IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2018

Project-Team ORPAILLEUR

Knowledge discovery, knowledge engineering

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

RESEARCH CENTER
Nancy - Grand Est

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team ORPAILLEUR

Creation of the Project-Team: 2008 January 01

Keywords:

Computer Science and Digital Science:

A3. - Data and knowledge
A3.1.1. - Modeling, representation
A3.1.7. - Open data
A3.2. - Knowledge
A3.2.1. - Knowledge bases
A3.2.2. - Knowledge extraction, cleaning
A3.2.3. - Inference
A3.2.4. - Semantic Web
A3.2.5. - Ontologies
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.5. - Bayesian methods
A3.4.8. - Deep learning
A3.5.2. - Recommendation systems
A4. - Security and privacy
A4.1. - Threat analysis
A8.1. - Discrete mathematics, combinatorics
A8.7. - Graph theory
A9. - Artificial intelligence
A9.1. - Knowledge
A9.2. - Machine learning
A9.6. - Decision support

Other Research Topics and Application Domains:

B1.1.2. - Molecular and cellular biology
B2. - Health
B2.3. - Epidemiology
B2.4.1. - Pharmaco kinetics and dynamics
B2.4.2. - Drug resistance
B3.1. - Sustainable development
B3.5. - Agronomy
B3.6. - Ecology
B3.6.1. - Biodiversity
B6.3.4. - Social Networks
B6.4. - Internet of things
B8.5.2. - Crowd sourcing

B9. - Society and Knowledge
B9.5.6. - Data science

1. Team, Visitors, External Collaborators

Research Scientists
Amedeo Napoli [Team leader, CNRS, Senior Researcher, HDR]
Esther Catherine Galbrun [Inria, Researcher]
Chedy Raïssi [Inria, Researcher]

Faculty Members
Miguel Couceiro [Univ. de Lorraine, Professor, HDR]
Adrien Coulet [Univ. de Lorraine, Associate Professor (on sabbatical leave at University of Stanford, USA)]
Nicolas Jay [Univ. de Lorraine, Professor, HDR]
Jean Lieber [Univ. de Lorraine, Associate Professor, HDR]
Jean-François Mari [Univ. de Lorraine, Professor, HDR]
Emmanuel Nauer [Univ. de Lorraine, Associate Professor]
Sébastien Da Silva [Univ. de Lorraine, Associate Professor]
Malika Smaïl-Tabbone [Univ. de Lorraine, Associate Professor, until Apr 2018, HDR]
Yannick Toussaint [Univ. de Lorraine, Professor, HDR]

Post-Doctoral Fellows
Joël Legrand [Univ. de Lorraine, until Jun 2018]
Abdelkader Ouali [Inria, from Oct 2018]

PhD Students
Nacira Abbas [Inria, from Oct 2018]
Guilherme Alves Da Silva [Inria, from Nov 2018]
Quentin Brabant [Univ. de Lorraine]
Laurine Huber [Univ. de Lorraine, from Oct 2018]
Nyoman Juniarta [CNRS]
Tatiana Makhalova [Inria]
Pierre Monnin [Univ. de Lorraine]
François Pirot [Univ. de Lorraine]
Justine Reynaud [Inria, until Aug 2018, ATER from Sep 2018]

Technical staff
Eric Biagioli [Inria, from Oct 2018]
Jérémie Nevin [Inria]

Interns
Clement Bellanger [Univ. de Lorraine, from Apr 2018 until Jun 2018]
Marie Bogusz [Univ. de Lorraine, from May 2018 until Jul 2018]
Adrien Claudel [Univ. de Lorraine, from May 2018 until Jun 2018]
Yohann Fransot [CNRS, from Feb 2018 until Jun 2018]
Andre Giang [Univ. de Lorraine, from Apr 2018 until Jun 2018]
Corentin Jobard [Univ. de Lorraine, from Apr 2018 until Jun 2018]
Damien Levy [Univ. de Lorraine, from Apr 2018 until Jun 2018]
Hue Nam Ly [Univ. de Lorraine, from Apr 2018 until Jun 2018]
Reem Mathbout [Univ. de Lorraine, from Mar 2018 until Jul 2018]
Baptiste Mounier [Univ. de Lorraine, from Apr 2018 until Aug 2018]
Yelen Per [CNRS, until Aug 2018]

Administrative Assistants
Emmanuelle Deschamps [Inria]

Delphine Hubert [Univ. de Lorraine]
Annick Jacquot [CNRS, from Jul 2018]
Martine Kuhlmann [CNRS]

Visiting Scientist

Martin Trnecka [Visiting Scientist, University of Olomouc, Czech Republic, from Jun 2018 until Sep 2018]

External Collaborators

Alexandre Blansché [Univ. de Lorraine, Metz, Associate Professor (External Collaborator)]
Lydia Boudjeloud-Assala [Univ. de Lorraine, Metz, Associate Professor (External Collaborator), HDR]
Brieuc Conan-Guez [Univ. de Lorraine, Metz, Associate Professor (External Collaborator)]
Alain Gély [Univ. de Lorraine, Metz, Associate Professor (External Collaborator)]
Florence Le Ber [ENGEES Strasbourg, Professor (External Collaborator), HDR]
Frédéric Pennerath [Centrale-Supelec Metz, Associate Professor (External Collaborator)]

2. Overall Objectives

2.1. Introduction

Knowledge discovery in databases (KDD) consists in processing large volumes of data in order to discover knowledge units that are significant and reusable. Assimilating knowledge units to gold nuggets, and databases to lands or rivers to be explored, the KDD process can be likened to the process of searching for gold. This explains the name of the research team: in French "orpailleur" denotes a person who is searching for gold in rivers or mountains. The KDD process is based on three main operations: data preparation, data mining and interpretation of the extracted units as knowledge units. Moreover, the KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the analyst. The analyst selects and interprets a subset of the extracted units for obtaining knowledge units having a certain plausibility. In this view, KDD is an exploratory process similar to "exploratory data analysis".

As a person searching for gold may have a certain experience about the task and the location, the analyst may use general and domain knowledge for improving the whole KDD process. Accordingly, the KDD process may be associated with knowledge bases –or domain ontologies– related to the domain of data for implementing *knowledge discovery guided by domain knowledge* (KDDK). In KDDK, extracted units may have "a life" after the interpretation step for becoming "actionable": they are represented as knowledge units using a knowledge representation formalism and integrated within an ontology to be reused for problem-solving needs. In this way, knowledge discovery extends and updates existing knowledge bases, materializing a complementarity between knowledge discovery and knowledge engineering.

3. Research Program

3.1. Hybrid and Exploratory Knowledge Discovery

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining, data exploration, formal concept analysis, classification, pattern mining, numerical methods in data mining.

Knowledge discovery in databases (KDD) aims at discovering patterns in large databases. These patterns can then be interpreted as knowledge units to be reused in knowledge systems. From an operational point of view, the KDD process is based on three main steps: (i) selection and preparation of the data, (ii) data mining, (iii) interpretation of the discovered patterns. The KDD process –as implemented in the Orpailleur team– is based on data mining methods which are either symbolic or numerical. Symbolic methods are based on pattern mining (e.g. mining frequent itemsets, association rules, sequences...), Formal Concept Analysis (FCA [74]) and extensions of FCA such as Pattern Structures [79] and Relational Concept Analysis (RCA [84]). Numerical methods are based on Random Forests, SVM, Neural Networks, and probabilistic approaches such as second-order Hidden Markov Models (HMM [80]).

Domain knowledge, when available, can improve and guide the KDD process, materializing the idea of *Knowledge Discovery guided by Domain Knowledge* or KDDK. In KDDK, domain knowledge plays a role at each step of KDD: the discovered patterns can be interpreted as knowledge units and reused for problem-solving activities in knowledge systems, implementing the exploratory process "mining, interpreting (modeling), representing, and reasoning". In this way, knowledge discovery appears as a core task in knowledge engineering, with an impact in various semantic activities, e.g. information retrieval, recommendation and ontology engineering. Usual application domains include agronomy, astronomy, biology, chemistry, and medicine.

One main operation in the research work of Orpailleur on KDDK is *classification*, which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. Classification problems can be formalized by means of a class of objects (or individuals), a class of attributes (or properties), and a binary correspondence between the two classes, indicating for each individual-property pair whether the property applies to the individual or not. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Formal Concept Analysis (FCA) relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting a set of formal concepts then organized within a concept lattice [74] (concept lattices are also known as "Galois lattices" [71]).

In parallel, the search for frequent itemsets and the extraction of association rules are well-known symbolic data mining methods, related to FCA (actually searching for frequent itemsets can be understood as traversing a concept lattice). Both processes usually produce a large number of items and rules, leading to the associated problems of "mining the sets of extracted items and rules". Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow finding interesting subsets of association rules, e.g. informative association rules. This explains why several algorithms are needed for mining data depending on specific applications [86].

For being able to deal with complex and large data, numerical data mining methods can be associated with symbolic methods, for improving applicability and efficiency of knowledge discovery. This is particularly true in classification, where supervised and unsupervised approaches may be combined with benefits [77].

3.2. Text Mining

Keywords: text mining, knowledge discovery from texts, text classification, annotation, ontology engineering from texts.

The objective of a text mining process is to extract useful knowledge units from large collections of texts [67]. The text mining process shows specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making text mining a difficult task. A text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods.

From a knowledge discovery perspective, text mining aims at extracting "interesting units" (nouns and relations) from texts with the help of domain knowledge encoded within a knowledge base. The process is roughly similar for text annotation. Text mining is especially useful in the context of semantic web for ontology engineering. In the Orpailleur team, we work on the mining of real-world texts in application domains such as biology and medicine, using numerical and symbolic data mining methods. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.3. Knowledge Systems and Web of Data

Keywords: knowledge engineering, web of data, semantic web, ontology, description logics, classification-based reasoning, case-based reasoning, information retrieval.

The web of data constitutes a good platform for experimenting ideas on knowledge engineering and knowledge discovery. Following the principles of semantic web, a software agent may be able to read, understand, and manipulate information on the web, if and only if the knowledge necessary for achieving those tasks

is available: this is why knowledge bases (domain ontologies) are of main importance. OWL is the knowledge representation language used to design ontologies and knowledge bases, which is based on description logics (DLs [68]). In OWL, knowledge units are represented by classes (DL concepts) having properties (DL roles) and instances. Concepts can be organized within a partial order based on a subsumption relation, and the inference services are based on satisfiability, classification-based reasoning and case-based reasoning (CBR).

Actually, there are many interconnections between concept lattices in FCA and ontologies, e.g. the partial order underlying an ontology can be supported by a concept lattice. Moreover, a pair of implications within a concept lattice can be adapted for designing concept definitions in ontologies. Accordingly, we are interested here in two main challenges: how the web of data, as a set of potential knowledge sources (e.g. DBpedia, Wikipedia, Yago, Freebase) can be mined for helping the design of definitions and knowledge bases and how knowledge discovery techniques can be applied for providing a better usage of the web of data (e.g. LOD classification).

Accordingly, a part of the research work in Knowledge Engineering is oriented towards knowledge discovery in the web of data, as, with the increased interest in machine processable data, more and more data is now published in RDF (Resource Description Framework) format. Particularly, we are interested in the completeness of the data and their potential to provide concept definitions in terms of necessary and sufficient conditions [69]. We have proposed algorithms based on FCA and Redescription Mining which allow data exploration as well as the discovery of definition (bidirectional implication rules).

4. Application Domains

4.1. Life Sciences: Biology, Chemistry and Medicine

Participants: Miguel Couceiro, Adrien Coulet, Nicolas Jay, Joël Legrand, Jean Lieber, Pierre Monnin, Amedeo Napoli, Abdelkader Ouali, Chedy Raïssi, Malika Smaïl-Tabbone, Yannick Toussaint.

Keywords: knowledge discovery in life sciences, biology, chemistry, medicine, pharmacogenomics and precision medicine.

One major application domain which is currently investigated by the Orpailleur team is related to life sciences, with particular emphasis on biology, medicine, and chemistry. The understanding of biological systems provides complex problems for computer scientists, and the developed solutions bring new research ideas or possibilities for biologists and for computer scientists as well. Indeed, the interactions between researchers in biology and researchers in computer science improve not only knowledge about systems in biology, chemistry, and medicine, but knowledge about computer science as well.

Knowledge discovery is gaining more and more interest and importance in life sciences for mining either homogeneous databases such as protein sequences and structures, or heterogeneous databases for discovering interactions between genes and environment, or between genetic and phenotypic data, especially for public health and precision medicine (pharmacogenomics). Pharmacogenomics is one main challenge for the Orpailleur team as it considers a large panel of complex data ranging from biological to medical data, and various kinds of encoded domain knowledge ranging from texts to formal ontologies.

On the same line as biological data, chemical data are presenting important challenges w.r.t. knowledge discovery, for example for mining collections of molecular structures and collections of chemical reactions in organic chemistry. The mining of such collections is an important task for various reasons among which the challenge of graph mining and the industrial needs (especially in drug design, pharmacology and toxicology). Molecules and chemical reactions are complex data that can be modeled as labeled graphs. Graph mining methods may play an important role in this framework and Formal Concept Analysis can also be used in an efficient and well-founded way [81]. Graph mining as considered in the framework of FCA is an important task on which we are working, whose results can be transferred to text mining as well.

Finally, the so called “projet de recherche exploratoire” (PRE) HyGraMi for “Hybrid Graph Mining for the Design of New Antibacterials” is about the fight against resistance of bacteria to antibiotics. The objective of HyGraMi is to design a hybrid data mining system for discovering new antibacterial agents. This system should rely on a combination of numeric and symbolic classifiers, that will be guided by expert domain knowledge. The analysis and classification of the chemical structures is based on an interaction between symbolic methods e.g. graph mining techniques, and numerical supervised classifiers based on exact and approximate matching.

4.2. Other Application Domains

Participants: Florence Le Ber, Jean Lieber, Jean-François Mari, Amedeo Napoli, Emmanuel Nauer, Sébastien Da Silva.

4.2.1. Cooking

Keywords: cooking, knowledge engineering, case-based reasoning, semantic web

The origin of the Taaable project is the Computer Cooking Contest (CCC). A contestant to CCC is a system that answers queries about recipes, using a recipe base; if no recipe exactly matches the query, then the system adapts another recipe. Taaable is a case-based reasoning system based on knowledge representation, semantic web and knowledge discovery technologies. The system enables to validate scientific results and to study the complementarity of various research trends in an application domain which is simple to understand and which raises complex issues at the same time.

4.2.2. Agronomy

Keywords: simulation in agronomy, graph model in agronomy

Research in agronomy is based on a cooperation between Inria and INRA. The research work is related to the characterization and the simulation of hedgerow structures in agricultural landscapes, based on Hilbert-Peano curves and Markov models [72].

4.2.3. Digital Humanities

Keywords: digital humanities, semantic web, SPARQL, approximate search, case-based reasoning

Members of the Orpailleur team are collaborating with a group of researchers working in history and philosophy of science and technologies (they are located in Brest, Montpellier and Nancy). The idea is to reuse semantic web technologies for better access and better representation of their text corpora.

5. Highlights of the Year

5.1. Highlights of the Year

This year we would like to mention two publications as highlights of the year.

- The first highlight is related to the Snowball Inria Associated Team supervised by Adrien Coulet (see § 8.4.1). The participants to Snowball have obtained very good results in AI and Medicine which have been recently published in the selective journal “Scientific Reports” [4]. In addition, the same participants have obtained a “Grant Seed” funded by Stanford University, to pursue their research efforts in building fair and equitable predictive models for medicine (see http://medicine.stanford.edu/news/current-news/standard-news/presenceannouncesseedgrantawardees.html).
- The second highlight is related to the stay of Chedy Raïssi at NASA lab in 2018 (see § 8.4.3.1). Chedy Raïssi worked with some other researchers on a machine-learning model for classifying signals from local and global views of the light curves. The researchers had the idea of associating expert domain knowledge with the model and they were able to obtain very good results unseen until now (see https://aasnova.org/2018/12/07/using-machine-learning-to-find-planets/?fbclid=IwAR0UI9LcjlSYKh8JNDiJzztwK0OUqxkhtzdTGod20U10JLKO4vm6sPPU990). A publication on this work was accepted and published [2].

6. New Software and Platforms

6.1. ARPEnTAge

Analyse de Régularités dans les Paysages : Environnement, Territoires, Agronomie
KEYWORDS: Stochastic process - Hidden Markov Models
FUNCTIONAL DESCRIPTION: ARPEnTAge is a software based on stochastic models (HMM2 and Markov Field) for analyzing spatio-temporal data-bases. ARPEnTAge is built on top of the CarottAge system to fully take into account the spatial dimension of input sequences. It takes as input an array of discrete data in which the columns contain the annual land-uses and the rows are regularly spaced locations of the studied landscape. It performs a Time-Space clustering of a landscape based on its time dynamic Land Uses (LUS). Displaying tools and the generation of Time-dominant shape files have also been defined.

- Partner: INRA
- Contact: Jean-François Mari
- URL: http://carottage.loria.fr/index_in_english.html

6.2. CarottAge

KEYWORDS: Stochastic process - Hidden Markov Models
FUNCTIONAL DESCRIPTION: The system CarottAge is based on Hidden Markov Models of second order and provides a non supervised temporal clustering algorithm for data mining and a synthetic representation of temporal and spatial data. CarottAge is currently used by INRA researchers interested in mining the changes in territories related to the loss of biodiversity (projects ANR BiodivAgrim and ACI Ecoger) and/or water contamination. CarottAge is also used for mining hydromorphological data. Actually a comparison was performed with three other algorithms classically used for the delineation of river continuum and CarottAge proved to give very interesting results for that purpose.

- Participants: Florence Le Ber and Jean-François Mari
- Partner: INRA
- Contact: Jean-François Mari
- URL: http://carottage.loria.fr/index_in_english.html

6.3. CORON

KEYWORDS: Data mining - Closed itemset - Frequent itemset - Generator - Association rule - Rare itemset
FUNCTIONAL DESCRIPTION: The Coron platform is a KDD toolkit organized around three main components: (1) Coron-base, (2) AssRuleX, and (3) pre- and post-processing modules.

The Coron-base component includes a complete collection of data mining algorithms for extracting itemsets such as frequent itemsets, closed itemsets, generators and rare itemsets. In this collection we can find APriori, Close, Pascal, Eclat, Charm, and, as well, original algorithms such as ZART, Snow, Touch, and Talky-G. AssRuleX generates different sets of association rules (from itemsets), such as minimal non-redundant association rules, generic basis, and informative basis. In addition, the Coron system supports the whole life-cycle of a data mining task and proposes modules for cleaning the input dataset, and for reducing its size if necessary.

- Participants: Adrien Coulet, Aleksey Buzmakov, Amedeo Napoli, Florent Marcuola, Jérémie Bourseau, Laszlo Szathmary, Mehdi Kaytoue, Victor Codocedo and Yannick Toussaint
- Contact: Amedeo Napoli
- URL: http://coron.loria.fr/site/index.php

6.4. LatViz: Visualization of Concept Lattices

- Contact: Amedeo Napoli
- URL: http://latviz.loria.fr/
- KEYWORDS: Formal Concept Analysis, Pattern Structures, Concept Lattice, Implications, Visualization

FUNCTIONAL DESCRIPTION.

LatViz is a tool allowing the construction, the display and the exploration of concept lattices. LatViz proposes some noticeable improvements over existing tools and introduces various functionalities focusing on interaction with experts, such as visualization of pattern structures for dealing with complex non-binary data, AOC-poset which is composed of the core elements of the lattice, concept annotations, filtering based on various criteria and a visualization of implications [70]. This way the user can effectively perform interactive exploratory knowledge discovery as often needed in knowledge engineering.

The LatViz platform can be associated with the Coron platform and extends its visualization capabilities (see http://coron.loria.fr). Recall that the Coron platform includes a complete collection of data mining algorithms for extracting itemsets and association rules.

6.5. OrphaMine: Data Mining Platform for Orphan Diseases

- Contact: Chedy Raïssi
- URL: http://orphamine.inria.fr/
- KEYWORDS: Bioinformatics, data mining, biology, health, data visualization, drug development.

FUNCTIONAL DESCRIPTION.

The OrphaMine platform enables visualization, data integration and in-depth analytics in the domain of "orphan diseases", where data is extracted from the OrphaData ontology (http://www.orpha.net/consor/cgi-bin/index.php). At present, we aim at building a true collaborative portal that will serve different actors: (i) a general visualization of OrphaData data for physicians working, maintaining and developing this knowledge database about orphan diseases. (ii) the integration of analytics (data mining) algorithms developed by the different academic actors. (iii) the use of these algorithms to improve our general knowledge of rare diseases.

6.6. Siren: Interactive and Visual Redescription Mining

- Contact: Esther Catherine Galbrun
- URL: http://siren.gforge.inria.fr/main/
- KEYWORDS: Redescription mining, Interactivity, Visualization.

FUNCTIONAL DESCRIPTION.

Siren is a tool for interactive mining and visualization of redescriptions. Redescription mining aims to find distinct common characterizations of the same objects and, vice versa, to identify sets of objects that admit multiple shared descriptions. The goal is to provide domain experts with a tool allowing them to tackle their research questions using redescription mining. Merely being able to find redescriptions is not enough. The expert must also be able to understand the redescriptions found, adjust them to better match his domain knowledge and test alternative hypotheses with them, for instance. Thus, Siren allows mining redescriptions in an anytime fashion through efficient, distributed mining, to examine the results in various linked visualizations, to interact with the results either directly or via the visualizations, and to guide the mining algorithm toward specific redescriptions.

New features, such as a visualization of the contribution of individual literals in the queries and the simplification of queries as a post-processing, have been added to the tool.

7. New Results

7.1. Mining of Complex Data

Participants: Nacira Abbas, Guilherme Alves Da Silva, Alexandre Blansché, Lydia Boudjeloud-Assala, Quentin Brabant, Brieuc Conan-Guez, Miguel Couceiro, Adrien Coulet, Alain Gély, Laurine Huber, Nyoman Juniarta, Florence Le Ber, Joël Legrand, Pierre Monnin, Tatiana Makhalova, Amedeo Napoli, Abdelkader Ouali, François Pirot, Frédéric Pennerath, Justine Reynaud, Chedy Raïssi, Sébastien Da Silva, Yannick Toussaint.

Keywords: formal concept analysis, relational concept analysis, pattern structures, pattern mining, association rule, redescription mining, graph mining, sequence mining, biclustering, hybrid mining, meta-mining

7.1.1. FCA and Variations: RCA, Pattern Structures and Biclustering

Advances in data and knowledge engineering have emphasized the needs for pattern mining tools working on complex data. In particular, FCA, which usually applies to binary data-tables, can be adapted to work on more complex data. In this way, we have contributed to two main extensions of FCA, namely Pattern Structures and Relational Concept Analysis. Pattern Structures (PS [73]) allow building a concept lattice from complex data, e.g. numbers, sequences, trees and graphs. Relational Concept Analysis (RCA) is able to analyze objects described both by binary and relational attributes [84] and can play an important role in text classification and text mining. Many developments were carried out in pattern mining and FCA for improving data mining algorithms and their applicability, and for solving some specific problems such as information retrieval, discovery of functional dependencies and biclustering.

We got several results in the discovery of approximate functional dependencies [8], the mining of RDF data and the and visualization of the discovered patterns [1], and redescription mining (detailed later). Moreover, we have also investigated the use of the MDL principle ("Minimum Description Length") for the selection of interesting and diverse patterns [37], [39].

In the framework of the CrossCult European Project about cultural heritage, we worked on the mining of visitor trajectories in a museum or a touristic site. We presented a theoretical and practical research work about the characterization of visitor trajectories and the mining of these trajectories as sequences [32], [33]. The mining process is based on two approaches in the framework of Formal Concept Analysis (FCA). We focused on different types of sequences and more precisely on subsequences without any constraint and frequent contiguous subsequences. In parallel, we introduced a similarity measure allowing us to build a hierarchical classification which is used for interpretation and characterization of the trajectories. In addition, for completing the research work on the characterization of trajectories, we also studied how biclustering may be applied to trajectory recommendation [31], [52].

7.1.2. Redescription Mining

Among the mining methods developed in the team is redescription mining. Redescription mining aims to find distinct common characterizations of the same objects and, vice versa, to identify sets of objects that admit multiple shared descriptions [82]. It is motivated by the idea that in scientific investigations data oftentimes have different nature. For instance, they might originate from distinct sources or be cast over separate terminologies. In order to gain insight into the phenomenon of interest, a natural task is to identify the correspondences that exist between these different aspects.

A practical example in biology consists in finding geographical areas that admit two characterizations, one in terms of their climatic profile and one in terms of the occupying species. Discovering such redescriptions can contribute to better our understanding of the influence of climate over species distribution. Besides biology, applications of redescription mining can be envisaged in medicine or sociology, among other fields.

This year, we used redescription mining for analyzing and mining RDF data with the objective of discovering definitions of concepts and as well disjunctions (incompatibilities) of concepts, for completing knowledge bases in a semi-automated way [49], [44].

7.1.3. Text Mining

In the context of the PractikPharma ANR Project, we study how cross-corpus training may guide the task of relationship extraction from texts, and especially, how large annotated corpora developed for alternative tasks may improve the performance of biomedical tasks, for which only a few annotated resources are available [34].

Transfer learning proposes to enhance machine learning performance on a problem, by reusing labeled data originally designed for a related problem. This is particularly relevant to the applications of deep learning in Natural Language Processing, because those usually require large annotated corpora that may not exist for the targeted domain, but exist for side domains. In a recent work, we experimented the extraction of relationships from biomedical texts with two deep learning models. The first model combines locally extracted features using a Multi Channel Convolutional Neural Network (MCCNN) model, while the second model exploits the syntactic structure of sentences using a Tree-LSTM (Long Short-Term Memory) architecture. The experiments show that the Tree-LSTM model benefits from a cross-corpus learning strategy, i.e. performances are improved when training data are enriched with off-target corpora, whereas it is not the case with MCCNN.

Indeed our approach leads to state of the art performances in four biomedical tasks for which only a few annotated resources are available (less than 400 manually annotated sentences) and even surpass state of the art performances in two of these four tasks. We particularly investigated how the syntactic structure of a sentence, which is domain independent, participates in the increase of performance when adding additional training data. This may have a particular impact in specialized domains in which training resources are scarce, because it means that these resources may be efficiently enriched with data from other domains for which large annotated corpora exist.

7.1.4. Mining subgroups as a single-player game

Discovering patterns that strongly distinguish one class label from another is a challenging data-mining task. The unsupervised discovery of such patterns would enable the construction of intelligible classifiers and to elicit interesting hypotheses from the data. Subgroup Discovery (SD) is one framework that formally defines this pattern mining task. However, SD still faces two major issues: (i) how to define appropriate quality measures to characterize the uniqueness of a pattern; (ii) how to select an accurate heuristic search technique when exhaustive enumeration of the pattern space is unfeasible. The first issue has been tackled by the Exceptional Model Mining (EMM) framework. This general framework aims to find patterns that cover tuples that locally induce a model that substantially differs from the model of the whole dataset. The second issue has been studied in SD and EMM mainly with the use of beam-search strategies and genetic algorithms for discovering a pattern set that is non-redundant, diverse and of high quality. Consequently,

In our current work [9], we proposed to formally define pattern mining as a single-player game, as in a puzzle, and to solve it with a Monte Carlo Tree Search (MCTS), a technique mainly used for artificial intelligence and planning problems. The exploitation/exploration trade-off and the power of random search of MCTS lead to an any-time mining approach, in which a solution is always available, and which tends towards an exhaustive search if given enough time and memory. Given a reasonable time and memory budget, MCTS quickly drives the search towards a diverse pattern set of high quality. MCTS does not need any knowledge of the pattern quality measure, and we show to what extent it is agnostic to the pattern language.

7.1.5. Consensus and Aggregation Functions

Aggregation and consensus theory study processes dealing with the problem of merging or fusing several objects, e.g., numerical or qualitative data, preferences or other relational structures, into a single or several objects of similar type and that best represents them in some way. Such processes are modeled by so-called aggregation or consensus functions [76], [78]. The need to aggregate objects in a meaningful way appeared naturally in classical topics such as mathematics, statistics, physics and computer science, but it became

increasingly emergent in applied areas such as social and decision sciences, artificial intelligence and machine learning, biology and medicine.

We are working on a theoretical basis of a unified theory of consensus and to set up a general machinery for the choice and use of aggregation functions. This choice depends on properties specified by users or decision makers, the nature of the objects to aggregate as well as computational limitations due to prohibitive algorithmic complexity. This problem demands an exhaustive study of aggregation functions that requires an axiomatic treatment and classification of aggregation procedures as well as a deep understanding of their structural behavior. It also requires a representation formalism for knowledge, in our case decision rules and methods for discovering them. Typical approaches include rough-set and FCA approaches, that we aim to extend in order to increase expressivity, applicability and readability of results. Applications of these efforts already appeared and further are expected in the context of three multidisciplinary projects, namely the "Fight Heart Failure" (research project with the Faculty of Medicine in Nancy), the European H2020 "CrossCult" project, and the "ISIPA" (Interpolation, Sugeno Integral, Proportional Analogy) project.

In the context of the project RHU "Fighting Heart Failure" (that aims to identify and describe relevant bio-profiles of patients suffering from heart failure) we are dealing with biomedical data, highly complex and heterogeneous, that include, among other, sociodemographical aspects, biological and clinical features, drugs taken by the patients, etc. One of our main challenges is to define relevant aggregation operators on this heterogeneous patient data that lead to a clustering of the patients. Each cluster should correspond to a bio-profile, i.e. a subgroup of patients sharing the same form of the disease and thus the same diagnosis and medical care strategy. We are working on ways for comparing and clustering patients, namely, by defining multidimensional similarity measures on this complex and heterogeneous biomedical data. To this end, we recently proposed a novel approach, that we named "unsupervised extremely randomized trees" (UET) [27], that is inspired by the frameworks of unsupervised random forests (URF) [85] and of extremely randomized trees (ET) [75]. The empirical study of UET showed that it outperforms existing methods (such as URF) in running time, while giving better clustering. However, UET was implemented for numerical data only, and this is a drawback when dealing with biomedical data. We are now working on the adaptation of UET for heterogeneous data (both numerical and symbolic), possibly, with missing values.

In the context of the project ISIPA, we mainly focused on the utility-based preference model in which preferences are represented as an aggregation of preferences over different attributes, structured or not, both in the numerical and qualitative settings. In the latter case, the Sugeno integral is widely used in multiple criteria decision making and decision under uncertainty, for computing global evaluations of items based on local evaluations (utilities). The combination of a Sugeno integral with local utilities is called a Sugeno utility functional (SUF). A noteworthy property of SUFs is that they represent multi-threshold decision rules. However, not all sets of multi-threshold rules can be represented by a single SUF. We showed how to represent any set of multi-threshold rules as a combination of SUFs and studied their potential advantages as a compact representation of large sets of rules, as well as an intermediary step for extracting rules from empirical datasets [51]. For further results in the qualitative approach to decision making see, e.g., [10] [3]; and see also [24] for a survey chapter on new perspectives in ordinal evaluation.

7.2. Knowledge Discovery in Healthcare and Life Sciences

Participants: Miguel Couceiro, Adrien Coulet, Nicolas Jay, Joël Legrand, Pierre Monnin, Amedeo Napoli, Abdelkader Ouali, Chedy Raïssi, Malika Smaïl-Tabbone, Yannick Toussaint.

7.2.1. Ontology-based Clustering of Biological Data

Biomedical objects can be characterized by ontology annotations. For example, Gene Ontology annotations provide information on the functions of genes, while Human Phenotype Ontology (HPO) annotations provide information about phenotypes associated with diseases. It is usual to consider such annotations in the analysis of biomedical data, most of the time annotations from only one single ontology. However, complex objects such as diseases can be annotated at the same time w.r.t. different ontologies, making clear distinct dimensions. We are investigating how annotations from several ontologies may be cooperating in disease classification. In

particular, we classified Genetic Intellectual Disabilities (GID), on the basis of their HPO annotations and of GO annotations of genes known for being responsible for these diseases [43]. We used clustering algorithms based on semantic similarities and enabling to compare sets of annotations. This experiment illustrates the fact that considering several ontologies provides better results, while selecting the best set of ontologies to combine is dependent on the dataset and on the classification task.

7.2.2. Validation of Pharmacogenomic Knowledge

State of the art knowledge in pharmacogenomics is heterogeneous w.r.t. validation. A part is well validated, observed on a large population and already used in clinical practice, while a large majority of this knowledge is lacking validation and reproducibility, mainly because of scarce observation. Accordingly, validating state of the art knowledge in pharmacogenomics by mining Electronic Health Records (EHRs) is one objective of the ANR project "PractiKPharma" initiated in 2016 (http://practikpharma.loria.fr/).

To lead this validation, we define a minimal data schema for pharmacogenomic knowledge units (PGxO ontology), which is instantiated with data of various provenance (e.g. biomedical databases, literature and EHR). Such an instantiation produces a unique knowledge graph named PGxLOD (https://pgxlod.loria.fr/). We defined and applied a first set of reconciliation rules that compare and align whenever possible knowledge elements of various provenance. A journal article on the construction of PGxLOD and its use in knowledge comparison is currently under evaluation. We are continuing this effort by studying methods which enable a more flexible knowledge comparison.

In addition, we took part to the Biohackathon 2018 Paris (https://bh2018paris.info/) during which we worked on two tasks. Firstly we updated PGxLOD for improving its quality, completeness and interconnection with other resources. Secondly we mined PGxLOD and searched for explanations of the molecular mechanism of adverse drug responses. PGxLOD is under evaluation for being registered as a resource of the IBF (*French Institute for Bioinformatics*) and Elixir (an international organization that supports and structures bioinformatics efforts in Europe).

7.2.3. Mining Electronic Health Records

In the context of the Snowball Inria Associate Team, we developed an approach based on pattern structures to identify frequently associated ADRs (Adverse Drug Reactions) from patient data either in the form of EHR or ADR spontaneous reports. Pattern structures provide an expressive representation of ADR, taking into account the multiplicity of drugs and phenotypes involved in such reactions. Additionally, pattern structures allow considering diverse biomedical ontologies used to represent or annotate patient data, enabling a "semantic" comparison of ADRs. Up to now, this is one of the first research attempts considering such representations to mine rules between frequently associated ADRs. We illustrated the generality of the approach on two patient datasets, each of them linked to distinct biomedical ontologies. The first dataset corresponds to anonymized EHRs, extracted from "STRIDE", the EHR data warehouse of Stanford Hospital and Clinics. The second dataset is extracted from the U.S. FDA (for Food & Drug Administration) "Adverse Event Reporting System" (FAERS). Several significant association rules have been extracted, analyzed and may be used as a basis for a recommendation system.

In collaboration with Stanford University and the CHRU Nancy, we studied the use of Electronic Health Records to predict at first prescription the need for a patient to be prescribed with a reduced drug dose [4]. We particularly focused on drugs whose dosage is known to be sensitive and variable. We used data from the Stanford Hospital to construct cohorts of patients that either did or did not need a dose change for each considered drug. After feature selection, we trained Random Forest models which successfully predict whether a new patient will or not require a dose change after being prescribed one of 23 drugs among 22 drug classes. Several of these drugs are related to clinical guidelines that recommend dose reduction exclusively in the case of adverse reaction. For these cases, a reduction in dosage may be considered as a surrogate for an adverse reaction, which our system could help predicting and preventing.

7.3. Knowledge Engineering and Web of Data

Participants: Nicolas Jay, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer, Justine Reynaud, Yannick Toussaint.

Keywords: knowledge engineering, web of data, definition mining, classification-based reasoning, case-based reasoning, belief revision, semantic web

7.3.1. Current Trends in Case-Based Reasoning

Case-based reasoning (CBR) aims at solving a new problem, called the target problem, by exploiting past experiences (i.e. source cases) as well as other knowledge sources: domain knowledge, similarity knowledge and adaptation knowledge.

Two research works were carried out about how exploiting at the best the source cases. A first work addresses the exploitation of negative cases for adaptation knowledge discovery. Usually CBR exploits positive source cases consisting of a source problem and its solution that is known to be correct for the problem. However, negative cases, i.e. problem-solution pairs where the solution is an incorrect answer to the problem, which can be acquired when CBR process fails, are useful, especially for adaptation knowledge discovery. In [29], we propose an adaptation knowledge discovery approach exploiting both type of cases (positive and negatives cases), using closed itemsets built on variations between cases. Experiments show that exploiting negative cases in addition to positive ones improves the quality of the adaptation knowledge being extracted and, so, improves the results of the CBR system.

A second work addresses the issue of the selection of source cases used to solve a target problem. Three approaches have been studied to better exploit source cases: (1) approximation, which considers the use of one source case (the most similar to the target problem) to solve the target problem, (2) interpolation, which considers the use of two source cases (such as the target problem is between these two similar source problems), and (3) extrapolation, which considers the use of three source cases, linked to the target problem by an analogical proportion, where the analogical proportion handles both similarity and dissimilarity between cases. Experiments show that interpolation and extrapolation techniques are of interest for reusing cases, either in an independent or in a combined way [36], [47].

Using analogical proportion has also been used to find relevant pathology-gene pairs [28]. This first study to infer pathology-gene relation is based on the following hypothesis: if a target pathology is in analogy with three other pathologies for which associated genes are known, then it is plausible that the gene to be associated with the target pathology is in analogy with the genes associated to the three pathologies involved in the analogical proportion.

Another use of analogical proportion is its application to machine translation and is based on a similar principle: if four sentences form an analogical proportion in a language, then it is plausible that their translations in another language also form an analogical proportion. This was the idea developed by Yves Lepage (Waseda University), a few years ago. Now, a starting work on case-based machine translation aims at developing these ideas by incorporation other knowledge sources to the CBR system than the cases (domain knowledge, retrieval knowledge and adaptation knowledge) [35].

Another work on CBR is its application to medical coding. Cancer registries are important tools in the fight against cancer. At the heart of these registries is the data collection and coding process. Ruled by complex international standards and numerous best practices, operators are easily overwhelmed. In [54], [55], a system is presented to assist operators in the interpretation of best medical coding practices.

There has been another work on CBR related to an application in agronomy developed some time ago that has been synthesized in [60].

7.3.2. Exploring and Classifying the Web of Data

A part of the research work in Knowledge Engineering is oriented towards knowledge discovery in the web of data, following the increase of data published in RDF (Resource Description Framework) format and the interest in machine processable data. The quick growth of Linked Open Data (LOD) has led to challenging aspects regarding quality assessment and data exploration of the RDF triples that shape the LOD cloud. In the

team, we are particularly interested in the completeness of the data viewed as their their potential to provide concept definitions in terms of necessary and sufficient conditions [69]. We have proposed a novel technique based on Formal Concept Analysis which classifies subsets of RDF data into a concept lattice [83]. This allows data exploration as well as the discovery of implication rules which are used to automatically detect possible completions of RDF data and to provide definitions. Moreover, this is a way of reconciling syntax and semantics in the LOD cloud. Experiments on the DBpedia knowledge base shows that this kind of approach is well-founded and effective [44].

In the same way, FCA can be used to improve ontologies associated with the Web of data. Accordingly, we proposed a method to build a concept lattice from linked data and compare the structure of this lattice with an ontology used to type the considered data. The result of this comparison makes clear some alternative axioms to be proposed to ontology developers. We extended and reused this work in ontology alignment tasks [41].

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. AGREV-3

Participant: Jean-François Mari.

The AGREV 3 project (for "Agriculture Environment Vittel") is part of "Agrivair" –a subsidiary of Nestlé Waters– in actions to protect the natural resources of natural mineral water. We used ARPEnTAge to mine survey data about the Vittel-Contrexéville territory, which is suspected of groundwater quality risks [5]. This allowed to locate regions having the same behavior. In addition, this provided a more contrasted simulation by eliminating the influence of stable zones (forests, permanent grasslands) and a more precise definition of a "neutral" model.

8.1.2. Hydreos

Participants: Jean-François Mari, Chedy Raïssi.

Hydreos is a state organization, so-called "Pôle de compétitivité", aimed at monitoring and evaluating the quality of water and its delivery (http://www.hydreos.fr/fr). Actually, data about water resources rely on many agronomic variables, including land use successions. The data to be analyzed are obtained by surveys or by satellite images and describe the land use at the level of the agricultural parcel. Then there is a search for detecting changes in land use and for correlating these changes to groundwater quality. Accordingly, one main challenge in our participation in Hydreos is to process and analyze space-time data for reaching a better understanding of the changes in the organization of a territory. The systems ARPEnTAge and CarottAge are used in this context, especially by agronomists of INRA (ASTER Mirecourt http://www6.nancy.inra.fr/sad-aster).

On other aspects, we tested new deep graph convolutional learning over data provided by the SEDIF "Syndicat des eaux d'Île-de-France" to predict the likelihood of water leaks in a network of pipes and compared it with a master thesis where spatial point process techniques were used (master thesis of Nicolas Dante, M2 IMSD Nancy).

8.1.3. The Smart Knowledge Discovery Project

Participants: Jérémie Nevin, Amedeo Napoli, Chedy Raïssi.

The SKD project for "Smart Knowledge Discovery" aims at analyzing complex industrial data for troubleshooting and decision making, and is funded by "Grand Est Region". We are working on exploratory knowledge discovery with the Vize company, which is based in Nancy and specialized in visualization-based data mining. The data which are under study are provided by the Arcelor-Mittal Steel Company and are related to the monitoring of rolling mills. Data are complex time series and the problem is related to a so-called "predictive maintenance", or how to anticipate problems in the furnaces and avoid their stop. In this way, one main objective of SKD is to combine sequence mining and visualization tools for recognizing temperature problems in the furnaces, and thus preventing the occurrences of defects in the outputs of the rolling mills.

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. Elker (2017–2020)

Participants: Nacira Abbas, Miguel Couceiro, Amedeo Napoli, Chedy Raïssi.

The objectives of the ELKER ANR Research Project is to study, formalize and implement the search for link keys in RDF data. Link keys generalize database keys in two independent directions, i.e. they deal with RDF data and they apply across two relation datasets. Then we study the automatic discovery of link keys and reasoning with link keys, in taking an FCA point of view. The projects relies on the competencies of Orpailleur in FCA for solving the problem using FCA and pattern structures algorithms, partition pattern structures which are related to the discovery of functional dependencies. This project involves the EPI Orpailleur at Inria Nancy Grand Est, the EPI MOEX at Inria Rhône Alpes, and LIASD at Université Paris 8.

8.2.1.2. PractiKPharma (2016–2020)

Participants: Adrien Coulet, Joël Legrand, Pierre Monnin, Amedeo Napoli, Malika Smaïl-Tabbone, Yannick Toussaint.

PractiKPharma for "Practice-based evidences for actioning Knowledge in Pharmacogenomics" is an ANR research project (http://practikpharma.loria.fr/) about the validation of domain knowledge in pharmacogenomics. Pharmacogenomics is interested in understanding how genomic variations related to patients have an impact on drug responses. Most of the available knowledge in pharmacogenomics (state of the art) lies in biomedical literature, with various levels of validation. An originality of PractiKPharma is to use Electronic Health Records (EHRs) to constitute cohorts of patients. These cohorts are then mined for extracting potential pharmacogenomics patterns to be then validated w.r.t. literature knowledge for becoming actionable knowledge units. More precisely, firstly we should extract pharmacogenomic patterns from the literature and secondly we should confirm or moderate the interpretation and validation of these units by mining EHRs. Comparing knowledge patterns extracted from the literature with facts extracted from EHRs is a complex task depending on the EHR language –literature is in English whereas EHRs are in French– and on knowledge level, as EHRs represent observations at the patient level whereas literature is related to sets of patients. The PractiKPharma involves three other laboratories, namely LIRMM in Montpellier, SSPIM in St-Etienne and CRC in Paris.

8.2.2. CNRS Mastodons Projects: HyQual, HyQualiBio and QCM-BioChem (2016–2018)

Participants: Nacira Abbas, Guilherme Alves Da Silva, Miguel Couceiro, Alain Gély, Nyoman Juniarta, Tatiana Makhalova, Amedeo Napoli, Chedy Raïssi, Justine Reynaud.

The HyQual project was proposed in 2016 in response to the Mastodons CNRS Call about data quality in data mining (see http://www.cnrs.fr/mi/spip.php?article819&lang=fr). This project is interested in the mining of nutritional data for discovering predictive biomarkers of diabetes and metabolic syndrome in elder populations. The considered data mining methods are hybrid, and they combine symbolic and numerical methods for mining complex and noisy metabolic data [77]. Regarding the mining process, we are interested in the quality of the data at hand and in the discovered patterns. In particular, we check the incompleteness of the data, the quality of the extracted rules and the possible existence of redescriptions.

Initially, the project involved researchers from the EPI Orpailleur, with researchers from LIRIS Lyon, ICube Strasbourg, and INRA Clermont-Ferrand. Then, the project was merged the other Mastodons project named QualiBioConsensus, about the "ranking of biological data using consensus ranking techniques". The joint Mastodons project was called "HyQualiBio". The year after, the project was a new time merged with the PEPS Decade project to form the new "QCM-BioChem" (https://www.lri.fr/~cohen/QCM-BioChem.html). The topics of interest for the participants are the mining of complex biological data, rankings and ties in rankings, and the search of dependencies in the web of data.

8.3. European Initiatives

8.3.1. FP7 & H2020 Projects

8.3.1.1. CrossCult (H2020 Project, 2016-2020)

Participants: Miguel Couceiro, Nyoman Juniarta, Amedeo Napoli, Chedy Raïssi.

CrossCult aims at making reflective history a reality in the European cultural context, by enabling the re-interpretation of European (hi)stories through cross-border interconnections among cultural digital resources, citizen viewpoints and physical venues. The project has two main goals. The first goal is to lower cultural EU barriers and create unique cross-border perspectives, by connecting existing digital historical resources and by creating new ones through the participation of the public. The second goal is to provide long-lasting experiences of social learning and entertainment that will help for achieving a better understanding and re-interpretation of European history. To achieve these goals, CrossCult aims at using cutting-edge technology to connect existing digital cultural assets and to combine them with interactive experiences that all together are intended to increase retention, stimulate reflection and help European citizens appreciate their past and present in a holistic manner. CrossCult has to be implemented on four real-world flagship pilots involving a total of 8 sites across Europe.

The role of the Orpailleur Team (in conjunction with the LORIA Kiwi Team) is to work on knowledge discovery and recommendation. The focus is on the mining of visitor trajectories for analysis purposes [32], [33] and on the definition of a visitor profile in connection with domain knowledge for recommendation [31].

The numerous partners of the Orpailleur team in the CrossCult project are: Luxembourg Institute for Science and Technology and Centre Virtuel de la Connaissance sur l'Europe (Luxembourg, leader of the project), University College London (England), University of Malta (Malta), University of Peloponnese and Technological Educational Institute of Athens (Greece), Università degli Studi di Padova (Italy), University of Vigo (Spain), National Gallery (London, England), and GVAM Guìas Interactivas (Spain).

8.4. International Initiatives

8.4.1. Inria International Labs

Inria@SiliconValley

Associate Team involved in the International Lab:

8.4.1.1. Snowball

Title: Discovering knowledge on drug response variability by mining electronic health records

International Partner (Institution - Laboratory - Researcher):

Stanford (United States) - Department of Medicine, Stanford Center for Biomedical Informatics Research (BMIR) - Nigam Shah

Start year: 2017

See also: http://snowball.loria.fr/

Snowball (2017-2019) is an Inria Associate Team and the continuation of the preceding Associate Team called Snowflake (2014-2016). The objective of Snowball is to study drug response variability through the lens of Electronic Health Records (EHRs) data. This is motivated by the fact that many factors, genetic as well as environmental, imply different responses from people to the same drug. The mining of EHRs can bring substantial elements for understanding and explaining drug response variability.

Accordingly the objectives of Snowball are to identify in EHR repositories groups of patients which are responding differently to similar treatments, and then to characterize these groups and predict patient drug sensitivity. These objectives are complementary to those of the PractiKPharma ANR project. Moreover, it should be noticed that Adrien Coulet is continuing a two-years sabbatical

stay in the lab of Nigam Shah at Stanford University since September 2017 (granted by an "Inria délégation").

Participants of the Snowball Associate Team have been awarded with a Grant Seed funded by Stanford University, to pursue their efforts in AI in Medicine. The granted project will particularly focus on the building of fair and equitable predictive models for medicine (see http://medicine.stanford.edu/news/current-news/standard-news/presenceannouncesseedgrantawardees.html).

8.4.2. Informal International Partners: Research Collaboration with HSE Moscow

Participants: Nacira Abbas, Guilherme Alves Da Silva, Miguel Couceiro, Alain Gély, Nyoman Juniarta, Tatiana Makhalova, Amedeo Napoli, Chedy Raïssi, Justine Reynaud.

An on-going collaboration involves the Orpailleur team and Sergei O. Kuznetsov at Higher School of Economics in Moscow (HSE). Amedeo Napoli visited HSE laboratory several times while Sergei O. Kuznetsov visits Inria Nancy Grand Est every year. The collaboration is materialized by the joint supervision of students (such as the thesis of Aleksey Buzmakov defended in 2015 and the on-going thesis of Tatiana Makhalova), and the the organization of scientific events, as the workshop FCA4AI with six editions between 2012 and 2018 (see http://www.fca4ai.hse.ru).

This year, we participated in the writing of common publications around the thesis work of Tatiana Makhalova and the organization of one main event, namely the sixth edition of the FCA4AI workshop in July 2018 at the ECAI-IJCAI Conference which was held in Stockholm, Sweden (see http://ceur-ws.org/Vol-2149, [58]).

8.4.3. Participation in other International Programs

8.4.3.1. A stay at NASA Frontier Development Lab

In July and August 2018, Chedy Raïssi visited NASA Ames and SETI Institute as part of the Frontier Development Lab, where he worked on mentoring teams and developing meaningful research opportunities, as well as support the work of the planetary defense community and show the potential of this kind of applied research methodology to deliver breakthrough of significant value.

During the eight-week research incubator he aimed at applying cutting-edge machine-learning algorithms to challenges in the space sciences. He worked with two machine-learning students (PhD and post-doc level) that were paired with two space-science researchers (post-doc level) on the improvement of machine-learning models for exoplanet transit classification. This small team started initially from a machine-learning model that classified signals based on straightforward local and global views of the light curves that was developed by Google Brain engineer Chris Shallue. To improve upon it, the team added scientific domain knowledge –staying true to the Orpailleur idea of injecting domain knowledge– that was provided by domain experts. Using the resulting model, the team managed to classify a Kepler data set with 97.5% accuracy and 98% average precision [2].

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organization, General Chairs, Scientific Chairs

- Amedeo Napoli and Yannick Toussaint were the general chairs of the "21th International Conference on Knowledge Engineering and Knowledge Management" (EKAW 2018, https://project.inria.fr/ekaw2018/) held on November 12–16 2018 at Inria NGE/LORIA Nancy.
- Amedeo Napoli was the program chair with Sergei O. Kuznetsov (HSE Moscow) and Sebastian Rudolph (TU Dresden) of the sixth workshop FCA4AI ("What can do FCA for Artificial Intelligence") co-located with the IJCAI-ECAI Conference in Stockholm, July 13 2018 (http://fca4ai.hse.ru/ and http://ceur-ws.org/Vol-2149/).

- Amedeo Napoli was the co-chair with Sergei Kuznetsov of the track “General Topics of Data Analysis” at the AIST Conference in Moscow on July 5–7 2018 (7th International Conference on Analysis of Images, Social Networks, and Texts http://aistconf.org/ and http://aistconf.org/board/).
- Miguel Couceiro was an organizer of the tutorial on “Majority Logic Synthesis” at the International Conference On Computer Aided Design (ICCAD 2018, https://iccad.com/agenda/embedded-tutorials, [23]).
- Miguel Couceiro and Jérôme David (Inria Rhône Alpes, MOEX) were the organizers of the workshop “Symbolic methods for data-interlinking” co-located with EKAW 2018 (https://project.inria.fr/ekaw2018/workshops/). This workshop was organized in the framework of EKAW 2018 and of the ANR project ELKER.

9.1.2. Scientific Animation

- The scientific animation in the Orpailleur team is based on the Team Seminar which is called the “Malotec” seminar (http://malotec.loria.fr/). The Malotec seminar is held in general twice a month and is used either for general presentations of members of the team or for invited presentations of external researchers.
- Members of the Orpailleur team are all involved, as members or as head persons, in various national research groups.
- The members of the Orpailleur team are involved in the organization of conferences and workshops, as members of conference program committees (AAAI, ECAI, ECML-PKDD, ESWC, ICCBR, ICDM, ICFCA, IJCAI, ISWC, KDD, SDM...), as members of editorial boards, and finally in the organization of journal special issues.

9.2. Teaching - Supervision - Juries

- All the permanent members of the Orpailleur team are involved in teaching at all levels and mainly at University of Lorraine. Actually, most of the members of the Orpailleur team are employed on “Université de Lorraine” positions.
- The members of the Orpailleur team are also involved in student supervision, at all university levels, from under-graduate until post-graduate students, engineers, PhD, postdoc students.
- Finally, the permanent members of the Orpailleur team are involved in HDR and thesis defenses, being thesis referees or thesis committee members.

10. Bibliography

Major publications by the team in recent years

[1] M. ALAM, A. BUZMAKOV, A. NAPOLI. *Exploratory Knowledge Discovery over Web of Data*, in "Discrete Applied Mathematics", 2018, vol. 249, pp. 2-17, https://hal.inria.fr/hal-01673439

[2] M. ANSDELL, Y. IOANNOU, H. OSBORN, M. SASDELLI, J. SMITH, D. CALDWELL, J. JENKINS, C. RAÏSSI, D. ANGERHAUSEN. *Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning*, in "The Astrophysical Journal Letters", December 2018, vol. 869, n^{o} 1, L7 p. [*DOI :* 10.3847/2041-8213/AAF23B], https://hal.inria.fr/hal-01957950

[3] M. COUCEIRO, M. MARÓTI, T. WALDHAUSER, L. ZADORI. *Computing version spaces in the qualitative approach to multicriteria decision aid* , in "International Journal of Foundations of Computer Science", 2018, https://hal.inria.fr/hal-01404590

[4] A. COULET, N. H. SHAH, M. WACK, M. CHAWKI, N. JAY, M. DUMONTIER. *Predicting the need for a reduced drug dose, at first prescription*, in "Scientific Reports", October 2018, vol. 8, n° 1 [*DOI :* 10.1038/s41598-018-33980-0], https://hal.inria.fr/hal-01901566

[5] J.-F. MARI, A. GOBILLOT, M. BENOÎT. *Time Space Simulation of Land Use changes by stochastic modeling*, in "Revue Internationale de Géomatique", August 2018, vol. 28, n° 2, pp. 219 - 242, https://hal.inria.fr/hal-01662140

Publications of the year

Articles in International Peer-Reviewed Journals

[6] M. ALAM, A. BUZMAKOV, A. NAPOLI. *Exploratory Knowledge Discovery over Web of Data*, in "Discrete Applied Mathematics", 2018, vol. 249, pp. 2-17 [*DOI :* 10.1016/J.DAM.2018.03.041], https://hal.inria.fr/hal-01673439

[7] M. ANSDELL, Y. IOANNOU, H. OSBORN, M. SASDELLI, J. SMITH, D. CALDWELL, J. JENKINS, C. RAÏSSI, D. ANGERHAUSEN. *Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning*, in "The Astrophysical journal letters", December 2018, vol. 869, n° 1, L7 p. [*DOI :* 10.3847/2041-8213/AAF23B], https://hal.inria.fr/hal-01957950

[8] J. BAIXERIES, V. CODOCEDO, M. KAYTOUE, A. NAPOLI. *Characterizing Approximate-Matching Dependencies in Formal Concept Analysis with Pattern Structures*, in "Discrete Applied Mathematics", 2018, vol. 249, pp. 18-27 [*DOI :* 10.1016/J.DAM.2018.03.073], https://hal.inria.fr/hal-01673441

[9] G. BOSC, J.-F. BOULICAUT, C. RAÏSSI, M. KAYTOUE. *Anytime Discovery of a Diverse Set of Patterns with Monte Carlo Tree Search*, in "Data Mining and Knowledge Discovery", 2018, vol. 32, n° 3, pp. 604-650 [*DOI :* 10.1007/S10618-017-0547-5], https://hal.archives-ouvertes.fr/hal-01662857

[10] Q. BRABANT, M. COUCEIRO. *k-maxitive Sugeno integrals as aggregation models for ordinal preferences*, in "Fuzzy Sets and Systems", 2018, vol. 343, pp. 65-75 [*DOI :* 10.1016/J.FSS.2017.06.005], https://hal.archives-ouvertes.fr/hal-01657107

[11] Q. BRABANT, M. COUCEIRO, J. R. FIGUEIRA. *Interpolation by lattice polynomial functions: a polynomial time algorithm*, in "Fuzzy Sets and Systems", 2018 [*DOI :* 10.1016/J.FSS.2018.12.009], https://hal.archives-ouvertes.fr/hal-01958903

[12] M. COUCEIRO, J. DEVILLET, J.-L. MARICHAL. *Characterizations of idempotent discrete uninorms*, in "Fuzzy Sets and Systems", 2018, vol. 334, n° 60-72, https://arxiv.org/abs/1701.07253v1 [*DOI :* 10.1016/J.FSS.2017.06.013], https://hal.inria.fr/hal-01447513

[13] M. COUCEIRO, J. DEVILLET, J.-L. MARICHAL. *Quasitrivial semigroups: Characterizations and enumerations*, in "Semigroup Forum", 2018, 22 p. [*DOI :* 10.1007/S00233-018-9928-3], https://hal.inria.fr/hal-01826868

[14] M. COUCEIRO, L. HADDAD, K. SCHÖLZEL. *On the lower part of the lattice of partial clones*, in "Journal of Multiple-Valued Logic and Soft Computing", 2018, https://hal.inria.fr/hal-01826870

[15] M. COUCEIRO, E. LEHTONEN. *Majors of functions*, in "Order", 2018, vol. 35, n° 2, pp. 233-246, https://hal.inria.fr/hal-01519377

[16] M. COUCEIRO, M. MARÓTI, T. WALDHAUSER, L. ZADORI. *Computing version spaces in the qualitative approach to multicriteria decision aid* , in "International Journal of Foundations of Computer Science", 2018, https://hal.inria.fr/hal-01404590

[17] M. COUCEIRO, B. TEHEUX. *Pivotal decomposition schemes inducing clones of operations*, in "Beiträge zur Algebra und Geometrie / Contributions to Algebra and Geometry", 2018, vol. 59, n° 1, pp. 25-40, https://hal.inria.fr/hal-01450835

[18] A. COULET, N. H. SHAH, M. WACK, M. CHAWKI, N. JAY, M. DUMONTIER. *Predicting the need for a reduced drug dose, at first prescription*, in "Scientific Reports", October 2018, vol. 8, n° 1 [*DOI :* 10.1038/s41598-018-33980-0], https://hal.inria.fr/hal-01901566

[19] E. GALBRUN, P. MIETTINEN. *Mining redescriptions with Siren*, in "ACM Transactions on Knowledge Discovery from Data (TKDD)", 2018, vol. 12, n° 1, pp. 6:1–6:30 [*DOI :* 10.1145/3007212], https://hal.archives-ouvertes.fr/hal-01399213

[20] E. GALBRUN, H. TANG, M. FORTELIUS, I. ŽLIOBAITĖ. *Computational biomes: The ecometrics of large mammal teeth*, in "Palaeontologia Electronica", 2018, vol. 21, n° 21.1.3A, pp. 1-31 [*DOI :* 10.26879/786], https://hal.archives-ouvertes.fr/hal-01726076

[21] J. KALOFOLIAS, E. GALBRUN, P. MIETTINEN. *From sets of good redescriptions to good sets of redescriptions*, in "Knowledge and Information Systems (KAIS)", 2018, pp. 1-34 [*DOI :* 10.1007/s10115-017-1149-7], https://hal.archives-ouvertes.fr/hal-01726071

[22] J.-F. MARI, A. GOBILLOT, M. BENOÎT. *Time Space Simulation of Land Use changes by stochastic modeling*, in "Revue Internationale de Géomatique", August 2018, vol. 28, n° 2, pp. 219 - 242, https://hal.inria.fr/hal-01662140

International Conferences with Proceedings

[23] L. AMARÙ, E. TESTA, M. COUCEIRO, O. ZOGRAFOS, G. DE MICHELI, M. SOEKEN. *Majority logic synthesis*, in "ICCAD 2018 - IEEE/ACM International Conference on Computer-Aided Design", San Diego, United States, November 2018 [*DOI :* 10.1145/3240765.3267501], https://hal.inria.fr/hal-01925946

[24] Q. BRABANT, M. COUCEIRO, D. DUBOIS, H. PRADE, A. RICO. *Extracting Decision Rules from Qualitative Data via Sugeno Utility Functionals*, in "International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2018)", Cadiz, France, Communications in Computer and Information Science book series (CCIS), Springer, Cham, June 2018, vol. 853, pp. 253-265, https://hal.inria.fr/hal-01670924

[25] V. CODOCEDO, J. BAIXERIES, M. KAYTOUE, A. NAPOLI. *Characterizing Covers of Functional Dependencies using FCA*, in "CLA 2018 - The 14th International Conference on Concept Lattices and Their Applications", Olomouc, Czech Republic, D. I. IGNATOV, L. NOURINE (editors), CEUR-WS, June 2018, pp. 279-290, https://hal.archives-ouvertes.fr/hal-01856516

[26] B. CONAN-GUEZ, A. GÉLY, L. BOUDJELOUD, A. BLANSCHÉ. *K-spectral centroid: extension and optimizations*, in "ESANN 2018 - 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning", Bruges, Belgium, April 2018, pp. 603-608, https://hal.archives-ouvertes.fr/hal-01901251

[27] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Unsupervised extremely randomized trees*, in "PAKDD 2018 - The 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining", Melbourne, Australia, May 2018, https://hal.inria.fr/hal-01667317

[28] M.-D. DEVIGNES, Y. FRANSOT, Y. LEPAGE, J. LIEBER, E. NAUER, M. SMAÏL-TABBONE. *First steps toward finding relevant pathology-gene pairs using analogy*, in "EvoCBR 2018 : Workshop on Evolutionary Computation and CBR at the International Conference on Case-Based Reasoning (ICCBR 2018)", Stockholm, Sweden, July 2018, https://hal.inria.fr/hal-01906547

[29] T. GILLARD, J. LIEBER, E. NAUER. *Improving Adaptation Knowledge Discovery by Exploiting Negative Cases: First Experiment in a Boolean Setting*, in "ICCBR 2018 - 26th International Conference on Case-Based Reasoning", Stockholm, Sweden, July 2018, https://hal.inria.fr/hal-01905077

[30] A. GÉLY, M. COUCEIRO, A. NAPOLI. *Steps Towards Achieving Distributivity in Formal Concept Analysis*, in "CLA 2018 - The 14th International Conference on Concept Lattices and Their Applications", Olomouc, Czech Republic, June 2018, 291 p. , https://hal.inria.fr/hal-01889163

[31] N. JUNIARTA, V. CODOCEDO, M. COUCEIRO, A. NAPOLI. *Biclustering Based on FCA and Partition Pattern Structures for Recommendation Systems*, in "FCA4AI@IJCAI2018 - 6th International Workshop "What can FCA do for Artificial Intelligence?"", Stockholm, Sweden, July 2018, https://hal.inria.fr/hal-01858409

[32] N. JUNIARTA, M. COUCEIRO, A. NAPOLI, C. RAÏSSI. *Sequence Mining within Formal Concept Analysis for Analyzing Visitor Trajectories*, in "SMAP 2018 - 13th International Workshop on Semantic and Social Media Adaptation and Personalization", Zaragoza, Spain, September 2018, https://hal.inria.fr/hal-01887927

[33] N. JUNIARTA, M. COUCEIRO, A. NAPOLI, C. RAÏSSI. *Sequential Pattern Mining using FCA and Pattern Structures for Analyzing Visitor Trajectories in a Museum*, in "CLA 2018 - The 14th International Conference on Concept Lattices and Their Applications", Olomouc, Czech Republic, June 2018, https://hal.inria.fr/hal-01887914

[34] J. LEGRAND, Y. TOUSSAINT, C. RAÏSSI, A. COULET. *Syntax-based Transfer Learning for the Task of Biomedical Relation Extraction*, in "LOUHI 2018 - The Ninth International Workshop on Health Text Mining and Information Analysis", Brussels, Belgium, Proceedings of LOUHI 2018: The Ninth International Workshop on Health Text Mining and Information Analysis, October 2018, https://hal.inria.fr/hal-01869071

[35] Y. LEPAGE, J. LIEBER. *Case-Based Translation: First Steps from a Knowledge-Light Approach Based on Analogy to a Knowledge-Intensive One*, in "ICCBR 2018 - 26th International Conference on Case-Based Reasoning", Stockholm, Sweden, July 2018, https://hal.inria.fr/hal-01906528

[36] J. LIEBER, E. NAUER, H. PRADE, G. RICHARD. *Making the Best of Cases by Approximation, Interpolation and Extrapolation*, in "ICCBR 2018 - 26th International Conference on Case-Based Reasoning", Stockholm, Sweden, July 2018, https://hal.inria.fr/hal-01905058

[37] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *A First Study on What MDL Can Do for FCA*, in "CLA 2018 - The 14th International Conference on Concept Lattices and Their Applications", Olomouc, Czech Republic, D. I. IGNATOV, L. NOURINE (editors), June 2018, https://hal.archives-ouvertes.fr/hal-01888453

[38] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *How to improve itemset assessment using minimum description length principle*, in "RCAI-2018 - Russian Conference on Artificial Intelligence", Moscou, Russia, September 2018, https://hal.archives-ouvertes.fr/hal-01889791

[39] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *MDL for FCA: is there a place for background knowledge?*, in "IJCAI ECAI 2018 - 6th International Workshop "What can FCA do for Artificial Intelligence?"", Stockholm, Sweden, July 2018, https://hal.archives-ouvertes.fr/hal-01888440

[40] P. MONNIN. *Discovering and Comparing Relational Knowledge, the Example of Pharmacogenomics*, in "EKAW 2018 - 21st International Conference on Knowledge Engineering and Knowledge Management", Nancy, France, November 2018, https://hal.inria.fr/hal-01955424

[41] P. MONNIN, A. NAPOLI, A. COULET. *Combining Concept Annotation and Pattern Structures for Guiding Ontology Mapping*, in "FCA4AI@IJCAI2018 - 6th International Workshop "What can FCA do for Artificial Intelligence?"", Stockholm, Sweden, S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors), Proceedings of the 6th International Workshop "What can FCA do for Artificial Intelligence"? co-located with International Joint Conference on Artificial Intelligence and European Conference on Artificial Intelligence (IJCAI/ECAI 2018), Stockholm, Sweden, July 13, 2018, July 2018, vol. CEUR Workshop Proceedings, n^{O} 2149, https://hal.inria.fr/hal-01858391

[42] F. PENNERATH. *An Efficient Algorithm for Computing Entropic Measures of Feature Subsets*, in "ECML-PKDD 2018 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Dublin, Ireland, September 2018, https://hal-centralesupelec.archives-ouvertes.fr/hal-01897734

[43] G. PERSONENI, M.-D. DEVIGNES, M. SMAÏL-TABBONE, P. JONVEAUX, C. BONNET, A. COULET. *Cooperation of bio-ontologies for the classification of genetic intellectual disabilities : a diseasome approach*, in "Proceedings of the 11th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS 2018)", Antwerp, Belgium, December 2018, https://hal.inria.fr/hal-01925471

[44] J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Three Approaches for Mining Definitions from Relational Data in the Web of Data*, in "FCA4AI@IJCAI2018 - 6th International Workshop "What can FCA do for Artificial Intelligence"?", Stockholm, Sweden, July 2018, https://hal.inria.fr/hal-01887838

National Conferences with Proceedings

[45] A. GÉLY, M. COUCEIRO, Y. NAMIR, A. NAPOLI. *Contribution à l'étude de la distributivité d'un treillis de concepts*, in "EGC 2018 - Extraction et Gestion des Connaissances", Paris, France, Extraction et Gestion des Connaissances, Editions RNTI, January 2018, vol. RNTI-E-34, 478 p. , https://hal.inria.fr/hal-01889149

[46] N. JUNIARTA, V. CODOCEDO, M. COUCEIRO, A. NAPOLI. *Biclustering Based on FCA and Partition Pattern Structures for Recommendation Systems*, in "SFC 2018 - XXVèmes Rencontres de la Société Francophone de Classification", Paris, France, September 2018, https://hal.inria.fr/hal-01889309

[47] J. LIEBER, E. NAUER, H. PRADE, G. RICHARD. *Tirer parti au mieux des cas sources en raisonnement à partir de cas : approximation, interpolation et extrapolation*, in "JIAF 2018 - 12èmes Journées d'Intelligence Artificielle Fondamentale", Amiens, France, June 2018, pp. 1-9, https://hal.inria.fr/hal-01906519

[48] J. REYNAUD, E. GALBRUN, M. ALAM, Y. TOUSSAINT, A. NAPOLI. *Définir les catégories de DBpedia avec des règles d'associations et des redescriptions*, in "EGC 2018 - Extraction et Gestion des Connaissances", Paris, France, January 2018, https://hal.inria.fr/hal-01887801

[49] J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Trois approches pour classifier les données du web des données*, in "CNIA/RJCIA 2018 - Conférence Nationale d'Intelligence Artificielle et Rencontres des Jeunes Chercheurs en Intelligence Artificielle", Nancy, France, July 2018, https://hal.inria.fr/hal-01887820

[50] J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Trois approches pour classifier les données du web des données*, in "SFC 2018 - XXVèmes Rencontres de la Société Francophone de Classification", Paris, France, September 2018, https://hal.inria.fr/hal-01887884

Conferences without Proceedings

[51] Q. BRABANT, M. COUCEIRO, D. DUBOIS, H. PRADE, A. RICO. *Sugeno Integral for Rule-Based Ordinal Classification*, in "IJCAI-ECAI 2018 - Workshop on Learning and Reasoning: Principles and Applications to Everyday Spatial and Temporal Knowledge", Stockholm, Sweden, July 2018, https://hal.archives-ouvertes.fr/hal-01889785

[52] N. JUNIARTA, V. CODOCEDO, M. COUCEIRO, A. NAPOLI. *Biclustering Based on FCA and Partition Pattern Structures for Recommendation Systems*, in "NFMCP 2018 - 7th International Workshop on New Frontiers in Mining Complex Patterns", Dublin, Ireland, September 2018, https://hal.inria.fr/hal-01889384

[53] G. PERSONENI, E. BRESSO, M.-D. DEVIGNES, M. DUMONTIER, M. SMAÏL-TABBONE, A. COULET. *Découverte d'associations entre Evénements Indésirables Médicamenteux par les structures de patrons et les ontologies*, in "Journée I.A. et Santé", Nancy, France, July 2018, https://hal.inria.fr/hal-01872312

[54] M. SCHNELL, S. COUFFIGNAL, J. LIEBER, S. SALEH, N. JAY. *Interpretation of Best Medical Coding Practices by Case-Based Reasoning - A User Assistance Prototype for Data Collection for Cancer Registries*, in "JWAIH 2018 - Joint Workshop on Artificial Intelligence in Health", Stockholm, Sweden, July 2018, https://hal.archives-ouvertes.fr/hal-01907093

[55] M. SCHNELL, S. COUFFIGNAL, J. LIEBER, S. SALEH, N. JAY. *Interprétation de bonnes pratiques de codification médicale par du raisonnement à partir de cas - Application à la saisie de données pour les registres du cancer*, in "Journée I.A. et Santé", Nancy, France, July 2018, https://hal.archives-ouvertes.fr/hal-01907088

Scientific Books (or Scientific Book chapters)

[56] M. COUCEIRO, D. DUBOIS, H. FARGIER, M. GRABISCH, H. PRADE, A. RICO. *New directions in ordinal evaluation: Sugeno integrals and beyond*, in "New Perspectives in Multiple Criteria Decision Making", M. DOUMPOS, J. FIGUEIRA, S. GRECO, C. ZOPOUNIDIS (editors), Springer, 2018, https://hal.inria.fr/hal-01941776

Books or Proceedings Editing

[57] C. FARON ZUCKER, C. GHIDINI, A. NAPOLI, Y. TOUSSAINT (editors). *Knowledge Engineering and Knowledge Management*, Lecture Notes in Computer Science, Springer, Nancy, France, 2018, vol. 11313 [*DOI :* 10.1007/978-3-030-03667-6], https://hal.inria.fr/hal-01948604

[58] S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors). *Workshop Notes of the Sixth International Workshop "What can FCA do for Artificial Intelligence?"*, CEUR Proceedings, Stockholm, Sweden, 2018, vol. 2149, 150 p. , https://hal.inria.fr/hal-01956367

[59] W. M. VAN DER AALST, D. I. IGNATOV, A. V. SAVCHENKO, S. WASSERMAN, M. KHACHAY, S. O. KUZNETSOV, V. LEMPITSKY, I. A. LOMAZOVA, N. LOUKACHEVITCH, A. NAPOLI, A. PANCHENKO, P. M. PARDALOS (editors). *Analysis of Images, Social Networks and Texts*, Lecture Notes in Computer Science, Springer, Moscow, Russia, 2018, vol. 10716, 412 p. [*DOI :* 10.1007/978-3-319-73013-4], https://hal.inria.fr/hal-01962199

Scientific Popularization

[60] F. LE BER, J. LIEBER, M. BENOÎT. *Case-based Reasoning for Forecasting the Allocation of Perennial Biomass Crops*, in "ERCIM News", April 2018, vol. 113, 2 p. , https://hal.archives-ouvertes.fr/hal-01773571

Other Publications

[61] Q. BRABANT, M. COUCEIRO. *Sugeno Utility Functionals for Monotonic Classication & Decision Rules*, July 2018, ISWS 2018 - International Semantic Web Research Summer School, Poster, https://hal.archives-ouvertes.fr/hal-01906052

[62] M. COUCEIRO, P. MERCURIALI, R. PÉCHOUX, A. SAFFIDINE. *On the complexity of minimizing median normal forms of monotone Boolean functions and lattice polynomials*, 2018, working paper or preprint, https://hal.inria.fr/hal-01905491

[63] N. JUNIARTA, M. COUCEIRO, A. NAPOLI, C. RAÏSSI. *Application of Biclustering to the Discovery of Constant and Gradual Patterns*, November 2018, APIL 2018 - Annual PhD students conference IAEM Lorraine, Poster, https://hal.inria.fr/hal-01935849

[64] N. JUNIARTA, M. COUCEIRO, A. NAPOLI, C. RAÏSSI. *Sequential pattern mining for analyzing visitor trajectories*, July 2018, ISWS 2018 - International Semantic Web Research Summer School 2018, Poster, https://hal.inria.fr/hal-01890429

[65] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *What MDL can bring to Pattern Mining*, July 2018, ISWS 2018 - International Semantic Web Research Summer School, Poster, https://hal.archives-ouvertes.fr/hal-01889792

[66] P. MONNIN, A. NAPOLI, A. COULET. *Data-Interlinking: the Seed of Knowledge Reconciliation in Pharmacogenomics*, 2018, working paper or preprint, https://hal.inria.fr/hal-01955262

References in notes

[67] C. C. AGGARWAL, C. ZHAI (editors). *Mining Text Data*, Springer, 2012

[68] F. BAADER, D. CALVANESE, D. MCGUINNESS, D. NARDI, P. PATEL-SCHNEIDER (editors). *The Description Logic Handbook*, Cambridge University Press, Cambridge, UK, 2003

[69] M. ALAM, A. BUZMAKOV, V. CODOCEDO, A. NAPOLI. *Mining Definitions from RDF Annotations Using Formal Concept Analysis*, in "International Joint Conference in Artificial Intelligence", Buenos Aires, Argentina, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, July 2015, https://hal.archives-ouvertes.fr/hal-01186204

[70] M. ALAM, T. N. N. LE, A. NAPOLI. *LatViz: A New Practical Tool for Performing Interactive Exploration over Concept Lattices*, in "CLA 2016 - Thirteenth International Conference on Concept Lattices and Their Applications", Moscow, Russia, July 2016, https://hal.inria.fr/hal-01420751

[71] M. BARBUT, B. MONJARDET. *Ordre et classification – Algèbre et combinatoire (2 tomes)*, Hachette, Paris, 1970

[72] S. DA SILVA, F. LE BER, C. LAVIGNE. *Structures de haies dans un paysage agricole : une étude par chemin de Hilbert adaptatif et chaînes de Markov* , in "EGC 2016 – 16èemes Journées Francophones ”Extraction et Gestion des Connaissances”", Reims, France, Revue des Nouvelles Technologies de l’Information, January 2016, vol. RNTI-E-30, pp. 279–290, https://hal.archives-ouvertes.fr/hal-01266344

[73] B. GANTER, S. O. KUZNETSOV. *Pattern Structures and Their Projections*, in "Proceedings of ICCS 2001", LNCS 2120, Springer, 2001, pp. 129–142

[74] B. GANTER, R. WILLE. *Formal Concept Analysis*, Springer, Berlin, 1999

[75] P. GEURTS, D. ERNST, L. WEHENKEL. *Extremely Randomized Trees*, in "Machine Learning", 2006, vol. 63, n^{o} 1, pp. 3–42

[76] M. GRABISCH, J.-L. MARICHAL, R. MESIAR, E. PAP. *Aggregation Functions*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2009

[77] D. GRISSA, B. COMTE, E. PUJOS-GUILLOT, A. NAPOLI. *A Hybrid Knowledge Discovery Approach for Mining Predictive Biomarkers in Metabolomic Data*, in "ECML PKDD", Riva del garda, Italy, September 2016, pp. 572 - 587 [*DOI :* 10.1007/978-3-319-46128-1_36], https://hal.archives-ouvertes.fr/hal-01421011

[78] O. HUDRY, B. MONJARDET. *Consensus Theories. An oriented survey*, in "Mathématiques et Sciences Humaines", 2010, vol. 190, n^{o} 2, pp. 139–167

[79] M. KAYTOUE, V. CODOCEDO, A. BUZMAKOV, J. BAIXERIES, S. O. KUZNETSOV, A. NAPOLI. *Pattern Structures and Concept Lattices for Data Mining and Knowledge Processing*, in "Machine Learning and Knowledge Discovery in Databases", Porto, Portugal, A. BIFET, M. MAY, B. ZADROZNY, R. GAVALDA, D. PEDRESCHI, F. BONCHI, J. CARDOSO, M. SPILIOPOULOU (editors), Lecture Notes in Computer Science, Springer International Publishing, 2015, vol. 9286, pp. 227-231 [*DOI :* 10.1007/978-3-319-23461-8_19], https://hal.archives-ouvertes.fr/hal-01188637

[80] J.-F. MARI, F. LE BER, E.-G. LAZRAK, M. BENOÎT, C. ENG, A. THIBESSARD, P. LEBLOND. *Using Markov Models to Mine Temporal and Spatial Data*, in "New Fundamental Technologies in Data Mining", K. FUNATSU, K. HASEGAWA (editors), Intech, 2011, pp. 561–584, http://hal.inria.fr/inria-00566801/en

[81] J.-P. METIVIER, A. LEPAILLEUR, A. BUZMAKOV, G. POEZEVARA, B. CRÉMILLEUX, S. O. KUZNETSOV, J. LE GOFF, A. NAPOLI, R. BUREAU, B. CUISSART. *Discovering structural alerts for mutagenicity using*

stable emerging molecular patterns, in "Journal of Chemical Information and Modeling", 2015, vol. 55, n^{o} 5, pp. 925–940 [*DOI :* 10.1021/CI500611V], https://hal.archives-ouvertes.fr/hal-01186716

[82] N. RAMAKRISHNAN, D. KUMAR, B. MISHRA, M. POTTS, R. F. HELM. *Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions*, in "Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", New York, NY, USA, KDD '04, ACM, 2004, pp. 266–275

[83] J. REYNAUD, M. ALAM, Y. TOUSSAINT, A. NAPOLI. *A Proposal for Classifying the Content of the Web of Data Based on FCA and Pattern Structures*, in "International Symposium on Methodologies for Intelligent Systems", Warsaw, Poland, June 2017, https://hal.inria.fr/hal-01667437

[84] M. ROUANE-HACENE, M. HUCHARD, A. NAPOLI, P. VALTCHEV. *Relational Concept Analysis: Mining Concept Lattices From Multi-Relational Data*, in "Annals of Mathematics and Artificial Intelligence", January 2013, vol. 67, n^{o} 1, pp. 81-108 [*DOI :* 10.1007/S10472-012-9329-3], http://hal.inria.fr/lirmm-00816300

[85] T. SHI, S. HORVATH. *Unsupervised Learning With Random Forest Predictors*, in "Journal of Computational and Graphical Statistics", 2006, vol. 15, n^{o} 1, pp. 118-138

[86] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN, A. BOC, V. MAKARENKOV. *A fast compound algorithm for mining generators, closed itemsets, and computing links between equivalence classes*, in "Annals of Mathematics and Artificial Intelligence", 2014, vol. 70, pp. 81 - 105 [*DOI :* 10.1007/S10472-013-9372-8], https://hal.inria.fr/hal-01101140