Activity Report 2018

Team LACODAM

Large scale Collaborative Data Mining

Joint team with Inria Rennes – Bretagne Atlantique

D7 – Data and Knowledge Management

Table of contents

Project-Team LACODAM

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

Keywords:

Computer Science and Digital Science:

A2.1.5. - Constraint programming
A3.1.1. - Modeling, representation
A3.1.6. - Query optimization
A3.1.11. - Structured data
A3.2.1. - Knowledge bases
A3.2.2. - Knowledge extraction, cleaning
A3.2.3. - Inference
A3.2.4. - Semantic Web
A3.3. - Data and knowledge analysis
A3.3.1. - On-line analytical processing
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A9.1. - Knowledge
A9.2. - Machine learning
A9.3. - Signal analysis
A9.6. - Decision support
A9.7. - AI algorithmics
A9.8. - Reasoning

Other Research Topics and Application Domains:

B1.2.2. - Cognitive science
B2.3. - Epidemiology
B2.4.1. - Pharmaco kinetics and dynamics
B3.5. - Agronomy
B3.6. - Ecology
B3.6.1. - Biodiversity

1. Team, Visitors, External Collaborators

Research Scientists

Louis Bonneau de Beaufort [Ecole nationale supérieure agronomique de Rennes, Researcher]
Luis Galárraga Del Prado [Inria, Researcher]
René Quiniou [Inria, Researcher, until Mar 2018]
Torsten Schaub [Univ de Postdam, Chair, until Sep 2018]

Faculty Members

Marie-Odile Cordier [Univ de Rennes I, Emeritus, until Aug 2018, HDR]
Elisa Fromont [Univ de Rennes I, Professor, HDR]
Thomas Guyet [Agrocampus Ouest, Associate Professor]
Christine Largouët [Agrocampus Ouest, Associate Professor]
Véronique Masson [Univ de Rennes I, Associate Professor]
Laurence Rozé [INSA Rennes, Associate Professor]
Alexandre Termier [Team Leader, Univ de Rennes I, Professor, HDR]

PhD Students

Erwan Bourrand [Inria, from Sep 2018 until Nov 2018]
Yann Dauxais [Univ de Rennes I, until Apr 2018]
Kevin Fauvel [Inria]
Clément Gautrais [Univ de Rennes I, until Sep 2018]
Maël Guillemé [Energiency]
Colin Leverger [Orange Labs]
Gregory Martin [Inria, from Nov 2018]
Anh Duong Nguyen [Vietnamese Ministry of Education and Training, from Mar 2018]
Alban Siffer [AMOSSYS]
Yichang Wang [Univ de Rennes I, from Apr 2018]
Heng Zhang [ATERMES, from Dec 2018]

Interns

Marine Antigny [Univ de Rennes I, from Jun 2018 until Aug 2018]
Erwan Bourrand [Inria, from Mar 2018 until Jul 2018]
Nicolas Buton [Univ de Rennes I, from May 2018 until Jul 2018]
Folatchegoun Chabi [Inria, from May 2018 until Aug 2018]
Julien Delaunay [Univ de Rennes I, from Apr 2018 until Aug 2018]
Manon Derocles [Univ de Rennes I, from May 2018 until Aug 2018]
Diane Valerie Kouadio Comoe [Univ de Rennes I, from May 2018 until Aug 2018]
Devang Kulshreshtha [Univ de Rennes I, from May 2018 until Aug 2018]
Theo Losekoot [Univ de Rennes I, from May 2018 until Jul 2018]
Gregory Martin [Groupe PSA, from Mar 2018 until Sep 2018, contractuel Inria depuis Oct 2018]
Etienne Menager [Ecole normale supérieure de Rennes, from May 2018 until Jul 2018]
Loic Mosser [Ecole normale supérieure de Rennes, from May 2018 until Jul 2018]
Gregoire Pacreau [Ecole normale supérieure de Rennes, from May 2018 until Jul 2018]
Olivier Pelgrin [Inria, from Feb 2018 until Aug 2018]
Muaz Twaty [Univ Saint-Étienne, from Apr 2018 until Jun 2018]

Administrative Assistant

Marie-Noëlle Georgeault [Inria]

Visiting Scientist

Alexandre Sahuguede [Univ Paul Sabatier, from Nov 2018]

External Collaborators

Johanne Bakalara [Univ de Rennes I, from Oct 2018]
Philippe Besnard [CNRS, HDR]
Romaric Gaudel [Ecole nationale de la statistique et de l'analyse de l'information, from Nov 2018]
Raphael Gauthier [INRA]
Anne-Isabelle Graux [INRA]

2. Overall Objectives

2.1. Overall Objectives

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

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.

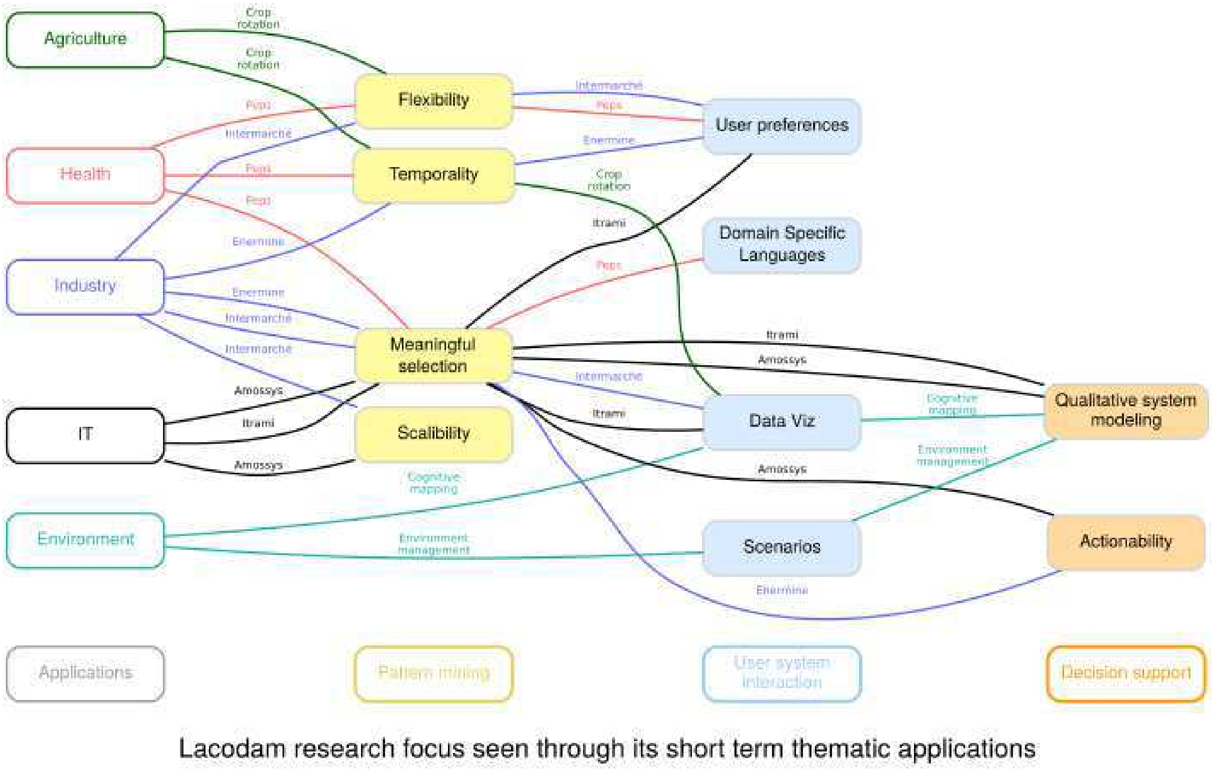

Figure 1. Lacodam research topics organized by axis and application

3.2. Pattern mining algorithms

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

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

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

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

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

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

3.3. User/system interaction

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

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

make significant advances in the other field (this is detailed in [Keim 2010]). Thus, our strategy is to establish collaborations with prominent data visualization teams for this line of research, with a long term goal to recruit a specialist in data visualization if the opportunity arises.

3.4. Decision support

Patterns have proved to be quite useful for decision-aid. Predictive sequential patterns, to give an example, have a direct application in diagnosis. Itemsets and contrast patterns can be used for interpretable machine learning (ML). In regards to diagnosis, LACODAM inherits, from the former DREAM team, a strong background in decision support systems with internationally recognized expertise in this field. This 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 (e.g., a ML classifier), a living being, etc. In relation to interpretable machine learning (ML), this subfield is concerned with the conception of models whose answers are understandable by users. This can be achieved by inducing inherently white-box models from data such as rule-based classifiers/regressors, or by mining rules and explanations from black-box models. The latter setting is quite common due to the high accuracy of black-box models compared to natively interpretable models. Pattern mining is a powerful tool to mine explanations from black-box systems. Those explanations can be used to diagnose biases in systems, either to debug and improve the model, or to generate trust in the verdicts of intelligent software agents.

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

Three main research sub-axes are investigated in LACODAM:

- *Diagnosis-based approaches.* We are exploring how to integrate knowledge found from pattern mining approaches, possibly with the help of interactive methods, into the qualitative models. The goal of such work is to automate as much as possible the construction of prediction models, which can require a lot of human effort.
- *Actionable patterns and rules.* In many settings of “exploratory data mining”, the actual interestingness of a pattern is hard to assess, as it may be subjective. However, for some applications there are well defined measures of interestingness and applicability for patterns. Patterns and rules that can lead to actual actions –that are relevant to the user– are called "actionable patterns" and are of vital importance to industrial settings.
- *Mining explanations from ML systems.* Interpretable ML and AI is a current trend for technical, ethical, and legal reasons [28]. In this regard, pattern mining can be used to spot regularities that arise when a complex black-box model yields a particular verdict. For instance, one may want to know the conditions under which the control module of a self-driving car decided to stop without apparent reason, or which factors caused a ML-based credit assessor to reject a loan request. Patterns and conditions are the building blocks for the generation of human-readable explanations for such black-box systems.

3.5. 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. Energiency, Amossys).

- **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 have collaborated 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.
- **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.

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 [32], [30] or medical devices [25], 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 [29]. Contrary to local hospital EHR or even to other national initiatives, the SNDS data warehouse covers a huge population. It makes possible studies on unfrequent drugs or diseases in real conditions of use. To tackle the volume and the diversity of the SNDS data warehouse, a research program has been established to design an innovative toolbox. This research program is focused first on the modeling of care pathways from the SNDS database and, second, on the design of tools supporting the extraction of insights about massive and complex care pathways by clinicians. In such a database a care pathway is an individual sequence of drugs exposures, medical procedures and hospitalizations.

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.

  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

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

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.

- **Numerical Rule Mining for Prediction of Wheat and Vine diseases.** Wheat and vine crops are crucial for the economy of France. Alas, they both suffer from threatening diseases. The fight against crop diseases is often implemented through the use of myriads of phytosanitary products, which raise concerns in regards to public health and environmental impact. In order to control the use of these products, agronomists have developed statistical models to understand the dynamics of diseases and reduce the utilization of phytosanitary products. The internship of Olivier Pelgrin, financed by #DigitAg and supervised in collaboration with the Acta [2] and the IFV [3], was concerned with the development of a data mining method capable of extracting hybrid expert rules from observations of vine and wheat diseases. Hybrid rules combine patterns such as $variety = \text{“}Grenache\text{”}$ with regression models, e.g., $incidence = \alpha \times temperature + \beta$. Such rules are conceived to aid the study of wheat and vine diseases. The rules are meant to be interpretable, i.e., as concise as possible, and globally accurate, thus they constitute a pattern-aided regression method that has shown good prediction performance. The resulting method, called HIPAR (Hierarchical Interpretable Pattern-aided regression), is currently under submission at the SIAM Conference on Data Mining (SDM19).

4.5. Others

- **Mining Referring Expressions in Knowledge Bases.** A *referring expression* (RE) is a description that identifies a concept unambiguously in a domain of knowledge. For example, the expression “X is the capital of France” is an RE for Paris, because no other city holds this title. Mining REs from data is a central task in natural language generation, and is also applicable to automatic journalism and query generation (e.g., for benchmarking purposes). A common requirement for REs is to be “intuitive”, that is, to resort to concepts that are easily understandable by users. For this reason, existing methods required users to provide a lexical ranking of concepts that conveys their preferences for certain predicates and entities in descriptions. In addition, state-of-the-art methods are not tailored for large current knowledge bases and, due to data incompleteness, are often unable to provide an answer. The internship of Julien Delaunay was conceived to tackle these issues by designing a parallel method to mine intuitive REs on large knowledge bases. The system extends the state-of-the-art language bias for REs to deal with incompleteness and proposes a notion of intuitiveness based on information theory that does not require a lexical ranking from the user. The description of the system, named REMI, is under review at the Extended Semantic Web Conference (ESWC) 2019.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

- Honorable Mention in the ACM SIGMOD Jim Gray Dissertation Award. ACM SIGMOD conference, June 2018 (L. Galárraga).

6. New Software and Platforms

6.1. REMI

Mining Intuitive Referring Expressions in Knowledge Bases

[2] http://www.acta.asso.fr/

[3] Institut Français de la Vigne

KEYWORDS: RDF - Knowledge database - Referring expression
FUNCTIONAL DESCRIPTION: REMI takes an RDF knowledge base stored as an HDT file, and a set of target entities and returns a referring expression that is intuitive, i.e., the user is likely to understand it.

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

6.2. HIPAR

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

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

6.3. NegPSpan

Negative pattern mining with PrefixSpan
KEYWORDS: Pattern discovery - Data mining - Sequential patterns - Traces
SCIENTIFIC DESCRIPTION: Mining frequent sequential patterns consists in extracting recurrent behaviors, modeled as patterns, in a big sequence dataset. Such patterns inform about which events are frequently observed in sequences, i.e. what does really happen. Sometimes, knowing that some specific event does not happen is more informative than extracting a lot of observed events. Negative sequential patterns (NSP) formulate recurrent behaviors by patterns containing both observed events and absent events. Few approaches have been proposed to mine such NSPs. In addition, the syntax and semantics of NSPs differ in the different methods which makes it difficult to compare them. This article provides a unified framework for the formulation of the syntax and the semantics of NSPs. Then, we introduce a new algorithm, NegPSpan, that extracts NSPs using a PrefixSpan depth-first scheme and enabling maxgap constraints that other approaches do not take into account. The formal framework allows for highlighting the differences between the proposed approach wrt to the methods from the literature, especially wrt the state of the art approach eNSP. Intensive experiments on synthetic and real datasets show that NegPSpan can extract meaningful NSPs and that it can process bigger datasets than eNSP thanks to significantly lower memory requirements and better computation times.
FUNCTIONAL DESCRIPTION: NegPSpan is software to extract patterns from sequential data (traces, sequences of events, client pathways, etc.). The NegPSpan software extracts two types of patterns: the classical sequential patterns and the negative sequential patterns. Sequential patterns describe recurrent behaviors described as a sequence of events (e.g. event A occurs, then event B occurs and finally C occurs) while negative sequential patterns hold information about absent event (e.g. event A occurs, then event B occurs but without any C in between).

The user has to provide a dataset in the IBM sequence format and, at least, a parameters corresponding to the minimal number of occurrences in the dataset (and possible additional parameters). The software efficiently extracts the patterns and output them (in a text or JSON format). The software can use different strategies for exploring negative sequential patterns, and also specify some constraints about the expected patterns.
NEWS OF THE YEAR: NegPSpan has been developed in 2018.

- Participants: Thomas Guyet and René Quiniou
- Contact: Thomas Guyet
- Publication: NegPSpan: efficient extraction of negative sequential patterns with embedding constraints
- URL: http://people.irisa.fr/Thomas.Guyet/negativepatterns/evalnegpat.php

7. New Results

7.1. Introduction

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

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 mining of negative patterns [24], [14] and periodic patterns [18].
- *Data Mining for the masses.* In [11], 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.
- *Efficient pattern mining.* In [10], we propose a method to sample itemsets efficiently on streaming data. This contribution tackles two limitations of the state of the art in pattern mining: (1) the so-called pattern explosion —the user is confronted to too many patterns—, and (2) the assumption of static data.
- *Data Mining for Data Science*. One of the most basic types of patterns is to know if the data makes one single group, i.e., is *unimodal*, or can be clustered into several groups. In [13], we propose a new statistical test of unimodality, that is both independent of the input distribution and computationally efficient.

7.1.2. Decision Support

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

- *Forecasting & prediction.* In [15], [12], we propose solutions to automate the task of capacity planning in the context of a large data network as the one available at Orange. The work in [19] offers a tool to predict the nutritional needs of sows in lactation.
- *Anomaly Detection.* The work in [20] tackles the problem of fraud detection under imbalanced data.

7.1.3. Others

- *Machine Learning.*[16] proposes a novel algorithm to weight the importance of classification errors when training a classifier. [8] proposes a classification algorithm optimized for highly imbalanced data.
- Query optimization. In [9] we propose a query-load-agnostic caching approach to speed-up provenance-aware queries in RDF data cubes.

7.2. Mining Periodic Patterns with a MDL Criterion

Participants: E. Galbrun, P. Cellier, N. Tatti, A. Termier, B. Crémilleux

The quantity of event logs available is increasing rapidly, be they produced by industrial processes, computing systems, or life tracking, for instance. It is thus important to design effective ways to uncover the information they contain. Because event logs often record repetitive phenomena, mining periodic patterns is especially relevant when considering such data. Indeed, capturing such regularities is instrumental in providing condensed representations of the event sequences. The work in [18] presents an approach for mining periodic patterns from event logs while relying on a Minimum Description Length (MDL) criterion to evaluate candidate patterns. Our goal is to extract a set of patterns that suitably characterises the periodic structure present in the data. We evaluate the interest of our approach on several real-world event log datasets.

7.3. NegPSpan: Efficient Extraction of Negative Sequential Patterns with Embedding Constraints

Participants: T. Guyet, R. Quinou

Mining frequent sequential patterns consists in extracting recurrent behaviors, modeled as patterns, in a big sequence dataset. Such patterns inform about which events are frequently observed in sequences, i.e., what does really happen. Sometimes, knowing that some specific event does not happen is more informative than extracting a lot of observed events. Negative sequential patterns (NSP) formulate recurrent behaviors by patterns containing both observed events and absent events. Few approaches have been proposed to mine such NSPs. In addition, the syntax and semantics of NSPs differ in the different methods which makes it difficult to compare them. [24] provides a unified framework for the formulation of the syntax and the semantics of NSPs. Then, it introduces a new algorithm, NegPSpan, that extracts NSPs using a PrefixSpan depth-first scheme and enabling maxgap constraints that other approaches do not take into account. The formal framework allows for highlighting the differences between the proposed approach w.r.t. to the methods from the literature, especially w.r.t. the state of the art approach eNSP. Intensive experiments on synthetic and real datasets show that NegPSpan can extract meaningful NSPs and that it can process bigger datasets than eNSP thanks to significantly lower memory requirements and better computation times.

7.4. NTGSP: Mining Negative Temporal Patterns

Participants: K. Tsesmeli, M. Boumghar, T. Guyet, R. Quiniou, L. Pierre

In [14] the authors study the problem of extracting frequent patterns containing positive events, negative events specifying the absence of events as well as temporal information on the delay between these events. [14] defines the semantics of such patterns and proposes the NTGSP method based on state-of-the-art approaches. The performance of the method is evaluated on commercial data provided by EDF (Électricité de France).

7.5. Accelerating Itemset Sampling using Satisfiability Constraints on FPGA

Participants: M. Gueguen, O. Sentieys, A. Termier

Finding recurrent patterns within a data stream is important for fields as diverse as cybersecurity or e-commerce. This requires to use pattern mining techniques. However, pattern mining suffers from two issues. The first one, known as “pattern explosion”, comes from the large combinatorial space explored and is the result of too many patterns output for analysis. Recent techniques, called *output space sampling* solve this problem by outputting only a sample of the results, with a target size provided by the user. The second issue is that most algorithms are designed to operate on static datasets or low throughput streams. In [10], the authors propose a contribution to tackle both issues, by designing an FPGA accelerator for pattern mining with output space sampling. They show that their accelerator can outperform a state-of-the-art implementation on a server class CPU using a modest FPGA product.

7.6. Are your data data gathered? The Folding Test of Unimodality

Participants: A. Siffer, C. Largouët, A. Termier

Understanding data distributions is one of the most fundamental research topics in data analysis. The literature provides a great deal of powerful statistical learning algorithms to gain knowledge on the underlying distribution given multivariate observations. We are likely to find out a dependence between features, the appearance of clusters or the presence of outliers. Before such deep investigations, [13] proposes the folding test of unimodality. As a simple statistical description, it allows to detect whether data are gathered or not (unimodal or multimodal). To the best of our knowledge, this is the first multivariate and purely statistical unimodality test. It makes no distribution assumption and relies only on a straightforward p-value. Experiments on real world data show the relevance of the test and how to use it for the task of clustering.

7.7. Day-ahead Time Series Forecasting: Application to Capacity Planning

Participants: C. Leverger, V. Lemaire, S. Malinowski, T. Guyet, L. Rozé
In the context of capacity planning, forecasting the evolution of server usage enables companies to better manage their computational resources. The work in [12] addresses this problem by collecting key indicator time series. The article proposes a method to forecast the evolution of server usage one day-ahead. The method assumes that data is structured by a daily seasonality, but also that there is typical evolution of indicators within a day. Then, it uses the combination of a clustering algorithm and Markov Models to produce day-ahead forecasts. Our experiments on real datasets show that the data satisfies our assumption and that, in the case study, our method outperforms classical approaches (AR, Holt-Winters).

7.8. PerForecast: A Tool to Forecast the Evolution of Time Series for Capacity Planning.

Participants: C. Leverger, R. Marguerie, V. Lemaire, T. Guyet, S. Malinowski
The work published in [15] presents PerForecast, a tool for automatic capacity planning. The tool relies on univariate temporal data and automatically configured predictive models. The aim is to anticipate *capacity problems* in the infrastructure of Orange in order to ensure the delivery of services to customers. For example, PerForecast can predict the overhead of a server at the earliest possible stage, so that new machines can be ordered before the deterioration of the service in question. The purchase procedures being long and costly, the earlier they are done, the better the quality of service.

7.9. Tree-based Cost-Sensitive Methods for Fraud Detection in Imbalanced Data

Participants: G. Metzler, X. Badiche, B. Belkasmi, E. Fromont, A. Habrard, M. Sebban
Bank fraud detection is a difficult classification problem where the number of frauds is much smaller than the number of genuine transactions. The authors of [20] present cost sensitive tree-based learning strategies applied in this context of highly imbalanced data. The paper first proposes a cost sensitive splitting criterion for decision trees that takes into account the cost of each transaction. Then the criterion is extended with a decision rule for classification with tree ensembles. The authors then propose a new cost-sensitive loss for gradient boosting. Both methods have been shown to be particularly relevant in the context of imbalanced data. Experiments on a proprietary dataset of bank fraud detection in retail transactions show that the presented cost sensitive algorithms increase the retailer's benefits by 1,43% compared to non cost-sensitive ones and that the gradient boosting approach outperforms all its competitors.

7.10. An Algorithm to Optimize the F-measure by Proper Weighting of Classification Errors

Participants: K. Bascol, R. Emonet, E. Fromont, A. Habrard, G. Metzler, M. Sebban
[16] proposes an F-Measure optimization algorithm with theoretical guarantees that can be used with any error-weighting learning method. The algorithm, iteratively generates a set of costs from the training set so that the final classifier has an F-measure close to optimal. The optimality of the F-measure is expressed using a finer upper bound as presented in [31]. Furthermore, we show that the costs obtained at each iteration of our method can drastically reduce the search space and thus converge quickly to the optimal parameters. The efficiency of the method is shown both in terms of F-measurement but also in terms of speed of convergence on several unbalanced datasets.

7.11. Learning Maximum excluding Ellipsoids from Imbalanced Data with Theoretical Guarantees

Participants: G. Metzler, X. Badiche, B. Belkasmi, E. Fromont, A. Habrard, M. Sebban

[8] addresses the problem of learning from imbalanced data. The authors consider the scenario where the number of negative examples is much larger than the number of positive ones. This work proposes a theoretically-founded method, which learns a set of local ellipsoids centered at the minority class examples while excluding the negative examples of the majority class. This task is addressed from a Mahalanobis-like metric learning point of view, which allows deriving generalization guarantees on the learned metric using the uniform stability framework. The experimental evaluation on classic benchmarks and on a proprietary dataset in bank fraud detection shows the effectiveness of the approach, particularly when the imbalance is huge.

7.12. Answering Provenance-Aware Queries on RDF Data Cubes under Memory Budgets

Participants: L. Galárraga, K. Ahlstrøm, K. Hose, T. B. Pedersen

The steadily-growing popularity of semantic data on the Web and the support for aggregation queries in SPARQL 1.1 have propelled the interest in Online Analytical Processing (OLAP) and data cubes in RDF. Query processing in such settings is challenging because SPARQL OLAP queries usually contain many triple patterns with grouping and aggregation. Moreover, one important factor of query answering on Web data is its provenance, i.e., metadata about its origin. Some applications in data analytics and access control require to augment the data with provenance metadata and run queries that impose constraints on this provenance. This task is called provenance-aware query answering. The work in [9] investigates the benefit of caching some parts of an RDF cube augmented with provenance information when answering provenance-aware SPARQL queries. [9] proposes provenance-aware caching (PAC), a caching approach based on a provenance-aware partitioning of RDF graphs, and a benefit model for RDF cubes and SPARQL queries with aggregation. The results on real and synthetic data show that PAC outperforms significantly the LRU strategy (least recently used) and the Jena TDB native caching in terms of hit-rate and response time.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

- **AdvisorSLA 2018 - Inria**
  Participants: E. Bourrand, L. Galárraga, E. Fromont, A. Termier
  Contract amount: 7,5k€
  Context. AdvisorSLA is a French company headquartered in Cesson-Sévigné, a city located in the outskirts of Rennes in Brittany. The company is specialized in software solutions for network monitoring. For this purpose, the company relies on techniques of network metrology. AdvisorSLA's customers are carriers and telecommunications/data service providers that require to monitor the performance of their communication infrastructure as well as their QoE (quality of service). Network monitoring is of tremendous value for service providers because it is their primary tool for proper network maintenance. By continuously measuring the state of the network, monitoring solutions detect events (e.g., an overloaded router) that may degrade the network's operation and the quality of the services running on top of it (e.g., video transmission could become choppy). When a monitoring solution detects a potentially problematic sequence of events, it triggers an alarm so that the network manager can take actions. Those actions can be preventive or corrective. Some statistics gathered by the company show that only 40% of the triggered alarms are conclusive, that is, they manage to signal a well-understood problem that requires an action from the network manager. This means that the remaining 60% are presumably false alarms. While false alarms do not hinder network operation, they do incur an important cost in terms of human resources.
  Objective. We propose to characterize conclusive and false alarms. This will be achieved by designing automatic methods to "learn" the conditions that most likely precede the fire of each type of alarm, and therefore predict whether the alarm will be conclusive or not. This can help adjust existing monitoring solutions in order to improve their accuracy. Besides, it can help network

managers automatically trace the causes of a problem in the network. The aforementioned problem has an inherent temporal nature: we need to learn which events occur before an alarm and in which order. Moreover, metrology models take into account the measurements of different components and variables of the network such as latency and packet loss. For these two reasons, we resort to the field of multivariate time sequences and time series. The fact that we know the "symptoms" of an alarm and whether it is conclusive or not, allows for the application of supervised machine learning and pattern mining methods.
Additional remarks. This is a pre-doctoral contract signed with AdvisorSLA to start the work for the PhD of E. Bourrand (Thèse CIFRE) while the corresponding administrative formalities are completed.

- **ATERMES 2018-2021 - Univ Rennes 1**
  Participants: H. Zhang, E. Fromont
  Contract amount: 45k€
  Context. ATERMES is an international mid-sized company, based in Montigny-le-Bretonneux with a strong expertise in high technology and system integration from the upstream design to the long-life maintenance cycle. It has recently developed a new product, called BARIERTM ("Beacon Autonomous Reconnaissance Identification and Evaluation Response"), which provides operational and tactical solutions for mastering borders and areas. Once in place, the system allows for a continuous night and day surveillance mission with a small crew in the most unexpected rugged terrain. BARIER™ is expected to find ready application for temporary strategic site protection or ill-defined border regions in mountainous or remote terrain where fixed surveillance modes are impracticable or overly expensive to deploy.
  Objective. The project aims at providing a deep learning architecture and algorithms able to detect anomalies (mainly the presence of people or animals) from multimodal data. The data are considered "multimodal" because information about the same phenomenon can be acquired from different types of detectors, at different conditions, in multiple experiments, etc. Among possible sources of data available, ATERMES provides Doppler Radar, active-pixel sensor data (CMOS), different kind of infra-red data, the border context etc. The problem can be either supervised (if label of objects to detect are provided) or unsupervised (if only times series coming from the different sensors are available). Both the multimodal aspect and the anomaly detection one are difficult but interesting topics for which there exist few available works (that take both into account) in deep learning.
- **PSA - Inria**
  Participants: E. Fromont, A. Termier, L. Rozé, G. Martin
  Contract amount: 15k€
  Context. Peugeot-Citroën (PSA) group aims at improving the management of its car sharing service. To optimize its fleet and the availability of the cars throughout the city, PSA needs to analyze the trajectory of its cars.
  Objective. The aim of the internship is (1) to survey the existing methods to tackle the aforementioned need faced by PSA and (2) to also investigate how the techniques developed in LACODAM (e.g., emerging pattern mining) could be serve this purpose. A framework, consisting of three main modules, has been developped. We describe the modules in the following.
  - A town modelisation module with clustering. Similar towns are clustered in order to reuse information from one town in other towns.
  - A travel prediction module with basic statistics.
  - A reallocation strategy module (choices on how to relocate cars so that the most requested areas are always served). The aim of this module is to be able to test different strategies.

  Additional remarks. This is a pre-doctoral contract to start the work for the PhD of G. Martin (Thèse CIFRE) while the corresponding administrative formalities are completed.

9. Partnerships and Cooperations

9.1. National Initiatives

- **Hyptser: Hybrid Prediction of Time Series**

  Participants: T. Guyet, S. Malinowski (LinkMedia), V. Lemaire (Orange)

  HYPTSER is a collaborative project between Orange Labs and LACODAM funded by the Fondation Mathématique Jacques Hadamard (PGMO program). It aims at developping new hybrid time series prediction methods in order to improve capacity planning for server farms. Capacity planning is the process of determining the infrastructure needed to meet future customer demands for online services. A well-made capacity planning helps to reduce operational costs, and improves the quality of the provided services. Capacity planning requires accurate forecasts of the differences between the customer demands and the infrastructure theoretical capabilities. The HYPTSER project makes the assumption that this information is captured by key performance indicators (KPI), that are measured continuously in the service infrastructure. Thus, we expect to improve capacity planning capabilities by making accurate forecasts of KPI time series. Recent methods about time series forecasting make use of ensemble models. In this project, we are interested in developing hybrid models for time series forecasting. Hybrid models aim at jointly partitioning the data, learning forecasting models in each partition and learning how to combine their outputs. We are currently developing two different approaches for that purpose, one based on the MODL framework and the other based on neural networks. We describe these approaches below:

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

  In the next steps of this project, we will analyze the performance of this two strategies on KPI time series provided by Orange and compare them to classical ensemble methods.

9.1.1. ANR

- **#DigitAg: Digital Agriculture**

  Participants: A. Termier, V. Masson, C. Largouët, A.I. Graux

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

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

9.1.2. National Platforms

- **PEPS: Pharmaco-epidemiology for Health Products**
  Participants: Y. Dauxais, T. Guyet, V. Masson, R. Quinou, A. Samet
  The PEPS project (Pharmaco-epidemiology des Produits de Santé) is funded by the ANSM (National Agency for Health Security). The project leader is E. Oger from the clinical investigation center CIC-1414 INSERM/CHU Rennes. The other partners located in Rennes are the Institute of Research and Technology (IRT), B<>Com, EHESP and the LTSI. The project started in January 2015 and is funded for 4 years. The PEPS project consists of two parts: a set of clinical studies and a research program dedicated to the development of innovative tools for pharmaco-epidemiological studies with medico-administrative databases. Our contribution to this project will be to propose pattern mining algorithms and reasoning techniques to analyse the typical care pathways of specific groups of insured patients. Since last year we have been working on the design and development of algorithms [27], [26] to mine patterns on care pathways.

9.2. International Research Visitors

9.2.1. Internships

From May to August 2018 we hosted Devang Kulshreshtha, a computer science student from the Indian Institute of Technology (BHU) Varanasi, who worked on “Debugging Deep Learning Algorithms via Pattern Mining Methods”. His work aimed at mining patterns of neuron activation that precede misclassifications in deep neural networks (DNN). The goal of this effort is to predict when a DNN will likely err. This can be used e.g., to obtain hints on how to retrain the network to improve its accuracy.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

- General co-chair (E. Fromont) of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD’19, http://www.ecmlpkdd2019.org/)
- Journal-track co-chair (E. Fromont) of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD’18) (http://www.ecmlpkdd2018.org/)
- Organization co-chair (T. Guyet) of the technical track “Geoinformatic analytics” of the 34rd ACM/SIGAPP Symposium On Applied Computing (https://gia.sciencesconf.org/)
- Organization co-chair (T. Guyet) of the 3nd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data @ ECML (https://project.inria.fr/aaldt18/)
- Organization chair (T. Guyet) of GAST workshop at EGC 2018 (https://gt-gast.irisa.fr/)
- Organization chair and program committee member (L. Galárraga) of the AIMLAI workshop (https://project.inria.fr/aimlai/) that will take place at EGC 2019.
- Organization of the first workshop of the Data Mining axis of the #DigitAg project (A. Termier)

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- Senior PC member of the Symposium on Intelligent Data Analysis 2018 (E. Fromont)
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2018 (E. Fromont, T. Guyet)
- AAAI Conference on Artificial Intelligence 2018 (T. Guyet)

- International Conference on Data Mining (ICDM) 2018 (A. Termier)
- International Joint Conference on Artificial Intelligence (IJCAI) 2018 (A. Termier)
- International Conference on Information Management and Big Data (SimBig) 2018 (T. Guyet)
- International Symposium on Computing and Networking (CANDAR) 2018 (A. Termier)
- Extraction et Gestion de Connaissances (EGC) 2018 (T. Guyet, A. Termier)
- Conférence Nationale en Intelligence Artificielle (CNIA) 2018 (T. Guyet, A. Termier)
- Conférence Française en Photogramétrie et Télédétection (CFPT) 2018 (T. Guyet)

10.1.2.2. Reviewer

- Conference on Information and Knowledge Management (CIKM) 2018 (L. Galárraga)
- Conference on Practice of Knowledge Discovery in Databases (PKDD) 2018 (L. Galárraga)
- Conference on Very Large Databases (VLDB) 2018, demo track (L. Galárraga)
- Extended Semantic Web Conference (ESWC) 2018, demo track (L. Galárraga)
- The Web Conference (WWW) 2018 (L. Galárraga)
- International Workshop on the Web and Databases (WebDB) 2018 (L. Galárraga)
- Symposium on Theoretical Aspects of Computer Science (STACS) 2019, (A. Termier)
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) 2018 (A. Termier)

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- Machine Learning journal for the ECMLPKDD 2018 special issue (E. Fromont)
- Data Mining journal for the ECMLPKDD 2018 special issue (E. Fromont)
- Revue d'Intelligence Artificielle (RIA, T. Guyet)

10.1.3.2. Reviewer - Reviewing Activities

- Remote Sensing (T. Guyet)
- Journal of Biomedical Informatics (T. Guyet)
- Journal of Machine Learning (T. Guyet)
- Artificial Intelligence Review (T. Guyet, L. Galárraga)
- Annals of Mathematics and Artificial Intelligence (T. Guyet)
- Data Mining and Knowledge Discovery (L. Galárraga, A. Termier)
- Information Systems (L. Galárraga)

10.1.4. Invited Talks

- Invited to the Dagstuhl Seminar 18401 on "Automating Data Science", Germany (E. Fromont and A. Termier)
- Invited talk for the IA^2 (Autumn school) organized by the GDR IA. (E. Fromont)
- Invited talk for the Labex Henri Lebesgue at Technicolor, Rennes. (E. Fromont)
- Invited talk for the "Transfer Learning: From Theory to Applications" summer school in Cachan. (E. Fromont)
- Invited talk for the conference "Brain Hack", Rennes. (E. Fromont)
- Invited talk for the Machine Learning Meetup, Rennes. (E. Fromont)
- Invited talks by ONERA, LAAS, Strasbourg University (T. Guyet)
- Keynote talk at CITT 2018 (International Conference on Technology Trends), Universidad Técnica de Babahoyo, Ecuador (L. Galárraga)

10.1.5. Scientific Expertise

- Member of an ANR (CE 23) evaluation committee (E. Fromont)
- Working group initiated by ALLISTENE about a Research Infrastructure for AI in France. It was proposed in the "Villani report" under the name GENIAL (Grand Equipement National pour l'Intelligence Artificielle)
- Evaluation of an ERC proposal (A. Termier)
- Evaluation of a project proposal for Université Catholique de Louvain (A. Termier)
- Evaluation of projects proposals for ANR (T. Guyet: 3 projects, A. Termier: 1 project)
- Evaluation of a project proposal for Pays de Loire Region (A. Termier)
- Audit of the AI activities of a major French industrial group with a group of Inria experts (A. Termier)

10.1.6. Research Administration

- Elected (college A) at IRISA research institute scientific council (E. Fromont)
- Nominated at the scientific council of the "Fondation Blaise Pascal", dedicated to scientific mediation (E. Fromont)
- Co-chair of the national evaluation committee of INRA Engineers (section Mathematics and Informatics) (T. Guyet)
- Member of INRA (National Agronomy Institute) Scientific Committee (A. Termier)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

1. Master 2 IL, CCN: Option Machine Learning, Istic, University of Rennes 1, 32h (E. Fromont)
2. Master 2 DMV Module: Data Mining and Visualization, 13h, M2, Istic, University of Rennes 1 (A. Termier)
3. Master 1 SIF: Option IA, Istic, 20h University of Rennes 1 (A. Termier, E. Fromont)
4. Master 2 DataViz with R 10h, Computer Science for BigData 30h, Master Datascience, Agrocampus Ouest Rennes (Louis Bonneau, C. Largouët)
5. Master 1 Scientific Programming, Data Management, Agrocampus Ouest Rennes (C. Largouët)
6. INSA 4 Project Interpretability of Time series Machine Learning algorithms. (L. Rozé, M. Guillemé)

10.2.2. Supervision

PhD defended in 2018: Yann Dauxais, "Query-language for Care-pathway Mining and Analysis", 01/02/2015, D. Gross-Amblard, T. Guyet, A. Happe

PhD defended in 2018: Clément Gautrais, "Mining Massive Data from Client Purchases", 01/10/2015, A. Termier, P. Cellier, T. Guyet and R. Quiniou

PhD defended in 2018: Romain Deville, "Fouille de Grilles Spatio-temporelles Appliquée à la Classification d'Images et à l'Analyse d'Automates Cellulaires", 30/05/2018, E. Fromont, C. Solnon (Lyon), B. Jeudy (Saint-Etienne).

PhD in progress: Maël Guillemé, "New Data Mining Approaches for Improving Energy Consumption in Factories", 03/10/2016, L. Rozé, V. Masson, A. Termier

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

PhD in progress: Alban Siffer, “Data Mining Approaches for Cyber Attack Detection”, 03/2016, P-A Fouque, A. Termier, C. Largouët.

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

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

PhD in progress: Anh Duong Nguyen, “Compression Based Pattern Mining”, 01/03/2018, R. Gaudel, P. Cellier, A. Termier

PhD in progress: Johanne Bakalara, “Temporal Model of Care pathway to Query Medico-administrative Databases", 01/10/2018, T. Dameron O., Oger E., Guyet, S. A. Happe

PhD in progress: Erwan Bourrand, “Interactive Data Mining for Root Cause Analysis of Performance Issues in Networks”, 03/12/2018, L. Galárraga, E. Fromont, A. Termier

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

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

10.2.3. Juries

- PHD defenses (E. Fromont): Jean Ogier, Paris (reviewer); Maxime Chabert, Lyon (committee member); Wissam Siblini, Nantes (committee member, president); Michaël Blot, Paris (committee member, president); Nicolas Audebert, Paris (committee member, president); Hoang Viet Tuan Nguyen, Annecy (reviewer); Soufiane Belharbi, Rouen (reviewer); Romain Deville, Lyon (supervisor); Bastien Moysset, Lyon (reviewer); Géraud Le Falher, Lille (committee member).
- Reviewer for the HDR defense of Claude Pasquier (Univ. Nice), defended on 20/09/2018 (A. Termier)
- Examiner for the HDR defense of Marc Plantevit (Univ. Lyon 1), defended on 14/12/2018 (A. Termier)
- Committee member for the PhD defense of Clément Gautrais (Univ. Rennes), defended on 16/10/2018 (A. Termier, T. Guyet, R. Quiniou)
- Committee member for the PhD defense of Yann Dauxais (Univ. Rennes), defended on 13/04/2018 (T. Guyet)
- Member of the hiring committee for Lille MCF27-0076, Paris-Sud PR27-0122, Saint-Etienne TSE MCF27-511 as president (E. Fromont)
- Member of the INRA research hiring committee on 25/04/2018 (MIA/INRA) (T. Guyet)
- Member of the associate professor hiring committee on 23/04/2018 (LERIA/Univ. Angers) (T. Guyet)
- Member of the associate professor hiring committee MCF0942 of Polytech Nancy (A. Termier)
- Thesis advisory committee member:
  - T. Guyet: Caglayan Tuna (UBS)
  - A. Termier: Mathilde Chen (INRA), Lucas Bourneuf (Univ. Rennes)

10.3. Popularization

10.3.1. Interventions

- Public debate on Artificial Intelligence at Café des Champs-Libres, Rennes (E. Fromont)

- Radio C-LAB, invited to the "Mars 2081" show on AI. (E. Fromont)
- Invited member to the focus group "Transition numérique et pratiques de recherche et d'enseignement en agriculture, alimentation, environnement et sciences vétérinaires à l'échéance 2040" (T. Guyet)
- Intervention at the Pint of Science 2018 event, topic: Artificial Intelligence and Data Science (A. Termier)
- Invited talk at "Matinale Rennes Atalante", Rennes (A. Termier, 22/12/2018) [4].
- Invited talk at Lacroix-Sofrel industrial group (A. Termier)
- Invited talk at the kickoff the OpenLab PSA-Inria, where Lacodam if actively involved (A. Termier)
- Invited talk at the conference *IN'Sciences "Big data : comment combiner intelligence artificielle et gestion de données ?"* (L. Galárraga, chaired by L. Rozé)

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

Doctoral Dissertations and Habilitation Theses

[3] C. Gautrais. *Signatures : detecting and characterizing complex recurrent behavior in sequential data*, Université Rennes 1, October 2018, https://tel.archives-ouvertes.fr/tel-01984629

Articles in International Peer-Reviewed Journals

[4] J. Davis, B. Bringmann, E. Fromont, D. Greene. *Guest editors introduction to the special issue for the ECML PKDD 2018 journal track*, in "Machine Learning", September 2018, vol. 107, n^o 8-10, pp. 1207-1208 [*DOI :* 10.1007/s10994-018-5745-x], https://hal.archives-ouvertes.fr/hal-01952495

[5] 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", 2018, vol. 32, n^o 1, pp. 78–80 [*DOI :* 10.1111/fcp.12323], https://hal.inria.fr/hal-01631802

[6] K. Fauvel, V. Masson, P. Faverdin, A. Termier. *Data Science Techniques for Sustainable Dairy Management*, in "ERCIM News", April 2018, vol. 113, pp. 29-30, https://hal.archives-ouvertes.fr/hal-01951807

[7] D. Greene, B. Bringmann, E. Fromont, J. Davis. *Introduction to the special issue for the ECML PKDD 2018 journal track*, in "Data Mining and Knowledge Discovery", September 2018, vol. 32, n^o 5, pp. 1177-1178 [*DOI :* 10.1007/s10618-018-0586-6], https://hal.archives-ouvertes.fr/hal-01952487

[4] Video: http://www.rennes-atalante.fr/actualites-technopole/captations-matinales-rennes-atalante/2018/visionnez-la-matinale.html

[8] G. Metzler, X. Badiche, B. Belkasmi, E. Fromont, A. Habrard, M. Sebban. *Learning maximum excluding ellipsoids from imbalanced data with theoretical guarantees*, in "Pattern Recognition Letters", September 2018, vol. 112, pp. 310-316 [*DOI :* 10.1016/j.patrec.2018.08.016], https://hal.archives-ouvertes.fr/hal-01878830

International Conferences with Proceedings

[9] L. Galárraga, K. Ahlstrøm, K. Hose, T. B. Pedersen. *Answering Provenance-Aware Queries on RDF Data Cubes under Memory Budgets*, in "ISWC 2018 - 17th International Semantic Web Conference", Monterey, United States, LNCS, Springer, October 2018, vol. 11136, pp. 547-565 [*DOI :* 10.1007/978-3-030-00671-6_32], https://hal.inria.fr/hal-01931333

[10] M. Gueguen, O. Sentieys, A. Termier. *Accelerating Itemset Sampling using Satisfiability Constraints on FPGA*, in "DATE 2019 - IEEE/ACM Design, Automation and Test in Europe", Florence, Italy, March 2019, pp. 1-6, https://hal.inria.fr/hal-01941862

[11] S. Kelly. *A Communication Model that Bridges Knowledge Delivery between Data Miners and Domain Users*, in "HICSS 2018 - 51th Hawaii International Conference on System Sciences", Hawaii, United States, January 2018, pp. 192-198, https://hal.archives-ouvertes.fr/hal-01651737

[12] C. Leverger, V. Lemaire, S. Malinowski, T. Guyet, L. Rozé. *Day-ahead time series forecasting: application to capacity planning*, in "AALTD'18 at ECML 2018", Dublin, Ireland, September 2018, https://arxiv.org/abs/1811.02215 , https://hal.inria.fr/hal-01912002

[13] A. Siffer, P.-A. Fouque, A. Termier, C. Largouët. *Are your data gathered? The Folding Test of Unimodality*, in "KDD 2018 - 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Minin", London, United Kingdom, August 2018, pp. 2210-2218 [*DOI :* 10.1145/3219819.3219994], https://hal.archives-ouvertes.fr/hal-01951676

[14] 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. 263-268, https://hal.inria.fr/hal-01657540

National Conferences with Proceedings

[15] C. Leverger, R. Marguerie, V. Lemaire, T. Guyet, S. Malinowski. *PerForecast : un outil de prévision de l'évolution de séries temporelles pour le planning capacitaire*, in "EGC 2018 - Conférence Extraction et Gestion des Connaissances", Paris, France, vol. RNTI, January 2018, vol. E, n^{o} 34, pp. 455-458, https://hal.inria.fr/hal-01911243

Conferences without Proceedings

[16] K. Bascol, R. Emonet, E. Fromont, A. Habrard, G. Metzler, M. Sebban. *Un algorithme de pondération de la F-Mesure par pondération des erreurs de classfication*, in "Conférence pour l'Apprentissage Automatique", Saint-Etienne du Rouvray, France, June 2018, https://hal.archives-ouvertes.fr/hal-01803183

[17] Y. Dauxais, C. Gautrais. *Predicting Pass Receiver In Football Using Distance Based Features*, in "5th Workshop on Machine Learning and Data Mining for Sports Analytic of ECML/PKDD", Dublin, Ireland, September 2018, https://hal.archives-ouvertes.fr/hal-01912616

[18] E. Galbrun, P. Cellier, N. Tatti, A. Termier, B. Crémilleux. *Mining Periodic Patterns with a MDL Criterion*, in "ECML/PKDD 2018 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Dublin, Ireland, September 2018, pp. 1-16, https://hal.archives-ouvertes.fr/hal-01951722

[19] R. Gauthier, F. Guay, L. Brossard, C. Largouët, J.-Y. Dourmad. *Precision feeding of lactacting sows: development of a decision support tool to handle variability*, in "EAAP 2018 - 69th Annual Meeting of the European Federation of Animal Science", Dubrovnik, Croatia, Book of Abstracts of the 69th Annual Meeting of the European Federation of Animal Science, Wageningen Academic Publishers, August 2018, vol. 24, https://hal.archives-ouvertes.fr/hal-01949645

[20] G. Metzler, X. Badiche, B. Belkasmi, E. Fromont, A. Habrard, M. Sebban. *Tree-based Cost-Sensitive Methods for Fraud Detection in Imbalanced Data*, in "IDA 2018 - 17th International Symposium on Intelligent Data Analysis", ‘s-Hertogenbosch, Netherlands, LNCS, Springer, October 2018, vol. 11191, pp. 213-224 [*DOI :* 10.1007/978-3-030-01768-2_18], https://hal.archives-ouvertes.fr/hal-01895967

Scientific Books (or Scientific Book chapters)

[21] Y. Dauxais, D. Gross-Amblard, T. Guyet, A. Happe. *Discriminant chronicle mining*, in "Advances in Knowledge Discovery and Management (Vol. 8)", 2018, pp. 1-30, https://hal.inria.fr/hal-01940146

Other Publications

[22] P. Besnard, T. Guyet, V. Masson. *Admissible Generalizations of Examples as Rules*, October 2018, pp. 1-29, One-day Workshop on Machine Learning and Explainability 2018, https://hal.inria.fr/hal-01940129

[23] R. Gauthier, F. Guay, L. Brossard, C. Largouët, J.-Y. Dourmad. *Développement d'un outil de prévision des besoins nutritionnels des truies en lactation. Application à l'alimentation de précision*, April 2018, 1 p. , Journées d'animation scientifiques du département Phase, Poster, https://hal.archives-ouvertes.fr/hal-01949603

[24] T. Guyet, R. Quiniou. *NegPSpan: efficient extraction of negative sequential patterns with embedding constraints*, July 2018, https://arxiv.org/abs/1804.01256 - working paper or preprint, https://hal.inria.fr/hal-01743975

References in notes

[25] 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

[26] 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

[27] Y. Dauxais, T. Guyet, D. Gross-Amblard, A. Happe. *Discriminant chronicles mining: Application to care pathways analytics*, in "Artificial Intelligence in Medicine", Vienna, Austria, 16th Conference on Artificial Intelligence in Medicine, June 2017 [*DOI :* 10.1007/978-3-319-59758-46], https://hal.archives-ouvertes.fr/hal-01568929

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

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

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

[31] S. A. P. PARAMBATH, N. USUNIER, Y. GRANDVALET. *Optimizing F-measures by Cost-sensitive Classification*, in "Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2", Cambridge, MA, USA, NIPS'14, MIT Press, 2014, pp. 2123–2131, http://dl.acm.org/citation.cfm?id=2969033.2969064

[32] 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$^{\text{o}}$ 11, pp. 1161–1169