FabLe : Framework for Automatic Interpretability in Machine Learning

Summary table of persons involved in the project:

Partner	Family name	First name	Current position	Role & responsibilities in the project (4 lines max)	Involvement (person.month) through the project's total duration
Inria Rennes / LACODAM	*GALÁRRAGA*	*Luis*	*Researcher (CR2)*	*Scientific coordinator and thesis advisor*	*70 % * 48 = 33.6*
Univ. Rennes 1 / LACODAM	*TERMIER*	*Alexandre*	*Professor*	*Collaborator and thesis advisor*	*25 % * 48 = 12*
Univ Rennes 1 / LACODAM	*FROMONT*	*Elisa*	*Professor*	*Collaborator and thesis advisor*	*25 % * 48 = 12*
Univ Rennes 1 / DRUID	*BOUADI*	*Tassadit*	*Associate Professor*	*Collaborator for the design of user studies with crowd-sourcing platforms*	*12.5 %* 48 = 6*
ENSAI / LACODAM	*GAUDEL*	*Romaric*	*Associate Professor*	*Collaborator for the selection of machine learning algorithms and use cases*	*6.25 % * 48 = 3*
Télécom ParisTech / DIG	*DESSALLES*	*Jean-Louis*	*Associate Professor*	*Collaborator for the design of a simplicity model for ML explanations*	*6.25 % * 48 = 3*

I. Proposal’s context, positioning and objective(s)

a. Context and motivation

Recent technological advances rely on accurate decision support systems that often operate as black boxes. That is, the rationale behind the system's answers cannot be explained to users. This can happen due to business strategical reasons, or because the system's logic is too complex. Take as examples accurate machine learning (ML) methods such as deep neural networks or random forests. While they normally deliver very high accuracy for classification and regression tasks, their verdicts can hardly be explained to users. For instance, a deep convolutional neural network (ConvNet) trained for image classification may model each pixel in the image as a neuron in the input layer. These input neurons are the starting points of a complex composition of linear and non-linear operations that allow the network to classify an image as a dog or as a cat. Even if the mathematical formulation that drove the design of the ConvNet is well-understood by machine learning experts, its inner workings on particular instances are extremely intricate.

This lack of explanation can lead to technical, ethical, and legal issues. For example, if the black-box control module of a self-driving car fails at detecting a pedestrian, it becomes crucial to inquire the cause of the error[1], which implies to *open the black box*. Recent work [1, 2] has shown the utility of explanations in debugging complex ML classifiers. In [2] it is shown that explanations can be very effective when end users want to personalize a machine learning system –a task known as exploratory debugging. Users with access to high level explanations about the behavior of a text classifier were able to boost the system’s performance with fewer interactions than those users who had access only to the classifier’s answer. In addition, the work in [1] shows how explanations can help (i) understand the poor performance of image and text classifiers on unseen instances, and (ii) spot biases in the training data. Indeed, a system trained on biased data may reflect unacceptable behaviors that can generate distrust and even perpetuate prejudice and unfairness. That was the case of the Tay Twitter chatbot developed by Microsoft, which had to be deactivated 24 hours after its publication due to its racist and antisemitic statements[2]. But Tay is not an isolated case. The authors of [3], for example, found out that text classifiers based on word embeddings and trained on Web data exhibit racial and gender biases. The study in [3] shows that last names common among black Americans are “closer” to terms about unpleasant concepts in the embedding space, whereas the opposite is true for last names associated to white Americans. A similar phenomenon was observed between female and male first names and certain activities and professions. A recently published article[3] showed that COMPAS, a predictive model for the risk of crime recidivism, has a strong ethnic bias. Among non-reoffenders, COMPAS is **almost twice more likely** to signal black people as high risk. Furthermore, COMPAS makes also the opposite mistake, in other words, white reoffenders are labeled as low risk much more often than black reoffenders. COMPAS is used by judges in USA to decide whether to grant or deny probation to offenders, hence, understanding how this model reaches its verdicts is of utmost importance towards a fair, transparent, and trustworthy justice system.

The interest in interpretable algorithms has gained momentum in Europe thanks to the General Data Protection Regulation (GDPR)[4]. This law, approved in May 2018 by the European Parliament, regulates

[1] https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

[2] https://www.bbc.com/news/technology-35902104

[3] http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] https://eugdpr.org/

the use of the personal data of European citizens. One of its clauses (see recital 71[5]) introduces the idea that humans should have the possibility to obtain meaningful explanations of decisions made via algorithmic means. In the golden age of data, this clause concerns a multitude of private and public organizations that rely on black-box systems to make decisions. While there exist methods, such as decision/regression trees and rule-based classifiers, that are per-se interpretable, they lag behind its black-box fellows in terms of accuracy. Since this interpretability-accuracy trade-off seems ineluctable, a trendy line of research focuses on interpretability modules for supervised ML algorithms. These modules act as interpreters between these black-box accurate systems and the end users. Some modules are designed for specific black-box algorithms, such as neural networks [5], however the trend is moving towards black-box agnostic explanations [4]. In regards to the scope of the explanation, there exist two notable families of modules. *Global interpretability modules* provide an integral explanation of the decisions of a system for all possible outcomes [4], i.e., *they tell the whole truth*. Alas, global explanations are very sensitive to the interpretability-accuracy trade-off. A simplification of a complex mathematical model will inevitably incur a loss in accuracy and fidelity[6], which will diminish the user's trust in the explanation [7]. Vice versa, an increase in accuracy and fidelity can be only achieved at the expense of longer and more complex explanations. For this reason, the most recent body of research in interpretable modules for ML algorithms concentrates on *local explanations*, namely explanations for the answers of the black-box in the vicinity of an individual instance of interest. By focusing on a region of the instance space, we can reduce the complexity of the target model and diminish the impact of the interpretability-accuracy trade-off.

In all cases, explanations take the form of white-box models such as decision/regression trees, rule ensembles, and linear models on a space of interpretable features. This surrogate space of features consists of artifacts, i.e., concepts, that are easy to understand by humans. For example, explanations for image classifiers [1, 6] are expressed in terms of the areas of the image that contributed most to the classifier's answer (called superpixels in [1]), even though the underlying neural networks use the different color channels of a pixel as input features. In text classification, a classifier trained on word embeddings may be explained in terms of the occurrences of words. To generate such explanations, global interpretability modules induce a white-box model on the interpretable features from the answers of the black box on the whole training data. For local interpretability modules, the induction is limited to a neighborhood around the instance of interest (pink circle in Figure 1). In this line of thought, approaches such as LIME [1] and Anchors [4] explain the answers of a black-box classifier by means of *local* linear models (Figure 1) and rules on regions (Figure 3) respectively. The coefficients of a local linear model provide both (1) a quantification of feature local importance, and (2) a hyperplane that approximates the black-box decision boundary as shown in Figure 1. However, if an instance lies far from the actual black-box decision boundary (Figure 2), a linear explanation may become uninformative and sensitive to the neighborhood construction strategy. The latter phenomenon can lead to multiple feasible explanations as in Figure 2. While methods such as [17] guarantee the existence of a unique linear explanation with strong guarantees of local fidelity and consistency, distant instances still require the construction of larger neighborhoods, which undermines the advantages of locality. Moreover, linear explanations have not been tested on multi-class settings where rules on regions may be more suitable. For example, the blue box in Figure 3, called an *anchor* in [4], induces this rule explanation: "*if the age is between 20 and 40, and the salary is between 200 and 500, the system's answer will likely be* ☆**"**. Alternatively, the regions defined in Figure 3 may be used to induce the simple decision tree explanation in Figure 4.

[5] http://www.privacy-regulation.eu/en/r71.htm

[6] Fidelity is the accuracy of the explanation w.r.t. the answers of the black-box system

b. Research hypothesis and questions

How can we fully **automatically** choose the **best explanation** for a ML classifier? Answering this question is the raison d'être of FAbLe[7] (Framework for Automatic Interpretability in Machine Learning). By "best explanation", we mean the one with best *interpretability-fidelity* trade-off in a universe of possible explanations. While fidelity is well-defined as the accuracy of the explanation w.r.t. the answers of the black box [3], interpretability is a subjective concept that has not yet been formalized. Existing work [9] suggests that interpretability possesses two dimensions: an operational objective dimension, called *comprehensibility*, and a subjective dimension called *plausibility*. Comprehensibility is concerned with how fast or accurately users can use the explanation to perform the classification task of the underlying black box. Using the explanation in this way requires users to understand the explanation's logic, hence the general consensus favors short and compact explanations as they are easier to retain and master. While studies performed on decision trees of different depth confirm this intuition [11], users may still prefer longer explanations if they are more plausible. Plausibility is related to how likely users will accept and understand an explanation because it seems reasonable. In [9], it is suggested that plausibility depends heavily on the users' background and psychology: much like explanations of low fidelity generate distrust among users [7], explanations that contradict users' intuition or prior knowledge will likely be rejected regardless of their length. In fact, the study conducted in [12] showed that medical experts rejected the rules learned by a regression tree because they found them oversimplified, and presumably unreliable.

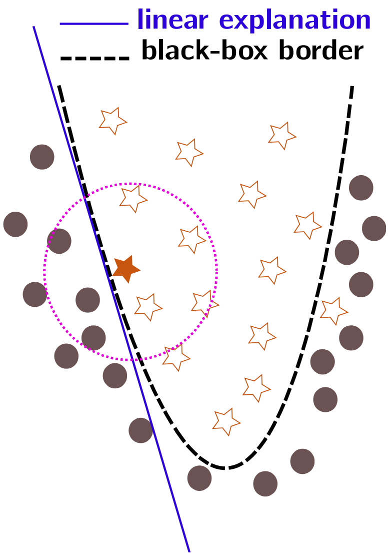

Figure 1: A linear explanation

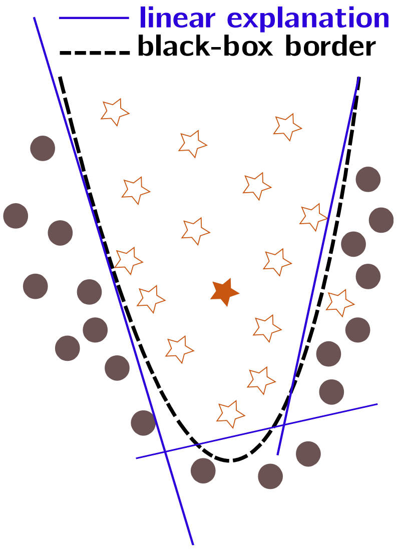

Figure 2: Multiple possible linear explanations

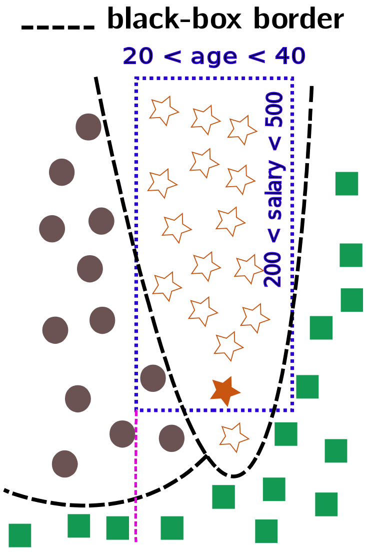

Figure 3: An anchor

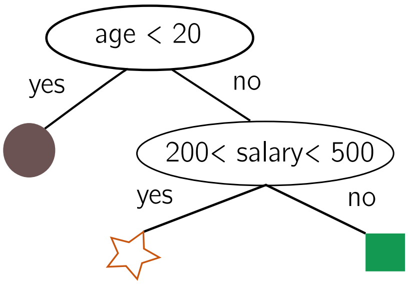

Figure 4: Explanation based on a decision tree

[7] We pronounced it /ˈfeɪbəl/ and much like in the literary genre, we aim at telling stories about classifiers. The stories always convey a moral and have interpretable features (instead of animals) as characters.

It follows that in order to answer the prime question we first need to answer a second question: "How can we formalize and quantify comprehensibility and plausibility?". Given the plethora of available explanation methods [1, 4, 8], automatically choosing the best explanation for a given use case requires us to quantify interpretability across models. We must therefore inquire under which circumstances a linear attribution model is more interpretable than an anchor or a decision tree. For this purpose our quantification framework should (i) measure all existing models using a common unit, and (ii) take into account both dimensions of comprehensibility and plausibility. Once interpretability across models has been quantified, we can design search strategies that will explore the space of possible explanations and report the most interpretable and faithful. In FAbLe we will focus on local interpretable models due to their higher resilience to the interpretability-accuracy trade-off.

c. Research Objective

Much like research in automatic machine learning has delegated the task of accurate model selection to computers [10], FAbLe aims at fully delegating the selection of interpretable explanations to computers. Our goal is to produce a suite of algorithms that will compute suitable explanations for ML algorithms based on our insights of what is interpretable. The algorithms will choose the best explanation method based on the data, the use case, and the user's background. We will implement our algorithms so that they are fully compatible with the body of available software for data science (e.g., AutoML [10], Scikit-learn[8]).

d. Position of the project as it relates to the state of the art

To best of our knowledge FAbLe is the first attempt to study and quantify interpretability of explanations across models in a user-driven fashion. Existing works [9, 11, 12] have studied the effect of some parameters on the comprehensibility and/or plausibility of models of the same nature, e.g., rules list vs. rules ensembles, decisions trees of different depth, etc. Thus, these studies can answer questions such as "when are longer rules deemed more interpretable than shorter rules?" or "does interpretability in decision trees correlate with tree depth?". FAbLe, on the other hand, intends to translate all these models into a common unit: quantity of information. We claim that such a formalization can capture both the dimension of comprehensibility and plausibility and account for the cases when long explanations are deemed more interpretable. In this spirit, a short decision tree composed of statements that contradict the user's expertise or composed of concepts unknown to the user, could be encoded with more bits and deemed complex and inadequate. Methods such as [13] have used the notion of information quantity (via the criterion of minimum description length) to guide the learning of decision trees, however such approaches define quantity of information neither in regards to the user's prior knowledge, nor across models of different nature. Moreover, given the recent rise of linear explanations [1], there are no comparative studies of such models with traditional explanations based on rule ensembles and decision trees.

It is crucial to mention that FAbLe is also the first initiative to bring the generation of explanations to the realm of fully automatic data science. The most remarkable efforts in automatic ML [10] focus on automating the selection of the right black-box and its parametrization. In our vision, the generation of explanations can also be fully automated: users (e.g., data scientists) can therefore provide the training data to an approach such as [10], generate an accurate black-box that will be sent as input to FAbLe, and obtain a comprehensible high-level explanation of its logic when tested on instances[9].

[8] http://scikit-learn.org/

[9] Recall that we focus on local interpretability, thus explanations are constructed for an instance of interest.

e. Methodology and risk management

Methodology

In order to answer FAbLe's prime research question, we must solve two sub-problems detailed in the following.

A) Quantification of interpretability. In a first stage we will conduct user studies aimed at shedding light on the perceived interpretability of the different existing explanation models, namely linear functions, anchors (which are tantamount to rule ensembles), and decision trees. These studies will make use of crowd-sourcing platforms such as Amazon Mturk[10] or Foule Factory[11]. Two members of the research consortium (Tassadit Bouadi and Luis Galárraga) have already experience in conducting experiments with this type of platforms. In regards to the data used to learn the black-box classifiers, we will make use of publicly available datasets from sites such as Kaggle[12]. Even though data can take the form of tables (structured data), unstructured text, images, or time series, we will focus on tabular and textual data, because such data types accept multiple explanation models[13]. The user studies will allow us, for example, to understand whether (or when) a decision tree like the one in Figure 4 is more interpretable than the rule induced by Figure 3 or the linear attribution model in Figure 1. We will resort to information theory to quantify the amount of conveyed information of an explanation model in bits. Paradigms such as MDL (minimum description length) offer an attractive theoretical ground for this purpose. The goal of MDL is to find an encoding **H** for the symbols of an alphabet such that the length of the encoding **L(H)** plus the size of the data **D** when compressed with H, i.e., **L(D|H)**, is minimal. Analogously, we propose a model-agnostic encoding for explanations so that interpretable explanations compress well. This encoding can be seen as a formalization of the notion of interpretability and its comprehensibility and plausibility components. In this line of thought, the length of an explanation's code will depend both on the number of artifacts (e.g., number of conditions of a rule, number of attributes in a linear attribution model) and on how much those artifacts contradict the users' intuition and preferences. For example, if we know that users of a movie recommendation system perceive the actors and the prizes of a movie as better indicators of quality than the language of a movie, our encoding will assign fewer bits to the first attributes. It follows that, all else being equal, explanations based on the actors of a movie will be preferred over those based on the language as they are intuitively "simpler" for users, i.e., they convey less information. Our user studies will, thus, take care of collecting such preferences for the chosen use cases. We highlight that the design of a universal encoding for explanations falls within the domains of competence of the consortium (See Section II, Coordinator and Consortium). We now list all the tasks associated to this sub-problem in the following:

1. Selection of use cases. This includes the selection of datasets with clear classification tasks, as well as the selection of different accurate black-box models to carry out those tasks. We will focus on classification problems that do not require specialized domain expertise. Examples are sentiment analysis in text, and recommendation tasks, among others.

2. Design and execution of user studies to shed light on the (i) the users' explanation preferences, and (ii) the model traits with the highest impact on the perceived interpretability among **models of the same type**. Point (i) accounts for the dimension of

[10] https://www.mturk.com/

[11] https://www.foulefactory.com/

[12] https://www.kaggle.com/

[13] Image classification can be only explained via linear attribution models on superpixels, and –to the best of our knowledge– agnostic explanations for time series classifiers have not yet been studied.

plausibility, whereas point (ii) aims at measuring the comprehensibility of the explanations. As suggested in [15], we can measure comprehensibility by asking users to apply the explanation to make predictions and registering both their invested time and the fidelity of their answers. Such experiments can helps us find out, for example, whether tree depth has a positive impact in the comprehensibility of decision trees or up to how many attributes a user is willing to read when confronted with a linear attribution model. Another interesting question is how much cognitive load an extra condition in a rule does carry compared to an additional attribute in a linear model.

3. Design and execution of user studies to compare the comprehensibility of the different explanation models. For instance, we could compare the best-ranking models of each type identified in the previous round of experiments and see how they rank when they are offered to users as explanations for a given black box.

4. Design of an encoding based on our findings in the user studies. Such encoding could rely on a cost model for the different components of an explanation. If we figure out, for instance, that an additional attribute in a linear explanation model affects users' performance more that an additional condition in a rule, we should ensure that a rule condition is encoded with more bits than an attribute in a linear model.

B) Automating the generation of explanations. Once we understand what users deem interpretable and we can measure interpretability for models in bits, we will design search algorithms that can report the most accurate and understandable explanations from a universe of alternatives. This universe consists of explanations generated via various methods (anchors, linear models, and decision trees), with different levels of interpretability and fidelity, and a myriad of possible parametrizations (e.g., decision trees of different depths, linear explanations on different neighborhoods). Confronted with such a large and heterogeneous space, we may resort to search methods based on beam search, or even more complex paradigms such as branch and bound or Monte-Carlo tree search [14]. To search among explanation models of the same nature (e.g., two linear attribution models), we may borrow inspiration from the literature in automatic ML. The work in [6] applies Bayesian optimization to find the hyper-parameters of a ML model that deliver the best accuracy (fidelity in our case) in cross-validation. The search for good hypotheses in large spaces is one of the specialties of FAbLe's host team (LACODAM).

In a final stage, we will evaluate the performance of FAbLe with the users in order to determine to which extent our explanations correlate with users' notion of interpretability.

Risks associated to project

We identify two types the risks in the execution of FAbLe. On the one hand, there exist logistic risks related to timing such as delays in recruitment of human resources such as PhD students and interns. On the other hand, there exist risks related to the research hypotheses and the experimental setup.

Logistic risks. We envision to hire one PhD student, an intern, and an engineer for the design of FAbLe. Knowing that the availability of suitable candidates does not depend on us, we are requesting a project duration of 48 months that includes 36 months for the PhD thesis plus one year of backup to account for recruiting eventualities.

Research risks. The validity of our research proposal is subject to the validity of our experimental setup. For this reason, we have contemplated a final round of experiments to empirically assess the validity of the FAbLe approach. We observe that this phase might require more than one iteration. If our cost model and encoding does not correlate enough with users' notion of interpretability, we may have to carry out adjustments in the model and re-evaluate it. Such event would incur more resources, already contemplated in our proposal (12 months of backup, 2k of supplementary costs for user studies). In regards to the quality of the answers provided by the participants of the user studies, we recall that crowd-sourcing platforms offer multiple mechanisms to assess the participants before accepting their answers. These mechanisms include a filtering by profile and introductory tests to estimate the reliability of the participants.

II. Organisation and implementation of the project

a. Scientific coordinator and its consortium / its team

The project will be coordinated by Luis Galárraga and will be hosted by the LACODAM[14] project-team at the IRISA/Inria research center. LACODAM brings together researchers from the University of Rennes 1, Inria, CNRS, INSA, ENSAI, INRA, and Agrocampus Ouest. Besides hosting the project, LACODAM will also contribute with its expertise in the different challenges of this research proposal. The team specializes in pattern mining and its different aspects, namely search space modeling, scalability, and user-system interactions. LACODAM also counts on recognized competences in ML becoming the leader team of the Inria Project Laboratory HyAIAI[15]. This research initiative –financed by Inria from 2019 to 2022 and composed of five other Inria teams–, aims at answering multiple open research questions in the domain of interpretable ML. All these questions are complementary to the objective of FAbLe and are concerned with issues such as integrating explicit user constraints in explanations, the semantification of the explanations in image classification (i.e., give meaning to super-pixels), or the design of multi-instance local explanations, among others. Three members of the consortium, namely Alexandre Termier, Elisa Fromont, and Luis Galárraga are part of the research committee of HyAIAI (see Table 1).

We elaborate on the research consortium in the following.

- **Luis Galárraga (project coordinator).** Luis holds a PhD in Computer Science, defended in 2016, from Télécom ParisTech. Currently, he is an Inria researcher and member of LACODAM since October 2017. Before, he was a post-doctoral researcher in the DAISY team of Aalborg University. His research interests comprise mainly three domains: pattern mining, semantic web, and interpretable ML. During his PhD and post-doctoral fellowship, he worked on topics that reconciliate the two first domains [15, 16], whereas at LACODAM he is involved with problems in the area of interpretable ML such as the mining of natively interpretable regression models (work under submission). As member of LACODAM he has also been in charge of multiple supervision duties, including four internships, and two PhD theses (one in co-supervision with the DAISY team from Aalborg University). He is part of the direction of the research axis "Visual query answering" of the HyAIAI initiative for the semantification of explanations for image classifiers. He is also one of the organizers of the first edition of the AIMLAI[16] workshop co-located with the conference EGC 2019 (Extraction et Gestion des Connaissances). The objective of this workshop is to become a discussion venue for the French-speaking community about the advent of novel interpretable algorithms and interpretability modules that mediate the communication between complex ML/AI systems and users.

[14] https://team.inria.fr/lacodam/

[15] Hybrid Approaches for Interpretable AI

[16] https://project.inria.fr/aimlai/

- **Alexandre Termier.** Alexandre is a professor at the University of Rennes 1 and the head of the LACODAM team. His main research area is pattern mining. As a pattern mining specialist, he has an important expertise in the exploration of large search spaces. Recently he has started to work on the application of information theory tools, namely MDL, for model selection tasks in pattern mining [18].
- **Elisa Fromont.** Elisa is a professor at the University of Rennes 1 and member of LACODAM since September 2017. She is specialized in ML (notably deep learning) for fraud and anomaly detection and semantic scene labelling. Elisa has also worked on graph mining for object tracking in videos.
- **Romaric Gaudel.** Romaric is an assistant professor at ENSAI and external collaborator of LACODAM. He is specialized in ML (notably clustering), online recommender systems, and bandit theory.

The consortium will additionally count on the external support of:

- **Tassadit Bouadi.** Tassadit is an associate professor at the University of Rennes 1 and member of the DRUID[17] team at the IRISA research center. Tassadit's main research interests comprise data warehousing and skyline query optimization. As part of DRUID, she also works on the management of uncertain, user-generated, and crowd-sourced data, notably from the perspective of privacy. She co-organized the workshop AIMLAI at EGC in collaboration with Luis Galárraga.
- **Jean-Louis Dessalles.** Jean-Louis is professor from the DIG[18] team of Télécom ParisTech. His research interests comprise the fields of cognitive sciences and information theory. He has worked on cognitive modeling of different behavioral phenomenons such as emotional intensity, narrative interest, and relevance in argumentative discussions.

We highlight the team's diversity and complementary of skills. The quantification of interpretability (sub-problem A in Section "Methodology") requires expertise in information theory, MDL, and cognitive modeling for the explanation preferences of users. LACODAM counts on competences in the two first fields [18, 19], whereas we count on our partnership with Télécom ParisTech for the cognitive modeling of users' bias. LACODAM also has plenty of experience in search in large spaces (sub-problem B) as this is a central task in pattern mining. We also emphasize the pertinence of our experience in rule mining, which will be handy at optimizing and personalizing existing rule-based explanations methods such as [8]. Furthermore, our competence in skyline query optimization is also pertinent since we aim at optimizing two potentially conflicting objectives, namely fidelity and interpretability (quantified in number of bits). Last but not least, our competence in ML becomes crucial when studying the behavior of complex classification algorithms. All things considered, the issues that motivate FAbLe define a clear application of the research lines and competences of LACODAM and the proposed research consortium and will lay the groundwork for further research in the not-yet-fully developed topic of interpretable ML.

As detailed in the next section, we will count on the support of a PhD student and an intern to carry out the experiments required to answer FAbLe's research questions. Figure 2 presents a Gantt chart of FabLe's work plan and assignment of tasks to human resources. The tasks are detailed in Table 3.

[17] https://www-druid.irisa.fr/

[18] https://dig.telecom-paristech.fr/

Table 1 : Implication of the scientific coordinator in on-going project(s)

Name of the researcher	Person.month	Call, funding agency, grant allocated	Project's title	Name of the scientific coordinator	Start - End
Luis Galárraga	20% * 48 = 9.6 months	Inria Project Labs[19]	Hybrid Approaches for Interpretable AI	Alexandre Termier	2019-2022

b. Implemented and requested resources to reach the objectives

The requested funding (195 208 EUR) will essentially allow us to recruit one **full-time PhD student** for a period of 3 years and a **full-time level 1 engineer** (master degree M2, with 0 to 2 years of experience) for one year. This differs from the amount requested in the pre-proposal (170k EUR) as we did not initially contemplate the need of an engineer. Given our goal of fostering the adoption of FabLE by organizations and individuals, we believe that it is crucial to release a well-designed, well-documented, and user-friendly implementation of our algorithms. Moreover, platforms such as Scikit-learn have very high engineering and research standards[20], which suggests us to decouple the research part (conducted mainly by the PhD student) from the engineering and software development part.

As explained in the previous section, we estimate at least three rounds of user studies: studies for explanation models of the same type, studies for explanation models of different types, and validation studies for the evaluation of FAbLe. The PhD student will be responsible of (i) conducting those experiments, (ii) compiling the knowledge produced, and (iii) translating this knowledge into algorithms for the automatic generation of explanations for ML classifiers. Because of the significant amount of work on design, execution and analysis required by the user studies, we envision to hire an (M1 or M2) intern **from LACODAM's budget (3k)** for a period of six months. The intern will help us in the execution of the first round of experiments, as we expect to master the process for the subsequent rounds of user studies. Once the algorithmic principles of FAbLe are well-defined and evaluated, we will hire an engineer to implement the final versions of the algorithms so that they are compatible with major ML software libraries such as Scikit-learn. We estimate then a cost of approximately 168k EUR ("staff expenses" in Table 1) in human resources comprising approx. 120k EUR for the PhD student, and 48k for the engineer. The engineer, the PhD student, and the intern will be supervised by the project's coordinator in collaboration with other members of the consortium.

The execution of the user studies will also incur important costs due to the need of crowd-sourcing. The user studies will take the form of questionnaires asking for the preferences of users among different explanation models for ML algorithms. If we estimate one hour of total work per participant at a cost of at least 10.03 EUR (minimal hourly wage salary in France), we would incur in costs of at least 1k EUR for a round of experiments with 100 candidates. This amounts to 3k EUR for 3 rounds of experiments. We consider 5k to account for additional rounds of user studies caused by late adjustments of FAbLe. This amount is considered in the category "Outsourcing/subcontracting" in Table 1. We also contemplate a cost of 3k EUR for the computers assigned to the PhD student and the engineer. Conversely, the intern will make use of the computer resources (workstations) available in the IRISA laboratory. In addition, we estimate a cost of 2 times 2.5k EUR for conference missions as we envision to publish two conference articles in international venues (we estimate 1k for conference registration and 1.5k for transportation, food and hotel). Finally our financial service estimates 14.4k EUR in administrative and management costs.

[19] https://www.inria.fr/en/research/research-teams/inria-project-labs

[20] https://scikit-learn.org/stable/faq.html

Table 2 : Requested means by item of expenditure and by partner*

		FabLE’s consortium	**Partner**	**Partner** ***Intitulé***	**Partner** ***Intitulé***
Staff expenses		167 748 €			
Instruments and material costs (including the scientific consumables)		3k €			
Building and ground costs					
Outsourcing / subcontracting		5k €			
General and administrative costs & other operating expenses	Travel costs	5k €			
	Administrative management & structure costs**	14 459.84 €			
Sub-total		195 207.84 €			
Requested funding		**195 207.84 €**			

* The amounts indicated here must be strictly identical to those entered on the website. If both information are not consistent, if they were badly filled in or lacking, the information entered online will prevail on those reported in the submission form / scientific document.

** For marginal cost beneficiaries, these costs will be a package of 8% of the eligible expenses. For full cost beneficiaries, these costs will be a sum of max. 68% of staff expenses and max. 7% of other expenses.

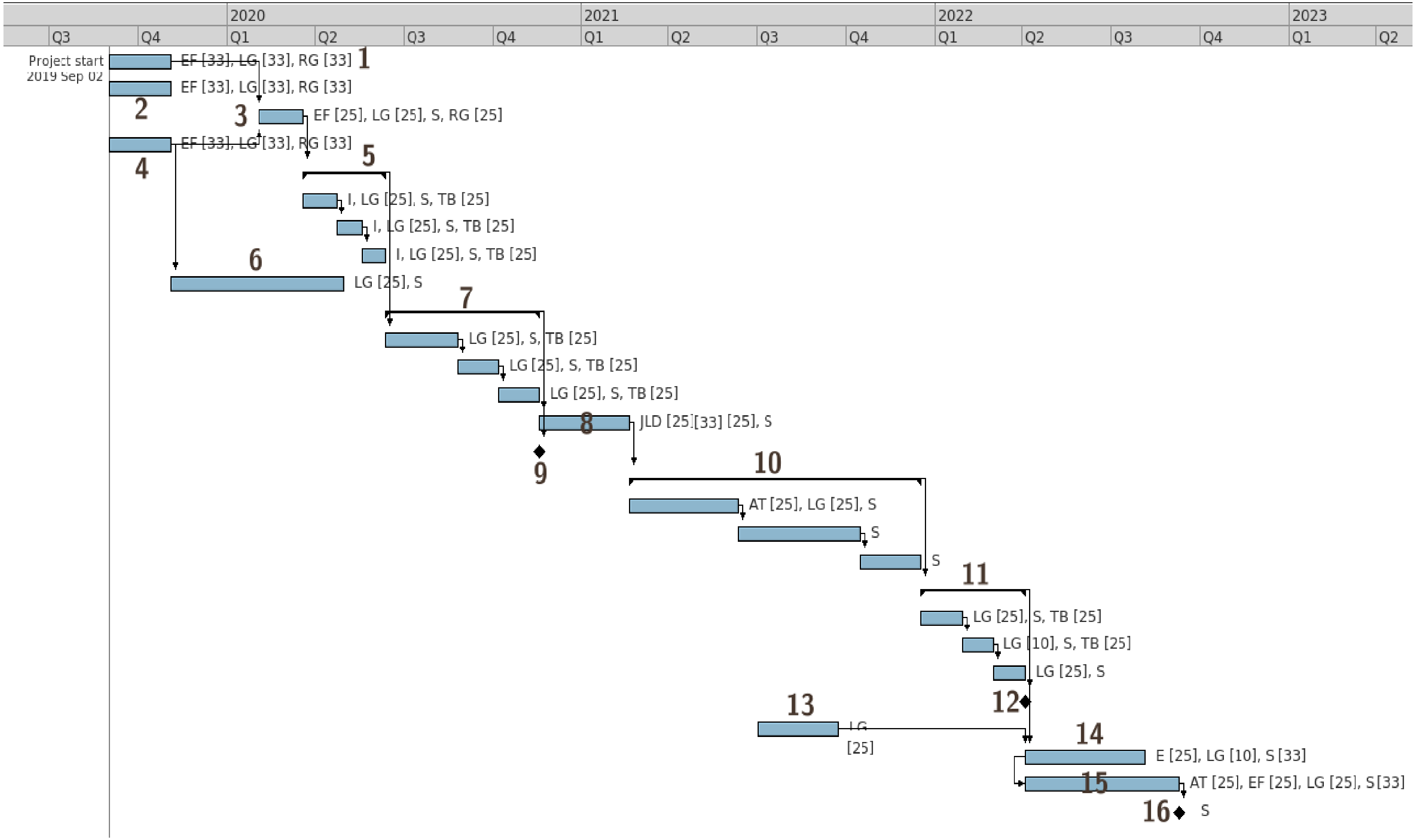

Figure 1 : FabLe’s Gantt chart. The numbers correspond to the tasks defined in Table 3. The initials denote the involved human resources, **EF**=Elisa Fromont, **LG**=Luis Galárraga, **AT**=Alexandre Termier, **JLD**=Jean-Louis Dessalles, **RG**=Romaric Gaudel, **TB**=Tassadit Bouadi, **S**=Phd Student, **I** = intern, **E**=engineer. The numbers in brackets are estimates of the person’s involvement (%) during a particular task.

Task	Description
1	Selection of use cases
2	Recruitment of intern (I)
3	Training of accurate classifiers on use cases
4	Recruitment of PhD student
5	User studies for explanations models of the same type
5.1	Experiments’ design and implementation
5.2	Experiments’ execution
5.3	Analysis of results
6	Revision of the state of the art in interpretable ML
7	User studies for explanation models of different type
8	Design of cost model and encoding for explanation models (Problem A in methodology)
9	Publication of the results of the user studies
10	Automatic generation of explanations (design of FAbLe)
10.1	Revision of the state of the art in search algorithms
10.2	Design and implementation of the search strategy for explanations (Problem B in methodology)
10.3	Quantitative evaluation of FAbLe (runtime, number of admissible explanations)
11	User studies on FAbLe (qualitative evaluation)
11.1	Experiments’ design and implementation
11.2	Experiments’ execution
11.3	Analysis of results
12	Publication of FAbLe
13	Recruitment of engineer (E)
14	Integration of FAbLe with major ML software libraries
15	Preparation towards PhD defense
16	PhD defense

Table 3 : Fable’s tasks, milestones are highlighted.

III. Impact and benefits of the project

Social and Economic Impact

As the need for interpretable ML becomes compelling, we expect FAbLe to have a timely economical and societal impact. Automatic simple explanations will make data science even more accessible to people with any background, which will fuel the adoption of the GDPR by individuals and organizations. In this spirit, organizations relying on automatic decision-making will be able to generate high-level explanations for the answers of their systems. This will not only allow them to deliver explanations if a customer or client requires it, but it will also provide a deeper understanding of the system's inner workings and promote a culture of self-audit and transparency. The availability of automatic explanations will also provide a thrust to the productivity of data scientists, which will undoubtedly have a positive economic impact for the concerned organizations.

Diffusion Strategy

Our diffusion strategy consists of three major axes:

Publication. We envision to publish our research results in international conferences (2 papers) and journals (1 paper) in the domain of machine learning (ECML/PKDD, ICML, JMLR) and data management (KDD, CIKM, ICDE).

Software availability. In order to facilitate the adoption of FAbLe by data scientists (and their corresponding organizations), we will provide an implementation of our algorithms that is compatible with popular software libraries such as Scikit-learn.

Scientific activities. Given the consortium's privileged position as coordinator of the HyAIAI initiative and the AIMLAI workshop, we envision to leverage these opportunities to share our research findings with the research community in interpretable ML/AI. Moreover, we have proposed a second edition of the AIMLAI workshop for the ECML/PKDD conference 2019 in order to target a broader research community, and widen our contact network.

IV. References related to the project

1. Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ***Why Should I Trust You?: Explaining the Predictions of Any Classifier***. ACM SIGKDD Conference on Knowledge Discovery. DOI: 10.1145/2939672.2939778

2. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. ***Principles of Explanatory Debugging to Personalize Interactive Machine Learning***. ACM Conference on Intelligent User Interfaces. At: http://openaccess.city.ac.uk/13819/1/paper326.pdf

3. Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. ***Semantics Derived Automatically from Language Corpora Contain Human-like Biases***. 2017. Journal *Science*, *356*(6334), 183-186. DOI: https://doi.org/10.1126/science.aal4230

4. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. ***A Survey of Methods for Explaining Black Box Models***. ACM Computing Survey. DOI: 10.1145/3236009

5. M. G. Augasta and T. Kathirvalavakumar. ***Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems***. 2012. Neural Process Letters, 35(2), 131–150. DOI: https://doi.org/10.1007/s11063-011-9207-8

6. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. ***Learning Deep Features for Discriminative Localization***. 2016. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. At: *http://cnnlocalization.csail.mit.edu/*

7. Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, Weng-Keen Wong. ***Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models***. 2013. Proceedings of the 2013 IEEE Symposium on Visual Languages and Human-Centric Computing. At: http://openaccess.city.ac.uk/6344/3/VLHCC2013.pdf

8. Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. ***Anchors: High-Precision Model-Agnostic Explanations***. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. At: https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

9. Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. ***On Cognitive Preferences and the Plausibility of Rule-based Models***. 2018. At: https://arxiv.org/abs/1803.01316

10. Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost T. Springenberg, Manuel Blum, and Frank Hutter. ***Efficient and Robust Automated Machine Learning***. 2015. Annual Conference on Neural Information Processing Systems. At: http://dl.acm.org/citation.cfm?id=2969442.2969547

11. Rok Piltaver, Mitja Luštrek, Matjaž Gams, and Sanda Martinčić-Ipšić. ***What Makes Classification Trees Comprehensible?*** 2016. Expert Systems with Applications, Volume 62, 333-346, DOI: https://doi.org/10.1016/j.eswa.2016.06.009

12. Igor Kononenko. ***Inductive and Bayesian Learning in Medical Diagnosis***. 1993. Applied Artificial Intelligence. DOI:*10.1080/08839519308949993*

13. Manish Mehta, Jorma Rissanen, and Rakesh Agrawal. ***MDL-based Decision Tree Pruning***. 1995. ACM SIGKDD Conference on Knowledge Discovery. https://www.aaai.org/Papers/KDD/1995/KDD95-025.pdf

14. Cameron B. Browne *et al*. ***A Survey of Monte Carlo Tree Search Methods***. 2012. IEEE Transactions on Computational Intelligence and AI in Games, Volume 4(1), 1-43. DOI: https://doi.org/10.1109/TCIAIG.2012.2186810

15. Luis Galárraga, Christina Tefloudi, Katja Hose, and Fabian Suchanek. ***AMIE : Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases***. 2013. World Wide Web Conference. At: http://luisgalarraga.de/docs/amie.pdf

16. Danai Symeonidou, Luis Galárraga, Nathalie Pernelle, Fatiha Saïs, and Fabian Suchanek. ***VICKEY: Mining Conditional Keys on Knowledge Bases***. 2017. International Semantic Web Conference. At: https://hal.archives-ouvertes.fr/hal-01647597

17. Scott M. Lundberg, Su-In Lee. **A Unified Approach to Interpreting Model Predictions**. 2017. Neural Information Processing Systems (NIPS). At : http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

18. Esther Galbrun, Peggy Cellier, Nikolaj Tatti, Alexandre Termier, and Bruno Crémilleux. **Mining Periodic Patterns with a MDL Criterion**. 2018. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD). At https://hal.archives-ouvertes.fr/hal-01951722v1

19. Clément Gautrais, René Quiniou, Peggy Cellier, Thomas Guyet, Alexandre Termier. **Purchase Signatures of Retail Customers**. 2017. The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). At: https://hal.archives-ouvertes.fr/hal-01639795v1