RESEARCH CENTRE

Rennes - Bretagne Atlantique

IN PARTNERSHIP WITH:

Institut national des sciences appliquées de Rennes, Institut national supérieur des sciences agronomiques, agroalimentaires, horticoles et du paysage, Université Rennes 1

2020
ACTIVITY REPORT

Project-Team

LACODAM

Large scale Collaborative Data Mining

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team LACODAM

Creation of the Team: 2016 January 01, updated into Project-Team: 2017 November 01

Keywords

Computer sciences and digital sciences

A2.1.5. – Constraint programming

A3.1.1. – Modeling, representation

A3.1.6. – Query optimization

A3.1.11. – Structured data

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.3. – Data and knowledge analysis

A3.3.1. – On-line analytical processing

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A9.1. – Knowledge

A9.2. – Machine learning

A9.3. – Signal analysis

A9.6. – Decision support

A9.7. – AI algorithmics

A9.8. – Reasoning

Other research topics and application domains

B1.2.2. – Cognitive science

B2.3. – Epidemiology

B2.4.1. – Pharmaco kinetics and dynamics

B3.5. – Agronomy

B3.6. – Ecology

B3.6.1. – Biodiversity

1   Team members, visitors, external collaborators

Research Scientist

- Luis Galarraga Del Prado [Inria, Researcher]

Faculty Members

- Alexandre Termier [Team leader, Univ de Rennes I, Professor, HDR]

- Tassadit Bouadi [Univ de Rennes I, Associate Professor]

- Elisa Fromont [Univ de Rennes I, Professor, HDR]

- Thomas Guyet [Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, Associate Professor, HDR]

- Christine Largouët [Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, Associate Professor, HDR]

- Véronique Masson [Univ de Rennes I, Associate Professor]

- Laurence Rozé [INSA Rennes, Associate Professor]

Post-Doctoral Fellow

- Neetu Kushwaha [Inria, from Mar 2020]

PhD Students

- Johanne Bakalara [Univ de Rennes I]

- Erwan Bourrand [Advisor Sla, until Sep 2020]

- Lenaig Cornanguer [Univ de Rennes I, from Sep 2020]

- Julien Delaunay [Inria, from Sep 2020]

- Kevin Fauvel [Inria, until Sep 2020]

- Samuel Felton [Univ de Rennes I]

- Camille Sovanneary Gauthier [Louis Vuitton]

- Maël Guillemé [Univ de Rennes I, until Jan 2020]

- Victor Guyomard [Orange Labs, CIFRE, from Nov 2020]

- Colin Leverger [Orange Labs]

- Gregory Martin [Groupe PSA]

- Josie Signe [Univ de Rennes I, from Sep 2020]

- Antonin Voyez [Inria]

- Yichang Wang [China Scholarship Council]

- Heng Zhang [Atermes, CIFRE]

Technical Staff

- Remi Adon [Inria, Engineer]

- Louis Bonneau de Beaufort [Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, Engineer]

Interns and Apprentices

- Thibaud Balem [Happy Blue Fish, Intern, from Feb 2020 until Jul 2020]

- Lenaig Cornanguer [Univ de Rennes I, from Feb 2020 until Jul 2020]

- Simon Coumes [École normale supérieure de Rennes, from May 2020 until Jul 2020]

- Julien Delaunay [Inria, from Feb 2020 until Jul 2020]

- Mansor Gueye [Univ de Rennes I, from Jun 2020 until Sep 2020]

- Sami Rahhaoui [Agro Campus Ouest, from Jun 2020 until Aug 2020]

- Josie Signe [Inria, from Feb 2020 until Jul 2020]

Administrative Assistant

- Gaelle Tworkowski [Inria]

External Collaborators

- Philippe Besnard [CNRS, HDR]

- Romaric Gaudel [École nationale de la statistique et de l'analyse de l'information]

- Raphael Gauthier [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, until Nov 2020]

- Anne-Isabelle Graux [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]

2   Overall objectives

Data collection is ubiquitous nowadays and it is providing our society with tremendous volumes of knowledge about human, environmental, and industrial activity. This ever-increasing stream of data holds the keys to new discoveries, both in industrial and scientific domains. However, those keys will only be accessible to those who can make sense out of such data. This is, however, a hard problem. It requires a good understanding of the data at hand, proficiency with the available analysis tools and methods, and good deductive skills. All these skills have been grouped under the umbrella term "Data Science" and universities have put a lot of effort in producing professionals in this field. "Data Scientist" is currently an extremely sought-after job, as the demand far exceeds the number of competent professionals. Despite its boom, data science is still mostly a "manual" process: current data analysis tools still require a significant amount of human effort and know-how. This makes data analysis a lengthy and error-prone process. This is true even for data science experts, and current approaches are mostly out of reach of non-specialists.

The objective of the LACODAM is to facilitate the process of making sense out of (large) amounts of data. This can serve the purpose of deriving knowledge and insights for better decision-making. Our approaches are mostly dedicated to provide novel tools to data scientists, that can either performs tasks not addressed by any other tools, or that improve the performance in some area for existing tasks (for instance reducing execution time, improving accuracy or better handling imbalanced data).
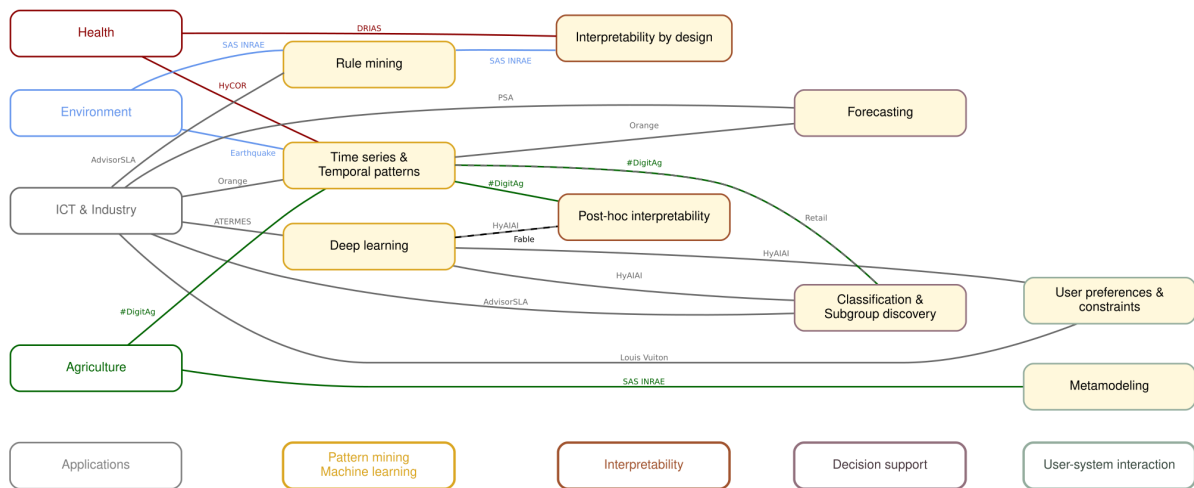
Figure 1: LACODAM research topics organized by axis and applications

3   Research program

3.1   Introduction

The research of the Lacodam project team is organized around four research axes, the first three being based on the long standing research interests of the team members, and the last one (interpretability) being a topic that emerged more recently in the machine learning community. The team rapidly committed itself to this novel research area, as it sits perfectly at the intersection of our traditional research interests, and has strong societal expectations.

- The first research axis (Section 3.2) is dedicated to the design of *novel pattern mining and machine learning methods*. Pattern mining is one of the most important approaches to discover novel knowledge in data, and one of our strongest areas of expertise. Machine learning (ML), especially classification, is the basis for many decision support systems. The work on this axis will thus serve as foundations for work on the other two axes.

- The second axis (Section 3.3) tackles another aspect of knowledge discovery in data: the *interaction between the user and the system* in order to co-discover novel knowledge. Our team has plenty of experience collaborating with domain experts, and is therefore aware of the need to improve such interaction.

- The third axis (Section 3.4) concerns *decision support*. With the help of methods from the two previous axes, our goal here is to design systems that can either assist humans with making decisions, or make relevant decisions in situations where extremely fast reaction is required.

- The fourth axis (Section 3.5) concerns *interpretable machine learning* and can be considered as a transversal axis that has roots in all the other three axes.

Figure 1 sums up the detailed work presented in the next few pages: we show the four research axes of the team (X-axis) on the left and our main applications areas (Y-axis) below. The colored nodes in the middle represent the precise research topics of the team aligned with their corresponding axis and main application area. These research topics will be described in this section. Lines represent projects that can link several topics and application areas.

3.2   Novel Pattern Mining and Machine Learning Algorithms

Twenty years of research in pattern mining have resulted in efficient approaches to handle the algorithmic complexity of the problem. Existing algorithms are now able to efficiently extract patterns with complex

structures (ex: sequences, graphs, co-variations) from large datasets. However, when dealing with large, real-world datasets, these methods still output a large number of patterns, which is impractical for human analysis. This problem is called *pattern explosion.* The ongoing challenge of pattern mining research is to extract fewer but more meaningful patterns. The LACODAM team is committed to solve the pattern explosion problem by pursuing the following research topics:

- the design of dedicated algorithms for mining temporal patterns

- the design of flexible pattern mining approaches

- the automatic selection of interesting data mining results

The originality of our contributions relies on the exploration of knowledge-based approaches whose principle is to incorporate dedicated domain knowledge (aka application background knowledge) deep into the mining process. While most data mining approaches are based on agnostic approaches designed to cope with pattern explosion, we propose to develop data mining techniques that rely on knowledge-based artificial intelligence techniques. This entails the use of structured knowledge representations, as well as reasoning methods, in combination with mining.

The first topic concerns classical pattern mining in conjunction with expert knowledge in order to define new pattern types (and related algorithms) that can solve applicative issues. In particular, we investigate how to handle temporality in pattern representations which turns out to be important in many real world applications (in particular for decision support) and deserves particular attention.

The next two topics aim at proposing alternative pattern mining methods to let the user incorporate, on their own, knowledge that will help define their domain of interest. Flexible pattern mining approaches enable analysts to easily incorporate extra knowledge, for example domain related constraints, in order to extract only the most relevant patterns. On the other hand, the selection of interesting data mining results aims at devising strategies to filter out the results that are useless to the data analyst. Besides the challenge of algorithmic efficiency, we are interested in formalizing the foundations of interestingness, according to background knowledge modeled with either logical knowledge representation paradigms or with statistical approaches.

The team is also active in machine learning research. Several of our decision support applications rely on supervised classification algorithms. While supervised classification is a well researched area, there are several settings where the state of the art still requires improvements:

- when the classes are extremely unbalanced, it is difficult to learn a correct classifier. We are interested in techniques such as cost-sensitive learning to handle such situations.

- a lot of our applications deal with time series data: we are thus especially involved on proposing novel time series classifiers.

3.3   User/System Interaction

As we pointed out before, there is a strong need to present relevant patterns to the user. This can be done by using more specific constraints, background knowledge and/or tailor-made optimization functions. Due to the difficulty of determining these elements beforehand, one of the most promising solutions is that the system and the user co-construct the definition of relevance, i.e., to have a human in the loop. This requires to have means to present intermediate results to the user, and to get user feedback in order to guide the search space exploration process in the right direction. This is an important research axis for LACODAM, which will be tackled in several complementary ways as detailed next.

- *Domain Specific Languages:* One way to interact with the user is to propose a domain specific language (DSL) tailored to the domain at hand and to the analysis tasks. The challenge is to propose a DSL allowing the users to easily express the required processing workflows, to deploy those workflows for mining on large volumes of data and to offer as much automation as possible.

- *What if / What for scenarios:* We also investigate the use of scenarios to query results from data mining processes, as well as other complex processes such as complex system simulations or model predictions. Such scenarios are answers to questions of the type "what if [situation]?" or "what [should be done] for [expected outcome]?".

- *User preferences:* In exploratory analysis, users often do not have a precise idea of what they want, and are not able to formulate such queries. Hence, in LACODAM we investigate simple ways for users to express their interests and preferences, either during the mining process – to guide the search space exploration –, or afterwards during the filtering and interpretation of the most relevant results.

- *Data visualization:* All our research axes deal with scenarios where a system displays information or some sort of distilled knowledge (e.g., patterns, predictions, explanations) to users. For example, the output of most pattern mining algorithms is a (usually long) list of patterns. Interpretable machine learning and applications in decision support may require to visualize the output of an algorithm in a comprehensible way. Fueled by our work in interpretable and explainable machine learning, data visualization techniques (referred as DataVis) have become a transversal research topic in LACODAM. We are particularly interested in the effective visualization of patterns and explanations for machine learning algorithms. Since DataVis is beyond the expertise of LACODAM, our strategy is to establish collaborations with prominent data visualization teams for this line of research, with a long term goal to recruit a specialist in data visualization if the opportunity arises.

3.4   Decision support

The increasing volumes of data coming from a range of different systems (ex: sensor data from agro-environmental systems, log data from software systems and ML models, biological data coming from health monitoring systems) can help human and software agents make better decisions. Hence, LA-CODAM builds upon the idea that decision support systems should take advantage of the available data.

Our works in decision support are thus rooted in applications and conducted in partnership with applicative partners: mainly Inrae researchers for agro-environmental systems, but also companies in various domains (such as cyber-security, retail, transportation) for other applications. Our aim is to help users reach a better decisions by exploiting data mining and/or machine learning techniques over the available data. In some ways, the first axis of the team can be seen as the "engine" leading to the extraction of information from data, while this axis is interested in the complete chain from raw data to the knowledge provided to the user to help them make a better decision.

3.5   Interpretability

The pervasiveness of complex decision support systems, as well as the general consensus about the societal importance of understanding their logic[1], has given momentum to the field of interpretable and explainable ML. Being a team specialized in data science, we are fully aware that many problems can be solved by means of complex and accurate ML models. Alas, this accuracy sometimes comes at the expense of interpretability, which can be a major requirement in some contexts (e.g., regression using expertise/rule mining). For this reason, one of the interests of LACODAM is the study of the interpretability-accuracy trade-off. Our studies may be able to answer questions such as "how much accuracy can a model lose (or perhaps gain) by becoming more interpretable?". Such a goal requires us to define interpretability in a more principled way —an endeavour that has been very recently addressed, still not solved. LACODAM is interested in the two main currents of research in interpretability, namely the development of natively interpretable methods, as well as the construction of interpretable mediators between users and black-box models, known as post-hoc interpretability.

We highlight the link between interpretability and LACODAM's axes of decision support, and user/system interaction. In particular, interpretability is a prerequisite for proper user/system interaction and is a central incentive for the advent of data visualization techniques for ML models. This convergence has motivated our interest in *user-oriented post-hoc interpretability*, a sub-field of interpretable ML that adds the user into the formula when generating proper explanations of black-box ML algorithms. This rationale is supported by existing work [37] that suggests that interpretability possesses a subjective component known as plausibility. Moreover, our user-oriented vision meets with the notion of semantic interpretability, where an explanation may resort to high level semantic elements (objects in image

---

[1]General Data Protection Regulation, recital 71 http://www.privacy-regulation.eu/en/r71.htm

classification, or verbal phases in natural language processing) instead of surrogate still-machine-friendly features (such as super-pixels). LACODAM tackles these unaddressed aspects of interpretable ML with other Inria teams through the IPL HyAIAI.

3.6 Long-term goals

The following perspectives are at the convergence of the four aforementioned research axes and can be seen as ideal towards our goals:

- *A crisp definition for interpretability.* The formal definition of what is "interpretable" is still work in progress in the international research community. We would therefore like to find a more principled way to define interpretability. This would allow us to answer questions such as "is an explanation based on decision trees more suitable to a given user than an explanation based on linear attribution?". This poses multiple challenges such as characterizing the user's background and knowledge, defining the right metrics to quantify the interpretability of a model or explanation, and understanding the cognitive aspects of the existing methods and its visual representations.

- *Logic argumentation based on epistemic interest.* Having increasingly automated approaches will require better and better ways to handle the interactions with the user. A term goal is to explore the use of logic argumentation, i.e., the formalisation of human strategies for reasoning and arguing, in the interaction between users and data analysis tools. Alongside visualization and interactive data mining tools, logic argumentation can be a way for users to query both the results and the way they are obtained. Such querying can also help the expert to reformulate her query in an interactive analysis setting.

- *User-guided MDL model selection.* To mitigate the pattern explosion, the research in pattern mining has turnd the attention to methods borrowed from information theory, in particular the Minimum Description Length (MDL) principle. The idea is to select a small number of patterns that together allow to compress best the data. The team now has a solid expertise and reputation in this area. A problem of MDL approaches in pattern mining is that they are challenging to develop. Especially, they require to properly design an encoding for the patterns and the data, an activity that is mostly designed in an ad-hoc way by researchers, and which has some pitfalls. A long term objective would be to formalize general principles of a "good" MDL encoding for pattern mining, in order to be able to automatically produce encodings for different pattern mining problems. It could allow, for example, to incorporate user preferences in the encoding, in order to take them into account in the selection made by MDL, without breaking the good properties of the approach.

4 Application domains

The current period is extremely favorable for teams working in Data Science and Artificial Intelligence, and LACODAM is not the exception. We are eager to see our work applied in real world applications, and have thus an important activity in maintaining strong ties with industrials partners concerned with marketing and energy as well as public partners working on health, agriculture and environment.

4.1 Industry

We present below our industrial collaborations. Some are well established partnerships, while others are more recent collaborations with local industries that wish to reinforce their Data Science R&D with us (e.g. Energiency, Amossys).

- **Car Sharing Data Analysis**. Peugeot-Citroën (PSA) group's know-how encompasses all areas of the automotive industry, from production to distribution and services. Among others, its aim is to provide a car sharing service in many large cities. This service consists in providing a fleet of cars and a "free floating" system that allows users to use a vehicle, then drop it off at their convenience in the city. To optimize their fleet and the availability of the cars throughout the city, PSA needs to

analyze the trajectory of the cars and understand the mobility needs and behavior of their users. We tackle this subject together through the CIFRE PhD of Gregory Martin.

- **Multimodal Data Analysis for the Supervision of Sensitive Sites.** ATERMES is an international mid-sized company with a strong expertise in high technology and system integration from the upstream design to the long-life maintenance cycle. It has recently developed a new product, called BARIER TM ("Beacon Autonomous Reconnaissance Identification and Evaluation Response"), which provides operational and tactical solutions for mastering borders and areas. Once in place, the system allows for a continuous night and day surveillance mission with a small crew in the most unexpected rugged terrain. The CIFRE PhD of Heng Zhang aims at developing a deep learning architecture and algorithms to detect anomalies (mainly people) from multimodal data. The data are "multimodal" because information about the same phenomenon can be acquired from different types of detectors, at different conditions, in multiple experiments.

- **Root Cause Analysis in Networks**. AdvisorSLA is a French company specialized in software solutions for network monitoring. For this purpose, the company relies on techniques of network metrology. By continuously measuring the state of the network, monitoring solutions detect events (e.g., overloaded router) that may degrade the network's operation and the quality of the services running on top of it (e.g., video transmission could become choppy). When a monitoring solution detects a potentially problematic sequence of events, it triggers an alarm so that the network manager can take actions. Those actions can be preventive or corrective. Some statistics show that only 40% of the triggered alarms are conclusive, that is, they manage to signal a well-understood problem that requires an action from the network manager. This means that the remaining 60% are presumably false alarms. While false alarms do not hinder network operation, they do incur an important cost in terms of human resources. Thus, the CIFRE PhD of Erwan Bourrand proposes to characterize conclusive and false alarms. This will be achieved by designing automatic methods to "learn" the conditions that most likely precede each type of alarm, and therefore predict whether the alarm will be conclusive or not. This can help adjust existing monitoring solutions in order to improve their accuracy. Besides, it can help network managers automatically trace the causes of a problem in the network.

4.2 Health

- **Care Pathways Analysis for Supporting Pharmaco-Epidemiological Studies**. Pharmaco-epidemiology applies the methodologies developed in general epidemiology to answer to questions about the uses and effects of health products, drugs [41, 40] or medical devices [33], on population. In classical pharmaco-epidemiology studies, people who share common characteristics are recruited to build a dedicated prospective cohort. Then, meaningful data (drug exposures, diseases, etc.) are collected from the cohort within a defined period of time. Finally, a statistical analysis highlights the links (or the lack of links) between drug exposures and outcomes (*e.g.,* adverse effects). The main drawback of prospective cohort studies is the time required to collect the data and to integrate them. Indeed, in some cases of health product safety, health authorities have to answer quickly to pharmaco-epidemiology questions.

  New approaches of pharmaco-epidemiology consist in using large EHR (Electronic Health Records) databases to investigate the effects and uses (or misuses) of drugs in real conditions. The objective is to benefit from nationwide available data to answer accurately and in a short time pharmaco-epidemiological queries for national public health institutions. Despite the potential availability of the data, their size and complexity make their analysis long and tremendous. The challenge we tackle is the conception of a generic digital toolbox to support the efficient design of a broad range of pharmaco-epidemiology studies from EHR databases. We propose to use pattern mining algorithms and reasoning techniques to analyse the typical care pathways of specific groups of patients.

  To answer the broad range of pharmaco-epidemiological queries from national public health institutions, the PEPS [2] platform exploits, in secondary use, the French health cross-schemes

---

[2]PEPS: Pharmaco-Epidémiologie et Produits de Santé – Pharmacoepidemiology of health products

insurance system, called SNDS. The SNDS covers most of the French population with a sliding period of 3 past years. The main characteristics of this data warehouse are described in [39]. Contrary to local hospital EHR or even to other national initiatives, the SNDS data warehouse covers a huge population. It makes possible studies on unfrequent drugs or diseases in real conditions of use. To tackle the volume and the diversity of the SNDS data warehouse, a research program has been established to design an innovative toolbox. This research program is focused first on the modeling of care pathways from the SNDS database and, second, on the design of tools supporting the extraction of insights about massive and complex care pathways by clinicians. In such a database a care pathway is an individual sequence of drugs exposures, medical procedures and hospitalizations.

- **Care Sequences for the Exploration of Medico-administrative Data**. The difficulty of analyzing medico-administrative data is the semantic gap between the raw data (for example, database record about the delivery at date t of drug with ATC 2 code N 02BE01) and the nature of the events sought by clinicians ("was the patient exposed to a daily dose of paracetamol higher than 3g?"). The solution that is used by epidemiologists consists in enriching the data with new types of events that, on the one side, could be generated from raw data and on the other side, have a medical interpretation. Such new abstract events are defined by clinician using proxies. For example, drugs deliveries can be translated in periods of drug exposure (drug exposure is a time-dependent variable for non-random reasons) or identify patient stages of illness, etc. A proxy can be seen as an abstract description of a care sequence.

  Currently, the clinicians are limited in the expression of these proxies both by the coarse expressivity of their tools and by the need to process efficiently large amount of data. From a semantic point of view, care sequences must fully integrate the temporal and taxonomic dimensions of the data to provide significant expression power. From a computational point of view, the methods employed must make it possible to efficiently handle large amounts of data (several millions care pathways). The aim of the PhD of Johanne Bakalara is to study temporal models of sequences in order 1) to show their abilities to specify complex proxies representing care sequences needed in pharmaco-epidemiological studies and 2) to build an efficient querying tool able to exploit large amount of care pathways.

4.3 Agriculture and Environment

- **Dairy Farming**. The use and analysis of data acquired in dairy farming is a challenge both for data science and animal science. The goal is to improve farming conditions, i.e., health, welfare and environment, as well as farmers' income. Nowadays, animals are monitored by multiple sensors giving a wealth of heterogeneous data such as temperature, weight, or milk composition. Current techniques used by animal scientists focus mostly on mono-sensor approaches. The dynamic combination of several sensors could provide new services and information useful for dairy farming. The PhD thesis of Kevin Fauvel (#DigitAg grant), aims to study such combinations of sensors and to investigate the use data mining methods, especially pattern mining algorithms. The challenge is to design new algorithms that take into account data heterogeneity —in terms of nature and time units—, and that produce useful patterns for dairy farming. The outcome of this thesis will be an original and important contribution to the new challenge of the IoT (Internet of Things) and will interest domain actors to find new added value to a global data analysis. The PhD thesis, started on October 2017, takes place in an interdisciplinary setting bringing together computer scientists from INRIA and animal scientists from INRA, both located in Rennes.

- **Optimizing the Nutrition of Individual Sow**. Another direction for further research is the combination of data flows with prediction models in order to learn nutrition strategies. Raphaël Gauthier started a PhD thesis (#DigitAg Grant) in November 2017 with both INRIA and INRA supervisors. His research addresses the problem of finding the optimal diet to be supplied to individual sows. Given all the information available, e.g., time-series information about previous feeding, environmental data, scientists models, the research goal is to design new algorithms to determine the optimal ration for a given sow in a given day. Efficiency issues of developed algorithms will be considered

since the proposed software should work in real-time on the automated feeder. The decision support process should involve the stakeholder to ensure a good level of acceptance, confidence and understanding of the final tool.

- **Ecosystem Modeling and Management**. Ongoing research on ecosystem management includes modelling of ecosystems and anthroprogenic pressures, with a special concern on the representation of socio-economical factors that impact human decisions. A main research issue is how to represent these factors and how to integrate their impact on the ecosystem simulation model. This work is an ongoing cooperation with ecologists from the Marine Spatial Ecology of Queensland University, Australia and from Agrocampus Ouest.

4.4   Education

- **Data-oriented Academic Counseling.** Course selection and recommendation are important aspects of any academic counseling system. The Learning Analytics community has long supported these activities via automatic, data-based tools for recommendation and prediction. LACODAM, in collaboration with the Ecuadorian research center CTI [3] has contributed to this body of research with the design of a tool that allows students to select multiple courses and predict their academic performance based on historical academic data. The tool resorts to visualization and interpretable machine learning techniques, and is intended to be used by the students before the counseling sessions to plan their upcoming semester at the Ecuadorian university ESPOL. In our ongoing collaboration with CTI we have now studied the impact of academic predictions, explanations in the behavior and decision of the students. Moreover, we have studied whether those effects change when the predictions are conveyed using a different visual representation. The results of this study have been published in a paper that will appear in the proceeding of the CHI conference in 2021.

4.5   Others

- **RDF Archiving.** The dynamicity of the Semantic Web has motivated the development of solutions for RDF archiving, i.e., the task of storing and querying all previous versions of an RDF dataset. Notwithstanding the value of RDF archiving for data maintainers and consumers, this field of research remains under-developed for multiple reasons. These include notably (i) the lack of usability and scalability of the existing systems, (ii) no archiving support for multi-graph RDF datasets, (iii) the absence of a standard SPARQL extension for RDF archives, and (iv) a disregard of the evolution patterns of RDF datasets. The PhD thesis of Olivier Pelgrin aims at developing techniques towards a scalable and full-fledged archiving solution for the Semantic Web. This PhD thesis is a collaboration between LACODAM and the DAISY team at Aalborg University.

5   Social and environmental responsibility

5.1   Footprint of research activities

There are two main axes that characterize the bulk of LACODAM's environmental impact: work trips, and computing resources utilisation.

Work trips.   The sanitary crisis around the COVID-19 pandemic has strongly disturbed the team's activities. Since the beginning of the first lockdown (March 2020), almost all relevant scientific events (conferences, workshops, team seminars. etc.) have been held online. This has drastically reduced the environmental footprint of our scientific missions.

Utilisation of computing resources.   LACODAM contributed last year with a new server (abacus12) to the Igrida computing platform. Being a team specialized in data science and machine learning, a recurrent task in LACODAM is to run CPU intensive algorithms on large data collections, for example,

---

[3]Centro de Tecnologías de Información, http://cti.espol.edu.ec/

to train deep neural networks. Two of our ongoing PhD projects (the theses of Yichang Wang and Heng Zhang) concern deep learning technologies, and the increasing place of eXplanaible AI in our research program will boost our reliance on Igrida (notably with the PhD of Julien Delaunay). This will increase the energetic and environmental footprint of our activities in a non-negligible way. We are therefore willing to collaborate with the institute's direction in any initiative that could mitigate such an impact.

5.2 Impact of research results

We estimate that the research work can have actual impact in three different ways:

- In the short/medium term, a significant part of our research work is conduced in collaboration with companies, through CIFRE PhDs. Hence, the research problems addressed concern an important challenge for the company, and the solutions proposed are evaluated on their relevance to tackle this challenge. For example, the CIFRE PhD of Colin Leverger (defended this year), funded by Orange, was concerned with forecasting seasonal time series evolution, with application to the server load of the company. Its results could help better provision server resources during the year. Colin was hired by Orange after his PhD.

- In the medium/long term, we also have potential impactful research work with scientists from other domains, especially in environment and agriculture. Some earlier work of the team, conduced with INRAE SAS team, helped better understand nitrate pollution in Brittany, and important envirnomental issue. And recent work from the PhD of Kevin Fauvel, conduced with the INRAE Pegase team, allowed for the design of the current best method to detect oestrus in dairy cows from cheap sensors. This is important as reproduction of dairy cows is directly related to milk production, and also to cow culling in case of repetitive failures to inseminate at the right time.

- Last, in the longer term, the team has a fundamental line of work on machine learning and interpretability. This is a critical topic nowadays due to the emergence of the GDPR. Given the increasing use of machine learning solutions in most areas of human activity, work on interpretability is of utmost societal importance, as it will help in designing more useful and also more acceptable machine learning approaches. This will require a sustained effort from the community: LACODAM is taking part in this effort, both on its own, as the coordinator of the Inria HyAIAI project, and last by having several of its members in the large European Project TAILOR dedicated to this topic.

6 Highlights of the year

6.1 Awards

The paper of Kevin Fauvel *A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning* got an "Oustanding paper award" in the special session "AI for social good" of the prestigious AAAI 2020 conference. We are especially happy of this work as it is a collaboration with the KerData team and the Rutgers University (US).

7 New software and platforms

7.1 New software

7.1.1 REMI

Name: Mining Intuitive Referring Expressions in Knowledge Bases

Keywords: RDF, Knowledge database, Referring expression

Functional Description: REMI takes an RDF knowledge base stored as an HDT file, and a set of target entities and returns a referring expression that is intuitive, i.e., the user is likely to understand it.

URL: http://gitlab.inria.fr/lgalarra/remi

Contact: Luis Galarraga Del Prado

7.1.2 PyChronicle

Keywords: Sequence, Sequential patterns, Pattern matching

Functional Description: Python library containing classes for representing sequences and chronicles, ie a representation of a temporal pattern. It implements efficient recognition algorithms to match chronicles in a long sequence.

URL: https://gitlab.inria.fr/tguyet/pychronicles

Publication: hal-02422796

Contact: Thomas Guyet

Participant: Thomas Guyet

7.1.3 AMIE

Name: Association Rule Mining Under Incomplete Evidence

Keywords: Pattern discovery, Knowledge database

Functional Description: AMIE takes as input a file that contains a knowledge base. This file must have one of the following formats:

subject DELIM predicate DELIM object [whitespace/tabulation .] NEWLINE factid DELIM subject DELIM predicate DELIM object [whitespace/tabulation .] NEWLINE

The default delimiter DELIM is the tabulation (.tsv files) but can be changed using the -d option. Any trailing whitespaces followed by a point are ignored. This allows parsing most NT files using the option: -d" ".

However make sure the factid, subject, predicate nor the object contains the delimiter used (particularly in literal facts files). Otherwise the parsing may fail or facts may be wrongfully recognized as the second format.

In the near future, AMIE will be able to parse the W3C Turtle format as well.

Authors: Luis Galarraga Del Prado, Lajus Jonathan, Fabian Suchanek

Contact: Luis Galarraga Del Prado

Partner: Telecom Paris

8   New results

We present the research contributions and our scientific achievements along the axes detailed in Section 3.

8.1   Novel Pattern Mining and Machine Learning Algorithms

We frame our research achievements primarily with respect to fundamental research problems and questions they tackle and then make the link with the practical applications they do or may serve.

Efficient chronicle mining for classification.   A *chronicle* is an episode that consists of a set of events and temporal constraints between the occurrences of those events. Chronicle mining is the task of finding recurrent chronicles in data, and has found plenty of applications in pharmaco-epidemiology, e.g., seizure prediction in patients, patient profiling, etc. LACODAM has a long experience in chronicle mining applied to such contexts and our contributions comprise (a) novel heuristic-based techniques to efficiently enumerate chronicles in a database of sequences [15], and (b) methods to use those chronicles for classification of temporal sequences [6].

Negative sequential patterns.   A negative sequential pattern can express the absence of an event or set of events within a sequence. Mining such patterns is extremely challenging. There exist several semantics for what an absent event actually means [11], and even when we stick to one semantic interpretation, we are confronted with a very large search space. NegPSpan [3] offers a solution to this problem. NegPSpan is an algorithm that can efficient mine negative sequential patterns with embedding constraints. Moreover, the authors show the applicability of such negative sequential patterns in care pathway analysis.

MDL for signature discovery   The principle of Mininum Description Length is a paradigm that resorts to the analogy of compression to tackle the problem of *pattern explosion* in pattern mining. This is achieved by selecting a subset of the patterns of minimal size such that they "compress" the data at best, i.e., they minimize the description length of the data when used for description. The work of Gautrais et al. [10] applies the MDL principle to the problem of mining signature patterns in retail databases. A signature is a set of items co-occurring recurrently in an event sequence with unbounded or loosely defined periods of time, e.g., a set of products bought by a customer during a season. To find such signatures, the data analyst must define in advance the maximal size of the time segments fixed to search for the signatures. The work presented in [10] applies the principle of MDL to find that size automatically and return a compact set of meaningful signatures that can be used to characterize the purchase profiles of retail customers.

Knowledge Bases.   A knowledge base (KB) is a collection of facts or binary statements of the form $p(s, o)$ such as *capital*(*France*, *Paris*). Knowledge bases find applications in a plethora of AI related tasks, e.g., smart assistants, question answering, information retrieval, etc. Given the large amounts of information contained in today's KBs, applications such as smart assistants, or automatic query generation can benefit from mining referring expressions from KBs. A referring expression (RE) is a statement that describes an entity or set of entities uniquely, e.g., " is the capital of France" is an RE for Paris. The work of Galárraga et. al [9] proposes REMI, a method to mine referring expressions from KBs. However, since entities can be uniquely described in several ways (Paris is also the resting place of Victor Hugo), REMI aims at intuitive REs, that is, descriptions that resort to prominent entities. That way we can make sure that the users will relate to the descriptions. The pertinence of REMI's descriptions was evaluated via user studies. Moreover the system outperformed a solution based on rule mining.

The experimental evaluation conducted for REMI [9] showed, among other things, that existing rule mining systems still do not scale to the most challenging settings. On those grounds, Lajus et. al [12], presents AMIE 3, the new generation of the AMIE rule mining system that presents a series of (query) optimizations tailored at achieving exhaustive and exact rule mining on large knowledge bases. Unlike its predecessors [35, 36], AMIE 3 can mine rules in the Wikidata KB in less than one day, becoming the new state of the art in rule miners on KBs.

8.2   User/System interaction

We report three major contributions to the axis of user-system interaction:

- An approach based on model checking that can detect faulty patterns in systems modeled by Petri nets. The proposed approach [4] can tell users if a scenario containing an event sequence of interest can lead to a fault in a systes. The approach has been evaluated and validated in the context of product transportation systems.

- A declarative approach for mining negative sequential patterns [5]. Continuing our research line on negative sequential patterns, we propose to encode their different semantics [11] using a formal language, namely Answer Set Programming [38]. These encodings enable to compare the outputs and computing performances of the different semantics for negative sequential patterns.

- A data-oriented counseling system for term selection that allows students to select multiple courses and predict their academic performance based on historical academic data [32]. The tool is intended to be used prior to counseling sessions in which students plan their upcoming semester at the Ecuadorian university ESPOL. We have conducted a series of user studies that shed light on the impact of the system's advise on the behavior and decisions of the students.

8.3   Decision Support

Differently from the previous section we describe the axis of decision support in terms of the application domains of our research contributions.

Earthquake detection.   Fauvel et al. [8] proposes a novel approach that uses information from multiple sensors, i.e., high-precision GPS stations and seismometers, to predict earthquake severity with higher accuracy than the state of the art. The method, called DMSEEW[4], combines data from multiple sensors and plugs it into a stacking ensemble method, which has been evaluated on a real-world dataset and validated by geoscientists. The system builds on a geographically distributed infrastructure, ensuring an efficient computation in terms of response time and robustness to partial infrastructure failures.

Computer Vision.   In the domain of object detection, Zhang et al. [24] questions the well-established use of IoU[5] predefined anchor boxes to evaluate matching quality, and proposes a new anchor matching criterion. This novel criterion guides the training phase in object detectors by optimizing mutually both the localization and the classification tasks: the predictions related to one task are used to dynamically assign sample anchors and improve the model on the other task, and vice versa. Experiments on different state-of-the-art deep learning architectures on PASCAL VOC and MS COCO datasets demonstrate the effectiveness and generality of this mutual guidance strategy.

Again regarding the task of object detection in a multispectral setting (e.g., pairs of visible and infrared spectral band images), Zhang et al. [25, 26] build upon the evidence of the advantage of mid-term intermediate information fusion over early or late fusion, and propose the methods PS-Fuse and Cyclic Fuse-and-Refine. The experimental evaluation shows these two contributions lead to significant improvements in precision compared to existing methods in object detection on multispectral datasets.

Online handwriting recognition is challenging but an already well-studied topic. However, recent advances in the development of convolutional neural networks (CNN) make us believe that these networks could still improve the state of the art especially in the much more challenging context of online children handwritten letters recognition. This is because, children handwriting is, at an early stage of learning, approximate and includes deformed letters. To evaluate the potential of these networks, Corbille et al. [18] study the early and late fusions of different input channels that can provide a CNN with information about the handwriting dynamics in addition to the static image of the characters. The experiments on a real children handwriting dataset with 27 000 characters acquired in primary schools, show that using multiple channels with CNN, improves the accuracy performance of different CNN architectures and different fusion settings for character recognition.

Recommendation systems.   Gauthier et al. [20] present PB-MHB, a new algorithm for product recommendations to users in websites. PB-MHB makes multiple recommendations following a position-based model. The algorithm is based on the principle of one-armed bandits and uses Thompson sampling coupled with a Metropolis-Hastings algorithm to derive the parameters of the probabilistic laws used by the recommendation engine. This has never been done in the context of a position-based model. Unlike similar methods, PB-MHB does not require the user view probabilities on each position of the web page as input. This gives PB-MHB a practical advantage over existing online recommendation methodas as these probabilities are difficult to obtain a priori. Experiments made on simulated data and on data from real databases (KDD-CUP2012 and Yandex) show that PB-MHB yields better recommendations than the state of the art.

Product transportation systems.   The work of Bakalara et al. [4] presents two methods based on model checking for diagnosing fault patterns in discrete event systems modeled as Petri nets. The proposed approaches are evaluated in the context of product transportation systems, and can provide a proof of the existence of faults for runs of the system characterized by a given sequence of events.

---

[4]Distributed Multi-Sensor Earthquake Early Warning
[5]Intersection over Union

Intrusion detection.   Siffer et al. [13] extended their earlier work on anomaly detection with Extreme Value Theory, and integrated them in a simple yet efficient intrusion detection system callet Netspot. Unlike all the previous works, Netspot is not an end-to-end solution aimed to detect all cyber-attacks with packet resolution. It is rather a module providing a behavioral information which can be integrated in a more general monitoring system. Netspot is simple: it has few (simple) parameters, it adapts along time to the monitored network and it is as fast as current rulebased methods. But most importantly, it is able to detect realworld cyber-attacks, making it a credible practical anomaly based NIDS (Network Intrusion Detection System).

8.4   Interpretability

LACODAM's achievements in this domain include the proposal of an explainable by design classifier for time series, a study of techniques to improve the fidelity of a post-hoc explainable method, and an approach to benchmark machine learning methods taking into account the dimension of interpretability.

Interpretable time series classification   The work of Wang et al [14] proposes to embed interpretability constraints in a shapelet-based time series classifier. Such constraints guarantee that the learned shapelets (the features used for the classification) ressemble as much as possible some actual regions of the learned time series. That way users of the classifier are provided hints of the shapes and time segments that may have played an important role in the classification.

Improving rule-based explanations   An *anchor* is a rule that can explain the behavior of a classifier for a given class with respect to a target instance. An example is the rule that says "when the salary is greater than 50K USD and the client is married the system grants the loan" (with probability greater than 95%) in reference to a classifier that processes loan requests. To mine those anchors, existing methods must first discretize the existing numerical variables (since the explanations lie on a symbolic space) and generate examples around the instance for which we want to generate an explanation. Those "neighbor" instances are generating by perturbing the target instance. That perturbation depends on the type of data. The work of Delaunay et al. [7] studies the impact of discretization and different perturbation strategies in the fidelity of the resulting anchor-based explanations for text classifiers. It shows, among other things, that entropy-based discretizations can greatly improved the fidelity of the resulting explanations.

Interpretability-aware benchmarking   The work of Fauvel et al [19] proposes a new performance-explainability analytical framework to assess and benchmark machine learning methods. The framework details a set of characteristics that systematize the performance-explainability assessment of existing machine learning methods. [19] illustrates the applicability of the benchmark to the current state-of-the-art multivariate time series classifiers.

9   Bilateral contracts and grants with industry

9.1   Bilateral contracts with industry

- **ENEDIS - Univ Rennes 1**
  Participants: E. Fromont
  Contract amount: 9k€

  This is a short pre-PhD contract with Enedis, that allowed to prepare the CIFRE PhD of Anthonin Voyez.

- **Hyptser: Hybrid Prediction of Time Series**
  Participants: T. Guyet, S. Malinowski (LinkMedia), V. Lemaire (Orange)
  Contract amount: 25k€

HYPTSER is a collaborative project between Orange Labs and LACODAM funded by the Fondation Mathématique Jacques Hadamard (PGMO program). It aims at developing new hybrid time series prediction methods in order to improve capacity planning for server farms. Capacity planning is the process of determining the infrastructure needed to meet future customer demands for online services. A well-made capacity planning helps to reduce operational costs, and improves the quality of the provided services. Capacity planning requires accurate forecasts of the differences between the customer demands and the infrastructure theoretical capabilities. The HYPTSER project makes the assumption that this information is captured by key performance indicators (KPI), that are measured continuously in the service infrastructure. Thus, we expect to improve capacity planning capabilities by making accurate forecasts of KPI time series. Recent methods about time series forecasting make use of ensemble models. In this project, we are interested in developing hybrid models for time series forecasting. In the next steps of this project, we will analyze the performance of this two strategies on KPI time series provided by Orange and compare them to classical ensemble methods.

Additional remarks. This project has financed the PhD of Colin Leverger who defended his thesis in November 2020.

- **AdvisorSLA 2018 - Inria**
  Participants: E. Bourrand, L. Galárraga, E. Fromont, A. Termier
  Contract amount: 25k€

  Context. AdvisorSLA is a French company headquartered in Cesson-Sévigné, a city located in the outskirts of Rennes in Brittany. The company is specialized in software solutions for network monitoring. For this purpose, the company relies on techniques of network metrology. AdvisorSLA's customers are carriers and telecommunications/data service providers that require to monitor the performance of their communication infrastructure as well as their QoE (quality of service). Network monitoring is of tremendous value for service providers because it is their primary tool for proper network maintenance. By continuously measuring the state of the network, monitoring solutions detect events (e.g., an overloaded router) that may degrade the network's operation and the quality of the services running on top of it (e.g., video transmission could become choppy). When a monitoring solution detects a potentially problematic sequence of events, it triggers an alarm so that the network manager can take actions. Those actions can be preventive or corrective. Some statistics gathered by the company show that only 40% of the triggered alarms are conclusive, that is, they manage to signal a well-understood problem that requires an action from the network manager. This means that the remaining 60% are presumably false alarms. While false alarms do not hinder network operation, they do incur an important cost in terms of human resources.

  Objective. We propose to characterize conclusive and false alarms. This will be achieved by designing automatic methods to "learn" the conditions that most likely precede the fire of each type of alarm, and therefore predict whether the alarm will be conclusive or not. This can help adjust existing monitoring solutions in order to improve their accuracy. Besides, it can help network managers automatically trace the causes of a problem in the network. The aforementioned problem has an inherent temporal nature: we need to learn which events occur before an alarm and in which order. Moreover, metrology models take into account the measurements of different components and variables of the network such as latency and packet loss. For these two reasons, we resort to the field of multivariate time sequences and time series. The fact that we know the "symptoms" of an alarm and whether it is conclusive or not, allows for the application of supervised machine learning and pattern mining methods.
  Additional remarks. This contract financed the PhD thesis of Erwan Bourrand that was interrupted in September 2020.

- **ATERMES 2018-2021 - Univ Rennes 1**
  Participants: H. Zhang, E. Fromont

Contract amount: 45k€
Context. ATERMES is an international mid-sized company, based in Montigny-le-Bretonneux with a strong expertise in high technology and system integration from the upstream design to the long-life maintenance cycle. It has recently developed a new product, called BARIERTM ("Beacon Autonomous Reconnaissance Identification and Evaluation Response"), which provides operational and tactical solutions for mastering borders and areas. Once in place, the system allows for a continuous night and day surveillance mission with a small crew in the most unexpected rugged terrain. BARIER™ is expected to find ready application for temporary strategic site protection or ill-defined border regions in mountainous or remote terrain where fixed surveillance modes are impracticable or overly expensive to deploy.

Objective. The project aims at providing a deep learning architecture and algorithms able to detect anomalies (mainly the presence of people or animals) from multimodal data. The data are considered "multimodal" because information about the same phenomenon can be acquired from different types of detectors, at different conditions, in multiple experiments, etc. Among possible sources of data available, ATERMES provides Doppler Radar, active-pixel sensor data (CMOS), different kind of infra-red data, the border context etc. The problem can be either supervised (if label of objects to detect are provided) or unsupervised (if only times series coming from the different sensors are available). Both the multimodal aspect and the anomaly detection one are difficult but interesting topics for which there exist few available works (that take both into account) in deep learning.

- **PSA - Inria**
  Participants: E. Fromont, A. Termier, L. Rozé, G. Martin
  Contract amount: 75k€
  Context. Peugeot-Citroën (PSA) group aims at improving the management of its car sharing service. To optimize its fleet and the availability of the cars throughout the city, PSA needs to analyze the trajectory of its cars.

  Objective. The aim of the internship is (1) to survey the existing methods to tackle the aforementioned need faced by PSA and (2) to also investigate how the techniques developed in LACODAM (e.g., emerging pattern mining) could be serve this purpose. A framework, consisting of three main modules, has been developped. We describe the modules in the following.

  - A town modelisation module with clustering. Similar towns are clustered in order to reuse information from one town in other towns.

  - A travel prediction module with basic statistics.

  - A reallocation strategy module (choices on how to relocate cars so that the most requested areas are always served). The aim of this module is to be able to test different strategies.

  Additional remarks. This is the doctoral contract for the PhD of Gregory Martin (Thèse CIFRE).

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Visits of international scientists

- **Benoît Frénay**, associate professor at the Faculty of Computer Science of the University of Namur (11 to 13/02/2020).

10.2   European initiatives

10.2.1   FP7 & H2020 Projects

- **TAILOR**
  Participants: E. Fromont, L. Galárraga, A. Termier

  TAILOR (https://tailor-network.eu/) is one of the four H2020 ICT48 networks of AI excellence centers. The main focus of this network is the advancement of research on "Trustworthy AI". Lacodam is strongly involved in the WP3 of this project (one of the core scientific WP, titled "Trustworthy AI"), with E. Fromont leading task 3.8 "Fostering the AI Scientific community around Trustworthy AI" and L. Galárraga leading task 3.9 "Synergies with Industry, Roadmap and Challenges".

10.3   National initiatives

- **HyAIAI: Hybrid Approaches for Interpretable AI**
  Participants: E. Fromont (leader), A. Termier, L. Galárraga

  The Inria Project Lab HyAIAI is a consortium of Inria teams (Sequel, Magnet, Tau, Orpailleur, Multispeech, and LACODAM) that work together towards the development of novel methods for machine learning, that combine numerical and symbolic approaches. The goal is to develop new machine learning algorithms such that (i) they are as efficient as current best approaches, (ii) they can be guided by means of human-understandable constraints, and (iii) their decisions can be better understood.

- **#DigitAg: Digital Agriculture**
  Participants: A. Termier, V. Masson, C. Largouët, A.I. Graux

  #DigitAg is a "Convergence Institute" dedicated to the increasing importance of digital techniques in agriculture. Its goal is twofold: First, making innovative research on the use of digital techniques in agriculture in order to improve competitiveness, preserving the environment, and offer correct living conditions to farmers. Second, preparing future farmers and agricultural policy makers to successfully exploit such technologies. While #DigitAg is based on Montpellier, Rennes is a satellite of the institute focused on cattle farming.

  LACODAM is involved in the "data mining" challenge of the institute, which A. Termier co-leads. He is also the representative of Inria in the steering comittee of the institute. The interest for the team is to design novel methods to analyze and represent agricultural data, which are challenging because they are both heterogeneous and multi-scale (both spatial and temporal).

- **DRIAS: Accountability of Artificial Intelligence in Health**
  Participants: T. Guyet, T. Allard (DRUID)

  DRIAS is an inter-disciplinary project funded by MITI-CNRS. It gathers researches in computer science, public law, and health informatics to investigate the current limits of the legal frameworks (mainly French and European laws) facing the changes induced by the use of artificial intelligence in healthcare systems. This project is mainly focused on the notion of accountability, that is a motivating concept for addressing issues related to machine learning interpretability.

10.3.1   ANR

- **FAbLe: Framework for Automatic Interpretability in Machine Learning**
  Participants: L. Galárraga (holder), C. Largouët

  How can we fully automatically choose the best explanation for a given use case in classification?.

Answering this question is the raison d'être of the JCJC ANR project FAbLe. By "best explanation" we mean the explanation that yields the best trade-off between interpretability and fidelity among 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, interpretability is a subjective concept that has not been formalized yet. Hence, in order to answer our prime question we first need to answer the question: "How can we formalize and quantify interpretability across models?". Much like research in automatic machine learning has delegated the task of accurate model selection to computers [34], 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., Scikit-learn).

11 Dissemination

11.1 Promoting scientific activities

- Board member of the French Association of Artificial Intelligence (AFIA) (T. Guyet)

11.1.1 Scientific events: organisation

General chair, scientific chair

- Co-chairs of the third edition of the AIMLAI workshop (Advances in Interpretable Machine Learning and Artificial Intelligence)[6] colocated with the conference CIKM 2020[7] (T. Bouadi, L. Galárraga)

- Co-chair of the Fifth AALTD workshop [8] colocated with the conference ECML 2020 (T. Guyet)

- Co-chair of the second AGEE (AGriculture, Environnent, Ecologie) workshop of GdR MADICS (T. Bouadi)

11.1.2 Scientific events: selection

Member of the conference program committees

- ISWC 2020 (L. Galárraga)

- IJCAI–PRICAI 2020 (L. Galárraga, T. Guyet, T. Bouadi)

- KDD 2020 (A. Termier)

- CIKM 2020 (T. Guyet, T. Bouadi)

- ECML/PKDD 2020 (T. Guyet, T. Bouadi, A. Termier)

- ECAI 2020 (T. Guyet, T. Bouadi)

- NeurIPS 2020 (R. Gaudel)

- SAC 2020 (T. Guyet)

- CAp 2020 (R. Gaudel)

- XKDD 2020 workshop (L. Galárraga, R. Gaudel)

- AIMLAI 2020 workshop (L. Rozé, R. Gaudel)

---

[6] https://project.inria.fr/aimlai/
[7] http://cikm2020.org/
[8] https://project.inria.fr/aaltd20/

- EGC 2020 (T. Bouadi)

- EDA 2020 (T. Bouadi)

- APIA 2020 (C. Largouët)

Reviewer

- Poster Session at ESWC 2020 (L. Galárraga)

11.1.3 Journal

Member of the editorial boards

- Data Mining Journal (E. Fromont)

- Revue Ouverte d'Intelligence Artificielle (T. Guyet)

Reviewer - reviewing activities

- Journal on Very Large Databases, VLDBJ (L. Galárraga)

- Pattern Recognition (R. Gaudel)

- Data Mining Journal (A. Termier)

- Agronomy for Sustainable Development (A. Termier)

11.1.4 Invited talks

- Fête de la science, Bruz, conference on "Mythes et réalités de l'intelligence artificielle"

- Invited speaker about "Explainable Time Series Classification with Regularized Shapelets" at the Spring Workshop on Mining and Learning, Titisee, Germany.

- Rencontre des acteurs de médiation de la région Auvergne Rhône Alpes (organisé par fondation Blaise Pascal), Lyon, talk on "L Codent L Créent"

11.1.5 Leadership within the scientific community

- E. Fromont is an IUF member (Junior) from 2019 to 2024

- Member of the steering committee of the conference ECML/PKDD (the European conference on Machine Learning and Data Mining) (E.Fromont)

- Member of the steering committee of the TIME Symposium [9] (T. Guyet)

11.1.6 Scientific expertise

- Reviewer of the call for proposals *FAIR Data and Models for AI and ML* of the *Advanced Scientific Computing Research Program* financed by the U.S. Department of Energy (L. Galárraga)

- ANR project Review (E. Fromont)

- ICT-AGRI-FOOD (ERA-NET Cofund) Projet Review (C. Largouët)

- Inria Rennes MI Jury (E. Fromont)

- Inria Rennes CR Jury (C. Largouët)

- ENSTBB MCF Jury (C. Largouët)

- Member of the working group "software" for the Health Data Hub (T. Guyet)

---

[9] http://time.di.unimi.it/TIME_Home.html

11.1.7 Research administration

- IPL HyAIAI: E. Fromont (director), A. Termier (member), L. Galárraga (member)

- IRISA/cross-cutting theme on Health and Biology[10]: T. Guyet (leader)

- Member of the research group of the Environmental Intelligence project (led by OSUR) (T. Bouadi)

- Elected member (substitute, college B) of the research commission- IUT Lannion (T. Bouadi)

- Member of the scientific committee of INRAE AgroEnv department (A. Termier)

11.2 Teaching - Supervision - Juries

Teaching Some members of the project-team LACODAM are also faculty members and are actively involved in computer science teaching programs in ISTIC, INSA and Agrocampus-Ouest. Besides these usual teachings LACODAM is involved in the following programs:

- Master 2 IL, CCN: Option Machine Learning, Istic, University of Rennes 1, 32h (E. Fromont)

- Master 2 DMV Module: Data Mining and Visualization, 13h, M2, Istic, University of Rennes 1 (A. Termier)

- Master 2 Big Data, 30h, Master Datascience, University of Rennes 2 & Agrocampus Ouest (C. Largouët)

- Master 2 DataViz with R, 10h, Master Datascience, University of Rennes 2 & Agrocampus Ouest (L. Bonneau de Beaufort)

- Master 1 SIF: Option IA, Istic, 20h, University of Rennes 1 (A. Termier, E. Fromont, T. Bouadi)

- Master 1 Scientific Programming, 25h, Agrocampus Ouest (C. Largouët)

- Master 1 Data Management, 25h, Agrocampus Ouest (C. Largouët)

- Master : R. Gaudel, Apprentissage profond, 9h eq. TD, M2, ENSAI, France

- Master : R. Gaudel, Systèmes de recommandation, 9h eq. TD, M2, ENSAI, France

- Master : R. Gaudel, Bandits Theory, 9h eq. TD, M2, ENSAI, France

- Master : R. Gaudel, INFO2, 30h eq. TD, M1, ENS, France

- Master : T. Guyet, Spatial data science and GIS programming, Institut Agro/Rennes University, France

- V. Masson is the head of the L3 of Computer Science at Universiy of Rennes 1 (112 students)

Supervision PhD Defenses

- Colin Leverger, "Management and forecasting of measures for server optimization in a cloud context", 20/11/2020, supervised by A. Termier, R. Marguerie, S. Malinowski, T. Guyet, L. Rozé

- Maël Gueguen, "Improving the Performance and Energy Efficiency of Complex Heterogeneous Manycore Architectures with On-chip Data Mining", 23/10/2020, supervised by O. Sentieys, A. Termier

- Kévin Fauvel, "Enhancing Performance and Explainability of Multivariate Time Series Machine Learning Methods: Applications for Social Impact in Dairy Resource Monitoring and Earthquake Early Warning", 13/10/2020, supervised by V. Masson, A. Termier, P. Faverdin

---

[10]http://www.irisa.fr/en/health-and-biology

- Yiru Zhang, "Modeling and management of imperfect preferences with the theory of belief functions", 03/02/2020, supervised by T. Bouadi, A. Martin

PhD in Progress

- Lenaïg Cornanguer, "Apprentissage automatique d'automates temporisés à partir de séries temporelles" (*Learning timed automata from time series*), since 01/09/2020, supervised by C. Largouët, A. Termier and L. Rozé

- Josie Signe, "Animal welfare: characterizing the diversity between and within livestock farming situations with data mining methods used on information from dairy herd sensors", since 01/09/2020, supervised by A. Termier, P. Cellier, C. Largouët and V. Masson

- Julien Delaunay, "Automatic Construction of Explanations for AI Models", since 01/09/2020, supervised by C. Largouët and L. Galárraga

- Simon Corbillé, "Explainable deep learning-based methods for children handwriting analysis in education", since 1/10/2019, supervised by E. Fromont (25%), E. Anquetil.

- Antonin Voyez, "Privacy-Preserving Power Consumption Time-Series Publishing", since 1/06/2020, supervised by E. Fromont, G. Avoine, T. Allard.

- Abderaouf Nassim Amalou, "Machine learning for timing estimation", since 1/10/2020, supervised by E. Fromont (25%), I. Puaut.

- Victor Guyomard, "Explanation of decisions taken by machine learning algorithms", since 02/11/2020, supervised by T. Bouadi, F. Fessant, T. Guyet

- Raphaël Gauthier, "Modelling of Nutrient Utilization and Precision Feeding of Lactating Sows", since 01/11/2017, supervised by C. Largouët, J.-Y. Dourmad

- Heng Zhang, "Deep Learning on Multimodal Data for the Supervision of Sensitive Sites", since 03/12/2018, supervised by E. Fromont, S. Lefevre

- Yichang Wang, "Interpretable Shapelet for Anomaly Detection in Time Series", since 15/04/2018, supervised by E. Fromont, S. Malinowski, R. Tavenard, R. Emonet

- Camille-Sauvanneary Gauthier, "Session-aware Recommendation System", since 15/03/2019, supervised by É. Fromont, R. Gaudel, B. Guilbot

- Gregory Martin, "Data mining to optimize a free-floating car sharing service", since 8/04/2019, supervised by E. Fromont, L. Rozé, A. Termier.

- Johanne Bakalara, "Temporal models of care sequences for the exploration of medico-administrative data", since 1/10/2018, supervised by E. Oger, O. Dameron, T. Guyet, A. Happe

- Nicolas Sourbier, "Log Analysis Using Reinforcement Learning for Intrusion Detection", since 1/10/2019, supervised by M. Pelcat, O. Gesny, F. Majorczyk, T. Ollivier, T. Guyet.

- Maëva Durand, "Real-time welfare and health assessment of gestating sows using heterogeneous data", since 1/09/2020 supervised by C. Largouët, J.-Y. Dourmad (INRAE) and C. Gaillard (INRAE).

Juries HDR Juries

- Thomas Guyet, "Enhance sequential pattern mining with time and reasoning", 10/02/2020, Université de Rennes (A. Termier, examiner)

- Peggy Cellier, "Towards Usable Pattern Mining", 30/10/2020, Université de Rennes 1 (A. Termier, examiner)

PhD Defenses

- Adrien Bibal, "Interpretability and Explainability in Machine Learning with Application to Nonlinear Dimensionality Reduction", 16/11/2020, Namur University (L. Galárraga, examiner)

- Mélanie Münch, "Improving Uncertain Reasoning Combining Probabilistic Relational Models and Ex- pert Knowledge", 17/11/2020, AgroParisTech/PSL (T. Guyet, examiner)

- Alexandre Sahuguède, "Un algorithme de découverte de chroniques pertinentes pour le diagnostic par identification et reconstitution", 12/03/2020, Université Toulouse 3 Paul Sabatier (T. Guyet, examiner)

- Qiushi Cao, "Semantic Technologies for the Modelling of Preventive Maintenance for a SME Network in the Framework of Industry 4.0", 26/06/2020, INSA Rouen (T. Guyet, reviewer)

- Matthieu Jedor, "Bandit Algorithms for Recommender System Optimization", 18/12/2020, Univ. Paris-Saclay (R. Gaudel, examiner)

- Mohammed Ali Hammal, "Contribution à la découverte de sous-groupes corrélés : Application à l'analyse des systèmes territoriaux et des réseaux alimentaires", 05/06/2020, INSA Lyon (A. Termier, reviewer)

- Xin Du, "The Uncertainty in Exceptional Model Mining", 28/09/2029, TU Eindhoven - Netherlands (A. Termier, examiner)

- Romain Mathonat, "Rule Discovery in Labeled Sequential Data : Application to Game Analytics', 29/09/2020, INSA Lyon (A. Termier, reviewer)

- Nicolas Wagner, "Détection des modifications de l'organisation circadienne des activités des animaux en relation avec des états pré-pathologiques, un stress ou un évènement de reproduction", 27/10/2020, Université Clermont-Auvergne (A. Termier, reviewer)

- Joris Dugueperoux, "Protection des travailleurs dans les plateformes de crowdsourcing: une perspective technique", 1/10/2020, Université de Rennes 1 (A. Termier, examiner)

- Marine Louarn, "Analysis and Integration of Heterogeneous Large-scale Genomics Data – Application to B-cell Differentiation and Follicular Lymphoma Non-coding Mutations", 26/11/2020, Université de Rennes 1 (A. Termier, examiner)

11.3   Popularization

11.3.1   Internal or external Inria responsibilities

- Participation in the network "Data Science et Modélisation pour l'Agriculture et Agroalimentaire"[11] leaded by the ARVALIS institute and aimed to organize scientifique events (workshops, tutorials) around the topic of Digital Agriculture (L. Galárraga)

11.3.2   Education

- A. Termier and E. Fromont are together responsible of the second session of the DIU EIL ("Enseigner l'Informatique au Lycée"), whose goal is to prepare high school teachers to teach Computer Science. This year there are 24 students following this program.

11.3.3   Interventions

- A. Termier gave 4h of lectures at Emile Zola High School to college math teachers, to present broadly the topics of AI and Big Data.

---

[11]http://www.modelia.org/moodle/mod/forum/discuss.php?d=70

12 Scientific production

12.1 Major publications

[1] K. Fauvel, V. Masson, E. Fromont, P. Faverdin and A. Termier. 'Towards Sustainable Dairy Management - A Machine Learning Enhanced Method for Estrus Detection'. In: *KDD 2019 - ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 25th SIGKDD Conference on Knowledge Discovery and Data Mining proceedings. Anchorage, United States, Aug. 2019, pp. 1–9. DOI: 10.1145/3292500.3330712. URL: https://hal.archives-ouvertes.fr/hal-0219079 0.

[2] E. Galbrun, P. Cellier, N. Tatti, A. Termier and B. Crémilleux. 'Mining Periodic Patterns with a MDL Criterion'. In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. Dublin, Ireland, 2018. URL: https://hal.archives-ouve rtes.fr/hal-01951722.

12.2 Publications of the year

International journals

[3] T. Guyet and R. Quiniou. 'NegPSpan: efficient extraction of negative sequential patterns with embedding constraints'. In: *Data Mining and Knowledge Discovery* (2020). URL: https://hal.inr ia.fr/hal-03025572.

International peer-reviewed conferences

[4] J. Bakalara, Y. Pencolé and A. Subias. 'How to use Model Checking for Diagnosing Fault Patterns in Petri nets'. In: WODES 2020 - 15th IFAC Workshop on Discrete Event Systems. IFAC-PapersOnLine. Virtual, Brazil, 2020, pp. 1–6. URL: https://hal.archives-ouvertes.fr/hal-02997908.

[5] P. Besnard and T. Guyet. 'Declarative mining of negative sequential patterns'. In: *Procedings of DPSW@ECAI*. 1st Declarative Problem Solving Workshop - DPSW 2020. Santiago de Compostela, Spain: https://sites.google.com/view/dpsw2020/, 29th Aug. 2020. URL: https://hal.inr ia.fr/hal-03025560.

[6] Y. Dauxais and T. Guyet. 'Generalized chronicles for temporal sequence classification'. In: Advanced Analytics and Learning on Temporal Data. Ghent, Belgium, Sept. 2020. URL: https://hal.inria .fr/hal-03025590.

[7] J. Delaunay, L. Galárraga and C. Largouët. 'Improving Anchor-based Explanations'. In: CIKM 2020 - 29th ACM International Conference on Information and Knowledge Management. Galway / Virtual, Ireland: https://www.cikm2020.org/, 19th Oct. 2020, pp. 3269–3272. DOI: 10.1145/3340531 .3417461. URL: https://hal.inria.fr/hal-03133223.

[8] K. Fauvel, D. Balouek-Thomert, D. Melgar, P. Silva, A. Simonet, G. Antoniu, A. Costan, V. Masson, M. Parashar, I. Rodero and A. Termier. 'A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning'. In: In Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, United States, 7th Feb. 2020, pp. 403–411. DOI: 10.1609/aaai.v34i01.5376. URL: https://hal.archives-ouvertes.fr/hal-02373429.

[9] L. Galárraga, J. Delaunay and J.-L. Dessalles. 'REMI: Mining Intuitive Referring Expressions on Knowledge Bases'. In: EDBT 2020 - 23rd International Conference on Extending Database Technology. Virtual Event, Denmark, 30th Mar. 2020, pp. 387–390. DOI: 10.5441/002/edbt.2020.39. URL: https://hal.inria.fr/hal-03084627.

[10] C. Gautrais, P. Cellier, M. Van Leeuwen and A. Termier. 'Widening for MDL-based Retail Signature Discovery'. In: IDA 2020 - Symposium on Intelligent Data Analysis. Konstanz, Germany, 27th Apr. 2020, pp. 1–12. URL: https://hal.inria.fr/hal-02505913.

[11] T. Guyet and P. Besnard. 'Semantics of negative sequential patterns'. In: ECAI 2020 - 24 th European Conference on Artificial Intelligence. Santiago de Compostela, Spain, 29th Aug. 2020, pp. 1–11. URL: https://hal.inria.fr/hal-02481240.

[12] J. Lajus, L. Galárraga and F. Suchanek. 'Fast and Exact Rule Mining with AMIE 3'. In: ESWC 2020: The Semantic Web. Virtual Event, Greece, 27th May 2020, pp. 36–52. DOI: 10.1007/978-3-030-49461-2_3. URL: https://hal.inria.fr/hal-03084606.

[13] A. Siffer, A. Termier, P.-A. Fouque and C. Largouët. 'Netspot: a simple Intrusion Detection System with statistical learning'. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Guangzhou, China, 2021. DOI: 10.1109/TrustCom50675.2020.00122. URL: https://hal-agrocampus-ouest.archives-ouvertes.fr/hal-03145963.

[14] Y. Wang, R. Emonet, E. Fromont, S. Malinowski and R. Tavenard. 'Adversarial Regularization for Explainable-by-Design Time Series Classification'. In: ICTAI 2020 - 32th International Conference on Tools with Artificial Intelligence. online, Greece, 9th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03025671.

National peer-reviewed Conferences

[15] T. Guyet, P. Besnard, A. Samet, N. Ben Salha and N. Lachiche. 'Énumération des occurrences d'une chronique'. In: EGC 2020 - 20ème édition de la Conférence Extraction et Gestion des Connaissances. Bruxelles, Belgium, 27th Jan. 2020, pp. 1–8. URL: https://hal.inria.fr/hal-02422796.

Conferences without proceedings

[16] J. Bakalara, T. Guyet, O. Dameron, A. Happe and E. Oger. 'An extension of chronicles temporal model with taxonomies -Application to epidemiological studies'. In: HEALTHINF2021. online, France, 11th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03096846.

[17] V. Bellon-Maurel, L. Brossard, F. Garcia, N. Mitton, E. Prados and A. Termier. 'Livre blanc INRIA / INRAE sur le numérique en Agriculture'. In: PFIA 2020 - (Plateforme de l'Intelligence Artificielle) - Journée Agronomie et IA. Virtual, France: http://pfia2020.fr/, 29th June 2020, pp. 1–26. URL: https://hal.inrae.fr/hal-02887646.

[18] S. Corbillé, E. Fromont, E. Anquetil and P. Nerdeux. 'Integrating Writing Dynamics in CNN for Online Children Handwriting Recognition'. In: 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). Dortmund, Germany, 7th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-02940236.

[19] K. Fauvel, V. Masson and E. Fromont. 'A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers'. In: In Proceedings of the IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence. Yokohama, Japan, 8th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03094885.

[20] C.-S. Gauthier, R. Gaudel and E. Fromont. 'Bandits manchots avec échantillonnage de Thompson pour des recommandations multiples suivant un modèle fondé sur les positions'. In: CAp 2020 - Conférence sur l'Apprentissage automatique. Vannes, France, 23rd June 2020, pp. 1–2. URL: https://hal.archives-ouvertes.fr/hal-02872149.

[21] R. Gauthier, C. Largouët, L. Rozé and J.-Y. Dourmad. 'Algorithme de prédiction en temps réel de la consommation alimentaire journalière chez la truie en lactation'. In: 52 èmes Journées de la Recherche Porcine. Paris, France, 4th Feb. 2020. URL: https://hal.inria.fr/hal-03134418.

[22] L. Rousvoal, T. Allard and E. Fromont. 'Profilage et prédiction : de l'usage des algorithmes en matière pénale'. In: 2020 - Festival des libertés numériques. Rennes, France: https://fdln.insa-rennes.fr/, 1st Feb. 2020. URL: https://hal.archives-ouvertes.fr/hal-02905186.

[23] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon. 'Guided Attentive Feature Fusion for Multispectral Pedestrian Detection'. In: IEEE Winter Conference on Applications of Computer Vision. Waikoloa/Virtual, United States, 5th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03119907.

[24] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon. 'Localize to Classify and Classify to Localize: Mutual Guidance in Object Detection'. In: ACCV 2020 - The 15th Asian Conference on Computer Vision. Kyoto / Virtual, Japan, 30th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03025756.

[25] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon. 'Méthodes de Fusion de Cartes de Caractéristiques pour la Détection Multispectrale par Réseaux de Neurones Profonds'. In: RFIAP 2020 - Reconnaissance des Formes, Image, Apprentissage et Perception. Vannes, France, 23rd June 2020, pp. 1–2. URL: https://hal.archives-ouvertes.fr/hal-02872123.

[26] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon. 'Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks'. In: ICIP 2020 - IEEE International Conference on Image Processing. Abou Dabi, United Arab Emirates, 25th Oct. 2020, pp. 1–5. URL: https://hal.archives-ouvertes.fr/hal-02872132.

Scientific book chapters

[27] T. Guyet and Y. Dauxais. 'Generalized Chronicles for Temporal Sequence Classification'. In: *Extended articles of the International Workshop on Advanced Analytics and Learning on Temporal Data*. Vol. 12588. Lecture Notes in Computer Science. 16th Dec. 2020. URL: https://hal.inria.fr/hal-03090211.

Doctoral dissertations and habilitation theses

[28] K. Fauvel. 'Enhancing performance and explainability of multivariate time series machine learning methods : applications for social impact in dairy resource monitoring and earthquake early warning'. Université Rennes 1, 13th Oct. 2020. URL: https://tel.archives-ouvertes.fr/tel-03140673.

[29] M. Gueguen. 'Frequent Itemset Sampling of High Throughput Streams on FPGA Accelerators'. Université de Rennes 1 (UR1), 23rd Oct. 2020. URL: https://tel.archives-ouvertes.fr/tel-03120148.

[30] T. Guyet. 'Enhancing sequential pattern mining with time and reasoning'. Université de Rennes 1, 10th Feb. 2020. URL: https://tel.archives-ouvertes.fr/tel-02495270.

[31] C. Leverger. 'Investigation of a framework for seasonal time series forecasting'. Université Rennes 1, 20th Nov. 2020. URL: https://tel.archives-ouvertes.fr/tel-03118382.

12.3 Cited publications

[32] J. Castells, P.-D. Mohammad, L. Galárraga, G. Méndez, M. Ortiz-Rojas and A. Jiménez. 'A Student-oriented Tool to Support Course Selection in Academic Counseling Sessions'. In: *Proceedings of the Workshop on Adoption, Adaptation and Pilots of Learning Analytics in Under-represented Regions co-located with the 15th European Conference on Technology Enhanced Learning 2020*. Virtual Event, Germany, Sept. 2020. URL: https://hal.inria.fr/hal-03084671.

[33] S. Colas, C. Collin, P. Piriou and M. Zureik. 'Association between total hip replacement characteristics and 3-year prosthetic survivorship: A population-based study'. In: *JAMA Surgery* 150.10 (2015), pp. 979–988.

[34] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter. 'Efficient and Robust Automated Machine Learning'. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett. Curran Associates, Inc., 2015, pp. 2962–2970. URL: http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.

[35] L. Galárraga, C. Teflioudi, K. Hose and F. Suchanek. 'AMIE: Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases'. In: *WWW*. 2013.

[36] L. Galárraga, C. Teflioudi, K. Hose and F. M. Suchanek. 'Fast Rule Mining in Ontological Knowledge Bases with AMIE+'. In: *VLDB Journal* 24.6 (2015).

[37]   T. Kulesza, S. Stumpf, M. M. Burnett, S. Yang, I. Kwan and W.-K. Wong. 'Too much, too little, or just right? Ways explanations impact end users' mental models'. In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing* (2013), pp. 3–10.

[38]   V. Lifschitz. 'What Is Answer Set Programming?' In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*. 2008, pp. 1594–1597.

[39]   G. Moulis, M. Lapeyre-Mestre, A. Palmaro, G. Pugnet, J.-L. Montastruc and L. Sailler. 'French health insurance databases: What interest for medical research?' In: *La Revue de Médecine Interne* 36.6 (2015), pp. 411–417.

[40]   E. Nowak, A. Happe, J. Bouget, F. Paillard, C. Vigneau, P.-Y. Scarabin and E. Oger. 'Safety of Fixed Dose of Antihypertensive Drug Combinations Compared to (Single Pill) Free-Combinations: A Nested Matched Case–Control Analysis'. In: *Medicine* 94.49 (2015), e2229.

[41]   E. Polard, E. Nowak, A. Happe, A. Biraben and E. Oger. 'Brand name to generic substitution of antiepileptic drugs does not lead to seizure-related hospitalization: a population-based case-crossover study'. In: *Pharmacoepidemiology and drug safety* 24.11 (2015), pp. 1161–1169.