Activity Report 2021

Team LACODAM

Large scale Collaborative Data Mining

Joint team with Inria Rennes – Bretagne Atlantique

D7 – Data and Knowledge Management

Contents

Project-Team LACODAM

Creation of the Project-Team: 2017 November 01

Keywords

Computer sciences and digital sciences

A2.1.5. – Constraint programming

A3.1.1. – Modeling, representation

A3.1.6. – Query optimization

A3.1.11. – Structured data

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.3. – Data and knowledge analysis

A3.3.1. – On-line analytical processing

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A5.2. – Data visualization

A9.1. – Knowledge

A9.2. – Machine learning

A9.3. – Signal analysis

A9.5. – Robotics

A9.6. – Decision support

A9.7. – AI algorithmics

A9.8. – Reasoning

Other research topics and application domains

B1.2.2. – Cognitive science

B2.3. – Epidemiology

B2.4.1. – Pharmaco kinetics and dynamics

B3.5. – Agronomy

B3.6. – Ecology

B3.6.1. – Biodiversity

1 Team members, visitors, external collaborators

Research Scientist

- Luis Galarraga Del Prado [Inria, Researcher]

Faculty Members

- Alexandre Termier [Team leader, Univ de Rennes I, Professor, HDR]
- Johanne Bakalara [Univ de Rennes I, from Sep 2021]
- Tassadit Bouadi [Univ de Rennes I, Associate Professor]
- Peggy Cellier [INSA Rennes, Associate Professor, from Jul 2021, HDR]
- Elisa Fromont [Univ de Rennes I, Professor, HDR]
- Thomas Guyet [Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, Associate Professor, until Aug 2021, HDR]
- Christine Largouët [Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, Associate Professor, HDR]
- Véronique Masson [Univ de Rennes I, Associate Professor]
- Laurence Rozé [INSA Rennes, Associate Professor]

Post-Doctoral Fellow

- Neetu Kushwaha [Inria, until Feb 2021]

PhD Students

- Abderaouf Nassim Amalou [Univ de Rennes I, With PACAP Team]
- Johanne Bakalara [Univ de Rennes I, until Aug 2021]
- Lenaig Cornanguer [Inria]
- Julien Delaunay [Inria]
- Samuel Felton [Univ de Rennes I, With RAINBOW Team]
- Olivier Gauriau [ACTA Association, CIFRE]
- Camille Sovanneary Gauthier [Louis Vuitton]
- Elodie Germani [Univ Rennes I, With EMPENN Team]
- Victor Guyomard [Orange Labs, CIFRE]
- Gregory Martin [Groupe PSA]
- Josie Signe [Inria]
- Antonin Voyez [ENEDIS, CIFRE]
- Yichang Wang [China Scholarship Council, until Feb 2021]
- Heng Zhang [Atermes, CIFRE, until Nov 2021]

Technical Staff

- Remi Adon [Inria, Engineer, until Jul 2021]
- Louis Bonneau de Beaufort [Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, Engineer]
- Yichang Wang [Univ de Rennes I, Engineer, from Mar 2021 until Jul 2021]

Interns and Apprentices

- Aymen Bazouzi [Inria, from Apr 2021 until Sep 2021]
- Lola Beuchee [Univ de Rennes I, from May 2021 until Aug 2021]
- Ezanin Bile [Inria, from Apr 2021 until Oct 2021]
- Antoine Cellier [STIME (entreprise), Intern, from Mar 2021 until Sept 2021]
- Mathieu Chambard [École normale supérieure de Rennes, from May 2021 until Jul 2021]
- Jacques Lacourt [École centrale de Marseille, from Feb 2021 until Mar 2021]
- Guillaume Latour [Inria, from Feb 2021 until May 2021]
- Thomas Leguay [Univ de Rennes I, from May 2021 until Aug 2021]
- Simon Mattens [Inria, from Feb 2021 until May 2021]
- Mohammed Qureshi [Inria, from Mar 2021 until Aug 2021]
- Laurent Spillemaecker [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, from May 2021 until Aug 2021]

Administrative Assistant

- Gaelle Tworkowski [Inria]

Visiting Scientists

- Bruno Cremilleux [Univ de Caen Basse-Normandie, from Apr 2021 until Oct 2021]
- Yichang Wang [NC, from Aug 2021 until Sep 2021]

External Collaborators

- Philippe Besnard [CNRS, HDR]
- Yazid Boumarafi [Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement, From 09/2021]
- Romaric Gaudel [École nationale de la statistique et de l'analyse de l'information]

2 Overall objectives

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

The objective of the LACODAM is to facilitate the process of making sense out of (large) amounts of data. This can serve the purpose of deriving knowledge and insights for better decision-making. Our approaches are mostly dedicated to provide novel tools to data scientists, that can either perform 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

LACODAM is a research team on data science methods and applications, composed of researchers with a background in symbolic AI, data mining, databases, and machine learning. For this year’s activity report, we have updated our presentation of the research axes of the team so that they correspond better to our work. This new presentation of the team is organized into three research axes:

- **Symbolic methods** (Section 3.2) is the first fundamental research axis on methods that operate in symbolic domains, that usually take as input discrete data (ex: event logs, transactional data, RDF data) and output symbolic results (ex: patterns, concepts).
- **Interpretable Machine Learning** (Section 3.3) is the other fundamental research axis of the team. It aims at providing interpretable machine learning approaches, mostly by proposing *post-hoc interpretability* for state-of-the-art numerical machine learning methods. *Interpretable by design* machine learning approaches that do not fall into the "Symbolic methods" axis are also studied here.
- **Real world AI** (Section 3.4) deals with the application or adaptation of the methods developed in the aforementioned fundamental axes to real world problems. These works are conducted in collaboration with either industrial or academic partners from other domains. For example, one important application area for the team is numerical agriculture with colleagues from Inrae.

3.2 Symbolic methods

LACODAM’s core symbolic expertise is in methods for exploring efficiently large combinatorial spaces. Such expertise is used in three main research areas:

- Pattern mining, a field of data mining where the goal is to find regularities in data (in an unsupervised way);
- Semantic web, where the goal is to reason over the contents of the Web;
- Skyline queries, where the goal is to find solutions to multiple criteria optimization queries.

In the pattern mining domain, the team is well known for tackling problems where the data and expected patterns have a temporal components. Usually the data considered are timestamped event logs, an ubiquitous type of data nowadays. The patterns extracted can be more or less complex subsequences, but also patterns exhibiting temporal periodicity.

A well-known problem in pattern mining is pattern explosion: due to either underspecified constraints or the combinatorial nature of the search space, pattern mining approaches may produce millions of patterns of mixed interest. The current best approach to limit the number of output patterns is to produce a small size *pattern set*, where the set optimizes some quality criteria. The best pattern set methods so far are based on information theory and rely on the principle of Minimum Description Length (MDL). LACODAM is the leading French team on MDL-based pattern mining, especially for complex patterns. This situation has been strengthened by the formal integration of Peggy Cellier in the team, with whom the team has a long history of collaboration, especially on MDL-based pattern mining. On this subject she brings her expertise on (MDL-based) graph mining on which she works with Sebastien Ferré, who will join the team officially in April 2022.

The contribution of the team in the Semantic Web domain focuses on different problems related to knowledge graphs (KGs) – usually extracted (semi-)automatically from the Web. These include applications such as mining and reasoning, and data management tasks such as provenance and archiving. Reasoning can resort to either symbolic methods such as Horn rules or numeric approaches such as KG embeddings that can be explained via post-hoc explainability modules. The integration of Sébastien Ferré in 2022 will reinforce the SW axis by extending our expertise on general graph mining, relation extraction, and semantic data exploration.

Skyline queries is a research topic from the database community, and is closely related to multicriteria optimization. In transactionnal data, one may want to optimize over several different attributes of equal importance, which means discovering a Pareto Front (the "skyline"). The team has expertise on skyline queries in traditionnal databases as well as applied to pattern mining (extraction of *skypatterns*). Recently, the team started to tackle the extraction of skyline *groups*, i.e. groups of records that together optimize multiple criteria.

3.3 Interpretable ML

Making Machine Learning more interpretable is one of the greatest challenges for the AI community nowadays. LACODAM contributes to the main areas of explainable AI (XAI):

- From a fundamental point of view, the team is trying to deepen the understanding of state-of-the-art post-hoc interpretability approaches (LIME/SHAP), in order to improve these methods or adapt them to novel domains. The team has also started working on the generation of counterfactual explanations. Both lines of work have in common the need for novel notions of neighborhood of points in the model's data space.

- The team is also working on "interpretable-by-design" machine learning methods, where the decision taken can immediately be explained by the (part of) the model that took the decision. Approaches used can as well be deep learning architectures or hybrid numeric/symbolic models relying on pattern mining techniques.

- Last, the team has a special interest in time series data, which arises in many applications but has not yet received enough attention from the interpretability community. We have proposed both post-hoc and "by design" approaches for interpretable ML for time series.

More generally, LACODAM is interested in 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—a challenge that has very recently been addressed, not yet overcome.

3.4 Real world AI

LACODAM's research work is firmly rooted in applications: on the one hand the data science tools proposed in our fundamental work need to prove their value at solving actual problems. And on the other hand, working with practitioners allows us to understand better their needs and the limitations of existing approaches w.r.t. those needs. This can open new and fruitful (fundamental) research directions.

Our objective, in that axis, is to work on challenging problems with interesting and pertinent partners. We target problems where off-the-shelf data science approaches either cannot be applied or do not give satisfactory results: such problems are the most likely to lead to new and meaningful research in our field. For some problems, collaborative research may not necessarily lead to fundamental breakthroughs, but can still allow making progress in the practitioners' field. We also value such work, which contributes to the discovery of new knowledge and helps industrial partners innovate.

Due to the team expertise in handling temporal data, a lot of our applicative collaborations revolve around the analysis of time series or event logs. Naturally, our work on interpretability is also present in most of our collaborations, as experts want accurate models, but also want to understand the decisions of those models.

The precise application domains are described in more details in the next section (Section 4).

4 Application domains

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

4.1 Industry

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

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

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

- **Recommender Systems.** The CIFRE PhD of Camille-Sovanneary Gauthier at Louis Vuitton is concerned with the identification of the right click behavioral models for clients in order to optimize the arrangement of the items presented to potential customers in Web pages. This work builds upon new bandit algorithms to infer the parameters that model the customers' behavioral patterns accurately.

4.2 Health

- **Care Pathways Analysis for Supporting Pharmaco-Epidemiological Studies**. Pharmaco-epidemiology applies the methodologies developed in general epidemiology to answer to questions about the uses and effects of health products, drugs [44, 43] or medical devices [40], 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 [1] 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 [42]. Contrary to local hospital EHR or even to other national initiatives, the SNDS data warehouse covers a huge population. It makes possible studies on unfrequent drugs or diseases in real conditions of use. To tackle the volume and the diversity of the SNDS data warehouse, a research program has been established to design an innovative toolbox. This research program is focused first on the modeling of care pathways from the SNDS database and, second, on the design of tools supporting the extraction of insights about massive and complex care pathways by clinicians. In such a database a care pathway is an individual sequence of drugs exposures, medical procedures and hospitalizations.

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

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

4.3 Robotics

- **Visual Servoing.** Visual servoing (VS) is the task of controlling a robot by means of a camera, and is a common way to provide instructions to robots nowadays. The PhD thesis of Simon Felton has for purpose the exploration of novel deep learning techniques, and unsupervised learning to

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

improve the quality of VS settings and reduce the amount of human work to provide training data to such systems. This project is joint work with the RAINBOW team (IRISA).

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 of our research is the combination of data flows learning methods with mechanistic nutritional models in order to predict daily feed intake of lactating sows. Raphaël Gauthier (co-supervized with both INRIA and INRAE directors) defended his PhD thesis (#DigitAg Grant) in March 2021. 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 develop a forecasting procedure supported 40 by unsupervised learning of consistent clusters, specifically designed to make one-day-ahead forecasts of sow feed intake during lactation. Efficiency issues of developed algorithms have been considered since the proposed software should work in real-time on the automated feeder. The decision support process involves 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.

  Prediction of the Dynamics of Crop Diseases. The PhD thesis of Olivier Gauriau focuses on the prediction of the dynamics of crop diseases by means of pattern-aided regression techniques. Such techniques are known to strike an interesting trade-off between accuracy and interpretability, which can help agronomers understand the best predictors of high disease incidence, and therefore optimize the usage of phytosanitary products. This project is funded by #DigitAg and the Ecophyto program and constitutes a collaboration with the ACTA of Toulouse and the INRAE.

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 [2] has contributed to this body of research with the design of a tool that allows students to select multiple courses and predict their academic performance based on historical academic data. The tool resorts to visualization and interpretable machine learning techniques, and is intended to be used by the students before the counseling

[2] Centro de Tecnologías de Información

sessions to plan their upcoming semester at the Ecuadorian university ESPOL. In our ongoing collaboration with CTI we are studying the impact of academic predictions, explanations in the behavior and decision of the students and counselors.

- **Online Children Handwriting Recognition.** The PhD thesis of Simon Corbillé adresses the problem of online handwriting recognition, a problem that enjoys satisfactory solutions for adults, but remains a challenge for children. This is because, children handwriting is, at an early stage of learning, approximate and includes deformed letters. This is a joint effort between the LACODAM and IntuiDoc (IRISA) teams.

4.6 Semantic Data Management

- **RDF Archiving and Provenance.** Archiving and provenance tracking are two crucial tasks in the management of large collaborative RDF knowledge bases, such as Wikidata or DBpedia. This is a consequence of the dynamicity and source heterogeinity of such data collections. Notwithstanding the value of RDF archiving and provenance tracking for both data maintainers and consumers, this field of research remains under-developed for multiple reasons. These include, among others, the lack of usability and scalability of the existing systems, a disregard of the evolution patterns of RDF datasets, and a weaker focus on data processes involving non-monotone operations[3]. These challenges are tackled in our ongoing collaboration with the DAISY team of Aalborg University, namely thanks the PhD thesis of Olivier Pelgrin on scalable RDF archiving, and the post-doctoral fellowship of Daniel Hernández on how-provenance computation for SPARQL queries.

5 Social and environmental responsibility

5.1 Footprint of research activities

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

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

Utilisation of computing resources. LACODAM contributed last year with a new server (abacus12) to the Igrida computing platform. Being a team specialized in data science and machine learning, a recurrent task in LACODAM is to run CPU intensive algorithms on large data collections, for example, to train deep neural networks. Some of our ongoing PhD research topics (e.g., the theses of Heng Zhang, Simon Corbillé, and Simon Felton) concern deep learning technologies, and the increasing place of eXplanaible AI in our research program will boost our reliance on Igrida (notably with the PhD of Julien Delaunay and Victor Guyomard). This will increase the energetic and environmental footprint of our activities in a non-negligible way. We are therefore willing to collaborate with the institute's direction in any initiative that could mitigate such an impact.

5.2 Impact of research results

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

- In the short/medium term, a significant part of our research work is conducted in collaboration with companies, through CIFRE PhDs. Hence, the research problems addressed concern an important challenge for the company, and the solutions proposed are evaluated on their relevance to tackle this challenge. For example, the CIFRE PhD of Colin Leverger (defended last year), funded by

[3] Processes where there is some sort of data difference

Orange, was concerned with forecasting seasonal time series evolution, with application to the server load of the company. Its results could help better provision server resources during the year. Colin was hired by Orange after his PhD.

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

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

6 Highlights of the year

The main highlight of this year was the integration of Peggy Cellier in the team in June. Peggy has for long been a strong collaborator of the team on pattern mining approaches, especially the approaches based on the MDL (Minimum Description Length) principle. Her integration is thus natural, and will secure the leadership of Lacodam in this domain for the years to come.

Conversely, Thomas Guyet was awarded a “detachement Inria” and left the team, he is now working as an Inria researcher at Inria Lyon (Beagle team).

From a scientific result point of view, the team is especially proud that Camille-Sovanneary Gauthier, Romaric Gaudel and Elisa Fromont, together with their collaborator XXX, got a paper accepted in the prestigious International Conference on Machine Learning (ICML) [15].

7 New software and platforms

In this section we describe the new software productions of the team.

7.1 New software

7.1.1 HIPAR

Name: 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.

URL: `https://gitlab.inria.fr/lgalarra/hipar`

Contact: Luis Galarraga Del Prado

7.1.2 SNDS Data Generator

Name: SNDS Synthetic Data Generator

Keywords: Databases, Benchmarking

Functional Description: The SNDS, formerly SNIIRAM, is a huge database (consisting of several TBs of data and about 700 tables) that contains information about healthcare reimbursements of about 60 million French insured patients. This database is used to carry out epidemiological and medical-economic studies. Due to its sensitive medical content, identifying information (names, social security numbers) is removed or replaced by spurious information. This repository contains a solution to generate a synthetic version of the database.

The software generates relational data compliant with the original database schema with realistic distributions. This guarantees privacy preservation thanks to the use of open data only.

This work is currently under review to the AIME conference.

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

Contact: Thomas Guyet

7.1.3 SIERRA

Name: Seasonal tIme sERies foRecaster Api

Keywords: Time Series, Forecasting

Functional Description: SIERRA is a software dedicated to forecasting of seasonal time series. It includes a set of prediction models that are trained from the history of a time series and that can be applied to new data. The input of these models is a seasonal time series. It predicts its evolution over one or several seasons ahead. Models can also be used to describe historical data by providing access to characteristics about typical seasons in the dataset. Two main categories of models are proposed: deterministic models that predict the exact evolution of a time series, and probabilistic models that predict a probability distribution over time.

The SIERRA software is an API (Applied Programming Interface) which offers functionalities in the form of python module that can be embedded in a data analysis chain. The methods are accompanied by additional tools for the identification of seasonality, for the identification of the best configurations and the visualisation of the (probabilistic) predictions.

News of the Year: During 2021, a version of the SIERRA software has been released. This version has been sell to Orange SA which is now in charge of doing the official APP deposit. Orange SA is currently continuing the development of the sierra software based on this release.

Publication: hal-02371221

Contact: Thomas Guyet

Participants: Thomas Guyet, Colin Leverger

Partner: Orange Labs

8 New results

We organize the scientific results of the research conducted at LACODAM according to the axes described in our research program (Section 3).

8.1 Symbolic Methods

8.1.1 Pattern Mining

Regarding this research sub-axis, our contributions this year cover sequential or timestamped data for the sake of predicting future events, efficient enumeration of skyline groups, and subgroup discovery.

Efficient Enumeration of Skyline Groups. Skyline queries are multicriteria queries that are of great interest for decision applications. Skyline Groups extend the idea of skyline to groups of objects. In the recent years, several algorithms have been proposed to extract, in an efficient way, the complete set of skyline groups. Due to the novelty of the skyline group concept, these algorithms use custom enumeration strategies. [22] highlights the observation that a skyline group corresponds to the notion of ideal of a partially ordered set. From this observation, the authors of [22] propose a novel and efficient algorithm for the enumeration of all ideals of a given size k (i.e. all skyline groups of size k) of a poset. This algorithm, called GenIdeals, has a time delay complexity of $O(w^2)$, where w is the width of the poset, which improves the best known time output complexity for this problem: $O(n^3)$ where n is the number of elements in the poset. This work present new theoretical results and applications on skyline queries.

Novel Distance Metric for Evidential Preferences. In [9], the authors focus on measuring the dissimilarity between preferences with uncertainty and imprecision, modelled by evidential preferences based on the theory of belief functions. Two issues are targeted: The first concerns the conflicting interpretations of incomparability, leading to a lack of consensus within the preference modelling community. This discord affects the value settings of dissimilarity measures between preference relations. After reviewing the state of the art, [9] proposes to distinguish between two cases: indecisive and undecided, respectively modelled by a binary relation and union of all relations. The second concerns a flaw that becomes apparent when measuring the dissimilarity in the theory of belief functions. Existing dissimilarity functions in the theory of belief functions are not suitable for evidential preferences, because they measure the dissimilarity between preference relations as being identical. This is counter-intuitive and conflicting with almost all the related works. This work proposes a novel distance named Unequal Singleton Pair (USP) distance, able to discriminate specific singletons from others when measuring the dissimilarity. The advantages of USP distances are illustrated by the evidential preference aggregation and group decision-making applications. The experiments show that USP distance effectively improves the quality of decision results.

MDL-Based Sequential Mining. In [20], we propose COSSU, an algorithm to mine compact sets of sequential rules from long sequences of symbols. A sequential rule is an event of the form $A \Rightarrow B$ where A and B are sequences of events and B is seen to occur after A. Such rules can be used in a handful of applications such as next-element prediction and classification. Alas, mining those rules is prone to pattern explosion, which diminishes their applicability in real settings. COSSU tackles this challenge by relying on the principle of Minimum Description Length that resorts to the analogy of compression to choose a small set of rules that compresses (describes) the long sequence efficiently. Our experimental evaluation show that the rules found by COSSU can be successfully used for predicting future elements in a life-log, for text classification, and in general for drawing human-readable insights from long sequences of events.

Timed Automata Learning. The work in [11] proposes the TAG[4] Algorithm that learns timed automata from log event sequences. Those automata are behavioral models that describe a system, e.g., a biological ecosystem, an industrial system, etc., in terms of a set of (potentially recurring) events. TAG outperforms the state of the art by learning representations that are more precise and can therefore predict future events with higher accuracy.

Exceptional Model Mining for Time Series. Exceptional model mining task finds interesting subgroups according to several target attributes in tabular data. In [27] this approach is extended to time series data,

[4] Timed Automata Generator

using time series as a target attribute to evaluate subgroups. A subgroup is characterized by a description based on a model of time series computed from all series of the subgroup. Evaluation of subgroup quality is based on the difference between the subgroup model and the model of the whole dataset.

8.1.2 Semantic Web

Our contributions to the domain of Semantic Web cover different topics of interest for knowledge producers, namely archiving and provenance management, as well as one application for data consumers on epidemiology.

How-provenance for SPARQL Queries. The authors of [4] propose SPARQLprov, a method to compute how-provenance explanations for SPARQL query results. These explanations are semiring polynomials that convey the records (triples in the RDF model) and the operations that participate in the computation of a given query result. These annotations are of great value to a plethora of applications, e.g., dynamic updates, fact checking, etc. To compute them, SPARQLprov applies query rewriting techniques applicable to any standard triple store with a SPARQL interface. Contrary to state-of-the-art approaches, SPARQLprov method does not require any specialized engine, and can compute sound polynomial annotations for non-monotone queries, e.g., queries with operators involving some sort of difference. The experimental results show that the overhead induced by the rewriting remains reasonable, which makes SPARQLprov a viable solution for applications that require provenance tracking for queries.

RDF Archiving. The work in [8] provides a survey of the existing approaches for archiving the history of very large RDF knowledge graphs. The paper shows that among the existing solutions for RDF archiving, only one can be actually deployed on large RDF datasets. The survey studies the limitations of this solution and provides a framework to evaluate the evolution of large RDF datasets. This sets the ground for devising a set of design lessons and perspectives to address the problem of large-scale archiving in RDF. A first step towards this goal is materialized by the work in [18], where the authors propose TrieDF, an in-memory architecture to store and query metadata-augmented RDF in an efficient way. The experimental evaluation suggests that TrieDF's architecture is a promising solution for handling evolving RDF datasets with provenance and temporal (e.g., revision) metadata. Future work envision to combine this architecture with existing and/or novel in-disk storage approaches.

Chronicles and SPARQL for Querying Care Trajectories. Medico-administrative databases contain information about patients' medical events, i.e. their care trajectories. Semantic Web technologies are used by epidemiologists to query these databases in order to identify patients whose care trajectories conform to some criteria. In [19] the authors are interested in care trajectories involving temporal constraints. In such cases, Semantic Web tools lack computational efficiency while temporal pattern matching algorithms are efficient but lack of expressiveness. The authors therefore propose to use a temporal pattern called *chronicles* to represent temporal constraints on care trajectories. This work also proposes a hybrid approach, combining the expressiveness of SPARQL and the efficiency of chronicle recognition to query care trajectories. We evaluate our approach on synthetic data and real large data. The results show that the hybrid approach is more efficient than pure SPARQL, and validate the interest of the proposed tool to detect patients having venous thromboembolism disease in the French medico-administrative database.

8.2 Interpretable Machine Learning

Pattern-aided Regression. HiPaR[5] [33] is a pattern-aided regression method that mines accurate and compact sets of hybrid rules. These are rules of the form $p \Rightarrow y = f(X)$ where p is a set of conditions on categorical attributes and the right-hand side is a regression model that is only valid for the points that match p. HiPaR consists of two phases: an enumeration stage that explores the space of hybrid rules efficiently, and a selection phase that picks a small set of rules that guarantees both joint prediction accuracy and coverage. Our experimental evaluation shows that this method strikes a best trade-off

[5] Hierarchical Pattern-aided

between model simplicity (measured by the number of rules and their length) and prediction performance among existing interpretable and non-interpretable regression methods.

Explainable Multivariate Time Series (MTS) Classification. The current state-of-the-art MTS classifier is a heavyweight deep learning black-box approach that outperforms the second-best MTS classifier only on large datasets. To account for the need for explainability in MTS classification, [1] presents XCM, an eXplainable Convolutional neural network for MTS classification that extracts information relative to the observed variables and time directly from the input data. Thus, XCM architecture enables a good generalization ability on both large and small datasets, while allowing the full exploitation of a faithful post hoc model-specific explainability method (Gradient-weighted Class Activation Maps) by precisely identifying the observed variables and timestamps of the input data that are important for predictions. Experiments on real and synthetic data show the superiority of XCM in terms of performance and interpretability.

Performance-Explainability Benchmarking. [23] proposes a new performance-explainability analytical framework to assess and benchmark machine learning methods. The framework details a set of characteristics that systematize the performance-explainability assessment of existing machine learning methods. The framework has been applied to MTS classifiers.

User Impact of Performance Predictions for Term Planning. [13] offers a user study on the impact of showing academic performance predictions to students using the iCoRA term planning tool [39]. This tool allows students to plan the courses to be taken the next semester, and provides ML-based recommendations in the form of GPA performance predictions for each course based on different factors, e.g., workload, course stringency factors, the student's academic history, etc. The study sheds light on the impact of showing those predictions according to three criteria: behavior, decisions, and preferences. Moreover the article studies the effect of the visual representation used to convey the predictions, as well as the effect of providing global GPA predictions and feature-attribution explanations for the system's outcomes. The study shows, among other things, that users praise simplicity and credibility for predictions, and that specific visual representation (e.g., value ranges, punctual predictions) encourages the users to invest more time and effort in the task.

8.3 Real World AI

Our contributions in this axis cover a wide spectrum of applications from robotics, computer vision, logistics, forecasting, privacy, and recommender systems to precision agriculture, epidemiology, and natural environment.

8.3.1 Computer Vision and Robotics

Visual Servoing. Visual servoing (VS) is a common way in robotics to control a robot motion using information acquired by a camera. This approach requires to extract visual information from the image to design the control law. The resulting servo loop is built in order to minimize an error expressed in the image space. In [2, 12], the authors consider a direct visual servoing (DVS) from whole images. They propose a new framework to perform VS in the latent space learned by a convolutional autoencoder. It is shown that this latent space avoids explicit feature extraction and tracking issues and provides a good representation, smoothing the cost function of the VS process. Besides, the experiments show that this unsupervised learning approach allows us to obtain, without labelling cost, an accurate end-positioning, often on par with the best DVS methods in terms of accuracy but with a larger convergence area.

Deep Active Learning on Multispectral Data. Data from multiple sensors provide independent and complementary information, which may improve the robustness and reliability of scene analysis applications. While there exist many large-scale labelled benchmarks acquired by a single sensor, collecting labelled multi-sensor data is more expensive and time-consuming. In [29], the authors explore the construction of an accurate multispectral (here, visible and thermal cameras) scene analysis system with minimal annotation efforts via an active learning strategy based on the cross-modality prediction inconsistency. Experiments on multiple multispectral datasets and vision tasks demonstrate the effectiveness of the method. In particular, with only 10% of labelled data on KAIST multispectral pedestrian detection dataset, the system obtains comparable performance as other fully supervised state-of-the-Art methods.

Feature Fusion on Multispectral Data. Multispectral image pairs can provide complementary visual information, making pedestrian detection systems more robust and reliable. To benefit from both RGB and thermal IR modalities, [30] introduces a novel attentive multispectral feature fusion approach. Under the guidance of the inter- and intra-modality attention modules, the proposed deep learning architecture learns to dynamically weigh and fuse the multispectral features. Experiments on two public multispectral object detection datasets demonstrate that the proposed approach significantly improves the detection accuracy at a low computation cost.

Multispectral Scene Analysis with Modality Distillation. Despite its robust performance under various illumination conditions, multispectral scene analysis has not been widely deployed due to two strong practical limitations: 1) thermal cameras, especially high-resolution ones are much more expensive than conventional visible cameras; 2) the most commonly adopted multispectral architectures, twostream neural networks, nearly double the inference time of a regular mono-spectral model which makes them impractical in embedded environments. In [31], the authors aim to tackle these two limitations by proposing a novel knowledge distillation framework named Modality Distillation (MD). The proposed framework distills the knowledge from a high thermal resolution two-stream network with feature level fusion to a low thermal resolution one-stream network with image-level fusion. It is shown on different multispectral scene analysis benchmarks that the proposed method can effectively allow the use of low-resolution thermal sensors with more compact one-stream networks.

Knowledge distillation still faces some challenges: Due to the extreme imbalance between the foreground and the background of images, when traditional knowledge distillation methods are directly applied to the object detection task, there is a large performance gap between the teacher model and the student model. The work in [32] tackles this imbalance problem from a sampling perspective, and proposes to include the teacher-student prediction disagreements into a feature-based detection distillation framework, called PDF-Distil. This is done by dynamically generating a weighting mask applied to the knowledge distillation loss, based on the disagreements between the predictions of both models. Extensive experiments on PASCAL VOC and MS COCO datasets demonstrate that, compared to state-of-the-art methods, PDF-Distil is able to better reduce the performance gap between the teacher and student models.

8.3.2 Recommender Systems

Bandits for Optimal Item Display on Web Pages. Multiple-play bandits aim at displaying relevant items at relevant positions on a web page. [14] introduces a new bandit-based algorithm, PB-MHB, for online recommender systems which uses the Thompson sampling framework with Metropolis-Hastings approximation. This algorithm handles a display setting governed by the position-based model. Our sampling method does not require as input the probability of a user to look at a given position in the web page, which is difficult to obtain in some applications. Experiments on simulated and real datasets show that PB-MHB, with fewer prior information, delivers better recommendations than state-of-the-art algorithms.

Parametric Graph for Unimodal Ranking Bandit. In [15, 24], the authors tackle the online ranking problem of assigning L items to K positions on a web page in order to maximize the number of user clicks. The authors propose an original algorithm, easy to implement and with strong theoretical guarantees to tackle this problem in the Position-Based Model (PBM) setting, well suited for applications where items are displayed on a grid. Besides learning to rank, the proposed algorithm, GRAB (for parametric Graph for unimodal RAnking Bandit), also learns the parameter of a compact graph over permutations of K items among *L*. The logarithmic regret bound of this algorithm is a direct consequence of the unimodality property of the bandit setting with respect to the learned graph. Experiments against state-ofthe-art learning algorithms which also tackle the PBM setting, show that GRAB is more efficient while giving regret performance on par with the best known algorithms on simulated and real life datasets.

8.3.3 Forecasting

Seasonal Time-series. In [17] the authors propose a framework for seasonal time series probabilistic forecasting. It aims at forecasting the whole next season of a time series, rather than only the next value. Probabilistic forecasting consists in forecasting a probability distribution function for each future position. The proposed framework is implemented combining several machine learning techniques 1) to identify typical seasons and 2) to forecast a probability distribution of the next season. This framework is evaluated using a wide range of real seasonal time series. On the one side, we intensively study the alternative combinations of the algorithms composing our framework (clustering, classification), and on the other side, we evaluate the framework forecasting accuracy. As demonstrated by our experiences, the proposed framework outperforms competing approaches by achieving lower forecasting errors.

8.3.4 Logistics

Fleet Relocation. The success of a free-floating car-sharing service depends on a good allocation of the vehicles across the city, i.e. where and when they are needed by citizens. This requires predicting the demand across the geographical regions and across time, which is challenging due to the sparsity and variability of the data. Furthermore, the purpose of these predictions is to help computing the best possible car positions for the next day, hence the need to model both the prediction task and the optimisation task in a compatible way. As the allocation optimisation involves reasoning about the number of cars to assign to geographical regions, [26] proposes to predict the expected utilisation of a car when added to a region. The authors discuss the challenges in modeling both the machine learning and the relocation problem, and propose an integer linear programming method that solves the relocation problem while taking into account the model predictions and relocation distances. We experiment with the datasets from a citywide car sharing company and show how the proposed method can increase the allocation strategies and hence the profitability of the service.

8.4 Privacy

Interference Attacks on Aggregated Time Series. Aggregation is widely used as a privacy protection method. Aggregate membership inference attacks aim to determine whether or not a given target participated in the computation of the aggregate under attack. In [28], the authors study the vulnerability of aggregated time series—where each point is a time-stamped aggregate—to membership inference attacks. The considered attacker has auxiliary knowledge about a superset of the aggregated data (e.g., from a data leak). [28] proposes a new attack leveraging this type of auxiliary knowledge and the multiple points forming the aggregate time series. The attack is modeled as an integer linear optimization problem, allowing the attacker to benefit from the power of dedicated solvers (e.g., Gurobi). This attack, tested on public datasets, shows the vulnerability of an aggregated time series publication if the number of aggregated series is too small compared to the number of points constituting the series.

8.5 Medicine

Temporal Phenotyping applied to Care Pathways for COVID-19 patients. During the COVID19 crisis, Intensive Care Units admitted many patients with breathing disorders up to respiratory insufficiency. The care strategy of such patients was difficult to find and preventing patients to drift away toward a critical situation was one of the first challenges of physicians. In [21], the authors characterize care pathways of patients that required a mechanical ventilation (invasive treatment for the most critical respiratory insufficiencies). Through the analysis of the sequence of cares, the goal of this work is to support physicians in understanding the evolution of patients, and let them propose new medical procedures for preventing ventilation. This article proposes a method which combines a tensor factorization and sequence clustering. The tensor factorization enables to represent the care sequences as a sequence of daily phenotypes. Then, the sequences of phenotypes is clustered to extract typical care trajectories. This method is experimented on real data from Greater Paris university Hospital and is compared to a direct clustering of the sequences. The results show that the outputs are more easily interpretable with the proposed method.

Benchmarking for Healthcare Systems. The recent development of data analysis provides opportunities for improving healthcare systems through analysis of health databases. However this thirst for data conflicts with the preservation of the privacy of individuals. In that vibe, the generation of synthetic datasets may foster research on healthcare data analytics. Synthetic data is mostly based on generative statistical models fitted on real data. [25] proposes a probabilistic relational model fitted on publicly available datasets. More specifically, the authors propose to generate a synthetic version of the national database of French insured patients. The authors do not only provide synthetic datasets, but a generator of datasets that can be used without any data access request. Experiments compare official statistics with those computed on synthetic datasets and show the potential use of plausible synthetic data.

8.5.1 Agriculture

Future Improvements on Precision Feeding. Taking into account individual variability while feeding a group of sows allows feed cost reductions and therefore improves animal efficiency. This precision feeding strategy is based on 1) nutritional models, which are able to predict daily individual nutrient requirements; 2) automatons, that can deliver individual rations; and 3) new technologies such as sensors which provide real-time information on the animal performance and life conditions that should be integrated into the estimation of requirements. Up to now, only production data (body weight, backfat thickness) have been integrated into the calculation of individual nutrient requirements. However, the literature reported that health status and behavior, such as physical activity, social behavior, and location in the pen, can strongly influence nutrient requirements. A change in the feeding or drinking behavior can also indicate a health or welfare problem. Sensors, automatons and cameras are now able to detect some diseases or injuries, and record certain onfarm behaviors. Therefore, nutrient requirements should be adjusted based on these health and behavioral parameters. Environmental factors such as thermal conditions, housing type and noise level have also been reported to affect nutrient requirements. On-farm sensors can easily be installed to record these parameters to be integrated into the nutritional model and improve its precision. A decision support system can be used to integrate these new measurements into the nutritional model for gestating sows. It would also be helpful to trigger alerts and propose corrective actions when behavior changes or health issues are detected. All these perspectives are detailed in [3].

Offline Time-series Clustering for Precision Feeding. According to precision livestock farming principles, it is essential to apply feed intake forecasting processes to real time precision feeding strategies in order to improve the overall efficiency of the livestock feeding chain. Considering the lack of a mechanistic model that predicts daily feed intake in lactating sows, a novel approach combining an online forecasting procedure with an offline learning procedure is proposed in [5, 15]. The authors used a database of 39,090 lactations—from 6 different farms and containing the first 20 daily feed intake records after farrowing—(a) to identify consistent sets of clusters and trajectory curves offline, and (b) to test 3 predictive functions of daily feed intake online. Though variability in feed intake among sows and over the lactation period is high, online forecasting of feed intake can be improved by the use of feed intake trajectory curves. These trajectory curves may be computed on a regular basis with data obtained directly on the farm or on farms with similar practices. The online forecasting procedure requires few computing resources, and could easily be embedded in smart feeder control systems as a practical application in precision feeding systems for lactating sows.

8.6 Natural Environment

Fishing and River Resources. [6] studies the perceptions of fishers on the interactions between their practices and their environment. The study was performed in the upper section of the Maroni River in French Guiana, a relatively remote region in the tropical rainforest where subsistence fishing still occurs. We assessed the fishers' perceptions of their relationship with the river by asking them about the state of natural resources, their fishing practices, nearby activities, and their way of life. Cognitive mapping was used to capture their individual viewpoints, especially those that formed a consensus with the other fishers. Regardless of their ethnic group (Aluku vs. Amerindian) or way of life (subsistence vs. commercial fishing), most fishers generally shared the same views. The main perception was that fishing is threatened by illegal gold mining, increasing use of fishing nets, and a loss of knowledge of fish behavior by younger generations of villagers. Furthermore, fishers perceived an ongoing shift in their role and relationship with the fish resource, which is becoming increasingly commercially oriented, and since the river is no longer the only source of food. Detailed analysis of arguments put forward to explain these threats shows that this process originates from ill-managed Westernization, which has caused painful changes in lifestyles of local populations, especially Amerindians. This analysis can also provide local governments with mechanisms for action. Our results raise questions about the future of this region and suggest ways to protect its natural resources better. They can help decision makers respond to poorly understood informal fisheries and motivate local residents to contribute to sustainable management of a river's natural resources.

This work is complemented by a survey [7] that sheds light on the fishing practices in the Maroni River in French Guiana. The authors surveyed 754 boat landings in seven villages located in the upper half of the watershed, representing > 6300 fish during the study period (November 2013 - September 2014). Fishers used canoes with outboard engines almost exclusively (75 %) and fished within 32 km of their villages. Most fish were caught in trammel nets (81 %); the 20 most-landed species represented more than 87% of catches. Depending on the village, daily catches and biomass averaged 6–14 fish and 1.7–13 kg per boat landing, respectively. Seven control sites located outside of the fishing grounds were fished to identify potential differences in catch per unit effort and fish size. Per 100 m2 of trammel net, mean catches ranged from 4 to 13 and 8–29 fish in the villages and control sites, respectively, while fish biomass ranged from 0.9 to 4 and 3.2–7 kg in villages and control sites, respectively. For all species combined, fish caught at control sites were bigger than those landed in villages. This difference was significant for nine of the most-landed species. Differences in fishing techniques and fish catches between villages illustrated the gradual disappearance of the ancestral subsistence fishing. Our results support indications that the fish community in the upper Maroni River is harvested intensively, address the issue of sustainability of the fishery there, and call attention to the need to conserve the river's remarkable biodiversity.

Edge-to-Cloud Computing for Earthquake Detection. The growth of the Internet of Things is resulting in an explosion of data volumes at the Edge of the Internet. To reduce costs incurred due to data movement and centralized cloud-based processing, it is becoming increasingly important to process and analyze such data closer to the data sources. Exploiting Edge computing capabilities for stream-based processing is however challenging. It requires addressing the complex characteristics and constraints imposed by all the resources along the data path, as well as the large set of heterogeneous data processing and management frameworks. Consequently, the community needs tools that can facilitate the modeling of this complexity and can integrate the various components involved. The work in [10] introduces MDSC, a hierarchical approach for modeling distributed stream-based applications on Edge-to-Cloud continuum infrastructures. The article demonstrates how MDSC can be applied to a concrete real-life ML-based application—early earthquake warning—to help answer questions such as: when is it worth decentralizing the classification load from the Cloud to the Edge and how?

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

- **ORANGE - Univ. Rennes I**
  Participants: T. Bouadi, T. Guyet, V. Guyomard
  Contract amount: 30k€ + Phd Salary
  Context. This project is a collaboration with Orange Labs Lannion about interpretable machine learning. The Orange company aims to develop the use of machine learning algorithms to enhance

the services they proposed to their customers (for instance, credit acceptance or attribution prediction). It ensues the development of *generic approaches for providing interpretable decisions* to customers or client managers.
Objective. The GDPR, implemented by the EU in 2018, stipulates the right for explanations for EU citizens in regards to decisions made from personal data. In a society where many of those decisions are computer-assisted via machine learning algorithm, interpretable ML is crucial. A promising way to convey explanations for the outcomes of ML models are *counterfactual explanations.* The focus of the PhD thesis financed by this project is the generation of usable and actionable counterfactual explanations for ML classifiers, which are intensively used by Orange within their services.
Additional remarks. This contract finances the PhD of Victor GUYOMARD by Orange.

- **Louis Vuitton - Univ. Rennes I**
  Participants: R. Gaudel, E. Fromont, C. Sovanneary Gauthier
  Contract amount: 60k€ (shared with CREST)
  Context. Louis Vuitton is a French high-end luxury fashion house with a large catalog of products available for purchase on their site web via keyword research. Hence, pertinence of search results as well as useful recommendations are of paramount importance for their business.

  Objective. To tackle the particular recommendation use cases encountered at Louis Vuitton, we focus on the *online learning to rank* problem: we identify the right click behavioral models for clients and we develop new bandit algorithms to efficiently infer the parameters of such click behavioral models.
  Additional remarks. This contract finances the PhD of Camille-Sovanneary Gauthier.

- **Intermaché - Univ. Rennes I**
  Participants: T. Bouadi, P. Cellier, A. Termier
  Contract amount: around 3k€

  Although not an actual contract, we had a collaboration with Intermarché (Groupement des Mousquetaires) which directly financed the M2R internship of Antoine Cellier. In this work, we studied the analysis of retail data in order to propose methods for the early discovery of significant life changes (large home renovation projects, arrival of a baby for examples). We used subgroup discovery methods.

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

  Context. 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.

  Objective. Capacity planning requires accurate forecasts of the differences between the customer demands and the infrastructure theoretical capabilities. The HYPTSER project makes the assumption that this information is captured by key performance indicators (KPI), that are measured continuously in the service infrastructure. Thus, we expect to improve capacity planning capabilities by making accurate forecasts of KPI time series. Recent methods about time series forecasting make use of ensemble models. In this project, we are interested in developing hybrid models for time series forecasting. In the next steps of this project, we will analyze the performance of this two strategies on KPI time series provided by Orange and compare them to classical ensemble methods.

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

- **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.
  Additional remarks. This project has financed the PhD of Heng ZHANG who defended his thesis in December 2021.

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

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

  - A town modelisation module with clustering. Similar towns are clustered in order to reuse information from one town in other towns.
  - A travel prediction module with basic statistics.
  - A reallocation strategy module (choices on how to relocate cars so that the most requested areas are always served). The aim of this module is to be able to test different strategies.

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

10 Partnerships and cooperations

10.1 European initiatives

10.1.1 FP7 & H2020 projects

TAILOR (Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization) H2020-ICT-2019-3 Elisa Fromont, Luis Galárraga and Alexandre Termier are all involved in the European project in particular in WP3 about Trustworthy AI.

10.2 National initiatives

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

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

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

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

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

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

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

- **Bourse IUF - Elisa FROMONT**

10.2.1 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 [41], FAbLe aims at fully delegating the selection of interpretable explanations to computers. Our goal is to produce a suite of algorithms that will compute suitable explanations for ML algorithms based on our insights of what is interpretable. The algorithms will choose the best explanation method based on the data, the use case, and the user’s background. We will implement our algorithms so that they are fully compatible with the body of available software for data science (e.g., Scikit-learn).

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Elisa Fromont is General chair of the Symposium on Data Analysis (IDA 2022). All LACODAM (including Gaëlle Tworkowski, the project assistant) is involved in the organization of this event.
- Elisa Fromont is member of the steering committee of the “Conférence sur l’Apprentissage automatique” (CAp since 2020) and of the steering committee of the European Conference in Machine Learning and Knowledge Discovery (ECML PKDD) since 2019, and until the end of 2021.

11.1.2 Scientific events: selection

Chair of conference program committees

- Luis Galárraga was co-chair of the special session “Fair and Explainable Models”, in collaboration with Miguel Couceiro (LORIA), at the European Conference on Operational Research (EURO 2021).
- Tassadit Bouadi and Luis Galárraga were co-organizers of the workshop on Advances in Interpretable Machine Learning and Artificial (AIMLAI) co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2021).
- Tassadit Bouadi is Program Chair of the Symposium on Data Analysis (IDA 2022)
- Elisa Fromont is co-program chair of the “Conférence sur l’Apprentissage automatique” (CAP 2022) in Vannes.
- Peggy Cellier is co-Journal track chair of the European Conference in Machine Learning and Knowledge Discovery ECML PKDD 2022 in Grenoble

Member of the conference program committees

- Elisa Fromont was in the program committee of KDD 2021 ("ACM SIGKDD Conference on Knowledge Discovery and Data Mining"); ECMLPKDD 2021 ("European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases") as senior PC; AAAI 2021 ("AAAI Conference on Artificial Intelligence"); WACV 2021(IEEE’s and the PAMI-TC’s "Winter Conference on Applications of Computer Vision"); BMVC 2021 ("The British Machine Vision Conference"); IDA 2021 ("Intelligent Data Analysis").
- Peggy Cellier was in the program committee of ECMLPKDD 2021 ("European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases"); EGC 2021 ("Extraction et Gestion de la Connaissances") as senior PC; ICCS 2021 ("International Conferences on Conceptual Structures"); ICFCA 2021("International Conference on Formal Concept Analysis"); RealDataFCA 2021 (workshop at ICFCA); TALN 2021 ("Conférence sur le Traitement Automatique des Langues Naturelles").

- Luis Galárraga was in the program committee of ECML PKDD 2021 (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases); IJCAI 2021 (International Joint Conference on Artificial Intelligence); ISWC 2021 (International Semantic Web Conference); AKBC 2021 (Workshop on Automatic Knowledge Base Construction).
- Romaric Gaudel was in the program committee of NeurIPS 2021 ("Conference on Neural Information Processing Systems"); AISTATS 2021 ("International Conference on Artificial Intelligence and Statistics"); AAAI 2021 ("AAAI Conference on Artificial Intelligence"); AIMLAI 2021 ("workshop on Advances in Interpretable Machine Learning and Artificial Intelligence"); CAP 2021 ("Conférence sur l'Apprentissage automatique").
- Alexandre Termier was in the program comittee of KDD 2021 ("ACM SIGKDD Conference on Knowledge Discovery and Data Mining"); ECMLPKDD 2021 ("European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases"); AAAI 2021 ("AAAI Conference on Artificial Intelligence"); AISTATS 2021 ("International Conference on Artificial Intelligence and Statistics"); SDM 2021 ("SIAM International Conference on Data Mining"); ITICE 2021 ("International conference on Time Series and Forecasting"); EGC 2021 ("Extraction et Gestion des Connaissances").

11.1.3 Journal

Member of the editorial boards

- Elisa Fromont is Co-Specialty Chief Editor of Frontiers in Artificial Intelligence specialty Machine Learning and Artificial Intelligence.
- Elisa Fromont and Alexandre Termier are Members of the Editorial Board of Data Mining journal (DMKD).
- Elisa Fromont is a guest editor of the Mathematics journal for a special issue on Time Series Analysis.
- Peggy Cellier is a Member of the Editorial Board of ICFCA ("International Conference on Formal Concept Analysis").

Reviewer - reviewing activities

- Luis Galárraga served as a reviewer for the Semantic Web Journal, the Journal on Data Semantics, and the Tutorial Paper Track of the Artificial Intelligent Review.
- Luis Galárraga served as Expert reviewer for the ANR project call 2021.
- Romaric Gaudel served as a reviewer for Pattern Recognition, and Transactions on Information Systems.
- Christine Largouët served as a reviewer for BioSystems and Computational Biology.
- Alexandre Termier served as a reviewer for the Data Mining journal (DMKD) and IEEE Transactions on Knowledge and Data Engineering (TKDE).

11.1.4 Invited talks

Elisa Fromont did the following invited talks:

- 2/10/2021: Invited speaker (in French) for the "Journée Recherche du Labex DigiCosme" on the topic "Explainable Time Series Classification" (virtual).
- 05/10/2021: Keynote speaker (in French) for the Conférence sur l'Intelligence Artificielle organized by ESIR, ENSSAT et IGR-IAE Rennes.

- 29/09/2021: Invited speaker (in French) GDR IA fall school on "(Deep) neural networks: basic principles and explanation issues", Paris.
- 16/09/2021: Keynote speaker (in French) for the conference ORASIS (Journées francophones des jeunes chercheurs en vision par ordinateur) on the topic "Multispectral Object Detection" (virtual).
- 29/06/2021: Invited speaker (in French) for JSI (Journées Scientifiques INRIA) on the topic "Challenges in computer vision and a few contributions" (virtual).
- 21/06/2021: Invited speaker (in French) Colloque DRIAS (Droit, IA et Santé) "Apprentissage Machine Explicable", Rennes (virtual).
- 22/04/2021: Invited speaker (in French) Cycle IA & IMAGES du Breizh Data Club "Analyse de scènes extérieures par des méthodes de deep learning: problèmes de fusion de données, de déséquilibres de classes et d'adaptation de domaine", Rennes (virtual).
- 08/03/2021: Invited speaker for MIAI legality & inclusion day, Grenoble (virtual).

Christine Largouët did the following talk:

- 27/03/2021 Invited Talk : "Data vs. Concept driven models" in "Journée ATMOSPHASE", INRAE organized by AgroParisTech.

Alexandre Termier did the following invited talks:

- 25/11/2021: Invited speaker at the workshop "Explainable and Responsible AI" of the Multidisciplinary Institute in Artificial Intelligence of Grenoble-Alpes University, title: "HiPaR: Hierarchical Pattern-Aided Regression"
- 08/12/2021: Invited speaker at ENS Rennes seminar for 2nd year students, title: "Pattern Mining: a biased perspective".

11.1.5 Scientific expertise

- Elisa Fromont is a member of the scientific council of the GDR IA, the Machine Learning College at AFIA, and of the "Société Savante Francophone d'Apprentissage Machine" (SSFAM).
- Elisa Fromont was a member of a recruitement committee in Leuven, Belgium; Position MCF_27MCF0950 at (LIFO; University); Position PR_4641 at Nantes (LS2N, University).
- Alexandre Termier and Christine Largouët were members of an associate professor recruitement committee in Rennes (Institut Agro, Agrocampus Ouest).
- Peggy Cellier was member of an associate professor recruitement committee in Orléans (Université d'Orléans).
- Alexandre Termier was member of the recruitement committees for permanent researcher (CRCN) positions at Inria Renne and for the MCF_1261 associate professor position at ESIR, University of Rennes 1.

11.1.6 Research administration

- Since September 2018, Peggy Cellier is in charge of the Irisa Ph.D. students at IRISA, i.e. she is involved in the "commission du personnel" and organizes the selection of Ph.D. students for ministerial grants (contrats doctoraux). She is also an elected member of the "Conseil de Composante IRISA/INSA" at INSA and an elected member of the "Conseil de laboratoire" at IRISA. Peggy Cellier is "secrétaire" of "Revue de Traitement automatique des langues" since 2019.
- Since 2017, Elisa Fromont is elected at IRISA lab council (she is a member of the gender equality group and responsible for the anti-harassment group). She is elected, since 2020, at the scientific council of the University (UR1) and is a member of the HDR committee for the University. She is the head of the D7 scientific departement at IRISA since Sept 2021 (and part of the direction scienfic board).

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

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

- Alexandre Termier is responsible for the following courses at ISTIC (Univ. Rennes 1): Object Programming (L2 info, elec, maths), AI (M1 info), Data Mining and Visualization (M2 SIF).
- ENSAI's third year (M2 level): Machine Learning modules (Deep Learning, large-scale Learning, Recommender Systems), 21h, ENSAI (R. Gaudel)
- Master Smart Data: Bandits Theory et Recommender Systems, 9h, ENSAI (R. Gaudel)
- At INSA, Peggy Cellier is responsible of four courses: "Databases and web development" (Licence 3 INFO), "Databases" (Licence 3 Math), "Data Mining" (Licence 3) and "Advanced Database and Semantic Web" (Master 2). She also teaches some other courses: "Database" (Licence 2), "Use and functionalities of an operating system" (Licence 3). At master 2 SIF, she teaches in English 4 hours in the data mining course (DMV). In addition she gives a lecture of 2 hours also in master 2 SIF about "Qu'est-ce qu'une thèse, un doctorat, un·e doctorant·e ?".
- Participation in the module "Case Study in Data Science" with the seminar "Interpretable AI", M2 EIT DSC, ISTIC, Univ. Rennes I (L. Galárraga, 4h).

11.2.2 Supervision

Postdocts

- Alexandre Termier and Elisa Fromont were supervisors of the post-doc of **Neetu Kushwaha** (until February 2021) and the M2 Internship of **Ezanin Christian Bile** (Polytechnique, from April to September 2021)
- Luis Galárraga co-supervises the following post-doctoral fellows: **Daniel Hernández** working on how-provenance for SPARQL queries in collaboration with Katja Hose from Aalborg University, and **Mohit Mihal** post-doctoral fellow from the Sequel team at Inria Lille in collaboration with Philippe Preux.

PhD. Students

- **Elodie Germani**, 2021-2024; supervisors: Élisa Fromont and Camille Maumet; title: Metric learning for robust FMRI pipelines.
- **Abderaouf Nassim Amalou** (PhD, UR1/Projects) 2021-2024; supervisors: Élisa Fromont and Isabelle Puaut; title: Machine Learning for Timing Estimation.
- **Olivier Gauriau**, (Inria, DigitAg, Acta Toulouse) 2021-2024; supervisors: Alexandre Termier, Luis Galárraga, François Brun, and David Makowski; title: Numerical Rule Mining for the Prediction of the Dynamics of Crop Diseases.
- **Antonin Voyez**, (PhD, CIFRE Enedis) 2020-2023; supervisors: Élisa Fromont, Tristan Allard and Gildas Avoine; title: Privacy-preserving Power Consumption Time-series Publishing.
- **Julie Boudebs**, 2021-2024; supervisors: Peggy Cellier and Sébastien Ferré (SemLIS), title: Un assistant en langue naturelle pour interroger le Web sémantique, ED MathStic.
- **Simon Corbillé**, (PhD, UR1) 2019-2022; supervisors: Élisa Fromont and Eric Anquetil; title: Explainable Deep-learning-based Methods for Children Handwriting Analysis in Education.
- **Samuel Felton**, (PhD, UR1) 2019-2022; supervisors: Élisa Fromont and Erich Marchand; title: Deep Learning for End-to-end Visual Servoing.

- **Gregory Martin**, (PhD, CIFRE PSA) 2019-2022 supervisors: Élisa Fromont, Alexandre Termier, and Laurence Rozé; title: Data mining to Optimize a Free-floating Car Sharing Service.
- **Camille-Sovanneary Gauthier**, (PhD, CIFRE Vuitton) 2019-2022; supervisors: Romaric Gaudel and Élisa Fromont, title: Bandit-based Recommender Systems.
- **Heng Zhang**, (PhD, CIFRE Atermès) 2018-2021; supervisors: Élisa Fromont and Sébastien Lefèvre. Rémi Emonet, Romain Tavenard, and Simon Malinowski; title: Multispectral Object detection
- **Yichang Wang**, (PhD, CSC) 2018-2021; supervisors: Élisa Fromont, Rémi Emonet, Romain Tavenard and, Simon Malinowski; title: Interpretable Time Series Classification
- **Victor Guyomard**, 2020-2023; supervisors: Tassadit Bouadi, Thomas Guyet, Françoise Fessant (Orange Labs) and Alexandre Termier, title: Explaining individual decisions made by an AI algorithm.
- **Josie Signe**, 2020-2023; supervisors: Peggy Cellier, Yannick Le Cozler (Inrae), Véronique Masson and Alexandre Termier, title: Animal Welfare, Characterizing the Diversity between and within Livestock Farming Situations with Data Mining Methods used on Information from Dairy Herd Sensors, ED MathStic.
- **Lénaïg Cornanguer**, 2020-2023, supervisors: Christine Largouët, Alexandre Termier and Laurence Rozé; title: Timed Automata Learning, ED MathSTIC
- **Julien Delaunay**, (Inria, ANR) 2020-2023; supervisors: Christine Largouët and Luis Galárraga; title: Automatic Construction of Explanations for AI Models, ED MathSTIC.
- **Maëva Durand**, 2020-2023; supervisors: Christine Largouët, Charlotte Gaillard (INRAE) and Jean-Yves Dourmad (INRAE); title: Real-time Integration of Gestating Sow Welfare and Health from Heterogeneous Data for Precision Feeding, ED EGAAL.
- **Hugo Ayats**, 2020-2023; supervisors: Peggy Cellier and Sébastien Ferré (SemLIS), title: De la prédiction à l'automatisation avec une IA explicable et centrée-utilisateur – application à la construction de graphes de connaissances, ED MathStic.
- **Olivier Pelgrin**, (PhD, AAU) 2019-2022; supervisors: Katja Hose and Luis Galárraga; title: Fully-fledged Archiving for RDF Datasets.
- **Johanne Bakalara**, 2018-2021; supervisors: Thomas Guyet, E. Oger, O. Dameron, A. Happe; title: Temporal Models of Care Sequences for the Exploration of Medico-administrative Data.
- **Francesco Bariatti**, 2018-2021; supervisors: Peggy Cellier and Sébastien Ferré (SemLIS); title: Mining Tractable Sets of Graph Patterns with the Minimum Description Length Principle, ED MathStic.
- **Raphaël Gauthier**, 2017-2021, supervisors: Christine Largouët and Jean-Yves Dourmad (INRAE), title: "Precision feeding system for lactating sows, using modeling and machine learning", ED EGAAL.

11.2.3 Juries

- Alexandre Termier was a member of the following PhD juries in 2021: Yichang Wang, 20/09 Rennes (committee member, president); Amaury Bouchra Pilet, 10/11 Rennes (committee member, president); Francesco Bariatti, 23/11 Rennes (committee member, president); Luong Nguyen, 10/12 Tours (reviewer). He was a member of the mid-term evaluation juries of Grégory Siekaniec (University of Rennes 1) and Daniel Rosendo (University of Rennes 1).
- Elisa Fromont was a member of the following PhD juries in 2021: Ian Jeantet, 5/01 Rennes (committee member, president); Corentin Lonjarret, 12/01 Lyon (committee member, president); Eduardo Hugo Sanchez, 17/02 Toulouse (committee member, president); Yann-Raphael Lifchitz, 20/04,

Rennes (committee member, president); Ahmed Nassar, 21/05 Vannes (committee member, president); Alexandre Araujo, 1/06, Paris (committee member, president); Loïck Bonniot, 10/06, Rennes (committee member, president); Luca Erculiani, 10/06 Trento (reviewer); Tatiana Makhalova, 23/06 Nancy (committee member); Elies Gherbi, 5/07 Evry (reviewer); Benjamin Lucas, Monash, Australia (reviewer); Guillaume Renton, 08/07 Rouen (committee member); Yichang Wang, 20/09 Rennes (co-surpervisor); Nitika Verma, 20/10 Grenoble (committee member); Vincent Le Guen, 30/11 Paris (committee member); Tanguy Kerdoncuff, 9/12 Saint-Etienne (committee member, president); Grégoire Siekaniec, 10/12 Rennes (committee member, president); Victor Bouvier, 13/12 Paris (committee member); Heng Zhang, 14/12 Rennes (co-supervisor).

- Christine Largouët was a member of the following PhD juries: Raphaël Gauthier 05/03/2021 (co-supervisor), Yang Su 22/10/2021 (Committee member).
- Luis Galárraga was a member of mid-term evaluation committee of the following PhD candidates: Armand Boschin (Télécom ParisTech, 05/2021) working on learning and exploiting knowledge graph representations, Hugo Ayats (Univ. Rennes I, 05/2021) working on relation extraction via concepts of neighbors, Louis Béziaud (Univ. Rennes I/UAQM, 05/2021) working on privacy and fairness in law, and Armita KHajeh Nassiri (Univ. Paris Saclay, 11/2021) working on discovering expressive rules in knowledge graphs
- Peggy Cellier was a member of the following PhD juries: Cheikh Brahim EL VAIGH, 07/01/2021, Univ. Rennes I (committee member); Francesco Bariatti, 23/11 Rennes (co-supervisor). She was a member of mid-term evaluation committee of the following PhD candidates: Cyrielle Mallart (Univ. Rennes I), Priscilla Keip (Université de Montpellier), Albeiro Espinal (IMT Atlantique), Grégory Martin (Univ. Rennes I).

11.2.4 Education

- Veronique Masson is the head of the L3 studies in Computer Science at University of Rennes 1
- Since September 2021, Alexandre Termier is co-head of Master 2 SIF (Science Informatique - research master in Computer Science) at University of Rennes 1, with Bertrand Coüasnon (INSA Rennes).
- Elisa Fromont is reponsible for the Master MEEF NSI (a master dedicated to future high school teachers of computer sciences)
- Christine Largouët is responsible for the computer-science educational program in Institut Agro, Agrocampus Ouest (2 engineering schools)
- Romaric Gaudel is responsible for the department of computer-science at ENSAI (École nationale de la statistique et de l'analyse de l'information).
- Since October 2021, Peggy Cellier is responsible of the last year at Computer Science Department at INSA (master 2 level, about 70 students).

11.2.5 Interventions

- Camille-Sovanneary Gauthier participated to the challenge "ma thèse en 180 secondes" and was selected for the inter-regional final.
- Lénaïg Cornanguer participated to the program **L codent L créent** that raises awareness of programming among middle school girls during creative workshops.

12 Scientific production

12.1 Publications of the year

International journals

[1] K. Fauvel, T. Lin, V. Masson, É. Fromont and A. Termier. 'XCM: An Explainable Convolutional Neural Network for Multivariate Time Series Classification'. In: *Mathematics* 9.23 (5th Dec. 2021), pp. 1–20. DOI: 10.3390/math9233137. URL: https://hal.inria.fr/hal-03469487.

[2] S. Felton, P. Brault, E. Fromont and E. Marchand. 'Visual Servoing in Autoencoder Latent Space'. In: *IEEE Robotics and Automation Letters* (2022). DOI: 10.1109/LRA.2022.3144490. URL: https://hal.inria.fr/hal-03506036.

[3] C. Gaillard, M. Durand, C. Largouët, J.-Y. Dourmad and C. Tallet. 'Effects of the environment and animal behavior on nutrient requirements for gestating sows: Future improvements in precision feeding'. In: *Animal Feed Science and Technology* 279 (2021), pp. 1–17. DOI: 10.1016/j.anifeedsci.2021.115034. URL: https://hal.inrae.fr/hal-03347926.

[4] L. Galárraga, D. Hernández and K. Hose. 'Computing How-provenance for SPARQL queries via Query Rewriting'. In: *Proceedings of the VLDB Endowment (PVLDB)* 14.13 (Sept. 2021), pp. 3389–3401. DOI: 10.14778/3484224.3484235. URL: https://hal.inria.fr/hal-03500656.

[5] R. Gauthier, C. Largouët, L. Rozé and J.-Y. Dourmad. 'Online forecasting of daily feed intake in lactating sows supported by offline time-series clustering, for precision livestock farming'. In: *Computers and Electronics in Agriculture* 188 (2021), p. 106329. DOI: 10.1016/j.compag.2021.106329. URL: https://hal.archives-ouvertes.fr/hal-03315102.

[6] G. Longin, L. Bonneau De Beaufort, G. Fontenelle, R. Rinaldo, J.-M. Roussel and P.-Y. Le Bail. 'Fishers' perceptions of river resources: case study of French Guiana native populations using contextual cognitive mapping'. In: *Cybium : Revue Internationale d'Ichtyologie* 45.1 (2021), pp. 5–20. DOI: 10.26028/cybium/2021-451-001. URL: https://hal.inrae.fr/hal-03148402.

[7] G. Longin, G. Fontenelle, L. Bonneau De Beaufort, C. Delord, S. Launay, R. Rinaldo, G. Lassalle, P.-Y. Le Bail and J.-M. Roussel. 'When subsistence fishing meets conservation issues: Survey of a small fishery in a neotropical river with high biodiversity value'. In: *Fisheries Research* 241 (Sept. 2021), pp. 1–9. DOI: 10.1016/j.fishres.2021.105995. URL: https://hal-agrocampus-ouest.archives-ouvertes.fr/hal-03228509.

[8] O. Pelgrin, L. Galárraga and K. Hose. 'Towards Fully-fledged Archiving for RDF Datasets'. In: *Semantic Web – Interoperability, Usability, Applicability* 12.6 (12th Apr. 2021), pp. 903–925. DOI: 10.3233/SW-210434. URL: https://hal.inria.fr/hal-03500522.

[9] Y. Zhang, T. Bouadi, Y. Wang and A. Martin. 'A distance for evidential preferences with application to group decision making'. In: *Information Sciences* 568 (2021), pp. 113–132. DOI: 10.1016/j.ins.2021.03.011. URL: https://hal.archives-ouvertes.fr/hal-03171577.

International peer-reviewed conferences

[10] D. Balouek-Thomert, P. Silva, K. Fauvel, A. Costan, G. Antoniu and M. Parashar. 'MDSC: Modelling Distributed Stream Processing across the Edge-to-Cloud Continuum'. In: DML-ICC 2021 workshop (held in conjunction with UCC 2021). Leicester, United Kingdom, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03510012.

[11] L. Cornanguer. 'Passive Learning of Timed Automata from Logs (Student Abstract)'. In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*. AAAI 2021 - 35th AAAI Conference on Artificial Intelligence. Vancouver (virtual), Canada, 2nd Feb. 2021, pp. 1–2. URL: https://hal.inria.fr/hal-03201649.

[12] S. Felton, E. Fromont and E. Marchand. 'Siame-se(3): regression in se(3) for end-to-end visual servoing'. In: ICRA 2021 - IEEE International Conference on Robotics and Automation. Xi'an, China: IEEE, 30th May 2021, pp. 14454–14460. URL: https://hal.inria.fr/hal-03173684.

[13] L. Galárraga, G. Mendez and K. Chiluiza. 'Showing Academic Performance Predictions during Term Planning: Effects on Students' Decisions, Behaviors, and Preferences'. In: CHI 2021 - ACM CHI Conference on Human Factors in Computing Systems. Yokohama, Japan: ACM, 8th May 2021, pp. 1–17. DOI: 10.1145/3411764.3445718. URL: https://hal.inria.fr/hal-03500534.

[14] C.-S. Gauthier, R. Gaudel and E. Fromont. 'Bandit Algorithm for Both Unknown Best Position and Best Item Display on Web Pages'. In: IDA 2021 - 19th International Symposium on Intelligent Data Analysis. Advances in Intelligent Data Analysis XIX. Porto (virtual), Portugal, 2021, p. 12. URL: https://hal.archives-ouvertes.fr/hal-03163763.

[15] C.-S. Gauthier, R. Gaudel, E. Fromont and B. A. Lompo. 'Parametric Graph for Unimodal Ranking Bandit'. In: ICML 2021 - International Conference on Machine Learning. Vol. 139. Proceedings of the 38th International Conference on Machine Learning. Virtual, Canada, 2021, pp. 3630–3639. URL: https://hal.archives-ouvertes.fr/hal-03256621.

[16] R. Gauthier, C. Largouët, L. Rozé and J.-Y. Dourmad. 'Algorithm for real-time prediction of daily feed intake in lactating sows'. In: *Journées de la Recherche Porcine en France*. 53. Journées de la Recherche Porcine. Vol. 53. 53èmes Journées de la recherche porcine. En ligne, France: Ifip, 1st Feb. 2021, pp. 127–132. URL: https://hal.inria.fr/hal-03134418.

[17] C. Leverger, T. Guyet, S. Malinowski, V. Lemaire, A. Bondu, L. Rozé, A. Termier and R. Marguerie. 'Probabilistic forecasting of seasonal time series Combining clustering and classification for forecasting'. In: ITISE 2021 - 7th International Conference on Time Series and Forecasting. Gran Canaria, Spain, 19th July 2021, pp. 1–13. URL: https://hal.archives-ouvertes.fr/hal-03326626.

[18] O. Pelgrin, L. Galárraga and K. Hose. 'TrieDF: Efficient In-memory Indexing for Metadata-augmented RDF'. In: CEUR-WS.org 2021 - CEUR Workshop Proceedings. Virtual Event, France, 18th Oct. 2021, pp. 1–10. URL: https://hal.inria.fr/hal-03500647.

Conferences without proceedings

[19] J. Bakalara, T. Guyet, O. Dameron, A. Happe and E. Oger. 'An extension of chronicles temporal model with taxonomies -Application to epidemiological studies'. In: HEALTHINF 2021 - 14th International Conference on Health Informatics. online, France, 11th Feb. 2021, pp. 1–10. URL: https://hal.archives-ouvertes.fr/hal-03096846.

[20] E. Bourrand, L. Galárraga, E. Galbrun, E. Fromont and A. Termier. 'Discovering Useful Compact Sets of Sequential Rules in a Long Sequence'. In: ICTAI 2021 - 33rd IEEE International Conference on Tools with Artificial Intelligence. Virtual, United States: IEEE, 1st Nov. 2021, pp. 1–5. URL: https://hal.archives-ouvertes.fr/hal-03494520.

[21] M. Chambard, T. Guyet, Y.-L. Nguyen and E. Audureau. 'Temporal phenotyping for characterisation of hospital care pathways of COVID19 patients'. In: AALTD 2021 - The 6th International Workshop on Advanced Analytics and Learning on Temporal Data. Bilbao / Virtual, Spain, 13th Sept. 2021, pp. 1–16. URL: https://hal.archives-ouvertes.fr/hal-03326636.

[22] S. Coumes, T. Bouadi, L. Nourine and A. Termier. 'Skyline Groups Are Ideals. An Efficient Algorithm for Enumerating Skyline Groups'. In: IWOCA 2021 - 32nd International Workshop on Combinatorial Algorithms. Vol. 12757. Lecture Notes in Computer Science. Ottawa, Canada: Springer International Publishing, 30th June 2021, pp. 223–236. DOI: 10.1007/978-3-030-79987-8_16. URL: https://hal.archives-ouvertes.fr/hal-03519794.

[23] K. Fauvel, V. Masson and E. Fromont. 'A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers'. In: IJCAI-PRICAI 2020 - Workshop on Explainable Artificial Intelligence (XAI). Yokohama, Japan, 8th Jan. 2021, pp. 1–8. URL: https://hal.archives-ouvertes.fr/hal-03094885.

[24] C.-S. Gauthier, R. Gaudel, E. Fromont and A. B. Lompo. 'Ordonnancement d'objets par bandits unimodaux sur des graphes paramétriques'. In: CAp 2021 - Conférence sur l'Apprentissage automatique. Saint-Etienne (en ligne), France, 14th June 2021, p. 1. URL: https://hal.archives-ouvertes.fr/hal-03236458.

[25] T. Guyet, T. Allard, J. Bakalara and O. Dameron. 'An open generator of synthetic administrative healthcare databases'. In: *Actes de l'atelier Intelligence Artificielle et Santé (IAS)*. IAS 2021 - Atelier Intelligence Artificielle et Santé. Bordeaux (virtuel), France, 29th June 2021, pp. 1–8. URL: https://hal.archives-ouvertes.fr/hal-03326618.

[26] G. Martin, M. Donain, E. Fromont, T. Guns, L. Roze and A. Termier. 'Prediction-Based Fleet Relocation for Free Floating Car Sharing Services'. In: ICTAI 2021 - 33rd International Conference on Tools with Artificial Intelligence. Virtual, United States, 1st Nov. 2021, pp. 1–5. URL: https://hal.archives-ouvertes.fr/hal-03491955.

[27] J. Signe. 'Extraction de sous-groupes exceptionnels de séries temporelles'. In: RJCIA 2021 - Rencontres des Jeunes Chercheurs en Intelligence Artificielle. Bordeaux / Virtual, France, 1st July 2021, pp. 89–90. URL: https://hal.archives-ouvertes.fr/hal-03298742.

[28] A. Voyez, T. Allard, G. Avoine, P. Cauchois, E. Fromont and M. Simonin. 'Attaque par inférence d'appartenance sur des séries temporelles agrégées en utilisant la programmation par contraintes'. In: BDA 2021 - 37ème Conférence sur la Gestion de Données – Principes, Technologies et Applications. Paris, France, 25th Oct. 2021, p. 1. URL: https://hal.archives-ouvertes.fr/hal-03499977.

[29] H. Zhang, E. Fromont, S. Lefevre and B. Avignon. 'Deep Active Learning from Multispectral Data Through Cross-Modality Prediction Inconsistency'. In: ICIP 2021 - 28th IEEE International Conference on Image Processing. Anchorage, United States: IEEE, 19th Sept. 2021, pp. 1–5. URL: https://hal.archives-ouvertes.fr/hal-03236409.

[30] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon. 'Guided Attentive Feature Fusion for Multispectral Pedestrian Detection'. In: WACV 2021 - IEEE Winter Conference on Applications of Computer Vision. Waikoloa /Virtual, United States: IEEE, 5th Jan. 2021, pp. 1–9. URL: https://hal.archives-ouvertes.fr/hal-03119907.

[31] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon. 'Low-cost Multispectral Scene Analysis with Modality Distillation'. In: IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa /Virtual, United States, 4th Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03491950.

[32] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon. 'PDF-Distil: including Prediction Disagreements in Feature-based Distillation for object detection'. In: *British Machine Vision Conference (BMVC) 2021 proceedings*. BMVC 2021 - 32nd British Machine Vision Conference. Virtual, United Kingdom, 22nd Nov. 2021, pp. 1–13. URL: https://hal.archives-ouvertes.fr/hal-03487128.

Scientific book chapters

[33] L. Galárraga, O. Pelgrin and A. Termier. 'HiPaR: Hierarchical Pattern-Aided Regression'. In: *Advances in Knowledge Discovery and Data Mining*. Vol. 12712. Lecture Notes in Computer Science. Springer International Publishing, 9th May 2021, pp. 320–332. DOI: 10.1007/978-3-030-75762-5_26. URL: https://hal.inria.fr/hal-03500548.

Doctoral dissertations and habilitation theses

[34] Y. Wang. 'Interpretable time series classification'. Université Rennes 1, 20th Sept. 2021. URL: https://tel.archives-ouvertes.fr/tel-03509607.

[35] H. Zhang. 'Multispectral object detection'. Rennes 1, 14th Dec. 2021. URL: https://hal.archives-ouvertes.fr/tel-03530257.

Reports & preprints

[36] S. Felton, P. Brault, E. Fromont and E. Marchand. *Visual Servoing in Autoencoder Latent Space: Supplementary material*. 25th Nov. 2021. URL: https://hal.inria.fr/hal-03448667.

Other scientific publications

[37] G. Gravier, E. Fromont, N. Courty, T. Furon, C. Guillemot and P. Robuffo Giordano. 'Rennes - une IA souveraine au service de la vie publique'. In: *Bulletin de l'Association Française pour l'Intelligence Artificielle* (2021). URL: https://hal.archives-ouvertes.fr/hal-03313161.

[38] T. Guyet. 'Génération d'un SNDS synthétique à partir de données ouvertes'. In: EGC 2021 - Actes de la conférence Extraction et Gestion des connaissances. Montpellier, France, 25th Jan. 2021, pp. 1–2. URL: https://hal.archives-ouvertes.fr/hal-03326620.

12.2 Cited publications

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

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

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

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

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

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