# IPL HyAIAI : Hybrid Approaches for Interpretable AI

**Project lead:**

Elisa Fromont efromont@inria.fr

**Project lead EP and center:**

Lacodam Rennes Bretagne Atlantique

**Inria partners (EP, center, representative):**

- Magnet: Lille, Marc Tommasi, marc.tommasi@inria.fr
- Multispeech: Nancy, Emmanuel Vincent, emmanuel.vincent@inria.fr
- Orpailleur: Nancy, Amedeo Napoli, amedeo.napoli@inria.fr
- SequeL: Lille, Philippe Preux, philippe.preux@inria.fr
- TAU: Saclay, Marc Schoenauer, marc.schoenauer@inria.fr

## Motivation

This IPL is a follow-up of the work on the Inria PS2018 challenge "*Data Science for Everyone*"[1]. There is currently a huge regain of attention for the Artificial Intelligence field (AI), which is often erroneously assimilated to Machine Learning (ML). The goal of supervised ML is, given a sufficient number of training examples, to learn to perform a task (ex: differentiate normal and spam mails). The learning must be sufficiently generic to be able to perform the task satisfactorily on unseen examples. The ML algorithm "learns" by building a *model* from the training examples. Depending on the approaches, the model can either be *symbolic* (e.g. a set of rules) or *numerical* (e.g. a set of weights of a neural network, the weight of examples in an SVM). Recent numerical approaches based on Deep Learning (DL) have significantly pushed the boundaries of the state of the art, and reached human and even superhuman performance on difficult tasks such as recognizing objects in images. In practice, due to the complexity of the tasks to be learned, numerical models often prove to be more flexible and can better capture this complexity. ML research has mostly focused on these models, while symbolic models have been more studied either in traditional AI fields (*e.g.* planning) or in Data Mining (DM) where researchers work on models understandable by human experts.

ML approaches based on numerical models are victims of their success: due to their good performance, they are used for more and more tasks, and are more and more likely the basis of decisions having a strong impact on human users (*e.g.* accept or reject a loan). The recent European General Data Protection and Regulation (GDPR, see in particular recital 71) introduces the idea that humans should have the possibility to obtain an explanation of a decision proposed by automated processing, and to challenge that proposed decision. It is extremely difficult to produce such explanations with modern Machine Learning approaches, especially those based on Deep Learning.

There is thus an emerging research trend aiming to provide *interpretations* for the decision of "black box" ML algorithms such as DL. In the HyAIAI IPL, we claim that there is a need for *two-way*

1 Inria Strategic Plan 2018-2022, challenge 6, page 46 (English version)

communication between a DL model and a user: of course the user must understand the DL decisions, but when the user participates in the training of the DL model, she/he must also be able to provide expressive feedback to the model. We believe that this two-way communication requires a hybrid approach: complex numerical models must play the role of the learning engine due to their performance, but they must be combined with symbolic models in order to ensure an effective communication with the user.

This combination is difficult to achieve, as symbolic and numerical communities have somehow evolved in different directions, and use different theoretical tools. There exist only a few works on hybrid approaches, hence the need for this IPL. Inria is well positioned to lead such an ambitious program, thanks to its excellent teams in both domains. The teams involved in this IPL include numerical machine learning researchers (Magnet, Multispeech, SequeL, TAU) and symbolic data mining researchers (Lacodam, Magnet, Orpailleur) who wish to collaborate together in order to pioneer a new family of hybrid approaches.

## Context

Nowadays, research on AI is having a huge momentum. A large part of this research is focused on Machine Learning, and especially on Deep Learning approaches. One of the most visible part of that research is the work of the GAFAM-BATX[2], which exploits data of millions of users and learns accurate models for pushing ads or product recommendations to these users. AI research is also driven by the scientific community, where the learning capabilities of DL approaches, especially for complex data such as images and videos, may lead to new progress in fields as diverse as medicine or astronomy. However, the lack of explanations of DL proposed decisions is becoming a problem as it makes its way in more and more applications with critical stakes for human users (medicine, banking, automated cars,...), and the opacity of the models and the decisions can rapidly lead to a societal rejection of the technology. Hence, the recent years have seen the development of high profile projects toward explanability.

The most well known is DARPA's eXplainable AI[3] (XAI) project, whose goal is to replace traditional ML "opaque" models with explainable models, which will help users understand and trust ML decisions. Several research directions are being explored: on the one hand, "deep explanations" (generate explanations directly from a Deep Network), and on the other hand, alternative models which are more easily explainable. In parallel, the human factor is also taken into account: what makes a good explanation for a human, from a psychological point of view? HyAIAI has similar goals to XAI, but focuses solely on hybrid numerical/symbolic methods.

In Europe, the ERC has funded several AI/ML projects. Most of these projects currently focus on the task that stands "before" interpretability: helping a Data Scientist to design a ML/DM approach, by providing more automation to parts of the Data Science pipeline.

- The Franck Hutter's ERC (BeyondBlackBox, 2017-2022) focuses on automating the design of DL approaches. The hyperparameters of a Deep Network govern its structure: number of neurons, number of layers, architecture of the network, activation functions, etc. The performance of

2 American and Chinese major web players: Google, Apple, Facebook, Amazon, Microsoft / Baidu, Alibaba, Tencent, Xiami

3 https://www.darpa.mil/program/explainable-artificial-intelligence

the network is highly dependent on these hyperparameters, and the design space is huge: it is thus important to provide some automated help to explore this space.

- The Luc De Raedt's ERC (SYNTH, 2016-2021) has the more general goal of automating or semi-automating Data Science tasks, with initial work on "data wrangling" (the tasks preliminary to any data analysis process). The starting point of that ERC is the estimation that 80% of the time in Data Science is spent in preprocessing the data, i.e. in writing scripts to select the right parts of the data and to convert them to a format suitable for consumption by analysis tools. (Partly) automating such data formatting / conversions would save a lot of time in the data science process.
- The Tijl de Bie's ERC (FORSIED, 2014-2019) proposes to "formalize subjective interest in Exploratory Data Mining", i.e., find ways of better determining what human analysts are interested in when analyzing a dataset for discovering interesting patterns. It combines recent pattern mining techniques such as subgroup discovery and work on information visualization.

While none of these projects directly focuses on interpretability, they enable progress on the numerical or symbolic models of ML/DM, with the goal of helping users: the techniques they develop will allow making more interpretable approaches.

Apart from the XAI project, most works on interpretable ML are conducted in a less structured way by independent groups. Recent and notable contributions in this field (mostly from the University of Washington) have tilted the research trend towards "local and agnostic interpretability modules": methods to generate explanations for any ML algorithm on a given instance. The explanations take the form of traditionally interpretable models such as linear attribution models [1, 2], rules [3] or decision trees. This contrasts with previous efforts that specialized on a particular type of ML model (ex: DL) and focused on global interpretability, i.e., compile a simpler interpretable model that explains the black box on all instances. There is a flurry of publications [4, 5] and a convergence of research and societal interest that witnesses the emergence of a strong community working on "interpretable ML". It would be strategic for Inria to quickly take action to get expertise in this novel domain, in order to become a leader in this emerging community.

Inside Inria, the HPC-BigData IPL involves, among others, the SequeL and TAU teams. This IPL is focused on the interplay between HPC, ML and data mining algorithms, and is mostly three-fold: use HPC to improve the running speed of existing ML algorithms; determine how ML algorithms can be tailored to best fit available HPC resources; design HPC architectures to benefit ML algorithms. The overlap with HyAIAI is minimal, although progress on one IPL may be beneficial to the other.

## Targeted challenges and expected impact

Our aim is to propose a principled two-way communication strategy between a human user and some black-box ML model, with a specific emphasis on DL models. Depending on the task at hand, we mainly consider two types of human users: on the one hand, domain analysts, end-users of ML methods, and on the other hand, ML system designers. For expressiveness and interpretability purposes, we want to use symbolic models as a "translation layer" between the human and the learning model (called the *system* below). This is our major distinction with XAI: due to the considerable difference of manpower, we consider that a more efficient use of our resources is to focus on a single family of solutions, namely the use of hybrid numerical/symbolic approaches.

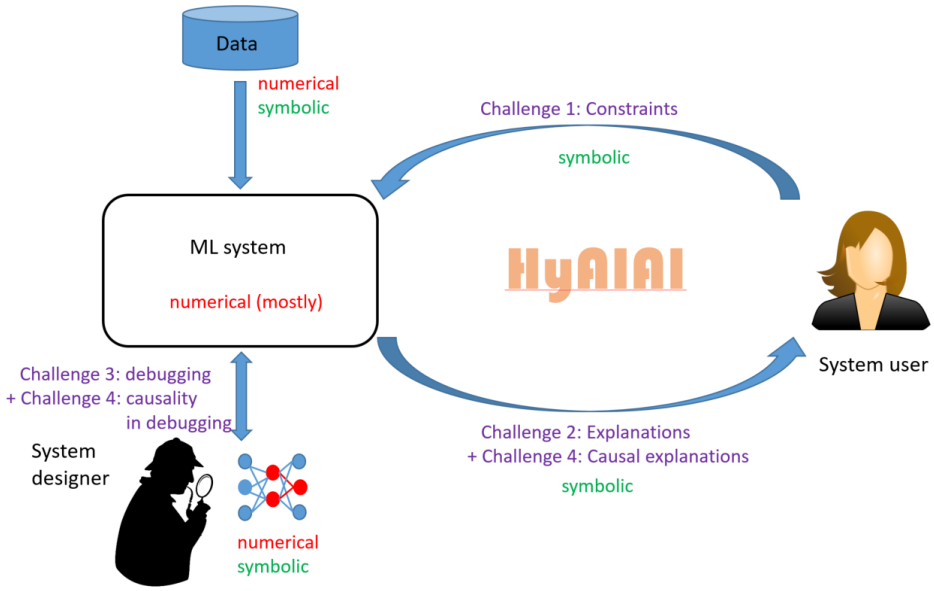



Figure: Diagram of HyAIAI challenges

HyAIAI wants to make progress on the four following challenges:

1. **System must understand human requirements:** As of now, the only information that can be given to a Deep Learning system is the set of training examples, which can be seen as a very low level description of the task to be performed. An important challenge is to enable the human using the system to provide high level knowledge and constraints to the DL system. Knowledge is typically encoded as a set of concepts and relations (*ex: a cat is a kind of animal, it has whiskers and fur, it can jump but not fly*), while constraints are expressed in terms of the probability, or relative probability, of certain events defined via such concepts (ex: *autonomous car must not hit a human, drone may not fly over restricted zone, system must perform similarly for female and male users,* or *how to take counterexamples into account in a more explicit way*).
2. **Human must understand system responses:** A huge challenge nowadays is to enrich ML systems with the ability to *explain* their outputs to their human users. In many scenarios, it may be risky, unacceptable, or simply illegal, to let artificial intelligent systems make decisions without any human supervision. Hence, it is necessary for ML systems to provide an explanation of their decisions to all the humans concerned. However, the notion of explainability is user- and domain-dependent, and depends on i) the level of abstraction desired/mastered by the user; ii) the preferred mode of explanation (e.g., by means of rules or by means of examples; many other options exist in cognitive modeling, and are beyond our expertise). Some form of dialog between the user and the system thus seems necessary to reach a satisfactory level of explainability. One direction for designing this dialog is to embed the results of classifiers in a symbolic classification methodology such as Formal Concept Analysis [6], as already explored by the ORPAILLEUR team. Regarding relation with the previous challenge, while challenge 1 addresses the creation of a new family of hybrid methods that are interpretable by design, challenge 2 addresses the interpretation of existing methods which were not initially designed with interpretability in mind.
3. **Human must understand the inner working of the system ("debugging"):** Apart from non-expert system users, the other humans involved in ML systems are the system designers. Designing ML systems involves a complex choice of methods, and the tuning of a large number of hyperparameters. There are many works to automatically tune these hyperparameters (e.g., Hutter's ERC). However, this induces a huge computational cost (even more so for Deep Learning models), and does not empower the human designers. It would be much more efficient to provide

human designers with understandable clues about the reasons for the bad performance of the system, in order to benefit from their creativity to quickly reach more promising regions of the hyperparameter search space. This would allow human experts to tailor an ML system to their needs in an easier way, while reducing the ecological and economical footprint of AI.

4. **Causality, yet another dimension of explainability**: The celebrated link from knowledge to prediction, from prediction to prescription, is one of the roots for the fame of Big Data. The implicit promise of Data Science relies on two conjectures. The first one is that massive data allows building accurate predictive models. This conjecture holds true under some conditions on the quality of these data and the nature of the target phenomenon. The second conjecture is that predictive models support interventions, e.g., if I could predict in which case someone is ill, I would know how to act (intervention) in order not to become ill. This conjecture is plainly wrong: correlation-based models can be accurate predictive models (umbrellas in the street allow to infer that it rains), but do not support interventions (rain will not happen by the virtue of bringing umbrellas in the street). It is thus important to work on *causal models.*

**Impact.** Due to the ubiquity of ML in everyday applications, progress in any of the 4 challenges above is likely to have a large societal impact. The most obvious is challenge 2 which, regarding important decisions (such as loan attribution or predictive justice), may prevent citizens from feeling victims of completely arbitrary decisions (leading to frustration and rejection of ML approaches). Challenge 1 will increase trust in ML approaches, thus enabling their use in more applications. Challenge 3 will participate in accelerating the design of ML approaches, with a better understanding from the designers. This will also contribute to have more trusted approaches, that can be deployed in less time. Lastly, challenge 4 is more fundamental, and in the context of HyAIAI will allow more robust solutions for challenges 2 and 3.

## Scientific approach

In practice, we will first focus on Deep Learning models, due to their recent highly visible successes on numerous tasks.

**Challenge 1: System must understand human requirements (LACODAM, MAGNET, MULTISPEECH, ORPAILLEUR, TAU)**. Understanding human requirements is important to address the acceptability of AI-based systems. In most cases, system performance is the first and main requirement users address for a new ML system. Thus, obtaining guaranties on the accuracy and the speed of new systems is of paramount importance. However, given the acceptable performances of systems in certain classes of problems and the societal impacts of ML, new requirements have appeared. ML-based systems must protect privacy and must avoid reproducing unacceptable biases and stereotypes when proposing decisions. A first aspect of this challenge is therefore to deal with such symbolic constraints in the design of systems. New methods have been proposed to address fairness in several domains [7, 8] but this research domain is still in its infancy and not really mature for real world applications.[4] The same observation can be made for privacy. In this project we want to address the privacy problem from a decentralized learning perspective but also study the impact of this new technology and protocols on the legal side, for instance in the context of necessity appearing in the GDPR.

4 https://www.reuters.com/article/us-alphabet-google-ai-gender/fearful-of-bias-google-blocks-gender-based-pronouns-from-new-ai-tool-idUSKCN1NW0EF

While applicable to ML in general, let us explain a second aspect of this challenge using the specific case of Deep Learning (DL). Current DL methods excel at supervised classification with a fixed set of classes, where information is provided at training time in the form of costly human labeling of training data. This approach faces several serious limitations: it does not scale with the thousands of fine-grained concepts experienced in daily life, it performs poorly when only a few labeled data is available for a given class, and it operates as a black-box and hence does not allow users to easily express desirable constraints. Humans learn in a totally different way. Rather than treating each concept as a separate class, they understand concepts in relation to each other. This enables them to discover new concepts, to achieve good performance with little or no supervision, and to take specific constraints into account on the fly. This knowledge is typically represented as a knowledge graph, which encodes all known concepts along with their attributes and their relations. We claim that using and enriching knowledge graphs in the DL training and exploitation stages is a key towards addressing the above limitations. The resulting hybrid ML systems will not only understand human requirements, but also scale and perform better and be explainable by design.

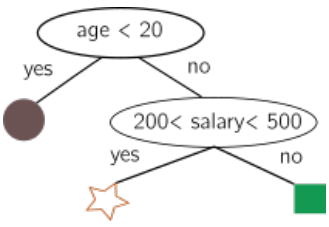
For example, considering the task of animal image recognition, knowing that "*cats have whiskers and fur*" and "*sea lions have whiskers but no fur*" shall help the ML system recognize whiskers even when the training images are labeled in terms of "*cat*" or "*sea lion*" only. Training a "*whisker*" classifier on images of multiple animals is then expected to improve the recognition accuracy for these animals, and to enable the discovery of other animals with whiskers which are neither a cat nor a sea lion and may be added to the knowledge graph after suitable interaction with the user.

Despite their ability to learn a composite representation of the data, conventional DL systems do not disentangle the concepts underlying the target classes very well [10]. Work has focused on improving this ability by constraining the network weights [9], introducing adversarial objectives [10], or exploiting spatial relationships [11]. Nevertheless, these approaches still assume a flat list of classes and they do not exploit the available human knowledge. Much less work has been devoted to integrating existing symbolic knowledge into DL. Some attempts have attempted designing the network structure according to that of the knowledge graph [12, 13], which is limited to few concepts and does not allow on-the-fly introduction of new concepts. Another solution consists in adding regularization layers based on a given class hierarchy [14], which is limited to hierarchical relations as opposed to other relations between classes. It is also possible to design an end-to-end network that verifies whether a relation (specified in natural language) between its inputs is true or not [15], which requires supervised training data and does not allow transparent use and composition of relations.

Drawing inspiration from these methods and especially the last one, we aim to design a hybrid ML method that will enable the exploitation of several relations (not only spatial or hierarchical) between concepts, and the learning of new concepts on the fly. User constraints may then be specified in terms of these concepts, and enforced in the training phase or in the exploitation phase.

**Challenge 2: Human must understand system response (LACODAM, MAGNET, MULTISPEECH, ORPAILLEUR, SequeL, TAU)**. This challenge is concerned with the several ways in which a human can understand a learned model and the predictions it makes.

The first approaches for interpreting DL proposed to apply reverse engineering on the outputs of a Deep NN. The idea is to construct a training set from the answers of the DNN and induce a simpler interpretable model (often a decision tree, see figure on the left) that mimics the DNN's behavior. Other methods, such as RxREN [16], carried out reverse engineering by removing the input neurons that exhibit little impact on the DNN's answer. Rule induction is then applied on the simplified DNN and its answers. More generally, there are several works aiming at extracting rules from DNNs [17] or from



simpler network surrogates [18]. All these methods compute "global explanations" that summarize the logic of the DNN. Since explaining a complex NN with a simple model carries a loss in fidelity (accuracy w.r.t the original model), recent methods propose to derive explanations on a *local view* of the DL model. This local view is computed around an instance of interest, i.e., the instance on which the ML algorithm has been applied and for which we want an explanation. These explanations are expressed by means of a simplified model on an interpretable space (see LIME [1], SHAP [2], and Anchors [3]) in the vicinity of the instance of interest. Examples of such models are linear functions and anchors as illustrated in Figures 1 and 2 (the instance of interest is highlighted).

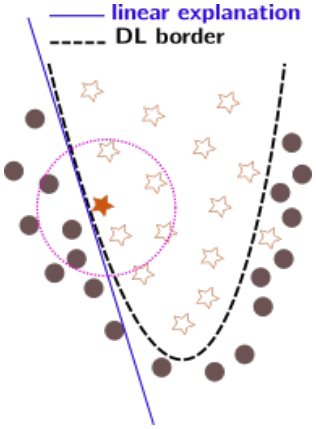



**Figure 1:** A linear approximation (LIME) defines the local border between 2 classes around an instance of interest. ML is the border defined by the black-box classifier we want to explain.

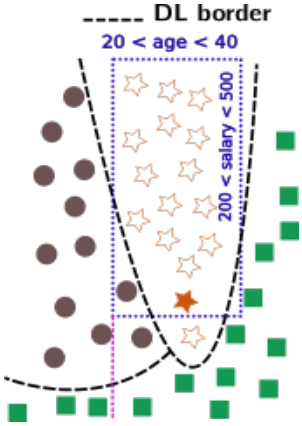



**Figure 2:** An anchor is a region of the space where the answer of a classifier is the same as the instance of interest. The description of this region constitutes an explanation

These methods can however be extended to richer symbolic models and adapted to other settings. For instance, most approaches have focused on two-class settings, hence we could induce explanations on multiclass settings in the form of decision trees as illustrated on the left. This also suggests the need for research that can shed light on the pertinence of the existing methods in different use cases. In other words, we aim at answering questions such as, *under which circumstances is a linear model more informative and understandable than an anchor?*

In other respects, an interesting and still unexplored line of research is the construction of multi-instance local explanations. Such a scenario can arise when we aim at explaining the outcome of an ML algorithm on multiple instances at the same time, e.g., why the system denied a credit to multiple people. Such setting can be extremely valuable when evaluating the coherence of an algorithm in a broader region of the instance space, e.g., to see if and/or why it provides different answers for opposite credit profiles. On the technical side, this idea is tantamount to implementing local interpretability on a broader area of the instance space. More generally, we are interested in explaining the behavior of a model on a (sub)population. This could be relevant when an algorithm wants to explain what fairness properties a model satisfies and why, or to what extent a model or a prediction preserves privacy. Finally, we highlight that most existing approaches on interpretable ML focus on explanations for classifiers on images and text, thus we aim at extending the state of the art in interpretable ML to other settings such as time series.

**Challenge 3: Human must understand the inner working of the system (LACODAM, MULTISPEECH, ORPAILLEUR, SequeL, TAU).** This challenge does not focus on understanding the predictions of neural networks (as in the previous challenge) but on understanding, using symbolic methods, which parts of a network do not behave as expected and are, for example, responsible for certain errors. Such knowledge could help us reduce the size of the networks, improve their performance, make them more robust to adversarial attacks (see Figure below), design their architecture, etc.

Examples of adversarial images from [22] generated for a given DNN (here AlexNet). (Left) is a correctly predicted sample, (center) difference between correct image and image predicted incorrectly (right) as an"**ostrich**" (in both cases)

Understanding the inner working of deep neural networks (DNN) has attracted a lot of attention in the past years [19, 20] and most problems were detected and analyzed using visualization techniques [21, 22, 23]. Those techniques help to understand what an individual neuron or a layer of neurons are computing. We would like to go beyond this by focusing on groups of neurons which are commonly highly activated when a network is making wrong predictions on a set of examples. In the same line as [19], where the authors theoretically link how a training example affects the predictions for a test example using the so called "influence functions", we would like to design a tool to "debug" neural networks by identifying, using symbolic data mining methods (in particular discriminative graph mining), (connected) parts of the neural network architecture associated with erroneous or uncertain outputs. To do that, we would benefit from the expertise of the LACODAM and ORPAILLEUR teams on the pattern mining domain. Besides, the MULTISPEECH team will provide testbeds for the proposed method to improve some currently low performing speech recognition tasks with DNN. Likewise, SequeL will provide their Visual Question Answering system as a testbed aiming at studying how such a system may be equipped to explain e.g. why it failed to answer correctly to a question. Identifying such "faulty" groups of neurons could lead to the decomposition of the DL network into "blocks" encompassing several layers. "Faulty" blocks may be the first to be modified in the search for a better design. This requires precise information on the parts of the data space that are not well predicted. In this regard, we will study adaptive sampling methods. This problem may be seen as black-box optimization of a noisy function, a problem which can be tackled with bandit methods on which SequeL has a lot of experience: their work on optimization with bandit algorithms will be used to discover and focus the search in the most promising parts of the search space[5].

**Challenge 4: Causal models are (more) explainable (LACODAM, MAGNET, TAU).** Most ML studies are concerned with learning predictive models, that allow to predict the value of some variables, based on the values of others. However, this amounts to learn 'only' correlations: another approach to understanding and explaining a complex system is to discover the causal relationships between variables [29]. Indeed, distinguishing causes from effects is of paramount importance in domains such as bioinformatics (e.g., to infer network structures from gene expression data [30]), social sciences and econometrics (e.g., to model the impact of monetary policies [31]), climatology, and epidemiology, both for providing a better

5 see the whole line of research and the many related publications in SequeL since HOO [24], to StoSOO [25] for the optimization of noisy functions, and successors, both theoretical aspects and practical aspects having been studied (see e.g. the participation to a challenge on function optimization at CEC'2015, [26] applications to computational advertising [27], and to recommendation systems [28])

understanding of the world that surrounds us and to assist decision makers that need answers to 'what if' questions (e.g., *what if I open my umbrella. Will it rain?*). While randomized controlled experiments will remain the gold standard to determine causal relationships, such experiments are hampered due to practical, economical, or even ethical reasons. In such cases, determining causal relationships from observational data is of primary importance.

Let us consider the simplest case of two variables: there exists various independence tests on a sample of their joint distribution to decide whether or not they are probabilistically independent. However, determining their causal relationship is much more of an open problem [32]. Firstly, considering two variables in isolation raises the question of the existence of possible confounders, i.e., other variables that would be the actual cause of both variables under study (hence the observed correlation). Secondly, even if confounders are known, and taken into account (e.g. through conditioning), inferring the causal relation between two variables, a.k.a. causation, from observational data, raises many issues. Formally, considering an empirical sample drawn i.i.d. from the joint distribution of two variables (X,Y), which can be represented as a scatter plot, one aims to classify this scatter plot according to the causal relationship between X and Y , distinguishing four classes: X causes Y ($X \rightarrow Y$ ), Y causes X ($X \leftarrow Y$ ), presence of a common cause or confounder ($X \leftrightarrow Y$) or independence ($X \perp Y$ ). This was addressed in the context of the cause-effect pair challenges, and several algorithms have been developed with success to address the causation problem [33]. Note that other algorithms were also proposed with very different approaches, e.g., based on Information Theory, and the fact that if X causes Y, then transmitting X then Y will represent less information than transmitting Y then X [34]. Their use in practice, however, faces several hurdles: the need to identify more than pairwise interactions, the obvious scaling issue when handling high dimensional problems, the lack of sufficient datasets of cause-effect pairs.

Another approach was proposed in TAU [35], that addresses the problem globally, removing the need for combinatorial search to identify the causal network, and borrowing ideas to generative adversarial networks (GANs). However, this approach does not prevent cycles in the causal graph, and requires large datasets. The challenges ahead are the following:

1- Robustify the cause-effect identification for pairs of variables when few examples of neighbor distributions are available, in particular, using techniques inspired by transfer learning.
2- Automatically identify confounders when they are present in the set of variables describing the datasets.
3- Identify some reduced latent representation (e.g., using autoencoders) and apply existing techniques in this small-dimension space, possibly reducing the complexity of the problem.

We hence argue that research into causal models, i.e., models that only exploit patterns for which there is evidence that they relate to a causal relation, will help humans understand the produced models. As one may not be able to be certain about all causal relations (due to the nature or limited amount of data), we are also interested in considering intermediate levels, e.g., settings where models are limited to exploit correlations that are statistically significant, rather than any correlations which together happen to empirically improve prediction performance.

## Evaluation and applications

To evaluate our approaches, we will require close collaboration with domain experts. We will rely on the applications where the collaborations are the most mature in the teams of the consortium. Some examples are speech recognition, metabolomics and agriculture.

Some of the work proposed in the context of this IPL will be relatively easy to evaluate:

- For challenge 1, once constraints will be added on top of a DL approach, the evaluation will consist in verifying that the constraints are indeed enforced, and what is the possible loss in error and computation time incurred by the enforcement of these constraints. The discovery of concepts and their relations (second part of the challenge) can be validated in settings where these concepts and relations are already known, or through user studies (in a similar setting as Challenge 2 validation, developed below).
- For challenge 3, an indirect validation will be the performance gains allowed by correcting parts of the network identified as "faulty" by the proposed methods. Note that here careful experimental design will be required, as the identification and the correction of the "faulty" part will be simultaneously evaluated and their effects will be difficult to disentangle.
- For challenge 4, a first validation uses artificial datasets, for which the ground truth is known (see e.g., the experimental section in [33]). Acyclic or cyclic causal graphs can be built for any given dependency between the variables. There also exist complex real-world systems for which the causation is known (or is at least consensual among researchers), see e.g., the protein network in [30]. ROC curves, or precision-recall plots can be used to illustrated the comparative results.

Challenge 2, on the other hand, is much harder to evaluate, as its evaluation is based on how well a human user understands an explanation, and how well this explanation conveys the rationale for a black box classifier's proposed decision. This requires user studies, a method that data science is not yet used to employ. To conduct such user studies, a first question is to determine who are the users concerned. The recent literature focuses on well known tasks such as object identification in images, allowing the use of "normal people" for user-studies. These studies can then be conducted via crowdsourcing platforms. Such studies are likely to become a norm as the field matures: we will also investigate them, especially with applications in speech understanding. This can be an interesting point of convergence with CHI teams of Inria, which have a long experience of user studies.

All teams involved in the IPL have strong collaborations with industry and/or academics in other domains, on challenging applications. These collaborations are prime targets to make an in-depth evaluation of the approaches proposed in HyAIAI, and see how much benefit they can bring in an actual setting where the end-users have a practical interest in the results produced.

For example, one of these applications is agriculture. In the context of the #Digitag Convergence Institute, LACODAM is collaborating the the INRA unit PEGASE for analyzing dairy cow sensor data (temperature, activity, milk composition…). Machine Learning methods applied to this data can help to detect diseases or estimate the optimal period for reproduction. These decisions have a cost, so it is important for the farmer to understand the reasons that lead to the decision proposed by the Machine Learning system, in order to build trust and exploit human expertise. This is thus an ideal testbed for HyAIAI approaches. On a related topic, the ORPAILLEUR team has gained a strong experience in mining metabolomic data for nutrition purposes within the Diapason INRA project involving a team of biologists and chemists working on nutrition habits. A special combination of supervised classifiers and pattern mining methods has been designed for that purpose and solving two interrelated problems, namely discrimination (i.e. markers separating classes) and prediction (markers

predicting class membership). Moreover, visualisation based on pattern mining appeared to be a good means for guiding knowledge discovery and interpretation by domain analysts.

Another example is the Visual Question Answering (VQA) task, consisting of open-ended questions about real images, studied in the SequeL team. Answering these questions requires an understanding of vision, language and commonsense knowledge. Their solution (named MODERN) is based on the idea of modulation introduced in [36] and trained on a 614K questions on 204K images. MODERN combines an LSTM with a Resnet-50.

The MULTISPEECH team is specialist in speech recognition, which consists of transcribing a speech signal as a word sequence. Two systems will be considered as benchmarks for Challenge 3: a classical system[6], which combines a time-delayed neural network (TDNN) acoustic model outputting phonetic targets and a language model helping to decode these targets into words, and an end-to-end system[7] which directly outputs a letter sequence. Another application from MULTISPEECH team is audio event detection: it aims to find and classify sound objects (e.g., car alarm, footsteps, door) in an ambient audio recording. They used Youtube data and part of an existing taxonomy of 632 classes[8] as the basis for their system and for a challenge they organized[9]. In real life, the variety of sounds is larger and a complete ontology can be built using, e.g., WordNet[10], that qualifies sounds in terms of verbs and adjectives, e.g., “closing a heavy door”. HyAIAI will use this task as a benchmark in Challenge 1.

## Organization

We are planning to have a general meeting with representatives of each team of the consortium every 6 months, and a visio/audio meeting every month. The goal of the face-to-face meetings will be to exchange about the progress made and the issues encountered. Due to the fast pace of the research in the DL area, the monthly video/phone meetings will mostly be devoted to a reading group where the most recent literature will be presented and discussed. The postdocs will be in charge of curating these discussions and producing final documents for diffusion and archival.

The approaches envisioned by HyAIAI have an important interest, not only for the “core AI” community, but also for the large number of colleagues of Inria that exploit AI/ML methods in their area of Computer Science. For example, a colleague from robotics is likely to use Deep Learning for image analysis, and may be interested either in simple ways to constrain the learning (challenge 1), in better understanding the network decisions (challenge 2), or in finding more easily what goes wrong in the network (challenge 3). We will thus organize a yearly “HyAIAI workshop” open to all the interested teams of Inria and more broadly to the national / European community. This workshop will allow us to both disseminate the results of HyAIAI, and to create a strong community on hybrid approaches, fostering stimulating scientific exchanges. We are confident in our collective capacity to attract top-level international speakers, which will further increase the scientific interest of this workshop.

6 https://github.com/kaldi-asr/kaldi
7 https://github.com/espnet/espnet
8 https://research.google.com/audioset/
9 http://dcase.community/challenge2018/index
10 https://wordnet.princeton.edu/

## Ressources

This IPL proposal is a 48 month project. The main request of HyAIAI is in funding non-permanent research staff in order to work on the topics of the IPL. We propose the hiring of 6 non-permanent research staff. In this document, we do not make a distinction between PhD student positions and 2-year postdoc positions (both are referred to as "Position" below). Currently there is a huge demand on well trained AI/ML personnel, making the hiring of good postdocs difficult. Depending on opportunities, we will hire either PhD students or postdocs in order to maximize the quality of the candidates. The organisation of 1 workshop per year with international invited speakers, a laptop for each non-permanent staff, and 3 international missions per year will require an annual operating budget of 18 Keuros.

**Position 1:** *Integration of symbolic knowledge into DL* (MULTISPEECH, ORPAILLEUR, TAU)

We aim to design a general methodology for exploiting the relations between concepts known from an existing ontology into the design and the learning phase of an ML system, and for learning new concepts and relations on the fly. Experiments will focus on DL applied to audio event detection, considering the Audioset taxonomy and augmenting it with WordNet and new concepts and relations discovered from the data, possibly using pattern mining methods. To achieve this two-way interaction between an ontology and a DL system, we seek both to encode relations into the network structure, and to discover salient dependencies between its outputs. The resulting hybrid ML system will be explainable by design (Challenge 1) and it is expected to scale and perform better than existing purely DL-based systems. A symbolic model of the input-output of such a hybrid system can be further built, e.g. based on a two dimensional data table, and analyzed in terms of extracted implications (or functional dependencies). The latter can be used either for verifying existing explanations or providing new explanations (challenge 2).

**Position 2:** *Visual query answering* (LACODAM, SequeL)

We consider a deep neural network that has been trained up to a certain satisfactory level of performance on a given dataset. We want to design a mechanism that lets the network explain its output in response to an input instance. As of now, the only explanation that we can obtain is given by the whole architecture of the net along with its parameters, i.e., weights, activation function, etc. Our goal is to go towards an explanation which is of a higher level of significance for humans (Challenge 2). The state of the art provides this to some extent: local explainability modules offer explanations in terms of weights in a simplified, interpretable feature space. This space consists of features that are more comprehensible than the original features. Examples are superpixels instead of pixels or word occurrences instead of word embeddings. Nevertheless, we remark that superpixels and word occurrences are devoid of semantics. Thus, we propose to semantify these explanations by generating explanations in terms of concepts (Challenge 1). For instance, if the neural network detected a cat, a possible explanation could be the presence of whiskers (existing methods may highlight the area that contains such conceptual feature as explanation). In the same way, word occurrences may be replaced by bags of word and their topical interpretation (e.g., religion, sports, an entity from a knowledge base, etc.). This semantification opens the door to a range of visualization techniques that will boost the quality and understanding of the explanations.

**Position 3:** *Legal transparency and verifiability* (MAGNET, TAU)

Recently, there is an increasing interest in machine learning, which satisfies some qualities of fairness and privacy-friendliness. To improve the trust of data subjects in machine learning processes, an understandable, symbolic explanation is needed which convincingly shows why an algorithm is fair or privacy-friendly. While some ontologies have been proposed for legislation such as GDPR, no integration with learning algorithms has been made. This postdoc project will develop an ontology and associated tool for reasoning about properties of the knowledge discovery process, and provide understandable formal verifications of claimed guarantees (Challenge 1).

The HyAIAI partners are already involved in projects on formal verification, on privacy-preserving machine learning and speech processing, and on a better understanding of data protection legislation. Together, this will provide the postdoc with the necessary expertise to perform the described action.

**Position 4:** *Causality discovery from observational data* (LACODAM, TAU)

Learning causal models from observational data is a key challenge in Machine Learning Research (Challenge 4). Tau has recently proposed two original approaches using Deep Neural Networks: the first one starts from a skeleton of the causal relations and identifies the exact nature of these causations; the second one borrows ideas for generative adversarial networks (GANs) [35]. In this position, we would like to bridge the gap between both approaches, and to extend the second approach by reducing the volume of data it requires, and reduce the complexity of the problem through the use of latent representations (e.g., using autoencoders).

Another line of research will be to particularize the notion of causation: existing algorithms discover if, globally, variable X is the cause of variable Y. Tau and Lacodam will investigate more complex causation relations, like "When condition C holds, then X causes Y", or "The temporal sequence <A,B,C>, anywhere in time, causes an alarm".

**Position 5:** Pattern mining for Neural Nets debugging (LACODAM, MULTISPEECH)

The inner working of deep neural networks is still not well understood. With this position, we would like to go beyond the current common visualization techniques that help to understand what an individual neuron or a layer of neurons is computing, by focusing on groups of neurons that are commonly highly activated when a network is making wrong predictions on a set of examples. We propose to use symbolic data mining methods (here, discriminative pattern mining) to better understand the behavior of neural networks and propose solutions that improve their performance (Challenge 3).

**Position 6:** Audio-based predictive car maintenance (LACODAM, MAGNET)

Predictive car maintenance is the task of predicting in advance if/when a car will need maintenance. Car equipment suppliers and have promoted the use of dedicated sensors that record vibrations and sounds (Carfit, Xee[11]).

However, it is useful to not only have some statistical prediction but also to understand what exactly can be heard in the audio. In particular, experts desire to interact with the learning algorithm to decompose the prediction in a more detailed explanation (challenge 2), to constrain the model with their domain knowledge (challenge 1) and to understand to what extent predictions are based on statistical correlations (which may less likely hold in new contexts) or on causal effects (e.g. where there is a mechanical explanation for the sound pattern) (challenge 4). At the same time, the machine learning performed need to satisfy some privacy preserving properties as audio could contain

11 Magnet collaborates with Xee (https://www.xee.com) and has contacts thanks to alumni with Carfit (https://car.fit/)

passenger conversations. It will hence be needed to be able to explain why the algorithms, the models and the predictions preserve privacy (challenge 2).

## Objectives for the 4 years

At the end of HyAIAI, we expect fundamental contributions to the domains of Interpretable ML and Hybrid AI, backed up by open source prototypes showcasing our research work. These prototypes, possibly restricted to a few application domains, will demonstrate the feasibility of providing the user a complete and understandable interaction with a ML model: provide it with complex high-level constraints, receive information on ill-performing parts of the model in order to improve them, and ultimately get explanations on the returned results.

Some of these prototypes may have a large interest for the data science community at large: a longer term objective will then be to submit ADT proposal(s) in order to fund the necessary engineering work to mature these prototypes in order to propose their integration in highly visible data science libraries such as Scikit-learn.

## Future possibilities (at the end of the IPL)

Positioning ourselves as serious contributors of hybrid, interpretable ML approaches, a field that is likely to expand in the coming years, should allow us to establish strong collaborations with international partners. This will then lead to the submission of further projects (Horizon Europe, ERC, etc) to fund future research in this area.

Moreover, there will likely be a very strong industrial interest for the results of HyAIAI. We purposely did not involve any industrial partner in the consortium of the IPL, in order to conduct as fundamental research as we deem necessary in this project. However, given the strong industrial demand for collaborations with the teams of HyAIAI, it will be straightforward for the teams to transfer the most readily applicable results of HyAIAI to their existing industrial collaborators (CIFRE, FUI, bilateral contracts, etc), and to prepare more applied projects at the end of HyAIAI.

## References

[1] Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: *"Why Should I Trust You?": Explaining the Predictions of Any Classifie*r. KDD 2016: pp 1135-1144

[2] Scott M. Lundberg, Su-In Lee: *A Unified Approach to Interpreting Model Predictions*. NIPS 2017: pp 4768-4777

[3] Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: *Anchors: High-Precision Model-Agnostic Explanations*. AAAI 2018: pp 1527-1535

[4] *"Summit on Machine Learning meets Formal Methods"* at the FLOC 2018 conference: https://www.floc2018.org/summit-on-machine-learning/

[5] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, Fosca Giannotti: *A Survey Of Methods For Explaining Black Box Models*. CoRR abs/1802.01933 (2018)

[6] Dhouha Grissa, Blandine Comte, Estelle Pujos-Guillot, Amedeo Napoli: *A Hybrid Knowledge Discovery Approach for Mining Predictive Biomarkers in Metabolomic Data*. ECML/PKDD (1) 2016: pp 572-587

[7] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, Cynthia Dwork: *Learning Fair Representations*. ICML (3) 2013: pp 325-333

[8] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, Kai-Wei Chang: *Learning Gender-Neutral Word Embeddings*. EMNLP 2018: 4847-4853

[9] Kevin Bascol, Rémi Emonet, Élisa Fromont, Jean-Marc Odobez: *Unsupervised Interpretable Pattern Discovery in Time Series Using Autoencoders*. S+SSPR 2016: pp 427-438

[10] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, Marc'Aurelio Ranzato: *Fader Networks: Manipulating Images by Sliding Attributes*. NIPS 2017: pp 5969-5978

[11] Carl Doersch, Abhinav Gupta, Alexei A. Efros: *Unsupervised Visual Representation Learning by Context Prediction*. ICCV 2015: pp 1422-1430

[12] Hao Wang, Dejing Dou, Daniel Lowd: *Ontology-Based Deep Restricted Boltzmann Machine*. DEXA (1) 2016: pp 431-445

[13] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, Jimeng Sun: *GRAM: Graph-based Attention Model for Healthcare Representation Learning*. KDD 2017: pp 787-795

[14] Wonjoon Goo, Juyong Kim, Gunhee Kim, Sung Ju Hwang: *Taxonomy-Regularized Semantic Deep Convolutional Neural Networks*. ECCV (2) 2016: pp 86-101

[15] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, Tim Lillicrap*: A simple neural network module for relational reasoning*. NIPS 2017: pp 4974-4983

[16] M. Gethsiyal Augasta, T. Kathirvalavakumar: *Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems*. Neural Processing Letters 35(2): pp 131-150 (2012)

[17] Tameru Hailesilassie: *Rule Extraction Algorithm for Deep Neural Networks: A Review*. CoRR abs/1610.05267 (2016)

[18] Jimmy Ba, Rich Caruana: *Do Deep Nets Really Need to be Deep?* NIPS 2014: pp 2654-2662

[19] Pang Wei Koh, Percy Liang: *Understanding Black-box Predictions via Influence Functions*. ICML 2017: pp 1885-1894 (best paper)

[20] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals*: Understanding deep learning requires rethinking generalization*. ICLR 2017

[21] Anh Mai Nguyen, Jason Yosinski, Jeff Clune: *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*. CVPR 2015: pp 427-436

[22] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus: *Intriguing properties of neural networks*. ICLR 2014.

[23] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, Wenchang Shi: *Deep Text Classification Can be Fooled*. IJCAI 2018: pp 4208-4215

[24] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, Csaba Szepesvári: *Online Optimization in X-Armed Bandits*. NIPS 2008: pp 201-208

[25] Michal Valko, Alexandra Carpentier, Rémi Munos: *Stochastic Simultaneous Optimistic Optimization*. ICML (2) 2013: pp 19-27

[26] Bilel Derbel, Philippe Preux*: Simultaneous optimistic optimization on the noiseless BBOB testbed*. CEC 2015: pp 2010-2017

[27] Sertan Girgin, Jérémie Mary, Philippe Preux, Olivier Nicol: *Advertising Campaigns Management: Should We Be Greedy?* ICDM 2010: pp 821-826

[28] Frédéric Guillou, Romaric Gaudel, Philippe Preux: *Large-Scale Bandit Recommender System*. MOD 2016: pp 204-215

[29] Judea Pearl: *Causality: models, reasoning and inference.* Cambridge University Press, 2009 (2nd edition)

[30] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, Garry P. Nolan: *Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data*. Science, vol 308, Issue 5721, pp 523-529 (2005)

[31] Pu Chen, Chihying Hsiao, Peter Flaschel, Willi Semmler*: Causal Analysis in Economics: Methods and Applications*. Australasian Macroeconomics Workshop, 2008.

[32] Isabelle Guyon: *Causality Challenge #3: Cause-effect pairs*. 2013.

[33] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, Bernhard Schölkopf: *Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks*. Journal of Machine Learning Research 17: 32:1-32:102 (2016)

[34] Alexander Marx, Jilles Vreeken: *Telling Cause from Effect Using MDL-Based Local and Global Regression*. ICDM 2017: pp 307-316

[35] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, Michèle Sebag: *SAM: Structural Agnostic Model, Causal Discovery and Penalized Adversarial Learning*. CoRR abs/1803.04929 (2018)

[36] Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, Aaron C. Courville: *Modulating early visual processing by language*. NIPS 2017: pp6597-6607