UMR IRISA

Activity Report 2019

Team LACODAM

Large scale Collaborative Data Mining

Joint team with Inria Rennes – Bretagne Atlantique

D7 – Data and Knowledge Management

CentraleSupélec cnrs rennes IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom Inria INSA RENNES Université Bretagne Sud ubs: UNIVERSITÉ DE RENNES 1

Table of contents

Project-Team LACODAM

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

Keywords:

Computer Science and Digital Science:

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, Visitors, External Collaborators

Research Scientists

Louis Bonneau de Beaufort [École nationale supérieure agronomique de Rennes, Researcher]
Luis Galárraga Del Prado [Inria, Researcher]

Faculty Members

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

Tassadit Bouadi [Univ de Rennes I, Associate Professor, from Jul 2019]
Elisa Fromont [Univ de Rennes I, Professor, HDR]
Thomas Guyet [École nationale supérieure agronomique de Rennes, Associate Professor]
Christine Largouët [École nationale supérieure agronomique de Rennes, Associate Professor, HDR]
Véronique Masson [Univ de Rennes I, Associate Professor]
Laurence Rozé [INSA Rennes, Associate Professor]

PhD Students

Erwan Bourrand [Advisor SLA, PhD Student]
Kevin Fauvel [Inria, PhD Student]
Samuel Felton [Univ de Rennes I, PhD Student, from Oct 2019]
Camille Sovanneary Gauthier [Louis Vuitton, PhD Student, from Apr 2019]
Maël Guillemé [Eniergency, PhD Student, granted by CIFRE]
Colin Leverger [Orange Labs, PhD Student]
Gregory Martin [PSA Automobiles, PhD Student, granted by CIFRE]
Anh Duong Nguyen [Vietnam, PhD Student, until May 2019]
Alban Siffer [AMOSSYS, PhD Student, granted by CIFRE]
Antonin Voyez [Inria, PhD Student, from Dec 2019]
Yichang Wang [CSC Grant, PhD Student]
Heng Zhang [Atermes, PhD Student, granted by CIFRE]

Technical staff

Remi Adon [Inria, Engineer, from Dec 2019]

Interns and Apprentices

Mohamed Abdel Wedoud [Univ de Rennes I, from Apr 2019 until Aug 2019]
Vaishnavi Bhargava [Inria, from Sep 2019 until Dec 2019]
Thomas Dahmen [INRA, from Jun 2019 until Jul 2019]
Issei Harada [Univ de Rennes I, from Apr 2019 until Sep 2019]
Sophie Le Bars [Inria, from January until Jul 2019]
Naima Mazari [Inria, from May 2019 until Jul 2019]
Mohammad Poul Doust [Univ de Rennes I, from Apr 2019 until Aug 2019]
Corentin Raphalen [INRA, from Jul 2019 until Sep 2019]
Josie Signe [Univ de Rennes I, from May 2019 until Jul 2019]
Allan Vitre [Univ de Rennes I, from Jun 2019 until Aug 2019]

Administrative Assistant

Gaelle Tworkowski [Inria, Administrative Assistant]

Visiting Scientist

Alexandre Sahuguede [Univ Paul Sabatier, until Jan 2019]

External Collaborators

Johanne Bakalara [Univ de Rennes I]
Philippe Besnard [CNRS, HDR]
Romaric Gaudel [École nationale de la statistique et de l'analyse de l'information]
Raphael Gauthier [INRA]
Anne-Isabelle Graux [INRA]

2. Overall Objectives

2.1. 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. Making sense out of data is 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 the most sought-after job in the USA, 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).

3. Research Program

3.1. Introduction

The three original research axes of the LACODAM project-team are the following. First, we briefly introduce these axes, as well as their interplay. We then introduce the axis of *Interpretable AI* (Section 3.4), whose emergence is a response to the current societal needs.

- The first research axis (Section 3.2) is dedicated to the design of *novel pattern mining methods*. Pattern mining is one of the most important approaches to discover novel knowledge in data, and one of our strongest areas of expertise. The work on this axis will serve as foundations for work on the other two axes. Thus, this axis will have the strongest impact on our overall goals.
- 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.

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

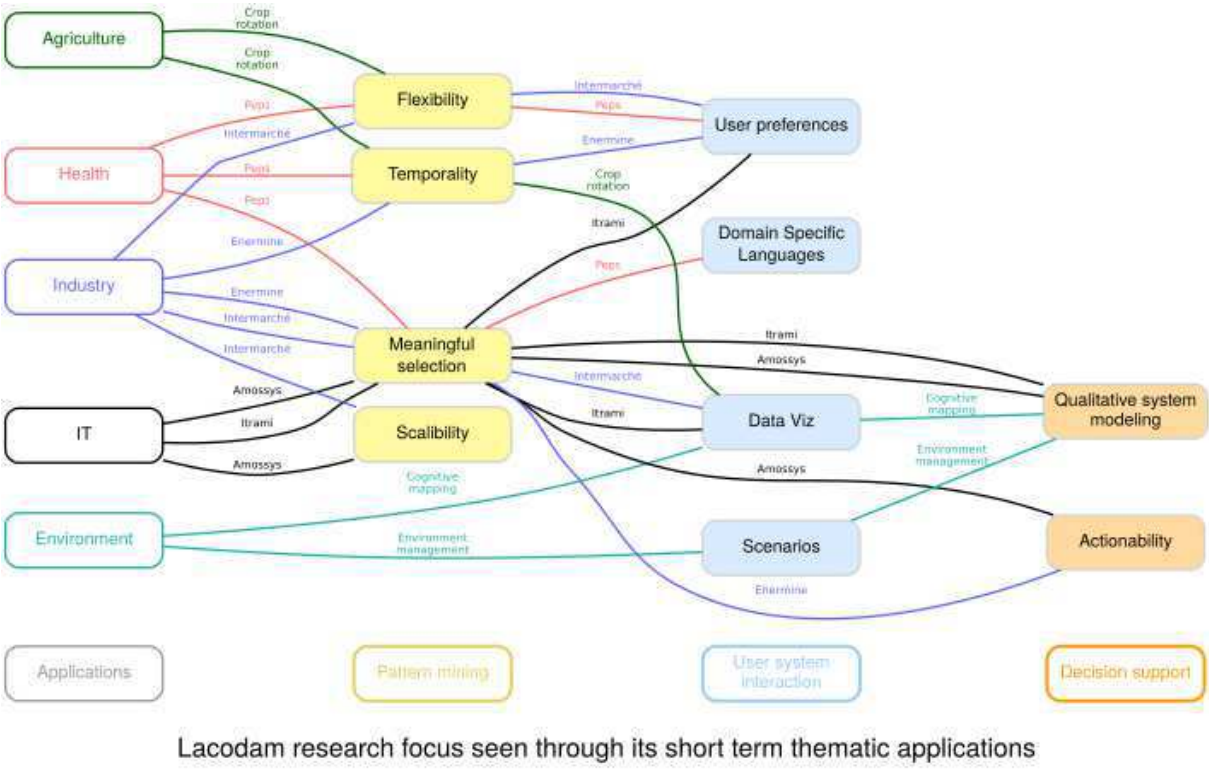

Figure 1. LACODAM research topics organized by axis and application

3.2. Pattern mining 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 huge set 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 four research topics:

1. the design of dedicated algorithms for mining temporal patterns
2. the design of flexible pattern mining approaches
3. the automatic selection of interesting data mining results
4. the design of parallel pattern algorithms to ensure scalability

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 her own, knowledge that will help define her pattern 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 logical knowledge representation paradigms.

Last but not least, pattern mining algorithms are compute-intensive. It is thus important to exploit all the available computing power. Parallelism is for a foreseeable future one of the main ways to speed up computations, and we have a strong competence on the design of parallel pattern mining algorithms. We will exploit this competence in order to guarantee that our approaches scale up to the data provided by our partners.

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:

- *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:* Most of the research directions presented in this document require users to examine patterns at some point. The output of most pattern mining algorithms is usually a (long) list of patterns. While this presentation can be sufficient for some applications, often it does not provide a complete understanding, especially for non-experts in pattern mining. A transversal research topic that we want to explore in LACODAM is to propose data visualization techniques that are adequate for understanding output results. Numerous (failed) experiments have shown that data mining and data visualization are fields, which require distinct skills, thus researchers in one field usually do not make significant advances in the other field (this is detailed in [Keim 2010]). Thus, 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

Patterns have proved to be quite useful for decision-aid. Predictive sequential patterns, to give an example, have a direct application in diagnosis. Itemsets and contrast patterns can be used for interpretable machine learning (ML). In regards to diagnosis, LACODAM inherits, from the former DREAM team, a strong background in decision support systems with internationally recognized expertise in this field. This subfield of AI (Artificial Intelligence) is concerned with determining whether a system is operating normally or not, and the cause of faulty behaviors. The studied system can be an agro- or eco-system, a software system (e.g., a ML classifier), a living being, etc. In relation to interpretable machine learning (ML), this subfield is concerned with the conception of models whose answers are understandable by users. This can be achieved by inducing inherently white-box models from data such as rule-based classifiers/regressors, or by mining rules and explanations from black-box models. The latter setting is quite common due to the high accuracy of black-box models compared to natively interpretable models. Pattern mining is a powerful tool to mine explanations from black-box systems. Those explanations can be used to diagnose biases in systems, either to debug and improve the model, or to generate trust in the verdicts of intelligent software agents.

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, LACODAM builds upon the idea that decision support systems (an interest bequeathed from DREAM) should take advantage of the available data. This third and last research axis is thus a meeting point for all members of the team, as it requires the integration of AI techniques for traditional decision support systems with results from data mining techniques.

Three main research sub-axes are investigated in LACODAM:

- *Diagnosis-based approaches.* We are exploring how to integrate knowledge found from pattern mining approaches, possibly with the help of interactive methods, into the qualitative models. The goal of such work is to automate as much as possible the construction of prediction models, which can require a lot of human effort.
- *Actionable patterns and rules.* In many settings of “exploratory data mining”, the actual interestingness of a pattern is hard to assess, as it may be subjective. However, for some applications there are well defined measures of interestingness and applicability for patterns. Patterns and rules that can lead to actual actions –that are relevant to the user– are called "actionable patterns" and are of vital importance to industrial settings.

- *Mining explanations from ML systems.* Interpretable ML and AI is a current trend for technical, ethical, and legal reasons [27]. In this regard, pattern mining can be used to spot regularities that arise when a complex black-box model yields a particular verdict. For instance, one may want to know the conditions under which the control module of a self-driving car decided to stop without apparent reason, or which factors caused a ML-based credit assessor to reject a loan request. Patterns and conditions are the building blocks for the generation of human-readable explanations for such black-box systems.

3.5. Interpretability

The pervasiveness of complex decision support systems, as well as the general consensus about the societal importance of understanding the rationale embedded in such systems [1], has given momentum to the field of interpretable 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 [28] 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 classification, or verbal phases in natural language processing) instead of surrogate still-machine-friendly features (such as super-pixels). LACODAM will tackle all 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:

- *Automating data science workflow discovery.* The current methods for knowledge extraction and construction of decision support systems require a lot of human effort. Our three research axes aim at alleviating this effort, by devising methods that are more generic and by improving the interaction between the user and the system. An ideal solution would be that the user could forget completely about the existence of pattern mining or decision support methods. Instead the user would only loosely specify her problem, while the system constructs various data science / decision support workflows, possibly further refined via interactions.

  We consider that this is a second order AI task, where AI techniques such as planning are used to explore the workflow search space, the workflow itself being composed of data mining and/or decision support components. This is a strategic evolution for data science endeavors, were the demand far exceeds the available human skilled manpower.

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

- *Logic argumentation based on epistemic interest.* Having increasingly automated approaches will require better and better ways to handle the interactions with the user. Our second long 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.

  This research direction aims at exploiting principles of interactive data analysis in the context of epistemic interestingness measures. Logic argumentation can be a natural tool for interactions between the user and the system: display of possibly exhaustive list of arguments, relationships between arguments (e.g., reinforcement, compatibility or conflict), possible solutions for argument conflicts, etc.

  The first step is to define a formal argumentation framework for explaining data mining results. This implies to continue theoretical work on the foundations of argumentation in order to identify the most adapted framework (either existing or a new one to be defined). Logic argumentation may be implemented and deeply explored in ASP, allowing us to build on our expertise in this logic language.
- *Collaborative feedback and knowledge management.* We are convinced that improving the data science process, and possibly automating it, will rely on high-quality feedback from communities on the web. Consider for example what has been achieved by collaborative platforms such as StackOverflow: it has become the reference site for any programming question.

  Data science is a more complex problem than programming, as in order to get help from the community, the user has to share her data and workflow, or at least some parts of them. This raises obvious privacy issues that may prevent this idea to succeed. As our research on automating the production of data science workflows should enable more people to have access to data science results, we are interested in the design of collaborative platforms to exchange expert advices over data, workflows and analysis results. This aims at exploiting human feedback to improve the automation of data science system via machine learning methods.

4. Application Domains

4.1. Introduction

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.

4.2. 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).

- **Resource Consumption Analysis for Optimizing Energy Consumption and Practices in Industrial Factories (Energiency)**. In order to increase their understanding of factory operation, companies introduce more and more sensors in their factories. Thus, the resource (electricity, water, etc.) consumption of engines, workshops and factories are recorded in the form of times series or temporal sequences. The person who is in charge of resource consumption optimization needs better software than classical spreadsheets for this purpose. He/she needs effective decision-aiding tools with statistical and artificial intelligence knowledge. The start-up Energiency aims at designing and

offering such pieces of software for analyzing energy consumption. The CIFRE PhD thesis of Maël (defended 16/12/2019) aimed at proposing new approaches and solutions from the data mining field to tackle this issue.

- **Security (Amossys)**. Current networks are faced with an increasing variety of attacks, from the classic "DDoS" that makes a server unusuable for a few fours, to advanced attacks that silently infiltrate a network and exfiltrate sensitive information months or even years later. Such intrusions, called APT (Advanced Persistent Threat) are extremely hard to detect, and this will become even harder as most communications will be encrypted. A promising solution is to work on "behavioral analysis", by discovering patterns based on the metadata of IP-packets. Such patterns can relate to an unusual sequencing of events, or to an unusual communication graph. Finding such complex patterns over a large volume of streaming data requires to revisit existing stream mining algorithms to dramatically improve their throughput, while guaranteeing a manageable false positive rate. We collaborated on this topic with the Amossys company and the EMSEC team of Irisa through the co-supervision of the CIFRE PhD of Alban Siffer (located in the EMSEC team, defended 19/12/2019). Our goal was in particular to design novel anomaly detection methods making minimal assumptions on the data, and able to scale to real traffic volumes.
- **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 able to detect anomalies (mainly persons) 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.

4.3. 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 [31], [30] or medical devices [23], 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 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 [29]. 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.

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

4.4. 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.

4.5. 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 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. The data visualization aspects of the tool, as well as the data science considerations and our emphasis on explainability are compiled in a paper that is under revision at the Learning Analytics and Knowledge Conference (LAK'20). The data science component of the tool was developed during the M1 internship of Mohammad Poul-Doust.

4.6. Others

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

- **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. Highlights of the Year

5.1. Highlights of the Year

- Elisa Fromont was awarded a Junior Member position at the Institut Universitaire de France (IUF). This is a prestigious position given for 5 years (2019-2024), the selection process is especially competitive.
- Tassadit Bouadi (MCF Univ Rennes 1) joined the team in July 2019. Her research topics are skyline queries and preference mining. Her work will especially contribute to the design of approaches having results easier to grasp by human users.

6. New Software and Platforms

6.1. REMI

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.

- Contact: Luis Galarraga Del Prado
- URL: http://gitlab.inria.fr/lgalarra/remi

6.2. HIPAR

Hierarchical Interpretable Pattern-aided Regression

KEYWORDS: Regression - Pattern extraction

FUNCTIONAL DESCRIPTION: Given a (tabular) dataset with categorical and numerical attributes, HIPAR is a Python library that can extract accurate hybrid rules that offer a trade-off between (a) interpretability, (b) accuracy, and (c) data coverage.

- Contact: Luis Galarraga Del Prado
- URL: https://gitlab.inria.fr/opelgrin/hipar

6.3. 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.

- Participant: Thomas Guyet
- Contact: Thomas Guyet
- Publication: Énumération des occurrences d'une chronique
- URL: https://gitlab.inria.fr/tguyet/pychronicles

7. New Results

7.1. Introduction

In this section, we organize the bulk of our contributions this year along two of our research axes, namely Pattern Mining and Decision Support. Some other contributions lie within the domain of machine learning.

7.1.1. Pattern Mining

In the domain of pattern mining we can categorize our contributions along the following lines:

- *Efficient Pattern Mining (Sections 7.2-7.4).* In [9], we propose a method to accelerate itemset sampling on FPGAs, whereas [18] proposes SSDPS, an efficient algorithm to mine discriminant patterns in two-class datasets, common in genetic data. Finally [11] presents a succinct data structure that represents concisely a cube of skypatterns.
- *Semantics of Pattern Mining (Sections 7.5-7.6).*[14] discusses the ambiguity of the semantics of pattern mining with absent events (negated statements). Likewise [8] shows formal properties of admissible generalizations in pattern mining and machine learning.

7.1.2. Decision Support

In regards to the axis of decision support, our contributions can be organized in two categories: forecasting & prediction, and modelisation.

- *Forecasting & Prediction (Sections 7.7-7.9).* In [10], we propose solutions to automate the task of capacity planning in the context of a large data network as the one available at Orange. [17] applies machine learning techniques for estrus detection in diary farms. [21] proposes a machine learning architecture in multi-sensor environments for earthquake early warning.
- *Modelling (Section 7.10).* In [5] we present a modeling approach for the nutritional requirements of lactating sows.
- *Data Exploration (Section 7.11).*[6] proposes a formal framework for the exploration of care trajectories in medical databases.

7.1.3. Others

- *Machine Learning (Section 7.12-7.14).*[7], [16] proposes novel methods to optimize the F-measure in ML, and to improve the task of domain adaptation by source selection. [19] proposes the use of GANs to make time series classification more interpretable.

7.2. Accelerating Itemset Sampling using Satisfiability Constraints on FPGA

Finding recurrent patterns within a data stream is important for fields as diverse as cybersecurity or e-commerce. This requires to use pattern mining techniques. However, pattern mining suffers from two issues. The first one, known as "pattern explosion", comes from the large combinatorial space explored and is the result of too many patterns output to be analyzed. Recent techniques called output space sampling solve this problem by outputting only a sampled set of all the results, with a target size provided by the user. The second issue is that most algorithms are designed to operate on static datasets or low throughput streams. In [9], we propose a contribution to tackle both issues, by designing an FPGA accelerator for pattern mining with output space sampling. We show that our accelerator can outperform a state-of-the-art implementation on a server class CPU using a modest FPGA product.

7.3. Statistically Significant Discriminative Patterns Searching

In [18], we propose a novel algorithm, named SSDPS, to discover patterns in two-class datasets. The SSDPS algorithm owes its efficiency to an original enumeration strategy of the patterns, which allows to exploit some degrees of anti-monotonicity on the measures of discriminance and statistical significance. Experimental results demonstrate that the performance of the SSDPS algorithm is better than others. In addition, the number of generated patterns is much less than the number of the other algorithms. Experiment on real data also shows that SSDPS effciently detects multiple SNPs combinations in genetic data.

7.4. Compressing and Querying Skypattern Cubes

Skypatterns are important since they enable to take into account user preference through Pareto-dominance. Given a set of measures, a skypattern query finds the patterns that are not dominated by others. In practice, different users may be interested in different measures, and issue queries on any subset of measures (a.k.a subspace). This issue was recently addressed by introducing the concept of skypattern cubes. However, such a structure presents high redundancy and is not well adapted for updating operations like adding or removing measures, due to the high costs of subspace computations in retrieving skypatterns. In [11], we propose a new structure called Compressed Skypattern Cube (abbreviated CSKYC), which concisely represents a skypattern cube, and gives an efficient algorithm to compute it. We thoroughly explore its properties and provide an efficient query processing algorithm. Experimental results show that our proposal allows to construct and to query a CSKYC very efficiently.

7.5. Semantics of Negative Sequential Patterns

In the field of pattern mining, a negative sequential pattern expresses behavior by a sequence of present and absent events. In [14], we shed light on the ambiguity of this notation and identify eight possible semantics with the relation of inclusion of a motif in a sequence. These semantics are illustrated and we are studying them formally. We thus propose dominance and equivalence relationships between these semantics, and we highlight new properties of anti-monotony. These results could be used to develop new efficient algorithms for mining frequent negative sequential patterns.

7.6. Admissible Generalizations of Examples as Rules

Rule learning is a data analysis task that consists in extracting rules that generalize examples. This is achieved by a plethora of algorithms. Some generalizations make more sense for the data scientists, called here admissible generalizations. The purpose of our work in [8] is to show formal properties of admissible generalizations. A formalization for generalization of examples is proposed allowing the expression of rule admissibility. Some admissible generalizations are captured by preclosure and capping operators. Also, we are interested in selecting supersets of examples that induce such operators. We then define classes of selection functions. This formalization is more particularly developed for examples with numerical attributes. Classes of such functions are associated with notions of generalization and they are used to comment some results of the CN2 algorithm [22].

7.7. Towards a Framework for Seasonal Time Series Forecasting Using Clustering

Seasonal behaviours are widely encountered in various applications. For instance, requests on web servers are highly influenced by our daily activities. Seasonal forecasting consists in forecasting the whole next season for a given seasonal time series. It may help a service provider to provision correctly the potentially required resources, avoiding critical situations of over or under provision. In [10], we propose a generic framework to make seasonal time series forecasting. The framework combines machine learning techniques 1) to identify the typical seasons and 2) to forecast the likelihood of having a season type in one season ahead. We study this framework by comparing the mean squared errors of forecasts for various settings and various datasets. The best setting is then compared to state-of-the-art time series forecasting methods. We show that it is competitive with then.

7.8. Towards Sustainable Dairy Management - A Machine Learning Enhanced Method for Estrus Detection

Our research tackles the challenge of milk production resource use efficiency in dairy farms with machine learning methods. Reproduction is a key factor for dairy farm performance since cows milk production begin with the birth of a calf. Therefore, detecting estrus, the only period when the cow is susceptible to pregnancy, is crucial for farm efficiency. Our goal is to enhance estrus detection (performance, interpretability), especially on the currently undetected silent estrus (35% of total estrus), and allow farmers to rely on automatic estrus detection solutions based on affordable data (activity, temperature). In [17] we first propose a novel approach with real-world data analysis to address both behavioral and silent estrus detection through machine learning methods. Second, we present LCE, a local cascade based algorithm that significantly outperforms a typical commercial solution for estrus detection, driven by its ability to detect silent estrus. Then, our study reveals the pivotal role of activity sensors deployment in estrus detection. Finally, we propose an approach relying on global and local (behavioral versus silent) algorithm interpretability (SHAP) to reduce the mistrust in estrus detection solutions.

7.9. A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning

Our research [21] aims to improve the accuracy of Earthquake Early Warning (EEW) systems by means of machine learning. EEW systems are designed to detect and characterize medium and large earthquakes before their damaging effects reach a certain location. Traditional EEW methods based on seismometers fail to accurately identify large earthquakes due to their sensitivity to the ground motion velocity. The recently introduced high-precision GPS stations, on the other hand, are ineffective to identify medium earthquakes due to its propensity to produce noisy data. In addition, GPS stations and seismometers may be deployed in large numbers across different locations and may produce a significant volume of data consequently, affecting the response time and the robust-ness of EEW systems. In practice, EEW can be seen as a typical classification problem in the machine learning field: multi-sensor data are given in input, and earthquake severity is the classification result. In this paper, we introduce the Distributed Multi-Sensor Earthquake Early Warning (DMSEEW) system, a novel machine learning-based approach that combines data from both types of sensors (GPS stations and seismometers) to detect medium and large earthquakes. DMSEEW is based on a new stacking ensemble method which has been evaluated on a real-world dataset validated with geoscientists. The system builds on a geographically distributed infrastructure, ensuring an efficient computation in terms of response time and robustness to partial infrastructure failures. Our experiments show that DMSEEW is more accurate than the traditional seismometer-only approach and the combined-sensors (GPS and seis-mometers) approach that adopts the rule of relative strength.

7.10. Dynamic Modeling of Nutrient Use and Individual Requirements of Lactating Sows

Nutrient requirements of sows during lactation are related mainly to their milk yield and feed intake, and vary greatly among individuals. In practice, nutrient requirements are generally determined at the population level based on average performance. The objective of the present modeling approach was to explore the variability in nutrient requirements among sows by combining current knowledge about nutrient use with on-farm data available on sows at farrowing [parity, BW, backfat thickness (BT)] and their individual performance (litter size, litter average daily gain, daily sow feed intake) to estimate nutrient requirements. The approach was tested on a database of 1,450 lactations from 2 farms. The effects of farm (A, B), week of lactation (W1: week 1, W2: week 2, W3+: week 3 and beyond), and parity (P1: 1, P2: 2, P3+: 3 and beyond) on sow performance and their nutrient requirements were evaluated. The mean daily ME requirement was strongly correlated with litter growth ($R2 = 0.95$; $P < 0.001$) and varied slightly according to sow BW, which influenced the maintenance cost. The mean daily standardized ileal digestible (SID) lysine requirement was influenced by farm, week of lactation, and parity. Variability in SID lysine requirement per kg feed was related mainly to feed intake

(R2 = 0.51; P < 0.001) and, to a smaller extent, litter growth (R2 = 0.27; P < 0.001). It was lowest in W1 (7.0 g/kg), greatest in W2 (7.9 g/kg), and intermediate in W3+ (7.5 g/kg; P < 0.001) because milk production increased faster than feed intake capacity did. It was lower for P3+ (6.7 g/kg) and P2 sows (7.3 g/kg) than P1 sows (8.3 g/kg) due to the greater feed intake of multiparous sows. The SID lysine requirement per kg of feed was met for 80% of sows when supplies were 112 and 120% of the mean population requirement on farm A and B, respectively, indicating higher variability in requirements on farm B. Other amino acid and mineral requirements were influenced in the same way as SID lysine. In [5], we present a modeling approach that allows us to capture individual variability in the performance of sows and litters according to farm, stage of lactation, and parity. It is an initial step in the development of new types of models able to process historical farm data (e.g., for ex post assessment of nutrient requirements) and real-time data (e.g., to control precision feeding).

7.11. Temporal Models of Care Sequences for the Exploration of Medico-administrative Data

Pharmaco-epidemiology with medico-administrative databases enables the study of the impact of health products in real-life settings. These studies require to manipulate the raw data and the care trajectories, in order to identify pieces of data that may witness the medical information that is looked for. The manipulation can be seen as a querying process in which a query is a description of a medical pattern (e.g. occurrence of illness) with the available raw features from care trajectories (e.g. occurrence of medical procedures, drug deliveries, etc.). The more expressive is the querying process, the more accurate is the medical pattern search. The temporal dimension of care trajectories is a potential information that may improve the description of medical patterns. The objective of this work [6] is to propose a formal framework that would design a well-founded tool for querying care trajectories with temporal medical patterns. In this preliminary work, we present the problematic and we introduce a use case that illustrates the comparison of several querying formalisms.

7.12. Improving Domain Adaptation By Source Selection

Domain adaptation consists in learning from a source data distribution a model that will be used on a different target data distribution. The domain adaptation procedure is usually unsuccessful if the source domain is too different from the target one. In [16], we study domain adaptation for image classification with deep learning in the context of multiple available source domains. This work proposes a multi-source domain adaptation method that selects and weights the sources based on inter-domain distances. We provide encouraging results on both classical benchmarks and a new real world application with 21 domains.

7.13. From Cost-Sensitive Classification to Tight F-measure Bounds

The F-measure is a classification performance measure, especially suited when dealing with imbalanced datasets, which provides a compromise between the precision and the recall of a classifier. As this measure is non-convex and non-linear, it is often indirectly optimized using cost-sensitive learning (that affects different costs to false positives and false negatives). In [7], we derive theoretical guarantees that give tight bounds on the best F-measure that can be obtained from cost-sensitive learning. We also give an original geometric interpretation of the bounds that serves as an inspiration for CONE, a new algorithm to optimize for the F-measure. Using 10 datasets exhibiting varied class imbalance, we illustrate that our bounds are much tighter than previous work and show that CONE learns models with either superior F-measures than existing methods or comparable but in fewer iterations.

7.14. Time Series Classification Based on Interpretable Shapelets

[19] proposes a new architecture, called AI⟷PR-CNN, composed of generative adversarial neural networks (GANs), which addresses the problem of the lack of interpretability of the existing methods for time series classification. Our network has two components: a classifier and a discriminator. The classifier is a CNN, it serves to classify series. Convolutions are discriminant patterns learned from the data that allow for a more

discriminating representation of time series (similar to a shapelet). To be able to explain the decision of the classifier, we would like to impose that the convolutions used are real "shapelets", that is to say that they are close to real sub-series present in the training set. This constraint is implemented by a GAN whose purpose will determine how much the weight matrices classifier convolutions are close to subset of the training set.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

- **AdvisorSLA 2018 - Inria**
  Participants: E. Bourrand, L. Galárraga, E. Fromont, A. Termier
  Contract amount: 7,5k€
  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 is a pre-doctoral contract signed with AdvisorSLA to start the work for the PhD of E. Bourrand (Thèse CIFRE) while the corresponding administrative formalities are completed.
- **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: 15k€
  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 a pre-doctoral contract to start the work for the PhD of G. Martin (Thèse CIFRE) while the corresponding administrative formalities are completed.

9. Partnerships and Cooperations

9.1. 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) its decisions can be better understood.
- **Hyptser: Hybrid Prediction of Time Series**
  Participants: T. Guyet, S. Malinowski (LinkMedia), V. Lemaire (Orange)
  HYPTSER is a collaborative project between Orange Labs and LACODAM funded by the Fondation Mathématique Jacques Hadamard (PGMO program). It aims at developping 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. Hybrid models aim at jointly partitioning the data, learning forecasting models in each partition and learning how to combine their outputs. We are currently developing two different approaches for that purpose, one based on the MODL framework and the other based on neural networks. We describe these approaches below:

- MODL is a mathematical framework that turns the learning task into a model selection problem. It aims at finding the most probable model given the data. The MODL approach has been applied on numerous learning tasks. In all cases, this approach leads to a regularized optimization criterion. We formalize a new MODL criterion able to learn hybrid models on time series in order to: i) make a partition of time series; ii) learn local regression models. This approach formalizes these two steps in a unified way.
- We are also developing an hybrid neural network structure that is able to learn automatically a soft partitioning of the data together with local models on each partition.

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.

- **#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, make innovative research on the use of digital techniques in agriculture in order to improve competitiveness, preserve the environment, and offer correct living conditions to farmers. Second, prepare 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).

9.1.1. 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 [26], 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).

9.1.2. National Platforms

- **PEPS: Pharmaco-epidemiology for Health Products**
  Participants: J. Bakalara, Y. Dauxais, T. Guyet, V. Masson, R. Quinou, A. Samet
  The PEPS project (Pharmaco-epidemiology des Produits de Santé) is funded by the ANSM (National Agency for Health Security). The project leader is E. Oger from the clinical investigation center CIC-1414 INSERM/CHU Rennes. The other partners located in Rennes are the Institute of Research and Technology (IRT), B<>Com, EHESP and the LTSI. The project started in January 2015 and is funded for 4 years. The PEPS project consists of two parts: a set of clinical studies and a research program dedicated to the development of innovative tools for pharmaco-epidemiological studies with medico-administrative databases. Our contribution to this project will be to propose pattern mining algorithms and reasoning techniques to analyse the typical care pathways of specific groups of insured patients. Since last year we have been working on the design and development of algorithms [25], [24] to mine patterns on care pathways.

9.2. International Research Visitors

9.2.1. Internships

From September to December 2019 we hosted Vaishnavi Bhargava, a computer science student from the Birla Institute of Technology and Science in Pilani, who worked on "Automatic Neighborhood Design for Localized Model-interpretation". Her work aimed at finding a set of metrics and procedures to determine the best parameterization of the method LIME for local post-hoc interpretability of machine learning models. The goal of this effort is to inform users of the parameter values (if any) for which a LIME explanation should be trusted because it can faithfully reproduce the behavior of the black-box it tries to explain.

10. Dissemination

10.1. Scientific Events: Organisation

10.1.1. General Chair, Scientific Chair

- Elisa Fromont was one of the general chairs of ECML/PKDD 2019 that took place in Würzburg, Germany (Sept).
- Tassadit Bouadi was co-chair of the ECML/PKDD'2019 PhD Forum
- Tassadit Bouadi and Luis Galárraga are organizers of the two editions of the AIMLAI workshop (Advances in Interpretable Machine Learning and Artificial Intelligence) colocated with the conferences EGC (https://project.inria.fr/aimlai/) and ECML/PKDD (https://kdd.isti.cnr.it/xkdd2019/) 2019 respectively.
- AALTD@ECML (T. Guyet), https://project.inria.fr/aaltd19/
- GAST@EGC (T. Guyet), https://gt-gast.irisa.fr/gast-2019/
- GeoInformation Analytics Technical Track at the ACM/SIGAPP Symposium On Applied Computing (T. Guyet), https://gia.sciencesconf.org/
- Time series days, https://project.inria.fr/tsdays/ (T. Guyet)

10.1.2. Scientific Events: Selection

10.1.2.1. Member of the Conference Program Committees

- KDD (A. Termier, E. Fromont)
- ICTAI (T. Guyet)
- ECAI (T. Bouadi)
- ECML-PKDD (T. Guyet, A. Termier, E. Fromont, T. Bouadi)

- IJCAI (T. Guyet, A. Termier)
- SDM (A. Termier)
- MedInfo (T. Guyet)
- EGC (T. Bouadi, T. Guyet)
- HiPC (A. Termier)
- APIA (C. Largouët)
- CAp (R. Gaudel)
- ISWC (L. Galárraga)
- LDK (L. Galárraga)

10.1.2.2. Reviewer

- ECML-PKDD (C. Largouët)
- ICML (R. Gaudel)
- NeurIPS (R. Gaudel)
- LOD (R. Gaudel)
- ICTAI (L. Galárraga)
- BDA (L. Galárraga)

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing Activities

- Pattern Recognition (R. Gaudel)
- Revue d'intelligence artificielle (R. Gaudel)
- Data Mining and Knowledge Discovery (L. Galárraga, A. Termier)
- Semantic Web Journal (L. Galárraga)
- Remote Sensing (T. Guyet)
- Journal of Biomedical Informatics (T. Guyet)

10.1.4. Invited Talks

- Signatures: detecting and characterizing recurrent behavior in sequential data (A. Termier, March 2019). Invited talk to the Exploratory Data Analysis group of Jilles Vreeken, Saarebruck University.
- Pattern Mining with MDL (A. Termier, May 2019). Talk at the Dagstuhl seminal on Enumeration in Data Management.
- (Random) Search in a Structured Space (R.Gaudel, Oct. 2019). GdR IA course, Paris, France.
- Recommendation as a sequential process (R. Gaudel, June 2019). Obelix, Vannes, France.
- Litterature review: From computational complexity of function optimization to large scale Machine Learning (R. Gaudel, June 2019). Thematic season on the topic COMPLEXITY, IRISA, Rennes, France
- Interpretability in Classifiers (L. Galárraga, May 2019). Journée EGC&IA, Orsay, France.
- Mining sequential patterns with ASP (T. Guyet, January 2019). LERIA, Angers, France. 6/12/2019: Master Class at La Digital Tech Conference on "IA, Machine learning et Deep learning : démêlons les concepts !", Rennes.
- (E. Fromont) 25/11/2019: Participation in the panel organized by EIT Digital for the "Industrial Doctorate meetup: Boosting business with data science", Rennes.
- (E. Fromont) 8/10/2019: Invited speaker for Planète conférence on "Mythes et réalités de l'I.A.", Vannes.

- (E. Fromont) 9/09/2019: Participation in the workshop "Smart-cities et mobilités" of the Forum Européen d'Intelligence Artificielle Territoriale organised by the CHEMI, Rambouillet, Fr.
- (E. Fromont) 5/07/2019: Participation in the panel organized by Google numérique on "Intelligence artificielle, opportunités diverses et défis communs?", Rennes.
- (E. Fromont) 1/07/2019: Participation in the panel organized by EIT summer school "Unleashing the power of data for better cities" on "Future of smart cities and data innovation", Rennes.
- (E. Fromont) 6/06/2019: Invited speaker for Breizh Conseil on "Mythes et réalités de l'I.A.", Rennes (with a video teaser)
- (E. Fromont) 5/06/2019: Invited speaker for the "journées nationales MIAGE" (JNM2019) on "Mythes et réalités de l'I.A.", Rennes.
- (E. Fromont) 23/05/2019: Invited speaker for ESTIM NUMERIQUE on "Qu'est ce que l'I.A.", Rennes.
- (E. Fromont) 31/01/2019: Invited speaker for the Ecole Navale Science Day 2019 JSN'19 on "New Challenges in Computer Vision", Brest.
- (E. Fromont) 26/01/2019: Invited speaker for the Digital transformer Challenge (Introduction to AI), Rennes.

10.1.5. Scientific Expertise

Scikit-Mine. The team recently secured a two-year engineer position (Inria ADT grant) for working on MDL-based pattern mining algorithms. The goal of the project is to design a modern pattern mining library fully compatible with scikit-learn (the major data science library). The library will contain major MDL-based pattern mining algorithms of the state of the art, including those developed in the team (periodic pattern mining). The project will start in fall 2020.

10.1.6. Research Administration

- Alexandre Termier is member of the Scientific Committee of the "Environment and Agriculture" department of INRA (now INRAe).
- (E. Fromont) Member of an HCERES evaluation committee.
- (E. Fromont) Member of the executive board of the CominLabs LABEX for the "Data, IA and Robotics" scientific axis.
- (E. Fromont) Elected at the scientific council of the Société Savante Francophone d'Apprentissage Machine.
- (E. Fromont) Nominated at the CNRS INS2I ("INstitut des Sciences de l'Information et de leurs Interactions") scientific council (CSI)

10.2. Teaching - Supervision - Juries

10.2.1. 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, Neural Networks, 4,5h eq. TD, école d'été, ENSAI, France

Master : R. Gaudel, Systèmes de recommandation, 27h eq. TD, M2, ENSAI, France

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

Master : R. Gaudel, Bandits Theory, 4,5h eq. TD, école d'été, ENSAI, France

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

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

10.2.2. Supervision

PhD (defended 25/09/2019): Guillaume Metzler, "Learning from Imbalanced Data: An Application to Bank Fraud Detection", 15/01/2016, E. Fromont, M. Sebban, A. Habrard.

PhD (defended 16/12/2019): Kevin Bascol, "Multi-source domain adaptation on imbalanced data:application to the improvement of chairlifts safety", 1/11/2016, E. Fromont,R. Emonet.

PhD (defended 16/12/2019): Maël Guillemé, "New Data Mining Approaches for Improving Energy Consumption in Factories", 03/10/2016, L. Rozé, V. Masson, A. Termier

PhD (defended 19/12/2019): Alban Siffer, "Data Mining Approaches for Cyber Attack Detection", 03/2016, P-A Fouque, A. Termier, C. Largouët.

PhD in progress: Maël Gueguen, "Improving the Performance and Energy Efficiency of Complex Heterogeneous Manycore Architectures with On-chip Data Mining", 01/11/2016, O. Sentieys, A. Termier

PhD in progress: Raphaël Gauthier, "Modelling of Nutrient Utilization and Precision Feeding of Lactating Sows", 01/11/2017, C. Largouët, J.-Y. Dourmad

PhD in progress: Kévin Fauvel, "Using Data mining Techniques for Improving Dairy Management", 01/10/2017, V. Masson, A. Termier, P. Faverdin

PhD in progress: Heng Zhang, "Deep Learning on Multimodal Data for the Supervision of Sensitive Sites", 03/12/2018, E. Fromont, S. Lefevre

PhD in progress: Yichang Wang, "Interpretable Shapelet for Anomaly Detection in Time Series", 15/04/2018, E. Fromont, S. Malinowski, R. Tavenard, R. Emonet

PhD in progress: Camille-Sauvanneary Gauthier, Session-aware Recommendation System, since March 15. 2019, advisors: É. Fromont, R. Gaudel, & B. Guilbot

PhD in progress: Erwan Bourrand, Interactive Data Mining for Root Cause Analysis of Performance Issues in Networks, advisors: E. Fromont, L. Galárraga, A. Termier.

PhD in progress: Gregory Martin, Data mining to optimize a free-floating car sharing service: E. Fromont, L. Rozé, A. Termier.

PhD in progress: Johanne Bakalara, Temporal models of care sequences for the exploration of medico-administrative data, advisors: E. Oger, O. Dameron, T. Guyet, A. Happe

PhD in progress: Colin Leverger, Management and forecasting of measures for server optimization in a cloud context, advisors: A. Termier, R. Marguerie, S. Malinowski, T. Guyet, L. Rozé.

PhD in progress: Nicolas Sourbier, Log Analysis Using Reinforcement Learning for Intrusion Detection, advisors: F. Majorczyke, O. Gesny, T. Ollivier, M. Pelcat, T. Guyet.

PhD in progress: Yiru Zhang, Modeling and management of imperfect preferences with the theory of belief functions, advisors: T. Bouadi, A. Martin.

10.2.3. Juries

- PhD Justine Reynaud (Orpailleur, Inria Nancy, Université de Lorraine): Découverte de définitions dans le web des données (L. Galárraga, examiner, december 2019).
- PhD Anes Bendimerad (INSA Lyon): Mining useful patterns in attributed graphs (A. Termier, reviewer, september 2019)
- HDR Frédéric Flouvat (Université de Nouvelle Calédonie): Extraction de motifs spatio-temporels: co-localisations, séquences et graphes dynamiques attribués (A. Termier, reviewer, october 2019)
- HDR Christine Largouët (Université de Rennes 1): Intelligence Artificielle pour l'aide à la décision des systèmes dynamiques: diagnostic, prévision, recommandation d'actions (A. Termier, examiner, december 2019)
- PhD Ricardio Sperandio (Université de Rennes 1): Time series retrieval using DTW-preserving shapelets (A. Termier, examiner, december 2019)
- PhD Raphaël Jakse (Université Grenoble Alpes): Interactive runtime verification (A. Termier, reviewer, december 2019)
- (E. Fromont) Ziyu GUO, Marseille 5/11/2019 (committee member); Jordan Frery, Saint-Etienne 26/09/2019 (committee member, president); Julien Salotti, Lyon 24/09/2019 (reviewer); Xuhong Li, Compiègne (committee member, president) 10/99/2019; Michele Linardi, Paris (committee member) 21/08/2019; Pauline Luc, Grenoble (committee member, president) 25/06/2019, Diogo Luvizon, Cergy (reviewer) 8/04/2019; Zaruhi Alaverdyan, Lyon (committee member, president) 18/01/2019
- PhD Trung-Dung Le (Shaman, IRISA, Université de Rennes 1): Gestion de masses de données dans une fédération de nuages informatiques (T. Bouadi, examiner, july 2019)

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

- (E. Fromont) 2018- ?: Nominated at the scientific council of the "Fondation Blaise Pascal" (dedicated to scientific mediation, funded by Inria)

10.3.2. Education

- In 2019, Alexandre Termier and Elisa Fromont are co-heads of a special training for high school teachers, ordered by the Ministry of Education. From September 2019, French high school students will have the opportunity to have reinforced Computer Science courses, with the same importance as Mathematics or Physics for example. As new teachers are not yet hired to teach theses courses, there was thus an urgent need to train existing high school teachers so that they can train the new teachers. A national formation has been set up, with some coordination between all universities. Alexandre and Elisa represent the University of Rennes, which is the formation center for all of Britany. The formation load is 125 teaching hours, to a group of 71 high school teachers.
- Elisa Fromont is also leading the set up of a CAPES in University of Rennes 1, which will allow to prepare future high school teachers in computer science.

10.3.3. Interventions

- A. Termier participated at a mediation action on Artificial Intelligence for a general technical public, the "Trial of metro car number 42", at the Digital Tech in Rennes (november 2019, https://www.ladigital.tech/proces-rame-metro-numero-42/)
- Alexandre Termier did two 2h interventions at the Emile Zola high school, aimed at math teachers. The topics were "Artificial Intelligence" and "Big Data".
- (E. Fromont) 6/12/2019: Master Class at La Digital Tech Conference on "IA, Machine learning et Deep learning : démêlons les concepts !", Rennes.

11. Bibliography

Major publications by the team in recent years

[1] K. FAUVEL, V. MASSON, E. FROMONT, P. FAVERDIN, 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", Anchorage, United States, 25th SIGKDD Conference on Knowledge Discovery and Data Mining proceedings, August 2019, pp. 1-9 [*DOI :* 10.1145/3292500.3330712], https://hal.archives-ouvertes.fr/hal-02190790

[2] E. GALBRUN, P. CELLIER, N. TATTI, A. TERMIER, 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, https://hal.archives-ouvertes.fr/hal-01951722

Publications of the year

Doctoral Dissertations and Habilitation Theses

[3] C. LARGOUËT. *Artificial Intelligence for decision support in dynamic systems: Diagnosis, Prediction, Recommendation of actions*, Université de Rennes 1, December 2019, Habilitation à diriger des recherches, https://hal.inria.fr/tel-02437159

Articles in International Peer-Reviewed Journals

[4] A. AYADI, A. SAMET, F. D. B. DE BEUVRON, C. ZANNI-MERK. *Ontology population with deep learning-based NLP: a case study on the Biomolecular Network Ontology*, in "Procedia Computer Science", 2019, vol. 159, pp. 572-581 [*DOI :* 10.1016/J.PROCS.2019.09.212], https://hal.archives-ouvertes.fr/hal-02357839

[5] R. GAUTHIER, C. LARGOUËT, C. GAILLARD, L. CLOUTIER, F. GUAY, J.-Y. DOURMAD. *Dynamic modeling of nutrient use and individual requirements of lactating sows*, in "Journal of Animal Science", May 2019, vol. 97, n^{o} 7, pp. 2822–2836 [*DOI :* 10.1093/JAS/SKZ167], https://hal.archives-ouvertes.fr/hal-02142870

International Conferences with Proceedings

[6] J. BAKALARA, T. GUYET, O. DAMERON, E. OGER, A. HAPPE. *Temporal models of care sequences for the exploration of medico-administrative data*, in "2019 - Workshop IA&Santé, PFIA", Toulouse, France, July 2019, pp. 1-8, https://hal.archives-ouvertes.fr/hal-02265743

[7] K. BASCOL, R. EMONET, E. FROMONT, A. HABRARD, G. METZLER, M. SEBBAN. *From Cost-Sensitive Classification to Tight F-measure Bounds*, in "AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics", Naha, Okinawa, Japan, The 22nd International Conference on Artificial Intelligence and Statistics, April 2019, vol. 89, n^{o} 1, pp. 1245-1253, https://hal.archives-ouvertes.fr/hal-02049763

[8] P. BESNARD, T. GUYET, V. MASSON. *Admissible Generalizations of Examples as Rules*, in "ICTAI 2019 - 31st International Conference on Tools with Artificial Intelligence", Portland, United States, November 2019, pp. 1480-1485, https://hal.inria.fr/hal-02267166

[9] M. GUEGUEN, O. SENTIEYS, A. TERMIER. *Accelerating Itemset Sampling using Satisfiability Constraints on FPGA*, in "DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe", Florence, Italy, IEEE, March 2019, pp. 1046-1051 [*DOI :* 10.23919/DATE.2019.8714932], https://hal.inria.fr/hal-01941862

[10] C. LEVERGER, S. MALINOWSKI, T. GUYET, V. LEMAIRE, A. BONDU, A. TERMIER. *Toward a Framework for Seasonal Time Series Forecasting Using Clustering*, in "IDEAL 2019", Manchester, United Kingdom, October 2019, pp. 328-340 [*DOI :* 10.1007/978-3-030-33607-3_36], https://hal.inria.fr/hal-02371221

[11] W. UGARTE, S. LOUDNI, P. BOIZUMAULT, B. CRÉMILLEUX, A. TERMIER. *Compressing and Querying Skypattern Cubes*, in "IEA/AIE-2019 - 32nd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems", Graz, Austria, Advances and Trends in Artificial Intelligence. From Theory to Practice 32nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2019, Graz, Austria, July 9–11, 2019, Proceedings, Springer, July 2019, pp. 406-421 [*DOI :* 10.1007/978-3-030-22999-3_36], https://hal.archives-ouvertes.fr/hal-02190788

National Conferences with Proceedings

[12] R. GAUTHIER, C. LARGOUËT, C. GAILLARD, L. CLOUTIER, F. GUAY, J.-Y. DOURMAD. *Dynamic modelling of nutrient use and individual requirements in lactating sows*, in "JRP 2019 - 51èmes Journées de la Recherche Porcine", Paris, France, IFIP - Institut du Porc, 2019, pp. 117-122, https://hal.archives-ouvertes.fr/hal-02010788

[13] T. GUYET, P. BESNARD, A. SAMET, N. BEN SALHA, N. LACHICHE. *Énumération des occurrences d'une chronique*, in "EGC 2020 - 20ème édition de la Conférence Extraction et Gestion des Connaissances", Bruxelles, Belgium, January 2020, pp. 1-8, https://hal.inria.fr/hal-02422796

[14] T. GUYET. *Semantic(s) of negative sequential patterns*, in "JIAF 2019 - 13èmes Journées d'Intelligence Artificielle Fondamentale", Toulouse, France, July 2019, pp. 1-10, https://hal.inria.fr/hal-02188501

Conferences without Proceedings

[15] J. BAKALARA. *Temporal models of care sequences for the exploration of medico-administrative data*, in "AIME 2019 - 17th Conference on Artificial Intelligence in Medicine", Poznan, Poland, June 2019, pp. 1-7, https://hal.archives-ouvertes.fr/hal-02265731

[16] K. BASCOL, R. EMONET, E. FROMONT. *Improving Domain Adaptation By Source Selection*, in "ICIP 2019 - IEEE International Conference on Image Processing", Taïpei, Taiwan, IEEE, September 2019 [*DOI :* 10.1109/ICIP.2019.8803325], https://hal.archives-ouvertes.fr/hal-02136378

[17] K. FAUVEL, V. MASSON, E. FROMONT, P. FAVERDIN, 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", Anchorage, United States, 25th SIGKDD Conference on Knowledge Discovery and Data Mining proceedings, August 2019, pp. 1-9 [*DOI :* 10.1145/3292500.3330712], https://hal.archives-ouvertes.fr/hal-02190790

[18] H. S. PHAM, G. VIRLET, D. LAVENIER, A. TERMIER. *Statistically Significant Discriminative Patterns Searching*, in "DaWaK 2019 - 21st International Conference on Big Data Analytics and Knowledge Discovery", Linz, Austria, Springer, August 2019, pp. 105-115 [*DOI :* 10.1007/978-3-030-27520-4_8], https://hal.archives-ouvertes.fr/hal-02190793

[19] Y. WANG, R. EMONET, E. FROMONT, S. MALINOWSKI, E. MENAGER, L. MOSSER, R. TAVENARD. *Classification de séries temporelles basée sur des "shapelets" interprétables par réseaux de neurones antagonistes*, in "CAp 2019 - Conférence sur l'Apprentissage automatique", Toulouse, France, July 2019, pp. 1-2, https://hal.archives-ouvertes.fr/hal-02268004

Scientific Books (or Scientific Book chapters)

[20] Y. DAUXAIS, D. GROSS-AMBLARD, T. GUYET, A. HAPPE. *Discriminant chronicle mining*, in "Advances in Knowledge Discovery and Management (vol 8)", B. PINAUD, F. GUILLET, F. GANDON, C. LARGERON (editors), Advances in Knowledge Discovery and Management, Springer, Cham, June 2019, pp. 89–118 [*DOI :* 10.1007/978-3-030-18129-1_5], https://hal.inria.fr/hal-01940146

Other Publications

[21] K. FAUVEL, D. BALOUEK-THOMERT, D. MELGAR, P. SILVA, A. SIMONET, G. ANTONIU, A. COSTAN, V. MASSON, M. PARASHAR, I. RODERO, A. TERMIER. *A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning*, November 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02373429

References in notes

[22] P. CLARK, T. NIBLETT. *The CN2 Induction Algorithm*, in "Mach. Learn.", March 1989, vol. 3, n^{o} 4, pp. 261–283, https://doi.org/10.1023/A:1022641700528

[23] S. COLAS, C. COLLIN, P. PIRIOU, M. ZUREIK. *Association between total hip replacement characteristics and 3-year prosthetic survivorship: A population-based study*, in "JAMA Surgery", 2015, vol. 150, n^{o} 10, pp. 979–988

[24] Y. DAUXAIS, D. GROSS-AMBLARD, T. GUYET, A. HAPPE. *Extraction de chroniques discriminantes*, in "Extraction et Gestion des Connaissances (EGC)", Grenoble, France, January 2017, https://hal.inria.fr/hal-01413473

[25] Y. DAUXAIS, T. GUYET, D. GROSS-AMBLARD, A. HAPPE. *Discriminant chronicles mining: Application to care pathways analytics*, in "Artificial Intelligence in Medicine", Vienna, Austria, 16th Conference on Artificial Intelligence in Medicine, June 2017 [*DOI :* 10.1007/978-3-319-59758-46], https://hal.archives-ouvertes.fr/hal-01568929

[26] M. FEURER, A. KLEIN, K. EGGENSPERGER, J. SPRINGENBERG, M. BLUM, F. HUTTER. *Efficient and Robust Automated Machine Learning*, in "Advances in Neural Information Processing Systems 28", C. CORTES, N. D. LAWRENCE, D. D. LEE, M. SUGIYAMA, R. GARNETT (editors), Curran Associates, Inc., 2015, pp. 2962–2970, http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf

[27] R. GUIDOTTI, A. MONREALE, S. RUGGIERI, F. TURINI, F. GIANNOTTI, D. PEDRESCHI. *A Survey of Methods for Explaining Black Box Models*, in "ACM Computing Surveys", 2018, vol. 51, n^{o} 5, pp. 93:1–93:42

[28] T. KULESZA, S. STUMPF, M. M. BURNETT, S. YANG, I. KWAN, 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

[29] G. MOULIS, M. LAPEYRE-MESTRE, A. PALMARO, G. PUGNET, J.-L. MONTASTRUC, L. SAILLER. *French health insurance databases: What interest for medical research?*, in "La Revue de Médecine Interne", 2015, vol. 36, n^{o} 6, pp. 411 - 417

[30] E. NOWAK, A. HAPPE, J. BOUGET, F. PAILLARD, C. VIGNEAU, P.-Y. SCARABIN, E. OGER. *Safety of Fixed Dose of Antihypertensive Drug Combinations Compared to (Single Pill) Free-Combinations: A Nested Matched Case–Control Analysis*, in "Medicine", 2015, vol. 94, n^{o} 49, e2229 p.

[31] E. POLARD, E. NOWAK, A. HAPPE, A. BIRABEN, 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", 2015, vol. 24, n^{o} 11, pp. 1161–1169