

# 2025 Activity Report

RESEARCH CENTRE: Inria Paris Centre

IN PARTNERSHIP WITH: Ecole normale supérieure de Paris, CNRS

  
Project-Team

# VALDA

Value from Data  


*In collaboration with* Département d'Informatique de l'Ecole Normale Supérieure



## **Project-Team VALDA**

*Creation of the Project-Team: 2018 January 01*

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

## Keywords

### Computer sciences and digital sciences

- A3.1. – Data
  - A3.1.1. – Modeling, representation
  - A3.1.2. – Data management, querying 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. – Data and knowledge analysis
  - A3.3.1. – On-line analytical processing
  - A3.3.2. – Data mining
  - A3.3.3. – Big data analysis
- 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
  - A9.2.3. – Reinforcement learning
  - A9.2.5. – Bayesian methods
- A9.8. – Reasoning

**Other research topics and application domains**

- B2. – Digital health
- B3.3. – Geosciences
- B4. – Energy
  - B4.2. – Nuclear Energy Production
- B6.3.1. – Web
- B6.3.5. – Search engines
- B9.3. – Medias
- 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

## Contents

<b>Project-Team VALDA</b>	<b>1</b>
<b>1 Team members, visitors, external collaborators</b>	<b>6</b>
<b>2 Overall objectives</b>	<b>7</b>
<b>3 Research program</b>	<b>8</b>
3.1 Research axis 1: Foundations of data management . . . . .	8
3.2 Research axis 2: Uncertainty, provenance, and explainability in data management . . . . .	8
3.3 Research axis 3: Knowledge discovery at scale . . . . .	8
<b>4 Application domains</b>	<b>8</b>
<b>5 Highlights of the year</b>	<b>9</b>
5.1 Awards . . . . .	9
<b>6 Latest software developments, platforms, open data</b>	<b>9</b>
6.1 Latest software developments . . . . .	9
6.1.1 ProvSQL . . . . .	9
6.1.2 VUS . . . . .	10
6.1.3 TSB-UAD . . . . .	11
6.1.4 ADecimo . . . . .	11
6.1.5 MSAD . . . . .	12
6.1.6 apxproof . . . . .	13
<b>7 New results</b>	<b>13</b>
7.1 Research axis 1: Foundations of data management . . . . .	13
7.2 Research axis 2: Uncertainty, provenance, and explainability in data management . . . . .	14
7.3 Research axis 3: Knowledge discovery at scale . . . . .	15
<b>8 Bilateral contracts and grants with industry</b>	<b>18</b>
8.1 Bilateral contracts with industry . . . . .	18
<b>9 Partnerships and cooperations</b>	<b>18</b>
9.1 International initiatives . . . . .	18
9.1.1 Participation in other International Programs . . . . .	18
9.2 International research visitors . . . . .	19
9.2.1 Visits of international scientists . . . . .	19
9.2.2 Visits to international teams . . . . .	19
9.3 National initiatives . . . . .	19
9.3.1 ANR . . . . .	19
9.3.2 Others . . . . .	20
<b>10 Dissemination</b>	<b>21</b>
10.1 Promoting scientific activities . . . . .	21
10.1.1 Scientific events: organisation . . . . .	21
10.1.2 Scientific events: selection . . . . .	21
10.1.3 Journal . . . . .	21
10.1.4 Invited talks . . . . .	21
10.1.5 Leadership within the scientific community . . . . .	22
10.1.6 Research administration . . . . .	22
10.2 Teaching - Supervision - Juries - Educational and pedagogical outreach . . . . .	23
10.2.1 Supervision . . . . .	23
10.2.2 Juries . . . . .	24

---

10.3 Popularization . . . . .	24
10.3.1 Specific official responsibilities in science outreach structures . . . . .	24
10.3.2 Productions (articles, videos, podcasts, serious games, ...) . . . . .	24
10.3.3 Participation in Live events . . . . .	25
<b>11 Scientific production . . . . .</b>	<b>25</b>
11.1 Major publications . . . . .	25
11.2 Publications of the year . . . . .	26
11.3 Cited publications . . . . .	29

# 1 Team members, visitors, external collaborators

## Research Scientists

- Serge Abiteboul [Inria, Emeritus, HDR]
- Paul Boniol [Inria, ISFP]
- Camille Bourgaux [CNRS, Researcher]
- Luc Segoufin [Inria, Senior Researcher, HDR]
- Michael Thomazo [Inria, Researcher, HDR]

## Faculty Member

- Pierre Senellart [Team leader, ENS-PSL, Professor, HDR]

## PhD Students

- Felix Chavelli [Inria]
- Anatole Dahan [Université Paris-Cité, until Jul 2025]
- Antoine Gauquier [ENS-PSL]
- Robin Jean [CNRS]
- Lucas Larroque [ENS-PSL]
- Magali Parrino [EDF, CIFRE, from Jul 2025]
- Aryak Sen [CNRS & Université de Grenoble]
- Marijan Soric [Inria, from Mar 2025]
- Emmanouil Sylligardos [ENS-PSL]

## Technical Staff

- Louis Chanaron [Inria, Engineer, from Oct 2025]

## Interns and Apprentices

- Arushi Goyal [IIT Delhi & ENS-PSL, Intern, until May 2025]
- Adam Rozzio [ENS Paris-Saclay & ENS-PSL, Intern, from Feb 2025 until Jul 2025]
- Marijan Soric [Centrale Lyon & Inria, Intern, until Feb 2025]

## Administrative Assistant

- Meriem Guemair [Inria]

## Visiting Scientist

- Victor Vianu [UC San Diego, from Jun 2025]

## 2 Overall 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 [50, 52] 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 [49][51], and possibly distributed [56] 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*<sup>1</sup>: 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* [54] or *distributed query optimization* [55].

Among all these aspects, intensionality is perhaps the one that has least been studied, so let us expand a bit on this. 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 [53] or aggregated search [59]. 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 [58] (train an agent to accomplish a task in a partially known world based on rewards obtained) and of active learning [57] (decide which 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*.

<sup>1</sup>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*.

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.

## 3 Research program

### 3.1 Research axis 1: Foundations of data management

This axis covers the theory of data management, broadly taken, and in particular the fields of *database theory*, *knowledge representation*, and some *symbolic* aspects of *artificial intelligence* (especially, *reasoning on data*).

The goal is to define solid and high-level foundations of data management tasks (query evaluation and optimization of various forms of queries, counting, reasoning, verification of data-centric processes, etc.) through formal tools, such as logics (esp., finite model theory), automata theory, complexity theory; we occasionally have contributions in these areas as well, though most of our work is motivated by data applications. We are especially interested in clean specifications of key aspects of database systems and data management tasks (e.g. confidentiality, access control, robustness), whether they are properties of the data or appropriate (query) languages for these tasks. We study expressive power of languages, computability and complexity of deciding or computing results, as well as the design of appropriate structures (e.g., indexes) to optimize these tasks.

### 3.2 Research axis 2: Uncertainty, provenance, and explainability in data management

This research axis deals with the modeling and efficient management of data that come with some uncertainty (probabilistic distributions, logical incompleteness, missing values, inconsistencies, open-world assumption, etc.) and with provenance information (indicating where the data originates from), as well as with the extraction of uncertainty and provenance annotations from real-world data. Provenance is also linked to explainability: determining where the result of a data management task comes from, how and why it was produced, helps explaining it. Interestingly, the foundations and tools for uncertainty management often rely on provenance annotations. For example, a typical way to compute the probability of query results in probabilistic databases is the so-called *intensional* approach: first generate the provenance of these query results (in some appropriate framework, e.g., that of Boolean functions or of provenance semirings), and then compute the probability of the resulting provenance annotation. For this reason, we deal with uncertainty and provenance in a unified manner, and with explainability as an application thereof.

### 3.3 Research axis 3: Knowledge discovery at scale

Our final axis deals with knowledge discovery at scale. The goal is to use techniques such as data mining, information extraction, data cleaning, information integration, machine learning, to derive knowledge from raw, dirty, inconsistent, heterogeneous, rapidly changing, data from real-world application scenarios.

We intend to leverage our expertise on data management to focus on the scalability of the approaches and tools developed. This is also in some sense an application axis for techniques developed in the other two axes; in particular, we have a focus on intensionality of data (i.e., cost to data access), on the trade-off between data uncertainty and its cost, on data provenance and explanations.

This axis is typically very changing in subtopics, depending on projects, collaborations, application partners.

## 4 Application domains

A large part of Valda's research is foundational in nature and not tailored to any specific application domain. Some applied works target certain application domains however:

**Web data** in a broad-sense (semi-structured, structured or unstructured content extracted from Web databases; knowledge bases from the Semantic Web; social networks; Web archives and Web crawls; Web applications and deep Web databases; crowdsourcing platforms). This is a historical domain of interest of Valda researchers, and we have expertise in the acquisition, extraction, and management of this kind of data.

**Open science** (publication databases, scientific publications, open-source software).

**Clinical data** (notably inconsistent or incomplete hospital records).

**Energy** (notably data from power stations, in collaboration with industrial partners).

**Geoscience** (seismology or vulcanology time series, structured data about geological campaigns).

**Data journalism** (statistical datasets, fact checking data).

Finally, transversal concerns which occur in different applications area and motivate some of our theory work are ethics of data management and privacy.

## 5 Highlights of the year

The Inria–BRGM **Géolaug** challenge, which Valda contributes to, was launched in September 2025.

### 5.1 Awards

Camille Bourgaux, Anton Gnatenco (Free University of Bozen–Bolzano), and Michael Thomazo have received a *best contribution award* at DL 2025 and an *outstanding paper award* at ECAI 2025 for their work on analysing temporal reasoning in description logics using formal grammars [28, 27].

## 6 Latest software developments, platforms, open data

### 6.1 Latest software developments

#### 6.1.1 ProvSQL

**Keywords:** Databases, Provenance, Probability

**Scientific Description:** ProvSQL is a general and easy-to-deploy provenance tracking and probabilistic database system implemented as a PostgreSQL extension. ProvSQL's data and query models closely reflect that of a large core of SQL, including multiset semantics, the full relational algebra, and aggregation. A key part of its implementation relies on generic provenance circuits stored in memory-mapped files.

**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:** Compatibility with PostgreSQL 18. Support for PROV-XML output. Partial support of HAVING queries. Support for compiled semirings, including the counting, Boolean, and Why semirings. Basic documentation infrastructure. Temporal semiring and temporal database support. Various minor enhancements and bug fixes.

**URL:** <https://github.com/PierreSenellart/provsql>

**Publications:** [hal-05037471](#), [hal-05072212](#), [hal-04930705](#), [hal-04911715](#), [hal-04561331](#), [hal-04393781](#), [hal-01672566](#), [hal-01851538](#)

**Contact:** Pierre Senellart

**Participants:** Aryak Sen, Pierre Senellart

**Partners:** Université Grenoble Alpes, CNRS, National University of Singapore

### 6.1.2 VUS

**Name:** Volume Under the Surface

**Keywords:** Time Series, Anomaly detection, Measures, Performance measure, Python

**Scientific Description:** Anomaly detection (AD) is a fundamental task for time-series analytics with important implications for the downstream performance of many applications. In contrast to other domains where AD mainly focuses on point-based anomalies (i.e., outliers in standalone observations), AD for time series is also concerned with range-based anomalies (i.e., outliers spanning multiple observations). Nevertheless, it is common to use traditional point-based information retrieval measures, such as Precision, Recall, and F-score, to assess the quality of methods by thresholding the anomaly score to mark each point as an anomaly or not. However, mapping discrete labels into continuous data introduces unavoidable shortcomings, complicating the evaluation of range-based anomalies. Notably, the choice of evaluation measure may significantly bias the experimental outcome. Despite over six decades of attention, there has never been a large-scale systematic quantitative and qualitative analysis of time-series AD evaluation measures. This paper extensively evaluates quality measures for time-series AD to assess their robustness under noise, misalignments, and different anomaly cardinality ratios. Our results indicate that measures producing quality values independently of a threshold (i.e., AUC-ROC and AUC-PR) are more suitable for time-series AD. Motivated by this observation, we first extend the AUC-based measures to account for range-based anomalies. Then, we introduce a new family of parameter-free and threshold-independent measures, VUS (Volume Under the Surface), to evaluate methods while varying parameters. Our findings demonstrate that our four measures are significantly more robust in assessing the quality of time-series AD methods.

**Functional Description:** The receiver operator characteristic (ROC) curve and the area under the curve (AUC) are widely used to compare the performance of different anomaly detectors. They mainly focus on point-based detection. However, the detection of collective anomalies concerns two factors: whether this outlier is detected and what percentage of this outlier is detected. The first factor is not reflected in the AUC. Another problem is the possible shift between the anomaly score and the real outlier due to the application of the sliding window. To tackle these problems, we incorporate the idea of range-based precision and recall, and suggest the range-based ROC and its counterpart in the precision-recall space, which provides a new evaluation for the collective anomalies. We finally introduce a new measure VUS (Volume Under the Surface) which corresponds to the averaged range-based measure when we vary the range size. We demonstrate in a large experimental evaluation that the proposed measures are significantly more robust to important criteria (such as lag and noise) and also significantly more useful to separate correctly the accurate from the the inaccurate methods.

**News of the Year:** We recently published in 2025 a new paper introducing two optimized implementations of VUS that significantly reduce the execution time of the initial implementation.

Publication: <https://inria.hal.science/hal-05076186>

**URL:** <https://github.com/TheDatumOrg/VUS>

**Publication:** [hal-05076186](https://hal.archives-ouvertes.fr/hal-05076186)

**Contact:** Paul Boniol

**Participants:** Paul Boniol, Emmanouil Sylligardos, 9 anonymous participants

**Partners:** Ohio State University, Université Paris-Descartes

### 6.1.3 TSB-UAD

**Keywords:** Time Series, Anomaly detection, Python, Library

**Scientific Description:** The detection of anomalies in time series has gained ample academic and industrial attention. However, no comprehensive benchmark exists to evaluate time-series anomaly detection methods. It is common to use (i) proprietary or synthetic data, often biased to support particular claims, or (ii) a limited collection of publicly available datasets. Consequently, we often observe methods performing exceptionally well in one dataset but surprisingly poorly in another, creating an illusion of progress. To address the issues above, we thoroughly studied over one hundred papers to identify, collect, process, and systematically format datasets proposed in the past decades. We summarize our effort in TSB-UAD, a new benchmark to ease the evaluation of univariate time-series anomaly detection methods. Overall, TSB-UAD contains 13766 time series with labeled anomalies spanning different domains with high variability of anomaly types, ratios, and sizes. TSB-UAD includes 18 previously proposed datasets containing 1980 time series and we contribute two collections of datasets. Specifically, we generate 958 time series using a principled methodology for transforming 126 time-series classification datasets into time series with labeled anomalies. In addition, we present data transformations with which we introduce new anomalies, resulting in 10828 time series with varying complexity for anomaly detection. Finally, we evaluate 12 representative methods demonstrating that TSB-UAD is a robust resource for assessing anomaly detection methods. TSB-UAD provides a valuable, reproducible, and frequently updated resource to establish a leaderboard of univariate time-series anomaly detection methods.

**Functional Description:** TSB-UAD is a new open, end-to-end benchmark suite to ease the evaluation of univariate time-series anomaly detection methods. Overall, TSB-UAD contains 12686 time series with labeled anomalies spanning different domains with high variability of anomaly types, ratios, and sizes. Specifically, TSB-UAD includes 18 previously proposed datasets containing 1980 time series from real-world data science applications. Motivated by flaws in certain datasets and evaluation strategies in the literature, we study anomaly types and data transformations to contribute two collections of datasets. Specifically, we generate 958 time series using a principled methodology for transforming 126 time-series classification datasets into time series with labeled anomalies. In addition, we present a set of data transformations with which we introduce new anomalies in the public datasets, resulting in 10828 time series (92 datasets) with varying difficulty for anomaly detection.

**URL:** <https://tsb-uad.readthedocs.io/en/latest/>

**Contact:** Paul Boniol

**Participants:** Paul Boniol, Emmanouil Sylligardos, 5 anonymous participants

**Partners:** Université Paris-Descartes, Ohio State University

### 6.1.4 ADecimo

**Name:** A Web-app for the Evaluation of Model selection for Anomaly Detection in Time Series

**Keywords:** Time Series, Anomaly detection, Web Application

**Scientific Description:** Anomaly detection is a fundamental task for time-series analytics with important implications for the downstream performance of many applications. Despite increasing academic interest and the large number of methods proposed in the literature, recent benchmark and evaluation studies demonstrated that there exists no single best anomaly detection method when applied to heterogeneous time series datasets. Therefore, the only scalable and viable solution to solve anomaly detection over very different time series collected from diverse domains is to propose a model selection method that will choose, based on time series characteristics, the best anomaly detection method to run. This paper describes ADecimo, a modular and extensible web application that helps users understand the performance of time series classification algorithms used as model selection methods for time series anomaly detection. Overall, our system enables users to compare 17 different classifiers over

1980 time series, and decide on the most suitable time series classification method for their own time series and use cases.

**Functional Description:** We present here ADecimo, a modular and extensible web application that helps users understand the performance of time series classification algorithms used as model selection methods for time series anomaly detection. Overall, our system enables users to compare 17 different classifiers over 1980 time series, and decide on the most suitable time series classification method for their own time series and use cases.

**URL:** <https://adecimots.streamlit.app/>

**Publication:** [hal-04590326](#)

**Contact:** Paul Boniol

**Participants:** Paul Boniol, Emmanouil Sylligardos, 3 anonymous participants

### 6.1.5 MSAD

**Name:** Model Selection for Anomaly Detection

**Keywords:** Time Series, Machine learning, Classification, Ensemble classifier, Python

**Scientific Description:** Anomaly detection is a fundamental task for time-series analytics with important implications for the downstream performance of many applications. Despite increasing academic interest and the large number of methods proposed in the literature, recent benchmark and evaluation studies demonstrated that no overall best anomaly detection methods exist when applied to very heterogeneous time series datasets. Therefore, the only scalable and viable solution to solve anomaly detection over very different time series collected from diverse domains is to propose a model selection method that will select, based on time series characteristics, the best anomaly detection method to run. Existing AutoML solutions are, unfortunately, not directly applicable to time series anomaly detection, and no evaluation of time series-based approaches for model selection exists. Towards that direction, this paper studies the performance of time series classification methods used as model selection for anomaly detection. Overall, we compare 17 different classifiers over 1800 time series, and we propose the first extensive experimental evaluation of time series classification as model selection for anomaly detection. Our results demonstrate that model selection methods outperform every single anomaly detection method while being in the same order of magnitude regarding execution time. This evaluation is the first step to demonstrate the accuracy and efficiency of time series classification algorithms for anomaly detection, and represents a strong baseline that can then be used to guide the model selection step in general AutoML pipelines.

**Functional Description:** MSAD proposes a pipeline for model selection based on time series classification and an extensive experimental evaluation of existing classification algorithms for this new pipeline. Our results demonstrate that model selection methods outperform every single anomaly detection method while being in the same order of magnitude regarding execution time.

**News of the Year:** In 2025, we published a new paper that extended the model selection pipeline, improving performance in Out-of-Distribution (OoD) settings.

Paper: <https://inria.hal.science/hal-05343228>

**URL:** <https://github.com/boniolp/MSAD>

**Publication:** [hal-05343228](#)

**Contact:** Paul Boniol

**Participants:** Emmanouil Sylligardos, Paul Boniol, Pierre Senellart, 2 anonymous participants

**Partners:** Ohio State University, Université Paris-Descartes

### 6.1.6 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:** Fix forward linking when used in conjunction with aliascnt (e.g., in Springer classes), Compatibility with recent versions of acmart.cls

**News of the Year:** - Fix forward linking when used in conjunction with aliascnt (e.g., in Springer classes) - Compatibility with recent versions of acmart.cls - Support for user-defined claimproof environments - Remove forward linking command from PDF bookmarks

**URL:** <https://github.com/PierreSenellart/apxproof>

**Contact:** Pierre Senellart

**Participant:** Pierre Senellart

## 7 New results

### 7.1 Research axis 1: Foundations of data management

**Participants:** Camille Bourgaux, Anatole Dahan, Jean Robin, Lucas Larroque, Arthur Lombardo, Michaël Thomazo, Luc Segoufin.

**Knowledge representation and knowledge bases** In [27, 28], we establish a correspondence between (fragments of)  $\mathcal{TEL}^\circ$ , a temporal extension of the  $\mathcal{EL}$  description logic with the LTL operator  $\circ^k$ , and some specific kinds of formal grammars, in particular, conjunctive grammars (context-free grammars equipped with the operation of intersection). This connection implies that  $\mathcal{TEL}^\circ$  does not possess the property of ultimate periodicity of models, and further leads to undecidability of query answering in  $\mathcal{TEL}^\circ$ , closing a question left open since the introduction of  $\mathcal{TEL}^\circ$ . Moreover, it also allows to establish decidability of query answering for some new interesting fragments of  $\mathcal{TEL}^\circ$ , and to reuse for this purpose existing tools and algorithms for conjunctive grammars.

**Consistent query answering** In [17], we consider the dichotomy conjecture for consistent query answering under primary key constraints. It states that, for every fixed Boolean conjunctive query  $q$ , testing whether  $q$  is certain (i.e. whether it evaluates to true over all repairs of a given inconsistent database) is either polynomial time or coNP-complete. This conjecture has been verified for self-join-free and path queries. We propose a simple inflationary fixpoint algorithm for consistent query answering which, for a given database, naively computes a set  $\Delta$  of subsets of facts of the database of size at most  $k$ , where  $k$  is the size of the query  $q$ . The algorithm runs in polynomial time and can be formally defined as: (1) Initialize  $\Delta$  with all sets  $S$  of at most  $k$  facts such that  $S \models q$ . (2) Add any set  $S$  of at most  $k$  facts to  $\Delta$  if there exists a block  $B$  (i.e., a maximal set of facts sharing the same key) such that for every fact  $a \in B$  there is a set  $S' \subseteq S \cup \{a\}$  such that  $S' \in \Delta$ . For an input database  $D$ , the algorithm answers "q is certain" iff  $\Delta$  eventually contains the empty set. The algorithm correctly computes certainty when the query  $q$  falls in the polynomial time cases of the known dichotomies for self-join-free queries and path queries. For arbitrary Boolean conjunctive queries, the algorithm is an under-approximation: the query is guaranteed to be certain if the algorithm claims so. However, there are polynomial time certain queries (with self-joins) which are not identified as such by the algorithm.

**The Chase and Existential Rules** [29] The chase is a fundamental algorithm with ubiquitous uses in database theory. Given a database and a set of existential rules (aka tuple-generating dependencies), it iteratively extends the database to ensure that the rules are satisfied in a most general way. This process may not terminate, and a major problem is to decide whether it does. This problem has been studied for a large number of chase variants, which differ by the conditions under which a rule is applied to extend the database. Surprisingly, the complexity of the universal termination of the restricted (aka standard) chase is not fully understood. We close this gap by placing universal restricted chase termination in the analytical hierarchy. This higher hardness is due to the fairness condition, and we propose an alternative condition to reduce the hardness of universal termination.

In [34], we address one of the fundamental open questions in the realm of existential rules: the conjecture on the finite controllability of bounded derivation depth rule sets ( $\text{bdd} \Rightarrow \text{fc}$ ). We take a step toward a positive resolution of this conjecture by demonstrating that universal models generated by  $\text{bdd}$  rule sets cannot contain arbitrarily large tournaments (arbitrarily directed cliques) without entailing a loop query,  $\exists x E(x, x)$ . This simple yet elegant result narrows the space of potential counterexamples to the ( $\text{bdd} \Rightarrow \text{fc}$ ) conjecture.

**Other aspects of theoretical computer science** Our research occasionally touches other aspects of theoretical computer science not related to data management.

In [31], we introduce an extension of fixed-point logic (FP) with a group-order operator ( $\text{ord}$ ), that computes the size of a group generated by a definable set of permutations. This operation is a generalization of the rank operator ( $\text{rk}$ ). We show that  $\text{FP} + \text{ord}$  constitutes a new candidate logic for the class of polynomial-time computable queries ( $\text{P}$ ). As was the case for  $\text{FP} + \text{rk}$ , the model-checking of  $\text{FP} + \text{ord}$  formulae is polynomial-time computable. Moreover, the query separating  $\text{FP} + \text{rk}$  from  $\text{P}$  exhibited by Lichter in his recent breakthrough is definable in  $\text{FP} + \text{ord}$ . Precisely, we show that  $\text{FP} + \text{ord}$  canonizes structures with Abelian colors, a class of structures which contains Lichter’s counter-example. This proof involves expressing a fragment of the group-theoretic approach to graph canonization in the logic  $\text{FP} + \text{ord}$ .

## 7.2 Research axis 2: Uncertainty, provenance, and explainability in data management

**Participants:** Camille Bourgaux, Robin Jean, Pierre Senellart, Aryak Sen.

**Inconsistent knowledge bases** Repair-based semantics have been extensively studied as a means of obtaining meaningful answers to queries posed over inconsistent knowledge bases (KBs). While several works have considered how to exploit a priority relation between facts to select optimal repairs, the question of how to specify such preferences remains largely unaddressed. This motivates us in [22, 23] to introduce a declarative rule-based framework for specifying and computing a priority relation between conflicting facts. As the expressed preferences may contain undesirable cycles, we consider the problem of determining when a set of preference rules always yields an acyclic relation, and we also explore a pragmatic approach that extracts an acyclic relation by applying various cycle removal techniques. Towards an end-to-end system for querying inconsistent KBs, we present a preliminary implementation and experimental evaluation of the framework, which employs answer set programming to evaluate the preference rules, apply the desired cycle resolution techniques to obtain a priority relation, and answer queries under prioritized-repair semantics.

In [24, 25], we explore the issue of inconsistency handling in DatalogMTL, an extension of Datalog with metric temporal operators. Since facts are associated with time intervals, there are different manners to restore consistency when they contradict the rules, such as removing facts or modifying their time intervals. Our first contribution is the definition of relevant notions of conflicts (minimal explanations for inconsistency) and repairs (possible ways of restoring consistency) for this setting and the study of the properties of these notions and the associated inconsistency-tolerant semantics. Our second contribution is a data complexity analysis of the tasks of generating a single conflict / repair and query entailment under repair-based semantics.

**Provenance and probability management** Ensemble methods aggregate the predictions of multiple models by some form of weighted voting. In [33], we consider the impact of the choice of the assignment of voting power to every individual model on the performance of ensemble methods. We empirically and comparatively evaluate the accuracy and running time of the different power voting ensemble methods using standard classifiers and mainstream classification benchmarks. The results show that power ensemble voting outperforms the equal-power baseline, and that unsupervised learning of the voting power can be competitive with respect to supervised learning; within supervised approaches, learning voting power through Shapley values and regression outperforms simply using accuracy.

The Shapley value provides a principled framework for attributing marginal contributions to players in coalitional games. While its axiomatic fairness guarantees have made it a cornerstone of value distribution in economics and multi-agent systems, recent computational advances have extended its applicability to data-driven domains. [32] bridges game-theoretic foundations with probabilistic reasoning by studying Shapley-like scores in stochastic environments. We prove that the expected Shapley value (EShap) – player’s average impact in a game with an independent probabilistic setting – coincides with the Shapley value of the game whose utility is the expected utility of the original game (ShapE). This equality, however, fails for other power indices, such as the Banzhaf index, underscoring the Shapley value’s specificity of consistency in uncertain settings. We further identify that for a certain class of coefficients (including normalized Banzhaf indices) the equality persists, broadening the scope of reliable attribution mechanisms.

ProvSQL is a PostgreSQL extension implementing provenance management and probabilistic database features. ProvSQL seamlessly extends relational database functionality to support the storage, tracking through derivations and transformations, and querying of metadata that explain and qualify the data and query results. In [40], ProvSQL is used to implement a content-based image retrieval system. A deep learning object detection model identifies objects of selected classes located within the images of a large-scale image data set. The uncertainty associated with object detection is recorded. ProvSQL’s provenance model incorporates this uncertainty into the retrieval process, thus facilitating the generation of accurate and reliable results and allowing for decision-making in scenarios with incomplete or uncertain information. The demonstration illustrates how ProvSQL handles query processing, uncertainty tracking, and probability computation. It highlights the utility of a probabilistic database for applications dealing with uncertain data, compared to traditional threshold-based approaches.

In [39], we further enhance ProvSQL by enabling provenance tracking for update operations (DELETE, INSERT, UPDATE). We illustrate the practical utility of update provenance by implementing a temporal database capable of standard operations, including time travel (inspecting past database states), history tracking (monitoring tuple states over time), and undo (reversing previous updates). These features rely on a provenance formalism based on the union-of-intervals m-semiring. Additionally, we emphasize a key advantage of using semiring-based provenance model: its generality allows the same semiring structure to seamlessly support various applications, such as probabilistic databases, by simply modifying the semiring definition.

### 7.3 Research axis 3: Knowledge discovery at scale

**Participants:** Paul Boniol, Felix Chavelli, Antoine Gauquier, Magali Parrino, Pierre Senellart, Marijan Soric, Emmanouil Sylligardos.

**Mining time series** Recent advances in data collection technology, accompanied by the ever-rising volume and velocity of streaming data, underscore the vital need for time series analytics. In this regard, time-series anomaly detection has been an important activity, entailing various applications in fields such as cyber security, financial markets, law enforcement, and health care. While traditional literature on anomaly detection is centered on statistical measures, the increasing number of machine learning algorithms in recent years calls for a structured, general characterization of the research methods for time-series anomaly detection. In [36], we present a process-centric taxonomy for time-series anomaly detection methods, systematically categorizing traditional statistical approaches and contemporary machine learning techniques. Beyond this taxonomy, we conduct a meta-analysis of the existing literature to identify broad research trends. Given the absence of a one-size-fits-all anomaly detector, we also introduce emerging trends for time-series anomaly

detection. Furthermore, we review commonly used evaluation measures and benchmarks, followed by an analysis of benchmark results to provide insights into the impact of different design choices on model performance. Through these contributions, we aim to provide a holistic perspective on time-series anomaly detection and highlight promising avenues for future investigation.

Anomaly detection is a fundamental task for time series analytics with important implications for the downstream performance of many applications. Despite increasing academic interest and the large number of methods proposed in the literature, recent benchmarks and evaluation studies demonstrated that no overall best anomaly detection methods exist when applied to very heterogeneous time series datasets. Therefore, the only scalable and viable solution to solve anomaly detection over very different time series collected from diverse domains is to propose a model selection method that will select, based on time series characteristics, the best anomaly detection methods to run. Existing AutoML solutions are, unfortunately, not directly applicable to time series anomaly detection, and no evaluation of time series-based approaches for model selection exists. Towards that direction, [19] studies the performance of time series classification methods used as model selection for anomaly detection. In total, we evaluate 234 model configurations derived from 16 base classifiers across more than 1980 time series, and we propose the first extensive experimental evaluation of time series classification as model selection for anomaly detection. Our results demonstrate that model selection methods outperform every single anomaly detection method while being in the same order of magnitude regarding execution time. This evaluation is the first step to demonstrate the accuracy and efficiency of time series classification algorithms for anomaly detection, and represents a strong.

In contrast to other domains where AD mainly focuses on point-based anomalies (i.e., outliers in standalone observations), AD for time series is also concerned with range-based anomalies (i.e., outliers spanning multiple observations). Nevertheless, it is common to use traditional point-based information retrieval measures, such as Precision, Recall, and F-score, to assess the quality of methods by thresholding the anomaly score to mark each point as an anomaly or not. However, mapping discrete labels into continuous data introduces unavoidable shortcomings, complicating the evaluation of range-based anomalies. Notably, the choice of evaluation measure may significantly bias the experimental outcome. Despite over six decades of attention, there has never been a large-scale systematic quantitative and qualitative analysis of time-series AD evaluation measures. [15] extensively evaluates quality measures for time-series AD to assess their robustness under noise, misalignments, and different anomaly cardinality ratios. Our results indicate that measures producing quality values independently of a threshold (i.e., AUC-ROC and AUC-PR) are more suitable for time-series AD. Motivated by this observation, we first extend the AUC-based measures to account for range-based anomalies. Then, we introduce a new family of parameter-free and threshold-independent measures, Volume Under the Surface (VUS), to evaluate methods while varying parameters. We also introduce two optimized implementations for VUS that reduce significantly the execution time of the initial implementation. Our findings demonstrate that our four measures are significantly more robust in assessing the quality of time-series AD methods.

Motif Discovery involves identifying recurring patterns and locating their occurrences within a time series without prior knowledge about their shape or location. In practice, Motif Discovery faces several data-related challenges, leading to various definitions of the problem and multiple algorithms addressing these challenges to different extents. However, there has been no systematic evaluation and comparison of these diverse approaches. Consequently, [18] presents a comprehensive literature review covering data-related challenges, motif definitions, and algorithms. We also analyze the strengths and limitations of algorithms carefully chosen to represent the literature diversity. The analysis is structured around key research questions identified from our review. Our experimental findings provide practical guidelines for selecting Motif Discovery algorithms suitable for a given task and suggest directions for future research.

Time series clustering poses a significant challenge with diverse applications across domains. A prominent drawback of existing solutions lies in their limited interpretability, often confined to presenting users with centroids. In addressing this gap, [16] presents k-Graph, an unsupervised method explicitly crafted to augment interpretability in time series clustering. Leveraging a graph representation of time series subsequences, k-Graph constructs multiple graph representations based on different subsequence lengths. This feature accommodates variable-length time series without requiring users to predetermine subsequence lengths. Our experimental results reveal that k-Graph outperforms current state-of-the-art time series clustering algorithms in accuracy, while providing users with meaningful explanations and interpretations of the clustering outcomes.

Time series clustering is important for identifying patterns in these datasets. However, prevailing methods

often encounter obstacles in maintaining data relationships and ensuring interpretability. We present in [26] Graphint, an innovative system based on the  $k$ -Graph methodology that addresses these challenges. Graphint integrates a robust time series clustering algorithm with an interactive tool for comparison and interpretation. More precisely, our system allows users to compare results against competing approaches, identify discriminative subsequences within specified datasets, and visualize the critical information utilized by  $k$ -Graph to generate outputs. Overall, Graphint offers a comprehensive solution for extracting actionable insights from complex temporal datasets.

Time series segmentation is a fundamental task in analyzing temporal data across various domains, from human activity recognition to energy monitoring. While numerous state-of-the-art methods have been developed to tackle this problem, the evaluation of their performance remains critically limited. Existing measures predominantly focus on change point accuracy or rely on point-based measures such as Adjusted Rand Index (ARI), which fail to capture the quality of the detected segments, ignore the nature of errors, and offer limited interpretability. In [30], we address these shortcomings by introducing two novel evaluation measures: WARI (Weighted Adjusted Rand Index), that accounts for the position of segmentation errors, and SMS (State Matching Score), a fine-grained measure that identifies and scores four fundamental types of segmentation errors while allowing error-specific weighting. We empirically validate WARI and SMS on synthetic and real-world benchmarks, showing that they not only provide a more accurate assessment of segmentation quality but also uncover insights, such as error provenance and type, that are inaccessible with traditional measures.

In recent years, electricity suppliers have installed millions of smart meters worldwide to improve the management of the smart grid system. These meters collect a large amount of electrical consumption data to produce valuable information to help consumers reduce their electricity footprint. However, having non-expert users (e.g., consumers or sales advisors) understand these data and derive usage patterns for different appliances has become a significant challenge for electricity suppliers because these data record the aggregated behavior of all appliances. At the same time, ground-truth labels (which could train appliance detection and localization models) are expensive to collect and extremely scarce in practice. [37] introduces DeviceScope, an interactive tool designed to facilitate understanding smart meter data by detecting and localizing individual appliance patterns within a given time period. Our system is based on CamAL (Class Activation Map-based Appliance Localization), a novel weakly supervised approach for appliance localization that only requires the knowledge of the existence of an appliance in a household to be trained.

Improving smart grid system management is crucial in the fight against climate change, and enabling consumers to play an active role in this effort is a significant challenge for electricity suppliers. In this regard, millions of smart meters have been deployed worldwide in the last decade, recording the main electricity power consumed in individual households. This data produces valuable information that can help them reduce their electricity footprint; nevertheless, the collected signal aggregates the consumption of the different appliances running simultaneously in the house, making it difficult to apprehend. Non-Intrusive Load Monitoring (NILM) refers to the challenge of estimating the power consumption, pattern, or on/off state activation of individual appliances using the main smart meter signal. Recent methods proposed to tackle this task are based on a fully supervised deep-learning approach that requires both the aggregate signal and the ground truth of individual appliance power. However, such labels are expensive to collect and extremely scarce in practice, as they require conducting intrusive surveys in households to monitor each appliance. In [38], we introduce CamAL, a weakly supervised approach for appliance pattern localization that only requires information on the presence of an appliance in a household to be trained. CamAL merges an ensemble of deep-learning classifiers combined with an explainable classification method to be able to localize appliance patterns. Our experimental evaluation, conducted on 4 real-world datasets, demonstrates that CamAL significantly outperforms existing weakly supervised baselines and that current SotA fully supervised NILM approaches require significantly more labels to reach CamAL performances.

**Information Extraction** [35], which is situated within the TheoremKB [41] project, presents TheoremView, a novel framework for extracting proofs and theorems from raw PDF scientific papers without requiring LaTeX source files. Our approach combines three modalities (font, text, and vision) with sequential modeling to capture long-term dependencies and layout information. By eliminating OCR preprocessing, TheoremView reduces computational overhead for real-time applications while providing robust automated theorem extraction.

Graphs, and notably RDF graphs, are a prominent way of sharing data. As data usage democratizes,

users need help figuring out the useful content of a graph dataset. In particular, journalists with whom we collaborate are interested in identifying, in a graph, the connections between entities, e.g., people, organizations, emails, etc. In [14], we present a novel method for exploring data graphs through their data paths connecting Named Entities (NEs, in short); each data path leads to a tabular-looking set of results. NEs are extracted from the data through dedicated Information Extraction modules. Our method builds upon the pre-existing ConnectionLens platform and follow-up work in the Abstra project, which builds simple, visual ER-style summaries of semi-structured data. The contribution of the present work, and its novelty, is twofold. First, we propose a novel analysis of entity-to-entity paths contained in datasets of any nature, and propose a new method for ranking paths, leveraging a novel Information Extraction module we built on top of ChatGPT. Second, we present an efficient approach to enumerate and compute NE paths, based on an algorithm which automatically recommends sub-paths to materialize, and rewrites the path queries using these subpaths. Our experiments demonstrate the interest of NE paths and the efficiency of our method for computing and ranking them.

## 8 Bilateral contracts and grants with industry

### 8.1 Bilateral contracts with industry

**Participants:** Paul Boniol, Magali Parrino, Pierre Senellart.

Magali Parrino started her PhD in 2025, under a CIFRE agreement between Valda (Paul Boniol and Pierre Senellart) and EDF (Chatou lab).

## 9 Partnerships and cooperations

### 9.1 International initiatives

#### 9.1.1 Participation in other International Programs

##### DesCartes

**Participants:** Pierre Senellart.

**Title:** Intelligent Modelling for Decision-making in Critical Urban Systems

**Partner Institution(s):** CNRS@CREATE, National University of Singapore

**Duration:** 2021–2026

**Additional info:** DesCartes 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, and became in 2025 Lead PI for one of the workpackages.

##### International ANR project EQUUS

**Participants:** Luc Segoufin.

**Title:** Efficient query answering under updates

**Partner Institution(s):** TU Ilmenau, Uni. Bayreuth, HU Berlin, CNRS

**Duration:** 2020–2025

## 9.2 International research visitors

### 9.2.1 Visits of international scientists

#### Other international visits to the team

##### **Anton Gnatenko**

**Status:** PhD students

**Institution of origin:** Free University of Bozen–Bolzano

**Country:** Italy

**Dates:** December 2024 to May 2025

**Mobility program:** PhD research visit

##### **Amélie Marian**

**Status:** Professor

**Institution of origin:** Rutgers University

**Country:** USA

**Dates:** March 2026 to April 2026

**Mobility program:** ENS Visiting Professor

##### **Victor Vianu**

**Status:** Professor

**Institution of origin:** UC San Diego

**Country:** USA

**Dates:** June 2025 to January 2026

**Mobility program:** Sabbatical

### 9.2.2 Visits to international teams

#### Research stays abroad

- Pierre Senellart was an invited participant to the *Logic and Algorithms in DB Theory and AI Reunion* seminar at UC Berkeley, CA, USA (January 2025)
- Camille Bourgaux was an invited participant to the *Semirings in Databases, Automata, and Logic* seminar in Dagstuhl, Germany (February 2025)

## 9.3 National initiatives

### 9.3.1 ANR

**PRC EXPAND (coordinator)****Participants:** Michael Thomazo, , Camille Bourgaux.**Title:** Expanding the reach of ontology-based data access: EXpressivity, exPlanation, and Algorithms**Partner Institution(s):** LIRMM, LaBRI, LIMOS, Inria Lille (SPIRALS & D-DAL), IRISA**Duration:** 2025–2030**Budget for Valda:** 55 k€ (Inria budget)**PR[AI]RIE-PSAI AI Cluster****Participants:** Pierre Senellart, Camille Bourgaux.**Title:** Paris Artificial Intelligence Research Institute – Paris School of AI**Duration:** 2025–2029**Funding for Valda:** 575 k€ (ENS budget)**Megyn's Bienvenu INTENDED Chair in Artificial Intelligence****Participants:** Camille Bourgaux.**Title:** Intelligent handling of imperfect data**Partner Institution(s):** LaBRI**Duration:** 2020–2026**9.3.2 Others****France 2030 i-Demo Cyberté project****Participants:** Paul Boniol.**Partner institution(s):** Scality, Inria Rennes (CIDRE)**Duration:** 2025–2030**Funding for Valda:** 499 k€ (Inria budget)**CNRS MITI nanoNet project****Participants:** Paul Boniol.**Title:** Méthodologie avancée pour la détection des nanoparticules dans les séries temporelles spICP-ToF-MS**Partner Institution(s):** IPGP**Duration:** 2025–2026

## 10 Dissemination

### 10.1 Promoting scientific activities

#### 10.1.1 Scientific events: organisation

##### General chair, scientific chair

- Paul Boniol, IEEE BigData 2025, Chair of the Industrial & Government track
- Paul Boniol, International Workshop on Multivariate Time Series Analytics (MulTiSA) 2025, Panel Chair

##### Member of the organizing committees

- Camille Bourgaux, member of the DL steering committee
- Camille Bourgaux, co-responsible for the MaDICS/RADIA RECAST working group(organization of a thematic day in November, and two sessions of the GDR MaDICS symposium in May)
- Luc Segoufin, member of the STACS steering committee
- Pierre Senellart, editorial board of the LIPIcs series of conference proceedings

#### 10.1.2 Scientific events: selection

##### Member of the conference program committees

- Paul Boniol, VLDB 2025, EDBT 2025, ICDE 2025 (Industry & applications track), Multisa Workshop of ICDE 2025, BDA 2025, BERT2S Workshop of NeurIPS 2025
- Camille Bourgaux, IJCAI 2025, KR 2025, DL 2025
- Antoine Gauquier, WASP 2025
- Pierre Senellart, SIGMOD 2026, Provenance Week 2025, SDProc 2025
- Michael Thomazo, KR 2025, RuleML+RR 2025, IJCAI 2025

#### 10.1.3 Journal

##### Member of the editorial boards

- Luc Segoufin, associate editor, *ACM Transactions on Computational Logic*
- Victor Vianu, editor, Database Theory Column, *SIGACT News*

##### Reviewer - reviewing activities

- Pierre Senellart, review for *Transactions on Graph Data and Knowledge*

#### 10.1.4 Invited talks

- Paul Boniol, *Anomaly Detection in Time Series: Overview and New Trends*, Invited speaker at Orange Innovation
- Paul Boniol, *An introduction to Time series anomaly detection (a data-driven perspective)*, Invited speaker for SIDOS at EGC 2025, Strasbourg, France (January 2025)
- Pierre Senellart, *Quels horizons de pratique pour la recherche en IA?*, Invited speaker at Printemps Couperin, Paris, France (March 2025)

- Silviu Maniu (Univ. Grenoble–Alpes) & Pierre Senellart, *Making Provenance and Probabilistic Database Theory Work in Practice*, Invited talk at ICDT 2025 (Database Theory in Practice), Barcelona, Spain (March 2025) [21]
- Pierre Senellart, *Qualitative Evaluation of Academic Careers in Computer Science at CNRS. Global Forum on Development of Computer Science, Tsinghua University, Beijing, Chine*, Invited keynote speaker at the *Global Forum on Development of Computer Science* of Tsinghua University, Beijing, China (April 2025) [44]
- Pierre Senellart, *Artificial Intelligence. A Personal View*. Invited speaker at INSP Days, Paris, France (July 2025)
- Pierre Senellart, *Les BD pourront-elles sauver l'IA?*, Panel participant, BDA 2025, Toulouse, France (October 2025)
- Pierre Senellart, *Intelligence artificielle: Concepts, modèles et enjeux*, Invited speaker at Séminaire scientifique et technique de l'Inrap, Chartres, France (November 2025)
- Paul Boniol, *Time Series Anomaly Detection: The Road to Automatic Solutions*, Invited Speaker at the 3rd Macau Symposium on Data Science, Macau SAR, China (December 2025)

### 10.1.5 Leadership within the scientific community

- Serge Abiteboul is a member of the French Academy of Sciences, of the Academia Europaea, and an ACM Fellow.
- Pierre Senellart was until August 2025 is a junior member of the Institut Universitaire de France.

### 10.1.6 Research administration

- Serge Abiteboul is a member of the scientific committee of the Programme Inria Quadrant (PIQ).
- Antoine Gauquier is an elected member of the *Conseil d'Administration* of ENS-PSL
- Antoine Gauquier is an elected member of the DIENS lab council
- Luc Segoufin is a member of the *Formation Spécialisée de Site (FSS)* of the Inria Paris research centre.
- Pierre Senellart is Vice-President of PSL University in charge of Digital infrastructure and IT convergence. [48]
- Pierre Senellart was until August 2025 the president of section 6 of the National Committee for Scientific Research. [43] As a representative of CoNRS, Pierre Senellart was in the Hcéres evaluation committee of the IRIT research unit, and president of the evaluation committee of the LIRMM research unit.
- Pierre Senellart was until August 2025 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).
- Pierre Senellart is deputy director of the DI ENS laboratory, joint between ENS, CNRS, and Inria.
- Pierre Senellart is the scientific resource person for *Scientific information & edition* of the Inria Paris centre.
- Pierre Senellart is the vice-president of the **Gilles Kahn PhD award** of Société Informatique de France.
- Pierre Senellart is a member of the strategic orientation committee of **ISIMA**.
- Michael Thomazo is a deputy director of the *École Doctorale Sciences Mathématiques de Paris-Centre* (ED386)
- We participated in the following hiring committee within universities:
  - Camille Bourgaux, Maître de conférences, ENSEIRB-MATMECA-Bordeaux INP

## 10.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

- Licence: The Art of Computer Programming, L1, International Bachelor of Science in Artificial Intelligence, PSL – Pierre Senellart
- Licence: Algorithms, L1, CPES, PSL – Antoine Gauquier
- Licence: Differential calculus, L2, CPES, PSL – Antoine Gauquier
- Licence: Formal Languages, Computability, Complexity, L3, ENS – Michael Thomazo, Lucas Larroque
- Licence: Databases, L3, ENS – Pierre Senellart, Paul Boniol, Lucas Larroque
- Licence: Practical Computing, L3, École normale supérieure – Pierre Senellart
- Master: Logiques de description, M1, DCI – Camille Bourgaux
- Master: Data acquisition, extraction, and storage, M2, IASD – Pierre Senellart
- Master: Knowledge graphs, description logics, and reasoning on data, M2, IASD – Michael Thomazo
- Master: NoSQL databases, M2, IASD – Paul Boniol
- Professional training: Web Security, PESTO (Corps des Mines professional training) – Pierre Senellart

As a professor at ENS, Pierre Senellart held various teaching responsibilities (M2 administration, entrance competition) at ENS. Pierre Senellart is the academic director of the graduate program in Computer Science of PSL.

As an adjunct professor at PSL, Michaël Thomazo is in charge of PhD committees within DI ENS and deputy director of the École doctorale.

We also gave invited courses in summer schools:

- Camille Bourgaux, *Inconsistency-Tolerant Semantics Based on (Preferred) Repairs*, 21st Reasoning Web Summer School (RW 2025) – Istanbul, Turkey [20]
- Paul Boniol, *Time Series Anomaly Detection*, Summer school on Artificial Intelligence for Aerospace – GSSI, L'Aquila, Italy
- Paul Boniol, *Time Series Anomaly Detection: Foundations and Practice*, TwinODIS 1st Summer School – FORTH-ICS, Heraklion, Greece

Most permanent 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.

### 10.2.1 Supervision

- PhD defended: Anatole Dahan, *The Role of Permutation Groups in the Search for a Logic for Polynomial Time*, 2020–2025, Arnaud Durand (Université Paris-Cité) & Luc Segoufin [42]
- PhD in progress: Antoine Gauquier, *Intelligent construction of a multimodal and heterogeneous data warehouse, with data traceability*, started in September 2023, Pierre Senellart & Ioana Manolescu (Inria Cedar)
- PhD in progress: Lucas Larroque, *Extension of rewriting procedures for reasoning using existential rules*, started in September 2023, Michaël Thomazo
- PhD in progress: Robin Jean, *Integration of preferences and domain knowledge in inconsistency-tolerant ontology-based data access*, started in October 2023, Meghyn Bienvenu (CNRS LaBRI) & Camille Bourgaux

- PhD in progress: Aryak Sen, *Scalability of a data provenance and probability management system*, started in February 2024, Silviu Maniu (Université Grenoble Alpes) & Pierre Senellart
- PhD in progress: Emmanouil Sylligardos, *Accuracy and execution time trade-off in ensembling and model selection for time series analytics*, started in February 2024, Paul Boniol & Pierre Senellart
- PhD in progress: Felix Chavelli, *Graph representations for multivariate time series analytics*, started in October 2024, Paul Boniol & Michaël Thomazo
- PhD in progress: Pratik Karmakar, *Quality, uncertainty, and lineage of data*, Stéphane Bressan (NUS, deceased), Tan Kian-Lee (NUS), & Pierre Senellart (as he is based in Singapore, he is not considered a Valda member)
- PhD in progress: Marijan Soric, *Exploitation et structuration des données et des connaissances géologiques hétérogènes*, started in March 2025, Pierre Senellart, Ioana Manolescu (Inria Cedar), & Cécile Gracianne (BRGM)
- PhD in progress: Magali Parrino, *Détection non-supervisée d'anomalies dans des flux continus de séries temporelles multivariées*, started in July 2025, Paul Boniol, Emmanuel Remy (EDF), & Pierre Senellart
- PhD in progress: Arthur Lombardo; started in October 2025, Pierre Senellart, Antoine Amarilli (Inria D-DAL) & Mikaël Monet (Inria D-DAL) (as he is based in Lille, he is considered a D-DAL member)
- Master's internship: Arushi Goyal; Pierre Senellart
- Master's internship: Marijan Soric; Pierre Senellart and Ioana Manolescu (Inria Cedar) [45]
- M1 research project: Jeanne Coschieri; Michael Thomazo & David Carral (Inria Boreal)
- M1 research project: Paul Raphaël; Michael Thomazo & Lucas Larroque

### 10.2.2 Juries

- PhD: François Amat [reviewer], Institut polytechnique de Paris, Pierre Senellart

## 10.3 Popularization

### 10.3.1 Specific official responsibilities in science outreach structures

- Serge Abiteboul, President of the scientific steering committee of ANR
- Serge Abiteboul, President of the AFNIC Foundation
- Pierre Senellart is a scientific expert advising the Scientific and Ethical Committee of Parcoursup and MonMaster, the platforms for the selection of higher-education students at the first-year level and the Master's level. As such, he contributed to the [7th yearly report of the committee to the French parliament](#)

### 10.3.2 Productions (articles, videos, podcasts, serious games, ...)

- Serge Abiteboul, editor of the [binaire](#) blog, which moved from the blog platform of *Le Monde* to that of *La Recherche*
- Serge Abiteboul, codirector of the *Parlez-moi d'IA* podcast on [Cause commune](#)
- Serge Abiteboul, co-author of articles on theatre and computer science [46, 47]

### 10.3.3 Participation in Live events

- Serge Abiteboul, co-organizer with French Senator Ghislaine Senée of a Colloquium at the Senate : *Les données au service des territoires intelligents*
- Serge Abiteboul, co-organizer with Isabelle Hilali from Datacraft of eh conference: *Quantum & Intelligence artificielle : vers une convergence des ruptures technologiques ?*

## 11 Scientific production

### 11.1 Major publications

- [1] 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](https://hal.inria.fr/hal-02307999). URL: <https://hal.inria.fr/hal-02307999>.
- [2] 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>.
- [3] P. Boniol, E. Sylligardos, J. Paparrizos, P. Trahanias and T. Palpanas. ‘ADecimo: Model Selection for Time Series Anomaly Detection’. In: *ICDE 2024 - IEEE 40th International Conference on Data Engineering*. Utrecht, Netherlands, 13th May 2024. URL: <https://inria.hal.science/hal-04590326>.
- [4] C. Bourgaux, P. Bourhis, L. Peterfreund and M. Thomazo. ‘Revisiting Semiring Provenance for Datalog’. In: *KR 2022 - 19th International Conference on Principles of Knowledge Representation and Reasoning*. Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning. Haifa, Israel, 31st July 2022, pp. 91–101. DOI: [10.24963/kr.2022/10](https://hal.science/hal-03771031). URL: <https://hal.science/hal-03771031>.
- [5] 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>.
- [6] 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>.
- [7] N. Carmeli and L. Segoufin. ‘Conjunctive Queries With Self-Joins, Towards a Fine-Grained Complexity Analysis’. In: *PODS’23*. Seattle, United States, 18th June 2023. URL: <https://inria.hal.science/hal-04136055>.
- [8] 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](https://hal.inria.fr/hal-03127726). URL: <https://hal.inria.fr/hal-03127726>.
- [9] N. Grosshans, P. McKenzie and L. Segoufin. ‘Tameness and the power of programs over monoids in DA’. In: *Logical Methods in Computer Science* 18.3 (2nd Aug. 2022), 14:1–14:34. DOI: [10.46298/lmcs-18\(3:14\)2022](https://hal.science/hal-03114304). URL: <https://hal.science/hal-03114304>.
- [10] P. Karmakar, M. Monet, P. Senellart and S. Bressan. ‘Expected Shapley-Like Scores of Boolean Functions: Complexity and Applications to Probabilistic Databases’. In: *Proceedings of the ACM on Management of Data* 2.2 (PODS) (12th Jan. 2024). DOI: [10.1145/3651593](https://inria.hal.science/hal-04393781). URL: <https://inria.hal.science/hal-04393781>.
- [11] N. Schweikardt, L. Segoufin and A. Vigny. ‘Enumeration for FO Queries over Nowhere Dense Graphs’. In: *Journal of the ACM (JACM)* 69.3 (30th June 2022), pp. 1–37. DOI: [10.1145/3517035](https://hal.inria.fr/hal-03809754). URL: <https://hal.inria.fr/hal-03809754>.

- [12] 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](https://doi.org/10.14778/3229863.3236253). URL: <https://hal.inria.fr/hal-01851538>.
- [13] E. Toussaint, P. Guagliardo, L. Libkin and J. Sequeda. ‘Troubles with nulls, views from the users’. In: *Proceedings of the VLDB Endowment (PVLDB)* 15.11 (July 2022), pp. 2613–2625. DOI: [10.14778/3551793.3551818](https://doi.org/10.14778/3551793.3551818). URL: <https://hal.inria.fr/hal-03934346>.

## 11.2 Publications of the year

### International journals

- [14] N. Barret, A. Gauquier, J.-J. Law and I. Manolescu. ‘Finding meaningful paths in heterogeneous graphs with PathWays’. In: *Information Systems* 127 (Jan. 2025), p. 102463. DOI: [10.1016/j.is.2024.102463](https://doi.org/10.1016/j.is.2024.102463). URL: <https://hal.science/hal-04727209> (cit. on p. 18).
- [15] P. Boniol, A. K. Krishna, M. Bruel, Q. Liu, M. Huang, T. Palpanas, R. S. Tsay, A. Elmore, M. J. Franklin and J. Paparrizos. ‘VUS: Effective and Efficient Accuracy Measures for Time-Series Anomaly Detection’. In: *The VLDB Journal* 34.32 (18th Feb. 2025). DOI: [10.1007/s00778-025-00907-x](https://doi.org/10.1007/s00778-025-00907-x). URL: <https://inria.hal.science/hal-05076186> (cit. on p. 16).
- [16] P. Boniol, D. Tiano, A. Bonifati and T. Palpanas. ‘ $k$ -Graph: A Graph Embedding for Interpretable Time Series Clustering’. In: *IEEE Transactions on Knowledge and Data Engineering* (2025), pp. 1–14. DOI: [10.1109/TKDE.2025.3543946](https://doi.org/10.1109/TKDE.2025.3543946). URL: <https://inria.hal.science/hal-04981926> (cit. on p. 16).
- [17] D. Figueira, A. Padmanabha, L. Segoufin and C. Sirangelo. ‘A Simple Algorithm for Consistent Query Answering under Primary Keys’. In: *Logical Methods in Computer Science* 21.1 (21st Feb. 2025). DOI: [10.46298/lmcs-21\(1:18\)2025](https://doi.org/10.46298/lmcs-21(1:18)2025). URL: <https://hal.science/hal-04966879> (cit. on p. 13).
- [18] V. Guerrini, T. Germain, C. Truong, L. Oudre and P. Boniol. ‘Time Series Motif Discovery: A Comprehensive Evaluation’. In: *Proceedings of the VLDB Endowment (PVLDB)* 18.7 (1st Aug. 2025), pp. 2226–2239. DOI: [10.14778/3734839.3734857](https://doi.org/10.14778/3734839.3734857). URL: <https://inria.hal.science/hal-05218910> (cit. on p. 16).
- [19] E. Sylligardos, J. Paparrizos, T. Palpanas, P. Senellart and P. Boniol. ‘MSAD: A Deep Dive into Model Selection for Time series Anomaly Detection’. In: *The VLDB Journal* 34.6 (30th Oct. 2025), p. 72. DOI: [10.1007/s00778-025-00949-1](https://doi.org/10.1007/s00778-025-00949-1). URL: <https://inria.hal.science/hal-05343228> (cit. on p. 16).

### Invited conferences

- [20] C. Bourgaux. ‘Inconsistency-Tolerant Semantics Based on (Preferred) Repairs’. In: *RW 2025 - 21st Reasoning Web International Summer School*. Reasoning Web Summer School (RW 2025). Istanbul, Turkey, 25th Sept. 2025. DOI: [10.4230/OASICS.RW.2024/2025.5](https://doi.org/10.4230/OASICS.RW.2024/2025.5). URL: <https://inria.hal.science/hal-05291421> (cit. on p. 23).
- [21] S. Maniu and P. Senellart. ‘Database Theory in Action: Making Provenance and Probabilistic Database Theory Work in Practice (Invited Talk)’. In: *ICDT 2025 - International Conference on Database Theory*. Barcelona, Spain, 25th Mar. 2025. DOI: [10.4230/LIPICS.ICDT.2025.35](https://doi.org/10.4230/LIPICS.ICDT.2025.35). URL: <https://inria.hal.science/hal-04911715> (cit. on p. 22).

### International peer-reviewed conferences

- [22] M. Bienvenu, C. Bourgaux, K. Inoue and R. Jean. ‘A Rule-Based Approach to Specifying Preferences over Conflicting Facts and Querying Inconsistent Knowledge Bases’. In: *Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning (KR 2025)*. KR 2025 - 22nd International Conference on Principles of Knowledge Representation and Reasoning. Melbourne, Australia, 11th Nov. 2025. URL: <https://inria.hal.science/hal-05291388> (cit. on p. 14).

- [23] M. Bienvenu, C. Bourgaux, K. Inoue and R. Jean. ‘A Rule-Based Approach to Specifying Preferences over Conflicting Facts and Querying Inconsistent Knowledge Bases (Extended Abstract)’. In: *Proceedings of the 38th International Workshop on Description Logics (DL 2025)*. DL 2025 - 38th International Workshop on Description Logics. Opole, Poland, 3rd Sept. 2025. URL: <https://inria.hal.science/hal-05291482> (cit. on p. 14).
- [24] M. Bienvenu, C. Bourgaux and A. Khodadaditaghanaki. ‘Inconsistency Handling in DatalogMTL’. In: *IJCAI 2025 - Thirty-Fourth International Joint Conference on Artificial Intelligence*. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 16th Aug. 2025. DOI: [10.24963/ijcai.2025/487](https://doi.org/10.24963/ijcai.2025/487). URL: <https://inria.hal.science/hal-05291362> (cit. on p. 14).
- [25] M. Bienvenu, C. Bourgaux and A. Khodadaditaghanaki. ‘Inconsistency Handling in DatalogMTL (Extended Abstract)’. In: *Proceedings of the 38th International Workshop on Description Logics (DL 2025)*. DL 2025 - 38th International Workshop on Description Logics. Opole, Poland, 3rd Sept. 2025. URL: <https://inria.hal.science/hal-05291501> (cit. on p. 14).
- [26] P. Boniol, D. Tiano, A. Bonifati and T. Palpanas. ‘Graphint: Graph-based Time Series Clustering Visualisation Tool’. In: *ICDE 2025 - IEEE 41th International Conference on Data Engineering*. Hong Kong, Hong Kong SAR China, 10th Mar. 2025. URL: <https://inria.hal.science/hal-05076185> (cit. on p. 17).
- [27] C. Bourgaux, A. Gnatenco and M. Thomazo. ‘Analysing Temporal Reasoning in Description Logics Using Formal Grammars’. In: *Proceedings of the 28th European Conference on Artificial Intelligence. ECAI’25 – 28th European Conference on Artificial Intelligence*. Bologna, Italy, 25th Oct. 2025. URL: <https://inria.hal.science/hal-05273645> (cit. on pp. 9, 13).
- [28] C. Bourgaux, A. Gnatenco and M. Thomazo. ‘Analysing Temporal Reasoning in Description Logics Using Formal Grammars (Extended Abstract)’. In: *Proceedings of the 38th International Workshop on Description Logics (DL 2025)*. DL 2025 - 38th International Workshop on Description Logics. Opole, Poland, 3rd Sept. 2025. URL: <https://inria.hal.science/hal-05291517> (cit. on pp. 9, 13).
- [29] D. Carral, L. Gerlach, L. Larroque and M. Thomazo. ‘Restricted Chase Termination: You Want More than Fairness’. In: *ACM digital library. PODS 2025 - ACM SIGMOD/PODS International Conference on Management of Data*. Vol. 3. Proceedings of the ACM on management of data 2. Berlin, Germany, 22nd June 2025, pp. 1–17. DOI: [10.1145/3725246](https://doi.org/10.1145/3725246). URL: <https://hal.science/hal-05240721> (cit. on p. 14).
- [30] F. Chavelli, P. Boniol and M. Thomazo. ‘Toward Interpretable Evaluation Measures for Time Series Segmentation’. In: *NeurIPS 2025 - 39th Annual Conference on Neural Information Processing Systems*. San Diego, United States: arXiv, 2025. URL: <https://inria.hal.science/hal-05334499> (cit. on p. 17).
- [31] A. Dahan. ‘Group Order Logic’. In: *LICS 2025 - Logic in Computer Science*. Singapore, Singapore: arXiv, 21st May 2025. DOI: [10.48550/arXiv.2505.15359](https://doi.org/10.48550/arXiv.2505.15359). URL: <https://hal.science/hal-05083596> (cit. on p. 14).
- [32] P. Karmakar, A. Gauquier and P. Senellart. ‘Expected Shapley Value is Shapley Value for Expected Utility Game’. In: *ECSQARU 2025 - 18th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Hagen, Germany, 23rd Sept. 2025. URL: <https://inria.hal.science/hal-05177852> (cit. on p. 15).
- [33] P. Karmakar, A. Saadeh, P. Senellart and S. Bressan. ‘Discovering Voting Power for Ensemble Methods’. In: *DEXA 2025 - International Conference on Database and Expert Systems Applications*. Bangkok, Thailand, 25th Aug. 2025. URL: <https://inria.hal.science/hal-05108835> (cit. on p. 15).
- [34] L. Larroque, P. Ostropolski-Nalewaja and M. Thomazo. ‘No Cliques Allowed: The Next Step Towards BDD/FC Conjecture’. In: *Proceedings of the ACM on Management of Data. PODS 2025 - ACM SIGMOD/PODS International Conference on Management of Data*. Vol. 3. Berlin, Germany, 9th June 2025, pp. 1–20. DOI: [10.1145/3725238](https://doi.org/10.1145/3725238). URL: <https://inria.hal.science/hal-05273623> (cit. on p. 14).

- [35] S. Mishra, N. Sharma, A. Gauquier and P. Senellart. ‘TheoremView: A Framework for Extracting Theorem-Like Environments from Raw PDFs’. In: ECIR 2025 - European Conference on Information Retrieval. Lucca, Italy: Springer, 6th Apr. 2025, p. 6. DOI: [10.1007/978-3-031-88720-8\\_5](https://doi.org/10.1007/978-3-031-88720-8_5). URL: <https://inria.hal.science/hal-04894570> (cit. on p. 17).
- [36] J. Paparrizos, P. Boniol, Q. Liu and T. Palpanas. ‘Advances in Time-Series Anomaly Detection: Algorithms, Benchmarks, and Evaluation Measures’. In: KDD ’25: The 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Toronto (ON), Canada: ACM, 3rd Aug. 2025, pp. 6151–6161. DOI: [10.1145/3711896.3736565](https://doi.org/10.1145/3711896.3736565). URL: <https://inria.hal.science/hal-05218929> (cit. on p. 15).
- [37] A. Petralia, P. Boniol, P. Charpentier and T. Palpanas. ‘DeviceScope: An Interactive App to Detect and Localize Appliance Patterns in Electricity Consumption Time Series’. In: ICDE 2025 - IEEE 41th International Conference on Data Engineering. Hong Kong, Hong Kong SAR China, 19th May 2025. URL: <https://inria.hal.science/hal-05076225> (cit. on p. 17).
- [38] A. Petralia, P. Boniol, P. Charpentier and T. Palpanas. ‘Few Labels are all you need: A Weakly Supervised Framework for Appliance Localization in Smart-Meter Series’. In: ICDE 2025 - IEEE 41th International Conference on Data Engineering. Hong Kong, Hong Kong SAR China, 19th May 2025. URL: <https://inria.hal.science/hal-05076222> (cit. on p. 17).
- [39] A. A. Widiaatmaja, B. Djeflal, A. Dandekar and P. Senellart. ‘Demonstration of Provenance through Temporal Databases’. In: PW25 - ProvenanceWeek. Berlin, Germany, 27th June 2025. DOI: [10.1145/3736229.3736253](https://doi.org/10.1145/3736229.3736253). URL: <https://inria.hal.science/hal-05072212> (cit. on p. 15).
- [40] F. Yunus, P. Karmakar, P. Senellart, T. Abdesslem and S. Bressan. ‘Using a Probabilistic Database in an Image Retrieval Application’. In: EDBT 2025 - 28th International Conference on Extending Database Technology. Barcelona, Spain, 25th Mar. 2025. URL: <https://inria.hal.science/hal-04930705> (cit. on p. 15).

#### National peer-reviewed Conferences

- [41] S. Mishra, A. Gauquier and P. Senellart. ‘Apprentissage multimodal modulaire pour l’extraction de théorèmes et de preuves dans des documents scientifiques longs’. In: *Revue des Nouvelles Technologies de l’Information*. Extraction et Gestion des Connaissances, EGC’2025. Strasbourg, France, 27th Jan. 2025. URL: <https://inria.hal.science/hal-04806300> (cit. on p. 17).

#### Doctoral dissertations and habilitation theses

- [42] A. Dahan. ‘The Role of Permutation Groups in the Search for a Logic for Polynomial Time’. Université Paris Cité (UPC), 1st July 2025. URL: <https://hal.science/tel-05413898> (cit. on p. 23).

#### Reports & preprints

- [43] N. Appel, J. Bourdon, N. Bousquet, J. Cohen, A. Genitrini, P. Georgeon, Y. Grandvalet, K. Jaffrès-Runser, A. Legrand, D. Markham, A. Muscholl, A. Paparrizou, L. Paulevé, M. Poss, M. Gradinariu Potop-Butucaru, J.-F. Raymond, R. Rouvoy, Y. Sallent, P. Senellart, T. Seiller, Y.-Q. Song, A. Tchana and H. Waeselynck. *Section 06 Sciences de l’information : fondements de l’informatique, calculs, algorithmes, représentations, exploitations: Rapport de conjoncture 2024*. CNRS, 2025, pp. 1–20. URL: <https://inria.hal.science/hal-05238890> (cit. on p. 22).

#### Other scientific publications

- [44] P. Senellart. ‘Qualitative Evaluation of Academic Researchers in Computer Science: Practices and Reflections from CNRS’. In: *Computer Education* 12.372 (2025), pp. 36–41. DOI: [10.16512/j.cnki.jsjy.2025.12.012](https://doi.org/10.16512/j.cnki.jsjy.2025.12.012). URL: <https://inria.hal.science/hal-05421318> (cit. on p. 22).
- [45] M. Soric. ‘Understanding and Extracting Table Information from BRGM Documents’. Ecole centrale de Lyon, Feb. 2025. URL: <https://inria.hal.science/hal-04974179> (cit. on p. 24).

### Scientific popularization

- [46] R. Ronfard, C. Truchet and S. Abiteboul. ‘Informatique théâtrale’. In: *Binaire* (17th Jan. 2025), pp. 1–12. URL: <https://inria.hal.science/hal-05328671> (cit. on p. 24).
- [47] R. Ronfard, C. Truchet and S. Abiteboul. ‘Régie, captation, mise en scène. . . Quand l’informatique s’invite au théâtre’. In: *The Conversation France* (17th Jan. 2025). URL: <https://inria.hal.science/hal-04956688> (cit. on p. 24).
- [48] P. Senellart and N. Vieira. ‘OnePSL30 : transformation’. In: *Collection numérique de l’AMUE, Agence de mutualisation des universités et établissements d’enseignement supérieur* 39 (Dec. 2025). URL: <https://inria.hal.science/hal-05424514> (cit. on p. 22).

### 11.3 Cited publications

- [49] S. Abiteboul, P. Buneman and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, 1999 (cit. on p. 7).
- [50] S. Abiteboul, R. Hull and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995. URL: <http://webdam.inria.fr/Alice/> (cit. on p. 7).
- [51] S. Abiteboul, I. Manolescu, P. Rigaux, M. Rousset and P. Senellart. *Web Data Management*. Cambridge University Press, 2011. URL: <http://webdam.inria.fr/Jorge> (cit. on p. 7).
- [52] 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](https://doi.org/10.1007/978-1-4614-1168-0_10). URL: [https://doi.org/10.1007/978-1-4614-1168-0\\_10](https://doi.org/10.1007/978-1-4614-1168-0_10) (cit. on p. 7).
- [53] A. Deshpande, Z. G. Ives and V. Raman. ‘Adaptive Query Processing’. In: *Foundations and Trends in Databases* 1.1 (2007), pp. 1–140. DOI: [10.1561/19000000001](https://doi.org/10.1561/19000000001). URL: <https://doi.org/10.1561/19000000001> (cit. on p. 7).
- [54] A. Y. Halevy. ‘Answering queries using views: A survey’. In: *VLDB J.* 10.4 (2001), pp. 270–294. DOI: [10.1007/s007780100054](https://doi.org/10.1007/s007780100054). URL: <https://doi.org/10.1007/s007780100054> (cit. on p. 7).
- [55] 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](https://doi.org/10.1145/371578.371598). URL: <http://doi.acm.org/10.1145/371578.371598> (cit. on p. 7).
- [56] M. T. Özsu and P. Valduriez. *Principles of Distributed Database Systems, Third Edition*. Springer, 2011. DOI: [10.1007/978-1-4419-8834-8](https://doi.org/10.1007/978-1-4419-8834-8). URL: <https://doi.org/10.1007/978-1-4419-8834-8> (cit. on p. 7).
- [57] B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012. DOI: [10.2200/S00429ED1V01Y201207AIM018](https://doi.org/10.2200/S00429ED1V01Y201207AIM018). URL: <https://doi.org/10.2200/S00429ED1V01Y201207AIM018> (cit. on p. 7).
- [58] 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> (cit. on p. 7).
- [59] 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](https://doi.org/10.1145/2505515.2505691). URL: <http://doi.acm.org/10.1145/2505515.2505691> (cit. on p. 7).