RESEARCH CENTRE

Inria Saclay Centre at Institut Polytechnique de Paris

IN PARTNERSHIP WITH:

Institut Polytechnique de Paris, CNRS

2024
ACTIVITY REPORT

Project-Team

CEDAR

Rich Data Exploration at Cloud Scale

IN COLLABORATION WITH: Laboratoire d'informatique de l'école polytechnique (LIX)

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team CEDAR

Creation of the Project-Team: 2018 April 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.3. – Distributed data

A3.1.6. – Query optimization

A3.1.7. – Open data

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

A3.1.9. – Database

A3.2.1. – Knowledge bases

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.3. – Data and knowledge analysis

A3.3.1. – On-line analytical processing

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A9.1. – Knowledge

A9.2. – Machine learning

Other research topics and application domains

B6.5. – Information systems

B8.5.1. – Participative democracy

B9.5.6. – Data science

B9.7.2. – Open data

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Ioana Manolescu [Team leader, INRIA, Senior Researcher]
- Oana-Denisa Balalau [INRIA, ISFP]
- Oana Goga [Inria, Senior Researcher, from Oct 2024, HDR]
- Oana Goga [CNRS, Researcher, until Sep 2024, HDR]
- Madhulika Mohanty [INRIA, Researcher, from Oct 2024]

Faculty Member

- Yanlei Diao [ECOLE POLY PALAISEAU]

Post-Doctoral Fellows

- Garima Gaur [INRIA, Post-Doctoral Fellow, from Feb 2024]
- Chadi Helwe [INRIA, Post-Doctoral Fellow, from Oct 2024]
- Guillaume Lachaud [ECOLE POLYTECHNIQUE]

PhD Students

- Nardjes Amieur [CNRS]
- Nelly Barret [INRIA, until Mar 2024]
- Abir Benzaamia [CNRS, from Mar 2024]
- Theo Bouganim [INRIA]
- Tom Calamai [AMUNDI TECHNOLOGY, CIFRE]
- Salim Chouaki [CNRS, until Sep 2024]
- Asmaa El Fraihi [CNRS]
- Qi Fan [ECOLE POLYTECHNIQUE, until Aug 2024]
- Vincent Jacob [ECOLE POLYTECHNIQUE, until Sep 2024]
- Hritika Kathuria [INRIA, from Oct 2024]
- Muhammad Khan [INRIA]
- Nazim Mezhoudi [Ecole Polytechnique, BNP Paribas, CIFRE, from Sep 2024]
- Kun Zhang [INRIA]

Technical Staff

- Ines Abdelaziz [INRIA, Engineer, from Dec 2024]
- Simon Ebel [INRIA, Engineer]
- Theo Galizzi [INRIA, Engineer]
- Madhulika Mohanty [INRIA, Engineer, until Sep 2024]
- Georgios Siachamis [INRIA, Engineer, from Oct 2024]

Interns and Apprentices

- Ines Abdelaziz [CNRS, Intern, until Jul 2024]
- Sarra Bendaho [LIX, Intern, until Jul 2024]
- Pablo Bertaud-Velten [INRIA, Intern, until Mar 2024]
- Minh Hoang Duong [INRIA, Intern, from Jun 2024 until Aug 2024]
- Mohammed Younes El Fraihi [CNRS, Intern, from Feb 2024 until Aug 2024]
- Nada Hanad [LIX, Intern, until Jul 2024]
- Ilyes Nourredine Hattabi [LIX, Intern, until Jul 2024]
- Jagan Mahadevan Koipallil [INRIA, Intern, from Jul 2024 until Sep 2024]
- Hiba Louzzani [LIX, Intern, until Jul 2024]
- Anis Mahmahi [LIX, Intern, until Jul 2024]
- Mija Pilkaite [FX CONSEIL, Intern, until Mar 2024]
- Brahim Saadi [CNRS, Intern, from Feb 2024 until Aug 2024]
- Flaviu-Cristian Verde [INRIA, Intern, from Jul 2024 until Sep 2024]

Administrative Assistant

- Michael Barbosa [INRIA]

External Collaborators

- Angelos Anadiotis [Oracle Switzerland]
- Alexandre Barlot [Radio France, from Jun 2024]
- Nelly Barret [ECOLE POLYT. MILAN, from May 2024]
- Antoine Deiana [Radio France]
- Helena Galhardas [Université de Lisbonne, from Sep 2024]
- Emilie Gautreau [Radio France]
- Samuel Guimaraes [CNRS, from Jul 2024]
- Samuel Guimaraes [CNRS, until Apr 2024]
- Stéphane Horel [LE MONDE, until Aug 2024]
- Chenghao Lyu [University of Massachusetts, Amherst, until Aug 2024]
- Adrien Maumy [Radio France]
- Fatemeh Nargesian [University of Rochester, from Apr 2024, USA]
- Thomas Pontillon [Radio France, from Apr 2024]
- Gerald Roux [Radio France]
- Prajna Upadhyay [BITS Hyderabad, India]
- Joanna Yakin [Radio France, from Apr 2024]

2 Overall objectives

Our research aims at **models, algorithms and tools for highly efficient, easy-to-use data and knowledge management**; throughout our research, **performance at scale** is a core concern, which we address, among other techniques, by designing algorithms for a **cloud (massively parallel)** setting. In addition, we explore and mine rich data via machine learning techniques. Our scientific contributions fall into four interconnected areas:

Optimization and performance at scale. This topic is at the heart of Y. Diao's ERC project "Big and Fast Data", which aims at optimization with performance guarantees for real-time data processing in the cloud. Machine learning techniques and multi-objective optimization are leveraged to build performance models for cloud data analytics. The same goal is shared by our work on the efficient evaluation of queries in dynamic knowledge bases.

Data discovery and exploration. Today's Big Data is complex; understanding and exploiting it is daunting, especially to novice users such as journalists or domain scientists. To help such users, in the AI Chair "SourcesSay: Intelligent Data Analysis and Interconnection in Digital Arenas", we explore efficient and keyword search techniques to find answers in the data its highly heterogeneous structure makes standard (e.g., SQL) queries inapplicable. Further, we propose novel data abstraction methods, which, given a dataset, automatically compute a simple, human-understandable model thereof. Finally, we study heterogeneous graph exploration, blending graph querying, and natural language summarization.

Natural language understanding for analyzing and supporting digital arenas. In this area, we are focused on new natural langauge processing tools and their applications to problems such as argumentation mining, factfulness evaluation and information extraction. We are interested in particular on applications with high social value, such as analysing public discourse with the goal of finding elements that could bias the world view of citizens, such as false claims, fallacious arguments, propaganda, or greenwashing.

Safeguarding information systems. Recent events have brought to light the easiness of using current online systems to propagate information (that is sometimes false) and that we are facing an information war. We create knowledge and technology in this area to make the online information space safer. In O. Goga's ERC project "Momentous: Measuring and Mitigating Risks of AI-driven Information Targeting", we seek to use AI for good to help fact-checking and journalists, we develop natural language processing techniques to detect malicious online content (e.g., propaganda, manipulation), and we develop measurement methodologies and controlled experiments to assess risks with online systems.

3 Research program

3.1 Multi-model querying

As the world's affairs get increasingly more digital, a large and varied set of data sources becomes available: they are either structured databases, such as government-gathered data (demographics, economics, taxes, elections), legal records, stock quotes for specific companies, un-structured or semi-structured, including in particular graph data, sometimes endowed with semantics (see e.g., the Linked Open Data cloud). Modern data management applications, such as data journalism, are eager to combine in innovative ways both static and dynamic information coming from structured, semi-structured, and unstructured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lengthy rigid cycle of data integration and consolidation in a warehouse. Thus, we need flexible tools allowing us to interconnect various kinds of data sources and query them together.

3.2 Exploratory querying of data graphs

Semantic graphs, including data and knowledge, are hard to apprehend for users due to the complexity of their structure and, often to their large volumes. To help tame this complexity, our research follows several avenues. First, we build compact summaries of Semantic Web (RDF) graphs suited for a first-sight interaction with the data. Second, we devise fully automated methods of exploring RDF graphs using interesting aggregate queries, which, when evaluated over a given input graph, yield interesting results (with interestingness understood in a formal, statistical sense). Third, we study the exploration of highly heterogeneous data graphs resulting from integrating structured, semi-structured, and unstructured (text) data. In this context, we develop data abstraction methods, showing the structure of any dataset to a novice user, as well as searching on the graph through (i) keyword queries and (ii) exploration leveraging graph structure and linguistic contents.

3.3 An unified framework for optimizing data analytics

Data analytics in the cloud has become an integral part of enterprise businesses. Big data analytics systems, however, still lack the ability to take user performance goals and budgetary constraints for a task collectively referred to as task objectives, and automatically configure an analytic job to achieve the objectives. Our goal is to develop a data analytics optimizer that can automatically determine a cluster configuration with a suitable number of cores and other runtime system parameters that best meet the task objectives. To achieve this, we also need to design a multi-objective optimizer that constructs a Pareto optimal set of job configurations for task-specific objectives and recommends new job configurations to best meet these objectives.

3.4 Elastic resource management for virtualized database engines

Database engines are migrating to the cloud to leverage the opportunities for efficient resource management by adapting to the variations and heterogeneity of the workloads. Resource management in a virtualized setting, like the cloud, must be enforced in a performance-efficient manner to avoid introducing overheads to the execution. We design elastic systems that change their configuration at runtime with minimal cost to adapt to the workload every time. Changes in the design include both different resource allocations and different data layouts. We consider different workloads, including transactional, analytical, and mixed, and we study the performance implications on different configurations to propose a set of adaptive algorithms.

3.5 Argumentation mining

Argumentation appears when we evaluate the validity of new ideas, convince an addressee, or solve a difference of opinion. An argument contains a statement to be validated (a proposition also called claim or conclusion), a set of backing propositions (called premises, which should be accepted ideas), and a logical connection between all the pieces of information presented that allows the inference of the conclusion. In our work, we focus on fallacious arguments, where evidence does not prove or disprove the claim, for example, in an "ad hominem" argument, a claim is declared false because the person making it has a character flaw. We study the impact of fallacies in online discussions and show the need for improving tools for their detection. In addition, we look into detecting verifiable claims made by politicians. We started a collaboration with RadioFrance and with Wikidébats, a debate platform focused on proving quality arguments for controversial topics.

3.6 Measuring and mitigating risks of AI-driven information targeting

We are witnessing a massive shift in the way people consume information. In the past, people had an active role in selecting the news they read. More recently, the information started to appear on people's social media feeds as a byproduct of one's social relations. We see a new shift brought by the emergence of online advertising platforms where third parties can pay ad platforms to show specific information to particular groups of people through paid targeted ads. AI-driven algorithms power these targeting technologies. Our goal is to study the risks with AI-driven information targeting at three levels: (1) human-level-in which conditions targeted information can influence an individual's beliefs; (2) algorithmic-level in which conditions AI-driven targeting algorithms can exploit people's vulnerabilities; and (3) platform-level are targeting technologies leading to biases in the quality of information different groups of people receive and assimilate. Then, we will use this understanding to propose protection mechanisms for platforms, regulators, and users.

4 Application domains

4.1 Cloud computing

Cloud computing services are strongly developing and more and more companies and institutions resort to running their computations in the cloud, in order to avoid the hassle of running their own infrastructure. Today's cloud service providers guarantee machine availabilities in their Service Level Agreement (SLA), without any guarantees on performance measures according to a specific cost budget. Running analytics on big data systems require the user not to only reserve the suitable cloud instances over which the big data system will be running, but also setting many system parameters like the degree of parallelism and granularity of scheduling. Chosing values for these parameters, and chosing cloud instances need to meet user objectives regarding latency, throughput and cost measures, which is a complex task if it is done manually by the user. Hence, we need need to transform cloud service models from availabily to user performance objective rises and leads to the problem of multi-objective optimization. Research carried out in the team within the ERC project "Big and Fast Data Analytics" aims to develop a novel optimization framework for providing guarantees on the performance while controlling the cost of data processing in the cloud.

4.2 Computational journalism

Modern journalism increasingly relies on content management technologies in order to represent, store, and query source data and media objects themselves. Writing news articles increasingly requires consulting several sources, interpreting their findings in context, and crossing links between related sources of information. CEDAR research results directly applicable to this area provide techniques and tools for rich Web content warehouse management. Within the SourcesSay AI Chair project, we work to devise concrete algorithms and platforms to help journalists perform their work better and/or faster. This work is in collaboration with the journalists from RadioFrance, the team Le vrai du faux.

4.3 Computational social science

Political discussions revolve around ideological conflicts that often split the audience into two opposing parties. Both parties try to win the argument by bringing forward information. However, often this information is misleading, and its dissemination employs propaganda techniques. We investigate the impact of propaganda in online forums and we study a particular type of propagandist content, the fallacious argument. We show that identifying such arguments remains a difficult task, but one of high importance because of the pervasiveness of this type of discourse. We also explore trends around the diffusion and consumption of propaganda and how this can impact or be a reflection of society.

4.4 Online targeted advertising

The enormous financial success of online advertising platforms is partially due to the precise targeting features they offer. Ad platforms collect large amounts of data on users and use powerful AI-driven algorithms to infer users' fine-grain interests and demographics, which they make available to advertisers to target users. For instance, advertisers can target groups of users as small as tens or hundreds and as specific as "people interested in anti-abortion movements that have a particular education level". Ad platforms also employ AI-driven targeting algorithms to predict how "relevant" ads are to particular groups of people to decide to whom to deliver them. While these targeting technologies are creating opportunities for businesses to reach interested parties and lead to economic growth, they also open the way for interested groups to use user's data to manipulate them by targeting messages that resonate with each user.

5 Social and environmental responsibility

5.1 Impact of research results

Our work on Big Data and AI techniques applied to data journalism and fact-checking have attracted attention beyond our community and was disseminated in general-audience settings, for instance through Ioana Manolescu's participation in invited talks at Sciences Po Paris and at the EU Joint Research Center (JRC) Workshop on Disinformation, and through invited keynotes, e.g., at GRADES-NDA-2024 and ISWC 2024.

Our work in the SourcesSay project (Section 8.1.1), on propaganda detection (Section 8.3.1), and on ad transparency (Section 8.2), goes towards making information sharing on the Web more transparent and more trustworthy.

5.2 Contribution to Diversity, Equity and Inclusion

Nelly Barret and Madhulika Mohanty co-lead the SCOUT action of the Diversity, Equity and Inclusion initiative (website) for the DB research community. This action provided a checklist of items to be checked before submitting a paper to promote and ensure more DEI-compliant papers. This will be integrated within the standard submission systems for DB conferences. This has led to the publication of [32].

6 Highlights of the year

6.1 Awards

The team is proud of the following awards received by the team members in 2024:

- Oana Goga received the CNRS Bronze Medal.
- Nelly Barret received the runner-up for Best PhD thesis [29] in Data Management awarded by the BDA Association in 2024.
- Oana Goga received the runner-up for the CNIL-Inria Award for Privacy Protection, for the article titled "Marketing to Children Through Online Targeted Advertising: Targeting Mechanisms and Legal Aspects" by Tinhinane Medjkoune, Oana Goga, Juliette Senechal in ACM Conference on Computer and Communications Security (CCS), November 2023.
- Asmaa El Fraihi, Nardjes Amieur and Oana Goga received the Andreas Pfitzmann Best Student Paper Award for their article titled "Client-side and Server-side Tracking on Meta: Effectiveness and Accuracy" at PETS 2024.
- The team has hired two permanent members through Inria national competition: Oana Goga (DR) and Madhulika Mohanty (CRCN).

7 New software, platforms, open data

7.1 New software

7.1.1 ConnectionLens

Name: Integration of heterogeneous data using information extraction

Keyword: Data analysis

Functional Description: ConnectionLens treats a set of heterogeneous, independently authored data sources as a single virtual graph, whereas nodes represent fine-granularity data items (relational tuples, attributes, key-value pairs, RDF, JSON or XML nodes. . .) and edges correspond either to structural connections (e.g., a tuple is in a database, an attribute is in a tuple, a JSON node has a parent. . .) or to similarity (sameAs) links. To further enrich the content journalists work with, we also apply entity extraction which enables to detect the people, organizations etc. mentioned in text, whether full-text or text snippets found e.g. in RDF or XML. ConnectionLens is thus capable of finding and exploiting connections present across heterogeneous data sources without requiring the user to specify any join predicate.

URL: https://team.inria.fr/cedar/connectionlens/

Publications: hal-02934277, hal-02904797, hal-01841009

Contact: Manolescu Ioana

7.1.2 Abstra

Name: Abstra: Toward Generic Abstractions for Data of Any Model

Keywords: Heterogeneous Data, Data Exploration, Data analysis, Databases, LOD - Linked open data

Functional Description: Abstra computes a description meant for humans, based on the idea that, regardless of the syntax or the data model, any dataset holds some collections of entities/records, that are possibly linked with relationships. Abstra relies on a common graph representation of any incoming dataset, it leverages Information Extraction to detect what the dataset is about, and relies on an original algorithm for selecting the core entity collections and their relations. Abstractions are shown both as HTML text and a lightweight Entity-Relationship diagram. A GUI also allows to tune the abstraction parameters and explore the dataset.

URL: https://team.inria.fr/cedar/projects/abstra/

Contact: Ioana Manolescu

7.1.3 StatCheck

Name: Fact-checking Multidimensional Statistic Claims in French

Keywords: Machine learning, Databases, Natural language processing, Software engineering

Scientific Description: To strengthen public trust and counter disinformation, computational fact-checking, leveraging digital data sources, attracts interest from the journalists and the computer science community. A particular class of interesting data sources comprises statistics, that is, numerical data compiled mostly by governments, administrations, and international organizations. Statistics are often multidimensional datasets, where multiple dimensions characterize one value, and the dimensions may be organized in hierarchies. This paper describes STATCHECK, a statistic fact-checking system jointly developed by the authors, which are either computer science researchers or fact-checking journalists working for a French-language media with a daily audience of more than 15 millions (aud, 2022). The technical novelty of STATCHECK is twofold: (i) we focus on multi-dimensional, complex-structure statistics, which have received little attention so far, despite their practical importance, and (ii) novel statistical claim extraction modules for French, an area where few resources exist. We validate the efficiency and quality of our system on large statistic datasets (hundreds of millions of facts), including the complete INSEE (French) and Eurostat (European Union) datasets, as well as French presidential election debates.

Functional Description: StatCheck firstly allows the collection of data for its operation. Two types of data are collected: statistical tables and posts from social networks: - Acquisition of statistical files on the site of referent organisations (INSEE, Eurostat) - Extraction of statistical tables from these files, and storage of the extracted tables - Acquisition of political tweets from a list of accounts

The application allows the detection, extraction and search of statistical facts: - Detection and extraction of statistical facts from Twitter posts (e.g. "Unemployment rate increased by 30% in 2023) - Search for statistical facts in our database. Display of the twenty most relevant statistical tables for a statistical fact - Automatic transcription of audio files to detect and extract transcripts of statistical facts.

Release Contributions: - Redesign of the user interface - Modification of the software architecture - Addition of audio transcription

URL: `https://cedar-rf.saclay.inria.fr/`

Publications: hal-01496700, hal-01745768, hal-02121389, hal-01915148, hal-03767992, hal-03791175

Contact: Ioana Manolescu

Participants: Tien Duc Cao, Ioana Manolescu, Xavier Tannier, Oana-Denisa Balalau, Simon Ebel, Theo Galizzi

7.1.4 ConnectionStudio

Keywords: Heterogeneous Data, Data Exploration

Functional Description: ConnectionStudio integrates highly heterogeneous data into graphs, enriched with extracted entities. Studio users can discover the entities in their data, navigate across connections between datasets, explore and query the data in many ways. The Studio currently supports: CSV, JSON, XML, RDF, text, property graphs, all Office formats, and PDF datasets.

ConnectionStudio is a novel front-end to ConnectionLens, Abstra and PathWays (see also the respective Web sites). Its own novel features are outlined in a CoopIS 2023 article.

URL: `https://connectionstudio.inria.fr/`

Contact: Ioana Manolescu

7.1.5 FactSpotter

Keywords: Factual Faithfulness, Text generation

Functional Description: We propose a new metric that correctly identifies factual faithfulness, i.e., given a triple (subject, predicate, object), it decides if the triple is present in a generated text. We show that our metric FactSpotter achieves the highest correlation with human annotations on data correctness, data coverage, and relevance. In addition, FactSpotter can be used as a plug-in feature to improve the factual faithfulness of existing models.

Contact: Kun Zhang

Partner: Ecole Polytechnique

7.1.6 PathWays

Name: PathWays: finding entity paths in heterogeneous data graphs

Keywords: Named entities, Data Journalism, Heterogeneous Data

Functional Description: PathWays models heteroegenous datasets in a graph (see ConnectionLens). To identify interesting paths in this graph, Pathways works on its (smaller) summary (see Abstra) for efficiency and optimisation. Then, it sorts paths by their potential interest (metric based on the entity found and the information diluation along the path) before evaluating them with the help of a new multi-query optimisation algorithm. Finally, PathWays shows the most interesting (evaluated) paths in the form of tables, wich are very easy to understanf for journalists who are at the initiative of this scenario.

URL: https://team.inria.fr/cedar/projects/pathways/

Contact: Ioana Manolescu

7.1.7 OpenIEEntity

Name: Open Information Extraction with Entity Focused Constraints

Keyword: Information extraction

Functional Description: This tool takes in input a sentence and outputs the facts contained in the sentence, in the format (subject,predicate,object).

Contact: Oana-Denisa Balalau

7.1.8 FactCheckBureau

Name: FactCheckBureau: Build Your Own Fact-Check Analysis Pipeline

Keywords: Fact Check Retireval, Fact-checking

Functional Description: FactCheckBurea is an end-to-end solution that enables researchers to easily and interactively design and evaluate Fact Check retrieval pipelines. Further, it provides a query interface for non-technical users to find relevant Fact Checks for the input query in the form of a key phrase, social media post, or an image.

URL: https://gitlab.inria.fr/cedar/factcheckbureau

Publication: hal-04684068

Contact: Ioana Manolescu

8 New results

8.1 Data management for analyzing digital arenas

8.1.1 Graph integration of heterogeneous data sources for data journalism

Participants: Oana-Denisa Balalau, Nelly Barret, Theo Bouganim, Simon Ebel, Theo Galizzi, Ioana Manolescu, Madhulika Mohanty.

Work carried within the ANR AI Chair SourcesSay project has focused on developing a platform for integrating arbitrary heterogeneous data into a graph, then exploring and querying that graph using simple, intuitive query interfaces. The main technical challenges are: (i) how to interconnect structured and semi-structured data sources? We address this through information extraction (when an entity appears in two data sources or two places in the same graph, we only create one node, thus interlinking the two locations) and through similarity comparisons7.1.1; (ii) how to find all connections between nodes matching specific search criteria, or certain keywords? The question is particularly challenging in our context since ConnectionLens graphs can be pretty large, and query answers can traverse edges in both directions(iii) how to convert this graph into standard graph data models like property graphs, etc.

In this context, the following new contributions have been brought:

1. **ConnectionStudio: a user-friendly data lake for data exploration.** ConnectionStudio [16, 19, 27] is a user-friendly data lake, ingesting structured, semi-structured and un-structured documents (XML, JSON, CSV, RDF, PG, PDFs and Office files). It provides users different ways to explore and query the underlying data. For instance, it provides a set of lake-level statistics about the entities found in the data and Entity-Relationship diagrams showing the datasets structures. Further, it

allows users to enumerate and browse a set of (interesting) paths connecting entities of interest in the data. It also provides a querying interface 7.1.4 to let users formulate their queries using few variables and inspect the data lake without having to write SQL queries. Finally, users can save, export and share results they produce in the lake, e.g., data journalists can share them in newsrooms.

2. **Computing Generic Abstractions from Application Datasets** The unprecedented creation, use and share of data have led to large, complex, more or less structured, and heterogeneous datasets in terms of model and schema. Therefore, users crucially need tools to consolidate and interact with sources gathered from various actors in order to compute practical results. Users of ConnectionLens7.1.1 who are not familiar with the graph formalism or who know very little about the structure and/or content of their datasets can quickly become lost. Therefore, in [21], we propose a novel generic method, named Abstra 7.1.2, to automatically generate data descriptions, in the form of diagrams like Entity-Relationship schemas, for structured and semi-structured data, such as JSON, XML, RDF and PG, without requiring to define new methods for each model.

3. **Keyword Query Batch Evaluation on Graphs** In this work, we consider the setting where many keyword queries are issued simultaneously. To solve them, one could launch an algorithm such as MoLESP [33] individually on each query. However, this approach may be inefficient, especially when some queries in the batch share common keywords. In such cases, computing the query results separately would repeat the exploration of paths starting from the common keywords. In this work, we develop algorithms for answering simultaneously several keyword queries without duplicating exploration work while utilising effective pruning techniques to stay practical with formally proven guarantees on the results.

4. **Turning Heterogeneous Data Into Property Graphs** Digital data proliferation has lead to the multiplication of data sources of various formats, organized in different ways, yet which may be used together. Following up on our ConnectionLens project which integrates any data sources into a simple graph, and Abstra which analyzes such graphs and casts them into Entity-Relationship diagrams, we have investigated the problem of automatically converting any (set of) datasets into a Property Graph. The benefits of such a conversion are: (a) a single format simplifies data processing, (b) Property Graphs tend to enable more efficient query processing by attaching attributes to nodes, which removes some of the joins needed by the processing of queries over RDF graphs; (c) there is now widespread research interest in Property Graphs, in particular leading to the development of a host of PG data storage and query processing systems. Towards this goal, we have described a preliminary approach for generating PG Schemas from Abstra data abstractions in [20]. Separately, we have proposed a method for interactively visualizing any data source based on the ConnectionLens graph conversion followed by abstraction and finally PG conversion [28].

ConnectionLens is available online at: ConnectionLens Gitlab repository, while ConnectionStudio is available at ConnectionStudio Gitlab repository.

8.1.2 Decentralized semantic data sharing with access control

Participants: Maxime Buron, Hritika Kathuria, Ioana Manolescu, Georgios Siachamis.

Data exchange and sharing is a common need in virtually all modern applications. To achieve interoperability among different, heterogeneous databases, graph databases, and in particular knowledge graphs, are preferred due to their flexibility, which enables them to describe different structures of the underlying databases. Ontology-Based Data Access (OBDA) is the name commonly given to data integration systems based on knowledge graphs and ontologies; they have been successfully deployed in a variety of applications. Our team is a partner in DXP (Data Exchange Platform), a collaborative project between several Inria teams and Amadeus, technology provided for the travel industry. Within DXP, we are working to develop scalable, decentralized, and secure OBDA mechanisms for exchanging data across the different partners involved in a travel application: providers of services such as transport and lodging, travel operators, individual travelers, etc.

8.1.3 Finding Subgraphs with Maximum Total Density and Limited Overlap in Weighted Hypergraphs

Participants: Oana-Denisa Balalau.

Finding dense subgraphs in large (hyper)graphs is a key primitive in a variety of real-world application domains, encompassing social network analytics, event detection, biology, and finance. In most such applications, one typically aims at finding several (possibly overlapping) dense subgraphs, which might correspond to communities in social networks or interesting events. While a large amount of work is devoted to finding a single densest subgraph, perhaps surprisingly, the problem of finding several dense subgraphs in weighted hypergraphs with limited overlap has not been studied in a principled way, to the best of our knowledge. In our work [12], we define and study a natural generalization of the densest subgraph problem in weighted hypergraphs, where the main goal is to find at most k subgraphs with maximum total aggregate density, while satisfying an upper bound on the pairwise weighted Jaccard coefficient, i.e., the ratio of weights of intersection divided by weights of union on two nodes sets of the subgraphs. After showing that such a problem is NP-Hard, we devise an efficient algorithm that comes with provable guarantees in some cases of interest, as well as, an efficient practical heuristic. Our extensive evaluation on large real-world hypergraphs confirms the efficiency and effectiveness of our algorithms.

8.1.4 FactCheckBureau: Build Your Own Fact-Check Analysis Pipeline

Participants: Oana-Denisa Balalau, Garima Gaur, Oana Goga, Ioana Manolescu.

Fact-checkers are overwhelmed by the volume of claims they need to pay attention to fight misinformation. Even once debunked, a claim may still be spread by people unaware that it is false, or it may be recycled as a source of inspiration by malicious users. Hence, the importance of fact-check retrieval (FCR) as a research problem: given a claim and a database of previous checks, find the checks relevant to the claim. Existing solutions addressing this problem rely on the strategy of retrieve and re-rank relevant documents. Leveraging this pattern, we have built FactCheckBureau [17], an end-to-end solution for quickly designing, evaluating and deploying FCR pipelines. FactCheckBureau 7.1.8 serves the dual purpose – on one hand, researchers can interactively built and analyze their FCR pipelines and, on the other hand, non-technical users can use the FC query interface to fetch relevant fact checks from our corpus. Along with the tool, we also present fact check corpus we have built, which can be used in further research to test fact-check retrieval tools.

8.1.5 STaR: Space and Time-aware Statistic Query Answering

Participants: Oana-Denisa Balalau, Simon Ebel, Helena Galhardas, Theo Galizzi, Ioana Manolescu.

High-quality data is essential for informed public debate. Highquality statistical data sources provide valuable reference information for verifying claims. To assist journalists and fact-checkers, user queries about specific claims should be automatically answered using statistical tables. However, the large number and variety of these sources make this task challenging. STaR [18] is a novel method for Space and Time-aware STatistic Retrieval, based on a user natural language query. STaR is deployed within our system StatCheck, which we developed and shared with fact-checking journalists. STaR improves the quality of statistic fact retrieval by treating space and time separately from the other parts of the statistics dataset. Specifically, we use them as dimensions of the data (and the query), and focus the linguistic part of our dataset search on the rich, varied language present in the data. Our demonstration uses statistic datasets from France, Europe, and a few beyond, allowing users to query and explore along space and time dimensions.

8.1.6 Structured Discourse Representation for Factual Consistency Verification

Participants: Oana-Denisa Balalau, Ioana Manolescu, Kun Zhang.

Analysing differences between how events are represented in different texts, or verifying if large language models hallucinate, requires the ability to compare two texts. For this, a more structured representation of language can help by identifying distinct facts, and the discourse relationships connecting them, in each text. In this work, we first present an approach for extracting structured atomic facts and the relations between them, and then we present how to leverage this information to compare two texts. In addition, we show that the task of evaluating if the relation between facts has changed is challenging for SOTA LLMs. We believe that this work advances research on sentence similarity, by grounding it in linguistics and decomposing it in explainable modules. We demonstrate its potential in evaluating the factual consistency of two downstream tasks: text generation and text summarization.

8.1.7 Open information extraction for scientific text

Participants: Oana-Denisa Balalau, Garima Gaur, Ioana Manolescu, Prajna Upadhyay.

A conflict of interest occurs when a person or an organization is in a position to take actions or decisions in their official capacity to gain personal benefits—for instance, an oil company funds environmental researchers who are also on the company's external advisory committee. In the interest of uncovering such incompatible situations, it is essential to have a knowledge base that captures relationships among different research studies, authors, and funding agencies. We work towards building such a knowledge base by leveraging the power of LLMs for the Open Information Extraction (OpenIE) from the scientific text. In particular, we focus on the acknowledgment sections or the metadata provided with the academic articles. We propose an end-to-end in-context learning-based framework to extract structured relationships from the unstructured text.

8.1.8 User intent modelling and information extraction in datalakes

Participants: Guillaume Lachaud, Fatemeh Nagesian.

Datalakes have the potential to help in many fields, with the ability to access rich and diverse information on a wide variety of topics. At the same time, the lack of structure inside a data lake make the joinability of the data a vital issue. Currently, data lakes remain vastly underexploited. In our work, we try to address this challenge in a two step process: (i) (ML for DB): we try to model user intent in an intermediate representation that can exploited by database systems; and (ii) (DB for ML): to answer a user query, we enrich the data with relevant information collected from the data lake. The user query can take the form of a machine learning model, such as a classification task.

8.2 Online targeted advertising

Participants: Ines Abdelaziz, Nardjes Amieur, Abir Benzaamia, Salim Chouaki, Asmaa El Fraihi, Oana Goga.

8.2.1 Client-side and Server-side Tracking on Meta: Effectiveness and Accuracy

Growing concern over digital privacy has led to the widespread use of tracking restriction tools, such as ad blockers, Virtual Private Networks (VPN), and privacy-focused web browsers. All major browser vendors have also deprecated, or plan to deprecate, third-party cookies to reduce tracking. Despite these efforts, advertising companies continuously innovate to overcome these restrictions. Recently, advertising platforms, like Meta, have been promoting server-side tracking solutions to bypass traditional browser-based tracking restrictions. This paper explores how server-side tracking technologies can link website visitors with their user accounts on Meta products. The goal is to assess the effectiveness and accuracy of employing this technology, as well as the effect of tracking restrictions on online tracking. Our methodology involves a series of experiments where we integrate Meta's client-side tracker (the Meta Pixel) and server-side technology (the Conversions API) on different web pages. We then drive traffic to these pages and evaluate the success rate of linking website visitors to their profiles on Meta products. Our findings show that Meta's server-side technology can match between 34% and 51% of website visitors to user profiles on Meta products using basic information like the visitor's IP address, user agent, and location data. This is comparable to Pixel-based user matching in optimal conditions (i.e., in the absence of tracking restrictions), which links between 42% and 61% of user profiles. Nevertheless, we see a considerable difference in accuracy: while the Pixel-based tracking achieves 100% accuracy, less than 65% of the profiles matched by server-side tracking are accurate. The findings of this work [23] are extremely relevant in the current context of online tracking, where we recently saw Google allowing advertisers to use device fingerprinting to track and advertise to users. This paper has won the Andreas Pfitzmann Best Student Paper Award.

8.2.2 A Comparative Study of News Exposure and Consumption On and Off Facebook

Very large online platforms such as Meta, Google, and X use algorithms to feed users posts that match their interests. These algorithms have come under public scrutiny, as they can bias the information users consume in unwanted ways and consequently impact the formation of their political opinions and voting decisions. Many scholars have questioned whether social media promote misinformation and lead to echo chambers. To contribute to this debate, this study contrasts news exposure on Facebook (with an expected stronger algorithmic component) with news consumption off Facebook (with an expected weaker algorithmic component and stronger user behavior component) through the following questions: (1) Are users exposed to more/less misinformation on Facebook compared with their off-platform misinformation consumption? (2) Is news exposure on Facebook more/less diverse than off-platform news consumption? (3) To what extent do socio-demographic and psychological factors influence misinformation exposure and consumption both on and off Facebook? (4) Is there a relationship between socio-demographic and psychological factors and news diversity on and off Facebook? and (5) Does off-platform news consumption impact users' exposure to misinformation on Facebook?

8.2.3 Analyzing Exposure and Consumption of News through Data Donations

Understanding how exposure to news on social media impacts public discourse and exacerbates political polarization is a significant endeavor in both computer and social sciences. Unfortunately, progress in this area is hampered by limited access to data due to the closed nature of social media platforms. Consequently, prior studies have been constrained to considering only fragments of users' news exposure and reactions. To overcome this obstacle, we present an innovative measurement [22] approach centered on donating personal data for scientific purposes, facilitated through a privacy-preserving tool that captures users' interactions with news on Facebook. This approach offers a nuanced perspective on users' news exposure and consumption, encompassing different types of news exposure: selective, incidental, algorithmic, and targeted, driven by the diverse underlying mechanisms governing news appearance on users' feeds. Our analysis of data from 472 participants based in the U.S. reveals several interesting findings. Furthermore, our study uncovers that users are open to engaging with news sources with opposite political ideology as long as these interactions are not visible to their immediate social circles. Overall, our study showcases the viability of data donation as a means to provide clarity to longstanding questions in this field, offering new perspectives on the intricate dynamics of social media news consumption and its effects.

8.2.4 NewsDB: An Automated Approach to Build an Extensive Database of Self-Proclaimed News Providers

The credibility of news obtained online has become a concern due to the ease with which individuals or groups can claim to be news publishers and share news-related content. Unfortunately, research on monitoring misleading information in the online news ecosystem is hindered because the community lacks a comprehensive and up-to-date list of social media pages and domains claiming to be news media. In this work, we propose an automated approach that uses Google's GNews API and Meta's CrowdTangle API (now replaced by Meta's Content Library API) to identify self-proclaimed news providers. Our method was able to discover 19k self-proclaimed news providers in the United States active in June 2022 and 23k active in October 2019. This is significantly more than the known 1,553 U.S.-based sources listed by Media Bias Fact Check and News Guard. We are currently compiling the largest database of self-proclaimed news providers worldwide. We have compiled 776 sources in the UK, 1629 in France, 1012 in Germany, 1188 in Italy, 2655 in Brazil, and 764 in India.

8.2.5 Privacy-oriented personalization of online services

The digital advertising industry has rapidly grown as a key revenue source not only for smaller content publishers and developers, such as mobile app creators, but also for tech giants like Google and Facebook. Over time, Advertisers have shifted towards behavioral advertising, where users are targeted based on their demographics, interests, and preferences. This approach relies heavily on extensive data collection to build user profiles, often conducted by third-party tracking scripts. But, these mechanisms have been repeatedly shown to come at the expense of user privacy with invasive tracking across websites. Therefore, web browsers are adopting privacy-preserving alternatives to existing advertising mechanisms, ensuring the advertising ecosystem can continue to operate effectively while preserving the user privacy. Third-Party Cookies (3PCs) have already been deprecated by browsers such as Safari, Brave, and Mozilla Firefox, while others, like Google Chrome, have announced plans to follow suit through a series of proposals under the "Privacy Sandbox Initiative". At this stage, advertisers and publishers still have doubts regarding the effectiveness of these new technologies in targeting users compared to traditional 3PCs tracking and how they will affect their business. Conversely, it remains uncertain to what extent these technologies protect users privacy and respect their choices. Our work aims to analyze and evaluate the effectiveness, as well as assess the privacy risks associated with these new privacy-preserving advertising technologies.

8.3 Improving the quality of public debate with AI

Participants: Oana-Denisa Balalau, Tom Calamai, Chadi Helwe.

8.3.1 Finding Conflicts of Opinion in Citizen Participation Platforms

Online citizen participation platforms are powerful democratic tools that allow large numbers of contributors to be involved in public decision-making. However, for large groups of contributors to collaborate, we need to provide tools for users and decision makers to navigate and understand high volumes of content. Towards this goal, we introduce in [15] an approach based on natural language processing to detect pairs of contradictory and equivalent proposals in online citizen participation contexts. We apply this approach on two major national citizen consultations: the République Numérique and Revenu Universel d'Activité consultations. We highlight the potential of our method in two use cases. First, our method is a high-quality tool for finding idea communities in online content. Second, we demonstrate that the method improves on the state-of-the-art for finding relevant complementary content for a user, by identifying new relevant views for 76% of the proposals test.

8.3.2 Navigating the Political Compass: Evaluating Multilingual LLMs Across Languages and Nationalities

Large Language Models (LLMs), particularly decoder models like GPT-4, Gemma, and LLaMA, have revolutionized applications ranging from creative writing and article summarization to political debate analysis. While these models excel in multilingual support, enabling accessibility to diverse user groups, concerns about inherent biases, including gender, stereotypical, and political biases, remain significant. This study specifically investigates political biases in multilingual LLMs, focusing on their tendencies toward particular political ideologies. We evaluate the political inclinations of these models using the Political Compass Test, examining the impact of model size, prompt language, and assigned nationality personas. Our findings reveal that the political biases of LLMs are influenced by these factors, highlighting the complex interplay between linguistic and cultural contexts in shaping model outputs.

8.3.3 Automatic detection of greenwashing

This work [25] focuses on the automatic detection of greenwashing using advanced NLP techniques, aiming to address the growing need for accountability in corporate communication. In collaboration with journalists, we explored real-world challenges and contextualized our efforts to ensure practical relevance. The project includes a comprehensive literature review, identifying key trends and gaps in existing approaches. A significant outcome is the development of a toolbox of NLP tools tailored for detecting greenwashing in corporate disclosures. Looking ahead, we aim to expand this research by developing a tool for analyzing corporate carbon transition plans, leveraging the capabilities of large language models (LLMs) to assess credibility and alignment with stated sustainability goals.

8.4 Efficient Big Data analytics

Participants: Angelos Anadiotis, Yanlei Diao, Qi Fan, Muhammad Khan, Guillaume Lachaud, Chenghao Lyu, Ioana Manolescu.

8.4.1 Efficient Version Space Algorithms for Human-in-the-loop Model Development

When active learning (AL) is applied to help the user develop a model on a large dataset through interactively presenting data instances for labeling, existing AL techniques often suffer from two main drawbacks: First, to reach high accuracy they may require hundreds of data instances to be labeled by the user, which is an onerous task for the user. Second, retrieving the next instance to label from a large dataset can be time- consuming, making it incompatible with the interactive nature of the human exploration process. To address these issues, we introduce a novel version-space-based active learner for kernel classifiers, which possesses strong theoretical guarantees on performance and efficient implementation in time and space. In addition, by leveraging additional insights obtained in the user labeling process, we can factorize the version space to perform active learning in a set of subspaces, which further reduces the user labeling effort. Evaluation results show that our algorithms [13] significantly outperform state-of-the-art version space strategies, as well as a recent factorization-aware algorithm, for model development over large data sets.

8.4.2 Efficient and robust active learning methods for interactive database exploration

There is an increasing gap between fast growth of data and the limited human ability to comprehend data. Consequently, therehas been a growing demand of data management tools that can bridge this gap and help the user retrieve high-value contentfrom data more effectively. In this work, we propose an interactive data exploration system as a new database service, usingan approach called “explore-by-example.” Our new system is designed to assist the user in performing highly effective dataexploration while reducing the human effort in the process. We cast the explore-by-example problem in a principled “activelearning” framework. However, traditional active learning suffers from two fundamental limitations: slow convergence andlack of robustness under label noise. To overcome the slow convergence and label noise problems, we bring the propertiesof important classes of database queries to bear on the design of new algorithms and optimizations for active learning-baseddatabase exploration. Evaluation results using real-world datasets and user interest patterns show that our new system, both inthe noise-free case and in the label noise case, significantly outperforms state-of-the-art active learning techniques and dataexploration systems in accuracy while achieving the desired efficiency for interactive data exploration.

8.4.3 A Spark Optimizer for Adaptive Fine-Grained Parameter Tuning

As Spark becomes a common big data analytics platform, its growing complexity makes automatic tuning of numerous parameters critical for performance. Our work on Spark parameter tuning is particularly motivated by two recent trends: Spark's Adaptive Query Execution (AQE) based on runtime statistics, and the increasingly popular Spark cloud deployments that make cost-performance rea- soning crucial for the end user. This paper presents our design of a Spark optimizer that controls all tunable parameters of each query in the new AQE architecture to explore its performance benefits and, at the same time, casts the tuning problem in the theoretically sound multi-objective optimization (MOO) setting to better adapt to user cost-performance preferences. To this end, we propose a novel hybrid compile-time/runtime approach [26] to multi-granularity tuning of diverse, correlated Spark parameters, as well as a suite of modeling and optimization techniques to solve the tuning problem in the MOO setting while meeting the stringent time constraint of 1-2 seconds for cloud use. Evaluation results using TPC-H and TPC-DS benchmarks demonstrate the superior performance of our approach: (i) When prioritizing latency, it achieves 63% and 65% reduction for TPC-H and TPC-DS, respectively, under an average solving time of 0.7-0.8 sec, outperforming the most competitive MOO method that reduces only 18-25% latency with 2.6-15 sec solving time. (ii) When shifting preferences between latency and cost, our approach dominates the solutions of alternative methods, exhibiting superior adaptability to varying preferences.

8.4.4 Scalable Analytics on Multi-Streams Dynamic Graphs

Several real-time applications rely on dynamic graphs to model and store data arriving from multiple streams. In addition to the high ingestion rate, the storage and query execution challenges are amplified in contexts where consistency should be considered when storing and querying the data. This work [30] addresses the challenges associated with multi-stream dynamic graph analytics. We propose a database design that can provide scalable storage and indexing, to support consistent read-only analytical queries (present and historical), in the presence of real-time dynamic graph updates that arrive continuously from multiple streams. This work has been accepted in VLDB 2025.

Participants: Ioana Manolescu, Oana Balalau, Yanlei Diao, Ghufran Khan, Maxime Buron, Hritika Kathuria, Georgios Siachamis.

9 Bilateral contracts and grants with industry

9.1 Bilateral Grants with Industry

Ioana Manolescu is involved in the BPI-funded project CodeCommons, in collaboration with the Software Heritage Foundation (SWF). The project started in October 2024, for a duration of two years. We work to generalize, enlarge, and enable the efficient processing of the world's largest repository of free software. CodeCommons funds the last year of PhD of Ghufran Khan.

Ioana Manolescu is involved in the BPI-funded project DXP (Data Exchange Project), with Amadeus, the international tourism services operator. We participate in this project in collaboration with Maxime Buron, former team member, now an Assistant Professor at UCA. Our contribution here is to devise an architecture for decentralized, access-controled data sharing, allowing tourism service providers and clients to exchange their information via Amadeus' platform. The project started in October 2024, for a duration of 4 years. It has lead to the hire of Hritika Kathuria (PhD student) and Georgios Siachamis (engineer).

Yanlei Diao has started a CIFRE project with the Data & AI Innovation Lab of BNP Paribas (PhD of Nazim Mezhoudi).

Oana Balalau is involved in a CIFRE project with Amundi (PhD of Tom Calamai).

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

MediumAI

Title: Responsible AI for Journalism

Duration: 2024 ->

Coordinator: Davide Ceolin (Davide.Ceolin@cwi.nl)

Partners:

- CWI Amsterdam (Pays-Bas)

Inria contact: Oana-Denisa Balalau

Summary: From recommender systems to large language models, data-driven AI tools have shown different forms of limitations and bias. Bias in AI tools may stem from multiple factors, including bias in the input data the AI tools are trained on, the algorithm and the individuals responsible for designing the AI tools, and bias in the evaluation and interpretation of AI tool outputs. Limitations are due to technical difficulties in achieving specific tasks. Media outlets use different algorithmic aids in their workflow: keyword extraction, entities and relations extractions, event extraction, sentiment analysis, automatic summarization, newsworthy story detection, semi-automatic production of news using text generation models, and search, among others. Given the importance of the media sector for our democracies, shortcomings in the tools they use could have severe consequences. Both Inria and CWI have partnerships with large media groups and can help them address bias and limitations in their AI workflows.

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

Davide Ceolin

Status researcher

Institution of origin: Human-Centered Data Analytics team, CWI

Country: Netherlands

Dates: Oct 1-3, 2024

Context of the visit: Associated team MediumAI

Mobility program/type of mobility: research stay

Manel Slokom

Status post-doc

Institution of origin: Human-Centered Data Analytics team, CWI

Country: Netherlands

Dates: Nov 18-22, 2024

Context of the visit: Associated team MediumAI

Mobility program/type of mobility: research stay

Sanne Vrijenhoek

Status PhD

Institution of origin: Human-Centered Data Analytics team, University of Amsterdam

Country: Netherlands

Dates: Nov 19-20, 2024

Context of the visit: Associated team MediumAI

Mobility program/type of mobility: research stay

10.2.2 Visits to international teams

Research stays abroad

Oana Balalau

Visited institution: Human-Centered Data Analytics team at CWI

Country: Netherlands

Dates: Dec 2-6, 2024

Context of the visit: Associated team MediumAI

Mobility program/type of mobility: research stay

Chadi Helwe

Visited institution: Human-Centered Data Analytics team at CWI

Country: Netherlands

Dates: Dec 2-6, 2024

Context of the visit: Associated team MediumAI

Mobility program/type of mobility: research stay

10.3 European initiatives

10.3.1 Horizon Europe

ELIAS

Participants: Ioana Manolescu, Oana Goga, Madhulika Mohanty, Garima Gaur.

ELIAS project on cordis.europa.eu

Title: European Lighthouse of AI for Sustainability

Duration: From September 1, 2023 to August 31, 2027

Partners:

- ECOLE POLYTECHNIQUE (EP), France
- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- ROBERT BOSCH KFT, Hungary
- BITDEFENDER SRL (Bitdefender), Romania
- ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS (CENTRE FOR RESEARCH AND TECHNOLOGY HELLAS CERTH), Greece
- THE UNIVERSITY OF MANCHESTER (UNIVERSITY OF MANCHESTER), United Kingdom
- ROBERT BOSCH GMBH (BOSCH), Germany
- INSTITUT JOZEF STEFAN (JSI), Slovenia
- INSTITUT POLYTECHNIQUE DE PARIS, France
- UNIVERSITAT DE VALENCIA (UVEG), Spain
- PROMETEIA SOCIETA PER AZIONI (Prometeia), Italy
- IBM IRELAND LIMITED, Ireland
- KOBENHAVNS UNIVERSITET (UCPH), Denmark
- AALTO KORKEAKOULUSAATIO SR (AALTO), Finland
- IDEAS NCBR SP Z O.O., Poland
- UMEA UNIVERSITET, Sweden
- INSTITUT MINES-TELECOM, France
- FONDAZIONE ISTITUTO ITALIANO DI TECNOLOGIA (IIT), Italy
- FONDATION DE L'INSTITUT DE RECHERCHE IDIAP (IDIAP), Switzerland
- EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH (ETH Zürich), Switzerland
- CESKE VYSOKE UCENI TECHNICKE V PRAZE (CVUT), Czechia
- FUNDACION DE LA COMUNITAT VALENCIANA UNIDAD ELLIS ALICANTE, Spain
- FONDAZIONE BRUNO KESSLER (FBK), Italy
- UNIVERSITATEA POLITEHNICA DIN BUCURESTI (POLITEHNICA UNIVERSITY FROM BUCHAREST), Romania
- POLITECNICO DI MILANO (POLIMI), Italy
- UNIVERSITE DE TOULOUSE (UNIVERSITE DE TOULOUSE), France
- UNIVERSITA DEGLI STUDI DI TRENTO (UNITN), Italy
- UNIVERSITA DEGLI STUDI DI MILANO (UMIL), Italy

- HASSO-PLATTNER-INSTITUT FUR DIGITAL ENGINEERING GGMBH (HPI), Germany
- ENGINEERING - INGEGNERIA INFORMATICA SPA (ENG), Italy
- EBERHARD KARLS UNIVERSITAET TUEBINGEN (UT), Germany
- UNIVERSITA DEGLI STUDI DI GENOVA (UNIGE), Italy
- MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV (MPG), Germany
- UNIVERSITA DEGLI STUDI DI MODENA E REGGIO EMILIA (UNIMORE), Italy
- UNIVERSITEIT VAN AMSTERDAM (UvA), Netherlands

Inria contact: Ioana Manolescu

Coordinator: Nicu Sebe

Summary: We live in a crucial historical moment, with tremendous challenges ahead, from climate change to the energy crisis. ELIAS emerges from the belief that AI will be a key discipline to help us tackle these challenges. At the same time, the development of AI entails deep ethical and societal concerns that need to be addressed. As for fundamental research, ELIAS will address key scientific questions about how AI can reduce computational costs, serves to model effects of policy decisions on society, and impacts individuals. ELIAS will strive for a deep integration of the fundamental research that takes place in academia and the more applications-focused research from industry.

ELIAS builds on and expands the highly successful and internationally recognized European

Laboratory for Learning and Intelligent Systems (ELLIS). ELIAS will further develop the excellence criteria and the pillars in ELLIS and implement actions that will support AI researchers and young talents at different stages of their careers. Furthermore, ELIAS will develop a Sciencentrepreneurship track, with the purpose of attracting and empowering talents at the interface of scientific innovation and business and establish original AI solutions that move towards a sustainable long-term future for our planet, contribute to a cohesive society, and respect individual rights.

The outcome of ELIAS will be to establish Europe as a leader in AI research in which impact on the environment, society and the individual are integral considerations during development. We will measure the success of this endeavor in terms of key indicators, including the number of new cross-institutional collaborations, the number of cross-disciplinary collaborations, the number of industry-academic partnerships, publications in top conferences and journals, patents, and the number of projects that have resulted in deployed technologies.

MOMENTOUS Oana Goga is the PI of ERC MOMENTOUS (Measuring and Mitigating Risks of AI-driven Information Targeting), 2022-2027

10.3.2 H2020 projects

Oana Goga is the Local PI for H2020 Trustaware (June 2021 - May 2024)

10.4 National initiatives

10.4.1 ANR

- Oana Goga is the local PI for LIX partner – ANR PRC 2022 – 2026 “FeedingBias: A multi-platform mixed-methods approach to news exposure on social media” (our part: 128,000 €)
- Oana Goga is the local PI for LIX partner – ANR PRCE 2021 – 2025 “PROPEOS: Privacy-oriented Personalization of Online Services” (our part: 202,720 €)
- Ioana Manolescu is the local PI for Inria Saclay in CQFD (2019-2024), an ANR project coordinated by F. Ulliana (U. Montpellier). Its research aims at investigating efficient data management methods for ontology-based access to heterogeneous databases (polystores).

- Ioana Manolescu is the PI of SourcesSay (2020-2024), an AI Chair funded by Agence Nationale de la Recherche and Direction Générale de l'Armement. The project goal is to interconnect data sources of any nature within digital arenas. In an arena, a dataset is stored, analyzed, enriched and connected, graph mining, machine learning, and visualization techniques, to build powerful data analysis tools.

10.4.2 Others

- Associated Inria team MediumAI was accepted for a period of 3 years (2024-2026), PI: Oana Balalau and Davide Ceolin (CWI, Amsterdam).
- A 2-years post-doc funding was obtained from Hi!Paris for Garima Gaur with an application by Ioana Manolescu, Oana Balalau, and Oana Goga.
- A 2-years post-doc funding was obtained from Inria for Chadi Helwe with an application by Oana Balalau and Davide Ceolin (CWI, Amsterdam).

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees Oana Goga and Ioana Manolescu co-organized Infox sur Seine, Paris in March 2024.

11.1.2 Scientific events: selection

Chair of conference program committees Ioana Manolescu was the Chair of the Demonstration track at ACM SIGMOD 2024, and of the Tutorial track at EDBT 2025.

Yanlei Diao is the Program Committee Co-Chair of VLDB 2026, where work for the review board and conference website started in October 2024.

Member of the conference program committees The team members have been part of the following program committees:

- Oana-Denisa Balalau- ICSWM 2024, BDA 2024
- Oana Goga- The Web Conference 2024, CCS 2024, ConPro 2024, IC2S2 2024, Infox sur Seine 2024
- Ioana Manolescu- ACM SIGMOD 2024, CIDR 2024
- Yanlei Diao- VLDB Industry Track 2024
- Madhulika Mohanty- CoDS-COMAD 2024 (Distinguished PC Member Award), SIGMOD Availability 2023-2024, WSDM 2024, ICDE (Demo) 2024, ICWE 2024, VLDB 2024, SIGIR (Demo) 2024, VLDB (Tutorial) 2024
- Garima Gaur- SIGMOD (Demo) 2024

Reviewer Guillaume Lachaud has served as an external reviewer for SIGMOD (Demo) 2024. Madhulika Mohanty has served as an external reviewer for SIGMOD 2024 and CIDR 2024.

11.1.3 Journal

Member of the editorial boards Yanlei Diao served as the Editor-in-Chief of PVLDB Volume 19.

Oana Goga served as an Associate Editor at Transactions on Privacy and Security.

Ioana Manolescu served as an Associate Editor for PVLDB Journal.

Reviewer - reviewing activities Kun Zhang has served as a reviewer for TKDE journal.

Garima Gaur and Madhulika Mohanty have served as reviewers for the VLDB Journal.

11.1.4 Invited talks

The team members have given invited talks at various venues to national and international audience.

Oana Balalau has delivered the following talks:

- Keynote at InfoxSurSeine 2024, Paris
- Presentation and round table at I-EXPO 2024, Paris
- Presentation at the European Cyber Week, in the Cyber cognitive threats seminar, Rennes
- Presentation at LIX seminar, Palaiseau
- Presentation at the AI & Social Sciences Seminar, CREST, Palaiseau

Oana Goga has delivered the following national (N) and international (I) talks:

- (N) Plenary talk at Journée Inria - Ministére des Armées, June 2024
- (N) Plenary talk at Congrès Annuel de la Société Informatique de France, June 2024
- (I) Plenary talk at CNIL Privacy Research Day, June 2024
- (N) Plenary talk at Academie des Sciences, May 2024
- (N) Panel at CPDP The Role of Research and Researchers in AI Governance, May 2024
- (I) Plenary talk Meeting of the OECD Working Party on Digital Economics, Measurement, and Analysis, April 2024
- (I) Plenary talk at Forum International de la Cybersécurité – FIC 2024, March 2024
- (I) Panel at OECD report release Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity, March 2024
- (I) Plenary talk at DSA and Platform Regulation Conference, Feb 2024
- (N) Plenary talk at Musée des Arts et Métiers – L'aventure des inventions, Feb 2024, general audience

Ioana Manolescu has delivered the following talks:

- Invited keynote at the ISWC 2024 conference: "Retrieve (and Leverage)the Inner Graph Behind the Data"
- Invited keynote at the GRADES 2024 workshop in conjunction with SIGMOD 2024: "The Real World is Graph-Structured: Retrieving Meaning from Heterogeneous Data" [14]
- Keynote at the Junior Data Science and Engineering Conference at U. Paris Saclay: "Data and AI techniques for Fact-checking and Investigative journalism"
- Invited talk at a Colloquium on Conflict of Interests at Sciences Po Paris, January 2024: "Intégration de données hétérogènes et détection de liens/conflits d'intérêt: Système ConnectionLens"
- Invited talk at the EU Joint Research Center (JRC) Workshop on Disinformation, September 2024: "Harnessing data and AI to catch and unmask fakes"

Yanlei Diao has delivered an invited keynote at AWS Machine Learning Workshop on "Transformers and Foundation Models for Big Data Analytics Systems", September 2024.

11.1.5 Leadership within the scientific community

Ioana Manolescu has been the president of the informal French Data Management Association (BDA).

11.1.6 Scientific expertise

Oana Goga has provided her scientific expertise under the following roles:

- Member of the Science Advisory Committee for the NSF-funded Mid-scale RI-1 project (a 15 million US dollars project): Observatory for online human and platform behavior (2024).
- Member of the OECD Task Force on Global Challenge to Build Trust in the Age of Generative AI (2024).
- External Expert with contract for the European Commission (DG Connect & ECAT), to advise on the Delegated Act addressing the implementation of the DSA Article 40 that gives vetted researchers access to data from very large online platforms and very large search engines (2024).

11.1.7 Research administration

Ioana Manolescu has contributed to the following:

- She represents Inria in the Comité Operationnel of Hi!Paris, an AI Pole of Excellency comprising IP Paris and HEC.
- She is an elected member of IP Paris' Comité Académique and serves on its Scientific Committee.
- She is the head of the Data Analytics and Machine Learning axis of research within LIX.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

Oana Balalau is a part-time (33%) assistant professor at Ecole Polytechnique, where she teaches in two courses:

- INF473G "Mining, learning and reasoning on Web Graphs", L3, Ecole Polytechnique
- INF583 "Systems for Big Data", M1, Ecole Polytechnique

Yanlei Diao holds a part-time (50%) position at Ecole Polytechnique. She teaches INF583 "Systems for Big Data ", M1, Ecole Polytechnique

Garima Gaur taught 6h TP for the course ECE5DA04TP "Big Graph Databases", M1/M2, Télécom

Ioana Manolescu is a part-time professor (50%) at Ecole Polytechnique. She taught:

- Courses, labs and TDs in CSC51053EP "Database Management Systems", M1, Ecole Polytechnique
- She is in charge of the M1 Internship program in Artificial Intelligence and Data Science (CSC51092EP).
- She is also in charge of the Artificial Intelligence and Data Science program (M1) at Ecole Polytechnique
- Further, Ioana Manolescu has taught 9h of lectures for the new course ECE5DA04TP "Big Graph Databases", M1/M2, Télécom.

Madhulika Mohanty has taught:

- 3h of lectures and 3h of TP for the course ECE5DA04TP "Big Graph Databases", M1/M2, Télécom.
- 28 h of lectures and 28h of labs for INF540 "Databases", M1, Ecole Polytechnique.
- 18h of TP for CSC51053EP "Database Management Systems", M1, Ecole Polytechnique

11.2.2 Supervision

PhD supervision The team has supervised the following PhDs:

1. Nardjes Amieur, February 2023 - December 2024, advised by Oana Goga
2. Nelly Barret, January 2024 - March 2024, advised by Ioana Manolescu
3. Abir Benzaamia, March 2024 - December 2024, advised by Oana Goga
4. Théo Bouganim, January 2024 - December 2024, advised by Ioana Manolescu and Emmanuel Pietriga
5. Tom Calamai, January 2024 - December 2024, advised by Fabian Suchanek and Oana Balalau
6. Salim Chouaki, November 2021 - December 2024, advised by Oana Goga
7. Asmaa Elfraihi, February 2023 - December 2024, advised by Oana Goga
8. Antoine Gauquier (at ENS Paris), January 2024 - December 2024, advised by Ioana Manolescu and Pierre Senellart
9. Hritika Kathuria, October 2024 - December 2024, advised by Ioana Manolescu and Maxime Buron (UCA)
10. Ghufran Muhammad Khan, Jan 2024 - December 2024, advised by Angelos Anadiotis and Ioana Manolescu
11. Kun Zhang, January 2024 - December 2024, advised by Ioana Manolescu and Oana Balalau
12. Vincent Jacob, January 2024 - September 2024, advised by Yanlei Diao
13. Qi Fan, January 2024 - September 2024, advised by Yanlei Diao
14. Nazim Mezhoudi, September 2024 - December 2024, co-advised by Yanlei Diao and Mariam Barry (BNP Paribas)

Postdocs The team has supervised the following postdocs:

1. Guillaume Lachaud, January 2024 - December 2024, advised by Yanlei Diao
2. Chadi Helwe, October 2024 - December 2024, advised by Oana Balalau and Davide Ceolin
3. Garima Gaur, February 2024 - December 2024, advised by Oana Balalau and Ioana Manolescu

Engineers The team has supervised the following engineers:

1. Ines Abdelaziz, November 2024 - December 2024, advised by Oana Goga
2. Simon Ebel, January - December 2024, supervised by Oana Balalau and Ioana Manolescu
3. Théo Galizzi, January - December 2024, supervised by Oana Balalau and Ioana Manolescu
4. Georgios Siachamis, October - December 2024, supervised by Ioana Manolescu and Maxime Buron (UCA)
5. Madhulika Mohanty, January - September 2024 supervised by Ioana Manolescu.

Interns The team has supervised the following interns:

1. Pablo Bertaud-Velten, Polytechnique Bachelor student, Jan. 2024 - Feb. 2024
2. Minh-Hoang Duong, Polytechnique Bachelor student, July 2024 - Aug. 2024
3. Marijan Soric (at ENS Paris), Ecole Centrale de Lyon, Sept 2024 - Dec. 2024
4. Samuel Guimaraes, visiting student from UFMG Brazil, Jan. 2024 – July 2024
5. Nada Hanad, ESI Alger (6 months internship for eng. school diploma), Jan. 2024 – July 2024
6. Sarra Bendaho, ESI Alger (6 months internship for eng. school diploma), Jan. 2024 – July 2024
7. Anis Mahmahi, ESI Alger (6 months internship for eng. school diploma), Jan. 2024 – July 2024
8. Noureddine Ilyes Hattabi, ESI Alger (6 months internship for eng. school diploma), Jan. 2024 – July 2024
9. Hiba Louzzani, ESI Alger (6 months internship for eng. school diploma), Jan. 2024 – July 2024
10. Ines Abdelaziz, ESI Alger (6 months internship for eng. school diploma), Jan. 2024 – July 2024
11. Younes El Fraihi, ESI Alger (6 months internship for eng. school diploma), Feb. 2024 – August 2024
12. Brahim Saadi, ESI Alger (6 months internship for eng. school diploma), Feb. 2024 – August 2024

Part-time projects The team has supervised the following part-time research projects:

1. Luca Fechete, L2 bachelor Ecole Polytechnique, advised by Oana Balalau, worked on "Mining Arguments Across Languages: A Study of Political Bias in Translated and Generated Debates", January - June 2024
2. Yining Chen, Zhicheng Hui, Paul-Marie Jasson, Maxime Roth, Yuming Zhang, engineering cycle L3 Ecole Polytechnique, advised by Oana Balalau, worked on "Analysing media discourse around climate change", September 2023 - May 2024
3. Tudor Enache, M1 student at Ecole Polytechnique, advised by Nelly Barret, Madhulika Mohanty and Ioana Manolescu, worked on "Automatically deriving a Property Graph schema from data abstractions", in October 2023-March 2024.
4. Shay Pripstein, visiting M1 student at Ecole Polytechnique, advised by Nelly Barret, Madhulika Mohanty and Ioana Manolescu, worked on "Focused Information Gathering from Knowledge Bases", in October 2023-Jan 2024.

11.2.3 Juries

The team members have served on various juries.

Oana Balalau was a:

- member of the recruitment committee for part-time teaching position for Ecole Polytechnique
- member of the recruitment comittee for assistant professor at Ecole Polytechnique
- part of the PhD defense committee of Hadi Abdine, Ecole Polytechnique

Oana Goga was the part of the following committees:

- Award comitee: CNIL-Inria Privacy Protection Award 2024
- Selection comitee: tenure committee for Mobin Javed at LUMS (2024), Assistant Professor for Telecom ParisTech (2024)
- Ph.D. Committee jury member: Paul Bouchaud (2024), Evan Dufraisse (September 2024), Alireza Mohammadinodooshan (June 2024), Naif Mehanna (May 2024), Moitree Basu (Jan 2024)

Ioana Manolescu reported on the Ph.D. of Alexandra Rogova (co-supervised by Leonid Libkin and Amélie Gheerbrant), U. Paris 6, defended in November 2024.

11.3 Popularization

11.3.1 Internal or external Inria responsibilities

- Oana Balalau is a member of Inria Saclay's Scientific Commission.
- Oana Goga is a member of Conseil Scientifique of Regalia (2024).
- Ioana Manolescu is an elected member of Inria's Comité d'Evaluation.

11.3.2 Scientific Interventions

- Oana Goga appeared in an interview for CNRS Bronze Medal Award: "Oana Goga récompensée pour l'étude des risques liés aux plateformes en ligne" in March 2024.
- Oana Goga delivered a seminar at Musée des Arts et Métiers – L'aventure des inventions in February 2024.
- Oana Balalau participated in "L'intelligence artificielle au service du bien public", for young students (*seconde*) during their Inria internship, 2024.
- The Franco-British Council in partnership with Inria and the Franco-British Data Society hosted a one-day seminar on "Artificial Intelligence and Information Warfare" in July 2024 at University College London. Simon Ebel and Theo Galizzi presented StatCheck and Garima Gaur presented FactCheckBureau. Garima Gaur also participated in a panel discussion with Benjamin Guedj (Director of Research at Inria, Professor of Machine Learning and Foundational AI at University College London, and Turing Fellow). The full report is available here.
- Nelly Barret delivered a talk on "User-friendly data lake exploration" at LIB (Université de Dijon) in Jan. 2024.
- Nelly Barret talked about "Des données au journalisme" at Lycée international de Palaiseau Paris Saclay (LIPPS), in Jan. 2024.
- Salim Chouaki delivered a talk on "Analyzing Risks with Information Exposure on Social Media" at National Taipei University (Taiwan) in November 2024.
- Madhulika Mohanty has delivered talks on "Effective Exploration of Graph-Structured Data" at LIFO, Orléans and at the BOREAL team of Inria at Montpellier in January 2024.

12 Scientific production

12.1 Major publications

[1] R. Alotaibi, D. Bursztyn, A. Deutsch, I. Manolescu and S. Zampetakis. 'Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue'. In: *SIGMOD 2019 - ACM SIGMOD International Conference on Management of Data*. Amsterdam, Netherlands, June 2019. URL: https://hal.inria.fr/hal-02070827.

[2] O. Balalau, S. Ebel, T. Galizzi, I. Manolescu, Q. Massonnat, A. Deiana, E. Gautreau, A. Krempf, T. Pontillon, G. Roux and J. Yakin. 'Fact-checking Multidimensional Statistic Claims in French'. In: TTO 2022 - Truth and Trust Online. Boston [Hybrid Event], United States, 12th Oct. 2022. URL: https://hal.science/hal-03791175.

[3] O. Balalau and R. Horincar. 'From the Stage to the Audience: Propaganda on Reddit'. In: EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics. Online, France, 19th Apr. 2021. URL: https://hal.inria.fr/hal-03351621.

[4] M. Buron, F. Goasdoué, I. Manolescu and M.-L. Mugnier. 'Reformulation-based query answering for RDF graphs with RDFS ontologies'. In: *ESWC 2019 - European Semantic Web Conference*. Portoroz, Slovenia, Mar. 2019. URL: https://hal.archives-ouvertes.fr/hal-02051413.

[5] D. Bursztyn, F. Goasdoué and I. Manolescu. ‘Teaching an RDBMS about ontological constraints’. In: *Very Large Data Bases*. New Delhi, India, Sept. 2016. URL: https://hal.inria.fr/hal-01354592.

[6] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu and X. Tannier. ‘A Content Management Perspective on Fact-Checking’. In: *The Web Conference 2018 - alternate paper tracks "Journalism, Misinformation and Fact Checking"*. Lyon, France, Apr. 2018, pp. 565–574. URL: https://hal.archives-ouvertes.fr/hal-01722666.

[7] S. Cebiric, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou and M. Zneika. ‘Summarizing Semantic Graphs: A Survey’. In: *The VLDB Journal* (2018). URL: https://hal.inria.fr/hal-01925496.

[8] Y. Diao, P. Guzewicz, I. Manolescu and M. Mazuran. ‘Spade: A Modular Framework for Analytical Exploration of RDF Graphs’. In: *VLDB 2019 - 45th International Conference on Very Large Data Bases*. Proceedings of the VLDB Endowment, Vol. 12, No. 12. Los Angeles, United States, Aug. 2019. DOI: 10.14778/3352063.3352101. URL: https://hal.inria.fr/hal-02152844.

[9] E. Huang, L. Peng, L. D. Palma, A. Abdelkafi, A. Liu and Y. Diao. ‘Optimization for active learning-based interactive database exploration’. In: *Proceedings of the VLDB Endowment (PVLDB)* 12.1 (Sept. 2018), pp. 71–84. DOI: 10.14778/3275536.3275542. URL: https://hal.inria.fr/hal-01969886.

[10] A. Roy, Y. Diao, U. Evani, A. Abhyankar, C. Howarth, R. Le Priol and T. Bloom. ‘Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study’. In: *SIGMOD ’17 Proceedings of the 2017 ACM International Conference on Management of Dat*. SIGMOD ’17 Proceedings of the 2017 ACM International Conference on Management of Data. SIGMOD ACM Special Interest Group on Management of Data. Chicago, Illinois, United States: ACM, May 2017, pp. 187–202. DOI: 10.1145/3035918.3064048. URL: https://hal.inria.fr/hal-01683398.

[11] S. Y. Sahai, O. Balalau and R. Horincar. ‘Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions’. In: ACL-IJCNLP 2021 - Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online, France, 2nd Aug. 2021. URL: https://hal.inria.fr/hal-03351649.

12.2 Publications of the year

International journals

[12] O. Balalau, F. Bonchi, T.-H. H. Chan, F. Gullo, M. Sozio and H. Xie. ‘Finding Subgraphs with Maximum Total Density and Limited Overlap in Weighted Hypergraphs’. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 18.4 (12th Feb. 2024), pp. 1–21. DOI: 10.1145/3639410. URL: https://inria.hal.science/hal-04870686 (cit. on p. 12).

[13] L. Di Palma, Y. Diao and A. Liu. ‘Efficient Version Space Algorithms for Human-in-the-loop Model Development’. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 18.3 (12th Jan. 2024), pp. 1–49. DOI: 10.1145/3637443. URL: https://inria.hal.science/hal-04414855 (cit. on p. 16).

Invited conferences

[14] I. Manolescu. ‘The Real World is Graph-Structured: Retrieving Meaning from Heterogeneous Data (invited keynote)’. In: GRADES-NDA 2024 - 7th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA). Santiago (CL), Chile, 14th June 2024. URL: https://hal.science/hal-04614511 (cit. on p. 23).

International peer-reviewed conferences

[15] W. Aboucaya, O. Balalau, R. Angarita and V. Issarny. ‘Finding Conflicts of Opinion in Citizen Participation Platforms’. In: 2024 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Bangkok, Thailand, 9th Dec. 2024, p. 8. URL: https://inria.hal.science/hal-04870985 (cit. on p. 15).

[16] O. Balalau, N. Barret, S. Ebel, T. Galizzi, I. Manolescu and M. Mohanty. ‘Graph lenses over any data: the ConnectionLens experience’. In: ICDE 2024 - 40th IEEE International Conference on Data Engineering. Utrecht, Netherlands, 13th May 2024. URL: https://inria.hal.science/hal-04591897 (cit. on p. 10).

[17] O. Balalau, P. Bertaud-Velten, Y. El-Fraihi, G. Gaur, O. Goga, S. Guimaraes, I. Manolescu and B. Saadi. ‘FactCheckBureau: Build Your Own Fact-Check Analysis Pipeline’. In: CIKM 2024 - 33rd ACM International Conference on Information and Knowledge Management. Boise Idaho, United States, 23rd Oct. 2024. DOI: 10.1145/3627673.3679220. URL: https://inria.hal.science/hal-04684068 (cit. on p. 12).

[18] O. Balalau, S. Ebel, H. Galhardas, T. Galizzi and I. Manolescu. ‘STaR: Space and Time-aware Statistic Query Answering’. In: *ACM Conference on Information and Knowledge Management (CIKM)*. CIKM 2024 - 33rd ACM International Conference on Information and Knowledge Management. Boise, Idaho, United States, 24th Oct. 2024. DOI: 10.1145/3627673.3679209. URL: https://hal.science/hal-04689206 (cit. on p. 12).

[19] N. Barret, S. Ebel, T. Galizzi, I. Manolescu and M. Mohanty. ‘Exploration utilisateur de lacs de données très hétérogènes’. In: EGC 2024 - 24ème conférence francophone sur l’Extraction et la Gestion des Connaissances. Dijon, France, 24th Jan. 2024. URL: https://hal.science/hal-04388609 (cit. on p. 10).

[20] N. Barret, T. Enache, I. Manolescu and M. Mohanty. ‘Finding the PG schema of any (semi)structured dataset: a tale of graphs and abstraction’. In: SEAGraph Workshop 2024 - 3rd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores - 40th IEEE International Conference on Data Engineering (ICDE 2024). Utrecht, Netherlands, 2024. URL: https://inria.hal.science/hal-04591933 (cit. on p. 11).

[21] N. Barret, I. Manolescu and P. Upadhyay. ‘Computing Generic Abstractions from Application Datasets’. In: *OpenProceedings*. EDBT 2024 - 27th International Conference on Extending Database Technology. Vol. 27. Paestum, Italy, 25th Mar. 2024, pp. 94–107. URL: https://hal.science/hal-04131974 (cit. on p. 11).

[22] S. Chouaki, A. Chakraborty, O. Goga and S. Zannettou. ‘What News Do People Get on Social Media? Analyzing Exposure and Consumption of News through Data Donations’. In: WWW 2024 - The ACM International World Wide Web Conference. Singapore, Singapore: ACM, 13th May 2024, pp. 2371–2382. DOI: 10.1145/3589334.3645399. URL: https://inria.hal.science/hal-04618579 (cit. on p. 14).

[23] A. El fraihi, N. Amieur, W. Rudametkin and O. Goga. ‘Client-side and Server-side Tracking on Meta: Effectiveness and Accuracy’. In: *Proceedings on Privacy Enhancing Technologies*. PETS 2024 - 24th Privacy Enhancing Technologies Symposium. Vol. 2024. 3. Bristol, United Kingdom, July 2024, pp. 431–445. DOI: 10.56553/popets-2024-0086. URL: https://hal.science/hal-04665102 (cit. on p. 14).

[24] A. Gauquier. ‘Towards Efficient Construction of a Traceable, Multimodal, and Heterogeneous Data Warehouse’. In: *CEUR Workshop Proceedings*. VLDB 2024 PhD Workshop - The 50th International Conference on Very Large Data Bases. Guangzhou, China, 26th Aug. 2024. URL: https://inria.hal.science/hal-04617269.

[25] C. Helwe, T. Calamai, P.-H. Paris, C. Clavel and F. M. Suchanek. ‘MAFALDA: A Benchmark and Comprehensive Study of Fallacy Detection and Classification’. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL 2024 - North American Chapter of the Association for Computational Linguistics. Vol. 1: Long Papers. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 4810–4845. DOI: 10.18653/v1/2024.naacl-long.270. URL: https://hal.science/hal-04631163 (cit. on p. 16).

[26] C. Lyu, Q. Fan, P. Guyard and Y. Diao. ‘A Spark Optimizer for Adaptive, Fine-Grained Parameter Tuning’. In: VLDB 2024 - 50th International Conference on Very Large Databases. Guangzhou, China, 26th Aug. 2024. DOI: 10.14778/3681954.3682021. URL: https://inria.hal.science/hal-04906931 (cit. on p. 17).

Conferences without proceedings

[27] N. Barret, N. Dobričić, S. Ebel, T. Galizzi, I. Manolescu and M. Mohanty. ‘ConnectionStudio: User-friendly Exploration of Highly Heterogeneous Data Lakes’. In: BDA « Gestion de Données – Principes, Technologies et Applications ». Orleans (France), France, 21st Oct. 2024. URL: https://inria.hal.science/hal-04912842 (cit. on p. 10).

[28] T. Bouganim, I. Manolescu and E. Pietriga. ‘A Unified Visual Exploration Framework for (Semi-)structured Data’. In: BigVis 2024 - 7th International Workshop on Big Data Visual Exploration and Analytics (VLDB 2024). Guangzhou, China, 29th Aug. 2024. URL: https://hal.science/hal-04662129 (cit. on p. 11).

Doctoral dissertations and habilitation theses

[29] N. Barret. ‘User-oriented exploration of semi-structured datasets’. Institut Polytechnique de Paris, 15th Mar. 2024. URL: https://theses.hal.science/tel-04672899 (cit. on p. 7).

Reports & preprints

[30] A. C. Anadiotis, M. Ghufran Khan and I. Manolescu. *Dynamic Graph Databases with Out-of-order Updates (extended version)*. Institut Polytechnique de Paris; INRIA, 30th Oct. 2024. URL: https://hal.science/hal-04759818 (cit. on p. 17).

Scientific popularization

[31] B. Vanderborght, M. Colom, R. Binkytė, O. Balalau, O. Goga, H. Debar, M. Coupechoux and J. Herrera. *AIvolution - al and digital technologies in the European Union: Conclusions*. Ed. by E. Poptcheva. 1st Apr. 2024. URL: https://hal.science/hal-04645463.

12.3 Cited publications

[32] S. Amer-Yahia, D. Agrawal, Y. Amsterdamer, S. S. Bhowmick, R. Borovica-Gajic, J. Camacho-Rodríguez, J. Cao, B. Catania, P. K. Chrysanthis, C. Curino, A. El Abbadi, A. Floratou, J. Freire, S. Idreos, V. Kalogeraki, S. Maiyya, A. Meliou, M. Mohanty, F. Özcan, L. Peterfreund, S. Sahri, S. Sellami, R. Shraga, U. Sirin, W.-C. Tan, B. Thuraisingham, Y. Tian, G. Vargas-Solar, M. Zhang and W. Zhang. ‘Diversity, Equity and Inclusion Activities in Database Conferences: A 2023 Report’. In: *SIGMOD Rec.* 53.2 (July 2024), pp. 63–67. DOI: 10.1145/3685980.3685996. URL: https://doi.org/10.1145/3685980.3685996 (cit. on p. 7).

[33] A. C. Anadiotis, I. Manolescu and M. Mohanty. ‘Integrating Connection Search in Graph Queries’. In: *ICDE 2023 - 39th IEEE International Conference on Data Engineering*. Anaheim (CA), United States, Apr. 2023. URL: https://inria.hal.science/hal-04110779 (cit. on p. 11).