

A Pinch of XAI from a KR Perspective

Pierre Marquis
and the EXPEKCTATION team
(Gilles Audemard, Steve Bellart, Louenas Bounia,
Frédéric Koriche, Jean-Marie Lagniez)

CRIL, Université d'Artois & CNRS
Institut Universitaire de France

Perspectives et défis de l'IA : journée "explicabilité",
organisée par l'AFIA, April 8th 2021



- ▶ Progress in ML techniques has **revolutionized** vision, speech, language understanding, and other fields for the past decade

- ▶ Progress in ML techniques has **revolutionized** vision, speech, language understanding, and other fields for the past decade
- ▶ **Classification** (defining $\mathbf{C} : \mathbf{X} \rightarrow \mathbf{Y}$ from $\mathcal{T} \subseteq \mathbf{X} \times \mathbf{Y}$) is a major task

- ▶ Progress in ML techniques has **revolutionized** vision, speech, language understanding, and other fields for the past decade
- ▶ **Classification** (defining $\mathbf{C} : \mathbf{X} \rightarrow \mathbf{Y}$ from $\mathcal{T} \subseteq \mathbf{X} \times \mathbf{Y}$) is a major task
- ▶ Many **families of predictors** (alias ML models) have been investigated so far
 - ▶ Decision trees
 - ▶ Decision lists
 - ▶ Random forests
 - ▶ Bayes nets
 - ▶ Neural networks (of many different types)
 - ▶ ...

- ▶ Progress in ML techniques has **revolutionized** vision, speech, language understanding, and other fields for the past decade
- ▶ **Classification** (defining $\mathbf{C} : \mathbf{X} \rightarrow \mathbf{Y}$ from $\mathcal{T} \subseteq \mathbf{X} \times \mathbf{Y}$) is a major task
- ▶ Many **families of predictors** (alias ML models) have been investigated so far
 - ▶ Decision trees
 - ▶ Decision lists
 - ▶ Random forests
 - ▶ Bayes nets
 - ▶ Neural networks (of many different types)
 - ▶ ...

- ▶ However, efficient predictors are often **black boxes**

- ▶ However, efficient predictors are often **black boxes**
- ▶ This is **an issue for a number of applications** (e.g., in medicine)
 - ▶ The classifier should **explain the predictions made**:
*“Hey, C , you told me **that** $C(\mathbf{x}) = \mathbf{y}$, but please tell me **why** $C(\mathbf{x}) = \mathbf{y}$!”*

- ▶ However, efficient predictors are often **black boxes**
- ▶ This is **an issue for a number of applications** (e.g., in medicine)
 - ▶ The classifier should **explain the predictions made**:
*“Hey, C , you told me **that** $C(\mathbf{x}) = \mathbf{y}$, but please tell me **why** $C(\mathbf{x}) = \mathbf{y}$!”*
 - ▶ The classifier should be **amenable to inspection** (e.g., ensuring that the predictions made are not biased is expected)

- ▶ However, efficient predictors are often **black boxes**
- ▶ This is **an issue for a number of applications** (e.g., in medicine)
 - ▶ The classifier should **explain the predictions made**:
*“Hey, C , you told me **that** $C(x) = y$, but please tell me **why** $C(x) = y!$ ”*
 - ▶ The classifier should be **amenable to inspection** (e.g., ensuring that the predictions made are not biased is expected)
- ▶ The ability of providing explanations is **required in Europe** since May 2018 (GDPR, Recital 71)

- ▶ However, efficient predictors are often **black boxes**
- ▶ This is **an issue for a number of applications** (e.g., in medicine)
 - ▶ The classifier should **explain the predictions made**:
*"Hey, C , you told me **that** $C(x) = y$, but please tell me **why** $C(x) = y$!"*
 - ▶ The classifier should be **amenable to inspection** (e.g., ensuring that the predictions made are not biased is expected)
- ▶ The ability of providing explanations is **required in Europe** since May 2018 (GDPR, Recital 71)
- ▶ **The XAI field: explaining predictions, verifying predictors**
- ▶ A **major topic** in AI for a couple of years

- ▶ Explanations take much of the time a **symbolic** form: they are based on **concepts** expressed in some language
- ▶ Explaining is basically a **multi-faceted reasoning activity** (abduction, diagnosis, postdiction, goal regression, etc.)
- ▶ Explaining is a **social process**, a model of the explainee (the concepts she knows, the beliefs she has, etc.) must be taken into account
- ▶ Human beings have **limited knowledge** and are **not perfect reasoners** (the structure, the size, the concepts used in explanations make them more or less intelligible)

- ▶ Reasoning about knowledge in theory ... and in practice!

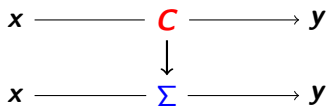
- ▶ Reasoning about knowledge **in theory ... and in practice!**
- ▶ Much progress in SAT solving for the past 20 years
- ▶ Used in AI and outside AI (formal verification, software engineering, etc.)
- ▶ Can be leveraged for **solving computationally harder problems**
- ▶ The era of **deep solving** (alias beyond NP) has got started

- ▶ A major topic of the research project developed at CRIL for 2020-2024

- ▶ A major topic of the research project developed at CRIL for 2020-2024
- ▶ The ANR AI chair EXPEKCTATION (started from September 2020)
 - ▶ www.cril.fr/expekctation/
 - ▶ Leveraging KR techniques (especially knowledge compilation) for XAI
 - ▶ From the theory side to the practical side

- ▶ A major topic of the research project developed at CRIL for 2020-2024
- ▶ The ANR AI chair EXPEKCTATION (started from September 2020)
 - ▶ www.cril.fr/expekctation/
 - ▶ Leveraging KR techniques (especially knowledge compilation) for XAI
 - ▶ From the theory side to the practical side
- ▶ The TAILOR project (“Trustworthy AI - Integrating Learning, Optimisation and Reasoning”), an H2020 ICT-48 European network of AI excellence centres

- ▶ **Key observation:** XAI tasks about a predictor \mathbf{C} can be delegated to a circuit $\Sigma \in \mathcal{L}$ exhibiting the same input-output behaviour as \mathbf{C}



- ▶ In this approach \mathbf{C} has been learnt first (both its hyper-parameters and its parameters are set)
- ▶ Boolean circuits or arithmetic circuits Σ can be targeted
- ▶ The translation from \mathbf{C} to Σ is done **once for all**: the same Σ can be used for all the $\mathbf{x} \in \mathbf{X}$

- ▶ **Defining encodings** to go from \mathbf{C} to Σ for several families of classifiers
- ▶ **Defining XAI queries** of interest
- ▶ Identifying the **computational complexities** of those queries depending on the language \mathcal{L} used to represent Σ
- ▶ **Showing how the XAI queries** can be addressed by combining queries and transformations over Boolean circuits Σ
- ▶ Exhibiting sufficient conditions on \mathcal{L} for **ensuring tractability** of XAI queries
- ▶ Pointing out KC languages \mathcal{L} **satisfying those conditions**
- ▶ Designing techniques to derive **intelligible explanations**

Several encodings have been defined so far to associate classifiers from several families with Boolean or arithmetic circuits

- ▶ Decision trees
- ▶ Random forests
- ▶ Bayes nets
- ▶ Binary neural networks
- ▶ ...

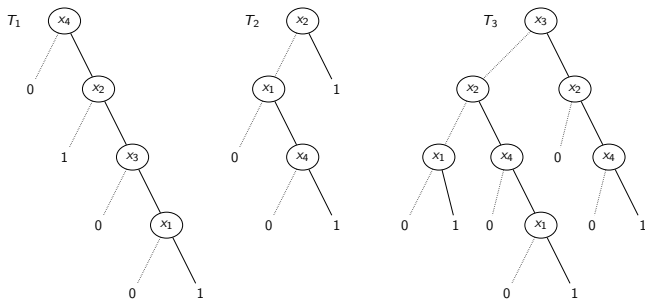


Recognizing *Cattleya* orchids using the following features:

- ▶ x_1 : "has fragrant flowers"
- ▶ x_2 : "has one or two leaves"
- ▶ x_3 : "has large flowers"
- ▶ x_4 : "is sympodial"
- ▶ x_5 : "has white flowers"

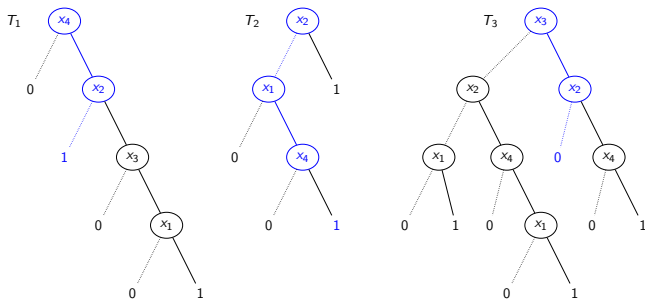
A Toy Example: The Flower Power

- ▶ $X = \{x_1, x_2, x_3, x_4, x_5\}$ (Boolean features)
- ▶ $Y = \{y\}$ (Boolean label: 1 for Cattleya orchids)
- ▶ $C = \{T_1, T_2, T_3\}$ (random forest)



x_1 : "has fragrant flowers" x_2 : "has one or two leaves" x_3 : "has large flowers" x_4 : "is sympodial" x_5 : "has white flowers"

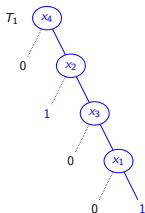
Is $\mathbf{x} = (1, 0, 1, 1, 1)$ a Cattleya orchid?



Yes, $\mathbf{C}(\mathbf{x}) = 1$ since 2 decision trees (T_1, T_2) of \mathbf{C} (out of 3) agrees with it

x_1 : "has fragrant flowers" x_2 : "has one or two leaves" x_3 : "has large flowers" x_4 : "is sympodial" x_5 : "has white flowers"

- ▶ **Introducing auxiliary variables:** One per class plus one per class and decision tree (here, 3 new variables)
- ▶ **Encoding each decision tree of C**



$$y^1 \Leftrightarrow ((\bar{x}_2 \wedge x_4) \vee (x_1 \wedge x_2 \wedge x_3 \wedge x_4))$$

...

- ▶ **Encoding majority voting:** $y \Leftrightarrow (y^1 + y^2 + y^3 \geq 2)$

x_1 : "has fragrant flowers" x_2 : "has one or two leaves" x_3 : "has large flowers" x_4 : "is sympodial" x_5 : "has white flowers"

- ▶ **Explanation queries:** explaining why x has been classified by C as such, or not classified by C as expected
- ▶ **Verification queries:** determining the extent to which classes as identified by C comply with the expectations of the user

- ▶ **Explanation queries**
 - ▶ Computing sufficient reasons
 - ▶ Computing counterfactual (contrastive) explanations
 - ▶ ...
- ▶ **Verification queries**
 - ▶ Identifying irrelevant features for a given class
 - ▶ Identifying mandatory / forbidden features for a given class
 - ▶ Identifying monotone features for a given class
 - ▶ Measuring the frequency of features in a given class
 - ▶ Counting the instances associated with a given class
 - ▶ Measuring how much classes are close to each other
 - ▶ ...

► Sufficient reasons

- A **sufficient reason** for \mathbf{x} given \mathbf{C} is a minimal subset t of the characteristics of \mathbf{x} such that every instance \mathbf{x}' that agrees with them is classified by \mathbf{C} in the same way as \mathbf{x}
- $x_1 \wedge x_4$ is a sufficient reason for $\mathbf{x} = (1, 0, 1, 1, 1)$ given \mathbf{C}

► Counterfactual explanations

- A **counterfactual explanation** for \mathbf{x} given \mathbf{C} is a minimal subset t of the characteristics of \mathbf{x} such that the instance \mathbf{x}' obtained by flipping t in \mathbf{x} is classified by \mathbf{C} in a different way than \mathbf{x}
- $\mathbf{x} = (0, 1, 1, 0, 0)$ is not recognized as a Cattleya orchid by \mathbf{C}
- x_4 is a counterfactual explanation for $\mathbf{x} = (0, 1, 1, 0, 0)$ given \mathbf{C} since $\mathbf{x}' = (0, 1, 1, 1, 0)$ is recognized as a Cattleya orchid by \mathbf{C}

x_1 : "has fragrant flowers" x_2 : "has one or two leaves" x_3 : "has large flowers" x_4 : "is sympodial" x_5 : "has white flowers"

▶ Irrelevant features

- ▶ $x_i \in X$ is **irrelevant** for \mathbf{C} when flipping it in any instance \mathbf{x} does not change the way \mathbf{x} is classified by \mathbf{C}
- ▶ x_5 is irrelevant for \mathbf{C}

▶ Mandatory features

- ▶ $x_i \in X$ is **mandatory** for the class of positive (resp. negative) instances associated with \mathbf{C} when every instance \mathbf{x} such that $\mathbf{C}(\mathbf{x}) = 1$ (resp. 0) contains the characteristics x_i
- ▶ x_4 is mandatory for the class of positive instances associated with \mathbf{C}

x_1 : "has fragrant flowers" x_2 : "has one or two leaves" x_3 : "has large flowers" x_4 : "is sympodial" x_5 : "has white flowers"

▶ Monotone features

- ▶ $x_i \in X$ is **monotone** for the class of positive (resp. negative) instances associated with \mathbf{C} if for every instance \mathbf{x} that does not contain the characteristics x_i and is such that $\mathbf{C}(\mathbf{x}) = 1$ (resp. 0), the instance \mathbf{x}' that coincides with \mathbf{x} but contains the characteristics x_i is such that $\mathbf{C}(\mathbf{x}') = 1$ (resp. 0)
- ▶ x_1, x_2, x_3, x_4 are monotone features for the class of positive instances associated with \mathbf{C}

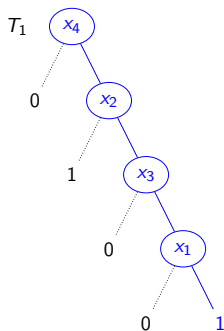
▶ Frequent features

- ▶ The **frequency** of $x_i \in X$ in the class of positive (resp. negative) instances associated with \mathbf{C} is the number of positive (resp. negative) instances that contain the feature, divided by the number of positive (resp. negative) instances
- ▶ The frequency of x_3 in the class of positive instances associated with \mathbf{C} is

$$\frac{6}{10} = \frac{3}{5}$$

- ▶ Using Σ to address the queries over \mathcal{C}
- ▶ Computational problems of various types (decision, counting, enumeration, etc.)
- ▶ **Theorem** XAI queries are **NP-hard** in the broad sense when Σ is any Boolean classification circuit
- ▶ Three questions arise then
 - ▶ Does the complexity of some queries **fall down** when Σ results from the encoding of a classifier from a given family?
 - ▶ How **much inconvenient** is this intractability result from the **practical** side?
 - ▶ How to **circumvent this intractability**?
- ▶ The complexity of XAI queries (and the interpretability of ML models) **turns out to heavily depend on the model at hand**

Because a **direct reason** can be associated with each prediction made, that explains it somehow

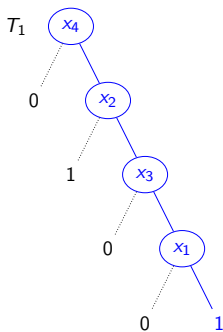


- ▶ The direct reason for $\mathbf{x} = (1, 1, 1, 1, 1)$ given T_1 is $x_1 \wedge x_2 \wedge x_3 \wedge x_4$
- ▶ It can be computed in **linear time** given \mathbf{x} and T_1
- ▶ It does **not always coincide** with a sufficient reason
- ▶ $x_1 \wedge x_3 \wedge x_4$ is a sufficient reason for $\mathbf{x} = (1, 1, 1, 1, 1)$ given T_1

Decision Trees are Interpretable Models ... for Many More Reasons!

21

Theorem XAI queries are in P when Σ corresponds to a decision tree



For decision trees, computing a sufficient reason from the direct reason **in polynomial time using a greedy algorithm**

One can efficiently derive $x_1 \wedge x_3 \wedge x_4$ from $x_1 \wedge x_2 \wedge x_3 \wedge x_4$

x_1 : "has fragrant flowers" x_2 : "has one or two leaves" x_3 : "has large flowers" x_4 : "is sympodial" x_5 : "has white flowers"

- ▶ They appear as **far less interpretable** than decision trees
- ▶ **Theorem** XAI queries are **NP-hard** in the broad sense when Σ corresponds to
 - ▶ a decision list
 - ▶ a random forest
 - ▶ a binary neural network
 - ▶ ...

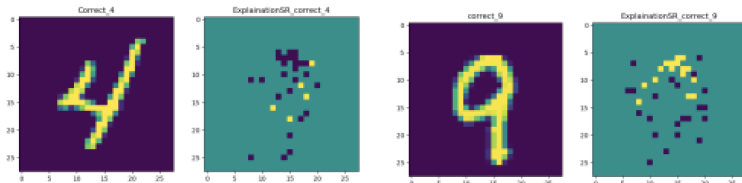
- ▶ Is the game over? Not really ...
- ▶ **Intractability** (NP-hardness) is likely to preclude the existence of a polynomial-time (deterministic) algorithm for solving the XAI query
- ▶ It concerns the **worst case scenario**, but *le pire n'est pas toujours sûr* ...
- ▶ Experiments are needed

Example: Deriving Sufficient Reasons given Random Forests

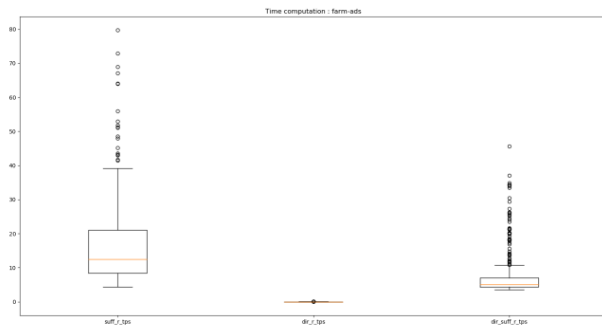
- ▶ Computing a sufficient reason for an input instance given a random forest is **NP-hard**
- ▶ Sufficient reasons can nevertheless be **characterized using automated reasoning concepts**
- ▶ This paves the way for **deriving sufficient reasons using SAT solvers**, which can prove very efficient in practice
- ▶ **Experiments have been made**
- ▶ Generating random forests using Scikit-learn for many standard datasets (coming from open ML, Kaggle or the UCI repository)
- ▶ Computing sufficient reasons for many instances
- ▶ **Distribution of the computation times**

Though computing sufficient reasons is **NP-hard**, this looks as **feasible in practice in a number of cases**

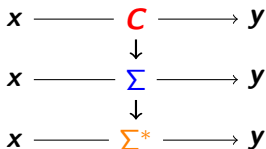
- ▶ Separating “4” from “9” in MNIST dataset ($28 \times 29 = 784$ pixels, viewed as binary features)
- ▶ Using a random forest consisting of 10 decision trees (accuracy: 88%)



- ▶ A dataset based on more features: Farm-ads (54 877 binary features)
- ▶ Using a random forest consisting of 100 decision trees (accuracy: 92,7%)
- ▶ Statistics based on 400 instances



- ▶ Translating the circuit Σ into a more tractable form
- ▶ A matter of **knowledge compilation!**
- ▶ **Principle:**
 - ▶ Turn Σ into another data structure Σ^* during an off-line phase (done once)
 - ▶ Solve the XAI queries using Σ^* instead of Σ , the other inputs (instances, features, class) varying



Identify for each XAI query a set of KC queries and transformations that, when offered, are **sufficient to make the XAI query tractable**

▶ Queries

- ▶ CO: consistency
- ▶ ME: model enumeration
- ▶ IM: prime implicant
- ▶ EQ: equivalence
- ▶ SE: sentential entailment
- ▶ CT: model counting
- ▶ OPT: optimization

▶ Transformations

- ▶ CD: conditioning
- ▶ FO: forgetting
- ▶ $\wedge BC$: bounded conjunction
- ▶ OPT: optimization
- ▶ $\wedge DC$: decomposable conjunction

XAI query	Tractability conditions on \mathcal{L}	Candidate languages \mathcal{L}
EMC	CD, OPT, ME	DNNF
DPI	CD, FO, IM	(*) Decision-DNNF
ECO	CD, OPT, ME	DNNF
CIN	CD, CT	d-DNNF
EIN	CD, ME	DNNF
CAM	CD, CT	d-DNNF
EAM	CD, ME	DNNF
MFR	CD, CT	d-DNNF
IMA	CD, CO	DNNF
IIR	CD, FO, EQ	(*) structured Decision-DNNF
IMO	CD, FO, SE	(*) structured Decision-DNNF
MCJ	CD, CT	d-DNNF
MCH	CD, $\wedge BC$, $\wedge DC$, OPT, ME	structured DNNF
MCP	CD, OPT, ME	DNNF

One Step Further: From Explanations to Intelligible Explanations

Intelligibility is a matter of

- ▶ **structure**: explanations must be **structurally simple** ✓

One Step Further: From Explanations to Intelligible Explanations

Intelligibility is a matter of

- ▶ **structure**: explanations must be **structurally simple** ✓
- ▶ **size**: explanations must be **short**
 - ▶ George Miller (1956): *“The magical number seven, plus or minus two: Some limits on our capacity for processing information”*
 - ▶ When human beings “chunk” items (i.e., group them together as a unit), due to human memory limitations, the size of chunks is limited to 7, plus or minus 2
 - ▶ Ever since then, many experiments in cognitive science have confirmed this limitation

One Step Further: From Explanations to Intelligible Explanations

Intelligibility is a matter of

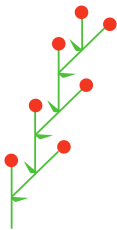
- ▶ **structure**: explanations must be **structurally simple** ✓
- ▶ **size**: explanations must be **short**
 - ▶ George Miller (1956): *“The magical number seven, plus or minus two: Some limits on our capacity for processing information”*
 - ▶ When human beings “chunk” items (i.e., group them together as a unit), due to human memory limitations, the size of chunks is limited to 7, plus or minus 2
 - ▶ Ever since then, many experiments in cognitive science have confirmed this limitation
- ▶ **concepts involved**: explanations must be **understandable**
- ▶ ...

The Sizes of the Reasons for Decision Trees

Dataset	#I	#F	%A	#B	#DR	#SR
Ad-data	3279	1558	96.58	141.1	33.8±14.9	30.3±10.7
Adult	48842	14	81.41	2973.2	17.4±5.9	16.5±5.1
AllBooks	590	8266	71.02	88.8	15.0±13.5	14.1±12.1
Arcene	200	10000	73.00	11.7	4.1±0.9	4.1±0.9
Christine	5418	1636	62.77	419.0	16.1±9.1	15.8±9.1
CNAE	1079	856	86.00	113.9	14.5±13.7	13.7±12.5
Dexter	600	20000	86.50	36.2	7.2±2.8	6.9±2.8
Dorothea	1150	100000	90.70	32.1	16.7±3.9	16.6±4.2
Farm-ads	4143	54877	86.75	264.6	25.9±21.4	24.7±20.6
Gina	3153	970	87.54	164.5	14.4±6.4	14.3±6.5
Gina-p	3168	970	86.77	186.7	13.4±4.7	13.3±4.7
Gina-a	3468	784	85.29	186.0	13.9±5.9	13.8±6.0
Gisette	7000	5000	93.67	173.3	25.2±10.4	25.0±10.5
Madelon	2600	500	76.00	181.9	10.6±3.5	10.4±3.6
Malware	6248	1084	99.09	43.0	7.3±1.6	7.1±1.4
p53mutant	31420	5407	99.36	85.1	37.4±4.7	37.4±4.8
Pd-speech	756	755	81.10	44.3	11.2±5.2	10.9±5.3
Reuters	2000	249	92.05	89.8	16.7±6.3	16.4±6.3
Shuttle	58000	9	99.98	32.3	7.2±1.7	7.2±1.7
Spambase	4601	58	92.05	261.1	15.9±6.3	15.3±6.1

Results for 20 datasets. For each dataset, we indicate the number of instances (#I), the number of features (#F), the mean accuracy over the 10 decision trees (%A) that have been generated, the average number of binary features they are based on (#B). The average size is provided for direct reasons (#DR) and sufficient reasons (#SR).

- ▶ Explanations are expected to be based on concepts that are understandable
- ▶ $x_1 \wedge x_4$ is a sufficient reason for $\mathbf{x} = (1, 0, 1, 1, 1)$ given the random forest \mathbf{C} considered at start
- ▶ x_4 means “is sympodial”
- ▶ Is this helpful for you?
- ▶ “The stem has a zigzag form” must be better!



- ▶ KR has developed concepts and tools to deal with reformulation
- ▶ Amounts to a **definability** issue
- ▶ A domain theory K **defines** a concept x in terms of a vocabulary U if and only if there exists a formula φ over U such that

$$K \models \varphi_U \Leftrightarrow x$$

- ▶ Defining **new encodings** dedicated to other families of classifiers (e.g., CNN)
 - ▶ Implementing and evaluating programs for **addressing other XAI queries for other families of classifiers**
 - ▶ Designing dedicated knowledge compilation techniques for XAI
 - ▶ Developing open source libraries for XAI
 - ▶ Taking advantage of them for specific applications (confiance.ai)
 - ▶ Using KR techniques to better learn (e.g., the data frugality issue) and ML techniques to better reason
 - ▶ ...
- ⇒ **Developing approaches combining ML and KR techniques, to take the best of each, towards hybrid AI**