IN PARTNERSHIP WITH:
Institut national des sciences appliquées de Rennes

Institut national supérieur des sciences agronomiques, agroalimentaires, horticoles et du paysage

Université Rennes 1

Activity Report 2017

Project-Team LACODAM

Large scale Collaborative Data Mining

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

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team LACODAM

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

Keywords:

Computer Science and Digital Science:

A2.1. - Programming Languages
A2.1.5. - Constraint programming
A2.1.11. - Proof languages
A3. - Data and knowledge
A3.1. - Data
A3.1.1. - Modeling, representation
A3.2. - Knowledge
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. - Machine learning and statistics
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A4. - Security and privacy
A4.9.1. - Intrusion detection
A7.1. - Algorithms
A7.2. - Logic in Computer Science
A7.2.1. - Decision procedures
A7.2.2. - Automated Theorem Proving
A9. - Artificial intelligence
A9.1. - Knowledge
A9.2. - Machine learning
A9.3. - Signal analysis
A9.6. - Decision support
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B1.1.6. - Genomics
B2. - Health
B2.3. - Epidemiology
B2.4. - Therapies

B2.4.2. - Drug resistance
B3. - Environment and planet
B3.4. - Risks
B3.4.3. - Pollution
B3.5. - Agronomy
B3.6. - Ecology
B3.6.1. - Biodiversity
B4. - Energy
B6. - IT and telecom
B6.2. - Network technologies
B8. - Smart Cities and Territories
B8.1. - Smart building/home
B8.1.1. - Energy for smart buildings
B8.1.2. - Sensor networks for smart buildings
B8.2. - Connected city
B8.3. - Urbanism and urban planning
B9. - Society and Knowledge
B9.4.5. - Data science
B9.9.1. - Environmental risks

1. Personnel

Research Scientists

Luis Galárraga Del Prado [Inria, Researcher, from Oct. 2017]
René Quiniou [Inria, Researcher]
Torsten Schaub [University of Postdam, Inria International Chair, Researcher, HDR]

Faculty Members

Alexandre Termier [Team leader, Univ. Rennes I, Professor, Partial secondment Inria from September 2016, HDR]
Marie-Odile Cordier [Univ. Rennes I, Faculty Member, Emeritus Professor, HDR]
Elisa Fromont [Univ. Rennes I, Faculty Member, Professor, from Sep 2017, HDR]
Véronique Masson [Univ. Rennes I, Associate Professor, Partial secondment Inria from September 2017]
Laurence Rozé [INSA Rennes, Associate Professor]
Thomas Guyet [AGROCAMPUS OUEST, Associate Professor]
Christine Largouët [AGROCAMPUS OUEST, Associate Professor]

Post-Doctoral Fellow

Ahmed Samet [Univ. Rennes I, until May 2017]

PhD Students

Yann Dauxais [Univ. Rennes I]
Kevin Fauvel [Inria, from Oct. 2017]
Clément Gautrais [Univ. de Rennes I]
Maël Guillemé [Energiency, granted by CIFRE]
Colin Leverger [Orange Labs, from Oct. 2017]
Alban Siffer [Amossys, from May 2016]

Technical staff

Louis Bonneau de Beaufort [AGROCAMPUS OUEST, Engineer]

Interns

Scarlett Kelly [Inria, from May 2017 until Sept. 2017]
François Laferrière [Inria, from Feb. 2017 until June 2017]
Julien Laurent [Univ. de Rennes I, from May 2017 until July 2017]
Grégory Martin [Univ. de Rennes I, from May 2017 until Sept. 2017]
François Mentec [Univ. de Rennes I, from May 2017 until Aug. 2017]
Aikaterini Tsesmeli [Inria, from Feb. 2017 until July 2017]

Administrative Assistant

Marie-Noëlle Georgeault [Inria]

External Collaborators

Philippe Besnard [CNRS, Researcher, HDR]
Anne-Isabelle Graux [INRA, Researcher]

2. Overall Objectives

2.1. Overall Objectives

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

We claim that nowadays, Data Science is in its “Iron Age”: Good tools are available, however skilled craftsmen are required to use them in order to transform raw material (the data) into finished products (knowledge, decisions). We foresee that in a decade from now, we should be in an “Industrial Age” of Data Science, where more elaborate tools will alleviate a lot of the human work required in Data Science. Basic Data Science tasks will no longer require a skilled data scientist; instead software tools will enable small companies or even individuals to get valuable knowledge from their data. Skilled data scientists will thus be fully available to work on the hard tasks that matter. This will entail a drastic improvement in productivity thanks to a new generation of tools that will do the tedious work for data analysts and scientists.

The objective of the LACODAM team 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. Since data science in its current state involves lots of human intervention, we envision a novel generation of data analysis and decision support tools that require significantly less tedious human work. Such solutions will rely only on a few interactions between the user and the system with high added value. We foresee solutions that bridge data mining techniques with artificial intelligence (AI) approaches, in order to integrate existing automated reasoning techniques in knowledge discovery workflows. Such solutions can be seen as “second order” AI tasks: they exploit AI techniques (for example, planning) in order to pilot more classical AI tasks such as data mining and decision support.

3. Research Program

3.1. Introduction

The three research axes of the LACODAM project-team are the following. First, we briefly introduce these axes, as well as their interplay:

- The first research axis is dedicated to the design of *novel pattern mining methods*. Pattern mining is one of the most important approaches to discover novel knowledge in data, and one of our strongest areas of expertise. The work on this axis will serve as foundations for work on the other two axes. Thus, this axis will have the strongest impact on our goals overall.
- The second axis tackles another aspect of knowledge discovery in data: the *interaction between the user and the system* in order to co-discover novel knowledge. Our team has plenty of experience collaborating with domain experts, and is therefore aware of the need to improve such interaction.
- The third axis concerns *decision support*. With the help of methods from the two previous axes, our goal here is to design systems that can either assist humans with making decisions, or make relevant decisions in situations where extremely fast reaction is required.

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

3.2. Pattern mining algorithms

Twenty years of research in pattern mining have resulted in efficient approaches to handle the algorithmic complexity of the problem. Existing algorithms are now able to efficiently extract patterns with complex structures (ex: sequences, graphs, co-variations) from large datasets. However, when dealing with large, real-world datasets, these methods still output a huge set of patterns, which is impractical for human analysis. This problem is called *pattern explosion*. The ongoing challenge of pattern mining research is to extract fewer but more meaningful patterns. The LACODAM team is committed to solve the pattern explosion problem by pursuing the following four research topics:

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

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

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

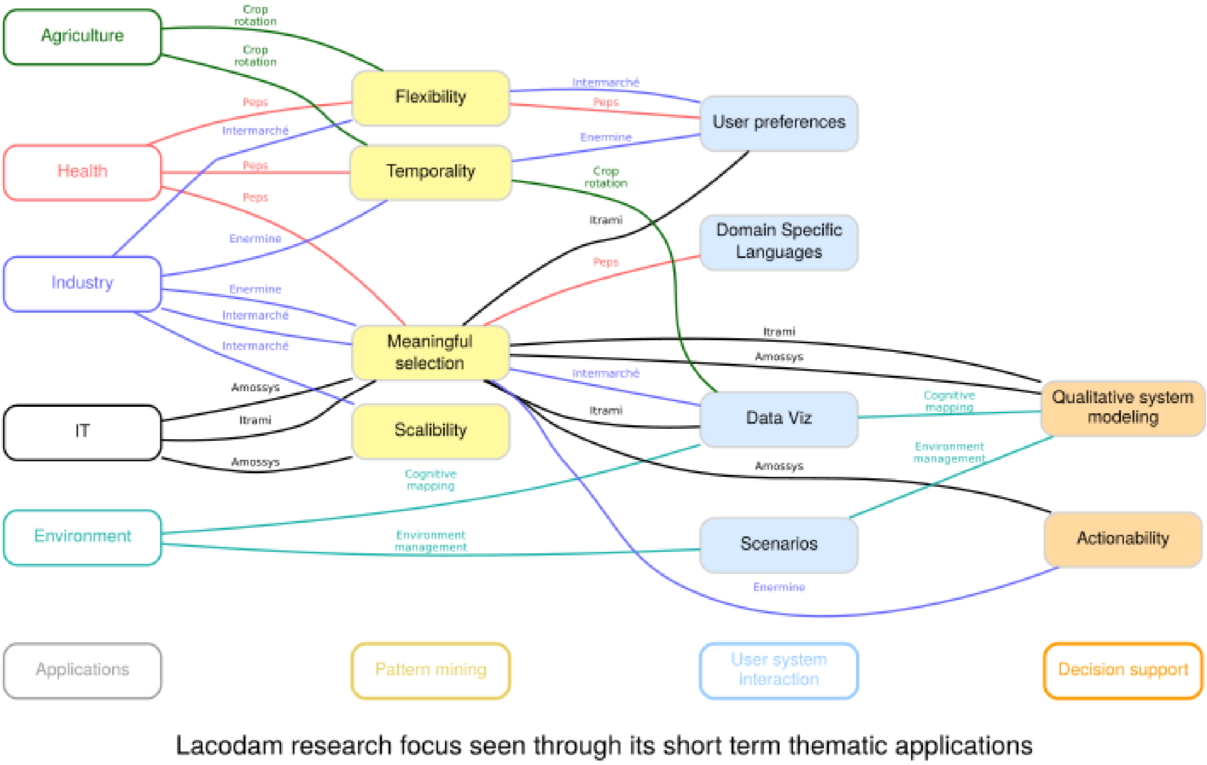

Figure 1. Lacodam research topics organized by axis and application

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

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

3.3. User/system interaction

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

- *Domain Specific Languages:* One way to interact with the user is to propose a Domain Specific Language (DSL) tailored to the domain at hand and to the analysis tasks. The challenge is to propose a DSL allowing the users to easily express the required processing workflows, to deploy those workflows for mining on large volumes of data and to offer as much automation as possible.
- *What if / What for scenarios:* We also investigate the use of scenarios to query results from data mining processes, as well as other complex processes such as complex system simulations or model predictions. Such scenarios are answers to questions of the type “what if [situation]?” or “what [should be done] for [expected outcome]?”.
- *User preferences:* In exploratory analysis, users often do not have a precise idea of what they want, and are not able to formulate such queries. Hence, in LACODAM we investigate simple ways for users to express their interests and preferences, either during the mining process – to guide the search space exploration –, or afterwards during the filtering and interpretation of the most relevant results.
- *Data visualization:* Most of the research directions presented in this document require users to examine patterns at some point. The output of most pattern mining algorithms is usually a (long) list of patterns. While this presentation can be sufficient for some applications, often it does not provide a complete understanding, especially for non-experts in pattern mining. A transversal research topic that we want to explore in LACODAM is to propose data visualization techniques that are adequate for understanding output results. Numerous (failed) experiments have shown that data mining and data visualization are fields, which require distinct skills, thus researchers in one field usually do not make significant advances in the other field (this is detailed in [Keim 2010]). Thus, our strategy is to establish collaborations with prominent data visualization teams for this line of research, with a long term goal to recruit a specialist in data visualization if the opportunity arises.

3.4. Decision support

Predictive sequential patterns have a direct application in diagnosis. LACODAM inherits a strong background in decision support systems with internationally recognized expertise in diagnosis from the former DREAM team. This AI subfield is concerned with determining whether a system is operating normally or not, and the cause of faulty behaviors. The studied system can be an agro- or eco-system, a software system, a living being, etc.

The increasing volumes of data coming from a range of different systems (ex: sensor data from agro-environmental systems, log data from software systems, biological data coming from health monitoring systems) can help human and software agents make better decisions. Hence, LACODAM builds upon the idea that decision support systems (an interest bequeathed from DREAM) should take advantage of the available data. This third and last research axis is thus a meeting point for all members of the team, as it requires the integration of AI techniques for traditional decision support systems with results from data mining techniques.

Two main research sub-axes are investigated in LACODAM:

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

3.5. Long-term goals

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

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

  We consider that this is a second order AI task, where AI techniques such as planning are used to explore the workflow search space, the workflow itself being composed of data mining and/or decision support components. This is a strategic evolution for data science endeavors, were the demand far exceeds the available human skilled manpower.
- *Logic argumentation based on epistemic interest.* Having increasingly automated approaches will require better and better ways to handle the interactions with the user. Our second long term goal is to explore the use of logic argumentation, i.e., the formalisation of human strategies for reasoning and arguing, in the interaction between users and data analysis tools. Alongside visualization and interactive data mining tools, logic argumentation can be a way for users to query both the results and the way they are obtained. Such querying can also help the expert to reformulate her query in an interactive analysis setting.

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

  The first step is to define a formal argumentation framework for explaining data mining results. This implies to continue theoretical work on the foundations of argumentation in order to identify the most adapted framework (either existing or a new one to be defined). Logic argumentation may be implemented and deeply explored in ASP, allowing us to build on our expertise in this logic language.

- *Collaborative feedback and knowledge management.* We are convinced that improving the data science process, and possibly automating it, will rely on high-quality feedback from communities on the web. Consider for example what has been achieved by collaborative platforms such as StackOverflow: it has become the reference site for any programming question.

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

4. Application Domains

4.1. Introduction

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

4.2. Industry

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

- **Execution trace analysis for SOC debugging (STMicroelectronics)**. We have an ongoing collaborations with STMicroelectronics, which is one of top-5 electronic chip makers worldwide. Nowadays, set-top boxes, smartphones and onboard car computers are powered by highly integrated chips called System-on-Chip (SoC). Such chips contain on a single die, processing units, memories, IO units and specialized accelerators (such as audio and video encoding/decoding). Programming SoC is a hard task due to their inherent parallelism, leading to subtle bugs when several components do not deliver their results within a given time frame. Existing debuggers and profilers are ill-adapted in this case because of their high intrusivity that modifies the timings. Hence the most used technique is to capture a trace of the execution and analyze it post-mortem. While Alexandre Termier was in Grenoble he initiated several works for analyzing such traces with pattern mining techniques, which he is now pursuing with his colleagues of the LACODAM project-team.
- **Resource consumption analysis for optimizing energy consumption and practices in industrial factories (Energiency)**. In order to increase their benefits, companies introduce more and more sensors in their factories. Thus, the resource (electricity, water, etc.) consumption of engines, workshops and factories are recorded in the form of times series or temporal sequences. The person who is in charge of resource consumption optimization needs better software than classical spreadsheets for this purpose. He/she needs effective decision-aiding tools with statistical and artificial intelligence knowledge. The start-up Energiency aims at designing and offering such pieces of software for analyzing energy consumption. The starting CIFRE PhD thesis of Maël Guillemé aims at proposing new approaches and solutions from the data mining field to tackle this issue.
- **Security (Amossys)**. Current networks are faced with an increasing variety of attacks, from the classic “DDoS” that makes a server unusuable for a few fours, to advanced attacks that silently infiltrate a network and exfiltrate sensitive information months or even years later. Such intrusions, called APT (Advanced Persistent Threat) are extremely hard to detect, and this will become even

harder as most communications will be encrypted. A promising solution is to work on "behavioral analysis", by discovering patterns based on the metadata of IP-packets. Such patterns can relate to an unusual sequencing of events, or to an unusual communication graph. Finding such complex patterns over a large volume of streaming data requires to revisit existing stream mining algorithms to dramatically improve their throughput, while guaranteeing a manageable false positive rate. We are collaborating on this topic with the Amossys company and the EMSEC team of Irisa through the co-supervision of a CIFRE PhD (located in the EMSEC team). Our goal is to design novel anomaly detection methods that can detect APT, and that scales on real traffic volumes.

- **Market basket data analysis (Intermarché) and multi-channel interaction data analysis (EDF) for better Customer Relationship Management (CRM)**. An important application domain of data mining for companies that deal with large numbers of customers is to analyze customer interaction data, either for marketing purposes or to improve the quality of service. We have activities in both settings. In the first case, we collaborate with a major french retailer, Intermarché, in order to detect customer churn by analyzing market basket data. In the second case, we collaborate with the major french power supplier, EDF, to discover actionable patterns for CRM that aim at avoiding undesirable situations. We use logs of user interactions with the company (e.g., web clicks, phone calls, etc.) for this purpose.

4.3. Health

- **Care pathways analysis for supporting pharmaco-epidemiological studies**. Pharmaco-epidemiology applies the methodologies developed in general epidemiology to answer to questions about the uses and effects of health products, drugs [33], [32] or medical devices [27], 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 algorithm 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 [31]. 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.

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

4.4. Agriculture and environment

- **Dairy farming**. The use and analysis of data acquired in dairy farming is a challenge both for data science and animal science. The goal is to improve farming conditions, i.e., health, welfare and environment, as well as farmers' income. Nowadays, animals are monitored by multiple sensors giving a wealth of heterogeneous data, ex: 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 the 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.

  Similar problems are investigated with the veterinary department of the University of Calgary in the context of cattle monitoring from multiple sensors placed on calves for the early detection of diseases.
- **Optimizing the nutrition of individual sow**. Another direction for further research is the combination of data flows with prediction models in order to learn nutrition strategies. Raphaël Gauthier started a PhD thesis (#DigitAg Grant) in November 2017 with both Inria and INRA supervisors. His research addresses the problem of finding the optimal diet to be supplied to individual sows. Given all the information available, e.g., time-series information about previous feeding, environmental data, scientists models, the research goal is to design new algorithms to determine the optimal ration for a given sow in a given day. Efficiency issues of developed algorithms will be considered since the proposed software should work in real-time on the automated feeder. The decision support process should involve the stakeholder to ensure a good level of acceptance, confidence and understanding of the final tool.
- **Ecosystem modeling and management**. Ongoing research on ecosystem management includes modelling of ecosystems and anthroprogenic pressures, with a special concern on the representation of socio-economical factors that impact human decisions. A main research issue is how to represent these factors and how to integrate their impact on the ecosystem simulation model. This work is an ongoing cooperation with ecologists from the Marine Spatial Ecology of Queensland University, Australia and from Agrocampus Ouest.

5. Highlights of the Year

5.1. Highlights of the Year

This year's highlight was the exceptional success of LACODAM in the recruiting process: we could hire two new staff members.

- Elisa Fromont joined as a Professor of University of Rennes 1. She brings to the team her skills in Machine Learning, which include a precious expertise on dealing with numerical data.
- Luis Galárraga Del Prado joined as an Inria Researcher. He brings to the team his skills in exploiting knowledge bases, which will be strongly needed for including domain knowledge in our approaches. He will also reinforce our work on rule mining.

6. New Software and Platforms

6.1. EcoMata

KEYWORD: Environment perception
FUNCTIONAL DESCRIPTION: The EcoMata toolbox provides means for qualitative modeling and exploration of ecosystems in order to aid the design of environmental guidelines. We have proposed a new qualitative approach for ecosystem modeling based on the timed automata (TA) formalism combined to a high-level query language for exploring scenarios.

- Participants: Christine Largouët, Marie-Odile Cordier, Thomas Guyet and Yulong Zhao
- Contact: Christine Largouët
- URL: https://team.inria.fr/dream/fr/ecomata/

6.2. PATURMATA

KEYWORDS: Bioinformatics - Biology
SCIENTIFIC DESCRIPTION: The Paturmata tool-box provides means for qualitative modeling and the exploration of agrosystems, specifically management of herd based on pasture. The system is modeled using a hierarchical hybrid model described in the timed automata formalism.
FUNCTIONAL DESCRIPTION: In the PaturMata software, users can create a pasture system description by entering herds and plots information. For each herd, the only parameter is the number of animals. For each plot, users should enter the surface, the density, the herb height, the distance to the milking shed, a herb growth profile and an accessibility degree. Users then specify pasturing and fertilization strategies. Finally, users can launch a pasture execution. PaturMata displays the results and a detailed trace of pasture. Users can launch a batch of different strategies and compare the results in order to find the best pasture strategy. PaturMata is developed in Java (Swing for the GUI) and the model-checker that is called for the timed properties verification is UPPAAL.

- Participants: Christine Largouët and Marie-Odile Cordier
- Contact: Christine Largouët

6.3. Promise

KEYWORDS: Data mining - Monitoring
FUNCTIONAL DESCRIPTION: Promise is a software that predicts rare events in industrial production systems from data analysis of energy consumption data. The data is represented as a time series. The program takes as input the temporal series of energy consumption, an abnormal pattern (rare event) and a temporal dilatation, and outputs a set of sub-series similar (according to a similarity metric) to the abnormal pattern.

- Participants: Véronique Masson, Laurence Rozé and Mael Guilleme
- Contact: Véronique Masson

6.4. GWASDM

Genome Wide Association Study using Data Mining strategy
KEYWORDS: GWAS - Data mining
FUNCTIONAL DESCRIPTION: From two cohorts of genotyped individuals (case and control), the GWASDM software performs a Genome Wide Association Study based on data mining techniques and generates several patterns of SNPs that correlate with a given phenotype. The algorithm implemented in GWASDM directly uses relative risk measures such as risk ratio, odds ratio and absolute risk reduction combined with confidence intervals as anti-monotonic properties to efficiently prune the search space. The algorithm discovers a complete set of discriminating patterns with regard to given thresholds or applies heuristic strategies to extract the largest statistically significant discriminating patterns in a given dataset.

- Contact: Dominique Lavenier

6.5. DCM

Discriminant Chronicle Mining
KEYWORDS: Pattern extraction - Sequence - Classification
FUNCTIONAL DESCRIPTION: DCM is a temporal sequences analysis tool. It extracts discriminant chronicles from a large set of labeled sequences. A sequence is made of timestamped events. Each sequence of events is associated to a label (e.g. positive and negative sequences). A chronicle is a temporal model that characterizes a behavior by a set of events linked by temporal constraints. The DCM algorithm extracts chronicles that occurs more in positive sequences than in negative sequences.

- Participants: Yann Dauxais and Thomas Guyet
- Partners: REPERES - Université de Rennes 1
- Contact: Yann Dauxais
- Publications: Discriminant chronicles mining: Application to care pathways analytics - Extraction de chroniques discriminantes
- URL: https://gitlab.inria.fr/ydauxais/DCM

6.6. NTGSP

Negative Time-Gap Sequential Patterns
KEYWORDS: Pattern discovery - Sequence
FUNCTIONAL DESCRIPTION: The NTGSP algorithm is a sequential pattern mining algorithm. It analyses a large database of temporal sequences, i.e., events with timestamps, by extracting its regularities (the patterns). A pattern describes the behavior as a sequence of events that frequently occurred in sequences. What makes NTGSP novel is its ability to handle patterns with negations, i.e., the description of a behavior that specifies the absence of an event. More precisely, it extracts frequent sequences with positive and negative events, as well as temporal information about the delay between these events.

- Participants: Thomas Guyet and René Quiniou
- Partner: Edf
- Contact: René Quiniou
- Publication: Fouille de motifs temporels négatifs

6.7. Relevant Interval Rules Miner

KEYWORDS: Association rule - Pattern discovery - Formal concept analysis
FUNCTIONAL DESCRIPTION: This software extracts relevant rules from a dataset of labeled numerical attributes (tabular datasets). A rule is an interval-based pattern associated to a predicted label. The tool extracts a subset of rules based on the accuracy and relevance criteria where most of the algorithms are simply based on accuracy. This allow us to extract the best rules that capture the data behavior.

- Participants: René Quiniou, Véronique Masson and Thomas Guyet
- Contact: Thomas Guyet
- Publication: Mining relevant interval rules

6.8. OCL

One click learning
KEYWORDS: Data mining - Interactivity

FUNCTIONAL DESCRIPTION: This pattern mining software builds a user model preference from implicit feedback of the user in order to automatically choice the type of patterns and algorithms used. The principle builds upon the algorithm introduced by M. Boley et al, "One click mining: interactive local pattern discovery through implicit preference and performance learning". In addition OCL integrates algorithms dealing with temporal series.

- Contact: Laurence Rozé
- URL: https://github.com/Gremarti/OneClickLearning

7. New Results

7.1. Introduction

In this section, we organize our contributions this year along two of our research axes, namely Pattern Mining and Decision Support. These correspond to the contributions that has been accepted for publication this year.

7.1.1. Pattern Mining

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

- *Mining of novel types of patterns.* This includes temporal pattern mining, signature mining, opinion mining in uncertain databases, interval rules, and top-k item-centric mining. All these contributions have been proposed as solutions to problems in the domains of pharmaco-epidemiology, retail databases, biomedical databases, and analysis of speech corpora. We provide more details about these results in Sections 7.2 to 7.9.
- *Data Mining with ASP.* Answer Set Programming is a powerful search tool in combinatorial spaces, which can be naturally ported to pattern mining, as the latter is a specific type of search problem. Our contributions include the application of ASP in the discovery of frequent, constrained, condensed, and rare sequential patterns. Sections 7.11 and 7.12 elaborate on our new research insights.
- *Data Mining for the masses.* In [14], we propose a communication model that bridges knowledge delivery between data miners and domain users in the field of library science. Our model proposes a five-steps process in order to achieve effective knowledge synthesis and delivery of insights to the domain users.

7.1.2. Decision Support

In regards to the axis of decision support, our contributions can be organized in two categories: exploration and diagnosis.

- *Exploration.* We propose two exploration methods in the context of analysis of trajectories and agro-environmental systems. We propose customized data models and resort to data-warehousing and multidimensional data representations to facilitate the querying, and thus the exploration and understanding of the data, for the sake of decision making. Our results in this line are further detailed in Sections 7.13 to 7.15.
- *Diagnosis.* In Section 7.16 we propose a novel method for anomaly detection in time series by resorting to Extreme Value Theory. In addition, [21] offers a formalization of diagnosis based on automata with focus on discrete event systems.

7.2. Discriminant chronicles mining: Application to care pathways analytics

Participants: Yann Dauxais, Thomas Guyet, David Gross-Amblard [Druid], André Happe [Brest University Hospital/REPERES].

Pharmaco-epidemiology (PE) is the study of uses and effects of drugs in well defined populations. As medico-administrative databases cover a large part of the population, they have become very interesting to carry PE studies. Such databases provide longitudinal care pathways in real condition containing timestamped care events, especially drug deliveries. Temporal pattern mining becomes a strategic choice to gain valuable insights about drug uses. We propose DCM [8], [7], a new discriminant temporal pattern mining algorithm. It extracts chronicle patterns that occur more in a studied population than in a control population. We present satisfactory results on the identification of possible associations between hospitalizations for seizure and anti-epileptic drug switches in care pathway of epileptic patients.

A stable release of the DCM algorithm (see Section 6.5) have been deposed to the Program Protection Agency (APP) and is available online.

7.3. Purchase Signatures of Retail Customers

Participants: Clément Gautrais, Peggy Cellier [SemLis], Thomas Guyet, René Quiniou, Alexandre Termier.

In the retail context, there is an increasing need for understanding individual customer behavior in order to personalize marketing actions. We propose the novel concept of customer signature, that identifies a set of important products that the customer refills regularly [10]. Both the set of products and the refilling time periods give new insights on the customer behavior. Our approach is inspired by methods from the domain of sequence segmentation, thus benefiting from efficient exact and approximate algorithms. Experiments on a real massive retail dataset show the applicability of the signatures for understanding individual customers.

7.4. Topic Signatures in Political Campaign Speeches

Participants: Clément Gautrais, Peggy Cellier [SemLis], René Quiniou, Alexandre Termier.

Highlighting the recurrence of topics usage in candidates speeches is a key feature to identify the main ideas of each candidate during a political campaign. In this study [9], we develop a method combining standard topic modeling with signature mining for analyzing topic recurrence in speeches of Clinton and Trump during the 2016 American presidential campaign. The results show that the method extracts automatically the main ideas of each candidate and, in addition, provides information about the evolution of these topics during the campaign.

7.5. Expert Opinion Extraction from a Biomedical Database

Participants: Ahmed Samet, Thomas Guyet, Benjamin Négrevergne, Tien-Tuan Dao, Tuan Nha Hoang, Marie-Christine Ho Ba Tho.

This work tackles the problem of extracting frequent opinions from uncertain databases. This problem is encountered in real-world applications, such as the opinions of medical experts to evaluate the reliability level of biomedical data. We introduce the foundation of an opinion mining approach with the definition of pattern and support measure. The support measure is derived from the commitment definition. In [15], we proposed a new algorithm called OPMINER that extracts the set of frequent opinions modeled as a mass functions. We applied it on a real-world biomedical opinion database. Performance analysis showed that our proposal generated better patterns compared to literature-based methods.

7.6. Mining Relevant Interval Rules

Participants: Philippe Besnard, Thomas Guyet, Véronique Masson, René Quiniou.

Rule mining is a classical data mining task. Numerical rule mining consists of extracting decision rules from a dataset with numerical attributes. In this work, we are interested in extracting a subset of accurate rules, called relevant rules. This selection criteria was introduced by Garriga et al. for categorical attributes [28]. In [13] we extend the method of Garriga et *al.* for mining relevant rules on numerical attributes by extracting interval-based pattern rules. We proposed an algorithm that extracts such rules from numerical datasets using the interval-pattern approach from Kaytoue et *al.* [29]. The algorithm has been implemented and intensively evaluated on real datasets. This study on numerical rules mining leads us to initiate a study about admissible generatizations of examples as rules [18].

7.7. Time Series Rule Matching: Application to Energy Consumption

Participants: Maël Guillemé, Véronique Masson, Laurence Rozé, René Quiniou, Alexandre Termier.

Pattern mining in time series is an important subfield of Data Mining. In various applications, patterns exhibit distortion in time (or time elasticity) that requires using specific distance measures. In this work, we extend an algorithm proposed by Shokoohi et *al.* [35] by improving the performance of rule matching in the detection of energy consumption patterns. Nowadays companies are more and more equipped with sensors in order to trace losses of energy resources. Detecting dysfunctions from time series recorded by these sensors becomes a crucial problem for reducing energy consumption. Locating specific patterns related to dysfunctions in time series requires handling with time elasticity (i.e. distortion in time) of patterns. We propose a detection of predictive rules based on several variations of Dynamic Time Warping (DTW) and show the superiority of subsequence DTW [11]. We study now multivariate time series classification to predict dysfunctions as soon as possible.

7.8. Negative Temporal Sequence Mining

Participants: Katerina Tsesmeli, Thomas Guyet, René Quiniou, Manel Boumghar [EDF R&D], Laurent Pierre [EDF R&D].

Temporal pattern mining is one of the important tasks in the data mining research field. It aims at extracting interesting sequences of occurring events from timestamped event sequences as well as their temporal constraints relating sequence events. Little research has focused on mining sequential patterns with non-occurring (negative) events, though they can bring much value and relevance to extracted patterns. In this context, we are interested in formalizing normal and undesirable situations, that can be defined in terms of negative temporal patterns. We proposed the NTGSP algorithm [17] that extracts frequent sequences with positive and negative events, as well as temporal information about the delay between these events. The method performance has been evaluated on synthetic sequences and on commercial data provided by EDF, a major french power distribution company.

7.9. TopPI: An efficient algorithm for item-centric mining

Participants: Vincent Leroy, Martin Kirchgessner, Alexandre Termier, Sihem Amer-Yahia.

In this paper [6], we introduce item-centric mining, a new semantics for mining long-tailed datasets. Our algorithm, TopPI, finds for each item its top-k most frequent closed itemsets. While most mining algorithms focus on the globally most frequent itemsets, TopPI guarantees that each item is represented in the results, regardless of its frequency in the database.

TopPI allows users to efficiently explore Web data, answering questions such as “what are the k most common sets of songs downloaded together with the ones of my favorite artist?”. When processing retail data consisting of 55 million supermarket receipts, TopPI finds the itemset “milk, puff pastry” that appears 10,315 times, but also “frangipane, puff pastry” and “nori seaweed, wasabi, sushi rice” that occur only 1120 and 163 times, respectively. Our experiments with analysts from the marketing department of our retail partner, demonstrate that item-centric mining discover valuable itemsets. We also show that TopPI can serve as a building-block to approximate complex itemset ranking measures such as the p-value.

Thanks to efficient enumeration and pruning strategies, TopPI avoids the search space explosion induced by mining low support itemsets. We show how TopPI can be parallelized on multi-core architectures and Hadoop clusters. Our experiments on datasets with different characteristics show the superiority of TopPI when compared to standard top-k solutions, and to Parallel FPGrowth, its closest competitor.

7.10. Declarative Sequential Pattern Mining of Care Pathways

Participants: Thomas Guyet, André Happe [Brest University Hospital/REPERES], Yann Dauxais.

Sequential pattern mining algorithms are widely used to explore care pathways database, but they generate a deluge of patterns, mostly redundant or non-informative. Clinicians need tools to express complex mining queries in order to generate less but more significant patterns. These algorithms are not versatile enough to answer complex clinician queries. This work [12] proposes to apply a declarative pattern mining approach based on the Answer Set Programming paradigm. It is exemplified by a pharmaco-epidemiological study investigating the possible association between hospitalization for seizure and antiepileptic drug switch from a French medico-administrative database.

7.11. Efficiency Analysis of ASP Encodings for Sequential Pattern Mining Tasks

Participants: Thomas Guyet, Yves Moinard, René Quiniou, Torsten Schaub.

This study [22] presents the use of Answer Set Programming (ASP) to mine sequential patterns. ASP is a high-level declarative logic programming paradigm that allows for representation of combinatorial and optimization problems, as well as knowledge and reasoning tasks. Thus, ASP is a good candidate for implementing pattern mining with background knowledge, which has been a data mining issue for a long time. We propose encodings of the classical sequential pattern mining tasks within two representations of embeddings (fill-gaps vs skip-gaps) and for various kinds of patterns: frequent, constrained and condensed. We compare the computational performance of these encodings with each other to get a good insight into the efficiency of ASP encodings. The results show that the fill-gaps strategy is better on real problems due to lower memory consumption. Finally, compared to a constraint programming approach (CPSM), another declarative programming paradigm, our proposal showed comparable performance.

7.12. Mining Rare Sequential Patterns with ASP

Participants: Ahmed Samet, Thomas Guyet, Benjamin Négrevergne.

This work [20] presents an approach of meaningful rare sequential pattern mining based on the declarative programming paradigm of Answer Set Programming (ASP). The setting of rare sequential pattern mining is introduced. Our ASP approach provides an easy manner to encode expert constraints on expected patterns to cope with the huge amount of meaningless rare patterns. Encodings are presented and quantitatively compared to a procedural baseline. An application on care pathways analysis illustrates the applicability of our method in the encoding of constraints provided by experts.

7.13. From Medico-administrative Databases Analysis to Care Trajectories Analytics: An Example with the French SNDS

Participants: Erwan Drezen [Rennes University Hospital/REPERES], Thomas Guyet, André Happe [Brest University Hospital/REPERES].

Medico-administrative data like SNDS (Système National de Données de Santé) are not collected initially for epidemiological purposes. Moreover, the data model and the tools proposed to SNDS users make their in-depth exploitation difficult. We propose a data model, called the ePEPS model, based on health care trajectories to provide a medical view of raw data [4]. A data abstraction process enables the clinician to have an intuitive medical view of raw data and to design a study-specific views. This view is based on a generic model of care trajectory, i.e. a sequence of timestamped medical events for a given patient. This model is combined with tools to manipulate care trajectories efficiently.

7.14. A Data Warehouse to Explore Multidimensional Simulated Data from a Spatially Distributed Agro-hydrological Model to Improve Catchment Nitrogen Management

Participants: Tassadit Bouadi, Marie-Odile Cordier, Pierre Moreau, Jordy Salmon-Monviola, Chantal Gascuel-Odoux.

Spatially distributed agro-hydrological models allow researchers and stakeholders to represent, understand and formulate hypotheses about the functioning of agro-environmental systems and to predict their evolution. These models have guided agricultural management by simulating effects of landscape structure, farming system changes and their spatial arrangement on stream water quality. Such models generate many intermediate results that should be managed, analyzed and transformed into usable information. We introduce [3] a data warehouse (N-Catch) built to store and analyze simulation data from the spatially distributed agro-hydrological model TNT2. We present scientific challenges to and tools for building data warehouses and describe the three dimensions of N-Catch: space, time and an original hierarchical description of cropping systems. We show how to use OLAP to explore and extract all kinds of useful high-level information by aggregating the data along these three dimensions. We also show how to facilitate exploration of the spatial dimension by coupling N-Catch with GIS. Such tool constitutes an efficient interface between science and society, simulation remaining a research activity, exploration of the results becoming an easy task accessible for a large audience.

7.15. Extended Automata for Temporal Planning of Interacting Agents

Participants: Christine Largouët, Omar Krichen, Yulong Zhao.

In this paper [5], we consider the planning problem for a system represented as a set of interacting agents evolving along time according to explicit timing constraints. Given a goal, the planning problem is to find the sequence of actions such that the system reaches the goal state in a limited time and in an optimal manner, assuming actions have a cost. In our approach, the planning problem is based on model-checking and controller synthesis techniques while the goal is defined using temporal logic. Each agent of the system is represented using the formalism of Priced Timed Game Automata (PTGA). PTGA is an extension of Timed Automata that allows the representation of cost on actions and the definition of uncontrollable actions. We define a planning algorithm that computes the best strategy to achieve a goal. To experiment our approach, we extend the classical Transport Domain with timing constraints, cost on actions and uncontrollable actions. The planning algorithm is finally presented on a marine ecosystem management problem.

7.16. Anomaly Detection in Streams with Extreme Value Theory

Participants: Alban Siffer [EMSEC], Pierre-Alain Fouque [EMSEC], Christine Largouët, Alexandre Termier.

Anomaly detection in time series has attracted considerable attention due to its importance in many real-world applications including intrusion detection, energy management and finance. Most approaches for detecting outliers rely on either manually set thresholds or assumptions on the distribution of data. In [16], we propose a new approach to detect outliers in streaming univariate time series based on Extreme Value Theory that does not require to handpick thresholds and makes no assumption on the distribution: the main parameter is only the risk, controlling the number of false positives. Our approach can be used for outlier detection, but more generally for automatically setting thresholds, making it useful in wide number of situations. We also test our algorithms on various real-world datasets which confirm the soundness and efficiency of our methods.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. ITRAMI: Interactive Trace Mining

Participant: Alexandre Termier.

ITRAMI is a Nano2017 project. Such projects are designed to support joint research efforts between STMicrolectronics and academic partners in the domain of embedded systems. Alexandre Termier is the PI of this project whose goal is to design novel data mining methods for interactive analysis of execution traces. Such methods aim at considerably reducing the time that STMicroelectronics developers spend at understanding, debugging and profiling applications running on STMicrolectronics chips. The work is done at University Grenoble Alps, in collaboration with LACODAM researchers. Two contractual staff members are working on the project in Grenoble: Willy Ugarte as postdoc, and Soumaya Ben Alouane as engineer.

8.1.2. Hyptser: Hybrid Prediction of Time-Series

Participants: Thomas Guyet, Vincent Lemaire [Orange Labs], Simon Malinowski [LinkMedia].

HYPTSER is a project funded by the Gaspard Monge Program for Optimisation and Operational Research (PGMO). It is dedicated to the development of innovative methods for predictions in time series. In the field of machine learning, *ensemble methods* have gained popularity in the last years. These methods combine several algorithms that solve the same task in order to improve the performance of the outcome. Two main families of ensemble methods can be found in the literature : The first family makes use of different models and combine their results a posteriori. The methods Bagging and Boosting are examples of methods in this family [26], [34]. The second family is based on a smart selection of the local algorithms in order to create a global hybrid algorithm. Logistic Model Tree [30] or Extreme Learning Machine Tree [36] are examples of such hybrid algorithms. In this project, starting at the end of 2017 for one year, we envision to explore the second family of methods in order to analyze how efficiently hybrid models can perform on the task of time series prediction. We plan to apply these methods to predict resource usage for cloud computing (CPU, memory) so as to minimize their infrastructure.

8.1.3. Particular Contract of the Strategic Action EDF/Inria

Participants: Manel Boumghar [EDF R&D], Laurent Pierre [EDF R&D], Thomas Guyet, René Quiniou.

The analysis of customer pathways has become a strategic issue for many businesses. The interaction traces left by clients when connecting to the customer services can be combined with data from other communication channels (phone, web form, e-mail, mail, fax, SMS, shop, etc.) and allow to analyse the customer pathways in details.

Pattern mining tools are able to extract the frequent customer behaviors in very large databases of client pathways. Nevertheless, taking into account the duration and the delay between the customer actions in the mining remains a challenge. The objective of this one-year contract was to design and develop a frequent mining tool that accounts for temporal patterns with negations for analysis of multichannel customer pathways. In this line, we developed and implemented the NTGSP algorithm [17].

8.2. Bilateral Grants with Industry

Maël Guillemé has obtained a CIFRE PhD grant with the Energiency startup, supervised by V. Masson and L. Rozé. The goal of Maël's thesis is to propose new approaches to improve industrial energy performance by integrating both numerical and symbolic attributes. An M2 internship from 2016 explored an approach based on an algorithm proposed by Shokoohi and al,. and proposed several improvements: avoid data normalisation, detect patterns as fast as possible, enhance functions like distance and score.

Another CIFRE thesis has started, this time with the Amossys company, which specializes in cyber-security. This is the PhD of Alban Siffer, located in the EMSEC team of IRISA and co-supervised between EMSEC (P.A. Fouque) and LACODAM (A. Termier, C. Largouët). The goal of this PhD is to propose new methods for intrusion detection in networks. The novel insight is to consider only IP flow as input (metadata of packets and not packet contents) and detect intrusion via unusual traffic patterns.

On October 2017, Colin Leverger started a thesis funded by Orange and co-supervised between Orange Labs (R. Marguerie), LACODAM (A. Termier, T. Guyet) and LinkMedia (S. Malinowski). The goal of this thesis is to propose new methods to forecast time series in order to support capacity planning tasks.

Elisa Fromont is still involved in the supervision of two PhD students through her former employer: the University of Saint-Etienne. One of the students is Guillaume Metzler, who works with the sponsorship of the Blitz company on bank fraud detection. On the other hand, Kevin Bascol (financed by a FUI project) works in collaboration with Bluecime (Grenoble) and works on improving ski-lift security.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. SePaDec: Declarative approaches for Sequential Pattern mining

Participants: Benjamin Negrevergne, Thomas Guyet, Ahmed Samet, Alexandre Termier.

The SePaDec project is funded by the Region Bretagne. During the execution of this project we explored the application of declarative pattern mining (specifically ASP) in the field of care pathway analysis. The goal was to model domain knowledge to enrich raw data with medical expert knowledge and to develop a toolbox that smoothly integrates both expert knowledge and declarative pattern mining.

We developed a new approach for mining rare sequential mining with ASP [20] and we also proposed a general framework based on ASP for flexibly mine care pathways [12].

9.2. National Initiatives

9.2.1. ANR

9.2.1.1. #DigitAg: Digital agriculture

Participants: Alexandre Termier, Véronique Masson, Christine Largouët, Anne-Isabelle Graux.

#DigitAg is a "Convergence Institute" dedicated to the increasing importance of digital techniques in agriculture. Its goal is twofold: First, make innovative research on the use of digital techniques in agriculture in order to improve competitiveness, preserve the environment, and offer correct living conditions to farmers. Second, prepare future farmers and agricultural policy makers to successfully exploit such technologies.

While #DigitAg is based on Montpellier, Rennes is a satellite of the institute focused on cattle farming. LACODAM is involved in the "data mining" challenge of the institute, that A. Termier co-leads. He is also the representative of Inria in the steering comittee of the institute.

The interest for the team is to design novel methods to analyze and represent agricultural data, which are challenging because they are both heterogeneous and multi-scale (both spatial and temporal).

9.2.2. National Platforms

9.2.2.1. PEPS: Pharmaco-epidemiology for Health Products

Participants: Yann Dauxais, Thomas Guyet, Véronique Masson, René Quiniou, Ahmed Samet.

The PEPS project (Pharmaco-epidemiology des Produits de Santé) is funded by the ANSM (National Agency for Health Security). The project leader is E. Oger from the clinical investigation center CIC-1414 INSERM/CHU Rennes. The other partners located in Rennes are the Institute of Research and Technology (IRT), B<>Com, EHESP and the LTSI. The project started in January 2015 and is funded for 4 years.

The PEPS project consists of two parts: a set of clinical studies and a research program dedicated to the development of innovative tools for pharmaco-epidemiological studies with medico-administrative databases.

Our contribution to this project will be to propose pattern mining algorithms and reasoning techniques to analyse the typical care pathways of specific groups of insured patients. This year we worked on the design and development of the DCM algorithm [8], [7] to mine patterns on care pathways.

9.3. International Research Visitors

9.3.1. Internships

This year, we hosted Scarlett Kelly, a student of Dalhousie University (Canada) from May to the end of August. Her internship was funded by a joint Mitacs Globalink (Canada) / Inria grant. Scarlett Kelly is a student of social sciences, thus she has a different profile that the computer science students who usually do internships at LACODAM. We were interested in such profile in order to gain a critical view on the current approaches of *interactive data mining*. Scarlett quickly picked up the literature of the domain, and could write a report and make interesting propositions that were unexpected from a computer science point of view, i.e., introduce a specially trained "data liaison" person between practitioners and data scientists. Her proposition led to a paper [14] accepted at the HICSS conference (an IT conference ranked “A” at CORE2017).

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

Organization co-chair (T. Guyet) of the first conference on “usage of the medico-administrative database of french insured for health research” (http://www.rennes-donnees-sante-2017.fr/)

Organization chair (T. Guyet) of GAST workshop at EGC 2018 and at EGC 2017 (https://gt-gast.irisa.fr/)

Organization co-chair (T. Guyet) of the technical track “Knowledge Extraction from Geographical Data” of the 33rd ACM/SIGAPP Symposium On Applied Computing

Webmaster (L. Galárraga) of the ICDE 2018 conference (https://icde2018.org/)

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

AAAI Conference on Artificial Intelligence 2017, 2018 (T. Guyet)

Conférence Nationale sur les Applications Pratiques de l’Intelligence Artificielle (APIA) 2017 (C. Largouët)

IEEE International Conference on Data Science and Advanced Analytics (DSAA) 2017 (A. Termier)

European Conference Dedicated to the Future Use of ICT in the Agri-food Sector, Bioresource and Biomass sector (EFITA) 2017 (A. Termier, C. Largouët)

Extraction et Gestion de Connaissances (EGC) 2018 (T. Guyet, R. Quiniou, A. Termier).

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (EMLCPKDD) 2017 (E. Fromont)

International Conference on Data Mining (ICDM) 2017 (A. Termier)

Conference on Intelligent Data Analysis (IDA) 2017 (E. Fromont)

International Joint Conference on Artificial Intelligence (IJCAI) 2017 (T. Guyet, A. Termier)

Symposium on Information Management and Big Data (SimBig) 2017 (T. Guyet)

10.1.2.2. Reviewer

Y. Dauxais, C. Gautrais: IJCAI 2017, ICDM 2017, EGC 2018

E. Fromont: ECMLPKDD 2017, AAAI 2017, IDA2017, ECMLPKDD demo track 2017

L. Galárraga: The Web Conference 2018, WebDB workshop 2017, AKBC workshop 2017

T. Guyet: IJCAI 2017, ICDM 2017, AAAI 2018, AAAI 2017

R. Quiniou: IJCAI 2017, ICDM 2017, AAAI 2018

A. Termier: EuroPar 2017

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

E. Fromont: Machine Learning and Data Mining ECMLPKDD special issue

T. Guyet: Revue d'Intelligence Artificielle (RIA)

10.1.3.2. Reviewer - Reviewing Activities

E. Fromont: Data Mining and Knowledge Discovery, Machine Learning Journal

L. Galárraga: Data Mining and Knowledge Discovery, Artificial Intelligence Review, Semantic Web Journal

C. Gautrais: Data Mining and Knowledge Discovery

T. Guyet: Journal of Biomedical Informatics, Journal of Intelligent Information Systems, Artificial Intelligence Review

C. Largouët: Natural Computing, International Journal of Agricultural and Environmental Information Systems (IJAEIS)

A. Termier: Knowledge and Information Systems, The Very Large DataBases Journal

10.1.4. Invited Talks

E. Fromont gave an invited talk for the Machine Learning seminar of the KULEUVEN University in May 2017, Ostende, Belgium.

A. Termier gave an invited talk at KU Leuven on 09/06/2017.

E. Fromont gave an invited talk "Big Data and Business" summer school oragnised by the IRIXYS consortium and ATOS, Chiemsee, Germany.

A. Termier gave an invited talk at the esaconnect event of ESA agriculture school (Angers), on 26/10/2017.

L. Galárraga gave an invited talk at the LTCI Data Science Seminar (https://ltci.telecom-paristech.fr/data-science-seminar/) at Télécom ParisTech on 16/11/2017.

T. Guyet gave an invited talk at the SPECIF Campus association (https://www.specifcampus.fr/les-supports-de-la-journee-du-24-novembre-2016/) on 24/11/2017.

10.1.5. Leadership within the Scientific Community

A. Termier is the representative of Inria for the #DigitAg Convergence Institute

10.1.6. Scientific Expertise

Evaluation of one project proposal for IFREMER: T. Guyet

Evaluation of two projects proposals for ANR: A. Termier

Comité d'évaluation scientifique ANR CE23 - Données, Connaissances, Big data, Contenus multimédias, Intelligence Artificielle: E. Fromont

Groupe de travail national "infrastructure de recherche pour l'Intelligence Artificielle" (demandé par Allistene): E. Fromont

10.1.7. Research Administration

Member of INRA CEI (Engineers Evaluation Committee): T. Guyet

Member of the scientific board of department EA of INRA: A. Termier

Member of the scientific board of Agrocampus Ouest - COREGE: C. Largouët

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

A. Termier, DMV Module: Data Mining and Visualization, 18h, Master 2, Istic, Univ. Rennes 1

L. Bonneau and C. Largouët, DataViz with R, 10h, Master Datascience, Agrocampus Ouest Rennes

L. Bonneau and C. Largouët, Computer Science for BigData, 30h, Master Datascience, Agrocampus Ouest Rennes

C. Largouët, Scientific Programming, Master 1, Agrocampus Ouest Rennes

C. Largouët, Data Management, Master 1, Agrocampus Ouest Rennes

10.2.2. Supervision

PhD in progress: Maël Guillemé, “New data mining approaches for improving energy consumption in factory”, 03/10/2016, A. Termier, V. Masson, L. Rozé and R. Quiniou

PhD in progress: Clément Gautrais, “Mining massive data from client purchases”, 01/10/2015, A. Termier, P. Cellier, T. Guyet and R. Quiniou

PhD in progress: Yann Dauxais, “Query-language for care-pathway mining and analysis”, 01/02/2015, D. Gross-Amblard, T. Guyet, A. Happe

PhD in progress: Colin Leverger, “Cluster resources optimization through forecasting and management of metric time series”, 01/10/2017, T. Guyet, S. Malinowski, R. Marguerie, A. Termier

PhD in progress: Alban Siffer, “DataMining approaches for cyber attack detection”, 05/2016, P.-A. Fouque, A. Termier, C. Largouët

PhD in progress: Kevin Fauvel, “Using data mining techniques for improving dairy management”, 10/2017, V. Masson, P. Faverdin, A. Termier

PhD in progress: Raphaël Gauthier, "Modelling of nutrient utilization and precision feeding of lactating sows", 11/2017, C. Largouët, J.-Y. Dourmad

10.2.3. Juries

Committee member of Georges Nassopoulos’ PhD defense on 22/05/2017 (Université de Nantes): R. Quiniou

Committee member of Chemseddine NABTI’s PhD defense (Université Claude Bernard Lyon 1): A. Termier

Reviewer of Vladimir Dzyuba’s Phd (KU Leuven): A. Termier

Committee member of Hoang Son Pham’s PhD defense (Université de Rennes 1): A. Termier

Thesis advisory committee member of Jean Coquet (Univ. Rennes 1): A. Termier

Thesis advisory committee member of Mathilde Chen (INRA): A. Termier

Thesis advisory committee member of Benoit Bellot (INRA/IGEPP): T. Guyet

Thesis advisory committee member of Zhi Cheng (UNC/PPME): T. Guyet

Thesis advisory committee member of Vanel Siyou (Univ. Blaise Pascal/LIMOS): T. Guyet

Committee member of Asma Dachraoui’s PhD defense on 31/01/2017 (ABIES/AgroParisTech): T. Guyet

Member of the associate professor hiring commitee on 12/05/2017 (LINK/AgroParisTech) : T. Guyet

Committee member of Jordy Salmon-Monviola’s PhD defense on 05/04/2017 (AgroCampus Ouest): V. Masson

Committee member (as co-supervisor) of Damien Fourure on 12/12/2017 (Univ. Saint-Etienne) : E Fromont

Committee member of Van-Tinh TRAN, (University of Lyon) on 11/07/2017 : E. Fromont

Committee member of Maxime Gasse, (University of Lyon on 13/01/2017: E. Fromont

Committee member of Pauline Wauquier, (University of Lille 3) on 29/05/2017 : E. Fromont

Committee member of Ouadie Gharroudi, (University of Lyon) on 21/12/2017 : E. Fromont

Thesis advisory committee member of Mohamed Ali Hammal on 30/04/2017 (University of Lyon), Maxime CHABERT on 11/04/2017 and 2/06/2017 (University of Lyon), Tuan Nguyen on 21/06/2017 (Université Savoie Mont Blanc), Jean-Jacques Ponciano on 17/05/2017 (Univ. Saint-Etienne), Julien Tissier, Jordan Fréry, Carlos Arango, Dennis Diefenbach on 12/07/2017 (Univ. Saint-Etienne): E. Fromont

10.3. Popularization

M.-O. Cordier is editorial board member of Interstices webzine.

C. Largouët and L. Galárraga participated in a science diffusion talk at the educational institution "Assomption" on 20/10/2017.

L. Galárraga participated as AI specialist in a discussion forum at the general election assembly of the "Pôle Images et Reseaux" on 23/11/2017.

T. Guyet gave a public talk about Artificial Intelligence and Data Mining in MDA Rennes (FranceIA events) on 01/03/2017.

A. Termier gave a public talk about Artificial Intelligence at the Esaconnect event on 26/10/2017 https://www.chaire-mutations-agricoles.com/evenements/esaconnect-2017/programme/panorama-de-lintelligence-artificielle-en-2017-alexandre-termier/link.

A. Termier was interviewed for the December podcast of the Interstices webzine.

11. Bibliography

Major publications by the team in recent years

[1] M. Gebser, T. Guyet, R. Quiniou, J. Romero, T. Schaub. *Knowledge-based Sequence Mining with ASP*, in "IJCAI 2016- 25th International joint conference on artificial intelligence", New-york, United States, AAAI, July 2016, 8 p. , https://hal.inria.fr/hal-01327363

[2] A. Siffer, P.-A. Fouque, A. Termier, C. Largouët. *Anomaly Detection in Streams with Extreme Value Theory*, in "KDD 2017 - Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", Halifax, Canada, August 2017 [*DOI :* 10.1145/3097983.3098144], https://hal.archives-ouvertes.fr/hal-01640325

Publications of the year

Articles in International Peer-Reviewed Journals

[3] T. Bouadi, M.-O. Cordier, P. Moreau, R. Quiniou, J. Salmon-Monviola, C. Gascuel-Odoux. *A data warehouse to explore multidimensional simulated data from a spatially distributed agro-hydrological model to improve catchment nitrogen management*, in "Environmental Modelling and Software", November 2017, vol. 97, pp. 229 - 242 [*DOI :* 10.1016/j.envsoft.2017.07.019], https://hal.inria.fr/hal-01597840

[4] E. DREZEN, T. GUYET, A. HAPPE. *From medico-administrative databases analysis to care trajectories analytics: an example with the French SNDS*, in "Fundamental and Clinical Pharmacology", September 2017 [*DOI :* 10.1111/FCP.12323], https://hal.inria.fr/hal-01631802

[5] C. LARGOUËT, O. KRICHEN, Y. ZHAO. *Extended Automata for Temporal Planning of Interacting Agents*, in "International Journal of Monitoring and Surveillance Technologies Research", April 2017, vol. 5, n^{o} 1, pp. 30 - 48 [*DOI :* 10.4018/IJMSTR.2017010102], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01640137

[6] V. LEROY, M. KIRCHGESSNER, A. TERMIER, S. AMER-YAHIA. *TopPI: An efficient algorithm for item-centric mining*, in "Information Systems", 2017, vol. 64, pp. 104 - 118 [*DOI :* 10.1016/J.IS.2016.09.001], https://hal.archives-ouvertes.fr/hal-01479067

International Conferences with Proceedings

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

[8] Y. DAUXAIS, T. GUYET, D. GROSS-AMBLARD, A. HAPPE. *Discriminant chronicles mining: Application to care pathways analytics*, in "Artificial Intelligence in Medicine", Vienna, Austria, 16th Conference on Artificial Intelligence in Medicine, June 2017, https://arxiv.org/abs/1709.03309 [*DOI :* 10.1007/978-3-319-59758-46], https://hal.archives-ouvertes.fr/hal-01568929

[9] C. GAUTRAIS, P. CELLIER, R. QUINIOU, A. TERMIER. *Topic Signatures in Political Campaign Speeches*, in "EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing", Copenhagen, Denmark, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, September 2017, https://hal.archives-ouvertes.fr/hal-01640498

[10] C. GAUTRAIS, R. QUINIOU, P. CELLIER, T. GUYET, A. TERMIER. *Purchase Signatures of Retail Customers*, in "PAKDD 2017 - The Pacific-Asia Conference on Knowledge Discovery and Data Mining", Jeju, South Korea, Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 2017, https://hal.archives-ouvertes.fr/hal-01639795

[11] M. GUILLEME, L. ROZÉ, V. MASSON, C. CARTON, R. QUINIOU, A. TERMIER. *Improving time-series rule matching performance for detecting energy consumption patterns*, in "DARE 2017 - 5th International Workshop on Data Analytics for Renewable Energy Integration", Skopje, Macedonia, Springer, September 2017, vol. 10691, pp. 59-71 [*DOI :* 10.1007/978-3-319-71643-5_6], https://hal.inria.fr/hal-01654890

[12] T. GUYET, A. HAPPE, Y. DAUXAIS. *Declarative Sequential Pattern Mining of Care Pathways*, in "Conference on Artificial Intelligence in Medicine in Europe", Vienna, Austria, 16th Conference on Artificial Intelligence in Medicine, June 2017, vol. 24, pp. 1161 - 266, https://arxiv.org/abs/1707.08342 [*DOI :* 10.1007/978-3-319-59758-4_29], https://hal.inria.fr/hal-01569023

[13] T. GUYET, R. QUINIOU, V. MASSON. *Mining relevant interval rules*, in "International Conference on Formal Concept Analysis", Rennes, France, Supplementary proceedings of International Conference on Formal Concept Analysis (ICFCA), June 2017, https://arxiv.org/abs/1709.03267 , https://hal.inria.fr/hal-01584981

[14] S. KELLY. *A Communication Model that Bridges Knowledge Delivery between Data Miners and Domain Users* , in "51th Hawaii International Conference on System Sciences (HICSS)", Hawaii, United States, January 2018, https://hal.archives-ouvertes.fr/hal-01651737

[15] A. SAMET, T. GUYET, B. NEGREVERGNE, T.-T. DAO, T. NHA HOANG, M.-C. HO BA THO. *Expert Opinion Extraction from a Biomedical Database*, in "Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)", Lugano, Switzerland, Proceedings of 14th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, July 2017, vol. 31, n^{o} LNCS 10369, pp. 1 - 12, https://arxiv.org/abs/1709.03270 [*DOI :* 10.1016/S0888-613X(02)00066-X], https://hal.inria.fr/hal-01584984

[16] A. SIFFER, P.-A. FOUQUE, A. TERMIER, C. LARGOUËT. *Anomaly Detection in Streams with Extreme Value Theory*, in "KDD 2017 - Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", Halifax, Canada, August 2017 [*DOI :* 10.1145/3097983.3098144], https://hal.archives-ouvertes.fr/hal-01640325

[17] K. TSESMELI, M. BOUMGHAR, T. GUYET, R. QUINIOU, L. PIERRE. *Fouille de motifs temporels négatifs*, in "EGC 2018 - 18ème Conférence Internationale sur l'Extraction et la Gestion des Connaissances", Paris, France, January 2018, pp. 1-6, https://hal.inria.fr/hal-01657540

Conferences without Proceedings

[18] P. BESNARD, T. GUYET, V. MASSON. *Admissible generalizations of examples as rules*, in "11e Journées d'Intelligence Artificielle Fondamentale", Caen, France, July 2017, https://hal.archives-ouvertes.fr/hal-01576047

[19] C. GAUTRAIS, Y. DAUXAIS, M. GUILLEME. *Multi-Plant Photovoltaic Energy Forecasting Challenge: Second place solution*, in "Discovery Challenges co-located with European Conference on Machine Learning - Principle and Practice of Knowledge Discovery in Database", Skopje, Macedonia, September 2017, https://hal.archives-ouvertes.fr/hal-01639813

[20] A. SAMET, T. GUYET, B. NEGREVERGNE. *Mining rare sequential patterns with ASP*, in "ILP 2017 - 27th International Conference on Inductive Logic Programming", Orléans, France, September 2017, https://hal.archives-ouvertes.fr/hal-01569582

Scientific Books (or Scientific Book chapters)

[21] M.-O. CORDIER, P. DAGUE, Y. PENCOLÉ, L. TRAVÉ-MASSUYÈS. *Diagnosis and supervision: model-based approaches*, in "A guided tour of artificial intelligence research", H. P. PIERRE MARQUIS (editor), Knowledge representation and reasoning, Springer, 2018, n^{o} 1, https://hal.archives-ouvertes.fr/hal-01483436

[22] T. GUYET, Y. MOINARD, R. QUINIOU, T. SCHAUB. *Efficiency Analysis of ASP Encodings for Sequential Pattern Mining Tasks*, in "Advances in Knowledge Discovery and Management", B. PINAUD, F. GUILLET, B. CREMILLEUX, C. DE RUNZ (editors), Springer, October 2017, vol. 7, pp. 41–81, https://arxiv.org/abs/1711.05090 , https://hal.inria.fr/hal-01631879

[23] A. NAPOLI, A. TERMIER. *La fouille de données*, in "Les Big Data à découvert", M. BOUZEGHOUB, R. MOSSERI (editors), CNRS Editions, 2017, pp. 1-3, https://hal.inria.fr/hal-01673437

Scientific Popularization

[24] M.-O. CORDIER. *Regard sur « Le mythe de la Singularité. Faut-il craindre l'intelligence artificielle ? »*, in "Interstices", June 2017, https://hal.inria.fr/hal-01616345

[25] A. TERMIER, J. JONGWANE. *Vers une démocratisation des outils pour l'exploration de données ?*, in "Interstices", December 2017, https://hal.inria.fr/hal-01688785

References in notes

[26] L. BREIMAN. *Bagging predictors*, in "Machine Learning", Aug 1996, vol. 24, n^{o} 2, pp. 123–140, https://doi.org/10.1007/BF00058655

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

[28] G. C. GARRIGA, P. KRALJ, N. LAVRAC. *Closed Sets for Labeled Data*, in "Journal of Machine Learning Research", 2008, vol. 9, pp. 559–580, http://doi.acm.org/10.1145/1390681.1390700

[29] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI. *Revisiting numerical pattern mining with formal concept analysis*, in "IJCAI Proceedings-International Joint Conference on Artificial Intelligence", 2011, vol. 22, n^{o} 1, 1342 p.

[30] N. LANDWEHR, M. HALL, E. FRANK. *Logistic model trees*, in "Machine learning", 2005, vol. 59, n^{o} 1-2, pp. 161–205

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

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

[33] E. POLARD, E. NOWAK, A. HAPPE, A. BIRABEN, E. OGER. *Brand name to generic substitution of antiepileptic drugs does not lead to seizure-related hospitalization: a population-based case-crossover study*, in "Pharmacoepidemiology and drug safety", 2015, vol. 24, n^{o} 11, pp. 1161–1169

[34] R. E. SCHAPIRE. *The Boosting Approach to Machine Learning: An Overview*, Springer New York, New York, NY, 2003, pp. 149–171

[35] M. SHOKOOHI-YEKTA, Y. CHEN, B. CAMPANA, B. HU, J. ZAKARIA, E. KEOGH. *Discovery of meaningful rules in time series*, in "Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", ACM, 2015, pp. 1085–1094

[36] R. WANG, Y.-L. HE, C.-Y. CHOW, F.-F. OU, J. ZHANG. *Learning ELM-tree from big data based on uncertainty reduction*, in "Fuzzy Sets and Systems", 2015, vol. 258, pp. 79–100