

eXplainable AI: Why and How

Pierre Marquis

Univ. Artois, CNRS, CRIL
Institut Universitaire de France

Conférence des Jeunes Chercheurs en Génie Électrique, Le Croisic,
June 19th, 2024



A Short Introduction to XAI

The Various Facets of Explaining

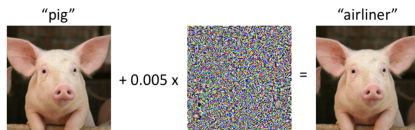
A Glimpse at Formal XAI

- ▶ AI systems empowered by ML techniques are **now all around us in everyday life**
- ▶ ML is also revolutionizing Science by its capacity to **automate discovery** in many domains (e.g., diagnosing a system, anticipating the occurrence of faults in a system, in a perspective of predictive maintenance, etc.)

- ▶ However, ML has a number of **limitations**
 - ▶ Ensuring 100% correct predictions: No way!
 - ▶ **Sensitivity to data** (quality, quantity), garbage in, garbage out...
 - ▶ The most accurate ML models are **black boxes (opacity)**
 - ▶ They suffer from a clear lack of **common-sense**

- ▶ However, ML has a number of **limitations**
 - ▶ Ensuring 100% correct predictions: No way!
 - ▶ **Sensitivity to data** (quality, quantity), garbage in, garbage out...
 - ▶ The most accurate ML models are **black boxes (opacity)**
 - ▶ They suffer from a clear lack of **common-sense**
- ▶ Those limitations must be kept in mind (outputs must be checked!)
- ▶ **“Guard rails” are needed** in many cases
- ▶ **Using the right AI tool for the right purpose** (a context-dependent issue)

Classification: $M : X \rightarrow Y$





Not a big deal...

- ▶ Very good friends, useful for many tasks!
- ▶ Always ready to help you



- ▶ Very good friends, useful for many tasks!
- ▶ Always ready to help you
- ▶ **But not very liable:** the outputs are often impressive ...
but they can be wrong nevertheless!
- ▶ Many limitations w.r.t.
 - ▶ Counting
 - ▶ Reasoning
 - ▶ Planning


May I Count on You?





 daisyedda 2  Feb 2023

ChatGPT failed to calculate $241 - (-241) + 1$.

 $241 - (-241) + 1$ 

 $241 - (-241) + 1$ is equivalent to $241 + 241 + 1$, which simplifies to $483 + 1$. So $241 - (-241) + 1$ is equal to 484.

[Taken from Gary Marcus' blog]

On a strange island, the inhabitants only express themselves through closed questions, i.e., whose answers are yes or no. Each resident is either a Positive or a Negative. Positives can only ask questions where the exact answer is yes, while Negatives can only ask questions where the exact answer is no. During a visit to this island, you meet an inhabitant, Arthur. He asks you the question: "Am I a Positive? ". What is Arthur's type?

Given that Arthur asks the question "Am I a Positive?" and considering the rules of the island's inhabitants:

- **Positives** can only ask questions that have the answer "yes."
- **Negatives** can only ask questions that have the answer "no."

Let's analyze Arthur's question "Am I a Positive?":

1. **If Arthur is a Positive:**

- He can only ask questions where the answer is "yes."
- Therefore, if Arthur is a Positive, the answer to his question "Am I a Positive?" must be "yes."

2. **If Arthur is a Negative:**

- He can only ask questions where the answer is "no."
- Therefore, if Arthur is a Negative, the answer to his question "Am I a Positive?" must be "no."

Given that Arthur asked "Am I a Positive?", the nature of the question itself can help us determine his type:

- If Arthur were a Negative, he would not be able to ask "Am I a Positive?" because the answer to that question would be "no," which violates the rule that Negatives can only ask questions with the answer "no."

Therefore, Arthur must be a Positive, as only a Positive could ask a question where the answer is "yes."

Conclusion:

Arthur is a Positive.

Planning: Playing with Cubes

A stack of two cubes A and B, with A on top on B, is on a table. The goal is to get a stack with B on top of A, by moving the cubes. Cube A cannot be moved. Please give me a solution plan.

Given the constraint that Cube A cannot be moved, here is a solution plan to achieve the goal of having Cube B on top of Cube A:

1. **Move Cube A from on top of Cube B to a temporary holding spot:**

- Since Cube A cannot be moved off the table, we need to find a way to temporarily hold Cube A without changing its position relative to the table.
- One way to do this is to lift Cube A slightly and place a temporary support (such as a small object or stack of smaller items) under Cube A to hold it in place while Cube B is moved.

2. **Remove Cube B from under Cube A:**

- With Cube A temporarily supported, slide Cube B out from under Cube A and place it next to the stack on the table.

3. **Lower Cube A back to its original position on the table:**

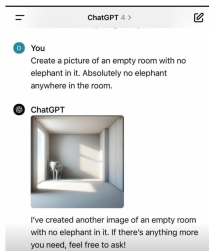
- Carefully remove the temporary support from under Cube A, allowing it to rest directly on the table again.

4. **Place Cube B on top of Cube A:**

- Finally, move Cube B from its temporary position on the table and place it on top of Cube A.

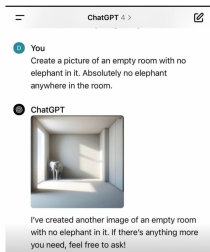
By following these steps, you will achieve the goal of having Cube B on top of Cube A without ever moving Cube A from its original position on the table.

► Handling negation properly



[Taken from Gary Marcus' blog]


► Handling negation properly



ChatGPT 4 >

You
Create a picture of an empty room with no elephant in it. Absolutely no elephant anywhere in the room.

ChatGPT

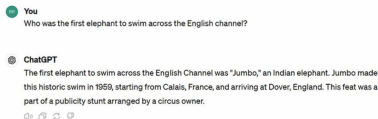


I've created another image of an empty room with no elephant in it. If there's anything more you need, feel free to ask!

[Taken from Gary Marcus' blog]

► Connection to truth

ChatGPT 4o ▾



You
Who was the first elephant to swim across the English channel?

ChatGPT
The first elephant to swim across the English Channel was "Jumbo," an Indian elephant. Jumbo made this historic swim in 1950, starting from Calais, France, and arriving at Dover, England. This feat was a part of a publicity stunt arranged by a circus owner.

👍 🗨️ 🔄 🗑️

[Taken from Gary Marcus' blog]

- ▶ **Watson Health**
(diagnosing diseases, pointing out treatments)
 - ▶ Sold a couple of years ago!
 - ▶ No use of medical expertise as such
 - ▶ No explanation facilities
 - ▶ Training set and test set too close

- ▶ **Watson Health**
(diagnosing diseases, pointing out treatments)
 - ▶ Sold a couple of years ago!
 - ▶ No use of medical expertise as such
 - ▶ No explanation facilities
 - ▶ Training set and test set too close
- ▶ **PredPol** in LA (predictive policing)
 - ▶ Determine where to send the police in order to reduce crime
 - ▶ Based on a very debatable analogy between crime and earthquake predictions
 - ▶ Self fulfilling prophecy

[Angwin and Larson, "Machine Bias", ProPublica, 2016]

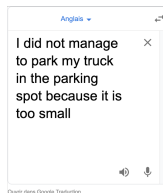
- ▶ **Correctional Offender Management Profiling Alternative Sanctions (COMPAS)**
 - ▶ Assess the likelihood of a defendant becoming a recidivist
 - ▶ **Biased data:**
 - ▶ Blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend
 - ▶ Whites are much more likely than blacks to be labeled lower-risk but go on to commit other crimes

[Angwin and Larson, "Machine Bias", ProPublica, 2016]

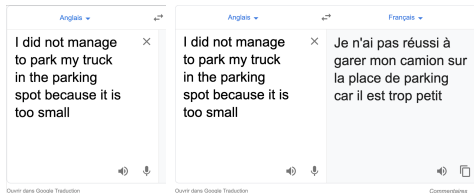
- ▶ **Correctional Offender Management Profiling Alternative Sanctions (COMPAS)**
 - ▶ Assess the likelihood of a defendant becoming a recidivist
 - ▶ **Biased data:**
 - ▶ Blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend
 - ▶ Whites are much more likely than blacks to be labeled lower-risk but go on to commit other crimes
- ▶ **Social credit** in PRC: the Dong Mingzhu case



- ▶ *"I did not manage to park my truck in the parking spot because **it** is too small"*
- ▶ Translation to French using Google Translate



- ▶ *"I did not manage to park my truck in the parking spot because **it** is too small"*
- ▶ Translation to French using Google Translate



- ▶ Traffic laws & common-sense



- ▶ Traffic laws & common-sense



- ▶ Ethical issues



- ▶ ...

- ▶ People will use ML-based AI systems only if they can trust in them
- ▶ Especially, trustworthy AI is mandatory for high-risk or safety-critical applications

- ▶ People will use ML-based AI systems only if they can trust in them
- ▶ Especially, trustworthy AI is mandatory for high-risk or safety-critical applications
- ▶ Trustworthy AI has **a number of facets** (interpretability, explainability, transparency, confidentiality, fairness, reliability, safety, etc.)

- ▶ People will use ML-based AI systems only if they can trust in them
- ▶ Especially, trustworthy AI is mandatory for high-risk or safety-critical applications
- ▶ Trustworthy AI has **a number of facets** (interpretability, explainability, transparency, confidentiality, fairness, reliability, safety, etc.)
- ▶ Explaining the decisions made became **a legal issue** in a number of countries, especially in Europe (General Data Protection Regulation – GDPR – since May 2018, European AI Act since December 2023, etc.)

XAI is the part of Trustworthy AI focusing on the **interpretability and explainability** issues

DARPA, at the origin of the buzz word “XAI”, pointed out the following purpose for XAI in 2019:

“to provide users with **explanations** that enable them to understand the system’s overall strengths and weaknesses, convey an **understanding** of how it will behave in future or different situations, and perhaps permit users to **correct** the system’s mistakes”

XAI is the part of Trustworthy AI focusing on the **interpretability and explainability** issues

DARPA, at the origin of the buzz word “XAI”, pointed out the following purpose for XAI in 2019:

“to provide users with **explanations** that enable them to understand the system’s overall strengths and weaknesses, convey an **understanding** of how it will behave in future or different situations, and perhaps permit users to **correct** the system’s mistakes”

As human beings, a truly intelligent system **should not persist in error**

Designing methods for

- ▶ **Allowing users to understand the behaviour of the AI system at hand**
 - ▶ Explaining its global behaviour
 - ▶ Understanding how it has been conceived (and trained), and how to use it
- ▶ **Verifying / inspecting the AI system**
 - ▶ Testing whether it behaves as the user expects
 - ▶ Entering (to some extent) into the black box

Designing methods for

- ▶ Explaining the predictions made using the AI system
 - ▶ Making the generation of the outputs of the system intelligible
 - ▶ Providing explanations suited to the explainee (human in the loop!)
- ▶ Correcting the AI system when it goes wrong
- ▶ Ultimately, making the user able to decide whether to trust (or not to trust) the AI system or the predictions made

A Short Introduction to XAI

The Various Facets of Explaining

A Glimpse at Formal XAI

- ▶ **Global explanations**
 - ▶ Explaining the way the ML model **has been learnt**
 - ▶ Making precise the way the ML model can be **used to classify** instances
 - ▶ Ability **to derive classification rules, aka decision sets, in an efficient way**
- ▶ **Local explanations**
 - ▶ About **instances**
 - ▶ Identifying the **importance of the characteristics of the instance** at hand in the prediction achieved
 - ▶ Many definitions of an explanation (offering more or less guarantees) exist

- ▶ Explanations depend on how instances are described
- ▶ Explanations are often based on **descriptors of the same types as instances**
 - ▶ **Subsymbolic information** (raw data, e.g., pixels in a picture, features in a speech signal): **no intrinsic meaning**
 - ▶ **Symbolic information** (e.g., logical formulae, tabular data, attribute/value pairs): **concept-based** representations, can be viewed as pieces of **transferable knowledge**, paves the way to **reasoning from explanations** (especially, correcting the model)

A **large variety**, in terms of

- ▶ **Nature**
 - ▶ model-agnostic
 - ▶ model-based
 - ▶ ...
- ▶ **Type**
 - ▶ abductive
 - ▶ contrastive
 - ▶ ...
- ▶ **Format**
 - ▶ graphical
 - ▶ textual
 - ▶ logical
 - ▶ ...

Estimating the importance of each feature of the instance in the prediction made

- ▶ Typically suited to explaining predictions derived from raw data using black-box models
- ▶ Perturbation-based feed forward approaches (with occlusion)
- ▶ Gradient-based backpropagation approaches
- ▶ Heat (saliency) maps
- ▶ Many popular XAI techniques (LIME, Anchor, ...)
- ▶ ...

Explaining How a Picture is Classified



(a) Original Image

(b) Explaining *Electric guitar*

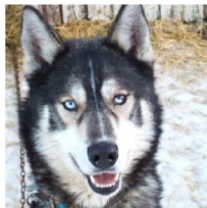
(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*

[Ribeiro et al., ACM SIGKDD'16]

- ▶ The explainer (here LIME) determines which pixels are important in the classification made
- ▶ Feature importance can be displayed as **saliency maps** when dealing with images
- ▶ **The interpretation of the explanation is achieved by the explainee**
- ▶ **No concepts** (e.g., fretboard) are involved in the explanation!









Explanations are sometimes unexpected, reflecting learning issues (biased data)



[Ribeiro et al., ACM SIGKDD'16]

- ▶ **Model-agnostic approaches to XAI: dealing with a surrogate model**
 - ▶ LIME [Ribeiro et al., ACM SIGKDD'16]
 - ▶ LORE [Guidotti et al., ACM Computing Surveys (2019)]
 - ▶ Anchor [Ribeiro et al., AAAI'18]
- ▶ Local explanations are **computed heuristically**
- ▶ Those approaches are **scalable** but they ensure **no guarantee** w.r.t. the underlying ML model
[Garreau and von Luxburg, AISTATS'20] [Narodytska et al., SAT'19]
- ▶ Especially, two instances associated with distinct classes **can share the same explanation!** [Ignatiev, IJCAI'20]

Co-12 Properties

Correctness Match between model and explanation. 	Completeness How much of the model is explained? 	Consistency Robustness to small changes in model and implementation. $g(x) = g(x)$	Continuity Robustness to small changes input. $g(x) = g(x')$
Contrastivity Discriminative to other events or targets? $g(x Cat) \neq g(x Dog)$	Covariate Complexity Complexity of features in the explanation 	Compactness Size of the explanation 	Composition Presentation format 
Confidence Probability information available? $p = ?$	Context Useful for users? 	Coherence Match with domain knowledge. $g(x) =$ 	Controllability Can user influence explanation? $g(x)$ 

Explanation / Model / User

[Nauta et al., ACM Computing Survey 2023]

- ▶ Many criteria must be taken into account
- ▶ Some of them are user-dependent
- ▶ Some criteria are antagonistic

A Short Introduction to XAI

The Various Facets of Explaining

A Glimpse at Formal XAI

- ▶ Associating a circuit C equivalent to the ML model M in terms of inputs/outputs
- ▶ Delegating XAI queries to the circuit C
- ▶ **Ensuring correctness**, which is of the utmost value when dealing with safety-critical applications: model-based explanations are faithful by design

[Breiman et al 1984; Quinlan, 1986]

- ▶ Decision trees are both **classifiers AND representation languages (Boolean circuits)**
- ▶ Almost not effort is needed to go from the model **M** to an equivalent circuit **C**
- ▶ Decision trees are a key component of more sophisticated ML models (random forests, boosted trees) that are SOTA for learning from tabular data

A Toy Example: The Flower Power



f		x_3x_4			
		00	01	10	11
x_1x_2	00	0	0	0	0
	01	0	0	0	0
	10	0	0	0	1
	11	1	1	1	1

Recognizing **common hollyhocks** (*alcea rosea*) from other roses using the following Boolean features:

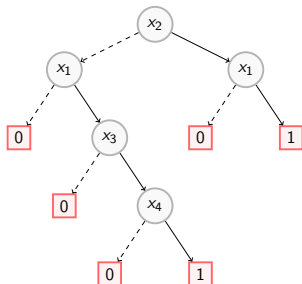
- ▶ x_1 : “has a deciduous foliage”
- ▶ x_2 : “has heart-shaped leaves”
- ▶ x_3 : “has large flowers”
- ▶ x_4 : “has a light green stem”

In general, tabular instances are described using **pairs feature/value** involving **several types of features**:

- ▶ **Boolean features** (values in $\{0, 1\}$)
- ▶ **Numerical features** (values are numbers: integers, rationals, real numbers in floating point format, etc.)
- ▶ **Categorical features** (values are taken from a finite domain, that is not “naturally ordered”, e.g., $\{red, blue, green\}$)

A Toy Example: The Flower Power

- ▶ $X = \{x_1, x_2, x_3, x_4\}$ (Boolean features)
- ▶ $Y = \{y\}$ (Boolean label: 1 for common hollyhocks)
- ▶ $M =$



x_1 : "has a deciduous foliage" x_2 : "has heart-shaped leaves" x_3 : "has large flowers" x_4 : "has a light green stem"

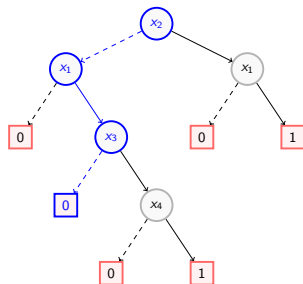
The learning issue

- ▶ The starting point is a **dataset**, i.e., a set of training instances where each instance is associated with a prediction (typically, a class, a value)
- ▶ Leaves of decision trees are **prediction labels** (class identifiers, numerical values)
- ▶ Internal nodes of decision trees are **decision nodes** labelled by Boolean conditions of the form $A > v$ or $A = v$ where v is an element of the domain of feature A
- ▶ Top-down learning algorithms determine decision nodes using **information theory criteria**
- ▶ To **avoid overfitting**, the depth of the trees is either bounded *a priori* or a post-pruning phase is achieved once the tree has been learned

How to Make a Prediction using a Decision Tree?

The inference issue

Is $\mathbf{x} = (1, 0, 0, 0)$ a common hollyhock?



No, since $M(\mathbf{x}) = 0$

x_1 : "has a deciduous foliage" x_2 : "has heart-shaped leaves" x_3 : "has large flowers" x_4 : "has a light green stem"

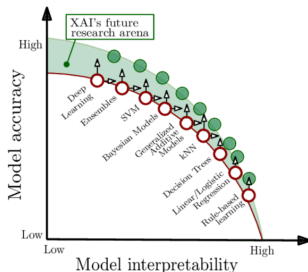
- ▶ **Explanation queries:** Explaining why x has been classified by M as such, or not classified by M as expected [Ignatiev et al., CoRR'20]
 - ▶ Abductive explanations (explain “Why?”): find a subset t of the characteristics of x such that every instance x' satisfying t is such that $M(x') = M(x)$
 - ▶ Contrastive explanations (explain “Why not?”): find how to modify a subset t of the characteristics of x so as to get an instance x' such that $M(x') \neq M(x)$

- ▶ **Explanation queries:** Explaining why x has been classified by M as such, or not classified by M as expected [Ignatiev et al., CoRR'20]
 - ▶ Abductive explanations (explain “Why?”): find a subset t of the characteristics of x such that every instance x' satisfying t is such that $M(x') = M(x)$
 - ▶ Contrastive explanations (explain “Why not?”): find how to modify a subset t of the characteristics of x so as to get an instance x' such that $M(x') \neq M(x)$
- ▶ **Inspection / verification queries:** Determining the extent to which the predictions made by M comply with the expectations of the user

- ▶ **Explanation queries:** Explaining why x has been classified by M as such, or not classified by M as expected [Ignatiev et al., CoRR'20]
 - ▶ Abductive explanations (explain “Why?”): find a subset t of the characteristics of x such that every instance x' satisfying t is such that $M(x') = M(x)$
 - ▶ Contrastive explanations (explain “Why not?”): find how to modify a subset t of the characteristics of x so as to get an instance x' such that $M(x') \neq M(x)$
- ▶ **Inspection / verification queries:** Determining the extent to which the predictions made by M comply with the expectations of the user
 - ▶ Relevance of a feature
 - ▶ Monotonicity of a feature
 - ▶ Proportion of instances in a given class sharing some characteristics
 - ▶ ...

The Accuracy/Interpretability Trade-Off

- ▶ To which extent are decision trees interpretable?
- ▶ Are they really “interpretable by design”?

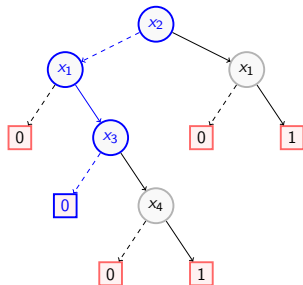


[Barredo Arrieta et al., Information Fusion (2020)]

- ▶ What does it mean? How to make it formal?

“Decision Trees are Interpretable by Design”?

Each path of a decision tree corresponds to a classification rule



The blue path in the tree corresponds to the rule $(\bar{x}_2 \wedge x_1 \wedge \bar{x}_3) \Rightarrow \bar{y}$

- ▶ **Global interpretability**

- ▶ Quid if the number of paths (leaves) is quite large? E.g. > 100 ?
- ▶ Quid if the depth of the tree is quite large? E.g. > 100 ?
- ▶ Global interpretability is **debatable** because it depends on the explainee (subjective evaluation)

▶ Global interpretability

- ▶ Quid if the number of paths (leaves) is quite large? E.g. > 100 ?
- ▶ Quid if the depth of the tree is quite large? E.g. > 100 ?
- ▶ Global interpretability is **debatable** because it depends on the explaineer (subjective evaluation)

▶ Local interpretability

- ▶ The path followed for inferring a prediction explains the prediction made for the instance at hand
- ▶ It corresponds to an abductive explanation for the instance
- ▶ However, it may contain arbitrarily many redundant characteristics [Izza et al., CoRR 2022]
- ▶ On the running example, the abductive explanation $\bar{x}_2 \wedge x_1 \wedge \bar{x}_3$ for \mathbf{x} is redundant ($\bar{x}_2 \wedge \bar{x}_3$ also is an abductive explanation for \mathbf{x})

Towards a **formal definition of interpretability**: viewing families of ML models as **representations languages** [Audemard et al., KR'20]

- ▶ **Identifying XAI queries** (explanation and verification) of interest
- ▶ The set of XAI queries to be considered is **user-dependent**
- ▶ Determining those queries that are **tractable** (and those that are not)
- ▶ The computational interpretability of an ML model is defined as **the set of tractable XAI queries it supports**
- ▶ **Choosing an ML model** accordingly
(taking into account its predictive performance as well)

Decision Trees are Computationally Interpretable Models...

Not only because of the existence of an abductive explanation that is easy to compute (path-restricted explanation, direct reason)

[Izza et al., CoRR 2010] [Audemard et al., KR'21]

- ▶ Removing redundant characteristics from the direct reason to derive an irredundant abductive explanation (aka a sufficient reason) is tractable
- ▶ Computing all subset-minimal contrastive explanations is tractable [Huang et al., KR'21] [Audemard et al., DKE 2022]
- ▶ Many inspection / verification queries are tractable as well

- ▶ [Audemard et al., KR'21]
Many explanation queries (e.g., computing a sufficient reason) and verification / inspection queries of interest are **NP-hard** in the broad sense when the classifier is a decision list, a random forest, a boosted tree, a multilayer perceptron, a binarized neural network, etc.
- ▶ **NP-hard** means that efficient (i.e., polynomial-time) algorithms to answer those queries are unlikely
- ▶ This does not mean that no answers can be obtained in any case, but that there is **no guarantee** that the algorithms will be efficient enough in practice
- ▶ **Experiments are needed** to evaluate the extent to which those algorithms are practical enough

No, **limitations exist**

- ▶ An instance may have exponentially many sufficient reasons, and even exponentially many sufficient reasons of minimal size
[Audemard et al., IJCAI'22]
- ▶ Such reasons may heavily differ one another (no shared features)
[Audemard et al., DKE 2022]
[Izza et al., CoRR 2022]
- ▶ Sufficient reasons (even those of minimal size) can be too large to be considered as intelligible
- ▶ Relaxing the irredundancy condition of sufficient reason to consider only a probabilistic one changes the picture, computationally speaking

- ▶ Computing a single sufficient reason is not enough in general
- ▶ An output-polynomial enumeration algorithm for sufficient reasons is **unlikely** (it would give an output-polynomial enumeration algorithm for the minimal transversals of hypergraphs) [de Colnet and M., IJCAI'22]
- ▶ Computing a minimum-size abductive explanation is **NP-hard** [Barceló et al., NeurIPS'20]
- ▶ Computing a probabilistic sufficient reason is **NP-hard** [Arenas et al., NeurIPS'22]
- ▶ One can nevertheless “synthesize” the set of sufficient reasons when dealing with decision trees [Audemard et al., DKE 2022]

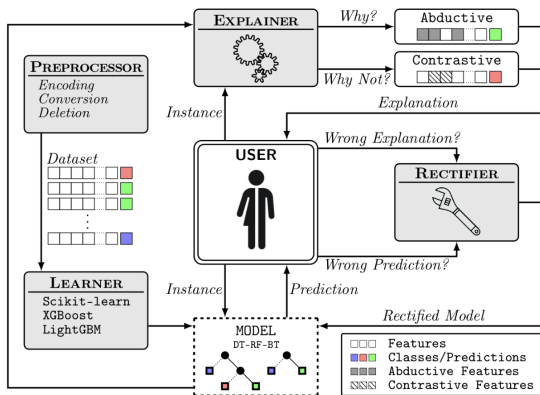
Theory vs. practice: computing useful explanations is often feasible in practice for tree-based models

- ▶ Computing **preferred** sufficient reasons for decision trees (and preferred abductive explanations for random forests) [Audemard et al., AAAI'22]
- ▶ Computing **example-based** abductive explanations for random forests [Audemard et al., IJCAI'24]
- ▶ Computing abductive explanations for **boosted trees** [Audemard et al., AISTATS'23]
- ▶ Computing abductive explanations when dealing with **regression problems** [Audemard et al., IJCAI'23]
- ▶ Computing contrastive explanations for **random forests** [Audemard et al., ECAI'23]

www.cril.univ-artois.fr/expekctation/

- ▶ EXPEKCTATION is an acronym for “EXPLainable artificial intelligence: a KnowlEdge CompilaTion FoundATIOn”
- ▶ It is the name of a research and teaching chair in AI (ANR-19-CHIA-0005-01), funded by ANR, the French Agency for Research (2020-2025)
- ▶ The objective is the the **development of approaches to eXplainable AI for interpretable and robust machine learning**, using constraint-based automated reasoning methods, in particular knowledge compilation

www.cril.univ-artois.fr/pyxai/



User interaction with PyXAI.

eXplainable AI: Why and How

Pierre Marquis

Univ. Artois, CNRS, CRIL
Institut Universitaire de France

Conférence des Jeunes Chercheurs en Génie Électrique, Le Croisic,
June 19th, 2024

