RESEARCH CENTRE

Saclay - Île-de-France

IN PARTNERSHIP WITH:

Ecole Polytechnique

2021 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

A3.3.1. – 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

B8.5.1. – Participative democracy

B9.5.6. – Data science

B9.7.2. – Open data

1 Team members, visitors, external collaborators

Research Scientists

- Ioana Manolescu [Team leader, Inria, Senior Researcher, HDR]
- Oana Balalau [Inria, ISFP]

Faculty Members

- Angelos Anadiotis [École polytechnique, Associate Professor]
- Yanlei Diao [École polytechnique, Professor]

Post-Doctoral Fellows

- Madhulika Mohanty [Inria, from Feb 2021]
- Le Ha Vy Nguyen [Institut Polytechnique de Paris, from Feb 2021 until Apr 2021]
- Fei Song [École polytechnique]

PhD Students

- Nelly Barret [Inria]
- Luciano Di Palma [École polytechnique, until Sep 2021]
- Qi Fan [École polytechnique]
- Pawel Guzewicz [École polytechnique, until Sep 2021]
- Mhd Yamen Haddad [Inria]
- Enhui Huang [École polytechnique, until Jun 2021]
- Vincent Jacob [École polytechnique]
- Muhammad Khan [Inria, from Oct 2021]
- Khaled Zaouk [École polytechnique, until Feb 2021]
- Kun Zhang [Inria, from Sep 2021]
- Fanzhi Zhu [Institut Polytechnique de Paris]

Technical Staff

- Francesco Chimienti [Inria, Engineer, from Mar 2021 until Sep 2021]
- Moustafa Latrache [Inria, Engineer, until Apr 2021]
- Tayeb Merabti [Inria, Engineer, Jan 2021]
- Arnab Sinha [École polytechnique, Engineer]
- Joffrey Thomas [CNRS, Engineer, Sep 2021]
- Prajna Devi Upadhyay [Inria, Engineer, from Mar 2021]

Interns and Apprentices

- Theo Bouganim [Inria, Apprentice]
- Abdenour Chaoui [Inria, from Apr 2021 until Sep 2021]
- Francesco Chimienti [Inria, until Feb 2021]
- Alexandre Hirsch [Inria, until Mar 2021]
- Ali Keshavarzi [École polytechnique, from Apr 2021 until Sep 2021]
- Lucas Maia Morais [Inria, from Apr 2021 until Aug 2021]
- Thomas Meunier [Inria, from Apr 2021 until Aug 2021]
- Kaylynn Pellicer [Ecole Polytechnique, until Mar 2021]
- Ashley Rakotoarisoa [Inria, from Apr 2021 until Jun 2021]
- Maya Touzari [Inria, from Jun 2021 until Sep 2021]

Administrative Assistant

- Alexandra Merlin [Inria]

External Collaborators

- Rana Alotaibi [UCSD]
- Stephane Horel [Le Monde]
- Roxana Horincar [Thales Research]
- Chenghao Lyu [École polytechnique, from Oct 2021]
- Tayeb Merabti [Univ de Provence]
- Saumya Yashmohini Sahai [Université d'État de l'Ohio - Columbus USA]
- Joffrey Thomas [École polytechnique, from Feb 2021 until Jun 2021]

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 in four interconnected areas:

Expressive models for new applications As data and knowledge applications keep extending to novel application areas, we work to devise appropriate data and knowledge models, endowed with formal semantics, to capture such applications' needs. This work mostly concerns the domains of data journalism and journalistic fact checking;

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-objectives optimization are leveraged to build performance models for data analytics the cloud. The same boal is shared by our work on efficient evaluation of queries in dynamic knowledge bases.

Data discovery and exploration Today's Big Data is complex; understanding and exploiting it is difficult. To help users, we explore: compact summaries of knowledge bases to abstrac their structure and help users formulate queries; interactive exploration of large relational databases; techniques for automatically discovering interesting information in knowledge bases; and keyword search techniques over Big Data sources.

Machine learning for text and graphs In this area, we are focused on new tools and their applications to problems such as questions answering and argumentation mining. Question answering is the task of automatically providing answers to user questions expressed in natural language via the exploration of a knowledge base. Argumentation mining is the task of extracting and modelling arguments from human discussions.

3 Research program

3.1 Scalable Heterogeneous Stores

Big Data applications increasingly involve *diverse* data sources, such as: structured or unstructured documents, data graphs, relational databases etc. and it is often impractical to load (consolidate) diverse data sources in a single repository. Instead, interesting data sources need to be exploited "as they are", with the added value of the data being realized especially through the ability to combine (join) together data from several sources. Systems capable of exploiting diverse Big Data in this fashion are usually termed *polystores*. A current limitation of polystores is that data stays captive of its original storage system, which may limit the data exploitation performance. We work to devise highly efficient storage systems for heterogeneous data across a variety of data stores.

3.2 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 un-structured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lenghy rigid cycle of data integration and consolidation in a warehouse. Thus, we see a need for flexible tools allowing to interconnect various kinds of data sources and to query them together.

3.3 Natural language question answering

We investigate methods for finding useful information in large datasets, to provide support for investigative journalism. Real-world events such as elections, public demonstrations, disclosures of illegal or surprising activities, etc. are mirrored in new data items being created and added to the global corpus of available information. Making sense of this wealth of data by providing a natural language question answering framework will facilitate the work of journalists.

3.4 Interactive Data Exploration at Scale

In the Big Data era we are faced with an increasing gap between the fast growth of data and the limited human ability to comprehend data. Consequently, there has been a growing demand of data management tools that can bridge this gap and help users retrieve high-value content from data more effectively. To respond to such user information needs, we aim to build interactive data exploration as a new database service, using an approach called "explore-by-example".

3.5 Exploratory Querying of Semantic Graphs

Semantic graphs including data and knowledge are hard to apprehend for users, due to the complexity of their structure and oftentimes to their large volumes. To help tame this complexity, in prior research (2014), we have presented a full framework for RDF data warehousing, specifically designed for heterogeneous and semantic-rich graphs. However, this framework still leaves to the users the burden of chosing the most interesting warehousing queries to ask. More user-friendly data management tools are needed, which help the user discover the interesting structure and information hidden within RDF graphs. This research has benefitted from the arrival in the team of Mirjana Mazuran, as well as from the start of the PhD thesis of Paweł Guzewicz, co-advised by Yanlei Diao and Ioana Manolescu.

3.6 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 come up with a data analytics optimizer that can automatically determine a cluster configuration with a suitable number of cores as well as 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.

3.7 Elastic resource management for virtualized database engines

Database engines are migrating to the cloud in order to leverage the opportunities for efficient resource management by adapting to the variations and the heterogeneity of the workloads. Resource management in a virtualized setting, like the cloud, need to be enforced in a performance-efficient manner in order to avoid introducing overheads to the execution.

We design elastic systems which change their configuration at runtime with minimal cost in order to adapt to the workload every time. Changes in the design include both different resource allocation and different data layouts. We consider different workloads including transactional, analytical and mixed and we study the performance implications on different configurations in order to eventually propose a set of adaptive algorithms.

3.8 Argumentation mining for social good

Harmful information is being spread online, in many forms, from fake news, to propaganda, and in particular fallacious arguments. 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 have started a collaboration with a online platform for debates on controversial topics, Wikidébats.

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's 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 ANR ContentCheck project, and following through the SourcesSay AI Chair, we work to devise concrete algorithms and platforms to help journalists perform their work better and/or faster. This work is in collaboration with journalists form Le Monde's: its fact-checking team "Les Décodeurs" was involved in ANR ContentCheck, while our current work continues in collaboration with award-winning investigative journalist Stéphane Horel.

4.3 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. False information, or fake news, has increasingly polluted public discourse. However, false information is just the tip of the iceberg of misleading information shared online and offline every day. In our work we focus on fallacial 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 discussion and we show the need for improving tools for its detection. In addition, we started a collaboration with the collaborative platform Wikidébats, a debate platform focused on proving quality arguments for controversial topics.

5 Social and environmental responsibility

5.1 Impact of research results

Our research has contributed to solving several concrete problems raised in the French public administration, through the projects realized within LabIA (Section 7.5).

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 I. Manolescu's participation in panels at *Festival des Idées*, at the Applied Mathematics seminar of INSA Rouen, and through the press, e.g., in a *Libération article on deep fakes*.

Our work in the SourcesSay project (Section 7.1.2) and on propaganda detection (Section 7.1.1) goes towards the goal of making information sharing on the Web more transparent and more trustworthy.

6 New software and platforms

6.1 New software

6.1.1 ConnectionLens

Keywords: Data management, Big data, Information extraction, Semantic Web

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

Authors: Tayeb Merabti, Helena Galhardas, Julien Leblay, Ioana Manolescu, Oana-Denisa Balalau, Catarina Pinto Conceicao

Contact: Manolescu Ioana

6.1.2 AIDEme

Keywords: Active Learning, Data Exploration

Scientific Description: AIDEme is a large-scale interactive data exploration system that is cast in a principled active learning (AL) framework: in this context, we consider the data content as a large set of records in a data source, and the user is interested in some of them but not all. In the data exploration process, the system allows the user to label a record as "interesting" or "not interesting" in each iteration, so that it can construct an increasingly-more-accurate model of the user interest. Active learning techniques are employed to select a new record from the unlabeled data source in each iteration for the user to label next in order to improve the model accuracy. Upon convergence, the model is run through the entire data source to retrieve all relevant records.

A challenge in building such a system is that existing active learning techniques experience slow convergence in learning the user interest when such exploration is performed on large datasets: for example, hundreds of labeled examples are needed to learn a user interest model over 6 attributes, as we showed using a digital sky survey of 1.9 million records. AIDEme employs a set of novel techniques to overcome the slow convergence problem:

• Factorization: We observe that a user labels a data record, her decision making process often can be broken into a set of smaller questions, and the answers to these questions can be combined to derive the final answer. This insight, formally modeled as a factorization structure, allows us to design new active learning algorithms, e.g., factorized version space algorithms [2], that break the learning problem into subproblems in a set of subspaces and perform active learning in each subspace, thereby significantly expediting convergence.

• Optimization based on class distribution: Another interesting observation is that when projecting the data space for exploration onto a subset of dimensions, the user interest pattern projected onto such a subspace often entails a convex object. When such a subspatial convex property holds, we introduce a new "dual-space model" (DSM) that builds not only a classification model from labeled examples, but also a polytope model of the data space that offers a more direct description of the areas known to be positive, areas known to be negative, and areas with unknown labels. We use both the classification model and the polytope model to predict unlabeled examples and choose the best example to label next. • Formal results on convergence: We further provide theoretical results on the convergence of our proposed techniques. Some of them can be used to detect convergence and terminate the exploration process. • Scaling to large datasets: In many applications the dataset may be too large to fit in memory. In this case, we introduce subsampling procedures and provide provable results that guarantee the performance of the model learned from the sample over the entire data source.

Functional Description: There is an increasing gap between fast growth of data and limited human ability to comprehend data. Consequently, there has been a growing demand for analytics tools that can bridge this gap and help the user retrieve high-value content from data. We introduce AIDEme, a scalable interactive data exploration system for efficiently learning a user interest pattern over a large dataset. The system is cast in a principled active learning (AL) framework, which iteratively presents strategically selected records for user labeling, thereby building an increasingly-more-accurate model of the user interest. However, a challenge in building such a system is that existing active learning techniques experience slow convergence when learning the user interest on large datasets. To overcome the problem, AIDEme explores properties of the user labeling process and the class distribution of observed data to design new active learning algorithms, which come with provable results on model accuracy, convergence, and approximation, and have evaluation results showing much improved convergence over existing AL methods while maintaining interactive speed.

Release Contributions: Project code can be found over: https://gitlab.inria.fr/ldipalma/aideme

URL: http://www.lix.polytechnique.fr/aideme

Contact: Yanlei Diao

Participants: Luciano Di Palma, Enhui Huang

6.1.3 ConnectionLensInMem

Keywords: Data management, Graph processing

Functional Description: In-memory graph-based keyword search. It works in collaboration with ConnectionLens and it focuses on parallelization of the query execution. The software includes a module to export a graph from ConnectionLens PostgreSQL warehouse which can then be loaded in the main memory for querying.

Contact: Angelos Anadiotis

6.1.4 Spade

Name: Efficient Exploration of Interesting Aggregates in RDF Graphs

Keywords: Data analytics, Data Exploration, RDF

Functional Description: To help users discover the most interesting facets of an RDF graph, we devised Spade, a tool for automatically identifying the k most interesting aggregate queries that can be evaluated on an RDF graph, given an integer k and a user-specified interestingness function. We implemented an extensible end-to-end framework that enables the identification and evaluation of interesting aggregates based on a new RDF-compatible one-pass algorithm for efficiently evaluating a lattice of aggregates, and a novel early-stop technique (with probabilistic guarantees) that can prune uninteresting aggregates.

Release Contributions: First deposit of Spade

URL: https://team.inria.fr/cedar/projects/spade-efficient-exploration-of-interesting-aggregates-in-rdf-graphs/

Publications: hal-02152844v2, hal-03320929v1

Contact: Manolescu Ioana

Partner: Ecole Polytechnique

6.1.5 RDFQuotient

Name: Quotient summaries of RDF graphs

Keywords: RDF, Graph algorithmics, Graph visualization, Graph summaries, Semantic Web

Functional Description: RDF graphs can be large and heterogeneous, making it hard for users to get acquainted with a new graph and understand whether it may have interesting information. To help users figure it out, we have devised novel equivalence relations among RDF nodes, capable of recognizing them as equivalent (and thus, summarize them together) despite the heterogeneity often exhibited by their incoming and outgoing node properties. From these relations, we derive four novel summaries, called Weak, Strong, Typed Weak and Typed Strong, and show how to obtain from them compact and enticing visualizations.

Publications: hal-01325900v6, hal-01808737

Contact: Manolescu Ioana

Participants: Ioana Manolescu, Pawel Guzewicz, François Goasdoué

Partner: Université de Rennes 1

6.1.6 Butterfly

Keywords: Data management, Databases, Graph processing

Functional Description: Integrated system for data science workload processing. Butterfly includes operators for relational and graph processing, as well as different data layouts and execution models.

Contact: Angelos Anadiotis

6.1.7 Exathlon

Keywords: Anomaly detection, Explainability, Time Series, Machine learning, Deep learning

Functional Description: A Pipeline for Explainable Anomaly Detection over Time Series

URL: `https://github.com/exathlonbenchmark`

Publication: hal-03381732

Contact: Vincent Jacob

6.1.8 FallacyMining

Keyword: Argument mining

Functional Description: People debate on a variety of topics on online platforms such as Reddit, or Facebook. Debates can be lengthy, with users exchanging a wealth of information and opinions. However, conversations do not always go smoothly, and users sometimes engage in unsound argumentation techniques to prove a claim. These techniques are called fallacies. Fallacies are persuasive arguments that provide insufficient or incorrect evidence to support the claim. We construct a new annotated dataset of fallacies, using user comments containing fallacy mentions as noisy labels, and cleaning the data via crowdsourcing. Finally, we study the task of classifying fallacies using neural models. We find that generally the models perform better in the presence of conversational context. We have released the data and the code.

Contact: Oana-Denisa Balalau

7 New results

7.1 Data management for analysing digital arenas

7.1.1 Graph integration of heterogeneous data sources for data journalism

Work carried within the ANR AI Chair SourcesSay projects has focused on developing a platform for integrating arbitrary heterogeneous data into a graph, then exploring and querying that graph in a simple, intuitive manner through keyword search. The main technical challenges are: (i) how to interconnect structured and semistructured 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 comparisons; (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.

Intense work has been invested in the ConnectionLens prototype, in particular due to the following contributions:

1. **Scaling the Grow and Aggressive Merge (GAM) algorithm**, which discovers connections among nodes in a graph [9]. The project considers scale-up, in-memory deployments and introduces Parallel GAM (P-GAM), an algorithm that parallelizes GAM operations by relying on novel, concurrent data structures. The algorithm has been implemented within a prototype graph processing engine, which scales across several threads under different graph topologies. At the same time, it provides interactive response times for real-world graphs [10, 14].

2. **Richer structured querying support for graph-structured data.** We have studied the problem of building an optimized graph query system supporting both structured and unstructured fragments in a single query. The structured fragment corresponds to a conjunctive query of edge patterns in the graph, whereas the unstructured fragment corresponds to querying for a (minimal) connected path/tree between input nodes. The state-of-the-art structured query systems only support reachability queries between any two nodes in a graph, with most of them requiring the specification of permitted edge labels. Our extended support removes these limitations to support applications requiring combining exact querying with general connectivity. For example, we can now ask, “How were the 2004 German, French and the US Presidents related to each other?” without specifying the expected connections. The work involved developing the grammar of the query language, a parser for the same and an execution engine to execute the query. We used javacc to generate the parser for our language. For the execution engine, we strategically used postgres for structured fragments and joins, and employed GAMSearch (also developed in the team before) for unstructured fragments. During building the system, we have improved the runtime of GAMSearch by almost 5x using algorithmic optimizations. In addition, we have identified the graph parameters affecting the system’s performance and used them to parameterize and develop a synthetic benchmark generator for the runtime experiments. Going forward, we will now focus on optimized execution of such queries exploiting the potential of shared computations in the unstructured fragments of the query.

3. **Improving the abstraction module of ConnectionLens**, which aims at creating small descriptions out of a dataset of any format (JSON, XML, RDF, property graphs etc.) to help researchers in the quest of the right dataset to be used in an experiment but also data providers which need to document/share their data in an informative way in the open data era. We leverage ConnectionLens to obtain a graph out of any input dataset. Then, our work is composed of four main parts: (1) summarise the graph to obtain a structural summary the graph (2) find records, possibly organised in collections, where a record is a thing/concept and a collection is a set of similar records (3) classify records and collections among a set of expressive categories (e.g. person, location, organisation, event) (4) present important facts of the abstraction to the user, i.e. largest collections, most frequent categories. We have implemented the abstraction module into ConnectionLens as a plug. We have done scalability experiments on three data formats (JSON, XML and RDF) to check that the code can handle large files. We are doing quality experiments, based on real-life data and use-cases.

Preliminary examples show the good quality of the abstraction. This work has been published as two papers to BDA 2021 [22, 21].

4. **Extracting relations between entities and its application to conflict of interests.** The ConnectionLens tool already uses an entity recognition module, so currently we are working on the adapting a relation extraction tool for the use cases that we are interested in, in particular, for extracting the different types of conflict of interests (COI) mentioned in the COI statements. To achieve this, we first extract triples of the form <subject. predicate, object> from COI statements in the Pubmed data. For instance, we want to extract the triples i) <Shang Gin Wu, has received speaking honoraria from, Roche> ii) <Shang Gin Wu, has received speaking honoraria from, AstraZeneca> iii) <Shang Gin Wu, has received speaking honoraria from, Pfizer> from a statement "Shang Gin Wu has received speaking honoraria from Roche , AstraZeneca and Pfizer ". Using tools such as OpenIE 6 [33] to extract triples may not always lead to the best extractions because they are domain-independent — they do not know the entities, such as names of the people or the organizations. Hence, we improve the quality of extractions by adding the knowledge about the entities to tools such as OpenIE6. These entities can be identified using entity recognition tools such as SNER or FLAIR. Adding knowledge about the entities involves imposing more constraints to identify the subject, predicate or object from a COI statement, such as, an entity cannot belong to the predicate, or that it has to appear in the subject or the object of the extraction. We implement such constraints in the OpenIE 6 neural network. We also used heuristics to identify sentences having a similar pattern to the sentences in the Pubmed data, modified their labels and re-trained OpenIE6 neural network. Our preliminary experiments show that the constraints help in retrieving better quality extractions. OpenIE6 also relies on a coordinate boundary detector module to split a conjunctive sentence into its constituent parts. We are also working towards improving the result of the coordinate boundary detection model by retraining the neural network with more data. This will further improve the quality of extractions from OpenIE6.

ConnectionLens is available online at: https://gitlab.inria.fr/cedar/connection-lens. This research has also lead to seven invited talks presented in Section 9.1.4.

7.1.2 Argumentation mining in online forums

The goal of this research is to understand online discussions better, and in particular why conversations can become highly contentious. We focused on political discussions, as politics reflect our societies. The first research direction was on propaganda, a very prevalent communication style in politics. Propaganda is defined as a communication strategy deliberately designed to influence the opinions or the actions of other individuals or groups concerning a predetermined end. Propaganda is an ensemble of strategies, which can be divided into invalid argumentation, when we try to support or attack a claim (invalid arguments are known as fallacies), and emotional manipulation techniques (see Propaganda techniques). Our goal was to understand how propaganda content is received on online forums. For this, we analyzed the impact of propaganda on six major political forums that target a diverse audience in two countries, the US and the UK. This work has been accepted as a long paper at EACL 2021 [15].

In the project's sequel, we focused on fallacy detection in social media. Fallacy detection is part of argumentation mining, a field that is concerned with computational tasks around argumentation. Fallacies do not appear only in the context of propaganda, but in any argumentative discussion, in which the goal is to determine the truth value of a claim. In this work, we investigated several neural architectures for classifying fallacies and we proposed a cost-efficient methodology for creating an annotated dataset of fallacies. This work was accepted as a long paper at ACL 2021 [19]

7.1.3 Machine learning for graph data

The first topic we explored in our work is **automatically generating suitable questions for existing knowledge bases**. A knowledge graph (KB) is represented by a set of triples, where each triple is composed of subject, predicate, and object. The purpose of our question generation model is to generate a set of answerable questions from a given knowledge graph KB, where each question corresponds to a subgraph of KB. In the first part of this work, we align Question-Answering datasets across similar KBs. In particular,

we have focused on aligning Freebase QA datasets to another general KG, YAGO4. This is especially interesting because (i) even though YAGO has been around for quite some time, there is a dearth of QA datasets on it, and, (ii) there are many QA datasets on Freebase but Google is no longer maintaining Freebase. At the core, this problem involves aligning the entities, classes and predicates of Freebase to those of YAGO4. We have used the paraphrase model of BERT for computing predicate matching with some success. Inspired from previous works, we have also developed a Greedy Matching algorithm for iteratively aligning the two KGs. Going forward, we will now be evaluating the performance of the Bert model, improving the results from the Greedy Matcher and finally generating a QA dataset on YAGO4.

In the second part of this work, we will train Transformers or Graph Neural Networks (GNNs) to generate questions from knowledge graphs, as it is very likely we will not be able to align every KB with Freebase. This approach is based on existing datasets that match questions to subgraphs (e.g. SimpleQuestions, GraphQuestions, GrailQA, etc.). The subgraphs in these datasets are usually from existing KBs such as FreeBase and WikiData. An important challenge of this approach is that the neural network usually fails to predict the correct sentences under zero-shot settings, i.e., when it encounters unseen predicates or entities in the test set. This problem is especially acute in the case of zero-shot predicates. In the future, we intend to use unsupervised methods for the neural network to learn to understand the unseen predicates and entities to help the question generation under zero-shot settings.

The second topic on which we worked is on **citation intent in citation networks**. Many metrics have been proposed over the years to understand influential authors and influential articles. However, recent advancements in NLP have relaunched the discussion on what makes a paper influential and how a scientific field evolves over the year. In particular, recent works have looked at the intent of authors when citing other papers, highlighting six functions of a citation: citation of background work - an article relevant for the field, a motivation work which illustrates the need for the current work, an article containing a method used in the current paper, a state-of-the-art competitor, an article which the work extends, or a paper that could be an inspiration for future work. In our work we predict the intent of a citation. We depart from previous literature that considered only the linguistic information in an article for the task. We incorporate more context by representing articles and citations as a heterogeneous graph, with node and edge types and labels, and predict new citations as new links in our graph.

7.2 Data exploration

7.2.1 Semantic graph exploration

In this area, the Ph.D. thesis of P. Guzerwicz has lead to two main contributions [17, 25].

Large (Linked) Open Data are increasingly shared as RDF graphs today. However, such data does not yet reach its full potential for sharing and reuse. Therefore, we provide new methods to meaningfully summarize data graphs, with a particular focus on RDF graphs. One class of tools for this task are structural RDF graph summaries, which allow users to grasp the different connections between RDF graph nodes. To this end, we introduce our novel RDFQuotient tool 6.1.5 that finds compact yet informative RDF graph summaries that can serve as first-sight visualizations of an RDF graph's structure.

We also consider the problem of automatically identifying the k most interesting aggregate queries that can be evaluated on an RDF graph, given an integer k and a user-specified interestingness function. Aggregate queries are routinely used to learn insights from relational data warehouses, and some prior research has addressed the problem of automatically recommending interesting aggregate queries. This work lead to the tool Spade 6.1.4.

In our recent and ongoing collaboration with Matteo Lissandrini and the Daisy team from Aalborg University, Denmark, we target "human-in-the-loop" data exploration. The starting point of this work is a limitation shared by all fully automated tools meant to help RDF graph data exploration, including RDFQuotient and Spade. Despite universal answers, which our systems provide to their respective problems, they do not consider user preferences. With this modified objective, we are currently designing a new framework. Our goal is to reverse-engineer a SPARQL aggregate query on an RDF graph based on a user input table. We begin by presenting an empty table to the user; then, we ask them to fill in some initial values for the cells and/or headers. Next, we propose some hints to the user to guide further exploration and find the target query. We require our future system to allow interactivity and leave as much flexibility to the user as possible.

7.2.2 A factorized version space algorithm for interactive database exploration

In a recent trend known as "Machine Learning for Everyone", IT companies deliver cloud platforms to help every data user develop machine learning models for their data sets with minimum effort. However, a key question is how to obtain a high-quality training data set for model development with minimum user effort. In other words, at the center of the new IT trend lies a critical "training data" problem. While industry solutions to this problem are limited to manual labeling or crowdsourcing, recent research on interactive data exploration (IDE) for model development bridges the gap between the data exploration and machine learning communities, and brings active learning-based data exploration to bear on the new process of model learning. In this setting, active learning is applied to select a small sequence of data instances for the user to label in order to derive an accurate model, while at the same time, offering interactive performance in presenting the next data instance for the user to review and label.

However, existing active learning techniques often fail to provide satisfactory performance when such models need to be built over large data sets. Not only do such models often require hundreds of labeled data instances to reach high accuracy (slow convergence), but retrieving the next instance to label can be time-consuming (inefficiency), making it incompatible with the interactive nature of the human exploration process. To address the slow convergence and inefficiency issues, we have developed two main ideas: First, we introduce a novel version space based active learning algorithm for kernel classifiers, which not only has strong theoretical guarantees on convergence, but also allows for an efficient implementation in time and space. Second, by leveraging additional insights obtained in the user exploration and labeling process, we explore a new opportunity to factorize an active learner so that active learning can be performed in a set of low-dimensional subspaces, which further expedites convergence and reduces the user labeling effort.

More specifically, we have developed the following contributions:

1. A new theoretical framework for version space (VS) algorithms over kernel classifiers: We developed a new theoretical framework that efficiently implements the Generalized Binary Search strategy over kernel classifiers, offering both strong theoretical guarantees on performance and an efficient implementation in time and space. We also proved generalization error bounds on accuracy and F-score, enabling our techniques to run over a sample from the original large data set with minimal performance loss.

2. Implementation and Optimizations: Based on our theoretical results, we devised an optimized VS algorithm called OptVS, which uses the hit-and-run algorithm for sampling the version space. However, hit-and-run may require thousands of iterations to output a high-quality sample, which can incur a high time cost. To reduce the cost, we develop a range of sampling optimizations to improve both sample quality and running time. In particular, we provide a highly efficient version of the rounding technique for improving the sample quality from the version space.

3. A Factorized Version Space Algorithm: Additionally, we developed a new algorithm that leverages the factorization structure provided by the user to create low-dimensional subspaces, and factorizes the version space accordingly to perform active learning in the subspaces. Compared to recent work that also used factorization for active learning, our work explores it in the new setting of VS algorithms and eliminates the strong assumptions made in prior work such as convexity of user interest patterns, resulting in significant performance improvement while increasing the applicability in real-world problems. We also managed to prove theoretical results on the optimality of our factorized VS algorithm.

Using real-world data sets and a large suite of user interest patterns, we have empirically observed that our optimized version space (VS) algorithms outperform existing VS algorithms, as well as DSM, a factorization-aware algorithm, often by a wide margin while maintaining interactive speed.

The results of our work are presented in the PhD thesis of Luciano Di Palma [24].

7.2.3 Learning with label noise

In active learning based data exploration, theory of active learning is applied to select a small sequence of data instances for the user to label in order to derive an accurate model, while at the same time,

offering interactive performance in presenting the next data instance for the user to review and label. Several algorithms that have been proposed in the literature to address the slow convergence and inefficiency problems, making the assumption that the user-provided labels are uncorrupted. However, practical experience shows that users may mislabel some examples. Given limited labeled examples in the interactive data exploration scenario, we improve the robustness of learning algorithms in the presence of label noise by (i) applying advanced methods to collect distilled examples out of noisy examples automatically, (ii) leveraging the polytope-based model learnt on distilled examples to further filter noisy labels, and (iii) developing new sample acquisition strategies that are less sensitive to label noise. Evaluation results using real-world datasets and user interest patterns show that our proposed algorithm is far more robust than alternative algorithms and it achieves desired accuracy and efficiency under different noise levels.

The results of our work are presented in the Ph.D. thesis of Enhui Huang [26].

7.3 Efficient Big Data analytics

7.3.1 Integrated graph-relational processing

We started this project by studying the literature about the Relational Database management system internals (query optimiser, query execution engine), in-memory databases (more specifically the column oriented database architecture like Monet db and C-store), and the different query execution paradigms most notably: the volcano model, the vectorization model (such as vectorwise) and the code generation model (such as Hyper). We studied the comparison between these different models and where each model has better performance than the other. Next, we worked on implementing a prototype of an in-memory column based execution engine called Butterfly 6.1.6. Butterfly can execute relational queries in a vectorized manner. It has been tested for correctness against the TPC-H benchmark. We are continuously working on enhancing its performance. We then added some graph algorithms to butterfly. We have two different approaches that we want to compare the graph algorithms on: Using a native graph representation by building an adjacency list (using CSR) and running native graph algorithm over this representation. Using the already existing butterfly relational execution engine and run the graph query as a recursive join query. We implemented both approaches in Butterfly and tested the implementation against another known benchmark for graph databases called SNB (Social Network Benchmark). The comparison leads us to propose the incremental CSR index building.

7.3.2 Distributed storage and indexing for large scale graph analytics

The goal of the project is to provide a set of scalable storage and indexing technologies, which should enable powerful analytics operations on top of graphs built by ingesting heterogeneous data sources. At this point, the work in this project focuses on reviewing the relevant literature as well as implementing and experimenting with existing approaches [32, 35, 34] to investigate their performance and functionality. The prototypes are designed for parallel deployments, and state-of-the art threading and memory allocation libraries are used to accomplish our project needs [35, 34].

7.3.3 Operator fusion for large scale complex analytics pipelines

In this work, we study data science pipelines that include a mix of relational and machine learning operators. During the last year, the project focused on the interaction between several relational database management systems, such as MySQL, Hive, and MonetDB and a machine learning framework, TensorFlow. By using either microbenchmarks, or the TPCx-BB benchmark, the project evaluated the cost of data transformations when data are exchanged between different systems and processing operators. The result of this work has given indicators on the bottlenecks that need to be addressed in order to scale mixed relational/machine-learning workflows.

7.3.4 Scaling up graph pattern matching

Graph pattern matching brings high computational complexity, as it is equivalent to the subgraph isomorphism problem. Accordingly, several heuristics have been proposed in the literature, which mostly

rely on graph partitioning, either statically performed in the beginning of the search, or dynamically adjusted at execution time. This project takes a scale-up approach and introduces concurrent, dynamic data structures that balance the workload at runtime among the worker threads and eliminate duplicate work, which is the main bottleneck in this setting. The preliminary results show that the proposed approach scales in several different datasets and workloads and performs better than the related work by several times. The project is currently ongoing.

The project is joint work with the EPFL.

7.3.5 GPU-accelerated scale-out data analytics

GPUs are becoming increasingly popular in modern data analytics. However, despite their obvious benefits stemming from their high computational throughput, they also bring complexity to the deployment of the system. Especially in a scale-out setting, there are several bottlenecks that need to be considered, like the network and the interconnects between the network controller, the GPU, the memory and the CPU. This project studies the impact of every interconnect with different workloads and proposes an end-to-end framework that achieves load balancing and high resource utilization of all the available compute devices in the system. The project is currently ongoing.

The project is joint work with the EPFL.

7.3.6 Multi-grained garbage collection in multi-version concurrency control

Most modern online transaction processing (OLTP) engines rely on multi-version concurrency control (MVCC) to synchronize access to the database by several parallel worker threads. MVCC brings increased concurrency in the presence of read-write workloads at the cost of increased storage use. Redundant storage allocations are thus removed during a garbage collection phase, which often becomes a bottleneck due to random data accesses. This project proposed a multi-grained storage scheme for storing the different versions introduced by MVCC. In doing so, it reduces and, even eliminates in some cases, the garbage collection cost. The project is currently ongoing.

7.3.7 View-based rewriting for hybrid (DB-ML) pipelines

Hybrid complex analytics workloads typically include (i) data management tasks (joins, filters, etc.), easily expressed using relational algebra (RA)-based languages, and (ii) complex analytics tasks (regressions, matrix decompositions, etc.), mostly expressed in linear algebra (LA) expressions. Such workloads are common in a number of areas, including scientific computing, web analytics, business recommendation, natural language processing, speech recognition. Existing solutions for evaluating hybrid complex analytics queries – ranging from LA-oriented systems, to relational systems (extended to handle LA operations), to hybrid systems – fail to provide a *unified optimization framework* for such a hybrid setting. These systems either optimize data management and complex analytics tasks separately, or exploit RA properties only while leaving LA-specific optimization opportunities unexplored. Finally, they are not able to exploit *precomputed (materialized) results* to avoid computing again (part of) a given mixed (LA and RA) computation. We devised HADAD, an *extensible lightweight approach for optimizing hybrid complex analytics queries*, based on a common abstraction that facilitates unified reasoning: a relational model endowed with integrity constraints, which can be used to express the properties of the two computation formalisms. Our approach enables full exploration of LA properties and rewrites, as well as semantic query optimization. Importantly, our approach does not require modifying the internals of the existing systems. Our experimental evaluation shows significant performance gains on diverse workloads, from LA-centered ones to hybrid ones [13]. This article has been also informally presented at the BDA conference 2021, where it received a Best Paper Award. A SIGMOD Repeatability submission has been made subsequently.

HADAD is available online: https://gitlab.inria.fr/cedar/hadad.

7.3.8 Boosting cloud data analytics using multi-objective optimization

In our work, we present a data analytics optimizer and a principled multi-objective optimization approach. This approach computes a Pareto optimal set of job configurations to reveal tradeoffs between different

user objectives, recommends a new job configuration that best explores such tradeoffs, and employs novel optimizations to enable such recommendations within a few seconds. Using benchmark workloads, our experimental results show that our MOO techniques outperform existing MOO methods in speed and coverage of the Pareto frontier, as well as simpler optimization methods, and can indeed recommend configurations that lead to better performance or explore interesting tradeoffs.

This work has been accepted to the proceedings of the ICDE 2021 for inclusion as a full paper [20].

7.3.9 Workload tuning using recommender systems

Spark is a widely popular massively parallel processing system which is used to execute various types of data analysis workloads. Accordingly, tuning its performance is a hard problem. In this research thread, we have casted the problem of tuning Spark workloads into a recommender systems framework.

In [28],[27], we have introduced representation learning architectures from three families of techniques: (i) encoder/decoder architectures; (ii) siamese neural networks; and (iii) a new family that combines the first two in hybrid architectures. These representation learning based techniques explicitly extract encodings from runtime traces before feeding them to a neural network dedicated for the end regression task on the runtime latency. We have covered deterministic and generative auto-encoders and proposed extensions of them in order to satisfy different desired encoding properties. We have also explained why the siamese neural networks are particularly interesting when a job is admitted with an arbitrary configuration and we have trained these architectures using two types of losses: a triplet loss and a soft nearest neighbor loss.

We have extended a previous benchmark of streaming workloads and sampled traces from these workloads as well as workloads from the TPCx-BB benchmark. We have provided comparative results between different modeling techniques and provided end-to-end comparative results with related work, which demonstrate the efficiency of our approach.

Our project is available online on Github.

7.4 Anomaly detection

7.4.1 Explainable Anomaly Detection benchmark

Anomaly Detection (AD) refers to the task of finding patterns in data that do not conform to expected behavior. Due to the ubiquity and complexity of time-oriented datasets, anomaly detection over time series data is widely studied and applied across modern systems. In such systems, the task of Explanation Discovery (ED), the ability to discover why given anomalies have occurred, is sometimes equally important to anomaly detection to enable faster corrective actions. Access to high-quality data repositories and benchmarks have been instrumental in advancing the state of the art in many research domains, and lack of such community resources severely limits scientific progress. There are several open datasets and benchmarks for time series analysis, but none focus on detecting and explaining anomalies in high dimension.

In this work, we introduced Exathlon, the first public benchmark for explainable anomaly detection over high-dimensional time series data. Exathlon consists of (1) a labeled dataset, (2) an evaluation framework for anomaly detection and explanation discovery, as well as a (3) modular and extensible pipeline to build and evaluate techniques. The dataset, code, and documentation for Exathlon are publicly available 6.1.7. This benchmark project led to two publications, one as an article and one as a demonstration, in the Research and Demonstration tracks of VLDB 2021 [11, 12].

7.4.2 DL-based feature selection for Anomaly Detection and explanation discovery

As enterprise information systems are collecting event streams from various sources, the ability of a system to automatically detect anomalous events and further provide human-readable explanations is of paramount importance. In this project, we aim to integrate anomaly detection and explanation discovery, which is able to leverage state-of-the-art Deep Learning (DL) techniques for anomaly detection, while being able to recover human-readable explanations for detected anomalies. At the core of the framework is a new human-interpretable dimensionality reduction (HIDR) method that not only reduces

the dimensionality of the data, but also maintains a meaningful mapping from the original features to the transformed low-dimensional features.

Such transformed features can be fed into any DL technique designed for anomaly detection, and the feature mapping will be used to recover human-readable explanations through a suite of new feature selection and explanation discovery methods.

Evaluation using a recent explainable anomaly detection benchmark demonstrates the efficiency and effectiveness of HIDR for AD, and the result that while all three recent ED techniques failed to generate quality explanations on high-dimensional data, our HIDR-based ED framework can enable them to generate explanations with dramatic improvements in the quality of explanations and computational efficiency.

Our work has been accepted for publication at the DEBS 2021 conference [18].

7.4.3 VIPR-based Anomaly Detection

As enterprise information systems are collecting event streams from various sources, the ability of a system to automatically detect anomalous events and further provide human-readable explanations is of paramount importance.

In this project, we aim to address anomaly detection and explanation discovery in a single, integrated system, which not only offers increased business intelligence, but also opens up opportunities for improved solutions. In particular, we propose several new loss terms, together with the VIPR framework to build such a system. The VIPR framework has the ability to identify the most important low-dimensional sub-projections. However, the original framework is only for the explanation discovery and will require the labeled data. Our current modification is to add the few-shot learning based anomaly detection on top of this framework.

7.5 LabIA collaborative projects

LabIA is a French government initiative to bring the benefits of artificial intelligence to public administrations. As part of this initiative, O. Balalau has co-supervised 3 master thesis together with several administrations:

- Question answering for the website service-publique.fr. Public administrations offer many platforms where citizens can find answers to their questions. This is the case, for example, of service-public.fr. When users cannot find the answer to their question directly, they can contact an agent to help them in their process, however, these requests can become time-consuming for public servants. The project aims at providing an automated response service to users' questions. We focused on text retrieval, that is given a query from a user, retrieve the most pertinent document from a dataset. While in the first part we focused on unsupervised techniques of text retrieval using standard metrics such as BM25, in this second part we focused on more recent approaches that leverage neural networks. Thesis manuscripts is available [29].

- Question answering for Hub'eau. Hub'eau gives access to APIs that can be used for information related to water sources, water quality, etc. Hub'eau was created by BRGM, Bureau de Recherches Géologiques et Minières, to allow citizens interested in environmental issues to query the currently available data. At the end of this project, we provided a software that mapped a question given in natural language to a correct query for the right API service. Thesis manuscripts is available [31].

- Detecting companies that are at risk of bankruptcy. Signaux Faibles[1] is a state start-up which has the goal of detecting using machine learning techniques the companies that are at risk of bankruptcy. In this project, we collaborated with the developers from Signaux Faibles to improve their model, whose prediction accuracy has been greatly affected by the current economic situation. In this work, we benchmarked many machine learning algorithms and did feature engineering to determine a good model to improve upon in future work. Thesis manuscripts is available [30].

[1] https://beta.gouv.fr/startups/signaux-faibles.html

Still as part of our involvement in LabIA, I. Manolescu has supervised three projects which we elicited through DINUM, and on which work 2nd (L3) and 3rd year (M1) students of Ecole Polytechnique. These projects are still ongoing in January 2022, as they span over two terms. Specifically:

1. The DARES (*Direction de l'Animation de la recherche, des Études et des Statistiques*) studies among other topics employment offers in order to form a vision of the French labor market. They work with multiple sest of job offers. In collaboration with Davide Buscaldi (U. Paris Nord and Ecole polytechnique), I. Manolescu supervises a project where Polytechnique M1 students (Yuhan Xiong, Michael Fotso Fotso) extract information from job announcements, classify the announcements according to the job qualifications, and extract a set of interesting fields.
2. HAS (*Haute Autorité de la Santé*) issues high-quality recommendations related to the French public health. These are available through an API as XML documents. A Polytechnique M1 student (Yunhao Yu) works to extract named entities from these documents, using ConnectionLens, and then to restructure the content into an RDF knowledge base. This will enable HAS to better reuse its own recommendations.
3. MESRI (*Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation*) builds a database of academic publications, which is fed automatically from a certain number of bibliographic sources. This database may contain duplicates when a publication is declared in more than one data source. Four L3 students (Aymen Echarghaoui, François Gatine, Changqing Liu et Biao Shi) work to deduplicate a very large publication dataset, by adapting the Hierarchical Heuristic Clustering algorithm to the specificities of the MESRI dataset.

8 Partnerships and cooperations

8.1 International research visitors

Other international visits to the team

Chenghao Lyu:

Visiting PhD

University of Massachusetts, Amherst, USA

Research stay from Oct 2021 - Dec 2021

8.1.1 Visits to international teams

Angelos Anadiotis holds a visiting professor position at EPFL.

8.2 National initiatives

8.2.1 ANR

- AIDE ("A New Database Service for Interactive Exploration on Big Data") is an ANR "Young Researcher" project led by Y. Diao, started at the end of 2016.
- CQFD (2019-2024) is an ANR project coordinated by F. Ulliana (U. Montpellier), in collaboration with U. Rennes 1 (F. Goasdoué), Inria Lille (P. Bourhis), Institut Mines Télécom (A. Amarilli), Inria Paris (M. Thomazo) and CNRS (M. Bienvenu). Its research aims at investigating efficient data management methods for ontology-based access to heterogeneous databases (polystores).
- SourcesSay (2020-2024) is 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.

8.2.2 Others

- The goal of the iCODA project is to develop the scientific and technological foundations for knowledge-mediated user-in-the-loop collaborative data analytics on heterogenous information sources, and to demonstrate the effectiveness of the approach in realistic, high-visibility use-cases. The project stands at the crossroad of multiple research fields—content analysis, data management, knowledge represen- tation, visualization—that span multiple Inria themes, and counts on a club of major press partners to define usage scenarios, provide data and demonstrate achievements. This is a project funded directly by Inria ("Inria Project Lab"), and is in collaboration with GraphIK, ILDA, LINKMEDIA (coordinator), as well as the press partners AFP, Le Monde (Les Décodeurs) and Ouest-France.

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

Chair of conference program committees:

- A. Anadiotis was the web chair of VLDB 2021.

9.1.2 Scientific events: selection

Chair of conference program committees

- O. Balalau was a co-chair of the Workshop Track of ICWSM 2021.
- I. Manolescu was a Program Chair for the International Conference of Web Engineering (ICWE) 2021.
- I. Manolescu was a Program Chair for the First International Conference on Artificial Intelligence and Machine Learning Systems (AI/ML Systems).
- I. Manolescu was a Tutorial Track Chair at VLDB 2021.

Member of the conference program committees I. Manolescu was a PC member for:

- ACM SIGMOD (Special Interest Group on the Management of Data) Conference 2021
- CIDR (Conference on Innovative Database Research) 2021
- EDBT (Extending Database Technologies) 2021
- FAccT (Fairness, Accountability, and Transparency) 2021
- TTO (Truth and Trust Online) 2021
- The Web Conference 2021 (yearly conference of the W3C)
- BDA (Bases de Données Avancées), 2021

A. Anadiotis was a PC member for:

- SIGMOD Reproducibility 2021
- SIGMOD Student Research Competition
- AIML Systems 2021 (The First International Conference on AI-ML-Systems)
- GLOBECOM 2021 (IEEE Global Communications Conference)

O. Balalau was a PC member for:

- ACL-IJCNLP 2021 (59th Annual Meeting of the Association for Computational Linguistcs and the 10th International Joint Conference on Natural Language Processing)
- EMNLP 2021 (2021 Conference on Empirical Methods in Natural Language Processing)
- The Web Conference 2022
- AIMLSystems 2021 (The First International Conference on AI-ML-Systems)
- AutoDS 2021 (ECMLPKDD Workshop on Automating Data Science)
- ICWSM 2021 (International Conference on Web and Social Media)
- BDA (Bases de Données Avancées), 2021

M. Mohanty was a PC member for:

- AIMLSystems 2021 (The First International Conference on AI-ML-Systems)
- CoDS-COMAD 2022 (ACM India Joint International Conference on Data Science and Management of Data)
- The Web Conference 2022

Reviewer

- O. Balalau, N. Barret, and M. Mohanty were external reviewers for ICWE 2021.
- P. Devi Upadhyay and M. Mohanty were external reviewers for CIDR 2022.
- P. Guzewicz was an external reviewer for SIGMOD 2021.
- Y. Diao was a jury member in the SWIFT Hachathon, Challenges 1 and 2, October 2021.

9.1.3 Journal

A. Anadiotis was a reviewer for the VLDB Journal and ACM Transactions on Database Systems. O. Balalau was a reviewer for the Journal of Affective Computing. M. Mohanty was a reviewer for Information Systems Journal.

9.1.4 Invited talks

Y. Diao has given invited talks:

- “UDAO: A Multi-Objective Optimizer for Cloud Data Analytics via Large-Scale Machine Learning”, Distinguished Lecture, Max Planck Institut (MPI) Informatik, December 16, 2021
- “Proactive Monitoring on High-Volume Event Streams through Large-Scale Machine Learning”, Keynote, 15th ACM International Conference on distributed and event-based systems (DEBS), 2021
- “Model Learning and Explanation Discovery for Exploring Large Datasets”, Distinguished Lecture, IBM Almaden Research Center, San Jose, April 30, 2021

I. Manolescu has given invited talks:

- “What do the Sources Say? Exploring Heterogeneous Journalistic Data As a Graph", invited talk at the DKM department of IRISA, March 18, 2021.
- "What do the Sources Say? Exploring Heterogeneous Journalistic Data As a Graph", keynote talk at DOLAP 2021, March 22, 2021.

- "Fact-checking and Computational Lead Finding in ContentCheck: from Data to the Press", invited keynote at the International Workshop on Knowledge Graphs for Online Discourse Analysis (KnoD), April 15, 2021
- "From fact-checking to Journalistic Data Integration", Invited talk in the “Trustworthy Data Science and AI” at U. Simon Fraser, Canada, on April 21, 2021.
- "What do the Sources Say? Exploring Heterogeneous Journalistic Data As a Graph", keynote of the LIG lab, May 6, 2021.
- "Data Integration for journalism: challenges and steps toward pragmatic solutions", invited keynote at the International Conference on Formal Concept Analysis (ICFCA), July 1, 2021.
- "Next-Generation Query Federation in the Cloud", invited talk at the Intelligent Data Systems session of the Huawei Cloud InnovWave conference, July 8, 2021.
- “What do the Sources Say? Interweaving Data Sources with the Help of AI for Journalistic Data Integration” at the “Intelligence Artificielle & Entreprises – Applications et défis mathématiques” meeeting at INSA Rouen, on Sept 30, 2021

M. Mohanty gave a talk and hands-on session titled “Effective exploration of graph-structured data” in the Karyashala workshop (28 June - 2 July, 2021) at IIT Dhanbad sponsored by Government of India, SERB through Accelerate Vigyan Scheme (AVS).

9.1.5 Leadership within the scientific community

Y. Diao is a member of the Proceedings of Very Large Databases Endowment Board. M. Mohanty is part of the D&I initiative in DB as a core member where she is co-leading the SCOUT action.

9.1.6 Research administration

- A. Anadiotis is a member of the Cloud and Intensive Computing task force of Inria Saclay–Île-de-France.
- I. Manolescu is the scientific director of LabIA, a collaboration between Inria and the DINUM focused on applying Artificial Intelligence techniques to address concrete problems encountered by public administrations. Oana Balalau is also devoting part of her activity to LabIA research projects.
- I. Manolescu is a member of the scientific committee of RegalIA, a pilot project of Inria aiming at algorithm regulation.
- I. Manolescu is responsible of the "Data Analytics and Machine Learning" research axis of LIX (Computer Science Lab of Ecole Polytechnique)

9.2 Teaching - Supervision - Juries

9.2.1 Teaching

A. Anadiotis is full-time Assistant Professor at Ecole Polytechnique, where he is in charge of two courses:

- Master: A. Anadiotis, “Systems for Big Data”, M1, Ecole Polytechnique
- Master: A. Anadiotis, “Systems for Big Data Analytics”, M2, Ecole Polytechnique

I. Manolescu is a part-time (50%) professor at Ecole Polytechnique, where she is in charge of two courses:

- Master: I. Manolescu, “Database Management Systems”, 45h, M1, École Polytechnique
- Master: I. Manolescu, "Research Internship in Data Science", 30h, M1, Ecole Polytechnique

I. Manolescu is also co-responsible of the "Data Science" M1 program of Ecole Polytechnique/Institut Polytechnique de Paris.

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

- Bachelor: “Graphe Global Géant”, L3, Ecole Polytechnique
- Master: “Systems for Big Data”, M1, Ecole Polytechnique

Team members also collaborate in teaching a course of Institut Polytechnique de Paris:

- P. Guzewicz, I. Manolescu, “TPT-DATAAI921 Architectures for Big Data”, Institut Polytechnique de Paris, January 2021.
- I. Manolescu, M. Mohanty, "TPT-DATAAI921 Architectures for Big Data", Institut Polytechnique de Paris, November-December 2021.

P. Guzewicz has also been serving as teaching assistant at Ecole Polytechnique in CSE204 Machine Learning (taught by J. Read and A. Ehrhardt). Y. Hadad has been a teaching assistant in INF553 Database Management Systems, taught by I. Manolescu.

9.2.2 Supervision

PhD supervision The team has supervised the following PhDs students:

- N. Barret, from January to December 2021 (I. Manolescu)
- P. Guzewicz, from January to September 2021 (Y. Diao 50%, I. Manolescu 50%)
- Y. Haddad, from January to December 2021 (A. Anadiotis 50%, I. Manolescu 50%)
- G. Khan, from October to December 2021 (A. Anadiotis 50%, I. Manolescu 50%)
- K. Zhang, from September to December 2021 (O. Balalau 50%, I. Manolescu 50%)
- E. Huang, from January to July 2021 (Y. Diao, A. Liu from U. Massachussets at Amherst USA)
- L. Di Palma, from January to June 2021 (Y. Diao, A. Liu from U. Massachussets at Amherst USA)
- K. Zaouk, from January to March 2021 (Y. Diao)
- Qi Fan, from January to December 2021 (Y. Diao)
- F. Zhu, from September to October 2021 (A. Anadiotis 50%, and Y. Diao 50%)
- P. Chrysogelos, PhD student at EPFL, PhD advisor: Anastasia Ailamaki, project-level technical supervision: A. Anadiotis
- A. Raza, PhD student at EPFL, PhD advisor: Anastasia Ailamaki, project-level technical supervision: A. Anadiotis
- S. Giannakopoulou, PhD student at EPFL, PhD advisor: Anastasia Ailamaki, project-level technical supervision: A. Anadiotis

Intern supervision The team has supervised the following interns:

- L. Maia Morais (M1), “Novel cost function and optimizations for graph search in ConnectionLens”, supervised by I. Manolescu
- A. Rakotoarisoa (L3), “Web applications for managing an internship Web site”, supervised by I. Manolescu
- M. Touzari (M1), “A chatbot for water quality data” (LabIA project), supervised by O. Balalau (80%) and I. Manolescu (20%).
- T. Meunier (M2), “Automatic or semi-automatic detection of companies in difficulty or weakened by the crisis” (LabIA project), supervised by O.Balalau (60%) and S. Lebastard (40%)
- A. Chaoui (M2), “Improving a Search Engine for Answering User Questions in Natural Language ” (LabIA project), supervised by O. Balalau (60%), R. Reynaud (20%) and P. Soriano (20%).
- A. Keshavarzi (M1), “Efficient model maintenance under data drift”, supervised by Y. Diao and Peter Haas.

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

- Yunhao Yu (M1 Ecole Polytechnique): “Knowledge graph construction from reports by Haute Autorité de la Santé” (I. Manolescu, project part of the team’s involvement in LabIA)
- A. Echargaoui, F. Gatine, C. Liu, B. Shi (L3 Ecole Polytechnique): “Deduplication of bibliographic records in MESRI bibliographic data” (I. Manolescu, project part of the team’s involvement from LabIA)
- P. Fotso-Fotso, Y. Xiong: “Information extraction from DARES employment offers” (I. Manolescu, project inspired from LabIA)
- K. Zheng (M1 Ecole Polytechnique), “Learning to rank in a knowledge base” (O. Balalau)
- D. Berrebbi and N. Huynh (M1 Ecole Polytechnique), “Citation intent in scientific articles” (O. Balalau)
- A. Keshavarzi (M2 IPP), “Argumentation Mining in a Low Resource Setting” (O.Balalau)

9.2.3 Juries

I. Manolescu has been a member of the following PhD juries:

- P. N. Sawadogo, "*Des lacs de données à l’analyse assistée de documents textuels et tabulaires*", Université de Lyon, October 2021
- A. Gillet, "*Modélisation et développement d’un observatoire générique pour la collecte et l’analyse de données massives*", Université de Bourgogne, December 2021

Y. Diao has been a member in the PhD jury of Pedro Timbó Holanda, CWI & Leiden University, September 2021.

O. Balalau has been a member in the PhD jury of Yang Qiu, "*Methods for optimizing customer prospecting in automated display advertising with Real-Time Bidding*", Ecole Polytechnique, October 2021.

9.3 Popularization

9.3.1 Articles and contents

- I. Manolescu was interviewed for an article by Lucie Ronfaut, on deep fakes and the role they may play in the upcoming French political campaign. The article appeared in Libération, on May 14, 2021.

9.3.2 Interventions

- “Festival des Idées” was held in Paris on November 18-20, 2021. I. Manolescu was a panelist in the session "#NousSachons : le futur de l’information en péril ?" , on Nov 19, 2021.
- I. Manolescu was a panelist in "What is AI?" at the yearly meeting of Association Française de Marketing (AFM) on Dec 1, 2021.

10 Scientific production

10.1 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

10.2 Publications of the year

International journals

[9] A. C. Anadiotis, O. Balalau, C. Conceicao, H. Galhardas, M. Y. Haddad, I. Manolescu, T. Merabti and J. You. ‘Graph integration of structured, semistructured and unstructured data for data journalism’. In: *Information Systems* (18th July 2021), p. 42. DOI: 10.1016/j.is.2021.101846. URL: https://hal.inria.fr/hal-03150441.

[10] A.-C. Anadiotis, O. Balalau, T. Bouganim, F. Chimienti, H. Galhardas, M. Y. Haddad, S. Horel, I. Manolescu and Y. Youssef. 'Empowering Investigative Journalism with Graph-based Heterogeneous Data Management'. In: *Bulletin of the Technical Committee on Data Engineering* (30th Sept. 2021). URL: https://hal.archives-ouvertes.fr/hal-03337650.

[11] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao and N. Tatbul. 'A Demonstration of the Exathlon Benchmarking Platform for Explainable Anomaly Detection'. In: *Proceedings of the VLDB Endowment (PVLDB)* (Aug. 2021). URL: https://hal.archives-ouvertes.fr/hal-03383535.

[12] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao and N. Tatbul. 'Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series'. In: *Proceedings of the VLDB Endowment (PVLDB)* (July 2021). URL: https://hal.archives-ouvertes.fr/hal-03381732.

International peer-reviewed conferences

[13] R. Alotaibi, B. Cautis, A. Deutsch and I. Manolescu. 'HADAD: A Lightweight Approach for Optimizing Hybrid Complex Analytics Queries'. In: ACM SIGMOD 2021 - International Conference on Management of Data. Xi'an / Online, China, 20th June 2021. URL: https://hal.inria.fr/hal-03347677.

[14] A. C. Anadiotis, O. Balalau, T. Bouganim, F. Chimienti, H. Galhardas, M. Y. Haddad, S. Horel, I. Manolescu and Y. Youssef. 'Discovering Conflicts of Interest across Heterogeneous Data Sources with ConnectionLens'. In: ACM International Conference on Information and Knowledge Management (CIKM 2021). Online, Australia, 1st Nov. 2021. DOI: 10.1145/3459637.3481982. URL: https://hal.inria.fr/hal-03337765.

[15] 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.

[16] 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.

[17] Y. Diao, P. Guzewicz, I. Manolescu and M. Mazuran. 'Efficient Exploration of Interesting Aggregates in RDF Graphs'. In: SIGMOD/PODS '21 - International Conference on Management of Data. Virtual Event China, China: ACM, 20th June 2021, pp. 392–404. DOI: 10.1145/3448016.3457307. URL: https://hal.inria.fr/hal-03320929.

[18] B. Rad, F. Song, V. Jacob and Y. Diao. 'Explainable Anomaly Detection on High-Dimensional Time Series Data'. In: The 15th ACM International Conference on Distributed and Event-based Systems (DEBS '21). virtual event, Italy, 28th June 2021. DOI: 10.1145/3465480.3468292. URL: https://hal.inria.fr/hal-03522878.

[19] 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.

[20] F. Song, K. Zaouk, C. Lyu, A. Sinha, Q. Fan, Y. Diao and P. Shenoy. 'Spark-based Cloud Data Analytics using Multi-Objective Optimization'. In: ICDE - 37th IEEE International Conference on Data Engineering. Chania, Greece, 19th Apr. 2021. URL: https://hal.inria.fr/hal-02549758.

Conferences without proceedings

[21] N. Barret. 'Facilitating Heterogeneous Dataset Understanding'. In: BDA 2021 - informal publication only. Paris, France, 25th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03344102.

[22] N. Barret, I. Manolescu and P. Upadhyay. 'Toward Generic Abstractions for Data of Any Model'. In: BDA 2021 - Informal publication only. Paris, France, 25th Oct. 2021. URL: https://hal.inria.fr/hal-03344041.

[23] T. Bouganim, H. Galhardas and I. Manolescu. 'Efficiently identifying disguised nulls in heterogeneous text data'. In: BDA (Conférence sur la Gestion de Données – Principles, Technologies et Applications). Paris, France, 25th Oct. 2021. URL: https://hal.inria.fr/hal-03347947.

Doctoral dissertations and habilitation theses

[24] L. Di Palma. 'New Algorithms and Optimizations for Human-in-the-Loop Model Development'. Institut Polytechnique de Paris, 7th July 2021. URL: https://hal.inria.fr/tel-03319743.

[25] P. Guzewicz. 'ExpRalytics: Expressive and Efficient Analytics for RDF Graphs'. École polytechnique, 6th Oct. 2021. URL: https://hal.inria.fr/tel-03478282.

[26] E. Huang. 'Active Learning Methods for Interactive Exploration on Large Databases'. Institut Polytechnique de Paris, 6th July 2021. URL: https://hal.inria.fr/tel-03339951.

[27] K. Zaouk. 'Neural-Based Modeling for Performance Tuning of Cloud Data Analytics'. Institut Polytechnique de Paris, 11th Mar. 2021. URL: https://tel.archives-ouvertes.fr/tel-03284173.

Reports & preprints

[28] K. Zaouk, F. Song, C. Lyu and Y. Diao. *Neural-based Modeling for Performance Tuning of Spark Data Analytics*. 20th Jan. 2021. URL: https://hal.inria.fr/hal-03116831.

Other scientific publications

[29] A. Chaoui. 'Improving a Search Engine for Answering User Questions in Natural Language'. University of Paris Saclay, 27th Sept. 2021. URL: https://hal.inria.fr/hal-03524281.

[30] T. Meunier. 'Automatic or semi-automatic detection of companies in difficulty or weakened by the crisis'. Ecole nationale des ponts et chaussées, 30th Aug. 2021. URL: https://hal.inria.fr/hal-03523010.

[31] M. Touzari. 'Réalisation d'un système Q&A spécialisé dans la qualité des eaux'. Paris: Sorbonne Université - Faculté des sciences, 1st Sept. 2021. URL: https://hal.inria.fr/hal-03523094.

10.3 Cited publications

[32] M. Besta, E. Peter, R. Gerstenberger, M. Fischer, M. Podstawski, C. Barthels, G. Alonso and T. Hoefler. 'Demystifying graph databases: Analysis and taxonomy of data organization, system designs, and graph queries'. In: *arXiv preprint arXiv:1910.09017* (2019).

[33] K. Kolluru, V. Adlakha, S. Aggarwal, Mausam and S. Chakrabarti. 'OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3748–3761. DOI: 10.18653/v1/2020.emnlp-main.306. URL: https://aclanthology.org/2020.emnlp-main.306.

[34] P. Macko, V. J. Marathe, D. W. Margo and M. I. Seltzer. 'Llama: Efficient graph analytics using large multiversioned arrays'. In: *2015 IEEE 31st International Conference on Data Engineering*. IEEE. 2015, pp. 363–374.

[35] J. Malicevic, B. Lepers and W. Zwaenepoel. 'Everything you always wanted to know about multicore graph processing but were afraid to ask'. In: *2017 USENIX Annual Technical Conference (USENIX ATC 17)*. 2017, pp. 631–643.