Departement of Research Evaluation

RESEARCH UNIT SELF-ASSESSMENT DOCUMENT

—

2023-2024 EVALUATION CAMPAIGN
GROUP D

Team Data Science
LAMSADE, PSL Université & CNRS

Contents

1 GENERAL INFORMATION FOR THE CURRENT CONTRACT: DATA SCIENCE TEAM

Collecting and analysing information related to activities of human and organisations is vital for effectively supporting and modeling decision making processes. While nowadays most of the data can be stored, their efficient access and analysis become the main bottlenecks with the following challenges: how to efficiently query data lakes to retrieve information in order extract knowledge; and how to design models at scale to support decision makers. On the one hand, there is a strong research interest on data collection and analysis. On the other hand, the recent promises of machine learning have introduced the core issue of how to make them actionable, acceptable and meaningful.

In this context, the Data Science team is built on 4 projects to gather a wide spectrum of skills and to foster scientific interactions:

- the Management, Analysis and Exploration of Big Data (MADAX) project objective is to propose, study and experimentally analyse techniques of management and analysis of massive data, with a particular focus on web and graph data;
- the Web Services Discovery, Composition and Analysis (WISE) project objective is to propose models and algorithms for discovering and composing reliable services as well as analysing logs produced by service-based systems;
- the Policy Analytics project focuses on decision support in the design, implementation and evaluation of public policies;
- the Machine Intelligence and Learning Systems (MILES) project addresses the challenge of trustworthy machine learning.

The Data Science team is composed of 17 permanent members (1 DR, 6 PU, 8 MCF, 1 PAST). The leader of the team was Dario Colazzo (MADAX) between 2017 and 2020. Since 09/2020, the executive team is composed of Alexandre Allauzen (MILES) and Joyce El Haddad (WISE).

During the period, the organisation and the scientific animation of the team has evolved with two initiatives. First, we organize regular meetings, three per semester, to favors discussions at the team level and share scientific presentations. Then, an annual day workshop of the team was set up to foster scientific interactions and build new shared research topics. Given the sanitary conditions, only one edition was unfortunately organised last year, but we plan from now to have it on a yearly basis.

1.1 Scientific subjects and their implications

The research conducted in the Data Science team is about collecting, managing, and analysing large volumes of heterogeneous, highly dynamic, and distributed data for decision support. Indeed, for decision-makers, extracting, synthesising, and modelling information are very important challenges to effectively support their decision-making processes. In this context, the research activities of the Data Science team focus on identifying, formalising, and designing algorithms and models for decision support.

On a national and international level, the uniqueness of our Data Science team lies in its broad spectrum of research in data science. Our contributions ranged from large (semi-)structured data collecting and management to the development of models and algorithms for decision support, all from the perspectives of trustworthiness, ethics, regulation, and responsibility. Moreover, the research in the Data Science team cover a wide variety of real life applications such as Web services, game playing, art and humanities.

These research activities are carried out in four projects including two cross-disciplinary projects in collaboration with the Decision Aiding team namely Policy Analytics project and Machine Intelligence and Learning Systems project. Moreover, members of Data Science team participate/have some interactions to/with other transversal scientific projects of the LAMSADE such as AGAPE, and Social Choice and Game Theory. Several team members are involved in multiple projects.

The scientific achievements of the Data Science team over the reference period are presented below, grouped by project:

Massive Data Management, Analysis and Exploration (MADAX).

The topics of the project are related to data provenance and reproducibility, analysis and evolutionary learning of heterogeneous massive data, and structure inference and schema-driven data generation.

Scientific workflow: data provenance, anonymisation, transparency, and reproducibility. Scientific workflows are now an established mature technology in a number of modern sciences, including life sciences, biodiversity and astronomy. They are used as a means to specify and implement *in silico* experiments and data analysis. In this context, we have examined a number of issues that arise when using workflows, namely the collection and querying of provenance traces gathered as a result of workflow execution [1, 2, 3], their annotation [4, 5, 6], summarization and exploitation for reporting [7]. We also studied the problem of anonymizing provenance traces that contain information about individuals in order to limit their sharing and exploitation by third-party scientists and communities [8, 9, 10]. Furthermore, we have explored ways to curate reusable data products from workflow provenance traces as well as the problem of reproducibility of experiments and data analysis [11, 12, 13].

Analysis of heterogeneous and massive data. As for the explosion of the enormous amount of data generated to be analyzed by various applications, it has led to the existence of a vast landscape of architectural solutions. Non-expert users who have to decide which analytical solutions are the most appropriate for their particular constraints and specific requirements in a Big Data context are today at a loss, faced with a panoply of disparate and diverse solutions. In this context, in [14] we proposed a generic architecture to classify Big Data analytic approaches, focusing on the data storage layer, the parallel programming model, the type of database and query language that can be used; and a set of comparison/evaluation criteria, such as OLAP support, scalability and fault tolerance support. In [15] on the one hand, we extend this generic architecture to consider a broader typology of Big Data approaches and tools, and enhance the evaluation criteria by including other important aspects for analytical processing, such as Machine Learning support. Moreover, we classified different existing Big Data analytics solutions according to our generic architecture and evaluated them qualitatively based on the comparison criteria. Finally, we proposed a preliminary design of a decision support system oriented to generate suggestions to non-expert users based on such classification and qualitative evaluation.

Evolutionary learning from massive data. Machine learning applications are diversifying and the amount of available data to process is constantly increasing, which raises the difficulty for ML algorithms. Evolutionary Algorithms (EAs), as powerful meta-heuristics, can provide effective tools in this context [16]. We focus on the application of EAs for solving supervised and unsupervised learning problems and their scaling to massive databases. Several scaling solutions have been proposed based on Active Sampling [17, 18, 19, 20] and horizontal parallelization with Apache Spark [21, 22]. Otherwise, EAs, as a universal machine learning technique, were chosen to solve an important problem in biology related to the detection of gene communities in PPI networks, corresponding to sets of proteins/genes collaborating in the same cellular function. The objective is to combine two levels of information: the semantic level which represents the information stored in biological ontologies, and the functional level which represents the information stored in public databases. The contribution of EAs in this context is twofold. First, the communities are built in an iterative and evolutive way, which allows a better exploitation of the data [23, 24, 25, 26, 27]. Secondly, EAs are able to take into account a large volume of data with sampling techniques, thus able to exploit very marge PPI networks, such as that of homo-sapiens.

Structure Inference and schema-driven data generation. Concerning semi-structured data, one of the main topics of the project, they play a crucial role today in many applications context, going from traditional data analytic to machine-learning pipelines and stream processing. Our attention has focused on XML, JSON and RDF data and in this context our results are about techniques for: i) streaming saturation of RDF data [28, 29], efficient techniques for membership checking for regular expressions with counting and interleaving [30], efficient, scalable and interactive techniques for schema inference from massive JSON datasets [31, 32, 33, 34], iii) sound and complete techniques for checking non-emptiness for JSON Schemas [35, 36, 37]. The expertise we have gained in these research fields have been recognized at highest international level, as testified by the acceptance of tutorials at EDBT and SIGMOD [38, 39].

Web Services Discovery, Composition and Analysis (WISE).

The management of ultra large number of services in the global Internet creates many open problems from discovering to composing services and offering a reliable service execution. The aim of the project is to tackle each of these issues as well as the discovery and analysis of contextual behavioural patterns from event logs produced by service-based applications.

Discovery, Composition and Execution. Discover services consists in identifying the services that are able to accomplish a given goal. Composing them correspond to grouping or aggregating existing services to create a new composite one. A service execution is reliable when it is fault-tolerant or self-healing. To tackle each of these challenges, in [40], we proposed a service description enrichment approach based on I/O relations for improving Web service discovery. While the problem of process similarity search has been extensively studied in the literature, existing techniques do not use semantic annotations that are proposed by the semantic web services languages and models. In [41], we proposed a framework for semantic annotated process models discovery, using indexing and matching techniques taking into account semantic annotations. In order to facilitate workflow/process reuse [42], we addressed also the issue of finding workflows, or fragments thereof, that are of interest to the user and proposed a keyword-based search of workflow fragments and composition operations to combine them [43]. In [44], we focused on QoS service composition, and we analysed and compared the theoretical complexity induced by each QoS criterion, stating that optimal solution for execution time or throughput QoS criteria can be determined in polynomial time but optimality is no more guaranteed in polynomial time for QoS criteria like cost or reliability. This aforementioned project is transversal between “Combinatorial optimization, algorithms” and “Data Sciences” research areas. In [45], we propose an architecture based on a multi-agent system exhibiting a self-adaptive behavior to address the dynamic resource allocation in Cloud computing. This self-adaptive system follows a MAPE-K approach to reason and act, according to QoS, service information, and propagated run-time information, to detect QoS degradation and make efficient resource allocation decisions. In [46], we proposed a trust-based dynamic coalition formation process for service composition in social networks. In [47], we showed how it is possible to relax the *retriable*[1] property in composite service execution, based on our fuzzy

[1] guarantee of a successfully termination after a finite number of invocations

atomicity model, allowing to self-adapt the CS execution, taking into account the state of the CS execution and user preferences (ie, the acceptable fuzzy atomicity expressed in the user requirements). More recently, we have been interested in Microservice architectures. In [48], we applied our research work on polyglot persistence microservices, defining a weak global consistency definition for microservice architectures and presenting a recovery protocol after disaster recovery. In [49], we proposed an approach for regulating microservices' deployment and execution over clouds, edges and Internet-of-things platforms using contracts. We also tackled the problem of enriching local data sources with data provided by data services, given user requirements in terms of queries. In doing so, we have addressed data integration problem by considering two contrasting dimensions, namely the quality of data services and the cost they incur in [50, 51, 52].

Process mining. Other research direction was process mining, analysing event logs produced by transactional or service-based applications. While many approaches exist for structured and processes, the analysis of unstructured logs (text, emails) and flexible processes is still an open problem. Owing to its wide use in personal, but especially professional contexts, e-mail represents a valuable source of information that can be harvested to understand, reorganize and reallocate undocumented business processes in companies and institutions. In this perspective, we have investigated a number of solutions that leverage, among other techniques, unsupervised and supervised learning, entity recognition, process mining for the identification of process topics [53, 54], process instances [55], activities [56, 57], metadata characterizing activities [58]. More recently, we have explored the use of relational learning to investigate the interplay between the identification of activities and the extraction of process instances [59]. Existing algorithms for discovering end-to-end process models, are not appropriate for highly flexible process, as the unstructuredness of the resulting models renders them meaningless. It has therefore been suggested to derive insights about flexible processes by mining behavioral patterns, i.e., models of frequently recurring episodes of a process' behavior. However, existing algorithms to mine such patterns suffer from imprecision and redundancy of the mined patterns and a comparatively high computational effort. We proposed an approach for discovering behavioral patterns that overcomes the limits of existing approaches by discovering patterns that are maximal and compact (avoiding to provide redundant information). Our approach, coined COBPAM (COmbination based Behavioral Pattern Mining) [60] improves theefficiency by evaluating patterns only on parts of the log and by exploiting that complex patterns can be characterized as combinations of simpler patterns. In order to understand flexible process execution, in [61] we aimed to discover factors that affect the process flow. The proposed methodology tackles challenges related to the general correlation problem of process mining, like dealing with general process behavior (not just local decisions) and relaxing the independence assumption among the elements of behavior. In [62] we identified potential of behaviors for improving organizational performance by applying the Method of Reflections and analysing specialization of organisations. Similarly, in software development there is a vast amount of data available in repository managers such as provided artifacts, number of vulnerabilities, and quality information. All these valuable data can be used to improve software development by supporting the maintenance of software ecosystems. In this perspective, we proposed two metrics related to artifacts and releases evolution in time, namely rhythm and speed, to track dynamics of Maven ecosystem [63].

Policy Analytics.

This project is an interdisciplinary approach to decision-making that is a joint project with the Decision Aiding team. Its presentation is done in the self-assessment document of Decision Aiding team.

Machine Intelligence and Learning Systems (MILES).

The MILES (Machine Intelligence and Learning Systems) project, established in early 2016, continues the cross-disciplinary Machine Learning project that was reviewed during the previous HCERES evaluation. The core idea of this project is to bring together researchers from the three teams of LAMSADE (Decision-Making, Optimization, and Data Science) with a shared goal: advancing the theoretical and algorithmic foundations of machine learning with a focus on trustworthiness. While machine learning becomes ubiquitous, these research topics are today crucial.

On a national and international level, the MILES project's identity and uniqueness lie in its ability to bring together researchers specializing in game theory, computational social choice, applied mathematics, computer science and more recently Natural Language Processing. This diversity builds upon an established tradition of LAMSADE and has enabled the project to develop cutting-edge research. By leveraging recent advancements in these fields, our contributions tackle critical issues in modern machine learning, such as robustness against adversarial attacks and fairness in recommender systems. This singular approach places the project in a prominent position within the national and international research community, as evidenced by the high number of CNRS applications received annually.
Since its establishment, the MILES project has seen an exceptional growth due to the continued and robust support from its stakeholders (Université Paris Dauphine - PSL, CNRS). The project has welcomed two university professors, an assistant professor, and a CNRS researcher during the period, adding further depth to its research themes. Currently, the project is comprised of 11 permanent members, including 4 professors, 2 CNRS researchers, 3 assistant professors, 1 associated professor, and 1 research fellow. The following table provides an overview of the project's research topics:

Theoretical foundations	**Algorithmic aspects**	**Applications**
Trustworthy ML: - Robustness to adversarial attacks - Fairness and privacy - Explainability Online learning & game theory Reinforcement learning Theory of Deep Learning	Frugality in deep learning High dimensional data streams Optimization methods Sketching and clustering Structured representations	Health Game playing Robotics Computer Vision Natural Language Proc. Art and Humanities

At the institutional level, the MILES project is highly involved in shaping and executing the strategies of its stakeholders at the local and national levels. Several project members hold 3IA PRAIRIE chairs, and one is a co-director. One member of the project serves as the director of the PSL transverse Data program, which provides AI training to the entire PSL community. This program was awarded 8.8 million euros from the CMA-IA call. The project was instrumental in the creation of the PSL Master IASD (Artificial Intelligence, Systems and Data) and coordinates the Computer Science graduate program. At the University of Paris Dauphine - PSL, the project is involved in coordinating the Dauphine Numérique transverse program and was instrumental in the creation of the double degree in AI and Management. At the national level, the project is active in the PEPR Cybersecurity, Digital Health, and AI initiatives. The MILES project has established a range of national and international academic and industrial partnerships. For example, CIFRE theses are conducted with the Meta FAIR research center, and the project has collaborated with Google Brain NY, among others. The project's exchanges with Riken in Japan are ongoing and expected to intensify. Since 2016, the project's efforts have been recognized with numerous achievements, including:

- The DGA thesis prize for Anne Morvan for her work in sketching
- The AAAI 2022 best paper award for Virginie Do for her work in equity
- 1st place in the BCI Challenge competition at the IEEE WCCI 2020 conference for Florian Yger and his collaborators
- Several prizes for Tristan Cazenave for his work in games

At the scientific level, the project's work has made significant advancements in the field of trustworthy machine learning. To name a few examples:
Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier: Online certification of preference-based fairness for personalized recommender systems. AAAI 2022. Best Paper award.
Geovani Rizk, Albert Thomas, Igor Colin, Rida Laraki, Yann Chevaleyre: Best Arm Identification in Graphical Bilinear Bandits, International Conference on Machine Learning. ICML 2021. Bandit algorithms are traditionally applied to cases where a single agent has to choose actions to take to maximise its payoffs, interacting with a stochastic environment. This paper constitutes the first extension of linear bandit algorithms to situations where several agents interact. The paper also proposes bounds on the number of iterations required to achieve a certain quality in the solution.
Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, Jamal Atif : Randomization matters. How to defend against strong adversarial attacks. ICML 2020 This paper introduces an original theoretical framework combining game theory, statistical learning theory and information theory to derive strong theoretical results on the nature of equilibria between attackers and defenders in machine learning. In particular, it has provided theoretical arguments for defences that rely on noise addition techniques in learning models.
Ikko Yamane, Florian Yger, Jamal Atif, Masashi Sugiyama : Uplift Modeling from Separate Labels. NeurIPS 2018 : 9949-9959 This paper presents a remarkable advance in the context of learning. It introduces a theoretical framework for the study of a realistic instance of the estimation of the effect of a treatment on an individual in the situation where one has observation and control data but with separate labels.

1.2 Consideration of the recommendations in the previous report

Next, we first resume the recommendations concerning the Data Science team made in the previous evaluation report. After, we present the actions taken by the team to implement the recommendations.

–*"risque que certaines personnes fortement sollicitées (avec la création d'une nouvelle offre de formation autour des Data Sciences) s'éloignent de la recherche à cause des charges."*

The charges created by new teaching programs was spread to avoid an overwhelming workload for a reduced number persons. The inclusion in PSL allowed us to better distribute the burden across the involved institutions. For instance the tasks associated with the Computer Science Graduated program of PSL, along with the IASD master program, were shared with ENS.

–*"positionnement du projet Web Services à l'intérieur du pôle restreint significativement la part du champ investigué. Les contours doivent être mieux explicités et le nom du projet légèrement modifié."*

Following this recommendation, the name of the project have been changed from "Web Services" to "Web Services Discovery, Composition and Analysis". Moreover, as described later in the report, the research scope of the field investigated by the members of this project have evolved, either towards the MADAX themes or towards mining processes with the rise of AI.

–*“un grand nombre de thèses (et notamment des thèses CIFRE) vont débuter. Il faut être vigilant quant aux risques de dispersion thématique, compte tenu de la multiplicité des questions et des champs d’application considérés.”*

We still have a high number of ongoing PhD thesis. However, we kept in mind the recommendation of the previous committee and tried to ensure the consistency between the associated PhD topics at the project level.

2 PORTFOLIO INTRODUCTION : DATA SCIENCE TEAM

The portfolio of the Data Science team is composed of 4 papers, at least one per project to demonstrate the quality of our research work, and a serie of *Ex-Machina* posdcasts:

- The paper entitled *Online certification of preference-based fairness for personalized recommender systems* received the Outstanding Paper Award in AAAI 2022. Online platforms are facing an increased public scrutiny and stronger regulations on how recommendations are delivered to different users. This paper introduces a new framework for user-side fairness in recommender systems based on the envy-freeness and proposes a new auditing algorithm which is sample efficient and with theoretical guarantee.

- The interaction between Machine Learning and Physics constitutes a new and important research trend with promising perspectives. As an illustration, the paper entitled *A Dynamical System Perspective for Lipschitz Neural Networks* published at ICML 2022 build upon the analogy between Dynamical Systems and the inference in Deep Neural Networks. The idea is to design a stable architecture for Deep Network by drawing insipiration from continuous dynamical systems and their discretization schemes.

- The paper entitled *Witness generation for json schema* [64] published at VLDB 2022 presents an original sound and complete algorithm for checking the satisfiability of an input schema, generating a witness when the schema is satisfiable. While the problem was proven to be EXPTIME-complete, this work is the first to explicitly describe an algorithm that has the potential to work in reasonable time over schemas of realistic size.

- The paper entitled *Discovering and Analyzing Contextual Behavioral Patterns From Event Logs* [65] published in TKDE 2022 presents an original approach to discover contextual behavior patterns and to analyse causal relations between context information and the patterns. The work has been done in the context of a PhD thesis and in collaboration with Humboldt University of Berlin.

- Four members of the team were involved in the podcast *Ex-Machina* created by *Dauphine Numérique*. This podcast develops a scientific discussion between researchers from Dauphine Université about the impact of AI and algorithm on our lives, the society and the future. After a first season with more than 10 000 listeners, the awarded podcast has started a second season. We participated to the following programs:
  - Ethics and algorithm: How to keep the control on AI systems (Éthique des algorithmes : comment garder la main sur les IA ? janvier 2023);
  - AI and History (L'IA, quelle Histoire! novembre 2022);
  - Art & AI : Robots, ”draw me a sheep” (Art & IA : Robot, dessine-moi un mouton octobre 2022);
  - Algorithms get the power, what is the impact of the democracy ? with a member of the Decision Aiding team (Les algorithmes au pouvoir - Numérique, information et démocratie décembre 2021);
  - AI : the humans workers behind the machines ? (IA : qui sont les humains derrière la machine ? octobre 2021)

3 SELF-ASSESSMENT DOCUMENT : DATA SCIENCE TEAM

Evaluation area 2. Attractiveness

Standard 1. The unit has an attractive scientific reputation and contributes to the construction of the European research area.

Invitations
Members of the Data Science team have international reputation and have be invited to institutions/stays in foreign research centers, as well as to national and international conferences and workshops.

- *Invitations to academic institutions/stays in foreign research centers:*
  During the reference period, members of the Data Science team were invited to stays in the following European and international institutions: Polytechnic University of Turin (2022), Politecnico di Torino (2022), Imperial College (London, 2018), University Metropolitaine de Manchester (2017), DIMACS, Rutgers University (USA, 2022), 3A Institute at Australian National University (2019), IBM Almaden California (2020), University of California (2020), University Mohamed V à Rabat (2021), Lebanese American University (2018, 2019), Faculty of Sciences of Tunis (2018), University Central of Venezuela (2019), and University Simon Bolivar (Venezuela, 2017, 2018).

- *Invitations to international conferences, workshops, doctoral schools,...*
  Several members have been invited to give key talks in international venues: keynote panel at the International Conference on Computational Science and its Applications in 2020, several lectures at MCDM Summer School in 2018, tutorial speaker at the International conference on Knowledge Management, Information and Knowledge in 2017, keynote at the workshop DARLIAP@EDBT in 2022, keynote speaker at French-Brazilian School on Big Data and Smart Cities in 2017.

Scientific Expertise
During the reference period, several members of Data Science team participated in research steering boards and are scientific expertise to European and national evaluation and funding agencies.

- *Expert for European and national evaluation and funding agencies:*
  Chargé de mission ”Science des données et intelligence artificielle” INS2I CNRS (one member since 2016), CNU (two members since 2019), ANR, HCERES, H2020, FNRS (Belgium), FNRS (Luxembourg), Polish Academy of Sciences, and ANRT.

- *Expert for MESRI and the French government:*
  Co-moderator and main writer of the national report GENIAL-Allistene, Animation of the Allistene Cloud AI working group, Member of the board of ”France is AI”

- *Participation in research steering board:*
  Chair of the Steering Committee of the Algorithmic Decision Theory series of conferences, (Université Paul Sabatier, FR, 2021; Duke University, USA, 2019; University of Luxembourg, 2017), member of the GDR MaDICS steering committee (since 2020), coordinator of the IEEE GRSS Database Working Group on Data Management (since 2022), external member of scientific council of INSA Centre Val de Loire (since 2019), steering committees in Hawaii International Conference on System Sciences (since 2015), International Conference on Smart Cities and Green ICT Systems (since 2017), National

EDA Conference (Data Warehouses and Online Analytics, 2011-2017), ISCRAM International Conference on Information Systems for Crisis Response and Management (since 2015), International Journal Transactions on Knowledge and Data Engineering (2017), International Conference Connected Smart Cities (since 2017), President of ”Maître de Conférences” recruitement committee at Université de Caen Basse-Normandie, externel member of ”Maître de Conférences” recruitement committee at Université de Rouen (2022), Université de Belfort-Montbéliard (2021), member of the working group on Knowledge Graph Construction of W3C community, member of the ReProVirtuFlow action of GDR MaDICS (2016-2019), member of a working group on Knowledge Sharing in PSL University.

Program and Organisation committees

Members of the Data Science team enjoy high international visibility as shown by their participation to program committees of prestigious venues and by organising international conferences.

Several members of Data Science team were program committee members (or program chair of a track) of the following international events:

- Track Program (co)-Chair for ”Data and Knowledge Management” in the international conference on Future Internet of Things and Cloud (Prague, 2017), Crisis Management Mini Track à Hawaii International Conference on System Sciences (since 2016), Smart City Mini Track (2014-2018)
- Senior PC member of ”Extraction et Gestion des Connaissances” conference (since 2019)
- Members of program committee in prestigious international and national conferences and workshops since 2017: ACL, NAACL, NEURIPS, ICML, IJCAI, AAAI, ICLR, SSDBM, BPM, ICWS, ICSOC, CoopIS, ICPM, WISE, VLDB, SIGMODn, ICBI, CIKM, EDBT, ODBASE, C&TC, AI-PA@ICSOC, BDA, EGC, EA, ICIKS, AISTAT, SPC IJCAI-ECAI
- Session Chair ”Anomaly and change point detection” at ESANN 2022, and ”Workshop on Riemannian Geometry” at BCI Meeting 2021
- Main chair of the first symposium on ”Big Data Analytics”, organized in Tunis, 2018

Some of events that our team organized during the reference period at Université Paris-Dauphine or abroad:

- Organisation of IEEE GRSS second workshop on Remote Sensing Data Management Technologies in GeoScience (RSDM-GeoSci), Universit Paris-Dauphine, 50 participants, in 2022
  Committee members: Khalid Belhajjame, Maude Manouvrier
- Chair of the organizing committee and organiation of the Informatics of Organizations and Information and Decision Systems (INFORSID) congress, 100 participants, 2019
  Chair : Elsa Negre
- Member of Organisation Committee of the french conference on databases, BDA, Bucarest, 2018
  Commitee member: Daniela Grigori
- Organisation of two editions of a Winter Thematic School on ”Microservices and Big Data” at Université Paris Nanterre Ouest La défense, 2018 and 2019
  Chair : Marta Rukoz
  committee members: Joyce El Haddad, Sonia Guehis-Saadaoui, Maude Manouvrier

- Celebrate Erasmus+ Day and prelaunch of Light-Code Project, octobre 2022
  Chair : Alexis Tsoukias
  Committee members: Michel Zam

Editorial board
Several members of Data Science team served on editorial boards of international journals

- Journal of Multicriteria Decision Analysis provides an international forum for the presentation and discussion of all aspects of research, application and evaluation of multicriteria decision analysis, published by Wiley Online Library
- EURO Journal of Decision Processes publishes scientific knowledge on the theoretical, methodological, behavioural and organizational topics that contribute to the understanding and appropriate use of operational research in supporting different phases of decision making processes, published by Elsevier
- Decision Analysis, a peer-reviewed international journal dedicated to advancing the theory, application, and teaching of all aspects of decision analysis, published by INFORMS
- MethodsX, a multidisciplinary open access peer-reviewed journal published by Elsevier
- Data in Brief, a multidisciplinary open access peer-reviewed journal published by Elsevier (2017-2019)
- Data journal, a peer-reviewed open access journal on data in science published by MDPI (since 2020)
- Big Data and Cognitive Computing, a peer-reviewed open access journal published by MDPI (since 2020)

Prize and distinctions

- Outstanding Paper Award at the prestigious international conference AAAI in 2022 for Jamal Atif and Virginie Do and their co-authors for their work on online certification of preference-based fairness for personalized recommender systems
- Most Reproducible Paper Award at VLDB in 2022 for Dario Colazzo and his co-authors for their work on witness generation for json schema
- Best Paper Award at the national conference BDA in 2020 for Dario Colazzo and his co-authors for their work on witness generation for json schema
- Winner of the best score for motion prediction at the Clinical BCI Challenge (click to see web page) of IEEE WCCI in 2020 for Florian Yger and his team
- Florian Yger holded a ”Tremplin” Chair from PaRis Artificial Intelligence Research InstitutE (Prairie) (click to see web page) (2019-2021)
- Best Scientific Poster Award at Dauphine Digital Days in 2022 to Virginie Do for her work on ”Optimizing generalized Gini indices for fairness in rankings”.

Standard 2. The unit is attractive for the quality of its staff hosting policy.

PhD and post-doctorate
During the reference period, a total of 36 students started their Ph.D. in the Data Science team. 18 of them have defended their thesis. The average length of a thesis is 43.7 months. The funding of the Ph.D. thesis are as follows: 12 CIFRE, 9 funded by different ministry (7 MESRI,

2 other ministries), 10 fundings based on contract with industry, 2 ”co-tutelle” with Tunisia, and 3 auto-funding. All the PhD students follow a training in ethics and scientific integrity via the doctoral school. The data science team received during the reference period 5 post-doctorate fellows Paul Beaujean (2018-2021), Sileno Giovanni (2017-2018), Yamane Ikko (2020-2022), Mehdi Acheli (2021-2022), Emna Beligith (2022-2023), and Muni Pydi, a Junior Fellow (5-year contract) funded by PSL under the CMA-IA project (MESRI).

Senior researchers

The Data Science tema is composed of 17 permanent members (1 Directeur de Recherche CNRS, 6 Professeurs des Universités, 8 Maîtres de Conférences, 1 PAST, and 1 Professeur Emérite). Yann Chevaleyre was transfered (mutation) in 2017. Alexandre Allauzen was recruited as Professeur des Universités in 2019. Pierre Ablin was recruited as a Chargé de Recherche CNRS in November 2021, he did a short stay at the Data Science team before his resignation from CNRS in august 2022. Elsa Negre defended her Habilitation thesis in November 2017 and Khalid Bellahjame defended his Habilitation thesis in October 2022. The head of the team was Dario Colazzo until September 2020, followed by Alexandre Allauzen and Joyce El Haddad in September 2020 until now.

Visiting researchers

During the reference period, the Data Science team hosted 33 researchers from abroad each for a month. In addition, there were other visits that were funded by projects. The invited Professors were: Weidlich Mathias [Humboldt-Universität, Germany]; Sartiani Carlo [Università degli Studi della Basilicata, Italy]; Van Der Aa Johannes [University of Mannheim, Germany]; Giorgio Ghelli [Università di Pisa, Italy]; Stefano Bistarelli [University of Perugia, Italy]; Cesare Pautasso [Lugano, Switzerland]; Pedreschi Dino [Università di Pisa, Italy]; Delias Pavlos [Eastern Macedonia and Thrace Institute of Technology, Greece]; Comes Tina [Delft Technology, Netherlands]; Chackar Salem [University Laval, Canada]; Pino Perez Ramon [Los Andes University, Venezuela]; David Rios Insua [Spanish Royal Academy of Science, Spain]; Mike Papazoglou [University of Tillburg, Netherlands]; Boualem Benatallah [UNSW, Asutralia]; Ventre Carmine [King’s College London, UK]; Yudith Cardinale [Universidad Simón Bolívar, Venezuela].

Standard 3. The unit is attractive because of the recognition gained through its success in competitive calls for projects.

During the reference period, the members of Data Science team are PI of the following funded projects:

- The PEPR (*Programmes et équipements prioritaires de recherche*) on Artificial Intelligence will start around fall 2023. During the evaluation period the team were involved in building 4 projects of this program. The projects are Reinforcement Learning and Digital Health (as PI for both of them), along with Smart (AI and Frugality) and Cybersecurity as partner. These project will have a great impact on the future of the team.

- Project Erasmus+ LightCode, (2022-2025), PI: Alexis Tsoukias, 400K€, partners: University of Macedonia, Univerzitet U NISU, Karmic Software Research, REACH Innovation Consultancy, Universite de Zagreb, SYMPLEXIS. The project aims to strengthening the digital transformation of higher education through low-code.

- CNRS project 80Prime-MITI: PI, 1 PhD with the institute Chemistry Biology Innovation (or CBI, UMR ESPCI-PSL) on the application of deep-learning to Epistasis and biological interaction.

- Mining evOlving sOftwaRe Ecosystems (2022-2024), Appel Unique INS2I CNRS, PI: Joyce El Haddad, 12400 euros. The objective of this project is to propose new solutions, based on machine learning, helping organisations to analyse their software ecosystems and to realise the evolution in their dependencies.

- PHC CEDRE project, Algorithmes efficaces pour la réparation de la qualité dans les réseaux dynamiques(2018-2020), PI: Joyce El Haddad, 16000 euros, partners : LAMSADE (porteur), LIP6, LAU (Liban), Université Haigazian (Liban). The objective of the project was to propose new, efficient algorithms for quality repair in dynamic networks such as cloud services composition or influence propagation in social networks.

The table 3 gathers the ANR funded projects started during the evaluation period along with the amount allocated to the LAMSADE. Moreover, ANR projects allows the team to develop new research topics and to build strong partnership with other labs. The following list provides an overview of these collaborations:

- FlowCon on Turbulent flow closed-loop control with machine learning (PI: LISN Université Paris-Saclay)

- ACDC on Adversarial training for controlled data-to-text generation (PI: ISIR, Sorbonne Université)

- E-SSL on Efficient self-supervised learning for inclusive and innovative speech technologies(PI: LIA, Université d'Avignon)

- STAP on Spatio-temporal analysis of pediatric magnetic resonance images (PI: Telecom ParisTech)

- DELCO on Deep Learning for Combinatorial Optimization (PI, JCJC)

- DeepIntegrOmics on End-to-End Deep learning for Precision Medicine through Metagenomics and cost-sensitive data integration (PI: UMMISCO)

- CARE for Capabilities for risk Acceptability and REsilience (PI/ INERIS),

Name	Start	End	Budget Dauphine en K€
STAP	01/10/2017	31/12/2022	137
DELCO	01/10/2019	30/11/2024	284
SPEED	01/04/2021	31/12/2024	44
CARE	01/05/2021	30/09/2023	47
ACDC	01/10/2021	31/03/2026	155
DeepIntegrOmics	01/10/2021	31/01/2026	107
E-SSL	01/10/2022	31/03/2026	133
PROTEUS	01/10/2022	30/09/2026	159

Table 1: Summary of ANR projects on the evaluation period

Standard 4. The unit is attractive for the quality of its major equipment and technological skills.

The Data Science team is not concerned by this point.

Evaluation area 3. Scientific production

Standard 1. The scientific production of the unit meets quality criteria.

Members of the Data Science team have obtained results published in highly selective conferences and journals which attest the solid theoretical and methodological foundations. Next, we give some publications in journals ranked Q1 and Q2 in Scimago, and in international conferences ranked A* or A on CORE:

- Publications in high impact journals ranked Q1 in Scimago: VLDB Journal, Future Generation Computer Systems, Socio-Economic Planning Sciences, International Journal of Web and Grid Services, Machine Learning, Computers in Industry, Journal of Environmental Management, Discrete Applied Mathematics, European Journal of Operational Research, Autonomous Agents and Multi-Agent Systems, Annals of Operations Research, GigaScience, IEEE Access, IEEE Transactions on Biomedical Engineering, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Neural Systems and Rehabilitation Engineering, International Journal of Approximate Reasoning, Journal of Environmental Management, Journal of Neural Engineering, Sustainable Cities and Society

- Publications in leading conferences ranked A* or A on CORE: AAAI, IJCAI, EDBT, ICML, NeurIPS, AISTATS, ICDM, ECAI, ICALP, GECCO, ECML-PKDD, ICAPS, SIGMOD, IROS, ICSOC, ECCV, PERCOM, PLDI, SIGIR, ICDAR, ICWS, Interspee, FOGA, MICCAI, CaiSE, ICCS.

Moreover, members of the Data Science team have long-term academic collaborations and many publications with international partners. Let us illustrate this with the following examples:

- Researchers: Prof. Dr. Tina Comes (TU Delft), Giorgio Ghelli (Université de Pise, Italy), Carlo Sartiani, (Université de la Basilicata, Italy), Stefanie Scherzinger (Université de Passau, Germany), Matthias Weidlich (Humboldt University of Berlin), Pavlos Delias (International Hellenic University), Dr. Ron Fagin (Almaden, CA, USA), Prof. Darrell Long (UCSC, CA, USA), Prof. Th. Schwarz (Marquette U., Milwaukee, Wisconsin), Prof. S. Jajodia (CSIS, GMU, Virginia).

- Institutions: University of Campinas on Database versions & geospatial, University of Luxembourg, University of Brest, University of Mons (on DecisionDeck), University of macedonia, University of Zagreb, University of Nis, University of Milwaukee on LightCode, ANU Centre for European Studies, ANU School of Cybernetics, ANU Fenner School of Environment and Society, DIMACS at Rutgers University, University of Canberra and EnsadLab, laboratory of the École Nationale Supérieure des Arts Décoratifs – PSL University on Social Responsibility of Algorithms (click to see web page), DIMACS, ANU on Algorithmic Futures Policy Lab (click to see web page), ICMAT (Spain) following several Ph.D. students, RIKEN Center for Advanced Intelligence Project, University Simon Bolivar, University Central (Venezuela) on interoperability and decision making in pervasive information systems, University Tunis El Manar, and University of Manouba.

Standard 2. Scientific production is proportionate to the research potential of the unit and shared out between its personnel.

During the reference period, the total number of publication of the Data Science team in HAL collection is 341 articles as shown in the following table:

journal	conf	book	book chap.	these	report	preprint	Proc & special issues
75	178	3	22	15	6	40	2

The distribution among publication categories as presented in the pie chart of Figure 1 are as follows: 22% of journal articles, 64% of articles published in conference proceedings, and 7% of books and book chapters.

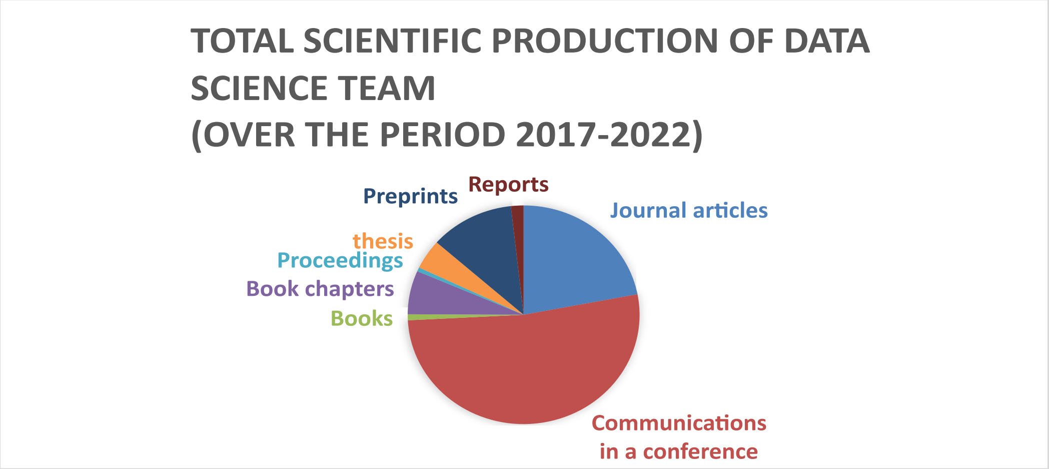

Figure 1: Overall scientific production of Data Science team during the evaluation period (HAL collection LAMSADE-DAUPHINE).

Moreover, as shown in Table 2, members of Data Science team have published 275 articles in international journals or conference proceedings. Publications reflect collaborative works with an average of 4.0 authors per publication. 38, 1% of publications involve Ph.D. students and 18,54% of publications are joint with members of Decision Aiding team or Combinatorial Optimization, Algorithms team.

# publications (journal, conf, book chapter)	average # authors per publications	# papers involving Ph.D. students	# papers involving at least two teams
275	4.0	105	51

Table 2: Some values about Data Science team scientific production over the evaluation period 2017-2022.

Figure 2 shows the production for each year of the reference period for three main categories (journal articles, communications in conferences, and book chapters). For these main categories, the total number of contributions is 275. The average number of publications per year is 45.8.

Standard 3. The scientific production of the unit respects the principles of scientific integrity, ethics and open science. It complies with the applicable guidelines in this field.

Details about our unit policy regarding scientific integrity, ethics and open science are available in the LAMSADE self-assessment document.

Evaluation area 4. Contribution of Research Activities to Society

Standard 1. The unit stands out by the quality and quantity of its non-academic interactions.

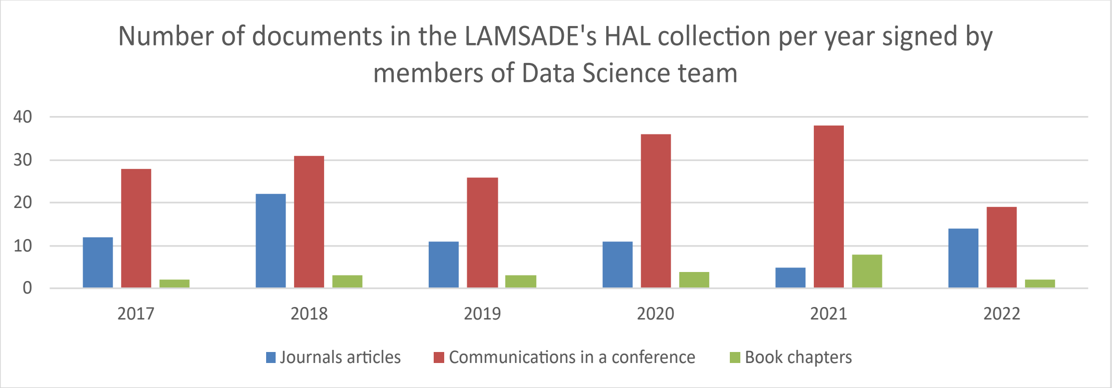

Figure 2: Scientific production of Data Science team during the evaluation period.

Relations and partnership relations with the economic, social and health worlds.

Members of Data Science team have sustained relations with the economic, social and health worlds, as well as long-term partnership relations with large groups, small groups and start-ups.

- Projects with economic and social world: Convention with the Ministry of Sustainable Development (Crisis management unit) on the Decision support in critical situations, Convention with IRSN (Institut de Radioprotection et de Sûreté Nucléaire) on decision support for the management of a major nuclear issue in a sea context, Project *France Relance* of postdoc funding with the Logpickr startup (currently part of iGraphx) (2021-2022), Project on ”Aggregative research using external sources” and ”Reading recommendation” with Credit Agricole CIB (CACIB), Collaboration with Emvista a stratup specialized on NLP.
- Projects with health world: Participation to the ANR Deep Integromics on health and collaboration in this framework with practitioners of the APHP, partnership with CHU Lyon on research evaluation support for medical diagnostic support META-Conseil.
- Research collaborations: with the consulting companies ADWay and Square, Senior consultant (one member) in evolutionary optimization and learning in the R&D team of InstaDeep, a multinational startup company specialized in Artificial Intelligence, Coordination and Participation in the thematic lecture series on emerging topics for the Dauphine Digital Partner Circle.

Platforms developed or shared or used by the external actors.
One member of the Data Science team was involved in the development of Decision Deck and Cloud platform (click for the web page dedicated to Open Source software tools and led by Decision Aiding team.

Societal and technological issues and the impact of the team.
Data Science and Artificial Intelligence have both developed strongly in recent years and have emerged as crucial technologies for public and private organisations. Both play and will continue to play a central role in countless aspects of life affecting healthcare, education, commerce, finance and justice, not to mention everyday life. Data Science members contributes through their research to the study of social responsibility of algorithms, to the development of concepts, theories and methodologies on decision support in the design, implementation and evaluation of public policies. Moreover, through our partnerships, our research contributes to

model interactions between genetic mutations (collaborative research with ESPCI), and to develop walking algorithms for exoskeletons dedicated to the rehabilitation of paraplegic patients (Wandercraft CIFRE thesis), and models to predict the onset of certain diseases based on the metagenomes of patients (ANR Deep Integromics).

Hosting Ph.D. students whose research is funded by non-academic partners.
During the reference period, the Data Science team hosted 12 CIFRE thesis in collaboration with large groups, small groups and start-ups such as META (Facebook), Google Brain, Foxstream, TalentSoft, SAP, Air France, Criteo, Coheris, STmicroelectronics, Wavestone, Huawei, and Wandercraft.

Standard 2. The unit develops products for the cultural, economic and social world.

Some members of Data Science team have results in terms of product development for the cultural and economic (patent) worlds:

Rasta is a system built to automatically identify the artistic style of a painting given its image file (or url). The painting can be well known or home-made and the identification covers 25 different styles: Abstract Art, Abstract Expressionism, Art Informel, Art Nouveau (Modern), Baroque, Color Field Painting, Cubism, Early Renaissance, Expressionism, High Renaissance, Impressionism, Magic Realism, Mannerism (Late Renaissance), Minimalism, Naïve Art (Primitivism), Neoclassicism, Northern Renaissance, Pop Art, Post-Impressionism, Realism, Rococo, Romanticism, Surrealism, Symbolism and Ukiyo-e. Based on Deep Learning, the underlying model was trained with a set of 60 000 paintings annotated with their style. When you submit a painting, Rasta will display what it considers to be the three most probable artistic style (represented as a probability distribution). This work was published in a machine learning conference, but more importantly it has created an interaction with the artistic community and with a wider audience:

- an analysis of Rasta's mistakes by Erik Levesque was published in the journal *Réalités Nouvelles*;
- an article in Science et Avenir, a popular science magazine;
- an article in Artension 152, a magazine for art enthusiasts.

FlauBERT is one of the first large scale language model dedicated to French, and the first as an open science project: every step is shared, from the data collection and pre-processing to how to use such kind of model. This resource is publicly available and its purpose is to offer a powerful tool for researchers (in computer science or linguistic) and beyond to startups.
A *patent* were registered on methods for learning parameters of a neural network able to generate a trajectory of an exoskeleton and for setting the exoskeleton in motion. This "invention" includes:

- learning parameters of a first neural network suitable for generating periodic elementary trajectories of the exoskeleton.
- Learning, using parameters from the first neural network, parameters of a second neural network suitable for generating periodic elementary trajectories of the exoskeleton.

This patent is the result of the collaboration with the company Wandercraft.

Standard 3. The team shares its knowledge with the general public and takes part in debates in society.

Several members of the Data Science team are involved in actions to share knowledge with a general and wider audience:

- PSL has created in 2021 a working group on the topic on sharing scientific knowledge with a large audience and through several initiatives. Several members of the team are involved in this working group.
- One member has written two articles on Recommender systems in the journal *Interstices* (2018) and *Grand Angle-CGE* (2017)
- Podcast ”Dauphine Numérique: Ex Machina”: 4 programs (see the portfolio)
- Chapter in the collective book on Big Data edited by the CNRS : Bernd Amann, Daniela Grigori, ”L'intégration de données massives, hétérogènes et distribuées”, Les Big Data à découvert, pp. 104-105, (ISBN : 978-2-271-11464-8) (2017)

4 TRAJECTORY : DATA SCIENCE TEAM

In this section, we first develop an analysis that depicts the strengths, weaknesses, threats and opportunities that the Data Science team will have to face during the next period. Then research perspectives and evolution are summarized.

Strengths. The research topics covered by the Data Science team are well identified and positioned in a lively national and international context. The team's results are of a very high standard, as showcased by awards and the publications in top-level journals and conferences. Some members are seasoned researchers having long term experiences. They have been involved as consultants for government agencies and outside bodies. Some members are active in producing prototypes demonstrating the team's ability to be effective in both theoretical and practical aspects. The team has benefited from numerous research grants and contracts with industry. These features are emphasized by the attractiveness for Ph.D. students, CNRS researchers and postdocs.

Weaknesses. There are three major weaknesses which need to be considered in the very near future. The first concerns the members increasing workload. Almost all the members of the team are deeply involved in a lot of teaching and pedagogical activities (in charge of masters, certification, networks), in management activities (unit director, Deputy Vice President, head of PSL programs) and in consultancy activities (H2020, HCERES, ANR). The second weakness is the lack of links between projects of the team. There is a need to improve the cohesion between the projects which, although encouraged by calls for “inter-project internships”, has not yet resulted in joint publications between the projects. The third weakness concerns the lack of CNRS researchers in the team (only one DR CNRS) and research engineers who could speed up implementation tasks, and allow researchers to better focus on theoretical aspects.

Opportunities. From one hand, the environment and the location of our team is clearly favorable to attractiveness of good students and visiting professors: the context of both Dauphine University and PSL is an asset for attractiveness. From the other hand, there is growing agreement that research topics of all our team projects: machine learning, data and process management are of great importance to many other fields. We therefore plan to take advantage of this trend and our environment in order to increase our funding and research activities. This includes the possibility of creating Post-doc and Ph.D. positions, as well as new computing infrastructures, particularly useful for massive data processing.

Threats. The main threat is the low number of new researchers being recruited. Two new CNRS researchers left during the last period for industry. Thus, the risk of a reduction or a loss in skills in data management and engineering is not negligible, with a consequence on our students too. Our experience has shown that courses in these areas are crucial for them, as companies need strong and advanced skills in these areas, both for research and business development. Another threat is the involvement in implementation activities that takes a lot of time and energy. We fear that without the support of research engineers, our product development for cultural, economic and social worlds will be in decline.

The trajectory of the Data Science team for the next five years derives from elements we highlighted in the present document. Indeed, we plan to maintain the current organizational structure of our team, however, as our research topics evolves, the following changes and developments are expected in the next five years:

- The research topics of our team will revolve around three projects: MADAX, MILES, and a new project MINER built on ongoing work.

- The Policy Analytics project will be transferred to the Decision Aiding team, where the majority of the members involved in this project are already located. Therefore, the trajectory of this project will be presented in the Decision Aiding team self-assessment document.
- The MADAX project will benefit from the participation of some members from Data Science team, who plan to work on aspects related to Knowledge Graphs. Recently, this domain has attracted a lot of attention in the research community, as knowledge graphs have the strength of putting advantages coming from a graph-bases representation of data, and the presence and use of analogies enabling forms of reasoning that are complementary to many other kinds of data analysis, including graph querying and machine learning.

The remaining part of this section presents the research perspectives for each the three project.

Massive Data Management, Analysis and Exploration (MADAX).

Concerning Knowledge Graphs (KGs) the main research lines we plan to develop are the following ones.

- KGs can represent useful information about connections among users and products/items in Context Aware Recommendation Systems. By leveraging the additional presence of semantic information about those connections and possible contexts, we plan to study and devise techniques to discover latent aspects that can improve both recommendation quality and their explanation.
- Since 2015, the United Nations, in conjunction with other collaborators, have defined the Sustainable Development Goals (SDGs) as a set of 17 goals aimed at improving living conditions on the planet by 2030. Achievement of targets and indicators for SDGs are of key interest to policy-makers worldwide and KGs are particularly adapted for data-driven application, such as effective policies and initiatives for environmental sustainability. In this context we plan to study the use of KGs with machine learning techniques to provide an updated and synthetic panel on actions taken and their impacts on the planet. We will study the evolution of the graphs, the problems of updating and refreshing them from real-time sources (e.g., sensors, United Nations reports).
- One of the most significant areas of current computational biology research is the community detection in protein-protein interaction (PPI) networks. In this context we plan to identify and propose an improved property-graph representation of a PPI, including topological information, such as the different protein-protein interactions levels, and semantic information stored in biological ontologies such as the functional similarities between proteins. Then, we plan to devise advanced mining graph techniques that will be applied to detect genes' communities in the PPI and their hidden features.
- Several problems arise when managing KGs, ranging from their construction to their exploration and exploitation. We will mainly focus on the management of dynamic KGs. Indeed, knowledge is intrinsically dynamic: the sources that feed the KG can undergo changes that have an impact on the KG itself. Moreover, new promising sources can be added to the list of sources used to enrich the KG, and other sources that are no longer relevant can be dropped, which in turn has an impact on the facts (nodes and relations) composing the KG. In this context, we plan to design new solutions to assist KG providers and users to better handle the effects of dynamic KGs. A particular attention will be given to the profiling of KGs by relying and making available useful provenance information.

After the past results obtained in the context of semi-structured data (JSON representation and JSON Schemas description), the MADAX project plans to develop the following research activities:

- Data generation for JSON Schema; generating many instances of a JSON Schema is crucial for software verification and bench-marking systems for analytic and machine learning over data coming from JSON Data collections. Important problems that remain open are how to identify properties and strategies ensuring a good covering of a given schema, or of constraints posed by the user, both on the structure of data and on the distribution of base values. In this context on problem is to find a good balance between expressive and cost of data generation, since complete and very precise techniques incur in exponential time complexity, while for data generation we need systems that are fast and flexible enough to meet new needs that may arise.

- Rather than starting from a JSON Schema for data-generation, we plan to consider the problem of *exploding* a given data-set. The typical situation is the following: a user has a an existing JSON data-set, but she needs a bigger, different one, preserving some features of the existing data set and, in addition, featuring new aspects that are important for testing, learning and benchmarking purposes. To this goal, we plan to study, formalize and implement techniques for JSON data explosion, by leveraging distribution and parallelism, in order to tame the high volume of generated data and the use of time expensive approaches for data expansion.

- A crucial problem in JSON data management is validation wrt a given JSON Schema. Recently JSON Schema has been enriched with new mechanisms which we believe drastically impact the validation process, and user interactions. So we plan to dedicate research efforts in understanding what are the consequences of these extensions, first at the level of complexity of the validation problem, and then at the level of the formalization of systems that are able to produce rich information during the validation process that can be used by users and/or applications for managing schema evolution tasks or to automatically repair invalid data-sets.

Concerning evolutionary learning from massive data, we will continue to propose solutions to scale up Evolutionary Algorithms (EAs) in order to handle, in addition to the volume, the imbalance character of massive data, which is a big challenge for machine learning techniques. Thanks to the iterative engine of EAs, it is possible to extend some scaling solutions based on active and ensemble learning with sampling techniques for imbalanced data, such as SMOTE (Synthetic Minority Oversampling TEchnique) or BUS (Balanced Under Sampling). These solutions will be applied essentially for medical diagnosis and high imbalanced detection problems such as fraud detection.
Otherwise, a new project is started in the context of biodiversity monitoring of marine ecosystems, that is a matter of great importance in our days. Biodiversity monitoring is essentially carried out by estimating the biotic index based on environmental DNA extracted from the collected samples. This process provides a large amount of information that can be used as input to a machine learning system in order to generate a predictive model of the biotic index. In this context, we suggest the use of a hybrid EA, scaled with some previously approved solutions for handling the volume and the variety of the data. Providing generic predictive models can allow a more frequent and expanded monitoring.

Mining Business Processes and Software (MINER).

The MINER project research topics follows on our work on process mining and our work on mining evolving software dependencies. The main research line we plan to develop are the following ones:

- We will address current open problems in process mining. While many process analytics techniques have been proposed that are effective for repetitive processes automated by using process management systems, many challenges have to be tackled in order to apply them in real contexts, for different types of processes and ensuring some non-functional guarantees. Majority of techniques work on structured event logs. But process related information can be found in databases logs, emails, social networks exchanges, logs of outsourced web services. We will propose analytics techniques for unstructured logs (emails,..) and related to multiple objects (object-centric logs). Analyzing very flexible processes is also challenging due to the large variability of the behavior of different process instances that lead to complex models. In order to help analysts understanding the process, we plan to extend our work on behavior patterns in order to be able to discover contextual patterns, i.e., to be able to automatically discover correlation between contexts (combinations of process variables) and patterns and to characterize the patterns as specific to a context or generic (common in every context). Other future direction is addressing ethical issues like privacy and fairness, that have just been recently brought into discussion in the field of process analytics.

- Machine learning techniques offer a new approach to smells detection, technical debt analysis and quality prediction in software. However, they concern the lower levels of the software, whereas we are interested in the ecosystem level. Therefore, in this context, we plan to study how existing machine learning solutions for code smell detection, technical debt analysis and quality prediction can be adapted to the analysis of dependency evolution in software ecosystems. Another challenge is to collect data sets to form models. In the literature, the main source for experimentation remains open source repositories. The advantage is that with the history of software versions, it might be possible to predict the evolution on an older version and to check whether this happens or not on later ones. Another research challenge, that we will tackle, arises in terms of how to retrieve training data for the machine learning application.

Machine Intelligence and Learning Systems (MILES).

During the last five years, the machine learning field has drastically changed. New paradigms have appeared, such as large scale (language) models along with self-supervised learning and the emergence of new architectures like transformers. Maybe a crucial point is also the rate of change and how fast new models can move from research lab to our everyday life. Important consequences can result from this important lack of insight. In this context, **trustworthy machine learning** is a genuine social issue and is the grounding research topic of the MILES team. The notion of trust in our context can be explored along different tracks: **fairness, robustness, privacy, and explicability** along with their relations to decision making. As described in the progress report part, the team has established the foundations of long term research on these tracks with a national and international visibility.
For the next period, the team plans to further explore this topic and the newcomers during the past period enable us to explore new extensions (on the optimization side) and applications (natural and speech processing, along with Physics). More precisely, the team is involved in four different PEPR (Programmes et équipements prioritaires de recherche). These programs are consistent with our research goals explained above and will structure our effort. They are expected to start soon (summer or fall 2023) and will cover the next period. Here is a short description for each of them.

Theoretical Foundations of Machine Learning: The interaction between theory and applications of machine learning is at the core of our approach. The trust in machine learning must rely on theoretical guarantee, while the recent progress on some downstream tasks completely changes many perspectives. For instance, recent observations reveal that large scale generative models can amplify biaises present in the training data, hence redrawing the challenge of fairness: how to train fair models without ”perfect sampling” and how to measure the fairness.

The second challenge is to carry on our exploration of provable robustness by connecting optimal transport game theory, and adversarial attacks. Finally, in decision making situations involving multiple agents, it is often assumed that each agent is perfectly rational. However, this assumption is not realistic for humans or machines. To address this issue, various concepts of bounded rationality have been developed and studied, especially in game theory. However, very little research has been done on extending this notion to sequential decision making environments such as stochastic games. This research will be pursued with Rida Laraki, who has extensive experience in the field of stochastic games.

Cyber-security: Multimedia data protection is certainly the area of cyber-security that has benefited the most from AI in the last decade. However the verification of the security level of these new tools has been neglected. Then AI has become one of the weakest links in multimedia data protection, and a scientific challenge. More than a downstream application, this notion of cyber-security is nowadays a real opportunity to study the security of deep learning by mapping its cardinal values (integrity, confidentiality and identification) to deep learning data (training data, test data, learned model). This parallelism between the protection of multimedia data and the security of deep learning creates a virtuous cross-fertilization that ensures end-to-end security, from the data to the algorithms that process it. This cross-fertilization is firmly grounded in MILES as explained before. and the team plans to build on recent work for deeper exploration of how game theory can improve adversarial training, how to build adversarial attack against randomized models to reach some theoretical guarantee, and how to design relevant loss functions for the adversarial framework.

Frugality in AI: The ecological impact of AI has emerged as a major concern for researchers in machine learning, and soon, this will also be the case for the society as an acceptability criterion. The energy consumption of large scale models has become the downside of the recent success. The major challenge is therefore to design, analyze and deploy intrinsically frugal models (neural or not) able to achieve the versatility and performance of the best models while requiring only a vanishing fraction of the resources currently needed. The long term vision is that the resources required to train machine learning models can be decreased by several orders of magnitude, with negligible performance drop, compared to the state of the art. A first axis focuses on the efficient introduction of prior knowledge. Of course it depends on the data and the type of knowledge available, e.g., physical models, temporal structure, graph properties or expected expert knowledge. While this topic is not new, the recent neural architectures give rise to new challenges but also to new promises. For the second axis, since the training cost of such models is so huge, they should be designed for sharing and evolution. Hence the team will investigate lifelong learning and reusable foundation models. At last but not least, the design of frugal and unbiased generative models for Multimedia Understanding (like e.g text and speech) is clearly today one of the main challenge that the team can address.

Digital Health: The healthcare sector generates a vast amount of data from various sources. Along with advanced machine learning algorithms, it has a huge potential for future developments of personalized treatments and policies, as well as for healthcare delivery and outcomes. However, the highly sensitive nature of personal health data limits their widespread exploitation and hinders the potential benefits. More specifically, a first challenge for the team is differential privacy in the context of continual release of statistics. The recent COVID-19 outbreak highlights the significant risks for individuals' privacy, such as re-identification, loss of confidentiality, discrimination, and misuse of data. The second challenge arises from the emergence of large scale and over-parametrized models (like Chat-GPT). The notion of privacy must be redefined appropriately in this new context of interpolation regime. Finally, it is important for both robustness and privacy to be considered and addressed jointly in order to exploit synergies between the two, but also to overcome the following pitfall: algorithms and models that are designed to be adversarially robust can be more vulnerable to privacy attacks and vice et versa.

These four projects together constitute an ambitious research program. Its consistency relies on the shared topics and their cross-fertilization: privacy and robustness, frugality and efficient training, game theory, optimal transport and adversarial attack. Of course, the team needs to grow to address these challenges. For this purpose, the PEPR programs and industrial collaborations will provide funding to hire PhD students as well as postdocs. It will also help us in terms of attractiveness for future permanent positions.

5 BIBLIOGRAPHY : DATA SCIENCE TEAM

[1] Khalid Belhajjame. On Answering Why-Not Queries Against Scientific Workflow Provenance. In *Proceedings of the 21st International Conference on Extending Database Technology, EDBT 2018*, Vienna, Austria, March 2018.

[2] Sarah Cohen-Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, and Christophe Blanchet. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, 75:284–298, October 2017.

[3] Alban Gaignard, Khalid Belhajjame, and Hala Skaf-Molli. SHARP: Harmonizing and Bridging Cross-Workflow Provenance. In *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events*, volume 10577 of *Lecture Notes in Computer Science*, pages 219–234, Portoro, Slovenia, June 2017. Springer International Publishing.

[4] Pinar Alper, Khalid Belhajjame, and Carole Goble. Static analysis of Taverna workflows to predict provenance patterns. *Future Generation Computer Systems*, 75:310–329, October 2017.

[5] Khalid Belhajjame and Mahmoud Barhamgi. Querying Data Preparation Modules Using Data Examples. In *9th International Provenance and Annotation Workshop*, volume 12839 of *Lecture Notes in Computer Science*, pages 211–217, Virtual, United States, July 2021. Springer International Publishing.

[6] Khalid Belhajjame. On Discovering Data Preparation Modules Using Examples. In *18th International Conference Service-Oriented Computing*, pages 56–65, Dubai, United Arab Emirates, December 2020.

[7] Pinar Alper, Khalid Belhajjame, Vasa Curcin, and Carole Goble. LabelFlow Framework for Annotating Workflow Provenance. *Informatics*, 5(1):11, March 2018.

[8] Khalid Belhajjame. On the Anonymization of Workflow Provenance without Compromising the Transparency of Lineage. *Journal of data and information quality*, 14(1):1–27, March 2022.

[9] Khalid Belhajjame. Lineage-Preserving Anonymization of the Provenance of Collection-Based Workflows. In *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT*, Copenhagen, Denmark, 2020.

[10] Khalid Belhajjame, Noura Faci, Zakaria Maamar, Vanilson Burégio, Edvan Soares, and Mahmoud Barhamgi. On privacy-aware eScience workflows. *Computing*, 102(5):1171–1185, May 2020.

[11] Alban Gaignard, Hala Skaf-Molli, and Khalid Belhajjame. Findable and reusable workflow data products: A genomic workflow case study. *Semantic Web – Interoperability, Usability, Applicability*, pages 1–13, May 2020.

[12] Alban Gaignard, Hala Skaf-Molli, and Khalid Belhajjame. Découvrabilité et réutilisation de données produites par des workflows : un cas d'usage en génomique. In Maxime Lefrançois, editor, *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA'21)*, pages pp 73–80, Bordeaux, France, June 2021.

[13] Alban Gaignard, Khalid Belhajjame, and Hala Skaf-Molli. SHARP: Harmonizing cross-workflow provenance. In *SeWeBMeDA 2017 : Semantic Web solutions for large-scale BioMedical Data Analtics - 14th ESWC 2017*, Portoroz, Slovenia, May 2017.

[14] Yudith Cardinale, Sonia Guehis, and Marta Rukoz. Big Data Analytic Approaches Classification. In *12th International Conference on Software Technologies (ICSOFT 2017)*, pages 151–162, Madrid, Spain, July 2017.

[15] Yudith Cardinale, Sonia Guehis, and Marta Rukoz. Classifying Big Data Analytic Approaches: A Generic Architecture. In *12th International Joint Conference (ICSOFT 2017)*, pages 268–295, Madrid, Spain, July 2017.

[16] Alain Petrowski and Sana Ben Hamida. *Evolutionary Algorithms*. John Wiley & Sons, Ltd, 2017.

[17] Hmida Hmida, Sana Ben Hamida, Amel Borgi, and Marta Rukoz. Scale Genetic Programming for large Data Sets: Case of Higgs Bosons Classification. *Procedia Computer Science*, 126:302–311, 2018.

[18] Hmida Hmida, Sana Ben Hamida, Amel Borgi, and Marta Rukoz. A new adaptive sampling approach for Genetic Programming. In *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pages 1–8, Marrakech, France, October 2019. IEEE.

[19] Sana Ben Hamida, Hmida Hmida, Amel Borgi, and Marta Rukoz. Adaptive sampling for active learning with genetic programming. *Cognitive Systems Research*, 65:23 – 39, January 2021.

[20] Sana Ben Hamida and Ghita Benjelloun. Extending DEAP with Active Sampling for Evolutionary Supervised Learning. In *Proceedings of the 16th International Conference on Software Technologies, ICSOFT 2021*, Online Streaming, France, August 2021.

[21] Hmida Hmida, Sana Ben Hamida, Amel Borgi, and Marta Rukoz. Genetic Programming over Spark for Higgs Boson Classification. In *22nd International Conference Business Information Systems*, Business Information Systems 22nd International Conference, BIS 2019, Seville, Spain, June 26–28, 2019, Proceedings, pages 300–312, Seville, Spain, June 2019.

[22] Sana Ben Hamida, Ghita Benjelloun, and Hmida Hmida. Trends of Evolutionary Machine Learning to Address Big Data Mining. In Inès Saad, Camille Rosenthal-Sabroux, Faiez Gargouri, and Pierre-Emmanuel Arduin, editors, *Information and Knowledge Systems. Digital Technologies, Artificial Intelligence and Decision Making*, volume 425 of *Lecture Notes in Business Information Processing*, pages 85–99. Springer International Publishing, 2021.

[23] Marwa Ben M'barek, Amel Borgi, Walid Bedhiafi, and Sana Ben Hamida. Genetic Algorithm for Community Detection in Biological Networks. *Procedia Computer Science*, 126(6):195–204, 2018.

[24] Marwa Ben M'Barek, Amel Borgi, Sana Ben Hamida, and Marta Rukoz. Genetic Algorithm to Detect Different Sizes' Communities from Protein-Protein Interaction Networks. In *14th International Conference on Software Technologies*, pages 359–370, Prague, Czech Republic, July 2019. SCITEPRESS - Science and Technology Publications.

[25] Marwa Ben M'barek, Amel Borgi, Sana Ben Hmida, and Marta Rukoz. Generic GA-PPI-Net: Generic Evolutionary Algorithm to Detect Semantic and Topological Biological Communities. In *15th International Conference on Software Technologies*, pages 295–306,

Lieusaint - Paris, France, July 2020. SCITEPRESS - Science and Technology Publications.

[26] Marwa Ben M'barek, Amel Borgi, Sana Ben Hmida, and Marta Rukoz. GA-PPI-Net: A Genetic Algorithm for Community Detection in Protein-Protein Interaction Networks. In *Communications in Computer and Information Science book series (CCIS, volume 1250)*, pages 133–155. July 2020.

[27] Marwa Ben M'barek, Sana Ben Hmida, Amel Borgi, and Marta Rukoz. GA-PPI-Net Approach vs Analytical Approaches for Community Detection in PPI Networks. In *the 25th International Conference KES-2021*, volume 192, pages 903–912, Szczecin, Poland, September 2021.

[28] Mohammad Amin Farvardin, Dario Colazzo, Khalid Belhajjame, and Carlo Sartiani. Scalable saturation of streaming RDF triples. *Trans. Large Scale Data Knowl. Centered Syst.*, 44:1–40, 2020.

[29] Mohammad Amin Farvardin, Dario Colazzo, Khalid Belhajjame, and Carlo Sartiani. Streaming saturation for large RDF graphs with dynamic schema information. In Alvin Cheung and Kim Nguyen, editors, *Proceedings of the 17th ACM SIGPLAN International Symposium on Database Programming Languages, DBPL 2019, Phoenix, AZ, USA, June 23, 2019*, pages 42–52. ACM, 2019.

[30] Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Linear time membership in a class of regular expressions with counting, interleaving, and unordered concatenation. *ACM Trans. Database Syst.*, 42(4):24:1–24:44, 2017.

[31] Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Counting types for massive JSON datasets. In Tiark Rompf and Alexander Alexandrov, editors, *Proceedings of The 16th International Symposium on Database Programming Languages, DBPL 2017, Munich, Germany, September 1, 2017*, pages 9:1–9:12. ACM, 2017.

[32] Mohamed Amine Baazizi, Houssem Ben Lahmar, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Schema inference for massive JSON datasets. In Volker Markl, Salvatore Orlando, Bernhard Mitschang, Periklis Andritsos, Kai-Uwe Sattler, and Sebastian Breß, editors, *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017*, pages 222–233. OpenProceedings.org, 2017.

[33] Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. A type system for interactive JSON schema inference (extended abstract). In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, volume 132 of *LIPIcs*, pages 101:1–101:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[34] Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Parametric schema inference for massive JSON datasets. *VLDB J.*, 28(4):497–521, 2019.

[35] Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, and Stefanie Scherzinger. An empirical study on the "usage of not" in real-world JSON schema documents. In Aditya K. Ghose, Jennifer Horkoff, Vítor E. Silva Souza, Jeffrey Parsons, and Joerg Evermann, editors, *Conceptual Modeling - 40th International Conference, ER 2021, Virtual Event, October 18-21, 2021, Proceedings*, volume 13011 of *Lecture Notes in Computer Science*, pages 102–112. Springer, 2021.

[36] Lyes Attouche, Mohamed Amine Baazizi, Dario Colazzo, Yunchen Ding, Michael Fruth, Giorgio Ghelli, Carlo Sartiani, and Stefanie Scherzinger. A test suite for JSON schema containment. In Roman Lukyanenko, Binny M. Samuel, and Arnon Sturm, editors, *Proceedings of the ER Demos and Posters 2021 co-located with 40th International Conference on Conceptual Modeling (ER 2021), St. John's, NL, Canada, October 18-21, 2021*, volume 2958 of *CEUR Workshop Proceedings*, pages 19–24. CEUR-WS.org, 2021.

[37] Lyes Attouche, Mohamed amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani andStefanie Scherzinge. Witness generation for json schema. *Proc. VLDB Endow.*, 15(3):4002 – 4014, 2022.

[38] Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Schemas and types for JSON data: From theory to practice. In Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska, editors, *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 2060–2063. ACM, 2019.

[39] Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Schemas and types for JSON data. In Melanie Herschel, Helena Galhardas, Berthold Reinwald, Irini Fundulaki, Carsten Binnig, and Zoi Kaoudi, editors, *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, pages 437–439. OpenProceedings.org, 2019.

[40] Mohamed Lamine Mouhoub, Daniela Grigori, and Maude Manouvrier. Towards an Automatic Enrichment of Semantic Web Services Descriptions. In *Confederated International Conferences: CoopIS, C&TC, and ODBASE 2017*, pages 681–697, Rhodes, Greece, October 2017.

[41] D. Grigori and Ahmed Gater. PSearch: a framework for semantic annotated process model search. *Service Oriented Computing and Applications*, 11(3), 2017.

[42] Khalid Belhajjame and Daniela Grigori. On Reuse in Service-Based Workflows. In *Next-Gen Digital Services. A Retrospective and Roadmap for Service Computing of the Future*, volume 12521 of *Lecture Notes in Computer Science*, pages 77–87. Springer International Publishing, April 2021.

[43] Khalid Belhajjame, Daniela Grigori, Mariem Harmassi, and Manel Ben Yahia. Keyword-Based Search of Workflow Fragments and Their Composition. In *Transactions on Computational Collective Intelligence XXVI*, pages 67–90. 2017. LNCS, volume 10190; TCCI, volume 10190.

[44] Virginie Gabrel, Maude Manouvrier, Kamil Moreau, and Cecile Murat. QoS-aware Automatic Syntactic Service Composition problem: complexity and resolution. *Future Generation Computer Systems*, 80:311–321, March 2018.

[45] Merzoug Soltane, Yudith Cardinale, Rafael Angarita, Philippe Rosse, Marta Rukoz, Derdour Makhlouf, and Kazar Okba. A self-adaptive agent-based system for cloud platforms. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–8, 2018.

[46] Amine Louati, Joyce El Haddad, and Suzanne Pinson. Formation de coalitions pour une composition de services Web fondée sur la confiance dans les réseaux sociaux. In *Journées Francophones sur les Systèmes Multi-Agents (JFSMA 2017)*, Caen, France, July 2017.

[47] Yudith Cardinale, Joyce El Haddad, Maude Manouvrier, and Marta Rukoz. Fuzzy ACID properties for self-adaptive composite cloud services execution. *Concurrency and Computation: Practice and Experience*, 31(2), January 2019.

[48] Maude Manouvrier, Cesare Pautasso, and Marta Rukoz. Microservice Disaster Crash Recovery: A Weak Global Referential Integrity Management. In *Computational Science – ICCS 2020. 20th International Conference*, pages 482–495, Amsterdam, Netherlands, June 2020. LNCS volume 12138.

[49] Zakaria Maamar, Noura Faci, Joyce El Haddad, Fadwa Yahya, and Mohammad Askar. Multi-party Contract Management for Microservices. In *17th International Conference on Software Technologies*, pages 276–287, Lisbon, Portugal, July 2022. SCITEPRESS - Science and Technology Publications.

[50] Hiba Alili, Khalid Belhajjame, Rim Drira, Daniela Grigori, and Henda Hajjami Ben Ghezala. Quality Based Data Integration for Enriching User Data Sources in Service Lakes. In *IEEE International Conference on Web Services (ICWS 2018)*, pages 163–170, San Francisco, United States, July 2018.

[51] Hiba Alili, Rim Drira, Khalid Belhajjame, Henda Ben Ghezala, and Daniela Grigori. A Model-Driven Framework for the Modeling and the Description of Data-as-a-Service to Assist Service Selection and Composition. In *30th International Conference on Database and Expert Systems Applications (DEXA 2019)*, pages 396–406, Linz, Austria, August 2019.

[52] Hiba Alili, Khalid Belhajjame, Daniela Grigori, Rim Drira, and Henda Hajjami Ben Ghezala. On Enriching User-Centered Data Integration Schemas in Service Lakes. In *Business Information Systems 20th International Conference (BIS 2017)*, pages 3–15, Poznan, Poland, June 2017.

[53] Diana Jlailaty, Daniela Grigori, and Khalid Belhajjame. A framework for mining process models from emails logs. working paper or preprint, May 2019.

[54] Diana Jlailaty, Daniela Grigori, and Khalid Belhajjame. Multi-level clustering for extracting process-related information from email logs. In *11th IEEE International Conference on Research Challenges in Information Science (RCIS 2017)*, pages 455–456, Brighton, United Kingdom, May 2017.

[55] Diana Jlailaty, Daniela Grigori, and Khalid Belhajjame. Business Process Instances Discovery from Email Logs. In *2017 IEEE International Conference on Services Computing (SCC)*, pages 19–26, Honolulu, Hawaii, United States, June 2017.

[56] Diana Jlailaty, Daniela Grigori, and Khalid Belhajjame. On the elicitation and annotation of business activities based on emails. In *SAC '19: The 34th ACM/SIGAPP Symposium on Applied Computing*, pages 101–103, Limassol Cyprus, France, June 2019. ACM.

[57] Diana Jlailaty, Daniela Grigori, and Khalid Belhajjame. Mining Business Process Activities from Email Logs. In *2017 IEEE 1st International Conference on Cognitive Computing (ICCC 2017)*, pages 112–119, Honolulu, Hawaii, United States, June 2017.

[58] Diana Jlailaty, Daniela Grigori, and Khalid Belhajjame. Email Business Activities Extraction and Annotation. In *12th International Workshop, ISIP 2018*, pages 69–86, Fukuoka, Japan, May 2018. Communications in Computer and Information Science book series (CCIS, volume 1040).

[59] Raphael Azorin, Daniela Grigori, and Khalid Belhajjame. A Reproducible Approach for Mining Business Activities from Emails for Process Analytics. In *Service-Oriented Computing – ICSOC 2021 Workshops. ICSOC 2021*, volume 13236 of *Lecture Notes in Computer Science*, pages 77–91, Dubai, United Arab Emirates, November 2021. Springer International Publishing.

[60] Mehdi Acheli, Daniela Grigori, and Matthias Weidlich. Efficient Discovery of Compact Maximal Behavioral Patterns from Event Logs. In *31st International Conference on Advanced Information Systems Engineering (CAiSE 2019)*, pages 579–594, Rome, Italy, June 2019.

[61] Pavlos Delias, Athanasios Lagopoulos, Grigorios Tsoumakas, and Daniela Grigori. Using multi-target feature evaluation to discover factors that affect business process behavior. *Computers in Industry*, 99, 2018.

[62] Pavlos Delias, Mehdi Acheli, and Daniela Grigori. Applying the Method of Reflections through an Event Log for Evidence-based Process Innovation. In *2019 International Conference on Process Mining ICPM 2019*, pages 105–112, Aachen, Germany, June 2019.

[63] Damien Jaime, Joyce El Haddad, and Pascal Poizat. A Preliminary Study of Rhythm and Speed in the Maven Ecosystem. In *21st Belgium-Netherlands Software Evolution Workshop*, Mons, Belgium, September 2022.

[64] Lyes Attouche, Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, and Stefanie Scherzinger. Witness Generation for JSON Schema. *Proceedings of the VLDB Endowment (PVLDB)*, 15(13):4002–4014, September 2022.

[65] Mehdi Acheli, Daniela Grigori, and Matthias Weidlich. Discovering and Analyzing Contextual Behavioral Patterns From Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5708–5721, December 2022.