RESEARCH CENTRE

Paris

IN PARTNERSHIP WITH:

Ecole normale supérieure de Paris, CNRS

2021

ACTIVITY REPORT

Project-Team

VALDA

Value from Data

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team VALDA

Creation of the Project-Team: 2018 January 01

Keywords

Computer sciences and digital sciences

A3.1. – Data

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.3. – Distributed data

A3.1.4. – Uncertain data

A3.1.5. – Control access, privacy

A3.1.6. – Query optimization

A3.1.7. – Open data

A3.1.8. – Big data (production, storage, transfer)

A3.1.9. – Database

A3.1.10. – Heterogeneous data

A3.1.11. – Structured data

A3.2. – Knowledge

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.2.6. – Linked data

A3.3.2. – Data mining

A3.4.3. – Reinforcement learning

A3.4.5. – Bayesian methods

A3.5.1. – Analysis of large graphs

A4.7. – Access control

A7.2. – Logic in Computer Science

A7.3. – Calculability and computability

A9.1. – Knowledge

A9.8. – Reasoning

Other research topics and application domains

B6.3.1. – Web

B6.3.4. – Social Networks

B6.5. – Information systems

B9.5.6. – Data science

B9.6.5. – Sociology

B9.6.10. – Digital humanities

B9.7.2. – Open data

B9.9. – Ethics

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Serge Abiteboul [Inria, Emeritus, HDR]
- Camille Bourgaux [CNRS, Researcher]
- Olivier Cappé [CNRS, Senior Researcher, until Sep 2021]
- Luc Segoufin [Inria, Senior Researcher]
- Michael Thomazo [Inria, Researcher]
- Victor Vianu [Inria, Advanced Research Position, from Jun 2021 until Aug 2021]

Faculty Members

- Pierre Senellart [Team leader, École Normale Supérieure de Paris, Professor]
- Leonid Libkin [École Normale Supérieure de Paris, Professor]

Post-Doctoral Fellows

- Nofar Carmeli [École Normale Supérieure de Paris]
- Liat Peterfreund [CNRS, until Aug 2021]

PhD Students

- Juliette Achddou [1000 Mercis, CIFRE, until Sep 2021]
- Anatole Dahan [Université de Paris]
- Baptiste Lafosse [École Normale Supérieure de Paris, from October 2021]
- Shrey Mishra [École Normale Supérieure de Paris]
- Yann Ramusat [École Normale Supérieure de Paris, ATER]
- Alexandra Rogova [Université de Paris, from October 2021]
- Yoan Russac [École Normale Supérieure de Paris, until Sep 2021]

Interns and Apprentices

- Salome Attar [École Normale Supérieure de Paris, Intern, from Feb 2021 until Jul 2021]
- Riccardo Corona [École Normale Supérieure de Paris, Intern, from Mar 2021 until Aug 2021]
- Gregoire Lothe [Inria, Intern, from Apr 2021 until Jul 2021]
- Demir Renaux [Inria, Intern, from Mar 2021 until Aug 2021]

Administrative Assistant

- Meriem Guemair [Inria]

2 Overall objectives

2.1 Objectives

> Valda's focus is on both *foundational and systems aspects of* complex *data management*, especially *human-centric data*. The data we are interested in is typically heterogeneous, massively distributed, rapidly evolving, intensional, and often subjective, possibly erroneous, imprecise, incomplete. In this setting, Valda is in particular concerned with the optimization of complex resources such as computer time and space, communication, monetary, and privacy budgets. The goal is to extract *value from data*, beyond simple query answering.

Data management [39, 48] is now an old, well-established field, for which many scientific results and techniques have been accumulated since the sixties. Originally, most works dealt with static, homogeneous, and precise data. Later, works were devoted to heterogeneous data [37] [40], and possibly distributed [70] but at a small scale.

However, these classical techniques are poorly adapted to handle the new challenges of data management. Consider human-centric data, which is either produced by humans, e.g., emails, chats, recommendations, or produced by systems when dealing with humans, e.g., geolocation, business transactions, results of data analysis. When dealing with such data, and to accomplish any task to extract value from such data, we rapidly encounter the following facets:

- *Heterogeneity*: data may come in many different structures such as unstructured text, graphs, data streams, complex aggregates, etc., using many different schemas or ontologies.

- *Massive distribution*: data may come from a large number of autonomous sources distributed over the web, with complex access patterns.

- *Rapid evolution*: many sources may be producing data in real time, even if little of it is perhaps relevant to the specific application. Typically, recent data is of particular interest and changes have to be monitored.

- *Intensionality* [1]: in a classical database, all the data is available. In modern applications, the data is more and more available only intensionally, possibly at some cost, with the difficulty to discover which source can contribute towards a particular goal, and this with some uncertainty.

- *Confidentiality and security*: some personal data is critical and need to remain confidential. Applications manipulating personal data must take this into account and must be secure against linking.

- *Uncertainty*: modern data, and in particular human-centric data, typically includes errors, contradictions, imprecision, incompleteness, which complicates reasoning. Furthermore, the subjective nature of the data, with opinions, sentiments, or biases, also makes reasoning harder since one has, for instance, to consider different agents with distinct, possibly contradicting knowledge.

These problems have already been studied individually and have led to techniques such as *query rewriting* [61] or *distributed query optimization* [66].

Among all these aspects, intensionality is perhaps the one that has least been studied, so we pay particular attention to it. Consider a user's query, taken in a very broad sense: it may be a classical database query, some information retrieval search, a clustering or classification task, or some more advanced knowledge extraction request. Because of intensionality of data, solving such a query is a typically dynamic task: each time new data is obtained, the partial knowledge a system has of the world is revised, and query plans need to be updated, as in adaptive query processing [54] or aggregated search [78]. The system then needs to decide, based on this partial knowledge, of the best next access to perform. This is reminiscent of the central problem of reinforcement learning [76] (train an agent to accomplish a task in a partially known world based on rewards obtained) and of active learning [72] (decide which

[1] We use the spelling *intensional*, as in mathematical logic and philosophy, to describe something that is neither available nor defined in *extension*; *intensional* is derived from *intension*, while *intentional* is derived from *intent*.

action to perform next in order to optimize a learning strategy) and we intend to explore this connection further.

Uncertainty of the data interacts with its intensionality: efforts are required to obtain more precise, more complete, sounder results, which yields a trade-off between *processing cost* and *data quality*.

Other aspects, such as heterogeneity and massive distribution, are of major importance as well. A standard data management task, such as query answering, information retrieval, or clustering, may become much more challenging when taking into account the fact that data is not available in a central location, or in a common format. We aim to take these aspects into account, to be able to apply our research to real-world applications.

2.2 The Issues

We intend to tackle hard technical issues such as query answering, data integration, data monitoring, verification of data-centric systems, truth finding, knowledge extraction, data analytics, that take a different flavor in this modern context. In particular, we are interested in designing strategies to *minimize data access cost towards a specific goal, possibly a massive data analysis task*. That cost may be in terms of communication (accessing data in distributed systems, on the Web), of computational resources (when data is produced by complex tools such as information extraction, machine learning systems, or complex query processing), of monetary budget (paid-for application programming interfaces, crowdsourcing platforms), or of a privacy budget (as in the standard framework of differential privacy).

A number of data management tasks in Valda are inherently intractable. In addition to properly characterizing this intractability in terms of complexity theory, we intend to develop solutions for solving these tasks in practice, based on approximation strategies, randomized algorithms, enumeration algorithms with constant delay, or identification of restricted forms of data instances lowering the complexity of the task.

3 Research program

3.1 Scientific Foundations

We now detail some of the scientific foundations of our research on complex data management. This is the occasion to review connections between data management, especially on complex data as is the focus of Valda, with related research areas.

Complexity & Logic Data management has been connected to logic since the advent of the relational model as main representation system for real-world data, and of first-order logic as the logical core of database querying languages [39]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [65], recursive queries (Datalog), or querying of XML databases [48]. Logical formalisms facilitate reasoning about the expressiveness of a query language or about its complexity.

The main problem of interest in data management is that of query evaluation, i.e., computing the results of a query over a database. The complexity of this problem has far-reaching consequences. For example, it is because first-order logic is in the AC_0 complexity class that evaluation of SQL queries can be parallelized efficiently. It is usual [77] in data management to distinguish *data complexity*, where the query is considered to be fixed, from *combined complexity*, where both the query and the data are considered to be part of the input. Thus, though conjunctive queries, corresponding to a simple SELECT-FROM-WHERE fragment of SQL, have PTIME data complexity, they are NP-hard in combined complexity. Making this distinction is important, because data is often far larger (up to the order of terabytes) than queries (rarely more than a few hundred bytes). Beyond simple query evaluation, a central question in data management remains that of complexity; tools from algorithm analysis, and complexity theory can be used to pinpoint the tractability frontier of data management tasks.

Automata Theory Automata theory and formal languages arise as important components of the study of many data management tasks: in temporal databases [38], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [44], queries are naturally given as automata; typical query and schema languages for XML databases such as XPath and XML Schema can be compiled to tree automata [69], or for more complex languages to data tree automata[34]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [52] that show that very expressive queries (from the language of monadic second-order language) can be evaluated as tree automata over *tree decompositions* of the original databases, yielding linear-time algorithms (in data complexity) for a wide variety of applications.

Verification Complex data management also has connections to verification and static analysis. Besides query evaluation, a central problem in data management is that of deciding whether two queries are *equivalent* [39]. This is critical for query optimization, in order to determine if the rewriting of a query, maybe cheaper to evaluate, will return the same result as the original query. Equivalence can easily be seen to be an instance of the problem of (non-)satisfiability: $q \equiv q'$ if and only if $(q \wedge \neg q') \vee (\neg q \wedge q')$ is not satisfiable. In other words, some aspects of query optimization are static analysis issues. Verification is also a critical part of any database application where it is important to ensure that some property will never (or always) arise [50].

Workflows The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of data management applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and to statically verify critical properties of the system. Different approaches to workflow specifications exist: automata-based, logic-based, or predicate-based control of function calls [36].

Probability & Provenance To deal with the uncertainty attached to data, proper models need to be used (such as attaching *provenance* information to data items and viewing the whole database as being *probabilistic*) and practical methods and systems need to be developed to both reliably estimate the uncertainty in data items and properly manage provenance and uncertainty information throughout a long, complex system.

The simplest model of data uncertainty is the NULLs of SQL databases, also called Codd tables [39]. This representation system is too basic for any complex task, and has the major inconvenient of not being closed under even simple queries or updates. A solution to this has been proposed in the form of *conditional tables* [63] where every tuple is annotated with a Boolean formula over independent Boolean random events. This model has been recognized as foundational and extended in two different directions: to more expressive models of *provenance* than what Boolean functions capture, through a semiring formalism [59], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [60]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [51, 45]. Research in the past ten years has focused on a better understanding of the tractability of query answering with provenance and probabilistic annotations, in a variety of specializations of this framework [75] [64, 42].

Machine Learning Statistical machine learning, and its applications to data mining and data analytics, is a major foundation of data management research. A large variety of research areas in complex data management, such as wrapper induction [71], crowdsourcing [43], focused crawling [58], or automatic database tuning [46] critically rely on machine learning techniques, such as classification [62], probabilistic models [57], or reinforcement learning [76].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [67] system result in probabilistic annotations that need to be properly modeled, stored, and queried.

Finally, complex data management also brings new twists to some classical machine learning problems. Consider for instance the area of *active learning* [72], a subfield of machine learning concerned with how to optimally use a (costly) oracle, in an interactive manner, to label training data that will be used to build a learning model, e.g., a classifier. In most of the active learning literature, the cost model is very basic (uniform or fixed-value costs), though some works [73] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [55]. These assumptions usually break when applied to complex data management problems on real-world data, such as crowdsourcing.

3.2 Research Directions

At the beginning of the Valda team, the project was to focus on the following directions:

- foundational aspects of data management, in particular related to query enumeration and reasoning on data, especially regarding security issues;
- implementation of provenance and uncertainty management, real-world applications, other aspects of uncertainty and incompleteness, in particular dynamic;
- development of personal information management systems, integration of machine learning techniques.

We believe the first two directions have been followed in a satisfactory manner. The focus on personal information management has not been kept for various organizational reasons, however, but the third axis of the project is reoriented to more general aspects of Web data management.

New permanent arrivals in the group since its creation have impacted its research directions in the following manner:

- Camille BOURGAUX and Michaël THOMAZO are both specialists of knowledge representation and formal aspects of knowledge bases, which is an expertise that did not exist in the group. They are also both interested in, and have started working on aspects related to connecting their research with database theory, and investigating aspects of uncertainty and incompleteness in their research. This will lead to more work on knowledge representation and symbolic AI aspects, while keeping the focus of Valda on foundations of data management and uncertainty.
- Olivier CAPPÉ is a specialist in statistics and machine learning, in particular multi-armed bandits and reinforcement learning. He is also interested in applications of these learning techniques to data management problems. His arrival in the group therefore complemented the expertise of other researchers, and led to more work on machine learning issues. In September 2021, Olivier CAPPÉ and his students left the Valda group to join the newly created *Center for Data Science* at ENS.
- Leonid LIBKIN is a specialist of database theory, of incomplete data management, and has a line of current research on graph data management. His profile fits very well with the original orientation of the Valda project.

We intend to keep producing leading research on the foundations of data management. Generally speaking, the goal is to investigate the borders of feasibility of various tasks. For instance, what are the assumptions on data that allow for computable problems? When is it not possible at all? When can we hope for efficient query answering, when is it hopeless? This is a problem of theoretical nature which is necessary for understanding the limit of the methods and driving research towards the scenarios where positive results may be obtainable. Only when we have understood the limitation of different methods and have many examples where this is possible, we can hope to design a solid foundation that allowing for a good trade-off between what can be done (needs from the users) and what can be achieved (limitation from the system).

Similarly, we will continue our work, both foundational and practical, on various aspects of provenance and uncertainty management. One overall long-term goal is to reach a full understanding of the interactions between query evaluation or other broader data management tasks and uncertain and annotated data models. We would in particular want to go towards a full classification of tractable (typically polynomial-time) and intractable (typically NP-hard for decision problems, or #P-hard for probability evaluation) tasks, extending and connecting the query-based dichotomy [53] on probabilistic query evaluation with the instance-based one of [41, 42]. Another long-term goal is to consider more dynamic scenarios than what has been considered so far in the uncertain data management literature: when following a workflow, or when interacting with intensional data sources, how to properly represent and update uncertainty annotations that are associated with data. This is critical for many complex data management scenarios where one has to maintain a probabilistic current knowledge of the world, while obtaining new knowledge by posing queries and accessing data sources. Such intensional tasks requires minimizing jointly data uncertainty and cost to data access.

As application area, in addition to the historical focus on personal information management which is now less stressed, we target Web data (Web pages, the semantic Web, social networks, the deep Web, crowdsourcing platforms, etc.).

We aim at keeping a delicate balance between theoretical, foundational research, and systems research, including development and implementation. This is a difficult balance to find, especially since most Valda researchers have a tendency to favor theoretical work, but we believe it is also one of the strengths of the team.

4 Application domains

4.1 Personal Information Management Systems

We recall that Valda's focus is on human-centric data, i.e., data produced by humans, explicitly or implicitly, or more generally containing information about humans. Quite naturally, we have used as a privileged application area to validate Valda's results that of personal information management systems (Pims for short) [35].

A Pims is a system that allows a user to integrate her own data, e.g., emails and other kinds of messages, calendar, contacts, web search, social network, travel information, work projects, etc. Such information is commonly spread across different services. The goal is to give back to a user the control on her information, allowing her to formulate queries such as "What kind of interaction did I have recently with Alice B.?", "Where were my last ten business trips, and who helped me plan them?". The system has to orchestrate queries to the various services (which means knowing the existence of these services, and how to interact with them), integrate information from them (which means having data models for this information and its representation in the services), e.g., align a GPS location of the user to a business address or place mentioned in an email, or an event in a calendar to some event in a Web search. This information must be accessed intensionally: for instance, costly information extraction tools should only be run on emails which seem relevant, perhaps identified by a less costly cursory analysis (this means, in turn, obtaining a cost model for access to the different services). Impacted people can be found by examining events in the user's calendar and determining who is likely to attend them, perhaps based on email exchanges or former events' participant lists. Of course, uncertainty has to be maintained along the entire process, and provenance information is needed to explain query results to the user (e.g., indicate which meetings and trips are relevant to each person of the output). Knowledge about services, their data models, their costs, need either to be provided by the system designer, or to be automatically learned from interaction with these services, as in [71].

One motivation for that choice is that Pims concentrate many of the problems we intend to investigate: heterogeneity (various sources, each with a different structure), massive distribution (information spread out over the Web, in numerous sources), rapid evolution (new data regularly added), intensionality (knowledge from Wikidata, OpenStreetMap...), confidentiality and security (mostly private data), and uncertainty (very variable quality). Though the data is distributed, its size is relatively modest; other applications may be considered for works focusing on processing data at large scale, which is a potential research direction within Valda, though not our main focus. Another strong motivation for the choice of Pims as application domain is the importance of this application from a societal viewpoint.

A Pims is essentially a system built on top of a user's *personal knowledge base*; such knowledge bases are reminiscent of those found in the Semantic Web, e.g., linked open data. Some issues, such as ontology alignment [74] exist in both scenarios. However, there are some fundamental differences in building personal knowledge bases vs collecting information from the Semantic Web: first, the scope is quite smaller, as one is only interested in knowledge related to a given individual; second, a small proportion of the data is already present in the form of semantic information, most needs to be extracted and annotated through appropriate wrappers and enrichers; third, though the linked open data is meant to be read-only, the only update possible to a user being adding new triples, a personal knowledge base is very much something that a user needs to be able to edit, and propagating updates from the knowledge base to original data sources is a challenge in itself.

4.2 Web Data

The choice of Pims is not exclusive. We also consider other application areas as well. In particular, we have worked in the past and have a strong expertise on Web data [40] in a broad sense: semi-structured, structured, or unstructured content extracted from Web databases [71]; knowledge bases from the Semantic Web [74]; social networks [68]; Web archives and Web crawls [56]; Web applications and deep Web databases [49]; crowdsourcing platforms [43]. We intend to continue using Web data as a natural application domain for the research within Valda when relevant. For instance [47], deep Web databases are a natural application scenario for intensional data management issues: determining if a deep Web database contains some information requires optimizing the number of costly requests to that database.

A common aspect of both personal information and Web data is that their exploitation raises ethical considerations. Thus, a user needs to remain fully in control of the usage that is made of her personal information; a search engine or recommender system that ranks Web content for display to a specific user needs to do so in an unbiased, justifiable, manner. These ethical constraints sometimes forbid some technically solutions that may be technically useful, such as sharing a model learned from the personal data of a user to another user, or using blackboxes to rank query result. We fully intend to consider these ethical considerations within Valda. One of the main goals of a Pims is indeed to empower the user with a full control on the use of this data.

5 Highlights of the year

5.1 Awards

- Camille Bourgaux and Michaël Thomazo received the Ray Reiter Best Paper Award for their KR 2021 paper [20].
- Victor Vianu has been invited to give a *Gems of PODS* talk on Datalog at PODS 2021. [28]

5.2 Others

Pierre Senellart was elected president of section 6 (foundations of computer science, computing, algorithms, representations, exploitations) of the French national committee for scientific research (CoNRS).

6 New software and platforms

6.1 New software

6.1.1 ProvSQL

Keywords: Databases, Provenance, Probability

Functional Description: The goal of the ProvSQL project is to add support for (m-)semiring provenance and uncertainty management to PostgreSQL databases, in the form of a PostgreSQL extension/module/plugin.

News of the Year: Merging of the in-memory storage of the provenance circuit in the main branch, including data persistence when restarting the PostgreSQL server. Support for PostgreSQL 14. Support for new version of d4. Miscellaneous enhancements and bug fixes.

URL: `https://github.com/PierreSenellart/provsql`

Publications: hal-01672566, hal-01851538

Contact: Pierre Senellart

Participants: Pierre Senellart, Silviu Maniu, Yann Ramusat

6.1.2 apxproof

Keyword: LaTeX

Functional Description: apxproof is a LaTeX package facilitating the typesetting of research articles with proofs in appendix, a common practice in database theory and theoretical computer science in general. The appendix material is written in the LaTeX code along with the main text which it naturally complements, and it is automatically deferred. The package can automatically send proofs to the appendix, can repeat in the appendix the theorem environments stated in the main text, can section the appendix automatically based on the sectioning of the main text, and supports a separate bibliography for the appendix material.

Release Contributions: Compatibility fixes with xypic, fancyvrb, memoir, natbib

News of the Year: 1.2.2 release: compatibility with AMS document classes, avoid creating useless appendix section when files are included, fix handling of optional arguments of repeated theorems containing optional arguments

1.2.3 release: compatibility with tocbibind, forwardlinking option

URL: `https://github.com/PierreSenellart/apxproof`

Contact: Pierre Senellart

Participant: Pierre Senellart

6.1.3 TheoremKB

Keyword: Information extraction

Functional Description: TheoremKB is a collection of tools to extract semantic information from (mathematical) research articles.

News of the Year: Add computer-vision based object detection. Add NLP-based techniques such as transformers and LSTM networks for sequence prediction.

URL: `https://github.com/PierreSenellart/theoremkb`

Publications: hal-02956526, hal-02940819, hal-03293643

Contact: Pierre Senellart

Participants: Pierre Senellart, Theo Delemazure, Lucas Pluvinage, Shrey Mishra

7 New results

We present the results we obtained and published in 2021 in three areas: statistical machine learning; knowledge representation; data management and data science.

7.1 Statistical aspects of machine learning

Our research on statistics and machine learning dealt with several objects of interest: dimensionality reduction, bandits, auctions, and privacy.

Dimensionality reduction

Non-negative matrix factorization (NMF) has become a well-established class of methods for the analysis of non-negative data. In particular, a lot of effort has been devoted to probabilistic NMF, namely estimation or inference tasks in probabilistic models describing the data, based for example on Poisson or exponential likelihoods. When dealing with time series data, several works have proposed to model the evolution of the activation coefficients as a non-negative Markov chain, most of the time in relation with the Gamma distribution, giving rise to so-called temporal NMF models. In [14], we review four Gamma Markov chains of the NMF literature, and show that they all share the same drawback: the absence of a well-defined station- ary distribution. We then introduce a fifth process, an overlooked model of the time series literature named BGAR(1), which overcomes this limitation. These temporal NMF models are then compared in a MAP framework on a prediction task, in the context of the Poisson likelihood.

Bandits

Motivated by A/B/n testing applications, we consider in [27] a finite set of distributions (called *arms*), one of which is treated as a *control*. We assume that the population is stratified into homogeneous subpopulations. At every time step, a subpopulation is sampled and an arm is chosen: the resulting observation is an independent draw from the arm conditioned on the subpopulation. The quality of each arm is assessed through a weighted combination of its subpopulation means. We propose a strategy for sequentially choosing one arm per time step so as to discover as fast as possible which arms, if any, have higher weighted expectation than the control. This strategy is shown to be asymptotically optimal in the following sense: if τ_δ is the first time when the strategy ensures that it is able to output the correct answer with probability at least $1-\delta$, then $E[\tau_\delta]$ grows linearly with $\log(1/\delta)$ at the exact optimal rate. This rate is identified in three different settings: (1) when the experimenter does not observe the subpopulation information, (2) when the subpopulation of each sample is observed but not chosen, and (3) when the experimenter can select the subpopulation from which each response is sampled. We illustrate the efficiency of the proposed strategy with numerical simulations on synthetic and real data collected from an A/B/n experiment. Contextual sequential decision problems with categorical or numerical observations are ubiquitous and Generalized Linear Bandits (GLB) offer a solid theoretical framework to address them. In contrast to the case of linear bandits, existing algorithms for GLB have two drawbacks undermining their applicability. First, they rely on excessively pessimistic concentration bounds due to the non-linear nature of the model. Second, they require either non-convex projection steps or burn-in phases to enforce boundedness of the estimators. Both of these issues are worsened when considering non-stationary models, in which the GLB parameter may vary with time. In [26], we focus on self-concordant GLB (which include logistic and Poisson regression) with forgetting achieved either by the use of a sliding window or exponential weights. We propose a novel confidence-based algorithm for the maximum-likehood estimator with forgetting and analyze its perfomance in abruptly changing environments. These results as well as the accompanying numerical simulations highlight the potential of the proposed approach to address non-stationarity in GLB.

There has been a recent surge of interest in nonparametric bandit algorithms based on subsampling. One drawback however of these approaches is the additional complexity required by random subsampling and the storage of the full history of rewards. Our first contribution in [18] is to show that a simple deterministic subsampling rule, proposed in recent work of under the name of “last-block subsampling”, is asymptotically optimal in one-parameter exponential families. In addition, we prove that these guarantees also hold when limiting the algorithm memory to a polylogarithmic function of the time horizon. These findings open up new perspectives, in particular for non-stationary scenarios in which the arm distributions evolve over time. We propose a variant of the algorithm in which only the most recent observations are used for subsampling, achieving optimal regret guarantees under the assumption of a known number of abrupt changes. Extensive numerical simulations highlight the merits of this approach, particularly when the changes are not only affecting the means of the rewards.

Auctions

First-price auctions have largely replaced traditional bidding approaches based on Vickrey auctions in programmatic advertising. As far as learning is concerned, first-price auctions are more challenging because the optimal bidding strategy does not only depend on the value of the item but also requires some knowledge of the other bids. They have already given rise to several works in sequential learning, many of which consider models for which the value of the buyer or the opponents' maximal bid is chosen in an adversarial manner. Even in the simplest settings, this gives rise to algorithms whose regret grows as $\sqrt{T}$ with respect to the time horizon T. Focusing on the case where the buyer plays against a stationary stochastic environment, we show in [16] how to achieve significantly lower regret: when the opponents' maximal bid distribution is known we provide an algorithm whose regret can be as low as $\log_2(T)$; in the case where the distribution must be learnt sequentially, a generalization of this algorithm can achieve $T^{1/3+\epsilon}$ regret, for any $\epsilon > 0$. To obtain these results, we introduce two novel ideas that can be of interest in their own right. First, by transposing results obtained in the posted price setting, we provide conditions under which the first-price biding utility is locally quadratic around its optimum. Second, we leverage the observation that, on small sub-intervals, the concentration of the variations of the empirical distribution function may be controlled more accurately than by using the classical Dvoretzky–Kiefer–Wolfowitz inequality. Numerical simulations confirm that our algorithms converge much faster than alternatives proposed in the literature for various bid distributions, including for bids collected on an actual programmatic advertising platform.

Developing efficient sequential bidding strategies for repeated auctions is an important practical challenge in various marketing tasks. In this setting, the bidding agent obtains information, on both the value of the item at sale and the behavior of the other bidders, only when she wins the auction. Standard bandit theory does not apply to this problem due to the presence of action-dependent censoring. In [15], we consider second-price auctions and propose novel, efficient UCB-like algorithms for this task. These algorithms are analyzed in the stochastic setting, assuming regularity of the distribution of the opponents' bids. We provide regret upper bounds that quantify the improvement over the baseline algorithm proposed in the literature. The improvement is particularly significant in cases when the value of the auctioned item is low, yielding a spectacular reduction in the order of the worst-case regret. We further provide the first parametric lower bound for this problem that applies to generic UCB-like strategies. As an alternative, we propose more explainable strategies which are reminiscent of the Explore Then Commit bandit algorithm. We provide a critical analysis of this class of strategies, showing both important advantages and limitations. In particular, we provide a minimax lower bound and propose a nearly minimax-optimal instance of this class.

Privacy

The calibration of noise for a privacy-preserving mechanism depends on the sensitivity of the query and the prescribed privacy level. A data steward must make the non-trivial choice of a privacy level that balances the requirements of users and the monetary constraints of the business entity. In [13], firstly, we analyze roles of the sources of randomness, namely the explicit randomness induced by the noise distribution and the implicit randomness induced by the data-generation distribution, that are involved in the design of a privacy-preserving mechanism. The finer analysis enables us to provide stronger privacy guarantees with quantifiable risks. Thus, we propose privacy at risk that is a probabilistic calibration of privacy-preserving mechanisms. We provide a composition theorem that leverages privacy at risk. We instantiate the probabilistic calibration for the Laplace mechanism by providing analytical results. Secondly, we propose a cost model that bridges the gap between the privacy level and the compensation budget estimated by a GDPR compliant business entity. The convexity of the proposed cost model leads to a unique fine-tuning of privacy level that minimizes the compensation budget. We show its effectiveness by illustrating a realistic scenario that avoids overestimation of the compensation budget by using privacy at risk for the Laplace mechanism. We quantitatively show that composition using the cost optimal privacy at risk provides stronger privacy guarantee than the classical advanced composition. Although the illustration is specific to the chosen cost model, it naturally extends to any convex cost model. We also provide realistic illustrations of how a data steward uses privacy at risk to balance the trade-off between utility and privacy.

7.2 Knowledge representation

A large part research on knowledge representation is motivated by ontology-mediated query answering, and revolves around the study of existential rules.

Ontology-mediated query answering (OMQA) employs structured knowledge and automated reasoning in order to facilitate access to incomplete and possibly heterogeneous data. While most research on OMQA adopts (unions of) conjunctive queries as the query language, there has been recent interest in handling queries that involve counting. In [19], we advance this line of research by investigating cardinality queries (which correspond to Boolean atomic counting queries) coupled with DL-Lite ontologies. Despite its apparent simplicity, we show that such an OMQA setting gives rise to rich and complex behaviour. While we prove that cardinality query answering is tractable (TC^0) in data complexity when the ontology is formulated in $\mathrm{DL}-\mathrm{Lite}_{core}$, the problem becomes coNP-hard as soon as role inclusions are allowed. For $\mathrm{DL}-\mathrm{Lite}^{\mathcal{H}}_{pos}$ (which allows only positive axioms), we establish a P-coNP dichotomy and pinpoint the TC^0 cases; for $\mathrm{DL}-\mathrm{Lite}^{\mathcal{H}}_{core}$ (allowing also negative axioms), we identify new sources of coNP complexity and also exhibit L-complete cases. Interestingly, and in contrast to related tractability results, we observe that the canonical model may not give the optimal count value in the tractable cases, which led us to develop an entirely new approach based upon exploring a space of strategies to determine the minimum possible number of query matches.

Existential rules are a very popular ontology-mediated query language for which the chase represents a generic computational approach for query answering. It is straightforward that existential rule queries exhibiting chase termination are decidable and can only recognize properties that are preserved under homomorphisms. In [20], we show the converse: every decidable query that is closed under homomorphism can be expressed by an existential rule set for which the standard chase universally terminates. Membership in this fragment is not decidable, but we show via a diagonalisation argument that this is unavoidable.

In [22], we consider again existential rules. The chase is a fundamental tool to do reasoning with existential rules as it computes all the facts entailed by the rules from a database instance. We introduce parallelisable sets of existential rules, for which the chase can be computed in a single breadth-first step from any instance. The question we investigate is the characterization of such rule sets. We show that parallelisable rule sets are exactly those rule sets both bounded for the chase and belonging to a novel class of rules, called pieceful. The pieceful class includes in particular frontier-guarded existential rules and (plain) datalog. We also give another characterization of parallelisable rule sets in terms of rule composition based on rewriting.

In the search for knowledge graph embeddings that could capture ontological knowledge, geometric models of existential rules have been recently introduced. It has been shown that convex geometric regions capture the so-called quasi-chained rules. Attributed description logics (DL) have been defined to bridge the gap between DL languages and knowledge graphs, whose facts often come with various kinds of annotations that may need to be taken into account for reasoning. In particular, temporally attributed DLs are enriched by specific attributes whose semantics allows for some temporal reasoning. Considering that geometric models and (temporally) attributed DLs are promising tools designed for knowledge graphs, [21] investigates their compatibility, focusing on the attributed version of a Horn dialect of the DL-Lite family. We first adapt the definition of geometric models to attributed DLs and show that every satisfiable ontology has a convex geometric model. Our second contribution is a study of the impact of temporal attributes. We show that a temporally attributed DL may not have a convex geometric model in general but we can recover geometric satisfiability by imposing some restrictions on the use of the temporal attributes.

7.3 Data management and data science

We studied different aspects of data management and data science: data provenance, data streams, information extraction, as well as connection with sociological studies.

Data provenance

In [25], we investigate the efficient computation of the provenance of rich queries over graph databases. We show that semiring-based provenance annotations enrich the expressiveness of routing queries over graphs. Several algorithms have previously been proposed for provenance computation over graphs, each yielding a trade-off between time complexity and generality. Here, we address the limitations of these algorithms and propose a new one, partially bridging a complexity and expressiveness gap and adding to the algorithmic toolkit for solving this problem. Importantly, we provide a comprehensive taxonomy of semirings and corresponding algorithms, establishing which practical approaches are needed in different cases. We implement and comprehensively evaluate several practical applications of the problem (e.g., shortest distances, top-shortest distances, Boolean or integer path features), each corresponding to a specific semiring and algorithm, that depends on the properties of the semiring. On several real-world and synthetic graph datasets, we show that the algorithms we propose exhibit large practical benefits for processing rich graph queries.

Data streams

Mining high-dimensional data streams poses a fundamental challenge to machine learning as the presence of high numbers of attributes can remarkably degrade any mining task's performance. In the past several years, dimension reduction (DR) approaches have been successfully applied for different purposes (e.g., visualization). Due to their high-computational costs and numerous passes overlarge data, these approaches pose a hindrance when processing infinite data streams that are potentially high-dimensional. The latter increases the resource-usage of algorithms that could suffer from the curse of dimensionality. To cope with these issues, some techniques for incremental DR have been proposed. In [17], we provide a survey on reduction approaches designed to handle data streams and highlight the key benefits of using these approaches for stream mining algorithms.

Information extraction

We propose in [24] a new grammar-based language for defining information-extractors from documents (text) that is built upon the well-studied framework of document spanners for extracting structured data from text. While previously studied formalisms for document spanners are mainly based on regular expressions, we use an extension of context-free grammars, called extraction grammars, to define the new class of context-free spanners. Extraction grammars are simply context-free grammars extended with variables that capture interval positions of the document, namely spans. While regular expressions are efficient for tokenizing and tagging, context-free grammars are also efficient for capturing structural properties. Indeed, we show that context-free spanners are strictly more expressive than their regular counterparts. We reason about the expressive power of our new class and present a pushdown-automata model that captures it. We show that extraction grammars can be evaluated with polynomial data complexity. Nevertheless, as the degree of the polynomial depends on the query, we present an enumeration algorithm for unambiguous extraction grammars that, after quintic preprocessing, outputs the results sequentially, without repetitions, with a constant delay between every two consecutive ones.

Scholarly articles in mathematical fields often feature mathematical statements (theorems, propositions, etc.) and their proofs. In [23], we present preliminary work for extracting such information from PDF documents, with several types of approaches: vision (using YOLO), natural language (with transformers), and styling information (with linear conditional random fields). Our main task is to identify which parts of the paper to label as theorem-like environments and proofs. We rely on a dataset collected from arXiv, with LaTeX sources of research articles used to train the models

Sociological aspects of data science

Technology characterizes and facilitates our daily lives, but its pervasive use can result in the introduction or the exacerbation of social problems. Because of their intrinsic complexity, these issues require to be addressed from different but complementary perspectives, which are provided to us by two disciplines of very different nature: data science and sociology. Specifically, Riccardo Corona's Master thesis [33] would like to be a bridge between the technical field of data analysis and a specific category of social problems, namely that of discrimination, and, in particular, gender discrimination. To move within this context, we use an approach that has data analysis as its starting point, and which finds in sociology a useful supporting instrument, as well as a source of requirements. We investigate in depth the sociological reasons behind gender discrimination in the specific society of our interest – the American one – introducing and exploring what is commonly referred as 'gender gap', and we carry out several experiments on data related to U.S. employees, focusing on the economic perspective (gender pay gap) but taking into account the different other facets of the problem. The main contributions of this thesis derive from the application of preprocessing techniques and the use of tools created with the aim of detecting bias in data, with which we try to understand which design choices have the greatest impact on the so-called 'fairness' of the results, and of which we highlight strengths and weaknesses, emphasizing the importance of a multidisciplinary approach to problems of this kind, that is essential to obtain information on the complex context in which data are embedded.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Numberly:

Participants: Olivier Cappé, Juliette Achddou.

- Duration: 2019–2022
- Local coordinator: Olivier Cappé
- Juliette Achddou's PhD research is set up as a CIFRE contract and supervision agreement between her employer, the Numberly company, and École normale supérieure.

Neo4j:

Participants: Leonid Libkin, Liat Peterfreund, Alexandra Rogova.

- Duration: 2020–2021
- Local coordinator: Leonid Libkin
- A contract has been established with Neo4j, the leading company in the field of graph databases, to work towards the creation of a new standard for graph languages called GQL, building on Neo4j's Cypher query language. Leonid Libkin is chairing a working group on the formal semantics of GQL. In addition to Valda, it involves researchers from Edinburgh, Santiago, Warsaw, and other universities in Paris (UPEM, Université de Paris). This project is supported by a grant from Neo4j. Leonid Libkin is also a scientific advisor of Neo4j.

8.2 Standardization activities

Leonid Libkin is involved in the standardization process of the GQL and SQL query languages. In particular, he is a chair of the LDBC working group on semantics of GQL, and a member of ISO/IEC JTC1 SC32 WG3 (SQL committee).

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Participation in other International Programs

DesCartes (2021–2026) is a project managed by CNRS@CREATE, a CNRS subsidiary in Singapore and funded by Singapore's National Research Foundation, with 50 million total budget. Pierre Senellart is involved in the project as one of the French PIs. See `https://www.cnrsatcreate.cnrs.fr/descartes/`.

9.1.2 Informal international partners

Valda has strong collaborations with the following international groups:

Univ. Edinburgh, United Kingdom: Paolo Guagliardo, Andreas Pieris

Univ. Oxford, United Kingdom: Michael Benedikt and Georg Gottlob

TU Dresden, Germany: Markus Krötzsch and Sebastian Rudolph

Dortmund University, Germany: Thomas Schwentick

Bayreuth University, Germany: Wim Martens

Univ. Bergen, Norway: Ana Ozaki

Univ. Roma La Sapienza, Italy: Marco Console

Warsaw University, Poland: Mikołaj Bojańczyk and Szymon Toruńczyk

Tel Aviv University, Israel: Daniel Deutch and Tova Milo

NYU, USA: Julia Stoyanovich

Univ. California San Diego, USA: Victor Vianu

Pontifical Catholic University of Chile: Marcelo Arenas, Pablo Barceló

National University of Singapore: Stéphane Bressan

9.2 International research visitors

9.2.1 Visits of international scientists

Other international visits to the team Victor Vianu, Professor at UCSD, visited the group during several months in 2021. He was also hired on an Inria Advanced Research Position.

9.3 European initiatives

9.3.1 Other european programs/initiatives

A bilateral French–German ANR project, entitled EQUUS – Efficient Query answering Under UpdateS has started in 2020. It involves CNRS (CRIL, CRIStAL, IMJ), Télécom Paris, HU Berlin, and Bayreuth University, in addition to Inria Valda.

9.4 National initiatives

9.4.1 ANR

Valda has been part of four national ANR projects in 2021:

HEADWORK (2016–2021; 38 k€ for Valda, budget managed by Inria), together with IRISA (Druid, coordinator), Inria Lille (Links & Spirals), and Inria Rennes (Sumo), and two application partners: MNHN (Cesco) and FouleFactory. The topic is workflows for crowdsourcing. See http://headwork.gforge.inria.fr/.

BioQOP (2017–2021; 66 k€ for Valda, budget managed by ENS), with Idemia (coordinator) and GREYC, on the optimization of queries for privacy-aware biometric data management. See http://bioqop.di.ens.fr/.

CQFD (2018–2022; 19 k€ for Valda, budget managed by Inria), with Inria Sophia (GraphIK, coordinator), LaBRI, LIG, Inria Saclay (Cedar), IRISA, Inria Lille (Spirals), and Télécom ParisTech, on complex ontological queries over federated and heterogeneous data. See http://www.lirmm.fr/cqfd/.

QUID (2018–2022; 49 k€ for Valda, budget managed by Inria), LIGM (coordinator), IRIF, and LaBRI, on incomplete and inconsistent data. See https://quid.labri.fr/home.html

Camille Bourgaux has been participating in the AI Chair of Meghyn Bienvenu on *INTENDED (Intelligent handling of imperfect data)* since 2020.

9.4.2 Others

Dissemin (2021–2024; 124€ for Valda, budget managed by ENS), sole partner, on the development of the https://dissem.in/ platform for open science promotion. Funded by the Fonds National Science Ouverte.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- Leonid Libkin, general chair of PODS 2021 and chair of the PODS Executive Committee
- Luc Segoufin, chair (until September 2021) and member of the steering committee of the conference series Highlights of Logic, Games and Automata

Member of the organizing committees

- Camille Bourgaux & Michaël Thomazo, organizers of the in-person day at BDA 2021
- Nofar Carmeli, judge at the ICPC (International Collegiate Programming Contest) Southwestern Europe 2020-2021 competition.
- Leonid Libkin, member of the LICS Steering Committee.
- Leonid Libkin, member of the SIGMOD Executive Committee.
- Pierre Senellart, member of the steering committee of BDA, the French scientific community on data management.
- Pierre Senellart, co-organizer and secretary of the ICPC (International Collegiate Programming Contest) Southwestern Europe 2020-2021 competition.

10.1.2 Scientific events: selection

Chair of conference program committees

- Leonid Libkin, PC chair of LICS 2021

Member of the conference program committees

- Camille Bourgaux, AAAI 2022, IJCAI 2021, KR 2021
- Nofar Carmeli, ICDT 2021
- Pierre Senellart, PODS 2021, ICDT 2021 Test-of-Time Award committee
- Michaël Thomazo, AAAI 2022, IJCAI 2021, KR 2021

10.1.3 Journal

Member of the editorial boards

- Olivier Cappé, *Annals of the Institute of Statistical Mathematics*
- Leonid Libkin, *Bulletin of Symbolic Logic*
- Leonid Libkin, *Acta Informatica*
- Leonid Libkin, *RAIRO Theoretical Informatics and Applications*
- Leonid Libkin, *Journal of Applied Logic*
- Leonid Libkin, *SN Computer Science*
- Luc Segoufin, *ACM Transactions on Computational Logics*

10.1.4 Leadership within the scientific community

- Serge Abiteboul is a member of the French Academy of Sciences, of the Academia Europaea, of the scientific council of the Société Informatique de France, and an ACM Fellow.
- Leonid Libkin is a Fellow of the Royal Society of Edinburgh, a member of the Academia Europaea, of the UK Computing research committee, and an ACM Fellow.
- Pierre Senellart is a junior member of the Institut Universitaire de France.

10.1.5 Research administration

- Olivier Cappé is a scientific deputy director of CNRS division of Information Sciences and Technologies (INS2I).
- Luc Segoufin is a member of the CNHSCT of Inria.
- Pierre Senellart was a member of the board (till August 2021) and is now president (since September 2021) of section 6 of the National Committee for Scientific Research.
- Pierre Senellart is a member of the board of the conference of presidents of the national committee (CPCN) and as such a member of the coordination of managing parties of the national committee (C3N) Site d'information sous l'autorité du porte-parole de la C3N
- Pierre Senellart is deputy director of the DI ENS laboratory, joint between ENS, CNRS, and Inria.
- Pierre Senellart is a member of the board of the DIM RFSI (Réseau Francilien en Sciences Informatiques).

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Licence: *Algorithms*, 42 heqTD, L2, CPES, PSL – Pierre Senellart
- Licence: *Databases*, 74 heqTD, L3, École normale supérieure – Leonid Libkin, Yann Ramusat, Michaël Thomazo
- Master: *Data wrangling, Data privacy*, 36 heqTD, M2, IASD – Leonid Libkin, Pierre Senellart
- Master: *Anonymization, privacy*, 36 heqTD, M2, IASD – Ashish Dandekar, Pierre Senellart
- Master: *Knowledge graphs, description logics, reasoning on data*, 36 heqTD, M2, IASD – Camille Bourgaux, Michaël Thomazo

Pierre Senellart has had various teaching responsibilities (L3 internships, M1 projects, M2 administration, entrance competition) at ENS. Leonid Libkin was responsible of the graduate program in computer science of PSL University until 2021; Leonid Libkin and Pierre Senellart are in the managing board of the graduate program. Leonid Libkin is co-responsible of the international entrance competition at ENS. Yann Ramusat was the secretary of the entrance competition at ENS for computer science. Most members of the group are also involved in tutoring ENS students, advising them on their curriculum, their internships, etc. They are also occasionally involved with reviewing internship reports, supervising student projects, etc.

10.2.2 Supervision

- PhD in progess: Juliette Achddou, Application of reinforcement learning strategies to the context of Real-Time Bidding, started in September 2018, Olivier Cappé & Aurélien Garivier; Juliette and Olivier left the group in September 2021
- PhD in progress: Anatole Dahan, Logical foundations of the polynomial hierarchy, started in October 2020, Arnaud Durand & Luc Segoufin
- PhD in progress: Baptiste Lafosse, Compiler dedicated to the evaluation of SQL queries, started in October 2021, Pierre Senellart & Jean-Marie Lagniez
- PhD in progress: Shrey Mishra, Towards a knowledge base of mathematic results, started in January 2021, Pierre Senellart
- PhD in progress: Yann Ramusat, Provenance-based routing in probabilistic graphs, started in September 2018, Silviu Maniu & Pierre Senellart
- PhD in progress: Alexandra Rogova, Query analytics in Cypher, started October 2021, Amelie Gheerbrant & Leonid Libkin
- PhD in progess: Yoan Russac, Sequential methods for robust decision making, started in December 2018, Olivier Cappé; Yoan and Olivier left the group in September 2021

10.2.3 Juries

- HdR: Stefan Mengel [reviewer], Université d'Artois,Pierre Senellart
- HdR: Federico Ulliana [reviewer], Université de Montpellier,Pierre Senellart
- PhD: Amit Kumar [president], Université Caen Normandie, Pierre Senellart
- PhD: Valentin Iovene [reviewer], Université Paris-Saclay, Pierre Senellart
- PhD: Wissam Kouadri Maamar [reviewer], Université de Paris, Pierre Senellart

10.3 Popularization

10.3.1 Internal or external Inria responsibilities

- Serge Abiteboul is a member of the strategic committee of the Blaise Pascal foundation for scientific mediation.
- Pierre Senellart is a scientific expert advising the Scientific and Ethical Committee of Parcoursup, the platform for the selection of first-year higher education students.

10.3.2 Articles and contents

- Serge Abiteboul is a founding editor of the binaire blog for popularizing computer science. See `https://www.lemonde.fr/blog/binaire/`.
- Serge Abiteboul has written two science popularization articles: on digital identifiers [11] and on transparency and explainability of algorithmic decisions [12].

11 Scientific production

11.1 Major publications

[1] S. Abiteboul, P. Bourhis and V. Vianu. 'Explanations and Transparency in Collaborative Workflows'. In: *PODS 2018 - 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles Of Database Systems*. Houston, Texas, United States, June 2018. URL: https://hal.inria.fr/hal-01744978.

[2] M. Benedikt, P. Bourhis, G. Gottlob and P. Senellart. 'Monadic Datalog, Tree Validity, and Limited Access Containment'. In: *ACM Transactions on Computational Logic* 21.1 (2020), 6:1–6:45. DOI: 10.1145/3344514. URL: https://hal.inria.fr/hal-02307999.

[3] M. Bienvenu, Q. Manière and M. Thomazo. 'Answering Counting Queries over DL-Lite Ontologies'. In: *IJCAI 2020 - Twenty-Ninth International Joint Conference on Artificial Intelligence*. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. Reportée de juillet 2020 à janvier 2021 en raison de la COVID. Yokohama, Japan, July 2020. URL: https://hal.inria.fr/hal-02927913.

[4] C. Bourgaux, D. Carral, M. Krötzsch, S. Rudolph and M. Thomazo. 'Capturing Homomorphism-Closed Decidable Queries with Existential Rules'. In: KR 2021 - 18th International Conference on Principles of Knowledge Representation and Reasoning. Virtual, Vietnam, 3rd Nov. 2021, pp. 141–150. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03345614.

[5] C. Bourgaux, A. Ozaki, R. Peñaloza and L. Predoiu. 'Provenance for the Description Logic ELHr'. In: *IJCAI-PRICAI-20 - Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence*. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. Reportée de juillet 2020 à janvier 2021 en raison de la COVID. Yokohama, Japan, July 2020, pp. 1862–1869. DOI: 10.24963/ijcai.2020/258. URL: https://hal.archives-ouvertes.fr/hal-02899464.

[6] M. Console, P. Guagliardo, L. Libkin and E. Toussaint. 'Coping with Incomplete Data: Recent Advances'. In: *SIGMOD/PODS 2020 - International Conference on Management of Data*. Portland / Virtual, United States: ACM, June 2020, pp. 33–47. DOI: 10.1145/3375395.3387970. URL: https://hal.inria.fr/hal-03127726.

[7] W. Kazana and L. Segoufin. 'First-order queries on classes of structures with bounded expansion'. In: *Logical Methods in Computer Science* 16.1 (2020). URL: https://hal.inria.fr/hal-01706665.

[8] L. Peterfreund. 'Grammars for Document Spanners'. In: ICDT 2021 - 24th International Conference on Extending Database Technology. Nicosia / Virtual, Cyprus, 23rd Mar. 2021. URL: https://hal.inria.fr/hal-03104144.

[9] N. Schweikardt, L. Segoufin and A. Vigny. 'Enumeration for FO Queries over Nowhere Dense Graphs'. In: *PODS 2018 - Principles Of Database Systems*. Houston, United States, June 2018. URL: https://hal.inria.fr/hal-01895786.

[10] P. Senellart, L. Jachiet, S. Maniu and Y. Ramusat. 'ProvSQL: Provenance and Probability Management in PostgreSQL'. In: *Proceedings of the VLDB Endowment (PVLDB)* 11.12 (Aug. 2018), pp. 2034–2037. DOI: 10.14778/3229863.3236253. URL: https://hal.inria.fr/hal-01851538.

11.2 Publications of the year

International journals

[11] S. Abiteboul. 'Dans la jungle des identifiants numériques'. In: *Acteurs Publics* (2021). URL: https://hal.inria.fr/hal-03172025.

[12] S. Abiteboul. 'Qualité, équité, transparence, vérification et explicabilité des décisions algorithmiques'. In: *Annales des Mines - Enjeux Numériques* (Mar. 2021). URL: https://hal.inria.fr/hal-03117322.

[13] A. Dandekar, D. Basu and S. Bressan. 'Differential Privacy at Risk: Bridging Randomness and Privacy Budget'. In: *Proceedings on Privacy Enhancing Technologies* 2021.1 (2021), pp. 64–84. DOI: 10.2478/popets-2021-0005. URL: https://hal.inria.fr/hal-02942997.

[14] L. Filstroff, O. Gouvert, C. Févotte and O. Cappé. 'A Comparative Study of Gamma Markov Chains for Temporal Non-Negative Factorization'. In: *IEEE Transactions on Signal Processing* (19th Feb. 2021). DOI: 10.1109/TSP.2021.3060000. URL: https://hal.archives-ouvertes.fr/hal-02883800.

International peer-reviewed conferences

[15] J. Achddou, O. Cappé and A. Garivier. 'Efficient Algorithms for Stochastic Repeated Second-price Auctions'. In: ALT 2021. Paris, France, 16th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-02997579.

[16] J. Achddou, O. Cappé and A. Garivier. 'Fast Rate Learning in Stochastic First Price Bidding'. In: ACML 2021 - Proceedings of Machine Learning Research 157, 2021. SIngapore, Singapore, 17th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03277164.

[17] M. Bahri, A. Bifet, S. Maniu and H. M. Gomes. 'Survey on Feature Transformation Techniques for Data Streams'. In: IJCAI-PRICAI 2020 - 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence. Yokohama / Virtual, Japan, 7th Jan. 2021, pp. 4796–4802. DOI: 10.24963/ijcai.2020/668. URL: https://hal.archives-ouvertes.fr/hal-03189968.

[18] D. Baudry, Y. Russac and O. Cappé. 'On Limited-Memory Subsampling Strategies for Bandits'. In: ICML 2021- International Conference on Machine Learning. Vienna / Virtual, Austria, 18th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03265442.

[19] M. Bienvenu, Q. Manière and M. Thomazo. 'Cardinality Queries over DL-Lite Ontologies'. In: IJCAI 2021 - 30th International Joint Conference on Artificial Intelligence. Montreal, Canada, 19th Aug. 2021. URL: https://hal.inria.fr/hal-03405769.

[20] C. Bourgaux, D. Carral, M. Krötzsch, S. Rudolph and M. Thomazo. 'Capturing Homomorphism-Closed Decidable Queries with Existential Rules'. In: KR 2021 - 18th International Conference on Principles of Knowledge Representation and Reasoning. Virtual, Vietnam, 3rd Nov. 2021, pp. 141–150. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03345614.

[21] C. Bourgaux, A. Ozaki and J. Z. Pan. 'Geometric Models for (Temporally) Attributed Description Logics'. In: DL 2021 - 34th International Workshop on Description Logics. DL 2021 - 34th International Workshop on Description Logics. Bratislava, Slovakia, 28th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03345699.

[22] M. Buron, M.-L. Mugnier and M. Thomazo. 'Parallelisable Existential Rules: a Story of Pieces'. In: KR 2021 - 18th International Conference on Principles of Knowledge Representation and Reasoning. Virtual, Vietnam, 3rd Nov. 2021. URL: https://hal.inria.fr/hal-03405745.

[23] S. Mishra, L. Pluvinage and P. Senellart. 'Towards Extraction of Theorems and Proofs in Scholarly Articles'. In: DocEng '21 - 21st ACM Symposium on Document Engineering. Limerick, Ireland, 24th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03293643.

[24] L. Peterfreund. 'Grammars for Document Spanners'. In: ICDT 2021 - 24th International Conference on Extending Database Technology. Nicosia / Virtual, Cyprus, 23rd Mar. 2021. URL: https://hal.inria.fr/hal-03104144.

[25] Y. Ramusat, S. Maniu and P. Senellart. 'Provenance-Based Algorithms for Rich Queries over Graph Databases'. In: EDBT 2021 - 24th International Conference on Extending Database Technology. Nicosia / Virtual, Cyprus, 23rd Mar. 2021. URL: https://hal.inria.fr/hal-03140067.

[26] Y. Russac, L. Faury, O. Cappé and A. Garivier. 'Self-Concordant Analysis of Generalized Linear Bandits with Forgetting'. In: AISTATS 2021 - International Conference on Artificial Intelligence and Statistics. San Diego / Virtual, United States, 13th Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-02984117.

[27] Y. Russac, C. Katsimerou, D. Bohle, O. Cappé, A. Garivier and W. M. Koolen. 'A/B/n Testing with Control in the Presence of Subpopulations'. In: NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems. Virtual, France, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03407247.

[28] V. Vianu. 'Datalog Unchained'. In: SIGMOD/PODS '21- International Conference on Management of Data. Xi'an, Shaanxi / Virtual Event China, China: ACM, 20th June 2021, pp. 57–69. DOI: 10.1145/3452021.3458815. URL: https://hal.inria.fr/hal-03381199.

Reports & preprints

[29] D. D. Freydenberger and L. Peterfreund. *The theory of concatenation over finite models*. 8th Jan. 2021. URL: https://hal.inria.fr/hal-03104159.

[30] N. Grosshans, P. McKenzie and L. Segoufin. *Tameness and the power of programs over monoids in DA*. 3rd Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03114304.

[31] L. Libkin and L. Peterfreund. *Handling SQL Nulls with Two-Valued Logic*. 8th Jan. 2021. URL: https://hal.inria.fr/hal-03104130.

[32] Y. Ramusat, S. Maniu and P. Senellart. *A Practical Dynamic Programming Approach to Datalog Provenance Computation*. 3rd Dec. 2021. URL: https://hal.inria.fr/hal-03465813.

Other scientific publications

[33] R. Corona. 'Gender Discrimination in Data Analysis: a Socio-Technical Approach'. Politecnico di Milano. Dipartimento di elettronica, informazione e bioingegneria (Milano, Italie), 7th Oct. 2021. URL: https://hal.inria.fr/hal-03374130.

11.3 Cited publications

[34] F. Jacquemard, L. Segoufin and J. Dimino. 'FO2(<, +1, ~) on data trees, data tree automata and branching vector addition systems'. In: *Logical Methods in Computer Science* 12.2 (2016). DOI: 10.2168/LMCS-12(2:3)2016. URL: https://doi.org/10.2168/LMCS-12(2:3)2016.

[35] S. Abiteboul, B. André and D. Kaplan. 'Managing your digital life'. In: *Commun. ACM* 58.5 (2015), pp. 32–35. DOI: 10.1145/2670528. URL: http://doi.acm.org/10.1145/2670528.

[36] S. Abiteboul, P. Bourhis and V. Vianu. 'Comparing workflow specification languages: A matter of views'. In: *ACM Trans. Database Syst.* 37.2 (2012), 10:1–10:59. DOI: 10.1145/2188349.2188352. URL: http://doi.acm.org/10.1145/2188349.2188352.

[37] S. Abiteboul, P. Buneman and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, 1999.

[38] S. Abiteboul, L. Herr and J. Van den Bussche. 'Temporal Versus First-Order Logic to Query Temporal Databases'. In: *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, 1996, Montreal, Canada*. 1996, pp. 49–57. DOI: 10.1145/237661.237674. URL: http://doi.acm.org/10.1145/237661.237674.

[39] S. Abiteboul, R. Hull and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995. URL: http://webdam.inria.fr/Alice/.

[40] S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset and P. Senellart. *Web Data Management*. Cambridge University Press, 2011. URL: http://webdam.inria.fr/Jorge.

[41] A. Amarilli, P. Bourhis and P. Senellart. 'Provenance Circuits for Trees and Treelike Instances'. In: *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part II*. 2015, pp. 56–68. DOI: 10.1007/978-3-662-47666-6_5. URL: https://doi.org/10.1007/978-3-662-47666-6_5.

[42] A. Amarilli, P. Bourhis and P. Senellart. 'Tractable Lineages on Treelike Instances: Limits and Extensions'. In: *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016*. 2016, pp. 355–370. DOI: 10.1145/2902251.2902301. URL: http://doi.acm.org/10.1145/2902251.2902301.

[43] Y. Amsterdamer, Y. Grossman, T. Milo and P. Senellart. 'CrowdMiner: Mining association rules from the crowd'. In: *PVLDB* 6.12 (2013), pp. 1250–1253. URL: http://www.vldb.org/pvldb/vol6/p1250-amsterdamer.pdf.

[44] P. B. Baeza. 'Querying graph databases'. In: *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013*. 2013, pp. 175–188. DOI: 10.1145/2463664.2465216. URL: http://doi.acm.org/10.1145/2463664.2465216.

[45] D. Barbará, H. Garcia-Molina and D. Porter. 'The Management of Probabilistic Data'. In: *IEEE Trans. Knowl. Data Eng.* 4.5 (1992), pp. 487–502. DOI: 10.1109/69.166990. URL: https://doi.org/10.1109/69.166990.

[46] D. Basu, Q. Lin, W. Chen, H. T. Vo, Z. Yuan, P. Senellart and S. Bressan. 'Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning'. In: *T. Large-Scale Data- and Knowledge-Centered Systems* 28 (2016), pp. 96–132. DOI: 10.1007/978-3-662-53455-7_5. URL: https://doi.org/10.1007/978-3-662-53455-7_5.

[47] M. Benedikt, G. Gottlob and P. Senellart. 'Determining relevance of accesses at runtime'. In: *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011, June 12-16, 2011, Athens, Greece*. 2011, pp. 211–222. DOI: 10.1145/1989284.1989309. URL: http://doi.acm.org/10.1145/1989284.1989309.

[48] M. Benedikt and P. Senellart. 'Databases'. In: *Computer Science, The Hardware, Software and Heart of It*. Springer, 2011, pp. 169–229. DOI: 10.1007/978-1-4614-1168-0_10. URL: https://doi.org/10.1007/978-1-4614-1168-0_10.

[49] M. Bienvenu, D. Deutch, D. Martinenghi, P. Senellart and F. M. Suchanek. 'Dealing with the Deep Web and all its Quirks'. In: *Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012*. 2012, pp. 21–24. URL: http://ceur-ws.org/Vol-884/VLDS2012_p21_Bienvenu.pdf.

[50] M. Bojańczyk, L. Segoufin and S. Toruńczyk. 'Verification of database-driven systems via amalgamation'. In: *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013*. 2013, pp. 63–74. DOI: 10.1145/2463664.2465228. URL: http://doi.acm.org/10.1145/2463664.2465228.

[51] P. Buneman, S. Khanna and W.-C. Tan. 'Why and Where: A Characterization of Data Provenance'. In: *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*. 2001, pp. 316–330. DOI: 10.1007/3-540-44503-X_20. URL: https://doi.org/10.1007/3-540-44503-X_20.

[52] B. Courcelle. 'The Monadic Second-Order Logic of Graphs. I. Recognizable Sets of Finite Graphs'. In: *Inf. Comput.* 85.1 (1990), pp. 12–75. DOI: 10.1016/0890-5401(90)90043-H. URL: https://doi.org/10.1016/0890-5401(90)90043-H.

[53] N. N. Dalvi and D. Suciu. 'The dichotomy of probabilistic inference for unions of conjunctive queries'. In: *J. ACM* 59.6 (2012), 30:1–30:87. DOI: 10.1145/2395116.2395119. URL: http://doi.acm.org/10.1145/2395116.2395119.

[54] A. Deshpande, Z. G. Ives and V. Raman. 'Adaptive Query Processing'. In: *Foundations and Trends in Databases* 1.1 (2007), pp. 1–140. DOI: 10.1561/1900000001. URL: https://doi.org/10.1561/1900000001.

[55] P. Donmez and J. G. Carbonell. 'Proactive learning: cost-sensitive active learning with multiple imperfect oracles'. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*. 2008, pp. 619–628. DOI: 10.1145/1458082.1458165. URL: http://doi.acm.org/10.1145/1458082.1458165.

[56] M. Faheem and P. Senellart. 'Adaptive Web Crawling Through Structure-Based Link Classification'. In: *Digital Libraries: Providing Quality Information - 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015, Proceedings*. 2015, pp. 39–51. DOI: 10.1007/978-3-319-27974-9_5. URL: https://doi.org/10.1007/978-3-319-27974-9_5.

[57] L. Getoor. *Introduction to statistical relational learning*. MIT Press, 2007.

[58] G. Gouriten, S. Maniu and P. Senellart. 'Scalable, generic, and adaptive systems for focused crawling'. In: *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*. 2014, pp. 35–45. DOI: 10.1145/2631775.2631795. URL: http://doi.acm.org/10.1145/2631775.2631795.

[59] T. J. Green, G. Karvounarakis and V. Tannen. 'Provenance semirings'. In: *Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*. 2007, pp. 31–40. DOI: 10.1145/1265530.1265535. URL: http://doi.acm.org/10.1145/1265530.1265535.

[60] T. J. Green and V. Tannen. 'Models for Incomplete and Probabilistic Information'. In: *IEEE Data Eng. Bull.* 29.1 (2006), pp. 17–24. URL: http://sites.computer.org/debull/A06mar/green.ps.

[61] A. Y. Halevy. 'Answering queries using views: A survey'. In: *VLDB J.* 10.4 (2001), pp. 270–294. DOI: 10.1007/s007780100054. URL: https://doi.org/10.1007/s007780100054.

[62] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf. 'Support vector machines'. In: *IEEE Intelligent Systems* 13.4 (1998), pp. 18–28. DOI: 10.1109/5254.708428. URL: https://doi.org/10.1109/5254.708428.

[63] T. Imielinski and W. Lipski Jr. 'Incomplete Information in Relational Databases'. In: *J. ACM* 31.4 (1984), pp. 761–791. DOI: 10.1145/1634.1886. URL: http://doi.acm.org/10.1145/1634.1886.

[64] B. Kimelfeld and P. Senellart. 'Probabilistic XML: Models and Complexity'. In: *Advances in Probabilistic Databases for Uncertain Information Management*. Springer, 2013, pp. 39–66. DOI: 10.1007/978-3-642-37509-5_3. URL: https://doi.org/10.1007/978-3-642-37509-5_3.

[65] A. C. Klug. 'Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions'. In: *J. ACM* 29.3 (1982), pp. 699–717. DOI: 10.1145/322326.322332. URL: http://doi.acm.org/10.1145/322326.322332.

[66] D. Kossmann. 'The State of the art in distributed query processing'. In: *ACM Comput. Surv.* 32.4 (2000), pp. 422–469. DOI: 10.1145/371578.371598. URL: http://doi.acm.org/10.1145/371578.371598.

[67] J. D. Lafferty, A. McCallum and F. C. N. Pereira. 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data'. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. 2001, pp. 282–289.

[68] S. Lei, S. Maniu, L. Mo, R. Cheng and P. Senellart. 'Online Influence Maximization'. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. 2015, pp. 645–654. DOI: 10.1145/2783258.2783271. URL: http://doi.acm.org/10.1145/2783258.2783271.

[69] F. Neven. 'Automata Theory for XML Researchers'. In: *SIGMOD Record* 31.3 (2002), pp. 39–46. DOI: 10.1145/601858.601869. URL: http://doi.acm.org/10.1145/601858.601869.

[70] M. T. Özsu and P. Valduriez. *Principles of Distributed Database Systems, Third Edition*. Springer, 2011. DOI: 10.1007/978-1-4419-8834-8. URL: https://doi.org/10.1007/978-1-4419-8834-8.

[71] P. Senellart, A. Mittal, D. Muschick, R. Gilleron and M. Tommasi. 'Automatic wrapper induction from hidden-web sources with domain knowledge'. In: *10th ACM International Workshop on Web Information and Data Management (WIDM 2008), Napa Valley, California, USA, October 30, 2008*. 2008, pp. 9–16. DOI: 10.1145/1458502.1458505. URL: http://doi.acm.org/10.1145/1458502.1458505.

[72] B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012. DOI: 10.2200/S00429ED1V01Y201207AIM018. URL: https://doi.org/10.2200/S00429ED1V01Y201207AIM018.

[73] B. Settles, M. Craven and L. Friedland. 'Active learning with real annotation costs'. In: *NIPS 2008 Workshop on Cost-Sensitive Learning*. 2008. URL: http://burrsettles.com/pub/settles.nips08ws.pdf.

[74] F. M. Suchanek, S. Abiteboul and P. Senellart. 'PARIS: Probabilistic Alignment of Relations, Instances, and Schema'. In: *PVLDB* 5.3 (2011), pp. 157–168. URL: http://www.vldb.org/pvldb/vol5/p157_fabianmsuchanek_vldb2012.pdf.

[75] D. Suciu, D. Olteanu, C. Ré and C. Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011. DOI: 10.2200/S00362ED1V01Y201105DTM016. URL: https://doi.org/10.2200/S00362ED1V01Y201105DTM016.

[76] R. S. Sutton and A. G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. URL: http://www.worldcat.org/oclc/37293240.

[77] M. Y. Vardi. 'The Complexity of Relational Query Languages (Extended Abstract)'. In: *Proceedings of the 14th Annual ACM Symposium on Theory of Computing, May 5-7, 1982, San Francisco, California, USA*. 1982, pp. 137–146. DOI: 10.1145/800070.802186. URL: http://doi.acm.org/10.1145/800070.802186.

[78] K. Zhou, M. Lalmas, T. Sakai, R. Cummins and J. M. Jose. 'On the reliability and intuitiveness of aggregated search metrics'. In: *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*. 2013, pp. 689–698. DOI: 10.1145/2505515.2505691. URL: http://doi.acm.org/10.1145/2505515.2505691.