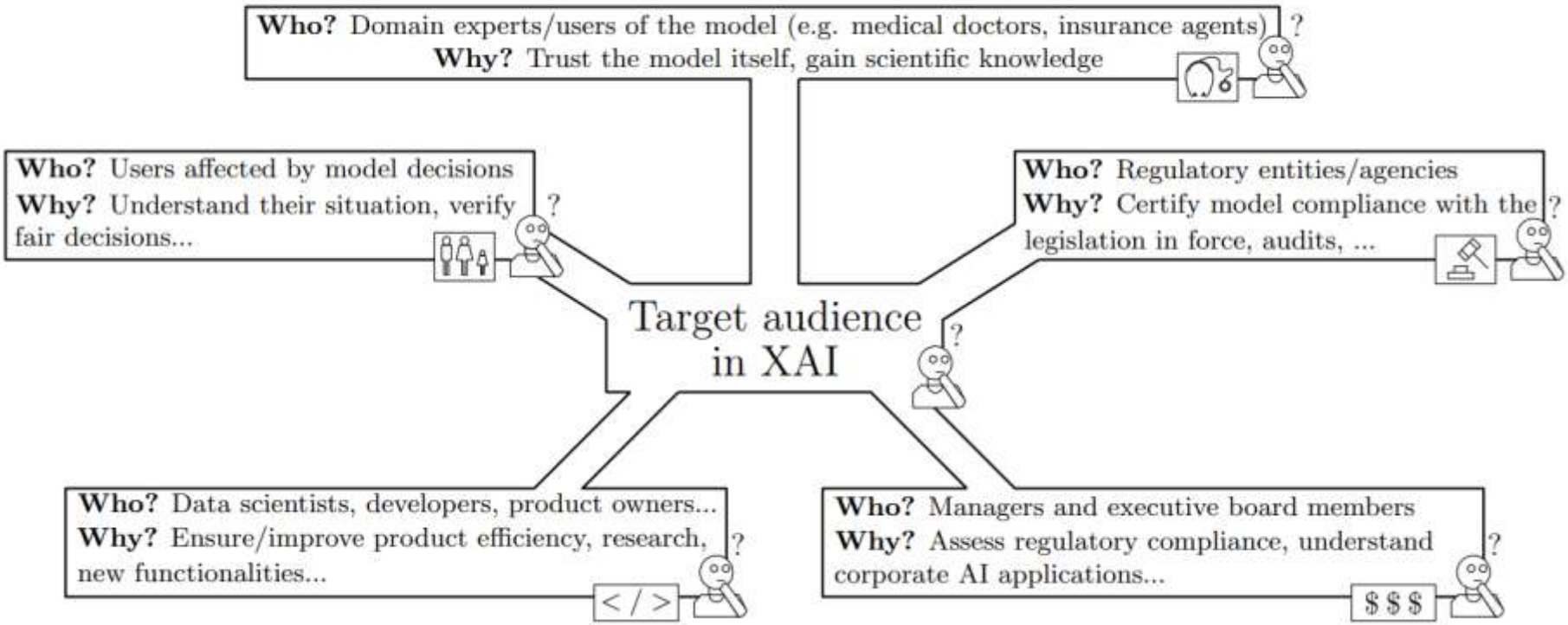


Explanable Model :



Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI - AB Arrieta et al. - Information Fusion, 2020

Who needs XAI



Why do we need XAI - Trustworthiness/Causality

Trustworthiness might be considered as the confidence of **whether a model will act as intended** when facing a given problem

Explainable models might ease the task of finding relationships that could be tested further for a stronger **causal link between the involved variables**

Why do we need XAI - Transferability

- Elucidate the **boundaries** that might affect a model
- Understand the inner relations taking place within a model facilitates the ability of a user to **reuse this knowledge** in another problem

Why do we need XAI - Fairness

Highlighting **bias** in the data a model was exposed to

Example

Wrong

*A **man** sitting at a desk with a laptop computer.*

Right for the wrong reasons

*A **man** holding a tennis racquet on a tennis court.*

Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A. (2018). Women Also Snowboard: Overcoming Bias in Captioning Models. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11207. Springer, Cham. https://doi.org/10.1007/978-3-030-01219-9_47

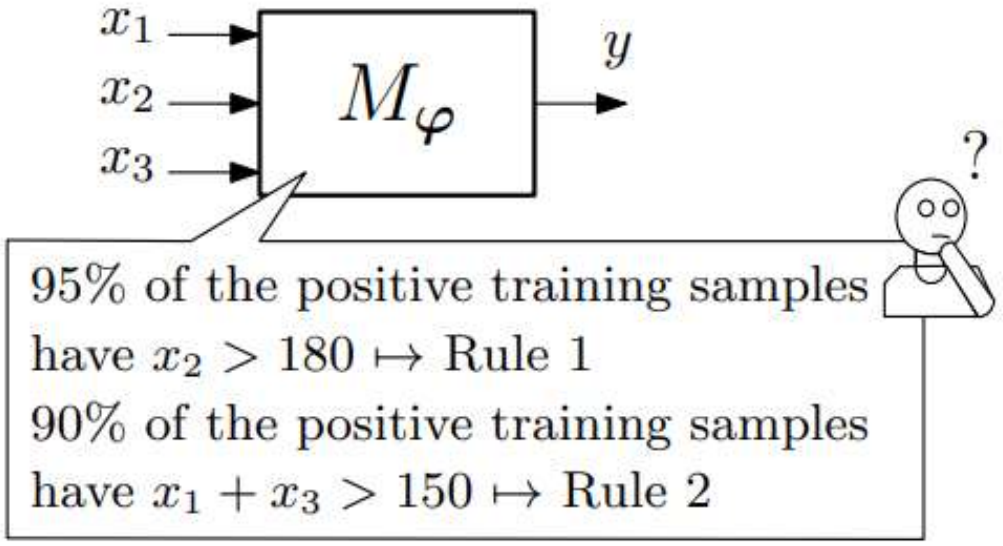
How?

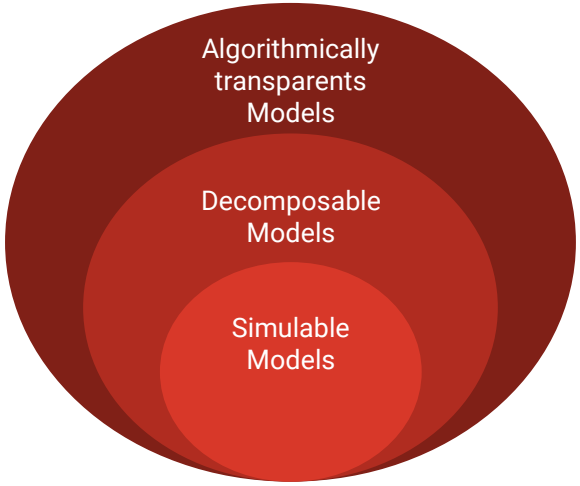
Clear distinction in the literature among models that:

- Are interpretable by design → **Transparent Models**
- Should be explained by XAI techniques → **Post-hoc explainability**

Levels of Transparency - Algorithmic Transparency

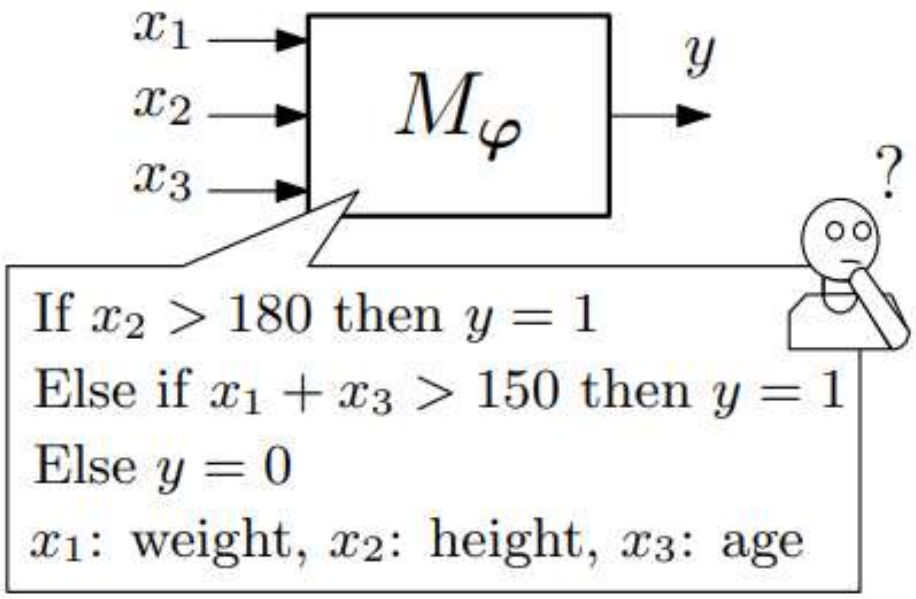
Algorithmic Transparency deals with the ability of the user to understand the process followed by the model to produce any given output from its input data

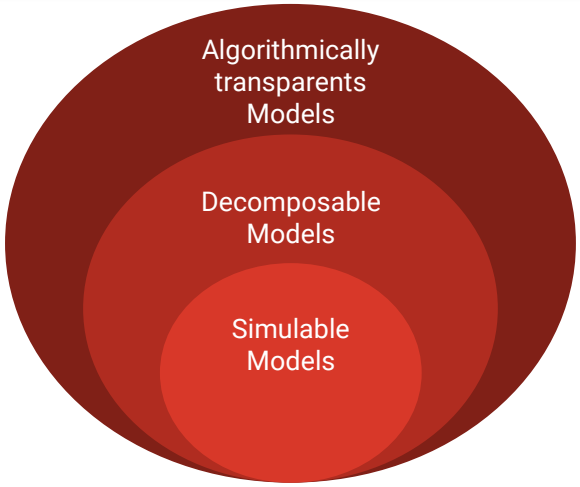




Levels of Transparency - Decomposability

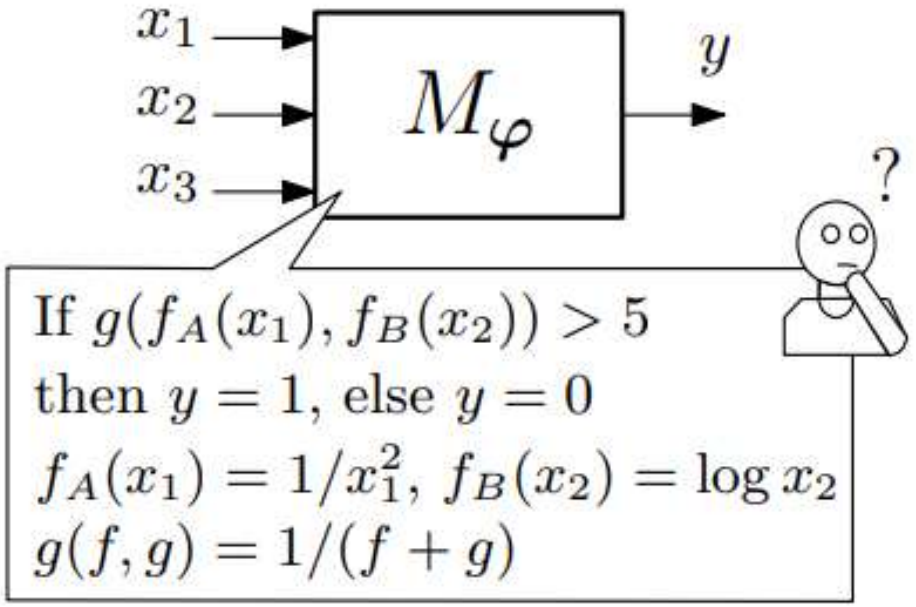
Decomposability stands for the ability to explain each of the parts of a model (input, parameter and calculation)

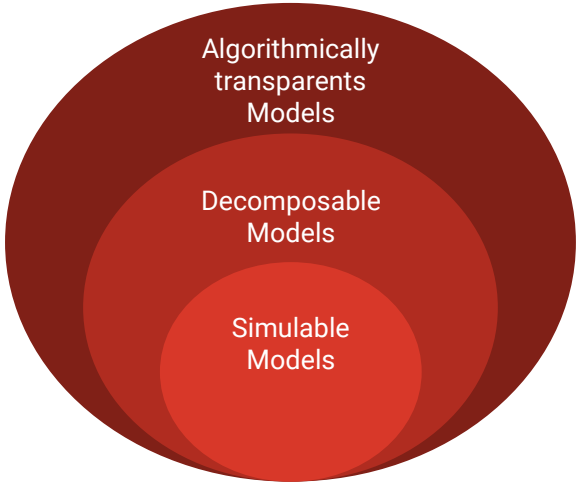




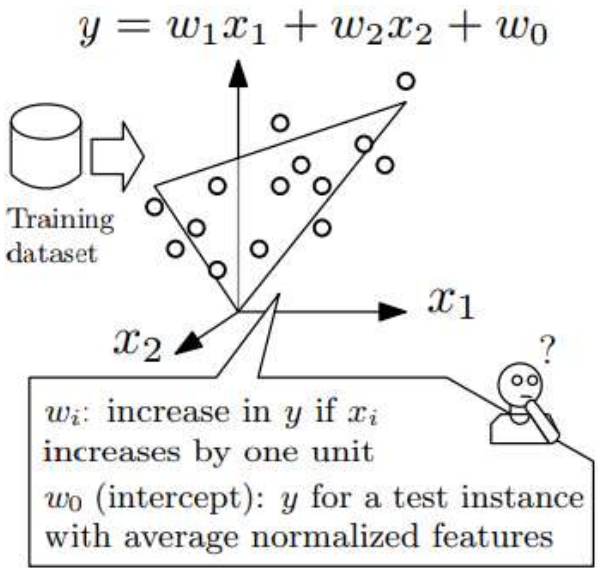
Levels of Transparency - Simulatability

Simulatability denotes the ability of a model of being simulated or thought about strictly by a human

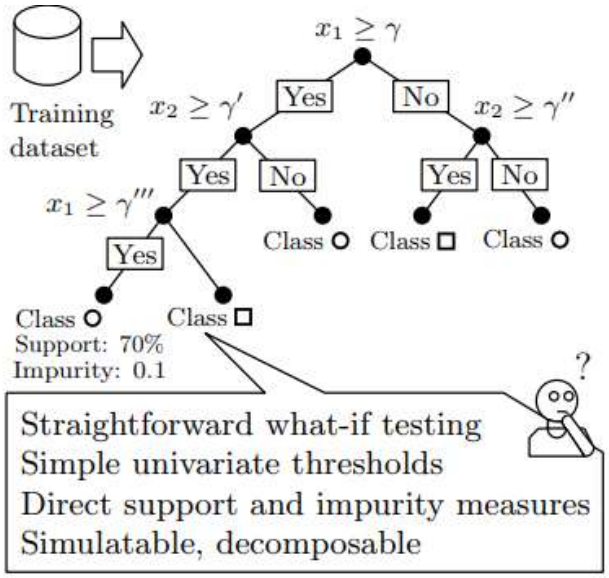




Examples of Transparent Models 1/4

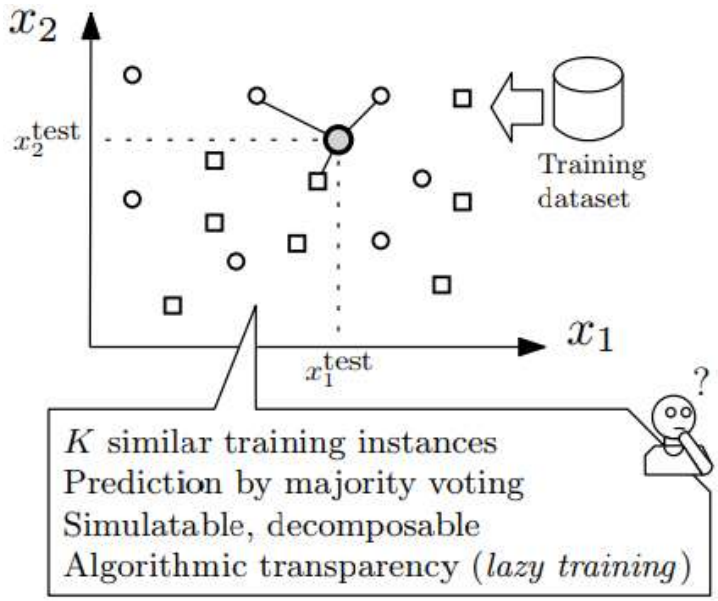


Linear Regression

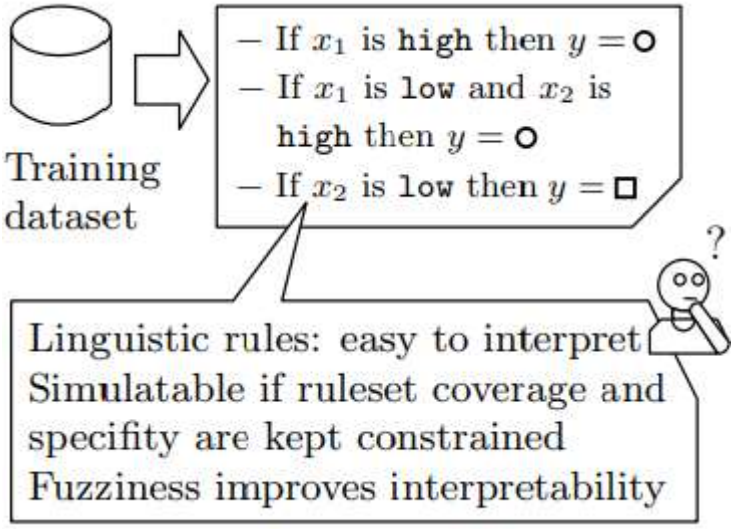


Decision Trees

Examples of Transparent Models 2/4



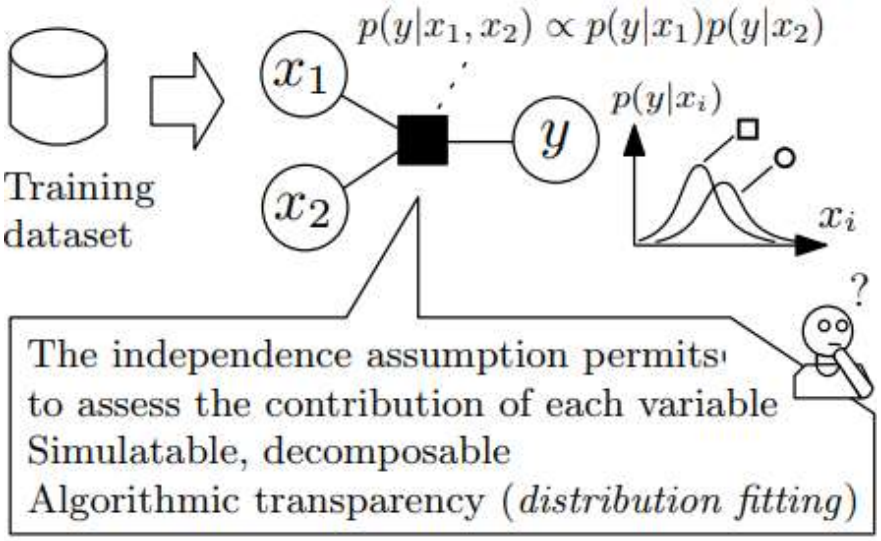
K-Nearest Neighbors



Rule-based Learner

Examples of Transparent Models 3/4

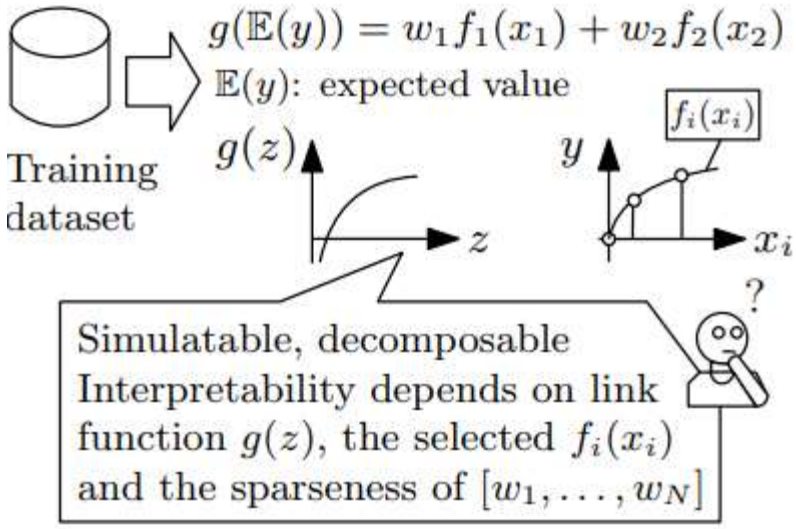
Usually takes the form of a **probabilistic model** whose links represent the conditional dependencies between a set of variables



Bayesian Models

Examples of Transparent Models 4/4

Linear model in which the value of the variable to be predicted is given by the **aggregation** of a number of **unknown smooth functions** defined for the predictor variables



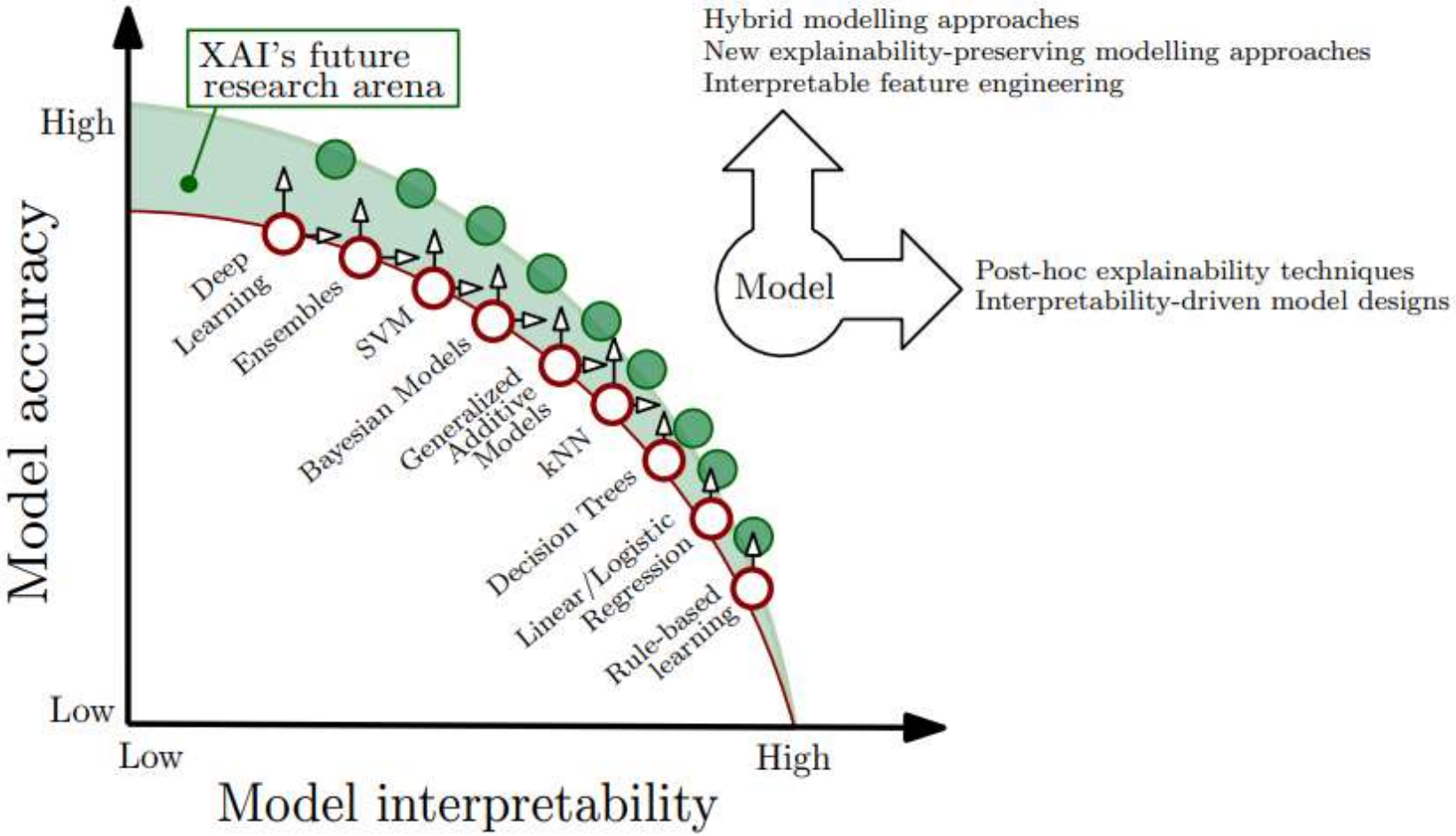
Generalized Additive Models

Opaque Models

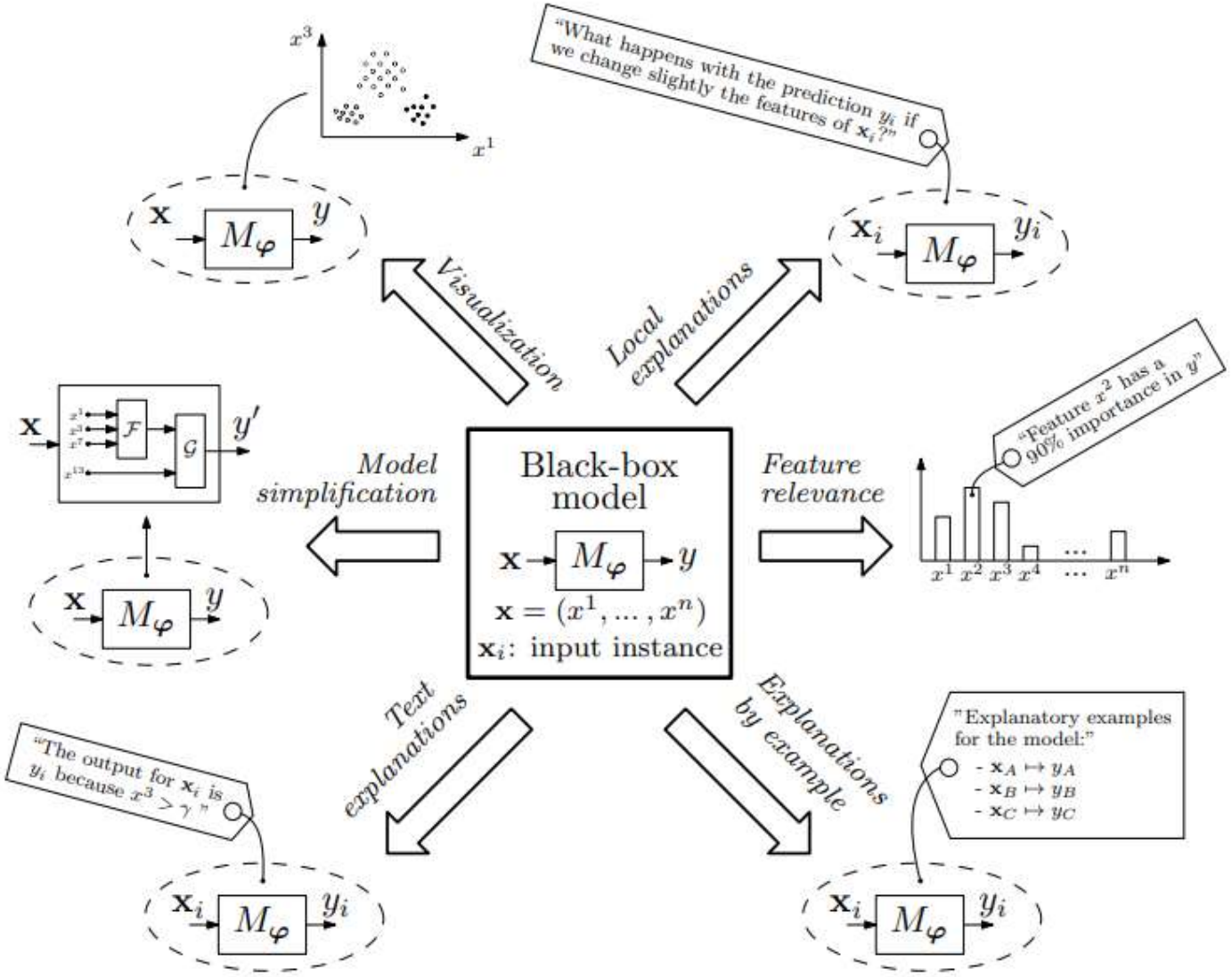
5 categories of Models that are **not transparent**

- Tree Ensembles
- Support Vector Machines
- Multi-layer Neural Networks
- Convolutional Neural Networks
- Recurrent Neural Networks

Trade-off between model interpretability and performance



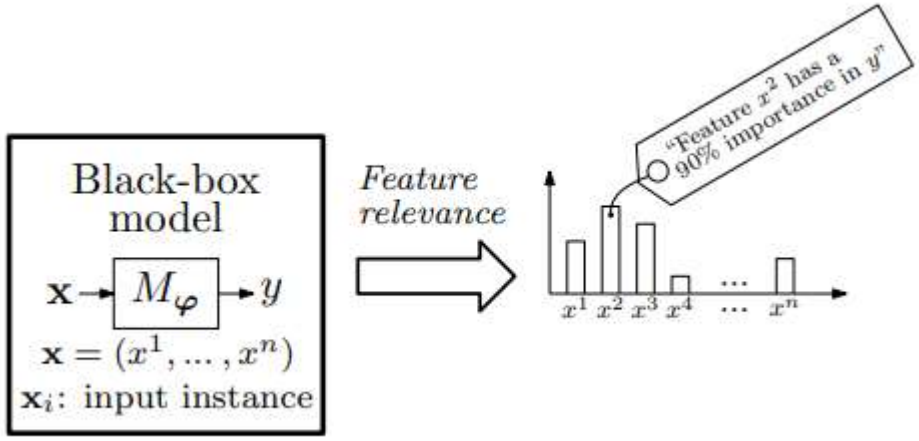
Post-hoc Methods



Post-hoc Methods - Feature Relevance

Computes a **relevance score** for the model's features.

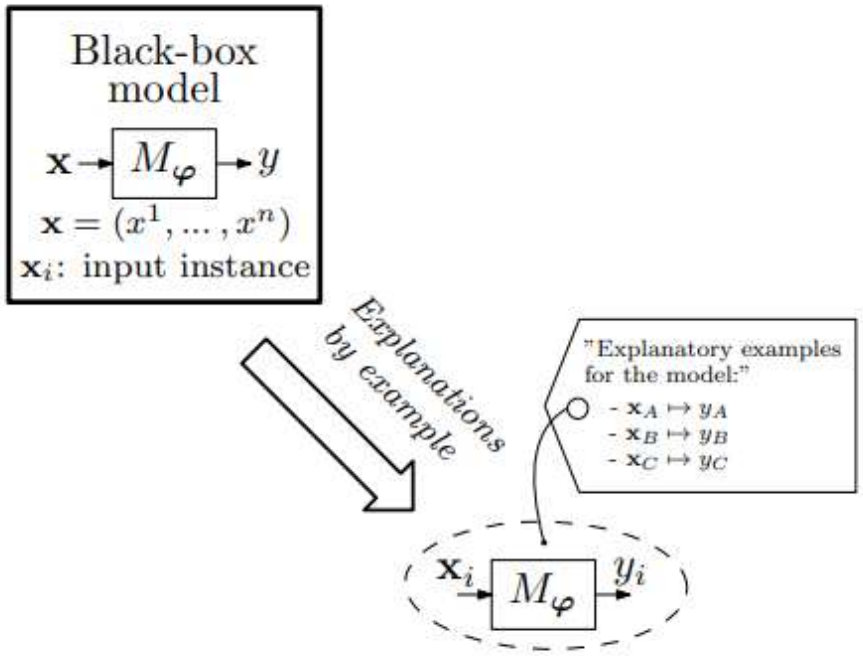
These scores quantify the **sensitivity** a feature has upon the output of the model.



Post-hoc Methods - Explanation by example

Extraction of **data examples** that relate to the result generated by a certain model

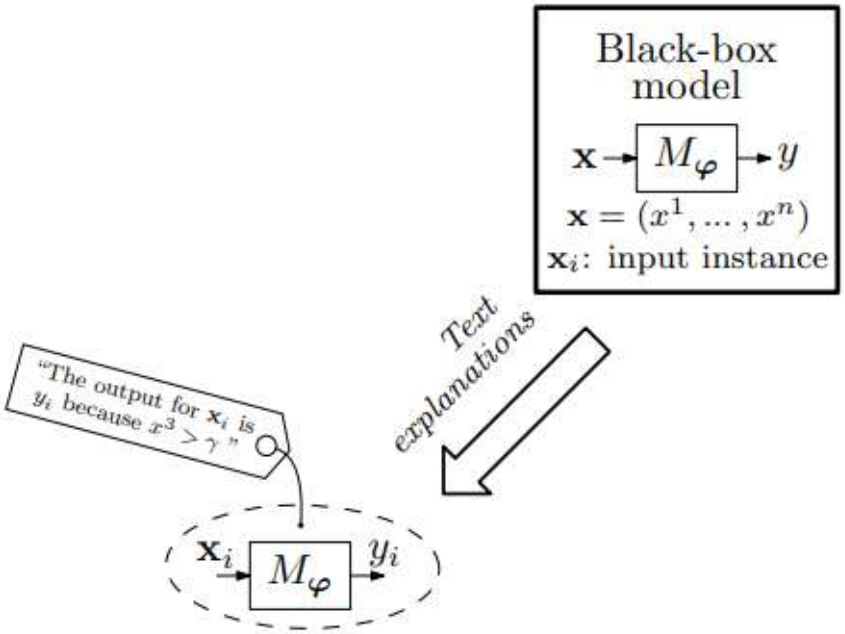
Find **representative** examples that grasp the inner relationships and correlations found by the model



Post-hoc Methods - Textual Explanations

The model **learns to generate** text explanations that help explaining the results from the model

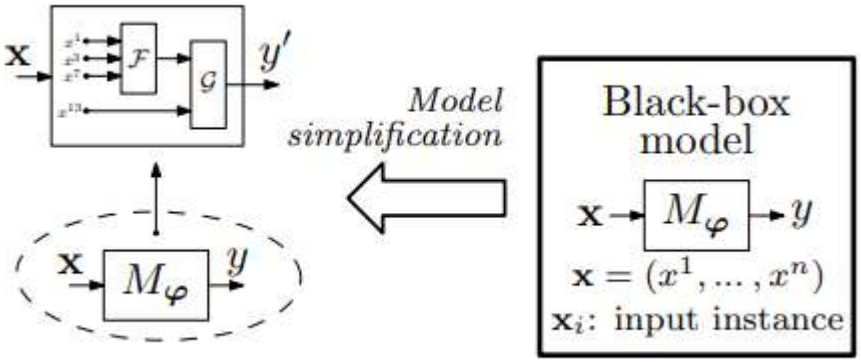
Generates **symbols** that represent the functioning of the model



Post-hoc Methods - Model Simplification

A whole **new system is rebuilt** based on the trained model to be explained

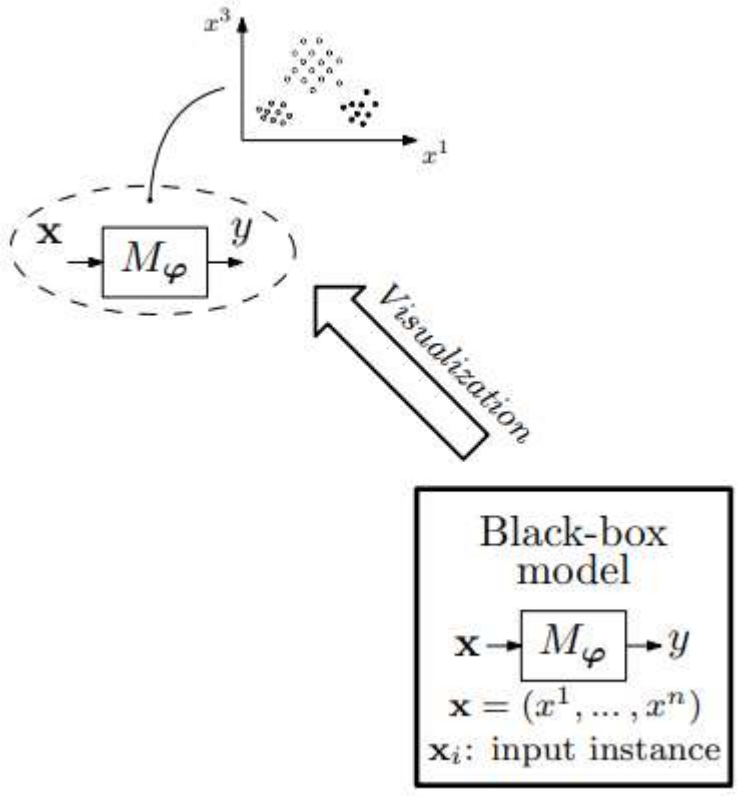
Simplified model attempts at **optimizing its resemblance** to its antecedent functioning



Post-hoc Methods - Visual Explanation

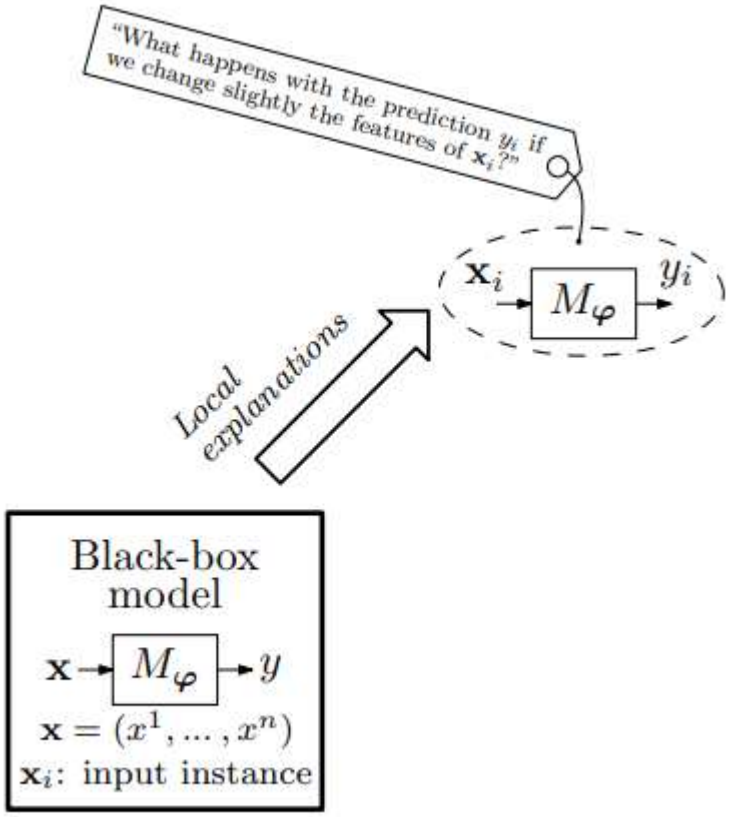
Visualizing the model's behavior

Dimensionality reduction techniques that allow for a human interpretable simple visualization



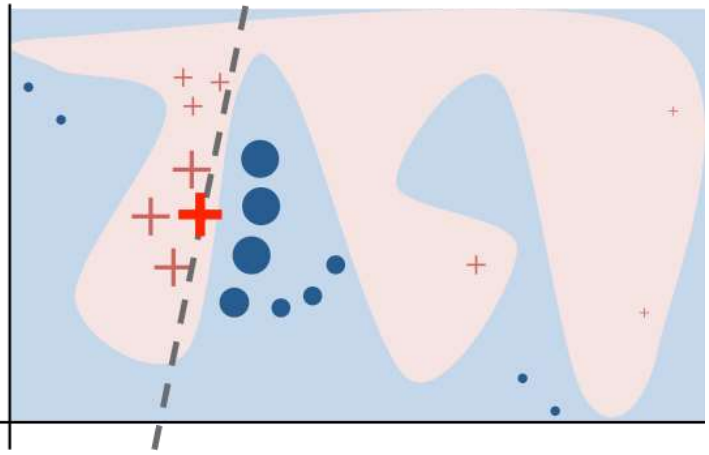
Post-hoc Methods - Local Explanation

Segments the solution space and giving explanations to **less complex solution subspaces** that are relevant for the whole model



Local Interpretable Model-Agnostic Explanations (LIME)

1. From a given input, generate a **new dataset** consisting of **perturbed samples near the input** and the corresponding predictions of the opaque model
2. On this new dataset LIME **trains** an **interpretable model**



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

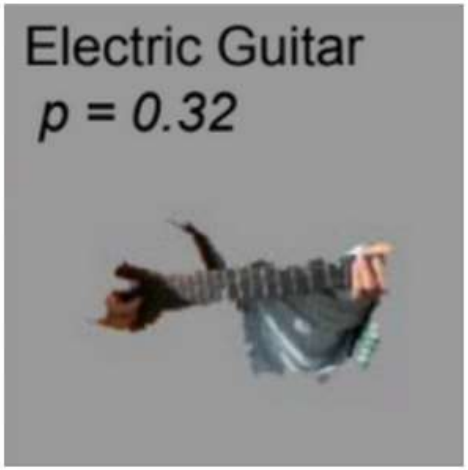
LIME

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

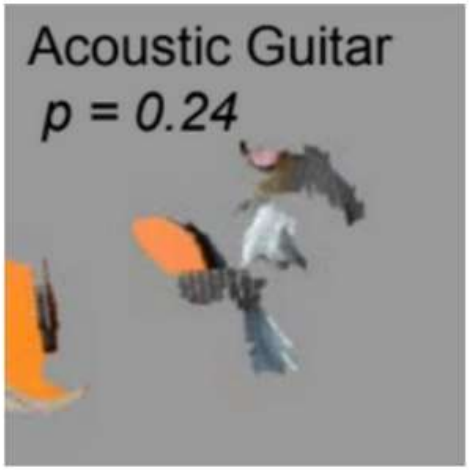
- $g \in G$: set of **transparent functions** (e.g. linear regression model)
- Loss :
  - $L(f, g, \pi_x)$: **closeness** between g and the opaque model f
    - π_x : **proximity measure** (complexity)
  - $\Omega(g)$: Model **complexity** (depth of the tree for decision trees number of non-zero weights for linear)

LIME

(a) Original image

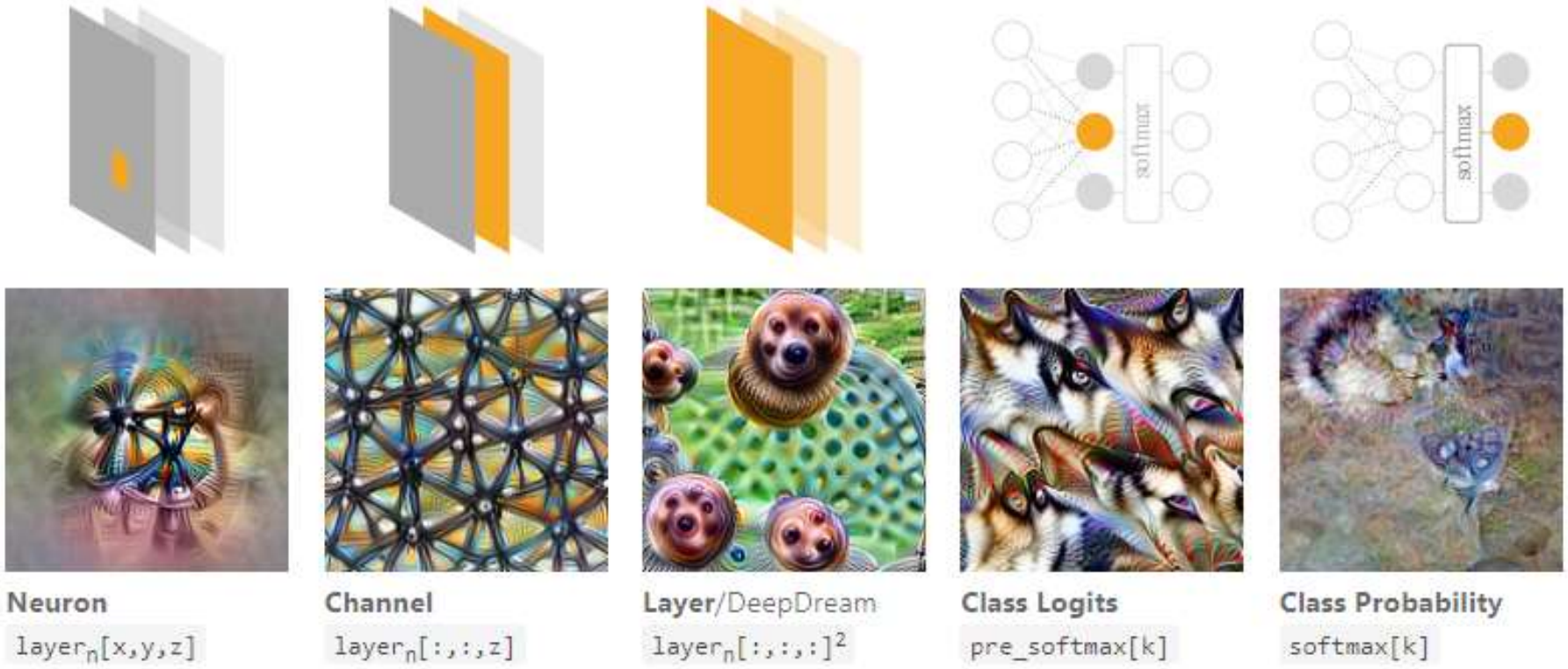


(b) Explaining *electric guitar*



(c) Explaining *acoustic guitar*

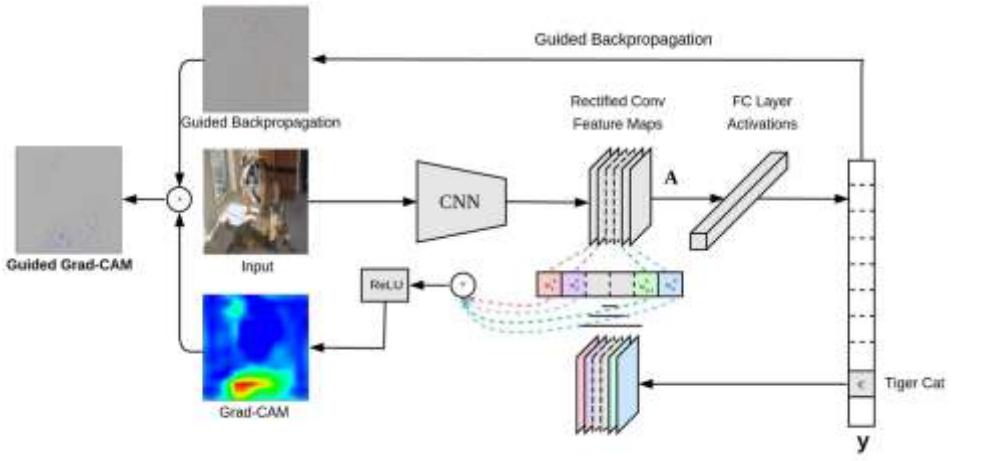
Feature Visualization



Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7

Gradient-weighted Class Activation Mapping (Grad-CAM)

1. Compute the **activation maps** given the input image
2. For a given class compute the **importances weights** by computing the average of the gradient for each features maps
3. Plot a **heatmap** from the linear combination between activations maps and importances weights



Selvaraju, et al.. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Gradient-weighted Class Activation Mapping (Grad-CAM)

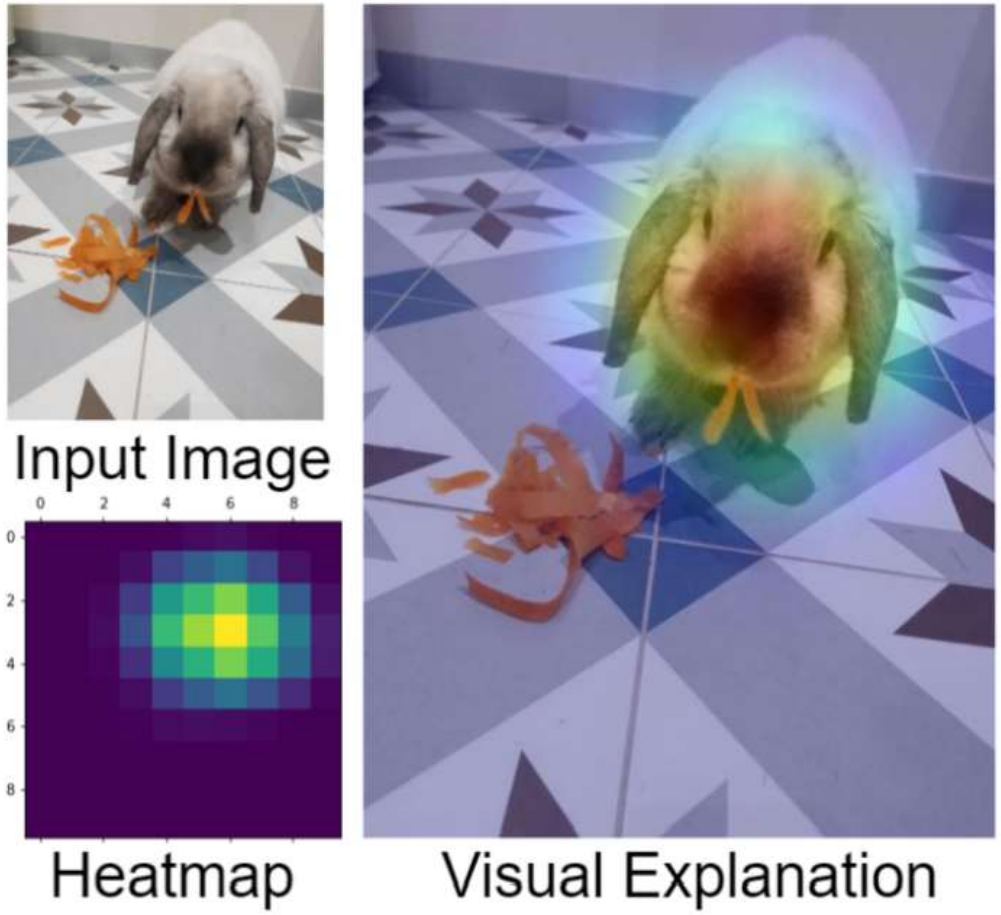
$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha^c_k A^k \right)}_{\text{linear combination}}$$

- A^k : **activations maps**

- α^c_k : **importance weight** of the activation map k

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{i,j}}$$

Gradient-weighted Class Activation Mapping (Grad-CAM)



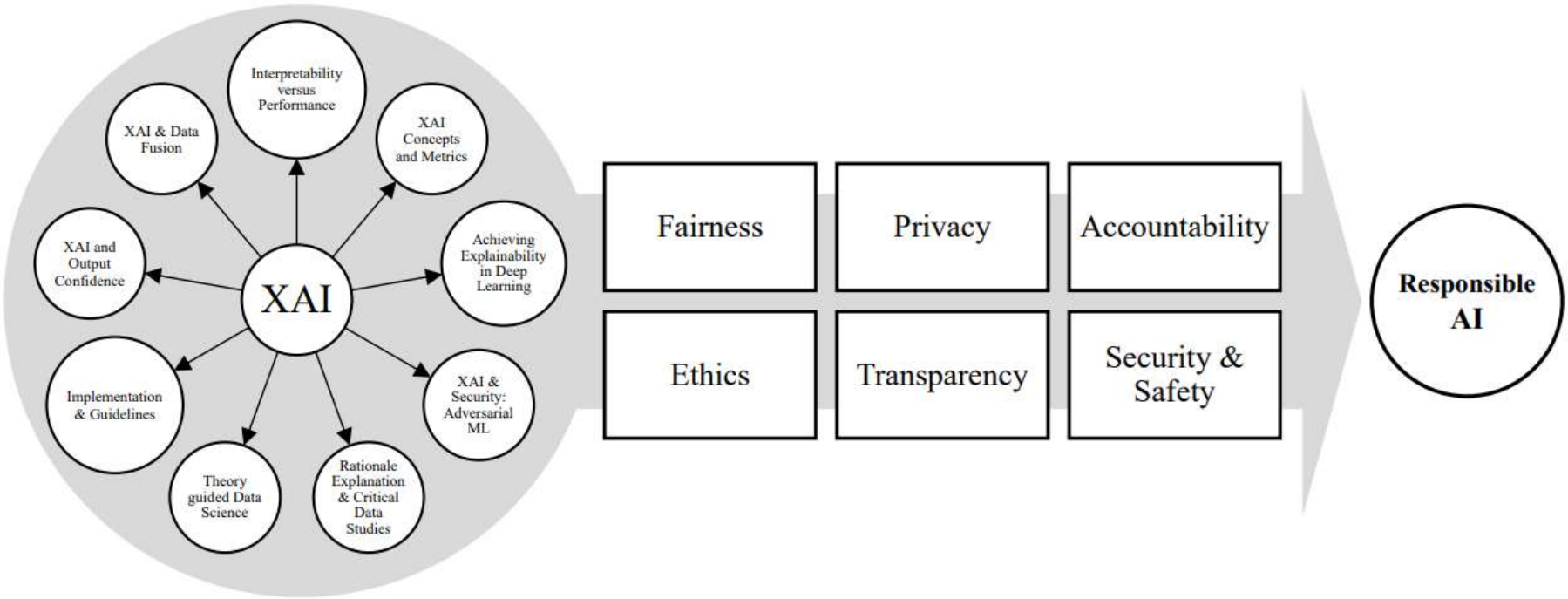
Open question about metrics

How to judge the **quality** of an explanation?

How to ensure that an explanation is **faithful** to what really happened inside the model?

How to ensure that an explanation is **valid** to what it should be?

XAI challenges



Bibliography

1. 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).
2. Olah, et al., "The Building Blocks of Interpretability", Distill, 2018.
3. Selvaraju, et al., Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
4. Bennetot, et al., A Practical Tutorial on Explainable AI Techniques
5. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI - AB Arrieta et al. - Information Fusion, 2020
6. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A. (2018). Women Also Snowboard: Overcoming Bias in Captioning Models. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11207. Springer, Cham. https://doi.org/10.1007/978-3-030-01219-9_47